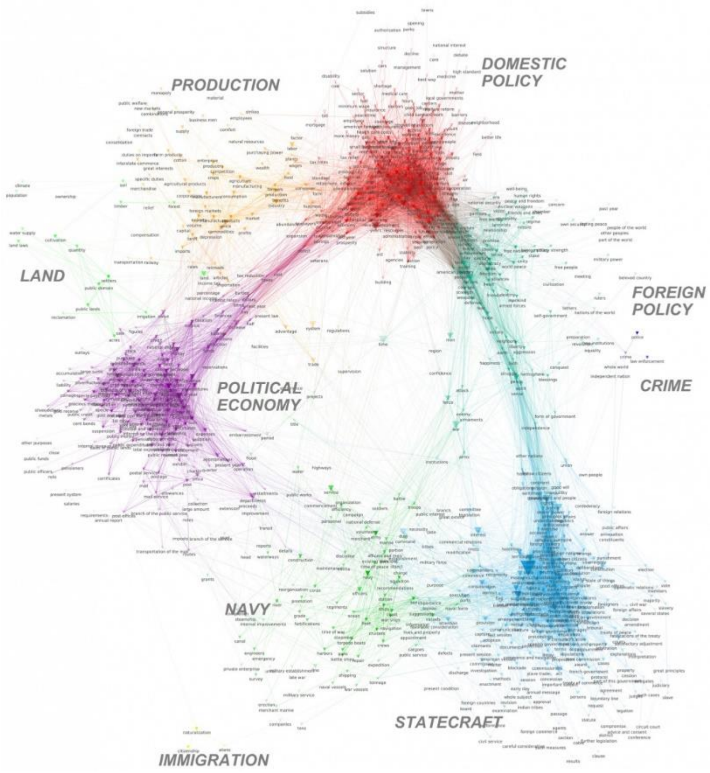
Text Mining & Reproducible Research

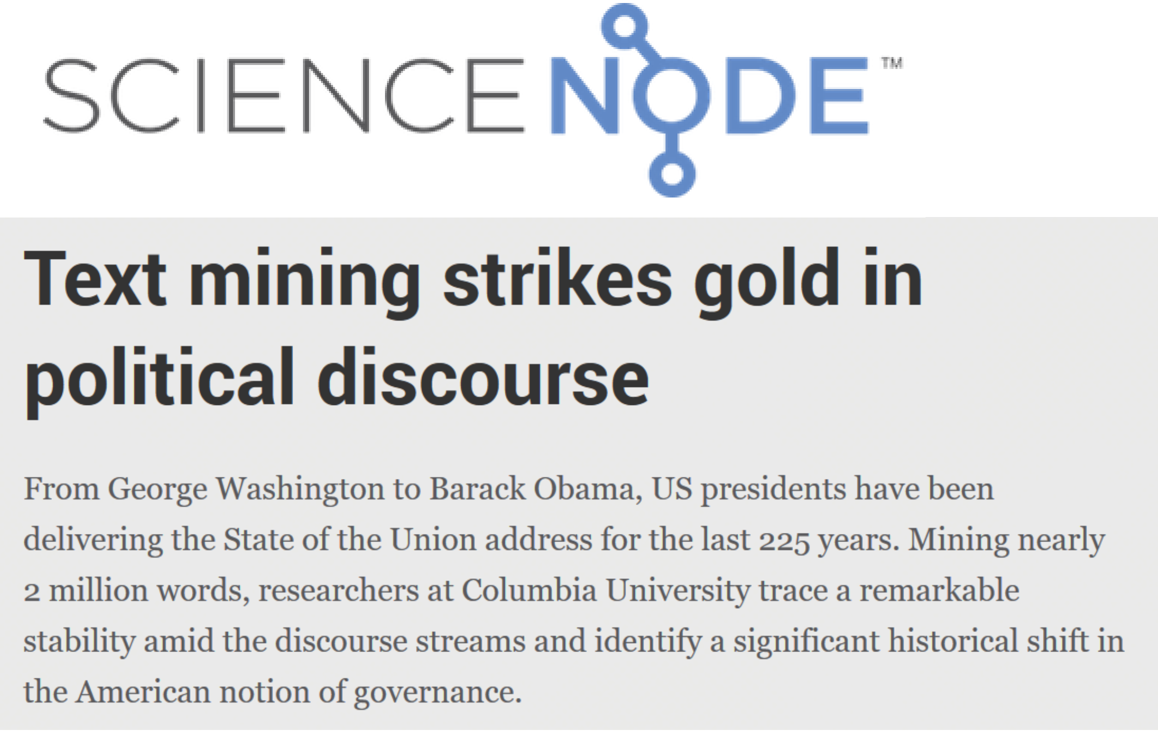
PEETER TINITS

#DIGMET SUMMER SCHOOL, TARTU

28.08.2019

Text mining



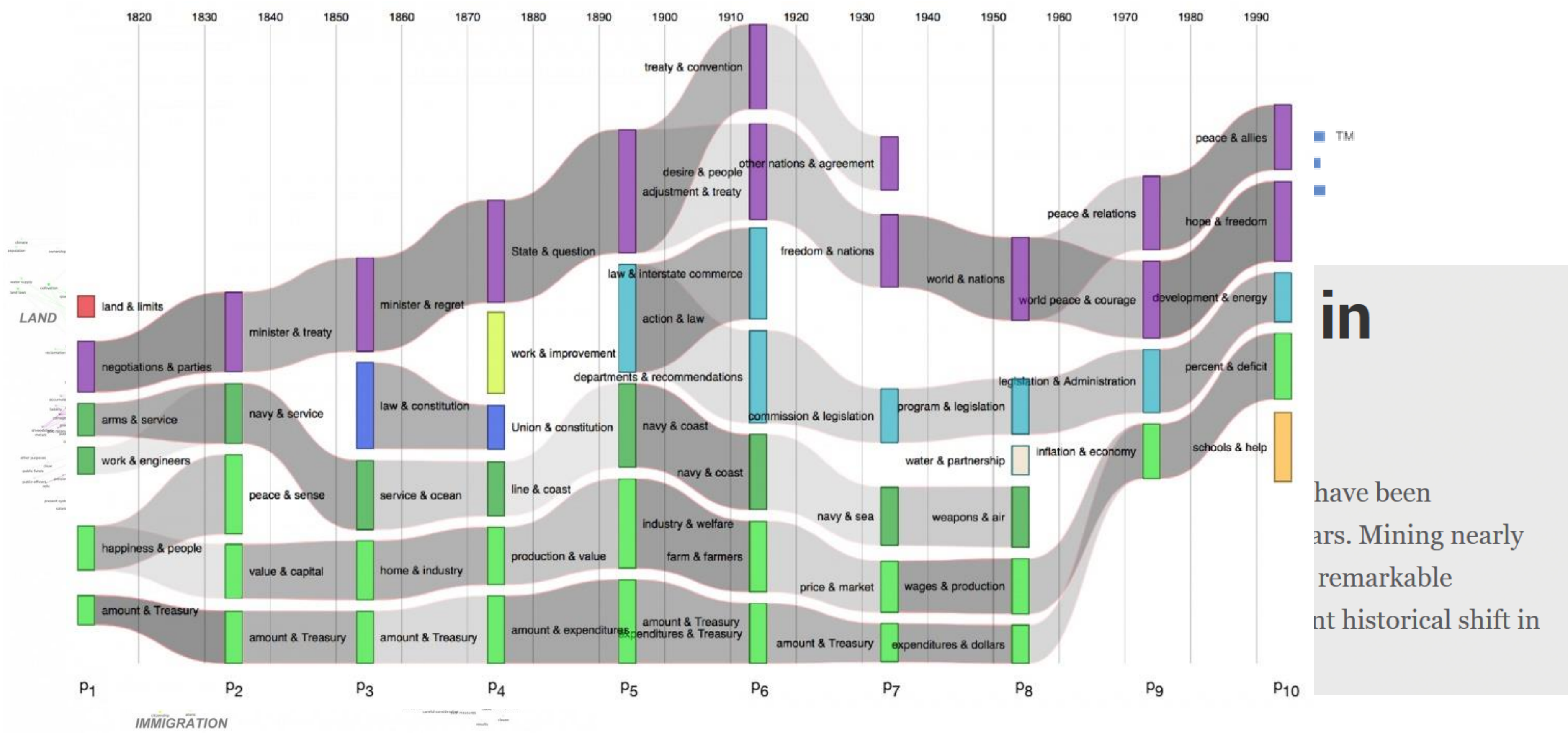


Text mining strikes gold in political discourse

From George Washington to Barack Obama, US presidents have been delivering the State of the Union address for the last 225 years. Mining nearly 2 million words, researchers at Columbia University trace a remarkable stability amid the discourse streams and identify a significant historical shift in the American notion of governance.

