I/S: A JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY

Lessons from the Avalanche of Numbers: Big Data in Historical Perspective

MEG LETA AMBROSE, JD, PHD*

Abstract: The big data revolution, like many changes associated with technological advancement, is often compared to the industrial revolution to create a frame of reference for its transformative power, or portrayed as altogether new. This article argues that between the industrial revolution and the digital revolution is a more valuable, yet overlooked period: the probabilistic revolution that began with the avalanche of printed numbers between 1820 and 1840. By comparing the many similarities between big data today and the avalanche of numbers in the 1800s, the article situates big data in the early stages of a prolonged transition to a potentially transformative epistemic revolution, like the probabilistic revolution. The widespread changes in and characteristics of a society flooded by data results in a transitional state that creates unique challenges for policy efforts by disrupting foundational principles relied upon for data protection. The potential of a widespread, lengthy transition also places the law in a pivotal position to shape and guide big data-based inquiry through to whatever epistemic shift may lie ahead.

* Assistant Professor, Communication, Culture & Technology, Georgetown University. Many thanks to my Governing Algorithms course and Samantha Fried, whose thesis *Quantify This: Statistics, The State, and Governmentality* (on file at CCT, available upon request by emailing cctprogram@georgetown.edu) directed me to many of the wonderful historical secondary sources in this article. Additional thanks to John Grant at Palantir, Professor Paul Ohm, Solon Barocas and all the wonderful participants at the 2014 Privacy Law Scholars Conference for their insightful comments and suggestions.

Paris Vaudeville, March 16, 1861:

Magis: Statistics, Madame, is a modern and positive science. It sheds light on even the most obscure facts. Thus recently, thanks to laborious research, we now know the exact number of widows that crossed the Pont Neuf during 1860.

Horace, rising: Bah.

Desambois: This is prodigious. And how many?

Magis: Thirteen thousand, four hundred and ninety eight, plus one doubtful case.[1]

Twitter, @BigDataBorat, January 1, 2014:

#BigData2014Predictions #1: Falling #bigdata hype force 3 Vs model to be reduced to 2 Vs model. "Variety" temporary loan to 3-D Printing.[2]

INTRODUCTION

In the spirit of big data, this article is a prediction based on data collected from the past. The data comes from a twenty-year period in the mid-1800s, and the prediction is that data protection has a long, challenging road ahead. Big data is often referred to as a revolution and the article asks and provides insight into just how revolutionary big data may be – and what that means for data protection.

The industrial revolution is the revolution of choice for understanding the transformative nature of the big data revolution.[3] Articles detailing the story of big data begin just after World War II

---

[1] Ian Hacking, *Nineteenth Century Cracks in the Concept of Determinism*, 44:3 J. HIST. IDEAS 455, 473 (1983).

[2] Big Data Borat, Twitter, (Jan. 1, 2014), https://twitter.com/BigDataBorat/status/418473520100683776.

[3] *See e.g.*, VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 182 (2013); Neil Richards & Jonathan King, *Big Data Ethics*, WAKE FOREST L. REV. (forthcoming, 2014).

with the birth of modern computing.[4] The end of the story generally tells us of an impending big data revolution that will change the world and that the law must figure out how to weigh the pros and cons of big data practices.[5]

But, just after the industrial revolution came a revolution that has been overlooked and perhaps is more pertinent: the probabilistic revolution launched by the avalanche of printed numbers that flooded Europe between 1820 and 1840. By focusing solely on the technological transformations at the beginning of each period, an informative epistemic and ethical transformation that occurred between the industrial revolution and the digital revolution has gone unnoticed. Looking at big data through this lens allows us to assess it as an epistemic revolution, as opposed to simply a technological or economic revolution. Comparing the flood of data that washed over society after a technical revolution two hundred years ago to the flood of data we are experiencing today after the computer revolution offers insight into our attempts to govern it.

Taking a closer look at the avalanche of numbers (what I will call first wave big data) reveals remarkable similarities to the second wave of big data occurring today. Between 1820 and 1840, a flood of data from across society became available, aggregated, analyzed, and acted upon. From this period, a series of similarities to big data can be extracted: datafication issues, big data lures, and structural changes. A number of social issues surfaced in the 1800s that have resurfaced today: governability, classification effects, and data-based knowledge.[6] Enthusiasm for big data during both periods was driven by particular lures: standardized sharing, objectivity, control through feedback, enumeration, and the discovery versus production of knowledge.[7] Both periods also experience(d) structural changes: division of data labor, methodological changes, and a displacement of theory.[8]

---

[4] *See e.g.*, Richards & King, supra note 3.

[5] *See e.g.*, Jules Polonetsky & Omer Tene, *Privacy and Big Data Making Ends Meet* 66 STAN. L. REV. ONLINE 25 (2013); Jules Polonetsky & Omer Tene, *Big Data for All: Privacy and User Control in the Age of Analytics,* 11:5 NW. J. TECH. & INTELL. PROP. 239 (2013); Cukier & Mayer-Schönberger, *supra* note 3, at 172-183.

[6] *Infra* Section III (B).

[7] *Infra* Section III (A).

[8] *Infra* Section III (C).

With these similarities established, it appears that big data represents another influx of data following the development of momentous technological power and may serve as the springboard for significant epistemic transformation. Thus, big data is not the revolution - big data is the avalanche of numbers, which was intimately intertwined with the probabilistic revolution. Any proceeding revolution could certainly be more rapid or more sluggish than the course taken by the probabilistic revolution - or may never come to fruition. Because big data's revolutionary status is in question in the article, I will discontinue using the term 'big data revolution' to avoid confusion and reserve the term revolution for either technical revolutions (industrial or computer) or an epistemic revolution (probabilistic). Instead, I will refer to a 'big data transition.' To get a better understanding of the nature of any big data transition underway - and what that means for data protection - requires another comparative exercise: comparing the probabilistic revolution to other scientific revolutions.

First, the structure of a scientific revolution, outlined by Thomas Kuhn,[9] and related concepts are discussed. Then, relying on the similarities extracted in the earlier section, a comparison is made between the probabilistic revolution as a scientific revolution and the big data transition as a scientific revolution. Both are more accurately described as emergent, rather than revolutionary.[10] The challenging characteristics of scientific revolutions - namely value disputes, multiple perspectives, theoretical quarrels, and methodological changes[11] that eventually settle into a new view of the world – are

---

[9] THOMAS KUHN, THE STRUCTURE OF SCIENTIFIC REVOLUTIONS (1962).

[10] *Infra* Section IV.

[11] Silvio Funtowicz and Jerome Revetz coined the term Post-Normal Science in the 1980s to describe new approaches to "wicked problems," characterized by uncertain facts, disputed values, high stakes, and urgent political decisions. Silvio Funtowicz & Jerome Revetz, *Three Types of Risk Assessment and the Emergence of Post-Normal Science*, *in* SOCIAL THEORIES OF RISK 251–274, 253 (Sheldon Krimsky & Dominic Golding, eds., 1992). These are post-normal problems for post-normal science as opposed to normal problems presented by normal science. The means to developing sound scientific answers to wicked problems are an extended peer community, a new quality-oriented reference system, and the consideration of extended facts. This concept is utilized almost exclusively by those working in environmental and ecological issues. The term "post-normal" is intended to address the period of transition between normal sciences. For a discussion on the various ways "post-normal science" has been used and expanded upon see John Turnpenny, Mavis Jones, & Irene Lorenzoni, *Where Now for Post-Normal Science: A Critical Review of its Development, Definitions, and Uses*, 36:3 SCI., TECH., & HUMAN VALUES 287-306 (2011).

present in these emergences, which differ from revolutions because they develop across society over a long period of time.

Data protection in this type of transition is incredibly difficult. Foundational data protection regimes, like the Fair Information Practices Principles (FIPPs), have already begun to feel the strain of this transition. This is because FIPPs protects values and resolved datafication issues by prescribing a process that no longer fits attempts to understand the world, which are presented in various forms from various directions and have yet to find consensus. The law takes on a role of legitimizing practices as they progress without a crystal ball to know what, if any, epistemic revolution lies ahead. In addition to this role, the law may also mitigate the growing pains of this shift by revisiting the datafication issues and lures of big data in light of the second wave of big data changes with the lessons from the first wave of big data in hand.

The article makes three specific contributions and proceeds in five parts. The first contribution is methodological, offering an approach to new socio-technical issues that situates the subject in historical context to assess larger social, ethical, and legal implications. By locating and comparing big data to the avalanche of numbers, I offer a new perspective of the technological practice that has proven difficult to pin down. The second is categorical, providing a set of similarities between big data and the avalanche of numbers and then categorizing both as a pre-epistemic emergence, as compared to a scientific revolution. By providing this type of categorization we may distinguish governance approaches for scientific and technological revolutions, emerging technologies, emerging uses, and emergences, to name only a few potential labels. The third is substantive, suggesting a cyclical relationship between technological innovation, an outpour of data, and epistemic shifts as well as extracting the associated legal and ethical questions.

Section I further outlines the methodology developed to better situate big data in historical context. Section II describes big data and the avalanche of numbers. Section III categorizes their relevant transitional similarities. Section IV compares these transitions to other scientific revolutions. Section V discusses the challenging and opportune role of law and lessons that can be learned from the avalanche of numbers and the probabilistic revolution.

I. METHOD

"History does not repeat, but it does rhyme." – Mark Twain[12]

As socio-technical legal researchers tackling cyberlaw and emerging technology issues across a range of legal contexts, we regularly begin with a socio-technical problem. A new technology or a new social use or unexpected consequence triggers some kind of interesting legal problem that provokes an argument or research question. This functional approach begins by de-black boxing the technology or practice and then describes the social issues using hypotheticals, social science research, or some other evidence that convinces the reader that the problem should raise eyebrows. Next the shortcomings of existing law are outlined followed by a normative proposal, assessing and disregarding those that are legally, technically, or practically unacceptable. The proposals for emerging socio-technical legal issues are often combinations of Lawrence Lessig's four forms of governance (norms, markets, code, and law). Much socio-technical legal research has followed this path to analyze and provide solutions for social and ethical issues arising from big data practices.[13] This article attempts to enrich the beginning and inform the end of these pursuits. Big data is not a technology but is

---

[12] The attribution to Twain has been around since the 1970s but no evidence that he spoke or wrote the words has been revealed. *See* JAMES GEORGE EAYRS, DIPLOMACY AND ITS DISCONTENTS 121 (1971) ("(When Mark Twain declared 'History does not repeat itself, but it rhymes,' he went about as far as he could go) Even the closest historical parallels are only analogues; they are never identical.").

[13] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010); Danielle Keats Citron, *Technological Due Process,* 85 WASH. L. REV. 1249 (2008); Daniel J. Solove, Privacy and Power: Computer Databases and Metaphors for Information Privacy 53 STAN. L. REV. 1393 (2001); Paul M. Schwartz and Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. REV. 1814 (2011); Jane R. Bambauer, *Tragedy of the Data Commons*, 25 HARV. J. LAW & TECH. (2011); Ryan Calo, *Digital Market Manipulation*, 82 GEO. WASH. L. REV. (forthcoming 2014); Omer Tene and Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INT. PROP. 239 (2013); Ira Rubinstein, *Big Data: The End of Privacy or a New Beginning*, 3:2 INT'L DATA PRIVACY L. 74 (2013); Julie Cohen, *What Privacy is For*, 126 HARV. L. REV. 1904 (2013); Lior Strahilevitz, *Toward a Positive Theory of Privacy Law*, 113 HARV. L. REV. 2010 (2013); Jonas Lerman, *Big Data and Its Exclusions*, 66 STAN. L. REV. ONLINE 55 (2013); Woodrow Hartzog and Evan Selinger, *Big Data in Small Hands*, 66 STAN. L. REV. ONLINE 81 (2013); Neil M. Richards, *Three Paradoxes of Big Data*, 66 STAN. L. REV. ONLINE 41 (2013); Mayer-Schönberger and Cukier, *supra* note 3; Richards and King, *supra* note 3; Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum (eds.), PRIVACY BIG DATA, AND THE PUBLIC GOOD (2014).

necessarily technological. It is not an unintended use or consequence of a technology but is a use of technology and may have unintended consequences. In an attempt to better understand big data and its social implications, the article analyzes the socio-technical phenomenon by looking to the past for clarity on how to proceed with governance.

Although "those that cannot remember the past are condemned to repeat it,"[14] Richard Neustadt and Ernest May argue that history can misinform as well as inform policy-makers.[15] The authors, while teaching Decision Making at Harvard's John F. Kennedy School of Government, developed methods for policy-makers to utilize history effectively. While their text *Think in Time* is not necessarily intended to be a scholarly methodology, the exercises are nonetheless utilized in this text. For instance, the authors suggest asking "What's the story?" as opposed to "What's the problem?" and draft a timeline to understand sequences or causation. It also only attempts to reframe some of the issues and ask better questions, as opposed to suggesting specific action based on what should have or could have been done in the past. The comparative exercise is not intended to suggest that big data and the avalanche of numbers are the same, but that big data is new in a way that is similar to the way the avalanche of numbers was new. But, what type of newness are we dealing with? This is a vital question for a field investigating the governance of emerging socio-technical issues. We must assess and predict the nature of technological innovation and integration to choose what we spend time on, determine how those subjects are framed, and craft viable and sustainable governance solutions.

Of course, the article is not intended to be a prediction in any real sense, but a way of situating the data protection policy debate to view the work being undertaken from the forest instead of through the trees. Time periods, innovations, and phenomena are reflected upon by historians who have established markers of what signifies a scientific revolution. Although it would be uncouth for a historian to predict whether a scientific revolution is approaching, legal and technology scholars have no choice but to make predictions about emerging technologies and social change[16] and so, I will perform the

---

[14] George Santayana, *Reason in Common Sense*, *in* THE LIFE OF REASON 284 (1905).

[15] RICHARD E. NEUSTADT AND ERNEST R. MAY, THINKING IN TIME: THE USES OF HISTORY FOR DECISION-MAKERS (1986).

[16] So often we are combatting the complaint that the law cannot keep up with technological change. *See e.g.*, Vivek Wadwa, *Laws and Ethics Can't Keep Pace with Technology*, MIT

exercise. My approach is historical and comparative and loosely resembles comparative-historical research, which is rarely attempted in anything shorter than a book-length analysis and seeks to answer incredibly large-scale questions, like what factors shaped the development of state-regimes in Western Christendom during the eighteenth century.[17]

This article departs from the comparative-historical analysis in numerous ways. The first difference is the purpose of the analysis. Historical-comparative research seeks to "ask questions and formulate puzzles about specific sets of cases that exhibit sufficient similarity to be meaningfully compared with one another."[18] While historical-comparative research does not seek to create general knowledge, and neither does this article, the field does seek to establish causal configurations by comparing similarities and divergences during clearly delineated historical periods.[19] The specific perspective taken to view the historical cases here is similar but intended to inform policy approaches to emerging technology and associated social changes – namely, big data and associated epistemic and ethical issues.

The analysis is streamlined based on this precisely focused goal. This article takes a single case[20] (the avalanche of numbers which

---

TECH. REV. (Apr. 15, 2014), available at http://www.technologyreview.com/ view/ 526401/laws-and-ethics-cant-keep-pace-with-technology/.

[17] THOMAS ERTMAN, BIRTH OF THE LEVIATHAN: BUILDING STATES AND REGIMES IN MEDIEVAL AND EARLY MODERN EUROPE (1997). Parallels outside social scientists actively pursuing historical-comparative methodologies, Tim Wu's *Master Switch* is comparative historical work that loosely resembles this method as well. He identifies a communication innovation cycle (moving from the openness of tinkering in basements and sharing freely to centralized closed systems, which launches another cycle) and in doing so places the newness of the internet in context with prior novel technologies' newness. "Illuminating the past to anticipate the future is the raison d'être of this book." TIM WU, THE MASTER SWITCH: THE RISE AND FALL OF INFORMATION EMPIRES 7 (2010). Tom Standage's *The Victorian Internet* does not identify a cycle but does look back to draw important comparisons between the newness of the telegraph and the newness of the internet. TOM STANDAGE, THE VICTORIAN INTERNET: THE REMARKABLE STORY OF THE TELEGRAPH AND THE NINETEENTH CENTURY'S ON-LINE PIONEERS (2014). Perhaps a book is the only way to perform these types of research endeavors in a satisfactory way, in which case, I consider this article a jumping off point for such a project.

[18] James Mahoney and Dietrich Rueschemeyer, *Comparative Historical Analysis*, *in* COMPARATIVE HISTORICAL ANALYSIS IN THE SOCIAL SCIENCES 8 (James Mahoney and Dietrich Rueschemeyer, eds., 2003).

[19] *Id.*

[20] Looking at a single case can actually have its own benefits. The single time period focused on in this article is made up of numerous cases – a number of countries

occurred across Europe between 1820 and 1840), relying entirely on secondary sources[21] from renowned experts in the field of the history of statistics, and compares it to an ongoing phenomenon (big data). And so, I refer to my comparative conclusions as predictions instead of explanations.

The basic tenets of comparative-historical research (causal analysis, processes over time, and systematic and contextualized comparison[22]) are observed but also manipulated. While I suggest a causal relationship in the process of technological change, a wave of data, and epistemic revolution and draw contextualized comparisons, there are two important caveats. The first is that the wave of data both causes and is caused by the epistemic revolution. A linear causal relationship is not supported by the research. The second is that the goal of the analysis is not to prove such a cycle exists, but to better contextualize and inform the role of law in periods of socio-technical change. To meet that goal, I compare the most related historical process of socio-technical change to one at issue today. The probabilistic revolution (1800-1930) blossomed out of and drove the avalanche of numbers (1820-1840) that followed the industrial revolution (1760-1820); big data has followed the computer revolution with similar pacing. This comparison commences first by sketching out the two periods and then by extracting and organizing relevant changes and characteristics shared by the two periods.

---

experienced the avalanche of numbers and shaped and were shaped by probabilistic thinking. While I reference the two countries with the most useful similarities (and avoid France, as a country with a government stronghold on statistical practices during the time, and the U.S., as an early adopter of the census but a late adopter of probability). While an older wave of big data would certainly be interesting to include, it would be difficult to compare such a case to today. Nonetheless, I continue to search for such a relevant case to incorporate. "[T]he study of single historical cases can do much more than merely generate initial hypotheses. It not only can develop new theoretical ideas, but it can also put them to the test and use the results in the explanation of outcomes. Moving beyond the first case yields often particularly powerful new insights. At the same time, cross-case variation presents difficult methodological problems for macrosocial analysis, both quantitative and qualitative… [T]estable and tested explanatory propositions are not the only gains we can derive from the analysis of a limited number of cases." *Id.*, at 307.

[21] It is common to use a mix of sources in comparative historical analysis, but the research often relies heavily on secondary historical and social sources. Mahoney & Rueschemeyer, *supra* note 18, at 3-24; Edwin Amenta, *What We Know about the Development of Social Policy*, *in* COMPARATIVE HISTORICAL ANALYSIS IN THE SOCIAL SCIENCES 91-123 (James Mahoney and Deitrich Rueschemeyer, eds., 2006). Those scholars working on emerging technology will likely find this deference most appropriate.

