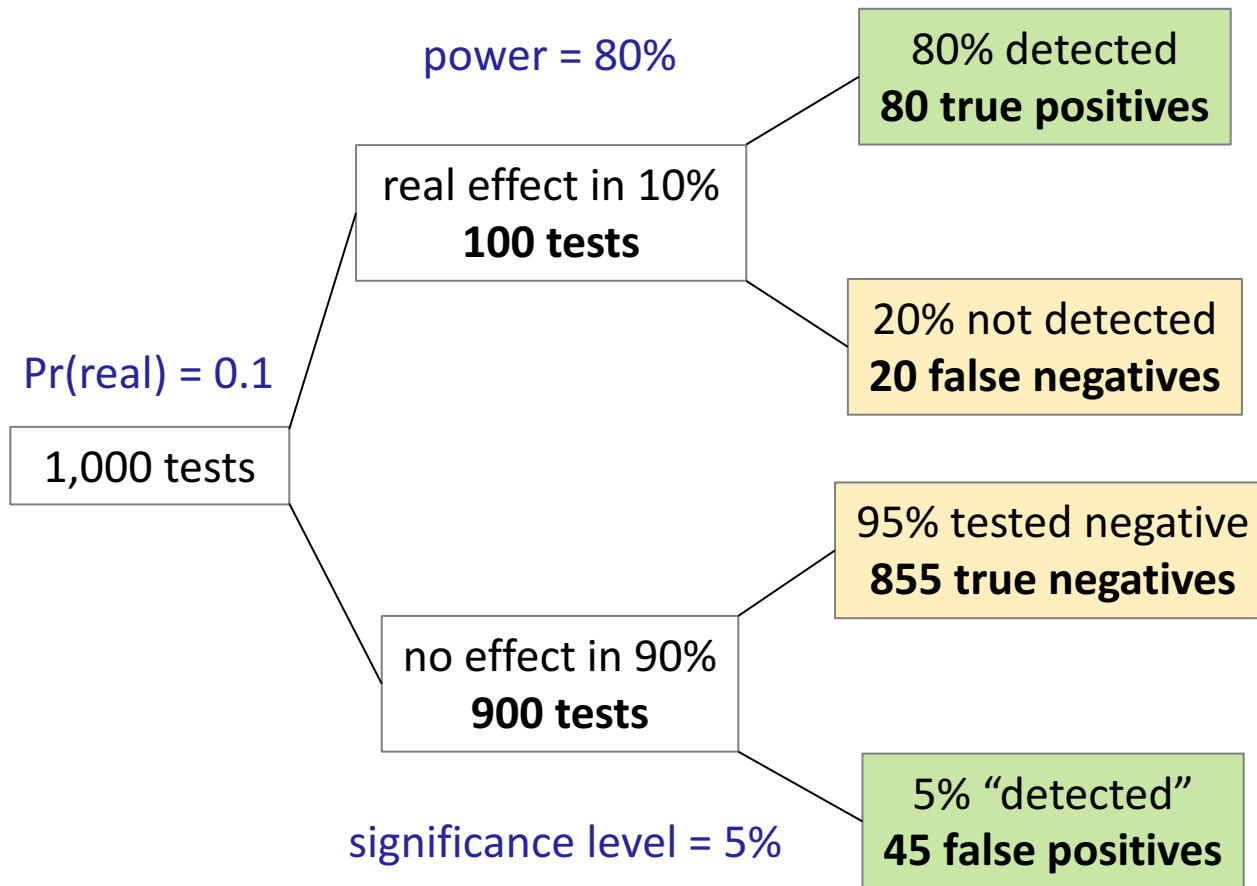


Lies
damn lies
statistics

Marek Gierliński
Division of Computational Biology



False positive rate

$$FPR = \frac{\text{false positives}}{\text{no effect}}$$

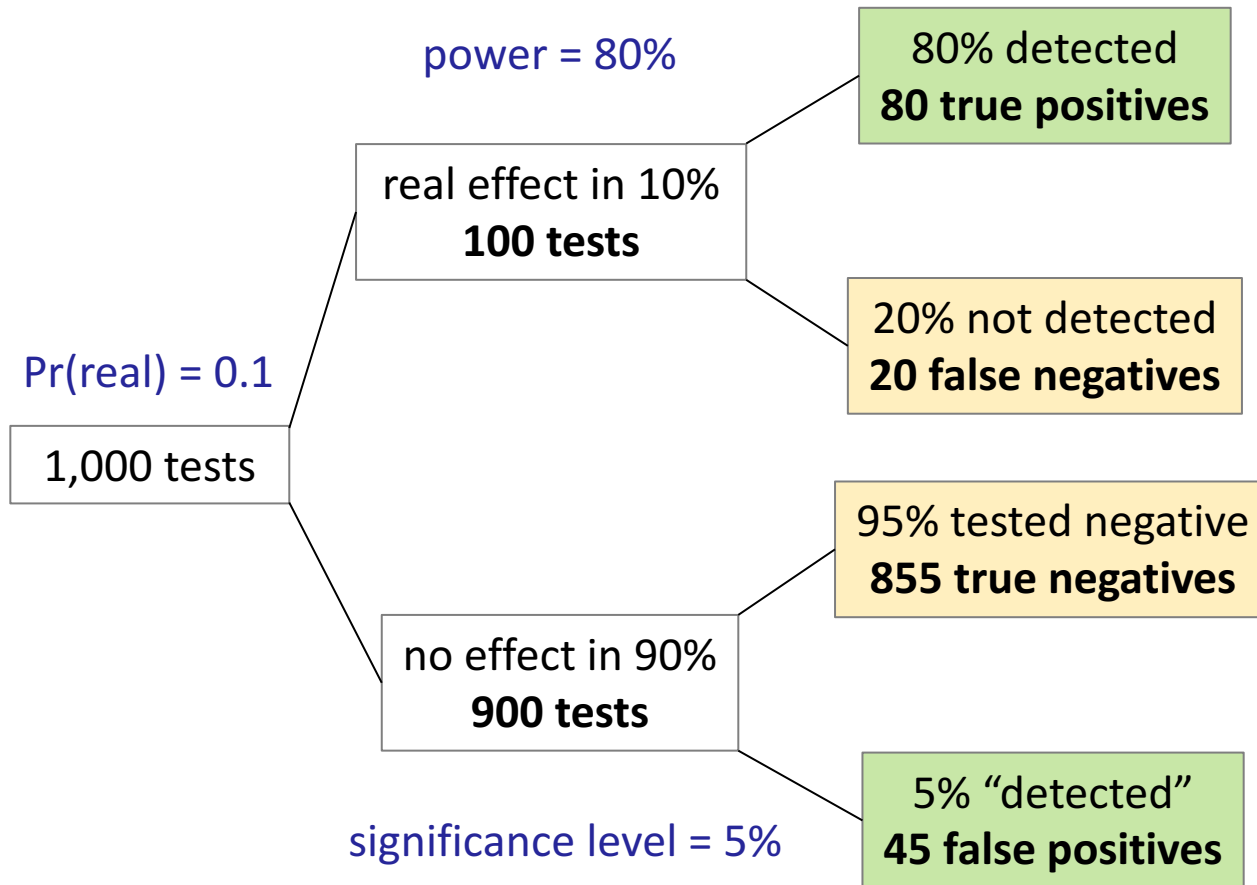
$$FPR = \frac{45}{900} = 0.05$$

False discovery rate

$$FDR = \frac{\text{false positives}}{\text{discoveries}}$$

$$FDR = \frac{45}{45 + 80} = 0.36$$

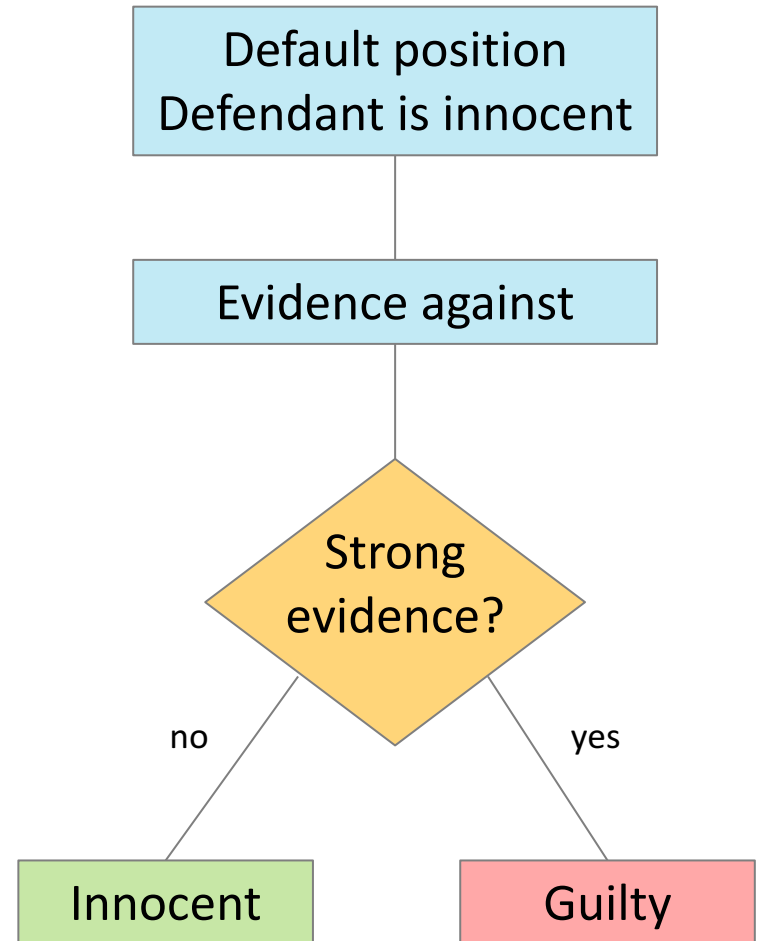
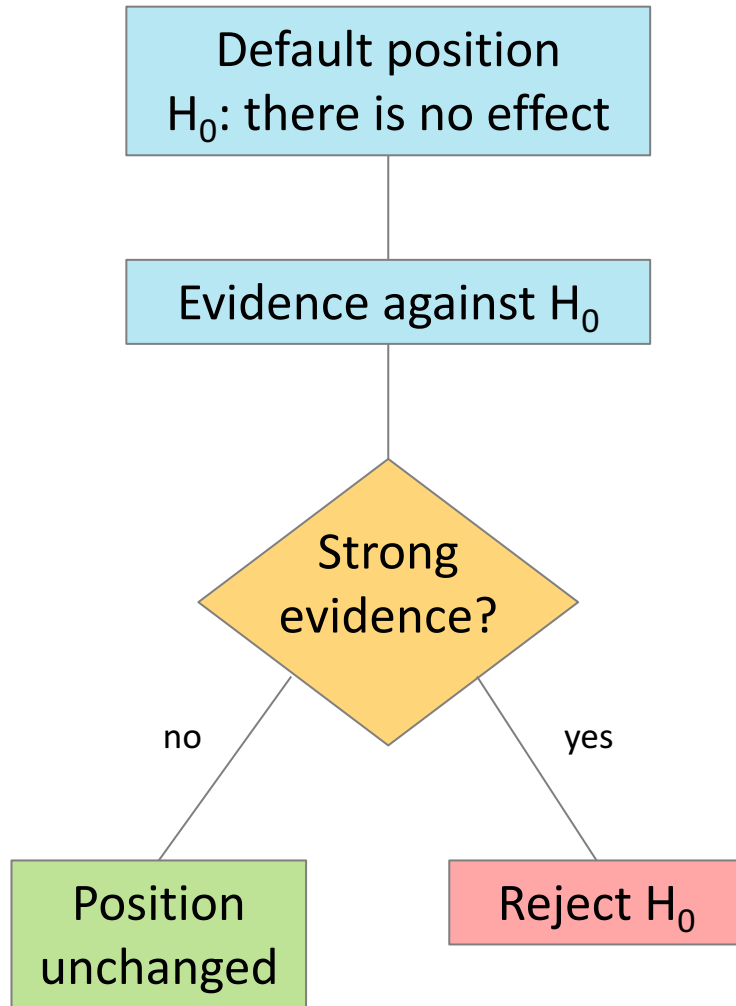
Colquhoun D., 2014, "An investigation of the false discovery rate and the misinterpretation of *p*-values", *R. Soc. open sci.* **1**: 140216.



