I expected a Model T, but instead I got a loom: Awaiting the second big data revolution

Mark Huberty*

February 11, 2014

Abstract

"Big data" has been heralded as the agent of a third industrial revolution–one with raw materials measured in bits, rather than tons of steel or barrels of oil. Yet the industrial revolution transformed not just how firms made things, but the fundamental approach to value creation in industrial economies. To date, big data has not achieved this distinction. Instead, today's successful big data business models largely use data to scale old modes of value creation, rather than invent new ones altogether. Moreover, today's big data cannot deliver the promised revolution. In this way, today's big data landscape resembles the early phases of the first industrial revolution, rather than the culmination of the second a century later. Realizing the second big data revolution will require fundamentally different kinds of data, different innovations, and different business models than those seen to date. That fact has profound consequences for the kinds of investments and innovations firms must seek, and the economic, political, and social consequences that those innovations portend.

*Prepared for the 2013 BRIE-ETLA Conference, September 6-7, Claremont California. This paper has benefited from extended discussions with Chris Diehl, David Gutelius, Joseph Reisinger, Sean Taylor, Sean Gerrish, Drew Conway, Cathryn Carson, Georg Zachmann, and John Zysman. All errors commmitted, and opinions expressed, remain solely my own.

1 Introduction

We believe that we live in an era of "big data". Firms today accumulate, often nearly by accident, vast quantities of data about their customers, suppliers, and the world at large. Technology firms like Google or Facebook have led the pack in finding uses for such data, but its imprint is visible throughout the economy. The expanding sources and uses of data suggest to many the dawn of a new industrial revolution. Those who cheer lead for this revolution proclaim that these changes, over time, will rival the industrial revolution in scope and consequences for economic and social prosperity.

Yet this "big data" revolution has so far fallen short of its promise. Precious few firms transmutate data into novel products. Instead, most rely on data to operate, at unprecedented scale, business models with long pedigree in the media and retail sectors. Big data, despite protests to the contrary, is thus an incremental change–and its revolution one of degree, not kind.

The reasons for these shortcomings point to the challenges we face in realizing the promise of the big data revolution. Today's advances in search, e-commerce, and social media relied on the creative application of marginal improvements in processing power and storage. In contrast, tomorrow's hopes for transforming real-world outcomes in areas like health care, education, energy, and other complex phenomena pose scientific and engineering challenges of an entirely different scale.

2 The implausibility of big data

Our present enthusiasm for big data stems from the confusion of data and knowledge. Firms today can gather more data, at lower cost, about a wider variety of subjects, than ever before. Big data's advocates claim that this data will become the raw material of a new industrial revolution that will alter how we govern, work. play, and live. These raw materials are so cheap and abundant that, we are told, the horizon is bounded only by the supply of smart people capable of molding these materials into the next generation of innovations (Manyika et al., 2011).

This utopia of data is badly flawed. Those who promote it rely on a series of bad assumptions about the origins and uses of data, none of which hold up to serious scrutiny. Taken together, those mistakes point out the limits of a revolution built on the raw materials that today seem so abundant.

Four assumptions need special attention: First, $N = all$, or the claim that our data allow a clear and unbiased study of humanity; second, that today equals tomorrow, or the claim that understanding online behavior today implies that we will still understand it tomorrow; third, that understanding online behavior offers a window into offline behavior; and fourth, that complex patterns of social behavior, once understood, will remain stable enough to become the basis of new data-driven, predictive products and services. Each of these has its issues. Taken together, those issues limit the future of a revolution that relies, as today's does, on the "digital exhaust" of social networks, e-commerce, and other online services. The true revolution must lie elsewhere.

2.1 N=all

Gathering data via traditional methods has always been difficult. Small samples were unreliable; large samples were expensive; samples might not be representative, despite researchers' best efforts; monitoring the same sample over many years posed all sorts of difficulties. None of this, moreover, was very scalable: researchers needed a new sample for every question, or had to divine in advance a battery of questions. No wonder social research proceeded so slowly.

Mayer-Schönberger and Cukier (2013) argue that big data will eliminate these problems. Instead of having to rely on samples, online data allows us to measure the universe of online behavior, where N (the number of people in the sample) is basically *All* (the entire population of people we care about). Hence we no longer need worry, they claim, about the problems that have plagued researchers in the past. When $N = all$, large samples are cheap and representative, new data on individuals arrives constantly, monitoring data over time poses no added difficulty, and cheap storage permits us to ask new questions of the same data again and again.

But $N \neq All$. Most of the data that dazzles those infatuated by "big data" comes from what McKinsey & Company termed "digital exhaust" (Manyika et al., 2011): the web server logs, e-commerce purchasing histories, social media relations, and other data thrown off by systems in the course of serving web pages, online shopping, or person-to-person communication. The N covered by that data concerns only those who use these services–not society at large.

Hence the uses of that data are limited. It's very relevant for understanding web search behavior, purchasing, or how people behave on social media. But the N here is skewed in ways both known and unknown–perhaps younger than average, or more tech-savvy, or wealthier than the general population. That we have enormous quantities of data about these people says nothing about whether that data tells us anything about society.

2.2 All (today) = All (tomorrow)

But let's say that we truly believe this assumption–that everyone is (or soon will be) online. Surely the proliferation of smart phones and other devices is bringing that world closer, at least in the developed world. This brings up the second assumption–that we know where to go find all these people. Several years ago, MySpace was the leading social media website, a treasure trove of new data on social relations. Today, it's the punchline to a joke. The rate of change in online commerce, social media, search, and other services undermines any claim that we can actually know that our $N = all$ sample that works today will work tomorrow. Instead, only actually know about new developments–and the data and populations they cover–well after they have already become big. Hence our $N = all$ sample is persistently biased in favor of the old.

