POLICY FORUM

SOCIAL SCIENCE

Computational social science: Obstacles and opportunities

Data sharing, research ethics, and incentives must improve

By **David M. J. Lazer[1,2], Alex Pentland[3], Duncan J. Watts[4], Sinan Aral[3], Susan Athey[5], Noshir Contractor[6], Deen Freelon[7], Sandra Gonzalez-Bailon[4], Gary King[2], Helen Margetts[8,9], Alondra Nelson[10,11], Matthew J. Salganik[12], Markus Strohmaier[13,14], Alessandro Vespignani[1], Claudia Wagner[14,15]**

The field of computational social science (CSS) has exploded in prominence over the past decade, with thousands of papers published using observational data, experimental designs, and large-scale simulations that were once unfeasible or unavailable to researchers. These studies have greatly improved our understanding of important phenomena, ranging from social inequality to the spread of infectious diseases. The institutions supporting CSS in the academy have also grown substantially, as evidenced by the proliferation of conferences, workshops, and summer schools across the globe, across disciplines, and across sources of data. But the field has also fallen short in important ways. Many institutional structures around the field—including research ethics, pedagogy, and data infrastructure—are still nascent. We suggest opportunities to address these issues, especially in improving the alignment between the organization of the 20th-century university and the intellectual requirements of the field.

We define CSS as the development and application of computational methods to complex, typically large-scale, human (sometimes simulated) behavioral data (*1*). Its intellectual antecedents include research on spatial data, social networks, and human coding of text and images. Whereas traditional quantitative social science has focused on rows of cases and columns of variables, typically with assumptions of independence among observations, CSS encompasses language, location and movement, networks, images, and video, with the application of statistical models that capture multifarious dependencies within data. A loosely connected intellectual community of social scientists, computer scientists, statistical physicists, and others has coalesced under this umbrella phrase.

MISALIGNMENT OF UNIVERSITIES

Generally, incentives and structures at most universities are poorly aligned for this kind of multidisciplinary endeavor. Training tends to be siloed. Integrating computational training directly into social science (e.g., teaching social scientists how to code) and social science into computational disciplines (e.g., teaching computer scientists research design) has been slow. Collaboration is often not encouraged, and too often is discouraged. Computational researchers and social scientists tend to be in different units in distinct corners of the university, and there are few mechanisms to bring them together. Decentralized budgeting models discourage collaboration across units, often producing inefficient duplication.

Research evaluation exercises such as the United Kingdom's Research Excellence Framework, which allocate research funding, typically focus within disciplines, meaning that multidisciplinary research may be less well recognized and rewarded. Similarly, university promotion procedures tend to underappreciate multidisciplinary scholars. Computational research infrastructures at universities too often cannot fully support analysis of large-scale, sensitive data sets, with the requirements of security, access to a large number of researchers, and requisite computational power. To the extent these issues have been partially resolved in the academy (e.g., with genomic data), lessons have not fully made their way into practice in CSS.

INADEQUATE DATA-SHARING PARADIGMS

Current paradigms for sharing the kinds of large-scale, sensitive data used in CSS offer a mixed bag. There have been successes built on partnerships with government, especially in economics, from the study of inequality (*2*) to the dynamics of labor markets (*3*). There are emerging, well-resourced models of administrative data research facilities serving as platforms for analyzing microlevel data while preserving privacy (*4*). These offer important lessons for potential collaboration with private companies, including the development of methodologies to keep sensitive data secure, yet accessible for analyses (e.g., innovations in differential privacy).

The value proposition for private companies is different and there has been predictably less progress. Data possessed by government agencies are held in trust for the public, whereas data held by companies are typically seen as a key proprietary asset. Public accountability inherent in sharing data is likely seen as a positive for the relevant stakeholders for government agencies, but generally, far less so for shareholders for private companies. Access to data from private companies is thus rarely available to academics, and when it is, it is typically granted through a patchwork system in which some data are available through public application programming interfaces (APIs), other data only by working with (and often physically in) the company in question, and still other data through personal connections and one-off arrangements, often governed by nondisclosure agreements and subject to potential conflicts of interest. An alternative has been to use proprietary data collected for market research (e.g., Comscore, Nielsen), with methods that are sometimes opaque and a pricing structure that is prohibitive to most researchers.

We believe that this approach is no longer acceptable as the mainstay of CSS, as pragmatic as it might seem in light of the apparent abundance of such data and limited resources available to a research community in its infancy. We have two broad concerns about data availability and access.

