In the data: interdisciplinary modes of machine learning

Adrian Mackenzie

Sociology Department, Lancaster University

a.mackenzie@lancaster.ac.uk

Abstract

This paper explores ways of thinking about digital data that lie somewhere between blithe faith and critical dismissal. It focuses on the machine learning, an increasingly widespread bundle of techniques and approaches that lies at the centre of contemporary data processing. Machine learning is used to program computers to find patterns, associations, and correlations, to classify events and make predictions on a large scale. As a set of techniques for classifying and predicting, machine learning lies close to centre of calculation in social network media, finance markets, robotics, and contemporary sciences such as genomics and epidemiology. This paper will discuss who is doing machine learning, who could do machine learning, and how they might do it differently.

The problem

- if you had all the data, what then?
  - this is the horizon of the contemporary lifeworld. Someone, perhaps not us, not yet, has all the data. Probably they are a business like Walmart, amazon, google or Facebook. They might be a State, like the United States or the United Kingdom. Perhaps we will have the data. There is, after all, a lot of data already available, and more all the time. It may be that civil society, non-governmental groups like Wikileaks, or various forms of massive data leak will render valuable data available. Or maybe just the increasingly powerful techniques of record linkage will allow comprehensive data streams to appear.
  - if we had it, what would we do? pattern finding vs models?
    * the problem is that at the moment we have no good ways of thinking about what you might do, or ways of articulating optimism about data with understandings of what it might cost to have the data. There are many different facets to this problem. Here I focus on a more academic side of the problem.

* Andrew Abbott, whose work has received much more attention in the wake of (M. Savage 2009), argued over that a decade ago – I don't know whether he is still arguing it – that we should adopt pattern-based approaches to working with data:

    > 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 (Abbott 2001, 241)

* There are two problems with Abbott's recommendation. The first is an empirical one: 'big money social science' such as business analytics intimately and increases uses linear regression models. In the meantime, however, the practices of linear modelling have been heavily renovated, and proliferated in a number of different directions. There is really no dispute on this point, and hence the whole opposition that Abbott sets up between pattern-based approaches and linear model or regression-based approaches to data is not empirically well-grounded.
* The second observation flows from this. The largely North American social sciences that Abbott criticises for their linear modelling of social data may or may not continue in their practices (augmented by social network analysis and heavy dosed of Bayesian statistics). The broader point here is that the simple opposition between pattern or cluster-based approaches and linear models of reality does not sufficiently to my mind orient us to the shifts in data practice that have been happening in the last decade or so. We need other analytic filters.

Learning from data

if you had all the data, what then?

I want to present several vignettes that illustrate what people do when they have all the data. They nearly always do what is variously called data mining, knowledge discovery or currently, *machine learning*. These terms reflect different scientific, govermental and commercial interests in data. Formally, they are linked by the idea that they are *finding an approximation to the function that generated the data*. Nearly all of the models and techniques used, whether it is clustering, linear modelling, neural networks, support vector machines, random

forests, topic modelling, etc., etc., can be seen as attempts to find the function that generated the data.

infrastructural re-dimensioning

- cancer genomes + google compute
- begin with someone/something that does have all the data – Google and certain genomic scientists work together on cancer genomes;
- Let me illustrate what I mean by infrastructural re-dimensioning:

  Google Compute Engine + Institute of Systems Biology, June 2012

  The world's 3rd largest supercomputer *learns associations* between genomic features' (Anthony 2012)

  The Google Compute Engine, a globally-distributed ensemble of computers, was briefly turned over to exploration of cancer genomics during 2012, and publicly demonstrated during the annual Google I/O conference. Midway through the demonstration, visualized as a circular genome, the speaker, Urs Hölzle, Senior Vice President of Infrastructure at Google 'then went even further and scaled the application to run on 600,000 cores across Google's global data centers' (GoogleInc. 2012). The world's '3rd largest supercomputer', as it was called by TechCrunch, a prominent technology blog, 'learns associations between genomic features' (Anthony 2012).

