Reproducible Data Science

Fundraising Analytics with GitHub and R Markdown

Paul Hively, Northwestern University

The DRIVE/ Conference

March 11-13, 2019 Baltimore, Maryland

What to expect today

- Paul Hively, Director of Analytics at Kellogg School of Management, Northwestern University
- A few bad jokes
- Plenty of real-world examples and analogies
- How the reproducible approach can **greatly increase** future efficiency and impact

Outline

- The case for reproducibility
  - What?
  - Why?
  - How?
- Tools & use cases
  - GitHub
  - R Markdown
- Demonstration

The case for reproducibility

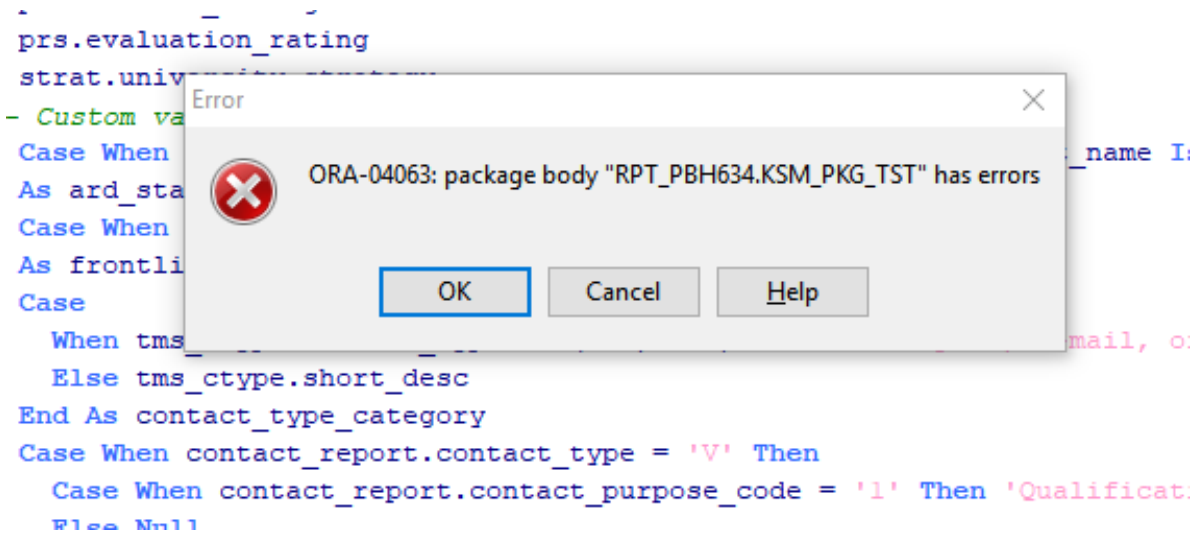
Do any of these situations look familiar?

Sudden show-stoppers: “But it worked yesterday!”

The case for reproducibility

Do any of these situations look familiar?

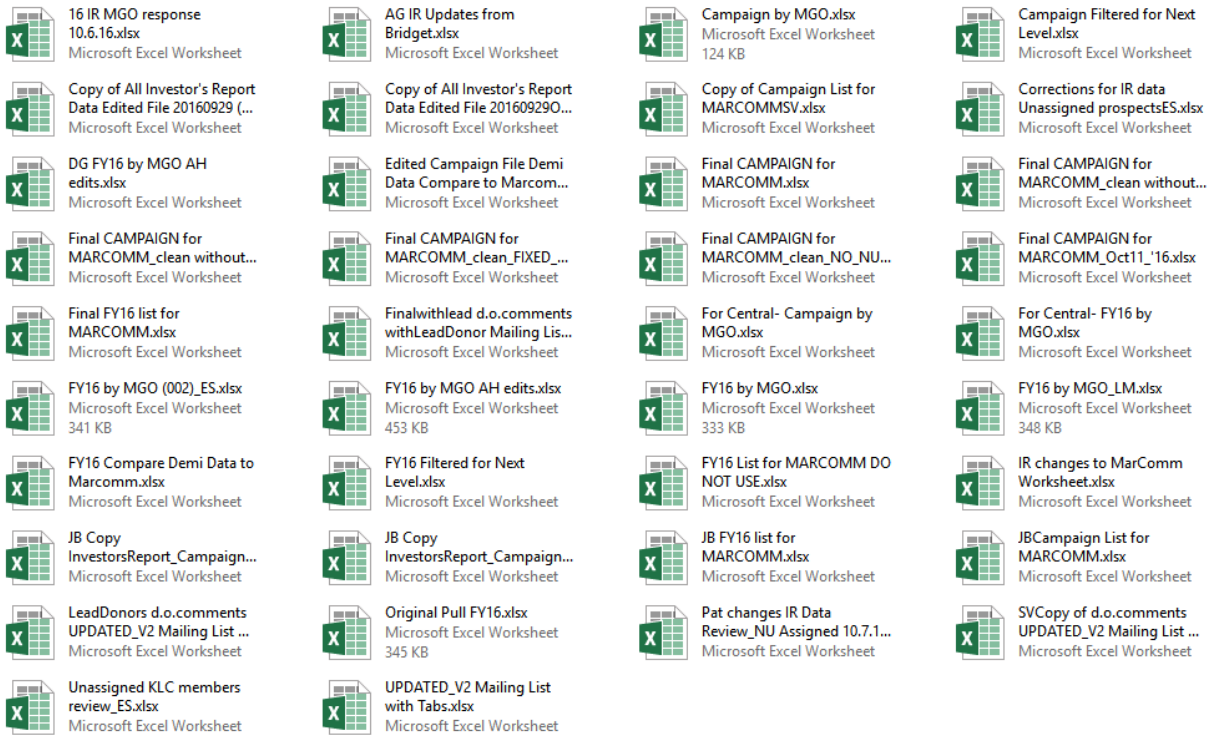
Even better: “But it worked until I hit Save!”



