Book Proposal

TODO

- rewrite the chapter outline
- highlight the scientific/technical literature as key working material
- add data science
- add other literature on data -- Beer, Burrows, Savage,

Into the data: learning from machine learning

Adrian Mackenzie Sociology, Lancaster University Bailrigg, LA14YD, UK

Overview

> The key question isn't `How much will be automated?' It's how we'll conceive of whatever *can't* be automated at a given time. [@Lanier_Who Owns the Future_, 2012, 77]

This book sets out to construct some ways of critically thinking about data that neither blithely affirm beliefs in the power of data, nor reject beliefs in data as pure hype. It focuses on practices and techniques at the heart of contemporary media, sciences, government, commerce and industry: *machine learning.* Machine learning is widely used in industry, science, commerce, media and government to program computers to find patterns, associations, and correlations, to classify events and make predictions on a large scale. The Microsoft Kinect motion-sensing system is a mundane example. It uses a machine learning technique called `decision trees' to identify and classify gestures and poses. Similar predictive and classificatory mechanisms appear in many gadgets (for instance, the face recognition feature in many digital cameras). Almost any search engines' results

are shaped by machine learning algorithms, as are many of the recommendations and suggestions generated by social media platforms. The detection of abnormal tissue growths in medical scans or the classification of biological function in whole genome sequencing projects. Machine learning, as a set of techniques for classifying and predicting, is widely heralded in the form of data mining, predictive analytics or knowledge discovery as a vital component of contemporary innovation and economic growth. Machine learning is heavily used in search engines, social network media, high-frequency trading, in increasingly data-intensive scientific practices in astrophysics, genomics, ecology or material science, in manifold devices and control systems, and of course, by government intelligence and security agencies in massive data surveillance programs.

There is intense technical interest and investment in these techniques, and they have developed rapidly in the last two decades. Yet they have hardly been discussed in humanities and social science literature, even in digital humanities, which heavily relies on them (for instance, in the growing use of `topic models'). Examining key machine learning techniques and practices drawn from social network media, finance markets, image processing, robotics, and contemporary sciences such as genomics and epidemiology, *Into the data* does not exhaustively describe who does machine learning, where and how. It seeks rather to identify how and why such techniques have moved or circulated so widely. At the same time, *In the Data* is an experiment in writing for humanities and social science audiences that combines code, data and diagram, text, and number with the goal of imagining doing machine learning differently. There is something quite algorithmically important in machine learning that may well have already changed both the objects and methods of humanities and social sciences. In both analysing and re-purposing techniques found at the intersection of contemporary sciences and network media, *In the Data* attempts to both make potent data practices more visible, and to facilitate greater overlaps and entanglements between various science and political, economic and cultural processes associated with data.

Key concerns for the book are:

- The modes of existence of data (in the forms of `big data,' `open data,' the rise of `data analytics' and `data science') should be analysed critically by situating them in relation to specific settings, techniques and practices. These settings, techniques and practices have complex genealogies, criss-crossing sciences, industries, military, commercial and governmental domains. While critical accounts and indeed skepticism about the value of data are appearing, the provenance and distribution of data practices such as learning algorithms, predictive modelling and classification needs much more critical attention. This book focuses on the role of machine learning (or data-mining, as it is called in some domains) because the dynamic and varied life of these techniques themselves have received least attention of all. So much depends on and is decided by the models, algorithms and techniques of machine learning, yet they are hardly ever discussed in their own right.

- Humanities and social science responses to data techniques should be methodologically and conceptually inventive, and include appropriation and re-purposing of the techniques and practices. This is major undertaking for the book. It seeks to a broad-ranging way of thinking about data, and what is `in data' that both soberly appraises beliefs about data, and offers ways of evaluating what is at stake in data as it is processed, explored, organised, filtered and classified by machine learning techniques. The question here is: what can humanities and critical social sciences learn from machine learning?

Approaches used in the book

A broad ethico-political concern underpins *In the Data*. Much contemporary data practice is closely allied to the predictive ambitions of business, the military and states, as well as sciences and media. The recent upswing in data talk continues and intensifies the techno-scientific `Regime of Computation' (Hayles 2005). It is no accident that autonomous milil-

tary vehicles, large-scale analysis of sentiment social media for commercial or security purposes, or face recognition for national border control are iconic examples of machine learning in action. A key question for critical humanities and social science researchers, as well as activists, non-government groups and civil society actors of many kinds, is how to make sense of such data practices. They are hard to render visible since they take place largely on platforms that are not publicly accessible. Rendering such practices visible, learning to track their workings, and inventing different ways of working with them: these concerns lie at the core of the analytical and experimental writing practices of *In the Data.*

Broadly speaking, the writing seeks to respond to the long-standing call for what in a widely cited passage Donna Haraway more than a decade ago termed `diffraction': `What we need is to make a difference in material-semiotic apparatuses, to diffract the rays of technoscience so that we get more promising interference patterns on the recording films of our lives and bodies' (Haraway 1997, 17). There are a growing number of attempts to adapt and reinvent data practices such as machine learning for less overtly biopolitically laded, security-minded or commercially-motivated purposes (the growth in university `data science' courses might be one example [@schutt_doing_2013]; see the `OccupyData' group in New York, N.Y. for one such example (OccupyData 2013); many citizen science projects have something of this flavours to them too). Some of these will be discussed in the course of the book.

