**Chapter 2**

**The haystack fallacy, or why Big Data provides little security**.

Ann Rudinow Sætnan (000-0003-4147-7836)

**Abstract**

This chapter examines the promises made on behalf of Big Data and compares these to its potential for errors. The author concludes that while Big Data collections and analyses very likely confer net marginal advantages in settings where small margins matter while inaccuracies are of little consequence, and while it also may be useful for generating hypotheses in science, Big Data carries significant dangers in settings where life and death decisions are made on the basis of the data, often at a distance from where the decisions are carried out and have their lethal effects.

**Introduction**

Our increasingly digitised lives leave myriad electronic ‘breadcrumbs’ behind us, morsels of information about many of our activities and even thoughts.[[1]](#endnote-1) Take for example the simple act of accessing the day’s news. Only a few decades ago, media information streams were almost exclusively one-way – from media to the public. Up-stream traces left by media consumption were minimal and mostly voluntary. Newspapers were predominantly bought anonymously, for cash. Subscriptions left data links between media and address but said nothing about who in the household read what in the paper, listened to what on the radio, or watched what on the television (if indeed anything at all). Traces were left if you wrote a letter to the editor, but most readers never did. Broadcast audience numbers were estimated based on survey panels comprising minute fractions of the total audience – participation voluntary and responses self-registered. Advertisers had to base marketing decisions on media sales summaries, survey estimates of audience size, and whatever changes they could see in their own sales after a campaign. Certainly some subscriptions were tracked for political surveillance. Back in pre-internet times, the detour certain checks took via the FBI was a large part of what constituted internal espionage overreach. All this is not so long ago, but for younger generations it is already buried in history.

Now media information streams are definitely two-way, and not narrow streams but broad, silt-laden rivers. Accessing an on-line ‘newspaper’ sends back traces not only of having logged onto a given front page, but also of which articles we opened, which advertisements we clicked to follow, which videos we watched and for how long, whether we ‘skipped’ the ads or not, whether we ‘shared’ or ‘liked’ any items, whether and what we commented. Some programs track each keystroke, or tap into the resulting comment texts. Some even track eye movements, showing what catches viewers’ attentions and how fast or slowly readers scan through an article. The volume and granularity of information flow from audience back to media has increased monumentally. We have become more than consumers of media content. Now we, or at least our ‘data doubles’ (Haggerty & Ericson 2000), are also its products.

Furthermore, media information streams are not merely two-way, but flared – frazzling out into multiple strands at either end that can then be braided together with other streams. At both ends of the stream, data are not simply used for one pre-defined purpose. They are recycled, repurposed, shared, modified, commodified, capitalised, requisitioned, appropriated, expropriated, stolen, etc.

Of course, media usage is not the only example of this. More and more aspects of our lives are becoming digitised, with the resulting information crumbs being stored and made reusable in this way. I recently spent some weeks as a visiting scholar in Århus, Denmark. Early on, I tried to find a map of the bus routes around the city, but the bus company no longer makes them. Instead, passengers are told they must download the bus company’s mobile app. With the app, you can find all you need to know about bus routes, and buy and download tickets. But of course, such a system is no longer anonymous. With a paper map, no one would know when I consulted it, what I looked for on it or whether I followed up my inquiry with a journey. A bus app records all these things: When did I look up a route on the map? Where was I when I did so? From where to where did I want to go? Did I download a ticket? How did I pay for it? Did I flash the electronic ticket on my screen to the reader instrument on the bus, and if so when? In a city of some 300 000 inhabitants, that rapidly builds a vast bank of travel information the bus company previously did not have and that now potentially offers them commercial value in its own right. My cell phone provider may also be collecting some or all of these data in parallel with the bus company, and both may be farming out data for analysis, selling it on for profit, storing it in ways not well secured against hacking ... In short, these data may be spread in ways I have little control over.

When the granular ‘silt’ of each such data stream is amassed over time and geo-social space, and especially if masses from various data streams (telecommunications data, GPS signals, library loans, credit card transactions, merchandise sales, acceleration and braking registered on your car’s computer, RFID tags passing by recording portals, steps counted by your activity bracelet, etc. etc.) can then be combined and repurposed, analysed for commercial, medical, scientific, political, policiary, security … well, whatever goals – then we are in the realm of the phenomenon now known as ‘Big Data’.[[2]](#endnote-2)

 Big Data technophiles tell us that we are entering a marvelous new age of data-supported convenience with vast commercial and scientific opportunities; technophobes that we are entering a ‘Brave New World’ of lost privacy and increased opportunities for oppression. Agnostically, we could ask these questions: *What purposes are well served by Big Data analyses and what goals might suffer? Can Big Data be used to make us safer, or does it inherently make us less safe***?** This chapter addresses those questions by critically examining the claims made for Big Data accuracy and usefulness in general, and then taking a closer look at how these accuracy and usefulness issues might affect our security when Big Data are applied towards that particular goal.

**Setting chapter parameters**

*What are data?*

The empiricist view of data, and very likely the popular view among those of us not trained in science theory, is that data are simply the recorded, uncontestable, observed facts about things, people and events around us. For instance, one on-line dictionary defines data as:

1. *Facts* [emphasis added] that can be analysed or used in an effort to gain knowledge or make decisions; information.
2. Statistics or other information represented in a form suitable for processing by computer.

