Running head: OPEN COMPUTATIONAL SOCIAL SCIENCE

Open Computational Social Science

Jan G. Voelkel[a]

https://orcid.org/0000-0003-4275-9798

Jeremy Freese[a]

https://orcid.org/0000-0002-9451-6432

[a]: Department of Sociology, Stanford University, Stanford, USA

Author Notes

Please cite this paper as follows:

Voelkel, J. G., & Freese, J. (2022). Open computational social science. In U. Engel, A. Quan-Haase, S. Xun Liu, & L.E. Lyberg (Eds.), *Handbook of computational social science. Volume 1 theory, case studies and ethics*. Routledge. ISBN 9780367456528.

Please note that this is the non copy edited version.

We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Jan G. Voelkel, Department of Sociology, Stanford University, Stanford, CA 94305, USA. Contact: jvoelkel@stanford.edu.

Open Computational Social Science

The rise of computational social science is a recent major development in the social sciences. The possibility of using computational techniques to analyze large datasets has provided researchers from many different fields with new tools to solve puzzles once thought impossible to study. Furthermore, computational social science often uses social media or other digital data, paving the way for new theoretical insights into societal phenomena (Lazer et al., 2009; Edelmann, Wolf, Montagne, & Bail, 2020). The prospects of how computational social science will change and improve social scientific answers to important theoretical and practical problems are exciting.

At the same time, the data typically used by computational social scientists also create challenges to its openness. Open Science is an umbrella term that summarizes a movement of researchers who aim to improve the fundamental features of science, including accessibility, transparency, rigor, reproducibility, replicability, and accumulation of knowledge (Crüwell et al., 2019). Insights from the Open Science movement can help computational social science build on praiseworthy developments in the field to institutionalize transparency and reproducibility. In the following subsections, we introduce four principles of Open Science, explain why each is important, describe some of the challenges of their implementation, and suggest how computational social scientists could address these challenges. While these four Open Science principles are well known and we think they are the most relevant for computational social scientists, we do not claim that this way of specifying principles is either exhaustive or definitive of an evolving science (for alternative classifications see for example Crüwell et al., 2019; Munafo et al., 2017).

Open Practices

Open Practices refer to honest and transparent specification of all steps in the data processing and data analysis. One example for computational social scientists is posting their study's materials and the code used to process and analyze the data. Computational social scientists sometimes test hypotheses they had before starting their data analyses and sometimes build hypotheses based on their data analyses. Here, an important Open Practice is clearly demarcating exploratory and confirmatory parts of the research process, so that findings generated post hoc are not presented as testing pre-existing hypotheses.

Open Practices are an important goal for (social) science to increase replicability, i.e. the replication of prior findings with "new data". It is well known that many findings in the social sciences cannot be replicated (Anderson et al., 2016; Baker, 2016; Camerer et al., 2016; Camerer et al., 2018; Ioannidis, 2005; Open Science Collaboration, 2015; Shiffrin, Boerner, & Stigler, 2018). Many explanations for these low replication rates have been discussed in the literature, including publication bias (e.g., Bakker, van Dijk, & Wicherts, 2012; Ioannidis, Munafo, Fusar-Poli, Nosek, & David, 2014), questionable research practices (e.g., John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011), statistical errors (Gelman & Loken, 2014; Greenland et al., 2016), procedural errors in replication (Gilbert, King, Pettigrew, & Wilson, 2016; Loken & Gelman, 2017), contextual sensitivity of findings (R. Klein et al., 2018; Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016), and lack of a cumulative theoretical framework (Muthukrishna & Henrich, 2019).

Open Practices are also an important goal for (social) science to increase reproducibility, i.e. the verification of prior findings from the same data and code (Freese & Peterson, 2017). Although submissions to journals often require authors to agree to share their data and materials upon request, many researchers either cannot or choose not to provide the necessary files when requests are actually made (Wicherts, Borsboom, Kats, & Molenaar, 2006). This is particularly problematic because evidence indicates that the less willing researchers are to share their data, the more likely it is that the original results cannot be reproduced (Wicherts, Bakker, & Molenaar, 2011).

