A SAGE White Paper

Who Is Doing Computational Social Science?

Trends in Big Data Research

Katie Metzler
Publisher for SAGE Research Methods, SAGE Publishing

David A. Kim
Stanford University, Department of Emergency Medicine

Nick Allum
Professor of Sociology and Research Methodology, University of Essex

Angella Denman
University of Essex

www.sagepublishing.com

SAGE Publishing

Contents

Suggested Citation: Metzler, K., Kim, D. A., Allum, N., & Denman, A. (2016). *Who is doing computational social science? Trends in big data research* (White paper). London, UK: SAGE Publishing. doi: 10.4135/wp160926. Retrieved from https://us.sagepub.com/sites/default/files/CompSocSci.pdf

Overview

Information of all kinds is now being produced, collected, and analyzed at unprecedented speed, breadth, depth, and scale. The capacity to collect and analyze massive data sets has already transformed fields such as biology, astronomy, and physics, but the social sciences have been comparatively slower to adapt, and the path forward is less certain. For many, the big data revolution promises to ask, and answer, fundamental questions about individuals and collectives, but large data sets alone will not solve major social or scientific problems. New paradigms being developed by the emerging field of "computational social science" will be needed not only for research methodology, but also for study design and interpretation, cross-disciplinary collaboration, data curation and dissemination, visualization, replication, and research ethics (Lazer et al., 2009).

SAGE Publishing conducted a survey with social scientists around the world to learn more about researchers engaged in big data research and the challenges they face, as well as the barriers to entry for those looking to engage in this kind of research in the future. We were also interested in the challenges of teaching computational social science methods to students. The survey was fully completed by 9412 respondents, indicating strong interest in this topic among our social science contacts. Of respondents, 33 percent had been involved in big data research of some kind and, of those who have not yet engaged in big data research, 49 percent (3057 respondents) said that they are either "definitely planning on doing so in the future" or "might do so in the future."

What Have We Learned About Those Doing Big Data Research?

Of the 33 percent of our respondents who have been involved in big data research, 60 percent have done so recently, within the last 12 months, and 23 percent (744 respondents) said that all or most of their research involved big data or data science methods. Our survey shows that early career researchers are no more likely to have done big data research than respondents who had had their PhDs for over 10 years.

We asked researchers which data sources they used in their last big data research project and found that 55 percent (1690 respondents) had used administrative data, the most common data type, followed by 29 percent (927 respondents) having used some kind of social media data and 23 percent (697 respondents) having used commercial data in their research. One of the biggest problems cited by researchers doing big data research was getting access to commercial or proprietary data, suggesting that more needs to be done to unlock data sets for social science research.

A characteristic of researchers doing big data research is that they are more likely to collaborate with other academics (79 percent of big data researchers in our survey). Considering that a large number of social science papers are single authored (about 40 percent, according to Thomson Reuters (King, 2013), this information is significant. The top three disciplines of collaborators were social and behavioral science, biological and medical science, and computer science. These interdisciplinary collaborations may be influencing the nature of funding sources and publication outlets sought: our survey respondents named science-funding bodies in addition to social science funders, and research results are being published in science, technical, and medical (STM) publications, as well as traditional social science journals. A trend seen in STM is for big data researchers to share their code or software openly via GitHub; however, only 54 respondents to our survey said that they shared code this way, suggesting that social science may be slower to adopt this practice.

What Have We Learned About Those Who Want to Engage in Big Data Research in the Future?

Of respondents, 49 percent (3057 respondents) not currently doing big data research said that they are either "definitely planning on doing so in the future" or "might do so in the future." This response

suggests that there is an appetite to engage with big data research but that there are barriers to entry. Our survey respondents listed finding collaborators with the right skills and the amount of time required to learn a new field as the biggest barriers to entry.

To overcome their skills gap, 40 percent of respondents (3750 respondents) would like to attend big data training in the future. Most respondents would like to undertake basic introductory training on big data analytics or data science, although many other respondents also listed specific topics, such as text mining and R and Python programming. A large number of those who had already carried out big data training in the last 12 months had done so via massive open online courses (MOOCs) and online courses.

What Have We Learned About Those Teaching Research Methods?

Forty-three percent (4026) of respondents are currently teaching research methods or statistics. Of those, 31 percent cover big data analytics or data science methods in their research methods or statistics course. The biggest problems for educators trying to teach big data methods to students are that students do not have the appropriate level of programming knowledge or the appropriate level of statistical knowledge and that there is a limited amount of time available in the methods syllabus to overcome students' lack of existing knowledge.

Methodology

After internal and external pretesting, the survey was deployed in two stages—an initial deployment to 10,000 contacts and a subsequent deployment to 543,819 social science contacts. The completion rate was higher for those that said they have not been involved in big data research: 75 percent of those who said yes to having been involved in big data research reached the end of the survey while 93 percent of those who said that they had not been involved in big data research reached the end.

Although the survey was pretested, from the responses given to free-text answers, a number of respondents did not seem to understand the screener question regarding big data and said "yes" despite not having done big data research. A number of these respondents' responses were recoded as "no" during analysis when it was possible to determine from other item responses that they had misunderstood the question. The definition of big data given was probably not specific enough as it did not specify how big data has to be to be included in our definition (e.g., more than a terabyte). However, by including an arbitrary cutoff point in terms of size, we would have introduced other problems as those doing research with very large data sets under the specified size may have had useful feedback to share that would have been missed. "Data science" is also a problematic term because some people would consider all methods to fit under the umbrella term of data science, while we had a more specific meaning in mind denoting big data analytics.

Analysis

Of the respondents who opened the survey link, **9412** reached the end of the survey and have been included in the analysis. The respondents represented a range of social science disciplines, with a majority from education, psychology, and health sciences (see Figure 1); 84 percent of the respondents were based in a university or college (see Table 1), and 8 percent were graduate students. The great majority (75 percent) were employed full-time (see Table 2).

Figure 1 Primary discipline—all respondents

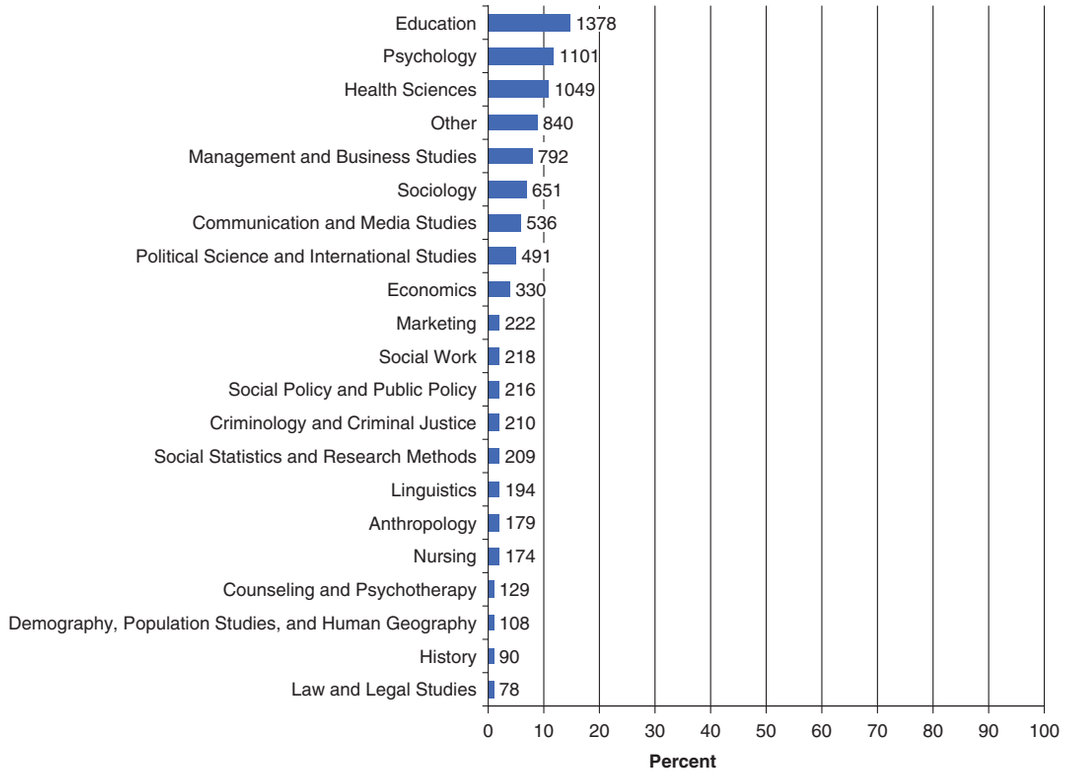


Table 1 Sector—all respondents

Sector	*N*	%
University or college	7933	84
Government	527	6
Nonprofit	341	4
Business or industry	301	3
Other	280	3

Table 2 Employment status—all respondents

Employment	*N*	%
Full-time	7005	75
Part-time	764	8
Self-employed	287	3
Graduate student	842	8
Retired	319	3
Other	171	2

The survey was sent out to a global list of social science contacts. Table 3 shows the number of responses compared with the number of invitations sent out to contacts in each of these countries and the response rate by country (which does not account for undelivered emails). The majority of the respondents were from the United States (3302 respondents) and the United Kingdom (728 respondents), with a large number of Indian and Canadian respondents also completing the survey. The response rates were in the 1 to 2 percent range. (See Figure 2.)