2.3 Online behavior = offline behavior

But let's again assume that problem away. Let's assume that we have all the data, about all the people, for all the online behavior, gathered from the digital exhaust of all the relevant products and services out there. Perhaps, in this context, we can make progress understanding human behavior online. But that is not the revolution that big data has promised. Most of the "big data" hype has ambitions beyond improving web search, online shopping, socializing, or other online activity. Instead, big data should help cure disease, detect epidemics, monitor physical infrastructure, and aid first responders in emergencies.

To satisfy these goals, we need a new assumption: that what people do online mirrors what they do offline. Otherwise, all the digital exhaust in the world won't describe the actual problems we care about.

There's little reason to think that offline life faithfully mirrors online behavior. Research has consistently shown that individuals' online identities vary widely from their offline selves. In some cases, that means people are more cautious about revealing their true selves. danah boyd's work (Boyd and Marwick, 2011) has shown that teenagers cultivate online identities very different from their offline selves–whether for creative, privacy, or other reasons. In others, it may mean that people are more vitriolic, or take more extreme positions. Online political discussions–another favorite subject–suffers from levels of vitriol and partisanship far beyond anything seen offline (Conover et al., 2011). Of course, online and offline identity aren't entirely separate. That would invite suggestions of schizophrenia among internet users. But the problem remains–we don't know what part of a person is faithfully represented online, and what part is not.

2.4 Behavior of all (today) = Behavior of all (tomorrow)

OK, but you say, surely we can determine how these distortions work, and incorporate them into our models? After all, doesn't statistics have a long history of trying to gain insight from messy, biased, or otherwise incomplete data?

Perhaps we could build such a map, one that allows us to connect the observed behaviors of a skewed and selective online population to offline developments writ large. This suffices only if we care primarily about describing the past. But much of the promise of big data comes from predicting the future– where and when people will get sick in an epidemic, which bridges might need the most attention next month, whether today's disgruntled high schooler will become tomorrow's mass shooter.

Satisfying these predictive goals requires yet another assumption. It is not enough to have all the data, about all the people, and a map that connects that data to real-world behaviors and outcomes. We also have to assume that the map we have today will still describe the world we want to predict tomorrow.

Two obvious and unknowable sources of change stand in our way. First, people change. Online behavior is a culmination of culture, language, social norms and other factors that shape both people and how they express their identity. These factors are in constant flux. The controversies and issues of yesterday are not those of tomorrow; the language we used to discuss anger, love, hatred, or envy change. The pathologies that afflict humanity may endure, but the ways we express them do not.

Second, technological systems change. The data we observe in the "digital exhaust" of the internet is created by individuals acting in the context of systems with rules of their own. Those rules are set, intentionally or not, by the designers and programmers that decide what we can and cannot do with them. And those rules are in constant flux. What we can and cannot buy, who we can and cannot contact on Facebook, what photos we can or cannot see on Flickr vary, often unpredictably. Facebook alone is rumored to run up to a thousand different variants on its site at one time. Hence even if culture never changed, our map from online to offline behavior would still decay as the rules of online systems continued to evolve.

Compounding this problem, we cannot know, in advance, which of these social and technological changes will matter to our map. That only becomes apparent in the aftermath, as real-world outcomes diverge from predictions cast using the exhaust of online systems.

Lest this come off as statistical nihilism, consider the differences in two papers that both purport to use big data to project the outcome of US elections. DiGrazia et al. (2013) claim that merely counting the tweets that reference a Congressional candidate can provide leverage on whether that candidate will win his or her election. They make no adjustment for the demographics of the Twitter user base, the possibility of spam, the sentiment directed at the candidates, or other factors. This is a purely "digital exhaust" approach. As Huberty (2013a) shows, that approach fails. The Twitter data add no predictive value above and beyond party incumbency, already known to be a good predictor of election outcomes. Models trained on one election provide little traction on predicting future elections.

Contrast this with Wang et al. (2014). They use Xbox as a polling instrument, which they hope might help compensate for the rising non-response rates that have plagued traditional telephone polls. As with Twitter, $N \neq All$: the Xbox user community is younger, more male, or less politically involved. But the paper nevertheless succeeds in generating accurate estimates of general electoral sentiment. The key difference lies in their use of demographic data to re-weight respondents' electoral sentiments to look like the electorate at large. The Xbox data were no less skewed than Twitter data; but the process of data collection provided the means to compensate. The black box of Twitter's digital exhaust, lacking this data, did not.

2.5 The implausibility of Big Data 1.0

Taken together, the assumptions that we have to make to fulfill the promise of today's big data hype appear wildly implausible. To recap, we must assume that:

1. Everyone we care about is online
2. We know where to find them today, and tomorrow
3. That they represent themselves online consistent with how they behave offline
4. That they will continue to represent themselves online–in behavior, language, and other factors–in the same way, for long periods of time

Nothing in the history of the internet suggests that even one of these statements holds true. Everyone was not online in the past; and likely will not be online in the future. The constant, often wrenching changes in the speed, diversity, and capacity of online services means those who are online move around constantly. They do not, as we've seen, behave in ways necessarily consistent with their offline selves. And the choices they make about how to behave online evolve in unpredictable ways.

