CHAPTER 5

TAKING SHORTCUTS: CORRELATION, NOT CAUSATION, AND THE MORAL PROBLEMS IT BRINGS

Kevin Macnish

ABSTRACT

Large-scale data analytics have raised a number of ethical concerns. Many of these were introduced in a seminal paper by boyd and Crawford and have been developed since by others (boyd & Crawford, 2012; Lagoze, 2014; Martin, 2015; Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016). One such concern which is frequently recognised but under-analysed is the focus on correlation of data rather than on the causative relationship between data and results. Advocates of this approach dismiss the need for an understanding of causation, holding instead that the correlation of data is sufficient to meet our needs. In crude terms, this position holds that we no longer need to know why *$X+Y=Z$. Merely acknowledging that the pattern exists is enough.*

In this chapter, the author explores the ethical implications and challenges surrounding a focus on correlation over causation. In particular, the author focusses on questions of legitimacy of data collection, the embedding of persistent bias, and the implications of future predictions. Such concerns are vital for understanding the ethical implications of, for example, the collection and use of 'big data' or the covert access to 'secondary' information ostensibly 'publicly

Ethical Issues in Covert, Security and Surveillance Research
Advances in Research Ethics and Integrity, Volume 8, 55–70

ISSN: 2398-6018/doi:10.1108/S2398-601820210000008006

available'. The author's conclusion is that by failing to consider causation, the short-term benefits of speed and cost may be countered by ethically problematic scenarios in both the short and long term.

Keywords: Correlation; causation; ethics; big data analytics; legitimacy; bias

INTRODUCTION

> This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behaviour, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves. (Chris Anderson, then Editor-in-Chief of *Wired*, writing in 2008. Cited in boyd & Crawford, 2012; see also Ananny, 2016; Hildebrandt, 2011).[1]

Large-scale data analytics have raised a number of concerns in recent years. Many of these were helpfully introduced by boyd and Crawford and have been developed since by others (boyd & Crawford, 2012; Lagoze, 2014; Martin, 2015; Mittelstadt et al., 2016). One such concern which has led to only limited discussion is that of a focus on correlation of data to results in preference to a focus on the causative relationship between the data and the results (boyd & Crawford, 2012; Mittelstadt et al., 2016). Adherents to this approach, captured eloquently in the above quote by Anderson, dismiss the need for an understanding of causation. Instead the correlation of data is assumed sufficient to meet our needs.

An example of the focus on correlation rather than causation can be found in a TED talk by Jennifer Golbeck (2013). Research linked the enjoyment of curly fries to a higher level of intelligence than 'normal'. This is almost certainly not because eating curly fries *makes* a person more intelligent, nor because being intelligent *causes* a person to like curly fries. Rather, it may simply be because one person with a higher-than-average intelligence liked curly fries on Facebook as a joke, and her friends continued that joke. As people who have a higher-than-average intelligence tend to be friends on social media with others with higher-than-average intelligence, this leads to the strange correlation between intelligence and liking curly fries. The consequences, however, are that companies wishing to market products to people of higher-than-average intelligence may choose to target people who like curly fries, knowing that there is an established correlation and irrespective of the reason for that correlation. Worse, recruiters may start to look for apparently random but established correlations in seeking candidates for jobs. While being the subject of a marketing campaign is not typically intrusive (although it can be: see Ebeling, 2016, pp. 49–66), being turned down for a job in preference for another person simply because *they* like curly fries is definitely objectionable.

At the same time, large-scale data analytics can provide a wealth of benefits to individuals and society. This extends well beyond personalised coupons

or targeted advertising to significant advances in public health and welfare. If research can establish clear connections on a national scale with existing data regarding links between, for example, red meat consumption and bowel cancer, or between obesity and particular foods, then governments may have an obligation to carry out this research. The purpose of this chapter is not to dismiss data analytics as unethical *en masse*, nor to provide an overview of ethical concerns. It is rather to draw attention to the ethical issues arising from one particular aspect of data analytics, namely the attention paid to correlation in preference to (or instead of) causation.

The curly fries example is relatively easy for non-statisticians to understand. In that case, a correlation occurs between people with a higher-than-average IQ and people who 'like' curly fries. However, data analytics is typically far more opaque than in this example. A more complex case is the famous case of retail chain Target discovering that a teenage girl was pregnant before she told her father (Duhigg, 2012). In his article, Charles Duhigg suggests a number of items in the teenager's shopping basket would, when taken together, indicate to the data analytics team at Target that she is pregnant. These include buying more moisturiser, switching to unscented moisturiser, and buying a large bag that could double as diaper bags when the baby arrives. Independently none of these would bear much significance. It is when taken together that they take on a relevance not seen in the individual parts.

The items Duhigg lists bear an obvious relation to the pregnancy of the teen, but this need not be the case. The data analytics team at Target were not deciding which items were relevant based on their intuitions and then tracking these. Rather they were reverse engineering the known shopping baskets of customers who were also known to be pregnant to see what goods they had bought, and then searching for these prospectively with customers whose state of pregnancy was not known.