(thefreedictionary.com, no date)

In our daily lives, we of necessity act, at least for the most part, as if facts and data were such simple things. When we need to make snap decisions, however vital they may be, we of necessity take the same stance – a viewpoint called empiricism in science theory. Empiricism is a convenient simplification. To actively doubt and question each observation would be impractical, to say the least. But in a science context, or when we have time to take some care about vital decisions, we need to be aware that empiricism is flawed. Facts are not so obvious, observable, or uncontestable.

Firstly, how do we turn our endless and boundless flow of observations/sensations/emotions into discreet facts? The cosmos is a continuum without uncontestable boundaries. I may anthropocentrically think of myself as a neatly bounded body, separate from the air I breathe, the water I drink, the bacteria that form an essential part of my digestion and my skin … but I am not neatly bounded. I am porous. I-as-I-anthropocentrically-think-of-myself and my-surroundings-as-I-anthropocentrically-think-of-them emerge through intra-actions with one another (that is, with our shared selves) and are thoroughly intra-dependent (Barad 2007). To consider conceptual units such as my-anthropocentric-self as one discreet unit about which discreet traits might be measured and recorded and the air as another such unit, or to similarly distinguish between ‘different’ species is to draw boundaries which, however conventional and/or well-reasoned, are nevertheless at some level arbitrary. Such boundaries could always have been drawn differently, and sometimes *are* drawn differently. And this drawing of boundaries is just the first step of turning the cosmos into data.

Turning these constructed samenesses and differences into data involves further layers of constructed conventions and/or individual choices: what features of each category are worth recording (what counts?), and how they are to be counted (by what means and in what measurement units?). Each of these layers is equally arbitrary and changeable. The social processes of reaching a degree of consensus, layer by layer, are often fraught with contention (Bowker and Star 1999). It is only once a degree of consensus has been reached and/or enforced, once units and measurements have been conventionalised and standardised, that the resulting data can be made to appear obvious, natural, observable, uncontestable – until, that is, someone again makes the effort to contest them (Latour 1987). Furthermore, explicit and implicit contestations, however marginal, are frequent. Therefore – as also discussed by Noorman et al. (this volume) – data require custodianship in order for standardisations to retain a degree of stability.

All this is not to say that data are useless or erroneous fictions. It only means that they could *always* have been otherwise and that it pays, especially in some contexts, to keep this in mind. The same goes for their (in)stability. It is not that data are too unstable, too fragile, to be of use. It is only that it pays to be aware of their potential instablity and fragility and of the procedures being used to maintain and stabilise them.

*What makes data ‘big’?*

Big Data is big along one or more of several dimensions that characterise data sets. First and foremost, it is big in terms of numbers of cases or instances recorded – numbers of persons, numbers of vehicles, numbers of transactions … whatever the initial units triggering recording might be. Secondly, data sets referred to as Big Data often contain records of many variables about each case/instance/unit. Thirdly, such sets often include linkage potentials to further data sets, big and/or small, thus allowing expansion of the original data set. Fourthly, such data sets are usually big on speed – that is, the data they contain are often recorded automatically and more or less instantaneously as each unit occurs: as a person enters a shop through an RFID portal, as a credit card transaction occurs, as a vehicle passes an autopay toll station, as an email is sent, and so on. Fifth, such data sets are often structured and regulated to be big on flexibility, making the data available for a variety of analytical purposes. Separately and together, these features generally thought to make Big Data big on power: statistical power, power to produce knowledge, power to influence events. However, as we shall see, each of these features also entails potential weaknesses.

When learning statistical research methods, one of the earliest lessons is that more data is better. More data (more cases) means more statistical power, more likelihood of statistically significant results, more widely generalisable results, less risk of findings being distorted by a handful of cases with entry errors or outlier values. But … more data also means greater difficulty checking for errors.

Similarly for more detailed data (more variables and more linkages). Here too there are both advantages and risks. The more detail we have, the less likely we are to get stuck with an exciting finding, one that raises new hypotheses, and then not have relevant variables to explore that finding further. We’re also less likely to get stuck with no variables that can explain whatever variations we see in outcomes. But … more detailed data also means greater risk of finding spurious patterns.

Flexibility is also a source of both advantages and risks. The varied sources and types of data available in Big Data sets, the level of data detail and the potential for linkages all invite analysts to pose a wide variety of questions the data might perceivably offer answers to. This increases the practical, intellectual and market value of the data. But … the repurposing of data to answer questions they were not initially collected to address carries a number of risks. Remember that data, however conventionalised and seemingly natural, are socio-material constructs. Their social and material construction histories leave an imprint on them. Data collected for one purpose are shaped by that purpose, which may render them mis-shapen in the context of some new purpose. While data expressions (a code or number) may well be both immutable and mobile (Latour 1987: 227), their meanings are not. At the very least, a process of reinterpretation must take place, if only an instantaneous (and mistaken) one of declaring the data to be pure and universal reflections of truths that travel effortlessly and unchangingly from one context to another.

Furthermore, in Europe at any rate, the repurposing of data may run afoul of privacy law.Privacy law in Norway (and Norwegian law is in this regard typical for European laws more widely) demands that any collection of personal data – i.e. data that can be traced back to an identifiable person – must be mandated by either a specific law, a vital need, or the informed consent of each person concerned. Each of these sources for mandate requires an identified purpose for the data collection. When legitimately collected and stored data are moved to a new user and/or reused for some new purpose, the original mandate may not be presumed to follow. The law that mandated their collection and storage may not apply. The vital need may not be relevant. The informed consent cannot be presumed unless the new user and purpose were included in the original information given and consented to.[[3]](#endnote-3)

In evaluating the claims of Big Data advocates, I will be considering both the advantages and risks of Big Data’s characteristics. I will be asking how these advantages and risks seem likely to play out given what Big Data advocates claim regarding what Big Data can and should be used for and how.

