Big Data – big trouble?

Meanderings in an uncharted ethical landscape

Robindra Prabhu
Norwegian Board of Technology
robindra.prabhu@teknologiradet.no

Few concepts have made as many headlines in the past few years as the term «Big Data». From its nascent beginnings in technology circles, the term has rapidly catapulted into mainstream society. In recent years it has even become a household notion in the higher echelons of government: the Obama administration has launched its multi-million dollar «Big Data Initiative» (Office of Science and Technology Policy 2012), the United Nations has established a network of experimental labs exploring the possibility of leveraging Big Data for humanitarian purposes (United Nations Global Pulse), and recently the Australian government became one of the first in the world to launch a «Big Data Strategy» for its public service (Australian Government 2013).

With a promise to fundamentally «transform the way we live, work and think» through extensive «datafication» of all things human (Mayer-Shönberger, Cukier 2013: 73–97), Big Data vows to give us unprecedented insight into complicated problems and a powerful toolkit to better understand the world around us. With the

excessive hyperbole that is often associated with novel technology trends, one would be forgiven for mistaking the buzz for hype with little substance. And despite its ostensible potential, the significant amount of excitement it generates and the widespread agreement that Big Data will impact our society in crucial ways, there appears to be surprisingly little clarity concerning what Big Data actually is and what it entails for society at large.

So what is Big Data, anyway?

To be sure, there is no shortage of definitions. The «Big» alludes to unfathomable troves of digital data, in various shapes and forms, which we deliberately or passively generate in our daily interactions with technology. Then there is our enhanced ability to store, manage and extract insight from these data troves using powerful computing technology and the latest in advanced analytical techniques. But «Big Data» does not refer to a fixed quantitative threshold or clear-cut technological constraint. Indeed what is considered «big», «complex» and «advanced» varies widely. So much so that researchers have found it necessary to collate various definitions of the term «Big Data» and furnish the following meta-definition:

> Big Data is a term describing the storage and analysis of large and complex datasets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning. (Ward, Barker 2013)

Perhaps it is only natural that early attempts to capture and define an allusive concept will come in many guises and possibly fall along a «moving technological axis». But while we struggle to pin down this new technology, it is important to recognise that Big Data's entry into the mainstream is equally about cultural changes in how we think about data, it's capture and analysis, and their rightful place in the fabric of society.

Whether hype or substance, and however most appropriately defined, the Big Data discourse is taking place against some profound (and I would argue exciting) changes in how we interact with our physical and social surroundings. These interactions invariably involve technologies and result in digital traces manifested in such various ways such as Internet clickstreams, location data from cell phones interacting with phone towers, data streams from credit card transactions, the logging of purchasing patterns in shops, the rich and multifaceted sensor data from an Airbus A380 in flight or vast detectors at research labs like CERN. Moreover, the emergent proliferation of low-cost sensors allows us to track and monitor objects and mechanisms in ways that were previously impossible. Farmers employ moisture sensors to monitor moisture levels in fields, shipments of fish and fruit are monitored for temperature and location in real-time as they are moved between continents, and people log personal health indicators using their smartphones.

Not only do we spend more time «online», but as we continue to add more «things» to the Internet, our lives become increasingly more entwined with the virtual world. And the digital traces we constantly leave behind in the virtual world now give us new handles on complex problems in the physical world.

The sceptic might demur that Big Data still has some way to go to deliver on its promise; targeted advertising and tailored movie recommendations may not appear to be the stuff of «revolutions». But fascinating applications are beginning to emerge: mobile phone data is being leveraged to map and track the spread of disease (Talbot 2013), law enforcement agencies are using predictive data-driven tools to determine the «where and when» of the next crime (The Economist 2013) and aggregate citizen sentiments are mined from large-scale social media feeds (Social Media and Post-2015) (Anderson 2008). In the future, more applications are likely to appear. And as we begin to deliberate on the wider ramifications of the rather nebulous

phenomenon of «Big Data», important clues for the ethical challenges ahead are likely to be found through close examination of the existing Big Data landscape. An appreciation of these challenges is not only of relevance to researchers who are dappling with new and vast troves of data, but also to private enterprises and governmental agencies looking to harness the power and potential of Big Data.

Treading ground between the enthusiasts and the sceptics

The nascent debate around Big Data may appear to be quite polarised. As Sandra González-Bailón remarks, the discussion on the proper governance and use of all these novel data sources has bifurcated public opinion into a two-pronged needle:

> … the sceptics, who question the legitimate use of that data on the basis of privacy and other ethical concerns; and the enthusiasts, who focus on the transformational impact of having more information than ever before. (González 2013: 147)

Both camps have extremists that will either dismiss the Big Data phenomenon as overhyped and underwhelming, or espouse the view that we are witnessing a new era in which the proliferation of data will render theory and interpretation superfluous (Anderson 2008) (Richards and King 2013).

