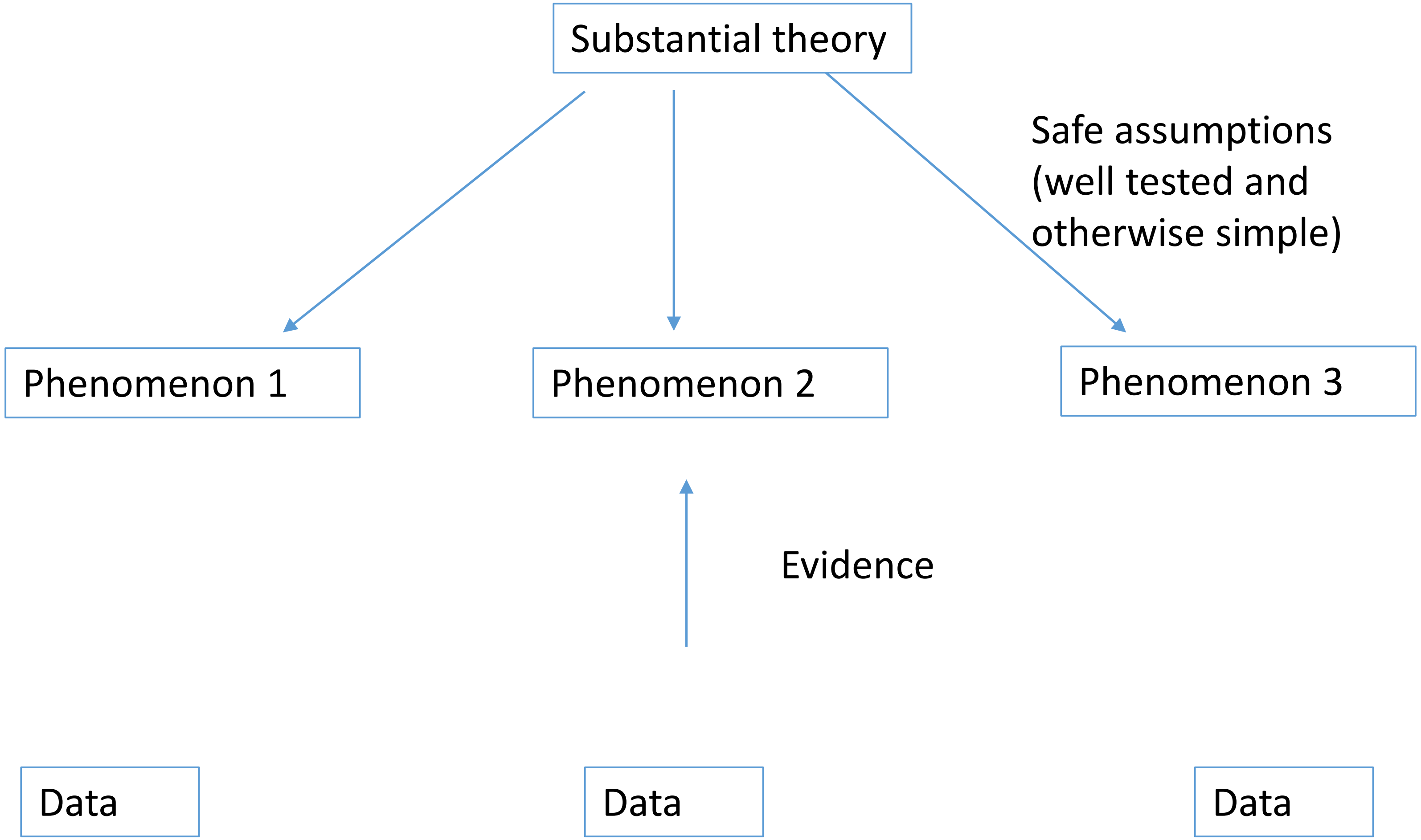
The inner workings of Registered Reports

Zoltan Dienes

University of Sussex



Substantial theory

Safe assumptions (well tested and otherwise simple)

Phenomenon 1

Phenomenon 2

Phenomenon 3

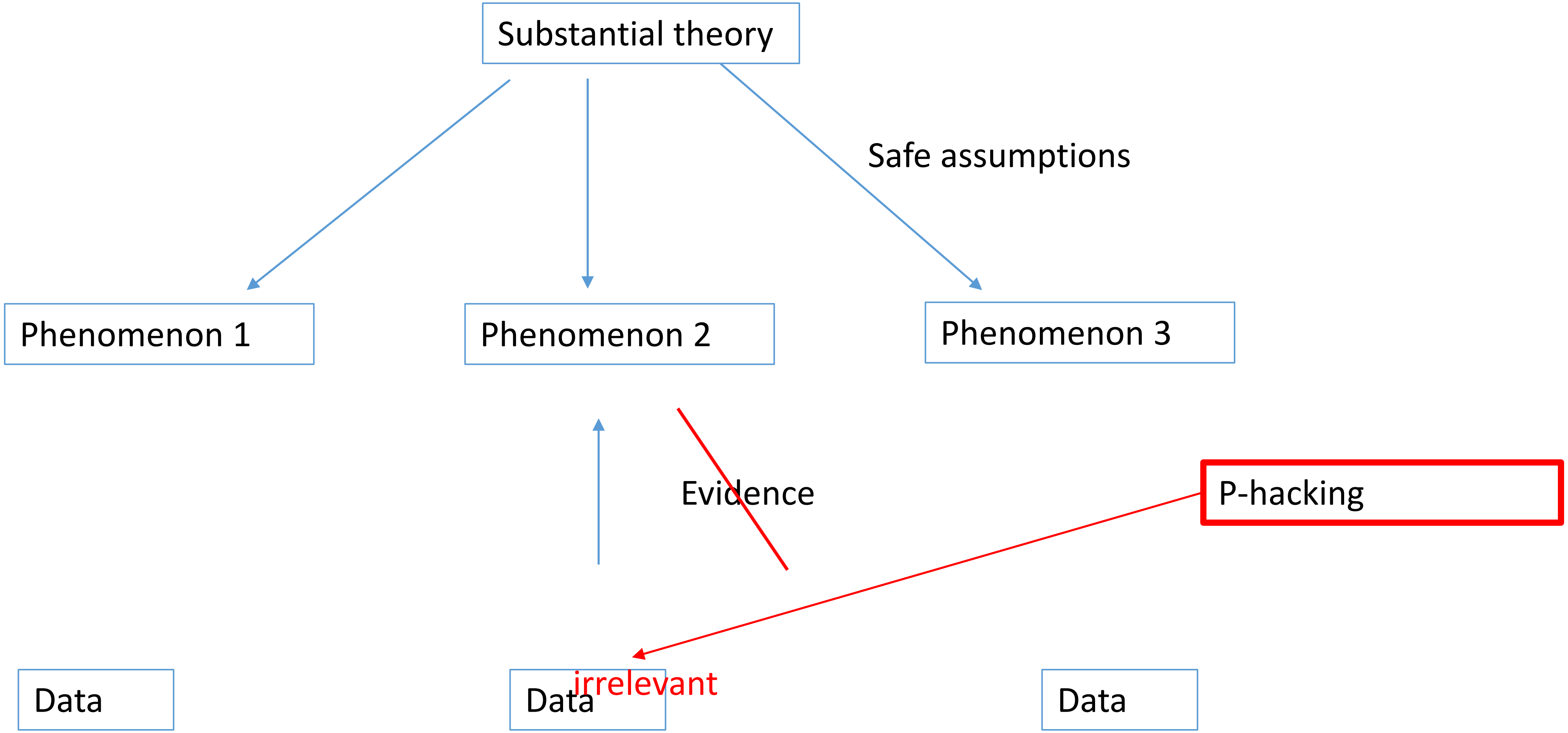
Evidence

Data

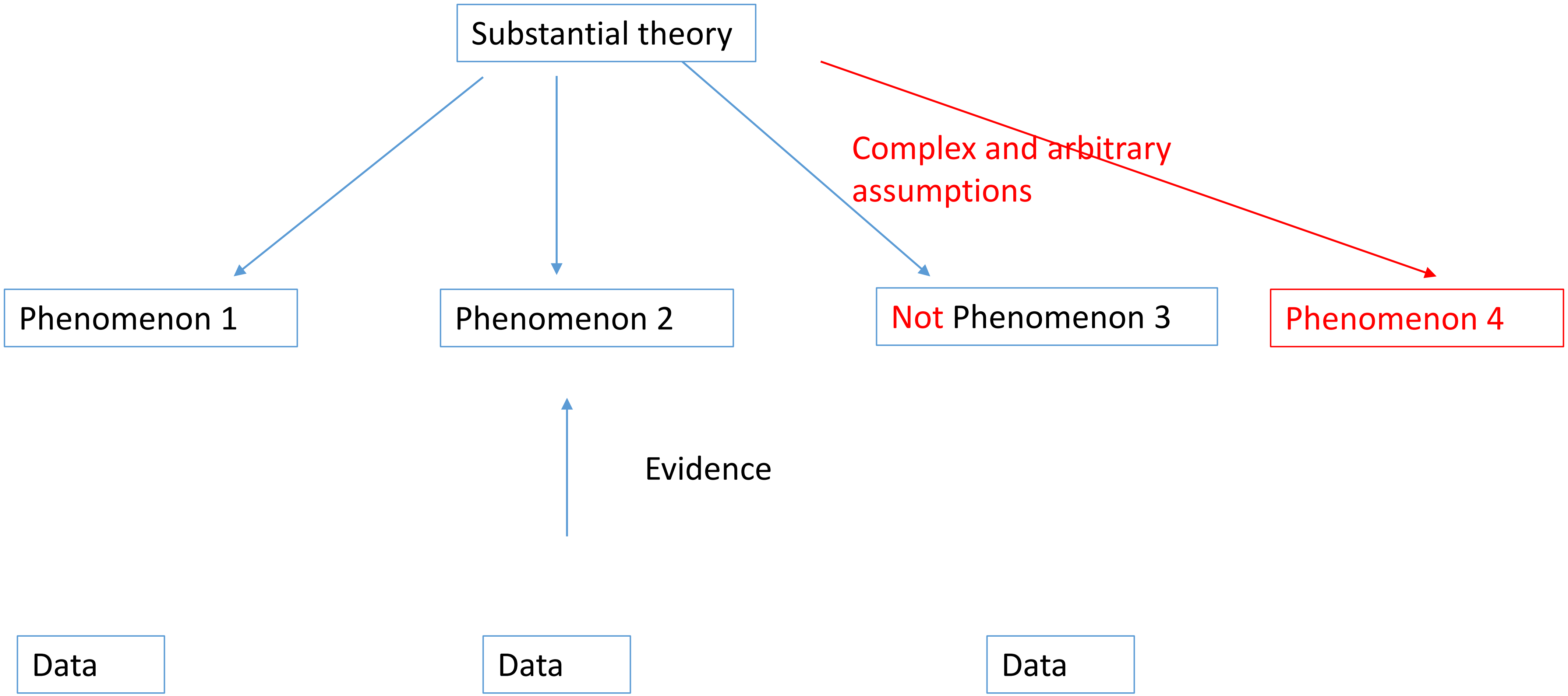
Data

Data

When data are seen first:



