Human-behavioral challenges in reproducibility & replicability

@jtleek

jtleek.com/talks

@jtleek

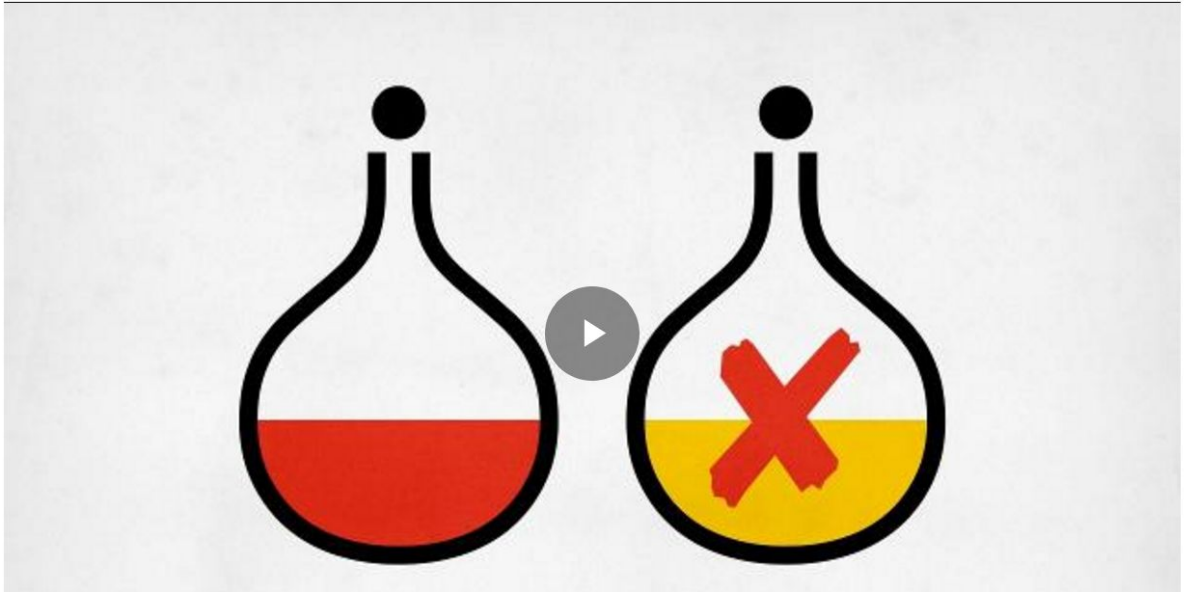
FIXING SCIENCE

Most science research findings are false. Here's how we can change that

POLICY & ETHICS

Is There a Reproducibility Crisis in Science?

By Nature Video on May 28, 2016



NATURE | NEWS FEATURE

1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.

Monya Baker

25 May 2016 | Corrected: 28 July 2016

Google

most science is wrong

All Images Videos News Shopping More Settings Tools

About 176,000,000 results (0.43 seconds)

Most Scientific Findings Are Wrong or Useless - Reason.com

reason.com/archives/2016/08/26/most-scientific-results-are-wrong-or-use

Aug 26, 2016 - ScientistYanlevDreamstime Yanlev/Dreamstime"**Science**, the pride of modernity, our one source of objective knowledge, is in deep trouble.

Sign in

PLOS Medicine: Why Most Published Research Findings Are False

journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124

by JPA Ioannidis - 2005 - Cited by 4846 - Related articles

Aug 30, 2005 - Moreover, for many current **scientific** fields, claimed research findings ... Citation: Ioannidis JPA (2005) Why **Most** Published Research Findings Are False. ..... what might have gone **wrong** with their data, analyses, and results.

Is Most Published Research Wrong? - YouTube

https://www.youtube.com/watch?v=42QuXLucH3Q

Aug 11, 2016 - Uploaded by Veritasium

Why **Most** Published Research Findings Are False: ..... The problem with the approach to **science** is that ...

Believe It Or Not, Most Published Research Findings Are Probably ...

bigthink.com/.../believe-it-or-not-most-published-research-findings-are-probably-fals...

Ten years ago, a researcher claimed **most** published research findings are false; ... of the Internet has worked wonders for the public's access to **science**, but this ... the case, experiments are underpowered,

176,000,000!

Replication crisis

From Wikipedia, the free encyclopedia

The **replication crisis** (or **replicability crisis**) refers to a methodological crisis in science in which scientists have found that the results of many scientific studies are difficult or impossible to replicate on subsequent investigation, either by independent researchers or by the original researchers themselves.[1] While the crisis has long-standing roots, the phrase was coined in the early 2010s as part of a growing awareness of the problem.

Since the reproducibility of experiments is an essential part of the scientific method, the inability to replicate the studies of others has potentially grave consequences for many fields of science in which significant theories are grounded on unreproduceable experimental work.

The replication crisis has been particularly widely discussed in the field of psychology (and in particular, social psychology) and in medicine, where a number of efforts have been made to re-investigate classic results, and to attempt to determine both the validity of the results, and, if invalid, the reasons for the failure of replication.[2][3]

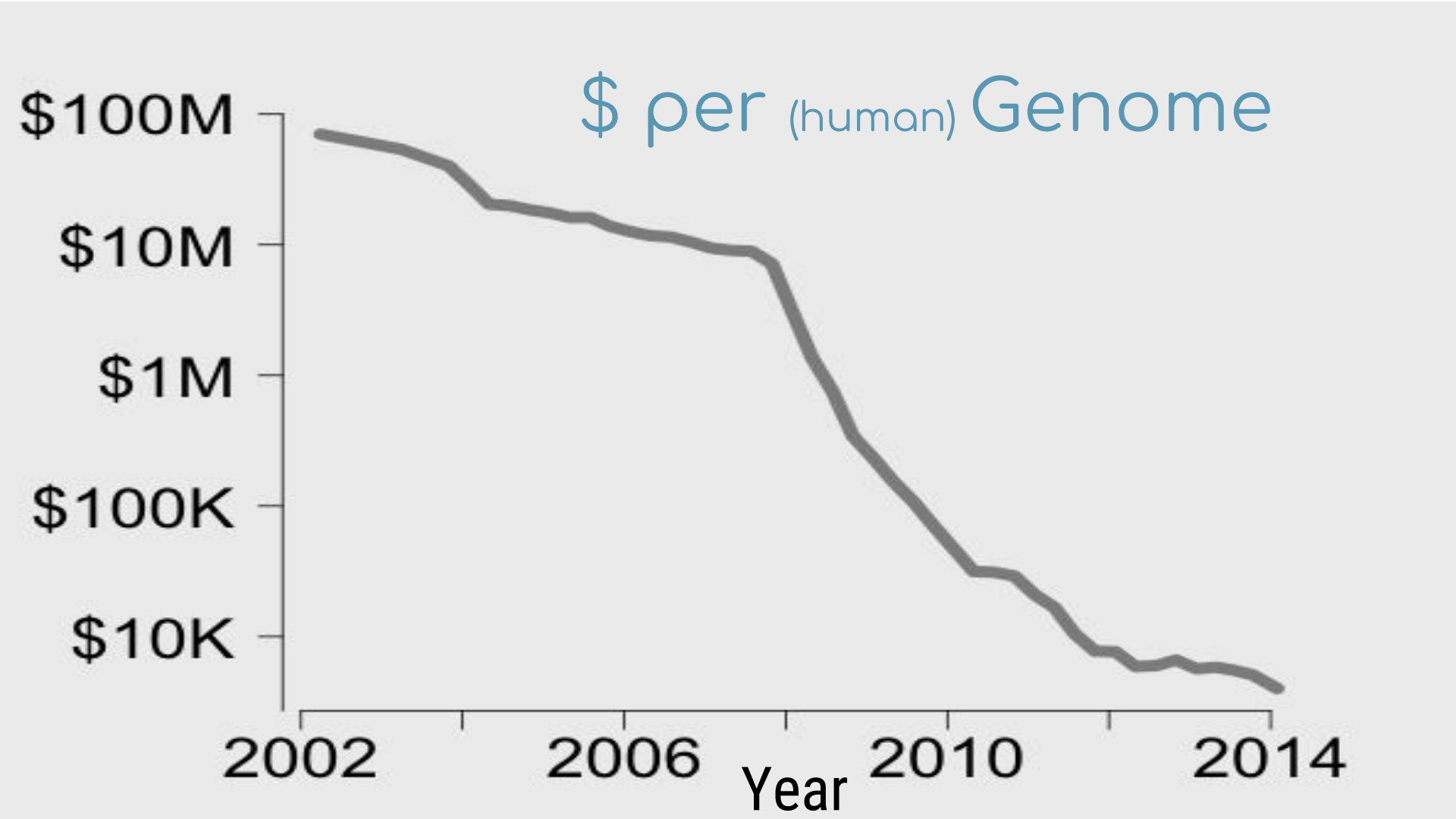
Contents [hide]

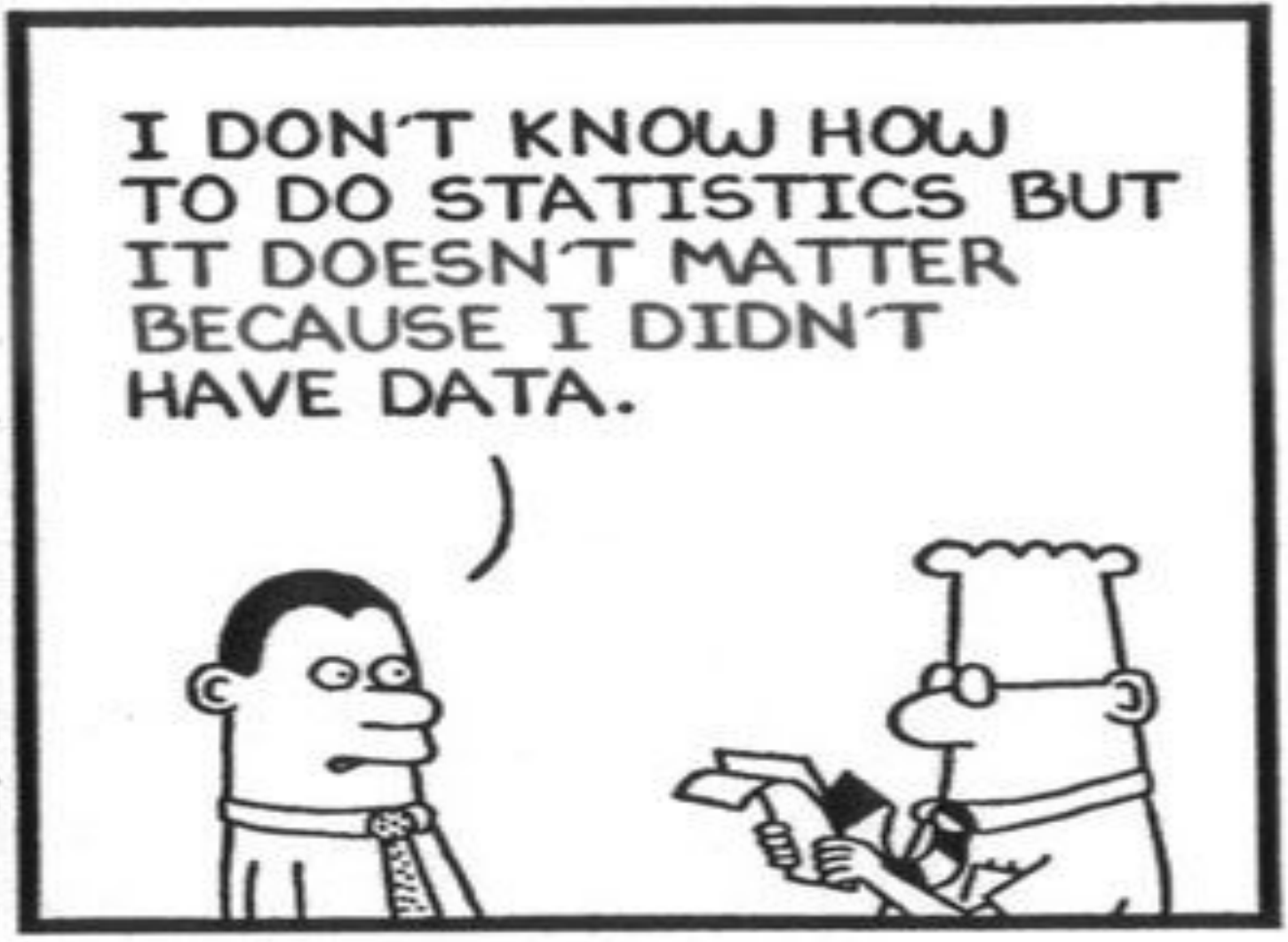
A hypothesis


I HAVE AN IDEA.
A GOOD ONE?
LET`S NOT GET AHEAD OF OURSELVES.
poorlydrawnlines.com

N = SAMPLE SIZE

$$N = \frac{(€ \text{ YOU HAVE})}{(€ \text{ PER SAMPLE})}$$


$ per (human) Genome
$100M
$10M
$1M
$100K
$10K
2002
2006
2010
2014
Year


I DON'T KNOW HOW TO DO STATISTICS BUT IT DOESN'T MATTER BECAUSE I DIDN'T HAVE DATA.

Medical school entrance requirements (U.S.)

One year of biology

One year of physics

One year of English

Two years of chemistry

simplystats

- Home
- About
- Archive
- Conferences
- Courses
- Interviews
- Contributing
- Twitter
- GitHub

Built with blogdown and Hugo. Theme Blackburn.

The vast majority of statistical analysis is not performed by statisticians

Jeff Leek 2013/06/14

Whether you know it or not, everything you do produces data - from the websites you read to the rate at which your heart beats. Until pretty recently, most of the data you produced wasn't collected, it floated off unmeasured. The only data that were collected were painstakingly gathered by scientists one number at a time in small experiments with a few people. This laborious process meant that data were expensive and time-consuming to collect. Yet many of the most amazing scientific discoveries over the last two centuries were squeezed from just a few data points. But over the last two decades, the unit price of data has dramatically dropped. New technologies touching every aspect of our lives from our money, to our health, to our social interactions have made data collection cheap and easy (see e.g. Camp Williams).

To give you an idea of how steep the drop in the price of data has been, in 1967 Stanley Milgram did an experiment to determine the number of degrees of separation between two people in the U.S. In his experiment he sent 296 letters to people in Omaha, Nebraska and Wichita, Kansas. The goal was to get the letters to a specific person in Boston, Massachusetts. The trick was people had to send the letters to someone they knew, and they then sent it to someone they knew and so on. At the end of the experiment, only 64 letters made it to the individual in Boston. On average, the letters had gone through 6 people to get there. This is where the idea of "6-degrees of Kevin Bacon" comes from. Based on 64 data points. A 2007 study updated that number to "7

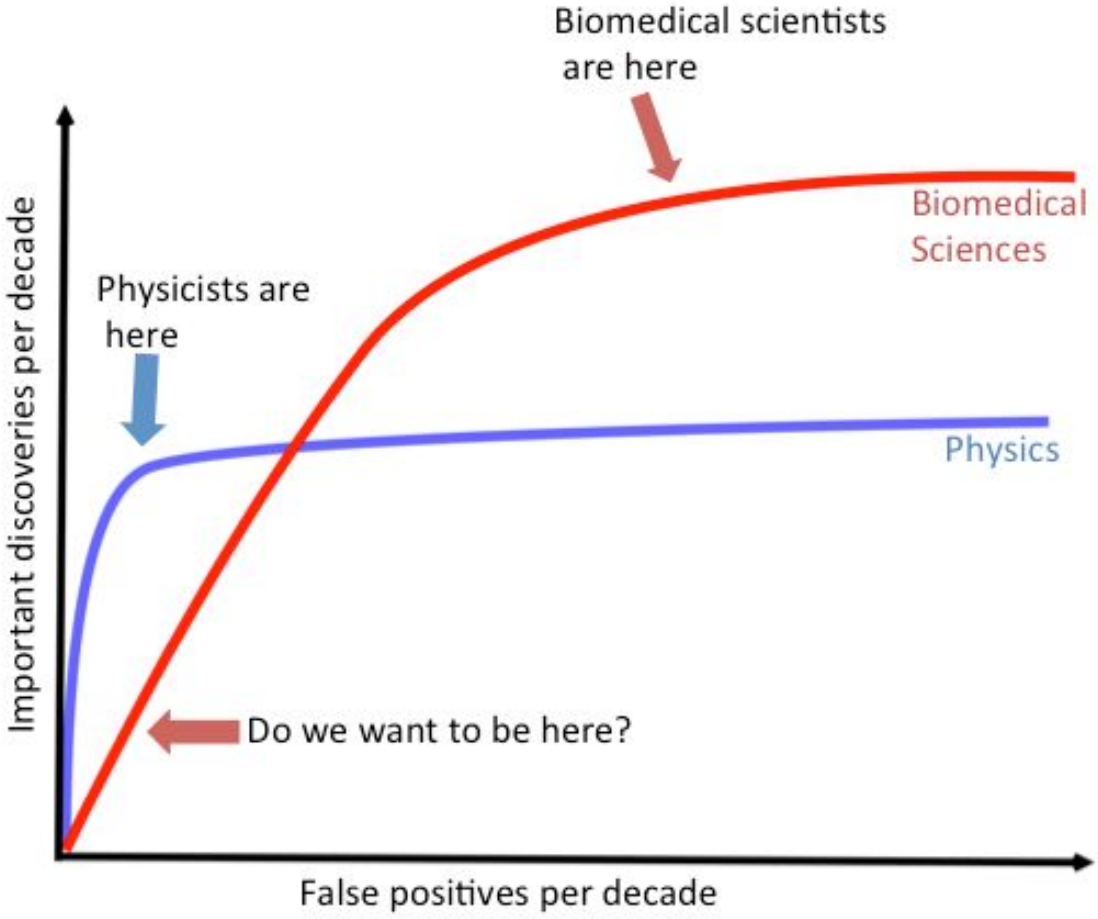
https://simplystatistics.org/2013/06/14/the-vast-majority-of-statistical-analysis-is-not-performed-by-statisticians/

3

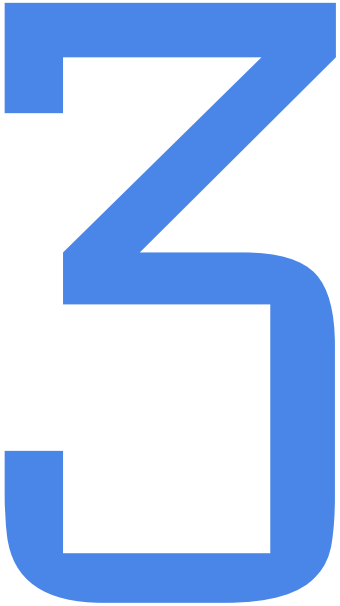
We need to say what we mean

The tools already exist

Humans are the problem



https://simplystatistics.org/2013/08/01/the-roc-curves-of-science/



We need to say what we mean

The tools already exist

Humans are the problem


Population


Question


Hypothesis


Experimental Design


Experimenter

Data


Analysis Plan


Analyst


Code


Estimate


Claim

Patil, Peng and Leek biorXiv 2016

PERSPECTIVE | SCIENTIFIC INTEGRITY

What does research reproducibility mean?

