Harvard Data Science Review • Issue 2.3, Summer 2020

Data Science and Cities: A Critical Approach

Fábio Duarte¹ Priyanka deSouza¹

¹Department of Urban Studies and Planning, School of Architecture, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

Published on: Jul 30, 2020

DOI: <https://doi.org/10.1162/99608f92.b3fc5cc8>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

Sensors increasingly permeate our lives and generate a plethora of data, which has transformed the way we live in cities. Planners have been using data-science to improve our understanding of urban issues. While other domains have highlighted concerns with big data collection, aggregation, and analytical methods to understand different phenomena, urban planning has an additional aspiration: not only to understand, but to *transform* society through planning. Thus, on top of critically approaching data collection and analytical methods, for the emergent field of urban science to become a distinctively unique body of knowledge, it must examine the ontological and epistemological boundaries of the big data paradigm and how it affects urban decision-making processes and their short- and long-term consequences in cities.

Keywords: urban science, urban informatics, sensors, data analytics, data politics

Data-driven approaches have transformed the way we analyze, design and make policy decisions in cities. This has been true during the COVID-19 pandemic, where countries have used self-reported information and tracing apps to map infected people. [South Korea Corona Map](#), for example provides the addresses of all infected residents, and [Singapore COVID19](#) maps each case and their social networks, to help other people identify if they had contact with an infected person, took the same flight or used the same urban facilities to be aware of their risk of contagion.

There are many examples of data-driven approaches in other aspects of city management. Urban greenery brings several benefits to city dwellers, including countering heat-island effects, improving air quality, and decreasing stress levels. However, mapping street trees is labor-intensive, and there are not cheaply available technologies that allow a comparative analysis among cities across the world. Research has been conducted using Google Street View images and machine learning techniques to quantify green canopy in cities at the pixel level (Cai et al., 2018; Li et al., 2015 et al.). This algorithm is open-source, and has allowed cities around the world to use this technique to plan how to increase greenery in their environs.

Another example of data-drive urbanism is an effort to address health problems related to air pollution. Cities such as New York have been implementing a network of stationary air quality monitors. However, when it comes to measuring a resident's exposure to pollutants, research usually relies on people's home location. Using cell phone data as a proxy to people's movements across the boroughs in New York, researchers identified residents' exposure to pollutants with finer spatial and temporal resolution, considering not only their home address but also their commute and work or school location (Nyhan et al., 2016).

Finally, in regions undergoing rapid urbanization, such as in China, it takes a long time to accurately quantify urban growth. Researchers found that restaurant data (including the number of seats, type of cuisine and consumer's rates) available on open platforms such as Dianping, the Chinese equivalent to Yelp, is a strong

predictor of population growth in Chinese cities at the neighborhood level (Dong et al., 2019). In countries that don't undertake a regular census of their populations, or where public data is unreliable, such an approach using crowd-sourced open data can help companies and public officials map areas that are transforming rapidly that will be in need of public infrastructure and services.

All these initiatives aim to improve the understanding of urban issues in ways that were not possible with previous methods and tools; Thus, the emergence of a field that has been called *urban science*. Recognizing the need to professionalize urban science, data science initiatives have been introduced in academic programs, private practices and public agencies that focus on urban issues. Such initiatives mainly involve familiarizing scholars and practitioners with new methods and tools to gather and analyze such data abundance, such as machine learning, data mining, and data visualization (French et al., 2017). Planning schools have also initiated training programs centered on the use of advanced data analytic methods to understand urban issues, thereby signaling a significant transformation in planning practice. Examples include New York University's Center for Urban Science and Progress, the University of Michigan's graduate certificate in Urban Informatics, and the Massachusetts Institute of Technology's major in Urban Science.

