Neurobiology Research Unit
Rigshospitalet
Copenhagen University Hospital

Hacking, HARKing and SHARKING your research: a tutorial

Cyril Pernet, PhD

Based on my lecture: 10.6084/m9.figshare.5451067

WARNING: If you have issues 'reading' me when I talk, this is meant to be sarcastic and you must not engage in such practices

October 2023

Intended Learning Outcomes

- Being able to explain what is and what is not a p-value
- List various procedures that lead to p-hacking

Rules of engagement

- you can stop me anytime to ask questions
- whenever a gray rectangle appears, an activity/answer is needed, often in small groups (say 3 people)
- group yourselves now

Question box

Definitions

Let's start by you telling me what you think I'll be talking about!

- **What is p-hacking?**
- **Have you heard of HARKING and SHARKING**

Definitions

- **p-hacking:** employ methods and techniques that allow you to achieve statistically significant *p*-values

- **HARKing**: Kerr (1998) **H**ypothesizing **A**fter the **R**esults are Known. Personality and Social Psychology Review, 2, 196-217

- **SHARKing**: Poldrack et al. (2017) **S**electing **H**ypothesized **A**reas after **R**esults are **K**nown. Nature Reviews Neuroscience, 18, 115-126.

How to p-hack? the basics

Simmons et al. (2011) let play with the **'researcher degrees of freedom'**
Szucs D (2016) A Tutorial on Hunting Statistical Significance by Chasing

- Do multiple analyses 'adjusting' the data selection (within subject):
  - use different thresholds to clean data (e.g reaction times)
  - use different outlier methods to remove to low/high values

→ iterate until you get the right p value (no need to think and justify why a threshold or method over another one)

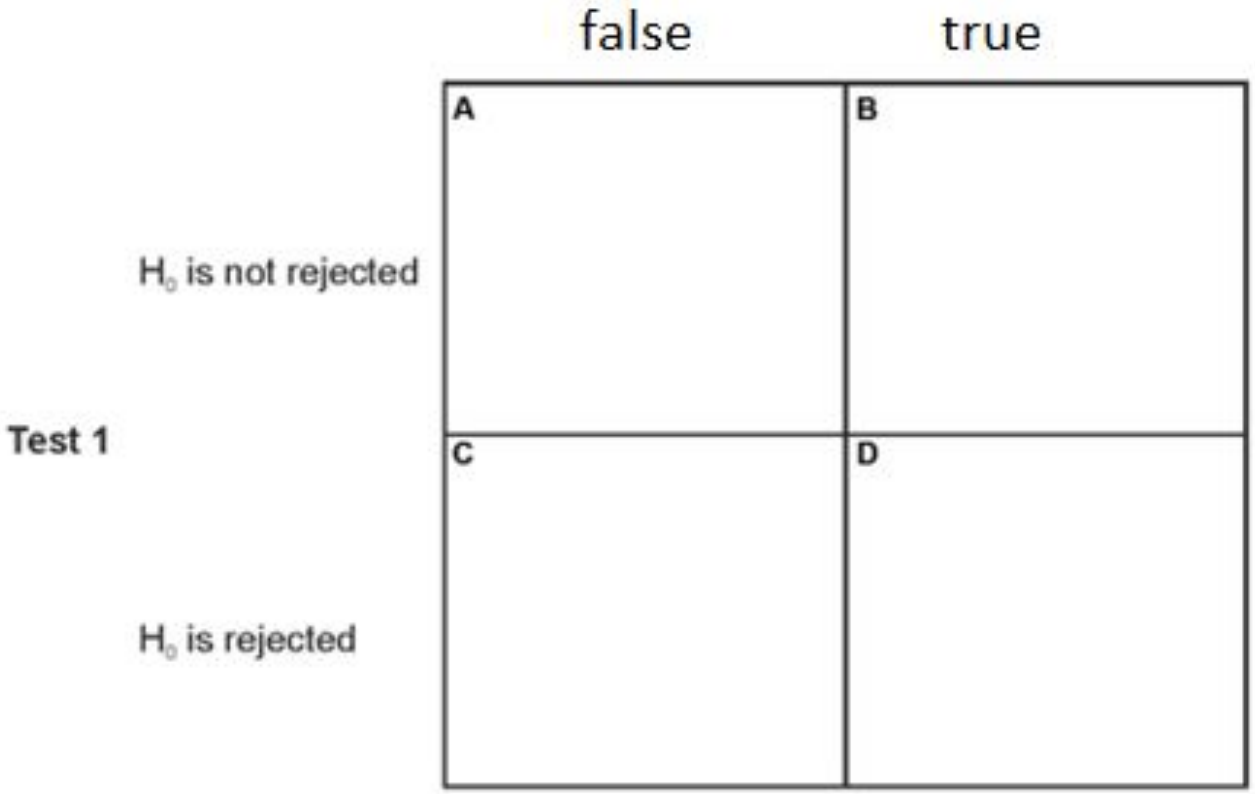
- Employ different statistical tests: your t-test did not work? Have you tried LM, HLM, LMM? Surely it's a distribution problem, use non-parametric in case you can get a significant result