When data are seen first:



Substantial theory

Complex and arbitrary assumptions

Phenomenon 1

Phenomenon 2

Not Phenomenon 3

Phenomenon 4

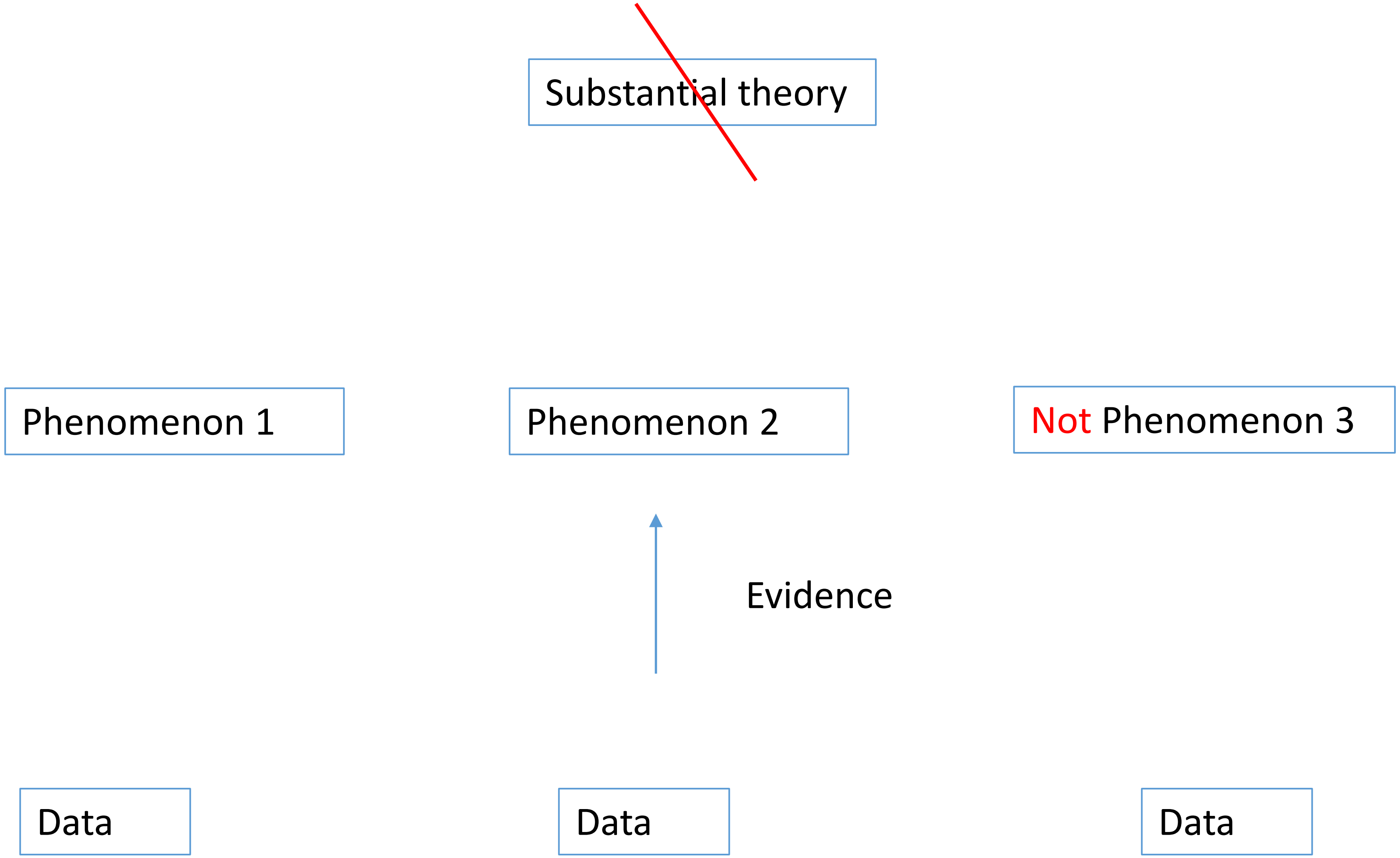
Evidence

Data

Data

Data

When data are seen first:



~~Substantial theory~~

Phenomenon 1

Phenomenon 2

Not Phenomenon 3

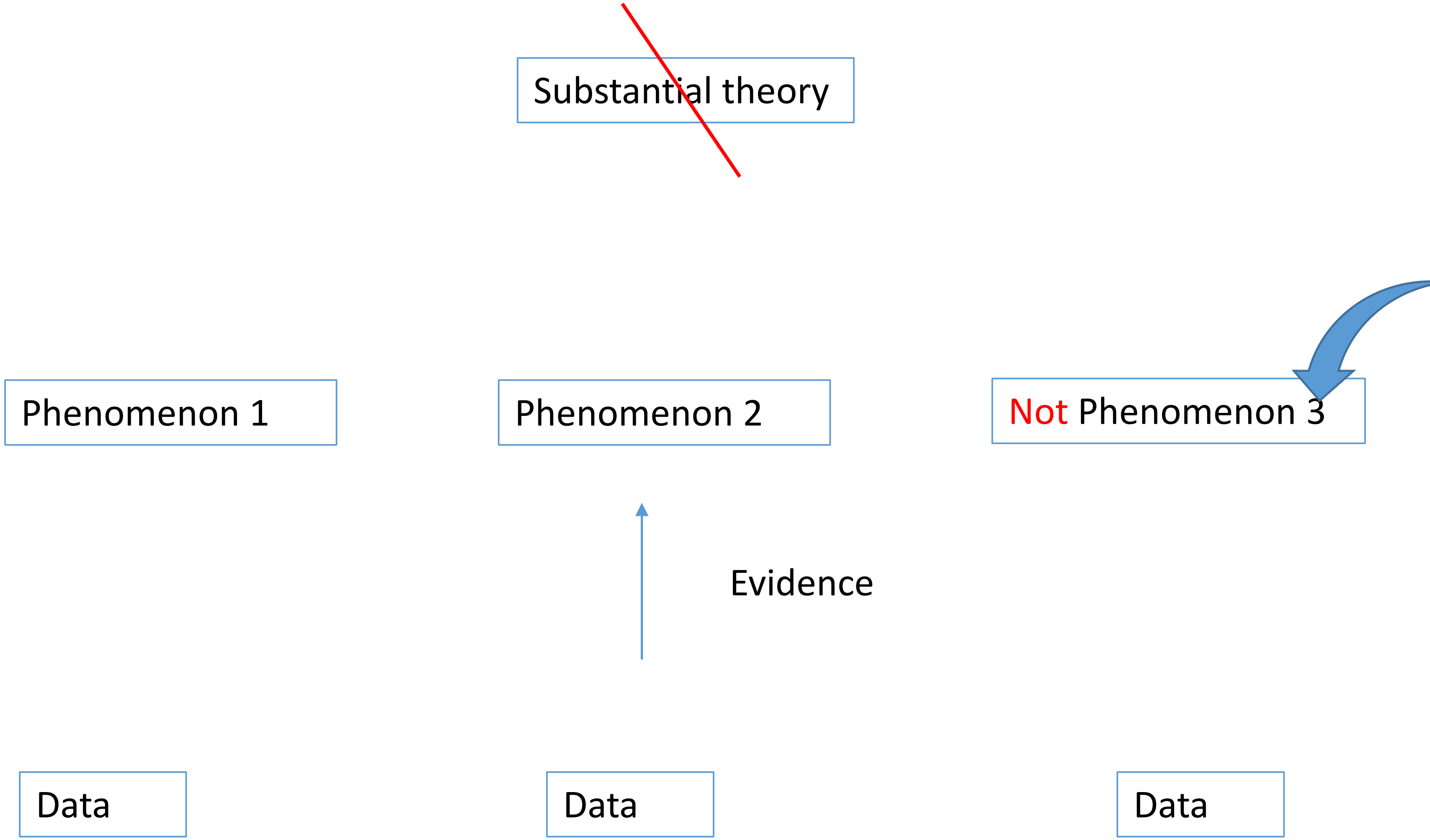
Evidence

Data

Data

Data

When data are seen first:



~~Substantial theory~~

Phenomenon 1

Phenomenon 2

Not Phenomenon 3

Evidence

Data

Data

Data

A phenomena chaser

Writing papers in light of data can lead to:

1) p-hacking (/B-hacking) the predicted result -> data are no longer evidential regarding phenomenon

2) Introducing complex assumptions to make theory fit phenomena-> data no longer evidential regarding theory

3) Giving up theory and chasing phenomena -> mindless research

A paper should be accepted if it helps in an aspect of

1) Setting up a substantial theory

2) Which uses safe assumptions to make predictions (predict phenomena)

3) Which are severely tested

(But should not be accepted on the basis of whether results support a theory or not)

Registered Reports:

Accept paper before data are collected based on:

1) A substantial theory being tested.

2) Assumptions connecting theory to predictions being safe.

3) Analytic flexibility being tied down while ensuring sensitive results.

Registered Reports:

Accept paper before data are collected based on:

1) A substantial theory being tested.

Cortex: Submissions will be evaluated with respect to “the importance of the research question(s)”

Registered Reports:

Accept paper before data are collected based on:

2) Assumptions connecting theory to predictions being safe.