Steven N. Goodman*, Daniele Fanelli and John P. A. Ioannidis

+ See all authors and affiliations

Science Translational Medicine 01 Jun 2016:
Vol. 8, Issue 341, pp. 341ps12
DOI: 10.1126/scitranslmed.aaf5027

Abstract

The language and conceptual framework of “research reproducibility” are nonstandard and unsettled across the sciences. In this Perspective, we review an array of explicit and implicit definitions of reproducibility and related terminology, and discuss how to avoid potential misunderstandings when these terms are used as a surrogate for “truth.”

Science Translational Medicine

Vol 8, Issue 341
01 June 2016

Table of Contents

SHARE

PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

+ See all authors and affiliations

Science 02 Dec 2011:
Vol. 334, Issue 6060, pp. 1226-1227
DOI: 10.1126/science.1213847

Abstract

Computational science has led to exciting new developments, but the nature of the work has exposed limitations in our ability to evaluate published findings. Reproducibility has the potential to serve as a minimum standard for judging scientific claims when full independent replication of a study is not possible.

. . .

The rise of computational science has led to exciting and fast-moving developments in many scientific areas. New technologies, increased computing power, and methodological advances have dramatically improved our ability to collect complex high-dimensional data (*1*, *2*). Large data sets have led to scientists doing more

Science

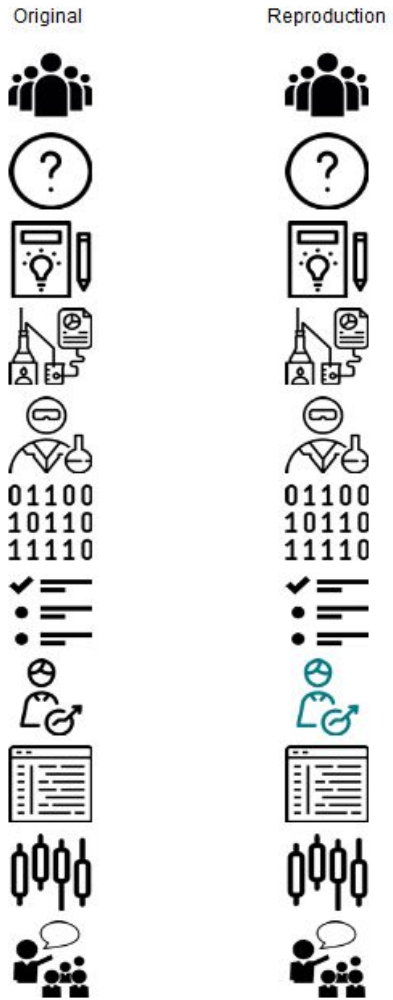
Vol 334, Issue 6060
02 December 2011

Table of Contents
Print Table of Contents
Advertising (PDF)
Classified (PDF)
Masthead (PDF)

http://science.sciencemag.org/content/334/6060/1226.full

`...I’m actually not sure we all agree on meaning of “replication” / “replicability” / “replicate”, nor is this term necessarily used consistently in the literature’

Reproduce



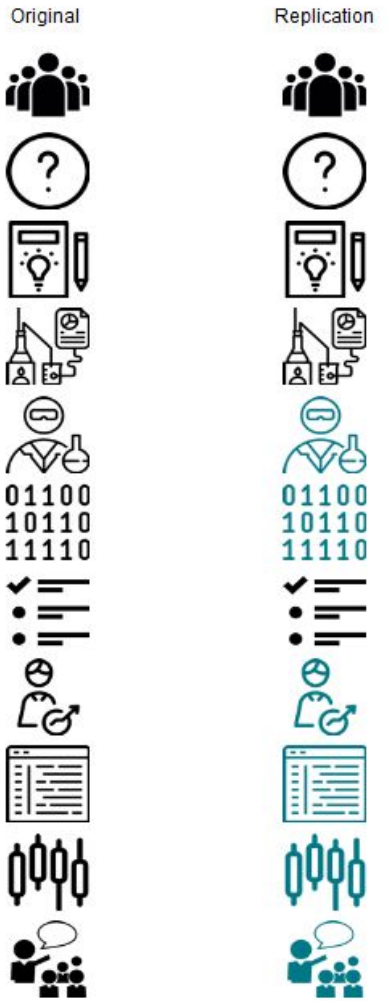
Original

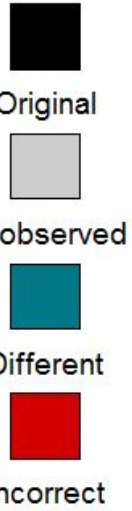
Unobserved

Different

Incorrect

Patil, Peng and Leek biorXiv 2016

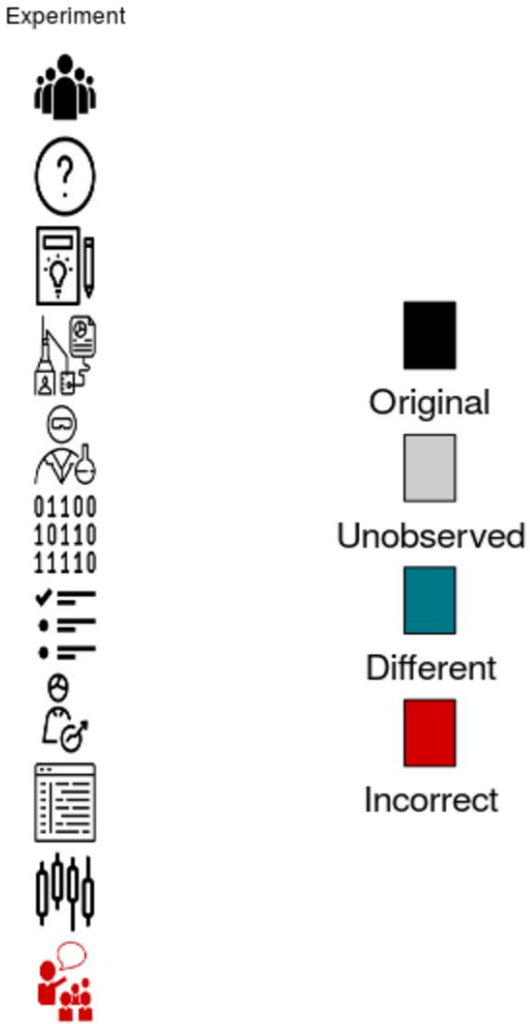




Replicate

Patil, Peng and Leek biorXiv 2016

False discovery



Patil, Peng and Leek biorXiv 2016

Science AAAS

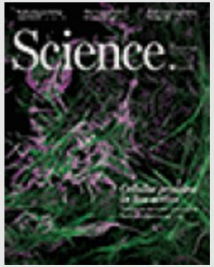
RESEARCH ARTICLE

Estimating the ~~reproducibility~~ replicability of psychological science

Open Science Collaboration[*,†]

+ See all authors and affiliations

Science 28 Aug 2015:
Vol. 349, Issue 6251, aac4716
DOI: 10.1126/science.aac4716



Science

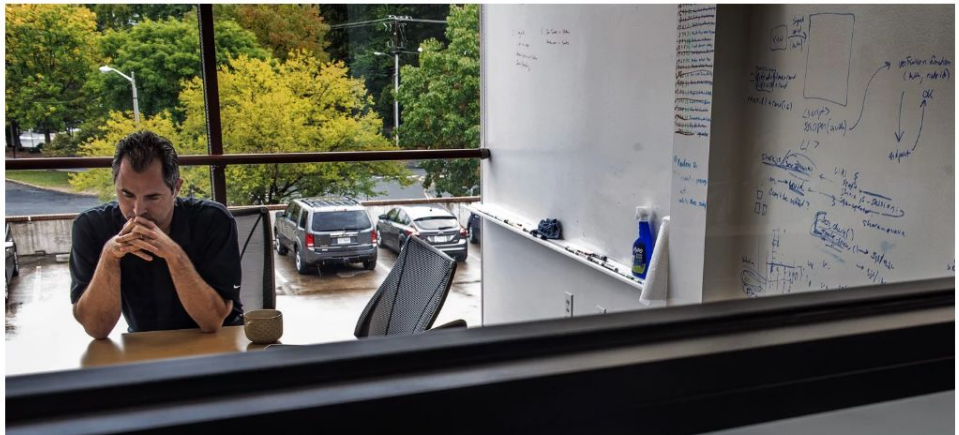
Vol 349, Issue 6251
28 August 2015

Table of Contents
Print Table of Contents
Advertising (PDF)
Classified (PDF)
Masthead (PDF)

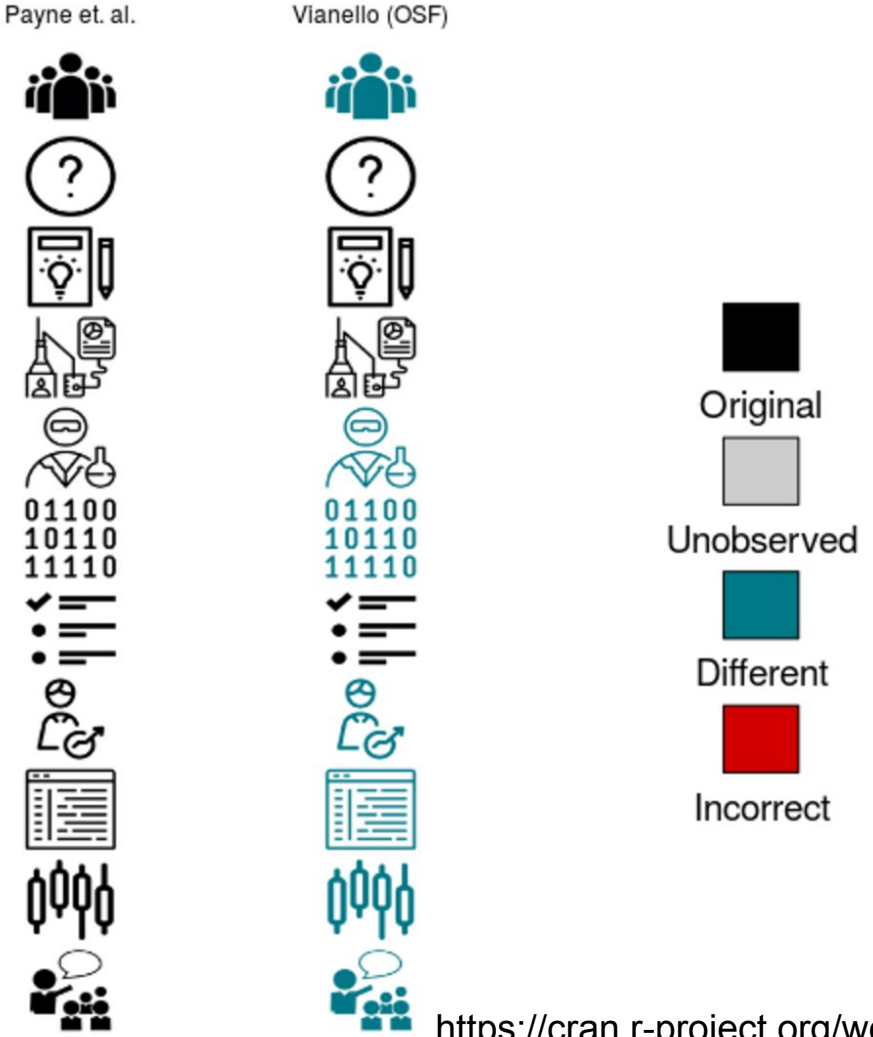
Speaking of Science

Many scientific studies can't be replicated. That's a problem.

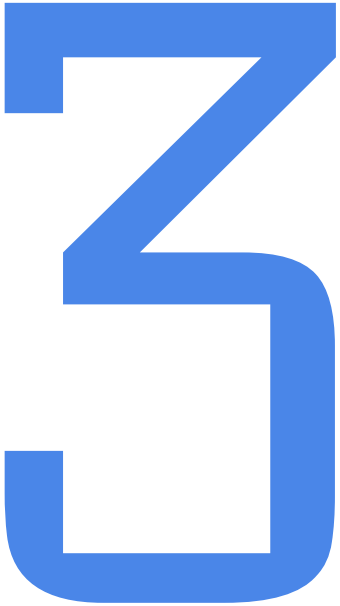
By Joel Achenbach August 27, 2015



Over the course of four years, 270 researchers attempted to ~~reproduce~~ replicate the results of 100 experiments that had been published in three prestigious psychology journals. It was awfully hard. They ultimately concluded that they'd succeeded just 39 times.



https://cran.r-project.org/web/packages/scifigure/index.html



We need to say what we mean

The tools already exist

Humans are the problem

Jeff L.

jtleek

Add a bio

Developer Program Member

Baltimore,MD

http://biostat.jhsph.edu/~jleek/

Overview | Repositories 81 | Stars 7 | Followers 4.5k | Following 6

Popular repositories

Customize your pinned repositories

datasharing

The Leek group guide to data sharing

★ 4k ⑂ 175k

dataanalysis

The lecture slides for Coursera's Data Analysis class

JavaScript ★ 636 ⑂ 631

rpackages

R package development - the Leek group way!