[22] MAHONEY & RUESCHEMEYER, *supra* note 18 at 8.

From the historical comparison, further contextual comparison is performed. The epistemic revolution that followed the avalanche of numbers and may follow big data is compared with other scientific revolutions to better understand the role of policy in any potential revolutionary developments. Relying on the criteria laid out by Kuhn and on the assessment by Ian Hacking of the probabilistic revolution, big data is situated similarly to the avalanche of numbers as the early stages in an emergent type of transition, with its own characteristics relevant to policy keeping up with technological change. It is important to note that this is not a legal-historical comparison. Although such a comparison would be incredibly valuable, the role of law during the probabilistic revolution has not been the focus of any legal historians and so, beyond the scope of this article.

After pulling back (twice) to attain this view of big data and the role of law in such a possibly long, likely uncertain transition, the final section poses a number of questions based on the societal challenges faced during the probabilistic revolution that may be pursued without a clear grasp of the benefits and harms of big data.

II. DATA, DATAFICATION & BIG DATA

In this section I offer a definition of "data" as a representation that is mediated through social practices and discuss the phenomenon of "datafication." Then the two waves of big data are described, followed by a section devoted to unpacking their similarities.

A. *Data and Datafication*

Data is defined as "a representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process."[23] This definition of data is fairly modern. It evolved with changes in philosophical theories of knowledge (epistemology) and with developments in the science of statistics. The term "data" was used in the early 17th century to refer to irrefutable or self-evident truths that were beyond the realm of empirical verification.[24] Today the term data continues to carry an air of irrefutability. But rather than refer to axioms, data are now understood to be representations of

---

[23] Ian H. Gould, *I.F.I.P. Guide to Concepts and Terms in Data Processing*, INT'L FEDERATION OF INFO. PROCESSING (1971).

[24] Daniel Rosenberg, *Data Before the Fact*, *in* RAW DATA IS AN OXYMORON 19 (Lisa Gitelman ed.,2013).

small pieces of information observed, sensed, and collected in daily life. The acceptance of data is (and was) the product of surrounding culture and mediated through layers of context, language, and tools.

> To mediate an object, a digital or computational device requires that this object be translated into the digital code that it can understand… [A] computer requires that everything is transformed from the continuous flow of our everyday reality into a grid of numbers that can be stored as a representation of reality which can then be manipulated by algorithms.[25]

According to this definition, data must be communicated or manipulated by a process. According to this definition there is no raw data. "Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care."[26] Perhaps not with care, data is always cooked.

If data is the *representation* of facts in a formalized manner, datafication is the *act of* rendering these representations into a format that can be communicated or manipulated by some process. Datafication can also be explained slightly differently. According to Viktor Mayer-Schoenberger and Kenneth Cukier, "To datafy a phenomenon is to put it in a quantified format so it can be tabulated and analyzed."[27] Datafication does not necessarily need to be quantified - represented as a numerical value - but it does need to be standardized through classification or categorization to be aggregated, processed, and analyzed computationally. "Counting is hungry for categories. Many of the categories we now use to describe people are byproducts of the needs of enumeration."[28] Datafication must be systematic categorization so that its objects can be counted and quantitatively analyzed, but need not be quantitative material.[29] With this terminological foundation in place, I turn to big data.

---

[25] David M. Berry, *The Computational Turn: Thinking About the Digital Humanities*, 12 CULTURE MACHINE (2011).

[26] GEOFFREY BOWKER, MEMORY PRACTICES IN THE SCIENCES 183 (2005).

[27] CUKIER & MAYER-SCHÖNBERGER, *supra* note 3, at 78.

[28] Ian Hacking, *Biopower and the Avalanche of Printed Numbers*, 5 HUMANITIES IN SOC'Y 279, 280 (1982). For instance, cookies may be descriptive, not quantified, but are lines of code that are recognized by a program that can aggregate and analyze the data.

[29] In fact, statistics did not become quantitative until around 1850 or 1860. Theodore Porter, *Lawless Society: Social Science and the Reinterpretation of Statistics in Germany*,

B. *Big Data*

While big data is a new term, a comparison with the avalanche of numbers (the first wave of big data) reveals many similarities that provide insight into the socio-technical changes being experienced today. In this section, two pictures are painted - one of the second wave of big data we are experiencing now and one of the first wave of big data. I will spend significantly more time on the first wave, as it is a less familiar scene, but will expand on both in the sections to follow. The similarities in these pictures, categorized and further developed in Section III, serve as the basis for relying on this time period as a way of framing big data policy action.

1. *Second Wave Big Data*

A great deal of scholarship has described big data. Ira Rubenstein offers a useful and concise description of the phenomenon: "'Big data' refers to novel ways in which organizations including government and businesses, combine diverse digital datasets and then use statistics and other datamining techniques to extract from them both hidden and surprising correlation."[30]

Today, big data is often described in terms of current computing power. It refers to information sources that cannot be processed or analyzed using commonly available tools or techniques and has been characterized as "datasets whose size is beyond the ability of typical database software to capture, store or analyze."[31] This relative definition suggests that big data must always exist - there must always (or often) be some collection of data available that challenges commonly available processing equipment. The three commonly referenced characteristics of big data are relative as well: the three Vs (volume, velocity, and variety). In the big data transition, massive amounts of data are available (*higher* volume) from the number of sensors carried by individuals and objects and those that fill our physical space. Data comes in as a flood (*faster* velocity) across the

---

in THE PROBABILISTIC REVOLUTION: VOL. 1: IDEAS IN HISTORY 352 (Lorenz Kruger, Lorraine J. Daston, & Michael Heidelberger, eds., 1987).

[30] Ira S. Rubenstein, *Big Data: The End of Privacy or a New Beginning?*, 3:2 INT'L DATA PRIVACY L. 74 (2013).

[31] McKinsey Global Institute, "Big Data: The Next Frontier for Innovation, Competition, and Productivity," (May 2011), available at http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

network; it moves as quickly as the connection. And data comes in numerous forms from numerous new sources (*more* variety). Sometimes a fourth V, veracity, is included, which refers to the usability of the data, whether it is contextualized, accurate, up to date, clean, and generally maintained. Another notable attribute should be mentioned as well. Beyond the speed at which big data can move because of the networked world, boyd and Crawford explain that its relationality to other data is one of its main characteristics. "Big Data is fundamentally networked."[32] Data that is collected is intended to be combined with other data and aggregated across a number of contexts, and new ways of linking datasets has continued to generate new insights.[33]

While the attributes of big data are impressive on their own, the big data transition is marked by what can be done with big data. Big data is mined, poked, and prodded to predict what will happen next. Processing vast stores of "historical" data (now virtually real-time data) leads to more accurate models – the time span has shrunk. The real-time speed at which many predictions can be made allows for the integration of predictions into everyday life. Companies, governments, and individuals take actions based on these predictions, not guarantees, which increase the likelihood of a desired result. Netflix predicts our movies, our houses change temperatures, cash registers print us coupons based on the past – these are things we regularly experience in our day to day lives. But more systematic adoption of predictive analytics is happening as well, like state parole boards using analytics to decide which inmates to release based on potential for recidivism,[34] school systems integrating analytics in kindergarten to guide students through high school graduation based on their potential for success,[35] and human resource departments choose and

---

[32] danah boyd & Kate Crawford, *Critical Questions for Big Data*, 15:5 INFO., COMM. & SOC'Y 662, 664 (2012).

[33] Jonathan Shaw, "Why 'Big Data' is a Big Deal," HARV. MAG. 30 (Mar.–Apr., 2014).

[34] "Parole and Technology: Prison Breakthrough," ECONOMIST (Apr. 19, 2014), available at http://www.economist.com/news/united-states/21601009-big-data-can-help-states-decide-whom-release-prison-prison-breakthrough.

[35] Jennifer Zaino, "What Big Data Means for K-12," EDTECH MAG. (Summer 2013), available at http://www.edtechmagazine.com/k12/article/2013/06/what-big-data-means-k-12-0.

track talent to make hiring and firing decisions based on predicted performance.[36]

The innovative tools that have come with big data relate more to improved statistical and computational methods than the steady increase in computing power.[37] Professor Gary King, social statistician at Harvard, uses "big algorithms" to analyze data on a laptop in twenty minutes that would have until recently required a $2 million computer.[38] These tools and techniques are transferrable across disciplines, commercially, publicly, and individually, and to the masses. "There is a movement of quantification rumbling across fields in academia and science, industry and government and nonprofits… It is hard to find an area that hasn't been affected."[39]

The commonly articulated characteristics of big data are relative in nature, suggesting only a difference in degree. But, the big data "revolution" suggests a fundamental change - difference in kind. danah boyd and Kate Crawford explain that big data is about more than scale, "It is a profound change at the levels of epistemology and ethics. It reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality."[40] This was certainly true of the period following the first wave of big data, which is detailed next to reveal remarkable similarities between the current second wave of big data and the avalanche of numbers, suggesting something big is on the way.

2. *First Wave Big Data*

"Almost no domain of human enquiry is left untouched by the events that I call the avalanche of numbers…"[41] Ian Hacking explains

---

[36] Don Peck, "They're Watching You at Work," THE ATLANTIC (Dec. 2013), available at http://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/.

[37] Shaw, *supra* note 33, at 30.

[38] *Id.*

[39] *Id.*, quoting Professor Gary King of Harvard University.

[40] boyd & Crawford, *supra* note 32, at 665.

[41] Ian Hacking, *How Should We do the History of Statistics? in* THE FOUCAULT EFFECT: STUDIES IN GOVERNMENTALITY 181, 189 (Graham Burchell, Colin Gordon, & Peter Miller, eds., 1991).

that the avalanche of printed numbers took place between 1820 and 1840[42] across Europe.[43] When the problems of society became visible and concentrated with urbanization and migration in the nineteenth century, "middle class and aristocratic reformers, philanthropists, professionals, civil servants, and humanitarians, gathered in statistical societies and voluntary associations, set out to do so by means of social surveys and investigations and pressured governments to do likewise."[44] Public agencies responded with official investigations and hearings, collection and dissemination of data, and creating new agencies that could focus on the task.[45]

The systematic study of sickness rates was increasingly pursued after the plague, but was beginning to uncover correlation that provided some rhyme and reason to outbreak.[46] These rates then began to be tied to data about regions and occupations, including daily employment activities supplied by companies like the East India Company.[47] This led to massive collection and reporting on the poor, criminals, mental illness, etc., and ministries of justice data on crime, insanity, prostitution, vagabondage, vagrancy, and suicide were organized and disseminated.[48] Commercial tables of figures were printed and rich education data was collected.[49] The data sources from the avalanche of numbers were quickly combined to find correlations.[50] New bureaucracies and jobs were created to collect data and to organize data-banks.[51] "Disease, madness, and the state of

---

[42] Harald Westergaard called the period between 1830 and 1850 "the first 'Era of Enthusiam'" for statistics. Hacking explains that Westergaard is noticing "the fulfillment of a fetishism for numbering," while he is studying "its inception." Hacking, *Biopower and the Avalanche of Printed Numbers*, *supra* note 28, at 281.

[43] Ian Hacking, *Biopower and The Avalanche of Printed Numbers*, *supra* note 28.

[44] Anthony Oberschall, *Empirical Roots of Social Theory*, *in* THE PROBABILISTIC REVOLUTION: VOLUME 2, IDEAS IN SCIENCE 105-106 (Lorenz Kruger, Gerg Gigerenzer, & Mary S. Morgan, eds., 1987).

[45] *Id.*, at 108.

[46] Hacking, *Biopower and the Avalanche of Printed Numbers, supra* note 28, at 279.

[47] Oberschall, *supra* note 44, at 284-85.

[48] *Id.*, at 286-87.

[49] *Id.*, at 287.

[50] *Id.*

[51] *Id.*, at 285.

the threatening underworld, *les misérables*, created a morbid and fearful fascination for numbers upon which the bureaucracies fed."[52] In other words, 1820-1840 could be labeled as first wave big data. The similarities to today are striking and are described in this section by looking at two countries: Great Britain and Germany.[53]

Systematic statistical practices are traced to 18th century, when Prussian monarchies collected demographic qualitative data on their regions.[54] The initially qualified data was used to make financial and military policy decisions for the state and shifted to quantitative data around the mid-19th century.[55] Statistical analysis and control were taken up by many others beyond the state:

> [T]here were a large number of enthusiasts who measured and counted and put it all down in local, national, and international dailies, weeklies, monthlies, and yearbooks. No travel book was complete without its little summaries of statistics information about the town being traversed.[56]

Prior to 1800, these parties were distinct. The Prussian authorities did not interfere with the statistical work performed by amateurs or universities but they refused, when asked, to release data or findings.[57] Then, in 1805, the Prussian Bureau of Statistics was established as a state operated statistical office that published data and findings to the

---

[52] Hacking, *Biopower and the Avalanche of Printed Numbers, supra* note 28, at 288. Hacking additionally remarks, "*Les misérables* is not merely the title of Hugo's masterpiece, but a standard set of pages in statistical reports, and when the first international scientific congress commenced… *les misérables* was a regular section at which learned papers would be presented.").

[53] While the U.S. census was an inspiration for these countries to implement their own governmental data collection of citizens, the probabilistic revolution did not commence until later in the nineteenth century, once probability had been firmly established as mathematical and practice, as opposed to the study of subjective chance, in Europe.

[54] ALAIN DESROSIERES, THE POLITICS OF LARGE NUMBERS: A HISTORY OF STATISTICAL REASONING 179 (2002).

[55] *Id.*

[56] Ian Hacking, *Prussian Numbers, 1860-1882, in* THE PROBABILISTIC REVOLUTION 377, 378 (Lorenz Kruger, Lorraine J. Daston, & Michael Heidelberger, eds., 1987).

[57] *Id.*, at 377-379.

public. Many of the Bureau's directors were university professors[58] and amateurs continued to publish their statistics as well[59] creating an overlap between the public, the state, and academics that had previously not existed.

In 1860, the Bureau was directed by Ernst Engel who sought complete enumeration and direct collection. This meant that *everyone* should be counted, as opposed to sampling, through a process called individual bulletins, which were sent directly to individuals, rather than being collected by churches or municipal organizers.[60]

Today's parallels to the avalanche of numbers and associated datafication practices are notable, as are the level and nature of criticism and debate in their own time. Germans were reluctant to understand the world through a quantified lens, because such a mechanical perspective "flattened the delicate social contours."[61] They were unconvinced it could capture the essence and quirks of society.

Embraced in a different way, parallels can be found in 19th century England, but with a stronger push from the bottom in a decreasingly top-down world in which a middle class had emerged. Data as a democratic tool was a concept that took hold during this period. As Karl Metz explains:

> Having been an amalgamation of descriptive information on the peculiarities and 'curiosities' of a given country, [statistics] now became synonymous with gathering figures and arranging them to form averages. It was no longer… exclusively the statesman for whom such work was done; the public was also regarded as a legitimate audience for the statistical message.[62]

Industrialization spurred massive migration shifts to urban areas that placed significant strain on infrastructure, leading to new issues

---

[58] DESROSIERES, *supra* note 54, at 180.

[59] *Id.*, at 179.

[60] *Id.*, at 383

[61] Porter, *Lawless Society: Social Science and the Reinterpretation of Statistics in Germany, 1850-1880*, *supra* note 29, 352 (1987).

[62] Karl H. Metz, *Social Thoughts and Social Statistics of the Early Nineteenth Century*, 29:2 INT'L REV. SOC. HIST. 348 (1984).

of disease and sanitation, as well as new forms of work, financial exchange, food importation challenges, and burdens on the poor.[63] People reached for quantification when chaos from a massive shift in the sociotechnical world ensued. The middle class found these social metrics refreshing as objective ways to measure a man, particularly appealing in a time when social mobility had become newly imaginable. Charles Babbage's appeal for a 19th century equivalent of a data center expresses the enthusiasm for universal datafication in 1832:

> Amongst those works of science which are too large and too laborious for individual efforts, and are therefore fit objects to be undertaken by united academies, I wish to point out one which seems eminently necessary at the present time, and which would be of the greatest advantage to all classes of the scientific world. I would propose that its title should be '*The Constants of Nature and of Art.*' It ought to contain all those facts which can be expressed by numbers in various sciences and arts.[64]

Of course, the State was tightly tied to this movement. Rejected in 1753 just prior to the industrial revolution, a bill was passed in 1800 in England that required the regular collection of population data to ensure "certainty [over] conjecture."[65] And, again great effort was put into enumeration, not sampling, in order to heighten accuracy and increase control, but these practices led to less accuracy. The categories chosen, method of collection, level of ability to respond, bias of the designer and collector, language and literacy limitations, etc., all led to inaccuracy with the arrogant assumption of total information awareness, to use a modern term.[66] Through the establishment of the General Register Office in 1837, many of the kinks in data collection were worked out or acknowledged and the

---

[63] DAVID VICTOR GLASS, NUMBERING THE PEOPLE 90 (1973).

[64] IAN HACKING, THE TAMING OF CHANCE 55 (1990) (citing Charles Babbage, "On the Advantage of a Collection of Numbers, to be Entitled the Constants of Nature and Art," V:1 EDINBURGH J. SCI. 334 (July 1831)).

[65] Glass, *supra* note 63, at 91 (quoting Statesman Charles Abbot).

[66] *Id.*

findings and conclusions qualified.[67] Enumeration was thereafter defined as "being in respect of the individuals present on a specified night."[68]

Critics voiced their concerns about these limitations and dangers as datafication enthusiasm swept Britain. "The attitude that counting men instead of weighing them, which was generally ascribed to the statisticians by their critics, was unbearable… Statistics, they claimed, reduced men to averages, to a mean man that only existed in one's fantasy."[69] As early as 1790 Edmund Burke wrote, "But the age of chivalry is gone – That of sophisters, economists, and calculators, has succeeded; and the glory of Europe is extinguished for it."[70] Even in the 19th century, these critics felt the ironic and familiar pressure to "bolster their moral objections with some statistics to support their case."[71] French social economist Jean-Baptiste Say argued that statistics were descriptive only and could not identify underlying laws about the universe.[72] Doctors were particularly opposed to the use of statistical probability, because medicine was founded on the judgment of human physicians dealing with individual patients and her complexities. In 1836 at the Academy of Medicine in Paris, probability in medicine was argued by Risueno d'Amador to be anti-scientific and anti-medical seeking "not to cure this or that disease, but to cure the most possible out of a certain number."[73]

As these two examples reveal, in many ways the big data issues confronting us today confronted Europe in the 1800s. In a landscape where data seemed to be coming from every direction and new sources at tremendous speeds, new techniques for gathering and analyzing data were created to turn the numbers into actionable knowledge. Private data was mixed with public data, commercial with

---

[67] *Id.*, at 94.

