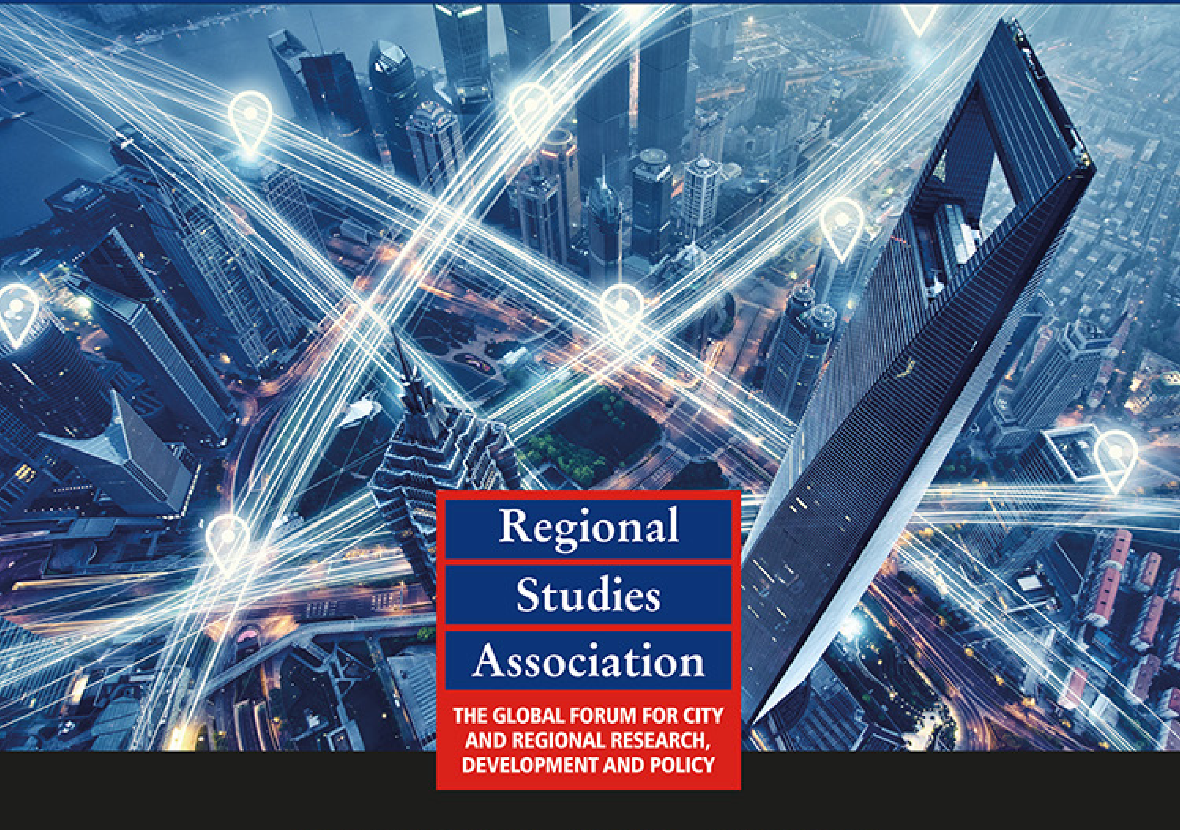
ROUTLEDGE

# DATA AND THE CITY



Regional
Studies
Association
THE GLOBAL FORUM FOR CITY AND REGIONAL RESEARCH, DEVELOPMENT AND POLICY

REGIONS AND CITIES

EDITED BY ROB KITCHIN, TRACEY P. LAURIAULT AND GAVIN McARDLE

# Data and the City

There is a long history of governments, businesses, science and citizens producing and utilizing data in order to monitor, regulate, profit from and make sense of the urban world. Recently, we have entered the age of big data, and now many aspects of everyday urban life are being captured as data and city management mediated through data-driven technologies.

*Data and the City* is the first edited collection to provide an interdisciplinary analysis of how this new era of urban big data is reshaping how we come to know and govern cities, and the implications of such a transformation. This book looks at the creation of real-time cities and data-driven urbanism and considers the relationships at play. By taking a philosophical, political, practical and technical approach to urban data, the authors analyse the ways in which data is produced and framed within socio-technical systems. They then examine the constellation of existing and emerging urban data technologies. The volume concludes by considering the social and political ramifications of data-driven urbanism, questioning whom it serves and for what ends.

This book, the companion volume to 2016's *Code and the City*, offers the first critical reflection on the relationship between data, data practices and the city, and how we come to know and understand cities through data. It will be crucial reading for those who wish to understand and conceptualize urban big data, data-driven urbanism and the development of smart cities.

**Rob Kitchin** is Professor and European Research Council (ERC) Advanced Investigator at Maynooth University, Ireland. He is also (co)Principal Investigator of the Programmable City project, the Building City Dashboards project, the All-Island Research Observatory (AIRO) and the Digital Repository of Ireland (DRI).

**Tracey P. Lauriault** is Assistant Professor of Critical Media and Big Data in the School of Journalism and Communication at Carleton University, Canada. She is also Research Associate with the Programmable City project at Maynooth University, Ireland, and the Geomatics and Cartographic Research Centre at Carleton University.

**Gavin McArdle** is Assistant Professor in the School of Computer Science at University College Dublin (UCD), Ireland. He is also Research Associate with the National Centre for Geocomputation (NCG) and the Programmable City project at Maynooth University, Ireland.

Regional Studies Association
THE GLOBAL FORUM FOR CITY AND REGIONAL RESEARCH, DEVELOPMENT AND POLICY

Routledge
Taylor & Francis Group

## Regions and Cities

### Series Editor in Chief

**Joan Fitzgerald**, *Northeastern University, USA*

**Editors**
**Maryann Feldman**, *University of North Carolina, USA*
**Gernot Grabher**, *HafenCity University Hamburg, Germany*
**Ron Martin**, *University of Cambridge, UK*
**Kieran P. Donaghy**, *Cornell University, USA*

In today's globalised, knowledge-driven and networked world, regions and cities have assumed heightened significance as the interconnected nodes of economic, social and cultural production, and as sites of new modes of economic and territorial governance and policy experimentation. This book series brings together incisive and critically engaged international and interdisciplinary research on this resurgence of regions and cities, and should be of interest to geographers, economists, sociologists, political scientists and cultural scholars, as well as to policy-makers involved in regional and urban development.

For more information on the Regional Studies Association visit www.regionalstudies.org

There is a **30% discount** available to RSA members on books in the ***Regions and Cities*** series, and other subject related Taylor and Francis books and e-books including Routledge titles. To order just e-mail Joanna Swieczkowska, Joanna.Swieczkowska@tandf.co.uk, or phone on +44 (0) 20 3377 3369 and declare your RSA membership. You can also visit the series page at www.routledge.com/Regions-and-Cities/book-series/RSA and use the discount code: **RSA0901**

**124 The Rural and Peripheral in Regional Development**
An Alternative Perspective
*Peter de Souza*

**123 In The Post-Urban World**
Emergent Transformation of Cities and Regions in the Innovative Global Economy
*Edited by Tigran Haas and Hans Westlund*

**122 Contemporary Transitions in Regional Economic Development**
Global Reversal, Regional Revival?
*Edited by Turok et al.*

**121 The Illicit and Illegal in Regional and Urban Governance and Development**
Corrupt Places
*Edited by Francesco Chiodelli, Tim Hall and Ray Hudson*

**120 The Political Economy of Capital Cities**
*Heike Mayer, Fritz Sager, David Kaufmann and Martin Warland*

**119 Data and the City**
*Edited by Rob Kitchin, Tracey P. Lauriault and Gavin McArdle*

**118 The Unequal City**
Urban Resurgence, Displacement and The Making of Inequality in Global Cities
*John Rennie Short*

**117 Urban Transformations**
Geographies of Renewal and Creative Change
*Edited by Nicholas Wise and Julie Clark*

**116 The Scottish Economy**
A Living Book
*Edited by Kenneth Gibb, Duncan Maclennan, Des McNulty and Michael Comerford*

**115 Reanimating Regions**
Culture, Politics, and Performance
*Edited by James Riding and Martin Jones*

**114 Territorial Policy and Governance**
Alternative Paths
*Edited by Iain Deas and Stephen Hincks*

**113 Economics of Planning Policies in China**
Infrastructure, Location and Cities
*Wen-jie Wu*

**112 The Empirical and Institutional Dimensions of Smart Specialisation**
*Edited by Philip McCann, Frank van Oort and John Goddard*

**111 EU Cohesion Policy**
Reassessing performance and direction
*Edited by John Bachtler, Peter Berkowitz, Sally Hardy and Tatjana Muravska*

**110 Geography of Innovation**
*Edited by Nadine Massard and Corinne Autant-Bernard*

**109 Rethinking International Skilled Migration**
*Edited by Micheline van Riemsdijk and Qingfang Wang*

**108 The EU's New Borderland**
Cross-border relations and regional development
*Andrzej Jakubowski, Andrzej Miszczuk, Bogdan Kawałko, Tomasz Komornicki, and Roman Szul*

**107 Entrepreneurship in a Regional Context**
*Edited by Michael Fritsch and David J. Storey*

106 **Governing Smart Specialisation**
*Edited by Dimitrios Kyriakou, Manuel Palazuelos Martínez, Inmaculada Periáñez-Forte, and Alessandro Rainoldi*

105 **Innovation, Regional Development and the Life Sciences**
Beyond clusters
*Kean Birch*

104 **Unfolding Cluster Evolution**
*Edited by Fiorenza Belussi and Jose Luis Hervás-Olivier*

103 **Place-based Economic Development and the New EU Cohesion Policy**
*Edited by Philip McCann and Attila Varga*

102 **Transformation of Resource Towns and Peripheries**
Political economy perspectives
*Edited by Greg Halseth*

101 **Approaches to Economic Geography**
Towards a geographical political economy
*Ray Hudson*

100 **Secondary Cities and Development**
*Edited by Lochner Marais, Etienne Nel and Ronnie Donaldson*

99 **Technology and the City**
Systems, applications and implications
*Tan Yigitcanlar*

98 **Smaller Cities in a World of Competitiveness**
*Peter Karl Kresl and Daniele Ietri*

97 **Code and the City**
*Edited by Rob Kitchin and Sung-Yueh Perng*

96 **The UK Regional–National Economic Problem**
Geography, globalisation and governance
*Philip McCann*

95 **Skills and Cities**
*Edited by Sako Musterd, Marco Bontje and Jan Rouwendal*

94 **Higher Education and the Creative Economy**
Beyond the campus
*Edited by Roberta Comunian and Abigail Gilmore*

93 **Making Cultural Cities in Asia**
Mobility, assemblage, and the politics of aspirational urbanism
*Edited by Jun Wang, Tim Oakes and Yang Yang*

92 **Leadership and the City**
Power, strategy and networks in the making of knowledge cities
*Markku Sotarauta*

91 **Evolutionary Economic Geography**
Theoretical and empirical progress
*Edited by Dieter Kogler*

90 **Cities in Crisis**
Socio-spatial impacts of the economic crisis in Southern European cities
*Edited by Jörg Knieling and Frank Othengrafen*

**89 Socio-Economic Segregation in European Capital Cities**
East meets West
*Edited by Tiit Tammaru, Szymon Marcińczak, Maarten van Ham, Sako Musterd*

**88 People, Places and Policy**
Knowing contemporary Wales through new localities
*Edited by Martin Jones, Scott Orford and Victoria Macfarlane*

**87 The London Olympics and Urban Development**
The mega-event city
*Edited by Gavin Poynter, Valerie Viehoff and Yang Li*

**86 Making 21st Century Knowledge Complexes**
Technopoles of the world revisited
*Edited by Julie Tian Miao, Paul Benneworth and Nicholas A. Phelps*

**85 Soft Spaces in Europe**
Re-negotiating governance, boundaries and borders
*Edited by Philip Allmendinger, Graham Haughton, Jörg Knieling and Frank Othengrafen*

**84 Regional Worlds: Advancing the Geography of Regions**
*Edited by Martin Jones and Anssi Paasi*

**83 Place-making and Urban Development**
New challenges for contemporary planning and design
*Pier Carlo Palermo and Davide Ponzini*

**82 Knowledge, Networks and Policy**
Regional studies in postwar Britain and beyond
*James Hopkins*

**81 Dynamics of Economic Spaces in the Global Knowledge-based Economy**
Theory and East Asian cases
*Sam Ock Park*

**80 Urban Competitiveness**
Theory and practice
*Daniele Letri and Peter Kresl*

**79 Smart Specialisation**
Opportunities and challenges for regional innovation policy
*Dominique Foray*

**78 The Age of Intelligent Cities**
Smart environments and innovation-for-all strategies
*Nicos Komninos*

**77 Space and Place in Central and Eastern Europe**
Historical trends and perspectives
*Gyula Horváth*

**76 Territorial Cohesion in Rural Europe**
The relational turn in rural development
*Edited by Andrew Copus and Philomena de Lima*

**75 The Global Competitiveness of Regions**
*Robert Huggins, Hiro Izushi, Daniel Prokop and Piers Thompson*

**74 The Social Dynamics of Innovation Networks**
*Edited by Roel Rutten, Paul Benneworth, Dessy Irawati and Frans Boekema*

73 **The European Territory**
From historical roots to global challenges
*Jacques Robert*

72 **Urban Innovation Systems**
What makes them tick?
*Willem van Winden, Erik Braun, Alexander Otgaar and Jan-Jelle Witte*

71 **Shrinking Cities**
A global perspective
*Edited by Harry W. Richardson and Chang Woon Nam*

70 **Cities, State and Globalization**
City-regional governance
*Tassilo Herrschel*

69 **The Creative Class Goes Global**
*Edited by Charlotta Mellander, Richard Florida, Bjørn Asheim and Meric Gertler*

68 **Entrepreneurial Knowledge, Technology and the Transformation of Regions**
*Edited by Charlie Karlsson, Börje Johansson and Roger Stough*

67 **The Economic Geography of the IT Industry in the Asia Pacific Region**
*Edited by Philip Cooke, Glen Searle and Kevin O'Connor*

66 **Working Regions**
Reconnecting innovation and production in the knowledge economy
*Jennifer Clark*

65 **Europe's Changing Geography**
The impact of inter-regional networks
*Edited by Nicola Bellini and Ulrich Hilpert*

64 **The Value of Arts and Culture for Regional Development**
A Scandinavian perspective
*Edited by Lisbeth Lindeborg and Lars Lindkvist*

63 **The University and the City**
*John Goddard and Paul Vallance*

62 **Re-framing Regional Development**
Evolution, innovation and transition
*Edited by Philip Cooke*

61 **Networking Regionalised Innovative Labour Markets**
*Edited by Ulrich Hilpert and Helen Lawton Smith*

60 **Leadership and Change in Sustainable Regional Development**
*Edited by Markku Sotarauta, Ina Horlings and Joyce Liddle*

59 **Regional Development Agencies: The Next Generation?**
Networking, knowledge and regional policies
*Edited by Nicola Bellini, Mike Danson and Henrik Halkier*

58 **Community-based Entrepreneurship and Rural Development**
Creating favourable conditions for small businesses in Central Europe
*Matthias Fink, Stephan Loidl and Richard Lang*

57 **Creative Industries and Innovation in Europe**
Concepts, measures and comparative case studies
*Edited by Luciana Lazzeretti*

56 **Innovation Governance in an Open Economy**
Shaping regional nodes in a globalized world
*Edited by Annika Rickne, Staffan Laestadius and Henry Etzkowitz*

55 **Complex Adaptive Innovation Systems**
Relatedness and transversality in the evolving region
*Philip Cooke*

54 **Creating Knowledge Locations in Cities**
Innovation and integration challenges
*Willem van Winden, Luis de Carvalho, Erwin van Tujil, Jeroen van Haaren and Leo van den Berg*

53 **Regional Development in Northern Europe**
Peripherality, marginality and border issues
*Edited by Mike Danson and Peter De Souza*

52 **Promoting Silicon Valleys in Latin America**
*Luciano Ciravegna*

51 **Industrial Policy Beyond the Crisis**
Regional, national and international perspectives
*Edited by David Bailey, Helena Lenihan and Josep-Maria Arauzo-Carod*

50 **Just Growth**
Inclusion and prosperity in America's metropolitan regions
*Chris Benner and Manuel Pastor*

49 **Cultural Political Economy of Small Cities**
*Edited by Anne Lorentzen and Bas van Heur*

48 **The Recession and Beyond**
Local and regional responses to the downturn
*Edited by David Bailey and Caroline Chapain*

47 **Beyond Territory**
*Edited by Harald Bathelt, Maryann Feldman and Dieter F. Kogler*

46 **Leadership and Place**
*Edited by Chris Collinge, John Gibney and Chris Mabey*

45 **Migration in the 21st Century**
Rights, outcomes, and policy
*Kim Korinek and Thomas Maloney*

44 **The Futures of the City Region**
*Edited by Michael Neuman and Angela Hull*

43 **The Impacts of Automotive Plant Closures**
A tale of two cities
*Edited by Andrew Beer and Holli Evans*

42 **Manufacturing in the New Urban Economy**
*Willem van Winden, Leo van den Berg, Luis de Carvalho and Erwin van Tuijl*

41 **Globalizing Regional Development in East Asia**
Production networks, clusters, and entrepreneurship
*Edited by Henry Wai-chung Yeung*

40 **China and Europe**
The implications of the rise of China as a global economic power for Europe
*Edited by Klaus Kunzmann, Willy A Schmid and Martina Koll-Schretzenmayr*

39 **Business Networks in Clusters and Industrial Districts**
The governance of the global value chain
*Edited by Fiorenza Belussi and Alessia Sammarra*

38 **Whither Regional Studies?**
*Edited by Andy Pike*

37 **Intelligent Cities and Globalisation of Innovation Networks**
*Nicos Komninos*

36 **Devolution, Regionalism and Regional Development**
The UK experience
*Edited by Jonathan Bradbury*

35 **Creative Regions**
Technology, culture and knowledge entrepreneurship
*Edited by Philip Cooke and Dafna Schwartz*

34 **European Cohesion Policy**
*Willem Molle*

33 **Geographies of the New Economy**
Critical reflections
*Edited by Peter W. Daniels, Andrew Leyshon, Michael J. Bradshaw and Jonathan Beaverstock*

32 **The Rise of the English Regions?**
*Edited by Irene Hardill, Paul Benneworth, Mark Baker and Leslie Budd*

31 **Regional Development in the Knowledge Economy**
*Edited by Philip Cooke and Andrea Piccaluga*

30 **Regional Competitiveness**
*Edited by Ron Martin, Michael Kitson and Peter Tyler*

29 **Clusters and Regional Development**
Critical reflections and explorations
*Edited by Bjørn Asheim, Philip Cooke and Ron Martin*

28 **Regions, Spatial Strategies and Sustainable Development**
*David Counsell and Graham Haughton*

27 **Sustainable Cities**
*Graham Haughton and Colin Hunter*

26 **Geographies of Labour Market Inequality**
*Edited by Ron Martin and Philip S. Morrison*

**25 Regional Innovation Strategies**
The challenge for less-favoured regions
*Edited by Kevin Morgan and Claire Nauwelaers*

**24 Out of the Ashes?**
The social impact of industrial contraction and regeneration on Britain's mining communities
*Chas Critcher, Bella Dicks, David Parry and David Waddington*

**23 Restructuring Industry and Territory**
The experience of Europe's regions
*Edited by Anna Giunta, Arnoud Lagendijk and Andy Pike*

**22 Foreign Direct Investment and the Global Economy**
Corporate and institutional dynamics of global-localisation
*Edited by Jeremy Alden and Nicholas F. Phelps*

**21 Community Economic Development**
*Edited by Graham Haughton*

**20 Regional Development Agencies in Europe**
*Edited by Charlotte Damborg, Mike Danson and Henrik Halkier*

**19 Social Exclusion in European Cities**
Processes, experiences and responses
*Edited by Judith Allen, Goran Cars and Ali Madanipour*

**18 Metropolitan Planning in Britain**
A comparative study
*Edited by Peter Roberts, Kevin Thomas and Gwyndaf Williams*

**17 Unemployment and Social Exclusion**
Landscapes of labour inequality and social exclusion
*Edited by Sally Hardy, Paul Lawless and Ron Martin*

**16 Multinationals and European Integration**
Trade, investment and regional development
*Edited by Nicholas A. Phelps*

**15 The Coherence of EU Regional Policy**
Contrasting perspectives on the structural funds
*Edited by John Bachtler and Ivan Turok*

**14 New Institutional Spaces**
TECs and the remaking of economic governance
*Martin Jones, Foreword by Jamie Peck*

**13 Regional Policy in Europe**
*S. S. Artobolevskiy*

**12 Innovation Networks and Learning Regions?**
*James Simmie*

**11 British Regionalism and Devolution**
The challenges of state reform and European integration
*Edited by Jonathan Bradbury and John Mawson*

**10 Regional Development Strategies**
A European perspective
*Edited by Jeremy Alden and Philip Boland*

**9 Union Retreat and the Regions**
The shrinking landscape of organised labour
*Ron Martin, Peter Sunley and Jane Wills*

**8 The Regional Dimension of Transformation in Central Europe**
*Grzegorz Gorzelak*

**7 The Determinants of Small Firm Growth**
An inter-regional study in the United Kingdom 1986–90
*Richard Barkham, Graham Gudgin, Mark Hart and Eric Hanvey*

**6 The Regional Imperative**
Regional planning and governance in Britain, Europe and the United States
*Urlan A. Wannop*

**5 An Enlarged Europe**
Regions in competition?
*Edited by Louis Albrechts, Sally Hardy, Mark Hart and Anastasios Katos*

**4 Spatial Policy in a Divided Nation**
*Edited by Richard T. Harrison and Mark Hart*

**3 Regional Development in the 1990s**
The British Isles in transition
*Edited by Ron Martin and Peter Townroe*

**2 Retreat from the Regions**
Corporate change and the closure of factories
*Stephen Fothergill and Nigel Guy*

**1 Beyond Green Belts**
Managing urban growth in the 21st century
*Edited by John Herington*

# Data and the City

## Edited by Rob Kitchin, Tracey P. Lauriault and Gavin McArdle

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2018
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
711 Third Avenue, New York, NY 10017

*Routledge is an imprint of the Taylor & Francis Group, an informa business*

© 2018 selection and editorial matter, Rob Kitchin, Tracey P. Lauriault and Gavin McArdle; individual chapters, the contributors

The right of Rob Kitchin, Tracey P. Lauriault and Gavin McArdle to be identified as the authors of the editorial material, and of the authors for their individual chapters, has been asserted in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

*Trademark notice*: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

*British Library Cataloguing in Publication Data*
A catalogue record for this book is available from the British Library

*Library of Congress Cataloging in Publication Data*
Names: Kitchin, Rob, editor. | Lauriault, Tracey P., editor. | McArdle, Gavin, editor.
Title: Data and the city / edited by Rob Kitchin, Tracey P. Lauriault and Gavin McArdle.
Description: Abingdon, Oxon ; New York, NY : Routledge, 2018. | Includes index.
Identifiers: LCCN 2017014690| ISBN 9781138222625 (hardback) | ISBN 9781138222632 (pbk.) | ISBN 9781315407388 (ebook)
Subjects: LCSH: Cities and towns—Statistics. | Urbanization—Statistics. | City planning—Statistical methods. | City planning—Data processing.
Classification: LCC HT153 .D27 2018 | DDC 307.76—dc23
LC record available at https://lccn.loc.gov/2017014690

ISBN: 978-1-138-22262-5 (hbk)
ISBN: 978-1-138-22263-2 (pbk)
ISBN: 978-1-315-40738-8 (ebk)

Typeset in Times New Roman
by Swales & Willis Ltd, Exeter, Devon, UK

# Contents

**PART III**
**Urban data technologies** 109

**PART IV**
**Urban data cultures and power** 187

# Figures

# Tables

# Contributors

**Pouria Amirian**, Ordnance Survey, Southampton, UK.

**Anahid Basiri**, Department of Geography and Environment, University of Southampton, UK.

**Jo Bates**, Information School, University of Sheffield, UK.

**Michael Batty**, Centre for Advanced Spatial Analysis, University College London, UK.

**Craig Dalton**, Department of Global Studies and Geography, Hofstra University, New York, USA.

**Mordechai (Muki) Haklay**, Department of Geography, University College London, UK.

**Rob Kitchin**, NIRSA, National University of Ireland, Maynooth, Ireland.

**Tracey P. Lauriault**, School of Journalism and Communication, Carleton University, Canada.

**Deborah Maxwell**, Department of Theatre, Film and Television, University of York, UK.

**Gavin McArdle**, School of Computer Science, University College Dublin, Ireland.

**Dietmar Offenhuber**, Art + Design, Public Policy and Urban Affairs, Northeastern University, Massachusetts, USA.

**Larissa Pschetz**, School of Design, University of Edinburgh, UK.

**Evelyn Ruppert**, Department of Sociology, Goldsmiths, University of London, UK.

**Teresa Scassa**, Faculty of Law, University of Ottawa, Canada.

**Chris Speed**, School of Design, University of Edinburgh, UK.

**Till Straube**, Department of Human Geography, Goethe University Frankfurt, Germany.

**Jim Thatcher**, Urban Studies, University of Washington – Tacoma, USA.

**Martijn de Waal**, Lectorate of Play and Civic Media, Amsterdam University of Applied Sciences, The Netherlands.

**James Merricks White**, NIRSA, National University of Ireland, Maynooth, Ireland.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

# 1 Data and the city

*Rob Kitchin, Tracey P. Lauriault and Gavin McArdle*

## Introduction

There is a long history of governments, businesses, science and citizens producing and utilizing data in order to monitor, regulate, profit from, and make sense of the urban world. Data have traditionally been time-consuming and costly to generate, analyse and interpret, and generally have provided static, often coarse, snapshots of urban phenomena. Recently, however, we have entered the age of big data, with data related to knowing and governing cities increasingly becoming a deluge; a wide, deep torrent of timely, varied, resolute and relational data (Kitchin 2014a; Batty 2016). This has been accompanied by an opening up of state data, and to a much lesser degree, business data, the production of volunteered geographic information, and the emergence of open data cultures and practices (Goodchild 2007; Bates 2012). As a result, evermore aspects of everyday life – work, consumption, travel, communication, leisure – and the worlds we inhabit are being captured and stored as data, made sense of through new data analytics, mediated through data-driven technologies, normalized through data-driven infrastructures, and shared through data infrastructures and data brokers (Amoore 2013; Kitchin 2014b; Offenhuber and Ratti 2014).

This data revolution has produced multiple challenges that require critical and technical attention – how best to produce, manage, analyse and act on urban big and open data, make sense of data infrastructures, data cultures and practices, and understand their consequences with respect to city governance, economy, politics and everyday life. However, to date, there has been relatively little *critical* reflection on the new emerging relationship between data and the city, and how we come to know and understand cities through data in the present era.

In the rush to create so-called ‘smart cities’, wherein core city services and infrastructures become digitally mediated and data-driven – generating, processing and acting on data in real-time to algorithmically manage systems and calibrate performance – much of the attention has been on how to technically create and implement suitable smart city technologies, and associated institutional and infrastructural supports such as data standards, protocols, policies, and a variety of telecom networks. Such data-driven technologies include: urban control rooms, e-government systems, city operating systems, coordinated emergency response systems, intelligent transport systems, integrated ticketing, real-time passenger information, smart parking, fleet and logistics management, city dashboards, predictive policing, digital surveillance, energy smart grids, smart meters, smart lighting, sensor networks, building management systems and a wide plethora of locative and spatial media. Collectively these technologies are generating an ever-growing tsunami of indexical data (uniquely linked to people, objects, territories, transactions) that can be repurposed in diverse ways – for example, in predictive profiling and social sorting of citizens and neighbourhoods, creating urban models and simulations, for policing and security purposes, etc. (CIPPIC 2006; Batty 2013; Kitchin 2014b; 2016). These data are in addition to large quantities of administrative and statistical data, more traditional sampled survey data, polling and public opinion data, and any other data the city may collect as part of reporting and delivering services.

Rather less attention has been paid to more epistemological, normative, ethical and political questions concerning how data-driven cities and urban issues are framed and approached; how city development and progress are envisaged; what kinds of data are being produced and to what purposes they are being employed; what kinds of cities we ideally want to create and live in (not simply from an instrumental perspective – solving particular issues such as traffic congestion; but with respect to issues such as fairness, equity, justice, citizenship, democracy and governance); how these data-driven technologies and processes work in practice on the ground; what kinds of social and spatial relations they produce; whom they benefit and disadvantage or exclude; what kinds of subjectivity, citizenship, participation and political action they support; and how they reshape many aspects of urban life. This is not to say that there has been no consideration of such questions – as the chapters that follow and the work they reference attest, there is a growing body of research that critically examines urban data and their use. However, the work to date is still relatively formative in theoretical and empirical terms, often considers data-driven systems within the context of smart cities in general terms rather than focusing specifically on the unfolding relationship between data and cities, and the development and rollout of data-driven urbanism is largely outpacing critical reflection and interventions.

## Data and the city

This volume is designed to help to fill this lacuna through an interdisciplinary examination of the relationship between data and contemporary urbanism. The focus is not smart city technologies per se, but rather the essays concentrate on how to make sense of urban data and the emerging era of data-driven urbanism. As well as providing synoptic analyses and new conceptual thinking, the chapters detail a number of illustrative examples of urban data, data-driven systems and related issues, including data infrastructures, urban blockchains, mapping, urban modelling, data provenance, data quality, data citizenship, citizen science, data practices, data cultures, data frictions and city dashboards. Importantly, given the wide-ranging, diverse and complex relationship between data and the city, and the need to bring various expertise and knowledge into dialogue, the contributors are drawn from a number of disciplines (Geography, Geographic Information Science, Planning, Sociology, Information Science, Design, Media Studies, Law and Computer Science).

All but three of the chapters were prepared initially for a workshop at the National University of Ireland Maynooth in September 2015, funded by the European Research Council through an Advanced Investigator Award to Rob Kitchin for The Programmable City project (ERC-2012-AdG-323636-SOFTCITY). Each essay was pre-prepared and submitted in advance of the meeting, then extensively discussed at the workshop, and subsequently revised for publication. While the book is designed to work as a standalone text, there is a companion book, *Code and the City* (Kitchin and Perng 2016), that focuses predominately on the relationship between software and the city. To provide a structure, we have divided the book into four parts.

## *Data-driven cities*

The first part considers the relationship between data and the city in a broad sense, focusing on the creation of real-time cities and data-driven urbanism and how the ever-greater flows of data are transforming city services, infrastructures, urban life and how we understand and govern cities.

In the opening chapter, Martijn de Waal examines the creation of ‘real-time cities’, wherein computation is embedded into the fabric of cities producing real-time data flows that can be used to know and manage city services in the here-and-now. He argues that such data-driven systems are changing how we understand cities in three ways. The first is the adoption of an action-orientated epistemology wherein the production of real-time data, along with machine learning techniques, enables a new kind of scientific knowledge about cities that treats them as complex systems which can be made actionable through smart city technologies. The second approach is more critical in orientation and, on the one hand, challenges the scientific principles and epistemology of the first, and on the other, considers more ontological questions concerning how real-time data and data-driven systems transform the production of space, the nature of place, and the experience of living in the city. The third approach asks more normative questions and argues that cities cannot be conceptualized and approached as being analogous to other complex systems, such as galaxies and rainforests, because they are social-cultural-political in nature. Instead, it is contended that a new science of cities needs to frame data-driven cities with respect to wider concerns about the kinds of cities we want to create and how to produce particular kinds of ‘cityness’. De Waal argues that more attention needs to be paid to this third kind of knowledge making and its praxes.

Mike Batty considers the nature of urban big data and the epistemological challenges of using them to make sense of the city, placing his discussion in historical context. Adopting an approach that is perhaps characterised as fitting within de Waal’s first mode of understanding data-driven cities, Batty argues that we have always been struggling to extract insights from ever-larger and more dynamic data as urban technologies evolve and urban computational research struggles to keep up. He notes that what might be considered small data – sampled in time, space and by category – soon become very large once the interactions between data points are examined. Using the concept of a data cube, Batty examines the characteristics of urban flow data between locations. In particular, he illustrates his arguments by detailing the difficulties of making sense of traditional transport interaction data, such as origin (home) to destination (work) flows across a city, and more dynamic and massive datasets, such as the tap-in and tap-outs of travellers on the London Underground (one of his datasets consists of nearly 10 billion records generated over 86 days in the summer of 2012). In both cases, urban science is still struggling to extract and communicate meaningful insight. He concludes that rather than abandoning theory for an empiricist form of data science, there is a pressing need to develop a theoretically insightful urban science.

In his chapter, Rob Kitchin argues that while there has long been forms of urbanism that are data-informed, a new era of data-driven urbanism unfolding as cities become ever more instrumented and networked, their systems interlinked and integrated, and vast troves of big urban data are being generated and used to manage and control urban life in real-time. He contends that data-driven urbanism is the key mode of production for what have widely been termed smart cities. Adopting an approach that largely maps onto de Waal's third approach, Kitchin critically examines a number of urban data issues, including: the politics of urban data and production of data assemblages; data ownership, data control, data coverage and access; the creation of buggy, brittle, hackable urban systems (data security, data integrity); and social, political, ethical effects (data protection and privacy, dataveillance, and data uses including social sorting and anticipatory governance). He concludes that whilst data-driven urbanism purports to produce a common-sense, pragmatic, neutral, apolitical, evidence-based form of responsive urban governance, it is nonetheless selective, crafted, flawed, normative and politically inflected. Consequently, whilst data-driven urbanism provides a set of solutions for urban problems, it does so within limitations and in the service of particular interests or there is an overreliance on mathematically and engineered models that do not factor in a city's social, cultural, historical, institutional and political complexities; those very things that give cities their character.

## *Urban data*

The second part focuses attention on the nature of urban data, examining them from ontological, political, practical and technical points of view. Importantly, the analysis does not conceive of urban data from a common-sense, essentialist position, wherein they are seen to faithfully and validly represent the state of the world, but rather consider the ways in which data are produced and framed within socio-technical systems.

Teresa Scassa provides a critical overview of crime data and their sharing through open data sites, interactive visualizations, and other media. She details how crime data are far from neutral, objective records of criminal, policing and legal activity, but rather are shaped significantly by legal, institutional and cultural factors. She argues that crime data are subjective and contested, record certain kinds of information but excludes others, and are known to be full of gaps and errors. Moreover, capturing, analysing and acting upon crime data requires human interpretation and judgement, framed with societal and institutional contexts. And yet, despite these issues, crime data are often taken at face value and are used to drive social, policing, security and legal policy and programmes and to underpin new interventions such as predictive policing. While the data do hold value and are important in revealing levels of crime and society's institutional response, she contends that they need to be treated with caution, with users considering how, by whom, and for what purposes the data were generated to gauge their veracity and trustworthiness.

Jim Thatcher and Craig Dalton similarly consider the issues of data veracity and trustworthiness by considering data provenance. They note that data provenance is presently largely instrumental in nature and concerns information about the production and history of a dataset. Such information allows users to know how the data were captured, by whom, using what techniques and technologies, how they were processed and handled, and so on, enabling them to judge their quality, shortcomings and suitability for use. Typically, such information is stored as a metadata – that is, data about the data. However, they contend that such an instrumental approach to data provenance is limited and too technically orientated, ignoring the wider context in which the data are produced and used. Instead, they suggest the use of a more-than-technical form of provenance that not only documents traditional metadata, but also includes situated contextual factors such as motivation, value and power. They formulate this version of data provenance as the recording of 'data encounters' which capture the always already-cooked nature of data and the contextual nature of its use.

Jim Merricks White likewise is interested in data encounters, but rather than focus on provenance, he seeks to follow data from their generation through to their various uses, exposing how they are cleaned, recombined and put to work. Using an empirical example of infant mortality and their use in city indicator initiatives he charts the translation and circulation of data, seeking to document what he terms 'data threads', highlighting the entanglement of data infrastructures and geography, and their inherent materiality and relationality. He traces how infant mortality data are generated by messy human and computational practices shaped by a framework of definitions and standards. These data are then used in varying ways, reworked to create new derived data, and used in ways not anticipated with respect to their original generation. He notes that the devastating loss of a child's life is rendered first as trace, then as data point, and then as input to derivative calculations and distant ambitions, in this case various health and city indicator initiatives. With each transformation, he argues the data become increasingly alienated from their material associations and their meaning mutate to reflect new discourses and ideologies. Comparing his notion of data threads to that of 'data journeys' detailed by Bates *et al.* (2016), White provides a useful epistemological avenue for thickening the description of data assemblages and how data translate and are woven together across such assemblages.

Considering the nature of urban data further, Dietmar Offenhuber examines what makes urban data meaningful, the extent to which data are always cooked and never raw, and concerns with respect to the repurposing data. Utilizing the concept of 'data friction' he examines the issues that arise when data and metadata generated by different organizations, that utilize different formats and standards, are moved or bought into contact. He notes that despite difficulties and limitations, data sets can develop a life of their own and be repurposed in diverse ways, often as data proxies for other phenomena. Offenhuber examines these issues with respect to Twitter data, which have become widely used in social science research, and satellite imagery generated by the Operational Linescan System (OLS) of the US Air Force's Defense Meteorological Satellite Program (DMSP). He contends that Twitter data, despite its widespread repurposing, are 'sticky data', that is meaningful when discussed in their original context, but problematic to interpret, extrapolate and generalize otherwise. In contrast, OLS/DMSP data are relatively non-sticky, being used extensively to identify city street lighting and act as a proxy for population density and economic activity, though it is not without problems. Offenhuber thus concludes that as proxies for urban phenomena, both data sources offer only partial perspectives and need to be used with caution.

## *Urban data technologies and infrastructures*

The third part examines the constellation of existing and emerging urban data technologies and infrastructures. The chapters explore a range of political, practical and technical issues and epistemological and theoretical approaches with respect to building, operating and making sense of such data-driven systems.

One way in which a plethora of urban data are made sense of by city managers and shared with citizens is through city dashboards that provide a variety of visualization and analytic tools which enable these data to be explored. While such dashboards provide useful tools for evaluating and managing urban services, understanding and formulating policy, and creating public knowledge and counter-narratives, Rob Kitchin and Gavin McArdle's analysis reveals a number of conceptual and practical shortcomings. They critically examine six issues with respect to the building and use of city dashboards: epistemology, scope and access, veracity and validity, usability and literacy, use and utility, and ethics. Drawing on their experience of building the Dublin Dashboard, they advocate a shift in thinking and praxis that openly situates the epistemology and instrumental rationality of city dashboards and addresses more technical shortcomings.

Pouria Amirian and Anahid Basiri also consider the sharing and analysis of urban big data, though their focus is more technical in nature. Given the wide variety of different data-driven platforms being utilized across a number of organizations and domains, and the need to be able to share and integrate such data so they can be used by many systems and actors, it is necessary to create platform-independent principles and mechanisms to ensure interoperability. They contend that such interoperability is best achieved through Service Orientation Principles (SOP) along with a new architecture, Organizational Service Layer, that uses polyglot binding. They detail three core SOP approaches, and their benefits and shortcomings, currently being utilized to share data and analysis (Web Services, RESTful services and Geoservices), as well detailing how four types of bindings can be used to provide loose couplings between backend implementation and other software applications. These bindings enable platform independency and agile and straightforward communication between systems, thus creating accessible, flexible, scalable and interoperable smart city platforms and more easily implementable city data portals, urban control rooms and city dashboards.

An alternative and emerging form of data infrastructure for city dashboards and services are blockchains. Blockchains are sealed and encrypted distributed ledgers of all transactions ever conducted within a system. Each block records key metadata regarding a transaction such as information about sender and receiver, time, value, fees and IP address, and once recorded cannot be altered, thus creating trust. Each block adds to the sequence of transactions forming a chain that leads back to the start of the database. While blockchains are most commonly associated with new financial currencies such as Bitcoin, Chris Speed, Deborah Maxwell and Larissa Pschetz examine their utility for recording and sharing other kinds of transactions. To illustrate how blockchains work as economic, social and cognitive ledgers they discuss their use with regards to finance and work. They then detail the development of two prototype city ledgers produced in a design workshop that utilize Bitcoin technology demonstrating how blockchains offer opportunities to capture diverse social practices and transactions in city ledgers. They contend that the blockchain has the potential to create trusted city ledgers (databases), and thus trusted city dashboards, and provide the foundation for dealing with complexity and predicting future outcomes.

Rather than focus on the form, operation, building and shortcomings of building data infrastructures, Till Straube focuses on how best to theoretically and empirically make sense of them. He proposes a materialist approach to understanding the constitution and work of data infrastructures and data-driven systems. Instead of concentrating on the relational effects of such infrastructures – how they produce space–time compression or a space of flows – he argues that attention needs to be paid to the materiality and spatiality of the infrastructures themselves (programming languages, database software, data formats, protocols, APIs, etc.). Such a focus, he argues, foregrounds data technologies and infrastructures, their make-up and practices, and how they are materially embedded into the fabric of cities and everyday lives. The approach he advocates is a topological reading of data technologies, underpinned by assemblage theory. Here, emphasis is put on charting the network of relations between potentially dispersed socio-technical systems, rather than the topography of their territories; that is, it is concerned with material connections and power relations that operate across and produce a relational rather than Euclidean space. He thus forwards an epistemology, what he terms an applied materialist topology, that seeks to pay close attention to how data technologies and infrastructures articulate, perform and translate time-spaces within a socio-political context. Such an approach also makes clear that as well as having a materiality, data technologies are never neutral in formulation, operation and effects.

Tracey Lauriault also aims to make sense of data infrastructures and offers a nested methodological approach to study the power/knowledge of data models and ontologies. Drawing on ethnographic work in which she was embedded at Ordnance Survey Ireland (OSi) examining how the organization introduced a fundamentally new data model, Prime2, that replaced a map layers model with an object-orientated model, she considers how cities are captured within data models and how these models transduce the city. She advances three interlinked methodological approaches for making sense of the diverse range of empirical materials she amassed, including interviews, technical documents, procedure and training manuals, databases, in-field observation and news reports. The first is the application of her modified version of Hacking's dynamic nominalism to assess how the city is 'made-up' through the new Prime2 ontology. The second is a genealogical mapping of the development of the Prime2 data model over time and the key events in its production. The third is an application of Kitchin's socio-technical assemblage as a framing tool to study how the model constitutes one part of a national spatial data infrastructure. She argues that using these methodological approaches together enables an unpacking of the discursive and material production of data models and data infrastructures and how these models and infrastructures produce space.

## *Urban data cultures and power*

The fourth part considers the social and political configurations of urban data infrastructures and data-driven systems and who they are operated by, their purposes and who they serve. Far from being neutral and objective in nature and serving the public good in a general sense, this part examines their data cultures and data power.

As Jo Bates notes, data do not arise from nowhere. Rather, data are produced by people and technology embedded within socio-material relations situated within time and space. They are the result of data practices and modes of data governance operating within specific data cultures. In other words, data production and use is shaped by cultural norms, value systems and beliefs, as well as the wider political economy and institutional and legal landscape. Data cultures, and their sites of practice and governance, are historically constituted, dynamic, open and porous, and thus mutate over time. Bates notes that for each city there are a multitude of interrelated data cultures operating within and across public organizations, private enterprise and civic bodies, though these cultures are not all created equal, with some dominating and subverting others. She argues that it is important to unpack these data cultures and their sites of data practice to reveal their assumptions, values, participants, rhetorical and material work, the power dynamics at play, how they shape the domain on which they operate, and how they interconnect with other data cultures. In so doing, the inherent politics and power of such systems are revealed, enabling us to challenge and reconsider how they are conceived and work in practice.

Given that data about the city and its citizens are produced to enable the functioning of city systems, monitor and regulate populations, to underpin markets, or to provide counter-systems they are fundamentally instruments of power and capital (even when they seemingly enable diverse communication, communities and play – there are always inclusions and exclusions in their production and whom benefits from their operation). They therefore raise important questions concerning citizenship and political subjectivity in the digital age. Evelyn Ruppert examines the extent to which people are data subjects or data citizens in the contemporary era and how data citizenship is constituted. She argues that to understand the relationship between data and the city necessitates asking political questions concerning the framing, identity and positioning of digital subjects and the conduits of power that systems work to reproduce. The data of cities, she notes, are produced by technologies in the employ of public institutions and companies that confer differing forms of citizenship, though these are not accepted uncontested. This is evident in ongoing debates concerning the production of big data and surveillance, privacy, confidentiality, anonymity, security, policing, governance and data markets. Rather than focusing on the substantive nature of digital data rights, Ruppert concentrates on who are the subjects of these rights, their political subjectivity, and the role of subjects in the making and shaping of data, developing a theory of data citizens.

Muki Haklay approaches the question of citizenship through the emergence of citizen science. His starting questions are to ask: whether the future being produced within the smart city vision by data-driven technologies is the one citizens want? And whether such technologies integrate and foster meaningful and purposeful social and communal activities or create feelings of alienation? His concern is that smart city systems represent the interests of city governments and corporate interests and focus on instrumental issues rather than human values and desires; on technocratic constraints and management rather than imagination and serendipity. He explores these issues drawing on the ideas of Albert Borgmann, especially those relating to the difference between device paradigms (instrumental, technically mediated engagements) and focal practices (meaningful social engagement). He argues that smart city technologies tend towards the former, being automated and autocratic, whereas as citizen science initiatives tend towards the latter, being more social and community engaged. He thus argues for a more open, democratic and participatory vision of data-driven city systems in which people play an active role as citizens, not simply subjects. Such meaningful participatory and collective action centred on focal practices, he contends, has the potential to transform the present smart city paradigm.

## Future agendas

Taken together, the chapters highlight the diverse ways in which data and cities are becoming ever more intertwined, transforming how we come to know, manage, govern and live in cities. There are several themes that cut across the essays and in conclusion we want to highlight three that we believe require particular theoretic and empirical attention.

## *Data politics and power*

Collectively, the chapters that follow make a compelling argument that urban data are always cooked and never raw, and the data-driven systems and infrastructures that produce, manage, share and act on them are socio-technical systems not simply technical ones. Urban data and systems then are never neutral, objective and common-sensical, but rather are inherently political – invested with values and judgements, are formed and operated within cultural milieu, and are designed to produce certain effects. This is as much the case for initiatives that seek to be inclusive and enable citizen-engaged data projects, such as open data sites, as it is for systems designed for state surveillance or corporate profit. Certainly, many data-driven city systems and their data practices work to manage, regulate and control urban activities; they inherently capture certain kinds of data and use them to enact particular power relations. Much rhetorical and material work is invested in reproducing the logic and legitimacy of these systems, for example through smart city discourses, but they always remain open to resistance, subversion and transgression.

*Data and the City* performs important work, we believe, in exploring urban data politics, cultures and power. However, there is still much empirical and theoretical research needed to unpack the specific ways in which data are cooked and utilized to perform political work, however subtle that may be – to examine: how data are generated, processed, shared, translated and used; how data cultures form and are reproduced; how data practices operate within and across networks of actors and data-driven technologies; how data-driven systems produce political subjectivity and data citizens; how data cultures, politics and practices create ethical dilemmas, especially with regards to dataveillance and the work of data brokers; and the forms and practices of alternative data-driven systems that seek to enact more participatory and emancipatory politics. Moreover, further research is required to understand how data influence digital labour, investigating issues such as how institutional and organizational structures change with the introduction of new databased regimes, how data ecosystems change government and corporate work practices, and how the database managers and data scientists become more important within institutions with their knowledge and expertise becoming privileged over others.

## *Epistemology*

As the chapters make clear, there are a diverse set of epistemologies being deployed to make sense of urban data, data-driven systems, and the relationship between data and the city. This varies from more computational approaches, such as the urban science practised by Batty and the technical mapping of data-driven systems by Amirian and Basiri, through to the Kitchin's unpacking of data assemblages, Straube's applied materialist topology, Lauriault's modified use of dynamic nominalism and genealogical approach, Merrick White's strategy of following data threads, Bates's mapping of data cultures, and Haklay's charting of citizen science. These are by no means the only epistemological approaches being used, as illustrated by de Waal's chapter. The sheer variety of disciplinary and philosophical traditions, technologies and issues make this epistemological diversity inevitable, and we would not be in favour of trying to advocate for a single epistemological paradigm. We do, however, believe that much more attention needs to devoted to the epistemological challenge of providing useful insights into urban data systems and infrastructure and data-driven urbanism.

These challenges include trying to make sense of highly dynamic, complex and capricious domains that are full of various actors and actants, interlinked systems, diverse practices and processes, competing politics and interests, and are often black-boxed (in terms of the technical processes, but also institutional access). Moreover, these domains work across scales from single devices to entire cities. Indeed, there is a need for a balance between detailed and empirically rich mappings of individual systems that tease apart their complex relations and workings, and wider synoptic analysis of how these data-driven systems and cultures are working together or in conflict to produce data-driven urbanism. The pressing task then is to, on the one hand, develop conceptual tools for making sense of data-driven technologies and urbanism, their architecture and workings, and the transformations they are producing, and on the other to identify suitable methodologies for grounding such tools through empirical research. While the chapters provide some useful starting points, building on longer legacies of related research, it is clear that there is much more to be done.

### *Normative questions*

Most analysis of urban data and data-driven city systems grapple little with normative questions – that is, consider in-depth questions about how things should be as opposed to how they are, or for whom and what purpose are data-driven cities being created? Instead, analysis is concerned with detailing how systems are configured and work, either from a technical or social perspective. For those that develop such systems, the goals are usually defined in instrumental terms – to make a city more sustainable, resilient, efficient, secure, competitive, and so on. As a consequence, a fundamental question such as 'what kind of cities do we want to create and live in?' has largely been framed technically and instrumentally, rather philosophically in relation to issues such as fairness, equity, justice, citizenship and democracy. By highlighting issues such as data cultures, data power and data citizens, the chapters in *Data and the City* point to the need to consider wider normative questions about the goals of data-driven urbanism and whose interests they should serve. For example, should data-driven systems be primarily about creating new markets and profit? Facilitating state control and regulation? Improving the quality of life of citizens? Or all three and in what balance? And in what form should they be conceived and implemented? Exploring, debating and answering these normative questions is important because they frame how the urban data revolution will unfold and how policy and law will need to be formulated to produce the kinds of cities desired. In fact, framing the debate in instrumental terms has been a useful rhetorical strategy for avoiding such normative considerations because it shifts the debate into a post-political and seemingly common-sensical register. We believe it is time to challenge such a positioning.

## Conclusion

As we have noted above, there are many political, ethical, epistemological and normative questions still to be asked and answered with respect to urban data, data-driven city systems and urbanism, yet the urban data revolution continues to unfold at pace. There is thus a pressing need for empirical research and conceptual thought to make sense of the changes taking place. Collectively, we believe the chapters in *Data and the City* provide a productive set of routes into thinking about these questions that help advance our understanding of the evolving relationship between data and urban life and forms of data-driven urbanism. As such, it adds to an emerging interdisciplinary body of work and should hopefully make for an illuminating and stimulating read.

## Acknowledgements

The research for this chapter and the Data and the City workshop were funded by a European Research Council Advanced Investigator Award to Rob Kitchin, entitled ‘The Programmable City’ (ERC-2012-AdG-323636-SOFTCITY).

## References

Amoore, L. (2013) *The Politics of Possibility: Risk and Security Beyond Probability*. Durham, NC: Duke University Press.

Bates, J. (2012) ‘This is what modern deregulation looks like: Co-optation and contestation in the shaping of the UK’s Open Government Data Initiative’, *The Journal of Community Informatics* 8(2), available from: www.ci-journal.net/index.php/ciej/article/view/845/916 [accessed 6 February 2017].

Bates, J., Lin, Y-W. and Goodale, P. (2016) ‘Data journeys: Capturing the socio-material constitution of data objects and flows’, *Big Data & Society* 3(2): 1–12.

Batty, M. (2013) *The New Science of Cities*. Cambridge, MA: MIT Press.

Batty, M. (ed.) (2016) ‘Big data and the city.’ Special issue of *Built Environment* 42(3), available from: www.alexandrinepress.co.uk/built-environment/big-data-and-city.

CIPPIC (2006) *On the Data Trail: How detailed information about you gets into the hands of organizations with whom you have no relationship. A Report on the Canadian Data Brokerage Industry*. The Canadian Internet Policy and Public Interest Clinic, Ottawa. https://cippic.ca/sites/default/files/May1-06/DatabrokerReport.pdf [accessed 6 February 2017].

Goodchild, M.F. (2007) 'Citizens as sensors: the world of volunteered geography', *GeoJournal* 69: 211–221.

Kitchin, R. (2014a) 'The real-time city? Big data and smart urbanism', *GeoJournal* 79(1): 1–14.

Kitchin, R. (2014b) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences.* London: SAGE.

Kitchin, R. (2016) *Getting smarter about smart cities: Improving data privacy and data security*. Data Protection Unit, Department of the Taoiseach, Dublin, Ireland, available from: www.taoiseach.gov.ie/eng/Publications/Publications_2016/Smart_Cities_Report_January_2016.pdf.

Kitchin, R. and Perng, S.Y. (eds) (2016) *Code and the City*. New York: Routledge.

Offenhuber, D. and Ratti, C. (eds) (2014) *Decoding the City: Urbanism in the Age of Big Data.* Basel: Birkhauser.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

# Part I

# Data-driven cities

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

# 2 A city is not a galaxy

## Understanding the city through urban data

*Martijn de Waal*

## Introduction

In a 2013 report to the UK Economic and Social Research Council, Michael Batty, the director of the Centre for Advanced Spatial Analysis (CASA), looks back at the time when computers were first being used in urban planning:

> Fifty years ago if you had asked the question 'what can we do with computers with respect to cities?' the answer would have been we can build computer models of cities – abstractions – that can then be used to pose conditional questions such as 'What If . . .'
>
> (Batty 2013a: 22)

Half a century later, Batty argues this vision has been turned inside out. Computers are no longer seen as mere tools to analyse the city, rather they have become part of the city, embedded into its very fabric. From electronic tolling on roads, to CCTV cameras with facial recognition detection, to buildings managed by software systems, to citizens wielding their cell phones to find a nearby restaurant, computers have become active agents in the shaping of urban life.

The rise of these various urban computing systems has contributed to what Rob Kitchin (2014b: xiii) has called a 'data revolution' – the availability of 'a wide, deep torrent of timely, varied, resolute and relational data that are relatively low in cost and, outside of business, increasingly open and accessible'. From citizens posting on social networks to traffic data aggregated by navigation service providers, a constellation of computer systems has started to generate a broad variety of real-time 'urban data', producing what some have called the 'real-time city', wherein the city can be known and managed in the here-and-now through control rooms and urban dashboards (Townsend 2008; Kitchin 2014c; Kloeckl *et al.* 2012).

This chapter explores how the creation of the so-called 'real-time city' is changing our understanding of cities and creating new scientific approaches to urban studies. At least three different (partially overlapping) ways of understanding the city through urban data have emerged. The first can be understood as a new 'action oriented epistemology' of the city. Researchers in academia and business consultancy have started to claim that real-time data can give us a new kind of knowledge in which cities can be understood as complex systems, not unlike galaxies or rainforests. In turn, these insights can be made actionable through the deployment of smart city technologies. A second approach has a more critical and often also an ontological orientation and seeks to understand the production and experience of urban space mediated by computation. The third approach has focused on normative theories of urban culture at large. The main argument to be made here is that cities are different from other complex systems such as galaxies or rainforests, in that they are social-cultural-political systems that can be framed and evaluated normatively. After all, it is humans themselves that set – and can change – many of the rules that govern urban life. What kind of city do we want to live in? And what do we make of the changes brought about by the various assemblages that employ software and urban data to manage urban life in new ways? Besides providing us insights into the workings of a city, as with the first approach, it is contended that a 'new science of cities' should play a role in addressing these kinds of questions.

## An action oriented epistemology

In 2008, in the introduction to a seminal anthology on the then newly emerging discipline of urban informatics, Anthony Townsend proclaims that the rise of real-time urban data might lead to a paradigm shift in the way we understand our cities:

> if aerial photography showed us the muscular and skeletal structure of the city, the revolution in urban informatics is likely to reveal it's circulatory and nervous systems. I like to call this vision the 'real-time city' because for the first time we'll see cities as a whole the way biologists see a cell – instantaneously and in excruciating detail but also alive . . .
>
> (Townsend 2008: xxvi)

More recently, Townsend notes that this line of thought has given rise to at least a dozen new academic labs, departments and schools that explore this new understanding of the city (Townsend 2015b). What is remarkable is that many of these institutes are not grounded in disciplines such as planning or urban sociology, but – as Townsend alluded in 2008 – rather seek inspiration in biology, physics and astrophysics. A case in point: the director of Singapore's Future Cities Lab was trained as a rainforest ecologist; the director of the Centre for Urban Science and Progress in New York is a physicist. What these new institutes seek, according to Townsend is to pursue 'deeply quantitative and computational approaches to understanding the city' (Townsend 2015a; 2015b).

The ecological and physical understanding of cities that we find in the new science of cities is not completely new. The beginning of the twentieth century already witnessed scientists like the evolutionary biologist Patrick Geddes starting to map cities in order to gain an 'objective' understanding of them. Likewise, the sociologists of the Chicago School in the 1920s were inspired by evolutionary theories, and sought to understand the 'human ecology' of cities as a complex system (Sennett 1969; Park 1969). A second wave of this approach emerged with the rise of cybernetics after the Second World War. The social problems of cities, it was believed by, amongst others, the newly founded United States Department of Housing and Urban Development (HUD), could be tackled by modelling cities with the aid of computers. One of the projects in this program was one of the first geodemographic profiling systems, designed by Jonathan Robbin, that was later turned into the commercial PRIZM database of zip-code based lifestyle clusters (Burrows and Gane 2006). However, the enthusiasm for the models waned quickly when they failed to live up to their promises. Both the data sets used as well as the models were too crude and received much criticism (Lee Jr 1973; Townsend 2015a, 2015b).

What is new this time around is the availability of massive amounts of real-time data generated by all kinds of assemblages of hardware, software, algorithms and institutions in the city itself, plus increased computational power and data analytics utilizing machine learning. Batty (2013a) argues that these may change the logic of the city and at the same time could give us a new understanding of cities as complex systems in which the decisions of millions of heterogeneous individual actors add up to a hard-to-understand system that nevertheless seems to have an order. This system is not static: as cities grow, they also change qualitatively, yet how exactly remains undertheorized. This new understanding is based on flows and networks, shifting our thinking from the city as a system in place to a system in time (Batty 2013a, 2013b).

Whereas some of the new institutes addressing the city as a system of flows are mainly oriented toward finding new theoretical models to understand the city, others are linking the new insights the real-time city may produce to an agenda of urban improvement and citizen empowerment. ‘Giving people visual and tangible access to real-time information about their environment’, claim Nabian and Ratti (2012: 76), ‘enables them to make decisions that are more in sync with what is actually happening around them.’ The research projects in their Senseable City Lab aim to explore this idea. For instance, their project Trash Track, that reveals the ecology of trash collection and waste disposal by adding tracking sensors to items that are thrown away, gives clues on how to ‘create a more efficient removal chain’. In addition, the data could be used by local governments ‘to promote behavioral changes among its citizens’ (Ratti and Townsend 2011: 45).

Outside academia we have seen somewhat related (but not completely similar) claims by professional communities working on theories for the now widely discussed ‘smart city’ (Allwinkle and Cruickshank 2011; Caragliu *et al.* 2011; Hemment and Townsend 2013; de Waal 2014; Kourtit *et al.* 2012). As a research paper by IBM proclaims:

> Smart Cities provide a new form of instrumentation for observing in fine detail the way that people use the city and so may enable new approaches to theories of cities. Through new sources of information cities hope to create insight, innovation, opportunity and real jobs that will increase prosperity and quality of life.
>
> (Harrison and Donnelly 2011: 5)

Here, the city is mainly understood as a series of infrastructural services that can be optimized by better understanding their dynamics.

According to some proponents of this trend, this will lead to a paradigm shift in our knowledge about cities. Rob Kitchin (2014a) has described how a new empiricist school of thought has emerged that takes the data these computer systems are generating at face value to produce direct insights in (amongst others) urban patterns. As one of their protagonists, Chris Anderson (2008), claims:

> We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot . . . Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

In their vision, Batty's observation about the use of computers in the city has come full circle: computer systems produce data about the city that allegedly give us a transparent look into the city's dynamics. In turn, these data can be used to analyse the city in order to optimize that system. In the feedback loop that emerges from these assemblages of computer systems, the institutions that manage them and their users, the software might even start fine-tuning its own algorithms, producing a new form of an autopoieitic city (Kryssanov *et al.* 2002).

Where the new empiricists see a system that makes the workings of a city more transparent, their critics point to the fact that they overlook the social and ideological dimensions active in the production of data through these systems (Kitchin 2014b; Greenfield 2013; Batty 2013a; Hollands 2008; de Waal 2014; see Chapters 2, 5, 10, 12, 15). As such, these systems may reinforce social, political and economic power relations rather than providing ways to challenge them or come up with alternatives.

In addition, Batty argues, whereas the new science of cities could provide new insights into the complexity of the city and the feedback loops between individual agents and the workings of the system as a whole, many of the smart city approaches seem to reduce the city to a set of seemingly simple technical problems that can be monitored, analysed and solved, falling into the trap of modernism. Rather, he claims cities should be understood as wicked problems:

> Moreover the notion that urban problems are simple to solve should by now have been dispelled for the experience in everything from garden cities to green belts, from the provision of public housing to the provision of transport systems over the last 50 to 100 years, has been salutary and sobering. Problems in cities are 'wicked' in the terminology of Rittel and Webber (1973) in that they are more likely to get worse than better if you attempt to address them in directly obvious ways which seek simple solutions. The smart city movement has to yet address this question.
>
> (Batty 2013a: 11)

To be fair, a number of actors involved in smart city developments – especially some local governments – have not been deaf to these criticisms, and have started to look for alternative approaches. One of them is bringing 'smart citizens' to the processes of knowledge production about the city (Saunders and Baeck 2015; Hemment and Townsend 2013). Others have invited citizens to the design process of interventions, turning the city into living labs (Concilio *et al.* 2013; Pallot *et al.* 2010; Friedrich *et al.* n.d.; Coenen *et al.* 2014). One way these programs now try to evade the modernist trap is to use urban data to bring various perspectives to the table in the design process. For example, outside academia, an alternative urban data movement has emerged in the form of citizen sensing communities who frame urban issues and the categories of data they need to get a hold on them in different ways (see Chapter 11). A number of new labs and city programs have currently moved into this direction. Here, data are not understood as indexical registrations of urban reality, but as social constructs. Data can be used to produce insights into urban issues, but they are understood as potentially contested, and produced in relation to social and political conditions.

## Data and the production and experience of urban space

What unites the approaches discussed so far is that they seek a new understanding of the city through the analysis of urban data in order to intervene in the city. What has not been fully addressed is this: the rise of urban data does not only produce a new way of making sense of the city. At the same time, the interventions that result from the analysis of these data may produce new kinds of spatial organizations and experiences. If computers are indeed used to run the city, rather than merely to analyse it, this could lead to new spatial regimes. What are the power structures and ideologies operative in the ways data are collected, categorized, analysed and acted upon? How does that affect the production and experience of place and our ability to act?

A number of scholars have taken up this issue in their research, resulting in a variety of approaches that seek to understand the role of software and urban data in the production of urban space and the urban experience. To do this, a theory is needed that understands the city as a complex system and provides ways to detangle the various individual actors and their discursive, social, political and social contexts coming together in the production of space. In this line of thought, Kitchin and Dodge (2005) argue that space needs to be theorized as 'ontogenetic'. (Urban) space should not be understood as a given, but is continuously reproduced through the interaction of various actors.

> This ontogenetic conception of space acknowledges that the forms and spatial relations of the world around us are clearly not static and fixed; they are constantly being altered, updated, and constructed in ways that alter sociospatial relations. . . . space is not a container with pregiven attributes frozen in time; rather, space gains its form, function, and meaning in practice.
>
> (Kitchin and Dodge 2005: 171–172)

In their article 'Code and the transduction of space' they provide a number of examples of the way software has become a part of this process. Code they assert 'mediates, supplements, augments, monitors, regulates, operates and facilitates many everyday task and routines related to domestic living, travel, work, communication, and consumption' (Kitchin and Dodge 2005). However, this code is not a given or a neutral factor, but it is produced in assemblages of institutions, governments, companies and/or individuals that seek to manage particular processes with particular aims, as well as the discursive and material practices and the economic and political context around these processes. Following Latour, they describe how these (power) relations may become encoded into the software, and how in turn this software plays a role in the constitution of material and discursive practices.

The strength of this approach is that it allows for the un-black-boxing of the production and use of software by 'following the actors' involved. It seeks to understand the situational context in which tools are produced and used. In *The Data Revolution*, Kitchin (2014b) develops a somewhat similar framework for the understanding of (urban) data. Data itself can be thought of as produced in particular assemblages: 'amalgams of systems of thought, forms of knowledge, finance, political economies, governmentalities and legalities, materialities and infrastructures, practices, organisations and institutions, subjectivities and communities, places, and marketplaces' (Kitchin 2014b: xiv), and as such can be problematized as one of the factors contributing in the production of space.

This ontogenetic approach is a welcome addition to the epistemological ones described above, as it forwards the construction of data in complex assemblages and opens up the debate for critical understandings of smart city epistemologies. As such it fits within a growing interest in urban studies towards ANT methodologies (Farias and Bender 2010), bringing in the production and consumption of data as one of the many aspects of urban assemblages.

Whereas these approaches focus on (our understanding of) the production of space, and the role of data in it, others have started to focus on the role of urban data in the experience of space, shifting the vantage point from the production of space to the experience of the subject. Urban space, as has been theorized widely, should now be understood as 'hybrid' (de Souza e Silva 2006), meaning that its experience is no longer confined to the physical conditions of a particular site, but now includes the networks of communication that can be tapped into through a variety of devices. Gordon and de Souza e Silva (2011: 2) have called this 'networked locality', or 'net locality':

> Net locality implies a ubiquity of networked information – a cultural approach to the web of information as intimately aligned with the perceptual realities of everyday life. We don't enter the web anymore; it is all around us.

On the one hand these theories foreground the connective affordance of mobile media networks, providing the ability to connect with others in remote locations. On the other hand, a variety of digital media interfaces also provide us access to (real-time) information about the city or to stored representations of experiences. In fact, De Souza e Silva and Gordon argue, the mobile phone is a device that turns our urban experiences into data, allowing others to access these in real-time or at a later point in time, either as individual experiences or aggregated in data sets or streams that could reveal particular urban conditions – as for instance in live traffic information that partly consists out of the aggregation of data generated by the networks of mobile phone providers.

