Victor Mayer-Schönberger/ Kenneth Cukier, Big Data

John Murray 2013, 242 pages, ISBN 978-1-84854-792-6

Book Review

by **Thomas Dreier**, Director, Institute for Information and Economic Law, Karlsruhe Institute of Technology (KIT), Karlsruhe

Recommended citation: Thomas Dreier, Book Review: Victor Mayer-Schönberger/Kenneth Cukier, Big Data, 5 (2014) JIPITEC 60, para1

1 Every once in a while and at unpredictable intervals, books are published which sum up an emerging trend in digital technology and explain its future impact on society and the regulatory system. Amongst such books one might list Nicholas Negroponte's *Being Digital* (1995), Hal Varian's and Carl Shapiro's *Information Rules* (1999), Jeremy Rifkin's *Age of Access* (2000), Lawrence Lessig's *Code and Other Laws of Cyberspace* (also 2000) and now Victor Mayer-Schönberger's and Kenneth Cukier's *Big Data* (2013). Of course, to qualify a book as 'important' in the sense that it spots a major trend, correctly describes this trend's future impact upon society and makes an imprint on subsequent discussion is only possible in retrospect. The saying traditionally attributed to Nils Bohr according to which "prediction is very difficult, especially about the future" also holds true in this respect. After all, 'big data' might just be another one of those buzz-words succeeding the rather short-lived 'cloud computing' and already being supplanted, at the time of publication of Meyer-Schönberger's and Kenneth Cukier's book, by the term 'smart data'.[1] However, there is some credible evidence that big data does indeed "mark an important step in humankind's quest to quantify and understand the world", as the authors – the first a professor at the Oxford Internet Institute and author of *Delete: The Virtue of Forgetting in the Digital Age* (2009), the second *The Economist's* data editor – claim at the end of their introductory chapter.[2]

2 What are the reasons why 'big data' – which suggests a mere increase in the amount of data collected and processed – will lead to a fundamental change as the authors pretend? The answer is that rather than resulting in a quantum leap, the increase of data results in a qualitative change of data collection and analysis. This qualitative change is threefold. First, there is more – as a matter of fact, much more, and in some cases all – data relating to a particular phenomenon that can be analysed.[3] This represents a marked shift from earlier times when only samples of data were available that merely represented the total reality analysed. Second, in the authors' words, data will be "messier", i.e. "looking at vastly more data ... permits us to loosen up our desire for exactitude",[4] which again contrasts with the days when the basis for analysis was representative data, which had to be as accurate as possible in order not to produce incorrect results. Third, and perhaps most importantly, big data analysis merely searches for correlation rather than for causality, which is a decisive "move away from the age-old search for causality".[5] This move away will lead to a change in the way we explain the world (think of the new field of computational social sciences which supplanted earlier empirical methods based on sample statistics). It will likewise result in changes in the information economy and the way we organize our institutions. This "datafication" of society, as the authors call it, is driven by digital data collection undertaken both by public authorities and private companies, from public sector information, customer data, satellite data to data collected by the increasing number of geo-positioned devices.[6]

3 As regards the economy, a new "treasure hunt" has just begun, which is "driven by the insights to be expected from data and the dormant value that can be unleashed by a shift from causation to correlation".[7] While new markets are emerging, the question is whether companies that possess huge amounts of data should keep them for themselves, whether they should hand them over to big data analysts who aggregate them with data resources from other companies thus creating added-value to be sold back to the initial producers/owners of data, whether companies should license their data to third parties or even competitors, or whether they should make them openly – and freely – available to everyone (as has been opted for, one might add, by the legislature with regard to public sector information[8]). Last but not least, there is the tricky yet important issue of how to price data. It appears that as of yet, little clarity exists regarding the answer to the question of which model should be adopted in which case. However, almost certainly, the shift from traditional modes of data analysis to the analysis of big data will produce both winners and losers. According to the authors, in the big data value chain composed of big data holders, intermediaries – i.e. data specialists with expertise or technologies to carry out complex analysis – and "companies and individuals with a big data mindset",[9] data owners and those with a big data mindset will most likely be on the winner's side. In contrast, according to the authors, in many areas, we'll see the "demise of the expert" whose decisions are mainly based on year-long experience, whereas newly emerging data analysts who often come from fields outside of the area analysed will take over. But these intermediaries also operate on shaky ground, the more the tools for analysing data will become generally available. Also, data owners are in a position to keep their data as property. Summing up, the authors conclude that it will be the data itself which will be the most important asset in the big data value chain. In the book, the authors describe and explore each of these trends in separate chapters under the rather simple and straightforward headings "Now", "More", "Messy", "Correlation", "Datafication", "Value" and "Implications".

