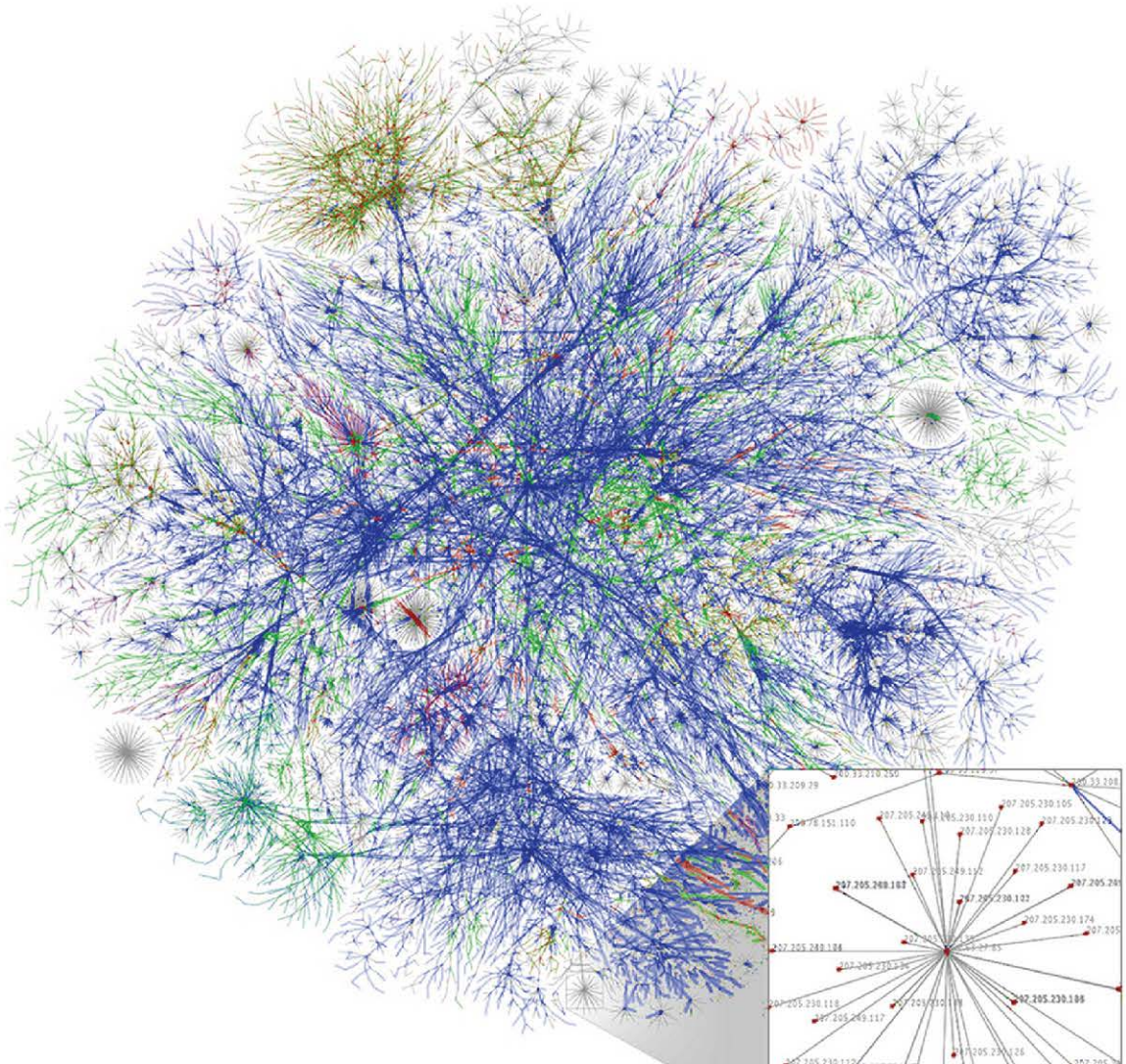
BIG DATA, FALSE DATA, SMART DATA, DUMB DATA

By Nicole Lambrou

Measuring the human condition.

Published November 18, 2014

Big data is immediate. It is laden with the technological and informational ubiquity of the digital age. As much as 90 percent of the planet's data was collected between 2010 and 2012. This is hardly surprising, given the fact that our human genome can be stored in less than 1.5 gigabytes. Yet big data is useless as raw material. It is valuable only after a program, or a search engine, is written to pull out information necessary for the task at hand. The researcher, acting as a human conduit, is introduced into the equation after information has been harvested. He or she is responsible for processing and distilling this crude data into meaningful information. This is the semiotic process referred to as data analytics.