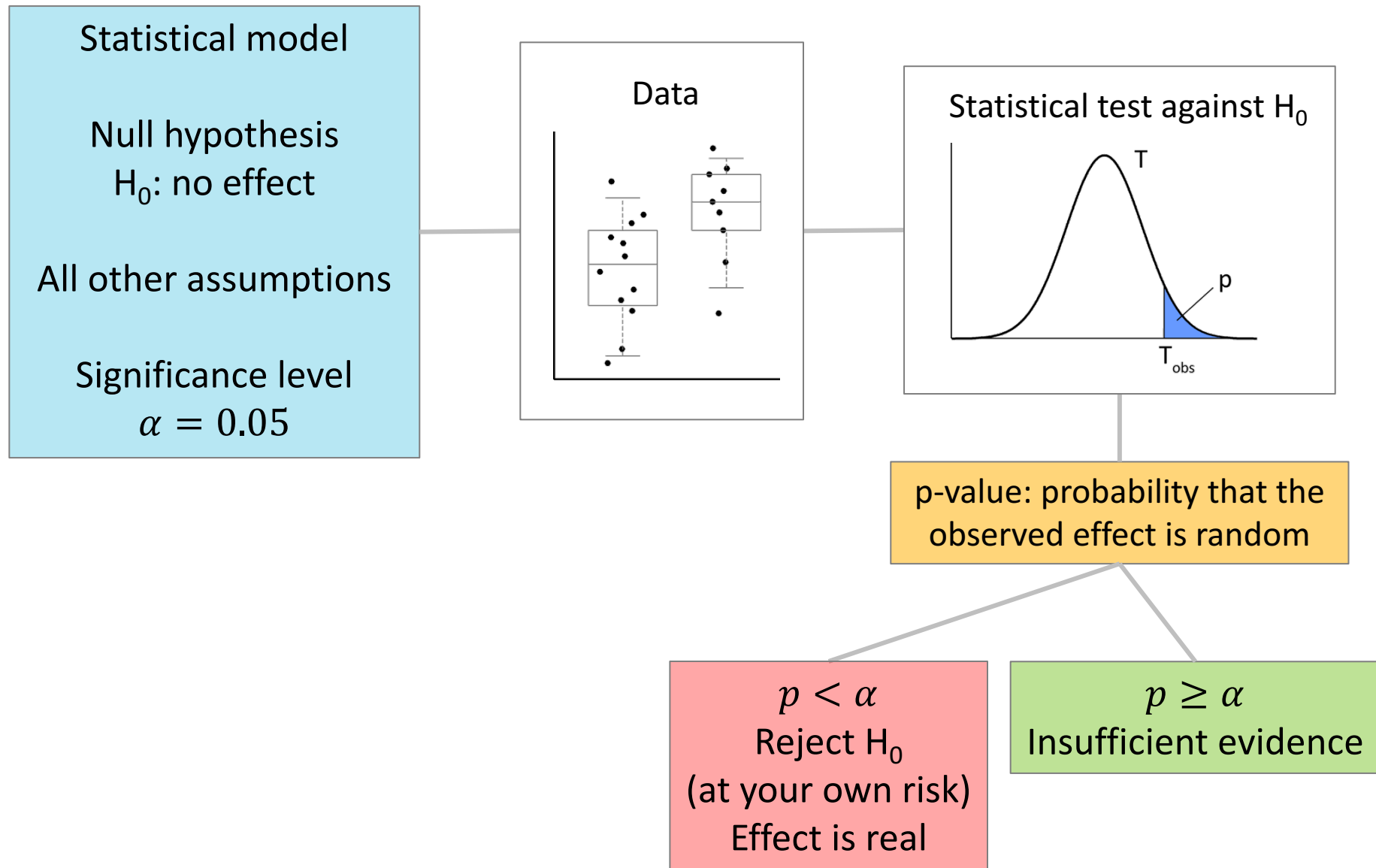
If you publish a $p < 0.05$ result, you have a 36% chance of making a fool of yourself

Colquhoun D., 2014, "An investigation of the false discovery rate and the misinterpretation of p -values", *R. Soc. open sci.* **1**: 140216.

Null hypothesis



Statistical testing



p-value:

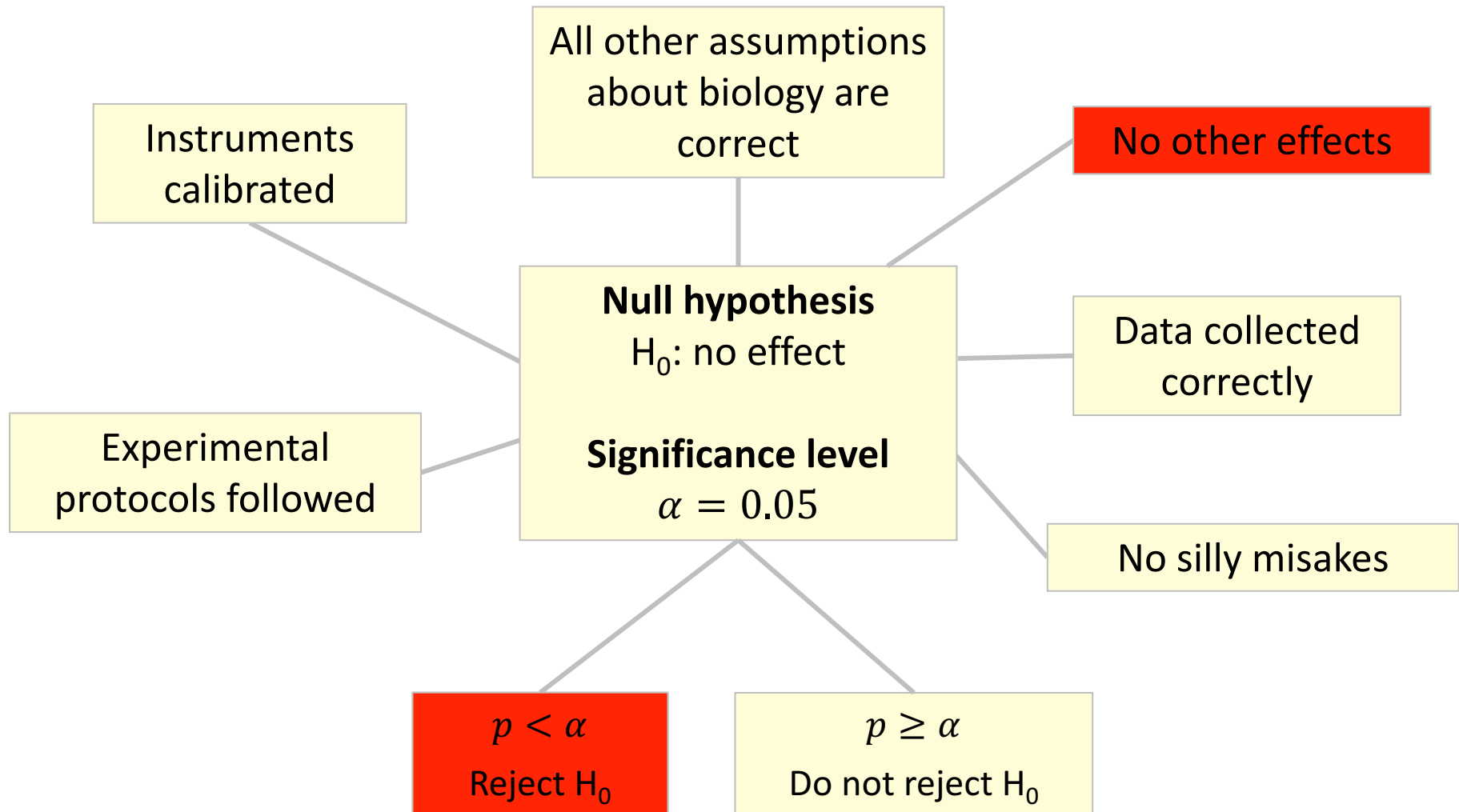
Given that H_0 is true, the probability of observed, or more extreme, data

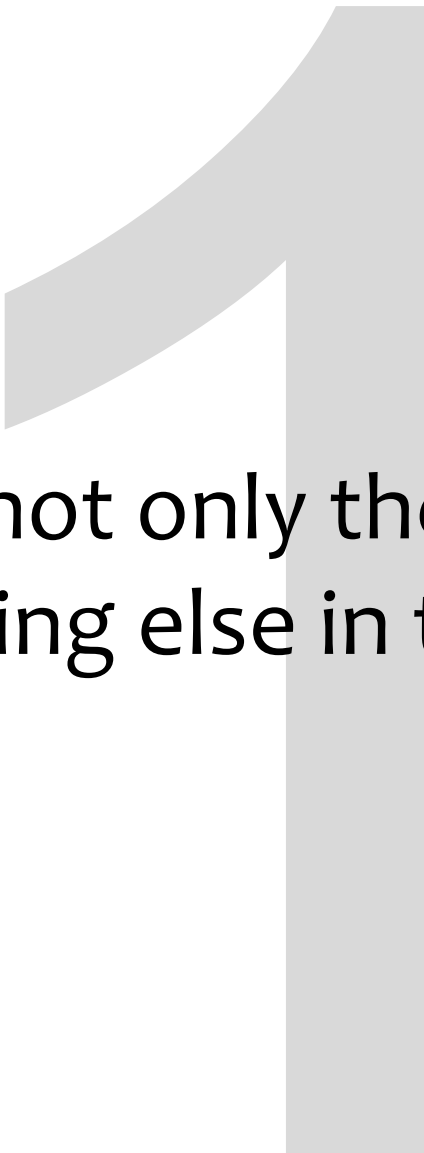
It is **not** the probability that H_0 is true

P-value is the degree to which the data are embarrassed by the null hypothesis

Nicholas Maxwell

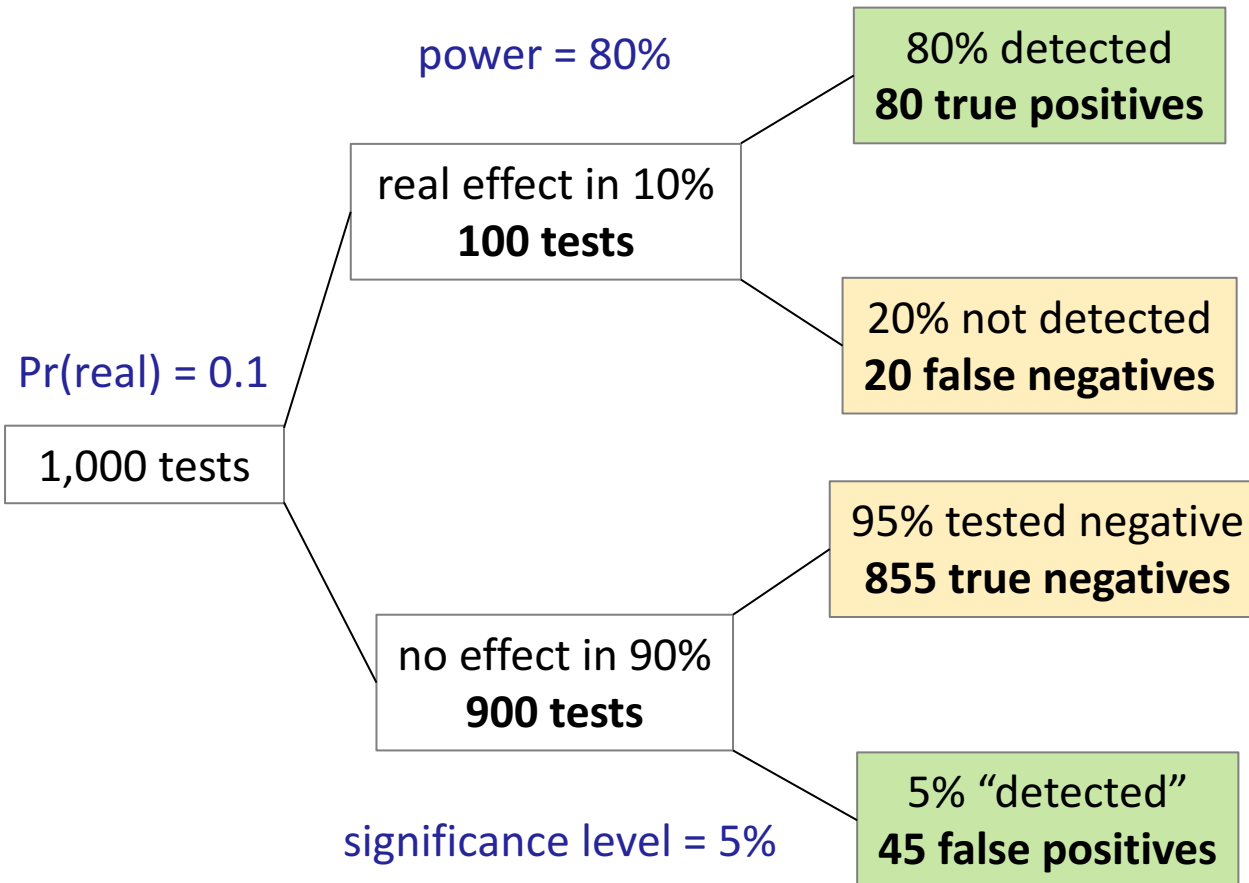
“All other assumptions”





