Training Computational Social Science PhD Students for Academic and Non-Academic Careers

Aniket Kesari, *Fordham University, USA*
Jae Yeon Kim, *Code for America, USA*
Sono Shah, *Pew Research Center, USA*
Taylor Brown, *Meta, USA*
Tiago Ventura, *Georgetown University, USA*
Tina Law, *City University of New York Graduate Center, USA*

ABSTRACT Social scientists with data science skills increasingly are assuming positions as computational social scientists in academic and non-academic organizations. However, because computational social science (CSS) is still relatively new to the social sciences, it can feel like a hidden curriculum for many PhD students. To support social science PhD students, this article is an accessible guide to CSS training based on previous literature and our collective working experiences in academic, public-, and private-sector organizations. We contend that students should supplement their traditional social science training in research design and domain expertise with CSS training by focusing on three core areas: (1) learning data science skills, (2) building a portfolio that uses data science to answer social science questions, and (3) connecting with computational social scientists. We conclude with practical recommendations for departments and professional associations to better support PhD students.

As more social scientists have gained training and experience in data science during their graduate studies, an increasing number have assumed positions as computational social scientists in academic and non-academic organizations. We define computational social science (CSS) as a field that engages the social sciences and data science by applying novel digital and digitized data and computational methods to advance social scientific understanding of human behavior (Edelmann et al. 2020; Salganik 2019). What distinguishes computational social scientists from social scientists and data scientists is their ability to combine research design, domain knowledge, and computational methods to generate scientific knowledge about human behavior (Grimmer 2015).

Computational social scientists currently work in academic departments, professional schools, nonprofit organizations (e.g.,

Aniket Kesari *is associate professor of law at Fordham University Law School. He can be reached at akesari@fordham.edu.*
Jae Yeon Kim *is senior data scientist at Code for America's Safety Net Innovation Lab. He can be reached at jykim@codeforamerica.org.*
Sono Shah *is senior computational social scientist on the Pew Research Center's Data Labs team. He can be reached at sshah@pewresearch.org.*
Taylor Brown *is research scientist and engineer at Meta. She can be reached at taylorwbrown@meta.com.*
Tiago Ventura *is assistant professor of computational social science at the McCourt School of Public Policy at Georgetown University. He can be reached at tv186@georgetown.edu.*
Tina Law *is a postdoctoral scholar at the Stone Center on Socio-Economic Inequality at the City University of New York Graduate Center. She can be reached at tlaw@gc.cuny.edu.*

doi:10.1017/S1049096523000732

https://doi.org/10.1017/S1049096523000732 Published online by Cambridge University Press

Code for America, Pew Research Center's Data Labs, and Urban Institute); tech companies (e.g., Meta, X, Google, Amazon, and Microsoft); international organizations (e.g., the World Bank and UN Global Pulse Labs); and government agencies (e.g., the US Federal Reserve, Census Bureau, and Office of Evaluation Sciences). Yet, how to become a computational social scientist still feels like a "hidden curriculum" (Barham and Wood 2022; Calarco 2020) to many social science PhD students. Because the field is so new, most social science PhD programs do not yet offer systematic training or dedicated advising to help students navigate careers as computational social scientists. Moreover, access to training and support remains unequal across institutions, and cross-institutional activities such as conferences and workshops continue to lack substantive racial and gender representation in terms of speakers and participants.

To initiate conversations on how training, professionalization, and support of computational social scientists can be more systematic and inclusive, this article makes explicit the informal knowledge of navigating CSS for PhD students in the social sciences. We offer this guide, drawing on previous studies that offer constructive guidelines for innovating political science graduate training in the digital era (Barham and Wood 2022; Grimmer 2015; Grimmer, Roberts, and Stewart 2021) as well as our collective experiences as computational social scientists working in academic, public-, and private-sector organizations. We break down the CSS professionalization process into three core areas: (1) learning data science skills, (2) building a CSS portfolio, and (3) connecting with computational social scientists. For each area, we identify and elaborate on core competencies and additional useful skills specific to the academic and non-academic job markets (table 1).

