Article accepted by Nature magazine in September 2019, for publication on 17 October 2019

How data came to rule
At the heart of science is where political, social and economic interests meet, shows Sabina Leonelli

Data. The cornerstone of research. The grounding for a scientific understanding of the world. Lightening rods for the negotiation of scientific, political, social and economic interests.

Over the past 150 years ideas of what counts as data or reliable data and who owns it have shifted dramatically. Once data were regarded as stable objects whose scientific significance was determined by a handful of professional interpreters. Now they are re-usable goods whose mettle depends on the extent to which they are mobilised across contexts and aggregated with others. Growing in volume, variety and value, data have come to drive the very process of discovery.

This explicit designation as intrinsically valuable assets has only become possible through a complex web of institutional, technological and economic developments. The history and consequences of how this web has been woven has repeatedly transformed research practice and its role in society.

Collecting commodities

Until the start of 19th century, efforts to collect facts and objects of relevance to research were spearheaded by visionary individuals, typically backed by wealthy patrons. Naturalists roamed the globe in search of new biological specimens. Court astronomers devised tools to observe new parts of the cosmos. The large quantities of data accumulated were systematised and analysed through simple and powerful models (think Kepler's laws) and classification systems (such as Linnaeus'). Thus was born the myth of the heroic theoretician, mining order from the chaos of observations. This individualistic view was tied to an understanding of data as fundamentally private – its scientific value residing in conceptual interpretation.

The 19th century marked a shift. Data were recognised and institutionalised as social commodities. Their scientific, financial and political value, arose from

investments and required regulation and oversight. The botanical wonder cabinet that was Paris' natural history museum was re-organised as a world-leading, publicly accessible repository of objects of potential scientific value. By the 1850s, the natural history museums of Berlin, London and New York followed suit.

The centralisation of food markets spawned standardised approaches to the valuation and trade of organisms – such the crop measures devised by the Chicago Board of Trade. Cholera epidemics in Europe spurred large scale collection of information on the spread and targets of disease. New methods of visualisation and analysis emerged, such as John Snow's famous maps of how contaminated water spread cholera in central London.

National weather services started to build links between data collected regionally. The 1853 Brussels Convention on naval meteorology coordinated ship logbooks into the first quasi-global data records for climate science. In Berlin the first real bureau of standards, the Physikalische-Technische Reichsanstalt, appeared in 1871 with physicist Hermann von Helmholtz as its founding director and a mandate to generate data needed for society as a whole. That same decade, the U.S. Army tasked the Library of the Surgeon-General's office to collect as many disease case reports as possible. Within twenty years it had become the largest medical library in the world.

National treasures

By the turn of the century, the rise of nation states and the increasing demands of international trade drove initiatives to measure nature and society in a more systematic, objective way. National information infrastructures helped regions share data, marking the start of a new informational globalism (Hewson 1999). International entities such as the League of Nations and the International Monetary Fund, yearned to globalise data collection and analysis for many purposes and across all scientific domains.

For example, the Permanent Commission on Biological Standardisation was created to monitor drug tests and biological assays from 1924. Information on employment, unemployment, wages and migration was amassed by the new International Statistical Commission. Such initiatives were fostered by an ever-expanding group of researchers, administrators, merchants and politicians.

All this fuelled the development of sophisticated approaches to quantification. Statistics emerged as a separate discipline -- the main source of information for emerging insurance practices and public health monitoring systems (Desrosieres 1993, Porter 1995). Techniques were developed to match the complexity of

social exercises such as the census (van Oertzen 2018). Population-level thinking gripped the life sciences too – **f**or good and ill, given its significance for the eugenic movement. A new type of data collection focused on genetic mutants of the same model species (Kohler 1994) – a trend that exploded in the molecular era (Strasser 2019).

The two world wars severely disrupted data collection and sharing in the short term. But from the 1940s on, the huge military investment in intelligence and information technologies kick-started the post-war drive towards mechanised computing. The space race was perhaps the most dramatic Cold War contribution to globalised data systems and practices, particularly satellite technology. This produced the first global view of the planet and spurred the inauguration of the Intelsat system for worldwide civil communications networks in 1967.

The World Meteorological Organisation was founded in 1950 to oversee the international linkage of regional weather services, for instance in the Global Atmospheric Research Program. The International Geophysical Year of 1957-58 marked a step change in the commitment of geophysical and oceanographic sciences to global data exchange. In the midst of the Cold War, it was also a diplomatic achievement (Anorova et al 2010).

Global goods

From the 1970s, virtually every scientific field was building global, digitalised infrastructures for data sharing. The United Nations consolidated its global environmental monitoring system just as the World Health Organisation systematised its efforts to map the spread of infectious diseases. The holy grail became the development of tools, such as computer models, that could crunch numbers at a previously unimaginable scale.