Open Practices are intended to reduce the negative effect of questionable research practices on both replicability and reproducibility. Other researchers can only reproduce and retrace the original results if the original authors have published their materials and data and provided a comprehensible and detailed script of their analysis. Furthermore, other researchers can only estimate the extent to which results are replicable if all steps in the data processing and data analysis process were described honestly and transparently. Otherwise, the apparent evidence might be undermined by the influence of invisible questionable research practices. That is, researchers might have taken advantage of chance variability or other vicissitudes in the data to cherry-pick the strongest-appearing set of results (Simmons, Nelson, & Simonsohn, 2011). Viewed in this way, questionable research practices may be understood as a motivated form of overfitting data, with the subsequent failure to replicate being analogous to the prediction error that results from overfitting.

The most prominent solution to limit questionable research practices is preregistration. Preregistration entails the specification of hypotheses, exclusion rules, data processing, and statistical models used to evaluate the hypotheses before the collection of the data (Nosek, Ebersole, DeHaven, & Mellor, 2018; van 't Veer & Giner-Sorolla, 2016; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Such preregistrations are then time-stamped and published so that other researchers can examine whether the actual data analysis corresponds to the preregistered data analysis. Preregistration works particularly well for simple designs, such as experiments, in which researchers know a priori how they will analyze their data.

A main challenge for institutionalizing Open Practices among computational social scientists is that computational techniques often work inductively and can often be impossible to specify a priori. Luckily, computational social scientists who work with prediction tasks on very large datasets are already familiar with an arsenal of practices to reduce overfitting. Central among them is dividing data into a training set and test (or validation) set (e.g., Anderson & Magruder, 2017; Egami, Fong, Grimmer, Roberts, & Stewart, 2018). The training set is used for the *exploratory* work of developing a model, while the test set is used for the *confirmatory* work of determining how well the model predicts the outcome.

Usually, when researchers do projects that involve this sort of cross-validation, they have access in principle to the test set while they are fitting data on the training set. Often the researcher has no incentive to overfit the data — so cross-validation is fully in their interests — and, when they do, the researcher is trusted not to iterate predictions over the test set to improve the apparent performance of their model. However, when incentives to obtain certain (e.g. statistically significant or highly predictive) results are strong, access to the test dataset before committing to a model is recognized as unviable.

One solution is that a trusted third-party controls access to the test dataset until researchers publicly commit to a model. For example, when sites like Kaggle host competitions with rewards for the best-predicting model, the data that will be used to determine the winner is not available to the competitors. Salganik et al. (2020) hosted a similar competition in which researchers were tasked with submitting predictions about outcomes of respondents in the Fragile Families dataset, and again the structure was that competitors would submit predictions that would be evaluated against a sequestered portion of the original data to which no competitor had access. Outside of competitions, it has been suggested that researchers could send their data to a trusted third-party to randomly divide data into a training set and a withheld test set (Fafchamps & Labonne, 2017). One could imagine ways that this could be institutionalized, for example by being run by a professional association, but we are not aware of any efforts to do so at present.

Cross-validation and pre-registration can be used in tandem. van Loon and Freese (2019) divided existing data into a training and validation set for the purposes of evaluating the success of their model for predicting questionnaire-based ratings of the affective connotations of concepts from word-embeddings. Afterward, they used the same model to generate predictions for a new set of concepts and posted these predictions publicly prior to collecting the new data used to evaluate them. A different idea for combining cross-validation and pre-registration is proposed by Anderson and Magruder (2017), who note that one problem with pre-registration is that it can discourage including more adventurous supplementary hypotheses, as the addition of these hypotheses may imply expectations to correct for multiple comparisons which may be seen as reducing the chances of success for the main hypotheses. They propose using pre-registration and the full sample for hypotheses that are regarded either as especially central or posed with high confidence by the researcher, and then using the split-sample approach to explore more speculative ideas.

