1

Data Stories

Data. Big data. Open data. Database. Personal data. Data-driven. Data brokers. Data analytics. Data is the new oil. Data plan. Data deluge. Data activism. Data science. Data infrastructure. Data justice. General Data Protection Regulation. Data. Data. Data …

A decade ago, the word 'data' seemed to be largely the preserve of researchers and a few professional administrators and workers. Now, it has entered our everyday lexicon. We are aware that our data are routinely harvested, that our digital world generates and consumes vast volumes of data, and our economy and government are becoming data-driven. Yet, it is not always clear what data are, how best to make sense of them, and what is at stake.

Before this data revolution, data were largely treated in a common sense way. They were gathered facts about the world; the raw material generated through scientific measurements, government surveys, administrative systems, academic research, explorations and fieldwork. They were the building blocks that enabled the creation of information and knowledge. However, just as we pay little attention to the bricks that enable us to build complex buildings, which collectively create vibrant cities, we rarely focused critical attention on data. Consequently, we have found ourselves living in a data-driven world that, on the one hand, seems intuitive to navigate – we use digital technologies, click here and there, fill out online forms, agree to terms and conditions, share our thoughts and photos. On the other hand, we have little comprehension as to the full spectrum of data harvested about us, what happens to them, and how they are used.

Digital systems are often black-boxed, meaning we have little sense of how they work beyond their interface and immediate effects. Data-rich organizations, such as private companies, data brokers or state bodies, are by design secretive, either to protect commercial advantage, hide their activities, or to protect the integrity of their work. We simply have to trust that they are only generating the data they need, and are treating, using, and acting on these data in ethical ways.

In recent years, this trust that has been eroded. Our faith in digital systems and data-rich organizations has been undermined by a series of scandals and data hacks. The Snowden revelations exposed how our data are being gathered on a massive scale by governments in order for them to be able to spy on their own and other nations' citizens. The Facebook/Cambridge Analytica scandal demonstrated how the data profiles of millions of people were used to target them with posts designed to shape their voting preference. Not a week seems to go by without a major data breach in which millions of personal records are accessed, including sensitive ID and financial information.[1]

Despite these events, the data revolution continues apace, and we remain either largely ignorant or indifferent to their consequences, or feel impotent to be able to do anything about them. We have little choice in our interactions with government if we want to receive entitlements and services and comply with legislation. Data is the price we have to pay for using commercial platforms, or receiving benefits and rewards for purchases. The only alternative appears to be to avoid the use of digitally mediated transactions, which is all but impossible when we routinely shop online, pay with credit/debit cards, use mobile phone apps and converse with friends using social media.

While most people interested in data are focused on how to use them productively in their work – as a means to create insight and innovation, gain efficiencies and optimization, extract value and make money, and manage and control systems – a number of academics, journalists and civil rights advocates have started to question the logics and consequences of the data revolution. Under the banner of critical data studies they focus their attention on the nature and production of data, and how everyday life is being transformed by data.

Rather than seeing data as simply the raw material – the bricks – used to create information and knowledge, they cast data as a manufactured material that intrinsically has value. Data, they argue, is never raw but always cooked to some recipe. Data do not pre-exist their generation; they are not simply waiting to be collected or harvested in a technical, passive and objective manner. Instead, they are produced – that is, actively created via procedures and instruments of our devising. Through their handling and sharing data can mutate, becoming cleaned, wrangled, transformed and combined. Ultimately, they can be deleted, completing what can be a complex data life cycle.

Data then are not benign, neutral measures that reflect the world as it is, within technical constraints. What data are generated, and how they are produced, handled and used, is the result of choices and decisions by people. These, in turn, are shaped by intended outcomes, theories and concepts, research designs, procedures, protocols, standards, resources and finance, regulations and laws, organizational processes, ethics reviews, political context and so on. The cooking of data does not take place in vacuum, but within context. Data-driven endeavours are not simply technical systems, but are socio-technical systems. That is, they are as much a result of human values, desires and social relations as they are scientific principles and technologies.

The sociality of data is also evident with respect to how we have come to live with data. The data revolution has been transforming work and the economy, the nature of consumption, the management and governance of society, how we communicate and interact with media and each other, and forms of play and leisure. Our lives are saturated with digital devices and services that generate, process and share vast quantities of data.

