Big Data Factory for Big Decisions

The call to data prayer

Claire SOMERVILLE

lecturer, international affairs executive director of the gender centre, Graduate Institute Geneva

"We need data"- comes the call to nearly every pre-decision making process in policy, programmes, politics and processes be it in public panels or internal meetings- the requests for more data keep on coming. And even if some data are available, the call comes "*we need **more** data*".

To what extent is the echoing ode to data a new phenomenon? Or is the song a response to the multiplying sources of new data purportedly at our fingertips- the raw material of our social, economic and political lives.

Either way, decisions-makers- those with executive and also corporate powers- appear to have become paralysed in making (public) decisions without recourse to "the data". We have moved from the mantra of evidence-based policy, through policy-based evidence to data-driven evidence and policy. The United Nations Sustainable Development Goals 2030 Agenda (SDGs) is a base in point on the call to data. "Trusted, accurate data is key to make sure we move forward on the right track" cried the head of the UN Global Working Group on Big Data at 4th International Conference on Big Data (UNDESA). With 17 goals, 169 targets and 230 indicators – the institutional framework to measure success only on the basis of what can only be "big data" is a mammoth enumeration endeavour that are supposed to hold those in power accountable for their investments. The monitoring and evaluation of the SDGs over a 15-year time span may become the world largest Big Data experiment.

But have data always been so important?

To reflect on the pre-digital revolution when data were something mainly scientists generated with often little connection to the worlds of action and decision--making, brings us to question what we define as data and, furthermore, wonder about how our worlds are reproduced in these new data streams. Are our decisions and actions any better now with all these so--called big data at hand?

And what is this data to which the calls to prayer are invoked? What are data and where is this mystical reservoir of magical essence that will make all human decisions better? Why do leaders and decision-makers feel they need data to move forward?

In an era of unprecedented mistrust, fake news and weakened governance decision-makers appear to lack confidence to act and take responsibility without first letting the data orb whisper it's guidance not unlike the oracles of the Azande much studied in early anthropology (Evans-Pritchard, 1937).

The following chapter takes these questions as a landscape against which to respond to some rather more prosaic questions of our time. What are data these days? What biases might such data hold and reproduce over time?

What is Data?

What scientists inside the academy call data often differs from the kind of things referred to as data, especially "big data", by those outside academia. Data is, perhaps traditionally, something that is generally purposefully collected to respond to a fundamental research question. It follows method and methodology, speaks to ontological and epistemology claims and is "typically obtained by scientific work and used for reference, analysis and calculation" (OED). Raw data, as Dourish and Cruz (2018) recently commented, is an oxymoran – and further still, data, they argue, must be narrated to give shape and meaning. To present meta or big data, as is so often the case, as the "raw material" of human life is to eschew the complexities of data capture and collection, sampling, representativeness, bias, and what van Dijck calls "datafication" (2014). These "big" concepts underpin the science behind data and without which the production of any data risks spurious claims.

Most of that which is described as "big data" is one way or another closely tied with the digitalization of everyday processes and activities. Where once a telephone directory was just that: a place to identify a name, address and landline telephone number in order to contact another person; now it is a searchable, codeable, analysable source of understanding and mapping of geographies, genealogies, migrations, health, education, social and economic status, religion, voting patterns… and the list goes on as such directories are augmented with other big data sources such as google maps, ISP analytics, electricity and water meters, bus timetables and more. The possibilities are exponential and the augmentation apparently seamless. The more we digitalize the deeper and broader our data sets become. So much so that modern data scientists no longer analyse but "mine" their data seeking out that gold nugget on which they can claim their fortunes. The data-rush is here to stay- and so we must become more cogent of the status, use and mining of such sources and the limitation as well as

opportunities they afford us.

What marks out these "new big" data is what McFarland calls "found data" (2015). They arise from observational sources; what I call the outputs of the digitalization of everyday life. They are not purposely sought under scientific rules of design and rigor. They are just "there"; the by-product of other efficiency saving processes of everyday life. Returning to the telephone example: itemized billing of mobile calls together with mobile 4G tracking means it is possible to document the everyday movements and contacts of ordinary people: it is possible – but for what purpose? Whereas, somewhat in contrast to these digital outputs, the academic scientist commences with a research question(s) and considers what data would need to be collected to respond to the question (and then goes out and collects it with a suitable sampling methods and size and appropriate method in a specified period of time), these new sets of mass data are collected for very different purposes with no guiding research question. Their purpose is, for example, to ensure a user pays the correct bill in full knowledge of their usage. Before such invoices were itemized, households simply trusted their service provider to request the correct amount! Trust it would seem, and its corollary, confidence, are two of the unanticipated and undesirable side effects of the rise of these big data.

Whilst not purposeful or designed with the sorts of methodological rigor that scientists inside the academy have spent several hundred years developing, these new forms of big data, these "found data" or by-products of 21st century life, are thought to hold some hidden, previously unknown, possibly un-named aspect of humanity. We can see our lives in ways never known before; we examine streams of sensor data emanating from devices we wear, use, engage and interact with- not even always knowingly.

Data produced as by-products of the ever-growing digital world can and should be subjected to the same or similar processes of rigor and method as those collected purposefully as the raw material of the human sciences. All data, purposeful or by-product, harbour bias – and it is these that I shall discuss by asking whether the sorts of bias that we find in purposeful data are also replicated in by-product data and secondly do are some of the data cleaning and bias reduction strategies employed in the social sciences applicable to big data.

Big Data Bias?

