Intro I: Motivating reproducibility

Reproducible Science Workshop

Tagline and mission

Tagline: Accelerating scientific progress through reproducible science

Mission: To train researchers in the best practices and approaches of reproducible research and accelerate scientific progress.

Intro to Reproducible Research outline

- Getting to know you
- Recognize the problems that reproducible research helps address via a case studies
- **Exercise**: Identify pain points in getting your analysis to be reproducible
- Introduce tools that address these pain points
- **Demo**: Computational toolkit for reproducibile data analysis

Getting to know you. . .

with respect to reproducibility

Science retracts gay marriage paper without agreement of lead author

- Science retracted a study of how canvassers can sway people's opinions about gay marriage.
- Science Editor-in-Chief: Original survey data not made available for independent reproduction of results + Survey incentives misrepresented + Sponsorship statement false
- Two Berkeley grad students attempted to replicate the study and discovered that the data must have been faked.
- Methods we'll discuss today can't prevent this, but they can make it easier to discover such issues.

Source: `http://news.sciencemag.org/policy/2015/05/science-retracts-gay-marriage-paper-without-lead-author-s-`

Seizure study retracted after authors realize data got "terribly mixed"

From the authors of **Low Dose Lidocaine for Refractory Seizures in Preterm Neonates**:

"*The article has been retracted at the request of the authors. After carefully re-examining the data presented in the article, they identified that data of two different hospitals got terribly mixed. The published results cannot be reproduced in accordance with scientific and clinical correctness.*"

Source: `http://retractionwatch.com/2013/02/01/seizure-study-retracted-after-authors-realize-data-got-ter`

Bad spreadsheet merge kills depression paper, quick fix resurrects it

- The authors informed the journal that the merge of lab results and other survey data used in the paper resulted in an error regarding the identification codes.
- **Original conclusion:** Lower levels of CSF IL-6 were associated with current depression and with future depression [. . .].
- **Revised conclusion:** Higher levels of CSF IL-6 and IL-8 were associated with current depression [. . .].

Source: `http://retractionwatch.com/2014/07/01/bad-spreadsheet-merge-kills-depression-paper-quick-fix-res`

Exercise: Motivating reproducibility

This is a two-part exercise:

Part 1: Analyze + document

Part 2: Swap + discuss

Part 1: Analyze + document

Complete the following tasks and **write instructions / documentation** for your collaborator to reproduce your work starting with the original dataset (`data/gapminder-5060.csv`).

1. Visualize life expectancy over time for Canada in the 1950s and 1960s using a line plot.
2. Something is clearly wrong with this plot! Turns out there's a data error in the data file: life expectancy for Canada in the year 1957 is coded as `999999`, it should actually be `69.96`. Make this correction.
3. Visualize life expectancy over time for Canada again, with the corrected data. *Stretch goal:* Add lines for Mexico and United States.

Part 2: Swap + discuss

Introduce yourself to your collaborator and tell them why you're here.

1. Swap instructions / documentation with your collaborator, and try to reproduce their work, first **without talking to each oher**. If your collaborator does not have the software they need to reproduce your work, we encourage you to either help them install it or walk them through it on your computer in a way that would emulate the experience. (Remember, this could be part of the irreproducibility problem!)
2. Then, talk to each other about challenges you faced (or didn't face) or why you were or weren't able to reproduce their work.

Reflection

- What tools did you use (Excel / R / Python / Word / plain text etc.)?
- Were you successful in reproducing each others' work?

- What would happen if your collaborator is no longer available to walk you through their analysis?
- What made it easy / hard for reproducing your partners' work?
- What would have to happen if
  - you had to swap out the dataset or extend the analysis further?
  - you caught further data errors and had to re-create the analysis with corrections?
  - you had to revert back to the original dataset?

Summary

- Everyone struggles with reproducibility and it is a hindrance to moving science forward
- Evan with a fairly simple analysis challenges were faced in four main areas: organization, documentation, automation, and dissemination
- Over the two day workshop, data analysis tasks will become more complex as we gather more data and ask more complicated questions, so we need better tools and workflows to combat issues arising in these areas

Four facets of reproducibility

1. **Documentation:** difference between binary files (e.g. docx) and text files and why text files are preferred for documentation
   - *Protip:* Use markdown to document your workflow so that anyone can pick up your data and follow what you are doing
   - *Protip:* Use literate programming so that your analysis and your results are tightly connected, or better yet, unseperable
2. **Organization:** tools to organize your projects so that you don't have a single folder with hundreds of files
3. **Automation:** the power of scripting to create automated data analyses
4. **Dissemination:** publishing is not the end of your analysis, rather it is a way station towards your future research and the future research of others