Research Writing Guide

Derek Aguiar

July 5, 2023

Prepared for CSE5825-BML[1]

Contents

1 Before writing: defining a project 3

1.1 Selecting a topic and data . . . 3

1.1.1 General . . . 3

1.1.2 AI and machine learning . . . 3

1.1.3 Examples of biological data . . . 3

1.2 Selecting a journal . . . 4

1.3 You have a project, now what? . . . 5

1.4 Resources . . . 5

2 General writing advice 5

2.1 Unhelpful words and phrases . . . 5

2.2 Sentence structure . . . 6

2.3 Semantics . . . 6

2.4 Detailed outline . . . 6

2.5 Finishing a document . . . 7

3 Abstract 8

3.1 Format . . . 8

4 Introduction 8

4.1 Structure . . . 9

4.2 Writing . . . 9

4.3 Resources . . . 9

5 Related work 9

5.1 Goal . . . 9

5.2 Structure . . . 10

6 Methods 10

6.1 Writing style . . . 10

6.2 Experimental description . . . 10

6.3 Methods development . . . 11

6.4 Metrics for validation . . . 11

6.5 Methods resources . . . 11

[1] significant portions borrowed from Barbara Engelhardt and some parts from Sorin Istrail

7 Results **12**
7.1 Structuring your Results section 12
7.1.1 Results on Simulated Data; Methods Comparison 12
7.1.2 Results on Experimental Data 12
7.2 Presenting results in the Results section 13
7.2.1 Making an observation about your data or results. 13
7.2.2 Referring to a figure or a table 13
7.3 Reproducibility 13

8 Discussion **14**

9 Citations **14**
9.1 Captions 14
9.2 Citing in text 14

10 Reviewing **15**
10.1 How to Read a Manuscript 15
10.2 How to Review a Manuscript 15
10.2.1 Technically Sound 15
10.2.2 Originality 15
10.2.3 Manuscript presentation and layout 16
10.3 Some references 16

1 Before writing: defining a project

In these tips, we focus on research projects that fall into the *Methods development* category. This means that the main contribution of the paper is often methodological; the results should include comparisons to relevant competitive methods and, ideally, novel domain-specific findings that were enabled by the new method.

Methods development projects can be partitioned into four categories.

- **Old data, new methods.** Most typical case. The data has been used in previous studies, maybe in many different ways. But you are developing methods that allow new signals to be identified in these data to find what others have not found yet.
- **New data, new methods.** With new data, you are guaranteed to find signals that others have not found yet. But you still need to justify the development of novel methods, typically with superior comparisons on evaluation metrics.
- **New data, old methods.** The novelty should come from applying an old method in an interesting way that others have not thought about. The presentation of the new data is also important to the community.
- **Old data, old methods.** No one has thought to apply this model to these data yet, even though both have been analyzed. What will this show in these data that other methods have not already?

1.1 Selecting a topic and data

The order in which you select a topic or data set depends on the nature of the project. If the topic is the development of novel methodology, the data set should be chosen to exemplify the methods. Typically, however, the topic is defined relative to some data and a *specific* hypothesis. These data source aggregators are useful:

1.1.1 General

1. KDNuggets
2. Google Data
3. Hadoop public data

1.1.2 AI and machine learning

1. Kaggle
2. UCI Machine Learning Repository
3. Scikit learn datasets
4. Imagenet
5. MNIST handwritten digits
6. Great article on data sources for ML.

1.1.3 Examples of biological data

1. dbGaP: repository for genotype and other genomic data
2. Protein Data Bank (PDB): database of tertiary structures of tens of thousands of proteins
3. Gene Expression Omnibus (GEO): gene expression and RNA-sequencing repository

4. The Cancer Genome Atlas (TCGA): gene expression and RNA-seq repository for cancer studies
5. 1000 Genomes/HapMap: 40M SNPs from > 1000 individuals worldwide, well curated and phased
6. UCSC Genome Browser: visualization for all of these data sets and more

Hypotheses should be specific, for example:

- **Comparing results from different available methods.** What software and models perform best for identifying underlying ancestral structure in genotype data?
- **Testing a specific hypothesis.** Is there evidence of racial bias in machine learning algorithms trained to assist convictions in misdemeanor?
- **Identifying underlying structure.** What is the ancestry of individuals? Can I identify signals of natural selection in genotypes? Can I reconstruct the communities in a social network? What are the signatures of denial of service attacks in network traffic data?
- **Searching for specific relationships.** Which Twitter accounts are fake and principally used to sow discord abroad? Which genes are differentially expressed in smokers?
- **Develop a method to perform a specific task.** Can I write a method to identify which U.S. politically districts are illegally gerrymandered? Can I develop a model that will allow me to predict gene expression from methylation status and genotype data?