★ 337 ⑂ 247

genomicspapers

The Leek group guide to genomics papers

★ 245 ⑂ 114

reviews

Writing reviews of academic papers

★ 206 ⑂ 61

capitalIn21stCenturyinR

Piketty in R

HTML ★ 197 ⑂ 125

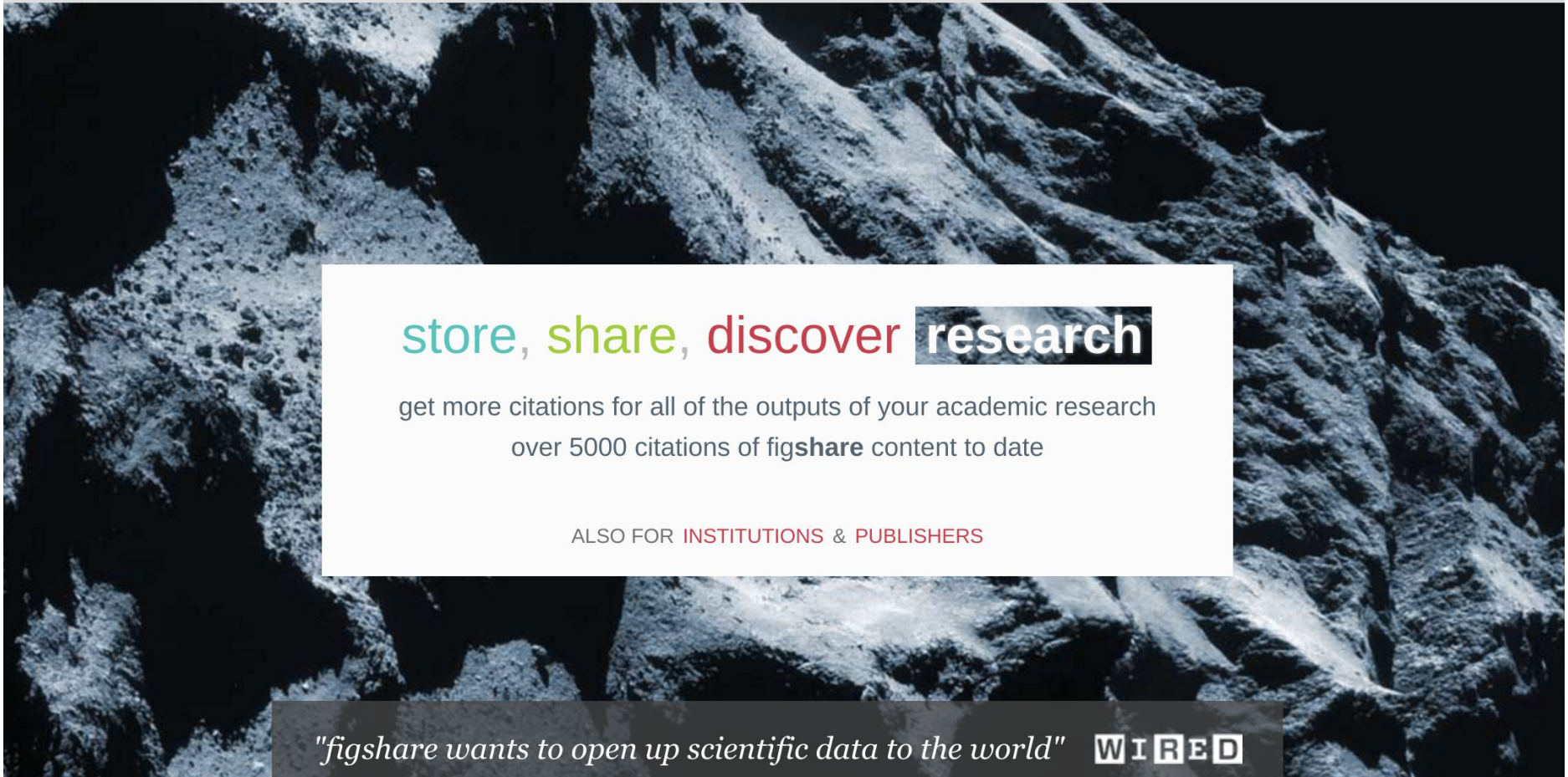
R Markdown

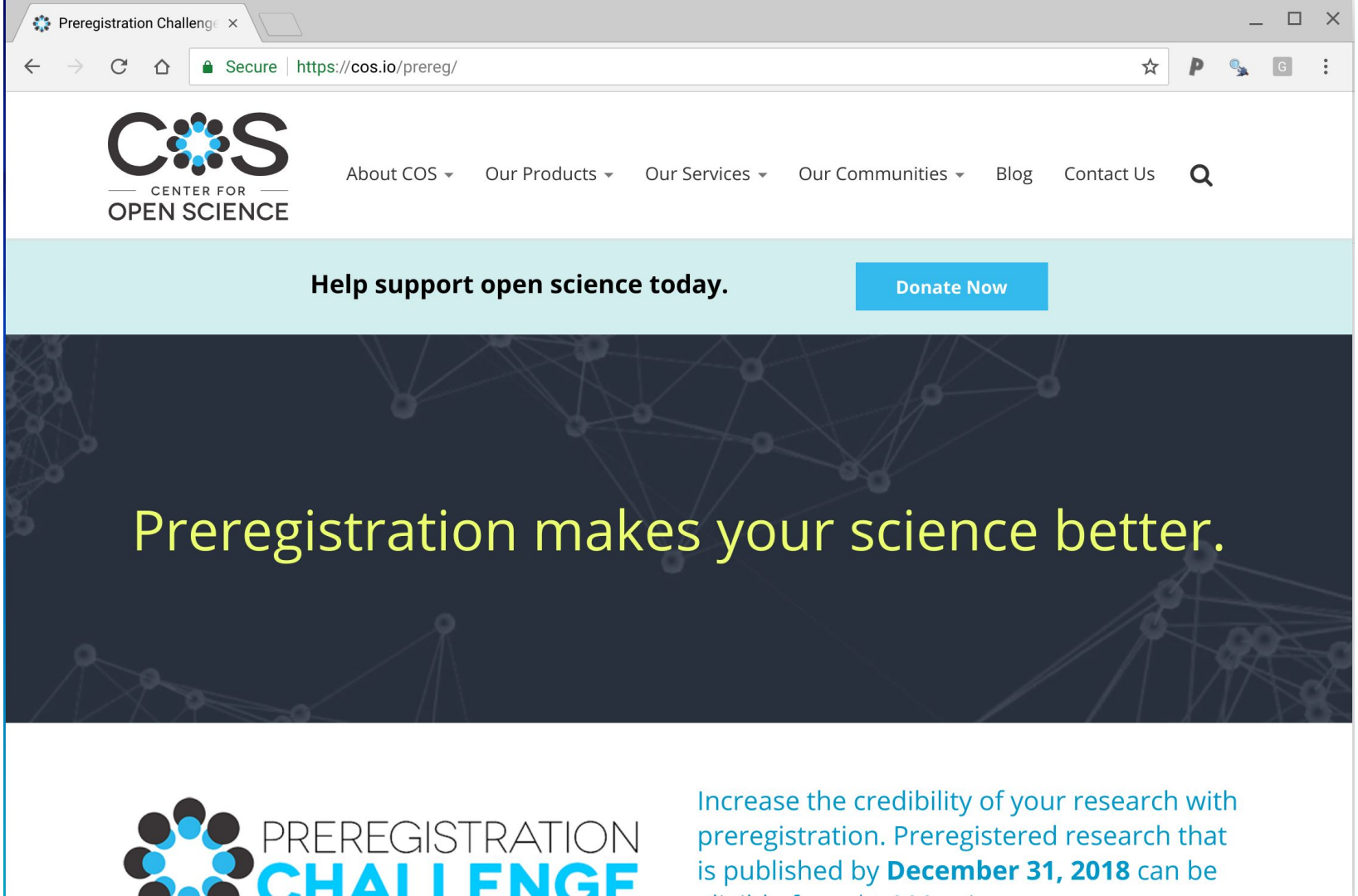
from RStudio

Get Started Gallery Formats Articles

Analyze. Share. Reproduce.

Your data tells a story. Tell it with R Markdown. Turn your analyses into high quality documents, reports, presentations and dashboards.


figshare
search on figshare
Browse
Upload
Sign up
Log in
store, share, discover research
get more citations for all of the outputs of your academic research
over 5000 citations of figshare content to date
ALSO FOR INSTITUTIONS & PUBLISHERS
"figshare wants to open up scientific data to the world" WIRED


Preregistration Challenge
Secure
https://cos.io/prereg/
CENTER FOR
OPEN SCIENCE
About COS
Our Products
Our Services
Our Communities
Blog
Contact Us
Help support open science today.
Donate Now
Preregistration makes your science better.
PREREGISTRATION
CHALLENGE
Increase the credibility of your research with preregistration. Preregistered research that is published by December 31, 2018 can be

About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 1560 software packages, and an active user community. Bioconductor is also available as an AMI (Amazon Machine Image) and a series of Docker images.

News

- BioC 2018 conference web site.
- **Core team job opportunities for scientific programmer / analyst and senior programmer / analyst!**
- Bioconductor 3.7 is available.
- Bioconductor F1000 Research Channel available.
- Orchestrating high-throughput genomic analysis with *Bioconductor* (abstract) and other recent literature.
- View recent course material.
- Use the support site to get help installing, learning and using Bioconductor.

Install »

Get started with *Bioconductor*

- Install *Bioconductor*
- Explore packages
- Get support
- Latest newsletter
- Follow us on twitter
- Install R

Learn »

Master *Bioconductor* tools

- Courses
- Support site
- Package vignettes
- Literature citations
- Common work flows
- FAQ
- Community resources
- Videos

Use »

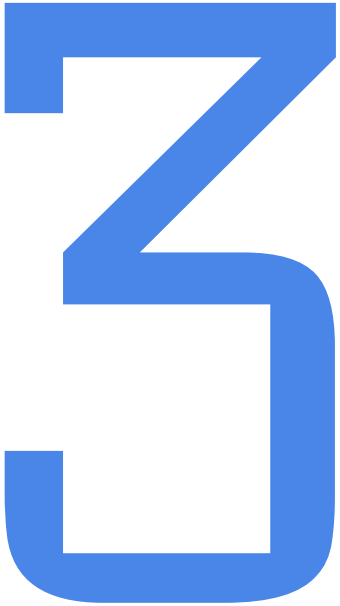
Create bioinformatic solutions with *Bioconductor*

- Software, Annotation, and Experiment packages
- Amazon Machine Image
- Latest release annoucement
- Support site

Develop »

Contribute to *Bioconductor*

- Developer resources
- Use Bioc 'devel'
- 'Devel' Software, Annotation and Experiment packages
- Package guidelines
- New package submission
- Git source control
- Build reports



We need to say what we mean

The tools already exist

Humans are the problem

reproducibility

JAMA. 2016 Mar 15;315(11):1141-8. doi: 10.1001/jama.2016.1952. Sign in

Evolution of Reporting P Values in the Biomedical Literature, 1990-2015.

Chavalarias D[1], Wallach JD[2], Li AH[3], Ioannidis JP[4].

Author information

Abstract

IMPORTANCE: The use and misuse of P values has generated extensive debates.

OBJECTIVE: To evaluate in large scale the P values reported in the abstracts and full text of biomedical research articles over the past 25 years and determine how frequently statistical information is presented in ways other than P values.

DESIGN: Automated text-mining analysis was performed to extract data on P values reported in 12,821,790 MEDLINE abstracts and in 843,884 abstracts and full-text articles in PubMed Central (PMC) from 1990 to 2015. Reporting of P values in 151 English-language core clinical journals and specific article types as classified by PubMed also was evaluated. A random sample of 1000 MEDLINE abstracts was manually assessed for reporting of P values and other types of statistical information; of those abstracts reporting empirical data, 100 articles were also assessed in full text.

MAIN OUTCOMES AND MEASURES: P values reported.

RESULTS: Text mining identified 4,572,043 P values in 1,608,736 MEDLINE abstracts and 3,438,299 P values in 385,393 PMC full-text articles. Reporting of P values in abstracts increased from 7.3% in 1990 to 15.6% in 2014. In 2014, P values were reported in 33.0% of abstracts from the 151 core clinical journals (n = 29,725 abstracts), 35.7% of meta-analyses (n = 5620), 38.9% of clinical trials (n = 4624), 54.8% of randomized controlled trials (n = 13,544), and 2.4% of reviews (n = 71,529). The distribution of reported P values in abstracts and in full text showed strong clustering at P values of .05 and of .001 or smaller. Over time, the "best" (most statistically significant) reported P values were modestly smaller and the "worst" (least statistically significant) reported P values became modestly less significant. Among the MEDLINE abstracts and PMC full-text articles with P

Hi John

I read with interest your recent paper in JAMA on p-values:

http://jama.jamanetwork.com/article.aspx?articleid=2503172#

But could not find the data or code. Would you mind letting me know where they are?

Thanks!

Dear Jeff,

I still have to publish the code (I managed it on a private git). I plan to do it early june since I am quite busy until then. I just want to properly explain how it works when I release it. I hope this won't be too long

“So if I have time I will make a website with an API to retrieve data on requests.”

As fo time,
I will t's
the M

Rega

David

Hi,

The dataset is now online on dataverse http://dx.doi.org/10.7910/DVN/6FMTT3

After import of the sql you should have
- 1,985,670 rows for the table `medline_full_txt_list`
- 12,436,631 rows for the table `medline_full_txt_pv`
- 16,116,061 rows for the table `medline_pt`
- 9,088,701 rows for the table `medline_pvalues`

Tell me if there is any issue. Source code will follow.

7 Files

Download

medline_full_txt_list.sql
Unknown - 228.3 MB - Apr 12, 2016 - 4 Downloads
MD5: 5d78f42859d4044660cf636675281384
List of all PMC papers processed
Data P-values

Download

medline_full_txt_pv.sql
Unknown - 1.0 GB - Apr 12, 2016 - 9 Downloads
MD5: 9e3ee983cf8bee7d75153124abfae6e2
sql table of all P-values extracted from PMC full text
Data P-values

Download

medline_full_txt_pv.tab
Tabular Data - 174.3 MB - Apr 12, 2016 - 14 Downloads
7 Variables, 3438298 Observations -
UNF:6:tRQMiS7wwbyq7H4scJ6DAQ==
CSV of all P-values extracted from PMC full text
Data P-values

Explore Download

medline_pt.sql
Unknown - 1.5 GB - Apr 12, 2016 - 1 Download
MD5: 68ebdedabc670c188d9b49629db931d3
sql of all P-values extracted from Medline abstracts
Data P-values

Download

```
> library(readr)
> dat = read_csv("~/data/medicine/medline_full_txt_pv.csv")
Parsed with column specification:
cols(
  `7669595` = col_integer(),
  `0370635` = col_character(),
  `=` = col_character(),
  `0.14` = col_double(),
  `1995` = col_integer(),
  plain = col_character(),
  `1` = col_integer()
)
|==================================================================| 100%  174 MB
=====================================                              |  64%  112 MB
> head(dat)
# A tibble: 6 x 7
  `7669595` `0370635`   `=` `0.14` `1995` plain   `1`
      <int>     <chr> <chr>  <dbl>  <int> <chr> <int>
1   7669596   0370635     =  0.001   1995 plain     1
2   8611396   0370635     <  0.010   1996 plain     1
3   8611396   0370635     <  0.010   1996 plain     1
4   8611396   0370635     <  0.010   1996 plain     1
5   8611397   0370635     <  0.010   1996 plain     1
6   8611398   0370635     <  0.010   1996 plain     1
> 
```

7 Files

Download

medline_full_txt_list.sql
Unknown - 228.3 MB - Apr 12, 2016 - 4 Downloads
MD5: 5d78f42859d4044660cf636675281384
List of all PMC papers processed
Data P-values

Download

medline_full_txt_pv.sql
Unknown - 1.0 GB - Apr 12, 2016 - 9 Downloads
MD5: 9e3ee983cf8bee7d75153124abfae6e2
sql table of all P-values extracted from PMC full text
Data P-values

Download

medline_full_txt_pv.tab
Tabular Data - 174.3 MB - Apr 12, 2016 - 14 Downloads
7 Variables, 3438298 Observations -
UNF:6:tRQMiS7wwbyq7H4scJ6DAQ==
CSV of all P-values extracted from PMC full text
Data P-values

Explore Download

medline_pt.sql
Unknown - 1.5 GB - Apr 12, 2016 - 1 Download
MD5: 68ebdedabc670c188d9b49629db931d3
sql of all P-values extracted from Medline abstracts
Data P-values

Download

P-values from Chavalarias et al. 2016 for the tidypvals package

Jeff Leek

26 July 2017

Contents

These p-values come from the paper: Evolution of Reporting P Values in the Biomedical Literature. The csv file for the p-values from medline did not have column names, so to ensure we had the right data we downloaded the MySQL dump from the Dataverse https://dataverse.harvard.edu/file.xhtml;jsessionid=94274f10cbdbecaaaf6da71ca209?fileId=2801917&version=RELEASED&version=.0 on on 2017-07-24. We re-loaded it into a MySQL database and that is where the code starts.

1 Set up

1.1 Load packages

How to share data for collaboration

Shannon E. Ellis
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
and
Jeffrey T. Leek *
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

September 1, 2017

replicability

Original Contribution FREE

March 8, 2006

Fine Particulate Air Pollution and Hospital Admission for Cardiovascular and Respiratory Diseases

Francesca Dominici, PhD; Roger D. Peng, PhD; Michelle L. Bell, PhD; et al

Article Information

“In the first stage, single lag and distributed lag overdispersed Poisson regression models were used for estimating county-specific RRs of hospital admissions associated with ambient levels of $PM_{2.5}$”

Source: Dominici, Peng et al. 2006 JAMA

log(E[Hospital Admissions | X]) =

PM2.5

+ log(# at risk) + factor(day of week) + ns(day of year, 8) + ns(temperature,6)

+ ns(dew point temperature, 3) + ns(lagged temperature, 6)

+ ns(lagged dew point temperature, 3)

Monographs
on Statistics and
Applied Probability 37

Generalized Linear Models

SECOND EDITION

P. McCullagh and
J.A. Nelder FRS

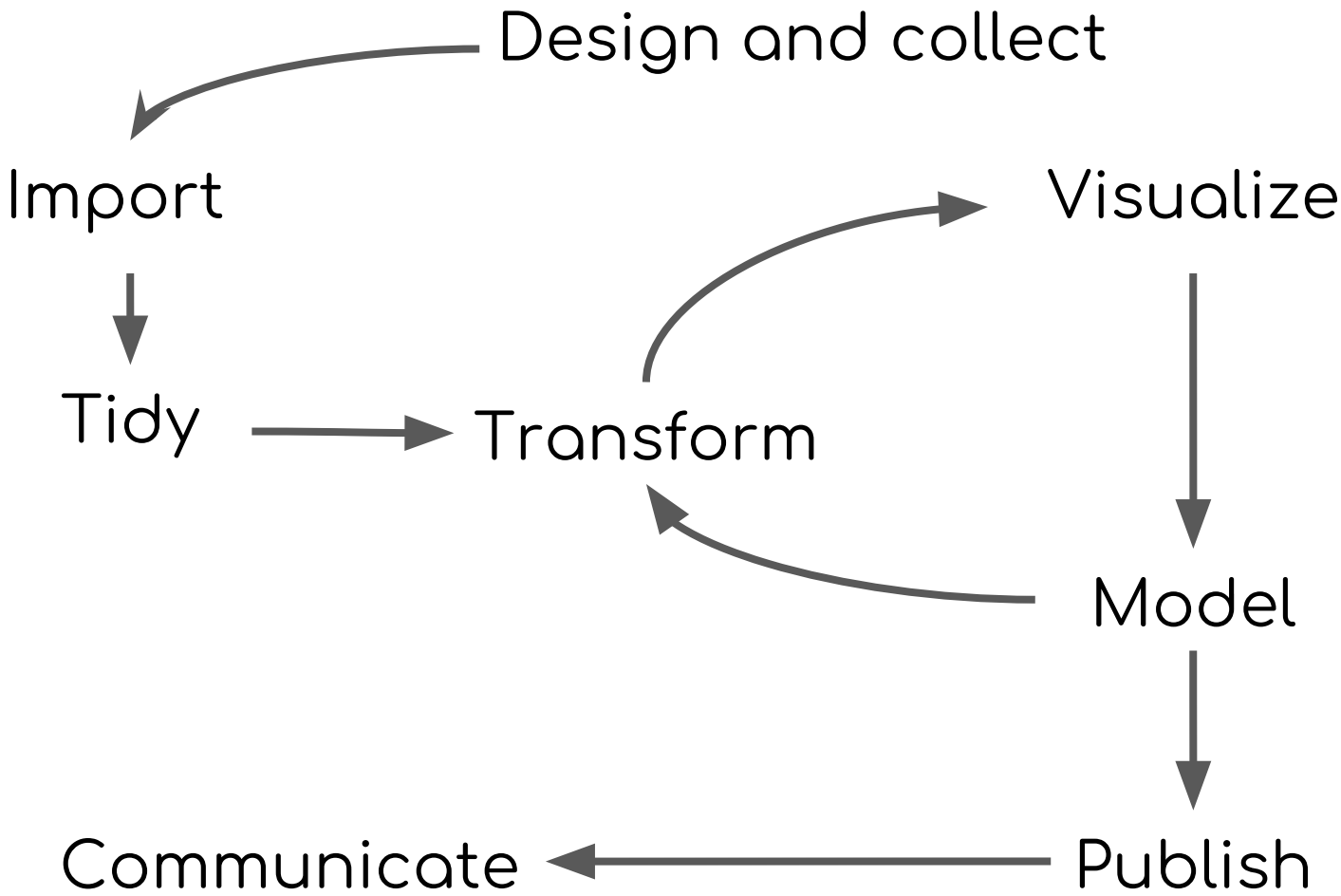
Springer-Science+Business
Media, B.V.

Eight outcomes were considered based on the ICD-9 codes for 5 cardiovascular outcomes (heart failure [428], heart rhythm disturbances [426-427], cerebrovascular events [430-438], ischemic heart disease [410-414, 429], peripheral vascular disease [440-448]), 2 respiratory outcomes (chronic obstructive pulmonary disease [COPD; 490-492], respiratory tract infections [464-466, 480-487]), and hospitalizations caused by injuries and other external causes (800-849). The county-wide daily hospitalization rates for each outcome for 1999-2002 appear in Table 1.