Figure 2 Number of respondents per country

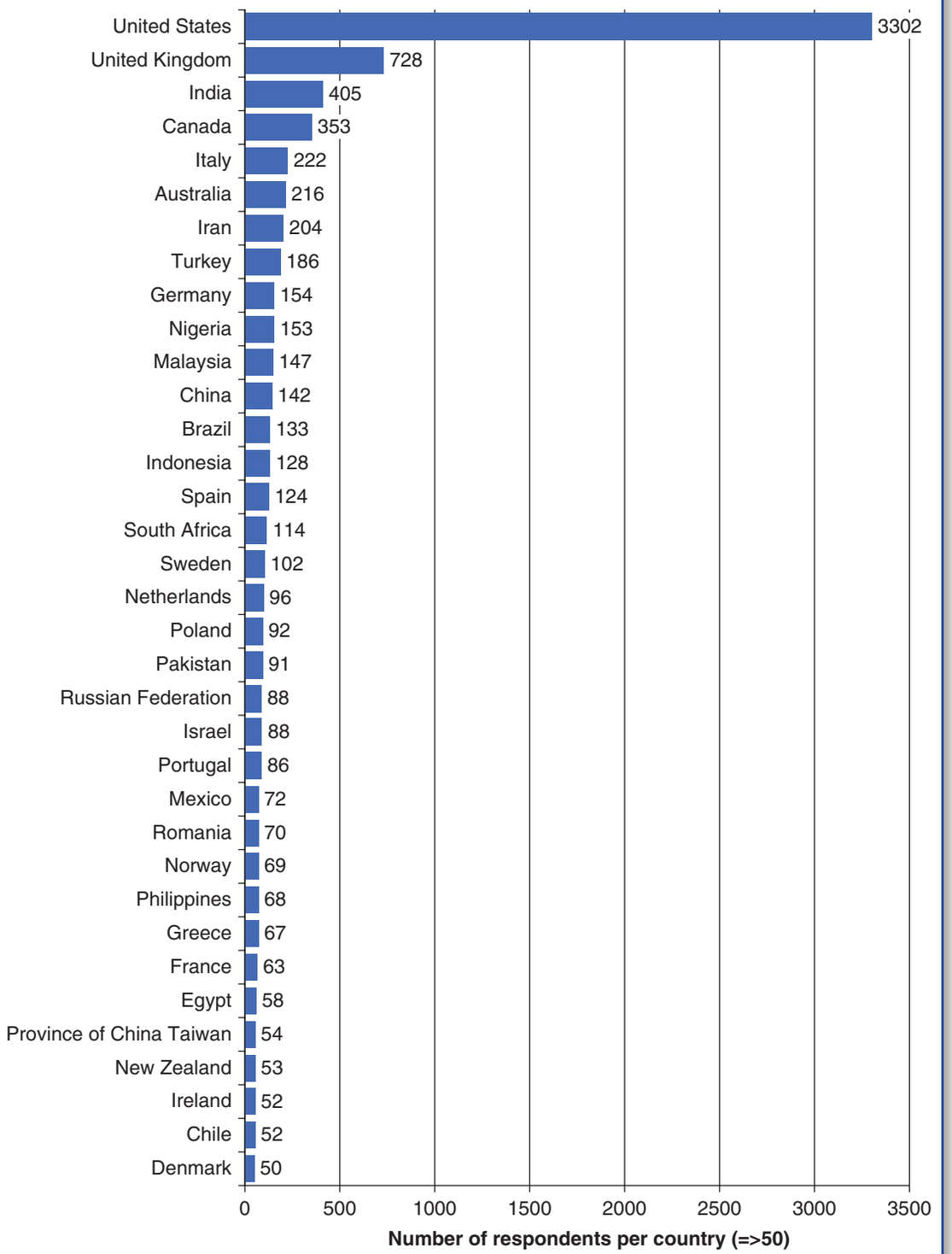


Table 3 Response rates by country

	Completed Survey	Invitation Issued	Response Rate
United States	3316	280,854	1.2%
United Kingdom	728	72,586	1.0%
India	405	20,089	2.0%
Canada	353	18,566	1.9%

The screener question gave the following definitions of big data and data science and asked respondents whether they had ever been involved in research of this kind:

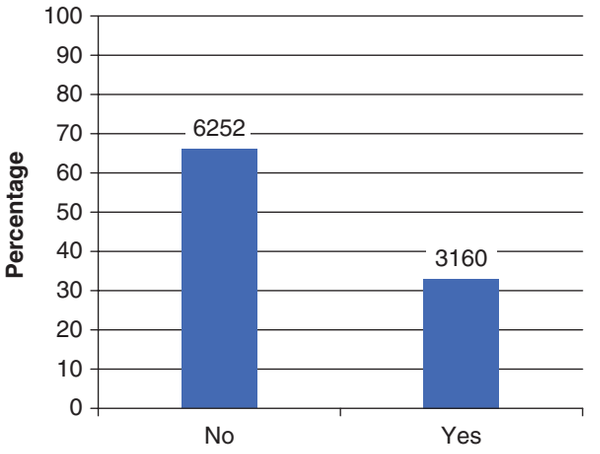
Research involving "big data" is becoming more common. By big data, we mean data sets that are too large and complex to be analyzed using traditional software and methods. Examples of these data include social media data, data generated from online transactions, administrative data, mobile phone data, and audio, visual, text, and sociometric sensor data. These data sets have given rise to new methods and analytic tools, evolving from the interdisciplinary fields of social science, statistics, computer science and design, that are sometimes collectively referred to as "data science" or "big data analytics."

Essex University with partners SAGE Publishing are conducting this survey in order to find out about your interest in and experience of big data and data science. Even if you are not involved in this type of research, we would still like to hear your views.

First of all, then, what about you? Have you ever been involved in any research using big data or data science methods?

Among respondents, 3160 (33 percent) reported that they had been involved in research using big data and the remaining 66 percent said that they had not; see Figure 3. We expect nonresponse bias to be present here as those doing big data research were probably more inclined to complete the survey than were those with no interest in big data, so this cannot be taken as representative of the larger social science population.

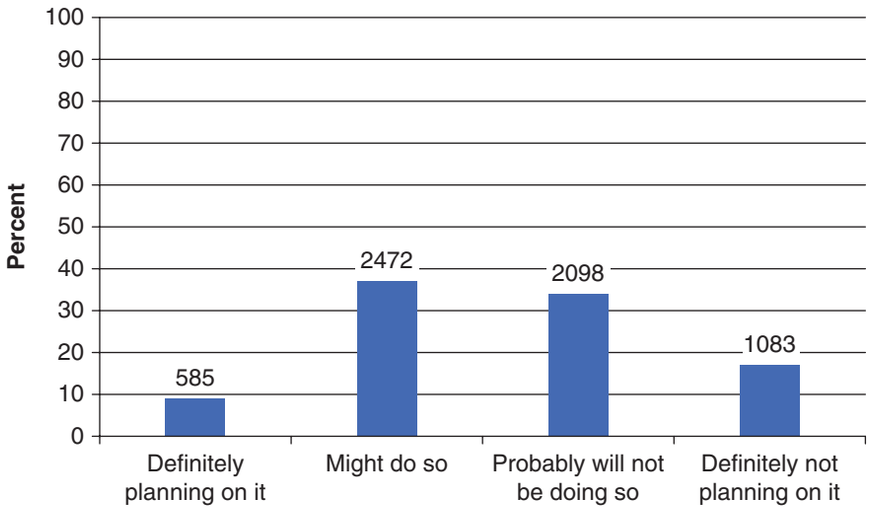
Figure 3 Percentage of respondents who have been involved in research using big data (*n* = 9412)



Of the four countries with the highest number of respondents (United States, United Kingdom, India, and Canada), India had the highest proportion of respondents who answered "yes" to having been involved in big data research (45 percent). 33 percent of U.S. respondents said "yes" whereas 24 percent of Canadian and 23 percent of U.K. respondents answered "yes."

All those who responded saying they had not been involved in big data research to date were asked a follow-up question about whether they intended to do big data research in the future, to which 3057 respondents (49 percent) said they were "definitely planning on it" or "might do so" in the future (see Figure 4).

Figure 4 Percentage of respondents planning on doing big data in the future (*n* = 6238)



In total, 744 respondents said that all or most of their research involved big data (see Figure 5).