How to p-hack? the basics

- Test several times ‘adjusting’ for participants (between subjects)
  - tests various outlier methods again or even ad-hoc subjects’ elimination until the right selection of subjects gives the expected results
  - collect new subjects and test regularly until significance is achieved (it’s just you did not have enough statistical power)
- Do these analyses on many variables:
  - when possible, collect several dependent variables and test them all separately, then pick those for which you got good results and do not mention the others (remember those drug trials per 2000/mandatory registration, they know how to do it)

Why is it bad for science?

- Because the type 1 error rate is not controlled

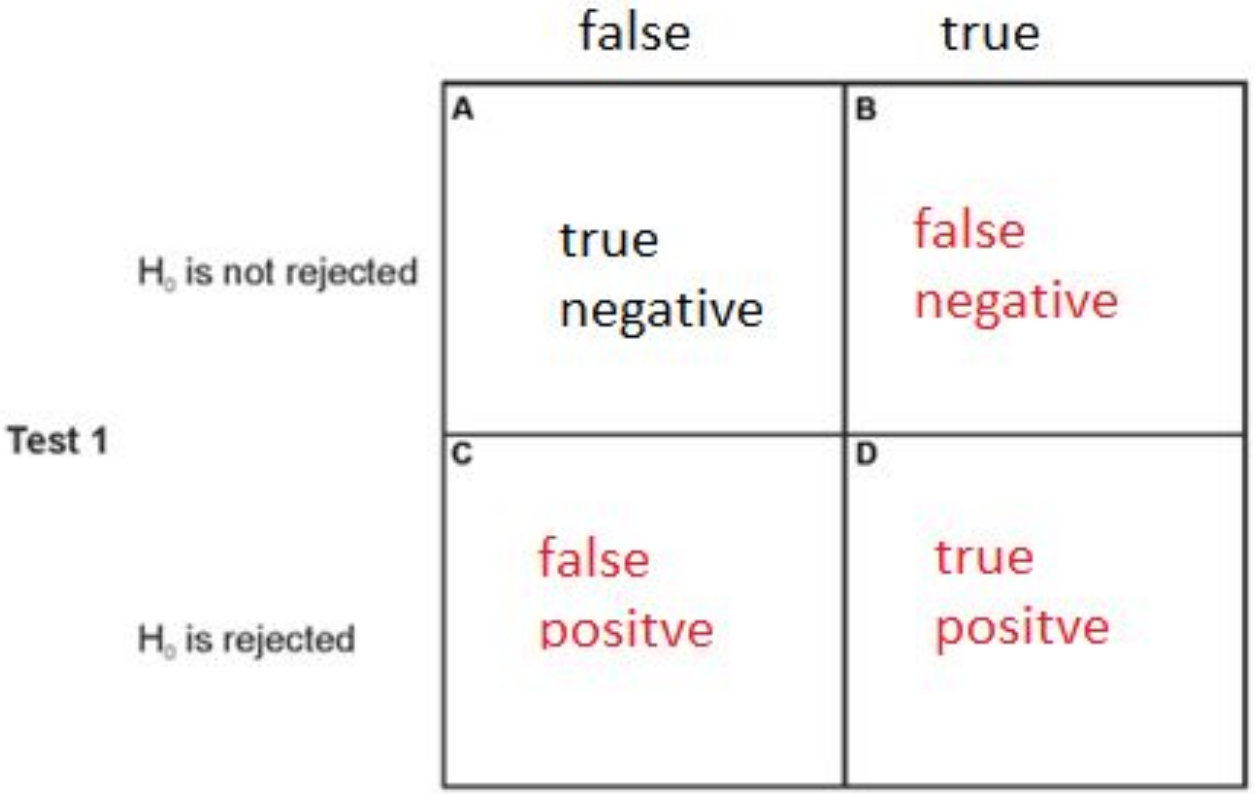


Remind me where those go:

- **True negative**
- **True positive**
- **False negative**
- **False positive**