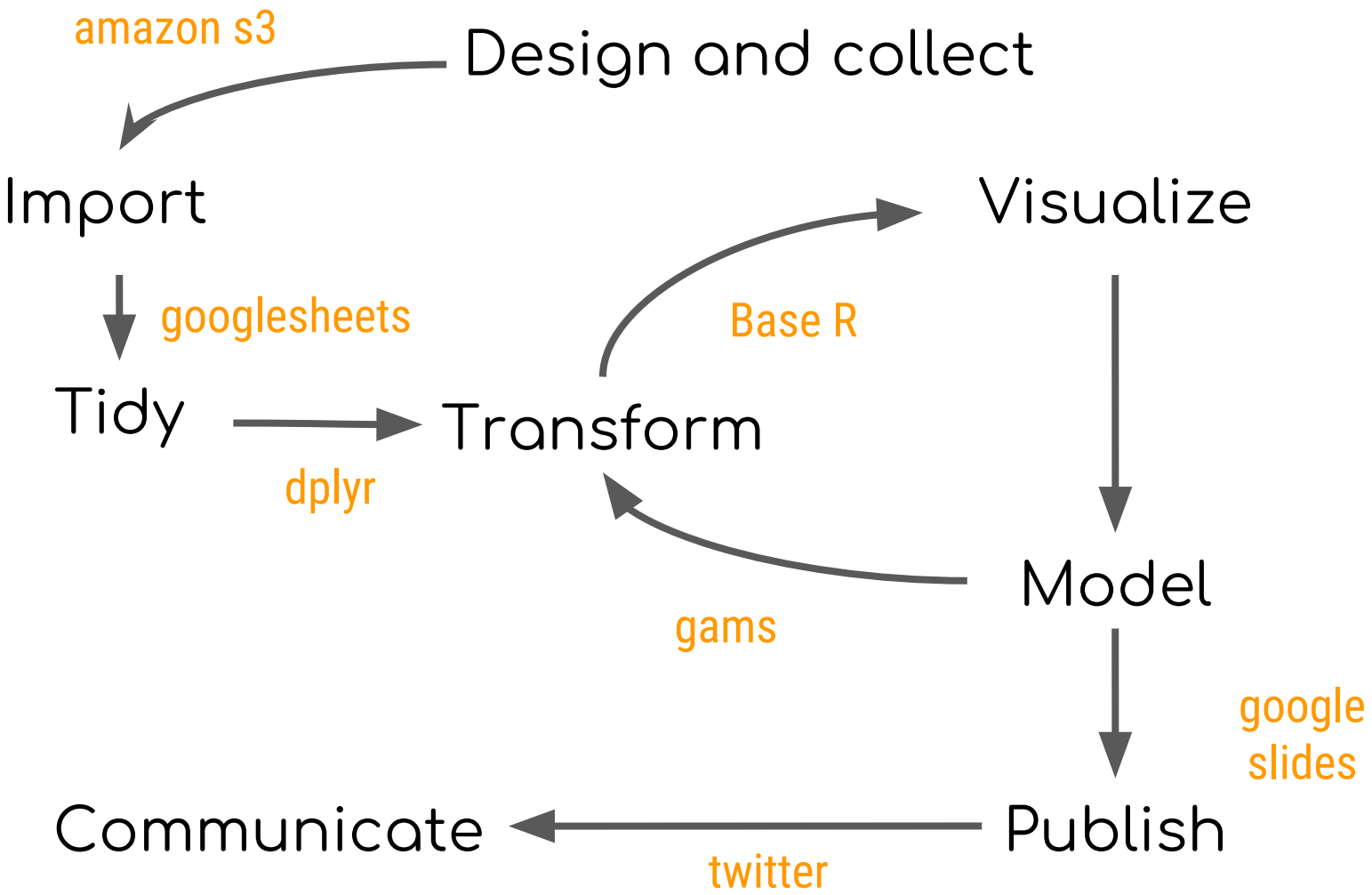
The study population includes 11.5 million Medicare enrollees residing an average of 5.9 miles from a PM2.5 monitor. The analysis was restricted to the 204 US counties with populations larger than 200 000. Of these 204 counties, 90 had daily PM2.5 data across the study period and the remaining counties had PM2.5 data collected once every 3 days for at least 1 full year. The locations of the 204 counties appear in Figure 1. The counties were clustered into 7 geographic regions by applying the K-means clustering algorithm to longitude and latitude for the counties.10,11

Eight outcomes were considered based on the ICD-9 codes for 5 cardiovascular outcomes (heart failure [428], heart rhythm disturbances [426-427], cerebrovascular events [430-438], ischemic heart disease [410-414, 429], peripheral vascular disease [440-448]), 2 respiratory outcomes (chronic obstructive pulmonary disease [COPD; 490-492], respiratory tract infections [464-466, 480-487]), and hospitalizations caused by injuries and other external causes (800-849). The county-wide daily hospitalization rates for each outcome for 1999-2002 appear in Table 1.

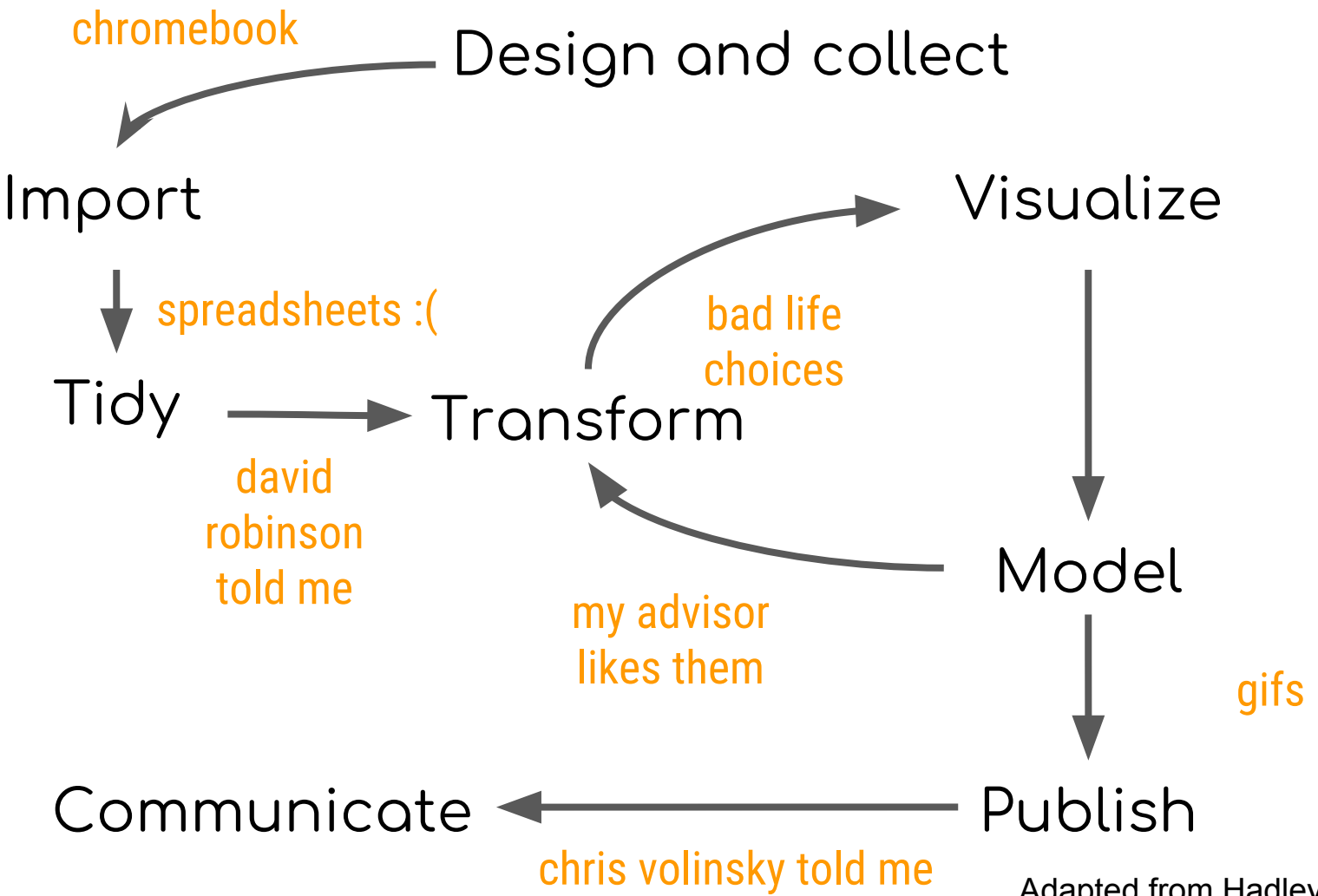
The study population includes 11.5 million Medicare enrollees residing an average of 5.9 miles from a PM2.5 monitor. The analysis was restricted to the 204 US counties with populations larger than 200 000. Of these 204 counties, 90 had daily PM2.5 data across the study period and the remaining counties had PM2.5 data collected once every 3 days for at least 1 full year. The locations of the 204 counties appear in Figure 1. The counties were clustered into 7 geographic regions by applying the K-means clustering algorithm to longitude and latitude for the counties.10,11



Adapted from Hadley Wickham



Adapted from Hadley Wickham


chromebook
Design and collect
Import
Visualize
spreadsheets :(
bad life
choices
Tidy
Transform
david
robinson
told me
Model
my advisor
likes them
gifs
Communicate
Publish
chris volinsky told me
Adapted from Hadley Wickham

Suppose we have a sample with 30 lawyers and 70 engineers.

What is the probability Dick is a lawyer?

Source: Tversky and Kahneman 1984

Dick is a 30 year old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues.

What is the probability Dick is a lawyer?

Source: Tversky and Kahneman 1984

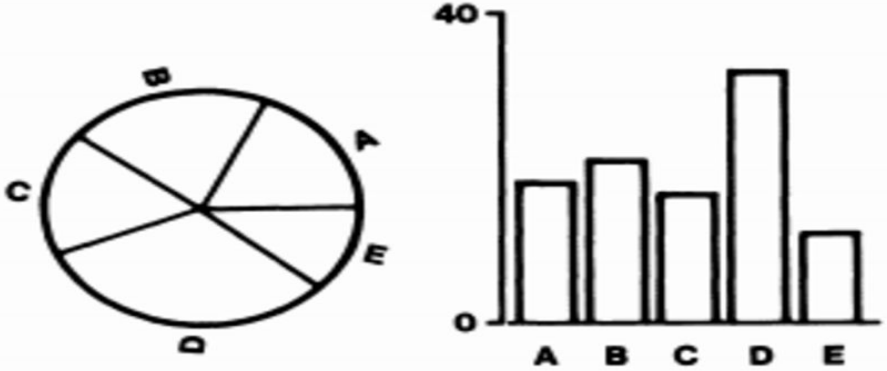


Figure 3. Graphs from position–angle experiment.

Source: Cleveland and McGill JASA 1984

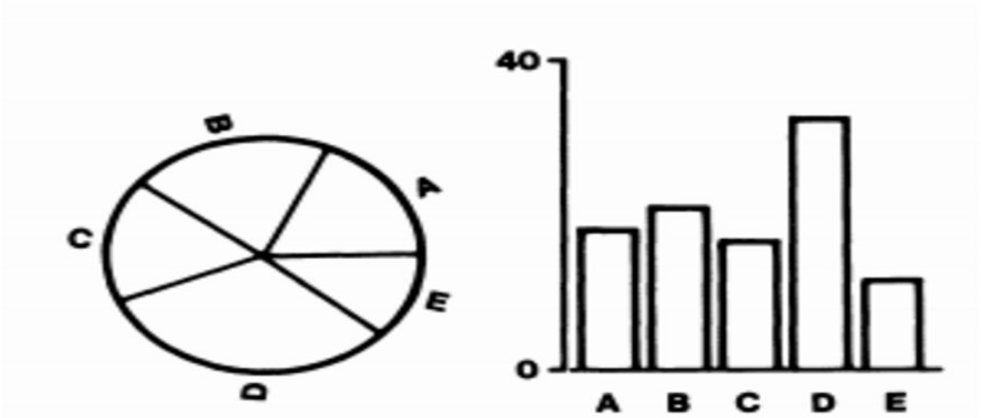
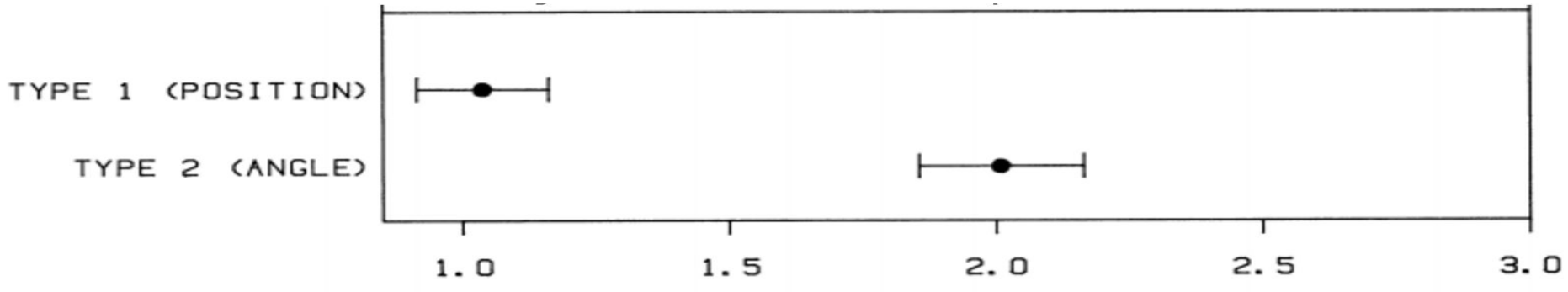


Figure 3. Graphs from position–angle experiment.



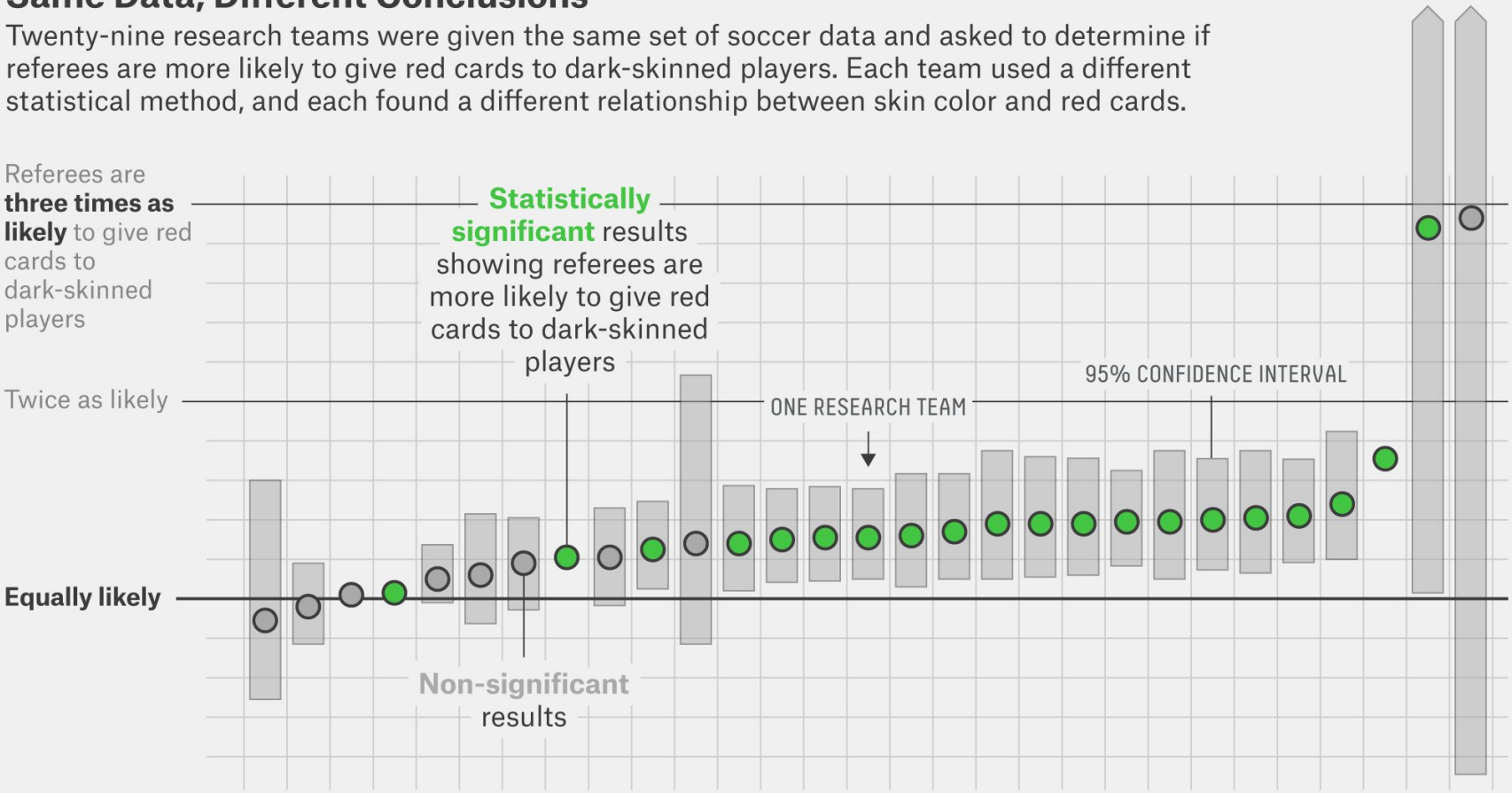
Source: Cleveland and McGill JASA 1984

Human data interaction

(observational epidemiology)

Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.



FIVETHIRTYEIGHT

SOURCE: BRIAN NOSEK ET AL.