**What are Big Data good for given how they are created, processed and used?**

Ideally, I should be answering this question by delving directly into Big Data collection processes and analysis algorithms. Those, however, are for the most part proprietary and/or confidential. It will take time to gain access, and then time to explore them deeply enough for interpretation – possibly more time than I at 67 have remaining in my career. Algorithms are discussed by Matzner (this volume), but for my part I will now turn to the more general claims of Big Data advocates.

Authors such as Mayer-Schönberger and Cukier (2013) and Anderson (2008) make a number of claims as to why and how Big Data can ‘extract new insights or create new forms of value’ that cannot be achieved at a smaller scale (Mayer-Schönberger and Cukier 2013: 242). Anderson’s version is the most concise. His claims can be enumerated as follows:

1. The massive availability of data causes a new analytical approach.
2. Models, including causal models, are no longer relevant.
3. Theory is no longer relevant.
4. Data speak for themselves.
5. Correlation is enough*.*
6. Correlation results are correct to an unprecedented degree.

To this we can add a seventh point, not explicitly mentioned by Anderson, but implicit there and posited explicitly elsewhere (e.g. critically by Harford (2014)) as an argument for Big Data, namely that N (the number of instances included in the database for a given phenomenon) now equals very nearly all and that since N=all (or very nearly all) we are no longer dealing with samples nor, therefore, with sampling error. This is purported to explain why ‘correlation is enough’: it is enough because it is the whole.

So how do these seven arguments hold up to scrutiny? Not entirely well.

*Because petabytes*

Ignoring the irony that Anderson (2008) begins with a causal argument just before declaring causality moot, his causal argument is historically wrong. The availability of mass data is not the *cause* of Google’s or others’ inductivist/empiricist[[4]](#endnote-4) approach to data analysis; it is a *result*. A substantial portion of those data were intentionally created by, for instance, Google and Facebook, whose business plans at the outset were to produce, harvest, analyse and market masses of data. Yes, there is also a vast amount of data that has been produced incidentally, as a by-product from other purposes – for instance communications traffic data initially produced for billing and/or for reconstituting data packets back into entire messages at the receiving end. However the mere existence of that data does not force us to use it for other purposes than those for which it was created. In fact, as discussed above, such repurposing of data may be illegal.

Furthermore, even if legal, storing and using those data for new purposes implies both empiricism (as discussed above, the view that data are simply the registered ‘mirror’ images of observable and uncontestable truths, endlessly mobile and immutable) and inductivism (the view that knowledge is best built from data upwards by observing patterns in the data and then examining whether those same patterns hold for larger bodies of data).

Empiricism has already been discussed above as a useful, even necessary, but erroneous simplification. Inductivism is only ‘wrong’ if practiced absolutely and exclusively. Many disciplines practice inductivism as their preferred style of research. However, inductivism practiced thoroughly involves testing ones initial interpretations of patterns by looking at larger and larger data sets to see whether and under what circumstances the pattern still holds. In other words, good inductivism entails rather than precludes deductive steps, working ones way gradually upwards in data set size and/or outwards to data sets involving more and more categories of units. What Andersson and other Big Data advocates are claiming is that, given the vastness of the data sets available, one can skip those pesky, time-consuming deductive and/or iterative steps by going straight to the biggest data set and looking for patterns there inductively. But no, it is wrong to conclude that the ultimate data set in terms of size is the ultimate proof of any observed pattern and that work with smaller data sets is a waste of time. For instance, a pattern that holds true for a large data set viewed as a whole, might not hold true for some subset of that data set. Suffice it to say that iterative research, moving back and forth between data and tentative conclusions, is a form of ‘due diligence’, that jumping to conclusions in a single step is rarely wise, that inductive approaches are fine but best when combined with deductive ones as well, and that regardless of data availability we still have methodological choices to make. Soaring data handling capacities and falling data handling prices have created an *opportunity* (but not a requirement) which actors of an inductivist bent have seized. But, as I intend to show, even that opportunity has its limits.

*[Causal] models are moot*

Not so. Or at least, not entirely so. Yes, there are situations where we needn’t care about causality, only about how acting on our predictions and understandings affects bottom line outcomes. One such situation is when gambling. In gambling, all we need is a marginal improvement in outcome probabilities, yielding us better odds than those taking our bets have calculated. We don’t need to understand why, as long as we stop gambling before our predictions cease to produce that statistical advantage. Similarly with advertising: Even a small margin saved on advertising costs or a small margin of increased sales will produce increased profits, without necessarily having to understand the reasons for customers’ preferences. But, if we do happen to have a reliable causal understanding of why certain customers buy our product, we may be better able to anticipate shifts in the market and thus maintain our market advantage longer. Furthermore, there are other situations where an understanding of causality provides not just a statistical advantage, but precisely the sought-for outcome.

For instance, research is often directed towards offering policy advice, and policy is inherently a causality-motivated exercise. Policy is, ideally and practically, based on a balance between assessments of means and of desired outcomes. Each of these assessments is not only about pragmatics, but also about morality and politics as well. The question is not only what can we do, but also what should we do (or not do) and what is it socially acceptable that we do (or not do). Of course, pragmatics are also a part of the latter two questions, since what we should do and what it would be accepted that we do are to a significant extent matters of causal assessment: how effective are the means we are considering in terms of producing the outcomes we are aiming for? If we decline to discuss causality, we render ourselves irrelevant to policy discourse.