In order to bring data, code, images and text together more fluidly, the book relies on some straightforward `executable paper' formats developed in recent scientific publishing. It mingles code written in R, an important statistical modelling, data manipulation and visualization programming language, with code written in `Python` and `Javascript`, two of the most popular general programming languages in use today. Some of the empirical materials in the book have been garnered, ordered, analysed and displayed using `R`, `Python` and `Javascript`. More importantly, since all machine learning techniques are implemented in code, working with code is an important way of navigating and exploring

the architecture of these techniques. Code allows movement, visualization and demonstration of machine learning in practice. Code excerpts form part of the text, and will be the object of commentary and analysis, alongside diagrams and graphs generated by the code. All of the code will be available at a public code repository (github.com).

The motivation for the executable format of this book is partly ethnographic and partly experimental. A long line of ethnographers have learned to do what they are observing (as in `observant participation' (Wacquant 2004)). This has include working in factories, going to prison, spending time in isolated, far-flung or ostensibly boring places, learning techniques ranging from weaving and cooking to playing the piano or programming robots. Ethnographic presence in a particular setting is normally documented through text, photographs, diagrams and occasionally film or audio recordings, and aims to make sense of this setting in ways that both resonate with the people who live there and with people who don't. The forms of observant participation in this book include competing in machine learning competitions, reconstructing and implementing algorithms, building predictive models and visualization as a way to present machine learning. It treats reading and writing code as an ethnographic practice, and code as part of the ethnographic writing process. Hence forms of data practice used in producing this book are also the objects of its analysis. Several versions of this recursivity will appear in the chapters.

The experimental character of this writing entails both practical and theoretical challenges. Practically, the book experiments with a range of code constructs, some key mathematical formulae as well as data tables and data graphics. Such constructs are not typically found in humanities and qualitative social science research, although they are extremely common in many scientific fields. The presence of code, formulae and graphics in *In the Data* is not meant to instruct readers in machine learning algorithms or statistical inference. Accompanied by forms of explication and commentary, they are intended to allow readers to pay close attention to the forms of thought or contemporary equipment [@rabinow_anthropos_2003] at work in the manifold data practices of sciences or business analytics, and to begin to borrow, appropriate and re-purpose some of the patterns of

thought for different purposes. The theoretical ambition here is to treat the code writing also as a way of constructing concepts, metaphors and ways of speaking about contemporary entanglements of subjectivity and computation.

The architecture of the book

The book is organised around two different axes.

1. On one axis, the `technique axis,' the chapters of the book catalogue, document and analyse some of the most widely used machine learning techniques of working with data (Hastie, Tibshirani, and Friedman 2009). As mentioned above, the techniques analysed on this axis -- linear and logistic regression models, decision trees, clustering algorithms, neural networks, support vector machines and Markov Chain Monte Carlo simulation, -- are used across scientific, industrial, biomedical, commercial and military settings. Their extraordinary success in populating these domains cannot be explained in terms of IT or digitisation in general. The case studies explore how these techniques, and their implementation as `learning algorithms,' rely on widely shared assumptions about the problems of knowing, acting, responding or predicting how things happen. To the extent that a situation can be reshaped to conform to these assumptions, these techniques gain traction.

2. The other axis of the book is `recursive reconstruction:' the attempt to show how specific situated entanglements of subjectivity and data practice might open up different ways of thinking about contemporary experience as it is increasingly pervaded and subtly (or not subtly) modulated by data-driven processes. Along this axis, chapters of the book engage with the messiness, complications, and frictions of working with datasets, with predictive models and forms of visualization ranging from standard plots of curves to network graphics. The diagrams, functions and code constructs arrayed along this axis are drawn from scientific fields, or from commercial applications where data is made available through APIs (Application

Programmer Interfaces). The reconstruction of data practices draws on the pragmatist philosopher John Dewey's notion of philosophy as an empirical reconstruction of experience (Dewey 1957; Dewey 2004). The kinds of experience reconstructed range from encounters with databases, with stream of numbers of varying kinds, with statistical predictions, with various engines that classify, recommend or in general find patterns. Each chapter seeks to address a facet of this. At various points, these reconstructive moves will be linked to broader debates around politics, ethics, publics, democracy, power, equality and differences.

