Building Research Infrastructures to Study Digital Technology and Politics: Lessons from Switzerland

Fabrizio Gilardi, *University of Zurich, Switzerland*
Lucien Baumgartner, *University of Zurich, Switzerland*
Clau Dermont, *Swiss Federal Office of Culture, Bern, Switzerland*
Karsten Donnay, *University of Zurich, Switzerland*
Theresa Gessler, *University of Zurich, Switzerland*
Maël Kubli, *University of Zurich, Switzerland*
Lucas Leemann, *University of Zurich, Switzerland*
Stefan Müller, *University College Dublin, Ireland*

ABSTRACT The relationship between digital technology and politics is an important phenomenon that remains poorly understood due to several structural problems. A key issue is the lack of adequate research infrastructures or the lack of access. This article discusses the challenges many social scientists face and presents the infrastructure we built in Switzerland to overcome them, using COVID-19 as an example. We conclude by discussing seven lessons we learned: automatization is key; avoid data hoarding; outsource some parts of the infrastructure but not others; focus on substantive questions; share data in the context of collaborations; engage in targeted public outreach; and collaboration is more promising than competition. We hope that our experience is helpful to other researchers pursuing similar goals.

Fabrizio Gilardi *is professor of policy analysis at the University of Zurich. He can be reached at gilardi@ipz.uzh.ch.*
Lucien Baumgartner *is a PhD student in philosophy at the University of Zurich. He can be reached at lucien.baumgartner@philos.uzh.ch.*
Clau Dermont *is employed at the Swiss Federal Office of Culture. He can be reached at clau.dermont@pm.me.*
Karsten Donnay *is assistant professor of political behavior and digital media at the University of Zurich. He can be reached at donnay@ipz.uzh.ch.*
Theresa Gessler *is a postdoctoral researcher in political science at the University of Zurich. She can be reached at gessler@ipz.uzh.ch.*
Maël Kubli *is a PhD student in political science at the University of Zurich. He can be reached at kubli@ipz.uzh.ch.*
Lucas Leemann *is assistant professor of comparative politics and empirical democracy research at the University of Zurich. He can be reached at leemann@ipz.uzh.ch.*
Stefan Müller *is assistant professor and Ad Astra Fellow in the School of Politics and International Relations at University College Dublin. He can be reached at stefan.mueller@ucd.ie.*

Digital technology affects politics in many ways. The role of social media in elections, especially in connection with their potential to spread disinformation, has been one of the most visible aspects of the phenomenon. It also is one of the most researched in political science and political communication (Grinberg et al. 2019; Guess, Nagler, and Tucker 2019; Jungherr, Rivero, and Gayo-Avello 2020). However, digital technology also affects how public administration works (i.e., "e-government") and, more generally, how the state interacts with its citizens (and potentially surveils them). Moreover, digital tools and platforms promise to facilitate new forms of participation and citizen involvement in decision-making processes (i.e., "civic tech").

The connections between digital technology and politics are complex, multifaceted, and—despite a surge of high-quality research—not as well understood as they should be. The research community lacks clear answers to many questions that are highly

doi:10.1017/S1049096521000895

salient to the public and decision makers alike: What is the prevalence of disinformation on different platforms and countries? How do online political ads affect behavior and are they similar to offline ads? How can we strike a balance between data protection and the transparency of digital platforms? How can digital technology improve political participation?

One reason why answering these questions is difficult is the existence of several structural challenges. We argue that a key problem is the lack of adequate research infrastructures or the lack of access to them. We outline the nature of these challenges and then present the infrastructure that we built to overcome them in the Swiss context, which we illustrate using the example of the public debate on COVID-19. We conclude by offering recommendations for other scholars interested in replicating our efforts in other contexts.