As Leighton Evans (2015) has pointed out, this availability of communication and information networks should not only be understood instrumentally – as in a way to solve a particular problem by accessing information databases, say: where can I find a nearby restaurant? These devices also change our experience of being in place in a phenomenological sense, revealing the world 'poetically' (Evans 2015), meaning that they can be used to make meaning of a particular space. A number of researchers have found for instance how the notion of 'presence' changes (Okabe and Ito 2006; Matsuda 2006), or how citizens use mobile devices to 'tune' their experience of place, by tuning in and out layers of information revealed by their devices (Coyne 2010). Evans himself has shown how devices, such as a smart phone, allow users to 'dwell' (find themselves at home) even in so called non-places (Auge 1995) through the connectivity of their devices (Evans 2015). Evans (2015: 6) makes use of Jameson's concept of 'social cartography' to explain how these technologies can be used 'as a means of understanding and regaining a capacity to act'. The apps, maps, social graphs and network updates on our mobile devices provide us with access to all kinds of urban data, enabling particular ways to act, and this changes our experience of urban space.

These studies can be seen in a broader framework of urban studies that has taken an interest in 'situatedness' that goes all the way back to (at least) the studies of Goffman in the 1950s (Goffman 1959). As Goffman demonstrated in his work, subjects take clues from their surroundings as to what cultural codes are present, and they might attune their behaviour accordingly. As theories on performativity have shown, in turn every instantiation of these clues into a particular behaviour, speech act, or more lasting act of (re)design, reinforces the cultural code as others might take it as another clue. At the same time, this system is never stable. Dominant codes may be challenged radically, or the performance of them might be modulated in more subtle ways, leading to gradual change. Similarly, other theories have given us insights in how people develop a sense of place, including feelings of belonging to a particular place (Geertz 2000; Gordon 2008).

What is new in our present-day experience of space is that these contexts and behaviours are no longer limited to the physical scene. Our situatedness is mediated through mobile media networks, giving us on the one hand clues that are absent in the physical location, and at the same time turning our performances into data that can be circulated within these networks, both within and outside the original situated contexts.

The theories referred to here provide us with an inroad to redefine the ontology of the urban experience as one that is partly constituted through this production, processing and representation of urban data. On the one hand, these theories allow us to understand our situated experience of and performances in urban spaces as partly mediated through urban media systems and their data. On the other hand, they also allow us to un-black-box these media systems and the role of urban data in them to understand what power or interest may be operative in them, and how these might produce a particular urban spatiality.

## Urban data and urban culture

So far, we have seen how the 'revolution' in urban data has led to new ways to gain insight into the city as a complex system, as well as to new understandings of the production of place and our experience of it. The first may provide us with important insights in the complex logic of urban systems. Likewise, theories that seek to unpack the production and experience of place are valuable in their own right and give us tools to understand our cities and evaluate and critique the way it is constantly being remade. A third approach that focuses on urban culture at large could complement these new understandings of the city. Here the issue at stake is the often normative question: what makes a city 'work' as a social-cultural system? What makes cities different from other kinds of complex systems such as galaxies or rainforests? In other words: what constitutes the 'cityness' in the city? And how is this 'cityness' affected now that production and experience of space has become hybridized and the computational production and analysis of urban data have started to play a role in them?

Discussion on what it is that makes a city a special form of social organization go back to the German and Chicago schools of sociology that flourished in the first quarter of the previous century, when for instance Robert Park wrote: 'The city is not in other words merely a physical mechanism and an artificial construction. It is involved in the vital process of the people who compose it. It is a product of nature, and particularly of human nature' (Park 1969: 91). As many have argued, what makes cities special is that they consist of constellations of strangers: people who do not know each other, not personally nor categorically, yet who have to find a way to live together (Simmel 1969; Jacobs 2000; Blokland 2003; Lofland 1973; 1998). That condition is both an opportunity (these strangers are potential customers for our services, they might teach us something we would like to learn, bring us excitement, love or consolation), as well as a challenge (can we trust these strangers? Will they not thwart our ambitions?).

Cultural critics have argued that the city is a cultural system that balances these two sides of the equation. As Lewis Mumford wrote:

> Now, the great function of the city is . . . to permit, indeed to encourage and incite the greatest possible number of meetings, encounters, challenges between all persons, classes and groups providing, as it were, a stage upon which the drama of social life may be enacted, with the actors taking their turns as spectators, and the spectators as actors.
>
> (Mumford, cited in Goldberger 2004)

Mumford describes the city both as a market place and as a theatre. In the city, supply and demand in various spheres of life are brought together spatially. Strangers come together physically so they can interact. At the same time, these interaction spaces function as theatres: we act out our lives for others to see. This is how we get familiar with the strangers around us and it provides us with opportunities to identify or distance ourselves from them (Jacobs 2000; Blokland 2003). At the same time, the 'sets' or the 'scenes' we find ourselves in, may give us clues as to how to behave there.

More contemporary, we find a similar argument in the work of Manuel Castells (2002: 382):

> Cities have always been communication systems, based on the interface between individual and communal identities and shared social representations. It is their ability to organize this interface materially in forms, in rhythms, in collective experience and communicable perception that makes cities producers of sociability, and integrators of otherwise destructive creativity.

Particular social practices, Castells argues, become spatially institutionalized ('a material organization') at particular places in the city, making the chaos of urban experience legible and manageable: by experience we learn what to expect where in the city. At the same time, these forms and rhythms produce collective experiences that in some way or another connect all those individuals that co-inhabitate the city into a community of strangers (Boomkens 1998). To put it in the words of Paul Goldberger (2001): 'The role of the city . . . is to be a common place, to be common ground, and as such, to support us and to stimulate us . . . the urban impulse is an impulse toward community – an impulse toward being together, and toward accepting the idea that however different we may be, something unites us.'

What unites these theories is that they see the urban public sphere as a crucial ingredient that makes cities 'work', that integrate the otherwise 'destructive creativity'. The organization of a broad variety of urban practices in the nineteenth and early twentieth centuries produced a type of space that brought strangers together and allowed them to interact, and at the same time experience a sense of community. To quote Goldberger (2001) once more:

> In a sense, [the city] is the original Internet, the original hyperlink – since cities are places in which random connections, rather than linear order, often determines what will happen. Cities aren't linear, even though they exist in real space. Random connections are what make them work, and surprise and a sense of infinite choice is what gives them their power.

What's interesting in these theories, is that they understand the city itself as an interface, as a mechanism that through its spatial organization connects its citizens to each other, producing amongst others trust and a sense of community, as well as economic opportunities. Cityness then, lies in the spatial organization of density and diversity, and the somewhat chaotic interaction that results from it, as well as in the social goods that this may produce: solidarity, creativity, innovation, trust, community. What's important in this line of thinking is that it's best understood as a normative assumption: a city works best when it functions as an open and somewhat disorderly system with ample public spaces as catalysts for chance encounters (Sennett 2001; Jacobs 2000; Boomkens 1998; 2006), as it is in these public spaces that 'urban publics' emerge (de Waal 2014).

As we have seen in the previous paragraphs, the datafication of urban life and the emergence of the real-time city have led to digital interfaces that have started to represent what Castells referred to as the rhythms, collective experience and shared social representations. Even more so: these interfaces have started to function as the market places and theatre spaces through which citizens perform part of their lives and forge connections with others. If Castells's city can be understood as an offline interface that produces urban publics, our digital interfaces have taken over some of the functions of the city. Whether it is finding a date through Tinder, a ride through Uber, a power drill to borrow through Peerby, funders through Kickstarter, or a plumber through Taskrabbit, the network society has been turning into a platform society. To come back to Batty's insight: computers are now not just tools that automate and optimize existing urban functions such as traffic flows, they have partially taken over essential characteristics of the cityness we find in cities: their functioning as a 'market place' and a 'theatre'.

Research in cultural geography or urban sociology has just started to give us some first insights into what this may mean for the way our urban societies come into being. An interesting example is for instance Manuel Tironi's research into the experimental music scene in Santiago de Chile. Tironi describes how the practitioners in this subculture do not have a single hangout in the city, a site where their practices have materialized and is recognizable for both in- and outsiders as the locus of a particular scene. In the course of three years Tironi counted 45 different venues in which the scene performed and numerous rehearsal spaces scattered throughout the city. Yet, he did find the scene had a central meeting place: the online platform Myspace:

> MySpace has become the scene's place of publicness. In the absence of a geographical realm in which competing agents can map out the industry's innovations, MySpace has become the site where the members of the scene can watch each other, check their innovations and hear their new products – to defy, emulate or transubstantiate them. To be sure, MySpace is not just a promotional platform, a social network on which the scene can observe itself. MySpace, more radically, is a condition of possibility for the scene.
>
> (Tironi 2010: 46)

Members of the scene used the digital media platform to plan and communicate activities, to meet up with the likeminded as well as to perform by uploading audio samples. The website functioned as a market place and a theatre space, and

Tironi goes as far as to state that without this platform the scene could not have come into existence at all.

The point here is that whereas a digital platform has taken over some of the central functions of the 'cityness' we find in our cities, it does not substitute spatial activities. Members of the experimental music scene still meet up for live gigs and rehearsal sessions. But it does lead to a different, networked cultural geography and new ways of building trust and community organized through the datafications of musical activities and the algorithms of the platform.

As of yet, we are unsure as to what kind of urban culture this may produce. Various scenarios abound. On the one hand, this could lead to 'software sorted cities', where public spaces no longer function as Goldberger 'original hyperlinks', forging random connections between citizens. Rather, the platforms will connect us mainly with the likeminded and guide us to those places in the city where we will encounter them (Shepard and Greenfield 2007; Graham 2005; Pariser 2011). On the other hand, this could lead to a situation in which, as William Mitchell predicted in the end of the 1990s, function starts to follow code (Mitchell 2005). Once a connection is made through a digital platform, a physical meeting could take place anywhere in the city, rather than in a specialized location per se. And particular locations could be usurped by a variety of urban publics at the same time. As such, this might increase the density and heterogeneity of our cities even further (de Waal 2014).

## Conclusion

It is beyond the scope of this chapter to provide a full overview of the possible implications of the datafication of urban life in the constitution of urban culture. In fact, much more research is needed in this domain, and the new ways of understanding the city brought out in the first two sections can help to get a better understanding of these processes. The point here is to demonstrate that the 'data revolution' in urban life should be approached from multiple perspectives. The emergence of computing systems in our everyday life does not just produce a new way of understanding our cities, and new ways in which space is produced and experienced. It may also bring about a shift in the 'cityness' of our cities, in the functioning of our urban culture. In that respect, cities are different from galaxies or rainforests. As Rob Kitchin has pointed out, many of the epistemological approaches 'wilfully ignore the metaphysical aspects of human life (concerned with meanings, beliefs, experiences) and normative questions (ethical and moral dilemmas about how things should be as opposed to how they are)' (Kitchin 2014b: 136).

Cities and almost everything they behold from masterplans to algorithms are different from galaxies in that they are human constructs, that partially operate according to rules and laws created as outcomes of political (rather than natural) processes. As such, these rules are not given, but can be adapted to the normative preferences of those in power. Likewise, the urban data and algorithms that operate upon them are not neutral products that in a natural way produce a particular instance of urban culture, but they can be attuned to a particular normative idea of cityness. In the end, insights in the workings of the city as a complex system should be combined with such normative discussions: what kind of city do we want to live in, and what do we make of these changes brought about by the various assemblages that employ software and urban data to manage urban life in new ways? A new science of cities should therefore contribute to our understanding of cities in addressing these kinds of questions as well.

## References

Allwinkle, S. and Cruickshank, P. (2011) 'Creating smart-er cities: an overview', *Journal of Urban Technology* 18(2): 1–16.

Anderson, C. (2008) 'The end of theory: The data deluge makes the scientific method obsolete', *Wired* magazine online, 23 June, available at: www.wired.com/2008/06/pb-theory/.

Auge, M. (1995) *Non-places: Introduction to an Anthropology of Supermodernity*. London: Verso.

Batty, M. (2013a) *Urban Informatics and Big Data. A Report to the ESRC Expert Group.* London: CASA, UCL, available from: www.spatialcomplexity.info/files/2015/07/Urban-Informatics-and-Big-Data.pdf [accessed 24 November 2016].

Batty, M. (2013b) *The New Science of Cities*. Cambridge, MA: MIT Press.

Blokland, T. (2003) *Urban Bonds: Social Relationships in an Inner City Neighbourhood*. Cambridge, UK: Polity.

Boomkens, R. (1998) *Een drempelwereld: moderne ervaring en stedelijke openbaarheid*. Rotterdam: NAi Uitgevers.

Boomkens, R. (2006) *De nieuwe wanorde. Globalisering en het einde van de maakbare samenleving*. Amsterdam: Van Gennep.

Burrows, R. and Gane, N. (2006) 'Geodemographics, software and class', *Sociology* 40(5): 793–812.

Caragliu, A., del Bo, C. and Nijkamp, P. (2011) 'Smart cities in Europe', *Journal of Urban Technology* 18(2): 65–82.

Castells, M. (2002) 'The culture of cities in the information age', in I. Susser and M. Castells (eds), *The Castells Reader on Cities and Social Theory*. Malden, MA: Blackwell, pp. 367–389.

Coenen, T., van der Graaf, S. and Walravens, N. (2014) 'Firing up the city: a smart city living lab methodology', *Interdisciplinary Studies Journal (Special Issue on Smart Cities)* 3(4): 118–128.

Concilio, G., Puerari, E. and Rizzo, F. (2013) *Living Labs Models for co-designing in Urban and Public Space*. Report from My Neighborhood | My City Research Project.

Coyne, R. (2010) *The Tuning of Place, Sociable Spaces and Pervasive Digital Media*. Cambridge, MA: MIT Press.

de Souza e Silva, A. (2006) 'Mobile technologies as interfaces of hybrid spaces', *Space and Culture* 9(3): 261–278.

Evans, L. (2015) *Locative Social Media: Place in the Digital Age*. New York: Palgrave Macmillan.

Farias, I. and Bender, T. (2010) *Urban Assemblages: How Actor-Network Theory Changes Urban Studies*. New York: Routledge.

Friedrich, P., Karlsson, A. and Federley, M. (n.d.) *Report 2.1 Boundary conditions for successful Urban Living Labs*. Report from SubUrbanLab research project.

Geertz, C. (2000) *Local Knowledge: Further Essays in Interpretive Anthropology*. New York: Basic Books.

Goffman, E. (1959) *The Presentation of Self in Everyday Life*. New York: The Overlook Press.

Goldberger, P. (2001) 'Cities place and cyberspace', lecture at University of California, Berkeley, available from: www.paulgoldberger.com/lectures/cities-place-and-cyberspace/ [accessed 24 November 2016].

Goldberger, P. (2004) 'New School University Convocation', available from: www.paulgoldberger.com/lectures/new-school-university-convocation/ [accessed 24 November 2016].

Gordon, E. (2008) 'Towards a theory of networked locality', *First Monday* 13(10).

Gordon, E. and de Souza e Silva, A. (2011) *Net Locality: Why Location Matters in a Networked World*. Malden, MA: Wiley-Blackwell.

Graham, S. (2005) 'Software-sorted geographies', *Progress in Human Geography* 29(5): 562–580.

Greenfield, A. (2013) *Against the Smart City*. New York: Do projects.

Harrison, C. and Donnelly, I.A. (2011) 'A theory of smart cities', in *Proceedings of the 55th Annual Meeting of the ISSS – 2011, Hull, UK*, available from: http://journals.isss.org/index.php/proceedings55th/article/view/1703 [accessed 24 November 2016].

Hemment, D. and Townsend, A. (2013) *Smart Citizens*. Manchester: FutureEverything Publications.

Hollands, R.G. (2008) 'Will the real smart city please stand up? Intelligent, progressive or entrepreneurial?', *City* 12(3): 303–320.

Jacobs, J. (2000) *The Death and Life of Great American Cities*. London: Pimlico.

Kitchin, R. (2014a) 'Big data, new epistemologies and paradigm shifts', *Big Data & Society* 1(1): 1–12.

Kitchin, R. (2014b) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: SAGE.

Kitchin, R. (2014c) 'The real-time city? Big data and smart urbanism', *GeoJournal* 79(1): 1–14.

Kitchin, R. and Dodge, M. (2005) 'Code and the transduction of space', *Annals of the Association of American Geographers* 95(1): 162–180.

Kloeckl, K., Senn, O. and Ratti, C. (2012) 'Enabling the real-time city: LIVE Singapore!', *Journal of Urban Technology* 19(2): 89–112.

Kourtit, K., Nijkamp, P. and Arribas, D. (2012) 'Smart cities in perspective – a comparative European study by means of self-organizing maps', *Innovation: The European Journal of Social Science Research* 25(2): 229–246.

Kryssanov, V., Okabe, M., Kakusho, K. and Minoh, M. (2002) 'Communication of social agents and the digital city – a semiotic perspective', in K. Tanabe, P. van den Besselaar and T. Ishida (eds), *Digital Cities II: Computational and Sociological Approaches*. Berlin: Springer, pp. 56–70.

Lee Jr, D.B. (1973) 'Requiem for large-scale models', *Journal of the American Institute of Planners* 39(3): 163–178.

Lofland, L. (1973) *A World of Strangers: Order and Action in Urban Public Space*. New York: Basic Books.

Lofland, L.H. (1998) *The Public Realm : Exploring the City's Quintessential Social Territory*. Hawthorne, NY: Aldine de Gruyter.

Matsuda, M. (2006) 'Mobile communication and selective sociality', in M. Ito, D. Okabe and M. Matsuda (eds), *Personal, Portable, Pedestrian: Mobile Phones in Japanese Life*. Cambridge, MA: MIT Press, pp. 123–142.

Mitchell, W. (2005) *E-topia: 'Urban life, Jim – But Not as We Know It'*. Cambridge, MA and London: MIT Press.

Nabian, N. and Ratti, C. (2012) 'Top-down/bottom-up urbanism', in D. Offenhuber and K. Schechtner (eds), *Inscribing a Square: Urban Data as Public Space*. Vienna: Springer, pp. 76–79.

Okabe, D. and Ito, M. (2006) 'Technosocial situations: emergent structuring of mobile e-mail use', in M. Ito, D. Okabe and M. Matsuda (eds), *Personal, Portable, Pedestrian: Mobile Phones in Japanese Life*. Cambridge, MA: MIT Press, pp. 257–276.

Pallot, M., Trousse, B., Senach, B. and Scapin, D. (2010) 'Living Lab research landscape: From user centred design and user experience towards user cocreation', in *First European Summer School 'Living Labs'*, Paris.

Pariser, E. (2011) *The Filter Bubble: What the Internet is Hiding from You*. New York: Penguin.

Park, R. (1969) 'The city: Suggestions for investigation of human behavior in the urban environment', in R. Sennett (ed.), *Classic Essays on the Culture of Cities*. New York: Appleton-Century-Crofts, pp. 91–130.

Ratti, C. and Townsend, A. (2011) 'The social nexus', *Scientific American* 305(3): 42–48.

Saunders, T. and Baeck, P. (2015) *Rethinking Smart Cities from the Ground Up*. Nesta, public report, London.

Sennett, R. (1969) *Classic Essays on the Culture of Cities*. New York: Appleton-Century-Crofts.

Sennett, R. (2001) 'A flexible city of strangers', *Monde Diplomatique*, February.

Shepard, M. and Greenfield, A. (2007) *Situated Technologies Pamphlet 1: Urban Computing and its Discontents*. New York: The Architectural League of New York.

Simmel, G. (1969) 'The metropolis and mental life', in R. Sennett (ed.), *Classic Essays on the Culture of Cities*. New York: Appleton-Century-Crofts, pp. 47–60.

Tironi, M. (2010) 'Gelleable spaces, eventful geographies. The case of Santiago's experimental music scene', in I. Farias and T. Bender (eds), *Urban Assemblages: How Actor-Network Theory Changes Urban Studies*. New York: Routledge, pp. 27–52.

Townsend, A. (2008) 'Foreword', in M. Foth (ed.), *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City*. Hershey, PA: Information Science Reference, pp. xxiii–xxvii.

Townsend, A. (2015a) 'Cities of data: Examining the new urban science', *Public Culture* 27(2): 201–212.

Townsend, A. (2015b) *Making Sense of the New Urban Science*. Rudin Center and Data & Society Research Institute, New York, available from: www.citiesofdata.org/wp-content/uploads/2015/04/Making-Sense-of-the-New-Science-of-Cities-FINAL-2015.7.7.pdf [accessed 24 November 2016].

Waal, M. de (2014) *The City as Interface*. Rotterdam: Nai010 Publishers.

# 3 Data about cities

## Redefining big, recasting small

*Michael Batty*

## Introduction

Prior to the industrial revolution, record-keeping was an intensive but modest affair with manual technologies constraining the growth of data. The development of mechanical technologies from the late eighteenth century began to change this and local records gradually became more automated during the nineteenth century. The Population Census was in fact one of the only systematic catalogues of data produced on a continuing basis at a national level until national accounts and related economic data began to be collected seriously and routinely in the 1920s (Bos 2011). Automation, however, using mechanical devices continued apace in the early twentieth century and the first digital computers in mid-century embraced the challenge of dealing with ever larger data volumes that now form the basis of all kinds of development in electronic media and communications technologies.

Historically, data were always big with respect to the available means by which they could be manipulated. There is a wonderful story from the 1950s about the use of spare cycles in the early computers developed for the Lyons Tea Company (Ferry 2010) where these computers were used to compute shortest routes for freight in the rail system so that British Railways could price these goods accordingly. Dramatic and ingenious manipulations had to be devised to make this possible, such as stuffing data and intermediate calculations into all corners of memory and Scotland needing to be treated separately from the rest of Britain and then stitched back together after separate computation. In the process, those involved actually invented the well-known Dijkstra algorithm a year before Dijkstra did so himself and some four years before he published it (Graham-Cumming 2012). Countless examples such as these exist, which show how the limits of computation were reached with new algorithms, and data mining techniques were invented on the back of data which were then viewed as ‘big’.

So ‘big’ with respect to data is a relative concept and some data have always been big with regards to how they might be manipulated using state-of-the-art computation. But apart from the sheer volume of data, in cities data have always been big in another sense. Here, our concern is no longer with location but with interactions (Batty 2013): relationships between locations are best expressed by flows. The volume of data contained in flows is, in general, the square of the elements that define the locations between which the flows are generated. If there are *n* locations, then there are $n^2$ possible interactions between them and thus the data associated with interactions increases exponentially as the number of locations increases or as locations get finer and finer in terms of their resolution. Here, the contention is that big data can be generated from small data through interactions, and that higher order effects are in fact big data. Although I do not conclude that the big data revolution is a red herring, we will conclude that 'bigness' is never what it seems and that 'bigness' in terms of computational time taken to explore data, which might be quite small in size, is as important as dealing with massive data volumes.

## Classifying city data: the data cube

Introduced by Brian Berry (1964), an early data typology that has withstood the test of time is the 'geographic matrix'. This consisted of an array of places – locations – and their attributes, which he called characteristics. Such a matrix, he argued, was the essence of geographical analysis in that the dimension of place and its characteristics or attributes defined the central qualities of any location. To this he added another dimension, time, though this rarely had the same level of detail of the other two. In fact, he envisaged these additional time slices to be limited in number, though in principle each of these dimensions could take on any number of categories. Although he did not use the term, the geographic matrix in its three-dimensional form is close, if not identical to, what in data science is now called the 'data cube' (Han *et al.* 2011). Berry then proceeded to use this matrix to explode a spatial data set. In one sense, the focus was on place rather than its characteristics or its temporal positioning, but by concatenating these dimensions one might envisage a series of relationships in single, pairwise or in three-wise fashion. If we label characteristics by their volume as $M$, places as $N$, and time slices by $T$, then there are seven possible combinations of relations: $M$, $N$ and $T$ by themselves, $M \otimes N$, $M \otimes T$ and $N \otimes T$, and $M \otimes N \otimes T$. Unpacking these further, we might consider relations between $M \otimes M$, $N \otimes N$, and $T \otimes T$. Significant for this discussion is the relation between $N$ and itself which essentially is spatial interaction – linkages or flows between locations. Berry's focus however was on another kind of data explosion that comes from generating relationships between the dimensions. We will illustrate these here with respect to relationships between places – spatial interactions – which can also be tagged to quite fine resolutions of time.

In fact, it is important to be clear as to the way the data cube might be used in the analysis of city data. Even though it is based on three dimensions, which can in fact be extended to many more, usually any analysis takes one of these as being the anchor point – place, characteristics or time – and conducts analysis with respect to relationships associated with this anchor. Although the data cube is generic, whenever data are considered in these terms, the problem is usually structured from one of these perspectives and thus it is important to see the size of data, its volume and its variety at least in terms of the particular perspective adopted. It is worth indicating how traditional urban data – urban populations collected from traditional sources such as complete Population Censuses – can explode into big data. This was possible long before the current era and it is very obvious when spatial interaction is considered. In 1964, Lowry built a state-of-the-art urban model for Pittsburgh which divided the region into 456 zones between which the flows of people moving to work, shop and so on were collected. The data were collected from household interviews intended for traffic studies, but the volume when considered with respect to the matrix of interactions was huge by the standards of those times where $456^2 = 207{,}936$ possible interactions (trips) was standard. This was in an era when many mainframe computers could barely store more than 64K numbers and most of the transport models then built always pushed up against these limits. Indeed, it was one of the main reasons for the enormous problems associated with the earliest urban models, which Lee (1973) in his famous paper defined as one of data 'hungriness' (Batty 2014).

### *The emergence of big data in cities*

Before turning to examples, it is important to get a tangible sense of what the term 'big data' means, for it has only become significant in the last decade. This has coincided with the development and dissemination of countless digital devices that sense characteristics of objects in the physical environment with respect to their type, positioning and the time when they are observed. These are, of course, the three dimensions of our data cube and big data thus tends to be data that are dimensioned in at least these three ways – by their attributes or characteristics, by their spatial positioning or location, and by the time instant at which the relevant objects are observed. The objects can be human or physical, indeed of any type as long as they are associated with a relevant sensing device.

There are many definitions of big data. The cliché is that big data are defined by volume, variety, velocity, veracity and value. This simply roots the data in questions of size (bigness), variety (diversity and extent), velocity (temporal frequency of collection or observation), veracity (level of accuracy and/or uncertainty) and value (what it brings to various purposes), but it might be objected that all these criteria apply to small data. However, the implication is that it is size, scale and scope that pertain to these characteristics (IBM n.d.). In fact, big data are much more than these four or five 'Vs'. Dutcher (2014) has collected together some 40 definitions from 'thought leaders' across the industry and one of the main conclusions is that big data are more about the tools that are needed to process them than their size or volume.

Often big data are hard to understand because they have little structure, they are sometimes but not always large, and traditional tools are very difficult to use in their processing. For example, very large quantities of household census data, although not any larger in the volumetric sense than at any time in the last half century, often stretch and confuse traditional multivariate techniques. Even plotting a scatter diagram relating, say, population income to level of education at the individual or household level for a country the size of the UK requires visualizations of more than 20 million points and most if not all statistical packages and even statistical interpretations break down when confronted with such data volumes. Even so, such data would not be regarded as big data by contemporary standards for the usual rule of thumb is that the data must be giga- and upwards in size for it to be classed as big data.

Big data which are streamed in real time represents the cutting edge of new data about the functioning of cities. Much of these data are streamed from devices that are simply embedded in the physical environment and transmit data in continuous fashion with little human interference or management, such as loop counters which record traffic volumes, digital weather stations, and such like. Much of these are captured in the various dashboards that have been set up to pull together such data and make them intelligible to interested observers and policymakers. These dashboards have mainly been produced so far to demonstrate that by visually synthesizing such data one can gain an immediate impression of the state of the city (O'Brien *et al.* 2014; Kitchin *et al.* 2014). In fact, the synthesis that is required to make sense of this is very hard to develop as many of the data sources cannot be easily integrated. Moreover, much of these streamed, real-time data reflect very different concerns for cities from more traditional data sets.

Real-time data pertaining to the socio-economic structure of the city are much more problematic to collect using sensing devices. Unambiguous answers to queries which involve the human condition are almost impossible to link to real-time sensors. Information on people's choices are fraught with difficulty in terms of collection and interpretation. The reason why so much data in real time are transit data is that travel is a relatively routinized activity, whereas collecting data about unemployment, income, employment activity, migration and so on requires human and related agencies to put in place systems where people are required to respond by answering or registering. Some data are being picked up in retailing with respect to sales data from smart, credit, loyalty cards and so on, but invariably where these data are collected (and sometimes available) in real-time, various sensing devices are used. Data which are compiled from registrations are increasingly being made in near real-time, such as house prices. In these cases, the frequency at which such data are produced is monthly, possibly weekly at best to date, but these kinds of data depend on the frequency of changes – people make changes in these phenomena over matters of days and weeks and months rather than seconds and minutes (Batty *et al.* 2015).

To illustrate these issues, we will focus on transport where data are intrinsically big, including traditional data collected from questionnaires about travel patterns administered to individual travellers or households, smart card usage for collecting fares, real-time movement data from vehicles themselves, and data captured by monitoring passengers using automated observations. Not only are transport data big in that much of them deal with how travellers move between origins and destinations, thus generating spatial interactions, but they are also big temporally because automated methods can capture data continuously.

### *Traditional transport interaction data: big data generating complex visualizations*

Ever since transportation planning formally began in the 1950s, the focus has been on potential interactions or flows between origins and destinations. Different types of traffic form the essence of transport models, usually based on different modes, but the class of models that we will allude to deal with many other kinds of flow from social networks, to input–output trade relations, to patterns of migration, and so on. The concatenations that we are focusing on here are flows between places, that is $N \otimes N$ which generate travel volumes that can be substantial as the number of places $N$ increases, as we noted above for Lowry's (1964) model of Pittsburgh. Until quite recently visualizing flows has been stymied by constraints imposed on graphics. To consider the nature of the problem, in Figure 3.1(a) we show London divided into 33 separate but contiguous zones for which a journey to work matrix – flows from any zone (which is an administrative borough) to any other – is almost impossible to plot clearly. Thirty-three zones generate a total possible number of trips $33^2 = 1089$ which may not appear to be a large number, but is hard to plot clearly. We show this plot in Figure 3.1(b) where plotting all links from any zone to another, but excluding the intra-zonal trips and also suppressing the asymmetry of the matrix where the flow from zone $i$ to $j$ is generated by adding the flows as $T_{ij} + T_{ji}$ , still produces a map which is hard to interpret. Plotting individual trips from one origin to all destinations is the only way to make the map clear but we get no sense of the polycentricity of the system from this visualization and this is what we really need to detect in the data.

Now this is a very crude characterization of the journey to work in Greater London. Even 50 years ago, we would not be content with this level of resolution and therefore we will need to work with a much bigger data set by dividing these 33 zones into their constituent wards – typically local electoral districts which have on average around 13,500 residents living within them. There are 633 such zones and immediately the data have exploded to $633^2 = 400689$ potential interactions, which is quite large. We usually calibrate a model for this kind of data so that we predict each of these flows, but many of the flows for a system of this size and resolution will be small and quite a few zero in terms of the observations.

In Figure 3.2, we show the more disaggregate zoning system. It is not worth showing a plot for the full trip matrix as this is simply a mess with no way of detecting the complexity of the physical form. What we want to do is detect how close different patterns from different parts of the metropolis are and a first way into this problem is visualization. The notion of examining trips origin by origin or destination by destination is an obvious way forward and we do this in

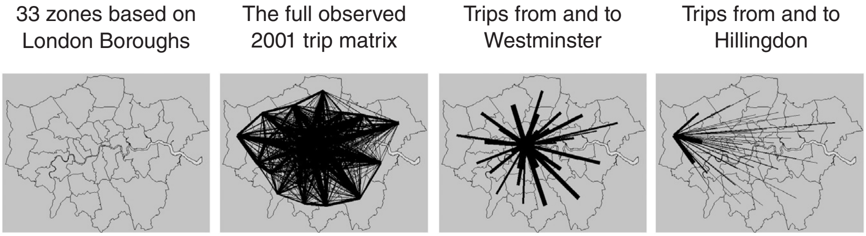


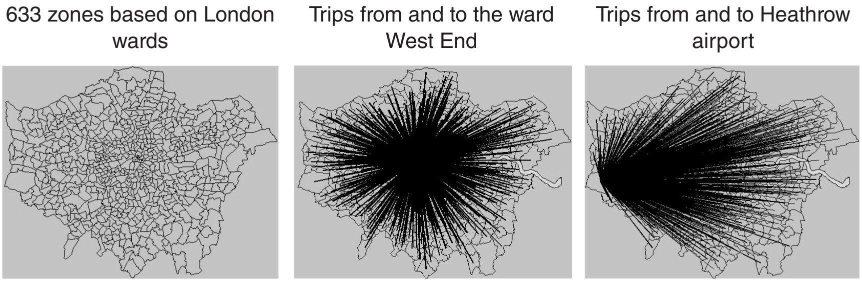
*Figure 3.1* Total two-way trips: a) the zoning system, b) all trips plotted, c) trips associated with Westminster (the centre), d) trips associated with Hillingdon (Heathrow). Note that intra-zonal trips are not plotted.

Figures 3.2(b) and (c) as we did in Figure 3.1 for the coarser resolution system. Aggregation and animation are ways of dealing with these data in terms of building up a structured understanding of this complexity, but the problem really becomes serious once we wish to test comparisons and compute correlations between the observed trip matrix and any other matrix such as a predicted one. To show how this kind of problem explodes into big data, which need new methods, we will compare the 633 × 633 matrix with one that is predicted by the model.

We now need to note the model that we will build to produce the predictions to be compared against the data in Figure 3.2. The model predicts trips $T'_{ij}$ between origins $O_i^{obs}$ and destinations $D_j^{obs}$ which are then compared against observed trips $T_{ij}^{obs}$. Observed origin and destination volumes $O_i^{obs}$ and $D_j^{obs}$ are computed from the observed data as $O_i^{obs} = \sum_j T_{ij}^{obs}$ and $D_j^{obs} = \sum_i T_{ij}^{obs}$. The model is an unconstrained gravity model that computes predicted trips as a function of the observed origin and destination volumes and an inverse function of distance $d_{ij}$ between each origin and destination pair. The model can be stated as $T'_{ij} = K\ O_i^{obs} D_j^{obs} \exp(-\beta d_{ij})$ where $K$ and $\beta$ are parameters that meet normalizing constraints. From the model, we clearly derive predicted trips but also predicted origin and destination totals $O'_i = \sum_j T'_{ij}$ and $D'_j = \sum_i T'_{ij}$. To measure the goodness of fit of the model with the data, we need to examine the scatter plots which contain the correlations between $O'_i$ and $O_i^{obs}$, $D'_j$ and $D_j^{obs}$, and $T'_{ij}$ and $T_{ij}^{obs}$.

The scatter plots for origins and destinations are easy enough to visualize as there are 633 observations in each. However, for the trips, there are a possible total of 400,869. In terms of the observed trip data, some 64 per cent of these are zero, and as the data are taken from a 10 per cent sample, this poses a problem. Should we compare zero cells with predicted ones, which will always be positive, and should we compare cells with a fractional number with integers? If we exclude the zero cells, then we still have some 142,291 to deal with, implying that only 36 per cent of our data matrix is occupied. We illustrate these patterns in Figures 3.3 and 3.4.

Figure 3.3 is revealing. The three scatters are very different with employment being predicted rather well, residential population less well, and trips showing


633 zones based on London wards

Trips from and to the ward West End

Trips from and to Heathrow airport

*Figure 3.2* Total two-way trips: a) the fine-scale zoning system, b) trips associated with an inner-city ward, c) trips associated with Heathrow airport.

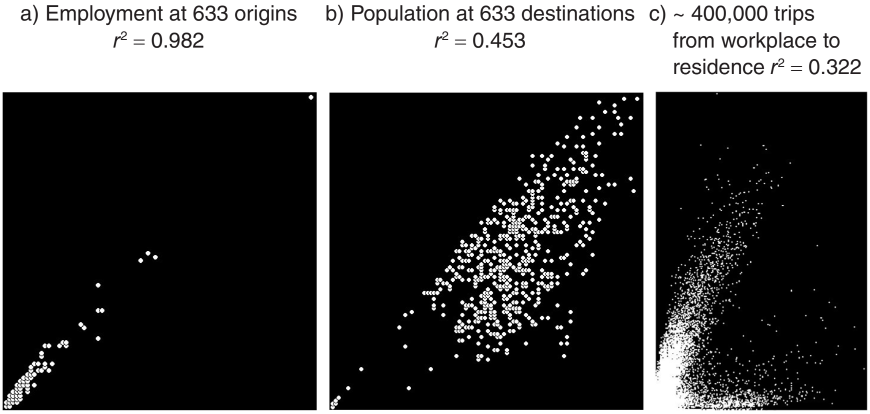


*Figure 3.3* Predicted against observed data: a) origin employments, b) destination working populations, c) trips from work to home.

that there are at least two regimes characterizing travel in London. In fact, the scatter of trips in Figure 3.3 reveals a clear density map and in Figure 3.4 we show this as best we can. The intensity of very small trips is much greater than larger ones for the distribution of trip volumes follows some sort of power law.

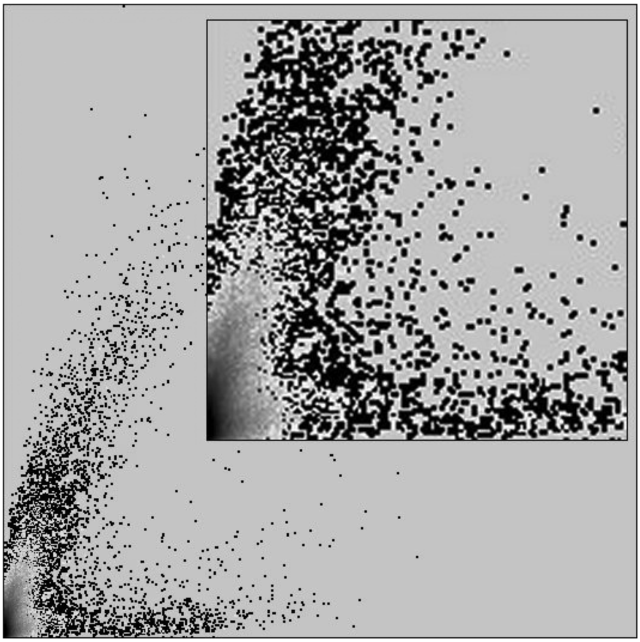
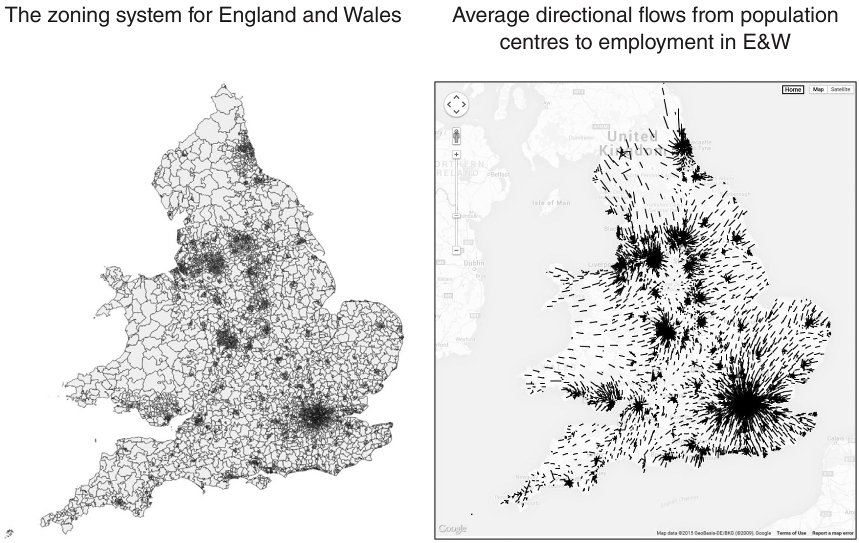


*Figure 3.4* The density of the scatter: different patterns at different scales.

In Figure 3.4, we have blown up the lower portion of the scatter to reveal this intensity and this reveals that this kind of data mining must be supplemented by many other kinds of visualization and analysis so that the true patterning of a system with this kind of complexity can be laid bare.

Now all this may not look very much like big data, but our current extensions of these models are equivalent to entire systems of cities at the same level of resolution as the Greater London model zoning system in Figure 3.2. We are now working on a model with 7,201 zones which have an average population for England and Wales of some 7,000. Our model is built for all these zones and immediately there comes a problem of visualizing the scatter of origins and destinations as well as trips of which there are a total possible cells in the matrix of $7,201^2 = 51,854,401$. Visualizing nearly 52 million points on a scatter graph is well beyond our capabilities and although only 10 million or so of these points are likely to be above zero, this is still beyond the capabilities of this kind of analysis. We show the zoning system in Figure 3.5(a) and when we move to flows, it is impossible to use the single origin, many destination tool to visualize a set of flows one by one. What we have done here is to produce a single flow for each origin to all its destinations using a weighted directional vector. For each origin $i$, we compute the average vector as a single arrow showing the average strength and direction as $[\vec{x}_i, \vec{y}_i] = [(x_i y_i), (\Sigma_j T_{ij}[x_i - x_j]/n, \Sigma_j T_{ij}[y_i - y_j]/n)]$. Much information is lost in our visualization but in the system we are developing, there is zoom capability that is able to illustrate the overall pattern at a coarse spatial scale and the detail at the finest scale of the zones themselves. We show the coarser visualization for England and Wales in Figure 3.5(b).


The zoning system for England and Wales

Average directional flows from population centres to employment in E&W

*Figure 3.5* Visualizing big data in tens of millions or more of transport flows.

Much of this has been possible in terms of data available for the last 30 years or more but only now that we have computers large enough are we able to exploit the bigness of these data. This is very different from the big data that we will present in the next section where the volume comes largely from the temporal and individual rather than spatial dimension. It does reveal, however, that big data have been with us for a while and it is computation more than anything else that determines the size of data set that we can handle, interpret and use fruitfully.

### *Real-time streamed transportation data at the micro-level*

Since the 1950s, data have been collected in continuous time for traffic flow analysis. Much of these data have been hard to link to origin-destination data of the kind just examined largely because they are supply-side data pertaining to vehicular movement and not to intentional trip-making. However, with the advent of RIFD and related technologies, it is now possible to collect data on where people enter and exit a transit system or where they embark and end any journey if the relevant collector is in place. Devices which are specially devised for the data collection in question are by far the best as the data that they produce are unambiguous (although there may be substantial noise still to be filtered out). Mobile devices for other purposes, such as phones, can also be used to extract data from call detail records which locate the phone when a call is made (Chen *et al*. 2015).

Because these data are recorded at the exact time when the smart card or mobile device is linked to the system in question, there is a continuous or at least continual record of activations which represent real-time collection, either accessible in real-time itself or for *post hoc* analysis. In short, the data are as voluminous as the number of activations. If this is phone calls, then it is the number of calls made from that device per day or over whatever unit of time and space the data are aggregated to. Here, we will use data generated by the Oyster card, a RFID smart card used on all public transport in Greater London. This card stores the money that travellers use to pay for journeys and the system is designed to recognize the category of payer as well as the time and place where the traveller taps in or out of the system. Travellers tap in and out on trains but only tap in on buses.

We have several tranches of data from this system. Our largest set is for 86 days in the summer of 2012 where there were 9,902,266,857 (nearly 10 billion) taps. Of these taps, 44 per cent were on buses and 56 per cent on rail, which is tube and overground with some being on the mainline network rail. As there is only tap-in on buses, we can guess that if round trips are made by rail, then this is about half of all rail trips meaning that there are about 60 per cent more bus trips than rail. The data also show that 11,535,090 different Oyster cards are used for these 10 billion taps, which is 86 taps per unique card, on average about one per card per day.

These data are quite unstructured. They come as a flat files where each tap is recorded by place and time – subway station, location of bus by stop, etc., and some classification of the traveller such as whether the card is free, and what the payment category is. Generally, it is possible to trace the behaviours of an individual cardholder through time and space. The degree of heterogeneity in the data set is enormous and this is a feature that makes them usable for all kinds of temporal modelling at the level of the cardholder conceived of as an agent. However, there are critical problems. The analysis of one day's worth of data in November 2010 from a series we have of three weeks' data for the 660 tube and overground rail stations revealed that 6.2 million travellers tapped in but only 5.4m tapped out. Essentially this was because barriers were up. A large class of Oyster users with free passes are not fined for not tapping in or out while season ticketholders are also not fined as their cards are loaded with a fixed amount of money for a period. This is quite a large loss of data. If you combine this with travellers using more than one card, then this confounds the data set for transport analysis.

It is possible with some analysis to figure out how many journeys are made by tracing different travellers in terms of the tap-in and -out activity during the working day, for rail at least. We have attempted some analysis of buses with respect to travellers who have a unique identifier and who hop onto buses and trains within a certain time interval, which we assume captures some multi-modal journeys, but our analysis is limited and our confidence in extracting multimodal journeys is low. In terms of the rail system, we are able to produce distinct trips in terms of segments although the analysis of round trips is more limited. For example, in the 2012 data, we can identify 291 million trips between one station and another in terms of a tap-in and tap-out with the most popular segment in the system the trip from Victoria to Oxford Circus and vice versa. Waterloo to Canary Wharf is the most frequent during the morning and evening peak with Waterloo and Victoria the two biggest volume hubs in the system.

In understanding cities, origins and destinations of trips, indeed of any flow, is essential for understanding the rationale of the location where those creating the flow are based. One of the problems with smart card data that is orientated to transit systems, such as fixed rail, is that the locations which anchor these infrastructures do not have the same meaning as origins and destinations in terms of work, shopping, residences, schools and so on which generate trips. It is extremely difficult to tie places where people enter such systems to the comprehensive patterns of locations that are described by traditional data. We can quite easily assemble flow matrices and assign trips to network segments such as lines between stations – although the precise paths of travel have to be inferred, but tying these to places of work, residence and so on is difficult. Some headway has been made using smart card data for Singapore (Zhong *et al.* 2014) but the problem is perennial and requires additional data to link points of fixed infrastructure to ultimate origins and destinations.

We have assembled several pictures of transit systems in operation from our Oyster card data. Using shortest path algorithms, Reades (2013) has worked on finding the best routes between stations identified in the data and pieced together actual flows by assigning origin data from tap-ins to the network, then finding the shortest routes on lines linking the origin to the destination. He has produced a computer movie of a typical week from the 2012 data by adding data for several typical weeks – excluding the Olympic Games weeks – thence producing an averaged version which shows the peaks and troughs in the data from Sunday to Saturday. The weekend days are very different with much less pronounced morning and evening peaks while typical workdays show very distinct morning and evening peaks that in themselves are very different with a small blip in the central area in the late evening (see Figure 3.6).

We are developing several projects using the Oyster card data but so far these tend to examine very different aspects of the city from those that pertain to traditional flow data. The focus is inevitably on questions of disruption and smooth flowing on a fine-scale temporal basis, but we are not able to relate these to links between home and work. We are able of course to examine the variability of the tap-in and tap-out data with respect to the station hubs through two interlocking patterns of entries and exit volumes that reflect two layers of polycentricity which vary through time and are reflected in the peak and off-peak flow patterns. The essential challenge is to tie this to other data, such as activity volumes of employment retailing, residential populations and so on, that come from more traditional sources.

## Conclusions and next steps

Big data are never what they seem. The multiple Vs that have become their signature definition do not capture the fact that quite small data when elaborated into their second, third and higher order effects can become big in the sense that conventional techniques and models fail to deal with their extended volumes. Our first illustrations here focus on quite modest data sets and we are conscious that really big data volumes that come from interaction patterns

Clips from the **YouTube** Movie: *Oyster Gives Up Its Pearls*, made by UCL Engineering from Jon Reades's movies of the data

Morning peak hour Monday 8 am | Lunchtime Monday 2 pm | Evening peak hour Monday 6pm

*Figure 3.6* Visualizations of the flows on the rail segments during a working day.

Movie available at YouTube (www.youtube.com/watch?v=9sAugcb2Qj4) are hard to measure in terms of their complexity through visualization. The visualization of data in countless ways has proceeded in parallel to the big data revolution, which is focused more on data mining through machine learning and in essence involves iterative techniques for searching for patterns in such data that may or may not have substantive meaning. For example, our illustration of the quality of the fit of our spatial interaction model of journey to work in Greater London (see Figures 3.3 and 3.4), suggests several features of our model and data that are quite counter to one another. In fact, the intensity of points in Figure 3.4 – the fact that a large proportion of points are inside the core of the scatter – probably need to be separated out.

Our continuing work on contemporary big data is taking many forms but so far it is mainly dealing with transit. Data on energy flows and usage in the smart city are not focal as yet, while the analysis of big data associated with social media may well remain in some preliminary form for many years. Representativeness is the key issue, as is meaning in such data, and it is not clear as yet the extent to which these social media data pertain to the social and economic functioning of the city. In another sense, big data are being created or rather extended and conflated through mashups. These kinds of integration are as important as the search for pattern in such data and as the big data revolution proceeds it is increasingly clear that the pronouncements on the end of theory, made so vociferously by commentators such as Anderson (2008), are not being borne out in any sense. The need to approach big data with clear theory has never been more important.

## References

Anderson, C. (2008) ‘The end of theory: the data deluge makes the scientific method obsolete’, *Wired Magazine* 16-07, 23 June, available from: http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory [accessed 24 November 2016].

Batty, M. (2013) *The New Science of Cities*. Cambridge, MA: MIT Press.

Batty, M. (2014) ‘Can it happen again? Planning support, Lee’s requiem and the rise of the smart cities movement’, *Environment and Planning B: Planning and Design* 41(3): 388–391.

Batty, M., Hudson-Smith, A., Hugel, S. and Roumpani, F. (2015) ‘Visualising data for smart cities’, in A. Vesco and F. Ferrero (eds), *Handbook of Research on Social, Economic, and Environmental Sustainability in the Development of Smart Cities*. Hershey, PA: IGI Global, pp. 339–362.

Berry, B.J.L. (1964) ‘Approaches to regional analysis: a synthesis’, *Annals of the Association of American Geographers* 54: 2–11.

Bos, F. (2011) ‘Three centuries of macro-economic statistics’, Munich Personal, RePEc Archive (MPRA) Paper No. 35391, available from: http://mpra.ub.uni-muenchen.de/35391/ [accessed 4 December 2016].

Chen, C., Batty, M. and van Vuren, T. (2015) ‘Editorial’, *Transportation* 42: 537–540.

Dutcher, J. (2014) ‘What is big data?’, *DataScience Berkeley Blog*, available from: http://datascience.berkeley.edu/what-is-big-data/ [accessed 24 November 2016].

Ferry, G. (2010) *A Computer Called LEO*. New York: Harper Perennial.

Graham-Cumming, J. (2012) ‘The great railway caper: big data in 1955’, available from: www.youtube.com/watch?v=pcBJfkE5UwU and see http://bigdata.blogweb.casa.ucl.ac.uk/2012/10/03/big-data-problems/ [accessed 24 November 2016].

Han, J., Kamber, M. and Pei, J. (2011) *Data Mining: Concepts and Techniques*. Waltham, MA: Morgan Kauffman.

IBM (n.d.) 'The four V's of big data', available from: www.ibmbigdatahub.com/infographic/four-vs-big-data/ [accessed 24 November 2016].

Kitchin, R., Lauriault, T.P. and McArdle, G. (2014) 'Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards', *Regional Studies, Regional Science* 2(1): 6–28.

Lee, D.B. (1973) 'Requiem for large-scale models', *Journal of the American Institute of Planners* 39: 163–178.

Lowry, I.S. (1964) *A Model of Metropolis*. RM-4035-RC, Santa Monica, CA: The Rand Corporation, available from: https://www.rand.org/pubs/research_memoranda/RM4035.html [accessed 24 November 2016].

O'Brien, O., Batty, M., Gray, S., Cheshire, J. and Hudson-Smith, A. (2014) 'On city dashboards and data stores', a paper presented to the Workshop on Big Data and Urban Informatics, 11–12 August, University of Illinois at Chicago, Chicago, IL, available from: http://urbanbigdata.uic.edu/proceedings/ [accessed 24 November 2016].

Reades, J. (2013) 'Pulse of the city', *Vimeo*, available from https://vimeo.com/41760845, original at http://simulacra.blogs.casa.ucl.ac.uk/2011/08/pulse-of-the-city/ [accessed 24 November 2016].

Zhong, C., Arisona, S.M., Huang, X., Batty, M. and Schmitt, G. (2014) 'Detecting the dynamics of urban structure through spatial network analysis', *International Journal of Geographical Information Science* 28(11): 2178–2199.

# 4 Data-driven urbanism

*Rob Kitchin*

## Introduction

There is a rich history of data being generated about cities concerning their form, their citizens, the activities that take place, and their connections with other locales. These data have been generated in a plethora of different ways, including audits, cartographic surveying, interviews, questionnaires, observations, photography and remote sensing, and are quantitative and qualitative in nature, stored in ledgers, notebooks, albums, files, databases and other media. Data about cities provide a wealth of facts, figures, snapshots and opinions that can be converted into various forms of derived data, transposed into visualizations, such as graphs, maps and infographics, analysed statistically or discursively, and interpreted and turned into information and knowledge. As such, urban data form a key input for understanding city life, solving urban problems, formulating policy and plans, guiding operational governance, modelling possible futures and tackling a diverse set of other issues. For as long as data have been generated about cities then, various kinds of data-informed urbanism have been occurring.

A new era is, however, presently unfolding wherein data-informed urbanism is increasingly being complemented and replaced by data-driven urbanism. Here, urban operational governance and city services are becoming highly responsive to a form of networked urbanism in which big data systems are prefiguring and setting the urban agenda and are influencing and controlling how city systems respond and perform. In short, we are moving into an era where cities are becoming ever more instrumented and networked, their systems interlinked and integrated, and the vast troves of data being generated used to manage and control urban life. Computation is now routinely being embedded into the fabric and infrastructure of cities producing a deluge of contextual and actionable data which can be processed and acted upon in real-time. Moreover, data that used to be the preserve of a single domain are increasingly being shared across systems enabling a more holistic and integrated view of city services and infrastructures. As such, cities are becoming knowable and controllable in new dynamic ways, responsive to the data generated about them (Kitchin *et al.* 2015). I thus argue that data-driven urbanism is the key mode of production for what have widely been termed smart cities.

In this chapter I provide a critical overview of data-driven urbanism focusing in particular on the relationship between data and the city, rather than network infrastructure, computational or urban issues. The chapter starts by setting out how cities are being instrumented and captured as big urban data, how these data are being used to manage and control cities, and how data-driven urbanism is underpinning the emergence of smart cities. This is then followed by a critical examination of a number of problematic issues related to data-driven urbanism, including: the corporatization of governance (data ownership, data control, data coverage and access); the creation of buggy, brittle, hackable urban systems (data security, data integrity); and social, political, ethical effects (data protection and privacy, dataveillance, and data uses including social sorting and anticipatory governance). More technical data issues such as data quality, the veracity of urban data models and data analytics, and data integration and interoperability are discussed in Chapters 9 and 10.

## Big data and smart cities

Since the start of computing era urban data have been increasingly digital in nature, either digitized from analogue sources (manually entered or scanned) or born digital, generated by digital devices, stored as digital files and databases, and processed and analysed using various software systems such as information management systems, spreadsheets and stats packages, and geographic information systems. From the 1980s onwards, public administration records, official statistics and other forms of urban data were released predominately in digital formats and processed and analysed through digital media. However, these data were (and continue to be) generated and published periodically and often several months after generation.

In cases such as exhaustive datasets – for example, detailed framework mapping data or national censuses – new surveys are very infrequent (e.g. 10 years for censuses) and their publication might be 18–24 months after collection, and longer for specific subsets. For domain specific issues, such as transport and traffic flows or public transportation usage, surveys are conducted every few years, using a limited spatial and temporal sampling framework (selected locations for a short period of time). Only a handful of datasets are published monthly (e.g. unemployment rates) or quarterly (e.g. GDP), with most being updated annually due to the effort required to generate them. These data typically have poor spatial resolution, referring to large regions or a nation, and have little disaggregation (e.g. by population classes or economic sectors). In cases where data generation is more frequent, such as remote sensing, only occasional snapshots are bought by city administrations due to their licensing costs. In other cases, such as consumer purchasing (as evidenced in credit card transactions), data have largely been black-boxed within financial institutions. In other words, whilst there has been a range of urban digital data available to urban managers and policymakers from the 1970s through to 2000s, along with increasingly sophisticated software such as GISs to make sense of them, sources of data were temporally, spatially and domain (scope) limited.

Post-Millennium, the urban data landscape has been transformed, with a massive step-change in the nature and production of urban data, transitioning from small data to big data, wherein the generation of data is continuous, exhaustive to a system, fine-grained, relational and flexible across a range of domains (Kitchin 2014a). From a position of relative data scarcity, the situation is turning to one of data deluge. This is particularly the case with urban operational data wherein traditional city infrastructure, such as transportation (e.g. roads, rail lines, bus routes, plus the vehicles/carriages) and utilities (e.g. energy, water, lighting), have become digitally networked, with grids of embedded sensors, actuators, scanners, transponders, cameras, meters and GPS (constituting what has been called the Internet of Things) producing a continuous flow of data about infrastructure conditions and usage. Many of these systems are generating data at the individual level, tracking travel passes, vehicle number plates, mobile phone identifiers, faces and gaits, buses/trains/taxis, meter readings, etc. (Dodge and Kitchin 2005). These are being complemented with big data generated by: (a) commercial companies such as mobile phone operators (location/movement, app use, activity), travel and accommodation sites (reviews, location/movement, consumption), social media sites (opinions, photos, personal info, location/movement), transport providers (routes, traffic flow), website owners (clickstreams), financial institutions and retail chains (consumption, in-store movement, location), and private surveillance and security firms (location, behaviour) that are increasingly selling and leasing their data through data brokers, or making their data available through APIs (e.g. Twitter and Foursquare); (b) crowdsourcing (e.g. OpenStreetMap) and citizen science (e.g. personal weather stations) initiatives, wherein people collaborate on producing a shared data resource or volunteer data. Other kinds of more irregular urban big data include digital aerial photography via planes or drones, or spatial video, LiDAR (light detection and ranging), thermal or other kinds of electromagnetic scans of environments that enable the mobile and real-time 2D and 3D mapping of landscapes. And whilst official statistics are largely still waiting to undergo the data revolution (Kitchin 2015), the generation of public administration data has been transformed through the use of e-government online transactions that produce digital data at the point-of-collection.

We are at start of this new big data era and the flow and variety of urban data is only going to grow and diversify. Moreover, whilst much of these data presently remain in silos and are difficult to integrate and interlink due to varying standards and formats, they will increasingly be corralled into centralized systems such as inter-agency control rooms for monitoring the city as a whole or what have been termed city operating systems. With regards to the former, the Centro De Operacoes Prefeitura Do Rio in Rio de Janeiro, Brazil, a data-driven city operations centre pulls together into a single location real-time data streams from 30 agencies, including traffic and public transport, municipal and utility services, emergency and security services, weather feeds, information generated by employees and the public via social media, as well as administrative and statistical data, and is overseen by a staff of 400 data operatives (see Figure 4.1 for two examples of urban control rooms). City operating systems are effectively Enterprise Resource Planning (ERP) systems designed to coordinate and operate the activities of large companies repurposed for cities. Examples include Microsoft's *CityNext*, IBM's *Smarter City*, Urbiotica's *City Operating System* and PlanIT's *Urban Operating System*. With the advent of the open data movement some of these data also feed into public-facing urban dashboards that provide a mix of interactive visualizations of real-time, public administration and official statistical data (Kitchin *et al.* 2015; see Chapter 9).

Further, the production of these new big data has been accompanied by a suite of new data analytics designed to extract insight from very large, dynamic datasets, consisting of four broad classes: data mining and pattern recognition; data visualization and visual analytics; statistical analysis; and prediction, simulation and optimization (Miller 2010; Kitchin 2014b). These analytics rely on machine learning (artificial intelligence) techniques and vastly increased computational power to process and analyse data. Moreover, they enable a new form of data-driven science to be deployed that rather than being theory-led seeks to generate hypotheses and insights 'born from the data' (Kelling *et al.* 2009). This is leading to the development of 'urban informatics' (Foth 2009), an informational and human–computer interaction approach to examining and communicating urban processes, and 'urban science', a computational modelling approach to understanding and explaining city processes that builds upon and radically extends quantitative forms of urban studies that have been practised since the 1950s, blending in geocomputation, data science and social physics (Batty 2013). Whereas urban informatics is more human-centred, interested in understanding and facilitating the interactions between people, space and technology, urban science promises to not only make sense of cities as they presently are (by identifying relationships and urban 'laws'), but to also predict and simulate likely future scenarios under different conditions, potentially providing city managers with valuable insight for planning and development decision-making and policy formulation.

*Figure 4.1* Urban control rooms: (a) Rio de Janeiro, (b) Dublin.

Urban big data, city operating systems, urban informatics and urban science analytics provide the basis for a new logic of urban control and governance – data-driven urbanism – that enables real-time monitoring and steering of urban systems and the creation of what has widely been termed smart cities. The notion of a smart city can be traced back to experiments with urban cybernetics in the 1970s (Flood 2011; Townsend 2013), the development of new forms of city managerialism and urban entrepreneurship, including smart growth and new urbanism, in the 1980s and 1990s (Hollands 2008; Wolfram 2012; Söderström *et al.* 2014; Vanolo 2014), and the fusing of ICT and urban infrastructure and development of initial forms of networked urbanism from the late 1980s onwards (Graham and Marvin 2001; Kitchin and Dodge 2011). As presently understood, a smart city is one that strategically uses networked infrastructure and associated big data and data analytics to produce a:

- *smart economy* by fostering entrepreneurship, innovation, productivity, competitiveness, and producing new forms of economic development such as the app economy, sharing economy and open data economy;
- *smart government* by enabling new forms of e-government, new modes of operational governance, improved models and simulations to guide future development, evidence-informed decision-making, better service delivery, and making government more transparent, participatory and accountable;
- *smart mobility* by creating intelligent transport systems, efficient inter-operable multi-modal public transport, smart parking and sharing services related to taxis and bikes;
- *smart environments* by promoting sustainability and resilience and the development of green energy;
- *smart living* by improving quality of life, increasing safety and security and reducing risk;
- *smart people* by creating a more informed citizenry and fostering creativity, inclusivity, empowerment and participation (Giffinger *et al.* 2007; Cohen 2012).

In short, the smart city promises to solve a fundamental conundrum of cities – how to reduce costs and create economic growth and resilience at the same time as producing sustainability and improving services, participation and quality of life – and to do so in common-sense, pragmatic, neutral and supposedly apolitical ways by utilizing a fast-flowing torrent of urban data and data analytics, algorithmic governance and responsive networked urban infrastructure. Moreover, much more information is being placed into the hands of the public to aid decision-making, navigation and participation through a plethora of locative social media (apps that tell them about the city and which they can contribute to), open data sites, public dashboards, hackathons and so on.

The notion of smart cities, and the mode of data-driven urbanism, have not however been universally welcomed and have been subject to a number of critiques. First, smart city initiatives treat cities as a set of knowable and manageable systems that act in largely rational, mechanical, linear and hierarchical ways and can be steered and controlled, rather than dealing with cities as complex, messy, contingent cities full of wicked problems (Kitchin *et al.* 2015). Second, smart city initiatives are largely ahistorical, aspatial and homogenizing in their orientation and intent, treating cities as if they are all alike in terms of their political economy, culture and governance (Greenfield 2013). Third, an emphasis is placed on creating technical rather political/social/policy solutions to urban problems thus overly promoting technocratic forms of governance (Morozov 2013). Fourth, the project of producing smart cities tends to reinforce existing power geometries and social and spatial inequalities rather than eroding or reconfiguring them (Datta 2015). Fifth, the approach fails to recognize the politics of urban data and the ways in which they are the product of complex socio-technical assemblages (Kitchin 2014b). Sixth, the smart city agenda is being overly driven by corporate interests who are using it to capture government functions as new market opportunities rather than serve a public good (Hollands 2008). Seventh, networking city infrastructure potentially creates buggy, brittle and hackable urban systems (Kitchin and Dodge 2011; Townsend 2013). And finally, data-driven urbanism produces a number of activities that have profound social, political, ethical consequences, including dataveillance and extensive geosurveillance, social and spatial sorting, and anticipatory governance (Graham 2005; Kitchin 2014a).

In the rest of this chapter, I want to concentrate on the last four critiques, and in particular their associated data issues (rather than other aspects of the technical stacks of urban socio-technical assemblages, and wider political-economic framing and effects) as way of further illustrating some of the challenges posed by data-driven urbanism and the need to further examine the relationship between data and the city.

## Data and the city

### *The politics of urban data*

One of the key arguments for adopting a data-driven approach to urban governance is that it provides a strong evidence-based approach to decision-making, system control and policy formation, rather than one that is anecdotal, clientelist or localist. A data-driven approach, it is argued, is less susceptible to political influence and instead is driven by objective, neutral facts in a technocratic, common-sense, pragmatic way. Technical systems and the data they produce are objective and non-ideological and thus politically benign. Sensors, networked infrastructure and computers it is contended have no inherent politics – they simply measure a value, communicate those values, and process, analyse and display the data using scientific principles, thus producing measurements, records and information that reflect the truth about cities. And while data from social systems, such as social media platforms (e.g. Twitter), are inherently more subjective and noisy, they provide a direct reflection of the views, interactions and behaviour of people, in contrast to official surveys which reflect what people say they do or think (or what they think the surveyor wants to hear), thus providing better ground truthing of social reality. As such, big data about cities can be taken at face value and used unconditionally shed light on cities and to manage and control urban systems and infrastructure and guide urban policy.

The reality is somewhat different for two reasons. First, there are a number of technical issues concerning data coverage, access and quality that means that the view data presents of the city is always partial and subject to caution (see Chapters 2, 3, 10, 15, 17). Second, data are the products of complex socio-technical assemblages that are framed and shaped by a range of technical, social, economic and political forces and are designed to produce particular outcomes (Kitchin 2014b; see Figure 4.2). On the one hand, what data are produced, how they are handled, processed, stored, analysed and presented is the result of a particular technical configuration and how it is deployed (e.g. where sensors are located, their field of view, their sampling rate, their settings and calibration, etc.). On the other hand, how a system is designed and run is influenced by systems of thought, technical know-how, the regulatory environment, funding and resourcing, organizational priorities and internal politics, institutional collaborations and marketplace demand. In other words, a data assemblage possesses a '*dispositif*', defined by Foucault (1980 [1977]: 194) as a: 'thoroughly heterogeneous ensemble consisting of discourses, institutions, architectural forms, regulatory decisions, laws, administrative measures, scientific statements,

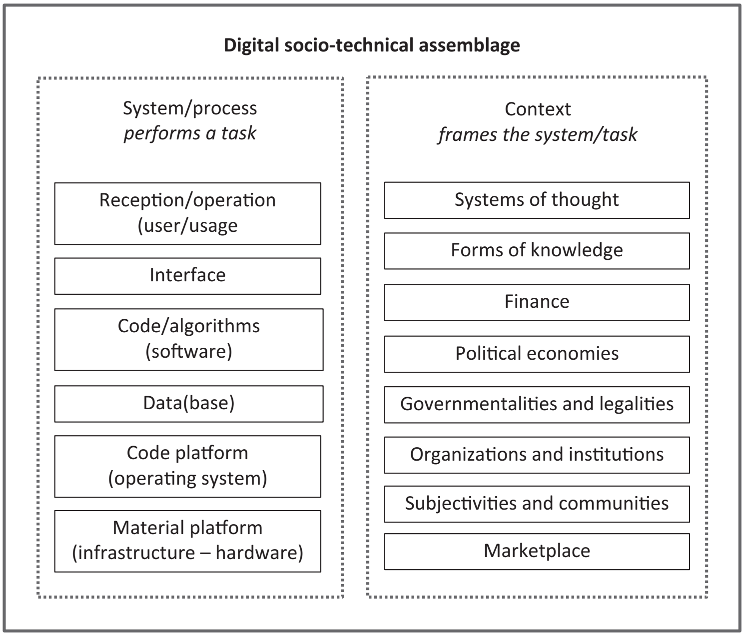


*Figure 4.2* A data assemblage.

philosophical, moral and philanthropic propositions'. For Foucault, a *dispositif* is inherently political producing what he terms 'power/knowledge', that is knowledge that fulfils a strategic function. In other words, urban big data are never neutral and objective, but rather are situated, contingent, relational and framed and used contextually to try and achieve certain aims and goals (to monitor, enhance, empower, discipline, regulate, control, produce profit, etc.). Or to put it another way, urban data are never raw but are always already cooked to a particular recipe for a particular purpose (Bowker 2005; Gitelman 2013). As such, data-driven urbanism is thoroughly political, seeking to produce a certain kind of city. It is thus necessary when examining urban big data to critically unpack their associated data assemblage (including the entire technical stack – infrastructure, platform, software/algorithms, data, interface) to document how it is constituted and works in practice to produce urban processes and formations, and for whose benefit.

