UNIVERSITY OF CAMBRIDGE

MRC Medical Research Council

Open science with large public datasets: Principles, practice and pitfalls

Rogier A. Kievit

Sir Henry Wellcome Fellow

Group leader, Lifespan Cognitive Dynamics lab, MRC-CBU

rogier.kievit@mrc-cbu.cam.ac.uk

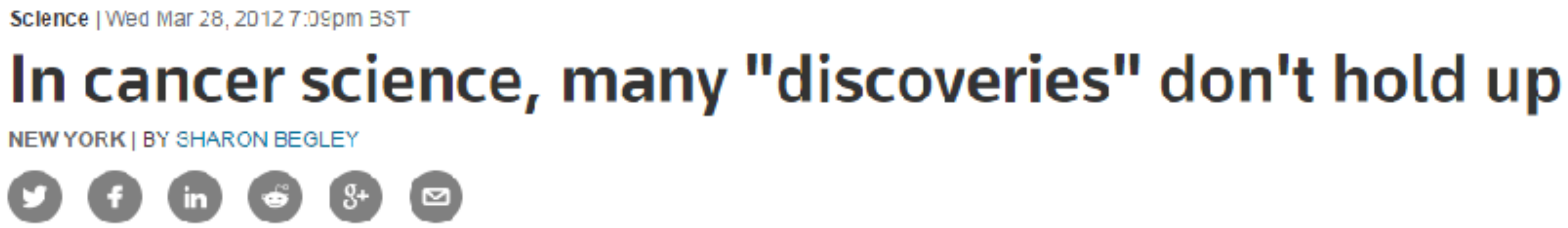
@rogierK

Outline

- Principles
  - Power
  - Preregistration
  - Practice
- Practice
  - Where to find public data?
  - Public data and open science
- Pitfalls
  - Getting data
  - The data

Principles

Problem 1: many findings don’t hold up



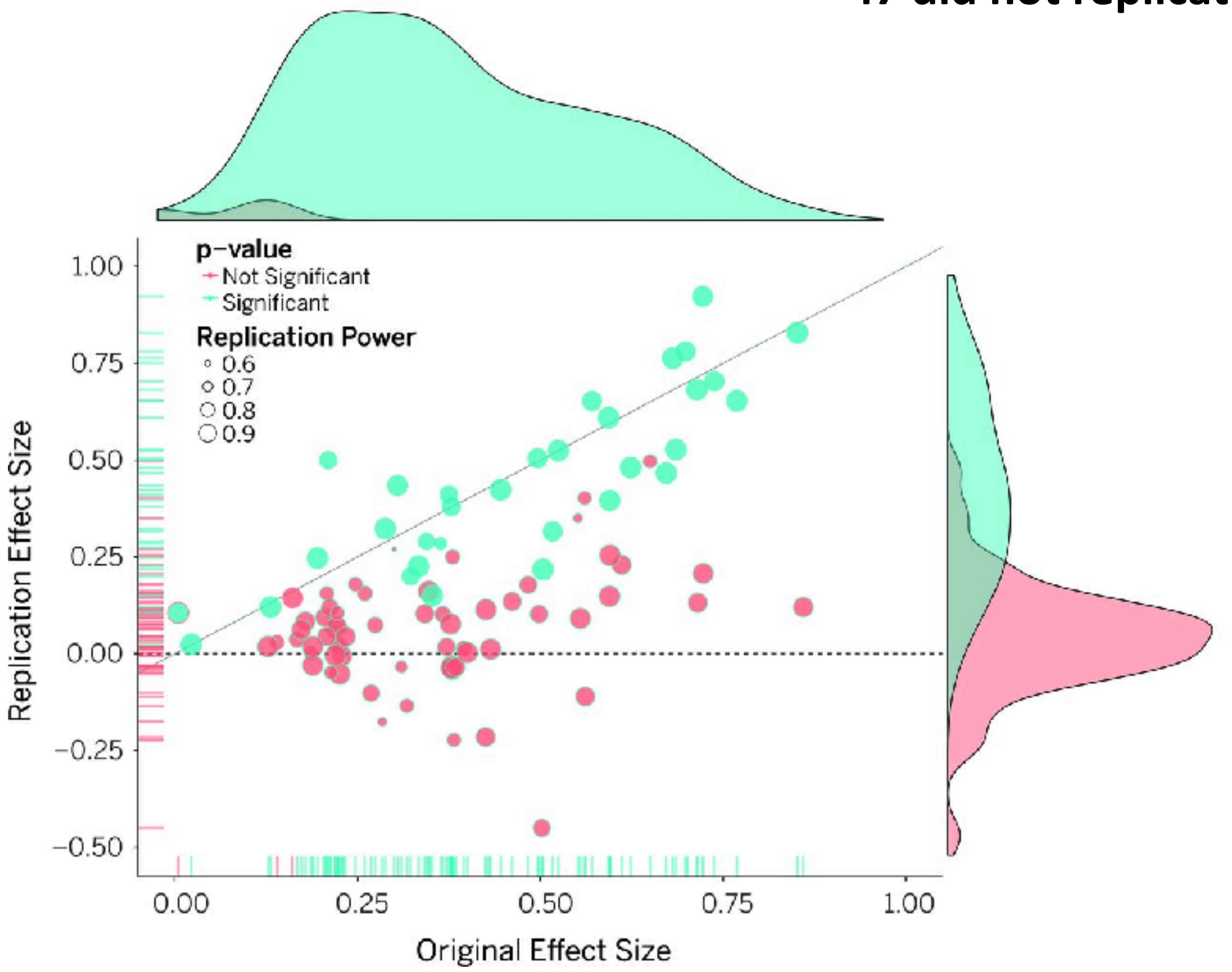
Science | Wed Mar 28, 2012 7:09pm BST

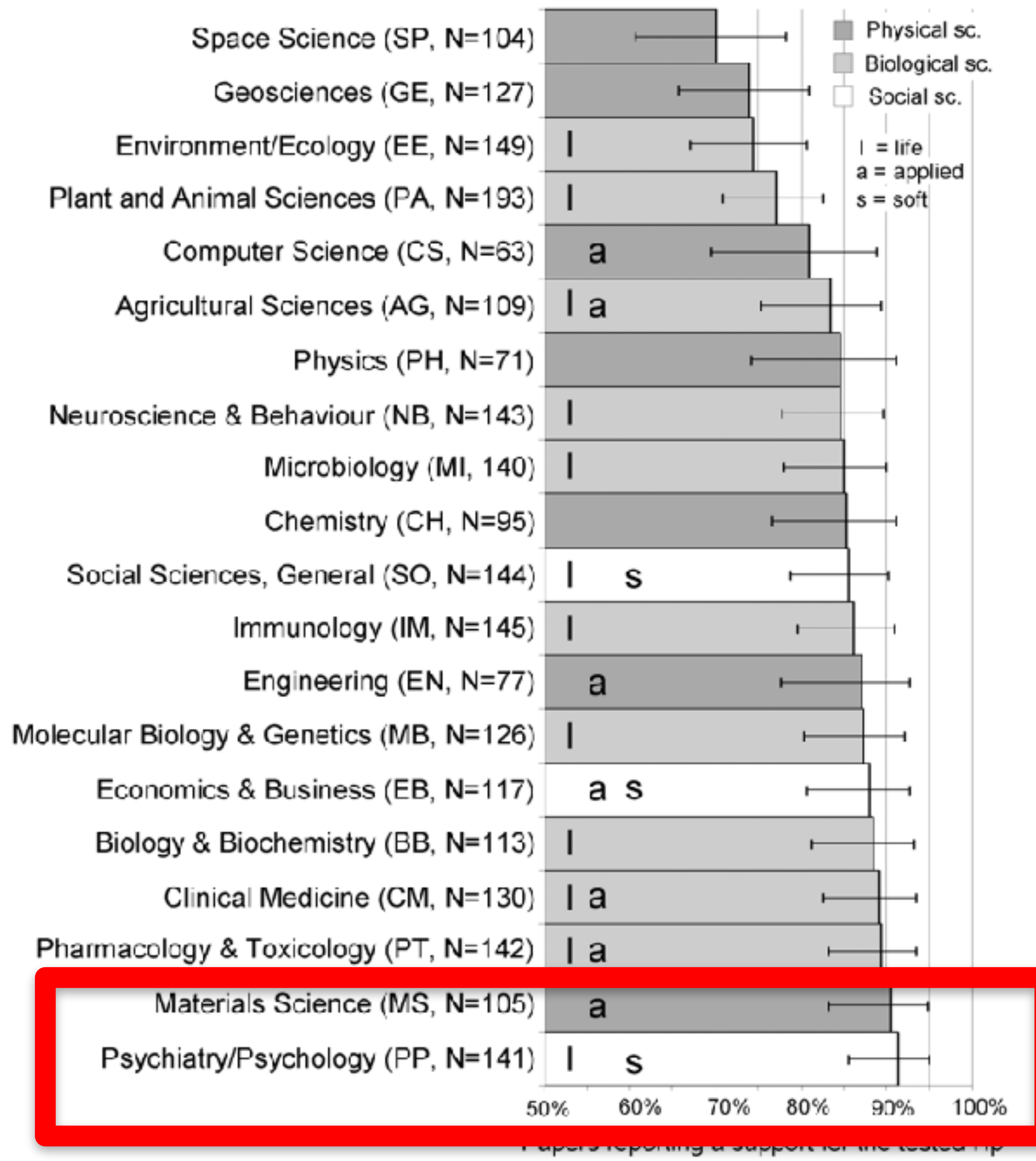
In cancer science, many "discoveries" don't hold up

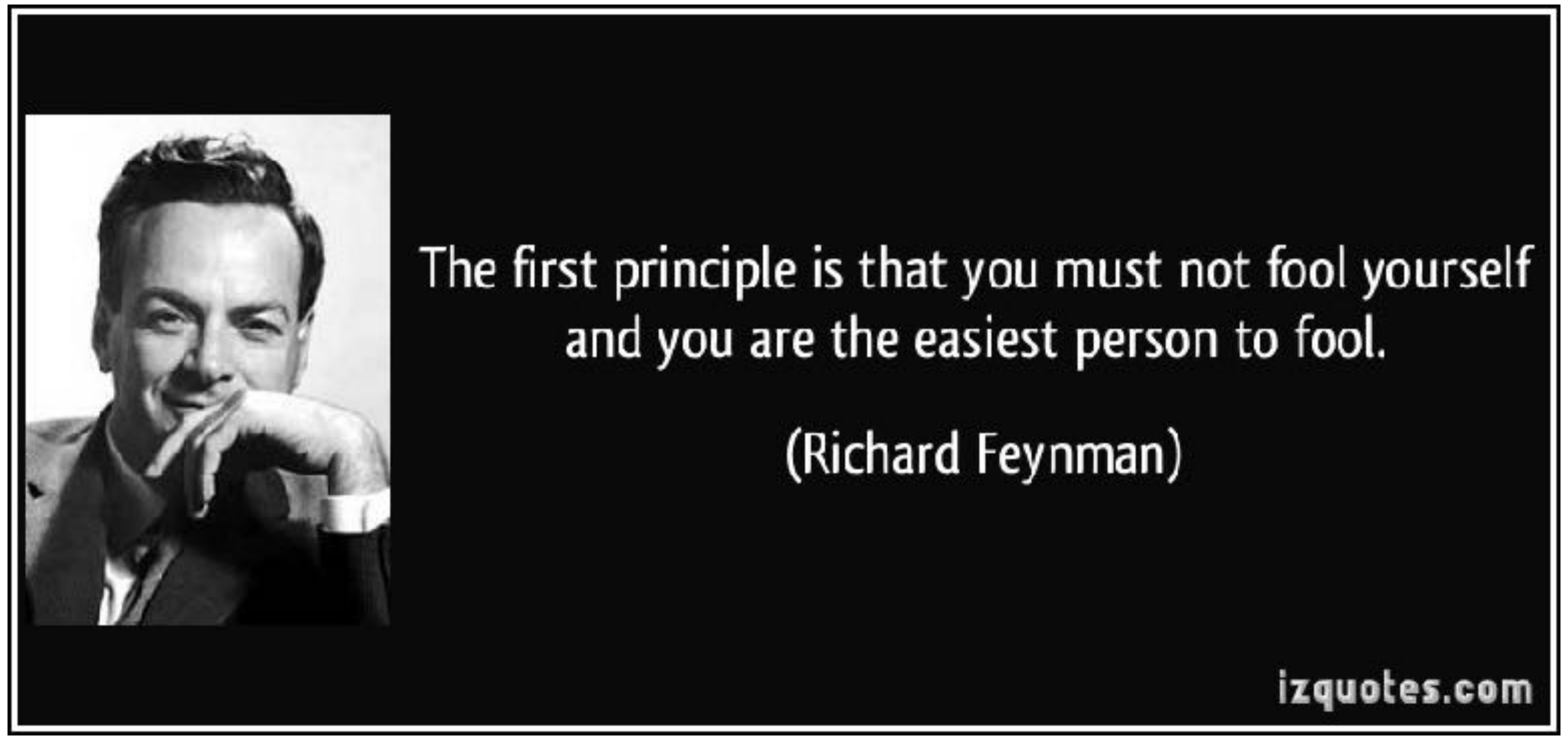
NEW YORK | BY SHARON BEGLEY

53 landmark papers on cancer

47 did not replicate



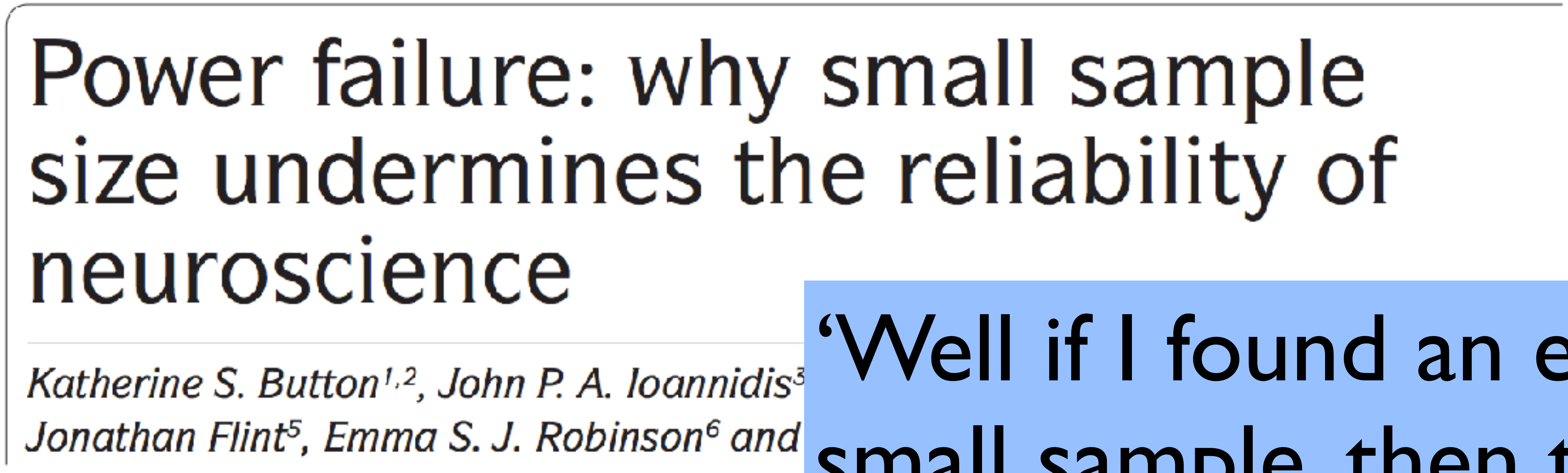




The first principle is that you must not fool yourself and you are the easiest person to fool.

(Richard Feynman)

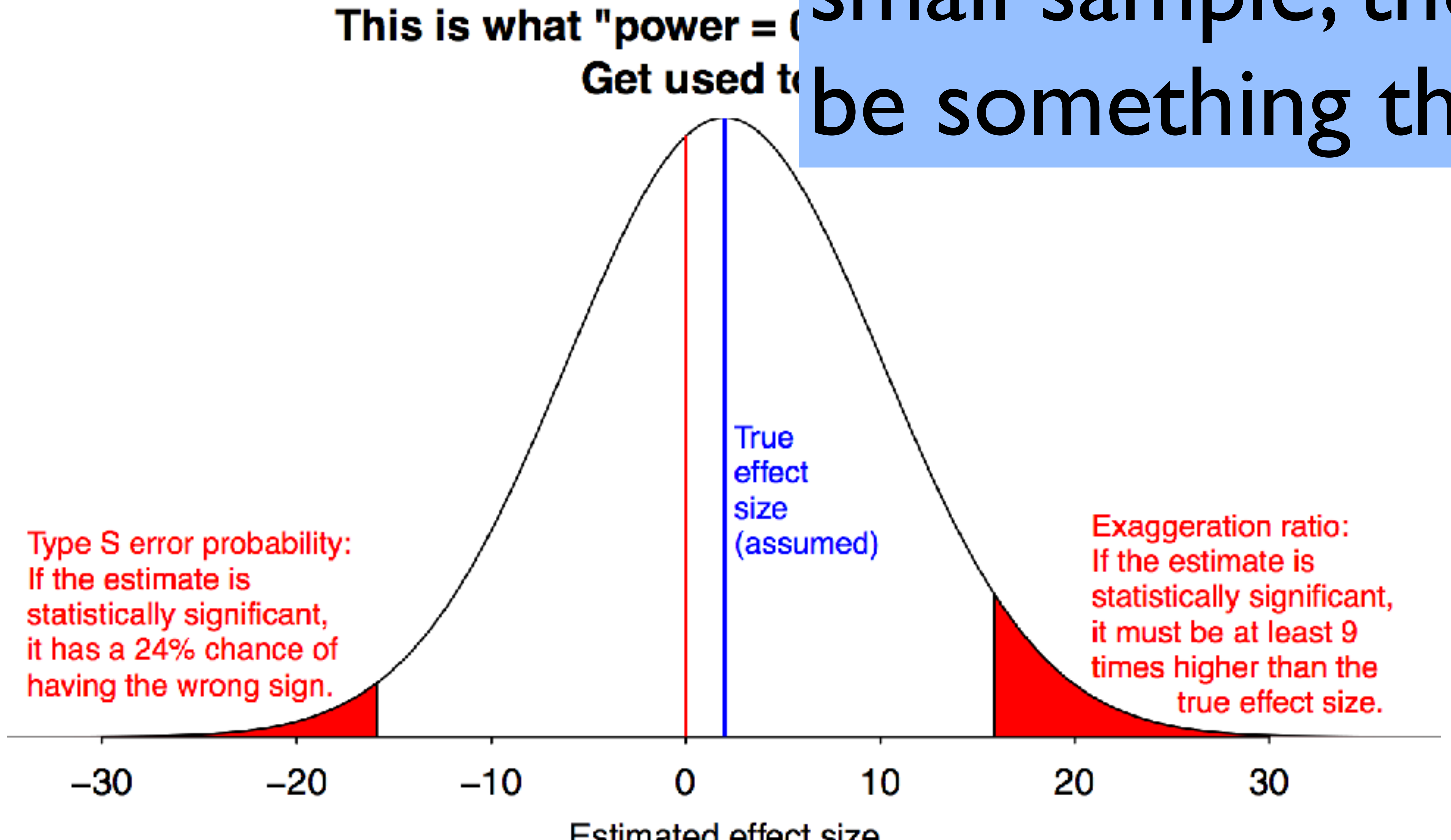
Problem 2: Type S and M errors



Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3]
Jonathan Flint[5], Emma S. J. Robinson[6] and

