# Fides: Towards a Platform for Responsible Data Science

Julia Stoyanovich, Bill Howe, Serge Abiteboul, Gerome Miklau, Arnaud Sahuguet, Gerhard Weikum

# Fides: Towards a Platform for Responsible Data Science

Julia Stoyanovich*
Drexel University, USA
stoyanovich@drexel.edu

Bill Howe†
University of Washington, USA
billhowe@cs.washington.edu

Serge Abiteboul‡
Inria, France
Serge.Abiteboul@inria.fr

Gerome Miklau§
UMass Amherst, USA
miklau@cs.umass.edu

Arnaud Sahuguet
Cornell Tech, USA
Arnaud.Sahuguet@gmail.com

Gerhard Weikum
MPI Informatics, Germany
weikum@mpi-sb.mpg.de

May 18, 2017

**Abstract**

Issues of responsible data analysis and use are coming to the forefront of the discourse in data science research and practice, with most significant efforts to date on the part of the data mining, machine learning, and security and privacy communities. In these fields, the research has been focused on analyzing the fairness, accountability and transparency (FAT) properties of specific algorithms and their outputs. Although these issues are most apparent in the social sciences where fairness is interpreted in terms of the distribution of resources across protected groups, management of bias in source data affects a variety of fields. Consider climate change studies that require representative data from geographically diverse regions, or supply chain analyses that require data that represents the diversity of products and customers. Any domain that involves sparse or sampled data has exposure to potential bias.

In this vision paper, we argue that FAT properties must be considered as database system issues, further upstream in the data science lifecycle:

*This work was supported in part by NSF Grants No. 1464327 and 1539856, and BSF Grant No. 2014391.

†This work was supported by the University of Washington Information School, Microsoft, the Gordon and Betty Moore Foundation (Award #2013-10-29) and the Alfred P. Sloan Foundation (Award #3835) through the Data Science Environments program.

‡Département d'informatique de l'ENS, École normale supérieure, CNRS, PSL Research University

§This work was supported in part by NSF Grant No. 1409143.

bias in source data goes unnoticed, and bias may be introduced during pre-processing (fairness), spurious correlations lead to reproducibility problems (accountability), and assumptions made during pre-processing have invisible but significant effects on decisions (transparency). As machine learning methods continue to be applied broadly by non-experts, the potential for misuse increases. We see a need for a data sharing and collaborative analytics platform with features to encourage (and in some cases, enforce) best practices at all stages of the data science lifecycle. We describe features of such a platform, which we term Fides, in the context of urban analytics, outlining a systems research agenda in responsible data science.

# 1 Introduction

In all areas of science, government and industry, the rate of data acquisition outpaces the rate of data analysis. New methods and systems have emerged to help make predictions from large, noisy, heterogeneous datasets and deploy results to automate decision-making. But as these technologies continue to be democratized, the potential for misuse increases. In particular, recent advances in data systems research in new uses of existing technology [20], new architectures [24, 33], new execution strategies [8], and language extensions and interfaces [10, 32], drive the democratization of scalable machine learning. Yet, this work has almost exclusively focused on supporting *exploratory research*, characterized by rapid iteration through data gathering, feature engineering, model selection, parameter adjustment, and assessment. This exploratory process can quickly generate promising hypotheses, but must eventually give way to *confirmatory analysis*, characterized by rigorous control for bias in the source data, management of multiple hypothesis testing issues, and considerations of alternative explanations.

We argue here for systematic support of *FAT-Aware Data Science*, provided by a hypothetical platform called Fides. This system must provide common data management and analytics features (scalable query-answering, storage management, access control), along with a set of new capabilities to enable a full FAT-aware data lifecycle. Consider these examples of "upstream" challenges that can complicate "downstream" fairness, accountability, and transparency, and the support that Fides will provide to mitigate them:

- To facilitate collaboration with a team of data scientists, a county official uploads a dataset $D$ of homeless citizens registered in a transitional housing program. The data scientists wish to predict which citizens are likely to find permanent housing based on job experience, geography, demographics, health, substance abuse, and other features. They are unaware that men tend to be less likely to provide their data than women, introducing a bias in the results. Capturing this kind of domain knowledge as an annotation of the dataset enables automatic correction during downstream analytics.

