computation and the sociological imagination

by james evans and jacob g. foster

what is computational sociology?

Computational sociology leverages new tools and data sources to expand the scope and scale of sociological inquiry. It opens up exciting frontiers for sociologists of every stripe—from theorists and ethnographers to experimentalists and survey researchers. Above all, it expands the sociological imagination.

Computational sociology is the unintended beneficiary of a massive data windfall. The internet has drawn a growing slice of social, economic, and political life into the digital domain. The sensors embedded in cell phones and social media platforms serve commercial purposes—but they also form a massive, distributed digital observatory.

This observatory routinely captures novel, large-scale data about real-world activities. In addition to "born digital" observations like tweets or geolocation data, digitization projects (think Google Books) make past cultural products available for consideration as well. Together, these sources allow sociologists to observe and analyze human discourse and interaction at unprecedented scales. Unlike astronomers (who must lobby for telescopes) or traditional quantitative sociologists (who must lobby for large-scale surveys), computational sociologists have gotten our first wave of "big data" for free. Computation provides the prosthetic lenses through which we can see this world of big social data, discovering novel patterns that challenge existing theories—or call out for new ones.

The internet is more than a source of observational data; it also furnishes a new platform for active social investigation. Web-scale surveys and experiments allow us to test theories and causal relationships, just like their offline analogs. Online investigations outpace traditional methods in scale, duration, and complexity.

Social researchers of the future will need to write code, wrangle data, and think computationally as well as sociologically.

What are we to make of this? Is computational sociology a passing fad, a tryst with hot new methods? Not at all. Computational sociology throws off both heat and light. It is a reactor. And it's just getting started.

This "computational reactor" has the potential to transform the study of any substantive problem, from political polarization or residential segregation to the emergence of cultural blockbusters or breakthrough scientific ideas. Computational methods extend and complement familiar methods. They also make new demands on our training. Social researchers of the future will need to write code, wrangle data, and think computationally as well as sociologically. We must teach them how to do so.

What, then, are the implications for the practice of sociology in the 21st century? We look to social theory, ethnography, surveys, experiments, and statistics for answers.

from social theory to formal worlds

Social theory is traditionally expressed in natural language, using a rich conceptual vocabulary. It pays for nuance with ambiguity. A rich theory can fit almost any outcome; in explaining everything, it risks explaining nothing. Computational sociologists often express theory in the language of mathematics or algorithm. The resulting formal model can be explored with empirical data to assess its validity. Alternatively, it can be explored by simulation, allowing scholars to probe alternative scenarios by "replaying history" in artificial worlds with different assumptions.

These two approaches can be combined. In one of our own papers, we developed a simple model to investigate how scientists choose what phenomena to study. Formalizing prior verbal theory, our model allowed scientists to be influenced by the past popularity of phenomena as well as by the conceptual distance between them. Using data from millions of articles and patents, we found that unfolding knowledge was most likely generated by a conservative strategy, which connects popular phenomena to less prominent ones nearby. We then used that same model to search for more efficient strategies. We replayed history over and over on a Cray supercomputer, tweaking scientists' strategies to see how quickly they would discover the same universe of knowledge. And we found strategies that would have been much more efficient!

Verbal theory is still essential. It inspires useful formalizations, puts models in a broader context, and allows for productive ambiguity. But explorations like ours are impossible with verbal theory alone.

Contexts, Vol. 18, Issue 4, p. 10-15. ISSN 1536-5042. © American Sociological Association.
http://contexts.sagepub.com. DOI 10.1177/1536504219883850.

