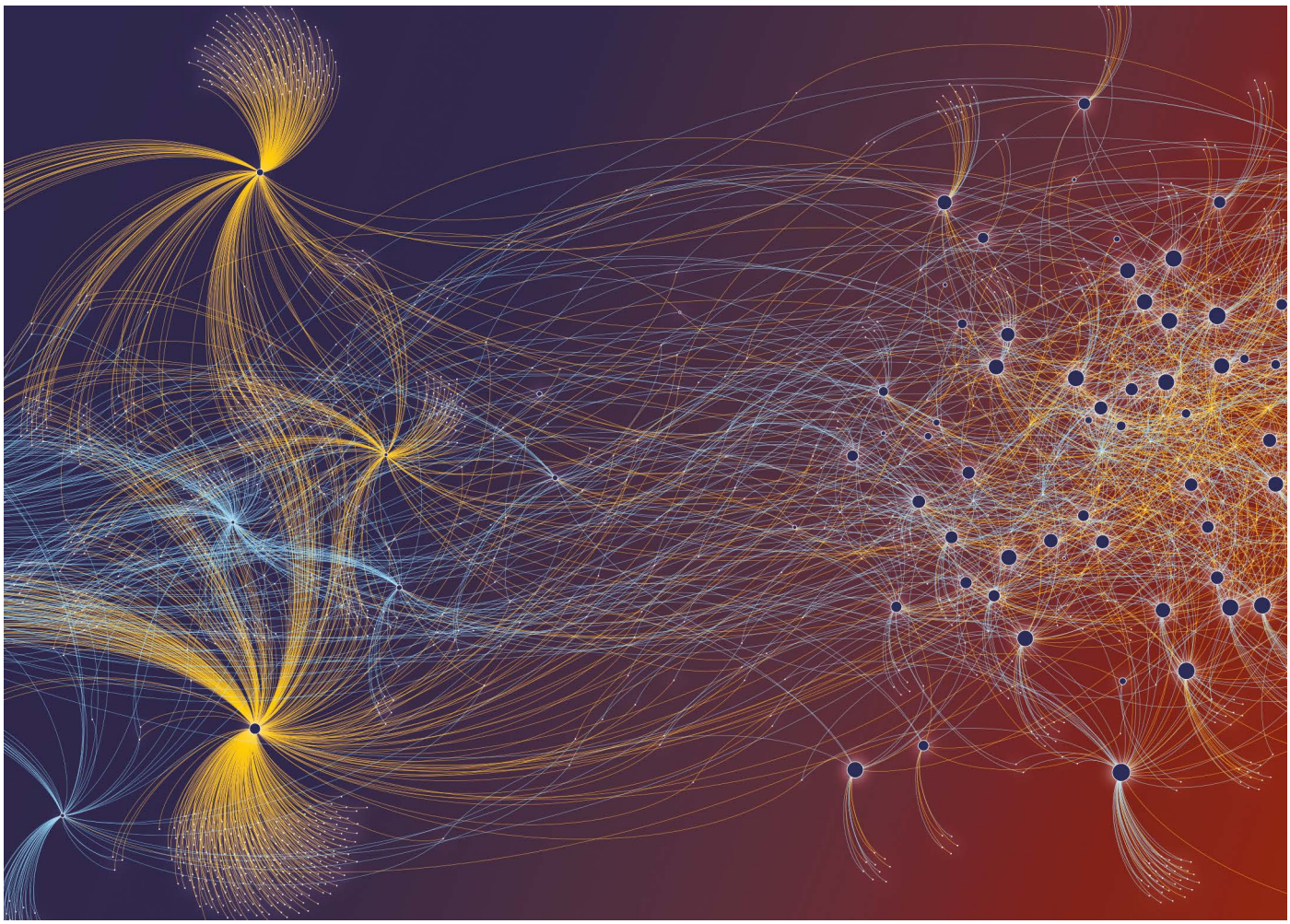

MINZHANG ZHENG AND NEIL JOHNSON

Scientists studied data from thousands of social-media users to analyse clusters perpetuating extremism.

