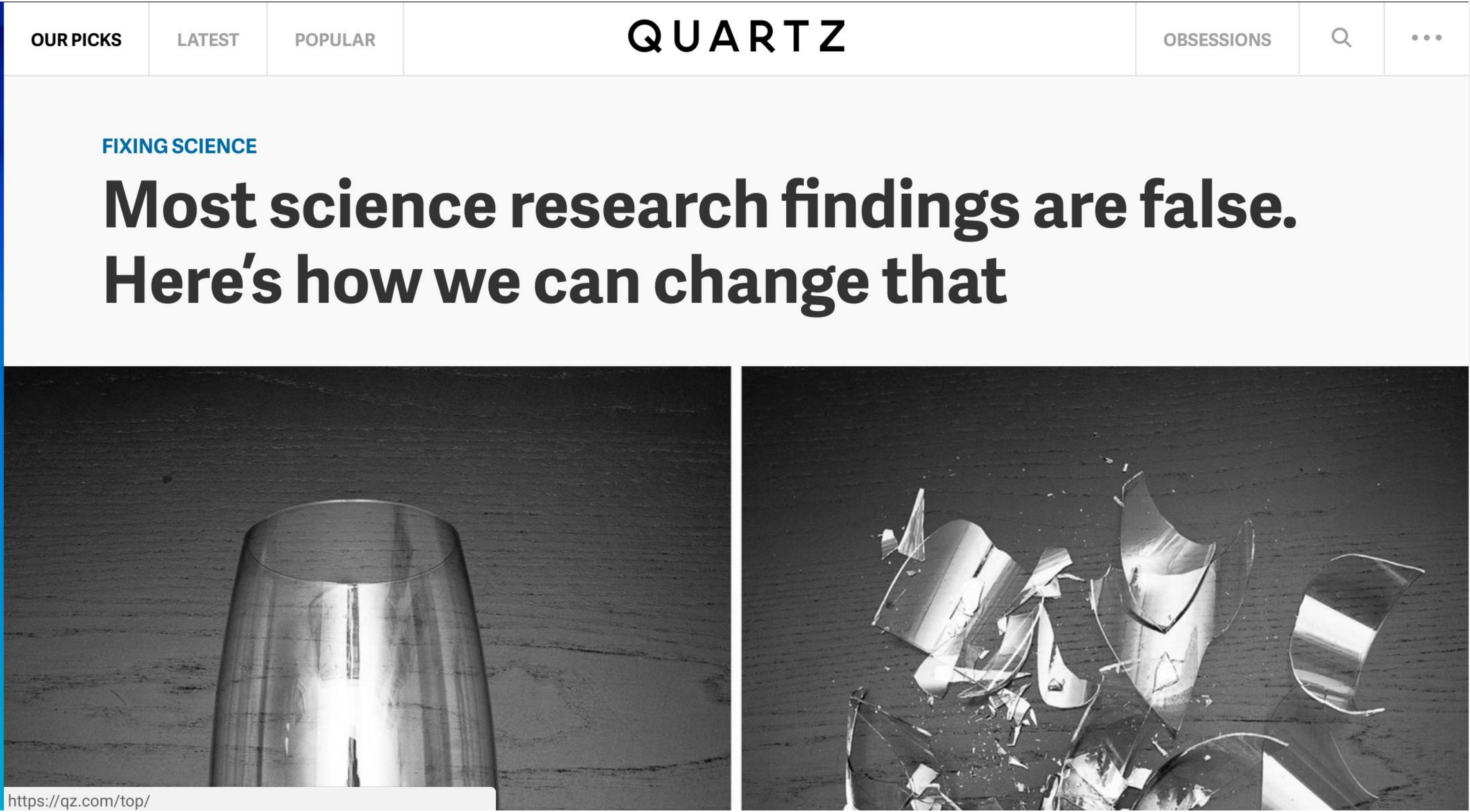
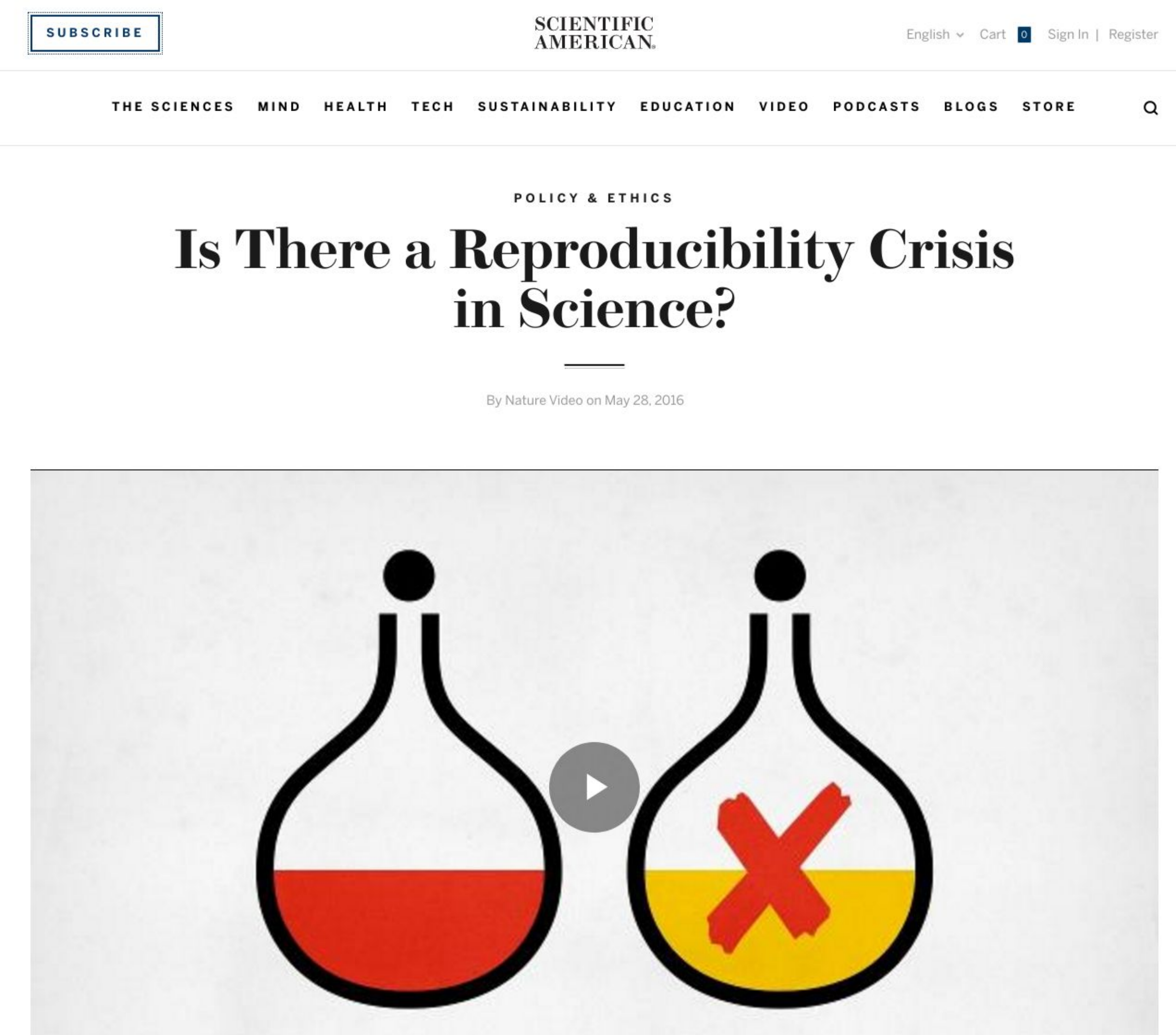
Coming to terms with data overload in science

@jtleek



FIXING SCIENCE

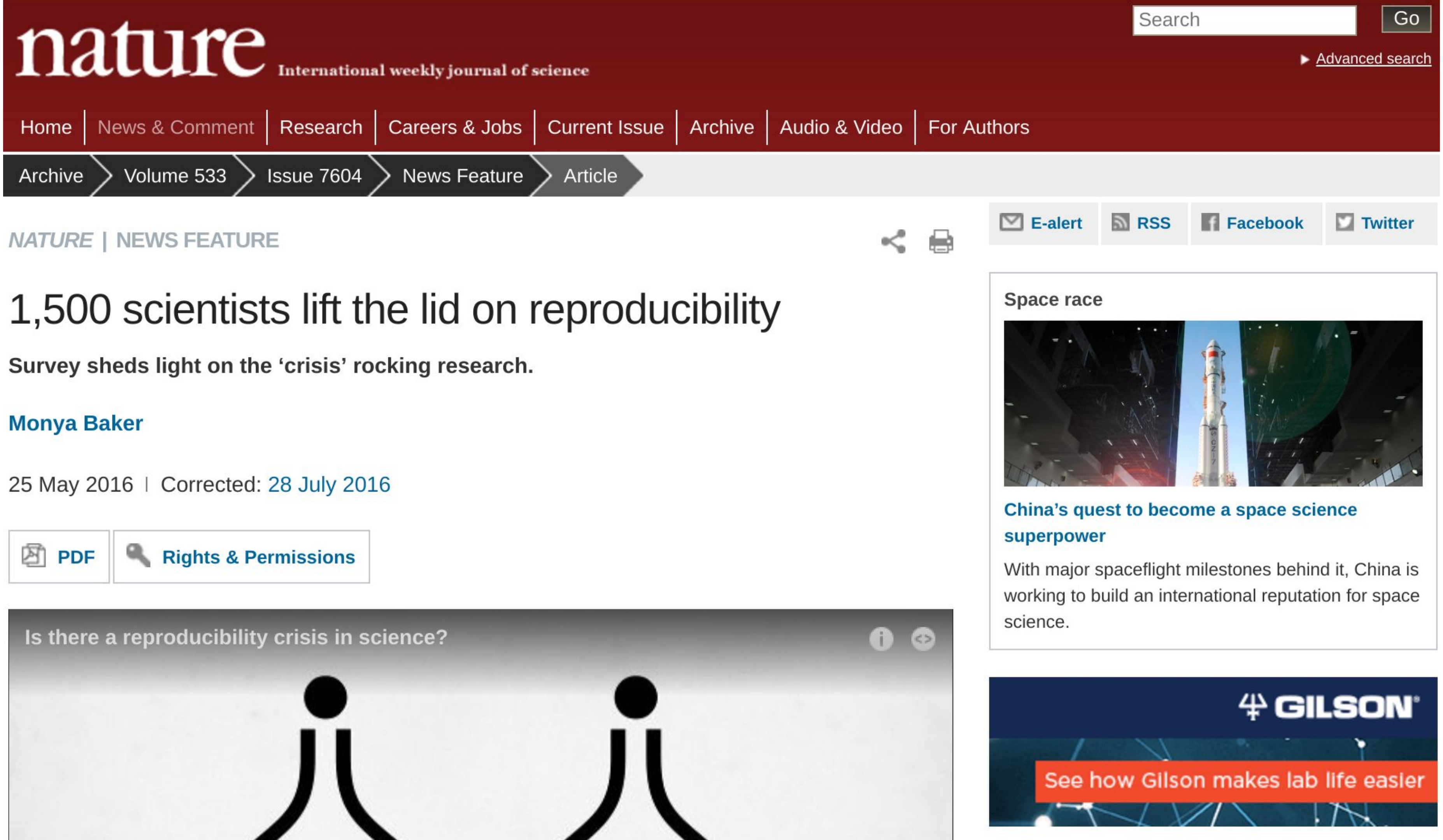
Most science research findings are false. Here's how we can change that



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.

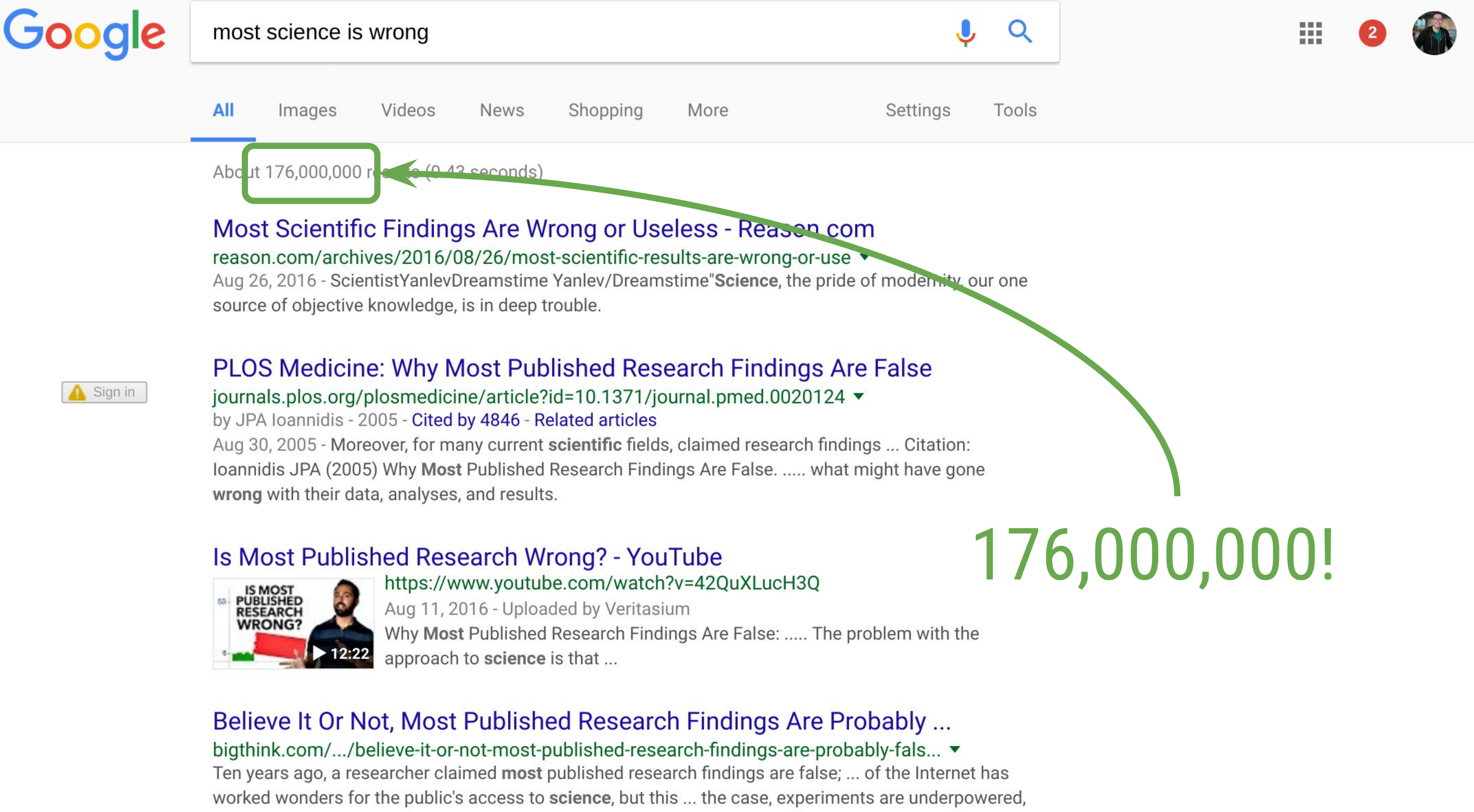
Monya Baker

25 May 2016 | Corrected: 28 July 2016

Space race

China's quest to become a space science superpower

With major spaceflight milestones behind it, China is working to build an international reputation for space science.



most science is wrong

About 176,000,000 results (0.43 seconds)

