

# New horizons at the interface of psychology and the computing sciences

Tal Yarkoni  
University of Texas at Austin

# Psychoinformagic

- The ability to conjure up amazing psychological results through sheer technical wizardry



# Psychoinformatics

- The ability to conjure up amazing psychological results through sheer technical wizardry



"Any sufficiently  
advanced technology is  
indistinguishable from  
magic."

- Arthur C. Clarke

# Precedent



# Biomedical Big Data

- One human genome: ~3 billion base pairs
- 1 - 100 GB of data per person
- It gets worse...
  - Gene expression, microbiomics, etc...

# Bioinformatics

- “An interdisciplinary field that develops methods and software tools for understanding biological data. Bioinformatics combines computer science, statistics, mathematics, and engineering to study and process biological data.” —Wikipedia

# Why?

- How else are we going to do this?
- We can't do a paired t-test on 20 data points to find gene variants implicated in disease
- If the reality is enormously complex, our models have to scale accordingly



# -informatics

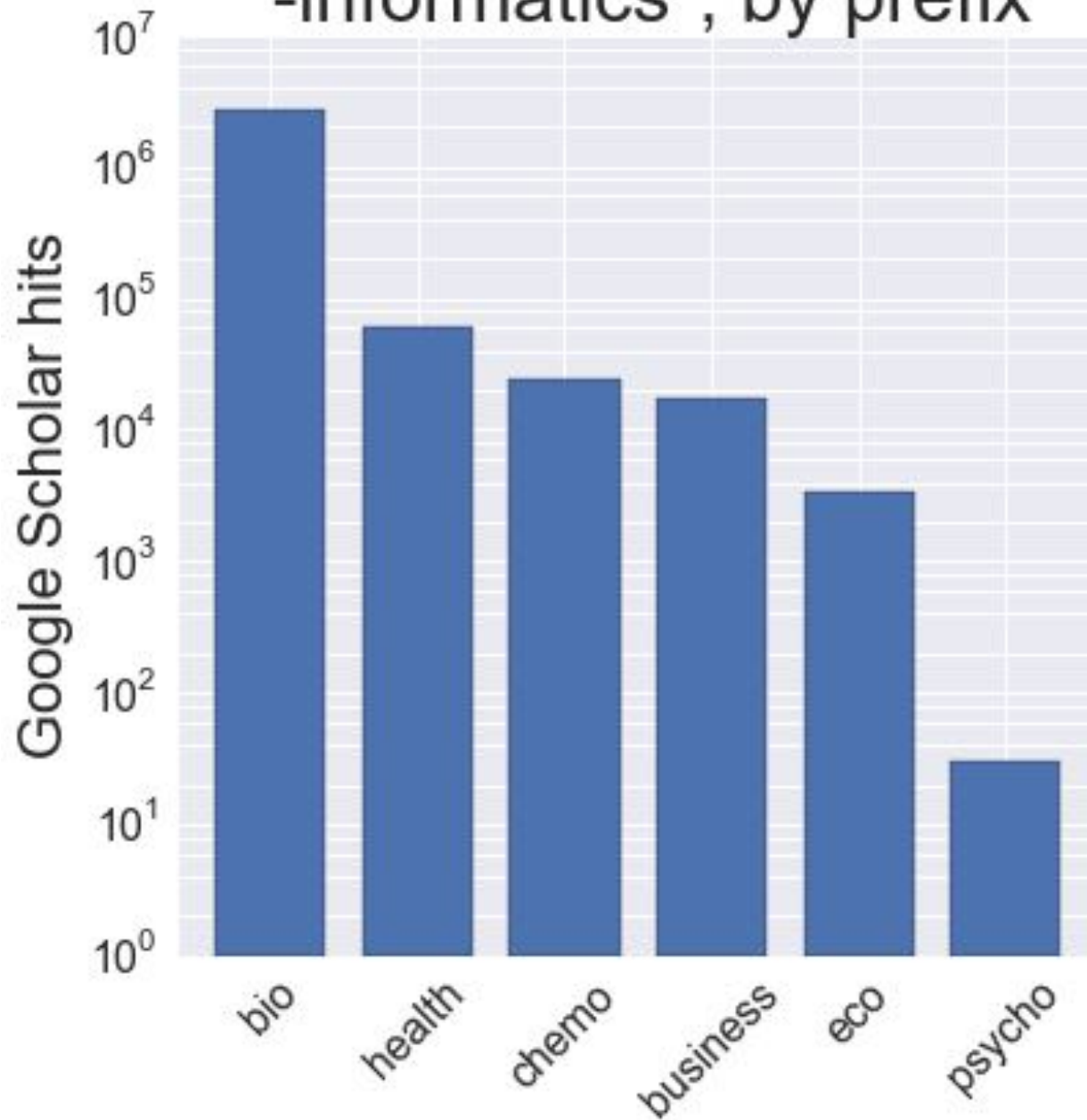
- The same holds in other fields...
  - Chemoinformatics
  - Ecoinformatics
  - Neuroinformatics
  - Health informatics
  - Etc. etc.

# We're late!

Psychology




## "-informatics", by prefix



# Psychoinformatics: New Horizons at the Interface of the Psychological and Computing Sciences

**Tal Yarkoni**

Institute of Cognitive Science, University of Colorado Boulder

Current Directions in Psychological Science  
21(6) 391–397  
© The Author(s) 2012  
Reprints and permission:  
[sagepub.com/journalsPermissions.nav](http://sagepub.com/journalsPermissions.nav)  
DOI: 10.1177/0963721412457362  
<http://cdps.sagepub.com>  


## Abstract

Psychologists live in an increasingly data-rich world, and our ability to make continued progress in understanding the mind and brain depends on finding new ways to organize and synthesize an ever-expanding body of knowledge. In this article, I review current research in psychoinformatics—an emerging discipline that uses tools and techniques from the computer and information sciences to improve the acquisition, organization, and synthesis of psychological data. I focus on several areas where the application of informatics approaches has already paid large dividends, leading to advances including novel data-collection approaches, the adaptation of computational techniques and insights, the enhanced aggregation and organization of psychological data, large-scale data mining and synthesis, and improved research and publication practices. I argue that in the coming years, informatics approaches are likely to play the same instrumental role in shaping psychological research that they have already played in other fields, such as genetics and neuroscience.

## Keywords

informatics, methods, data mining, information science



- Good news! This doesn't happen any more!
- Now we're up to 83 Google Scholar hits
- ...versus ~2.5 million for bioinformatics

# Why the gap?

- Data in psychology are historically “small”
- Theory dominates exploratory data analysis
- Most psychologists don’t get training in computing

# How are we doing?

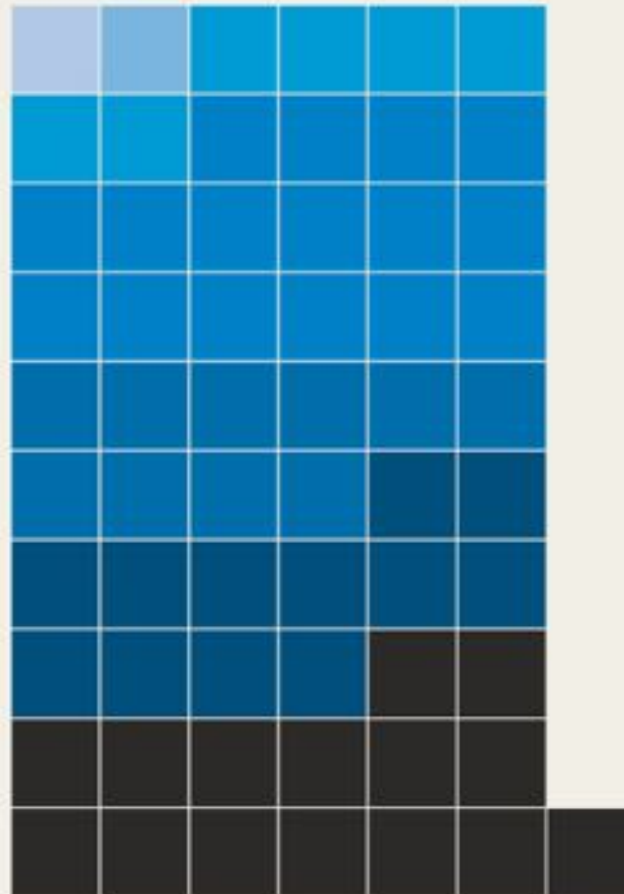
- Somewhere between “hard to say” and “pretty bad”
  - We’re in the middle of a replication crisis
  - Datasets are getting bigger
  - Effect sizes are getting smaller
  - Theoretical disputes are rarely resolved

