1.4 Flipbook: What Does Data Science Mean to the World?

_____

We've gathered some selective readings exploring what some contemporary mainstream media are writing about what data science means to the world.

We first turn to National Public Radio's Adam Frank for his piece, *Big Data Is The Steam Engine Of Our Time*. You can read the full piece here (http://www.npr.org/blogs/13.7/2013/03/12/174028759/big-data-is-the-steam-engine-of-our-time), or these excerpts below.

> "We inhabit a world of blinding technological change. New devices, new programs and new infrastructure rise up, dominate discourse and pass away before we even have time to comprehend their intent. But for all the change we've experienced, the the most profound transformation of the digital era is really just getting started. Welcome to the era of Big Data.
>
> For the sake of clarity, let's recall what Big Data means. At the intersection of the Internet, digital recording technologies and a profusion of small-scale wireless sensors comes an exponential increase in stored data. We're recording information of every kind: tweets, engine temperatures, Facebook photos, stock trades, grocery store purchases. These data points pile up, billions upon billions, trillions upon trillions, recording our lives and the life of the world we inhabit. They are very, very valuable to people with the ability to see patterns in, and extract insight from, the blizzard of information blowing out of the cloud.
>
> Ultimately, the promise of Big Data is the ability to understand (and control) a seemingly chaotic world on levels never before imagined. The dangers of Big Data stem from that very same promise. Its impact on society will be akin to the transformative effect of past technological revolutions."
>
> As strange as it may seem, Big Data may be the steam engine of our time.
>
> I believe there is something real and powerful happening in the Big Data revolution. It's more than just a fad. It's the next link in the long chain connecting culture and technology to human history.
>
> Now Big Data—seen and unseen—is hitting us in all corners of our lives, from the price of things to traffic patterns to who our social networks think we befriend. Through new fields like data science and network theory, Big Data will not only change the world we move through as individuals, it will change the world we imagine through science. Like it or not, Big Data will only get bigger (and bigger)."

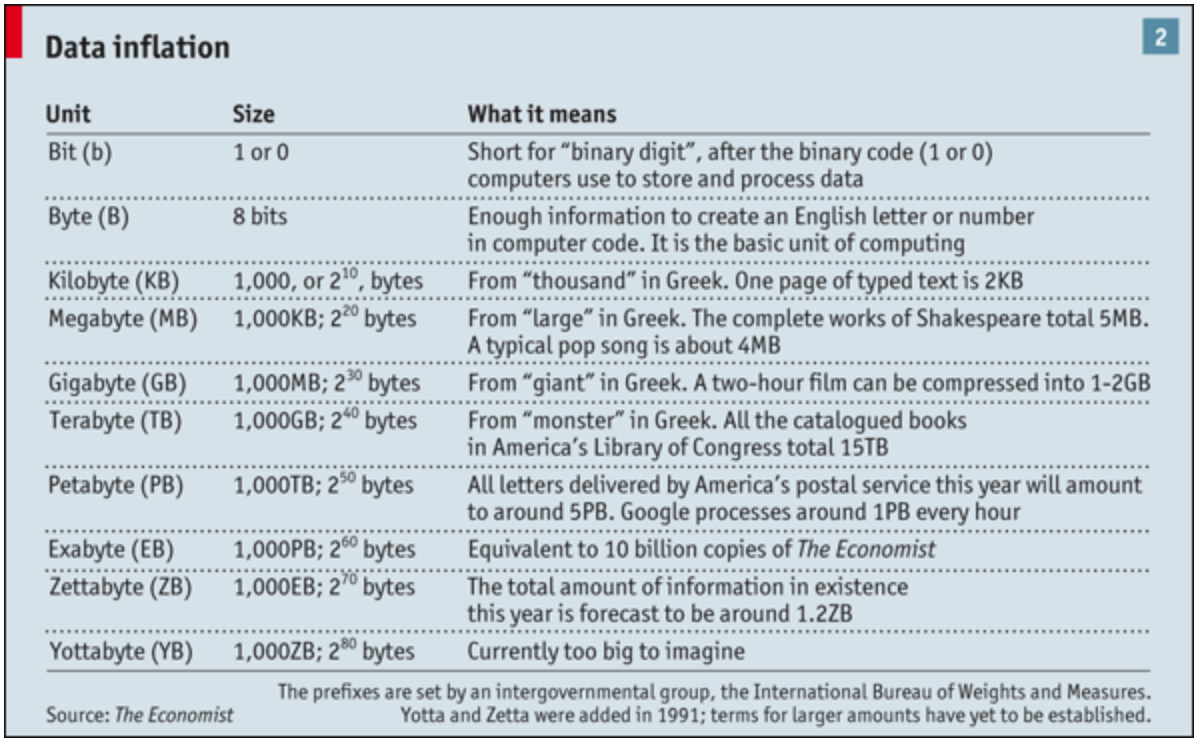
For a global perspective we now look to pieces from the Economist's 2010 special report on managing information. Let's start with a piece on the seemingly growing ubiquity of data across domains. Here's an excerpt from *Data, data everywhere* (http://www.economist.com/node/15557443):