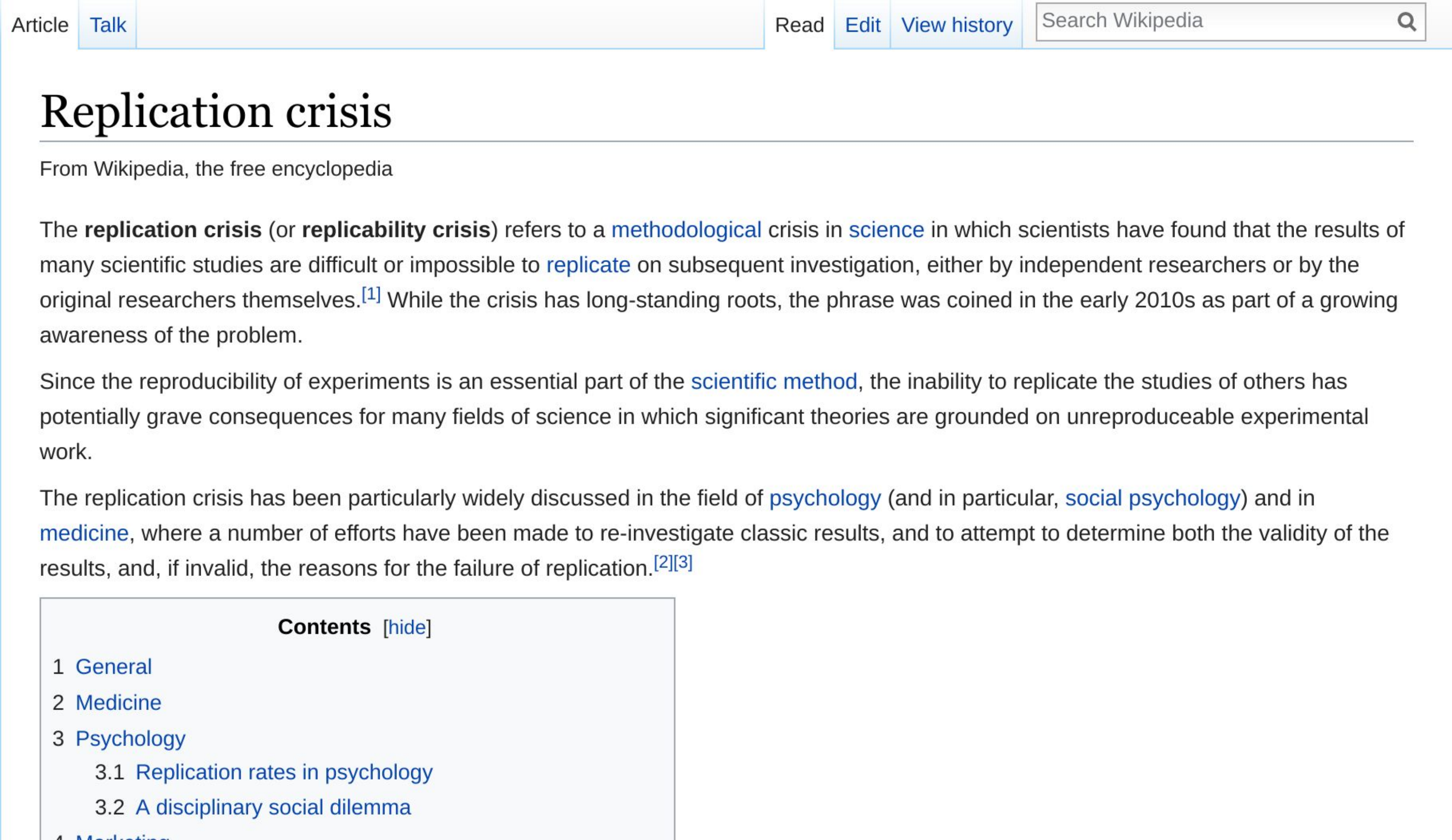
Most Scientific Findings Are Wrong or Useless - Reason.com
reason.com/archives/2016/08/26/most-scientific-results-are-wrong-or-use ▾
Aug 26, 2016 - ScientistYanlevDreamstime Yanlev/Dreamstime"**Science**, the pride of modernity, our one source of objective knowledge, is in deep trouble.

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]

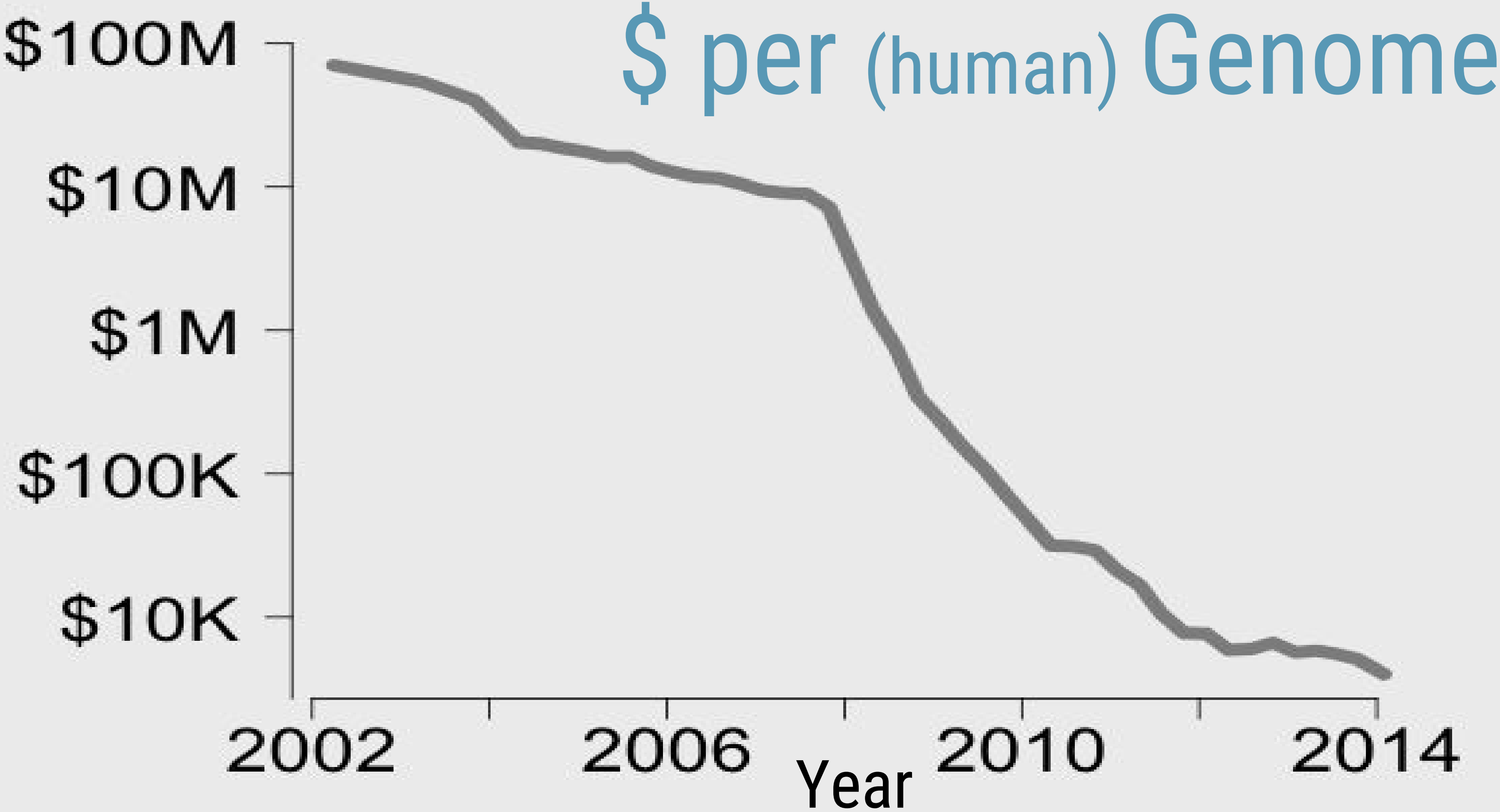
Contents [hide]

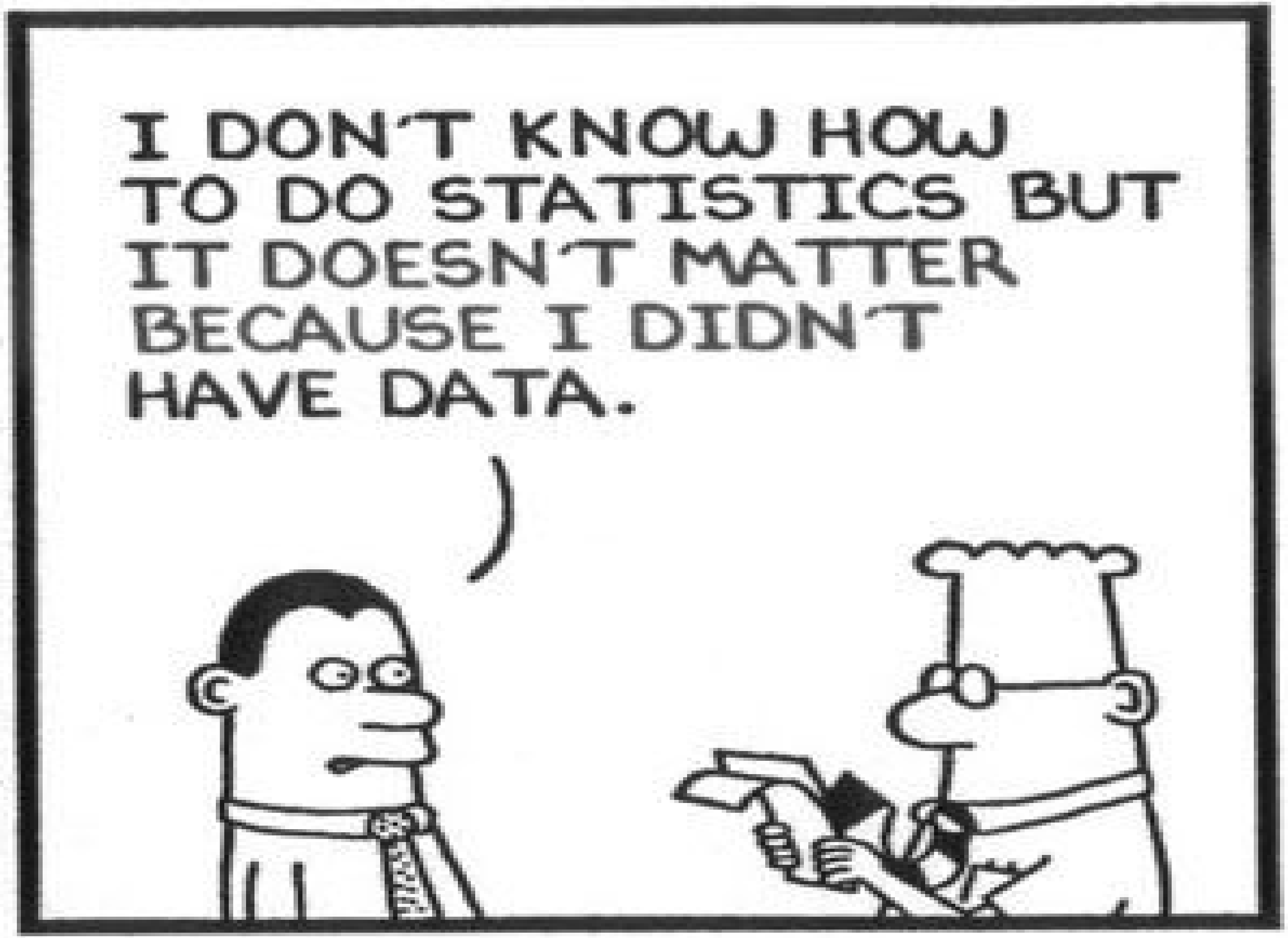
A hypothesis

N = SAMPLE SIZE

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

$ per (human) Genome




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

2

The tools to solve the “crisis” exist

The humans are the problem

2

The tools to solve the “crisis” exist

The humans are the problem

What is the “crisis”?

Population

Question

Hypothesis

Experimental Design

Experimentor

Data

Analysis Plan

Analyst

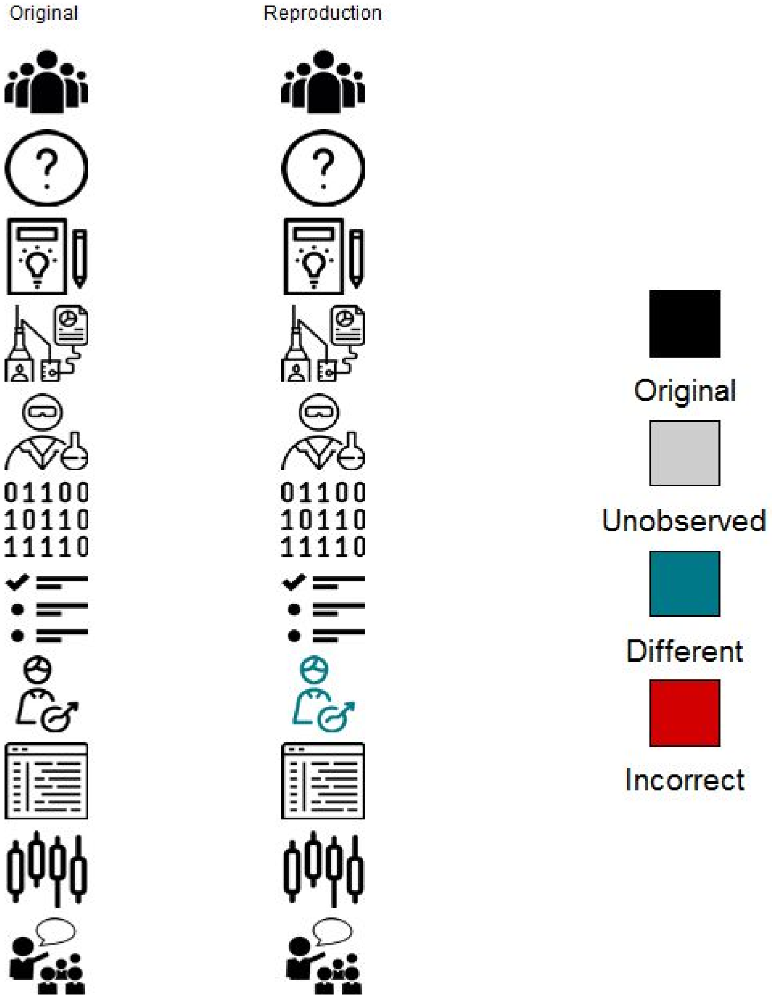
Code

Estimate

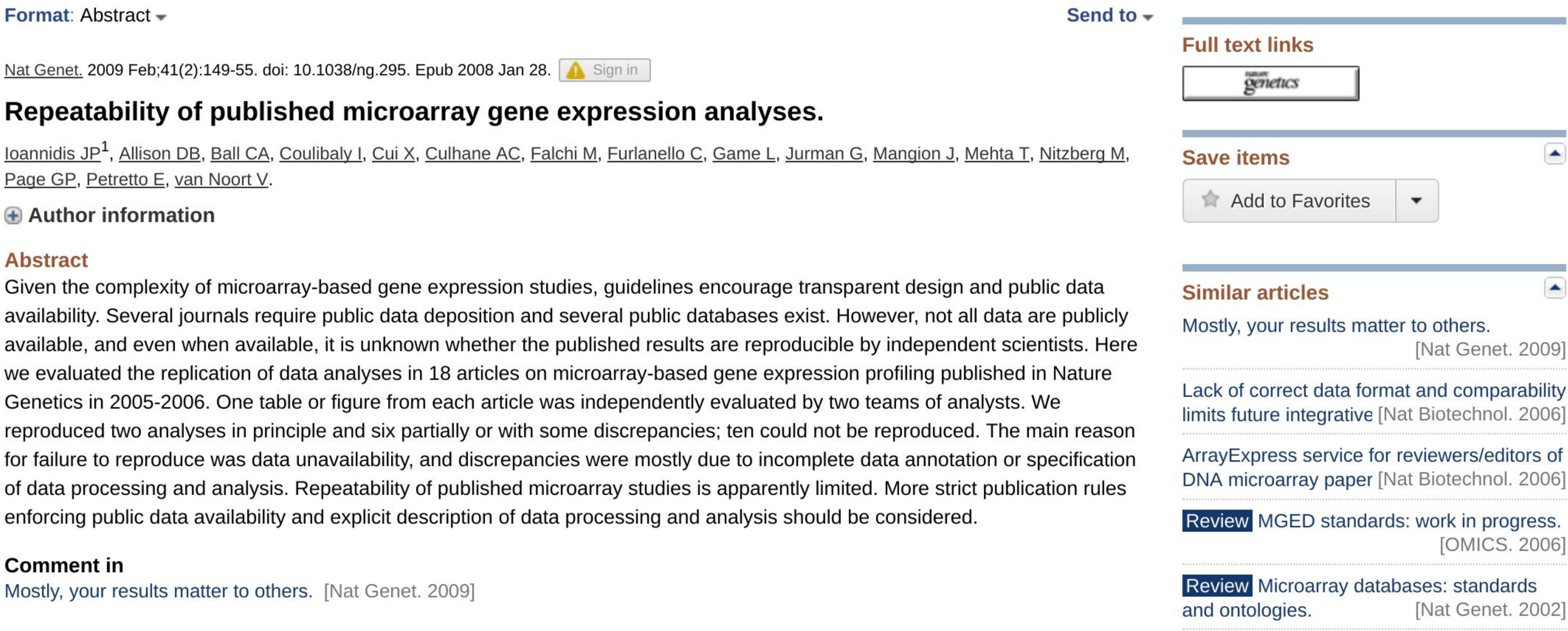
Claim

Patil, Peng and Leek biorXiv 2016

Reproduce



Patil, Peng and Leek biorXiv 2016



Nat Genet. 2009 Feb;41(2):149-55. doi: 10.1038/ng.295. Epub 2008 Jan 28.