Whenever researchers do preregister their projects, the preregistrations should be highly specific and followed precisely in order to be a solution against questionable research practices. Proponents of preregistration have acknowledged that while there are strong theoretical arguments why preregistrations should increase replicability, there is a lack of demonstrated evidence for this claim (Nosek, Ebersole, DeHaven, & Mellor, 2018). Recent research suggests that, in practice, preregistrations often remain too vague to be an effective shield against questionable research practices (Veldkamp et al., 2018). Furthermore, it is unclear to what extent adherence to the preregistration is enforced in the review process.

Overcoming some of the limitations of preregistration are registered reports (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014; Hardwicke & Ioannidis, 2018). Registered reports are protocols that are submitted for peer-reviewed evaluation before the data is collected, with the result provisionally accepted for publication as long as the registered methodology is followed. Reviewers thus have the opportunity to influence both the theory and the analysis, yet without increasing incentives for post-hoc motivated reasoning. Furthermore, registered reports have an explicit additional submission stage after the data was collected in which it is checked whether the theory and methods section still match the initial submission. Some registered reports have also included a “crowdsource” component, where the protocol is made public along with an invitation for other investigators also to collect data following the registration in exchange for joint-authorship, so that the resulting paper includes many simultaneous replications and much more strongly powered findings.

While we have focused on how computational social science can benefit from Open Science, there is also great potential for contributions from computational social science to increasing replicability. For example, although it is well-known that a study with larger effect sizes and smaller p-values is more likely to be replicated than a study with smaller effect sizes and p-values close to .05, we know relatively little about the relationship between the other, many, many features of research articles and replicability. We think that computational social scientists could make use of text-analysis and machine learning to identity which features do and which features do not predict replicability. Researchers have already started such projects. As one, the DARPA program Systematizing Confidence in Open Research and Evidence (SCORE) seeks to develop automatic tools to assess social science claims.

Open Data

Open Data is data that is publicly available. As one example, large-scale representative surveys such as the General Social Survey make their data available so that everyone can use the data. As another, the American Journal of Political Science requires authors to publish all files necessary to reproduce an article's results on the Harvard Dataverse Network, unless there is a specific reason why a study's data cannot be shared.

Open Data is an important goal for computational social science because it is a prerequisite for establishing the reproducibility of a finding, i.e. the verification of prior findings from the same data and code (Freese & Peterson, 2017). It is unclear if most computational research meets even this minimal standard. For example, Dacrema, Cremonesi, and Jannach (2019) sought to reproduce research that involved using deep-learning techniques on a certain type of recommendation task. They considered a study to be reproducible if at least one dataset from the paper was available and if the paper's results from these data could be obtained with no more than minimal modification to source code. Of the 18 papers they identified from 4 conference series, the results of only 7 were reproducible by these criteria.

Another benefit of Open Data is that investigators can interrogate findings directly and decide if the conclusions drawn by investigators are warranted. Sometimes this results in others realizing things about the data that those who originally published findings missed. For example, Back, Küfner, and Egloff (2010) analyzed pager transmissions from New York City after 9/11/2001 and reported a dramatic increase in anger-associated messages in the latter hours of the day. The data were publicly available not because of any Open Data initiative, but because they were part of a WikiLeaks leak. Looking at the same data, Pury (2011) demonstrated that the increase in "anger" was entirely driven by nearly 6,000 identically-formatted automated messages notifying a single pager about a "CRITICAL" computer system problem, with "critical" being counted as an anger-associated word.

The community benefits of Open Data are not limited to the ability to better evaluate the findings from a particular study. The accumulation of Open Data in a research community can also allow large-scale insights that were unavailable when investigators only have their own datasets to work with. In neuroimaging, overdue data-sharing initiatives finally allowed researchers to rigorously interrogate the performance of techniques with suitable amounts of real data that had previously only been validated with simulated data. Eklund, Nichols, and Knutsson (2016) demonstrate that three leading software packages for working with neuroimages all did not properly correct for multiple comparisons, so that false-positive results between purely random groupings of scans were obtained up to 70% of the time instead of the expected 5%. The problem has led to an unknown—but substantial—number of erroneous findings being published, and could have been detected earlier if the neuroscience community had been more proactive about making data available to other investigators.

