AI & Soc (2023) 38:2705–2708
DOI 10.1007/s00146-016-0662-7

CURMUDGEON CORNER

Big Data

Xin Wei Sha[1] · Gabriele Carotti-Sha[2]

Published online: 29 April 2016

Big Data is a term popular among administrators and business circles motivating quite a lot of investment today. Part of it is rebranding. But rather than old wine in new bottle, it may be more likely watered wine in new bottles branded to take old wine's market. David Donoho, one of the foremost statisticians in the world and a visionary of data science, observes that much of what passes for Big Data is a bit of software engineering plus a bit of statistics. Knowledge is power. But a smidgen of knowledge plus poor judgment can do a lot of damage, especially when billions of dollars implicate billions of lives.[1]

1 What is Big Data?

The definition of "Big Data" is inherently vague. It is the field that studies management and processing of any dataset too large for direct, individual interpretation. This limitation of interpretability bears the risk of conflation, for it really has two meanings. On the one hand, we refer to the current limit in computing power: As the quantity of collected data pushes against this limit, new techniques have to be developed. The other limit is of *human* interpretation: Human-guided development is necessary because only the scientist can declare what the data actually mean. As has been said, "Don't be worried about Big Data, be worried about who's using it."[2]

2 Big Data as IT sysad, rebranded

Oftentimes the storing transmission of decanting of large quantities of data itself introduces mechanical problems—the disk drives and RAM simply cannot contain it all in one processor, so the databases have to be chopped up and distributed across a network of computers. Donoho writes:

> The *new* skills attracting so much media attention are not skills for better solving the real problem of inference from data; they are coping skills for dealing with organizational artifacts of large-scale cluster computing. The new skills cope with severe new constraints on algorithms posed by the multiprocessor/networked world. In this highly constrained world, the range of easily constructible algorithms shrinks dramatically compared to the single-processor model, so one inevitably tends to adopt inferential approaches which would have been considered rudimentary or even inappropriate in olden times.

✉ Xin Wei Sha
shaxinwei@gmail.com

Gabriele Carotti-Sha
weiminsha@gmail.com

[1] Synthesis Center, School of Arts, Media + Engineering, Arizona State University, Phoenix, AZ, USA

[2] Stanford University, Stanford, CA, USA

[1] Data science attracts big capital, which in turn begins to draw academic administrators' attention: In September 2015, University of Michigan announced a $100 million "Data Science Initiative," with funding to hire 35 new faculties, and UC Berkeley has a new Masters in Data Science. http://midas.umich.edu/dsi/announcement, https://datascience.berkeley.edu/.

[2] http://lifehacker.com/what-is-big-data-and-whos-collecting-it-1595798695.

It's revealing that the figures in IBM's white paper about data science[3] are system architectures—diagrams of what machines plug to what—and org charts, rather than workflows of analytical techniques for modeling, inferencing, or predicting. Viewed with untinted glasses, such a mechanical notion of Big Data is like reducing the medicine to an inventory of tools: stethoscope, hypodermic needle, thermometer, or reducing music to diagrams for assembling tubas and violins or seating charts showing where musicians sit.

3 Big Data $\ll$ Media like video or film

MongoDB points out the monstrous volume, velocity, and variety of corporate databases: Facebook ingests 500 terabytes per day, and interestingly, a Boeing 737 generates 240 terabytes of flight data flying across the USA.[4] But data by itself are meaningless: What's the value of all the archives of video and music if nobody experiences them?

$$\text{Media} \neq \text{Data}, \text{Media} = \text{Data} + \text{Interpretation}.$$

4 Big Data is sparse compared with matter

The Apple iPhone's Retina display boasts a dot pitch of 401 ppi. But the inter-molecular distance between SiO_2 molecules in quartz glass is about 3.04 Angstroms, which yields a density of 291,261,058 SiO_2 molecules per inch. Therefore, each pixel on your iPhone screen is over 700,000 times coarser than the atoms of which it is made. And in terms of area, the atomic density is over 520 billion times greater than the pixel resolution of the screen. And this is even before we begin to account for the boggling number of gradations of intensity available to the levels of light energy that can be directed through a sheet of atoms (think Planck).