We carry smartphones around with us, using an array of apps and games throughout the day. We communicate using email, messaging, video calls and social media. We employ computers and digitally controlled technologies to perform work tasks. We monitor our health and activity using self-tracking devices. We drive around in cars chocked full of computers that mediate our driving experience, even if we're not aware that they are doing so. Our homes are increasingly full of digitally mediated appliances from personal assistants such as Alexa and Siri, to smart TVs and

home management systems. We pay for goods and services using electronic cash, often to buy digital products (games, streaming services, books). Even if we try to live analogue lives, we cannot escape surveillance systems and the databases of governments and companies with whom we have to interact.

The result is we leave a series of data footprints (that we choose to create) and shadows (that are captured whether we want them to be or not) as we live our everyday lives. These data are valuable commodities and are used for a range of purposes, including verification, profiling and decision-making. They are shared and traded, feeding into a vast global data market. Our data now often precede us, influencing what adverts we routinely see, whether we are approved for a loan, or tenancy, or job, and whether we receive special offers and rewards. As such, our data can be used to empower us, but also exploit, discriminate and persecute. Much of this consumption and monetization of our data is hidden from us, and many of us barely know the extent to which it is influencing outcomes critical for our future well-being. Consequently, while some of our life with data is clear, much is opaque.

Recognizing the sociality of data, and comprehending data and data systems as being socio-technical in nature, provides a conceptual position for challenging the teleological mantra of the data revolution. That is, it provides a platform to contest the seemingly inevitable, pre-ordained data landscape and how data-driven systems work, and to envisage different data futures. A critical aspect of divining such futures is to actively explore moral and ethical questions with respect to how data are produced, shared, traded and protected. Indeed, several debates – some of which are happening in the public eye, some behind closed doors – are ongoing about how data should be managed and governed through rules, principles, policies, licences and laws, and under what circumstances and to what ends data can be employed (GDPR (General Data Protection Regulation), for example, is a result of these debates). There are no easy answers, and the suggested approaches and outcomes are strongly contested by different parties given what is at stake. Yet, the results of these debates will have an effect on our data-driven world.

The objective of *Data Lives* is to reveal the myriad, complex, contested ways in which data are produced and circulated, and the consequences of living in a data-driven world. What follows reveals our data world to be full of potential dangers, but also benefits and hope. The life of data and living with data is bound up in contingencies and is open to new configurations and possibilities. And while the data revolution seems to have a relatively robust path dependency, it can be diverted onto new routes. We can create our own data lives.

Telling Stories

The usual approach to writing an article or book that examines the praxes and politics of data is to produce a rather sterile narrative that weaves together the theories, observations and findings from the academic and policy literature with empirical evidence and conceptual musings of the author. The voice is most often in the third person, impassionate and distant. The aim is to create a convincing argument in a neutral register that displays a strong degree of objectivity. I've written plenty of such accounts, which are usually impersonal, full of jargon and conceptual musings, and are aimed at academic peers.

In *Data Lives* I take a different approach to examining data, using the more familiar narratives of personal reminiscence, journalistic-style essays, and short stories. Rather than writing from a detached academic position, I adopt a different kind of voice and point-of-view to personalize the narrative both with respect to the writer and reader. In essence, I employ a more reflexive standpoint that draws on my many years of experience of researching and writing about data, building data infrastructures, serving on the boards of institutions that produce government data and telling data stories to influence public policy. The chapters that follow tell a set of interconnected stories about how data are produced, processed and interpreted, and the consequences of living in a data-driven world.

Storytelling has always been a powerful way of communicating ideas and providing a critical lens to consider society and social processes and change. Short stories, novels, comics, documentaries, biographies, television dramas and movies provide media that

can be more provocative and playful than academic accounts. They can set out different views and explore values, conflict and consequences using various forms of narrative devices. For example, science fiction uses extrapolation and speculation to explore possible futures given present trends. In particular, SF employs the tactics of estrangement (pushing a reader outside of what they comfortably know) and defamiliarization (making the familiar strange) as a way of creating a distancing mirror and to prompt critical reflection on society.[2] These tactics were used to good effect in a book I edited recently – *How to Run a City Like Amazon, and Other Fables* – in which the contributors used short stories to explore what cities would be like if they were run by, or using the business model of, different companies.