‘Well if I found an effect
small sample, then there
be something there righ



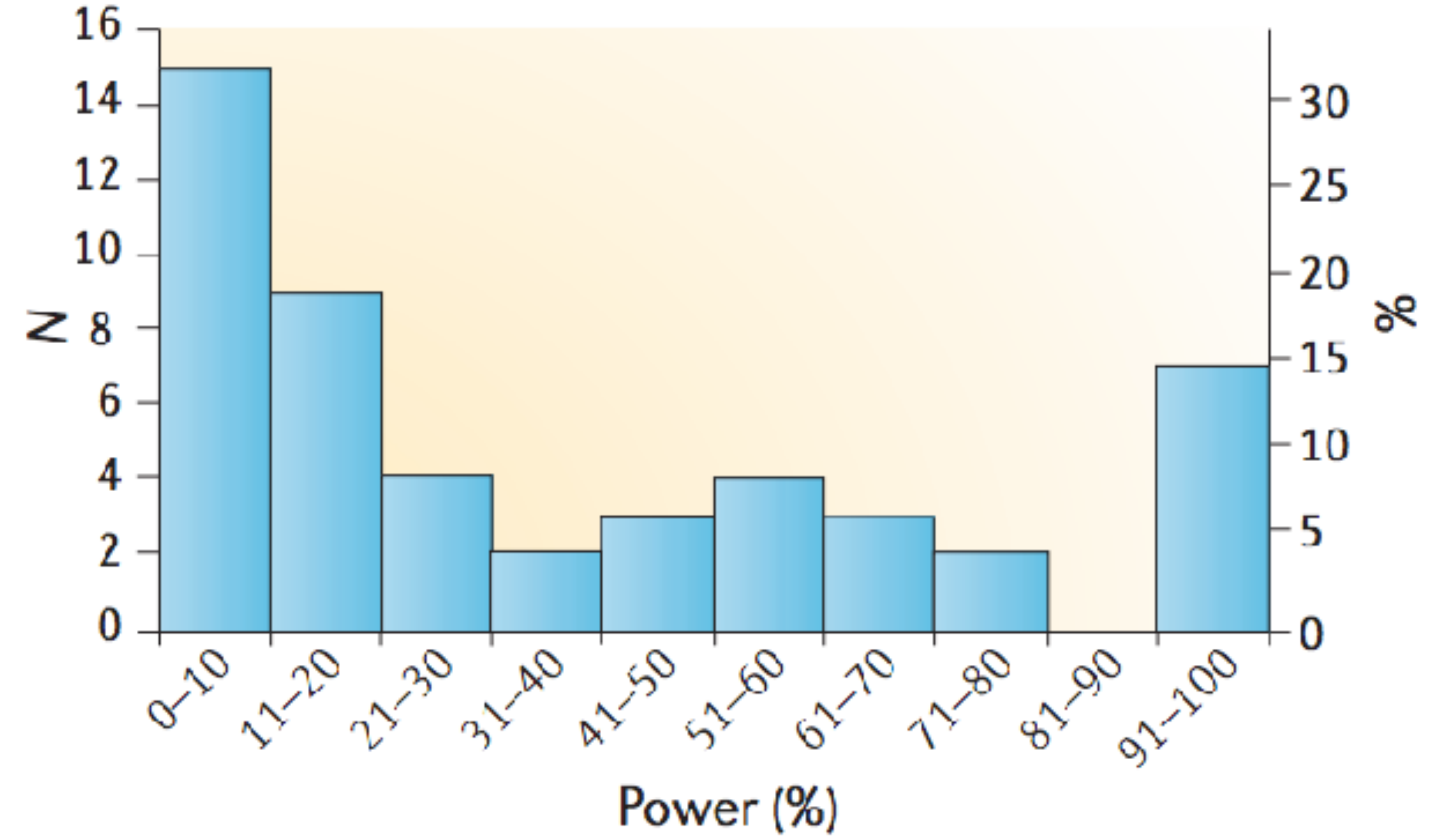
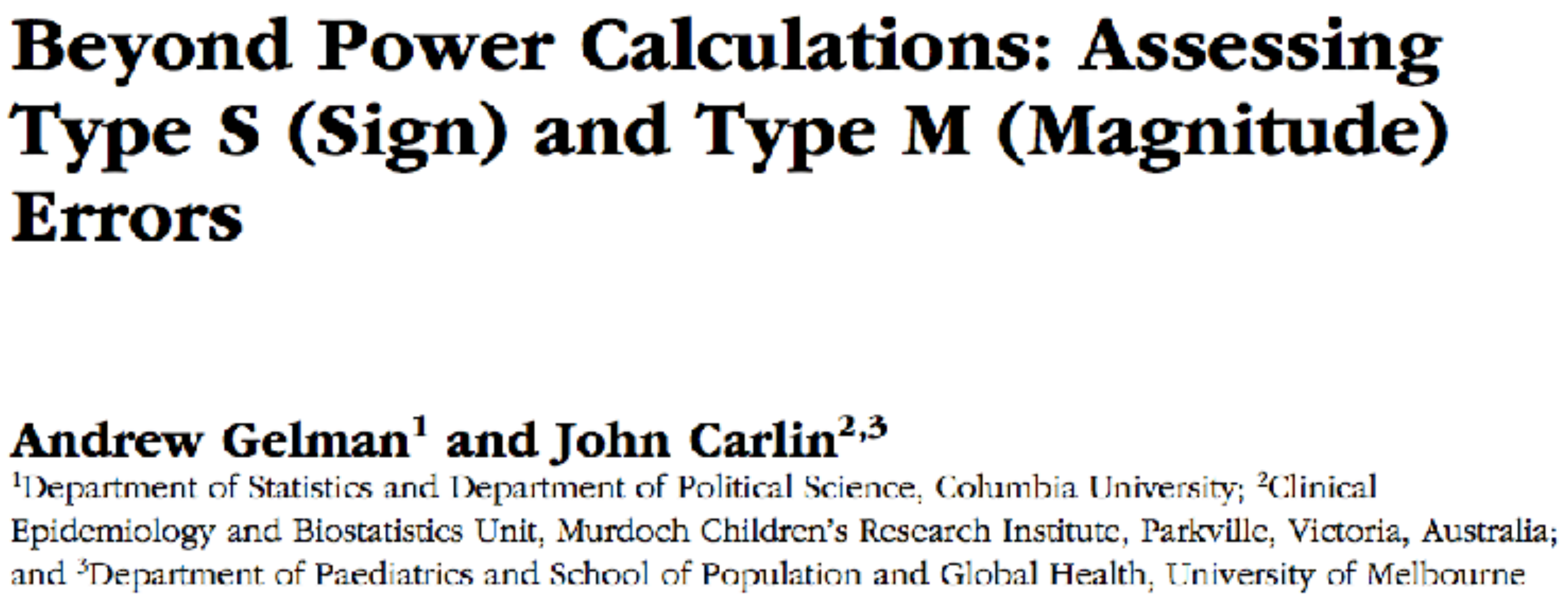


Figure 3 | **Median power of studies included in neuroscience meta-analyses.** The figure shows a histogram of median study power calculated for each of the $n = 49$ meta-analyses included in our analysis, with the number of meta-analyses (N) on the left axis and percent of meta-analyses (%) on the right axis. There is a clear bimodal distribution; $n = 15$ (31%) of the meta-analyses comprised studies with median power of less than 11%,



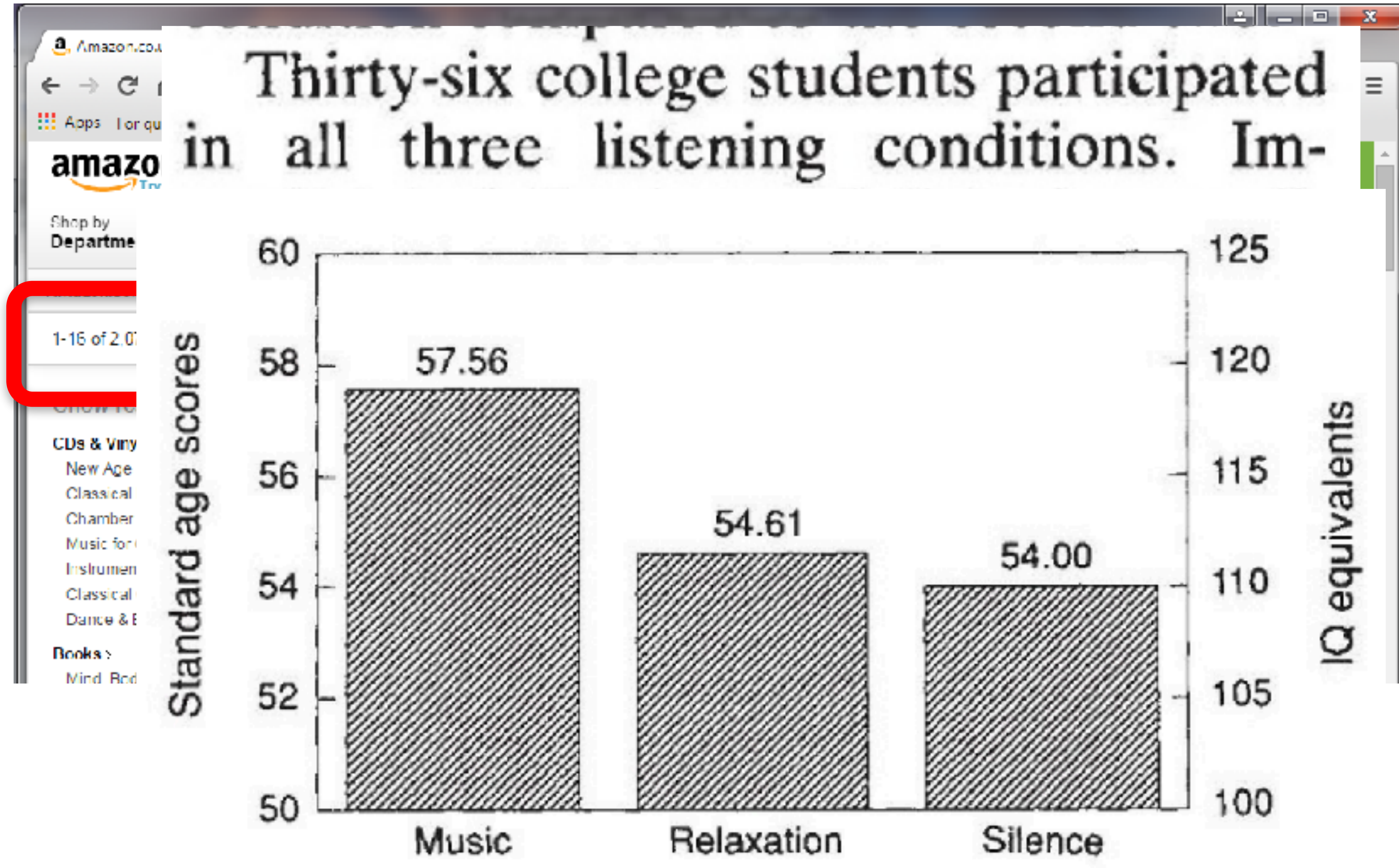
Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors

Andrew Gelman[1] and John Carlin[2,3]

[1]Department of Statistics and Department of Political Science, Columbia University; [2]Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Parkville, Victoria, Australia; and [3]Department of Paediatrics and School of Population and Global Health, University of Melbourne

‘But science is self-correcting’

- Listening to Mozart for 10 minutes leads to increase in spatial IQ of nine points (!)
  - Statewide funding scheme in Georgia (Cromie, 1999)
  - Trademark applications (Campbell, 1997)
  - Except… It’s not true
  - Dozens of failed, high power replications
- Comprehensive Soldier Fitness
  - Positive psychology training program
  - Weak studies, criticism absent
  - US army invested 125 million dollars
  - No empirical effect (small negative)



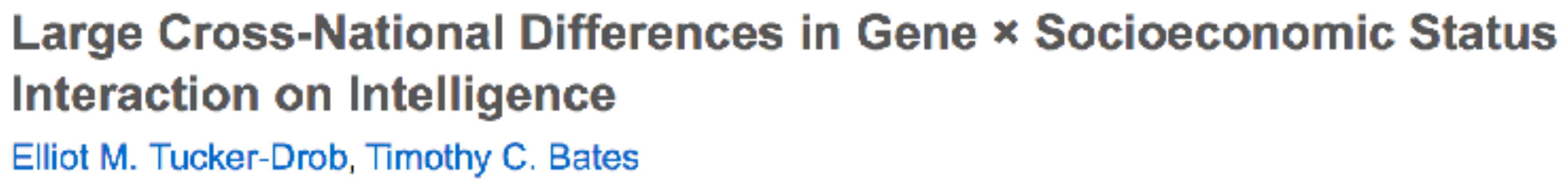
Thirty-six college students participated in all three listening conditions. Im-

Rauscher, F. H., Shaw, G. L., & Ky, K. N. (1993). Music and spatial task performance. *Nature*, (365), 611.

~~The Solution~~
a, partial, incremental solution

- If suitable for your research questions:
- 1) Use large datasets (when possible)
- 2) Preregister (when possible)
- 3) Embed in other open science practices (as much as you can)

Benefits



Large Cross-National Differences in Gene × Socioeconomic Status Interaction on Intelligence

Elliot M. Tucker-Drob, Timothy C. Bates

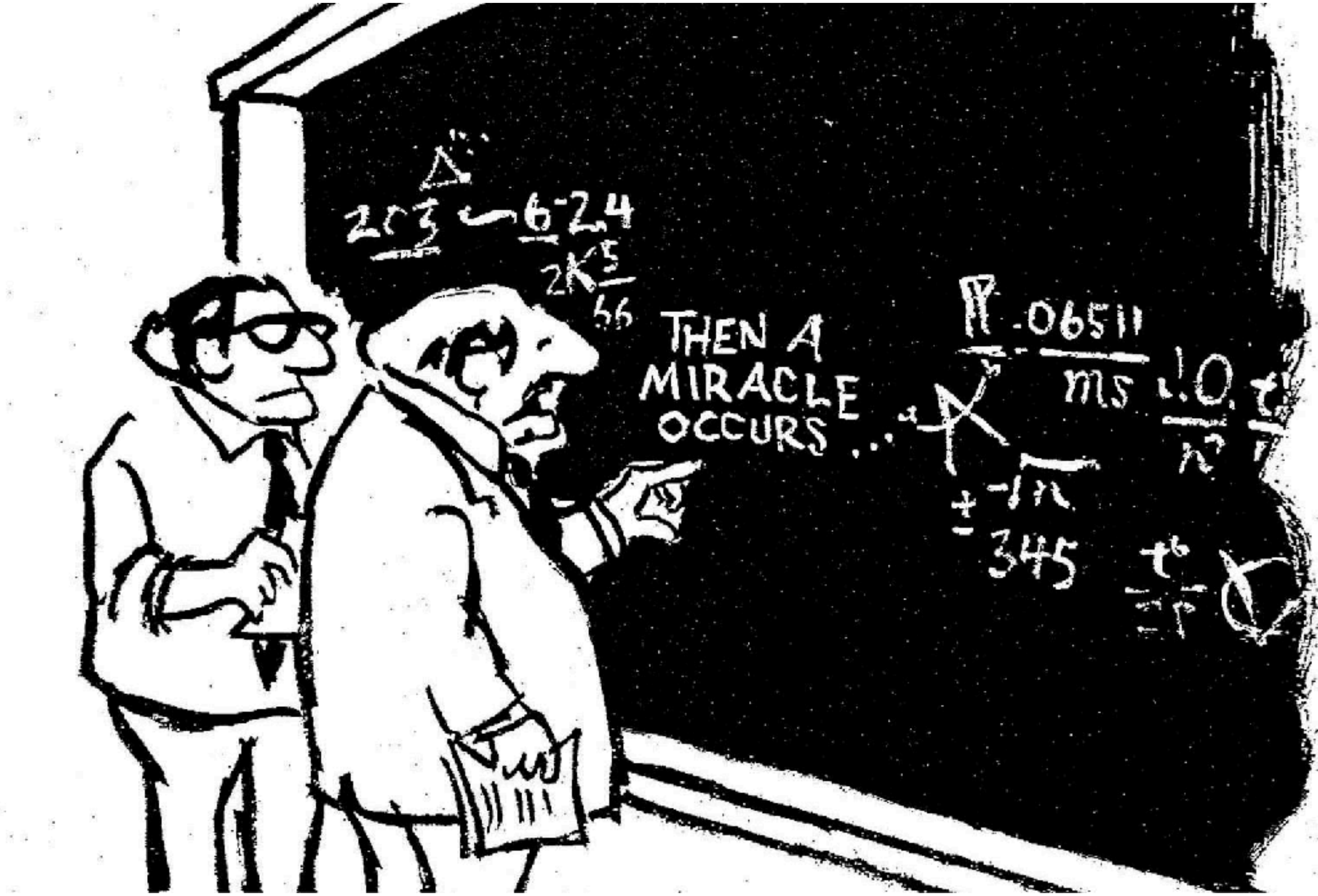
- 1) It's (almost) free
- 2) More statistical power is always better
- 3) Improve generalizability (across samples, countries etc.)
- 4) You can integrate with many/all OS practices
- 5) Scientifically: Think of new questions (Do effects vary by country? Age?)

Practice

LARGE DATASETS

Procedure

- 1) Find suitable data
- 2) Apply
- 3) Wait
- 4) (Wait some more)
- 5) Data!



Open data types:
Databases (cognitive neuro)

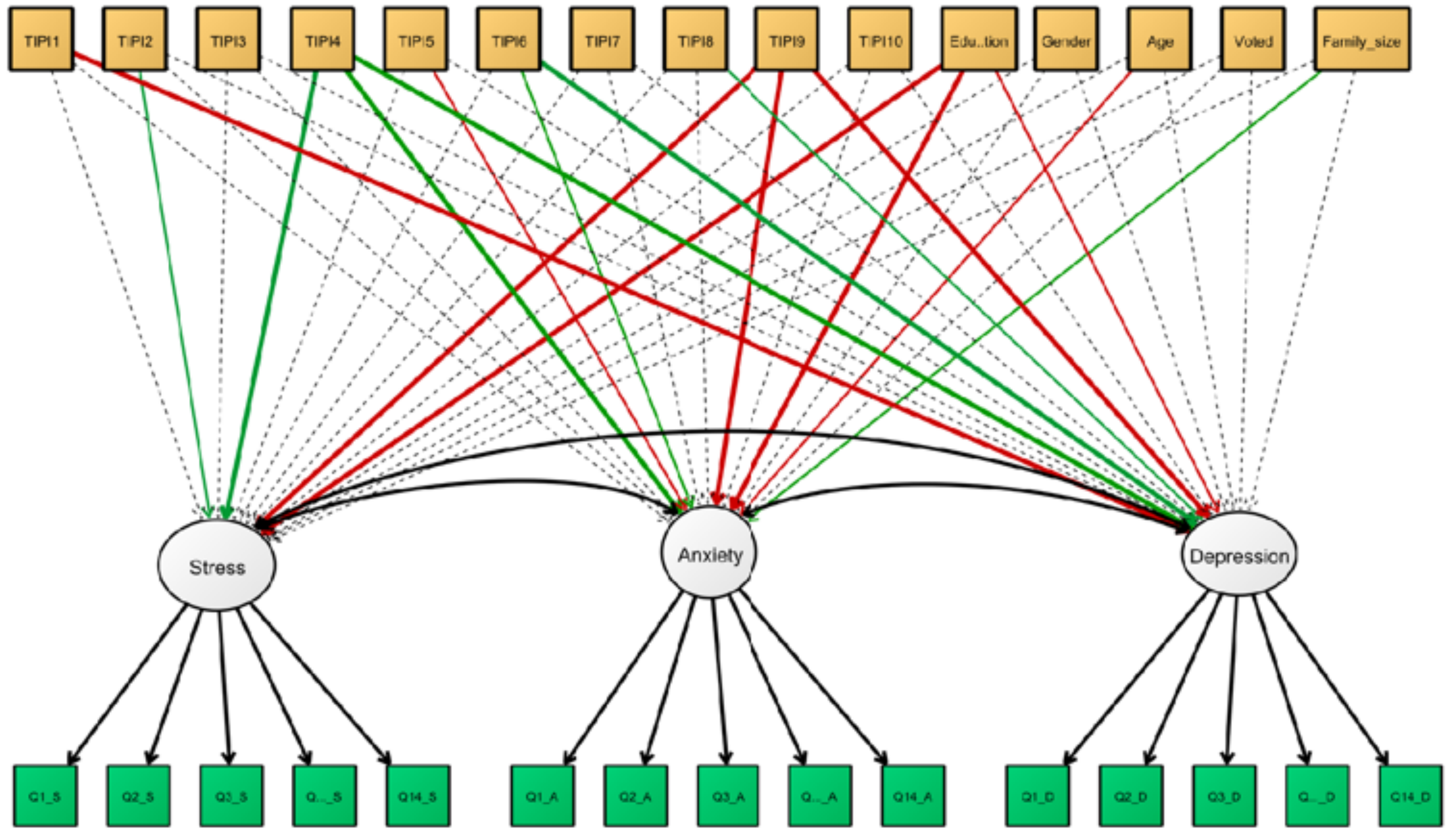
	Sample size	cost	age	data
Biobank	500.000	2000 £	43-73	everything
ABCD	10000	free	9-11	cognitive, neural, mental health
HCP	1000	<£1000	21-35	cognitive, neural, mental health
IMAGEN	2000	free	14-16	cognitive, behavioural, mental health, neural
PNC	800	free	11–17	cognitive, behavioural, neural
NKI Rockland	800	free	6-18	cognitive, behavioural, some neural

