CIS 602-01: Computational Reproducibility

Reproducibility Overview

Dr. David Koop
Technical Papers

- A document that describes scientific research
- Three general categories:
  - Survey: What has been done in a specific area
  - Research: a problem, related work, solution, and results
  - Position: ideas about the direction of a field, next steps
- 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
- 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]
Author List

THE AUTHOR LIST: GIVING CREDIT WHERE CREDIT IS DUE

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

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


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]
References

[Piled Higher and Deeper, J. Cham, 9/11/2015]
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!
Reading Papers

Jorge Cham © The Stanford Daily

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
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?
Common Technical Writing Issues

Tao Xie

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
• markdown -> LaTeX (via, e.g., pandoc)
Reproducible Research in Computational Science

R. D. Peng
Reproducible Research

• Science is verified by replicating work independently

• Replication Issues:
  - Requires many resources to replicate (Sloan Digital Sky Survey)
  - Requires significant computing power (Climate Model Simulation)
  - Requires too much time or very specific circumstances (Environment Epidemiology)

• Reproducibility
  - Replication of the analysis based on the collected data (not replicating the data collection itself)
  - Better if we have the actual code or available executables
Reproducibility Spectrum

Reproducibility Spectrum

Publication only

Publication +

Code

Code and data

Linked and executable code and data

Full replication

Not reproducible

Gold standard
Published Papers

• “It’s impossible to verify most of the results that computational scientists present at conference and in papers.” [Donoho et al., 2009]

• “Scientific and mathematical journals are filled with pretty pictures of computational experiments that the reader has no hope of repeating.” [LeVeque, 2009]

• “Published documents are merely the advertisement of scholarship whereas the computer programs, input data, parameter values, etc. embody the scholarship itself.” [Schwab et al., 2007]
Literate Programming

• Knuth’s WEB system
• Mathematica
• Code this is well-documented using comments
• Scientific Notebooks
Data and Code Availability

• Code Repositories:
  - Sourceforge
  - GitHub
  - Launchpad
  - ...

• Data Repositories:
  - DataONE, figshare, freebase, dryad
  - Also many domain-specific repositories
  - http://oad.simmons.edu/oadwiki/Data_repositories
Ten Simple Rules for Reproducible Computational Research

G. K. Sandve, A. Nekrutenko, J. Taylor, E. Hovig
Rules for Reproducible Computational Research

- Rule 1: For Every Result, Keep Track of How It Was Produced
- Rule 2: Avoid Manual Data Manipulation Steps
- Rule 3: Archive the Exact Versions of All External Programs Used
- Rule 4: Version Control All Custom Scripts
- Rule 5: Record All Intermediate Results, When Possible in Standardized Formats

[Sandve et al., 2013]
Rules for Reproducible Computational Research

• Rule 6: For Analyses That Include Randomness, Note Underlying Random Seeds
• Rule 7: Always Store Raw Data behind Plots
• Rule 8: Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
• Rule 9: Connect Textual Statements to Underlying Results
• Rule 10: Provide Public Access to Scripts, Runs, and Results
Computational Reproducibility: State-of-the-Art, Challenges, and Database Research Opportunities

J. Freire, P. Bonnet, and D. Shasha
Rules or Benefits?

- Laws to make sure people don't cheat or lie or steal
- Is that a good incentive? You won't be mislabeled as a criminal?
- Benefits of Reproducibility
  - Reproducible programs can be compared
  - Reproducible software and results are documented
  - Reproducible software is portable
  - Reproducible experiments are cited
Reproducible Experiments Classification

- Depth: how much is available?
  - figures
  - scripts
  - raw data
  - experiments
  - software system

- Portability: what machine specs are necessary?
  - same machine
  - similar machine
  - different OS

- Coverage: how much can be reproduced?
(Database) Research Topics

- Design and Management of Experiment Repositories
- Querying and Searching Experiments
- Mining Experiments