[68] *Id.*

[69] Metz, *supra* note 62, at 348.

[70] GERD GIGERENZER, ZENO SWIJTINK, THEODORE PORTER, LORRAINE DASTON, JOHN BEATTY, & LORENZ KRUGER, THE EMPIRE OF CHANCE: HOW PROBABILITY CHANGED SCIENCE AND EVERYDAY LIFE 37 (1990) (citing EDMUND BURKE, REFLECTIONS ON THE REVOLUTION IN FRANCE 113 (1790)).

[71] Metz, *supra* 62, at 341.

[72] GERD GIGERENZER, ET AL., *supra* note 70, at 46.

[73] *Id.*

government, health with education. Governments have never been the only entity numbering people. Companies, organizations, and individuals were important players in the first wave of big data.

Beyond their use by government to structure Britain and Germany (as well as other European countries and the United States), the numbers that fell in the avalanche of numbers were associated with a shift in science that reached nearly every field of study. A number of fields borrowed, commented on, and refined techniques from other disciplines during this transitive period.

One strand of probability theory served as the calculus by which an individual could and should form beliefs and direction under uncertainty, and another used by non-experimental natural scientists used it to establish best estimates and to minimize human error in measurements. Although the first strand of probability theory seems more naturally suited for a human science, it was the latter that took root in the social sciences in the form of social physics between 1830 and 1860. A good part of the nineteenth century was consumed by social scientists attempting to discover "the laws which govern men's habits and the principles of human nature, upon which the structure of society and its movement depends," explained Lord Brougham at a meeting of the Social Science Association in 1857.[74] Also called social physics by Adolphe Quetelet,[75] analogies between man and machine ran wild,[76] but were not without their critics, and quickly gave way to the similar and more popular ideas of Emile Durkheim, who continued to seek social laws through statistics.[77] "The world, it was said, might often look haphazard, but only because we do not know the inevitable workings of its inner springs. As for probabilities – whose mathematics was called the doctrine of chances – they were merely the defective but necessary tools of people who know too little."[78]

---

[74] Oberschall, *supra* note 44, at 106, citing quote from Stephen Cole, *Continuity in Institutionalization in Science*, in THE ESTABLISHMENT OF EMPIRICAL SOCIOLOGY 75 (Anthony Oberschall ed.,1972).

[75] Originally an astronomer, Quetelet makes numerous references to "celestial mechanics" and applied these ideas to the early have of the nineteenth century. Hacking, *Nineteenth-Century Cracks in Concept of Determinism*, *supra* note 1, at 470.

[76] *Id.*

[77] Oberschall, *supra* note 44, at 112-124, quote at 124.

[78] HACKING, THE TAMING OF CHANCE, *supra* note 64, at 1.

Despite fervent opposition from social scientists like Auguste Comte, who argued statistics smeared together too many aspects of society in transition and was annoyed by the appropriation of his term "social physics" instead coining the new term "sociology" to distance himself, statistics and probabilistic modeling became an integral part of sociology.[79] Debates within the social sciences about the appropriate use, meaning, and implications of utilizing large datasets, statistical analysis, and probabilistic theory were wide ranging. The regularity of crime statistics sparked a debate about whether responsibility lay on a small subset of the population (the "malevolence of certain individuals") or the community for the existence of a small general penchant for crime in each of us ("the average man").[80]

While social sciences adhered to a severe strand of determinism, it was highly influential to physics which was undoubtedly indeterministic by the twentieth century. Quetelet believed "The legislator must not seek to block the historical path of the social body, but he can hope to avoid the perturbations to which it is subject. It is the task of social physics to identify each force of perturbations, so that it can be nullified an equal and opposite force."[81] Quetelet's shifting interest from astronomy to sociology when he encountered anthropometric data that appeared to show distribution similar to the error curve. In this similarity, he saw nature aiming at a true value, but accidental influences had produced inaccuracies and errors. The same error curve was found in marriage, suicide, crime, etc., and Quetelet saw the same attempt at natural truth. Although he adhered to an eighteenth century sense of the regularities his techniques revealed, probabilities regularities, as understood in the nineteenth century, were quickly and effectively incorporated by physicists. Moral statistic had developed the building blocks for modeling higher level order from lower level processes. The kinetic theory of gases and statistical physics, and the field would progress to become decidedly indeterministic by the twentieth century.

Similarly, Francis Galton, who pioneered heredity and eugenics, also saw something special in Quetelet's application of the error curve, and it became the basis for his work on understanding variation in and transmission of individual characteristics within different

---

[79] GERD GIGERENZER, ET AL., *supra* note 70, at 46.

[80] *Id.*, at 47.

[81] *Id.*, at 43.

populations. Galton progressed along these lines to develop correlation and regression – important tools for sociologists. Of course, eventually the question of why deviations from the mean in moral statistics should be considered errors in the same way they are in observations of celestial bodies was posed to sociologists, conceptual divides between observations versus statistics and means versus deviation were presented, and distance grew between the fields again. Numerous fields contributed to the development of statistical methods and probabilistic ideas: for instance, biology studies launched the analysis of correlation, educational psychology launched analysis of factors, and eugenics and agronomy launched analysis of variance.

This may imply a kind of comradery across disciplines, but some, discussed below,[82] took issue with the application of their approaches and theories to other fields and some disciplines simply were not intellectually positioned to incorporate techniques or ideas until later in the century.[83] Still, all fields of study dealt with and handled the idea of probability, chance, determinism, and errors differently - some rejected it, some debated and split, and some became fully probabilistic.[84] Although the fields that became statistical in the nineteenth century were those that dealt with a single entity or phenomena that were overwhelmingly large or complex, probabilistic thinking shook all fields of inquiry and society in a way that has shaped our modern world.[85] We are now capable of seeing the world through probabilities: climate change, cancer, robberies, employment, sports, and dating all may be understood in probabilistic terms.

This capability is a product of the probabilistic revolution, and these first wave big data shifts were part of the probabilistic revolution, spanning 1840-1930. Prior to 1840, theories of probability were used to "manage the imperfections of human observation and reasoning."[86] While once only discussed in terms of gambling and

---

[82] *See e.g.,* Poisson, *infra* note 235, 236.

[83] STEPHEN M. STIGLER, THE HISTORY OF STATISTICS: THE MEASUREMENT OF UNCERTAINTY UNTIL 1900 2 (1991).

[84] *See* e.g., Lorenz Kruger, Gerd Gigerenzer, & Mary S. Morgan (eds.), THE PROBABILISTIC REVOLUTION: VOL. 2: IDEAS IN THE SCIENCE (1990), for a series on the various handling of probability by different fields between 1800-1920.

[85] *Id.*

[86] Theodore Porter, *Statistics and Physical Theories*, *in* THE CAMBRIDGE HISTORY OF SCIENCE, VOL. 5, THE MODERN PHYSICAL AND MATHEMATICAL SCIENCES 488 (Mary Jo Nye ed., 2003).

luck, Jakob Bernoulli (seeking to model the reasonableness and good sense of an "impartial judge") and his nephew Nicholas Bernoulli (seeking to economic self-interested "canny merchant")[87] turned their attention to more serious subject matters like wine, futures, annuities, insurance, dowry funds, and returns from mills – none of which were structured mathematically – to find broader applicability and more legitimacy for the study of probabilities.[88] Theories developed and refined by Bayes, Price, and Laplace over the eighteenth and early nineteenth century were highly contested; they often were understood as arguments against God, moral agency, and scientific certainty. Thomas Bayes sought to quantify what could be learned when starting with a guess or belief and updating it as more information became available in the early 1700s. Richard Price refined, disseminated, and found use for Bayes theorem in the mid-1700s, which was further refined and developed by Pierre-Simon Laplace in the late 1700s. As the ideas central to and related to probabilities took shape between these extraordinary men, as well as others, probability was associated with human belief, imperfection, subjectivity, and imprecision. It was therefore, not easily understood as a serious scientific pursuit. However, as attempts to model the "reasonable man" (which included asking well respected gentlemen what they would do in a given situation and comparing mathematical results to confirm their legitimacy) gave way to understanding large scale populations and systems, probability began to gain popularity and acceptability.[89]

Laplace was certainly a big data scientist. He was increasingly frustrated by inaccurate, imprecise, or conflicting data and sought to refine the data in his sets by aggregating large data sets from different time periods, sources, and locations.[90] He sought maximum objectivity so as to calculate and minimize human subjectivity. By the beginning of the nineteenth century it was widely held that the world was ruled by the same mechanics of physics and represented by numerical constants.[91] The early part of the period was dominated by

---

[87] GERD GIGERENZER, ET AL., *supra* note 70, at 15.

[88] *Id.*, at 20.

[89] *Id.*, at 4, 14, 31-32, 37.

[90] SHARON BERTSCHE MCGRAYNE, THE THEORY THAT WOULDN'T DIE: HOW BAYES' RULE CRACKED THE ENIGMA CODE, HUNTED DOWN RUSSIAN SUBMARINES, AND EMERGED TRIUMPHANT FROM TWO CENTURIES OF CONTROVERSY (2012).

[91] IAN HACKING, THE EMERGENCE OF PROBABILITY (1984).

Pierre Simon de Laplace's dictum, "All events, even those which by their insignificance, seem not to follow from the great laws of the universe, follow from them just as necessarily as the revolutions of the sun."[92] Statistics were valued for their potential to establish and support social order. "Ironically, the great improvement in accuracy of demographic, economic, anthropometric, and social records early in the nineteenth century reduced dependence on probability. The difference between statistics and political arithmetic, according to a common nineteenth-century view, was precisely that the former had been freed from reliance on estimates and conjectures, and could aspire to near-perfect accuracy."[93] Previously associated with fate or luck, by the eighteenth century chance had come to be understood as the absence of divine design and in the nineteenth century as lack of control or knowledge. By the early 1900s, statistical practices and theory developed around the idea of probabilities and indeterminacy took hold. The probabilistic revolution produced a post-1930s world "run at best by laws of chance."[94]

This is not to suggest that the landscape during the first wave big of data is identical to the one we have today. While the landscapes are significantly different, a comparison allows us to extract the issues that concerned the population two hundred years ago to understand what questions are (1) not new, but inherent, to datafication practices, (2) accompany waves of big data, and (3) associated with structural changes to data practices.

III. BIG DATA TRANSITIONS

These two time periods share a number of similarities, but one is still in its infancy and the other reflected upon as part of an epistemic and ethical shift that changed the world. In this section, I extract and categorize some of the similarities from the above description of each wave to better assess big data's revolutionary nature and the challenges posed to data protection.

---

[92] PIERRE SIMON DE LAPLACE, A PHILOSOPHICAL ESSAY ON PROBABILITIES 3 (F.W. Truscott and F. L. Emory, trans., 1951).

[93] GERD GIGERENZER, ET AL., *supra* note 70, at 38.

[94] IAN HACKING, *Was There a Probabilistic Revolution 1800-1930?*, *in* THE PROBABILISTIC REVOLUTION: VOL. 1: IDEAS IN HISTORY 45 (Lorenz Kruger, Lorraine J. Daston, & Michael Heidelberger, eds., 1987).

A. *Lures of Big Data*

Datafication's potential for novel, transformative, big data-based knowledge has been greeted with a level of enthusiasm that cannot be overstated. Gary King of the Harvard School of Public Health stated, "People are literally dying every day simply because data are not being shared." The quote reads as extreme but in retrospect, the avalanche of numbers produced nothing less.

> It is certainly not true that most applications of the new statistical knowledge were evil. One may suspect the ideology of the great Victorian social reformers and still grant that their great fight for sanitation, backed by statistical enquiries, was the most important single amelioration of the epoch. Without it most of you would not exist, for your great-great…-grandparents would never have lived to puberty.[95]

When articulated this way, the first wave of big data actually did save lives.

The social gains projected by the second wave of big data rely on universal datafication that is shared and applied without hurdles from the division between disciplines, public and private, commercial and non-commercial, macro and micro. Chris Anderson of Wired writes, "This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology."[96]

Five perceptions and prospects of datafication that promote new potential for progress are shared by both waves of big data. First, data provides a standard, common vocabulary that promotes sharing across disparate and previously siloed parties. Second, data enables numerical control and understanding of systems through feedback. Third, the widely held perception that knowledge expressed and supported numerically possesses a high level of objectivity. Fourth,

---

[95] Ian Hacking, *How Should we do the History of Statistics? supra* note 38, at 184.

[96] Chris Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," WIRED (June 23, 2008), available at http://archive.wired.com/science/discoveries /magazine/16-07/pb_theory. Whether theoretical or ontological underpinnings of knowledge will be considered useful in the future will be discussed in detail in Section III(C)(3).

knowledge is considered out there to be discovered, not something that is produced. And finally, the possibility of enumeration – the idea that so much knowledge can be gained through complete datafication. Although these lures also drove the first wave of big data, they are not inherent questions for datafication. They could be considered common traps that societies experiencing an influx of data fall into, but are not necessarily shared across time and cultures. They are an integral part of understanding the movement toward datafication and its governance.

1. *Standardized Sharing*

Datafication promises to enable sharing of information by expressing objects, characteristics, and phenomena in standard, commonly understood language. Such standardization promotes sharing across disciplines, geographies, industries and institutions. Using the common language of data to describe and analyze occurrences and events, researchers can more easily connect areas of research. Galton saw statistics as a way of connecting fields and applicable to many forms of knowledge, quickly realizing his correlation principle developed as part of his eugenics research could be a method for evaluating interrelation between variables in general.[97] By 1889, correlation was being used in anthropometry, sociology, economics, psychology, and education.[98] Tying fields together with data tested the applicability of statistical methods and theories to a range of areas.[99] For instance, Galton incorporate the data points created by French criminologist Alphonse Bertillon, which included the height and finger, arm, and foot length for criminals, in order to determine the nature of their entanglement with other data.[100]

Today big data enthusiasts are calling for the modern day equivalent of Babbage's '*The Constants of Nature and of Art*,' that would express all facts from the arts and science in numbers. Datafication is understood to make possible new connections between new sources of information and areas of inquiry – connecting data

---

[97] GERD GIGERENZER, ET AL., *supra* note 70, at 58.

[98] *Id.*

[99] *Id.*

[100] *Id.*

collected and maintained offline with that gathered and stored online; location data with purchasing data; government data with commercial data, and data about the humanities with data about the social sciences. Data.gov holds datasets from different agencies posted for public consumption[101] and individuals that monitor their daily health habits as part of the quantified self movement continue push for more ways to access and share their personal data.[102] The World Bank[103] and the United Nations[104] have created platforms for the innovative use of their data. Acxiom collects, stores, processes, and sells information on over 500 million active users with around 1,500 data points for each.[105] Initiatives like university sponsored MOOCs[106] collect and share data on tens of thousands of users to better understand how students learn different concepts and topics and how educational resources should be used, and recently partnered with Google.[107] Commercial markets, government, and universities have embraced the notion that datafication yield new forms of knowledge, creates research efficiencies, and eliminates barriers that currently exist between disciplines when shared.

2. *Feedback and Control*

Representing phenomena through data makes it possible to control a system though a feedback loop. In a feedback loop the output of a system is fed back into the controller which will adjust the

---

[101] "Data Catalog," Data.gov, http://catalog.data.gov/dataset.

[102] Sara M. Watson, "You Are Your Data: And You Should Demand the Right to Use It," SLATE (Nov. 12, 2013), http://www.slate.com/articles/technology/future_tense/2013/11/quantified_self_self_tracking_data_we_need_a_right_to_use_it.html.

[103] "World Bank Open Data: Free and Open Access to Data About Development in Countries Around the Globe," The World Bank, http://data.worldbank.org/.

[104] "United Nations Global Pulse," United Nations, http://www.unglobalpulse.org/.

[105] Natasha Singer, "You For Sale: Mapping, and Sharing, the Consumer Genome," N.Y. TIMES (June 16, 2012), http://www.nytimes.com/2012/06/17/technology/ acxiom-the-quiet-giant-of-consumer-database-marketing.html?pagewanted=all&_r=0.

[106] Massive Open Online Courses (MOOCs) are increasingly being created by universities to expand access to education and advance teaching and learning through practice and research.

[107] "We Are Joining the Open edX platform," Google Research Blog (Sept. 10, 2013), http://googleresearch.blogspot.com/2013/09/we-are-joining-open-edx-platform.html.

system to attain some goal.[108] Feedback loops most relevant to big data are controlled by computational algorithms. Today, datafication practices involve[109] aggregation, classification, association, segmentation, sequencing, and otherwise mining for knowledge that the algorithm assesses input data against and provides an output – and the output data is fed back into the algorithm which will adjust based on the feedback. But, feedback loops need not be a complex computational algorithm.

Although the census projects that popped up across Europe at the beginning of the probabilistic revolution represented an attempt to collect data, analyze it, act upon it, and feed the resulting data back into the resource distribution analysis to make further adjustments, this particular lure gained momentum toward the end of the probabilistic revolution and has not slowed down. In the late 1800s Frederick Winslow Taylor implemented a theory of scientific management with the goal of revolutionizing efficiency and dramatically improving productivity on the factory floor.[110] To do so, he separated tasks performed into discrete, unambiguous elements. Broken down in this way, productivity could be assessed and managed by expert managers continually comparing the maximum possible

---

[108] Feedback loops may best be explained by comparison of open versus closed loop control systems. For example, a clothes dryer that operates by tumbling clothing in a heated drum until the level of moisture measured in the drum reaches a certain level benefits from a feedback loop. The dryer constantly received information about the amount of moisture in the clothing and internally determines whether to continue operating. This is a closed loop control system. An example of an open loop control system is a dryer that runs for the amount of time input into the system by the operator.

[109] The importance of this closed system feedback loop as a form of knowing suggests that simply representing information in a transferrable format is too narrow a concept for big data discourse. Julie Cohen focuses on modulation, defined as "a set of processes in which the quality and content of surveillant attention is continually modified according to the subject's own behavior, sometimes in response to inputs from the subject but according to logics that ultimately are outside the subject's control." Datafication is a necessary step in modulation, but modulation refers to the algorithmic process, whereas datafication refers to the representational process. Reflecting on the history of statistics in *The Politics of Large Numbers*, Alain Desrosieres discusses this difference, "The rationality of a decision, be it individual or collective, is linked to its ability to derive support from things that have a stable meaning, allowing comparisons to be made and equivalences to be established." Alain Desrosieres, *supra* note 54, at 6. For the purposes of big data discussions (both now and from the past), the two are necessarily linked and contribute the growing perception that big data practices are synonyms with modern competitiveness. For the sake of simplicity, we will continue to use the term datafication to embody both the act of forming representations and the processes that follow.