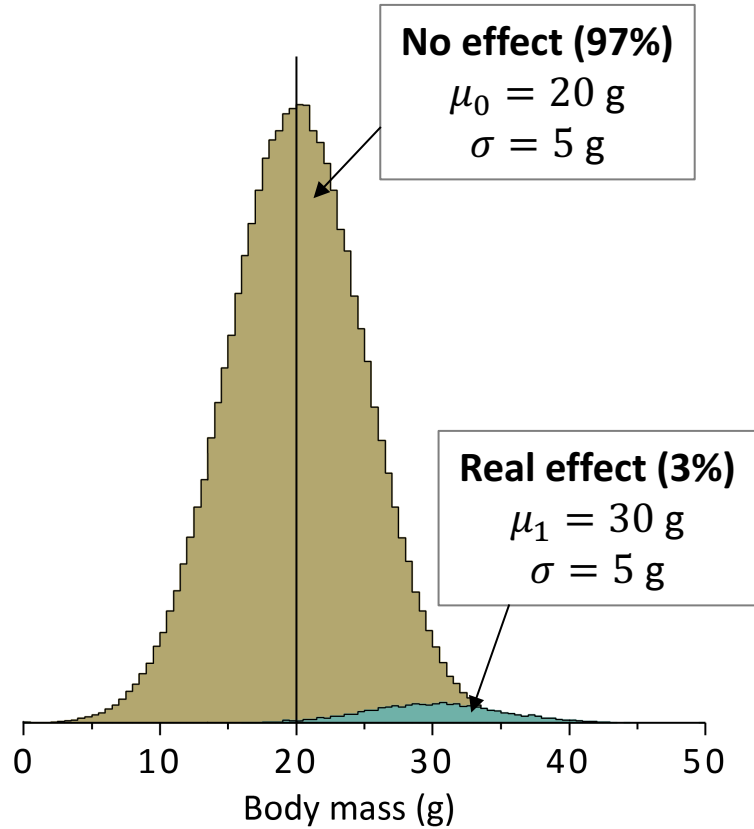
p-values test not only the null hypothesis,
but everything else in the experiment

Why large false discovery rate?



$$FDR = \frac{45}{45 + 80} = 0.36$$

Simulated population of mice



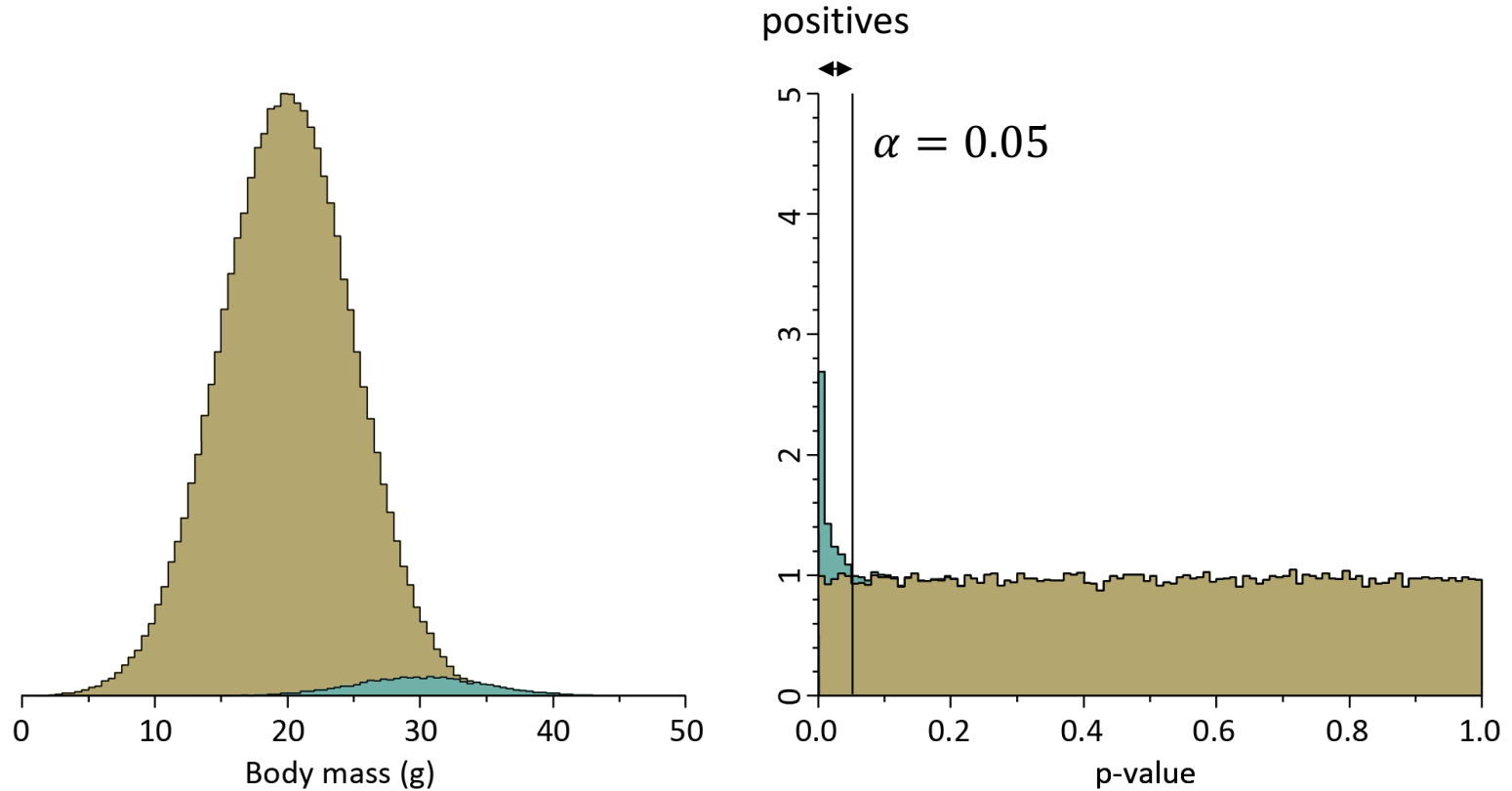
Null hypothesis $H_0: \mu = 20$ g

one-sample t-test

Power analysis

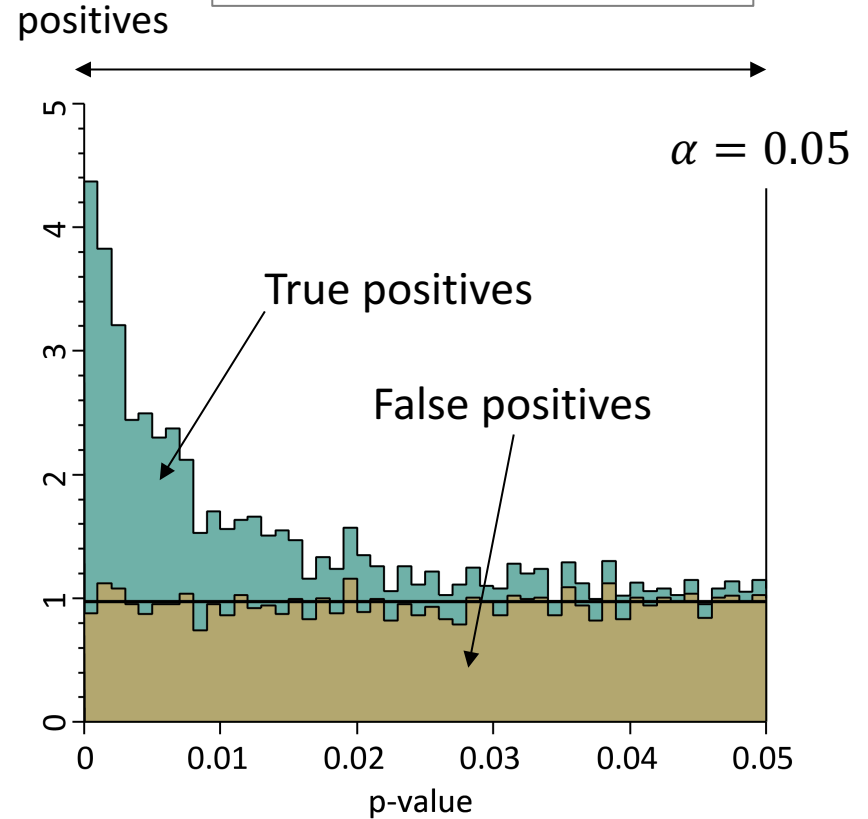
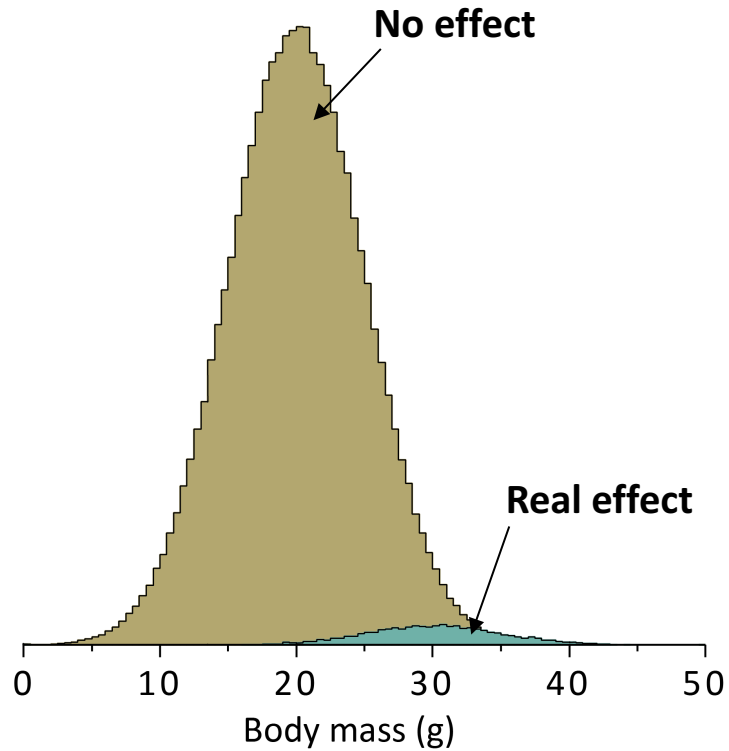
effect size	$\Delta m = 10$ g
power	$\mathcal{P} = 0.9$
significance level	$\alpha = 0.05$
sample size	$n = 5$

Gedankenexperiment: distribution of p-values



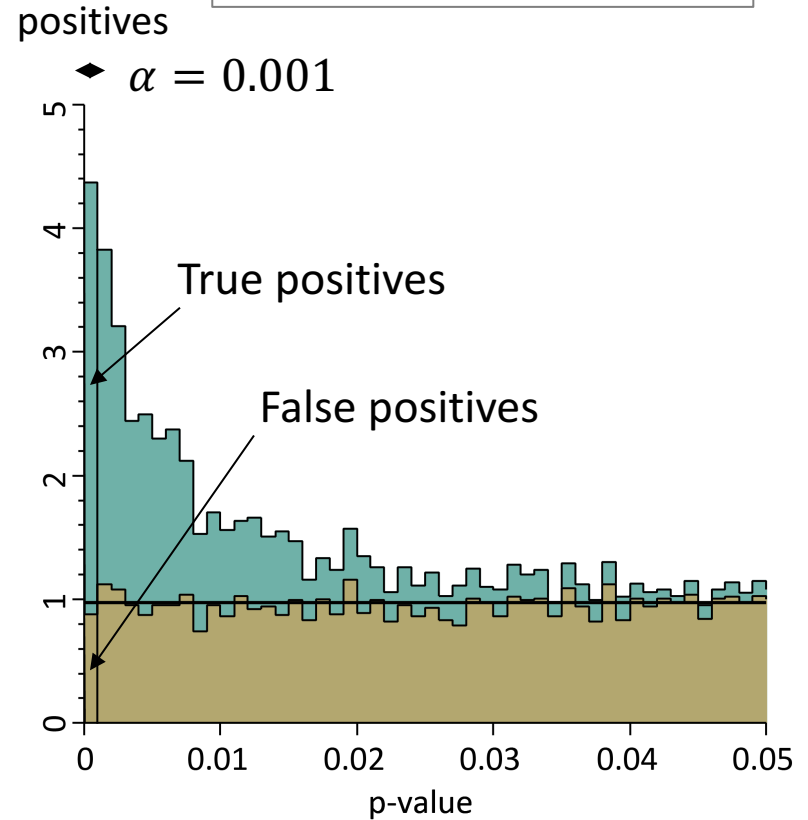
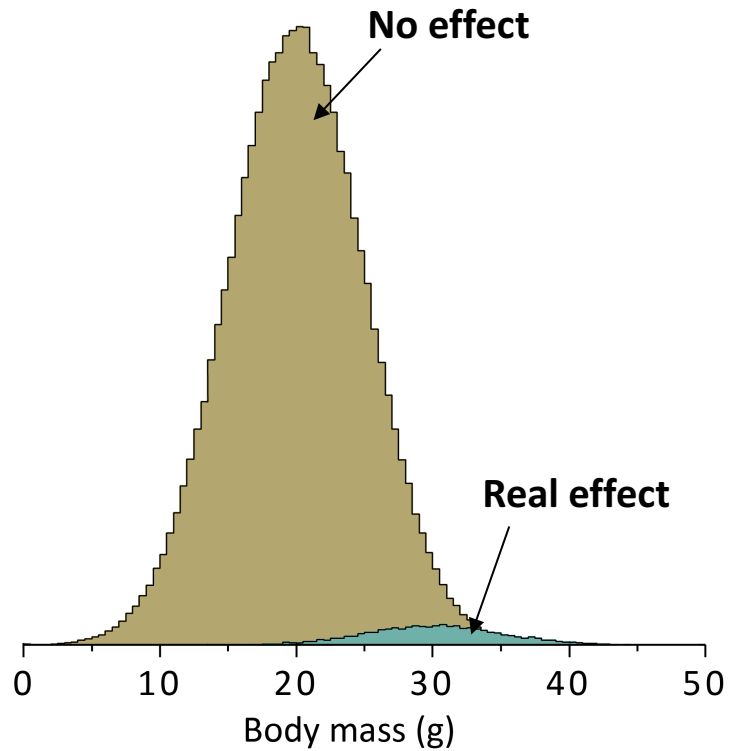
Gedankenexperiment: “significant” p-values