You must critically evaluate if the data can be used to support the chosen research topic. If there is no signal, a different topic, data set, or method will have to be chosen or developed.

Finally, motivate the work. If you answer the hypotheses, what will be the contribution to the project domain of the selected topics? What will be the contribution to the field (e.g. computer science, computational biology, sociology)?

1.2 Selecting a journal

Your choice of project should be decided principally on what interests you the most. When deciding on where to publish your work, however, a number of considerations come into play. There are many factors upon which a journal may be selected for a manuscript. These include:

- Relevancy: does this journal publish papers on related topics? Will their readers find your project interesting?
- Your career path: do you want a job in a stats department? Publish papers in stats journals.
- Open access: will your journal be put behind a paywall, or made public? Many journals have options for open access with a fee.
- Impact factor: do people read and cite articles in this journal?
- Review process: is this journal known for a careful but quick review process? Does it allow preprints?
- Overshooting: is this the type of work that belongs in this journal? If not – either in impact or in style – it is often better to avoid a lengthy review process and many resubmissions by submitting to a lower impact journal.

Once you have selected a journal, most offer a LaTeX templates for manuscript preparation and submission. Often the manuscript format for submission is suboptimal for preparation (e.g., figures at the end or submitted separately). Ignore the format and develop the manuscript in a way that is easy to read and prepare. At the time of submission, reread the submission format information. Note: some scientific journals make the claim that they do not support LaTeX submissions, but, in practice this is not likely. In the worst case, everyone either accepts PDFs or Pandoc can be used to convert between formats.

Journals that publish machine learning and AI research include Journal of Machine Learning Research (JMLR), and Annals of Applied Statistics (AOAS). Additionally, journals that publish genomics work which include applied ML, include Science, Nature, Cell, Nature Methods, Nature Genetics, Journal of Computational Biology, PLoS Computational Biology, PLoS Genetics, Genome Research, Genome Biology, Bioinformatics, PeerJ, eLife, PLoS One, and Biostatistics. Conferences meet once a year to present accepted work. They require shorter manuscripts with quicker turn-around than journals: RECOMB, ISMB, Pacific Symposium on Biocomputing, NIPS, ICML, AISTATS, UAI.

1.3 You have a project, now what?

If you are working in a group, decide how to organize the work. At minimum, you should:

- Set up a git repository for your project including folders for data, code, docs, and libraries.
- Do commit: .tex and .sty files for your manuscript, a supplemental information document (see PLOS-one or JMLR templates), source code, documentation, and small test case data. Do not commit: intermediate build files for LaTeX or source code, libraries (ideally, although I think it is fine to include libraries pre-release), large or access-restricted data.
- Build the general structure of the code to better allocate resources. Jupyter notebooks are useful for organizing data preprocessing or visualization.
- In the Methods section, detail your data set: where you acquired it, when, who developed it, on what platforms they developed it, and the numbers and features that it contains. We will describe the Methods section in detail later, but it is important to write as you work.
- Write a provisional title.

1.4 Resources

- Google Scholar: search the scientific literature
- PubMed: search the scientific literature (geared towards life and medical science)
- SciReader: helps find relevant research

2 General writing advice

The effective use of proper grammar is essential to scientific writing. You will gain enormous benefit from taking some time early in your career to learn about writing style[2] and effective writing practice[1]. This section contains rules for scientific writing that, when broken, can sometimes be a detriment of the paper. Some of these rules are more subjective than others, so they should be read with some skepticism.

2.1 Unhelpful words and phrases

The following word usages can make it difficult to comprehend sentences.

- etc.
- ellipses (...) except in mathematical contexts
- any use of backslashes, e.g., and/or
- & (use "and")
- vs. (use "versus")
- utilize (use "use"). In general, the easiest to comprehend word that effectively conveys your meaning is preferred.