CHALLENGES OF STUDYING DIGITAL TECHNOLOGY AND POLITICS

The first challenge is data access. Many data that researchers need to answer questions on digital technology and politics are difficult to obtain, for several reasons. First, the skills required to collect online data are different from what we traditionally train our students in (Salganik 2017). Several initiatives, such as the Summer Institutes in Computational Social Science,[1] have helped social scientists to close the skills-needs gap. With new graduate programs and more methods courses geared toward computational social science, many junior scholars now are trained in many of these essential skills. However, even with the required skills, data access remains problematic. Researchers are largely dependent on the goodwill of digital platforms. Some (e.g., Wikipedia) provide Application Programming Interfaces (APIs) that allow for extensive data access. Others (e.g., Twitter) recently expanded access with a new API for Academic Research, rolled out in 2021. Facebook announced plans for a similar API, but these developments remain uncertain, and access could be restricted at any time and without notice. This state of affairs was described as an "APIcalypse" and "postAPI age," preventing independent, critical research on digital platforms (Bruns 2019; Freelon 2018). Today, access to the most valuable data remains exceedingly difficult without significant resources or collaborations, effectively limiting many types of studies to a select group of researchers. Initiatives such as Social Science One (King and Persily 2020) have worked to provide transparent processes to gain access to Facebook data, but Social Science One "is not a one-size-fits-all model, nor is it intended to be" (Levi and Rajala 2020, 710). Moreover, all existing efforts "are both novel and experimental. Evaluation of which is best suited for what type of data and circumstances is still in the future" (Levi and Rajala 2020, 711).

The second challenge is data permanence. Typically, researchers collect the data that they need for their projects and, when funding runs out or the project ends, they stop. The data are not updated and other researchers do not have access to them—often dictated by the platforms' terms of use. New projects basically must start over from anew. This is inefficient, and significant resources are regularly wasted redoing what already has been done. A better system is for data to be collected continuously in a centralized way so that many researchers could access what they need when they need it, including for replication purposes. At the same time, hoarding vast amounts of data that nobody uses is not meaningful.

Third, data sharing often is constrained by more or less clearly defined rules. Twitter data, for example, can be shared freely only within research groups—although what counts as a research group is not entirely clear. Tweet IDs can be shared publicly but they are not an adequate solution. These IDs allow other researchers to identify relevant posts, but they still must be redownloaded and preprocessed. For replication purposes, the arrangement is ineffective because tweets (and accounts) may have been deleted since the original data collection, making it impossible to reproduce the original results (Zubiaga 2018).

The fourth challenge is related to data protection. The European Union (EU)'s General Data Protection Regulation (GDPR) has a global reach because it affects any researcher collecting data on EU citizens. Although the GDPR includes an explicit research exception, it is poorly defined (European Data Protection Supervisor 2020). Consequently, researchers must be aware of the constraints set by GDPR without clear guidelines on how to navigate them.

Fifth, most research in this area is focused on one specific country: the United States (Jungherr, Rivero, and Gayo-Avello 2020, 7). Its size, language, institutional context, and electoral and party system are not representative of other countries. Therefore, results based on the United States might overstate certain effects because they are bound to one context rather than controlled for in various settings. As Jungherr, Rivero, and Gayo-Avello (2020, 7) stated: "Any uncritical generalizations on the role of digital media in politics based on cases and findings from the United States is obviously deeply naïve." Relatedly, research in the US case does not have to be concerned about languages. In other contexts, however, multiple languages are relevant and constitute a challenge, despite the increasingly high quality of automatic translation and progress in natural-language-processing approaches for languages other than English (de Vries, Schoonvelde, and Schumacher 2018).

A RESEARCH INFRASTRUCTURE TO STUDY DIGITAL TECHNOLOGY AND POLITICS

This section describes the infrastructure that we built in the Digital Democracy Lab at the University of Zurich,[2] using a relatively simple COVID-19 analysis as an example. The next section discusses which broader lessons can be drawn from our experience.