## RELIABILITY TEST

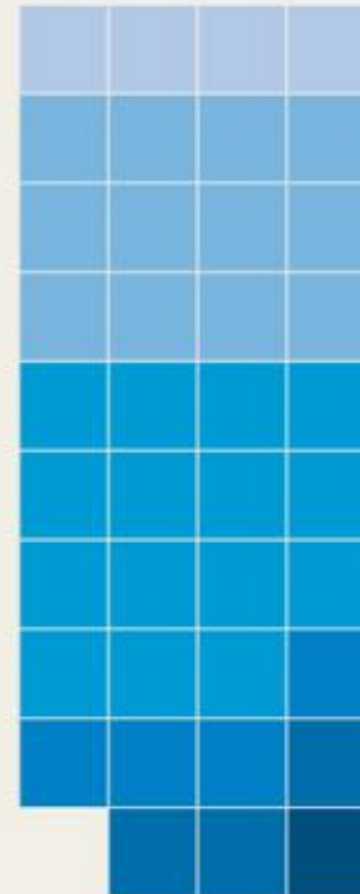
An effort to reproduce 100 psychology findings found that only 39 held up.\* But some of the 61 non-replications reported similar findings to those of their original papers.

Did replicate match original's results?

NO: 61



YES: 39

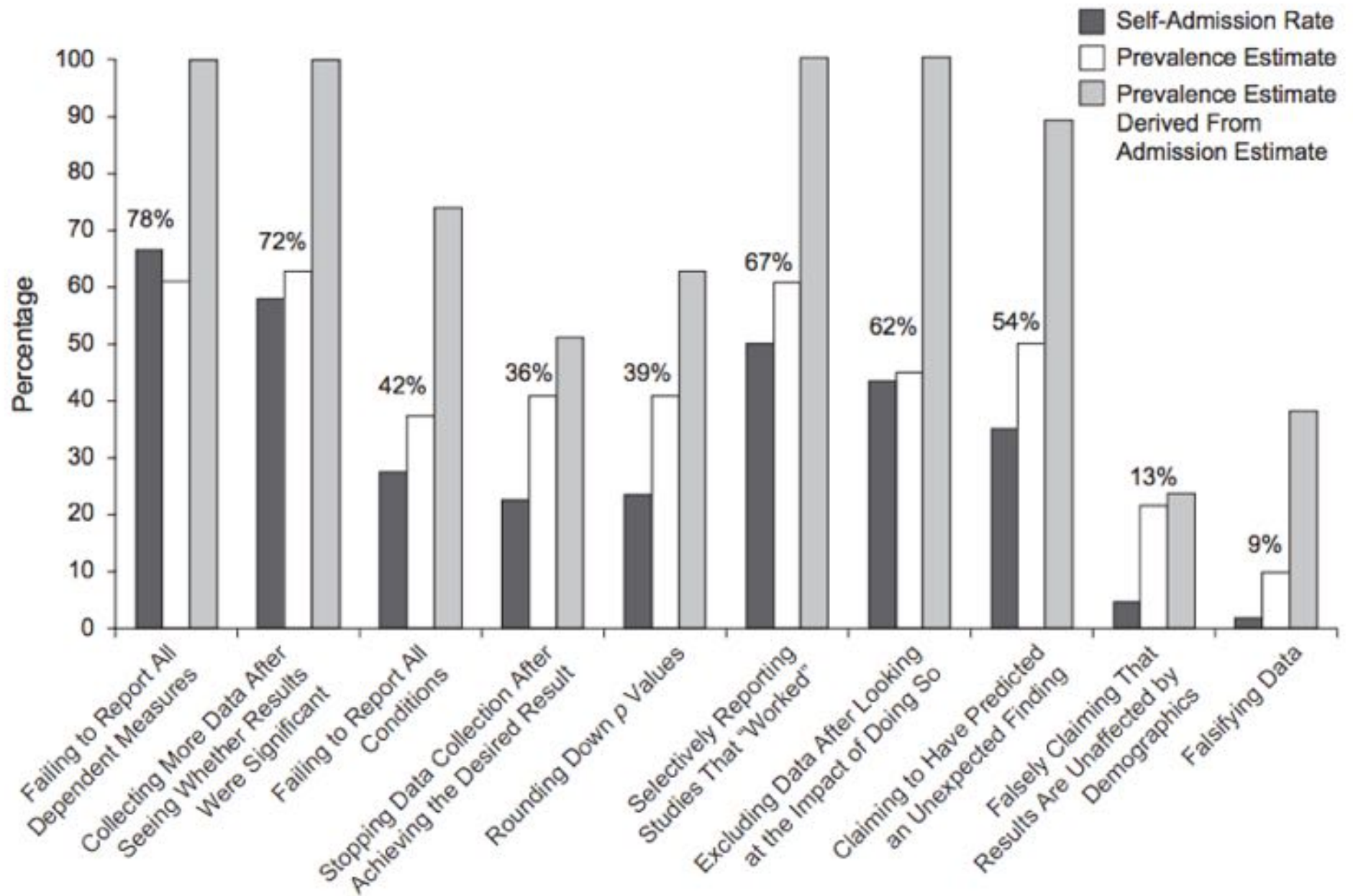


Replicator's opinion: How closely did findings resemble the original study:

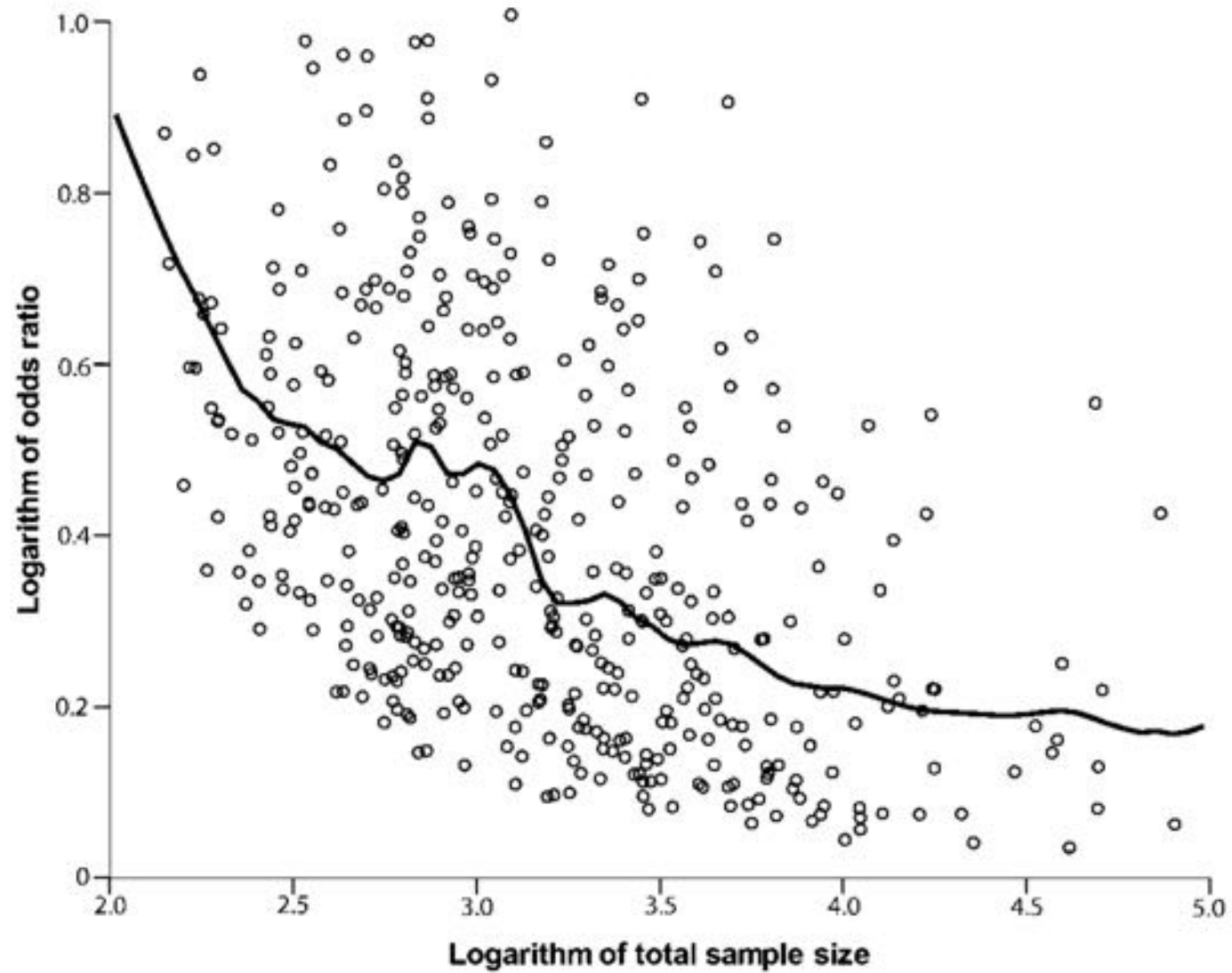
- Light blue: Virtually identical
- Light blue: Extremely similar
- Very similar
- Moderately similar
- Somewhat similar
- Slightly similar
- Not at all similar

