Section A01
CDS 101 *Introduction to Computational and Data Sciences*

Class 2: The data scientist's toolbox

May 22, 2018

General

Annoucements

- If you haven't introduced yourself on the Slack channel yet, please do so!
- Reading:
    - *R for Data Science:* Chapters 1, 26 (short) , and 27, complete by Wednesday, May 23rd at 9:00am
- RMarkdown mini-assignment (to be posted) due Wednesday, May 24th by 11:59pm

Class agenda

- Motivation for "Reproducible research"
- How does Github work?

Motivation for Reproducible Research

The scientific method

1. Review evidence

The scientific method

1. Review evidence
2. Hypothesis

The scientific method

1. Review evidence
2. Hypothesis
3. Formulate predictive test

The scientific method

1. Review evidence
2. Hypothesis
3. Formulate predictive test
4. Design/run experiment

The scientific method

1. Review evidence
2. Hypothesis
3. Formulate predictive test
4. Design/run experiment
5. Validate or revise hypothesis

The scientific method

1. Review evidence
2. Hypothesis
3. Formulate predictive test
4. Design/run experiment
5. Validate or revise hypothesis

- Key point: create a hypothesis and test it out

The scientific method

1. Review evidence
2. Hypothesis
3. Formulate predictive test
4. Design/run experiment
5. Validate or revise hypothesis

- Key point: create a hypothesis and test it out
- Validation by the natural world ("Nature")

The scientific method

1. Review evidence
2. Hypothesis
3. Formulate predictive test
4. Design/run experiment
5. Validate or revise hypothesis

- Key point: create a hypothesis and test it out
- Validation by the natural world ("Nature")
- Anyone can double check!

Reproducibility in practice

- Sometimes easier said than done, various reasons why

Reproducibility in practice

- Sometimes easier said than done, various reasons why
  - Lack of funding sources

Reproducibility in practice

- Sometimes easier said than done, various reasons why
  - Lack of funding sources
  - Lack of data sharing

Reproducibility in practice

- Sometimes easier said than done, various reasons why
  - Lack of funding sources
  - Lack of data sharing
  - Lack of interest

Reproducibility in practice

- Sometimes easier said than done, various reasons why
  - Lack of funding sources
  - Lack of data sharing
  - Lack of interest
  - "Top-tier" journals won't publish

Reproducibility in practice

- Sometimes easier said than done, various reasons why
  - Lack of funding sources
  - Lack of data sharing
  - Lack of interest
  - "Top-tier" journals won't publish
  - Vague methods

Reproducibility in practice

- Sometimes easier said than done, various reasons why
  - Lack of funding sources
  - Lack of data sharing
  - Lack of interest
  - "Top-tier" journals won't publish
  - Vague methods
- It's very important that we have reproduced research, because...

The Reproducibility Project

> *Brian Nosek of University of Virginia and colleagues sought out to replicate 100 different studies that all were published in 2008. The project pulled these studies from three different [psychology] journals... to see if they could get the same results as the initial findings. [...] Only 36.1% of the studies [were] replicated.*
>
> — Reproducibility Project Wikipedia entry

Science retracts gay marriage paper without agreement of lead author LaCour

- In May 2015 Science retracted a study of how canvassers can sway people's opinions about gay marriage published just 5 months ago.

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

Science retracts gay marriage paper without agreement of lead author LaCour

- In May 2015 Science retracted a study of how canvassers can sway people's opinions about gay marriage published just 5 months ago.
- Science Editor-in-Chief Marcia McNutt: Original survey data not made available for independent reproduction of results.

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

Science retracts gay marriage paper without agreement of lead author LaCour

- In May 2015 Science retracted a study of how canvassers can sway people's opinions about gay marriage published just 5 months ago.
- Science Editor-in-Chief Marcia McNutt: Original survey data not made available for independent reproduction of results.
    - Survey incentives misrepresented.

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

Science retracts gay marriage paper without agreement of lead author LaCour

- In May 2015 Science retracted a study of how canvassers can sway people's opinions about gay marriage published just 5 months ago.
- Science Editor-in-Chief Marcia McNutt: Original survey data not made available for independent reproduction of results.
    - Survey incentives misrepresented.
    - Sponsorship statement false.

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

Science retracts gay marriage paper without agreement of lead author LaCour

- In May 2015 Science retracted a study of how canvassers can sway people's opinions about gay marriage published just 5 months ago.
- Science Editor-in-Chief Marcia McNutt: Original survey data not made available for independent reproduction of results.
    - Survey incentives misrepresented.
    - Sponsorship statement false.
- Two Berkeley grad students who attempted to replicate the study quickly discovered that the data must have been faked.

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

Science retracts gay marriage paper without agreement of lead author LaCour

- In May 2015 Science retracted a study of how canvassers can sway people's opinions about gay marriage published just 5 months ago.
- Science Editor-in-Chief Marcia McNutt: Original survey data not made available for independent reproduction of results.
    - Survey incentives misrepresented.
    - Sponsorship statement false.
- Two Berkeley grad students who attempted to replicate the study quickly 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 issues.

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

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

> *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.*
>
> — Authors of **Low Dose Lidocaine for Refractory Seizures in Preterm Neonates**

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

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. Results of the analyses were based on the data set in which this error occurred. Further analyses established the results reported in this manuscript and interpretation of the data are not correct.

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

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. Results of the analyses were based on the data set in which this error occurred. Further analyses established the results reported in this manuscript and interpretation of the data are not correct.

> ***Original conclusion:*** *"Lower levels of CSF IL-6 were associated with current depression and with future depression [...]".*

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

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. Results of the analyses were based on the data set in which this error occurred. Further analyses established the results reported in this manuscript and interpretation of the data are not correct.

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-resurrects-it/

Reproducibility: why should we care?

Two-pronged approach

- Convince researchers to adopt a reproducible research workflow
- Train new researchers who don't have any other workflow

Reproducible data analysis

- Scriptability → R
- Literate programming → R Markdown
- Version control → Git / GitHub

Scripting and literate programming

Donald Knuth *Literate Programming* (1983)

> *Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do.*
>
> — Donald Knuth in *Literate Programming* (1983)

Donald Knuth *Literate Programming* (1983)

> *Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do.*
>
> — Donald Knuth in *Literate Programming* (1983)

- These ideas have been around for years!

Donald Knuth *Literate Programming* (1983)

> *Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do.*
>
> — Donald Knuth in *Literate Programming* (1983)

- These ideas have been around for years!
- and tools for putting them to practice have also been around

Donald Knuth *Literate Programming* (1983)

> *Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do.*
>
> — Donald Knuth in *Literate Programming* (1983)

- These ideas have been around for years!
- and tools for putting them to practice have also been around
- but they have never been as accessible as the current tools

Reproducibility checklist

- Are the tables and figures reproducible from the code and data?
- Does the code actually do what you think it does?
- In addition to what was done, is it clear **why** it was done? (e.g., how were parameter settings chosen?)
- Can the code be used for other data?
- Can you extend the code to do other things?

Credits

These slides were adapted from the following sources:

- The Introduction to R/Rstudio and git/GitHub slides developed by Mine Çetinkaya-Rundel and made available under the CC BY-NC-SA 4.0 license.