4 Of course, this new development will not come without "Risks" (the "dark side of big data" as the authors call it), and these risks call for "Control", if the future ("Next") will be mastered without loss of human freedom and individual responsibility. These risks are also threefold. First, with the new insights big data provides to those who analyse them, privacy and data protection are threatened even more than they already are on the Internet.[10] Second, the correlations found on the basis of big data between certain indicators and the behaviour of groups of people results in the "possibility of using big data predictions about people to judge and punish them even before they've acted". Needless to point out, such "penalties based on propensities ... negate ideas of fairness, justice and free will".[11] Third, the danger exists that data and numbers will be fetishized and relied on even in instances where the numbers are not the only factor on which an appropriate decision should be based. In sum, "handled responsibly, big data", the authors believe, "is a useful tool of rational decision-making". However, the authors fear, "wielded unwisely, it can become an instrument of the powerful, who may turn it into a source of repression".[12]

5 What do the authors propose in order to control the risks just described? What is lost and what will have to be preserved?

6 As regards privacy, it is obvious that existing data protection rules are at odds with big data. Data protection's three fundamental principles of (1) data avoidance, (2) specification of purpose of use and (3) prohibition on passing on data without consent, can hardly be maintained in view of the three fundamental conditions on which big data analysis rests, namely (1) to collect as much data as possible, which (2) are used for purposes other than those for which the initial consent was given, and which (3) are combined with data held by other sources. In addition, in many instances, anonymisation of personal data – the traditional means of redress – will not be of help when it comes to analysing big data. Since banning the collection and use of big data is not a viable alternative, the authors propose to move from privacy to accountability (in a way similar to the shift, in the Gutenberg era, from censorship to freedom of expression on the one hand, and legal responsibility in case of libel and slander on the other hand). In other words, in the alternative privacy framework Mayer-Schönberger and Cukier propose, big data users should as a rule have the permission to collect, store and analyse personal data as much and for as long and for whatever purpose they want. Of course, "legislators may choose different time frames for reuse, depending on the data's inherent risk, as well as on different societies' values".[13] As a counterpart, according to the authors, big data users should be held accountable for adverse results of their actions. In addition to this regulatory shift from "privacy by consent" to "privacy through accountability", the authors rely on technical innovation, mainly techniques of "differential privacy" which blur data so that correlations may still be detected without revealing results which make it possible to identify a particular individual.

7 Regarding the problem of judging individuals according to group propensities, the authors propose "a guarantee that we will continue to judge people by considering their personal responsibility and their actual behavior, not by 'objectively' crunching data to determine whether they are likely wrongdoers".[14] Most importantly, the authors call for monitoring and transparency of the algorithms which establish

the correlations and which in almost all cases constitute a black box. Inspired mainly by the German model of the internal data protection official and external auditing systems as well as the dual role of in-house accountants and outside auditors, the authors propose the mandatory creation of both internal and external "algorithmists". Their task should be to "monitor big data companies' activities", to act "as impartial auditors to review the accuracy or validity of big-data predictions whenever the government requires it" and to "perform audits for firms that want expert support".[15] Finally, "as the nascent big data industry develops, an additional critical challenge will be to safeguard competitive big-data markets". This challenge the authors want to meet by way of antitrust regulation preventing abusive power comparable to the regulatory systems that established competition and oversight in the area of earlier monopolistic or oligopolistic technologies such as railroads, steel manufacturing and telegraph networks.

8 Ultimately, the authors are "confident" that with these new strategies in place, "the dark side of big data will be contained".[16]

9 Most, if not all of this makes perfect sense, and the book addresses the major issues that can be spotted at present. However, a couple of additional issues can already be pointed out which the book does not yet address. For example, the aspect of nature of legal "ownership" of data is not dwelt on, nor is the issue discussed whether or not performing an analysis of someone else's big data infringes upon the extraction and reutilization right under the sui generis protection regime of the EU database Directive.[17] Undeniably, there is always factual "ownership" of data by those who have first collected them. But the authors only briefly mention possible strategies of benefiting from the economic value which these data may hold. Should a particular company keep those data for itself? Should it entrust a data-intermediary with its analysis and pay for the results of the analysis? Should it license the data or even make them generally available for free? Some additional guidance similar to the one given in the book by Varian and Shapiro mentioned above with regard to doing business on the Internet still seems to be called for regarding the economics of big data, both on the level of micro- and of macro-economics. The crucial question is, under what conditions will an individual firm and the society at large benefit from big data analysis? Most likely, this answer will depend on the amount of data collected and on the quality of the algorithms performing the analysis, as well as on the extent to which the data and the analysing software tools will become available to third parties.

