What Separates Big Data from Lots of Data

Gabriel Jones
Indiana University
107 S Indiana Ave
Bloomington, Indiana, USA 47405
gabejone@indiana.edu

ABSTRACT

We briefly analyze the history of data to show how having *Lots of Data* hardly differs from data storage and analysis in the early days of SQL, or even before computers. We then explain how *Big Data* represents a paradigmatic shift from conventional data analysis. We then begin to look at the potential limits of *Big Data* to assert that this paradigmatic shift does not mean the end of science. We conclude that misunderstanding *Big Data* prevents organizations from capitalizing on its potential and can lead them to spurious answers.

KEYWORDS

i523, hid104, Big Data, Lots of Data, Data Science, Data History, Sociotechnical

1 INTRODUCTION

In 2008, Wired.com published an article titled "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." They tried to assert that *Big Data*, at that point still a relatively new term, was such a revolutionary change that the scientific method would no longer exist.[**?**] Since at least 2008, professionals, scientists, and the public have flocked to the idea of *Big Data*, but many still struggle to understand both its grand potential and its realistic limits.[**?**] On one extreme, boasting terabytes of storage, they claim to be *Big Data* experts but only utilize *Lots of Data*, high quantities of traditional data. On the other extreme, they hyperbolize about how *Big Data* will change the world because it eliminates the need for educated hypothesizing, based on the fallacious assumption that *Big Data* is synonymous with *All Data*, implied by Wired.com and by some academics.[**?**]

To avoid the common data deluge delusions, we borrow from a 2014 article written by Professor Carl Logoze, "Big Data, Data Integrity, and the Fracturing of the Control Zone," that defines *Big Data* as "data that disrupt fundamental notions of integrity and force new ways of thinking and doing to reestablish it."[**?**] This definition both breaks the boundaries of *Lots of Data* and reins in the assumed panacea that leads people to believe they have *All Data*. Taking a brief look at the history of data clarifies what it means to have *Lots of Data*. A case study of the 1880s US Census Bureau demonstrates that mostly just volume and efficiency mark the difference between today's use of *Lots of Data* and the historical use of data in general, and how this differs from the definition and possibilities of *Big Data*. Having separated *Big Data* from an incorrectly limiting category, we then make the case for investigating what are the limits of *Big Data*. We briefly examine the basis for the argument that *Big Data* is not *All Data*, but a more rigorous analysis is beyond our current scope. We break down the first extreme, the synonymizing of *Big Data* with *Lots of Data*, by succinctly explaining how it represents a paradigmatic shift. We also hope to foster additional sociotechnical scholarly discussion and case studies of its limits, which would help break down the hyperbolic synonymizing of *Big Data* with *All Data*.

2 A BRIEF LOOK AT THE HISTORY OF DATA

The human ability to store and analyze data has evolved gradually over millennia. Although digital computer technology greatly accelerated this evolution, most mainstream uses of data still show signs of their historical roots. The formation of early libraries over 4,000 years ago signifies an important moment in methods of amassing data to be organized and processed by humans into knowledge.[**?**] Libraries still have prominence today both in the traditional sense, brick and mortar sites where one can study texts, and in a broader context, digital archives of algorithmically curated information. In either case, libraries are literal representations of information. If one wants to access a text, they can obtain a copy of it, physical or digital, and read the actual words of the text.[**?**] In contrast, another ancient technology, the abacus, demonstrates one of the first symbolic representations of data. The abacus uses an arrangement of beads to represent other numbers and calculations. The numbers themselves did not exist but were symbolically represented. This is an important early prerequisite to the emergence of statistics, which seeks to make accurate claims about a population based only on a sample.[**?**]

One of the first uses of the term business intelligence, a feature of statistical analysis, was used in the *1865 Encyclopaedia of Commercial and Business Anecdotes*. The book described how a banker, Henry Furnese, gained an advantage over competitors by applying a structured method to collect and analyze information relevant to his business activities. Furnese's data analysis is considered one of the first of its kind for commercial purposes. It builds from the fundamental idea that the real world can be represented and analyzed symbolically, as a sample, to produce insights.[**?**] This is the same idea that allows, for instance, modern companies to provide performance bonuses to employees based on how well they meet certain criteria called Key Performance Indicators. It would be impossibly inefficient to have supervisors accurately observe every activity of every employee and objectively judge who made the most contributions, so instead, companies define metrics of good employee behavior and use these metrics to symbolically represent who adds the most value.

While being able to store and analyze data increased in importance near the end of the 19th century, the physical limits of storage and analysis, paper documents and human eyes, created a problem of *Lots of Data*. The US Census Bureau found themselves faced with this problem. As the US population skyrocketed, they estimated

that with late 19th century methods, it would take an estimated 8 years to process the data collected in the 1880 census. Processing the 1890s census data, they predicted, would take over 10 years, so it would not be ready to study until becoming outdated by the 1900 census. The solution came from a young engineer named Herman Hollerith, eventual founder of IBM and creator of the Hollerith Tabulating Machine. His machine mechanically processed punch cards so efficiently it that reduced 10 years of work to three months.[**?**] Thus, he effectively solved the problem of volume, processing data for the entire US population, and of efficiency, since a few machines successfully completed what would have taken countless human hours.