Existing academic literature and framing of the book

The existing literature relevant to this monograph come from a variety of disciplines. The critical work on data is largely found in science and technology studies (STS), and some parts of information science. Software studies and anthropological accounts of software cultures are highly relevant in reading machine learning algorithms and data visualization software. A broader theoretical background here includes recent reappraisals of pragmatism (particularly William James, C.S Peirce), feminist and other work on materialities, as well as strands of largely European contemporary philosophy relating to experience, space-time, science, calculation and events. A final reference point comes from recent attempts in social sciences and humanities to reinvent methods of research.

In STS, work on calculation (Callon and Law 2005), data practice (Edwards et al. 2011), databases (Bowker 2005) and digital data more generally (Latour et al. 2012) have extensively discussed how science assembles numbers, observations, instruments, readings and databases. This work forms an important part of the background of this book since machine learning has heavy mathematical underpinnings and institutional dynamics. The STS work has broadly re-theorised many different aspects of data, ranging across collection, measurement, calculation, archiving, labelling and visualising. Much of this work is based on ethnographic case studies of laboratories, technical devices, standards and

controversies. It has notably developed ways of analysing its objects relationally (as in actor-network approaches), and with an eye on entanglements and hybridisation of human and non-human entities. While *In the Data* is not by any means a standard laboratory ethnography, it does rely on practices of participant observation and analytical approaches found in STS.

The history of statistics, number and mathematics also frame important aspects of *In the Data*. Works such as Theodore Porter's *Trust in Numbers* (Porter 1996), Lorraine Daston's work on probability (Daston 1988), or Alain Desrosiere's *The Politics of Large Numbers* (Desrosieres 1998) amongst others not only provide background for many of the statistical techniques used in machine learning, they suggest that numerical data and numbers have had an eventful course of development from the 18th to the 20th century. While much of this historical work leaves just around the time when machine learning approaches are emerging (1960-1970s), it provides an extremely useful way to contextualise key traits in the contemporary data practice, ranging from genres of visualization to underlying concepts of probability, chance or error.

The nascent field of software studies has begun to develop ways of analyzing software and code, ranging from source code files to large assemblages. Coupled with media studies and media archaeology-type approaches, software studies has developed geneaologies, critical framings and methods of reading many different aspects of software. Work in this field ranges from quite high-level analyses such as Wendy Chun's *Programmed Visions* (Chun 2011) or Lev Manovich's *Cultural Software* [@manovich_cultural_2011], or Alex Galloway and Eugene Thacker's work (Galloway and Thacker 2007) through to studies of specific code objects (as for instance in many of the entries in the *Software Studies: A Lexicon* volume (Fuller 2007)) or analysis of code as speech (Cox and MacLean 2012). Other work should be included here (along with related work on `platform studies'), but for present purposes, the key influence of software studies consists in its treatment of software, computer code, algorithms and protocols as first-ranking objects of social and cultural analysis. Some literature on data flow and data visualization has started to appear,

but machine learning doesn't figure in it [@beer_popular_2013]. A broader background of work on software cultures [@kelty_two_2008; @coleman_code_2009], with their important re-thinking of publics, property and value, is also relevant, but differs greatly in its emphasis on software production and social movements.

A broader range of theoretical approaches informs this book. These include on events, materiality, experience, and capitalism from scholars that include Brian Massumi on radical empiricism (Massumi 2000), Nigel Thrift on time-space signatures of calculation (Thrift 2005), Celia Lury on topological conceptions of culture [@lury_introduction_2012], Manuel Delanda on simulation in philosophy and social science (DeLanda 2002; DeLanda 2011), Anna Munster on conjunctive experience in networks (Munster 2013), or Luciana Parisi on the contagiousness of computation (Parisi 2013). Many of these authors share an interest in re-thinking notions of experience, body, event, time-space and materiality in the context of ongoing transformations of media and technology. Many of them draw on philosophers such as William James or A.N. Whitehead to question taken-for-granted concepts of nature, life or agency. Again, this loose coalescence of work cannot be adequately summarised or even limned here, but it indicates something of the theoretical registers on which *In the Data* will work.

The final framing body of work is even less coherent, but nevertheless important: it largely comprises threads of research and debate about methods today in social sciences and humanities. This literature tends to treat the growth of digital data as both posing a problem and an opportunity for research in social sciences and humanities. The problem, as framed by sociologists such as Andrew Abbott (Abbott 2001) or Mike Savage (Savage 2009),is that existing quantitative methods in social science cannot match the efficacy of quantitative methods in the natural or applied sciences, nor those used in business and marketing (e.g. as in analysis of transaction data). Some social scientists advocate the development of `computational sociology' (King 2011). A version of the same crisis can be found in digital humanities, and has prompted developments such as `cultural analytics' (Manovich 2009). Commonly, these responses advocate a pattern-based approach to working with so-

cial or cultural data, and in this respect, they mirror some of the commitments in machine learning to finding the function that generates the data. Whilst very sympathetic to and in some ways aligned with these debates, *In the Data* also aims to offer something other than a set of `better' data methods. It is more closely aligned with work that seeks to re-think social science methods in the light of new flows of data [@marres_redistribution_2012] and its temporalities [@uprichard_being_2012].