Example of Data Analytics Visualization. "Internet map 1024 – transparent" by *The Opte Project*. Licensed under Creative Commons Attribution 2.5 via Wikimedia Commons.

The immediacy of data overshadows a real set of issues that have been pushed aside due to an unbridled collective enthusiasm over the vastness and supposed objectivity of data. Big data already directly influences urban policies, from the distribution of public funds to infrastructural updates and zoning changes. But it also carries with it inherent flaws relating to its own feedback effects. In most instances of algorithmic city planning, outputs rather than outcomes are assessed.[1] The widely-held belief that an algorithm can directly reflect reality to predict growth, pattern, and infrastructure is problematic. Algorithms are written to serve the eye of their author: they look for what they want to see.

FROM DATA TO MODEL

The epistemological and real-world applications of big data are defined within the limitations of processing such quantitatively large and qualitatively complex systems of data. However "big" big data may be, it will always be incomplete and inconclusive. We can only measure that which can be measured, excluding a significant amount of information outside the parameters of zeros and ones.

In statistical terms, if data is consistent and does not shift based on trends, it is considered stationary. It is tempting, then, to create a digital model based on this data, since its predictive qualities are grounded in consistency. Conversely, non-stationarity, also referred to as concept drift, is defined as a changing target variable.

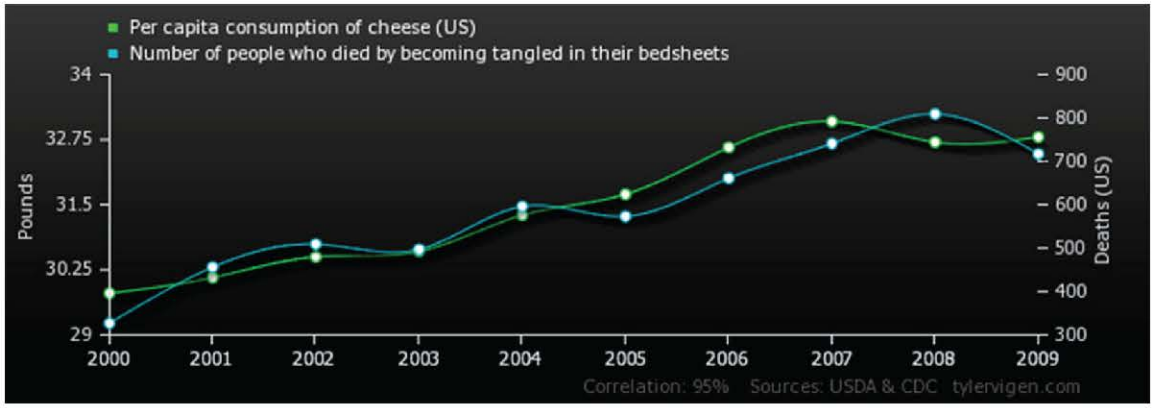
In law and policy-making we typically allow room for interpretation. For example, loitering is defined as lingering in a place for a period of time, the result of which one reasonably expects is criminal activity. That the definition of loitering does not include a specific amount of time—it is not measured in minutes—is necessary. It guarantees that the law itself does not become the target of human behavior.

The translation of a loitering law into measurable information, a series of zeros and ones, has certain quantitative requirements. Take for example, Barcelona's initiative to construct smart lampposts to "predict" suspicious behavior.[2] The digital model that serves this prediction must be fed data in seconds, minutes or hours in order to define suspicious activity. However, what is reasonable or unreasonable cannot be recorded as a series of numerical digits. The term "reasonable," which embeds interpretation into the system, is a necessary part of a system that both respects and accounts for openness, and is guided by checks and balances.

To avoid the implementation of data-driven decisions as surveillance tools, conceptual drift (or the ever-changing target variable of shifting human behavior) needs to be embraced. The model must allow data to be searched and assessed, and for successful variances (in the form of innovation or healthy deviance) to be identified and re-integrated into a new model. This new model can then recycle through deployment, testing, and assessment. It can be re-written to absorb innovation and deviation perpetually in real-time, while maintaining the rigor of validity and reliability.

Even the graphic form of this data shapes our vision and comprehension of its overall pragmatic application. Searchable real-time data may direct us towards the nearest empty parking spot, but when it is mined to serve automated surveillance, the leap from information to prediction is spurious–it is built upon false premises and false logics.

In the wake of the Boston marathon bombing, a Long Island family was erroneously investigated based on searches made from a family member's work computer for "the terms 'pressure cooker bombs' and 'backpacks.'"[3] In this case, the search of a singular data relationship was the target. Deceptive alliances increase with the availability of more data, forming "spurious correlations: associations that are statistically robust but happen only by chance."[4]



Per capita consumption of cheese (US) correlates with number of people who died by becoming tangled in their bedsheets, from tylervigen.com

In the example above, a false correlation between per capita consumption of cheese, and the number of people who have died by becoming entangled in their bed sheets is displayed to reveal how while the correlation rate is ostensibly and statistically high, it is truly aleatoric.