The case for reproducibility

Do any of these situations look familiar?

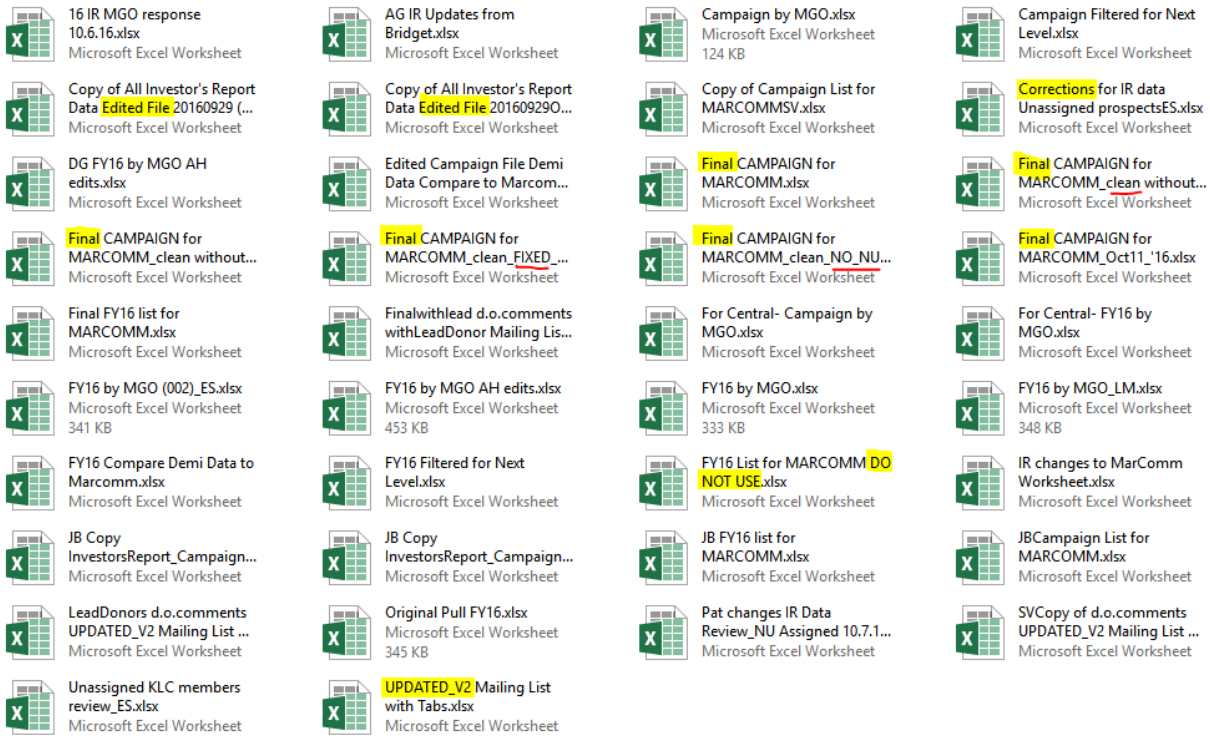
File mayhem: “Find the master list”



The case for reproducibility

Do any of these situations look familiar?

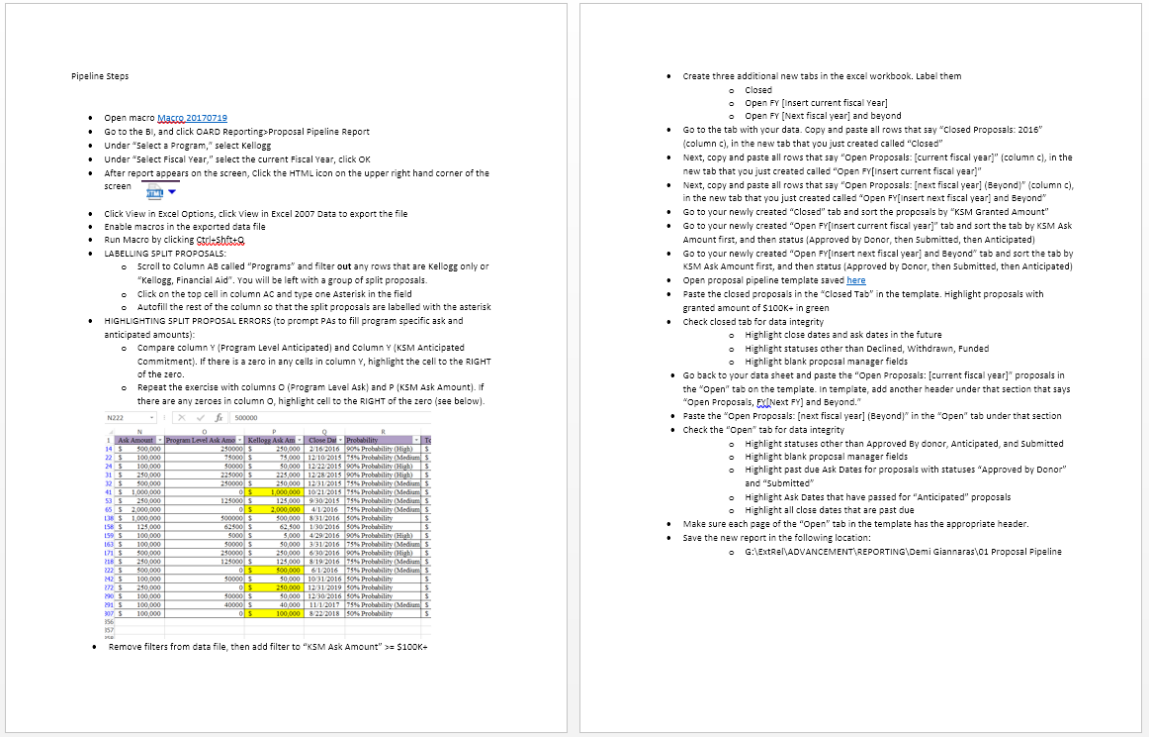
Some of my favorite features:



The case for reproducibility

Do any of these situations look familiar?

