The Values Challenge for Big Data

H. V. Jagadish
Univ. of Michigan

Abstract

As Big Data and analytics defined on top of Big Data have increasingly greater impacts on society, we humans are becoming incorporated in a Big Data loop: our activities, transactions, posts, and images, are all being recorded as Big Data; and in turn the analysis of Big Data is being used to make decisions that affect us. This paper explores characteristics of this grand loop of Big Data and begins the definition of a research agenda to address associated challenges.

1 Introduction

When we think about human in the loop for Big Data, we usually consider a human user of a database system. Most often, our concern is humans querying data in a database: How do they communicate their query intent? How do they make sense of returned results? Sometimes our concerns may go beyond these to humans who have to set up the system: How can one define a good schema? How can one integrate and clean data? How can one wrangle and load data? How can one tune the system for good performance? How usable is the database system [1]?

While all of these are important questions, and definitely worthy of study, the humans we focused on above do not include humans with perhaps the least voice: humans who generate data and are the targets of data analysis. These are human members of society at large. They are the humans in the grand loop of Big Data. Much of Big Data is obtained from their activities: consciously produced, as on social media; generated transactionally, as with credit card transactions; or recorded by third parties, as by a camera on the street. Big Data processing is all about taking these diverse sources of data, and analyzing them to produce value for the processor of Big Data. But this value is realized through advertising to the human, selecting for a job, or providing credit. All of these are activities that impact the human who was the source of the data in the first place. That is the grand loop of Big Data.

Big Data is commonly considered to pose challenges along four axes: Volume, Velocity, Variety, and Veracity. (Some sources consider only the first three of these four). While these are dimensions of technical challenge, there is also a socio-technical challenge, in terms of Values. By this, we do not mean Value (e.g. to the implementor of a Big Data system); rather we mean our Values as a society, and what impact Big Data has on these values. Ultimately, Big Data cannot be successful if this Values challenge is not adequately addressed. As such, this axis is perhaps even more important than the others. Furthermore, humans in the Big Data loop are indeed impacted by Volume, Variety, etc, but humans are centrally in the grand loop of the Big Data Values axis.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

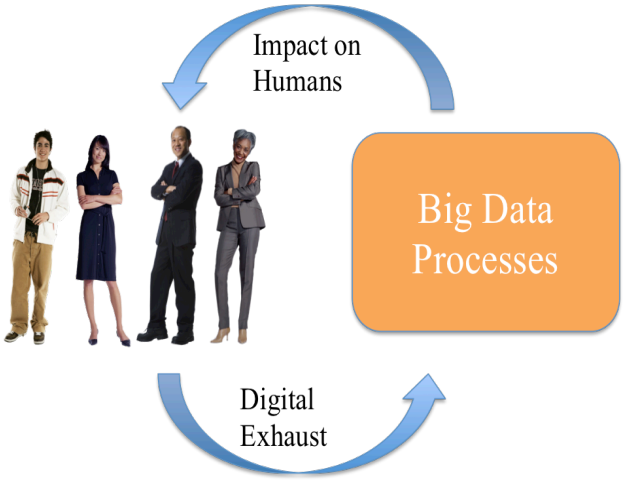

Figure 1: The Grand Loop of Big Data

In this paper, we describe some of the challenges in this grand loop. Naturally, the first challenges that come to mind are those of privacy and anonymity. We consider these in Sections 2 and 3 respectively. Issues of transparency are considered in Section 4. Somewhat surprisingly to many computer scientists, Big Data also has a significant validity challenge. This challenge is considered in Section 5. Many Data Scientists assume that the data are neutral and speak for themselves. In Sections 6 and 7, we discuss how the biases of data scientists can be reflected in the supposedly neutral results . Finally, there is a social impact of our work on Big Data, impact that we cannot ignore, as discussed in Section 8. We finally conclude with a short code of ethics in Section 9.

2 Privacy

Privacy is of course the first thing that anyone thinks of in the context of Big Data. And with good reason. There is no question that the whole point of Big Data is to suck in a great deal of information about us, to put together data from multiple sources, to analyze all this data, and thereby to generate insights that someone can use in their interactions with us: to sell something better or determine credit terms, for example [2].

Unfortunately, there are too many even today, and especially in the tech world, who do not appreciate the importance of privacy. Scott McNealy famously said [3], "You have zero privacy anyway. Get over it." Indeed, there are many who wish fondly for the old days in a small town where everyone knew everyone, and everyone also knew everything about everyone. There was a strong social fabric, and low privacy was a cost paid to achieve that strong social fabric. In big cities, we get privacy, but a much weaker social fabric. However, giving up privacy online does not cause a strong social fabric to be created. We end up with no privacy, and still the same, presumably weak, social fabric.