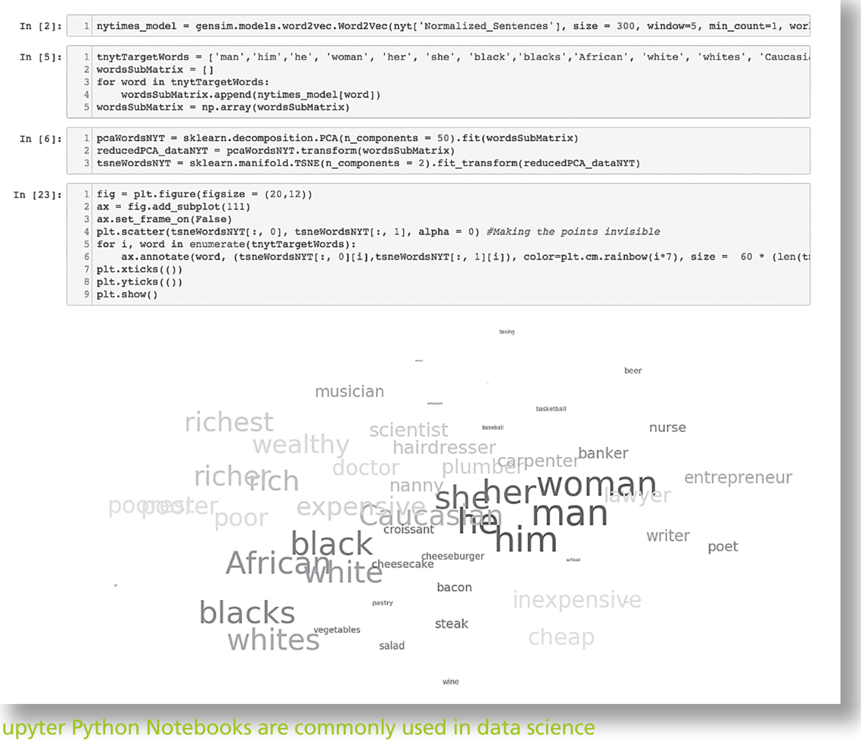
from ethnography to digital observatories

How can computational methods speak to ethnography? The ethnographer embedded in her particular case or community might seem worlds apart from the computational sociologist writing database queries or training deep neural networks. The former is out in the field, using her body as an instrument to capture the delicate texture of lived experience. The latter is bathed in the cool light of dancing data; she worries about eyestrain and carpal tunnel, not rough weather or interpersonal conflict.

To see the connection, recall ethnography's strengths: surfacing novel practices; accumulating rich detail, from which theory can emerge; hunting for surprises that challenge received wisdom and call for revised explanation. Computational sociologists can carry out similar tasks from their digital observatories, gathering massive data at scales beyond the reach of ethnography. This comes at a cost. Rather than the technicolor social reality of the ethnographer, digital traces are typically rendered as computational cartoons. They are reduced by computational methods from dozens, hundreds, or thousands of features per data point to a low-dimensional representation: categories, clusters, or aggregate scores.

Low-dimensional cartoons nevertheless facilitate the discovery of novel behaviors and practices. Kevin Lewis used the choices of online daters to discover that racial biases are temporarily attenuated when daters receive and reply to cross-race messages. Elizabeth Bruch and colleagues showed that online daters screen potential mates with "deal breakers" like age differences or smoking.

These cartoons also allow the testing of ethnographic insight at scale. Computational linguist Rob Voigt and his collaborators trained algorithmic systems to detect the respect that police officers show in their speech during routine traffic stops. By analyzing 36,738 utterances from 981 stops, they demonstrated that officers of all races are systematically less respectful to Black community members.



```
In [2]:  nytimes_model = gensim.models.word2vec.Word2Vec(nyt['Normalized_Sentences'], size = 300, window=5, min_count=1, wor

In [5]:  tnytTargetWords = ['man','him','he', 'woman', 'her', 'she', 'black','blacks','African', 'white', 'whites', 'Caucasi
         wordsSubMatrix = []
         for word in tnytTargetWords:
             wordsSubMatrix.append(nytimes_model[word])
         wordsSubMatrix = np.array(wordsSubMatrix)

In [6]:  pcaWordsNYT = sklearn.decomposition.PCA(n_components = 50).fit(wordsSubMatrix)
         reducedPCA_dataNYT = pcaWordsNYT.transform(wordsSubMatrix)
         tsneWordsNYT = sklearn.manifold.TSNE(n_components = 2).fit_transform(reducedPCA_dataNYT)

In [23]: fig = plt.figure(figsize = (20,12))
         ax = fig.add_subplot(111)
         ax.set_frame_on(False)
         plt.scatter(tsneWordsNYT[:, 0], tsneWordsNYT[:, 1], alpha = 0) #Making the points invisible
         for i, word in enumerate(tnytTargetWords):
             ax.annotate(word, (tsneWordsNYT[:, 0][i],tsneWordsNYT[:, 1][i]), color=plt.cm.rainbow(i*7), size =  60 * (len(t
         plt.xticks(())
         plt.yticks(())
         plt.show()
```