$$FDR = \frac{FP}{FP + TP} \approx 0.63$$



Small α doesn't help

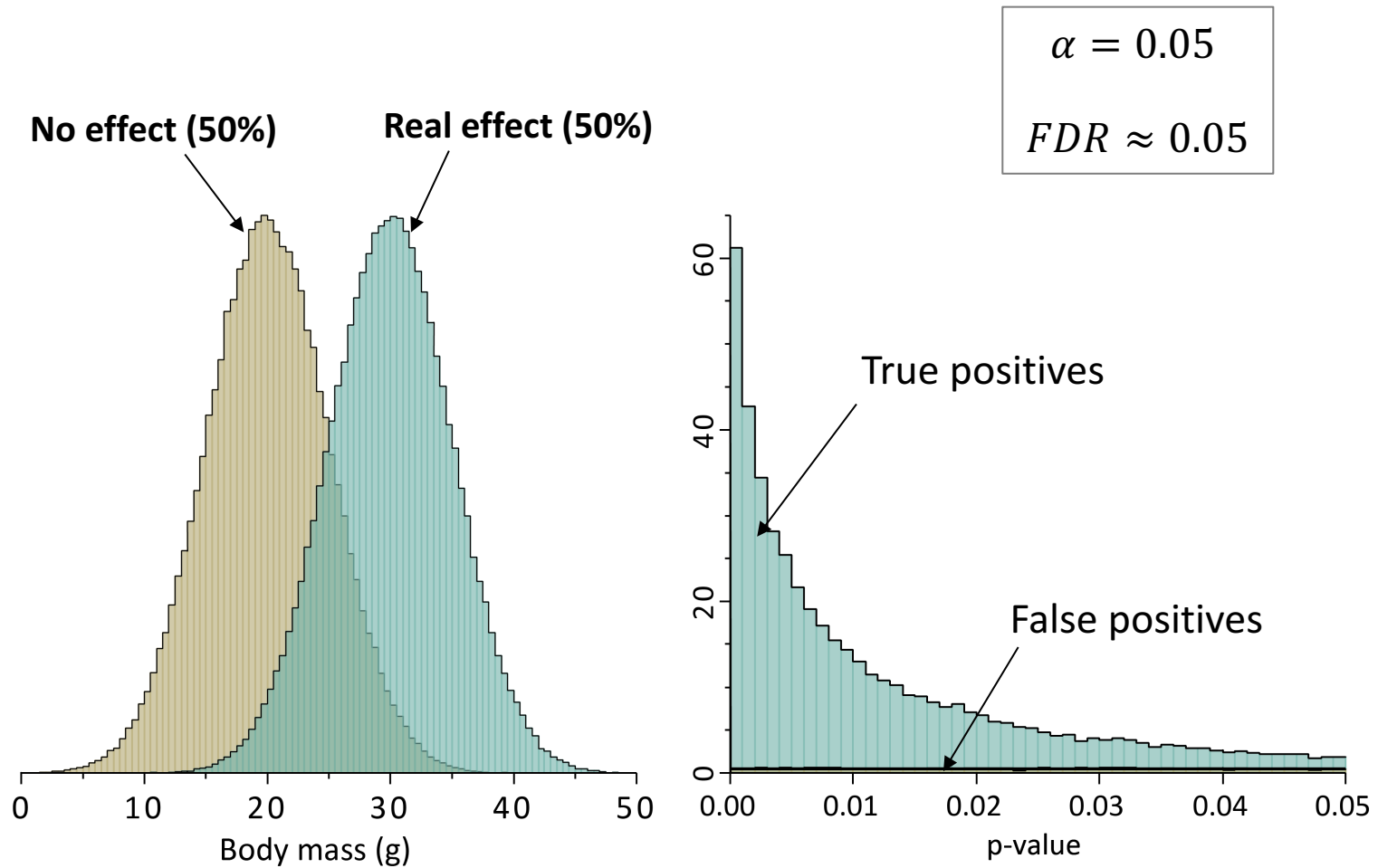
$$FDR = \frac{FP}{FP + TP} \approx 0.20$$





The chance of making a fool of yourself
is much larger than $\alpha = 0.05$

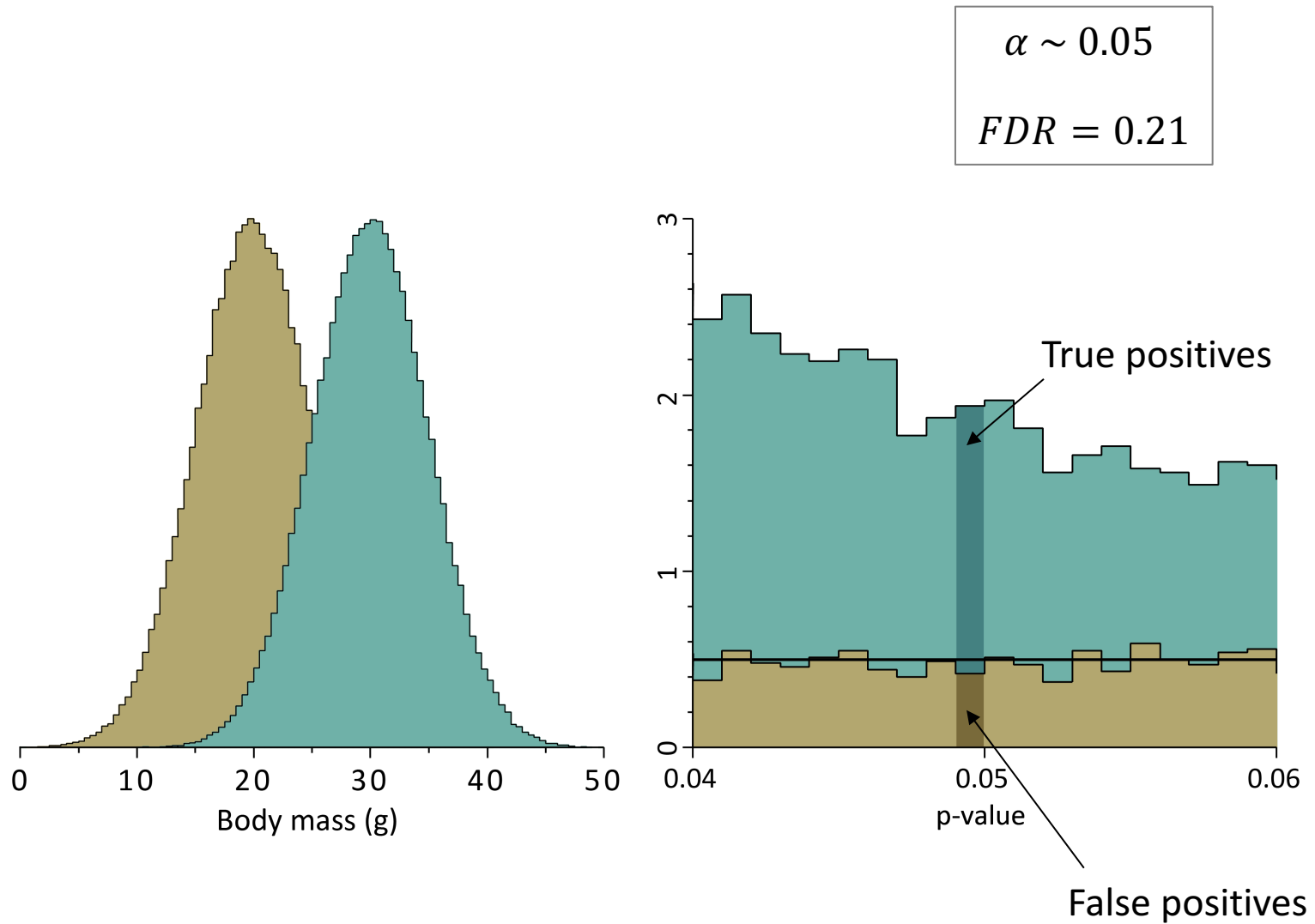
FDR depends on the probability of real effect



A large, light gray number '3' is centered on the page, serving as a background for the text.

When the effect is rare,
you are screwed

What does a p-value ~ 0.05 really mean?



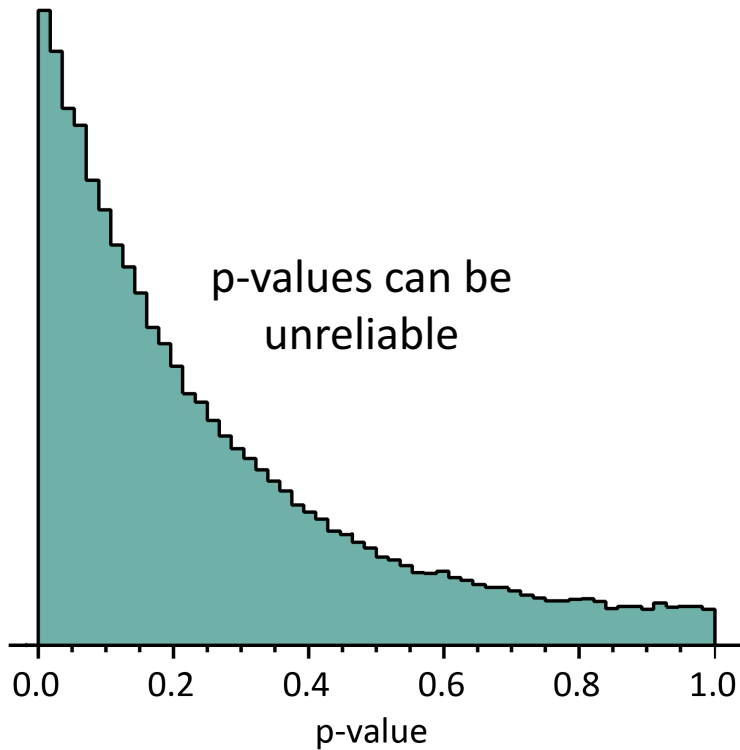


When you get a $p \sim 0.05$,
you are screwed

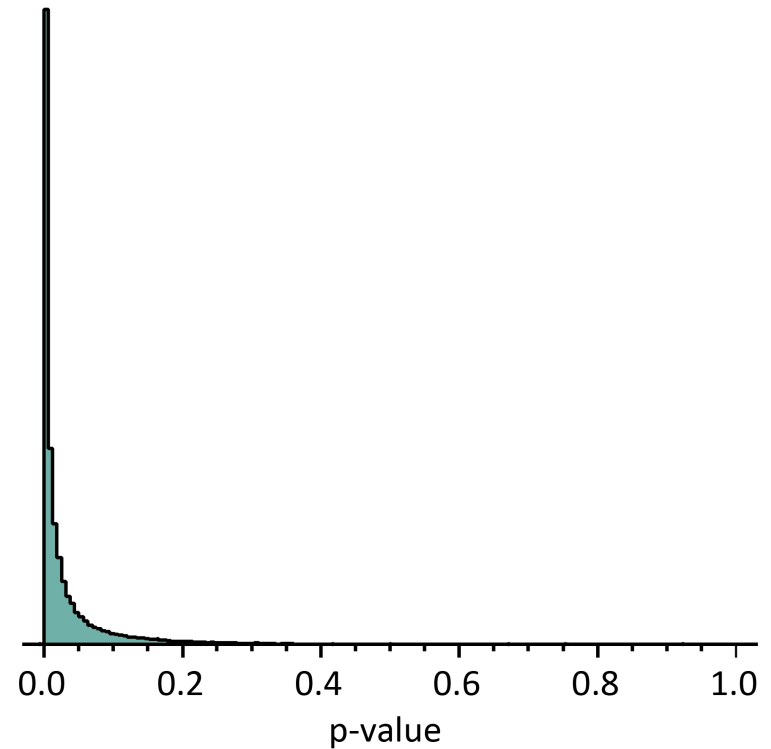
Gedankenexperiment: reliability of p-values