\* based on criteria set at the start of each study





John et al. (2012)



Ioannidis (2008)

**Table 1.** Some of Psychology's Theory Competitions

Phenomenon	Competing theories	Initial or early publication	Controversy-resolving publication
Sapir–Whorf hypothesis	Language/culture does (or does not) influence categorization	Whorf (1956)	
Structure of affect	Bipolar vs. independent positive and negative dimensions	Nowlis and Nowlis (1956)	
Counterattitudinal role playing	Dissonance vs. self-perception vs. impression management	Festinger and Carlsmith (1959)	
Memory search	Serial vs. parallel search	Sternberg (1966)	
Implicit learning	Rules vs. associative learning	Reber (1967)	
Mental rotation	Analog vs. propositional representation	Shepard and Metzler (1971)	
Semantic priming	Spreading activation vs. compound cueing	Meyer and Schvaneveldt (1971)	
Categorization	Features, exemplars, prototypes, rules	Labov (1973)	
Altruism	Intrinsic vs. extrinsic motivation	Cialdini, Darby, and Vincent (1973)	
Misleading information	Altered traces vs. independent traces	Loftus and Palmer (1974)	
Judgment under uncertainty	Heuristics and biases vs. rationality	Tversky and Kahneman (1974)	
Affect–cognition relationship	Affective primacy vs. cognitive primacy	Zajonc (1980)	
Memory dissociations	Modules vs. processes vs. thresholds	Jacoby and Dallas (1981)	

Note. The emptiness of the rightmost column is not an accident—see text.

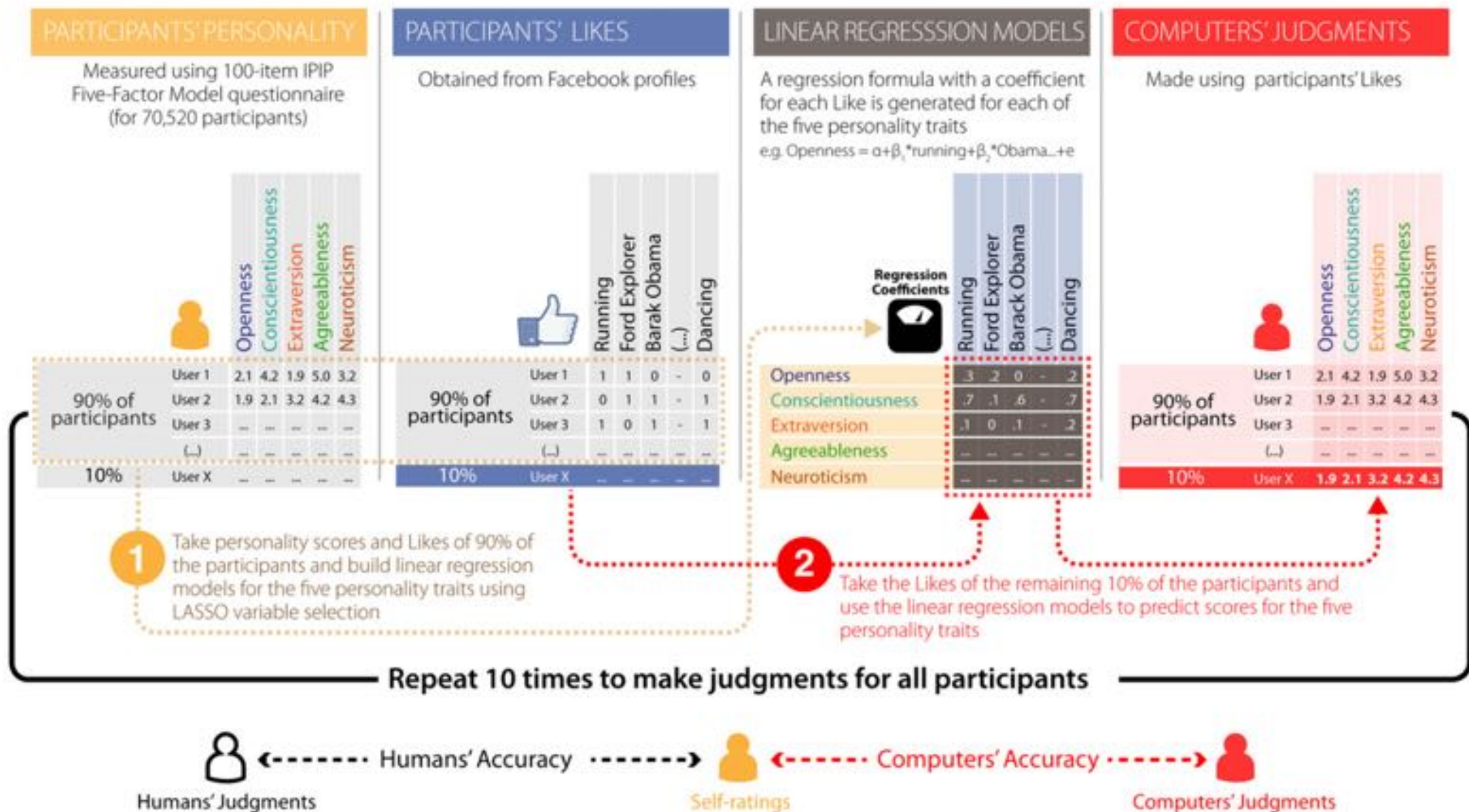
Greenwald (2012)