Beginning with our example, figure 1 shows the salience of COVID-19 in Switzerland from the end of December 2019 until August 2020, with a focus on traditional and social media. The analysis includes about 7 million Tweets for 300,000 users, 11,000 Facebook posts by political actors published on 169 public pages, and 1.4 million articles published in 84 newspapers (Gilardi et al. 2021a). These documents are multilingual, including Switzerland's three largest official languages (i.e., German, French, and Italian). Across the three platforms, we can observe the striking extent to which COVID-19 dominated public attention. The Swiss debate on COVID-19 began after the first cases were detected in Europe and achieved a high degree of salience when the Swiss federal government enacted the first measures against the spread of the virus. Salience reached a peak when Switzerland declared a state of emergency. The topic's salience gradually decreased, with spikes when the government announced new rules. It is interesting that attention to COVID-19 was higher in newspapers (with peaks of about 70% of all articles) than on social media. This analysis illustrates the basic workflow that is the basis of more

Figure 1

Salience of COVID-19 in Traditional and Social Media in Switzerland

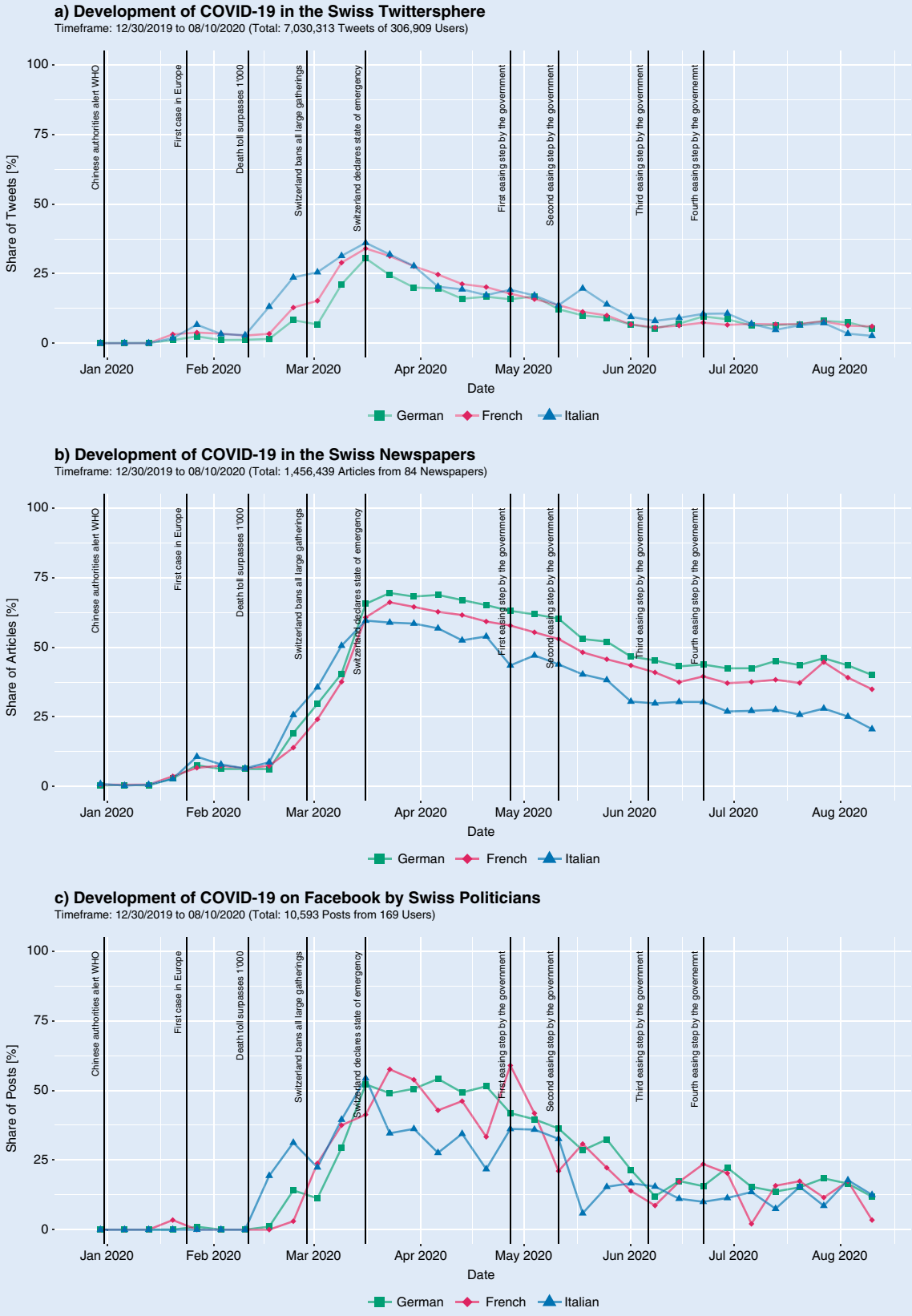
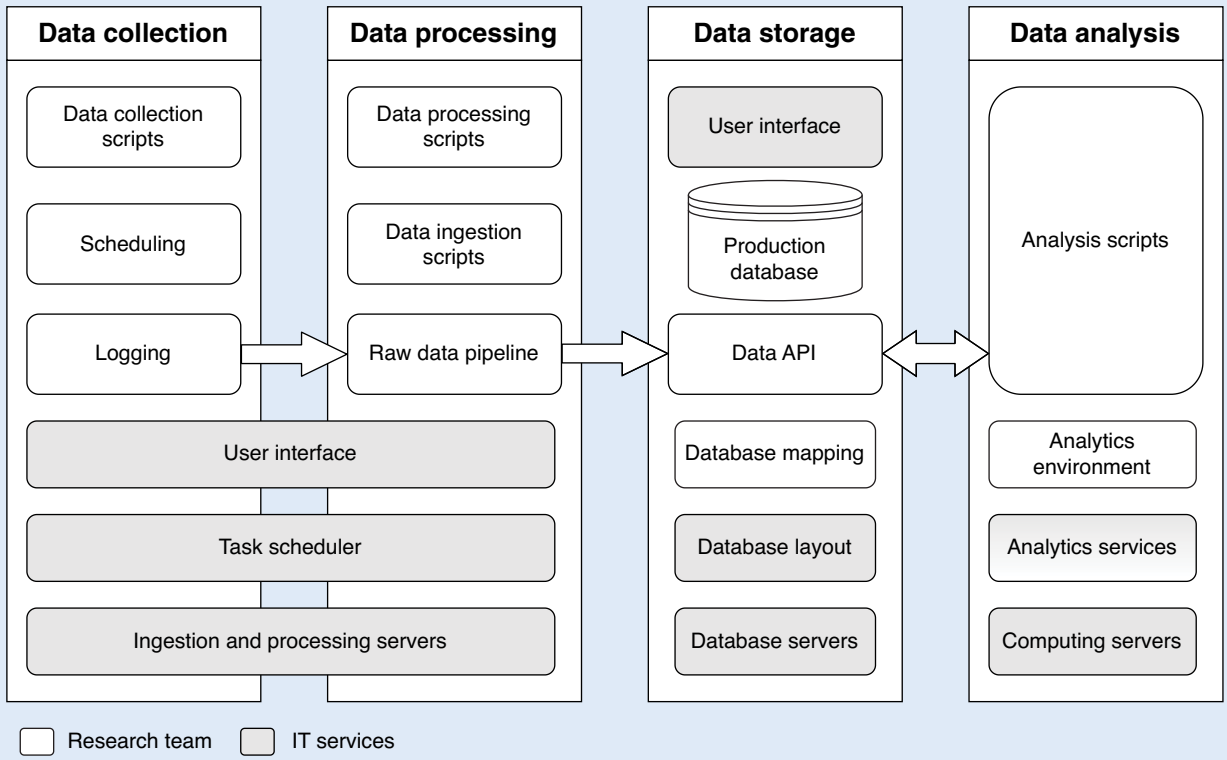


Figure 2

Overview of the Research Infrastructure



sophisticated studies. It requires overcoming several of the challenges discussed previously—in particular, data access and permanence.