First, many companies have been steadily cutting back data that can be pulled from their platforms (*5*). This is sometimes for good reasons—regulatory mandates (e.g., the European Union General Data Protection Regulation), corporate scandal (Cambridge Analytica and Facebook)—however, a side effect is often to shut down avenues of potentially valuable research. The susceptibility of data availability to arbitrary and unpredictable changes by private actors, whose cooperation with scientists is strictly voluntary, renders this system intrinsically unreliable and potentially biased in the science it produces.

[1]Northeastern University, Boston, MA, USA. [2]Harvard University, Cambridge, MA, USA. [3]Massachusetts Institute of Technology, Cambridge, MA, USA. [4]University of Pennsylvania, Philadelphia, PA, USA. [5]Stanford University, Stanford, CA, USA. [6]Northwestern University, Evanston, IL, USA. [7]University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [8]University of Oxford, Oxford, UK. [9]The Alan Turing Institute, London, UK. [10]Institute for Advanced Study, Princeton, NJ, USA. [11]Social Science Research Council, New York, NY, USA. [12]Princeton University, Princeton, NJ, USA. [13]RWTH Aachen University, Aachen, Germany. [14]GESIS–Leibniz Institute for the Social Sciences, Cologne, Germany. [15]University of Koblenz-Landau, Landau, Germany. Email: d.lazer@northeastern.edu

Second, data generated by consumer products and platforms are imperfectly suited for research purposes (*6*). Users of online platforms and services may be unrepresentative of the general population, and their behavior may be biased in unknown ways. Because the platforms were never designed to answer research questions, the data of greatest relevance may not have been collected (e.g., researchers interested in information diffusion count retweets because that is what is recorded), or may be collected in a way that is confounded by other elements of the system (e.g., inferences about user preferences are confounded by the influence of the company's ranking and recommendation algorithms). The design, features, data recording, and data access strategy of platforms may change at any time because platform owners are not incentivized to maintain instrumentation consistency for the benefit of research.

For these reasons, research derived from such "found" data is inevitably subject to concerns about its internal and external validity, and platform-based data, in particular, may suffer from rapid depreciation as those platforms change (*7*). Moreover, the raw data are often unavailable to the research community owing to privacy and intellectual property concerns, or may become unavailable in the future, thereby impeding the reproducibility and replication of results.

INADEQUATE RULES

Finally, there has been a failure to develop "rules of the road" for scientific research. Despite prior calls to develop such guidance, and despite major lapses that undermined public trust, the field has failed to fully articulate clear principles and mechanisms for collecting and analyzing digital data about people while minimizing the potential for harm. Few universities provide technical, legal, regulatory, or ethical guidance to properly contain and manage sensitive data. Institutional Review Boards are still generally not attuned, and consistent in their response, to the distinct ethical challenges around digital trace data. The recent modification of the Common Rule in the United States, which concerns ethics of human subjects research, did not fully address these problems.

For example, in a networked world, how should we deal with the fact that sharing information about oneself intrinsically provides signals about those with whom one is connected? The challenges around consent highlight the importance of managing security of sensitive data and also of reimagining institutional review processes and ethical norms; yet few universities integrate infrastructure and oversight processes to minimize the risks of security lapses.

Cambridge Analytica, and other, similar, events, have engendered an impassioned debate around data sovereignty. Battle lines have been drawn between privacy advocates and companies, where the former seek to minimize the collection and analysis of all individual data, whereas the latter seek to justify their collection strategies on the grounds of providing value to consumers.

Often missing in public debates are voices for policies that would encourage or mandate the ethical use of private data that preserves public values like privacy, autonomy, security, human dignity, justice, and balance of power to achieve important public goals—whether to predict the spread of disease, shine a light on societal issues of equity and access, or the collapse of the economy. Also often missing are investments in infrastructures in the academy that could power knowledge production and maintain privacy.

Resources and rules, incentives and innovations

Strengthen collaboration

- Develop enforceable guidelines in collaborations with industry around research ethics, transparency, researcher autonomy, and replicability.
- Develop secure data centers supplemented by an administrative infrastructure for granting access, monitoring outputs, and enforcing privacy and ethics rules.

New data infrastructures

- Develop large-scale, secure, privacy-preserving, shared infrastructures driven by citizen contributions of time and/or data. Capture the metadata that describe the collection process.
- Develop infrastructure to capture the dynamic, algorithm-driven behavior of the major platforms over time.
- Promote legal frameworks that allow and mandate ethical data access and collection about individuals and rigorous auditing of platforms.

