Big Data, Curly Fries and the Individual

Olly Abraham

“‘The individual has always had to struggle to keep from being overwhelmed by the tribe. If you try it, you will be lonely often, and sometimes frightened. But no price is too high to pay for the privilege of owning yourself.’”
Rudyard Kipling

“In God, we trust. All others must bring data.”
W. Edwards Deming

1. Introduction

To each of its more than 100 million customers, American discount retailer Target assigns a unique ‘Guest ID’ number, which is linked to their credit card number, email address and a variety of other personal identifiers. It should come as no surprise to even the most naive user of the internet that Target are able to use this information to track the purchases of individual consumers and present them with advertising targeting their habits. What might seem slightly more surprising is that the same data allows Target to predict, with a startling degree of accuracy, which of their customers are pregnant, along with their due date. As a result of this exercise in data analysis, a teenage girl in Minneapolis was forced to reveal to her father that she was pregnant because Target had mailed her a series of coupons and advertisements for maternity clothing and nursery furniture, addressed to her by name.

Concurrently, by identifying users of Microsoft’s Bing search engine whose activity unmistakably indicated that they had been diagnosed with pancreatic cancer and working algorithmically through their previous searches, researchers from Microsoft and Columbia University found that they were able to offer early diagnoses on between 5% and 15% of cases, returning almost no false positives. Pancreatic cancer is incredibly difficult to detect in its early stages, and as a result 94% of patients will die within five years of being diagnosed, while 74% will die within the first year. Early diagnosis, however, means that a patient is almost twice as likely to be alive five years down the track.

---

2 As far as statistical incursions into the personal lives of one’s customers go, Target’s efforts were rather a simple affair. Target’s statisticians identified twenty-five products that, when purchased at certain intervals, could be combined to develop a ‘pregnancy prediction score’ – these products included vitamin supplements in the first 20 weeks, scent-free soaps and large bags of cotton balls later in the pregnancy, and so on. According to statistician Andrew Pole, once you know that this kind of behaviour is indicative of pregnancy at certain stages, predicting a due date is a relatively straightforward matter. See Charles Duhigg, "How Companies Learn Your Secrets," *The New York Times*, February 16th, 2012.
3 Ibid.
When ethical questions regarding such uses of ‘big data’ are raised, they are often parsed in terms broadly relating to privacy: the requirement of consent for the collection of data, the right of individuals to withdraw that consent, the security of private data once collected, and so on. So it is that a recent volume on the ethics of big data in a biomedical setting suggests that the trend is to see “the demands of large-scale genomic research, which depends upon ready access to vast amounts of genetic information, as a threat to genetic privacy,” while the UK government has made much of a proposed law that will “give people more control over their data” by requiring higher levels of consent and broadening the definition of ‘personal data’ to include “IP addresses, internet cookies and DNA”.

We might, however, suggest a different emphasis. Rather than focussing on issues surrounding control, consent, privacy, storage and so on, we might focus on questions of application and use. Advanced techniques in data analysis and machine learning allow companies to make a variety of generalisations about our lives on the basis of information that we either voluntarily share (for example, on Facebook) or that we can hardly avoid sharing in the course of our interactions with various entities (in purchasing products and using search engines). The power of ‘big data’ lies precisely in what it allows us to do with huge quantities of this otherwise innocuous-looking data. In short: unless Kipling’s suggestion that “no price is too high for the privilege of owning yourself” inspires you enough to refrain from using social media and search engines while opting to make all purchases in cash, it is not the data that is worrying but the inferences that it allows.

The more delicate question is, then: what it is that companies, governments and institutions ought to be able to do with these generalisations? In some way, our various interactions with the institutions that colour our lives reflect a balance between our interests in achieving some measure of consistent treatment and taking into account the individual circumstances of people. We are happy for banks to make decisions on mortgage applications on the basis of our income, previous credit history, and so on, but might perhaps find it inappropriate if the decision on a married couple’s application was influenced by the perceived strength of their relationship. Yet advances in big data techniques mean that these more intimate pictures of our lives are not as insulated as they once were from such decision-making processes. There are algorithms, for instance, that can determine a person’s sexual orientation to a surprising degree of accuracy on the basis merely of their Facebook friends list or the TV shows they watch.