Excel Olympics: “Just follow these 60 quick steps!”



Pipeline Steps

- Open macro Macro 20170719
- Go to the BI, and click OARD Reporting>Proposal Pipeline Report
- Under “Select a Program,” select Kellogg
- Under “Select Fiscal Year,” select the current Fiscal Year, click OK
- After report appears on the screen, Click the HTML icon on the upper right hand corner of the screen
- Click View in Excel Options, click View in Excel 2007 Data to export the file
- Enable macros in the exported data file
- Run Macro by clicking Ctrl+Shift+Q
- LABELLING SPLIT PROPOSALS:
  - Scroll to Column AB called “Programs” and filter **out** any rows that are Kellogg only or “Kellogg, Financial Aid”. You will be left with a group of split proposals.
  - Click on the top cell in column AC and type one Asterisk in the field
  - Autofill the rest of the column so that the split proposals are labelled with the asterisk
- HIGHLIGHTING SPLIT PROPOSAL ERRORS (to prompt PAs to fill program specific ask and anticipated amounts):
  - Compare column Y (Program Level Anticipated) and Column Y (KSM Anticipated Commitment). If there is a zero in any cells in column Y, highlight the cell to the RIGHT of the zero.
  - Repeat the exercise with columns O (Program Level Ask) and P (KSM Ask Amount). If there are any zeroes in column O, highlight cell to the RIGHT of the zero (see below).

N222 | 500000

	N	O	P	Q	R	
1	Ask Amount	Program Level Ask Amt	Kellogg Ask Am	Close Dat	Probability	Ye
14	$ 500,000	250000	$ 250,000	2/16/2016	90% Probability (High)	$
22	$ 100,000	75000	$ 75,000	12/10/2015	75% Probability (Medium)	$
24	$ 100,000	50000	$ 50,000	12/22/2015	90% Probability (High)	$
31	$ 250,000	225000	$ 225,000	12/28/2015	90% Probability (High)	$
32	$ 500,000	250000	$ 250,000	12/31/2015	75% Probability (Medium)	$
41	$ 1,000,000	0	$ 1,000,000	10/21/2015	75% Probability (Medium)	$
53	$ 250,000	125000	$ 125,000	9/30/2015	75% Probability (Medium)	$
65	$ 2,000,000	0	$ 2,000,000	4/1/2016	75% Probability (Medium)	$
138	$ 1,000,000	500000	$ 500,000	8/31/2016	50% Probability	$
158	$ 125,000	62500	$ 62,500	1/30/2016	50% Probability	$
159	$ 100,000	5000	$ 5,000	4/29/2016	90% Probability (High)	$
163	$ 100,000	50000	$ 50,000	3/31/2016	75% Probability (Medium)	$
171	$ 500,000	250000	$ 250,000	6/30/2016	90% Probability (High)	$
218	$ 250,000	125000	$ 125,000	8/19/2016	75% Probability (Medium)	$
222	$ 500,000	0	$ 500,000	6/1/2016	75% Probability (Medium)	$
242	$ 100,000	50000	$ 50,000	10/31/2016	50% Probability	$
272	$ 250,000	0	$ 250,000	12/31/2019	50% Probability	$
290	$ 100,000	50000	$ 50,000	12/30/2016	50% Probability	$
591	$ 100,000	40000	$ 40,000	11/1/2017	75% Probability (Medium)	$
807	$ 100,000	0	$ 100,000	8/22/2018	50% Probability	$
156						
157						