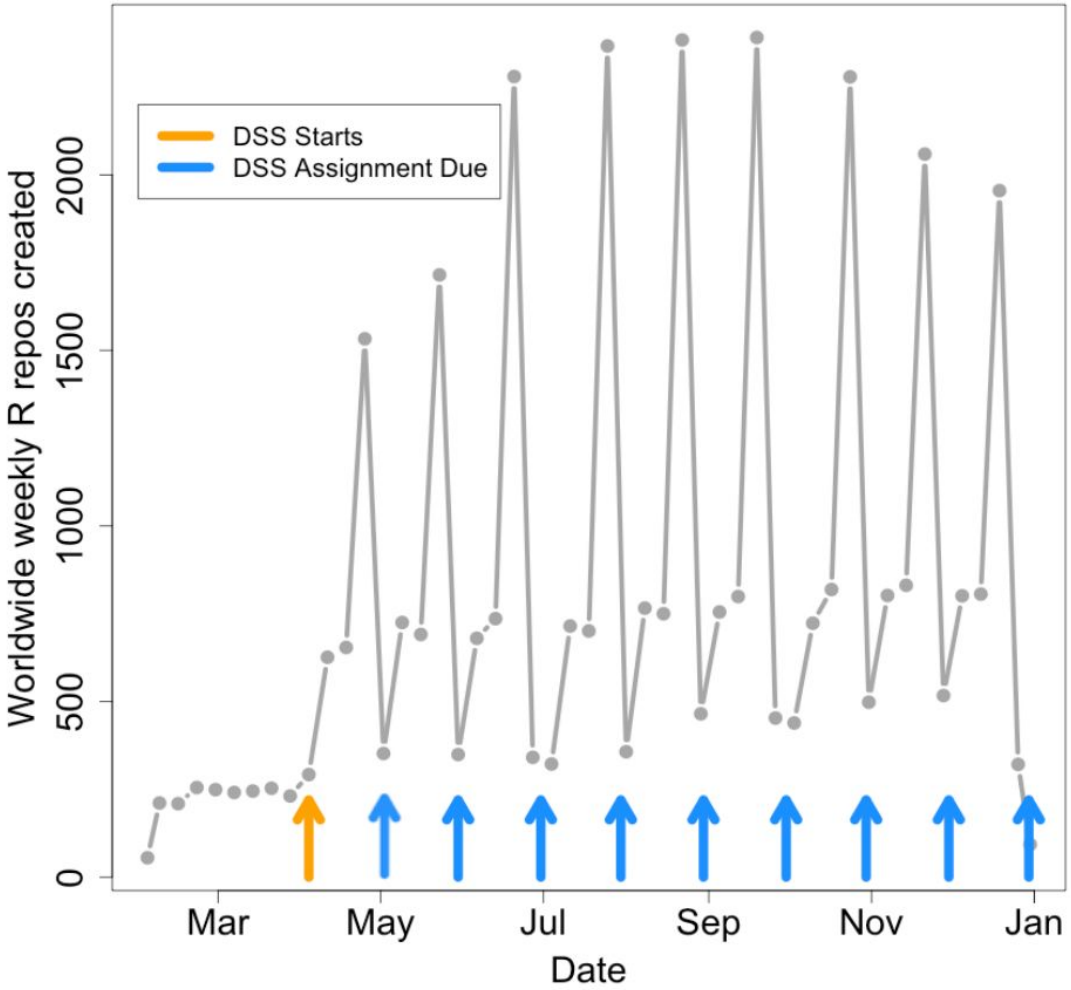
Source: https://fivethirtyeight.com/features/science-isnt-broken/

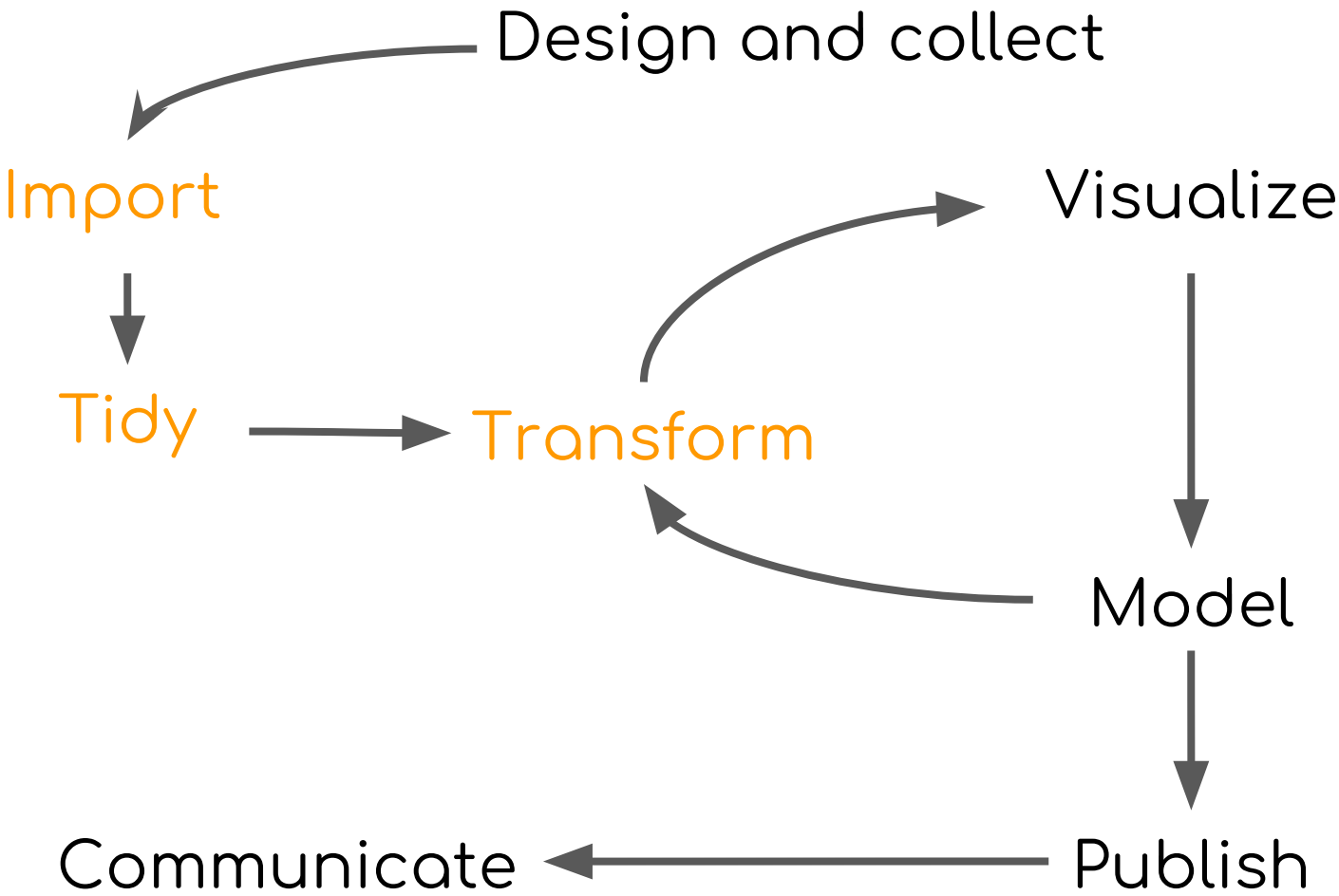
MSD in R
6 Courses
35K+ Enrollments

Executive Data Science
5 Courses
150K+ Enrollments

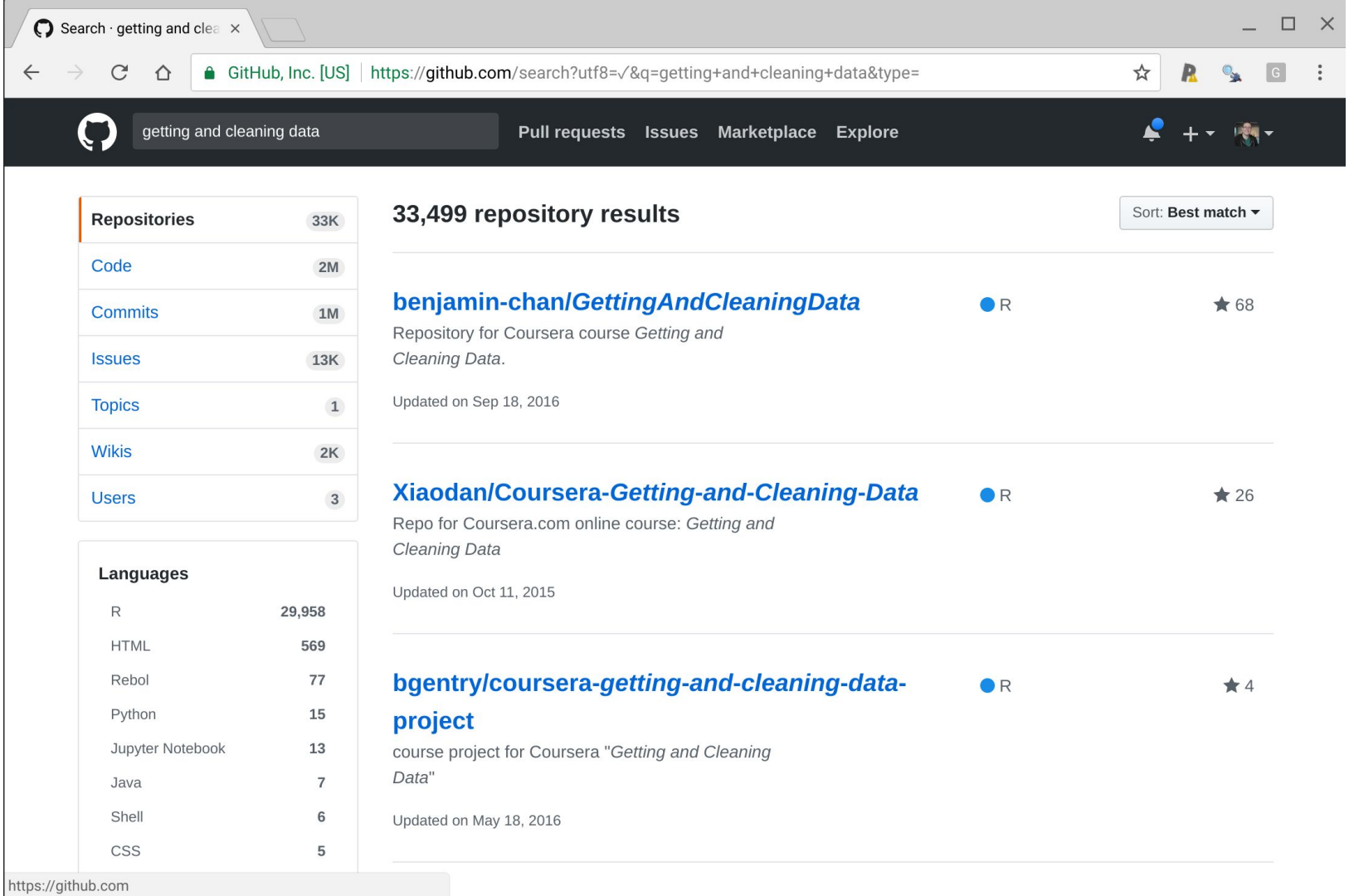
Genomic Data Science
9 Courses
230k+ Enrollments

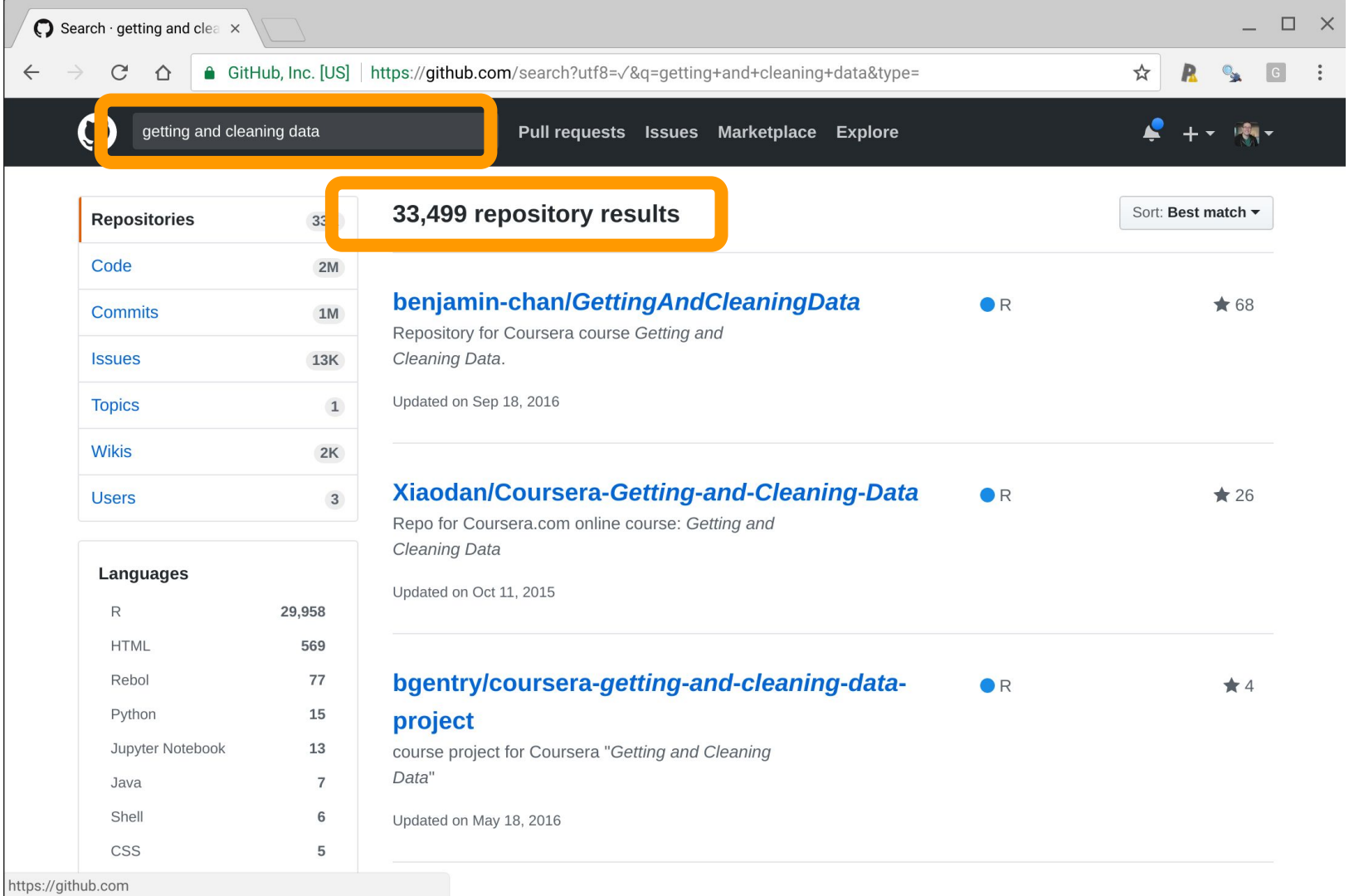
Data Science
10 Courses
4.4M+ Enrollments

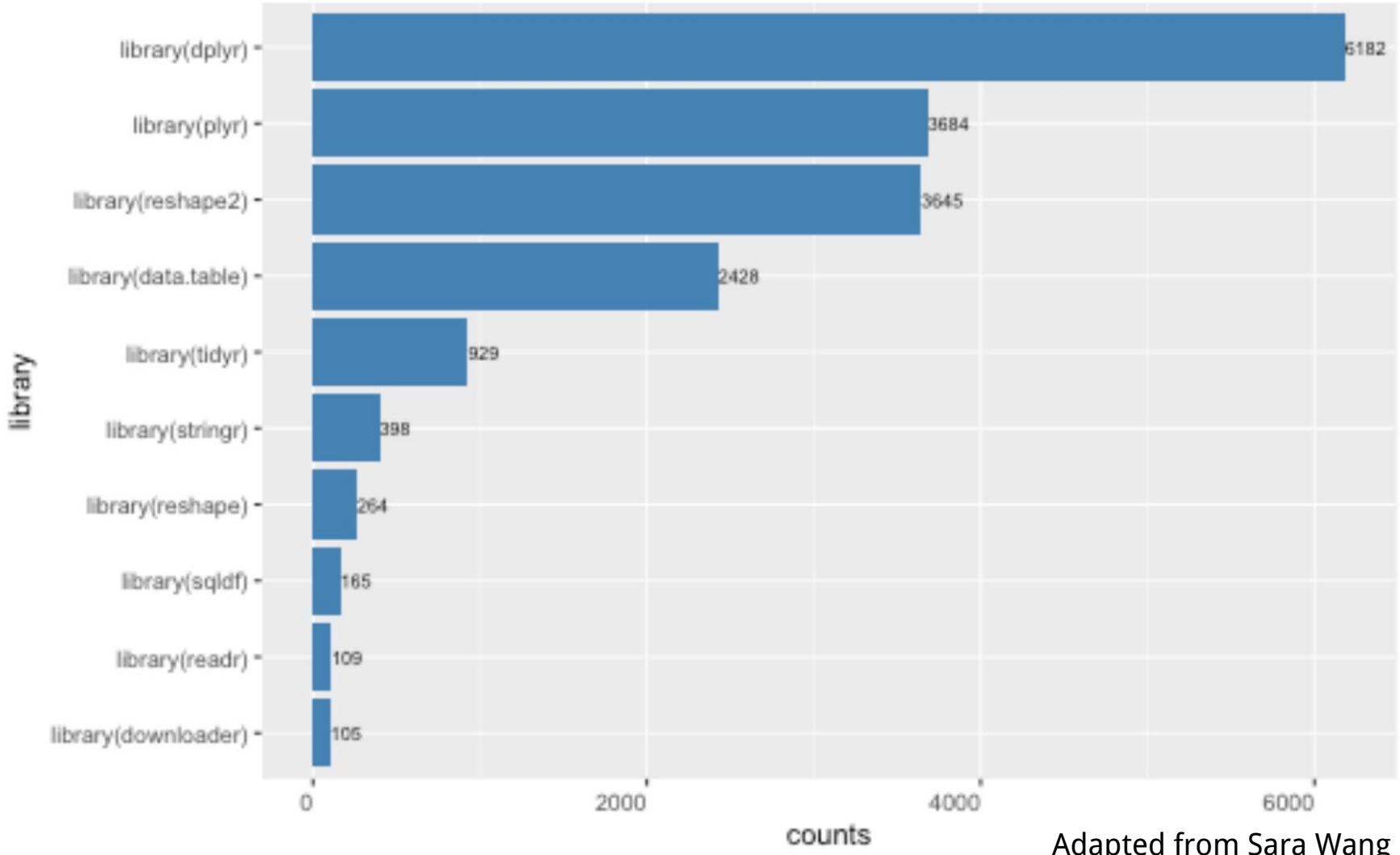

DSS Starts
DSS Assignment Due
Worldwide weekly R repos created
0
500
1000
1500
2000
Mar
May
Jul
Sep
Nov
Jan
Date