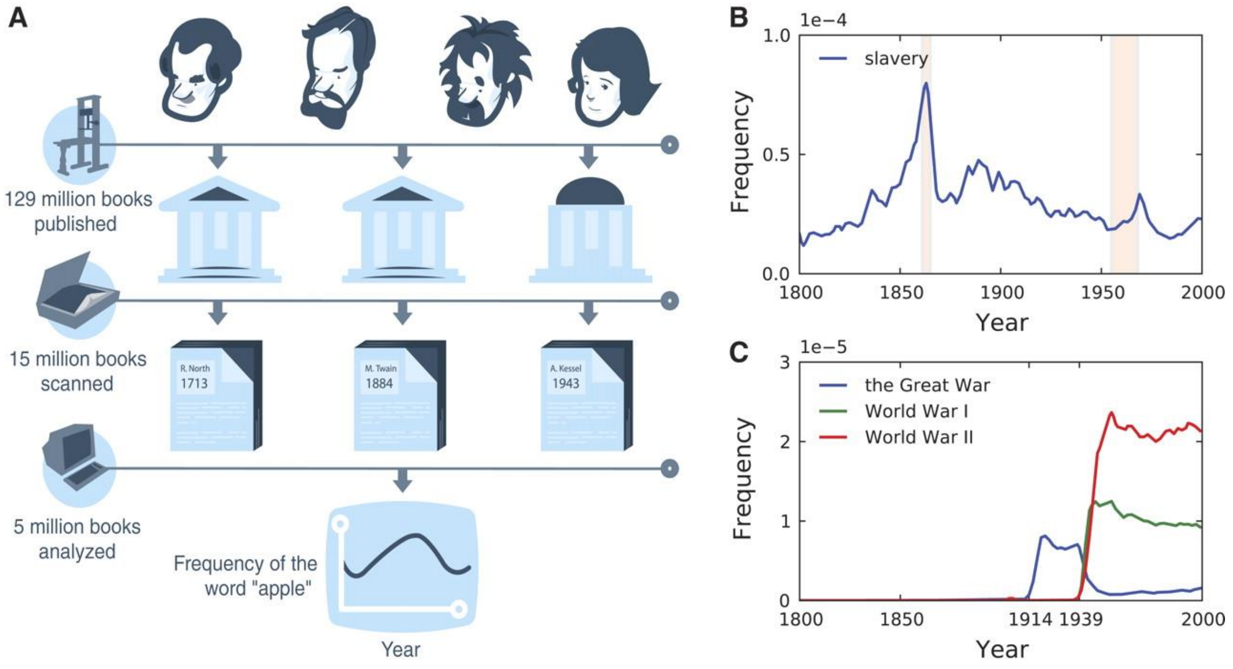
Rule et al. 2015 Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014

Text mining



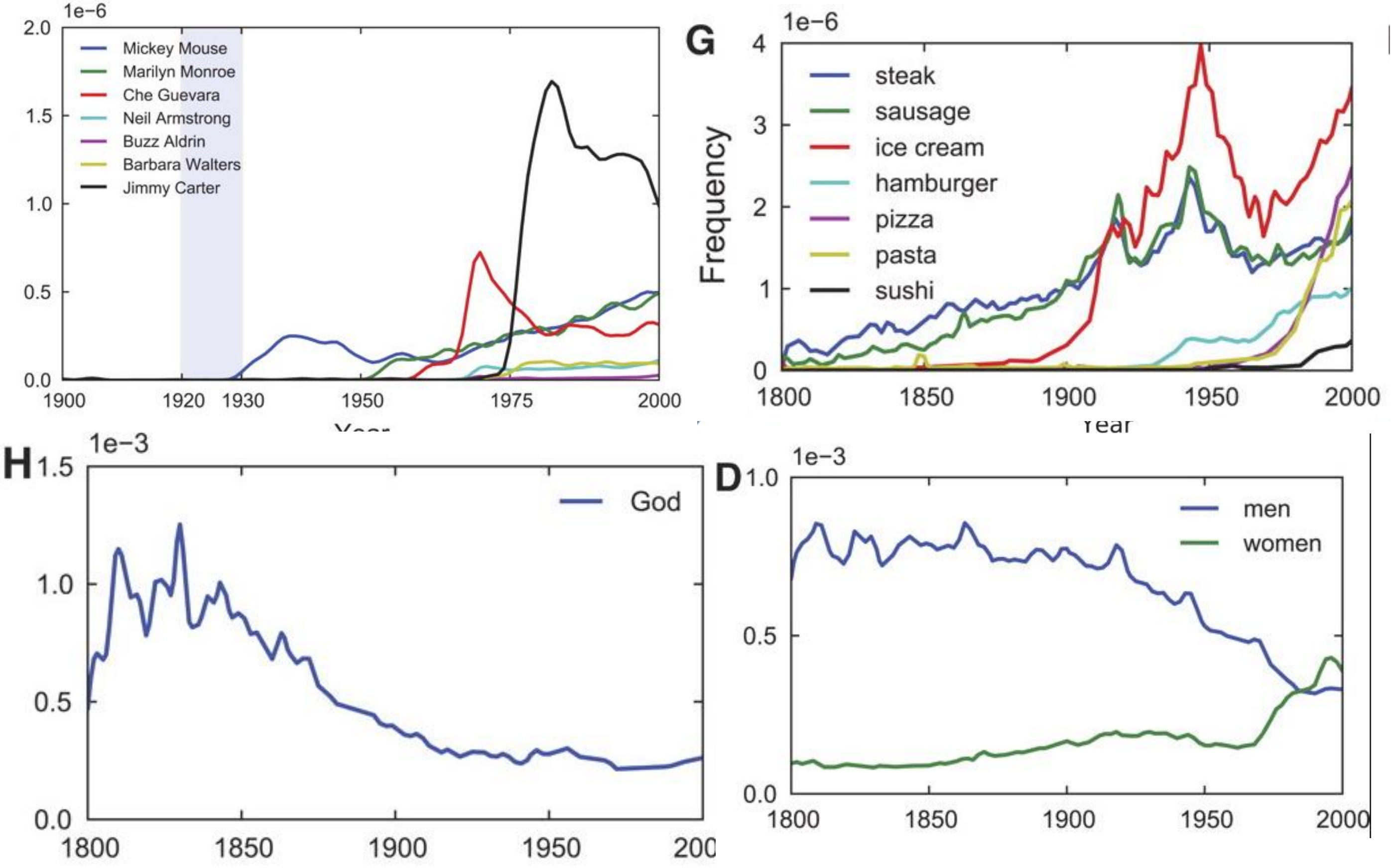
Rule et al. 2015 Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014

„All digitized texts“



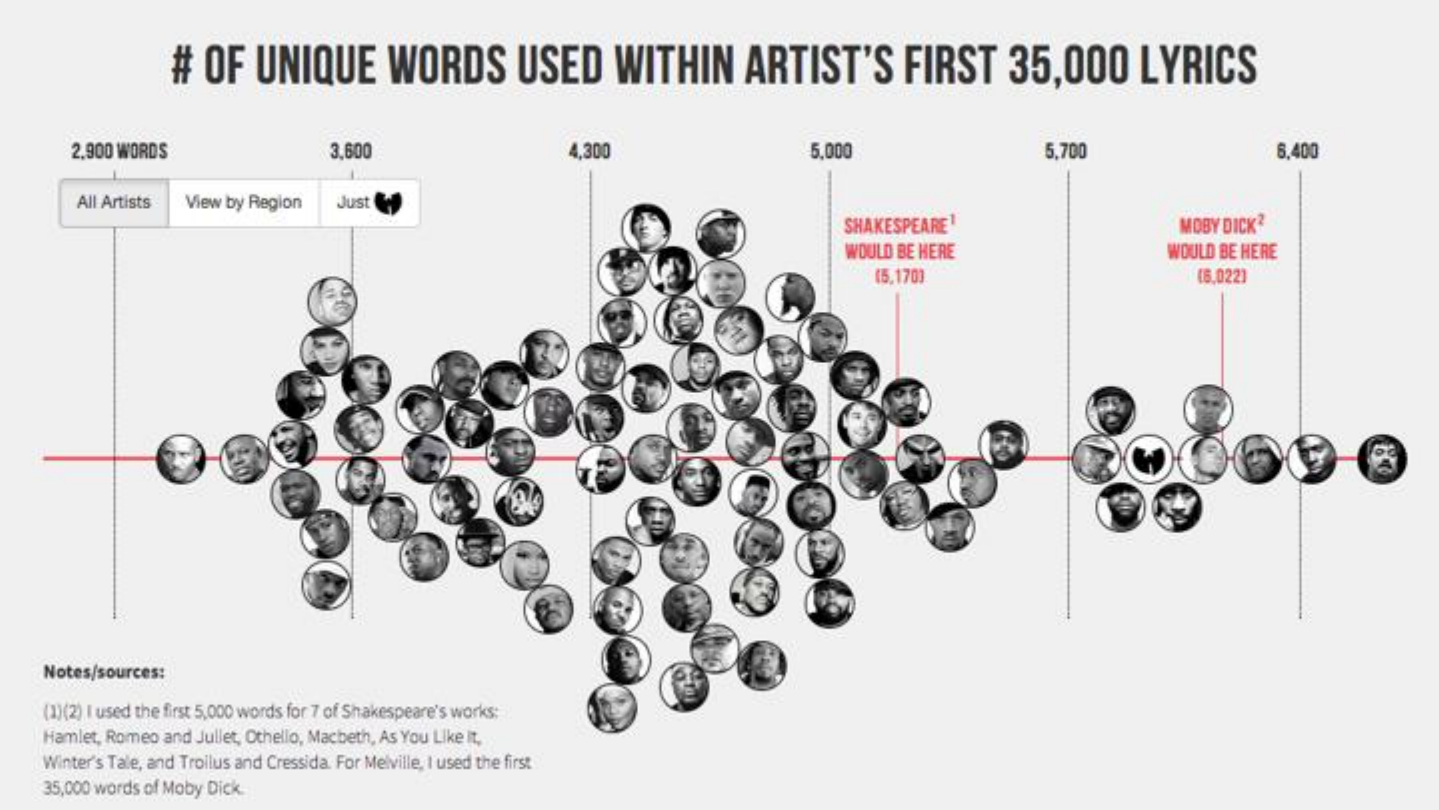
Michel et al. 2008 Quantitative Analysis of Culture Using Millions of Digitized Books

„All digitized texts“



Michel et al. 2008 Quantitative Analysis of Culture Using Millions of Digitized Books