[110] FREDERICK WINSLOW TAYLOR, PRINCIPLES OF SCIENTIFIC MANAGEMENT (1911).

output to the actual output and adjusting the work process accordingly. Planning and administering work processes in this way treated man and machine as one single feedback loop and became known as Taylorism.[111] In order to achieve highly efficient and precise production, factories relied increasingly on mechanical controllers. David Noble explained that this kind of numerical control "allowed management to achieve through mechanical methods objectives that had to that point been dealt with by organizational means."[112] Norbert Wiener expanded the application of feedback and control to all systems as he developed the field of cybernetics. "The language of servo-mechanisms, self-regulation, feedback loops and control loops, it was claimed, applied to all – mechanical, natural, social, human – processes."[113] Man-machine systems were constructed and analyzed as feedback loops, which required human actions and attributes to be machine readable (datafied).

Today, companies are beginning to engage in datafication practices by collecting data from sensors worn by employees that allow them to analyze data to identify problems, efficiencies, and correlations and enable them to make adjustments to their processes and utilize their employees in more effective ways. These practices trace their roots to Taylorism, numerical control, and cybernetics.[114] Just as Taylor intended, the practice is applied far beyond work environments. The quantified self-movement, wherein individuals incorporate technology to monitor their inputs, states, and performance (also explained as "self-knowledge through numbers")[115] can also be understood as individual efforts to establish control through feedback loops. The idea that the world is made up solely of information, as digital physics, for instance, suggests that everything

---

[111] This type of scientific management permeated throughout American culture, but remained more selectively dispersed throughout Europe, where the commitment to technological efficiency and productivity were referred to as "Americanism." Charles S. Maier, *Between Taylorism and Technocracy: European Ideologies and the Vision of Industrial Productivity in the 1920s*, 5:2 J. CONTEMPORARY HIST. 27-61 (1970).

[112] DAVID NOBLE, FORCES OF PRODUCTION 231 (1984).

[113] KEVIN ROBINS & FRANK WEBSTER, TIMES OF THE TECHNOCULTURE 179 (1999).

[114] Rachel Emma Silverson, "Tracking Sensors Invade the Workplace," WALL ST. J. (Mar. 7, 2013), http://online.wsj.com/news/ articles/SB10001424127 887324034804578344303429080678.

[115] Gary Wolf, "Know Thyself: Tracking Every Facet of Life, from Sleep to Mood to Pain, 24/7/365," WIRED (June 22, 2009).

in the universe is nothing but 1s and 0s controlled through a giant feedback loop.[116]

3. *Objectivity*

Related to standardized sharing and closed loop systems is the powerful perception of increased objectivity attendant to datafied understandings of the world. Objectivity is synonymous with realism, referring to one's ability to know things as they actually are, while subjectivity refers to beliefs that exist only in the mind.[117] The statistical societies that popped up in the later half of the nineteenth century were devoted to neutral knowledge, which required routine, mechanical, and precise data practices.[118] The London Statistical Society adhered to the rule that required the exclusion of all opinions.[119] Or as William Farr explained, "The dryer the better. Statistics should be the driest of all reading."[120] The perceived objectivity of the first wave of big data was attractive to societies not only because the early probability debates had left large scale objective statistics the winner, but also because statistical probability represented an alternative was the subjective whim of authority. The lure of objectivity is hard to resist, particularly because of its vital role in representative democracies.

"The American mind seems extremely vulnerable to the belief that any alleged knowledge which can be expressed in figures is in fact as final and exact as the figures in which it is expressed."[121] Mechanical objectivity has particular appeal to the wider public, because it implies following rules - a check on subjectivity.[122] Representative democracies are laden with mechanical objectivity (e.g., judicial impartiality and reliance on precedent), because "[t]hose whose authority is suspect, and who are obliged to deal with an involved and

---

[116] Kevin Kelly, "God Is the Machine" WIRED (Dec. 2002); MICHAEL ELDRED, THE DIGITAL CAST OF BEING (2009).

[117] THEODORE M. PORTER, TRUST IN NUMBERS 3 (1995).

[118] GERD GIGERENZER, ET AL., *supra* note 70, at 38.

[119] *Id.*

[120] *Id.*

[121] RICHARD HOFSTADTER, ANTI-INTELLECTUALISM IN AMERICAN LIFE 339 (1963).

[122] *Id.*, at 4.

suspicious public, are much more likely to make their decisions by the numbers than are those who govern by divine or hereditary right."[123]

Enthusiasm for increased objectivity in education policy, for instance, has roots in curriculum tracking[124] and SATs,[125] but spurred a number of recent efforts including "evidence-based" education policies[126] and education analytics for K-12 schools[127] to support decisions about where and how to spend resources that are beyond political reproach. It is no less true for companies, organizations, and individuals. "Quantification is a way of making decisions without seeming to decide."[128]

4. *Knowledge Discovery versus Production*

A search of Google Books suggests that "knowledge discovery," as a term used outside the education and law, took hold in the late 1970s with datamining. While a new big data-based nature of inquiry is being shaped - where knowledge is discovered, not produced – the avalanche of numbers was associated with the discovery of knowledge as well. The "erosion of determinism" Hacking refers to is the movement from social laws to "the taming of chance,"[129] or an understanding of the world in probabilistic terms that occurred over the long probabilistic revolution. The shift away from determinism

---

[123] Theodore M. Porter, *Objectivity as Standardization: The Rhetoric of Impersonality in Measurement, Statistics, and Cost-Benefit Analysis*, 9 ANNALS OF SCHOLARSHIP 28 (1992).

[124] JEANNIE OAKES, KEEPING TRACK: HOW SCHOOLS STRUCTURE INEQUALITY (1985).

[125] NICHOLAS LEMANN, THE BIG TEST: THE SECRET HISTORY OF THE AMERICAN MERITOCRACY (2000), ("[On February 4, 1945 Henry Chauncey] want[ed] to mount a vast scientific project that [would] categorize, sort, and route the entire population. It [would] be accomplished by administering a series of multiple-choice mental test to everyone, and then by suggesting, on the basis of the scores, what each person's role in society should be." Chauncey would later establish the Educational Testing Service, which develops and administers most standardized tests including the SAT.).

[126] Robert E. Slavin, *Evidence-Based Education Policies: Transforming Educational Practice and Research*, 31:7 EDUC. RES. 15-21 (2002).

[127] *See* e.g., "Analytics at Work in Education, IBM Analytics for Education, available at http://www-01.ibm.com/software/analytics/education/; "Education Analytics for Schools," Microsoft in Education, available at http://www.microsoft.com/education /ww/solutions/Pages/education_analytics.aspx.

[128] PORTER, TRUST IN NUMBERS, *supra* note 117, at 8.

[129] Hacking, THE TAMING OF CHANCE, supra note 59, at 185.

has important implications for our conversation today. Throughout the 1800s, if statistics showed that one in five people committed suicide, the figures were understood to mean that 10% of all people *do* kill themselves. Probability allows us to talk about these numbers in terms of chance and likelihood as opposed to laws and truths.

Today the movement toward universal datafication and knowledge discovery carries the tones of determinism and mathematical realism. Mathematical realism holds that humans do not invent mathematics but discover truth of the mathematical reality. The second wave of big data has revitalized optimism for understanding the laws that govern us as individuals and societies. Social physics, for example, is the term used by Alex Pentland in his 2014 book.[130] Acknowledging that it is an idea over two centuries old, Pentland argues that new tools for collecting and analyzing big data on human interaction can uncover the laws of social physics.[131]

5. *Enumeration*

Enumeration relates to all of the other lures – datafying everything is an important aspect of the drive to objectively share, discover, and control systems, as opposed to sampling or specifying what is and is not part of the inquiry. But enumeration has a distinctive, consistent characteristic that warrants its own category. Prior to the end of the probabilistic revolution, probabilistic inference from samples to population were not known or practiced in social or political statics. During the avalanche of numbers, both Britain and Germany were incredibly optimistic about collecting and processing data across their societies through systematic collection of *all* individuals or processes or transactions. Today, big data efforts attempt to capture *all* voters, *all* drivers, *all* relationships to discover new insights through correlations between previously uncaptured or unaggregated data. It may seem naïve for European countries to have assumed they could capture all their citizens and their needs in a census, but today we are attempting to use big data to map the brain.[132]

---

[130] ALEX PENTLAND, SOCIAL PHYSICS: HOW GOOD IDEAS SPREAD – THE LESSONS FROM A NEW SCIENCE (2014).

[131] *Id.*

[132] J. Nicholas Hoover, "Obama Brain Mapping Project Tests Big Data Limits, INFORMATIONWEEK (Apr. 2, 2013), http://www.informationweek.com/ software/

B. *Datafication Attributes*

Some of the issues that arise during big data periods have nothing to do with the influx of data. These are attributes that come along with datafication no matter its size, speed, or variety. Thus, datafication attributes refer to those questions that are inherent and ever-present and revisited during waves of big data.

1. *Governability*

Datafication makes us governable. It allows us to be both accounted for systematically, offering recognition and protection to those that may have been previously overlooked or underrepresented, as well as controlled and manipulated. Michel Foucault offers two important ideas relevant to the big data debate. First, his work explains that some of the issues are *very* old and the second is the concept he called governmentality (what I will call governability). In his lecture on governmentality, Foucault described the shift away from feudalism, focused on ruling territory prior to the 16th century, toward modern government, focused on ruling people after the 17th century - the development of the "art of government."[133] This 'art' relied on statistics, numbering individuals and understanding the aggregate of a population.[134] Early government utilization of statistical practices in the 16th century developed around health information initially as disease control, attempting to understand spread, death rates, and causes.

During the first wave of big data, the idea that governments could improve health, crime, education, and other important social concerns with data spurred novel collection and analysis efforts. The concerns around the modern quant business mantra "if you can't measure it,

---

information-management/obama-brain-mapping-project-tests-big-data-limits/d/d-id/1109355.

[133] Michel Foucault, *Governmentality*, *in* THE FOUCAULT EFFECT: STUDIES IN GOVERNMENTALITY 87-104 (Graham Burchell, Colin Gordon, & Peter Miller, eds., 1978).

[134] Sam Fried offers a fascinating critique of Foucault's theory of governmentality, arguing that not only is government use of statistics an exercise of control and form of power but the revocation and exclusive use of statistics and quantification can also exert control over and disempower the citizenship. Sam Fried, *Quantify This: Statistics, The State, and Governmentality*, unpublished thesis (2014) (available by request at cctprogram@georgetown.edu).

you can't manage it"[135] are reminiscent of Foucault's concern that recording and measuring human activity after the plague creates both order and control.

2. *Data-Based Knowledge*

The first wave of big data presented opportunities for intervention and thoughtful planning but carried significant, though overlooked, limitations and misunderstandings. To datafy something is to necessarily reduce it to make it useful as standardized and transferable. The limitations in what can be known given that inherent reduction as well as the current tools, techniques, and social foresight at our disposal will always be an issue for datafication.

Germany and Great Britain had different responses to data-based knowledge. Germans found this type of knowledge incredibly limited in its capacity to capture the essence of a person,[136] while the English middle class were enthusiastic about being portrayed by numbers instead of one's lot in life and the neat, orderly world quantification promised.[137] Because of the reductive nature of datafication, data-based knowledge is assessed differently by the cultures that it develops within. It may be accepted as just one way to produce knowledge, the best way to produce knowledge, or even an inherently flawed way to produce knowledge.

Today, cultures around the world are asking questions about what data-based knowledge should mean to them, while others may "miss out on the big data boom" and be further distanced from wealthier nations.[138] For instance, a study on the variation between American and European perceptions of big data found that 43% of Europeans versus 53% of Americans believe data collection can bring the general public more secure and 41% of Europeans believe that big data results

---

[135] Liz Ryan, "'If You Can't Measure It, You Can't Manage It': Not True," FORBES (Feb. 10, 2014), at http://www.forbes.com/sites/lizryan/2014/02/10/if-you-cant-measure-it-you-cant-manage-it-is-bs/.

[136] Porter, *Lawless Society: Social Science and the Reinterpretation of Statistics in Germany, 1850-1880*, *supra* note 29, at 352.

[137] Metz, *supra* note 57, at 337-350.

[138] Mike Orcutt, "Poorer Countries Stand to Miss Out on the Big Data Boom," MIT TECH. REV. (Apr. 25, 2014), http://www.technologyreview.com/ view/526941/poorer-countries-stand-to-miss-out-on-the-big-data-boom/.

in access to lower prices versus 66% of Americans.[139] Recently, there has been a backlash from traditional US news outlets to the big data hype pointing to its many flaws.[140] Reliance on and purposes for data-based knowledge continues to be a question today as new ways of collecting and processing data are developed.

3. *Classification Effects*

Datafication is both natural[141] (because classification based on prior experience is how we make sense of the world[142]) and unnatural (because the way information is represented is necessarily mediated by modern culture and tools[143]). Although intentional abuses by way of data-based social order that rely on classification[144] have certainly occurred, well-meaning yet harmful impacts have also occurred. Caused by the systematic and democratized provision of services through the classification aspect of datafication, these unintended consequences and harms are not easy to identify in their own time.

For instance, recidivism was a concept conceived in the early 1800s when criminal behavior first was studied statistically. But, that simple idea and the probabilities associated with it have had significant impacts on our social structure giving rise to changes in employment practices, penal codes, prison structures, and inter-personal relationships which fundamentally change an individual's life upon conviction. This is not something that has been resolved over the centuries. Danielle Citron analyzed Colorado's public benefits

---

[139] Dan Healy, Raoul Bhavnani, & Arne Koeppel, "Data Privacy in the EU and the US," FTI CONSULTING WHITE Paper (Dec. 2013), http://www.fticonsulting.com/global2/media/collateral/united-states/data-privacy-in-the-eu-and-the-us.pdf.

[140] Woodrow Hartzog & Evan Selinger, "What You Don't Say About Data Can Still Hurt You," FORBES (Nov. 21, 2013), http://www.forbes.com/sites/privacynotice/2013/11/21/what-you-dont-say-about-data-can-still-hurt-you/.

[141] Bowker & Star, *supra* note 26, at 1-32 (introduction entitled "To Classify is Human").

[142] GEORGE LAKOFF, WOMEN, FIRE, AND DANGEROUS THINGS: WHAT CATEGORIES REVEAL ABOUT THE MIND (1987).

[143] David M. Berry, *The Computational Turn: Thinking About the Digital Humanities*, 12 CULTURE MACHINE (2011).

[144] Foucault referred to this as "biopolitics" of the population, which "gave rise to the comprehensive measures, statistical assessments, and interventions aimed at the entire social body or at groups taken as a whole." MICHEL FOUCAULT, THE HISTORY OF SEXUALITY 146 (1978).

system that required eligibility workers to enter whether a potential recipient is a "beggar."[145] Today, anyone with a smart phone or participating online is filtered into classification systems encoded in machine-readable formats. "Classification is the foundation of targeting and tailoring information and experiences to individuals. Big data promises - or threatens - to bring classification to an increasing range of human activity."[146] Classification choices, determinations, and their impacts are inherent in aspects of datafication and are difficult to uncover and understand.

C. *Structural Changes*

Big data-based inquiry is not limited to "scientists." As Section I illustrates, data-based understanding has penetrated nearly every corner of society (except perhaps legal scholarship) from the most micro level to the most macro. Here the structural changes to data labor, method, and theory that occurred and are occurring are discussed.

1. *Division of Data Labor*

A hefty supply of data during the avalanche of numbers was available, in part, because of the division of data labor. Previously, data collection was performed by a priest, a mayor, a lone researcher, or a reformer.[147] By the mid-1800s, statistical societies had formed to pool their resources, like a modern meetup or hackathon. Professional interviewers and tabulators were hired and grants and prizes were awarded for statistical advancements.[148] Soon, the release of statistics to a data-hungry public was delegated to departments within agencies.[149]

Big data-based knowledge is derived from a work flow in which labor is divided amongst a number of parties. In the first wave of big

---

[145] Citron, *supra* note 13, at 1280.

[146] Cynthia Dwork and Deirdre K. Mulligan, *It's Not Privacy, And It's Not Fair*, 66 STAN. L. REV. ONLINE 35 (2013).

[147] Oberschall, *supra* note 44, at 106.

[148] *Id.*

[149] *Id.*, at 107

data, structural changes occurred when the state began to seek data directly from the source (households) by cutting out players in municipalities and parishes that served as middle men. The state created arrangements with corporate and civic organizations, and the data collected was analyzed by various parties and shared with the public. However, this was not necessarily an organized or efficient endeavor. It appears to have been ad hoc and project-specific.

Today, those who collect the data are not those who manage it or those who analyze it. Often analysis is performed on data the user has not directed in any way. "The new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline," explained Jim Gray in 2007.[150] Gray described a project where the designer of the instruments collecting the data is different than the designer of the software who is someone other than the scientist. Collection, curation, management, processing and analysis were previously done by a single team that had control over the direction of each. This statement does not sufficiently communicate the significant distance between parties involved in big data analytics. For some projects, those performing the analysis may have some say in collection, management, and processing by selecting, directing, or designing the tools used for those tasks, but often data is collected by someone, packaged by another, and analyzed by a data scientist across completely different social contexts. There has always been a division between those that discover statistical insights and those that apply them, but the division of labor in big data-based knowledge discovery today changes the role of each laborer involved and the process.

Those who collect will collect as much as possible. Those who package it will package based on their potential users. Those who analyze it will look for anything and everything the data may reveal. Those who use or apply data knowledge are further removed than ever. Standardized sharing and a data market drive this division, and the efficiency of such a system is hard to ignore. In 2009, President Obama signed the Memorandum on Transparency and Open Government,[151] which led to the creation of the data.gov site, where

---

[150] Tony Hey, Stewart Tansley, & Kristen Tolle, *Jim Gray on eScience: A Transformed Scientific Method*, *in* THE FOURTH PARADIGM: DATA-SENSITIVE SCIENTIFIC DISCOVERY xix, xxi (Tony Hey, Stewart Tansley, & Kristen Tolle eds., 2012), (based on the transcript of a talk given by Jim Gray to the NRC-CSTB in Mountain View, California on Jan. 11, 2007).

[151] THE WHITE HOUSE, "MEMORANDUM FOR HEADS OF DEPARTMENTS AND AGENCIES: PRESIDENT'S MEMORANDUM ON TRANSPARENCY AND OPEN GOVERNMENT" (Feb. 24, 2009).

federal agencies are required to deliver data sets.[152] The government, like many online sites and services, is a collector, cooking the data a certain way in hopes that it will be of value to someone downstream.[153] Unlike the government, companies will not openly share the data they collect but sell it to a broker or partner. Data markets have emerged as platforms for exchange between collectors, processors, and users, serving as the middle men that create value through resale and analytic services. Even traditional research universities have seen the value in outsourcing data collection and management[154] asking not "how can I produce this data?" but "where can I find it?" or "what is available?"

2. *Methodological Disruption*

Statistical methods grew in leaps and bounds during the probabilistic revolution: least squares,[155] frequentist statistics,[156] and

---

[152] "Open Data: A History," Data.gov (Apr. 4, 2013), https://www.data.gov/blog/open-data-history.

[153] The operators of these sites, as well as the government, may hire another group to actually collect at their collection points, adding another layer of division.

