CHANGE
19 Key Essays on How Internet Is Changing Our Lives

CHANGE

Michael Nielsen

Who Owns Big Data?
Who Owns Big Data?

Michael Nielsen
Writer, scientist, and programmer
Who Owns Big Data?
Michael Nielsen
Michael Nielsen
michaelnielsen.org
Michael Nielsen is a writer, scientist, and programmer. Originally trained as a theoretical physicist, he worked on quantum computing for 15 years at institutions that include the California Institute of Technology and the Los Alamos National Laboratory. As he pursued that work, Michael became fascinated by the potential of the Internet and collective intelligence to change the way scientific research is done, and in 2008 he left his job as a tenured professor to pursue that interest. He is the author of Reinventing Discovery (Princeton University Press, 2011), a book describing how the Internet is changing science. His current major project is a technical book about artificial intelligence and neural networks.

Sites and services that have changed my life
arxiv.org
github.com
scholar.google.com
Who Owns Big Data?

In 2010, the CEO of Google at the time, Eric Schmidt, made a remarkable statement at a media event in Abu Dhabi: “One day we had a conversation where we figured we could just [use Google's data about its users] to predict the stock market. And then we decided it was illegal. So we stopped doing that” (Fortt 2010).

The journalist John Battelle (2010) has described Google as “the database of [human] intentions.” Battelle noticed that the search queries entered into Google express human needs and desires. By storing all those queries—more than a trillion a year—Google can build up a database of human intent. That knowledge of intention then makes it possible for Google to predict the movement of the stock market (and much else). Of course, neither Google nor anyone else has a complete database of human intentions. But part of the power of Battelle’s phrase is that it suggests that aspiration. Google cofounder Sergey Brin has said that the ultimate future of search is to connect directly to users’ brains (Arrington 2009). What could you do if you had a database that truly contained all human intentions?

The database of human intentions is a small part of a much bigger vision: a database containing all the world’s knowledge. This idea goes back to the early days of modern computing, and people such as Arthur C. Clarke and H. G. Wells exploring visions of a “world brain” (Wikipedia 2013). What’s changed recently is that a small number of technology companies are engaged in serious (albeit early stage) efforts to build databases which really will contain much of human knowledge. Think, for example, of the way Facebook has mapped out the social connections between more than 1 billion people. Or the way Wolfram Research has integrated massive amounts of knowledge about mathematics and the natural and social sciences into Wolfram Alpha. Or Google’s efforts to build Google Maps, the most detailed map of the world ever constructed, and Google Books, which aspires to digitize all the books (in all languages) in the world (Taycher 2010). Building a database containing all the world’s knowledge has become profitable.
This data gives these companies great power to understand the world. Consider the following examples: Facebook CEO Mark Zuckerberg has used user data to predict which Facebook users will start relationships (O’Neill 2010); researchers have used data from Twitter to forecast box office revenue for movies (Asur and Huberman 2010); and Google has used search data to track influenza outbreaks around the world (Ginsberg et al. 2009). These few examples are merely the tip of a much larger iceberg; with the right infrastructure, data can be converted into knowledge, often in surprising ways.

What’s especially striking about examples like these is the ease with which such projects can be carried out. It’s possible for a small team of engineers to build a service such as Google Flu Trends, Google’s influenza tracking service, in a matter of weeks. However, that ability relies on access to both specialized data and the tools necessary to make sense of that data. This combination of data and tools is a kind of data infrastructure, and a powerful data infrastructure is available only at a very few organizations, such as Google and Facebook. Without access to such data infrastructure, even the most talented programmer would find it extremely challenging to create projects such as Google Flu Trends.

Today, we take it for granted that a powerful data infrastructure is available only at a few big for-profit companies,¹ and to secretive intelligence agencies such as the NSA and GCHQ. But in this essay I explore the possibility of creating a similarly powerful public data infrastructure, an infrastructure which could be used by anyone in the world. It would be Big Data for the masses.