Figure 5 Amount of respondent's research in the last five years that has involved big data (*n* = 3128)

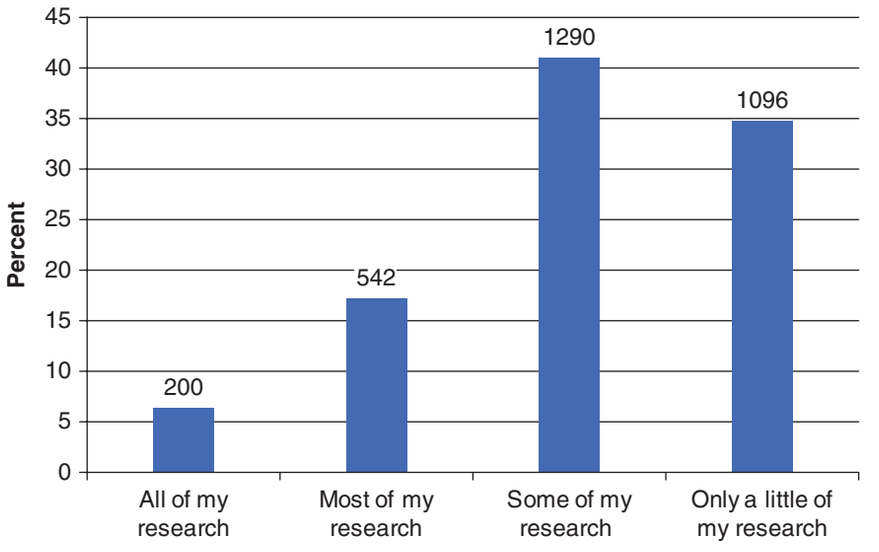
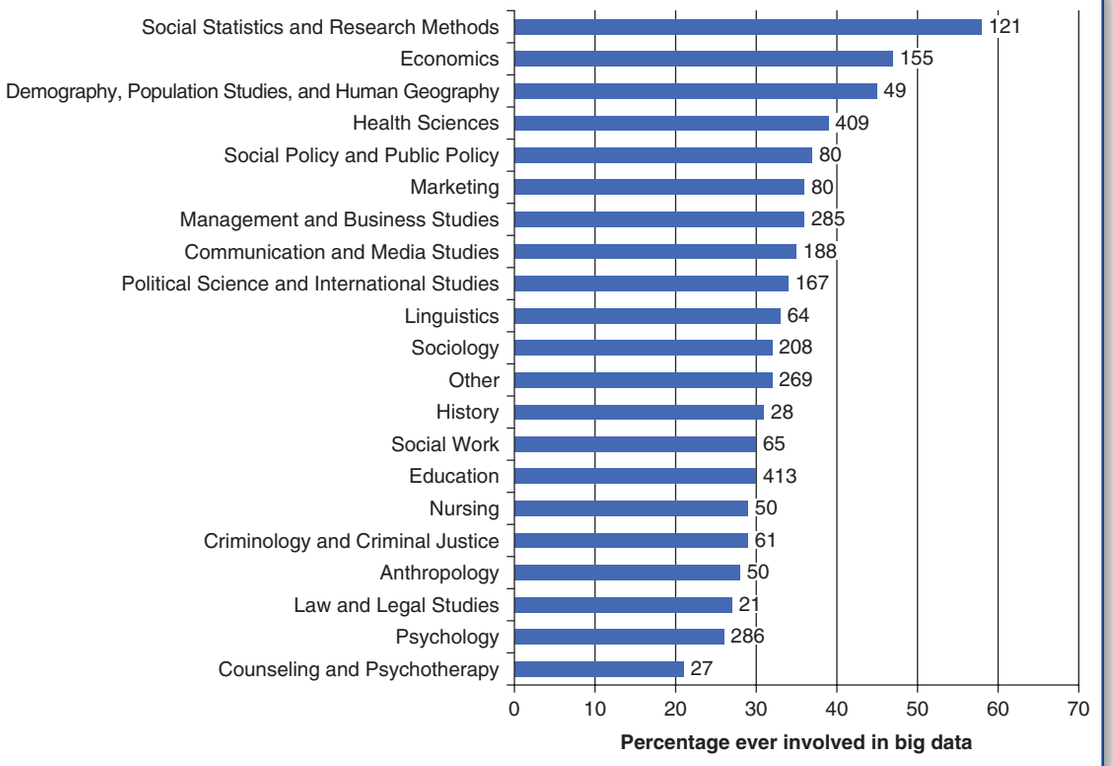


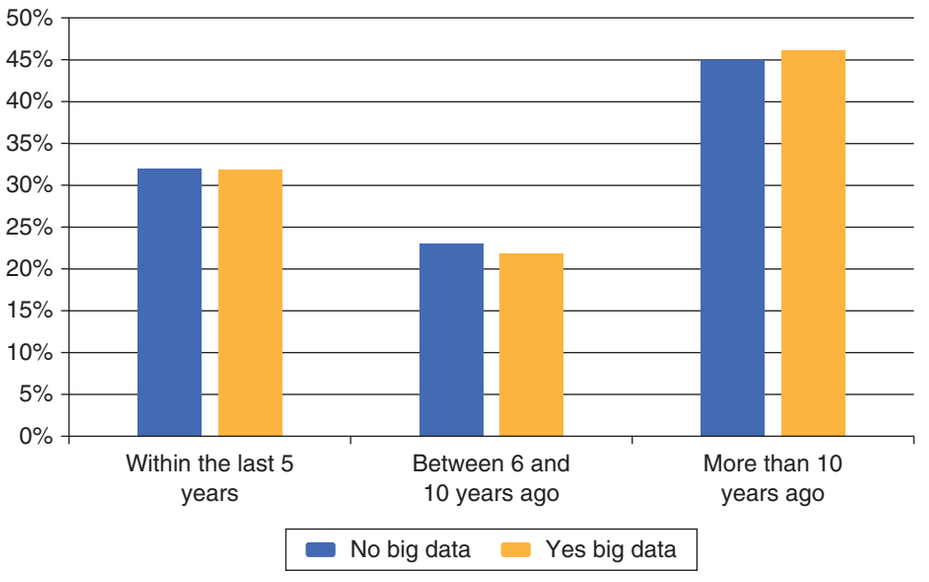
Figure 6 shows the prevalence of big data research by primary discipline (the variation in the raw numbers shown in Figure 6 reflects the varying proportion of respondents from each discipline in the sample). Of the social statistics and research methods, 60 percent of respondents said that they had been involved in big data research, and 21 percent of the counseling and psychotherapy respondents said they'd been involved in big data research. Overall, these percentages seem very high (especially in the case of history and anthropology, which are not typically disciplines associated with big data), and this further suggests that researchers who are very interested in big data and who are already engaged in big data research were more likely to complete the survey. It may also indicate ambiguity about what people understand by the terms *big data* and *data science*.

Figure 6 Primary discipline of respondents who have been involved in big data research (*n* = 9195)



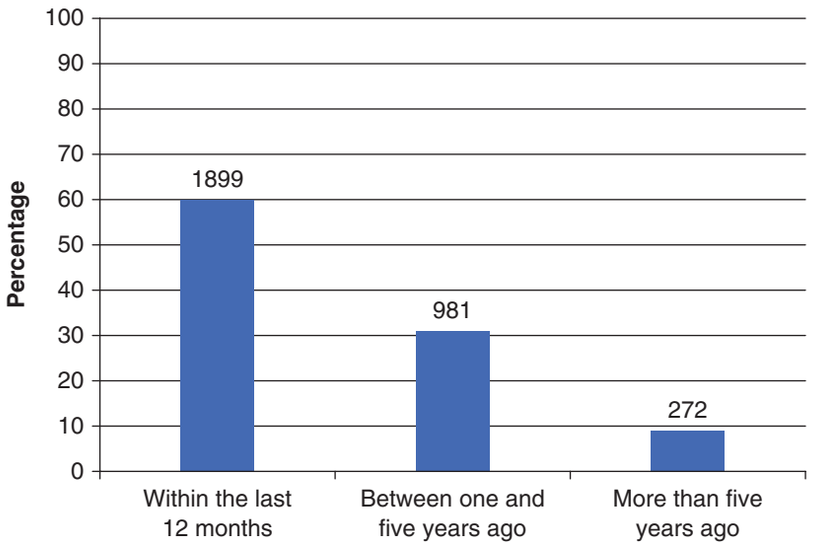
Our hypothesis was that big data research was more likely to be carried out by early-career researchers, as it's an emerging field and often these developments are led by early-career researchers. In fact, there is no difference by career stage of those doing big data and not doing big data research among our sample (see Figure 7).

Figure 7 Career stage (time since PhD) by involvement with big data research (*n* = 6200)



Our hypothesis was that researchers engaging in big data research are likely to have done so recently and this has been supported by the survey that found that 60 percent of those doing research involving big data had done so in the last 12 months (see Figure 8). However, we did not ask respondents to tell us how long ago they began doing big data research, which would have been helpful in determining the pace of growth of the field.

Figure 8 When respondent was involved in big data research (*n* = 3152)



In total, 985 respondents said their university had an interdisciplinary big data lab or center (more than 3000 respondents said they were not sure), and 281 respondents said they were affiliated with the lab or center (see Figure 9 for a selection of big data labs and centers listed in the survey).