- contractions, e.g., don't, wont, didn't. Contractions are informal and should be avoided in scientific writing.
- eg. (use "e.g.")
- Use a comma before "which" but not "that" when using nonrestrictive phrases (sentence clauses that add nonessential information to a noun phrase already mentioned in the sentence).
- Avoid adverbs, e.g., very, really, especially. There is usually a better word to describe your meaning and adverbs tend to be nonspecific to the point of not conveying information.
- Avoid using a negative phrase and a verb together, e.g., "not enriched" should be "depleted".
- Always use an Oxford comma: "A, B, and C" not "A, B and C".
- Be cautious using words suggested from a thesaurus if you are not familiar with exactly how the word is used and in what contexts.
- "This" as the subject of a sentence. e.g. "This will," "This has been," "This is,". the writer is placing a burden on the reader to figure out what portion of the material already introduced is being referred to. If you must use "this" at the start of a sentence, please give us some clues about what "this" might be, for example: "This tendency to use vague demonstrative pronouns as the subject of sentences drives me crazy."
- Unique. Statements like "unique expertise," a "unique endeavor," "uniquely prepared students," and "unique opportunities." do not provide many details that the reader would understand. It's better to be specific about what would make these unique.

2.2 Sentence structure

- Do not write long sentences. Break up run-on sentences with periods or semicolons.
- Avoid using passive voice.
- When possible, use simple words and sentence structures.

2.3 Semantics

The goal of this section is to promote the writing of meaningful, clear sentences and paragraphs. You should not use a word that does not capture the object, idea, or action you are trying to convey to the reader. For example, "We collected nutritional data for 943 participants." What does it mean to collect "nutritional data"? Do you mean that you collected UNHCR nutritional survey responses for these participants? Think about the purpose of each sentence and the most effective semantics to convey that purpose.

Logical flow is another concept that is difficult to master. Paragraphs read similar to mathematical proofs. Avoid writing a statement until all the required information leading up to that statement has been written.

Finally, the order of idea presentation is also a difficult and subjective area. Section and subsection headers direct the reader similar to comments in code and should have a logical flow throughout the whole document. Headers should be a descriptive as possible without being unreadably long. Some authors advise to write complete sentences as section names with a punchline, for example: "Linear regression results suggest ingesting brocolli increases lifespan".

2.4 Detailed outline

It can be a useful practice to write a paragraph-resolution outline of a manuscript including ideas for figures and caption headers prior to writing the document. Each journal accepts papers which tend to conform to a specific format. Find several high-quality related papers in the journal you identified in Section 1.2. Make sure that the *type* of paper matches, for example, if you are writing a methods paper, then search for methods papers. *Do not use the specific words or phrases that the related papers use. This is plagiarism.* However, you can use ideas from the related work like the general flow and organization in terms of presentation of ideas and details. You will have to make adaptations to the flow and organization for your specific work.

2.5 Finishing a document

After receiving feedback from your coauthors, it is time to submit. Here is a general checklist for finalizing a manuscript (some advice assumes you are using TeX)[2].

- □ Check author names and affiliations. Verify that none of your authors have recently moved or have additional affiliations they would prefer to be listed. If the submission is for a double blind review, double check that your manuscript follows the appropriate conventions.

- □ Thoroughly spell check the document. Most document preparation systems have built in spell-check but if you have scientific terms or jargon these will likely not be included in their dictionaries.

- □ Bibliography entries should be about two lines but there are exceptions.

- □ Bibliography entries should be consistent, e.g., each entry includes the first initial and last name only for each author. Omit "Proceedings of Conference" and simply name the conference. You can usually omit page numbers if they would add another line to the entry. You can force capitalization in LaTeX by surrounding wording with double curly braces in the .bib entry, e.g., {{Hierarchical Dirichlet Process}}.

- □ Bibliography should follow the format designated by the journal.

- □ Review every bibliography entry in the generated document for complete information (no missing authors, journal names, volume numbers, and so forth).

- □ Ensure that the colors, ordering, and inclusion of methods, experiments, and results are consistent across Methods, Figures, Tables, and Results.

- □ Read the Methods section again and make sure that a third party could repeat each step without having to go to your code. Methods sections should read more like a recipe than the other sections of the manuscript. There should not be motivation or interpretations in the Methods section.

- □ If space is a concern, look for paragraphs that end with a few words on a new line. You can often refactor and condense your wording to save space.