[154] At my own institution, the creation of the Massive Data Institute has been initiated in the McCourt School of Public Policy to facilitate "an innovative approach to shaping public policy by taking the data generated by government programs as part of their processes, analyzing it and using the research to implement future program planning." "Georgetown Receives $100M to Create New Public Policy School," GEORGETOWN UNIVERSITY (Sept. 18, 2013), http://www.georgetown.edu/mccourt-school-public-policy-announced.html.

[155] Least squares developed to more accurately describe the behavior of celestial bodies and are an approach to establishing an approximation of the best-fit line or curve for a series of data points by squaring the distance from an observed data point and a line on an axis, adding the differences and taking the lowest possible sum to determine the most suitable line. First published by Adriene-Marie Legendre in 1805 (*Nouvelles méthodes pour la détermination des orbites des comètes* or "New Methods for the Determination of the Orbits of Comets") but often attributed to Carl Friedrich Gauss who published on the subject in 1809 (*Theoria motus corporum coelestium in sectionibus conicis solem ambientum* or "Theory of motion of the celestial bodies moving in conic sections around the Sun").

[156] Frequentist statistics or probability describes the likelihood of an event occurring within a random experimentation or sampling. Working from Poisson's 1837 book (RECHERCHES SUR LA PROBABILITÉ DES JUGEMENTS EN MATIÈRE CRIMINELLE ET EN MATIÈRE CIVILE or "Researches into the Probabilities of Judgments in Criminal and Civil Cases," Robert Leslie Ellis (*On the Foundations of the Theory of Probabilities*) and Antoine Augustin Cournot (*Exposition de la théorie des chances et des probabilités*), as well as others published work on this approach in the late 1830s, but frequentist probability was heavily investigated and

correlation[157] were all introduced, refined, and practiced over the 1800s. The law of large numbers, originally proved by Jakob Bertoulli in 1713, was significantly refined and practiced by Siméon Denis Poisson who coined the phrase in 1837. Statistical societies that formed across France, Germany, and the U.S. were most abundant in Great Britain where the International Statistical Congresses met regularly after 1853.[158] These societies as well as the official statistical offices created at the end of the Eighteenth Century sought to craft and required collection and presentation be processed routinely, mechanically, and thoroughly.[159] However, these are not structural changes to methodology. Rather, this was a period where the scientific method taught to elementary school children was refined. The linear scientific method is to:

1. Define a question
2. Gather information and resources (observe)
3. Form an explanatory hypothesis
4. Test the hypothesis by performing an experiment and collecting data in a reproducible manner
5. Analyze the data
6. Interpret the data and draw conclusions that serve as a starting point for new hypothesis
7. Publish results

---

developed by John Venn in 1866 (THE LOGIC OF CHANCE: AN ESSAY ON THE FOUNDATIONS AND PROVINCE OF THE THEORY OF PROBABILITY).

[157] Broadly, correlation is statistical relationship involving dependence. Sir Francis Galton's diligent research on inherited characteristics of peas led to the conceptualization of linear regression (NATURAL INHERITANCE, 5TH ED. (1894)). Later efforts by both Galton and Karl Pearson developed the more sophisticated technique of multiple regression *Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia*, PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOC'Y OF LONDON, 187, 253-318 (1896).

[158] GERD GIGERENZER, ET AL., *supra* note 70, at 38.

[159] *Id.*

8. Retest

In the late 1800s randomization in experimentation and sampling and optimal experiment design for regression were developed around a new understanding of error rates and probability.[160]

In 1874, still prior to a sophisticated and accepted understanding of probability and the quantification of uncertainty, William Stanley Jevons wrote of the inappropriate use of methodology applied to social data:

> No one will be found to deny that there are certain uniformities of thinking and acting which can be detected in reasoning beings, and so far as we detect such laws we successfully apply scientific method. But those who attempt thus to establish social and moral sciences, soon become aware that they are dealing with subjects of enormous complexity. Take, for instance, the science of Political Economy. If a science at all, it must be a mathematical science, because it deals with quantities of commodities. But so soon as we attempt to draw out the equations expressing the laws of variation of demand and supply, we discover that they must have a complexity entirely surpassing our powers of mathematical treatment. We may lay down the general form of the equations, expressing the demand and supply for two or three commodities among two or three trading bodies, but all the functions involved are of so complicated a character that there is not much fear of scientific method making a rapid progress in this direction.[161]

Big data is scary because it has begun to make rapid scientific methodological changes, although not likely in the direction Jevons suggested. In order to grapple with the complexity of systems big data methodology is taking a different shape. Big data provides answers without questions. No hypothesis is needed to test. "Historically, social scientists would plan an experiment, decide what data to collect, and analyze the data. Now the low costs of storage… have caused a

---

[160] HACKING, THE TAMING OF CHANCE, *supra* note 59.

[161] WILLIAM STANLEY JEVONS, THE PRINCIPLES OF SCIENCE: A TREATISE ON LOGIC AND SCIENTIFIC METHOD 2:457-458 (1874).

rethinking, as people 'collect everything and then search for significant patterns in the data.'"[162] As data-mining matured and grew into analytics, it spread well beyond small computer science and computational sub-fields. Now, big data analytics can be used to mine incoming data in real-time to reveal correlations that may or may not be interesting or valuable. It is observational in a sense, but a departure from the concept of knowledge production which required significantly more upfront planning by the researcher.

This collect-all, discover-later approach has been called data dredging, which runs the risk of returning statistical significance by chance. "Big Data enables the practice of apophenia: seeing patterns where none actually exist, simply because massive quantities of data can offer connections that radiate in all directions."[163] Nathan Eagle argues that "No matter how much data exists, researchers still need to ask the right questions to create a hypothesis, design a test, and use the data to determine whether that hypothesis is true."[164] This may be true in a few corners of health research, but even in the high stakes arena of health data, discoveries in drug interactions[165] and genomics[166] have already moved beyond the delineated process. The requisite Google Flu Trend reference may serve as a commercial example:

> Not only was 'Google Flu Trends' quick, accurate and cheap, it was theory-free. Google's engineers didn't bother to develop a hypothesis about what search terms – 'flu symptoms' or 'pharmacies near me' – might be correlated with the spread of the disease itself. The

---

[162] Shaw, *supra* note 30, at 34.

[163] boyd & Crawford, *supra* note 32, at 668.

[164] Shaw, *supra* note 30, at 34.

[165] John Markoff, "Unreported Side Effects of Drugs Are Found Using Internet Search Data, Study Finds," N. Y. TIMES (Mar. 6 2013), available at http:// www.nytimes. com / 2013/ 03/07/science/unreported-side-effects-of-drugs-found-using-internet-data-study-finds.html?_r=1&.

[166] "Big Data Genomics Sequencing," *Intel*, http://www.intel.com/ content/www/us/ en/big-data/renci-peer-story.html.

> Google team just took their top 50 million search terms and let the algorithms do the work.[167]

From hypothesis-free to theory-free, the quote leads to the next structural change.

3. *Displacement of Theory*

Theory during the probabilistic revolution was full of tension between determinism and chance. The trend for those dealing with the numbers was to let "the facts" speak for themselves. The London Statistical Society's motto was *Aliis exterendum* or "to be threshed out by others." It refused to concern itself with cause or effect or chance or why.[168] Quetelet's approach required no knowledge of actual causes but only the identification of regularities and he renounced analysis.[169] A number of theories were articulated by champions of various camps, but slowly debates began to take shape around a few themes, such as the meaning of error, appropriate application of methods to particular problems, and the role of free will.

Not entirely dissimilar to the conversations today about the quality, applicability, and responsibility of the use of big data analytics. In an incredibly controversial 2008 *Wired* article, Chris Anderson wrote,

> [The Petabyte Age] forces us to view data mathematically first and establish a context for it later. For instance, Google conquered the advertising world with nothing more than applied mathematics. It didn't pretend to know anything about the culture and conventions of advertising — it just assumed that better

---

[167] When the Center for Disease Control data slowly came in, Google's prediction had overstated the severity of the outbreak by a factor of two. "The problem was that Google did not know – could not begin to know – what linked the search terms with the spread of flu. Google's engineers weren't trying to figure out what caused what." Tim Hartford, "Big Data: Are We Making a Big Mistake?" FINANCIAL TIMES (Mar. 28, 2014), available at http:// www.ft.com/cms/s/2/ 21a6e7d8-b479-11e3-a09a-00144feabdc0. html#axzz2xL1zS Uh9. Google Flu Trends is consistently used as an example of the power and potential of big data, but its appropriateness of such and accuracy has been questioned and scrutinized. *See* David Lazar, et al., *The Parable of -Google Flu: Traps in Big Data Analysis*, 343: 6176 SCIENCE 1203-05 (Mar., 2013).

[168] GERD GIGERENZER, ET AL., *supra* note 70, at 37-40.

[169] *Id.*, at 40-42.

> data, with better analytical tools, would win the day. And Google was right… With enough data, the numbers speak for themselves. Petabytes allow us to say: 'Correlation is enough.'[170]

danah boyd and Kate Crawford's response is worth citing in full:

> Significantly, Anderson's sweeping dismissal of all other theories and disciplines is a tell: it reveals an arrogant undercurrent in many Big Data debates where all other forms of analysis can be sidelined by production lines of numbers, privileged as having a direct line to raw knowledge. Why people do things, write things, or make things is erased by the sheer volume of numerical repetition and large patterns. This is not a space for reflection or the older forms of intellectual craft.[171]

Not only does a big data-based mindset require a computational perspective that may limit the way the world is explored, it may also cut inquiry short. This is the claim of the end of theory: once a correlation has been discovered, the important work is done. "Hard sciences" use "theory" differently than other disciplines. A scientific theory is a hypothesis supported by repeated testing, which may eventually elevate to a law. In computer science theories represent problems to be solved like randomness, interaction, and non-determinism. Theory beyond these realms is a system of ideas intended to explain some aspect of the world; theory comes first - data second. As computational turns are taken across society, theory may shift to mean something more closely aligned to hard science: data first - theory second.[172]

The role theory will play in data-based knowledge is unclear, but its displacement has already occurred in a number of fields. For instance, an international relations symposium entitled "The 'End of IR Theory?'" was held in November, 2013, where John J.

---

[170] Anderson, *supra* note 96.

[171] boyd & Crawford, *supra* note 32, at 666.

[172] This point is dangerously close to motivating a discussion about inductive versus deductive reasoning, which is beyond the scope of the article. Instead, I simply suggest that where the theoretical work is being done and who is doing it in the process is changing and, at least momentarily, theory may be disappearing.

Mearsheimer and Stephen M. Walt explain that theorists represent the fields most prestigious scholars and still, the discipline has moved away from theory toward efforts "devoted to collecting data and testing empirical propositions."[173] More conclusive is an analysis of 27,000 internet studies social science articles published between 2000 and 2009 that found less than a third referenced a single theoretical source.[174] These examples suggest that theory is currently displaced by big data-based knowledge discovery and that this difference in kind signifies a shift. Understanding the world through datafication is a different process designed to produce a different form of knowledge allowing for governance of different parts of individuals and society.

D. *Differences*

Beyond the similarities associated with increased volume, speed, variety, connectivity, technical and statistical sophistication, and widespread application, there are probably numerous differences between these periods that are far more relevant to the conversation today than the difference between printed and digital numbers. One that is likely relevant is the fact that there appears to have been no market for printed numbers during the avalanche in the 1800s. There were examples of exclusionary measures taken to keep others from data, but no indication of a market or small scale sales. Relatedly, money was not represented or traded using data, so security threats do not appear to have been an issue. Reidentification does not appear to have been a recognized threat – there was likely neither the incentive to reidentify nor the technical or time resources to make it a worthwhile endeavor. These may be couched as proprietary and security issues, which will not be addressed further in any depth, but point to the continued applicability of industrialization as a frame of reference.

Additionally, much of the work published on the meaning and appropriateness of probabilistic approaches were placed in philosophy journals or by men trained in philosophy, if not wholly labeled as

---

[173] Daniel Nexon, "Symposium – Leaving Theory Behind: Why Simplistic Hypothesis Testing is Bad for IR," THE DUCK OF MINERVA (Sept. 7, 2013), http://www.whiteoliphaunt.com /duckofminerva/ 2013/09/leaving-theory-behind-why-simplistic-hypothesis-testing-is-bad-for-ir.html.

[174] Tai-Quan Peng, et al., *Mapping the Landscape of Internet Studies: Text Mining of Social Science Journal Articles 2000-2009*, 15:5 NEW MEDIA & SOCIETY 644-664 (2012).

philosophers – who were also practitioners. Others like Comte, rejected these methods and argued against their incorporation into their disciplines. Today expertise is more heavily divided. Cutting edge data scientists are rarely aware of or arguing the finer points of human autonomy, dignity, or larger social ramifications of datafication. The political climate for reform was strong in Europe over this period, most notably in the United Kingdom, which can be distinguished in many ways from today, but civic unrest certainly permeates modern times, mostly notably since the financial crisis of 2008. Turning back to the similarities shared by the avalanche of numbers and big data, the following section addresses the question of just how revolutionary big data may be and what this revolutionary future means for data protection.

IV. BIG DATA REVOLUTION(S)

"Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means."
- Bertrand Russell, 1929 lecture[175]

"Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it…"
- Dan Ariely, 2013 Facebook post[176]

Again, the comparative exercise is not intended to suggest that big data and the avalanche of numbers are the same, but that big data is new in a way that is similar to the way the avalanche of numbers was new.[177] Determining the type of newness presented by big data is important to its governance, and any other type of emerging socio-technical issue. The probabilistic revolution that took place from 1800-1930 was pressed forward with the avalanche of numbers from the 1820s to 1840s, but it is not considered a "scientific revolution" under prominent criteria. In this section, I compare the probabilistic

---

[175] Quoted in STANLEY SMITH STEVENS, HANDBOOK OF EXPERIMENTAL PSYCHOLOGY 44 (1951).

[176] Dan Ariely, Facebook (Jan. 6, 2013), https://www.facebook.com/dan.ariely/posts/904383595868.

[177] Tom Standage similarly ties the telegraph to the internet in THE VICTORIAN INTERNET: THE REMARKABLE STORY OF THE TELEGRAPH AND THE NINETEENTH CENTURY'S ON-LINE PIONEERS (1989).

revolution to the early evidence of a big data transition so as to better understand its disruptive nature in historical context against markers that serve to signify scientific revolution.

A. *The Revolutionary Status of Probability*

In 1962, Thomas S. Kuhn published his landmark text, *The Structure of Scientific Revolutions*, in which he argued that periods of "normal science" were disrupted by episodes of "revolutionary science." Normal science is cumulative. It is business as usual. Key theories, metaphysics, tools, and values are fixed and research generates puzzle-solutions.[178] Novel, anomalous puzzles that cannot be tackled effectively by normal science incite revolution within the community. When confidence in normal science to solve these anomalies is lost, a crisis arises. During this pre-paradigm phase of crisis, research is carried out in a scientific nature, but methodology, terminology, and types of experimentation are all diverse, innovative, and unorganized without a clear understanding of which are likely to contribute insight relevant to the future. There is little consensus around a particular theory; science is pursued under incomplete, incompatible, and theory-disputed design.

Crisis is followed by resolution and the forming of new paradigms of normal science. Paradigms can be understood as consensus – the chaotic revolutionary science that is produced post-normal science[179] matures and paradigms develop as research traditions and conceptual frameworks begin to form out of the disparate and unorganized intellectual energy. The community is converted to the new paradigm based on its promise for future research, its ability to resolve gaps the old paradigm could not fill, and a novel aesthetic (these converts are often very young or new to the field).[180] New paradigms not only reinterpret old data in new ways, but ask new questions of old data.[181]

---

[178] Kuhn, *supra* note 9, at 35-41.

[179] Silvio Funtowicz and Jerome Revetz coined the term Post-Normal Science in the 1980s to describe new approaches to "wicked problems," characterized by uncertain facts, disputed values, high stakes, and urgent political decisions. Funtowicz & Revetz, *supra* note 12, at 253. These are post-normal problems for post-normal science as opposed to normal problems presented by normal science. The term "post-normal" is intended to address the period of transition between normal sciences. For a discussion on the various ways "post-normal science" has been used and expanded upon see John Turnpenny, Mavis Jones, & Irene Lorenzoni, *supra* note 12, at 287-306.

[180] Kuhn, supra note 9, at 165.

They move beyond the puzzle solving of the old paradigm,[182] change the rules of the research game,[183] and redirect the "map" of new research.[184]

The probabilistic revolution is not considered revolutionary under these criteria.[185] No single anomalous event threw the scientific world into upheaval. No single individual or group can claim credit. It did not develop within a single science to deal with a particular problem that required major theoretical alterations. Instead, it stems from and affects a wide range of fields, with a number of piecemeal contributors, and the new methods that developed seem to be a product of external social force, as opposed to internal specific problem solving.[186] Most importantly, the probabilistic revolution was slow, taking the better part of a century to attain and settle into its paradigmatic state.[187]

Ian Hacking refers to the probabilistic revolution as an "emergence."[188] It shared some of the same phases as scientific revolutions. Theories, methods, and possibilities were shaken up, but very slowly hashed out as the probabilistic paradigm took shape. Hacking explains, "The emergence of probability, however, was a change more fundamental than any revolution. A new thinking cap."[189] This cap is the result of slow, steady progress made through the development of practices surrounding data across a wide range of fields and contexts and decades-long debates about what those developments meant.

---

[181] *Id.*, at 121, 139, 159.

[182] *Id.*, at 36-42, 144.

[183] *Id.*, at 40-41, 175.

[184] *Id.*, at 109-111.

[185] HACKING, THE EMERGENCE OF PROBABILITY, *supra* note 91, at Introduction 2006: Archeology of Probable Reasoning; Hacking, *Was There a Probabilistic Revolution 1800-1930?*, *supra* note 94.

[186] *Id.*

[187] *Id.*

[188] *Id.*

[189] HACKING, THE EMERGENCE OF PROBABILITY, *supra* note 91, at Introduction 2006: The Archeology of Probable Reasoning.

B. *Revolutionary State of Second Wave Big Data*

The second wave of big data similarly resembles the beginning of an emergence, as opposed to a revolution. Like the probabilistic revolution, it is missing a key player or group that discovered an anomaly that requires a complete overhaul of understanding in a field. Nonetheless, the post-normal (the pre-paradigm/crisis phase) characteristics of value disputes, multiple perspectives, and methodological experimentation appear to be underway.[190] Enthusiasm surrounding big data has spread to all corners and fields. Value disputes about big data range from privacy concerns to the value of knowledge data practices produce and span across fields, industries, government agencies, communities, and individuals. Each assesses and takes on big data in a different way. The wide reaching enthusiasm relate to the problems that can be solved and answers that can be found using "new science." Old data is being reinterpreted and new questions asked of it.[191]

The extraordinary leaps made in experiment design that developed over the course of the probabilistic revolution were still situated in what we now understand as the classic scientific method, where "[o]bservation is always selective. It needs a chosen object, a definite task, a point of view, a problem."[192] Jim Gray argued that this was part of the second paradigm. The first paradigm was a "thousand years ago and involved describing natural phenomena (empirical). The second from the "last few hundred years" used models and generalization (theoretical). The third occurred over the "last few decades" and involved simulating complex phenomena (computational). These have culminated in the fourth paradigm, which Gray describes as data exploration (or eScience):

> "Originally, there was just experimental science, and then there was theoretical science, with Kepler's Laws, Newton's Laws of Motion, Maxwell's equations, and so on. Then for many problems, the theoretical models grew too complicated to solve analytically, and people

---

[190] FUNTOWICZ & REVETZ, *supra* note 12; TURNPENNY, JONES, & LORENZONI, *supra* note 12.