Ethical, legal, and social implications

- Professional associations should help develop ethical guidelines.
- Large investments are needed to develop regulatory frameworks and ethical guidance for researchers.

Reorganize the university

- Develop structures that connect researchers having shared interests in computational approaches.
- Fundamentally reconceive graduate and undergraduate curricula.
- Reward collaboration across silos.
- Appoint faculty with multi-unit affiliations
- Physically collocate faculty from different fields
- Allocate internal funding to support multidisciplinary collaboration.
- Empower and enforce ethical research practices—e.g., centrally coordinated, secure data infrastructures.

RECOMMENDATIONS

In response to these problems, we make five recommendations.

Strengthen collaboration

Despite the limitations noted above, data collected by private companies are too important, too expensive to collect by any other means, and too pervasive to remain inaccessible to the public and unavailable for publicly funded research (*8*). Rather than eschewing collaboration with industry, the research community should develop enforceable guidelines around research ethics, transparency, researcher autonomy, and replicability. We anticipate that many approaches will emerge in coming years that will be incentive compatible for involved stakeholders.

The most widespread and longest-standing model is open, aggregated data such as Census data. The aforementioned models developed to share government data, with an emphasis on security and privacy, offer promise in working with corporate data. The United Nations Sustainable Development Goals call for partnerships on public-private data sources to provide a wide variety of new, very rich neighborhood-by-neighborhood measures across the entire world (*9*), and national statistical offices in every corner of the world are quietly working on producing such products, but progress is slow owing to lack of funding. The development of secure administrative data centers supplemented by an administrative infrastructure for granting access, monitoring outputs, and enforcing compliance with privacy and ethics rules offers one model for moving forward. As noted above, this model has already been demonstrated in the domain of government administrative data; as well as in a few cases, by telecommunications and banking companies.

Similar models are rare—but becoming more common—for academic research. The Open Data Infrastructure for Social Science and Economic Innovations in the Netherlands is one example. Facebook has iterated through multiple models for collaboration with academics. In its early years, it focused on one-off collaborations, largely in-

formally negotiated. After the 2016 election, it launched Social Science One, providing access to aggregate data of news consumption, which, despite being well resourced, has faced challenges in providing data (*10*).

Coronavirus disease 2019 (COVID-19) has played a particular role in creating partnerships between researchers and companies to produce insights regarding the trajectory of the disease. (COVID-19 has, in many countries, including the United States, also illuminated the fractured and politically contingent nature of much public data regarding the disease.) Twitter has provided a streaming API regarding COVID-19 for approved researchers. Similarly, location data companies such as Cuebiq have provided access to anonymized mobility data. There remain open questions as to what extent these data-sharing efforts will continue after the disease settles into the history books and, if so, how to robustly align them with critical research norms in academia, such as transparency, reproducibility, replication, and consent.

The election examples with respect to Facebook highlight the potentially adversarial role between researchers and corporations. A central contemporary question for the field of CSS (as discussed below) is in what ways particular sociotechnical systems are playing positive and negative roles in society. This tension may partially (but not entirely) be resolved if companies feel that it is in their long-term interest to transparently study and anticipate these issues. Even in the most optimistic scenario, however, there will be a disjuncture between the public interest in the insights that research could produce, and corporate interests.

Academia, more generally, needs to provide carefully developed guidelines for professional practice. What control can companies have over the research process? It clearly is not acceptable for a company to have veto power over the content of a paper; but the reality of any data-sharing agreement is that there are negotiated domains of inquiry. What are the requirements for providing data for replication? What are the needs of researchers for access to a company's internal data management and curation processes?

New data infrastructures

Privacy-preserving, shared data infrastructures, designed to support scientific research on societally important challenges, could collect scientifically motivated digital traces from diverse populations in their natural environments, as well as enroll massive panels of individuals to participate in designed experiments in large-scale virtual labs. These infrastructures could be driven by citizen contributions of their data and/or their time to support the public good, or in exchange for explicit compensation. These infrastructures should use state-of-the-art security, with an escalation checklist of security measures depending on the sensitivity of the data. These efforts need to occur at both the university and cross-university levels. Finally, these infrastructures should capture and document the metadata that describe the data collection process and incorporate sound ethical principles for data collection and use. The Secure Data Center at the GESIS Leibniz Institute for the Social Sciences is an example of shared infrastructure for research on sensitive data. Further, it is important to capture the algorithm-driven behavior of the major platforms over time (*11*, *12*), both because algorithmic behavior is of increasing importance and because algorithmic changes create enormous artifacts in platform-based data collection. It is critical that legal frameworks allow and mandate ethical data access and collection about individuals and rigorous auditing of platforms.