Cortex: “Full descriptions must be provided of any outcome-neutral criteria that must be met for successful testing of the stated hypotheses. Such quality checks might include the absence of floor or ceiling effects in data distributions, positive controls, or other quality checks that are orthogonal to the experimental hypotheses.”

If predictions not confirmed, need to make sure assumptions safe, so theory takes the blame.

Substantial theory: “Belief in free will induces one to overcome automatic habits and hence behave prosocially”

Prediction:
“After reading a Francis Crick free will rather than control passage, people will give more milligrams of hot sauce in to someone who doesn’t like it”

Assumption:
The intervention – reading statements about free will – actually changes free will beliefs.

If we fail to confirm predictions could we just as plausibly reject this auxiliary as reject the substantial theory? If so, the test is not a good one.

Outcome neutral test: Belief in free will changes.

Outcome neutral tests: Those specified MUST be passed!

Distinguish:

Checks that are useful but not essential (did participants take equal amount of time to read intervention and control passages?)

Registered Reports:

Accept paper before data are collected based on:

3) Analytic flexibility being tied down while ensuring sensitive results.

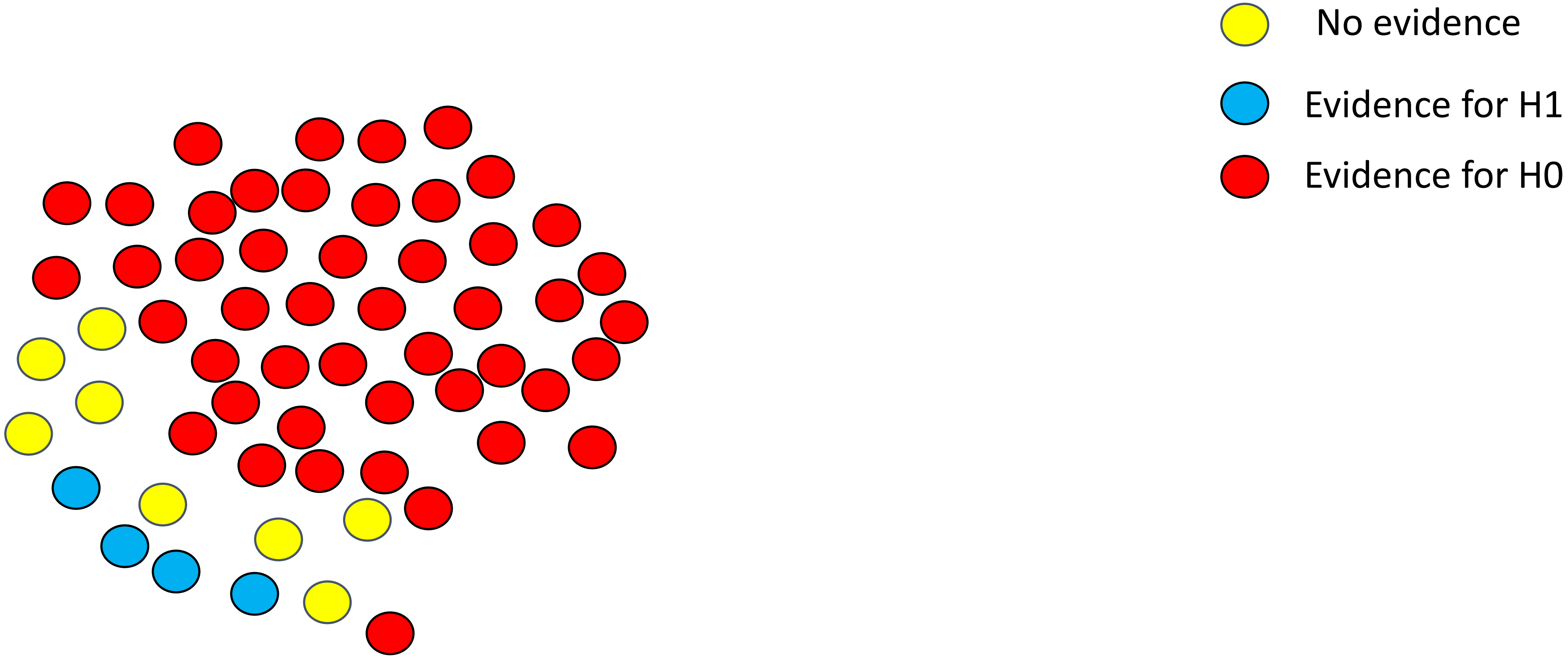
“Studies involving Neyman-Pearson inference should include a statistical power analysis, and please note that the default threshold for declaring statistical significance is α=.02 rather than the conventional α=.05. Estimated effect sizes should be justified with reference to the existing literature or theory. Since publication bias overinflates published estimates of effect size, power analysis must be based on the lowest available or meaningful estimate of the effect size. Where relevant, the a priori power must be 0.9 or higher for all proposed hypothesis tests.”

Contrast RSOS: No power requirements (but outcome neutral tests must be passed).

How is variable 1 coded? How is variable 2 coded?   Exclude if …? How ….

5   X 3   X 4   X ….

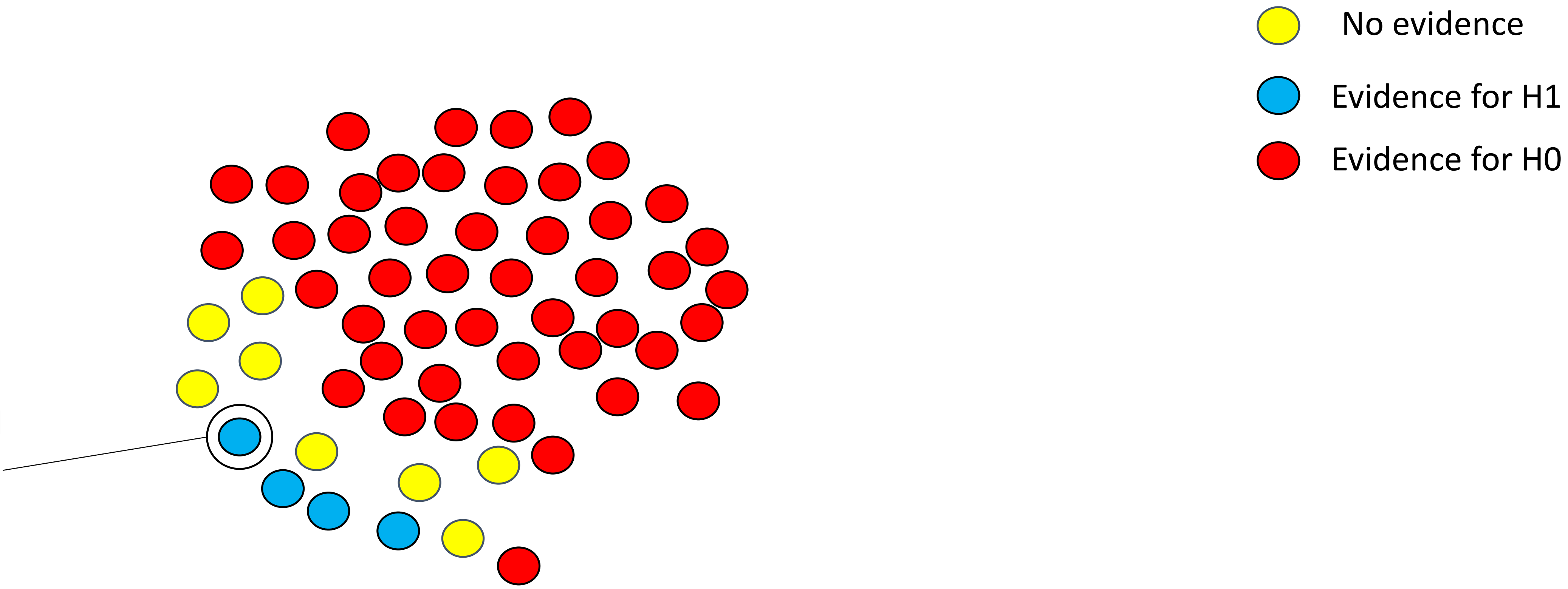
= the **multiverse**   (Carp 2012; Steegen et al 2016)



How is variable 1 coded? How is variable 2 coded?   Exclude if …? How ….

5 X 3 X 4 X ….

= the **multiverse** (Carp 2012; Steegen et al 2016)

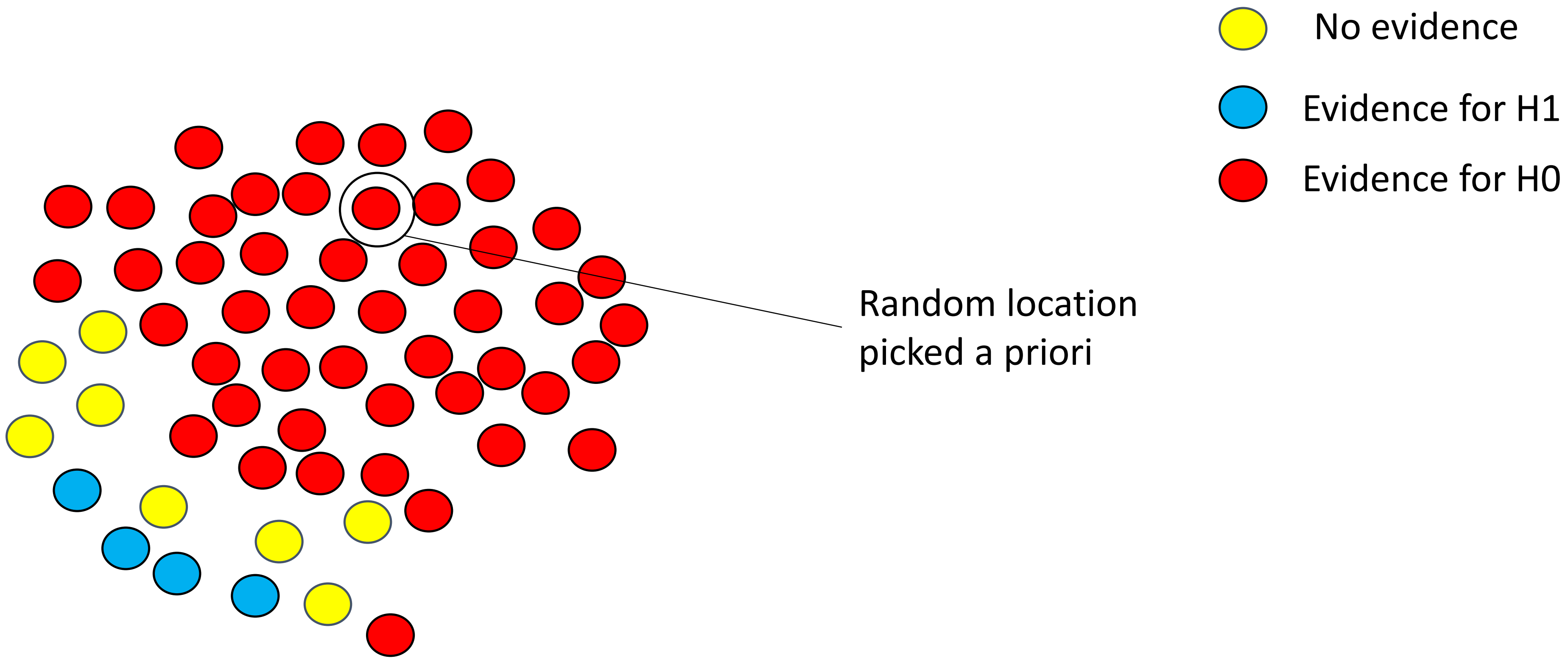


Post hoc stroll through the garden of forking paths

How is variable 1 coded? How is variable 2 coded? Exclude if …? How ….

