# CITY AS DATA

by Fábio Duarte

I'm writing this essay on a sunny morning at a cafe facing the Charles River full of sailing boats. Around me people are drinking coffee and pastries, some are working on their laptops, others are on their smartphones, and even talking with each other. Cars are rolling along Memorial Drive, bikers are riding on BlueBikes, the local bike sharing program, while people are jogging along the river, in pairs, with their dogs, pushing strollers. People are sitting outside, some enjoying the sun, others under the shade of the many trees on the patio in front of the glassed façade of the MIT School of Management, where too many lights are on although the sun is up in a perfect clear sky. Squirrels are jumping around, birds are chirping, and, if you get closer to the trees and bushes, you see insects flying and crawling. A joyous summer Friday.

This is a simple and quite uninspiring description of an urban scene, I know that. **Until we realize that every single element described above produces or is embedded with data, or can reveal some of their characteristics only when analyzed with cutting edge computational tools.**

Every coffee sold is tied to a credit card, which adds up with other items the person has purchased over time, creating a consumer profile. Since each machine has its specific location and each credit card is unique, in aggregate we have a rich and dynamic understanding of how people move and consume in the city. On the grass in front of the cafe I see a few beetles. I notice they are diverse, but that is about it. I snapped a photo with my phone and found one is an Asian multicolored lady beetle, another is a banded ash borer. Don't ask me further. I just know it would

be impossible for me to go beyond "insect" were it not for computer vision models that help me discover the entomological world that shares the city with us.

Since its inception, twenty years ago, **the MIT Senseable City Lab has the mission of understanding urban environments through the abundance of data we have around us, and developing new tools to collect data and analyse cities in novel and often unexpected ways**. Let me pick up some features of my initial description of my summer morning and see them from a data perspective in related projects.

In 2014, in _**Drinking Data**_, we collected data from more than 15,000 Coca-Cola Freestyle dispensers in the United States. Each time a person filled their cup, we recorded the time, location, and chosen beverage. With this huge dataset, spanning over a year, we could see expected spikes during weekends and dips during weekdays, and special celebrations such as Easter and 4th of July, the national holiday. But there were unexpected results, such as finding some consumption bumps on September 7th, for example. This date doesn't mean anything in the United States—but it is the Brazilian independence day. By seeing where Coca-Cola consumption spiked on September 7th we could infer where Brazilian communities exist. Or some small cities where consumption increased late at night in December and May—college towns during finals. Also in the consumption sphere, more recently MIT researchers (Fleder and Shah, 2021), using a large dataset of anonymized consumer credit card transactions, a sample

of itemized receipts, and employing machine learning techniques, were able to identify which products were purchased in each given transaction observing only the total amount of the bill.

As data is more easily available and analytics tools are more powerful, we can go from which drink or item a person is purchasing to the nutritional value of each of them—and connect it with socioeconomic characteristics and health outcomes. This is what we did in *Traveling Tastes*. We scrapped menus from thousands of restaurants in Dubai, London, and Boston. With the list of each dish sold in these restaurants, we ran it through large language models to describe the items of these dishes, and calculate their balance index of each meal, a measure of its nutritional quality. With this information, we explored the **correlations between food landscapes, socioeconomic profiles, and health outcomes**, such as neighborhoods at risk of obesity and cardio-metabolic diseases.

What about all joggers along the Charles? If they are carrying their smartphones, they are a rich source of information. If they have health apps installed, even more—and some smartphones capture basic health data, such as distance walked and calories burned, by default. By analyzing such data, we can understand the factors that **influence outdoor human activity**—such as weather, urban morphology, topography, traffic, or the presence of green areas. This is what we did in *City Ways*, based on billions of data points collected in Boston and San Francisco via activity monitoring apps. We found, for example, that during the winter months, a mere degree increase in temperature has a much larger effect in weekend trips in Boston—a 13.5% versus 2.5% increase in San Francisco, whereas the effect of a moderate rainfall of 5.0 mm per hour is associated with a 29.0% decrease in weekend trip counts San Francisco, whereas, for Bostonian—well, it's winter, after all (Vanky et al., 2017). We also found that 10% of pedestrians (not joggers) in Boston don't take the shortest path from their origin and destination. Intrigued by that, in *Pointiest Path* we decide to investigate whether there were any underlying rationale for such behavior. And we found a mathematical explanation for that: pedestrians who deviate from the shortest path chose routes that were slightly longer but more directly point towards the destination—that is, paths that allow them to more directly face their endpoint as they start the route, even if, in the end, the journey is longer (Bongiorno et al., 2021).