In what follows, the essays are more akin to personalized documentaries or memoir, or short explanatory accounts. The stories are purely fictional tales or modified dramatizations of events. Both the fiction and non-fiction are rooted in my own experiences, though they also draw from extensive wider research and reading. The use of personal recollection and reflection might create the impression that there is a somewhat anecdotal quality to the message. Undoubtedly, it is the case that some of the material is circumstantial; they are specific stories about data based on personal interactions and various kinds of data work. Many are data tales that I have told others over coffee or in a bar after workshops or to illustrate a point in the classroom; little vignettes of practices and encounters that I have found illustrative, disturbing or humorous.[3]

Some readers might be wary of this approach, casting it as 'unscientific'. It's certainly the case that the arguments presented are not based on the use of a rigorous, systematic, scientific method involving a representative sampling frame, impartial and rote analysis, and impassionate, descriptive interpretation (though some of the projects they describe are). However, I would contend that it does not render the arguments made invalid or without useful insight. Nor does it mean that the material and examples discussed are unrepresentative. Rather, what is presented are case studies/histories produced through a form of recovered auto-ethnography and contextualized with respect to the arguments and findings of the wider critical data

studies and data science literature. These case studies/histories are illustrative of situations and processes that I have witnessed many times and will be familiar to those who regularly work with data, or are concerned with how data affect our everyday lives. After all, they are rooted in 30 years of real-world projects, academic research, knowledge of the field, practical experience of data work, and the politics surrounding data initiatives.

Like all researchers, I am familiar with generating and handling data. Academics are trained with respect to how best to collect, process and interpret data. Whereas other academics are mostly using data to understand and explain particular phenomenon, much of my focus has been on data themselves. My doctoral work in the early 1990s focused on whether the methods used to measure individuals' geographic knowledge about the world produced valid data, and if the statistical tests used to analyze them created accurate and reliable insights.[4] My approach was scientific, using quantitative and statistical methods to test the veracity of the data produced and the analytics used to make sense of them. My conclusion was that significant amounts of overlooked error and bias were being introduced through the processes of generating and analyzing data and these shortcomings were leading to findings that lacked integrity.[5]

I still believed in the usefulness of data and science, but my faith in them was shaken a little. As a result, I have long been interested in issues of data quality and veracity, undertaking a number of projects that examine these in different contexts. For example, a recent project has explored whether it is possible to use real-time data from smart city technologies to create new official statistics.[6] This has involved examining the data with respect to issues such as representativeness, access, coverage, accuracy, reliability, cleanliness, consistency, completeness, provenance and transparency. Collectively, these qualities concern trust: do we believe in the integrity of a dataset to tell us something meaningful about the world?

Much of my own empirical research in the 2000s, and that of my colleagues in the institute I directed, concerned social, regional and cross-border policy. Government data was essential for performing this work. However, it was often hard to access, or provided in formats that were difficult to use. As a consequence,

throughout the 2000s and 2010s I was involved in a series of projects focused on creating open data infrastructures for storing and sharing data, as well as creating online interactive tools for visualizing and mapping data, including: the All-Island Research Observatory (AIRO) that compiles datasets that span the Republic of Ireland and Northern Ireland; the Irish Qualitative Data Archive (IQDA) that stores and shares qualitative social sciences data; the Digital Repository of Ireland (DRI), which is a national trusted digital repository for storing and sharing the digital collections of galleries, libraries, archives, museums and universities; and the Building City Dashboards (BCD) project which created the Dublin and Cork Dashboards that inform users how the cities are performing. Each of these initiatives has sought to formulate guidelines and best practices for building and maintaining data infrastructures. They are long-term endeavours and all still active.

From the late 1990s, I also became increasingly attentive to social theory and critiques of traditional science. My initial interest concerned the implications of the internet for everyday life and the nature of software. While I was trained to code as a key part of a Master's degree in Geographic Information Systems (GIS), and wrote dozens of programs to conduct my doctoral research, subsequently I learned how to deconstruct software and its embedded values and to think critically about what the code was being used for. In the early 2010s, my focus drifted from code to data, writing a book, *The Data Revolution*. This book adopted the lens of critical data studies to consider the nature of big data, open data, data infrastructures, and to map out their consequences for society. Just as the book went into production, I started a new five-year study, The Programmable City, which examined critically the role of software and data in managing and governing cities, teasing apart the discourses, technologies, practices and social implications of data-driven urbanism.[7]

In addition to academic work concerning data and data infrastructures, over the years I have served on various advisory boards/panels, such as the Data Forum of the Department of Taoiseach (Irish Prime Minister's Office), the Irish Census Advisory Board, the Irish Research Council, Dublinked (Dublin's open data portal) and Smart Dublin (the city's smart city body),

the Audit Committee for the Irish Central Statistical Office, the Irish Social Science Data Archive, and the National Consultative Panel on Open Data. I have also conducted public-facing data work. For example, between 2009 and 2015 I regularly wrote data stories for the blog *Ireland After NAMA*,[8] presenting just-in-time data analysis on topical issues – mainly housing, planning and population change. This involved undertaking and publishing an analysis of government or industry data within a couple of hours of its release. This was often followed by discussions with print journalists and radio interviews.