- □ Captions should be clear and consistent; they should begin (or end) with a sentence saying what you want the reader to take away from the figure or table.

- □ Avoid using citations as a noun, but, if you must, some LaTeX citation managing packages offer commands to generated citations in different formats. See \citet{} and \citep{}.

- □ Instead of Equation, Figure, and Table use Eq., Fig., and Tab.

- □ Write a draft of a cover letter if your selected journal requires it.

- □ Make sure the manuscript format follows the instructions exactly (e.g. formatting of figures, section headers, tables, titles).

- □ Every figure, table, and labeled equation should be referenced in the document at least once. They should be referenced in the text in the same order as the manuscript. This should be true of the Supplemental materials as well. In LaTeX an asterisk will usually suppress numbering in an environment, e.g., \begin{equation*} $E = MC^2$ \end{equation*}.

- □ Check the supplemental data, code, and methods. Is the code ready to be distributed to the reviewers? Is the appropriate license applied to the code? Does it include sample code and test cases for the reviewers to run?

[2] adapted from Barbara Engelhardt and David Blei

□ After building document search for "?" which is often used to indicate problems LaTeX encountered when building your document (e.g. missing citations). Correct LaTeX errors and warnings.

□ In some cases, you can suggest reviews to be included. If you do, make sure to ask your co-authors for their opinions. When appropriate, you can ask for reviewers to be excluded due to conflicts of interest (e.g. current advisors or collaborators).

□ Read the manuscript again out loud.

□ One last spelling and grammar check.

□ Get a suitably journal-anonymized version of the manuscript ready for arXiv or biorXiv after discussing with your coauthors.

3 Abstract

Journal abstract should be concise, describe the problem, the methods developed, and discuss the main one or two conclusions. All journals and conferences will have slightly different formats and lengths for abstracts. Most journals have character limits or specific formatting requirements, e.g., Motivation, Results, Implementation. Most journal prohibit citations (common) or acronyms (rare) but all will include an *instructions for authors* section on their website (e.g. https://www.nature.com/nature/for-authors/formatting-guide. View other published articles in the target journal, preferably concerning related problems, to get an idea of how to layout the abstract.

3.1 Format

The general format for an abstract follows (one sentence each):

- Motivation 1: What is the problem that the paper addresses?
- Motivation 2: Why is it an important problem?
- Motivation 3: Why are the current approaches (if they exist) insufficient to solve the problem?
- Methods: **In this work, we develop an approach to address these deficiencies...**
- Results 1: **We show that our approach** *(is it named?)* **improves on the state-of-the-art methods with respect to a scientific metric of the quality of results** *(which?)* **in simulated** *(how?)* **and real** *(which?)* **data.** This statement can be broken into a separate statement about simulated and real data results.
- Results 2: What are the domain-specific implications of the results? Preferably, these implications should have been enabled specifically by the new approach. What specific structure or mechanism is the method able to find? Are there any large-scale implications of the work? Do you introduce new problems or research areas?

More methods oriented manuscripts can devote more text to Methods sections of the abstract than results. Motivation sentences 1-3 can often be summarized in two sentences. Avoid self aggrandizing your work by using words like 'revolutionary'. Often the best description of you work is the simple, terse description. The abstract is not a place to compare specific details of novelty to other methods.

4 Introduction

The purpose of the introduction is to motivate the work, describe why solving a particular problem is important in the larger context, describe difficulties in solving this problem, and describe related approaches that have come before this work. The introduction section sometimes concludes with a formulaic paragraph giving a rough outline of the paper (this is mandatory in some types of journals, e.g., JMLR, JASA, AOAS). There should be no results or conclusions in the introduction. The introduction sets the stage to make the relevance of the results clear in a greater context but does describe research outcomes.

4.1 Structure

Before writing the introduction, enumerate the general ideas that are necessary for the work to be relevant. Each paragraph should cover on of those ideas, arranged in some logical order. For example, consider some sociological phenomenon P and a method M to characterize it.

1. Define P. What is it? Where does it happen?
2. Why is P important? What are its sociological and political implications?
3. How do we characterize P? What are the characterization methods limitations?
4. What types of methods exist to predict P? What are their limitations?
5. What is the basic experimental design of this paper? Why is this different than previous work?
6. How is the paper laid out (if applicable)?