Increasingly, data was seen as a sharable asset, for repurposing, the value of which could change depending on its use. This view owed much to the cybernetics movement, with its emphasis on modularity and complexity (Pickering 2010). It was also informed by the growth of international trade and the rising recognition of research as an engine of economic growth, military power and international relations.

Also in the 1970s, big science projects carried out at Los Alamos in the United States and CERN in Geneva took center stage as models for how to do research (Price 1967). Here the production and trade of data were no longer the responsibility of individual researchers. Rather they were the output of large investment and collective efforts carried out in centralised experimental

facilities. In many fields, of course, such centralisation was unfeasible, for instance in environmental, biological and climate sciences working with observational rather than experimental data. Yet even here there was a focus on building networks for sharing data to feed more information to novel computational tools.

Since the 1980s**,** portable computers, modelling and simulations have shaped the collection, manipulation and archiving of data. Climate scientists have developed sophisticated ways to use legacy data to reconstruct a history of the atmosphere at the global level. This effort drove the pooling of international data, culminating in 1992 in the Global Climate Observing System.

In biology, the quest to map biodiversity moved to the molecular level with the big sequencing projects, first in model organisms such as the worm *Caenorhabditis elegans*, then through the Human Genome Project (Hilgartner 2017). Sequencing databases were re-imagined as playgrounds for discovery to facilitate immediate, low-cost sharing, visualisation and analysis online, transforming the massive investment in genomic data production into useful biological knowledge.

Open season

As global data infrastructures and related institutions became ever more sophisticated, the resources needed to maintain them have mushroomed, and in ways that do not fit contemporary regimes of funding, credit and communication within science. For example, the curators of biological databases do essential work. But they do not routinely publish in top-ranking journals and may not be recognised nor rewarded as high-level researchers. Similarly, keeping digital platforms robust and fit for purpose requires serious investment. The more data move around and are used for a variety of purposes, the more vulnerable they are to unwarranted and even misleading forms of manipulation and enrichment.

Over the past few decades, the Open Science movement has called for widespread data sharing as fundamental to better science. These efforts prompted the birth of journals devoted principally to the publication of datasets; ambitious investment in data infrastructures, exemplified by the European Open Science Cloud; the FAIR guidelines for how data should be labelled and managed to be re-usable (Wilkinson et al 2016). There have also been calls to better reward data stewards (such as technicians, archivists and curators), to raise their professional status from support workers to knowledge creators (Directorate-General for Research and Innovation 2018).

These reforms are temporary solutions to a large-scale crisis of the contemporary research system, rooted in the inability to reconcile the diverse social and scientific valences of data. The crisis recalls how the 20th century reconfigured research data as political and economic assets, whose ownership can confer and signal power, and whose release may constitute a security threat – as in the Cold War efforts to contain geological data that may have signalled nuclear testing. Now new technologies are intersecting with emerging regimes of data ownership and trade. Starting from the 2000s, a handful of corporations has created – and wielded control over – new kinds of data left by billions of people as they meet, work, play, shop and interact.

As algorithms become ever more opaque, the transparency and accountability of techniques and tools used to interpret data is declining. While data curators remain the Cinderellas of academia, those who understand and control data management have climbed company ranks. And concerns are growing around data property rights, especially in the wake of misuses of personal data by the likes of Facebook and Cambridge Analytica.

Such tensions between data as public goods and private commodities have long shaped data practices and technologies. Consider for instance the acrimonious debate over the ownership and dissemination of genomic data (Maxson Jones et al 2018). That time free sharing won out, through the establishment of the Bermuda Rules – an agreement among publicly-funded researchers to deposit their sequences to public databases as soon as possible. Wildly successful, this paved the way for Open Data practices in other fields. Yet it also emphasised the financial advantages of owning genomic data (Parry 2004, Sunder Rajan 2006) – a lesson swiftly learnt by companies that sequence and claim to interpret clients' genomes. These typically retain and use the data.

Value added

The use of big data as input for AI systems relies on the promise of global, comprehensive, easily available data riches. In principle, the marriage of powerful analytic tools with big biological data can support personalised medicine and precision agriculture. Similarly, social data hoovered from internet platforms and social services can inform evidence-based policy, business strategies and education. And yet, history shows that moving research data around is not so simple. Underpinning technical questions around data integration and use, there are thorny social, ethical and semantic issues.

How to get different research cultures to communicate effectively? How to collect, share and interpret data generated by the state, industry or social media?

What expertise and stakeholders should have a say in data management and analysis? Who should have access to data, when and how? Addressing these issues required effective administration and monitoring, and a long-term vision of the research domain at hand (Edwards 2010, Daston 2016). It also demanded a repertoire of skills, methods and institutions geared to the study of specific research objects (Ankeny and Leonelli 2016).