Vocabulary of Rap

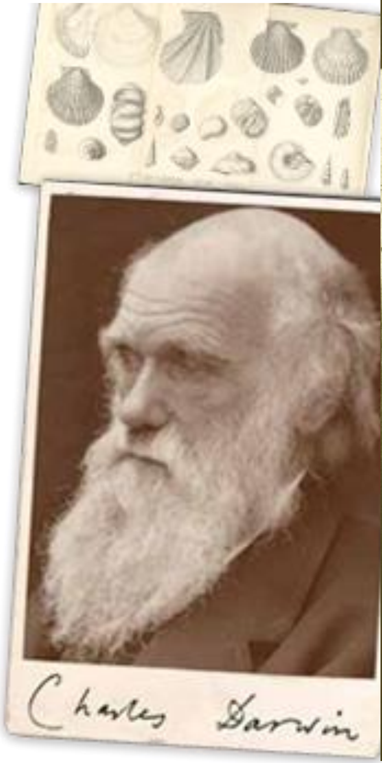


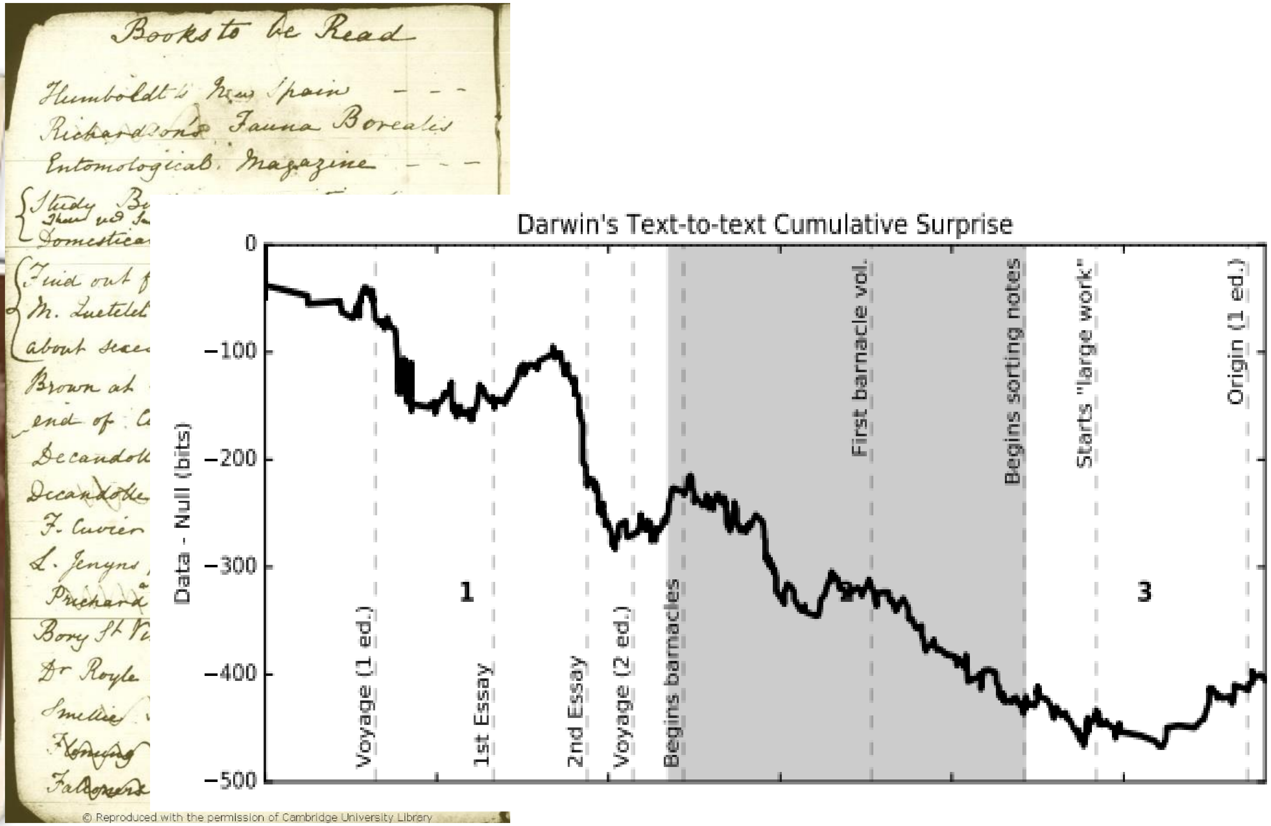
Notes/sources:

(1)(2) I used the first 5,000 words for 7 of Shakespeare's works: Hamlet, Romeo and Juliet, Othello, Macbeth, As You Like It, Winter's Tale, and Troilus and Cressida. For Melville, I used the first 35,000 words of Moby Dick.

https://pudding.cool/projects/vocabulary/

Darwin’s reading habits



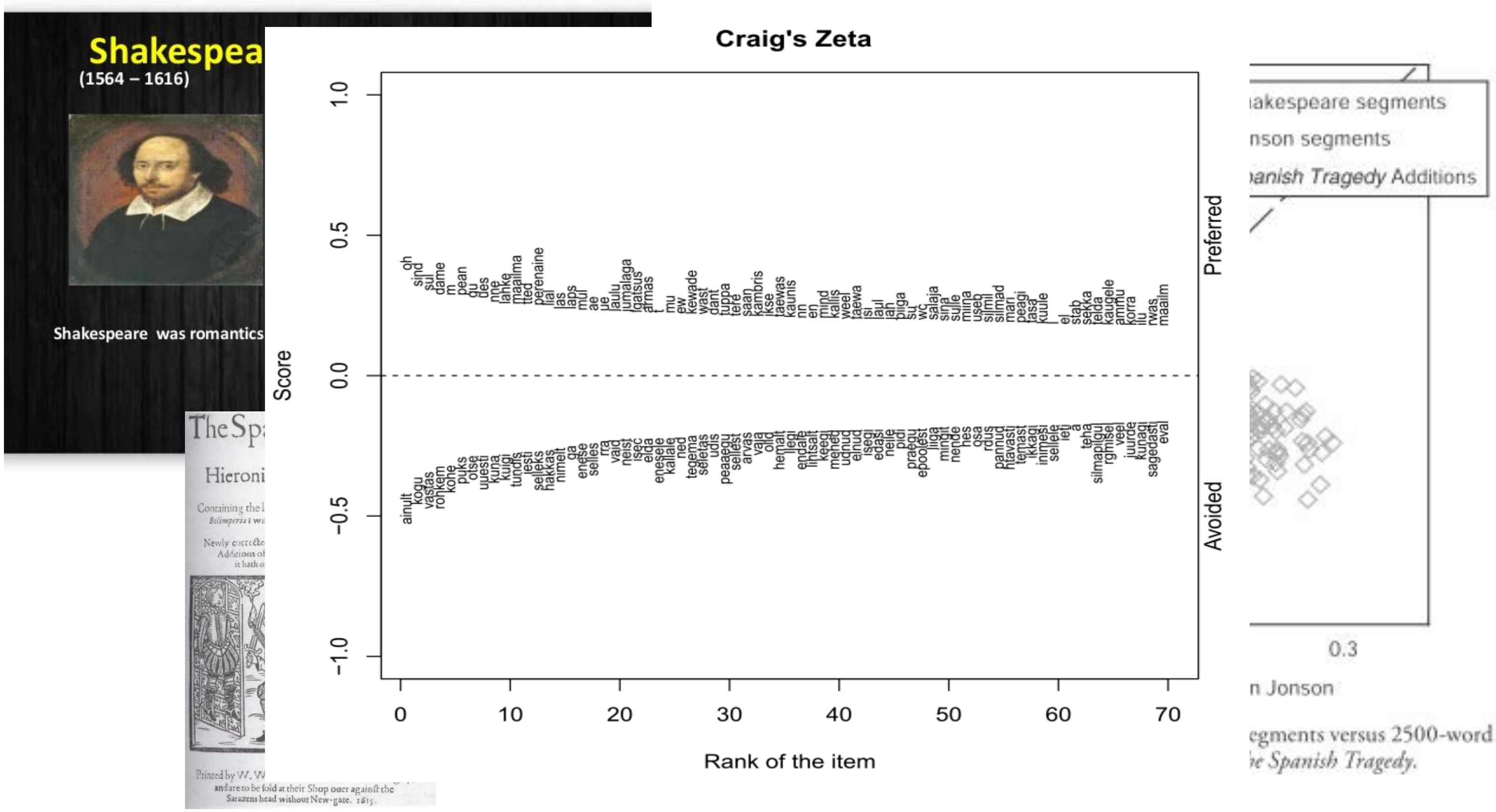

Books to be Read

Humboldt's New Spain — —
Richardson's Fauna Borealis
Entomological Magazine — —
Study B
Domestica
Find out f
M. Luetelet
about seec
Brown at
end of C
Decandoll
Decandolle
F. Cuvier
L. Jenyns
Prichard
Bory St V
Dr Royle
Smellie
Fleming
Falconers

© Reproduced with the permission of Cambridge University Library

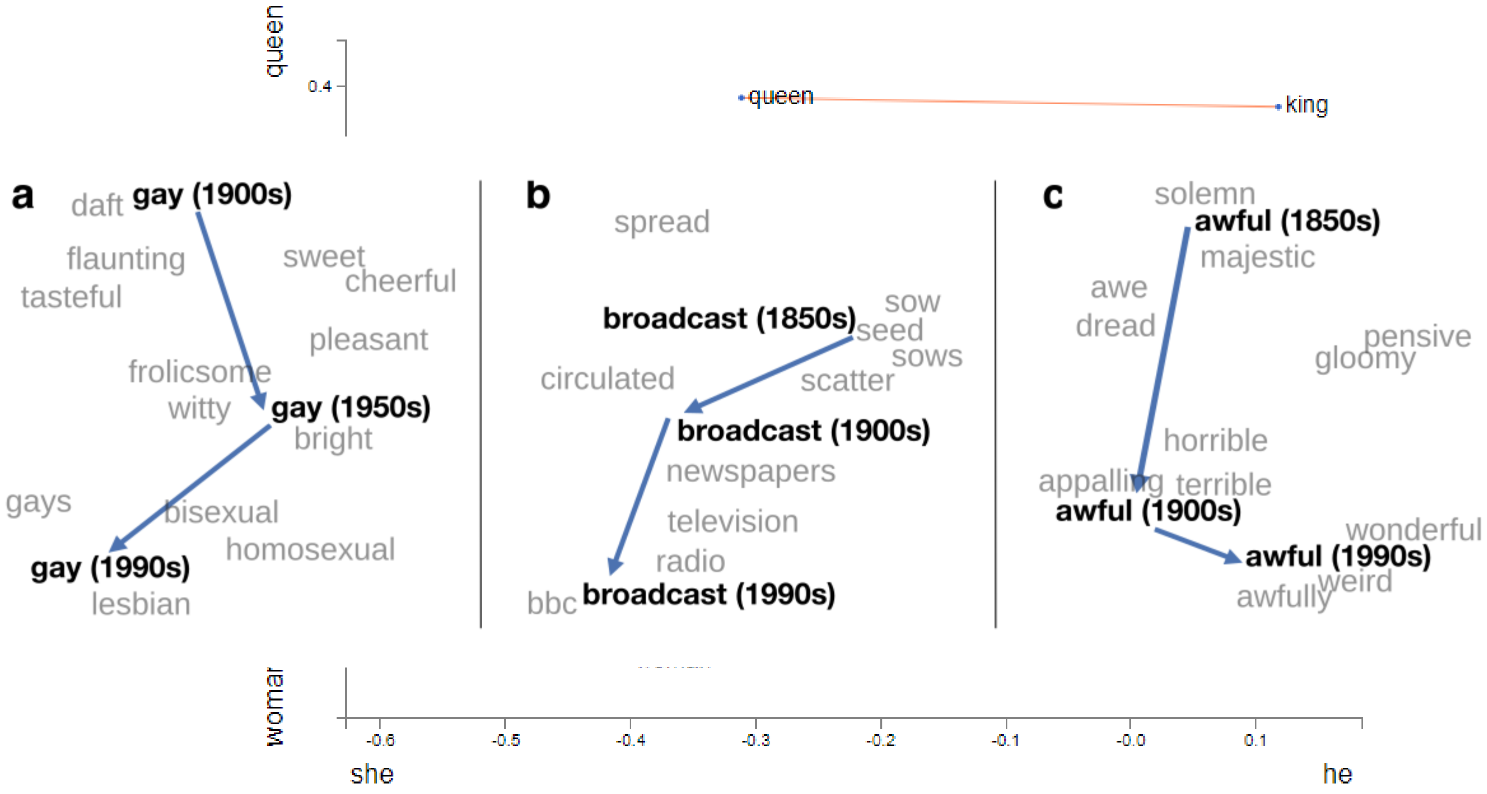
Murdock et al. 2017. Exploration and exploitation of Victorian science in Darwin's reading notebooks.

