blogs.lse.ac.uk
http://blogs.lse.ac.uk/impactofsocialsciences/2016/01/05/the-next-decade-of-data-science-rethinking-key-challenges-faced-by-big-data-researchers/

The Next Decade of Data Science: Rethinking key challenges faced by big data researchers

The vast availability of digital traces of unprecedented form and scale has led many to believe that we are entering a new data revolution. Will these new data sources and tools allow us to improve business processes in transformative ways? **Vyacheslav Polonski** *argues that the more data is available, the more theory is needed to know what to look for and how to interpret what we have found.*

Some data scientists dream of a time when massive datasets and distinctly accurate analytics will allow them to finally truly understand and predict human behaviours; a time when the emergent *"data revolution"* will come to its logical conclusion in providing scientist with a constant flow of real-time data on political, economic and social interaction, making theoretical assumptions about such phenomena obsolete. Over the course of the next few years of data research, there will be many temptations to give in to the wondrous promises of big data, but data scientists need to think one step ahead, delineate illusion from reality and stay true to the traditions of the scientific method. This blog aspires to convey precisely this argument.

The "end of theory" is an idea that has been popularised by Chris Anderson in a Wired-magazine article, where he proclaims that that the vast availability of digital traces and other data sources of unprecedented form and scale will fundamentally transform the realm of the social sciences and industry research. Anderson continues: "Petabytes [of data] allow us to say: 'Correlation is enough.' We can throw numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot." In other words, given the availability of fine-grained time-stamped records of human behaviour at the level of individual events, data analysts could increasingly succeed without "outdated" theoretical models.

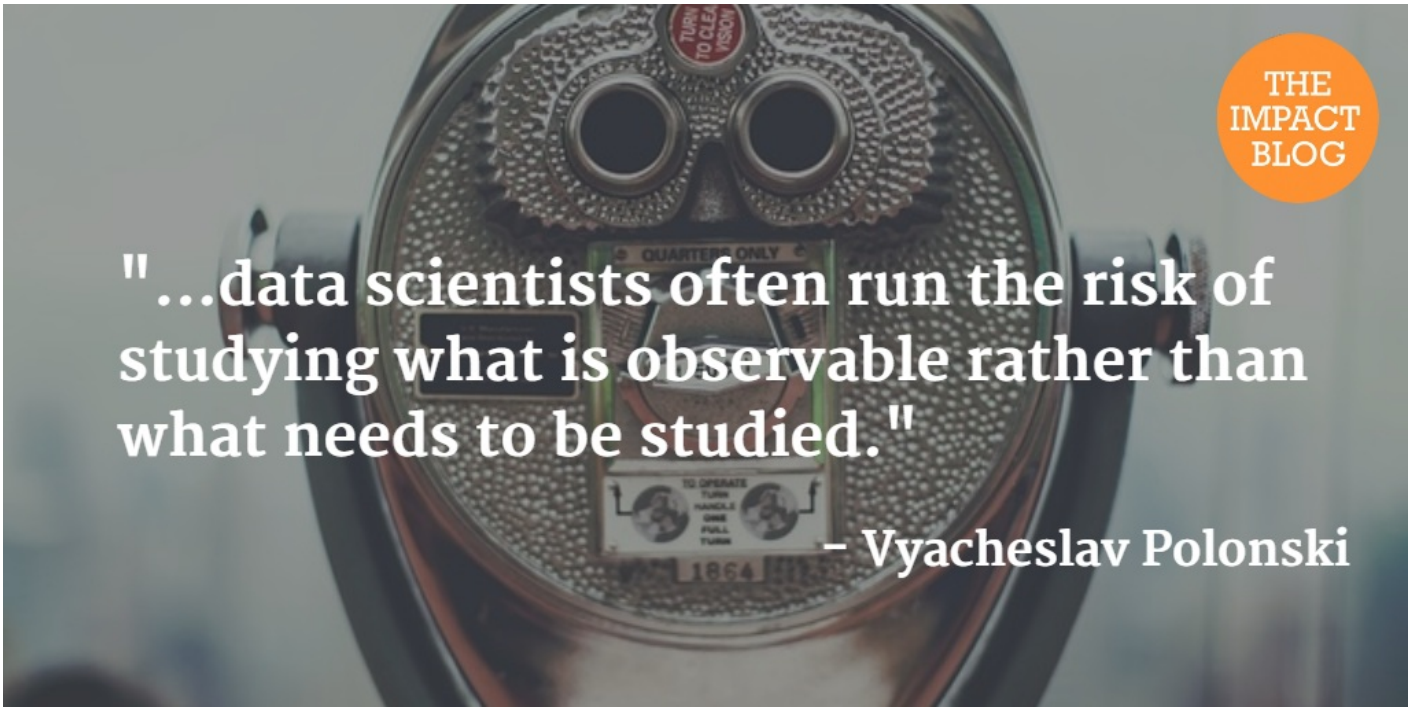
Image credit: markusspiske (CC0)

