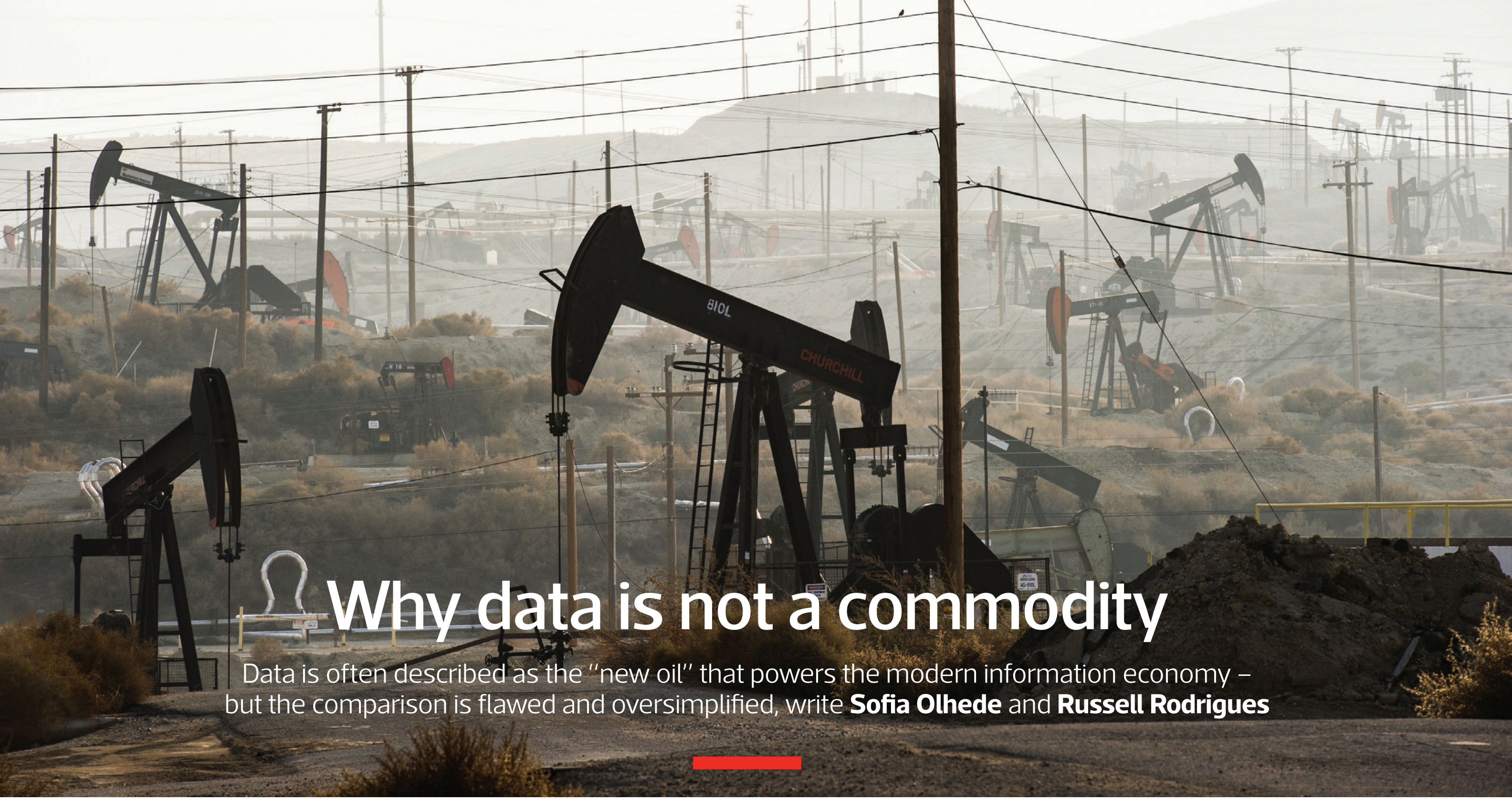

Why data is not a commodity

Data is often described as the "new oil" that powers the modern information economy – but the comparison is flawed and oversimplified, write **Sofia Olhede** and **Russell Rodrigues**