- A previous analysis was performed with a different version of $D$, raising the risk of inconsistent and incomparable results. By inspecting the datasets, the Fides system can detect the similarity between the two versions and prompt the data scientists accordingly, perhaps rerunning the analysis on the new version automatically.

- Several students participate in the project, each studying the likelihood of permanent housing for different subpopulations of homeless citizens. On a hunch, a student looks at recently homeless women aged 25-30 in a particular neighborhood and discovers a statistically significant result. But this signal could easily be attributed to chance, due to a number of closely related hypotheses the students are testing (different neighborhoods, different age groups, different genders). To control for these issues related to multiple hypothesis testing, Fides automatically establishes a reusable holdout set [17], and restricts access to the underlying data during the exploratory phase to preserve statistical validity.

- A data scientist collaborating with the county creates a model to produce a ranked list of at-risk families for targeted outreach. The data on which the model was trained was not representative of the county's population, leading to underrepresentation of protected groups among the top-ranked families. Fides is configured to automatically diagnose violations of statistical parity constraints along protected attributes, alert the data scientist, and propose ways to mitigate the violations with no change to the data analysis code.

- Civic groups question the validity of the model and demand evidence of equal treatment. Although the code can be shared publicly, the underlying data cannot. Fides generates a shareable dataset satisfying differential privacy [18] (where individual records may or may not hold true information) to support what-if analysis of the model without violating privacy.

These examples illustrate practical challenges and solutions to achieve FAT in realistic collaborative data science settings where the necessary domain knowledge, statistical expertise, and programming skill are rarely held by the one person. In such settings, the potential for spurious results, non-compliance with applicable laws, and reinforcement of existing disparities increases.

We consider these issues in the context of a four step model of the data-intensive research lifecycle:

- *data acquisition and curation* where relevant datasets are found in the repository or ingested from external sources, cleaned, transformed, combined, and annotated,

- *exploratory research* where false positives and biases may be tolerable in the interest of identifying promising leads,

- *confirmatory analysis* where rigor and reproducibility are paramount, and

- *operational deployment* where results of analysis are made available to users, and are used to enact decisions that affect the world.

While it is increasingly recognized that results of algorithmic analysis have material effects on people's lives [27], current research in fairness, accountability and transparency in machine learning tends to focus on the problems that arise once the models are deployed in practice. We argue here that these properties cannot be effectively enforced if they are not considered earlier in the lifecycle. We advocate for *FAT by design* — an approach where responsibility for ensuring these properties begins as soon as datasets with unknown biases and unclear provenance are collected, cleaned and integrated, continues through model development and deployment phases, and persists through result interpretation and iterative refinement. Statistical rigor can no longer be considered a secondary concern, and *systematic support* is needed to enable a full lifecycle of responsible data-intensive algorithmic processes, preventing their misapplication and mitigating the corresponding societal harms.

In this vision paper, we describe the capabilities of a hypothetical FAT-aware data system Fides, and outline a technical research agenda in this area. The capabilities we describe at each level are summarized in Figure 1.

We envision Fides as a multi-tenant software-as-a-service platform, for four reasons. First, some features (e.g., those relying on differential privacy) involve computing a query response as a function of the entire query history against a particular dataset, even across users. Only a shared platform has access to this global information. Second, some features require training models to automatically annotate datasets; the efficacy of these models is a function of the amount of training data to which they have access, motivating global access. Third, some features involve Fides serving as an honest broker for secure data access; the necessary trust relationship is difficult to enforce if the system can be installed locally. Fourth, centralized or managed installation and configuration of database systems has been shown to improve uptake for collaborative data science [7, 22].

## 2 Data Acquisition and Curation

Contextual information and rich metadata can help prevent misinterpretation of results downstream. There is a limited window of opportunity to attach this information when data is first collected (or first brought into a managed environment). Features to automate or facilitate metadata attachment, data curation, type inference, and annotation entry help maximize this opportunity.