Security measures also build, implicitly or explicitly, on causal models. When various US authorities repeatedly threaten to bomb some enemy or terrorist threat ‘back to the Stone Age’,[[5]](#endnote-5) this implies understandings (a.k.a. theories) about the causes of violent opposition, about what motivates or demotivates individuals to join in such opposition, and/or about what creates or destroys a group’s capability to carry out such opposition. Acknowledging the theoretical implications and critically examining them might help to identify a (more) successful way forward.

*Theory is irrelevant*

Again, not so. Or at least, that depends on what we mean by ‘theory’ and ‘relevant’. I would claim that theory is what invests observations with meaning, taking them from the realm of noise to the realm of data. Theory, in this sense, is our interpretive framework – however tentative, however much we treat that framework skeptically and are prepared to revise it. Of course, we may observe first and theorise later (the inductive path) or we may theorise first and observe on that basis (the deductive path). As discussed above, it is generally advisable to combine the two approaches in iterative alternations. Nevertheless, no matter which approach we have given to analysing numbers and other simplifications, sooner or later we must flesh out these numbers, skeletal simplifications that they necessarily are, with meanings. This brings us to the next point.

*Numbers speak for themselves*

Not so. Mathematics, I would argue, is an abstract language, a set of rules for thinking and communicating about anything socially deemed enumerable. Through those rules, numbers are positioned in relationship to one another, but are left empty of otherwise meaningful content. X can stand for anything; so can 1; so can 5347-56 or any set of values or place-holders in any mathematical equation however well their relationships to one another are defined by the equation. For a mathematical formula or any product of that formula to speak about anything beyond its own empty shells and grammatical rules, we have to fill those shells and relationships with meaning, we have to interpret, we have to add theory-relevant data.

Interpretation is a job numbers do not perform on their own. Indeed, it is a job that is never finally completed and settled. It is always possible to discuss the meanings of a set of data or the outcome of a data analysis. It is always possible to contest a meaning once agreed upon, or to add new layers of meaning to those already accepted, i.e. data are always ‘interpretatively flexible’ (Collins, 1981).

Sometimes interpretation is a purely personal matter (glass half full or half empty?). Often it is a social process. When researching with ready-produced data, it is important to know how each number/text string/classification/code was arrived at and what meaning(s!) it is understood to hold (for examples of this see Bowker and Star (2000), Sætnan, Lomell and Hammer (2010)).

*Correlation is enough*

This goes against experience as taught in Probability 101. The more data you throw together into a database, the more likely it is that some of those data will fall into patterns, some of which will be entirely random and meaningless, also known as *spurious correlations*. For instance, did you know that over the 6-year period 2008-2013 the number of lawyers in West Virginia corresponds almost perfectly (correlation coefficient[[6]](#endnote-6): 0.987938) with overall US on-line revenue on Thanksgiving? Over the 12-year period 1999-2010, it correlates nearly as precisely with the number of suicides by hanging, strangulation and suffocation in all the US (correlation coefficient: 0.968413) (tylervigen.com, no date)[[7]](#endnote-7)

Advocates for Big Data claim that, for trend analysis, it doesn’t matter if correlations make sense. You don’t have to know why; correlation alone matters. If communities hit by flu coincidentally had surges of on-line dog biscuit orders, that too could have been included in the Google Flu Trends algorithm[[8]](#endnote-8). Doctors do use population-level correlations when diagnosing. It matters that you have symptoms x and y but not z. It may also matter that you live in region a (where one possible diagnosis explaining your symptoms is more common than another), have job b (which exposes you to certain disease vectors), gender c (which entails a different hormonal spectrum), are d years old (and therefore more or less likely to get childhood diseases, more or less likely to show symptoms of degenerative diseases). But should I someday consult my doctor concerning suicidal thoughts, I hope s/he won’t base her/his diagnosis on the number of lawyers in West Virginia, but only on factors that make logical sense to her/him and to me.

*Results show unprecedented accuracy*

Ironically, this is a claim supported by anecdotal rather than statistical evidence. We are told of instances where non-spurious correlations were found between variables and instances where inferences about individuals in the data set were drawn on the basis of such correlations and proved correct, but we are not told how many times similar inferences were drawn and proved incorrect.

There is the famous story about when Google, using search term statistics, spotted a flu epidemic two weeks before the same wave of infection was spotted by the Centers for Disease Control on the basis of physicians’ reports. The story that Google Flu Trends later has *mis*-spotted flu epidemics (Harford 2014) is less referenced.

Equally famous is the story of how Target, tracking loyalty card links to sales of certain items, sent out baby product ads to a teen-aged girl. Her father, outraged at first, apologised two weeks later when the girl admitted to him that she was in fact pregnant. This story is referenced in informatics textbooks as a triumph for Big Data (e.g. Siegel, 2013; Bari, Chaouchi and Jung, 2014; Sanders 2014). The story is not always accompanied by discussion of why Target and others pursuing this form of advertising have since changed strategy (Greengard 2012), hiding such advertising by bundling it amongst numerous non-targeted ads and offers, thus protecting themselves against backlash should the targeted advertising prove hurtful because the algorithm’s ‘diagnosis’ is wrong, or unethical even when it is right.

*N=all*

N is never all. There are always some who are missing from (at least some parts of) the data set – for instance those too poor, too young, too old, or too infirm to own smart phones, hold credit cards, purchase electronic travel cards, or access the internet. There are also those who intentionally stay ‘off the grid’, be it for ideological, nefarious or other sorts of reasons. Furthermore, even if N=all at some given moment in time or space, that time-space is always a non-random sample from an infinite time-space continuum. A pattern may be real today, yet gone tomorrow.

