TransFAT: Translating Fairness, Accountability and Transparency into Data Science Practice

Julia Stoyanovich

New York University
New York NY, USA
stoyanovich@nyu.edu

Abstract. Data science holds incredible promise for improving peoples lives, accelerating scientific discovery and innovation, and bringing about positive societal change. Yet, if not used *responsibly* — in accordance with legal and ethical norms — the same technology can reinforce economic and political inequities, destabilize global markets, and reaffirm systemic bias. In this paper I discuss an ongoing regulatory effort in New York City, where the goal is to develop a methodology for enabling responsible use of algorithms and data in city agencies. I then highlight some ongoing work that makes part of the Data, Responsibly project, aiming to operationalize fairness, diversity, accountability, transparency, and data protection at all stages of the data science lifecycle. Additional information about the project, including technical papers, teaching materials, and open-source tools, is available at dataresponsibly.github.io.

Keywords: responsible data science · fairness · diversity · transparency

1 Introduction

Data science holds incredible promise for improving peoples lives, accelerating scientific discovery and innovation, and bringing about positive societal change. Yet, if not used *responsibly* — in accordance with legal and ethical norms — the same technology can reinforce economic and political inequities, destabilize global markets, and reaffirm systemic bias [1,4,6,7,14,17].

The public sector is under particular pressure to fulfill the mandate for responsibility: All decisions made by algorithms will be scrutinized by the affected individuals and groups, and by the taxpayers who are entitled to verify equitable resource distribution. Yet, recent reports on data-driven decision making, specifically in the public sector, underscore that fairness and equitable treatment of individuals and groups is difficult to achieve [15], and that transparency and accountability of algorithmic processes are indispensable but rarely enacted [1,5]. As a society, we cannot afford the status quo: Algorithmic bias in administrative processes limits access to resources for those who need these resources most, and amplifies the effects of systemic historical discrimination. Lack of transparency and accountability threatens the democratic process itself.

PIE 2019, June 4, 2019, Rome, Italy.

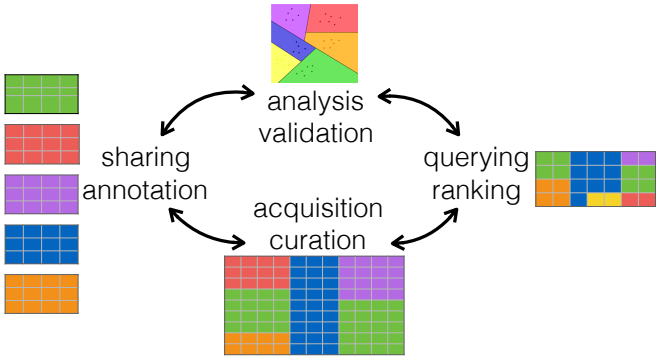

Fig. 1. The data usage lifecycle.

How can the technical community support responsible data science practices in complex administrative processes? Researchers are actively working on methods for enabling fairness, accountability and transparency (FAT) of specific algorithms and their outputs [9,10,11,13,18,28]. While important, these approaches focus solely on the analysis and validation step of the data science lifecycle (depicted in Figure 1), and operate under the assumption that input datasets are clean and reliable.

To realize the limitations of this assumption, observe that additional information and intervention methods are available if we consider the upstream process that generated the input data [21]. Appropriately annotating datasets when they are shared, and maintaining information about how datasets are acquired and manipulated, allows us to provide data transparency: to explain statistical properties of the datasets, uncover any sources of bias, and make statements about data quality and fitness for use. Put another way: if we have no information about how a dataset was generated and acquired, we cannot convincingly argue that it is appropriate for use by an automated decision system.

In the remainder of this paper, I will further motivate technical work on responsible data science in the context of an ongoing regulatory effort (Section 2). I will then highlight some work that makes part of the Data, Responsibly project (Section 3). For additional information about the project, including technical papers, teaching materials, and open-source tools, see `dataresponsibly.github.io`.