In Duhigg's example, then, there is an understandable causative role played by the items in the shopping basket: we can quickly see why at least some of these items, when taking together, might give an indication that the customer is pregnant. Likewise, in the curly fries example, there is a confounding variable (one not measured but which has an influence on seemingly unconnected results), namely the fact that the correlation is noted on a social media platform where like-minded people commonly cluster together.

However, it is not implausible that there could well be other items in a shopping basket which, again taken together, could indicate that a customer is pregnant. Imagine that these include 2lbs of carrots, 1lb of celery, scented candles, pillow cases, four AA batteries, and a CD. These could compare with a normal shopping basket of 3lbs of carrots, no celery, scented candles, eight AA batteries and two CDs. The first basket, through a long series of correlations, could provide an indication that the customer is pregnant, while the latter does not provide such an indication. That is, in this example, I want to stipulate a lack of *any* causative connection between the items bought and the results of the data analysis, and a lack of *any* confounding variables. Nonetheless, the correlation holds in 70% of cases and so it appears to be a reasonable assumption to make that this shopper is pregnant.[2]

In this chapter, I look at the ethical issues that arise from this focus on the correlative rather than the causative. I argue that this focus may be effective in the short term but it is ethically problematic. In particular, I hold that there are three reasons why we should be cautious of this approach. Firstly, this approach leads us away from considering the legitimacy of the data collected. Secondly, the approach lends itself to the embedding of persistent outliers and bias. Thirdly, the approach may miss the fact that, as any investor in financial services will say, past performance is no guarantee of future success. As noted above, there may be confounding variables that could be found to underpin a number of discoveries in big data. However, to stress the point that I am making about an unhealthy focus on correlation of data to the detriment of understanding causation, I will assume cases where there are no confounding variables.

LEGITIMACY

Central to the approach currently taken to big data analysis is that the data scientist frequently does not know which data will be relevant and which data will not. The response is to collect all of the data available in order to see which prove to be relevant. Does this matter? Clearly there are ethical questions that need to be raised about how the data are obtained, particularly whether the person providing the data has given valid and fully informed consent for that data to be collected and used in this way. Assuming this is the case, though, does it matter that all available data are collected and processed?

I want to argue that there is a significant difference here between what data are effectively legitimate, what are legally legitimate, and what are morally legitimate to collect. The current paradigm confesses its ignorance of what are effectively legitimate, and so seeks to collect all the data available, in order to discover which are effective in producing the results that interest the data scientist. This is tempered by legal restrictions as to what data are permissible to collect and what are not. Depending on the particular site or nation state, it may be illegal to collect data on a person's voting or health records. Also important of course is who is doing the collecting: it may be legal for a government to collect some data which it is not legal for a business to collect, or vice versa. One consideration that can be overlooked in this decision, though, is what data are *morally* legitimate to collect.

Whether it is morally legitimate to collect the data may never occur to the data scientist. It is not that they are necessarily malicious: their intentions are good and they want the best results for all concerned. However, intentions only go so far and consequences (intended and unintended) need to be considered as well. They may also choose not to see the collection of the data as their problem. Their job is to analyse the data provided. How that data are gathered is an issue for someone else.

Imagine a case in which data are collected by a university on the educational achievement of its students and the skin colour of its students. Then imagine that a clear correlation is drawn between the academic achievement of a student and the colour of their skin, so that it transpires that black students perform

worse than white. There could of course be any number of reasons for this, from the university running scholarship programmes for black students from deprived backgrounds who have further to go in order to catch up with white colleagues who tend to come from more privileged backgrounds, to racism among the university staff who grade papers.[3] The discovery of this correlation, though, could lead to a number of outcomes. Among these are that the university may choose to do nothing, or it could carry out an audit of staff and students to uncover any hitherto undiscovered racist assumptions, or it could choose to focus on recruiting white students in order to raise its results in national league tables.

While these consequences differ in their ethical acceptability, they also raise the question as to whether the data should have been collected in the first place. In this instance, the good that can be achieved by discovering staff with racist assumptions might be sufficiently advantageous to justify the collection of these data. However, this response implies that the *cause* of the correlation is of interest to the university. This seems to be morally unproblematic and not my focus here. By contrast, my starting assumption has been that the cause is not of interest but rather the results themselves (in Anderson's words, 'the numbers speak for themselves'). If that is the case, then the university would ask not '*why* is this data the way it is' but 'what could we *do* with these results?' The consequences of the collection are therefore focussed on action, which is itself governed by a series of values, rather than on research or discovery.

Even if the data were used for beneficent ends, it does not follow that all data that could be collected *would* be used for beneficent ends. Could a university morally collect the voting records of its students, or a list of their sexual partners? These seem to be more problematic. Certainly there might be a beneficent desire to uncover causes and help the students in some way, but the potential for abuse may increase with ever more intrusive data collection. Ultimately, the point may be reached at which the potential for abuse outweighs the potential for benefit to the students. At this point it would seem that the university would not be justified in collecting the data. The scale of intrusiveness of the data collected is clearly also a problem.