# What can we do?

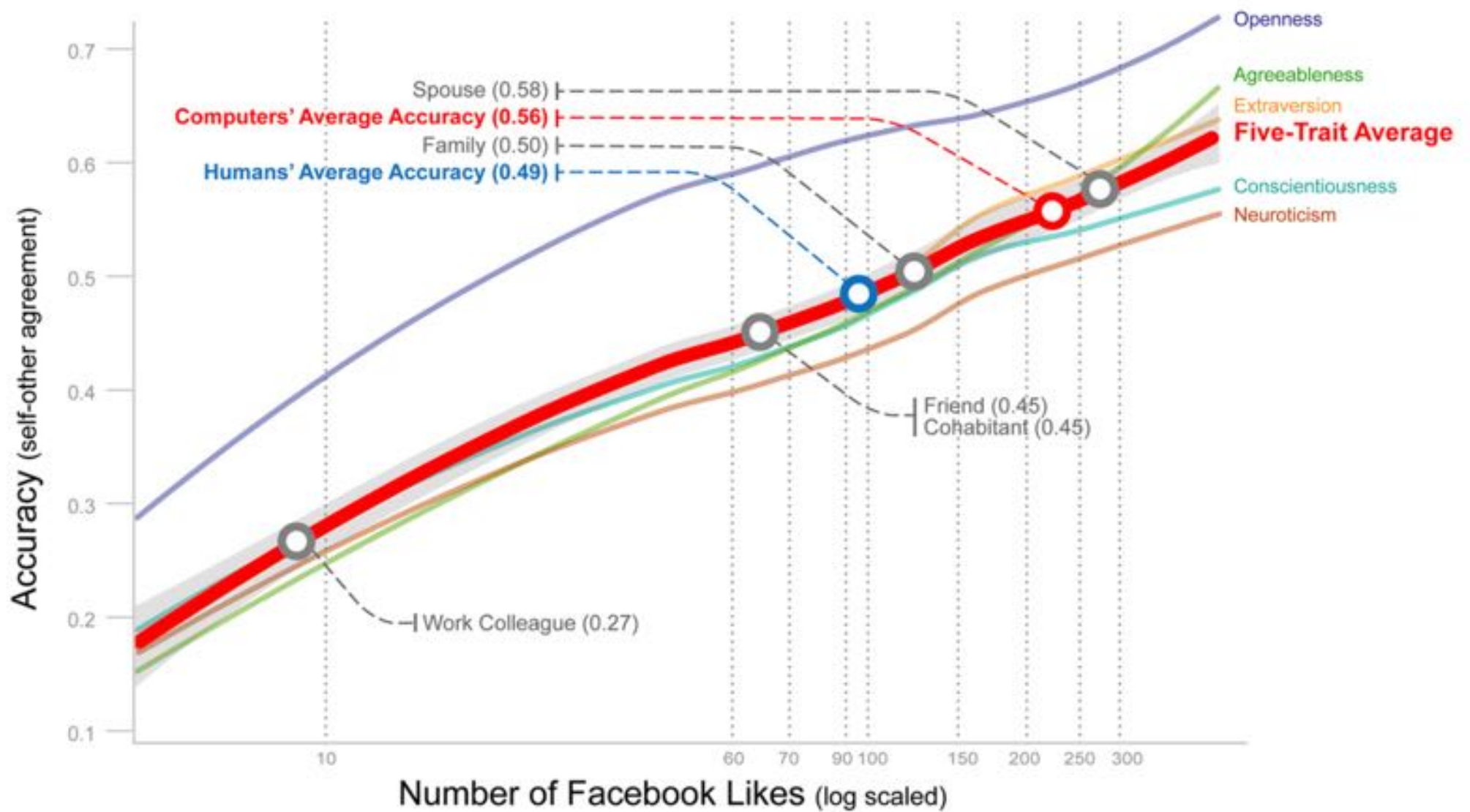
- Many things!
  - Restructure incentives, promote good practices, etc.
- Many potential solutions will be technical in nature
- Scientific computing can improve research at every stage of the work flow

# Data collection

- Conventional lab-based studies are small, time-consuming, artificial
- Sometimes unavoidable (e.g., physiological measures, special populations), but often not
- New platforms enable data acquisition on a massive scale
- With potentially greater ecological validity

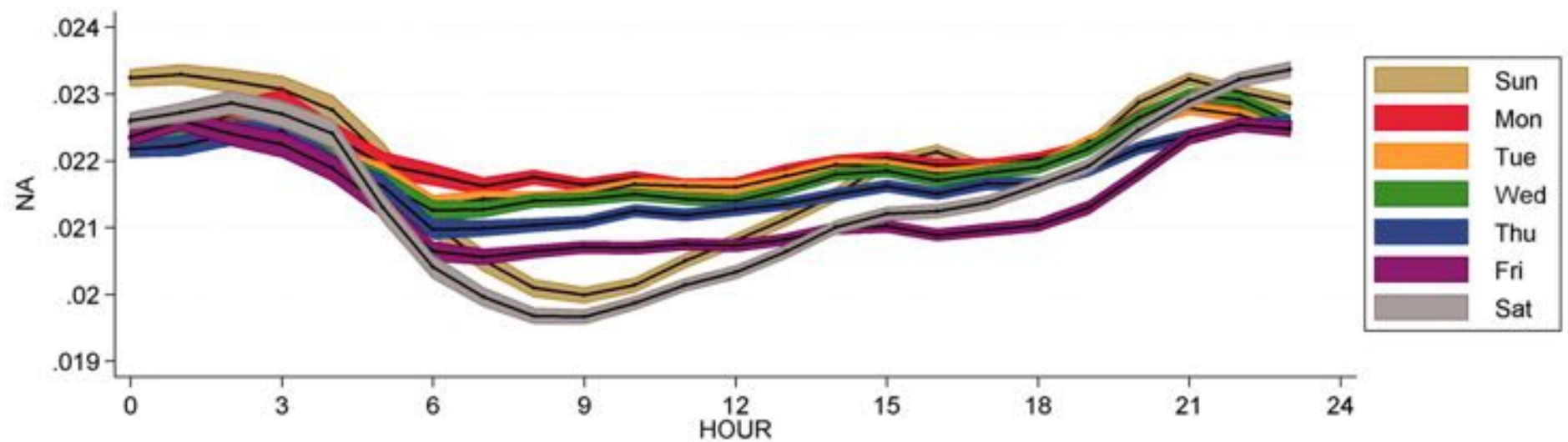
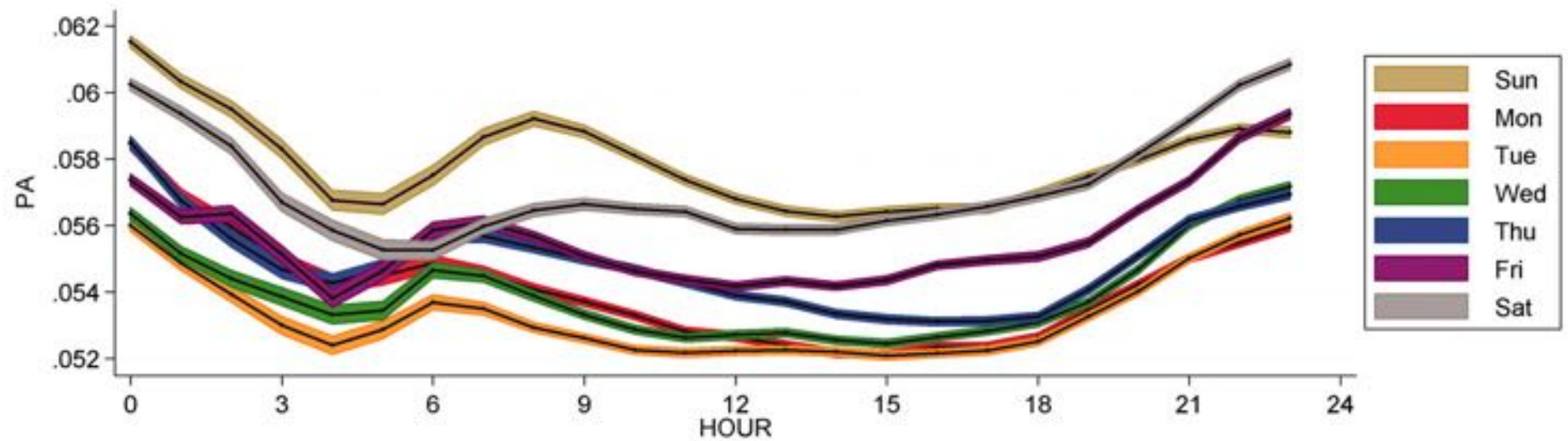


Youyou, Kosinski, & Stillwell (2015)



Youyou, Kosinski, & Stillwell (2015)





Golder & Macy (2011)