5 X 3 X 4 X ….

= the **multiverse** (Carp 2012; Steegen et al 2016)



🟡 No evidence

🔵 Evidence for H1

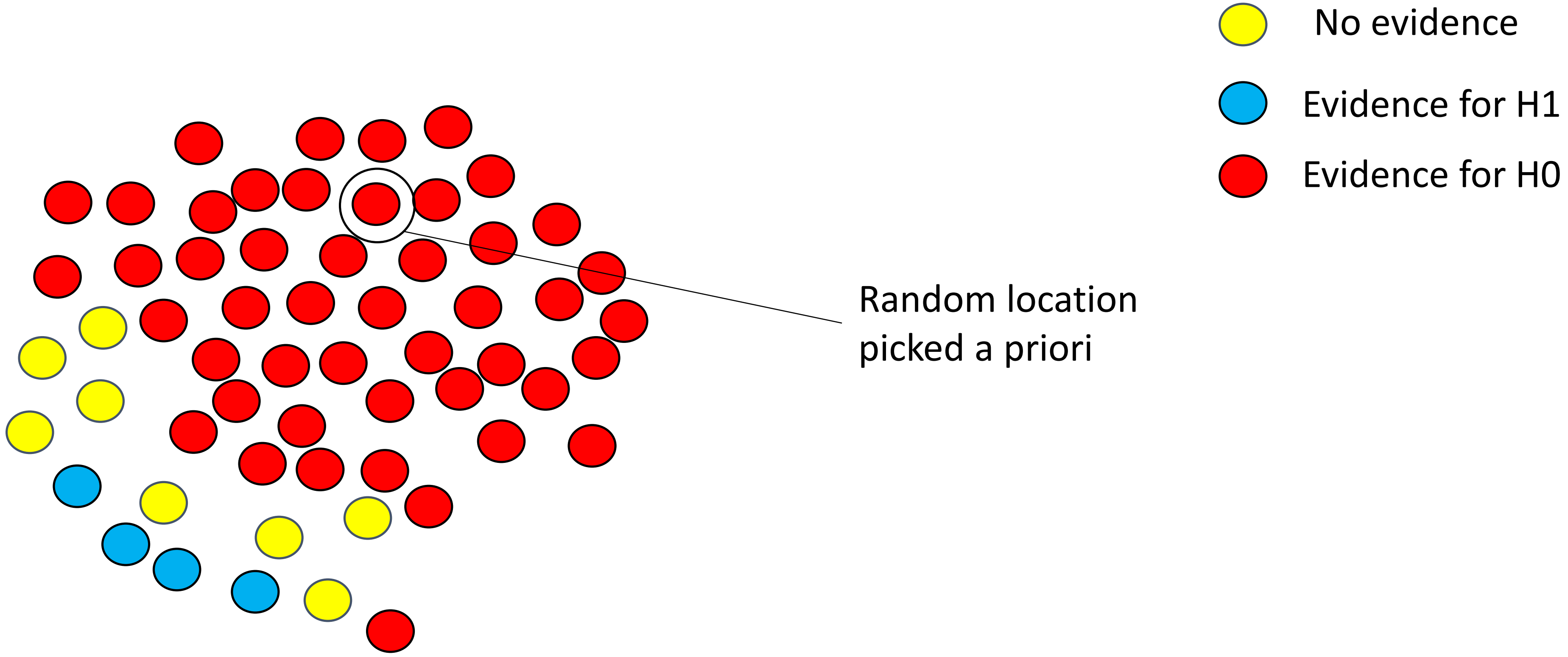
🔴 Evidence for H0

Random location
picked a priori

How is variable 1 coded? How is variable 2 coded? Exclude if …? How ….

5 X 3 X 4 X ….

= the **multiverse** (Carp 2012; Steegen et al 2016)



🟡 No evidence

🔵 Evidence for H1

🔴 Evidence for H0

Random location picked a priori

Choosing location of multiverse in advance probably yields most common conclusion: Objective evidential relation between data and hypotheses respected (probably)

Power: Minimally theoretically interesting effect size that is just plausible.

Past studies with a different DV found a Cohen’s d = 0.4.

Power: Minimally theoretically interesting effect size that is just plausible.

Past studies with a different DV found a Cohen’s d = 0.4.

But standardized effect sizes are measures of signal relative to noise – change number of trials, number of items, factors in analysis, … Cohen’s d will change.

Change RTs to % correct, why expect same signal to noise?

The N of past studies implies a minimal effect of interest …..
But how was that decided? Question only pushed back.

“The committee decided 3 units is minimal” But why? We need reasons that can be criticized.

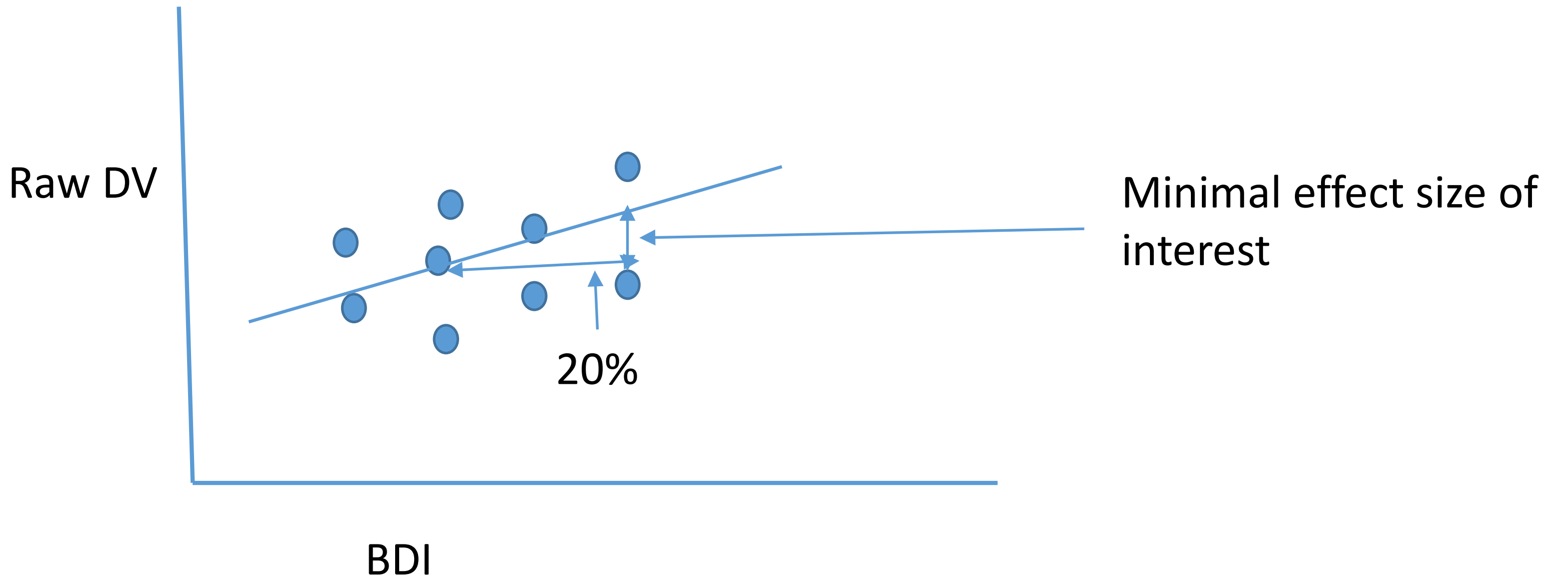
Power: Minimally theoretically interesting effect size that is just plausible.

1) Lower limit of 95% CI of raw effect from past studies. Is it still theoretically interesting?

Power: Minimally theoretically interesting effect size that is just plausible.

1) Lower limit of 95% CI of raw effect from past studies. Is it still theoretically interesting?

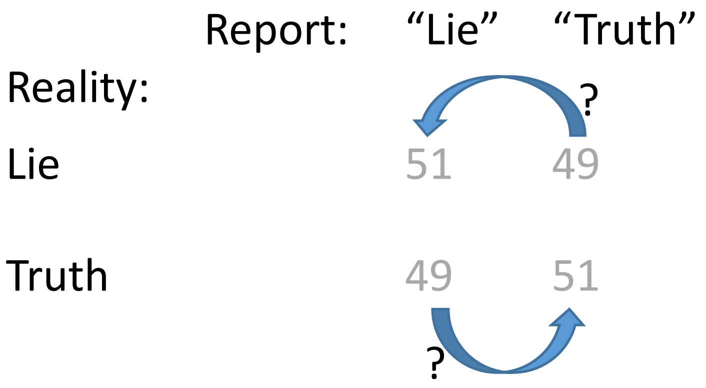
2) Clinical relevance. Button et al (2015): A minimal clinical significant effect according to depressed patients is a 20% change on the BDI.