Ethical, legal, and social implications

We need to develop ethical frameworks commensurate with scientific opportunities and emergent risks of the 21st century.

Social science can help us understand the structural inequalities of society, and CSS needs to open up the black box of the data-driven algorithms that make so many consequential decisions, but which can also embed biases (*13*). The Human Genome Project devoted more than $300 million as part of its Ethical, Legal, and Social Implications program "to ensure that society learns to use the information only in beneficial ways" (*14*). There are no off-the-shelf solutions on ethical research. Professional associations need to work on the development of new ethical guidelines—the guidelines developed by the Association of Internet Researchers offer one example of an effort to address a slice of the issue. Large investments, by public funders as well as private foundations, are needed to develop informed regulatory frameworks and ethical guidance for researchers, and to guide practice in the field in government and organizations.

Reorganize the university

Computation is adjacent to an increasing number of fields—from astronomy to the humanities. There needs to be innovation in the organization of typically siloed universities to reflect this, with the development of structures that connect diverse researchers, where collaborating across silos is professionally rewarded. Successful examples of institutional practice include the appointment of faculty with multi-unit affiliations (e.g., across computer science and social science disciplines) and of research centers that physically collocate faculty from different fields, as well as allocation of internal funding to support multidisciplinary collaboration. There needs to be a fundamental reconceiving of the development of undergraduate and graduate curricula for training a new generation of scientists. There must be pervasive efforts within the university to empower and enforce ethical research practices—e.g., centrally coordinated, secure data infrastructures.

Solve real-world problems

The preceding recommendations will require resources, from public and private sources, that are extraordinary by current standards of social science funding. To justify such an outsized investment, computational social scientists must make the case that the result will be more than the publication of journal articles of interest primarily to other researchers. They must articulate how the combination of academic, industrial, and governmental collaboration and dedicated scientific infrastructure will solve important problems for society—saving lives; improving national security; enhancing economic prosperity; nurturing inclusion, diversity, equity, and access; bolstering democracy; etc. Current applications of CSS in the global response to the pandemic are emblematic of the broader potential of the field. Beyond generating results that are meaningful outside of academia, the pursuit of this objective may also lead to more replicable, cumulative, and coherent science (*15*). ■

REFERENCES AND NOTES

1. D. Lazer *et al.*, *Science* **323**, 721 (2009).
2. R. Chetty, N. Hendren, P. Kline, E. Saez, *Q. J. Econ.* **129**, 1553 (2014).
3. J. J. Abowd, J. Haltiwanger, *J. Lane, Am. Econ. Rev.* **94**, 224 (2004).
4. A. Reamer, J. Lane, A Roadmap to a Nationwide Data Infrastructure for Evidence-Based Policymaking (2018); https://journals.sagepub.com/doi/abs/10.1177/0002716217740116.
5. D. Freelon, *Polit. Commun.* **35**, 665 (2018).
6. M. J. Salganik, Bit by Bit: Social Research in the Digital Age (Princeton Univ. Press, 2017).
7. K. Munger, *Soc. Media Soc* **5**, 205630511985929 (2019).
8. Social Science Research Council, To Secure Knowledge: Social Science Partnerships for the Common Good (2018); www.ssrc.org/to-secure-knowledge/.
9. IEAG, UN, "A World that Counts—Mobilising the Data Revolution for Sustainable Development." Independent Expert Advisory Group on a Data Revolution for Sustainable Development (2014).
10. G. King, N. Persily, "A New Model for Industry-Academic Partnerships" (Working Paper, 2018); http://j.mp/2q1lQpH.
11. A. Hannák *et al.*, in Proceedings of the 22nd International Conference on World Wide Web (ACM Press, New York, 2013), pp. 527–538.
12. I. Rahwan *et al.*, *Nature* **568**, 477 (2019).
13. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, *Science* **366**, 447 (2019).
14. J. E. McEwen *et al.*, *Annu. Rev. Genomics Hum. Genet.* **15**, 481 (2014).
15. D. J. Watts, *Nat. Hum. Behav.* **1**, 0015 (2017).

10.1126/science.aaz8170