Because of the infrastructure that we already had built, we were able to efficiently collect and analyze new data. The infrastructure, illustrated in figure 2, has four components: data collection, data processing, data storage, and data analysis. Specifically, the first and second components consist of servers and scripts to carry out data collection and processing tasks that, importantly, can be scheduled (e.g., download a Twitter timeline automatically once a day). The third component consists of a database in which all information is stored and checked automatically for integrity and duplicates, once daily. The database is distributed over several servers to ensure data permanence; if there is a problem with a server, the database remains fully operational, including backup capabilities. The fourth component, data analysis, runs on additional servers with graphical user interfaces for R and Python analyses that, as for data collection, can be scheduled. For example, new documents can be classified automatically using existing scripts as they are added to the database.

In our COVID-19 example, we needed to collect millions of tweets, hundreds of thousands of newspaper articles, and thousands of Facebook posts. Each source has its own document format and requires different data-collection and processing features. Our infrastructure provided us with a unique advantage: although we had to adapt our scripts (e.g., to collect Twitter timelines), we could build on the versions we already had implemented in the infrastructure described previously.

To build our infrastructure, we engaged with the relevant stakeholders in the information technology (IT) services of the University of Zurich to build scalable solutions that implement best practices—especially regarding database construction, task scheduling, and network structure. We considered a commercial cloud service but concluded that our in-house IT services have several advantages. First, the physical and institutional proximity to the service facilitates a smooth exchange of information. Second, IT services from one's own institution often are better suited than a cloud provider to help with the types of problems researchers typically encounter because they are used to working with researchers, although not necessarily with social scientists. Third, keeping the infrastructure in-house makes it easier to comply with local data-protection rules because IT services have experience with them from other fields (e.g., medical research).

Relying on professional IT services is important to ensure robust implementation of these features and to guarantee maintenance, data security, and data permanence. At the same time, we retain some tasks under our direct control to ensure that we can react as quickly as possible to new needs. Therefore, one question we were confronted with was the division of labor between IT services and the research team. Our setup is shown in figure 2. Tasks carried out by the research team are listed in boxes with a white background; those carried out by IT services have a gray background. IT services address server-related back-end tasks, such as setting up the servers (including operating systems and network infrastructure), combining different machines into computing clusters, troubleshooting, and hardware maintenance and

replacement. The research team does the rest: we write scripts for data collection (e.g., adding new sources), processing (e.g., transforming the raw data in the desired formats), storage (e.g., how data are written into the database and implementing the search functions), and—of course—analysis. This arrangement ensures a robust implementation of all of the features that we need while keeping as much control as possible over the tasks that are related most directly to research.

Because of our infrastructure, we could quickly adjust our data-collection and analysis workflow to study COVID-19. Regarding data collection, we had to adjust settings on an existing server to increase system memory to process the number of tweets, load scripts on the server, and schedule weekly data downloads. Then, we added all Twitter IDs of interest to a script using one of our functions to collect tweets from user timelines. Regarding data analysis, we adapted classifiers that we used for similar purposes, which already were implemented fully within our infrastructure. Specifically, we implemented a keyword search optimized for identifying texts related to COVID-19. It is a simple classification method that works well in this context because the topic is discussed using a unique set of words. However, our infrastructure can handle more sophisticated machine-learning classifiers (Gilardi et al. 2021b). It would have taken more time to implement them, but we could have done it efficiently if needed.

The research infrastructure described in this section allows us to not only continuously collect large amounts of unstructured data; it also is flexible and scalable so that we can quickly adjust or expand data-collection and analysis routines as new research needs arise.