Finding the author



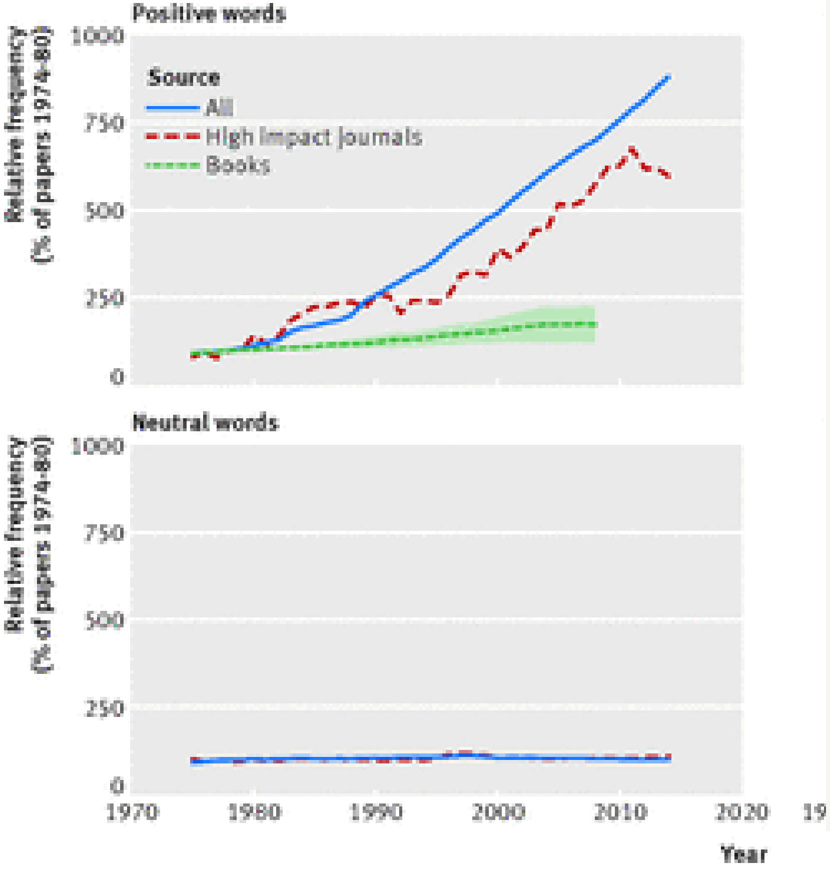
Craig & Kinney 2009. Shakespeare, Computers, and the Mystery of Authorship

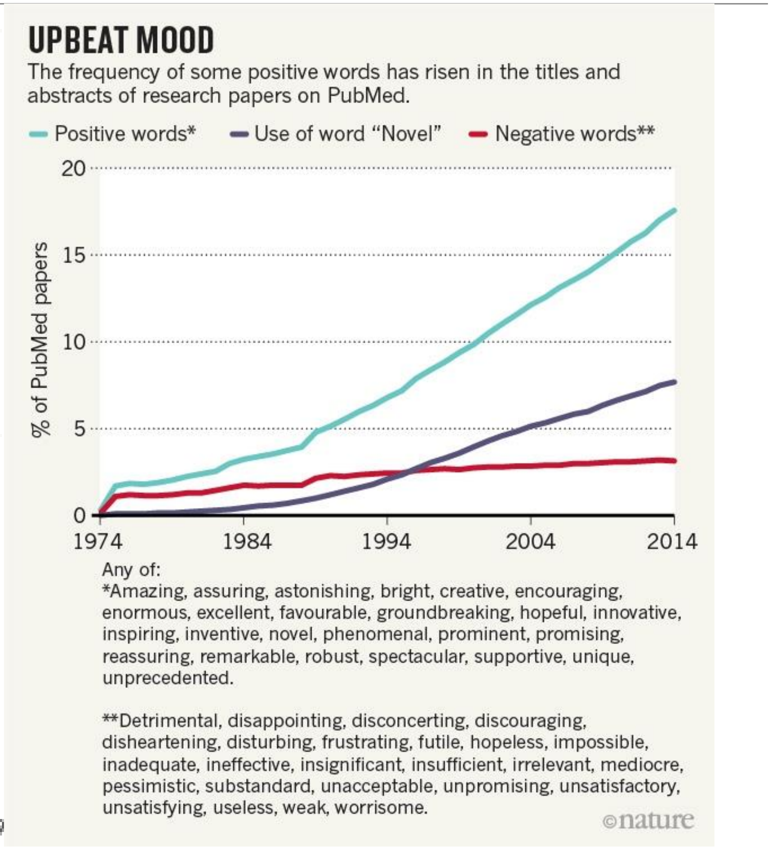
Word semantics



https://lamyiowce.github.io/word2viz/

Scientific language





UPBEAT MOOD

The frequency of some positive words has risen in the titles and abstracts of research papers on PubMed.

Any of:
*Amazing, assuring, astonishing, bright, creative, encouraging, enormous, excellent, favourable, groundbreaking, hopeful, innovative, inspiring, inventive, novel, phenomenal, prominent, promising, reassuring, remarkable, robust, spectacular, supportive, unique, unprecedented.

**Detrimental, disappointing, disconcerting, discouraging, disheartening, disturbing, frustrating, futile, hopeless, impossible, inadequate, ineffective, insignificant, insufficient, irrelevant, mediocre, pessimistic, substandard, unacceptable, unpromising, unsatisfactory, unsatisfying, useless, weak, worrisome.

Vinkers et al. 2015 Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis

Text mining

textual data <-> questions

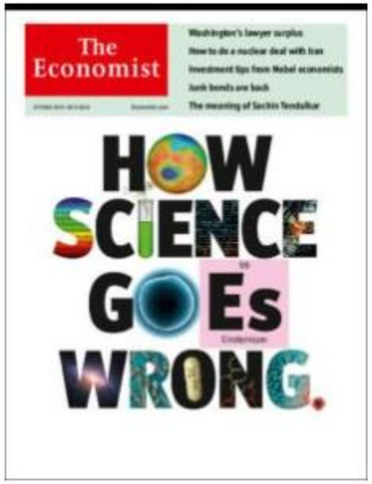
Questions matter.

Technology can be easy (if data is available).

Though not a magic bullet.

Reproducible research

Replication crisis



The Economist

Washington's lawyer surplus

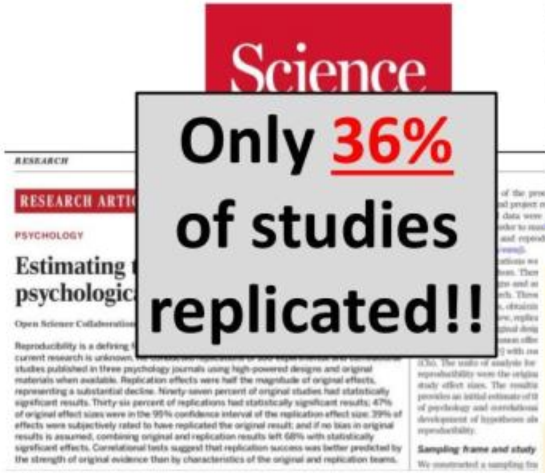
How to do a nuclear deal with Iran

Investment tips from Nobel economists

Junk bonds are back

The meaning of Sachin Tendulkar

HOW SCIENCE GOES WRONG.



Science

RESEARCH

RESEARCH ARTI

PSYCHOLOGY

Estimating t
psychologic

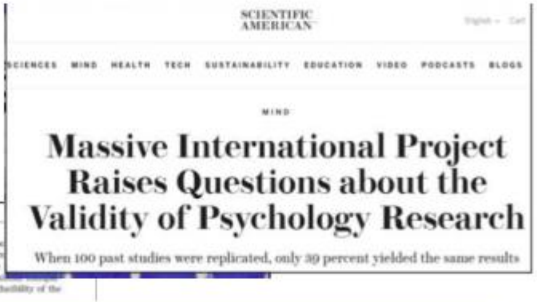
Open Science Collaboration

Reproducibility is a defining f current research is unknown. studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

Sampling frame and study

We constructed a sampling fra

Only **36%** of studies replicated!!



