

CS 2001: Research Topics
Conducting Research

Prof. Adriana Kovashka
University of Pittsburgh
September 29, 2021

Plan for this lecture

- Setting up
- Developing your method
- Experimental validation
- Bookkeeping
- What if results don't look good?
- Sometimes, you give up
- Getting help
- Ethical research
- Research integrity
- Tools (presented by Mingda and Erhan)

Homework

- Write a literature review on a topic of your choice
- Identify a claim you want to make: e.g. “X is a good approach”
- In class, we will pair up and give each other feedback on the literature review
- Aim for a 3-5 page review
- Due in 1 week

Setting Up

- Set up code for a baseline method—whether written from scratch, or code other researchers shared—get your “hands dirty”
- Set up evaluation—make sure it is correct, fair
 - Machine learning: Train/validation/test split (val to pick hyperparameters, use test set once)
 - Relevant metrics used—if proposing new metrics, explain why they are needed
 - Make sure your method isn’t using unfair advantage—anything that isn’t part of your contribution
 - Are the results what you expected? In both cases (yes/no), double-check your work

Developing Your Method

- Start with existing method (baseline), use existing code or implement from scratch (sometimes simpler, published code not always easy to use)
- Think about which parts of your full framework are your actual contribution that you will highlight in the paper, and which parts are standard

Experiments

- You need to show impact of your key contribution
 - Scientific method—change one variable at a time, to show its impact
 - Bad: 3 new method components (2 of which not very interesting), new metric, new dataset—and only compared against 1 baseline
- Gives rise to “ablation” experiments—compare a backbone to the backbone with each of your key ideas added, one at a time, then together
- Also need to compare to methods from the very recent literature—the more the merrier (if done properly)

Bookkeeping

- When developing your method and conducting validation (e.g. running experiments)—it's crucial that you document your work!
- You should be able to reproduce the exact settings and results that you obtained—apart from factors due to randomness (but can set random seed)
- You should know the complete setting of results you include in papers, and all results you show your advisor
- You should include as much information as possible to enable others to reproduce your results, even share code if possible—helps with getting your work cited

Reasons for Rejection

- “Only quantitative results; weak baselines; weak evaluation (automatic/intrinsic vs. human/extrinsic)”
- “Weak ablation study; weak experiments (dataset, competitors, baselines)”

What if Results Don't Look Good?

- Your baseline and evaluation is set up well, but your proposed method doesn't outperform the baseline?
- This is likely to happen, often—try to understand what isn't working well and why, then iterate on method
- Try simple changes first, e.g. tuning algorithm “knobs”
- Come up with simple environments where your new idea should work—e.g. a simpler test set, no-noise environment, attentive users, etc.
- It's ok to iterate on metrics too—perhaps one metric is too coarse to show impact of your method
- Keep reading papers to get more ideas on method

Sometimes, You Give Up

- Sometimes, you iterate multiple times, but nothing works
- When to give up on this overall idea, topic?
 - “Give up” = publish in a workshop
- Really hard to say, but important to know there may be a time you need to stop—1 year?
- Consider whether it is worth investing *even more time* in this particular project
- It doesn't reflect on your ability as a researcher!
- It's not wasted time—use the skills you acquired to tackle a related problem!

Getting Help

- Lots of learning resources on the web
 - Check out tutorials from relevant conferences
- Peers, senior students
 - Formulate your request clearly, make it efficient for them to help
 - Don't hesitate to ask for help—you're just starting, they're an expert!
- Your advisor
 - Often, it makes sense to try to figure things out on your own, unless it's something that advisor can easily and quickly resolve
 - Send email rather than waiting, or show up at their door—if you've discussed with your advisor that's ok; potentially use Slack group
- Reading group
 - Volunteer to present your work
 - Will get help with brainstorming
 - Sometimes just talking about your work helps!

Ethical Research

- Plagiarism—a big problem!
 - Don't claim others' work as your own
 - Don't copy text from others' work—ideally use quotes, but preferably say with your own words
- Verify your code and evaluation is sound!
 - Proofread code often, ask peers to review
 - Don't fabricate results!
- IRB human subjects research
 - Talk to your advisor

Ethical Research (cont'd)

- Keep in mind social impact
 - Who does my research help, hurt?
 - Example: Facial recognition software—help/hurt?
 - Software that can predict speaker's face from their voice—help/hurt?
 - Predicting political bias from politicians' portrayals (study how Obama's face may appear in left/right-leaning media)—help/hurt?

Research Integrity

- “**Research misconduct** is defined as fabrication, falsification, or plagiarism, including misrepresentation of credentials, in proposing, performing, or reviewing research, or in reporting research results. It does not include honest error or differences of opinion. Misconduct as defined above is viewed as a serious professional deviation that is subject to sanctions imposed both by the University, by many professional associations, and, in the case of research proposed to or funded by a federal agency, by that agency.”

