J Ind Compet Trade (2015) 15:35–47
DOI 10.1007/s10842-014-0190-4

# Awaiting the Second Big Data Revolution: From Digital Noise to Value Creation

**Mark Huberty**

Received: 14 April 2014 / Revised: 13 September 2014 /
Accepted: 3 December 2014 / Published online: 18 February 2015


**Abstract** "Big data"—the collection of vast quantities of data about individual behavior via online, mobile, and other data-driven services—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.

**Keywords** Big data · Digitalization · Value creation · Business models · Technological change

**JEL Classification** C80 · L86 · O33 · M15

## 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 bring about the same scope of change to economic and social prosperity in the 21st century, that rail, steam, or steel did in the 19th.

M. Huberty (✉)
Berkeley Roundtable on the International Economy, Berkeley, CA, USA
e-mail: mark.huberty@gmail.com

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 computational processing power and data 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. As with its 19th century predecessor, this revolution will alter how we govern, work, play, and live. But unlike the 19th century, we are told, the raw materials driving this revolution are so cheap and abundant that 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 dubious assumptions about the origins and uses of data, none of which hold up to serious scrutiny. In aggregate, these assumptions all fail to address whether the data we have actually provides the raw materials needed for a data-driven Industrial Revolution we need. Taken together, these failures point out the limits of a revolution built on the raw materials that today seem so abundant.

Four of these assumptions merit special attention: First, $N = all$, or the claim that our data allow a clear and unbiased study of humanity; second, that *today* = *tomorrow*, or the claim that understanding online behavior today implies that we will still understand it tomorrow; third, *offline* = *online*, the claim that understanding online behavior offers a window into economic and social phenomena in the physical world; 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 in sectors well beyond social and media markets. 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; tracking the same sample over many years required organizations and budgets that few organizations outside governments could justify. 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 and hope that this proved adequate. 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, they claim, 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. With this new database of what people are saying or buying, where they go and when, how their social networks change and evolve, and myriad other factors, the prior restrictions borne of the cost and complexity of sampling will melt away.

But $N \neq all$. Most of the data that dazzles those infatuated by "big data"—Mayer-Schönberger and Cukier included—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. In practice, this distinction turns out to matter quite a lot. The demographics of any given online service usually differ dramatically from the population at large, whether we measure by age, gender, race, education, and myriad other factors.

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. The fact that we have enormous quantities of data about these people may not prove very useful to understanding society writ large.

### 2.2 Today = 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, we only 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. Moreover, we have no way of systematically checking how biased the sample is, without resorting to traditional survey methods and polling—the very methods that big data is supposed to render obsolete.

### 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 of big data enthusiasts—suffer 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.

Furthermore, even where online behavior may echo offline preferences or beliefs, that echo is often very weak. In statistical terms, our ability to distinguish "significant" from "insignificant" results improves with the sample size—but statistical significance is not actual significance. Knowing, say, that a history of purchasing some basket of products is associated with an increased risk of being a criminal may be helpful. But if that association is weak—say a one-hundredth of a percent increase—it's practical import is effectively zero. Big data may permit us to find these associations, but it does not promise that they will be useful.

### 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 school student 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.

An anonymous reviewer pointed out, correctly, that social researchers have always faced this problem. This is certainly true but many of the features of social systems—political and cultural institutions, demography, and other factors—change on a much longer timeframe than today's data-driven internet services. For instance, US Congressional elections operate very differently now compared with a century ago; but change little between any two elections. Contrast that with the pace of change for major social media services, for which 2 years may be a lifetime.

A recent controversy illustrates this problem to a T. Facebook recently published a study (Kramer et al. 2014) in which they selectively manipulated the news feeds of a randomized sample of users, to determine whether they could manipulate users' emotional states. The revelation of this study prompted fury on the part of users, who found this sort of manipulation unpalatable. Whether they should, of course, given that Facebook routinely runs experiments on its site to determine how best to satisfy (i.e., make happier) its users, is an interesting question. But the broader point remains—someone watching the emotional state of Facebook users might have concluded that overall happiness was on the rise, perhaps consequence of the improving American economy. But in fact this increase was entirely spurious, driven by Facebook's successful experiment at manipulating its users.

Compounding this problem, we cannot know, in advance, which of the social and technological changes we do know about 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—with no adjustments for demography, or spam, or even name confusion—can forecast whether that candidate will win his or her election. This is a purely "digital exhaust" approach. They speculate—but cannot know—whether this approach works because (to paraphrase their words) "one tweet equals one vote", or "all attention on Twitter is better". Moreover, it turns out that the predictive performance of this simple model provides no utility. As Huberty (2013) shows, their estimates perform no better than an approach that simply guesses that the incumbent party would win—a simple and powerful predictor of success in American elections. Big data provided little value.

Contrast this with Wang et al. (2014). They use the Xbox gaming platform 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, 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. The difference? DiGrazia et al. (2013) sought to reuse data created for one purpose in order to do something entirely different; Wang et al. (2014) set out to gather data explicitly tailored to their purpose alone.