### *Data access, data ownership and data control*

As already noted, much of the data presently being generated about cities are produced by commercial companies, such as mobile phone operators, and private utility and transport companies. For them, their data are a valuable commodity that provides competitive advantage or an additional revenue stream if sold/leased, and they are under no obligation to share freely the data they generate through their operations with city managers or the public. As noted in 2014 by the British Minister for Smart Cities, Dan Byles MP,[1] the privatization of public services in the UK and elsewhere has also meant the privatization of their associated data unless special provision was made to ensure it was shared with the city or made open. Similarly, access to data within public–private partnerships and semi-state agencies, or state agencies operating as trading funds (such as the Met Office and Ordnance Survey in the UK who generate significant operating costs by selling data and services), can be restricted or costly to purchase. Consequently, key framework datasets (e.g. detailed maps) can have limited access and data concerning transportation (e.g. bus, rail, bike share schemes, private tolls), energy and water be entirely black-boxed. Even within the public sector, data can be siloed within particular departments and not be shared with other units within the organization, or be open for other institutions or the public to use. As such, whilst there might be a data revolution underway, access to much of that data is limited, and there are a number of issues that need to be explored with respect to data ownership and data control, especially with respect to procurement and the outsourcing or privatization of city services. Moreover, even if all data were to be open and shared it needs to be acknowledged that there are still many aspects about cities where data generation is weak or absent. For example, in a recent audit of Dublin datasets to determine whether the city was in a position to apply for ISO37120 (the ISO standard for city indicators) data could only be sourced for 11 of 100 indicators sought (predominately because the data sought was either privatized or released at an inappropriate scale).

### *Data security and data integrity*

One of the prime anxieties of networking infrastructure and ubiquitous urban computing is the creation of systems and environments which are inherently buggy and brittle and are prone to viruses, glitches, crashes and security hacks (Kitchin and Dodge 2011; Townsend 2013). As Mims (2013) notes, any networked device is open to be hacked and its data stolen and used for criminal purposes, or corrupted, or controlled remotely, or misdirected, or to spy on its users. The media report almost daily on large-scale data breaches of commercial companies and state agencies and the theft of valuable personal data, with several incidents of city infrastructure such as traffic management systems being hacked, disabled and controlled (Paganini 2013). As Townsend (2013) notes, the notion of smart cities takes two open, highly complex and contingent systems – cities and computing – and binds and networks them together, meaning that data-driven, networked urbanism has in-built vulnerabilities. Moreover, as urban systems evolve to become more complex, interconnected and interdependent these vulnerabilities potentially multiply (Townsend 2013; Kitchin 2016). Creating secure big urban data systems is thus set to be a significant ongoing task if public trust in their purported benefits are to be gained and maintained. Another significant element in upholding trust is how and to what purposes the data are deployed.

### *Data uses and ethics*

Urban big data are presently being used to undertake a diverse range of tasks, some of which seem relatively benign, such as monitoring city lighting with the aim of improving the quality of light and reducing its cost, and others more clearly political, such as directing policing activity. A significant concern is that as more and more data about cities and their citizens are generated, privacy becomes eroded (Kitchin 2016). Privacy is considered a basic human right, a condition that people expect and value in developed countries. Yet, as sensors, cameras, smartphones and other embedded and mobile devices generate ever more data it becomes increasingly difficult to protect, with individuals leaving ever greater quantities of digital footprints (data they themselves leave behind) and data shadows (information about them generated by others). Such troves of data are amenable to dataveillance, a mode of surveillance enacted through sorting and sifting datasets in order to identify, monitor, track, regulate, predict and prescribe (Clarke 1988; Raley 2013), and geosurveillance, the tracking of location and movement of people, vehicles, goods and services and the monitoring of interactions across space (Crampton 2003; see Table 4.1). Given the always-on nature of many of these systems, and the tracking of unique identifiers, such dataveillance and geosurveillance are becoming continuous and fine-grained with, for example, mobile phone companies always knowing the location of a phone (Dodge and Kitchin 2005). Moreover, as data minimization norms become relaxed there are anxieties that data are being shared, combined and used for purposes for which they were never intended (Kitchin 2014b). In particular, the last 20 years have witnessed the rapid growth of a number of data brokers who capture, gather together and repackage data for rent (for one time use or use under licensing conditions) or re-sale, and produce various derived data and data analytics (CIPPIC 2006).

Whilst focusing on different markets, data brokers seek to mesh together offline, online and mobile data to provide comprehensive views of people and places and to construct personal and geodemographic profiles (Goss 1995; Harris *et al.* 2005). These profiles are then used to predict behaviour and the likely value or worth of an individual and to socially sort them with respect to credit, employment, tenancy and so on (Graham 2005). The concern is that these firms practice a form of 'data determinism' in which individuals are not profiled and judged just on the basis of what they have done, but on the prediction of what they might do in the future using algorithms that are far from perfect, and yet are black-boxed and lack meaningful oversight and remediate procedures (Ramirez 2013). Such anticipatory governance can have far reaching effects. For example, a number of US police forces are now using predictive analytics to anticipate the location of future crimes and direct patrols, and to identify individuals most likely to commit a crime in the future, designating them pre-criminals (Stroud 2014). In such cases, a person's digital footprints and data shadow does more than follow them; it precedes them. Data assemblages then do not act as cameras reflecting the world as it is, but rather as engines shaping the world in diverse ways (Mackenzie 2008).

*Table 4.1* Movement and location tracking. Compiled from Kitchin (2016)

| | |
|---|---|
| Remote controllable digital CCTV cameras | Can zoom, move and track individual pedestrians and vehicles. Analysis and interpretation increasingly aided by facial, gait and automatic number plate recognition (ANPR) using machine vision algorithms. |
| Smartphone phone tracking | Location is communicated to telecommunications providers through the cell masts connected to the sending of GPS coordinates, or connections to wifi hotspots. |
| Sensor networks | Sensors deployed on street infrastructure such as bins and lampposts or in shops/malls capture and track phone identifiers such as MAC addresses. |
| Wifi meshes | The IDs of devices which access or try to access a wifi network are captured and tracked between wifi points. |
| Smart card tracking | Barcodes, magnetic strips or embedded RFID chips are tracked when they are scanned to gain entry to buildings or transportation. |
| Active GPS tracking | Embedded GPS in devices and vehicles communicate location and movement via cellular or satellite networks. |
| Transponder tracking | Transponders with embedded RFID chips broadcast their IDs and are tracked by scanning receivers, commonly used in automatic road tolling or electronic tagging of people on probation. |
| Other staging points | Such as using ATMs, credit cards or checking a book out of a library that leaves a digital record. |

## Conclusion

We are entering an era where computation is being routinely embedded into urban environments and networked together, and people are moving about with smartphones that ensure always available connectivity and access to information. These devices and infrastructures are producing and distributing vast quantities of data in real-time, and they are also responsive to these data and the analytics undertaken on them enabling new kinds of monitoring, regulation and control. Cities then are becoming data-driven and are enacting new forms of algorithmic governance. However, the data and algorithms underpinning them are far from objective and neutral, but rather are political, imperfect and partial. The smart cities that data-driven, networked urbanism purports to create are then smart in a qualified sense. Their production and operation is based on much more data and derived information than previous generations of urbanism, but it is a form of urbanism that is nonetheless still selective, crafted, flawed, normative and politically inflected. Moreover, while the instrumental rationality of data-driven, networked urbanism promotes urban knowledge and management rooted in a quite narrowly framed 'episteme (scientific knowledge) and teche (practical instrumental knowledge)', it is important that other forms of knowing, such as 'phronesis (knowledge derived from practice and deliberation) and metis (knowledge based on experience)' (Parsons 2004: 49) are not silenced, providing both a counter-weight to the limits of smart cities and positions from which to reflect on, critique and recast the production of data-driven urbanism. Indeed, whilst data-driven urbanism undoubtedly provides a set of solutions for urban problems, we also have to recognize that it has a number of shortcomings and a number of potential perils. The challenge facing urban managers and citizens in the age of smart cities is to realise the benefits of planning and delivering city services using a surfeit of data, evidence and real-time responsive systems whilst minimizing any pernicious effects. To do that we have to be as smart about data and data analytics as we would like to be about cities.

## Acknowledgements

The research for this chapter was provided by a European Research Council Advanced Investigator Award, 'The Programmable City' (ERC-2012-AdG-323636).

## Note

1 www.youtube.com/watch?v=3E3RpGMKbhg.

## References

Batty, M. (2013) *The New Science of Cities*. Cambridge, MA: MIT Press.

Bowker, G. (2005) *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.

CIPPIC (2006) *On the Data Trail: How detailed information about you gets into the hands of organizations with whom you have no relationship. A Report on the Canadian Data*

*Brokerage Industry*. The Canadian Internet Policy and Public Interest Clinic, Ottawa. Available from: https://cippic.ca/sites/default/files/May1-06/DatabrokerReport.pdf [accessed 6 February 2017].

Clarke, R. (1988) ‘Information technology and dataveillance’, *Communications of the ACM* 31(5): 498–512.

Cohen, B. (2012) ‘What exactly is a smart city?’, *Fast Co.Exist*, 19 September, available from: www.fastcoexist.com/1680538/what-exactly-is-a-smart-city [accessed 6 February 2017].

Crampton, J. (2003) ‘Cartographic rationality and the politics of geosurveillance and security’, *Cartography and Geographic Information Science* 30(2): 135–148.

Datta, A. (2015) ‘New urban utopias of postcolonial India: ‘Entrepreneurial urbanization’ in Dholera smart city, Gujarat’, *Dialogues in Human Geography* 5(1): 3–22.

Dodge, M. and Kitchin, R. (2005) ‘Codes of life: Identification codes and the machine-readable world’, *Environment and Planning D: Society and Space* 23(6): 851–881.

Flood, J. (2011) *The Fires: How a Computer Formula, Big Ideas, and the Best of Intentions Burned Down New York City – and Determined the Future of Cities*. New York: Riverhead.

Foth, M. (ed.) (2009) *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City*. New York: IGI Global.

Foucault, M. (1980 [1977]) ‘The confession of the flesh’, in C. Gordon (ed.), *Power/Knowledge.* New York: Pantheon Books, pp. 194–228.

Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanović, N. and Meijers, E. (2007) *Smart cities: Ranking of European medium-sized cities*. Centre of Regional Science, Vienna UT, available from: www.smart-cities.eu/download/smart_cities_final_report.pdf [accessed 12 October 2015].

Gitelman, L. (ed.) (2013) *‘Raw Data’ is an Oxymoron*. Cambridge, MA: MIT Press.

Goss, J. (1995) ‘“We know who you are and we know where you live”: the instrumental rationality of geodemographics systems’, *Economic Geography* 71: 171–198.

Graham, S. (2005) ‘Software-sorted geographies’, *Progress in Human Geography* 29: 562–580.

Graham, S. and Marvin, S. (2001) *Splintering Urbanism: Networked Infrastructures, Technological Mobilities and the Urban Condition*. New York: Routledge.

Greenfield, A. (2013) *Against the Smart City*. New York: Do Publications.

Harris, R., Sleight, P. and Webber, R. (2005) *Geodemographics, GIS and Neighbourhood Targeting*. Chichester: Wiley.

Hollands, R.G. (2008) ‘Will the real smart city please stand up?’, *City* 12(3): 303–320.

Kelling, S., Hochachka, W., Fink, D., Riedewald, M., Caruana, R., Ballard, G. and Hooker, G. (2009) ‘Data-intensive science: a new paradigm for biodiversity studies’, *BioScience* 59(7): 613–620.

Kitchin, R. (2014a) ‘The real-time city? Big data and smart urbanism’, *GeoJournal* 79(1): 1–14.

Kitchin, R. (2014b) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: SAGE.

Kitchin, R. (2015) ‘The opportunities, challenges and risks of big data for official statistics’, *Statistical Journal of the International Association of Official Statistics* 31(3): 471–481.

Kitchin, R. (2016) *Getting Smarter about Smart Cities: Improving Data Privacy and Data Security*. Data Protection Unit, Department of the Taoiseach, Dublin, Ireland, available from: www.taoiseach.gov.ie/eng/Publications/Publications_2016/Smart_Cities_Report_January_2016.pdf [accessed December 2016].

Kitchin, R. and Dodge, M. (2011) *Code/Space: Software and Everyday Life*. Cambridge, MA: MIT Press.

Kitchin, R., Lauriault, T. and McArdle, G. (2015) ‘Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards’, *Regional Studies, Regional Science* 2: 1–28.

Mackenzie, D. (2008) *An Engine*, *Not a Camera. How Financial Models Shape Markets*. Cambridge, MA: MIT Press.

Miller, H.J. (2010) ‘The data avalanche is here. Shouldn’t we be digging?’ *Journal of Regional Science* 50(1): 181–201.

Mims, C. (2013) ‘Coming soon: the cybercrime of things’, *The Atlantic*, 6 August, available from: www.theatlantic.com/technology/archive/2013/08/coming-soon-the-cybercrime-of-things/278409/ [accessed 7 August 2013].

Morozov, E. (2013) *To Save Everything, Click Here: Technology, Solutionism, and the Urge to Fix Problems That Don’t Exist*. New York: Allen Lane.

Paganini, P. (2013) ‘Israeli road control system hacked, causes traffic jam on Haifa highway’, *The Hacker News*, 28 October, available from: http://thehackernews.com/2013/10/israeli-roadcontrol-system-hacked.html [accessed 13 November 2013].

Parsons, W. (2004) ‘Not just steering but weaving: Relevant knowledge and the craft of building policy capacity and coherence’, *Australian Journal of Public Administration* 63(1): 43–57.

Raley, R. (2013) ‘Dataveillance and countervailance’, in L. Gitelman (ed.), *‘Raw Data’ is an Oxymoron*. Cambridge, MA: MIT Press, pp. 121–146.

Ramirez, E. (2013) ‘The privacy challenges of big data: A view from the lifeguard’s chair’, *Technology Policy Institute Aspen Forum*, 19 August, available from: http//ftc.gov/speeches/ramirez/130819bigdataaspen.pdf [accessed 11 October 2013].

Söderström, O., Paasche, T. and Klauser, F. (2014) ‘Smart cities as corporate storytelling’, *City* 18(3): 307–320.

Stroud, M. (2014) ‘The minority report: Chicago’s new police computer predicts crimes, but is it racist?’, *The Verge*, 19 February, available from: www.theverge.com/2014/2/19/5419854/the-minority-report-this-computer-predicts-crime-but-is-it-racist [accessed 4 December 2016].

Townsend, A. (2013) *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*. New York: W.W. Norton.

Vanolo, A. (2014) ‘Smartmentality: The smart city as disciplinary strategy’, *Urban Studies* 51(5): 883–898.

Wolfram, M. (2012) ‘Deconstructing smart cities: An intertextual reading of concepts and practices for integrated urban and ICT development’, in M. Schrenk, V.V. Popovich, P. Zeile and P. Elisei (eds), *Re-Mixing the City: Towards Sustainability and Resilience?* REAL CORP, pp. 171–181.

# Part II

# Urban data

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

# 5 Crime data and analytics

## Accounting for crime in the city

*Teresa Scassa*

## Introduction

Crime data are routinely collected in cities in the course of policing functions and are of great interest to many different constituencies from government and other public authorities to the private sector and private individuals. Lomell (2011: 191) refers to crime data as 'an old epistemological object', noting that they have been counted 'ever since society "became statistical"'. Crime data generally provide a record of incidents of crime, including the type of crime, and the outcome of any police investigation.

Crime data are made available to the public in different ways. Refined crime data in the form of national statistics are frequently available as open data; that is, they are made available in reusable digital formats with few or no restrictions on reuse. Crime data may also be made accessible in a variety of other ways, such as in reports, tables, charts and through other modes of representation and dissemination. In some cases, crime data may be made available to the public in ways that are neither open nor accessible in that the ability to extract the data for reuse is physically difficult or legally constrained. An example of this is through visualizations such as crime maps which may create both technological and legal barriers to extraction and reuse of data.

Crime data are a prime candidate for data analytics. As a result, whether or how they are made available to the public can be important in determining who will shape and control their analysis. A number of private sector companies now offer complex data analytics services to police forces in Canada, the US and the UK. Among the services offered by these companies are data management tools, dashboards, visualizations and predictive analytic services. Fundamentally, these services depend upon the data generated by police in the course of their operations. While many of the services offered by these companies are intended for internal police use, these companies also offer police agencies public-facing visualization tools that allow crime data to be made available to the public, although such data are neither open nor accessible in other ways (Scassa 2016).

The incidence of crime is linked to a broad range of social and economic factors that influence both victimization and involvement in criminal activity. In this way, crime data are frequently part of oppositional social justice or crime control narratives both within cities and at regional and national levels. Far from neutral and objective, crime data are subjective and contested, and reveal more in their inclusions and exclusions than their account of the incidence of crime. Crime data represent not just a point of contact between individuals and the state, but one which depends upon human judgment for their interpretation and categorization. While technology increasingly plays a role in capturing and analysing data about criminal activity, decisions as to whether to record incidents as 'crimes' and to identify them as specific types of crimes still rely upon human judgement. Thus, the 'sensors' that record crime data are human and their role is not simply to capture observations. Crime data involve interpretation, judgment and action; crime datasets are an artefact of the interaction of citizen and state, as understood by agents of the state operating within particular institutional cultures.

This chapter considers crime data within a context that increasingly relies upon data and data analytics for planning, decision-making and for informing public understanding of problems and their solutions. The second section defines crime data, while the third section explores those factors that limit them. The fourth section examines how crime data are communicated to the public, including as accessible, available and open data and through data visualization. The fifth section considers ways in which crime data can be used to increase both the transparency of the systems that produce them and our understanding of crime in the urban context. While more and different data may enhance insight into urban crime and policing that can be derived from crime data, these data remain significantly limited by their subjectivity and by the legal, institutional and cultural constraints that shape and control them.

## Crime data

Kitchin (2014: 2) describes the current concept of data as 'capta', or as 'those units of data that have been selected and harvested from the sum of all potential data'. Crime data are a good example of capta. They display considerable degrees of both choice and subjectivity in their recording and representation. Such data are 'partial, selective and representative, and the distinguishing criteria used in their capture has consequences' (Kitchin 2014: 3). At a very basic level crime data are data about the incidence of crime derived from police reports. They include the type of crime, the clearance status, and whether anyone was charged with an offence. In some cases, additional data such as the demographic characteristics of the perpetrator and/or victim may be included. Crime data also typically include a geographical reference point – usually the location of the incident in question. This spatial dimension makes the data more valuable analytically; for example, they can be used to determine crime 'hot spots' and to inform decisions about the appropriate allocation of police and other resources. The spatial dimensions of crime data also make them good candidates for visualization techniques.

Because crime data are derived from police reports, these data relate only to 'crimes known to law enforcement' (and not to unreported criminal activity) and thus represent ‘only a fraction of crimes that actually happen’ (Fisher *et al.* 2002: 72–73). They are also not necessarily crimes that are adjudicated in court. Thus, although they are typically referred to as crime data or crime statistics, these data are, more accurately, about policing.

Recognizing that crime data derived from police reports do not reflect a true picture of the incidence of crime, statistical agencies now try to correct for underreporting (and to measure underreporting) through victimization surveys (Fisher *et al.* 2002). In Canada, this takes the form of the General Social Survey – Victimization carried out by Statistics Canada. In the US, it is the National Crime Victim’s Survey, and in the UK it is the Crime Survey for England and Wales (CSEW) (with separate surveys for Scotland and for Northern Ireland). These surveys measure the public experience of crime. They are considered to provide a more accurate picture of crime trends than crime data derived from police reported crimes because they include crimes that have not been reported to the police, and they are not subject to variations in police data recording practices.[1]

Crime data are typically made available to the public in three main ways: in the form of statistics compiled and generated by national statistical (or other) agencies; as part of visualizations such as crime maps produced or commissioned by police forces; or as accessible and open data. While statistical data are generally now available as accessible and open data in some jurisdictions, the more textured local data used in these public-facing crime maps are much less widely available. It should also be noted that the data used in visualizations are often not the same as those used to generate crime statistics; some visualizations use emergency call data or police response data to populate their maps (Scassa 2016).

## The limits of crime data

The superficially descriptive nature of crime data lends them an aura of objectivity. They are used by public authorities in accounts of the city and may impact decisions regarding the allocation of resources; they are also used by the private sector, for example, in commercial decisions such as the setting of insurance rates. Crime data may also influence personal decisions such as where to walk, shop, live or attend school. However, crime data are affected by a variety of factors which make them imperfect measures of criminal activity.

### *Institutional factors*

Lomell (2011) observes that the objectives for collecting crime statistics can change over time, and these changes can impact the nature and quality of the statistical information. In her study of annual police reports of crime in Oslo from 1950 onwards she notes a shift in the reporting of crime data from ‘inputs’ to ‘outputs’. She attributes this shift to the impact of new public management approaches that have come to dominate the public sector in Europe and North America since the 1980s and 1990s (Lomell 2011; Hood 2007; Eterno and Silverman 2012). Prior to that time, crime data were recorded as measures of police workload (input);

that is, how much crime do police encounter. With the rise of public service performance-based metrics, crime data shifted to a measure of police performance (output); that is, how do police activities impact the incidence of certain crimes? Although in both cases what are recorded are crime data, a shift in motivation for recording the data affects what is recorded and how.

Eterno and Silverman (2012: 86) argue that performance-led policing in both the UK and the US arose 'from a fear of crime, drive for greater police managerial accountability, and enhanced business-oriented police operations'. The use of such systems in police management requires departments to set performance standards and then to rigorously monitor progress towards these goals (Eterno and Silverman 2012). Not surprisingly, the performance standards in the context of policing are typically expressed in terms of a reduction in crime or at least of certain types of crime (Chainey and Tompson 2012; Yung 2014). Police forces may have limited control over the incidence of crime – particularly where cuts to budgets and staffing make their jobs more difficult and where crime may be driven by circumstances well beyond police influence such as unaddressed economic inequality and social injustice. However, they do have considerably more control over data about crime. The manipulation of crime data by police forces – particularly in order to meet performance standards – is well documented (Eterno and Silverman 2012; Lomell 2011). Indeed, in 2014 the UK Statistics Authority ruled that crime statistics were no longer to be designated as national statistics because of concerns over 'some aspects of the police's recording of crime data' (UK Statistics Authority 2014: 1). These data may be specifically manipulated in ways that fit a public narrative around crime control and public money well spent.

### *Police as sensors*

Another influence on crime data is the manner in which such data are generated and recorded. Crime data are the result of reports made to the police and actions taken in response to those reports. At a very basic level therefore, crime data depend upon two key factors: the willingness of a victim to report the crime to police, and the responding officers' assessment of the incident (Fisher 1993).

The handling of sexual assault cases offers a good illustration of how each of these factors may affect the resultant data. In instances where there is a high distrust of the police, or of the criminal justice system, there may also be a significant underreporting of crimes. This is particularly (although not exclusively) the case with sexual assault (Russell 2010; Fisher *et al.* 2002). Reluctance to report may be linked to police attitudes towards victims; it may also be linked to concerns over the shortcomings of the justice system in dealing with such crimes. In some cases, shame or fear may prevent reporting. Whatever the reason, the substantial underreporting of sexual assault is well-documented (Johnson 2012).

Even if a sexual assault is reported to police, how that incident is recorded may be highly dependent upon factors that are unrelated to what actually occurred (Spohn *et al.* 2014). For example, a subjective belief on the part of the responding officer that women tend to fabricate stories of sexual assault can lead to the recording of incidents as 'unfounded' (Women's Law Project 2013; Johnson 2012; Hattem 2007; Gilsinan 2012). A belief that sex trade workers cannot be sexually assaulted, or that some women invite sexual assault can similarly lead to the unfounding of complaints (Women's Law Project 2013). Victim characteristics such as race, gender, socio-economic status or sexual orientation may also affect how complaints are dealt with and recorded. In some cases, police officers may be influenced by their perception of whether a conviction is likely in deciding whether to 'unfound' a complaint (Spohn *et al.* 2014; Police Executive Research Forum 2012).

If police are the 'sensors' that record crime data, then factors that remove crimes from police attention also limit the data (Gilsinan 2012). Police are often not the first point of official contact for victims of crime. Public or quasi-public authorities, such as public transit agencies or colleges and universities, may have their own security personnel who are frequently a first (and last) point of contact for complaints (Russell 2010). The growing privatization of police services (Stenning 2009), a result of neoliberal trends that shift the responsibility for security to the private sector and to ordinary individuals (Van Steden and Sarre 2007), also means that a high number of incidents may be dealt with by private security personnel. Van Steden and Sarre (2007) observe that the private security industry is diverse and multi-sectoral. Even in 2007, they found that the numbers of employees in private security industries exceeded those on police payrolls in the US, Canada and the UK, and privatization of these services has increased since that time (Rahall 2014). It may be difficult to ascertain how many complaints in these contexts are reported to police and how many are dealt with internally. A poor victim experience in a private context may also discourage the victim from taking his or her complaint to the police.

### *Choosing data points*

Although police files themselves may be rich sources of contextual data, not all information about incidents makes it into public-facing crime data. The data are thus also influenced by decisions as to what data points are considered useful or appropriate for sharing. Recognizing the growing importance of more contextual crime data, statistical agencies have revised their systems over time to permit the recording of additional data points, including some demographic information about victims and perpetrators.

The relevance or usefulness of particular data points may be contested. For example, data regarding the race or ethnicity of those stopped or investigated by police could be used to examine whether there are systemic biases in policing, although how such data are derived is fraught with challenges. Data about race/ethnicity is frequently reported from the perspective of the police officer and not the person who is detained or arrested. Such data can also be used to construct racist narratives around the criminality of certain groups or communities. Concerns over these issues led to data on race/ethnicity being excluded from official crime data in Canada, although some advocates have called for a return to the recording of race-based data as a means of monitoring the way in which some communities are policed (Owusu-Bempah and Millar 2010). Other data points may be omitted out of concern that their inclusion may lead to the reidentification of individuals, thus breaching data protection laws. Where preoccupations over how data may be used result in more limited data, their usefulness, including for transparency and accountability purposes, is diminished (Conroy and Scassa 2015).

## Communicating crime data to the public

According to Chainey and Tompson (2012: 230), crime data are made available to the public for a number of reasons, including 'to improve the credibility of crime statistics, address often over-inflated perceptions about local levels of crime, provide crime information that engages the public on local crime issues and empowers them to make decisions that improve their personal safety, and contributes to local community safety'. Notably, transparency and accountability are not in this list, perhaps because in most cases, crime data are presented to the public as part of official narratives about crime. Currently crime data are made available to the public in two main ways. The first is as statistical data; the second is through visualizations.

Canada, the UK and the US all collect crime data at local, regional and national levels. Typically, there is a centralized and standardized reporting mechanism which allows for data to be collected, compiled and analysed. The national compilation of these local statistics reveals the broader interest in crime data; indeed, these statistics are used by different levels of government in relation to criminal justice, crime control and security agendas. Statistical crime data are now routinely made available as open data at the national level in all three jurisdictions. This includes victimization survey data.

While national crime statistics tend to be available as open data, local data may or may not be as easily accessible. In the UK, national crime data are available in separate data sets for particular regions and by police force. In Canada and the US, whether crime data are made available as open data by a municipality will vary by municipality. The parameters of the data sets may also vary. The lack of consistent practice or standard formats may be due to the fact that both Canada and the US are federal states; federal systems allow for considerably more divergence at the regional level in terms of law, policy and practice. Although these differences can make useful analytical comparisons across jurisdictions almost impossible, even without these variations, such comparisons would be of dubious value. This is because of the sometimes significant variations from one police force to another in how data are required to be recorded, as well as issues of local and institutional culture.

One method of making crime data publicly available that has gained great popularity with police forces in Canada, the US and the UK is the use of visualization. A significant number of urban police forces now provide interactive online maps that display incidents of crime within city boundaries (Scassa 2016). In the UK, the mapping of crime data was part of a police pledge made at the national level (Sampson and Kinnear 2009). Mapping is seen as a vehicle for providing information to the public in an accessible and interactive format (Chainey and Thompson 2012). Crime maps allow users to view reported crimes on local maps, and to zoom in on particular streets or neighbourhoods. Typical functionality allows users to view reported crimes in specific temporal periods (e.g. the last 7, 14 or 28 days). Some maps allow users to register for email updates on crimes occurring within certain defined geographic parameters.

Visualizations of crime data do not solve the problems inherent in the data; these will necessarily be replicated in the visualization. As Wallace (2009: 16) notes, 'crime-mapping is entirely dependent on the categories and bureaucratic practices of local police forces'. Mapped crimes are only those that were brought to the attention of the police and that were recorded by police as a particular type of crime. Some maps rely entirely on emergency call and police response data. This may result in multiple reports of the same occurrence, and it may also mean that the mapped site relates to the location of the report and not that of the event (Scassa 2016). Further, not all crimes are featured on crime maps. Crimes that do not have a clear location such as fraud, counterfeiting, other white-collar crimes and Internet crimes tend to be absent from such maps (Wallace 2009). In cases where mapped data are based on call-for-service data, incidents that are not the subject of calls to the police are not represented. The absence of commercial and computer-based crimes from crime maps can lead to a skewed perception of the nature and incidence of criminal activity within the city, as well as the geographic dispersion of criminal activity.

Because crimes involve human victims and public sector authorities are constrained by privacy law in the release of data, many types of publicly accessible or available crime data, such as statistics and visualizations, do not include individually identifying information. Since location can be a powerful identifier, crime maps typically reflect steps taken to prevent re-identification. In the case of crimes where the victim's identity is protected by law (e.g. sexual assault or crimes involving children) or where the identity of the perpetrator is protected by law (e.g. crimes by juvenile offenders), such incidents or categories of crime may simply be excluded from the visualization (Sampson and Kinnear 2009), although this is not always the case. More general privacy concerns are typically addressed by reporting the crimes at the 100-block level. These alterations may have additional implications for the quality of the mapped data (Chainey and Tompson 2012).

The fact that not all crime is included in visualizations has been the subject of some criticism. Wallace (2009: 14) suggests that neoliberal biases lead to a focus on the reporting of vehicle and property-related offences, which are 'most central to a private property ownership society'; along with what she characterizes as 'quality of life' crimes such as vandalism or prostitution. Crime mapping may therefore integrate selected data into a particular narrative of crime and crime control, weaving that narrative into a geographic representation of the city. Wallace (2009: 16) notes that 'crime maps reduce the complexity of the space and the event' with the result that crimes are clustered along streets and their representation depends upon the ability to link a crime to a particular geographic space. Even though statistical agencies maintain that victimization surveys offer a better picture of the incidence of crime than data derived from police reports, it is this latter category of data that is used in crime maps. The visualizations are thus also incomplete in ways that may not be well-appreciated by the average user.

Wallace (2009: 5–6) also offers a broader critique of crime maps, arguing that these visualizations of crime data create for the public ‘a new aesthetic of danger’ and place the onus on individuals to ‘ultimately take responsibility for their own personal safety’. It is possible that the typical option to sign up for email alerts of crime in one’s neighbourhood also contributes to a subtext of individual responsibility. According to Wallace (2009: 7), ‘faced with an environment of failing social services and a crumbling state infrastructure, individual citizens now have the burden of using new media technology to supply information for protection against danger’.

Crime maps may also be offered to the public as an alternative to open crime data (Lofaro 2015; Wisnieski 2014). Some of the crime data mapping companies, such as *Bair Analytics* (RAIDS Online) and *Public Engines* (Crime Reports) provide public-facing crime maps to police department clients at no cost when those agencies contract for other data analytics services (Paulsen and LeBeau 2012). The terms of use for these sites specifically provide that the police departments supplying the underlying data remain the owners of any intellectual property rights in the data; however, the terms of use also specifically prohibit any re-use of the data embedded in the maps by any users of the visualizations (Scassa 2016). These terms of service create a barrier to data re-use (Hochberg 2014). It remains the law enforcement agencies’ choice to disclose (or not) the same data as open data. Where data are not provided as open data, the potential for alternative visualizations of urban crime is significantly limited.

## Transparency and crime data

In spite of their deficiencies, crime data do provide some account of police activity. Treated with the caution they deserve, they can be used in an assessment of that activity and in decisions regarding the allocation of resources. More critical analyses of crime data can be used to build arguments about the fairness of resource allocation, police practices, or about failures of the system to properly address certain types of crimes. For example, comparisons between crime data derived from police reports and victimization surveys can reveal problems such as a marked gap between the experience of certain types of crime such as sexual assault and the actual reporting of such crimes to police. A sharp rise in lesser theft offences accompanied by a surprising decrease in more serious related offences may be an indicator of changes not in crime itself but in the way in which it is being recorded, and may thus support critical challenges to the integrity of the data regarding police activity, as well as to the effectiveness of police efforts to control crime.

However, as noted above, the shortcomings of the official data suggest that open crime statistics and publicly available visualizations based upon incident data are not likely fully to achieve the goals of transparency or to meet the need for data to understand crime within the city. One way to improve the usefulness of the data is to add more data points. Crime data are becoming richer, with data about more types of crime being collected and made available. For example, in the UK open crime data are supplemented with other crime-related data sets, including stop and search data, forensic data and criminal justice statistics including data on women and the criminal justice system and race and the criminal justice system. Statistical agencies have also moved from aggregate counts of crime to more incident-based reporting, permitting a somewhat more contextual analysis of crime. In the US, the *Police Data Initiative*, launched by the White House in May 2015 was designed in part to make police data more available as open data and to encourage and support civil society in making use of these data.

Crime data can also be compared to other available data. Some such data may be in a format that facilitates analysis. Eterno and Silverman (2012) used hospital data, data on non-index crime and historical crime data (along with qualitative data) as part of their critical analysis of crime data recorded by the NYPD. Other data sources can include information from newspaper accounts of incidents. Community or advocacy groups may also compile and publish their own statistics. For example, many rape crisis centres provide data regarding how many calls they received or women they assisted (Lombardi 2009) creating sexual assault data that can reveal important gaps in official crime data (Yung 2014).

Greater transparency may also be sought from public authorities that carry out security or policing functions. For example, transit authorities can be pushed to produce more and better data regarding incidents handled by transit police services and the police. Martin (2011) notes that data regarding transit crime is sparse; there is no agreed upon definition; and with the exception of the British Transport Police, there is no organized reporting of data or public release of such data as open data.

While more crime-related data from other sources can enrich the understanding of urban crime as well as of responses to crime, new incident-based data will likely suffer from the same problems evident in police crime data. For example, in the US, federal legislation in the late 1990s imposed requirements on all colleges and universities receiving federal funds to report campus-based crime handled either by campus security or police. The legislation prescribed the required data points and set out the ways in which data had to be reported and made publicly available. The goal was to increase transparency and accountability and to improve public safety. The use of legislation to impose data reporting requirements on private or quasi-public actors is interesting. However, perhaps not surprisingly, the effectiveness of the Clery Act[2] in generating accurate and useful data has been questioned for many of the same reasons that the integrity of crime data has been challenged (Fisher *et al.* 2002). Inconsistencies in how crimes such as sexual assault are defined by universities and colleges affect reporting rates (National Institute of Justice 2005), and there are concerns that reporting may also be affected by institutional concerns about public image in the face of competition for enrolment (Lombardi 2009). The same issues of underreporting of crimes by victims – particularly sexual assault – have also arisen with respect to campus crime (Gardella *et al.* 2014; National Institute of Justice 2005).

The growing privatization of policing services also poses some significant challenges to developing fully accurate data about the incidence of crime. Where data are collected by private organizations, it is shielded from access to information legislation. Such data will not be compiled and made more broadly available without the imposition of reporting requirements, and these are likely to be resisted by the private sector on the grounds that the information about their activities is confidential business information. In any event, as evidenced by the experience under the Clery Act, such data are also likely to face the same issues of integrity as crime data. In addition, corporate culture, including accountability to shareholders rather than the public, and concerns over legal liability, will shape any such data sets. As public and private sectors become further intertwined in delivering 'smart cities' to the urban public, the clash between the need for data transparency and claims to confidential business information will need to be addressed.

In some cases, the crowd-sourcing of unofficial crime maps has been used as a means of providing a counter-narrative to official visualizations. For example, HarassMap[3] invites women to map and describe incidents of sexual harassment in Egypt, within a cultural context in which such harassment typically remains invisible and unreportable. While not a crime map *per se* (as it compiles data on harassment that ranges from criminal to non-criminal conduct), this is an example of how the crowd-sourcing of data can draw attention to a problem, and can be used in powerful visualization tools to create an alternative representation of the gendered experience of urban space.

While crowd-sourced maps are interesting, they present many challenges. Stability and continuity are one challenge – both in terms of management of the project as well as in terms of public participation. In the case of crimes like sexual assault there are also significant privacy issues, risks of further victimization, other liability issues, as well as risks that first-person reports could be used by defence lawyers in criminal prosecutions to discredit victim witnesses. Nevertheless, crowd-sourcing remains an interesting tool particularly in contexts where certain crimes are not dealt with effectively by public authorities (Friedman 2014).

## Conclusion

Crime data are a good example of 'capta' – data that have been selectively harvested from a broad pool of available data. They rely upon human 'sensors' for their recording and interpretation, and reporting and recording are shaped by bureaucratic demands and priorities. Crime data are also dependent upon the willingness of victims to report incidents to authorities. Systemic problems that lead to underreporting are made invisible in official statistics. Crime data are distilled from interactions between state and citizen as interpreted and recorded by agents of the state. As such, official crime data are also affected by institutional culture and legal constraints in both their recording and their representation.

In spite of these issues, crime data are used to drive decision-making around policing and are used by public authorities to justify spending, action or inaction. They are used as well in visualizations that communicate public narratives of urban crime. The multiple uses for such data suggest a need both for critical approaches and for analytical strategies to distil useful knowledge from the data and to improve it. While statistical crime data are more generally available as open data, urban crime data remains largely under the control of municipal police forces. Even where crime data visualizations are made publicly available, the underlying data may not be. Open data allows for different analyses, mashups and alternative narratives of crime. Both open data and alternative sources of crime data, including crowd-sourcing, may also assist in understanding urban crime.

Some or all of the characteristics of crime data are shared by other types of urban data. The lessons from crime data therefore have broader application. A critical perspective on how, by whom, and for what purposes data are collected is essential to moving past the superficial objectivity of data. Further, open data at all levels of government is crucial to allowing diverse perspectives and analyses that can inform public discourse and challenge official narratives.

## Acknowledgements

I gratefully acknowledge the support of the Canada Research Chairs programme. Thanks to Amy Conroy for her helpful comments on an earlier draft of this chapter.

## Notes

1 Office for National Statistics. Crime Survey for England and Wales: Guidance and Methodology, www.ons.gov.uk/ons/guide-method/method-quality/specific/crime-statistics-methodology/guide-to-finding-crime-statistics/crime-survey-for-england-and-wales--csew-/index.html.

2 20 U.S.C. §1092(f) (2000).

3 See http://harassmap.org/en/.

## References

Chainey, S. and Tompson, L. (2012) ‘Engagement, empowerment and transparency: Publishing crime statistics using online crime mapping’, *Policing* 6(3): 228–239. Doi: 10.1093/police/pas006.

Conroy, A. and Scassa, T. (2015) ‘Promoting transparency while protecting privacy in open government in Canada’, *Alberta Law Review* 53(1): 175–206.

Eterno, J.A. and Silverman, E.B. (2012) *The Crime Numbers Game: Management by Manipulation*. Boca Raton, FL: CRC Press.

Fisher, B.S., Hartman, J.L., Cullen, F.T. and Turner, M.G. (2002) Making campuses safer for students: The Clery Act as a symbolic legal reform. *Stetson Law Review* 32: 61–89.

Fisher, S.Z. (1993) ‘Just the facts, ma’am: Lying and the omission of exculpatory evidence in police reports’, *New England Law Review* 28(1): 1–62.

Friedman, U. (2014) ‘Crowdsourcing crime reporting in the world’s murder capitals’, *The Atlantic: CityLab*, available from: www.citylab.com/tech/2014/05/crowdsourcing-crime-reporting-worlds-murder-capitals/9058/.

Gardella, J.H., Nichols-Hadeed, C.A., Mastrocinque, J.M., Stone, J.T., Coates, C.A., Sly, C.J. and Cerulli, C. (2014) ‘Beyond Clery Act statistics: a closer look at college victimization based on self-report data’, *Journal of Interpersonal Violence* 30(4): 640–658.

Gilsinan, J.F. (2012) ‘The numbers dilemma: The chimera of modern police accountability systems’, *Saint Louis University Public Law Review* 32(1): 93–110.

Hattem, T. (2007) ‘Highlights from a preliminary study of police classification of sexual assault cases as unfounded’, *JustResearch* 14: 32–36, available from: www.justice.gc.ca/eng/rp-pr/jr/jr14/jr14.pdf.

Hochberg, A. (2014) ‘Disputes over crime maps highlight challenge of outsourcing public data’, *Poynter*, 24 November, available from: www.poynter.org/news/mediawire/213443/disputes-over-crime-maps-highlight-challenge-of-outsourcing-public-data/.

Hood, C. (2007) ‘Public service management by numbers: Why does it vary? Where has it come from? What are the gaps and the puzzles?’, *Public Money & Management* 27(2): 95–102.

Johnson, H. (2012) ‘Limits of a criminal justice response: Trends in police and court processing of sexual assault’, in Elizabeth A. Sheehy (ed.), *Sexual Assault in Canada: Law, Legal Practice and Women’s Activism*. Ottawa: University of Ottawa Press, pp. 613–634.

Kitchin, R. (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: SAGE.

Lofaro, J. (2015) ‘Adding crime data to city’s open data portal expensive, says Ottawa police CIO’, *Metro* (Ottawa), available from: http://metronews.ca/news/ottawa/1440605/adding-crime-data-to-citys-open-data-portal-expensive-says-ottawa-police-cio/.

Lombardi, K. (2009) ‘Campus sexual assault statistics don’t add up’, The Center for Public Integrity, available from: www.publicintegrity.org/2009/12/02/9045/campus-sexual-assault-statistics-don-t-add.

Lomell, H.M. (2011) ‘Making sense of numbers: The presentation of crime statistics in the Oslo Police annual reports 1950–2008’, in Ann Rudinow Saetnan, Heidi Mork Lomell, and Svein Hammer, *The Mutual Construction of Statistics and Society*. New York: Routledge, pp. 191–223.

Martin, J. (2011) ‘The incidence and fear of transit crime: A review of the literature’, University of the Fraser Valley: Centre for Public Safety and Criminal Justice Research, available from: www.ufv.ca/media/assets/ccjr/reports-and-publications/Transit_Crime_2011.pdf.

National Institute of Justice (2005) ‘Sexual assault on campus: What colleges and universities are doing about it’, US Department of Justice, Washington, DC.

Owusu-Bempah, A. and Millar, P. (2010) ‘Research note: Revisiting the collection of ‘justice statistics by race’ in Canada’, *Canadian Journal of Law and Society* 25(1): 97–104.

Paulsen, D. and LeBeau, J. (2012) ‘Survey and evaluation of online crime mapping companies’, Version 1.2., US Department of Justice, available from: www.ncjrs.gov/pdffiles1/nij/grants/239908.pdf.

Police Executive Research Forum (2012) *Improving the Police Response to Sexual Assault*. Washington, DC: Critical Issues in Policing Series.

Rahall, K. (2014) ‘The siren is calling: Economic and ideological trends toward privatization of public police forces’, *University of Miami Law Review* 68(3): 633–676.

Russell, L. (2010) ‘What women need now from police and prosecutors’, *Canadian Woman Studies* 28(1): 28–36.

Sampson, F and Kinnear, F. (2009) ‘Plotting crimes: Too true to be good? The rationale and risks behind crime mapping in the UK’, *Policing* 4(1): 15–27.

Scassa, T. (2016) ‘Police service crime mapping as civic technology: A critical assessment’, *International Journal of e-Planning Research* 5(3): 13–26.

Spohn, C., White, C. and Tellis, K. (2014) 'Unfounding sexual assault: Examining the decision to unfound and identifying false reports', *Law & Society Review* 48(1): 161–192.

Stenning, P. (2009) 'Governance and accountability in a plural policing environment – the story so far', *Policing* 3(1): 22–33.

UK Statistics Authority (2014) 'Compliance with the Code of Practice for Official Statistics: Statistics on crime in England and Wales', UK Statistics Authority, London, available from: www.ons.gov.uk/ons/rel/crime-stats/crime-statistics/period-ending-june-2014/sty-de-designation.html.

Van Steden, R. and Sarre, R. (2007) 'The growth of privatized policing: Some cross-national data and comparisons', *International Journal of Comparative and Applied Criminal Justice* 31(1): 51–71.

Wallace, A. (2009) 'Mapping city crime and the new aesthetic of danger', *Journal of Visual Culture* 8(1): 5–24.

Wisnieski, A. (2014) 'Is your city's crime data private property?' *The Crime Report*, 5 May, available from: www.thecrimereport.org/news/inside-criminal-justice/2014-05-is-your-citys-crime-data-private-property.

Women's Law Project (2013) 'Policy brief: Advocacy to improve police response to sex crimes', Women's Law Project, Philadelphia, PA, available from: www.womenslawproject.org/resources/Policy_Brief_Improving_Police_Response_to_Sexual_Assault_Feb2013_FINAL.pdf.

Yung, C-R. (2014) 'How to lie with rape statistics: America's hidden rape crisis', *Iowa Law Review* 99(3): 1197–1256.

# 6 Data provenance and possibility

## Thoughts towards a provenance schema for urban data

*Jim Thatcher and Craig Dalton*

## Introduction

Data suffuse society, underpinning how the world is known, planned and governed. The era of 'big data' marks a surge in the generation of diverse, exhaustive data streams in real-time that can be combined, analysed and exchanged. Data come as deluge: persistent, swelling streams of information flowing at such rates as to make storage and analysis Sisyphean feats of engineering. These data shape how we come to know and experience the world, guiding the systems with which we interact and tailoring the paths we take to a quantified representation of ourselves and others. For society, the ever-greater production and processing of data redefine efficient production, security, privacy discourses, the role of the state and scientific knowledge. In cities, the new big data era fosters 'smart' design and growth discourses that fetishize data in urban life, design, governance and planning. However, for all that data have come to influence, represent, enable and constrain, the *provenance* of urban big data is often left underexplored.

The provenance of a piece of art traces its possession from the artist's hand to its current owners. It is the key means of authenticating and assessing the value of a work. For data, provenance has many related meanings, but broadly refers to 'information about the origin, context, or history of the data' (Cheney *et al.* 2009: 959). The US Department of Homeland Security (2009) has identified data provenance as one of the 'hardest and most critical challenges that must be addressed' for information security (INFOSEC Research Council 2005), one whose solution would significantly improve the nation's national information security infrastructure. This understanding of data provenance conceptualizes it as a technical question about metadata, one that presumes a technical solution is not only possible, but desirable, not only for data practitioners, but as a general good for everyone.

In this chapter, we use a critical data studies approach to push back against such instrumental understandings of data provenance. We argue that if cities are to be 'smart', then they must move beyond the creation, manipulation and analysis of data that accepts data as ontologically given, static and discrete phenomena. To do so, we problematize technical definitions of data provenance to propose a more contextual form of provenance for urban spatial data that emphasizes its mediated and constantly changing nature amidst society and technology. The chapter proceeds in three parts. First, we introduce current technical approaches to data provenance. Second, we problematize those approaches in current urban spatial data by highlighting two interrelated issues: a failure to differentiate individual people from data points derived from their actions, which, in turn, results in an inscription of meaning into data that may not exist in the world. Finally, we conclude by outlining standards for a 'more than' technical approach to provenance. Building from critical ideas around metadata and the concept of the data-encounter, we propose a technically agnostic, contextual schema for the provenance of urban spatial data that involves both information traditionally found in metadata and situated considerations of motivation, value and power as the meaning and use of data develops through time and space.

## Provenance in current practice

Much like many other apparently unitary concepts, the terms *data* and *provenance* have multiple and contested meanings across a wide variety of contexts and professional fields. Data provenance as it is currently understood and practised varies a great deal depending on the social purposes and resulting standards behind their operationalization. Exploring these differences highlights the inherently 'more than technical' nature of provenance.

Provenance itself has a long and developed history in the archival and library sciences. In archival work more broadly, provenance refers to 'the origin or source of something, or . . . the person, agency or office of origin that created, acquired, used and retained a body of records in the course of their work or life' (Miller 2010: 98, in Poole 2015: 114).[1] This principle of organization has been formalized as the *respect des fonds* and requires a level of temporal and spatial control of the object that has been called into question for near-infinitely replicable digital data (Millar 2002). For scientists, Lauriault *et al.* (2007) note how provenance, along with terms like lineage and integrity, are used as a catch-all for data authenticity. Theoretically, at its strongest, provenance records not only the source, methods of collection, and any transformations, but also the biases and assumptions that went into said processes. Provenance, broadly refers to 'the source (or derivation) of an object and [to] the record of the derivation' (Moreau *et al.* 2008: 54). For data, it involves the origins and history of the data and is a core means by which data are trusted and reused (Cheney *et al.* 2009; Donaldson and Fear 2011).

While pieces of art with unestablished provenance are viewed with scepticism by researchers and buyers alike, the provenance of digital data is often less carefully vetted, especially when it is produced by what is considered to be an authoritative source, such as a government department or a mainstream news agency. However, there are numerous examples of weak metadata producing data and stories with suspect provenance. For example, in 2008, Google algorithms identified a news story from the *South Florida Sun Sentinel* about United Airlines 2002 bankruptcy that lacked a date, automatically assigned it the current date, and posted it on Google News (Zetter 2008). Investors, failing to check the factual provenance of the story (in this case, temporal accuracy), panicked and United's stock briefly lost three-quarters of its value. Such dangers are even greater in situations where experiments and data gathering are slow and difficult to replicate, such as policy research (Reichman *et al.* 2011: 704). Data provenance failures can result in massive societal and economic losses (see, for example, discussions around the liability of geographic information/data in Onsrud 1999 and Phillips 1999). While urban policy and planning have long been guided by data, big urban data, performance metrics and data analytics are increasingly shaping urban policy and planning (Kitchin 2014a; Townsend 2013). The provenance of data, particularly geographic data (Monmonier 1995), must then be firmly established to produce trust and the necessary information required to avoid possible misinterpretation or misuses of the data.

Definitions for and examples of data provenance can be found in methodological guides, data dictionaries, metadata abstracts and their accompanying articles, survey questionnaires and elsewhere. As noted, researchers tend to see data provenance as a technical problem with two basic paths to resolution. In an influential 2001 piece, Buneman *et al.* distinguish between 'where-provenance' and 'why-provenance'. Both are technical problems with discrete solutions. 'Where-provenance' refers to the original source of the data, while 'why-provenance' records all 'source data items that contributed to the creation of a result data item' (Glavic and Dittrich 2007: 229). Tan (2004) describes the situation slightly differently, marking a distinction between 'provenance of data' as a focus on the creation processes of the data (in raw form) in contrast to 'provenance of a data product' which focuses on how data have been processed and transformed into a derived form.[2] Additionally, provenance can be generated at the time of creation through the production of associated metadata, which Tan refers to as 'eager' provenance, or when it is requested, 'lazy' provenance. In professionally produced data sets provenance is usually formally recorded as metadata; however, Cheney *et al.* (2009) note that for many systems provenance is created in an ad hoc fashion, cobbled together to meet perceived or immediate needs.

We emphasize these multiple forms of provenance to highlight that, whether dealing with data transformations or focusing on the original source of data creation, current approaches to data provenance are predominantly technical, capturing factual, recorded information about a data set (e.g. date/time produced, instrument and standard used, data structures, file format, etc.). For example, the Provenance Aware Storage System developed by Harvard University creates provenance by 'capturing file system events in an operating system' (Moreau *et al.* 2008: 58) and the Chimera Virtual Data Catalog allows for the recording of 'transformations, data objects and derivations' (Glavic and Dittrich 2007: 230). While there have been attempts to standardize provenance (see, for example, ISO 19115), the standards are often not practised, as domain experts instead create bespoke systems to capture information most pertinent for intended use. Less clear and often elided from such forms of provenance are the recursive sociotechnical processes that go into the making of data, including handling practices, scientific assumptions, and methodological biases. These considerations are necessary to establish strong provenance (Lauriault *et al.* 2007). From a Critical Data Studies standpoint, the absence of a 'more than technical understanding' of provenance limits what can be known, much less can and should be done with urban data.

## Limited possibilities, problems from provenance

The rush to 'real-time' and 'smart' cities is predicated upon the ability to leverage big, urban data to make cities 'better': more efficient, more profitable and more livable, at least for certain sets of individuals. A litany of critiques push back against the unilinear myth of improvement sold via big data/smart city parables (Batty 2013; Kitchin 2014b; Crawford 2013; and others). Reduced to a technical problem, current approaches to data provenance reduce it to one more hurdle to overcome through standardization and computation. In contrast, drawing on Critical Data Studies (Dalton and Thatcher 2014; Kitchin 2014c; Taylor 2015), we highlight how current approaches to data provenance set previously unexamined limits on the use of urban data. Specifically, we identify two interrelated factors that circumscribe utility. First, many current sources of data are corporate in nature and, following imperatives for profit generation, provenance can be subsumed into privately owned, black-boxed systems. Second, as data are collected, they are categorically inscribed with meaning that may or may not align with the data creator's intent.

### *Provenance and profit*

At times, the rhetoric around data treats them in an essentialist manner as an almost natural by-product of existence; given off like body heat and lost to entropy unless captured and stored, as exhaust data. Such a by-product orientation is increasingly evident in the health care industry as it tries to capture every heartbeat of an as-yet born foetus through monitoring (Smith 2014). Similarly, the quantified-self movement promises better living through the routine, everyday capture and analysis of personal data (Wolcott 2013). However, actual data are not inherently produced as a function of being alive, but through the active and passive recording of actions. And just as a forest and a map of it are distinct, data about a person are merely representative rather than constitutive of them. Despite this, many current data practices seek to collapse said difference. In some cases, this collapse is necessary: if the result of a medical test is not true for the patient being tested, it serves no purpose. Further, in such situations a strong and uncompromised provenance is necessary to link the individual to test result. However, in many existing big data analytical practices, that correspondence is less integral, subsumed in the interest of profit generation.

Beyond the use value of a medical test, the market value of collapsing an individual to their representative digital identity, the assemblage of data points that refer to them, is potentially massive (Thatcher 2017; Dalton *et al.* 2016). For example, in 2011, Acxiom, a single firm selling aggregated personal data points, recorded $1.1 billion in revenues (Roderick 2014). Managing that much data necessarily involves metadata standards and provenance. However, the forms of data and thus considerations of provenance are structured in terms of a particular company's strategies and profit imperatives. For Acxiom, which sells data to advertisers, not consumers themselves, provenance is an issue to the extent that it validates data about consumers that appears to predict their (potential) consumptive behaviour in the future. This is quite different from data protocols for the results of a medical test in the past to be shared with a patient. A great deal of the value of companies like Google, Acxiom and Facebook is based on a strong presumed correspondence between someone and their data and metadata. For such corporate enterprises, considerations of data provenance are contingent upon the larger purpose of profit generation and the data practices it drives.

This is not an inherently new process as social researchers have been building data sets for the purpose of socially targeting people at small urban scales for decades, such as geodemographically targeted marketing segmented by neighbourhood. Digital data sources open a plethora of sources for collection and combination (mobile phones, government records, social media accounts, frequent shopper programs, credit card transactions and other data points). As the price for storing and processing massive amounts of data declines, companies are developing data 'doppelgängers' (Robinson 2008), highly individualized, data-based digital representations of individuals, pioneered by consumption and targeted advertising and numerous data segmentation procedures and services. Presently, Netflix, a service for renting and viewing media, uses 76,000 unique data categories to create individualized profiles of its more than 65 million subscriber accounts[3] (Madrigal 2014).

Market competition between data-driven corporations drives the black-boxing of data production and their provenance and attempts to rigorously control access and veracity. Such corporations not only include targeted advertisers, but also those involved with local public service provision such as trash collection, law enforcement analytics and transportation firms such as Uber. Furthermore, the purpose of leveraging such data is competitive advantage, not austere truth or societal good. As black-boxed trade secrets, the provenance of digital data produced or processed by corporations, or even the commodity chain that produced a single data set, is impossible to trace from the outside. The intentions and the biases they introduce underpinning data set creation are part of the 'added value' corporations bring to their client cities and/or users' lives (Thatcher 2014). For private corporations, provenance can be highly variable, depending on the domain and purpose of the data for the company. For example, financial data, such as invoices and bank transfers, must be closely vetted because such payments directly affect the bottom line and legal liability. However, data generated and used for other purposes that carry less risk and legal liability, such as consumer targeting, is often less closely vetted. On the one hand, this seems obvious as risk factors will differ: misdiagnosing an individual with cancer has far graver repercussions than sending them a discount on used car tires for which they have no use.

This distinction in significance of provenance resembles Kitchin and Dodge's (2011) demarcation between coded spaces and code/space. The former are examples of code used to enhance or extend the experience of space, the latter represent moments where spatial experience is fundamentally created *through* the function of code. While the provenance of data (and code) need not be spatial, these distinctions parallel some of the differing requirements for data provenance with respect to how the data may be used. Financial data, such as payment information, must directly correspond to exterior objects or precise amounts of existing value, which require a strong provenance to function. A company's systems grind to a halt without it. However, a lack of provenance for much of the data involved in the epistemological leap from individual to inscribed data does not stop the company's system of consumer profiling or its generation of profit (Dalton and Thatcher 2014). For example, in Foursquare, the intentionality underpinning user check-in data matters little as long as it generates business and profit: the provenance of the data is less important than its generative economic function. In practice, the users, and data pertaining to them, are the product and those buying advertising are the clients (Zittrain 2012). The bar for actionable, if not 'true', information is itself defined not by researched standards, but by the market. Even if data provenance and accuracy were not closely guarded trade secrets, the primary indicator of data worth is its literal exchange value: data are valuable because they, or products based upon them, can be sold. In many cases, if a company strives for more accuracy than the competition, they would not necessarily reap increased rewards. The price of a trip on a ride-share service depends not just company data, but also on the market as a whole. A company that charges much more than the competition is unlikely to survive. As long as the data are useful in leveraging a profit, as long as they are superior to their competitors in some way, the imperatives of capital are satisfied.

For open-data initiatives, particularly those pertaining to cities, this collapse has profound significance. Instead of seeking market value, open-data initiatives advocate making data free and available for all. The argument is that if more data, often paid for and maintained at taxpayer expense, were available, then more individuals would naturally leverage them to address problems within the city. While there have been both successes and failures in such an approach (Townsend 2013; Goldsmith and Crawford 2014), such data suffer an additional problem with regards to provenance. While for-profit endeavours often do not seek an ultimate truth only a market advantage, public data are purportedly meant to represent the 'truth' of a system/domain and to be used for a public good. If cities are to become 'smarter' and 'better' through data, then the provenance of urban data must become more than metadata; it must inscribe the biases and limitations held within the data set not simply as technical measurements of equipment calibration and limitations, as some provenance systems already track, but also socially, culturally and economically in terms of the intentions and meanings inscribed into the data. It must not leap from the individual to data point, but rather trace the intentions, design and social biases within a data system.

## *Inscribing meaning in data*

As an aspect of a digital object, data provenance becomes part of larger data infrastructures. In urban settings, these infrastructures store, analyse, share and host data deemed pertinent to the function of society (such as public administration record-keeping and programs, planning scenarios and disaster response). While the use and analysis of the myriad of new emerging forms of urban big data purportedly makes such data infrastructures austerely smarter, in reality they function within an apparatus that is as much social and political as technical, privileging certain forms of knowledge over others. Drawing on Foucault (1980), Kitchin and Lauriault (in press: 9) define data infrastructures as assemblages that are 'never neutral, essential, objective; their data never raw but always cooked to some recipe by chefs embedded within institutions'. Such a perspective is absent from current technical approaches to data provenance, which overlook the inscription of meaning into data that occurs through their creation, handling, processing and sharing. This inscription occurs on two levels: first, it is well documented how biases, goals, and intentions shape data capture (Gitelman 2013; Boellstorff 2013; Dalton and Thatcher 2014; Kitchin and Lauriault in press). Both institutionally and individually, data are cooked by the desires and values of those who create them. There is also a second, less documented dimension: the inscription of meaning that occurs in the leap between data creator and data analyst.

With respect to the latter, this leap occurs when the intentionality of data creation is assumed or inferred within data analysis. To return to the Foursquare example, a check-in may occur because the end-user wants to alert their friends as to where they are, wants to receive a discount for the meal they intend to purchase, or because they are attempting to commit 'location fraud' for any of a variety of reasons (Carbunar and Potharaju 2012). Foursquare can only infer the reason, but to extract value from the data they must inscribe a meaning and then act on it as if it is true. Foursquare can only act upon the individual that data can see, that data created in the epistemological leap from someone's actual everyday actions to partial inscriptions thereof (Thatcher *et al.* 2016). The creation, capture and analysis of those inscriptions typically do not allow for, much less incorporate, many forms of difference including identity, intercultural silences and even digital divides (Dalton *et al.* 2016; Thatcher *et al.* 2016). However, just as Harris states about geodemographic targeting, 'Given the nature of business decisions, the cost . . . would not be borne if the technique could not prove its worth' (Harris *et al.* 2005: 225), the provenance of said data is ultimately immaterial to Foursquare so long as the data, in some way, produced a profit.

When said data are used to 'discover [a city's] structure' or model the 'everyday experiences of real people', their purported truth takes on added significance (*Livehoods* 2012, in Thatcher 2014). A common critique of the idealized smart city shows its 'entangling of neoliberal ideologies with technocratic governance and the dystopian potential for mass surveillance' (Shelton *et al.* 2015: 14) and, regardless of how smart cities play out on the ground, such critiques illustrate the heady mixture of private data infrastructures for purported public good (see Kitchin 2014b). Like socially targeted advertising, this is hardly a new phenomenon, as social scientists and urban planners have long used data from one context to find meaning in another. However, as these systems seek less to interpret the world than to actively produce it (Kitchin *et al.* 2015), it is critical to remember that the reduction of society to numbers only works insofar as that society can be remade in the image of said numbers (Porter 1995). In the context of ubiquitous, multi-stream sources of digital urban data, this inscription occurs without traditional safeguards. Big, urban data become reality with a provenance that traces back to the act of inscription, but no further. The intentionality of the creators is effaced. Strong provenance, therefore, must account not only for the biases and assumptions that shape a data object, but also the intentionality of those that created the object, a recursive process of negotiation between society and technology. In the concluding section, we highlight how both data provenance and the very meaning of data itself, its ontological status, exist differently for distinct types of data. We use these distinctions to suggest a move towards a new schema which situates data and their provenance as socio-technical issues.

## Conclusions and possibilities: becoming provenance through data-encounter

Companies, non-profit organizations and government agencies will continue to use digitally born data to interpret, analyse and shape urban environments. Therefore, data provenance is always a more-than-technical question. It is never enough to simply document data characteristics and transformations once they are digital, rather, the contextual limitations and influences of said data must also be acknowledged. For archivists, Millar (2002: 14) has suggested a move from *respect des fonds* to *respect de provenance* with the latter emphasizing three areas: ‘creator history, records history, and custodial history’. In some ways, this shift suggests a return to emphasis upon the rigorous construction and recording of metadata. However, despite the efforts of scholars like Schuurman and Leszczynski (2006), metadata still are ‘not at the core of definitions for data structures or interwoven with the operations we execute on data’ (Bergmann 2016: 4). Given that data provenance is always intrinsically linked to what can be known and what can be done by and through the data in question, digital urban data require a provenance that is *more than* metadata.

Building from concepts of situated knowledges (Haraway 1991), vibrant matter (Bennett 2010), theories of relational space (Massey 2005), and more, Bergmann (2016) recognizes the ‘more than’ nature of the composition and interpretation of spatial data. Doing so moves from the ‘geo-atom’ (Goodchild *et al.* 2007) towards reflexive, polysemic interpretation and ‘geo-encounter’ (Bergmann 2016). In its unbounded form, the geo-encounter attempts to represent relations between individuals, objects, and space that account for reflexivity of interpretation within a more-than human world (ibid.: 10). In such a formulation, spatial coordinates become just one vector amongst potentially many that create the contexts ‘relevant for the object . . . to have come to participate in this particular *geo-encounter*’ (ibid.). This approach emphasizes the contextual, ontogenetic production (Dodge and Kitchin 2007; Dodge *et al.* 2009) of data, rather than accepting they exist, part and parcel of a neutral, objective science instantiated in the gaze from nowhere (Haraway 1991).

Critical data studies invokes data as a socio-technical practice, as more than digital data themselves. What constitutes data for a marine biologist may be quite different from what is acceptable for an ethnographer of migrant workers. Thus far, we explored both the current approaches to data provenance as a technical problem and, through the conceptual limitations of everyday urban data, the need for a 'more than' approach to data provenance. Mattern describes how this may be achieved through increased engagement by librarians and archivists and associated professional mores (2016). We adopt a different, less professionally oriented approach, focusing on how the concept of the *geo-encounter* highlights how the ontological nature of data need not be fixed, opening an additional, new approach to provenance.

In Bergmann's theoretically agnostic formulation, the *geo-encounter* grows unchecked, with each new observation, interpretation or analysis of said data adding a new layer of complexity. For urban data, whether leveraged for a 'smart' city or other purposes, this presents a problem. On one hand, the situated, contextual, ontogenetic nature of data as they exist in the world adds meaning to how those data can be and are being used. On the other, an individual accessing their smartphone to check when her next bus will arrive does not need a data-object that includes launch dates of the GPS satellites that serve the bus's GPS receiver. We therefore propose a mesoscale data-encounter that recognizes both that all data need not be explicitly spatial and that sets minimum standards for a 'more than' technical data provenance for urban spatial data.

