Big Data and Their Epistemological Challenge

Luciano Floridi

It is estimated that humanity accumulated 180 EB of data between the invention of writing and 2006. Between 2006 and 2011, the total grew ten times and reached 1,600 EB. This figure is now expected to grow fourfold approximately every 3 years. Every day, enough new data are being generated to fill all US libraries eight times over. As a result, there is much talk about "big data". This special issue on "Evolution, Genetic Engineering and Human Enhancement", for example, would have been inconceivable in an age of "small data", simply because genetics is one of the data-greediest sciences around. This is why, in the USA, the National Institutes of Health (NIH) and the National Science Foundation (NSF) have identified big data as a programme focus. One of the main NSF–NIH interagency initiatives addresses the need for core techniques and technologies for advancing big data science and engineering (see NSF-12-499).

Despite the importance of the phenomenon, it is unclear what exactly the term "big data" means and hence refers to. The aforementioned document specifies that: "The phrase 'big data' in this solicitation refers to large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future." You do not need to be an analytic philosopher to find this both obscure and vague. Wikipedia, for once, is also unhelpful. Not because the relevant entry is unreliable, but because it reports the common definition, which is unsatisfactory: "data sets so large and complex that they become awkward to work with using on-hand database management tools". Apart from the circular problem of defining "big" with "large", the definition suggests that data are too big or large only in relation to our current computational power. This is misleading. Of course, "big", as many other terms, is a relational predicate: a pair of shoes is too big for you, but fine for me. It is also trivial to acknowledge that we tend to evaluate things non-relationally, in this case as absolutely big, whenever the frame of reference is obvious enough to be left

L. Floridi (✉)
School of Humanities, University of Hertfordshire, de Havilland Campus, Hatfield, Hertfordshire AL10 9AB, UK
e-mail: l.floridi@herts.ac.uk

implicit. A horse is a big animal, no matter what whales may think. Yet, these two simple points may give the impression that there is no real trouble with "big data" being a loosely defined term referring to the fact that our current computers cannot handle so many gazillions of data efficiently. And this is where two confusions seem to creep in. First, that the *epistemological problem* with big data is that there is too much of them (the *ethical problem* concerns how we use them; see below). And second, that the *technological solution* to the epistemological problem is more and better techniques and technologies, which will "shrink" big data back to a manageable size. The epistemological problem is different, and it requires an equally epistemological solution.

Consider the problem first. "Big data" came to be formulated after other buzz expressions, such as "infoglut" or "information overload", began to fade away, yet the idea remains the same. It refers to an overwhelming sense that we have bitten off more than we can chew, that we are being forced-fed like geese, that our intellectual livers are exploding. This is a mistake. Yes, there is an obvious exponential growth of data on an ever-larger number of topics, but complaining about such overabundance would be like complaining about a banquet that offers more than we can ever eat. Data remain an asset, a resource to exploit. Nobody is forcing us to digest every available byte. We are becoming data-richer by the day; this cannot be the fundamental problem.

Since the problem is not the increasing wealth of data that is becoming available, clearly the solution needs to be reconsidered: it cannot be merely how many data we can technologically process. If anything, more and better techniques and technologies are only going to generate more data. If the problem were too many data, computers would only exacerbate it. Growing bigger digestive systems, as it were, is not the way forward.

The real, epistemological problem with big data is *small patterns*. Precisely because so many data can now be generated and processed so quickly, so cheaply, and on virtually anything, the pressure both on the data *nouveau riche*, such as Facebook or Walmart, Amazon or Google, and on the data *old money*, such as genetics or medicine, experimental physics or neuroscience, is to be able to spot where the new patterns with real added value lie in their immense databases and how they can best be exploited for the creation of wealth and the advancement of knowledge.

Small patterns matter because they represent the new frontier of competition, from science to business, from governance to social policies. In a Baconian open market of ideas, if someone else can exploit them earlier and more successfully than you do, you might be out of business soon, like Kodak, or miss a fundamental discovery. Small patterns may also be risky, because they push the limit of what is predictable and, therefore, may be anticipated, about not only nature's, but also people's, behaviour. This is the ethical problem. Target, an American retailing company, relies on the analysis of the purchasing patterns of 25 products in order to assign each shopper a "pregnancy prediction" score, estimate her due date, and send coupons timed to specific stages of her pregnancy. In a notorious case, it caused some serious problems when it sent coupons to a family in which the teenager daughter had not informed her parents about her new status.

Unfortunately, small patterns may be significant only if properly aggregated, e.g. in terms of loyalty cards and shopping suggestions, compared, as when a bank can

use big data to fight fraudsters, and timely processed, as in financial markets. And because information is indicative also when it is not there, small patterns can also be significant if they are absent. Sherlock Holmes solves one of his famous cases because of the silence of the dog, which should have barked. If big data are not "barking" when they should, something is going on, as the financial watchdogs (should) know.

The increasingly valuable undercurrents in the ever-expanding oceans of data are invisible to the computationally naked eye, so more and better techniques and technologies will help significantly. Yet, by themselves, they will be insufficient. And mere data hoarding, while waiting for more powerful computers and software, will not work. Since 2007, the world has been producing more data than available storage. We have shifted from the problem of what to save to the problem of what to erase. Something must be deleted or never be recorded. Think of your smart phone becoming too full because you took too many pictures, and make it a global problem. The infosphere run out of memory space to dump its data years ago. This is not as bad as it looks. Rephrasing a common saying in advertisement, half of our data is junk, we just do not know which half. So what we need is a better understanding of which data are worth preserving. And this is a matter of grasping which questions are or will be interesting. Which is just another way of saying that, because the problem with big data is small patterns, ultimately, the game will be won by those who "know how to ask and answer questions" (Plato, *Cratylus*, 390c) and therefore know which data may be useful and relevant, and hence worth collecting and curating, in order to exploit their valuable patterns. We need more and better techniques and technologies to see the small data patterns, but we need more and better epistemology to sift the valuable ones.

Big data are here to grow. The only way of tackling them is to know what you are or may be looking for. At the moment, such epistemological skills are taught and applied by a black art called *analytics*. Not exactly your standard degree at the university. Yet, so much of our well-being depends on it that it might be time to develop a philosophical investigation of its methods. Who knows, philosophers might have something to learn, but also a couple of lessons to teach. Plato would agree.