> "When the Sloan Digital Sky Survey started work in 2000, its telescope in New Mexico collected more data in its first few weeks than had been amassed in the entire history of astronomy. Now, a decade later, its archive contains a whopping 140 terabytes of information. A successor, the Large Synoptic Survey Telescope, due to come on stream in Chile in 2016, will acquire that quantity of data every five days.
>
> Such astronomical amounts of information can be found closer to Earth too. Wal-Mart, a retail giant, handles more than 1m customer transactions every hour, feeding databases estimated at more than 2.5 petabytes—the equivalent of 167 times the books in America's Library of Congress (see article for an explanation of how data are quantified). Facebook, a social-networking website, is home to 40 billion photos. And decoding the human genome involves analysing 3 billion base pairs—which took ten years the first time it was done, in 2003, but can now be achieved in one week.
>
> All these examples tell the same story: that the world contains an unimaginably vast amount of digital information which is getting ever vaster ever more rapidly. This makes it possible to do many things that previously could not be done: spot business trends, prevent diseases, combat crime and so on. Managed well, the data can be used to unlock new sources of economic value, provide fresh insights into science and hold governments to account."
>
> ...
>
> ""We are at a different period because of so much information," says James Cortada of IBM, who has written a couple of dozen books on the history of information in society. Joe Hellerstein, a computer scientist at the University of California in Berkeley, calls it "the industrial revolution of data". The effect is being felt everywhere, from business to science, from government to the arts. Scientists and computer engineers have coined a new term for the phenomenon: "big data"."
>
> …
>
> "The amount of digital information increases tenfold every five years. Moore's law, which the computer industry now takes for granted, says that the processing power and storage capacity of computer chips double or their prices halve roughly every 18 months. The software programs are getting better too. Edward Felten, a computer scientist at Princeton University, reckons that the improvements in the algorithms driving computer applications have played as important a part as Moore's law for decades."
>
> "This shift from information scarcity to surfeit has broad effects. "What we are seeing is the ability to have economies form around the data—and that to me is the big change at a societal and even macroeconomic level," says Craig Mundie, head of research and strategy at Microsoft. Data are becoming the new raw material of business: an economic

> input almost on a par with capital and labour. “Every day I wake up and ask, ‘how can I flow data better, manage data better, analyse data better?” says Rollin Ford, the CIO of Wal-Mart.”
>
> …
>
> “The way that information is managed touches all areas of life. At the turn of the 20th century new flows of information through channels such as the telegraph and telephone supported mass production. Today the availability of abundant data enables companies to cater to small niche markets anywhere in the world. Economic production used to be based in the factory, where managers pored over every machine and process to make it more efficient. Now statisticians mine the information output of the business for new ideas.
>
> “The data-centred economy is just nascent,” admits Mr Mundie of Microsoft. “You can see the outlines of it, but the technical, infrastructural and even business-model implications are not well understood right now.”“

In all this talk of big data and data science, we often see references to the volume of data. Just how much data are we talking about, particularly compared to volumes we’re familiar with? From the Economist’s *All too much* (http://www.economist.com/node/15557421)

> “Quantifying the amount of information that exists in the world is hard. What is clear is that there is an awful lot of it, and it is growing at a terrific rate (a compound annual 60%) that is speeding up all the time. The flood of data from sensors, computers, research labs, cameras, phones and the like surpassed the capacity of storage technologies in 2007. Experiments at the Large Hadron Collider at CERN, Europe's particle-physics laboratory near Geneva, generate 40 terabytes every second—orders of magnitude more than can be stored or analysed. So scientists collect what they can and let the rest dissipate into the ether.
>
> According to a 2008 study by International Data Corp (IDC), a market-research firm, around 1,200 exabytes of digital data will be generated this year. Other studies measure slightly different things. Hal Varian and the late Peter Lyman of the University of California in Berkeley, who pioneered the idea of counting the world's bits, came up with a far smaller amount, around 5 exabytes in 2002, because they counted only the stock of original content.”



