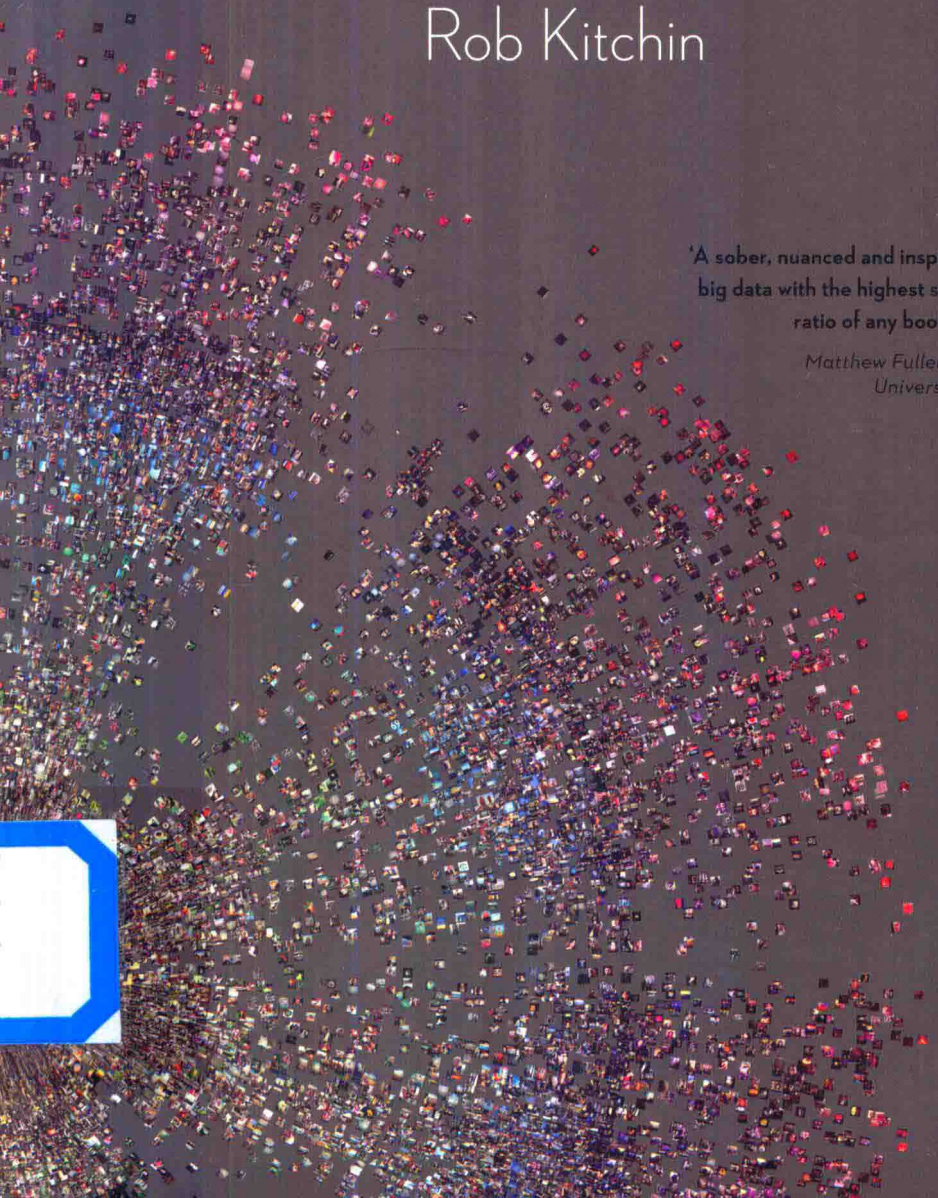
THE DATA REVOLUTION

BIG DATA, OPEN DATA, DATA INFRASTRUCTURES & THEIR CONSEQUENCES

Rob Kitchin



'A sober, nuanced and inspiring guide to big data with the highest signal to noise ratio of any book in the field.'