In the relevant paragraphs, define all concepts and acronyms to domain-specific ideas used in the paper.

4.2 Writing

The introduction should contain short, easy to read sentences. Nearly everything that is mentioned about definition, context, and existing approaches should be supported by citations (when in doubt, cite).

4.3 Resources

- Google Scholar: easy to search for articles and to pull references (find paper, click the quote symbol, click "Bibtex").
- Citeulike: CiteULike: Organize and gather bibliographies. Export collection of papers to .bib file. Free and on the web. The only issue with CiteULike is that it is crowd sourced so there are often errors in the bibliographies.
- Mendeley: Similar to CiteULike with more support. Interprets references from PDFs. Exports collections to .bib files. Integrates into Word. Some features cost money.

5 Related work

A related work section is essential to most research articles (and every methods paper). A fundamental contribution in methods papers is methodological novelty and editors will commonly ask how the method in questions differs from an already published method. A well written related work section will enable reviewers to confidently answer this question.

5.1 Goal

The overall goal is to describe the related research areas and to place your method's contributions to the field in this context. By clearly describing previous work, you can better describe the current limitations and the need for new methodology. It also gives you an opportunity to demonstrate knowledge of the area and helps others relate your current work to other scientific areas. The section can include methods if they formulated the problem, addressed a central or related problem, used a similar methodology as your work to a similar problem, or if your work was inspired by their work.

5.2 Structure

There should be a common theme to the section. For instance, you can describe related work chronologically starting from early work and their assumptions. Subsequent paragraphs can describe how newer methods improved on previous work, relaxed assumptions, or tried to solve similar problems. The final paragraph can then describe how the preceding research leads up to your work. As with the abstract, viewing other published articles in the target journal on related problems will give a good sense of what is expected.

6 Methods

The purpose of a Methods section is to explain the data acquisition, quality control, processing, models, algorithms, and analysis clearly enough so that a sufficiently sophisticated scientist could *replicate your study perfectly and find identical results*. Methods sections should be more formulaic than other sections and read similar to a recipe book. Because this is the unique purpose of a Methods section, these sections are dry and dull to someone who is not interested in the methods. However, these sections are the most important to quantitative researchers and methods developers.

6.1 Writing style

Writing style for a Methods section is declarative and in past tense. Say what you did, with no embellishment or motivation. For example:

1. "We used the hg18 reference genome sequence, removing non-autosomal chromosomes. We then mapped all reads to this genome sequence using MAQ v0.6.8 (Li et al., 2008) using the default parameters (the default settings allow two mismatches in the first 24 bases of a read)." *Note the version number of the software and the citation.*

2. "For each alignment and each of all six possible reading frames (three on the (+) strand and three on the (-) strand), we calculated a likelihood ratio for a model under which dN/dS is estimated versus a model in which dN/dS is 1 using PAML (Yang, 2007). To format alignments for PAML, gaps in the human sequence were removed, gaps in non-human sequences were converted to Ns, and each of the six possible frames was trimmed to be a multiple of three bases." *Note the clear explanation of method input and output.*

6.2 Experimental description

Methods for the experiment should include the following:

- State clearly how your data were acquired, this includes the website used to download the data, the specific ID number, the exact date, and the experimental or computational protocols used to generate the data.

- Describe data set in specific, quantitative terms, e.g. "The data set contained 2156 samples and 7,943,454 single nucleotide variants on autosomal chromosomes."

- State clearly how your data were quality controlled and processed, including citing software and other resources.

- State clearly if you had to reformat or modify any of the underlying data for any reason.

Methods for the analysis should include the following:

- Statistics used for validation and analysis.

- Software used for validation and analysis (R packages, Python packages).

- A description for how significance thresholds for p-values and FDR were identified.

