Proposal

Titles

Forms of data thought: entanglements of subjectivity and computation

Forms of data thought: learning from machines

Forms of data thought: what to do with machine learning?

Learning from data:

Reconstruction in data: forms of machine thought

Recursions and reconstructions:

What can you do with machine learning?

Envisaging data

Overview

This book explores and analyses the data practices and forms of knowledge associated with machine learning, an increasingly widely used way of programming computers to find patterns, associations, and correlations, and to make predictions. Through a series of key techniques and cases drawn from social network media, finance markets, contemporary sciences such as genomics and epidemiology, and electoral politics, it traces some of movements of techniques, data, decisions, desires and beliefs associated with machine learning. Key techniques are drawn from machine learning, from data visualization and database architectures. Importantly, *Data Forms of Thought* is an experiment in constructing recursive forms of textuality and writing that combine code, data and diagram, text, and number. This experiment draws on both recent scientific coding practices as well as aesthetic practices to demonstrate some different ways of thinking supported by code and

data. In both analysing and re-purposing techniques found at the intersection of contemporary sciences and network media, *Data Forms of Thought* is generally concerned to affirm and increase the overlaps and entanglements between science and political, economic and cultural processes of diverse kinds.

Key concerns for the book are:

- The recent prominence 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 techniques have complex genealogies, criss-crossing sciences, industries, military, commercial and governmental domains. Both the provenance and mobility of data practices such as learning algorithms, predictive modelling and data practices need critical attention. This book focuses on the role of machine learning (or data-mining, as it is called in some domains).

- Humanities and social science responses to data techniques should be methodologically inventive, and include appropriation and re-purposing of the techniques and practices. This is major undertaking for the book.

Approaches used in the book

The book draws on some straightforward `executable paper' formats developed in recent scientific publishing, in order to mingle code written in R, the statistical programming, data manipulation and visualization environment, with code written in Python and Javascript, two of the most popular programming languages in use today. Much of the empirical content of the book is either gathered, ordered, analysed and displayed using R, Python and Javascript. Not all of the code used in this process is printed (to avoid long boring printouts), although certain key portions of the code form part of the text, 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 `participant observation'). This has include working in factories, going to prison, spending time in isolated, far-flung or ostensible 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. This book treats coding, and in particular code for communicating with databases, for building predictive models and for data visualization as both ethnographic material to be analysed and itself an ethnographic practice forming form of a writing process.

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 *Data Forms of Thought* 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 at work in the manifold data practices of sciences or business analytics, and to begin to borrow, appropriate and repurpose 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.

A broader ethico-political concern underpins *Data Forms of Thought*. Much contemporary data practice is closely allied to the predictive ambitions of business, the military and states. It continues and intensifies the technoscientific `Regime of Computation' (Hayles 2005). It is no accident that autonomous mililtary vehicles, large-scale analysis of social

media for security purposes, or face recognition are iconic examples of machine learning practices 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 will be how to situate themselves in relation to data practices. This concern lies at the core of the ethnographic and experimental writing practices of *Data Forms of Thought,* and throughout, the writing seeks to respond to the long-standing call for what 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, Page 16]. 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 (see the `OccupyData' group in New York, N.Y. for one such example; many the citizen science projects have something of this flavour to them too). Some of these will be discussed in the course of the book.

The architecture of the book

The book is organised around two different axes.

On one axis, the `technique axis,' the chapters of the book catalogue, document and analyse some of the most visible or widely used machine learning techniques of working with data (Hastie, Tibshirani, and Friedman 2009). The techniques analysed on this axis -- linear regression models, decision trees, clustering algorithms, Markov Chain Monte Carlo simulation, neural networks and support vector machines -- 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.

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 enact engagements 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. Most of the diagrams, functions and code constructs arrayed along this access are drawn from scientific fields, of from commercial applications where data is made available through APIs (Application Programmer Interfaces). The reconstruction of data practices draws on the pragmatists philosopher John Dewey's notion of philosophy as an empirical reconstruction of experience (Dewey 1957; Dewey 2004). Dewey envisaged reconstruction as responding to 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.

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. Electronic versions of the book should have 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.