Readership and market

The readership for the book is quite diverse, since data practices and indeed machine learning itself are of interest to a growing audiences. One set of readers I have in mind for the book come from disciplines such as sociology, anthropology, media and cultural studies, and social geography who are grappling with the promise of data both as an object of analysis and in terms of a transformation of their own ways of researching. Another set of readers for the book come from the burgeoning `data science' courses being offered in North American, UK, SE-Asian/Pacific, and European universities. While these courses are largely focused on techniques of organising, visualising and modelling data, many of them are also open to thinking about the transformations in knowledge and value associated with contemporary data practice. The book is written very much with these kind of readers in mind. It will minimize reference to social theory in order to maintain more accessible to these readers. While I am keen to keep the social and cultural theory side of the book in the margins, the book will introduce some technical terminology, and indeed some mathematical formulations. But this technical material will be analysed rather than assumed as background.

Timetable

Many of the chapter exist in draft form, or as conference papers. Writing an introduction, conclusion, and revising the drafts will take roughly 11 months.

- draft conclusion: 1 month
- draft introduction: 1 month
- draft chapter 2: 2 months
- revise chapter 3,4,5,6,7,8 drafts: 3 months
- revise chapter 2: 1 month
- revise whole manuscript: 3 months

Format of the book

The book has a standard chapter format. It will include several dozen code-generated figures, diagrams or plots, as well as a number of tables. The Python and R code, and datasets used to generate these components of the text will be available through the public code repository github.com. The Markdown text of the book will be also part of this code repository. Electronic versions of the book will display colour versions of the plots, and be hyperlinked to both the code-data components on github, and to various relevant URLs. The predicted wordcount is 85,000 - 90,000 words. It will include approximately 20 diagrams or graphics.

Chapter outline

introduction: Into the Data

Key examples: kittydar, DARPA challenge, credit card checks, cancer prognosis
The introduction will begin with several relatively familiar examples drawn from a variety of fields over the last decade or so -- handwriting recognition, face recognition, autonomous robots (Thrun et al. 2006), credit card checks, and cancer prognosis. It will highlight these examples as symptoms of the wide-ranging investments in knowledge, control, prediction and decision-making associated with data flows, and at the same time,

suggest how these tracking some of the transformations might elicit changes in how humanities and social science researchers understand their own work.

These examples will also provide a preliminary overview of the techniques of machine learning discussed in the book -- supervised and unsupervised learning, the differences between classification, regression, and clustering and important notions such as learning and prediction. They will also highlight constrasts between disciplines such as computer science and statistics that develop machine learning techniques, as well as illustrate the overlaps between data-mining, pattern recognition, knowledge discovery, artificial intelligence, machine learning etc. Practically, these examples will also implicitly present some fo the methods used in the subsequent chapters, including the role of databases, data structures, code constructs, diagrams, and algorithms in typical scientific and industry practices of modelling.

These examples will also stage some of wider questions in the book about the promise of data. These include the oft-mentioned `end of theory' prediction (Chris Anderson, *Wired* magazine, 2008), and the many claims and controversies about data analytics, machine learning and the `power of big data' in physical, life and social sciences, in business, government and industry. Claims about power of data, and responses to these claims -- ranging from downright skepticism to enthusiastic embrace -- will be discussed here with an eye on what these debates about data mean for research practices in the social sciences and humanities themselves in terms of their topics of research and how they do research.

Finally, the introduction will sketch the themes of `in the data' and `modes of machine thought,' drawing on a range of work drawn from pragmatist philosophers such as C.S. Peirce (abduction and diagrams), William James on experience (James 1996), John Dewey on `reconstruction' (Dewey 1957), Alfred N. Whitehead on `abstraction' (Whitehead 1958) and from recent social and cultural theory such as Isabelle Stengers on experiment (Stengers 2008); Gilles Deleuze & Felix Guattari on scientific functions, and (Deleuze and Guattari 1994); Celia Lury on topology [@lury_introduction_2012]). In order to contextualise

forms of data thought, the introduction will also sketch some points of departure drawn from software studies work on algorithms and databases, science studies work on calculation, statistics, number, device, image and diagram, as well as accounts of subjectivity, experience [@berlant, 2007] or [@murphie, 2010] and materiality cross-cutting all of the above. This spectrum of work from across disciplines provide scaffolding and departure points for much of the book.

