LaTeX Template for BIDS Best Practices Reports

A Report from the Berkeley Institute for Data Science's
Best Practices in Data Science Series

R. Stuart Geiger[1†*], **Dan Sholler**[1,10†], **Aaron Culich**[3‡], **Ciera Martinez**[1,4‡], **Fernando Hoces de la Guardia**[5‡], **François Lanusse**[1,8,9‡], **Kellie Ottoboni**[1,2‡], **Marla Stuart**[1,6‡], **Maryam Vareth**[1,7‡], **Nelle Varoquaux**[1,2‡], **Sara Stoudt**[1,2‡], **Stéfan van der Walt**[1‡]

1 Berkeley Institute for Data Science, University of California, Berkeley
2 Department of Statistics, University of California, Berkeley
3 D-Lab, University of California, Berkeley
4 Department of Molecular and Cell Biology, University of California, Berkeley
5 Berkeley Initiative for Transparency in the Social Sciences, University of California, Berkeley
6 School of Social Welfare, University of California, Berkeley
7 Department of Radiology and Biomedical Imaging, University of California, San Francisco
8 Berkeley Center for Cosmological Physics, University of California, Berkeley
9 Foundations of Data Analysis Institute, University of California, Berkeley
10 rOpenSci

*Corresponding author: stuart@stuartgeiger.com

†These authors contributed equally to this work

‡These authors contributed equally to this work, order is alphabetical

Abstract:

This is the LaTeX template for the reports produced by the BIDS Best Practices in Data Science group, which has been modified from a freely-licensed template released by the editors of eLife (*eLife Editorial Team, 2017*). You should feel free to further adapt and modify this template for your own work, as you see fit. This field is for the abstract, which can be a decent length, but not so long — especially if you have a lot of authors. The front page also has space for a "recommended citation" element below, which is currently produced manually. The remainder of this article and the author list have been taken from our first publication, "Challenges of Doing Data-Intensive Research in Teams, Labs, and Groups: Report from the BIDS Best Practices in Data Science Series."

Introduction

This paper is a summary of the first meeting of the Best Practices lunch and discussion series, sponsored by the Berkeley Institute for Data Science (BIDS), in which we bring people together from across the UC-Berkeley campus and beyond to discuss a particular challenge or issue in doing data-intensive research. The goal of the series is to informally share experiences and ideas on how to do data science well (or at least better) from many disciplines and contexts. The topic for this week was doing data-intensive research in teams, labs, and other groups. For this first meeting, we focused on just identifying and diagnosing the many different kinds of challenges. In future meetings, we will dive deeper into some of these specific issues and try to identify best practices for dealing with them.

We prepared for this series by reviewing many of the papers and series around "best practices" in scientific computing (e.g. *Wilson et al., 2014*; *Noble, 2009*), "good enough practices" (*Wilson et al., 2017*), and PLOS Computational Biology's "ten simple rules" series (e.g. *Sandve et al., 2013*; *Goodman et al., 2014*; *Perez-Riverol et al., 2016*). There is also an extensive literature relevant to collaboration and teamwork in data-intensive groups, including work in ethnography of scientific labs (e.g. *Traweek, 1992*), library and information science (e.g. *Borgman, 2015*), team science (e.g. *National Research Council, 2015*), and industry guides (e.g. *Patil, 2011*). We also see this series as a successor to the collection of case studies in reproducible research published by several BIDS fellows (*Kitzes et al., 2018*). One reason we chose to identify issues with doing data science in teams and groups is because many of us felt like we understood how to best practice data-intensive research individually, but struggled with how to do this well in teams and groups.

Compute and data challenges

Getting on the same stack

Some of the major challenges in doing data-intensive research in teams is around technology use, particularly in using the same tools. Today's computational researchers have an overwhelming number of options to choose in terms of programming languages, software libraries, data formats, operating systems, compute infrastructures, version control systems, collaboration platforms, and more. One of the major challenges often faced is that members of a team often have been trained to work with different technologies, which also often come with their own ways of working on a problem. Getting everyone on the same technical stack often takes far more time than is anticipated, and new members can spend much time learning to work in a new stack.

One of the biggest divides our group had experienced was in the choice of using programming languages, as many of us were more comfortable with either R or Python. These programming languages have their own extensive software libraries, like the tidyverse (*Wickham, 2017*) vs. the numpy/pandas/Matplotlib stack (*van der Walt et al., 2011*; *Hunter, 2007*). There are also many different software environments to choose from at various layers of the stack, from development environments like Jupyter notebooks versus RStudio and RMarkdown to the many options for package and dependency management. While most of the people in the room were committed to open source languages and environments, many people are trained to use proprietary software like MATLAB or SPSS, which raises an additional challenge in teams and groups.

Another major issue is where the actual computing and data storage will take place. Members of a team often come in knowing how to run code on their own laptops, but there are many options for groups to work, including a lab's own shared physical server, campus clusters, national grid/supercomputer infrastructures, corporate cloud services, and more.