Existing academic literature and framing of the book

The existing literature on data is largely found in either science and technology studies (STS), and some parts of information science. Software studies and anthropological accounts of software are highly relevant. The broader theoretical background here includes recent reappraisal of pragmatism, feminist work on materialities, as well as strands of largely European contemporary philosophy relating to science, number (Badiou 2004; Badiou 2008), calculation and ontology .

In STS, work on calculation (Callon and Law 2005), data practice (Edwards et al. 2011), models, simulation, database and computation is a key resource.

- software studies
  - galloway (Galloway 2013)
  - *Speaking Code*
  - Manovich
  - (Chun 2011)
- platform studies
- sts
  - Bowker, Landecker, Kelty etc on reading texts
  - Adams, Murphie, Clarke on anticipation
  - Callon/Law on the power of calculation
    * On Qualcalculation, agency and otherness, 2005The power of a calculation depends on the number of entities that can be added to a list, to the number of relations between those entities, and the quality of the tools for classifying, manipulating, and ranking them. 720 (Latour 1990)
  - , etc on diagrams, etc

- history of statistics and numbers
  - Porter, Daston, Desrosieres, Mackenzie
- new materialisms
  - Massumi, thrift, Galloway
- speculative realism
- new media studies -
  - beer on data cultures;
  - (Munster 2011) on nerves of data;
  - (Manovich 2009a; Manovich 2009b) on cultural analytics
- sociology
  - governmentality-bio-surveillance Thrift
  - (Savage 2009)
  - Abbott, On the concept of turning point
    * If most things that could happen don't happen, then we are far better off trying first to find local patterns in data and only then looking for regularities among those patterns. Indeed, it is for this reason that cluster analysis and scaling, not regression, dominate big-money social science --- market research --- where the aim is to find, understand, and exploit strong local patterns. For these are methods that seek clumps and partitions of data and make not attempt to write general transformations. 241
    * Thus the real alternatives to Goldthorpe's variable approaches are not case-based approaches, but what I shall call, for want of a better term, ``pattern-based approaches." Pattern-based approaches begin by establishing local patterns among variables before setting out to generalize. …. This procedure will be most when much or most of the data clusters arund a few

types and a consideerable portions of the data space is more or less empty. 241

* The world is Markovian. But the past is encoded into the present in patterns of connections that we call structure. The production of the next moment of social life happens from the basis provided by that structure. And the arrangements of structure always leave openings for actions, which if they fit the situation can change the longest-enduring structures quite quickly. 257

  - interface with sciences -- borrowing of methods for viz and exploration;

- interface with network media -- acquisition of panoply of data (via APIs, etc), but also re-purposing of methods

- sciences and network media also in transformation

  - hot spots include pattern finding via machine learning; software libraries for data acquisition, exploration, and visualization;

Chapter outline

1. Introduction

The introduction will provide a couple of motivating cases and events of forms of data thought drawn from a variety of fields --- social media, finance, security, robotics, and biomedicine. It will highlight these cases as symptoms of the pervasive transformation in knowledge, control, and decision associated with data flows, and at the same time, suggest how these transformations also might elicit changes in how humanities and social science researchers understand their own work. At this point, the notion of `data forms of thought' will also be characterised, drawing on a range of theoretical work drawn from pragmatist philosophers such as C.S. Peirce (abduction and diagrams), William James (experience), and John Dewey (`reconstruction') (Debaise 2007), and from recent social and cultural

theory (Francois LaRuelle on generic science; Isabelle Stengers on experiment; Gilles Deleuze & Felix Guattari on functions; Celia Lury on topology).

In order to contextualise forms of data thought, the introduction will also sketch some points of departure drawn from software studies, platform and infrastructural studies, work on calculation, number, image and diagram, as well as the general background of science and technology studies, and accounts of subjectivity, experience and materiality cross-cutting all of the above. These sub-disciplines very provide the intellectual scaffolding and departure points for much of the book. They include anthropological work on number and calculation (Maurer, Verran) The recent work on subjectivity and experience such as [@berlant, 2007] or [@murphie, 2010] are particular important in thinking through what we hope and believe about forms of data thought.

The introduction will also provide a preliminary overview of the techniques of data thought discussed in the book -- clustering, linear modelling, Bayesian inference, etc -- but very much with a view to setting out the key conceptual theses of the book concerning data as a material-semiotic entity: dimensioning, diagrams and mapping, generating and discriminating, convolution and multiples, optimality and predictivity.