Reach out online/ email people/ Google data

OASIS, ADNI, HABS, ENIGMA, and many more

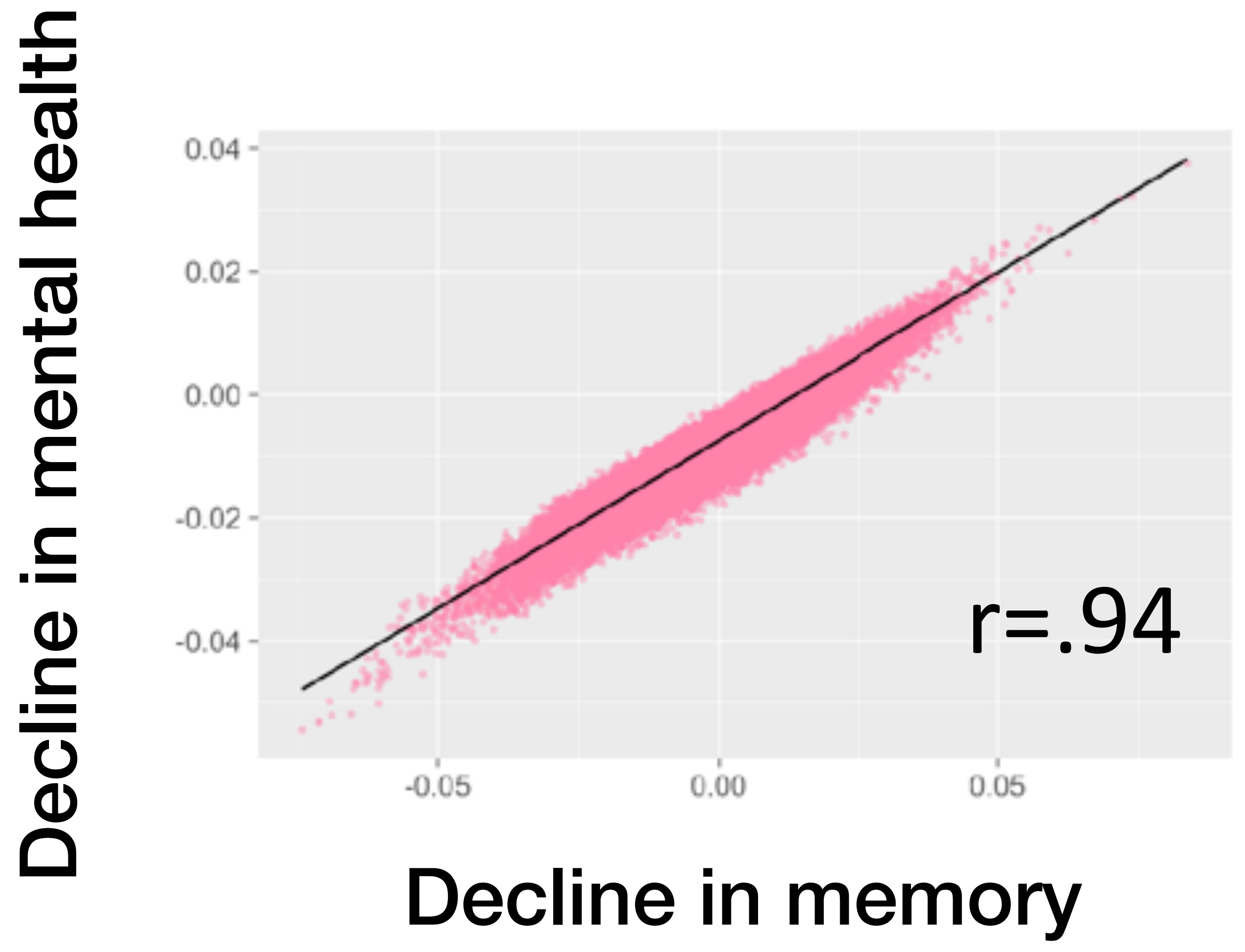
Databases (behavioural)

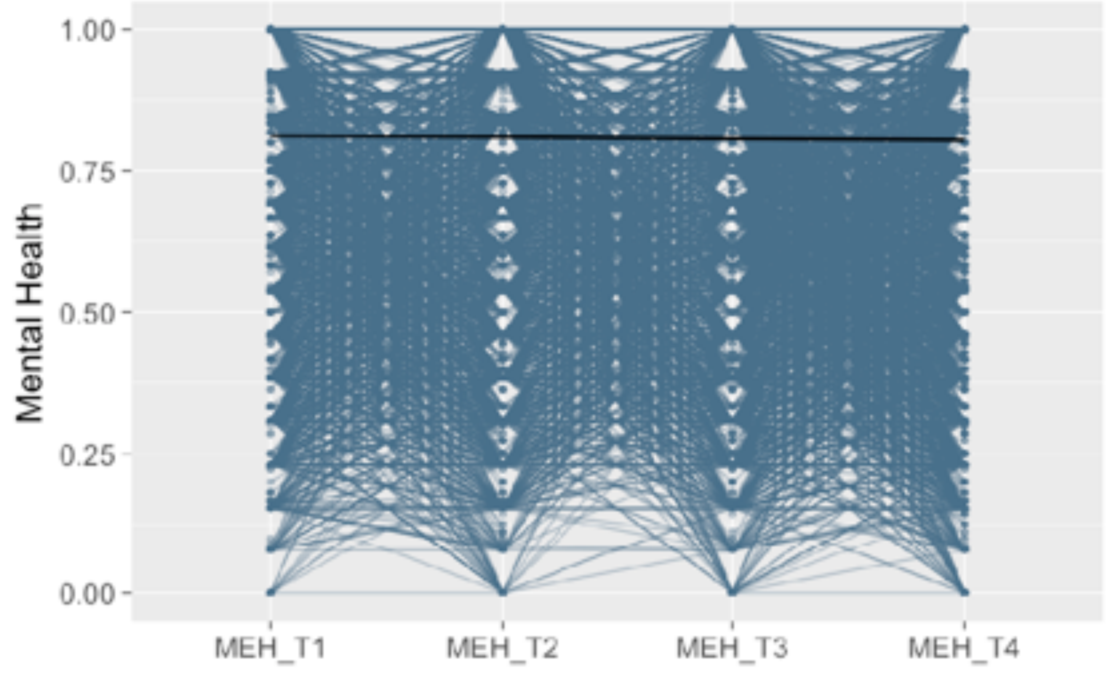
- Openpsychometrics (free)
- e.g.: Stress, anxiety, depression (DASS)
- N=48.000 in 5 seconds
- Model fit excellently
- Covariates explained >50% (!) of the variance in depression/anxiety/stress

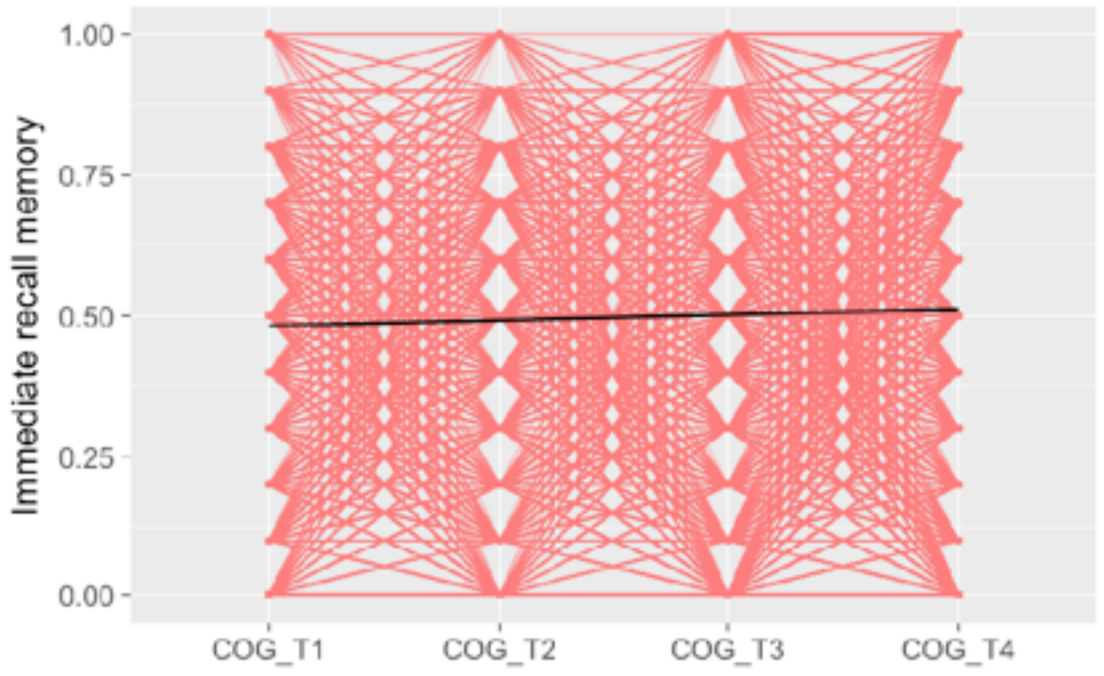


SHARE
Survey of Health, Ageing and Retirement in Europe
50+ in Europe

- Freely and easily available
- N=111.000 (!) in 60 minutes
- 6 waves



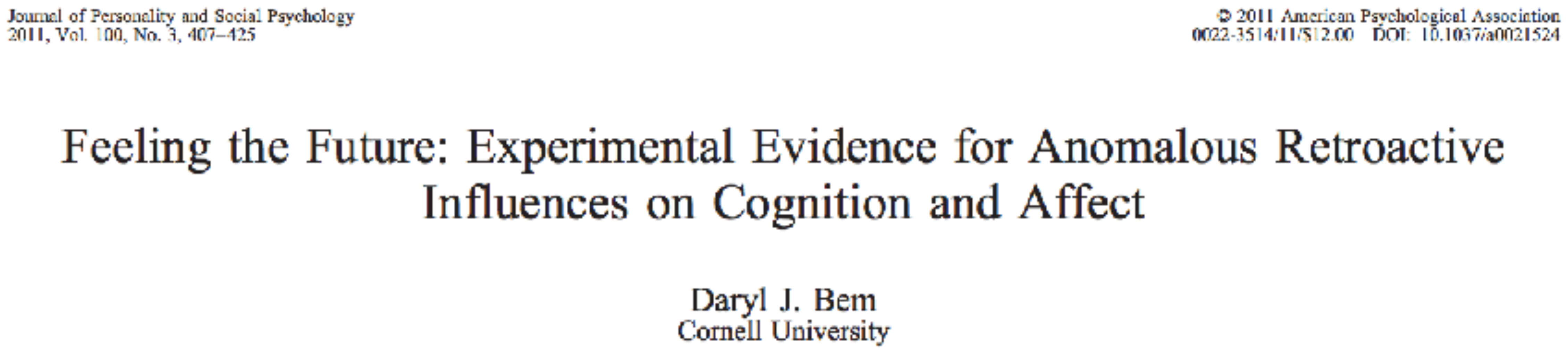




Jacobucci, R., Brandmaier, A. M., & Kievit, R. A. (2018). Variable selection in structural equation models with regularized MIMIC Models. In press, AMPPS

What about other Open Science practices?

Pre-registration

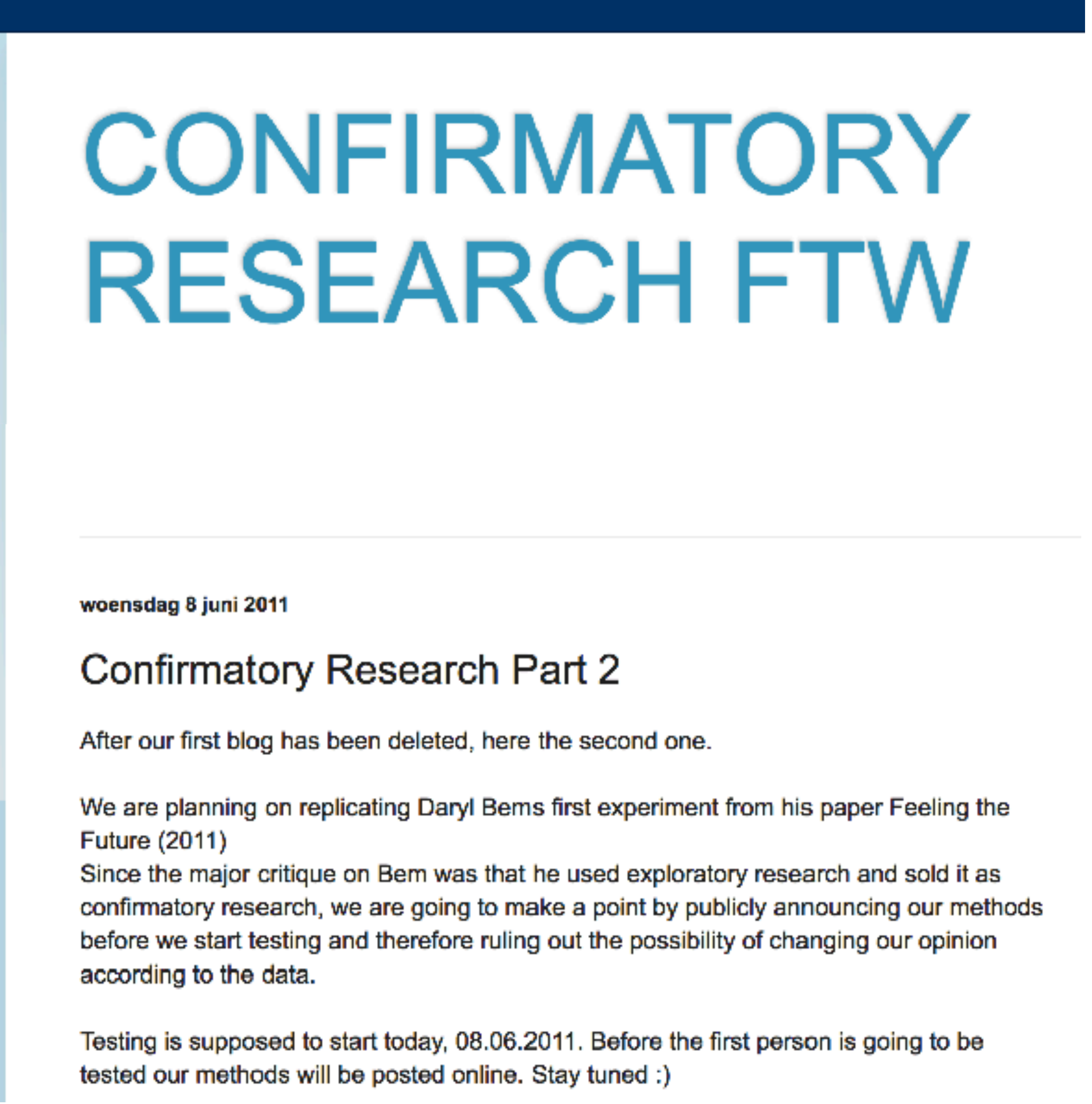


Journal of Personality and Social Psychology
2011, Vol. 100, No. 3, 407–425

© 2011 American Psychological Association
0022-3514/11/$12.00 DOI: 10.1037/a0021524

Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem
Cornell University



CONFIRMATORY RESEARCH FTW

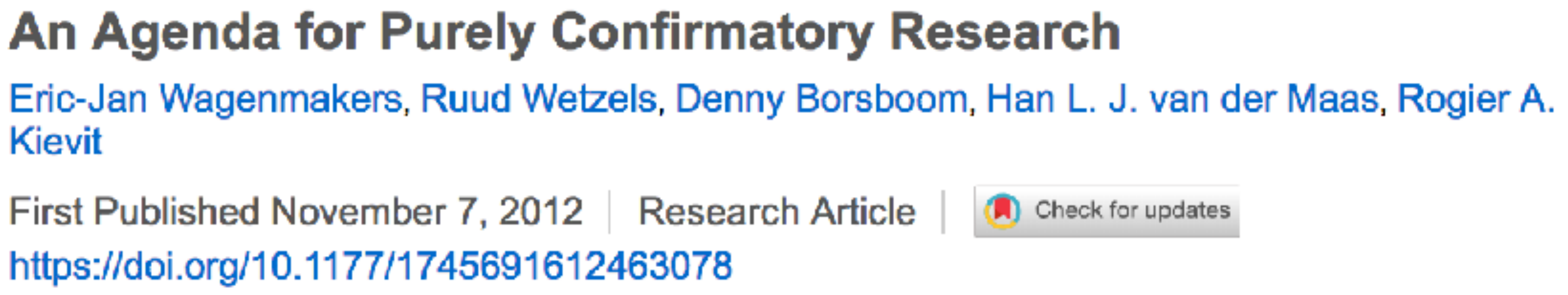
woensdag 8 juni 2011

Confirmatory Research Part 2

After our first blog has been deleted, here the second one.

We are planning on replicating Daryl Bems first experiment from his paper Feeling the Future (2011)
Since the major critique on Bem was that he used exploratory research and sold it as confirmatory research, we are going to make a point by publicly announcing our methods before we start testing and therefore ruling out the possibility of changing our opinion according to the data.

Testing is supposed to start today, 08.06.2011. Before the first person is going to be tested our methods will be posted online. Stay tuned :)

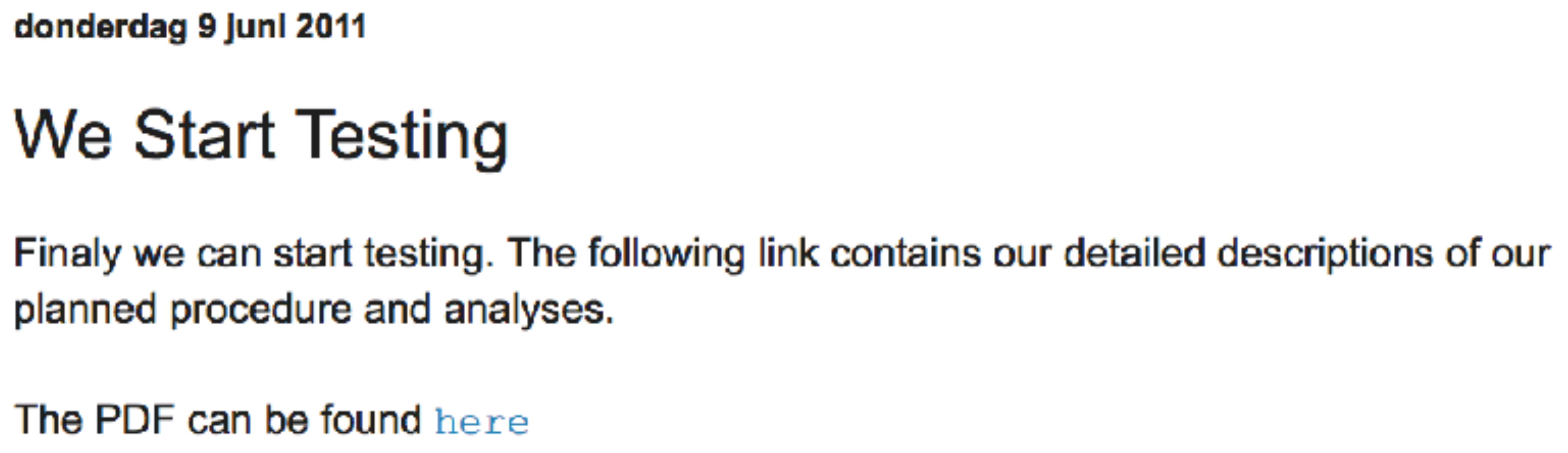


An Agenda for Purely Confirmatory Research

Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas, Rogier A. Kievit

First Published November 7, 2012 | Research Article | Check for updates

https://doi.org/10.1177/1745691612463078

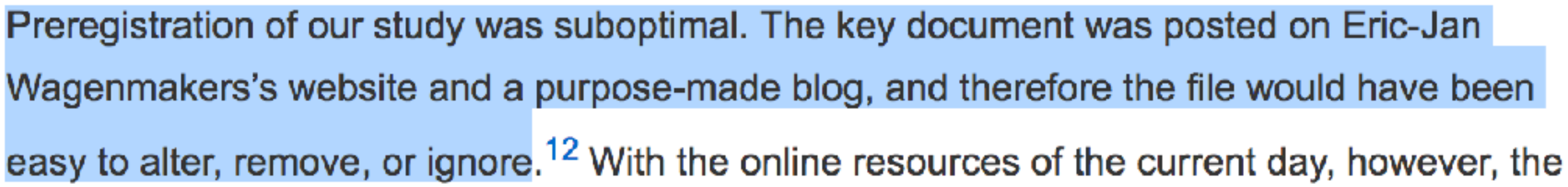


donderdag 9 juni 2011

We Start Testing

Finaly we can start testing. The following link contains our detailed descriptions of our planned procedure and analyses.