# Data sharing and re-use

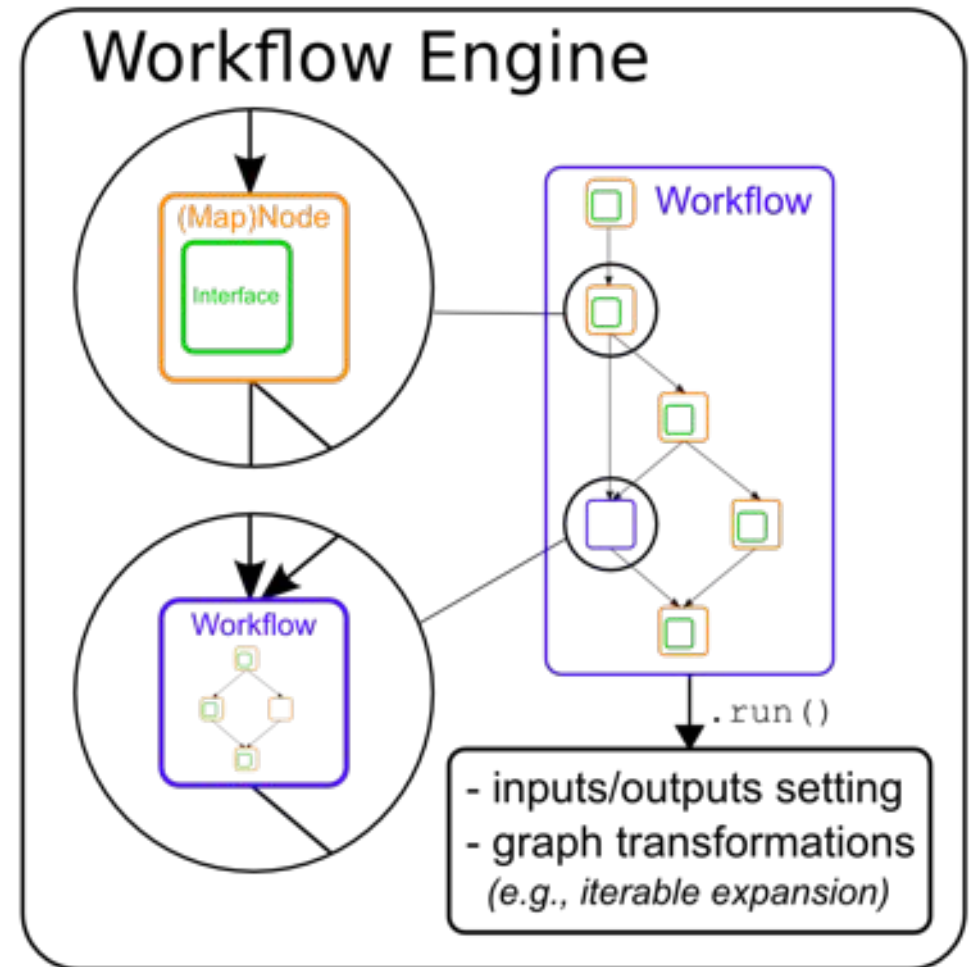
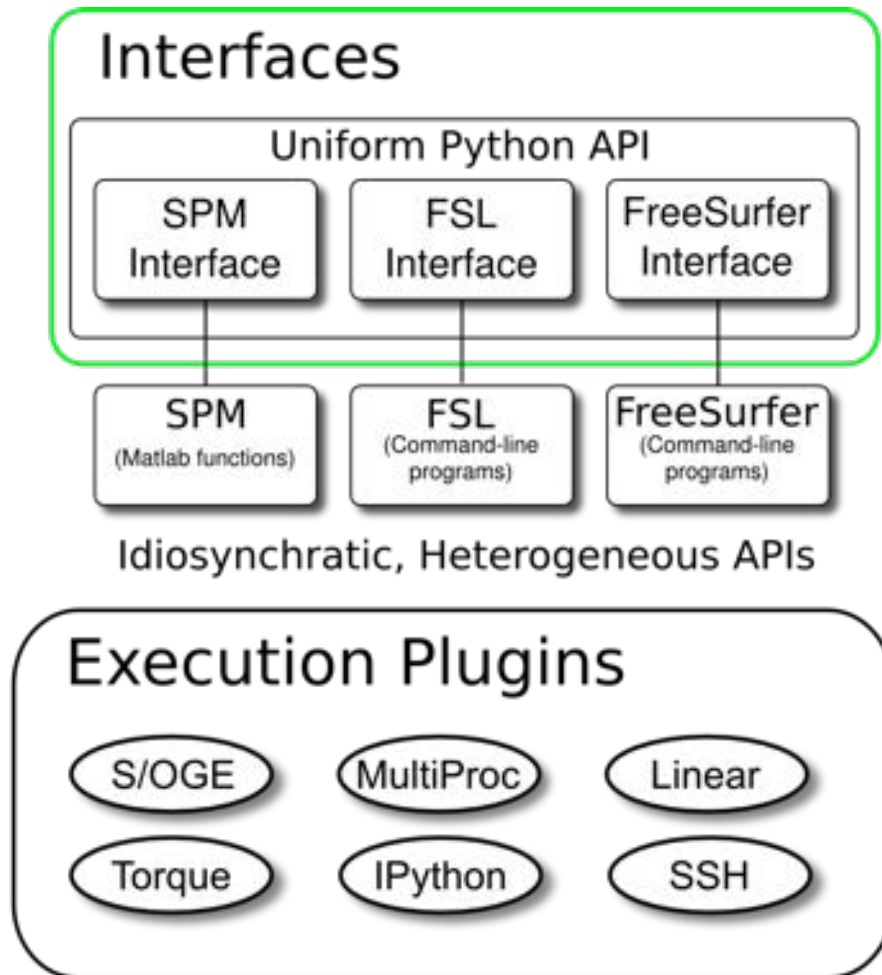
- In theory, we're supposed to share data
- In practice, not so much (Wicherts et al. 2006, 2011)
- We have the tools to solve these problems
  - GitHub, Dataverse, Open Science Framework...
  - Domain-specific repositories (NeuroVault, Databrary ...)

# Managing work flows

- Most scientists don't manage their data/analyses/results very systematically
- Who did what analysis? When? How were research decisions made? Why are there four versions of the data?
- Most experimental designs and findings are not annotated in a machine-readable way
- How do we find all studies done in population X? What's the relationship between concepts A and B?
- Compare with, e.g., genetics

# A better way

- Automate everything
- Version control everything
- Test everything



Gorgolewski et al (2011)

## SYNDROMES

Schizophrenia

Bipolar Disorder

ADHD

## COGNITION

WM Updating

Response Inhibition

Response inhibition is related to VLPFC activity through the contrast of stop vs. go trials on the Stop Signal Task

## NEURAL SYSTEMS

DL PFC

VL PFC

### SUPPORTED BY 6 REFERENCES

Aron, A.R. & Poldrack, R.A. (2006).  
Cortical and Subcortical Contributions to  
Stop Signal Response Inhibition: Role of  
the Subthalamic Nucleus.  
*Journal of Neuroscience*, 26, 2424-2433.

## SIGNALLING PATHWAYS

Dopamine D1

Dopamine D2

Adrenergic A2a

## GENES

COMT

DRD2

ADRA2a

Poldrack et al (2011)

# PyMVPA -- Multivariate Pattern Analysis in Python

build failing

coverage 79%

doi 10.3389/neuro.11.003.2009

For information how to install PyMVPA please see doc/source/installation.rst .