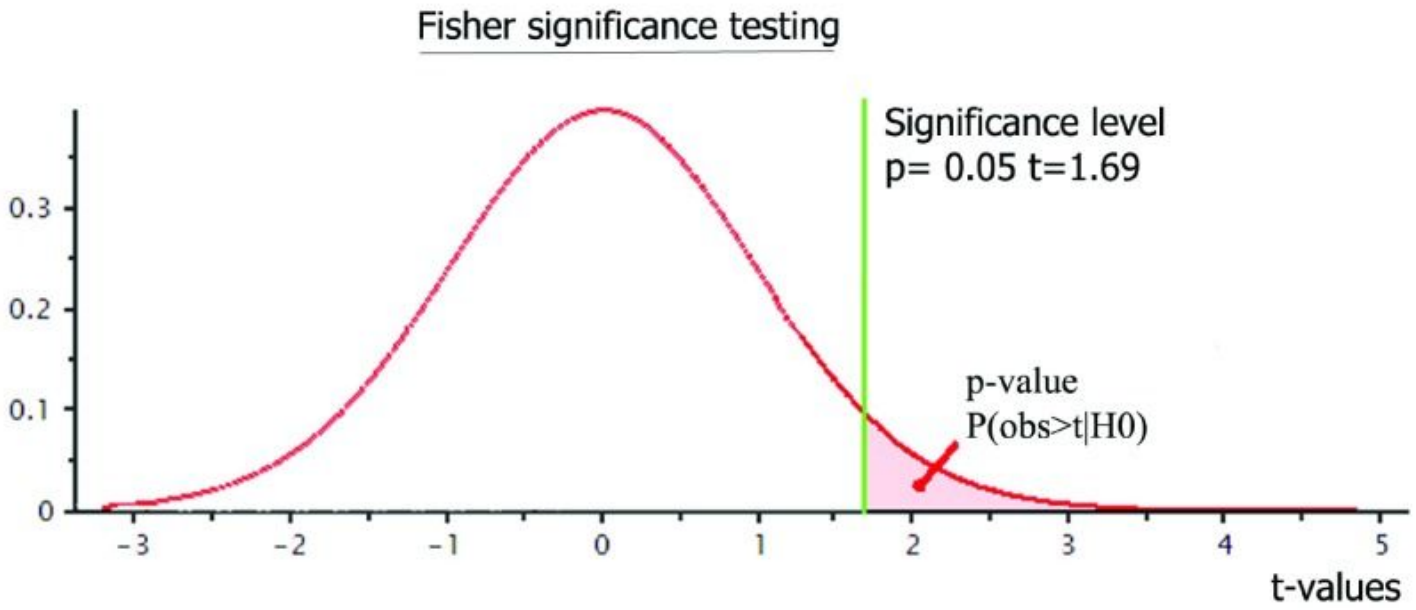
Why is it bad for science?

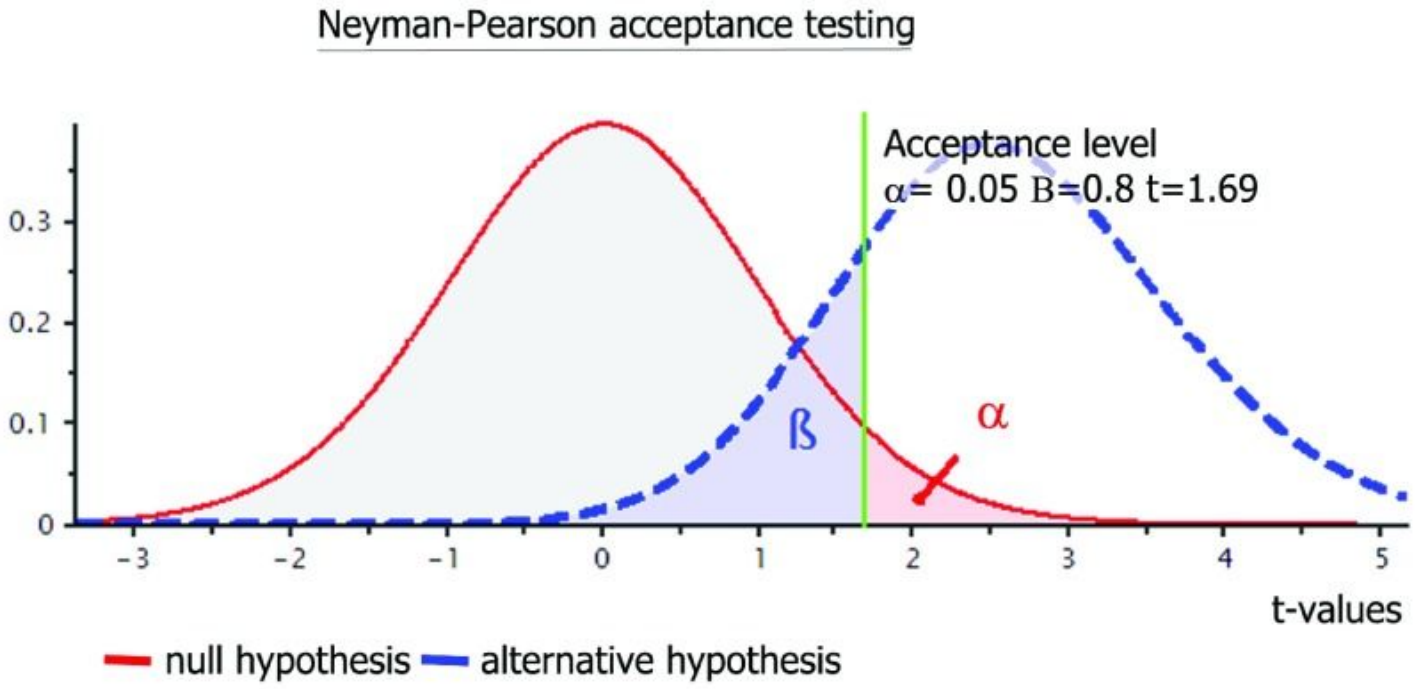
- Because the type 1 error rate is not controlled



