70. The use of big data in the analysis of inequality

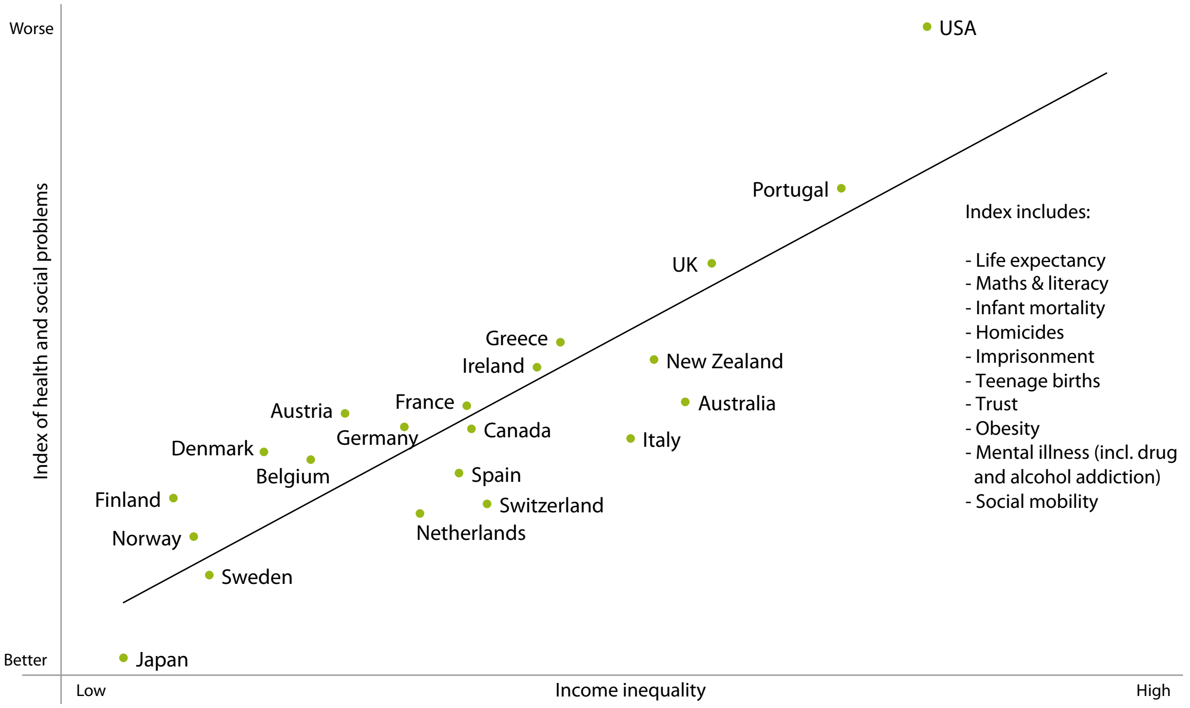
Mike Savage

Like it or not, big data is happening. How should social scientists engage with it?

There is a lively debate these days about how social scientists should engage with 'big data'. For some, it is a distraction which detracts from rigorous causal analysis. For others it offers more promise (Savage and Burrows, 2007). Like it or not, big data is happening.[1] It is part of the emergent world of new kinds of data assemblage which forms the terrain on which contemporary social science needs to stand. Failure to join in will leave social scientists without a foothold in the emerging devices which will produce knowledge and information in the twenty-first century. Big data is, and will be, particularly relevant to the analysis of inequalities.

We have seen a striking shift towards data-driven analysis in the past few years. In place of grand theories of inequalities, the highest-profile works have been based on large-scale data analysis, including the work of Robert Putnam (2000) on the impoverishment of our connections with each other, Wilkinson and Pickett (2009) on why more equal societies almost always do better, and Thomas Piketty (2013) on wealth and inequality in the twenty-first century. These authors, and the ways they describe the phenomena they are concerned with, make the social science of the past appear overly burdened by theory, philosophy and history, and social scientists' pronouncements too abstract to inform more than academic debates about an issue.