Fully Open Data also allows other researchers to make discoveries with data that were unanticipated by the original investigators, providing a plain public good for the research community. For example, a project on morality in everyday life (Hofmann, Wisneski, Brandt, & Skitka, 2014) created a publicly available dataset using experience sampling methodology to assess moral and immoral experiences in people's ordinary lives. A few years later, Crockett (2017) used the same data to explore whether immoral acts create more moral outrage online than in-person. Other initiatives collect and organize similar datasets. For example, the "experimentdatar" data package contains large-scale experiments that are suitable for causal inference techniques via machine learning (see also https://github.com/itamarcaspi/experimentdatar). New analyses on shared datasets may reduce the inefficient collection of similar and expensive datasets by many different researchers.

The crucial challenge for Open Data is that some data might not be shareable with others. In text analysis, pervasive examples include data protected by copyright or terms of service agreements. Legal boundaries are here contested, but one interpretation for the United States would be that analyzing copyright-protected materials can be considered permissible under "Fair Use" doctrine, while making that raw data publicly available to others is not. The Corpus of Contemporary American English includes considerable data under copyright, and degrades the data to count as "Fair Use": the regular distribution limits users to small snippets obtained via a fixed number of queries per day, and the downloadable version eliminates 5% of segments of text throughout.

Data may also be obtained from sources who are only willing to allow those investigators to use the data. For example, to evaluate hypotheses about how employees adapt to organizational culture, Srivastava and colleagues (2018) were able to obtain a complete corpus of internal e-mails among employees of a midsized company over a six-year period. It is remarkable enough when investigators are able to convince gatekeepers to allow them to use such data, and it is hardly surprising that gatekeepers would balk at permitting any broader dissemination of it.

In addition, data sharing may be restricted to protect the confidentiality and privacy of persons whose information is included in the data. The arsenal of data and tools available to computational social scientists has exacerbated some of the worries around data availability, as it is unclear what combination of information in a dataset can be combined with information elsewhere to deduce the identities of individuals. When New York City released “anonymized” data on 173 million taxi trips in response to a Freedom of Information Law request, a software developer figured out how to reverse hashed information about the taxi medallion number, and then an enterprising graduate student determined that this could be used to match the data to celebrity passengers from paparazzi photos of them getting in and out of taxis, allowing reports of how much those celebrities apparently tipped (Trotter, 2014). One need not be a celebrity to be reidentified using external data: Sweeney, Abu, and Winn (2013) report that a sizable percentage of sample participants with publicly available data in the Personal Genome Project could be identified by cross-referencing demographic variables in the dataset with voting and other public records. The US Census Bureau is controversially doing more to coarsen the data it releases publicly because of work indicating that Census responses can be sometimes deduced by cross-referencing Census tables with commercially-available data on individuals (Abowd, 2018).

Due to these concerns, the ideal of Open Data for every published study is unrealistic. Social science would be poorer if social science could only use data that may be shared with everyone. Nonetheless, journal policies that require researchers to either publicly share their data or require a clear statement why data cannot be shared that is published with the paper are instrumental to make progress towards the ideal of Open Data as much as possible. The inability for data to be shared is a flaw in a research design, about which judgments must be made by editors and audiences about how to weigh that flaw versus the strengths of the project. To make this judgment, of course, papers for which data cannot be shared need to be explicit about that fact, and it is distressing how many computational social science papers leave the reader to guess about the availability of the data on which findings are based. Throughout science, there has been a re-appraisal of the extent to which optimal data availability in research communities can be attained by professional norms or exhortation alone. Even in areas where codified norms about sharing data with qualified investigators after publication exist, studies of compliance with requests have led to disappointing results (see Christensen, Freese, & Miguel, 2019: 174-176 for a review). As a result, many have concluded that sharing standards are something that journals must enforce, by requiring that data that can be made broadly available be deposited in a third-party repository at the time of publication and that data that cannot be made broadly available include an explicit disclosure.