- **And the type 1 error rate is?**

- **Alpha vs p-value?**





- **Alpha vs p-value?**

C Pernet (2016) NHST A short tutorial
10.12688/f1000research.6963.3

Why is it bad for science?

- Because the type 1 error rate is not controlled

Test 1 \ Test 2	H_0 is not rejected	H_0 is rejected
H_0 is not rejected	A: 0.95 × 0.95 = 0.9025	B
H_0 is rejected	C	D

Let's try 2 tests now using alpha = 0.05.

Prob(test) not significant is thus 0.95, for 2 tests 0.9025
What about other cases?
What is the total type 1 error?

Why is it bad for science?

- Because the type 1 error rate is not controlled

		Test 2	
		H_0 is not rejected	H_0 is rejected
Test 1	H_0 is not rejected	A 0.95 × 0.95 = 0.9025	B 0.95 × 0.05 = 0.0475
	H_0 is rejected	C 0.05 × 0.95 = 0.0475	D 0.05 × 0.05 = 0.0025

At least one false positive = 1 - 0.95×0.95 = 0.0475 + 0.0475 + 0.0025 = 0.0975

Under the null (i.e. we know there is no effect), with alpha at 5% and 2 variables we have already 9% of false positives … you will find significant results by running many analyses !

FWER = $1-(1-\text{alpha})^n$
for n independent tests

Controlling type on error rate?

- Do you know how any techniques to do that?

Controlling type on error rate?

- Bonferroni correction $P(k) < \alpha/n$
- better → Holm-Bonferroni $P(k) < \alpha/(n+1-k)$
- For non independent tests? Max statistics
  If FWER is $p<\alpha$, then controlling the largest statistics at α ensures the FWER
- For spatial statistics: random field theory (ensure the assumptions apply!)

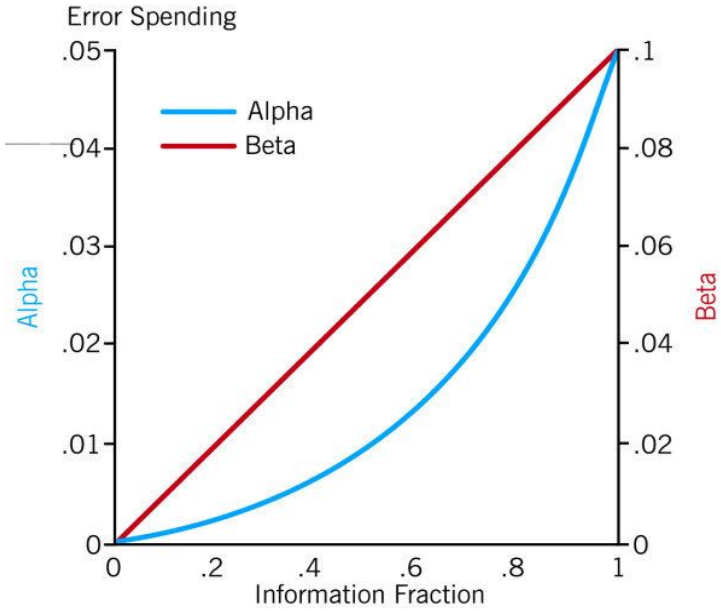
A popular approach is the False Discovery Rate correction -- a fine method

- **Why FDR does not control FWER?**

Controlling type on error rate?