Why should someone care about privacy if they are not doing anything wrong? Are there things you do that you do not want to discuss with someone close to you, like a spouse or a parent? Imagine how it would feel if you did not have the option of not discussing these things. Even worse, imagine that you do not have the option of not having to explain yourself to everyone in the world. How conformist a life would you be compelled to lead? Privacy is the cornerstone of our freedom to be ourselves [4, 5].

Two other common misconceptions about privacy are worth correcting. Privacy and security are often discussed in the same breath. Those in the know tell us all the time that these are two different things: the former is

a policy question regarding user data, and the latter is an implementation question regarding a computer system. As such, most discussions of security are at the system level, and do not consider data privacy explicitly; and vice versa. However, security breaches can lead to privacy violations. The two concepts, though distinct, are related. A problem that deserves greater attention is how to minimize privacy loss when the inevitable security breach occurs. This is a security-aware direction for data privacy research, focused on issues like data representation, very different from the current body of work on privacy-preserving analysis, which assumes only legitimate access.

A second misconception is that privacy is all or nothing. Once a thought has escaped my brain and stated to a second party, then some may think that ends my privacy rights regarding that thought. But in practice, we do share secrets with confidantes, and we expect them not to share the secrets further. The secret can remain private even after it has been shared within a select group of people. In fact, the right way to think about privacy is as control over information sharing [6]. This motivates the need to control information sharing or at least detect unauthorized sharing. For some simple cases, we have work in digital watermarking. But there is need for much more work to provide the needed assurances.

3 Anonymity

One important use case for sensitive data is to find patterns in the aggregate. A classic case here is medical research: individual patient health records are sensitive, but important progress in life-saving treatments can result from analyzing the aggregates. The traditional solution, even encoded in regulations today, is to use de-identified data. Typically, this is understood to mean removal of several specific sensitive attributes, such as name, address, and social security number. However, it is usually possible to re-identify the patient from the de-identified data with these fields removed. [7]. Similar exploits are well-known in the case of AOL [8] and Netflix [9]. In the medical scenario, astonishingly, genetic sequence data can be retained in the de-identified information. Clearly, a patient can uniquely be identified if we have their entire genome sequence, or even significant parts of it. But for many medical questions, access to the genome is essential. So a stricter rule, prohibiting access to the genome, would prevent addressing many important questions.

Computer scientists have studied privacy-preserving data mining for quite a while now [10]. Techniques such as k-anonymity and l-diversity have even found their way into some curricula. However, these simple techniques do not always work to protect private information. Differential privacy [11] has recently received considerable attention as a technique that can provably obfuscate the presence of any individual record in an aggregated data set. The basic idea is to add a carefully calibrated amount of noise into any release of (aggregated) data. The basic idea works very well, if the aggregate is queried exactly once. The theory gets much more involved, and the practical value diminishes (in terms of more noise to be added to achieve obfuscation to some level) if multiple queries are allowed.

Even with its current limitations, differential privacy works for a common privacy setting with private individual records and public analysis of aggregates. But there are many scenarios where we do need individual information, such as for marketing, for employment and credit decisions, for shipping goods, for providing mobile phone service. Appropriate models for privacy in these arenas are still open. For example, there is now beginning to be good work in obfuscating movement traces with location tracking [12]. But any obfuscation also relies on some model of expected use to ensure minimal loss of value to the analysis due to the noise added.

4 Transparency

How can a decision-maker trust the results of Big Data analysis? There is a long pipeline, with many steps, in many of which some technical expert has made decisions that could impact the final result. Furthermore, some steps may involve manipulations that are just too hard to explain. For example, most deep learning algorithms

are in this class. There is considerable work on algorithmic explanation [13]. There is certainly a great deal of work on data provenance [14]. But what we need is not just an explanation of what happened at any one step. Rather, we need to understand the entire Big Data pipeline. Workflow provenance [15] ideas are a good starting point, but require considerable extensions to be able to work with the Big Data pipeline.

It may be even more important for an affected individual, than the decision-maker, to trust the results of Big Data analysis. Difficulties can arise not only because of errors in the algorithms and data processing, but also because of errors in the source data. An instance of this problem that we have a great deal of experience with, in the US and several other countries, is the notion of a credit score. An analysis conducted by the US government itself, indicates that approximately 20% of us have at least one unwarranted negative 'credit affecting' report on file [16]. In addition, of course, almost all of us have some negative information on file that we wish were not there. To address these difficulties, there is a process, cumbersome and bureaucratic though it may be, to try to correct errors in credit reports. More important, the Fair Credit Reporting Act in the U.S. requires 'lenders', very broadly defined, to provide explanations when they deny credit, employment, or other benefits related to good credit. So lenders, and similar companies, do have systems in place to explain denials, but these systems leave a great deal to be desired, and are a fertile area for further research.