But if each of these statements fall down, then how have companies like Google, Facebook, or Amazon build such successful business models? The answer lies in two parts. First, most of what these companies do is self-referential: they use data about how people search, shop, or socialize online to improve and expand services targeted at searching, shopping, or socializing. Google by definition, has an $N = all$ sample of Google users' online search behavior. Amazon knows the shopping behaviors of Amazon users. Of course, that population is subject to change its behaviors, its self-representation, or its expectations at any point. But at least Google can plausibly claim to have a valid sample of the primary population it cares about.

Second, the consequences of failure are, on the margins, very low. Google relies heavily on predictive models of user behavior to sell the advertising that accounts for most of its revenue. But the consequences of errors in that model are low–Google suffers little from serving the wrong ad on the margins. The same is true of Facebook's recommendations for people you might know, or Amazon's quest to sell you complementary products. Of course, persistent and critical errors of understanding will undermine products and lead to lost customers. But there's usually plenty of time to correct course before that happens.

But if we move even a little beyond these low-risk, self-referential systems, the usefulness of the data that underpin them quickly erodes. Google Flu provides a valuable lesson in this regard. In 2008, Google announced a new collaboration with the Centers for Disease Control (CDC) to track and report rates of influenza infection. Historically, the CDC had monitored US flu infection patterns through a network of doctors that tracked and reported "influenza-like illness" in their clinics and hospitals. But doctors' reports took up to two weeks to reach the CDC–a long time in a world confronting SARS or avian flu. Developing countries with weaker public health capabilities faced even greater challenges. Google hypothesized that, when individuals or their family members got the flu, they went looking on the internet–via Google, of course–for medical advice. In a highly cited paper, Ginsberg et al. (2008) showed that they could predict region-specific influenza infection rates in the United States using Google search frequency data. Here was the true promise of big data–that we capitalize on data to better understand the world around us.

The subsequent history of Google Flu illustrates the shortcomings of the first big data revolution. Google Flu has failed twice since its launch. The patterns and reasons for failure speak to the limits of prediction. In 2009, Google Flu under-predicted flu rates during the H1N1 pandemic. Post-mortem analysis suggested that the different viral characteristics of H1N1 compared with garden-variety strains of influenza likely meant that individuals didn't know they had a flu strain, and thus didn't go looking for flu-related information (Cook et al., 2011). Conversely, in 2012, Google Flu over-predicted influenza infections. Google has yet to discuss why, but speculation has centered on the intensive media coverage of an early-onset flu season, which may have sparked interest in the flu among healthy individuals (Butler, 2013).

The problems experienced Google Flu provides a particularly acute warning of the risks inherent in trying to predict what will happen in the real world based on the exhaust of the digital one. Google Flu relied on a map–a mathematical relationship between online behavior and real-world infection. Google built that map on historic patterns of flu infection and search. It assumed that such patterns would continue to hold in the future. But there was nothing fundamental about those patterns, and Google had limited capacity to check whether they continued to hold, or to update its map in real time. Either a change in the physical world–a new virus–or the virtual one–media coverage–were enough to render the map inaccurate. The CDC's old reporting networks out-performed big data when it mattered most.

3 A revolution constrained: data, potential, and value creation

Despite ostensibly free raw materials, mass-manufacturing insight from digital exhaust has thus proven far more difficult than big data's advocates would let on. It's thus unsurprising that this revolution has had similarly underwhelming effects on business models. Google, Facebook, and Amazon are enormously successful businesses, underpinned by technologies operating at unprecedented scale. But they still rely on centuries-old business models for most of their revenue. Google and Amazon differ in degree, but not kind, from a newspaper or a large department store when it comes to making money. This is a weak showing from a revolution that was supposed to change the 21st century in the way that steam, steel, or rail changed the 19th. Big data has so far made it easier to sell things, target ads, or stalk long-lost friends or lovers. But it hasn't yet fundamentally reworked patterns of economic life, generated entirely new occupations, or radically altered relationships with the physical world. Instead, it remains oddly self-referential: we generate massive amounts of data in the process of online buying, viewing, or socializing; but find that data truly useful only for improving online sales and search.

Understanding how we might get from here to there requires a better understanding of how and why data–big or small–might create value in a world of better algorithms and cheap compute capacity. Close examination shows that firms have largely used big data to improve on existing business models, rather than adopt new ones; and that those improvements have relied data to describe and predict activity in worlds largely of their own making. Where firms have ventured beyond these self-constructed virtual worlds, the data have proven far less useful, and products built atop data far more prone to failure.

3.1 Locating the value in data

The Google Flu example suggests the limits to big data as a source of mass-manufactured insight about the real world. But Google itself, and its fellow big-data success stories, also illustrate the shortcomings of big data as a source of fundamentally new forms of value creation. Most headline big data business models have used their enhanced capacity to describe, predict, or infer in order to implement–albeit at impressive scale and complexity–centuries-old business models. Those models create value not from the direct exchange between consumer and producer, but via a web of transactions several orders removed from the creation of the data itself. Categorizing today's big data business models based on just how far they separate data generation from value creation quickly illustrates how isolated the monetary value of firms' data is from their primary customers. Having promised a first-order world, big data has delivered a third-order reality.

Realizing the promise of the big data revolution will require a different approach. The same problems that greeted flu prediction have plagued other attempts to build big data applications that forecast the real world. Engineering solutions to these problems that draw on the potential of cheap computation and powerful algorithms will require not different methods, but different raw materials. The data those materials require must originate from a first-order approach to studying and understanding the worlds we want to improve. Such approaches will require very different models of firm organization than those exploited by Google and its competitors in the first big data revolution.

