CIS 602-01: Computational Reproducibility

Scientific Writing

Dr. David Koop
http://www.cis.umassd.edu/~dkoop/cis602
Course Structure

- Topics rotate: not the same course as last year
- Designed to focus on current research topics
- Reading Papers
- Presenting and Discussing Material
- Relevant Assignments
- One Midterm
- Final Project
  - Reproduce existing work, or
  - Related research project
Course Information

• Registration
  - Make sure you have registered in COIN
  - If you have not registered, please see me after class

• Policies
  - Academic Honesty: Do not cheat, you will receive a zero \((0)\). Repeat offenders will fail the course.
  - Late Policy: 3 late days for assignments only
Gay Marriage Canvassing Study (2015)

Note: The first vertical dashed line represents the canvassing intervention, which was administered between Internet survey waves 1 and 2. The second vertical dashed line represents the U.S. Supreme Court decisions striking down California’s ban on same-sex marriage. The Y-axis is opinion change between the baseline survey and subsequent survey waves, with higher scores indicating more support for same-sex marriage. Points represent mean values, bars display 95% bootstrap confidence intervals.

[LaCour and Green, Science, 2014 (Retracted 2015)]
Gay Marriage Canvassing Study (2015)

Note: The first vertical dashed line represents the canvassing intervention, which was administered between Internet survey waves 1 and 2. The second vertical dashed line represents the U.S. Supreme Court decisions striking down California’s ban on same-sex marriage. The Y-axis is opinion change between the baseline survey and subsequent survey waves, with higher scores indicating more support for same-sex marriage. Points represent mean values, bars display 95% bootstrap confidence intervals.

[LaCour and Green, Science, 2014 (Retracted 2015)]
Note: The first vertical dashed line represents the canvassing intervention, which was administered between Internet survey waves 1 and 2. The second vertical dashed line represents the U.S. Supreme Court decisions striking down California’s ban on same-sex marriage. The Y-axis is opinion change between the baseline survey and subsequent survey waves, with higher scores indicating more support for same-sex marriage. Points represent mean values, bars display 95% bootstrap confidence intervals.

[LaCour and Green, Science, 2014 (Retracted 2015)]
Green, one of the two authors, requested retraction after these questions arose

• "...survey data **could not be traced** to any originating Qualtrics source files" (emphasis added)

• "...failure to produce the raw data coupled with the other concerns noted above **undermines the credibility of the findings**" (emphasis added)

• Science retracted the paper on May 28, 2015
fMRI
Determining Brain Activity from fMRI

• "[R]esearchers use statistical software to help them separate the signal from the noise. And when these malfunction, they lead to false positives: indications of brain activity when none exist"

  - N. Stockton, Wired

• IgNobel Prize in Neuroscience: The dead salmon study

[Bennet et al., 2009]
fMRI Study Conclusions

• Error rate was 70%!
• Found a bug in one of the software packages used for fMRI analysis (fixed in 2015)
• ~3,500 results may have issues
• "These results question the validity of a number of fMRI studies and may have a large impact on the interpretation of weakly significant neuroimaging results."
• "Due to lamentable archiving and data-sharing practices it is unlikely that problematic analyses can be redone."
• "Brain imagery has this tradition of showing a picture, but the data underlying that image is never shared" (via T. Nichols via Wired)

["Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates", A. Eklund et al., 2016, PNAS]
Gene names and Excel

- First cited in 2004
- Blog posts:
  - https://dontuseexcel.wordpress.com/2013/02/07/dont-use-excel-for-biological-data/
- Studied supplemental data from 18 journals, 35,175 Excel files
- Increased at an annual rate of 15% in the past five years
- Not just Excel: e.g. Apache OpenOffice Calc

["Gene name errors are widespread in the scientific literature", M. Ziemann et al., 2016,]
Technical Papers

• A document that describes scientific research
• Two general categories:
  - Survey: What has been done in a specific area
  - Research: a problem, related work, solution, and results
• Writing helps clarify your own thinking and communicate it to others [N. Feamster]
• "The purpose of research is to increase the store of human knowledge, and so even the very best work is useless if you cannot effectively communicate it to the rest of the world." — M. Ernst
• Research papers are primary sources, textbooks are secondary sources
• Most recent research is not in a textbook
• Technical Reports vs. Journal Articles/Conference Proceedings
Paper Structure