Jupyter Python Notebooks are commonly used in data science

Reid McIlroy and James Evans

These computational cartoons do not capture the rich multimodal inputs experienced by ethnographers. The human body is unparalleled in its ability to sense valuable social information, making the ethnographer's emotions and embodied responses essential data streams. At the same time, current computational methods are too blunt to capture interactional details of sequence and tone, which subfields like conversation analysis dissect through painstaking transcriptional methods.

Digital observatories do not replace the ethnographer. Instead, they demand her input and extend her insight. Approaching digital traces with the mindset of—and in collaboration with—an ethnographer or conversation analyst will be essential if computational sociologists aspire to the nuance and depth of these methods. Ethnographers may suggest new traces to collect, or new constructs to pull out of existing traces. Qualitative data can also justify some of the assumptions required by sophisticated computational methods.

from surveys to online interactions

The collection and analysis of surveys about opinions, preferences, beliefs, and behaviors has been the backbone of quantitative sociology for decades. Often drawing on substantial institutional support, survey researchers contact respondents door-to-door, through the telephone, or over the web. They use a sampling and question strategy designed to maximize the validity and generalizability of their findings. This often involves randomization and a finite set of questions exhaustively answered by respondents.

Instead of asking questions, many computational researchers predict attitudes, preferences, beliefs, and behaviors using the digital breadcrumbs that fall from online discourse. They sometimes use machine learning to impute signals of representativeness (e.g., demographic details like race, class, gender) from images and text. This method allows online data to benefit from the statistical techniques underlying classic survey designs. Other researchers use digital activity traces to select survey respondents. For example, one of us recently used imbalanced contributions to political pages on Wikipedia to sample respondents, and then validated the relationship between their online activities and political preferences using a survey.

Computational sociologists also build and analyze "simple surveys" like those used by Facebook and Twitter, Netflix and Spotify, or Amazon and Alibaba to generate recommendations. These surveys typically involve easy tasks, like rating or comparison. Researchers collect responses from informal, crowdsourced samples, recruiting participants through marketplaces like Mechanical Turk or CrowdFlower, commercial platforms like Facebook or XBox, or engaging websites they create themselves. Just as commercial recommender systems propose the products you want from the products you've bought, machine learning-driven surveys may avoid queries whose answers could have been predicted from other responses. Despite waning response rates to traditional surveys, simple surveys allow us to channel existing online activity by asking fewer and more customized questions.

Using these computational approaches, scholars have analyzed data from more people about more questions than could be exhaustively fielded through traditional surveys. And there's another advantage: many online traces capture what people actually do, think, and prefer in lived online contexts, rather than what they *say* they do, think, or prefer.

But the online data used to study attitudes, preferences, beliefs, and behaviors were often collected for another purpose—usually a commercial one. Digital traces were not designed to produce generalized insight, and can suffer from algorithmic confounding. Online actions are often presented to the user based on predictions about what users want; consider Google's query auto-complete service or Facebook's friendship suggestions. Interpreting the data produced by these online platforms requires accounting for the models that predicted user actions.