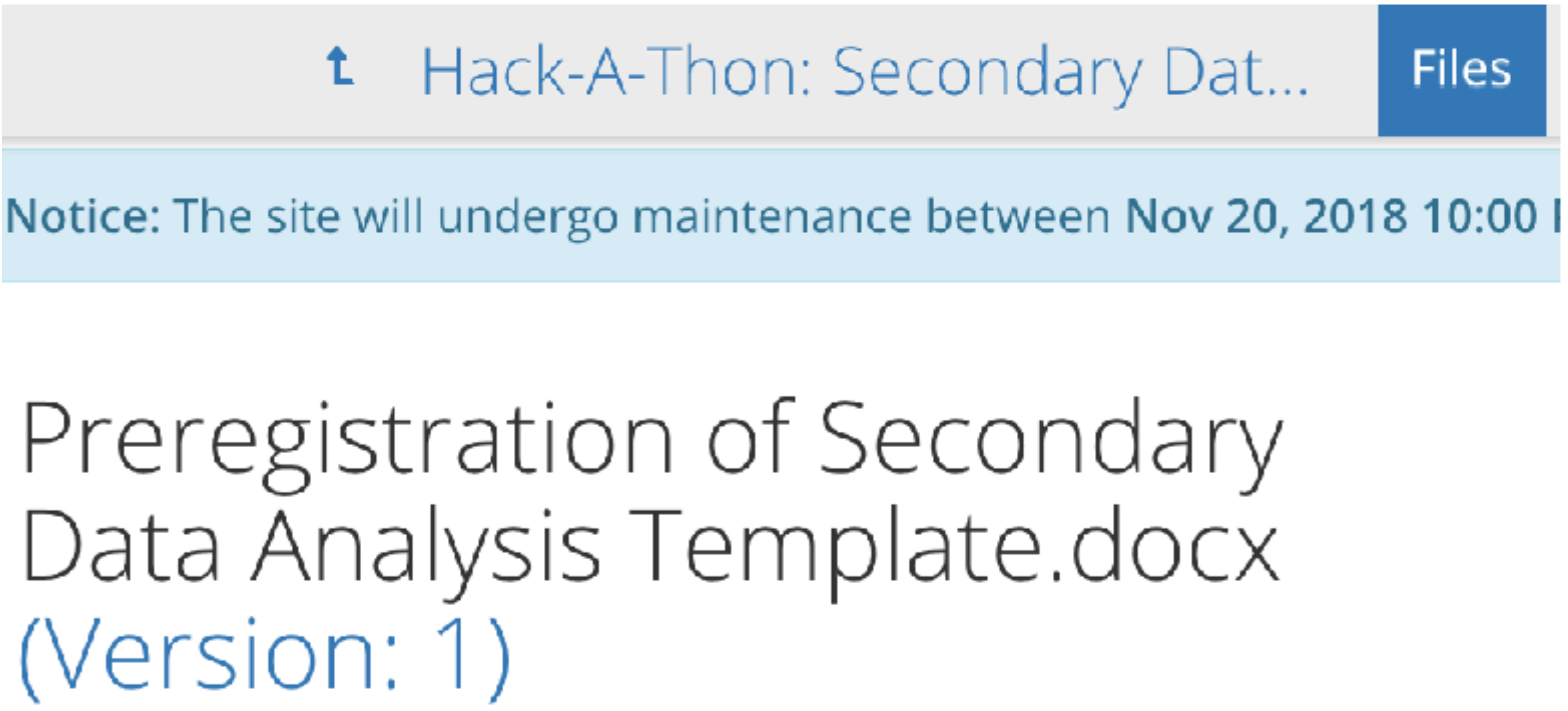
The PDF can be found here



Preregistration of our study was suboptimal. The key document was posted on Eric-Jan Wagenmakers's website and a purpose-made blog, and therefore the file would have been easy to alter, remove, or ignore.[12] With the online resources of the current day, however, the

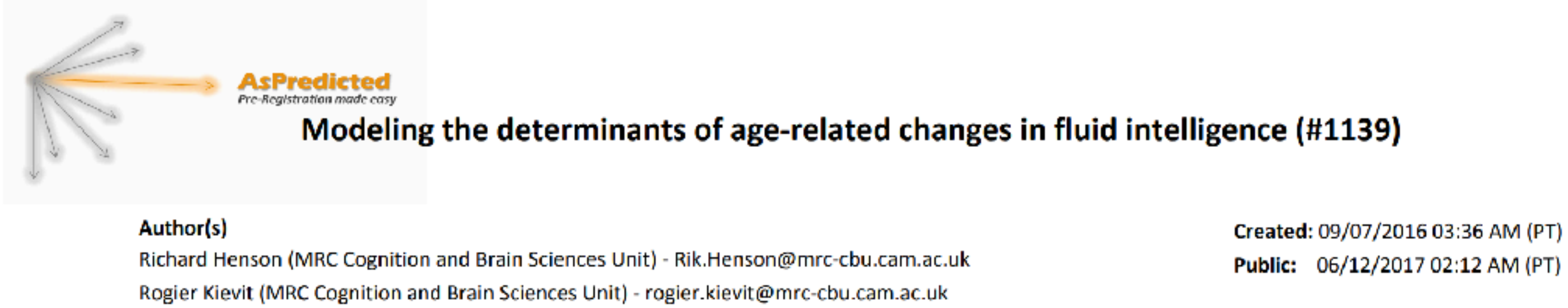
Open science and secondary data analysis

- Preregistration for secondary data: definitely worthwhile
- Emerging best practices
- Useful tool: 'aspredicted'
- Ideal: Shown access later than preregistration
- Our lab:
  - Currently 8 (imperfectly) preregistered studies on existing data



Notice: The site will undergo maintenance between Nov 20, 2018 10:00

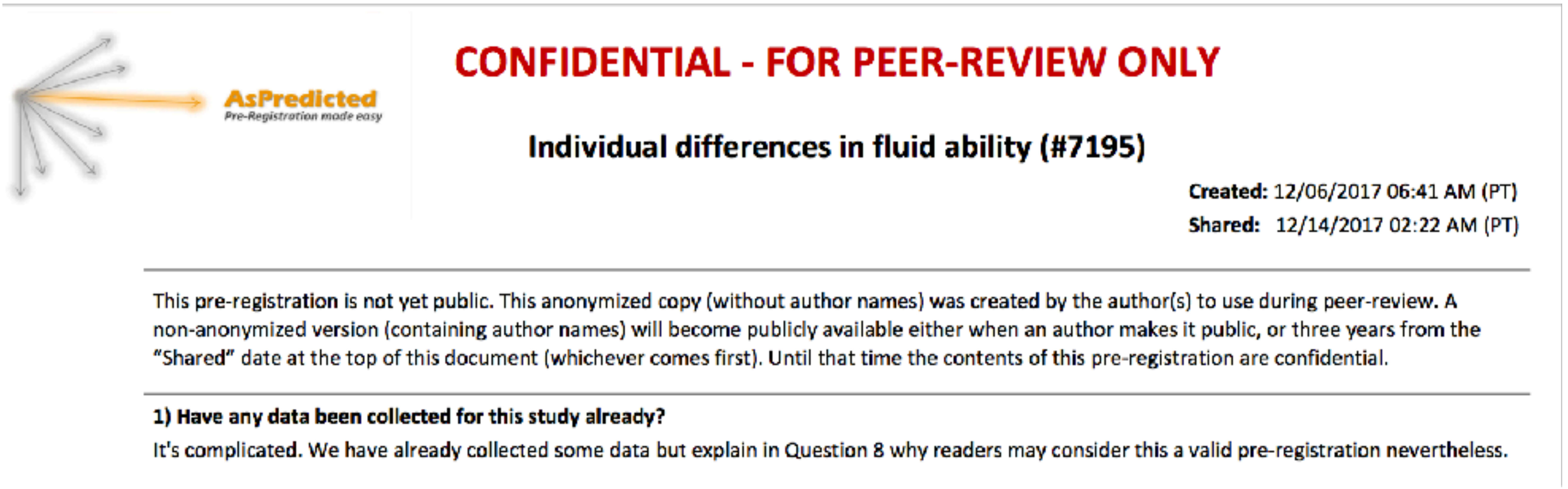
Preregistration of Secondary Data Analysis Template.docx (Version: 1)



Modeling the determinants of age-related changes in fluid intelligence (#1139)

Author(s)
Richard Henson (MRC Cognition and Brain Sciences Unit) - Rik.Henson@mrc-cbu.cam.ac.uk
Rogier Kievit (MRC Cognition and Brain Sciences Unit) - rogier.kievit@mrc-cbu.cam.ac.uk

Created: 09/07/2016 03:36 AM (PT)
Public: 06/12/2017 02:12 AM (PT)



CONFIDENTIAL - FOR PEER-REVIEW ONLY

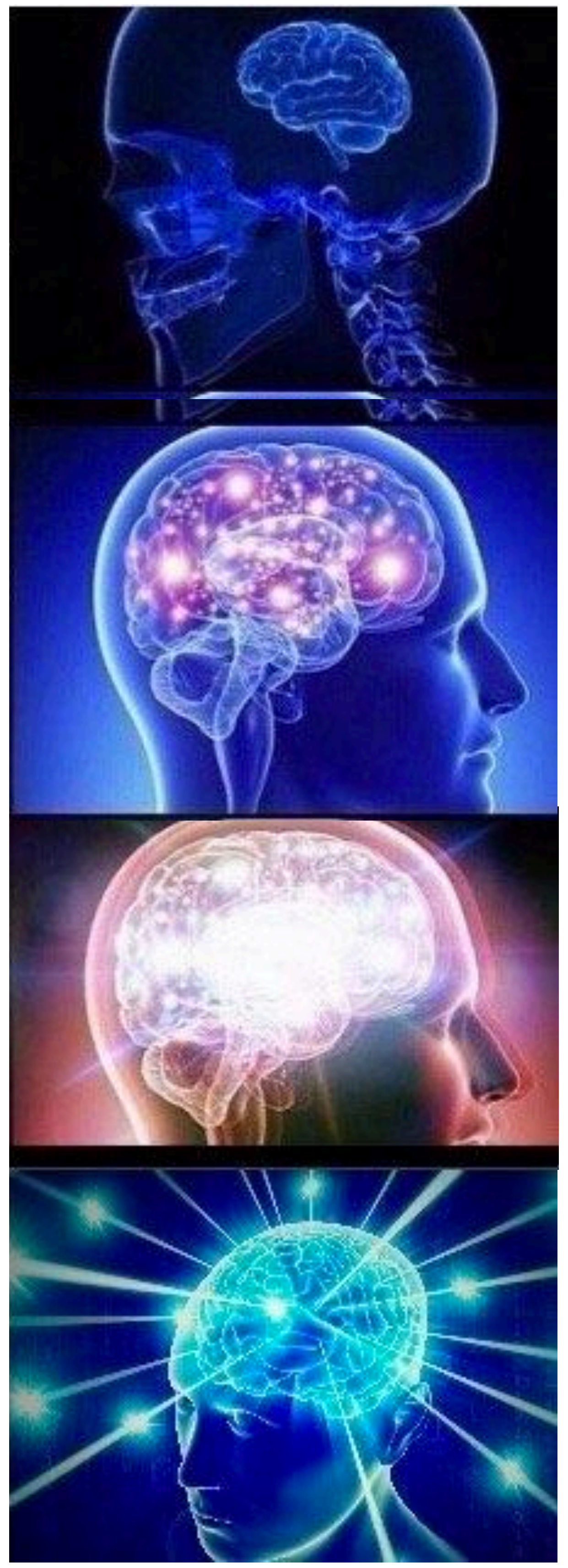
Individual differences in fluid ability (#7195)

Created: 12/06/2017 06:41 AM (PT)
Shared: 12/14/2017 02:22 AM (PT)

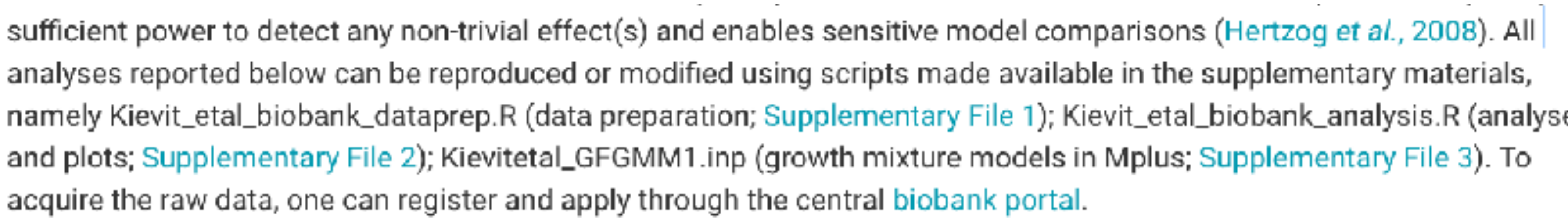
This pre-registration is not yet public. This anonymized copy (without author names) was created by the author(s) to use during peer-review. A non-anonymized version (containing author names) will become publicly available either when an author makes it public, or three years from the "Shared" date at the top of this document (whichever comes first). Until that time the contents of this pre-registration are confidential.

1) Have any data been collected for this study already?
It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.

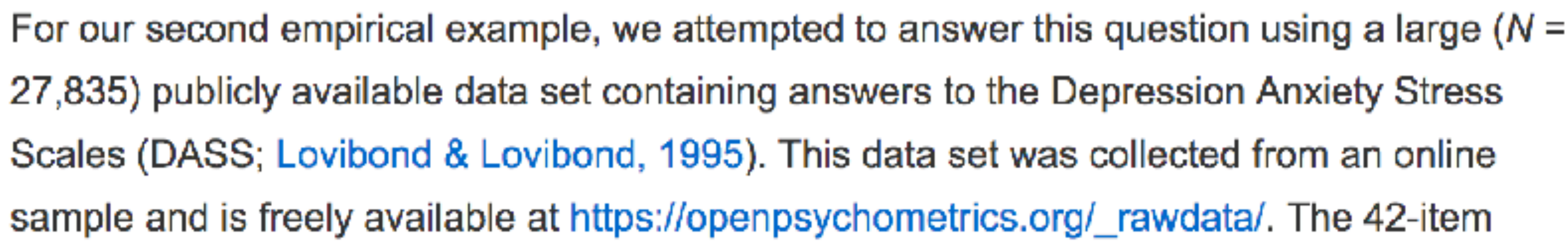
Secondary data analysis and data sharing



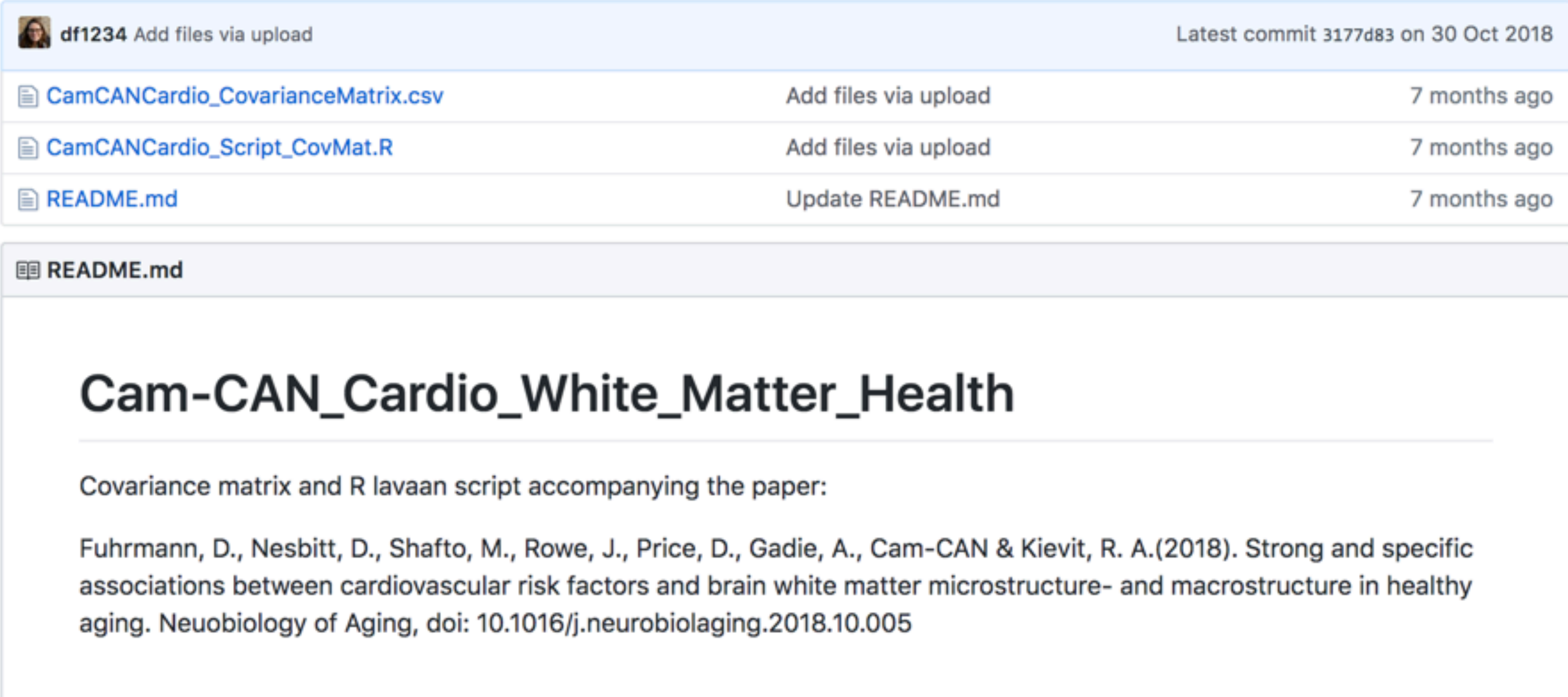
- Point to the data
- Share part of/derived the data
- Share synthetic data
- Share the actual data



sufficient power to detect any non-trivial effect(s) and enables sensitive model comparisons (Hertzog *et al.*, 2008). All analyses reported below can be reproduced or modified using scripts made available in the supplementary materials, namely Kievit_etal_biobank_dataprep.R (data preparation; Supplementary File 1); Kievit_etal_biobank_analysis.R (analyse and plots; Supplementary File 2); Kievitetal_GFGMM1.inp (growth mixture models in Mplus; Supplementary File 3). To acquire the raw data, one can register and apply through the central biobank portal.