Data Lives

The aim and the challenge of this book was twofold. First, to translate my knowledge of working with and thinking about data and recast it into stories that are more likely to chime with people's own experiences. Second, to use the essays and storytelling to highlight why we need to think and act critically with respect to data, both in terms of how they are produced and used. These ambitions to playfully and critically examine data is captured, I hope, in the title of the book, '*Data Lives*'. It is intended as a kind of pun, with 'lives' rhyming with both 'gives' (as in life of data) and 'hives' (as in living with data). About a third of the chapters discuss initiatives located in Ireland. That is inevitable given I have spent my entire career on the island, two years in the North and the rest in the Republic. The themes discussed in these essays and stories are equally applicable to elsewhere, as the other illustrative material related to the United Kingdom, United States, China, Hong Kong and Australia make clear.

Part II reveals the cooked nature of data and details the praxes and politics that shape the data life cycle from generation to destruction. It's divided into 12 short chapters, six of which are short stories and six essays. The part opens with a blind date between two researchers who have very different notions about the nature of data and the ethos and practices of science. An essay examining the nature of data from an etymological, philosophical and technical point of view then follows. 'Gridlock' charts the data journeys and transformations that take place across

the network of cameras, sensors and software that makes up a traffic control system. Issues of data quality and veracity in open datasets are then examined using a variety of examples from the Irish data system. 'How to Lose (and Regain) 3.6 Billion Euros' imagines a conversation between two senior civil servants when they realize that the Irish government has lost 3.6 billion euros through a spreadsheet error (which really happened and wasn't noticed for over a year). Data interoperability and the difficulties in harmonizing data across jurisdictions is explored using Ireland/Northern Ireland and Metropolitan Boston as case studies. A conversation between open data advocates and a civil servant in charge of the process reveals the challenges of getting government data made open. The technical and political trials involved in building a suite of open data tools is then scrutinized by charting the development of the Dublin Dashboard. This is followed by an argument between two researchers concerning the epistemology, methodology and ethics of data science versus traditional science in studying fertility. The role of finance and the politics of collaboration is made clear through charting the development of the DRI. How choices and decisions concerning the analytics applied to data shapes outcomes is revealed in an account of a working session between academics and a government minister to devise and implement an 'objective' method for allocating government funding. The final essay examines the transitory nature of data and its deletion, by either design or accident.

Part III details what is at stake living in a data–driven world. The part opens with an essay that charts the transition from an analogue to a digital world, its effect on data footprints and shadows, and the growth of data brokers and government use of data. How profiling and social sorting shape consumption and entertainment through recommendations and nudges is then detailed. This is followed by an account of sousveillance and how we produce, monitor and react to data relating to ourselves, thus creating a quantified self. The consequences of dataveillance is explored with respect to how personal data shadows led to unemployed, volunteer firefighters in Australia's recent fires losing their benefits. 'Management through Metrics' charts how public and private sector organizations are increasingly using key

performance indicators and technocratic procedures to manage work and workers and its consequences. The implications for citizens of data-driven management is examined by charting the issues of living in a smart city testbed area, demonstrated through a walking tour for local residents, led by a public official. How these technologies are deployed as mass surveillance and social credit scoring in China and their threat to democracy is then examined. This is followed by charting the Kafkaesque procedures involved in data-driven airport security. The benefits of evidence-informed policy over anecdote is explored through an account of the financial crash in Ireland and the effect of creating public data stories. 'Data theft' details the consequences of data breaches for a company and its customers. How citizens can take a more active role in using data for the public good through civic hacking, citizen science and data justice initiatives is then examined. The final story charts how a group of citizens seek to challenge systemic and institutional racism within their city by building their own datasets and tools.

Part IV opens with a chapter that pulls together the themes explored in Parts II and III in a discussion of the life of COVID-19 data, how it has been used to reshape our daily lives by directing intervention measures, and how new data-driven technologies have been deployed to try and help tackle the spread of the coronavirus. As a whole, the book argues that it is crucial to appreciate the praxes, politics and effects of both the life of data and living with data if we are to imagine and create different kinds of data-infused lives. Hopefully it prompts, on the one hand, a critical reflection on the role of data in your own lives and the data-driven world we are living in, and on the other, an impetus to take a more active role in shaping the terms in which those data are produced and used. With respect to the latter, the final chapter examines what kind of data future we want to create and strategies for realizing our visions.

I have tried to structure the book so that it can be read from cover to cover, with a long arc of argument threaded throughout, but also so that the reader can dip in and out and engage with particular issues and themes. Hopefully the essays and stories resonate and provide insight and food for thought in navigating your own data lives.