There is a deeper concern with using online data: The human subjects involved did not directly consent to the use of their data for research. This lack of consent raises complex ethical and legal questions. Compounding the challenges to research ethics, these new data sources often disclose (or allow inference of) unprecedented personal detail, demanding new techniques to protect research subjects.

from experiments to virtual laboratories

Laboratory and field experiments have been less prominent in sociology than in psychology and economics. They are nonetheless a powerful tool for testing theory and evaluating causal claims. In a classic field experiment, Devah Pager demonstrated the causal role of racial discrimination in hiring outcomes by sending out applicants for entry-level jobs. Applicants were matched in all characteristics but race and (purported) criminal record. Not only did employers discriminate against applicants with a criminal record; white applicants with a criminal record got more callbacks than Black applicants without one.

Online and mobile experiments retain the principal virtue of traditional lab or field experiments: The treatment and "environment" experienced by participants can be controlled (although the latter only through the narrow interface of a computer or mobile screen). Where do online methods shine, then? In the scale and complexity of designs that can be contemplated. In one of the earliest online experiments, Salganik, Dodds, and Watts demonstrated the unpredictability of cultural markets by constructing multiple independent "worlds" where users listened to and downloaded songs. The researchers manipulated whether and how prominently the previous choices of other users were featured. Their findings—that social information increases unpredictability and inequality in cultural markets—could only be cleanly demonstrated with digital controls. With over 13,000 participants, their study required an online experimental platform. It could not be done in the laboratory; the cost and logistics would have been prohibitive.

> Approaching digital traces with the mindset of—and in collaboration with—an ethnographer or conversation analyst will be essential if computational sociologists aspire to the nuance and depth of these methods.

Of course, these new experimental platforms have limits. Because they are delivered through computers or mobile devices, online experiments address only certain sensory modalities. It may also be hard, or impossible, to control what else a participant is doing. The characteristics of online participants may be skewed in ways that could be controlled in a laboratory setting. Some research designs, such as interventions performed on a pre-existing network, require more sophisticated approaches to randomize the assignment of interventions. And there is the looming question of ecological validity: What do online experiments tell us about "real life"?

These limitations are likely to erode. As our devices become higher bandwidth, cheaper, and more pervasive, researchers can control the environment and subject pool with more granularity. As more and more significant social interaction takes place online, the ecological validity of online experiments will certainly increase. For example, the experimental music market of Salganik, Dodds, and Watts is much less artificial today than when the experiment was initially performed! As new technologies like augmented and virtual reality become available for delivering experiments, computational sociologists will be able to carry out even more sophisticated designs. For example, augmented reality could deliver mobile experiments with much greater ecological validity, because the intervention is overlaid directly on the physical world, rather than simply on screen.

The complexity of experimental interventions will also increase. Most provocatively, artificially intelligent agents will be able to deliver precisely calibrated and interactive treatment to experimental participants. Indeed, Hirokazu Shirado and Nicholas Christakis recently deployed AI confederates in an online coordination game experiment—the first step in an exciting new direction for experimental sociology!

Because online experiments address many traditional reservations, the next decade may be a golden age for experimental sociology. In the face of this excitement, we must retain our disciplinary commitment to ethical research. With the surge of virtual laboratories will come greater responsibility to protect those with whom we engage.

Does sociology have a special vocation for computational research? We argue that it does… Embracing our special vocation will require key changes in the sociological curriculum. Graduate students should master languages like Python, R, or Julia in addition to Stata; these are the common tongues of machine learning and data science.

from statistical inference to machine discovery

Statistical inference is the quantitative method most familiar to researchers in the social sciences. Statistical methods co-evolved with survey design, experiments, and the analysis of observational data, always with the goal of identifying generalizable associations between social factors and outcomes of interest.

Machine learning is a more recent arrival. You might think of it as statistics' cool step-sister. Like statistics, it aims to discover patterns and regularities in the (social) world, but its heritage is not exclusively statistical. It draws heavily on computer science and engineering.