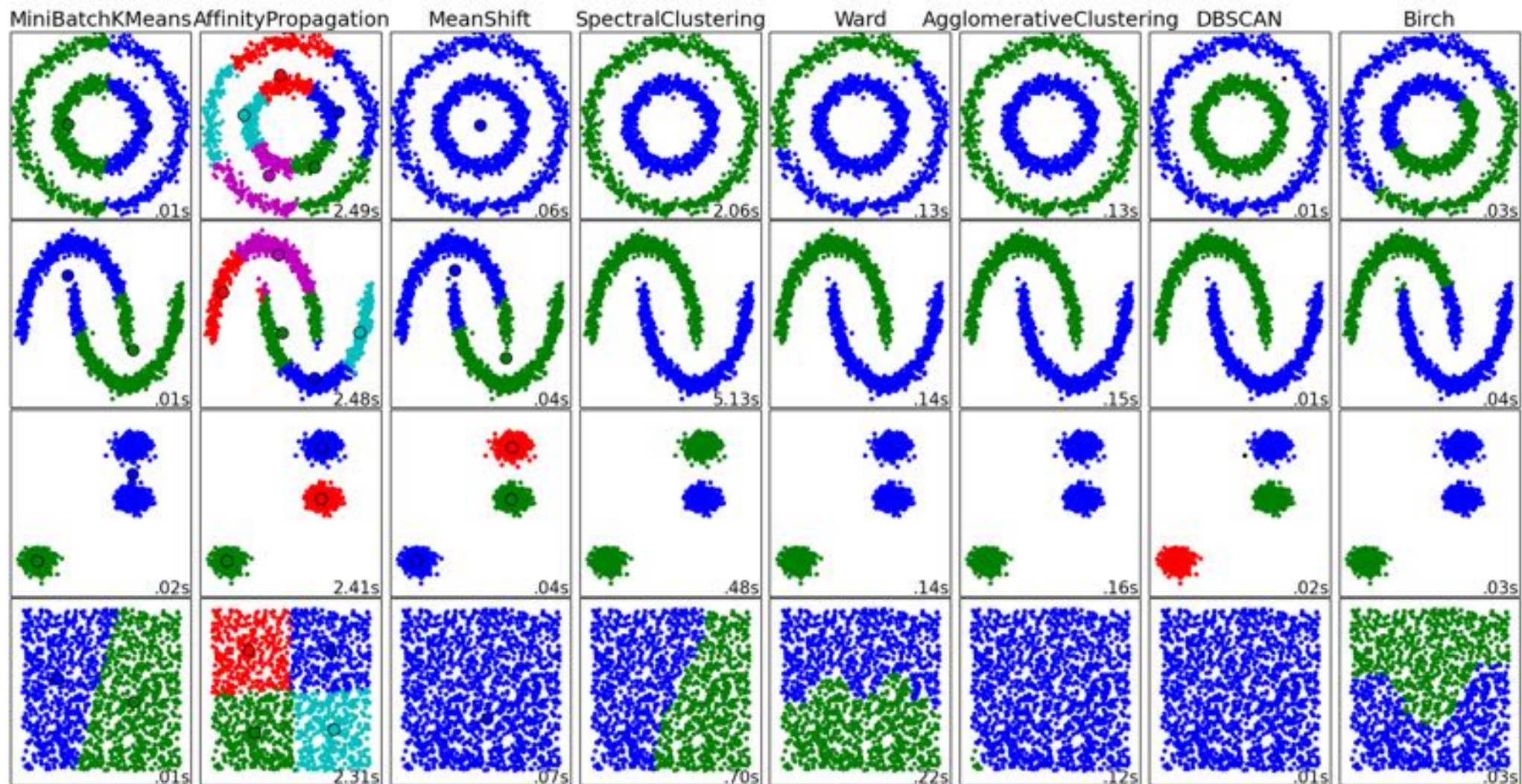
Further information and access to binary packages is available from the project website at <http://www.pymvpa.org> .

<https://github.com/PyMVPA/PyMVPA>

# Bigger, better analyses

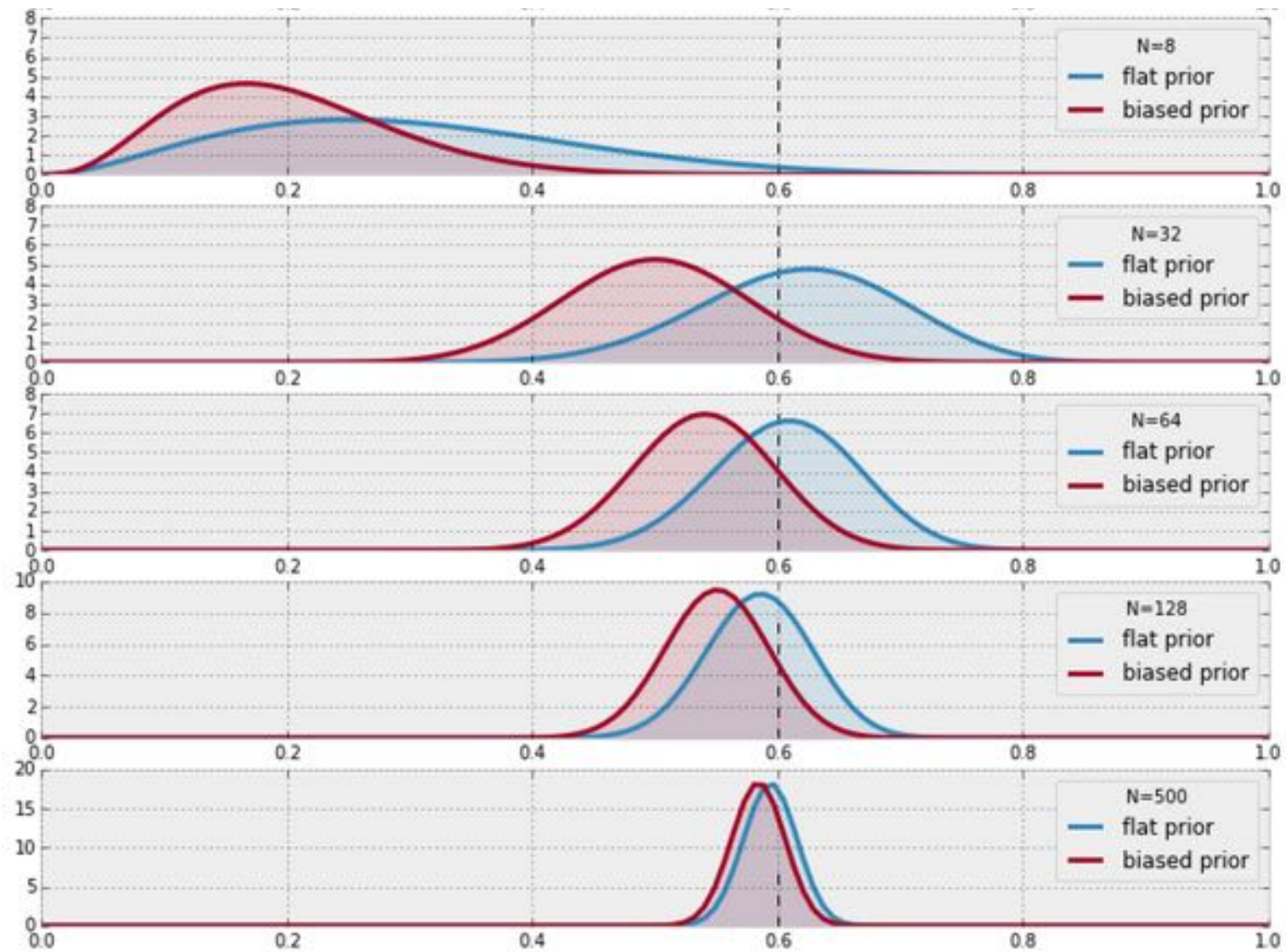
- Complement traditional focus on hypothesis testing via  $p$ -values
- Cross-validation and out-of-sample prediction
- Computationally-intensive multilevel models
- Computational simulations
- Novel analysis approaches (e.g., network analysis)