LEARNING DATA SCIENCE SKILLS

We argue that effective CSS training begins—first and foremost—with strong training in two areas that social science PhD programs already focus on: research design and domain expertise. On that foundation, social science PhD students interested in CSS also should focus on learning a specific set of data science skills: programming fluency, data management, collaborative research skills, machine learning paradigms, and the ability to engage with ethical concerns unique to massive data and computational methods.

> *We argue that effective CSS training begins—first and foremost—with strong training in two areas that social science PhD programs already focus on: research design and domain expertise.*

Research Design and Domain Knowledge Expertise

To conduct CSS research, social science PhD students first must learn how to effectively design and execute research designs and

Table 1

Computational Social Science Professionalization Process

Learning Data Science Skills	
Core Competencies	• Ability to design and execute research projects from end to end (data to report) • Domain expertise • Programming fluency in `R` and/or `Python` • Experience with data management, particularly with managing large, messy, and unstructured data • Effective communication and collaborative research skills with both technical and nontechnical colleagues (e.g., version control and documentation) • Practiced knowledge of machine learning and traditional quantitative social science paradigms • Engagement with ethical concerns about digital and digitized data and computational methods (e.g., privacy protection and algorithmic bias)
Additional Market-Specific Skills	• Ability to apply theory, methods, and findings to the practical aims of a product and/or organization (*non-academic*) • Proficiency with relational database languages (e.g., SQL) and cloud-based databases *(non-academic especially)*
Building a CSS Portfolio	
Core Competencies	• Publicly available research projects documented from end to end demonstrating engagement with social science and applied aspects of a research project via problem definition, hypothesis generation, data and outcome selection, and measurement and method application • Reproducible, efficient, and communicable code via GitHub • Publish and serve as reviewer for journal publications/conference proceedings
Additional Market-Specific Skills	• Sharing learnings through research notes *(non-academic)* and tutorials *(academic)*
Connecting with Computational Social Scientists	
Core Competencies	• Attend and know how to navigate cross-disciplinary computational social science conferences
Additional Market-Specific Skills	• Work with computational social scientists through internships and work with civic, social, and nonprofit organizations *(non-academic)* • Connect with computational social scientists working on similar topics in different sectors via online platforms (e.g., LinkedIn and Slack) *(non-academic)*

https://doi.org/10.1017/S1049096523000732 Published online by Cambridge University Press

develop and apply domain knowledge expertise. Computational social scientists are expected to draw on their doctoral training to lead and carry out research projects from end to end. For example, in a non-academic organization, computational social scientists often work on teams with engineers and user experience (UX) designers in which they are responsible primarily for research design. Computational social scientists also are expected in academic and non-academic settings to hold expertise not on all computational methods but instead on specific substantive and methodological topics, such as natural language processing and public opinion. We emphasize research design and domain expertise to underscore that CSS training should not replace but rather supplement the traditional social science PhD curriculum.

Programming Fluency

Programming fluency—defined as the ability to write code to collect, manage, and analyze data in open-source programming languages commonly used for data science (e.g., `R` and `Python`)—is fundamental to CSS research. Social science PhD programs often provide some training during statistics courses on how to use commercial point-and-click software such as `SPSS` and `Stata` for quantitative analysis. However, CSS students should prioritize learning a programming language for three reasons (Kim and Ng 2022). First, several core computational methods (e.g., machine learning and natural language processing) currently rely on packages developed and maintained in `Python` and `R`. Second, many sources of massive digital and digitized data (e.g., Tweets and Facebook posts) are accessible only through application programming interfaces (APIs), which can be queried easily through `R` and `Python`. Third, because `Python` and `R` are the *lingua franca* of computational social scientists, being fluent in a programming language will enable students to collaborate on research projects (in academic and non-academic organizations) and contribute to the development and maintenance of software and other tools used more widely in the field.