Table 6.1 presents what we consider the minimum necessary components for a data provenance given the existing limitations of urban data. Each new component of provenance is listed across from the technical question of provenance which it complements. In addition, there are two important and intentional caveats

*Table 6.1* The 'more than' requirements for a data-encounter model of urban data provenance

| Technical | 'More-than' technical |
|---|---|
| Time of creation, of collection, of derivation | Data structure creator and their purpose – profit generation, scientific analysis, state records, etc. |
| Method of creation, of collection, of derivation | Data user/producer's awareness of data |
| Databases in which data are stored | Data point geographical referent – individual, area, etc. |
| Database from which data were accessed | Data practitioner's intended purpose – profit generation, scientific analysis, state records, etc. |
| Database format of data | Other current and past uses of data |
| Transformations data have undergone | Assumptions in data creation and transformations |
| Current ownership, intellectual property restrictions | Lineage of ownership, with varying intellectual property restrictions along the way |
| Metadata standard | Context and purpose of that standard or the organization that manages it |

to the table. First, we are agnostic about how these new elements of data provenance are to be constituted within the data technically. This is done both because any attempt at formal procedure for an abstract urban system will inevitably be rendered useless by the progression of new technologies and data formats and, more significantly, because we recognize that there already exist multitudinous ways to achieve these goals. While we are sceptical of efficacy of simply creating stricture metadata protocols, such a formalization seems technically feasible at this juncture. We are more partial towards a reformatting of urban data within graph databases (Robinson *et al.* 2013), which allow for the data objects themselves to be ontologically constituted as diverse sets of objects rather than within prescribed tables. This latter point drives towards our second caveat, that this schema is meant as the beginning of a critical data studies approach to data provenance. Given the variability of data across space and time (Dalton *et al.* 2016), claiming an austere, objective schematic for urban data provenance would comically contradict the very situated, socio-technical, ontogenetic concept of data we are espousing. This meso-scale move towards data provenance as data encounter may begin a deeper conversation on the ontological nature of data, how it is used in urban systems, and what data provenance must become.

## Notes

1 For archivists, provenance is analogous to ideas of 'chain of control' for a data set; it helps create a provision of trust in the authenticity of an item. See www2.archivists.org/glossary/terms/p/provenance and www2.archivists.org/glossary/terms/a/authenticity for definitions.

2 In their survey of current data provenance approaches, Glavic and Dittrich (2007) refer to Tan's approaches as 'source provenance' and 'transformation provenance', respectively.

3 And even then their recommendations are often far from suitable.

## References

Batty, M. (2013) 'Big data, smart cities and city planning', *Dialogues in Human Geography* 3(3): 274–279.

Bennett, J. (2010) *Vibrant Matter: A Political Ecology of Things*. Durham, NC: Duke University Press.

Bergmann, L. (2016) 'Toward speculative data: Geographic information for situated knowledges, vibrant matter, and relational spaces', *Environment and Planning D: Society and Space* 34(6): 971–989.

Boellstorff, T. (2013) 'Making big data, in theory', *First Monday* 18(10). Available from: http://firstmonday.org/ojs/index.php/fm/article/view/4869/3750.

Buneman, P., Khanna, S. and Tan, W.C. (2001) 'Why and where: A characterization of data provenance', in J. Van den Bussche and V. Vianu (eds), *Proceedings of the 8th International Conference on Database Theory*. Berlin: Springer, pp. 316–330.

Carbunar, B. and Potharaju, R. (2012) 'You unlocked the Mt. Everest badge on Foursquare! Countering location fraud in geosocial networks', *Proceedings of the 9th IEEE International Conference on Mobile Adhoc and Sensor Systems (MASS)*, pp. 182–190.

Cheney, J., Chong, S., Foster, N., Seltzer, M. and Vansummeren, S. (2009) 'Provenance: A future history', *Proceedings of the 24th ACM SIGPLAN Conference Companion on Object Oriented Programming Systems Languages and Applications*, pp. 957–964.

Crawford, K. (2013) 'Think again: Big data', *Foreign Policy*. Available from: http://foreignpolicy.com/2013/05/10/think-again-big-data/ [accessed 10 February 2017].

Dalton, C. and Thatcher, J. (2014) 'What does a Critical Data Studies look like, and why do we care? Seven points for a critical approach to "big data"', *Society and Space*. Available from: http://societyandspace.org/2014/05/12/what-does-a-critical-data-studies-look-like-and-why-do-we-care-craig-dalton-and-jim-thatcher/ [accessed 10 February 2017].

Dalton, C., Thatcher, J. and Taylor, L. (2016) 'Critical data studies: A dialog on data and space', *Big Data and Society* 3(1). Available from: http://journals.sagepub.com/doi/pdf/10.1177/2053951716648346 [accessed 10 February 2017].

Department of Homeland Security (2009) 'A Roadmap for Cybersecurity Research'. Available from: www.dhs.gov/sites/default/files/publications/CSD-DHS-Cybersecurity-Roadmap.pdf [accessed 10 February 2017].

Dodge, M. and Kitchin, K. (2007) 'Rethinking maps', *Progress in Human Geography* 31: 1–14.

Dodge, M., Kitchin, K. and Zook, M. (2009) 'How does software make space? Exploring some geographical dimensions of pervasive computing and software studies', *Environment and Planning A* 41: 1283–1293.

Donaldson, D.R. and Fear, K. (2011) 'Provenance, end-user trust and reuse: an empirical investigation', *Proceedings of the Third USENIX Workshop on the Theory and Practice of Provenance*.

Foucault, M. (1980) *Power/Knowledge: Selected Interviews and Other Writings, 1972–1977*, ed. C. Gordon. New York: Pantheon Books.

Gitelman, L. (ed.) (2013) *'Raw Data' is an Oxymoron*. Cambridge, MA: MIT Press.

Glavic, B. and Dittrich, K.R. (2007) 'Data provenance: a categorization of existing approaches', *Datenbanksysteme in Business Technologie und Web (BTW '07)*, pp. 227–241.

Goldsmith, S. and Crawford, S. (2014) *The Responsive City: Engaging Communities Through Data-Smart Governance*. Chichester: Wiley.

Goodchild, M.F., Yuan, M. and Cova, T.J. (2007) 'Towards a general theory of geographic representation in GIS', *International Journal of Geographical Information Science* 21(3): 239–260.

Haraway, D. (1991) *Simians, Cyborgs, and Women*. New York: Routledge.

Harris, R., Sleight, P. and Webber, R. (2005) *Geodemographics, GIS, and Neighbourhood Targeting*. Hoboken, NJ: John Wiley.

INFOSEC Research Council (2005) 'Hard problem list'. Available from: www.nitrd.gov/cybersecurity/documents/IRC_Hard_Problem_List.pdf [accessed 10 February 2017].

Kitchin, R. (2014a) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: SAGE.

Kitchin, R. (2014b) 'The real-time city? Big data and smart urbanism', *Geojournal* 79(1): 1–14.

Kitchin, R. (2014c) 'Short presentation on the need for Critical Data Studies', *The Programmable City*. Available from: http://progcity.maynoothuniversity.ie/2014/04/short-presentation-on-the-need-for-critical-data-studies/ [accessed 10 February 2017].

Kitchin, R. and Dodge, M. (2007) 'Rethinking maps', *Progress in Human Geography* 31(3): 331–344.

Kitchin, R. and Dodge, M. (2011) *Code/Space: Software and Everyday Life*. Cambridge, MA: MIT Press.

Kitchin, R., Lauriault, T.P. and McArdle, G. (2015) 'Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards', *Regional Studies, Regional Science* 2(1): 6–28.

Kitchin, R. and Lauriault, T. (in press) 'Towards critical data studies: Charting and unpacking data assemblages and their work', in J. Thatcher, J. Eckert and A. Shears (eds), *The Geoweb and Big Data.* Lincoln: University of Nebraska Press. Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2474112.

Lauriault, T.P., Craig, B.L., Taylor, F.D.R. and Pulsifer, P.L. (2007) 'Today's data are part of tomorrow's research: Archival issues in the sciences', *Archivaria* 64: 123–179.

Madrigal, A.C. (2014) 'How Netflix reverse engineered Hollywood', *The Atlantic*. Available from: www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/ [accessed 10 February 2017].

Massey, D. (2005) *For Space*. London: SAGE.

Mattern, S. (2016) 'Public in/formation', *Places Journal*, November. Available from: https://placesjournal.org/article/public-information/#ref_17 [accessed 10 February 2017].

Millar, L. (2002) 'The death of the fonds and the resurrection of provenance: Archival context in space and time', *Archivaria* 53: 1–15.

Monmonier, M.S. (1995) *Drawing the Line: Tales of Maps and Cartocontroversy*. New York: Henry Holt.

Moreau, L., Groth, P., Miles, S., Vazquez-Salceda, J., Ibbotson, J., Jiang, S., Munroe, S., Rana, O., Schreiber, A., Tan, V. and Varga, L. (2008) 'The provenance of electronic data', *Communications of the ACM* 51(4): 52–58.

Onsrud, H.J. (1999) 'Liability in the use of geographic systems and geographic data sets', in D. Maguire, M. Goodchild, D. Rhind and P. Longley (eds) *Geographic Information Systems: Principles, Techniques, Management, and Applications*. Chichester: Wiley.

Phillips, J.L. (1999) 'Information liability: The possible chilling effect of tort claims against producers of geographic information systems data', *Florida State University Law Review* 26(3): 743–781.

Poole, A.H. (2015) 'How has your science data grown? Digital curation and the human factor: a critical literature review', *Archival Science* 15: 101–139.

Porter, T.M. (1995) *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.

Reichman, O.J., Jones, M.B. and Schildhauer, M.P. (2011) 'Challenges and opportunities of open data in ecology', *Science* 331: 703–705.

Robinson, I., Webber J. and Eifrem E. (2013) *Graph Databases*. Sebastopol, CA: O'Reilly Media.

Robinson, S.J. (2008) 'The doppelganger effect: Spaces, traces and databases and the multiples of cyberspace', doctoral dissertation, Carleton University Ottawa, Canada.

Roderick, L. (2014) 'Discipline and power in the digital age: The case of the US consumer data broker industry', *Critical Sociology* 40(5): 729–746.

Schuurman, N. and Leszczynski, A. (2006) 'Ontology-based metadata', *Transactions in GIS* 10(5): 709–726.

Shelton, T., Zook, M. and Wiig, A. (2015) 'The "actually existing smart city"', *Cambridge Journal of Regions, Economy and Society* 8: 13–25.

Smith, C. (2014) 'The digital doctor will see you now: how big data is saving lives', *TechRadar*. Available from: www.techradar.com/us/news/computing/the-digital-doctor-will-see-you-now-how-big-data-is-saving-lives-1253870 [accessed 10 February 2017].

Tan, W.C. (2004) 'Research problems in data provenance', *IEEE Data Engineering Bulletin* 27(4): 42–52.

Taylor, L. (2015) 'Towards a contextual and inclusive data studies: A response to Dalton and Thatcher', *Society and Space*. Available from: http://societyandspace.com/2015/05/12/linnet-taylor-towards-a-contextual-and-inclusive-data-studies-a-response-to-dalton-and-thatcher/ [accessed 10 February 2017].

Thatcher, J. (2014) 'Living on fumes: Digital footprints, data fumes, and the limitations of spatial big data', *International Journal of Communications* 8: 1765–1783.

Thatcher, J. (2017) 'Locative and sousveillant media', in R. Kitchin, T.P. Lauriault and M.W. Wilson (eds), *Understanding Spatial Media*. London: SAGE. pp. 56–65.

Thatcher, J., O'Sullivan, D. and Mahmoudi, D. (2016) 'Data colonialism through accumulation by dispossession: New metaphors for daily data', *Environment and Planning D*. doi:10.1177/0263775816633195.

Townsend, A.M. (2013) *Smart Cities*. New York: W.W. Norton.

Wolcott, J. (2013) 'Wired up! Ready to go!' *Vanity Fair*. Available from: www.vanityfair.com/culture/2013/02/quantified-self-hive-mind-weight-watchers [accessed 10 February 2017].

Zetter, K. (2008) 'Six-year-old news story causes united airlines stock to plummet – Update Google placed wrong date on story', *Wired*. Available from: www.wired.com/2008/09/six-year-old-st/ [accessed 10 February 2017].

Zittrain, J. (2012) 'Meme patrol: "When something online is free, you're not the customer, you're the product"', *The Future of the Internet and How to Stop It*. Available from: http://blogs.harvard.edu/futureoftheinternet/2012/03/21/meme-patrol-when-something-online-is-free-youre-not-the-customer-youre-the-product/ [accessed 10 February 2017].

# 7 Following data threads

## *James Merricks White*

## Introduction

In the introduction to their edited collection *Standards and their Stories*, Star and Lampland (2009: 11) suggest that one reason why standards have been neglected as a topic of study is that they are generally perceived as boring. Formal standards, classification schemes and practices of quantification do indeed seem to fade from view to become part of the invisibilities of infrastructure (Larkin 2013). If they are seen at all, it is as technical, detail-oriented and pragmatic – the very characteristics that confer on them a sense of trust. Like Star and Lampland, I too believe that these boring things are an overlooked but crucial component of everyday governance and economic life, and are worthy of examination.

Following Bates *et al.* (2016), I propose to follow data from their generation through to their various uses, exposing how they are cleaned, recombined and put to work. Rather than emphasize the formal development of data standards or the difficulties faced in their implementation, I attempt to cut across these geographies by exploring data in translation and circulation. Issues of provenance, storage, manipulation, standardization and licensing are all open to such an analysis, but only inasmuch as they relate to the particular data being observed. The approach thus seeks to make an epistemological and methodological contribution to studies of data by thickening the description of its assemblages (Kitchin 2014; Kitchin and Lauriault 2014). While Bates *et al.* (2016) propose a methodological commitment to uncovering 'data journeys', I reimagine the approach as a following of 'data threads', highlighting the entanglement of data infrastructures and geography, and their inherent relationality.

In the first section, I illustrate the data threads approach through a story of infant mortality statistics. Over a two-week period in early 2016, I conducted a set of interviews with public officials in Toronto, Canada, concerning the production and perceived value of indicators for international urban benchmarking. Drawing on these interviews as well as primary and secondary sources on health data in Canada, I follow how the tragic event of a loss of life is recorded, aggregated and used as an input to derivative calculations. Next, I compare data threads to data journeys (Bates *et al.* 2016) in terms of materiality and spatiality. While the metaphors are similar in many respects, they differ in their disposition to matter and meaning, and therefore diverge epistemologically. I conclude by reflecting briefly on the ethical and political work that data threads perform in revealing the invisibilities of infrastructure.

Before turning to the example, it is important to clarify some of the conceptual framing deployed in the chapter. Data can be thought of as a *boundary object* (Star and Griesemer 1989). Boundary objects are things that have different meanings to different communities of practice. They might sit comfortably between these different worlds or they might become a site of contestation. Far from being neutral and objective, data are always imbued with the assumptions and politics of their generators and calculators – data are never raw, they are always cooked (Bowker 2005). As they are manipulated, the meanings of data change to reflect new discourses and ideologies. Actor network theory, especially during its first decade, referred to these relational shifts as *translations* (Callon 1986). Actors enrol others to their particular habits, needs and desires through different mechanisms of translation. For my purposes, the concept offers a useful way to avoid scaled spatial imaginaries of city, country and globe (Marston 2000). Rather than attempt to identify obligatory passage points (actors occupying key mediating positions between two or more networks) or immutable mobiles (actor-networks that are relationally stabilized so that they can be spatially circulated), I use translation more loosely to refer to the interfolding of associations. Data threads are drawn into their various uses through processes of translation rather than scalar transition. Importantly, I want to resist the urge to conceive of this as a flat geography whereupon all spatial relations are afforded equal ontological status. Following Tsing (2015), I use the term *alienation* to stress the significance of political and economic associations in conditioning the possibility of data assemblage. The term is used playfully and provocatively, but with the serious intent of foregrounding the social and political lives of data. As data are translated and put to work in various settings, I suggest that they risk becoming further disentangled from the phenomena they are intended to represent. While this might be productive of financial and managerial knowledge practices, it also risks overriding local contingency and expertise.

## The data threads of infant mortality in Toronto, Canada

Whenever a live-born baby dies in Canada the circumstances of that death are carefully recorded according to the tenth revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10). This list of 14,400 symptoms, injuries and ailments attempts to organize social and physiological conditions to account the loss of life (Moriyama *et al.* 2011). Its purpose is to systematize and harmonize the generation of hospital data globally. The first edition of the ICD was developed in the early 1890s by French physician, Jacques Bertillon. It went through various revisions under the aegis of the International Statistical Institute before becoming the responsibility of the World Health Organization (WHO) upon its founding in 1948. The tenth edition was endorsed by the WHO in 1990, was published in 1992 (WHO 1992), and came into use in Canada ten years later.

The ICD is an idiosyncratic cultural artefact that carries the weight of its history in its more than 2,000 pages. While the ICD is intended to be a standardized and universal coding system, applicable in the instance of any death in any country around the world, it is biased towards the classification of life and death in the developed world, detailing a wide variety of ways of dying in societies with diverse material-consumer lifestyles (Bowker and Star 1999: 76). But even bodies in the global North do not always fit so easily into the ICD. For example, consider that it is not uncommon that the elderly in the developed world die due to a complex of disease symptoms. Precisely identifying the correct code to be used in such an instance is a difficult task (Rosenberg 1999). Furthermore, the coding practices that enact ICD-10 are contingent upon all sorts of social particularities (Bowker and Star 1999; Malley *et al.* 2005). Different medical institutions in different locations favour slightly different interpretations of disease symptoms (Timmermans and Berg 2003). Doctors are educated within their local and national medical traditions. Their trained gaze sees certain things that the untrained eye cannot, but also expects to see certain things and so overlooks the unexpected. Doctors are known to be hasty in diagnosing particular causes of death, preferring to spend their time and energy on the living (Bowker and Star 1999: 65). Further complicating matters, the medical record clerks and coders enact their own translations in turning medical records into an ICD code (Malley *et al.* 2005). Standards and databases of the causes of death are far from the neutral, scientific and objective account that they are often taken to be.

As these things go however, the coding of infant deaths in Canada is fairly straightforward. Child morbidity and mortality follow a recognizable pattern in the developed world (You *et al.* 2015). Certain congenital defects and genetic anomalies, present in a minority of births, invariably account for the majority of deaths occurring within the first year of life (Wen *et al.* 2000). Often, babies are only born with these conditions due to medical intervention. For this reason, there is a spike in the rate of mortalities within the first 24 hours of life, as some infants, sadly, are not strong enough to survive for longer than this. After 28 days, the rate of mortality plateaus and remains low over subsequent months. Between the first and fifth years of life, congenital anomalies can still be a cause of death, however other factors such as sudden infant death syndrome, violence, accident and disease also contribute to the total number (Ananth *et al.* 2009).

In addition to its ICD classification, statistics for infant mortality are further conditioned by the reporting practices in Canadian hospitals. Civic registrations of births, deaths and marriages (known as vital registrations) are collected by Ontario's Office of the Registrar General to produce official birth and death records. Statistics Canada (2015) pools data from each of the provinces and territories to compile a national death database. This includes fields for: age, sex, marital status, place of birth, place of residence, date of death, ICD codes for the cause of death and the province or territory in which the death occurred. Verification tables are used to check for instances of unexpected distribution, large absolute or relative changes in the data, outliers and the percentage and number of unknown data points (Statistics Canada 2015).

As we move into different communities of practice, the meanings and values attached to infant mortality change. For the WHO, accurate death statistics are part of a strategy to regularize and improve healthcare globally (Brown *et al.* 2006; Adams 2016). For Statistics Canada and the Office of the Registrar General in Ontario, the mechanics of compiling and processing data take centre-stage as part of the mandate of those government institutions. It is not necessarily the case however that those who compile administrative data are aware of how data generated from vital registrations affect decision-making. For example, in 1991, the Registrar General's data entry office was relocated from Toronto to Thunder Bay. As a result a lot of tacit knowledge was lost and coding errors were introduced into the data sets. A Toronto-based epidemiologist I spoke with described the effects of this change:

> [W]e had a problem with birth rate where they were rounding everything to the half-pound and then reporting things out in grams. Again it was like, this is looking strange, this is looking quite weird. [We] approached them again [and they replied] 'Oh we didn't know anybody used these data, we thought it was just for registering the death or the birth.' . . . The Ministry of Finance relies on the birth and death data to help figure out population projections. Businesses rely on these kinds of things once they get morphed out. Like, this is heavily used data which they sort of continually didn't recognize.

Vital statistics are taken up by local institutions in a range of different and meaningful ways. It is for this reason that I refer to data threads as boundary objects. In a general, colloquial sense, data are detached from their intended applications. They are 'both ambiguous and clear, at different moments, for different purposes' (Star 2002: 118). As such, it is important to pay attention to the translations which data undergo as they are put to work. Having discussed the collection and processing that originate the data threads of infant mortality, I now turn to the manipulations that ground their situated meanings.

According to an interviewee, the Surveillance and Epidemiology Unit of Toronto Public Health are less concerned with the rates of infant mortality than they are with analysis of low birth weights. Weights can be more closely correlated with differences in material living conditions within the city and so paint a far clearer picture of the social causes of disease and malnutrition. Public health is not simply a reflection of the quality of an area's formal healthcare system, but also must consider questions of who gets sick and why (Raphael 2016). Had Toronto's Public Health epidemiologists been more interested in analysing infant mortality rates, they could very easily have used the data produced in the coding and reporting of mortality to derive useful metrics. This might entail the separation and recombination of data representing deaths within the first day and within the first 28 days of life. They would not be inclined to aggregate all deaths occurring in the first five years of life. Nevertheless, this very figure is calculated by Toronto's Surveillance and Epidemiology Unit on a yearly basis and reported to the city's central administrative body.

Over the past five years Toronto's City Manager's office has collected data in accordance with a set of performance indicators developed by the Global City

Indicators Facility (GCIF) at the University of Toronto. To achieve this, they have enrolled the time, knowledge, expertise and database access of a range of departments within the city's administration. The employees of Toronto's City Manager's Office I spoke with describe participation in this exercise as a useful way in which to visualize and benchmark their performance across a number of governance areas. They hope, with time, to improve their metrics and to learn from good practices in other cities around the world.

This suite of performance indicators emerged from a World Bank funded research project to promote urban metrics in the developing world and thereby address their considerable global variation (Hoornweg and Blaha 2006). After comparing official statistics from 255 cities, researchers at the GCIF identified only two data points in common across the data sets. Where cities measured their performance, they typically did so with indicators attuned to ongoing assessment against local policy and planning documents. The GCIF conducted a survey of 1,015 indicators and discussed 53 in detail in their final report to the World Bank (The Global City Indicators Program 2008). In the years following the project these were adapted and extended, first in consultation with nine North American city administrations and then at the International Organization of Standardization (ISO). In 2014, a set of 100 indicators was published as international standard ISO 37120, 'Indicators for city services and quality of life' (ISO 2014). While formalizing the indicators with the ISO lends them credibility, this path is not without its drawbacks. International standardization stabilizes the measures, rendering them fixed for the near-term future. More adaptive and locally specific indicators might be informed by ISO 37120 but the standard does not replace them.

The formal documentation of ISO 37120 includes definitions for the calculation of its metrics and suggestions as to how city administrations might go about finding the data to meet them. In the case of the indicator compiled by the Surveillance and Epidemiology Unit, 'Under age five mortality per 1,000 live births', the standard adopts an indicator from the UN Millennium Development Goals (ISO 2014: 31). At the 2000 Millennial Summit, United Nations member states committed to eight goals for global health and well-being, the fourth of which was to decrease the under-five mortality rate. Progress on this goal is tracked using figures released annually by the UN Interagency Group for Child Mortality Estimation (UN IGME), whose calculations are based on vital registrations when they are available and survey data when they are not (UN IGME 2015). Many of the difficulties they have faced in this work are anticipated by ISO 37120 (ISO 2014: 31):

> In developing countries, household surveys are essential to the calculation of this indicator, but there are some limits to their quality. Survey data are subject to recall error, and surveys estimating under age five deaths require large samples, because such incidences are uncommon and representative households cannot ordinarily be identified by the sampling. Moreover, the frequency of the survey is generally only every three to five years. When using household surveys the user shall take sampling errors into account. Also, indirect estimates rely on estimated actuarial ('life') tables that may be inappropriate for the population concerned.

Two points are important here. The first relates to the use of life tables in the calculation: 'The under age five mortality rate, is strictly speaking, not a rate . . . but a probability of death derived from a life table and expressed as a rate per 1,000 live births' (ISO 2014: 31). A life table is a list of the probability of dying at each year of life, either at a moment in time or for a given population cohort (Statistics Canada 2016). Where most of the ISO 37120 indicators are found using a numerator and a denominator, in this case the figure is a sum of percentages multiplied by 1,000. While the final number ought not to be significantly different as a result – life tables are also derived from vital registrations and survey data – the use of life tables adds an additional layer of abstraction. The second point concerns the way in which the standard draws external expertise into its enactment. ISO 37120 bolsters its legitimacy by citing the Millennium Development Goals. The fact that the UN IGME also publishes numbers for neo-natal and infant mortality rates is not mentioned.

The ISO 37120 indicators are bound together in two ways. They are grouped into 17 themes intended to capture the principle responsibilities of a city's administration. These include areas such as education, finance, health, solid waste, transport, urban planning and wastewater. They are also separated into core indicators, which every city is expected to be able to report on, and supplementary indicators, which they may not presently be able to. 'Under age five mortality per 1,000 live births' is one of four core health indicators. The other three measure a city's 'Average life expectancy', its 'Number of in-patient hospital beds per 100,000 population' and its 'Number of physicians per 100,000 population'.

The World Council on City Data (WCCD), a non-profit set up by the GCIF, is the principal developer and sole certifier of ISO 37120. They work closely with cities that wish to become certified adherents to the standard. For this service they charge $7,500 for the first year and $5,000 for each year thereafter. One further step removed from the WCCD is their third-party auditor. Their task is to verify the data submitted by each of the cities, ironing out contingencies in the metadata and ensuring that the figures are of a sufficient regularity. Cities are awarded different levels of certification (aspirational, bronze, silver, gold and platinum) depending on their ability to meet the auditor's demands on all of the core indicators and a certain number of the supplementary indicators. Through the public use of the awards and the backgrounded work of the auditors, the WCCD encourage cities to constantly improve the quality of their submitted data (White 2016). While the purpose of this is presently to benchmark cities, in the future the WCCD expect the data to improve calculations of a city's insurance premiums and credit ratings. ISO 37120 clearly has the potential to deepen the financialization and globalization of cities and their governing bodies.

Three conclusions can be drawn from my analysis. First, to stabilize 100 measurements in the form of an international standard is to foreclose the ongoing feedback of local domain-area experts. The WCCD do not intend to address the misgivings of Toronto's public health epidemiologists. Their goal is to improve global transversal analysis (i.e. the comparison of Toronto to Makati, Makkah and Melbourne) rather than longitudinal analysis within a city. More nuanced and responsive metrics remain the prerogative of local or regional authorities. Second, in their efforts to make global city data commensurable, the WCCD prioritize certain urban geographies over others; ISO 37120 internalizes compromises made in the selection of what to measure and how. This is not unproblematic. While under-five mortality rates may be a useful measure in developing cities, in Toronto the data point is not particularly revealing. Calculations for neo-natal and infant mortality rates do exist for national-level comparisons made by the UN, however a choice has been made not to include these amongst the ISO 37120 indicators. This collapses epidemiological nuance into a single figure deemed fit for the purpose of international urban comparison. Finally, as we move along the thread, data become increasingly abstracted and alienated from the initial incident they represent. The devastating loss of a child's life is rendered first as trace, then as data point and then as input to derivative calculations and distant ambitions. Certainly, such data can be used to inform and improve public health services when conducted in a sensitive and domain-specific manner. However, to award certification on the ability to meet backgrounded auditing norms overshadows more embedded ways of knowing and caring for others (Adams 2016).

## Revealing data threads

Having shown how the analytical approach of data threads can be used to describe data standards and infrastructure, I want to shift now into a more theoretical frame, exploring the implications of the metaphor of the 'data thread' in terms of materiality and spatiality.

Many working within what has recently been called critical data studies (Dalton and Thatcher 2014) have asserted that data must be thought of as material or more-than-material (Dourish and Mazmanian 2011; Wilson 2011; Bates *et al.* 2016). For example, Dourish and Mazmanian (2011) argue that it is important to push beyond the mere physical fact of data and to encounter them in their use. By focusing on data practices they seek to bring the 'the historical particularities, cultural specificities, and political consequences' (Dourish and Mazmanian 2011: 4) of data in the world to the fore. Drawing on the conceptual framework developed by Haraway (1997), Wilson (2011) unpacks the material-semiotics of data and data production. Rather than detail data practices *per se*, his work interrogates how meaning comes to be associated with specific data within local communities. Somewhat at odds to these approaches however, Bates *et al.* (2016) push back against the blurring of matter and meaning. For them there remains a useful analytical distinction to be made between the material and the socio-cultural. Their attention is focused on the ways in which 'socio-material structures . . . [are] historically constituted through the actions of both historic and present-day human actors' (p. 3). Ideas and values are important, but they are dialectically distinct from matter. The perceiving and conceiving human is held apart.

Before positioning data threads within these debates, it is worth contextualizing the assertion that data are material. For what else could they be? Important here is genealogical work on the development of cybernetics and information theory. Through a close reading of post-war cyberneticians such as Norbert Weiner and Claude Shannon, Hayles (1999) shows how information came to be understood as a pattern separate from a material form. This strategy of disembodiment was important in allowing information theory to migrate between academic disciplines such that it could equally be applied to biology or cognitive psychology as to communications engineering. Implicit in this separation between matter and meaning, however, is a latent mind/body dualism. Cybernetics, at least in its early years, had an uneasy relationship with the tenants of liberal humanism. On the one hand, humans, animals and machines were all thought of as systems capable of being steered through manipulation of their inputs. On the other, early cybernetics rendered the mind as information separate from the body. In response to this conundrum, Hayles urges her readers to reconsider information at the intersection of dialectics of pattern/randomness and presence/absence. If this is to inform social scientific research, information, data and meaning all need to be thought through their various and complex embodiments. Hayles work is less interested in the histories of socio-material structures than the ways in which academic and fictional representations of information have their root within contingent and situated knowledges.

Following Hayles and others (Barad 2007; Orlikowski and Scott 2015), I have attempted to treat the data threads of infant mortality as both material and discursive by pressing genealogical analysis beyond its anthropocentric and representational biases. The point is not only to give data back their body but also to explore the richness of their more-than-material entanglements and the mechanisms by which they are disembodied and become alienated. Rather than focus my attention principally on data practices or the material histories of infrastructure then, I seek to move between the materializations of data and the apparatuses of signification, categorization and measurement that condition their possibility. Put differently, I refuse the *a priori* distinction between the material and the sociocultural as it is asserted by Bates *et al.* (2016). Matter and meaning are always already entangled; translation draws attention to how these entanglements unfold and evolve. By moving back-and-forth between the material and the discursive, my story attempts to disclose their co-constitutive diffractions (Barad 2007). In practice, this is most similar to material-semiotics (Haraway 1997; Wilson 2011), but with a heightened appreciation of the productive capacity of ideas and their historical context.

By situating the constitutive moment within relations, rather than in their perception, data threads also have implications for how we think about spatiality. Consider two spatial imaginaries. In the first, a thread might be seen to have a single dimension stretching from one end to the other. It can easily bend, loop, turn back on itself and be tied in knots. In the second, it might be observed that a thread occupies space. It has a length, a thickness and has starting and finishing positions. Its shape can be described in three-dimensional space and its movement understood through a dimension of time. While this second spatial imaginary lends itself to an absolute, Newtonian reckoning of spacetime, many relational analytics remain commensurate with this scaffolding.

The 'data journey' is a relational metaphor that nevertheless relies on this second imaginary. Journeys are sequential and like threads can be understood as linear, circuitous, with a beginning and an end. By observing how data are made and used in practice, the analytic aim is to 'situate data across interconnected sites of practice *distributed through time and space*, drawing attention to the movement of data between these sites' (Bates *et al.* 2016: 2; emphasis added). The metaphor exposes the mediation between differing social worlds. As such, it is in keeping with the concept of boundary objects. Further, the data journey is well attuned to the breaks, stoppages and disjunctions captured by Edwards notion of 'friction' (2010) – something that the 'data thread' fails to adequately capture. In foregrounding movement and mobility however, it inevitably requires a fixed frame of reference.

> Whilst academic research in this field has tended to refer to the 'flow' of data within a given context, the term 'flow' tends to suggest a disconnect of data from physical sites of data practice. The concept of a data journey aims to better locate data in physical space; places which should not be imagined as 'self-contained' units, but as sites constituted in part by social relations external to their particular locale.
>
> (Bates *et al.* 2016: 4)

The spatiality of the data journey is one in which data moves from location to location, intersecting with the complex relations that constitute place.

In incorporating this spatial imaginary into a consideration of the materiality data however I am left wondering what actually moves. At a technical level, data are only ever translated, manipulated and visualized with a computer's interface. Materially, it is the computer's memory hardware that stores data as digital representations; as voltage on a magnetic strip. These can be physically moved, as when a USB stick travels within a business person's suitcase, but it is much more common that data are sent between locations using network transmission protocols such as TCP/IP. Here, data are broken down into packets and passed along copper and fibre optic cables in an undetermined manner. But even 'passed' and 'sent' are metaphorical. What actually occurs, is that the presence or absence of charge is reinterpreted as the ones and zeros constituting the signal. Electrons do quite literally flow along cables, however I do not think this is what is meant by the mobility of data. More accurately, copies of data can be found fragmented and dispersed at any number of intermediate switches, routers and servers between the origin and the destination of a data transmission. The 'data journey' does not really capture this – even less so the 'data thread'. What these metaphors do convey is the more abstracted and relational movement of data; the way in which they are incorporated into material-discursive practices which take place. The fractious flow of the materialities of data suggests that the Newtonian imaginary may not adequately capture the spatiality of data.

In keeping with my effort to decentre the human observer, I want to develop what the first spatial imaginary outlined above might mean for data threads. What is important here is not the unbroken linearity of data nor the spaces which precede and envelope them. Rather, what I want to highlight are the spacetimes that are enacted by data practices, or more accurately, with which they are constituted. Barad (2007) offers a metaphysics in which the material and the discursive are formed through unfolding spacetimematter manifolds. She uses a metaphor of dough to help explain this:

> Imagine putting drops of colored dyes into a piece of bread dough. As you knead the dough, the dyes spread out in different patterns of entangled lines and surfaces. . . . Take a different kind of dough and make a different manifold with different lines, surfaces, and volumes of color. Intermingle the dough pieces: new entanglements form, new possibilities emerge.
>
> (p. 439)

The spread of the dyes through the dough, used to visualize the entanglements of matter and meaning within a phenomenal enactment, is consistent with my use of data threads. As we follow data through their permutations and calculations they are caught up in all sorts of strange and unanticipated matterings. These are produced along with spacetimes that reiterate and reconfigure the world – in my example infant mortality data in the city of Toronto. This is an inversion of Newtonian spatiality. Everything of concern is moving in and through the manifold but this is principally a relational space. Movement observed in the frame of absolute space (i.e. any journey between pre-existing and apparently stable locations) is ancillary to this.

The metaphoric use of data threads begs extension. We might imagine data knots, frays, breaks or loosenings in line with Edwards (2010). Or perhaps the warps, wefts and weaves outlined by Ingold (2007) might prove useful to other studies of data. Threads are a fundamentally relational conceptualization, not only in their entanglement of matter and meaning, nor in their fully relational spatiality, but also in the communities of association and dislocation they usher forth.

## Conclusion

In this chapter I have introduced the concept of data threads, used them to follow the material and discursive construction of infant mortality data in Toronto, and explored theoretical positions that underlie them. I hope that this has proven interesting (or at least, not boring) and might prove useful to other researchers. In making a commitment to the diffractive reading of matter and meaning, hopefully the notion of data threads opens up, rather than forecloses or obscures, thick descriptions of data.

In developing the concept of data threads I have drawn inspiration from actor-network theory, feminist technoscience, and the ways in which these literatures have been adopted by geographers. By closely observing the everyday work of scientists, technicians and engineers, these approaches move between naturalist and relativist accounts of science, describing scientific knowledge as constructed in practice. They take seriously the agency of nonhuman actors and attempt to break down the spatial structures of scale that are too easily assumed by more structuralist enquiry. Their weakness is that as they decentre and move away from human worlds they lose the capacity to argue convincingly in that register. The ethics that emerge are far less familiar and grandiose.

Data threads promise a way to draw attention to the bracketing-off of the materiality and material practices that bring about data, as well as the discursive and ideological regimes that allow them to take form. By foregrounding stories of inscription, translation and alienation in all their unusual geometries, the term allows one to denaturalize and perhaps even repoliticize the making of data. From a methodological perspective, the idea loosens the weave of managerial cultures, affording researchers the ability to: explore issues of data provenance, licensing and veracity; follow data threads through their various negotiations and dislocations; and expose the choices which are involved in the bundling and weaving of threads into a cohesive and strategic fabric of managerial governance. There is a politics in such a revealing of boring things.

## Acknowledgements

The research for this chapter was provided by a European Research Council Advanced Investigator Award, 'The Programmable City' (ERC-2012-AdG-323636). This chapter could not have been written without the help I received from my interlocutors in Toronto, Canada; my mother, Margaret White, who began her career as a medical records administrator and ended it a health information manager; and my colleagues Tracey Lauriault and Rob Kitchin.

## References

Adams, V. (ed.) (2016) *Metrics: What Counts in Global Health?* Durham, NC and London: Duke University Press.

Ananth, C.V., Shiliang, L., Joseph, K.S. and Kramer, M.S. (2009) 'A comparison of foetal and infant mortality in the United States and Canada', *International Journal of Epidemiology* 38(2): 480–489.

Barad, K. (2007) *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham, NC: Duke University Press.

Bates, J., Yu-Wei, L. and Goodale, P. (2016) 'Data journeys: Capturing the socio-material constitution of data objects and flows', *Big Data and Society* 3(2): 1–12.

Bowker, G.C. (2005) *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.

Bowker, G.C. and Star, S.L (1999) *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA and London: MIT Press.

Brown, M., Cueto, M. and Fee, E. (2006) 'The World Health Organization and the transition from international to global public health', *American Journal of Public Health* 96(1): 62–72.

Callon, M. (1986) ‘Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay’, in J. Law (ed.), *Power, Action and Belief: A New Sociology of Knowledge?* London: Routledge, pp. 196–223.

Dalton, C. and Thatcher, J. (2014) ‘What does a critical data studies look like, and why do we care? Seven points for a critical approach to “big data”’, *Society and Space*. Available from: https://societyandspace.com/material/commentaries/craig-dalton-and-jim-thatcher-what-does-a-critical-data-studies-look-like-and-why-do-we-care-seven-points-for-a-critical-approach-to-big-data [accessed 17 August 2016].

Dourish, P. and Mazmanian, M. (2011) ‘Media as material: Information representations as material foundations for organizational practice’, paper presented at the Third International Symposium on Process Organizational Studies, Corfu, Greece, June.

Edwards, P.N. (2010) *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA and London: MIT Press.

The Global City Indicators Program (2008) *Global City Indicators Program Report: Part of a Program to Assist Cities in Developing an Integrated Approach for Measuring City Performance*. Report submitted to The World Bank, 1 April.

Haraway, D.J. (1997) *Modest_Witness@Second_Millennium.FemaleMan©_Meets_OncoMouse™*. New York and London: Routledge.

Hayles, N.K. (1999) *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature and Informatics*. Chicago, IL and London: University of Chicago Press.

Hoornweg, D. and Blaha, D. (eds) (2006) *The Current Status of City Indicators: Part of a Study to Assist Cities in Developing an Integrated Approach for Measuring and Monitoring City Performance*. Report submitted to The World Bank, 8 December.

Ingold, T. (2007) *Lines: A Brief History*. London and New York: Routledge.

International Organization of Standardization (ISO) (2014) *ISO 37120: Sustainable Development of Communities – Indicators for City Services and Quality of Life*. Geneva: ISO copyright office.

Kitchin, R. (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: SAGE.

Kitchin, R. and Lauriault, T. (2014) ‘Towards critical data studies: Charting and unpacking data assemblages and their work’, The Programmable City Working Paper Series No. 2, Maynooth University.

Larkin, B. (2013) ‘The Politics and Poetics of Infrastructure’, *Annual Review of Anthropology* 42: 328–343.

Malley, K., Karon, J.O., Cook, F., Price, M.D., Wildes, K.R., Hurdle, J.F. and Ashton, J.A. (2005) ‘Measuring diagnoses: ICD code accuracy’, *Health Services Research* 40(5): 1620–1639.

Marston, S.A. (2000) ‘The social construction of scale’, *Progress in Human Geography* 24(2): 219–242.

Moriyama, I.M., Loy, R.M. and Robb-Smith, A.H.T. (2011) *History of the Statistical Classification of Diseases and Causes of Death*, edited and updated by H.M. Rosenberg and D.L. Hoyert. Hyattsville, MD: National Center for Health Statistics.

Orlikowski, W.J. and Scott, S.V. (2015) ‘Exploring material-discursive practices’, *Journal of Management Studies* 52(5): 697–705.

Raphael, D. (ed.) (2016) *Social Determinants of Health*, 3rd edn. Toronto: Canadian Perspectives.

Rosenberg, H.M. (1999) ‘Cause of death as a contemporary problem’, *Journal of History of Medicine* 54(2): 133–153.

Star, S.L. (2002) ‘Infrastructure and ethnographic practice: Working on the fringes’, *Scandinavian Journal of Information Systems* 14(2): 107–122.

Star, S.L. and Griesemer, J.R. (1989) ‘Institutional ecology, “translations” and boundary objects: Amateurs and professionals in Berkeley’s Museum of Vertebrate Zoology, 1907–39’, *Social Studies of Science* 19(3): 387–420.

Star, S.L. and Lampland, M. (2009) ‘Reckoning with standards’, in M. Lampland and S.L. Star (eds), *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*. Ithaca, NY and London: Cornell University Press, pp. 3–24.

Statistics Canada (2015) *Vital Statistics: Death Database*. Available from: www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey%7B%5C&%7DSDDS=3233 [accessed 16 August 2016].

Statistics Canada (2016) *Life Tables, Canada, Provinces and Territories: 2010 to 2012*. Available from: www.statcan.gc.ca/pub/84-537-x/84-537-x2016006-eng.htm [accessed 3 February 2017].

Timmermans, S. and Berg, M. (2003) *The Gold Standard: The Challenge of Evidence-Based Medicine and Standardization in Health Care*. Philadelphia, PA: Temple University Press.

Tsing, A.L. (2015) *The Mushroom at the End of the World: On the Possibility of Life in Capitalist Ruins*. Princeton, NJ and Oxford: Princeton University Press.

UN IGME (2015) *Mortality Estimation Explanatory Notes*. Report published by the United Nations Inter-Agency Group for Child Mortality Estimation.

Wen, S.W., Liu, S., Joseph, K.S., Rouleau, J. and Allen, A. (2000) ‘Regional patterns of infant mortality caused by major congenital anomalies’, *Teratology* 61(5): 342–346.

White, J.M. (2016) ‘ISO 37120, a medium for economies of translation’, paper presented at 4S/EASST Conference, Barcelona, Spain, September.

Wilson, M.W. (2011) ‘Data matter(s): Legitimacy, coding, and qualifications-of-life’, *Environment and Planning D: Society and Space* 29(5): 857–872.

World Health Organization (WHO) (1992) *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. Geneva: World Health Organization.

You, D., Hug, L., Ejdemyr, S. and Beise, J. (2015) *Levels and Trends in Child Mortality*. New York: United Nations Children’s Fund.

# 8 Sticky data

## Context and friction in the use of urban data proxies

*Dietmar Offenhuber*

## Introduction

What makes data meaningful? Is meaning hidden in the values and attributes of a data set or in the circumstances in which the data were collected and generated? In his general definition of information, philosopher Luciano Floridi (2011) describes information as data plus meaning, referring to a single datum as a lack of uniformity in the broadest sense. This definition implies that a data set by itself is not necessarily meaningful, since well-formed data can be generated from random numbers devoid of meaning. Only the imperfections of a random generator lead to biases and artefacts that can become meaningful when forensically analysed for breaking codes or identifying machines.

How does meaning emerge from data? A reasonable assumption would be that meaning is grounded in the numbers and symbols in a data set and the extent to which they serve as a useful representation of a phenomenon in the world. However, many scholars in the humanities will also point to the circumstances and conditions in which data were collected and which are usually not completely represented in data and metadata records. They remind us that data are already interpreted expressions (Drucker 2011), always already cooked, never raw (Bowker 2005: 183; Gitelman 2013). If data are understood as systematic observations that have been symbolically encoded and stored in some material form, it becomes clear that assembling a data set involves many human decisions. These decisions concern the aspects of a phenomenon to be observed, the method of observation, the rubrics and classification systems to encode the observations, and finally, how these encodings are to be stored or transmitted in a physical medium. Some of these decisions might have been provisional, and their justifications might have been forgotten once the collection mechanism is in place. When the circumstances of data generation are lost and we only have a data set without contextual information, we might end up with little in our hands.

Cities and public institutions generally keep accurate documentation specifying the circumstances under which they collect, encode and store their data records. But this is not always the case when data are the by-product of automatic mechanisms, or collected by private companies, who do not disclose their data sources and methods. The value of data is also diminished when data and metadata are separated, as we are beginning to see in open data portals, where data sets are often presented without explanation of their institutional context.

Analysts, administrators and the public might not always be aware of the implicit assumptions embedded in a data set and tempted to take data at face value. For example, block-level socio-economic data from the American Community Survey (ACS) is frequently used without considering the significant error ranges resulting from the small sample size, conveniently ignoring the values in the error column. In many cases, however, researchers are very familiar with these issues, but choose to live with the limitations, biases and uncertainties of a data set in the absence of more reliable sources. Coining the phrase 'data friction', historian Paul Edwards describes the struggles that ensue when attempting to move data and metadata from one format, organization, scale or context to another (Edwards 2010).

Data frictions concern all forms of data, including those relating to cities and urban life, and manifest in various ways. Due to different definitions of, for example, metropolitan areas, data sets from international agencies such as the UN and the World Bank often remain incompatible. Even within the same institution, methodical details can change over time and lead to data frictions. Such is the case in the US Census, where changing geographic boundaries require considerable effort to harmonize data from different decades.[1] Finally, the assumptions and definitions underlying urban data can be shaped by political agendas that are not always obvious (Litman 2014).

Despite these difficulties and limitations, data sets, once collected, can develop a life of their own, and may become useful in ways initially not anticipated. After all, if the value of data were strictly determined by the circumstances and original purposes of data generation, there would be little reason to collect data in the first place. Besides the internal, also the external context is relevant: how observations relate to various phenomena in the city beyond those of immediate interest to the observation. The use of proxy data allows studying issues indirectly through data describing related phenomena. Considering that social science works with abstract constructs that usually cannot be directly observed, one could argue that most urban data sources are in some sense data proxies. Like data, proxies are imperfect by definition, representing some aspect of a phenomenon, omitting others, often conflating multiple issues that are hard to disentangle.

## Twitter as sticky data

Among social media services, Twitter has become a favourite data source for investigating a broad range of urban phenomena. Several reasons explain this popularity: tweets are a public form of communication, programmatically accessible through a public API.[2] The data are well-structured, available in large quantities and easy to process even for researchers with limited technical background. Twitter is a medium that is widely used on mobile devices in a variety of different social contexts, activities and situations. Containing annotated textual content of a defined length, a timestamp and sometimes a geographic location, tweets lend themselves to quantitative, qualitative and spatial modes of analysis. Due to all of these properties, Twitter has been used as a proxy for studying a broad range of phenomena. In the domain of urban research, studies that take advantage of Twitter data can be categorized into three groups depending on how the research objective relates to the context of data generation.

The first group includes studies that investigate Twitter in its original role as a social platform instead of using it as a proxy and focuses on its communication processes as they take place in physical space. This group includes analyses of topics discussed on the network and their spatial (Lansley and Longley 2016) and temporal structures (Naaman *et al.* 2012). Subjects include social network and community formation, attitudes and sentiments (Hollander and Renski 2015), representation of identity (Bailey 2016), or the use of the platform for activism and collective action (Jackson and Foucault Welles 2016). Also the use of Twitter as a mode of surveillance falls into this category.

The second group uses Twitter as a proxy for communication, interaction or information production in a broader sense. It investigates spatial phenomena that previously could not be studied due to unavailable data or were limited to small-scale qualitative studies. This group includes studies investigating the perception of neighbourhood boundaries (Wakamiya *et al.* 2013), the spatial distribution of languages (Hong *et al.* 2011) and regional dialects (Huang *et al.* 2016), as well as the global cultural boundaries inferred from these distributions (Mocanu *et al.* 2013). Conversely, this group also includes work that investigates the blind spots or absences of information production, for example as a proxy for the inequalities of digital labour (Graham 2014).

The third group of studies uses Twitter data as a proxy for phenomena not directly related to communication, for example, to estimate where people are at a given moment. In this case, Twitter is often merely a choice of convenience, interchangeable with any other location-based media service capable of generating time-stamped geotags. In this group, Twitter data are often used in conjunction with other data sources to describe and predict phenomena that were previously modelled through other means. The group includes studies of the demographic structure of urban populations (Longley *et al.* 2015), estimations of the number of pedestrians in public space (Lansley 2014), predictions of influenza infections (Broniatowski *et al.* 2013), transport behaviour (Hawelka *et al.* 2014; Wang and Taylor 2016), or land use (Frias-Martinez and Frias-Martinez 2014).

While the first two approaches are directly or indirectly related to the context of online communication, the third group uses Twitter as an opportunistic data source unrelated to its purpose. In this third group, questions of validity and accuracy become especially pertinent, as data from different sources need to be aligned and inherent biases have to be controlled for. However, this does not always happen – issues of internal validity of Twitter data are ignored in many studies (Perng *et al.* 2016).

At present, it seems that data from Twitter and other social media sources resist the appropriation as data proxies. Many studies share a similar conclusion: due to the limitations of the data source no significant results can be reported at the moment, but great future potential is to be expected.[3] Social media platforms are used in a multitude of different contexts by different user groups with a demographic composition that is poorly known and constantly changing. The inequality in the use of social media among the various urban neighbourhoods is bigger than the inequality along socio-demographic and economic dimensions (Indaco and Manovich 2016).

Twitter data seem to be sticky – to introduce a provisional characterization – meaningful when discussed in their original context, but difficult to separate from this context, requiring the use of sharp statistical instruments. Tweets are meaningful in their own right, but it is not clear which aspects of Twitter data can be extrapolated or generalized. Even as a proxy for human presence, the local context remains sticky, inseparable from the data. Eric Fisher's maps that compare where in major cities people tweet and where they take pictures show distinct spatial patterns that can only be explained by the motivations for using the medium in a given situation.[4] People tend to take photos rather than tweet on the Golden Gate Bridge or the Alcatraz Island, but they tend to tweet rather than take pictures in suburban residential neighbourhoods. These are interesting, perhaps generalizable findings, but they also generate significant data friction. Stickiness does not prevent marketers, entrepreneurs or social activists from using Twitter data, but for urban researchers it is important to consider how each aspect of Twitter data set relates to different phenomena in the city and devise methods to disentangle those aspects.

## OLS city lights as non-sticky data

Other data sources, however, seem to involve less friction and serve as reliable proxies for phenomena that are seemingly unrelated to the original purpose of data collection. A case where a data set has entirely transcended its original context is the accidental history and the widespread use of satellite mosaics generated from the Operational Linescan System (OLS) of the US Air Force's Defense Meteorological Satellite Program (DMSP). The satellite composites, provided by the National Geophysical Data Center of the US National Oceanic and Atmospheric Administration (NOAA) for every calendar year since 1992, contain the brightness levels of nocturnal city lights, accompanied by the flares of oil and gas fields, wildfires and the lights of fishing fleets.[5]

During the past 20 years, OLS/DMSP data have become a 'workhorse' for geographers and economists alike. The radiance values from the satellite mosaics have found their way into data sets and studies that estimate population density (Sutton 1997; 2003), urbanization and suburban sprawl (Campanella 2012; Sutton 2003), economic productivity (Henderson *et al.* 2009), rural poverty (Jean *et al.* 2016; Elvidge *et al.* 2009), resource footprints and electrification rates (Elvidge *et al.* 2011), measles outbreaks (Bharti *et al.* 2011) and average wages (Mellander *et al.* 2013).

Considering the range of phenomena the night-time city lights data have been used to predict and explain, it is remarkable that even the capability of recording city lights was an accidental by-product of a system designed for different purposes.

The Defense Meteorological Satellite Program was launched in 1961 by the US Air Force after it became clear that an effective photo-surveillance by satellites require an accurate prediction of cloud cover over the target area (Hall 2001). At the time, photo-reconnaissance satellites depended on photographic film, which had to be arduously recovered from re-entry capsules ('film-buckets') dropped from the satellite back to Earth, only to discover, as it was often the case, that the images only contained clouds.

As detailed by the Army historian Cargill Hall, the development of optical instruments used for the DMSP satellites involved many iterations, and led to several discoveries such as the value of the infrared band for detecting clouds. In 1966, the development of the OLS imaging module started, which continuously recorded digital luminance data transmitted wirelessly back to Earth (Hall 2001). A year after DMSP data became available for certain civilian agencies in 1972, Thomas A. Croft, researcher at the Stanford Center for Radar Astronomy, expressed his amazement about the images in *Nature* magazine: 'The lights of cities are clearly visible, as are the aurora, surface features illuminated by moonlight, and fires such as those caused by burning gas from oil fields and refineries' (Croft 1973). At the height of the 1973 oil crisis, Croft read the data as a testament to the global waste of energy. Meanwhile, the army had also found a different use for the city lights – to locate and calibrate the recorded night-time scenes accurately, and to estimate the thickness and density of particles in the atmosphere by measuring the diffusion of their contours (Air Weather Service 1974).

By 1977, Croft had compiled a first global atlas of city lights that used methods for digitization, processing both original films and digital facsimiles, using pattern recognition to align different viewpoints (Croft and Colvocoresses 1979). A year later, he published the first global composite of night-time images in the Scientific American (Croft 1978). In 1992, DPMS data were opened to the general public, and the NASA Black Marble illustrations have become one of the most popular motifs of space imagery.[6]

Despite its wide use, OLS/DMSP data are limited in many ways. As a proxy for estimating human presence and activity the data are strongly biased, with the brightness values for a specific region dependent on many socio-economic, political and cultural factors. OLS data models therefore always use available country-level data as controls to allow predictions for places where no such data are available. Furthermore, since OLS was designed to identify clouds rather than measure illumination, the brightness values do not allow estimations of luminance at the source, the values are therefore dimensionless. Fully saturated pixels covering brightly lit urban agglomerations are a further concern[7] as are blooming artefacts spilling into neighbouring pixels. While the imaging sensor operates autonomously and stable, stitching the recorded observations together introduces additional issues that require human decisions. A set of heuristics regulates how to ensure the best resolution, avoid sun and moonlight and excluding clouds.[8] Some of these rules can be implemented algorithmically, such as removing flares by normalizing brightness over time, others require a human touch, such as the identification and exclusion of aurora borealis. Some populated regions are rarely ever cloud-free, reducing data quality. Studies based on OLS data have managed to control for some of these limitations; other limitations are accepted simply due to the lack of alternative sources that have comparable spatial and temporal coverage. For recent years, better alternatives exist. Since 2012, the Visible Infrared Imaging Radiometer Suite (VIIRS) has superseded the OLS instrument providing data in superior quality and resolution.[9]

OLS mosaics are powerful visual artefacts. At first sight, the correlation of OLS data with human population density appears to be self-evident, yet without additional statistical controls, it is much smaller than expected (Elvidge *et al.* 1997). The level of trust inspired by OLS satellite mosaics might be explained by their obvious realism: their similarity with photographic material from the Apollo missions and other examples of space photography. However, this assumption of realism is shaky, as anyone can confirm who has worked with raw satellite scenes, which usually look nothing like their vibrant published versions. Rob Simmon, map designer at NASA's Earth Observatory, eloquently demonstrates that satellite composites are elaborate information visualizations, designed to evoke the impression of photographs (Simmon 2011). What appears as the translucence of shallow coastal waters in delicate shades of blue is, in fact, the rendering of a data set of oceanic chlorophyll activity. The colourful transitions between lush forests and arid regions were never captured by a photographic lens, but are determined by carefully crafted colour palettes.

Considering the long way from taking pictures of clouds to measuring urban economies, OLS data appears to be a non-sticky data source. However, as described above, its mobilization for research involves a considerable amount of data friction. The successful use of OLS data as a proxy is only possible because the data source and its methods are extensively documented and the behaviour of the sensor, with all its limitations, is well understood – prerequisites for overcoming data friction.[10]

## Deconstructing stickiness

At first glance, data from OLS/DMSP and Twitter seem to be very different forms of data. The one an exemplar of mechanical objectivity (Daston and Galison 2007), continuously recording under stable conditions, the other reflecting human communication in its irreducible richness. On closer inspection, many of these apparent differences disappear. Data from both sources are indices of human behaviour, yet on vastly different spatial and temporal scales. OLS feeds and geolocated tweets both indicate where people are at a given time, each with their own representations and subject to their own biases. While tweets are initiated by the user, OLS data sets are not free from ad-hoc human decisions either. Such decisions involve, for example, trade-offs between cloudless coverage and data quality for some parts of the globe. The provisional categorization of sticky and non-sticky data seems increasingly untenable.

Yet, stickiness remains palpable in certain challenges that are unique to social media, which exist in many different contexts, used by various groups for various purposes. While the procedures of generating OLS data are explicated and their biases are mostly known, less information exists about the demographic composition of social media users. What is more important, is that the contexts of social media are not stable but perpetually evolving, as new platforms get adopted and the use of existing ones evolves. As David Lazer and his colleagues demonstrated with the example of the declining prediction quality of Google flu trends, a model that accurately predicts a phenomenon at a given time may quickly become obsolete (Lazer *et al.* 2014). Social media are feedback systems, the behaviour of their users adapts in response to how they perceive the system (Offenhuber 2014). As the contexts of data generation on social media platforms are ephemeral, the only constants are the users themselves. For online marketers, 'sticky data' include user-IDs, email addresses and other indices for personal identification, which allow them to track users across the multiple contexts of a constantly shifting social media landscape.[11]

As proxies for urban phenomena, both data sources offer only partial perspectives. They are susceptible to what journalist Joe Cortright describes as the 'drunk under the streetlamp' fallacy, expressed in the dialogue: 'Did you lose your keys here? No, but the light is much better here' (Cortright 2016). To correct for their inherent biases, both sources require the triangulation with other data sets. In this context, the law of large numbers, or 'more trumps better' (Mayer-Schönberger and Cukier 2013) is only partially helpful and should be contrasted with the disclaimer 'garbage in – garbage out'. Large data volumes allow for more statistical controls to take biases into account but do not compensate for missing information. Combining, comparing and integrating multiple data sources seems the most promising way to go. Many seemingly unrelated data sources can complete each other as they are already linked in various hidden ways. This explains why, for example, a survey of noise exposure of a marginalized community can predict the impacts of air pollution on the same community (Franklin and Fruin forthcoming).

Social media data offer new lenses for observing urban phenomena. They can complement existing data sources to provide a more fine-grained view into spatial and temporal processes. As cultural expressions, their value and richness go beyond narrow measures of accuracy of validity. But just as nocturnal city lights illustrate the unequal distribution of people, the landscape of social media includes data deserts as well as hotspots of activity. The data footprints of equivalent media services rarely align. When matching and contrasting the data artefacts of social media with other sources, the frictions and shifting contexts of data generation continue to play an elementary role.

The stickiness of social media data resists the operationalization in automatic pipelines for knowledge extraction and manifests itself in false positives that can only be identified and resolved by a close reading of the source. This has consequences for the use of big data in urban governance, urban operation centres and predictive policing – applications that often rely on decontextualized data and reductive modes of analysis, such as text mining based on trigger words or dictionary-based sentiment analysis. Ignoring stickiness of context can lead to cases where a terrorism suspect identified by unsupervised text analysis turns out to be the journalist who reported on the issue (Currier *et al.* 2015). In this sense, stickiness points to issues of privacy even within the realm of publicly accessible data sources. As social media scholar Judith Donath notes, privacy fails when something that is intended for a particular context gets shown in another where it acquires a different meaning (2014: 212). Ignoring stickiness can also increase the susceptibility to various forms of manipulation and hacking, such as the practice of feeding fake information to crowd-sourced traffic systems like Waze in an attempt to create virtual traffic jams and re-route traffic flows.

The complications of ambiguous data or missing context can rarely be avoided, since the best proxy is often simply a data source for which no alternative exists. In the case of OLS data, new remote sensing instruments may allow for more accuracy and resolution, but lack the historical reach of OLS composites that cover four decades of global urbanization. Nevertheless, OLS mosaics afford only a single perspective on urban phenomena that are reflected in many different representations. Working with stickiness means integrating multiple partial perspectives rather than simply relying on a larger amount of data and reducing them down to the common lowest denominator.

## Notes

1 Companies such as Geolytics provide harmonized data products based on the US census. For more information, see http://www.geolytics.com.

2 Acronym for application program interface, a set of defined methods for accessing data automatically through scripts or programs.

3 It should be noted that the number of active Twitter users has stagnated over the past few years. See for example, http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users.

4 For his See Something or Say Something project, see http://www.flickr.com/photos/walkingsf/sets/72157627140310742/.

5 For further information, see http://ngdc.noaa.gov/eog/dmsp.html.

6 See http://earthobservatory.nasa.gov/Features/NightLights.

7 The island of Singapore, for example, appears as a fully saturated blob of light in all yearly mosaics since 1992.

8 See http://ngdc.noaa.gov/eog/gcv4_readme.txt.

9 See http://jointmission.gsfc.nasa.gov/viirs.html.

10 See for example: http://ghrc.nsstc.nasa.gov/uso/ds_docs/ols/ols_dataset.html.

11 See for example: http://www.towerdata.com/blog/what-is-sticky-data-and-why-do-i-need-it.

## References

Air Weather Service (1974) ‘DMSP User’s Guide’, USAF.

Bailey, M. (2016) ‘Redefining representation: Black trans and queer women’s digital media production’, *Screen Bodies* 1(2).

Bharti, N., Tatem, A.J., Ferrari, M.J., Grais, R.F., Djibo, A. and Grenfell, B.T. (2011) ‘Explaining seasonal fluctuations of measles in Niger using nighttime lights imagery’, *Science* 334(6061): 1424–1427.

Bowker, G.C. (2005) *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.

Broniatowski, D.A., Michael, P.J. and Dredze, M. (2013) ‘National and local influenza surveillance through Twitter: An analysis of the 2012–2013 influenza epidemic’, *PloS One* 8(12), Public Library of Science: e83672.

Campanella, T.J. (2012) *The Concrete Dragon: China’s Urban Revolution and What It Means for the World*. New York: Princeton Architectural Press.

Cortright, J. (2016) ‘The dark side of data-based transportation planning’, *CityLab*. Available from: http://www.citylab.com/commute/2016/08/the-downside-of-data-based-transportation-planning/496250/ [accessed 10 February 2017].

Croft, T.A. (1973) ‘Burning waste gas in oil fields’, *Nature* 245(5425): 375–376.

Croft, T.A. (1978) ‘Nighttime images of the Earth from space’, *Scientific American*, July, pp. 86–98.

Croft, T.A. and Colvocoresses, A.P. (1979) ‘The brightness of lights on Earth at night, digitally recorded by DMSP satellite’, US Geological Survey, available from: https://pubs.er.usgs.gov/publication/ofr80167 [accessed 10 February 2017].

Currier, C., Greenwald, G. and Fishman, A. (2015) ‘U.S. government designated prominent Al Jazeera journalist as “member of Al Qaeda”’, *The Intercept*, available from: https://theintercept.com/2015/05/08/u-s-government-designated-prominent-al-jazeera-journalist-al-qaeda-member-put-watch-list/ [accessed 10 February 2017].

Daston, L. and Galison, P. (2007) *Objectivity*. New York: Zone Books.

Donath, J. (2014) *The Social Machine: Designs for Living Online*. Cambridge, MA: MIT Press.

Drucker, J. (2011) ‘Humanities approaches to graphical display’, *Digital Humanities Quarterly* 5(1), available from: http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html [accessed 10 February 2017].

Edwards, P.N. (2010) *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.

Elvidge, C.D., Baugh, K.E., Kihn, E.A., Kroehl, H.W. and Davis, E.R. (1997) ‘Mapping city lights with nighttime data from the DMSP Operational Linescan System’, *Photogrammetric Engineering and Remote Sensing* 63(6): 727–734.

Elvidge, C.D., Baugh, K.E., Sutton, P.C., Bhaduri, B., Tuttle, B.T., Ghosh, T., Ziskin, D. and Erwin, E.H. (2011) ‘Who’s in the dark-satellite based estimates of electrification rates’, in Xiaojun Yang (ed.), *Urban Remote Sensing*. Chichester: John Wiley, pp. 211–224.

Elvidge, C.D., Sutton, P.C., Ghosh, T., Tuttle, B.T., Baugh, K.E., Bhaduri, B. and Bright, E. (2009) ‘A global poverty map derived from satellite data’, *Computers and Geoscience* 35(8): 1652–1660.

Floridi, L. (2011) *The Philosophy of Information*. Oxford: Oxford University Press.

Franklin, M. and Fruin, S. (forthcoming) ‘The role of traffic noise on the association between air pollution and children’s lung function’, *Environmental Research.*

Frias-Martinez, V. and Frias-Martinez, E. (2014) ‘Spectral clustering for sensing urban land use using Twitter activity’, *Engineering Applications of Artificial Intelligence* 35: 237–245.

Gitelman, L. (2013) *‘Raw Data’ is an Oxymoron*. Cambridge, MA: MIT Press.

Graham, M. (2014) ‘Internet geographies: Data shadows and digital divisions of labour’, *SSRN*. Available from: http://papers.ssrn.com/abstract=2448222 [accessed 10 February 2017].

Hall, C.R. (2001) ‘A history of the military polar orbiting meteorological satellite program’, DTIC Document, available from: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA598477 [accessed 10 February 2017].

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P. and Ratti, C. (2014) 'Geo-located Twitter as proxy for global mobility patterns', *Cartography and Geographic Information Science* 41(3): 260–271.

Henderson, V.J., Storeygard, A. and Weil, D.N. (2009) 'Measuring economic growth from outer space', National Bureau of Economic Research, available from: www.nber.org/papers/w15199 [accessed 10 February 2017].

Hollander, J.B. and Renski, H. (2015) 'Measuring urban attitudes using Twitter: An exploratory study', Lincoln Institute of Land Policy, Cambridge, MA.