We argue that teaching planners and designers computer science concepts and tools, alone will not transform them into urban scientists. For urban science to become a distinctively unique body of knowledge, it must go beyond professionalizing urban science. Urban scientists must be acutely aware of the ways in which their science is used in different policy landscapes, and of the possible unintended environmental and social consequences of their work. This involves urban scientists engaging with the intrinsically political dimension of urban science, and the ways in which their data and predictive models produce results that embody and act on specific social relations.

Stephen M. Stigler (2019) recently argued in this journal that any data has a life span, and the way it is collected and the tools we employ to make sense of it are charged with ideological values, and reflect partial understandings of lived reality. In a nutshell, paraphrasing a classic paper in science, technology and society (Winner, 1980), data *do* have politics. Urban scientists need to constantly be asking ourselves who benefits from the new informational landscapes, and which populations slip through the cracks. For example, urban scientists tend to work in cities where data is easily accessible, which tend to be in Western countries. Geographies such as the global South are often left out of analyses. There is what David J. Hand (2020) calls 'dark data,' emerging from phenomena we are not prepared to observe directly, or data we cannot collect with current tools and do not fit within existing methods, but still can have major effects in our decisions and actions. Urban scientists *must* perform the "hard work of theory" to critically examine the ontological and epistemological boundaries of the big data paradigm (Pickles, 1997).

Furthermore, many datasets that urban scientists use are collected and aggregated by corporations to further their own profit-drive motives. For example, a large number of papers using image recognition and neural networks use street views available online, in particular Google Street View. However, such datasets are owned

by corporations, which can restrict access to data at any point—as Google did recently, charging for the use of these images even for research purposes. Moreover, as Google Street View only provides data on the visual physical features of cities, in places where social and racial segregation are frequently tied to ZIP codes, results from analyses that only rely on such data can reinforce stereotypes and segregation—in what Sarah Brayne (2017) calls another “quantified modality of social control.” Thus, urban researchers should constantly push for cities to develop open datasets for the public good, which are beneficial to everybody.

Urban scientists must also be cautious about the methods and models they use. For example, a general truism within the data science community was that machine learning algorithms are analogous to a black box: we cannot not precisely understand the models crunching the data, but it is worth sacrificing interpretability for accuracy. Cynthia Rudin and Joanna Radin (2019) discussed in this journal that such tradeoff is a fallacy, which has even been beneficial to companies marketing proprietary black box models. In urban studies, black-boxed findings risk to drive planning, policy, and design decisions that have the potential to reinforce detrimental *status quo* and further pre-existing bias. Cathy O’Neil (2016) discusses the lack of accountability in some predictive models used by police, with the unevenness treatment of social groups stemming from the input data (stop-and-frisk policing in New York inherently eschew what is collected), predictive policing models which focus on crimes that are usually tied with certain population groups, and evidence-based sentencing grounded on attributes more common on certain specific groups.

In parallel to getting hands on data and developing models, urban scientists need to address the limits and the unintended consequences of data-driven approaches, or the “unknown-unknowns” (Lakkaraju et al., 2017); when predictive models assign with high confidence incorrect labels to instances that often stem from incomplete models or datasets, but which raise ethical concerns about algorithmic intrinsic bias. In the politically-charged field of facial recognition, for example, Buolamwini and Gebru (2018) have shown that the training datasets for the most widely-used facial recognition algorithms systematically under-represent black people, and specifically black women. The training dataset is thus a poor reflection of the real world. In response, the IEEE [P7013 Inclusion and Application Standards for Automated Facial Analysis Technology](#) working group is developing standards that limit the scope of use of facial recognition software and are determining metrics for the success of algorithms. Urban scientists need to actively engage, and be a part of such initiatives.

Cities are socio-technical assemblages. Not all social aspects can be translated to discrete and numerical data convenient for use in current data-science methods. Often the methods and metrics we employ shape the phenomenon we observe. We need to be careful not to fall prey to the ‘tyranny of metrics’ (Muller, 2019), a reductionist, abstract view of reality—sometimes critical phenomena of our times do not produce data readily read by computers, and their social impacts would slip through the cracks of metrics-centered approaches. Urban scientists thus need to collaborate with social scientists, communities, artists to develop tools and models that benefit the people they serve.

