Issues in Ethical Data Management
Extended Abstract

Serge Abiteboul

Inria, École Normale Supérieure, Paris, France

October 23, 2017

Abstract Data science holds incredible promise of improving people's lives, accelerating scientific discovery and innovation, and bringing about positive societal change. Yet, if not used responsibly, this technology can generate economic inequality, destabilize global markets and worsen systemic bias. We consider issues such as bias and violation of data privacy in data analysis. We discuss desirable properties of data analysis such as fairness, transparency, neutrality, and diversity. Our goal is to draw the attention of the computer science community to the important emerging subject of responsible data management and analysis. We present our perspective on the issue, and motivate research directions.

The data management research field has traditionally been driven primarily by enterprise data and focused on developing more and more sophisticated data models, and tackling issues such as performance and reliability. We believe that, in the future, the field will be increasingly driven by personal and social data, and will need to deal with challenging ethical issues. We will have to design concepts and principles to guide us in determining which behaviors, in data management, help us and which are harmful.

To move computer science towards more responsible data analysis, the first issue is that of specifying desired properties. To make it operational in computers automatic decisions, the policy has to be made precise, to be formally stated in technical language. The computers role is also to help control the properties of data analysis by providing: tools to collect and analyze data responsibly, and tools to verify that data analysis was performed responsibly. To check the behavior of a program, one can either analyze its code (in the style of proving mathematical theorems) or test its effect (in the style of the experimental study of phenomena such as climate or the human heart). Both approaches have been intensely investigated, e.g., for guaranteeing security or performance, but much less for enforcing ethical properties such as fairness.

The research on data provenance is certainly relevant, beyond transparency. Also, the use of open data and open software (when possible) simplifies the task of verification.

Many societal and political disputes today are related to computer science and in particular to the management of data. Governments can ban improper behaviors and encourage ethical approaches via laws and regulations. If properly educated, users can leverage the power they have as consumers to choose the software they want to use, individually or via associations. An important role of computer scientists who understand the issues is to explain them to the general population, be they lawmakers, deciders, or simple citizens.

The computer science research community has brought fantastic new tools. Now is the time to learn how to perform this in ethical ways, so that it can best benefit the entire society.

Acknowledgments

The ideas presented here originate primarily from works with Julia Stoyanovich. In particular, they emerged from discussions with her and Gerome Miklau while preparing a tutorial for EDBT[1], as well with Gerhard Weikum while organizing a Dagstuhl workshop[2]. Finally, they have been influenced by works around the Fides Platform[3] with Julia, Gerhard and Gerome, together with Bill Howe, and Arnaud Sahuguet. I would also like to thank Amélie Marian, Benjamin André and Daniel Kaplan for discussions on Personal Data Management, and the members of the French Digital Council, in particular Valérie Peugeot, for lively discussions on topics such as neutrality, computer science education, and digital inclusion.

[1]Data Responsibly: Fairness, Neutrality and Transparency in Data Analysis. EDBT 2016
[2]Data, Responsibly, Dagstuhl Seminar 16291.
[3]Fides: Towards a Platform for Responsible Data Science. SSDBM 2017