For our second empirical example, we attempted to answer this question using a large (*N* = 27,835) publicly available data set containing answers to the Depression Anxiety Stress Scales (DASS; Lovibond & Lovibond, 1995). This data set was collected from an online sample and is freely available at https://openpsychometrics.org/_rawdata/. The 42-item



df1234 Add files via upload — Latest commit 3177d83 on 30 Oct 2018

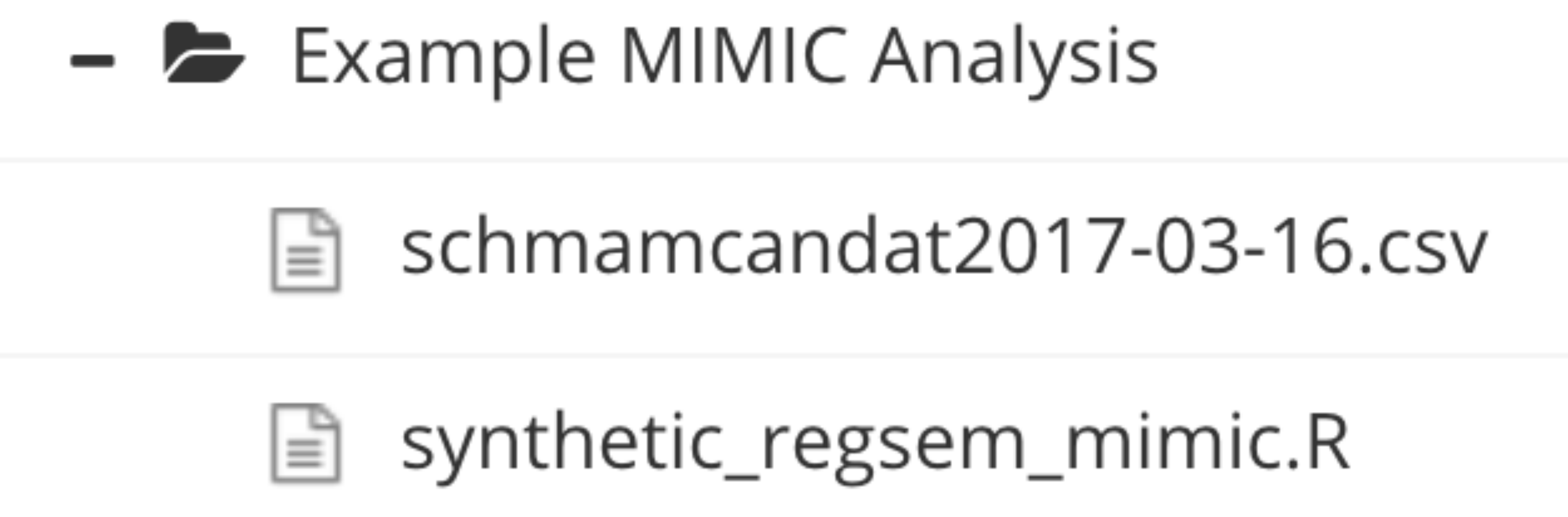
CamCANCardio_CovarianceMatrix.csv	Add files via upload	7 months ago
CamCANCardio_Script_CovMat.R	Add files via upload	7 months ago
README.md	Update README.md	7 months ago

README.md

Cam-CAN_Cardio_White_Matter_Health

Covariance matrix and R lavaan script accompanying the paper:

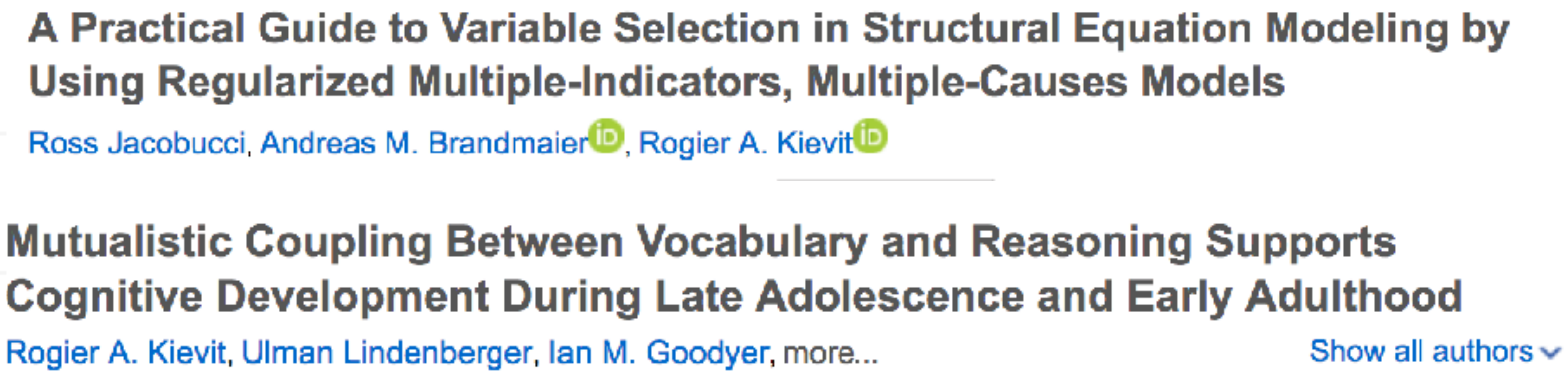
Fuhrmann, D., Nesbitt, D., Shafto, M., Rowe, J., Price, D., Gadie, A., Cam-CAN & Kievit, R. A.(2018). Strong and specific associations between cardiovascular risk factors and brain white matter microstructure- and macrostructure in healthy aging. Neuobiology of Aging, doi: 10.1016/j.neurobiolaging.2018.10.005



– Example MIMIC Analysis

schmamcandat2017-03-16.csv

synthetic_regsem_mimic.R

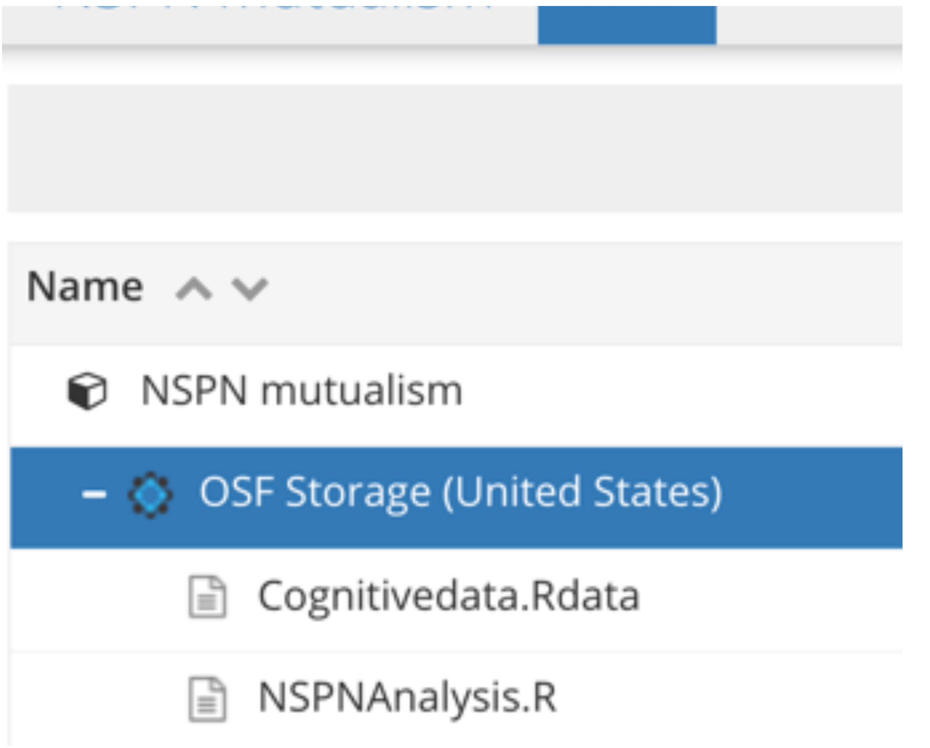


A Practical Guide to Variable Selection in Structural Equation Modeling by Using Regularized Multiple-Indicators, Multiple-Causes Models

Ross Jacobucci, Andreas M. Brandmaier, Rogier A. Kievit

Mutualistic Coupling Between Vocabulary and Reasoning Supports Cognitive Development During Late Adolescence and Early Adulthood

Rogier A. Kievit, Ulman Lindenberger, Ian M. Goodyer, more... Show all authors



Name

NSPN mutualism

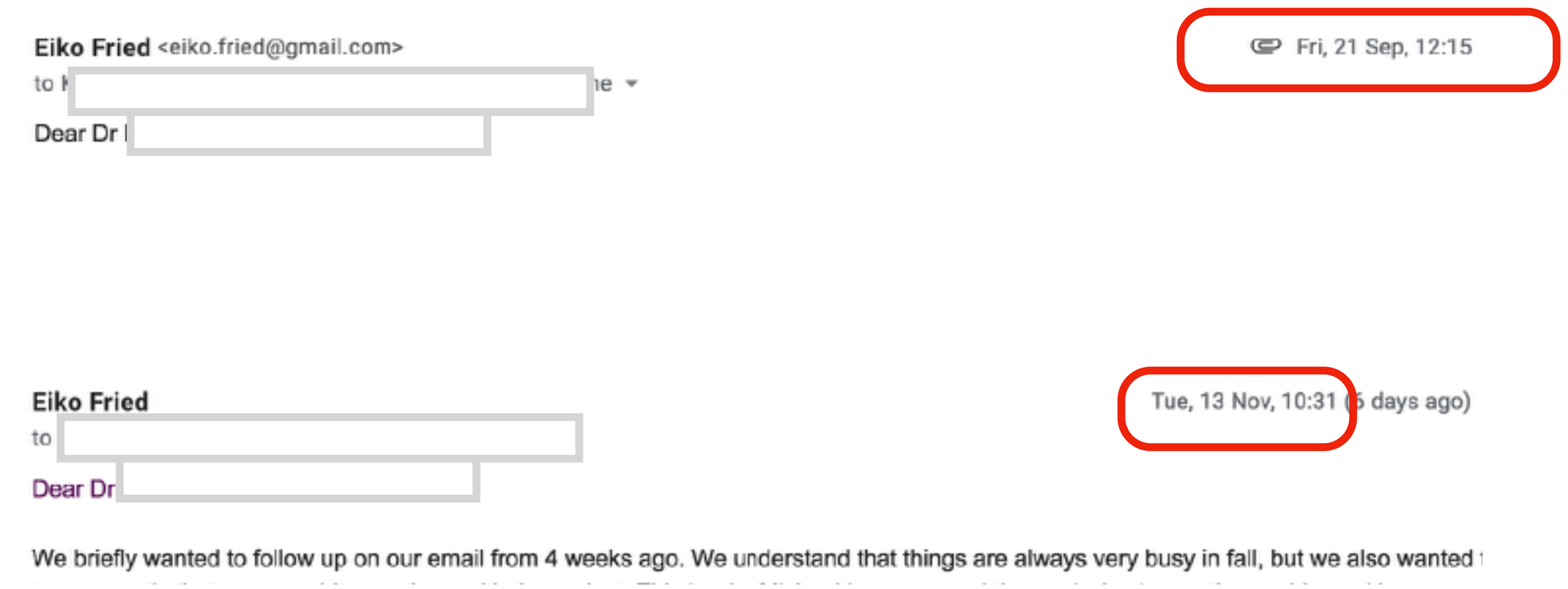
– OSF Storage (United States)

Cognitivedata.Rdata

NSPNAnalysis.R

Pitfalls

1) Getting the data
2) The data

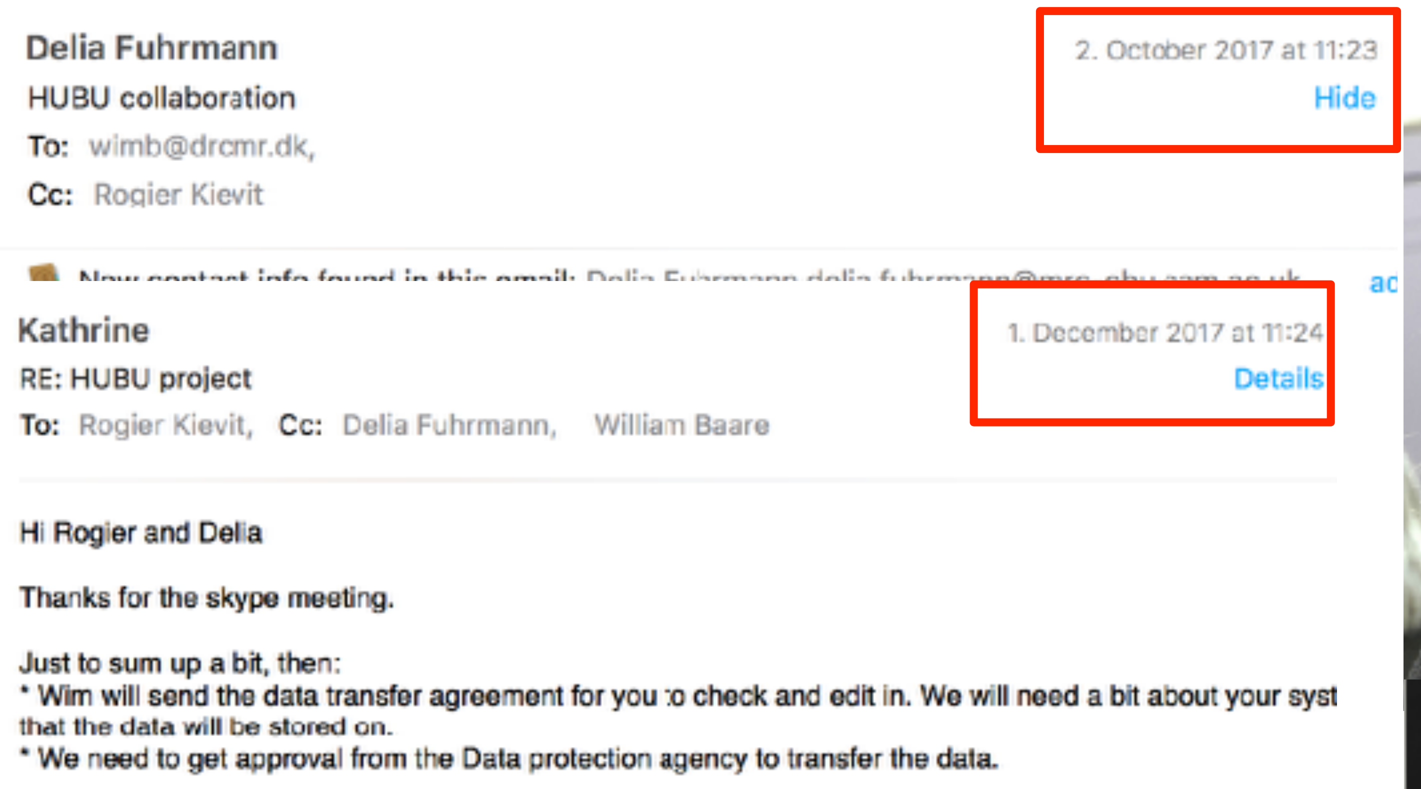


Eiko Fried <eiko.fried@gmail.com>
Fri, 21 Sep, 12:15
to [redacted] e
Dear Dr [redacted]

Eiko Fried
Tue, 13 Nov, 10:31 (6 days ago)
to [redacted]
Dear Dr [redacted]

We briefly wanted to follow up on our email from 4 weeks ago. We understand that things are always very busy in fall, but we also wanted

Hjernens
Udvikling hos
Børn og Unge



Delia Fuhrmann
HUBU collaboration
To: wimb@drcmr.dk,
Cc: Rogier Kievit
2. October 2017 at 11:23
Hide

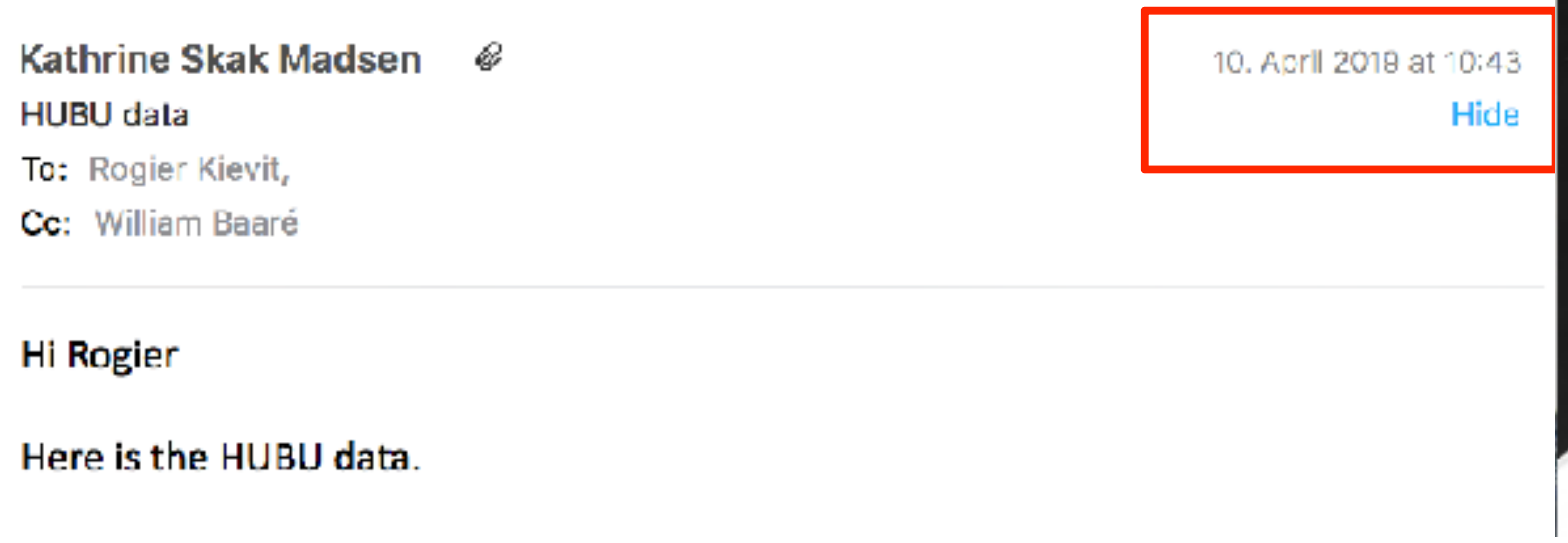
New contact info found in this email: Delia Fuhrmann delia.fuhrmann@mrc-cbu.cam.ac.uk

Kathrine
RE: HUBU project
To: Rogier Kievit, Cc: Delia Fuhrmann, William Baare
1. December 2017 at 11:24
Details

Hi Rogier and Delia

Thanks for the skype meeting.

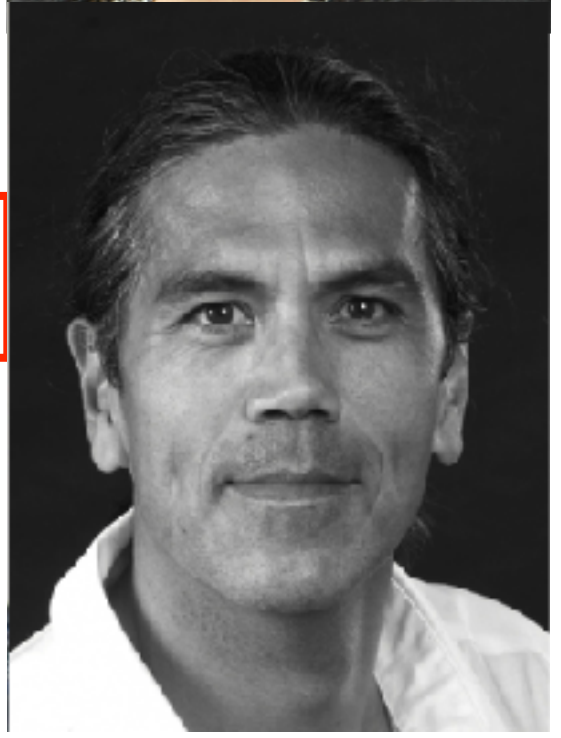
Just to sum up a bit, then:
* Wim will send the data transfer agreement for you to check and edit in. We will need a bit about your syst
that the data will be stored on.
* We need to get approval from the Data protection agency to transfer the data.



