Routledge
Taylor & Francis Group

Consumption Markets & Culture

ISSN: 1025-3866 (Print) 1477-223X (Online) Journal homepage: https://www.tandfonline.com/loi/gcmc20

Big data

John Deighton

To cite this article: John Deighton (2019) Big data, Consumption Markets & Culture, 22:1, 68-73, DOI: 10.1080/10253866.2017.1422902

To link to this article: https://doi.org/10.1080/10253866.2017.1422902

CONSUMPTION MARKETS & CULTURE
2019, VOL. 22, NO. 1, 68–73
https://doi.org/10.1080/10253866.2017.1422902

Routledge
Taylor & Francis Group

Big data

John Deighton

Harvard Business School, Boston, MA, USA

ABSTRACT

Big data is defined and distinguished from a mere moment in the "ancient quest to measure." Specific discontinuities in the practice of information science are identified which, the paper argues, have large consequences for the social order. The infrastructure that runs on big data is described as diffusing with unprecedented speed but as being difficult to analyze and critique, and therefore the designers of society's big data infrastructure, whether human or machines, play an unacknowledged legislative function of great consequence.

ARTICLE HISTORY
Received 13 July 2017
Accepted 26 December 2017

KEYWORDS
Big data; digital infrastructure; privacy; algorithm; data generators; marketplace icon

Introduction

This paper is a contribution to a series on, as the editor has put it, aspects of the marketplace so basic that we cannot imagine living without them. He calls them marketplace icons, and I shall take icons here to refer to representations receiving uncritical devotion.

Is[1] big data an icon? Are we devoted to it? I believe the answer is beyond dispute yes: not exactly that we are devoted to the thing itself, but that we embrace with remarkably little skepticism or caution the apps and engines that run on big data. When the thing itself is discussed in popular discourse, granted, it is not viewed uncritically. It gets a full range of treatments: dismissed as mere hype, condemned as a curse of our age, and analyzed, usually breathlessly, as the foundation of our economic future. And yet, modern life takes for granted that data, deployed at a scale inconceivable two decades ago, will let us drive the devices of modern life the way that bridges and roads let us drive motor vehicles. It is the results generated by devices, apps, and software, rather than the big data that they run on, that pass without much scrutiny.

To illustrate this blithe trust, consider a Google search. When a person searches the Internet, he or she takes for granted that the response will come from a trove of data so vast as to be indistinguishable from everything. The trust extends further to accepting without incredulity that the response comes after examining all the information on every page of the trove, even though Google announces at the head of the results page that it has performed the search in a decimal fraction of a second.

There are many examples of uncritical devotion to data-driven outcomes: we trust technology to assemble a page of the digital *New York Times* with editorial matter that is indifferent to our tastes and advertising matter that is specific to our tastes. We trust a ride-hailing app to match us instantly to the closest driver while in the same moment matching hundreds of thousands of other customers to their best driver options. We trust Facebook to link us to our little circle of friends from among the almost 30% of the earth's population that belongs to Facebook. We date people recommended to us by apps and sometimes marry them.

As a society, we are dimly aware that databases were not always so large, and data retrieval was not always so fast or so unerringly accurate, but over the past two decades, a sense of wonder has given

CONTACT John Deighton jdeighton@hbs.edu

[1]Yes, data is a plural noun, but not in this essay nor in any writing that wants not to come off as priggish.

way to a taken-for-grantedness. We do not cheer new additions to the data trove as earlier ages celebrated the mapping of new geographies. And yet the trove is growing, not slowing down. My own studies over eight years, of the revenue generated by firms that make up the Internet, find that the Internet is growing at an average of 20% per year, a slower rate in 2008 and slightly faster in 2016, as if the thing is still catching on.

The discontinuity that distinguishes big data from data

Not everyone agrees that big data is something discrete from mere data. Mayer-Schonberger and Cukier (2013, 78) contend that "datafication," the tendency to look for measurable quantities in a phenomenon, tabulate them and analyze them, long predates digitization. They call digitization simply, "a continuation of humankind's ancient quest to measure, record, and analyze the world." But I assert that digitization has led to something that is truly an economic and cultural discontinuity. At its most reductive, the discontinuity enables a fourth technological foundation for civilization: stone tools, agriculture, industry, and now the age of thinking machines. In less hyperbolic terms, former Treasury Secretary Lawrence Summers has said, "Data may be to the twenty-first century economy what oil was to the 20th, a hugely valuable asset essential to economic life … " Others call it "surveillance capitalism." Zuboff (2015, 75) warned of a "largely uncontested new expression of power," which "effectively exile(s) persons from their own behavior, while producing new markets of behavioral prediction and modification."