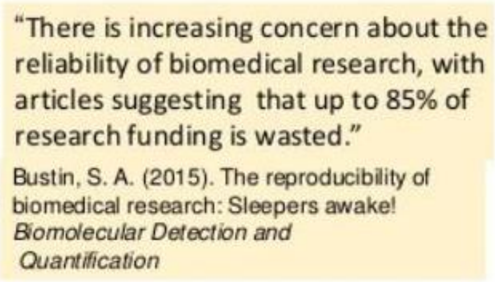
SCIENTIFIC AMERICAN

SCIENCES MIND HEALTH TECH SUSTAINABILITY EDUCATION VIDEO PODCASTS BLOGS

MIND

Massive International Project Raises Questions about the Validity of Psychology Research

When 100 past studies were replicated, only 39 percent yielded the same results



"There is increasing concern about the reliability of biomedical research, with articles suggesting that up to 85% of research funding is wasted."

Bustin, S. A. (2015). The reproducibility of biomedical research: Sleepers awake! *Biomolecular Detection and Quantification*

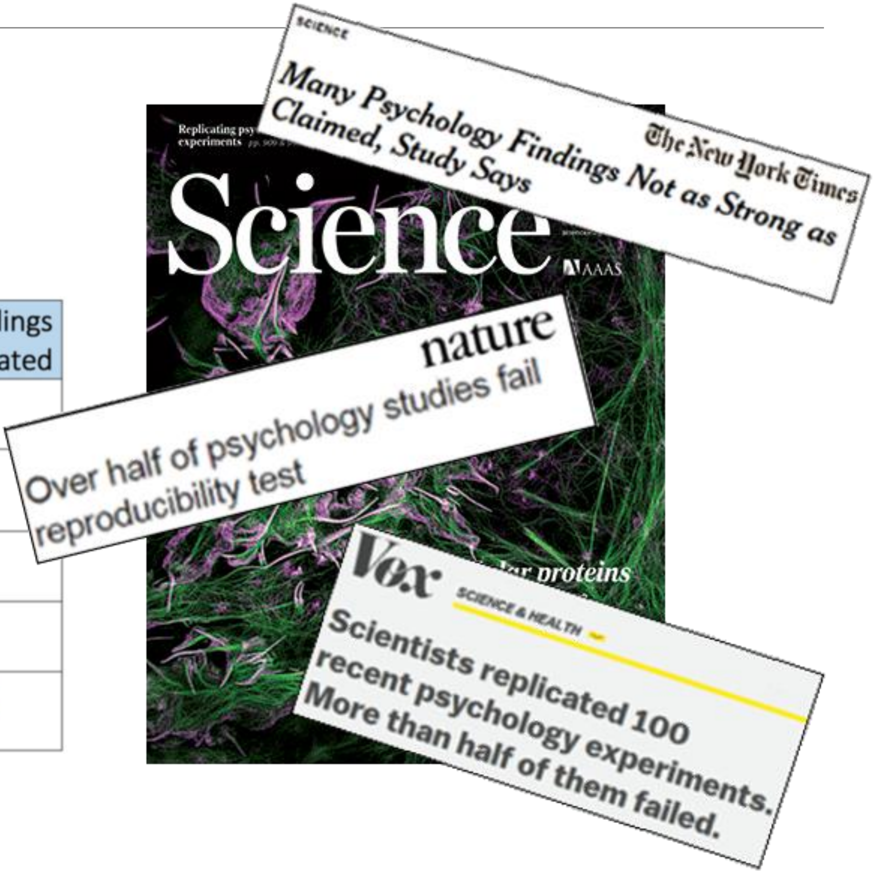
ERR novaator KOOSTÖÖS TARTU ÜLIKOOLIGA

Katsete korratavus kütab psühholoogias kirgi

Analüüs: teaduskirjandus on kiivas, kriisist pole mõtet rääkida

Replication crisis

Journal	% Findings Replicated
Journal of Personality and Social Psychology: Social	23
Journal of Experimental Psychology: Learning, Memory, and Cognition	48
Psychological Science, social articles	29
Psychological Science, cognitive articles	53
Overall	36


Replicating psy
experiments
Science
AAAS
lar proteins

SCIENCE

Many Psychology Findings Not as Strong as Claimed, Study Says

The New York Times

nature

Over half of psychology studies fail reproducibility test

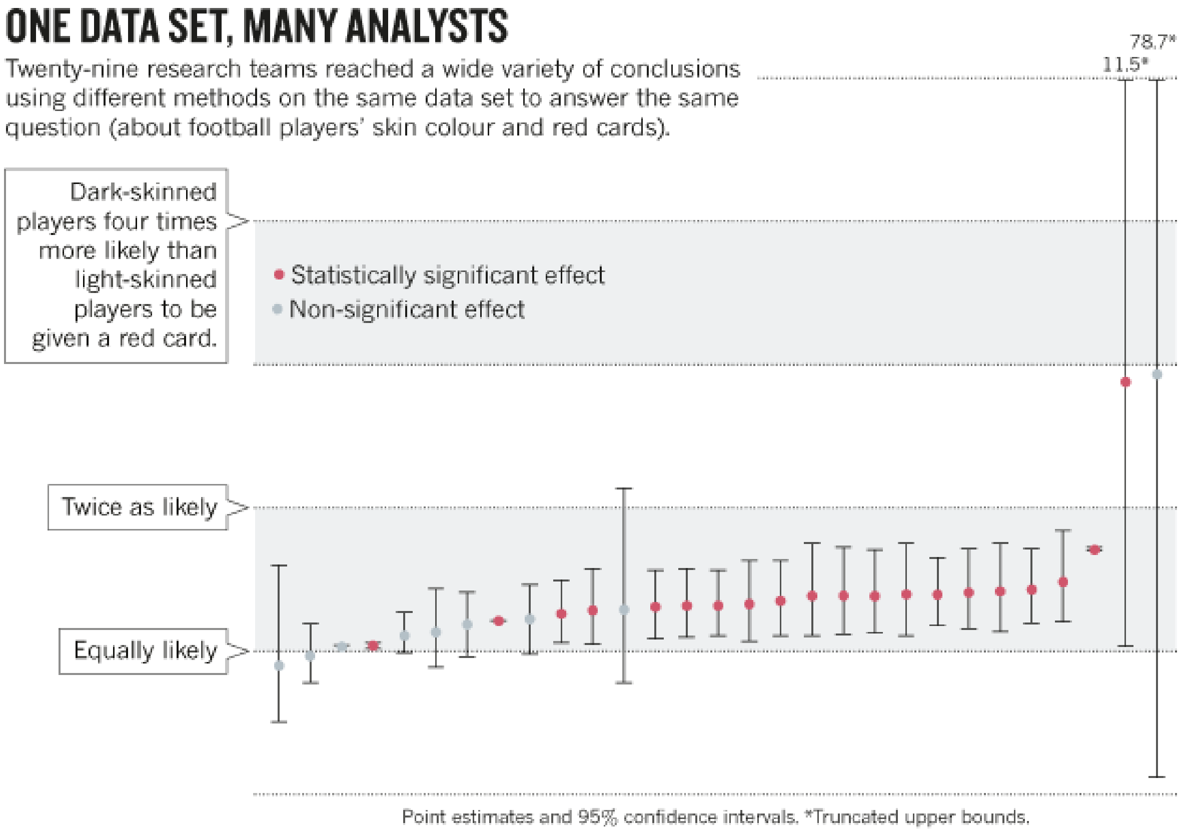
Vox

SCIENCE & HEALTH

Scientists replicated 100 recent psychology experiments. More than half of them failed.

Open Science Collaboration 2015. Estimating the reproducibility of psychological science.

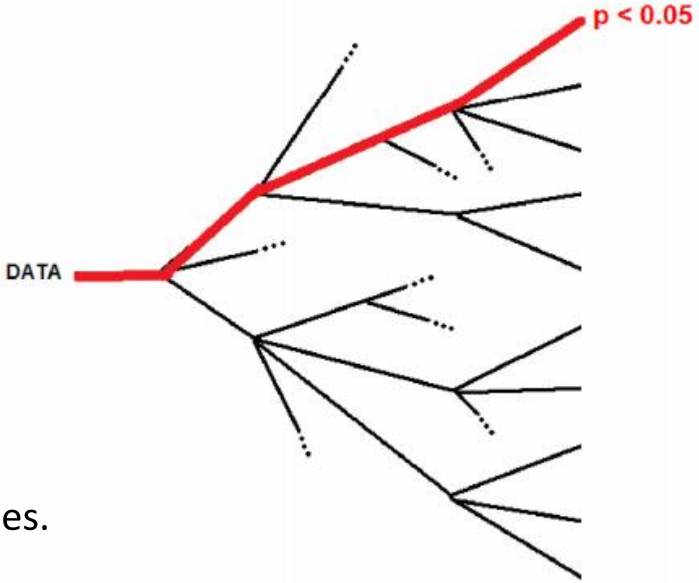
Same data, different conclusions



Silberzahn et al. 2018. [Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results](#)

