NEW YORK UNIVERSITY

Julia Stoyanovich
Assistant Professor of Computer Science and Engineering, Tandon School of Engineering

Assistant Professor of Data Science, Center for Data Science

Department Computer Science and Engineering
Tandon School of Engineering
2 MetroTech Center
Brooklyn, NY 11201
P: 212 998 3251
M: 917 470 8199
stoyanovich@nyu.edu

Testimony of Julia Stoyanovich before New York City Council Committee on Technology and the Commission on Public Information and Communication (COPIC)

February 12, 2019

Dear Speaker Johnson, Chair Koo, and members of the Committee and Commission:

The data revolution is transforming every sector of science and industry, but has been slow to reach local and municipal governments that deliver vital human services in health, housing, and mobility. The opportunities of data-driven algorithmic decision making in urban contexts have long been recognized, evidenced by the remarkable progress around open data, the digitization of government records and processes; and, perhaps most visibly, smart city efforts that emphasize using sensors to optimize city processes. Despite this progress, the public sector is slow to adopt data-driven technology, for two related reasons, both highly relevant to the topic of today's hearing.

The first reason concerns the legal and technical difficulties inherent in the sharing of sensitive data, both among government agencies and with external entities. The second reason is the government's mandate for responsibility — meaning that any decision made by algorithms will need to be scrutinized by the affected individuals, groups, and the general public. *In my testimony today, I will argue that both barriers to adoption of data-driven technology can be overcome by establishing a robust and flexible data-sharing infrastructure.* Consequently, establishing this infrastructure should be seen as a clear strategic and operational priority for New York City.

My name is Julia Stoyanovich, I am a resident of New York City (District 7). I hold a Ph.D. in Computer Science from Columbia University. I am an Assistant Professor of Computer Science and Engineering at New York University's Tandon School of Engineering, and an Assistant Professor of Data Science at the Center for Data Science. In my research and teaching, I focus on *responsible data science* — on incorporating legal requirements and ethical norms, including fairness, accountability, transparency, and data protection, into data-driven algorithmic decision making.[1] Some of the students enrolled in my Responsible Data Science course are here today.[2]

I am an appointed member of a Task Force established in response to Local Law 49 of 2018, in relation to automated decision systems used by agencies (the ADS Task Force). Opinions in this testimony, while informed to some extent by my work on the ADS Task Force, are my own, and do not represent the views of the Task Force.

I am thrilled that New York City is maintaining its leadership role in responsible data-driven governance. We are the first, and only, US city to pass an Open Data Law (Local Law 11 of 2012) and an Automated Decision Systems Law (Local Law 49 of 2018). Further, Local Laws 245 and 247 of 2017 establish the role of the Chief Privacy Officer, in support of responsible citywide data sharing

[1] See https://dataresponsibly.github.io/ for information about this work, funded by the National Science Foundation through the BIGDATA program (NSF Award #1741047).
[2] DS-GA 3001.009 Responsible Data Science, all course materials are publicly available at https://dataresponsibly.github.io/courses/spring19/

practices. Committee on Technology and the Commission on Public Information and Communication have an imperative to act jointly, to continue creating an environment in which legislative efforts and technological innovation act in concert, with the goal of *"improving government transparency, improving the public's access to government information, protecting personal information privacy, and facilitating data sharing between city agencies"* — the topic of today's hearing.

In my statement today, I would like to make three points:

1. Establishing a robust and flexible data-sharing infrastructure should benefit multiple stakeholders.
2. There is a continuum of data sharing modalities between open data and a secure data clean room that need to be explored as part of infrastructure design.
3. Developing a data-sharing infrastructure will require technological innovation, buy-in from city stakeholders, and public engagement.

I now briefly discuss each of these points in turn, and conclude with a set of recommendations. My testimony will be complemented by statements from my distinguished colleagues Julia Lane, Professor at the Wagner Graduate School of Public Service at New York University, and Stefaan Verhulst, Co-Founder and Chief or Research and Development at GovLab, an action research center at New York University.

My first point relates to the importance of establishing a data-sharing infrastructure that benefits multiple stakeholders. *Government agencies* need to share data to make decisions more effectively, and to enact policy in coordination. *Regulators* need access to agency data for purposes of oversight. In both cases, much of the data is sensitive and so is legally encumbered: This data either contains personally identifiable information, or is anonymized but still does not guarantee privacy when linked with other data.[3] Equally as importantly, but discussed less often, is *the public's* need to access data in support of algorithmic transparency.