2 Towards a Data Transparency Framework

New York City is the first municipality in the United States to attempt to regulate the use of data-driven algorithmic decision making in government. The City passed Local Law 49 of 2018 [23], requiring that a task force be put in place to survey the current use of "automated decision systems," defined as "computerized implementations of algorithms, including those derived from machine

learning or other data processing or artificial intelligence techniques, which are used to make or assist in making decisions," in City agencies. The task force will develop a set of recommendations for enacting algorithmic transparency by the agencies, and will propose procedures for:

- requesting and receiving an explanation of an algorithmic decision affecting an individual (Section 3(b));
- interrogating automated decision systems for bias and discrimination against members of legally protected groups, and addressing instances in which a person is harmed based on membership in such groups (Sections 3(c), 3(d));
- assessing how automated decision systems function and are used, and archiving the systems together with the data they use (Sections 3(e), 3(f)).

Local Law 49 of 2018 in effect mandates the development of an algorithmic transparency framework. In the remainder of this section, I argue that meaningful transparency of algorithmic processes cannot be achieved without transparency of data.

What is data transparency? In applications involving predictive analytics, data is used to customize generic algorithms for specific situations — we say algorithms are *trained* using data. The same algorithm may exhibit radically different behavior — make different predictions; make a different number of mistakes, and even different kinds of mistakes – when trained on two different datasets. In other words, without access to the training data, it is impossible to know how an algorithm would actually behave.

Algorithms and corresponding training data are used, for example, in predictive policing applications to target areas or people that are deemed to be high-risk. But as has been shown extensively, when the data used to train these algorithms reflects the systemic historical bias towards poor and predominately African American neighborhoods, the predictions will simply reinforce the status quo rather than provide any new insight into crime patterns. The transparency of the algorithm is neither necessary nor sufficient to understand and counteract these particular errors. Rather, the conditions under which the data was collected must be retained and made available to make the decision-making process transparent.

Even those decision-making applications that do not explicitly attempt to predict future behavior based on past behavior are still heavily influenced by the properties of the underlying data. For example, the VI-SPDAT [12] risk assessment tool, used to prioritize homeless individuals for receiving services, does not involve machine learning, but still assigns a risk score based on survey responses — a score that cannot be interpreted without understanding the conditions under which the data was collected. As another example: Matchmaking methods such as those used by the Department of Education to assign children to spots in public schools are designed and validated using datasets; if these datasets are not made available, the matchmaking method itself cannot be considered transparent.

What is data transparency, and how can we achieve it? One immediate interpretation of this term is "making the training and validation datasets publicly available." However, while data should be made open whenever possible, much of it is sensitive and cannot be shared directly. That is, *data transparency is in tension with the privacy of individuals who are included in the dataset.* In light of this, an alternative interpretation of data transparency is as follows:

- In addition to releasing training and validation datasets whenever possible, agencies shall make publicly available summaries of relevant statistical properties of the datasets that can aid in interpreting the decisions made using the data, while applying state-of-the-art methods to preserve the privacy of individuals.
- When appropriate, privacy-preserving synthetic datasets can be released in lieu of real datasets to expose certain features of the data, if real datasets are sensitive and cannot be released to the public.

An important aspect of data transparency is interpretability — surfacing the statistical properties of a dataset, the methodology that was used to produce it, and, ultimately, substantiating its "fitness for use" in the context of a specific automated decision system or task. This consideration of a specific use is particularly important because datasets are increasingly used outside the original context for which they were intended. This compels us to augment our interpretation of data transparency in the public sector to include:

- Agencies shall make publicly available information about the data collection and pre-processing methodology, in terms of assumptions, inclusion criteria, known sources of bias, and data quality.

Data transparency is important both when an automated decision system is interrogated for systematic bias and discrimination, and when it is asked to explain an algorithmic decision that affects an individual. For example, suppose that a system scores and ranks individuals for access to a service. If an individual enters her data and receives the result — say, a score of 42 — this number alone provides no information about why she was scored in this way, how she compares to others, and what she can do to potentially improve her outcome.