Repeatability of published microarray gene expression analyses.

Ioannidis JP[1], Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, Mangion J, Mehta T, Nitzberg M, Page GP, Petretto E, van Noort V.

Author information

Abstract

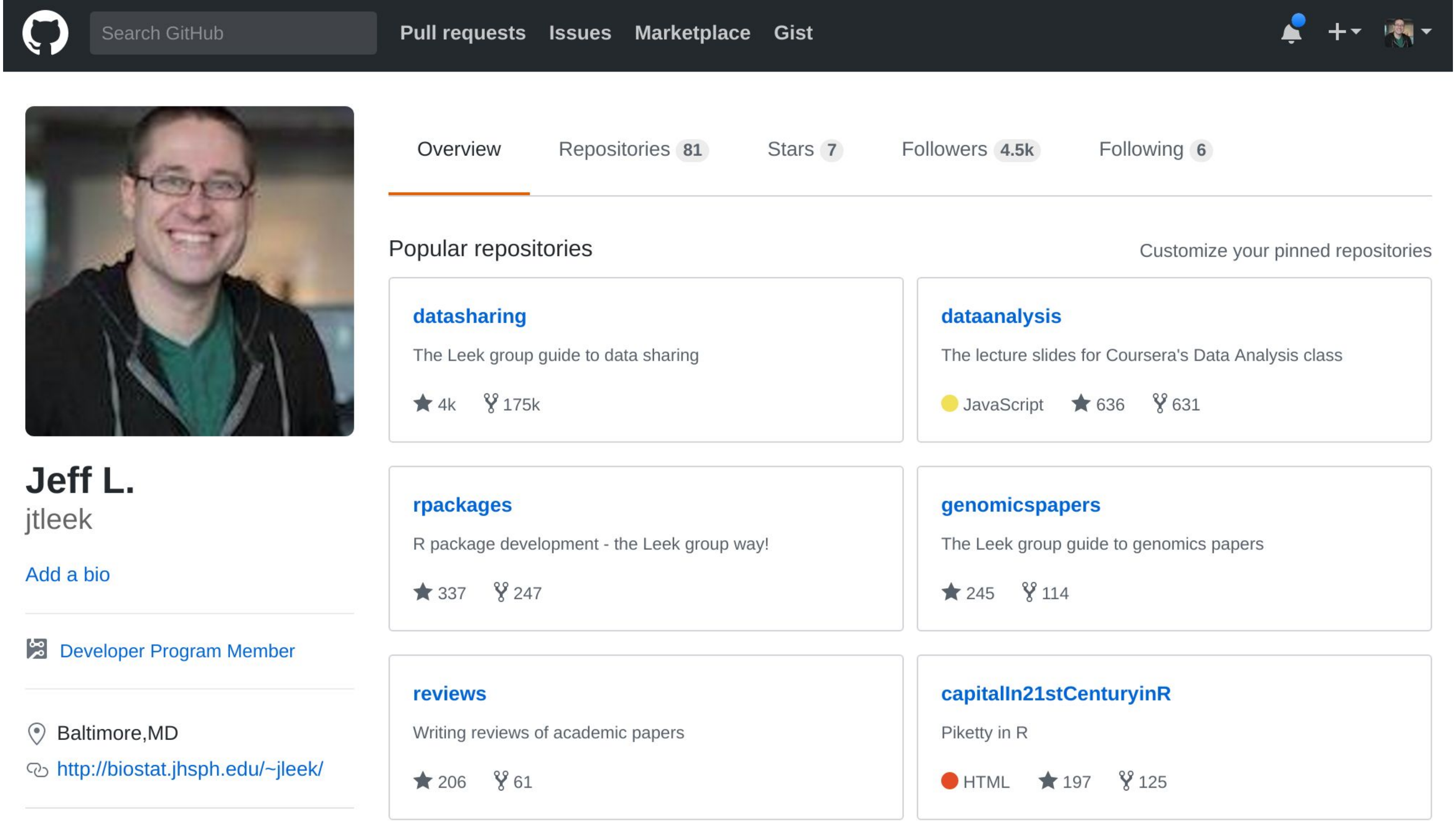
Given the complexity of microarray-based gene expression studies, guidelines encourage transparent design and public data availability. Several journals require public data deposition and several public databases exist. However, not all data are publicly available, and even when available, it is unknown whether the published results are reproducible by independent scientists. Here we evaluated the replication of data analyses in 18 articles on microarray-based gene expression profiling published in Nature Genetics in 2005-2006. One table or figure from each article was independently evaluated by two teams of analysts. We reproduced two analyses in principle and six partially or with some discrepancies; ten could not be reproduced. The main reason for failure to reproduce was data unavailability, and discrepancies were mostly due to incomplete data annotation or specification of data processing and analysis. Repeatability of published microarray studies is apparently limited. More strict publication rules enforcing public data availability and explicit description of data processing and analysis should be considered.

Comment in

Mostly, your results matter to others. [Nat Genet. 2009]

Similar articles

- Mostly, your results matter to others. [Nat Genet. 2009]
- Lack of correct data format and comparability limits future integrative [Nat Biotechnol. 2006]
- ArrayExpress service for reviewers/editors of DNA microarray paper [Nat Biotechnol. 2006]
- Review MGED standards: work in progress. [OMICS. 2006]
- Review Microarray databases: standards and ontologies. [Nat Genet. 2002]



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

Jeff L.

jtleek

Add a bio

Developer Program Member

Baltimore,MD

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

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

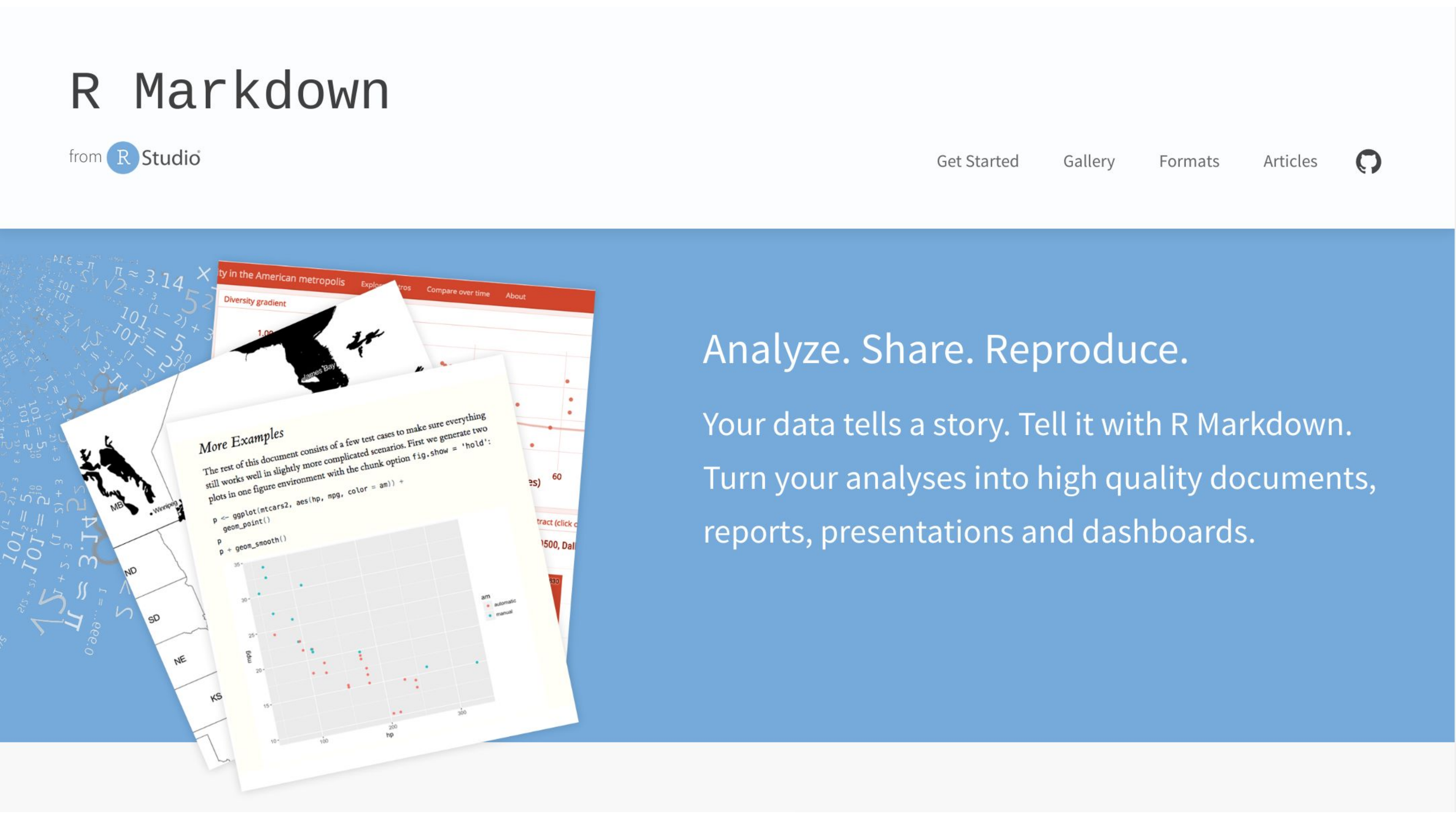
Writing reviews of academic papers

★ 206 ⑂ 61

capitalIn21stCenturyinR

Piketty in R

● HTML ★ 197 ⑂ 125



R Markdown

from R Studio

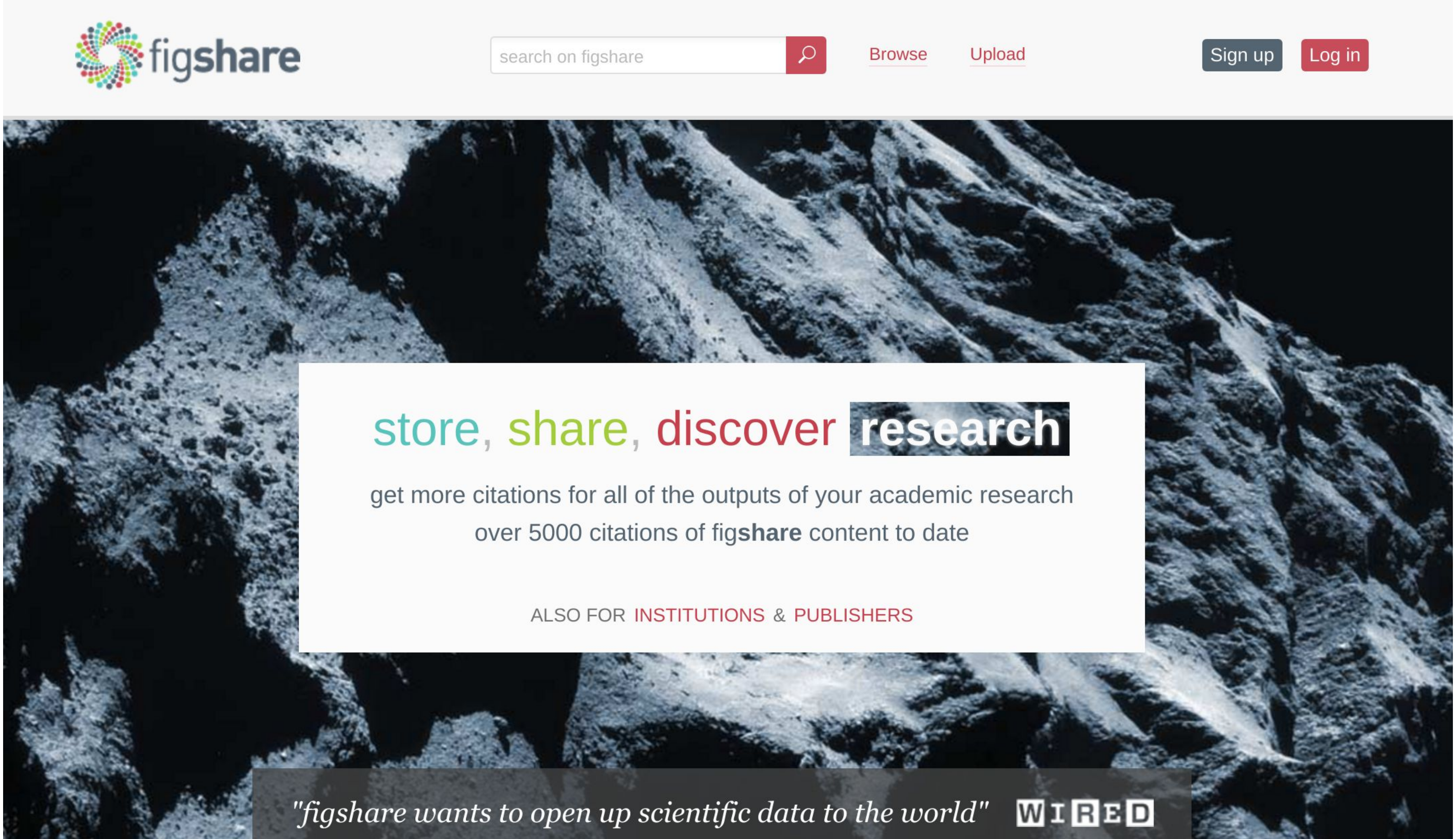
More Examples

The rest of this document consists of a few test cases to make sure everything still works well in slightly more complicated scenarios. First we generate two plots in one figure environment with the chunk option `fig.show = 'hold'`:

```
p <- ggplot(mtcars2, aes(hp, mpg, color = am)) +
  geom_point()
p
p + geom_smooth()
```

Analyze. Share. Reproduce.

