Data, Responsibly

Serge Abiteboul, Julia Stoyanovich

Data, Responsibly

Serge Abiteboul, INRIA and ENS Cachan, and Julia Stoyanovich, Drexel University

Our society is increasingly relying on algorithms in all aspects of its operation. We trust algorithms not only ***to help carry out routine tasks***, such as accounting and automatic manufacturing, but also ***to make decisions on our behalf***. The sorts of decisions with which we now casually entrust algorithms range from unsettling (killer drones), to tedious (automatic trading), or deeply personal (online dating). Computer technology has tremendous power, and with that power comes immense responsibility. Nowhere is the need to control the power and to judiciously use technology more apparent than in massive data analysis, known as big data.

Big data technology holds incredible promise of improving people's lives, accelerating scientific discovery and innovation, and bringing about positive societal change. The goal of big data analysis is to efficiently sift through oceans of data, identifying valuable knowledge. The more data is available, the more knowledge can be derived. This gives a strong incentive for data acquisition, as well as for data sharing. Data sharing may be fully unrestricted, as is the case with the Open Data movement, or more controlled, as is the case with medical data (for privacy) and scientific or commercial data (to preserve competitive advantage).

Our everyday activities are heavily monitored, generating more and more data, notably with Web cookies, smart phones and digital-self objects. We voluntarily participate in online social networks and use a plethora of Web platforms, allowing data about ourselves to be collected and analyzed. In addition, datasets gathered by different companies are being integrated on the large scale, enabling even richer analysis of our behavior and preferences, and possibly disclosing information that we prefer to keep private. We willingly partake of the digital ecosystem because we believe that on-line services help us make decisions, facilitate interaction and socialization, and, in general, improve our lives. In allowing Web platforms to help us decide what product to buy, which movie to watch, whom to befriend or date, and which news to read, we are effectively trading privacy and freedom for convenience. In doing so, we rarely stop to think that the primary business model of these platforms is in fact the monetization of our digital life.

Dimensions of responsible data analysis

If not used responsibly, big data technology can propel economic inequality, destabilize global markets and affirm systemic bias. While the potential opportunity of big data techniques is well accepted, the necessity of using these techniques responsibly should be discussed. ***What constitutes responsible data analysis practices?*** In the society we envision, data and data analysis are fair, transparent and available equally to all.

Fairness is interpreted here as ***lack of bias***. It is incorrect to assume that results obtained using computerized processing are unbiased by nature. Bias may come from the data, e.g., if a questionnaire contains biased questions, or if the answers are tampered with. It may come from the algorithm, reflecting political, commercial, sexual, religious, or other kinds of preferences of its designers. In the natural and social sciences, where data analysis has long been the cornerstone of discovery, conclusions must be based on reliable data and robust analysis methods that are not skewed by the research hypothesis. For example, it is unfair to use a crime dataset, in which some cities are under-represented, to draw conclusions about the relative incidence of crime among different racial groups. Conclusions can still be drawn from incomplete and imprecise datasets, but this requires honest and careful statistical analysis and typically more modest goals. Similarly, in consumer-facing applications, answers and recommendations should be guided by the user's questions, preferences and task-relevant attributes, and not purposefully skewed to offer commercial advantage to advertising platforms or re-sellers. As an example of commercial bias, consider the recent antitrust charges brought against Google by the European Union [1]. The company was accused of concealed advertising – skewing search results in favor of its own products, while failing to make this bias apparent to search engine users. A typical justification of this practice by Google has been that this is an attempt to personalize the search experience, and so is introduced for the user's own benefit, rather than for the benefit of the company.

A biased algorithm may reconstruct values of hidden variables, such as race, and then make decisions based on these variables. An algorithm may also attempt to hide its use of problematic variables, thereby introducing bias. Whether intentional or not, such biases may be unethical, or even illegal, as is, for example, disproportionately offering less advantageous financial products to members of minority groups, a practice known as steering. It is not intrinsically wrong for an algorithm to reconstruct values of hidden variables. This technique is at the heart of many methods in machine learning, and it is what makes these methods extremely powerful and appealing. Yet, this power comes with great responsibility. The fact that it is difficult to understand what these algorithms

compute, and to detect whether their choices are guided by the values of some hidden variables, cannot serve as a justification for violating our society's ethical principles.

In another related sense, ***fairness is non-discrimination***. When tackling a technically challenging problem such as relevance ranking of Web search results, or news article recommendation, it is rational to prioritize popular demands. However, it is important, and indeed fair, to also pay attention to the uncommon information needs. These are said to be "in the tail"; they may not be very popular individually, yet together constitute an important part of the alternatives. To illustrate, consider an on-line dating platform like Match.com, a crowdsourcing marketplace like Amazon Mechanical Turk, or a funding platform like Kickstarter. In these environments, it is often the case that a small subset of the items dominates rankings and so is given unfair advantage. This also hinders productivity, because most popular resources saturate quickly, while numerous tasks starve.