The consequences of changing an algorithm can be enormous. At a conference on big data in 2012, mathematician Patrick McSharry reflected, "[b]asically they just changed the science, and billions of dollars had to be moved around, not because the world changed in any way, but because the model changed."[5] This calls for an independent and objective body to evaluate computational models, especially in relation to public policy and planning. Analysts begin with chaotic data, use various software and algorithms to visualize that data, and then identify the predictive properties of that information. In the movement from data to model, McSharry insists that the perfect model does not exist. Instead, imperfections in the computational model will persist, which means assessing the limitations of those models intelligently. An "inspector general" of data modeling is needed to contain and curb the inherent chaos and game theory-esque attributes of big data.

While big data allows us to get closer to a "goal," however different our visions of that goal may be, the unbridled enthusiasm for big data should be filtered through discussions of privacy, processing, and predictions. The patterns that search engines cull from information are as much a product of our behaviors, our cultures, our language and our bodies as the data itself is.

THE WRITERS

Who writes these algorithms? Mathematicians, statisticians, computer scientists and software engineers, together or separately, seem to have the most facility in mining data. Who implements that data into action? Planners, public policy makers, politicians, private institutions, government. Yet the question remains: is there enough transparency from one side of the equation to the other, whereby the policy maker can not only see, but also understand the other side, and vice versa? Given a perpetually changing future, this is a slippery affair. How, after all, can we measure innovation in a model whose success is based on its predictive mechanism? Innovation becomes a deviation.

Predictions for future implementation of big data modeling systems are already being devised. Charlie Catlett, the director of Urban Center for Computation Data (UCCD), claims that architects using analog methods meet their limits of analysis at the scale of a 20-acre urban plan. Beyond this scale, as in the Lakeside Development project with 600 acres and over 500 buildings, Catlett argues that we must rely on computational models to better assess zoning and phasing effects.[6]

By Argonne National Laboratory via Creative Commons, link.

LakeSim, a project by UCCD, uses Lakeside Development as a model and simulates the effects of certain planning decisions on the city. Two dials at the bottom of the screen, one responsible for "zoning" and the other for "phasing," allow the user to assess the effects of these two categories. By changing a selected street from commercial to residential, for example, the effects on energy are immediately represented in a colorful graph on the same screen. The urban planner is assigned the role of data visualizer.

In an attempt to get to the socioeconomic heart of a neighborhood, the Data Science for Social Good project (also led by UCCD and organized by the University of Chicago and Argonne National Laboratory in collaboration with Cook County Land Bank) searches for abandoned properties that have been identified as investment opportunities.

By Argonne National Laboratory / U.S. Department of Energy. Licensed under Public domain via Wikimedia Commons, link.

Urban planning decisions are interdisciplinary, and must extend to funding organizations and ancillary institutions that have a hand in such planning modalities. Endeavors that link properties with institutions such as banks, which in turn facilitate property development, is a prime example of data ownership that needs to be questioned given the social and urban implications embedded in the algorithm.

The danger here can be extrapolated to policy decisions led by data. It is invaluable for a city to measure or assess the success of a specific policy, and to use that data as evidence to garner support among their tax base. However, as Campbell's law succinctly states, "the more any quantitative social indicator (or even some qualitative indicator) is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."[7] Once we know the rules, in other words, we can play the game to our advantage.

Data owners, coupled with those who own search engine and data mining capabilities, set into motion an economic agenda that has far-reaching repercussions. Aside from the question of whether property development is even a proper response to urban blight, particularly in neighborhoods where low property values provide home ownership solutions, one has to ask: who owns the data? Depending on how the algorithm is conceived, and how inherent biases are written into each model, investments approved by the bank as low risk and high return options may offer a singular, rather than pluralistic, response.

The city of Songdo, South Korea, is scheduled for completion in 2016 and meant to house more than 350,000 people. It is to be infiltrated with sensors. Promoted as a "startup city," people can expect to have nearly every gadget, structure and road equipped with mechanisms to monitor information, search for discrepancies in equilibrium, and make necessary real-time adjustments to achieve "ideal" conditions. While the 1,500-acre city is sprouting up from reclaimed land, its technological integration is marketed as a sustainable city.

Songdo Central Park and Lake, photo: Wikimedia Commons, link.