COMPUTING HUMANITY

How data from Facebook, Twitter and other sources are revolutionizing social science. **By Heidi Ledford**

Elizaveta Sivak spent nearly a decade training as a sociologist. Then, in the middle of a research project, she realized that she needed to head back to school.

Sivak studies families and childhood at the National Research University Higher School of Economics in Moscow. In 2015, she studied the movements of adolescents by asking them in a series of interviews to recount ten places that they had visited in the past five days. A year later, she had analysed the data and was feeling frustrated by the narrowness of relying on individual interviews, when a colleague pointed her to a paper analysing data from the Copenhagen Networks Study, a ground-breaking project that tracked the social-media contacts, demographics and location of about 1,000 students, with five-minute resolution, over five months[1]. She knew then that her field was about to change. "I realized that these new kinds of data will revolutionize social science forever," she says. "And I thought that it's really cool."

With that, Sivak decided to learn how to program, and join the revolution. Now, she and other computational social scientists are exploring massive and unruly data sets, extracting meaning from society's digital imprint. They are tracking people's online activities; exploring digitized books and historical documents; interpreting data from wearable sensors that record a person's every step and contact; conducting online surveys and experiments that collect millions of data points; and probing databases that are so large that they will yield secrets about society only with the help of sophisticated data analysis.

Over the past decade, researchers have

used such techniques to pick apart topics that social scientists have chased for more than a century: from the psychological underpinnings of human morality, to the influence of misinformation, to the factors that make some artists more successful than others. One study uncovered widespread racism in algorithms that inform health-care decisions[2]; another used mobile-phone data to map impoverished regions in Rwanda[3].

"The biggest achievement is a shift in thinking about digital behavioural data as an interesting and useful source", says Markus Strohmaier, a computational social scientist at the GESIS Leibniz Institute for the Social Sciences in Cologne, Germany.

Not everyone has embraced that shift. Some social scientists are concerned that the computer scientists flooding into the field with ambitions as big as their data sets are not sufficiently familiar with previous research. Another complaint is that some computational researchers look only at patterns and do not consider the causes, or that they draw weighty conclusions from incomplete and messy data — often gained from social-media platforms and other sources that are lacking in data hygiene.

The barbs fly both ways. Some computational social scientists who hail from fields such as physics and engineering argue that many social-science theories are too nebulous or poorly defined to be tested.

This all amounts to "a power struggle within the social-science camp", says Marc Keuschnigg, an analytical sociologist at Linköping University in Norrköping, Sweden. "Who in the end succeeds will claim the label of the social sciences."

But the two camps are starting to merge. "The intersection of computational social science with traditional social science is growing," says Keuschnigg, pointing to the boom in shared journals, conferences and study programmes. "The mutual respect is growing, also."

Computational revolution

In 2007, a small group of scientists with big ambitions convened a meeting to discuss the emerging art of social-science data crunching. They wanted to apply their skills to change the world. During his talk, political scientist Gary King at Harvard University in Cambridge, Massachusetts, said that the deluge of digital information "will make it possible to learn far more about society and to eventually start solving — actually solving — the major problems that affect the well-being of human populations".

By then, a smattering of computational social-science studies had already been published. A 2006 study had looked at the role of social influence on the popularity of music by creating an artificial online music market used by 14,341 people. The participants chose songs to download, sometimes with and sometimes without information about how popular those songs were among their fellow market users. The study found that the popularity of a song became harder to predict the more that users were influenced by others' behaviour[4], offering one explanation for why it is difficult to predict runaway success.

Two years later, a study analysed the movements of 100,000 mobile-phone users over six months, and found that people travel in simple and reproducible patterns[5]. The authors could calculate the likelihood of finding an individual in any particular location, and suggested that identifying similarities in travel patterns across a community could help with urban planning, understanding the spread of disease or preparing for emergencies.

"Over time, each side is understanding the other in terms of language and methods."

That same year, the technology magazine *Wired* published an article[6] arguing that the era of big data would spell an end to theory across the sciences. Although widely criticized as an oversimplification, the article struck a nerve: more than a decade later, social scientists repeatedly invoke the *Wired* article as a signal that the relevance of social-science theory was under attack.