3.1.1 Third-order value creation: the newspaper model

Most headline big data business models do make much money directly from their customers. Instead, they rely on third parties–mostly advertisers–to generate profits from data. The actual creation and processing of data is only useful insofar as it's of use to those third parties. In doing so, these models have merely implemented, at impressive scale and complexity, the very old business model used by the newspapers they have largely replaced.

If we reach back into the dim past when newspapers were viable businesses (rather than hobbies of the civic-minded rich), we will remember that their business model had three major components:

1. Gathering, filtering, and analyzing news

2. Attract readers by providing that news at far below cost
3. Profit by selling access to those readers to advertisers

As the newspaper model matured, the selling of access became more and more refined: newspapers realized that people who read the business pages differed from those who read the front page, or the style section. Front-page ads were more visible to readers than those buried on page A6. Newspapers soon started pricing access to their readers accordingly. Bankers paid one price to advertise in the business section, clothing designers another for the style pages. This segmentation of the ad market evolved as the ad buyers and sellers learned more about whose eyeballs were worth how much, when, and where.

Newspapers were thus third-order models. The news services they provided were valuable in their own right. But readers didn't pay for them. Instead, news was a means of generating attention and data, which was only valuable when sold to third parties in the form of ad space. Data didn't directly contribute to improving the headline product–news–except insofar as it generated revenue could be plowed back into news gathering. The existence of a tabloid press of dubious quality but healthy revenues proved the weakness of the link between good journalism and profit.

From a value creation perspective, Google, Yahoo, and other ad-driven big data businesses are nothing more than newspapers at scale. They too provide useful services (then news, now email or search) to users at rates far below cost. They too profit by selling access to those users to third-party advertisers. They too accumulate and use data to carve up the ad market. The scale of data they have available, of course, dwarfs that of their newspaper ancestors. This data, combined with cheap computation and powerful statistics, has enabled operational efficiency, scale, and effectiveness far beyond what newspapers could every have managed. But the business model itself–the actual means by which these firms earn revenues–is identical.

3.1.2 Second-order value creation: the retail model

Big-box retail ranks as the other substantial success for big data. Large retailers like Amazon, Wal-Mart, or Target have very effectively used data to optimize their supply chain, identify trends and logistical issues ahead of time, and maximize the likelihood of both initial sales and return business from their customers.

But at the end of the day, these businesses remain retailers. Big data may enable them to operate more efficiently. But that efficiency is in service of a model of value generation–retail–that has existed for a very long time. As with Google and ads, big data has enabled these retailers to attain scale and complexity heretofore unimaginable. In doing so, at least some of their profitability has come from market power over suppliers, who lack the access and data the retailers command. But the fundamental means by which they create value is no different than it was fifty years ago.

Retailers are thus second-order big data models. Unlike third-order models, the data they gather has a lot of direct value to the retailer. They don't need to rely on third party purchasers to give the data value. But the actual moneymaking transaction–the retail sale of goods and services–remains separated from the uses of data to improve operational efficiency.

3.1.3 First-order value creation: the opportunity

Second- and third-order models find value in data several steps removed from the actual transaction that generates the data. But, as the Google Flu example illustrated, that data may have far less value when separated from its virtual context. Thus while these businesses enjoy effectively free raw materials, the potential uses of those materials are in fact quite limited. Digital exhaust from web browsing, shopping, or socializing has proven enormously useful in the self-referential task of improving future web browsing, shopping, and socializing. But that success has not translated success at tasks far removed from the virtual world that generated this exhaust. Digital exhaust may be plentiful and convenient to collect, but it offers limited support for understanding or responding to real-world problems.

First-order models, in contrast, escape the Flu trap by building atop purpose-specific data, conceived and collected with the specific intent of solving specific problems. In doing so, they too capitalize on the cheap storage, powerful algorithms, and inexpensive compute power that made the first wave of big data firms possible. But they do so in pursuit of rather different problems.

First order products remain in their infancy. But some nascent examples suggest what might be possible. IBM's Watson famously used its natural language and pattern recognition abilities to win Jeopardy!. But now IBM has adapted Watson to medical diagnosis. By learning from disease and health data gathered from millions of patients, Watson can improve the quality, accuracy, and efficacy of medical diagnosis and service to future patients.[1] Watson closes the data value loop: patient data is made valuable because it improves patient services, not because it helps with insurance underwriting or product manufacturing or logistics or some other third-party service. The diversity of disease and the varied ways in which it can express itself means that Watson's big data capabilities help it improve on human judgment that could never incorporate such data alone.

Premise Corporation[2], provides another example. Premise has built a mobile-phone based data gathering network to generate very granular measurements of macroeconomic aggregates like inflation and food scarcity. This network allows them to monitor economic change at a very detailed level, particularly in regions of the world where official statistics are unavailable or unreliable. This sensor network is the foundation of the products and services that Premise sells to financial services firms, development agencies, and other clients. As compared with the attenuated link between data and value in second- or third-order businesses, Premise's business model links the design of the data generation process directly to the value of its final products

Optimum Energy (OE)[3] provides a final example. OE monitors and aggregates data on building energy use–principally data centers–across building types, environments, and locations. That data enables it to build models for building energy use and efficiency optimization. Those models, by learning building behaviors across many different kinds of inputs and buildings, can perform better than single-building models with limited scope. Most importantly, OE creates value for clients by using this data to optimize energy efficiency and reduce energy costs.