This essay will, then, examine some of the ways that the big data revolution has tipped the balance between individuals as general and as unique in two ways. First in the way that institutions may develop highly predictive profiles of individuals on the basis of seemingly irrelevant data, thus challenging notions of ‘relevant’ information that otherwise seems to mark various processes as ‘fair’. Second, in the depth and intimacy of information that can be inferred from such shallow stores of data. In this sense, pieces of information previously

---

5 Bryce Goodman, ”What’s Wrong with the Right to Genetic Privacy: Beyond Exceptionalism, Parochialism and Adventitious Ethics,” in The Ethics of Biomedical Big Data, ed. Brent Daniel Mittelstadt and Luciano Floridi (New York: Springer, 2016), p. 140.
8 Jeffrey Zaslow, ”If Tivo Thinks You Are Gay, Here’s How to Set It Straight,” November 26, 2002.
regarded as private may occupy prominent roles in statistical models for various decision-making processes. Before we can begin to answer questions of responsibility for the ethical production, curation and use of big data in any detail, we first need to take stock of the ethical challenges with which big data presents us. In concluding, I will suggest that what is important in this task is not necessarily what various institutions may come to know about us as a result of advanced algorithmic techniques, but rather the role that such data-driven inferences play in the various decision-making processes that shape our daily lives.

2. The Human Profile

During the 1960s, many credit bureaus in the United States found themselves searching for more sophisticated methods of assessing the credit worthiness of a would-be debtor. One of the results of this search was the collection of rudimentary ‘personality profiles’ of loan applicants, often augmented by amateur surveillance. As well as the standard financial data, these profiles tracked attributes such as tidiness, gardening habits, poorly kept yards and “effeminate gestures.”

In an attempt to curb such activity, the US Congress passed the Fair Credit Reporting Act in 1970, which required that bureaus “make their dossiers both accurate and relevant.”

While the landscape of credit profiling in the United States remains decidedly opaque, – Frank Pasquale refers to “the Heisenberg-meets-Kafka world of credit scoring” – there are at the very least few traces of the kind of psychological guesswork that was once prominent.

In Australia, the variety of information that may be included on an individual’s credit score is tightly regulated and limited to an individual’s relevant financial history. Implicit in some of our judgements regarding such situations might be the following notion: in order to reliably and accurately predict the likelihood that an individual will successfully meet loan repayments, we must only consider relevant kinds of information. Hence an individual’s income, assets, employment history and so on are fair game, but the tidiness of their yard and effeminateness of their gestures are not.

Big data, however, presents a considerable challenge to this intuitive distinction between relevant and irrelevant information. Researchers from Columbia University and the University of Delaware used text mining and machine learning tools to analyse over 18,000 loan requests, and found that individuals whose applications included religious vocabulary or the phrase

---


12 Although there are, of course, still deep problems. Pam Dixon and Robert Gellman cover in some detail how this corporate practice of boiling consumers down into their ‘scorable’ attributes takes place in America with almost no government oversight or accountability. See Pam Dixon and Robert Geelman, *The Scoring of America* (San Diego: World Privacy Forum, 2014).

‘thank you’ were especially likely to default on their loan.\textsuperscript{14} In fact, these researchers found that they could isolate certain words or phrases that were better predictors of loan default than the individual’s credit score.\textsuperscript{15} Other research has shown that there are strong correlations between particular kinds of Facebook activity and performance on a variety of personality and intelligence tests.\textsuperscript{16} Some of the strongest predictors of high intelligence included whether someone liked ‘Thunderstorms’ or ‘Curly Fries’, while liking the page of Nashville country music group ‘Lady Antebellum’ was a strong predictor of poor performance in IQ tests.\textsuperscript{17} The resulting model was capable of predicting substance abuse to an accuracy rate of between 65\% and 73\%.\textsuperscript{18}