While a healthy dose of both enthusiasm and scepticism is essential when dealing with new technologies, there are valuable lessons to be learned from the moderates on either side. Firstly, theory and interpretation are not likely to be discarded any time soon. Instead, their importance is reinforced as a sense-making tool in a growing sea of noisy data. Secondly, we would be wise to tread carefully, lest the critics are vindicated and we end up sleepwalking into a surveillance society.

As the debate and rhetoric advances and matures, it becomes important to capture the full range of nuanced challenges associated with the Big Data paradigm.

Moving beyond the hype: «Three paradoxes of Big Data»

An interesting turn in this direction is provided by Richards and King in their paper «Three Paradoxes of Big Data» (Richards and King 2013). While the authors do not deny the many benefits and the substantial potential inherent in Big Data, they advocate a more pragmatic discussion with due attention to the many faceted implications and inherent dangers of the Big Data paradigm by calling attention to three paradoxes in the current rhetoric:

1. *Transparency*: As sensors become ubiquitous and ever larger portions of our lives are mirrored onto a virtual world, enthusiastic proponents of Big Data argue that this pervasive data collection will serve to document the world as it is and make it more transparent.

   However, a fair portion of our personal data exhaust–small data inputs from sensors, cell phones, clickstreams and the like–are generally amassed into aggregated datasets «behind the scenes», largely without our knowledge. These datasets may in turn be saved in unknown and remote cloud services, where equally hidden algorithms mine the data for strategic insights. The paradox of this, they argue, is that if Big Data promises to make the world more transparent, then why is it that its «collection is invisible, and its tools and techniques are opaque, shrouded by layers of physical, legal, and technical privacy by design?» (Richards and King 2013: 42). Why, they argue, «is the Big Data revolution occurring mostly in secret?» (Richards and King 2013: 43).

While the authors acknowledge the need for trade secrets and the like, data collected from and used to make decisions about and on behalf of individuals merit the development of proper technical, commercial, ethical and legal safeguards. «We cannot have a system, or even the appearance of a system, where surveillance is secret, or where decisions are made about individuals by a Kafkaesque system of opaque and unreviewable decision-makers» (Richards and King 2013: 43).

2. *Identity*: Personalised services, exquisitely tailored to our individual tastes, needs and desires are a hallmark of the Big Data paradigm. Amazon leverages our browsing and purchasing history to group us with likeminded customers and provide us with customised shopping experiences. However, as these services gather information to identify «our true selves», there is a risk that the information is used to nudge us in a certain direction, different from where we would go if we were not under such influence. Google users, the authors argue, are «already influenced by big-data-fed feedback loops from Google's tailored search results, which risk producing individual and collective echo chambers of thought» (Richards and King 2013: 44). As Big Data actors leverage various data sources to identify «us», our right to define our own identity may be threatened. And without proper protections and safeguards against processes that minutely, incrementally and systematically undermine our intellectual choices, the authors argue «'you are' and 'you will like' risk becoming 'you cannot' and 'you will not'» (Richards and King 2013: 44).
3. *Power*: Enthusiasts often claim that Big Data will entail more transparency. Through the proper utilisation of new data streams, we are better placed than ever to shine a light on hidden processes and mechanisms – insight which in turn will allow us to generate an «X-ray» of the fabric of our society.

> However, as the authors point out, the tools and knowledge to wield these data streams, and to make inferences and decisions based on them, are currently in the hands of specialised intermediaries. They are not in the hands of the people who generate the data. Without a proper discourse around these challenges, the authors warn that this power asymmetry may result in «an uneasy, uncertain state of affairs that is not healthy for anyone and leaves individual rights eroded and our democracy diminished» (Richards and King 2013: 45).

The paradoxes framed around transparency, identity and power touch on more than one raw nerve in the current discourse on the ethical and societal implications of Big Data. A closer look at the various elements along the «Big Data chain» – namely data collection and storage, the application of analytical tools and finally action on the basis of insights mined–also reveals a host of potential shortcomings in current protective measures, as well as new challenges and problems.

Whose data is it anyway?