[191] *See* e.g., Michel Jean-Baptiste, et al., *Quantitative Analysis of Culture Using Millions of Digitized Books*, 331: 6014 SCI. 176-182 (Jan. 2011); Rosenberg, *supra* note 21 (discussing the use of google book analytics to understand the history of the word 'data.').

[192] HUGH G. GAUCH JR., SCIENTIFIC METHOD IN BRIEF 57 (2012).

> had to start simulating. These simulations have carried us through much of the last half of the last millennium. At this point, the simulations are generating a whole lot of data, along with a huge increase in data from the experimental sciences."[193]

Gray spoke of the fourth paradigm in traditional sciences as one marked by the combination of computational and informatics subdisciplines, but the combination of simulation and experimental data is only one of numerous combinations that are available for scientists, whom he explains see the data stores only much later in the discovery process.[194] The structural changes to inquiry – the division of data labor, methodological disruption, and the displacement of theory – all suggest that this is a moment of transition - a post-normal, pre-paradigmatic, crisis state of emergence, which do not necessarily take on a new shape quickly.

This fourth paradigm extends well beyond scientific discovery. Leading information ethicist Luciano Floridi's recent book *The Fourth Revolution*[195] echoes award winning historian Bruce Mazlish's 1993 *The Fourth Discontinuity*.[196] Both argue that we once understood humans as something special but that Nicholas Copernicus's theory dislodged us from the center of the universe, Darwin placed us on an evolutionary chain with the rest of life on Earth, and Freud Both argue that we again are having an identity crisis, trying to understand ourselves as distinct and special in the Digital Age. Floridi discusses humans in relation to information and Mazlish in relation to machines. The previous three phases took decades of debate and experimentation to resolve and have been embraced in various ways across fields and cultures. The fourth paradigm, revolution, and discontinuity all rely on big data and will represent important aspects of the second wave of big data emergence. For the purposes of governance or policy, the relevant difference is a prolonged state of

---

[193] HEY, TANSLEY, & TOLLE, *supra* note 123, at xvii.

[194] *Id.*

[195] LUCIANO FLORIDI, THE FOURTH REVOLUTION: HOW THE INFOSPHERE IS RESHAPING REALITY (2014).

[196] BRUCE MAZLISH, THE FOURTH DISCONTINUITY: THE CO-EVOLUTION OF HUMANS AND MACHINES (1993).

uncertainty about where we are headed, if anywhere – what is the new thinking cap?[197]

C. *Other Futures*

Of course a single set of events is not enough data to make a sound prediction.[198] Sketching a picture of the past is a futile exercise in many ways, and perhaps the overlooked differences make the comparison a fruitless one. Alternatively, there may have been other waves of technical advancement, followed by a wave of big data that resulted in no epistemic change.[199]

Perhaps nothing will come of the second wave of big data. Perhaps the similarities to the probabilistic revolution are too similar – this is an extension of the *exact* same process. The difference may only be digital, and therefore bigger, faster, and stronger, but will not yield a new understanding the world. Maybe the big data predictions we are working within today are simply a more sophisticated and saturated version of the probabilities from two hundred years ago. This is essentially Krishan Kumar's argument from 1995, in which he states, "The information society theorists can be attacked, firstly, for their short-sighted historical perspective… What seem to them novel and current can be shown to have been in the making for the past hundred years."[200] Kumar argues that the information society is simply a

---

[197] Hesitant to make any predictions about what kind of epistemic revolution may be approaching, it will necessarily be tied to computation, just as probability was tied to industrialization. I have found the most likely sources in machine learning. Tentatively I presume we will continue to be challenged by algorithmic living and its impact on society, but also our reliance on rules versus rationality and reason. For what I consider foreshadowing of some of the epistemic shifts in the computer age see PAUL ERICKSON, JUDY L. KLEIN, LORRAIN J. DASTON, REBECCA LEMOV, THOMAS STURM, AND MICHAEL D. GORDIN, HOW REASON ALMOST LOST ITS MIND: THE STRANGE CAREER OF COLD WAR RATIONALITY (2013).

[198] *See supra* note 17, for value of focusing on a single case.

[199] A potential place to look, although I have found nothing definitive or examples that extend big data beyond specific populations, are histories or the historizing of information overload. Daniel Rosenberg, *Early Modern Information Overload*, 64 J. HIST. IDEAS 1-9 (Jan. 2003), wherein the author discusses the scholarly publications that produced information overload throughout 1550 to 1750 and descriptions and measurements of the natural world that flooded scientists between 1550 and 1620.

[200] Krishan Kumar, *From Post-Industrial to Post-Modern Society*, in THE INFORMATION SOCIETY 109 (Frank Webster ed., 1995) (To that end, the "information society" "has not produced a radical shift in the way industrial societies are organized, or in the direction in which they have been moving. The imperatives of profit, power, and control seem as predominant now as they have ever been in the history of capitalistic industrialism. The

continuation of the industrial revolution.[201] Similarly, James Beniger argued in 1989 that the information society is an outgrowth of the control revolution that began in the 19th century, and that information technology will just be applied at higher levels of control.[202]

Perhaps we are all just computers. Big data may prove that, in fact, if everything is datafied, we are all just 1s and 0s governed by algorithms and the universe is one big computer. This would be quite an epistemic shift but essentially takes us back to determinism and so, not precisely an epistemic emergence. Disproving such notions has brought forth extraordinary gains in understanding our tools, innovations, and methods for interpreting and acting upon the world, as the debates and developments during the probabilistic revolution represent. In any case, everyone, including policy-makers, appears to be inclined to find out what big data is exactly and can do. So now, I turn to the data protection challenges such uncertainty presents.

V. DATA PROTECTION CHALLENGES

In some ways, it would be preferable, from a policy perspective, if big data were a revolution. Although a particular science would be in upheaval, at least there would be some central point to shift around. Emergence is slow, unintentional, uncertain, and widespread. The question for big data may not be how the law can balance benefits and

---

difference lies in the greater range and intensity of their applications… not in any change in the principles themselves." *Id.*, at 154).

[201] Kumar argued that the US is still an industrial society because it engages in the production and export of ideas and knowledge and that "In so far as Taylorism remains the master principle, information technology has a greater potential for proletarianization than for professionalization." *Id.*, at 112.

[202] JAMES BENIGER, THE CONTROL REVOLUTION 434 (1986). These arguments hinge on the machines and computers being understood the same way by those that apply them to humans and society. If we can all be broken down into discrete mechanical moving parts, we can be integrated into factory automation. But if we can all be broken down into discrete bits and algorithms, is there a difference? If not, Kumar is likely right and we are simply moving toward an increasing control world, but I believe the adaptability of computers (and humans) marks this as a (innovation-information-epistemic) cycle begun anew that may lead to a similarly momentous shift in understanding. In the 1800s we learned the many ways in humans are not machines and the world is not deterministic. I believe a pursuit that investigates the way in which humans are or are not computers could reveal an entirely new insight.

harms,[203] but what role law should play during times of emergence.[204] Too little is known at this point about the role of law in the emergence of probabilistic thought (as well as scientific revolutions in general). We are flying blind in two senses: we do not know enough about the role of law during different kinds of progress, revolution, evolution, or emergence, and we do not know what kind of understanding of the world may develop, if any, as the intellectual pursuits spurred by big data shake out and find footing. I turn now to the roles of law in emergence in order to better tailor difficult big data questions moving forward in light of reflections on the avalanche of numbers.

A. *Law as Regulator*

The Fair Information Practices Principles (FIPPs) have governed information use through its collection to disposal for decades. These principles have been in place since the 1970s, when they were originally crafted by a Health, Education, and Welfare advisory committee in response to growing concern about the use of data banks and systems holding and processing personal information.[205] Other committees, like the Committee on Privacy in Great Britain, were considering the same concerns and over the next decade FIPPs evolved and was incorporated into a number of data protection

---

[203] Jules Polonetsky and Omer Tene, *Privacy and Big Data Making Ends Meet*; *Big Data for All: Privacy and User Control in the Age of Analytics*, 66 STAN. L. REV. ONLINE 25 (2013).

[204] While I argue that the changing nature of inquiry is a strong justification for revisiting data protection policy in this section, it is important to note even differences in degree related to size, speed, and variety have strained basic concepts of data protection, like control. Sam Pfeifle, "Keynote: Forget Notice and Choice, Let's Regulate Use," IAPP PRIVACY ADVISOR (Dec. 12, 2013) at https://www.privacyassociation.org/ publications/ keynote_forget_notice_and_choice_lets_regulate_use. Control of one's information has served as a tenant of privacy until recently. Doubts surround privacy protections that rely on users to control their own information when there is so much, in so many places, being used for so many reasons, with so many potentials. The sheer number of different users also puts a strain on existing information protections where identification and enforcement of abuse are challenging. Total information awareness that leaves no room for emergent subjective, or the free spaces that protect the dynamic self's ability to develop and change, is another difference in degree worthy of motivating serious reflection on existing data laws. JULIE COHEN, RECONFIGURING THE NETWORKED SELF (2012). However, the differences in big data-based inquiry are also compelling justifications for retooling existing information policy for one main reason: the existing information policy (FIPPs) prescribes a process and the big data differences in kind are structural and methodological.

[205] COLIN J. BENNETT, REGULATING PRIVACY: DATA PROTECTION AND PUBLIC POLICY IN EUROPE AND THE UNITED STATES (1992).

regimes, including the Privacy Act of 1974 in the US, the Council of Europe's Convention for the Protection of Individuals with Regard to Automatic Processing of Personal data, and the Organization for Economic Cooperation and Development's Guidelines on the Protection of Privacy and Transborder Flows of Personal Data. FIPPs have been embraced by nearly every information policy, law, and regulation in one form or another. While the principles take numerous forms, I will use the OECD guidelines:[206]

> Collection Limitation Principle: There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject.
>
> Data Quality Principle: Personal data should be relevant to the purposes for which they are to be used and, to the extent necessary for those purposes, should be accurate, complete, and kept up-to-date.
>
> Purpose Specification Principle: The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfillment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.
>
> Use Limitation Principle: Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with the Purpose Specification Principle except: a) with the consent of the data subject; or b) by the authority of law.
>
> Security Safeguards Principle: Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorized access, destruction, use, modification or disclosure of data.
>
> Openness Principle: There should be a general policy of openness about developments, practices and policies with

---

[206] ORG. FOR ECON. CO-OPERATION AND DEV., OECD GUIDELINES ON THE PROTECTION OF PRIVACY AND TRANSBORDER FLOWS OF PERSONAL DATA (1980).

> respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.
>
> Individual Participation Principle: An individual should have the right: a) to obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to him; b) to have communicated to him, data relating to him within a reasonable time; at a charge, if any, that is not excessive; in a reasonable manner; and in a form that is readily intelligible to him; c) to be given reasons if a request made under subparagraphs (a) and (b) is denied, and to be able to challenge such denial; and d) to challenge data relating to him and, if the challenge is successful to have the data erased, rectified, completed or amended.
>
> Accountability Principle: A data controller should be accountable for complying with measures, which give effect to the principles stated above.

Big data practices have challenged these principles and caused many to call for the discontinuation or adjustment of FIPPs. Rationales for redrafting FIPPs usually involve the inefficiencies and difficulties in applying the principles and/or the benefits that are prevented by the application of the principles.

Rationales for revamping FIPPs often point to the principles that do not seem to represent the way big data analytics (and the data-based world they have emerged from) are being performed, and argue that applying the principles is inefficient. Notice and consent have become incredibly problematic,[207] which throws off data protection that relies on user control and participation.[208] These problems filter down through the other principles. The two principles most related to

---

[207] Daniel J. Solove, *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1879, 1888-93 (2013) ("The point is that it is virtually impossible for a person to make meaningful judgments about the costs and benefits of revealing certain data."); Fred H. Cate, *The Failure of Fair Information Practice Principles*, in CONSUMER PROTECTION IN THE AGE OF THE 'INFORMATION ECONOMY' 341 (Jane K. Winn ed.,2006).

[208] Alessandro Acquisti & Jens Grossklags, *Privacy and Rationality: A Survey*, *in* PRIVACY AND TECHNOLOGIES OF IDENTITY: A CROSS-DISCIPLINARY CONVERSATION 15, 16 (Katherine R. Strandburg & Daniela Stan Raicu eds., 2006); Paul M. Schwartz, *Privacy and Democracy in Cyberspace*, 52 VAND. L. REV. 1609, 1661 (1999).

big data processes are purpose specificity and use limitations.[209] Big data practices do not begin with data collected to be tested for a specific purpose but the aggregation of incongruent data to discover correlations that would otherwise go unnoticed. The use limitation is similarly problematic, because big data practices rely on the reuse and sharing of data to move data into new contexts and to understand how things change over time. It is argued that in order to maintain big data's three V's, FIPPs should be retooled so that data can remain voluminous (e.g., do not delete after specified use), fast (e.g., do not revisit user to get consent for new use or context), and various (e.g., share data with other partners not realized during collection). Even the data quality principle is problematic to big data, "with its emphasis on comprehensive datasets and messiness, [which] helps us get closer to reality than did our dependence on small data and accuracy."[210]

The related rationale is that these principles should be efficient for data practices because those data practices result in profound social benefits. Examples of the benefits range from medical advances to national security. These benefits often have immediate results. Cukier and Mayer-Schönberger cite to the way in which analytics are used to monitor and treat premature babies.[211] Analytics have also shed light on nudging individuals to vote, traveling patterns in urban areas, the impact of good teachers,[212] managing drought,[213] and identifying crimes and criminals.[214] With these benefits, many are inspired to reconsider FIPPs or other data protection regimes that would infringe on reaping and further developing these benefits.

---

[209] Fred H. Cate, Peter Cullen and Vikter Mayer-Schonberger, "Data Protection Principles for the 21st Century: Revising the 1980 OECD Guidelines," Oxford Internet Institute Report (March 2014); "Big Data and Analytics: Seeking Foundations for Effective Privacy Guidance, Centre for Information Policy Leadership Discussion Report (Feb. 2013).

[210] CUKIER & MAYER-SCHÖNBERGER, *supra* note 3, at 48.

[211] *Id.*, at 60.

[212] Erez Aiden & Jean-Baptiste Michel, "The Predictive Power of Big Data," NEWSWEEK (Apr. 22, 2014), http://www.newsweek.com/predictive-power-big-data-225125.

[213] Marcus Wohlsen, "Big Data Helps Farmers Weather Drought's Damage," WIRED (Sept. 6, 2012), http://www.wired.com/2012/09/big-data-drought/.

[214] Mark Ward, "Crime Fighting with Big Data Weapons," BBC (Mar. 18, 2014), http://www.bbc.com/news/business-26520013.

While these rationales are certainly worthy of consideration, they essentially relegate the law to weighing pros and cons. This has been a consistent theme in big data legal scholarship: the law must figure out how to balance the benefits with the harms.[215] The White House Big Data Report, released in May 2014, states, "Perhaps most important of all, a shift to focus on responsible uses in the big data context allows us to put our attention more squarely on the hard questions we must reckon with: how to balance the socially beneficial uses of big data with the harms to privacy and other values that can result in a world where more data is inevitably collected about more things."[216] As the above sections argue, the benefits of big data are uncertain.[217] Some projects provide seemingly beneficial results, but the overall benefits of big data are hard to predict. The harms are equally hard to predict, even when the assessment is restricted to a single project.

FIPPs intends to protect values in an increasingly automated world by prescribing a series of practices. By requiring a specified purpose, collection based on that purpose, and the destruction of data once the purpose has been met, FIPPs prescribes a way in which data practices must be performed. Risk-based accountability approaches have been proposed in the wake of FIPPs criticisms.[218] A risk-based approach increasingly allows data controllers and users to assess the risk the use poses to the data subject.[219] These efforts do not prescribe any particular process of inquiry, but have been criticized because

---

[215] *See supra* note 6.

[216] THE WHITE HOUSE, *supra* note 151.

[217] Paul Ohm, *The Underwhelming Benefits of Big Data*, 116 U. PA. L. REV. 339 (2013).

[218] Stuart S. Shapiro, "The Risk of the 'Risk-Based Approach,'" IAPP PRIVACY PERSPECTIVES (Mar. 31, 2014), https://www.privacyassociation.org/ privacy_perspectives/post/ the_risk_of_the_risk_based_approach (referring to a risk-based approach to privacy as the new black and accountability as less chic).

[219] Kenneth A. Bamberger, *Regulation as Delegation: Private Firms, Decisionmaking, and Accountability in the Administrative State*, 56 DUKE L. J. 377, 386-87 (2006).

they infringe the basic rights of individuals[220] and identifying risks under particular types of uncertainty is incredibly challenging.[221]

As those who have expressed concern about the applicability of FIPPs to performing big data practices suggest, the more pressing rationale is a movement away from the process FIPPs prescribes. The challenge for updating data protection so that it no longer adheres to an outdated process is that there is no new process in place to shift to – what will come of big data methods, practices, structures, theories is still unfolding. It is a moving target and any epistemic revolutions on the horizon are still unclear.

To continue to rely on the current version of FIPPs and the process it embodies effectively prohibits practices in post-normal big data science. To ignore the process set forth by FIPPs is to legitimize certain big data practices and potential, notably the high saturation of automated, data-based decision-making. Neither is unheard of. The departure from normal methods is not without its critics – it wouldn't be post-normal science without critics. On the contrary, the barrage of methodology, displacement of theory, and shifts in data labor that have come along with big data has provoked responses that range from hesitation to warnings to condemnation, and are discussed throughout Section II and III. Whether big data analytics and various immature approaches associated with it are short-sighted is one question. Whether they should be prohibited is another. Certainly history provides examples of swift legal responses to scientific inquiry that ranged from inhumane[222] to simply more harmful than beneficial,[223] as well as the research with the potential to know more

---

[220] Ann Cavoukian, "More Privacy Paternalism: 'We Know What's Best for You,'" IAPP PRIVACY PERSPECTIVES (Apr. 18, 2014), https://www.privacyassociation.org/ privacy_ perspectives/post/more_privacy_paternalism_we_know_whats_best_for_you. Also, the European Union Data Protection Directive continues to emphasize the important of rights held by data subjects as it is revamped to become the EU Data Protection Regulation.