To facilitate transparency, *the explanation given to an individual should be interpretable, insightful and actionable.* As part of the result, data that pertains to other individuals, or a summary of such data, may need to be released, for example, to explain which other individuals, or groups of individuals, receive a higher score, or a more favorable outcome. This functionality requires data transparency mechanisms discussed above.

3 Highlights of the Data, Responsibly Project

The goal of the Data, Responsibly project is to develop a foundational understanding of responsible data science at all stages of the data lifecycle [21], and to

translate that understanding into tools and platforms [16,27]. Such tools should be placed in the hands of data practitioners in the public sector. Importantly, the requirement of responsibility cannot be handled as an afterthought, but must be provisioned for at design time. In the remainder of this section, I highlight several recent technical results. To keep the discussion focused, I will discuss results that pertain to *ranking* and *set selection* tasks.

Algorithmic decisions often result in scoring and ranking individuals to determine credit worthiness, qualifications for college admissions and employment, and compatibility as dating partners. While automatic and seemingly objective, ranking algorithms can discriminate against individuals and protected groups, and exhibit low diversity. Furthermore, ranked results are often unstable — small changes in the input data or in the ranking methodology may lead to drastic changes in the output, making the result uninformative and easy to manipulate. Similar concerns apply in cases where items other than individuals are ranked, including colleges, academic departments, or products. Finally, even in cases where both the data and the ranking method are publicly available, ranked results may still be difficult to interpret. [20]

In addition to being commonly used in the analysis and validation stage of the data science lifecycle, set selection and ranking are also very common *upstream* from data analysis, in data sharing, acquisition, integration, and querying (see Figure 1), making this family of methods particularly important to study.

3.1 Fairness and diversity in ranking and set selection

In [25] we started an inquiry into fairness in ranked outputs. We considered the setting in which an institution, called a ranker, evaluates a set of individuals based on demographic, behavioral or other characteristics. The final output is a ranking that represents the relative quality of the individuals. While automatic and therefore seemingly objective, rankers can, and often do, discriminate against individuals and systematically disadvantage members of protected groups.

In this work we focused on datasets in which items have a single binary sensitive attribute, such as male or female gender, and minority or majority ethnic group, with one of the groups designated as the *protected group* (the groups that experienced a historical disadvantage). We proposed a family of fairness measures, quantifying the relative representation of protected group members at discrete points in the ranking (e.g., top-10, top-20, etc.), and compounding these proportions with a logarithmic discount, in the style of information retrieval.

Score-based set selection is a mechanism that closely related to ranking. Selection algorithms usually score individual items in isolation, and then select the top scoring items. However, often there is an additional diversity objective — selecting high-quality items that have different attributes (as in product recommendation systems), or highs-scoring individuals who belong to different demographic, geographic or socio-economic groups (as in college admissions and hiring). In a recent work [22] we proposed methods for enforcing diversity in online set selection, where a decision must be made on each item as it is presented.

We showed through experiments with real and synthetic data that diversity can be achieved, usually with modest costs in terms of quality.

Our experimental evaluation lead to several important insights in online set selection. Most importantly, we showed that if a difference in scores is expected between groups (e.g., due to historical disadvantage), then these groups must be treated separately during processing. Otherwise, a solution may be derived that meets diversity constraints, but that selects lower-scoring members of disadvantaged groups. This insight supports the argument of responsibility by design.

In a recent follow-up work [24], we studied an unintended consequence of applying diversity constraints to set selection and ranking, in datasets with multiple sensitive attributes (e.g., gender and race). We observed that maximizing utility (sum of item scores) subject to diversity constraints leads to reduced in-group fairness: the selected candidates from a given group may not be the best ones, and this unfairness may not be well-balanced across groups.

We studied this phenomenon using datasets that comprise multiple sensitive attributes. We then introduce additional constraints, aimed at balancing in-group fairness across groups, and formalized the induced optimization problems as integer linear programs. Using these programs, we conducted an experimental evaluation with real datasets, and quantified the feasible trade-offs between balance and overall performance in the presence of diversity constraints.