### 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. they represent themselves online consistent with how they behave offline, and;
4. 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, shaped by a complex and usually opaque amalgam of social norms and algorithmic influences.

But if each of these statements fall down, then how have companies like Amazon, Facebook, or Google built 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, these populations are subject to change their behaviors, their self-representation, or their expectations at any point. But at least Google or Amazon can plausibly claim to have a valid sample of the primary populations they care 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. 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. So long as Google does better than its competitors at targeting advertising, it will continue to win the competitive fight for advertising dollars.

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 2 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 virtual data to better understand, and react to, the physical world around us.

The subsequent history of Google Flu illustrates the shortcomings of the first big data revolution. While Google Flu has performed well in many seasons, it has failed twice, both times in the kind of abnormal flu season during which accurate data are most valuable. The patterns of and reasons for failure speak to the limits of prediction. In 2009, Google Flu underpredicted flu rates during the H1N1 pandemic. Post-hoc 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 by Google Flu provide 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 behavior. It assumed that such patterns would continue to hold in the future. But there was nothing fundamental about those patterns. 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. Amazon, Facebook, and Google 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 on 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 Refining Data into Value

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 not 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 wealthy), we will remember that their business model had three major components:

1. gather, filter, and analyze news;
2. attract readers by providing that news at far below cost, and;
3. profit by selling access to those readers to advertisers.

The market for access matured along with the newspapers that provided it. Both newspapers and advertisers 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 that 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 newsprint ancestors. This data, combined with cheap computation and powerful statistics, has enabled operational efficiency, scale, and effectiveness far beyond what newspapers could ever have managed. But the business model itself—the actual means by which these firms earn revenues—is identical.

Finally, that value model does not emerge, fully-formed, from the data itself. The data alone are no more valuable than the unrefined iron ore or crude oil of past industrial revolutions. Rather, the data were mere inputs to a production process that depended on human insight—that what people looked for on the internet might be a good proxy for their consumer interests.

### *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 harvested very fine-grained data about customer preferences to make increasingly accurate predictions of what individual customers wish to buy, in what quantities and combinations, at what times of the year, at what price. These predictions are occasionally shocking in their accuracy—as with Target's implicit identification of a pregnant teenager well before her father knew it himself, based solely on subtle changes in her purchasing habits.

From this data, these retailers can, and have, built a detailed understanding of retail markets: what products are complements or substitutes for each other; exactly how much more people are willing to pay for brand names versus generics; how size, packaging, and placement in stores and on shelves matters to sales volumes.

Insights built on such data have prompted two significant changes in retail markets. First, they have made large retailers highly effective at optimizing supply chains, identifying retail trends in their infancy, and managing logistical difficulties to minimize the impact on sales and lost competitiveness. This has multiplied their effectiveness versus smaller retailers, who lack such capabilities and are correspondingly less able to compete on price.

But it has also changed, fundamentally, the relationship of these retailers to their suppliers. Big box retailers have increasingly become monopsony buyers of some goods—books for Amazon, music for iTunes. But they are also now monopoly sellers of information back to their suppliers. Amazon, Target and Wal-Mart have a much better understanding of their suppliers' customers than the customers themselves. They also understand these suppliers' competitors far better. Hence their aggregation of information has given them substantial power over suppliers. This has had profound consequences for the suppliers. Wal-Mart famously squeezes suppliers on cost—either across the board, or by pitting suppliers against one another based on detailed information of their comparative cost efficiencies and customer demand.

Hence big data has shifted the power structure of the retail sector and its manufacturing supply chains. The scope and scale of the data owned by Amazon or Wal-Mart about who purchases what, when, and in what combinations often means that they understand the market for a product far better than the manufacturer. Big data, in this case, comes from big business—a firm that markets to the world also owns data about the world's wants, needs, and peculiarities. Even as they are monopsony buyers of many goods (think e-books for Amazon), they are correspondingly monopoly sellers of data. And that has made them into huge market powers on two dimensions, enabling them to squeeze suppliers to the absolute minimum price, packaging, size, and other product features that are most advantageous to them—and perhaps to their customers.

But big data has not changed the fundamental means of value creation in the retail sector. Whatever its distributional consequences, the basic retail transaction—of individuals buying goods from retail intermediaries, remains unchanged from earlier eras. The same economies of scale and opportunities for cross-marketing that made Montgomery Ward a retail powerhouse in the 19th century act on Amazon and Wal-Mart in the 21st. Big data may have exacerbated trends already present in the retail sector; but the basics of how that sector creates value for customers and generates profits for investors are by no means new. Retailers have yet to build

truly new products or services that rely on data itself—instead, that data is an input into a longstanding process of optimization of supply chain relations, marketing, and product placement in service of a very old value model: the final close of sale between a customer and the retailer.

### *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. However, 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 intent of solving specific problems. In doing so, they capitalize on the cheap storage, powerful algorithms, and inexpensive computing power that made the first wave of big data firms possible. But they do so in pursuit of a rather different class of 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 the Jeopardy! game show. Doing so constituted a major technical feat: the ability to understand unstructured, potentially obfuscated Jeopardy! game show answers, and respond with properly-structured questions based on information gleaned from vast databases of unstructured information on history, popular culture, art, science, or almost any other domain.