**Domain knowledge capture** To enable the system to automatically detect bias in statistical results, a model of the population from which a dataset is drawn must be available. In rigorous statistical contexts, a model of the underlying population may be assumed explicitly, but in data science contexts, data may be "inherited" from unmanaged (or even untrusted) sources.

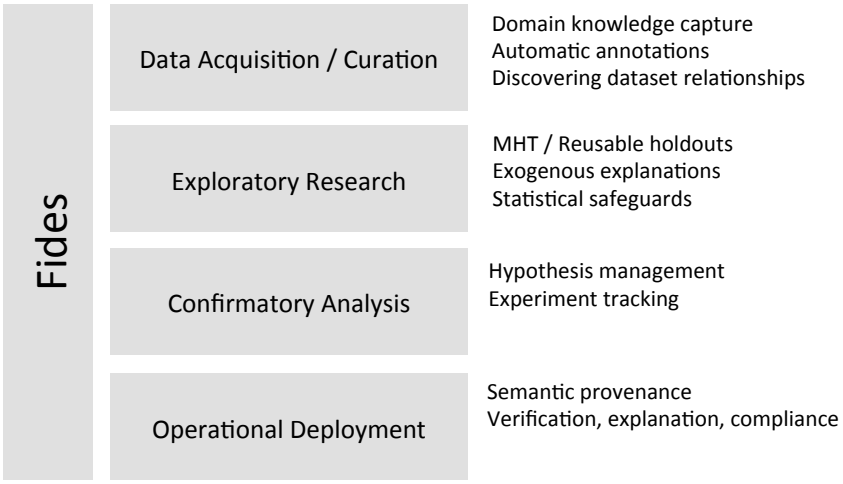


Figure 1: Proposed services for Fides, a FAT-aware data science platform. Fairness, accountability, and transparency must be managed at all stages of the lifecycle of data-intensive applications.

Fides can allow data owners to assert domain knowledge that can be used to quantify and correct for bias downstream. Datasets are modeled as select-project queries over a virtual population relation. Since the population dataset is not known in practice, domain knowledge is represented as an assertion of a conditional probability on specific attributes (or combinations of attributes). For example, Figure 2 shows three datasets derived from a common (virtual) dataset of all people.

Assertions about the underlying population are captured as (possibly multivariate) statements of prior probability; for example $P(gender = \texttt{F}, age < 30) = 0.23$. Assertions about population distributions of age and gender can be captured by the Fides system and used to assess bias in the sampled datasets. For example, if the school-age youth are disproportionately male, one might infer that the sampling was biased [6], and the results of downstream analyses could be automatically flagged. For complex multi-variate distributions, the priors may not be known about the population in general. In some of these cases, the priors may be computed from a reference dataset such as the Census. In other cases, domain experts may make assertions about the population based on their experience, possibly refining the global population model. Propagation of these flags throughout the analysis can perhaps be managed in a manner similar to taint analysis for secure Web programming [13].

**Automatic annotations of sensitive data** Data owners must have the means to effectively specify and verify how their data is disseminated and

used. Data use policies must be sufficiently expressive to specify fine-grained access [23, 26, 31], and to assign attribute labels indicating anonymity requirements or protected status. It is unrealistic to rely on the data owners to configure these annotations manually, or to ensure that annotations are maintained properly as data undergoes transformations due to cleaning, integration and analysis. For this reason, it is essential that Fides support (1) automatic or semi-automatic annotation of input data, (2) usable and flexible specification of access control policies, (3) automatic propagation of annotations through all stages of the data analysis lifecycle, (4) verification and explanation of data access and use.

To support semi-automatic curation, Fides can learn an annotation model from the shared corpus of sensitive data. When a new dataset is uploaded, semantic types of each attribute can be inferred automatically based on, for example, column names (e.g., `gender`, `M/F`, or `sex`) or values (e.g., "F," "female,"). Records with sensitive content (images with faces, medical conditions) can be flagged automatically and held back from downstream analysis. These annotations can be automatically propagated throughout complex queries as data is exchanged and used in computation [26].