Transparency. Users want to know and control both what is being recorded about them, and how the recorded information is being used, e.g., to recommend content or target advertisement to them. For example, there has been much outcry about the apparent lack of transparency in the way personal data is being handled by Facebook, leading to several changes in their privacy policy. However, while privacy is certainly an important part of the picture, there is far more to transparency than privacy. Transparent data analysis frameworks will require verification and auditing of datasets and algorithms for fairness, robustness, diversity, non-discrimination and privacy.

Transparency is a prerequisite for ***accountability***, and in particular for enabling a data provider, such as a social network user or a scientist, to specify authorized use for their data, and to verify that their data is being used responsibly. An important ingredient in transparency is availability of provenance metadata, which describes who created a dataset and how. In the sciences, provenance enables proper credit attribution, supports reasoning about result quality, and, more generally, encourages the publication of data.

Equal availability. The current model of data collection and usage, particularly in the social domain, but also in the sciences, leads to a concentration of datasets and data analysis methods in the hands of a few big players. Small companies and individuals lack access to data and to analysis capabilities. For small companies such as SMEs or NGOs, the trade-off is between open competition and equity. One may also argue that fully open competition is leading to an oligopoly context, thus towards less freedom for customers and eventually higher prices. For individuals, it is an issue of self-determination and empowerment. Users of on-line platforms often do not have full access to their own data. Both small companies and individuals typically lack the competency to understand data collection and analysis practices, and to correctly interpret results, let alone to carry out sophisticated data analysis of their own.

Towards data-responsible practices

What would it take to transform our data analysis practices, to make them responsible? It is our belief that this requires a coordinated effort along four dimensions: education, public policy, user organization, and technology.

Education

Basic command of mathematics and of the sciences, including physics, chemistry and biology, was considered essential for an informed citizen of the 20th century. In the 21st century, because of our society's increasing reliance on computer technologies, basic computational and data competencies are indispensable. US President Barak Obama called on every American to learn to code in his address during the 2013 Computer Science Education Week. The curriculum of computing education in schools is being discussed in most developed countries. We believe that there is an urgent need to include data literacy education into this discussion. To see how important data literacy is, recall how Fox News presented Obamacare enrollment statistics in March 2014 [2]. The conservative US TV channel showed a bar chart with two entries: enrollments as of March 27 (6,000,000) and enrollment goal (7,066,000). The chart visually conveyed that enrollments so far were about 2/3 short of the target. In response to an outcry about its misleading coverage of Obamacare enrollment statistics on Twitter, Fox News apologized for airing the chart and issued a correction.

What are the skills and competencies that citizens should possess to be considered data-literate? These skills should allow a person to critically judge data collection, the raw data that it is based on, the analysis processes, and the quality of the results. These skills are needed to help individuals make informed decisions about, e.g., the implications of disclosing some personal information to an application, the risk/benefit trade-offs of vaccines, or the comparative impact of government investment in foreign aid vs. military spending. Data literacy skills are particularly important for decision makers, who currently often lack them.

Important questions are how to teach data literacy in schools and which concepts to cover at what age. Children start accessing information on the Web at a very early age, and so it is important to expose them to simple data analysis experiments already in elementary school. At this age one can explain basic data quality concepts to students, demonstrating that they need to be skeptical of data found on the Web, and explaining the value of credit attribution. Middle school and high school students should be gradually exposed to additional data literacy concepts, including how to check the quality of analysis protocols and data visualization methods. Students should be taught basic principles of probability and statistics, including, e.g., the distinction between correlation and causality. This curriculum should also focus on turning students from informed spectators into actors, by asking them to practice building their own datasets and analyzing them.

Public policy

Government bodies are important for regulating the disclosure of assumptions underlying data collection and analysis, and for ensuring fair access. Nonetheless, meaningful top-down regulation is difficult to achieve and maintain, for several reasons. First, the problem is international: Data is generated and hosted in different countries, with typically different laws, and serves an international user base. Second, big data analysis technologies are developing very rapidly and it is difficult for governments to produce regulation that keeps up with the pace of technological changes. Finally, the questions involved in responsible data analysis are complex and very technical, and often involve deep statistical arguments. Even where the will to regulate exists, technology is not yet providing an appropriate level of support for what regulators may want. For example, it may not be possible to determine who produced a particular dataset, and under what assumptions, because the corresponding metadata is missing.

Nonetheless, while direct government regulation is definitely a complex issue, and may not be feasible today, there is still a need for government awareness and involvement, to ensure that responsible data analysis is receiving proper attention and to articulate high-level principles as guidelines. Government involvement should then be by a combination of regulation and incentives. For example, governments may provide incentives to organizations that share their data and code, in support of transparency and easier verification of data-responsible practices. Of course, making ***all*** data and code open cannot be required for reasons of protecting privacy of individuals and competitive advantage of companies. For this reason, governments should also support other means of facilitating responsibility verification. Finally, as already mentioned, the integration of data from different applications and different companies is the most serious issue. The control and limitation of data integration should also be part of the government mission.