2. Associating data

; writing code for data >description assemblage/Multivalent code - free association; learning to code; movement of methods; elementary operations

description; recursive embedding; autoethnography in code;

This chapter is partly a methodological discussion, in the form of 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.

It secondly develops an analysis of the transverse, cross-disciplinary moment of data methods 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.

Finally, this chapter is somewhat autoethnographic too, in that it reports on the author's own trajectory through coding competitions, `machine learning' courses, as well as more broadly on forms of empiricism associated with data. The forms of data empiricism used in producing this book are also the objects of its analysis. This recursivity is not exceptional. Versions of it can be found in many of the case studies discussed in later chapter. [TBA - say how this differs from the reflexive; rather evocative -- Bollas;]

3. The curve of curves

-- recursion, movement, evocative objects, partial observers, visualisation, etc; functions and states of things; linear regression

The visual form of the graph, plot or diagram lies at the centre of vision in contemporary data and knowledge economies. We might speak of a `curve of curves' in reference to the many and proliferating forms of curve, line and graph seen in data economies. The topography of curves, lines, points and diagrammatic elements convey views of data, and they are indispensable to many of the classification, decision and prediction techniques. They are themselves commonly used to convey expectations and predictions about the changes in the data practice, especially in the form of the curves showing growth of data.

This chapter treats the curve of curves as a process of proliferation that can itself be analysed recursively in terms of *functions*, the underlying generating mechanisms of curves. The chapter both introduces the key concept of the function as a mathematical construct, and 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 perhaps the most widely used data modelling technique, linear regression and its classificatory alternative, logistic regression. It will discuss these important functions from the perspective of the forms of relationality, referenciality and indexicality associated with them.

At the same time, it treats the production of curves through software packages and libraries, and through visualization techniques, as a practice worth investigating as a signifying social practice. 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.

4. Optimism about optimisation

: regularisation - dimensional reduction, dimensional explosion -- infinite dimensional spaces; recommender engine - svd as well; ebay; hunch.com

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 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, TBA], 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.

5. Self-reconstructions and algorithmic competition

- selfhood in Kaggle and Google compute - random forest; aggression, competition, and optimisation in the algorithmic

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.

6. Belief and desires in data

- probability and Bayesian inference - belief and desire in data - belief chance, Bayes, internal proliferation of numbers; event-belief oscillation

The topic in this chapter is 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.

7. Contagious numbers

- functions & supply chains; APIs; multiplication & convolution; states and functions of the lived;

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.

8. Genomic topologies

- doubling times, the auratic power of the instrument, and metacommunity, the topological turn

The final chapter of the book concerns data-generating instruments and data archives 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 -- 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 circuits. 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. The flatness of sequence data has been heavily leveraged by this positioning. This chapter experiments conceptually with the increasing topological character of machine learning (and particularly, the growth of `topological data analysis' [@carlsson_topology_2009; @singh_topological_2007] as well as the topological turn in culture (Lury, Parisi, and Terranova 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.

9. Conclusion

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.

Market

The market for the book is quite diverse, since data practices are of wide interest. 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. 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, 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 be relatively lightly-argued in

relation to social theory in order to facilitate access for them.

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 I'd like to deliver the whole manuscript mid-2014.

EXTRA STUFF

Chapter outline old

Introduction to platform pragmatism

Pragmatism used here in the sense that recent pragmatism has come to use it: not just what works, but what how a general account of experience can be derived from the irreducibility of practice to the forms/ideas/concepts that usually organise it. Platform pragmatism: points to the kinds of experience that relate to platforms as lifted-out places on which things work: living in data; included and perhaps belonging in data; Platform here has a relation to plane: not all platforms are planar, but planarity is a significant feature of the platforms I address Not just a theoretical pragmatism, but a pragmatism that comes from actuality taken against itself: how to counter-effectuate in practice; How to modify the practices of thinking so that data can be thought; Platforms used here to refer to two planes of reference -- the recording surfaces; the sampling surfaces, which themselves are involved in construction of functions meant to actualize variations on the recording surface; Implications for human and social sciences Galloway on the politics of theory

