Computational Social Science: Toward a Collaborative Future*

Hanna Wallach

Microsoft Research & University of Massachusetts Amherst†

This draft: February 21, 2015

Fifteen years ago, as an undergraduate computer science student in the UK, I read a popular science article (Matthews, 1999) profiling the research of my now colleague, Duncan Watts. This article, about the science of small-world networks, changed my life. To understand why, though, it's necessary to know that in the UK, there is (or at least was during the 1980s and 1990s) a profound "them-versus-us" split between the STEM[1] fields and all other disciplines. This split is amplified or perhaps even caused by the fact that people specialize very young—choosing, at age fifteen or sixteen, whether they will ever take another math course or write another essay again. I, like everyone else in my degree program, had chosen STEM, but my decision hadn't been easy—I had also wanted to study the social sciences. The article about Duncan's research changed my life because it had never before occurred to me that that math and computers could be used to study social phenomena. For the first time, I realized that rather than studying *either* computer science *or* the social sciences, perhaps I could study both. This, then, became my motivating goal.

*To appear in "Data Science for Politics, Policy, and Government," R. Michael Alvarez, editor.

†This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grant #IIS-1320219, and in part by NSF grant #SBE-0965436. Any opinions, findings and conclusions, or recommedations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

[1]Science, Technology, Engineering, and Mathematics

Ten years ago, as a PhD student studying machine learning, I wasn't really any closer to my goal. Sure, there was a growing number of researchers studying social networks, but for the most part these researchers were physicists, mathematicians, or computer scientists, or they were social scientists, with little interaction between the groups. Now, though, in 2015, we're on the cusp of a new era. Over the past five years, the nascent field of computational social science has really taken off, with universities and corporations alike creating interdisciplinary computational social science research institutes. This investment has, in part, been fueled an explosion of interest in "big data." Whereas this term used to refer to the massive data sets typically found in physics or biology, the data sets that fall under this new big data umbrella are, for the most part, granular, social data sets—that is, they document the attributes, actions, and interactions of individual people going about their everyday lives (Wallach, 2014). Consequently, research on aggregating and analyzing social data is more important (and better funded) than ever before. And, in turn, researchers have moved beyond the study of small-scale, static snapshots of networks, and onto nuanced, data-driven analyses of the structure, content, and dynamics of large-scale social processes. That said, we're still not quite there yet. Although this research is increasingly taking place in interdisciplinary environments, it's mostly being done by teams of socially minded computer scientists or teams of computationally minded social scientists. As a result, there's often a mismatch between the research being pursued by these teams, and, more importantly, a lack of agreement as to what even constitutes the big questions and scientific goals of computational social science as a field.

My aim for this chapter is therefore to characterize some of the differences between the computational social science research being done in computer science and that being done in the social sciences, as well as some paths for moving forward as a truly interdisciplinary field. While I am a computer scientist by training—specifically a machine learning researcher—most of my work over the past five years has been done in collaboration with social scientists, and thus this chapter constitutes a reflection of my own experiences, rather than an infallible statement of hard-and-fast fact. That said, I hope my observations resonate with and motivate researchers in both fields.

To date, much of the computational social science research coming out of computer science—

especially that appearing in machine learning, data mining, and data science conferences—has been driven by the fact that modern life is increasingly migrating online to new social platforms, often built by computer scientists, including Coursera, Ebay, Facebook, Foursquare, LinkedIn, OKCupid, Stack Overflow, Twitter, and Wikipedia, to name just a few. This research, therefore, is primarily concerned with understanding social behavior as it relates to these platforms—often with an explicit end goal of improving user experience or generating additional revenue. Unsurprisingly, most of this research is driven by partnerships with industry and often involves a tight feedback loop of aggregating user data, analyzing these data in order to reveal new insights about social behavior, and using these insights to inform the next iteration of platform design. This type of research is appealing to computer scientists for several reasons. Not only do companies provide convenient and compelling data sets, but professors often have close ties to industry since most computer science students work at these companies after graduation. In other words, these companies are already an integral part of the broader computer science ecosystem. Moreover, this style of research is familiar—it dovetails with and draws upon a long history of research on human–computer interaction and computer-supported cooperative work. And finally, for decades, computer science was internally focused—that is, focused on the computer itself—with researchers working on hardware, networks, operating systems, compilers, and so on. From a historical perspective, it's therefore unsurprising that socially minded computer scientists often choose to focus on analyzing platforms for social interaction built by other computer scientists.