- Title & Author List
- Abstract
- Introduction
- [Background/Preliminaries]
- Contribution (Approach/Theory/Specification/Implementation)
- Evaluation (Experiments, case studies)
- [Discussion]
- Related Work (here or after introduction)
- Conclusion [& Future Work]
- [Appendices]
Paper Titles

• Something of an art
• Want to informative and capture essence of the paper
• Sometimes catchy but should not be too "cute"
• Differs by domain:
  - "The histone lysine methyltransferase KMT2D sustains a gene expression program that represses B cell lymphoma development" [Nature Medicine]
  - "Time Curves: Folding Time to Visualize Patterns of Temporal Evolution in Data" [IEEE Vis]
Author List

THE AUTHOR LIST: GIVING CREDIT WHERE CREDIT IS DUE

The first author
Senior grad student on the project. Made the figures.

The second-to-last author
Ambitious assistant professor or post-doc who instigated the paper.

The third author
First year student who actually did the experiments, performed the analysis and wrote the whole paper. Thinks being third author is “fair”.

The second author
Grad student in the lab that has nothing to do with this project, but was included because he/she hung around the group meetings (usually for the food).

The middle authors
Author names nobody really reads. Reserved for undergrads and technical staff.


The last author
The head honcho. Hasn’t even read the paper but, hey, he got the funding, and his famous name will get the paper accepted.

[Piled Higher and Deeper, J. Cham, 3/13/2005]
Abstract

- Needs to summarize problem, approach, results, and conclusions
- Has the key contributions from the paper
- Should tell a reader whether they want to read further
- Vocabulary should describe work to a more general audience
- Does not attempt to cover everything in the paper but should highlight key points
Introduction

• Defines the problem
• Motivation: Why do I care about this?
• What is this paper doing that is new/different from what already has been done?
• Outlines the approach and results
Background vs. Related Work

- Paper may need to review notation, information from another domain, the existing work
- Related work serves to define **areas of interest** to the reader and how they relate to this paper
- Sometimes at the beginning (Section 2) or the end (Section n-1) of the paper
- Citations should occur **throughout** the paper not just in Background and Related Work sections
- References are located at the end of the paper
References

I've been reading the latest draft of your paper, Cecilia.

And?

You need to include more references.

Really?

Yes, in academic writing every single statement must be supported by data or credited to a reliable source.

 Says who?

Me, et al.

[Piled Higher and Deeper, J. Cham, 9/11/2015]
Contribution

- Theoretical or experimental
- May be broken into multiple sections
- For computer science techniques, often broken into a framework/specification, and the implementation
  - Framework describes the main contribution at a conceptual level, may include requirements
  - Implementation is secondary but gives readers an idea of the actual code (code can be made available on the Web)
  - Pseudocode is usually used for specific algorithms.
- Should provide details that allow other computer scientists to recreate the proofs or technique
- Not a daily journal—tell a story that argues for the importance of the results
Evaluation

• Need some way of judging whether the presented work matters
• Does it clearly support or refute a hypothesis (e.g. technique is faster, allows users to better understand data, etc.)?
• Comparisons to existing work are important
• What can be evaluated? [Widom]
  - Running time, parameter sensitivity, scalability, user perception
  - Absolute performance, relative performance (comparisons!)
Discussion

• What do the results suggest?
• Section to speculate
• Sometimes tied to related work
Conclusion & Future Work

• Summarize work more concretely
• Not just a rewrite of the abstract or introduction

• Gives a clue to others interested in the area about what else should be explored
• Sometimes stakes ground to show work is continuing
Citations and References

• Credit those who have done work already
• Always try to cite a paper or book to support a claim
• Any claim that someone may question or disagree with should be cited
• Cite refereed papers not blog posts!
• Citation formats:
  - ACM
  - IEEE
• Think of citation like a persistent web page link:
  - Allows readers to easily track down existing work
  - Makes it clear what the reference relates to
• Do not only put references at the end of a paper!
Citations in Presentations

• Applies to presentations as well
• Put citations inline ["CIS 602-01 Lecture 2", D. Koop, 2016]
• Cite figures and images taken from elsewhere
Quoting References

• In general, do not copy even a sentence from another source even if it is properly cited.
• Direct quotes are used to emphasize the specific terminology the author has used
  - Example: state an exact definition from another source
  - Surround all copied text by quotation marks
• Using a thesaurus to change words is not allowed either
• Your writing should reflect your own thinking
  - Read the paper, take notes of key points
  - Put the paper away
  - Write a response
Peer-reviewed Papers

• Scientific papers are reviewed by other scientists before being published
  - Papers are submitted to journal/conference
  - Assigned to external reviewers who provided critical feedback
  - The primary reviewer or editor writes a summary review
  - The editor or program committee decides on acceptance or what necessary rewrites should be done

• The fact that other experts have read and provided feedback on the science is important

• Citing Web URLs or unreviewed papers is usually done sparingly
Example Papers

http://www.cis.umassd.edu/~dkoop/
Reading Papers

[Image of a comic strip showing a person reading a paper, followed by looking at figures, and then searching the web. The strip concludes with a note from the author saying, "I did read that paper you gave me... I'll definitely add it to my reference list."]