Procedures for sequential testing
-- acquiring data and testing regularly until significant is p-hacking
-- at the same time it can save a lot of time and money

Alpha spending vs Bayes
-- alpha spending, give maximum sample size set alpha and beta
-- alternatively use a Bayesian procedure (no alpha, this is a freq. notion) and simply use the posterior probability of your effect (but you need good or conservative priors for this to work)



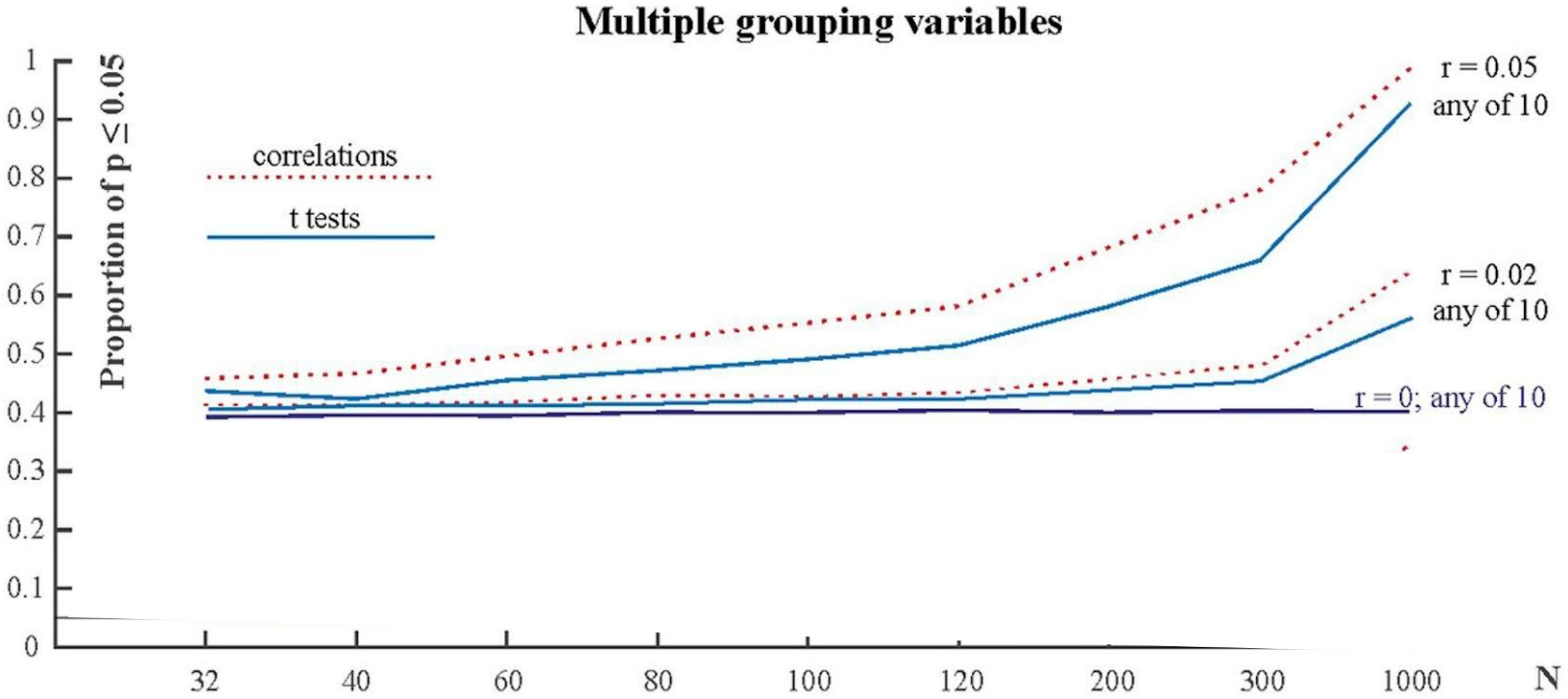
How to p-hack? - advanced

- Use multiple fitting procedures (e.g. raw data, distribution transform, generalized linear models / ML vs RML / frequentist vs Bayesian)