- the scientists are from Institute for Systems Biology (ISB), University of Washington. Its a fairly well-established and long-standing icon of big data science, with some Nobel Prizes, etc, associated with it. [SHOW Image of building]
- Urs Hölzle, the Vice-President of Infrastructure at Google shows the fruits of their collaboration to an audience of developers at Google I/O in San Francisco, 2012. His title is significant – VP of *Infrastructure*. [SHOW the video]
- On the one hand, the scientists at ISB have the data. They are somewhat like us – they are academics. They are different to us in that they have the best data possible on cancer genomes. Their data reaches down into clinics, it comes from the best scientific instruments money can buy, and they benefit from PhDs, programmers, engineers, disk drives and an assortment of other services at their beck and call.

- Yet, they need or want google. Or maybe Google wants them for something. What the Google I/O conference shows is the availability of flexible instrastructure that can be quickly, rapidly made available on demand. It can shift orders of magnitude in size in seconds. Not many infrastructures can do that. But even if it can, why are such infrastructures needed? It is as if Oxford St in London could suddenly switch between a pedestrian mall and an 8 lane motorway. What form of life or sociality would require that mutability?

- in this case, what the scientists want to do is to sift through massive numbers of associations between different parts of a genome to identify patterns and processes associated with mutagenesis. The algorithm they use is called RF-ACE, and an implementation in the R statistical programming language is freely available. While it is state-of-the-art, this algorithm is a typical machine learning algorithm. It is an example of unsupervised learning algorithm that tries to *find structure* in a data-set without any prior knowledge about the data.

data is wide, dirty and mixed: 3 million features

- In the Google Compute case, it is hard to see what machine learning is doing. We hear it is discovering associations, but associations between what? For present purposes, the issue here is the shape of the date.

- The shape of datasets is widely discussed in machine learning, and shape perhaps more than the size (as in 'big') matters to who algorithms go about finding structures or patterns. It is shape too that occasions much of the infrastructural re-dimensioning we have just seen displayed.

- The shape of contemporary datasets is sometimes described as 'wide, dirty and mixed'. Each of these terms carries interesting connotations, and a figurative analysis of data talk would be quite useful. I'm not going to offer that here, but just point to some examples of what they mean.

- Statistics textbooks and statistical practice are full of datasets like this one: classic dataset: 'iris' R.A. Fisher (1936)

```
## [1] 5
```

```
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrrl}
##   \hline
##  & Sepal.Length & Sepal.Width & Petal.Length & Petal.Width & Species \\
##   \hline
## 1 & 5.10 & 3.50 & 1.40 & 0.20 & setosa \\
```

```
##   2 & 4.90 & 3.00 & 1.40 & 0.20 & setosa \\
##   3 & 4.70 & 3.20 & 1.30 & 0.20 & setosa \\
##   4 & 4.60 & 3.10 & 1.50 & 0.20 & setosa \\
##   5 & 5.00 & 3.60 & 1.40 & 0.20 & setosa \\
##    \hline
## \end{tabular}
## \end{table}
```

- Fisher's widely cited 'iris' data is often used to demonstrate various statistical techniques.
- In machine learning textbooks such as (Hastie, Tibshirani, and Friedman 2009), iris is mentioned, but datasets are commonly much wider. That is, they have many more columns. For instance, in the late 1990s, when biomedical data from genechips or microarrays started to become widely available, machine learning techniques were applied to their analysis. A typical 'wide dataset' would be Golub's c.1999 'blood pressure' dataset

```
## columns in Golub dataset: 102
```

```
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrrrrl}
##   \hline
##  & \begin{sideways} RS2495368 \end{sideways} & \begin{sideways} RS2292857 \end{sideways}
##   \hline
## 1 &   0 &   0 &   0 &   0 &   0 & 171.19 & normal \\
##   2 &   0 &   0 &   0 &  &   0 & 179.25 & normal \\
##   3 &   0 &   0 &   0 &   1 &   0 & 188.23 & normal \\
##   4 &   0 &   0 &   0 &   0 &   0 & 175.44 & high \\
##   5 &   0 &   0 &   0 &   0 &   0 & 177.98 & normal \\
##    \hline
## \end{tabular}
## \end{table}
```

- the Golub blood pressure data from 1999 has over 100 columns rather than 5. This is not especially wide, but radically changes the structures that might be found in the data.

- as we can see, the number of different models that you might build to look at patterns between blood pressure, height and this genetic data very quickly becomes huge when there are numerous variables or 'features' in the data.