To date, most of the ethical concerns that have been raised relate to privacy challenges in the first link in the chain, namely that of the collection and storage of data. Many of these problems are not entirely new, but traditional mechanisms for ensuring privacy protection have come under increasing pressure with the advent of Big Data. Let us consider two cases:

1. The system of «notice and consent», whereby individuals are given the choice to opt out of sharing their personal data with third parties, has become a favoured mechanism of data protection. In practice, however, the online user is frequently met with lengthy privacy notices written in obscure legal language, where ultimately, she is presented with a binary choice to either accept

the complex set of terms or forsake the service in its entirety. The fatigue and apathy this generates is less than satisfying and it fails to bestow the individual with strong ownership over her data in any meaningful way. The problem is further exacerbated in the Big Data era because it places the onus of evaluating the consequences of data sharing on the individual generating the data. These evaluations can be both technical and complex, and individuals will necessarily be on unequal footing in terms of their ability to make informed choices. Therefore, researchers seeking to leverage e.g. social media data to study social systems cannot assume that they have tacit approval from users of these services – even if the consent agreement provides no legal impediments for such use. The key challenge lies in devising technical and regulatory frameworks that provide the user with tight and meaningful controls on personal data without compromising the practical utility of that same data.

Beyond the mere impracticability of the researcher having to actively seek consent from large swathes of people in all cases, some will argue that giving the data owner an absolute say in if and how her data is used runs the risk of interfering with the innovation potential of data use (Cate and Schönberger 2012; Narayana and Shmatikov 2006). All the possible use cases for a certain type of data (e.g. location data) are rarely apparent at the time of collection, which is when consent is typically sought. By tightly restricting the use of that data to certain predefined use cases, it is therefore likely that many useful applications we enjoy today, such as tracking tools that monitor traffic jams using cell phone movement, would never see the light of day (Sandberg 2012).

2. Another favoured privacy protecting measure is to anonymise datasets by stripping them of personally identifiable information before they are made available for analysis. While such

techniques might be privacy preserving when the dataset is treated in isolation, anonymised datasets have sometimes been shown to be easily de-anonymised when combined with other sources of information.

As part of a contest to improve its movie recommendation service, the online movie streaming service Netflix released an anonymised dataset containing the rental and rating history of almost half a million customers. By running the anonymised dataset against ratings on the online service «Internet Movie Database», researchers were not only able to identify individuals in Netflix's records, but also the political preferences of those people (Narayana and Shmatikov 2006).

There are similar examples of how an apparently anonymised dataset, when properly contextualised, is no longer truly anonymous. And the Big Data paradigm makes it increasingly more difficult to secure anonymity, because ever more data streams are generated, stored and made available for advanced data mining techniques (Navetta 2013). As the world becomes more data rich, researchers can no longer rest content with simplistic anonymisation to mitigate ethical risks.

Collect first, ask questions later ...

The re-identification problem does not only highlight the shortcomings of established protective measures, but also shows that focusing exclusively on the proper governance of datasets and their attributes will often fall short of capturing the nuanced ethical challenges associated with data analysis. In order to grab the bull by the horns and provide the individual with meaningful control over personal information, it is necessary to govern *data usage* – that is, the actual operations performed on and with the datasets – rather than focusing solely on the collection and retention of such data. Doing so, however, is challenging, to say the least.

For while the current Big Data scene may appear to be dominated by a handful of major players, such as Google, Facebook and Amazon, its ecosystem is in fact highly distributed, with a host of third party actors operating behind the scenes which «often piggyback on the infrastructure built by the giants» (Sandberg 2012). Data collection, curation and analysis do not necessarily take place at a single point which can be subjected to robust regulatory measures.

Moreover, the technical opacity of algorithms underpinning Big Data analysis, as well as the real-time nature of such analyses, does not easily lend itself to meaningful scrutiny by way of traditional transparency and oversight mechanisms. In a world where

> … highly detailed research datasets are expected to be shared and re-used, linked and analysed, for knowledge that may or may not benefit the subjects, and all manner of information exploited for commercial gain, seemingly without limit. (Dwork 2014)

it can be hard to gauge *a priori* which operations are socially and ethically sound and which are not. Researchers may find that seemingly innocuous operations reveal themselves as privacy-intrusive or otherwise ethically «sticky» only after they have been performed. As computational techniques become increasingly more sophisticated and systems are able to extract personal information from datasets that appear harmless, relying on human intuition to define which operations are privacy-intrusive and which are not seems unsatisfying.

Can technology help to fix the problems it creates?