So, Big Data is pretty trivial in size, density, and complexity compared to matter. And we have developed over the past 500 years quite powerful and systematic means to measure, account, model, and predict entities at the atomic scale. These means are called physics, chemistry, arithmetic, and statistics. Many of the techniques have been computationally implemented and made freely available in common languages such as FORTRAN, MATLAB, Mathematica, C++, S, and R.

"Wait wait!" you say, "By data, we mean human-readable stuff." Do we? Isn't the aura of Big Data in its superhuman scale? Aren't we talking about machines making, repackaging, and processing stuff? We have over 100,000 years of experience expertly and reproducibly manipulating stuff at far greater orders of density: fire, metallurgy, cooking.

Coming back to mundane human-generated data, Donoho points out that 200 years ago, census data were already on the order of what is called Big Data today. He points out that statistics was in fact invented to deal with such Big Data. Moreover, scientific data from astronomical to geophysical to high-energy physics instruments of observation all have been routinely generating oceans of data far larger than the relatively coarse data generated from human consumer behavior. And this has been happening for decades.

5 Data are intrinsically meaningless

Shannon's theory is applicable to any encoding scheme.

Fundamentally, the crux of Shannon's great invention of information—the concept of the bit, on which the data of Big Data are predicated—precisely cut datum (0s and 1s) apart from meaning. Defining "information" as $H = \log S^n = n \log S$ where S is the number of symbols in a code and n is the length of a message depends on the foundational conceptual step of considering only formal strings of symbols chosen from a fixed finite set of formal symbols, never mind what they mean. The very formality of this definition expressly excludes interpretation. All meaning has to come from interpretation. Someone, somebody, has to look at the data, do some operations with it, and make up some interpretation to be shared with other people. So, from the very get-go, data are meaningless, and pure, uninterpreted information is not knowledge.

Conflating Shannon's formal, abstract information with knowledge would be as magical as trying to revivify a dissected body by simply permuting its organs on a table (and we're not talking about the good kind of magic).

It gets worse.

6 Data are not facts of nature

Data are not just pieces of nature lying around for data scientists to pick up like shells on a beach. Data don't grow on trees. Data are constructed via very elaborate complexes of theory, politics, judgment calls, plus apparatuses, devices, technologies, and procedures that are themselves conditioned by theory, politics, and judgment calls. And every statistician and empirical scientist will tell you that it matters crucially how the data are constructed.

The degree of contingency built into the very data as collected before they ever show up in a spreadsheet or

[3] Fig 1: "Big Data analytics ecosystem" and Fig 2: "Second-generation tools."

[4] https://www.mongodb.com/big-data-explained.

database field is already profoundly intertwined with contingent factors like culture, prejudice, what the data collector ate for breakfast, how she was raised as a child, or the weather that day. Donoho notes: "At the Tukey Centennial, Rafael Irizarry gave a convincing example of exploratory data analysis of GWAS data, studying how the data row mean varied with the date on which each row was collected, [convincing] the field of gene expression analysis to face up to some data problems that were crippling their studies."[5]

It gets worse.

Google Flu Trends (GFT) was a tool that reputedly converted query information into diagnosis information by positing a "relationship" between how many people search for flu-related topics and how many people actually have flu symptoms. However, a "ground truth" study published in Science showed that GFT overestimated the incidence of flu in 2 years by over 50 %. During peak flu season, GFT claimed 11 % of Americans were sick with the flu when the incidence was 6 % (from the Center for Disease Control). And GFT failed to pick up major outbreaks like the H1N1-A flu pandemic. Simply extrapolating from CDC data 3 weeks back significantly out-performed GFT's opaque machinery.[6]

Google's machinery is opaque for non-mathematical and damning reasons: They tweak the search to increase advertising revenue. Recommended searches reflect what other users search for as well as what ad sponsors want to be displayed. So with advertisers and Google itself tweaking the search algorithms, such a method of "observation" is hopelessly entangled with contingency, politics, and market-rigging, about as far from "ground truth" as one can get.