Data inflation 2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

Source: *The Economist*

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

…

"Only 5% of the information that is created is "structured", meaning it comes in a standard format of words or numbers that can be read by computers. The rest are things like photos and phone calls which are less easily retrievable and usable. But this is changing as content on the web is increasingly "tagged", and facial-recognition and voice-recognition software can identify people and words in digital files.

"It is a very sad thing that nowadays there is so little useless information," quipped Oscar Wilde in 1894. He did not know the half of it."

Let's now look at what mainstream media suggests that data science means to various domains, from business to governments to broader society. We should also take note of what the media highlights as the potential risks associated with data science, particularly for individuals and their data. From the Economist' *The Data Deluge* (http://www.economist.com/node/15579717).

"Everywhere you look, the quantity of information in the world is soaring. According to one estimate, mankind created 150 exabytes (billion gigabytes) of data in 2005. This year, it will create 1,200 exabytes. Merely keeping up with this flood, and storing the bits that might be useful, is difficult enough. Analysing it, to spot patterns and extract useful information, is harder still. Even so, the data deluge is already starting to transform business, government, science and everyday life. It has great potential for good—as long as consumers, companies and governments make the right choices about when to restrict the flow of data, and when to encourage it.

A few industries have led the way in their ability to gather and exploit data. Credit-card

companies monitor every purchase and can identify fraudulent ones with a high degree of accuracy, using rules derived by crunching through billions of transactions. Stolen credit cards are more likely to be used to buy hard liquor than wine, for example, because it is easier to fence. Insurance firms are also good at combining clues to spot suspicious claims: fraudulent claims are more likely to be made on a Monday than a Tuesday, since policyholders who stage accidents tend to assemble friends as false witnesses over the weekend. By combining many such rules, it is possible to work out which cards are likeliest to have been stolen, and which claims are dodgy.

Mobile-phone operators, meanwhile, analyse subscribers' calling patterns to determine, for example, whether most of their frequent contacts are on a rival network. If that rival network is offering an attractive promotion that might cause the subscriber to defect, he or she can then be offered an incentive to stay. Older industries crunch data with just as much enthusiasm as new ones these days. Retailers, offline as well as online, are masters of data mining (or “business intelligence”, as it is now known). By analysing “basket data”, supermarkets can tailor promotions to particular customers' preferences. The oil industry uses supercomputers to trawl seismic data before drilling wells. And astronomers are just as likely to point a software query-tool at a digital sky survey as to point a telescope at the stars.

There's much further to go. Despite years of effort, law-enforcement and intelligence agencies' databases are not, by and large, linked. In health care, the digitisation of records would make it much easier to spot and monitor health trends and evaluate the effectiveness of different treatments. But large-scale efforts to computerise health records tend to run into bureaucratic, technical and ethical problems. Online advertising is already far more accurately targeted than the offline sort, but there is scope for even greater personalisation. Advertisers would then be willing to pay more, which would in turn mean that consumers prepared to opt into such things could be offered a richer and broader range of free online services. And governments are belatedly coming around to the idea of putting more information—such as crime figures, maps, details of government contracts or statistics about the performance of public services—into the public domain. People can then reuse this information in novel ways to build businesses and hold elected officials to account. Companies that grasp these new opportunities, or provide the tools for others to do so, will prosper. Business intelligence is one of the fastest-growing parts of the software industry.

“But the data deluge also poses risks. Examples abound of databases being stolen: disks full of social-security data go missing, laptops loaded with tax records are left in taxis, credit-card numbers are stolen from online retailers. The result is privacy breaches, identity theft and fraud. Privacy infringements are also possible even without such foul play: witness the periodic fusses when Facebook or Google unexpectedly change the privacy settings on their online social networks, causing members to reveal personal information unwittingly. A more sinister threat comes from Big Brotherishness of various kinds, particularly when governments compel companies to hand over personal information about their customers. Rather than owning and controlling their own personal data, people very often find that they have lost control of it.

The best way to deal with these drawbacks of the data deluge is, paradoxically, to make

more data available in the right way, by requiring greater transparency in several areas. First, users should be given greater access to and control over the information held about them, including whom it is shared with. Google allows users to see what information it holds about them, and lets them delete their search histories or modify the targeting of advertising, for example. Second, organisations should be required to disclose details of security breaches, as is already the case in some parts of the world, to encourage bosses to take information security more seriously. Third, organisations should be subject to an annual security audit, with the resulting grade made public (though details of any problems exposed would not be). This would encourage companies to keep their security measures up to date.

Market incentives will then come into play as organisations that manage data well are favoured over those that do not. Greater transparency in these three areas would improve security and give people more control over their data without the need for intricate regulation that could stifle innovation. After all, the process of learning to cope with the data deluge, and working out how best to tap it, has only just begun."