Thirdly, it is worth noting that different data may be collected for different reasons. The ethnicity of students may be collected for the morally legitimate end of ensuring that the university's recruitment process is not biasing against students on the grounds of their ethnicity, or to pass to government records aimed at monitoring social trends. Likewise, the university would be legitimate in collecting data about student performance. Indeed, it would be failing in its role as an educational institution if it neglected to pay attention to the academic performance of its students. However, the scenario becomes more problematic when these independent data sets are combined and subsequently used for a hitherto unforeseen or unanticipated end.

Ultimately, the collection of data on people is a form of surveillance, which raises a number of ethical issues (Lyon, 2002). In order to be ethical, surveillance should be subject to a number of limits regarding who is carrying it out and whether they are accountable, why they are doing so, whether it is proportionate, whether there are less intrusive ways of arriving at the same end, whether the

collection is likely to be successful in achieving the justifying cause, and whether the surveillance is discriminating between those who are liable and those who are not (Macnish, 2014). It does not seem unreasonable that the collection of data for analysis should be subject to the same ethical considerations as other forms of surveillance.

One response to the surveillance objection may relate not to the collection of new data, but the use of historic data already collected. This is not, it may be argued, an act of surveillance. The surveillance occurred at the point of collection. This is data arising from past surveillance which is now available for data scientists to use. Furthermore, given that the data exist, there may even be an obligation for it to be subject to analysis in the interests of, for example, public health.

The immediate counter to this response is to point out that surveillance does not involve merely the collection of data, but the collection *and processing* of that data. Surveillance may exist without the processing of data (i.e. it is not a necessary condition), but in this case, taking a broad understanding of what is meant by processing, there would be little purpose in the surveillance. As such it would not be justified on the grounds of the aforementioned need for a just cause. Furthermore, if my government collected my emails five years ago but did not read them at the time of collection but chose to read them today, I would argue that I have been under surveillance both at the time of collection and at the time of processing (i.e. reading) of my emails. This distinction gets to the heart of the revelations made by Edward Snowden in 2013 that the USA, UK and other governments were collecting large quantities of internet data relating to domestic citizens. While the intelligence communities at the time protested there was no violation of privacy as the internet data had not been accessed, it was nonetheless an act of surveillance (Macnish, 2016).

A further problem to arise from the processing of historic data is the possible lack of informed consent given by the owners of that data for its processing (see, e.g., Foster & Young, 2011). This has been one problem with the UK government's recent attempts to capture citizens' health data in a centralised database known as 'care.data'. This concerns health data which were initially given by patients to their GPs (general practitioners – their primary care physicians). Their reasons for doing so were probably legion, although one reason was almost certainly not the pooling of that data for future analysis. Certainly some may not have found the pooling of their data for analysis objectionable, just as many do not find this objectionable today. However, the fact remains that in giving the data to the GP, the patient did not give informed consent for this particular use of that data. As such, it is right that the patient be sought out in order to gain informed consent for this secondary use of the data.

There are understandable concerns with this focus on informed consent such that public health could suffer as a result of paying too much attention to finding and addressing the concerns of every citizen before accessing their data, a price that is too high to pay when the costs are minimal (Ganesh, 2014). However, this is to underplay the procedural importance of informed consent and the potential harms involved in the pooling of medical data.

Procedurally, the gaining of informed consent is central to the ethics surrounding the collection of data relating to individuals. This was emphasised in the Nuremburg Code and again in the Helsinki Declaration. There is a concern that a precedent will be established in ignoring the need for informed consent, with long-term ramifications. If liberal democracies cannot abide by their own standards in medical ethics, then they sacrifice any moral high ground in responding to others that do the same in more objectionable instances. To act on information over which people should have control without their consent is an abuse of their autonomy and can have a severe impact on their lives.

Secondly, there are a number of potential harms arising from the pooling of medical records, not least the discovering of health records of individuals, be they public figures, employees, or insurance applicants. It may be objected that this is not the intention of data scientists, and that there are security measures in place to prevent the leaking of information. However, it should be remembered that Edward Snowden managed to walk out of what is one of the most secure buildings in the world with millions of documents of top secret information on a memory stick. Given that Snowden had been subject to some of the most intense scrutiny before being allowed into those buildings, including polygraphs and background checks, it is naïve to suggest that health records will not be discovered and removed for the purpose of personal gain from large-scale pooled databases.

A final concern with the focus on legality rather than the moral legitimacy of the collection of data is that technology moves apace of legislation. Problems usually have to arise for a number of years before they lead to the introduction of legislation, and then a few more years before the response is passed into law. This enables unethical practices to continue unaffected for a considerable period before they are ultimately ended by legislation. It also overlooks the fact that while legislators generally try to ensure that laws are ethical, this is not always the case. Competing interests are often brought to bear on the legislative process and laws passed today have to be consistent with laws passed yesterday. As such, it may be that the law, when it is passed, does not go far enough to protect those it was designed to protect.