Matthew Fuller, Goldsmiths, University of London

THE DATA REVOLUTION

BIG DATA, OPEN DATA, DATA INFRASTRUCTURES & THEIR CONSEQUENCES

Rob Kitchin

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

First published 2014

Editor: Robert Rojek
Assistant editor: Keri Dickens
Production editor: Katherine Haw
Copyeditor: Rose James
Marketing manager: Michael Ainsley
Cover design: Francis Kenney
Typeset by: C&M Digitals (P) Ltd, Chennai, India
Printed and bound by CPI Group (UK) Ltd,
Croydon, CR0 4YY

Library of Congress Control Number: 2014932842

British Library Cataloguing in Publication data

A catalogue record for this book is available from
the British Library

MIX
Paper from responsible sources
FSC® C013604

ISBN 978-1-4462-8747-7
ISBN 978-1-4462-8748-4 (pbk)

THE DATA REVOLUTION

'This is a path-breaking book. Rob Kitchin has long been one of the leading figures in the conceptualisation and analysis of new forms of data, software and code. This book represents an important step-forward in our understanding of big data. It provides a grounded discussion of big data, explains why they matter and provides us with a framework to analyse their social presence. Anyone who wants to obtain a critical, conceptually honed and analytically refined perspective on new forms of data should read this book.'

David Beer, Senior Lecturer in Sociology, University of York

'Data, the newest purported cure to many of the world's most "wicked" problems, are ubiquitous; they're shaping discourses, policies, and practices in our war rooms, our board rooms, our classrooms, our operating rooms, and even around our dinner tables. Yet given the precision and objectivity that the *datum* implies, it's shocking to find such ***im***precision in how data are conceived, and such cloudiness in our understandings of how data are derived, analyzed, and put to use. Rob Kitchin's timely, clear, and vital book provides a much needed critical framework. He explains that our ontologies of data, or how we understand what data *are*; our epistemologies of data, or how we conceive of data as units of truth, fact, or knowledge; our analytic methodologies, or the techniques we use to process that data; and our data apparatuses and institutions, or the tools and (often huge, heavy, and expensive) infrastructures we use to sort and store that data, are all entwined. And all have profound political, economic, and cultural implications that we can't risk ignoring as we're led into our "smart," data-driven future.'

Shannon Mattern, Faculty, School of Media Studies, The New School

'A sober, nuanced and inspiring guide to big data with the highest signal to noise ratio of any book in the field.'

Matthew Fuller, Digital Culture Unit, Centre for Cultural Studies, Goldsmiths, University of London

'Data has become a new key word for our times. This is just the book I have been waiting for: a detailed and critical analysis that will make us think carefully about how data participate in social, cultural and spatial relations.'

Deborah Lupton, Centenary Research Professor News & Media Research Centre, University of Canberra

'By carefully analysing data as a complex socio-technical assemblage, in this book Rob Kitchin discusses thought-provoking aspects of data as a technical, economic and social construct, that are often ignored or forgotten despite the increasing focus on data production and usage in contemporary life. This book unpacks the complexity of data as elements of knowledge production, and does not only provide readers from a variety of disciplinary areas with useful conceptual framings, but also with a challenging set of open issues to be further explored and engaged with as the "data revolution" progresses.'

Luigina Ciolfi, Sheffield Hallam University

'Kitchin paints a nuanced and complex picture of the unfolding data landscape. Through a critique of the deepening technocratic, often corporate led, development of our increasingly data driven societies, he presents an alternative perspective which illuminates the contested, and contestable, nature of this acutely political and social terrain.'

Jo Bates, Information School, University of Sheffield

'*The Data Revolution* is a timely intervention of critical reflection into the hyperbolic and fast-paced developments in the gathering, analysis and workings of "big data". This excellent book

diagnoses the technical, ethical and scientific challenges raised by the data revolution, sounding a clarion for critical reflections on the promise and problematic of the data revolution.'

Sam Kinsley, University of Exeter

'Much talk of big data is big hype. Different phenomena dumped together, a dearth of definitions and little discussion of the complex relationships that give rise to and shape big data practices sums it up. Rob Kitchin puts us in his debt by cutting through the cant and offering not only a clear analysis of the range, power and limits of big data assemblages but a pointer to the crucial social, political and ethical issues to which we should urgently attend. Read this book.'

David Lyon, Queen's University, Canada

'Data matter and have matter, and Rob Kitchin thickens this understanding by assembling the philosophical, social scientific, and popular media accounts of our data-based living. That the give and take of data is increasingly significant to the everyday has been the mainstay of Kitchin's long and significant contribution to a critical technology studies. In *The Data Revolution*, he yet again implores us to think beyond the polemical, to signal a new generation of responsive and responsible data work. Importantly, he reminds us of the non-inevitability of data, articulating the registers within which interventions can and already are being made. Kitchin offers a manual, a set of operating instructions, to better grasp and grapple with the complexities of the coming world, of such a "data revolution".'

Matthew W. Wilson, Harvard University and University of Kentucky

'With a lucid prose and without hyperbole, Kitchin explains the complexities and disruptive effects of what he calls "the data revolution". The book brilliantly provides an overview of the shifting socio-technical assemblages that are shaping the uses of data today. Carefully distinguishing between big data and open data, and exploring various data infrastructures, Kitchin vividly illustrates how the data landscape is rapidly changing and calls for a revolution in how we think about data.'