To summarize: although we acknowledge at the outset that urban science has huge potential to improve cities, the socio-political context of the implementation of these technologies cannot be forgotten, for these technologies to improve cities. Urban science needs to go beyond the dexterity in using new methods to analyze the abundance of data in cities, and be prepared to interrogate every aspect of their work, from the dataset itself, and the methods they use to understand, predict, and inform emergent urban phenomena.

Disclosure Statement

Fábio Duarte and Priyanka deSouza have no financial or non-financial disclosures to share for this article.

References

- Brayne, S. (2017). Big data surveillance: The case of policing. *American Sociological Review*, 82(5), 977–1008. <https://doi.org/10.1177/0003122417725865>
- Buolamwini, J. and Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research: Vol. 81. Conference on fairness, accountability and transparency* (pp. 77–91). <https://proceedings.mlr.press/v81/buolamwini18a.html>
- Cai, B. Y., Li, X., Seiferling, I., & Ratti, C. (2018). Treepedia 2.0: Applying deep learning for large-scale quantification of urban tree cover. In *2018 IEEE International Congress on Big Data* (pp. 49–56). <https://doi.org/10.1109/bigdatacongress.2018.00014>
- Dong, L., Ratti, C., & Zheng, S. (2019). Predicting neighborhoods’ socioeconomic attributes using restaurant data. *Proceedings of the National Academy of Sciences*, 116(31), 15447–15452. <https://doi.org/10.1073/pnas.1903064116>
- French S. P., Barchers C., & Zhang W. (2017). How should urban planners be trained to handle big data? In P. Thakuriah, N. Tilahun, & M. Zellner (Eds.), *Seeing Cities Through Big Data* (pp. 209–217). Springer Geography. https://doi.org/10.1007/978-3-319-40902-3_12
- Hand, D. J. (2020). *Dark data: Why what you don’t know matters*. Princeton University Press.
- Li, X., Zhang, C., Li, W., Ricard, R., Meng, Q., & Zhang, W. (2015). Assessing street-level urban greenery using Google Street View and a modified green view index. *Urban Forestry & Urban Greening*, 14(3), 675–685. <https://doi.org/10.1016/j.ufug.2015.06.006>
- Lakkaraju, H., Kamar, E., Caruana, R. & Horvitz, E. (2017). Identifying unknown unknowns in the open world: representations and policies for guided exploration. In *Proceedings of the 31st Association for the*

Advancement of Artificial Intelligence Conference on Artificial Intelligence.

<https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/download/14434/14383>

Muller, J. Z. (2019). *The tyranny of metrics*. Princeton University Press.

Nyhan, M., Grauwin, S., Britter, R., Misstear, B., McNabola, A., Laden, F., Barrett, S. R., & Ratti, C. (2016). “Exposure Track”—The impact of mobile-device-based mobility patterns on quantifying population exposure to air pollution. *Environmental science & technology*, 50(17), 9671–9681.

<http://doi.org/10.1021/acs.est.6b02385>

Pickles, J. (n.d.). Tool or science? GIS, technoscience, and the theoretical turn. *Annals of the Association of American Geographers*, 87(2), 363–372. <https://doi.org/10.1111/0004-5608.00058>

Rudin, C., & Radin, J. (2019). Why are we using black box models in AI when we don’t need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, 1(2).

<https://doi.org/10.1162/99608f92.5a8a3a3d>

So, W. (2020). MIT senseable city lab [Photograph].

Stigler, S. M. (2019). Data have a limited shelf life. *Harvard Data Science Review*, 1(2).

<https://doi.org/10.1162/99608f92.f9a1e510>

Winner, L. (1980). Do artifacts have politics? *Daedalus*, 109(1), 121–136

©2020 Fábio Duarte and Priyanka deSouza. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.