Alongside this, we think that computational social scientists should not only follow journals' requirements but also actively aim to share their data whenever possible. Computational social scientists need to engage with ethical standards and evolving legal frameworks to make their decision to share or not share data (Hollingshead, Quan-Haase, & Chen, 2021). If researchers own their data, they can make anonymized versions of their data files available via a third-party repository like the Open Science Framework or Databrary. Anonymization of the data refers to the removal or recoding of variables that contain identifiable information. It is often overlooked that identifiable information does not only include obviously sensitive information such as names or email addresses but also combinations of common demographics. For example, it might be problematic to share the zip code and the birth date of participants because there might be certain birth dates that are unique in certain areas. If these variables are instrumental for the data analysis, researchers may recode the variables to avoid rare cases (ideally such decisions are made *before* data analysis). We advise researchers to always include a readme file that explains the steps of anonymization, ideally with a link to the analysis code that provides full details of how the anonymization was conducted. In order to minimize the long-term costs of data sharing, labs and departments should develop data management plans that standardize practices and reduce the necessity of individual decision making (Levenstein & Lyle, 2018). If researchers cannot share their data, we suggest that they make a readme file available that explains (a) how they accessed their data and (b) how other researchers can access the same dataset.

For computational social science, reproducibility demands not just sharing data but also sharing the code used to derive results from those data. Analysis code should contain a sufficient number of comments to explain each step and ideally be accessible in common formats. For example, R code can be transformed into Markdowns and then knitted to html or pdf files. Versions of any external dependencies should be tracked and documented, so that version differences can be ruled out as an explanation for any failure to reproduce results by others. Standard procedures within labs or departments can be developed for replication packages and should be enforced during the training of junior scholars (e.g., for dissertations). While this sort of work may often strike novices as cumbersome, the benefits of Open Science coding practices are realized not only by others attempting to reproduce work down the line, but often by collaborators and one's own future self when one later returns to the code.

Finally, we suggest that computational social scientists make use of openly shared data. As explained above, Open Data allows different forms of reproducibility analyses. Such analyses are important to ensure the validity and robustness of prior research. However, Open Data can also be used for original analyses and research questions.

Open Tools

Open Tools refer to software that is both free and has its source code available to others. The most prominent tools of computational social science – Python, R, Jupyter Notebooks, Git, Atom, Visual Studio Code – are all open software projects.

The availability and usage of Open Tools have many advantages for computational social science. First, software for computational research, in combination with high-quality educational material, is needed to allow as many researchers as possible to learn, understand, and use computational social science techniques. Free software is particularly suited for a quick diffusion because it does not impose monetary restrictions for the access of the software (von Krogh &

von Hippel, 2006). It also eliminates headaches from coordinating software licenses among collaborators or across machines. Second, source code availability can attract widespread collaborations of volunteers who can quickly improve and innovate the software (von Krogh & von Hippel, 2006). Finally, Open Tools also provide for the possibility of a layer of "peer review" to vet the accuracy of algorithms and allows researchers the possibility of averring that every step in a project's computational workflow is in principle available for inspection.

The main challenge for the provision of Open Tools is that there are no direct monetary incentives for the individual to provide software for the community at large. In this respect, open software is a classic public goods or collective action problem (Coleman, 1986), and of course most users of open software are effectively "free riders" in the sense that they reap the benefits of its availability without paying the costs for its development. Happily, however, compared to many real-life public goods problems, open software does not need a substantial proportion of users to contribute software in order to thrive. Meanwhile, the success of open source stems from a complex array of different paths of reward, monetary and otherwise, that have evolved to allow software to be developed and maintained without being proprietary. In academia, counting open software work as scientific contributions helps leverage the reward structure of academia to support the development of openly available research tools. The scientific research community has long benefitted from a strong online community that works admirably hard on providing and supporting open software. This community has effectively solved the collective action problem of providing Open Tools in computational social science.

However, not all Open Tools are good tools. Open software solutions range from venerable tools with large, active, and sophisticated user communities to obscure packages provided by creators with unknown skill and commitment to testing. A vexing scenario for researchers using any software tool is that it is not actually doing what the researcher thinks it is, creating the possibility for researchers to receive misunderstood or simply erroneous results. Researchers are thus encouraged to do what they can to verify results with simple examples when possible, and novice users may wish to be especially wary of unusual or poorly documented tools.