Normal population, 100% real effect
One-sample t-test

Sample size = 3, power = 0.18

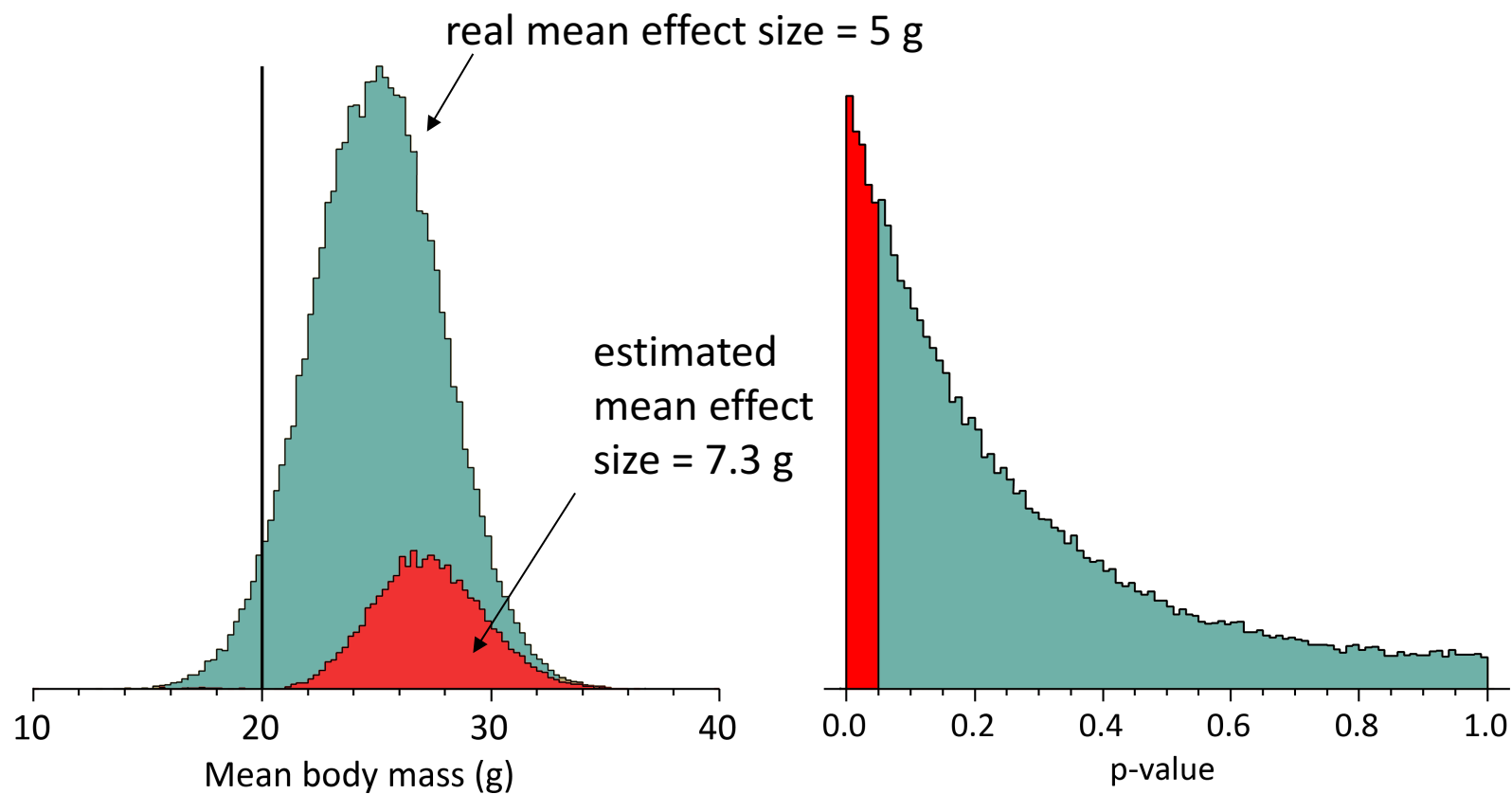


Sample size = 10, power = 0.80



Underpowered studies lead to
unreliable p-values

Inflation of the effect size



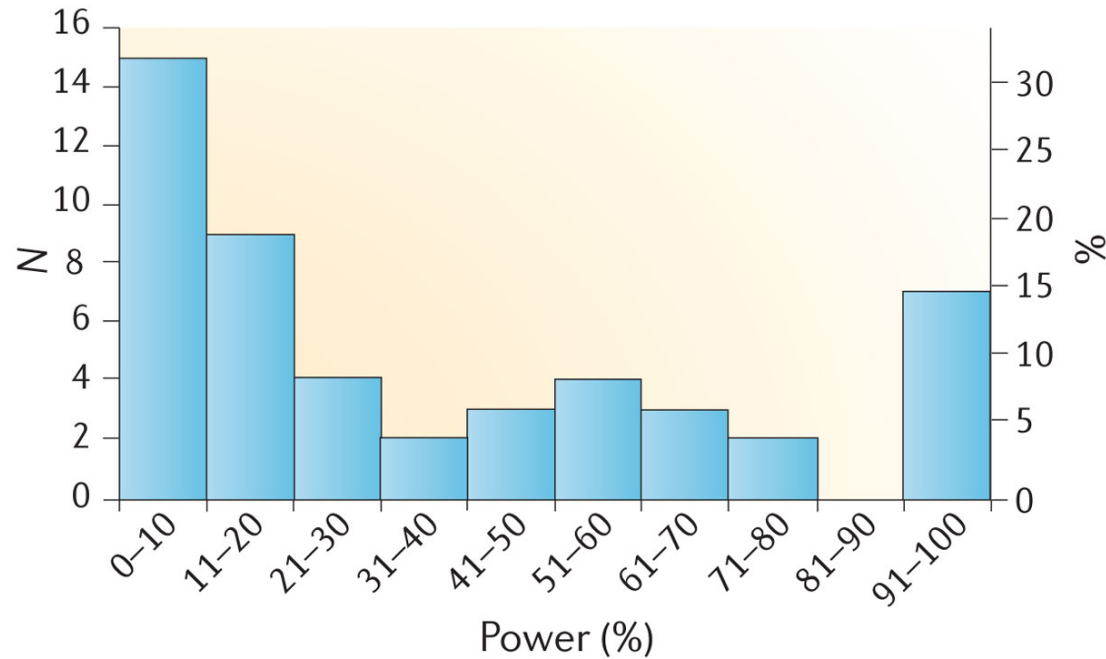
Underpowered studies lead to
unreliable p-values

Underpowered studies lead to
overestimated effect size



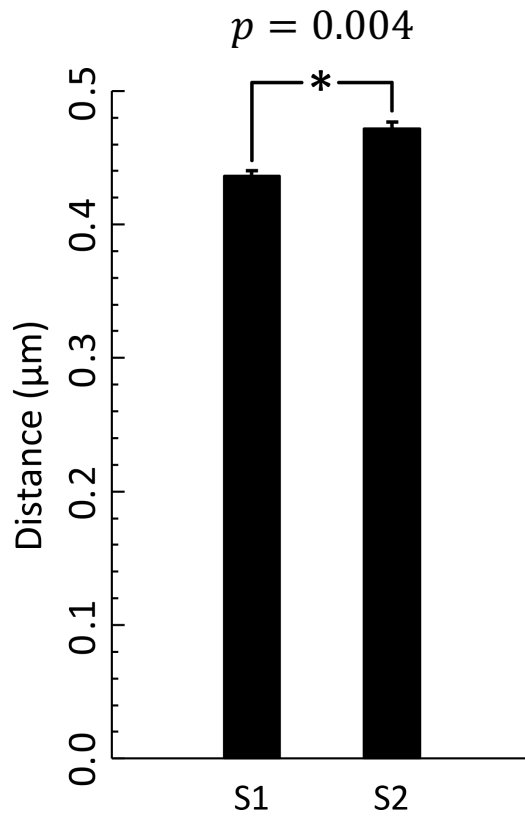
When your experiment is underpowered,
you are screwed

Neuroscience: most studies underpowered

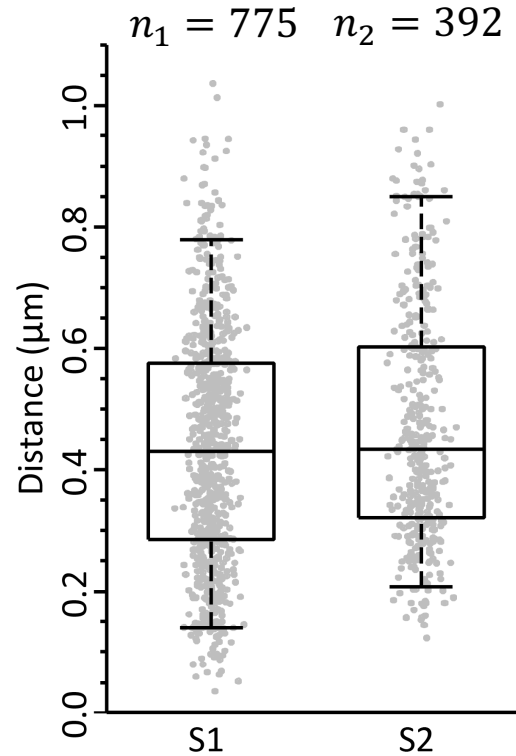
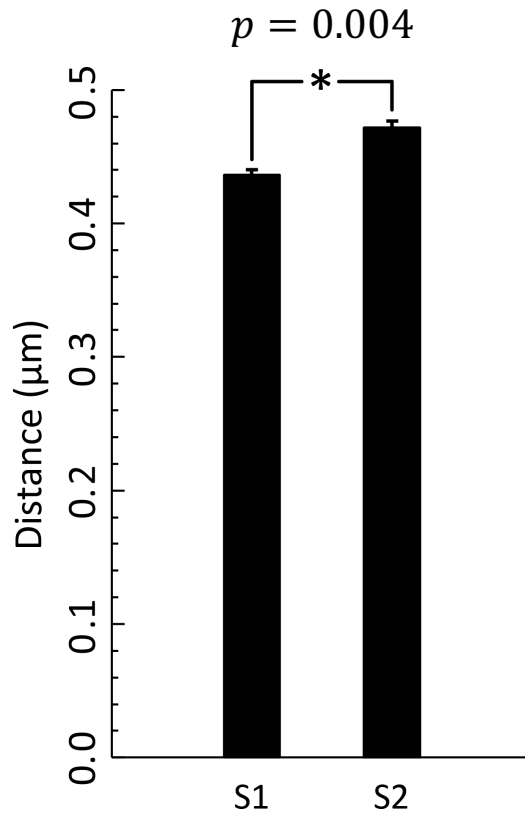


Button et al. (2013) "Power failure: why small sample size undermines the reliability of neuroscience", *Nature Reviews Neuroscience* **14**, 365-376

The effect size



The effect size



With sample size large enough everything is “significant”