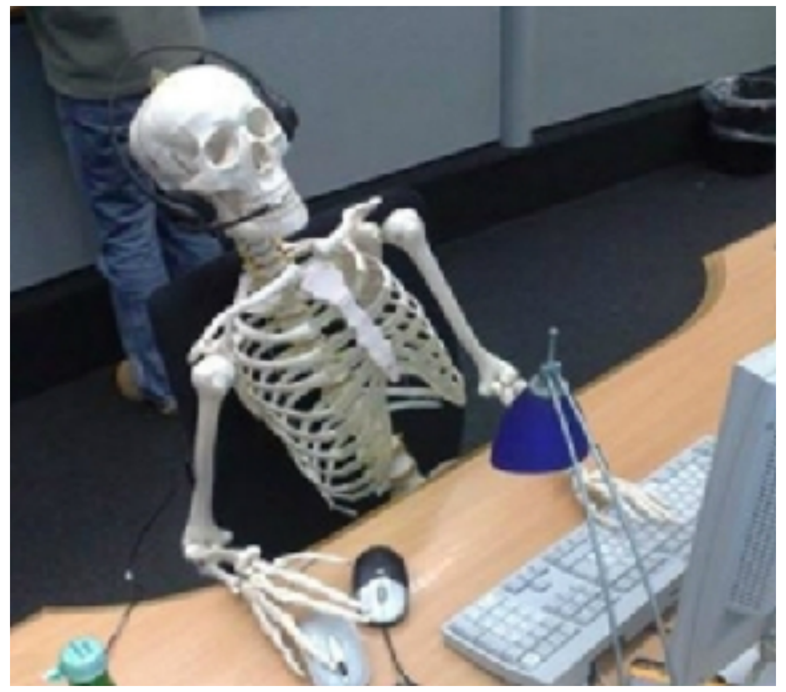
Kathrine Skak Madsen
HUBU data
To: Rogier Kievit,
Cc: William Baaré
10. April 2018 at 10:43
Hide

Hi Rogier

Here is the HUBU data.



- 1) Time
- 2) Effort
- 3) Requirements



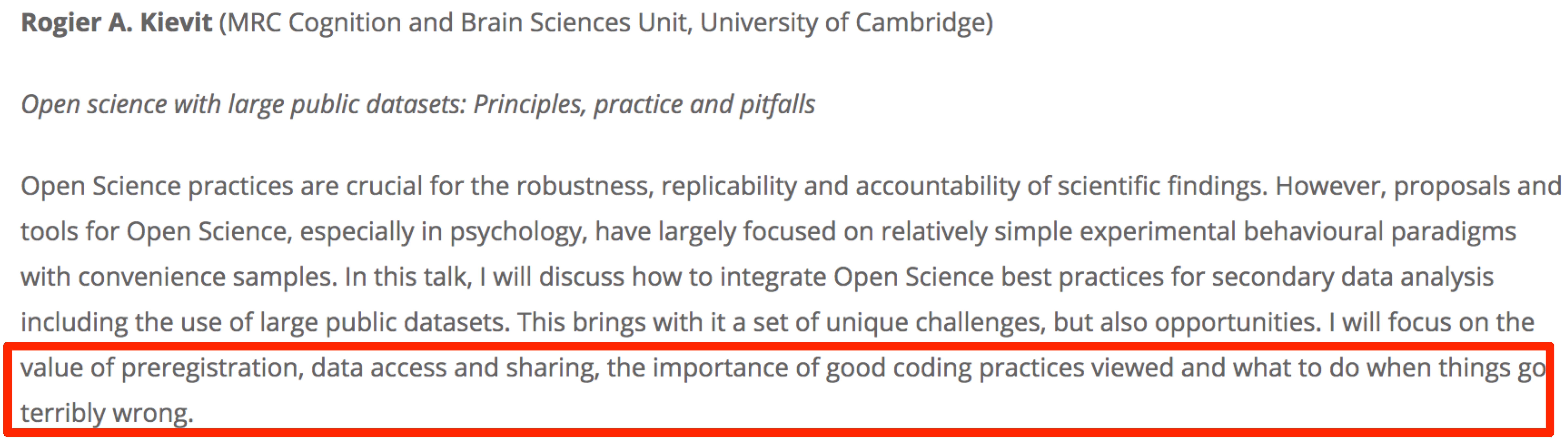
- Anyone who shares an office with you has to sign an NDA
- The computer cannot be on if anybody who has NOT signed the NDA is in the same room
- The computer with the data cannot connected to the internet or the CBU network
- You have to enter a password every time you load the data

36 emails…

10 phone calls…

3 months….

to get a single signature.



Rogier A. Kievit (MRC Cognition and Brain Sciences Unit, University of Cambridge)

Open science with large public datasets: Principles, practice and pitfalls

Open Science practices are crucial for the robustness, replicability and accountability of scientific findings. However, proposals and tools for Open Science, especially in psychology, have largely focused on relatively simple experimental behavioural paradigms with convenience samples. In this talk, I will discuss how to integrate Open Science best practices for secondary data analysis including the use of large public datasets. This brings with it a set of unique challenges, but also opportunities. I will focus on the value of preregistration, data access and sharing, the importance of good coding practices viewed and what to do when things go terribly wrong.

biobank uk
Imaging study

- Funded by Wellcome/MRC/DOH and many more
- Goals:
  - Scan 100.000
  - Phenotype/Genotype 500.000
- Strengths
  - Cognitive, lifestyle, biological sampling
  - (mental) health data: Both integrated with NHS and tailored measures
  - Total: (tens of) thousands of variables
  - *Open resource*

What does it look like?

biobank.ctsu.ox.ac.uk/crystal/
Apps
Inbox (2) - rogierki...
Google Scholar
Google Calendar
biobank uk
Data
3 Instances
Notes
4 Categories
2 Related Data-Fields
0
965

Browse by Primar

Category

- Population chara
- UK Biobank Ass
- Biological samp
- Genomics
- Online follow-up
- Additional expos
- Health-related o
- Returned datase

Summary generated 4 February 20

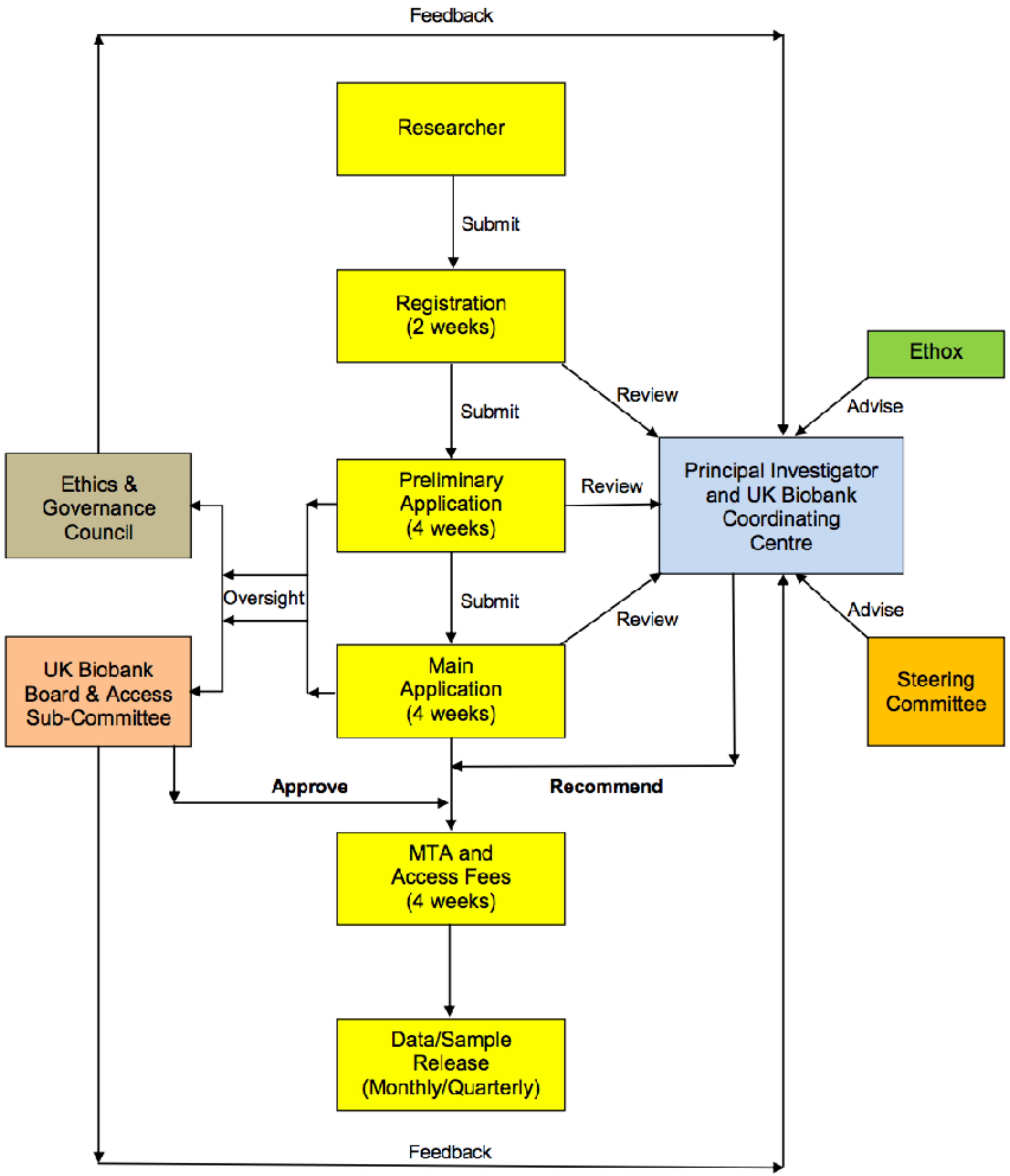
528,669 items of data are available, covering 496,895 participants.
Defined-instances run from 0 to 2, labelled using Instancing 2.
Units of measurement are milliseconds.

Maximum	2000
Decile 9	703
Decile 8	633
Decile 7	594
Decile 6	563
Median	539
Decile 4	516
Decile 3	493
Decile 2	469
Decile 1	441
Minimum	63

100004	Procedural metrics	11

How does it work?

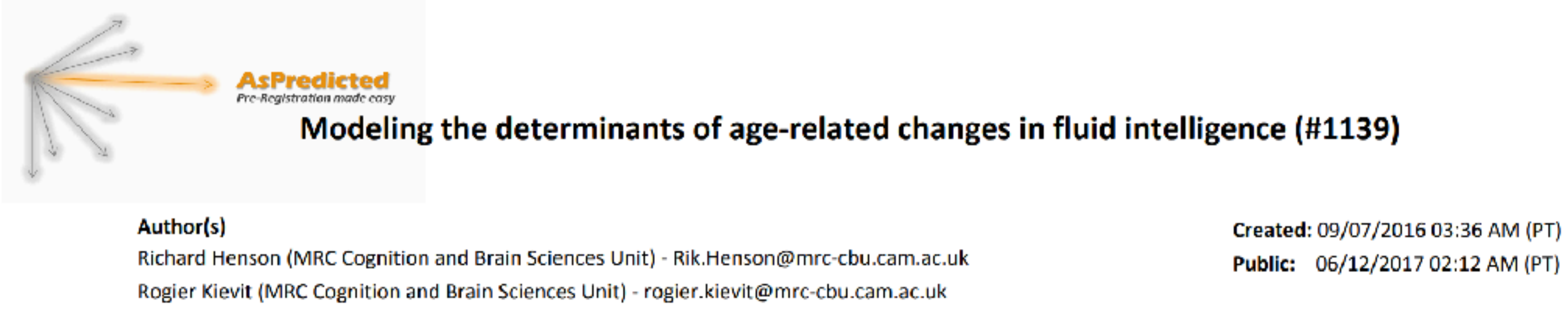
- It works. Very well.
- Competent professionals who understand the data and your project
- Feedback on proposal
- Approx 1-2 months start to finish
- Total cost: ~£2500 (can be more for 'bulk data')



Feedback
Researcher
Submit
Registration (2 weeks)
Review
Ethox
Advise
Submit
Ethics & Governance Council
Preliminary Application (4 weeks)
Review
Principal Investigator and UK Biobank Coordinating Centre
Oversight
Submit
Review
Advise
UK Biobank Board & Access Sub-Committee
Main Application (4 weeks)
Steering Committee
Approve
Recommend
MTA and Access Fees (4 weeks)
Data/Sample Release (Monthly/Quarterly)
Feedback

Stages in the application and review process (with the indicative timelines in parentheses); the roles of the different parties are described in Section C11

Our experiences

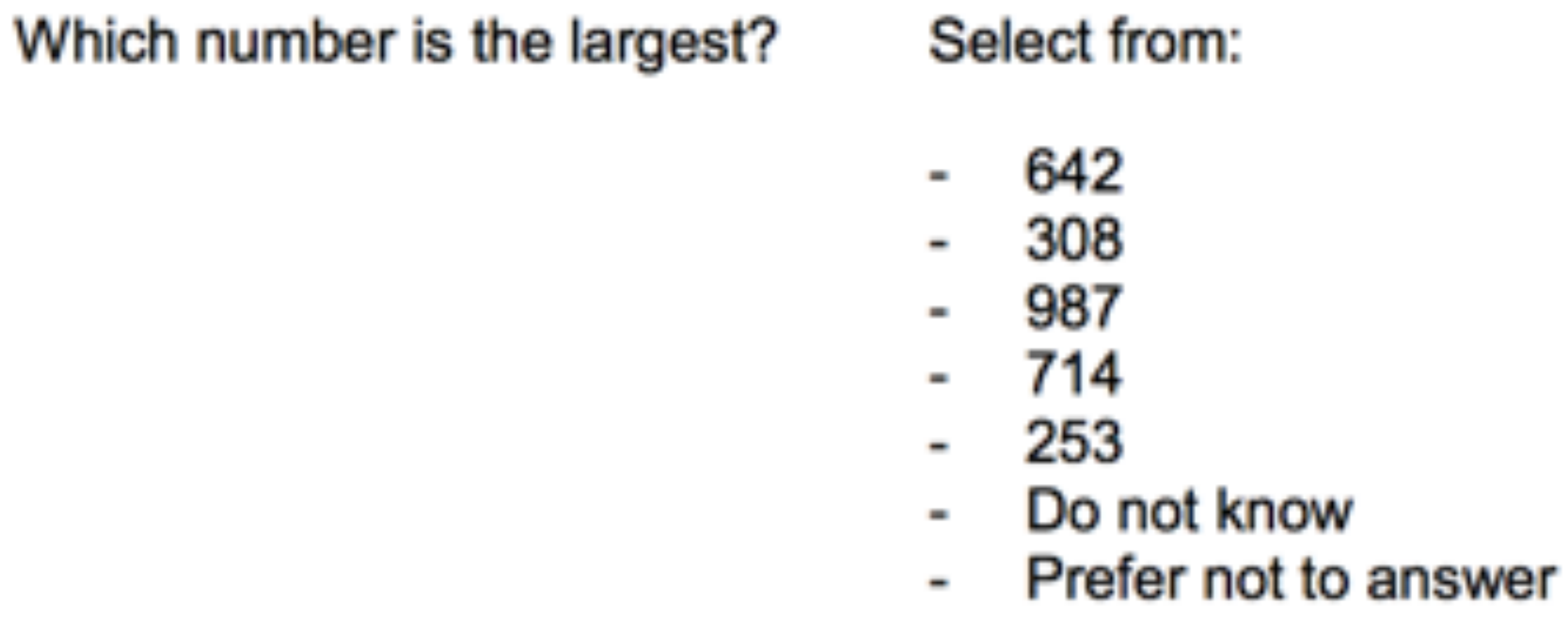

AsPredicted
Pre-Registration made easy

Modeling the determinants of age-related changes in fluid intelligence (#1139)

Author(s)
Richard Henson (MRC Cognition and Brain Sciences Unit) - Rik.Henson@mrc-cbu.cam.ac.uk
Rogier Kievit (MRC Cognition and Brain Sciences Unit) - rogier.kievit@mrc-cbu.cam.ac.uk

Created: 09/07/2016 03:36 AM (PT)
Public: 06/12/2017 02:12 AM (PT)

- Even specific plans are not specific until you write them down (and even then they're not)
- Complex models may fail to converge - Impossible to preregister every 'if then'
- The data/results:
  - Not 3 waves
  - Not fluid reasoning
  - Self-paced
  - Ceiling/Floor effects
  - Items (too) memorable
  - ***no significant slope variance in N=160.000***


Which number is the largest?

Select from:

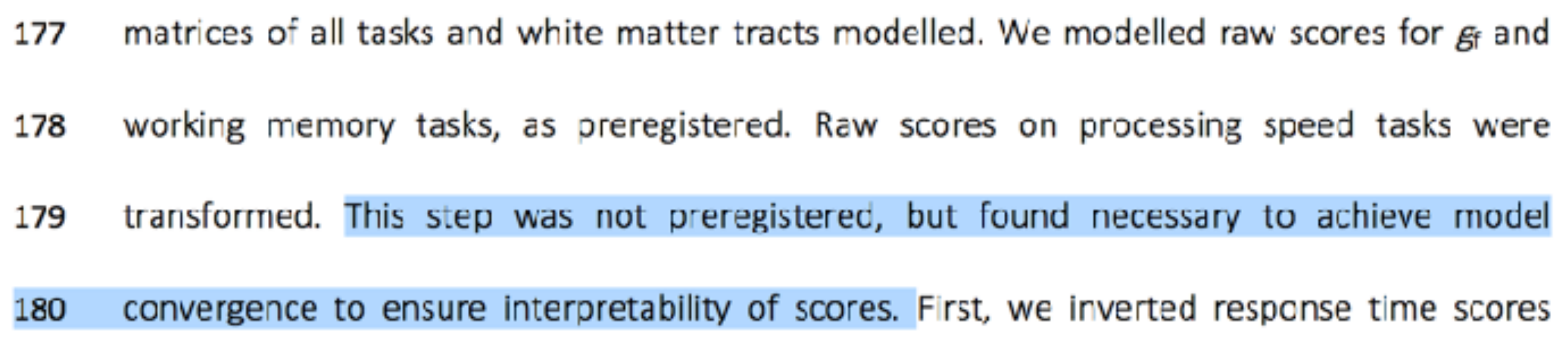
- 642
- 308
- 987
- 714
- 253
- Do not know
- Prefer not to answer

Me preregistering

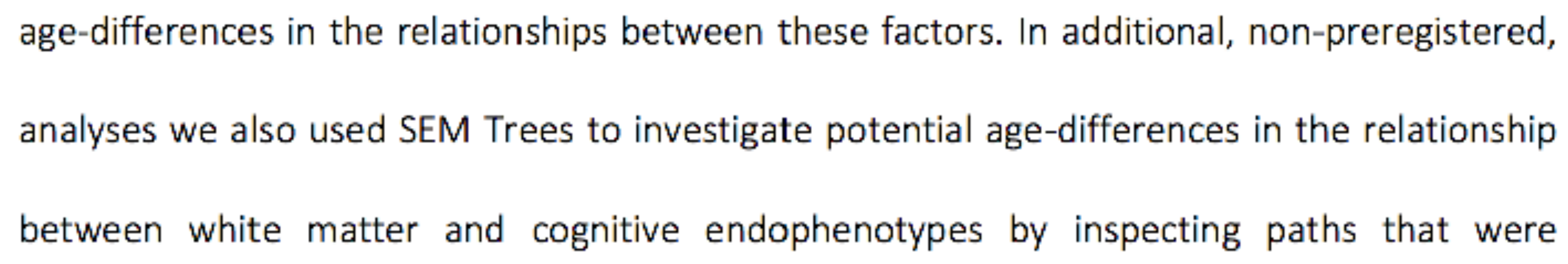
Me now

‘what to do when things go terribly wrong’

