Principles of Experimental Research in Computing in the Warnow Lab

Tandy Warnow

September 2024

1 Introduction

In designing a method, or in testing a new method in comparison to other methods, there are many issues to consider. Rigorous work requires care.

This document begins with an overview of the principles I use in my research and expect my students to follow. However, I realize some senior investigators have their own standards for how to do algorithm design and evaluation that may differ from mine. For that reason, a student who has been trained under a different system may consider some of these rules potentially excessive and unnecessarily restrictive. Nevertheless, please follow them, since this is what I think enables a high standard of rigor and academic honesty that will help in the review process and in developing reliable methods.

This document also includes some other advice that is not explicitly addressing rigor and scientific integrity so much as recommendations that increase the probability of success more generally.

Finally, all healthy lab cultures encourage and benefit from vigorous discussions, including differences of opinion, about interpretations of data and approaches for solving problems. However, the principles described in Section 2 are ones that my lab follows and that I consider not negotiable.

2 Principles of Experiment Design

Open Source Software Make your new method available in open-source software that is publicly available. This is certainly important to do *no later than when you submit a paper that uses the software.* Our lab uses GitHub, but if you prefer a different open-source repository, that is fine.

Design Stage and Testing Stage Typically method development operates in stages, where the first stage uses a set of datasets to design the algorithm, select default setting for parameters, etc. We can refer to the datasets used here as "training data", though it may not be any kind of machine learning task.

Use the design phase to set algorithmic parameters. You may want settings for fast versions and for slower but more accurate versions. Once the design phase is done, use separate data with the settings you picked in the design phase to evaluate your method in comparison to other methods.

During the testing phase you may find that your default parameters for your method aren't working well in some setting or some experiment. One option, and probably the right one in nearly every case, is to accept that as the outcome. However, if you really want to change the defaults, then it is probably best to add all the testing data you used to your training dataset, and then get new testing data.

Competing methods The choice of methods to compare to is essential. For sure, you should use the best competing methods to compare your new method to, and use them in such a way as to get the best results. Don't just use "default" modes if you know a way to run the method to get better results – whether that is runtime or accuracy, depending on your objective.

But there are also methods that are popular even if they don't perform well. Most likely you should include them, to provide the comparison. But if you do not include obvious competitors, you need to say why you are excluding them.

In general, if you do not include an obvious competitive method, you need to justify that decision, and it should not be only that you don't have time because the conference deadline is too soon. Valid reasons for excluding a method X can be that some prior study established that method X was not in the top group, and was specifically not as "good" (however you are measuring quality) as some other methods. If you are including the better methods already, then not including method X can be justified. Other valid reasons can be that method X requires larger computational resources than you have available (e.g., an HPC environment) or access to specific databases that you don't have access to, etc. In general, you need to justify not including a method that a reader or reviewer might consider natural to include.

This can get tricky, however, if you are working on a problem where new methods are being actively developed. For example, you might be writing up a paper for submission and have just discovered a paper (perhaps just in a preprint server, like arXiv) that describes a new method, and that method looks promising. If the method is still in preprint form, then it can be valid to mention the method and say that future work should compare to it; however, if it's in a journal already, this gets harder to justify. (Yet another reason to plan ahead and do a careful literature search before you engage in the project.)

Sometimes the first experiment is used to narrow down the set of competing methods to a smaller set. This is fine, but be careful to not exclude a method prematurely. Not only can this be a mistake (because the excluded method might be truly excellent in some other setting), but it can give the impression that you are *trying* to make the competing method look bad.

Finally, if you use a competing method in one experiment, in general you should use it in all your experiments where it is applicable. As with all of this

advice – excluding a method once you deem it relevant needs to be justified.

Datasets When you pick datasets for either the design stage or the testing stage, make sure the two sets are disjoint, and if possible independently generated.

Datasets for design or testing may be real-world or simulated. Depending on the problem, it may be difficult to pick good datasets of one or the other type, and you may need to generate new ones. If you generate new datasets for benchmarking, be careful and thoughtful about the process, and make it very clear (in your manuscript) how you did this, so the results can be interpreted.

Definitely consider those benchmarks that are used by others, but do apply your own judgment. If you exclude some popular benchmarks, you need to justify the exclusion.

One thing to be careful about is that you don't use datasets that have been rejected for a principled reason (e.g., unreliable ground truth). Knowing which published datasets have been considered reliable and which ones aren't requires some care.

As much as possible, look for a wide range of datasets, preferably ones that are realistic so that results on the datasets will be relevant to practice.

Finally, include datasets with a range of difficulty levels. The ones that are neither too easy nor too hard are likely the most useful, since these enable you to distinguish between methods. However, including ones that are easy *in addition* is helpful, since doing poorly on easy datasets is a clear indication of a problem.