Data Management

CSS research involves vast quantities of data and therefore requires strong data-management skills. In particular, there is a need for researchers who are well versed in working with large, messy, and often unstructured data. Many students are likely already familiar with the basics of spreadsheet software (e.g., Excel). They even may have practice with summarizing, aggregating, and manipulating data to perform basic calculations and visualizations. This knowledge is translated easily to the use of “DataFrames” in `R` and `Python`. Eventually, students may find that they need to work with increasingly larger datasets. In many cases, this involves using cloud computing resources (e.g., Amazon’s Web Services, Microsoft’s Azure, and Google’s BigQuery) for storing data or conducting analysis and learning how to query and manage relational databases using Structured Query Language (SQL). Indeed, many non-academic organizations store their data in large, relational databases, making proficiency in languages like SQL necessary for working in these settings.

Beyond familiarity with data structures, students interested in conducting CSS research are well served to develop good computing practices early, including developing clear workflows for data analysis and coding as well as setting procedures for documenting any changes to data and code. These practices ease ongoing individual or collaborative work and they maximize the usefulness of a dataset by making it legible to other users in the future. Learning version control also is key for ensuring that data are not easily destroyed or lost in a project that is long term and/or involves many collaborators.

Collaborative Research Skills

CSS research is a collaborative effort—especially in non-academic organizations. In collaborative research environments, familiarity with a version control framework such as git and GitHub is essential. We elaborate on this in the third section.

Machine Learning Paradigms

Most social science PhD curricula provide quantitative research training in the form of statistical and causal inference. Whereas CSS research certainly uses these paradigms, it also relies heavily on machine learning paradigms that currently are not taught in most social science PhD programs. There are many different approaches to machine learning, with specific applications—for example—to causal inference (Athey 2015; Varian 2016); text analysis (Grimmer, Roberts, and Stewart 2021; Grimmer and Stewart 2013); and measurement (Lundberg, Johnson, and Stewart 2021). Being able to apply machine learning paradigms and adroitly navigate the relationships between machine learning and more traditional quantitative social science paradigms (e.g., inferential statistics and causal inference) is an important skill for computational social scientists and one that will set them apart from social scientists and data scientists.

Research Ethics

Computational social scientists can and should have an important role in generating conversations about ethics in CSS research, particularly as it relates to machine learning and artificial intelligence (Noble 2018). Increasingly, technological advancements in CSS research—in terms of computational tools and data collection—outpace public and scholarly discussions on how these advancements may affect society. For example, the fairly recent development of large language models such as ChatGPT raises ethical concerns about everything from plagiarism to data rights to the safety of open-access code (Bender et al. 2021). Computational social scientists have the ability and responsibility to apply their research ethics training to proactively anticipate ethical concerns related to their work, especially when it affects vulnerable populations.

Learning Resources

How can social science PhD students learn these data science skills for conducting CSS research? Ideally, they can gain programming fluency and other data science skills through coursework and research projects. Because few social science departments currently have formal CSS curriculum, we recommend that students supplement their coursework with courses on programming, data structures, machine learning, and other computational methods of interest if they are offered by other departments or if there are faculty available who can supervise independent study.

For students who may not have access to data science training at their home institution, there is a growing number of external CSS training opportunities. The Summer Institutes in CSS (https://sicss.io) are free and held every year in locations around the world. Each location provides participants with two weeks of

https://doi.org/10.1017/S1049096523000732 Published online by Cambridge University Press

training on various CSS topics. Because the Summer Institutes emphasize breadth of CSS topics, they are especially well suited for early-stage PhD students. Participants are selected through a competitive application process. Interested students should prepare and submit an application—consisting of a curriculum vitae (CV), a statement of interest, and a writing sample—by early spring.

After students have gained programming fluency and a sense of their research interests, they can pursue additional external training on specific topics. The Inter-University Consortium for Political and Social Research regularly offers short courses on CSS topics including machine learning, network analysis, and agent-based modeling. Moreover, students can continue to hone their data science skills with free online tutorials and resources provided by groups such as Data Carpentry (https://datacarpentry.org) and free in-person workshops and meetups organized by groups such as R-Ladies (https://rladies.org).