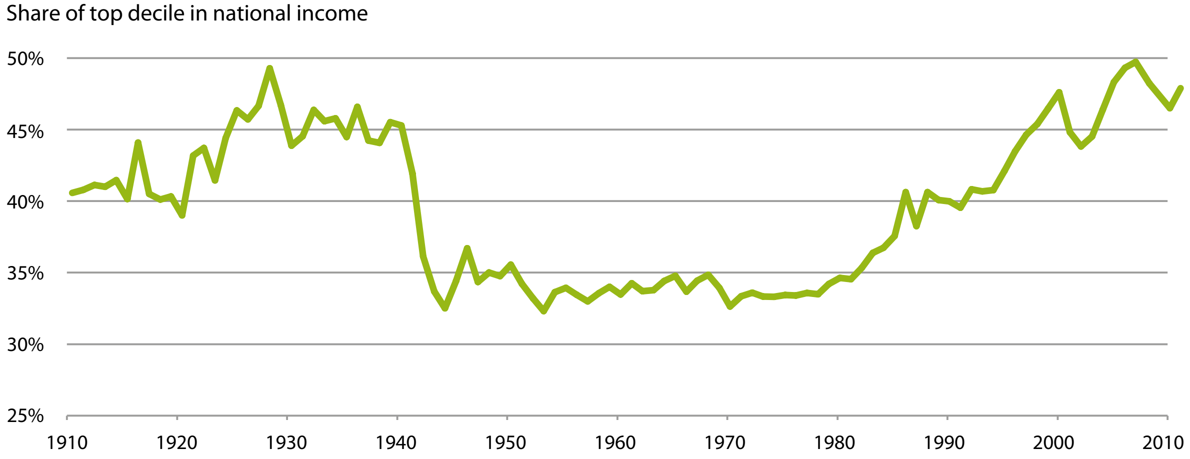
Figure 70.1 **Health and social problems are worse in more unequal countries**



Source: Wilkinson and Pickett (2009).

This article features in the *World Social Science Report 2016*, UNESCO and the ISSC, Paris. Click here to access the complete Report.

Figure 70.2 **Income inequality in the USA, 1910–2010**



Sources: See *http://piketty.pse.ens.fr/fr/capital21c.*

The current crop of influential social scientists make innovative use of multiple data sources. Some of these are digitally stored data, others are more conventional data sets, but regardless of format, these 'data assemblages' now form the vanguard of social scientific analysis. They have been used to great effect to relate important facts about our world and the inequalities of today. A good example is the graphic representations in the work of Wilkinson and Pickett (*Figure 70.1*), who plot one set of data (income inequality across countries), against other indices within a country (such as rates of drug abuse and imprisonment, or decreases in physical and mental health) and show the pattern of this covariation.

Social science has been at best ambivalent and at worst dismissive of the power of visualizations, and has characteristically preferred to deploy textual and numerical assemblages. Yet every picture can tell a story that a thousand words cannot. Piketty has become influential in part because of his skilful use of the U-shaped curve to tell a powerful story of the fall and rise of inequality over the past 120 years.

The fundamental challenge in analysing big data is to reduce its complexity to a clear pattern, and to depict this graphically in a succinct way. Putnam, Wilkinson and Pickett, and Piketty succeed in this regard not only by using new sets of data – though without calling it big data – but also in finding simple visualization tools which summarize complex data.

In order to clarify this, we need to recognize that there are different issues involved in what 'big data' means. I draw a distinction between data sources (such as tax returns), the power of descriptive analytic strategies, and the big vision (see e.g. Oxfam, 2015).

Data sources

Putnam uses changes in the membership of clubs and trade unions in the USA over time to grasp and demonstrate a decline in civic engagement. Wilkinson and Pickett's data sources are records of income inequality and various health and social problems. Piketty uses taxation data to plot, for example, the relationship over time between the top percentiles of earners as a proportion of the national income. This and other comparisons of a similar kind support his argument that we are now moving towards income inequality last seen in the nineteenth century. More particularly, he uses available data sources on tax returns to show that when the rate of return on capital is greater than the rate of economic growth over the long term, the result is further concentration of wealth. In all cases, the bedrock of the analysis is the skilful deployment of multiple data sources, which often exceed standard national representative surveys.