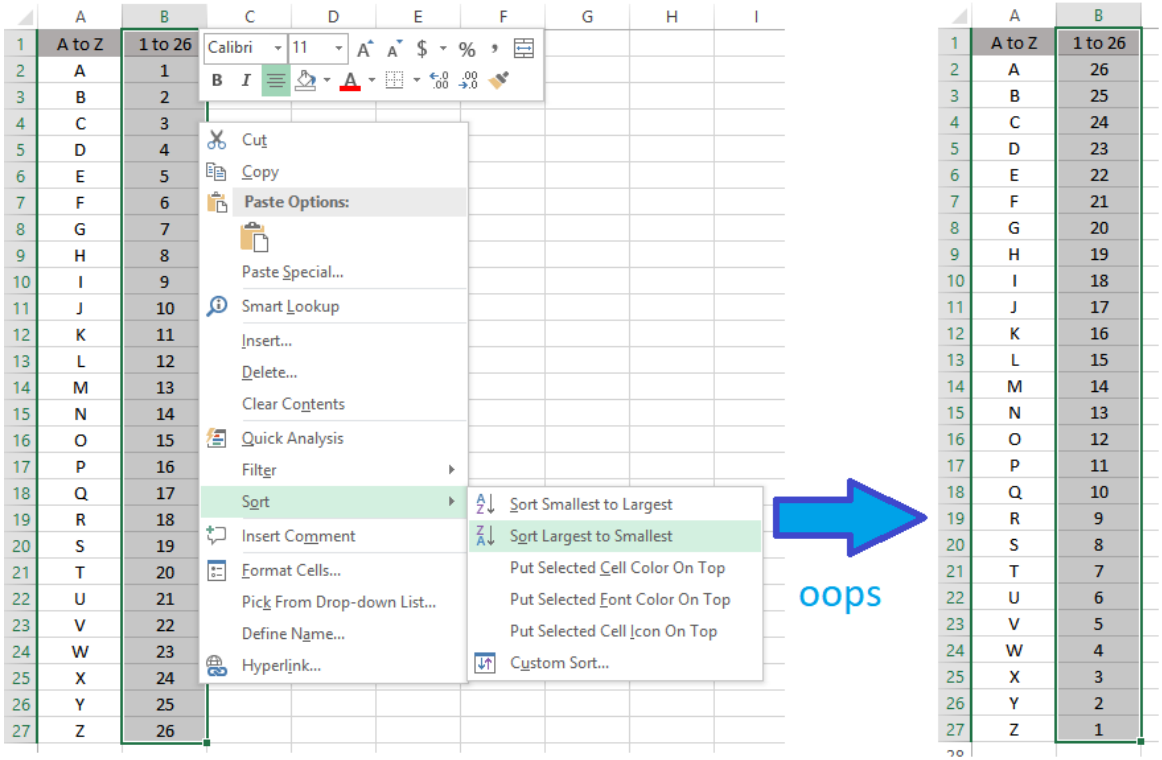
- Remove filters from data file, then add filter to “KSM Ask Amount” >= $100K+

- Create three additional new tabs in the excel workbook. Label them
  - Closed
  - Open FY [Insert current fiscal Year]
  - Open FY [Next fiscal year] and beyond
- Go to the tab with your data. Copy and paste all rows that say “Closed Proposals: 2016” (column c), in the new tab that you just created called “Closed”
- Next, copy and paste all rows that say “Open Proposals: [current fiscal year]” (column c), in the new tab that you just created called “Open FY[insert current fiscal year]”
- Next, copy and paste all rows that say “Open Proposals: [next fiscal year] (Beyond)” (column c), in the new tab that you just created called “Open FY[insert next fiscal year] and Beyond”
- Go to your newly created “Closed” tab and sort the proposals by “KSM Granted Amount”
- Go to your newly created “Open FY[insert current fiscal year]” tab and sort the tab by KSM Ask Amount first, and then status (Approved by Donor, then Submitted, then Anticipated)
- Go to your newly created “Open FY[insert next fiscal year] and Beyond” tab and sort the tab by KSM Ask Amount first, and then status (Approved by Donor, then Submitted, then Anticipated)
- Open proposal pipeline template saved here
- Paste the closed proposals in the “Closed Tab” in the template. Highlight proposals with granted amount of $100K+ in green
- Check closed tab for data integrity
  - Highlight close dates and ask dates in the future
  - Highlight statuses other than Declined, Withdrawn, Funded
  - Highlight blank proposal manager fields
- Go back to your data sheet and paste the “Open Proposals: [current fiscal year]” proposals in the “Open” tab on the template. In template, add another header under that section that says “Open Proposals, FY[Next FY] and Beyond.”
- Paste the “Open Proposals: [next fiscal year] (Beyond)” in the “Open” tab under that section
- Check the “Open” tab for data integrity
  - Highlight statuses other than Approved By donor, Anticipated, and Submitted
  - Highlight blank proposal manager fields
  - Highlight past due Ask Dates for proposals with statuses “Approved by Donor” and “Submitted”
  - Highlight Ask Dates that have passed for “Anticipated” proposals
  - Highlight all close dates that are past due
- Make sure each page of the “Open” tab in the template has the appropriate header.
- Save the new report in the following location:
  - G:\ExtRel\ADVANCEMENT\REPORTING\Demi Giannaras\01 Proposal Pipeline

The case for reproducibility

Do any of these situations look familiar?

Excel Olympics: Sorting SNAFU

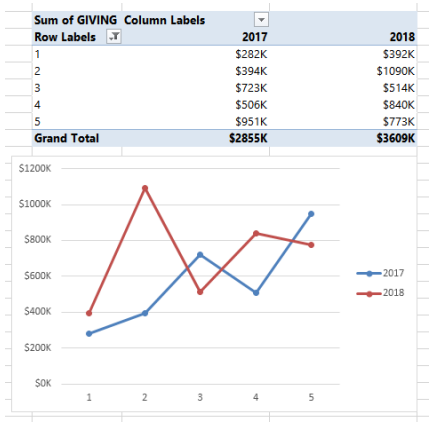


The case for reproducibility

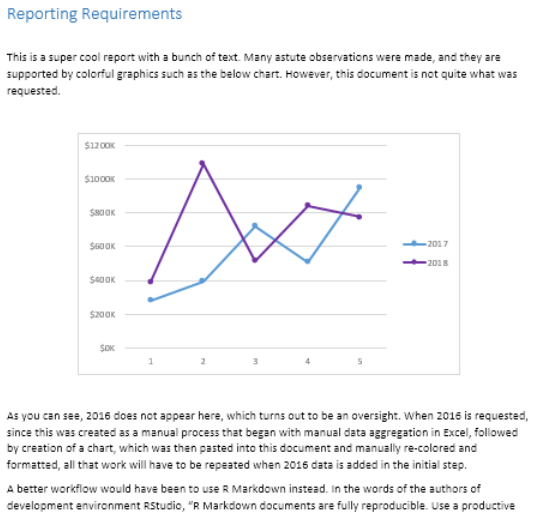
Do any of these situations look familiar?