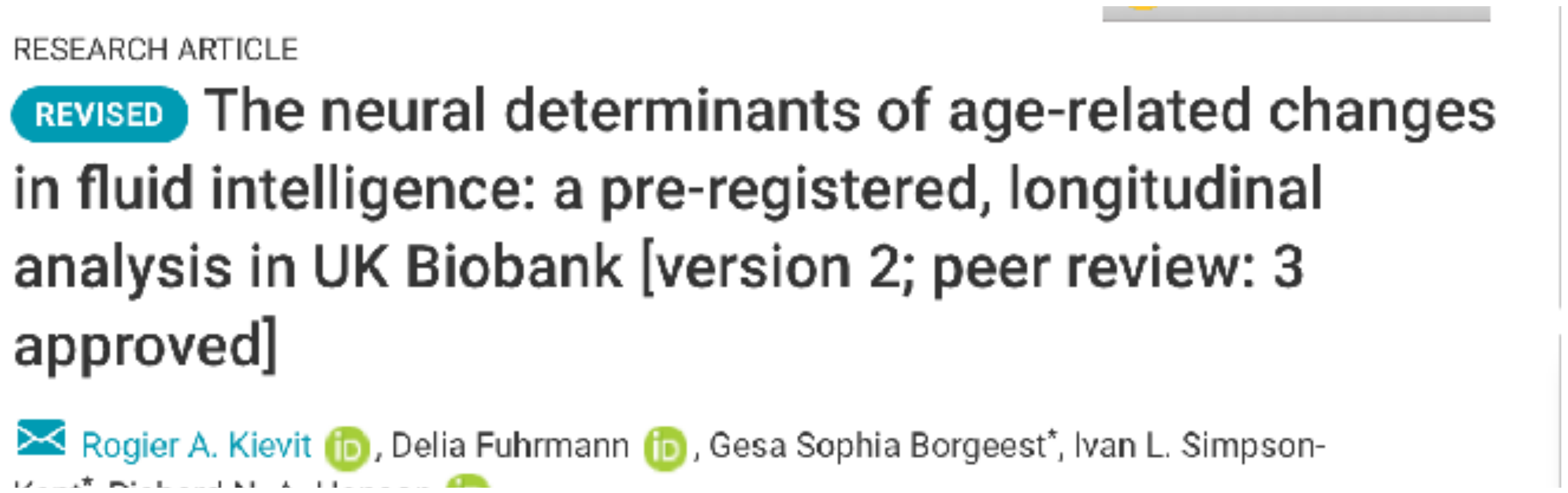
- 1) Deviate from preregistration when needed (and be transparant)
- OS practices to the rescue
- 2) Write-athon
- aka ‘shut up and write’
- Rules:
  - 1) Shut up
  - 2) Write



matrices of all tasks and white matter tracts modelled. We modelled raw scores for g_f and working memory tasks, as preregistered. Raw scores on processing speed tasks were transformed. This step was not preregistered, but found necessary to achieve model convergence to ensure interpretability of scores. First, we inverted response time scores



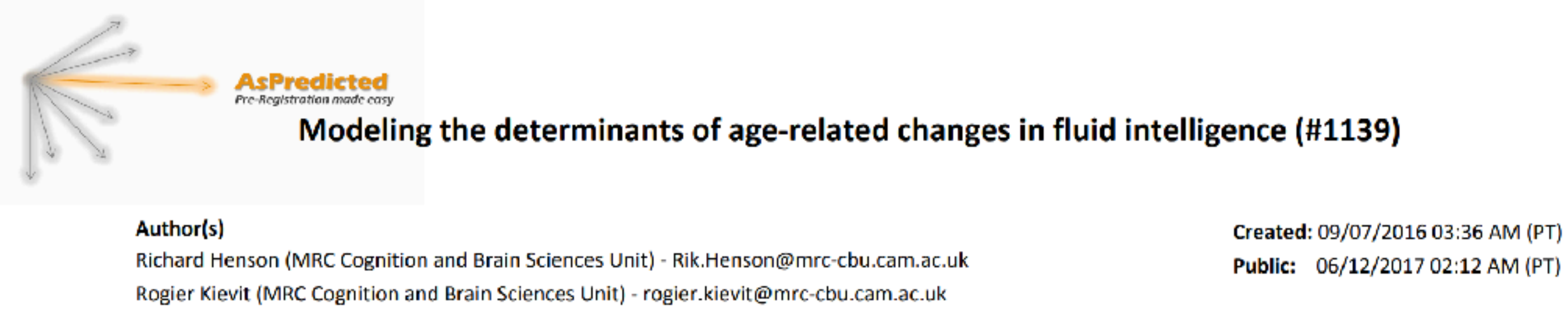
age-differences in the relationships between these factors. In additional, non-preregistered, analyses we also used SEM Trees to investigate potential age-differences in the relationship between white matter and cognitive endophenotypes by inspecting paths that were



RESEARCH ARTICLE

REVISED The neural determinants of age-related changes in fluid intelligence: a pre-registered, longitudinal analysis in UK Biobank [version 2; peer review: 3 approved]

Rogier A. Kievit, Delia Fuhrmann, Gesa Sophia Borgeest*, Ivan L. Simpson-

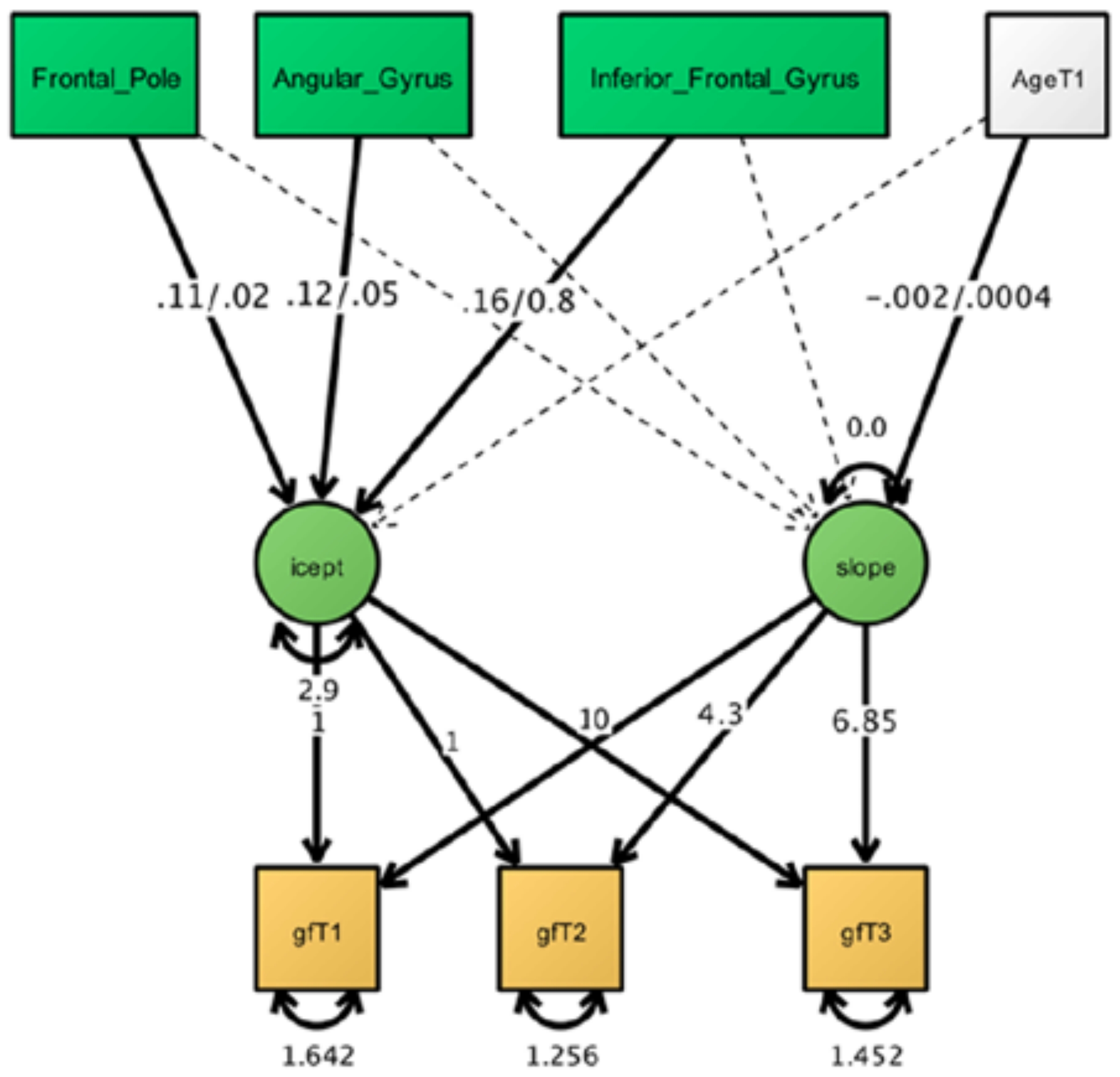


AsPredicted
Pre-Registration made easy

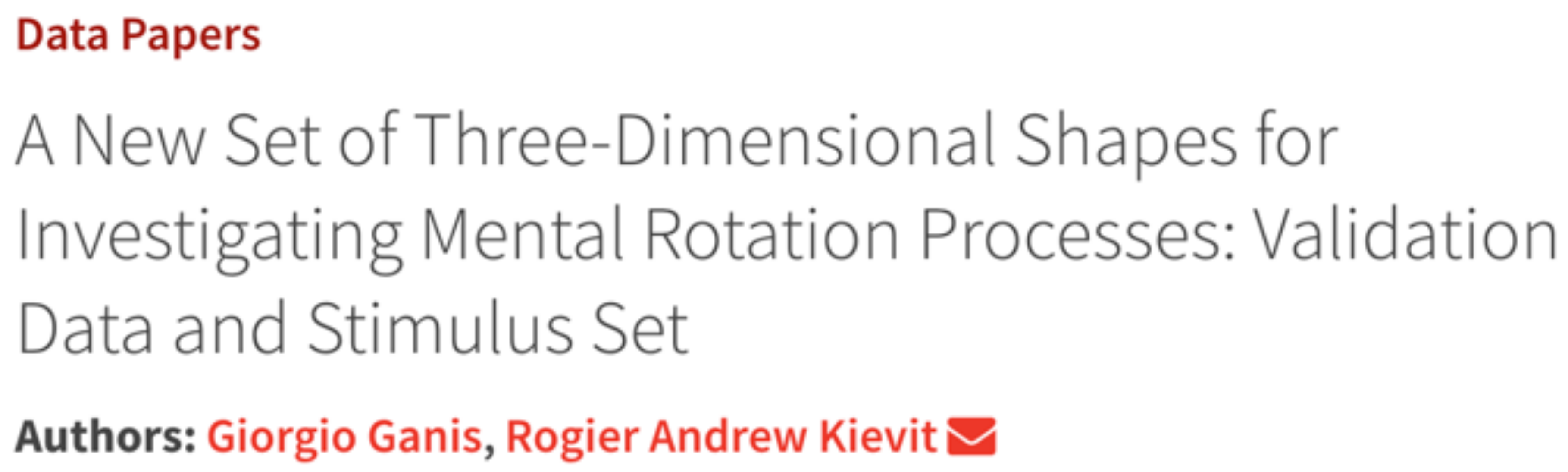
Modeling the determinants of age-related changes in fluid intelligence (#1139)

Author(s)
Richard Henson (MRC Cognition and Brain Sciences Unit) - Rik.Henson@mrc-cbu.cam.ac.uk
Rogier Kievit (MRC Cognition and Brain Sciences Unit) - rogier.kievit@mrc-cbu.cam.ac.uk

Created: 09/07/2016 03:36 AM (PT)
Public: 06/12/2017 02:12 AM (PT)



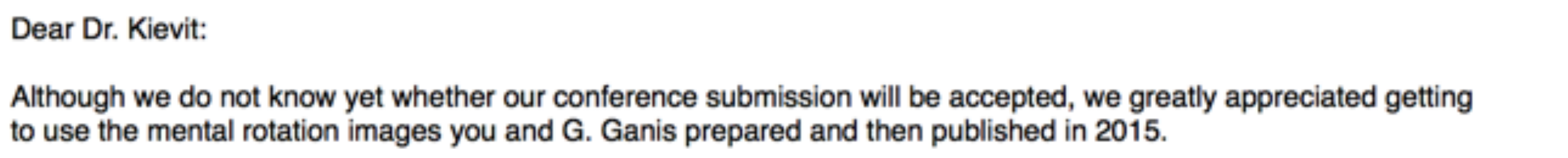
Where does all this data come from?



Data Papers

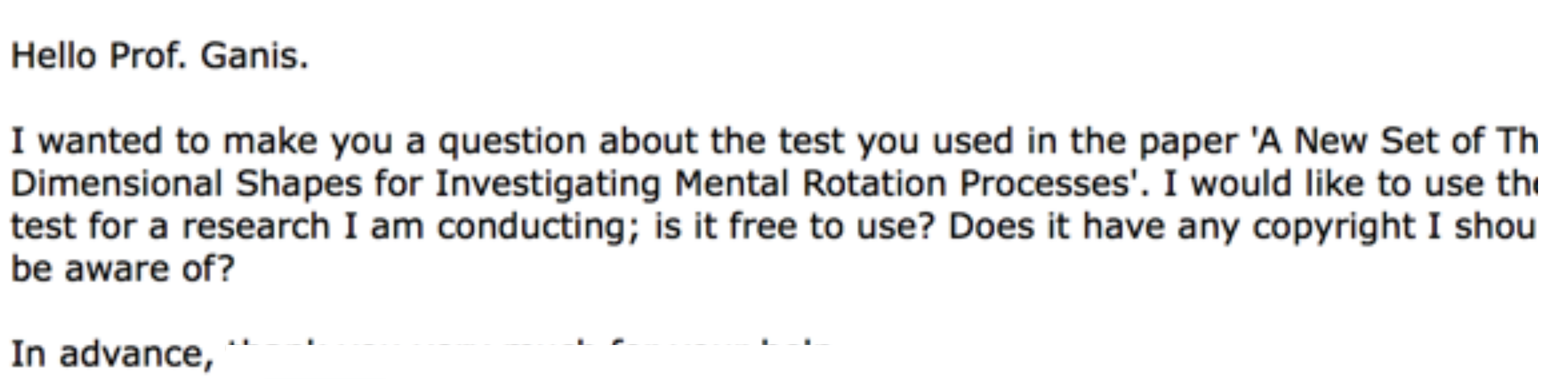
A New Set of Three-Dimensional Shapes for Investigating Mental Rotation Processes: Validation Data and Stimulus Set

Authors: Giorgio Ganis, Rogier Andrew Kievit



Dear Dr. Kievit:

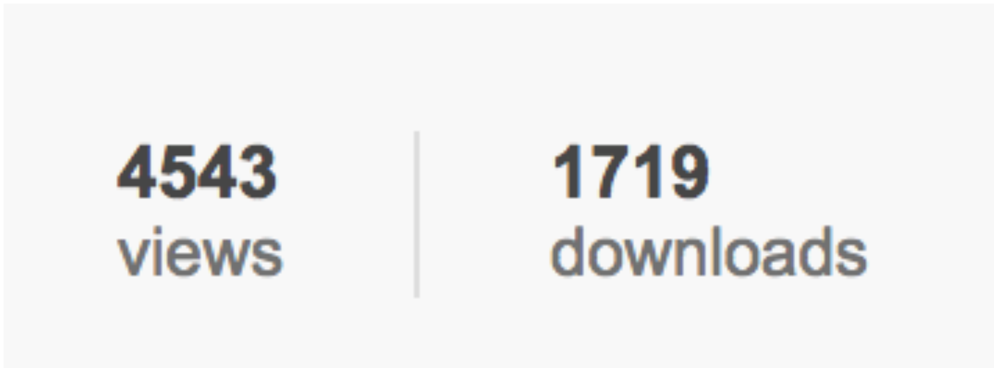
Although we do not know yet whether our conference submission will be accepted, we greatly appreciated getting to use the mental rotation images you and G. Ganis prepared and then published in 2015.



Hello Prof. Ganis.

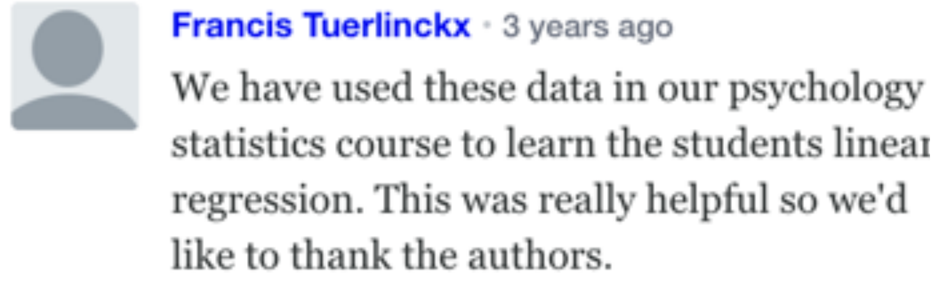
I wanted to make you a question about the test you used in the paper 'A New Set of Th
Dimensional Shapes for Investigating Mental Rotation Processes'. I would like to use th
test for a research I am conducting; is it free to use? Does it have any copyright I shou
be aware of?

In advance,



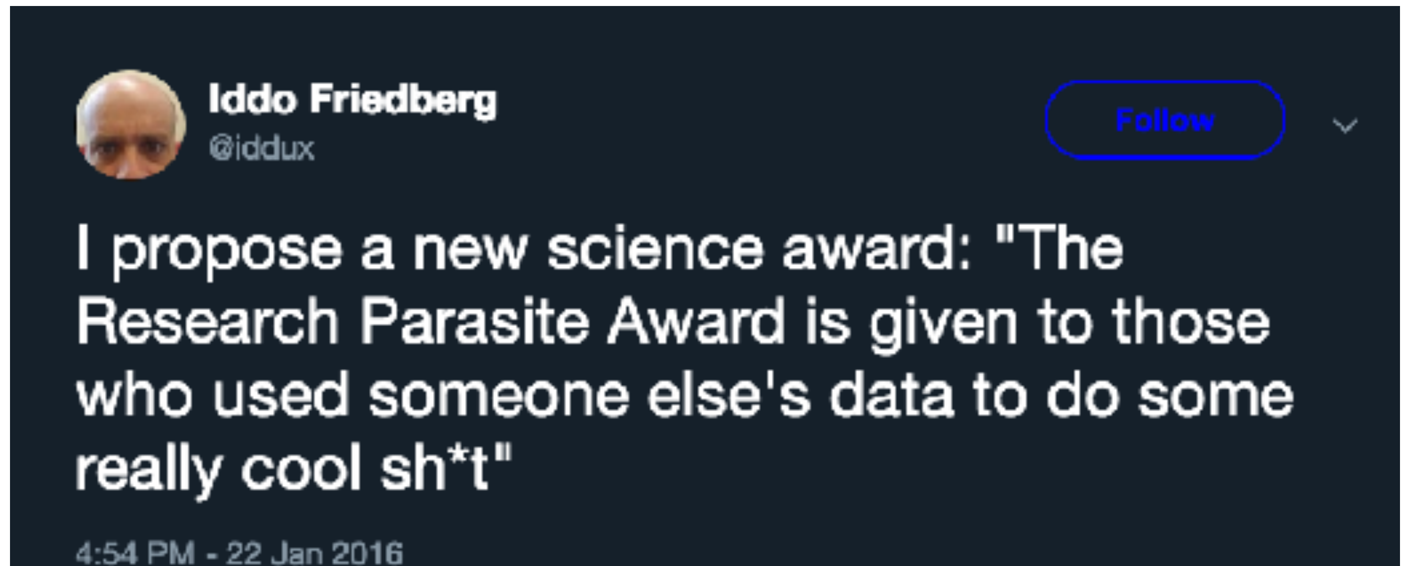
4543 views | 1719 downloads

25 citations



Francis Tuerlinckx · 3 years ago

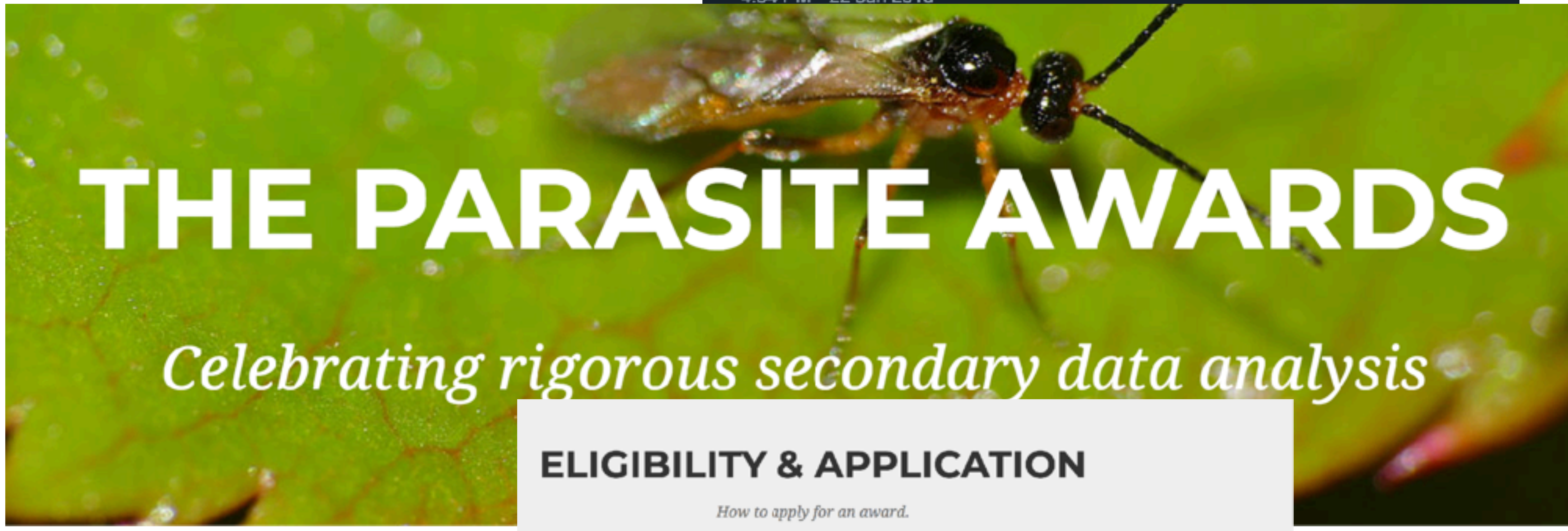
We have used these data in our psychology statistics course to learn the students linear regression. This was really helpful so we'd like to thank the authors.