Quantifying how much an increase of one degree in temperature influences a jogger's willingness to go outside, or the mathematics of pedestrian behavior, **was only possible because of the massive dataset of GPS traces collected by the smartphones** carried by people. Lady bugs don't use them. Still, they provide important information about the health of urban biodiversity. Here recent computational tools help us. There are several apps which can tell us which species of bird or tree or insect are there when you take a picture with your smartphone. In *B++* we decide to do it at scale. The scale of biodiversity loss is impressive: almost 70% decrease in biodiversity since 1970. Insects are a good proxy of biodiversity, which makes the rapid decline in insect populations particularly concerning, as around 65% of insect species could go extinct over the next one hundred years. We need to act fast. We have been developing a computer vision model to detect insect species in real time, with initial results jumping the

*Output map from the Pointiest Path project*

established models from 40 to 2,400 species. We are now expanding the model to more species and deploying it in different biodiversity contexts.

The initial description of what I see from the cafe also includes some subjective perception: it's a joyous summer Friday. River, trees, grass, small animals. But would anybody perceive it similarly? It seems like a silly question. And it might get even more complicated when it is charged with concepts such as biophilia, which claims that the contact with nature brings positive emotional effects. Although it makes sense, we were intrigued by that: nature varies widely across the world; so, would people in arid deserts, snowy mountains, or lush forests perceive the same

natural elements equally?; or would there be specific natural features that would trigger the biophilic benefits of people living in each of these environments?

In ***Feeling Nature*** we decided to test whether we value nature the same way across different terrestrial biomes. We combined two techniques: a visual artificial intelligence model and surveys with residents of eight cities in different biomes. We first employ an image classification model to quantify all natural features (water, rocks, grass, etc.) in street view images in Amsterdam, Barcelona, Buenos Aires, Dubai, Nairobi, Quebec, Singapore, and Trondheim. As expected, the

amount of sand and greenery differs greatly in Dubai and Singapore, for example. Concomitantly, we ran a survey with residents of each of these cities in which we showed them pairs of images of these cities and asked them to simply point which brought them more positive feelings. We used the results of the survey to retrain our visual AI model, we reran all street view images through it. **Now, the model not simply counted the amount of natural elements present in these images, but weighted them based on residents' perceptions**. For instance, we found that people living in arid or rocky biomes, such as Nairobi (savanna), Dubai (desert) and Trondheim (tundra), tend to be more attracted to land-related classes; Amsterdammers by waterscapes; and those in Nairobi showed a stronger affiliation towards people and wildlife. This is the first time the concept of biophilia, although strong and common sensical, has been quantified globally—and this has only been possible because of the abundance of visual data available globally and powerful computational techniques.

Even though we are excited with the immense possibilities of understanding cities in novel ways based on data and computational tools, we must be watchful of the risks. As we discussed elsewhere (2023), when data collected for one purpose is used in unexpected ways, we have novel insights about the city (as the Drinking Data above), but we might also see it in dangerous ways, and find that it might be too late to secure our right to anonymity, privacy, and prevent misuses. **It is definitely our role, as scholars, to be vigilant about these risks and ring the bell when any misuse happens**. But it is equally our role to explore new ways of collecting and analyzing data with the goals of improving our understanding of cities.

## References

Bongiorno, C., Zhou, Y., Kryven, M., Theurel, D., Rizzo, A., Santi, P., Tenenbaum, J., Ratti, C. (2021). Vector-based Pedestrian Navigation in Cities. Nature Computational Science 10(1): 1-8.

Duarte, F., Fröding, B. (2023). Watch out! Cities as data engines. AI & Society, 38, 1249–1250. https://doi.org/10.1007/s00146-022-01448-5.

Fleder, M., Shah, D. (2020). I know what you bought at Chipotle for $9.81 by solving a linear inverse problem. Proceedings of the ACM on Measurement and Analysis of Computing Systems, 4:3, 1 - 17. https://doi.org/10.1145/3428332.

Lefosse, D., Duarte, F., Sanatani, R., Kang, Y., van Timmeren, A., Ratti, C. (2025). Feeling Nature: Measuring perceptions of biophilia across global biomes using visual AI. npj Urban Sustainability, 5(4), 1-13. https://doi.org/10.1038/s42949-025-00192-1.

Tufano, M., Duarte, F., Mazzarello, M., Eshtiyagh, J., Ratti, C. and Camps, G. (2025). Data-Driven Nutritional Assessment of Urban Food Landscapes: Insights from Boston, London, and Dubai. Scientific Reports 15, 1-11. https://doi.org/10.1038/s41598-025-08098-9.

Vanky, A., Verma, S., Courtney, T., Santi, P., Ratti, C. (2017). Effect of weather on pedestrian trip count and duration: City-scale evaluations using mobile phone application data. Preventive Medicine Reports, 8, 30-37.

Venverloo, T., Duarte, F. (2024) Towards real-time monitoring of insect species populations. Scientific Reports 14, 1-9. https://doi.org/10.1038/s41598-024-68502-8.

# About Fábio Duarte



Fábio Duarte is the Associate Director, Research & Design, at the MIT Senseable City Lab, where he is also a Principal Research Scientist. Duarte is a Lecturer in DUSP and is affiliated with the Sustainable Urbanization Lab, in the Center for Real Estate.

**Webpage:**

https://dusp.mit.edu/people/fabio-duarte