In supervised machine learning, "labeled" data are used to train an algorithm that maps from the features of a data point to the appropriate label. The label can be a category, as in classification, or a quantity, as in regression. Rather than focusing on unbiased estimation of specific parameters in a given model, machine learning targets accurate prediction beyond the training data. The trick is finding a model that is complex enough to capture the relevant regularities but not so complex that it overfits, which means that the algorithm goes beyond the signal in the training data to capture noise as well.

Unsupervised methods try to discover and represent the hidden structure of social data. This might involve discovering latent clusters of similar cases; constructing combinations of variables that account for substantial variation; or mapping high-dimensional data to an efficient representation in a smaller number of dimensions, which can be used to reconstruct the original cases and to generate new ones. These methods make new social and cultural patterns available for discovery.

When compared with traditional social statistics, machine learning methods have one great virtue: breadth of application. They embrace both structured and unstructured data of many types, and when applied to social data, they can lead to surprises and new theory building. For example, Filiz Garip used unsupervised machine learning to discover four different clusters of Mexican migrants, three of them mapping onto established theories of migration, and one of them calling out for explanation. Joscha Legewie and Merlin Schaeffer used techniques from machine vision to demonstrate that conflict between ethnic communities tended to occur at "fuzzy boundaries," not sharp borders. Abdullah Almaatouq led a large team in the Fragile Families Challenge, using state-of-the-art machine learning methods to achieve best-in-class performance on the prediction of several important outcomes for both children (GPA and grit) and caregivers (layoff). Despite having over 12,000 features per data point, these methods only achieved modest improvement over the simplest possible prediction (i.e., predicting the average). Their findings suggest that our understanding of the social processes behind child and family wellbeing is quite limited, and call for the development of new methods, data, and theory.

The most powerful and expressive current machine learning models draw on deep neural networks. These algorithms search through complex combinations of data features to learn new composite features that are highly predictive and may be used to generate further features. The resulting models are so complex that they defy human interpretation, and predictive features do not necessarily correspond to causal factors. Hence cutting-edge machine learning may be of limited use in generating meaningful explanations or formulating effective social policy (which relies on the manipulation of causes).

These new methods live in tension with contemporary social statistics, which is geared towards a causal understanding of empirical phenomena (e.g., showing that X "really" causes Y). Machine learning methods produce models that describe the data but do not delve into the question of whether or why one variable might cause another. Still, modern social scientists have begun to recruit insights from machine learning into their quests for causal explanation, and many experts in machine learning recognize that causality is the next great frontier in their field.

imagining the future of computational sociology

In his classic book, C. Wright Mills argued that sociologists should cultivate an imagination that connects individual experiences with large-scale social currents and places them in historical perspective. Recent computational advances in the collection, analysis, and simulation of social data have cultivated an expanded sociological imagination. This computationally-enabled imagination links multiple scales of social life—microscopic and planetary, momentary and historical—while embracing new experiences, subtle social currents, and panoramic historical perspectives.

Does sociology have a special vocation for computational research? We argue that it does. Sociology is the science of societies, concerned with their structure and culture, at large scales. Sociologists are interested in prediction, explanation, and interpretation: three complementary modes of engaging with social data. And sociology is unique in its blend of qualitative and quantitative methods and frequent appreciation of both.

Embracing our special vocation will require key changes in the sociological curriculum. Graduate students should master languages like Python, R, or Julia in addition to Stata; these are the common tongues of machine learning and data science. Departments should offer boot camps on basic techniques for gathering and cleaning data, along with short courses to introduce foundational concepts from discrete math and linear algebra. Quantitative methods courses should discuss unstructured or complex data (like text or networks) at an early stage, as in Kosuke Imai's *Quantitative Social Science.* And sociologists should embrace the collaborative style that is pervasive in other computational sciences. In many cases, sociologists are better served by being able to communicate and collaborate with experts in cutting-edge machine learning approaches, rather than waiting for those methods to be black boxed and delivered.

Computation allows us to model, measure, and modify both social structures and the texture of individual experience—and to do so on bigger and smaller scales than ever before. The sociological imagination has been blown wide open. We must not forget that the new possibilities unleashed by the digital age—formal worlds and digital observatories, intelligent surveys, virtual laboratories, and machine discovery—require the sociological imagination to achieve their full potential.