1669: Belief in data and the invention of analytics: 1660

Belief in data and the invention of analytics: 1660 Supplies of random numbers, shaped by functions for almost the first time; But this problem continues today … Constitution of data in relation to notions of evidence, probability, error, prediction requires supplies of randomness; Plying numbers vs rolling numbers; Could introduction functions here -- fits with differentials and Leibniz

Dataset:

Curve of curves: 1828

Role of visual forms here -- density-shapes lashing out into the visual; Follows on from functions in D&G John Tukey -- exploratory data analysis NYT Graphics team; `Data is beautiful' vs `finding the signal in the data' Tarde's stuff on imitation useful here Logistic curves as key example here: both the role of curves, the role of linear models; Link between lines and curves described in terms of functions Tension between graphics and models continues (Fisher vs Tukey?) Matrices and hypervolumes Dataset: iris

1899 -- 1968: Aggregate data: more parts than elements

Has all the stuff on relationality, sets, etc;

Expands to include different scales apart from the meso-level databases: from spreadsheets to data centres; The excess parts over elements as another way in which full knowledge is inhibited constantly;

Plane of reference includes enterprises, states, etc; anywhere where information retrieval counts

Dataset: twitter, mongodb. couchdb

References: Manning & Co.

1971: Clustering and the curse of dimensionality

Machine learning chapter Many different ways to find the signal; as dimensions increase, more likely that points will lie near the boundaries of any sample. Also the many problems of bias and variance as the dimensions grow. If plane of reference is a hyperplane, then many such problems will arise This chapter goes through k-nearest neighbour, k-means, hierarchical clustering, decision trees, random forests, neural networks, etc Pattern recognition here and here it has ramified

Dataset:

Abyss of methods

What happens as methods become mobile: how are they recombined? This is a place where plane of reference is being folded onto itself Machine learning vs data geeks, etc; How to do deal with proliferation of methods in recombination?

Dataset: iris -- trace iris across different settings

Elusive variation

Variation becomes the norm; but this variation is well structured? Genomes + gwas The problem of well-structured data -- gives an explanation of certain biological forms attract so much attention. What happens to the messy ones? So, justification for talking about this is to try and capture why certain kinds of data matter more than others. Dataset:

Epidemiology and its problems with nubers: what cannot be observed vs what can be observed

Returns to population, but now with the idea of many populations interpenetrating Distributions of distributions Against the idea of full knowledge, etc

Datasets: Birrell, 2011; netflix

Conclusion

The overall argument:

1. what data can be for - analyse, control, find patterns;
2. creativity/curiosity to make algorithms to find things in data;
3. evaluating the results of the model ethically;
4. by accompanying models & code, transform them …

Things to fit in

- what am I describing actually?
- how does data empiricism differ from other empiricisms?
  - Lury & Adkins 2009; Gane 2009; Manovich, Trending 2011; Harrison White 2004; Gary King 2012; Clough 2009; Savage 2009; material-semiotic; Chatelet 2006 on indexation;
- the API
- the critiques of data that are appearing

- - The actuality of data needs to be counter-effectuated in methods - The shift from search to social media is also a shift to a data culture (beer) --- living in data --- this - relates to the Cambridge paper Data turn may be part of the ongoing destruction of practices, including of scientific practices (Stengers!- When data is live and when data is dead: how to find what is still living and what has been thought - through? - Forms of data thought plays on two senses of thought: thought as past tense of thinking; thought as - substantive form of thinking; - ``Data thought'' == something similar to what Munster calls `nerves of data' - Perhaps more important to link to practice than conceptuality, to those forms of