Garden of forking paths

„Researcher degrees of freedom“

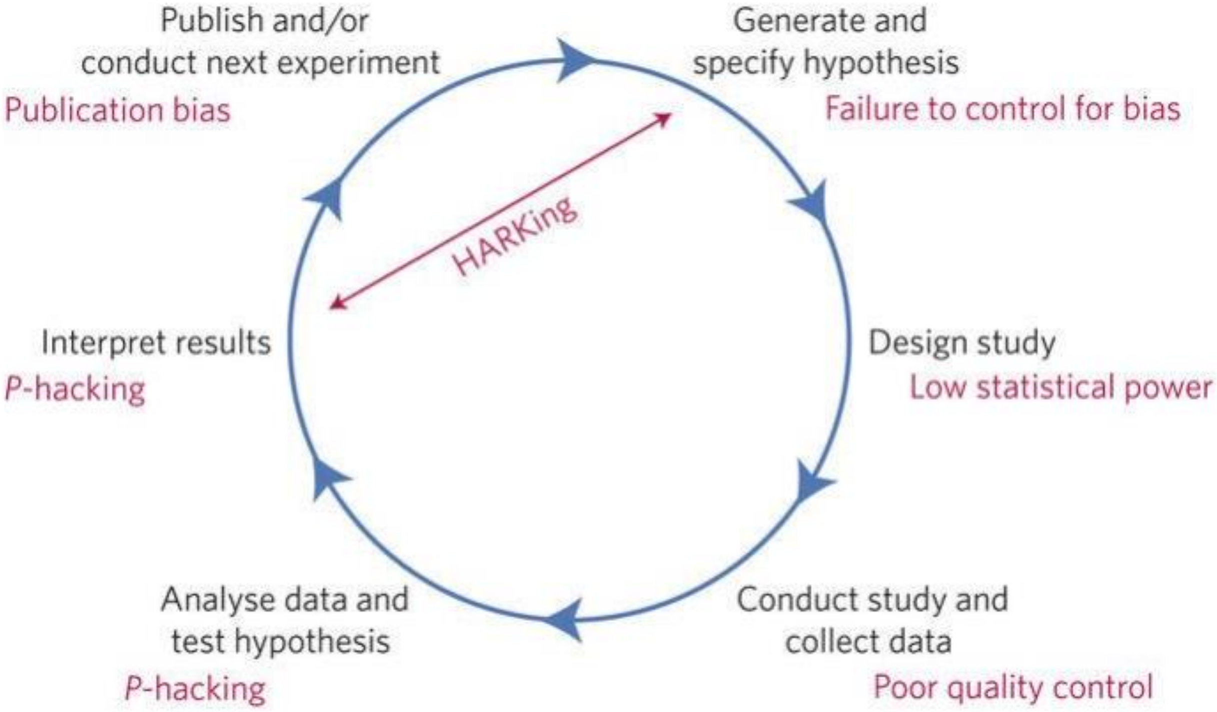


Problem is, if you pretend that there were no choices.

This was always the path we were going to take.

Andrew Gelman & Eric Loken 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no ``fishing expedition'' or ``p-hacking'' and the research hypothesis was posited ahead of time. (Unpublished.)

Problems all over the research cycle



Generate and specify hypothesis

Failure to control for bias

Design study

Low statistical power

Conduct study and collect data

Poor quality control

Analyse data and test hypothesis

P-hacking

Interpret results

P-hacking

Publish and/or conduct next experiment

Publication bias

Munafo et al. 2017. A manifesto for reproducible science

Open science solutions

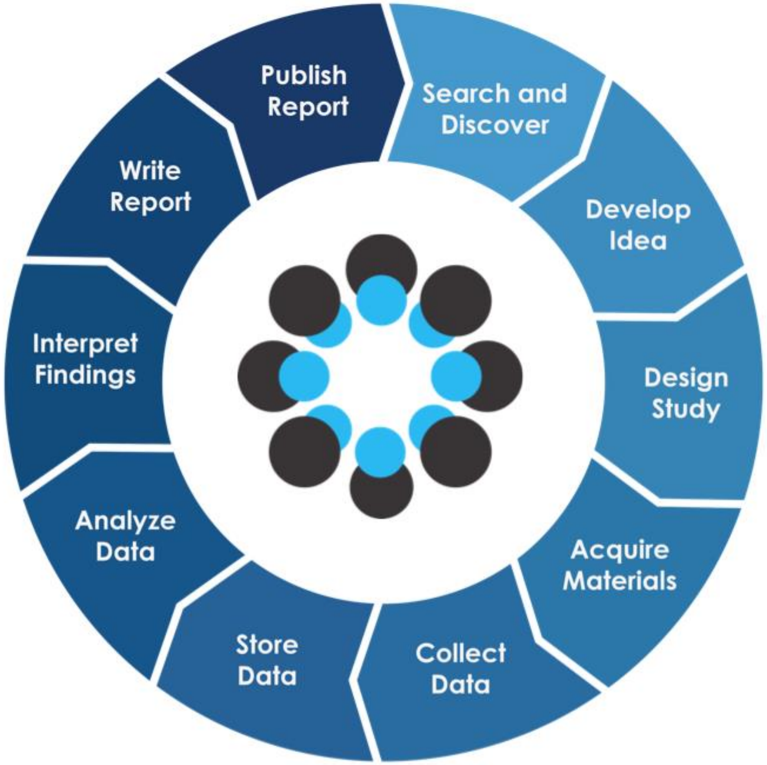
Openness

Fairness

Transparency

Open science solutions

CENTER FOR OPEN SCIENCE



Search and Discover

Develop Idea

Design Study

Acquire Materials

Collect Data

Store Data

Analyze Data

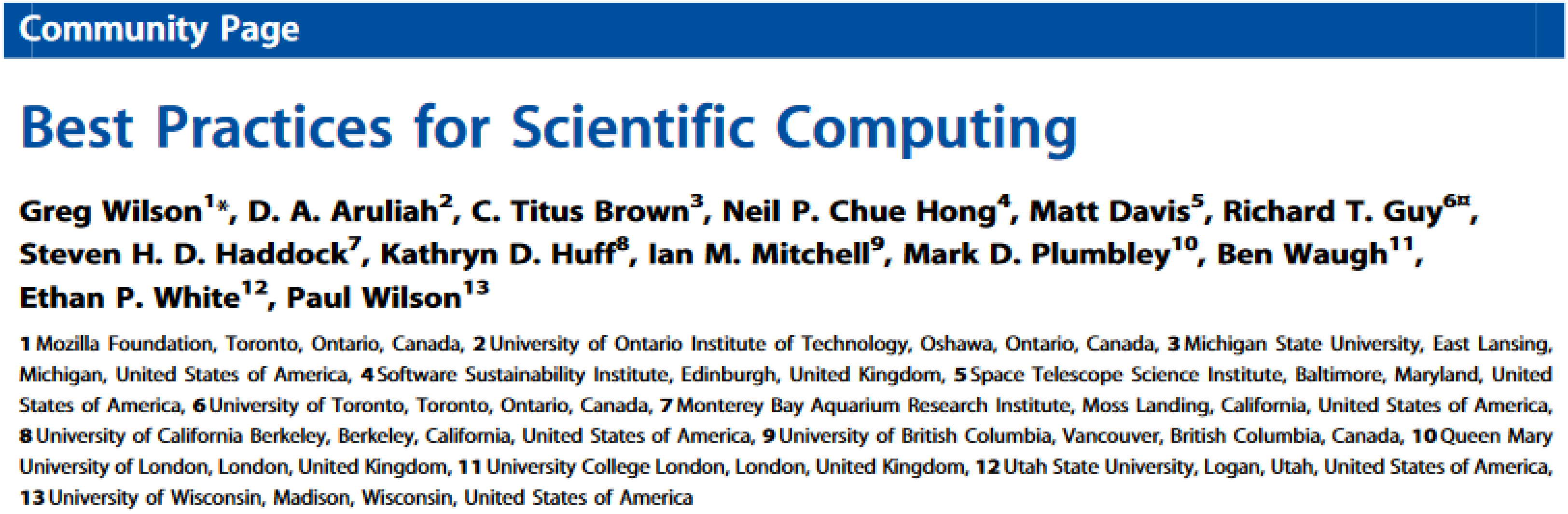
Interpret Findings

Write Report

Publish Report

https://osf.io/

Standards and practices



Community Page

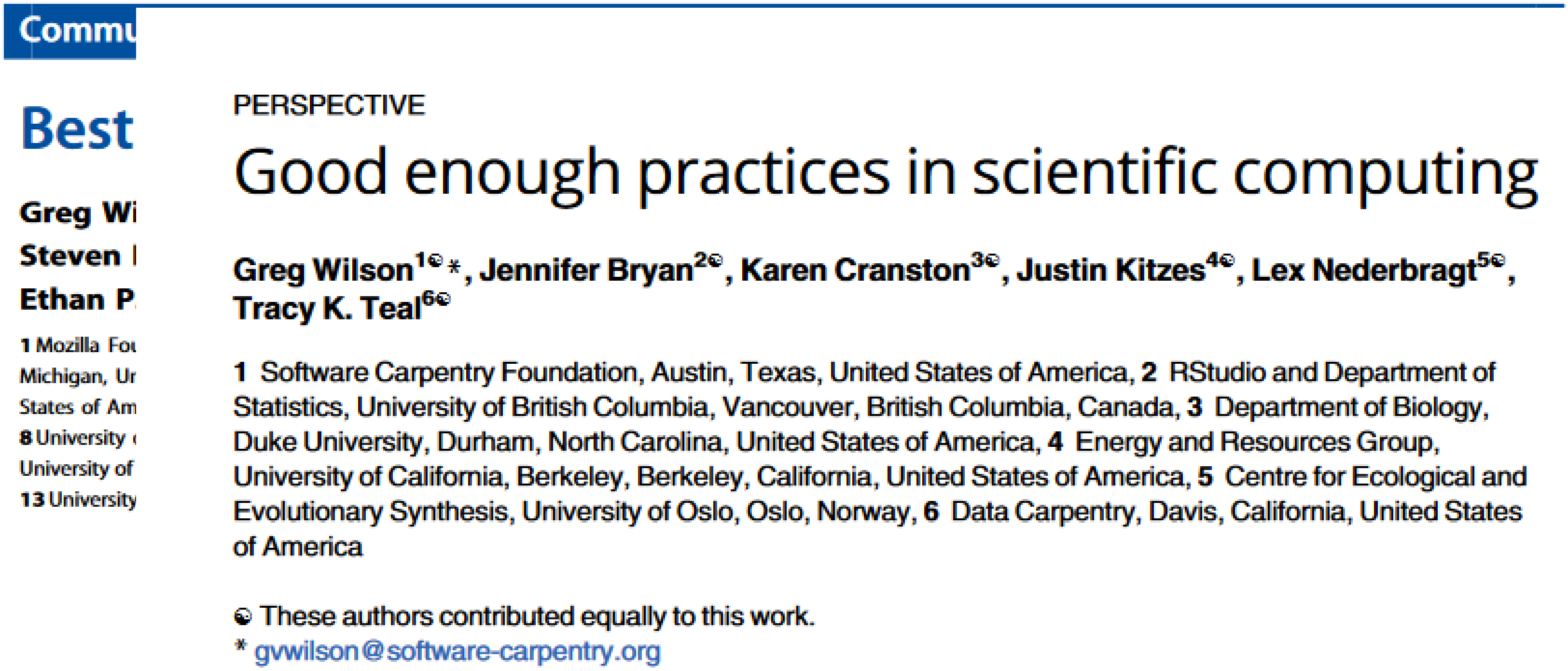
Best Practices for Scientific Computing

Greg Wilson[1]*, D. A. Aruliah[2], C. Titus Brown[3], Neil P. Chue Hong[4], Matt Davis[5], Richard T. Guy[6¤], Steven H. D. Haddock[7], Kathryn D. Huff[8], Ian M. Mitchell[9], Mark D. Plumbley[10], Ben Waugh[11], Ethan P. White[12], Paul Wilson[13]