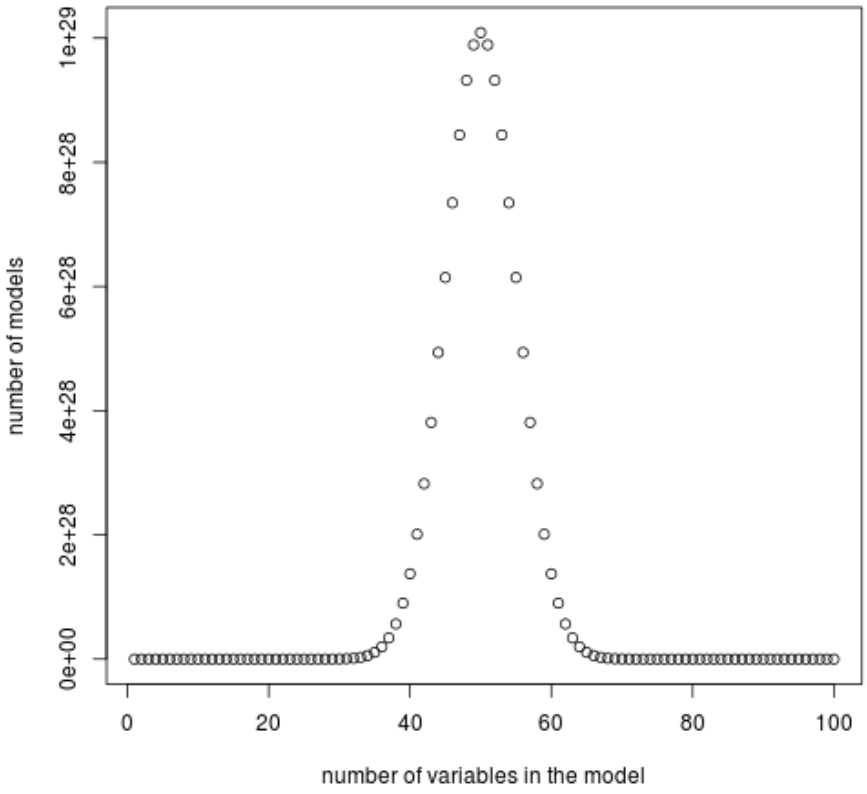

Figure 1: plot of chunk name

- but this kind of measurement data is quickly dwarfed by text or image-derived data. Even relatively tightly defined data forms such as URLS. Machine learning researchers spend quite a lot of time trying to build predictive models for mundane forms of data such as URLs. They can involve millions of variables, that could be interacting in countless combinations.
- Suspicious URLS dataset (Ma et al. 2009)

```
## \begin{table}[ht]
## \centering
## \begin{tabular}{lr}
##   \hline
## feature & value \\
##   \hline
## suspicious & -1.00 \\
##   4 & 0.08 \\
##   5 & 0.12 \\
##   6 & 0.12 \\
##   11 & 0.43 \\
##   155211 & 1.00 \\
##   155212 & 1.00 \\
##   155213 & 1.00 \\
##   945789 & 1.00 \\
##   1988571 & 1.00 \\
##   2139257 & 1.00 \\
##   2987739 & 1.00 \\
##    \hline
## \end{tabular}
## \end{table}
```

- The point here is that finding structure in data means working out which bits of data are related to which other bit. Typically a row of data in a dataset relates to one thing. The thing might be a person, an event, a sample, a place, but machine learning is rather indifferent to that. What matters is finding, amidst the many possible combinations of features or variables those combinations that carry some weight, that pattern the other data, or that stand out.
- While the data we have seen is wide, it is not necessarily very mixed or dirty. Mixed means that the types of data are varied. A mixed dataset includes numbers, categories, text, and so forth. Dirty data, as you can imagine, describes data lacks consistency, that might have holes in it or is somehow noisy.
- text data and shaping the world
  - James Lin, Twitter

  - e-discovery
  - leximancer
- signals – image and sound
  - driverless car - Thrum - winner of the Darpa challenge
  - kittydar: : kittydar vs. google cat
  - Eulerian video

machine learning as interdisciplinary method A couple of observations about these vignettes:

- machine learning move widely across disciplines. In some ways that is not suprising because statistical technqiues had already moved widely across natural sciences, social sciences and into various practices. There may be, however, a kind of hypermobility associated with machine learning that is partly due to infrastructural -re-dimensioning, and the associated shifting dimensionality of data.
- what is this hypermobility? I said above that machine learning tries to find the function that generated the data. Another way of viewing this is to see machine learning as a way of navigating high-dimensional spaces, spaces that are difficult for us to perceive, observe or represent. Many machine learning techniques try to address the fluxing dimensionality of patterns. We can see patterns easily when they are on surfaces, but hyperplanes or hypersurfaces can have patterns that we simply can't see, although we might be able to have some feeling of them. These spaces, generated by functions, are explored in machine learning, usually by finding lines or planes that cut through them, linking somethings together and separating others [1].
- Thirdly, if fluxing data dimensionality becomes navigable through machine learning, it becomes important to ask the usual kinds of science studies questions such as navigable or explorable at what cost.

The fluxing dimensionality of people

- So while I haven't yet talked about machine learning actually locates structure in data, I want to now turn to talking about people, companies and

[1] This is the key intuition developed by Luciana Parisi in her recent work (Parisi 2013). Her promotion of algorithms as objects of abstract experience is consonant with my argument here. I differ from her in the importance I attribute to algorithmic information theory and its notion of randomness. She sees algorithmic processes as ruptured by non-computable bursts of randomness generated by the axiomatic undecibility (cf. [@Mackenzie, 1997] on undecibility and non-computable numbers). Importantly, I differ in the cases I draw on and how I link practice and theory.

scientists doing machine learning. Who today does machine learning and at what cost?

- Academically, the epicentre of the field lies somewhere between computer science and statistics. For instance, one of the main textbooks in the field (Hastie, Tibshirani, and Friedman 2009), was written by three very well-known statisticians from Stanford University. On the other hand, the Stanford computer scientist Andrew Ng is one of the most well-known figures in the field. His Youtube lectures on machine learning CS229, a postgraduate course, show viewing figures peaking at 500,000.

- SHOW Andrew Ng, Stanford, CS229

- The tenor of these lectures is that understanding machine learning as a mathematically grounded process of finding the function that best approximates how the data was generated offers to do things that programmers and others can't do. Ng's refers often to the Silicon Valley companies he visits, and how often they are wasting their time trying to build systems that are statistically unreliable. He adjures his students to understand the mathematical and statistical underpinnings of machine learning so that when they graduate they will be to respond to the challenges of constructing and predicting increasingly complex processes. From the perspective of Stanford or MIT, the quintessential machine learning expert is a PhD, highly-maths literate and also able to program, who goes on to work in Silicon Valley, Wall Street or for some US government agency.

- This does not exhaust the constituency of machine learning practitioners. Beginning around 2006, machine learning methods started appearing more broadly in software cultures. For instance, the book *Programming Collective Intelligence* appeared in 2007 and quickly became popular as a way of thinking about how to reconfigure website and other network media platforms to deal with the greatly increased range of interactions. A series of software packages, instructional media, and events have ensued that promote machine learning technique as the way to manage, contain and optimise the flows of data not only in digital media, but in the sciences.

- More recently, the recourse to machine learning has become more explicit. Recent books such as *Machine Learning for Hackers*, written by two social science U.S. PhD students, are symptomatic. Many of the datascience organisations as well as hackers individually show increasing interest in machine learning as a way to program software to react in ways that cannot be anticipated in advance by programmers themselves.

- Yet the people who do machine learning are not the same as the typical programmer or hacker. Indeed machine learning seems to coincide with a shift not only in how data is handled, but in the ways in which data-driven systems are constructed and configured.

- Several vignettes outline this shift:
  - Hilary Mason at Bacon
    * [SHOW conference video]: Mason talks about the wonderful who lie someone between engineering, programming, mathematics and social science.
  - the second is the growth of data science courses that machine learning methods to the new coding classes. These courses are run at universities, industry conferences, hackathons and various online settings by statisticians as well as more conventional software industry trainers. For instance, the 'first' datascience course was run by by Rachel Schutt (formerly at Google) in late 2012 at Columbia University in New York. It attracted students from various departments and disciplines, and will appear as a book later this year. Not all of this course is machine learning, but substantial portions are.
    * SHOW Cath O'Neill & Rachel Schutt from Johnson Research Labs
  - Heather Arthur, the software developer I showed speaking earlier makes two claims that are relevant here. Heather Arthur on cat face detection [SHOW]
    * 'essentially machine learning algorithms are better programmers than you' (00:03:35)
    * 'what is cool is that this is all running on the client side. 'A few years ago this would not have been possible' (00:17:30);
    * Programmer's become better programmers via machine learning; at the same time, code in all settings becomes more powerful; 'all this is running on the client side' means that the machine learning algorithms (in the case of Kittydar, neural network algorithms) are much more widely distributed. They are not staying just in the server farms, the data centres such as Google Compute. Machine learning gets into 'the clients.'
  - We could also look at cases such as the Obama re-election data team