Imagine, for example, a 19-year-old intern at a health agency somewhere who has an idea like Google Flu Trends.² They could use the public data

---

¹. Many companies (including Google and Facebook) do, in fact, offer outsiders some limited access to their internal data. For example, the Facebook platform (http://en.wikipedia.org/wiki/Facebook_Platform) is a way for outside programmers to integrate applications with Facebook. Google products such as Google Maps offer “open APIs” (application programming interfaces) which allow outside programmers to use Google’s maps in their own applications. This kind of openness is valuable, but usually has stringent limitations that make it very different from, and far less powerful than, the direct access to infrastructure available to programmers at these companies.

². http://www.google.org/flutrends/
infrastructure to quickly test their idea. Or imagine a 21-year-old undergraduate with a new idea for how to rank search engine results. Again, they could use the public data infrastructure to quickly test their idea. Or perhaps a historian of ideas wants to understand how phrases get added to the language over time; or how ideas spread within particular groups, and die out within others; or how particular types of stories get traction within the news, while others don’t. Again, this kind of thing could easily be done with a powerful public data infrastructure.

These kinds of experiments won’t be free—it costs real money to run computations across clusters containing thousands of computers, and those costs will need to be passed on to the people doing the experiments. But it should be possible for even novice programmers to do amazing experiments for a few tens of dollars, experiments which today would be nearly impossible for even the most talented programmers.

Note, by the way, that when I say public data infrastructure, I don’t necessarily mean data infrastructure that’s run by the government. What’s important is that the infrastructure be usable by the public, as a platform for discovery and innovation, not that it actually be publicly owned. In principle, it could be run by a not-for-profit organization, or a for-profit company, or perhaps even by a loose network of individuals. Below, I’ll argue that there are good reasons such infrastructure should be run by a not-for-profit.

There are many nascent projects to build powerful public data infrastructure. Probably the best known such project is Wikipedia. Consider the vision statement of the Wikimedia Foundation (which runs Wikipedia): “Imagine a world in which every single human being can freely share in the sum of all knowledge. That’s our commitment.” Wikipedia is impressive in size, with more than 4 million articles in the English language edition. The Wikipedia database contains more than 40 gigabytes of data. But while that sounds enormous, consider that Google routinely works with data at the petabyte scale—a million gigabytes! By comparison, Wikipedia is miniscule. And it’s easy to see why there’s this difference. What the Wikimedia Foundation considers “the sum of all knowledge” is extremely narrow compared to the range of data about the world that Google finds useful—everything from scans of books to the data being generated by
Google’s driverless cars (each car generates nearly a gigabyte per second about its environment! [Gross 2013]) And so Google is creating a far more comprehensive database of knowledge.

Another marvelous public project is OpenStreetMap, a not-for-profit that is working to create a free and openly editable map of the entire world. OpenStreetMap is good enough that their data is used by services such as Wikipedia, Craigslist, and Apple Maps. However, while the data is good, OpenStreetMap does not yet match the comprehensive cover provided by Google Maps, which has 1,000 full-time employees and 6,100 contractors working on the project (Carlson 2012). The OpenStreetMap database contains 400 gigabytes of data. Again, while that is impressive, it’s miniscule by comparison to the scale at which companies such as Google and Facebook operate.

More generally, many existing public projects such as Wikipedia and OpenStreetMap are generating data that can be analyzed on a single computer using off-the-shelf software. The for-profit companies have data infrastructure far beyond this scale. Their computer clusters contain hundreds of thousands or millions of computers. They use clever algorithms to run computations distributed across those clusters. This requires not only access to hardware, but also to specialized algorithms and tools, and to large teams of remarkable people with the rare (and expensive!) knowledge required to make all this work. The payoff is that this much larger data infrastructure gives them far more power to understand and to shape the world. If the human race is currently constructing a database of all the world’s knowledge, then by far the majority of that work is being done on privately owned databases.