Hong, L. Convertino, G. and Chi, E.H. (2011) 'Language matters in Twitter: A large scale study', in *Fifth International AAAI Conference on Weblogs and Social Media*, pp. 518–521.

Huang, Y., Guo, D., Kasakoff, A. and Grieve, J. (2016) 'Understanding U.S. regional linguistic variation with Twitter data analysis', *Computers, Environment and Urban Systems* 59: 244–255.

Indaco, A. and Manovich, L. (2016) 'Urban social media inequality: Definition, measurements, and application', arXiv. Available from: http://arxiv.org/abs/1607.01845. [accessed 10 February 2017].

Jackson, S.J. and Foucault Welles, B. (2016) '#Ferguson is everywhere: Initiators in emerging counterpublic networks', *Information, Communication and Society* 19(3): 397–418.

Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B. and Ermon, S. (2016) 'Combining satellite imagery and machine learning to predict poverty', *Science* 353(6301): 790–794.

Lansley, G. (2014) 'Evaluating the utility of geo-referenced Twitter data as a source of reliable footfall insight', available from UCL Discovery: http://discovery.ucl.ac.uk/1437075/ [accessed 10 February 2017].

Lansley, G. and Longley, P.A. (2016) 'The geography of Twitter topics in London', *Computers, Environment and Urban Systems* 58: 85–96.

Lazer, D., Kennedy, R., King, G. and Vespignani, A. (2014) 'Big data. The parable of Google Flu: Traps in big data analysis', *Science* 343(6176): 1203–1205.

Litman, T. (2014) 'What is a "house"? Critiquing the demographia international housing affordability survey', *Planetizen*. Available from: http://www.planetizen.com/node/70829 [accessed 10 February 2017].

Longley, P.A., Adnan, M. and Lansley, G. (2015) 'The geotemporal demographics of Twitter usage', *Environment and Planning A* 47(2): 465–484.

Mayer-Schönberger, V. and Cukier, K. (2013) *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, MA: Eamon Dolan.

Mellander, C., Stolarick, K., Matheson, Z. and Lobo, J. (2013) 'Night-time light data: A good proxy measure for economic activity?', Royal Institute of Technology, CESIS – Centre of Excellence for Science and Innovation Studies. Available from: http://ideas.repec.org/p/hhs/cesisp/0315.html [accessed 10 February 2017].

Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q. and Vespignani, A. (2013) 'The Twitter of Babel: Mapping world languages through microblogging platforms', *PloS One* 8(4): e61981.

Naaman, M., Zhang, A.X., Brody, S. and Lotan, G. (2012) 'On the study of diurnal urban routines on Twitter', ICWSM. Available from: http://sm.rutgers.edu/pubs/naaman-twitterpatterns-icwsm2012.pdf [accessed 10 February 2017].

Offenhuber, D. (2014) 'Infrastructure legibility – a comparative study of Open311 citizen feedback systems', *Cambridge Journal of Regions, Economy and Society* 8(1): 93–112.

Perng, S-Y., Kitchin, R. and Evans, L. (2016) 'Locative media and data-driven computing experiments', *Big Data and Society* 3: 1–12.

Simmon, R. (2011) ‘Crafting the blue marble’, Elegant Figures, NASA Earth Observatory. Available from: http://earthobservatory.nasa.gov/blogs/elegantfigures/2011/10/06/crafting-the-blue-marble/ [accessed 10 February 2017].

Sutton, P. (1997) ‘Modeling population density with night-time satellite imagery and GIS’, *Computers, Environment and Urban Systems* 21(3): 227–244.

Sutton, P. (2003) ‘A scale-adjusted measure of “urban sprawl” using night-time satellite imagery’, *Remote Sensing of Environment* 86(3): 353–369.

Wakamiya, S., Lee, R. and Sumiya, K. (2013) ‘Social-urban neighborhood search based on crowd footprints network’, in *International Conference on Social Informatics*, pp. 429–442.

Wang, Q. and Taylor, J.E. (2016) ‘Patterns and limitations of urban human mobility resilience under the influence of multiple types of natural disaster’, *PloS One* 11(1): e0147299.

# Part III

# Urban data technologies

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

# 9 Urban data and city dashboards

## Six key issues

*Rob Kitchin and Gavin McArdle*

## Introduction

In this chapter we examine six key issues with respect to how we come to know and manage cities through urban data and city dashboards. We seek to provide an agenda for critically reflecting on urban dashboards by examining six related questions:

1. How are insight and value derived from city dashboards?
2. How comprehensive and open are city dashboards?
3. To what extent can we trust city dashboards?
4. How comprehensible and useable are city dashboards?
5. What are the uses and utility of city dashboards?
6. How can we ensure that dashboards are used ethically?

We start, however, by answering a more prosaic question: what are city dashboards?

City dashboards use visual analytics – dynamic and/or interactive graphics (e.g. gauges, traffic lights, meters, arrows, bar charts, graphs), infographics and icons, maps, 3D models and augmented landscapes – to display and communicate information about the performance, structure, pattern and trends of cities. In effect, key data about cities – related to urban systems and infrastructure, society, economy, environment, population, etc. – are displayed on a screen, updated as new data are released and, in many cases, can be interacted with (e.g. selecting, filtering and querying data; zooming in/out, panning and overlaying; changing type of visualization or simultaneously visualizing data in a number of ways) (see Figure 9.1). In some cases, key data are 'consolidated and arranged on a single screen so the information can be monitored at a glance' (Few 2006: 34). Analytical dashboards are more extensive in scope and are hierarchically organized to enable a plethora of interrelated dashboards to be navigated and summary-to-detail exploration within a single system (Dubriwny and Rivards 2004). Both types of dashboard are common in urban control rooms, but they are also increasingly being displayed in mayor's offices and public buildings and made accessible to the general public via dedicated websites.

Typically, city dashboards display seven kinds of data. First, public administration data generated by local government, state agencies and government

Weather Stations 1

| Station | Wind Speed | Wind Gusts | Direction | Temperature | Humidity | Rain Today | Pressure | Forecast |
|---|---|---|---|---|---|---|---|---|
| CASA Office: Bloomsbury W1 | 8.2 mph | 7 mph | E ← | 17.7 °C | 53% | 1.6 mm | 1016.5 mbar | Cloudy |
| Hampstead NW3 | 1.6 mph | 3.3 mph | NW ↘ | 17.2 °C | 58% | 2.3 mm | 1017.0 mbar | Mainly Fine |
| Ben Cook: Bow E2 | 2 mph | 0 mph | NW ↘ | 17.3 °C | 61% | 3.0 mm | 1017.7 mbar | Dry Clear |

Tube Line Status (TfL) 29

| | |
|---|---|
| Bakerloo | Good Service |
| Central | Good Service |
| Circle | Good Service |
| District | Good Service |
| H & C | Good Service |
| Jubilee | Good Service |
| Metropolitan | Good Service |
| Northern | Good Service |
| Piccadilly | Good Service |
| Victoria | Good Service |
| W & C | Good Service |
| Overground | Good Service |
| DLR | Good Service |

Air Pollution (DEFRA) 1738

| 1 = Low<br>10 = V High | Ozone | $NO_2$ | $SO_2$ | $PM_{2.5}$ | $PM_{10}$ |
|---|---|---|---|---|---|
| Bloomsbury | n/a | n/a | n/a | 1 | 1 |
| Marylebone Rd | 1 | 2 | 1 | 2 | 2 |
| N Kensington | 2 | 1 | 1 | 2 | 1 |

Twitter: London News and Events 1

**Londonist:** Don't know your Edinburgh previews from your improv? Take a look at our guide to live comedy in London http://t.co/9zjUwS4O

Twitter: London Universities 10

**Imperial Spark:** "As a scientist I feel really depressed, because we are losing the war against obesity" - Prof Jimmy Bell http://t.co/C3evTR9H

*Figure 9.1* City dashboards: (a) Dublin (an analytical dashboard), (b) London (a city at a glance dashboard).

departments (e.g. housing, planning, education, welfare records, budget information). Second, official statistical data typically generated through surveys (e.g. a census or household/business surveys) undertaken by a national statistical institution or compiled from public administration data. Third, operational data concerning the delivery of services by local government or specific agencies (e.g. road usage or repair, the collection of garbage). Fourth, scientific data relating to environmental conditions (e.g. weather, water levels, pollution, noise). Fifth, crowd-sourced data provided by citizens (e.g. reporting incidents). Sixth, locative and social media data (e.g. geo-referenced data from social media, such as Twitter, accessed via APIs). Seventh, derived data – that is, data that are created by combining and analysing the other six types of data (e.g. composite indicators, forecasts/predictions, benchmarks). Typically, most data within city dashboards – especially of the analytical variety – are traditional in their ontology. That is, they are sampled data generated on a set schedule (e.g. monthly, annually). Increasingly, however, city dashboards are incorporating big data, especially with respect to operational and scientific data. That is, data that are produced in real-time by sensors, actuators, meters, transponders, cameras and computational devices, but also through crowd-sourcing and locative and social media.

The use of urban indicator and city dashboard projects have grown in use since the early 1990s, driven by: the rise of new managerialism and the desire to reform the public sector management of city services; citizen and funder demands for evidence-based decision-making; the calls for open data to provide greater accountability and transparency of service delivery; and the development of smart city initiatives that seek to develop data-driven urbanism (Innes and Booher 2000; Holden 2006; Behn 2014; Kitchin *et al.* 2015). City dashboards are becoming increasingly popular with city governments and agencies because they collate diverse sets and streams of indicator and big data into one system and provide tools to visualize, query and analyse them. In particular, they allow a user to track and compare over time and space, and in the case of real-time data, the here-and-now, of different phenomena. As such, they permit the following questions to be answered: how is the city performing with respect to key concerns? What are the spatial/temporal patterns of different phenomena? How do different parts of the city compare or how does a city compare with other cities? What is happening in the city right now?

To date, city dashboards have received little critical attention. In the remainder of this chapter we consider the epistemology, scope and access, veracity and validity, use and utility, usability and literacy, and ethics of city dashboards. Our analysis draws on an engagement with the wider literature and our own experiences of building the Dublin Dashboard, an analytical dashboard for the city.

## Epistemology

What are the underlying scientific assumptions of city dashboards? How do dashboards work to generate insight and value? These are epistemological questions. Dashboards utilize visualizations and visual analytics in order to make data about a city legible and interpretable. Visualizations have long been used to summarize and describe data sets because they effectively reveal and communicate the structure, pattern and trends of data and their interconnections. Digital visualizations can also be used to navigate and query data, enabling users to gain an overview of the entire data set, zoom in on items of interest, filter out uninteresting data, select and query an item or group of data, view relationships among data, and extract sub-collections (Shneiderman 1996). These actions are particularly useful for making sense of very large data sets, revealing structure, clusters, gaps and outliers that might otherwise remain hidden. Visualizations can also be used as a form of analytical reasoning. Here, a visualization is not simply describing or displaying the data, but is used as a visual analytical tool to extract information, build visual models and explanation, and to guide further statistical analysis (Keim *et al.* 2010). Often several different types of visual graphics are used in conjunction with each other so that the data can be examined from more than one perspective simultaneously. In addition, data mining and statistical modelling, such as prediction, simulation and optimization, can be performed and outputted through visual interfaces and outputs (Thomas and Cook 2006). In the context of city dashboards, this epistemology is framed within the emerging field of urban informatics (Foth 2009) and urban science (Batty 2013).

Visual analytics, urban informatics and urban science – and thus city dashboards – adopt a realist epistemology that supposes the existence of an external reality which operates independently of an observer and which can be objectively and accurately measured as quantitative data and be tracked, statistically analysed, modelled and visualized to reveal the world as it actually is. In other words, urban data can be abstracted from the world in neutral, value-free, objective and mechanical ways and are understood to be essential in nature; that is, representative of that which is being measured (they faithfully capture its essence and are independent of the measuring process) (Desrosières 1998; Porter 1995). And these data when analysed in similarly objective ways reveal the truth about cities. As such, dashboards have scientific utility because they seemingly translate the messiness and complexities of cities into rational, detailed, systematic, ordered forms of knowledge; they enable a city to be known and explained and to assess how it is performing in a neutral, comprehensive and common-sense manner (Mattern 2014; Kitchin *et al.* 2015).

Such a framing has been criticized for being too closely aligned with positivist thinking, being reductionist, mechanistic, atomizing, essentialist, deterministic and parochial, collapsing diverse individuals and complex, multidimensional social structures and relationships to abstract data points and formulae (Mattern 2013; Kitchin *et al.* 2015). It also wilfully ignores the metaphysical aspects of human life which are difficult to capture as data suitable for inclusion in a dashboard and generally ignores the role of politics, ideology, social structures, capital and culture in shaping cities (Kitchin 2014b). Indeed, they generally deal with facts, not with intangibles, processes, and complex, multi-scalar phenomena, and if used in isolation they decontextualize a city from its history, its political economy, the wider set of social, economic and environmental relations, and its wider interconnections and interdependencies that stretches out over space and time (Craglia *et al.* 2004; Mori and Christodoulou 2012).

Moreover, it has been contended that city dashboards do not simply present urban data, but actively produce meaning, generating new visions and understandings of a city that re-shape policy formulation and decision-making. As such, a dashboard is not simply a mirror of a city (with varying levels of methodological imperfections and noise), but acts as a translator by setting the forms and parameters for how data are communicated, interpreted and acted upon (Kitchin *et al.* 2015). This translation is ideologically framed and inherently political, reflecting design decisions framed within its development context (Kitchin *et al.* 2016). Their makers might envisage them as detached, passive, neutral scientific instruments that communicate the world as is (as can best be scientifically measured, processed and analysed), or recognize their issues and practise a form of strategic essentialism, but dashboards are the product of the ideas, instruments, practices, contexts, knowledges and systems used to generate, process and analyse them and they actively frame and do work in the world (Kitchin 2014a; Kitchin *et al.* 2015). They are underpinned by normative assumptions about what should be measured and what should be revealed.

This epistemological critique is not to say that city dashboards do not produce valuable insights or are not useful. As noted above visual analytics do produce interesting knowledge about cities and, as discussed below, this knowledge can be deployed in the management and governance of cities. But it is to say that dashboards are not objective, neutral mirrors of cities and need to be understood as producing a particular kind of knowledge that has a number of shortcomings and silences that need to be appreciated and taken into account. As the following sections document, these limitations extend beyond epistemology.

## Scope and access

How comprehensive and open are city dashboards? The first part of this question concerns the scope of the data included in a city dashboard. The second concerns the extent to which a dashboard and the data it displays are open to dashboard builders and the wider public.

In general, dashboards process and display factual, quantitative data; that is, data such as counts, rates, monetary value and scientific measurements that are numeric in format. Much of these data are generated recurrently meaning they can be tracked over time/space and are thus termed ‘indicator data’. Indicators can be direct in nature (e.g. measuring R&D spend to reflect investment in innovation) or indirect (e.g. using a proxy, such as the number of patents registered). Composite indicators combine several indicators using a system of weights or statistics to create a single value, recognizing that most phenomena (e.g. social deprivation) are interrelated and multidimensional and cannot be captured through a single measure (Maclaren 1996). Similarly, urban big data are generally structured, recurrent quantitative measures.

The scope of the data that dashboards display is thus limited. This means that there is an enormous amount of information about cities that are not displayed in city dashboards. Indeed, as noted, there is a diverse range of everyday activities, forms of urban living, and the nature of the human urban condition that are difficult to capture as indicator data. There are also significant gaps and silences in the data that are displayed. Quantified measures are typically narrowly defined, sampled and non-exhaustive (do not represent all people, places, times) and aggregated (variance is suppressed). Even with big data it is important to appreciate that there remains, and will continue to remain, an unevenness in the deployment of technologies that generate them (e.g. not everyone has a smartphone or uses smart cards). In addition, there are data sets that are chosen not to be included, for example, data that are considered to be politically sensitive or embarrassing (e.g. homelessness or service delivery performance). Further, those data that are used are strongly shaped by the technologies and instruments (e.g. the quality and calibration of a sensor) used to generate them which prescribe their parameters and form.

A second limitation concerns access and whether the data that are generated are available for re-use. Up until recently all forms of data used within city dashboards have been relatively difficult to access. Government data were typically locked inside institutions and when made available their use was restricted by cost, copyright and licensing arrangements (Kitchin 2014a). This situation is starting to change with the move to open data, though it is clear that the level of openness varies across administrations and places (Lauriault and Francoli 2016). Data generated by private institutions continues to be a valuable asset and is generally not available for use without a licence – consequently much of the deluge of urban big data is not available for city dashboards (though some companies enable a limited amount of data to be accessed through an API). And when public institutions are privatized, their data are often similarly privatized and become closed (see Chapter 5).

Even when data are available there are often issues related to data measurement (e.g. different agencies using alternate instruments, sampling strategies or classification schemes), data formats and media (e.g. data being released in alternate file types or forms difficult to process such as pdfs), metadata (that is, data about the data concerning lineage, characteristics and quality, which are often missing), data standards (e.g. different agencies using alternate data and metadata standards), lack of methodological guides or data dictionaries to define categories and explain codes, and modes of sharing (e.g. different forms of API). This can make it tricky to process and manage data and to compile comparable and interoperable data sets.

Further, the dashboard itself might not be openly accessible, being used operationally by an organization but not shared publicly. And in cases where the dashboard is made publicly available, the underlying data might not be open to access for re-use, only being presented for viewing/analysis. In some cases, this is because the framework data (e.g. base maps) are used under licence (which might be the case if the base mapping data are sourced from a national mapping agency) or because the attribute data (e.g. indicators) are. Similarly, the software used to create the dashboard might be propriety (produced by a company and used under licence or provided as a service) or be open source. The Dublin Dashboard, for example, presently uses a mix of open and propriety software tools. All of the city dashboards we are aware of have a closed form of development and governance, meaning that how they are formulated, their underlying algorithms, data sources and how they are run has limited scrutiny.

## Veracity and validity

To what extent can we trust city dashboards? This question refers to data quality and veracity, the appropriateness of the methods used and the validity of the analysis produced by and interpretation drawn from a dashboard.

A common warning related to data analysis is 'garbage in, garbage out'. In other words, if the data used in a dashboard have little veracity, then the analysis presented has little validity. All data sets contain instrument and human error and bias; generating data always involves a process of abstraction (capturing particular measurements from the sum of all possible data), representation (converting what is being measured into a readable form; e.g. numbers, a wave pattern, a scatterplot, a stream of binary code, etc.), and often generalization (e.g. into a set of categories) or calibration (transformation to compensate for suspected error/bias). They are produced and shaped by technical instruments of varying specification and parameters, handling procedures, scientific norms and standards, scientist behaviour and organizational processes. While a fact seems immutable it is important to note that they are produced, not simply measured (Bowker and Star 1999). For example, how unemployment is calculated varies across jurisdictions and changes over time, with each new formula producing a different rate. Calculating the population of a city seems straightforward but varies depending on who are selected for inclusion/exclusion (e.g. seasonal migrants) or where the boundary of the city is drawn. Likewise, altering the relative weightings of data in composite indicators can have a profound effect on the resulting score (Gruppa and Mogee 2004). There are then with every data set concerns about data veracity and quality.

Ideally, these concerns should be minimized through well designed and tested processes of data generation and handling and be documented so that others using the data are aware of any issues (McArdle and Kitchin 2016). Indeed, the importance of reporting data quality is recognized by all scientific bodies. For example, the International Cartographic Association (ICA) details seven key data quality metrics that should be documented in relation to spatial data (which are often used in city dashboards) (Guptill and Morrison 1995):

*Lineage*: The history of the data including details of the source material and any transformations or processes applied in order to produce the final data.

*Positional accuracy*: An indication of the horizontal and vertical accuracy of the coordinates used in the data, both to absolute and relative locations.

> *Attribute accuracy*: The accuracy of the quantitative and qualitative data attached to the spatial data.
>
> *Completeness*: The degree to which spatial and attribute data are included or omitted from the data sets. It also describes how the sample is derived from the full population and presents the spatial boundaries of the data.
>
> *Logical consistency*: The dependability of relationships within the spatial data.
>
> *Semantic accuracy*: The quality with which geographical objects are described in accordance with the selected model. Semantic accuracy refers to the pertinence of the meaning of the geographical object rather than its geometry.
>
> *Temporal data*: The date of observation, the type of update and the validity period for the data.

Data quality issues are also mandated by several ISO standards, such as ISO 19115-1:2014 and 19157:2013. These standards do not indicate acceptable thresholds for data quality, but rather mandate the metadata that needs be generated with respect to data veracity in order to receive the standard (McArdle and Kitchin 2016).

Despite big data being known to be messy and dirty in nature (Mayer-Schonberger and Cukier 2013), some have argued that using big data does not require the same standards of data quality and veracity as traditional data because the exhaustive nature of the data set removes sampling biases and compensates for any errors, gaps, inconsistencies or weakness in the fidelity of the data. For example, Mayer-Schonberger and Cukier (2013: 13) contend that ‘more trumps better’. Helland (2011) suggests that with very large data sets ‘“good enough” is good enough’. And Franks (2012: 211) argues that big data just needs to be ‘clean enough data’. But what is good enough or clean enough data? And can we trust such data to underpin critical decisions about city policy, investments and public safety? What we really require is big data to have the same levels of veracity as traditional data sets, to have and comply with their own ISO standards, and for this information to be available as metadata. At present, such veracity measures cannot be determined for big urban data.

In fact, it is difficult to establish the veracity of much of the data that are made available through open data sites and displayed in city dashboards. This is because the data used in city dashboards have no or limited data quality metadata, or such metadata is not shared. Nor do they detail the process through which raw data are manipulated for publication. For example, the Dublin Dashboard includes no metadata beyond source and timeliness. One is simply asked to trust that the data have veracity and validity.

Similarly, one is also asked to trust that the visual analytics and any calculations and modelling undertaken are valid. As most methods textbooks highlight, it is relatively straightforward to either unintentionally or deliberately lie or mislead with statistics and maps. How data are transformed and presented can make a significant difference to how they are interpreted. Ecological validity concerns the legitimacy of making inferences based on the outputs presented.

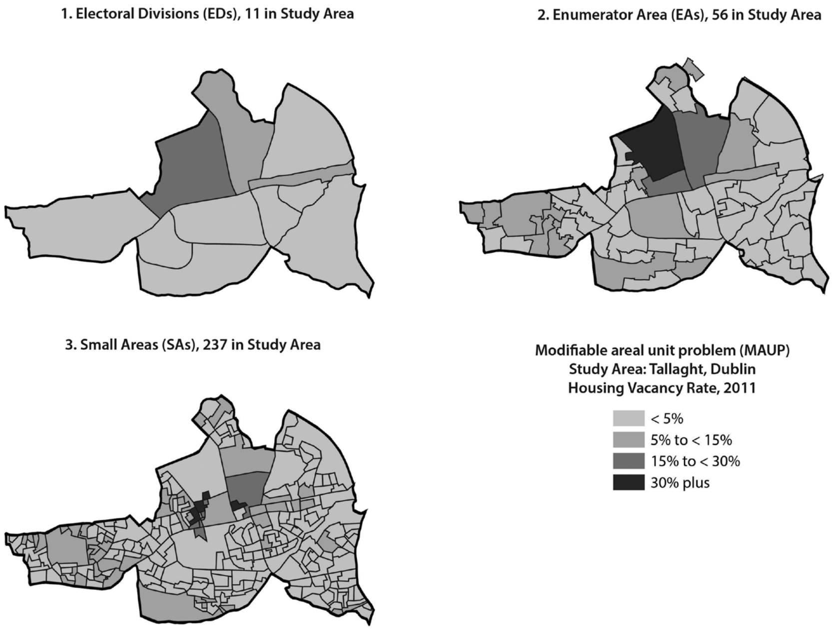


*Figure 9.2* Mapping the same data at three different administrative scales.

One of the most common types of ecological fallacy created within city dashboards is the Modifiable Areal Unit Problem (MAUP) (Openshaw 1984), wherein the statistical geography used to display aggregate data can have a marked effect on the pattern of observations. For example, Figure 9.2 displays the same housing vacancy data in three statistical geographies, producing varying patterns and interpretation. Likewise, altering the classification boundaries, or altering the number of classes, can have a similar effect. In other words, it is possible to draw very different conclusions depending on how the data are aggregated and presented. Similarly, the choice of analytics and models, and the selection and tweaking of parameters within them, can produce markedly different results. Yet how a calculation or model is formulated is often black-boxed, meaning that its workings are not made available to others to assess or to replicate. In these cases, users are asked to trust that the analytics are producing valid analysis which leads to sensible interpretation.

## Usability and literacy

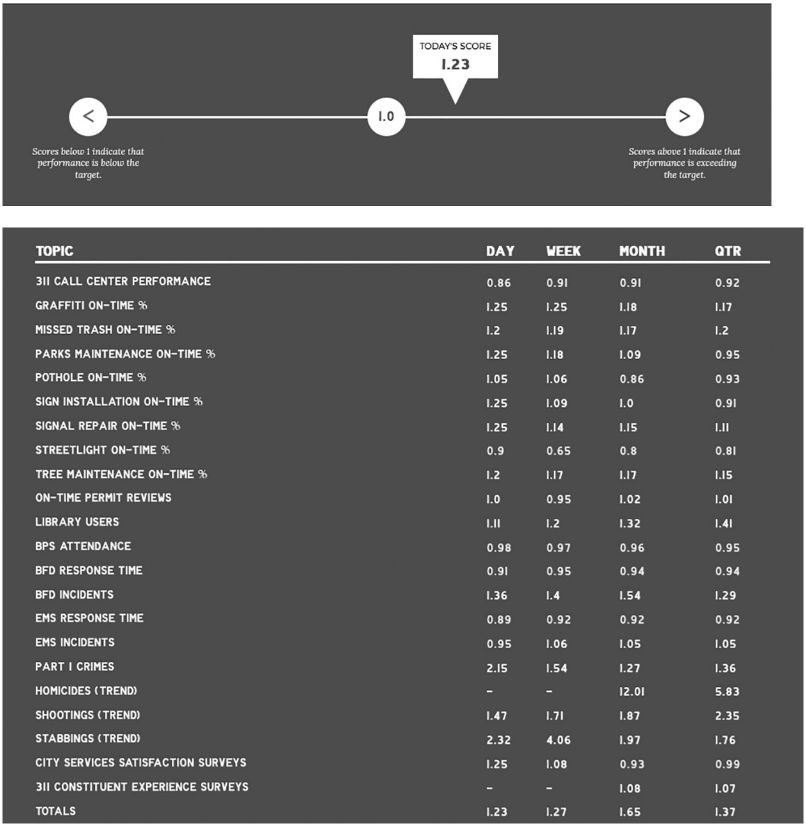
How comprehensible and useable are city dashboards? There is an assumption that city dashboards – especially those that are publicly available – enable a suite of urban data to be explored, analysed and interpreted in an easily digestible and intuitive way without the need for specialist skills or knowledge. In part, this is because the systems are point-and-click in nature and require no knowledge of how to produce interactive, dynamic graphics, maps or analytics. There are three issues here – navigation of site, use of tools, and data, analytics and visualization literacy.

For a city-at-a-glance dashboard navigation is straightforward. However, for an analytical dashboard that contains a number of modules navigation around the site and finding pertinent data and analysis can be more tricky. The Dublin Dashboard, for example, had 56 sub-modules at the time of writing, some that utilize hundreds of data variables (e.g. census mapping modules). At present, there is no detailed site map or data directory. There is also no deep sense of navigation, and there are search and browse issues as there has been no detailed user-testing. In terms of the tools presented, the Dublin Dashboard utilizes a number of different software data visualization and mapping programmes. It is sometimes not at all clear how to display data, change to new data layers, perform analysis, interact with data, etc. Nor are there any user guides to explain how to undertake different tasks. Again, there has been no user requirement or user testing analysis. This is not uncommon for city dashboards or open data sites. What this means is that city dashboards provide a sub-optimal experience for non-specialist users and their full utility is not being realized by most citizens and decision-makers alike.

Dashboards assume that their users understand what data are being presented (and to take into account issues of formulation, error/bias, etc.), can make sense of and validly interpret various forms of visualization and maps, and understand any analytics being undertaken. This, however, is not the case, and data, analytic and visualization literacy are highly variable across the general public, but also specialist users such as planners and policymakers. This is especially the case for analytics such as modelling where it is not clear how the model is calculated or what the output means. For example, in the case of Boston's City Score, the user is presented with a table of numbers, updated daily, that denote how well the city is performing in relation to a number of tasks (Figure 9.3; www.boston.gov/cityscore). In the accompanying text the viewer is told that a number above 1 means a target is being exceeded, whereas below 1 the service is deficient and will be reviewed. However, there is no detailed information on what the tasks being monitored are, with several unexplained acronyms, and no information on the form and veracity of underlying data or how the scores are calculated and targets are set. It is therefore quite difficult to interpret the information presented.

## Uses and utility

What are the uses and utility of city dashboards? City dashboards are used in three main ways: for monitoring performance and managing urban services; for contextually understanding and formulating policy; and for creating public knowledge and producing counter-narratives. In addition, they can be used tactically (e.g. for delaying a strategy, substitute for action, deflect criticism), symbolically (e.g. to provide reassurance or place promotion), politically (e.g. ammunition to support a particular position) (Hezri 2004), and if the data within them are open they can help facilitate and promote an open data economy and produce transparency and accountability.



Scores below 1 indicate that performance is below the target.

Scores above 1 indicate that performance is exceeding the target.

| TOPIC | DAY | WEEK | MONTH | QTR |
|---|---|---|---|---|
| 311 CALL CENTER PERFORMANCE | 0.86 | 0.91 | 0.91 | 0.92 |
| GRAFFITI ON-TIME % | 1.25 | 1.25 | 1.18 | 1.17 |
| MISSED TRASH ON-TIME % | 1.2 | 1.19 | 1.17 | 1.2 |
| PARKS MAINTENANCE ON-TIME % | 1.25 | 1.18 | 1.09 | 0.95 |
| POTHOLE ON-TIME % | 1.05 | 1.06 | 0.86 | 0.93 |
| SIGN INSTALLATION ON-TIME % | 1.25 | 1.09 | 1.0 | 0.91 |
| SIGNAL REPAIR ON-TIME % | 1.25 | 1.14 | 1.15 | 1.11 |
| STREETLIGHT ON-TIME % | 0.9 | 0.65 | 0.8 | 0.81 |
| TREE MAINTENANCE ON-TIME % | 1.2 | 1.17 | 1.17 | 1.15 |
| ON-TIME PERMIT REVIEWS | 1.0 | 0.95 | 1.02 | 1.01 |
| LIBRARY USERS | 1.11 | 1.2 | 1.32 | 1.41 |
| BPS ATTENDANCE | 0.98 | 0.97 | 0.96 | 0.95 |
| BFD RESPONSE TIME | 0.91 | 0.95 | 0.94 | 0.94 |
| BFD INCIDENTS | 1.36 | 1.4 | 1.54 | 1.29 |
| EMS RESPONSE TIME | 0.89 | 0.92 | 0.92 | 0.92 |
| EMS INCIDENTS | 0.95 | 1.06 | 1.05 | 1.05 |
| PART I CRIMES | 2.15 | 1.54 | 1.27 | 1.36 |
| HOMICIDES (TREND) | – | – | 12.01 | 5.83 |
| SHOOTINGS (TREND) | 1.47 | 1.71 | 1.87 | 2.35 |
| STABBINGS (TREND) | 2.32 | 4.06 | 1.97 | 1.76 |
| CITY SERVICES SATISFACTION SURVEYS | 1.25 | 1.08 | 0.93 | 0.99 |
| 311 CONSTITUENT EXPERIENCE SURVEYS | – | – | 1.08 | 1.07 |
| TOTALS | 1.23 | 1.27 | 1.65 | 1.37 |

*Figure 9.3* Boston City Score.

The realist epistemology that underpins the logic and workings of city dashboards promotes an instrumental rationality in which cities can be steered and managed through a set of data levers and analytics and that urban issues can be solved through a range of technical solutions (Mattern 2013; Kitchin *et al.* 2015). Here, city dashboards are used to: monitor and guide operational and policy practices with respect to specified targets; provide evidence of the success or failure of programmes and policies; discipline and reward performance; guide the development of new strategies; and shape spending patterns (Craglia *et al.* 2004; Behn 2014). An example of such an approach is Baltimore's use of CitiStat. Every week the mayor and city managers meet in a specially designed room using dashboards to review performance and set new targets for the city as a whole and for each department (Gullino 2009). This approach has been adopted in whole or part by a number of other US cities. For critics, this instrumental rationality promotes a technocratic form of governance that: forecloses other modes of governance and other forms of knowledge (such as phronesis – knowledge derived from practice and deliberation; and metis – knowledge based on experience) (Parsons 2004); fails to recognize that cities are complex, multifaceted, contingent, relational systems, full of contestation and wicked problems that are not easily captured or steered (Kitchin *et al.* 2015); and that urban issues are often best solved through political/social, public policy and public investment solutions and citizen-centred deliberative democracy rather than technical fixes (Kitchin 2014b).

In contrast, some municipalities use dashboards in a more contextual way. Here, it is recognized that cities are not mechanical systems that can be disassembled into its component parts and fixed, or steered and controlled through data levers. Instead, systems and governance are understood as complex and multi-level in nature, and the effects of policy measures are diverse and multifaceted, and neither is easily reducible to targets and performance metrics (Van Assche *et al.* 2010). Indicators highlight trends and potential issues, but do not show their causes or prescribe answers. Conceived in this way city dashboards provide useful contextual data – that can be used in conjunction with other data and initiatives – but are not used in a strongly instrumentalist, mechanistic way to direct management practices (Kitchin *et al.* 2015). A longstanding example of such an approach is that employed within Flanders, Belgium, where since the late 1990s a number of cities have employed a common City Monitor for Sustainable Urban Development, consisting of nearly 200 indicators, to provide contextual evidence for policymaking (Van Assche *et al.* 2010). The Dublin Dashboard follows this model. Nonetheless, in both managerial and contextual uses of dashboards, they are viewed as providing a stronger evidential base for city management than anecdote and occasional studies.

In cases where a city dashboard is publicly accessible it is hoped that it provides the same kinds of utility as open data in general – that is, it enables transparency, accountability and participation by providing the public with the data and the tools to extract insight and value from these data (Kitchin 2014a). In other words, it allows citizens to evaluate the work of city agencies in providing services and managing and governing the city, and it allows them to take an active role in contributing to evidence-informed debate and policymaking and to produce counter-narratives to those produced by authorities and other vested interest groups. Here, open city dashboards work to democratize the ability to produce information and knowledge, rather than the power of data being confined to its producers and those in a position to pay for data and tools.

## Ethics

How can we ensure that dashboards are used ethically? There has been a lot of concern with respect to the generation and use of personally identifiable information (PII) in the big data age, including those generated by smart city technologies (Kitchin 2016). However, city dashboards display aggregate and anonymous data or data that concerns a system rather than people. As such, ethical issues related to PII, such as individual level privacy and predictive privacy harms, are generally not pertinent. That is not to say that there are no ethical issues arising from city dashboards.

The data within city dashboards can be used to construct place profiles and histories that can be used as the basis for the social and spatial sorting of places and communities. Indeed, there is a multi-billion-dollar geodemographics industry that does precisely this, using place profiles to geo-target advertising/marketing and private investment and to calculate insurance premiums and online prices (Harris *et al.* 2005). Similarly, place profiles can be used by the public sector to determine which areas should receive place-targeted investment, additional policing or differential service provision. The data can also be used to discriminate areas of blight and problems and to reinforce territorial stigma, effecting public perception and affecting local community cohesion. In other words, the data and tools in dashboards can be used to treat places and the populations within them differentially in ways that can be discriminatory and affect quality of life. It is therefore important to consider the ways in which city dashboards are used and to consider whether their use is fair, equitable or prejudiced and how any issues might be addressed.

## Conclusion

In this chapter we have sought to document a number of key issues and questions with respect to the production and use of city dashboards. While we have provided a critical assessment that challenges some of the dominant thinking with respect to city dashboards, urban informatics and visual analytics, we also believe that dashboards provide useful insights and have much utility. Indeed, we have invested much time and effort into building the Dublin Dashboard and Cork Dashboard and working with Dublin's open data portal, Dublinked, and other data providers.

What is required, we believe, for city dashboards to reach their full potential as a smart city technology that can help produce more efficient, equitable, sustainable and resilient cities is a number of related shifts in thinking and praxes. First, there has to be a shift in the underlying epistemology of city dashboards to recognize that they conceive the urban in a particular way and seek to understand and explain the city using an approach which produces delineated and situated knowledge rather than communicating objective, scientific truths. Second, limitations with respect to the scope of data and accessing data sets need to be documented and also tackled by working with agencies and companies to open key data sets. Third, significant work needs to be undertaken to establish the veracity and validity of data sets and analytics and dashboards need to be populated with appropriate metadata and supporting documentation. Fourth, the usability of city dashboards in general and each specific tool needs to be established through user testing, as well as methods to improve user experience, and training and education tools to aid and improve data/analytics literacy need to be developed and included in dashboards. Fifth, the instrumental rationality of city dashboards needs to be reconceived, with dashboards always used in conjunction with other forms of knowledge and other modes of governance when evaluating, managing and formulating the delivery city services and policy. Sixth, the factors shaping the development of dashboards such as licensing, standards, institutional practices, local politics, choices about what data to included needs to be acknowledged and citizens need to be consulted about their configuration and deployment. And lastly, the ethics and potential harmful uses of using city dashboards need to be further examined and strategies need to be developed to minimize harm.

While this list is by no means comprehensive it provides an initial agenda for addressing the issues we have discussed. This shift in thinking will allow urban dashboards to reach their full potential and align them with the vision of the Open Knowledge Foundation (n.d.) who state that: 'Open knowledge is what open data becomes when it's *useful, usable and used* – not just that some data is open and can be freely used, but that it is useful – accessible, understandable, meaningful, and able to help someone solve a real problem.' Moreover, this agenda will provide useful insights for considering the wider relationship between data and the city.

## Acknowledgements

The research for this chapter was funded by a European Research Council Advanced Investigator grant, The Programmable City (ERC-2012-AdG-323636), and a Science Foundation Ireland grant, Building City Dashboards (15/IA/3090). The chapter draws extensively on three previously published papers: Kitchin *et al.* (2015 and 2016) and McArdle and Kitchin (2016).

## References

Batty, M. (2013) *The New Science of Cities*. Cambridge, MA: MIT Press.

Behn, R.D. (2014) *The Performance Stat Potential: A Leadership Strategy for Producing*. New York: Brookings Institution Press/Ash Center.

Bowker, G. and Star, L. (1999) *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Craglia, M., Leontidou, L., Nuvolati, G. and Schweikart, J. (2004) 'Towards the development of quality of life indicators in the "digital" city', *Environment and Planning B* 31(1): 51–64.

Desrosières, A. (1998) *The Politics of Large Numbers: A History of Statistical Reasoning*, trans. C. Naish. Cambridge, MA: Harvard University Press.

Dubriwny, D. and Rivards, K. (2004) 'Are you drowning in BI reports? Using analytical dashboards to cut through the clutter', *DM Review*, April. Available from: http://internal.advizorsolutions.com/press/Cut%20Through%20The%20Clutter.pdf [accessed 29 August 2016].

Few, S. (2006) *Information Dashboard Design: The Effective Visual Communication of Data*. Sebastopol, CA: O'Reilly Media.

Foth, M. (ed.) (2009) *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City*. Hershey, PA: IGI Global.

Franks, B. (2012) *Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics*. Hoboken, NJ: Wiley.

Gruppa, H. and Mogee, M.E. (2004) 'Indicators for national science and technology policy: How robust are composite indicators?', *Research Policy* 33(9): 1373–1384.

Gullino, S. (2009) ‘Urban regeneration and democratization of information access: CitiStat experience in Baltimore’, *Journal of Environmental Management* 90: 2012–2019.

Guptill, S.C. and Morrison, J.L. (eds) (1995) *Elements of Spatial Data Quality*. Amsterdam: Elsevier.

Harris, R., Sleight, P. and Webber, R. (eds) (2005) *Geodemographics, GIS and Neighbourhood Targetting*. Chichester: John Wiley.

Helland, P. (2011) ‘If you have too much data, then “good enough” is good enough’, *ACM Queue* 9(5).

Hezri, A.A. (2004) ‘Sustainability indicators system and policy processes in Malaysia: a framework for utilisation and learning’, *Journal of Environmental Management* 73(4): 357–371.

Holden, M. (2006) ‘Urban indicators and the integrative ideals of cities’, *Cities* 23(3): 170–183.

Innes, J. and Booher, D.E. (2000) ‘Indicators for sustainable communities: A strategy building on complexity theory and distributed intelligence’, *Planning Theory & Practice* 1(2): 173–186.

Keim, D., Kohlhammer, J., Ellis, G. and Mansmann, F. (2010) *Mastering the Information Age – Solving Problems with Visual Analytics*. Eurographics Association. Available from: www.vismaster.eu/book [accessed 29 August 2016].

Kitchin, R. (2014a) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: SAGE.

Kitchin, R. (2014b) ‘The real-time city? Big data and smart urbanism’, *GeoJournal* 79(1): 1–14.

Kitchin, R. (2016) *Getting smarter about smart cities: Improving data privacy and data security*. Data Protection Unit, Department of the Taoiseach, Dublin, Ireland. Available from: www.taoiseach.gov.ie/eng/Publications/Publications_2016/Smart_Cities_Report_January_2016.pdf [accessed 29 August 2016].

Kitchin, R., Lauriault, T.P. and McArdle, G. (2015) ‘Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards’, *Regional Studies, Regional Science* 2: 1–28.

Kitchin, R., Maalsen, S. and McArdle, G. (2016) ‘The praxis and politics of building urban dashboards’, *Geoforum* 77: 93–101.

Lauriault, T.P. and Francoli, M. (2016) ‘Openness, transparency, participation’, in R. Kitchin, T.P. Lauriault and M.W. Wilson (eds), *Understanding Spatial Media*. London: SAGE, pp. 188–203.

Maclaren, V.W. (1996) ‘Urban sustainability reporting’, *Journal of the American Planning Association* 62(2): 184–202.

Mattern, S. (2013) ‘Methodolatry and the art of measure: The new wave of urban data science’, *Places Journal*. Available from: https://placesjournal.org/article/methodolatry-and-the-art-of-measure [accessed 29 August 2016].

Mattern, S. (2014) ‘Interfacing urban intelligence’, *Places Journal*. Available from: https://placesjournal.org/article/interfacing-urban-intelligence [accessed 29 August 2016].

Mayer-Schonberger, V. and Cukier, K. (2013) *Big Data: A Revolution that will Change How We Live, Work and Think*. London: John Murray.

McArdle, G. and Kitchin, R. (2016) ‘Improving the veracity of open and real-time urban data’, *Built Environment* 42(3): 446–462.

Mori, K. and Christodoulou, A. (2012) ‘Review of sustainability indices and indicators: Towards a new City Sustainability Index (CSI)’, *Environmental Impact Assessment Review* 32: 94–106.

Open Knowledge Foundation (n.d.) ‘Vision and values’. Available from: http://okfn.org/about/vision-and-values [accessed 8 November 2016].

Openshaw, S. (1984) *The Modifiable Areal Unit Problem.* Concepts and Techniques in Modern Geography 38. Norwich: Geo Books.

Parsons, W. (2004) ‘Not just steering but weaving: Relevant knowledge and the craft of building policy capacity and coherence’, *Australian Journal of Public Administration* 63(1): 43–57.

Porter, T.M. (1995) *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.

Shneiderman, B. (1996) ‘The eyes have it: a task by data type taxonomy for information visualizations’, *Proceedings IEEE Visual Languages* 96: 336–343.

Thomas, J.J. and Cook, K.A. (2006) ‘A visual analytics agenda’, *IEEE Computer Graphics & Applications* 26: 10–13.

Van Assche, J., Block, T. and Reynaert, H. (2010) ‘Can community indicators live up to their expectations? The case of the Flemish city monitor for livable and sustainable urban development’, *Applied Research Quality Life* 5: 341–352.

# 10 Sharing and analysing data in smart cities

*Pouria Amirian and Anahid Basiri*

## Introduction

Nowadays, the successful and efficient management of a city depends on how data are collected, shared and transferred within and between various organizations in a city and how data analytics are used to extract actionable insights for decision-making. Such data include public administrative records, operational management information, as well as that produced by sensors, transponders and cameras that make up the internet of things, smartphones, wearables, social media, loyalty cards and commercial sources. In many cases, cities are turning to big data technologies and their novel distributed computational infrastructure for the reliable and fault tolerant storage, analysis and dissemination of data from various sources. In such systems, processing is generally brought to the data, rather than bringing data to the processing. Since each organization uses different platforms, operating systems and software to generate and analyse data, data sharing mechanisms should ideally be provided as platform-independent services so that they can be utilized by various users for different purposes, for example, for research, business, improving existing services of city authorities and organizations, and for facilitating communication between people and policymakers. Such sharing and communication amongst different entities in the city aligns with the vision of the smart city, which includes use of ICT technologies to improve the efficiency of service delivery, create sustainable development, as well as engaging citizens in decision-making.

Platform independency is necessary for providing interoperability from technical point of view. The interoperability of systems and services at various levels is an important requirement for public services and it is well defined in initiatives like the European Interoperability Framework (EIF) and many national interoperability frameworks. These frameworks ensure that the exchange of data is an ultimate enabler for sharing information and knowledge between organizations. This chapter has a wider view of interoperability in the context of smart cities, wherein it is also important for enabling service, citizen innovation and civic engagement, providing technical solutions to city problems, and producing sustainable development.

In addition to platform independency, in order to make the services as interoperable, resourceful and flexible as possible the services need to be designed based on certain principles. These principles are dependent on the type of application deployed and the users of those services. It is our view that data sharing principles for smart cities should be grounded in the concept of service orientation. This chapter describes the concept of service orientation principles (SOP) and explains three core approaches currently utilized for sharing data and analysis services (Web Services, RESTful services and Geoservices). The chapter demonstrates the need for and proposes a new architecture (Organizational Service Layer) to implement polyglot binding for flexible, scalable, efficient and interoperable implementation of data sharing and analysis services in a smart city.

## Service orientation principles

In service orientation, applications are constructed based on entities called 'services' (Erl *et al*. 2013). These services are underpinned by a set of design principles that are based on previous paradigms and practices in software engineering, such as component-based design, interface-based programming and distributed computing. The most widely referenced service orientation principles are loose coupling, abstraction, composability, standardized service contract, reusability, autonomy, statelessness and discoverability (Erl *et al*. 2013; Barry 2003; Erl *et al.* 2014) (see Table 10.1). The service orientation principles form

*Table 10.1* Service orientation principles

| *N* | *Principle* | *Brief explanation* |
|---|---|---|
| **1** | Standardized service contract | Any service must provide a formal contract that describes the service and defines the data exchange details (Amirian *et al*. 2010a). In order to consume services, the service contract is needed (URL of service contract). The service contract must be implemented and published using standard and well-defined technologies. |
| **2** | Abstraction | Services must be abstracted from the underlying logic and data. The underlying logic of services is invisible to service consumers. |
| **3** | Loose coupling | Services must be decoupled from their surrounding environment. In any software system, coupling is unavoidable. In fact, developers add value by implementing a system use case or a feature by coupling software functionality together. The loose coupling principle is about avoiding platform specific coupling. In other words, this principle means the interaction between services and users must be message-based. The principle of loose coupling is achieved through the use of service contracts that allow services to interact with the outside world via predefined parameters which are defined in the service contract. |

| | | |
|---|---|---|
| **4** | Composability | A service can represent any range of logic from any types of resources, including other existing services. Services can be composed of other services. However, services must be designed to participate as a member to the composed services if service compositions required. |
| **5** | Reusability | Services must be designed with reusability in mind. The reusability principle targets all forms of reusability, including inter-application, composition and utility services. Since a service encapsulates underlying logic, reusability in this context refers to generic operations and intelligent messages. In other words, when the messages contain enough metadata about the processing instruction, business rules and policies, the operations that comprise a service become more generic and as a result the service will be more reusable. |
| **6** | Autonomy | Services must have a high-level self-governance of their processing. The logic governed by a service resides within an explicit boundary. Services can share computing resources (such as a DBMS) with other services in that boundary and a service may depend on other services for execution of the underlying logic. However, at the time of execution, the service has control over whatever logic it exposes. |
| **7** | Statelessness | Services must minimize managing state information. The stateful services (the services that need to keep state data in order to process the requests) are usually less reusable and cannot be considered an efficient member in compositions. |
| **8** | Discoverability | Service requesters must be able to discover and understand contracts of services. Often this principle implemented using service registry systems. |

the basis for creating a Service Oriented Architecture (SOA) and cloud computing (Zimmermann *et al.* 2013). The SOA as a conceptual architecture is presently adopted by many organizations as an efficient means for integrating current enterprise applications and legacy applications (Amirian *et al*. 2010b). SOA is frequently characterized as a style that supports loose coupling, permitting extensibility and interoperability independent of the underlying technology (Erl *et al.* 2014). Service orientation principles and SOA can be implemented using any platform and technologies. However, Web Services and RESTful services are the most widely used technologies for the implementation of SOA in an open and standard manner (Daigneau 2011).

## Web Services and REST services

The World Wide Web Consortium (W3C) defines a Web Service as a software system designed to support interoperable machine-to-machine interaction over a network. Web Services are implemented using a collection of standards and technologies. SOAP and WSDL form the core implementation technologies

(Booth *et al.* 2004). SOAP is a lightweight, XML-based protocol for exchanging information in decentralized and distributed environments. SOAP is used for messaging among various SOA components and other systems interact with the Web Service in a manner prescribed by its description using SOAP messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards. WSDL is XML-based specification for describing the capabilities of a service in a standard and extensible manner. Technically, WSDL defines the software interface of a Web Service independently of the platform.

Web Services are based on open standards, so they provide interoperability in decentralized and distributed environments like the Web. These technologies can be developed using any software platform, operating system, programming language and object model. Web Services are not limited to the Web and can utilize any transport protocols such as TCP, HTTP and UDP. Web Service technologies can be extended using other related standards for supporting security, transaction management, composition, coordination and workflows through second generation Web Services standards which are called WS-*. Almost all these specifications are standard and maintained by W3C and OASIS organizations.

Web Services have been widely used for exposing data and functionality between systems since their introduction. They provide many advanced and complex features. The advanced features of Web Services are not necessary for exposing a resource (like reading a small set of data from a database) in many simple scenarios, such as public Web applications and connected mobile apps. In this context, REST (Representational State Transfer) has gained widespread acceptance across the Web as a simpler alternative to Web Services (SOAP and WSDL) (Erl 2008). Key evidence of this shift in interface design is the adoption of REST by mainstream Web 2.0 service providers. Microsoft, Google and Facebook have deprecated Web Service-based interfaces in favour of an easier-to-use, resource-oriented model of REST to expose their services. REST defines a set of architectural principles which allows designing services that focus on a system's resources. REST is primarily used to build services over the Web that are lightweight, maintainable and scalable.

A service based on REST is called a RESTful service. REST is not dependent on any protocol, but almost every RESTful service uses HTTP as its underlying protocol. Its importance is likely to continue to grow as all technologies move towards an API orientation. Based on REST architectural design principles, only HTTP methods need to be used explicitly for different purposes (Fielding 2000). This basic REST design principle establishes a one-to-one mapping between create, read, update and delete (CRUD) operations and HTTP methods. According to this mapping: POST is used to create a resource on the server, GET is used to retrieve a resource, PUT is used to update or change the state of a resource, and DELETE used for removing a resource.

RESTful services are easier to implement, maintain and utilize, but they do not have powerful and standard support for features like standard contract (or machine-readable service description) distributed transactions, composition and security, which are needed in most enterprise applications, and that is why most enterprise applications implement Web Services (Daigneau 2011). In addition, there is no standard discovery mechanism for RESTful services other than HTTP OPTION, which only provides the list of available methods that are supported by a service. In other words, implementation of RESTful services is done using documentation of the services and there is no such thing as a service contract for RESTful services.

However, both Web Services and RESTful services are not considered as standard approaches for sharing data and analysis in the geospatial community. Instead, there are other types of specifications which are maintained by OGC and they are called OGC Web services (or Geoservices) and are used for almost the same purpose, but with limited capabilities and scope.

## OGC Web Services (Geoservices)

Geoservices are considered the most promising technology for overcoming the non-interoperability problem associated with current geospatial processing systems (Sample *et al.* 2008). The OGC has defined a comprehensive framework of Geoservices which is known as the 'OGC Web Services framework' (OWS). The OWS allows distributed spatial processing systems to interact over the HTTP protocol, and provides a framework of interoperability for many Web-based services, such as accessing spatial data services, spatial processing services and data locating services (Percivall *et al.* 2003). The OWS framework consists of interface implementation specification for Geoservices and encodings which are openly available to be implemented by developers and companies. The interface implementation specifications are software technology neutral details. The encodings provide the standard glue among different parts of OWS. Each service

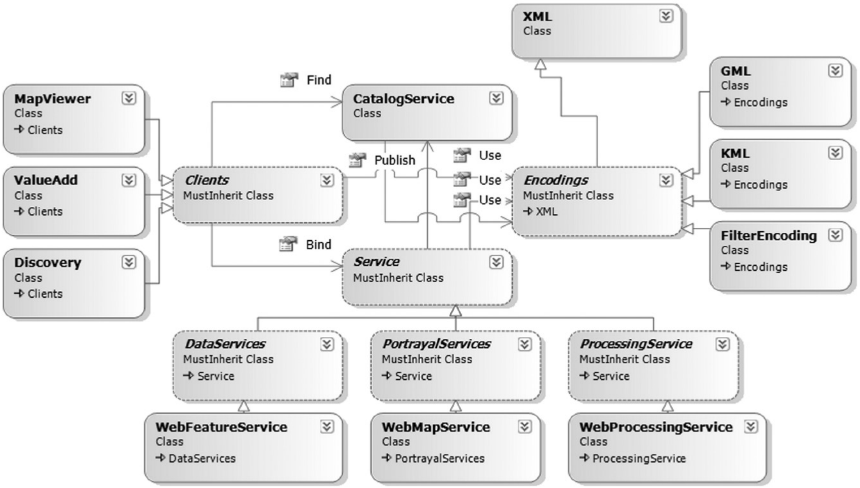


*Figure 10.1* Part of OGC Web Services framework (OWS).

of this framework can be implemented using various software technologies and systems. The most fundamental services and encodings of the OWS framework are Web Map Service (WMS), Web Feature Service (WFS), Web Processing Service (WPS) and Geography Markup Language (GML). The fundamental parts of the OWS are illustrated in Figure 10.1.

GML is an XML-based markup language that is used to encode information about real-world objects. GML has three main roles with respect to geospatial information. First, as an encoding for the transport of geospatial information from one system to another; second, as a modelling language for describing geospatial information types; and third, as a storage format for geospatial information (Lake 2005).

The WFS is the main Geoservice for publishing and requesting vector geospatial data in GML format. The WFS specification, also published as ISO 19142, allows a client to retrieve and update geospatial data encoded in GML from multiple Web Feature Services. A Basic WFS service implements three operations: GetCapabilities, DescribeFeatureType and GetFeature (see Figure 10.2). A client (usually a service or software) can request an XML-encoded capabilities document (containing the names of feature types that can be accessed via WFS service, the spatial reference system(s), the spatial extent of the data and information about the operations that are supported) by sending the GetCapabilities request to the WFS service. The GetCapabilities operation is required for any OGC Web service. The purpose of the GetCapabilities operation is to obtain service metadata, which is a machine-readable (and also human-readable) description of the server's information content and acceptable request parameter values.

The purpose of the DescribeFeatureType operation in the WFS standard is to retrieve an XML schema document with a description of the data structure (or schema) of the feature types served by that WFS service. The GetFeature operation allows for the retrieval of feature instances (with all or part of their attributes) as GML. A Transactional WFS (WFS-T) includes an optional Transaction operation to insert, update, or delete a feature (Vretanos 2010).

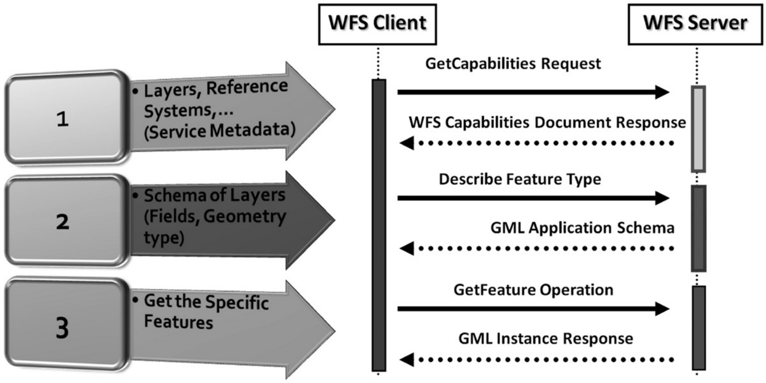


*Figure 10.2* Operations of WFS.

The WMS enables maps in graphical form to be delivered in response to queries from HTTP clients (de La Beaujardiere 2006). In the context of WMS, a map is a raster graphic picture of the data rather than the actual data itself. The WMS specification (ISO 19128), provides two mandatory operations (GetCapabilities and GetMap). The GetMap operation returns a map whose geospatial and dimensional parameters are specified in the GetMap request. The GetMap request allows the WMS client to specify distinct layers, the spatial reference system, the geographic area, and other parameters describing the returned map format (Figure 10.3).

The WPS is a Geoservice that enables the execution of computing processes and the retrieval of metadata describing their purpose and functionality. Typically, these processes combine raster, vector and/or coverage data with well-defined algorithms to produce new raster, vector and/or coverage information (Schut and Whiteside 2007). The WPS protocol supports both synchronous and asynchronous execution of processes. Synchronous execution may be used in simple and quick computation scenarios, where the data processing takes little to almost no time. Asynchronous processing is particularly well suited for complex computation scenarios, which may take significant time. In synchronous mode, the WPS must support GetCapabilities, DescribeProcess and Execute operations (see Figure 10.4). In addition to the mentioned operations, in asynchronous mode WPS must also implement GetStatus and GetResult operations.

Data exchange between WPS clients and servers requires an agreement on the general data exchange patterns and suitable communication protocols. Data may be sent to (and received from) a WPS server in two distinct ways: by reference (usually for large data sets) or by value (usually for atomic values or small data set). In contrast to the prior version, WPS 2.0 Interface Standard document provides a core conceptual model that may be used to specify a WPS in different architectures such as REST or SOAP (Mueller and Pross 2015). But in reality, no implementation details can be found about the support of REST or SOAP in the standard documentation.

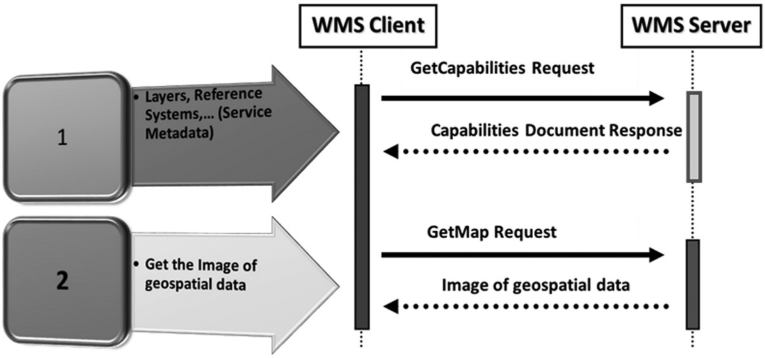


*Figure 10.3* Operations of WMS.

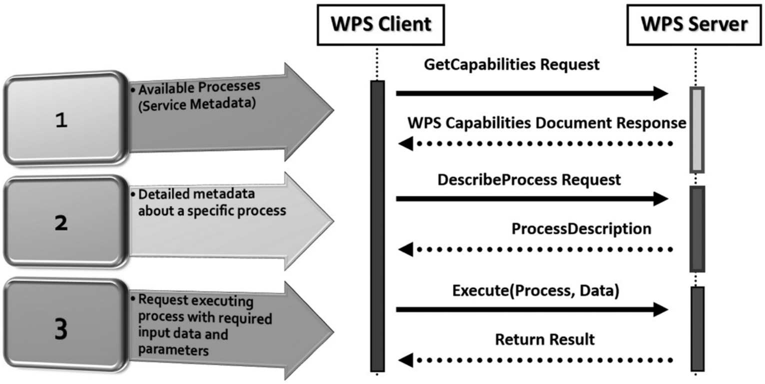


*Figure 10.4* Operations of WPS (synchronous mode).

## The need for integration of SOAP and REST services with OGC Web Services

Collaboration within and between organizations in a city through sharing data and analysis services is an important factor in the successful management of a city. While all types of services (Web Services, RESTful Services and Geoservices) detailed so far can provide the necessary interoperability, each type of service has been designed to address certain requirements.

Web services have several characteristics that make them one of the best and most efficient approaches for implementing enterprise applications in distributed environments and integration of legacy applications with current applications. In general, the most important characteristics of Web services are standardized contract publication, flexibility in composition, transmission protocol independency and platform neutrality. In order to maximize reusability, accessibility and interoperability the services in a smart city should provide functionality based on standard and discoverable interfaces. In other words, the service metadata of services or service contract must be published using platform neutral languages. In order to cover attributes of service, such as policy and business rules, security requirements, quality of service and supported operations of service, the service metadata must be based on flexible and comprehensive standards. In Web Services there are several specifications that provide the flexible and comprehensive foundation for disseminating a service contract, such as WSDL, WS-Policy and WS-Security. Since there might be various policies, agreements and security measures for data sharing within and between different organizations, this level flexibility is a must have feature. Also using a standard service contract, service discovery and potentially endpoint replacement is easier and can be done automatically using standard service registries.

Another important aspect of working with interoperable services in a city system is a capability to compose various services and chain their functionality to create repeatable and reproducible workflows. Based on SOP, services must be designed in a way that can participate as atomic and complex software entities in business workflows. The services need mechanisms for describing transactions, coordination and composition, as well as security and policies. Using Web services, WS-Transaction, WS-BPEL and WS-CDL are the most important specifications for providing such mechanisms in a standard and platform neutral manner. As mentioned previously, SOAP as a specification for transmitting messages between service consumers and providers is platform independent. This characteristic enables SOAP to provide high performance transmission (in internal secure networks using protocols like TCP) and interoperability with other enterprise applications over the Web (using HTTP protocol). Also messages in SOAP are self-contained. This means that they are intelligent enough to carry information about their processing steps, security and other aspects which are necessary in complex workflows. All these characteristics are in line with SOP. The above mentioned flexible and self-sufficient message-based communication is very important in data and analysis sharing within organizations in a city. Since different departments in an organization usually have access to an internal fast network, data and analysis within an organization can be provided using the fastest possible protocols. This might be also very useful in edge cases, such as disaster management in a city that needs the close collaboration of various organizations in a city in the fastest possible way. In addition, in complex scenarios and long-running jobs like batch processing of data using big data technologies, the features of Web services are very useful (Amirian *et al.* 2014).

REST services do not have the advanced capabilities of Web Services. Moreover, they do not have any means for supporting various security schemes, transactions, coordination, service composition and discovery. For all these features, RESTful services are usually limited to the capability of transport protocol (HTTP). The reason for this limitation is because they are intended for creating lightweight and easy to use services over the Web. Because of this simplicity RESTful services have been the de facto standard in Web and mobile application development in recent years. The most widely used online mapping services, such as Google and Microsoft (Bing) (which account for more than 95 per cent usage of online mapping services) provide their functionality through REST APIs. Most of time REST APIs and mapping services provide the foundation for city dashboards. In fact, for read-only and public exposition of resources (metadata, data and analysis), where there is no need for advanced features (like customized security, transaction control, composition, automatic discovery and coordination), REST services are the best approach. Moreover, because of simplicity and statelessness, RESTful services are highly scalable.

Unlike RESTful services and Web Services, Geoservices do not have strong support and popularity outside of geospatial community. Geoservices are not RESTful since they usually are not designed with architectural principles of REST (for example, they can use POST to get data or GET to update existing data) and they are not Web Services since in general they do not provide SOAP and WSDL bindings. However, like RESTful services, Geoservices are limited to HTTP.

Unlike RESTful services and Web Services, Geoservices provide a predefined set of requests/responses. In other words, based on Geoservices' specifications, the name of methods and input parameters are predefined by OGC. In contrast, developers can name the methods in RESTful services and Web Services freely. Unlike Web Services, in Geoservices there is no direct support for security and there is no specification for creating workflows and composition. Although there is some proposed solution for creating Geoservice chains and compositions in several research papers (Yue *et al.* 2011; Cruz *et al*. 2012; Weiser and Zipf 2007; Foerster *et al.* 2010), they are not applicable and efficient enough to be considered as a solution especially from performance and scalability for systems with a large number of users and huge volume of data (like cities) (Stollberg and Zipf 2008; Foerster *et al.* 2010).

In summary, Geoservices are important because they are designed and have widely been used for providing interoperability between various software systems in the geospatial community, but their non-conformance to trends and technologies in post-Web 2.0 era is a major issue constraining their wider popularity. There are some solutions for exposing the functionality of Geoservices using Web Service technologies, but they are not feasible in the real world. In most cases, especially when the response from a Geoservice is a XML document (like WFS), it is feasible to expose exactly same functionality of a Geoservice using a RESTful service or a Web Service (Erl *et al.* 2014). For some Geoservices it is difficult and complicated to provide the same functionality using a Web Service or a RESTful service. In complicated cases there is usually a need for a customized and non-standard solution. For example, the response of GetMap request in WMS is an image (in graphic formats like JPG or PNG formats) which is a binary file and so although Geoservices can suffer from limited capabilities and popularity (outside the geospatial community) they are essential in order to expose geospatial data and analysis to other systems.

Given that each class of service has various functionality, coupled with the fact that many legacy systems exist, we propose that for efficient and interoperable sharing of data and analysis all three types of services should be implemented by all organizations that are responsible for producing or updating data in a city which will ensure maximum usability of urban data. The next section provides a general architecture for this technical solution, which is called an organizational service layer.

## Organizational service layer

Figure 10.5 illustrates the organizational service layer (OSL). Ideally, each organization that is responsible for producing or maintaining data for a city should implement this layer in their IT infrastructure. Each organization that is responsible for building/maintaining a city infrastructure (like transportation, water and electricity) is responsible for producing digital data about that infrastructure. In order to share data and analysis in the most interoperable, flexible and secure way and in line with current Web trends, the OSL architecture suggests each organization in a smart city provide four types of bindings for their services. The bindings provide loose couplings between the backend implementation and its contract layer with other software applications. In other words, bindings in the OSL architecture separate the service implementation from service communication. By exposing functionality of a service using various bindings, service design, development, implementation and maintenance become more agile and straightforward. Another important advantage of these bindings is that they provide platform independency inside an organization. Based on the OSL, each department in an organization can provide different services and each service type should implement appropriate bindings based on the type of service and potential users (see Table 10.2). In general, bindings are defined by their underlying transport protocol, encoding, message format and security level as shown in Table 10.3.