Evelyn Ruppert, Goldsmiths, University of London

'Kitchin's powerful, authoritative work deconstructs the hype around the "data revolution" to carefully guide us through the histories and the futures of "big data". The book skilfully engages with debates from across the humanities, social sciences, and sciences in order to produce a critical account of how data are enmeshed into enormous social, economic, and political changes that are taking place. It challenges us to rethink data, information and knowledge by asking - who benefits and who might be left out; what these changes mean for ethics, economy, surveillance, society, politics; and ultimately, whether big data offer answers to big questions. By tackling the promises and potentials as well as the perils and pitfalls of our data revolution, Kitchin shows us that data doesn't just reflect the world, but also changes it.'

Mark Graham, University of Oxford

'This is an incredibly well written and accessible book which provides readers who will be curious about the buzz around the idea of big data with: (a) an organising framework rooted in social theory (important given dominance of technical writings) through which to conceptualise big data; (b) detailed understandings of each actant in the various data assemblages with fresh and novel theoretical constructions and typologies of each actant; (c) the contours of a critical examination of big data (whose interests does it serve, where, how and why). These are all crucial developments it seems to me and I think this book will become a trail blazer because of them. This is going to be a biggie citation wise and a seminal work.'

Mark Boyle, Director of NIRSA, National University of Ireland, Maynooth

LIST OF TABLES

LIST OF FIGURES

ABOUT THE AUTHOR

Professor Rob Kitchin is an European Research Council Advanced Investigator at the National University of Ireland Maynooth. He has authored or edited 23 other books and was the 2013 recipient of the Royal Irish Academy's Gold Medal for the Social Sciences. He is principal investigator for the Digital Repository of Ireland and the All-Island Research Observatory.

ACKNOWLEDGEMENTS

This book started life in early July 2012 as a discussion in a coffee shop in Edinburgh with Robert Rojek from Sage. I was suggesting he find someone to write a book on big data, open data, and data infrastructures, presenting ideas as to who might be well placed to draft such a text. He felt I was the right person for the job. A couple of months later I decided to juggle round my writing plans and started to draft what was to be a quite short, critical analysis of the changing data landscape. Over time the book developed into a full-length manuscript that sought to do justice to the emerging trends and debates. Along the way, Robert remained a keen sounding board and source of interesting material, and his help has been very much appreciated. At Sage, his colleague Keri Dickens helped shepherd the book into production, where it was admirably guided to production by Katherine Haw.

Martin Dodge and Tracey P. Lauriault kindly undertook a detailed read-through and critique of the entire manuscript. Mark Boyle read the entire second draft. Gavin McArdle and Evelyn Ruppert provided useful critique of individual chapters, and a number of other colleagues and peers engaged in useful discussions and guided me to relevant material, including Mark Graham, Taylor Shelton, Matt Zook, Matt Wilson, Lev Manovich, Cian O'Callaghan, Sung-Yueh Perng, Aileen O'Carroll, Jane Gray, Sandra Collins, John Keating, Sharon Webb, Justin Gleeson, Aoife Dowling, Eoghan McCarthy, Martin Charlton, Tim McCarthy, Jan Rigby, Rob Bradshaw, Alan Moore, Darach Mac Donncha and Jim White. I also received useful feedback at presentations at Durham University, Clark University and Harvard University. Rhona Bradshaw and Orla Dunne minded the office while I tried to keep my head down to conduct research and draft chapters. Justin Gleeson kindly produced some of the diagrams. I owe you all a debt of gratitude. I would also like to thank the many people on Twitter for pointing me to interesting material and engaging in relevant micro-discussions. Lastly, as ever, Cora kept me grounded and provided wonderful support.

The research conducted in writing this book was in part supported by a European Research Council Advanced Investigator Award, 'The Programmable City' (ERC-2012-AdG-323636; www. nuim.ie/progcity) and Programme for Research in Third Level Institutes Cycle 5 funding from the Higher Education Authority to create a Digital Repository for Ireland.

A hyperlinked version of the book's bibliography can be found at http://thedatarevolutionbook.wordpress.com/. Additional sources of information and stories about the data revolution are regularly scooped onto http://www.

scoop.it/t/the-programmable-city. Feedback is also welcome via email (Rob.Kitchin@nuim.ie) or Twitter (@robkitchin).