One additional consideration that can complicate matters is the proprietary nature of many models. If a company is too transparent, an adversary may find it easier to reverse engineer their model. This gives rise to the open problem of how to reveal information to customer that they care about without revealing anything (or at least as little as possible) about the model used.

5 Validity

Far too often today, computer scientists, with the tools to hammer away at Big Data, arrive at conclusions that are invalid, for a variety of possible reasons. While few statisticians, at least consciously, will make these mistakes, we see them all the time in our intense desire to get meaning out of Big Data.

A basic problem with much of Big Data analysis is the Streetlight effect [17]. In traditional analysis, data are collected for the explicit purpose of interest. There is a notion of completeness, or sample representativeness. With Big Data, most of the data we work with are repurposed – they may have been created for some purpose, but are used to throw light on some other purpose. For example, tweets have become an important source of data for many analyses (cf. [18]). Yet, we know that people with Twitter accounts are not representative of the population, and even among them, the actual tweets may not be representative sample of what they think. (E.g. I may have a clear preference for a presidential candidate, but may choose not to tweet about it).

Another problem with Big Data is that we often find patterns without understanding why. However, we know that correlation does not imply causality. What this means is that we can have a small change in the environment or the system being characterized, and this can be enough to cause any algorithmic deduction to fail. Perhaps the best known example of this problem is Google Flu [19].

Another potential problem with Big Data is that there often are a very large number of attributes or features that can be derived. Given enough possibilities, there are likely to be a few features that just happen to be correlated with the result we are trying to predict. Statisticians know this as the multiple hypothesis testing problem. Given a finite set of hypotheses, there are statistical techniques to correct for their multiplicity, but these are not often used by Big Data practitioners. Furthermore, large data sets help to ameliorate this problem by reducing the probability of chance correlations. But we often do not have a finite set of hypotheses, or at least not a count of them. The reason is that there are so many different ways in which models can be set up, data can be prepared, and features can be selected. It is commonplace to perform exploratory data analysis first to develop hypotheses, which are then tested on the very same data. Furthermore, these hypotheses, or prediction models, are sometimes iteratively tuned, compounding the difficulty. The consequence is that when we think we have learned something from Big Data, we actually may or may not have learned something that's true.

6 Fairness

People who work with data will often insist that they are just doing what the data tells them: that they are completely neutral and merely exposing the truths in the data they analyzed. Unfortunately, this is never the case. Data scientists make a myriad choices when constructing a model, and these choices can reflect unstated assumptions and unconscious biases even when the data scientist is trying her best to be neutral. Furthermore, there are additional choices to be made in the data collected, the population sampled, the attributes examined, and even the way the results are presented. [20]

There have recently been numerous accounts in the press about data-driven discrimination. [22, 23]. As there is growing recognition of these issues, a small sub-community has developed around issues of algorithmic fairness. There is now an annual workshop series on Fairness and Algorithmic Transparency in Machine Learning [21].

The first question, of course, is how we define fairness. All of us, at one time or another, have complained about how life is unfair. If algorithms are to be fair, we have to capture this subjective notion as a mathematical constraint. One simple notion is to require that algorithms not explicitly consider sensitive attributes, such as race or sex. In fact, there usually are laws preventing such consideration.

However, this notion is far too minimal. Even before we had Big Data, we had the concept of surrogate attributes used for redlining. When lenders were prevented from discriminating on the grounds of race, they began to discriminate based on zip code, which could serve as a proxy for race in a segregated society. With Big Data, similar correlations can become much easier to find, and also much more subtle (for example, involving a combination of correlates, rather than just one). Even when the algorithm-designer has no desire to discriminate, the algorithm may learn correlations, and produce discriminatory results [22, 23].

A commonly adopted notion of fairness, at the individual level, is that any two individuals with similar final outcomes get treated similarly. For example, if we predict something about creditworthiness or employee performance, that these predictions are not systematically biased for particular values of a sensitive attribute, such as race.

Individual fairness is hard to measure, and hence hard to guarantee, even though there is good work on techniques to achieve it (cf. [24]). A much simpler notion of fairness is simply to look at outcomes. In the selected set, e.g. of people considered creditworthy or employable, is the distribution of values for a sensitive attribute, such as race, the same as in the population as a whole? This is the notion we have in mind when we use the term under-represented. The goal of affirmative action is increase the representation of under-represented groups. Algorithms can be designed to accomplish this, if desired.