Effect size is more important

Looking at whole data is even more important

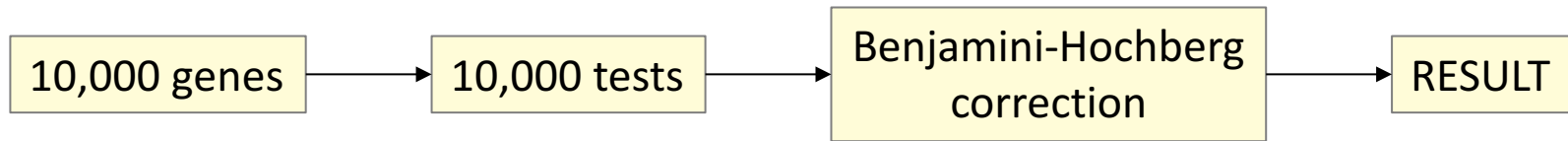


When you have lots of replicates,
p-values are useless

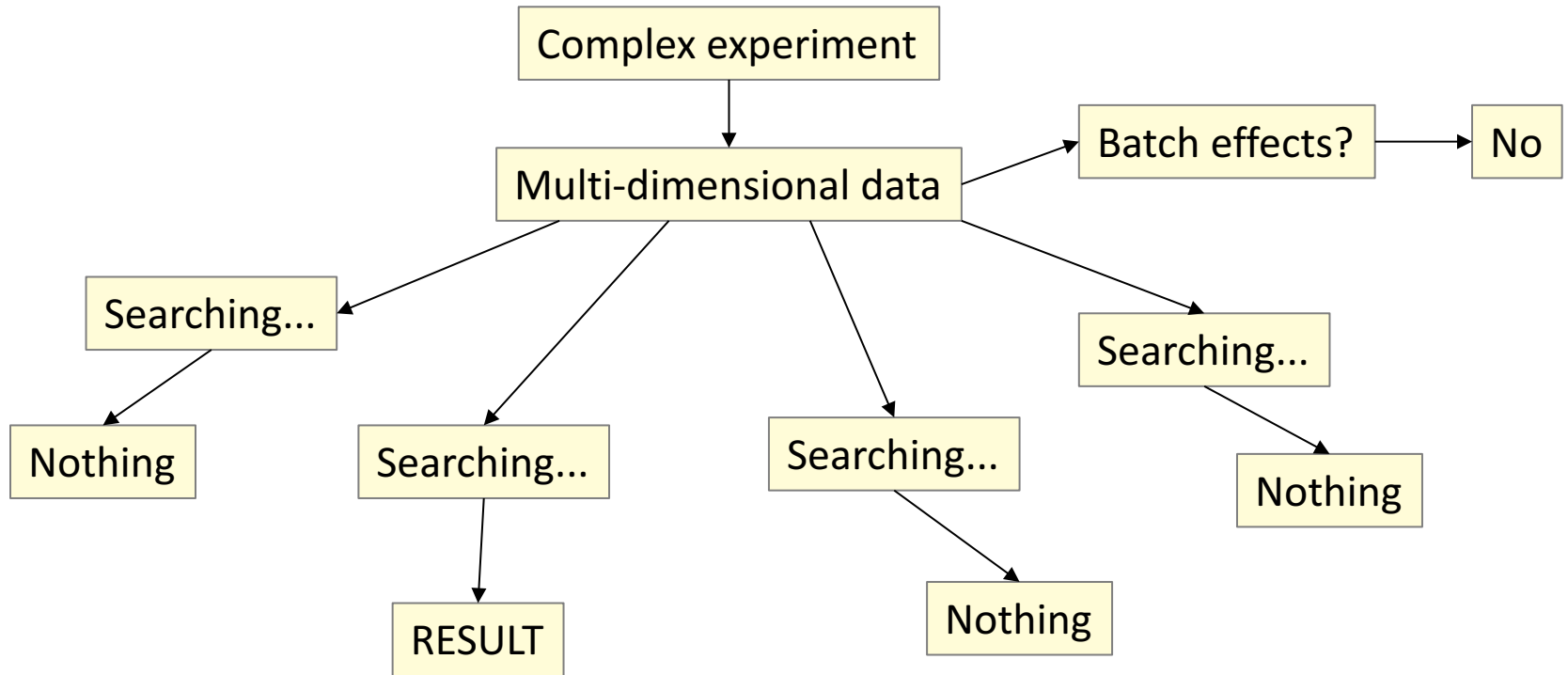
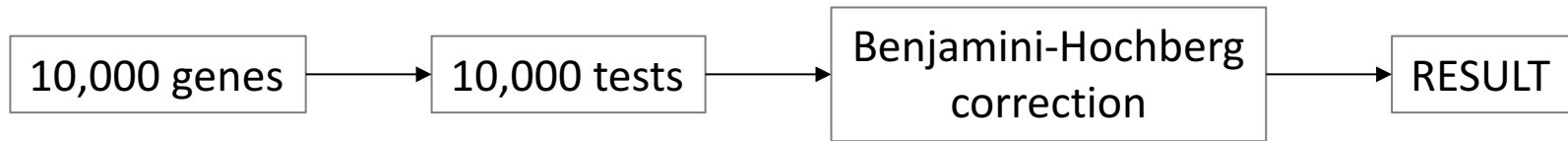


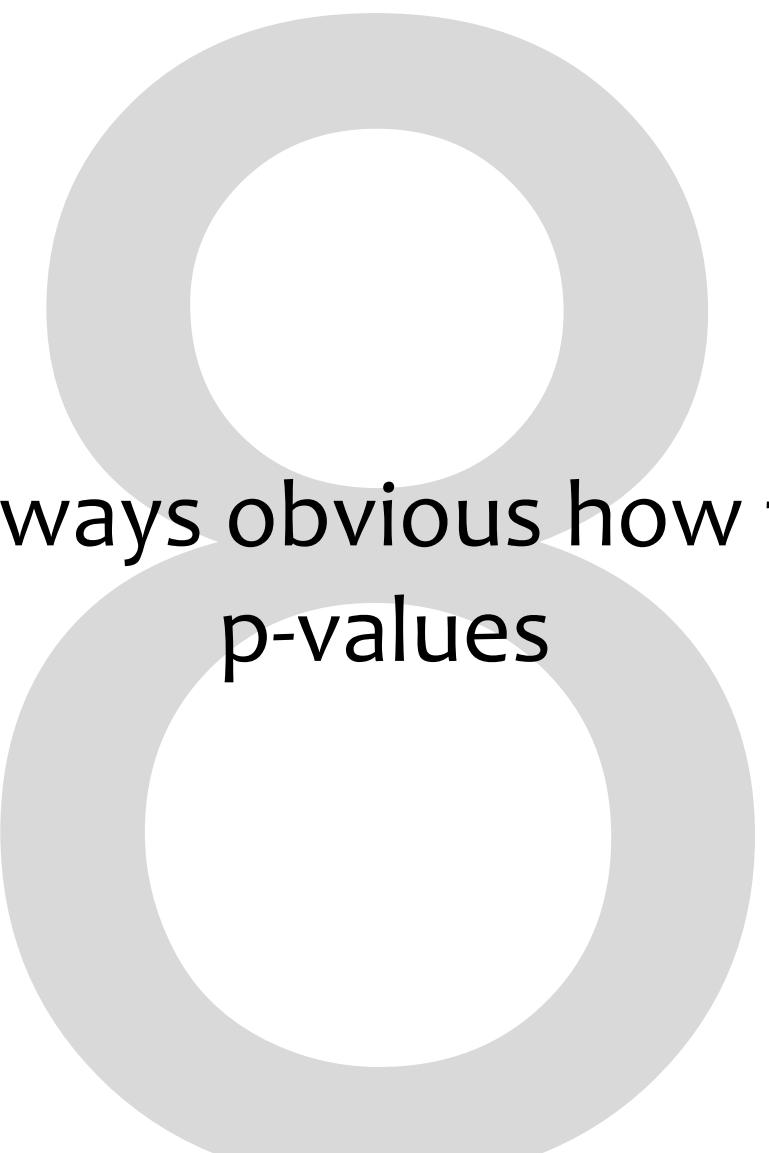
Statistical significance does not imply
biological relevance

Multiple test corrections can be tricky



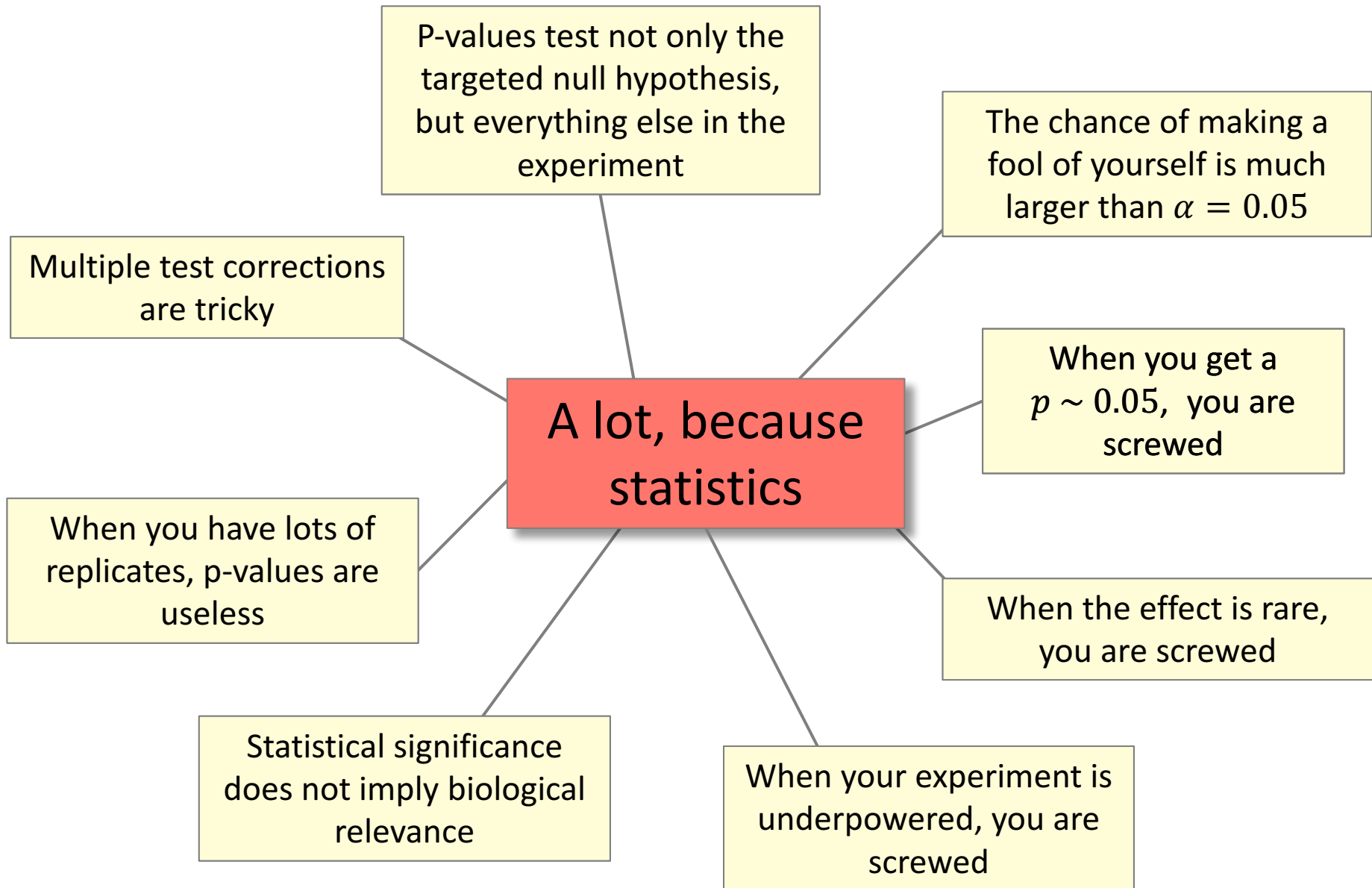
Multiple test corrections can be tricky





It is not always obvious how to correct
p-values

What's wrong with p-values?



***P*-Values: Misunderstood and Misused**

*Bertie Vidgen and Taha Yasseri**



MINI REVIEW
published: 04 March 2016
doi: 10.3389/fphy.2016.00006

The fickle *P* value generates irreproducible results

Lewis G Halsey, Douglas Curran-Everett, Sarah L Vowler & Gordon B Drummond

NATURE METHODS | VOL.12 NO.3 | MARCH 2015 | 179

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P.A. Ioannidis

Null hypothesis significance testing is a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring.

Paul Meehl, 1967, *Philosophy of Science*, 34, 103-115

The plain fact is that 70 years ago Ronald Fisher gave scientists a mathematical machine for turning baloney into breakthroughs, and flukes into funding. It is time to pull the plug.

Robert Matthews, *Sunday Telegraph*, 13 September 1998.

The widespread use of “statistical significance” as a license for making a claim of a scientific finding leads to considerable distortion of the scientific process.

ASA statement on statistical significance and p-values (2016)

Today's Random Medical News

from the New England
Journal of
Panic-Inducing
Gobbledygook

JIM BORGMAN
CINCINNATI INQUIRER



By Jim Borgman, first published by the Cincinnati Inquirer 27 April 1997

What's wrong with us?

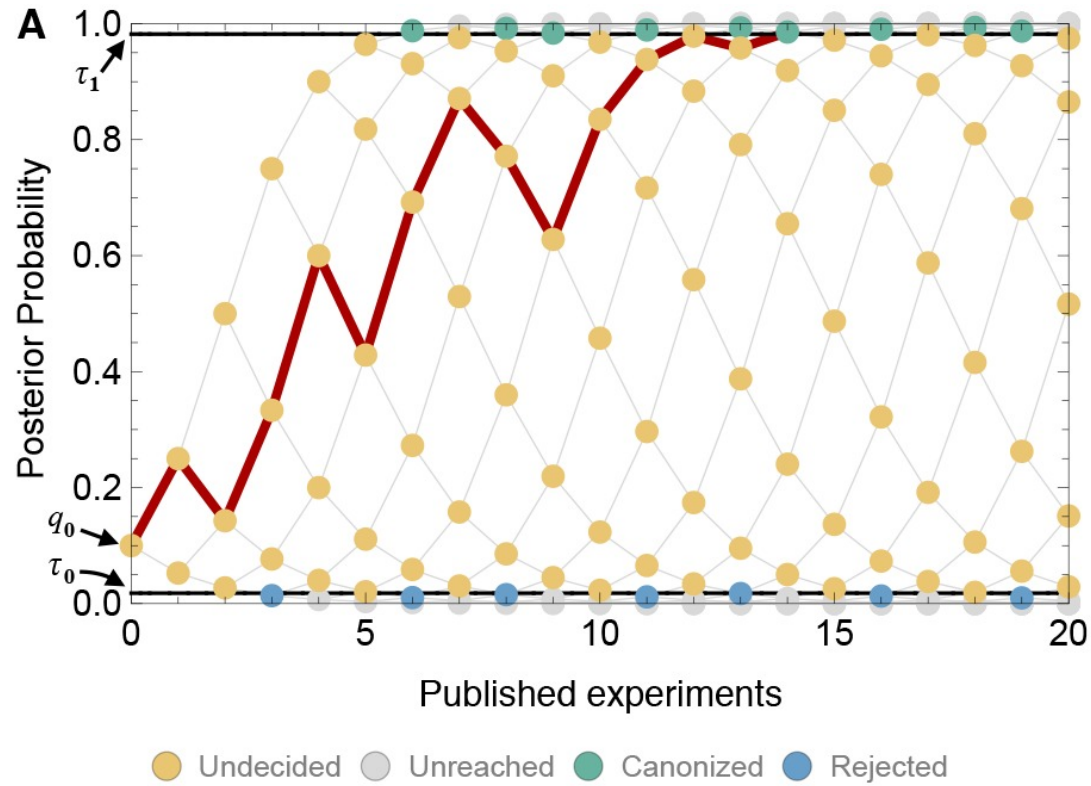
“There is some evidence that [...] research which yields nonsignificant results is not published. Such research being unknown to other investigators may be repeated independently until eventually by chance a significant result occurs [...] The possibility thus arises that the literature [...] consists in substantial part of false conclusions [...].”