6.3 Methods development

A description of developed methods should include the following:

- A clear definition of the random and observed variables, including dimension and type: "For n samples across p genes, we have a matrix $Y \in \mathbb{R}^{n \times p}$."
- For probabilistic modelling, a description of the generative model, joint probability, and, if necessary, plate diagram. "For $x_i \in \Re^p$ for sample $i \in \{1, \ldots, n\}$ and p genes, and latent variables $z_i \in \{0, 1\}$, we have
$$\begin{aligned} z_i \mid \boldsymbol{\pi} \sim & \quad \text{Mult}(\boldsymbol{\pi}) \\ \mathbf{x}_i \mid z_i = k \sim & \quad \mathcal{N}_D(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \end{aligned}$$
where k is the number of latent components, and $\boldsymbol{\Sigma}_k$ is a diagonal matrix with σ^2 on the diagonal.
- For algorithmic contributions, a clear description of the model of the input data and algorithm such that a well-versed software engineer in the area could reproduce your code.
- For parameter estimation: the starting point of the parameters, the updates to each of the parameters, the convergence criteria.
- For simulations: as in the generative model, describe how each random variable was generated.

A comparison against related methods should include the following:

- for related methods: version, date of download, parameter settings.
- a description for how the data were processed (e.g., "We thresholded all parameters below 0.05 to 0 to evaluate results against the sparse simulation.")
- always run related methods as smartly as possible without spending too much time fiddling with the parameters (cross validation over a hyperparameter grid is always well-appreciated).
- be fair with comparisons and, when possible, set parameters to give competitive methods the benefit of the doubt (e.g., "We set the number of clusters to match the known number of clusters in the simulation."). If, for instance, you vary the number of clusters *a priori*, then vary the number of clusters for all methods.
- describe all (hyper)parameter settings of all methods, including your own.

Whenever possible, methods should be open sourced and available to the research community under proper software licensing (e.g. through a GitHub repository).

6.4 Metrics for validation

Although likely not to be the focus of your paper, you should be very clear about what metrics you used to quantify or compare results. Writing out equations and citing specific packages or research removes uncertainty. The application of these should also be stated clearly (e.g., gene-wise, across samples, within each population, within each cluster). An example of a well-specified metric is "Pearson's Correlation between gene expression levels x and y was computed as $cor(x, y) = cov(x, y)/\sqrt{var(x)var(y)}$."

6.5 Methods resources

Some helpful Methods writing resources include

1. Purdue Online Writing Lab
2. A Dartmouth Professor's ideas
3. How to write methods

7 Results

The purpose of a Results section is to provide evidence for your specific conclusions through statistics quantified on the data. The Results section may come before the Methods section in, for example, some application or methods papers. The Results section should be self-contained as much as possible: do not assume the reader has read your Methods section, and start by motivating and describing your approach. Present the data in a clear, specific, and quantitative manner. Describe the results of applying methods to those data in the past tense (the experiments were done in the past). Finally, describe what the results individually and in aggregate suggest about your conclusions. A basic template for results paragraphs can take the following form: “In order to address the question of WWW, we applied our statistical model to the XXX data. We found that YYY. These results suggest that ZZZ.”

7.1 Structuring your Results section

7.1.1 Results on Simulated Data; Methods Comparison

Results on simulated data should contain the following information:

- how were the data simulated? If using an existing simulator, describe briefly and cite. If using a simulator created for this manuscript, describe briefly in the Results section and in great detail in the methods or supplemental materials.
- what other methods were run? How were they run? Was there any special post-processing or preprocessing of data or results necessary to make these results comparable in your analysis pipeline?
- what statistics are you using to compare the results from these methods on your data? why are those appropriate? what do each of them tell us about the methods (e.g., FDR vs power vs mean squared error, etc.)? why might we need more than one?
- how do each of the results from the methods compare? what assumptions implicit in the different methods are borne out in the results? What structure or predictions are hard or easy for each method? These interpretations can sometimes be more appropriate for the discussion section.

Be aware that the comparison is on the results from different methods, not the methods themselves. It is essential to be fair and honest. It is not useful to say that your method is better than all others in all tasks, but it is essential to show that your approach has some favorable qualities with respect to the other approaches. The authors of the related methods could likely be reviewers for this paper.

7.1.2 Results on Experimental Data

As with the simulated data (but often with fewer comparative methods), your Results for an experimental data application should include the following information:

- What do the experimental data look like (features, samples, missing data, outliers)? The experimental data was likely produced alongside a research article, cite this.
- What questions motivated the application of the method to the experimental data specifically?
- What were the results of applying this method to the data? This is often overlooked and instead the answers extracted directly. It is often useful to characterize what the results looked like visually, e.g., the number of selected features, plots of linear regression, imputed time-series data.
- What specific aspects of the results can be used to answer our domain-specific problems of interest?
- If an alternative method was run, why does the method in the paper produce results that are superior in terms of addressing specific questions about the data?