There is a technical marker of the discontinuity: a new way to store and retrieve data made possible by the conjunction of two technical leaps, the Hadoop distributed file system, and a protocol for analysis known as MapReduce. Unduly technical perhaps, but the marker serves to prop up the claim that the move from then to now is a jump, not a smooth glide.

In the beginning was the flat file. An Excel spreadsheet serves to illustrate. A flat file can store data and row and column labels can define the data, but retrieval is cumbersome. So, the second step in sophistication was the addition of a key, which made retrieving a row a matter of searching on the key. Student grades are commonly stored in Excel files with the student name serving as the key. The third step in sophistication was the relational database, a collection of tables each sharing common keys. So, for example student grades for many courses can be stored in many tables. As long as the student name continues to serve as the common key, a query can be written to report a particular student's grade in many courses and semesters. But the time to answer a query expands as the size of the database expands. The Big Data era started when the commonsense law, that time to find a needle in a haystack depends on the size of the haystack, stopped being true.

It stopped first because of Hadoop, developed at Yahoo in about 2003, which allowed unlimited amount of data to be stored on inexpensive servers at multiple locations around the world with replication. That in itself was no advantage as long as there was no way to assemble unlimited amount of data, so the second step was "crawling" all the sites on the Web and then indexing them, all in background. Ghemawat, Gobioff, and Leung in 2003 published "The Google File System" to implement the hypertextual search engine algorithm published by Brin in 1998 (Ghemawat, Gobioff, and Leung 2003). Its job was to crawl the entire World Wide Web as it then was, and assemble an index of every word on every page of every site. One circuit took up to four weeks, so to continue the haystack analogy, an index existed to show where to find the thing most like a needle, but the representation of the haystack could be up to four weeks out of date. For most purposes (finding the names of local restaurants, for example, but not for representing the cash balances in bank accounts), this picture of the haystack was quite satisfactory. Thus, were the limitations to the scaling of data transcended, and the possibilities of Big Data were born.

What began with the search for web pages containing words has expanded to include images and sounds, and files recording geographic movement, postings to social media, and retail transactions, and promises to include data generated by sensors. What began as a service for Internet users has expanded to include systems for surveillance of those users.

What we take for granted

These new practices have given rise to data-driven experiences among people who have broadband access to the Internet and have begun to constitute taken-for-granted facts of contemporary life. Consider three instances of data-driven lived experience: privacy, shopping, and politics.

The experience of privacy is paradoxical. On one hand, the surveillance that feeds big data engines is commonly viewed as a retrograde step, taking us from a life rich in independent agency toward a totalitarian future. It is taken to mark the closing down of a region of personal life into which we once could retreat when we wanted to be alone, to be off-stage as it were, to practice a more authentic self, or at least not be called upon to fend off or yield to intrusions. Although the motives of a police state can be distinguished from those of a nanny state or a hovering butler, they all are said to corrode personal agency with antidemocratic consequences (Palfrey and Gasser 2008; Richards and King 2013.) The Pew Report finds that 91% of US adults agree or agree strongly that consumers have lost control of their personal data.

On the other hand, big data enables us to be perpetually on-stage. In the personal, family, and workplace spheres of life, boasting and humblebragging abound. We use Facebook, Amazon, LinkedIn, Snap, check-in tools like Foursquare, and tools like TripAdvisor to review and thereby proclaim our participation in all manner of hedonic experiences. The fact that these services depend on sharing of personally identifying information seems to be the point, not the price. Adding a tool such as Ghostery to one's browser makes visible the number of entities that watch us as we browse, feeding on website visitation data to deliver services to numerous entities in the data ecosystem, but still we browse.

Shopping relies on big data to an accelerating extent. In 2010, the headline, "The Pants That Stalked Me on the Web" over a story on ad retargeting, became a meme in shopping culture (e.g. Learmonth 2010). It crystallized anxiety about the practice of using individual browsing histories to nudge shoppers to buy things they had revealed themselves to be interested in. Concern about tracking of online shopping is balanced by gratification that online suggestions, nudges, and wish lists work so well. But online shopping is a small part of the larger experience of shopping. Big data's role in offline shopping is less understood, and so less feared and less appreciated, but no less pervasive. Chain stores such as Target, Walmart, and Best Buy maintain data repositories that augment each customer's shopping history with thousands of fields of data from third parties, including online browsing on third-party sites. When Amazon bought the Whole Foods grocery chain in 2017, online and offline data became so blended that the on/offline distinction began not to matter.