In speaking of collecting legitimate data we may therefore be using the term in one of at least three ways: effective, legal, or moral. What is legal and effective to collect may not be moral. Or, it may be moral to collect data for one purpose but then unethical to use it for a different purpose. This latter concern is especially pertinent to historic data. A standard approach to alleviate the concerns regarding inappropriate collection or use of a person's data is to seek their informed consent. While this may be cumbersome and slow the process of analysis with a clear public benefit, there are sound ethical, procedural, and practical concerns that mean we should not sidestep this. Finally, there are dangers in an over-reliance on legislation to guide morally legitimate collection owing to the pace of change in the former, which can easily be outstripped by developments in technology.

EMBEDDING OF PERSISTENT OUTLIERS AND BIAS

A second problem for large-scale data analytics is the introduction of bias and discrimination. Bias can be obvious, such as the aforementioned case leading to

the increased recruitment of white students over black students, or through similar forms of social sorting such as upmarket stores sending coupons to regular customers but not occasional customers. The end result of this process is that the wealthier regular customers pay less to use the store than poorer occasional customers (Lyon, 2002). When this is seen to be the case, that bias can be guarded and, to some extent, legislated against.

More problematic is hidden bias which is by its nature less easy to discover. As an example of hidden bias, imagine a public transport system which is successfully designed to serve the needs of 90% of the public. This sounds laudable: no system is perfect and 100% use is unrealistic without coercion. What, though, if the 10% who are not served by the transport system all fall into this category because they are unable to use it, owing to some form of disability? Certainly one cannot please all of the people all of the time, but to discriminate against someone *purely* for that person's disability, even if the discrimination is unwitting, is clearly unethical.

In the case of everyday statistics, such hidden bias is a possibility. Given the scale and focus on likelihood of correlation of big data analytics, though, it is not only probable that hidden bias would be present, but probable also that such bias would remain hidden for longer. That a system has a 90% success rate may be indicative of a highly effective and desirable system. Yet, if the 10% of occasions when the system fails always involves the same people, or group of people, then there is a problem.

Returning to the scenario of university recruitment, imagine that a university is seeking to improve its place in the national league tables. To this end, it carries out data analysis on the students currently achieving the highest grades and looks at their behaviour in the last two years of school before attending university. This avoids the obvious bias encountered before of correlating grades with skin colour. The results, when they are collected, then indicate that the best-performing students all played polo, golf, and sailed in this crucial two-year period. This leads to the university increasing its recruitment efforts among school students who sail and play golf and polo. There is a reasonable likelihood at this point that the university would effectively be focussing its recruitment efforts on fee-paying schools rather than state schools, in the process recruiting more white than black students, as well as more wealthy than under-privileged students.

It is logically possible that most students who attended fee-paying and predominantly white schools are more successful at university than those who attended state schools. This could be because they are more intelligent or because they are advantaged in some way by the fee-paying education system, or a number of other reasons. It is demonstrably false that people have higher intelligence purely because their parents had sufficient money to send them to a fee-paying school. It might, though, be possible that students at fee-paying schools are trained better in critical thinking or independent work than those at state schools, and so are better prepared for university when they arrive. The result is that they 'hit the ground running' while those from state schools feel as if they are constantly trying to catch up. Alternatively, they may simply be better prepared by the school for the requirements to get into university and are as such better at 'playing the game' than those who attended state schools.

In this case, the university faces a choice as to whether to improve its place in the national rankings by recruiting from fee-paying schools, or by providing an opportunity for those less privileged to develop their skills in critical thinking and independent working. If the former is true, then the university's actions could be justified by (and may even genuinely be determined by) the results of the data analysis. They are not seeking to be elitist or racist, although these are likely to be the results of their actions.

The above scenario imagined one university engaged in this practice. However, if one imagines every university taking a similar approach then an obvious problem emerges. In the first instance, recruitment becomes focussed on and heavily competitive for students from fee-paying schools. With time, though, further accumulated data will predominantly come from students who have attended fee-paying schools, thus focussing recruitment still further on a few key fee-paying schools and not even considering state schools at all. It is also entirely feasible that some schools become aware of which extra-curricular activities are favoured by the best universities (some schools are capable of conducting their own data analytic processes on historic cases) and start to offer and endorse those activities to pupils. Once more, the better-resourced schools will be more successful than those with more stretched budgets. In both cases, though, rather than being liberated by education, those less privileged find that university education at least, and the opportunities that go with it, have returned to being a preserve of the wealthy.[4]

While this is a hypothetical example, one does not have to look far to uncover cases in which hidden bias persists in algorithmic approaches to social problem-solving. Perhaps the best known of these is the city of Boston's adoption of a smart phone app to automatically locate potholes in public roads (the city fixes 20,000 potholes every year). This meant that rather than wait for people to complain about the state of the roads, the city could respond more quickly. The app may even have been perceived as having a social levelling effect, given that underprivileged groups in society may be less likely to complain about the state of their roads than other groups. If this was the case, though, the app was unsuccessful. As Kate Crawford (2013) notes,

> People in lower income groups in the US are less likely to have smartphones, and this is particularly true of older residents, where smartphone penetration can be as low as 16%. For cities like Boston, this means that smartphone data sets are missing inputs from significant parts of the population – often those who have the fewest resources.