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

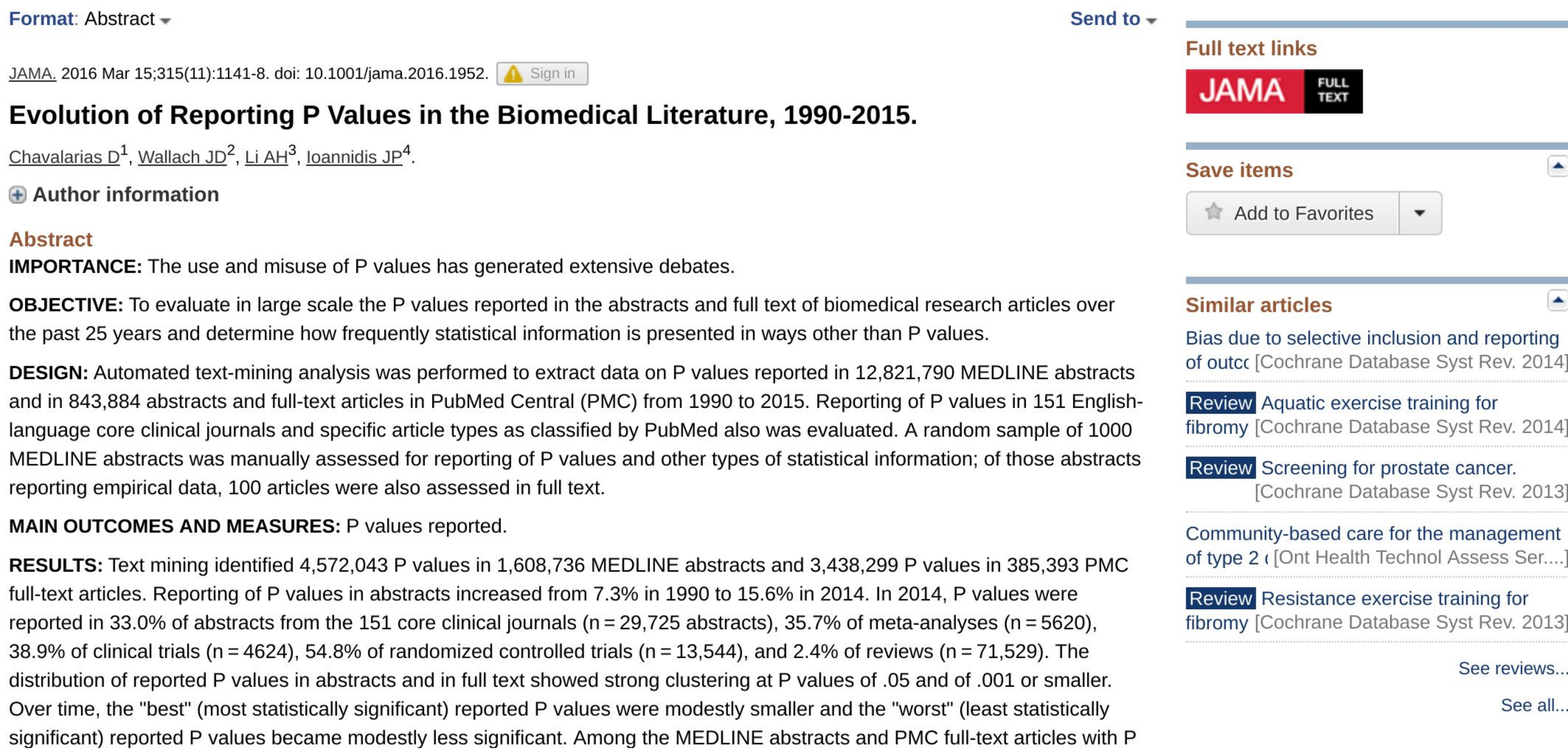


store, share, discover research

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



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

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

Similar articles

Bias due to selective inclusion and reporting of outc [Cochrane Database Syst Rev. 2014]

Review Aquatic exercise training for fibromy [Cochrane Database Syst Rev. 2014]

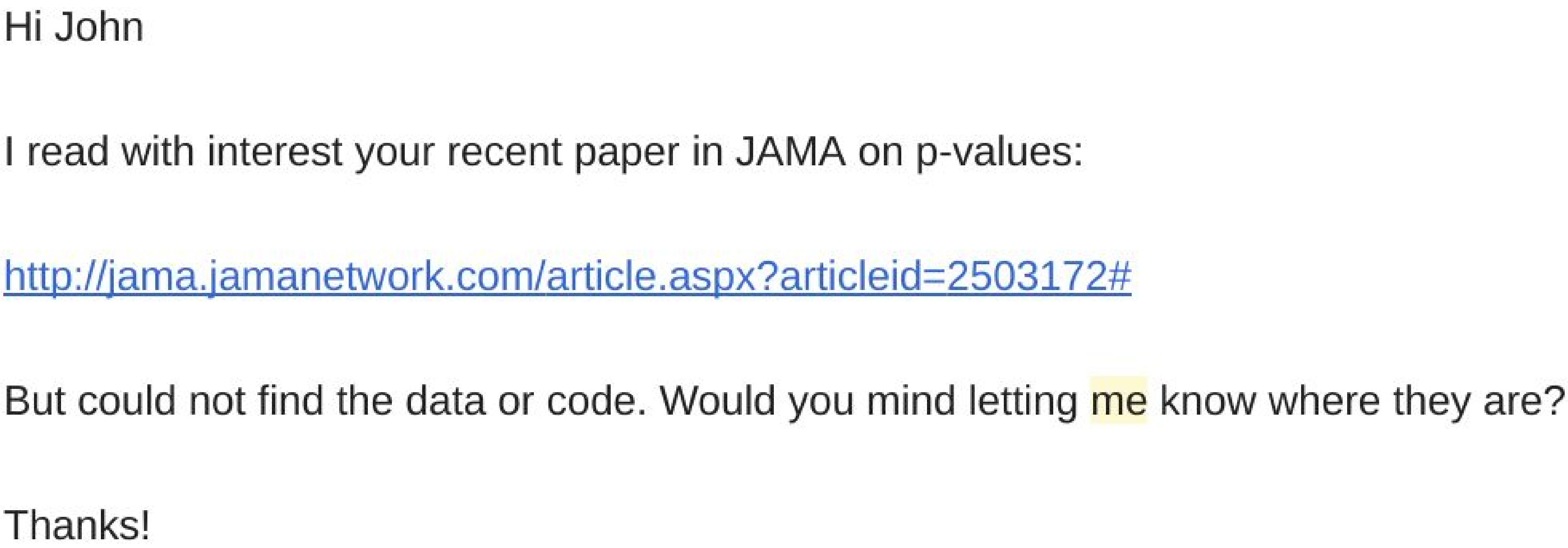
Review Screening for prostate cancer. [Cochrane Database Syst Rev. 2013]

Community-based care for the management of type 2 ([Ont Health Technol Assess Ser....]

Review Resistance exercise training for fibromy [Cochrane Database Syst Rev. 2013]

See reviews...

See all...



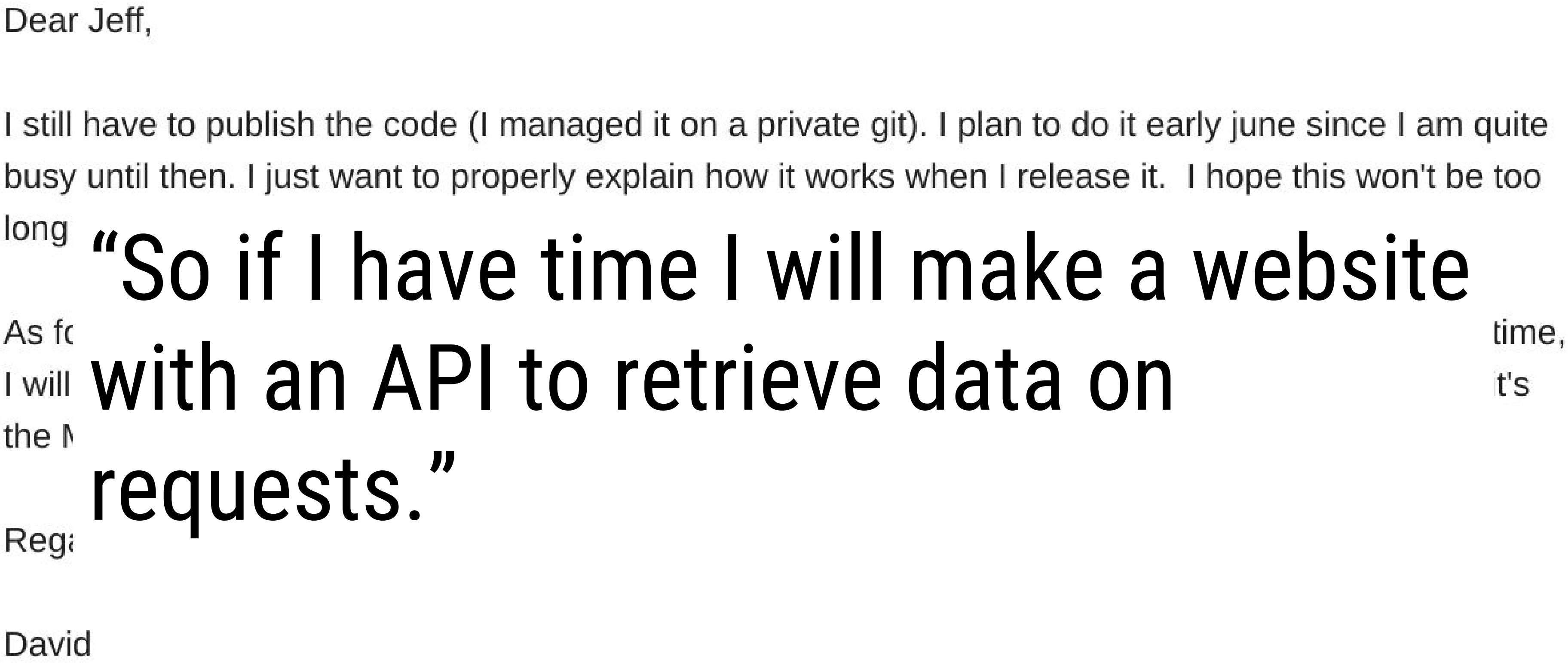
Hi John

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,

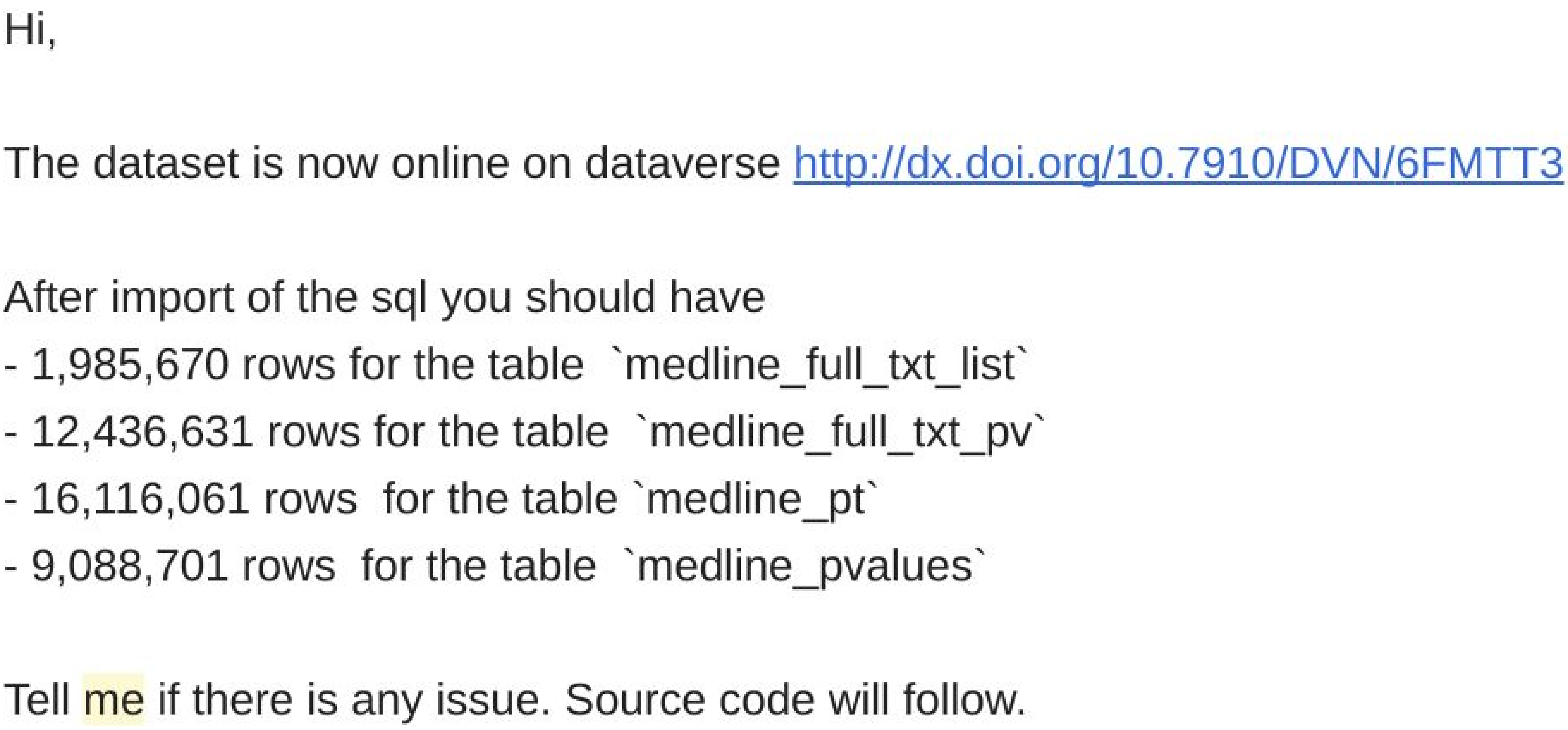
I still have to publish the code (I managed it on a private git). I plan to do it early june since I am quite busy until then. I just want to properly explain how it works when I release it. I hope this won't be too long

As fc time,
I will t's
the N

Rega

David

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

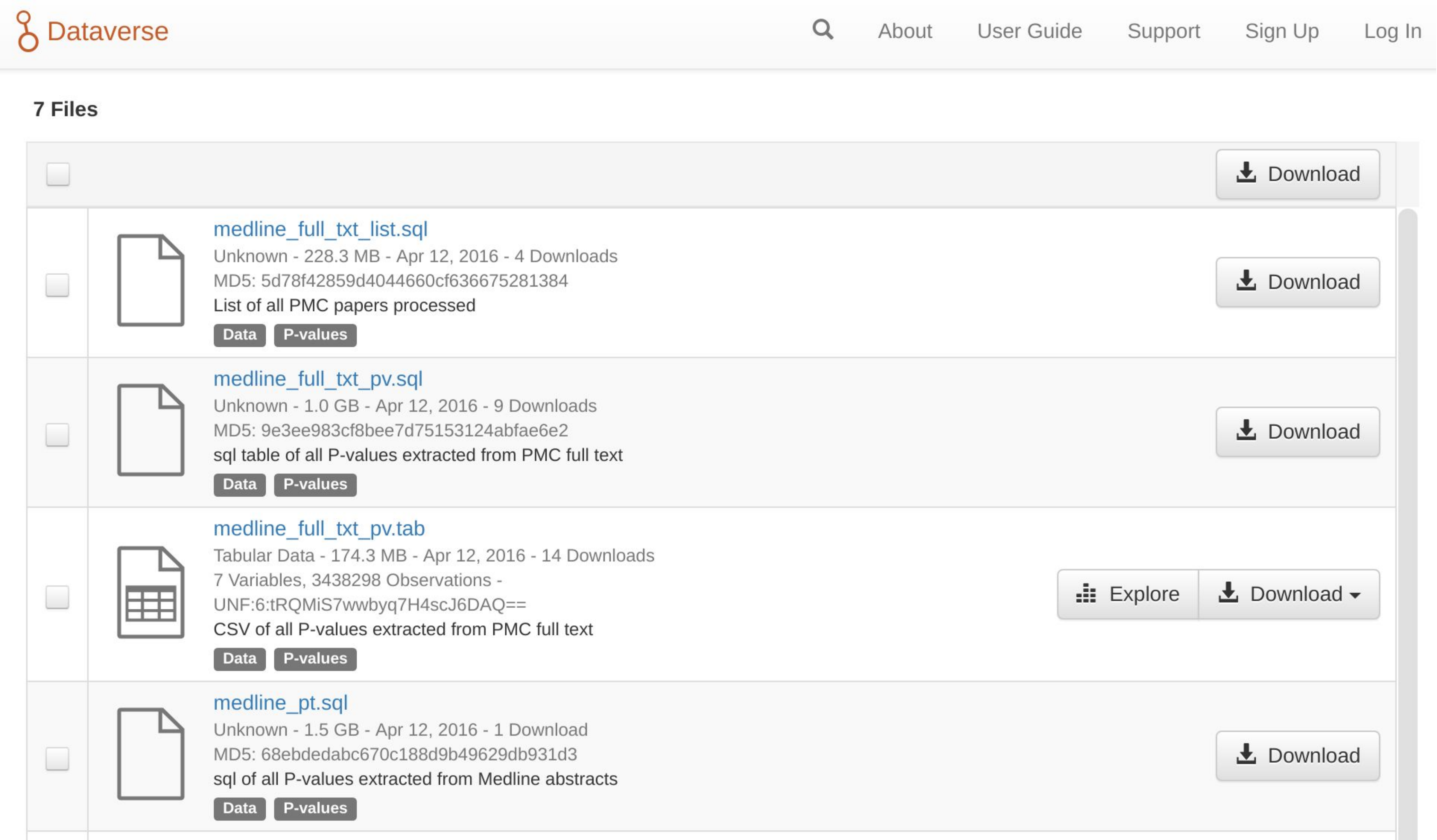


Hi,

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.