For instance, consider Boston’s experiment with using smartphones’ GPS sensors and accelerometers to spot and report potholes in city streets. The algorithm ‘works’ to some extent, but must be corrected for the bias that neighborhoods with predominantly poor and/or aging populations – areas that often have the  worst street conditions and/or greatest need for good conditions – have far fewer smart phones (Crawford 2013). The Boston pothole algorithm, then, is an example of the ‘streetlight fallacy’. The streetlight fallacy gets its name from a joke:

A man almost stumbles over another man on his hands and knees searching the ground under a streetlight. ‘What are you looking for?’ he asks. ‘I’m looking for my wallet,’ the searcher replies. ‘Oh! Did you lose it near here?’ ‘No, I lost it over in that alley.’ ‘Then why aren’t you searching in the alley?’ ‘Because the light’s better over here!’

Similarly with Big Data: We are tempted to look for more or less everything in the vast data bases that contain information about nearly everything, but the answers we need may require information that is *not* in those databases. That doesn’t mean that it’s wrong to look for answers through Big Data. It only means that it pays to think first what question you’re asking and what data would best answer that question, then use Big Data databases only if the data you need is there.

And then there’s the old adage ‘Garbage in, garbage out’, which holds true also for enormous databases. One of the problems of large databases is that their very size makes it impractical, if not impossible, to proofread, fact check and correct erroneous data. Even when data are automatically entered from digitised processes such as credit card transactions, there may be errors. If someone has hacked your credit card information, transactions attributed to you may not have been made by you at all, even if the data were correctly recorded as entered during the transaction itself. Such errors are non-random and will not cancel each other out in the vast mass of data; you have not accidentally charged other items to your hacker’s account, nor have you sought symmetrical revenge by hacking the hacker’s account. There will also, inevitably, be instances where data are wrongly entered more or less at random. A recent study of data quality in the Google Scholar, ISI Web of Science and Scopus databases (Franceschini et al. 2016) found that Google Scholar, though useful as a search engine, had far too many errors to be useful as a research tool. ISI and Scopus databases were perhaps useful, but results would have to be interpreted cautiously as they showed 6% and 4% errors respectively.[[9]](#endnote-9) Tüür-Fröhlich (2016) found even worse results for the social science database ISCI. For one of the original articles he traced, only 1% of citations were registered without errors. Furthermore, the errors were not randomly distributed, but were systematically skewed by factors such as publication language.

**So what?**

What does it matter that Big Data as a source of knowledge is riddled with myths and errors? The answer depends on what it’s used for. As a business model, it may or may not be a good gamble that increases its users’ profit odds at least marginally. Economic risks do also get passed on to others (workers, suppliers, taxpayers, etc.) and those unchosen risks can be quite serious, but there are ways of ameliorating them. Who winds up carrying the risks is a political battle we may choose to engage in, but probably not to the point of proposing a ban on legitimate, consent-based data analysis. This does presume that the data are only used for the purposes to which we originally consented, which is far from always the case when it comes to our data being appropriated and re-purposed by Big Data applications. But, what of security applications? What happens when Big Data are deployed as a key weapon in a strategy of what Strauβ (this volume) calls ‘securitisation’?

When it comes to using Big Data-based approaches to security issues, besides data collection being often involuntary and opaque and thus in itself a threat to democracy, the consequences risked by actions (and, for that matter, inactions) based on those data can be drastic and irrevocable. Four factors make security applications of Big Data particularly dangerous – the actions based on the data analyses, the frequency of errors in the analytical results, the mistaken faith put in those results’ accuracy, and the overall social consequences of Big Data practices. Before discussing the actions and their consequences, let’s take a closer look at the accuracy issues. What happens to needle-location accuracy when you build a bigger haystack? A simple percentage calculation can show how dramatically accuracy falls.

First we need to know the size of the haystack, an estimate for the number of needles in it, and an estimate as to how accurately our search algorithm identifies needles and hay respectively. Let’s suppose we are searching for terrorists. In 2014, CNN asked a number of security experts how many jihadist terrorist groups and group members they thought there were in the world (Bergen and Schneider 2014). Estimates varied from 85,000 to 106,000 members. Of course, not all terrorists are jihadists, not all consider jihadists to be terrorists, and numbers of members change over time,[[10]](#endnote-10) but let’s go with those figures for now, and for simplicity’s sake let’s call it a round 100,000. As for the size of the haystack, in an interview with Laura Poitras, William Binney (former NSA data analyst, now whistleblower) states that the NSA’s data storage facility in Bluffdale , Utah has the capacity to store 100 years’ worth of all the world’s electronic communications (Poitras, 2012, minutes 2:35-2:53). The world population currently stands at over 7 billion. Again for the sake of simplicity, let’s stipulate that the world population is 7,000,100,000, of which 100,000 are jihadist terrorists. Of course, not all ordinary citizens are communicating electronically, but neither are all jihadist terrorists doing so. We could estimate some average number of electronic transactions per person per day, but if we estimated the same average for jihadist terrorists as for the population in general, then the two factors would cancel one another out. So, we may stipulate our haystack numbers as 7,000,000,000 ‘straws’ and 100,000 ‘needles’.