Evaluation criteria Another important issue is how you evaluate methods. These criteria can include computational performance and accuracy measures. In some fields there are clear evaluation criteria, and if you don't have concerns about them, using them is fine. But if you disagree with the standard criteria and plan to use others, then you should state this and provide reasons. It may well be that your criteria are more sensitive and helpful than standard criteria, after all!

Make your work reproducible There are two aspects in which reproducibility comes up: one is in the data you generate and then use to evaluate methods, and the other is in the results of analyses of these data using methods. Both of these generate data that you want to make available, either through detailed instructions on how to produce the data, or by saving the data in a public repository.

If you want to enable this through providing information on how to generate the data, then take care to provide full information about what you did. For example, provide the commands and version numbers, at a minimum, for any software you used in generating the data. Since many methods employ randomness, exact reproduction of results may not be possible. On the other hand, if you can provide a random seed so that it becomes truly reproducible, that's good to do.

Don't cherry pick or give the impression of cherry picking Try to do what you can to not cherry pick what you show. All too often papers have only the positive results where the new method does well, and you have to look in the supplement for the other results. Either way, it can look like cherry picking or trying to mislead the reader.

Ethical challenges As a researcher, you will increasingly confront situations that are tricky and potentially challenging from an ethical perspective. These challenges arise in writing papers and in reviewing papers, but also in other conditions. Some of these have to do with the principles discussed above. Please take a look at [1] to see some examples of situations that can arise.

3 The importance of sticking to the schedule

For me, having a schedule for getting a paper done and then sticking to it is essential. Here are some specific guidelines about these issues for my lab.

In general, while I have overall preferences for how to get papers done (i.e., when to finish the design stage, when to finish the testing stage, etc.), for papers that are planned for a conference, I will usually specify the deadline for each stage to be done. Those deadlines will include one that says when no more experiments can be performed, no more data provided, and we only work on writing. Deviating from the schedule, and in particular doing experiments or generating new data after that "no more data" deadline has passed, is not something I want to consider. Specifically, with very few (and perhaps no exceptions), the only time I'd consider deviating from the schedule is if it is *essential* to do so. And for me, essential means something that a reviewer – if they noticed it – would recommend rejection of the paper.

Here are some examples of what I would consider a critical reason to violate the deadline:

- The experiments performed were not the ones that were planned (i.e., the wrong methods were used, the wrong datasets were analyzed, the wrong criteria were reported, the results doesn't match what we said we did, something like that).
- The experiments we performed were valid but we cannot include them in the paper, for example, if we lost the data so we cannot report it anymore.
- A planned experiment turns out to be deeply flawed in design, as in:
  - We used invalid benchmark datasets
  - We used data that are not public (and we don't have permission to publish results on these data)
  - We used competing methods poorly
  - We used criteria that are biased in favor of our own methods

– The training data and testing data overlap
– Another method was published and is absolutely essential to compare to, but we didn't include it and we cannot justify not including it

In other words, changing the schedule is *not* done for reasons other than these. And in particular, we don't change the schedule because we can improve the method by changing its design or changing the default parameter values.

If there are changes that might lead to improved methods, and the improvement is really essential for even this paper – then we could consider violating the schedule by *not* submitting it to the intended venue (conference or journal).

However, deciding not to submit to the conference or journal has an impact on other people (the other students in particular) and so it is something I would not do lightly. This is why, among other things, I don't approve changes to the schedule. (And yes, I know there are others who don't feel so strongly about sticking to a schedule, but I do.)

Finally, a change to the schedule without changing the venue likely means extra work at the last minute by me and by all the others on the project. This is added stress and often leads to things being done poorly (and mistakes being made that are not found before submission). In itself, this is another reason I don't want to change the schedule.

4 Other advice

I also have advice that is not about how to do rigorous experimental work, but aspects of being a researcher that increase the chances of having a successful career.

Being a good collaborator If you are collaborating with others, early in the collaboration (as early as possible) make your software and datasets easily accessible to them. The software can be initially in a private GitHub repository to which your collaborators have access, but should be made public no later than when you submit the paper. Document what you are doing. Good collaborations often involve people checking each other's results – this is not possible unless the software and datasets are available.

Please remember that collaborative research can take time to complete, sometimes a lot longer than you expect. Try to be patient and trust your lab PI about the timing. That said, if you are the person who is slowing down a project, make sure to talk with the project lead (i.e., your PI) about why it's slowing down, so that he or she can figure out how to handle the issue. In the same vein, if you have committed to providing something (software, datasets, figures, etc.) to a collaborator and are late, or expect to be delayed in completing the task, make sure to let your collaborators know. We all depend on each other, and communication is key to making collaborations work.