Part I: Data Form and Function

The four chapters of Part Iexplore underlying the major underlying spatial, ordering and counting practices of the many different techniques comprising machine-learning, pattern recognition and data-mining. It seeks to show how these practices diverge and multiply across a range of settings, and how they coalesce around a set of generic intuitions of difference that are both powerfully general yet highly constrained.

1. Writing about data

Key examples: house prices, Fisher's irises, R programming language; Python scikit-learn; the Titanic survivors;

Key techniques: linear models, perceptron This chapter is primary a methodological discussion that addresses several different problems in working with machine learning. These problems range from quite philosophical issues through to quite practical ones. At the philosophical end, it draws on various pieces of recent and not-so-recent theoretical work (for instance, the anthropologist Paul Rabinow's work on the concept of `equipment' [@rabinow_anthropos_2003]; the philosopher A.N. Whitehead on the spatial dimensionality of thought (Whitehead 1958); the philosopher Anne-Marie Mols' work praxiography -- writing about practices (Mol 2003)) to discuss how we might think about working on highly technical or scientific areas such as machine learning in ways that allow consider-

ation of their more general situation. It also draws on some cultural and psychoanalytic accounts of architecture and objects (Bollas 2009; Wilson 2010) to suggest how the researcher him or herself relate to objects of research. Finally in this vein, the chapter poses the methodological problem of working with large bodies of scientific and technical literature, and illustrates this by describing the growth of machine learning techniques in science and engineering research since the early 1960s, focusing on the growth of key techniques and algorithms [@kelty_ten_2009].

With these questions about thinking, techniques, subjectivity and literature on the table, the chapter then presents a series of vignettes that display some of the ways in which research and writing critical accounts of data cultures and data economies can make use of the tools, techniques, instruments and services of `data science' to generate textual, diagrammatic and modelised accounts of contemporary culture. A standard teaching example -- house-price prediction -- links these vignettes, but the real focus here is on two foundational issues: how machine learning treats data as a dimensional material that it seeks to reshape or recase in different dimensions (`models'); how implementing machine learning techniques shifts our relation to them.

Woven throughout this discussion of a praxiography of data, [TBC] Via a discussion of the development of R, the chapter analyses the transverse, cross-disciplinary implementation of machine learning techniques in recent decades. It describes some of the transformations in software, network and scientific cultures that underpin the recent growth in data techniques and methods. These range across transformations in statistical science associated with greater computational capacity; the mutations in network, database and digital device architectures and infrastructures that yield much greater abundance of data in various forms; and the intermeshing of knowledge economies with the media, communication, transaction, transport and logistics systems. It will trace how the lateral associations and multivalencies of data have developed through key software artefacts such as the widely used R programming language, and in generic programming languages such as Python. Coming from the author's own history of working with machine learning or online ac-

counts of machine learning, well as the ecology of thousands of software packages associated with the statistical programming language `R`.

2. Looking at data

Key examples: Boston house prices; cancer prognosis; digit recognition; credit scoring

Key techniques: logistic regression; *k*-nearest neighbours, neural networks -- recursion, movement, evocative objects, partial observers, visualisation, etc; functions and states of things; linear regression

Graphs and plots stand at the centre of vision in contemporary data and knowledge economies, whether in the time series plots of financial markets, the scatter plots of scientific publications or the network graphs of social media. The topography of curves, lines, points and network diagrams present views of data, and they are indispensable to many of the classification, decision and prediction techniques of machine learning. Such visual forms, with all their associate aesthetics (code, line, typography, animation) are themselves convey expectations and predictions about the changes in the data practice, especially in the form of the curves showing growth of data.

This chapter examines the proliferation of data-supported curves and lines in terms of *functions*, the underlying generating mechanisms of curves. Machine learning is conceptually framed as a form of function-finding. Drawing on statistical machine learning texts (Hastie, Tibshirani, and Friedman 2009), and more philosophical accounts of functions (e.g. (Deleuze and Guattari 1994; Whitehead 1958)), the chapter introduces the key instances of the function in machine learning, shows how functions underpin the generation of curves, and how movement along lines, curves and across planes. While later chapters will range across a variety of mathematical functions and forms, this chapter will focus on two of the most widely used machine-learning technique, linear regression model and

its classifier version, logistic regression model, and the *k* nearest neighbour algorithm. It will discuss these important techniques from the perspective of the forms of relationality, referenciality and indexicality associated with them.

Connecting aesthetic and mathematical data practices, this chapter suggests that finding the functions that generate lines and surfaces in data is a powerful form of imitation that tends to remake the world in certain ways. This re-making may be inimical to social life, or not. The chapter also suggests that the production of curves through software packages and libraries and through various visualization techniques is worth investigating as a signifying social practice in its own right. The architecture and practices associated with graphics and plotting libraries offer a way to trace some of the processes of imitation and invention associated with forms of data thought.