In summary, the research infrastructure described in this section allows us not only to continuously collect large amounts of unstructured data; it also is flexible and scalable so that we can quickly adjust or expand data-collection and analysis routines as new research needs arise.

CONCLUSION: LESSONS LEARNED

This article describes a research infrastructure that we built to address some of the challenges inherent in the study of digital technology and politics. We conclude by discussing the lessons that we learned that, we hope, can be helpful to other researchers pursuing similar goals.

First, automatization is key. Some types of data (e.g., social media) are difficult to obtain retrospectively. Therefore, there are substantial payoffs in setting up an "ingestion system" that collects data continuously. Once such a system is built, new data sources can be added efficiently and on short notice. We recommend adding new sources as soon as they appear to be potentially useful for current or future research. This advice, however, should be balanced against the risk of hoarding data with no clear purpose.

Second, when an efficient ingestion system is up and running, it is tempting to collect data simply because it is easy to do so. This is not a fruitful strategy. Although automatization reduces the marginal costs of data collection, there still are costs. Excessive collection may lead to a "one-size-fits-all" approach that neglects the specificities of individual data sources (e.g., different interactive features). Moreover, data collection risks becoming an end in itself. We recommend defining and regularly updating clear research areas that can prioritize data collection. The best protection against hoarding data that no one will ever use is to be constantly in an exchange with people who are working, or planning to work, with those data.

When an efficient ingestion system is up and running, it is tempting to collect data simply because it is easy to do so. This is not a fruitful strategy. Although automatization reduces the marginal costs of data collection, there still are costs.

Third, some parts of the infrastructure can be outsourced whereas others are better left under the direct control of the researchers. Most universities have a scientific IT service that can host the data, provide servers for computation, and support setting up and maintaining the ingestion system. One advantage of relying on a university's own service—compared to a cloud computing provider such as Amazon Web Services—is that it ensures compliance with local data-protection regulations. Moreover, it is helpful to have a partner onsite with whom a noncommercial relationship can be established. In our experience, university scientific IT services are motivated to collaborate with social scientists because it broadens their scope beyond traditional areas of operation, which may help them to gain additional resources. What is less amenable to outsourcing is the interface between research and IT. We recommend that the team include a social scientist with a strong technical background who can carry out some tasks directly (e.g., adding new sources to the ingestion system and ensuring that everything runs smoothly) as well as communicate effectively with the scientific IT service.

Fourth, to secure funding to set up the infrastructure, it is helpful to embed the infrastructure within a substantive project. Although it depends on the specific context, funding for infrastructure tends to be scarcer than for substantive projects. In terms of budget, one full-time position for a year might be sufficient, plus any additional costs that the university's scientific IT service may charge. After the infrastructure has been set up, a part-time position in many cases is sufficient to keep the system running, especially in smaller countries such as Switzerland.

Fifth, data sharing is not a trivial problem given various legal constraints. The types of data that these infrastructures collect are likely to be subject to restricted sharing due to the terms of service of the platform from which they were collected and in accordance with data-protection regulations. However, most of these issues

arise only when the data leave the research group. Therefore, if the data cannot go to the researcher, we recommend bringing the researcher to the data. In other words, data sharing can take place in the context of joint projects with other researchers. Moreover, this strategy helps to avoid becoming a pure service provider because data sharing is structurally tied to substantive research projects in which the core team members participate.

Sixth, the data collected with the infrastructure may be suited to public outreach, as our COVID-19 example demonstrates. We recommend that researchers develop clear expectations. As in the data-hoarding problem, there are many topics in the political news cycle that are amenable to analysis and visualization. To ensure that this type of work has an impact, we recommend reaching out to journalists before investing too much effort in a specific analysis. Impact depends on established media reporting on the analysis. This outreach is not necessarily a key component of the infrastructure, but it is helpful to increase visibility and, potentially, funding opportunities.