As Crawford notes, Boston's Office of New Urban Mechanics was aware of this problem and worked hard to adjust for it, but it does not take a leap of the imagination to consider the impact of this thinking had the Office been less careful.

One such instance has been noted in relation to motor insurance policies, which are increasingly being tied to the installation of a 'black box' in the owner's vehicle and which monitors driving habits. These can then be correlated with the habits of other drivers so that profiles are developed regarding 'safe' drivers and 'unsafe drivers' based on their driving. The immediate effect is to penalise unsafe

drivers for their more risky behaviour and, one hopes, in the long term promote safer driving for all.

One such example of an indicator of unsafe driving is the time of day or night at which a person is on the roads and the frequency with which they drive. According to Robinson and colleagues, one system operated by insurer Progressive favours drivers who do not go out at night and who drive infrequently. However, they go on to point out, drawing on research by Maria E. Enchautegui (2013), the unintended effects of this may serve to punish particular communities, and especially those on low incomes, who 'are more likely to work the night shift, putting them on the road late at night, and to live further from work' (Robinson, Yu, & Rieke, 2014, p. 6).

In essence, the Progressive system puts late night workers into a similar category as late night party-goers,

> forcing them to carry more of the cost of intoxicated and other irresponsible driving that happens disproportionately at night. Statistically speaking, this added cost does not simply reflect the risk that the late night commuter may be hit by a drunk driver. It also reflects the possibility that, as far as the insurer can tell, the late responsible night worker may be a drunk driver. (Robinson et al., 2014, p. 6)

Rather than spreading risk among the insured population, then, the system focusses that risk on particular groups who are already marginalised in society.

A final case which deserves mention is that of Latanya Sweeney's discovery that online searches for names typically associated with black people had a significantly greater chance of returning advertisements which related to arrests than names associated with white people. Names associated with black people

> generated ads suggestive of an arrest in 81 to 86 percent of name searches on one website and 92 to 95 percent on the other, while those assigned at birth primarily to whites, such as Geoffrey, Jill and Emma, generated more neutral copy: the word 'arrest' appeared in 23 to 29 percent of name searches on one site and 0 to 60 percent on the other. On the more ad trafficked website, a black-identifying name was 25% more likely to get an ad suggestive of an arrest record. (Sweeney, 2013, p. 1)

Sweeney notes that this occurs not because of an explicit bias in the software, but because that software

> learns over time which ad text gets the most clicks from viewers of the ad. It does this by assigning weights (or probabilities) based on the click history of each ad copy. At first all possible ad copies are weighted the same, they are all equally likely to produce a click. Over time, as people tend to click one version of ad text over others, the weights change, so the ad text getting the most clicks eventually displays more frequently. (Sweeney, 2013, p. 34)

In essence, then, the software learns over time to reflect the biases which exist in society (Robinson et al., 2014, p. 16).

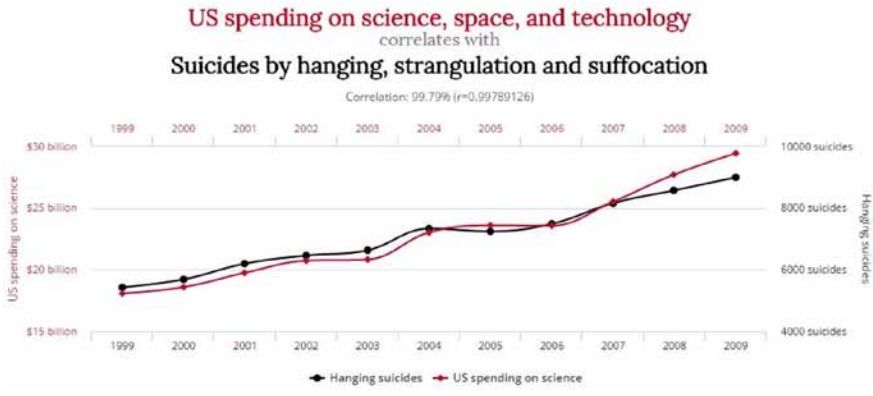
In this section, we have seen a number of cases, hypothetical and real, in which a focus on correlation and not causation of results is ethically problematic in that this can mask discrimination and bias. There is hence a concern that through focussing on correlations and failing to uncover the story behind those correlations, hidden biases might remain hidden for longer. Worse still, those biases could be exacerbated through decisions made on the basis of correlative data