3. Finding patterns in data

Key examples: hunch.com;

Key techniques: decision trees, neural networks, support vector machines For the last decade, the best-performing `off-the-shelf' machine learning algorithm has been a technique known broadly as `support vector machines' (SVM; see [@vapnik_nature_1999]). The chapter examines the architecture of this widely used algorithm both against the background of a spectrum of other statistical machine learning techniques, and more importantly, in terms of the *forms of movement* it brings to data practice. The key focus in this discussion is the dimensionality of data, and how dimensionality is managed in machine learning. While curves and functions, as discussed the previous chapter, engender senses of change and movement, the advent of increasingly extended and particularly `wide' datasets (many variables) implies models that embrace high-dimensional abstract spaces. Since the 1950s, scientists have been aware of the `curse of dimensionality' [@bellman_adaptive_1961], which arises when the dimensions of the data increase. Algorithms such as SVM, and implicitly other highly successful ML algorithms such as neural

networks, manage this dimensionality very differently to the regression models that have been the mainstay of statistical modelling for a century. Rather than trying to reduce the dimensionality of the model to a line, plane or hyperplane that best fits the datasets, SVM expands the dimensionality of the model massively, sometimes infinitely.

4. Believing in machine learning

Key examples: Microsoft TrueSkill; Obama election data team

Key techniques: Monte Carlo simulations and MCMC; Bayesian networks; The topic in this chapter is the role of randomness and chance in machine learning. While statistical techniques and practices have been discussed in previous chapters, this chapter foregrounds the changes in the so-called `Bayesian revolution' in statistical practice that took shape in the early 1990s, and in particular, the key algorithmic technique used in Bayesian statistics, Markov Chain Monte Carlo simulation (MCMC). The computationally intensive techniques of Bayesian analysis treat all numbers as potentially random variables; that is, as best described by probability distributions. The ensuing popularity of Bayesian inference is a striking example of transverse momentum of methods across fields, and the chapter will trace some of the ramifications of the heavily-used MCMC technique in fields ranging from nuclear physics, image processing to political science and epidemiology.

The chapter traces two important implications of this technique. First, because it is so computationally intensive, MCMC and Bayesian inference, although statistically powerful, are difficult to apply to many dimensional datasets. So Bayesian computation iconically figures the limits of contemporary data practices, with their ambitions to incorporate all available data into calculation. Second, in certain ways this technique challenges us to re-evaluate how we think about numbers. By following some of the ways numbers circulate through MCMC algorithms, we can discern to a semiotic-material faultline running through contemporary number formations. Numbers semiotically and materially embrace

both events and degrees of belief. If numbers are crucial in the data economy, then instabilities in their mode of existence will affect much of what happens to data. While much of the machine learning taking place in commercial and operational settings is decidedly non-Bayesian, the popularity of MCMC and Bayesian approaches in contemporary sciences suggests a tension in what counts as number.

Part II: Problems with Data

The three chapters of Part II explore what happens as the major intuitions of machine-learning encounter things, events and people. The chapter deal with transformation in science, governmentality and work.

5. What does data do for things?

Key examples: cancer prediction; the ENCODE project;

Key techniques: decision trees, random forests, self-organising maps This chapter of the book concerns perhaps the most concerted effort to count things every undertaken in the life sciences: sequencing in contemporary genomics (that is, post-Human Genome Project and after the advent of so-called `high-throughput' or `next generation sequencers'; this is roughly 2007 onwards). Genomics is a provocative form of data thought in several respects. First, it relentlessly treats one type of quite flat or mono-dimensional data -- nucleic acid sequences or `base sequences'-- as the key to potentially biological processes in all their plasticity and mutability. While it is not at all clear that this treatment will be effective, it has generated ways of generating shape or pattern from data that stand as a limit case for data-driven research more generally. Second, genomics is a scientific discipline almost overwhelmed by the effectiveness of its own instruments in generating data. The rate of production of sequence data from next generation sequencers exceeds Moore's Law, the standard 18-24 month doubling time for the number of transistors in an integrated cir-

cuits. This sequence data needs to be stored and analysed in rhythms that differ from many other settings where the growth of data can be managed through more memory and computer processing speed. Third, genomic researchers have been extraordinarily adaptive in positioning their work on the borders of cutting edge infrastructure development, machine learning and data-mining, and the life sciences. Genomics (and bioinformatics) loom large in the machine learning literature itself since the mid-1990s. The flatness of sequence data has been heavily leveraged by this positioning. The biological objects of genomics - genomes -- have been progressively transformed and re-shaped in ways that might be instructive for data more generally. This chapter explores how machine learning has imbued genomes with an increasingly topological character (and particularly, the growth of `topological data analysis' [@carlsson_topology_2009; @singh_topological_2007] as well as the topological turn in culture [@lury_introduction_2012]), and practically, with the rich ecology of programmatically accessible bioinformatics tools and archives that on the one hand permits sequence data to move relatively freely (especially in comparison to much commercial or even social media data), but on the one hand poses question as to who wants or needs the data.