Descriptive analytic strategies

The ways in which the diverse data assemblages are woven together and analysed, most typically using simple univariate and bivariate techniques, get to the heart of the new style of data-based social science.

The data sources and data points are big in that they reach across time (Putnam), across space (Wilkinson and Pickett), and across both space and time (Piketty), compared with the kind of data gathered and methods of analysis typical of older social science. Conventional social science emphasizes the need to focus analyses on specific 'dependent variables', which are then explicated through examining the potential causal power of numerous 'independent variables'. But here there are a great number of 'dependent variables', but only one main 'independent variable' (such as inequality). Rather than the 'parsimony' championed in mainstream social science, what matters here is 'prolixity', with the piling-up of examples of the same kind of relationship, punctuated by telling counterfactuals.

The big vision

The skill in using big data on inequality effectively is to stand back from the data, not get too involved in the detail, and allow the overall patterns, such as Piketty's U-shaped curves, to tell the story of a certain pernicious relationship between the structure of our world and the consequences. Rhythms of change have accelerated since the great social theorists of the past such as Marx, Durkheim and Weber first wrote about the structure of societies. The strength of the new style of doing social science depends on repeat visualizations with a recurring theme, each linked to an overarching story that effectively captures the central argument of the author, to deploy concepts with wide intellectual resonance and political implications. In doing this, social scientists can establish themselves as noteworthy commentators on our realities, and thus motivate change or transformation. This is in contrast to the often narrow and specialized ways in which big data is deployed by non-social scientists.

In conclusion, this is not the last word on the uses of big data (see Chang et al., 2014; González-Bailón, 2013) or the ways in which social scientists could use big data to inform citizens, policy-makers and debates more widely about the inequalities that mark our world. The most effective kinds of social science now are 'data-rich'. However, they are also theoretically sophisticated and offer an alternative form of data analysis to technocratic models derived from computation and information sciences.

The authors highlighted above are also deeply 'objective' in that they carefully report their data sources, their analysis of them, and thus where their findings come from. Yet all three also take a passionate, even politicized, view of their purposes. Putnam makes it clear that he wants to halt the decline of social capital. Wilkinson and Pickett are deeply perturbed by how socially damaging inequality is, while Piketty's concern to document the dynamics of wealth and income inequality throughout the past century and to reform capitalism is also clear. Data, theory and politics are richly and fruitfully combined in these three works. As we look to the future, it is important for social scientists to demonstrate their effectiveness in knowing how best to present and select data, in contrast to the data-driven and empiricist models often used in more technocratic visions of the 'big data' world, which are poorly placed to analyse inequality. Social scientists should not feel threatened by 'big data'. They should embrace their skills and sophistication in knowing how to deploy it to best effect.

Note

[1] For a very short history of 'big data' see Press (2013).

Bibliography

Chang, R. M., Kauffman, R. J. and Kwon, Y. 2014. Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems*, Vol. 63, pp. 67–80.

González-Bailón, S. 2013. Social science in the era of big data. *Policy and Internet*, Vol. 5, No. 2, pp. 147–60.

Oxfam. 2015. *Wealth: Having it All and Wanting More.* Oxford, UK, Oxfam GB for Oxfam International.

Piketty, T. 2013. *Capital au XXIe siècle* [*Capital in the Twenty-First Century*]. Paris, Editions du Seuil. (In French.)

Press, G. 2013. A very short history of big data. Forbes. *www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#2715e4857a0b180a284b55da* (Accessed 16 June 2016.)

Putnam, R. 2000. *Bowling Alone: The Collapse and Revival of American Community.* New York, Simon & Schuster.

Savage, M. and Burrows, R. 2007. The coming crisis of empirical sociology. *Sociology*, Vol. 41, No. 5, pp. 885–99.

Wilkinson, R. and Pickett, K. 2009. *The Spirit Level: Why More Equal Societies Almost Always Do Better.* London, Allen Lane.

■ ***Mike Savage*** *(UK) is professor of sociology at the London School of Economics (LSE) and co-director of the International Inequalities Institute at LSE. www.lse.ac.uk/sociology/whoswho/academic/savage.aspx*

This article features in the *World Social Science Report 2016*, UNESCO and the ISSC, Paris. Click here to access the complete Report.