**Discovering relationships between datasets** A dataset presented to Fides may be related to other datasets that were uploaded previously: different versions, different subsets, different transformations. Disambiguating the derivation structure of these datasets can help prevent analyses from using the incorrect version of a dataset, inadvertently changing the outcome. For example, Reinhart and Rogoff omitted values in a spreadsheet from their analysis that motivated a global economic policy in austerity; inclusion of the missing values changed the conclusions [21]. Aliwani et al. studied the relationship mining problem in the context of uploading spreadsheets to a shared repository [2].

# 3 Exploratory Research

Once the data is ingested into Fides, analysts and collaborators iteratively mine the data for interesting relationships. However, testing of many potential hypotheses can lead to spurious results: one should not be surprised when a p-value is less than 0.05 after running 20 different experiments! A shared system has the global scope required to provide protection against these issues.

**Multiple hypothesis testing and reusable holdouts** There are robust methods to account for multiple hypothesis testing errors, including controlling for the family-wise error rate and, less conservatively, the false discovery rate. Recently, Dwork et al. proposed a mechanism derived from differential privacy for shared use of common datasets that involves controlling the use of a reusable holdout set [17]. Enforcing the protection of this reusable holdout set can *only* be implemented in a shared environment; independent local copies of the same dataset would lead to independent reusable holdout sets.

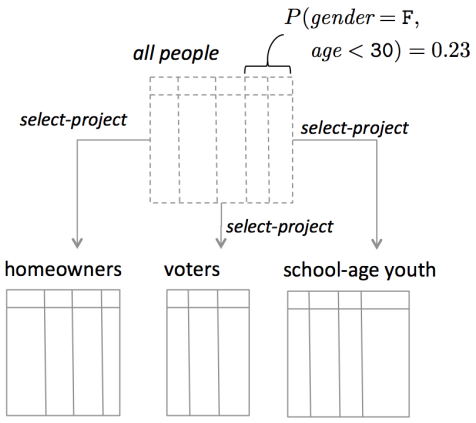


Figure 2: Datasets are represented as select-project queries over a virtual population relation. Domain knowledge is captured as a distribution of specific attributes of the population, allowing biased datasets to be detected. In this case, age and gender are attributes with known distribution; in some contexts the data model for the overall population may only be known by a few domain experts.

**Exogenous explanations** Results that are surprising to domain-agnostic analysts are not necessarily surprising to domain experts. For example, a significant reduction in ridership on public transportation during a three-day period may be tempting to interpret as an effect of a pricing change, but transportation engineers may know that inclement weather leads to similar patterns. These *exogeneous explanations* are important to bring into scope in preventing the publication of spurious results. Fides can support the discovery and application of exogenous explanations through finding similar signals in other data sources [12], crowdsourcing, or crawling the Web for events that coincide with spatio-temporal context. Fides will incorporate pattern finding services in the manner of Google Correlate, which given a time series will return other time series with similar patterns. For example, in our transportation example, Fides can return other instances of low ridership (e.g., bike) that can indicate inclement weather.

**Statistical safeguards** A linear regression model presumes the variables are normally distributed, but rarely will analysts take the time to check that the data satisfies typical normality tests. Fides could check these assumptions automatically as analyses are applied to prevent inappropriate usage.

# 4 Confirmatory Analysis

Once exploratory research has exposed a tantalizing relationship, a new level of statistical rigor is warranted to evaluate the hypothesis, perhaps requiring the use of a new dataset. System level support in hypothesis and experiment management is warranted.

**Hypothesis management** Management of a "stream of questions" is just as important as managing a stream of data. As mentioned, tracking questions posed by users allows the implementation of multiple hypothesis testing countermeasures, but also helps combat publication bias in which negative results are suppressed, giving a false confidence in positive results. By managing and exposing the questions being investigated, Fides can provide context for statistical results within a larger space of experiments, and help motivate others to design and conduct new experiments, and to collect datasets to answer the questions.