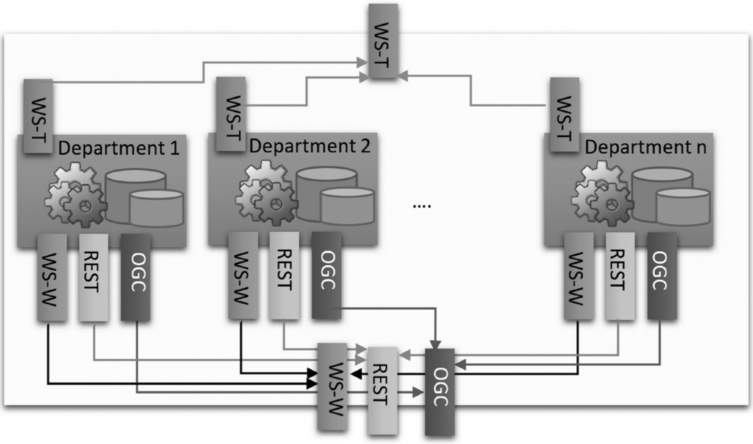


*Figure 10.5* Organizational Service Layer in an organization.

*Table 10.2* Potential users and client applications for various service types in a city

| *Service type* | *Potential client applications* | *Potential users* |
|---|---|---|
| **REST** | Web applications (like city dashboard), connected mobile applications | Citizen data scientists, researchers, citizens, developers |
| **WS-W** | Enterprise applications (for running long and/or complex jobs like service composition, communication with a big data technology) outside of an organization | Developers outside the organization |
| **WS-T** | Other software applications inside an organization (for sharing data and analysis in fastest possible way) | Developers inside the organization |
| **OGC** | GIS applications (mostly Desktop GIS applications) | GIS developers and GIS experts |

*Table 10.3* Details about various binding types

| *Service type* | *Transport* | *Encoding* | *Message format* | *Discovery/ service contract* | *Security* |
|---|---|---|---|---|---|
| **REST** | HTTP | Text | JSON, XML | HTTP OPTION /documentation | Transport level |
| **WS-W** | HTTP | Text, MTOM | SOAP | Service contract | Message level and Transport level |
| **WS-T** | TCP | Binary | SOAP | Service contract | Message level and Transport level |
| **OGC** | HTTP | TEXT, Image | GML, KML, JPG, PNG, . . . | Capabilities document | Transport level |

Since Web Services support advanced features (transaction support, message level security, service composition and coordination), services can use WS-W and WS-T endpoints in complex scenarios that need those features (for access to sensitive data, updating data, inserting new data or calling services to run complex processing jobs). The complex scenarios might be a long-running processing job using single service (for example, batch processing of a huge amount of data using a big data technology or executing a workflow based on composition of several services). In this context, WS-W services can be consumed (called) by the users outside of the organizations using Web or be utilized as standard communication point for push services. For users inside the organization the WS-T provides the fastest possible communication speed. Also WS-T can be consumed by other organizations in a city if they have an appropriate service level agreement. Developing applications using WS-W and WS-T bindings is easier for professional developers (or enterprise developers) because the services have a machine-readable contract and creating consumer applications (proxy classes) is almost automatic using integrated development environments.

REST endpoints can be used for accessing data and analysis which should be publicly available. The REST services are usually consumed for developing Web 2.0 applications and connected mobile applications (mobile apps that need to be always connected to the internet). A good example of such Web applications is a city dashboard which shows various aggregated metrics, reports and indicators from several organizations. The aim of city dashboard is to provide interactive data about all aspects of a city (see Chapter 9). A city dashboard can be used by citizens and companies to better understand city and by city managers to grasp the dynamics of the city and to monitor the progress of city projects. Also data scientists (in both academia and industry) can use the RESTful services to access public data about the city. Finally, OGC Web services which are very popular in the geospatial data community should be implemented using OGC binding to provide the same functionality of WS-W and RESTful services (and in accordance with OGC specifications) be consumed by mostly desktop applications in the geospatial community.

## Conclusion

In this chapter an ideal architecture for sharing data and analysis in a smart city has been proposed. The architecture intends to provide flexibility, scalability, accessibility and interoperability at the same time. Since no technology stack or standard stack provides the requirement of this architecture, we have advocated the use of a polyglot binding for implementing services contracts and service orientation for the design of services. Using polyglot binding and service orientation, functionality of a service within and between organizations can be invoked and composed in a flexible, interoperable and scalable manner. By implementing the OSL architecture, implementation of city dashboards, city data portals and city service hubs will be more agile.

Since the architecture provides flexible techniques for data and analysis sharing, communications within organizations, between organizations and between citizens and organizations can be improved. This is an important requirement for cities and citizens. With the rise of citizen data scientist and the corporate use of urban data, sharing data about cities using various bindings is advantageous. The architecture also can efficiently communicate with big data technologies. Current city dashboards are useful tool for now-casting. With big data technologies and data science, cities need to have other systems for sharing data using polyglot bindings and providing indicators and metrics about city for future (forecasting) using predictive analytics.

## References

Amirian, P., Alesheikh, A. and Bassiri, A. (2010a) ‘Standards-based, interoperable services for accessing urban services data for the city of Tehran’, *Computers, Environment and Urban Systems* 34(4): 309–321.

Amirian, P., Alesheikh, A. and Bassiri, A. (2010b) ‘Interoperable exchange and share of urban services data through geospatial services and XML database’, in *CISIS 2010 – The 4th International Conference on Complex, Intelligent and Software Intensive Systems*, pp. 62–68.

Amirian, P., Basiri, A. and Winstanley, A. (2014) ‘Evaluation of data management systems for geospatial big data’, *Computational Science and Its Applications – ICCSA 2014*, pp. 678–690.

Barry, D.K. (2003) *Web Services, Service-Oriented Architectures, and Cloud Computing*. Burlington, MA: Morgan Kaufmann.

Booth, D., Haas, H., McCabe, F., Newcomer, E., Champion, M., Ferris, C. and Orchard, D. (2004) *Web Services Architecture*. Available from: www.w3.org/TR/ws-arch/ [accessed 10 February 2017].

Cruz, S.A.B., Monteiro, A.M.V. and Santos, R. (2012) ‘Automated geospatial web services composition based on geodata quality requirements’, *Computers & Geosciences* 47: 60–74.

Daigneau, R. (2011) *Service Design Patterns: Fundamental Design Solutions for SOAP/WSDL and Restful Web Services*. Harlow: Addison-Wesley.

de La Beaujardiere, J. (2006) ‘OpenGIS® Web map server implementation specification’, *Open Geospatial Consortium Inc., OGC*, pp. 6–42.

Erl, T. (2008) *SOA Design Patterns*. Harlow: Pearson Education.

Erl, T., Gee, C., Chelliah, P., Kress, J., Normann, H., Maier, B., Shuster, L., Trops, B., Winterberg, T., Utschig, C. and Wik, P. (2014) *Next Generation SOA: A Concise Introduction to Service Technology & Service-Orientation*. Harlow: Pearson Education.

Erl, T., Puttini, R. and Mahmood, Z. (2013) *Cloud Computing: Concepts, Technology, & Architecture*. Harlow: Pearson Education.

Fielding, R.T. (2000) 'Architectural styles and the design of network-based software architectures', PhD dissertation, University of California, Irvine. Available from: www.ics.uci.edu/~fielding/pubs/dissertation/top.htm [accessed 10 February 2017].

Foerster, T., Schäffer, B., Baranski, B. and Brauner, J. (2010) 'Geospatial web services for distributed processing: Applications and scenarios', *Geospatial Web Services: Advances in Information Interoperability*, pp. 245–286.

Lake, R. (2005) 'The application of Geography Markup Language (GML) to the geological sciences', *Computers & Geosciences* 31(9): 1081–1094.

Mueller, M. and Pross, B. (2015) *OGC WPS 2.0 Interface Standard.* Available from: https://portal.opengeospatial.org/files/59944 [accessed 10 February 2017].

Percivall, G., Reed, C., Leinenweber, L., Tucker, C. and Cary, T. (2003) 'OGC reference model', *Open Geospatial Consortium Inc*, pp. 1–108.

Sample, J.T., Shaw, K., Tu, S. and Abdelguerfi, M. (2008) *Geospatial Services and Applications for the Internet*. Berlin: Springer Science & Business Media.

Schut, P. and Whiteside, A. (2007) 'OpenGIS Web Processing Service', *OGC Project Document*. Available from: http://xml.coverpages.org/OGC-05-007r7-WebProcessingService-WPS-V100.pdf [accessed 10 February 2017].

Stollberg, B. and Zipf, A. (2008) 'Geoprocessing services for spatial decision support in the domain of housing market analyses', in *Proceedings 11th AGILE Conference on GI Science*.

Vretanos, P.A. (2010) 'OpenGIS Web Feature Service 2.0 Interface Standard', OGC Implementation Specification 09-025r1, Open Geospatial Consortium, Wayland, MA.

Weiser, A. and Zipf, A. (2007) 'Web service orchestration of OGC Web services for disaster management', in *Geomatics Solutions for Disaster Management*. Berlin: Springer, pp. 239–254.

Yue, P., Wei, Y., Di, L., He, L., Gong, J. and Zhang, L. (2011) 'Sharing geospatial provenance in a service-oriented environment', *Computers, Environment and Urban Systems* 35(4): 333–343.

Zimmermann, A., Pretz, M., Zimmermann, G., Firesmith, D.G. and Petrov, I. (2013) 'Towards service-oriented enterprise architectures for big data applications in the cloud', in *2013 17th IEEE International Enterprise Distributed Object Computing Conference Workshops*, pp. 130–135.

# 11 Blockchain city

## Economic, social and cognitive ledgers

*Chris Speed, Deborah Maxwell and Larissa Pschetz*

## Introduction

City dashboards are typically representations of a city's accounts, manifest according to values set by the stakeholders. The currency of the data within a dashboard is typically reduced to an assessment of the performance of services (such as traffic flows and crime statistics) largely derived from quantitative sources. Whilst such databases may be useful for mayors to report on the performance of a local government, or to set targets that lead to penalties or bonuses, the city workers and inhabitants that are complicit in the production of data are rarely aware of the nature of how data are collected or the 'ledger' that they are contributing to. As a consequence, dashboards cannot describe many of the transactions that take place between people, nor can they make explicit the values that are brokered between the myriad of city occupants.

This chapter explores different perspectives upon economic and socio/geographical ledgers and the complexity that they involve as they inevitably collide with concepts of chronological time, representation and actions. Three means of approaching the concept and practice of the ledger are discussed: (1) *money, time and the blockchain*: an exploration of how the representation of money shifts from material representation within fiat currencies (i.e. those underpinned by governments or precious metals) to the blockchain, the sealed distributed ledger that supports the Bitcoin cryptocurrency; (2) *city as ledger*: a recovery of the role of time in the production of economic geographies with a focus upon Hägerstrand's approach to time-geography that accounted for personal and group actions within temporal and spatial frames, and inevitably a recovery of Marx and the obfuscation of histories and geographies; and (3) *cognitive and practice-based ledgers*: an introduction of the use of filmic storytelling as a cognitive ledger using the Dardennes' film *Two Days and One Night*.

These three theoretical perspectives on ledgers set the scene for two prototypes utilizing Bitcoin technology that emerged from a design workshop facilitated by the Design Informatics studio, University of Edinburgh. These prototypes begin to explore temporal and social potential for using ledgers within design experiences. By reflecting on the role of ledgers across different forms, this formative chapter establishes the complexity of capturing and producing data across a myriad of social practices using linear systems.

## Ledger 1: money, time and the blockchain

There are many elements that make Bitcoin an interesting alternative currency, but critically it is the development and implementation of the blockchain – a distributed ledger that contains all transaction records ever conducted. The Bitcoin blockchain is an encrypted, cumulative ledger composed of 'blocks' of transactions that are verified by miners and which lead back to the first 'Genesis' block whose instance is timed as 18:15:05 GMT, on 3 January 2009, signifying the start of the currency. Blocks can contain the social, economic and geographic information about the senders and receivers of Bitcoin wallets, time of transaction, amount of Bitcoins being transferred, fees and IP addresses from which location can also be identified. Transaction blocks are generated approximately every 10 minutes, a timing that is calibrated by the network – if blocks are completed quicker, the difficulty of the mining is increased, and vice versa. Each new block (and not a huge single list) provides an opportunity for transactions to be verified and thus takes place within a reasonable and anticipated amount of time – in many ways forming both the 'tick' and the check-sum of the platform. This process is verified by miners who compete to complete 'proof of work' functions, that is, computationally intensive algorithms, to check if every block that follows is legitimate (Maurer *et al.* 2013: 264). In addition, each new block essentially concatenates the previous block with the new set of transactions, creating the *chain*, which leads all the way back to the initial Genesis block. Once mined, the block is sealed and, currently, 12.5 Bitcoins are released as a reward to the winning miners, thus incentivizing the expensive mining activity and steadily populating the peer network with more currency.

This linear association, connecting one block to the next through the integrity of the encrypted mathematical codes, keeps the chain intact, and, along with the massively distributed, multiple copies of the currency system, helps to prevent fraud. The linear, cumulative nature of this system is of particular interest to the authors, and in particular how this differs from current, centrally controlled, fiat currencies that regulate the release and removal of money in the system (physical and virtual) to attempt to manage the market.

In stark comparison to the blockchain, fiat currencies are released as promissory coins, notes, mortgages or loans according to an assessment of how much money there should be within a society according to the values of that particular economic system. Monetary representation has become increasingly abstracted from the goods and services that it can be used to trade in, and this is central to Marx's concern for how value has become commodified, not in what is needed but what is desired. Since originating in the bartering of actual goods and tools such as animal skins, salt and weapons, for a long time the physical representation or tokenization of a currency corresponded with the goods being purchased.

The form that money takes, and its association with the value of the minerals that it is either made of or is connected to, has become increasingly slippery (Maurer 2006: 27). In 1816, the Bank of England changed the basis of English money from silver to gold through the Great Recoinage and at this point the value of silver in a silver coin was less than its representational value, and so coin transformed into a token. As global trade required 'modern' organization through the early part of the twentieth century, the Bretton Woods agreement was signed in 1944 by committed countries in order to maintain exchange rates to a fixed value in terms of gold. On its failure in 1971 – due to the dollar's inability to retain value in the light of a global recession – the detachment of monetary value from a mineral ore to a new system of floating exchange rates 'de-materialized' money (Harvey 1990). As the representation of value continues to become further abstracted from goods and services, for example, through electronic BACS transfers and online and mobile banking, we soon arrive at the role of money in society today.

In the abstraction of value from a material representation to a promissory token, both time and identity become obfuscated. Although the jurisdiction of English bills was encoded in such a way to manage the spatiality of economics, it mattered not who the bearer of the coin or note was and when it was exchanged. Once released into a system, the use of individual monies was not monitored or tracked – only the health of the system. In this way, there are significant differences with Bitcoin and its reliance on a blockchain. Given the nature of digital systems, perfect copies of money are conceptually even easier to make than the counterfeiting of physical money. The radical invention of the blockchain uses multiple copies of a single ledger distributed across a network to deal with the 'double spending' potential of digital money, that is, duplicating currency and spending it twice or more, is a central feature to the Bitcoin platform. In fiat currencies, third parties, for example, banks, balance the books at the close of each trading day. In Bitcoin, 'double spend' is prevented by ensuring digital scarcity through the verification of transactions through the mining process and transaction blocks.

These differences represent two entirely different models of time for each form of currency. The inflation and deflation of prices, the savings and overall growth within the system vary according to market values, goods or purchasing power of the currency and trading with foreign currencies. In this case, time is suspended and does not offer a metric through which individual transactions can be recorded.

The 'minting' of the Bitcoin currency is bound to the ledger that records the spend of the currency, resulting in a close relationship between time, value and ultimately power, that is, as time progresses the reward for mining depreciates and demand for more computing power increases. Furthermore, the algorithm knows there is a finite maximum amount of Bitcoins that can be created, a figure of 21 million. These are released at a fixed amount that halves every four years and are issued to winning miners who validate each transaction block. As the computing power has increased in line with the complexity of the maths, the distribution of miners has shifted from a distributed global community toward four mining companies based in China owning almost 70 per cent of the activity. Unlike the anonymous accounting of people spending the same material money over and over again (e.g. coins or notes) in the cash registers of disconnected shops, the spending of Bitcoins is inscribed in the blockchain and forever associated with specific transactions within a distributed network. The sealed, distributed nature of the blockchain means that the integrity of the currency is reliant on a linear model that looks back, before it generates money forward (DuPont and Maurer 2015). In contrast, fiat currencies project money forward and balance their books retrospectively according to the performance of spending across a system. Compared to the speed of the blockchain in checking the integrity of its system (approximately two hours), it is rumoured that it takes 58 days for the UK civil service to understand its GDP in any given month.

## Ledger 2: city as ledger

The introduction of the clock into mediaeval society was connected to the management of land and was closely tied to both the development of the written word and the use of ledgers to account for the production and trade. From an era when 'natural rhythms dictate the pace of life and work and the content of language', and any expectation of a future 'centres on a short lifespan and the imminence of the Day of Judgement' (Thrift 1996: 180), Thrift draws attention to the influence that writing technologies have upon our sense of time. Originating in the technology of the written word, Thrift argued that the linear process of writing and its evidence in the form of texts, revealed a 'consciousness of time past' (Thrift 1996: 180). This in turn informed time present, which became ripe for reorganizing, and consequently daily events became accountable, and inevitably associated with monetary values.

At the time, power of the controlling the ledger was in the hands of a new generation of literate monks and members of the King's Court who gradually began cataloguing the use of the land as a means of calculating profit and eventually monitoring performance and efficiency. 'Thus, financial accounts may now seem the most obvious way of stating time as money' (Thrift 1996: 184). Clocks were the next step toward a synchronization between the church day and the individual, instead of responding to the church bell, people could be organized in to more specific blocks of time. The term 'organized' is used because the owner of the clock is the one with power, and, as Harvey reveals, 'such time discipline crucially depended upon the construction of distinctive spaces of surveillance' (Harvey 1996: 225). Consequently, it is not long before a recognizable 'modern' system is in place.

In the late sixties and early seventies, Lund University in Sweden established important relationships between time and geography. Amongst many, Hägerstrand worked hard at eroding the 'compositional' view of the world that most social scientists were using to talk about how people make sense of space. Taking a situation as a snapshot, and seeing it as a complex construction of 'objects' that are acting upon one another, Hägerstrand suggested that the compositional view could only deal with context and was constrained by establishing fixed relations between artefacts, preventing the opportunity for movement and change. Whilst it aspired to objectivity, the compositional approach lacked point of view and subjects became items. In contrast, the time-geographic approach attempts 'to capture the complexity of interaction at the scale of the smallest indivisible unit which for human population is, of course, the individual' (Parkes and Thrift 1980: 244).

Much of the time-geography work concentrated upon developing methods of describing people and their journeys through space in time. Inevitably the work identified a linear series of events that make up a person's day, suggesting that the nature of these events (called a project) motivated individuals to move through space to see them fulfilled. The subsequent documentation of these various 'projects' formed a ledger of a group's activities and could be analysed to understand social, spatial and temporal relations.

Criticism of the time-geographic approach has since been targeted at the apparently linear, and indeed Cartesian, approach of conceptualizing time and space, although the Lund school argued that they merely absorbed given models in order to make their point clear. Another problem is the lack of focus that the models show for dealing with a psychological conception of space since the Cartesian parameters dominate the representation of events and places. Finally, and perhaps the most interesting problem with the time-geography model for considering the city as a database, was the participants' honesty (and apparent lack of interest) in what happens when the linear paths cross and cause conflicts in the completion of the tasks. The spatial and temporal ledger of social practices that Hägerstrand constructs provides a valuable insight into personal and group activities, offering a chance to reflect and consider how individual projects are connected to partners, groups and communities.

As databases across the city develop, Hägerstrand's time/space projections of the interweaving of social and material relations is to some extent possible as the potential of machine learning promises to uncover more and more correlations between datasets. A question remains though to the extent to which the city wants to reveal all of its social and geographical relations.

Constructing a form of digital ledger for geographical and social relations, Ian Cook's 'Follow the Things' project provides insight and discussion into the background of consumer products from food items to clothes, and electricals to health and beauty products (http://www.followthethings.com) (Cook 2015). For instance, the collection of reports on celebrity perfumes that was authored by Gethin Chamberlain and originally published in the *Guardian* in 2010.[1] The article primarily focuses upon the poor working conditions and pay of Indian employees of the Pragati company for packaging celebrity branded perfumes including Katie Price and Jade Goody. The financial markups of individual bottles retailed for £19.99 in UK pharmacies, whilst the average take-home for its 7,000 employees was a low as £2.05 per day. Follow The Things therefore becomes a form of socio-geographical ledger for a wide range of products, and supporting discussion and debate to better understand the 'veil' that is placed over desired artefacts that obfuscates their histories and boosts their economic value.

## Ledger 3: cognitive and practice-based ledgers

Beyond the various forms of metric ledgers that record financial transactions, citizens, times and space, cultures are underpinned by shared stories told through a series of passages, chapters and accounts. Within literate society, books and films share a great deal in common with ledgers as they rely upon a linear time base across which a story or an account is experienced by the recipient. Limited to only articulate one line of activity at a time, the author shapes the reader's experience by moving in and out of the activities of characters to reveal and hide circumstances that sustain the narrative.

However, the reading of a book or watching of a film constructs a 'cognitive ledger' in the mind of the reader who navigates the narrative to develop an individual understanding of the meaning and consequences of each of the interactions. In the case of a film taking place over a 90 to 120-minute period, the management of the director and the skill of the editor can offer a highly compelling experience whereby the audience moves through a visual ledger that suspends reality and invests them the actions and affairs of the onscreen characters.

One explicit manifestation of such a cognitive ledger (and perhaps closer to our theme) is the 2014 film *Two Days, One Night* by the Dardenne brothers. The film follows Sandra Bya, a working mother who returns to work on a Friday following a nervous breakdown to find that her 16 workmates have voted to take a €1,000 bonus in place of her job. Supported by her husband and her workmate Juliette, she lobbies her boss to ballot the workers again on Monday morning. Successful in her appeal, she has the weekend to canvass each of the workers at their homes in an attempt to persuade them to change their mind. Following 16 meetings on door steps, back gardens, launderettes and street corners, Sandra returns to work on Monday morning for the new ballot to understand her fate.

The film plays out as a ledger of interactions in which the audience develops a running balance of those who would prefer to keep Sandra, and those who would prefer to take the money. However, the brief insights into each co-worker's lives describe complex personal circumstances which in the mind of the viewer complicates the running total as to whether Sandra deserves to keep her job. However, Sandra's activities between negotiations are as interesting as the co-workers' lives and values that are represented through their partners, children and living conditions. Dominated through acts of consumption, we develop an understanding of Sandra and her family's economic disposition through her drinking of bottled water, take-away pizzas, eating of ice creams and purchase of artisan bread, whilst her mental state is portrayed through an attempted suicide as her encounters (positive and negative) all challenge her sense of identity. In the end, the audience is left divided according to how they balance the books between the welfare of Sandra and her co-workers, alongside wider politics of fairness.

In an interview with Larry Rohter for the *New York Times* (2014), the directors discuss the influence of a study by Michel Pialoux that became part of Pierre Bourdieu's edited book *The Weight of the World: Social Suffering in Contemporary Society* (1999). The book, a collection of studies that read like short stories, provides insight into the lives of a range of people whose lifestyles were affected and disrupted through the inequalities, politics and determinism of late twentieth-century economics. The Pialoux study, entitled 'The old worker and the new plant', reflects on a conversation between the author and two employees of the Sochaux Peugeot plant in Haute-Saône, a French department of the Franche-Comté region. For the Jean-Pierre Dardenne, the experiences of Gérard and Christophe became of particular interest: 'The book had probably 15 case studies and 15 analyses, and one of these stories was a worker cast aside because of the influence of managers, who got the other workers to agree to push him aside. This worker was probably a little less productive at his job, and therefore that team was never getting its bonuses.'

In many ways, the workers' experiences are situated in a particular epoch of transition for the automobile industry as linear car production began to struggle, and companies looked to Japan for a solution. The result was a move from Henry Ford's never-ending production line as a linear production ledger to the Toyota model in which production contained a reflexivity much closer how we might understand the blockchain; that is, how people become part of these systems and could develop practices within them.

In 1970, Toyota launched the Toyota Production System (TPS), a method that managed car manufacture and employees more effectively than the failing Fordist model, which had struggled in 1950s and 1960s in Japan. 'Just-In-Time' was the title of the manufacturing and conveyance model that informed the demand of car parts in terms of which part was needed, when it was needed, and how many were required. Just-In-Time used a Toyota model for time 'Takt-Time' that was used to monitor the production time against the volume required (Ohno 1995: 29). Coupled with 'Jidoka', a term referring to the ability to quickly stop and modify production lines if problems arise, TPS became a prime example of post-Fordist production models, and one that enabled Toyota to respond to consumer demands. Through TPS, both supply and quality were monitored constantly and allowed the company to build cars in such a way that consumers felt they had more control and individual choice, as colours and specifications could be relayed from the showroom to the factory (Ohno 1995: 30).

The transition from the traditional Fordist production lines to those influenced by TPS has not been easy for many Western manufacturers. Intrinsic differences including the role of unions and the speed of technological change made it difficult to change from old methods to new practices. Returning to the worker experiences at Peugeot, there is a genuine conflict in the power relations within the new teams that developed as the plant adapted to new manufacturing models.

Whilst the Dardennes' screenplay uses cinema to construct a cognitive ledger in the mind of the viewer, Pialoux's study of the worker experiences at the Peugeot plant describe the impact of more complex models of car manufacture as the production becomes part of a ledger of actions and check sums. TPS revolutionized Toyota, and subsequently other car manufacturers such as Peugeot and GM, and the empowerment of the teams within production areas to strive for quality over quantity has constructed a form of blockchain, as the control over the production line acts as a calibration within the system to ensure that mistakes are not passed on down the line.

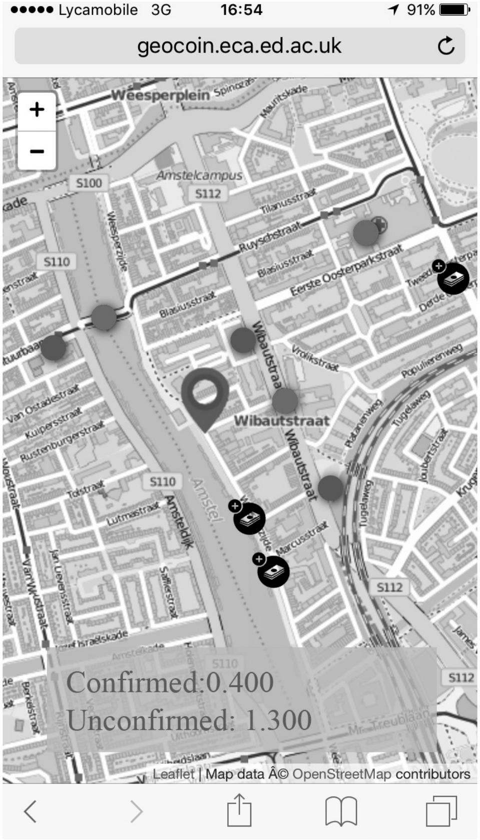
## Designing with ledgers: a design case study

In April 2016 the authors and the Centre for Design Informatics were invited to develop a 48-hour workshop for Martyn de Waal's 'Design & The City' programme in Amsterdam. The workshop entitled 'Blockchain City' was intended to expose 'contemporary design methodologies, and their relationship with living labs and smart cities'. A software platform was developed to enable design solutions from participants' exposure to the principles of ledgers as trust platforms, and programmable money such as Bitcoin. The simple premise was that whilst blockchain and the functions of Bitcoin remain abstract for many people, developing a platform to allow physical engagement would actualize characteristics of the technology, as well as lead to critical applications for social and or urban contexts.

In its own words, 'Design & The City' explored citizen-centred design approaches for the smart city. Central themes were the role of design to create opportunities and practices for citizens, (social) entrepreneurs and policymakers towards more liveable, sustainable and sociable urban futures. The workshop was located along the Amsterdam Knowledge Mile that runs from the Amstelplein to the Nieuwmarkt. The mile represents Amsterdam's digital economic initiative involving 'universities of applied sciences, citizens, municipality, organizations and companies to form an applied research ecosystem to develop, test and display smart solutions for metropolitan challenges in the area'. In this way, the social and economic context in which the lab was located was intended to stimulate ideas and reflections within the workshop. Physically located within the commercial Student Hotel, previously home to the editorial offices of some of the country's most important newspapers, Blockchain City aimed to engage participants with a location-based software platform that encouraged them to associate acts of trust in the local community with Bitcoin transactions.

A key stage in scaffolding the participants' ability to design new social economic experiences was the development of a piece of software called 'GeoCoin' that served as an introduction to what programmable currencies could offer in a technical sense, but also allow participants to test them in an urban context to support the development of their own ideas. 'GeoCoin' was a mobile application run from a web browser that used location information to pinpoint the participant within a map of Nieuwmarkt area of Amsterdam. Using the Bitcoin client Electrum, we were able to associate geofences (GPS locations) with transactional functions. On the map the participant was also able to see three types of icons: small bags of money scattered across the area, red hot spots and green hot spots (see Figure 11.1). In the bottom left corner of the screen two numerical amounts appeared preceded by the terms: Confirmed and Unconfirmed. Without further information, participants were asked to leave the workshop studio and venture out in to the surrounding area to discover what the three icons and the numerical values would do as they approached them.

Day 1: The structure of day 1 of the workshop was very simple, consisting of an introductory talk followed by a participatory exercise 'Block Exchange' that was developed by the Centre for Design Informatics that uses Lego to introduce the principles of the blockchain. More importantly however, it serves to dismantle cultural expectations between the representation of value, and the values that currency can potentially represent if we consider the role of a distributed ledger. Following a conversation over lunch regarding what the Lego activity had revealed we introduced the beta version of the GeoCoin software that would become the starting point for participants to redirect its purpose. Aware that designing applications that use a new representation of value (Bitcoin) is rather challenging, our software provided participants with a head-start toward designing critical applications based upon insights from the Block Exchange exercise. Introducing the software meant going outside and experiencing its functionality; following this, participants spent the remainder of the afternoon coming up with new applications of the software. At the end of first day the software was adapted overnight ready for testing on the second day.



Confirmed:0.400
Unconfirmed: 1.300

*Figure 11.1* Smartphone screenshot of the GeoCoins software featuring bags of coins, and red and green GPS hotspots.

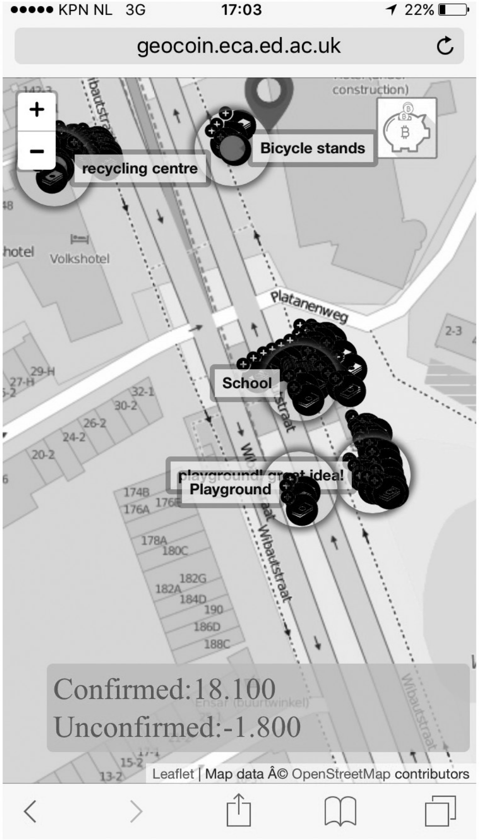
Day 2: Whilst Hadi (remote project software developer) continued to develop the two iterations of the platform on the morning of the second day, participants self-organized into two teams and were asked to develop very short explanatory videos that introduced, contextualized and demonstrated their new applications. By early afternoon the teams were able to test and refine the software with Hadi remaining online to troubleshoot bugs. The teams presented their videos during a short presentation and summary of the Blockchain City Lab during a reflective evening event.

Once outside, it became relatively clear to people that the small bags of money would disappear when a participant's location correlated with the GPS coordinates of an icon, and within moments the unconfirmed number would increase on their screen. On returning to the studio participants described their interpretations of how the red and green hot spots worked, and why Unconfirmed and Confirmed numbers fluctuated. Many had guessed that we had used a digital currency such as Bitcoin and distributed fractions of them across the landscape. Less easy to understand, because there was no instant feedback from the icons (unlike the bags of money that disappeared as you walked over them), the group began to realize that if their location corresponded with the GPS coordinates of a red hotspot then their Unconfirmed numbers would decrease, and that if they stood on a green hotspot their Unconfirmed numbers would increase. Whilst these elements were relatively easy to understand, the question of why numbers across the Unconfirmed and Confirmed lines fluctuated was less comprehensible. The difference in the two variables was explained as being the time it took for the blockchain to ratify a transaction within a block. At this point the value of experiencing the time between Unconfirmed and Confirmed transactions began to expose some of the characteristics of a currency that requires confirmation through an entire digital network. Body storming (Schleicher *et al.* 2010) the type of transactions that a programmable currency such as Bitcoin offers was an important step in supporting participants towards the design of their own derivations of the software, and based upon the results, the ability to perform economic software within an urban landscape informed both the conceptual development of ideas but also the representations of their work.

Following their forays into the local area surrounding the student hotel, the six participants formed two groups of three people, and began developing responses to both the Block Exchange Lego workshop and their experiences of the GeoCoin software. The two ideas that emerged corresponded to the ideas and values evoked during the Block Exchange workshop. During the final round of the Lego workshop participants are invited to trade anything that they desire as long as it can be valued by somebody else and written down on to the ledger. As facilitators we wrote down the subject of these exchanges because they tended to follow a pattern of participants realizing that they can trade anything as long as the ledger is trusted. The pattern follows that people begin with trading material goods, then they realize that they can trade services which tend to become increasingly outlandish, before settling down to trade services that are for the common good. This workshop was no exception with the teams moving from trading pens (that were vital to write transactions into the ledger and were therefore rare assets), through providing service such as kisses, singing songs for people, before finishing with Lego blocks that are required by everybody and finally a common fund. This pattern from materialist desires toward social projects reoccurs as participants place increasing faith in the trusted ledger, and for the two groups it provided the stimulus for two distinct iterations of the GeoCoin software.

### *Civic Blocks (Project Team: Dorota Kamrowska-Zaluska, Hanna Obracht-Prondzynska, Eileen Wagner)*

Civic Blocks transposed the value of a fraction of a Bitcoin into a vote for how a City Council should spend a proportion of its budget. The team suggested that a City Council could convert a proportion of its capital resource budget into Bitcoin, perhaps 10 per cent. Using the unique capabilities of Bitcoins to divide them into



Confirmed:18.100
Unconfirmed:-1.800

*Figure 11.2* Screenshot taken from smartphone displaying the Civic Blocks software in use. The position of the user is denoted by the marker who is spending their vote/coins on a bicycle rack project.

fractions, 10 per cent of the budget would be distributed to all citizens of a city that are eligible to vote. Citizens are then invited to spend their vote/coins by dropping them at locations generated by fellow residents including proposals for spending council monies on schools, parks and roads, or they can choose to generate their own spending project by creating a new geofence and naming it with their own cause. With the GPS coordinates, name of the project and the value of accumulated coins/votes inscribed in to the blockchain, council monies are locked into particular projects. Through the technical support of Hadi, the team were able to design a fully working prototype that allowed workshop attendees to spend their votes on social projects in the local area (see Figure 11.2). The team also produced a short video explaining the principles of the platform: https://vimeo.com/163760240.

### *HandFastr (Project Team: Corina Angheloiu, Max Dovey, James Stewart)*

The second group became very interested in the potential for the blockchain to record smart contracts that could reconfigure social pledges and transform spending powers. Adopting marriage as a social contract, the team designed a mechanism to support social economic bonds in the form of temporary mobile agreements using smartphones. As explained by Max Dovey, a member of the team:

> Marriage, with all its connotations, can be whittled down to one of the oldest forms of contract that binds two people from two families to create financial security. Arranged marriages, short-term fixed marriages or visa weddings all utilize the contract to secure wealth, security or freedom between different parties. We adapted the practical and functional aspects of marriage into the GeoCoin platform to enable impromptu financial commitments between people in public space.

*Figure 11.3* Still from the Handfastr video developed by participants to describe how their prototype software allows people to form temporary smart contracts for shared banking and spending.

Through negotiation with Hadi, a platform was developed that placed geofences in the vicinity of the workshop that when consenting participants agreed to ‘get married’, the software would transfer Bitcoins that were previously held in separate wallets, into a conjoined wallet. As long as the partners (can be any number) remained married, they could only spend the currency when they were in the same GPS location. The team also produced a short video to introduce the platform: https://vimeo.com/163565402 (see Figure 11.3).

## Conclusion

The database city is entirely based upon a multitude of ledgers – all owned by different parties, and all constructed to account for different transactions. Registering a database is as easy as buying a book, from signing up for Facebook to installing the MySQL databases that sit behind our personal blogs and local government Customer Relationship Management systems. All of them provide an account of social, spatial and economic interactions. However, rarely do they describe the city – that complex, messy, contested environment that is completed every night for some participants, and falls apart for others. It is impossible to escape the ledger as we are entangled in the accounting of ourselves, our friends and strangers. This chapter has set out to explore the experiences of being in a culture in which the ledger has become an intrinsic, if generally unrecognized, part of the data city. Not new, but now networked, ledgers offer to some extent the potential to resolve the cities greatest problems of complexity – prediction and instability. By looking for patterns within databases and building feedback loops to support public involvement in managing the city, the ledger is the history and the future of the city.

The chapter used three perspectives on the historical use of ledgers in the accounting, mediating and representation of value in order to better understand how the linear inscription of transactions forms is a habitual characteristic of social and economic practices. Through an introduction to the workings of the digital currency Bitcoin and the nature of the blockchain, the chapter explored the relationship between time and money. In particular, attention was drawn to the intrinsic and immutable association between bitcoins and the transactions that they are written into the blockchain, unlike material coins that become a proxy for value, and remain independent of the accounts that describe what they were used to purchase.

The growth of writing through the church, and in particular the accounting of labour and goods was used to demonstrate the close association between the documentation of time and the space. Hägerstrand’s time-geography explicitly used temporal and spatial ledgers to map the simultaneous activities of individuals within groups. The disaggregation of individual practices that are carried out through ‘projects’ within time and space reveal the social, economic and environmental constraints in which members of a group operate. The difference between members’ responsibility, mobility and freedom are laid bare in the time-geography ledgers to hint at levels of agency. As temporal, spatial and material data become increasingly associated with goods and products as they move through the value chains of production, distribution and consumption, projects such as Follow the Things extends the role of the ledger to involve not only humans, but the things that we buy.

In the third section, a dispute in a Peugeot factory that provided the inspiration for the Dardenne brothers' film *Two Days, One Night*, offered an opportunity to explore the tensions between established Fordist methods of car production in which the speed of production led to a ledger of poorly assembled cars, to the influence of the reflexive Toyota model that involved workers in the quality control of each car. Introduced as cognitive ledgers, the engagement of the individual to be involved in the evaluation of each transaction, whether it was the cinema audience who tried to retain a ledger of Sandra's interactions with her co-workers, or GM and Toyota employees that could halt the assembly line to rectify errors, the examples reminded us of how we enact ledgers.

At the time of writing, applications of distributed ledger technologies were still being developed and trialled within different sectors. As the technology finds its niche we can be confident that platforms that offer trust will support and encourage its members to trade any imaginable commodity. To illustrate, the authors used a case study to explore what designers would chose to exchange when they are given software that allows them to associate values to Bitcoin transactions in an urban context.

In summary, as the virtues and limitations of the blockchain and DLT technology begin to become part of the architectures of an internet in which trust is a key characteristic, the chapter reveals the complex relationship that we have between trusted lists, temporality and situated practices. Involving people in the co-authorship of a ledger suggests that we have an opportunity to understand the values that they intended to represent. Whatever form the city dashboard of the future takes, their function as visualizations of complex social, economic and environmental systems valued by all parties will require trusted methods for the capture and representation of values.

## Acknowledgements

The research for this chapter was funded by the UK Economic and Social Research Council grant 'After money: If you change the representation of value, does it change the values that you can represent?' Thanks to all workshop participants, Dave Murray-Rust, and software developer Hadi Mehrpouya.

## Note

1 http://www.followthethings.com/celebrityperfumes.shtml.

## References

Cook, I. (2015) 'Frequently asked questions', Follow The Things. Available from: www.followthethings.com/faq.shtml [accessed 26 June 2016].

DuPont, Q. and Maurer, B. (2015) *Ledgers and Law in the Blockchain*. Available from: http://kingsreview.co.uk/magazine/blog/2015/06/23/ledgers-and-law-in-the-blockchain/ [accessed 26 June 2016].

Harvey, D. (1990) *The Condition of Postmodernity*. Cambridge, MA: Blackwell.

Harvey, D. (1996) *Justice, Nature & the Geography of Difference*. Oxford: Blackwell.

Maurer, B. (2006) 'The anthropology of money', *Annual Review of Anthropology* 35: 15–36.

Maurer, B., Nelms, T.C. and Swartz, L. (2013) '"When perhaps the real problem is money itself!": the practical materiality of Bitcoin', *Social Semiotics* 23(2): 261–277.

Ohno, T. (1995) *Toyota Production System: Beyond Large-Scale Production*. Portland, OR: Productivity Press.

Parkes, D. and Thrift, N. (1980) *Times, Spaces and Places, A Chronogeographic Perspective*. Bath: Pitman Press.

Pialoux, M. (1999) 'The old worker and the new plant', in P. Bourdieu *et al.* (eds), *The Weight of the World*. Stanford, CA: Stanford University Press, pp. 267–281.

Rohter, L. (2014) 'Respect and awards, but still no Oscar, the Dardenne brothers discuss "Two Days, One Night"', *New York Times*. Available from: www.nytimes.com/2014/12/30/movies/the-dardenne-brothers-discuss-two-days-one-night.html [accessed 26 June 2016].

Schleicher, D., Jones, P. and Kachur, O. (2010) 'Bodystorming as embodied designing', *Interactions* 17(6): 47–51.

Thrift, N. (1996) *Spatial Formations*. London: SAGE.

# 12 Situating data infrastructures

*Till Straube*

## Introduction

When writing about digital technologies and the city as a human geographer it is a widely adopted convention to open with a brief fictional vignette: a woman is navigating urban space with her smartphone, friends run into each other because of Foursquare, somebody is lured into a store by a location-aware app, or a similar situation showcasing urban life being mediated by geodata in its various forms. These examples taken from (supposedly) everyday experience routinely serve as a point of departure for texts setting out to critically analyse the effects of new media and digital information on society at large, and their relation to urban spaces.

As a literary device, the vignette elegantly accomplishes several goals that should be prioritized by any good account. It offers anecdotal proof that digital devices influence the lives of real people, thereby establishing *relevance* of the technologies in question (and of the research presented). It additionally aims to facilitate the reader's intuitive understanding of an otherwise abstract, technically complex subject matter, advancing the text's *relatability*. Understood as a spatio-temporal *reference*, the vignette positions the suggested analysis in the here-and-now: at the time of this writing, an Uber ride would be a much more likely candidate for such an opening passage than a Facebook check-in. Finally, adhering to a style reminiscent of field notes from ethnographic research, the vignette also seems to add to the *authority* of the account.

However, the vignette also sets a narrative frame that makes some avenues of critical analysis appear more intuitive than others. Situating the lives and experiences of 'people like you and me' centre-stage reinforces an understanding of data and digital technologies as 'new *media*' – intermediaries, mediators – in relation to social processes.[1] The devices and technologies themselves remain largely 'black-boxed' and are considered only in regards to their effects on society (or as cultural expressions, for that matter; this is not a question of technological vs social determinism). This treatment of contemporary data-driven phenomena sustains the popular notion of 'the digital' as an immaterial sphere of relational intensities, and for some time underpinned some of the most influential analytical concepts guiding geographers' inquiries into this problematic – from time-space-compression (Harvey 1990) to the network society and the space of flows (Castells 1991; 1996).

Situating data in the city means to engage with spatial aspects of information and communication technologies (ICT). I propose to approach data-driven phenomena through *digital infrastructures* (programming languages, database software, data formats, protocols, APIs, etc.) and to understand their relation to space on their own terms: by closely reading documentation materials, technical specifications, and code. In this chapter, I outline a possible mode of inquiry that avoids relegating data devices to mere mediators of the social, or resorting to other grand abstractions. The entry point here is the problem of space: how can we account for the overflowing of spatial frames of reference by digital technologies without resorting to notions of immateriality?

This approach sets out by insisting on the constructed nature of the digital/physical dichotomy (among others) and proposes a radically materialist first analysis of ICT. In a next step, the problems of space and context are explored through an analytic lens building on assemblage theory, following the methodological principles of symmetry and free association. Taking further cues from science and technology studies (STS), the possibility of a topological reading of data devices is explored, and in a final step extended by reading it against a diverse set of additional texts and brief examples.

The present text, then, will have to make do without the all-but-obligatory opening vignette. *Relevance* is assumed: in a post-Snowden era, who would question the social and political importance of ICT and their critical analysis? *Relatability* comes at a too costly a price: when investigating technically complex digital infrastructures, there is a danger of losing nuance to simplification. Instead, this approach heeds Star's 'call to study boring things' (Star 1999: 377). *Reference* to specific technologies will not provide a precise spatio-temporal context: the programs and protocols in question are far more ubiquitous and longer-lasting than specific data sets, end-user applications or hardware generations. Finally, I make no claim to *authority* (or to a sound ontological stance): the veracity of the approach outlined here should be measured only against the quality of accounts it can facilitate.

## Digital technologies as a spatial problem

Binary computation is materially constituted by processes that 'take place' beyond human perception, be it almost-real-time interactions across globe-spanning distances, or computations 'at a scale of operation so small and fast as to be beyond direct human sensing' (Kitchin and Dodge 2005: 27). The mobilities and mutabilities facilitated by ICT are therefore greatly incongruent with those of (older) everyday things, and fundamentally challenge anthropocentric space-images. It is no surprise, then, that code and data are commonly thought of as consisting only of abstract connections, inhabiting a 'virtual' world that is separate from our tangible environment. Geodata may be conceived of as forming an annotating 'layer' to urban space, but its ethereal nature remains decidedly other to the hard materiality of storefronts, bricks and concrete.

Since the earliest attempts to make sense of the unfolding information revolution, the internet was often construed as an alternate immaterial realm where the laws of our physical world do not apply. The term ‘cyberspace’ was employed liberally to describe an exciting new reality brought forth by the internet – a world where bits and pixels seemed to float around freely and a whole new quality of possible mobilities and encounters emerged (Batty and Barr 1994; Graham 1998). At the time, thinking about digital phenomena followed ‘a paradigm . . . that was firmly ensconced in the notion that digital networks ran parallel to and remained separate from real life’ (Gordon and de Souza e Silva 2011: 8). Cyberspace was thus first and foremost a metaphorical space (Bell 2009).

To be sure, more recent efforts to theorize the relationship between the spheres of the digital/virtual and the physical/geographic insist on their mutually dependent relationship (Batty 1997; Aoyama and Sheppard 2003; Thrift and French 2002). For instance, the entirety of online geographical information is at times conceptualized as an additive dimension that is intricately intertwined with physical space: a ubiquitous virtual layer that is at once separate from, parallel to, and mutually constitutive with material places (Zook and Graham 2007). Research surrounding the ‘digital divide’ explores spatial (and other) inequalities in terms of internet availability, access to datasets and digital literacy (van Dijk 2006; Graham 2011). The increasingly popular (but fuzzy) notion of amalgamated virtual-material ‘hybrid spaces’ (Gordon and de Souza e Silva 2011: 86) that are imagined to flatten and fuse geometries of distance and scale through digital interfaces is intended to allow for consideration of both physical locales and digital processes simultaneously and coequally.

Without a question, these approaches to spatializing data and digital technologies have given rise to a wide range of fruitful critical research in geography and beyond. Ultimately, however, even the closest conceptual integration of digital information with physical settings relies on essentialized notions of both categories. Rather than any *specific* proposed relation between the digital and the physical, it is the *a priori distinction* between these two spheres that I wish to subject to scrutiny.[2]

## Assemblages of digital matter

Following Fuller’s assertion that ‘the idea of software’s “immateriality” is ultimately trivializing and debilitating’ (Fuller 2008: 4), I propose a materialist first analysis of digital technologies. This is to say nothing more than that digital communication, binary computation and data storage are strictly material phenomena (i.e. magnetic flux on a disk, electrical currents, photons in an optical cable, etc.) and that any proposed external reality (content, representation, meaning) cannot be taken for granted and should be treated with suspicion. The term *digital matter* is used here to indicate those tangible objects and infrastructures (like mobile phones, CPUs, screens, server parks, cables and so on) that do work in the world through ICT.

Insisting on the materiality of digital phenomena is hardly a novel move. Kittler famously stated that ‘[t]here is no software’ (Kittler 1995), pointing to the hard-wiredness of all computing processes in the last instance. Notions of the materiality of digital technologies also underlie the study of embodied practices of global financial traders in front of their screens (Knorr Cetina and Bruegger 2002). Kitchin and Dodge's work can be read along similar lines: drawing on Mackenzie's elaborations on technicity, transduction and individuation, they propose an understanding of embedded practices scripted by computer code that continuously and open-endedly structure 'code/spaces' (Kitchin and Dodge 2005; 2011; Mackenzie 2002). As Kinsley (2014) shows, these concepts – originating in deconstructionist philosophy and Simondonian thought – can be fruitfully employed to deconstruct 'virtual' geographies and engage with problems of materiality. Finally, Parikka calls for a 'multiplicity of materialisms' read against media theory:

> Such methodologies and vocabularies need to be able to talk not only of objects, but also as much about non-solid and the processual – the weird materiality inherent in the mode of technical media – so that we can understand what might be the specificity of this brand of materialism that we encounter (but not always perceive) in contemporary media culture.
>
> (Parikka 2012: 99)

As the above referenced accounts show a materialist reading of ICT does not preclude an engagement with *how* digital matter works – on the contrary: when trying to account for the unexpected, emergent configurations and associations brought into being by digital technologies, a close look at the actual mechanics and (coded) inscriptions becomes indispensable. As Latour remarks, 'the old dichotomy between function and form . . . is ridiculous when applied to a mobile phone. Where would you draw the line between form and function? The artefact is composed of writings all the way down' (Latour 2008: 4).

The dichotomies between *digital* and *physical* and between *form* and *function* are not the only ones that are called into question. In a strictly material analysis there can also be no natural distinction between *code* and *data* – they are handled exactly the same on the mechanical level of physical storage, transmission, the CPU and so on. Instead, this distinction is introduced by development practices and conventions, and supported by digital infrastructures like programming languages and database software. It should therefore be handled with care: data (e.g. a bitmap file) are always-also code (to display an image), and extensive infrastructures like Git exist for the sole purpose of handling code-as-data (or 'code as traffic'; Mackenzie 2016). Even in everyday development practice the line between code and data is blurry at best, and one would be utterly useless without the other.[3] Unlike notions of 'big data' or 'the algorithm', the comprehensive concept of digital matter avoids referring to only code or only data (and risk forgetting the integral role of the other).

Finally, this focal shift also draws on Whitehead's critique of the bifurcation of nature and culture – arguably the central starting point of STS and the schools of thought that followed (Stengers 2008; Latour 1993; Callon 1986; Law 1992).

Rather than media that unproblematically relay messages, or a type of resource, or instruments or circuits of power, digital matter comes into view here as the stuff of society itself. Following Latour (2005), mobile phones, lines of code and QR stickers are enrolled in associations just as complex as those of software engineers, users and other (human) members of society.

To account for this symmetry of humans and non-humans, and circling in on the problem of spatializing ICT, I further propose an examination of *assemblages* of digital matter. By employing the assemblage as an epistemological device, I want to draw attention to three defining ways in which it can inform thinking about digital technologies.[4]

First, the assemblage considers heterogeneous materials, including knowledge and practices. When deployed as an assemblage, devices of digital matter are defined by their relations not only to users and developers, but also to documentation, skills, best practices, standards and so on. Their analysis requires careful consideration of all the elements that are immediately involved in keeping them working.

Second, rather than static constellations, assemblages are never complete and always in the making. As an anti-structuralist concept, the assemblage allows for investigation into contingent arrangements that are temporarily stable, while taking into account their open-endedness and uncertainties regarding success or failure. The assemblage is derived from an active verb, and the intended purpose (for which a technology is deployed) and simultaneous over-determination of digital matter (resulting in unexpected behaviour or usage) are a source of tension that the proposed perspective draws on (Pickering 2003; 2008).

Third and most important for the problem of situating ICT, the assemblage not only challenges scalar thinking by rejecting the micro/macro distinction, but also blurs the line between inside and outside. In other words, there are no ‘parts’ to a ‘whole’, and assemblages are impossible to territorialize definitively. Looking at Google as an assemblage, for example, requires thinking about algorithms, server parks, developers, Web design, GPS satellites, stock prices, etc. – each for itself infinitely more complex than any imagined comprehensive entity: ‘The whole is always smaller than its parts’ (Latour *et al.* 2012: 590). What is more, each of these materials comes with its own spatio-temporal logic, and it is only within the assemblage that they are negotiated, translated and attuned to one another.

The proposed materialist reading of ICT is not at all intended to suggest, then, that digital matter can be unproblematically located in any ‘physical’ container space. But neither are arbitrary, generalized abstractions (like cyberspace, or the space of flows) productive in the envisioned goal of engaging digital technologies on their own terms. The contradictory spatialities and temporalities inscribed in and performed by assemblages of digital matter will be left unresolved and serve as a source of tension for the proposed topological solution to the problem of situating ICT.

## Topology: an ambiguation

In geography, topology is most commonly understood as an antithesis to topography; whereas topography locates things on projections of the surface of the

Earth, assigns coordinates, measures distances, and so on, topology is to describe connectivities and qualities of relations. If topography follows the logic of territory, topology is said to follow that of the network (Amin 2004). As an analytical tool, this reading of topology is employed to examine geographically dispersed networks and questions of power relations that escape Euclidian space. In shifting from topography to topology, measurable distances lose their meaning, and it is an 'other' form of connectedness that makes up a strictly relational space.

This reading of topology is at times illustrated by Serres's metaphor of the crumpled handkerchief (Serres and Latour 1995) or Euler's solution to the problem of the seven bridges of Königsberg (Shields 2012). Its usefulness has been demonstrated especially by accounts aiming to incorporate actor-network-theory into geography (Latham 2002; 2011; Allen 2011; McFarlane 2009; Murdoch 1998). The term 'network topology' is employed in computer sciences precisely in line with this understanding: it designates a description of how nodes in a computer network are connected to each other (rather than where they are located in relation to each other in Euclidian space), and it has become all but commonplace to remark that the connectivities introduced by ICT generally require topological rather than topographical analysis.

But this is not the only way (by far) in which topology is understood in its post-mathematical sense; a brief 'ambiguation' is in order. Martin and Secor identify a 'dizzying diversity' (Martin and Secor 2014: 2) of texts that use topology to mean vastly different things. They suggest that superficial readings of topology and its use as a vague heuristic device put the 'clarity and precision of [geographers'] theories of space' at stake (Martin and Secor 2014: 9). Specifically, they summarize a criticism of the topology-vs-topography debate, arguing that 'the point is to understand Euclidean space as *one possible topology among others*' (Martin and Secor 2014: 11; emphasis in original). Along those lines, Sha (2012) questions the suitability of the bridges of Königsberg as an example for topology. He posits that Euler's solution did not establish topology itself, but only graph theory, a precursor to topology. 'Topology is (much) more than graphs' (Sha 2012: 225). It seems indeed curious that 'topological space' is commonly understood as lacking metrics, when mathematicians routinely engage with metric topologies (Munkres 2000).

While an exhaustive discussion of post-mathematical topology is well beyond the scope of this text, I will briefly outline some alternative topology concepts that have been employed to think about data and digital technologies. Marres (2012), for example, uses topology synonymously to a Latourian symmetry in socio-technical arrangements. Rogers (2012) employs topology to recount specific historical geometries of the internet. Other scholars focus on effects of continuity and change in auto-spatializing practices of sorting, comparing and calculating: an algorithmic spatiality (Lury *et al.* 2012; Parisi 2013). Ruppert (2012) employs this perspective in her analysis of UK government databases and their conceptualization of individuality. Finally, researchers at the University of Technology in Darmstadt identify a series of relational spaces structured by (digital) technologies in their 'Topological Manifesto' (Graduiertenkolleg Topologie der Technik 2015).

From these diverse texts linking technology and topology, there are two aspects especially that will inform my proposed approach going forward. First is the notion of *auto-spatialization*, which is understood here as the capacity of (specifically data-oriented) technologies to impose their own spatial logic (of referencing, sorting, linking, categorizing) onto the world. It resonates strongly with the notion of transduction, and with the above proposed spatial deployment of assemblages of digital matter.

Second, as Martin and Secor assert: 'If there is something that unites geographers' uses of topology, it is a move to conceptualize the dialectic between continual change and enduring relations' (Martin and Secor 2014: 3). In a material analysis of digital infrastructures this tension between continuity and change takes centre-stage: when an e-mail is translated from a series of keystrokes to an electrical current, to an optical pulse, and to pixels on a screen, there is really no *material* continuity at all, and yet we can speak of the 'same' (textual) data. What changes, what stays the same, and how?

For articulating the mode of topological inquiry envisioned here, the main point of departure is the work of Law and Mol who employ topology to insist on a *multiplicity of space*, and propose a reading of *objects* in terms of the spaces folded into them (Law 2002; Law and Singleton 2005; Law and Mol 2001). This approach's analytic usefulness becomes clearest when they interpret the figure of the immutable mobile in this context:

> [W]e find that the immutable mobile achieves its character by virtue of *participation in two spaces*: it participates in *both network and Euclidean space*. And such is Latour's trick. To talk of an 'immutable mobile' is to elide the two.'
>
> (Law and Mol 2001: 612; emphases in original)

For the authors, this understanding of topology – that is, the deployment of concurrent spatialities – is not a heuristic device, or a result of epistemological uncertainty, but an ontological assertion. And yet, their topology does not amount to a model explaining the world, but should instead be read as a methodology with which to engage in research (Law and Singleton 2005). Along empirical examples as diverse as medieval Portuguese vessels, the Zimbabwe Bush Pump, aerodynamic calculations for a military aircraft and alcoholic liver disease, this group of authors develop notions of 'network', 'fluid' and 'fire spaces' which are as much at work in these objects as Euclidian space.

The intent of these alternate topologies was to trouble conventional notions of movement and continuity: for example, the form and function of the Zimbabwe Bush Pump changed as it travels through time and space through replication and recontextualization, and thus appears to the observer as a 'mutable mobile' (Laet and Mol 2000). But read through a topological lens, it is not the pump that changes: its mutability is simply an effect of a consistent object traveling through fluid space, observed from the vantage point of Euclidian space (Law and Mol 2001; Shields 2012).

For the remainder of this contribution, I will expand on this topology of multiplicity by reading it against some additional literature and short illustrative examples in order to put it to work in situating data infrastructures as assemblages of digital matter.

## Towards an applied materialist topology

The mode of inquiry I wish to develop here can be called an 'applied materialist topology', envisioned as an empiricist perspective that pays close attention to the concurrent articulation, performation and translation of multiple time-spaces by reading technological inscriptions in their socio-political contexts. It builds on Law and Mol's topological methodology and extends it by four epistemological moments, which are outlined below: arbitrary performation, post-human time-spaces, interlaced articulation and technical criticism.

### *Arbitrary performation*

There is a certain danger to looking at things through the lens of Law and Mol's topology: one runs the risk of seeing instances of fluid- or fire-like behaviour everywhere; of imposing them onto the object of study (or even introducing yet another, more appropriate associative spatiality; see Sheller 2004). This temptation should be resisted by steering clear of neatly fitting metaphorical spaces, and insisting instead on the fundamental *arbitrariness* of time-spaces as performed by digital matter. In a way, this means reading Law against himself: how could the clean-cut 'elemental' topologies (as in fire, fluid or solid networks) ever account for the full complexity – for the 'messiness' (Law 2004) – of the various spatio-temporal behaviours of digital matter?

For example, consider a topological reading of the Web's quintessential immutable mobile: the Internet Protocol (IP) is the technology that breaks up data (e.g. HTML files or bitmaps) into packets of manageable size, which are then independently routed through a network of computers, and finally reassembled at the recipient's end. Packet switching technology lies at the heart of how the internet works. IP packets are (super-)mobile in the sense that they traverse the World Wide Web in a matter of microseconds. They are (super-)immutable in that they require strict adherence to a predefined protocol. In IPv4 the destination address is determined by reading bits 128 through 159 of the packet. If one were to change just one bit in the header, the packet would become meaningless 'noise in a pipe'.

I argue that the IPv4 packet has nothing at all to do with Euclidian, network, fire or fluid spaces. Instead, I want to read the spaces it participates in directly from the object and its technical specifications: it is immutable in its specific sequence of ones and zeros – a one-dimensional topology of minimal (binary) difference – and mobile within a network of connected computers, each reading its header and passing it on towards its destination. Even its 'time to live' that prevents packets to travel around in circles (defined in bits 64–82, and originally intended as an integer value of seconds) has in practice become a 'hop count' reduced by one each time it passes through a node in the network. When the count reaches zero, the packet is discarded.

Taking cues from Callon's (2007; 2009) notion of performation, it can be said that this is an instance of an assemblage of digital matter *performing* the topologies of packet switching technology. These time-spaces are arbitrary in that they overflow – or rather are 'skew', or at odds with – preconceived notions of time and space (linear, Euclidian or other). To unearth them, to submit them to critical analysis, they need to be read for what they are rather than moulded to something that is already understood.

### *Post-human time-spaces*

Consider another example: Git is a popular free and open source versioning system developed by Linus Torvalds. It allows software engineers to keep track of changes made to a codebase over time and to cooperate on projects. Changes (or 'commits') are recorded for individual lines of code rather than files, allowing for a fine-grained transparency and flexible 'merging' of versions that have made changes to the same file. The technology follows a distributed approach, that is, instances of the same Git repository can exist independently on different machines and be selectively recombined without the necessity for a central server (like Github).

As previously noted, Git nicely illustrates the difficulty of the distinction between data and code. It also negotiates a variety of spatialities: hierarchical file systems, networks of developers, textual matrices of code, and of course its own tree-like logic of consecutive commits, branches and forks. Additionally, it can be understood as a tool for managing *temporality*. Git can be used to roll back to older (working) versions of a buggy software, and because individual commits contain only the incremental difference between two versions of individual lines of code, old commits can be removed or applied to newer branches, or entire branches can be 'rebased' to newer commits. Git allows for flexible re-ordering of a consecutive, incremental timeline of modifications – in other words, the changing of a code's history.

Inspiration for a critical analysis of such modulation of temporality comes from Amoore (2011), who shows how possible futures are folded into the present through risk scores calculated by self-normalizing security systems that are employed in air travel security. While geographers engaging with topology tend to focus on questions of space, others are concerned primarily with time (Serres and Latour 1995; Connor 2004). An applied materialist topology should be concerned, then, with non-linear *temporalities* as much as non-Euclidian spaces. Ideally, it transcends this distinction, and asks: what *time-spaces* are performed and negotiated by an assemblage (of digital matter)?

To further complicate matters, nothing warrants the expectation that the time-spaces performed be commensurable to human experience. For example, it is not at all out of the ordinary to employ more-than-three-dimensional calculative spaces, or nested loops and recursions in software programming. It is imperative, then, to be prepared for encounters with time-spaces that are difficult or impossible to relate to, without reducing them to some more intuitive framing (just as mathematics is equipped to handle topologies well beyond human capacity for imagination).

Literature surrounding feminist materialism and critical readings of cybernetics has a history of engaging with problems of agency and subjectivity in post-human ontologies and lend themselves to widen a topological perspective to include notions of time-spaces that escape human imagination (Pickering 2002; Barad 2003; Hayles 1999; Haraway 1987). Drawing on experiments in quantum physics, Barad develops the notion of 'spacetimemattering' to point to the inextricability of frames of reference and their 'dis/continuities' from those material arrangements that are employed in their observation (Barad 2010: 244). In order to further investigate the complex spatialities one invariably encounters in data devices, one would do well to draw on these literatures.

### *Interlaced articulation*

A superficial reading of Law and Mol's topology might suggest that Euclidian, network, fire and fluid spaces exist a priori and coequally as a sort of universal constant. Indeed, the authors never address the questions of how these spaces relate to each other, and where they originate. An applied materialist topology can engage this problem head on by examining the articulations and translations of time-spaces directly within the various inscriptions of digital matter.

What comes into view is more than just an interdependence, but a series of interlaced, convoluted articulations and translations. The rendering of a website – read as a matrix of colour values assigned to pixels in a 'screen space' – is articulated within the textual space of HTML and translated by the browser software, which is in turn articulated in the programming language C++, recurring back to the hierarchical space of the file system, articulated by the operating system, and so on. GPS would not function if the satellites' clocks did not take into account special and general relativity: surprisingly, the calculation of spacetime curvature is necessarily involved in the articulation of the 'simple' container space of a navigation system.

'We are thus confronted by an indefinite multitude of spaces, each one piled upon, or perhaps contained within, the next' (Lefebvre 1991: 8). Multiple topologies should be understood as articulated (or coded) within another and connected through diffuse passageways (Connor 2004; Serres and Latour 1995). The proposed mode of inquiry is therefore concerned with the 'pivot points' that translate between time-spaces: how smooth or disruptive are the shifts, or parallaxes, between the frames of reference performed by digital matter? Where are these 'topological operators' found, how do they put time-spaces in relation to each other, and in what spaces are they themselves articulated?

## Technical criticism

An applied materialist topology as proposed here is committed to technical detail in an effort to avoid generalized abstractions relating to the spatial and social being of digital technologies and information. It is easy to construe an objecting position that accuses this approach of being descriptive and 'forgetting about power relations'. But the critical quality of the envisioned mode of inquiry is not rooted in an account of embedded power dynamics (or 'geometries'). Following Latour (2005), I am instead interested in a perspective that decentres the subject, puts asymmetries under scrutiny, and understands power not as explanans, but as explanandum.

In his work on the interface as a cultural spatio-temporal operator that fragments and augments subjectivities, Hookway (2014: 23) proposes a conceptual distinction between sovereign *power* that 'define[s] and impose[s] order' on the one hand, and *control* that works 'on the threshold', through 'translations, transpositions, hybridizations, and phase shifts', on the other. An applied materialist topology rejects notions of the former, and concerns itself with the latter.