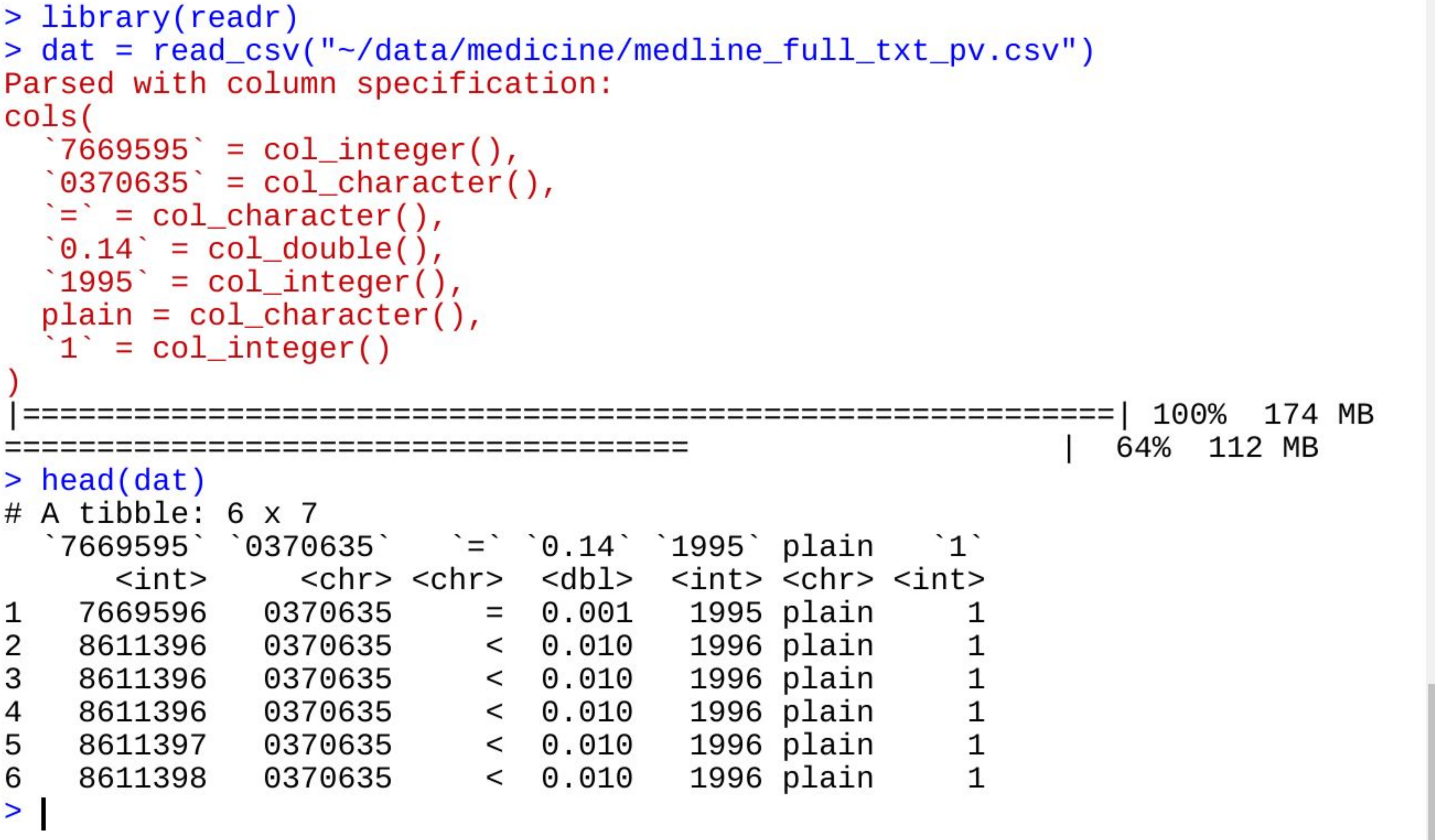
7 Files

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

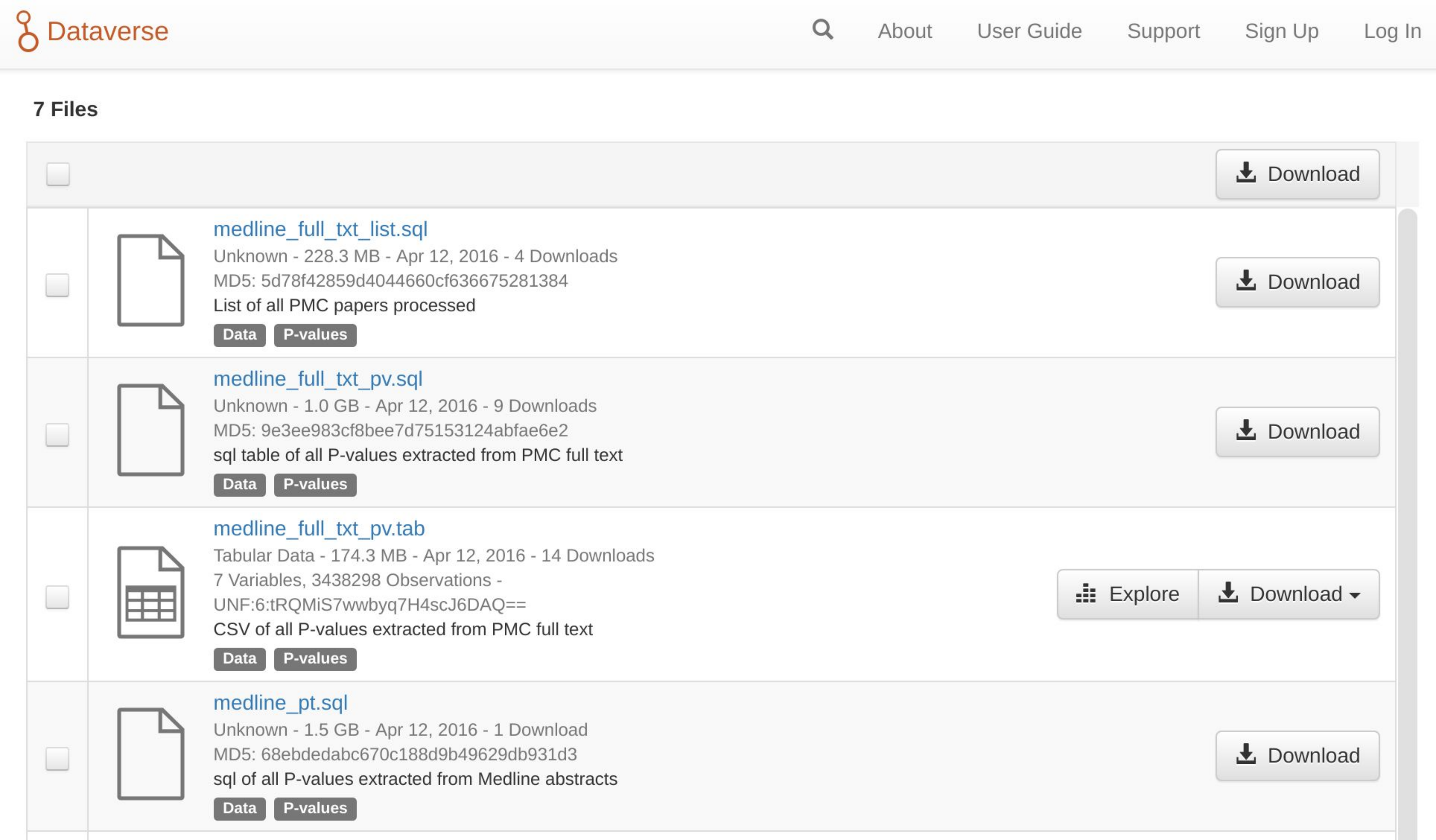
medline_full_txt_pv.sql
Unknown - 1.0 GB - Apr 12, 2016 - 9 Downloads
MD5: 9e3ee983cf8bee7d75153124abfae6e2
sql table of all P-values extracted from PMC full text
Data P-values