PUBLICATION DECISIONS AND THEIR POSSIBLE EFFECTS ON
INFERENCES DRAWN FROM TESTS OF SIGNIFICANCE
—OR VICE VERSA*

THEODORE D. STERLING
University of Cincinnati

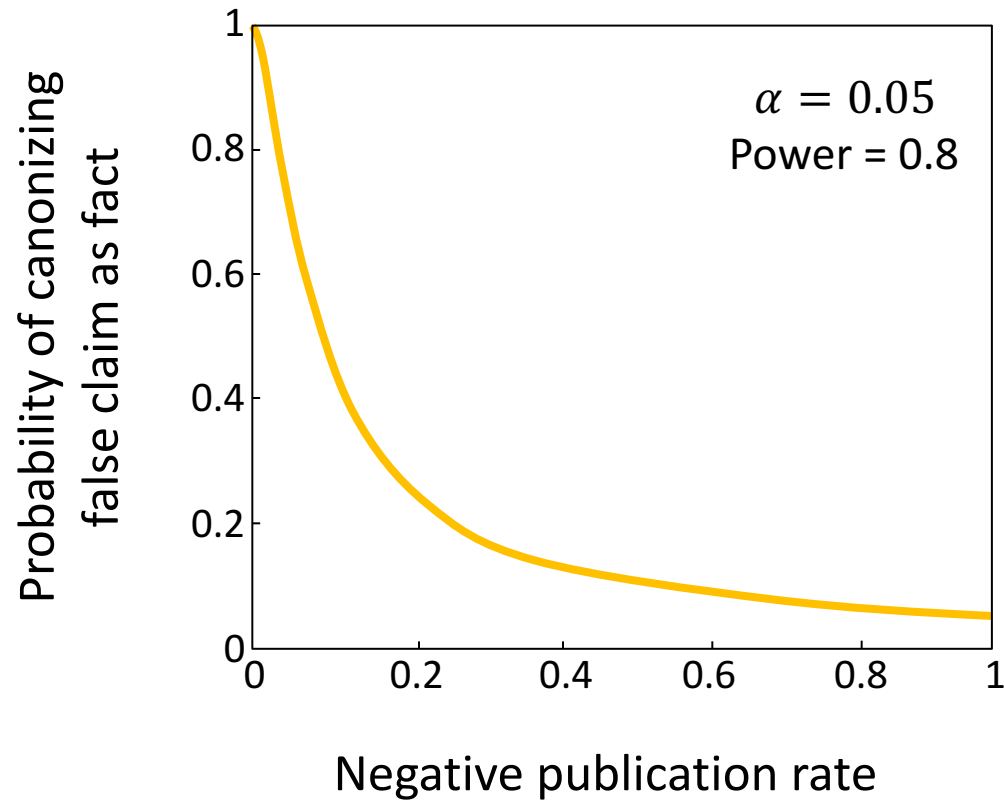
Journal of the American Statistical Association,
Vol. 54, No. 285 (Mar., 1959), pp. 30-34

Canonization of false facts



Nissen S.B., et al., "Research: Publication bias and the canonization of false facts", eLife 2016;5:e21451

Canonization of false facts



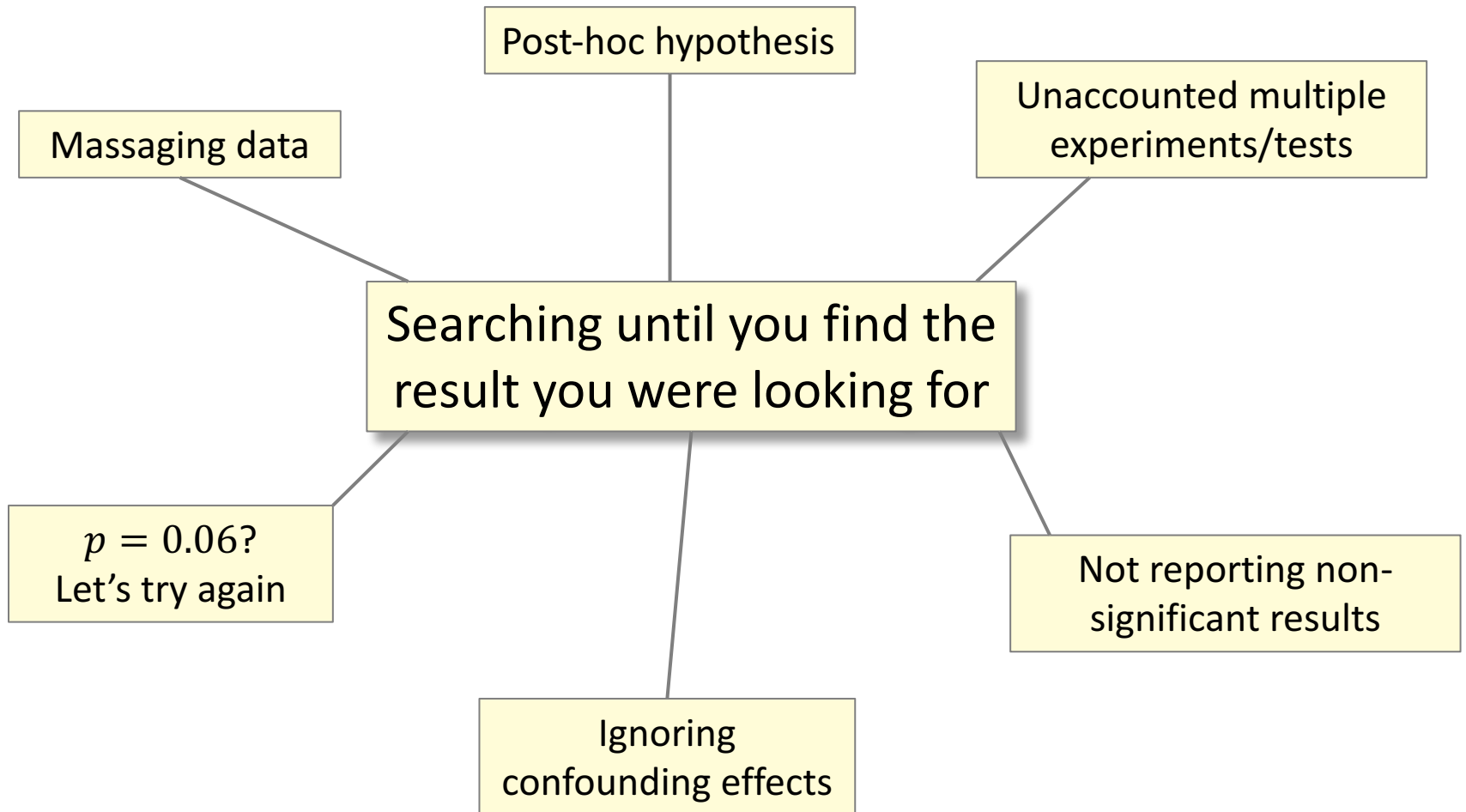
Nissen S.B., et al., "Research: Publication bias and the canonization of false facts", eLife 2016;5:e21451

If you don't publish negative results,
science is screwed

but...

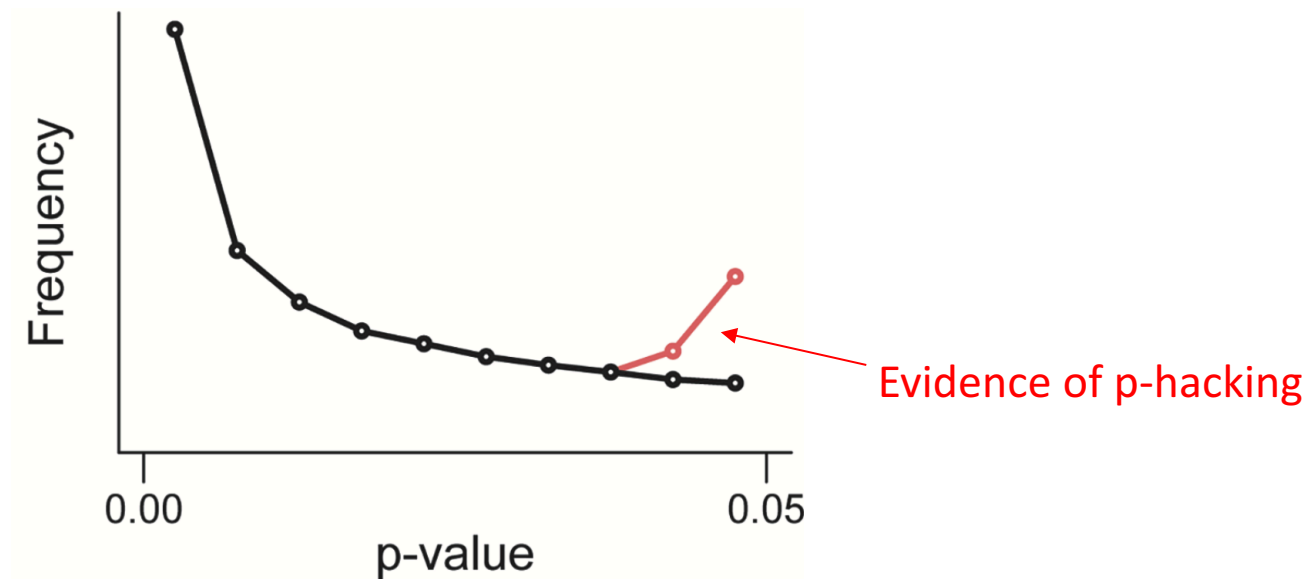
there is a thin line between “negative
result” and “no result”

Data dredging, p-hacking



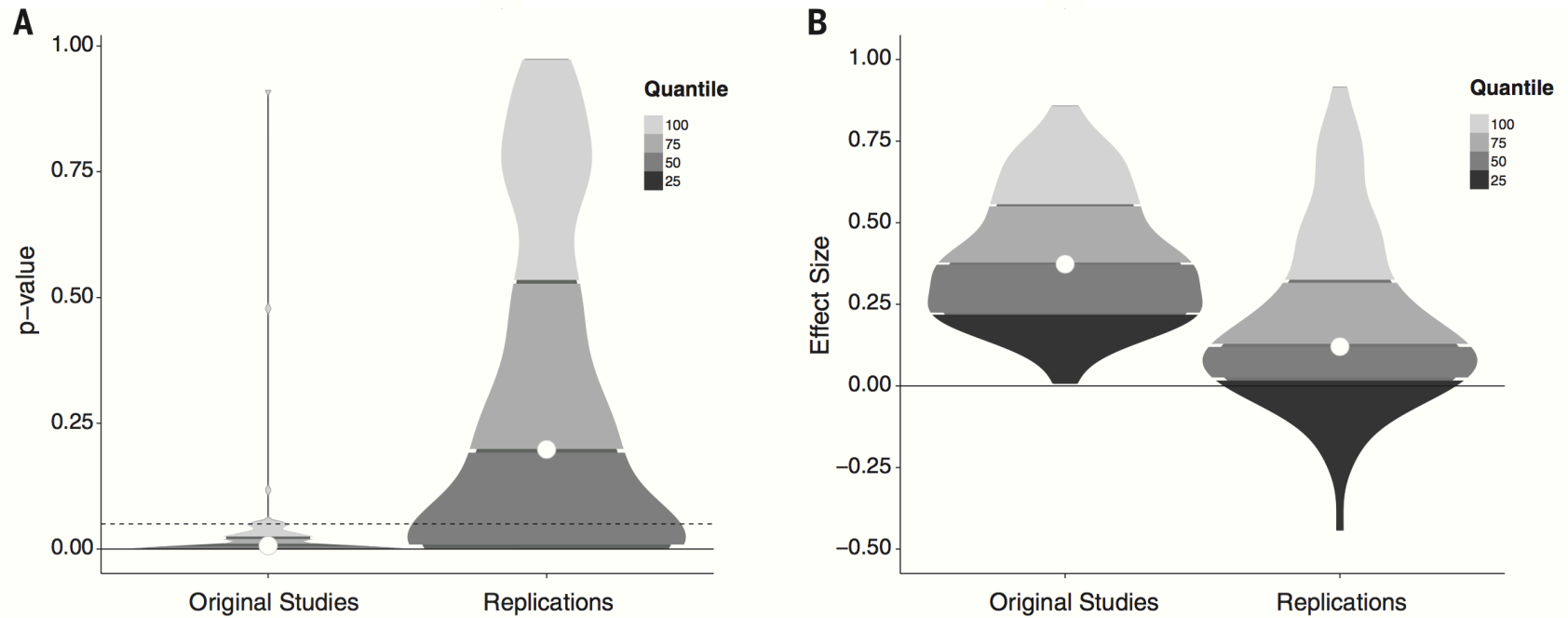
Evidence of p-hacking

Distribution of p-values reported in publications



Head M.L., et al. "The Extent and Consequences of P-Hacking in Science", PLoS Biol 13(3)