Technology is likely to be at least one part of the solution. Novel approaches such as «differential privacy» leverage mathematics to ensure both consistent and high standards for privacy protection in statistical operations on datasets involving sensitive data. Differentially

private algorithms satisfy mathematical conditions that allow the privacy risk involved in an operation on a dataset to be duly quantified. Once a threshold is passed the algorithm will intentionally blur the output so that individuals whose data are being analysed are ensured «plausible deniability». In other words, their presence or absence in the datasets in question has such a marginal impact on the aggregate result that there is no way of telling whether or not they were part of the dataset in the first place. Researchers can still draw value from the dataset because the «blurred» output differs only marginally from true output and the uncertainty, or «degree of blurring», is well known. Differentially private algorithms can also keep track of and appropriately quantify the cumulative privacy risk an individual sustains through repeated or multiple queries by iteratively adding more noise to mask any personal information residing in the data (Klarreich 2012).

Privacy and personal data protection are often touted as the central ethical challenges presented by Big Data. While technology may certainly help mitigate some of these risks, however, other challenges will require strong governance and legal protections. Furthermore, while we attend to the very pressing privacy concerns raised by Big Data, we should not lose sight of the many issues that fall outside the traditional privacy debate.

Looking beyond privacy

With recent technological advances, the cost of collecting, storing and analysing various kinds of data snippets has decreased quite dramatically. Novel methods also allow us to interlink and make sense of various kinds of data long after the data has been collected. As Alistair Croll remarks in an interesting blog-post: «In the old, data-is-scarce model, companies had to decide what to collect first, and then collect it. […] With the new, data-is-abundant model, we collect first and ask

questions later» (Croll 2012). This attitude was perhaps most amply illustrated by the mass surveillance activities of the NSA unravelled in the recent Snowden revelations, but it also holds true on a more general level. And this, Croll argues, is changing the way we use data.

Many remarkable successes of the Big Data paradigm, such as detecting disease outbreaks or predicting traffic jams, come from utilising data in ways that are very different from the original purpose of collection or the original context and meaning we bestowed on the data. However, Croll argues, this is a slippery slope fraught with ethical problems that go well beyond the regular privacy debate. Instead they deal with the inferences we are allowed to make and just how we act on or apply this insight. As the technical and financial barriers to what we can collect and do with data begin to crumble, the regulatory challenges to the proper use of data and analytics are likely to intensify (Soltani 2013).

As an example, Croll remarks on a study performed by the popular online dating service OkCupid, where the profile essays of some half a million users were mined for words that made each racial group in its member database statically distinguishable from other racial groups. According to the OkCupid blog post, «black people are 20 times more likely than everyone else to mention soul food, whereas no foods are distinct for white people» (Rudder 2010). The relatively simple study highlights just how easily information on race, sexual orientation, political standing or health can be inferred from innocuous information collected for very different purposes. Croll continues: «If I collect information on the music you listen to, you might assume I will use that data in order to suggest new songs, or share it with your friends. But instead I could use it to guess at your racial background. And then I could use that data to deny you a loan».

While such inferences may be partially construed as privacy issues–and legislative regulation can assist in preventing obvious transgressions–there are arguably deeper issues at play.

Inferences like the above are typically used to personalise and tailor adds, information, and online experiences to individuals. Such tailoring relies on classification–the algorithmic grouping of data points (people), and, as Dwork and Mulligan point out, such algorithms are a «messy mix of technical and human curating» and are «neither neutral nor objective», but always geared towards a specific purpose in a given context (Dwork and Mulligan 2013: 35). The objectionable or discriminatory outcome may not even be intentional or obvious to the providers of the service. As Sandberg remarks, a machine-learning algorithm trained on various data to determine the suitability of loan applicants, or even job applicants, may well «know» the race or political orientation of the applicant, even if it were not explicitly fed this information or programmed to use it. The information is baked into the data in non-obvious ways and ultimately «the algorithm will follow the data, not how we want to 'think'» (Sandberg 2012). Once differential treatment starts following certain social, political or religious patterns, the large scale effects on society can be profound. As Dwork and Mulligan argue, these issues have little to do with privacy and transparency, but are more about the «values embedded and reflected in classifications, and the roles they play in shaping public and private life» (Dwork and Mulligan 2013: 40).

Some cities across the U.S., notably Philadelphia, use statistical profiling techniques to determine the risk of criminal recidivism among parolees. The method relies on classification of offenders into groups for which certain statistical probabilities can be computed. While critics dismiss such methods as ethically questionable at best (should a cold calculus of past offences punish you for crimes you have not yet committed?), proponents argue that the method is not doing anything a parole board would not do, except with greater accuracy, with full absence of discriminatory urges and with more consistency and transparency.