Some of the material in this book has been previously published as papers and blog posts, though it has been updated, reworked and extended:

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): 851–81.

Kitchin, R. (2013) 'Big data and human geography: opportunities, challenges and risks', *Dialogues in Human Geography,* 3(3): 262–7.

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

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

Kitchin, R. and Lauriault, T. (2014) *Small Data, Data Infrastructures and Big Data.* The Programmable City Working Paper 1. Available at SSRN: http://ssrn.com/abstract=2376148.

Kitchin, R. and Lauriault, T. (in press) 'Small data in an era of big data,' *Geo Journal.*

Figure 1.1 is adapted from InformationisBeautiful.net with the permission of David McCandless.

Figure 1.2 is reproduced with the permission of The Statistical and Social Inquiry Society of Ireland.

Table 2.4 is included with the permission of Neil Beagrie, Brian Lavoie and Matthew Woollard and under a creative commons licence for Fry et al., http://repository.jisc.ac.uk/279/.

Table 3.1 is reproduced from http://opendefinition.org/od/ under a creative commons licence.

Table 3.3 is included with the permission of Michael Hausenblas, http://5stardata.info/.

Table 4.1 is reproduced with the permission of *The Economist.* The Economist Newspaper Limited, London, issued March 11, 2014.

Figure 6.1 is reproduced with the permission of Monica Stephens.

Table 6.1 is reproduced with the permission of Taylor and Francis.

Figure 6.2 is reproduced with the permission of Flightradar24.com.

Figure 6.3 is reproduced with the permission of Andrew Hudson-Smith.

Figures 6.4 and 6.5 are reproduced with the permission of Professor Mikael Jern, National Center for Visual Analytics, Linköping University, http://ncva.itn.liu.se.

Table 7.1 Forms of big data corporate intelligence is included with the permission of McKinsey & Company.

Table 7.2 and Figure 7.1 are reproduced courtesy of International Business Machines Corporation, © International Business Machines Corporation.

Figure 7.2 is reproduced from http://ipprio.rio.rj.gov.br/centro-de-operacoes-rio-usa-mapas-feitos-pelo-ipp/ under a creative commons license.

Tables 10.2 and 10.3 are included with the permission of John Wiley & Sons.

Table 10.4 is included with the permission of Ann Cavoukian, Ph.D., Information and Privacy Commissioner, Ontario, Canada.

NOTE

Throughout this book the term 'data' is expressed in the plural, with datum being used to denote a singular instance. As explained in the *Oxford English Dictionary* (OED):

> In Latin, **data** is the plural of **datum** and, historically and in specialized scientific fields, it is also treated as a plural in English, taking a plural verb, as in ***the data were collected and classified***.

However, the term is increasingly used in the singular form in popular media and everyday conversation. As the OED details:

> In modern non-scientific use, however, it is generally not treated as a plural. Instead, it is treated as a mass noun, similar to a word like **information**, which takes a singular verb. Sentences such as ***data was collected over a number of years*** are now widely accepted in standard English.

The book therefore follows scientific convention. However, where it is used in the singular in quoted passages, the original text has been retained. As to which version is correct, the grammarians would argue for the plural, but popular opinion is more open and flexible.

PREFACE

There is a long history of governments, businesses, science and citizens producing and utilising data in order to monitor, regulate, profit from, and make sense of the world. Data have traditionally been time-consuming and costly to generate, analyse and interpret, and generally provided static, often coarse, snapshots of phenomena. Given their relative paucity, good-quality data were a valuable commodity, either jealously guarded or expensively traded. Recently, this state of affairs has started to change quite radically. Data have lost none of their value, but in other respects their production and nature is being transformed through a set of what Christensen (1997) terms disruptive innovations that challenge the status quo as to how data are produced, managed, analysed, stored and utilised. Rather than being scarce and limited in access, the production of data is increasingly becoming a deluge; 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. A data revolution is underway, one that is already reshaping how knowledge is produced, business conducted, and governance enacted.

This revolution is founded on the latest wave of information and communication technologies (ICTs), such as the plethora of digital devices encountered in homes, workplaces and public spaces; mobile, distributed and cloud computing; social media; and the internet of things (internetworked sensors and devices). These new technical media and platforms are leading to ever more aspects of everyday life – work, consumption, travel, communication, leisure – and the worlds we inhabit to be captured as data and mediated through data-driven technologies. Moreover, they are materially and discursively reconfiguring the production, circulation and interpretation of data, producing what has been termed 'big data' – vast quantities of dynamic, varied digital data that are easily conjoined, shared and distributed across ICT networks, and analysed by a new generation of data analytics designed to cope with data abundance as opposed to data scarcity. The scale of the emerging data deluge is illustrated by the claim that '[b]etween the dawn of civilisation and 2003, we only created five exabytes of information; now we're creating that amount every two days' (Hal Varian, chief economist with Google, cited in Smolan and Erwitt 2012).