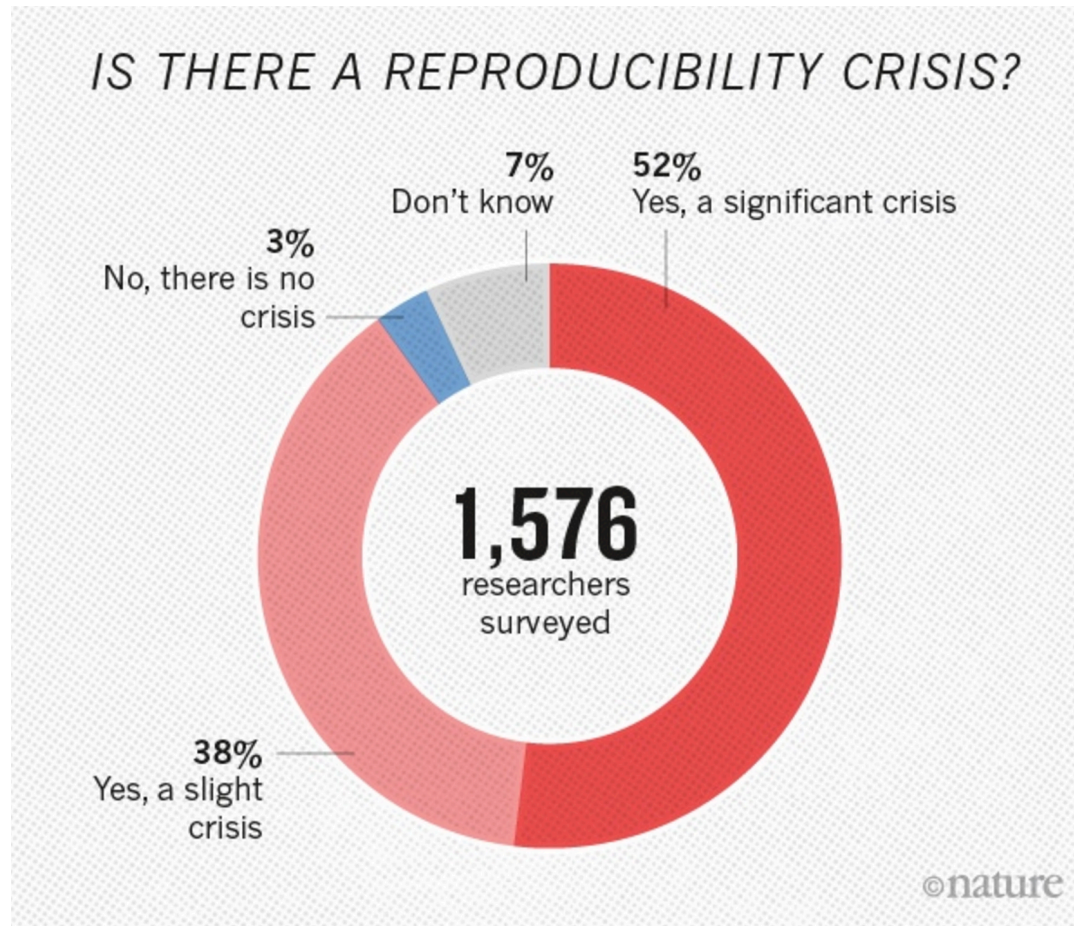
Reproducibility crisis



Open Science Collaboration, “Estimating the reproducibility of psychological science”, *Science*, **349** (2015)

Managed to reproduce only 39% results

Reproducibility crisis



Nature's survey of 1,576 researchers

WHAT FACTORS COULD BOOST REPRODUCIBILITY?

Respondents were positive about most proposed improvements but emphasized training in particular.



The great reproducibility experiment

Are referees more likely to give red cards to black players?



Mario Balotelli, playing for Manchester City, is shown a red card during a match against Arsenal.

Silberzahn et al., “Many analysts, one dataset: Making transparent how variations in analytical choices affect results”, <https://osf.io/j5v8f>

- one data set
- 29 teams
- 61 scientists
- task: find odds ratio

ONE DATA SET, MANY ANALYSTS

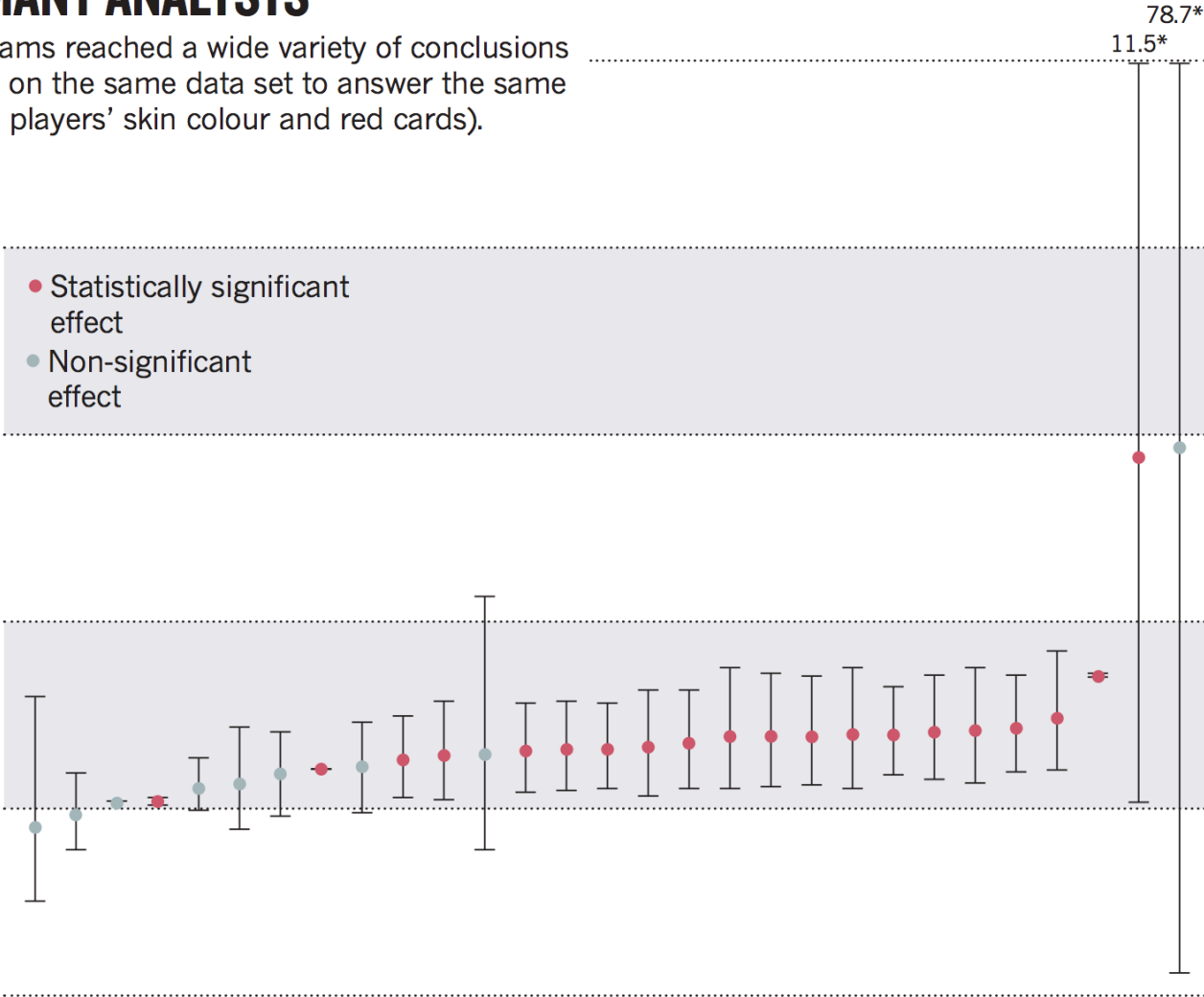
Twenty-nine research teams reached a wide variety of conclusions using different methods on the same data set to answer the same question (about football players' skin colour and red cards).

Dark-skinned players four times more likely than light-skinned players to be given a red card.

- Statistically significant effect
- Non-significant effect

Twice as likely

Equally likely



Point estimates and 95% confidence intervals. *Truncated upper bounds.

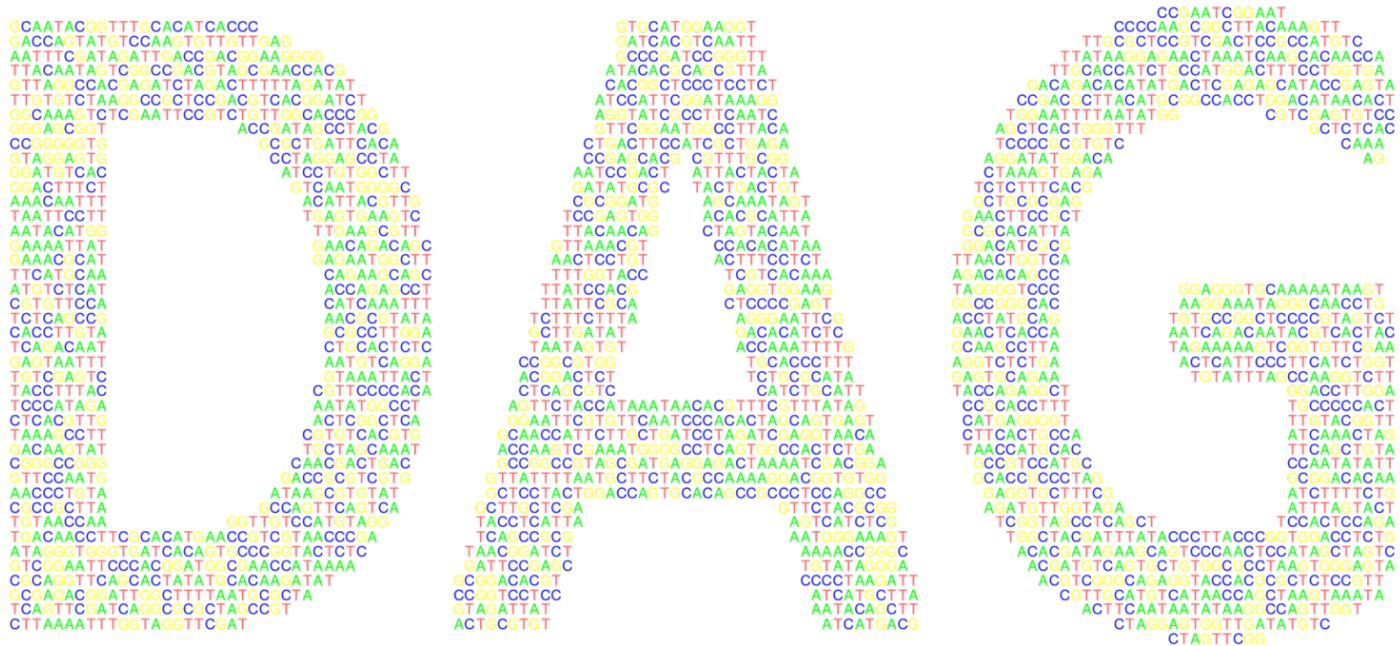
Science is broken

We are broken

What do we do?

What the hell do we do?

Before you do the experiment



talk to us

The Data Analysis Group

<http://www.compbio.dundee.ac.uk/dag.html>

Specify the null hypothesis

Design the experiment

- randomization
- statistical power

Quality control

some crap comes out in statistics

Ditch the α limit

use p-values as a continuous measure of data incompatibility with H_0

$p \sim 0.05$ only means '**worth a look**'

Reporting a discovery based only on $p < 0.05$ is **wrong**

We assumed the null hypothesis

Never, ever say that large p supports H_0

Use the three-sigma rule
that is $p < 0.003$, to demonstrate a discovery

Reporting

- Always report the effect size and its confidence limits
- Show data (not dynamite plots)
- Don't use the word 'significant'
- Don't use asterisks to mark 'significant' results in figures

Validation

Follow-up experiments to confirm discoveries

Publication

Publish negative results

ASA Statement on Statistical Significance and P-Values

1. P-values can indicate how incompatible the data are with a specified statistical model
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold
4. Proper inference requires full reporting and transparency
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis

https://is.gd/asa_stat



Hand-outs available at
<http://is.gd/statlec>