I haven’t yet said what I mean by a “database of all the world’s knowledge.” Of course, it’s meant to be an evocative phrase, not (yet!) a literal description of what’s being built. Even Google, the organization which has made most progress toward this goal, has for the most part not worked directly toward this goal. Instead, they’ve focused on practical user needs—search,

3.  http://www.openstreetmap.org/#map=5/51.500/-0.100
4.  An exception is the Google Knowledge Graph (http://en.wikipedia.org/wiki/Knowledge Graph), which really does seem to be a start on a database of all the world’s knowledge.
maps, books, and so on—in each case gathering data to build a useful product. They then leverage and integrate the data sets they already have to create other products. For example, they've combined Android and Google Maps to build up real-time maps of the traffic in cities, which can then be displayed on Android phones. The data behind Google Search has been used to launch products such as Google News, Google Flu Trends, and (the now defunct, but famous) Google Reader. And so while most of Google's effort isn't literally aimed at building a database of all the world's knowledge, it's a useful way of thinking about the eventual end game.

For this reason, from now I'll mostly use the more generic term public data infrastructure. In concrete, everyday terms this can be thought of in terms of specific projects. Imagine, for example, a project to build an open infrastructure search engine. As I described above, this would be a platform that enabled anyone in the world to experiment with new ways of ranking search results, and new ways of presenting information. Or imagine a project to build an open infrastructure social network, where anyone in the world could experiment with new ways to connect people. Those projects would, in turn, serve as platforms for other new services. Who knows what people could come up with?

The phrase a public data infrastructure perhaps suggests a singular creation by some special organization. But that's not quite what I mean. To build a powerful public data infrastructure will require a vibrant ecology of organizations, each making their own contribution to an overall public data infrastructure. Many of those organizations will be small, looking to innovate in new ways, or to act as niche platforms. And some winners will emerge, larger organizations that integrate and aggregate huge amounts of data in superior ways. And so when I write of creating a public data infrastructure, I'm not talking about creating a single organization. Instead, I'm talking about the creation of an entire vibrant ecology of organizations, an ecology of which projects like Wikipedia and OpenStreetMap are just early members.

I'll describe shortly how a powerful public data infrastructure could be created, and what the implications might be. But before doing that, let me make it clear that what I'm proposing is very different from the much-discussed idea of open data.
Many people, including the creator of the web, Tim Berners-Lee, have advocated open, online publication of data. The open data visionaries believe we can transform domains such as government, science, and the law by publishing the crucial data underlying those domains.

If this vision comes to pass then thousands or millions of people and organizations will publish their data online.

While open data will be transformative, it’s also different (though complementary) to what I am proposing. The open data vision is about decentralized publication of data. That means it’s about small data, for the most part. What I’m talking about is Big Data—aggregating data from many sources inside a powerful centralized data infrastructure, and then making that infrastructure usable by anyone. That’s qualitatively different. To put it another way, open publication of data is a good first step. But to get the full benefit, we need to aggregate data from many sources inside a powerful public data infrastructure.

Why a Public Data Infrastructure Should Be Developed by Not-for-Profits

Is it better for public data infrastructure to be built by for-profit companies, or by not-for-profits? Or is some other option even better—say, governments creating it, or perhaps loosely organized networks of contributors, without a traditional institutional structure? In this section I argue that the best option is not-for-profits.

Let’s focus first on the case of for-profits versus not-for-profits. In general, I am all for for-profit companies bringing technologies to market. However, in the case of a public data infrastructure, there are special circumstances which make not-for-profits preferable.

To understand those special circumstances, think back to the late 1980s and early 1990s. That was a time of stagnation in computer software,
a time of incremental progress, but few major leaps. The reason was Microsoft’s stranglehold over computer operating systems. Whenever a company discovered a new market for software, Microsoft would replicate the product and then use their control of the operating system to crush the original innovator. This happened to the spreadsheet Lotus 1-2-3 (crushed by Excel), the word processor Word Perfect (crushed by Word), and many other lesser-known programs. In effect, those other companies were acting as the research and development arms of Microsoft. As this pattern gradually became clear, the result was a reduced incentive to invest in new ideas for software, and a decade or so of stagnation.