Let's start dreaming!

recommended resources

Azoulay, Pierre, Christopher C. Liu, and Toby E. Stuart. 2017. "Social Influence Given (Partially) Deliberate Matching: Career Imprints in the Creation of Academic Entrepreneurs." *The American Journal of Sociology* 122 (4): 1223–71.

Bruch, Elizabeth, Fred Feinberg, and Kee Yeun Lee. 2016. "Extracting Multistage Screening Rules from Online Dating Activity Data." *Proceedings of the National Academy of Sciences of the United States of America* 113 (38): 10530–35.

Foster, Jacob G. 2018. "Culture and Computation: Steps to a Probably Approximately Correct Theory of Culture." *Poetics* 68 (June): 144–54.

Garip, Filiz. 2012. "Discovering Diverse Mechanisms of Migration: The Mexico-US Stream 1970-2000." *Population and Development Review* 38 (3): 393–433.

Imai, Kosuke. 2017. *Quantitative Social Science: An Introduction.* Princeton University Press.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, et al. 2009. "Computational Social Science." *Science* 323 (5915): 721–23.

Legewie, Joscha, and Merlin Schaeffer. 2016. "Contested Boundaries: Explaining Where Ethnoracial Diversity Provokes Neighborhood Conflict." *AJS; American Journal of Sociology* 122 (1): 125–61.

Lewis, Kevin. 2013. "The Limits of Racial Prejudice." *Proceedings of the National Academy of Sciences of the United States of America* 110 (47): 18814–19.

Mills, C. Wright (1959) 2000. *The Sociological Imagination.* Oxford University Press.

Pager, Devah. 2003. "The Mark of a Criminal Record." *The American Journal of Sociology* 108 (5): 937–75.

Rigobon, Daniel E., Eaman Jahani, Yoshihiko Suhara, Khaled AlGhoneim, Abdulaziz Alghunaim, Alex Pentland, and Abdullah Almaatouq. 2018. "Winning Models for GPA, Grit, and Layoff in the Fragile Families Challenge." *arXiv [stat.AP].* arXiv. http://arxiv.org/abs/1805.11557.

Rzhetsky, Andrey, Jacob G. Foster, Ian T. Foster, and James A. Evans. 2015. "Choosing Experiments to Accelerate Collective Discovery." *Proceedings of the National Academy of Sciences of the United States of America* 112 (47): 14569–74.

Salganik, Matthew J., Peter Sheridan Dodds, and Duncan J. Watts. 2006. "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market." *Science* 311 (5762): 854–56.

Sengupta, Nandana, Nati Srebro, and James Evans. 2019. "Simple Surveys: Response Retrieval Inspired by Recommendation Systems." *Social Science Computer Review.* https://doi.org/10.1177/0894439319848374.

Shi, Feng, Misha Teplitskiy, Eamon Duede, and James A. Evans. 2019. "The Wisdom of Polarized Crowds." *Nature Human Behaviour* 3 (4): 329–36.

Shirado, Hirokazu, and Nicholas A. Christakis. 2017. "Locally Noisy Autonomous Agents Improve Global Human Coordination in Network Experiments." *Nature* 545 (7654): 370–74.

Voigt, Rob, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. 2017. "Language from Police Body Camera Footage Shows Racial Disparities in Officer Respect." *Proceedings of the National Academy of Sciences of the United States of America* 114 (25): 6521–26.

James Evans is in the sociology department at the University of Chicago and at the Santa Fe Institute. His research uses large-scale data and computational methods to explore the collective system of thinking and knowing, ranging from the emergence of ideas to the distribution of attention, imagination and habits of reason. **Jacob G. Foster** is in the sociology department at the University of California–Los Angeles. He uses computational methods to study the production of collective intelligence, the evolutionary dynamics of ideas, and the co-construction of culture and cognition.