Widespread adoption of fossil fuels during the Industrial Revolution enabled significant advances in mechanisation that accelerated economic growth. Modern economies, by contrast, are increasingly information-based, with commentators frequently describing data as the "new oil".[1,2] It is sometimes suggested that data needs only to be tapped and purified, and one will easily distil key insights to inform and improve decision-making, cut costs and increase wellbeing – and thereby yield significant societal value.

Yet the comparison between data and commodities such as oil is not so straightforward. Commodities are the basic goods and raw materials – including agricultural products, metals and energy resources – whose exchange and use yield value and drive economic activity. While, like other commodities, access to data can in some sense be bought and sold, and data can be subsequently processed and manipulated to extract value, there are important differences which should prevent data from being considered a commodity in a typical sense. Overdependence on simplistic analogies can affect the manner in which we think about data, and its associated opportunities and risks. Imprecise or misleading views of the potential of data and its pitfalls can thereby generate false expectations, and lead to bad decisions.

Value and quality

The main goal of data analysis is to identify patterns or trends and thus gain (by conclusion or inference) insights about particular populations under study. However, insight is intangible, and is not an end-product in itself. The true value of data analysis arises through acting appropriately upon the insights gleaned. Insights are also liable to change over time, or with additional data, and given the limitless range of application areas, will also vary depending on the nature of the input data and on the question of interest. Thus, data is not a uniform, generic and static raw material; it is rather a product of several decisions on aggregation, filtering, deletion and recording, which are usually irreversible. This variability makes it challenging to assign consistent value to data.

By contrast, traditional commodities are essentially consistent and exchangeable, irrespective of the supplier. Two barrels of oil, or two bushels of corn, even from different sources, more or less conform to internationally accepted standards. Commodities of a particular type have identical uses, are marketed in discrete quantities, and, crucially, are uniform in quality (fungible) – essential requirements to facilitate trade. This is not the case with data, for which the manner of use and necessary quantities will vary according to the issue of interest. In any case, greater quantities of data do not guarantee better insights, because its quality is also important.

Sofia Olhede is a professor of statistics at University College London, and scientific director of the UCL Big Data Institute.
Credit: Kirill Photography

Russell Rodrigues is operations manager of the UCL Big Data Institute.

But quality too is difficult to standardise. It is constrained by the methods used in data collection, processing and handling. If these are not adapted to the particular questions of interest, the data gathered may embed diverse underlying assumptions, may fail to represent the populations of interest, and could contain biases, errors, or missing values. The Streetbump app, for example, is less likely to detect bumps and potholes in less affluent areas of Boston where smartphone ownership is lower, thereby giving a misleading impression of road conditions (bit.ly/2vPK8s6). Good experimental design can help to avoid scenarios like this, but as large quantities of data are increasingly gathered opportunistically, and incorporated into analyses for which they were not specifically collected, direct comparisons between amalgamated data sets may not even be possible. Collating, merging and/or splitting different data sets may introduce confounding effects from different populations or cause misalignment in the units of observation. Like crude oil, data may represent a mixture, but it is not so straightforward to refine data and remove contaminating factors. Thus, a good understanding of data provenance – where it comes from and the protocols that produced it – helps avoid intrinsic flaws that can produce incorrect conclusions.

Unfortunately, data is often not collected in formats readily conducive to analysis, and there may be significant obstacles to overcome before analysis can be performed. For instance, handwritten hospital records require conversion to machine-readable text before their information content can be probed – and such processes may cause information to be misinterpreted or lost. In other cases, the sheer volumes of data collected may pose challenges for storage, necessitating that only a subset be retained and the rest discarded, often irretrievably. Such decisions can inadvertently influence the computed outputs by masking patterns within the data and, being generally taken at the investigator's discretion, are very difficult to standardise. While the path from crude oil to petrol is well understood and consistent, the same is not true of the route from data to insight.