But big data only continued its ascendancy. To Duncan Watts, a sociologist at the University of Pennsylvania in Philadelphia, the changes in social science were reminiscent of what happened in biology during the 1990s, when high-throughput techniques began generating reams of data about DNA sequences and gene expression. "There was this avalanche in new data that required thinking about data in a very different way," he says.

But many conventional social scientists were unimpressed by the initial fruits of the revolution, and found some of its methods questionable. Sceptics viewed studies of social media as experiments conducted on thousands of unknowing and unconsenting participants. In 2018, news broke that the British consulting firm Cambridge Analytica had gathered data from millions of Facebook accounts without the consent of their owners. The aftermath of the scandal continues to bring added scrutiny and scepticism to social-media research, and some scientists have had their projects stymied as platforms institute new privacy policies.

Socially awkward

The field was also stigmatized by early papers that addressed 'toy' problems — questions that could be answered from the data, but did not address long-standing, fundamental issues in the social sciences, such as how to tackle inequality or influence public opinion. "There were a lot of Twitter studies in the beginning that I think social scientists were not very excited about," says Claudia Wagner, also a computational social scientist at the GESIS Leibniz Institute for the Social Sciences.

Some argue that the embrace of toy problems was at least in part the product of a young field finding its feet. As analyses have become more sophisticated and data sources more diverse, the field has started tackling more important issues, such as the roots of discrimination, inequality and radicalization, says Strohmaier. "Only now are we getting the kind of data that allow us to look at the big issues," he says.

Last year, for example, researchers from public health and from behavioural economics used health-care records for more than 50,000 patients in a US health-care system to analyse a commonly used algorithm that recommends people with complex medical needs for extra supervision and health interventions. The team used modelling to show that the algorithm was systematically discriminating against Black people — potentially influencing the care of millions of people[2]. The researchers then used knowledge of health-care disparities in the United States to track down the sources of that bias, and to suggest ways to remove it. For example, algorithms shouldn't assume that the amount spent on an individual's health care is a good proxy for how much care they need: because of unequal access to health care, less money is typically spent caring for Black Americans than white Americans, even when they have the same health-care needs.

But access to good data isn't the only challenge: scientists migrating from physics or computer science stand accused of failing to examine the theories that social scientists have formulated to explain human behaviour. "They tend to look for patterns," says Giulia Andrighetto, who trained as a philosopher but is now a computational social scientist at the Institute of Cognitive Sciences and Technologies, part of Italy's National Research Council in Rome. "But typically they don't look for the mechanisms through which those behaviours are generated."

To do that work requires a firm grasp of social-science theory. Jisun An, a computational social scientist at Hamad Bin Khalifa University in Doha, started her PhD in computer science in 2010, studying news sharing on social media just as the computational social-science movement began to bloom. At the start, she worked only with other computer scientists, and they struggled to wrap their heads around different social-science theories. Now, she collaborates with political scientists to study the influence of the media on public opinion — and vice versa — as well as

KEVIN FRAYER/GETTY

Mobile-phone data suggests that humans stick to simple, predictable movement patterns.

how to encourage people to boost the diversity of their news sources. "Over time, each side is understanding the other in terms of language and methods," says An.

There are now concrete signs of engagement. The first major conference bringing together the two approaches is scheduled for 2021. Universities are also creating institutes that bring together staff from different departments to bridge the divide. George Mason University in Fairfax, Virginia, has a dedicated department, for instance. A summer camp for computational social science runs in more than 30 locations around the world, and a bevy of enthusiastic young students — along with a boost to the number of jobs available — have given some hope that the power struggle could give way to richer collaborations.

Social gathering

The union of the two approaches can be powerful. Data scientist Joshua Blumenstock at the University of Washington in Seattle and his colleagues used mobile-phone data from millions of people in Rwanda to infer their socioeconomic status, then confirmed their results by comparing them with data collected using conventional surveys[3]. The resulting method could be used by policymakers to target poor regions of the country in need of interventions, for example, or to monitor the effects of policies that have been enacted.