6. Are there enough numbers in world?

Key examples: A/H1N1 London 2009, Google Flu;

Key techniques: transmission models, nested models A predominant narrative around data in many contemporary settings urges that more data makes all problems solveable. This narrative is usefully accompanied by an `abundance of data' (`big data', `data deluge', etc) narrative, in which the advent of data corresponds to a groundswell change in how we make sense of and intervene in events. Versions of these narratives surface in genomics, business analytics, and infrastructure management (e.g. in smart energy grids), as well as crisis-events such as financial collapses or epidemics. Via a case study of different data flows during the 2009 A/H1N1 `swine flu' epidemic, this chapter develops an alternative

narrative of data flow in terms of number supply chain logistics. The chapter reconstructs a real-time epidemiological model that combines clinical reports, laboratory test data, web surveys, urban population mixing patterns in order to disentangle biological and social forms of contagion and infection during the 2009 epidemic in London. In reconstructing this model, a model that is typical in complicated engagement with numbers of diverse origins, the chapter will suggest that the largely homogeneous data flows envisaged and embraced in many forms of data practice largely ignore the problem of the interactions between different agents. It specifically contrasts the much publicised Google Flu Trends approach to 'flu prediction, which is based on search query volumes, with epidemiological models based on multiple forms of surveillance data. The chapter argues that data practices during crises or times of great uncertainty, entail hybrid integrations of existing data practice and new forms of data.

7. Optimising machine learners

Key examples: Predictive Analytics World 2009; Kaggle facebook retention competition; Kaggle R recommendation engine; Sage Bionetworks

Key techniques: ensembles, RandomForests The chapter focuses on the forms of subjectivity associated with contemporary data practice, situated within plural data and knowledge economies. Software developers, hackers, statisticians, `data scientists,' as well as social scientists, are changed by forms of data thought. The case study in this chapter is data prediction contests run by the Kaggle.com as well as academic-based competitions. In these competitions, competitors from diverse technical and geographic backgrounds compete to construct predictive models for specific datasets -- the Netflix recommendation competition; the Facebook `find a friend' competition; or the Titanic survivor problem -- using whatever machine learning techniques they can bring to bear. These competitions, conducted on web-based platforms, are useful ways to track contemporary data practices. Combined with some examples of presentations by academic researchers

(for instance, Stanford University's Andrew Ng whose YouTube lectures haved attracted 100,000s of views), industry conferences (for instance, at the annual Predictive Analytics World events), this chapter will track the kinds of technical and affective investment associated with popular data modelling techniques such as Random Forest. It is possible, I will suggest, to read a technique as a partial subjectification, in that it affects how they experience and materially engage with data. In order to apprehend the character and texture of these subjectifications, the chapter links university research, commercial and non-commercial adoption, and flows of technical expertise. Again, this chapter has some auto-ethnographic vignettes, as the author has participated in some of these competitions.

Conclusion: Out of the Data

The conclusion draws together the main threads running through the previous chapter, and sets out a series of questions and provocations for thinking with data. Crucially, the conclusion will stand back from the much more hands-on approach to data and data practice adopted in the preceding chapters in order to think more about we -- social scientists, humanities scholars -- might invent or create in the midst of data. While this book has a critical angle to it (so many claims about and beliefs in data plainly deserve critique for their conservative and naive approach to things), it is principally concerned with conceptual invention through doing things with data. The work of learning about machine learning, and learning about it in a way that is deeply embodied or practically embodied, brings with it altered ways of thinking about, questioning and integrating what is happening to data more generally. It highlights the key argument that has run through the book about the plural dimensionality of data as it is aggregated, tabulated, summarised and modelled in contemporary data and signal processes, and as well as the extraordinary mobility or kinetic energy of generic machine learning methods. In discussing the shifting dimensionality of data, and the kinetics of methods, the conclusion will attempt to sketch out how some promising ways of thinking with data might proceed.

References

Abbott, Andrew. 2001. *Time matters: on theory and method*. University of Chicago press.

Bollas, Christopher. 2009. *The evocative object world*. Taylor & Francis. http://books.google.co.uk/books?h

Bowker, G. C. 2005. *Memory practices in the sciences*. MIT Press Cambridge, MA.

Callon, M., and J. Law. 2005. ``On qualculation, agency, and otherness." *Environment and Planning D* 23: 717.

Chun, Wendy. 2011. *Programmed visions: Software and memory*. The MIT Press.

Cox, Geoff, and Alex MacLean. 2012. *Speaking Code: Coding as Aesthetic and Political Expression*. Cambridge MA: MIT Press.

Daston, Lorraine. 1988. *Classical Probability in the Enlightenment*. New Brunswick, N.J.: Princeton University Press.