[Piled Higher and Deeper, J. Cham, 6/1/2001]
Questions when reading

- "How to Read and Evaluate Technical Papers", B. Griswold modified by G. Murphy
- Sometimes useful to read the paper "out of order"
- Five questions you should answer when reading a paper:
  1. What are the motivations for this work?
     - People problem
     - Technical problem
  2. What is the proposed solution?
  3. What is the evaluation of the proposed solution?
  4. What are the contributions?
  5. What are future directions for this research?
How to Read a Paper

• "How to Read a Paper", S. Keshav

• Make multiple passes over the paper
  - First pass: title, abstract, introduction, headings, conclusion, references
  - Second pass: read but ignore details, study figures
  - Third pass: virtually re-implement the paper
Missing Background

• You will get papers that do not explain all relevant background
  - May be outside the area you focus on
  - May involve an application you're not as familiar with
  - There isn't enough space to describe all details

• Strategies:
  - Search for key terms (e.g. Google Scholar)
  - Look at the referenced work and read/skim those papers
  - Check other work in the journal/venue
  - Read examples to clarify the proposed technique or goal
Writing a Critical Reading Response

1. Describe, in your own words, what the problem addressed is and what the key contributions are.

2. Respond to the paper:
   - How would you add to the work that was presented?
   - What evaluation was not done that should have been?
   - No vague statements like "The paper is well-written"
   - Does the direction of the work make sense?
   - Questions are fine, but they should be specific and show your understanding.
   - Keep track of points in favor, points against.
   - Should focus on specific parts of the paper, make sure you understand everything about that part of the technique/system.
Responding to a Paper [Griswold]

- What is your analysis of the identified problem, idea and evaluation?
  - Is this a good idea?
  - What flaws do you perceive in the work?
  - What are the most interesting points made?
  - What are the most controversial ideas or points made?
  - For work that has practical implications, you also want to ask: Is this really going to work, who would want it, what it will take to give it to them, and when might it become a reality?

- What questions are you left with?
  - What questions would you like to raise in an open discussion of the work?
  - What do you find confusing or difficult to understand?
Evaluating Responses

• Can I see that you understood the paper?
• Can I see that you thought critically about the paper?
• Style and grammar are important
  - Writing with spelling and grammar mistakes is of lower value
  - Use spelling and grammar checkers…
  - …but also read your own writing
Common Technical Writing Issues

Tao Xie

More advice on writing

• Practice
  - Just do it
  - Read back your own writing
  - Have others read your writing

• Read
  - Well-written papers
  - Other literature

• Look for how words are generally used
  - http://www.oxfordlearnersdictionaries.com
  - http://corpus.byu.edu/coca/

• Do not plagiarize
  - Use direct quotes when the exact words is important
Writing Tools

- LaTeX is the standard for research papers in computer science
  - Extension of TeX by Donald Knuth
- Compile your paper: computes character spacing, kerning
- BibTeX integrates with LaTeX to make keeping track of references much easier (similar to Endnote for Word)
- Collaborative writing: [https://www.sharelatex.com](https://www.sharelatex.com)
- markdown -> LaTeX (via, e.g., pandoc)
Reading Assignment

• Read the papers:
  - Reproducible Research in Computational Science, R. D. Peng, 2011
  - Ten Simple Rules for Reproducible Computational Research, G. K. Sandve et al., 2013
  - Computational Reproducibility, J. Freire et al., 2012

• Write your response to these papers:
  - Summarize themes and key contributions (1 paragraph)
  - Offer your reaction and critique (1-2 paragraphs): Describe why the paper is important, what concepts you agree or disagree with, and how the work might be extended or integrated with other work.
Writing Assignment

• Due next Tuesday, 9/20
• Read papers for Thursday's class
• We will discuss them in class on Thursday
• Please update your reaction and turn in Tuesday by 12pm (before class)