That is to say that one upshot of advancements in machine learning and data analysis has been the ability to generate highly predictive statistical profiles of people based on information that seems at best tangentially related to the activity in question. Buoyed by their results, the researchers behind the analysis of language in mortgage applications suggest that similar methods might be employed in other spheres – for instance, “universities and business schools might be able to predict students’ success based on the text in the application (beyond human manual reading the essays).”\textsuperscript{19} In other words, we are rapidly approaching a point in time where algorithmic models, powered by large amounts of aggregated data, might assess the likelihood of an individual to default on their loan on the basis of the language used in their application more accurately than a bank employee behind a desk can with the aid of income figures, credit history, asset lists and so on.

This possibility presents a deep ethical challenge. It would seem likely that even the most optimistic believer in the beneficial power of big data would admit that there is at least something troubling about the picture of university applications assessed by automated algorithms focused on a limited amount of ‘highly predictive’ phrases that ‘correlate strongly’ with student success. We might say that this prospect troubles us because the value that we place in the notion of ‘relevant information’ cuts both ways. We value the notion that university applicants, prospective debtors and job candidates ought to be evaluated on the basis of information that is at least loosely relevant to the situation not only because it strikes us as the best way of generating future predictions, but also because it is precisely what make such processes seem ‘fair’ to us.

Apart from any ethical disquiet we might feel at the notion of being judged on the basis of information that does not seem causally relevant, the employment of such general profiling

\textsuperscript{15} Ibid., p. 72.
\textsuperscript{16} It is remarkable, in fact, how much Facebook activity is not covered by any of the available privacy settings. Almost anyone, regardless of mutual connection, is able to see what pages you have ‘liked’, even though they might not be able to see more personal things like photos, biographical details, etc. See Michael Zimmer, “‘But the Data Is Already Public’": On the Ethics of Research in Facebook,” \textit{Ethics of Information Technology} 12(2010).
\textsuperscript{18} Ibid., p. 5803.
techniques can have very real effects on the already disadvantaged.\textsuperscript{20} On the basis of using the wrong kind of statistically predictive vocabulary, a person might find themselves unable to secure a small loan.\textsuperscript{21} Another algorithm in a separate field, perhaps when that person applies for health insurance, might take that credit-unworthiness as a data point itself, and so we find that disadvantage is compounded on the basis of observations that may have no causal connection to his or her likelihood of defaulting on their loan. It is in this spirit that Frank Pasquale warns that “runaway data can lead to cascading disadvantages as digital alchemy creates new analog realities.”\textsuperscript{22}

3. Digging Deep

When applying for a home loan or university entry, we are not simply concerned about whether we are to be judged on ‘relevant’ considerations – that is, assessed (in some sense) on the strength of our application and not on whether we took time to thank the person to whom the application was addressed or whether we happen to have liked ‘Curly Fries’ on Facebook. There is also a sense in which it is important to us that even the relevant considerations are restricted to a certain level of ‘depth’ or ‘appropriateness’. If we consider a young couple applying for a mortgage, we might certainly admit that the likelihood that they will remain a couple is extremely relevant to any consideration of their future capacity to service the loan. Yet it would seem invasive and inappropriate if the application process for a mortgage required that couples answer detailed and personal questions about the strength of their relationship, or submit to some kind of joint psychological evaluation. In other words, there are limits to the kind of relevant information we wish to admit to such decision-making processes.

Yet when we consider what these limits are or what they ought to be, we often have in mind \textit{explicit} requests for information. In the early 2000s, several Australian court cases went a long way to establishing the cases in which employer-mandated drug tests were ‘justified’ and those in which they were ‘invasive’.\textsuperscript{23} The result has been that the situations in which employers are legally entitled to inquire into the drug use of their employees are tightly regulated.\textsuperscript{24} For similar reasons, companies that offer private health insurance may not inquire into the medical history of a prospective consumer beyond a certain level of detail before issuing a quote.\textsuperscript{25} In short: when we say that certain intimate kinds of information are off-limits from the point of

\textsuperscript{20} Lori Andrews refers to this process of automated profiling as ‘weblining’, suggesting that “internet profiling can create new minorities who do not even know how they are being discriminated against.” See Lori Andrews, \textit{I Know Who You Are and I Saw What You Did} (New York: Free Press, 2011).