Politically, big data has begun to "fill in the gaps between society's three cardinal points: the self (individual, spectator, voter), the masses (population, audience, electorate) and power (party, spectacle, government)" (Meek 2017). The argument is that by precisely matching people according to tastes and preferences, messaging to them, and then encouraging them to share the messages with their social grid, algorithms can help people to see that what they might have thought of as their private disappointments and indignations are shared by others and constitute a political force. Could it be that the big data analytics deployed by Cambridge Analytica, combined with the ease of sharing facilitated by Facebook, created the sense of community and solidarity among people who were otherwise strangers that propelled Donald Trump into the presidency of the United States, as a series of articles in the *Guardian* newspaper have alleged? Did hoaxes pandering to private grievances and purporting to be real news, whether from political parties or foreign actors, constitute a significant share of the persuasive content that spread algorithmically on social grids? Few questions are more deserving of social science investigation than these.

What lies just ahead

We are on the brink of a step increase in the quantity of data in circulation: that what today we call big data is small by comparison with what is coming.

The argument rests on the assumption that the amount of data bears a relatively constant relation to the number of digital data generators in use. In the age of mainframe computers, there were perhaps 10,000 people playing the role of data generator. By the 1980s, with the arrival of personal computers, the number began to trend up and appears to have reached 2 billion worldwide by 2014 (Reference.com 2014). Now, as the most common data generator becomes a person on a mobile phone of which there are 4.8 billion in use (Statista 2017), there are perhaps 6 billion people generating digital data today. Soon people will give way to things, particularly sensors and chip-based cameras, as primary data generators. It is not unreasonable to conjecture that for each person in today's data-generating world, there will soon be one hundred sensors. As the cost of sensors falls with production at scale, they will become ubiquitous in personal and non-personal applications: to track vehicular and pedestrian traffic; to measure in-home energy consumption; in the wear-sensitive parts of automobiles and home appliances; to measure lawn soil moisture levels; to monitor pets, children, and disabled seniors; to control heat, lighting, and security in houses; to assess compliance with health practices from medical therapies to tooth brushing; to supervise out-of-office workers; and other uses and intrusions.

This explosion of data will challenge prevailing conceptions of a coordinated society. Consider for example the car as a mobile data center. All the cars in a region would make up a local area network, which would interact with the region's traffic network, the road construction database, the emergency services network, the retailer network, the gas station/charging station network, and the schools network. The car's navigation system, not the driver, would route vehicles to avoid traffic congestion, and the car's navigation system, not traffic police with flashlights, would let you know that there is no parking at the concert you were planning to attend and you are being sent home to watch a free streaming of the event with a stop at the store to pick up beer.

Adapting to change

I claimed at the start of this piece that modern life takes for granted the experiences made possible by data deployed at scale, and, in saying that, had in mind an analogy to the way that it consumes physical infrastructure like bridges. But it is safe, indeed practical, to take bridges for granted. Bridges evolve, improve, and decay on a time scale comparable to human lives. The way we live evolves comfortably in step with physical infrastructure: some elements, like housing construction methods, do not change in a lifetime; others, like methods for storing and playing music, change a handful of times. Even immaterial systems such as legal infrastructure evolve slowly, at the pace of legislators. Modernity depends on its popularity on the assurance that the way we live now is the same as or just a little better than the way it was, different but still intelligible. Between expressions of political opinion and exercise of choice in the marketplace, people, particularly well-off people, can feel uncritically devoted to the modernist drift.

It is altogether another thing to take digital infrastructure for granted. Data deployed at scale will not support a sense of comfort with this drift, if indeed there is drift and not a disruption. When change occurs in a data-driven system, the occasion for change, the rules governing change, and even the agencies responsible for change spread fast. No social disruption has spread to as many people on the planet in less time than have social networks such as Facebook and WeChat. Perhaps more people drink milk or use a battery-powered flashlight, but these global innovations diffused slowly. Digital innovations riding on the big data infrastructure have diffused to half the world's population in little over a decade.