The next challenge is that users often need support for learning how to use software. Computational social science deserves credit not just for the development of open research tools but also the openness of teaching materials. For example, the Russell Sage Foundation has sponsored the Summer Institute in Computational Social Science. The Institute has developed a curriculum that allows for it to be taught at satellite locations alongside its primary location, and it has also made lectures available as videos and other materials as well-annotated examples via GitHub. Providing teaching materials in this way leverages the technical acumen of computational social scientists to enable the field to offer state-of-the-art training to the broadest possible audience of aspiring researchers.

A related challenge is that even skilled users often need continuing support to deal with more complex challenges. Indeed, among the conventional arguments for using proprietary software over open software is the former offering more thorough documentation and more accessible technical support when users face problems. Especially for prominent open software solutions for computational social science, however, this has been largely neutralized by the development of a vast array of sites providing online answers to questions, like Stack Exchange, with different ways of crowdsourced curation of the effectiveness of different answers. Search engines are now the first strategy employed for all manner of technical queries for users of either proprietary or open software alternatives, and the large communities surrounding prominent open-source platforms often have produced much more high-quality content to help users than experts employed by a proprietary package.

Open Access

Open Access refers to the "free, public availability of a research product on the internet for distribution and re-use with acknowledgement" (Crüwell 2019, p. 7). Open Access now has different models. The model in which journal articles are made freely available is known as “Gold” Open Access. “Green” Open Access, meanwhile, refers to an author engaging in some kind of self-archiving alongside paywalled journal publication. Journals can prohibit even this as a condition of publication, but the appetite for doing so has waned considerably, especially now that various funding bodies require their investigators to provide some sort of open version of publications.

Open Access is an important goal for computational social science for several reasons. First, Open Access articles can be readily obtained both by non-researchers as well as researchers from countries and institutions with less resources (Tennant et al., 2016). Thus, publishing with Open Access increases and democratizes access to knowledge, which may in turn facilitate the more widespread adoption of computational techniques in the social sciences. Second, publishing with Open Access (more specifically, with a Creative Commons Attribution license CC-BY) enables other researchers, including computational social scientists, to employ automated text- and data-mining tools for analyzing scientific articles (Tennant et al., 2016). Third, even for those unmoved by the community appeals of openness, increasing the exposure and accessibility to one’s work provides a professional advantage. Making working papers available allows even faster transmission of knowledge, and also helps establish priority for findings in fast-moving fields. Finally, a happy side effect for the authors of Open Access articles is that Open Access articles are cited more often (Piwowar et al., 2018; Tennant et al., 2016).

There are two potential downsides that have been discussed in the context of Open Access. First, instead of publishers charging readers or readers’ institutions, Open Access publishers often charge the researchers who wrote the article (Tennant et al., 2016). This makes it more difficult for researchers with less resources to publish with Gold Open Access. However, Green Open Access (i.e. self-archival of papers) provides a cost-free model of Open Access that every researcher can adopt. Granted, it should be noted that some journals do not approve of submitting articles that have been made available as preprints; journal policies can be checked via SHERPA/RoMEO (http://www.sherpa.ac.uk/romeo/index.php). In addition, we urge (computational) social scientists to make sure that they choose a license for their paper that allows for automated text and data mining. An excellent resource for self-archiving papers is ArXiv (https://arxiv.org/), a widely used repository of preprints with different choices for licenses. Furthermore, while author charges are still the most common way that Gold Open Access is funded, there are journals in which charges are effectively underwritten by a benefactor, and some journals are piloting a model in which a journal is made open if and only if it sells a certain level of institutional subscriptions.

Second, and related to the first concern, publishers may exploit the Open Access model. If authors pay for publishing in an Open Access model, publishers maximize their income by maximizing the number of papers they publish. Thus, publishers have an incentive to reduce the rigor of the review process and publish every submitted article, undermining quality (Tennant et al., 2016). Researchers should carefully check the reputation of their target journal before submission. A helpful guide to identify predatory journals is available at https://predatoryjournals.com.