These first-order business models all rely on data specifically obtained for their products. They deploy sensors and data gathering networks with specific hypotheses in mind, and build products directly atop those hypothesis. Watson diagnoses disease with disease data; Premise estimates inflation through specifically-targeted price data; OE instruments data centers and then tunes those centers with the data from those instruments.

This reliance on purpose-specific data contrasts with third-order models that rely on data gathered for purposes wholly unrelated to the task at hand–the "digital exhaust" of conventional big data wisdom. To use the newspaper

[1] See Steadman (2013) for early results of experiments showing that Watson can improve the accuracy of cancer diagnoses.

[2] See `http://premise.is/`.

[3] See `http://optimumenergyco.com/`.

example, third-order models assume–but can't specifically verify–that those who read the style section are interested in purchasing new fashions. Hence they sold ads about fashions, rather than stocks or lawnmowers or funeral services. But it was still a guess. Google's success stemmed from closing this information gap a bit–showing that people who viewed web pages on fashion were likely to click on fashion ads. But that gap remains–But again, the data that supports this is data generated by processes unrelated to actual purchasing–activities like web surfing and search or email exchange.

3.1.4 Concept blurring

We should not overstate the separation among these models. Google arguably contains all three: a third-order ad business model that relies, in part, on products dependent on first-order goods like email with effective spam detection or maps with StreetView, tuned with second-order processes to optimize customer satisfaction with those products. Many retailers–particularly vertically-integrated manufacturer-retailers– gather customer data specifically to improve the quality or variety of the products they sell.

But the distinction remains valid insofar as we're waiting for a big data revolution that does more than sell us ads or optimize operational efficiency. Most of the most promising products–the genetic, medical, economic, and social products that will directly improve our lives, rather than just serve to sell us stuff–rely on the hypothesis that big data will permit new forms of first-order value generation. Yet to date it has not. We should ask why.

3.2 The unrealized promise of unreasonable data

We should remember the root of the claim about big data. That claim was perhaps best summarized by Halevy et al. (2009) in what they termed "the unreasonable effectiveness of data". Most have taken that to mean that data–and particularly more data–are unreasonably effective *everywhere*–and that, by extension, even noisy or skewed data could suffice to answer hard questions if we could simply get enough of it. But that mis-states the authors' claims. They did not claim that more data was always better. Rather, they argued that, for specific kinds of applications, history suggested that gathering more data paid better dividends than inventing better algorithms. Where data are sparse, or the phenomenon under measurement noisy, more data allow a more complete picture of what we are interested in. Machine translation provides a very pertinent example: human speech and writing varies enormously within one language, let alone two. Faced with the choice between better algorithms for understanding human language, and more data to quantify the variance in language, more data appears to work better.[4] But for other applications, the "bigness" of data may not matter at all. If I want to know who will win an election, polling a thousand people might be enough. Relying on the aggregated voices of a nation's Twitter users, in contrast, will probably fail (Gayo-Avello et al., 2011; Gayo-Avello, 2012; Huberty, 2013b). Not only are we are not in the "N=all" world that infatuated Mayer-Schönberger and Cukier (2013); but for most problems we likely don't care to be. Having the right data–and consequently identifying the right question to ask beforehand–is far more important than having a lot of data of limited relevance to the answers we seek.

These limits on the usefulness of third-order data point to real limits to the progress that today's big data revolution will make. The raw materials might well be free, thrown off from business operations that will occur anyway. But that data offers little support for the utopian dreams of a big data world. Realizing those dreams will require both the first-order data they require, and a range of fundamentally new business models capable of investing in, sustaining, and transforming that data to realize truly innovative products and services.

4 A loom, not a model T

Big data therefore falls short of proclamation that it represents the biggest change in technological and economic possibility since the industrial revolution. That revolution, in the span of a century or so, fundamentally transformed almost every facet of human life. An English peasant living in 1800 enjoyed relatively few advantages over his Roman predecessor of eighteen centuries prior. Textiles were still woven by hand; foodstuffs were perhaps reliable, but not easily stored nor necessarily very diverse; steel was valuable but expensive to make; animal power provided most non-human labor. Water was not safe to drink, particularly in urban areas. Transportation was

[4] Not everyone is convinced. Peter Norvig, head of research at Google, had a very public dispute with the linguist Noam Chomsky over whether progress in machine translation contributed anything at all to our understanding of human language. See `http://norvig.com/chomsky.html` for Norvig's account of this dispute and a link to Chomsky's position.

slow, and impossible in some seasons.

This peasant's Edwardian great grandchildren knew a very different world: of rapid transportation over air, sea, and land; of plentiful and cheap steel; of diverse and easily stored foodstuffs; of machine-made textiles; of municipal sewers that purged cholera from the cities; and of a world where the modern medicine was rapidly rendering the scourges of polio, smallpox, measles, and malaria treatable, if one contracted them at all.

Having ranked big data with the industrial revolution, we find ourselves wondering why our present progress seems so paltry in comparison. Google differs only in scale, but not mode of value creation, from a newspaper. Target still earns money by retail, not through some novel manipulation of nature or man. The fundamental advances in computation required to make these processes possible were largely made during and shortly after the Second World War.

We are in the position of someone who, in 1840, having been promised a Model T, looks around and sees only looms. Those looms were, to be sure, better than hand weaving. They made cloth cheaper, clothes more plentiful. But the innovations that turned the hand weaving of 1815 into the power looms of 1830 weren't that radical. They were mostly water-powered–a far cry from the giant steam- or electricity-driven factories of the late 19th century. The cloth they made wasn't dramatically different than what had been woven by hand–more plentiful and cheaper, to be sure, but not substantially different.