Finally, we considered the design of fair score-based ranking functions in [3]. Items from a database are often ranked based on a combination of criteria. The weight given to each criterion in the combination can greatly affect the fairness of the produced ranking, for example, systematically preferring men over women. A user may have the flexibility to choose combinations that weigh these criteria differently, within limits. In this work, we developed a system that helps users choose criterion weights that lead to greater fairness.

We considered ranking functions that compute the score of each item as a weighted sum of (numeric) attribute values, and then sort items on their score. Each ranking function can be expressed as a point in a multi-dimensional space. For a broad range of fairness criteria, including proportionality, we showed how to efficiently identify regions in this space that satisfy these criteria. Using this identification method, our system is able to tell users whether their proposed ranking function satisfies the desired fairness criteria and, if it does not, to suggest the smallest modification that does. Our extensive experiments on real datasets demonstrated that our methods are able to find solutions that satisfy fairness criteria effectively (usually with only small changes to proposed weight vectors) and efficiently (in interactive time, after some initial pre-processing).

3.2 Stability in ranking

Decision making is challenging when there is more than one criterion to consider. In such cases, it is common to assign a goodness score to each item as a weighted sum of its attribute values and rank them accordingly. Clearly, the ranking depends on the weights used for this summation. Ideally, one would want the ranked order not to change if the weights are changed slightly. We call

this property stability of the ranking. A consumer of a ranked list may trust the ranking more if it has high stability. A producer of a ranked list prefers to choose weights that result in a stable ranking, both to earn the trust of potential consumers and because a stable ranking is intrinsically likely to be more meaningful.

In a recent paper [2], we developed a framework that can be used to assess the stability of a provided ranking and to obtain a stable ranking within an acceptable range of weight values (called "the region of interest"). Using a geometric interpretation, we proposed algorithms that produce stable rankings, and experimentally validates our methods on real datasets. In our ongoing work we are developing methods to quantify and improve stability of rankings under slight changes to the data.

3.3 Interpretability with Nutritional Labels

In a recent paper we presented Ranking Facts, a Web-based application that generates a *"nutritional label"* for rankings. [26]. Ranking Facts is made up of a collection of visual widgets that implement our latest research results on fairness, diversity, stability, and transparency for rankings, and that communicate details of the ranking methodology, or of the output, to the end user. Figure 2 presents Ranking Facts for CS department rankings. The nutritional label consists of six widgets, each with an overview and a detailed view.

The Recipe widget succinctly describes the ranking algorithm. For example, for a linear scoring formula, each attribute would be listed together with its weight. The Ingredients widget lists attributes most material to the ranked outcome, in order of importance. For example, for a linear model, this list could present the attributes with the highest learned weights. Put another way, the explicit intentions of the designer of the scoring function about which attributes matter, and to what extent, are stated in the Recipe, while Ingredients may show additional attributes associated with high rank. Such associations can be derived with linear models or with other methods, such as rank-aware similarity in our prior work [19].

The Stability widget explains whether the ranking methodology is robust on the given dataset. An unstable ranking is one where slight changes to the data (e.g., due to uncertainty and noise), or to the methodology (e.g., by slightly adjusting the weights in a score-based ranker) could lead to a significant change in the output.

The Fairness widget quantifies whether the ranked output exhibits statistical parity (one interpretation of fairness) with respect to one or more sensitive attributes, such as gender or race. The Diversity widget shows diversity with respect to a set of demographic categories of individuals, or a set of categorical attributes of other kinds of items [8]. The widget displays the proportion of each category in the top-10 ranked list and over-all, and, like other widgets, is updated as the user selects different ranking methods or sets different weights.