The mode of critique envisioned here is most closely aligned to the notion of 'technical criticism' developed by Ong and Collier from Weberian thought:

> [This critical stance] involves neither a sociological reduction to 'structure' or a logic of power nor a cultural reduction or relativization of such 'universal' phenomena. Rather, it suggests a careful technical analysis – a *technical criticism*. Such a technical criticism would examine both the 'mechanical' foundations of these phenomena and the actual processes and structures that define their scope and significance.
>
> (Ong and Collier 2005: 10; emphasis in original)

Context matters. Precisely because the proposed perspective is invested in technical detail, and because it steers clear of far-reaching generalizations and grand abstractions, it should be applied within contexts that need critical understanding, rather than to observe everyday practices and extrapolate from there. As Tsing (2010) points out, every account necessarily deploys its own frame of reference (or 'worlding') that sorts, orders, links and confines its materials. Even if one tried, then, it would be impossible to write a 'neutral' description of code, data and their technical workings.

When investigating the multitude of ways in which cities in the twenty-first-century enter into complex relations with data devices and digital infrastructures, it is therefore advisable to consciously deploy the spatio-temporal workings of ICT towards a technical understanding of the socio-political implications of data. I propose that an applied material topology can help such analyses by challenging preconceived dualisms of code and data, or the digital and the physical realms, and by widening the array of permissible spatialities deployed in forthcoming research.

## Notes

1 McLuhan's work, famously stating that 'the medium is the message', can be read as an early intervention into this predominant framing of media as a passive carrier (McLuhan 1994 [1964]).

2 Latour makes an analogous argument regarding sociologists' various efforts to come to an arrangement between interaction- and structure-based perspectives: 'The combination of two artifacts could only produce a third, yet more annoying, one' (Latour 1996: 234).

3 Notably, some functional programming languages (e.g. Lisp) follow the principle of *homoiconicity*, that is, they allow any piece of code to be handled like data, and vice versa.

4 For a more encompassing discussion of the assemblage in the social sciences and geography more specifically, see Marcus and Saka (2006) and Anderson and McFarlane (2011), respectively.

## References

Allen, J. (2011) 'Topological twists: Power's shifting geographies', *Dialogues in Human Geography* 1(3): 283–298.

Amin, A. (2004) 'Regions unbound: Towards a new politics of place', *Geografiska Annaler. Series B, Human Geography* 86(1): 33–44.

Amoore, L. (2011) 'Data derivatives: On the emergence of a security risk calculus for our times', *Theory, Culture & Society* 28(6): 24–43.

Anderson, B. and McFarlane, C. (2011) 'Assemblage and geography', *Area* 43(2): 124–127.

Aoyama, Y. and Sheppard, E. (2003) 'The dialectics of geographic and virtual space', *Environment and Planning A* 35(7): 1151–1156.

Barad, K. (2003) 'Posthumanist performativity: Toward an understanding of how matter comes to matter', *Signs: Journal of Women in Culture and Society* 28(3): 801–831.

Barad, K. (2010) 'Quantum entanglements and hauntological relations of inheritance: Dis/continuities, spacetime enfoldings, and justice-to-come', *Derrida Today* 3(2): 240–268.

Batty, M. (1997) 'Virtual geography', *Futures* 29(4–5): 337–352.

Batty, M. and Barr, B. (1994) 'The electronic frontier: Exploring and mapping cyberspace', *Futures* 26(7): 699–712.

Bell, D. (2009) 'Cyberspace/cyberculture', in R. Kitchin and N.J. Thrift (eds), *International Encyclopedia of Human Geography*. Amsterdam: Elsevier, pp. 468–472.

Callon, M. (1986) 'Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay', in J. Law (ed.), *Power, Action and Belief: A New Sociology of Knowledge?* London: Routledge, pp. 196–223.

Callon, M. (2007) 'What does it mean to say that economics is performative?', in D. MacKenzie, F. Muniesa and L. Siu (eds) *Do Economists Make Markets? On the Performativity of Economics*. Princeton, NJ: Princeton University Press, pp. 311–357.

Callon, M. (2009) 'Elaborating the notion of performativity', *Le Libellio d'Aegis* 5(1): 18–29.

Castells, M. (1991) *The Informational City: Economic Restructuring and Urban Development*. Oxford and Cambridge, MA: Blackwell.

Castells, M. (1996) *The Rise of the Network Society*. Oxford and Cambridge, MA: Blackwell.

Connor, S. (2004) 'Topologies: Michel Serres and the shapes of thought', *Anglistik* 15(1): 105–117.

Fuller, M. (2008) 'Introduction: The stuff of software', in M. Fuller (ed.), *Software Studies: A Lexicon*. Cambridge, MA: MIT Press.

Gordon, E. and de Souza e Silva, A. (2011) *Net Locality: Why Location Matters in a Networked World*. Chichester and Malden, MA: Wiley-Blackwell.

Graduiertenkolleg Topologie der Technik (2015) *Topological Manifesto* [accessed 30 June 2015].

Graham, M. (2011) 'Time machines and virtual portals: The spatialities of the digital divide', *Progress in Development Studies* 11(3): 211–227.

Graham, S. (1998) ‘The end of geography or the explosion of place? Conceptualizing space, place and information technology’, *Progress in Human Geography* 22(2): 165–185.

Haraway, D. (1987) ‘A manifesto for cyborgs: Science, technology, and socialist feminism in the 1980s’, *Australian Feminist Studies* 2(4): 1–42.

Harvey, D. (1990) *The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change*. Oxford and Cambridge, MA: Blackwell.

Hayles, K. (1999) *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago, IL: University of Chicago Press.

Hookway, B. (2014) *Interface*. Cambridge, MA: MIT Press.

Kinsley, S. (2014) ‘The matter of “virtual” geographies’, *Progress in Human Geography* 38(3): 364–384.

Kitchin, R. and Dodge, M. (2005) ‘Code and the transduction of space’, *Annals of the Association of American Geographers* 95(1): 162–180.

Kitchin, R. and Dodge, M. (2011) *Code/Space: Software and Everyday Life*. Cambridge, MA: MIT Press.

Kittler, F. (1995) *There Is No Software*, available from: www.ctheory.net/articles.aspx?id=74 [accessed 9 May 2015].

Knorr Cetina, K. and Bruegger, U. (2002) ‘Inhabiting technology: The global lifeform of financial markets’, *Current Sociology* 50: 389–405.

Laet, M. de and Mol, A. (2000) ‘The Zimbabwe bush pump: Mechanics of a fluid technology’, *Social Studies of Science* 30(2): 225–263.

Latham, A. (2002) ‘Retheorizing the scale of globalization: Topologies, actor-networks, and cosmopolitanism’, in A. Herod and M.W. Wright (eds), *Geographies of Power*. Oxford: Blackwell, pp. 115–144.

Latham, A. (2011) ‘Topologies and the multiplicities of space-time’, *Dialogues in Human Geography* 1(3): 312–315.

Latour, B. (1993) *The Pasteurization of France*. Cambridge, MA: Harvard University Press.

Latour, B. (1996) ‘On interobjectivity’, *Mind, Culture, and Activity* 3(4): 228–245.

Latour, B. (2005) *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford and New York: Oxford University Press.

Latour, B. (2008) ‘A cautious Prometheus? A few steps toward a philosophy of design (with special attention to Peter Sloterdijk)’, Keynote lecture for the Networks of Design meeting of the Design History Society, Falmouth.

Latour, B., Jensen, P., Venturini, T., Grauwin, S. and Boullier, D. (2012) ‘“The whole is always smaller than its parts”: A digital test of Gabriel Tardes’ monads’, *The British Journal of Sociology* 63(4): 590–615.

Law, J. (1992) ‘Notes on the theory of the actor-network: Ordering, strategy, and heterogeneity’, *Systems Practice* 5(4): 379–393.

Law, J. (2002) ‘Objects and spaces’, *Theory, Culture & Society* 19(5–6): 91–105.

Law, J. (2004) *After Method: Mess in Social Science Research*. London and New York: Routledge.

Law, J. and Mol, A. (2001) ‘Situating technoscience: An inquiry into spatialities’, *Environment and Planning D: Society and Space* 19(5): 609–621.

Law, J. and Singleton, V. (2005) ‘Object lessons’, *Organization* 12(3): 331–355.

Lefebvre, H. (1991) *The Production of Space*. Oxford and Cambridge, MA: Blackwell.

Lury, C., Parisi, L. and Terranova, T. (2012) ‘Introduction: The becoming topological of culture’, *Theory, Culture & Society* 29(4–5): 3–35.

Mackenzie, A. (2002) *Transductions: Bodies and Machines at Speed*. London and New York: Continuum.

Mackenzie, A. (2016) 'Code traffic: Code repositories, crowds and urban life', in R. Kitchin and S. Perng (eds), *Code and the City*. London and New York: Routledge, pp. 72–87.

Marcus, G.E. and Saka, E. (2006) 'Assemblage', *Theory, Culture & Society* 23(2–3): 101–106.

Marres, N. (2012) 'On some uses and abuses of topology in the social analysis of technology (or the problem with smart meters)', *Theory, Culture & Society* 29(4–5): 288–310.

Martin, L. and Secor, A.J. (2014) 'Towards a post-mathematical topology', *Progress in Human Geography* 38(3): 420–438.

McFarlane, C. (2009) 'Translocal assemblages: Space, power and social movements', *Geoforum* 40(4): 561–567.

McLuhan, M. (1994) *Understanding Media: The Extensions of Man*. Cambridge, MA: MIT Press.

Munkres, J.R. (2000) *Topology*. Upper Saddle River, NJ: Prentice Hall.

Murdoch, J. (1998) 'The spaces of actor-network theory', *Geoforum* 29(4): 357–374.

Ong, A. and Collier, S.J. (2005) *Global Assemblages: Technology, Politics, and Ethics as Anthropological Problems*. Malden, MA: Blackwell.

Parikka, J. (2012) 'New materialism as media theory: Medianatures and dirty matter', *Communication and Critical/Cultural Studies* 9(1): 95–100.

Parisi, L. (2013) *Contagious Architecture: Computation, Aesthetics, and Space*. Cambridge, MA: MIT Press.

Pickering, A. (2002) 'Cybernetics and the mangle: Ashby, Beer and Pask', *Social Studies of Science* 32(3): 413–437.

Pickering, A. (2003) 'On becoming: Imagination, metaphysics and the mangle', in D. Ihde and E. Selinger (eds), *Chasing Technoscience: Matrix for Materiality*. Bloomington: Indiana University Press, pp. 96–116.

Pickering, A. (2008) 'New ontologies', in A. Pickering and K. Guzik (eds), *The Mangle in Practice: Science, Society, and Becoming*. Durham NC: Duke University Press, pp. 1–14.

Rogers, R. (2012) 'Mapping and the politics of web space', *Theory, Culture & Society* 29(4–5): 193–219.

Ruppert, E. (2012) 'The governmental topologies of database devices', *Theory, Culture & Society* 29(4–5): 116–136.

Serres, M. and Latour, B. (1995) *Conversations on Science, Culture, and Time*. Ann Arbor: University of Michigan Press.

Sha, X.W. (2012) 'Topology and morphogenesis', *Theory, Culture & Society* 29(4–5): 220–246.

Sheller, M. (2004) 'Mobile publics: Beyond the network perspective', *Environment and Planning D: Society and Space* 22(1): 39–52.

Shields, R. (2012) 'Cultural topology: The seven bridges of Konigsburg, 1736', *Theory, Culture & Society* 29(4–5): 43–57.

Star, S.L. (1999) 'The ethnography of infrastructure', *American Behavioral Scientist* 43(3): 377–391.

Stengers, I. (2008) 'A constructivist reading of process and reality', *Theory, Culture & Society* 25(4): 91–110.

Thrift, N. and French, S. (2002) 'The automatic production of space', *Transactions of the Institute of British Geographers* 27(3): 309–335.

Tsing, A. (2010) 'Worlding the Matsutake diaspora: Or, can actor-network theory experiment with holism?', in T. Otto and N. Bubandt (eds), *Experiments in Holism*. Oxford: Wiley-Blackwell, pp. 48–66.

van Dijk, J.A.G.M. (2006) 'Digital divide research, achievements and shortcomings', *Poetics* 34(4–5): 221–235.

Zook, M.A. and Graham, M. (2007) 'Mapping digiPlace: Geocoded internet data and the representation of place', *Environment and Planning B: Planning and Design* 34(3): 466–482.

# 13 Ontologizing the city

*Tracey P. Lauriault*

## Introduction

This chapter presents a methodological approach for critically examining a data model. The central question examined is, how is a city translated into code and data, and how does that code and data transduce and reshape the city (see Figure 13.1; Kitchin 2011)? More specifically, what are the technopolitical processes by which a city is modelled or translated into a database? What does that database model look like? In what ways does that model transduce space and reshape the city? Is the relationship between model and city recursive and can a city database eventually learn about itself from itself and simulate the city? What would be included and what would be left out of the database in order to avoid the similitude problem of Lewis Carroll's map of the city at the scale of a 'mile of a mile' (Carroll 1893), or where cartography is so perfect that a map includes each house, mountain or tree represented by just that, the houses, mountains and trees as Borges satirically wrote in 'On exactitude in science' (1946)? Finally, who decides?

Data models look deceptively simple (see Figure 13.2). A data model is an ontology, a formal description of the entities of a domain. At its simplest, the purpose of a data model is to logically classify and uniquely identify things – material artefacts, ideas, people, concepts – and their relation to each other according to a set of rules in order to create a database of the things that form part of the main

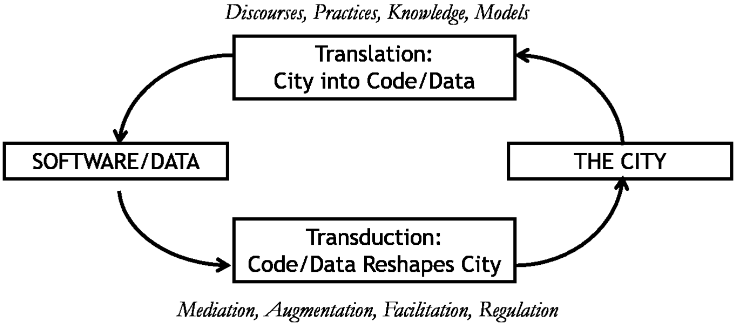


*Figure 13.1* Translation and transduction of data and the city (Kitchin 2011).

object of the model (i.e. Ireland). The object of a model or the domain could be the geography of a nation and its constituent elements, and in the case of a city its buildings, transport and utility infrastructure, waterways and land coverage, etc. Modelling in this sense entails the transformation of things into related conceptual objects and sorting them into classifications (Bowker and Star 1999). Once a data model is operationalized in a functional database, it gains momentum and becomes normalized throughout a socio-technological system (Hughes 1994) – such as the spatial database of a national mapping organization and the institutions that rely on its data (e.g. government departments, utilities or the education system). In turn, the classes and their relations seemingly become 'natural kinds of things' or 'real things' (Hacking 1991). For Hacking, a socially constructed class becomes known and acted upon as if it were an actual thing, and the thing (i.e. a tree) is only known by the qualities of how it has been classified.

Once a model becomes accepted infrastructure (Callon 1987; Star and Ruhleder 1996), it becomes difficult to change the model, and eventually harder to imagine the modelled objects and the object modelled in any other way (Lauriault 2012). The modelled database and its data become an objective reality, and once embedded into a data infrastructure, the model becomes largely invisible except for the rules articulated into algorithms encoded into the system. The machine however knows, and to a lesser extent the model becomes part of the tacit knowledge of its community of data producers and maintainers. The model eventually recedes into the background, becoming distant from its constructed and conceptual roots. Just like a well-functioning infrastructure, it becomes the substrate of other things

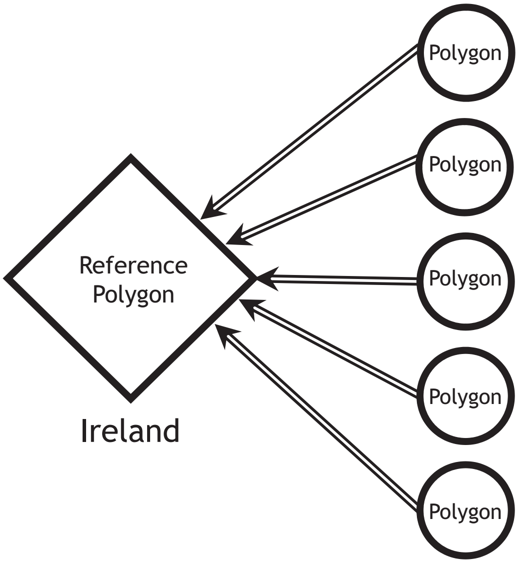


*Figure 13.2* Basic schematic of the OSi data model.

with agentric-like qualities (Hughes 2004; Allen and Hecht 2001) – the model produces the data infrastructure and is a product of it, and the use of the data infrastructure transduces the city.

Models abstract a world view and once operationalized into a database they can: influence how the world is viewed; can change the world of work, including tools, techniques and practices; and transform the structure of an organization and how that organization interconnects with others. If new standards are set, the organizations that are interconnected with it also change, and depending on the scale, they can influence the political economy of a particular databased ecosystem. Simple models are elegant (Figure 13.2), but are often deceptive, not being as simple as they appear given they capture and represent a complex world.

In order to illustrate data models and their utility, this chapter discusses a case study of the construction and deployment of the Prime2 data model and data platform created by Ordnance Survey Ireland (OSi). Prime2 is a real-world object data model that applies a ‘skin of the earth’ blanket metaphor (OSi 2014). In this model, the entire surface of Ireland is overlaid with a ‘consistent blanket’ that is shaped like the island and consists of millions of unique geometrically and relationally accurate (topological) patches of land covers and the outlines of structures (polygons) geographically stitched together by their topological relations and their geographic coordinates in such a way that there are no holes or gaps but for a few exceptions (such as cross over patterns, i.e. a bridge that crosses a river). Some features are overlain onto the blanket (i.e. buildings, structures and network lines), while polygons are seamed together with patterned stitching (i.e. networks of water, rail and ways), much like a modern quilt would be. This model is the guiding framework for the official digital geographic record for the state. These objects, however, need not be mapped or visualized, the data can be used in other ways. For example, they can be plugged into a direct mailing system to disseminate election ballots or census forms; questions can be asked of the database, such as how much of Ireland’s land surface is paved; and the paved surface data can be plugged into a pavement management system. The Prime2 data model underpins the design and construction of OSi’s data infrastructure and is the culmination of a seven-year national data re-engineering project.

The objective of the case study was to study the technological transformation at the OSi as a consequence of the implementation of this new data model in order to assess if it changed how space is understood, more specifically urban space, and if it produces different kinds of cities. In other words, does this model fundamentally alter how Dublin is known and if so how? The first section introduces Ordnance Survey Ireland, describes the Prime2 model, and details the case study. This is followed by a description of the theoretical assumptions guiding the study and how these informed data collection. The chapter ends with some preliminary observations given the analysis of the massive quantity of data collected is ongoing and concludes with some final remarks on the importance of empirically studying data models, and calls for critical data studies scholars to make data models a focus of analysis.

## Socio-technological transformation of Ordnance Survey Ireland (OSi): a case study

### *Ordnance Survey Ireland (OSi)*

Ordnance Survey Ireland (OSi) was established in 1824 and is the Republic of Ireland's national mapping organization (NMO). It has undergone a number of institutional transformations. Most notably, its mandate changed from being a colonial surveyor and military mapping organization under the British to a NMO just after independence in 1922. More recently, in 2002 it became a state body with both state and commercial functions under the Ordnance Survey Ireland Act 2001. In December 2016 a new National Mapping Agreement was officially signed with the government and the OSi's mandate will once again change when it merges with the Valuation Office and the Property Registration Authority sometime in 2017 to become *Tailte Éireann*. OSi is a well-established institution, every child studies its maps in school and all utilities and government offices rely on its maps and data to deliver services and programs. OSi has been actively engaged in the depiction of Ireland in maps and data for close to two centuries and has helped construct an empirical and scientific geographical imagination of the country. One of the questions of the case study is to consider whether or not Ireland will be imagined differently given the new ontology employed in its Prime2 database?

OSi is renowned for its mapping innovation internationally. Ireland was the first nation in the world to be completely scientifically surveyed in 1837–1846 by the British at the height of the empire's colonial era. The technique was exported throughout the colonies and surveying has become a norm and an established state apparatus globally. Like many NMOs it went digital in the mid-1970s and OSi implemented a series of geometrical framework models, created a new digital data collection, and developed new mapping techniques as computing power increased, the internet grew, new software and hardware tools were developed, and instrumentation improved and as earth observation (EO) technologies and imagery became more precise and accessible.

More recently, between 2007 and 2014 OSi and 1Spatial (a geospatial consultancy) created the Prime2 National Spatial Information Platform and together remodelled and re-engineered OSi data, institutionalized a new geospatial data management system, and built a new operational infrastructure. The objective was the provision of a single vision of the geographic 'truth' of the state, one that is standardized to align with OSi's mission 'to create, maintain and provide the State's definitive mapping and geospatial information services to support citizens, business and government', and its vision to be 'the national provider of trusted, maintained geospatial data and platforms to ensure the State's location data is easy to find, share and use' (OSi 2016). OSi, as one of the scientific arms of the state, typifies biopolitics both in terms of the management of territory and governmentality with respect to doing so (Foucault 2010; Garland 1997).

In September 2014, Prime2 was launched at the 1Spatial Big Data Roadshow in Dublin.[1] 1Spatial is a private sector firm that helps large data producing organizations transform their data using an enterprise-scale, rules-based approach that is

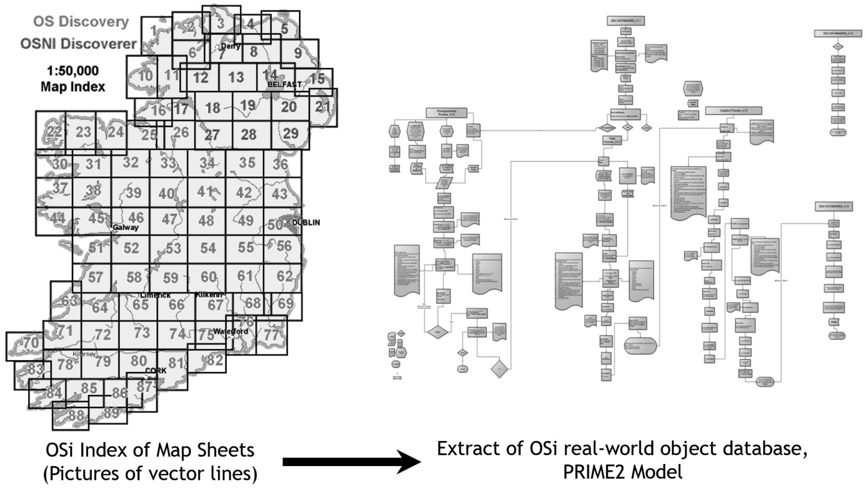


*Figure 13.3* From Ireland in maps to databased Ireland.

cross-platform and automated at all stages of the data lifecycle (1Spatial 2017). It was at this event that I realized that Ireland was being ontologized in an entirely new way. My hunch was that this was nothing less than a major socio-technological transformation, perhaps one of the largest OSi has ever undergone. Furthermore, this innovation, it seems, will influence the direction of other national mapping organizations around the world. This was made more poignant since Ireland was home to the world's first national survey.

The OSi was transitioning from the production and maintenance of a cartographic data system (Prime) produced exclusively for national mapping, toward spatial data as a big data database with mapping as just one application. In a seven-year period, the OSi re-engineered its data assets from sheet, tile and vector-based cartographic mapping to a seamless, scale independent real-world object oriented model approach. In a sense, Ireland became a real-world, feature-based information, national-scale spatial data platform (see Figure 13.3) which consisted of the sum of its materially defined parts defined by rules with topologically accurate, uniquely identified objects with spatial coordinates, attributes and spatial relations (OSi 2014). It is no longer a map, but a database of potential maps (Dodge *et al*. 2009).

### *High-level description of the OSi Prime2 data model*

In this model, Ireland is a set of objects that are topologically defined in relation to other objects, but also in relation to a reference object, which in this case is the geometry of the islands of Ireland (see Figure 13.2). These objects are topologically related, as seen in the few examples provided in Figure 13.4. While depicted as images for illustrative purposes, in the database, objects are described in code as having geometry, coordinates and attributes. In the OSi model, objects can be

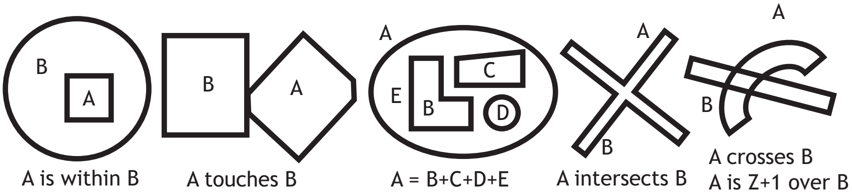


*Figure 13.4* Selection of polygon based topological relations in the Prime2 model.

contained within objects. For example, a building is located within a land area covered by a lawn. Objects can also intersect (e.g. a road intersects another) and they can touch and share a point or a boundary (e.g. the lawn surrounds the four sides of house and touches a fence). There are also grouped objects, which are a collection of objects that share common attributes but do not have a persistent geometry (e.g. a mall area). In this case, their geometry is derived from the union of all the objects in that group (e.g. a university campus and its parking lots and green areas, etc.). Finally, some objects may overlay others and are defined by a z-order priority, an object in this instance can be one or two or more orders above the skin of the Earth (e.g. a bridge over water) or below it (e.g. aqueduct).

The OSi model includes five skins of Earth objects (see Figure 13.5): (1) *way*, which is all driveable and walkable roads and pathways; (2) *water* objects, flowing, non-flowing, natural and human-made; (3) *vegetation*, which is all real-world vegetation cover from grass to forests; (4) *artificial* objects, which represent human-made ground cover such as concrete, rail beds and even gardens; and (5) *exposed* objects, which represent non-vegetative ground cover, which may be natural, such as sand, or human activity such as quarries. There are also superimposed objects that sit on top of these skins, such as buildings, building units, building groups, structures, divisions and service lines. In addition, there are networked objects,

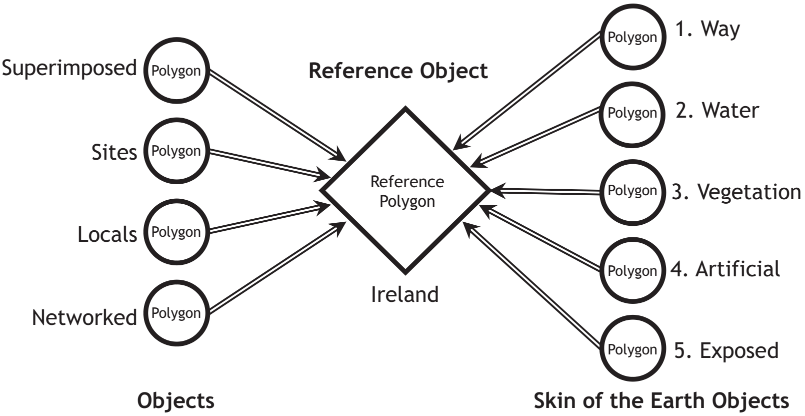


*Figure 13.5* Basic schematic of the OSi data model with object titles.

derived from the centreline of objects such a water and ways; and finally, there are sites (e.g. grouped objects) and locals (e.g. areas with fuzzy boundaries, such as a mountain range). Ireland becomes millions of classified and related objects which all refer to one master object, each uniquely identified and defined by rule sets.

### *Case study*

Rarely does one get to bear witness to such a massive technological change, and while at the Roadshow, I wondered what this meant in terms of understanding Ireland and if this new ontology was going to change how Ireland would be imagined (Lauriault 2012)? That same evening I spoke with representatives from 1Spatial who helped design the model and re-engineer the data as well as those involved with the process at OSi. They confirmed that this was in fact a massive change, part evolutionary and part revolutionary, and they were delighted that someone outside of their daily work observed the magnitude of this undertaking and took an interest. I proceeded to ask if they would be amenable to having a researcher examine the work they were doing, and in principle they thought this was a great idea. In consultation with Rob Kitchin, I designed a research case study proposal which was approved by the director at OSi, Colin Bray. Lorraine McNerney General Manager for Geospatial Systems and the person responsible for spatial data and systems helped me liaise with the institution. The resulting fieldwork included:

- conducting, recording and transcribing 30+ semi structured interviews and group meetings with executives, experts, data base managers, surveyors, cartographers, model designers, marketing managers, boundary experts, a variety of technological teams, accounting official and the procurement officer at the OSi, as well as designers and database engineers from 1Spatial;
- a tour of the OSi offices and facilities in Phoenix Park;
- a one-day visit of OSi offices in Sligo where the survey unit is based, and an on-site real-time demonstration of the data collection of a survey site and data flow line;
- studying data workflows, data lifecycles and data management before and after Prime2;
- studying ontologies and classification systems as these pertain to Dublin, cities/urban areas before and after Prime2;
- examining the data related to the lifecycle of properties and objects in Prime2 related to Dublin;
- studying the algorithmic approaches to map making before and after Prime2;
- studying the algorithms/processes of object creation and representation in a general sense as it pertains to Dublin and cities/urban environments in Prime2;
- tracing the social and material affects of this new mapping infrastructure, especially as it pertains to changes in how space has been reconceptualized.

In addition, grey literature was collected, such as:

- documentation related to the conceptualization and design process as well as the implementation of data models, concepts and objects, etc.;
- documentation related to timelines and key milestones on the implementation of the Prime2 National Spatial Platform;
- documentation (vision and implementation plan, requirements, specifications, instruction and operational manuals) related to the design, architecture and infrastructure (hardware, software, dbases, storage, standards, etc.) and to the implementation of the PRIME2 National Spatial Platform;
- software requirements, implementation process documents, documentation related to algorithms, how the system interconnects with other systems, installation processes, instruction manuals, operational manuals, policy and procedures, licences, databases, data storage, etc.;
- organizational structure diagrams, with roles, responsibilities and skill sets of OSI staff as a result of the implementation of Prime2 National Spatial Platform;
- catalogue/list of datasets/types collected and technologies related to the national surveying infrastructure which inform/populate PRIME2 National Spatial Platform;
- OSI ethical, normative and legal framework documents of the data;
- procedure and training manuals, reports, etc.

Finally, this was supplemented by the following tertiary data:

- reports, press releases, newsletters, web screen captures, presentations;
- news reports, clippings, videos, etc.;
- academic literature, theoretical, critical, pragmatic and methodological related to Object OP, OOD, software and database vendors, other implementations, standard, software studies, etc.

Most data collection work took place on site at the OSI between March and April 2015.

The OSi arranged private office space and full access to its staff. Being embedded in the organization and participating in meetings allowed me to get a sense of the place and to get to know the people working there.

## Theoretical approach

The approach taken to examining how a city is translated into code and data, and how these then reshape the city was a discourse analysis of the Prime2 data model and platform, with Dublin as the city. The analysis was structured by three nested frameworks: (1) an unpacking of the socio-technological assemblage (Kitchin 2014); (2) a Foucauldian genealogy of its development (Cosgrove 2001; Foucault 2003); and (3) an implementation of a modified version of Ian Hacking's dynamic nominalism and making-up people/places (Hacking 2001–2002; 2007; Lauriault 2012). These three approaches provide the means to study the OSi's national mapping infrastructure, of which the data model is a key component. Although these are different theoretical approaches, they share similarities and are interrelated through Foucault's ideas about the constitution and operation of power/knowledge.

### *Socio-technological assemblage*

Critical data scholars, situated broadly in the domains of critical social science and science and technology studies, accept that the usual technological conceptualization of data is limited and narrow and consider data to be more than neutral, unbiased, objective and scientific facts about the world. They contend that data do not exist independently from the context within which they were created, and the systems and processes that produce them (see Figure 13.6). The Prime2 Data model and platform is no exception. In order to study data in their 'habitat' and 'ecosystem', Kitchin (2014) offers a socio-technological assemblage approach to guide the empirical analysis of data (see also Kitchin and Lauriault 2014). The assemblage can be conceptualized as a constellation of co-functioning, loosely coupled heterogeneous elements, and it is these elements that guide data collection. Here, the assemblage is both a tool for research as well as a theoretical framing of data (Anderson *et al*. 2012). In essence, the elements in Kitchin's assemblage are the elements of an infrastructure, the hard and soft components, as well as the institutions and the environment within which it is situated. To make sense of an assemblage requires unpacking how it is constituted (Kitchin and Lauriault 2014). For this aspect of the case study, the databased transformation of the OSi infrastructure is the object of study. More specifically, infrastructure is understood as an assemblage, whereby the context that frames the model as well as the system is part of the infrastructure, as are the processes that perform the tasks of the model. For example, infrastructure is not simply hardware and software, it is the systems of thought that led to its creation, including how object-oriented modelling came to be and how that model materializes into code and algorithms which reformulated the entire data production flowline and its association with not only the equipment used by surveyors, but the entire database stack. It is only by looking at the model and how it came to be through database specifications and requirements, the observation of data production on-site in real time and in communication with database designers and mangers, that attributes of an infrastructure's assemblage can be observed in their state of play. What a cursory analysis shows is that the process of modelling is situated in the domain of object-oriented programming, the semantic web, GIScience, modelling software, taxonomies, the burgeoning database and GIS industry, modelling schemas, mathematics, consulting firms and offshore data re-engineering companies.

Furthermore, data modelling requires a particular form of logical abstract thinking, in the case of the OSi and 1Spatial those that were involved in the modelling exercise were very senior, experienced and renowned spatial data experts, all formally trained in spatial database design and maintenance as well as spatial

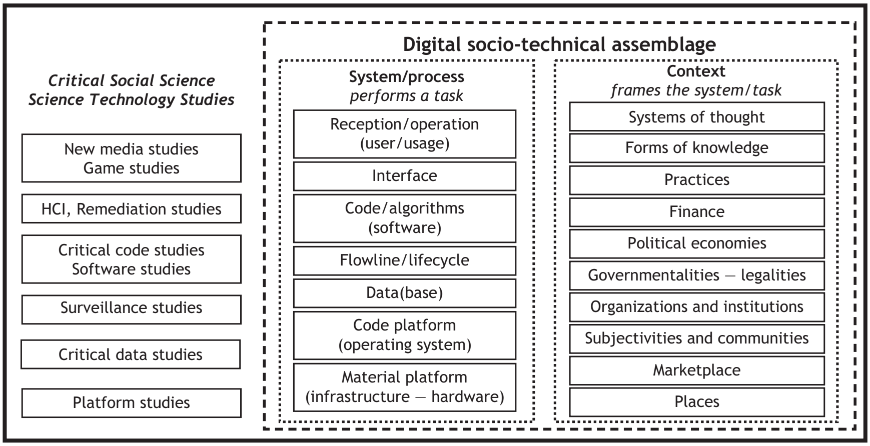


*Figure 13.6* Kitchin's socio-technological assemblage.

analysis at the enterprise level. The design and testing of a model is very labour intensive, recursive and incredibly expensive. At the OSi, this work was not done in-house, thus requiring the enactment of a procurement process to cover this major expenditure, and because of this, and because the model is key, it is a high-stakes tendering process.

## *Genealogy of a data model*

The Prime2 data model is part of the OSi's spatial data infrastructure, but the will and the act of modelling has its own provenance, discourse and language. Once a model is operational, it is hard to imagine its messiness and provenance. A genealogical approach provides for a deeper analysis of the evolution of power/knowledge of a data model and historically situates it in a very specific knowledge production process. Models do not just appear readymade. It took seven years to operationalize Prime2. The concept of remodelling at the OSi, however, goes back even further to the late 1990s, as part of early discussions between the CEOs and chief technology officers at OSi, OS Northern Ireland and OS Great Britain. Also some time was spent studying the trial and error of models deployed at other organizations, attending conferences and testing systems.

Figure 13.7 is an early schematic of the genealogy of the development of the OSi Prime2 data model. It is still a work in progress, but it nonetheless illustrates the sequence of key events and actions. For example, the research process of seeking documents and speaking with OSi employees revealed that details about models can be found in procurement documents since models need to be formally designed by experts, who will sketch out their version of a model requirements documents that will specify how that model can be encoded into software. A call for tender is required in order to hire those experts. The call for tender is a legal procurement process, as per public service regulation in Ireland but also according to European Union requirements for open, transparent, fair and objective tendering. Because modelling is complex, an education and communication strategy between OSi and those who are bidding on the contract is required, and this must be done equally with all those who answered the call, and this process is designed by a procurement specialist and comes with a set of documents. The decision to pick a specific solution is therefore not only contingent upon the art and science of databased modelling and database technology, but also on the procurement process, as well as contracting processes and bureaucratic knowledge. Applying a genealogical research methodology allows for these processes to become clear. In this case early model requirements were articulated in documents specific to the political economy of this form of procurement, in the proposals, the prototypes and in discussions between OSi and those bidding on the jobs.

Moreover, the new model does not completely separate itself from the old model, with Prime2 having echoes of Prime. New data uploaded into Prime2 are topologically situated in the past as Prime data remain the core having been re-engineered into Prime2, thus bridging the past within the present but also keeping the etymology of that change. Given the constant uploading of new data, the model is continuously and dynamically coming into being.

### *Dynamic nominalism and making up places*

Ian Hacking, as a philosopher of science and a contemporary of Foucault, deconstructed classification systems, primarily in the health sciences, in order to understand how these in turn produce knowledge about the work these do in the world, especially when classifications become understood as being the 'real thing' (1986; 1991). Hacking suggests that there are two interrelated processes at work within a data assemblage which both produce and legitimate a class, and those processes shape how that class does work in the world. In addition, he

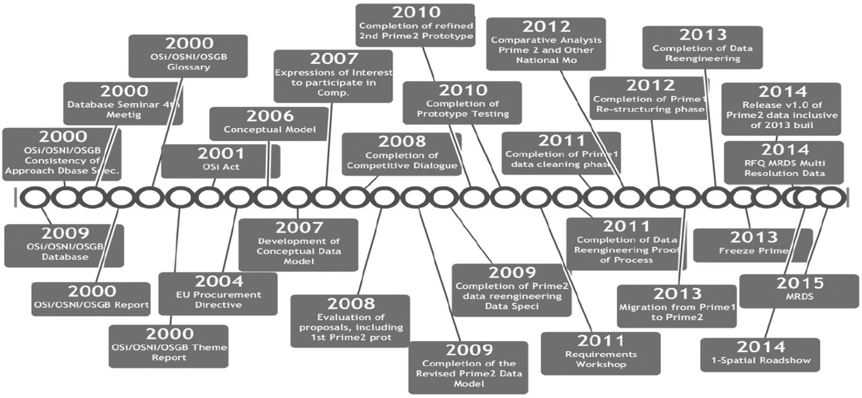


*Figure 13.7* A draft genealogy of the OSi Prime2 data model

observed that nominal classes are not firm constructs. He calls this dynamic nominalism, wherein there is an interaction between data classifications and what they represent that leads to mutual changes in the things classified and how classifications are understood across time and space. In the case of the Prime2 data model, Hacking's approach illustrates how 'real-world' objects and their attributes, and the things those objects represent, stay the same or change between the old Prime system and the new Prime2 system in terms of how Dublin is captured and represented. Hacking (2001–2002; 2007) calls the first part of this process 'the looping effect' (Figure 13.8). The looping effect concerns how data are classified and organized; in other words, how a data ontology or model comes into existence and how that can reshape that which has been classified. The loop has five stages:

1. *Classification*: or objects in this case (see Figure 13.8) are the grouping of things regarded as having shared characteristics.
2. *Objects of focus*: in this case, are the 'real' material things such as buildings, trees, or roads understood by people who then act toward these through the classification.
3. *Institutions*: that institutionalize classifications (ontologies, models) and manage the data infrastructures that operationalize these.
4. *Knowledge*: that is used to formulate, reproduce and tweak classifications (ontology, model).
5. *Experts*: those within institutions who produce and exercise knowledge, implementing the classification (database managers, modellers).

It is through this looping effect, Hacking argues, that the process of 'making people up' is found, or in the modified approach developed and applied by Lauriault (2012), where the process of 'making places up' occurs in data systems, such as a spatial data infrastructure. This is where the systems of classification work to reshape spaces and places in the image of a data ontology; for example, when 'sites' such as Phoenix Park become a prestigious city treasure that are acted upon

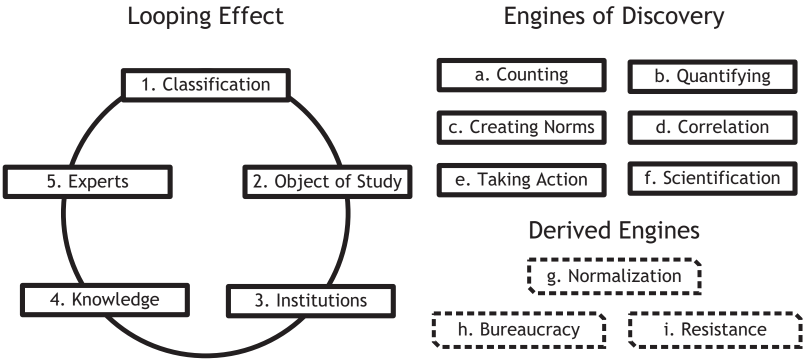


*Figure 13.8* Modified dynamic nominalism and making of spaces framework.

in such a way by citizens. In Prime2, Phoenix Park as an object is a site with a geometry derived from the sum of topologically related objects such as vegetation, buildings (including the President's residence, Dublin Zoo, military training grounds and the OSi offices), ways and other sites.

The second set of the processes is what Hacking terms 'engines of discoverability', that extend beyond simple methods, which he discusses using a medical lens, and which Lauriault (2012) has modified to incorporate the making up of spaces as well as people. Hacking (2007) posits that there are a number of such engines, the last three of which are derived engines and these are:

a *Counting*, the volumes of different phenomena (e.g. surfaces covered in pavement).
b *Quantifying*, turning counts into measures, rates, classifications (e.g. ratio of pavement to greenspace).
c *Creating norms*, establishing what might or should be expected (e.g. the ratio of pavement to greenspace for rural areas would be x while urban areas would be y).
d *Correlation*, determining relationships between measures (e.g. remote rural areas have a particular ratio and are considered underdeveloped).
e *Taking action*, employing knowledge to tackle and treat issues (e.g. the ratio suggests further development and funding).
f *Scientification*, establishing and adopting scientific knowledge (e.g. the right ratio for specific places is considered the objective formula for that kind of place).
g *Normalization*, seeking to fashion the world to fit norms (e.g. Phoenix Park is prestigious and therefore only prestigious institutions are associated with it, while the ratio of greenspace to pavement is optimal and other like places with that ratio should be created).
h *Bureaucratization*, putting in place institutions and procedures to administer the production of expectations and undertake action (e.g. ensure that the factors associated with the success of Phoenix Park are encoded in by-laws and adhered to).
i *Resistance*, to forms of knowledge, norms, bureaucracy by those that are affected in negative or positive ways and who suggest altering the form that makes up that (e.g. those who live on the outskirts of the Phoenix Park wall want to slow traffic speeds and increase traffic in the empty space of the park and want to increase the number of gates to make the park more accessible).

Hacking's framework, as applied here is a methodology to understand not only the ontology of the OSi model, and how these play out and shape the world, but also how they come into being and then change.

## Observations

In a sense, all three frameworks – assemblage, genealogy and dynamic nominalism – mutually reinforce the fact that data do not exist independently from the context they are situated in and the systems and process that bring them into being, are deployed and acted upon. All three frameworks in essence, employed as methodological approaches, are useful to empirically study Prime2 at different scales and from different angles, as an infrastructure or a large socio-technological system, from the historical roots of the model's provenance and the act of modelling as a system of thought, and at the micro-level to examine the objects themselves. Taken together the frameworks reveal the social construction of the Prime2 data model and platform. The re-modelling of the OSi data involved, among many things: the development of a data model; the adoption of a new ontology; the re-engineering of existing data; the restructuring of workflows; the hiring of and the retraining of personnel; the transformation and transmission of historical cartographical knowledge into a database but also the transfer of this knowledge to a new generation of employees via a database that is systemized and automated. It is also a process that: leaves behind tile-based cartography and layering and replaces them with the acquisition and implementation of new database technologies; involved the re-kitting of the surveyor's tools; and transformed the entire software and hardware environment. The new model changed the organizational structure of the OSi, where database managers become the scarcest and most desired resource. It also spearheaded the creation of a new e-commerce system and new relationships with old clients and a new set of relationships with new clients. This transformation aligns with OSi's stance that there should be one sole source of authoritative geographic 'truth' for Ireland, in this case a rules-based real-world, object-oriented geometrically accurate one.

This re-modelling entailed not only internal transformation, which will continue to have repercussions beyond the OSi as an institution since current clients, most notably those who rely on these data as foundational or framework data for their workflows and decision support systems (such as utilities, city planning and transportation) and will have to change their workflows. In addition to changing the technological and databased relationships with current clients, it augments big data possibilities, since the databased infrastructure is as much about topological and geometric accuracy, reliability, authenticity and standardization, as it is about location-based services, semantic interoperability and moving beyond a visual representation of the 'fixed' form of material entities. Function is now captured in the model, current and past, while changes in form are also traceable. While not an archive, the etymology or the provenance of things in the model is now possible, the objects in the system are therefore also records in the archival sense of the term, and the unique ID systems provides linkages to all the permutations of related objects across space and time.

The cartographic representation of the material (i.e. building) and conceptual (i.e. boundaries) objects which constitute Ireland is no longer the only 'representation' of what is, as unique identifiers allow for topologically consistent 'real-world objects' which are scale-independent and can to be connected with any number of other databases and attributes (such as archaeological information and artefacts, postcodes, place names, valuation and cadastre, social media, historical letters and maps, statistics, genealogical records, loyalty cards, mobiles devices, in-car navigation systems). This provides for the possibility for linked data for which the OSi is experimenting (Debruyne *et al.* 2017). A map need not be rendered for some applications, as the results of a database query may suffice. This re-modelling away from old school cartography therefore allows for new institutional relationships to emerge and a new political economy, which will be spatial and databased but not necessarily map-based.

Much more analysis of this treasure trove of data the OSi so generously offered to this case study is required. This will hopefully lead to answers to the research questions posed at the beginning of this chapter, but also, will provide empirical results to better critically study data; to better see how the elements of the assemblage are articulated in a large socio-technological system such as the OSi; and to see how data and code take shape and where power/knowledge is enlisted to produce places.

## Acknowledgements

The research for this chapter was funded by a European Research Council Advanced Investigator award (ERC-2012-AdG-323636-SOFTCITY). I would like to express my gratitude to all at the Ordnance Survey Ireland (OSi) for generously sharing their knowledge and time.

## Note

1 www.geoconnexion.com/news/1spatial-the-spatial-big-data-roadshow-dublin-17th-september/.

## References

1Spatial (2016) https://1spatial.com/.

Allen, M.T. and Hecht, G. (2001) 'Authority, political machines, and technology's history', in G. Hecht and M.T. Allen (eds), *Introduction in Technologies of Power: Essays in Honour of Thomas Parke Hughes and Agatha Chipley Hughes*. Cambridge, MA: MIT Press, pp. 1–24.

Anderson, B., Kearnes, M., McFarlane, C. and Swanton, D. (2012) 'On assemblages and geography', *Dialogues in Human Geography* 2(2): 171–189.

Borges, J.L. (1946) 'Del rigor en la ciencia [On exactitude in science]', *Los Anales de Buenos Aires* 1(3).

Bowker, G.C. and Star, S.L. (1999) *Sorting Things Out: Classification and its Consequences*. Cambridge, MA: MIT Press.

Callon, M. (1987) 'Society in the making: The study of technology as a tool for sociological analysis', in W.E. Bijker, T.P. Hughes and T.J. Pinch (eds), *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. Cambridge, MA: MIT Press, pp. 83–106.

Carroll, L. (1893) *Sylvie and Bruno Concluded*. London: Macmillan and Co.

Cosgrove, D. (2001) *Apollo's Eye: A Cartographic Genealogy of the Earth in the Western Imagination*. Baltimore, MD: Johns Hopkins University Press.

Debruyne, C., Clinton, É., McNerney, L., Lavin, P. and O'Sullivan, D. (2017) 'On the construction for a linked data platform for Ireland's authoritative geospatial linked data', available from: www.osi.ie/wp-content/uploads/2017/01/osi-eswc-2017-preprint.pdf [accessed 10 February 2017].

Dodge, M., Kitchin, R. and Perkins, C. (eds) (2009) *Rethinking Maps: New Frontiers in Cartographic Theory*. London: Routledge.

Foucault, M. (2003) *The Essential Foucault: Selections from Essential Works of Foucault, 1954–1984*. New York: The New Press.

Foucault, M. (2010) 'Biopower', in P. Rabinow (ed.), *The Foucault Reader*. New York: Vintage Books, pp. 257–273.

Garland, D. (1997) 'Governmentality and the problem of crime: Foucault, criminality, sociology', *Theoretical Criminology* 1(2): 173–214.

Hacking, I. (1986) 'Making up people', in T. Heller *et al*. (eds), *Reconstructing Individualism*. Stanford, CA: Stanford University Press, pp. 222–236.

Hacking, I. (1991) 'A tradition of natural kinds', *Philosophical Studies* 61: 109–126.

Hacking, I. (2001–2002) 'Façonner les gens, École de France Philosophie et histoire des concepts scientifiques (2001–2006)', available from: www.college-de-france.fr/site/ian-hacking/course-2001-2002.htm [accessed 19 January 2017].

Hacking, I. (2007) 'Kinds of people, moving targets', British Academy lecture, read at the Academy, 11 April 2006, available from: www.britac.ac.uk/sites/default/files/hacking-draft.pdf [accessed 19 January 2017].

Hughes, T.P. (1994) 'Technological momentum', in M.R. Smith and L. Marx (eds), *Does Technology Drive History? The Dilemma of Technological Determinism*. Cambridge, MA: MIT Press, pp. 101–114.

Hughes, T.P. (2004) *Human-Built World: How to Think About Technology and Culture*. Chicago, IL: University of Chicago Press.

Kitchin, R. (2011) 'The programmable city', *Environment and Planning B* 38: 945–951.

Kitchin, R. (2014) *The Data Revolution*. London: SAGE.

Kitchin, R. and Lauriault, T.P. (2014) 'Towards critical data studies: Charting and unpacking data assemblages and their work', available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2474112 [accessed 10 February 2017].

Lauriault, T.P. (2012) 'Data, infrastructures and geographical imaginations', unpublished PhD dissertation, Carleton University, available from: https://curve.carleton.ca/system/files/etd/7eb756c8-3ceb-4929-8220-3b20cf3242cb/etd_pdf/79f3425e913cc42aba9aa2b9094a9a53/lauriault-datainfrastructuresandgeographicalimaginations.pdf [accessed 10 February 2017].

Ordnance Survey Ireland (OSi) (2014) 'Prime2 – Data concepts and data model overview', available from: www.osi.ie/wp-content/uploads/2015/04/Prime2-V-2.pdf [accessed 10 February 2017].

Ordnance Survey Ireland (OSi) (2016) 'Statement of strategy 2016–2018', available from: www.osi.ie/wp-content/uploads/2015/04/OSi-Strategy-Statement-2016-2018-Final-Eng.pdf [accessed 10 February 2017].

Star, S.L. and Ruhleder, K. (1996) 'Steps toward an ecology of infrastructure: Design and access for large information spaces', *Information Systems Research* 7(1): 111–134.

# Part IV

# Urban data cultures and power

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

# 14 Data cultures, power and the city

*Jo Bates*

## Introduction

Data have consequences. The nature of these consequences is the outcome of people's interrelationships with the complex socio-material conditions they encounter as they think about and work with data. These relations frame people's engagement in practices of data production, processing, distribution and use, as well as their efforts to enable and restrict their own and others' data practice. At different *sites of data practice* and *data governance*, we can observe forms of *data culture* begin to coalesce in response to the socio-material conditions encountered by participants. These data cultures influence whether and how data are produced, processed, distributed and used; they shape what can, and cannot, be viewed through an informational lens, as well as the particular qualities of that lens. They shape the 'material properties' of data – their persistence, durability, spatiality, size, mobility, etc. (Dourish and Mazmanian 2011), and influence the development of the physical infrastructures that data depend upon. Through their practice these data cultures contribute to how we understand the world around us, and to the development of material conditions of production. In this chapter, I consider what questions need to be posed of the emergent data cultures found at different sites of data practice and governance, and begin to explore how participants in these spaces are influencing our perceptions of cities and the socio-material conditions of their future development.

## Big data and the city

In 2013, IBM (2013) calculated that 90 per cent of all data in existence had been created in the previous two years. Significant amounts of these data are the by-product of our everyday interactions with digital information and communication technologies. For many in business, government and research funding, this is the era of big data and new sources of data and data analysis techniques that will fundamentally change how societies are governed, and business and science conducted (Kitchin 2014a). The World Economic Forum (2016) has proclaimed that such developments, in part, constitute a 'Fourth Industrial Revolution'. Commercial organizations, universities, public bodies, governments and citizens are increasingly questioning how they should respond to developments within this shifting data landscape. Organizations currently face a variety of technical and labour barriers to processing these vast quantities of largely unstructured data. For many people, their priority is addressing how best to exploit these data in order to improve organizational intelligence, drive decision-making processes, and inform various other forms of value generation. However, these emerging practices of data production, processing, distribution and use also raise a multitude of complex social and ethical concerns that need to be addressed.

In cities, public and private efforts are converging to explore ways in which these data can be used to make cities more efficient, responsive and competitive within the global economy. A range of consultancy and data analytics firms, from major corporations such as IBM, CISCO and Pricewaterhouse Coopers (PwC) to smaller niche firms, are working with public authorities and other private sector organizations to assist them in extracting value from data in order to gain a deeper insight into urban dynamics, often generating even more data in the process. Across local and national government, and the wider public sector, organizations are beginning to explore the possibilities of data-informed public policy (Cabinet Office 2015), data-driven urban dashboards (Mattern 2015), and other 'Smart City' initiatives (Batty *et al.* 2012). Citizens and businesses have also demanded that data held by governments and public bodies be 'opened' for them to access and re-use so that they can develop their own information resources and applications using public data (Bates 2013; Kitchin 2014a).

The rationale behind many of these initiatives tends to be based upon a series of assumptions regarding the contribution data can make to developing and growing urban economies and improving the quality of life in cities, often with the intention of enhancing the city's competitive position in relation to other urban centres (Hollands 2008). Developments often aim to improve the management of a range of socio-material aspects of cities from democratic engagement to public transportation. Investments of time and money are being made by businesses and public bodies in order to explore how best to draw out these, and other forms of, perceived value.

Within universities, academics are being called upon to join these efforts through engagement in collaborative and interdisciplinary projects that aim to develop insight through data, and in some countries significant investments are being made to tackle a perceived quantitative skills deficit amongst social science students and researchers (British Academy 2012). In some cases, these developments have contributed to academics being encouraged by funders and universities to take on uncritical and enabling roles working closely with politically and economically powerful agents. In other cases, academics have aimed to understand and critically engage with the underlying assumptions and methodologies of collaborators, in order to try and influence the direction of projects away from the uncritical forms of empiricism that can be prevalent in data and computational science (Ruppert 2013; Kitchin 2014b). Such collaborative efforts are important, and whilst the agency of critical researchers engaged in them is necessarily constrained by the power dynamics of the collaboration and the wider societal context (Viseu 2015), such research has the potential not only to increase understanding about cultures of data practice and governance at particular sites, but also to become part of and influence the development of the projects they are embedded within.

## Data cultures

The concept of a data culture has been drawn upon in a variety of settings. In academia, the notion of a 'local data culture' was articulated by Bowker (2000) in relation to the diverse range of data coding and classification norms and practices that exist amongst biodiversity researchers. More recently, the concept has begun to emerge within anthropology, with ethnographers in the Research Data Alliance beginning to document and analyse the diverse data cultures that exist within the alliance.[1] The idea of a data culture is also recognized within the corporate sector. For example, Microsoft has adopted the term for a series of 'Data Culture' workshops, delivered in partnership with KPMG, Hortonworks and Hewlett Packard, which aim to assist data specialists in the development of 'game changing' data cultures within organizations.[2] All of these instances of the term data culture refer in some way to what others, including Kitchin (2014a) and Lauriault (2012), have identified as the different cultural norms, value systems and beliefs that inform, frame and justify people's practices of data production, processing, distribution or use (data practices), as well as their efforts to govern and shape particular forms of data practices through a variety of social and technical means. Sites of data practice and governance include small groups or teams, distributed networks, different types of organizations and other social collectives, each of which evolve their own complex data culture. Similar to Massey's (1994) conceptualization of space, we can recognize these sites of data practice and governance as historically constituted, dynamic, open and porous. Each local data culture develops in relation to the specific ways in which it interacts with the complex socio-material conditions that stretch beyond that particular site of data practice: 'the global as part of what constitutes the local, the outside as part of the inside' (Massey 1994: 5). From this perspective, a data culture can be understood as a specific articulation of socio-material relations situated within time and space. Whilst all data cultures are in some way interrelated, they are not all created equal. The socio-material conditions that these sites and cultures of data practice and governance emerge within enable and encourage some ideas and activities, whilst restricting and constraining others to varying degrees and in various ways.

Within the context of a single city, we can observe a multitude of interrelated data cultures across sites of data practice and governance located in public organizations, private enterprises, research settings and amongst citizens. Further, the development of a city is also heavily shaped by data cultures external to its geographical boundaries, for example those of finance, corporations, government, etc. Close examination of the cultures of these sites of data practice can, in part, help us to begin to answer questions regarding how participants in these spaces are influencing our perceptions of the city and the socio-material conditions of its future development, and ultimately to uncover some of the power dynamics at play in these processes.

In order to illuminate such processes in relation to a particular data-driven initiative in a city a number of different questions might be asked:

- First, what **sites of data practice and governance** are engaged in the data-driven initiative both inside and outside the city? Where are these sites located? How are these sites interconnected? What is the relationship between them? How do data move between these different sites? What socio-material factors (e.g. policies, infrastructures, finances, etc.) are influencing the development of, and relationships between, these sites?
- Second, what is the **data culture** of each of these sites? Who participates in each site? What are the demographics? How do participants perceive data? How do they imagine the relationship between data and the reality they aim to represent? What explicit and implicit values, assumptions and beliefs do participants bring to their data practice? What forms of value do they perceive in the data and practices they are engaged in? What are their overarching aims and how do participants perceive their practices contribute to them? What opportunities, risks and limitations do they perceive? How do these perceptions frame and give justification for participants' data practices?
- Third, how does **power** shape relations between these data cultures, and between them and the wider socio-material context? What tensions exist within and between data cultures and how do these play out? How does the wider socio-material context influence how data cultures imagine and engage with data? Who has more or less power to shape data cultures and their practices?
- Fourth, what are the **possible implications** for how we perceive the city? How do these factors contribute to the socio-material conditions of its future development? And, who might this advantage and disadvantage?

The following section will go on to begin to explore some of these questions in more depth, drawing on illustrative examples from ongoing research examining cultures of data practice and governance in the UK.

## Sites of data practice and governance

Sites of data practice and governance are hugely diverse. They vary in relation to size, structure, longevity, connection to physical place, participants, purpose and, crucially, power. Important sites to consider in relation to the development of cities include those within local government, public bodies, regulatory agencies, universities, citizen-led groups and private sector data analytics firms and consultancies that are engaged in new forms of data work. Each of these sites of data practice and governance are interrelated with the wider socio-material context – an 'assemblage' of historically constituted public policy, legislation, political economy and various other factors that inform and shape how ideas and practices around data unfold (Kitchin 2014a).

As an example, we can observe that in the last decade groups of citizens have formed open data groups in a number of cities around the world in an effort to extract data from public authorities and use it for a variety of ends – civic and commercial. These local groups can be conceptualised as sites of both data practice and governance. They are to varying degrees interconnected with other citizen-led open data groups and networks in different locations; other types of citizen-led interest groups – including open government and transparency campaigners who have been working for the release of public data for many years; policymakers at local, national and international levels; civil servants and employees of other public organizations; research and other staff based in universities; and commercial re-users of public data and other private interests – each of which have their own different data culture and different objectives regarding the opening of public data.

In the UK, the convergence of these different sites of data practice and governance, in a broader socio-material context of technological, political, economic, legislative and policy developments impacting the management of the national data infrastructure, led to these small city-based open data groups being able to have significant influence on other sites of data practice and governance within their respective cities. For example, the Manchester Open Data group played an important role in the development of the Data GM open data portal and a range of data-driven projects in the city. Similar observations can be made in other cities, for example, Sheffield and London. However, the specific changes in local authorities' data practices, for example which particular datasets have and have not been opened, have to a great extent been shaped by other sites of data practice and governance beyond the direct influence of these local open data groups. This demonstrates that it is important to map the complex network of sites and assemblages that are engaged in shaping emergent data practices, as well as considering the influence of the broader socio-material context that they exist within.

## Cultures of data practice and governance

Certain ways of thinking – beliefs, values systems, perceptions – tend to cohere as a form of 'common sense' (Hall 1987) at particular sites of data practice and governance as participants work together to respond to the wider socio-material conditions they encounter. These 'common sense' ways of thinking also emerge and spread across networks of interconnected sites. However, no site of data practice and governance is culturally homogenous. Common sense is 'necessarily fragmentary [and] contradictory' (Hall 1987), and tensions of varying degrees of magnitude emerge amongst participants as they work to construct shared understandings. The depth and nature of such tensions can help to illuminate the dynamics of power within a particular site, as well as its relationship with other sites of data practice and governance, and the wider socio-material context. The nature and relational dynamics of these 'common senses' that frame and give justification for data practices, can be identified as data cultures. As articulated above various questions can be asked of a data culture, here we draw out some of these in more depth, focusing specifically on philosophical beliefs, socio-cultural values and assumptions, and how the value of data practices is perceived.

### *Philosophical beliefs*

Important indicators of the nature of a data culture are the dominant epistemological and ontological beliefs it has about what data represent and how participants perceive the role of people in the construction and interpretation of data. Many that work with data on a regular basis are aware of the fallibility of data produced by poor quality methods or equipment, although this level of data literacy cannot be assumed across all data cultures. Beyond basic data literacy, it is also important to examine the more subtle philosophical assumptions of data cultures, such as whether there is a deep understanding of the ways in which data are socially constituted, or whether high quality data are perceived to be scientific or objective facts independent of social and cultural influence.

How a data culture perceives the nature of reality, and whether there is an assumption that all aspects of reality are empirically observable, can also be explored. Observers of sites of data practice have identified a dominant empiricist orientation amongst the participants of many emergent data cultures (Kitchin 2014c). Philosophical assumptions about the relationship between data and reality, and how these beliefs inform practice are therefore important to unpack. For example, is a simple relationship between empirical data and social reality assumed, or is the complexity of this relationship acknowledged? Is there an underlying assumption driving the data culture that social reality in its entirety is an observable phenomenon given the correct data collection and analysis techniques and tools?

Empiricist assumptions are prevalent, to a greater or lesser degree, within many data cultures, however such approaches are limited in terms of what they can observe and say about the nature of social reality (Archer 1998). In order to address questions of power, for example, the concepts and frameworks of the critical social sciences and humanities can contribute to explaining and theorizing patterns observed in empirical data. The degree of openness of a data culture to such forms of interdisciplinary engagement can illuminate how participants imagine the nature of social reality, as well as how they perceive their role as producers of knowledge, how they understand and relate to different ways of knowing, and how the data culture imagines its position within the power-infused processes of knowledge production (see Gaventa and Cornwall 2008).

Of course, to recognize that data practices are the product of a socio-material context and are limited in terms of what aspects of reality they can represent does not mean the scientific knowledge they might inform should be discredited (Edwards 2010: 436–438). Data can be very accurate and reliable enough for particular purposes. For example, the high-quality climate datasets generated by weather stations such as Sheffield Weston Park are reliable enough to give climate scientists a strong understanding of changes to urban climates over the last century and more (Jones *et al.* 2012). Whilst such accounts of reality remain partial and are an informational representation of the world emerging from a particular socio-material context, they are in many cases good enough for the ends to which they are being put.

### Socio-cultural values and assumptions

The socio-cultural values that are part of the data culture at Sheffield Weston Park Museum weather station have contributed significantly to the accuracy, completeness and reliability of its 135-year climate dataset. The museum curator who currently looks after the weather station, its historic written logs and more recent digital records, takes great pride in ensuring the quality of the data produced by the station. He recognizes the weather station as part of the fabric of the city, and the data it produces as part of the cultural heritage of Sheffield that belongs to the people of the city. Drawing on a cultural value system that champions public service, civic duty, scientific integrity and responsibility to his local community, data users and previous generations of curators who have maintained and run the station, the curator works hard to look after the physical infrastructure of the station and ensure the data it generates abide by international standards and are as accurate as possible.

Such cultural values are not unbiased in relation to the social world, however. When they surface in other contexts of data practice their influence may not always be as benign, for example when observing the social world such value frameworks may hide unacknowledged and implicit biases and assumptions that are brought into play as practitioners produce and engage with data (Greenwald and Krieger 2006). These biases might be political, social or cultural, and can impact significantly upon what people perceive to be of interest, what they prioritize, and whether and to what extent they look for and try to identify their own and others' underlying biases. It's important to gauge how much critical attention a data culture pays to the ways in which data practices are influenced by these subtle socio-cultural assumptions that enable some forms of data production and restrict others, and how they shape what data practitioners observe and desire to observe, how they perceive the relevance of the things they observe, whether they decide to try and capture these observations as data, and if so, how they influence how that data gets produced, processed, distributed and used.

As observed across diverse fields, from cartography to librarianship, critically reading the outputs of data practices as texts can begin to illuminate some of the hidden biases at play when people produce informational representations of the world, whether they be maps (Crampton and Krygier 2005), library catalogues (Bates and Rowley 2011), classification systems and standards (Bowker and Star 2000) or other forms of information resource. Often these hidden biases go unnoticed both within data cultures and amongst users of their products. Data produced by experts in trusted institutions such as national statistics offices are often assumed to be objective and complete representations of the world (Burkert 1992), with little recognition that those data are a representation of the world from a particular perspective. Whilst the perspective offered might be good enough to meet the needs of some data cultures, it may simultaneously be limited from the perspective others, particularly less powerful, marginalized subjectivities. As an example, it is clear that there is an often unacknowledged masculine culture driving many sites of data practice that are engaged in projects aimed at the datafication of cities (e.g. local authority technical teams, consultancy firms, open data groups, data analytics companies and academic researchers etc.). Where such demographic biases of gender, class, race and so on are perceived in a data culture, it is especially important that critical attention is directed at how dominant subjectivities may be becoming embedded into the data and data-driven insights that these cultures are producing.

### *Perceived value of data practices*

What value a data culture perceives in producing, processing, distributing and using data is also a product of its socio-material context, and will shape decisions about investments of time and money in data practices. A data culture's perception of the value of their and others' work can give insight into the underlying drivers behind data practices and forms of data governance, as well as how the data culture relates to other sites of data practice and governance, and the wider socio-material context. Different value frameworks may place greater or lesser emphasis on factors such as scientific, commercial, economic, social and cultural value of data practices.