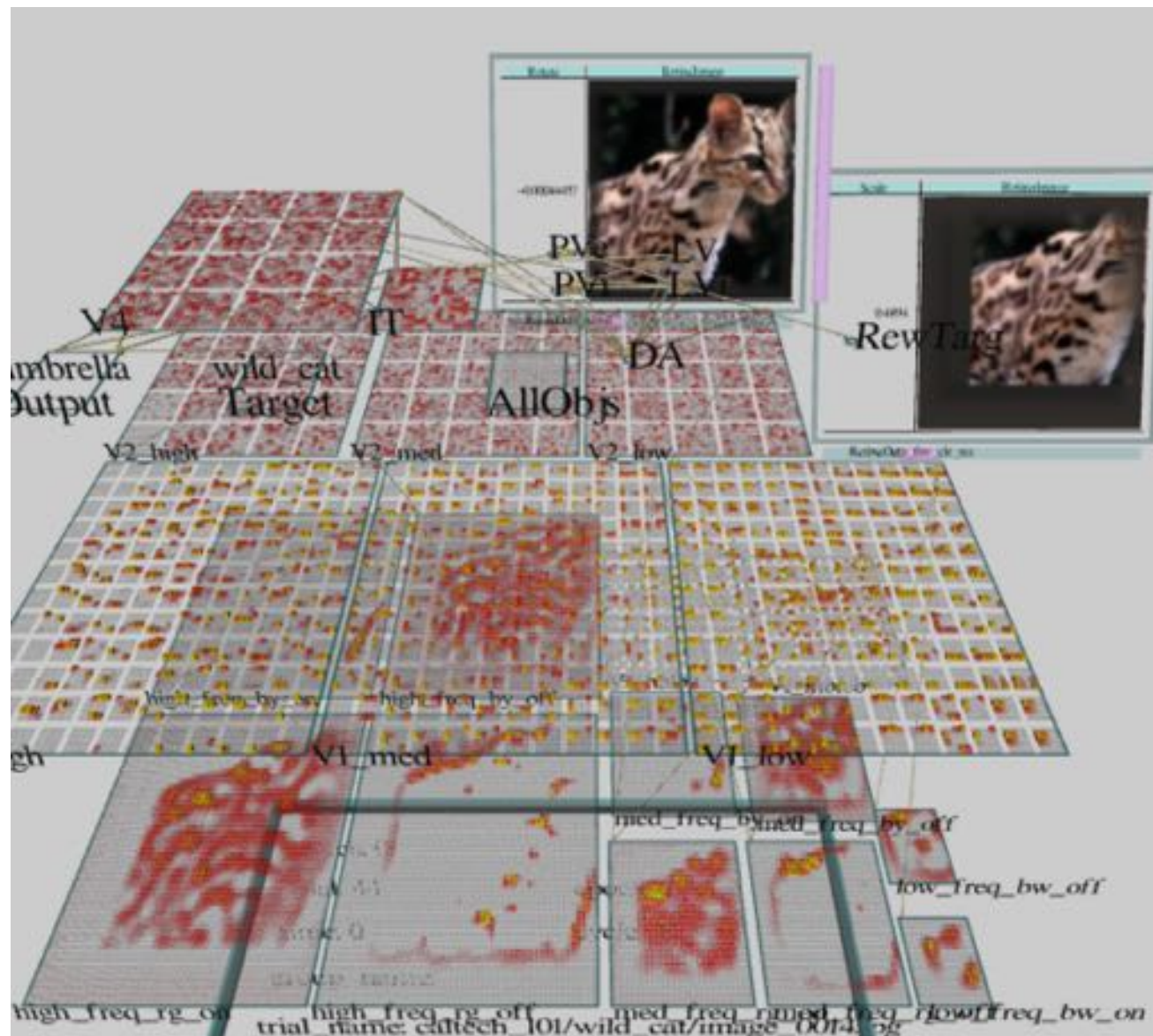


<http://scikit-learn.org/stable/modules/clustering.html>

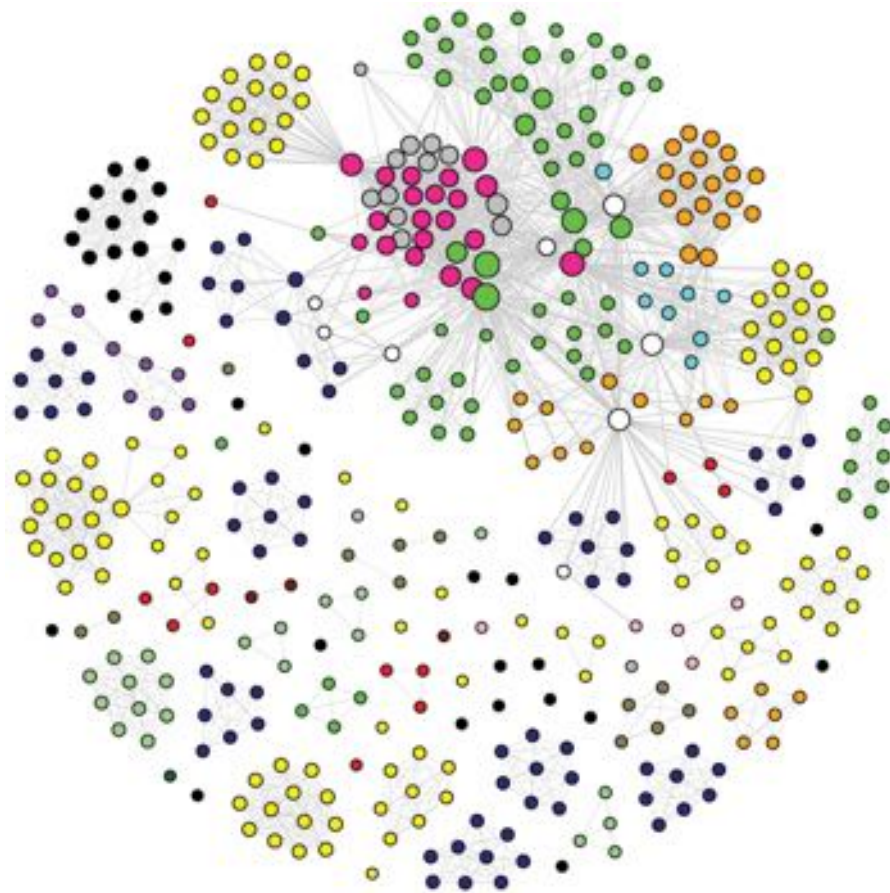




<http://nbviewer.ipython.org/github/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/>



<https://grey.colorado.edu/emergent/>



- Disorders usually first diagnosed in infancy, childhood or adolescence
- Delirium, dementia, and amnesia and other cognitive disorders
- Mental disorders due to a general medical condition
- Substance-related disorders
- Schizophrenia and other psychotic disorders
- Mood disorders
- Anxiety disorders
- Somatoform disorders
- Factitious disorders
- Dissociative disorders
- Sexual and gender identity disorders
- Eating disorders
- Sleep disorders
- Impulse control disorders not elsewhere classified
- Adjustment disorders
- Personality disorders
- Symptom is featured equally in multiple chapters

Borsboom et al. (2011)

# Faster, more effective evaluation

- Publishing is a frustratingly slow process
- Reliability of peer review is low (Bornmann et al., 2010)
- New publishing and evaluation models are continually emerging



# Registered Replication Report: Schooler and Engstler-Schooler (1990)

Perspectives on Psychological Science

2014, Vol. 9(5) 556–578

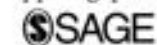
© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/1745691614545653

pps.sagepub.com



**Proposing Authors: This proposal was initiated by the editors**

**Contributing authors (alphabetical order):** Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., Bornstein, B. H., Bouwmeester, S., Brandimonte, M. A., Brown, C., Buswell, K., Carlson, C., Carlson, M., Chu, S., Cislak, A., Colarusso, M., Colloff, M. F., Dellapaolera, K. S., Delvenne, J.-F., Di Domenico, A., Drummond, A., Echterhoff, G., Edlund, J. E., Eggleston, C. M., Fairfield, B., Franco, G., Gabbert, F., Gamblin, B. W., Garry, M., Gentry, R., Gilbert, E. A., Greenberg, D. L., Halberstadt, J., Hall, L., Hancock, P. J. B., Hirsch, D., Holt, G., Jackson, J. C., Jong, J., Kehn, A., Koch, C., Kopietz, R., Körner, U., Kunar, M. A., Lai, C. K., Langton, S. R. H., Leite, F. P., Mammarella, N., Marsh, J. E., McConaughy, K. A., McCoy, S., McIntyre, A. H., Meissner, C. A., Michael, R. B., Mitchell, A. A., Mugayar-Baldocchi, M., Musselman, R., Ng, C., Nichols, A. L., Nunez, N. L., Palmer, M. A., Pappagianopoulos, J. E., Petro, M. S., Poirier, C. R., Portch, E., Rainsford, M., Rancourt, A., Romig, C., Rubínová, E., Sanson, M., Satchell, L., Sauer, J. D., Schweitzer, K., Shaheed, J., Skelton, F., Sullivan, G. A., Susa, K. J., Swanner, J. K., Thompson, W. B., Todaro, R., Ulatowska, J., Valentine, T., Verhoeven, P. P. J. L., Vranka, M., Wade, K. A., Was, C. A., Weatherford, D., Wiseman, K., Zaksasite, T., Zuj, D. V., Zwaan, R. A.

**Protocol vetted by:** Jonathan W. Schooler

**Protocol edited by:** Daniel J. Simons

**Multilab direct replication of:** Study 4 (modified) and Study 1 from Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, 22, 36–71.