This focus on improving social platforms is evidently conducive to fast-paced work with immediate real-world impact, but it also gives rise to certain biases. First, and most importantly, much of this research is concerned with modeling social behavior only insofar as it helps with predictive analyses—for example, modeling users' social networks in order to classify users into those who will make a purchase and those who won't, analyzing student discussion boards in order to predict who will drop an online course during the first two weeks, or quantifying users' preferences in order to select news stories for them. In other words, the emphasis is on "what" and "when" rather than "why." More generally, because studying social behavior is a means to an end, usu-

ally driven by short-term motivations rather than long-term, big-picture questions, the resultant findings do not necessarily transfer to other aspects—especially offline aspects—of society. Do we really want our understanding of social behavior to be tied to particular online platforms—with corresponding gaps in our knowledge relating to those aspects of life that take place elsewhere?

Second, it's often the case that this research is undertaken in the context of a single platform. This is understandable given the myriad of technical, legal, and ethical difficulties inherent in combining data sources—even for research purposes—that are the property of different companies, sometimes in direct competition with one another. But, as a result, the constituent analyses, and any methods or tools developed to perform them, are intended for relatively homogeneous data sets, which can limit their applicability outside of this "single-source" setting. Finally, because of the scale of the data sets involved, computational efficiency is often prioritized over making accurate and accountable predictions. While this is a justifiable tradeoff if the ultimate goal is to produce a platform that works well enough for the majority of a massive user base, this approach is not especially conducive to responsible, contextualized findings about social behavior.

In contrast, and as evidenced by the chapters in this book, the recent computational research arising from the social sciences has, for the most part, been fueled by longer-term, bigger-picture goals, usually with an explanatory bent. For example, Grimmer (chapter 6) discusses the analysis of a large collection of House press releases in order to answer long-standing questions about the ways in which legislators use political communication to influence their relationships with constituents. Even researchers whose work involves data arising from online social platforms tend to use these data sets to answer questions extrinsic to the platforms themselves. For instance, Tucker et al. (chapter 2) draw upon millions of geo-located Tweets in order to examine the extent to which Twitter use during political protests reflects and even influences on-the-ground protest participation, as opposed to general international attention. Importantly, in both of these pieces of work—Grimmer's and Tucker et al.'s—the research emphasis is firmly on "why" and "how," with "what" and "when" playing a secondary role. Moreover, Tucker et al.'s work explicitly focuses on the relationship between online and offline social behavior by studying whether and, more

importantly, how behavior on online social platforms can affect behavior in offline contexts.

Although the above two examples involved data arising from single data sources—House press releases and Tweets, respectively—many other questions of substantive interest to social scientists cannot be answered in a "single-source" setting, usually because there is no single platform capable of producing the kind of aggregate data needed to answer these questions. As a result, there is a growing body of computational social science research, driven by social scientists, focused on combining and drawing conclusions from multiple, diverse data sources, often at radically different granularities or scales. Again taking chapters from this book as examples, Warshaw (chapter 1) discusses the potential for using large, commercial public opinion data sets to augment small academic surveys, and the resultant need for new quantitative methods capable of analyzing the temporal dynamics of public opinion at a variety of scales. Meanwhile, Beieler et al. (chapter 4) provide an extensive discussion of the opportunities and obstacles involved in automatically generating—in near-real time—massive political event data sets (for subsequent use by international relations scholars, among others) from news reports arising from a diverse range of international outlets. As evidenced by both of these pieces of work, there are nontrivial computational challenges involved in aggregating and analyzing signals from data sources with varying emphases, granularities, and formats—many of which do not arise when working with even massive quantities of relatively homogeneous data obtained from a single platform or data source.

This focus on accurate and accountable explanations, often involving data from diverse sources, means that existing computer science methods—originally developed to facilitate efficient prediction or even data exploration, often at the expense of interpretability—aren't always immediately applicable. Most notably, social scientists are usually interested in identifying not just whether and, if so, when certain social behaviors occur, but also the conditions that explain those behaviors and any variation in them. As a result, analysis methods developed by social scientists typically provide the ability to incorporate arbitrary observed covariates, so that the effects of those covariates on the behaviors of interest may be quantified. Since it's comparatively rare for computer scientists to prioritize interpretability and explanation in this way, it's often the case that methods

arising from computer science have to be modified to capture the effects of observed covariates before they are adopted by social scientists. As a concrete example, the structural topic model of Roberts et al. (2014) (also the focus of chapter 3) extends latent Dirichlet allocation (Blei et al., 2003)—a statistical topic model developed by computer scientists for the exploratory analysis of large document collections—to allow users to include a wide range of document-level covariates, thereby facilitating nuanced investigation of the ways in which thematic content varies with these covariates. The end result is not only a model better suited to the explanatory goals of social scientists, but a novel contribution, in and of itself, to the statistical topic modeling literature.