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

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



```
> 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

□

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

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

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

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

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

Jeff Leek

26 July 2017

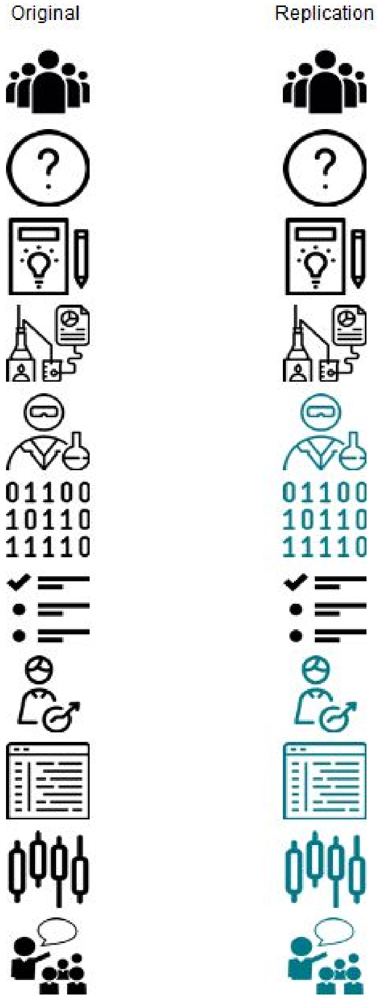
Contents

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

Replicate



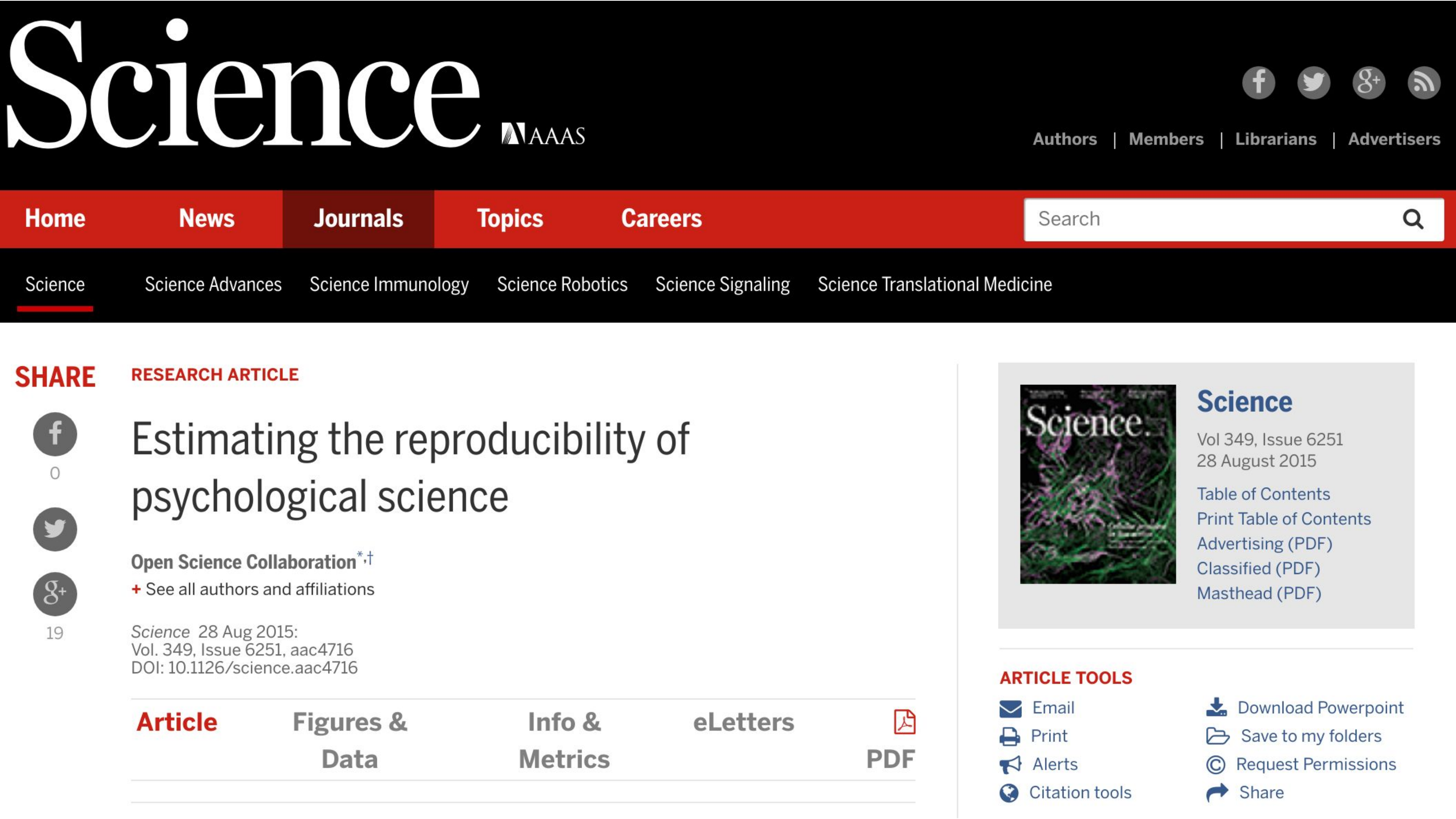
⬛ Original

⬜ Unobserved

🟦 Different

🟥 Incorrect

Patil, Peng and Leek biorXiv 2016



Science AAAS
Authors | Members | Librarians | Advertisers
Home News Journals Topics Careers
Search
Science Science Advances Science Immunology Science Robotics Science Signaling Science Translational Medicine
SHARE
0
19

RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration*,†
+ See all authors and affiliations

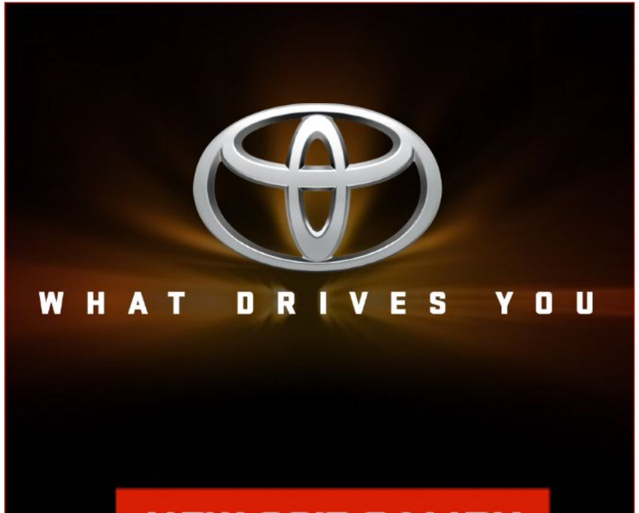
Science 28 Aug 2015:
Vol. 349, Issue 6251, aac4716
DOI: 10.1126/science.aac4716

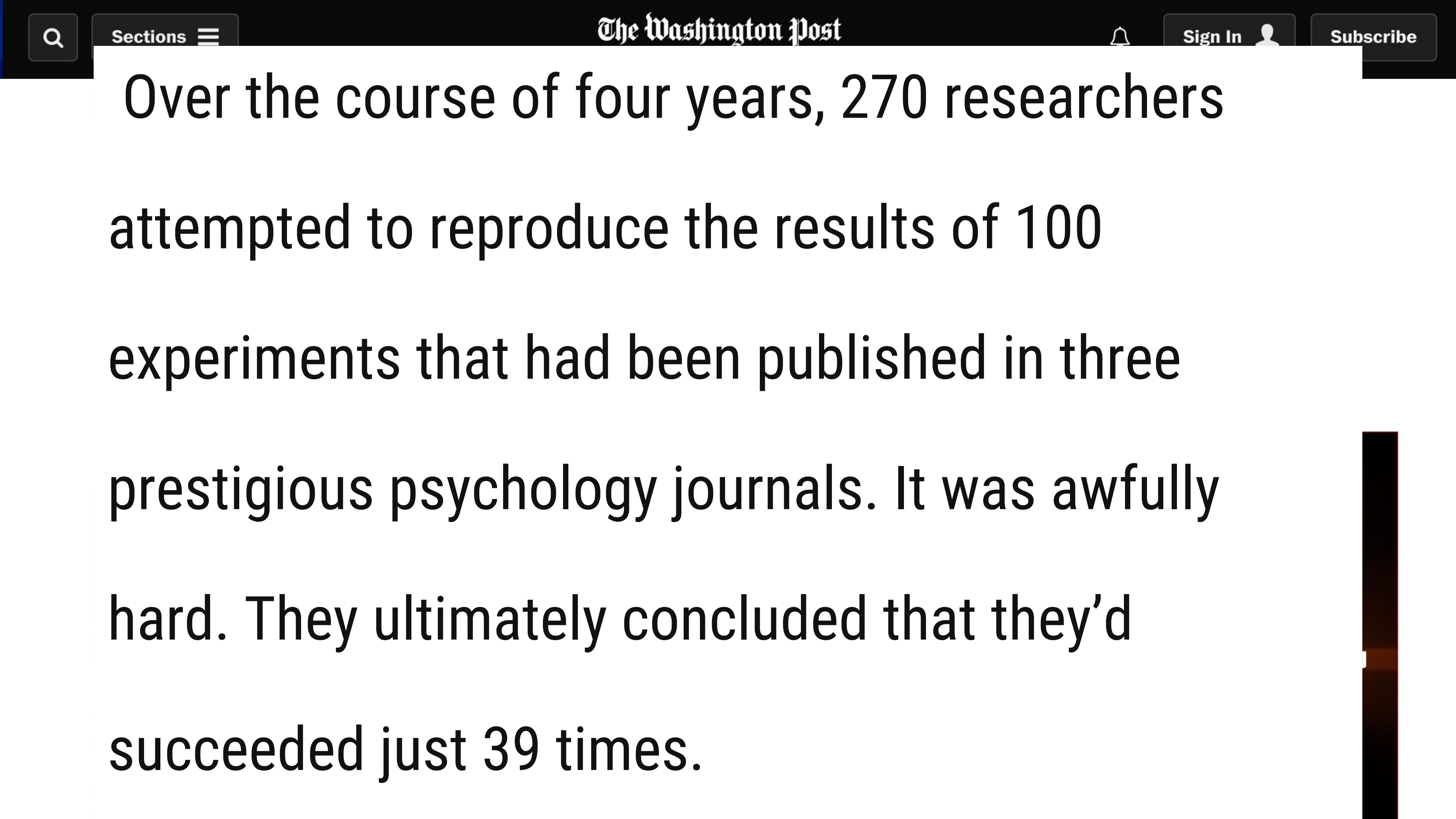
Science
Vol 349, Issue 6251
28 August 2015

Speaking of Science

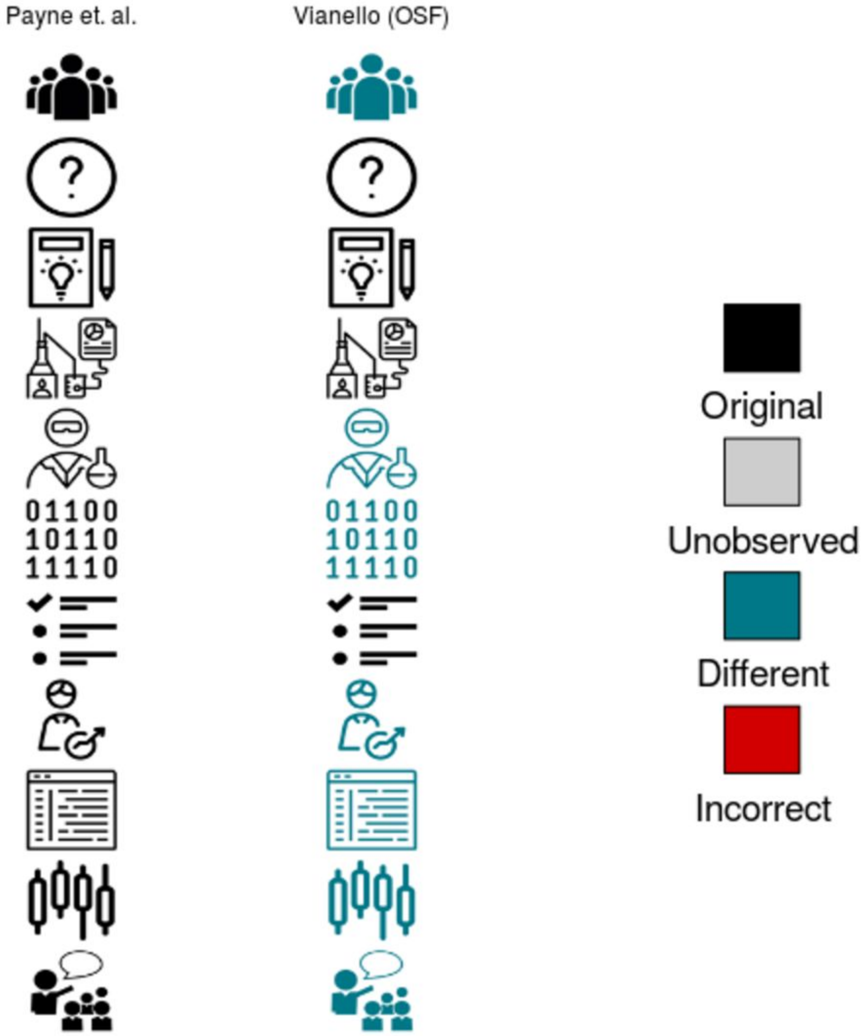
Many scientific studies can’t be replicated. That’s a problem.

By Joel Achenbach August 27, 2015





Over the course of four years, 270 researchers attempted to reproduce 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.



Payne et. al.
Vianello (OSF)
01100
10110
11110
01100
10110
11110
Original
Unobserved
Different
Incorrect

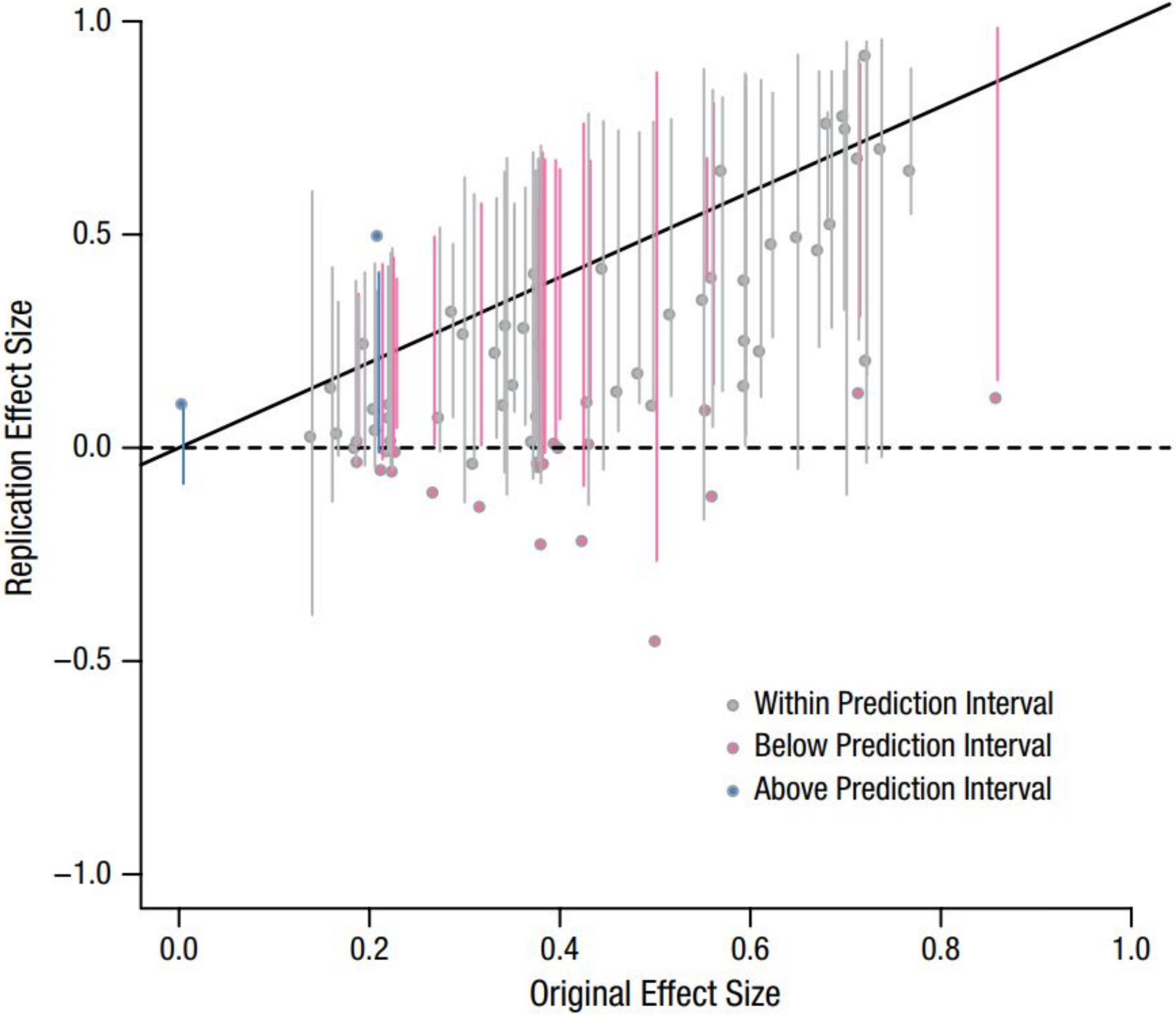
Replication Definition for 39

$P < 0.05$ in Original

$P < 0.05$ in Replicated Study

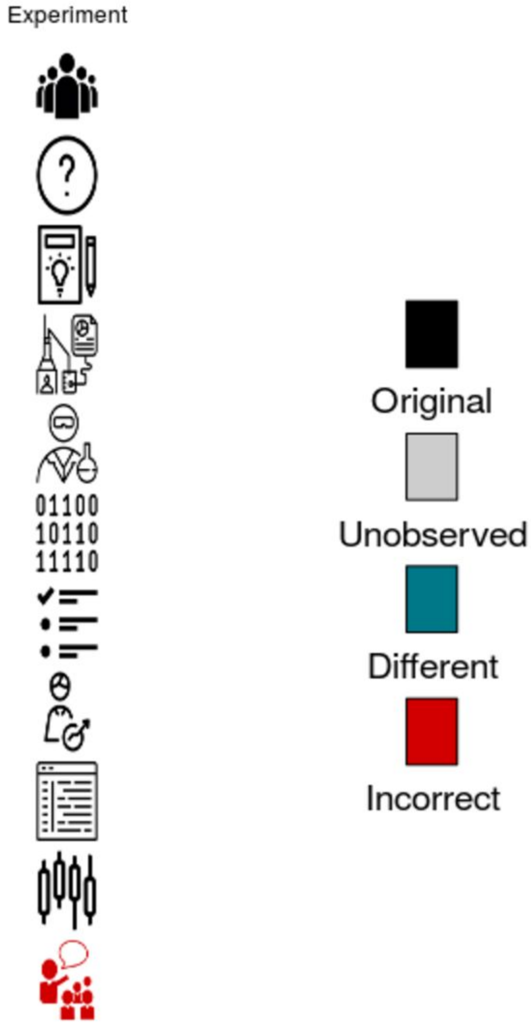
Alternative Definition

Effect size inside prediction interval for effect based on original study

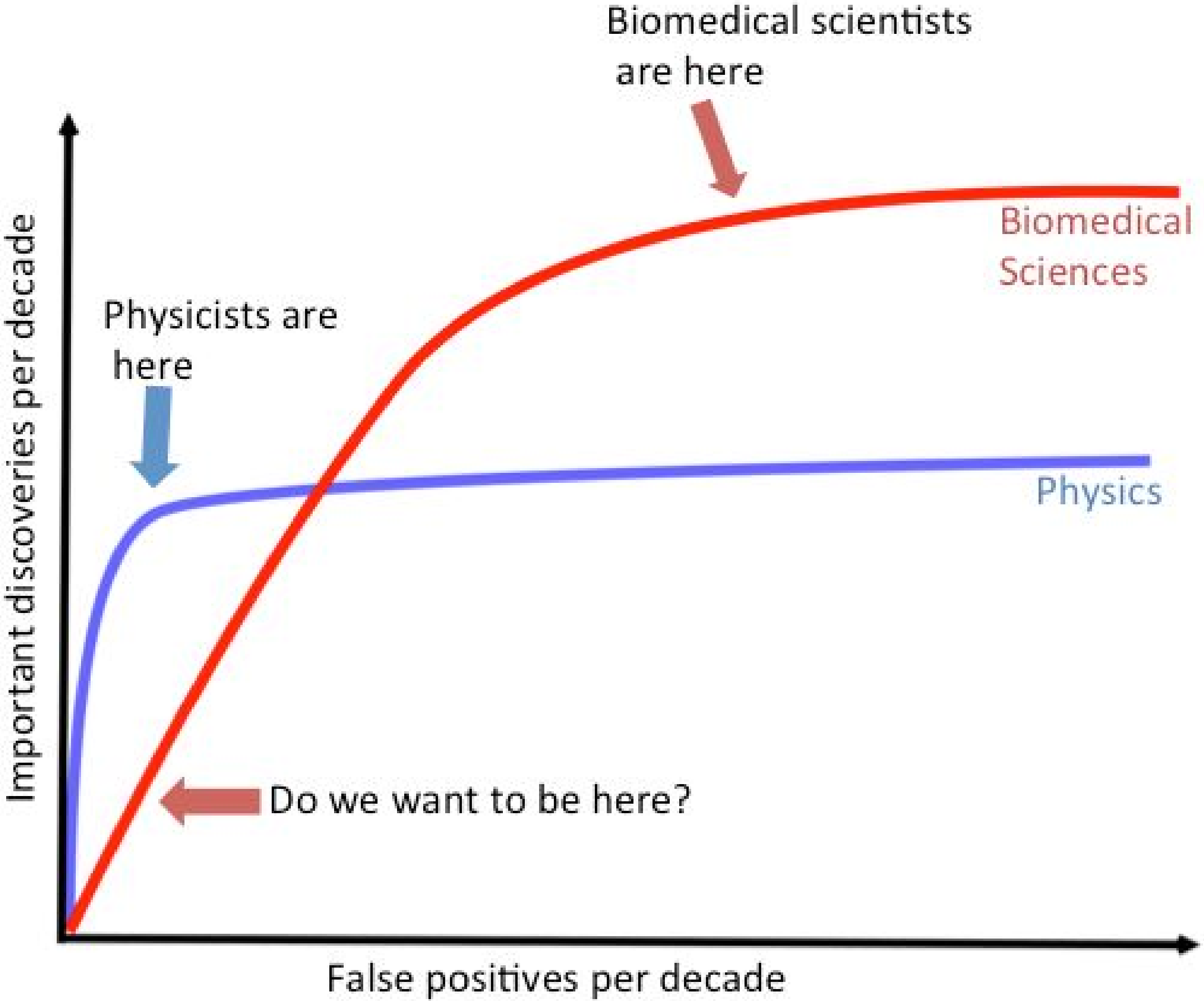


1.0
0.5
0.0
−0.5
−1.0
Replication Effect Size
0.0
0.2
0.4
0.6
0.8
1.0
Original Effect Size
● Within Prediction Interval
● Below Prediction Interval
● Above Prediction Interval

False discovery



Patil, Peng and Leek biorXiv 2016

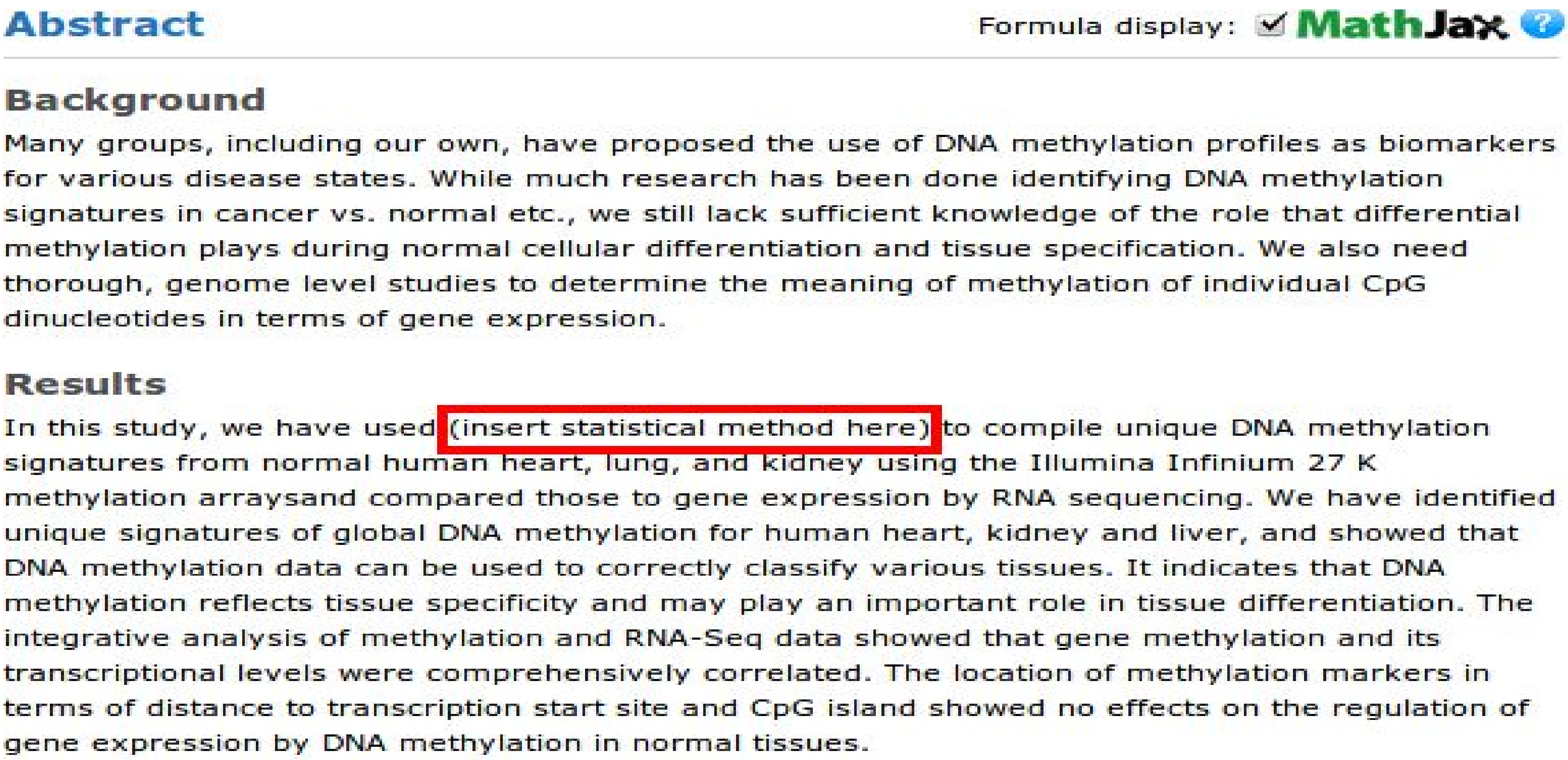

Biomedical scientists are here
Biomedical Sciences
Physicists are here
Physics
Do we want to be here?
Important discoveries per decade
False positives per decade

2

The tools to solve the “crisis” exist

The humans are the problem

How many people feel about statistics



Abstract

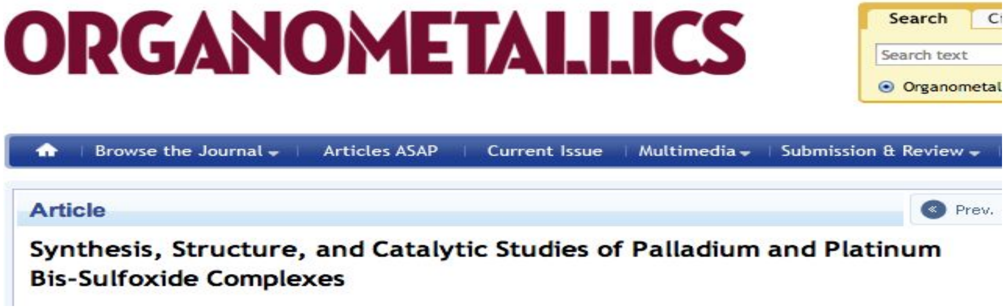
Background