Figure 9 A selection of big data labs and centers named by respondents

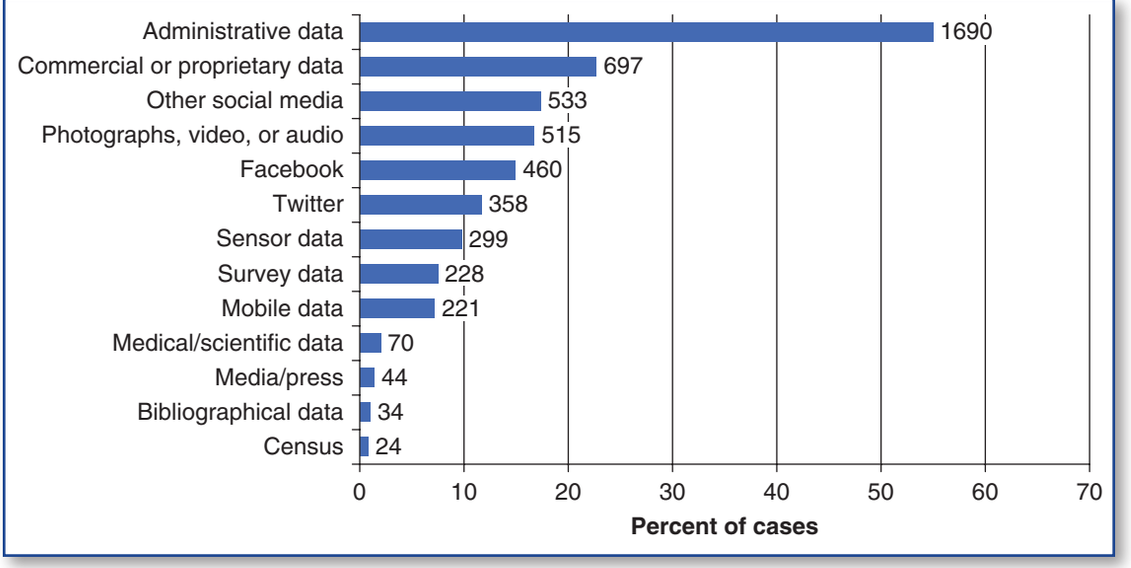
Big Data Consulting Services and Training Center, University of Georgia
Big Data Decision Analytics Research Centre, City University of Hong Kong
Big Data Institute (BDI), Oxford University
Cambridge Big Data, Cambridge University
Center for Customer Analytics and Big Data, Washington University in Saint Louis
Center for Data Science, University of Massachusetts Amherst
Center for Data Science and Big Data Analysis, Oakland University
Center for Human Dynamics in the Mobile Age, San Diego State University
Center for Internet Research, University of Haifa
Centre for Big Data Research in Health, University of Sydney
Centre for Smart Data Technologies, Robert Gordon University
Data Science Center TiU, Tilburg University
Data Science Institute, Columbia University
Delft Data Science, Technische Universiteit Delft
MIDAS, University of Michigan
Social Dynamics Lab, Cornell University
Supercomputer Center, University of California San Diego
Urban Big Data Centre (UBDC), Glasgow
Warwick Data Science Institute, Warwick
Web Science Institute, University of Southampton

Figure 10 presents the different types of data sources big data researchers have used. Respondents could select multiple answers for this question and options are not entirely mutually exclusive (e.g., Twitter is also commercial or proprietary data). Administrative data were the most widely used: 1690 respondents (55 percent) used this type of data in their most recent research involving big data. Administrative data includes data collected by government departments and can include health, educational, and income data.

Twenty-nine percent (927 respondents) have done research using some kind of social media data (including Facebook, Twitter, and other social media). In China, where Facebook and Twitter are banned, we unsurprisingly see a larger proportion of researchers choosing "other social media" which includes Weibo, Baidu, and WeChat.

The third most commonly used data type was commercial or proprietary data with 697 respondents (23 percent).

Figure 10 Data types used by respondents in most recent research involving big data (*n* = 3077)



One of the challenges researchers face when carrying out big data research can be that the data sets are so large that they require a distributed computing infrastructure. These systems are components of a software system shared among multiple computers to improve efficiency and performance. Figure 11 shows the respondents who answered "yes" to using one of the named distributed computing solutions given in the survey. Hadoop was the most commonly used, followed by subproducts within the Hadoop ecosystem: MapReduce and Spark.

An analysis of the free text answers given for "other distributed computing" suggested that there was confusion among respondents as to what counted as a distributed computing environment. Many respondents answered this question and the following questions regarding software in the same way, and so in order to get a clearer picture of the data, a variable was created that merged the free text software responses. Although 579 researchers answered with software that is used for big data research, 1248 respondents used traditional software (SPSS and STATA) for their research. While SPSS and STATA have both been enhanced to handle larger data sets, there is also a possibility that respondents who answered naming a traditional software package were either not working with very large data sets or were working with smaller subsets of a large data set, which is common among researchers in the social sciences engaging with social media data. Big data software or programming languages mentioned by the respondents include Python, R, PostgreSQL, SAS, Netezza, and Google Big Query (see Figure 12).

Figure 11 Respondents who used a distributed computing solution named in the survey (*n* = 238)

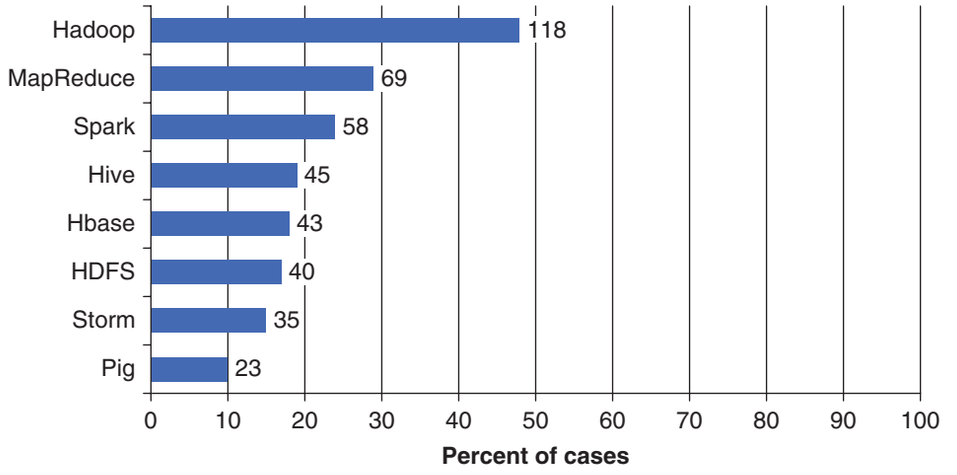


Figure 12 Big data software used

Sap Hana	R
Netezza	PostgreSQL
Google BigQuery	FORTRAN
GIS (geographic information system)	Mathematica
IBM Jam	Crimson Hexagon Forsight
Python	Netlytic
Galaxy (Computational Biology)	Cosmos (C# Open Source Managed Operating System)
Epi Infor.	KNIME
SAS	FSL
Pentaho	GAUSS
AWS Redshift	Artificial neural networks
The Issue Crawler	Pajek
NetViz	REDCap
EC2- Amazon Elastic Compute Cloud (Amazon EC2)	S-plus
SQL (structured query language)	ProM
Oracle Grid Engine	Talend
Mat labs (Matrix laboratory)	Statistica
Pulsar	Weka
ArcGIS	MaxQuant

We also asked researchers whether they had shared the code or the software they developed with other researchers (see Figure 13). Only 56 respondents had shared their code on GitHub, which is surprisingly low. The majority of researchers did not share anything. Those who said that all or most of their research involved big data were more likely to share code or software via e-mail (see Figure 14), but the majority still reported not sharing anything. Other ways researchers shared code were on request from individuals, internally, through publication of books or journals, at conferences, and one respondent used the code sharing platform, Sourceforge.

Figure 13 Sharing of code or software (*n* = 873)

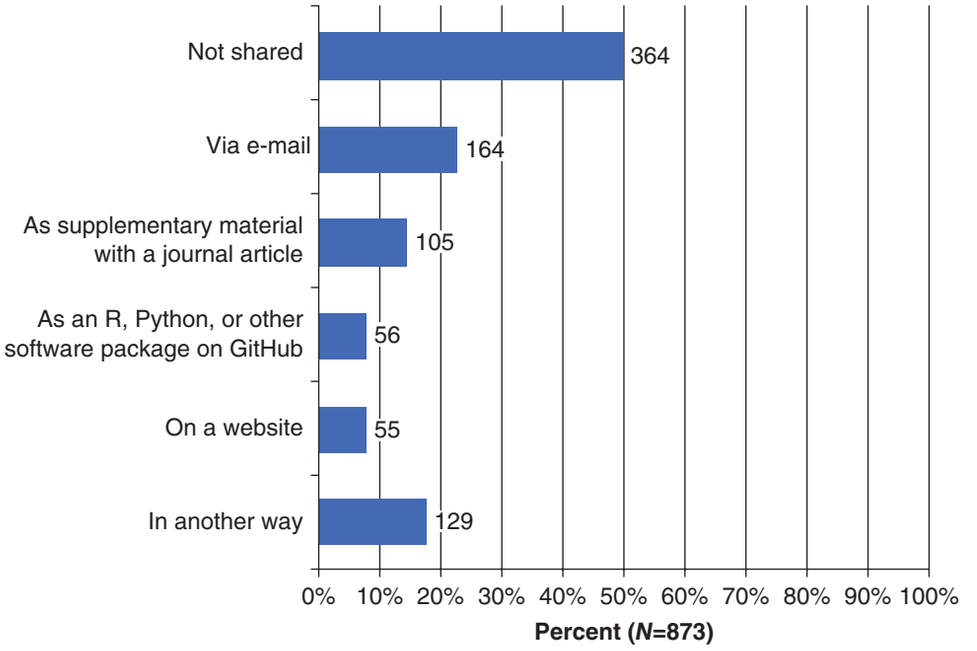
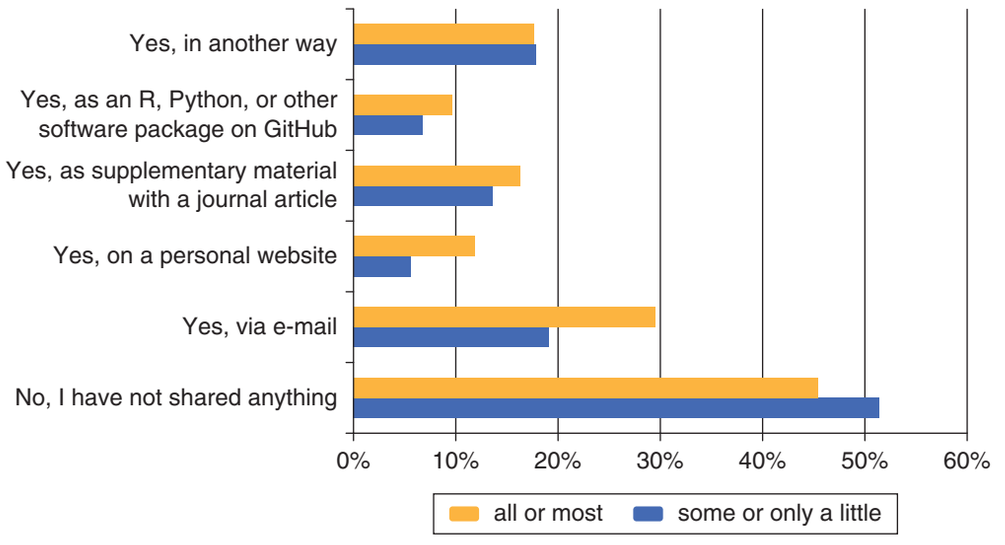