Copy-Paste Purgatory

FISCAL_YEAR	DATE_OF_RECORD	GIVING	MO	DAY
2017	2/8/2017	33805	2	8
2018	2/1/2018	8700	2	1
2018	12/31/2017	647935	12	31
2018	5/1/2018	97601.23	5	1
2017	4/13/2017	17853.87	4	13
2017	10/17/2016	21432.31	10	17
2018	11/3/2017	11574.5	11	3
2018	2/6/2018	19098.18	2	6
2018	9/29/2017	83676.02	9	29
2017	4/26/2017	5334.84	4	26
2017	8/1/2017	79601.5	8	1
2017	11/15/2016	33209.08	11	15
2017	6/30/2017	34598.49	6	30
2017	10/18/2016	16944.34	10	18
2017	11/17/2016	7366.87	11	17



Sum of GIVING	Column Labels	
Row Labels	2017	2018
1	$282K	$392K
2	$394K	$1090K
3	$723K	$514K
4	$506K	$840K
5	$951K	$773K
Grand Total	$2855K	$3609K



Reporting Requirements

This is a super cool report with a bunch of text. Many astute observations were made, and they are supported by colorful graphics such as the below chart. However, this document is not quite what was requested.

As you can see, 2016 does not appear here, which turns out to be an oversight. When 2016 is requested, since this was created as a manual process that began with manual data aggregation in Excel, followed by creation of a chart, which was then pasted into this document and manually re-colored and formatted, all that work will have to be repeated when 2016 data is added in the initial step.

A better workflow would have been to use R Markdown instead. In the words of the authors of development environment RStudio, "R Markdown documents are fully reproducible. Use a productive

Manual data pull → Manual aggregation → Manual formatting and content

Need to include more data

"Here we go again…"

The case for reproducibility

Reproducibility can help:

- Roll back and compare previous versions
- Effectively build upon past work
- Create unambiguous and automated procedures
- Save time, sanity, hairline

What is reproducibility?

“An analysis that can be passed from one person to another and, using the same data, generate the same results in an unambiguous manner.”
—Bray, Çetinkaya-Rundel, and Stangl

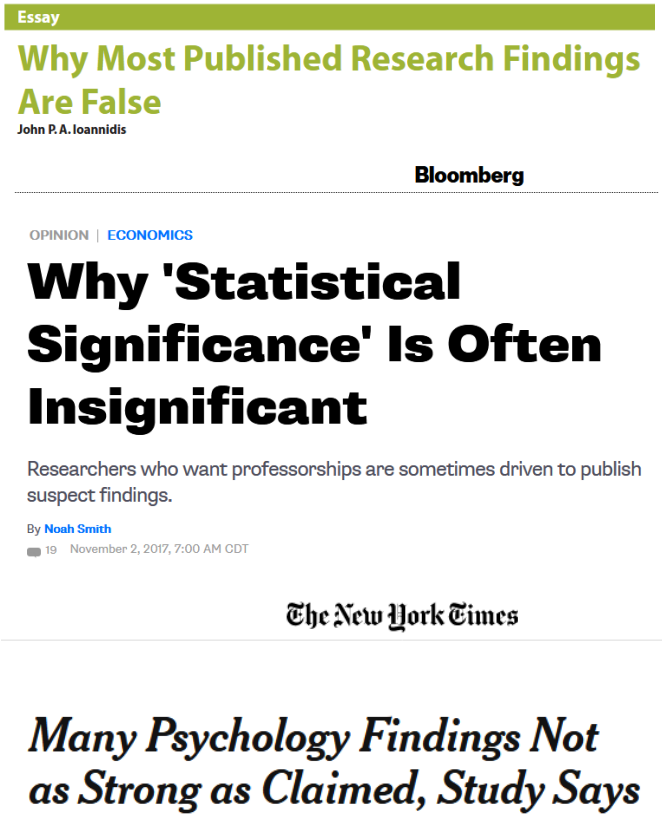
“Same data + Same script = Same results”
—Daniel Marcelino

What is reproducibility?

https://deevybee.blogspot.com/2018/02/improving-reproducibility-future-is.html

Whose idea was reproducibility?

Replication crisis



Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Bloomberg

OPINION | ECONOMICS

Why 'Statistical Significance' Is Often Insignificant

Researchers who want professorships are sometimes driven to publish suspect findings.

By Noah Smith

19 November 2, 2017, 7:00 AM CDT

The New York Times

Many Psychology Findings Not as Strong as Claimed, Study Says

General Article

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]

[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

Archive > Volume 497 > Issue 7447 > News > Article

NATURE | NEWS

Disputed results a fresh blow for social psychology

Failure to replicate intelligence-priming effects ignites row in research community.

Alison Abbott

30 April 2013 | Clarified: 17 May 2013

By Daniel Engber

TWEET SHARE

Daryl Bem Proved ESP Is Real

Which means science is broken.

MAY 17, 2017 • COVER STORY

Whose idea was reproducibility?

Replication crisis – what about ESP?!

- Psychologist Daryl Bem used large sample sizes and accepted experimental/statistical methods to demonstrate ESP exists
- Is ESP real or do the accepted methods have issues?

Lisa Larson-Walker for Slate magazine

Whose idea was reproducibility?

Replication crisis

- The results of many published scientific studies do not hold up to further scrutiny
- Methodology, statistical, and analysis issues
- Reproducibility can help: it requires a detailed methodology and provides evidence that correct results were produced

Whose idea was reproducibility?