Research Integrity (cont'd)

- “A researcher **must not report anticipated research results that had not yet been observed** at the time of submission of the report. In order to preserve accurate documentation of observed facts with which later reports or conclusions can be compared, every researcher has an obligation to **maintain a clear and complete record of data acquired**. As stated in the University's Guidelines on Research Data Management, "records should include sufficient detail to permit examination for the purpose of replicating the research, responding to questions that may result from unintentional error or misinterpretation, establishing authenticity of the records, and confirming the validity of the conclusions.”

Research Integrity (cont'd)

- “Authors who present the words, data, or ideas of others with the implication that they are their own, without attribution in a form appropriate for the medium of presentation, are committing **theft of intellectual property and may be guilty of plagiarism and thus of research misconduct**. This statement applies to reviews and to methodological and background/historical sections of research papers as well as to original research results or interpretations. If there is a word-for-word copying beyond a short phrase or six or seven words of someone else's text, that section should be enclosed in quotation marks or indented and referenced, at the location in the manuscript of the copied material, to the original source. The same rules apply to grant applications and proposals, to clinical research protocols, and to student papers submitted for academic credit. Not only does plagiarism violate the standard code of conduct governing all researchers, but in many cases it could constitute an infraction of the law by infringing on a copyright held by the original author or publisher. An author should cite the work of others even if he or she had been a co-author or editor of the work to be cited or had been an adviser or student of the author of such work.”

Research Integrity (cont'd)

- “One particularly serious form of plagiarism is the **misuse of privileged information taken from a grant application or manuscript received from a funding agency or journal editor for peer review**. In such a case, the plagiarism is a serious matter of theft of intellectual property because it not only deprives the original author of appropriate credit by citation but could also preempt priority of first publication or use of the original idea to which the source author is entitled. Also, one who breaches confidentiality by showing a privileged unpublished document to an unauthorized person can be held to a shared responsibility for any subsequent plagiarism of the document committed by that unauthorized person.”

Research Integrity (cont'd)

- “Research integrity requires not only that reported conclusions are based on accurately recorded data or observations but that **all relevant observations are reported. It is considered a breach of research integrity to fail to report data that contradict or merely fail to support the reported conclusions,** including the purposeful withholding of information about confounding factors. If some data should be disregarded for a stated reason, confirmed by an approved statistical test for neglecting outliers, the reason should be stated in the published accounts. A large background of negative results must be reported. Any intentional or reckless disregard for the truth in reporting observations may be considered to be an act of research misconduct.”

Research Integrity (cont'd)

- “Research data obtained in studies performed at the University of Pittsburgh and/or by employees of the University are **not the property of the researcher** who generated or observed them or even of the principal investigator of the research group. They belong to the University of Pittsburgh, which can be held accountable for the integrity of the data even if the researchers have left the University.”
- “Each student, postdoctoral fellow, or other investigator in a group project should come to an **understanding with the research director or principal investigator**, preferably in writing, about which parts of the project he or she might continue to explore after leaving the research group.”

Research Integrity (cont'd)

- “Since the scientific enterprise may be a cooperative endeavor encompassing many persons who now or in the future might pursue related research interests, and since it is in the interest of all to rely on the contributions and findings of others, every investigator has an obligation to the general scientific community to cooperate by **sharing of data**. Other virtues of sharing data include the facilitation of independent confirmation or refutation of reported outcomes. It is generally accepted that the data underlying a research publication should be made available to other responsible investigators upon request after the research results have been published or accepted for publication.”

Research Integrity (cont'd)

- “Publication must give appropriate **credit to all authors** for their roles in the research. If more than one person contributes significantly, the decision of which names are to be listed as co-authors should reflect the relative contributions of various participants in the research. Many professional associations and research journals have specified criteria for authorship. One common standard appearing in many of these statements is that each author should have participated in formulating the research problem, interpreting the results, and writing the research paper, and should be prepared to defend the publication against criticisms. Other statements require meeting two or three of the above criteria and, with respect to the last of these requirements, a more limited expectation is often prescribed that each author should be prepared to defend against criticism those portions of the publication falling within his or her particular area of expertise. A person's name should not be listed as author without his or her knowledge, permission, and review of the final version of the manuscript that includes the names of all co-authors.”

Research Integrity (cont'd)

- “Research protocols involving **human subjects** must be approved in advance by the University Institutional Review Board (IRB), which determines whether risks posed to subjects are acceptable and whether information describing risks and benefits of subject participation is conveyed to subjects in an accurate and intelligible manner.”
- “Research means ‘...a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge.’ Human subject means ‘...a living individual about whom an investigator (whether professional or student) conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information’ “
- “There may be a number of scholarly activities (e.g., an oral history project) which may not meet the federal definition of research, or other activities (e.g., secondary data analyses using de-identified data) which may meet the federal definition of research but may not meet the federal criteria for the involvement of ‘human subjects.’

Tools (next time)

- LaTeX
- Plotting
- Scripting
- AFS
- Department machines
- Github
- Docker
- etc.