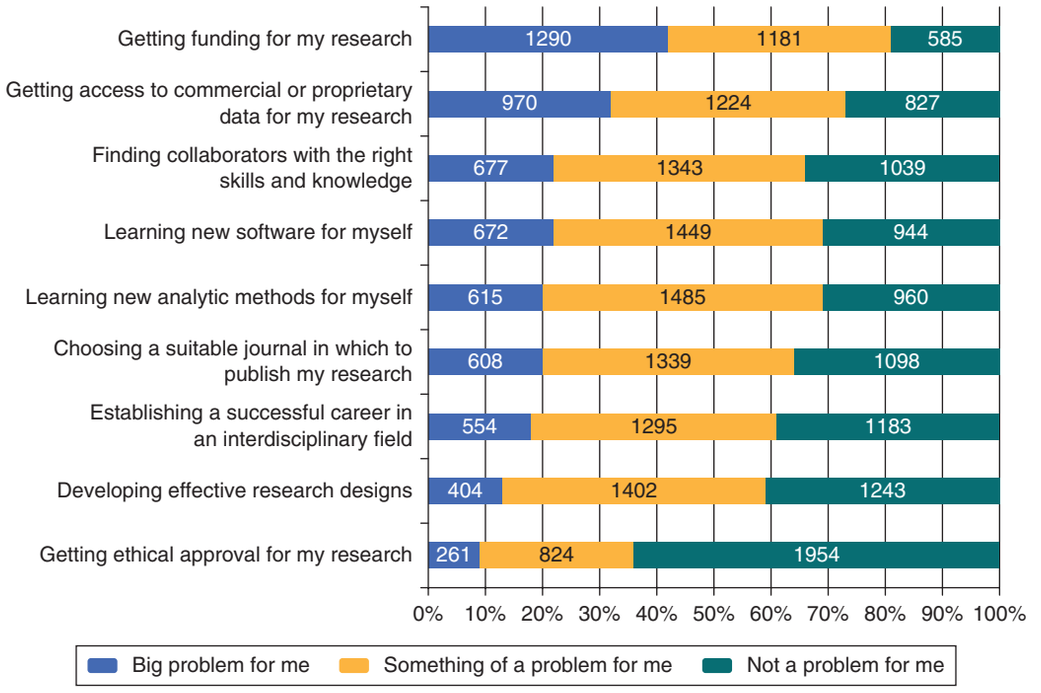
Figure 14 Sharing of software and code by amount of research using big data (*n* = 3056)



Challenges Facing Big Data Researchers in the Social Sciences

One of our hypotheses when designing the survey was that big data researchers face unique problems, in part due to the interdisciplinary nature of the field, and also as a result of its relative newness in the social sciences. Figure 15 presents a number of challenges faced by researchers who use big data. Of the respondents, 42 percent felt that getting funding was a "big problem" for them; however, we did not ask this question of non–big data researchers, and therefore, we do not know if this problem is specific to or more pronounced for big data researchers over other social science researchers: 32 percent said that getting access to commercial or proprietary data was a "big problem."

Figure 15 Challenges facing big data researchers (*n* = 2273)

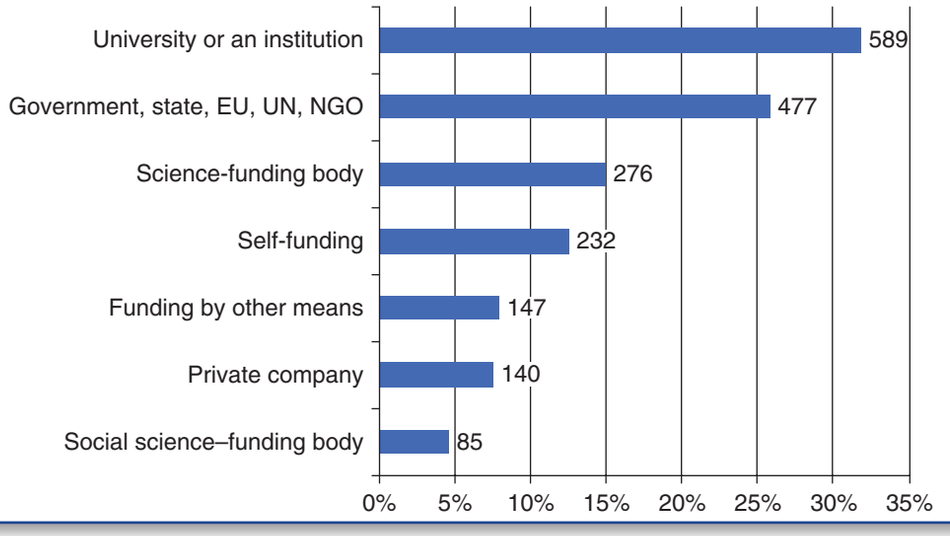


Other challenges mentioned were the following:

- Lack of time
- Lack of computing infrastructure required
- Challenges associated with working in interdisciplinary teams
- Concerns about data quality

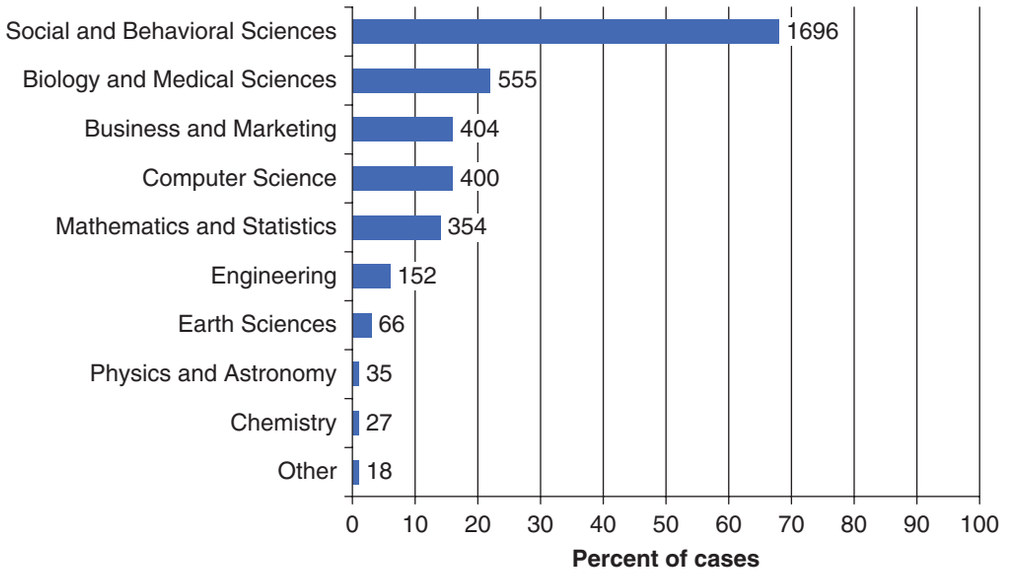
Big data researchers are currently being funded from a range of diverse sources, with the majority naming university or institutional funding as their main source, followed by government funding (see Figure 16), 15 percent naming a science-funding body, and fewer than 5 percent naming a social science–funding body.