Overcoming the Census Bureau challenge marks a key moment in the history of dealing with *Lots of Data*. With the advent of digital computing and languages like SQL, technologies have continually risen to the ever-greater demands for volume and efficiency.[**?**] But armed with new technologies like web and mobile, society has created new types of relatively easily accessible data.[**?**] The inherently messy, unstructured, rapidly changing nature of this new data goes beyond what an abacus, a library, a Hollerith Machine, or a simple SQL database can handle. In addition to data volume and efficiency, *Big Data* introduces challenges of velocity, the unstable, constantly changing nature, and variety, the unification of datasets as distinct as website-eye mapping and social media network analysis.[**?**] This distinguishes itself from *Lots of Data*, a term whose significance depends mostly on perspective. Processing the census data used to be a challenge of *Lots of Data*, but with modern computing technology, storing and analyzing simple demographic data is relatively straightforward. *Big Data* offers no such historical asymmetry. Even as technology improves its capability of dealing with volume, the other factors that comprise *Big Data* will still pose challenges. In other words, a *Big Data* problem of yesterday is still a *Big Data* problem of today.

3 THE BEGINNING OF A NEW ERA, BUT NOT THE END OF SCIENCE

As the history of data shows, *Big Data* is not just a buzz word. It has real meaning that separates it from past notions of data; it represents a paradigmatic shift in the way we approach the representation and analysis of information, so much so that notions of integrity have been revisited. But this realization can easily be taken too far. In their book, titled *Big Data*, Mayer-Schonberger and Cukier go as far as providing an omniscient mathematical formula for *Big Data*, *(n = all)*, where *n* is the sample size and all is the population. They claim that *Big Data* represents all the data possibly available, with no limits on time, size, or variety, and therefore represents objective, absolute truth. The correlations we derive from *Big Data* therefore do not need proof of causation; the existence of a relationship or pattern in Big Data must be true of reality because *Big Data* is *All Data*.[**?**]

While *Big Data* certainly does change the norms of what it means to prove causation, the *(n = all)* proposition falls short in theory and in practice. Numerous scholars argue that data, no matter what its size and complexity, is a sample, "with bias implicit due to choice of instrumentation, span of observation, units of measurement, and numerous other factors. In essence, *n* never equals all; all is a limit in mathematical terms that can be approached but never attained."[**?**] Ignoring the implicit uncertainty of dealing with a data sample can provide misleading conclusions. "A well-known example of the foibles of the reliance on informally collected data and algorithmic projection is the Google Flu Trends (GFT), which raised huge scientific optimism about the predictive utility of informally collected data when first published in Nature in 2009 (Ginsberg et al., 2009). This optimism suffered a serious setback in 2013 when the GFT predictions for that year were shown to be seriously exaggerated (Butler, 2013; Lazer et al., 2014). A complete accounting for this setback is beyond the scope of this paper. However, one acknowledged factor is an overconfidence in the veracity of the data as a true sample of reality, rather than a random snapshot in time and the result of algorithmic dynamics."[**?**] The grand miscalculations of GFT should not have come as a surprise. Researchers have long-since understood the fallibility of data samples. *Big Data*, while opening up new possibilities for discovery of new questions, still must be held to standards of methodological credibility. Despite the hyperbolic optimism of the 2008 Wired.com article, scientific methods, theories, and ways of thinking will still play an important role in discovery.

4 CONCLUSIONS

As the latest development in the long history of data, *Big Data* represents a paradigmatic shift. *Big Data* clearly distinguishes itself from its predecessors in definition and in possibility. But, despite its tremendous, paradigm-shifting potential, *Big Data* is still an evolution on the long history of symbolic representation. Like any such representation, it shows but a small sample of the real world, viewed through the distorted lens of various biases. Adding the aspects of velocity and variety expand our avenues of discovery, but they do not eliminate the need for establishing some sense of scientific integrity, even if the norms of integrity must adapt. To be fair, the argument against the *(n = all)* proposition comes mostly from the scientific community, whose entire existence relies on integrity. The business world offers a different context. Often, decision-makers must take actions while relying on nothing more than structured but easily fallible methods of analysis.[**?**] They try their best to produce reasonable insights with methods such as SWOT Analysis or Porter's Five Forces, because delivering timely, logical arguments often matters more than taking the time to find answers validated through scientific levels of scrutiny. In business, quickly finding reasonable answers often takes priority to slowly finding proven ones. Given their different priorities, businesses can perhaps afford to relax their standards of information integrity with *Big Data*, as long as they are cognizant of its inherent uncertainty. But it is this lack of cognizance that can lead people into the dangerous territory of making ill-advised decisions based on misleading data. In addition to clarifying what it means to go beyond *Lots of Data* to help people capitalize on the vast potential of *Big Data*, we hope to foster more sociotechnical research into what the dangerous territory of being incognizant looks like and how it can be avoided.

5 BIBTEX ISSUES

Warning–I didn't find a database entry for "Data-History"

Warning–no number and no volume in Keystone

Warning–page numbers missing in both pages and numpages fields in Keystone

(There were 3 warnings)

6 ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

6.1 Uncaught Bibliography Errors

Citations in text showing as [?]: this means either your report.bib is not up-to-date or there is a spelling error in the label of the item you want to cite, either in report.bib or in report.tex

6.2 Formatting

Other formatting issues - missing references section, likely due to encoding issues in report.bib

6.3 Citation Issues and Plagiarism

Need to paraphrase long quotations (whole sentences or longer)

6.4 Structural Issues

Acknowledgement section missing