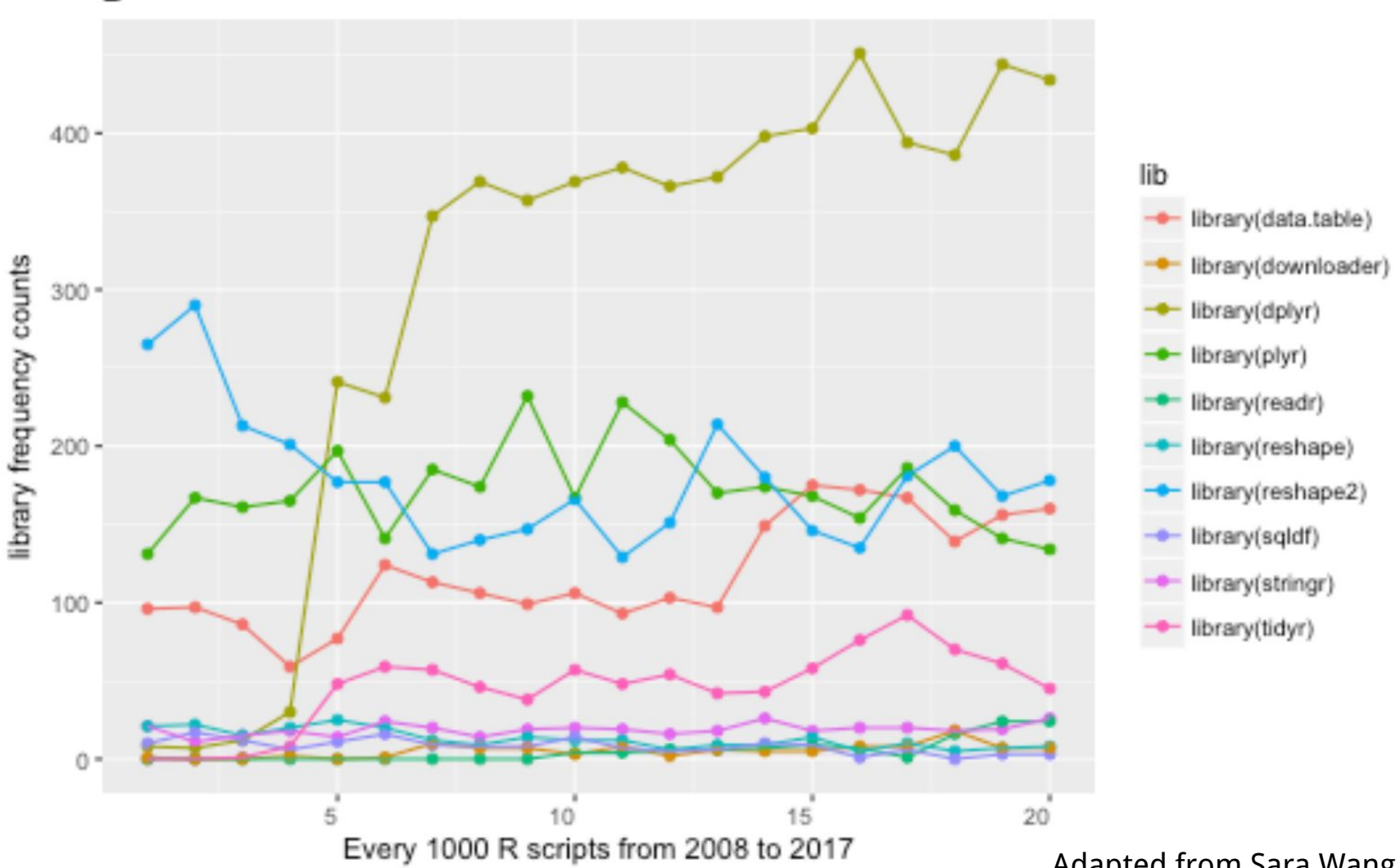
Adapted from Hadley Wickham


Search · getting and clea
GitHub, Inc. [US] | https://github.com/search?utf8=✓&q=getting+and+cleaning+data&type=
getting and cleaning data
Pull requests
Issues
Marketplace
Explore
Repositories 33K
Code 2M
Commits 1M
Issues 13K
Topics 1
Wikis 2K
Users 3
Languages
R 29,958
HTML 569
Rebol 77
Python 15
Jupyter Notebook 13
Java 7
Shell 6
CSS 5
33,499 repository results
Sort: Best match
benjamin-chan/GettingAndCleaningData
Repository for Coursera course Getting and Cleaning Data.
R
68
Updated on Sep 18, 2016
Xiaodan/Coursera-Getting-and-Cleaning-Data
Repo for Coursera.com online course: Getting and Cleaning Data
R
26
Updated on Oct 11, 2015
bgentry/coursera-getting-and-cleaning-data-project
course project for Coursera "Getting and Cleaning Data"
R
4
Updated on May 18, 2016
https://github.com

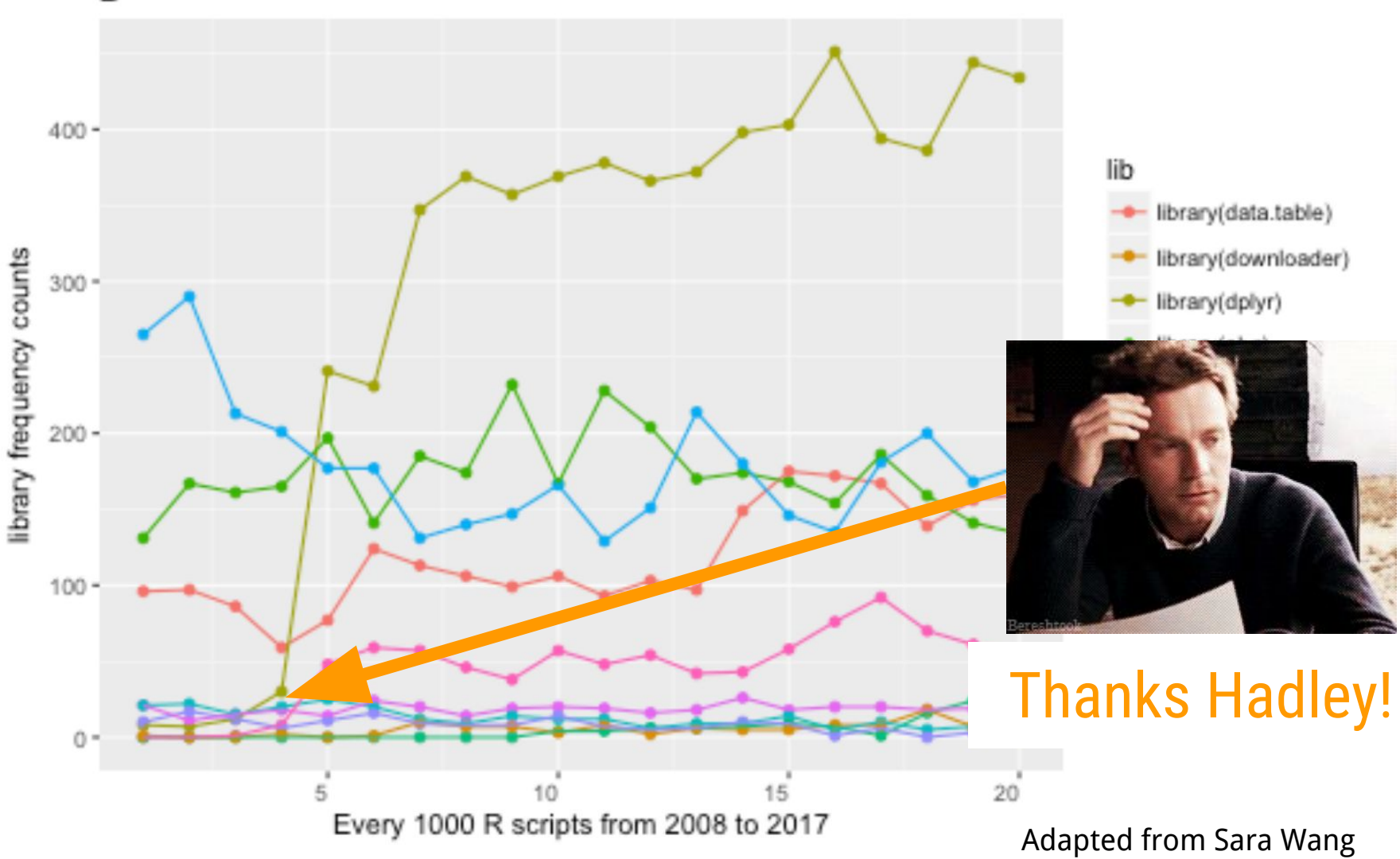

Search · getting and clea
GitHub, Inc. [US] https://github.com/search?utf8=✓&q=getting+and+cleaning+data&type=
getting and cleaning data
Pull requests
Issues
Marketplace
Explore
Repositories 33
Code 2M
Commits 1M
Issues 13K
Topics 1
Wikis 2K
Users 3
Languages
R 29,958
HTML 569
Rebol 77
Python 15
Jupyter Notebook 13
Java 7
Shell 6
CSS 5
33,499 repository results
Sort: Best match
benjamin-chan/GettingAndCleaningData
Repository for Coursera course Getting and Cleaning Data.
R
★ 68
Updated on Sep 18, 2016
Xiaodan/Coursera-Getting-and-Cleaning-Data
Repo for Coursera.com online course: Getting and Cleaning Data
R
★ 26
Updated on Oct 11, 2015
bgentry/coursera-getting-and-cleaning-data-project
course project for Coursera "Getting and Cleaning Data"
R
★ 4
Updated on May 18, 2016
https://github.com



Adapted from Sara Wang



Adapted from Sara Wang



Thanks Hadley!

Adapted from Sara Wang

Human data interaction

(randomized trials)

Quiz

(OPTIONAL) Data analysis practice with immediate feedback (NEW! 10/18/2017)

Question 1

1 Point

VARIATION 1

You are being asked to participate in a research experiment with the purpose of better understanding how people analyze data. If you complete this quiz, you are giving your consent to participate in the study. This quiz involves a short data analysis that gives you a chance to practice the regression concepts you have learned so far.

We anticipate that this will take about 15 minutes to complete. You will be receiving feedback on your work immediately after submission. For this reason, we ask that you do not post on the forums about this quiz to maintain the integrity of this experiment.

Thank you for helping us learn more about data science! -Brian, Roger, Jeff

---

Your assignment is to study how income varies across different categories of college majors. You will be using data from a study of recent college graduates. Make sure to use good practices that you have learned so far in this course and previous courses in the specialization.In particular, it is good practice to specify an analysis plan early in the process to avoid the “p-hacking” behavior of trying many analyses to find one that has desired results. If you want to learn more about “p-hacking”, you can visit

Quiz

(OPTIONAL) Data analysis practice with immediate feedback (NEW! 10/18/2017)

Question 1

1 Point

VARIATION 1

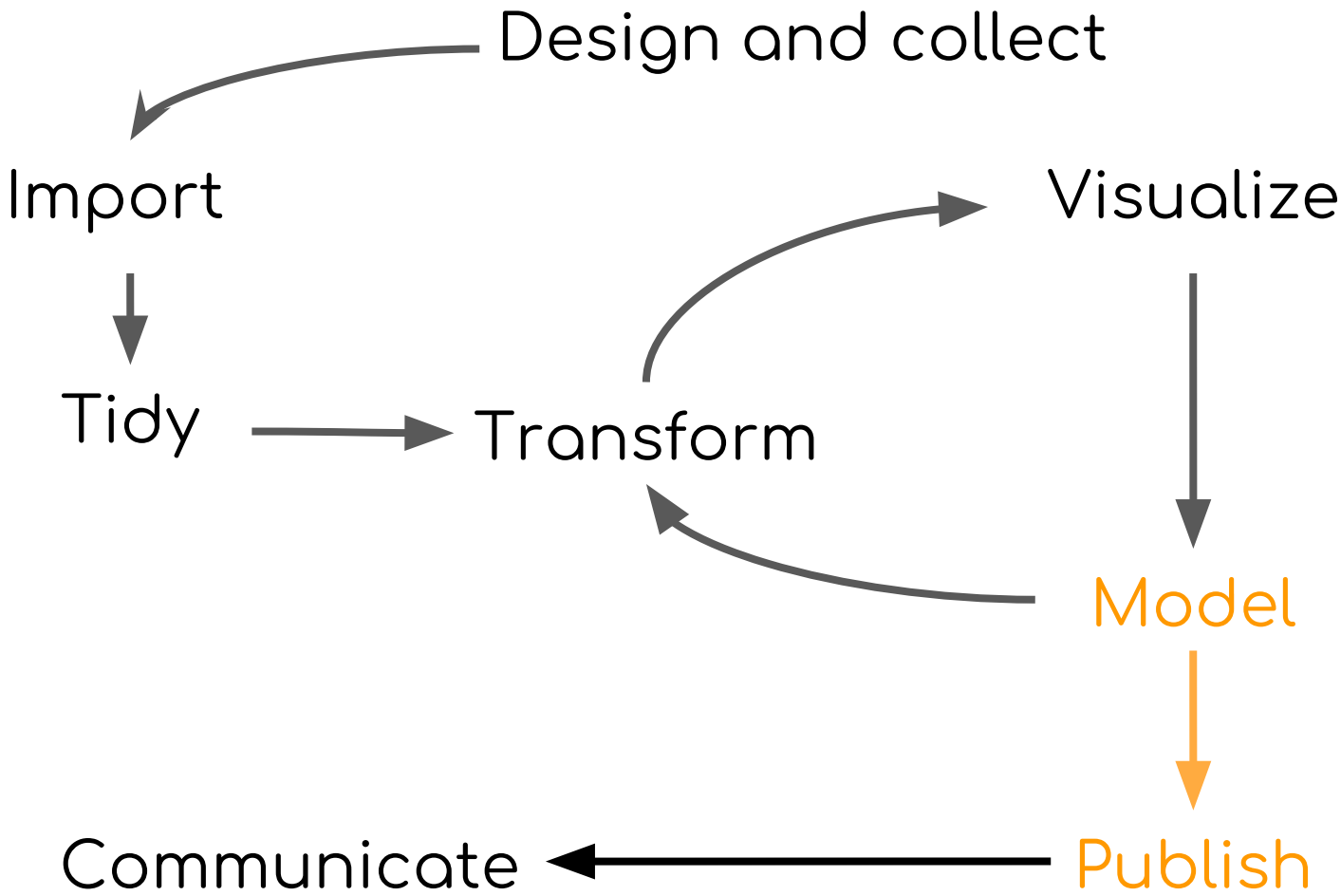
You are being asked to participate in a research experiment with the purpose of better understanding how people analyze data. If you complete this quiz, you are giving your consent to participate in the study. This quiz involves a short data analysis that gives you a chance to practice the regression concepts you have learned so far.

We anticipate that this will take about 15 minutes to complete. You will be receiving feedback on your work immediately after submission. For this reason, we ask that you do not post on the forums about this quiz to maintain the integrity of this experiment.

Thank you for helping us learn more about data science! -Brian, Roger, Jeff

---

Your assignment is to study how income varies across different categories of college majors. You will be using data from a study of recent college graduates. Make sure to use good practices that you have learned so far in this course and previous courses in the specialization.In particular, it is good practice to specify an analysis plan early in the process to avoid the “p-hacking” behavior of trying many analyses to find one that has desired results. If you want to learn more about “p-hacking”, you can visit



Adapted from Hadley Wickham

“...Coming up with nothing is unacceptable...”

Article Text

Article info

Citation Tools

Share

Responses

Article metrics

Ethics
Research

PDF

Identifying bioethical issues in biostatistical consulting: findings from a US national pilot survey of biostatisticians

Min Qi Wang[1], Alice F Yan[2], Ralph V Katz[3]

Author affiliations +

Abstract

Objectives The overall purposes of this first US national pilot study were to (1) test the feasibility of online administration of the Bioethical Issues in Biostatistical Consulting (BIBC) Questionnaire to a random sample of American Statistical Association (ASA) members; (2) determine the prevalence and relative severity of a broad array of bioethical violations requests that are presented to biostatisticians by investigators seeking biostatistical consultations; and (3) establish the sample size needed for a full-size phase II study.

Design A descriptive survey as approved and endorsed by the ASA.

Source: http://bmjopen.bmj.com/content/7/11/e018491

[It has been widely noted that there is a relationship between x1 and Y.] We are interested in studying the magnitude of this relationship and reporting results to the president of the company. Using the data available at **[data1/data2]**, use a linear regression model to investigate this relationship.

Would you be confident telling the company president that there is a meaningful relationship between x1 and y?