7 Data versus functions, operations, workflows

Even more important than the volume, velocity of Big Data is what is being done with that data. As any baker can tell you: whether you pour the water into the flour before or after kneading it with your fingers, and how you knead it, make a world of difference. Just so, what you do with the data, and how you do it, produce entirely different results. As principled data scientists will tell you, these variant workflows encode corporate interest, politics, law, and social and institutional habit.

Most fundamentally, taking a page from Wittgenstein, Saussure, Shannon, and everyone who recognized the abyss between abstract syntax and living pragmatics, a language capable of expression is not structured by symbol but by function and use. The token over which a function operates is meaningless by itself.

Moving to more professional waters, an extremely important question in the field of machine learning is always: How much data do we need? So much can go wrong by applying a model to massive amounts of information. Adding more samples could, based on the situation, do any of the three things: make the model worse, make it better, and not make a difference at all.

Machine learning engineers know that a model's usefulness lies not in its effective volume, but in the potential availability of the right kind of data when needed. If I don't have an accurate polling estimate in New Hampshire, it's not enough to get more data from a single location; I want to spread my survey across the territory to avoid overfitting. In fact, data scientists talk about Big Data problems being really dependent on "small data" when put into practice.

As for the hullabaloo over specific tools like Map/Reduce… (1) the toolkit trumpeted by Big Data evangelists is just a few basic tools rattling in a largely empty can: Map/Reduce, TensorFlow, ye olde neural nets fattened on big iron. Map/Reduce is just the capitalized name of a classic structure operator found in a host of programming languages like Mathematica and APL all the way back to the mother tongue, LISP. Albeit a very powerful operator, in context Map/Reduce is simply a bit of syntactic algebra, and a far cry from inference or prediction. TensorFlow black-boxes a mess of techniques. Some of these techniques are quite old, like neural nets, aka Deep Belief Networks (another example of hypostatization or imputing godlike power by Proper Name), whose fundamental limitations do not go away when you run them against larger datasets. Others, like tensors on algebraic structures, contain deep assumptions, such as multi-linearity, that profoundly empower and also limit the validity of their application. It's interesting that marketing evangelists capitalize common terms to reify them into intellectual property, whereas mathematicians and physicists lowercase their inventors—like Riemann and Dalton—into workaday tools and units of measure.

All this—judgment calls as to how the data are constructed (they are *always* constructed), which datasets to include (more data sometimes yield less knowledge), and generally which operation to apply before others—forms part of what data professionals call workflow. This is a critical hidden part of scientific analysis. "[T]here are [medical data] studies of the same dataset, and the same intervention and outcome, but with different analysis workflow, [where] the published conclusions about the risk of the intervention are reversed."[7]

[5] Donoho 2015, 23.

[6] Walsh, 2014. http://time.com/23782/google-flu-trends-big-data-problems/.

[7] David Donoho 2015, citing Carp 2012, Madigan 2014. Emphasis original.

8 Conclusion

If by information we mean predicates that humans can use to understand or manipulate the world, whereas data are what our computing machinery generates and manipulates, referencing past and future curmudgeons, we have a very useful set of inequalities

$$\text{Data} \neq \text{Information} \neq \text{Knowledge} \neq \text{Concern}$$

No doubt, all these are useful, but some are more useful than others. And when we come to matters of meaning and concern, data big or small are only a part of the equation, and a small part to boot.

Curmudgeon Corner Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for super-human intelligence promotes potential benefits to wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? -Editor.