[221] CALO, *supra* note 13; New Frontiers of Privacy Harms, Silicon Flatirons Center Symposium (Jan. 17, 2014).

[222] Patricia A. King, *The Dangers of Difference, Revisited*, in THE STORY OF BIOETHICS 197-214 (Jennifer K. Walter & Eran P. Klein eds., 2003) (detailing the "Tuskegee Study of Untreated Syphilis in the Negro Male" that came to light in 1972).

[223] Tom L. Beauchamp, *The Origins, Goals, and Core Commitments of the Belmont Report and Principles of Biomedical Ethics*, *in* THE STORY OF BIOETHICS 197-214 (Jennifer K. Walter & Eran P. Klein eds., 2003) (detailing the history of the Belmont Report and the National Research Act of 1974).

about people than is socially beneficial.[224] If the nature of inquiry blossoming in the data-based paradigm is to be discontinued due to individual and social harms, action should be taken to prevent its growth. If not, data protection must be revisited because it currently leaves data subjects exposed to great risks.[225]

B. *Law as Legitimizer*

Although the law will find balancing benefits and harms difficult with such uncertain developments, it will also find its role as legimitizer challenging. "One key role of the legal order in society, a role backed up by force if necessary, is to keep society on an even keel, to preserve it more or less as it is, and so far as change is concerned, to guarantee that change occurs in orderly and regular ways, in ways that society approves."[226] Based on the unique ways in which methods spread and advanced during the probabilistic revolution, the law holds an interesting role in big data development. In the nineteenth century big data practices were legitimized by being driven by government initiatives and collaborations, utilized by various agencies, but were also unrecognized when insufficient or unclear – well before probability had settled into what we understand it today.

Of course the census is a form of policy legitimizing the use of data to administer resources to alter the population, but the modern Western census was first conducted in Canada in 1666 to combat the threat of under-population and defense in the face of English colonialism.[227] The United States of America, named by statistician Richard Price, conducted its first constitutionally imposed census in 1790, but it was considered inaccurate and poorly administered. During the avalanche of numbers, census questions presented in the census grew rapidly, but the avalanche did not fall in the U.S. between

---

[224] Richard Delgado and David R. Millen, *God, Galileo, and Government: Toward Constitutional Protection for Scientific Inquiry*, 53 WASH. L. REV. 349 (1977-1978) (discussing prohibitions on DNA research in the 1970s).

[225] Paul Ohm, *Changing the Rules: General Principles for Data Use and Analysis*, in PRIVACY, BIG DATA, AND THE PUBLIC GOOD: FRAMEWORKS FOR ENGAGEMENT 96 (Julia Lane, Victoria Stodden, Stefan Bender, Helen Nissenbaum eds., 2014).

[226] LAWRENCE M. FRIEDMAN, GUARDING LIFE'S DARK SECRETS 2 (2007).

[227] Hacking, *Biopower and The Avalanche of Printed Numbers*, *supra* note 28, at 289.

1820 and 1840. The 1790 census only asked six questions[228] and not much changed until 1870, "[b]ut once the American bureaucrats caught on, they typically made 'avalanche of numbers' a mild understatement."[229] In 1870, 156 questions, then 13,010 in 1880, and in 1890 only a few more were asked, but they were calculated using Herman Hollerith's tabulation machine.[230]

Prior to the increase of questions, U.S. citizens were fined for non-compliance, communities were involved in shaming members into participation through public posting of returns, and marshals were fined if they did not meet their quotas.[231] Not only did the new enthusiasm for the census represent a form of legitimizing datafied governance, it was also associated with an important understanding regarding participation – confidentiality. The Census Office had a general policy abolishing public postings beginning in 1850 that stated, "No graver offense can be committed by Assistant Marshals than to divulge information acquired in the discharge of their duty. All disclosures should be treated as strictly confidential . . . . The Department is determined to protect the citizen in all his rights in the present Census."[232] But there were no penalties for doing so. Finally, at the behest of Representative James A. Garfield who explained "[t]he citizen is not adequately protected from the danger, or rather the apprehension, that his private affairs, the secrets of his family and his business, will be disclosed to his neighbors,"[233] Congress criminalized disclosure in 1890.[234] Trust, not pressure or coercion, appears to be an important aspect of legitimizing data practice.

---

[228] The questions were: (1) name of head of family, (2) number of free white males over sixteen, (3) number of free white males under 16, (4) number of free white females, (5) number of other free persons, and (6) number of slaves. These were not submitted on standardized forms, but whatever was convenient for the marshals so little uniformity existed in results. MARGO J. ANDERSON, THE AMERICAN CENSUS: A SOCIAL HISTORY 14 (1988).

[229] Hacking, *Biopower and The Avalanche of Printed Numbers*, *supra* note 28, at 290.

[230] *Id.*

[231] Douglas J. Sylvester and Sharon Lohr, *The Security of Our Secrets: A History of Privacy and Confidentiality in Law and Statistical Practice*, 83 DENVER U. L. REV. 147, 155-156 (2005).

[232] *Id.*, at 156.

[233] *Id.*, at 158 (citing House Comm. on the Ninth Census, H.R. Rep. No. 41-3, at 49 (1870)).

[234] *Id.*, at 159 (citing Act of Mar. 1, 1889, ch. 319, §§ 8, 13, 25 Stat. 760 (1889)).

Courtrooms were original spaces for investigation by the probabilists. Louis Poinsot considered Poisson's 1835 application of probability to judicial statistics a "false application of mathematical science… This singular idea of a calculus applicable to things where the ignorance and passion of men are intermingled in an imperfect light is dangerously illusive in several sense."[235] And later that "the application of this calculus to matters of morality is repugnant to the soul. It amounts, for example, to representing the *truth* of a verdict by a *number*, to thus treat men as if they were dice, each with many faces, some for error, some for truth."[236] Poisson, like Laplace and Nicolas de Condorcet before him, sought to optimally design a jury or tribunal so that the probability of an erroneous decision was minimized.[237]

A probabilistic degree of certainty was also apportioned to the probative weight of different types of evidence.[238] A probability for judging correctly was calculated for each decider, whose minds, if they were good minds, calculated and compared probabilities based on the regularity and frequency of experienced events.[239] Although probability theory's late eighteenth century and early nineteenth century goal to model reasonableness had tied it tightly psychology and law, it became instead affixed to habit, bias, self-interest, and ignorance rather than mental calculus over the next decades. "The story of the ill-fated probability of judgments might serve as an object lesson in the need to exercise caution in the choice of a suitable set of phenomena to mathematize. The 'good sense' of reasonable men turned out to be notoriously unstable, as probabilists bent on mathematically describing it discovered to their chagrin."[240] These

---

[235] Quote discussed by Siméon Denis Poisson in 1836 publication. *Note sur la loi des grands nombres*, COMPTES RENDUS HEBDOMADAIRES II, 377-382, 380 (1836). *See* Stigler, *supra* note 83, at 194.

[236] Further discussed by Poisson in another article in the same volume of the same publication. *Note sur le calcul des probabilités*, COMPTES RENDUS HEBDOMADAIRES II, 377-382, 380 (1836) 399 *See* Stigler, *supra* note 83, at 194.

[237] Lorraine J. Daston, *Fitting Numbers to the World: The Case of Probability Theory*, *in* HISTORY AND PHILOSOPHY OF MODERN MATHEMATICS 220, 228 (William Aspray & Philip Kitcher eds., 1988).

[238] *Id.*

[239] *Id.*, at 229.

[240] *Id.*, at 231.

ideas were motivated by an urgent interest in judicial reform, possible by the release of judicial statistics in the early nineteenth century, and conceptualized within existing notions of probabilities. The instability and criticism of these efforts likely influenced or at least stalled their incorporation into the courtroom. Although modeling and calculating the perfect trial faded as new conceptions of probability developed, the reasonable man remains a prominent fixture in legal systems.

The reasonable man was likely born in the nineteenth century, a period fixated and wrestled with averages, although is birth is heavily debated. The 1837 landmark case of *Vaughan v. Menlove* marks an early, if not original, moment in which the reasonable man took center stage in the law of negligence.[241] Menlove had been warned that his hay rack was dangerous and when it caught fire and destroyed a neighbor's building he was sued. Probability theory was in the eighteenth century "a formal description of the intuitions of a prototypical reasonable man, and as a prescription for the rest of us," meaning probability would model the minds of elite men who must be reasonable. This study of the rational individual faded when probability theory and statistics became the study of the irrationality of the masses. And so when Menlove argued in 1837 that he was not very intelligent and should be judged by a personal best effort standard, the court rejected it for an objective reasonableness standard, which "eliminates the personal equation and is independent of the idiosyncrasies of the particular person whose conduct is in question."[242] Early in the twentieth century Oliver Wendell Holmes' theory of tort liability reflects an American version of Quetelet's average man:

> The "reasonable man" knows, because "experience" tells him that a given behavior in a given circumstance—say, taking target practice in a populated area—carries the risk of injuring another person. Of course, any action in any circumstance carries some risk, however remote, of injuring another person; and reasonable people know this. But this knowledge is not what reasonableness consists in. What reasonableness consists in is the knowledge of the greater or lesser probability of an injury caused by

---

[241] Vaughan v. Menlove, (1837) 132 Eng. Rep. 490; 3 Bing (N.C.) 468.

[242] Glasgow Corp. v. Muir, [1943] A.C. 448 (H.L.) 457 (the reasonable man is "presumed to be free of both over-apprehension and from over-confidence").

> such and such an action in such and such circumstances. "[E]ven in the domain of knowledge," as Holmes put it, "the law applies its principle of averages."[243]

The history of the development of fault, risk, and reasonableness is disputed, but the reasonable man is a confusing fixture in law. Judges have poetically described this ordinary everyman as one on the Clapman omnibus or his lawn mower wearing shirt sleeves,[244] while scholars have mockingly endowed him with extraordinary abilities.[245] Even though the law, not the individual, sets the standard for negligent actions, big data may provide new tools and thinking about negligence, in which case the law will have to consider whether to embrace or reject big data.

Larger conversations regarding social responsibility and human rights consumed many of these visionaries. Free will debates occupied the writers of England, who responded to Henry Thomas Buckle's use of Quetelet's work to dismiss the notion of free will.[246] Fitzjames Stephen argued that statistical regularlity was an insufficient basis for conlcusions regarding human behavior and Robert Campbell chipped away at the rigor of Buckle's work revealing instability.[247] Although social and administrative statistics were strong in France, the French were leading opponents of statistics, with thinkers like Charles Bernard Renouvier who defended indeterminism, meaning that human freedom and free individual choices caused history, not vice versa.[248] The Germans offered the most heated and insightful debate on the subject.[249] Economist Adolph Wagner's round about support of Buckle's work, which was delivered in the form of a parable about social tendencies that great rulers could predict and expect but no

---

[243] LOUIS MENAND, THE METAPHYSICAL CLUB: A STORY OF IDEAS IN AMERICA 345-346 (2001).

[244] Hall v. Brooklands Auto Racing Club, [1933] I.K. B. 205, 224 (UK).

[245] DAVID MAXWELL WALKER, THE OXFORD COMPANION TO LAW 1038 (1980).

[246] GERD GIGERENZER, ET AL., *supra* note 70, at 45-50.

[247] *Id.*

[248] *Id.*, at 64.

[249] *Id.*, at 45-52.

individual or state could oppose, drove its contestation in Germany.[250] Although Wagner did not deny free will, his work sparked responses on the subject from other commenters who critiqued statistics of society in relation to the individuals within it.[251] To German thinkers, it was unacceptable to understand society as a mere sum of individuals. The French social physics and British economic liberalism were considered reductionistic and mechanic compared to the German holistic view of society, made up of varying autonomous individuals capable of acting in self-interest as well as to serve the larger social entity.[252] German statistical innovation was inspired by the free will debate and the possibility of meaningful reform. For instance, Wilhelm Lexis showed in 1879 that organized labor workers could compel an increase in wages even in the face of the English economic natural laws (e.g., David Ricardo's "iron law," which stated that wages could never rise above subsistence). These ideas have shaped and been embraced by liberal and socialist national policies creating entire identities for nations and regions. Moral responsibility, causality, and reasonableness are the things of law and they were central to developments in the probabilistic revolution.

The tumultuous development of consensus that follows the disruption of normal science (post-normal science) may work toward resolving uncertainty, complexity, and high stakes. Pressing policy choices extends far beyond the conventional use of the word 'science.'[253] Co-production of science[254] by joint efforts of decision-makers, regulators, experts and the public is encouraged during post-normal transitions.[255] As methodological and epistemological debates that grew out of a very culturally entrenched scientific shift during the probabilistic revolution, the law was more fully entrenched in emergent developments than in scientific revolutions, which are

---

[250] *Id.*, at 49-51.

[251] *Id.*

[252] *Id.*, at 50-52.

[253] John Turnpenny *Where Now for Post-Normal Science?*, 36:3 SCI., TECH., & HUMAN VALUES 287-306 (2011).

[254] SHEILA JASANOFF, STATES OF KNOWLEDGE: THE CO-PRODUCTION OF SCIENCE AND SOCIAL ORDER (2004).

[255] John Turnpenny *Where Now for Post-Normal Science?*, 36:3 SCI., TECH. & HUMAN VALUES 287-306 (2011); Karen Kastenhofer, *Risk Assessment of Emerging Technologies and Post-Normal Science*, 36:3 SCI., TECH., & HUMAN VALUES 307-222 (2011).

notable as more isolated scientific discoveries. For instance, free will played an integral part in the rise of probabilistic thinking, because determinism made no room for free will.[256] Similarly, if the law were to champion autonomy in an algorithmically operated world, would the law prevent us from realizing we are just machines or help guide us through a difficult deterministic phase of knowing? The law will not only place a role in develops surrounding big data and any impending emergence by legitimizing certain practices, processes, data labor structures, and theories as well as actively drawing lines and creating regulating boundaries within big data practices. As a social construct, law cannot be or remain isolated from epistemic shifts.

Today, the government's use of big data (e.g., for parole release[257]), funding of big data (e.g., National Science Foundation grant for Foundations of Data and Visual Analytics[258]), and contributing to big data (e.g., open government initiatives[259]) are all ways in which the law passively legitimizes an epistemic shift and certain related practices. Its hesitation to incorporate big data practices into education or convictions is an important kind of restraint. As these numerous forms and functions of big data mature, the law will play an important role in its direction. This role should be recognized and taken on with intention and reflection.

Far from a position of watching a technologically deterministic movement from the cheap seats, the law is vital to the development of and progress through any oncoming emergence. The role of legitimizer of epistemic shifts and practices is just one. The law's more comfortable charge, as regulator of particular initiatives and goals as well as safeguard of values throughout this transition are discussed next.

C. *Law as Defender*

Assuming that big data inquiry will not be prohibited by the application of FIPPs and the law takes a reflective approach to the way

---

[256] Ian Hacking, *Nineteenth Century Cracks in the Concept of Determinism*, *supra* note 1, at 455-475.

[257] "Parole and Technology: Prison Breakthrough," *supra* note 34.

[258] Foundations of Data and Visual Analytics (FODAVA), NSF Grant, Division of Computing and Communication Foundation, https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=501081.

[259] THE WHITE HOUSE, *supra* note 151.

in which it legitimizes big data practices, I turn to the challenges of adapting data protection that defends basic values to big data practices in transition. When normal science is disrupted, whether from an anomalous event or a slow emergence, accompanying safeguards may need to be questioned, sought, and established to prevent manipulation and abuse, over reliance on a particular, limited type of knowledge, and unfair and unwanted impacts of classification and prediction. FIPPs have articulated these choices and safeguards. But, as the previous section discusses, FIPPs protect these values by prescribing a particular process of inquiry, one that current data-based discovery does not adhere to and leaves individuals vulnerable to abuses, inappropriate conclusions, and unfair actions.

As proposals for revamping or discarding FIPPs are assessed, the transitional state of science should not fall too far into the backdrop. Focusing on the process or methods used may be an attempt to regulate a moving target, but as a legitimizing force, law will play an active role in the development of this shift.

Perhaps it is best to take the lessons from the first wave of big data, minimize the growing pains we can expect from the oncoming shift, and guide the development of new practices. In other words, without a specific practice to regulate, the focus on the principles is essential to guiding the development of responsible big data practices. Although there are many aspects of the big data debate that are inherent to datafication no matter its size, there are others that seem to be associated with the disruptive, particularly substantial influx in datafication practices. The same set of lures outlined in Section II(b)(2) for modern big data appear to have also spurred the probabilistic revolution two hundred years ago. These must be managed alongside the inherent concerns that always come with datafication in order to avoid some of the traps experienced during the first wave of big data. All this must be performed under a set of disrupted, disparate, and developing structural changes. A set of questions based on the lessons that may be taught by the first wave of big data are presented below. Working through each will help situate the role of law and policymakers in guiding the development of any progressing epistemic emergence.

1. *Structural Issues*

Structural issues are moving targets during an emergent phase. They change and shift as consensus develops and structures solidify. The law legitimizes these structural changes through traditional forms of regulation in its interaction with these issues. Determining what

big-data inquiry should look like, who can undertake it, under what terms, and how much deference to give to big data practitioners.

a. *Division of Data Labor*

The structure of the way in which data was collected and processed changed during the avalanche of numbers. Many more players from different parts of society were providing and processing new forms of data. Unlike France's centralized statistical practices, nineteenth century statistics were performed by scattered parties in Great Britain.[260] Britain's administrative bureaucracies had constant interaction with the social reformers, scholarly societies, and university professors, which led to numerous disputes, reliance, and organizational criticism.[261] Free-market reform in the country emphasized the local powers, adding another layer of disparate data collection. The General Register Office (dealing with social statistics) and the Board of Trade (dealing with economics) were created to coordinate the many moving parts and maintain consistency; the Central Statistical Office was created but 1941 to further coordinate the fragmented system.[262]

These offices intended to serve the public in a time when data did not hold the proprietary value it does today. Still, the fragmented data labor structure may be served by coordination from a government agency that coordinates the many moving parts, resolves inconsistencies, and cures inaccuracies amongst the many parties.[263] Questions for today may include whether such a form of governance would be an effective way to promote ethical progress and whether there is a division of data labor that should not be considered an available structure for data practices? There is no data protection agency in the United States, and of course, European Union member countries established data protection agencies in compliance with the 1995 Data Protection Directive. However, these agencies are not intimately involved with the data practices they govern. The potential for such intimacy is low considering hesitation toward government

---

[260] DESROSIERES, *supra* note 54, at 166.

[261] *Id.*

[262] *Id.*, at 166-67.