The case highlights the challenging problems that we are likely to face as Big Data moves out of its nascent stage of tailored ads to affect a wider range of human activity. It also shows how easy it is to fall prey to the temptation of framing the problem as one of man versus machine. Such an approach is likely to be counter-productive. The challenge of managing «ethical risk» in a Big Data world is one that is jointly technological and sociological, and as Dwork and Mulligan succinctly put it: «[…] Big Data debates are ultimately about values first, and about math and machines only second» (Dwork and Mulligan 2013: 38).

Moving forward

Like other technologies in the past, as Big Data unfolds and is absorbed into society we are likely to see adjustments and changes in our current notions of privacy, civil liberties and moral guidelines. And as the hype eventually wears off and a proper equilibrium between the role of human intuition and data-driven insight is established, we will need to develop tools, guidelines and legislation to govern this new world of data and data analysis. This process will require a wide perspective, coupled with constant and close scrutiny of all links along the Big Data chain. It is an arduous task, but also one that should be exciting for all involved. Future societies might be shaped by technological advances, but technology itself is moulded by human choices. And these choices are available for us to make now.

References

Anderson, C. 2008. «The End of Theory: The Data Deluge Makes the Scientific Method Obsolete». Available from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (accessed February 24, 2014).

Australian Government. «The Austrialian Public Service Big Data Strategy». Department of Finance and Deregulation, 2013.

Cate, F.H. and Mayer-Schönberger, V. 2013. Notice and Consent in a World of Big Data. International Data Privacy Law, 3 (2): 67–73.

Croll, A. «Big data is our generation's civil rights issue, and we don't know it.» *radar.oreilly.com*. August 2, 2012. http://radar.oreilly.com/2012/08/big-data-is-our-generations-civil-rights-issue-and-we-dont-know-it.html (accessed February 24, 2014).

Dwork, C. 2014 Differential Privacy: A Cryptographic Approach to Private Data Analysis (private communication).

Dwork, C. and Mulligan, D.K. 2013. It's Not Privacy, and It's Not Fair. *Stanford Law Review Online* 66, no. 35: 35–40.

González-Bailón, S. 2013. Social science in the era of big data. *Policy & Internet* 5: 147–160.

Klarreich, E. 2013. Privacy by the Numbers: A New Approach to Safeguarding Data. http://www.scientificamerican.com/article/privacy-by-the-numbers-a-new-approach-to-safeguarding-data/ (accessed February 24, 2014).

Mayer-Schönberger V. and Cuckier, K. 2013. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt.

Narayana, A. and Shmatikov, V. «How To Break Anonymity of the Netflix Prize Dataset.» *CoRR*, 2006. http://arxiv.org/pdf/cs/0610105.pdf (accessed February 24, 2014).

Navetta, D. 2013. The Privacy Legal Implications of Big Data: A Primer. http://www.infolawgroup.com/2013/02/articles/big-data/the-privacy-legal-implications-of-big-data-a-primer/ (accessed February 24, 2014).

Office of Science and Technology Policy, Executive Office of the President. 2012. «www.whitehouse.gov.» http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf (accessed February 24, 2014).

Richards, N.M. and Jonathan H.K. 2013. Three Paradoxes of Big Data. *Stanford Law Review Online* 66, no. 41: 41–46.

Rudder, C. 2010. The REAL «Stuff White People Like». http://blog.okcupid.com/index.php/the-real-stuff-white-people-like/ (accessed February 24, 2014).

Sandberg, A. 2012. Asking the right questions: big data and civil rights. August 6, 2012. http://blog.practicalethics.ox.ac.uk/2012/08/asking-the-right-questions-big-data-and-civil-rights/ (accessed February 24, 2014).

Social Media and Post-2015. 2013. post2015.unglobalpulse.net.

Soltani, A. 2013. Soaring Surveillance. http://www.technologyreview.com/view/516691/soaring-surveillance/ (accessed February 24, 2014).

Talbot, D. 2013. *Big Data from Cheap Phones*. http://www.technologyreview.com/featuredstory/513721/big-data-from-cheap-phones/ (accessed February 24, 2014).

The Economist. 2013. Predictive Policing: Don't even think about it. http://www.economist.com/news/briefing/21582042-it-getting-easier-foresee-wrongdoing-and-spot-likely-wrongdoers-dont-even-think-about-it (accessed February 24, 2014).

United Nations Global Pulse. www.unglobalpulse.org.

Ward, J.S. and Barker, A. 2013. Undefined By Data: A Survey of Big Data Definitions. http://arxiv.org/abs/1309.5821, (accessed February 24, 2014).