**Experiment tracking** It is rarely possible to detect and account for selection bias in observational data [6]. To answer causality questions of the form "What would happen if...," one must perturb the system and assess the results. Fides can support experiment management to collect new datasets: Setting up a randomized controlled trial, tracking the results, and bringing the data back into the managed environment. These datasets will be equipped with provenance about not only the experiment that generated them, but potentially the line of questions that motivated the experiment in the first place.

# 5 Operational Deployment

Once a finding is confirmed, the operationalization of the model into practice can have a direct impact on people's lives. As such, model deployment can and should lead to new scrutiny from citizens, law enforcement, legal scholars, legislators, watchdogs, and more. Techniques for black box verification of properties (via, for example, zero-knowledge proofs) are crucial, but systems capabilities based on provenance management offer a pragmatic and scalable approach.

**Semantic provenance.** Provenance is metadata that describes the origin and the history of derivation of a data item. Provenance tools and techniques can be used by Fides to enable the accountability and transparency aspects of FAT-aware data science. Provenance annotations that are captured during the data acquisition and curation phase should be propagated automatically thorough data analysis. Provenance traces accompanying the results can then be interrogated by the user who wishes to: (1) verify quality, fairness and robustness of processes and results; (2) check appropriateness of data access and use, and properly attribute credit; and (3) enable causal what-if analysis.

Provenance propagation and interrogation in Fides will build on a large body of work in the workflows community [3, 15], where provenance is typically coarse-grained and records the sequence of functional steps that led to a result, and in the database community [4, 9, 19], where provenance captures the fine-grained dependencies between tuples in the input and those in the result.

A practical challenge that must be addressed is that current provenance methods suffer from an information glut. The set of all fine-grained facts related to a particular result does not typically provide actionable intuition for a user. We refer to this "dumping" approach as *syntactic provenance*. In contrast, we advocate a research agenda in *semantic provenance*, where the relevant properties that affect interpretation of the results are automatically discovered and exposed.

A promising direction towards semantic provenance is identifying and displaying to the user the *core provenance* of a set of tuples [5] — provenance annotations that are both compact and informative in exposing the core of the derivation. It is also promising to derive compact and meaningful representations in the form of provenance views, which have so far been studied in the context of privacy [16] and should now be considered with the objectives of compactness, informativeness and, more generally, usability. Finally, work on causality in databases [25] has deep connections to provenance and can help answer causal what-if questions.

**Verification, Explanation and Compliance** Verification in a FAT-aware system addresses two questions: To what degree do required properties hold, and how do we convince external stakeholders that these properties hold? Data and its use in analyses must be scrutinized against best practice standards before it becomes actionable. Newly-developed and emerging methods for verifying and explaining algorithmic processes and their outcomes will benefit from systematic support in Fides. Support for this kind of compliance testing is absent in today's data management systems, but is crucial for next-generation data science.

*Verifying fairness properties.* Fairness involves ensuring that algorithmic decisions of hiring, resource allocation, sentencing, and more obey relevant laws and societal norms. Without explicit control for fairness and diversity, these algorithms frequently reinforce and amplify existing inequities. Sampling procedures for relational databases have been studied for decades [1, 11, 28], often focused on how sampling operators can be designed to commute with relational algebra operators to improve performance while maintaining statistical properties. New procedures for ensuring fairness via statistical parity with respect to protected attributes are warranted. There has also been significant recent work on quantifying the influence of properties of the input (e.g., values of individual attributes, or of combinations of attributes) on properties of the output (e.g., fairness) [14], and on explicit debugging for "fairness bugs" [30]. These techniques will be useful at different stages of the data science lifecycle, but become especially crucial during operational deployment when algorithmic decisions begin to affect people's daily lives.