Much of the value of the first industrial revolution came from such differences of degree, rather than kind. The advances in organic chemistry, rapid personal transportation, shipping, air travel, pain medication, and other products had to wait for the second industrial revolution. That revolution saw the emergence of fundamentally different kinds of firms. Rather than improve on or scale up pre-existing processes, these firms invested in huge industrial research and development operations to discover and then commercialize new scientific discoveries. Those firms were matched by significant government investment in basic research and development, particularly in the United States and Germany. The talented tinkerers, craftsmen, and inventors that built the first power looms were replaced by the trained engineers, scientists, and managers of the major industrial concerns. They, in turn, drew on the much more abstract advances in our fundamental understanding of physical, chemical, and information processes. These changes were expensive, complicated, and slow–so slow that John Stuart Mill despaired, as late as 1871, of human progress. But in time, they produced a world inconceivable to even the enthusiasts of the 1840s.

5 Consequences

We have a loom, but we envision the possibility of a Model T. Today, we can see glimmers of that possibility in IBM's Watson, Google's self-driving car, Nest's adaptive thermostats, and other technologies deeply embedded in, and reliant on, data generated from and around real-world phenomena. None rely on "digital exhaust". They do not create value by parsing customer data or optimizing ad click-through rates (though presumably they could). They are not the product of a relatively few, straightforward (if ultimately quite useful) insights. Instead, IBM, Google, and Nest have dedicated substantial resources to studying natural language processing, large-scale machine learning, knowledge extraction, and other problems. The resulting products represent an industrial synthesis of a series of complex innovations, linking machine intelligence, real-time sensing, and industrial design. These products are thus much closer to what big data's proponents have promised–but their methods are a world away from the easy hype about mass-manufactured insights from the free raw material of digital exhaust.

6 Dystopia

Each also, in its way, illustrates the dark side of the big data revolution. Big data and the Silicon Valley culture in which it emerged have suffered whithering criticism for overly optimistic techno-utopianism.[5] We forget too easily that social unrest, instability, and conflict were fellow-travelers with the industrial progress of the 19th century. Today, the promise of big data brings with it at least three dystopias, all visible even in today's early stages of technological change. Securing the gains of industrialization required careful choices about politics and policy. Big data will require no less.

[5] See here in particular Alex Payne's "Letter to a Young Programmer", and his related talks, focusing on the insularity of modern investing and its extreme risk-aversion. See `https://al3x.net/2013/05/23/letter-to-a-young-programmer.html`.

6.1 The Benthamite dystopia

> Yearly reminder: unless you're over 60, you weren't promised flying cars. You were promised an oppressive cyberpunk dystopia. Here you go.
>
> Kyle Marquis, via Twitter

First, a dystopia of abuse. During the first industrial revolution, Jeremy Bentham proposed that we solve the prison problem with technology, rather than better social policy. His Panopticon "[ground] rogues honest" with minimal human effort, allowing the jailers to watch their inmate charges while leaving the inmates unawares. Attempts to implement this innovation proved impractical with 19th century technology.

But the confluence of cloud computing and big data services have simplified the problem. The centralization of diverse data about large numbers of people onto the servers of a limited number of firms–usually for reasons having nothing to do with surveillance per se–has made the Panopticon trivial. Indeed, the very technologies that permit that data to serve better ads, predict consumer wants, or anticipate new trends are equally well suited to a mass surveillance state.

Sadly, it appears such a dystopia has crept up upon us, at least in the United States. The whistleblower disclosures about the NSA PRISM program portray a sophisticated effort to access, mine, and act on the communications traffic of correspondence between foreigners and American citizens. Whether by design or accident, that program spied on American citizens, possibly in violation of laws that prohibit the NSA from doing so. Congressional attempts to defund or otherwise curtail the program have, as of late 2013, been sparse and unsuccessful.

Hence, in the US and barring a change of law, "big data" business models focused on people and behavior–the second and third order models in particular–will likely find themselves unwilling participants in system that many US citizens regard as a gross violation of their right to privacy. Citizens and governments overseas may think twice before doing business with these firms, or corresponding with those who do.[6] Like Nobel and dynamite,

[6] This has already begun. The European Union has advised companies looking for cloud services to consider non-US providers. Early reports estimated that the US could

we marvel at the use of our new technologies for good, only to find them equally well suited for ill.

6.2 The Downton Abbey dystopia

Second, a dystopia of under-use. The origins of much of what we now know as big data lie in another era of artificial intelligence. And while we remain far from replicating consciousness on a chip, we've done quite well abstracting, standardizing, and automating what formerly were human activities. This process began with physical tasks like manufacturing[7] But it has now invaded formerly safe white-collar professions like accounting or legal services. Amazon has purchased a company that builds robots to automate its distribution centers–moving automation further up the retail value chain. Self-driving cars may, by the middle of the 21st century, replace taxis. The ability to learn and replicate a larger and larger share of both regular and (ostensibly) irregular means that a greater portion of human labor can now be automated. Like the first industrial revolution, that generates huge returns for those doing the automating. But it ravages employment for those being automated.

We have, of course, been through this before. The "creative destruction" of successive technological revolutions took us from a world where most of the population worked in agriculture, to one where most of it worked in factories, to one where most are now employed in services of some form. In each case, these transitions caused real hardship for those whose jobs were rendered superfluous. But they ultimately made all of society better off.