Workflow and pipeline management

Getting everyone to use an interoperable software and hardware environment is as much of a social challenge as it is a technical one, and we had a great discussion about whether a group leader should (or could) require members to use the same language, environment, or infrastructure. One of the technical solutions to this issue — working in staged data analysis pipelines — comes with its own set of challenges. With staged pipelines, data processing and analysis tasks are separated into modular tasks that an individual can solve in their own way, then output their work to a standardized file for the next stage of the pipeline to take as input.

The ideal end goal is often imagined to be a fully-automated (or 'one click') data processing and analysis pipeline, but this is difficult to achieve and maintain in practice. Several people in our group said they personally spend substantial amounts of time setting up these pipelines and making sure that each person's piece works with everyone else's. Even with groups that had formalized detailed data management plans, a common theme was that someone had to constantly make sure that team members were actually following these standards so that the pipeline keep running.

External handoffs to and from the team

Many of the research projects we discussed involved not only handoffs between members of the team, but also handoffs between the team and external groups.

The "raw" data a team begins with is often the final output of another research team, government agency, or company. In these cases, our group discussed issues that ranged from technical to social, from data formats that are technically difficult to integrate at scale (like Excel spreadsheets) to not having adequate documentation to be able to interpret what the data actually means. Similarly, teams often must deliver data to external partners, who may have very different needs, expectations, and standards than the team has for itself. Finally, some teams have sensitive data privacy issues and requirements, which makes collaboration even more difficult. How can these external relationships be managed in mutually beneficial ways?

Team management challenges

Beyond technical challenges, a number of management issues face research groups aspiring to implement best practices for data-intensive research. Our discussion highlighted the difficulties of composing a well-balanced team, of dealing with fluid membership, and of fostering generative coordination and communication among group members.

Composing a well-balanced team

Data-intensive research groups require a team with varied expertise. A consequence of varied expertise is varied capabilities and end goals, so project leads must devote attention to managing team composition. Whereas one or two members might be capable of carrying out tasks across the various stages of research, others might specialize in a particular area. How then can research groups ensure that no one member of the team departing would collapse the project and that the team holds the necessary expertise to accomplish the shared research goal? Furthermore, some members may participate simply to acquire skills, while others seek to establish or build an academic track record. How might groups achieve alignment between personal and team goals?

Dealing with voluntary and fluid membership

A practical management problem also relates to the quasi-voluntary and fluid nature of research groups. Research groups largely rely extensively on students and postdocs, with an expectation that they join the team temporarily to gain new skills and experience, then leave. Many members also only work part-time for a research group, with other professional obligations (like classes) that make real-time collaboration difficult. In the long-term, turnover becomes a problem when processes, practices, and tacit institutional knowledge are difficult to standardize or document. What strategies might project leads employ to alleviate the difficulties associated with voluntary, fluid membership? What kinds of collaboration plat-

forms, documentation practices, data management strategies, and workflows best support groups with regular turnover?

Fostering open and inclusive coordination and communication

The issues of team composition and voluntary or fluid membership raise a third challenge: fostering open and inclusive communication among group members. Previous research and guidelines for managing teams (*Edmondson, 1999*; *Google reWork, 2017*) emphasize the vital role of psychological safety in ensuring that team members share knowledge and collaborate effectively. Adequate psychological safety ensures that team members are comfortable speaking up about their ideas and welcoming of others' feedback. Yet fostering psychological safety is a difficult task when research groups comprise members with various levels of expertise, career experience, and, increasingly, communities of practice (as in the case of data scientists working with domain experts). How can projects establish avenues for open communication between diverse members?

Not abandoning best practices when deadlines loom

One of the major issues that resonated across our group was the tendency for a team to stop following various best practices when deadlines rapidly approach. In the rush to do everything that is needed to get a publication submitted, it is easy to accrue what software engineers call "technical debt." For example, substantial "collaboration debt" or "reproducibility debt" can be foisted on a team when a member works outside of the established workflow to produce a figure or fails to document their changes to analysis code. These stressful moments can also be difficult for the team's psychological safety, particularly if there is an expectation to work late hours to make the deadline.

Concluding thoughts and plans

Are there universal best practices for all cases and contexts?

At the conclusion of our first meeting, we evaluated topics for future discussions, thinking about identifying potential solutions to the challenges faced by data-intensive research groups. In doing so, we were quickly confronted with the diversity of technologies, research agendas, disciplinary norms, team compositions, and governance structures, and other factors that characterize scientific research groups. Are solutions that work for large teams appropriate for smaller teams? Do cross-institutional or inter-disciplinary teams face different problems than those working in the same institution or discipline? Are solutions that work in astronomy or physics appropriate for ecology or social sciences? Dealing with such diversity and

contextuality, then, might require adjusting our line of inquiry to the following question: At what level should we attempt to generalize best practices?