Standard interview

	Report:	“Lie”	“Truth”
Reality:			
Lie		51	49
Truth		49	51

New method



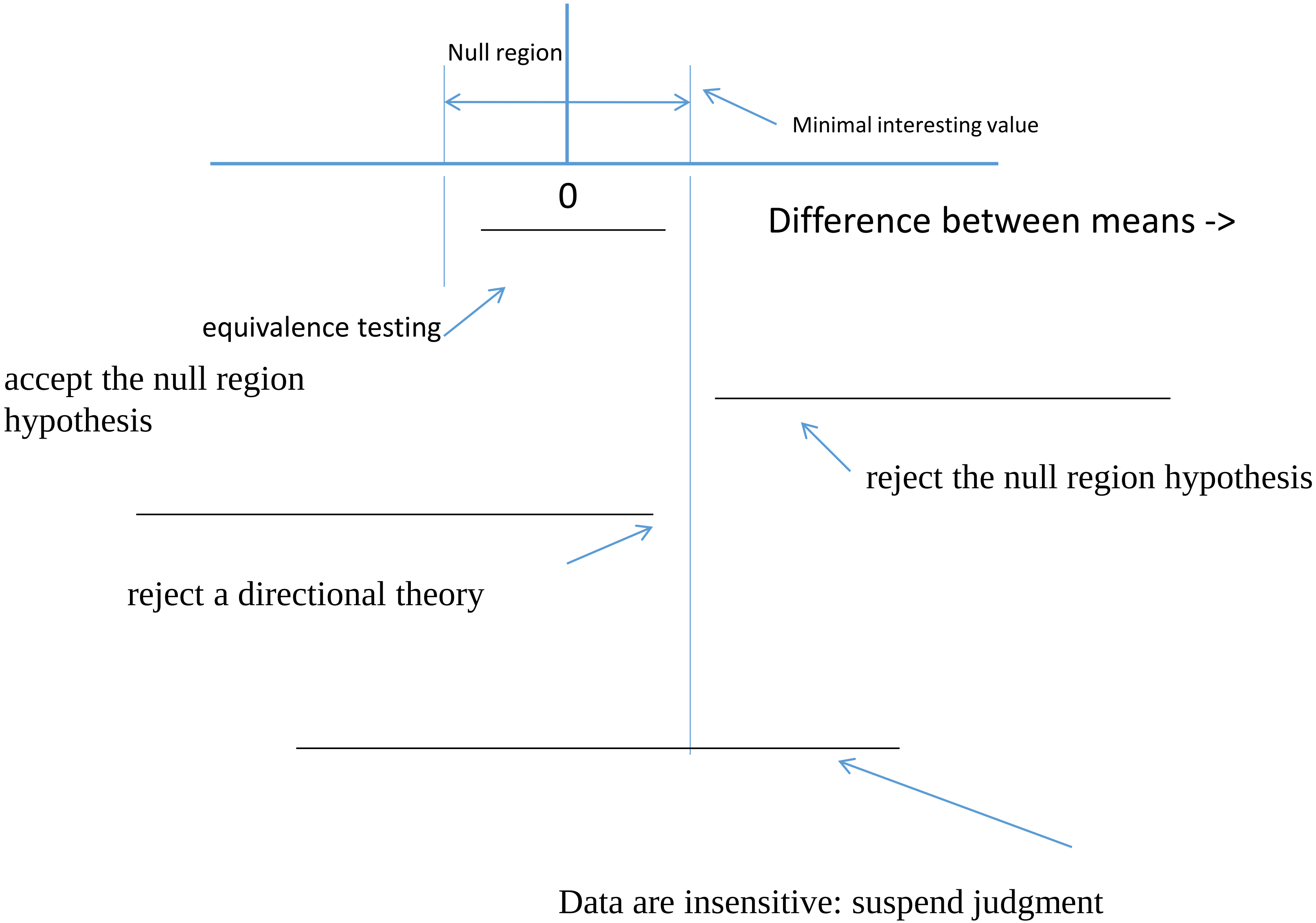
	Report:	“Lie”	“Truth”
Reality:			
Lie		51	49
Truth		49	51

When the point is an end user, the end user can decide how much is minimally enough (cf and contrast Freedman & Spiegelhalter 1983)

But for purely theoretical research minimal interesting effect sizes hard to pin down.

“For equivalence testing (as with classical power analysis), the minimally interesting effect size should be determined based on a justification for why that effect is theoretically or practically interesting/plausible, and not according to past sample sizes alone. “

The four principles of *inference by intervals* *(Greenwald, 1975)*:



Null region

Minimal interesting value

0

Difference between means ->

equivalence testing

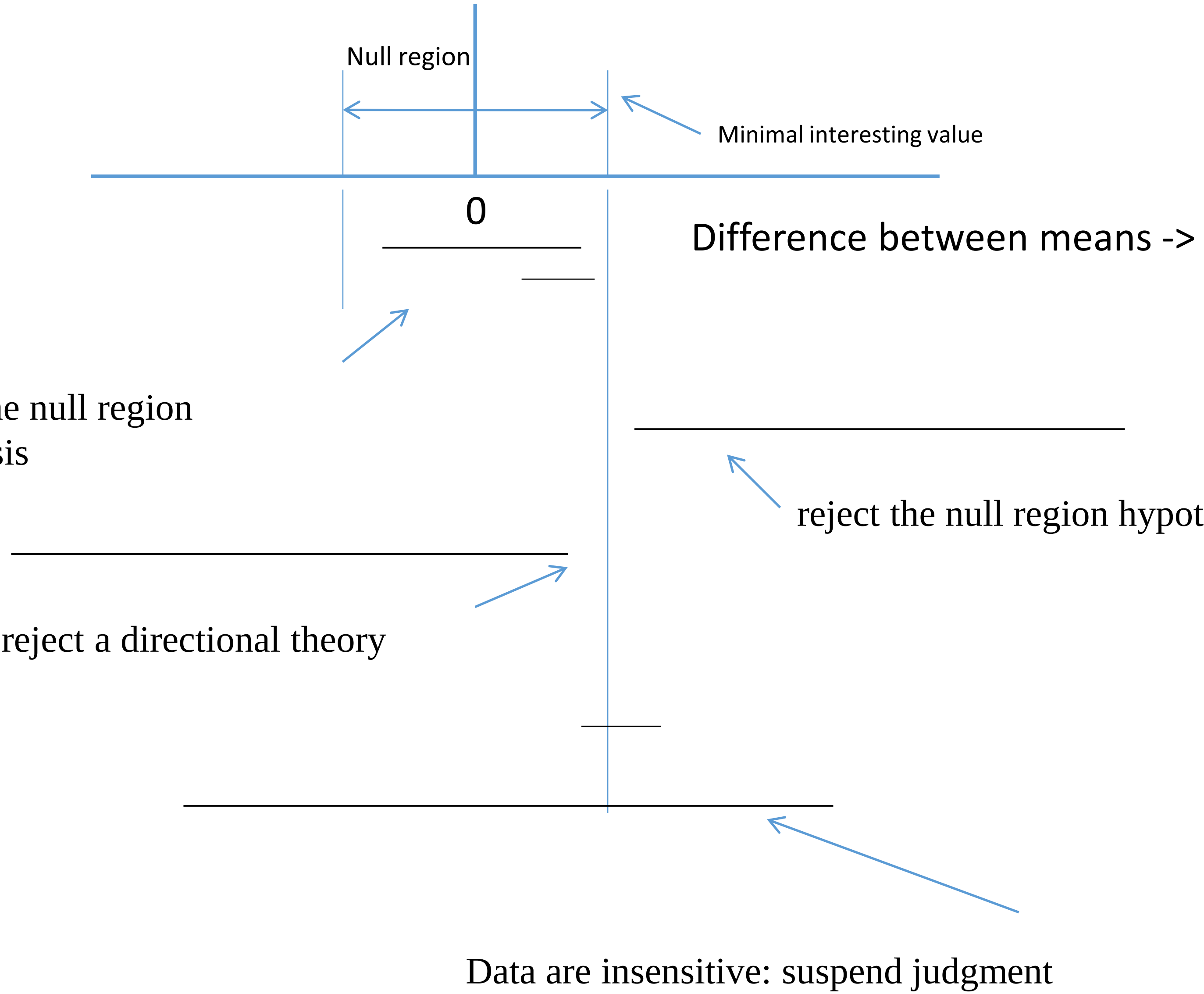
accept the null region hypothesis

reject the null region hypothesis

reject a directional theory

Data are insensitive: suspend judgment

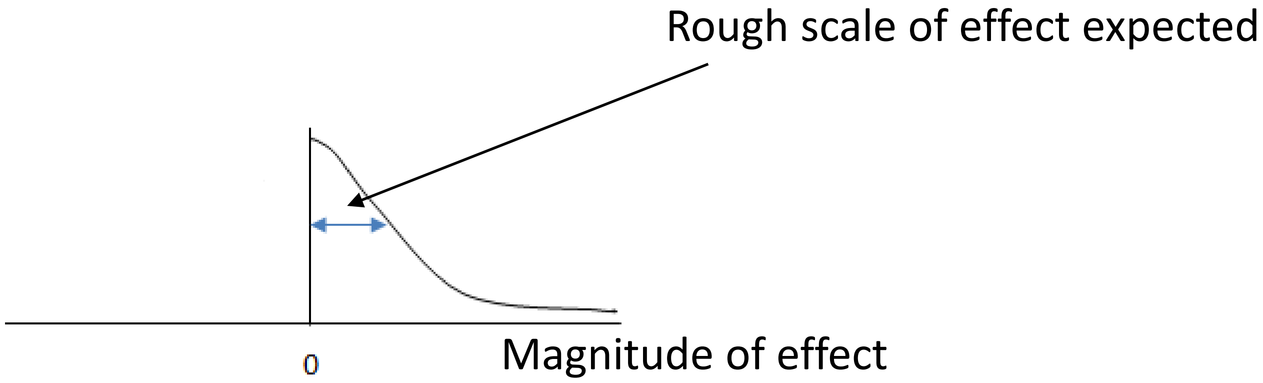
The four principles of *inference by interval:*


Null region
Minimal interesting value
0
Difference between means ->
accept the null region hypothesis
reject the null region hypothesis
reject a directional theory
Data are insensitive: suspend judgment