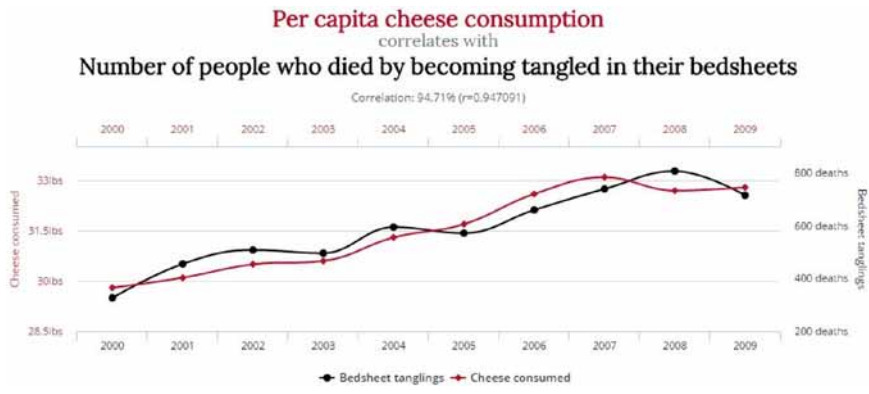
which has not been subject to adequate scrutiny regarding the causes of those correlations.

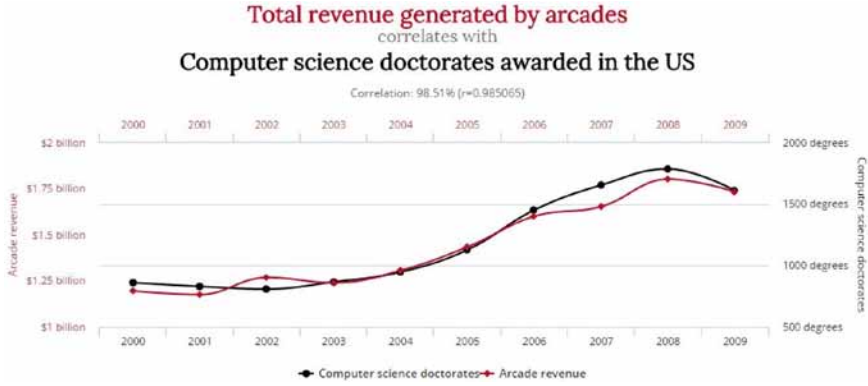
PREDICTING THE FUTURE ON THE STRENGTH OF THE PAST

It is an oft-repeated phrase in financial services that past performance is no guarantee of future success. Just because an investor has done well in the past there is no way of being sure that she will continue to do well in the future. Past results may have come about through luck or through a confluence of events that have since ceased to pertain.

The same is true in data analytics when causation is ignored for correlation. On an entertaining website, and now book, Tyler Vigen (2015a, 2015b) has provided graphs of a number of examples which fit the concerns raised in this chapter of seeking correlation without concern for causation. These include the following from among a list of 30,000:







In using these charts I am clearly oversimplifying, and drawing on oversimplifications, for ease of illustration. No one would attempt to predict the number of science doctorates to be awarded in the USA by basing the prediction purely on the total revenue generated by arcades. However, the focus on the correlation effect in big data analytics at the expense of causation could have similarly absurd, but less obvious, effects.

Imagine a case in which an analysis of all terrorists of any stripe, up to the present, have all walked at precisely 57 metres/minute (leaving to one side the thorny question of how we define a terrorist). Further analysis might also demonstrate that anyone who is not a terrorist has always walked faster or slower than 57 metres/minute. Does this mean that it would be reasonable to develop an automated security system that recognises and disables only those who walk at 57 metres/minute? To some degree this might be sensible: the statistics as stipulated appear to be overwhelming. However, there are a number of problems with this approach.

Firstly, returning to the opening quote of this section, the past is not always a reliable indicator of the future. Aside from standard challenges to inductive logic regarding geese at Christmas, whereby the goose is always welcomed into the farmhouse kitchen with food until Christmas morning when she is welcomed with a cleaver, when there is no confounding variable (such as the knowledge of Christmas traditions) or known reason for the correlation, there is no reason to presume that correlation will continue. Hence, just because all and only terrorists have until now walked at 57 metres/minute, there is no guarantee that all and only terrorists in the future will walk at 57 metres/minute. Indeed, should it become known that all and only terrorists walk at 57 metres/minute (and that this is being used as a means of identification of terrorists) then future terrorists are likely to consciously adopt a different walking speed.

Secondly, it is important to understand where the data are drawn from in these cases. It is, after all, impossible to measure the walking pace of every person on the planet and even if we were to do this it would only be relevant

to the time at which the measurement occurred. Walking speeds change with age and circumstance, as well as with culture. Currently, data sets regarding walking behaviour tend to be developed in the West and so have a predominantly western bias (Macnish, 2012). Hence, the aforementioned problem of hidden bias can enter the system through the choice of data set and have a significant impact on the system's applicability and ability to accurately predict the future. The implications of this are that the use of a seemingly strong correlation to identify terrorists could be flawed either over geography or over time, with the result that innocent people are harassed and stigmatised (Macnish, 2012).