Data wrangling

There are approaches to addressing data quality issues, both before and after processing, and these comprise the growing area known as data wrangling.[3] Wrangling aims to tidy up data – removing outliers, repetitions, and unreliable observations – and store it in a format that facilitates analysis. However, this process is time-consuming, occupying an estimated 80% of processing efforts.[4] Standard protocols exist for refining oil or purifying water, but data wrangling usually needs to be implemented more creatively, depending on the condition of the data, and often resembles more of an art than a science.

A further complication is that many modern sources of data evolve continuously – social media networks offer a good example. In such cases, it is not only the core variables of interest that matter, but also the associated metadata and timestamps. On Facebook or Twitter, where users can delete as well as add content, data analysed at one point may suggest a particular pattern, which differs at another point in time. Moreover, analysis algorithms can be designed to adapt in response to changing input data, and this often makes it challenging to replicate analysis at a given point. Data may also increase or decrease in value over time, as the practical utility of its information content changes. Timestamps can help analysts establish baselines in scenarios where data structures are dynamic and evolving, but if recordings are discontinuous, it can be difficult to draw reliable conclusions. By contrast, the value of commodities generally is unaffected by their time of origin.

The personal aspect

Another key consideration with data is that it very often comprises information on people,[5,6] and insights obtained through analysis often inform decisions made about them. Commodities, however, rarely affect specific individuals. The use of natural resources does raise broad ethical issues pertaining to sustainability and stewardship, but while oil leaks cause damage indiscriminately, the leakage of personal data, or the inappropriate use thereof, can cause irreversible harm to specific persons. The notion of informed consent – that individuals should be advised on how their data is used, and have the capacity to withhold or withdraw it – is being stretched by the dynamic nature of data. Often the precise uses for data may be unknown at the time of collection. As such, the concept of dynamic consent is emerging,[7] proposing that individuals can alter their consent over time, depending on the purpose of analysis. Further work will be required to develop analytic tools capable of handling data for which consent may fluctuate.

Further reading
This article is the third in an ongoing series of articles on data, algorithms and society. Go online to read "Fairness and transparency in the age of the algorithm" (bit.ly/2wZSgnL) and "The computer ate my personality" (bit.ly/2wZpkML).

Drilling down into the notion of data as a commodity, one finds the picture rather more complex than is sometimes presented. Caution must be exercised to avoid overly simplistic analogies, especially since much data is personal in nature. But it is helpful to think of data as the "new oil" in at least one respect. Fossil fuels powered the Industrial Revolution, but there was little concern at the time for the wider impacts this would have, particularly on health and the environment. Today we must be wary of rushing to embrace easy solutions to the usage and storage of data, and the implementation of algorithms, without due care for the consequences. ■

References

1. Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B. and Sethupathy, G. (2016) *The Age of Analytics Competing in a Data-Driven World*. McKinsey & Company, December.

2. The Economist (2017) The world's most valuable resource is no longer oil, but data. *The Economist*, 6 May.

3. Lohr, S. (2014) For big-data scientists, "janitor work" is key hurdle to insights. *New York Times*, 17 August.

4. Donoho, D. (2015) 50 years of data science. Paper presented at the Tukey Centennial Workshop, Princeton, NJ, 18 September.

5. National Academies of Sciences, Engineering, and Medicine (2017) *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: National Academies Press.

6. British Academy and Royal Society (2017) *Data Management and Use: Governance in the 21st Century*. London: British Academy and Royal Society. bit.ly/2wKroqD

7. Budin-Ljøsne, I., Teare, H. J. A., Kaye, J. *et al.* (2017) Dynamic consent: A potential solution to some of the challenges of modern biomedical research. *BMC Medical Ethics*, **18**, 4.

Christopher Halloran/Bigstock.com