For Bayes factors “Authors should indicate what distribution will be used to represent the predictions of the theory and how its parameters will be specified. For example, will you use a uniform up to some specified maximum, or a normal/half-normal to represent a likely effect size, or a JZS/Cauchy with a specified scaling constant? For inference by Bayes factors, authors must be able to guarantee data collection until the Bayes factor is at least 6 times in favour of the experimental hypothesis over the null hypothesis (or vice versa). “

Contrast:
Royal Society Open Science (no thresholds),
Nature Human Behaviour (B > 10)

Model of H1:



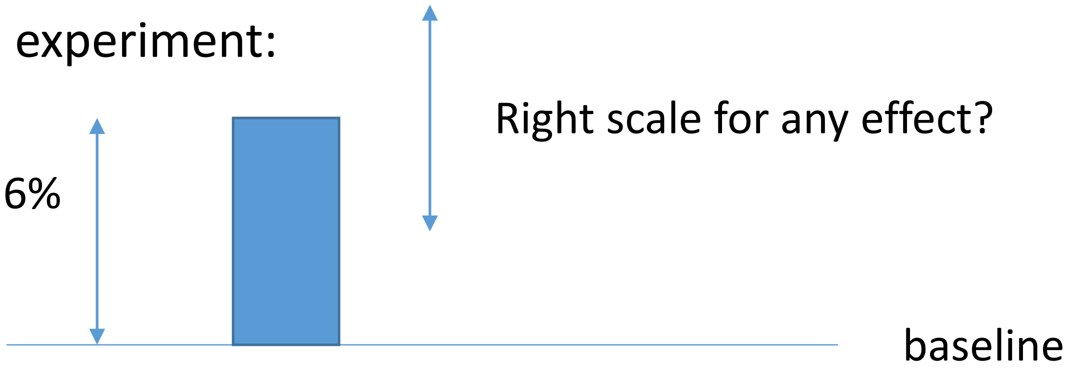
Determine rough size of effect expected on theory

Ziori & Dienes 2015:

Subjects learn whether a sequence of faces is rule governed:

Stimuli attractive vs normal
Gender of participants
Gender of faces

Average learning in experiment:



So used sd = 6% in half-normal for all effects in 3-way ANOVA

Cortex “Authors with resource limitations are permitted to specify a maximum feasible sample size at which data collection must cease regardless of the Bayes factor; however to be eligible for advance acceptance this number must be sufficiently large that inconclusive results at this sample size would nevertheless be an important message for the field. “

DV = bias
IV1 = time, four blocks (within)
IV2 = group, depressed vs nondepressed (between)

Hypothesis
“The bias will decrease over blocks more slowly for depressed than non-depressed participants”

Test with 2 X 4 ANOVA

“The power to detect bias being above zero is 0.90 with N = 30 for α= .02”

DV = bias
IV1 = time, four blocks (within)
IV2 = group, depressed vs nondepressed (between)

Hypothesis
“The bias will decrease over blocks more slowly for depressed than non-depressed participants”

Test with 2 X 4 ANOVA

“The power to detect bias being above zero is 0.90 with N = 30 for α= .02”

“The power for the two –way interaction (df = 3) is 0.90 with N = 150 for α= .02”

Note hypothesis is a 1-df.

DV = bias
IV1 = time, four blocks (within)
IV2 = group, depressed vs nondepressed (between)

Hypothesis
"The bias will decrease over blocks more slowly for depressed than non-depressed participants"

Linear contrast

$L = (-\frac{3}{4}) \times B1 + (-\frac{1}{4}) \times B2 + \frac{1}{4} \times B3 + \frac{3}{4} \times B4$

Test of theory = $L_{\text{non-depressed}} - L_{\text{depressed}}$

Calculate power/BFs for THIS test

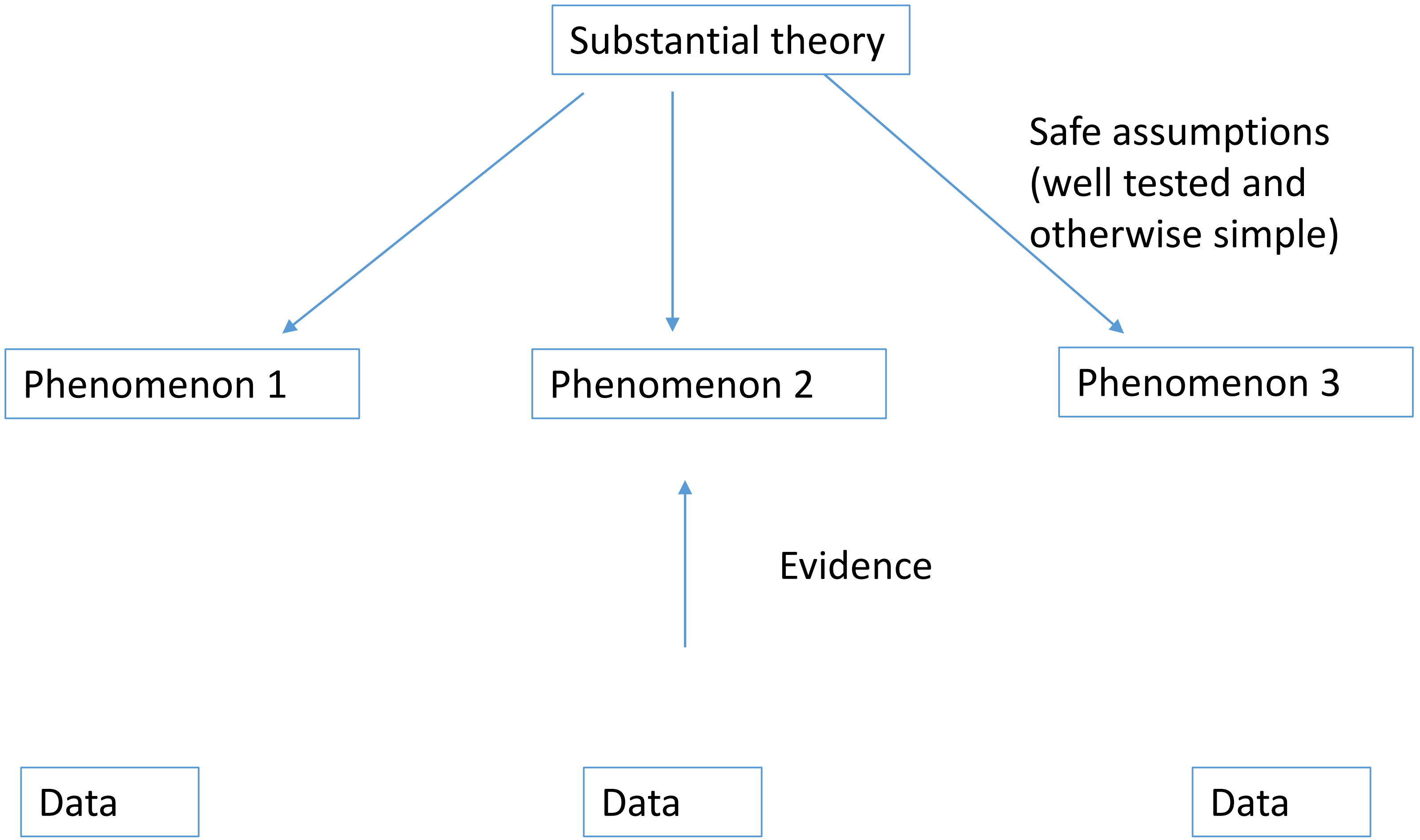
Substantial theory: What is the most general theory that could be disconfirmed?

Prediction: Is it 1-df?

Assumptions: How does prediction follow from theory? What test is needed?
(manipulation checks etc.)

Test: Need a statistical test for each prediction AND each assumption

EACH must have adequate power/reach BF threshold.


Substantial theory

Safe assumptions
(well tested and
otherwise simple)