\textsuperscript{21} All of this takes place, naturally, with a minimum of regulatory oversight. See Frank Pasquale, "Reputation Regulation: Disclosure and the Challenge of Clandestinely Commensurating Computing," in \textit{The Offensive Internet: Speech, Privacy, and Reputation}, ed. Saul Levmore and Martha C. Nussbaum (Cambridge: Harvard University Press, 2010).

\textsuperscript{22} Pasquale, \textit{The Black Box Society}, p. 32.


\textsuperscript{24} Fair Work Australia, "Alcohol and Drugs in the Workplace," Fair Work Australia, 2006.

view of a particular decision-making process, we very often mean that institutions ought not be entitled to ask for that information as part of the process.

With advances in big data, however, explicit requests become less and less necessary. Recall the statistical models that predicted various pieces of detailed social information about an individual on the basis of their public Facebook activity – the success rate of these models in predicting substance abuse fell between 65% and 73%. As techniques in machine learning become more advanced and larger stores of data become publicly available, such models will become more and more accurate. Suppose that such algorithmic crunching of public Facebook data becomes much more strongly predictive. Rather than navigating the legal challenges of asking employees to submit to drug tests, an employer may simply run an automated process which offers a verdict on the individual along with a numerical degree of confidence. Or imagine our young couple applying for a mortgage – if an automated statistical analysis of their social media activity and credit card history was capable of outputting a ‘Relationship Strength Index’ (represented by a number between 0 and 1), are we comfortable with the idea that the bank employee responsible for judging their application might have this number on her desk along with the rest of their willingly given information? Some American companies already track credit card purchases for indications of marital hardship or financial difficulty, reducing credit limits when purchases appeared that indicate that the couple has been attending marriage counselling.

That is to say that where our intuitive reactions to such cases are driven by norms regarding the kinds of questions one can and cannot ask in certain situations, we are approaching an era when such institutions do not need to ask such questions. We are no longer, as we might otherwise hope, the sole filter by way of which our most intimate and sensitive information might find its way into the world. Most perplexingly, it is not by way of invasions of privacy that such computational processes gain access to this information but, to borrow Sherlock Holmes’ description of his method, by “the observation of trifles.” It is only by coming to terms with the new digital realities that we inhabit that we might tackle the question of what is to be done with this information once it is inferred.

4. Conclusion: Lessons and Norms

So far we have gestured at two ways in which the use of big data challenges our traditional ethical intuitions and norms regarding our interactions with certain institutions. Not only do statistical models and machine learning allow institutions to make reliable predictions regarding our future behaviour on the basis of seemingly irrelevant information, but they also facilitate highly predictive inferences about relevant, personal information that they might otherwise be forbidden from explicitly requesting. In coming years, banks, government

27 “Other companies started cutting cardholders’ credit lines when charges appeared for pawnshops or marriage therapy because data indicated those were signs of desperation or depression that might lead to job loss.” See Charles Duhigg, "What Does Your Credit Card Company Know About You?,” May 12, 2009.
departments, and universities may find themselves in a position to generate increasingly general and intimate profiles of the people that pass before them in various capacities.

This possibility brings two ideas sharply into focus. The first is that the fixation on privacy that often characterises reactions to the advancing frontier of big data may be somewhat misplaced. In many cases, the interest that drives the development and employment of such techniques is not the discovery of information about one particular individual, but rather a desire to streamline various decision-making processes on the basis of highly predictive correlations. Causal connection to the situation at hand is decidedly optional. For instance, several US companies offer prescription reporting services that generate profiles of the risk associated with individuals seeking health insurance, and in doing so they do not tout the ‘accuracy’ of their models but rather the cost-saving benefits of a more automated assessment process. This allows companies to offer insurance to more people while avoiding those that are likely to incur higher future costs. Moreover, the data points that drive these predictive algorithms often appear at best tangentially relevant to the decision-making process to which they are applied.