The prevalent wisdom today is that fairness is a constraint that imposes a cost on algorithm performance. Mathematically, this may appear obvious. While this notion of constraint may be true for group fairness, further thought should make clear that individual fairness should not be a cost: we do want to make the best possible prediction for each individual, without regard to race, sex, etc. If we find that our algorithm output is consistently biased for any value of sensitive attribute, then we should be able to improve the algorithm by stratifying the population based on this sensitive attribute, and separately correcting for the bias in each stratum.

While individual notions of fairness are best dealt with through algorithm design, group notions of fairness are more difficult to handle purely within a scoring and classification algorithm. It would appear that we can profit from using set-oriented ideas commonly used in database management.

7 Diversity

Diversity is much lauded for the benefits it provides, in education settings, in work groups, and elsewhere [25]. If diversity is desired, then it can be worked into a dataset, or the algorithm, as a goal [26]. Just as with fairness, there are many different mathematical definitions of diversity: one has to choose the most appropriate

one for one's particular circumstance. Irrespective of the specific definition chosen, we note that diversity is a group concept: diversity measures apply to a set and not to an individual item. The challenge is that most algorithms are scoring or labeling individuals, so it is not easy to introduce a desired property of the result set when determining score or label for an individual. In fact, even the definition of result *set* is not straightforward: for university admissions it may be the yearly batch, for a company hiring it may be a moving window over some period, and so on.

A completely different aspect of diversity to consider is that human decision-makers are usually diverse, while the same decision-making algorithms may be used widely. Given many independent human decision-makers, not all will make the same decision. For example, given the same set of candidates, different human experts will likely differ on the best choice. Also, given the same circumstances, different judges will likely impose somewhat different sentences for a crime. In contrast, copies of an algorithm will all behave identically. Understanding and managing this lack of diversity is important. In some cases, such as criminal sentencing, the greater standardization may even lead to benefits. However, in many situations we may actually prefer diversity. It is an open research question how to build this diversity into algorithm design. We may be able to draw inspiration from extensive existing work on diverse result sets for queries.

8 Social Impact

We live in a world that evolves. We need our algorithms to evolve with us. Unfortunately, our algorithms are trained on Big Data from the past. (We do not have a choice, since we do not have training data from the future). We use these past data to make decisions about the future. The unstated assumption is that the future will look like the past. When this is not true, dependence on algorithms can lead to *ossification*, where the algorithms, like old curmudgeons, are resisting change. For example, consider a company that puts into place new programs to make their workplace friendlier to women than it was in the past. Such a company is obviously looking to change its employee make up, increasing the number of women employees; yet, its Big Data driven employee success prediction algorithm working on past data, may depress hiring of women by scoring their potential for success in the old company rather than the one with the new programs in place.

Another aspect of Big Data to consider is that addressing Big Data problems requires big resources. In many cases, only big companies may be able to afford these resources. In consequence, the benefits of Big Data may flow only to a few, at the expense of smaller companies and individual citizens. These issues are not well characterized, but underlie much of the worry that many feel about Big Data [27]. But technological progress need not result in such concentration of power. For example, technological innovations such as cloud computing, have made unprecedented computing resources available to all. The field is open for innovation in this dimension. An initial beginning has been made by efforts such as [28].

9 Conclusions

As we consider humans in the Big Data loop, perhaps the most important dimension of Big Data we must address is *Values*. Big Data is a powerful force with huge impacts on society. These impacts can be both good and bad. We need to work to make sure we get as much of the good with as little of the bad as possible.

Towards this end, I have proposed a code of conduct, with two simple rules:

- Do not surprise. Of course, the result of analysis can be surprising. In fact, we often look for surprises. But the process itself should not be a surprise. An individual should not be surprised with the manner of analysis carried out with data about them.
- Own the outcome. As technologists, we cannot have blinders on and disavow responsibility for the data we manage, the results we produce by analyzing this data, or what others do with our Big Data results.

We must teach this code of conduct, and the ethical framework to follow it, to every student of Data Science. To assist the community to do this, I have prepared a modular online course [29] that is available on the web with a Creative Commons license. I encourage you to use material from it in your classes as appropriate.

We must also initiate a body of research into the responsible use of data, and the technologies that assist us in such responsible use. A recent workshop [30] has started to develop a community around this topic.

This work was supported in part by the National Science Foundation, under grant IIS-1250880.