Some parts of the infrastructure can be outsourced whereas others are better left under the direct control of the researchers.

Seventh, competition is generally good but establishing one infrastructure per country may be a sensible strategy—but, of course, it depends on the size of the country. The entire point of the infrastructure is to avoid wasting resources in duplicating data-collection efforts. In this context, collaboration is more promising than competition.

The interplay between digital technology and politics is one of the most pressing challenges that our societies are facing. Research on this issue faces several specific constraints; we argue that building a dedicated research infrastructure is an important step to overcome them. We hope that our experience discussed in this article is helpful to other researchers pursuing similar goals.

ACKNOWLEDGMENTS

This project received funding from the Swiss National Science Foundation (Grant No. 10DL11_183120) and the European Research Council under the EU's Horizon 2020 research and innovation program (Grant Agreement No. 883121) and was supported by the Digital Society Initiative of the University of Zurich. We are grateful to Avi Bernstein for his support in the early stages of this project. We also thank Alix d'Agostino, Hannah Stenzler, and Rocco Leonardi for research assistance, and the Science IT team at the University of Zurich—especially Pim Witlox—for technical support.

DATA AVAILABILITY STATEMENT

Replication materials are available on Harvard Dataverse at https://doi.org/10.7910/DVN/BSQWOU. ■

NOTES

1. See https://compsocialscience.github.io/summer-institute.
2. See https://digdemlab.io.

REFERENCES

Bruns, Axel. 2019. "After the 'APIcalypse': Social Media Platforms and Their Fight Against Critical Scholarly Research." *Information, Communication & Society* 22 (1): 1544–66.

de Vries, Erik, Martijn Schoonvelde, and Gijs Schumacher. 2018. "No Longer Lost in Translation: Evidence That Google Translate Works for Comparative Bag-of-Words Text Applications." *Political Analysis* 26 (4): 417–30.

European Data Protection Supervisor. 2020. "A Preliminary Opinion on Data Protection and Scientific Research." https://edps.europa.eu/data-protection/our-work/publications/opinions/preliminary-opinion-data-protection-and-scientific_en.

Freelon, Deen. 2018. "Computational Research in the Post-API Age." *Political Communication* 35 (4): 665–68.

Gilardi, Fabrizio, Lucien Baumgartner, Clau Dermont, Karsten Donnay, Theresa Gessler, Maël Kubli, Lucas Leemann, and Stefan Müller. 2021a. "Replication Data for: Building Research Infrastructures to Study Digital Technology and Politics: Lessons from Switzerland." Harvard Dataverse. DOI:10.7910/DVN/BSQWOU.

Gilardi, Fabrizio, Theresa Gessler, Maël Kubli, and Stefan Müller. 2021b. "Social Media and Political Agenda Setting." *Political Communication* Online First: 1–22. DOI:10.1080/10584609.2021.1910390.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. "Fake News on Twitter During the 2016 US Presidential Election." *Science* 363 (6425): 374–78.

Guess, Andrew, Jonathan Nagler, and Joshua Tucker. 2019. "Less Than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook." *Science Advances* 5 (1): eaau4586.

Jungherr, Andreas, Gonzalo Rivero, and Daniel Gayo-Avello. 2020. *Retooling Politics: How Digital Media Are Shaping Democracy*. New York: Cambridge University Press.

King, Gary, and Nathaniel Persily. 2020. "A New Model for Industry: Academic Partnerships." *PS: Political Science & Politics* 53 (4): 703–709.

Levi, Margaret, and Betsy Rajala. 2020. "Alternatives to Social Science One." *PS: Political Science & Politics* 53 (4): 710–11.

Salganik, Matthew J. 2017. *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

Zubiaga, Arkaitz. 2018. "A Longitudinal Assessment of the Persistence of Twitter Datasets." *Journal of the Association for Information Science and Technology* 69 (8): 974–84.