Many groups, including our own, have proposed the use of DNA methylation profiles as biomarkers for various disease states. While much research has been done identifying DNA methylation signatures in cancer vs. normal etc., we still lack sufficient knowledge of the role that differential methylation plays during normal cellular differentiation and tissue specification. We also need thorough, genome level studies to determine the meaning of methylation of individual CpG dinucleotides in terms of gene expression.

Results

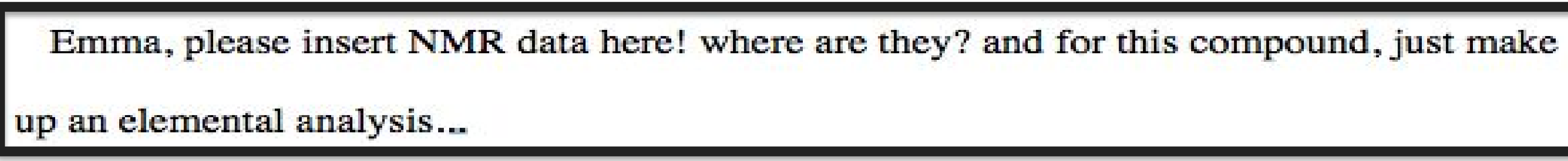
In this study, we have used (insert statistical method here) to compile unique DNA methylation signatures from normal human heart, lung, and kidney using the Illumina Infinium 27 K methylation arraysand compared those to gene expression by RNA sequencing. We have identified unique signatures of global DNA methylation for human heart, kidney and liver, and showed that DNA methylation data can be used to correctly classify various tissues. It indicates that DNA methylation reflects tissue specificity and may play an important role in tissue differentiation. The integrative analysis of methylation and RNA-Seq data showed that gene methylation and its transcriptional levels were comprehensively correlated. The location of methylation markers in terms of distance to transcription start site and CpG island showed no effects on the regulation of gene expression by DNA methylation in normal tissues.

http://bit.ly/OgW3xv



Article

Synthesis, Structure, and Catalytic Studies of Palladium and Platinum Bis-Sulfoxide Complexes



Emma, please insert NMR data here! where are they? and for this compound, just make up an elemental analysis...

Drinkel et al. *Oganometalics* 2013

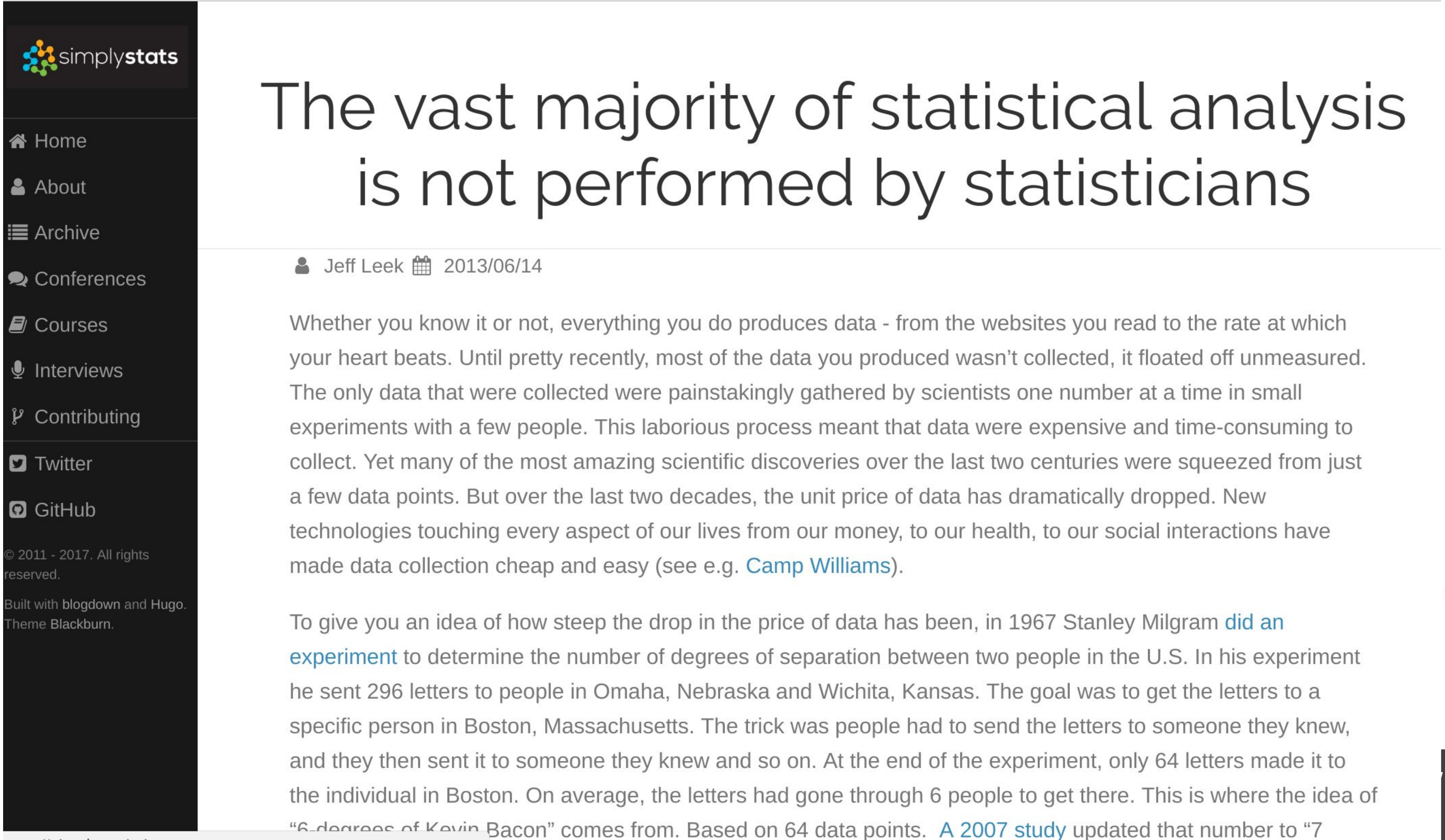
Medical school entrance requirements (U.S.)

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
© 2011 - 2017. All rights reserved.
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

Y = some outcome

X = some covariate

D = (**X**,**Y**)

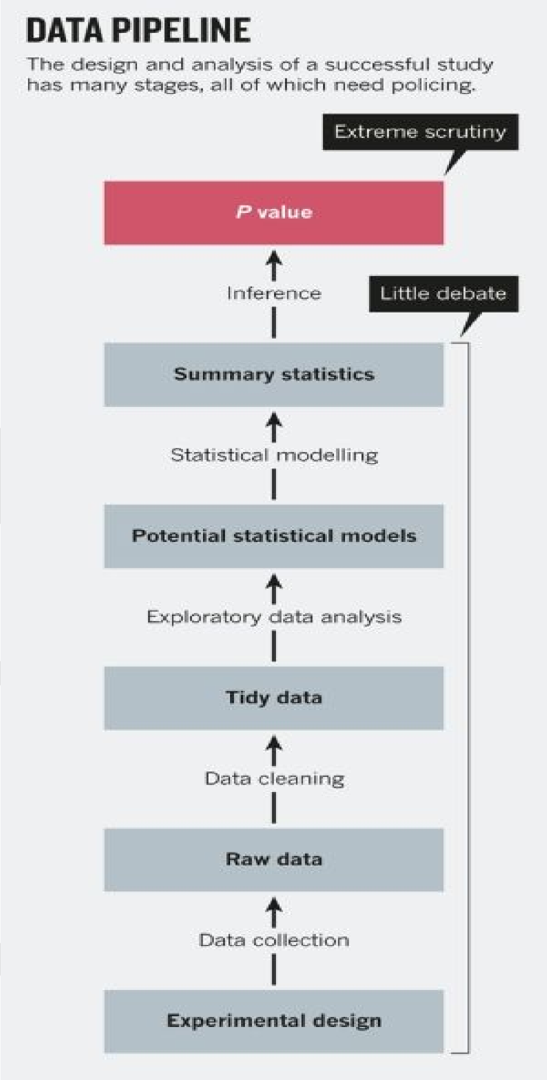
lm(**Y** ~ **X**)

Y = som come

X = som ariate

D = (**X**,**Y**

lm(



DATA PIPELINE

The design and analysis of a successful study has many stages, all of which need policing.

Extreme scrutiny

P value

Inference

Little debate

Summary statistics

Statistical modelling

Potential statistical models

Exploratory data analysis

Tidy data

Data cleaning

Raw data

Data collection

Experimental design

Leek and Peng, Nature 2015

Population

Question

Hypothesis

Experimental Design

Experimentor

Data

Analysis Plan

Analyst

Code

Estimate

Claim

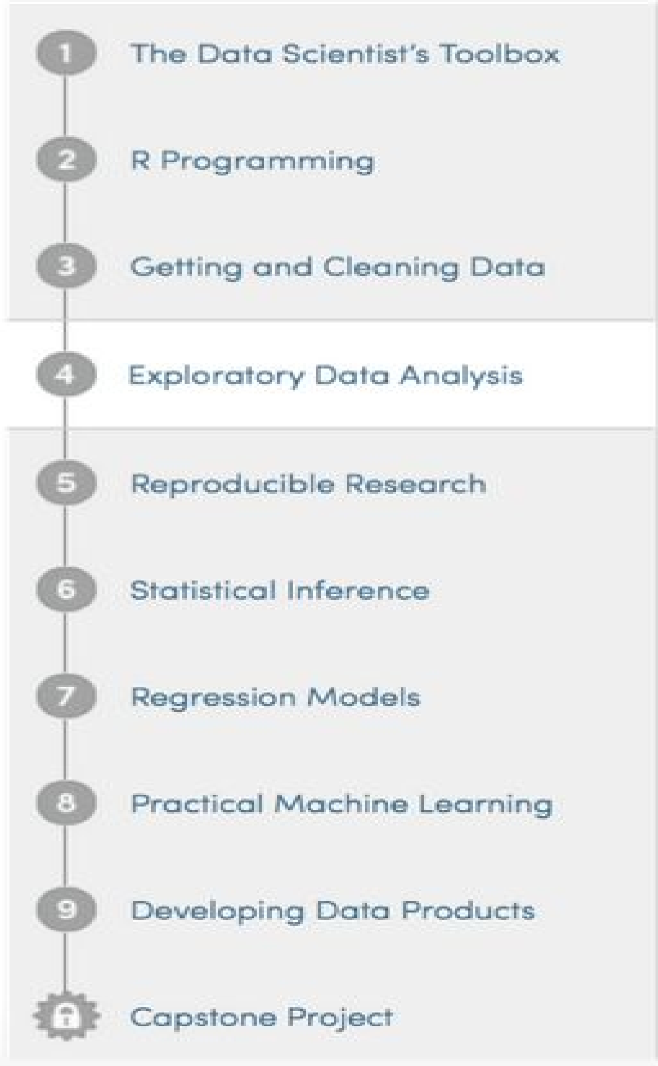
“Statistics”

Patil, Peng and Leek biorXiv 2016

9 classes

1 month long

Always open



1. The Data Scientist's Toolbox
2. R Programming
3. Getting and Cleaning Data
4. Exploratory Data Analysis
5. Reproducible Research
6. Statistical Inference
7. Regression Models
8. Practical Machine Learning