That all changed when a new platform for computing emerged—the web browser. Microsoft couldn’t use their operating system dominance to destroy companies such as Google, Facebook, and Amazon. The reason is that those companies’ products didn’t run (directly) on Microsoft’s operating system, they ran over the web. Microsoft initially largely ignored the web, a situation that only changed in May 1995, when Bill Gates sent out a company-wide memo entitled “The Internet Tidal Wave” (Letters of Note 2011). But by the time Gates realized the importance of the web, it was too late to stop the tidal wave. Microsoft made many subsequent attempts to get control of web standards, but those efforts were defeated by organizations such as the World Wide Web Consortium, Netscape, Mozilla, and Google. Effectively, the computer industry moved from a proprietary platform (Windows) to an open platform (the web) not owned by anyone in particular. The result was a resurgence of software innovation.

The lesson is that when dominant technology platforms are privately owned, the platform owner can co-opt markets discovered by companies using the platform. I gave the example of Microsoft, but there are many other examples—companies such as Apple, Facebook, and Twitter have all used their ownership of important technology platforms to co-opt new markets in this way. We’d all be better off if dominant technology platforms were operated in the public interest, not as a way of co-opting innovation. Fortunately, that is what’s happened with both the Internet and the web, and that’s why those platforms have been such a powerful spur to innovation.

Platforms such as the web and the Internet are a little bit special in that they’re primarily standards. That is, they’re broadly shared agreements on
how technologies should operate. Those standards are often stewarded by not-for-profit organizations such as the World Wide Web Consortium and the Internet Engineering Task Force. But it doesn’t really make sense to say the standards are owned by those not-for-profits, since what matters is really the broad community commitment to the standards. Standards are about owning hearts and minds, not atoms.

By contrast, a public data infrastructure would be a different kind of technology platform. Any piece of such an infrastructure would involve considerable capital costs, associated with owning (or leasing) and operating a large cluster of computers. And because of this capital investment there really is a necessity for an owner. We’ve already seen that if a public data infrastructure were owned by for-profit companies, those companies would always be tempted to use their ownership to co-opt innovation. The natural alternative solution is for a public data infrastructure to be owned and operated by not-for-profits that are committed to not co-opting innovation, but rather to encouraging it and helping it to flourish.

What about government providing public data infrastructure? In fact, for data related directly to government this is beginning to happen, through initiatives such as data.gov, the U.S. Government’s portal for government data in the U.S. But it’s difficult to believe that having the government provide a public data infrastructure more broadly would be a good idea. Technological innovation requires many groups of people to try our many different ideas, with most failing, and with the best ideas winning. This isn’t a model for development that governments have a long history of using effectively. With that said, initiatives such as data.gov will make a very important contribution to a public data infrastructure. But they will not be the core of a powerful, broad-ranging public data infrastructure.

The final possibility is that a public data infrastructure not be developed by an organization at all, but rather by a loosely organized network of contributors, without a traditional institutional structure. Examples such as OpenStreetMap are in this vein. OpenStreetMap does have a traditional not-for-profit at its core, but it’s tiny, with a 2012 budget of less than 100,000 British pounds (OMS 2013). Most of the work is done by a loose network of volunteers. That’s a great model for OpenStreetMap, but part of the reason it works is because of the relatively modest scale of the
data involved. Big Data involves larger organizations (and larger budgets), due to the scale of the computing power involved, as well as the long-term commitments necessary to providing reliable service, effective documentation, and support. All these things mean building a lasting organization. So while a loosely distributed model may be a great way to start such projects, over time they will need to transition to a more traditional not-for-profit model.

Challenges for Not-for-Profits Developing a Public Data Infrastructure

How could not-for-profits help develop such a public data infrastructure?

At first sight, an encouraging sign is the flourishing ecosystem of open-source software. Ohloh, a site indexing open-source projects, currently lists more than 600,000 projects. Open-source projects such as Linux, Hadoop, and others are often leaders in their areas.