[263] Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583 (2014).

access to personal information and government-corporate arrangements to share data. Many U.S. agencies govern data within their sector, but an overarching data protection agency may help to coordinate ethical and legal practices and baselines across these segmented efforts.

b. *Methodological*

The development of number crunching methods and a significant shift in the way questions are answered occurred in the probabilistic revolution. The refinement of methodology took place over a number of decades, but eventually produced a sophisticated level of scientific inquiry that is now required for sound findings. Today big data analytics are being practiced across society by the highly trained to the budding amateur for different purposes and with various outlets. The field of data science and the profession of the data scientist are just beginning to take shape. This speaks directly to questions surrounding FIPPs as a suitable means of protection moving forward. While the second wave of big data methodology is scattered, criticized, and difficult to assess, how proactive is the law willing to be to legitimize methodology in light of its transitory state and potential?[264] This is a classic question for innovation, but the distinction between regulating particular aspects of data practices and respecting the confidences and dignity of users may provide a way forward.

c. *The Displacement of Theory*

The first wave of big data shed light on the idea that the displacement of theory may be a product of transition - a characteristic of a changing nature of inquiry, not a characteristic of big data practices necessary or an end to theory. This is relevant to the way in which we describe the thing we are attempting to govern. There is a difference between big data as theory-less and big data theory-developing. Beyond describing big data more accurately, what is the role of law while theory is displaced? A correlation without causation form of inquiry may be dangerous when the causation is related to a protected class, vulnerable population, or particularly invasive insight. How may the law protect impacts of drawing correlations and changing notions of choice, free will, and autonomy as these theories develop? Legal scholars should continue to voice these concerns about

---

[264] CATE, CULLEN & MAYER-SCHONBERGER, *supra* note 209.

big data. The philosopher-mathematicians that debated these issues in the past may not be readily available in the over-specialized modern world. Data scientists and legal scholars can coordinate their efforts to theorize what big data.

2. *Datafication Issues*

Datafication issues are inherent issues associated with representing facts or ideas in a formalized manner that can be communicated or manipulated by a process. These are ever-present and must be revisited as the nature of inquiry transform over time. The law may legitimize big data practices as they develop by drawing lines in between acceptable and unacceptable uses, utilizing big data methods and relying on big data-based knowledge, and providing means for oversight and participation.

a. *Governability*

Datafication makes us governable as individuals who must be accounted for to receive benefits and vulnerable to the manipulation of those counting us. The first wave of big data began from a perspective of monarchial rule, which is seemingly less meritorious and abusive than data-based governance. The concern was that in order to be counted under data-based governance, you had to be measured and those that created the measuring stick could exercise power over a population. In 1859 John Stuart Mill wrote, "By virtue of its superior intelligence gathering and information processing capacities, there is a distinct role for the central state acquired in a liberal society over and above both local and individual claims to autonomy."[265] In order to serve a population, the government has to know its population, but should those with superior information collection and processing override individual autonomy? The same question can be posed to governments, organizations, and companies today. Motives matter to the question of governability. What are inappropriate motives for those that use big data? What are manipulative practices?[266]

---

[265] DESROSIERES, *supra* note 54, at 170 (citing JOHN STUART MILLS, REPRESENTATIVE GOVERNMENT (1859)).

[266] CALO, *supra* note 13.

b. *Data-Based Knowledge*

Numbers are always seen through the lens of knowledge provided by a particular time and culture. For instance, psychometrics, the study of psychological measurement, developed separately in Britain and Germany during the mid-1800s, but both led to intelligence testing.[267] The correlation between aptitude tests and children led to the 'discovery' of the 'g-factor,' a variable that summarizes positive correlation between different cognitive tasks.[268] Children were tested to determine their general intelligence based on the g-factor at age eleven; they were then divided into two very different educational tracks. The sociologists of the 1950s attacked the practice as unjust, but also that propensity to succeed could be interpreted as effects of numerous factors, such as family environment or social background.[269] How can the law remind itself and the rest of us that big data is but one of many ways to understand the world? And that it carries its own limitations? Reliance on and legitimization of big data practices through incorporation into the legal system will certainly reflect limitations.

c. *Classification Effects*

The way in which classification choices are made can have dramatic social impacts and this will continue to be true in the second wave of big data. "Perverts" did not exist before the late nineteenth century.[270] In the 1820s the science of 'analyse morale,' the statistics of deviance, led to thousands of categories for people and motives ranging from labels for mental illness to types of criminals.[271] "New slots were created in which to fit and enumerate people. Even national and provincial censuses amazingly show that the categories into which people fall change every ten years. Social change creates new categories of people, but the counting is not mere report of

---

[267] DESROSIERES, *supra* note 54, at 145-46.

[268] *Id.*

[269] *Id.*

[270] Ian Hacking, *Making Up People*, *in* RECONSTRUCTING INDIVIDUALISM: AUTONOMY, INDIVIDUALITY, AND THE SELF IN WESTERN THOUGHT 222-230 (Thomas C. Heller, David E. Wellbery, & Morton Sosna, eds., 1986).

[271] *Id.*

developments. It elaborately, often philanthropically, creates new ways for people to be."[272] While same sex activity has always been present in society, homosexual and heterosexual people were not "made up" until the end of the nineteenth century.[273] The classification effects at data collection and processing stages during the first wave of big data have informed the way in which we do research today. It revealed that the act of classifying individuals can in and of itself be detrimental. Today, and for the last two hundred years and before that, classification choices shape identities and perceptions.[274] How can the law ensure that classifications do not unfairly categorize or force individuals into categories that inaccurately describe them?[275] How should the law protect or encourage individual participate in their own classification? How should misclassification be resolved?[276] In many ways these are procedural questions that relate to existing laws like those that protect against discrimination. However, policies that promote user participation in classification may certainly help resolve some of these issues.

3. *Big Data Issues*

Looking back at the issues that are associated with substantial influxes of data can be particularly informative to legal responses to the enthusiasm for big data. Realizing that the zeal for big data is likely well-founded but also uncertain, the law can look to expected disputes, pitfalls, or overlooked harms associated with each lure of big data. As a big data participant, the use of big data by the government and legal institutions will play an important role in legitimizing and directing any oncoming epistemic revolution. The law will also play a role in legitimizing practices and perspectives that change the way

---

[272] *Id.*

[273] *Id.*, at 163, citing KENNETH PLUMMER, THE MAKING OF THE MODERN HOMOSEXUAL (1981).

[274] Bowker and Star, *supra* note 26, at 4.

[275] Dwork & Mulligan, *supra* note 146, at 35.

[276] *See* e.g., Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 101 (forthcoming 2014), arguing that procedural safeguards should be put in place for audits by regulators and opportunities for individuals to challenge adverse decisions based on miscategorization.

individuals, groups, and interactions are treated across society during this transition.

a. *Standardized Sharing*

While the benefits of standardized sharing – all of us speaking the common language of data – are alluring, it is not as simple or as true as it sounds. Datafication is still messy. "Society must be remade before it can be the object of quantification."[277] In many ways the digital age has remade society to conceive standard representations of money, labor, products, and measurements, but there are new aspects of the world being datafied that do not yet fall into a standardized format. Representing people as data in the digital age is less standardized; although the old demographic categories are still relied upon, new granularity has presented data without standardized measurement or coding. Anyone who has ever received a data set from another researcher can tell you that context matters. But, with enough data and enough time, standards and categories will develop so the data may be employed with "almost… unbounded applicability,"[278] further expanding and limiting data-based knowledge.

Babbage never got his *Constants of Nature and of Art*, but his proposal realized that the project would be too great for a single individual and that scientific societies would need to cooperate. The proposal also included Babbage's nineteen categories for tables to be developed. Under the category for "velocities," he included "arrow, musket ball at several distances, cannon ball, sound telegraph, light, birds."[279] The problems that arise with numerous individuals spread across distances datafying the velocity of birds and arrows shed some light on the significant issues related to context and inconsistencies that arise from datafying people. How the law can encourage contextualized sharing of data that retains as much of the original intention and choices made by the collector? Relying on context as a conceptualization of privacy has become popular and novel ways to retain context or punish its disruption should continue to develop. Consequences for not maintaining data quality should also continue to be pursued.

---

[277] Porter, *Objectivity as Standardization*, *supra* note 123, at 19, 23.

[278] *Id.*, at 23.

[279] Charles Babbage, *On the Advantage of a Collection of Numbers, to be Entitled the Constants of Nature and Art*, V:1 EDINBURGH J. SCI. 334 (July 1831).

b. *Feedback and Control*

The feedback loop allowed for by datafication provides for increased control of a system by feeding output data back into the controller of the process, but the controller is not the only aspect of the loop that adjusts. Those affected by the output of the data adjust to the feedback loop as well. There is always a risk of the counter creating the population that is it counting.[280] This is the issue of reflexivity. Although Taylorism sought to motivate workers to be their most efficient, many other datafication efforts intend only to give people what they want more efficiently. This is not necessarily intended to change what it is they want, but this is often the end result and very hard to detect or cure. Law professors can easily relate to the concept of reflexivity, as they are situated in institutions that exemplify (and often resist) the practice of adapting to the rankings that "make them count."[281] Those counted adjust to be accounted for and those left uncounted become invisible, forfeiting the benefits that being counted provide.[282]

It is said that Taylorism failed because it modelled the worker as simply self-interested, through the model of systematic soldering, and the managers as "heartily cooperative."[283] The asymmetry left no room for intelligence and innovation of employees. Education tracking has also been similarly criticized as limiting the potential of those students assessed in lower tracks and providing curriculum that reinforces inequalities.[284] Today algorithmic living has become the norm. From the filter bubble to modern scientific management of work, algorithms direct us. At what point does the further integration of human-machine systems reach a dehumanizing point at which law should address the issue? What kind of transparency in these systems should

---

[280] Anat E. Leibler & Daniel Breslau, *The Uncounted: Citizenship and Exclusion in the Israeli Census of 1948*, 28:5 ETHNIC & RACIAL STUD. 880-902 (2005).

[281] Wendy Nelson Espeland and Michael Sauder, *Rankings and Reflexivity: How Public Measures Recreate Social Worlds*, 113:1 AM. J. SOC'Y 1-40 (July 2007).

[282] Leibler and Breslau, *supra* note 211, at 880.

[283] Sigmund Wagner-Tsukamoto, *Scientific Management Revisited: Did Taylorism Fail Because of a Too Positive Image of Human Nature?*, 14:4 J.MGMT. HIST. 348-372 (2006).

[284] Jomills Henry Braddock, II and Robert E. Slavin, *Why Ability Grouping Must End: Achieving Excellence and Equity in American Education*, presented at Common Destiny Conference at Johns Hopkins University (Sept. 9-11, 1992), available at http://files.eric.ed.gov/fulltext/ED355296.pdf.

be required or encouraged? This is an area that needs significant contribution from modern legal scholars. Questions surrounding important distinctions between humans and machines should be engaged promptly.

c. *Objectivity*

Reliance on the impersonality of numbers provides a convincing way to make decisions while avoiding claims of coercion, improper bias, or stirring up the public.[285] Objectivity was an important driving force for scientific developments over the course of the nineteenth century, advocated strongly for by prominent figures like Pearson who saw… "Science claims unblemished character, not for the individuals who make it, but for the knowledge that results… The rhetoric of science is persuasive in large part because that knowledge is assumed not to depend on the fallible individuals who constitute the scientific community."[286]

Today it is well understood that the rules that govern scientific knowledge and the choices made within it are subject to human fallibility, but it hard to remember. In nineteenth century France, where public suspicion and interest was aroused by administrative decisions made on their own authority, numbers and statistics were indispensable as the credible language of disinterested French engineers that were also generally incomprehensible to the public at large.[287] Objectivity continues to permeate our efforts to solve social problems[288] in a setting where government decisions are scrutinized and data practices are beyond the reach of much of the public. How should the law reinforce the "human" aspects of big data? How can and should values in design be assessed and directed for developing big data practices? Avoiding overreliance by legal entities on the objectivity of these systems is one way. Another is further development of privacy by design concepts.

---

[285] *Id.*, at 20, 28.

[286] Theodore M. Porter, *Objectivity and Authority: How French Engineers Reduced Public Utility to Numbers*, 12:2 POETICS TODAY 250 (1991).

[287] *Id.*, at 254-55.

[288] *See* e.g., Cristóbal Romero, *Educational Data Mining: A Review of the State of the Art*, 40:6 SYSTEMS, MAN, AND CYBERNETICS 601, 602 (2010) (explaining that one goal of education data mining is "[t]o get objective feedback about instruction.").

d. *Knowledge Discovery*

When knowledge about people is discovered and not produced, it certainly carries an air of objectivity. Knowledge 'discovery' also appears to bring enthusiasm for identifying the 'rules' or sequences that make the world go round; in other words, a growing interest in determinism and waves of big data seem to go hand in hand. In 1836 Quetelet wrote, "The moral order falls in the domain of statistics… a discouraging fact for those who believe in the perfectibility of human nature. It seems as if free will exists only in theory."[289] When knowledge is discovered, like the law of gravity, entire systems can and must be built around such laws. It is only slightly comforting to know that statistical laws may apply to populations but not members of a population, but are consistently feel discomfort when presented with a defendant and crime statistics that bring concerns about responsibility and likelihood of guilt, such as the statistic that those physically abused are nine times more likely to be involved in criminal activity and a third will abuse their own children.[290]

Social physics is back, thanks to the second wave of big data.[291] Understanding the laws of social physics allows human life to be re-engineered for desired effect, or so the echoed argument goes.[292] Whether the big data transition will be a return to determinism, or knowledge "discovery" will be qualified or go through the painful phase that social laws and probability went through, will be a matter of guiding big data methodology and understanding through its development. Like every strategy to gain knowledge, it has its drawbacks, but even in its infancy, its strong appeal has accelerated big data-based knowledge discovery forwarded. What is the role of law when faced with determinism? Certainly, it legitimizes certain practices through its own use of analytics and evidence-based policy – avoiding "pre-crime" programs[293] – but how should the law impart

---

[289] Hacking, THE TAMING OF CHANCE, *supra* note 59, at 116.

[290] "Long-Term Consequences of Child Abuse and Neglect," Child Help, http://www.childhelp-usa.com/pages/statistics.

[291] Pentland, *supra* note 104.

[292] Matt Buchanan, *Speak Softly and Carry a Big Data*, THE NEW YORKER (June 7, 2013); Pentland, *supra* note 104.

[293] James Vlahos, *The Department of Pre-Crime*, 306 SCI. AM. 62-67 (2012); Gabe Mythen and Sandra Walklate, *Pre-crime, Regulation, and Counter-Terrorism: Interrogating Anticipatory Risk*, 81:1 CRIM. JUST. MATTERS 34-46 (2010).

any misgivings about determinism onto other big data users? Not only will the law be tempted to acknowledge determinism, it will also be asked to consider probability and prediction in new light as big data changes the way we understand the world.

e. *Enumeration*

Objectivity, feedback and control, and knowledge discovery, encourage enumeration, but enumeration has not gone well. Tempted by the idea that so much more can be datafied, scientists during waves of big data seek to datafy everyone and everything. Attempts to capture all data assume that all data is captured; this error often has to be resolved retroactively. Early attempts to universally datafy had to be tempered and results qualified, such as in Great Britain and Germany. Efforts to enumerate consistently fail to recognize those individuals and characteristics of people and the surrounding world that are invisible to the processes of datafication at our disposal.[294] How can the law encourage the acknowledgement of those not counted, by choice or situation, those people and aspects of the world that remain invisible?[295] Acknowledging and protecting those without a voice is often the law's job. Looking to the ways in which it has done so in other settings would support progress in the big data transition.

These questions do not present an exhaustive list of questions for big data today,[296] but those that may have longer historical references to guide us. The questions require no balance of the benefits and harms of big data. Where is the line between control and manipulation? No benefit should resolve that question. How can those that use big data-based knowledge recognize and express its bias, reductions, and limitations? How should errors in any of the questions be communicated and resolved? There are plenty of questions to answer that do not hinge on weighing pros and cons of big data, which has proven to be an exceptionally problematic exercise at the point in

---

[294] Leibler & Breslau, *supra* note 211, at 880 (the Israeli census example, the focus of this article, is an anomalous instance of enumeration because the census was implemented during wartime and collection took place over seven hours under significant social strain).

[295] Jonas Lerman, *Big Data and Its Exclusions*, 66 STAN. L. REV. ONLINE 55 (2013).

[296] For instance, it does not speak directly to machine learning or black box decision-making.

the transition.[297] Value disputes that do not rely on the outcome of this potential emergence, however, can be chipped away at and refined. This process has already begun with work by Ryan Calo on manipulation, Deirdre Mulligan and Cynthia Dwork on classification, Kate Crawford and Jason Schultz on feedback and control, and many others cited above. Often these normative works argue for prescriptive adjustments to big data practices – acting as legitimizer of certain directions for big data. This article has attempted to place those normative analyses and prescriptions in a broader perspective and enrich their contributions. The challenge is utilizing these contributions through an uncertain trajectory, recognizing the important role law plays in its direction.

CONCLUSION

> The truth of the matter is that our deficiency does not lie in the well-verified 'facts.' What we lack is our bearings. The contemporary experience of things technological has repeatedly confounded our vision, our expectations, and our capacity to make intelligent judgments.[298]

Often the pressure to look forward prevents us from looking back to get "our bearings." This article attempts to present data in a historical perspective to help establish those bearings in the exciting and tumultuous Digital Age. Many have pointed out the social problems presented by big data practices. Others have pointed out that existing data protection policies no longer work in a big data world. Most have acknowledged that big data is something different, whether in degree or kind, and that it offers great potential for a certain type of knowledge and innovation.

Relying on the avalanche of numbers as an appropriate comparison for big data (because of their similarities as titles for waves of data that flooded society following a technological revolution and the set of similarities extracted from the two periods) allows the big data phenomenon to be better understood historically and

---

[297] Ohm, *supra* note 175; Neil M. Richards & Jonathan H. King, *Three Paradoxes of Big Data*, 66 STAN. L. REV. ONLINE 41 (2013); The New Frontiers of Privacy Harm, Silicon Flatirons Center Symposium (Jan. 17, 2014).

[298] LANGDON WINNER, AUTONOMOUS TECHNOLOGY: TECHNICS-OUT-OF-CONTROL AS A THEME IN POLITICAL THOUGHT 7 (1978).

categorically. By comparing big data to the avalanche of numbers, an analysis of similarities reveals aspects of datafication that are ever-present, others that are lures associated with the excitement and optimism during both waves of big data, and structural changes to the nature of inquiry for the period. The avalanche of numbers was an important period in the early stages of an epistemic shift: the widespread recognition and understanding of probability and chance. Based on the similarities shared between big data and the avalanche of numbers and associated emergence of probability, the article situates big data in the early stages of a similarly prolonged, culturally entrenched, and still uncertain epistemic emergence.

Looking to the past for a richer understanding of the issues, I do not find that this is all a palaver over nothing – that big data is nothing new or that it is a lot of hype about nothing. Instead I find that there is potential for a significant epistemic shift on the horizon – one that presents the law with a great deal of uncertainty and responsibility. The role of law during times of optimism, disruption, and uncertainty is challenging, but there are lessons to be learned to from the avalanche of numbers that may mitigate the pains of progress presented by big data.

* * *