The question now is whether IBM can adapt this technology to other problems. Its first attempts at improving medical diagnosis appear promising. By learning from disease and health data gathered from millions of patients, initial tests suggest that Watson can improve the quality, accuracy, and efficacy of medical diagnosis and service to future patients (Steadman 2013). 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 activity.

Premise Corporation provides another example. Premise has built a mobile-phone based data gathering network to measure macroeconomic aggregates like inflation and food scarcity. This network allows them to monitor economic change at a very detailed level, 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) 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. This reliance on purpose-specific data contrasts with third-order models that rely on the "digital exhaust" of conventional big data wisdom. To use the newspaper example, third-order models assume—but can't specifically verify—that those who read the style section are interested in purchasing new fashions. 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 again, the data that supports this is data generated by processes unrelated to actual purchasing—activities like web surfing and search or email exchange. And so the gap remains. Google appears to realize this, and has launched Consumer Surveys as an attempt to bridge that gap. In brief, it offers people the chance to skip ads in favor of providing brand feedback.

### 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"—that, when seeking to improve the performance of predictive systems, more data appeared to yield better returns on effort than better algorithms. Most appear to 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 misstates 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. 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 2013). Not only are we not, as section 2 discussed, 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.

## 4 Consequences

Big data therefore falls short of the 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. Someone born in 1860, who lived to be 70 years old, grew up in a world of horses for travel, candles for light, salting and canning for food preservation, and telegraphs for communication. The world of their passing had cars and airplanes, electric light and refrigerators, and telephones, radio, and motion pictures. Having ranked big data with the industrial revolution, we find ourselves wondering why our present progress seems so paltry in comparison.

But much of what we associate with the industrial revolution—the advances in automobile transport, chemistry, communication, and medicine—came much later. The businesses that produced them were fundamentally different from the small collections of tinkerers and craftsmen that built the first power looms. Instead, these firms invested in huge industrial research and development operations to discover and then commercialize new scientific discoveries. 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 industrial enthusiasts of the 1840s.

In today's revolution, we have our looms, 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, or Nest's thermostats that learn the climate preferences of a home's occupants. These and other technologies are 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.

## 5 Towards 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. 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, 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. As Rao (2012) has argued, startups will increasingly become outsourced R&D, whose innovations are acquired to become features of existing products rather than standalone products themselves. 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.

**Acknowledgments** This research is a part of the ongoing collaboration of BRIE, the Berkeley Roundtable on the International Economy at the University of California at Berkeley, and ETLA, The Research Institute of the Finnish Economy. This paper has benefited from extended discussions with Cathryn Carson, Drew Conway, Chris Diehl, Stu Feldman, David Gutelius, Jonathan Murray, Joseph Reisinger, Sean Taylor, Georg Zachmann, and John Zysman. All errors committed, and opinions expressed, remain solely my own.



## References

Boyd D, Marwick AE (2011) Social privacy in networked publics: teens' attitudes, practices, and strategies. In: A decade in internet time: symposium on the dynamics of the internet and society. pp 1–29

Butler D (2013) When Google got flu wrong. Nature 494(7436):155

Conover MD, Ratkiewicz J, Francisco M, Goncalves B, Flammini A, 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 AL, Mohebbi MH (2011) Assessing Google Flu trends performance in the United States during the 2009 influenza virus a (H1N1) pandemic. PLoS One 6(8):1–8

DiGrazia J, McKelvey K, Bollen J, Rojas F (2013) More tweets, more votes: social media as a quantitative indicator of political behavior. PLoS ONE 8(11):1–5

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 PT, Mustafaraj E (2011) Limits of electoral predictions using Twitter. In: Proceedings of the International Conference on Weblogs and Social Media (ICWSM) 21:2011

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2008) Detecting influenza epidemics using search engine query data. Nature 457(7232):1012–1014

Halevy A, Norvig P, Pereira F (2009) The unreasonable effectiveness of data. Intell Syst IEEE 24(2):8–12

Huberty M (2013) Multi-cycle forecasting of congressional elections with social media. In: Proceedings of the 2nd Workshop on Politics, Elections, and Data (PLEAD), pp 23–30

Kramer A, Guillory J, Hancock J (2014) Experimental evidence of massive-scale emotional contagion through social networks. Proc Natl Acad Sci 111(24):8788–8790

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute Report

Mayer-Schönberger V, Cukier K (2013) Big data: a revolution that will transform how we live, work, and think. Eamon Dolan/Houghton Mifflin Harcourt

Rao V (2012) Entrepreneurs are the new labor. Forbes. http://www.forbes.com/sites/venkateshrao/2012/09/03/entrepreneurs-are-the-new-labor-part-i/. Accessed 3 Sept 2014

Steadman I (2013) IBM's Watson is better at diagnosing cancer than human doctors. Wired UK, February 11th

Wang W, Rothschild D, Goel S, Gelman A (2014) Forecasting elections with non-representative polls. Int J Forecast Forthcoming