“But we’re trying to fundraise, not publish research!”

Whose idea was reproducibility?

Collaboration

- Reproducibility makes it easier to catch mistakes
- Reproducibility enables others to catch mistakes
- Reproducibility is a teaching tool
- Reproducibility spreads the current state of the art

Whose idea was reproducibility?

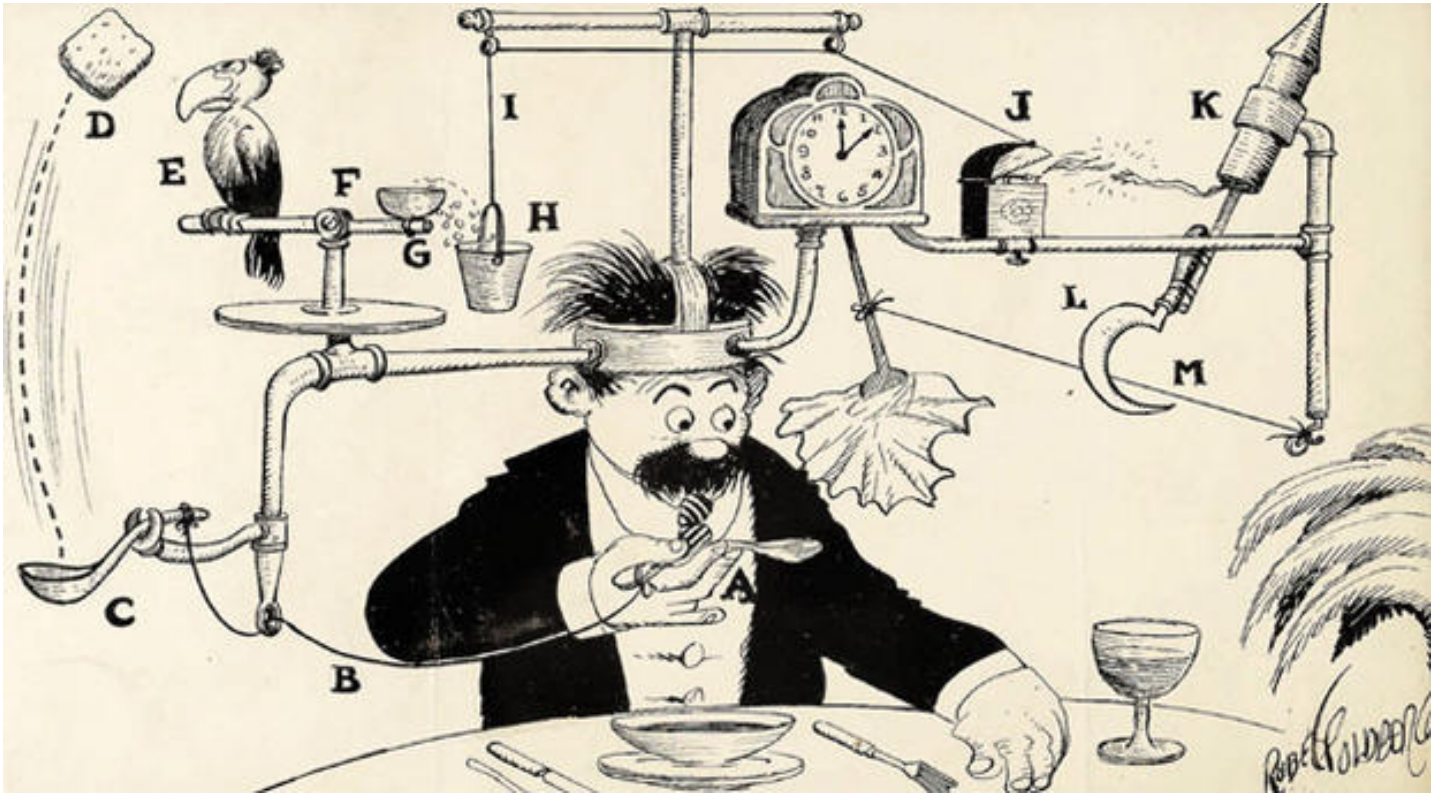
Collaboration

Recent prospect-dmm thread:

“I wanted to get some thoughts about how often others update thresholds/indicators of donors/prospects based on their models.... I would love insight on how y’all approach this part of the modeling workflow.”

Would you rather have thoughts or an example?

What does reproducibility require?



“Self-operating napkin” (Rube Goldberg, 1931)

What does reproducibility require?

The right mindset

- Willingness to plan ahead
- Go slow to go fast
- Don’t wait to document

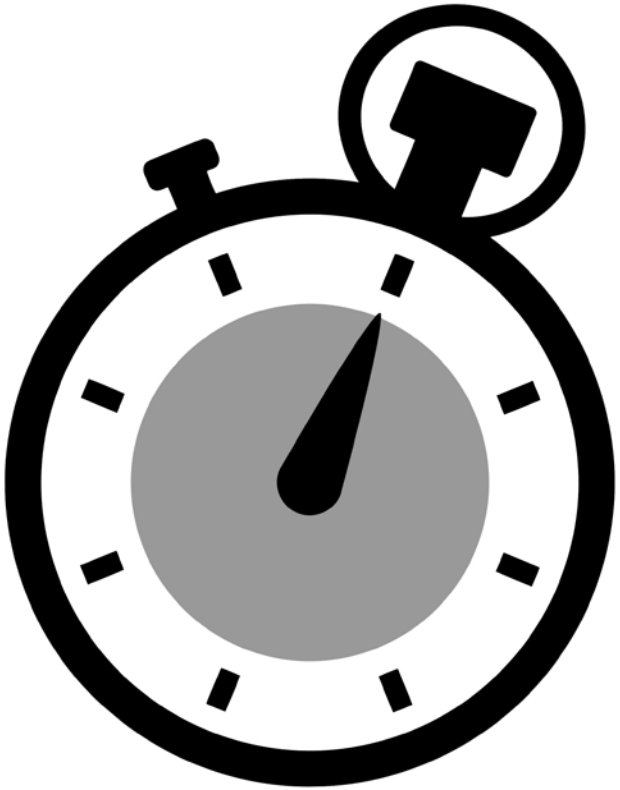
What does reproducibility require?