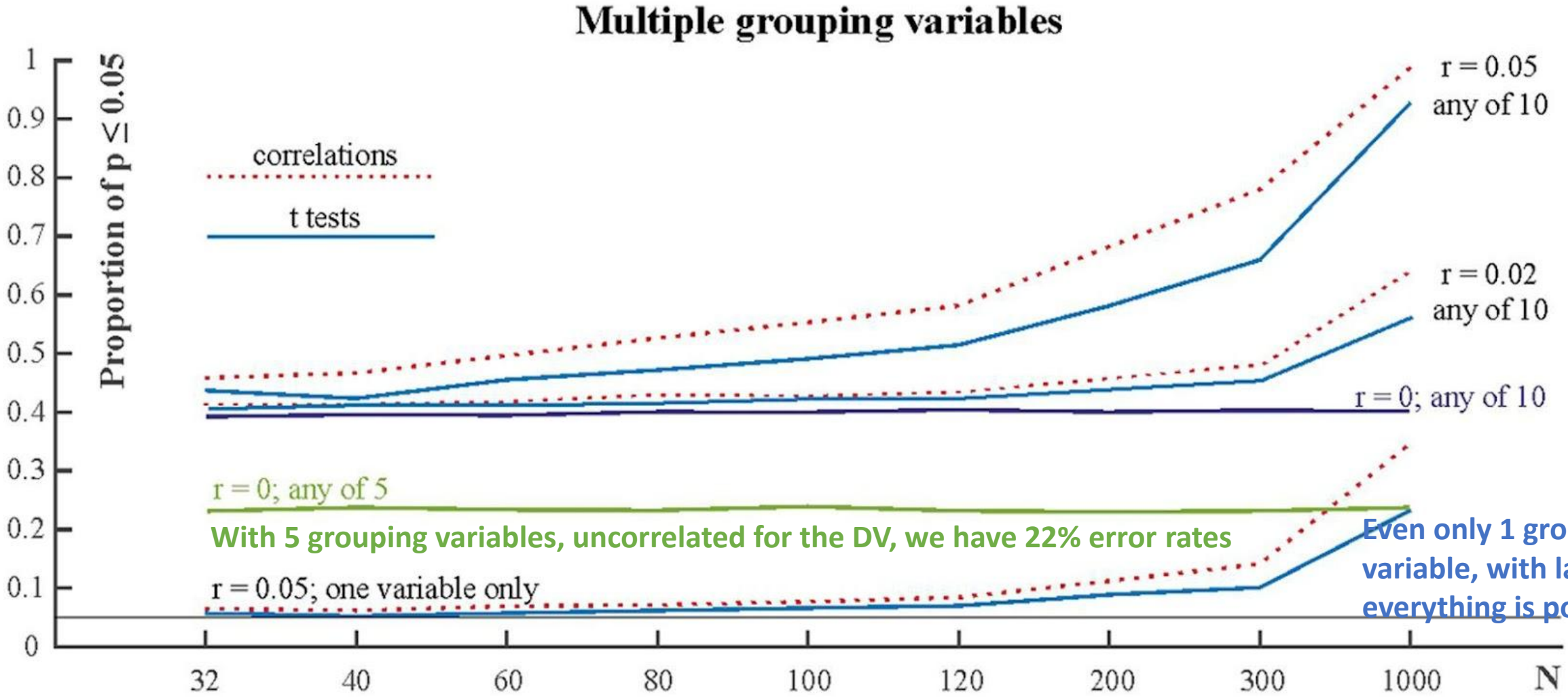
- Use ad-hoc grouping based on additional variables and/or use covariates: this approach is very ‘useful’ in large databases as you will always find some correlated variables to the DV to create ad-hoc sub-groups to run analyses on

Why is it bad for science?

Add reference here



Why is it bad for science?



With 5 grouping variables, uncorrelated for the DV, we have 22% error rates

Even only 1 grouping variable, with large N everything is possible

Harking as a way to p-hack

- Despite torturing the data, sometimes they just don’t want to confess !
- Luckily enough, we often have multiple dependent variables or complex design with many factors and some interactions come out significant – the problem with that is that some of these measures/effects were not really in our hypothesis

 Search the literature to find evidences that the new dependent variable/effect is influenced by your experimental factors and write a nice introduction and hypothesis

Why is it bad for science?

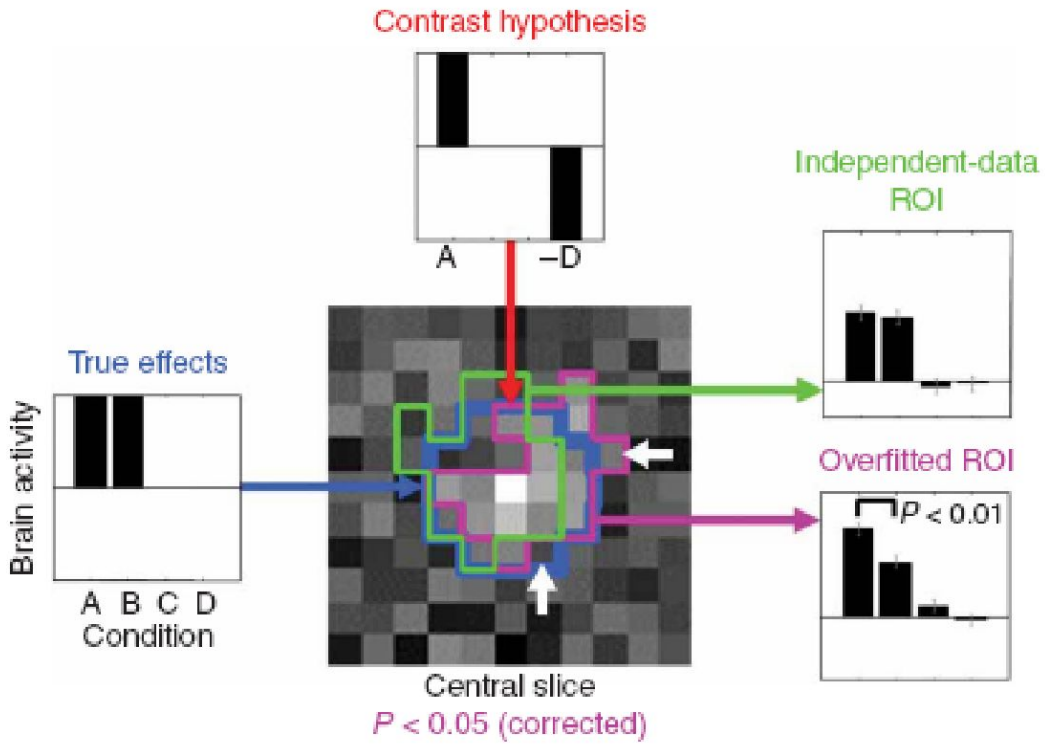
- Because the scientific model relies on falsification not confirmation
- New harked hypothesis is, after being made from observed data rather than theory/literature, unfalsifiable !
- Assuming the new hypothesis is true, the current study was not made to answer it at best (design and power issues)
- The old discarded hypothesis is not presented and it’s falsification (the actual essence of the scientific method) is not presented

