Tutorial

Data, Responsibly: Fairness, Neutrality and Transparency in Data Analysis

Julia Stoyanovich
Drexel University
stoyanovich@drexel.edu

Serge Abiteboul
INRIA Saclay & ENS Cachan
serge.abiteboul@inria.fr

Gerome Miklau
UMass Amherst
miklau@cs.umass.edu

ABSTRACT

Big data technology 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 propel economic inequality, destabilize global markets and affirm systemic bias. While the potential benefits of big data are well-accepted, the importance of using these techniques in a fair and transparent manner is rarely considered.

The primary goal of this tutorial is to draw the attention of the data management community to the important emerging subject of responsible data management and analysis. We will offer our perspective on the issue, will give an overview of existing technical work, primarily from the data mining and algorithms communities, and will motivate future research directions.

1. RESPONSIBLE DATA ANALYSIS

Big data technology 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 propel economic inequality, destabilize global markets and affirm systemic bias. While the potential benefits of big data are well-accepted, the importance of using these techniques in a fair and transparent manner is rarely considered.

We will start this tutorial with a brief introduction to foundational concepts of bias, positive and negative discrimination, redlining, and disparate impact. These legal and ethical issues have been attracting attention in the context of big data, and have been receiving coverage in the popular press. We will then identify key properties of responsible data analysis [2], outlined next.

The first property of responsible data analysis if **fairness**, by which we mean *lack of bias*. It is incorrect to assume that insights gained from computation on data are unbiased simply because data was gathered automatically or processing was performed algorithmically. Bias may come from the data, e.g., if a questionnaire contains biased questions, or from the algorithm, reflecting political, commercial, sexual, religious, or other kinds of preferences of its designers.

The second property 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 first focus on meeting common needs well. However to afford equal advantage to a wide variety of users, it is important to support uncommon information and data analysis needs. Such tasks are said to be "in the tail" — they may not be common individually, yet together constitute the overwhelming majority. For instance, Lerman [22] argues that the use of big data can lead to data exclusion and therefore poses risks to those it overlooks.

The third property of responsible data analysis is **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. 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. An important ingredient in transparency is availability of provenance meta-data, which describes who created a dataset and how.

2. OVERVIEW OF TECHNICAL WORK

In a paper that pre-dates big data, Friedman and Nissenbaum [15] give a systematic account of bias in computer systems. The authors identify several representative examples of bias, and develop a taxonomy, classifying bias as pre-existing (societal), technical and emergent (based on use).

More recently, two kinds of technical approaches have been developed. The first are empirical studies that serve to underscore the lack of fairness and transparency in current data analysis practices. In the second category are proposals from the data mining and machine learning communities that aim to make some common task unbiased.

The empirical study of current data-intensive applications aims to identify fairness violations in data analysis practices. This work is critical for understanding the current practice and for motivating research into responsible data use. We will give an overview of existing studies, including the XRay project [21] and the study by Datta et al. [7]. Both studies point to the lack of transparency in the way personal data is used for online ad targeting. We will also present a study by Sweeney [25], which identifies cases of racial discrimination

in online advertising.

Recently, work is beginning to emerge in the machine learning and data mining communities that concerns detecting and avoiding discrimination in classification. Fairness in classification is understood in terms of two goals, namely, individual fairness and group fairness. Individual fairness states that two individuals who are similar w.r.t. a particular classification task should be classified similarly, while group fairness states that the proportion of members of a protected group who are classified positively should be statistically indistinguishable from the proportion of members of the overall population.

Dwork et al. [11] propose a framework for fair classification, based on identifying a probabilistic mapping from individuals to an intermediate representation that achieves both individual and group fairness. This framework assumes that a distance function in the space of the classification task is given. In a follow-up work, Zemel et al. [26] propose a method for learning a class of distance functions and formulate fairness as an optimization problem that both encodes the data, preserving necessary attributes, and obfuscates membership in a protected group.

Feldman et al. [14] propose a formalization of the legal doctrine of disparate impact in the context of classification, and study the problems of disparate impact certification and removal, linking disparate impact to a particular loss function, namely, to the balanced error rate.