*Explanations.* Data collection, analysis and results must be comprehensible and defensible to a number of stakeholders, including end-users, commercial competitors, auditors, policy makers, and the public. *Syntactic transparency*, where the code and even the data on which the algorithms operate is fully disclosed, can still leave stakeholders in the dark [29]. Fides must enable *interpretability*, which rests on making explicit the interactions between the program and the data on which it acts, and exposes biases in data collection and analysis. This is particularly important when algorithms are performing a public function (e.g., allocation of public resources) or directly shaping the public sphere (e.g., ranking politicians).

When generating explanations, the system must be mindful of the trade-off between offering transparency and accountability on the one hand, and adhering to privacy and data use policies on the other hand. Further, great care must be taken to offer explanations that are both relevant to a particular user and interpretable by that user. To address both challenges, Fides will generate explanation that are tailored to a user's access rights, information need and level of expertise. With access to explanations, the public can understand and shape policy decisions, question the normative judgments that form basis of algorithmic processes and decisions, and evaluate the effectiveness of enacted policies.

# 6 Conclusions

We envision a research agenda in "FAT by design" data systems, where systems that support data management, sharing, and analysis are augmented with built-in support for FAT concerns at all stages of the data lifecycle. Research to consider these systems issues will help democratize access to FAT-aware machine learning methods.

With this paper, we hope to motivate research in data sharing systems that can provide a "delivery vector" for new methods and techniques for ensuring fairness, accountability, and transparency, and to encourage collaborative projects between systems, methods, and application researchers.

# References


[1] Sameer Agarwal, Barzan Mozafari, Aurojit Panda, Henry Milner, Samuel Madden, and Ion Stoica. Blinkdb: Queries with bounded errors and bounded response times on very large data. In *EuroSys*, pages 29–42, New York, NY, USA, 2013. ACM.

[2] Abdussalam Alawini, David Maier, Kristin Tufte, and Bill Howe. Helping scientists reconnect their datasets. In *SSDBM*, pages 29:1–29:12, 2014.

[3] Ilkay Altintas, Oscar Barney, and Efrat Jaeger-Frank. Provenance collection support in the kepler scientific workflow system. In *IPAW*, pages 118–132, 2006.

[4] Yael Amsterdamer, Susan B. Davidson, Daniel Deutch, Tova Milo, Julia Stoyanovich, and Val Tannen. Putting lipstick on pig: Enabling database-style workflow provenance. *PVLDB*, 5(4):346–357, 2011.

[5] Yael Amsterdamer, Daniel Deutch, Tova Milo, and Val Tannen. On provenance minimization. *ACM Trans. Database Syst.*, 37(4):30, 2012.

[6] Elias Bareinboim, Jin Tian, and Judea Pearl. Recovering from selection bias in causal and statistical inference. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pages 2410–2416. AAAI Press, 2014.

[7] Anant P. Bhardwaj, Souvik Bhattacherjee, Amit Chavan, Amol Deshpande, Aaron J. Elmore, Samuel Madden, and Aditya G. Parameswaran. Datahub: Collaborative data science & dataset version management at scale. In *CIDR*, 2015.

[8] Matthias Boehm, Shirish Tatikonda, Berthold Reinwald, Prithviraj Sen, Yuanyuan Tian, Douglas R. Burdick, and Shivakumar Vaithyanathan. Hybrid parallelization strategies for large-scale machine learning in systemml. *Proc. VLDB Endow.*, 7(7):553–564, March 2014.

[9] Peter Buneman, Sanjeev Khanna, and Wang Chiew Tan. Why and where: A characterization of data provenance. In *ICDT*, pages 316–330, 2001.

[10] Zhuhua Cai, Zografoula Vagena, Luis Perez, Subramanian Arumugam, Peter J. Haas, and Christopher Jermaine. Simulation of database-valued markov chains using simsql. In *ACM SIGMOD*, pages 637–648, 2013.

[11] Surajit Chaudhuri, Gautam Das, and Vivek Narasayya. Optimized stratified sampling for approximate query processing. *ACM Trans. Database Syst.*, 32(2), June 2007.