Now let’s give our hypothetical algorithm every benefit of the doubt. Let’s stipulate that we have an algorithm that analyses all these electronic transactions – phone calls, purchases by credit card, loyalty card swipes, internet searches, toll road passages, etc. etc. It is a complex and sophisticated algorithm. It analyses linkages among individuals and patterns of transactions. It is so incredibly precise that it correctly identifies 99% of the jihadist terrorist ‘needles’ and correctly exonerates 99% of the general populace ‘straws’. Only 1% (1000) of the ‘needles’ are missed – i.e. false negatives – correctly identifying the other 99,000 as true positives. Only 1% (70 million) of the ‘straws’ are drawn under suspicion (false positives), correctly exonerating the remaining 6 billion 930 million (true negatives).[[11]](#endnote-11) How accurate, then, are the algorithm’s positive ‘hits’ in the database?

As we can see in Table 1 below, the positive hits are far from 99% accurate. There are 70,099,000 (seventy million ninety-nine thousand) positive hits, of which only 99,000 are true positives. For this thought experiment, the rate of accuracy for each positive hit is 0.1412288335069% - *fourteen hundredths of one percent*. Feel free to experiment by changing the numbers in the Table 1 spread sheet. What happens when you increase the size of the haystack? What happens if you decrease it? What happens if the algorithm is less precise?

<Table 1 about here>

Usually there is a trade-off between false positives and false negatives. Suppose you agree with Blackstone that ‘it is better that ten guilty persons escape, than that one innocent suffer’ ([Harvard](http://library.law.harvard.edu/justicequotes/explore-the-room/south-4/) Law Library, no date). Then you would ‘tune’ your selection criteria to be as specific as possible, widening the openings in your search net so that only those highly likely to be guilty are marked for suspicion, even if that means that many others equally guilty go free. On the other hand, suppose you are seeking out asymptomatic sufferers of a deadly disease for which there are relatively painless and harmless tests and cures. Then you might tune your selection criteria to make the search net as tight as possible, even at the cost of frightening large numbers of healthy people and subjecting them to unnecessary further tests.  Choosing a balance point between sensitivity (identifying as many ‘needles’ as possible) and specificity (exonerating as many ‘straws’ as possible) is a moral and practical call. Is it better that ten guilty go free? This may depend on context and the associated consequences of an initial positive. It may well be acceptable that ten (or more) healthy are put through further tests for each extra case of breast cancer caught in a mammography screening program.[[12]](#endnote-12) In a context of identifying presumed terrorists, however, the consequences would be far different. Being mistaken for a suicide bomber doesn’t mean being subjected to further tests; it means being killed on the spot before you have a chance to set off a bomb … even if it turns out you weren’t carrying one after all. What then if, out of fear of false negatives, you allow for more false positives – or vice versa? The answer to that question may depend on how much trust authorities place in each positive hit from the database.

The math here is simple: Crunching big numbers can lead you astray – further astray the higher the ratio of ‘straw’ to ‘needles’ in the database you search through and further astray the higher the mass of numbers you crunch. Recall too, that the size and data accumulation speed of Big Data databases almost of necessity will result in more data errors and inadequate time for data proofreading and correction.

Furthermore, history shows that it may not be lack of information but inadequate interpretation of available information that has hampered prevention of terrorist episodes. For instance: various authorities had been warned about the 9/11 suicide pilots’ strange behavior; one of the brothers allegedly responsible for the Boston Marathon bombing was known to the FBI through two prior investigations as well as warnings from Russia; and, French anti-terrorist police knew of the men alleged responsible for the Charlie Hedbo and subsequent attacks, but explained that ‘there are far too many of them, and far too few of us’ (CNN 2015). Under such circumstances, it is hard to see how increasing the information load and increasing the number of suspects, especially by adding thousands of false suspects, can help in preventing future terrorist attacks.

And yet security officials – for instance NSA top brass, in interviews and Congressional committee hearings – continue to claim that results from sifting through masses of electronic transaction records are exactingly precise, that more data is what they need, that limiting their data collection practices means taking responsibility for future tragedies. NSA General Counsel Stewart Baker is cited as having said, ‘Metadata absolutely tells you everything about somebody’s life. If you have enough metadata, you don’t really need content.’ Michael Hayden, former director (at different times) of both the NSA and the CIA, concurs, saying, ‘Absolutely correct. We kill people based on metadata’ (Cole 2014). An otherwise anonymous drone operator, apparently attempting to de-dramatise the picture, said, ‘People get hung up that there’s a targeted list of people. It’s really like we’re targeting a cell phone. We’re not going after people – we’re going after their phones, in the *hopes* [emphasis added] that the person on the other end of that missile is the bad guy’ (Scahill and Greenwald 2014). It’s enough to give one pause any time you switch on your cell phone.

**In conclusion**

We have seen that in some ways Big Data simply doesn’t work. These weaknesses can be dangerous when applied to critical social functions such as security. But what of the social consequences of Big Data analysis more generally, in the areas where it does work? When I buy a book at Amazon.com, I am immediately told what other books previous purchasers of that book bought or ‘liked’. When I read certain articles on-line, or sign on-line petitions, newsfeeds and ad-feeds are adjusted, targeting me according to what the algorithm says are my interests and tastes. Sometimes the algorithm clearly gets it wrong. An accidental click on an advertisement now has me receiving endless tips on places to stay in Van Nuys, California. Irritating, but harmless. But let us assume for now that this ‘works’, that people have a better ‘Facebook experience’, more effective Google searches[[13]](#endnote-13), that Facebook and Google advertisers make more sales. Even if this doesn’t work, but Facebook’s and Google’s paying customers *think* that it works, then it at least ‘works’ to produce profits for Facebook and Google.  But what else might it produce? Does it contribute to a Balkanisation of society? Are we compartmentalised into ever-finer interest groups where we are increasingly exposed only to those who already agree with us? Does that create increasing polarisation of society, decreasing tolerance for ideas and values from outside our algorithm-defined groups (Tewksbury 2005, Schneider in this volume)? If so, then even when Big Data ‘works’, it may also work to make us less secure.