1 Mozilla Foundation, Toronto, Ontario, Canada, **2** University of Ontario Institute of Technology, Oshawa, Ontario, Canada, **3** Michigan State University, East Lansing, Michigan, United States of America, **4** Software Sustainability Institute, Edinburgh, United Kingdom, **5** Space Telescope Science Institute, Baltimore, Maryland, United States of America, **6** University of Toronto, Toronto, Ontario, Canada, **7** Monterey Bay Aquarium Research Institute, Moss Landing, California, United States of America, **8** University of California Berkeley, Berkeley, California, United States of America, **9** University of British Columbia, Vancouver, British Columbia, Canada, **10** Queen Mary University of London, London, United Kingdom, **11** University College London, London, United Kingdom, **12** Utah State University, Logan, Utah, United States of America, **13** University of Wisconsin, Madison, Wisconsin, United States of America

https://doi.org/10.1371/journal.pbio.1001745

Standards and practices



Commu

Best

Greg Wi

Steven I

Ethan P.

1 Mozilla Fou
Michigan, Ur
States of Am
8 University
University of
13 University

PERSPECTIVE

Good enough practices in scientific computing

Greg Wilson[1☯]*, Jennifer Bryan[2☯], Karen Cranston[3☯], Justin Kitzes[4☯], Lex Nederbragt[5☯], Tracy K. Teal[6☯]

1 Software Carpentry Foundation, Austin, Texas, United States of America, **2** RStudio and Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada, **3** Department of Biology, Duke University, Durham, North Carolina, United States of America, **4** Energy and Resources Group, University of California, Berkeley, Berkeley, California, United States of America, **5** Centre for Ecological and Evolutionary Synthesis, University of Oslo, Oslo, Norway, **6** Data Carpentry, Davis, California, United States of America

☯ These authors contributed equally to this work.
* gvwilson@software-carpentry.org

https://doi.org/10.1371/journal.pbio.1001745

https://doi.org/10.1371/journal.pcbi.1005510

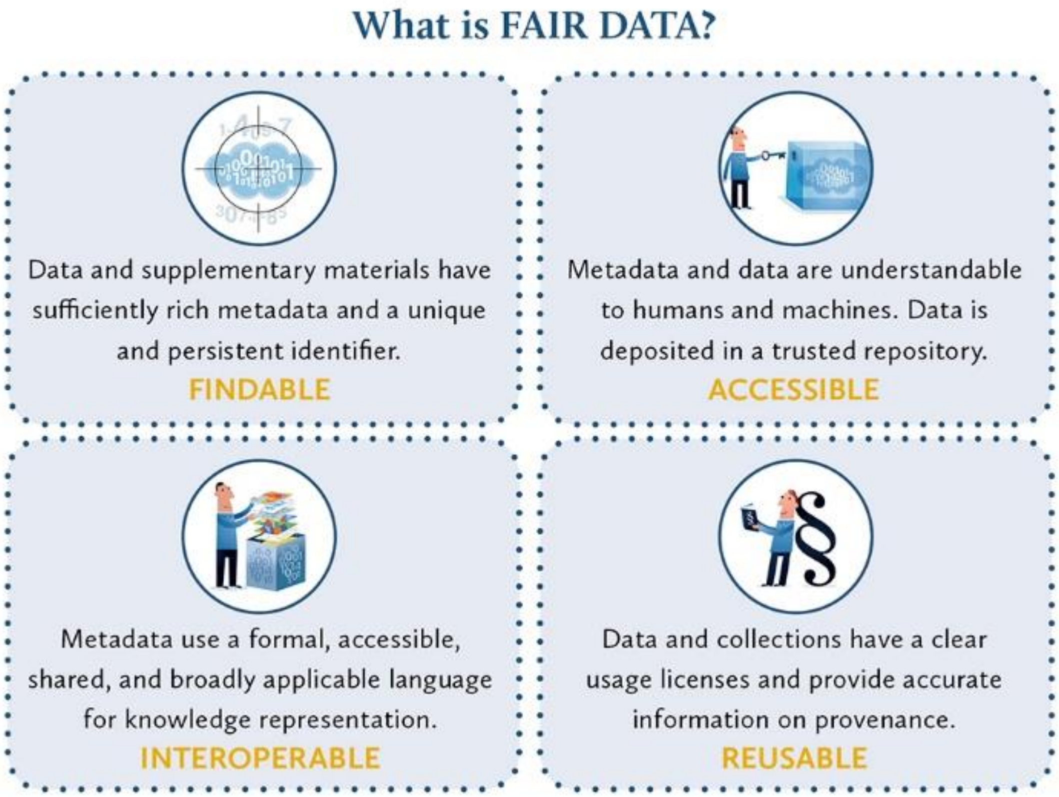
FAIR data

Findable

Accessible

Interoperable

Reusable



What is FAIR DATA?

Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

FINDABLE

Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

ACCESSIBLE

Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

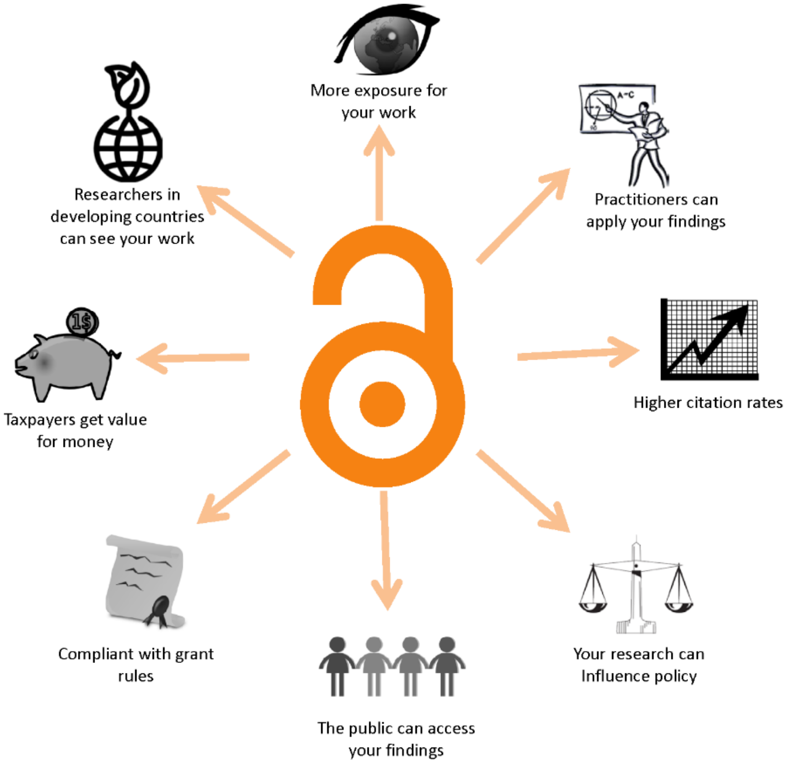
INTEROPERABLE

Data and collections have a clear usage licenses and provide accurate information on provenance.

REUSABLE

https://www.go-fair.org/fair-principles/

Open data



More exposure for your work

Practitioners can apply your findings

Higher citation rates

Your research can Influence policy

The public can access your findings

Compliant with grant rules

Taxpayers get value for money

Researchers in developing countries can see your work

CC-BY Danny Kingsley & Sarah Brown

https://scienceport.tut.fi/openaccess/whyoa

Practically, what it means

Open data

- Easy to use and access

Open scripts

- Published, verifiable, reusable

Multiple independent analysis

- Multilab studies: e.g. https://osf.io/tukby/, https://osf.io/j9ady/

Building on reusable elements

- Learning their limits

Primary motivations are selfish

Confidence

Comfort

Easy to share

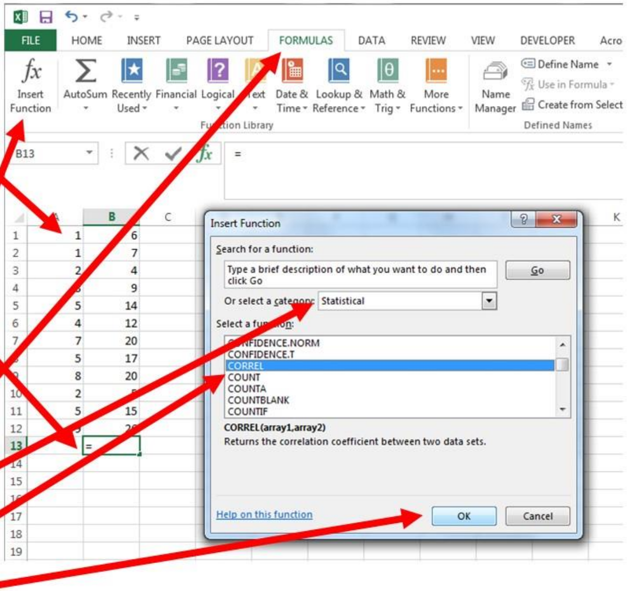
Easy to keep

Easy to reuse

Comfort

EXCEL

1. Enter the paired scores for each subject on an Excel spreadsheet.
2. After the data have been entered, place the cursor in an empty cell where you wish to have the correlation coefficient (Pearson's *r*) appear and click the mouse button.
3. Select **Insert Function** (f_x) from the **FORMULAS** tab.
4. A dialog box will appear. Select **Statistical**, select **CORREL**, and click **OK**.