The right tools

- A way to save earlier revisions
- A way to automate the process

Quick pause

Questions? Comments? Funny stories? Other stories?



GitHub

Key facts

- Cloud hosting solution for Git
- Free version control platform
- Easy to save and compare changes from any two points in history
- Mascot = Octocat

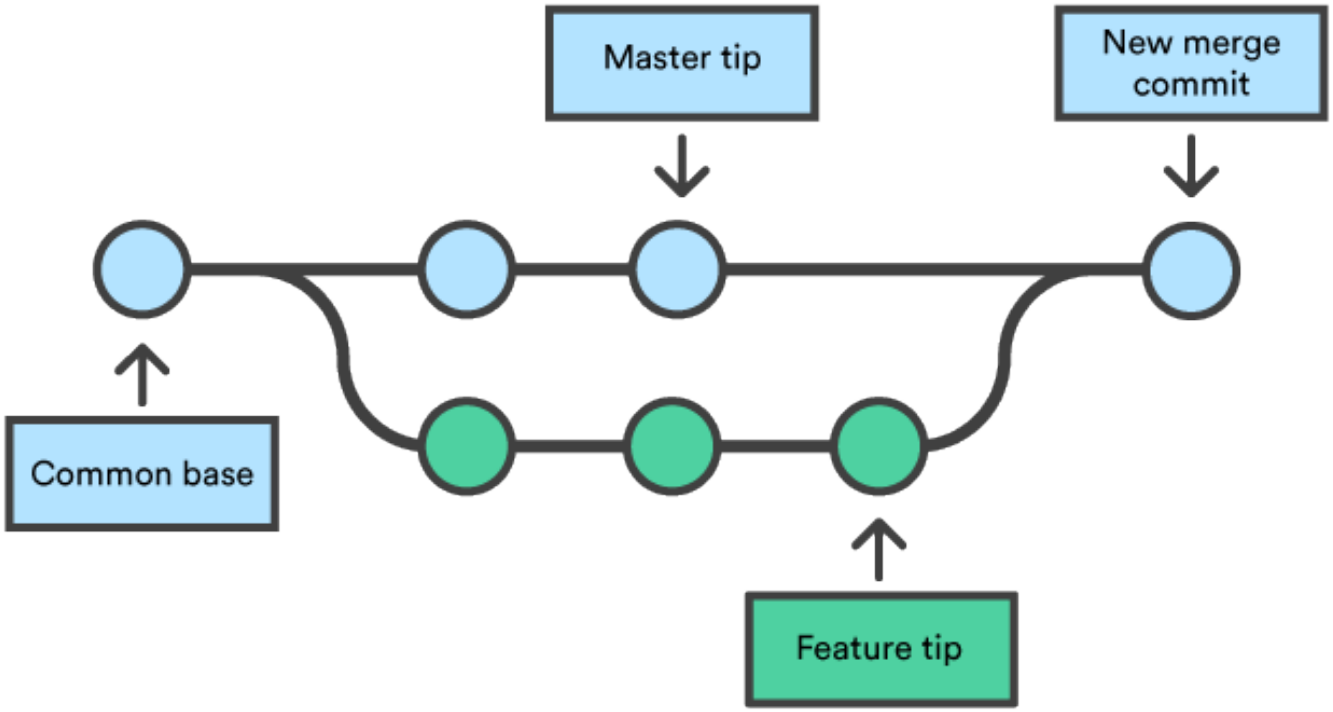
GitHub

Key concepts

- Repository = project
- Branch = copy
- Merge = reconcile revisions

GitHub

Key concepts



R Markdown

Key facts

- Built into the R Studio IDE
- Formatted text, R code, plots, tables, etc. can all exist in the same document
- Supports literate programming and self-documenting code
- Create HTML, pdf, and even Microsoft Office files (with plug-ins)