Figure 16 Sources of research funding (*n* = 1946)



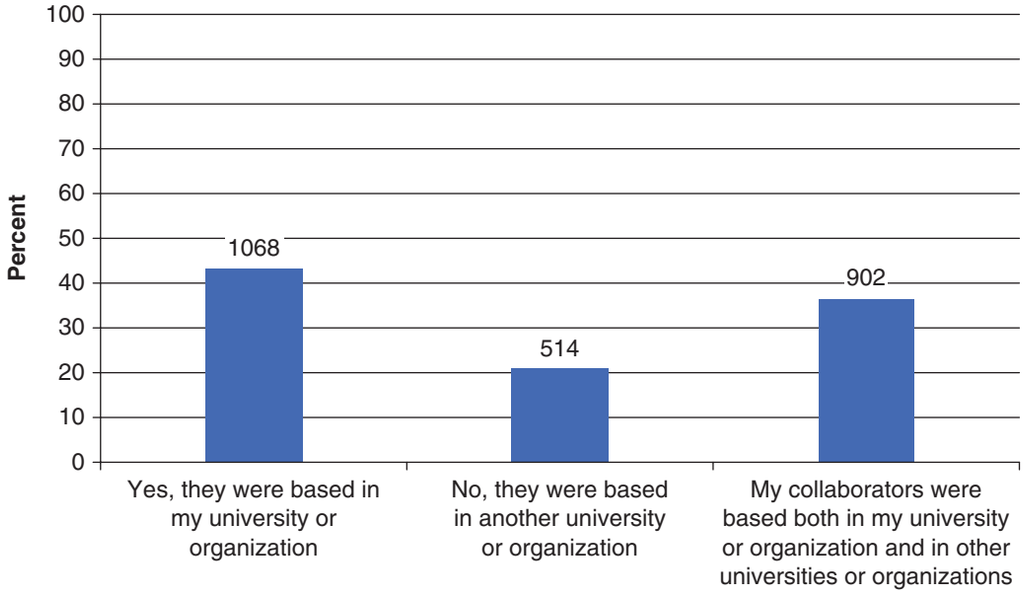
Sixty-six percent cited “finding collaborators with the right skills and knowledge” as ranging from “something of a problem” to a “big problem” (Figure 17). Those who have engaged in big data and worked with collaborators partnered with academics from the social and behavioral sciences primarily, biology and medical sciences, business and marketing, and computer science (16 percent).

Figure 17 Primary academic field of respondent’s big data research collaborators (*n* = 3707)



Of collaborators, 43 percent were based at the same university or institution as the survey respondent and 21 percent were based at another university or institution and 36 percent said they collaborated with those both inside and outside of their organization (Figure 18).

Figure 18 Where collaborators were based (*n* = 2484)

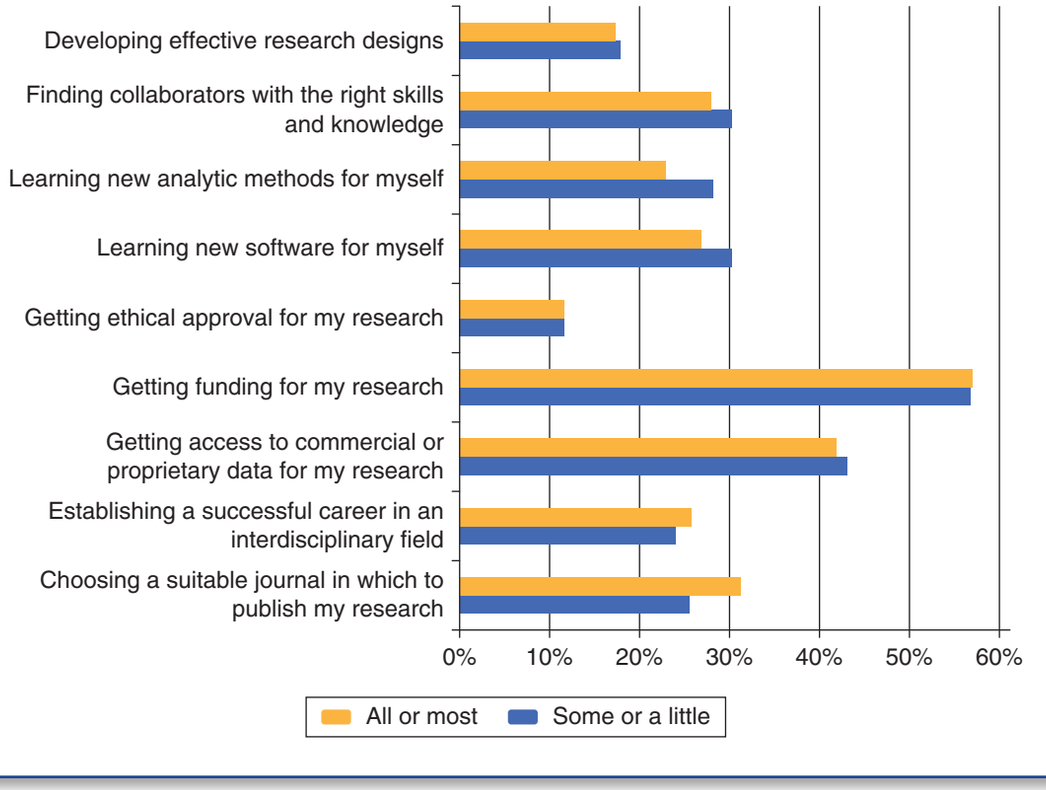


We were interested to know whether getting published posed challenges for big data researchers: 61 percent said "choosing an appropriate journal" was a "big problem" or "something of a problem" (Figure 19). Again, without a comparative question for non–big data researchers, we cannot say whether this is more of a problem for big data researchers, although our hypothesis is that it is because of the interdisciplinary nature of the field. Quotes from free-text answers related to this included the following:

> I would like to emphasize the difficulty in finding journals that are interested and willing to publish interdisciplinary research.
>
> Several of the top journals in business school disciplines have not yet embraced Big Data Analytics.

Interestingly, those who reported that most or all of their research was big data were more likely to say that "choosing a suitable journal" was a problem for them compared to those whose research is less focused on big data.

Figure 19 Problems encountered by amount of research using big data (*n* = 2266)



Of respondents who carried out research using big data, 48 percent have had their work published in a journal. The journals are wide ranging and include medical, social science, science, and methods journals, but few journals dedicated to publishing computational social science research. The following are a selection of journals mentioned by three or more respondents:

- *PLOS One* (13)
- *BMJ* and *BMJ Open* (7)
- *Urban Studies* (7)
- *JAMA* (5)
- *New Media and Society* (4)
- *Big Data and Society* (3)

- *International Journal of Humanities and Social Sciences* (3)
- *SAGE Open* (3)
- *Party Politics* (3)

Of the respondents who carried out research using big data, 33 percent presented a paper or a poster on their research or data science methods. The conferences named varied from all of the large U.S. society conferences (American Educational Research Association, American Sociological Association) to more specialized conferences such as Social Media and Society. Very few conferences named were big data specific, suggesting that researchers are presenting their research at established discipline conferences.

In the last 12 months, 12 percent (1133 respondents) had attended training on big data. The training reported included sessions at conferences, short courses, and courses run at the university and MOOCs. The MOOCs named in the survey included the following:

- Coursera (50)
- Edx (12)
- Future Learn (4)
- Udacity (2)
- Udemy (1)

An additional 17 respondents said they'd completed a MOOC but did not name the provider, and 24 said that they'd done an online course, which may also mean MOOCs. The following were popular topics for training:

- Text Mining
- Data Mining
- Social Network Analysis
- R
- Python
- Big Data Analytics

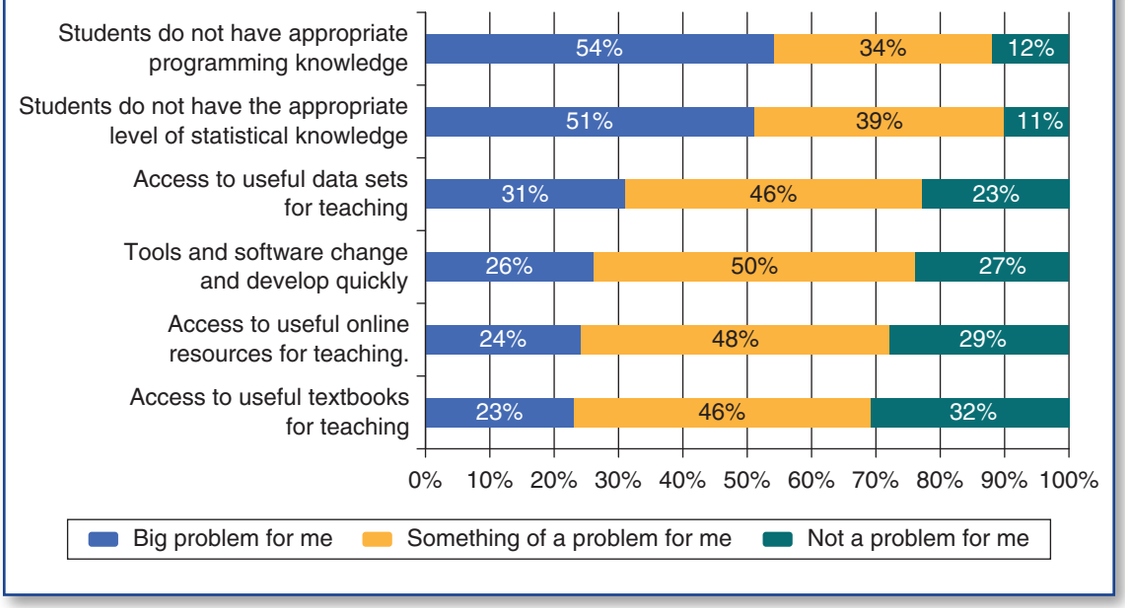
In the future, 40 percent (3750 respondents) would like to attend big data training. A large number of respondents requested introductory training on big data analytics. Other training requested included the following:

- Assessing quality of big data sets
- Analyzing social media data
- R
- SQL
- Data visualization
- Biostatistics and bioinformatics
- Corpus linguistics
- Data cleaning
- Data mining
- Distributed computing
- GIS
- Hadoop
- Machine learning
- Webscraping

Challenges Facing Educators

We asked all respondents to tell us about their teaching. Of the 9366 respondents who answered the question, 43 percent (4026) are currently teaching research methods or statistics. Of those, 31 percent cover big data analytics or data science methods in their research methods or statistics course. Figure 20 shows the challenges facing those teaching big data and data science methods to students. The two biggest problems named by educators were the levels of programming and statistical knowledge that students possess. We did not ask whether educators were teaching at the undergraduate, master's, or PhD level, but that there is a skills gap among students that is making it difficult for educators to include big data methods in their course is clear.