SHARKING or p-hacking in space

- PET, MRI, MEEG are great ! We measure many and many data points in space and/or time and/or frequencies
- Basic hacking = do all the statistics for each point without correcting for multiple testing (used to work well but people got fed-up with that)
- SHARKING = look at the signal for each data point and select regions showing the strongest effects ; then restrict the analysis to those (use HARKing to justify your choices)

SHARKING in fMRI and double-dipping

- Selection and tests must be independent – non independence create spurious effects (similar to sub-grouping using additional variables)

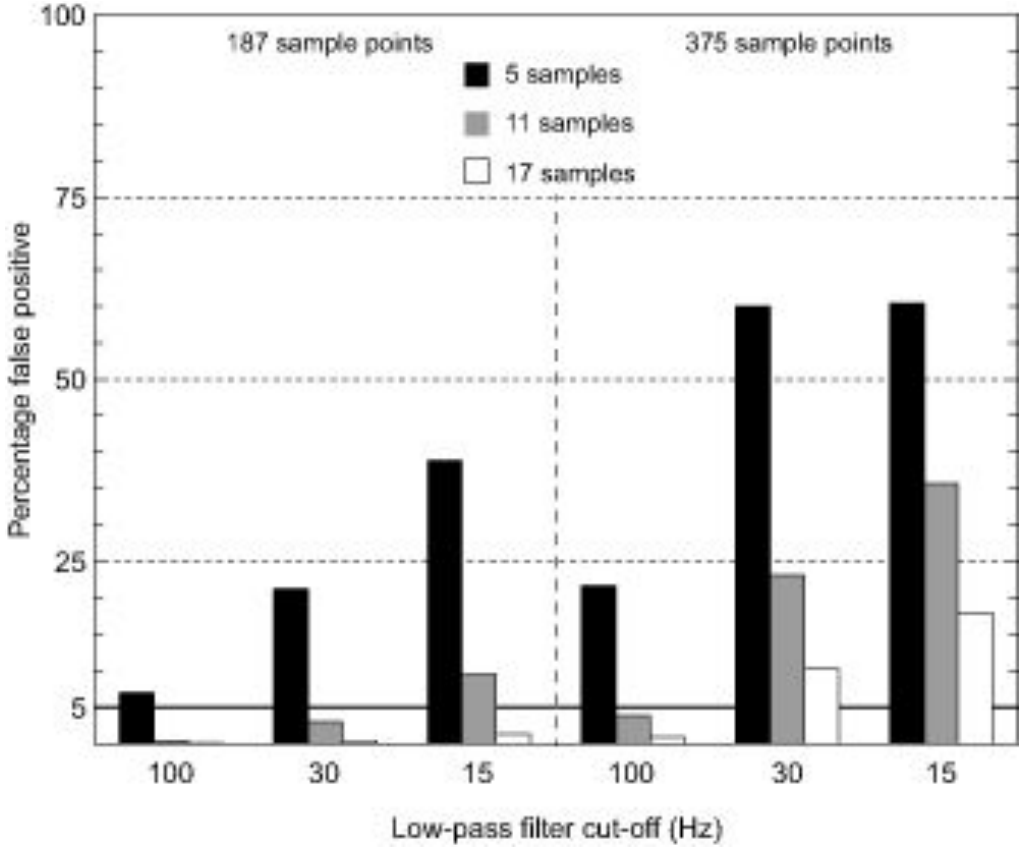


Kriegeskorte et al., 2009 *Nat. Neuroscience* 12

Make sure your 'a priori' regions fit well the observed signal it space as it will give more significant results and differences, since the noise is driving some of the response in the 'right' direction

SHARKING in time and the magic window

- Look at where you see differences in a time-series and perform the analysis only there, for a restricted ‘window’



Use standard stats and accept if multiple neighbours are significant

- choose ‘multiple’ as needed, eg 5 time points
- use the ‘based on autocorrelation’ argument if needed to justify your threshold
- Longer signal, strong filtering and small samples gives you more significance

Piai et al. Psychophysiology. 2015 Mar;52(3):440-3.