R Markdown

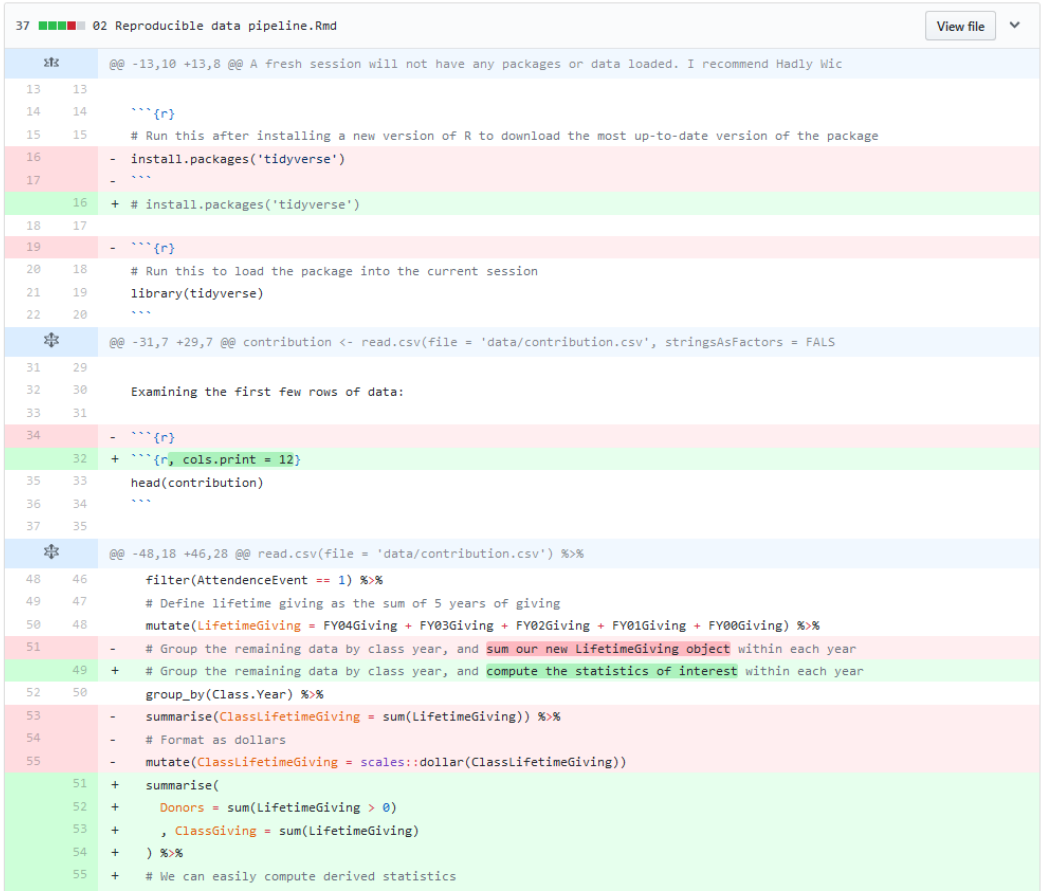
Coding rules of thumb

- Comment everything
- Use functions
- Parameters must be easy to find
- Use descriptive naming conventions
- Format consistently (including whitespace)

"There is no 'after' in which to write documentation."
—*Overheard on prospect-dmm*

Demo time!

https://github.com/phively/drive19



37 02 Reproducible data pipeline.Rmd — View file

```
@@ -13,10 +13,8 @@ A fresh session will not have any packages or data loaded. I recommend Hadly Wic
 
 ```{r}
 # Run this after installing a new version of R to download the most up-to-date version of the package
-install.packages('tidyverse')
-```
+# install.packages('tidyverse')
 
-```{r}
 # Run this to load the package into the current session
 library(tidyverse)
 ```
@@ -29,7 +27,7 @@ contribution <- read.csv(file = 'data/contribution.csv', stringsAsFactors = FALS

 Examining the first few rows of data:

-```{r}
+```{r, cols.print = 12}
 head(contribution)
 ```

@@ -46,8 +44,12 @@ read.csv(file = 'data/contribution.csv') %>%
   filter(AttendenceEvent == 1) %>%
   # Define lifetime giving as the sum of 5 years of giving
   mutate(LifetimeGiving = FY04Giving + FY03Giving + FY02Giving + FY01Giving + FY00Giving) %>%
- # Group the remaining data by class year, and sum our new LifetimeGiving object within each year
+ # Group the remaining data by class year, and compute the statistics of interest within each year
   group_by(Class.Year) %>%
- summarise(ClassLifetimeGiving = sum(LifetimeGiving)) %>%
- # Format as dollars
- mutate(ClassLifetimeGiving = scales::dollar(ClassLifetimeGiving))
+ summarise(
+   Donors = sum(LifetimeGiving > 0)
+   , ClassGiving = sum(LifetimeGiving)
+) %>%
+ # We can easily compute derived statistics
```
```

Key takeaways

- Automate everything you can
- Get help automating the things you can’t
- Go slow to go fast
- Don’t wait to document
- Share your work – let’s grow together!

References

Reproducibility in Science
https://ropensci.github.io/reproducibility-guide/sections/introduction/

Disputed results a fresh blow for social psychology (Alison Abbott)
https://www.nature.com/news/disputed-results-a-fresh-blow-for-social-psychology-1.12902

Five Concrete Reasons Your Students Should Be Learning to Analyze Data in the Reproducible Paradigm (Bray, Çetinkaya-Rundel, and Stangl)
http://chance.amstat.org/2014/09/reproducible-paradigm/

Many Psychology Findings Not as Strong as Claimed, Study Says (Benedict Carey)
https://www.nytimes.com/2015/08/28/science/many-social-science-findings-not-as-strong-as-claimed-study-says.html

Reproducible Research (John Cook)
https://reproducibleresearch.net

Daryl Bem Proved ESP Is Real (Daniel Engber)
https://slate.com/health-and-science/2017/06/daryl-bem-proved-esp-is-real-showed-science-is-broken.html

References

Why most published research findings are false (John Ioannidis)
http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124

What is reproducible research? (Daniel Marcelino)
https://www.r-bloggers.com/what-is-reproducible-research/

The Real Reason Reproducible Research is Important (Roger Peng)
https://simplystatistics.org/2014/06/06/the-real-reason-reproducible-research-is-important/

What is reproducible research? (Henry Rogalin)
https://bitesizebio.com/37187/reproducible-research/

False-Positive Psychology (Simmons, Nelson, and Simonsohn)
http://journals.sagepub.com/doi/abs/10.1177/0956797611417632

Why 'statistical significance' is often insignificant (Noah Smith)
https://www.bloomberg.com/view/articles/2017-11-02/why-statistical-significance-is-often-insignificant