BUILDING A CSS PORTFOLIO

A CSS portfolio, like a data science portfolio, includes projects and outputs (Robinson and Nolis 2020) and is an integral part of data science education (Nolan and Stoudt 2021) and career building (Craig et al. 2018). Preparing a portfolio is especially critical for pursuing a non-academic career in which publications are not the only—and far from the most important—metric for performance. Students do not need to take courses in everything related to CSS; at some point, building a portfolio that demonstrates applied knowledge will be a better use of their time and effort. One way to think about building a successful portfolio is to imagine it as a series of "deliverables" that demonstrate that one understands the CSS pipeline.

One way to think about building a successful portfolio is to imagine it as a series of "deliverables" that demonstrate that one understands the CSS pipeline.

It is helpful to first discuss how a CSS portfolio differs from a CV. First, a CSS portfolio defines outcomes more broadly by highlighting non-publication outputs, including open-source software development, interactive maps, and dashboards, as indicators of strong programming and public-engagement skills. Second, a CSS portfolio focuses on processes, not only outputs. In particular, a strong CSS portfolio will demonstrate that a researcher can write legible and reproducible code—a highly valued skill in non-academic research settings, where code often needs to be reproduced quickly and efficiently.

Networking is as valuable to computational social scientists in terms of finding collaborators and jobs; however, it operates slightly differently in CSS because the opportunities to connect span more spaces across disciplines and sectors.

How does one create and share an effective CSS portfolio? Graduate students should become familiar with version control tools such as git and make frequent use of open-source coding platforms such as GitHub. Git is a tool for managing and tracking changes to a codebase, thereby allowing users to exercise control over different versions of their work. GitHub is an online platform built on git that provides additional features for collaboration, tracking, and hosting code repositories. This latter function is especially important for developing and sharing a CSS portfolio. There is a learning curve in the beginning because these tools require familiarity with command-line tools. However, the payoff is enormous if graduate students are able to develop and share a varied and extensive portfolio throughout their PhD program.

With these skills, graduate students can share a CSS research project by making a repository openly accessible on GitHub. The repository should demonstrate not only technical skills but also illustrate substantive knowledge on a topic. To this end, `readme` files offer useful space to outline the theoretical and empirical motivations behind a research project. It also is worth noting that research projects do not need to be fully finished before they are shared through repositories; in fact, the point of an open-source repository is to show and share one's work and its value even when the project is not fully mature. Graduate students also can share their work by writing brief research notes in the form of blog posts and tutorials hosted on their personal website.

CONNECTING WITH COMPUTATIONAL SOCIAL SCIENTISTS

For any social scientist, networking is useful for developing collaborative research projects and learning about job opportunities (Kim, Lebovits, and Shugars 2022). Networking is as valuable to computational social scientists in terms of finding collaborators and jobs; however, it operates slightly differently in CSS because the opportunities to connect span more spaces across disciplines and sectors.

CSS Conferences

Social science PhD programs typically train and encourage students to attend flagship academic conferences within their disciplines. However, CSS PhD students can benefit from taking part in a broader range of conferences. There are many cross-disciplinary conferences focused on CSS topics where graduate students can meet other scholars as well as share and receive feedback on their research. Popular CSS conferences include the International Conference on CSS; the Association on Computing Machinery Conference on Human Factors in Computing Systems; the International Conference on Web and Social Media; the Text as Data Conference; the Network Science Society Conference; the International Social Networks Conference; the Politics and Computational Social Science Conference; and the Association on Computing Machinery Conference on Fairness, Accountability, and Transparency.