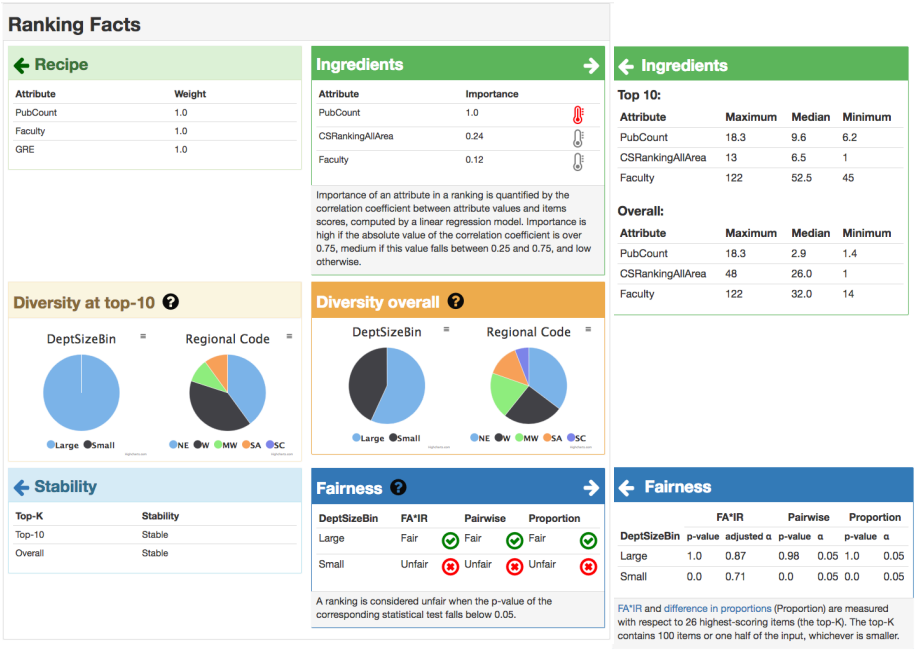

Fig. 2. Ranking Facts for the CS departments dataset (`https://github.com/emeryberger/CSRankings`). The Ingredients widget (green) has been expanded to show the details of the attributes that strongly influence the ranking. The Fairness widget (blue) has been expanded to show the computation that produced the fair/unfair labels.

4 Conclusions

Responsible data science — incorporating legal norms and ethical considerations into data-driven algorithmic decision making — presents significant challenges and exciting opportunities for both basic and applied research. Importantly, lasting impact in this area cannot be achieved by technology alone, but must combine technological advances with social science methodologies, regulatory efforts, and education and engagement of the stakeholders. Responsible data science is our new frontier.

5 Acknowledgements

This work was supported in part by NSF Grants No. 1926250 and 1916647.

References

1. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias: Risk assessments in criminal sentencing. ProPublica (May 23, 2016), `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`

2. Asudeh, A., Jagadish, H.V., Miklau, G., Stoyanovich, J.: On obtaining stable rankings. PVLDB **12**(3), 237–250 (2018), http://www.vldb.org/pvldb/vol12/p237-asudeh.pdf
3. Asudeh, A., Jagadish, H.V., Stoyanovich, J., Das, G.: Designing fair ranking schemes. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2019 (2019)
4. Barocas, S., Selbst, A.D.: Big Data's Disparate Impact. SSRN eLibrary (2014)
5. Brauneis, R., Goodman, E.P.: Algorithmic transparency for the smart city. Yale Journal of Law & Technology (forthcoming)
6. Citron, D.K., Pasquale, F.A.: The scored society: Due process for automated predictions. Washington Law Review **89** (2014), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2376209
7. Crawford, K.: Artificial intelligences white guy problem. New York Times (June 25, 2016), https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html
8. Drosou, M., Jagadish, H., Pitoura, E., Stoyanovich, J.: Diversity in Big Data: A review. Big Data **5**(2) (2017)
9. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.S.: Fairness through awareness. In: Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012. pp. 214–226 (2012). https://doi.org/10.1145/2090236.2090255, http://doi.acm.org/10.1145/2090236.2090255
10. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015. pp. 259–268 (2015). https://doi.org/10.1145/2783258.2783311, http://doi.acm.org/10.1145/2783258.2783311
11. Hajian, S., Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in data mining. IEEE Trans. Knowl. Data Eng. **25**(7), 1445–1459 (2013). https://doi.org/10.1109/TKDE.2012.72, http://dx.doi.org/10.1109/TKDE.2012.72
12. Homelessness, P.E.: Vulnerability Index - Service Prioritization Decision Assistance Tool (VI-SPDAT). http://pehgc.org/wp-content/uploads/2016/09/VI-SPDAT-v2.01-Single-US-Fillable.pdf, [Online; accessed on 14-September-2017]
13. Kamiran, F., Zliobaite, I., Calders, T.: Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowl. Inf. Syst. **35**(3), 613–644 (2013). https://doi.org/10.1007/s10115-012-0584-8, http://dx.doi.org/10.1007/s10115-012-0584-8
14. Muñoz, C., Smith, M., Patil, D.: Big data: A report on algorithmic systems, opportunity, and civil rights. The White House (May 2016)
15. Network, M.: First, do no harm: Ethical guidelines for applying predictive tools within human services. http://www.alleghenycountyanalytics.us/ (2017), [forthcoming]
16. Ping, H., Stoyanovich, J., Howe, B.: Datasynthesizer: Privacy-preserving synthetic datasets. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017. pp. 42:1–42:5 (2017). https://doi.org/10.1145/3085504.3091117, http://doi.acm.org/10.1145/3085504.3091117