Beyond classification, Pedreschi et al. [23] and Kamiran et al. [20] propose formalizations of discrimination in association rule mining and decision tree learning, respectively. The authors then develop ways to mediate the effects of discrimination in these settings.

3. RESEARCH DIRECTIONS

We will conclude the tutorial by surveying works that, while not specifically motivated by responsible data analysis, can be brought to bear on the problem.

We will mention works seeking to provide accurate data mining results about a population while protecting sensitive information about individuals, e.g., [9, 12]. We will also consider some extensive work on provenance [3, 17], especially in the context of data-intensive workflows [5, 8] and in distributed scenarios [18].

In general, the field of program verification is central to the issue of verifying properties such as fairness or non-discrimination. A broad survey of this field is beyond the scope of the tutorial. We will mention zero-knowledge proofs [6, 16], cryptographic techniques by which one party (the prover) can prove to another party (the verifier) that a given statement is true, without conveying any information apart from the fact that the statement is indeed true.

We will briefly discuss several topics related to supporting diverse preferences and information needs of users. This includes works on search result diversification [4] and rank-aware clustering [24]. We will consider another relevant line of work that concerns modeling, interpreting and aggregating user preferences, e.g., [10, 13, 19]. Finally, we will discuss recent work on personal information management [1], where the goal is to empower users to take control of their own data, so as to manage and disseminate it effectively.

4. REFERENCES

[1] S. Abiteboul et al. Managing your digital life. *Commun. ACM*, 58(5), 2015.

[2] S. Abiteboul and J. Stoyanovich. Plaidoyer pour une analyse "responsable" des données. *Le Monde*, October 12, 2015.

[3] P. Agrawal et al. Trio: A system for data, uncertainty, and lineage. In *VLDB*, 2006.

[4] R. Agrawal et al. Diversifying search results. In *WSDM*, 2009.

[5] Y. Amsterdamer et al. Putting lipstick on pig: Enabling database-style workflow provenance. *PVLDB*, 5(4), 2011.

[6] V. Cortier and S. Kremer. Formal models and techniques for analyzing security protocols: A tutorial. *Foundations and Trends in Programming Languages*, 1(3), 2014.

[7] A. Datta et al. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *CoRR*, abs/1408.6491, 2014.

[8] D. Deutch et al. Provenance-based analysis of data-centric processes. *VLDB J.*, 24(4), 2015.

[9] C. Dwork. A firm foundation for privacy. In *CACM*, volume 54, Jan 2011.

[10] C. Dwork et al. Rank aggregation methods for the web. In *WWW*, 2001.

[11] C. Dwork et al. Fairness through awareness. In *Innovations in Theoretical Computer Science*, 2012.

[12] C. Dwork and A. Roth. *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends in Theoretical Computer Science, 2014.

[13] R. Fagin et al. Optimal aggregation algorithms for middleware. In *PODS*, 2001.

[14] M. Feldman et al. Certifying and removing disparate impact. In *ACM SIGKDD*, 2015.

[15] B. Friedman and H. Nissenbaum. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3), 1996.

[16] S. Goldwasser et al. The knowledge complexity of interactive proof systems. *SIAM J. Comput.*, 18(1), 1989.

[17] T. J. Green et al. Provenance semirings. In *PODS*, 2007.

[18] T. J. Green et al. Provenance in ORCHESTRA. *IEEE Data Eng. Bull.*, 33(3), 2010.

[19] M. Jacob et al. A system for management and analysis of preference data. *PVLDB*, 7(12), 2014.

[20] F. Kamiran et al. Discrimination aware decision tree learning. In *ICDM*, 2010.

[21] M. Lécuyer et al. XRay: Enhancing the web's transparency with differential correlation. In *USENIX*, 2014.

[22] J. Lerman. Big data and its exclusions. *Stanford Law Review Online*, 66, 2013.

[23] D. Pedreschi et al. Discrimination-aware data mining. In *ACM SIGKDD*, 2008.

[24] J. Stoyanovich et al. Making interval-based clustering rank-aware. In *EDBT*, 2011.

[25] L. Sweeney. Discrimination in online ad delivery. *ACM Queue*, 11(3), 2013.

[26] R. S. Zemel et al. Learning fair representations. In *ICML*, 2013.