Phenomenon 1

Phenomenon 2

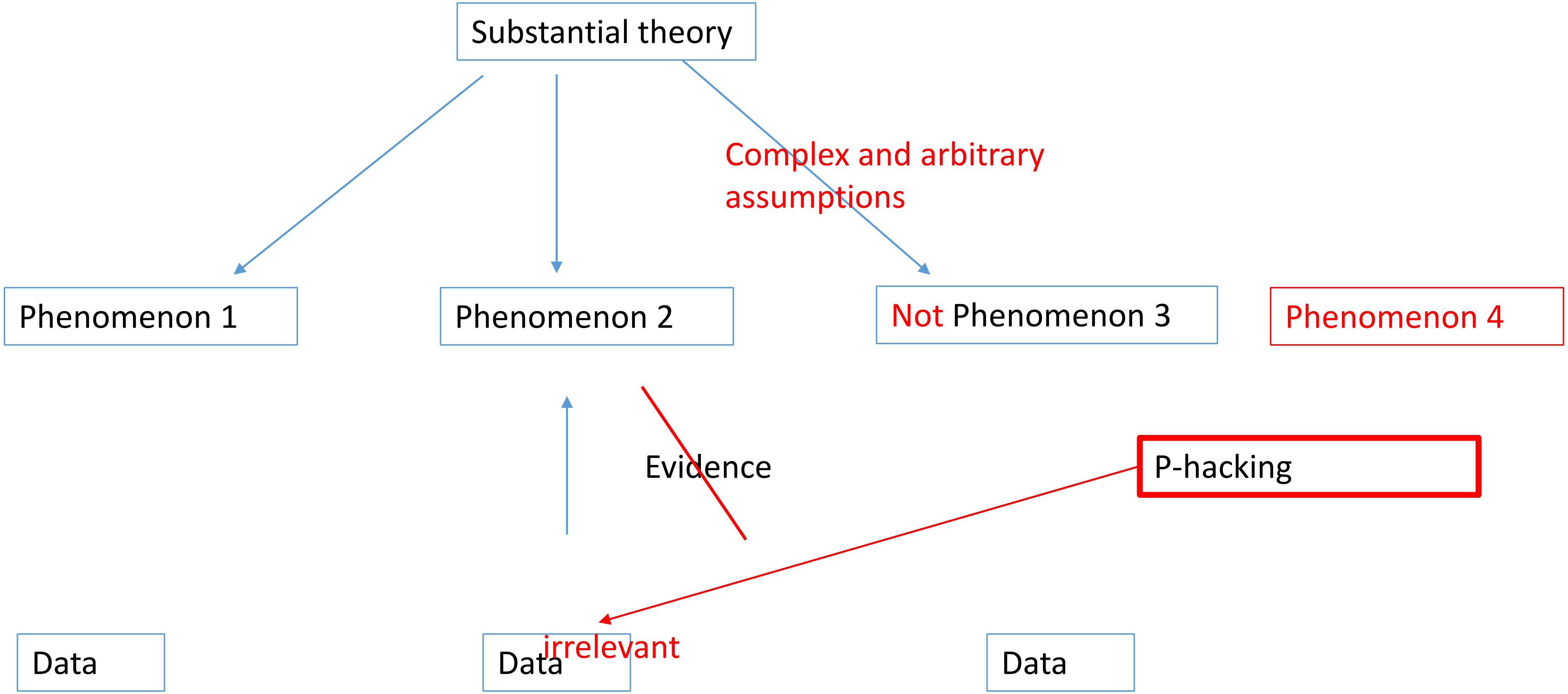
Phenomenon 3

Evidence

Data

Data

Data



Substantial theory

Complex and arbitrary assumptions

Phenomenon 1

Phenomenon 2

Not Phenomenon 3

Phenomenon 4

Evidence

P-hacking

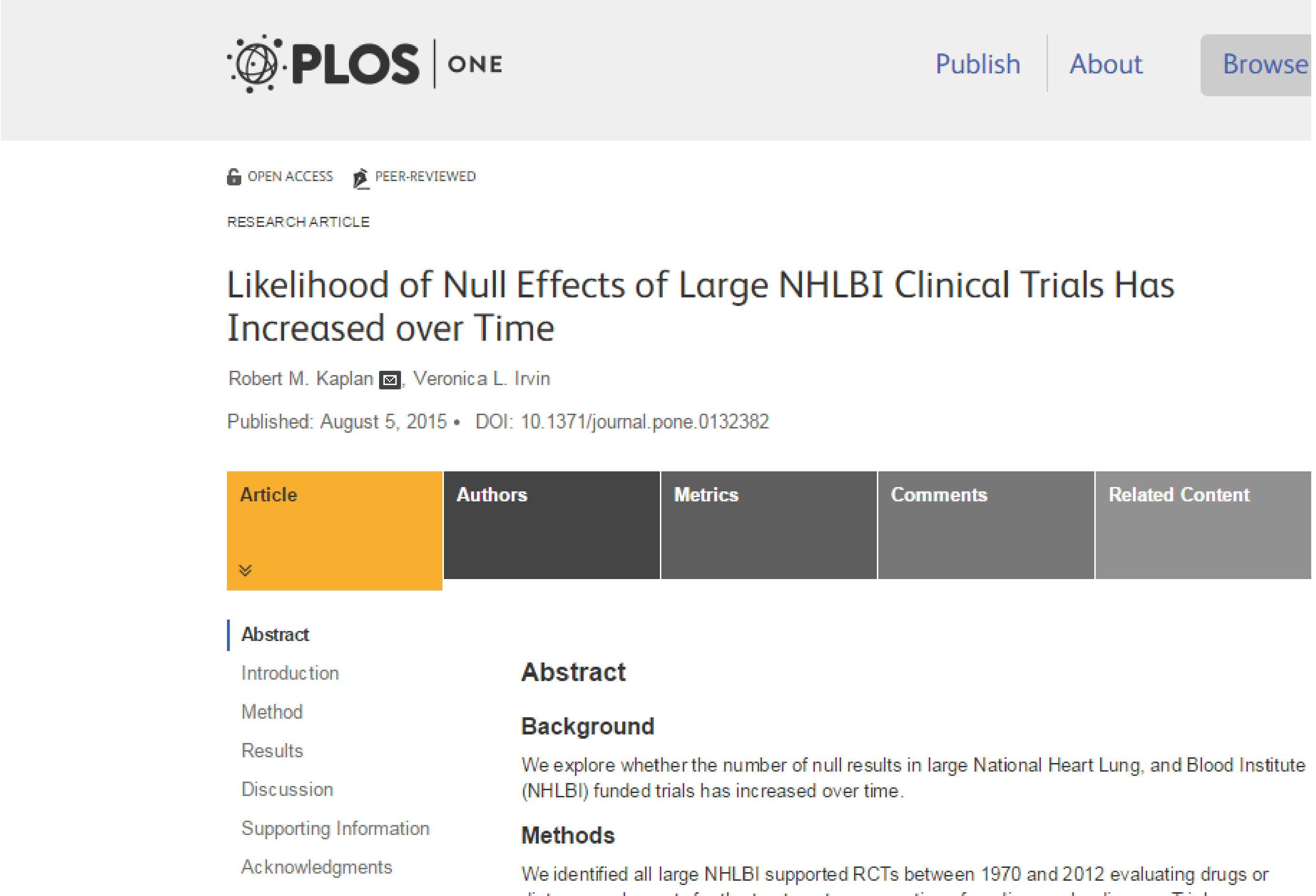
Data

Data irrelevant

Data

Number of studies finding medical interventions effective before preregistration introduced:
17/30 (55%)

Afterwards:



RESEARCH ARTICLE

Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time

Robert M. Kaplan, Veronica L. Irvin

Published: August 5, 2015 • DOI: 10.1371/journal.pone.0132382

Abstract

Background

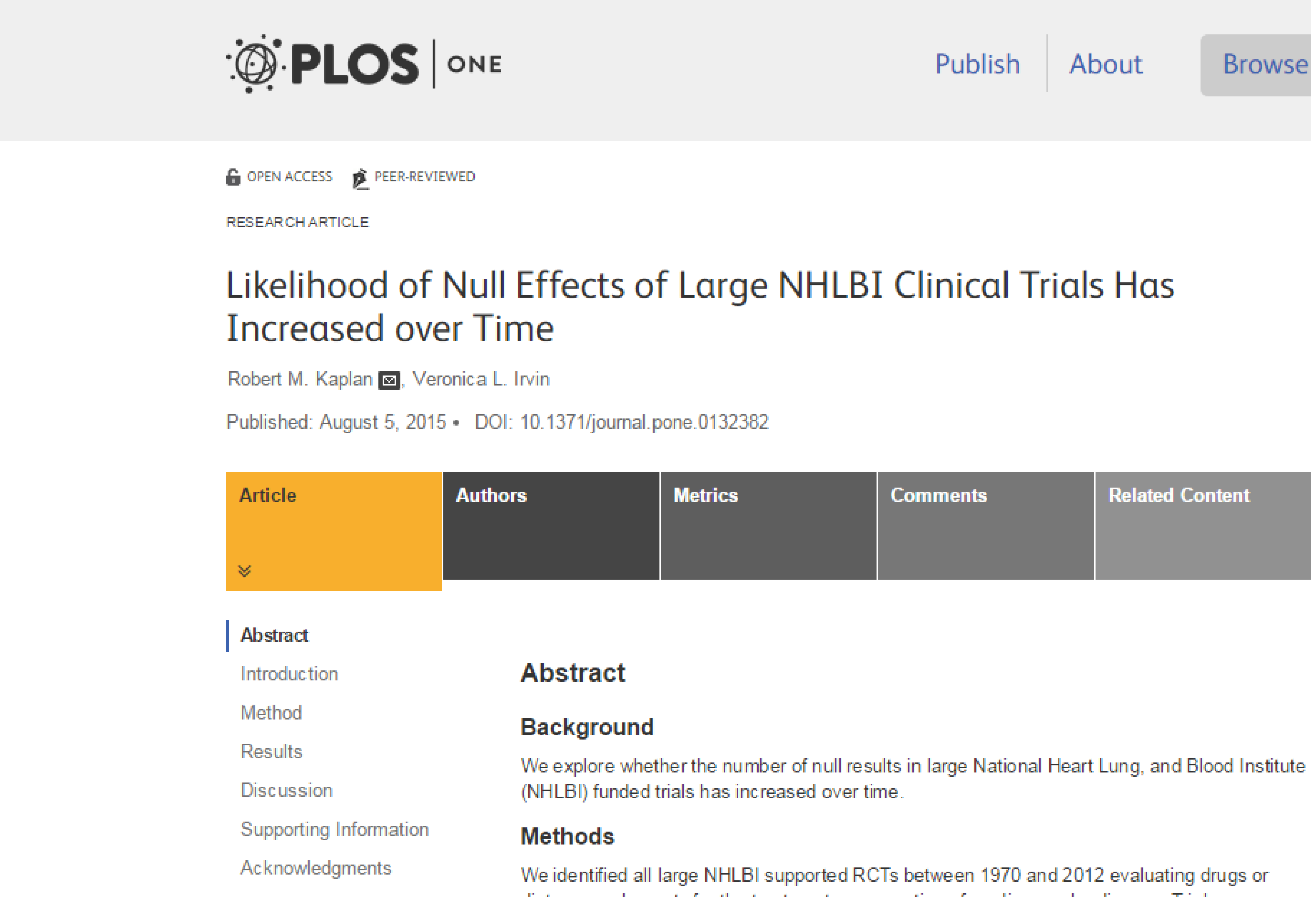
We explore whether the number of null results in large National Heart Lung, and Blood Institute (NHLBI) funded trials has increased over time.