A. Yes
B. No

data1: x2 confounds the relationship between Y and x1
data2: x2 is not a confounder but is predictive of Y

Source: Myint et al (in prep)

Confounded

Primed

68% found signal

(n = 263)

Not Primed

57% found signal

(n = 252)

Source: Myint et al (in prep)

Not confounded

Primed

78% found signal

(n = 263)

Not Primed

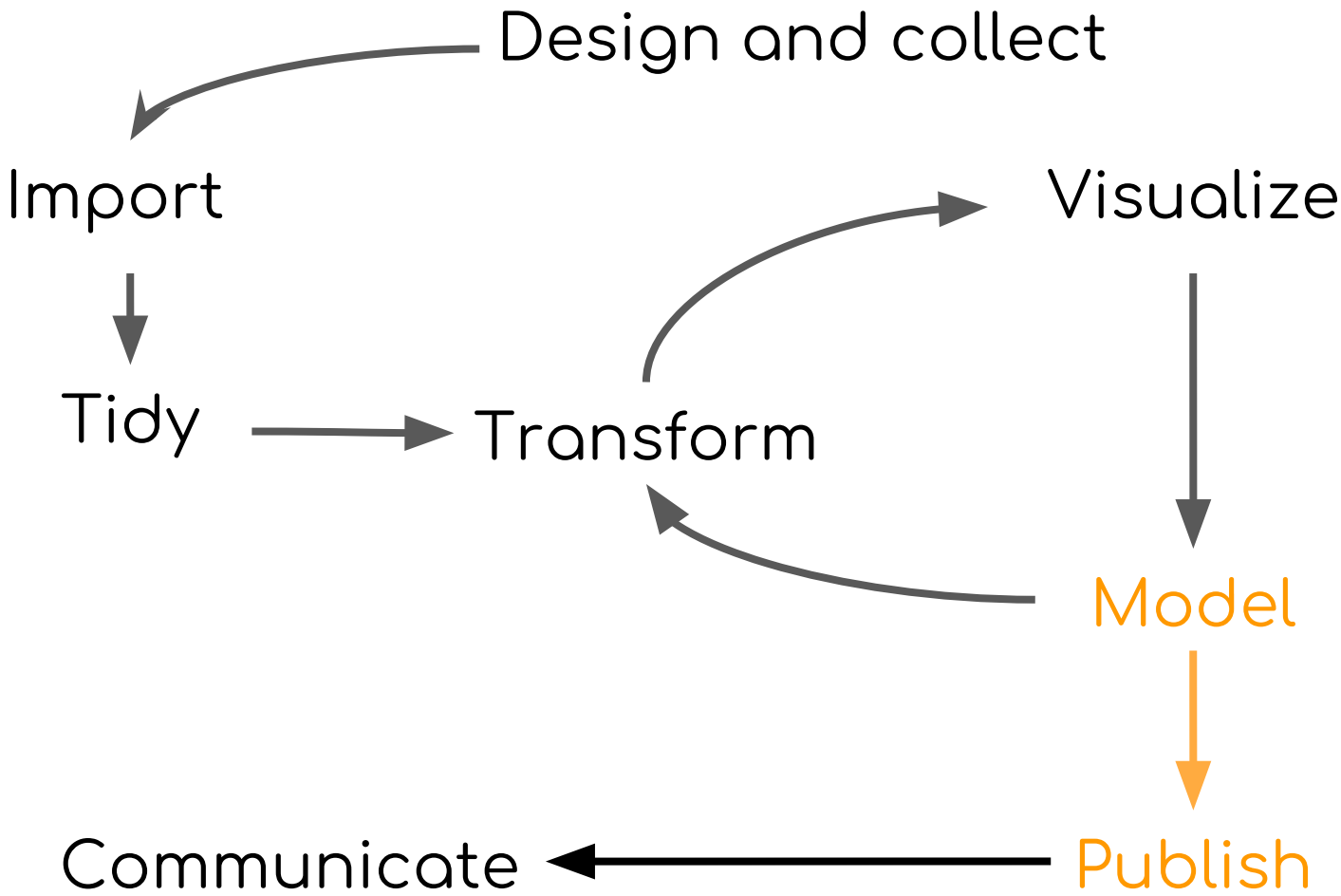
87% found signal

(n = 258)

Source: Myint et al (in prep)

Human data interaction

(interventional randomized trials)



Adapted from Hadley Wickham

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

General Article

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

Psychological Science
22(11) 1359–1366

DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

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

Abstract

In this article, we accomplish two things. First, we show that despite empirical psychologists' nominal endorsement of a low rate of false-positive findings (≤ .05), flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not. We present computer simulations and a pair of actual experiments that demonstrate how unacceptably easy it is to accumulate (and report) statistically significant evidence for a false hypothesis. Second, we suggest a simple, low-cost, and straightforwardly effective disclosure-based solution to this problem. The solution involves six concrete requirements for authors and four guidelines for reviewers, all of which impose a minimal burden on the publication process.

Keywords

methodology, motivated reasoning, publication, disclosure

Received 3/17/11; Revision accepted 5/23/11

Your assignment is to study how income varies across different categories of college majors. You will be using data from a study of recent college graduates. Make sure to use good practices that you have learned so far in this course and previous courses in the specialization. [In particular, it is good practice to specify an analysis plan early in the process to avoid the “p-hacking” behavior of trying many analyses to find one that has desired results. If you want to learn more about “p-hacking”, you can visit https://projects.fivethirtyeight.com/p-hacking/.]

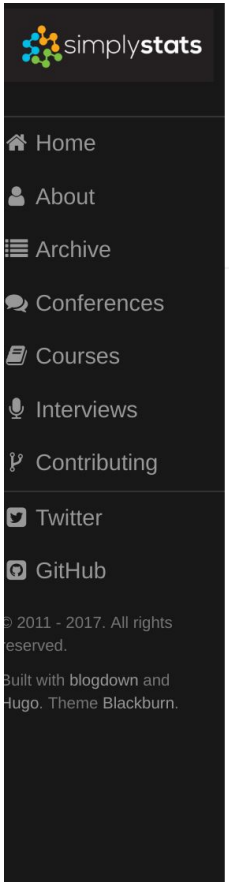
If you will proceed with the analysis, click “Yes”. Otherwise, click “No”.

Source: Myint et al (in prep)

	Significant	Not significant
Warned	66	90
Not Warned	59	88

Source: Myint et al (in prep)

$$Y = X + e$$



The researcher degrees of freedom - recipe tradeoff in data analysis

Jeff Leek 2013/07/31

An important concept that is only recently gaining the attention it deserves is researcher degrees of freedom. From Simmons et al.:

> The culprit is a construct we refer to as researcher degrees of freedom. In the course of collecting and analyzing data, researchers have many decisions to make: Should more data be collected? Should some observations be excluded? Which conditions should be combined and which ones compared? Which control variables should be considered? Should specific measures be combined or transformed or both?

So far, researcher degrees of freedom has primarily been used with negative connotations. This probably stems from the original definition of the idea which focused on how analysts could “manufacture” statistical significance by changing the way the data was processed without disclosing those changes. Reproducible research and distributed code would of course address these issues to some extent. But it is still relatively easy to obfuscate dubious analysis by dressing it up in technical language.

Source: https://simplystatistics.org/2013/07/31/the-researcher-degrees-of-freedom-recipe-tradeoff-in-data-analysis/


Y = X + e analyst + e sampling

MSD in R
6 Courses
35K+ Enrollments

Executive Data Science
5 Courses
150K+ Enrollments

Genomic Data Science
9 Courses
230k+ Enrollments

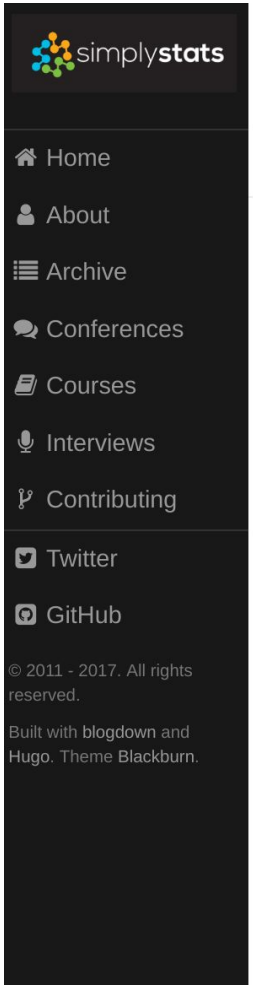
Data Science
10 Courses
4.4M+ Enrollments

MSD in R
6 Courses
35K+ Enrollments

Executive Data Science
5 Courses
150K+ Enrollments

Genomic Data Science
9 Courses
230k+ Enrollments

Data Science
10 Courses
4.4M+ Enrollments



Data analysis subcultures

Jeff Leek 2015/04/29

Roger and I responded to the controversy around the journal that banned p-values today in Nature. A piece like this requires a lot of information packed into very little space but I thought one idea that deserved to be talked about more was the idea of data analysis subcultures. From the paper:

> Data analysis is taught through an apprenticeship model, and different disciplines develop their own analysis subcultures. Decisions are based on cultural conventions in specific communities rather than on empirical evidence. For example, economists call data measured over time 'panel data', to which they frequently apply mixed-effects models. Biomedical scientists refer to the same type of data structure as 'longitudinal data', and often go at it with generalized estimating equations.

I think this is one of the least appreciated components of modern data analysis. Data analysis is almost entirely taught through an apprenticeship culture with completely different behaviors taught in different disciplines. All of these disciplines agree about the mathematical optimality of specific methods under very specific conditions. That is why you see methods like randomized trials [Roger and I responded to the controversy around the journal that banned p-values today in Nature. A piece like this requires a lot of information packed into very little space but I thought one idea that deserved to be talked about more was the idea of data analysis subcultures. From the paper:

Source: https://simplystatistics.org/2015/04/29/data-analysis-subcultures/

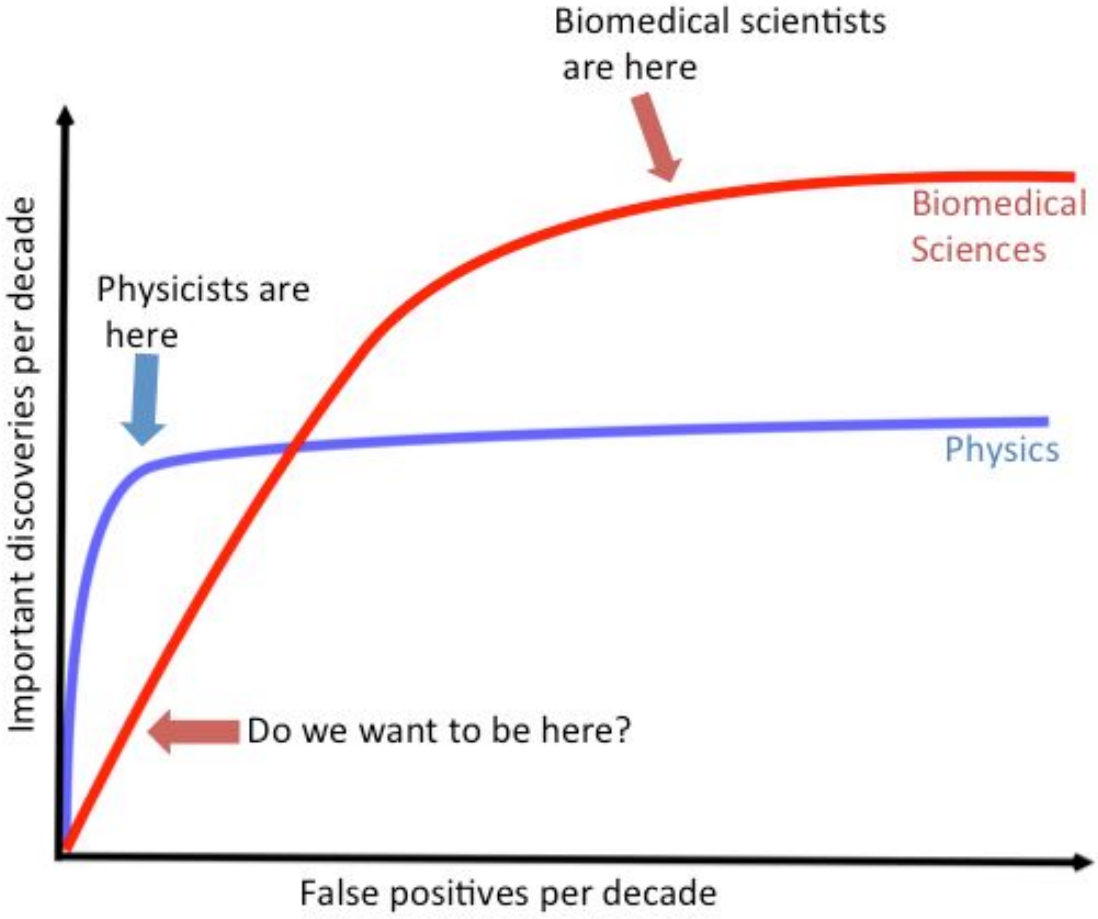
$$Y = X + e_{population} + e_{analyst} + e_{sampling}$$

3

We need to say what we mean

The tools already exist

Humans are the problem



https://simplystatistics.org/2013/08/01/the-roc-curves-of-science/

Leek group

- Shannon Ellis
- **Aboozar Hadavand**
- **Lucy D'Agostino McGowan**
- **Leslie Myint**
- Jack Fu
- Sara Wang
- Kayode Sosina
- You?

Alumni

- Simina Boca
- Hilary Parker
- Andrew Jaffe
- Alyssa Frazee
- Prasad Patil
- Leo Collado Torres
- Abhi Nellore
- Kai Kammers
- **Nick Carchedi**
- **Sean Kross**
- Divya Narayanan

Collaborators

- Kasper Hansen
- Margaret Taub
- **Leah Jager**
- Ben Langmead
- **Roger Peng**
- **Aaron Fisher**
- **Brooke Anderson**
- **Ira Gooding**

```
head(college)
fit <- lm(rank ~ major_category - 1, college)
summary(fit)
```

```
 str(college)
 library(ggplot2)
 library(dplyr)
 g1 <- ggplot(college, aes(x = major_category, y = median, fill = major_category))
 g1 <- g1 + geom_boxplot()
 g1 <- g1 + theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none")
 g1
 tapply(college$sample_size, college$major_category, sum)
 majormean <- tapply(college$sample_size, college$major_category, mean)
 max(majormean)
 college$major_category <- as.factor(college$major_category)
 college2 <- college
 college2$major_category <- relevel(college$major_category, ref = "Business")
 fit1 <- lm(median ~ major_category, data = college2)
 fit2 <- update(fit1, median ~ major_category + perc_men)
 fit3 <- update(fit1, median ~ major_category + perc_men + perc_employed)
 fit4 <- update(fit1, median ~ major_category + perc_men + perc_employed +
perc_employed_fulltime_yearround)
 fit5 <- update(fit1, median ~ major_category + perc_men + perc_employed +
perc_employed_fulltime_yearround + perc_college_jobs)
 anova(fit1, fit2, fit3, fit4)
 summary(fit5)
 tail(hatvalues(fit5)[order(hatvalues(fit5, decreasing = T))], 5)
 tail(cooks.distance(fit5)[order(cooks.distance(fit5, decreasing = T))], 5)
```

```
 str(college)
 library(ggplot2)
 library(dplyr)
 g1 <- ggplot(college, aes(x = major_category, y = median, fill = major_category))
 g1 <- g1 + geom_boxplot()
 g1 <- g1 + theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none")
 g1
 tapply(college$sample_size, college$major_categ
 majormean <- tapply(college$sample_size, college
 max(majormean)
 college$major_category <- as.factor(college$majo
 college2 <- college
 college2$major_category <- relevel(college$major_category, ref = "Business")
 fit1 <- lm(median ~ major_category, data = college2)
 fit2 <- update(fit1, median ~ major_category + perc_men)
 fit3 <- update(fit1, median ~ major_category + perc_men + perc_employed)
 fit4 <- update(fit1, median ~ major_category + perc_men + perc_employed +
perc_employed_fulltime_yearround)
 fit5 <- update(fit1, median ~ major_category + perc_men + perc_employed +
perc_employed_fulltime_yearround + perc_college_jobs)
 anova(fit1, fit2, fit3, fit4)
 summary(fit5)
 tail(hatvalues(fit5)[order(hatvalues(fit5, decreasing = T))], 5)
 tail(cooks.distance(fit5)[order(cooks.distance(fit5, decreasing = T))], 5)
```

Setup

```
str(college)
library(ggplot2)
library(dplyr)
g1 <- ggplot(college, aes(x = major_category, y = median, fill = major_category))
g1 <- g1 + geom_boxplot()
g1 <- g1 + theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none")
g1
tapply(college$sample_size, college$major_category, sum)
majormean <- tapply(college$sample_size, college$major_category, mean)
max(majormean)
college$major_category <- as.factor(college$major_category)
college2 <- college
college2$major_category <- relevel(college$major_category, ref = "Business")
fit1 <- lm(median ~ major_category, data = college2)
fit2 <- update(fit1, median ~ major_category + perc_men)
fit3 <- update(fit1, median ~ major_category + perc_men + perc_employed`
fit4 <- update(fit1, median ~ major_category + perc_men + perc_employed
perc_employed_fulltime_yearround)
fit5 <- update(fit1, median ~ major_category + perc_men + perc_employed
perc_employed_fulltime_yearround + perc_college_jobs)
anova(fit1, fit2, fit3, fit4)
summary(fit5)
tail(hatvalues(fit5)[order(hatvalues(fit5, decreasing = T))], 5)
tail(cooks.distance(fit5)[order(cooks.distance(fit5, decreasing = T))], 5)
```

EDA

```
 str(college)
 library(ggplot2)
 library(dplyr)
 g1 <- ggplot(college, aes(x = major_category, y = median, fill = major_category))
 g1 <- g1 + geom_boxplot()
 g1 <- g1 + theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none")
 g1
 tapply(college$sample_size, college$major_category, sum)
 majormean <- tapply(college$sample_size, college$major_category, mean)
 max(majormean)
 college$major_category <- as.factor(college$major_category)
 college2 <- college
 college2$major_category <- relevel(college$major_category, ref = "Business")
 fit1 <- lm(median ~ major_category, data = college2)
 fit2 <- update(fit1, median ~ major_category + perc_men)
 fit3 <- update(fit1, median ~ major_category + perc_men + perc_employed)
 fit4 <- update(fit1, median ~ major_category + perc_r
perc_employed_fulltime_yearround)
 fit5 <- update(fit1, median ~ major_category + perc_r
perc_employed_fulltime_yearround + perc_college_jobs)
 anova(fit1, fit2, fit3, fit4)
 summary(fit5)
 tail(hatvalues(fit5)[order(hatvalues(fit5, decreasing = T))], 5)
 tail(cooks.distance(fit5)[order(cooks.distance(fit5, decreasing = T))], 5)
```