Conclusion

In the current chapter, we have discussed how computational social scientists can achieve high standards of accessibility, transparency, replicability, and reproducibility by adhering to four Open Science principles, Open Practices, Open Data, Open Tools, and Open Access. Specifically, we suggest the following Open Science to-do list for computational social scientists who aim to publish a paper:

Open Practices

1. Be transparent about all steps in the data processing and analyses processes.
2. If you want to test a hypothesis and know how you want to test it, preregister your hypothesis and your analysis plan.
3. If you do not have hypotheses or do not know how to test them, divide your dataset into an exploratory part (for exploratory analyses) and a confirmatory part (for confirmatory analyses). Limit your own access to the confirmatory dataset until you enter the confirmatory stage.
4. Consider opportunities to write papers as registered reports or participate in crowdsourced registered reports projects.
5. Use computational techniques to provide insights into the features of a study or a paper that predict replicability.

Open Data

1. If possible, share your own data.
2. If possible, use a licence that allows other researchers to reuse your data.
3. If you cannot share your data yourself, explain how the data can be accessed, and if it cannot be shared at all, indicate this clearly in the paper.

Open Tools

1. Use open-source software tools.
2. Share your code, software, and educational materials so that other people can learn computational techniques.

Open Access

1. Publish your paper with Open Access, for example by posting it on a repository of preprints
2. Use a license that allows for automated text- and data-mining.

References

Abowd, J. M. The U.S. Census Bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2867-2867). doi:10.1145/3219819.3226070.

Anderson, M. L. & Magruder, J. (2017). *Split-sample strategies for avoiding false discoveries* (NBER Working Paper No. 23544). Retrieved from National Bureau of Economic Research website: http://www.nber.org/papers/w23544

Anderson, C. J., Bahnik, S., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., …, & Zuni, K. (2016). Response to Comment on "Estimating the reproducibility of psychological science". *Science, 351*(6277), 1037. doi:10.1126/science.aad9163

Back, M. D., Küfner, A. C. P., Egloff, B. (2010). The emotional timeline of September 11, 2001. *Psychological Science, 21*(10), 1417–1419. doi:10.1177/0956797610382124

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*(7604), 452-454. doi:10.1038/533452a

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543-554. doi:10.1177/1745691612459060

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433-1436. doi:10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and

Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637-644. doi:10.1038/s41562-018-0399-z

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of" playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*, *1*(1), 4-17. doi:10.3934/Neuroscience2014.1.4

Christensen, G., Freese, J., & Miguel, E. (2019). *Transparent and Reproducible Social Science Research: How to Do Open Science.* Berkeley and Los Angeles, CA: University of California Press.

Coleman, J. S. (1986). *Individual Interests and Collective Action: Selected Essays*. Cambridge University Press.

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769-771. doi:10.1038/s41562-017-0213-3

Crüwell, S., van Doorn, J., Etz, A., Makel, M. C., Moshontz, H., Niebaum, J. C., Orben, A., Parsons, S., & Schulte-Mecklenbeck, M. (2019). Seven easy steps to Open Science. *Zeitschrift für Psychologie*, 227, 237-248. doi:10.1027/2151-2604/a000387

Dacrema, M. F., Cremonesi, P., & Jannach, D. (2019). Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems* (pp. 101-109). doi:10.1145/3298689.3347058

Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational social science and sociology. *Annual Review of Sociology*, *46*, 61-81. doi:10.1146/annurev-soc-121919-054621

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., & Stewart, B. M. (2018). *How to make causal inferences using texts*. Retrieved from https://arxiv.org/abs/1802.02163

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, *113*(28), 7900-7905. doi:10.1073/pnas.1602413113

Fafchamps, M., & Labonne, J. (2017). Using split samples to improve inference on causal effects. *Political Analysis*, *25*(4), 465-482. doi:10.1017/pan.2017.22

Freese, J., & Peterson, D. (2017). Replication in social science. *Annual Review of Sociology*, *43*, 147-165. doi:10.1146/annurev-soc-060116-053450

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460-466. doi:10.1511/2014.111.460