Figure 20 Challenges facing educators teaching big data ($n \cong 1212$)



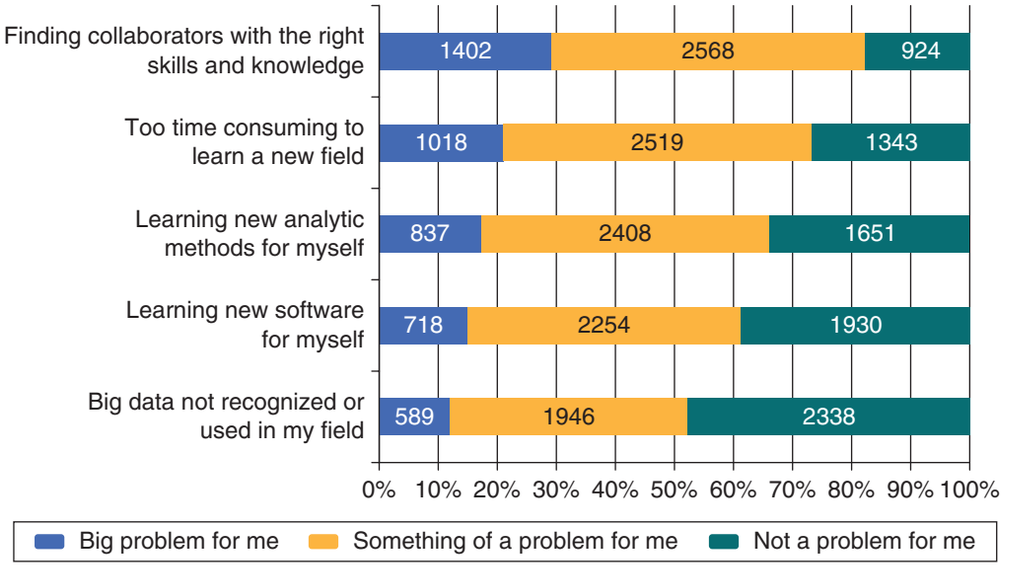
Other challenges were mentioned repeatedly:

- Resistance from students to research methods in general and especially to quantitative methods
- Poor infrastructure means the computing power or computers needed are not available
- A lack of staff with the right expertise means teaching big data would require teachers to skill up themselves first
- There is not enough time to teach big data within an existing methods course
- Limited access to the Internet, available software and resources in developing countries
- Teaching resources are not available in local languages

Barriers to Entry

For researchers who said they were not currently engaged in big data, but were interested in doing so in the future, we asked what the barriers to entry were (Figure 21). Finding collaborators with the right skills and the amount of time required to learn a new field were given as the biggest problems.

Figure 21 Challenges facing those wishing to enter into big data research ($n \cong 4894$)



Other problems mentioned included the following:

- Getting access to big data sets
- Lack of funding
- Lack of infrastructure
- Unconvinced of the value of big data research as it doesn't appear in the top journals
- Finding the right problem/question

Conclusion

In the natural sciences, the era of big data arose in the context of high-throughput instruments (e.g., new telescopes, particle accelerators, genome sequencers) designed specifically for analysis by scientists in the relevant field. These data were largely numerical and static; thus, the defining characteristic of big data was primarily its size (Lam, 2014).

In the social sciences, the new sources of data are similarly voluminous, but more importantly, derive overwhelmingly from mixed sources (e.g., social media, unstructured text, digital sensors, financial and administrative transactions) not designed to produce valid and reliable data for social scientific analysis (Lazer, Kennedy, King, & Vespignani, 2014), resulting in the challenge of harmonizing and extracting meaningful features from a variety of data streams. Moreover, many social scientific applications involve data generated dynamically, in which the quantities of interest are flows rather than stocks. In this sense, social scientific "big data" are notable less for absolute size per se than for the complexity that renders conventional methods inadequate (Doorn, 2014).

These data offer huge potential for social scientists, and at SAGE Publishing we believe that social research is at a turning point. However, the successful collection and rigorous analysis of this data require new skills, new collaborations, new research methods, and new computational tools. The findings of the survey suggest that many social scientists are already rising to some of the challenges posed by big data, and that a large number of social scientists are looking to engage in this kind of research in the future.

To find out more about what SAGE Publishing is doing to support researchers engaging or looking to engage in computational social science research, sign up to receive our monthly newsletter by e-mailing bigdataresearch@sagepub.com.

References

Doorn, P. (2014). Big data in the humanities and social sciences. Retrieved from https://sciencenode.org/feature/big-data-humanities-and-social-sciences.php

King, C. (2013). Single-author papers: A waning share of output, but still providing the tools for progress. Retrieved from http://sciencewatch.com/articles/single-author-papers-waning-share-output-still-providing-tools-progress

Lam, D. (2014). Big data challenges in social sciences & humanities research. Retrieved from http://www.datanami.com/2014/09/08/big-data-challenges-social-sciences-humanities-research/

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: Traps in big data analysis. *Science*, *343*, 1203–1205.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., . . . Van Alstyne, M. (2009). Computational social science. *Science*, *323*(5915), 721–723.

Suggestions for Further Reading

Aboab, J., Celi, L. A., Charlton, P., Feng, M., Ghassemi, M., Marshall, D. C., . . . Stone, D. J. (2016). A "datathon" model to support cross-disciplinary collaboration. *Science Translational Medicine, 8*(333), 333–338. doi:10.1126/scitranslmed.aad9072

Aral, S., & Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science, 337*(6092), 337–341. doi:10.1126/science.1215842

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509–512. doi:10.1126/science.286.5439.509

Barabási, A.-L., & Bonabeau, E. (2003). Scale-free networks. *Scientific American*, *288*(5), 60–69.

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, *23*(1), 76–91. doi:10.1093/pan/mpu011

Bauchner, H., Golub, R. M., & Fontanarosa, P. B. (2016). Data sharing: An ethical and scientific imperative. *JAMA, 315*(12), 1238–1240. doi:10.1001/jama.2016.2420

Blondel, V. D., Decuyper, A., & Krings, G. (2015). A survey of results on mobile phone datasets analysis. *arXiv*. Retrieved from https://arxiv.org/abs/1502.03406

Blumenstock, J. (2012). Inferring patterns of internal migration from mobile phone call records: Evidence from Rwanda. *Information Technology for Development, 18*(2), 107–125.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science, 350*(6264), 1073–1076. doi:10.1126/science.aac4420

Blumenstock, J., Eagle, N., & Fafchamps, M. (2016). Airtime transfers and mobile communications: Evidence in the aftermath of natural disasters. *Journal of Development Economics*, *120*, 157–181.

Bogomolov, A., Lepri, B., Larcher, R., Antonelli, F., Pianesi, F., & Pentland, A. (2016). Energy consumption prediction using people dynamics derived from cellular network data. *EPJ Data Science*, *5*(1), 1–15. doi:10.1140/epjds/s13688-016-0075-3

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature, 489*(7415), 295–298. doi:10.1038/nature11421

Cartwright, J. (2016). Smartphone science: Researchers are learning how to convert devices into global laboratories. *Nature, 531*, 669–671.

Conover, M. D., Ferrara, E., Menczer, F., & Flammini, A. (2013). The digital evolution of Occupy Wall Street. *PLoS One*, *8*(5), e64679. doi:10.1371/journal.pone.0064679

Cunningham, J. A. (2012). Using Twitter to measure behavior patterns. *Epidemiology*, *23*(5), 764–765. 10.1097/EDE.0b013e3182625e5d

D'Orazio, V., Landis, S. T., Palmer, G., & Schrodt, P. (2014). Separating the wheat from the chaff: Applications of automated document classification using support vector machines. *Political Analysis*. doi:10.1093/pan/mpt030

De Choudhury, M., Counts, S., & Horvitz, E. (2013a). *Predicting postpartum changes in emotion and behavior via social media*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France.