Recently we have also seen that Big Data may work in history-changing ways by marginally improving its users’ odds in a variety of odds-based situations. For instance, the Trump campaign, possibly together with Russian hackers, used Big Data analytics to pull out a narrow win in critical states (Doward and Gibbs 2017). They only needed to improve their margin by a few percentage points in the right places. When the voting was so close, it didn't matter if their targeted ads missed the mark in many cases, as long as they did hit the mark in a few thousand more; it didn't matter if targeted voter registration purges also took out some Republican voters, as long as they took out a few thousand more Democrats. In that sort of situation, if you don't care about the consequences for democracy, then Big Data can work. But, as discussed above, when it comes to making security decisions, relying on Big Data can be disastrous.

**References**

Anderson C. (2008). The end of theory: the data deluge makes the scientific method obsolete. *Wired Magazine*. <http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory>. (Accessed 22 Sep 2015)

Bari A., Chaouchi M. and Jung T. (2014). *Predictive analytics for dummies*. Hoboken, NJ: John Wiley & Sons.

Bergen P. and Schneider E. (2014). Jihadist threat not as big as you think. <http://www.cnn.com/2014/09/26/opinion/bergen-schneider-how-many-jihadists/>. (Accessed 22 Sep 2015)

Bowker G. C. and Star S. L. (2000). *Sorting things out. Classification and its consequences*. Cambridge, MA: MIT Press.

CNN (2015). <http://www.cnn.com/2015/01/11/world/terrorists-security-breaches-france/>. (Accessed 20 Oct 2017)

Cole D. (2014). ‘We kill people based on metadata’.  *The New York Review of Books*, 10 May 2014. <http://www.nybooks.com/blogs/nyrblog/2014/may/10/we-kill-people-based-metadata/>. (Accessed 22 Sep 2015)

Collins H. M. (1981). Stages in the empirical programme of relativism. *Social Studies of Science*, *11*(1), 3-10.

Crawford K. (2013). The hidden biases in big data. *Harvard Business Review*. <https://hbr.org/2013/04/the-hidden-biases-in-big-data>. (Accessed 22 Sep 2015)

Cullather, N. (2006). Bomb the back to the Stone Age: An etymology. <http://historynewsnetwork.org/article/30347> (Accessed 20 Oct 2017)

Desrosières, A. (2002) *The Politics of Large Numbers. A History of Statistical Thinking*. Cambridge, MA: Harvard University Press.

Doward J. and Gibbs A. (2017) Did Cambridge Analytica influence the Brexit vote and the US election? *The Guardian*, 4 March 2017. Available at <https://www.theguardian.com/politics/2017/mar/04/nigel-oakes-cambridge-analytica-what-role-brexit-trump> (Accessed 14 Aug 2017).

Elmore J. G., Miglioretti D. L., Reisch L. M., Barton M. B., Kreuter W., Christiansen C. L. and Fletcher S. W. (2002). Screening mammograms by community radiologists: Variability in false-positive rates. *Journal of the National Cancer Institute*, 94(18), 1373-1380.

Good, C. (2014). Cruz Invites Obama to Border, Calls for Bombing ISIS 'Back to the Stone Age' <http://abcnews.go.com/blogs/politics/2014/08/cruz-invites-obama-to-border-calls-for-bombing-isis-back-to-the-stone-age/>. (Accessed 20 Oct. 2017)

Greengard S. (2012). Advertising gets personal. *Communications of the ACM*. 55(8), 18-20. <http://cacm.acm.org/magazines/2012/8/153815-advertising-gets-personal/fulltext>. (Accessed 22 Sep 2015)

Franceschini, F., Maisano, D.A. and Mastrogiacomo, L. (2016) Empirical analysis and classification of database errors in Scopus and Web of Science, [*Journal of Informetrics*](https://www.researchgate.net/journal/1875-5879_Journal_of_Informetrics) 10(4): 933-953

Haggerty K. D. and Ericson R. V. (2000). The surveillant assemblage, *British Journal of Sociology*, 51(4), 605-22.

Harford T. (2014, March 28). Big Data – Are we making a big mistake? *Financial Times*. <http://www.stat.ucla.edu/~yuille/Courses/Stat161-261-Spring14/Big%20data_%20are%20we%20making%20a%20big%20mistake_%20-%20FT.pdf>. (Accessed 22 Sep 2015)

Harvard Law Library (no date). <http://library.law.harvard.edu/justicequotes/explore-the-room/south-4/>. (Accessed 20 Oct 2017)

Harzing, A.W. and van der Wal, R. (2008) Google Scholar as a New Data Source for Citation Analysis? *Ethics in Science and Environmental Politics*, 8(1): pp. 61-73.

Intel (2012). Peer research on big data analysis. <http://www.intel.com/content/www/us/en/big-data/data-insights-peer-research-report.html>. Accessed 22 Sep 2015

Laney, D. (2001). [3D data management: Controlling data volume, velocity and variety](http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf) (PDF). <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>. (Accessed 22 Sep 2015)

Mayer-Schönberger V. and Cukier K. (2013). *Big data: A revolution that will transform how we live, work, and think*. London: John Murray Publishers, Ltd.

Poitras L. (2012). *The Program*. Documentary film. Retrieved from <http://www.nytimes.com/video/opinion/100000001733041/the-program.html>. (Accessed 22 Sep 2015)

Saetnan A. R., Lomell H. M. and Hammer S., Eds. (2010). *The mutual construction of statistics and society*. New York: Routledge.