Iddo Friedberg
@iddux

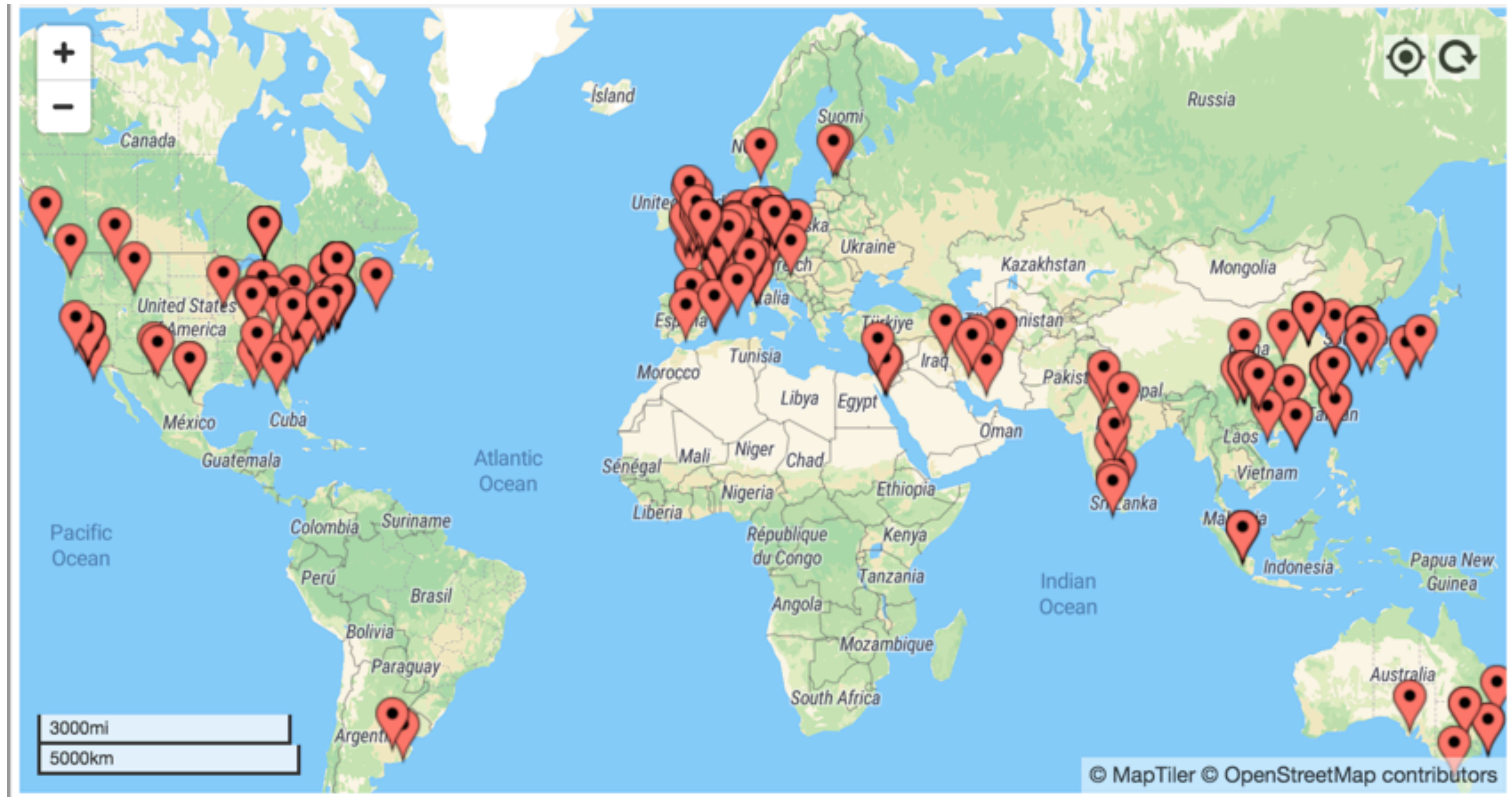
I propose a new science award: "The Research Parasite Award is given to those who used someone else's data to do some really cool sh*t"

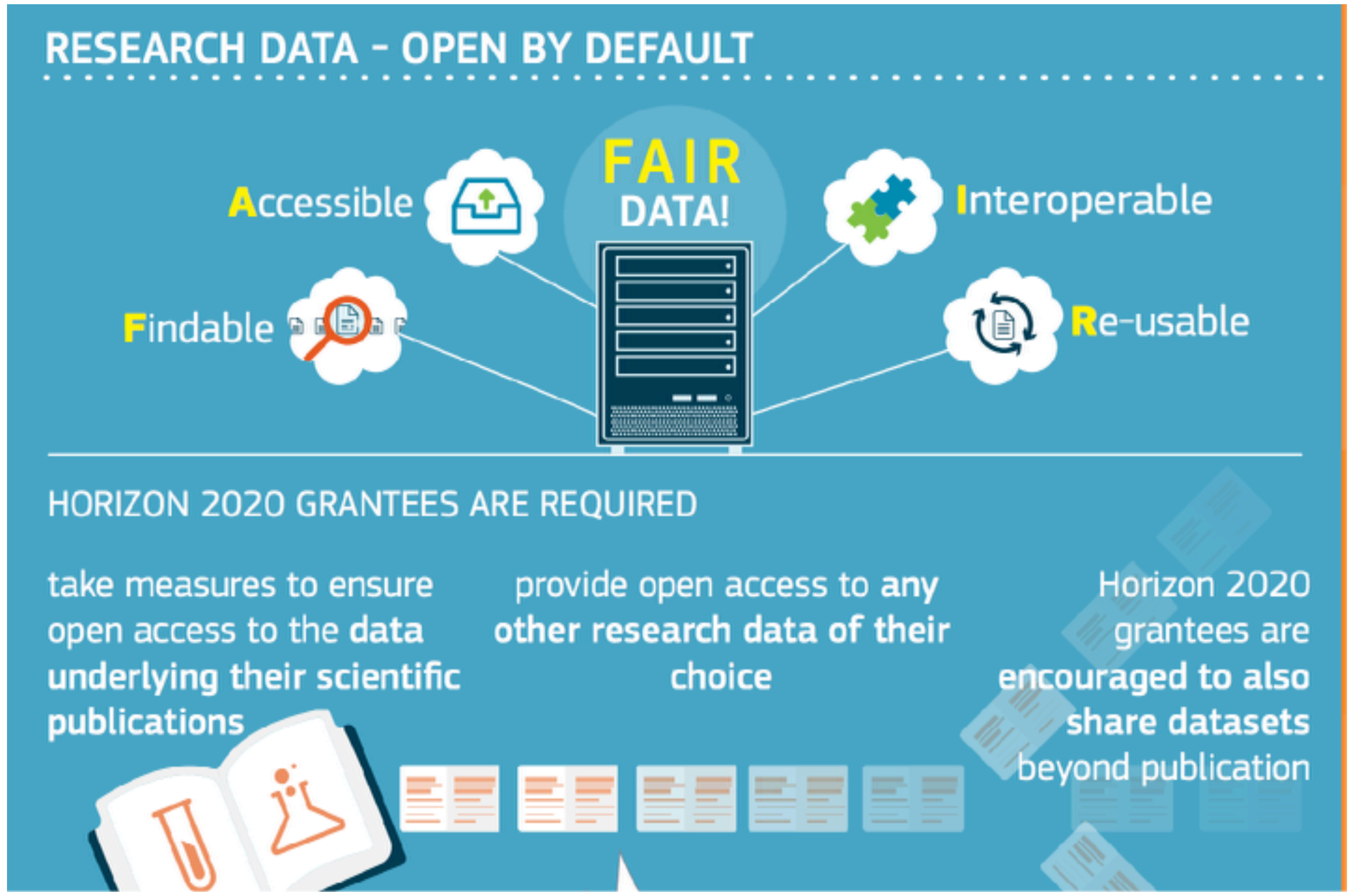
4:54 PM - 22 Jan 2016


THE PARASITE AWARDS
Celebrating rigorous secondary data analysis
ELIGIBILITY & APPLICATION
How to apply for an award.

Cam-CAN data portal

- Managed access
- 550 downloads





RESEARCH DATA - OPEN BY DEFAULT

FAIR DATA!

Accessible

Interoperable

Findable

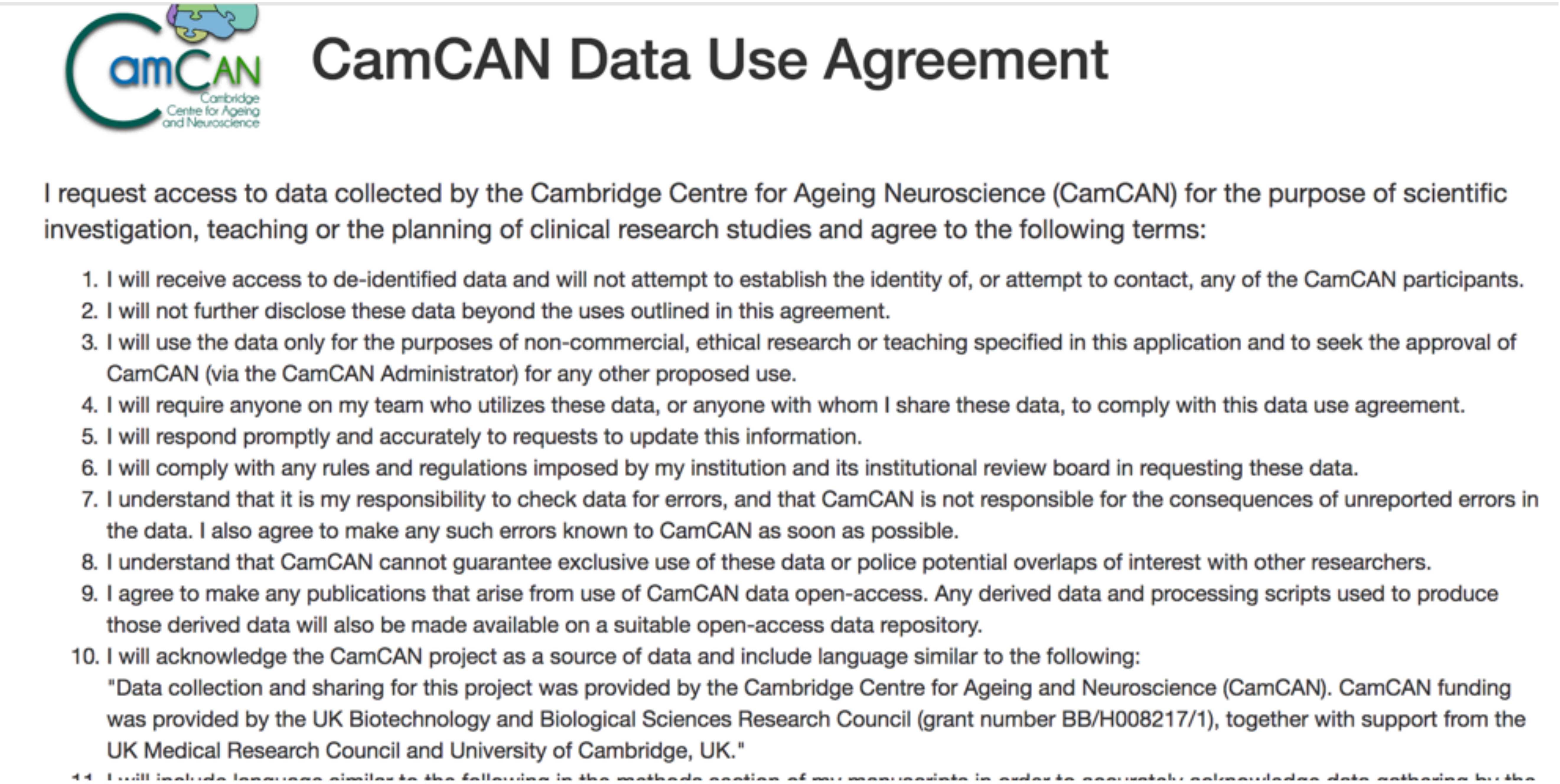
Re-usable

HORIZON 2020 GRANTEES ARE REQUIRED

take measures to ensure open access to the data underlying their scientific publications

provide open access to any other research data of their choice

Horizon 2020 grantees are encouraged to also share datasets beyond publication



CamCAN Cambridge Centre for Ageing and Neuroscience

CamCAN Data Use Agreement

I request access to data collected by the Cambridge Centre for Ageing Neuroscience (CamCAN) for the purpose of scientific investigation, teaching or the planning of clinical research studies and agree to the following terms:

1. I will receive access to de-identified data and will not attempt to establish the identity of, or attempt to contact, any of the CamCAN participants.
2. I will not further disclose these data beyond the uses outlined in this agreement.
3. I will use the data only for the purposes of non-commercial, ethical research or teaching specified in this application and to seek the approval of CamCAN (via the CamCAN Administrator) for any other proposed use.
4. I will require anyone on my team who utilizes these data, or anyone with whom I share these data, to comply with this data use agreement.
5. I will respond promptly and accurately to requests to update this information.
6. I will comply with any rules and regulations imposed by my institution and its institutional review board in requesting these data.
7. I understand that it is my responsibility to check data for errors, and that CamCAN is not responsible for the consequences of unreported errors in the data. I also agree to make any such errors known to CamCAN as soon as possible.
8. I understand that CamCAN cannot guarantee exclusive use of these data or police potential overlaps of interest with other researchers.
9. I agree to make any publications that arise from use of CamCAN data open-access. Any derived data and processing scripts used to produce those derived data will also be made available on a suitable open-access data repository.
10. I will acknowledge the CamCAN project as a source of data and include language similar to the following:
"Data collection and sharing for this project was provided by the Cambridge Centre for Ageing and Neuroscience (CamCAN). CamCAN funding was provided by the UK Biotechnology and Biological Sciences Research Council (grant number BB/H008217/1), together with support from the UK Medical Research Council and University of Cambridge, UK."
11. I will include language similar to the following in the methods section of my manuscripts in order to accurately acknowledge data gathering by the

Summary 1

- Secondary data increasingly rich, usable and important avenue for fundamental science
- Plenty of practical challenges, but:
- Improves power, precision, generalisation and replication
- Golden age for reproducible science and theory driven model testing

open science

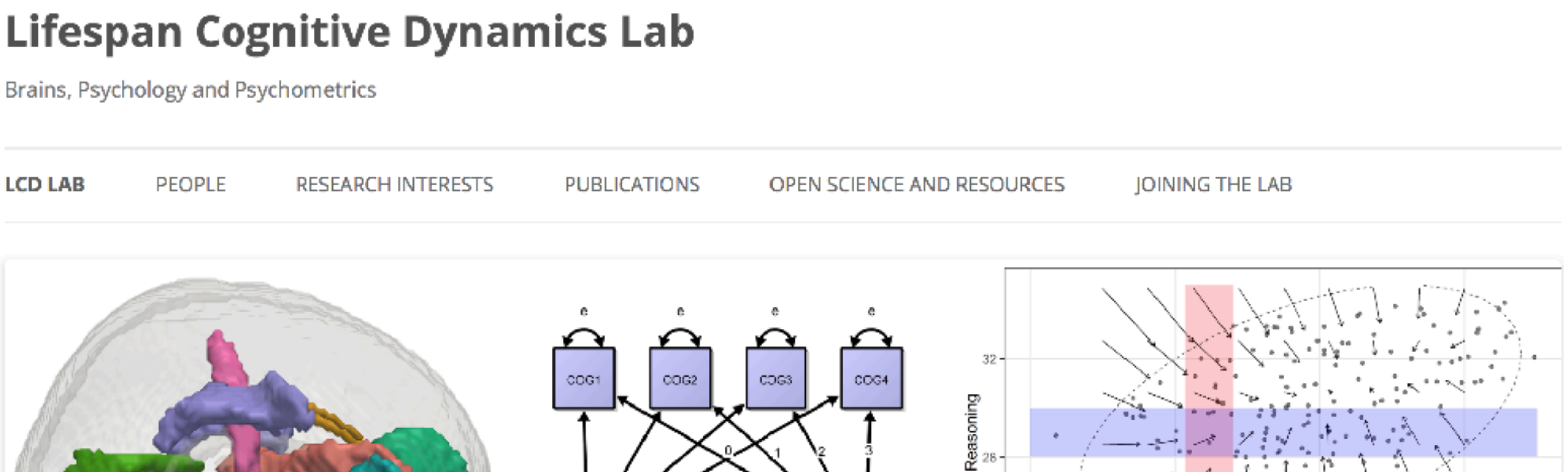
Summary 2

- 1) You don't have to do everything (OS) all the time (or pledge allegiance, sacrifice small animals etc.)
- 2) It doesn't have to (and it won't) be perfect
- 3) Start (today) with small steps and find your place
  - Preregister a study
  - Validate your findings with an open dataset
  - post your (synthetic) dataset online
  - Start a reproducibiliTea journal club
  - post a preprint
- Become a research symbiont
  - Contribute to, and benefit from, improved science

wellcome trust

MRC Medical Research Council

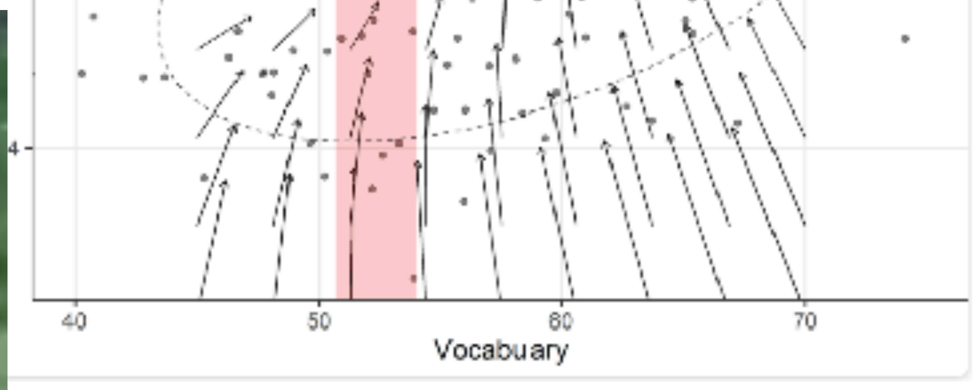
Thanks to funders, lab members & mascots



Lifespan Cognitive Dynamics Lab

Brains, Psychology and Psychometrics

LCD LAB PEOPLE RESEARCH INTERESTS PUBLICATIONS OPEN SCIENCE AND RESOURCES JOINING THE LAB



Questions?

rogier.kievit@mrc-cbu.cam.ac.uk/@rogierK