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on “Estimating the reproducibility of psychological science”. *Science*, *351*(6277), 1037-1037. doi:10.1126/science.aad7243

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337-350. doi:10.1007/s10654-016-0149-3

Hardwicke, T. E., & Ioannidis, J. P. (2018). Mapping the universe of registered reports. *Nature Human Behaviour*, *2*(11), 793-796. doi:10.1038/s41562-018-0444-y

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, *345*(6202), 1340-1343. doi:10.1126/science.1251560

Hollingshead, W., Quan-Haase, A., & Chen, W. (2021). Ethics and privacy in computational social science: A call for pedagogy. In Engel, U., Quan-Hasse, A., Liu, S. X., & Lyberg, L., *Handbook of Computational Social Science*. Routledge. Taylor & Francis Group.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235-241. doi:10.1016/j.tics.2014.02.010

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524-532. doi:10.1177/0956797611430953

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., ... & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443-490. doi:10.1177/2515245918810225

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... & Jebara, T. (2009). Computational social science. *Science*, *323*(5915), 721-723. doi:10.1126/science.1167742

Levenstein, M. C., & Lyle, J. A. (2018). Data: Sharing is caring. *Advances in Methods and Practices in Psychological Science*, *1*(1), 95-103. doi:10.1177/2515245918758319

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584-585. doi:10.1126/science.aal3618

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221-229. doi:10.1038/s41562-018-0522-1

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600-2606. doi:10.1073/pnas.1708274114

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi:10.1126/science.aac4716

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., ... & Haustein, S. (2018). The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, *6*, e4375. doi:10.7717/peerj.4375

Pury, C. L. (2011). Automation can lead to confounds in text analysis: Back, Küfner, and Egloff (2010) and the not-so-angry Americans. *Psychological Science*, *22*(6), 835-836. doi:10.1177/0956797611408735

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., ... & Datta, D. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*(15), 8398-8403. doi:10.1073/pnas.1915006117

Shiffrin, R. M., Börner, K., & Stigler, S. M. (2018). Scientific progress despite irreproducibility: A seeming paradox. *Proceedings of the National Academy of Sciences*, *115*(11), 2632-2639. doi:10.1073/pnas.1711786114

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366. doi:10.1177/0956797611417632

Srivastava, S. B., Goldberg, A., Manian, V. G., & Potts, C. (2018). Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations. *Management Science*, *64*(3), 1348-1364. doi:10.1287/mnsc.2016.2671

Sweeney, L., Abu, A., & Winn, J. (2013). *Identifying participants in the personal genome project by name (a re-identification experiment)*. Retrieved from https://arxiv.org/abs/1304.7605

Tennant, J. P., Waldner, F., Jacques, D. C., Masuzzo, P., Collister, L. B., & Hartgerink, C. H. (2016). The academic, economic and societal impacts of Open Access: an evidence-based review. *F1000Research*, *5*, 632. doi:10.12688/f1000research.8460.3

Trotter, J. K. (2014, October 23). Public NYC Taxicab Database Lets You See How Celebrities Tip. *Gawker*. Retrieved from https://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, *113*(23), 6454-6459. doi:10.1073/pnas.1521897113 van Loon, A. & J. Freese (2019). Can we distill fundamental sentiments from natural language use? Evaluating word embeddings as a complement to survey-based ratings of affective meaning. Retrieved from https://osf.io/preprints/socarxiv/r7ewx/

van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2-12. doi:10.1016/j.jesp.2016.03.004

Veldkamp, C. L. S., Bakker, M., Van Assen, M. A., Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., & Wicherts, J. (2018). *Ensuring the Quality and Specificity of Preregistrations*. Retrieved from https://psyarxiv.com/cdgyh

Von Krogh, G., & Von Hippel, E. (2006). The promise of research on open source software. *Management Science*, *52*(7), 975-983. doi:10.1287/mnsc.1060.0560

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632-638. doi:10.1177/1745691612463078

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726-728. doi:10.1037/0003-066X.61.7.726

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PloS One*, *6*(11), e26828. doi:10.1371/journal.pone.0026828