De Choudhury, M., Counts, S., & Horvitz, E. (2013b). *Social media as a measurement tool of depression in populations*. Paper presented at the Proceedings of the 5th Annual ACM Web Science Conference, Paris, France.

de Montjoye, Y.-A., Quoidbach, J., Robic, F., & Pentland, A. (2013). Predicting personality using novel mobile phone-based metrics. In A. M. Greenberg, W. G. Kennedy, & N. D. Bos (Eds.), Social Computing, Behavioral-Cultural Modeling and Prediction: 6th International Conference, SBP 2013, Washington, DC, April 2-5, 2013 (pp. 48-55). Berlin, Germany: Springer.

De Nadai, M., Staiano, J., Larcher, R., Sebe, N., Quercia, D., & Lepri, B. (2016). The death and life of great Italian cities: A mobile phone data perspective. *arXiv*. doi:10.1145/2872427.2883084

Dezső, Z., & Barabási, A.-L. (2002). Halting viruses in scale-free networks. *Physical Review E*, *65*(5), 055103.

Doshi, J. A., Hendrick, F. B., Gra, J. S., & Stuart, B. C. (2016). Data, data everywhere, but access remains a big issue for researchers: A review of access policies for publicly-funded patient-level health care data in the United States. *eGEMs, 4*(2). Retrieved from dx.doi.org/10.13063/2327-9214.1204

Dove, E. S., Townend, D., Meslin, E. M., Bobrow, M., Littler, K., Nicol, D., . . . Knoppers, B. M. (2016). Ethics review for international data-intensive research. *Science, 351*(6280), 1399–1400. doi:10.1126/science.aad5269

Eagle, N., Pentland, A., & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, *106*(36), 15274–15278. doi:10.1073/pnas.0900282106

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., . . . Seligman, M. E. P. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, *26*(2), 159–169. doi:10.1177/0956797614557867

Feick, R., & Robertson, C. (2015). A multi-scale approach to exploring urban places in geotagged photographs. *Computers, Environment and Urban Systems*, *53*, 96–109. doi:10.1016/j.compenvurbsys.2013.11.006

Felbo, B., Sundsøy, P., Pentland, A. S., Lehmann, S., & de Montjoye, Y.-A. (2015). Using deep learning to predict demographics from mobile phone metadata. *arXiv*. Retrieved from https://arxiv.org/abs/1511.06660

Fowler, J. H., Dawes, C. T., & Christakis, N. A. (2009). Model of genetic variation in human social networks. *PNAS*, *106*(6), 1720–1724.

Gao, J., Barzel, B., & Barabási, A.-L. (2016). Universal resilience patterns in complex networks. *Nature*, *530*(7590), 307–312. doi:10.1038/nature16948

Garcia-Herranz, M., Moro, E., Cebrian, M., Christakis, N. A., & Fowler, J. H. (2014). Using friends as sensors to detect global-scale contagious outbreaks. *PLoS One*, *9*(4), e92413. doi:10.1371/journal.pone.0092413

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, *104*(21), 8685–8690. doi:10.1073/pnas.0701361104

Grimmer, J. (2015). We are all social scientists now: How big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, *48*(01), 80–83. doi:doi:10.1017/S1049096514001784

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*. Advance online publication. doi:10.1093/pan/mps028

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.

Hidalgo, C. A., Blumm, N., Barabási, A.-L., & Christakis, N. A. (2009). A dynamic network approach for the study of human phenotypes. *PLoS Computational Biology*, *5*(4), e1000353. doi:10.1371/journal.pcbi.1000353

Hidalgo, C. A., & Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, *106*(26), 10570–10575. doi:10.1073/pnas.0900943106

Hidalgo, C. A., Klinger, B., Barabási, A.-L., & Hausmann, R. (2007). The product space conditions the development of nations. *Science*, *317*(5837), 482–487. doi:10.1126/science.1144581

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229–247. doi:10.1111/j.1540-5907.2009.00428.x

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning with applications in R*. New York: Springer.

Khoury, M. J., & Ioannidis, J. P. A. (2014). Big data meets public health. *Science*, *346*(6213), 1054–1055. doi:10.1126/science.aaa2709

King, G. (2014). Restructuring the social sciences: Reflections from Harvard's Institute for Quantitative Social Science. *PS: Political Science & Politics*, *47*(1), 165–172. doi:10.1017/S1049096513001534

King, G., & Grimmer, J. (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, *108*(7), 2643–2650.

King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, *107*(2), 326–343. doi:10.1017/S0003055413000014

King, G., Pan, J., & Roberts, M. E. (2014). Reverse-engineering censorship in China: Randomized experimentation and participant observation. *Science*, *345*(6199). doi:10.1126/science.1251722

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1218772110

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, 111*(24), 8788-8790. doi:10.1073/pnas.1320040111

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity. *Science Advances*, *2*(3). doi:10.1126/sciadv.1500779

Kuehn, B. M. (2014). Agencies use social media to track foodborne illness. *JAMA, 312*(2), 117-118. doi:10.1001/jama.2014.7731

Lazer, D. (2015). The rise of the social algorithm. *Science*, *348*(6239), 1090–1091. doi:10.1126/science.aab1422

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. A. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, *30*, 330–342.

Macy, M. W., & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, *28*, 143–166.

Pastor-Satorras, R., Castellano, C., Van Mieghem, P., & Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of Modern Physics*, *87*(3), 925–979.

Pastore y Piontti, A., Gomes, M. F. d. C., Samay, N., Perra, N., & Vespignani, A. (2014). The infection tree of global epidemics. *Network Science*, *2*(1), 132–137. doi:10.1017/nws.2014.5

Paul, M. J., Dredze, M., & Broniatowski, D. (2014). Twitter improves influenza forecasting. *PLoS Currents, 6*, ecurrents.outbreaks.90b99ed90f59bae94ccaa683a39865d39117. doi:10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117

Quercia, D., Schifanella, R., & Aiello, L. M. (2014). *The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city*. Paper presented at the Proceedings of the 25th ACM Conference on Hypertext and Social Media, Santiago, Chile.

Quercia, D., Schifanella, R., Aiello, L. M., & McLean, K. (2015). Smelly maps: The digital life of urban smellscapes. *arXiv*. Retrieved from https://arxiv.org/abs/1505.06851

Radford, J., Pilny, A., Ognyanova, K., Horgan, L., Wojcik, S., & Lazer, D. (2016). *Gaming for science: A demo of online experiments on VolunteerScience.com*. Paper presented at the Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion, San Francisco, CA.

Reis, B. Y., Kohane, I. S., & Mandl, K. D. (2009). Longitudinal histories as predictors of future diagnoses of domestic abuse: Modelling study. *BMJ, 339*. doi:10.1136/bmj.b3677

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., . . . Rand, D. G. (2013). Structural topic models for open-ended survey. *American Journal of Political Science*, *58*(4), 1064–1082.

Ronen, S., Gonçalves, B., Hu, K. Z., Vespignani, A., Pinker, S., & Hidalgo, C. A. (2014). Links that speak: The global language network and its association with global fame. *Proceedings of the National Academy of Sciences, 111*(52), E5616-E5622. doi:10.1073/pnas.1410931111

Rose, S. (2013). Mortality risk score prediction in an elderly population using machine learning. *American Journal of Epidemiology*, *177*(5), 443–452. doi:10.1093/aje/kws241

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). *Earthquake shakes Twitter users: Real-time event detection by social sensors*. Paper presented at the Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC.

Savage, N. (2015). Mobile data: Made to measure. *Nature*, *527*(7576), S12–S13. doi:10.1038/527S12a

Schich, M., Song, C., Ahn, Y.-Y., Mirsky, A., Martino, M., Barabási, A.-L., & Helbing, D. (2014). A network framework of cultural history. *Science, 345*(6196), 558–562. doi:10.1126/science.1240064

Schwalbe, M. (2016). *Statistical challenges in assessing and fostering the reproducibility of scientific results: Summary of a workshop*. Washington DC: The National Academies Press.

Servick, K. (2015). Proposed study would closely track 10,000 New Yorkers. *Science*, *350*(6260), 493–494. doi:10.1126/science.350.6260.493

Steinert-Threlkeld, Z. C., Mocanu, D., Vespignani, A., & Fowler, J. (2015). Online social networks and offline protest. *EPJ Data Science, 4*(1), 1–9. doi:10.1140/epjds/s13688-015-0056-y

Stopczynski, A., Pietri, R., Pentland, A., Lazer, D., & Lehmann, S. (2014). Privacy in sensor-driven human data collection: A guide for practitioners. *arXiv*. Retrieved from https://arxiv.org/abs/1403.5299

Ugander, J., Backstrom, L., Marlow, C., & Kleinberg, J. (2012). Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, *109*(16), 5962–5966. doi:10.1073/pnas.1116502109

Zhang, Q., Gioannini, C., Paolotti, D., Perra, N., Perrotta, D., Quaggiotto, M., . . . Vespignani, A. (2015). Social data mining and seasonal influenza forecasts: The FluOutlook Platform. In A. Bifet et al. (Eds.), *Machine learning and knowledge discovery in databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part III* (pp. 237–240). Cham, Switzerland: Springer International Publishing.

Zyskind, G., Nathan, O., & Pentland, A. (2015). Enigma: Decentralized computation platform with guaranteed privacy. *arXiv*. Retrieved from https://arxiv.org/abs/1506.03471