Finally, even those methods that are applicable in a supposedly "off-the-shelf" fashion often have "rough edges" that make using them to obtain accountable—even repeatable—findings a complex and time-consuming process. Continuing with the example of statistical topic modeling, Roberts et al. (chapter 3) tackle the issue of non-convexity and the resultant sensitivity of results to different initialization strategies. While this issue has been long-acknowledged in the topic modeling literature, it's seldom addressed head-on, with researchers tending to gloss over its role in producing unrepeatable results. This situation is unsurprising: if one is primarily concerned with feasibility and computational efficiency, then nuanced issues affecting model stability may not be a priority. But, if one is using these methods to explore substantive questions, whose answers have real-world social implications, then any issue that may affect repeatability and accountability matters a great deal. Roberts et al., like many others in the social sciences, have invested significant time into investigating not only when these kinds of rough edges occur and what causes their occurrence, but also effective techniques for addressing or engaging with them. In some ways, this relationship between the computer and social sciences is reminiscent of the symbiosis between the Debian[2] and Ubuntu[3] Linux distributions[4]. Debian developers produce a comprehensive base distribution, primarily used by relatively technical users, such as other developers and system administrators; Ubuntu builds upon this distribution, adding the interface and usability features

[2] http://debian.org/
[3] http://ubuntu.com/
[4] http://www.ubuntu.com/about/about-ubuntu/ubuntu-and-debian/

needed to make it a reliable and easy-to-use operating system. Since many of Ubuntu's developments are contributed back "upstream" to Debian, this relationship has resulted in not only a more accessible distribution (which, in turn, has increased Linux adoption), but also greater awareness of and attention to usability considerations within Debian—a net benefit to both groups.

To reiterate my earlier statement, we are on the cusp of a new era in computational social science. The past five years have seen some transformational changes to both the quantity and nature of research being undertaken by socially minded computer scientists and computationally minded social scientists. Personally, as someone whose long-term research goal is to develop computational methods for studying important social questions, I find these changes, and the foundation that they have laid, extremely exciting. Even more exciting to me, though, is the fact that computational social science is still evolving. As a result, I anticipate that the next few years will witness some even bigger changes, many of which will, at least from my viewpoint, move the field toward a collaborative future, characterized by truly interdisciplinary teams of computer scientists and social scientists and, increasingly, shared scientific goals. First, it's possible that engaging in platform-driven research will become increasingly difficult—at least for researchers in academic positions. Although privacy concerns have always made for a thorny path to navigate, especially for those researchers who do not have close ties to industry, last summer's controversial Facebook experiment highlighted the existence and importance of a range of ethical questions surrounding user experimentation (Grimmelmann, 2014). As long as these questions remain unanswered—and, realistically, they will for the foreseeable future while institutional review boards and other committees concerned with the oversight of human-subjects research rethink their policies, guidelines, and best practices—it's likely that many companies will be much more cautious about publicly engaging in experimentation for research purposes, effectively making it much harder for researchers to study social behavior as it relates to understanding and improving online platforms. At the same time, I expect that many more socially minded computer scientists will start turning their attention to bigger-picture social questions, most likely in collaboration with computationally minded social scientists. Here, the impetus is the outstanding methodological work coming

out of the social sciences (some by contributors to this book) that concretely demonstrates how methods developed by computer scientists can be adapted and used to undertake accountable, substantively important, explanatory analyses. These pieces of work, many of which have been presented at computer science conferences and workshops in addition to social science venues, make it abundantly clear to socially minded computer scientists that there are some truly exciting and challenging opportunities for collaboration that go far beyond "run method X on data set Y."

So what can we do to ensure that this future of computational social science—consisting of interdisciplinary teams of researchers contributing equally to common, mutually beneficial research goals—actually materializes? Unfortunately, I don't think there are any "quick fix" answers to this question—if there were, we'd likely have put them into practice already. That said, I do think there are a few things we can do that, taken together, may increase our chances of achieving this future.

First, and most importantly, we need to understand each other's disciplinary norms, incentive structures, and research goals. Although this sounds simple enough, it's not—mostly because it's time-consuming and we're all busy people who tend to prioritize sharing our own opinions (this chapter included!) over listening. But, by taking the time to engage in conversations or to read papers outside of our own disciplines, we're also taking the first steps toward innovative and genuinely interdisciplinary research ideas that no discipline could have produced in isolation. One way to jump-start this process is to attend each other's conferences. Small, interdisciplinary workshops, co-located with larger, disciplinary meetings, are a particularly effective way to facilitate this—not least because people can more easily justify the cost of attending. Personally, I've attended APSA[5] as a result of such a workshop[6], and have also persuaded political scientists, sociologists, and economists to attend NIPS[7] (one of the leading machine learning conferences) by organizing two co-located workshops on computational social science, one in 2010[8] and one in 2011[9]. Of course, stand-alone interdisciplinary events, such as the now annual "New Directions