The potential for big data to generate big bias and therefore inaccurate findings is a risk that must be addressed if the intrinsic value of scaled data is to be realized. If we just take a couple of standard methodological concepts from the academic sciences we begin to expose a few of the risks.

Take sampling. The natural and social sciences have developed multiple techniques applicable to quantitative, qualitative and mixed method data collection to ensure that sampling bias are limited in data sets- from calculating P-value significance in hypothesis testing in statistical data to immersive thematic saturation in thick ethnographic data. Academics have a toolbox filled with techniques to design and selection samples and ensure the connected concept of representativeness is fully implicated in all analyses processes. Sampling in the big data sciences is a nascent field and scholars report the "random" is the approach most typically adopted by big data science miners (Rojas et al 2017). Kandel et al, in an interview study of

data scientists, found that data miners were actually wary of using data sampling for fear of (ironically) bias it could introduce to their analysis – and furthermore is contrary to the goal of big data which seeks to use as much data as possible and run experiments at scale. And so we face an inherent contradiction as the new, and yet unproven, big data sciences and scientists try to define the old rules of science.

Let's take another well-founded source of bias in data: missing data. Anyone who has ever engaged in process of "cleaning" a dataset is well-aware the challenge of missing data. Be it as simple as a date of birth or inconclusive blood result on a patient record, to lost files, human error, data entry errors and technology glitches- even the most seemingly straight forward of data counts – the number of people currently living on the planet- are subject to unquantifiable levels of missing data (Wardrop et al., 2018). Missing data can take many forms in big digital data sets and are caused by both human and technological forces including infrastructure outages, update errors, trolling, spam bots even human non-compliance with data capture instruments such as wearable technologies. Sarah Pink and colleagues have begun conceptualizing the gaps in digital data as "broken data" (Pink et al 2018) – and include additional processes that can affect the quality of big data such as decay, repair, re-making and growth. Drawing on ethnographic detailing, Pink begins to demystify the multiple ways in which big data are constituted in all its fragmentations, incompleteness and contingent relations and entanglements with humans as producers but also technology and software as collectors. The materiality of these data cannot live in isolation and do not necessarily have objectively reliable predictive capacities. Missing, decaying and entangled data are intrinsic biases that need to be addressed in the new data paradigm and we will have to re-think our data cleaning processes- as despite its sanitized connotations, big data may well be creating a bigger laundry basket.

Final Thoughts

Taking just two of the key methodological concepts employed in science to understand bias in ordinary data shows us that these are also challenges for big data. The daily work of sampling, selection, accounting for missing and incorrect data are as relevant to big data as to small. Why then is there just now such a call to data as the source of all answers, the reservoir of solutions that have until now escaped our sight but now become visible through digitalization? To a large extent the "call to data" that opened this chapter, and one I hear so often among decision-makers, points to a deeper problem – one of trust, responsibility and belief. The compelling quest to make data-based decisions is in part constructed around the institutional scaffolding of Big Data thinking. The promises heralded by the big data firms peddling and mining away with supercomputer capacities are chipping away at the rather more human capabilities of intelligence and decision-making *sans* data. It is not data itself that renders and speaks but human analyses of data that typically try to simplify complexity and generate "readable, portable and tractable" (Latour 1987) insights. The data brokerage and mining of companies like Cambridge Analytica serve as salient reminders of the fragility and also ethics of using the by-products of digitalization as a source for action.

Since so few of the big data we refer are purposefully sought as part of a methodological design we are left with troubling situations where nearly anything counts as data, especially if it can be quantified or used

to first create and then operationalize algorithms; these mystical formulas known only to the data miners who profess their credibility. Big data comes with the allure of sanitized, objective mirroring of the world and implies as Jasanoff suggests "a panoptic viewpoint from which the entire diversity of human experience can be seen, catalogued, aggregated, and mined so that the narratives derived from the data speak as for themselves, compelling reasonable people to action" (2017). Yet taking just two examples of data bias make visible the cracks in big data "science". Should we be compelled to act in response to the facts produced by big data mining? Even if yes, we should bear in mind a healthy anthropological spin of all things factful: "Anyone can produce a new fact; the thing is to produce a new idea" said Evans-Pritchard of the claims of the witches, oracles and magicians of the Azande.

EVANS-PRITCHARD, E. E. 1937. *Witchcraft, oracles, and magic among the Azande.*

DOURISH, P., E. CRUZ. 2018. Datafication and data fiction: narrating data and narrating with data- Big Data and Society 1:10

JASANOFF, S. 2017. Virtual, visible, and actionable: Data Assemblage and the sightlines of justice. Big Data and Society 1:15

PINK, S., RUCKENSTEIN, M., WILLIM, R., M. DUQUE. 2018. Broken data: conceptualizing data in an emerging world. Big Data and Society 1:18

McFARLAND, D., R. McFARLAN. 2015 Big Data and the danger of being precisely inaccurate. Big Data and Society 1:4

UNDESA https://www.un.org/development/desa/en/news/nocat-uncategorized/big-data-for-sdgs.html (access 01/10/2018)

van DIJCK, J. 2014. Datafication, datism and dataveillance: Big Data between scientific paradigm and ideology. Surveillance and Society, 12 (2): 197-208

WARDROP, N. A., JOCHEM, W. C., BIRD, T. J., CHAMBERLAIN, H. R., CLARKE, D., KERR, D., BENGTSSON, L., JURAN, S., SEAMAN, V. & TATEM, A. J. 2018. Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences.*