More to the point, with new data sources and new algorithms, data analysts could now get to new answers that were essentially inconceivable before, overcoming the dominant methodological paradigms that analysed 21st century problems with 20th century methods. Within the social sciences, there is indeed much hope that the current advances in online social research will allow us to situate atomistic explanations within the relational context of complex social networks. This could help elucidate interesting phenomena such as the dynamics of social influence, the spread of information and the emergence of cultural norms. As Microsoft researcher Duncan Watts suggests, "we have finally found our telescope"; a beautiful metaphor that heralds a new age for data-driven research—except that it is somewhat inappropriate for the social sciences.

This approach is problematic for at least three reasons. First, data analysts are naturally part of the social world they seek to study. This is intuitive. Over time, humans may develop situational awareness about what is happening to their data and, consequently, may react to the direct or indirect presence of the researchers in a way that an electron never could. The implication is, therefore, that social scientists shape the social systems they inhabit, whether they like it or not. The metaphor of the telescope is, however, only relevant for the natural sciences and, for the most part, inappropriate for the study of human behaviour.

Second, it is no surprise that the most valuable datasets are proprietary and difficult to access. The severe constraints on publicly available datasets have meant that the researchers' ability to study specific social phenomena is impeded by asymmetries in data access and distribution. Given past examples of API changes and restrictions, it is evident that this introduces a distinct power relationship between the academic and corporate worlds over who gets access to data and for how much. In the future, this is expected to only get worse, as data analysts become increasingly dependent on such data sources. Furthermore, given the frequent lack of consent on behalf of the studied participants, ethical and privacy implications may play an increasingly critical role in this power relationship.

Third, and most importantly, even if we assume the universal availability of traced data at population scale, social scientists still need theory to make sense of this unstructured data deluge. Theoretically motivated hypotheses are instrumental to understanding where to aim the telescope *ex-ante*, as well as in order to move from mere measurement to explanation *ex-post*. As Golder and Macy note, big data research "may lack the theoretical grounding necessary to know where to look, what questions to ask, or what the results may imply".



On this basis, data scientists often run the risk of studying what is observable rather than what needs to be studied. This is especially problematic when researchers attempt to make causal claims from the data, neglecting the

possibility of confounding factors. Even large sample sizes cannot overcome this limitation. In fact, the more data is available, the more theory is needed to know what to look for and how to interpret what we have found.

Whatever the future holds, there are plenty of reasons to be optimistic about the next decade of data science and the potential impact of big data on society. However, it is equally important to be cognizant of the theoretical, methodological and practical challenges associated with a possible future of this research paradigm. Thus we have not yet reached the end of theory; as the world changes around us, we are instead entering a new renaissance for the scientific method in industry research and the social sciences.

Note: This article gives the views of the author, and not the position of the Impact of Social Science blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

About the Author

Vyacheslav Polonski *is a network scientist at the Oxford Internet Institute, researching complex social networks, the emergence of collective behaviours and the role of digital identity in technology adoption. He has previously studied a Harvard University, Oxford University and the London School of Economics and Political Science. Vyacheslav is actively involved in the World Economic Forum and its Global Shapers community, where he is the Vice-Curator of the Oxford Hub. He writes about the intersection of sociology, network science and technology on Medium and Twitter.*