[5] http://community.apsanet.org/annualmeeting/home/

[6] http://poliinformatics.org/index.php/the-challenge/scheduled-research-challenges/apsa-political-science-challenge/

[7] http://nips.cc/

[8] http://www.cs.umass.edu/~wallach/workshops/nips2010css/

[9] http://www.cs.umass.edu/~wallach/workshops/nips2011css/

in Text as Data" conference[10], are another great way to bring together researchers with potentially overlapping interests. Again using my own experiences to highlight the benefits of attending such events, I met two of my political science collaborators by attending the "New Directions" conference, in addition to many other political scientists, whose papers I now make sure to read. Naturally, since organizing a conference is potentially expensive (especially if participants are provided with travel funding), this is a great example of an area in which funding agencies and companies can help by covering either direct or indirect costs. Finally, even single-university seminar series can catalyze ongoing conversations and collaborations by bringing in external speakers whose research interests appeal broadly to both socially minded computer scientists and computationally minded social scientists. My own network of contacts in the social sciences exists, in large part, because of the seminar series run by the Computational Social Science Institute at the University of Massachusetts Amherst[11], plus numerous invited talks I've given at other institutions.

Ironically, one of the biggest obstacles to producing truly interdisciplinary research is the need—shared by all researchers, regardless of discipline—to publish in high-quality venues in a timely fashion. Unlike within-discipline collaborations, interdisciplinary collaborations are seldom "force multipliers" from a perspective of publishing quickly—mostly because of the time that must be spent defining shared research goals and establishing a common language for communicating efficiently about them, before any actual research can even take place. As a result, bringing an interdisciplinary project to publication can involve a much bigger time investment than that of a disciplinary project. Moreover, even when work is ready to be published, it's not always obvious where to publish it, as "standard" disciplinary venues may not be beneficial to all contributors, let alone appropriate for the work itself. A common strategy is therefore dual publication in a computer science conference and a social science journal, but this approach demands an even greater time investment. Unfortunately, these challenges are not always recognized by tenure and promotion committees, effectively disincentivizing researchers from pursuing this kind of work. Moving forward, it's therefore crucial that, at least in the short term, we—computational

[10] 2014 conference URL: `http://www.kellogg.northwestern.edu/research/ford-center/events/conferences/text-as-data-conf-2014.aspx`

[11] `http://cssi.umass.edu/`

social science researchers—explicitly manage expectations by acknowledging and articulating to others the fact that publishing interdisciplinary research can be slower than publishing single-discipline research. In turn, any academic institution wishing to support and encourage interdisciplinary researchers must also acknowledge these issues when considering promotion and tenure cases. Longer term, we also need better publication strategies than dual publication in disciplinary venues. The most obvious, albeit nontrivial, way to address this need is to create new, high-quality publication venues, explicitly focused on interdisciplinary computational social science research.

Finally, I want to conclude by noting that the best way to ensure the long-term success of computational social science as an genuinely interdisciplinary field, characterized by a set of unifying social questions and scientific goals, is to think carefully about the next generation of computational social scientists and their educational trajectories. With some serious thought and resource investment, undertaken now, we can ensure that unlike the current generation—people like me who had to choose between computer science and the social sciences—this new generation will consist of people with training in both areas: people who therefore possess a deep understanding of the norms of multiple disciplinary communities, and have been part of successful interdisciplinary collaborations long before they even graduate. For this to be possible though, academic departments—likely in different colleges or schools—will need to work together to create new educational opportunities. At the very least, students should be actively encouraged to enroll in dual degree programs, in which they produce a single, interdisciplinary dissertation, while satisfying the course requirements of two departments. Of course, much like the dual publication strategy mentioned above, dual degrees are time-consuming, and not all departments are willing to bear this hit to their "time-to-graduation" records, let alone the cost of supporting a student for the additional duration. As a result, dual degree enrollments are currently the exception rather than the rule, with faculty fighting for their students' rights to pursue such programs on a case-by-case basis. A better, and more sustainable, option is therefore the creation of new, interdisciplinary degree programs, devoted to training the next generation of computational social scientists. While this option constitutes a much bigger change, requiring significant institutional investment, both

in terms of financial and strategic support, the long-term benefit to society—namely, the success of computational social science as an innovative, interdisciplinary field, dedicated to collaboratively answering some of society's biggest questions—seems, to me at least, well worth it.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Grimmelmann, J. (2014). The Facebook emotional manipulation study: Sources. `http://laboratorium.net/archive/2014/06/30/the_facebook_emotional_manipulation_study_source`.

Matthews, R. E. (1999). Get connected. *New Scientist*, (2215).

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58:1064–1082.

Wallach, H. (2014). Big data, machine learning, and the social sciences: Fairness, accountability, and transparency. `https://medium.com/@hannawallach/big-data-machine-learning-and-the-social-sciences-927a8e20460d`.