A final point about being a good collaborator is that responding to emails quickly is good practice. Note: while "quickly" generally means (for me) within

24 hours, if you receive email on Friday and don't respond until Monday, that's perfectly reasonable. The main point is that delays in responding don't make a good impression.

Managing interactions with others My advice about responding to email from your collaborators doesn't apply as strongly to how you manage email with other people. For example, we all get way too much unsolicited email, often from complete strangers. You are not under an obligation to respond to all the email you receive! But if the email comes from someone whose opinion you might care about (or should care about), do make an attempt to respond reasonably quickly. For example, if someone writes to you about one of your papers with a question, do take the time to respond, even if just to say "Thank you for your question about this paper! I will get back to you in a few days after I have had a chance to look into this." And then, of course, do follow up.

Planning your experiments As I write in "How to write your first paper" [2], planning out your work well in advance can lead to improved results. Careful planning also leads to better development of the training phase and selection of datasets and alternative methods, and also helps you avoid the need to redo experiments. While that document was written up for new graduate students, it is relevant to senior students as well.

Have a lab notebook It is all too easy to make mistakes, and avoiding mistakes or catching them early, is essential. Therefore, I strongly recommend you plan ahead before doing analyses. Know what datasets you'll use, how you will run each method (what version number, what commands). When you do analyses, write down what you will do before you do it. Don't just copy and paste from previous analyses.

In some experimental sciences, this practice is done with a formal lab notebook, often a physical one (i.e., a book that you write into, with dates) rather than a digital one. To this date, I don't think any of my students have done a physical lab notebook, but at least one is using and maintaining a digital lab notebook. I wish they all did – and hope this encourages them to take up this practice.

Save your data It is important to do research that is reproducible. Therefore, full details of methods used (version numbers and commands) are important. But in addition, in many cases it is important to save your data, including analyses of datasets. Putting these data into a public repository may be required. Therefore, don't throw out your data until you are sure you don't need them anymore. And also, try to not overwrite your results, because then you end up probably needing to rerun experiments.

Having said this, it's also important to not use up your lab resources. Therefore, both during your analyses and afterwards, evaluate whether the data will be needed in the future by you or others. If the data are important to make

available to others, put them in a public repository. If you think not and don't want to use the data yourself, delete the data (free the space up for others). If you are unsure about whether you or others will need the data later, you can download the data and save it in your own digital storage. In other words, be considerate about others.

Read extensively It is very helpful to read the literature regularly (every week is best), and keep track of what you have read and what you thought about it. I have a latex document where I copy the abstract and full bibliographical information about every paper I have looked at long enough to gather an impression; in that document, I add notes that help me remember what the paper is about (i.e., theory vs experiment, what question they address, what they claim to achieve), as well as whether I think they succeed, whether I think I need to read it again, and what other papers are related to the paper. Maintaining a literature review document like this is tremendously helpful when it comes to writing your papers!

Citing the relevant literature How and what you cite in your research papers is surprisingly important. All too often, people cite papers they haven't read, just because others cited the paper. Here I strongly recommend you take care to cite literature appropriately. To begin with, make sure you have read the papers you cite, at least enough to verify what the paper is about and what it shows. Make sure you cite the most important relevant papers for results, and preferably the publication that established the result rather than one that cites those papers. All this requires a careful literature search, and should be done before you are finishing your paper!

But also, take care to not cite retracted papers (or justify why you need to), and avoid doing too much self-citation.

Remember that when you submit your paper for review, what you cite can either help or hurt you. The reviewers may be knowledgeable about the questions you address, potentially even people who wrote important papers on the subject.

Keep track of deadlines and be on time It's very easy to lose track of what you need to do, and this gets increasingly difficult as your responsibilities increase. My recommendation is to have a "To Do" list that you maintain. Personally, I have two: a digital one and also one that is on a whiteboard with the urgent ones. That way, I am reminded about things and don't forget them.

Related to this, if you are a reviewer on a paper or for a conference, submit your reviews on time, not a day or two after the deadline (which is all too common among conference committees). Believe it or not, many conference program committee chairs keep track of who submits reviews late, and they stop inviting them to join the PC (and word can get around).

Acknowledgments

I am glad to acknowledge feedback on earlier drafts of this document from both current and former students, including Luay Nakhleh (Rice University) and Siavash Mirarab (UCSD). Finally, extensive discussions with George Chacko (UIUC) led to additions and clarifications, and are greatly appreciated.

References

[1] T. Warnow. Ethics in science, 2015. Online, `https://tandy.cs.illinois.edu/ethics.pdf`.

[2] T. Warnow. How to write your first paper, 2015. Online, `https://tandy.cs.illinois.edu/ten-rules-writing-papers.pdf`.