DeLanda, Manuel. 2002. *Intensive Science and Virtual Philosophy*. London & New York: Continuum.

---------. 2011. *Philosophy and simulation: the emergence of synthetic reason*. Continuum. http://books.google.co.uk/books?hl=en&lr=&id=F5wvXkJwFwkC&oi=fnd&pg=PP7&dq=manuel+d

Deleuze, Gilles, and Félix Guattari. 1994. *What is philosophy?*. *European perspectives*. New York□; Chichester: Columbia University Press.

Desrosieres, Alain. 1998. *The Politics of Large Numbers: A History of Statistical Reasoning*. Cambridge, Mass: Harvard University Press.

Dewey, John. 1957. *Reconstruction in Philosophy*. Boston: Beacon Press.

---------. 2004. *Essays in experimental logic*. Mineola, N.Y.: Dover Publications. http://www.loc.gov/catdir/d.html.

Edwards, P. N., M. S. Mayernik, A. L. Batcheller, G. C. Bowker, and C. L. Borgman. 2011. ``Science friction: Data, metadata, and collaboration." *Social studies of science* 41:

667--690. http://sss.sagepub.com/content/41/5/667.short.

Fuller, Mathew, ed. 2007. *Software Studies: a Lexicon*. Cambridge, MA: MIT Press.

Galloway, Alexander R., and Eugene Thacker. 2007. *The exploit□: a theory of networks*. Minneapolis: University of Minnesota Press.

Haraway, Donna J. 1997. *Modest_Witness@Second_Millennium. FemaleMan©_Meets OncoMouse™*. New York: Routledge.

Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.

Hayles, N. Katherine. 2005. *My mother was a computer□: digital subjects and literary texts*. Chicago: University of Chicago Press.

James, William. 1996. *Essays in radical empiricism*. Lincoln: University of Nebraska Press.

King, Gary. 2011. ``Ensuring the Data-Rich Future of the Social Sciences." *Science* 331 (feb): 719--721. doi:10.1126/science.1197872. http://www.sciencemag.org/content/331/6018/719.abstract.

Latour, Bruno, Pablo Jensen, Tomasso Venturini, S. Grauwin, and D. Boullier. 2012. ``The Whole is Always Smaller than its Parts. How Digital Navigation May Modify Social Theory." *British Journal of Sociology* 63: 590--615.

Manovich, Lev. 2009. ``Cultural analytics: Visualing cultural patterns in the era of more media." *Domus (923)*. http://softwarestudies.com/cultural_analytics/Manovich_DOMUS.doc.

Massumi, Brian. 2000. ``Too-blue: colour-patch for an expanded empiricism." *Cultural Studies* 14: 177--226.

Mol, Annemarie. 2003. *The Body Multiple: Ontology in Medical Practice*. Durham, N.C: Duke University Press.

Munster, Anna. 2013. *An Aesthesia of Networks: Conjunctive Experience in Art and Technology*. MIT Press. http://mitpress.mit.edu/books/aesthesia-networks/.

OccupyData. 2013. ``Occupy Data." http://occupy-data.org/.

Parisi, Luciana. 2013. *Contagious Architecture: Computation, Aesthetics and Space*. MIT Press.

Porter, T. M. 1996. *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton Univ Pr.

Savage, Mike. 2009. ``Contemporary Sociology and the Challenge of Descriptive Assemblage." *European Journal of Social Theory* 12 (feb): 155--174. doi:10.1177/1368431008099650. http://est.sagepub.com/cgi/content/abstract/12/1/155.

Stengers, Isabelle. 2008. ``Experimenting with Refrains: Subjectivity and the Challenge of Escaping Modern Dualism." *Subjectivity* 22 (may): 38--59. doi:10.1057/sub.2008.6. http://www.palgrave-journals.com/doifinder/10.1057/sub.2008.6.

Thrift, N. J. 2005. *Knowing capitalism*. London: SAGE Publications.

Thrun, Sebastian, Mike Montemerlo, Hendrik Dahlkamp, David Stavens, Andrei Aron, James Diebel, Philip Fong, John Gale, Morgan Halpenny, and Gabriel Hoffmann. 2006. ``Stanley: The robot that won the DARPA Grand Challenge." *Journal of field Robotics* 23: 661--692. http://onlinelibrary.wiley.com/doi/10.1002/rob.20147/abstract.

Wacquant, Loic. 2004. *Body and Soul. Notebooks of an apprentice boxer*. Oxford: Oxford University Press.

Whitehead, Alfred North. 1958. *Modes of thought; six lectures delivered in Wellesley College, Massachusetts, and two lectures in the University of Chicago*. New York,: Capricorn Books.

Wilson, Elizabeth A. 2010. *Affect and artificial intelligence*. University of Washington Press.