- How can we quantify the results without ground truth? This can often be a difficult task. What experiments that can serve to validate our findings in the absence of experiments? For example, with ground truth, clustering can be evaluated by comparing the predicted class labels with the true class labels. Without ground truth, clustering methods can be evaluated by comparing the in-cluster homogeneity with the between-cluster heterogeneity.

7.2 Presenting results in the Results section

7.2.1 Making an observation about your data or results.

A statement of the general form:

> We found that there were substantially more quarks in a bloogle than in non-bloogles.

must be accompanied by a p-value or some measure of how likely it is that this observation could be made assuming a particular model of randomness. For example:

> We found that there were substantially more quarks in a bloogle than in non-bloogles ($p \leq 5.6 \times 10^{-4}$; Fisher's exact test).

Some notes about this significance presentation:

- always accompany the p-value statement with a short reference to the test. If it is your own test (e.g., from a logistic regression equation), cite that equation number, which should be written clearly in your Methods section.
- always write the quantified significance the same way throughout your results. Here, I specified p-value as p, and used the less than or equal sign.
- Format p-values using base 10 notation.

In general, words like *more, most, fewer, some* should be avoided; instead opt for a quantitative treatment of the observation or comparison.

7.2.2 Referring to a figure or a table

Refer to results from a table or a figure, not to figures or tables themselves.

-
- Bad: "As you can see in Figure 4, the first and second PCs clustered the ancestral groups."
- Instead: "The first and second PCs clustered the ancestral groups (Figure 4)."
- Bad: "Table 2 shows the accuracy of our method against all of the alternative methods."
- Instead: "We compared the accuracy of our method against four alternative methods on the simulated data (Table 2)."

7.3 Reproducibility

Just as we encourage software to be open source, the gold-standard for results transparency is to make your data, quality control processing pipelines, and results generation available as well. This can take the form of iPython or KnitR scripts or a set of scripts and programs run in sequence from a shell script. Maintaining a clean organization of your preprocessing steps from the start of the project facilitates dissemination and makes it easier to re-run experiments in the event that a mistake was made. Doing this will ensure that other researchers can replicate your data precisely and improves reader confidence in your work.

8 Discussion

Discussion sections can be difficult to write. The aim is to be clear about how your method works and the implications of the assumptions in your methods and analysis choices on the results. However, these questions can often be self-deprecating and be used against you by reviewers to extend the scope of your work beyond where this paper should go.

A general format for the discussion follows:

- Summarize the main contributions of the paper, the results, the conclusions in one paragraph.
- Bring up some (obvious) limitations of the assumptions. Describe why you have bounded the impact of these assumptions in your results.
- Bring up some (obvious) future steps for this approach, and motivate them in terms of the new domain-specific results from these future steps.

The use of "obvious" here refers to these limitations and future steps being obvious to you as you have been thinking and discussing this method for quote some time. But they will be, by no means, obvious to the readers and reviewers. You do not need to be very detailed; these can often be at the level of questions you have received at seminars where you discuss this work. If you dig too deeply to find these limitations, you will lose the readers and appear as if you overlooked the obvious limitations of your approach.

9 Citations

9.1 Captions

Table and figure captions should start with a bolded sentence or sentence fragment describing, at a high level, what is being shown. Figure captions should be followed up with a clear description of the figure: what each of the axes represent, the colors of the points/lines, etc. Each panel, if it is a multi-panelled figure, should be described separately. Some writers add a sentence at the end describing the main take home point of the figure. Table captions should include a careful description of what is meant by each row/column, and any details about data in the table (Why are cells missing or out-of-bounds numbers?).

9.2 Citing in text

These are general rules for citing in text that are not always adhered to by other authors.

- citations (both paper, and figures/tables) should not be used as nouns.
- citations to figures and tables should be in parenthesis at the end of the sentence describing the analysis that lead to that figure.
- borrowing significant portions of sentences (more than $\approx$ four words) from a paper means you put that phrase in quotes and cite the paper from which it came.
- cite often and freely. Giving respect to related work, or thoughtful work for that matter, can only increase the quality of the paper.
- any rules about limits to bibliography *can most often be broken*
- rules about limits to figures/tables cannot be broken as easily. Try to ascribe to these; use supplemental material if you have information-rich figures and tables that do not work in the main text.