In sum, data generation, processing and analysis are unavoidably value-laden. The scientific legitimacy of these activities depends on the extent to which such values are held up for public scrutiny. Indeed, the best examples of data-intensive research to this day include strategies and methods to explicitly account for the choices made during data collection, storage, dissemination and analysis. Model organism databases such as PomBase and FlyBase, for instance, clearly signal the provenance of the data that they store, including information about who created them, for which purpose and under which experimental circumstances. Users may then assess the quality and significance of data (Leonelli 2016). Similarly, the widely used Catalogue of Somatic Mutations in Cancer (COSMIC) captures the provenance of data and the interpretive decisions taken by its curators while processing them. This helps clinicians to re-assess the value of data (Forbes et al 2017).

The more such assumptions and judgement are filtered by large digital infrastructures, the easier it becomes to hide or lose them, making it impossible for future generations to adequately situate the data. Data are cultural artefacts whose significance is clear only once their provenance – and subsequent processing - is known.

Technological development - particularly digitisation – has revolutionised the production, methods, dissemination, aims, players and role of science. Just as important, however, are the broad shifts in the processes, rules and institutions which have determined who does what, under which conditions and why. Governance, in a word. Data emerge from this reading of history as relational objects. Objects whose very identity as sources of evidence – let alone their significance and interpretation - depends on the interests, goals and motives of the people involved, and their institutional and financial context. Extracting knowledge from data is not a neutral act.

Building robust records of the judgements baked into data systems, supplemented by explicit reflections on whom they represent, include and exclude will enhance the accountability of future uses of data. It also helps to bring questions of value to the heart of research, rather than pretending that they are external to the scientific process as has arguably happened in bioethics (Leonelli 2017). This is a crucial step towards making big data sciences into

reliable allies for tackling the grave social and environmental challenges of the 21st century.

References

Ankeny RA and Leonelli S (2016) Repertoires: A Post-Kuhnian Perspective on Scientific Change and Collaborative Research. *Studies in the History and the Philosophy of Science: Part A* 60: 18-28.

Anorova E, Baker KS and Oreskes N (2010) Big Science and Big Data in Biology : From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) Network. *Historical Studies of the Natural Sciences* 40 (2): 183–224.

Daston L (ed) (2016) *Science in the Archives*. Chicago, IL: University of Chicago Press.

Desrosieres A (1993) *La politique des grands nombres: Histoire de la raison statistique*. Paris: Editions La Découverte.

Directorate-General for Research and Innovation (European Commission) (2018) *OSPP-REC: Open Science Policy Platform Recommendations*. URL: https://ec.europa.eu/research/openscience/pdf/integrated_advice_ospp_recommendations.pdf

Edwards PN (2010) *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, ML: MIT Press.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L., Stefancsik, R., Harsha, B., Kok, C.Y., Jia, M., Jubb, H., Sondka, Z., Thompson, S., De, T., Campbell, P.J., 2017. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* 45, D777–D783.

Kohler RE (1994) *Lords of the Fly: Drosophila Geneticists and the Experimental Life*. Chicago, IL: Chicago University Press.

Kriege J (1996) *History of CERN. Volume III. The Years of Consolidation 1966-1980*. Amsterdam: North Holland.

Hewson M (1999) Did Global Governance Create Informational Globalism? In: Hewson and Sinclair (eds) *Approaches to Global Governance Theory*. NY: State University of New York Press.

Hilgartner S (2017) *Reordering Life: Knowledge and Control in the Genomics Revolution*. Cambridge, ML: MIT Press.

Leonelli S (2016) *Data-Centric Biology: A Philosophical Study*. Chicago, IL: Chicago University Press.

Leonelli S (2016) Locating Ethics in Data Science: Responsibility and Accountability in Global and Distributed Knowledge Production. *Philosophical Transactions of the Royal Society: Part A*. 374: 20160122.

Maxson Jones K, Ankeny RA and Cook-Deegan R (2018) The Bermuda Triangle: The Pragmatics, Policies, and Principles for Data Sharing in the History of the Human Genome Project. *Journal of the History of Biology*, no. March 1942.

Oertzen C (2017) Machineries of Data Power: Manual versus Mechanical Census Compilation in Nineteenth-Century Europe. *Osiris* 32 (1): 129–50.
Parry B (2004) *Trading the Genome: Investigating the Commodification of Bio-Information.* New York: Columbia University Press.

Pickering A (2010) *The Cybernetic Brain: Sketches of Another Future*. Chicago, IL: University of Chicago Press.

Porter T (1995) *Trust in Numbers*. Princeton University Press.

Price DJD (1963) Little science, big science. NewYork: Columbia University Press.

Strasser B (2019) *Collecting Experiments*. Chicago, IL: University of Chicago Press.

Sunder Rajan, K. (2006) *Biocapital: The Constitution of Postgenomic Life*. London: Duke University Press.