[12] Fernando Chirigati, Harish Doraiswamy, Theodoros Damoulas, and Juliana Freire. Data polygamy: The many-many relationships among urban spatio-temporal data sets. In *ACM SIGMOD*, pages 1011–1025, 2016.

[13] James Clause, Wanchun Li, and Alessandro Orso. Dytan: A generic dynamic taint analysis framework. In *ISSTA*, pages 196–206, 2007.

[14] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE SP*, pages 598–617, 2016.

[15] Susan B. Davidson and Juliana Freire. Provenance and scientific workflows: challenges and opportunities. In *ACM SIGMOD*, pages 1345–1350, 2008.

[16] Susan B. Davidson, Tova Milo, and Sudeepa Roy. A propagation model for provenance views of public/private workflows. In *ICDT*, pages 165–176, 2013.

[17] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *NIPS*, pages 2350–2358, 2015.

[18] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[19] Todd J. Green, Gregory Karvounarakis, and Val Tannen. Provenance semirings. In *PODS*, pages 31–40, 2007.

[20] Joseph M. Hellerstein, Christoper Ré, Florian Schoppmann, Daisy Zhe Wang, Eugene Fratkin, Aleksander Gorajek, Kee Siong Ng, Caleb Welton, Xixuan Feng, Kun Li, and Arun Kumar. The madlib analytics library: Or mad skills, the sql. *Proc. VLDB Endow.*, 5(12):1700–1711, August 2012.

[21] Thomas Herndon, Michael Ash, and Robert Pollin. Does high public debt consistently stifle economic growth? a critique of reinhart and rogo ff. Technical report, Political Economy Research Institute, UMass Amherst, 2013.

[22] Shrainik Jain, Dominik Moritz, Daniel Halperin, Bill Howe, and Ed Lazowska. SQLShare: Results from a multi-year SQL-as-a-service experiment. In *ACM SIGMOD*, pages 281–293, 2016.

[23] Kristen LeFevre, Rakesh Agrawal, Vuk Ercegovac, Raghu Ramakrishnan, Yirong Xu, and David J. DeWitt. Limiting disclosure in hippocratic databases. In *VLDB*, pages 108–119, 2004.

[24] Yucheng Low, Danny Bickson, Joseph Gonzalez, Carlos Guestrin, Aapo Kyrola, and Joseph M. Hellerstein. Distributed graphlab: A framework for machine learning and data mining in the cloud. *Proc. VLDB Endow.*, 5(8):716–727, April 2012.

[25] Alexandra Meliou, Sudeepa Roy, and Dan Suciu. Causality and explanations in databases. *PVLDB*, 7(13):1715–1716, 2014.

[26] Vera Zaychik Moffitt, Julia Stoyanovich, Serge Abiteboul, and Gerome Miklau. Collaborative access control in webdamlog. In *ACM SIGMOD*, pages 197–211, 2015.

[27] Cecilia Muñoz, Megan Smith, and DJ Patil. Big data: A report on algorithmic systems, opportunity, and civil rights. *The White House*, May 2016.

[28] Frank Olken and Doron Rotem. Simple random sampling from relational databases. In *VLDB*, pages 160–169, 1986.

[29] Julia Stoyanovich and Ellen P. Goodman. Revealing algorithmic rankers. *Freedom to Tinker*, August 5, 2016.

[30] Florian Tramèr, Vaggelis Atlidakis, Roxana Geambasu, Daniel J. Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Discovering unwarranted associations in data-driven applications with the fairtest testing toolkit. *CoRR*, abs/1510.02377, 2015.

[31] Qihua Wang, Ting Yu, Ninghui Li, Jorge Lobo, Elisa Bertino, Keith Irwin, and Ji-Won Byun. On the correctness criteria of fine-grained access control in relational databases. In *VLDB*, pages 555–566, 2007.

[32] Reynold S. Xin, Josh Rosen, Matei Zaharia, Michael J. Franklin, Scott Shenker, and Ion Stoica. Shark: Sql and rich analytics at scale. In *ACM SIGMOD*, pages 13–24, 2013.

[33] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. In *HotCloud*, 2010.