	A	B
1	1	6
2	1	7
3	2	4
4	3	9
5	5	14
6	4	12
7	7	20
8	5	17
9	8	20
10	2	5
11	5	15
12	[illegible]	20
13		=

Search for a function:

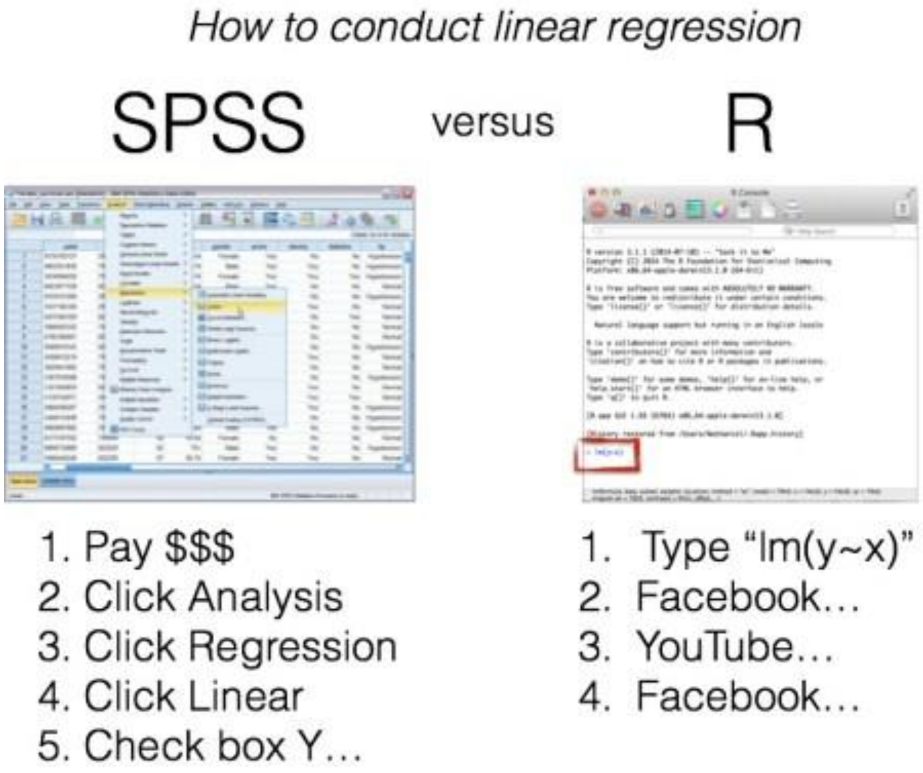
Type a brief description of what you want to do and then click Go

Or select a category: Statistical

Select a function:

CONFIDENCE.NORM
CONFIDENCE.T
CORREL
COUNT
COUNTA
COUNTBLANK
COUNTIF

CORREL(array1,array2)
Returns the correlation coefficient between two data sets.



How to conduct linear regression

SPSS versus **R**

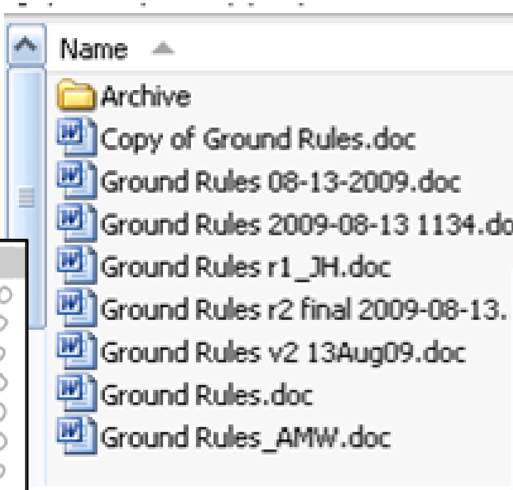
1. Pay $$$
2. Click Analysis
3. Click Regression
4. Click Linear
5. Check box Y…

1. Type "lm(y~x)"
2. Facebook…
3. YouTube…
4. Facebook…

Keeping track

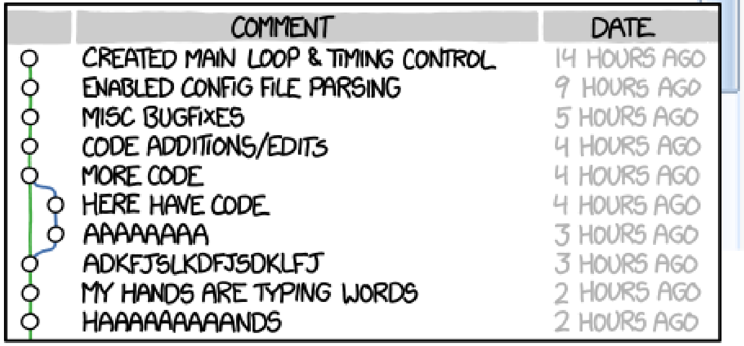
Open scripts is also documentation

- „Your best/worst collaborator is YOU from 6 months ago, and they do not answer e-mails!“ – Academia lore.



Name	Size	Date Modified
Archive		8/14/2009 8:38 PM
Copy of Ground Rules.doc	11 KB	8/14/2009 8:36 PM
Ground Rules 08-13-2009.doc	11 KB	8/14/2009 8:36 PM
Ground Rules 2009-08-13 1134.doc	11 KB	8/14/2009 8:36 PM
Ground Rules r1_JH.doc	11 KB	8/14/2009 8:36 PM
Ground Rules r2 final 2009-08-13.		
Ground Rules v2 13Aug09.doc		
Ground Rules.doc		
Ground Rules_AMW.doc		

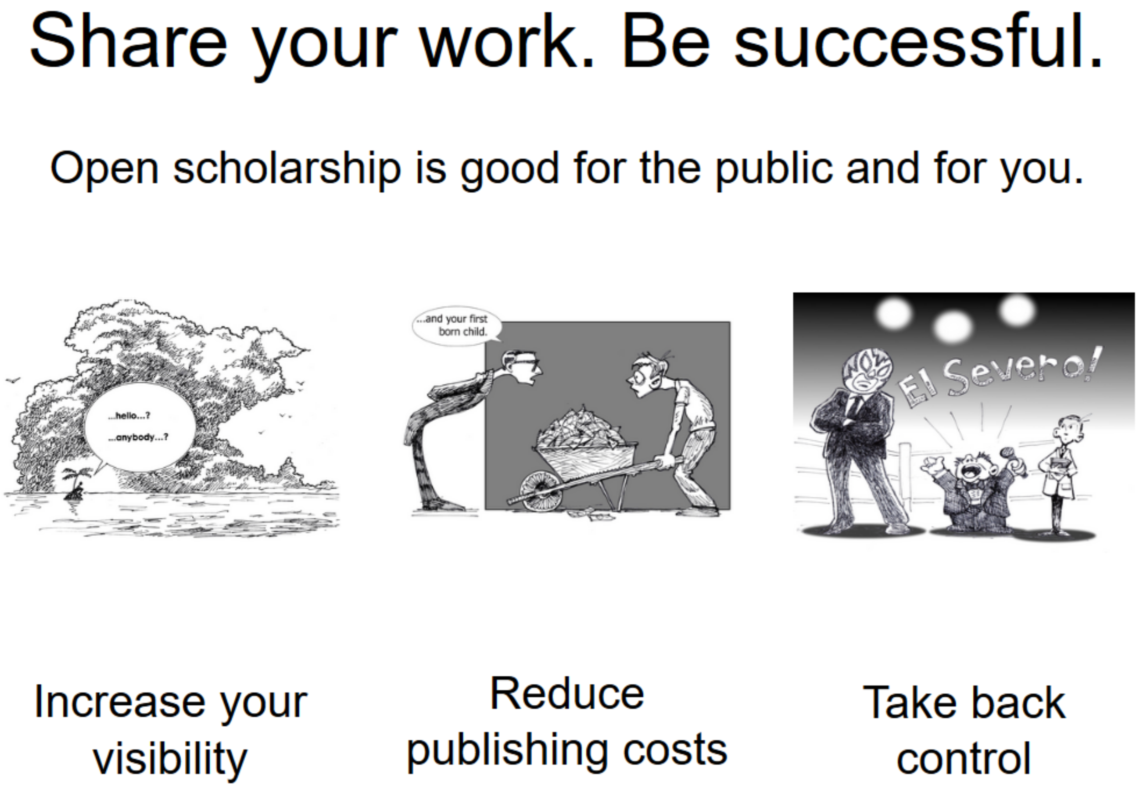
EVERY DESIGNER IN THIS WORLD
PSD
Ps
new.psd
newfinal.psd
newfinalfinal.psd
newfinalestfinal.psd
newfinalestfinal forsure.psd
newfinalestfuckthis shitfinal.psd



COMMENT	DATE
CREATED MAIN LOOP & TIMING CONTROL	14 HOURS AGO
ENABLED CONFIG FILE PARSING	9 HOURS AGO
MISC BUGFIXES	5 HOURS AGO
CODE ADDITIONS/EDITS	4 HOURS AGO
MORE CODE	4 HOURS AGO
HERE HAVE CODE	4 HOURS AGO
AAAAAAAA	3 HOURS AGO
ADKFJSLKDFJSDKLFJ	3 HOURS AGO
MY HANDS ARE TYPING WORDS	2 HOURS AGO
HAAAAAAAANDS	2 HOURS AGO

AS A PROJECT DRAGS ON, MY GIT COMMIT MESSAGES GET LESS AND LESS INFORMATIVE.

Sharing is caring



Share your work. Be successful.

Open scholarship is good for the public and for you.

Increase your visibility

Reduce publishing costs

Take back control

http://whyopenresearch.org/

Other opinions

Five selfish reasons

- Reason 1: reproducibility helps to avoid disaster
- Reason 2: reproducibility makes it easier to write papers
- Reason 3: reproducibility helps reviewers see it your way
- Reason 4: reproducibility enables continuity of your work
- Reason 5: reproducibility helps to build your reputation

Florian M. 2015. Five selfish reasons to work reproducibly
Open Science Training Handbook. 2018. Reproducible Research and Data Analysis
Gatto, L. 2019. Becoming a better scientist with open and reproducible research
Reproducible Research MOOC at coursera.org

Terminology

Technically:

- Replicability – a new experiment
- Reproducibility – can I rerun the code?

Today:

Reproducibility!