The second idea is that in many cases it is not the inferred information itself that poses the ethical challenge. The efforts of researchers at Microsoft and Columbia University to assist in the difficult task of early detection of pancreatic cancer might seem to typify the benevolent use of big data about which we might have fewer ethical worries. Yet exactly the same algorithms, statistical techniques, and data might be employed by a health insurance provider in order to avoid customers that are likely to prove expensive in the future. That is to say that responding to the ethical challenge of big data is not a matter of circumscribing the particular sorts of information that it is acceptable for institutions of various kinds to infer about us.

Rather, our ethical response to the growing predictive power of big data ought to rest on considerations of the roles that such a diverse array of inferences play in our social, financial, medical and political lives and the decisions that shape them. In different fields, we must understand what we are to make of the fact that talk of relevance and irrelevance does not provide us with the same sort of grip on difficult situations that it once did, and we must come to terms with the fact that intimate details of our lives may be borne out in the aggregate by otherwise innocuous wells of activity. Any regulative project that focuses on restricting how data can be collected and the kinds of inferences that may be drawn rather than the role played by such inferences in institutional settings is unlikely to form the kind of bulwark against corporate abuse that we might hope for.

In this sense, providing answers to the questions of responsibility, curation and use that surround big data is, more than anything, a process of developing new norms and standards. Some applications of big data might strike us as jarring precisely because our current notions

---

30 In the United States, 1% of patients account for over 20% of total health care costs, and 5% of patients account for half of costs—picking and choosing the customers that you insure is far more profitable than taking all comers. See William W. Yu and Trena M. Ezzati-Rice, "Concentration of Health Care Expenses in the U.S. Civilian Noninstitutionalized Population," 2005.
31 Richards and King suggest that “the potential for social change means that we are now at a critical moment; big data uses today will be sticky and will settle both default norms and public notions of what is “no big deal” regarding big data predictions for years to come.” See Neil M. Richards and Jonathan H. King, "Big Data Ethics," Wake Forest Law Review 49(2014): p. 393.
of privacy, relevance and appropriateness do not clearly suggest what these norms and standards ought to be in such a novel landscape. The task of adapting our notions and developing such norms extends far beyond this essay. It will need to be a conversation that considers, amongst other things, the (mostly) legitimate desire of various corporate entities to minimise exposure to risk, the desire of the individual to be judged by such institutions on the basis of information that is not merely correlative and the disquiet we might feel at the fact that the institutions with which we interact are more and more possessed of Sherlock Holmes-esque omniscience.

As this conversation takes place, it is important that we keep in mind that there are far deeper issues at stake than merely those that pertain to individual privacy and the storage and control of the data itself. Big data poses a challenge to many of the distinctions and intuitions that we use to articulate our reactions to certain situations and find a balance between competing interests. Given some of the applications of big data that we have seen, what role remains for distinctions between relevant and irrelevant information, or between explicit and implicit collection of data? To fixate on privacy is to enshrine at the centre of our considerations notions of control, information, and consent to which big data already poses enormous challenges.

Given the realities of big data, consideration of the way that institutions use statistical models and algorithms in making decisions that shape our lives in a variety of ways seems a better starting point for the conversations that must take place. In a turn of events that would come as no surprise to Rudyard Kipling, we find ourselves as individuals becoming lost in the aggregated data of the tribe. Yet as the frontier of big data advances, it is not clear that the “privilege of owning yourself” that Kipling describes can be bought at any price short of technological and financial hermitage. Perhaps some might find comfort in such a retreat, but the more prosaic amongst us might instead look to the impression made by big data on the various institutions that colour our lives. At the very least, you should make sure you like ‘Curly Fries’ on Facebook.

**Word Count:** 3492

**Works Cited**


Goodman, Bryce. "What’s Wrong with the Right to Genetic Privacy: Beyond Exceptionalism, Parochialism and Adventitious Ethics." In The Ethics of Biomedical Big Data, edited