Given this ecosystem of open-source software, it’s somewhat puzzling that there is comparatively little public data infrastructure. Why has so much important code been made usable by anyone in the world, and so little data infrastructure?

To answer this question, it helps to think about the origin of open-source software. Open-source projects usually start in one of two ways: (1) as hobby projects (albeit often created by professional programmers in their spare time), such as Linux; or (2) as by-products of the work of for-profit companies. By looking at each of these cases separately, we can understand why open-source software has flourished so much more than public data infrastructure.

Let’s first consider the motivations for open-source software created by for-profit companies. An example is the Hadoop project, which was created by Yahoo as a way of making it easier to run programs across

large clusters of computers. When for-profit companies open source projects in this way, it’s because they don’t view owning the code as part of their competitive business advantage. While running large cluster-based computations is obviously essential to Yahoo, they’re not trying to use that as their edge over other companies. And so it made sense for Yahoo to open-source Hadoop, so other people and organizations can help them improve the code.

By contrast, for many Internet companies owning their own data really is a core business advantage, and they are unlikely to open up their data infrastructure. A priori nothing says this necessarily has to be the case. A for-profit could attempt to build a business offering a powerful public data infrastructure, and find some competitive advantage other than owning the data (most likely, an advantage in logistics and supply chain management). But I believe that this hasn’t happened because holding data close is an easy and natural way for a company to maintain a competitive advantage. The investor Warren Buffet has described how successful companies need a moat—a competitive advantage that is truly difficult for other organizations to duplicate. For Google and Facebook and many other Internet companies their internal data infrastructure is their moat.

What about hobby projects? If projects such Linux can start as a hobby, then why don’t we see more public data infrastructure started as part of a hobby project? The problem is that creating data infrastructure requires a much greater commitment than creating open-source code. A hobby open-source project requires a time commitment, but little direct expenditure of money. It can be done on weekends, or in the evenings. As I noted already above, building effective data infrastructure requires time, money, and a long-term commitment to providing reliable service, effective documentation, and support. To do these things requires an organization that will be around for a long time. That’s a much bigger barrier to entry than in the case of open source.

What would be needed to create a healthy, vibrant ecology of not-for-profit organizations working on developing a public data infrastructure?

This question is too big to comprehensively answer in a short essay such as this. But I will briefly point out two significant obstacles to this
happening through the traditional mechanisms for funding not-for-profits: foundations, grant agencies, and similar philanthropic sources.

To understand the first obstacle, consider the story of the for-profit company Ludicorp. In 2003 Ludicorp released an online game called *Game Neverending*. After releasing the game, Ludicorp added a feature for players to swap photos with one another. The programmers soon noticed that people were logging onto the game just to swap photos, and ignoring the actual gameplay. After observing this, they made a bold decision. They threw out the game, and relaunched a few weeks later as a photo-sharing service, which they named Flickr. Flickr went on to become the first major online photo-sharing application, and was eventually acquired by Yahoo. Although Flickr has faded since the acquisition, in its day it was one of the most beloved websites in the world.

Stories like this are so common in technology circles that there’s even a name for this phenomenon. Entrepreneurs talk about *pivoting* when they discover that some key assumption in their business model is wrong, and they need to try something else. Entrepreneur Steve Blank, one of the people who developed the concept of the pivot, has devised an influential definition of a startup as “an organization formed to search for a repeatable and scalable business model” (Blank 2010). When Ludicorp discovered that photo sharing was a scalable business in a way that *Game Neverending* wasn’t, they did the right thing: they pivoted hard.

This pattern of pivoting makes sense for entrepreneurs who are trying to create new technologies and new markets for those technologies. True innovators don’t start out knowing what will work; they discover what will work. And so their initial plans are almost certain to be wrong, and will need to change, perhaps radically.