Ben Hecht, the president and CEO of Living Cities, finds exceptional promise in big data, specifically in its ability "to measure the human condition in real time." 'Humanity's Dashboard,' as some call it, would be comprised of key metrics, such as income, health, and connectedness, which can help us understand how we are doing as a people... The dashboard combined with the predictive possibilities that the data revolution has opened up could really enable us to better target our scarce resources and tackle our most complex problems, such as poverty and inequality."[8]

Ultimately, sustainable living modalities and big data modeling systems must be constantly assessed, beta-tested and revamped in order to achieve optimized sustainability and engendered degrees of happiness.

Take Dharavi, for example, an informal settlement in Mumbai, India, whose population ranges from 300,000 to 1 million residents of diverse ethnic and religious backgrounds, all co-existing in dwellings spread over 535 acres of land (with a density of anywhere from 600 to 2000 per acre). Depending on the source, Dharavi's informal economy generates over $500 million, $650 million, or $1 billion annually. This is in large part due to the entrepreneurial nature of the migrant population, which is responsible for a booming textile and pottery industry, and recycling center.

Dharavi, Mumbai. Photo by YGLvoices via Wikimedia Commons, link.

One cannot argue against the promise of using resources wisely and of tackling issues such as poverty, but what can be argued is the data itself: what are its composite measurements, how is it measured, and how are its alleged predictive capabilities pulled and extrapolated to fit most definitions of success, if any? That definitions of success vary is indicative of the need to question the criteria used in assessing modeling parameters. No matter what quantitative and qualitative yardstick is used to measure the efficacy of modeling, there must be a 1:1 ratio between the yardstick and the pragmatic manifestation of big data modeling in order to ordain it an objective success.

How can big data capture Dharavi? Does Dharavi align with the search engine's prescription for happiness and sustainable living? If not, how do our computational models account for this deviation? How can such an informal economy be measured, and how can the level of happiness of its inhabitants be tracked by a sensor, and included in a model for the way new cities are built? While Dharavi is not the model of sustainability, there are aspects of this unplanned settlement that are certainly worthy of study. If we can build a computational model that accounts for such deviations, big data's promise, while limited, can hold a place in the future of urban planning.

Taking Dharavi into consideration, can it be said that computational models of processed big data are measuring up to the idealistic tenets of the end-goal of the models themselves? If this is not the case, further heuristic and hermeneutic troubleshooting and diagnostic quality controls will be necessary in order to amplify the rates of efficacy for big data projections and schemata.

---

1. "City leaders should prioritize outcomes instead of just outputs." This is the first bullet point under the heading 'Actions for City Leaders' in a report released by Laura Lanzerotti, Jeff Bradach, Stephanie Sud, and Henry Barmeier titled *Geek Cities: How Smarter Use of Data and Evidence Can Improve Lives* (http://www.bridgespan.org/getmedia/4e91eca7-ed10-48b1-884d-b3a5f7100ab7/Geek-Cities_How-Smarter-Use-of-Data-and-Evidence-Can-Improve-Lives.aspx) ^

2. "Clever Cities: The Multiplexed Economy." *The Economist.* Accessed September 7, 2013. http://www.economist.com/news/briefing/21585002-enthusiasts-think-data-services-can-change-cities-century-much-electricity ^

3. The family recounts the story of the husband and father who was approached by officers inquiring about the pressure cooker. When the husband explained it was for quinoa, one of the officers responded "What the hell is quinoa?" Gabbatt, Adam. "New York woman visited by police after researching pressure cookers online." *The Guardian.* Accessed August 1, 2013. http://www.theguardian.com/world/2013/aug/01/new-york-police-terrorism-pressure-cooker?CMP=twt_fd&CMP=SOCxx212 ^

4. K.N.C. "The backlash against big data." *The Economist.* Accessed April 20, 2014. http://www.economist.com/blogs/economist-explains/2014/04/economist-explains-10) ^

5. McSharry, Patrick. Quoted from the conference "Potential and Challenges of Big Data for Public Policy-Making." In *IPP 2012 Plenary Panel.* Accessed August 12, 2014. https://www.youtube.com/watch?v=PiOzJlsQcnk ^

6. McSharry, Patrick. Accessed July 20, 2014. https://www.youtube.com/watchfeature=player_embedded&v=vXN3PjwYVjo ^

7. This is the often-quoted "Campbell's Law," attributed to Donald Campbell, a psychologist who worked in the latter half of the 20th century and who, while recognizing the benefits of the emerging technology at the time, also identified corruption as a by-product of our dependence on technology for predictive purposes. ^

8. Hecht, Ben. "Big Data Gets Personal in U.S. Cities." *Data-Smart City Solutions.* Accessed June 14, 2014. http://datasmart.ash.harvard.edu/news/article/big-data-gets-personal-in-u.s.-cities-473 ^