Not only do they propagate rapidly. Embedded in these data-driven apps are rules or algorithms that possess the ability to update without supervision. Algorithms "learn" quickly. They optimize functions set by their designers who can, disturbingly, include the systems themselves, and set a pace and direction for change that does not wait on democracy. Algorithms become alternatives to laws, regulations, and social processes generally. Steiner (2012) describes the broad range of policy

decisions already made algorithmically. The more big data applications we adopt, the more we rely on rules instead of human judgment.

Arguably the rules are conservative. O'Neil (2016) contends that in general algorithms propagate past practices. A value system, wittingly or not, is imported into the decisions, and at least today algorithmic learning updates on facts, not values. Particularly when the rules are generated by pattern recognition trained on known cases of past success and failure, then, no matter how large the data-set, its predictions are the result of historical practices and their embedded values. For example, algorithms to predict teacher performance are trained on cases where underresourced black teachers failed for structural not personal reasons, yet the algorithms affect personal lives.

So, infrastructure that evolves at the pace of data systems is difficult to analyze and critique. But worse, living with it alters the sense of personhood for people relying on it, living under its thumb, or seeking to stand apart from it. Data-informed infrastructure contaminates the notion of an autonomous self that uses infrastructure and replaces it with a subservient self that is used for infrastructure's purposes. Cheney-Lippold (2017) spells out some of the evidence for this position. If identity is construed constructively, as an interaction between an actor and society, and if society is to some degree voiced algorithmically, then the way we live now is already affecting who we think we are.

Conclusion

This essay drives toward this conclusion. Applications of big data are taken for granted because they are useful or seductively self-expressive, and their mechanisms are opaque. They create an infrastructure that expands into new areas of life far more rapidly than any previous physical, commercial, or legislative infrastructure has diffused, as it measures more and writes rules to act on what it measures. Over time, therefore, more of our lives are subject to rules. But these are not rules designed by visionaries to protect the pursuit of happiness or to promote human flourishing. They are designed instrumentally, indeed hegemonically, by agents of data science. Shelley proposed that poets were the unacknowledged legislators of the world (1840/2002), but in the age of big data, machine learning may have a stronger claim.

Big data is iconic, yes, but it has teeth and it will bite. It presages transformations of the social order that are radical in their consequences and will not be reversed. When we imagine living without it, we are delusional, and when we take it for granted we are naive.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Brin, Sergey. 1998. "The Autonomy of a Large-scale Hypertextual Web Search Engine." *Computer Networks and ISDN Systems [Proceedings of the Seventh World Wide Web Conference]* 30 (1-7): 107–117.

Cheney-Lippold, John. 2017. *We Are Data: Algorithms and The Making of Our Digital Selves*. New York: NYU Press.

Ghemawat, Sanjay, Howard Gobioff, and Shun-Tak Leung. 2003. "The Google File System." *ACM SIGOPS Operating Systems Review* 37 (5): 29–43.

Learmonth, Michael. 2010. "The Pants That Stalked Me on the Web" *Advertising Age*, August 2. http://adage.com/article/digitalnext/pants-stalked-web/145204/.

Mayer-Schonberger, Viktor, and Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, MA: Houghton Mifflin Harcourt.

Meek, James. 2017. "Short Cuts." *London Review of Books*, June 1. https://www.lrb.co.uk/v39/n11/james-meek/short-cuts.

O'Neil, Cathy. 2016. *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

Palfrey, John, and Urs Gasser. 2008. *Born Digital: Understanding the First Generation of Digital Natives*. New York: Basic Books.

Reference.com. 2014. "How Many Computers are there in the World?" https://www.reference.com/technology/many-computers-world-e2e980daa5e128d0.
Richards, Neil M., and Jonathan H. King. 2013. "Three Paradoxes of Big Data." *66 Stanford Law Review Online* 41, September 3. Social Science Research Network (SSRN), https://ssrn.com/abstract = 2325537.
Shelley, Percy Bysshe. 1840/2002. "A Defense of Poetry and Other Essays," Reproduced in Project Gutenberg, https://www.gutenberg.org/files/5428/5428-h/5428-h.htm.
Statista. 2017. "Number of Mobile Phone Users Worldwide from 2013 to 2019 (in Billions)." https://www.statista.com/statistics/274774/forecast-of-mobile-phone-users-worldwide/.
Steiner, Christopher. 2012. *Automate This: How Algorithms Came to Rule Our World*. New York: Portfolio/Penguin.
Zuboff, Shoshana. 2015. "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization." *Journal of Information Technology* 30 (1): 75–89.