10 Reviewing

The purpose of a peer review is two-fold: (1) give suggestions that improve the paper which is valuable information for the authors and (2) provide a recommendation to the editors that indicate whether the paper is accepted, returned for revisions, or rejected. Ultimately, it is the job of the Editor to decide whether a manuscript will be accepted or rejected, but your careful reading and suggestions for areas to improve or clarify will generally play a large role in that decision.

Reviews should be kind and be written as though they were not anonymous. Many reviewers sign their reviews. It is often said reviewers can be categorized as gate keepers, or those guarding the journal and literature from lower quality publications, and community builders, or those exposing the research world to new ideas and clear research. It is harder to be the latter, and impossible in certain circumstances.

10.1 How to Read a Manuscript

- How to Read a Technical Paper
- How to Read a CS Research Paper?

10.2 How to Review a Manuscript

Generally there are a number of criteria upon which a review is based. In general, the editor will decide whether or not a manuscript is relevant for the journal or conference but, in some cases, you will be required to determine relevancy.

10.2.1 Technically Sound

As a peer reviewer, your main review criterion is to determine if the methods and the data are *technically sound*. The editor generally does not have the technical background to evaluate each paper, but you do. The editor will rely heavily on your opinions and ideas in this area, and this is by far the most important aspect of a review.

- Are the data appropriately QC'd? Are the outliers controlled for?
- Are the model assumptions appropriate to make in this context?
- Were the number of significant results what you expect? Why or why not? Was there a proper experimental control? Was the null hypothesis evaluated appropriately, and false discovery rate evaluated in a reasonable way?
- Are the conclusions warranted from the results? This is a big question that weaves throughout the Results section.
- Simulation results: Is the method better on simulated data? Were important evaluations removed or avoided? Was the simulation difficult, and reasonable? Were the comparison methods really state-of-the-art and equivalent?
- Real results: what were the conclusions? Do you believe the results?

Much of this is thoughtful reflection on the ideas presented in the paper. Take your time with interesting papers to really understand how they came to their conclusions and what this means in terms of the bigger picture. Your feedback really will make a difference.

10.2.2 Originality

A second criterion is the *originality* of the manuscript which can be interpreted in a number of ways. Are the methods original, the data original, or the conclusions original? Some people disagree that originality should be a criterion of peer review. Regardless, proper domain-specific and methodological context should be a part of every paper to put the contributions of the paper in perspective.

- What are related papers? Are there any obvious fields missing from the related papers? Are they presented fairly? Are the discrepancies in these methods true?
- Is the contribution of the current manuscript appropriate, or an oversell? Is it stated clearly?
- Are the results original? Are they clearly presented in the context of domain-specific results from prior work? Are the original conclusions based on true data?
- Are relevant works properly cited?

If a method is not compared in contrast to many related methods, the paper is often quickly rejected. All it takes is a Google scholar search to determine if others have had similar ideas before. Context is critical to explaining originality.

That said, many reviews (especially for conferences) say:

> this model has two parts: A and B. Our community has studied both A and B well. There is nothing new here.

This is the worst kind of review: lazy and unactionable. Have people ever tried to put A and B together before? Is this a careful study of what happens when A and B are brought together? Is this model well motivated by the specific data application? Do they show that it really works on real data applications?

Related to originality, is the results of the work significant? Does it enable future developments?

10.2.3 Manuscript presentation and layout

The presentation of the ideas is important. If the ideas in the paper are not well presented or if the grammar or notation renders the manuscript difficult to parse, it may not be published.

- Is the paper clearly written and understandable? are all words, acronyms, abbreviations defined? Is technical jargon explained clearly? Do you have to re-read sentences or paragraphs multiple times to get the point?
- Are the Figures and Tables purposeful and helpful to the presentation? Are all axes labeled correctly? Are there simple ways to make them more helpful to the reader?
- Are the equations appropriate in the context of the paper? Are all variables well defined?
- is the structure of the paper appropriate? Are sections missing? Are concepts (e.g., Methods) explained clearly after they are used to draw conclusions from?
- Are there claims that they do not support with clear evidence?
- Are essential proofs missing?

10.3 Some references

Googling "how to review a journal article" yields many results that define this task and give helpful suggestions. Arjun Raj has given helpful advice on this topic.