The terrorist example is one with serious consequences. It is no light matter to be mistaken for a terrorist. However, if the outcomes of the analysis are comparatively trivial then this is less of a problem. One might be tempted to say that the pregnant teenager in Duhigg's story was a fairly trivial case of data analytics. The father did not presume that his daughter was pregnant but rather that Target was acting irresponsibly in sending her coupons for items that a pregnant woman might want. If he had thought that she was pregnant on the basis of the coupons alone, and if he had a particularly low view of pregnant teenagers, then the consequences for the daughter could have been far more severe. Taking this not to be the case in this instance, though, the worst that would happen in a scenario in which Target had a 60% success rate in identifying pregnant women, would be that 40% of those identified received coupons that they would never use. Furthermore, in signing up to a loyalty card programme, customers accept that they will get coupons through the post (indeed, many do it for this reason), often assuming that these will be fairly arbitrary and that some will therefore be of little interest to them. Indeed, Duhigg (2012) notes in his article the creepiness for a customer of realising that Target knows she is pregnant on receipt of such tailored coupons. The response, he claims, has been to include coupons for random items that it is known no pregnant woman is likely to want, such as lawn mowers or garden furniture. This is not to say that pregnant women would not want these items, merely that they do not relate to pregnancy in the way that other coupons might.

Serious cases are not restricted to security, though. They might also emerge in the health and public welfare sectors. For example, the discovery of a correlation between those who use a certain prescription drug and those who die when under general anaesthesia should rightly lead to hospitals warning patients not to use this drug when they are about to undergo an operation requiring a general anaesthetic. If, though, both the drug and the operation are significant to the patient's life, then the patient will be forced to choose between the two.

As things stand, this is a regrettable but not unconscionable scenario. Such things happen. However, it may be that the manufacturer of the drug used a particular compound in the composition of the drug which was inert in the delivery of the drug but which reacted negatively with a certain level of anaesthetic. If the supplier stopped using this compound, purely by chance, then the forced decision would cease to be an issue. Owing to the use of historic data in deriving the correlation, though, no one would know this without further tests being carried out.

It is not implausible that the manufacturer of the drug stopped using the compound after the data were collected but before the results were published, and so the ensuing warnings would be unnecessarily harmful.

Each of these cases is to a greater or lesser extent plausible. Furthermore, it is not unreasonable to draw conclusions and base future predictions on past data. Once more, though, the concern in this chapter is not the basing of future predictions on past data but the basing of such on past data alone without seeking the reasons for any correlations. For example, the correlation between the drug and death while under anaesthesia is plausible, and it is a valuable activity to notice this correlation and warn others in the light of the perceived pattern. However, without working to understand why there is a correlation, subsequent changes to the drug might go unperceived.

The practical challenge to this warning in particular is likely to be that such practices (seeking correlation as grounds for prediction without uncovering causative factors) are more likely to proliferate in business than in security or medicine. In such cases, the harms are more akin to receiving irrelevant coupons than people dying or terrorists evading capture. Yes, the response would come, the system is not perfect but then we do not seek perfection, and where is the harm?

To this my response would be that people's behaviour is governed in part by the expectations that are placed on them. If these expectations are derived from an arbitrary group then the expectations themselves risk being arbitrary. If the data set used is of current customers, who are overwhelmingly male, then the predictions may be significantly more pertinent to men than women. If this governs not only coupons but also marketing and design of stores, this may make it harder for women to use those stores (if maybe only for social reasons such as a predominance of certain styles and colours in the store windows and no female assistants such that the average woman needs to sum up greater courage in entering the store and prepare herself for a degree of mansplaining). This in turn has a societal impact of at best embedding and condoning existing social divisions, and at worst implicitly endorsing and furthering those divisions, which should be avoided. Hence, even apparently harmless or low risk uses of correlative data for predictions can have significant outcomes.

CONCLUSION

I have warned here against an *exclusive* focus on correlations in big data analytics. Failing to consider causation may be effective in the short term but it will prove to be ethically problematic in both the short and long term. Paying attention to causes could overcome the three problems explored here: the legitimacy of the data collected, the emergence of bias and persistent outliers, and the difficulty of predicting future events on the basis of current data.

The challenge is that in considering causes, the data scientist will lose the advantages of speed and complexity that accompany big data. Being forced to examine why a particular correlation occurs might slow the publishing or using of that data, with potentially significant social effects. If, for instance, a strong

correlation is found between eating a certain food and colon cancer, then it may well be prudent to advise people to refrain from eating that food prior to discovering the reason for that relationship, which could take years to uncover. At the same time, as I hope to have demonstrated, there are also potentially significant social effects that arise from ignoring causes as well. It is hence short-sighted to promote the quick returns that can be gained by ignoring causes as being justified by social benefit. The fuller picture shows social benefit *and harm* that can arise from this approach and so it is not one to be taken lightly.