Munging

```
str(college)
library(ggplot2)
library(dplyr)
g1 <- ggplot(college, aes(x = major_category, y = median, fill = major_category))
g1 <- g1 + geom_boxplot()
g1 <- g1 + theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none")
g1
tapply(college$sample_size, college$major_category,
majormean <- tapply(college$sample_size, college$maj
max(majormean)
college$major_category <- as.factor(college$major_ca
college2 <- college
college2$major_category <- relevel(college$major_category, ref = "Business")
fit1 <- lm(median ~ major_category, data = college2)
fit2 <- update(fit1, median ~ major_category + perc_men)
fit3 <- update(fit1, median ~ major_category + perc_men + perc_employed)
fit4 <- update(fit1, median ~ major_category + perc_men + perc_employed +
perc_employed_fulltime_yearround)
fit5 <- update(fit1, median ~ major_category + perc_men + perc_employed +
perc_employed_fulltime_yearround + perc_college_jobs)
anova(fit1, fit2, fit3, fit4)
summary(fit5)
tail(hatvalues(fit5)[order(hatvalues(fit5, decreasing = T))], 5)
tail(cooks.distance(fit5)[order(cooks.distance(fit5, decreasing = T))], 5)
```

Modeling

```
 str(college)
 library(ggplot2)
 library(dplyr)
 g1 <- ggplot(college, aes(x = major_category, y = median, fill = major_category))
 g1 <- g1 + geom_boxplot()
 g1 <- g1 + theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none")
 g1
 tapply(college$sample_size, college$major_cate
 majormean <- tapply(college$sample_size, colle
 max(majormean)
 college$major_category <- as.factor(college$ma
 college2 <- college
 college2$major_category <- relevel(college$major_category, ref = "Business")
 fit1 <- lm(median ~ major_category, data = college2)
 fit2 <- update(fit1, median ~ major_category + perc_men)
 fit3 <- update(fit1, median ~ major_category + perc_men + perc_employed)
 fit4 <- update(fit1, median ~ major_category + perc_men + perc_employed +
perc_employed_fulltime_yearround)
 fit5 <- update(fit1, median ~ major_category + perc_men + perc_employed +
perc_employed_fulltime_yearround + perc_college_jobs)
 anova(fit1, fit2, fit3, fit4)
 summary(fit5)
 tail(hatvalues(fit5)[order(hatvalues(fit5, decreasing = T))], 5)
 tail(cooks.distance(fit5)[order(cooks.distance(fit5, decreasing = T))], 5)
```

Evaluation

```
str(college)
library(ggplot2)
library(dplyr)
g1 <- ggplot(college, aes(x = major_category, y = median, fill = major_category))
g1 <- g1 + geom_boxplot()
g1 <- g1 + theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none")
g1
tapply(college$sample_size, college$major_category, sum)
majormean <- tapply(college$sample_size, college$major_cate
max(majormean)
college$major_category <- as.factor(college$major_category)
college2 <- college
college2$major_category <- relevel(college$major_category, ref = "
fit1 <- lm(median ~ major_category, data = college2)
fit2 <- update(fit1, median ~ major_category + perc_men)
fit3 <- update(fit1, median ~ major_category + perc_men + perc_employed)
fit4 <- update(fit1, median ~ major_category + perc_men + perc_em
perc_employed_fulltime_yearround)
fit5 <- update(fit1, median ~ major_category + perc_men + perc_em
perc_employed_fulltime_yearround + perc_college_jobs)
anova(fit1, fit2, fit3, fit4)
summary(fit5)
tail(hatvalues(fit5)[order(hatvalues(fit5, decreasing = T))],
tail(cooks.distance(fit5)[order(cooks.distance(fit5, decreasing = T))], 5)
```

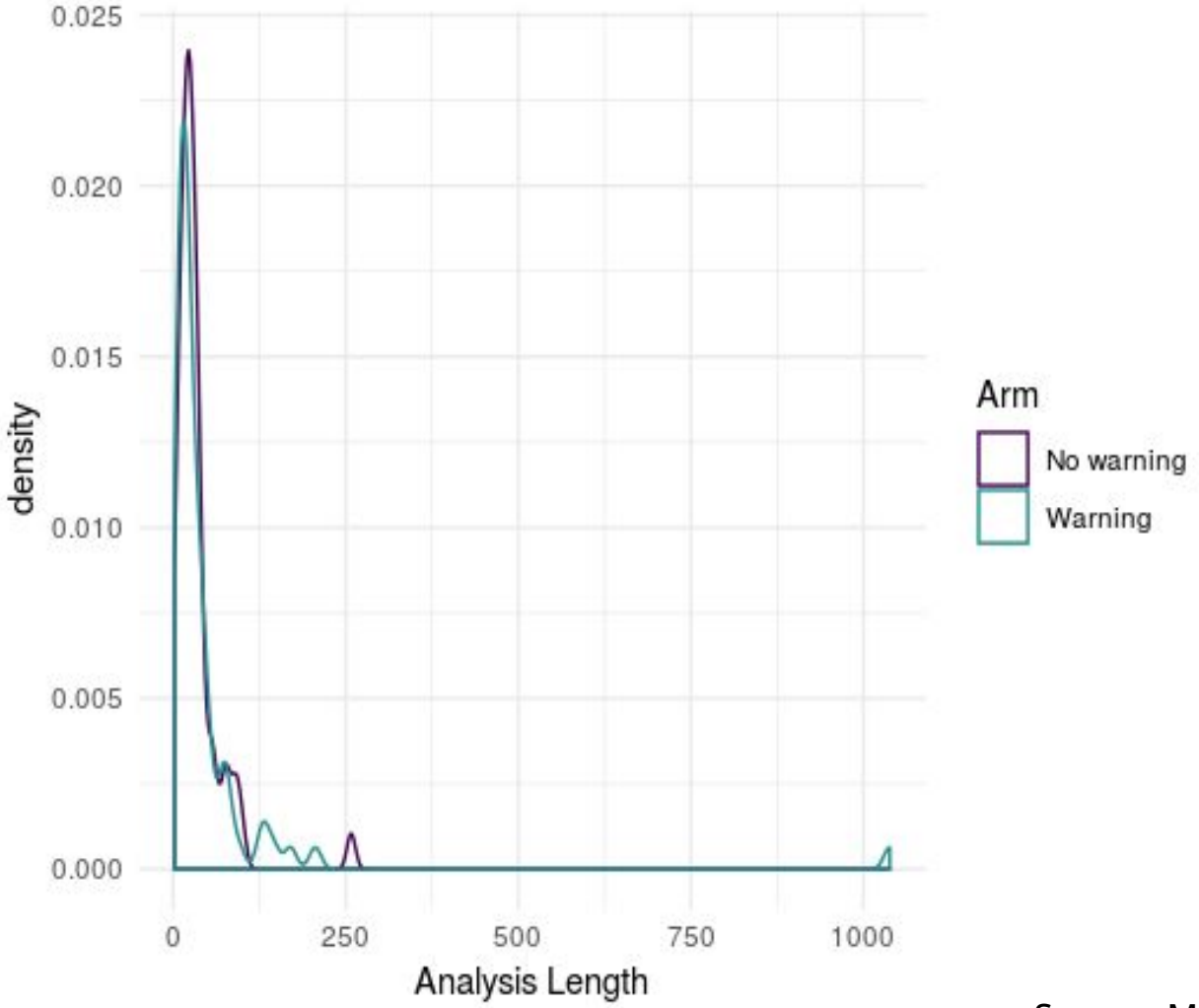
Setup

Exploratory

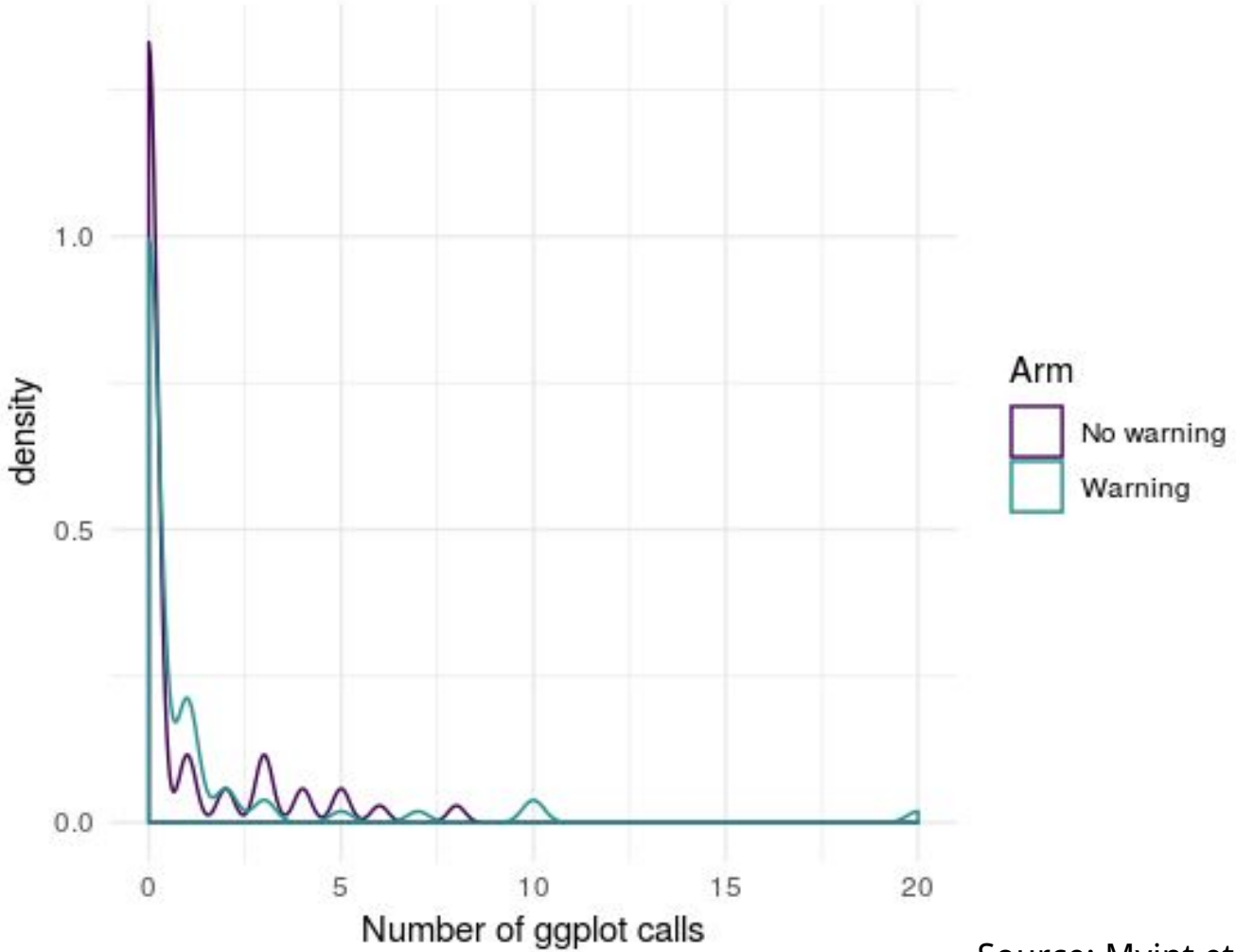
Munging

Modeling

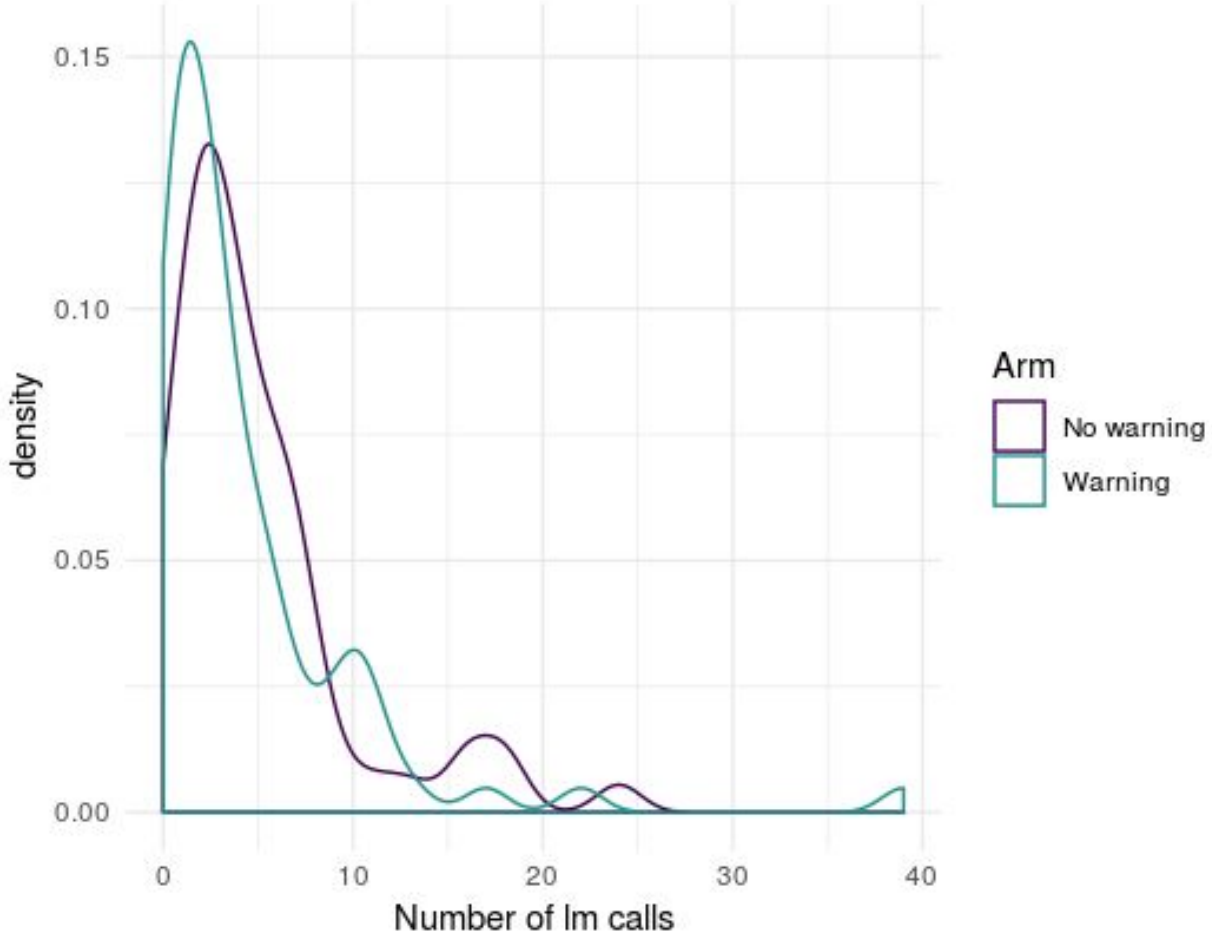
Evaluation



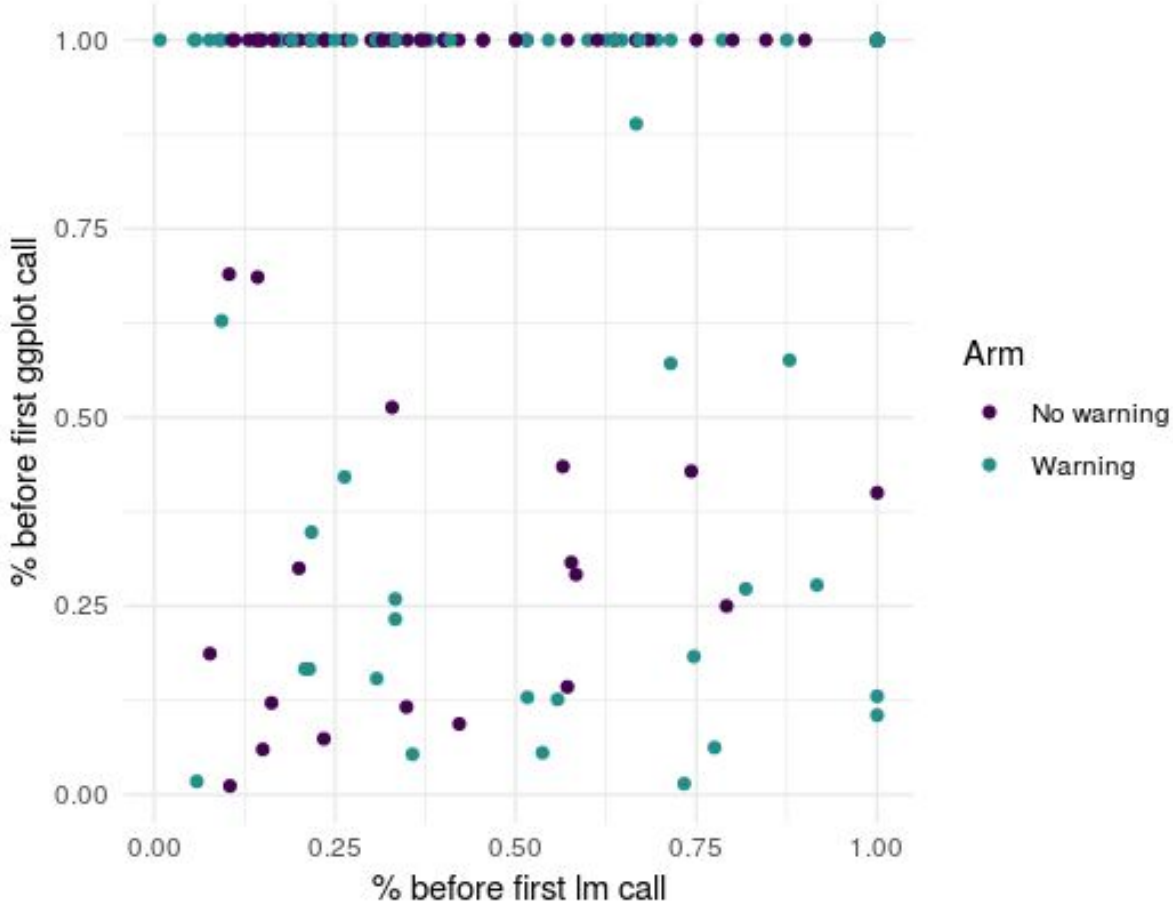
Source: Myint et al (in prep)



Source: Myint et al (in prep)



Source: Myint et al (in prep)



Source: Myint et al (in prep)