10 If, in this respect, the authors address the problem of judgment of individuals by propensities both as one of the individual vis-à-vis the state and vis-à-vis private firms, their proposed safeguard of procedural guarantees only seems to address the area of criminal law, i.e. the relationship between state and citizen. In contrast, they do not provide a hint as to how effects of scoring activities on individuals should be dealt with. Rather, in this respect the authors focus on the core problem that the algorithms designed to detect correlations in the mass of data from different sources are not transparent. In most cases, they are private property of the firms engaging in the business of big data analysis. Hence, even the authors cannot tell us how these algorithms work. They can only inform the reader about the fact that in order to predict the spread of the winter flu in the United States, it took Google "a staggering 450 million different mathematical models in order to test the search terms, comparing their predictions against actual flu cases"[18] in earlier years. Their call for transparency in this respect is of utmost importance and their proposal of data "algorithmists" – which at least in cases of dispute should be entrusted with advisory or auditing competencies – is at least one solution which might provide redress. However, this novel idea still needs to be propagated. Only recently, the German Federal Supreme Court granted the plaintiff, who had been refused credit on the basis of the German credit agency's big data calculation, a claim for information against the credit agency only concerning the data the agency had used for the calculation of the plaintiff's creditworthiness. In contrast, the Court denied a claim for information regarding the algorithm used by the credit agency which, in the eyes of the Court, constitutes a protected business secret.[19] This decision is not only a marked contrast from the call for transparency of the authors of *Big Data*; it also failed to take into account that the credit agency in question enjoys a de facto monopoly in Germany.

11 Finally, the non-transparency in this respect raises another problem. Decisions directly inflicted upon individuals meet with acceptance difficulties whenever it is not possible to understand how the decision was arrived at. This is a general problem of automated and computerized decisions which is aggravated by big data's complex algorithms and which affects more and more areas of society (think of the search results produced by Google search and, more generally, of how algorithms focus our attention via the use of computerized data in automated media processes[20]). But then, even before big data, we have become accustomed to the fact that a number of individual decisions are based on collective data and mathematical models (think about insurance premiums, airfare, etc.). Therefore, the question of transparency will have to be phrased differently. Rather than asking whether there should be transparency or no transparency, the question should be in what situations transparency is called for and in what situations non-transparency might

be acceptable. Ultimately, one might ask: Can the dark side of big data really be contained by implementing appropriate safeguards, or will we end up – if only for some time to come – accepting the individual "false positives" produced by non-transparent algorithms of big data analysis as "collateral damage" of a technological system in much the same way as we got used to those injured and killed by the system of motorized traffic?

12 The book is written in the Anglo-American style that appeals to the general public ("tell them what you will tell them, tell them, tell them what you just told them"). It contains the most famous as well as lesser known real-life examples of big data analyses, such as Google's predicting the spread of the flu in the US on the basis of 45 search terms used by the users of Google's search engine some days before the actual outbreak of the flu in a particular area, or the discovery of an individual woman's pregnancy on the basis of a change in her buying pattern that correlates to most women's third month of pregnancy, to name just two of these examples. The book is not an academic one, but as a *New York Times* and *Wall Street Journal* bestseller (as the paperback cover proudly announces), it will get all the attention it deserves.

1 See, e.g., the call for proposals by the German Ministry for Economic Affairs and Energy, www.bmwi.de/DE/Service/wettbewerbe,did=596106.html.

2 Mayer-Schönberger/Cukier, Big Data, London 2013, p. 17.

3 Ibid., p. 12 and pp. 19 et seq.

4 Ibid., p. 13 and pp. 32 et seq.

5 Ibid. p. 14 and pp. 50 et seq.

6 Ibid. p. 15 and pp. 73 et seq.

7 Ibid. p. 15 and pp. 98 et seq.

8 Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information, OJ L 345, 31.12.2003, p. 90, as amended by Directive 2013/37/EU of the European Parliament and of the Council of 26 June 20, OJ L 175, 27.6.2013, p.1.

9 Mayer-Schönberger/Cukier, op. cit. (footnote 2), p. 129.

10 Ibid., p. 151 and pp. 152 et seq.

11 Ibid., p. 151 and pp. 157 et seq.

12 Ibid., p. 151 and pp. 163 et seq.

13 Ibid. p. 174.

14 Ibid., p. 178.

15 Ibid., pp. 178 et seq. and p. 181.

16 Ibid. p. 183.

17 Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJ L 77 of 27.3.1996, p. 20.

18 Mayer-Schönberger/Cukier, op. cit. (footnote 2), pp. 2 and 179.

19 German Federal Supreme Court (Bundesgerichtshof) of 28 January 2014, case no. VI ZR 156/13.

20 See, e.g., www.groundbreaking-journalism.com/#konferenz.