But a lack of communication is still evident. Joan Donovan, a social scientist at Harvard, points to a study published last year in which researchers mapped out a network of online hate groups on the Facebook and VKontakte platforms, and showed how the structure of the network changed over time[7]. The physicists and computer scientists who carried out the study failed to cite key social-science studies in their work, she says, and as a result, their interpretation of their findings wasn't as rich as it could be. They also surveyed too few social-media platforms, when past research had shown that hate groups follow charismatic leaders across many domains. And the team came to what she considers a dangerous conclusion: that social-media platforms could try to steer discussion in hate groups, for instance by creating false accounts or engineering in-fighting between hate clusters. This could backfire by increasing the volume of discussion in the group and boosting its ranking on search algorithms, she says. A better strategy, she thinks, would be to check the spread of hate messages by having search engines limit the visibility of such groups.

Physicist Neil Johnson at George Washington University in Washington DC, and lead author of the hate study, is accustomed to criticism from social scientists. He says he cited the most relevant references. And as for search algorithms, social-media companies have the power to manipulate them, he says, "just as they are doing now to suppress the prominence of anti-vaccine and COVID-19 misinformation pages and groups". He has studied misinformation, conflict and extremism and says he gets complaints every time he publishes a high-profile paper. But his work has struck a chord with policymakers: he is frequently asked to consult by organizations who like the quantitative nature of his work and the ability to model what impact interventions might yield. "We can really look at concrete questions in a way that I think they haven't experienced in interactions with other academics," he says. Johnson, for his part, is concerned that too many social scientists are rushing into computational approaches without proper training.

Johnson isn't the only scientist sceptical of the importance of theory to their projects. Giangiacomo Bravo, who trained as a socioeconomist and is now a computational social scientist at Linnaeus University in Växjö, Sweden, says that many social-science theories are too nebulous to be tested using big data. The idea of social capital, for instance, is sometimes defined as the shared understanding and values in a society that allow individuals to work together. "The original formulation of this concept of social capital was just too vague to be tested," he says. "How could I measure it?"

Some theories, however, are more concrete. Andrighetto, who studies social norms — the shared rules that govern what is or is not acceptable behaviour in a society — says that researchers have spent a decade piecing together clear definitions and theories for this topic. For example, the theory suggests that when social norms shift, that should prompt changes in how a person responds to a given situation. Social norms are also thought to change only slowly and through the course of intensive social interactions. Testable statements such as these allow Andrighetto to combine computational work with social-science theory: she uses online experiments[8] to test whether simulated changes in social norms influence behaviour.

She is not alone in wanting to use social sciences to change the world. Too often, Watts says, he and other academic researchers are chasing publications rather than real-world solutions. "I felt like my job was done at the moment when the paper was published," he says. "It was my job to put these ideas out there, and it was somebody else's job to come along and figure out how to translate them into meaningful interventions in the real world."

For that shift to happen, researchers from both camps must sustain the momentum towards collaboration, says Watts. Some can already feel it happening. "Traditional social science and computational social science are actually becoming closer over time," says Wagner. "In 20 years, there will be no divide."

Heidi Ledford is a senior reporter with *Nature* in London.

1. Sekara, V., Stopczynski, A. & Lehmann, S. *Proc. Natl Acad. Sci. USA* **113**, 9977–9982 (2016).
2. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. *Science* **366**, 447–453 (2019).
3. Blumenstock, J., Cadamuro, G. & On, R. *Science* **350**, 1073–1076 (2015).
4. Salganik, M. J., Dodds, P. S. & Watts, D. J. *Science* **311**, 854–856 (2006).
5. González, M., Hidalgo, C. & Barabási, A. *Nature* **453**, 779–782 (2008).
6. Anderson, C. 'The end of theory: The data deluge makes the scientific method obsolete.' (*Wired*, 23 June 2008).
7. Johnson, N. F. *et al. Nature* **573**, 261–265 (2019).
8. Realpe-Gómez, J., Vilone, D., Andrighetto, G., Nardin, L. G. & Montoya, J. A. *et al. Games* **9**, 90 (2018).