Our future plans

The differences within and between research groups are meaningful and deserve adequate attention, but commonalities do exist. This semester, our group will aggregate and develop input from a diverse community of practitioners to construct sets of thoughtful, grounded recommendations. For example, we will aim to provide recommendations on issues such as how to build and maintain pipelines and workflows, as well as strategies for achieving diversity and inclusion in teams. In our next post, we will offer some insights on how to manage the common problem of perpetual turnover in team membership. On all topics, we welcome feedback and recommendations.

Combating impostor syndrome

Finally, many people who attended told us afterwards how positive and valuable it was to share these kinds of issues and experiences, particularly for combating the "impostor syndrome" that many of us often feel. In scientific research, we typically only present the final end-product of research. Even sharing one's final code and data in perfectly reproducible pipelines can smooth over and obscure all the messy, complex, and challenging work that inevitably takes place in any research process (*Neff et al., 2017*). And terms like "data science" and "machine learning" are often presented to the public as kinds of magic rather than as kinds of methods, which can set dangerous expectations (*Elish and boyd, 2018*).

In our group, people appreciated hearing others talk openly about the difficulties and challenges that come with doing data-intensive research and how they tried to deal with them. The format of sharing challenges followed by strategies for dealing with those challenges may be a meta-level best practice for this work, versus the more standard approach of listing more abstract rules and principles. Through these kinds of conversations, we hope to continue to shed light on the doing of data science in ways that will be constructive and generative across the many fields, areas, and contexts in which we work.

Acknowledgments

References

Borgman C. Big Data, Little Data, No Data: Scholarship in the Networked World. The MIT Press; 2015.

eLife Editorial Team, eLife Publication Template. eLife; 2017. https://submit.elifesciences.org/html/elife_author_instructions.html.

Edmondson A. Psychological Safety and Learning Behavior in Work Teams. Administrative Science Quarterly. 1999 Jun; 44(2):350. https://www.jstor.org/stable/2666999, doi: 10.2307/2666999.

Elish MC, boyd d. Situating methods in the magic of Big Data and AI. Communication Monographs. 2018; 85(1):57–80. https://doi.org/10.1080/03637751.2017.1375130.

Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, Di Stefano R, Gil Y, Groth P, Hedstrom M, Hogg DW, Kashyap V, Mahabal A, Siemiginowska A, Slavkovic A. Ten Simple Rules for the Care and Feeding of Scientific Data. PLoS Computational Biology. 2014 Apr; 10(4):e1003542. doi: 10.1371/journal.pcbi.1003542.

Google reWork, Guide: Understand team effectiveness. Google; 2017. https://rework.withgoogle.com/guides/understanding-team-effectiveness/steps/introduction/.

Hunter JD. Matplotlib: A 2D graphics environment. Computing In Science & Engineering. 2007; 9(3):90–95. doi: 10.1109/MCSE.2007.55.

Kitzes J, Turek D, Deniz F. The Practice of Reproducible Research : Case Studies and Lessons from the Data-Intensive Sciences. Oakland: University of California Press; 2018. http://practicereproducibleresearch.org.

National Research Council. Enhancing the Effectiveness of Team Science. Cooke N, Hilton M, editors, Washington, DC: The National Academies Press; 2015. doi: 10.17226/19007.

Neff G, Tanweer A, Fiore-Gartland B, Osburn L. Critique and contribute: A practice-based framework for improving critical data studies and data science. Big Data. 2017; 5(2):85–97.

Noble WS. A quick guide to organizing computational biology projects. PLoS Comput Biol. 2009 Jul; 5(7):e1000424.

Patil DJ. Building Data Science Teams. Sebastopol: O'Reilly Media; 2011.

Perez-Riverol Y, Gatto L, Wang R, Sachsenberg T, Uszkoreit J, Leprevost FdaV, Fufezan C, Ternent T, Eglen SJ, Katz DS, Pollard TJ, Konovalov A, Flight RM, Blin K, Vizcaino JA. Ten Simple Rules for Taking Advantage of Git and GitHub. PLoS Comput Biol. 2016 07; 12(7):e1004947.

Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten Simple Rules for Reproducible Computational Research. PLoS Computational Biology. 2013; 9(10). doi: 10.1371/journal.pcbi.1003285.

Traweek S. Beamtimes and Lifetimes: The World of High Energy Physicists. Cambridge: Harvard University Press; 1992.

van der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. CoRR. 2011; abs/1102.1523. http://arxiv.org/abs/1102.1523.

Wickham H. Tidyverse: Easily install and load tidyverse packages. Tidyverse. 2017; http://tidyverse.tidyverse.org.

Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, Haddock SHD, Huff KD, Mitchell IM, Plumbley MD, Waugh B, White EP, Wilson P. Best Practices for Scientific Computing. PLoS Biology. 2014 Jan; 12(1):e1001745. doi: 10.1371/journal.pbio.1001745.

Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK. Good enough practices in scientific computing. PLOS Computational Biology. 2017 Jun; 13(6):e1005510. doi: 10.1371/journal.pcbi.1005510.