Recent reports on data-driven decision-making underscore that fairness and equitable treatment of individuals and groups is difficult to achieve[4], and that transparency and accountability of these processes in government are indispensable but rarely enacted[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.

New York City's ADS Transparency Law (Local Law 49 of 2018) initiates a meaningful response to these threats, and other US municipalities are likely to follow with similar legal frameworks or recommendations. Of utmost importance as this happens is recognizing the central role of data transparency in any algorithmic transparency framework. *Meaningful transparency of algorithmic processes cannot be achieved without transparency of data!*[6] Data transparency in turn cannot be achieved without a robust and flexible data-sharing infrastructure.

---

[3] See https://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/ for a descritpion of a now-classic 1997 re-identification attack, in which Latanya Sweeney, a graduate student at the time, re-identified Massachusetts Governor Weld by linking anonymized hospital visit records and public voter rolls.

[4] MetroLab Network, "First, Do No Harm: Ethical Guidelines for Applying Predictive Tools within Human Services," 2017, https://metrolabnetwork.org/data-science-and-human-services-lab/

[5] Robert Brauneis and Ellen P. Goodman, "Algorithmic Transparency for the Smart City," *Yale Journal of Law & Technology* 20, no. 103 (2018), http://dx.doi.org/10.2139/ssrn.3012499

[6] Julia Stoyanovich and Bill Howe, "Follow the Data! Algorithmic Transparency Starts with Data Transparency," *The Ethical Machine*, November 27, 2018, https://ai.shorensteincenter.org/ideas/2018/11/26/follow-the-data-algorithmic-transparency-starts-with-data-transparency

What is data transparency? In applications involving predictive analytics, data is used to customize generic algorithms for specific situations—that is to say that 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. Decision-making applications that do not use machine learning technology, such as explicitly stated decision procedures like the Public Safety Assessment and Decision Making Framework (PSA), or that do not attempt to predict future behavior based on past behavior, such as matchmaking methods used by the Department of Education to assign children to spots in public schools, are still heavily influenced by the properties of the underlying data — they are designed and validated using data. We cannot understand whether these methods work, and what impacts they have on individuals and population groups, if we don't have access to the training and validation datasets. An important role of a data-sharing infrastructure is to support access to these training and validation datasets for the purpose of inter-agency coordination, auditing and oversight, and transparency and accountability to the public.

My second point is that there is continuum of data sharing modalities between open data and secure data sharing environments like clean rooms (also known as secure data enclaves). My colleague Julia Lane will discuss her extensive expertise in developing solutions of this kind for access to administrative data.

Let me continue with my argument about the need for data transparency in support of algorithmic transparency, and observe that, while we require access to the training and validation datasets of a particular automated decision system, these datasets may well be sensitive and so cannot be easily shared or released to the public. That is, data transparency is in tension with the privacy of individuals who are included in the dataset. In light of this, a data-sharing infrastructure can offer an alternative data sharing modality. When raw datasets cannot be exchanged or released, relevant statistical properties of the datasets can be exposed through statistically similar synthetic datasets or data summaries. These can in turn be generated using state-of-the-art methods to preserve the privacy of individuals included in the data.[7]

In addition to privacy-preserving data sharing techniques, appropriate for environments in which a trusted relationship between stakeholders cannot be established, it is possible to develop access control and usage control mechanisms for trusted environments. A carefully designed data-sharing infrastructure can be made to support multiple such modalities.

My third and final point is brief. When developing a data-sharing infrastructure, we must consider the legal, societal, and technical aspects of the challenge. A solution will entail engaging technology experts, building competencies and incentives within the City, and developing governance structures. My colleagues Julia Lane and Stefaan Verhulst will discuss these aspects in their statements.

To conclude, I recommend that the City consider the development of a data-sharing infrastructure as a strategic and operational priority, with the goals of (1) increasing efficiency of delivery of human services, and (2) supporting transparency and accountability to the public, thus increasing the public's trust in government. Developing this infrastructure will require significant investment, which should be amortized so as to benefit multiple City and external stakeholders. Different data sharing scenarios will require different sharing modalities, including open data, privacy-preserving synthetic data and summaries, access and usage control mechanisms, and secure data clean rooms.

---

[7] Haoyue Ping, Julia Stoyanovich, and Bill Howe, "DataSynthesizer: Privacy-Preserving Synthetic Datasets," in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, Chicago, Illinois, June 27–29, 2017, 42:1–42:5. https://dataresponsibly.github.io/tools/