References

[1] H. V. Jagadish, Adriane Chapman, Aaron Elkiss, Magesh Jayapandian, Yunyao Li, Arnab Nandi, Cong Yu. Making database systems usable. *SIGMOD Conference*, Beijing, China, 2007, pp. 13–24.

[2] Bruce Schnier. *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World*. WW Norton and Company, 2015.

[3] Stephen Manes. Private Lives? Not Ours! *PC World* 18 (6): 312, April 18, 2000.

[4] Louis Brandeis, Samuel Warren. The Right To Privacy *Harvard Law Review* IV(5), Dec.15, 1890

[5] Daniel Solove. A Taxonomy of Privacy. 154 *U. Pennsylvania Law Review* 477, 2006.

[6] Helen Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009.

[7] Khaled El Emam, Sam Rodgers, and Bradley Malin. Anonymizing and Sharing Patient-Level Data. *BMJ* 350:h1139, 2015.

[8] Michael Barbaro,Tom Zeller Jr. A Face Is Exposed for AOL Searcher No. 4417749. *The New York Times* Aug. 9, 2006. www.nytimes.com/2006/08/09/technology/09aol.html,

[9] A. Narayanan, V. Shmatikov. Myths and Fallacies of "Personally Identifiable Information". *CACM* 53(6): 24-26, June 2010.

[10] Jaideep Vaidya. *Privacy Preserving Data Mining*. Springer-Verlag Berlin, Heidelberg, 2009.

[11] C. Dwork, A. Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science* 9(3-4):211-407, 2014 .

[12] John Krumm. A Survey of Computational Location Privacy. *Personal and Ubiquitous Computing* 13(6): 391-399, 2009.

[13] Bryce Goodman,Seth Flaxman. European Union regulations on algorithmic decision-making and a right to explanation. https://arxiv.org/pdf/1606.08813.pdf, 31Aug 2016.

[14] James Cheney, Laura Chiticariu, Wang-Chiew Tan. Provenance in Databases: Why, How, and Where. *Foundations and Trends in Databases* 1(4):379-474, 2007.

[15] Susan Davidson, Sarah Cohen-Boulakia, Anat Eyal, Bertram Ludascher, Timothy McPhillips, Shawn Bowers, Manish Kumar Anand, Juliana Freire. Provenance in Scientific Workflow Systems. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 30(4): 44-50, 2007.

[16] Federal Trade Commission. In FTC Study, Five Percent of Consumers Had Errors on Their Credit Reports That Could Result in Less Favorable Terms for Loans. https://www.ftc.gov/news-events/press-releases/2013/02/ftc-study-five-percent-consumers-had-errors-their-credit-reports.

[17] David Freedman. Why Scientific Studies Are So Often Wrong: The Streetlight Effect. *Discover* Magazine, July-August 2010, http://discovermagazine.com/2010/jul-aug/29-why-scientific-studies-often-wrong-streetlight-effect

[18] D Antenucci, MJ Cafarella, MC Levenstein, C Re, M Shapiro. Using Social Media to Measure Labor Market Flows. NBER Working Paper No. 20010, issued Mar 2014. See http://econprediction.eecs.umich.edu/

[19] David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343(6176): 1203-1205, 14 Mar 2014.

[20] Solon Barocas, Andrew D. Selbst Big Data's Disparate Impact 104 *California Law Review* 671, 2016.

[21] Fairness, Accountability, and Transparency in Machine Learning. See http://www.fatml.org.

[22] Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner. Machine Bias. *ProPublica*, May 23, 2016. See https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[23] H. V. Jagadish. Why We are Hard on Amazon and Should Be. Blog post at http://www.bigdatadialog.com/fairness/why-we-are-hard-on-amazon-and-should-be. Aug 19, 2016.

[24] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* pages 214-226, 2012.

[25] Scott Page. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press, 2007.

[26] Marina Drosou, Evaggelia Pitoura. Multiple Radii DisC Diversity: Result Diversification Based on Dissimilarity and Coverage. *ACM Transactions on Database Systems* 40(1): 4, Mar 2015.

[27] Cathy O'Neil. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown, 2016.

[28] Data Science for Social Good. See http://dssg.uchicago.edu/.

[29] H V Jagadish. *Data Science Ethics*. EdX MOOC at https://www.edx.org/course/data-science-ethics-michiganx-ds101x, 2016.

[30] Serge Abiteboul, Gerome Miklau,Julia Stoyanovich, Gerhard Weikum. Data, Responsibly. *Dagstuhl Seminar Proceedings*, vol. 16291, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Germany, 2016.