The pivot has been understood and accepted by many technology investors. It’s expected and even encouraged that companies will change their mission, often radically, as they search for a scalable business model. But in the not-for-profit world this kind of change is verboten. Can you imagine a not-for-profit telling their funders—say, some big foundation—that they’ve decided to pivot? Perhaps they’ve decided that they’re no longer working with homeless youth, because they’ve discovered that their technology has
a great application to the art scene. Such a change won’t look good on the end-of-year report! Yet, as the pivots behind Flickr and similar companies show, that kind of flexibility is an enormous aid (and arguably very nearly essential) in developing new technologies and new markets.

A second obstacle to funding not-for-profits working on a public data infrastructure is the risk-averse nature of much not-for-profit funding. In the for-profit world it’s understood that technology startups are extremely risky. Estimates of the risk vary, but typical estimates place the odds of failure for a startup at perhaps 70 to 80 percent (Gompers et al. 2008). Very few foundations or grant agencies would accept 70 to 80 percent odds of failure. It’s informative to consider entrepreneur Steve Blank’s startup biography. He bluntly states that his startups have made “two deep craters, several ‘base hits,’ [and] one massive ‘dot-com bubble’ home run” (Blank 2013). That is, he’s had two catastrophic failures, and one genuine success. In the for-profit startup world this can be bragged about; in the not-for-profit world this rate of success would be viewed as disastrous. The situation is compounded by the difficulty in defining what success is for a not-for-profit; this makes it tempting (and possible) for mediocre not-for-profits to scrape by, continuing to exist, when it would be healthier if they ceased to operate, and made space for more effective organizations.

One solution I’ve seen tried is for foundations and grant agencies to exhort applicants to take more risks. The problem is that any applicant considering taking those risks knows failure means they will still have trouble getting grants in the future, exhortation or no exhortation. So it still makes more sense to do low-risk work.

One possible resolution to this problem would be for not-for-profit funders to run failure audits. Suppose programs at the big foundations were audited for failures, and had to achieve a failure rate above a certain number. If a foundation were serious about taking risks, then they could run a deliberately high-risk grant program, where the program had to meet a target goal of at least 70 percent of projects failing. Doing this well would require careful design to avoid pitfalls. But if implemented well, the outcome would be a not-for-profit culture willing to take risks. At the moment, so far as I am aware, no large funder uses failure audits or any similar idea to encourage genuine risk taking.
I’ve painted a bleak picture of not-for-profit funding for a public data infrastructure (and for much other technology). But it’s not entirely bleak. Projects such as Wikipedia and OpenStreetMap have found ways to be successful, despite not being started with traditional funding. And I am optimistic that examples such as these will help inspire funders to adopt a more experimental and high-risk approach to funding technological innovation, an approach that will speed up the development of a powerful public data infrastructure.

Two Futures for Big Data

We’re at a transition moment in history. Many core human activities are changing profoundly: the way we seek information; the way we connect to people; the way we decide where we want to go, and who we want to be with. The way we make such choices is becoming more and more dominated by a few technology companies with powerful data infrastructure. It’s fantastic that technology can improve our lives. But I believe that we’d be better off if more people could influence these core decisions about how we live.

In this essay, I’ve described two possible futures for Big Data. In one future, today’s trends continue. The best data infrastructure will be privately owned by a few large companies who see it as a competitive advantage to map out human knowledge. In the other future, the future I hope we will create, the best data infrastructure will be available for use by anyone in the world, a powerful platform for experimentation, discovery, and the creation of new and better ways of living.

Acknowledgments

Thanks to Jen Dodd, Ilya Grigorik, and Hassan Masum for many conversations about these ideas.
References


Gross, Bill. “Google’s Self-Driving Car gathers almost 1 GB per SECOND.” Twitter, April 13, 2013. https://twitter.com/Bill_Gross/statuses/329069954911580160


Gross, Bill. “Google’s Self-Driving Car gathers almost 1 GB per SECOND.” Twitter, April 13, 2013. https://twitter.com/Bill_Gross/statuses/329069954911580160