Sanders N. R. (2014). *Big data driven supply chain management*. Upper Saddle River, NJ: Pearson Education.

Scahill J. and Greenwald G. (2014). The NSA’s secret role in the U.S. assassination program, *The//Intercept*, 2/10/2014. <https://firstlook.org/theintercept/2014/02/10/the-nsas-secret-role/>. (Accessed 22 Sep 2015)

Siegel, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*. Hoboken, NJ: John Wiley & Sons.

Tewksbury D. (2005). The seeds of audience fragmentation: Specialization in the use of online news sites. *Journal of Broadcasting & Electronic Media*. [http://www.researchgate.net/publication/238400693\_The\_Seeds\_of\_Audience\_Fragmentation\_Specialization\_in\_the\_Use\_of\_Online\_News\_Sites. (Accessed 22 Sep 2015](http://www.researchgate.net/publication/238400693_The_Seeds_of_Audience_Fragmentation_Specialization_in_the_Use_of_Online_News_Sites.%20(Accessed%2022%20Sep%202015))

Thefreedictionary.com (no date) <http://www.thefreedictionary.com/data>. (Accessed 07 Oct 2016)

Tüür-Fröhlich, T. (2016) *The Non-trivial Effects of Trivial* *Errors in Scientific Communication and Evaluation*, Glückstad: Verlag Werner Hülsbusch

tylervigen.com (no date). <http://www.tylervigen.com/>. (Accessed 20 Oct 2017)

Ward J. S. and Barker A. (2013). Undefined by data: A survey of big data definitions. <http://arxiv.org/abs/1309.5821>. Accessed 22 Sep 2015

Wikipedia. (no date). Retrieved from <http://en.wikipedia.org/wiki/Big_data>. (Accessed 22 Sep 2015)

Wikiquote.org (no date). <http://en.wikiquote.org/wiki/Curtis_LeMay> (Accessed 20 Oct 2017)

1. But not *all* our thoughts and activities! We’ll return to this point later. [↑](#endnote-ref-1)
2. I use quotation marks here to signify that I am introducing Big Data as a concept, as opposed to simply big (adjective) data. Hereafter, when referring to that concept as a fact or a set of practices, I will only capitalise it. If I do use quotation marks around the concept later, it will be to highlight irony in contexts where I and/or others I cite raise doubts as to Big Data’s existence or capacities.  [↑](#endnote-ref-2)
3. See Personopplysningsloven [Personal Data Law] §§8 and 9. https://lovdata.no/dokument/NL/lov/2000-04-14-31   [↑](#endnote-ref-3)
4. Inductivism is a variant of empiricism. Both share the notion that the cosmos is as we observe it and that observation is the route to knowledge, but where some branches of empiricism tout hypothesis-testing through experimentation, inductivism argues that knowledge is built through the accumulation of experience. For more on the history and variations on this position, see for instance <http://www.vetenskapsteori.se/ENG/m1eepist.htm> or <http://www.philosophybasics.com/movements_british_empiricism.html>  [↑](#endnote-ref-4)
5. Among others, purportedly General LeMay re Vietnam in 1968 (Wikiquote.org, no date); Secretary of State Richard Armitage in 2011 re Pakistan’s resistance to joining the US in its ‘war on terror’, (Cullather 2006); and Senator Ted Cruz re ISIS (aka ISIL, IS, DAESH) in 2014, (Good 2014). [↑](#endnote-ref-5)
6. For those who have not taken Probability 101, the correlation coefficient is a measure of shared variability. Coefficients range from –1 to +1. A coefficient of –1 means that for every change in one variable, an equal and opposite change occurs in another. A coefficient of +1 means that an equal change occurs in the same direction for the second variable. A coefficient of 0 means that the two variables vary completely at random to one another. [↑](#endnote-ref-6)
7. The website is interactive, in case you wish to explore further spurious correlations. [↑](#endnote-ref-7)
8. While that hypothetical correlation, which I initially made up out of thin air, was meant to seem silly, it later occurred to me that it could have generated some researchable and potentially non-trivial hypotheses. For instance, feverish dog owners might need more treats to control their dogs while less able to give the dogs enough exercise, or might need to buy treats on line since they weren’t up to going out to the store. [↑](#endnote-ref-8)
9. Others have reached opposite conclusions. For instance, Harzing (2016) finds systematic errors in both Google Scholar and ISI but regards those in Google Scholar as less critical. [↑](#endnote-ref-9)
10. Here is another example of where causal models would be useful. Imagine if we knew what motivated people to seek political change through non-violent action and what motivated them to seek it through dramatic violence. Media reports indicate the effectiveness of terrorist groups in using such knowledge to recruit members. Media reports also indicate that, by comparison, governments are less knowledgeable and/or less effective at using such knowledge to slow or reverse such recruitment. See for instance *NY Times* June 13 2015 ‘US sees failure in fighting ISIS on social media’. [↑](#endnote-ref-10)
11. Given the flaws of Big Data discussed above, this estimate is obviously a gross exaggeration of the imagined algorithm’s accuracy. [↑](#endnote-ref-11)
12. Regarding actual false positive rates of screening mammography, see for instance Elmore et al. (2002). [↑](#endnote-ref-12)
13. More ‘effective’ in the sense of finding what you already tended to look for more rapidly, but at the same time reducing your chances of serendipitous findings you didn’t imagine might interest you? [↑](#endnote-ref-13)