The combination of big data and algorithms threatens to break this chain of improvement. The industrial revolution endowed machines with the capacity to perform repetitive physical tasks. To do so, those machines embodied the knowledge of their human designers and operators. But the operators remained a critical part of the process, forming the crucial link between abstract engineering and physical production. Big data now appears poised to displace the operators, in both physical tasks and repetitive white collar

lose $35-45 billion dollars in contracts by 2015 due to foreigners withdrawing for PRISM-related reasons.(Babcock, 2013). Brazil has openly stated the intent to insulate its internal communications from traffic passing through US networks (BBC, 2013).

[7] The old joke about the factory applies here. A perfect factory is said to require only a man and a dog to staff it. The man feeds the dog, and the dog makes sure the man doesn't touch anything.

work. As our capacity to abstract mental tasks into codified steps improves, a larger and larger share of formerly middle-class tasks will become the purview of machines.

What would remain of the people that used to perform this work? The danger may lie in a kind of Downton Abbey society: one where, as in late Victorian and early Edwardian England, a few very wealthy individuals who can command the personal attention of a large servant class.[8] The Downton Abbey version of this world isn't necessarily so bad. Unlike the reality of Edwardian domestic service, which could be quite difficult and cruel, Downton Abbey supplied a vision of a relatively benevolent upstairs class, and a sensitive and thoughtful, if less well-off, downstairs class. Lord Grantham may be a landed aristocrat, but at least within his household he comes off as responsive to the needs of his staff and interested in the development of his local community–so long as it doesn't require voting Labour. Such households surely existed, but were far from the norm.

These changes may extend far beyond white collar middle management. As Rao (2012) has argued, the second industrial revolution transformed entrepreneurship as well. Where the first phase had seen the proliferation of small-scale entrepreneurship, the second replaced small owner-managers with a professionalized managerial class in large industrial concerns. He argues that entrepreneurship is going through just such a transformation. Vast riches from startups are, and always have been, rare. Instead, venture capital sees startup exit via acquisition–into a large and stable firm like Google or Facebook–as the preferred path for most investments. Those acquisitions turn entrepreneur-managers into product managers for one small niche of a large enterprise. In doing so, it mitigates some startup risk. But, as Rao points out, it also puts these entrepreneurs in the position of being, effectively, labor. Nowhere is this more apparent than in commodity startup incubators like YCombinator, for which entrepreneurship is a pure volume play.

[8] Lindert and Williamson (1983) and Lindert (2000) show that income inequality in industrial England worsened over the course of the 19th century, before beginning a correction in the early 20th. Inequality peaked in the late 1860s and early 1870s. The subsequent decline of the English gentry and their Downton Abbey existence occurred in part due to the growing expense of servant labor.

6.3 The Lehman dystopia

Finally, a dystopia of misuse. Finance has a long lead on first-order business models for big data. Financial firms interest in data to inform more, faster, or better trading dates to well before computers–the house of Rothschild profited immensely from having received early warning, courtesy carrier pigeons, of the outcome of the Battle of Waterloo. More recently, financial firms have led their peer institutions in investing in information technology and sensor systems to improve the quality, quantity, and speed of data gathering. Entire classes of products–from high-frequency trading to derivatives and crop insurance–rely on such data and data analysis capabilities.

Yet in 2008, this system, for all its analytic prowess, utterly failed. For a few days that September, normally sober people in high office openly wondered whether the financial system would collapse and take the rest of the economy with it. Several years later, JP Morgan Chase found itself liable for a billion dollars in regulatory fines, courtesy a faulty risk management model that exposed the bank to several billion dollars in losses.(Kopecki, 2013) Around the same time, Knight Capital lost a half-billion dollars in less than an hour consequence of accidentally using out-of-date computer models for high-frequency trading (Strasburg and Bunge, 2012). More information, faster, does not seem to brought stability to even the largest and most sophisticated institutions.

Most of the blame here lies in a combination of failed macro-prudential regulation and very cheap money. But the misuse of data plays an important part as well. Firms, having built sophisticated models of risk atop reams of historical data, thought themselves protected against the ravages of the market. Yet the very data that enabled their products had simultaneously blinded them to their extraordinary risk exposure. Whether the ensuing failures were the product of willful misdeeds or naïve assumptions does not avoid the fact that data lay at their core.

The Lehman dystopia is thus a dystopia of misuse: of data-driven models that fail to deliver on their promise, consequence of poorly-designed, badly-executed ideas. It's a dystopia wherein organizations want the upside of big data without thinking too hard about where it might go wrong. And it's a dystopia of yes-men, where data becomes a means of confirming one's own biases–usually those coincident with short-term profits–rather than a means of understanding and operating in the world.

The danger here is already visible. Many companies now claim to offer "big data" solutions that cut out the messy and expensive step of hiring statisticians and engineers–and the inconvenience of listening to them afterwards.[9] Instead, they promise to put big data in the directly in the hands of day-to-day business operations. But doing so is the equivalent of asking for the outputs of an industrial research lab without the investment. These tools can no more turn marketing people into data scientists than a chemistry set would have turned a 19th century craftsman into an industrial dye chemist. These business models–data science as a service, if you will–will almost certainly fail substantively, even if they profit in the short term.. We can only hope that they don't, as Lehman did, cause widespread damage along the way.