Methods

We identified all large NHLBI supported RCTs between 1970 and 2012 evaluating drugs or

Number of studies finding medical interventions effective before preregistration introduced:
17/30 (55%)

Afterwards:
2/25 (8%)



RESEARCH ARTICLE

Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time

Robert M. Kaplan, Veronica L. Irvin

Published: August 5, 2015 • DOI: 10.1371/journal.pone.0132382

Abstract

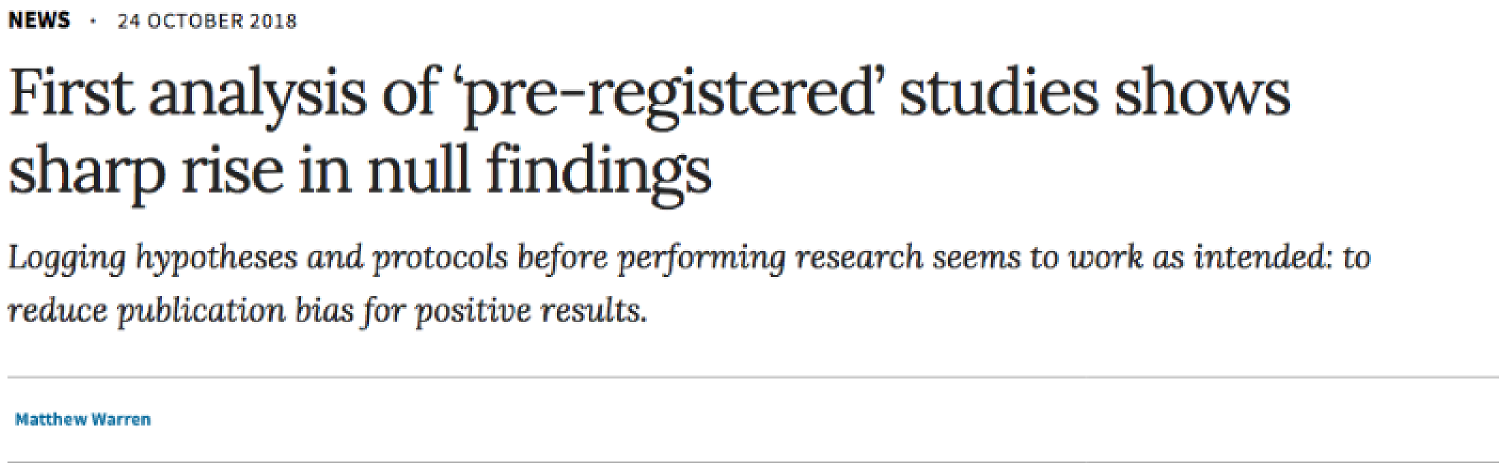
Background

We explore whether the number of null results in large National Heart Lung, and Blood Institute (NHLBI) funded trials has increased over time.

Methods

We identified all large NHLBI supported RCTs between 1970 and 2012 evaluating drugs or

Registered Reports appear to be working as intended

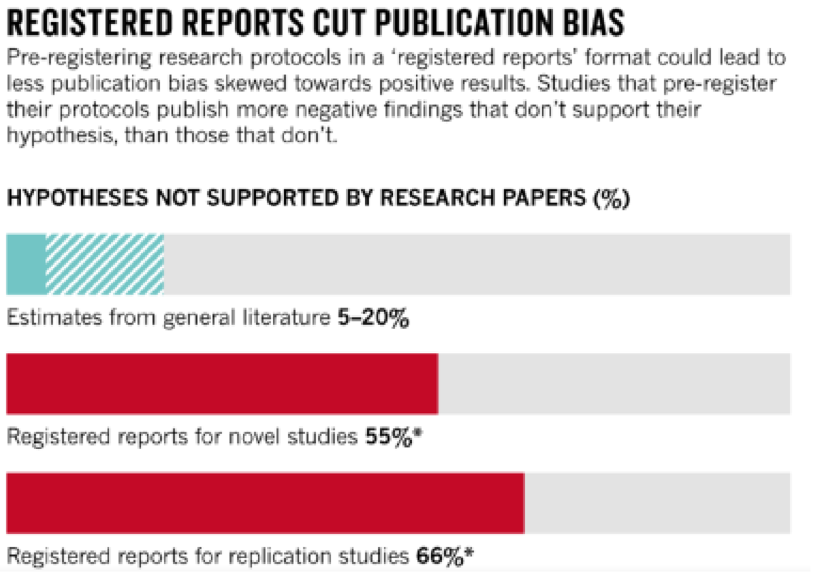


NEWS · 24 OCTOBER 2018

First analysis of ‘pre-registered’ studies shows sharp rise in null findings

Logging hypotheses and protocols before performing research seems to work as intended: to reduce publication bias for positive results.

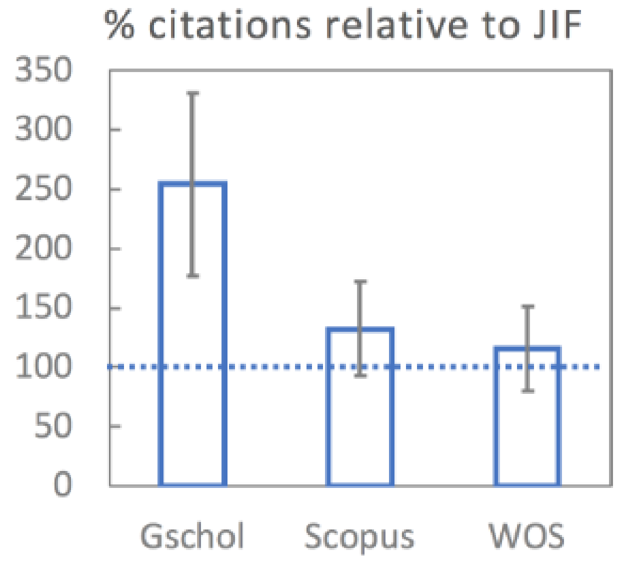
Matthew Warren



REGISTERED REPORTS CUT PUBLICATION BIAS

Pre-registering research protocols in a ‘registered reports’ format could lead to less publication bias skewed towards positive results. Studies that pre-register their protocols publish more negative findings that don’t support their hypothesis, than those that don’t.

HYPOTHESES NOT SUPPORTED BY RESEARCH PAPERS (%)



% citations relative to JIF

Hypotheses at at least three times more likely to be **disconfirmed** in Registered Reports compared with regular articles

Well cited -- at or above respective journal impact factor

https://tinyurl.com/RR-citations

PCI :

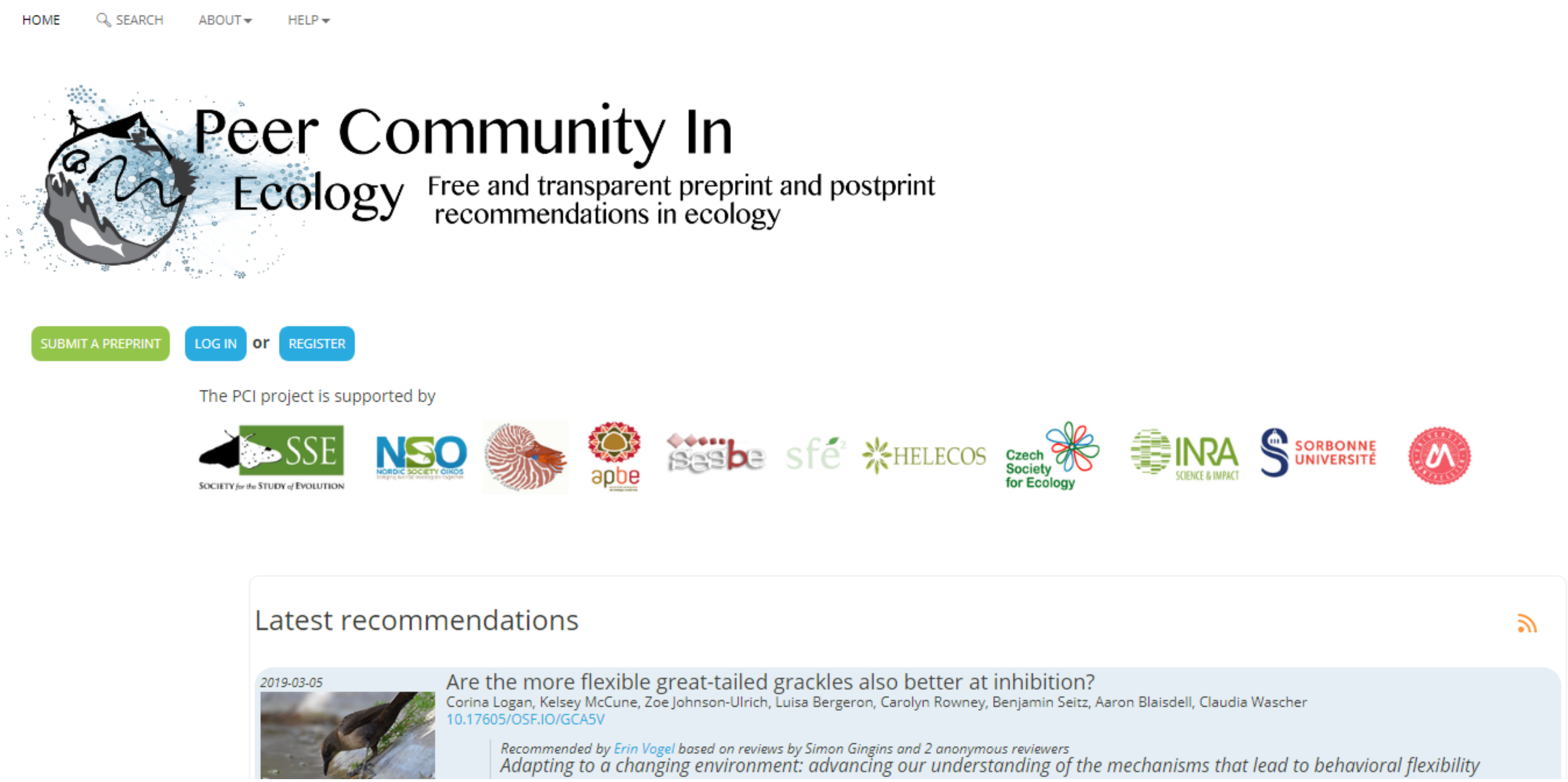
Submit pre-print

“Recommenders” send to review

Edited until accepted

Journals then consider paper

> 100 submissions, > 30 now accepted



Free and transparent preprint and postprint recommendations in ecology

The PCI project is supported by

Latest recommendations

2019-03-05

Are the more flexible great-tailed grackles also better at inhibition?

Corina Logan, Kelsey McCune, Zoe Johnson-Ulrich, Luisa Bergeron, Carolyn Rowney, Benjamin Seitz, Aaron Blaisdell, Claudia Wascher

10.17605/OSF.IO/GCA5V

Recommended by Erin Vogel based on reviews by Simon Gingins and 2 anonymous reviewers

Adapting to a changing environment: advancing our understanding of the mechanisms that lead to behavioral flexibility

PLAN:

PCI Registered Reports

Submissions dealt with to Stage I or II by PCI editorial board

Journals can consider accepting paper
Back-up: an overlay journal will definitely accept.