**Data and registered protocols:** <https://osf.io/ybeur/>

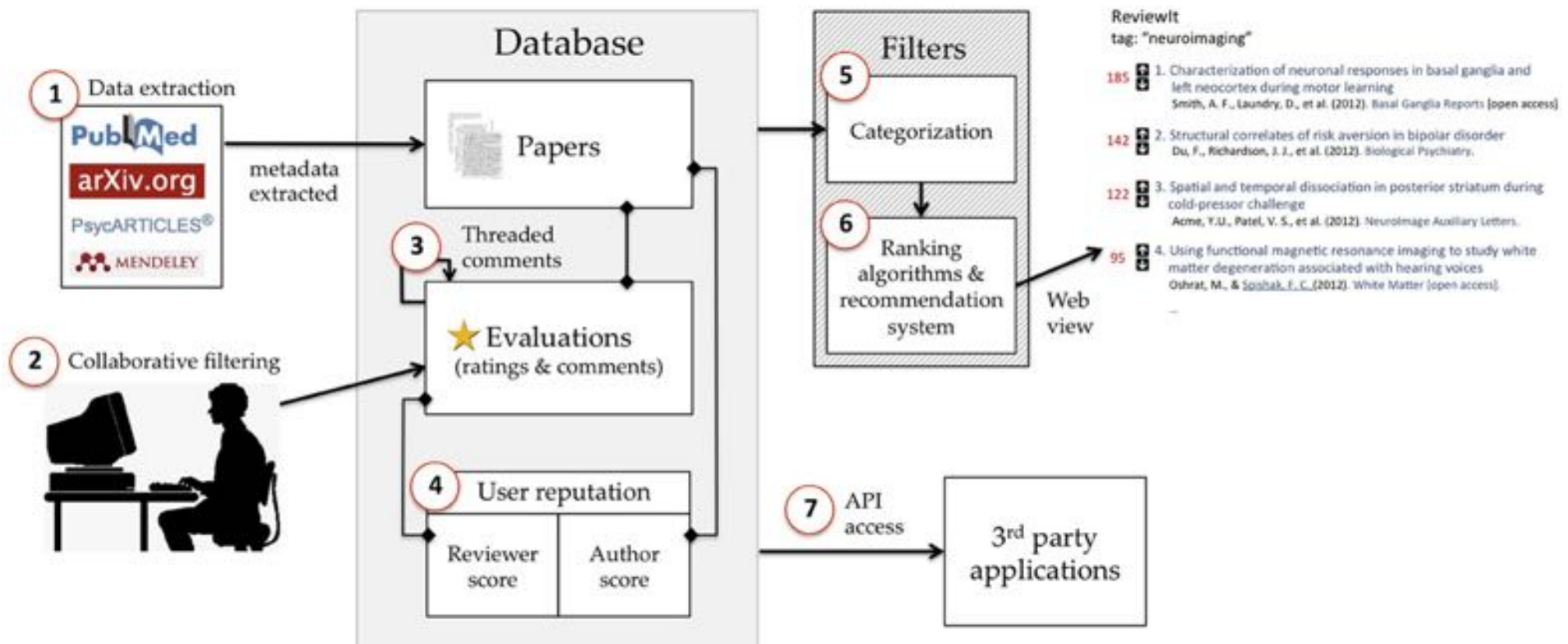
**Citation:** Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, S., Birch, S., Birt, A. R., ... Zwaan, R. A. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science*, 9, 556–578.



*the*  
**WINNOWER**

F1000  
Research





Yarkoni (2012)

I am a Birmingham Fellow (Assistant Professor) in the School of Biosciences at the University of Birmingham, UK. My research focuses on marine metagenomics, environmental sequencing, and computational biology + data visualization.



 Open Access

 Global Reach



## Selected works

### An Introduction to Social Media for Scientists

(2013) Bik, Goldstein. *PLoS Biol*

 read fulltext

highly cited

highly saved

highly viewed

highly discussed

highly viewed

### Dramatic Shifts in Benthic Microbial Eukaryote Communities following the Deepwater Horizon Oil Spill

(2012) Bik, Halanych, Sharma, et al. *PLoS ONE*

 read fulltext

highly cited

highly saved

highly viewed

highly discussed

highly viewed

cited

### PhyloSift: Phylogenetic analysis of genomes and metagenomes

(2013) figshare.

 view dataset

highly viewed

+4

highly viewed

discussed

discussed

## Key profile metrics

164.8k

 views across

 25 articles



# How much psycho-?

- How much of this is specific to psychology?
- Not that much
  - And that's okay!
- Still need to tailor methods/resources to audience

# Psychoinformatics vs...

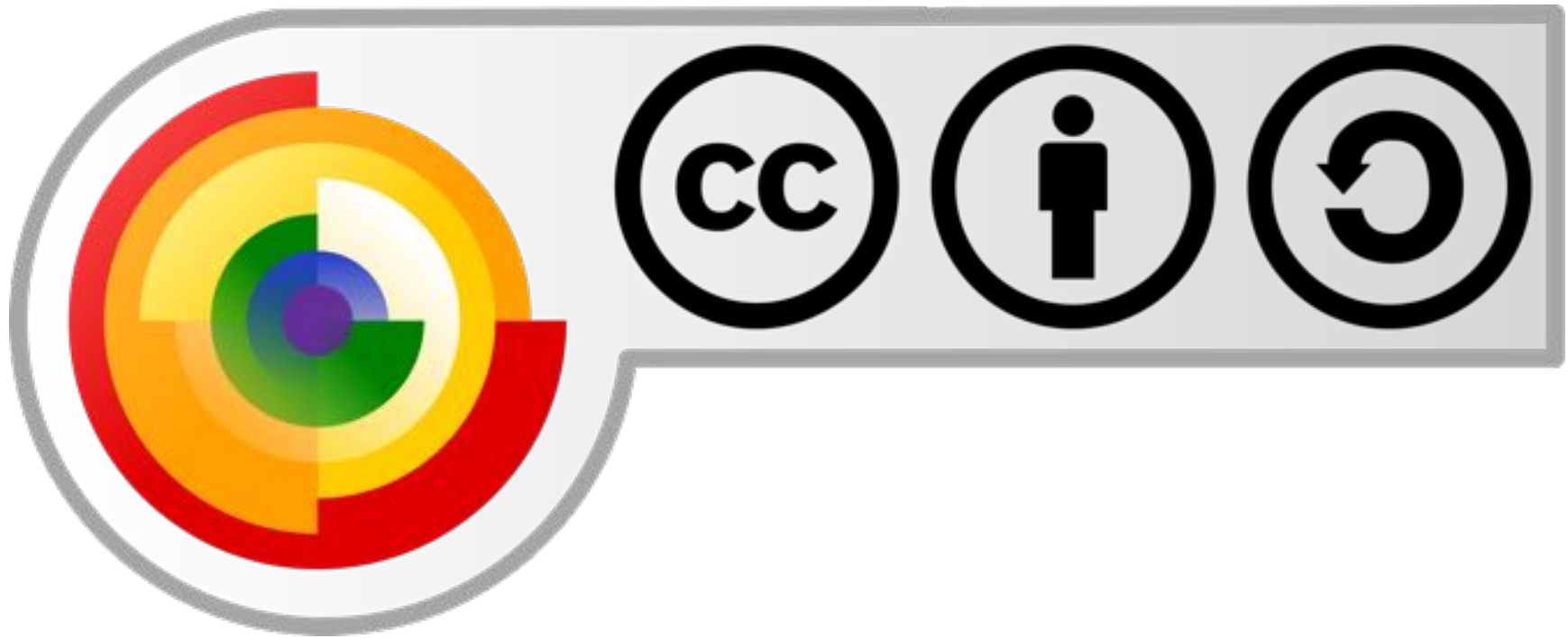
- Quantitative psychology?
- Computational psychology/social science?
- Data science?
- Scientific computing?

# Target: statistics

- If we do everything right, psychoinformatics should become invisible
- Follow the example of statistics
  - Basic statistical literacy is required for all psychologists
  - No one says “did you use statistics to analyze your data?”
  - Scientific computing skills are just as (more?) important

# Let's make some magic





These slides may be freely distributed and re-used under the terms of the CC BY-SA 3.0 license.

<https://creativecommons.org/licenses/by-sa/3.0/>