User organizations

Users can get organized to specify guidelines, best practices and standards of responsible data analysis, and to engage in a dialogue with companies to prevent particularly unfair or opaque data practices. After all, success of big data technologies critically depends on data availability. Users express support for a service by offering their participation, and speak against a service by changing service providers, or by otherwise withholding participation. User organizations have the potential to become a powerful regulation mechanism for responsible data analysis, since a grassroots movement can start much faster and have more immediate impact than government regulation. For a recent example of the power of users, recall the 2012 Instagram controversy [3]. The company, which had at that time recently been acquired by Facebook, made changes to its privacy policy, allowing image-based targeting of advertising, and did not communicate these changes clearly to its then 100 million-strong user base. The change angered users, prompting Instagram to rapidly revert its privacy policy changes.

A series of initiatives is converging towards giving individual users more control over how others gather and use their personal data [4]. Examples of such initiatives are Smart Disclosure in the U.S., enabling more than 40 million Americans to download their own health data using the “Blue Button” on their health insurance provider’s website, and MesInfos in France, where several large companies including network operators, banks and retailers have agreed to explore and experiment with the sharing with customers of the personal data they hold about them. Such moves clearly improve transparency.

Technology

Two kinds of technological advances are needed to make our society data-responsible. On the one hand, organizations must be given tools to ***design*** data collection and analysis methods that are fair and transparent. On the other hand, users (both individuals and organizations) must be given the means to ***verify*** that data collection and data analysis methods of an organization are responsible, and to detect violations. Note that software design typically also includes verification. Here we focus on ***a posteriori*** verification, which occurs after data analysis.

While design and verification are related, they also differ in an important way. In the case of design, data and algorithms are typically fully available (responsibility from within, in ***white-box*** mode). In contrast, verification may have to work with limited access to data and code, typically through a predefined set of access methods (responsibility from the outside, in ***black-box*** mode). In both cases, specific properties to be designed or verified must be specified by laws, contracts or commitments by a company to its users. These properties must be translated into technical specifications and must in turn be verified automatically by algorithms. This raises a number of technical challenges, e.g., design a setting where the responsibility of a recommendation engine is controlled with limited disclosure of its code and its data, or design a fair ranking algorithm that does not favor only the most popular items, but also strive for diversity in results.

To verify the responsibility of an algorithm ***a posteriori***, one can either analyze its program or test its behavior on different inputs. Program analysis is closely related to theorem proving in mathematics, while observation of behavior is related to how real-world phenomena such as a heart, a galaxy or an atmospheric cloud are studied in the sciences. Program analysis is complex and, with current technology, limited to small critical pieces of software, for example in the aerospace or nuclear industries. However, this task becomes more feasible if it is provisioned for at application design time. This approach, which we term ***responsibility by design***, is in-line with a recent approach to ensuring privacy, called ***privacy by design***. It is also encouraging that it is usually easier to check a proof that a program is behaving in a particular way than it is to find such a proof. Testing the behavior of a complex program by observation is also not easy because, depending on the input, very many different outputs are possible. Powerful verification methods will have to rely on a combination of program analysis and observation, both of which are complex and computationally expensive. This is why today most organizations rarely invest in verifying security and privacy properties, and more generally in properties that ensure responsibility.

As is apparent from the discussion so far, algorithms that are important here are decidedly data-centric. They may even rely on data during their development. For example, a machine-learning algorithm is ***trained*** on data before it can be used. Data is the basic material on which everything in this environment is based, and it is thus extremely important to reason about the responsibility properties of the data itself. Another essential ingredient is ***metadata*** accompanying the data that guarantees its authenticity, explains its origin and history of derivation (known as ***provenance***), and, more generally, assigns a meaningful interpretation to data. To both design and verify responsible data analysis environments, technology must be developed to answer the following questions. ***Is a particular dataset biased? Are the results of a particular data analysis method reliable with high confidence?***

Towards a data-responsible society

Big data technology has immense ***power***. This power comes with a great ***risk*** to our society if technology is used irresponsibly. All stakeholders of the big data ecosystem, including scientists and engineers, but also commercial companies, users and governments, have a ***responsibility*** to ensure that technology is used in a way that is fair and transparent, and that it is equally available to all. Let us coordinate our efforts along the dimensions of education, public policy, user organization and technology to turn the promise of big data into societal progress!

Pointers

[1] http://www.theguardian.com/technology/2015/apr/14/european-commission-antitrust-charges-google
[2] http://www.businessinsider.com/fox-news-obamacare-chart-2014-3
[3] http://www.telegraph.co.uk/technology/facebook/9760151/Instagram-retreats-over-privacy-outcry.html
[4] http://cacm.acm.org/magazines/2015/5/186024-managing-your-digital-life/fulltext