CSS conferences mostly operate like any other academic conference but differ in that they tend to (1) include participants not

https://doi.org/10.1017/S1049096523000732 Published online by Cambridge University Press

only from academia but also from industry, (2) use blinded review processes to select papers and sometimes publish papers as part of conference proceedings, and (3) organize well-attended poster sessions. These unique aspects of CSS conferences provide excellent opportunities for students to learn about research trends and professionalization norms across different disciplines and sectors. CSS conferences also are great places to meet and connect with peers (Kim, Lebovits, and Shugars 2022), many of whom may be seeking a cross-disciplinary collaborator or may provide helpful information about private- and public-sector internships.

In addition to the increasing number of CSS conferences, many disciplinary conferences recently have added preconferences focused on CSS topics. For example, the political networks section of the American Political Science Association also convenes PolNet, an annual convening that consists of workshops and panels. In recent years, the American Sociological Association also has organized a computational sociology preconference ahead of its annual conference.

Internships

Social science PhD students often spend their summers preparing for program milestones and conducting research. For CSS PhD students, these may serve as especially opportune periods to pursue additional training and engage in career exploration through internships. Whereas summer internships have long been a common part of the undergraduate experience and PhD training in the information and computer sciences, they also are increasingly an important part of the graduate experience for CSS PhD students. Participating in an internship—especially early in their PhD program—can help students to understand whether they want to remain in academia or pursue a career in the private or public sector.

Several organizations now offer internships to CSS PhD students. Many tech and social media companies, including Meta, X, Google, Amazon, and Microsoft Research, have internship programs. Some public agencies and nonprofit organizations also have internship programs, such as the Civic Digital Fellowship and the Data Science for Social Good Fellowship. Moreover, there are internship and job opportunities in the private and public sectors for social science PhD students who use mixed or qualitative methods; their ability to conduct interviews and focus groups are valued in—for example—UX and community-oriented research.

Internships may vary in their structure, length, residency requirements, and compensation, but they all typically provide opportunities for students to engage in research in an applied and often team-based setting. Students should be aware that internships are highly competitive and dedicate a minimum of several weeks to apply and prepare for multiple rounds of interviews. Interviews generally include aptitude assessments in research design, statistical methods, and coding. We encourage students to reach out to former interns to learn about their interviewing and internship experiences. Recruitment for internships in the private sector typically occurs in the fall, with most internships transpiring throughout the summer months. However, many companies offer internships year round, so students should request a time that works best for their academic schedule. Deadlines for public-sector internship applications are less clustered and often are set closer to internship start dates.

Online Networking Tools

CSS PhD students also can use online tools to connect with other scholars. For example, LinkedIn is a useful platform for connecting with CSS researchers who are working on similar topics in the private or public sector. LinkedIn may be especially useful if students are interested in connecting with CSS researchers who are working in nonprofit organizations and think tanks because they may be less likely to attend CSS conferences compared to industry researchers. Other useful platforms for connecting with computational social scientists currently include Slack and X. There are many groups and organizations such as Black in AI, Women in Machine Learning, and R-Ladies in which CSS PhD students may find community. In general, we encourage graduate students to reach out to CSS scholars, particularly those working in non-academic organizations, because they often have important insight about the aims, needs, and priorities of their organization.

CONCLUSION

CSS creates many new career opportunities for social scientists. This article shares advice and resources to provide PhD students a guide for navigating these emerging career paths in academic and non-academic job markets. We encourage PhD students interested in pursuing a career as a computational social scientist to build on their research design skills and domain expertise by (1) learning data science skills, (2) building a CSS portfolio, and (3) connecting with computational social scientists.

As the field of CSS continues to grow, departments and professional associations likely will need to assume a more coordinated and proactive role in supporting graduate students. For now, we draw on recent work on social science PhD training and job placement (Berdahl, Malloy, and Young 2020) to suggest the following changes that departments may consider adopting to better support PhD students who are interested in CSS. Notably, these same changes are likely to attract and retain CSS faculty:

- provide information on non-academic career opportunities, including internships, to students at the beginning of PhD training
- integrate data science skills building into existing curriculum (e.g., integrating `R` or `Python` in introductory statistics courses)
- offer new courses on computational methods and data management
- identify relevant data science coursework in other departments and recognize earned credits
- identify relevant data science faculty in other departments who can serve on dissertation committees
- offer options for students to substitute a program requirement (e.g., one field exam) for an internship or advanced CSS training
- provide support for current faculty to pursue CSS training
- hire more CSS faculty and recruit computational social scientists from industry and nonprofit organizations for faculty and visiting-scholar positions
- evolve publication standards to increasingly value CSS conference proceedings, journals, and the value of collaborative CSS project work

Surfacing the hidden curriculum is an important step toward democratizing access to CSS and supporting all students who consider themselves a computational social scientist in achieving their goals. We do not claim to have the final word on what makes

https://doi.org/10.1017/S1049096523000732 Published online by Cambridge University Press

a computational social scientist or a CSS curriculum; instead, we hope that this guide generates important and necessary conversations about the long-term development of this field.

ACKNOWLEDGMENTS

The authors of this article are listed alphabetically because all contributed equally to it. We thank Milan de Vries, Eric Giannella, Brian Heseung Kim, Rebecca Johnson, Joscha Legewie, and Sarah Shugars, as well as three anonymous reviewers for their constructive comments on an earlier draft of this article. We would love to hear from those who use this article for course syllabi and PhD curriculum development.

CONFLICTS OF INTEREST

The authors declare that there are no ethical issues or conflicts of interest in this research. ■

REFERENCES

Athey, Susan. 2015. "Machine Learning and Causal Inference for Policy Evaluation." *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 5–6. Sydney, Australia; August 10–13.

Barham, Elena, and Colleen Wood. 2022. "Teaching the Hidden Curriculum in Political Science." *PS: Political Science & Politics* 55 (2): 324–28.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the 2021 Association for Computing Machinery Conference on Fairness, Accountability, and Transparency*, 610–23. Virtual event, March 3–10.

Berdahl, Loleen, Jonathan Malloy, and Lisa Young. 2020. "Faculty Perceptions of Political Science PhD Career Training." *PS: Political Science & Politics* 53 (4): 751–56.

Calarco, Jessica McCrory. 2020. *A Field Guide to Grad School: Uncovering the Hidden Curriculum*. Princeton, NJ: Princeton University Press.

Craig, Michelle, Phill Conrad, Dylan Lynch, Natasha Lee, and Laura Anthony. 2018. "Listening to Early-Career Software Developers." *Journal of Computing Sciences in Colleges* 33 (4): 138–49.

Edelmann, Achim, Tom Wolff, Danielle Montagne, and Christopher A. Bail. 2020. "Computational Social Science and Sociology." *Annual Review of Sociology* 46 (1): 61–81.

Grimmer, Justin. 2015. "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together." *PS: Political Science & Politics* 48 (1): 80–83.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annual Review of Political Science* 24:395–419.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97.

Kim, Jae Yeon, and Yee Man Margaret Ng. 2022. "Teaching Computational Social Science for All." *PS: Political Science & Politics* 55 (3): 605–609.

Kim, Seo-Young Silvia, Hannah Lebovits, and Sarah Shugars. 2022. "Networking 101 for Graduate Students: Building a Bigger Table." *PS: Political Science & Politics* 55 (2): 307–12.

Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart. 2021. "What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory." *American Sociological Review* 86 (3): 532–65.

Noble, Sofiya Umoja. 2018. *Algorithms of Oppression*. New York: New York University Press.

Nolan, Deborah, and Sara Stoudt. 2021. "The Promise of Portfolios: Training Modern Data Scientists." *Harvard Data Science Review* 3 (3). https://doi.org/10.1162/99608f92.3c097160.

Robinson, Emily, and Jacqueline Nolis. 2020. *Build a Career in Data Science*. Shelter Island, NY: Manning Publications.

Salganik, Matthew J. 2019. *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

Varian, Hal R. 2016. "Causal Inference in Economics and Marketing." *Proceedings of the National Academy of Sciences* 113 (27): 7310–15.

https://doi.org/10.1017/S1049096523000732 Published online by Cambridge University Press