-

- the social scientists
  - Gary King
  - manovich – cultural analytics
  - Savage - descriptive assemblage

What to do

- Three observations then about the significance of machine learning:

- its ubiquity in science, business and government renders it a key contemporary control practice, that differs in powerful yet subtle ways from existing ways of knowing, predicting, anticipating or controlling
- its mobility as a cluster of methods has strongly performative effects – parts of the world are heavily re-configured through it
- the mobility of its methods has no clear boundaries. The spread of machine learning into social sciences and humanities has started. And much of our research already depends on it anyway (google search, etc). The question is how we move in relation to machine learning.

- Machine learning is a hard field to get into in some ways for social science and humanities researchers. There are lots of statistical, mathematical and infrastructural subleties to deal with. On the one hand, it is becoming enormously available and increasingly as Thrift would say, part of the contemporary time-space signature. Myriad mundane examples could be given. On the other hand, to get into it, to occupy the fluxing dimensionality of the data is technically difficult, and conceptually challenging. It takes very significant investments in time and attention (for instance, going through 27 hours of Youtube lectures with lots of equations written on blackboards is not trivial).

- I am suggesting that getting into it analytically and practically by whatever might be worthwhile, at least for some people for several reasons:

  1. It offers a way of accompanying the fluxing dimensionality of data. In *Modes of Thought*, Alfred North Whitehead writes:

     > Perhaps our knowledge is distorted unless we can comprehend its essential connection with happenings which involve spatial relationships of fifteen dimensions (Whitehead 1958, 78)

     In this passage, Whitehead's choice of 15 is arbitrary. I guess it just refers to a spatiality that it is hard for to imagine, even though we no doubt inhabit it from time. From the standpoint of machine learning, we often do move through high dimensional spaces. Many machine learning techniques seek to reduce the dimensionality of spatial relationships in data. These would include the many dimensional reduction strategies. But others seek to expand the dimensionality of data. And the feild as a whole tends to augment rather than diminish data dimensionality. Certain techniques artificially inject infinite dimensional spaces into the models in order to find hyperplanes that separate data.
  2. Like all sciences and technologies, machine learning must contain zones of slippage, inconsistency or friction where things can happen. While it might not be us as researchers who occupy or can identify

those zones most easily, in making sense of machine learning all the way down, and pointing to the structures, processes or relations at play, we help free up the possibilities of gaming the models. And actually, who else is going to do it? That is, machine learning provides a way to contest the asymmetrical distributions of agency admidst re-dimensioned infrastructures.

3. To some extent, as researchers we are encountering a version of the quandary I posed at the outset: what happens if you have all the data? The question is how we are going to comprehend 15D happenings, especially when a good number of those dimensions are occluded from us. A couple of scenarios occur to me here:
   - using machine learning to impute what kind of machine learning is going on in a given setting
   - using machine learning techniques as a way of thinking about differences, movement, shape and change
4.

- Jaron Lanier
- occupydata
- animation
- text/coding/reproducibility

References

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

Anthony, Sebastian. 2012. "Google Compute Engine: For $2 million/day, your company can run the third fastest supercomputer in the world \textbar ExtremeTech." http://www.extremetech.com/extreme/131962-google-compute-engine-for-2-milliondday-your-company-can-run-the-third-fastest-supercomputer-in-the-world.

GoogleInc. 2012. "Behind the Compute Engine demo at Google I/O 2012 Keynote - Google Compute Engine — Google Developers." https://developers.google.com/compute/io.

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

Ma, Justin, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. 2009. "Identifying suspicious URLs: an application of large-scale online learning." In

Proceedings of the 26th Annual International Conference on Machine Learning, 681–688.

Parisi, Luciana. 2013. *Contagious Architecture: Computation, Aesthetics and Space*. Cambridge ; Malden, MA: MIT Press.

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.

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.