NOTES

1. I believe that Anderson has since retracted this statement, although cannot find reference to this retraction.
2. Clearly, the dark arts of big data (Target's chief data analyst was prevented from communicating with Duhigg when Target discovered that he had discussed his work with a reporter) are vastly more complex than I have suggested here (Duhigg, 2012). However, to attempt to capture this complexity would detract from the central argument. I shall therefore continue to use simplified cases to press home the concerns with this approach of favouring correlation over causation.
3. Although intended as a hypothetical example, this is sadly the case in UK universities at least. See Alexander and Arday (2015). I am grateful to Rosemary Hill for drawing my attention to this report as well as for comments on an earlier draft.
4. A similar situation in the workplace is imagined by boyd, Levy, & Marwick (2014).

REFERENCES

Alexander, C., & Arday, J. (2015). *Aiming higher: Race, inequality and diversity in the academy*. Runnymede Perspectives. London: Runnymede.

Ananny, M. (2016). Toward an ethics of algorithms convening, observation, probability, and timeliness. *Science, Technology & Human Values*, *41*(1), 93–117. doi:10.1177/0162243915606523

boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679. doi:10.1080/1369118X.2012.678878

boyd, d., Levy, K., & Marwick, A. (2014). The networked nature of algorithmic discrimination. In S. P. Gangadharan, V. Eubanks, & S. Barocas (Eds.), *Data and discrimination: Collected essays* (pp. 53–57). Washington, DC: Open Technology Institute.

Crawford, K. (2013). The hidden biases in big data. *Harvard Business Review*. April 1. Retrieved from https://hbr.org/2013/04/the-hidden-biases-in-big-data

Duhigg, C. (2012). How companies learn your secrets. *The New York Times*, February 16. Retrieved from http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Ebeling, M. (2016). *Healthcare and big data: Digital specters and phantom objects*. New York, NY: Springer.

Enchautegui, M. E. (2013). *Nonstandard work schedules and the well-being of low-income families*. Paper 26, Urban Institute, Washington DC. Retrieved from http://www.urban.org/research/publication/nonstandard-work-schedules-and-well-being-low-income-families

Foster, V., & Young, A. (2011). The use of routinely collected patient data for research: A critical review. *Health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medicine*, *16*(4), 448–463. doi:10.1177/1363459311425513

Ganesh, J. (2014). Big data may be invasive but it will keep us in rude health. *Financial Times*, February 21. Retrieved from http://www.ft.com/cms/s/62a5aaaa-9a55-11e3-8232-00144feab7de,Authorised=false.html?siteedition=uk&_i_location=http%3A%2F%2Fwww.ft.com%2Fcms%2Fs%2F0%2F62a5aaaa-9a55-11e3-8232-00144feab7de.html%3Fsiteedition%3Duk&_i_referer=&classification=conditional_standard&iab=barrier-app#axzz4JeevuGEk

Golbeck, J. (2013). The curly fry conundrum: Why social media 'likes' say more than you might think. *TED*. Retrieved from http://www.ted.com/talks/jennifer_golbeck_the_curly_fry_conundrum_why_social_media_likes_say_more_than_you_might_think

Hildebrandt, M. (2011). Who needs stories if you can get the data? ISPs in the era of big number crunching. *Philosophy and Technology*, *24*(4), 371–390. doi:10.1007/s13347-011-0041-8

Lagoze, C. (2014). Big data, data integrity, and the fracturing of the control zone. *Big Data and Society*, *1*(2), 1–11. doi:10.1177/2053951714558281

Lyon, D. (2002). *Surveillance as social sorting: Privacy, risk and automated discrimination*. New York, NY: Routledge.

Macnish, K. (2012). Unblinking eyes: The ethics of automating surveillance. *Ethics and Information Technology*, *14*(2), 151–167. doi:10.1007/s10676-012-9291-0

Macnish, K. (2014). Just surveillance? Towards a normative theory of surveillance. *Surveillance and Society*, *12*(1), 142–153.

Macnish, K. (2016). Government surveillance and why defining privacy matters in a post-Snowden world. *Journal of Applied Philosophy*, *35*(2), 417–432. doi:10.1111/japp.12219

Martin, K. E. (2015). Ethical issues in the big data industry. *MIS Quarterly Executive*, *14*(2), 67–85.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 1–21. doi:10.1177/2053951716679679

Robinson, D., Yu, H., & Rieke, A. (2014). *Civil rights, big data, and our algorithmic future*. Washington, DC: Upturn.

Sweeney, L. (2013). *Discrimination in online ad delivery*. SSRN Scholarly Paper ID 2208240. Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2208240

Vigen, T. (2015a). 15 insane things that correlate with each other. Retrieved from http://tylervigen.com/spurious-correlations

Vigen, T. (2015b). *Spurious correlations*. New York, NY: Hachette Books.