These dystopias each have their precedent in the industrial revolution to which big data is the supposed successor. Thyssen, Krupp, and IG Farben were innovative industrial firms as well as the incubators of the *Wehrmacht* and Zyklon B. The early years of the industrial revolution saw widespread wage declines and enormous displacement of agricultural labor into the cities. We should remember that the Luddites weren't idle layabouts, but skilled craftsmen angry at the displacement of their hard-won knowledge and ability. Consumer products were a font of unreliability, danger, and malfunction before product liability, testing, and standards emerged to prohibit the worst of the snake oils and help verify the rest.

7 Awaiting the second big data revolution

We're stuck in the first industrial revolution. We have the power looms and the water mills, but wonder, given all the hype, at the absence of the Model Ts and telephones of our dreams. We wonder still at why our attempts to apply the equivalent of water power or steam engines to these dreams so consistently fall short. The answer is a hard one. The big gains from big data will require a transformation of organizational, technological, and economic operations on par with that of the second industrial revolution. Then, as now,

[9] "Big data" tools range from generic technologies designed to handle data, such as Hadoop; to tools highly tuned to specific problems, such as HortonWorks' server log processing technology. These tools facilitate, but don't promise to replace, human judgement in quantitative inference. Other companies, such as Tableau, promise to make big data accessible through interfaces that look like common spreadsheets. In an essay that bordered on parody, Mehta (2013) claimed that these tools would commodify the production of quantitative knowledge.

firms had to invest heavily in industrial research and development to build the foundations of entirely new forms of value creation. Those foundations permitted entirely new business models, in contrast to the marginal changes of the first industrial revolution. And the raw materials of the first revolution proved only tangentially useful to the innovations of the second.

These differences portend a revolution of greater consequence and complexity. Firms will likely be larger. Innovation will rely less on small entrepreneurs, who lack the funds and scale for systems-level innovation. Where entrepreneurs do remain, they will play far more niche roles. The success of systems-level innovation will threaten a range of current jobs–white collar and service sector as well as blue collar and manufacturing–as expanding algorithmic capacity widens the scope of digitizeable tasks. But unlike past revolutions, that expanding capacity also begs the question of where this revolution will find new forms of employment insulated from these technological forces; and if it does not, how we manage the social instability that will surely follow. With luck, we will resist the temptation to use those same algorithmic tools for social control. But human history on that point is not encouraging.

Regardless, we should resist the temptation to assume that a world of ubiquitous data means a world of cheap, abundant, and relevant raw materials for a new epoch of economic prosperity. The most abundant of those materials today turn out to have limited uses outside the narrow products and services that generate them. Overcoming that hurdle requires more than just smarter statisticians, better algorithms, or faster computation. Instead, it will require new business models capable of nurturing both new sources of data and new technologies into truly new products and services.

References

Babcock, C. (2013). Nsa's prism could cost u.s. cloud companies $45 billion. *Informationweek*, 15 August.

BBC (2013). Brazil data plan aims to keep us spies at bay. *www.bbc.co.uk*, 18 September.

Boyd, D. and Marwick, A. (2011). Social privacy in networked publics: Teens attitudes, practices, and strategies.

Butler, D. (2013). When google got flu wrong. *Nature*, 494(7436):155.

Conover, M. D., Ratkiewicz, J., Francisco, M., Goncalves, B., Flammini, A., and Menczer, F. (2011). Political polarization on twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Cook, S., Conrad, C., Fowlkes, A. L., and Mohebbi, M. H. (2011). Assessing google flu trends performance in the united states during the 2009 influenza virus a (h1n1) pandemic. *PLoS One*, 6(8):e23610.

DiGrazia, J., McKelvey, K., Bollen, J., and Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. Working paper, Indiana University.

Gayo-Avello, D. (2012). I wanted to predict elections with twitter and all I got was this lousy paper: a balanced survey on election prediction using twitter data. *arXiv preprint arXiv:1204.6441*.

Gayo-Avello, D., Metaxas, P. T., and Mustafaraj, E. (2011). Limits of electoral predictions using twitter. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM) 2011, July 17*, volume 21, page 2011.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2008). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.

Halevy, A., Norvig, P., and Pereira, F. (2009). The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24(2):8–12.

Huberty, M. (2013a). Multi-cycle forecasting of congressional elections with social media. In *Workshop on Politics, Elections, and Data (PLEAD) 2013*.

Huberty, M. (2013b). Multi-cycle forecasting of congressional elections with social media. In *Proceedings of the 2nd workshop on Politics, Elections, and Data*, pages 23–30.

Kopecki, D. (2013). Jpmorgan pays $920 million to settle london whale probes. *Bloomberg.com*, 19 September.

Lindert, P. H. (2000). Three centuries of inequality in britain and america. *Handbook of income distribution*, 1:167–216.

Lindert, P. H. and Williamson, J. G. (1983). Reinterpreting britain's social tables, 1688–1913. *Explorations in Economic History*, 20(1):94–109.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity.* Mckinsey & Company.

Mayer-Schönberger, V. and Cukier, K. (2013). *Big Data: A Revolution that Will Transform how We Live, Work, and Think.* Eamon Dolan/Houghton Mifflin Harcourt.

Mehta, C. (2013). Commoditizing data science. *CloudAve*, 15 February.

Rao, V. (2012). Entrepreneurs are the new labor. *Forbes.com*, 3 September.

Steadman, I. (2013). Ibm's watson is better at diagnosing cancer than human doctors. *Wired UK*, 11 February.

Strasburg, J. and Bunge, J. (2012). Loss swamps trading firm: Knight capital searches for partner as tab for computer glitch hits $440 million. *The Wall Street Journal*, 2 August.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2014). Forecasting elections with non-representative polls. *Forthcoming, International Journal of Forecasting.*