9. Developing Data Products

Capstone Project

The core problem

Who?

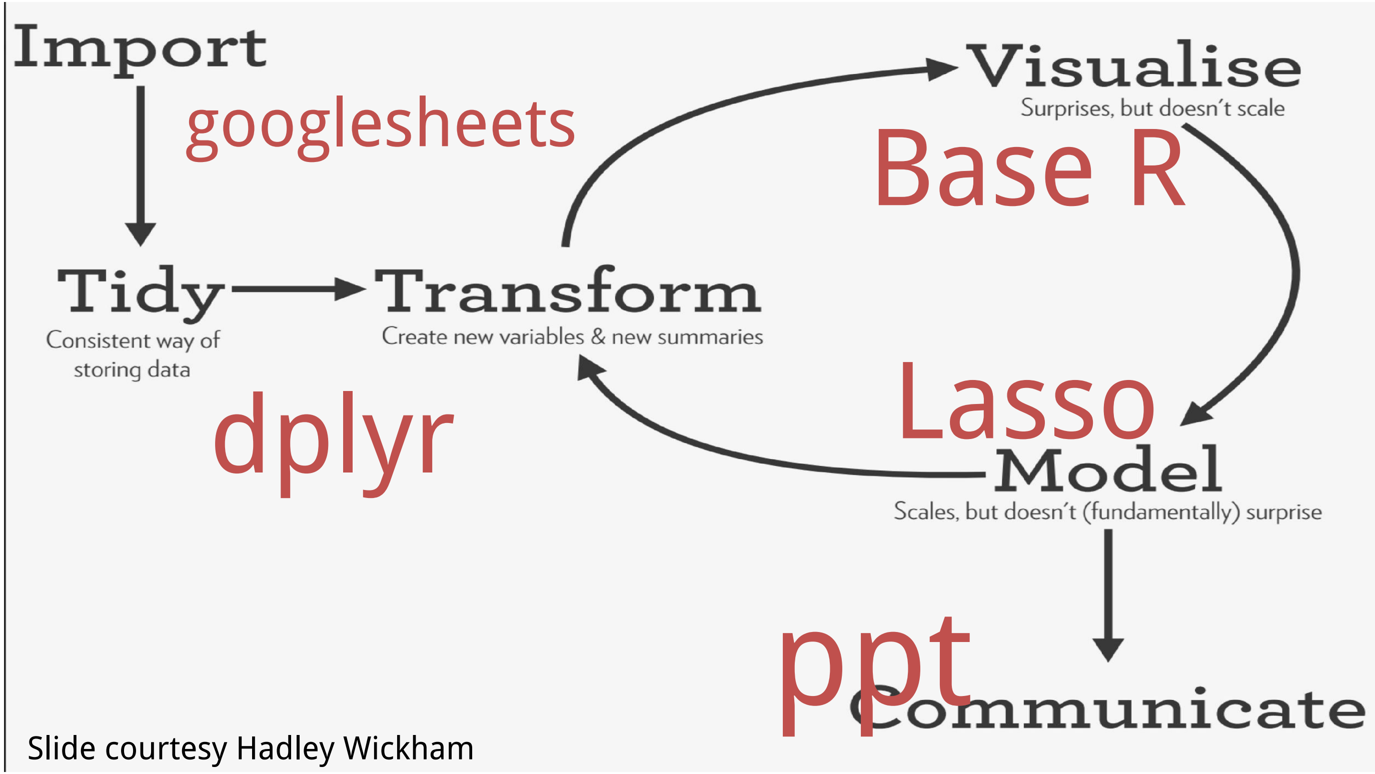
What?

When?

Why?

Where?

How?



Import

googlesheets

Visualise

Surprises, but doesn't scale

Base R

Tidy

Consistent way of storing data

Transform

Create new variables & new summaries

dplyr

Lasso

Model

Scales, but doesn't (fundamentally) surprise

ppt

Communicate

Slide courtesy Hadley Wickham

Who?

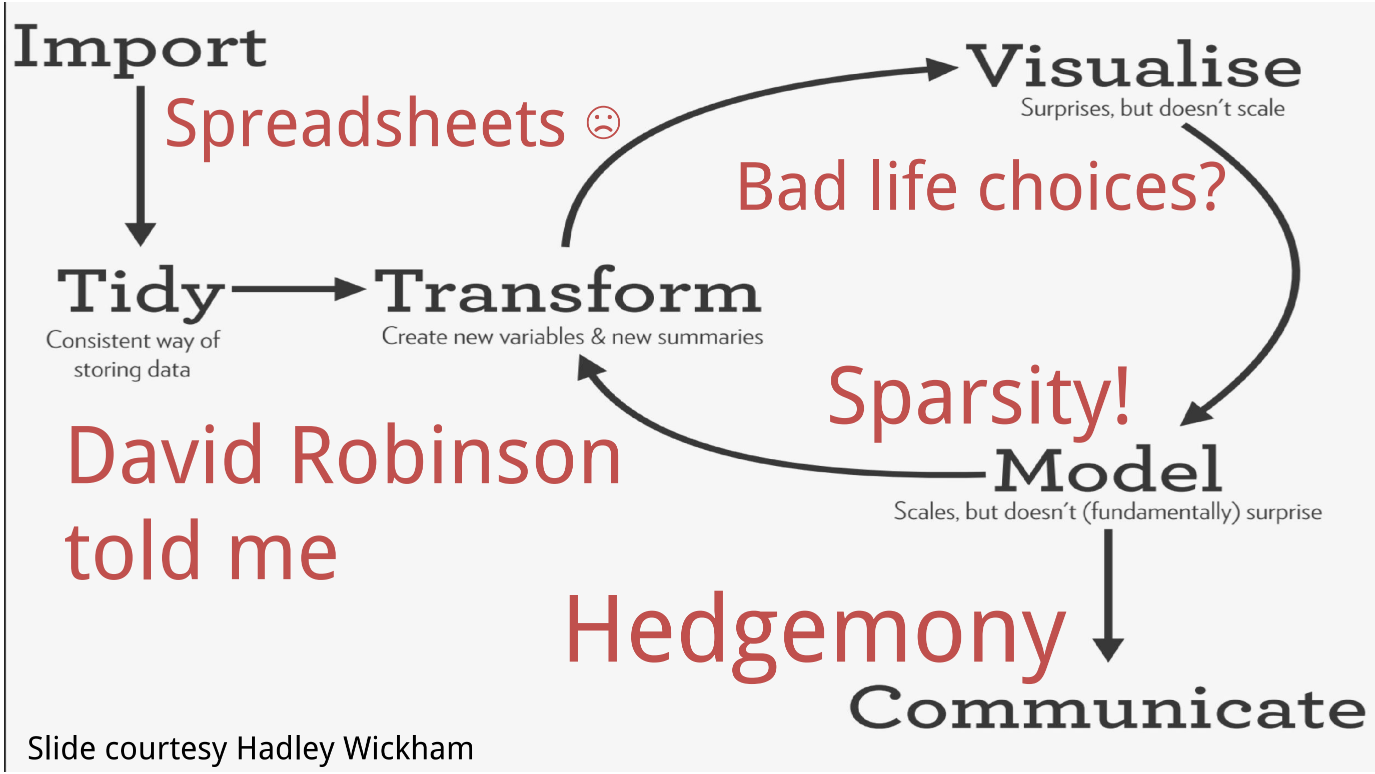
What?

When?

Why?

Where?

How?



Import

Spreadsheets ☹

Visualise

Surprises, but doesn't scale

Bad life choices?

Tidy

Consistent way of storing data

Transform

Create new variables & new summaries

Sparsity!

David Robinson told me

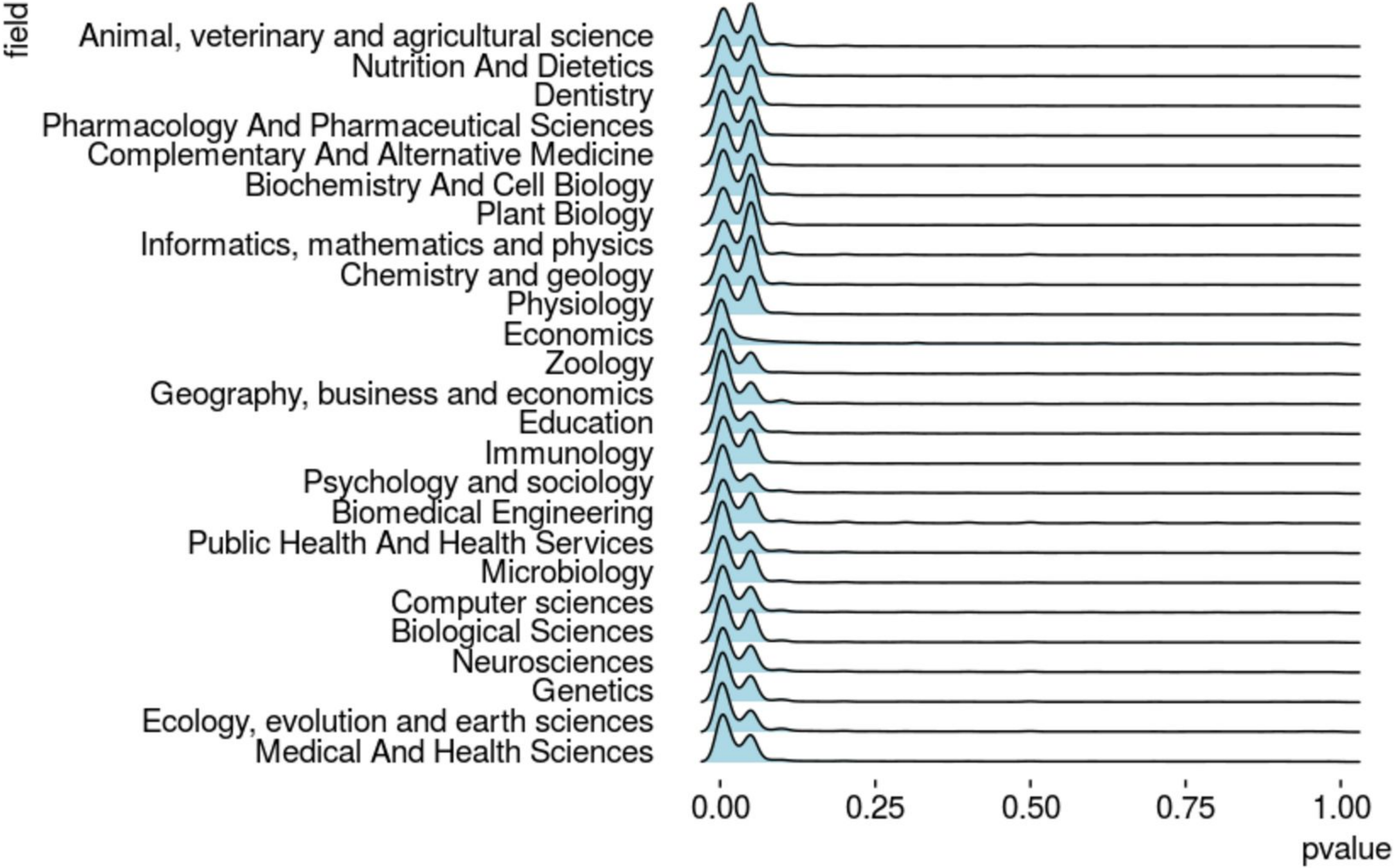
Model

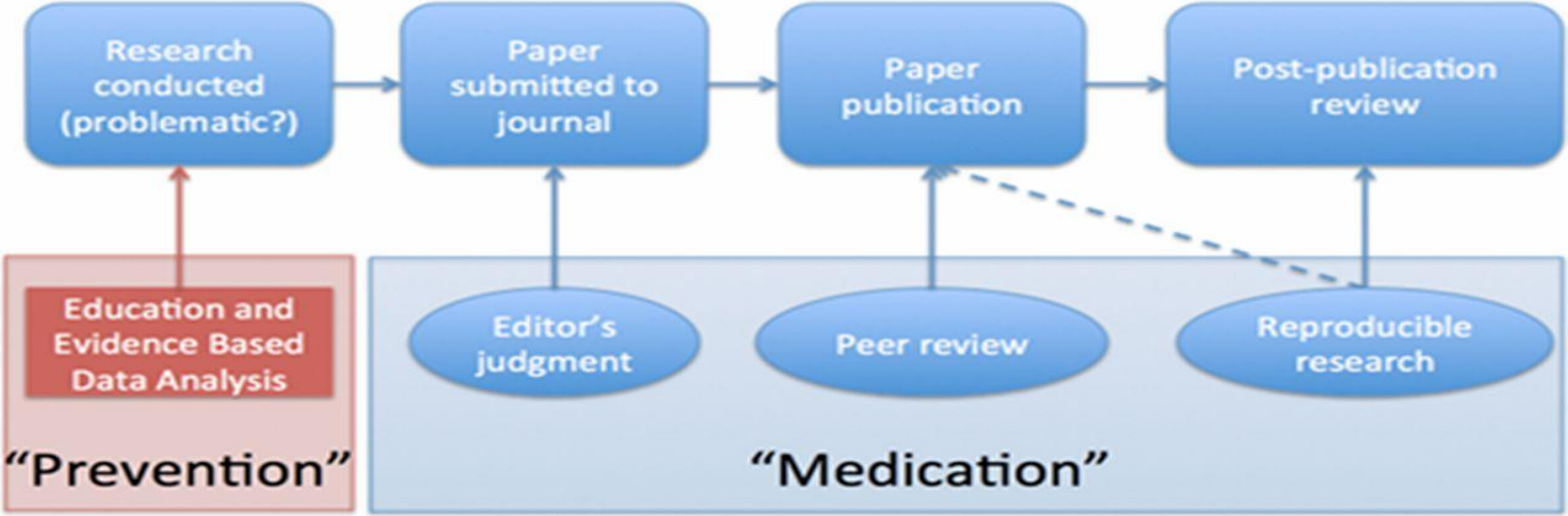
Scales, but doesn't (fundamentally) surprise

Hedgemony

Communicate

Slide courtesy Hadley Wickham


field
Animal, veterinary and agricultural science
Nutrition And Dietetics
Dentistry
Pharmacology And Pharmaceutical Sciences
Complementary And Alternative Medicine
Biochemistry And Cell Biology
Plant Biology
Informatics, mathematics and physics
Chemistry and geology
Physiology
Economics
Zoology
Geography, business and economics
Education
Immunology
Psychology and sociology
Biomedical Engineering
Public Health And Health Services
Microbiology
Computer sciences
Biological Sciences
Neurosciences
Genetics
Ecology, evolution and earth sciences
Medical And Health Sciences
0.00
0.25
0.50
0.75
1.00
pvalue



Leek & Peng 2015 PNAS

We take a random sample of individuals in a population and identify whether they smoke and if they have cancer. We observe that there is a strong relationship between whether a person in the sample smoked or whether they have lung cancer. We claim that smoking is related to lung cancer in the larger population.

79%

Inferential

vs

17%

Causal

n=47,141

We take a random sample of individuals in a population and identify whether they smoke and if they have cancer. We observe that there is a strong relationship between whether a person in the sample smoked or whether they have lung cancer. We claim that smoking is related to lung cancer in the larger population. We explain we think that the reason for this relationship is because cigarette smoke contains known carcinogens

65% vs 32 %

Inferential Causal

n=47,141

HOW SHOULD I ANALYZE DATA?
USE DATA!
imgflip.com

Data science as a Science

The Leek group

- Claire Ruberman
- Jack Fu
- Divya Narayanan
- Shannon Ellis
- Sean Kross
- **Leslie Myint**

Collaborators

- Andrew Jaffe
- Kasper Hansen
- Margaret Taub
- **Leah Jager**
- **Roger Peng**
- Ben Langmead
- Abhi Nellore
- Kai Kammers
- Leo Collado Torres
- **Prasad Patil**

jtleek.com/talks

@jtleek

@simplystats