17. Podesta, J., Pritzker, P., Moniz, E.J., Holdern, J., Zients, J.: Big data: seizing opportunities, preserving values. Executive Office of the President, The White House (May 2014), https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf
18. Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. Knowledge Eng. Review **29**(5), 582–638 (2014). https://doi.org/10.1017/S0269888913000039, http://dx.doi.org/10.1017/S0269888913000039
19. Stoyanovich, J., Amer-Yahia, S., Milo, T.: Making interval-based clustering rank-aware. In: EDBT (2011)
20. Stoyanovich, J., Goodman, E.P.: Revealing algorithmic rankers. Freedom to Tinker (August 5, 2016), http://freedom-to-tinker.com/2016/08/05/revealing-algorithmic-rankers/
21. Stoyanovich, J., Howe, B., Abiteboul, S., Miklau, G., Sahuguet, A., Weikum, G.: Fides: Towards a platform for responsible data science. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017. pp. 26:1–26:6 (2017). https://doi.org/10.1145/3085504.3085530, http://doi.acm.org/10.1145/3085504.3085530
22. Stoyanovich, J., Yang, K., Jagadish, H.V.: Online set selection with fairness and diversity constraints. In: Proceedings of the 21th International Conference on Extending Database Technology, EDBT 2018, Vienna, Austria, March 26-29, 2018. pp. 241–252 (2018). https://doi.org/10.5441/002/edbt.2018.22, https://doi.org/10.5441/002/edbt.2018.22
23. The New York City Council: Int. No. 1696-A: A Local Law in relation to automated decision systems used by agencies (2017)
24. Yang, K., Gkatzelis, V., Stoyanovich, J.: Balanced ranking with diversity constraints. In: Proceedings of the Twenty-Eighths International Joint Conference on Artificial Intelligence, IJCAI (2019)
25. Yang, K., Stoyanovich, J.: Measuring fairness in ranked outputs. FATML **abs/1610.08559** (2016), http://arxiv.org/abs/1610.08559
26. Yang, K., Stoyanovich, J., Asudeh, A., Howe, B., Jagadish, H.V., Miklau, G.: A nutritional label for rankings. In: Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018. pp. 1773–1776 (2018). https://doi.org/10.1145/3183713.3193568, https://doi.org/10.1145/3183713.3193568
27. Yang, K., Stoyanovich, J., Asudeh, A., Howe, B., Jagadish, H., Miklau, G.: A nutritional label for rankings. In: ACM SIGMOD 2018 (2018)
28. Zemel, R.S., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: ICML. pp. 325–333 (2013), http://jmlr.org/proceedings/papers/v28/zemel13.html