In many data cultures, emphasis is placed upon the commercial and economic value of data practices, for example their contribution to generating private profits, research funding, economic growth or organizational efficiencies and cost-savings (e.g. see Manyika *et al.* 2013; Kitchin 2014a). At some sites of data practice, for example, in private firms and some parts of the state, extracting economic value from data is likely to be a primary concern, albeit tempered in some cases with consideration of social, ethical, environmental and other values. Similarly, at sites of data governance, creating a policy environment that promotes data's contribution to economic growth might be of central importance. However, at other sites the relationship between different value systems can be more complex, and not immediately discernible to either participants or researchers. For example, surface level political or social values may obscure underlying economic drivers, as seen in some forms of data-driven corporate well-being initiatives or data-driven innovations in public service design which are framed by a need to respond to deep public spending cuts. Similarly, participants in a data culture might recognise and welcome the economic value of data as important given these socio-material contexts they encounter, but ultimately see this value as secondary to other forms of value production that they are engaged in, for example, social, cultural, political. Observations of data cultures shaping urban spaces in the UK suggest that many people perceive that data analytics offers a valuable opportunity for tackling the complex challenges faced by public organizations and citizens in the context of late neoliberal capitalism, particularly the challenge of providing more personalized services, to more people, for often significantly less money.

The development of urban dashboards that present a range of quantified metrics about the functioning of a city are one way in which people are using data to inform decisions and drive efficiency savings within such a context. Whilst used differently in different cities, in some cities these dashboards are being shaped by data cultures in an effort to supplant more complex, messy and situated forms of knowledge with the cleaner, more efficient quantified metrics offered by dashboards, without full consideration of their relationship with social reality or their efficacy for addressing the complex challenges cities currently face (Kitchin *et al.* 2015). Similarly, some smart city initiatives aim to utilize data analytics and other emergent technologies supplied by the global technology industry in order to drive efficiencies in urban systems and management, without critical examination of the underlying drivers for such efficiencies, or their compatibility with the development of sustainable, ecologically sound forms of political economy (Sadowski and Pasquale 2015).

Differences in perceptions of what is important and valuable across sites of data practice can also impact upon the production of data about cities, and feed into how people choose to respond to encroaching forms of 'dataveillance' (Clarke 1988). In sectors with heavily quantified audit cultures, it can be observed that the value of such data production and processing is seen differently by different actors, and that power relations between such actors play out in people's data practices. For example, people will fail to provide accurate data for a variety of reasons including resistance or apathy towards requests, attempts to game the system, or simply an inability to produce the data being demanded. These subtle forms of struggle and resistance identifiable at such sites impact upon the representation of social reality provided by the data they generate. Similar forms of agency are also seen when people engage in activities that manipulate the data traces that their digital devices generate, for example through selectively limiting or subverting the production of accurate by-product data that capture particular forms of behaviour. These various biases and exclusions have a significant impact upon data-driven representations of cities that are increasingly informing various forms of urban governance and decision-making.

## Data cultures, power and the city

As data-driven insights, decision-making and automation become more deeply embedded in the development and governance of cities, it is important to step back and address critical questions about what sort of future these practices are in the process of creating. The consequences of our data practices are not independent of the complex and contested socio-material contexts from which they emerge. The nature of these consequences is the product of people's interrelationships with the socio-material conditions they encounter as they think about and work with data. At various sites of data practice, people are attempting to extract different forms of value from data in an effort to prosper, survive, understand, engage with and explain the world around them given the socio-material conditions that they find themselves within and which, in some cases, are attempting to ameliorate. It is in these spaces, where structure meets agency, that the data cultures that frame and give justification for data practices coalesce.

Through illuminating the nature of these data cultures, we can begin to understand how power is being worked out in data-driven urban developments. We can observe the ways in which subtle biases and philosophical assumptions arise and advance within and across data cultures, and how they become embedded into digital artefacts such as urban dashboards and open data platforms that are increasingly being used to inform urban processes and developments. We can also recognize how different types of value framework influence the development of data practices, and better understand how the material conditions of data's production, processing and use, for example external funding priorities and the need to be commercially competitive, frame and give justification for data practices, and whether and how social, ethical and cultural factors are taken into consideration when economic considerations are driving practice.

It is important that rather than focusing predominantly on what these data-driven platforms and initiatives can tell us about cities, we also read them for gaps and silences, biases and underlying agendas. In this way, we can bring to the surface the ways in which these data-driven accounts are always partial, and always framed by both the limited possibilities of data and the particular subjectivities of the data culture that generate and process them. In so doing, we can draw attention to the data-driven gaze of these cultures of data practice, and increase understanding of the implications of such biases on the social world.

It is important to address such questions, not because the cultures of data practice are interesting for their own sake, but because they are shaping the world we live in. In asking such questions of the data cultures that are responsible for generating and using these data-driven insights about cities, we can begin to illuminate how power relations are being reproduced, disrupted, hidden and made visible through these practices, and ultimately contribute to the development of more critical and reflexive forms of data practice.

## Acknowledgements

This chapter is in part based on research conducted on the Arts and Humanities Research Council (UK) funded project 'The Secret Life of a Weather Datum' (AH/L009978/1).

## Notes

1 www.rd-alliance.org/bof-session-data-across-disciplines-ethnographic-project-understand-diverse-data-cultures-practices.

2 www.microsoft.com/en-gb/enterprise/event/microsoft-data-culture-series.aspx#fbid=Hwa_58PJp9L.

## References

Archer, M. (1998) 'Introduction: Realism in the social sciences', in M. Archer, R. Bhaskar, A. Collier, T. Lawson and A. Norrie (eds), *Critical Realism: Essential Readings*. London: Routledge, pp. 189–205.

Bates, J. (2013) ‘The domestication of open government data advocacy in the United Kingdom: a Neo-Gramscian analysis’, *Policy and Internet* 5(1): 118–137.

Bates, J. and Rowley, J. (2011) ‘Social reproduction and exclusion in subject indexing: a comparison of public library OPACs and Library Thing folksonomy’, *Journal of Documentation* 67(3): 431–448.

Batty, M., Axhausen, K.W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G. and Portugali, Y. (2012) ‘Smart cities of the future’, *The European Physical Journal Special Topics* 214(1): 481–518.

Bowker, G. (2000) ‘Biodiversity datadiversity’, *Social Studies of Science* 30(5): 643–683.

Bowker, G. and Star, S. (2000) *Sorting Things Out: Classification and its Consequences*. Cambridge, MA: MIT Press.

British Academy (2012) ‘Society Counts – Quantitative Studies in the Social Sciences and Humanities’, A British Academy Position Statement available from: www.britac.ac.uk/policy/Society_Counts.cfm [accessed 9 December 2016].

Burkert, H. (1992) ‘The legal framework of public sector information: Recent legal policy developments in the EC’, *Government Publications Review* 19(5): 483–496.

Cabinet Office (2015) ‘Open policy making toolkit: Data science’, available from: www.gov.uk/open-policy-making-toolkit-data-science [accessed 9 December 2016].

Clarke, R. (1988) ‘Information technology and dataveillance’, *Communications of the ACM* 31(5): 498–512.

Crampton, J. and Krygier, J. (2005) ‘An introduction to critical cartography’, *ACME: An International E-Journal for Critical Geographies* 4(1), available from: http://ojs.unbc.ca/index.php/acme/article/view/723/585 [accessed 9 December 2016].

Dourish, P. and Mazmanian, M. (2011) ‘Media as material: Information representations as material foundations for organizational practice’, Working Paper for the Third International Symposium on Process Organization Studies, Corfu, Greece, June, available from: www.douri.sh/publications/2011/materiality-process.pdf [accessed 9 December 2016].

Edwards, P. (2010) *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.

Gaventa, J. and Cornwall, A. (2008) ‘Power and knowledge’, in P. Reason and H. Bradbury (eds), *The SAGE Handbook of Action Research: Participative Inquiry and Practice*. London: SAGE, pp. 172–189.

Greenwald, A. and Krieger, L. (2006) ‘Implicit bias: Scientific foundations’, *California Law Review* 94(4): 945–967.

Hall, S. (1987) ‘Gramsci and us’, *Marxism Today*, June, pp. 16–21.

Hollands, R. (2008) ‘Will the real smart city please stand up?’, *City* 12(3): 303–320.

IBM (2013) ‘What is big data?’, available from: www-01.ibm.com/software/data/bigdata/what-is-big-data.html [accessed 9 December 2016].

Jones, P.D., Lister, D.H., Osborn, T.J., Harpham, C., Salmon, M. and Morice, C.P. (2012) ‘Hemispheric and large-scale land-surface air temperature variations: an extensive revision and an update to 2010’, *Journal of Geophysical Research: Atmospheres* 117(D05127).

Kitchin, R. (2014a) *The Data Revolution*. London: SAGE.

Kitchin, R. (2014b) ‘Making sense of smart cities: addressing present shortcomings’, *Cambridge Journal of Regions, Economy and Society* 8(1): 131–136.

Kitchin, R. (2014c) ‘Big data, new epistemologies and paradigm shifts’, *Big Data and Society* 1(1): 1–12.

Kitchin, R., Lauriault, T. and McArdle, G. (2015) ‘Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards’, *Regional Studies, Regional Science* 2: 1–28.

Lauriault, T. (2012) ‘Data, infrastructures and geographical imaginations: Mapping data access discourses in Canada’, PhD thesis, Carleton University, Ottawa.

Manyika, J., Chui, M., Farrell, D., Van Kuiken, S., Groves, P. and Almasi Doshi, E. (2013) ‘Open data: Unlocking innovation and performance with liquid information’, McKinsey Global Institute, available from: www.mckinsey.com/business-functions/business-technology/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information [accessed 9 December 2016].

Massey, D. (1994) *Space, Place and Gender*. Minneapolis: University of Minnesota Press.

Mattern, S. (2015) ‘Mission control: A history of the urban dashboard’, *Places Journal*, March, available from: https://placesjournal.org/article/mission-control-a-history-of-the-urban-dashboard/[accessed 9 December 2016].

Ruppert, E. (2013) ‘Rethinking empirical social sciences’, *Dialogues in Human Geography* 3(3): 268–273.

Sadowski, J. and Pasquale, F. (2015) ‘The spectrum of control: a social theory of the smart city’, *First Monday* 20(7), available from: http://firstmonday.org/ojs/index.php/fm/article/view/5903/4660#p3 [accessed 9 December 2016].

Viseu, A. (2015) ‘Caring for nanotechnology? Being an integrated social scientist’, *Social Studies of Science* 45(5): 642–664.

World Economic Forum (2016) ‘The Fourth Industrial Revolution: What it means, how to respond’, available from: www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond [accessed 9 December 2016].

# 15 Where are data citizens?

*Evelyn Ruppert*

## Introduction

I pose the question of where are data citizens in two senses. The first concerns conceptions of and relations between 'online' and 'offline' lives, and 'virtual' and 'real' spaces. To pose the question of where are data citizens involves problematizing these conceptions and then asking where is the space of their becoming? The second concerns conceptions of agency and power and assumptions that subjects are either controlled or free. To pose the question of where are data citizens involves problematizing this binary and then asking through what processes do data subjects become data citizens? Those are my questions for this chapter and they start from the proposition that studies of the internet and empirical analyses of specific digital platforms are proliferating, yet we lack concepts for framing and interpreting what these mean for political subjectivities and their relation to the data that are generated and interpreted. My objective is to provide a theory of digital acts and digital citizens that builds on this proposition to provide an approach for more detailed empirical investigations of data, citizens and the city.

The claim of this chapter is that to know and understand cities through data requires a conception of who and where are digital subjects and the power relations through which they become digital citizens and data about them are generated. The data of cities are ultimately the outcome of relations and struggles between and amongst people and technologies, relations and struggles that bring into being both digital citizens and what I will define as cyberspace. Such a move is politically important in the context of current debates about digital data rights such as privacy and anonymity. While important these debates usually do not attend to the various power relations through which that data are constituted. It is also conceptually important in light of questions about the making and meaning of digital data for governing, research and other uses. To address these debates and questions we need a political and conceptual understanding of the acting subject and the power relations of which she is a part in the making of data of the city. Thus, rather than investigating digital rights in terms of their substance my focus is on who is the subject of these rights, or more precisely, who are constituting themselves as political subjects of especially data rights.

The consideration of rights is important especially in relation to critiques of data-driven urban governance and what is commonly termed ‘smart cities’. As other contributors to this collection note, rather than neutral, data of the city are political in myriad ways from questions of data ownership and control to those of privacy and surveillance. While critically attending to how data of the city is generated through socio-technical arrangements of usually ‘interested’ and ‘powerful’ political actors, what is marginalized or not elaborated is how we might understand the acting subject of these arrangements. That is, beyond noting the arrangements, contexts and provenance of data, how might we conceive of the role of subjects in the making and shaping of that data? That is what I bring into focus through a conceptualization of the power relations that bring both digital citizens and cyberspace into being. It is a conceptualisation that draws from and builds on Isin and Ruppert’s (2015) book, *Being Digital Citizens*.

## Who is the digital subject? Who is the digital citizen?

During the 1990s and early days of the internet, ‘cyberspace’ was heralded as a new space and it enjoyed considerable popularity to describe being ‘online’ (e.g. Barlow 1996; Gibson 1984; Katz 1997). While variously defined, it has continued to function in contemporary discourse as evident in its many derivatives such as cyberwar, cybersecurity, cyberfeminism and cybercrime. It has also been subject to critical debate and resignification especially by feminist scholars for its connection to cybernetics (e.g. Braidotti 1996; Haraway 1991). While it has somewhat fallen out of fashion and concepts such as online versus offline have become more common, I think it is useful to return to some of the promises in early conceptions of cyberspace. As I will argue, I will (re)appropriate the concept and its early conceptions because they expressed two possibilities that I think are worth reclaiming: the first is a conception of cyberspace as space of relations and the second that its subjects can also be understood as citizens.

Cyberspace was initially conceived in literary texts as an-other world (e.g. Gibson 1984) and in political texts as an independent space where ‘digital citizens’ were inventing new ways of conducting themselves politically. Rather than emphasizing technology early debates focused on its inhabitant as a digital citizen. Indeed, for some authors such as Jon Katz (1997) the internet marked the birth of this new political subjectivity. He thought that although digital citizens were libertarian, they were not alienated nor were they isolated but instead they made up a political movement and a common cause based on values of sharing, prosperity, exchange, knowledge and openness (Katz 1997). In another well-known declaration, John Parry Barlow (1996) of the Electronic Frontier Foundation imagined a space without social distinctions and which anyone could enter. Early and well recognized books contributed to such euphoria such as Sherry Turkle’s (1995) *Life on the Screen* and Nicholas Negroponte’s (1995) *Being Digital*. Both celebrated the digital lives of sovereign subjects who were understood to be doing things through the internet. Such optimism of course has not translated into experience. Rather, we are now inundated with determinist analyses that imagine people as relatively passive subjects who participate ‘online’. Notably, Sherry Turkle (2011) no longer celebrates but instead critiques the internet for isolating people from more meaningful and ‘real’ face-to-face human interactions such that especially young people are now ‘alone together’.

Numerous other popular critiques such as Nicholas Carr’s (2010) *The Shallows* and Evgeny Morozov’s (2011) *Net Delusion* also critique digital lives. While such declarations have been a good correction to utopian visions, they have replaced sovereign subjects with obedient ones. In this way, they reflect a reversal of the understanding of power advanced in modern political theory, which posits a divide between modernity and tradition where a subject to power (tradition) was replaced by a subject of power (modernity). However, such binaries are problematic as critical political theorists contend. Indeed, both the divide and its displacement need to be questioned and in its place an understanding that captures subjects as composites of multiple forces, identifications and associations. This is a very brief summation of an understanding of power advanced by Etienne Balibar (1991) through his reading of Michel Foucault. Balibar argued that being a subject to power involves domination by and obedience to a sovereign whereas being a subject of power involves being an agent of power even if this requires participating in one’s own submission. He conceived of the citizen as not merely a subject to power or subject of power but one who embodies both. But critically it is through this combination and the subject’s participation in submission that the possibility of subversion is made possible and this is what distinguishes the citizen from the subject: she is not already formed and inhabited by external forces but a composite subject of obedience, submission and subversion where all three are always-present dynamic potentialities. This is an understanding advanced in critical citizenship studies, which I discuss later in this chapter.

This critical conception of political subjectivity moves away from asking how subjects are being ‘liberated’ or ‘controlled’ in relation to the internet to instead inquiring into the complexities of ‘acting’. Who then is the digital citizen? She is both a subject to and of power that comes into being by acting in relation to the mediations, regulations and monitoring of the platforms, devices and algorithms or, more generally, the sociotechnical conventions that format, organize and order what we do, how we relate, act, interact and transact through the internet. This is the configuration of power relations and, in short, the meaning of what I will refer to as ‘acting through the internet’. It is these complexities of acting and their relation to political subjectivity that are often ignored when digital data generated by devices and platforms are harvested and interpreted.

Where then is the digital citizen located when she acts and when data about her acts is generated? Where is she when she acts through the internet? In short, as I will elaborate below, she is an embodied subject who is both in and part of the relations that bring cyberspace into being. Rather than networks or arrangements, cyberspace can be understood as a space of relations of embodied subjects who act through the sociotechnical arrangements that make up the internet. That in brief is the argument I will develop in the next section.

## Where is the digital citizen?

As noted, cyberspace has been understood as independent and separate from 'real' space. This dominant understanding has been challenged by critical scholars such as Dodge and Kitchin (2001), who argue against not only utopian and determinist claims but also social constructivist and political economic analyses, and Jayne Rodgers (2003) who also complicates this view. Much more has been written on cyberspace since the 1990s; here I will summarize two recent examples from legal scholars since their focus on cyberspace concerns the rights and agencies of subjects, which is the concern of my argument.

Julie Cohen (2007) summarizes and challenges metaphoric uses of cyberspace especially for how they have treated it as a separate space that is different (exceptional) or the same (unexceptional) as physical space. Cohen is a law professor who critiques exceptionalist understandings of cyberspace that have formed the basis of cyberlaws. As she notes, while legal scholars no longer accept that cyberspace is more 'free' than 'real space', legal theories have been classified in ways that treat 'cyberspace' as different from 'real space' and this has come to affect the formulation of legal rules. However, while critical of the uses of the term, she implicitly if not inadvertently accepts a distinction between cyberspace and an ostensibly 'real' space. Cohen rightly notes that '[c]yberspace is in and of the real-space world, and is so not (only) because real-space sovereigns decree it, or (only) because real-space sovereigns can exert physical power over real-space users, but also and more fundamentally because cyberspace users are situated in real space' (Cohen 2007: 217–218). She concludes that:

> theories of cyberspace as space fail not because they lack the proper understanding of whether 'cyberspace' is different from 'real space,' and indeed that debate simply muddies the issue. Rather, they fail because they lack appreciation of the many and varied ways in which cyberspace is connected to real space and alters the experience of people and communities whose lives and concerns are inextricably rooted in real space.
>
> (Cohen 2007: 225)

She describes this connection as an interplay between real and digital geographies. While usefully challenging the separation, she nonetheless continues to maintain it by thinking of cyberspace as connected to and interacting with embodied and lived spaces and through her continued use of online and offline to describe the two in her later work (Cohen 2012).

Another legal scholar, Laurence Lessig, while routinely questioning the uses of cyberspace, similarly maintains a separation by insisting that cyberspace is fundamentally apart from and exceptional to other spaces. In 1996 he considers the internet and cyberspace to be more or less the same thing and uses the metaphor of cyberspace to understand the internet and the ways in which it is different from what he called 'real' or non-virtual space (Lessig 1996). What Lessig suggests is that cyberspace constitutes a new mode of power: one submits to code.

In this way, he attributes a sovereign power to code and it is because of code that freedom is lost in cyberspace. He later develops a slightly more nuanced idea of the difference between cyberspace and the internet, yet he still insists on a basic difference between cyberspace and real space (Lessig 2006). Lessig thinks cyberspace, like geographic space, has architecture, and this architecture is the code: algorithms govern hardware and software switches and regulate access to its specific zones. The difference between 'real' space and cyberspace is that real space is structured around public spaces that have access to everyone. By contrast, cyberspace includes many zones that are off limits to many and is constituted by code, which means 'You can resist this code – you can resist how you find it, just as you can resist cold weather by putting on a sweater. But you are not going to change how it is' (Lessig 2006: 93).

While Cohen affords more agency to the subject and her understanding of power is less determinist than Lessig's, both conflate the internet with cyberspace and maintain a separation between cyberspace and real space. However, the internet and cyberspace are not the same and their conflation is problematic. By elaborating the meaning of 'acting through the internet', I distinguish between the two in a way that is also helpful for transcending a separation between offline and online.

As advanced by others, the internet is an interconnected network of computers (and information and communication technologies (ICTs)) using standard and negotiated protocols to transmit information converted into binary numeric form known as digital objects which can be sounds, images, words or numbers (Deibert 2009). It includes governments, corporations and organizations that own and operate infrastructures that transmit digital objects as well as internet service providers (ISPs) who own and operate infrastructure that connects users to the internet; software such as operating systems, code and cryptography; and hardware such as routers, switches, cables, transmitters, receivers and servers. Such infrastructures also include all of the people who maintain, operate and configure them, as DeNardis (2012) notes in a useful description of its complexity and layers. However, the internet is only one part of the relations that make up cyberspace; Dodge and Kitchin (2001), for example, argue that ICTs that make up the internet support a cyberspace, a conceptual space that extends the relationship between people and place. However, it is more than a conceptual space. The people who relate to the internet are embodied subjects who act through its sociotechnical arrangements that are made up of conventions that include humans, devices, norms, values, affects, laws, ideologies and technologies. Conducting ourselves means to act with others as we take up and establish our social positions – something that Foucault captured by defining power as 'action upon action' or 'conduct of conduct'. We are not online or offline but part of a space of relations that we bring into being as embodied subjects who act through the internet. Cyberspace is thus the outcome of subjects' relations to and struggles with the conventions that make up the internet, struggles that also occur in relation to the actions and struggles of other subjects. It does not pre-exist these struggles but it is through those power relations and struggles that cyberspace comes into being as a contested space.

What then are the forces that configure the complexities of acting and also in turn the production of cyberspace? Here Henri Lefebvre's (1991) work, which has been advanced by many critical geography scholars, continues to provide a conceptual approach for conceiving of these forces and for undoing the separation between online and offline worlds. Lefebvre elaborated three registers involved in the production of space – conceived, perceived and lived. Conceived spaces are rendered by objectifying practices that code, recode, present and represent space to make it legible and intelligible. Perceived spaces are symbolic representations that guide imaginative relations to space. Lived spaces are those that we inhabit through the things we do in or by living and are the spaces through which we act. Lefebvre argued that we experience being-in-the-world through these simultaneous but asynchronous registers that are distinct yet overlapping and interacting: by inhabiting them, we produce them.

Scholars who study cultural, social, legal, economic or political spaces also advance such an understanding of the relations that make up these different spaces, such as Bourdieu's (1988) discussion of social space. Their assumption is not that such spaces are separate and independent from other spaces people inhabit. To speak of a social space, for example, is an analytic means to concentrate on the subset of relations that make up this space and thereby open it up to a deeper understanding of how people inhabit it as simultaneously conceived, perceived and lived. The point is that just as critical geographers understand geographic space as not only physical but involving three registers in its production, so, too, can we understand cyberspace. But rather than registers, these three can be conceived as subjectifying forces that are neither sequential nor parallel but simultaneous and intertwined relations of power. The *legality* of cyberspace consists of the forces of rules, regulations and other codes that govern (or attempt to govern) it and which are often unseen but embedded in the organization and requirements of sociotechnical arrangements. The *performativity* of cyberspace is the lived spaces that are brought into being by subjects who act through the internet and bring it alive as it were. *Imaginaries* then are the perceived images, ideals and ideologies of cyberspace that organize how we think, what we desire and want it to be. Like other social spaces then, cyberspace is not designed and arranged and then experienced by passive subjects but is brought into being through acting bodies and the interplay of legal, imaginary and performative forces.

This is different from dominant conceptions of online versus offline worlds. The difference is exemplified in how the role of social media has been interpreted in relation to protests, uprisings and occupations staged in cities. Christian Fuchs (2014), for example, outlines four ways the role has been interpreted – from social media as technological determinants, to their having no impact, or to being just 'useful'. Instead he offers a fourth interpretation, that the subjects of these events had already been formed into social groups able to recognize each other and it is that formation and recognition that enabled them to mobilize each other for action through social media. In other words, the objective conditions that led people to protest found mechanisms for expressing subjective positions, thereby helping organize these protests. However, while Fuchs is critical of assuming that a space exists that is separate and independent of squares, he does not offer a way of conceptualizing this space. Cyberspace is that space that comes into being by relations between and among embodied subjects who act through the internet. These bodies can be collective (institutions, organizations, corporations, groups), cybernetic or social, as advanced by Judith Butler (1993) and Donna Haraway (1991). As both authors contend, bodies as inherently social and not prior to their socialization; collective, technological and biological bodies are all social bodies. Embodied subjects acting through the internet are engaged in struggles that are no less or more 'real' than those that occur in social space or cultural space which are indeed inextricably connected to and through embodied subjects.

## How do digital subjects become digital citizens?

How then does the digital citizen come into being? Building on the conception of power outlined above, they are subjects who act in ways that submit to but also go beyond and transgress the conventions of the internet. In doing so they are not simply obedient and submissive but also subversive in the making of cyberspace. This follows a conception of the citizen advanced in critical citizenship studies, which positions the citizen beyond its modern configuration as simply a member of the nation-state (Clarke *et al.* 2014). Instead citizens are understood as subjects who make rights claims by contesting or struggling against existing regimes such that citizenship is a site of contestation rather than made up of bundles of given rights and duties.

How then do subjects make digital rights claims and become digital citizens? Words are of course one way that they make claims to rights such as speech, access and privacy. As John Austin famously argued, language is a means of social action: it can be performative such that people *do* things with words (Austin 1962). However, making claims was not one of the five classes of speech acts (judgments, decisions, commitments, acknowledgements and clarifications) that he identified as having performative force. Claims are thus a sixth speech act and key to the becoming of a citizen. To be sure subjects make rights claims through what they say as many individual and collective declarations attest. Chelsea Manning says, 'We're citizens, not subjects. We have the right to criticize government without fear' (Manning 2015). While these words may not have legal, if not performative force, their imaginary force can be powerful. There are many other examples of how subjects make claims such as those who call upon authorities to inscribe digital rights through regulations and legislation and give them legal force. The declaration of the UN World Summit on the Information Society, and the International Charter of Human Rights and Principles for the Internet that followed, are two examples of declarations. They pronounce digital rights claims such as the right to access, liberty, security and freedom of expression, or right to information, freedom from censorship or hate speech to right to privacy and data protection and many more (see Franklin [2013] for a detailed account of these declarations and claims). Collectively, they not only create a cumulative force but also disseminate this force through the internet.

However, what subjects do through the internet involves not only doing things with words but also the reverse of Austin's principle: they also say words with things of the internet. Bruno Latour (2000), while not referencing Austin or the internet, reversed the principle to state that people 'do words with things' and that this is also a social activity. In relation to acting through the internet, subjects challenge, subvert or resignify conventions through their deeds, by doing words with the things that make it up. From downloading, uploading, forwarding and blocking to encrypting and cloaking their actions, subjects make claims to rights such as to access, share or make private what they do through the internet and in making these claims they become citizens. While whistleblowers and hactivists are often highlighted as the vanguards of digital rights, there are many more political subjects of the internet who not only make rights claims by saying things but also by doing things through the internet. The everyday social life of communicating, interacting and networking therefore can be understood as part of the struggles and contestations over the emergence of this new political subjectivity, that of the digital citizen. So, when we study conventions such as microblogging we can ask: how do such platforms both configure everyday social actions and at the same time create possibilities for subjects to appropriate the platform in creative (and possibly subversive ways) and act differently and become digital citizens? What are the possibilities of thinking, speaking and acting differently, of challenging and resignifying conventions of the internet and thereby enacting digital rights through what we do and not only say?

Taken together subjects make rights claims both in and by what they say and what they do through the internet – their digital acts, be they virtuous, malicious, righteous or indifferent – and it is through making claims that they move from being data subjects to being digital citizens.

## Where are data citizens?

The question that I posed for this chapter has two meanings. It asks about the space of relations of which data citizens are a part and at the same time about their absence in much current work concerning the data of the internet. We know about the power of platform owners, corporations and governments and ways in which data are being commodified, traced, analysed and traded; how issues and concerns about data ownership, privacy and protection are being debated and contested; how people use various platforms to *do* politics such as organize and mobilize political protests and engage in citizen journalism or forms of digital activism; how the sociotechnical arrangements that make up the internet seek to organize what people do and format the data that is generated; and how internet data are being analysed by corporations, governments and researchers to enact and know social worlds in ways that are challenging other sources and methods. But what do we know about the subjects and citizens of the internet?

I have argued that we know a lot about the internet and have many good critical investigations, but we have yet to provide a theoretical conception of who are the political subjects and citizens of data. One conception we do have is that of citizen scientists who challenge dominant data regimes by formulating, organizing and operating their own data gathering devices and platforms. As such they constitute one kind of data citizen, one who expresses rights to data through what she does by engaging subjects and enabling them to shape and influence how data are generated. Yet such a conception does not attend to the varying subject positions that they take up and the ways of acting that such devices and platforms also configure and bring into being. To turn instead to what subjects are saying and doing through the internet draws attention to what varying acts in relation to platforms and devices mean for the data that are generated. Bringing the political subject into the centre of concern interferes with determinist analyses of these data and hyperbolic assertions about its emancipatory potential or impact, both which tend to imagine subjects as passive data subjects. It also interferes with interpretations of data that implicitly assume that they are a simple description of the action and behaviour of subjects. Instead, I have offered a way to investigate how subjects are composites of obedience, submission and subversion and that this matters in the making of data. Rather than singular, a range of subject positions can come into being through the interplay of imaginary, legal and performative forces.

As Lefebvre theorized, urban spaces are composed of struggles involving rules and laws, ideals and the practices and experiences of subjects. Many studies of urban spaces have taken this up to reject a flat ontology that posits the existence of an objective, natural, or physical space separate and independent from represented or lived spaces (Soja 1996). What they argue is that when the conduct of subjects in urban spaces are observed, recorded and analysed – from the everyday to protests and occupations – such conduct can be interpreted as the product of the interplay of these forces. Nicholas Blomley (2004; 2011), for example, has advanced a similar understanding in his analyses of how urban spaces are enacted and ‘taken’. The same can be said of cyberspace and the conduct of subjects. The data that are generated are also products of the actions, interruptions and appropriations of subjects.

How subjects act is not only in relation imaginary and legal forces but also performative ones that involve the recitation, repetition and invention of new conventions and meanings. Consider how specific internet platforms shape but are reformed and transformed through the performance of subjects, their repeated actions, demands, interruptions and appropriations and who do not just follow but play with and subvert or seek to modify their conventions. Tweeters have, for example, transformed the meaning of ‘what’s happening’ to ‘what do we want to make happen’ through their calls to action and protest. Tweeters are embodied subjects acting among and with other bodies through the internet and that acting involves not simply following or being determined by the technical configurations and conventions of the platform. They act in relation to rules and laws about what is sayable (e.g. as recent rulings on racist, offensive or misogynist tweets attest) or respond to imaginaries of the internet (e.g. libertarian ideas). And it is through their recitation, repetition and invention of new conventions that they engage in a whole series of relations as embodied subjects that act with others.

The data of the internet are not apart from these relations and forces; how to attend to these and their consequences for the interpretation of data is one of our many empirical challenges. Interrogating the multiple ways that subjects act through the internet and as the product of the play of forces and diverse subjectivities would be one approach. For example, with ever more tracing and tracking and selling and trading of data, how are subjects making rights claims by taking actions such as blocking and filtering, encrypting communications, creating multiple identities, deploying bots, gaming trending algorithms and so on? Do we ignore or account for them when we interpret data? Even for those subjects who obey and adhere to conventions, how do they act in multiple and not only preformed or expected ways? That is not to ignore dominant forms of subjectivation but even these call for closer analysis in relation to how subjects are formed and act. This is similar to an argument advanced by scholars of surveillance studies who have critiqued the strong tendency of researchers to focus on the impact of powerful institutional actors and relations of domination; however, in doing so conceptions of how this might be otherwise are not developed (Monahan *et al.* 2010). To reduce all forms of acting to one unexamined mode is to overlook the potentialities and possibilities of acting otherwise.

This challenge of how data are constituted is a question that surely can and has been asked of other forms of data. Survey data are often critiqued for being shaped by the way questions are formulated, the mode of asking questions (telephone/face-to-face interview, self-completion), the medium (paper, internet) and variations in the performances of subjects (memory, disinterest, misunderstanding, deceiving). Such questions can be raised about internet data when researchers seek to understand the relation between the configuring work of platforms and the actions of subjects. This is the case when researchers interrogate the multiple possible meanings and intentions behind the use of like buttons, search engine queries or locations saved on Google maps or when they seek to understand varying patterns in mappings of mobile phone data. In these instances, subjects do not merely follow but interpret, invent and circumvent ways of acting.

These questions are especially critical given the tendency to treat digital data as raw. We now have myriad studies that are challenging this by accounting for the relations between people and technologies that come to make it up. But we also need a conceptual framework to understand how these also involve power relations between and amongst embodied subjects and citizens who act through the internet and in doing so are part of the making of cyberspace and data through which we come to know cities.

## Acknowledgements

The framing that I develop in this chapter assembles a number of arguments and then builds on Isin and Ruppert (2015) *Being Digital Citizens*. Here I especially have elaborated some of the theoretical arguments in that book through a focus on what it means for data. Funding from a European Research Council (ERC) Consolidator Grant (615588) supported the research and writing of this chapter.

## References

Austin, J.L. (1962) *How to Do Things with Words*. Oxford: Oxford University Press.

Balibar, E. (1991) ‘Citizen subject’, in E. Cadava, P. Connor and J.-L. Nancy (eds), *Who Comes after the Subject?* London: Routledge, pp. 33–57.

Barlow, J.P. (1996) ‘A declaration of the independence of cyberspace’, available from: www.eff.org/cyberspace-independence [accessed 11 July 2014].

Blomley, N. (2004) *Unsettling the City: Urban Land and the Politics of Property*. London: Routledge.

Blomley, N. (2011) *Rights of Passage: Sidewalks and the Regulation of Public Flow*. London: Routledge.

Bourdieu, P. (1988) ‘Social space and symbolic power’, *Sociological Theory* 7(1): 14–25.

Braidotti, R. (1996) ‘Cyberfeminism with a difference’, available from: www.let.uu.nl/womens_studies/rosi/cyberfem.htm [accessed 3 September 2015].

Butler, J. (1993) *Bodies That Matter: On the Discursive Limits of ‘Sex’*. London: Routledge.

Carr, N.G. (2010) *The Shallows: What the Internet Is Doing to Our Brains*. New York: W.W. Norton.

Clarke, J., Coll, K.M., Dagnino, E. and Neveu, C. (eds) (2014) *Disputing Citizenship*. Bristol: Policy Press.

Cohen, J.E. (2007) ‘Cyberspace as/and space’, *Columbia Law Review* 107(1): 210–256.

Cohen, J.E. (2012) *Configuring the Networked Self: Law, Code, and the Play of Everyday Practice*. New Haven, CT: Yale University Press.

Deibert, R. (2009) ‘The geopolitics of internet control: Censorship, sovereignty, and cyberspace’, in A. Chadwick and P.N. Howard (eds), *Routledge Handbook of Internet Politics*. London: Routledge, pp. 323–336.

DeNardis, L. (2012) ‘Hidden levers of internet control’, *Information, Communication and Society* 15: 720–738.

Dodge, M. and Kitchin, R. (2001) *Mapping Cyberspace*. New York: Routledge.

Franklin, M.I. (2013) *Digital Dilemmas: Power, Resistance, and the Internet*. Oxford: Oxford University Press.

Fuchs, C. (2014) *Social Media, A Critical Introduction*. London: SAGE.

Gibson, W. (1984) *Neuromancer*. New York: Ace Books.

Haraway, D.J. (1991) *Simians, Cyborgs and Women: The Reinvention of Nature*. London: Free Association.

Isin, E.F. and Ruppert, E. (2015) *Being Digital Citizens*. London and New York: Rowman and Littlefield International.

Katz, J. (1997) ‘The digital citizen’, *Wired* 5(12): 68–82.

Latour, B. (2000) ‘The Berlin key or how to do words with things’, in P. Graves-Brown (ed.), *Matter, Materiality and Modern Culture*. London: Routledge, pp. 10–21.

Lefebvre, H. (1991) *The Production of Space*, trans. D. Nicholson-Smith. Oxford: Blackwell.

Lessig, L. (1996) ‘The zones of cyberspace’, *Stanford Law Review* 48(5): 1403–1411.

Lessig, L. (2006) *Code: Version 2.0*. New York: Basic Books.

Manning, C.E. (2015) ‘We’re citizens, not subjects. We have the right to criticize government without fear’, *The Guardian*, available from: http://bit.ly/1GORENO [accessed 10 May 2015].

Monahan, T., Phillips, D.J. and Wood, D.M. (2010) ‘Editorial: Surveillance and empowerment’, *Surveillance & Society* 8(2): 106–112.

Morozov, E. (2011) *The Net Delusion: How Not to Liberate the World*. New York: Public Affairs.

Negroponte, N. (1995) *Being Digital*. London: Hodder and Stoughton.

Rodgers, J. (2003) *Spatializing International Politics: Analysing Activism on the Internet*. London: Routledge.

Soja, E.W. (1996) *Thirdspace*. Oxford: Blackwell.

Turkle, S. (1995) *Life on the Screen : Identity in the Age of the Internet*. London: Weidenfeld & Nicolson.

Turkle, S. (2011) *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York: Basic Books.

# 16 Beyond quantification

## A role for citizen science and community science in a smart city

*Mordechai (Muki) Haklay*

## Introduction

The underlying assumptions of smart cities and the production of urban big data, namely the over-valuation of efficiency and productivity need to be examined and critically assessed (Su *et al.* 2011; Chourabi *et al.* 2012; Greenfield 2013; Nam and Pardo 2011). This requires an examination of how specific values get embedded in technologies deployed in the city. This chapter argues that human and environmental values should also be part of the design and implementation of smart city systems, especially since these systems influence the way cities operate.

A good starting point for unpacking embedded values is to take notice of how cities are portrayed within smart city discourses. Historically, cities have been portrayed as either tame or feral, as ordered and chaotic, as natural or engineered. Most often smart city discourses enlist these same tropes to promote the procurement of and investment in these technologies in order to support a particular political and developmental path. For example, the proliferation of closed circuit television (CCTV) cameras progressed from individual in-store cameras, to networked feeds from multiple cameras in public urban spaces, to the integration of image-processing algorithms such as number plate or face recognition software (Coleman and Sim 2000; Graham 2005), all in order to improve urban 'security and efficiency'. Such regimes of data collection are often glossed over and presented apolitically, framed as a form of urban 'intelligence' coupled with the 'smart' application of information and communication technology, such as environmental sensors, the measuring of city dweller digital footprints, analytical and statistical techniques, including the use of complexity modelling and advanced visualization. These smart city data and technology assemblages (see Chapters 2, 3, 4) are deployed with the ideal to promote efficiency, productivity and safety and to reduce uncertainty. Is this the future that citizens want? Is the future envisaged as one where efficiency and minimizing risk is considered more important than chance human encounters, imagination and dialogue; where the city is so numerically and technologically managed and where serendipity is considered a threat to efficiency?

Are proponents of smart city technology and data modelling designing a future in which the city and its citizens are 'all watched over by machines of loving grace' (Brautigan 1967)? Do city managers and politicians assume that the societal impacts of technology are benign and beneficial, while technology and data are value neutral? More critical and philosophical approaches (e.g. Feenberg 2002; Dusek 2006) have demonstrated that technologies have built-in values and can lead to the 'black-boxing' of ideologies and conceptions of how society should run, which then impact on daily life. This is the case for many smart city platforms, software and code (Kitchin and Dodge 2011). To return to the CCTV example, the location of cameras combined with advances in image-processing algorithms represent a specific conceptualization of place, where some places and people are worthy of protection and order and others are considered a threat to be managed and policed (see Fussey 2002).

This promise of the smart city is predicated on knowledge, information and data, such as that collected from sensors on personal devices and wearables as well as from environmental sensors installed by city governments, engineers and researchers. And like technology, data and their related algorithms are not neutral. Smart city data and related infrastructures pose deep epistemological and ontological problems as these promote a mostly quantitative approach to the study of societies, seemingly making the city 'knowable and controllable' (Kitchin 2014). Similar issues were debated in the wake of the first 'quantitative revolution' in the social sciences in the 1960s and 1970s (Marshall 2006). Smart cities in a sense represent a form of 'quantitative revolution 2.0' with many of the critiques (e.g. Greenfield 2013) resembling those seen in earlier discussions of positivism in geography and urban studies (Wyly 2014).

One of the central questions of this chapter is: in the context of smart cities how we can ensure that the computing and sensing abilities that are being developed are integrated with meaningful and purposeful social and communal activities? This question is explored paying special attention to the meaning given to the data that are collected. This is framed with respect to concepts from philosophy of technology and, in particular, the ideas of Albert Borgmann (1984; 1999; 2010) concerning *device paradigm* and *focal practices*. These are used as conceptual tools to assess if participatory sensing can form a more meaningful and complete approach to collecting data to produce a smart city, one which emphasizes social practices in addition to the technical and the quantitative.

## Device paradigm and focal practices

In the early 1980s, Albert Borgmann observed that modern technology tended to adopt a myopic 'device paradigm' in which specific interpretation of efficiency, productivity and a reductionist view of human actions was taking precedence over 'focal practices' which bring people together in a more meaningful way. For example, Facebook textual messages is one form of communication that demands attention from the sender and the receiver, but only for a fraction of time, while meeting a friend for coffee with phones switched off and put aside, and paying full attention to mutual needs for a chunk of time might more conducive to fostering companionship and a deeper more engaging social interaction. By reducing human interaction to moments of communication it can be argued that Web-based social networking offers a more 'efficient' way of maintaining social links. As Sherry Turkle demonstrated in *Alone Together* (2012), this reductionist view of social interaction is limited and, indeed, meaningful social relationships degrade and are being lost by relying on information and communication technology (ICT) as the main conduit for social relations. Consider, for example, the seemingly efficient sharing of images of grandchildren on Facebook, leaving them for anyone, compared to handwritten letters which includes a physical drawing from said grandchildren. The issue is not one of sentimentality only – for example, a synchronous engagement through a video call over Skype between the grandparent and grandchildren, which requires mutual presence and concentration, is qualitatively different from the Facebook sharing and fleeting 'likes'.

Borgmann's analysis is especially important to the question of technology and the city since he frames his investigations toward the development of a meaningful and fulfilling human life – thereby addressing the age-old philosophical question of 'the good life' within technological societies (Higgs *et al.* 2000; Verbeek 2002). He notes that modern technology operates by disburdening human effort from activities that are laborious and by doing so turns them into commodities. For example, a hearth, which requires a wood supply, a fire to be attended to and regular cleaning, is replaced with central heating, which is now controlled remotely from an app on a smartphone. While the very narrow result of 'warm and comfortable room' is achieved with both technological settings, something more profound happens. The hearth is a 'thing' that requires tending to and effort, it is also part of a group of objects and activities that are 'focal things' – things that facilitate wider human activity and make sitting in front of the fire especially meaningful in comparison to the 'device' of central heating, wherein the service of heating is commodified and easily accessible to the degree that it become invisible. Importantly, this move from 'things' to 'devices' is changing the way people relate to reality. 'Focal things' facilitate 'focal practices', such as gathering in front of the hearth in the evening and having a conversation about the events of the day – something that is lost with the convenience and availability of the app-controlled central heating system. A device paradigm is therefore a generalized trend in which technology promises to enrich and disburden people's lives. However, while successfully delivering on these promises it takes away a fuller engagement with others and the material reality. Bormann emphasizes that this is not to say that we should ignore the toil, efforts and the many costs of pre-technological generation of warmth, rather than to be aware of the central paradigm of modern technology, which commodifies and separates the ends (warmth) from the means. Such a separation opens the door for far-reaching manipulations of the means, and while enjoying the fruit of technology, we should consider its fuller societal impacts.

With ICTs, a device paradigm increases and, as the social networking platform companies have demonstrated, once other aspects of human life have been commodified (heat, housing, transport, communication), human relationships themselves are seen as ripe to be reduced to their technical essence and monetized.

This is what Borgmann (1999) predicted in his differentiation between natural and cultural information, which for him are focal things and practices, while technological information, which, like devices, can mislead us to think that, because it is available and easy to access, it makes the world knowable and controllable. We need to pay attention to three classes of information, which he defines specifically. According to Borgmann, *natural information* is the information that we receive from the natural world such as that received from a meandering river which direct us to walk a specific way or direction; *cultural information* is information that we used to construct reality, to act and do things in the world – music sheets for example are used to play music, a map is used to construct and plan a route; finally, he qualified *technological information* as something different – information as reality. This is when information claims to be such a detailed representation that it can replace reality for all intents and purposes, as demonstrated by the highly detailed images and visualizations in digital globes such as Google Earth. Borgmann's (1999) use of the words natural, cultural and technological is very specific and differs from their everyday use, and this difference is important. What is important to note is that according to his definition, technological information obfuscates our ability to understand the world and to deal with it in a meaningful way (see Sieber and Haklay 2015). 'Big data' that the smart city produces are a kind of technological information, claiming to make material and social reality transparent and knowable. However, by necessity they fail in this task since a perfect system that will include all the data from sensors are providing partial descriptions of the world will lead to a situation in which 'nothing any longer presents itself with any authority . . . Anything might as well be an impediment to inquiry' (Borgmann 1999: 177). In other words, if we are capturing reality fully, we are back in our starting point, trying to decipher signals from the overall cacophony and complexity of the city.

Borgmann is not being nostalgic or suggesting that we destroy our central heating systems. Instead he asks us to consider how technology is altering life and then find the ways to protect or restore the focal practices and things that we have lost. His approach opens the possibility to reform technology and to allow for a wider social discussion about its future directions and applications (see Feenberg 2002; Haklay 2013).

## Data creation as a focal practice in citizen science and participatory mapping

Cities offer opportunities for deep and meaningful, yet 'inefficient', human encounters – and we should be attentive to how technologies are developed, the assumptions that we put forward in support of them, and to ensure that those types of encounters remain plentiful.

Is it possible to nurture those types of connections within smart city agendas, either by subverting them or by using the data resources that are available? One such way is to use smart city assemblages of sensors, data sources and algorithms to address problems and challenges that individuals and communities are faced with in cities, such as urban agriculture, monitoring pollution or addressing energy use. Citizen science is a scientific practice where non-professional researchers are involved in the process of conducting research (Silvertown 2009), and it is a type of science which can insert agency and control into the smart city. It is possible to imagine groups coming together in an inclusive and open way, discussing urban issues they would like to address and using existing sources of data combined with their own reporting and analysis to address them.

The emergence of community/crowd/user-generated digital maps (Haklay *et al.* 2008) provide some evidence for activities that, at their worst, fall into the trap of a device paradigm and at their best demonstrate the potential of new focal practices that are facilitated by technology. Projects such as OpenStreetMap (OSM) (Haklay and Weber 2008) exhibit complex relationships between the contributor to the mapping product and the user of the map in terms of their understanding of data, as well as making decisions about what will be captured and how. For the OSM mapper, who is commonly interested in her local area and walks through it to record specific objects, the process of mapping is an example of a novel way to engage with the world (Budhathoki and Haythornthwaite 2013). In a project such as OSM, in which mappers state that their affiliation to the project is linked to the project's goal, which is the production of a freely available accurate digital map of the world (Budhathoki 2010), this is especially true, although there is some evidence that people who update Google Map Maker are also doing so because they identify an error in the map in their local area and are concerned with the way it is represented to the world. In both these cases, the process is about creating an empirical representation of reality in a digital format, of identifying a road or amenity in reality and creating a representation of it using the location information from a GPS receiver or identifying objects on detailed satellite images and describing them. Moreover, for the OSM mappers themselves, the process of mapping can become a focal practice. While a very small minority of the total volunteer mapper community attends meetings, for those who contribute significantly to these projects, face-to-face meetings and discussions about the practice of mapping are significant and meaningful events. Arguably, even the unruly and often impolite discussions on the projects' 'Internet Relay Chat' (IRC) channel or on mailing lists demonstrate how meaningful the activity is in the life of the mappers. The act of mapping itself can be an act of asserting presence, rights or expressions of personal belief in how the world should evolve and operate (see Gerlach 2015).

Even the solitary activity of a mapper, or a citizen scientist, can be deeply meaningful and transformative, as Russell (2014) described so vividly. Russell, a citizen scientist, shared her experience of deciding to study an unknown detail about the life of tiger beetles by studying them in the Gila River, near her home. The tasks that she took upon herself (and her family) included chasing beetles and capturing them, growing them in terrariums at home, dismembering some and analysing them under a microscope and so on. This quest was sparked by a statement from Dick Vane-Wright, then the Keeper of Entomology at the Natural History Museum, that: 'You could spend a week studying some obscure insect and you would know more than anyone else on the planet. Our ignorance is profound' (Russell 2014: 15). This is not only true about insects, or animals, but also the night sky, or our understanding of urban air pollution. Russell explored many other aspects of citizen science, from online activities to observing the changes in nature over the seasons (phenology) and recording wildlife footprints in the sand. Her love of nature in her area comes through in the descriptions of her scientific observations and also when she describes a coming storm or other aspects of her local environment. In her journey, she overcame difficulties from following instructions that seem obvious to professional entomologists, to figuring out what the jargon meant, to the critical importance of supportive mentoring by professional scientists. As her book title, *Diary of a Citizen Scientist: Chasing Tiger Beetles and Other New Ways of Engaging the World*, expresses, citizen science is a focal activity for those who participate in it.

Meaningful engagement can also be true even for what might seem, at first sight, to be the epitome of a device paradigm within citizen science activities – volunteer computing. Volunteer computing – the act of participating in a scientific project by downloading and installing software that utilizes the unused processing cycles of a computer or a smartphone – is the automation of the process of participating in a scientific project. Inherently, the level of engagement of the participants is assumed to be very low – merely downloading a piece of software and configuring it once in a while. Since 2010, I have been involved in volunteer computing as part of the IBM World Community Grid (WCG) project, as a way of experiencing volunteer computing on my work desktop, laptops and later on my smartphone. Even though I am only one of 378,000 participants in the project, I am part of the long tail – ranking 20,585 with my top contributions being for FightAIDS@Home and Computing for Clean Water projects.

The operation of WCG transformed my volunteering into a 'device' and it disburdened me from actively dedicating time to support the project. From time to time, I notice the screensaver on my computers and am pleased to see the IBM WCG icon on my smartphone in the morning, knowing that it has used time since being fully charged for some processing. I also notice it when I reinstall a computer, or get a new one, and remember that I need to reset it. I do not check my ranking, and I do not log-in more than twice a year to adjust the projects that I'm contributing to. I have therefore self-diagnosed myself as being a passive contributor in volunteer computing. When compared to Russell's experience, my participation in the WCG project would not be a focal practice (see also Nov *et al.* 2014).

But then came the downtime of the project on 28 February 2015. I missed an advanced message. When I looked at my computer that day, I noticed a 'No Work Available to Process' message. This eventually bothered me enough to check the state of processing on my smartphone, which was also not processing. I later searched the internet to find out what was going on with the system and discovered that the main site was down and continued to look around until I found a Twitter message announcing scheduled maintenance. Even so, I could not stop looking at the screensaver and was relieved when processing resumed. What surprised me about this episode was how much I cared about it. The lack of processing annoyed me enough to spend over half an hour trying to find out what was wrong. For that afternoon, and only for a short moment, volunteered computing got elevated from a device to a focal thing.

The difference between Russell's deep engagement and my fleeting WCG one can be associated with the nature of the data and information produced. For Russell, data were captured through the intimate connection with the tiger beetles and seasonal change in her area. In contrast, in the WCG project, I have no control over the data that are produced, nor will I have the ability to scrutinize them. These data are created by algorithms set by scientists who access to them and are able to control collection and use. My engagement with the WCG projects lacks a relationship with the data that are produced.

## DIY science as focal practice

The city is also a place for collective action and communal activities with good potential for developing new focal practices around data collection, processing and use. The degree to which users of existing technology are allowed to change the meaning of how it is used, or to apply it toward other unintended uses (Haklay 2013), is central to its potential to serve as a 'thing' and not only as a 'device'. This is especially true in the area of Do-It-Yourself (DIY) science.

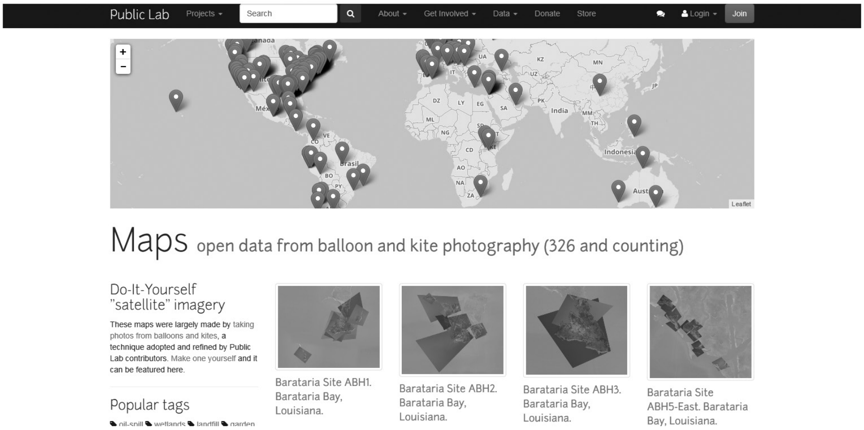
DIY science is emerging from the same technological trends that make the smart city a possibility, but with a fundamentally different ethos, focus and processes. The continuing decrease in the cost of electronics and sensors has enabled people from all walks of life to access and use devices either within embedded commodified devices, or as components that are ready for prototyping and experimentation. Consider, for example, the sensing ability of an average smartphone. It is, in effect, a sophisticated sensing machine, with sensors for sound (microphone), visible light (camera), location (GPS receiver), direction (compass), speed of movement (accelerometer), air pressure (barometer) and many other functions. The smartphone became widespread, the costs of sensors dropped and became widely available, and it is now possible to find a GPS chipset for less than $3. Sensors also appear in other industrial areas, such as the automotive industry and office machinery, and these also increasingly became cheap. These components provide the basis for new forms of DIY electronics where participants use open source licences, procedures and tools to share knowledge about the development of devices that can sense and act in the world. Combined with the growing availability of small-scale and local manufacturing facilities (known as fab lab, makerspaces and hackerspaces), technically able participants construct from these cheap components affordable sensing devices. The practice of sharing the code that drives the devices, as well as device blueprints, allows other people to take existing designs and adapt them to their own needs.

The Public Laboratory of Open Technology and Science (aka Public Lab) is a demonstration of this (Dosemagen *et al.* 2011; Wylie *et al.* 2014; see Figure 16.1). Born out of environmental activism resulting from the 'Deep Horizon' oil spill, Public Lab mixes online and offline communities of interest, in which members develop tools that can be used by any community to monitor different types of pollution and carry out various scientific investigations. Public Lab activities focus on:

> civic science in which we research open source hardware and software tools and methods to generate knowledge and share data about community environmental health. Our goal is to increase the ability of underserved communities to identify, redress, remediate, and create awareness and accountability around environmental concerns. Public Lab achieves this by providing online and offline training, education and support, and by focusing on locally-relevant outcomes that emphasize human capacity and understanding.
>
> (Public Lab 2015)

In practice, they rely on open hardware and software in which both the blueprints (in the case of hardware) and the code are available for anyone, free of charge, and open to modification. The technologies that they are developing are inexpensive (many well below $100) and, recognizing that not every community or individual would want to build the tools from scratch, they sell kits that can be used with detailed instructions provided on the Web. Finally, they encourage members to share their experience in developing tools through 'research notes' on the organization's website, as well as during an annual gathering that is called 'barn raising' after the communal practice of building a barn together.

A major theme of Public Lab activities is the development of very cheap aerial imagery tools. These enable participants to use a standard digital camera, plastic bottle, string and a balloon or kite to take highly detailed aerial photography. After the flight, images are stitched together and linked to existing geographical information using the 'Map Knitter' software, they can either be printed on paper or shared in Google Maps. Members of the Lab emphasize the value of mapping



Maps open data from balloon and kite photography (326 and counting)

Do-It-Yourself "satellite" imagery

These maps were largely made by taking photos from balloons and kites, a technique adopted and refined by Public Lab contributors. Make one yourself and it can be featured here.

Barataria Site ABH1. Barataria Bay, Louisiana.

Barataria Site ABH2. Barataria Bay, Louisiana.

Barataria Site ABH3. Barataria Bay, Louisiana.

Barataria Site ABH5-East. Barataria Bay, Louisiana.

*Figure 16.1* Public Lab map archive.

with balloons and kites, in which the operator is tethered to the data capture device as a demonstration of transparency in contrast to the hidden operators of CCTV, satellites or drones. This makes the act of capturing the imagery itself a purposeful demonstration of a civic data collection activity. The process of selecting images that will be used to create the mosaic, stitching the images and annotating the resulting maps was designed with tools to ensure that the data and information are owned by those who create them as well as expressing the message that they want them to convey. Thus, the balloons, cameras and software that are used within Public Lab activities are focal things and practices.

Another example concerns 'meaning hacking' and 'deep technical hacking' (see Haklay 2013) which also enable focal practices as evidenced in the EveryAware community project monitoring noise around Heathrow airport (Becker *et al.* 2013). Here, the process starts with an app that utilizes the sensing abilities of a smartphone. The app, WideNoise, records the level of sound in decibels (dB) and qualitative observations submitted by participants (such as the emotional scale of love/hate and adding a description through tagging), in addition to location and time information from the phone's sensors. The tool can accurately indicate sound ranges as being low, medium or high, though acoustic laboratory testing demonstrated that it could not be relied upon to capture exact dB values. Once the app was presented to a community organization in the area of Heathrow its use was welcomed with enthusiasm. Even though participants were aware of the technical limitations of the device, the activity of going out and recording incidents of airplane noise with emotional tags was considered meaningful. In turn, the community used these data to demonstrate the level of community concern to a governmental committee that is considering the expansion of the airport.

As a result, community-led data collection and the potential of new DIY electronic devices was discussed within the community (Nold 2015). Among the proposed devices was a noise meter that is programmed to send a Short-Messaging-System (SMS) message every time the level of noise breaches a predefined value to be used to alert a local or national decision-maker to the event. While the devices on offer were created as prototype interventions to spark a debate (see Nold 2015 for a full discussion), they led to interest within the community to construct a noise monitor that is accurate enough (within less than 2–3 dB from calibrated meters), can be installed in their attics and can record the nuisance throughout the day and can be attributed to specific flight events. The effort to construct the devices and install them are ongoing.

Both aspects of the process – the use of the WideNoise app and the development of a noise monitor – demonstrate that, even in minor participatory sensing events, the devices can act as a 'focal thing', bringing people together in a purposeful and meaningful social activity that is significant to participants.

## Towards meaningful data production

In this chapter, Albert Borgmann's concepts of device paradigm and focal practice were used to challenges normalized understandings of the smart city and pointed to some of the shortcomings of deploying sensors and collecting data as an efficient instrumentalist process in lieu of participatory sensing activities.

The crowdsourced mapping, citizen science and DIY science examples discussed, as well as numerous other emerging examples around the world, demonstrate the potential for reconsidering smart city technologies and their social role and the possibility of them functioning as focal things and practices. To make the smart city socially meaningful, however, requires technical support and active interventions by government actors and those who develop the technologies or hold the know-how to use data sources and turn them into useful and meaningful information. It is also important to get people together to develop technologies, discuss data collection protocols, or understand the analysis, as these activities can provide meaningful communal events that can nurture new and existing links between individuals and communities.

Borgmann's typology of information offers an alternative option for digital engagement. Generally, digital data and information are considered as merely technological (information *as* reality); through citizen science and participatory sensing (see Haklay 2016) it was argued here that it can be also cultural data and information (information *for* reality). By becoming cultural information used through meaningful participatory and collective action, the smart city paradigm can be transformed – or at least enhanced – with focal practices that bring people together. The Public Lab examples, and especially the Heathrow and the Gila River examples, demonstrate that this is possible. By opening up smart cities to allow for such data – and data collection practices – as an integral part of decisions about what will be collected and how, the potential of expressing competing visions and values of the city can be accommodated.

Although these citizen science approaches may develop new avenues for discussing alternatives to the efficiency and productivity logic of smart cities, it remains important to not absolve those with resources, power and knowledge from responsibility. There is an urgent need to ensure that the development and use of the smart city technologies be created open to democratic and societal control, and that they are not being developed only because technologists and scientists think that it is possible to do so or to capture cities as new markets of accumulation.

## Acknowledgement

A shorter version of this chapter appears in *Urban Pamphleteer No. 1* by UCL Urban Lab (2013), edited by Ben Campkin and Rebecca Ross. Some material also appears on my blog (povesham.wordpress.com). I would like to thank Chris Perkins and Sybille Lammes for the 2013 workshop on 'Thinking and doing digital mapping' where some of the ideas for this chapter were discussed. The research was supported by EPSRC 'Extreme' Citizen Science grant (EP/I025278/1) and FP7 EveryAware project.

## References

Becker, M., Caminiti, S., Fiorella, D., Francis, L., Gravino, P., Haklay, M., Hotho, A., Loreto, V., Mueller, J., Ricchiuti, F., Servedio, V.D., Sîrbu, A. and Tria, F. (2013) 'Awareness and learning in participatory noise sensing', *PLoS One* 8(12).

Borgmann, A. (1984) *Technology and the Character of Contemporary Life: A Philosophical Inquiry*. Chicago, IL: University of Chicago Press.

Borgmann, A. (1999) *Holding on to Reality: The Nature of Information at the Turn of the Millennium*. Chicago, IL: University of Chicago Press.

Borgmann, A. (2010) 'Orientation in technological space', *First Monday* 15(6), available from: http://firstmonday.org/ojs/index.php/fm/article/view/3037/2568 [accessed 9 December 2016].

Brautigan, R. (1967) *All Watched Over by Machines of Loving Grace*. San Francisco, CA: The Communication Company.

Budhathoki, N.R. (2010) 'Participants' motivations to contribute geographic information in an online community', doctoral dissertation, University of Illinois at Urbana-Champaign.

Budhathoki, N.R. and Haythornthwaite, C. (2013) 'Motivation for open collaboration crowd and community models and the case of OpenStreetMap', *American Behavioral Scientist* 57(5): 548–575.

Chourabi, H., Nam, T., Walker, S., Gil-Garcia, J.R., Mellouli, S., Nahon, K., Pardo, T. and Scholl, H.J. (2012) 'Understanding smart cities: an integrative framework', in *System Science (HICSS) 2012 45th Hawaii International Conference*, pp. 2289–2297.

Coleman, R. and Sim, J. (2000) '"You'll never walk alone": CCTV surveillance, order and neo-liberal rule in Liverpool city centre', *The British Journal of Sociology* 51(4): 623–639.

Dosemagen, S., Warren, J. and Wylie, S. (2011) 'Grassroots mapping: Creating a participatory map-making process centered on discourse', *Journal of Aesthetics and Protest* 8, available from: www.joaap.org/issue8/GrassrootsMapping.htm [accessed 9 December 2016].

Dusek, V. (2006) *Philosophy of Technology: An Introduction*. Oxford: Blackwell.

Feenberg, A. (2002) *Transforming Technology: A Critical Theory Revisited*. Oxford: Oxford University Press.

Fussey, P. (2002) 'An interrupted transmission? Processes of CCTV implementation and the impact of human agency', *Surveillance & Society* 4(3), available from: www.surveillance-and-society.org/articles4(3)/transmission.pdf [accessed 10 February 2017].

Gerlach, J. (2015) 'Editing worlds: Participatory mapping and a minor geopolitics', *Transactions of the Institute of British Geographers* 40(2): 273–286.

Graham, S.D. (2005) 'Software-sorted geographies', *Progress in Human Geography* 29(5): 562–580.

Greenfield, A. (2013) *Against the Smart City*. London: Do projects; 1.3 edition [accessed 20 December 2013].

Haklay, M. (2013) 'Neogeography and the delusion of democratisation', *Environment and Planning A* 45(1): 55–69.

Haklay, M. (2016) 'Making participatory sensing meaningful', in Y. Beebeejaun (ed.), *The Participatory City*. Berlin: Jovis, pp. 154–161.

Haklay, M. and Weber, P. (2008) 'OpenStreetMap: User-generated street maps', *IEEE Pervasive Computing* 7(4): 12–18.

Haklay, M., Singleton, A. and Parker, C. (2008) 'Web mapping 2.0: the neogeography of the Geoweb', *Geography Compass* 2(6): 2011–2039.

Higgs, E., Light, A. and Strong, D. (eds) (2000) *Technology and the Good Life?* Chicago, IL: University of Chicago Press.

Kitchin, R. (2014) 'The real-time city? Big data and smart urbanism', *GeoJournal* 79(1): 1–14.

Kitchin, R. and Dodge, M. (2011) *Code/Space: Software and Everyday Life*. Cambridge, MA: MIT Press.

Marshall, A. (2006) 'A critique of the development of quantitative methodologies in human geography', *Radical Statistics* 92, available from: www.radstats.org.uk/no092/marshall92.pdf [accessed 20 December 2013].

Nam, T. and Pardo, T.A. (2011) 'Conceptualizing smart city with dimensions of technology, people, and institutions', in *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, ACM, pp. 282–291.

Nold, C. (2015) 'Micro/macro prototyping', *International Journal of Human-Computer Studies* 81(C): 72–80.

Nov, O., Arazy, O. and Anderson, D. (2014) 'Scientists@ home: What drives the quantity and quality of online citizen science participation', *PloS One* 9(4).

Public Laboratory of Open Technology and Science (2015) 'About', available from: http://publiclab.org/about [accessed 10 June 2015].

Russell, S.A. (2014) *Diary of a Citizen Scientist: Chasing Tiger Beetles and Other New Ways of Engaging the World*. Corvallis: Oregon State University Press.

Sieber, R.E. and Haklay, M. (2015) 'The epistemology of volunteered geographic information', *Geo: Geography and Environment* 2(2): 122–136.

Silvertown, J. (2009) 'A new dawn for citizen science', *Trends in Ecology & Evolution* 24: 467–471.

Su, K., Li, J. and Fu, H. (2011) 'Smart city and the applications', in *Electronics Communication and Control (ICECC), 2011 International Conference*, pp. 1028–1031.

Turkle, S. (2012) *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York: Basic Books.

Verbeek, P.-P. (2002) 'Devices of engagement: On Borgmann's philosophy of information and technology', *Techné* 6(1): 69–92.

Wylie, S.A., Jalbert, K., Dosemagen, S. and Ratto, M. (2014) 'Institutions for civic technoscience: How critical making is transforming environmental research', *The Information Society* 30(2): 116–126.

Wyly, E. (2014) 'Automated (post)positivism', *Urban Geography* 35(5): 669–690.

# Index