thinking that are not - fixated on conceptualisation, idealisation, etc. - Possible to do reconstructions of knowledge using data and methods because these are so widely available. - Possible in doing these reconstructions to highlight both the radical contingencies and the embodied - materialities of these - This means that reconstruction can also be counter-effectuation, since it can take place using the very - same methods, materials, practices, and techniques that are constructing the plane of reference; but note that countereffectuation is not a beautiful Stoic or Spinozist one (again, Stengers is useful on thisS) - Has populations, evolution, life-death, reproduction, metabolism, decay, mutation, hybridity, semiosis, symbiosis, transduction, variability, heritability, --- all things that involve non-linear, multiply super-- imposed, biopolitically invested, promissory and speculative, rates of realization, etc. - This book is about living in large numbers, and what that means. How small numbers are being reconfigured - through large numbers. - Could use the stochia, and stochastic understandings of events found in stoic and epicurean thought to - think about the ethics of numbers. (cf above) - Could check Deleuze on this, and well as Foucault - Need to work out what this means for me and then connect it to some other theories - Idea of *separating hyperplane* as a way of making sense of many attempts to rectangularize and regularize data. Hyperplane can be understood partly in terms of Deleuze and Guattari's concept of the plane of reference on which scientific functions map matters of fact. It can also be understood in relation to the vectors and movements associated with network exploits and the vectoral movements (McKenzie; Galloway & - Thacker) - Idea of *reduction of dimensionality* -- actually track some of the many dimensional shifts that go on as data moves; it moves in and out of dimensions rapidly, and in some cases, the dimensions proliferate wildly; in other cases they are heavily restricted. - Hold together quite extreme things --- like scientists at stanford & ebi, with very commercial or institutional settings.

The ways in - which I have learned to use R are manifold. They include working through textbooks in various fields, attending training courses, tracking some of the many R-related blogs, and looking at print and online materials produced using R.

- Haraway - situated knowledges (Haraway 1999)
- Haraway - modest witness
  - Page 16 - What we need i- s 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.
  - Could argue that this is what I am trying to do with R, and also with PureData; Look at several kinds of `rays of technoscience' - cf Gabriel Tarde on rayons d'-imitation (Laws of Imitation). Tarde doesn't really want to diffract them, but only follow their diffractions. Maybe the notion of ray could be replaced by signals as a more contemporary form of propagation. Is a signal less idealised than a ray, with its quasi-geometrical connotations?
- Could turn *signal* into a guiding concept: how to deal with low signal to noise ratios?
- The givenness of data needs to be theorised more in order to get away from thinking that its an object in the world. Instead, I should draw on the James stuff to say more about what it means to work with data, especially in his account of knowing as what the end of experience says about the beginning; Or put more simply, to around the cognitivist framing that governs most understandings of dat- a.
- how to do the fevered projection + basic basics of everyday life
- Working through examples, trying out code that addresses both infrastructures and abstractions, and showing how they slide into each other
- Key precept only write about what you - can write about: only write about what you can code … If I can't code against it, then I don't write about it.
- Code here is then a mode of participation in t- he occurrence ….

Whitehead

References

Badiou, Alain. 2004. *Theoretical writings*. Ed. Ray Brassier and Alberto Toscano. London□; New York: Continuum.

---------. 2008. *Number and Numbers*. Cambridge: Polity Press.

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.

Debaise, Didier. 2007. *Vie et Expérimentation Peirce, James, Dewey*. Paris: Libraire Philosophique J.Vrin.

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.

Galloway, Alexander R. 2013. ``The Poverty of Philosophy: Realism and Post-Fordism." *Critical Inquiry* 39 (jan): 347--366. doi:10.1086/668529. http://www.jstor.org/stable/10.1086/668529.

Haraway, Donna. 1999. ``Situated Knowledges. The Science Question in Feminism and the Privilege of Partial Perspective." In *The science studies reader*, ed. Mario Biagioli, 172--188. 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.

Latour, Bruno. 1990. ``Drawing things together." In *Representation in Scientific Practice*, ed. Michael Lynch and Steve Woolgar, 20--68. Cambridge, MA London: MIT Press.

Lury, Celia, Luciana Parisi, and Tiziana Terranova. 2012. ``Introduction: The Becoming Topological of Culture." *Theory, Culture & Society* 29: 3--35. doi:10.1177/0263276412454552. http://tcs.sagepub.com/content/29/4-5/3.

Manovich, Lev. 2009a. ``Software takes command. 2008." *Published online by the author at http://lab. softwarestudies. com/2008/11/softbook. html*.

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

Munster, Anna. 2011. ``Nerves of data: the neurological turn in/against networked media□: Computational Culture" 1. http://computationalculture.net/article/nerves-of-data.

Savage, M. 2009. ``Contemporary Sociology and the Challenge of Descriptive Assemblage." *European Journal of Social Theory* 12: 155.

Vapnik, Vladimir. 1999. *The Nature of Statistical Learning Theory*. Springer.