Big data are not the only components of the data revolution. Rather, there are related initiatives such as the digitisation, linking together, and scaling-up of traditionally produced datasets (small data) into networked data infrastructures; the open data movement that seeks to make as much data as possible openly available for all to use; and new institutional structures that seek to secure common

guidelines and policies with respect to data formats, structures, standards, metadata, intellectual property rights, licensing and sharing protocols. Together, these constitute a set of new data 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 – that frame how data are produced and to what ends they are employed.

The impact of big data, open data and data infrastructures is already visible in science, business, government and civil society. Used to operating in data deserts, seeking to extract information and draw conclusions from relatively small numbers of observations, established disciplines are now starting to grapple with a data avalanche (H.J. Miller 2010). They are accompanied by new fields, such as data science, social computing, digital humanities, and computational social sciences, that are explicitly concerned with building data infrastructures and finding innovative ways to analyse and make sense of scaled and big data. In business, big data are providing a new means to dynamically and efficiently manage all facets of a company's activities and to leverage additional profit through enhanced productivity, competitiveness, and market knowledge. And data themselves have become an important commodity, actively bought and sold within a global, multi-billion dollar market. For governments, widespread, dynamic data are providing new insights about their own operations, as well as reshaping the means to govern and regulate society. Through examining open datasets, citizens and non-governmental organisations (NGOs) are drawing their own conclusions, challenging corporate and government agendas, and forwarding alternative visions of how society should be organised and managed.

These new opportunities have sparked a veritable boom in what might be termed 'data boosterism'; rallying calls as to the benefits and prospects of big, open and scaled small data, some of it justified, some pure hype and buzz. In turn, the terms big data and open data have become powerful memes, not just a way of describing data but symbolic of a wider rhetoric and imaginary that is used to garner support and spread their roll-out and adoption. Such boosterism and memes can make it easy to drift into uncritically hyping the changes taking place, many of which raise numerous ethical, political and legal concerns. History, though, does reveal earlier precedents of disruptive information-related innovations – the radical transformation of knowledge production in the wake of the printing press, for example. Indeed, every new era of science has had at its inception new technologies that lead to an information overload and spark a transition to new ways of generating, organising, storing, analysing and interpreting data (Darnton 2000). For example, Strasser (2012) notes, the explorations of the Renaissance, enabled by better navigation, mapping and scientific instruments, yielded vast quantities of new discoveries that led to new methods of categorisation, new technologies of analysis and storage, and new scientific insights.

Given the relatively early point in the present data revolution, it is not at all certain how the present transformations will unfold and settle, and what will be the broader consequences of changes taking place. What is clear is that there is an urgent need to try and make sense of what is happening. Thus, the aim of this book is to provide a synoptic, conceptual and critical analysis of data and the data revolution underway. It seeks, on the one hand, to chart the various ways in which the generation, processing, analysis and sharing of data is being reconfigured, and what this means for how we produce and use information and knowledge; and, on the other, to open up debate and critical reflection about data: their nature, how they are framed technically, philosophically, ethically and economically, and the technological and institutional assemblages that surround them. Rather than setting out a passionate case for the benefits of big data, open data and data infrastructures, or an entrenched critique decrying their more negative consequences, the book provides a contextual, critical appraisal of the changes taking place.

The analysis presented is based on an extensive engagement with the literature from across humanities, social sciences and the sciences, and from popular culture, journalism, and industry publications, and on first-hand experience of working on large-scale data archiving/infrastructure and data analytics projects. The book is divided into eleven chapters. The first provides an overview and critical reflection on the concept of data and how to make sense of databases and data infrastructures. The second examines the continued role of small data and how they are being scaled up into digital archives and infrastructures, and sold through data brokers. Chapter 3 discusses the drive towards creating open and linked data that are more widely shared and reused. Chapters 4 and 5 detail the nature of big data and its enablers and sources. Chapter 6 provides an overview of a new set of data analytics designed to make sense of scaled small data and big data. The next two chapters examine the arguments used to promote big data and their impact on governance and business, and the ways in which the data revolution is reshaping how research is conceptualised and practised. Chapters 9 and 10 discuss the technical, organisational, ethical, political and legal challenges of the data revolution. The final chapter sets out some overarching conclusions and provides a road map for further research and reflection.

CONTENTS