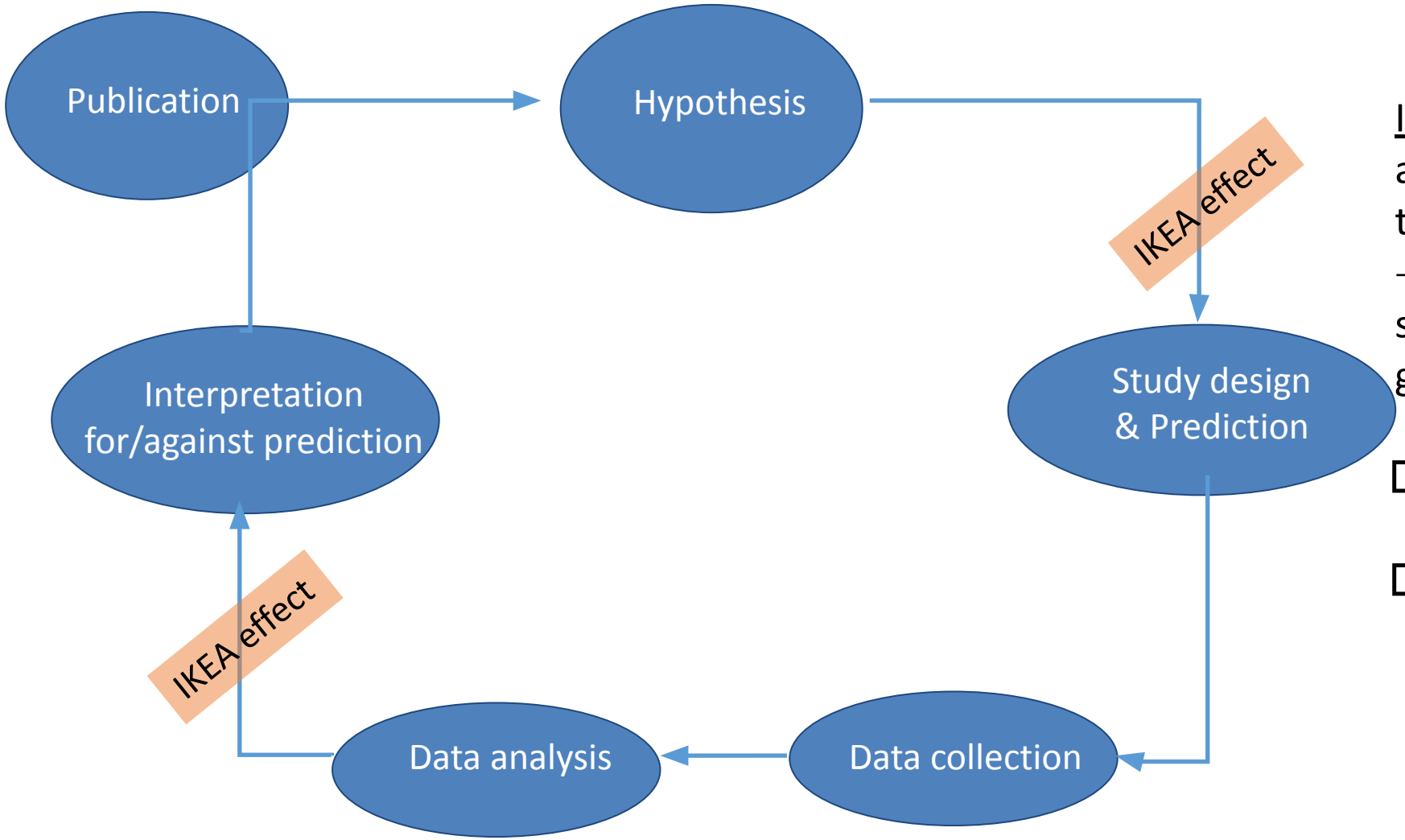
Are people cheating deliberately?

What's a cognitive bias?

- A **cognitive bias** refers to the systematic pattern of deviation from norm or rationality in judgment, whereby inferences about other people and situations may be drawn in an illogical fashion (Wikipedia).
- A **cognitive bias** is a mistake in reasoning, evaluating, remembering, or other **cognitive** process, often occurring as a result of holding onto one's preferences and beliefs regardless of contrary information. Psychologists study **cognitive biases** as they relate to memory, reasoning, and decision-making.

'Let's think about cognitive bias' *Nature editorial Oct 2015*

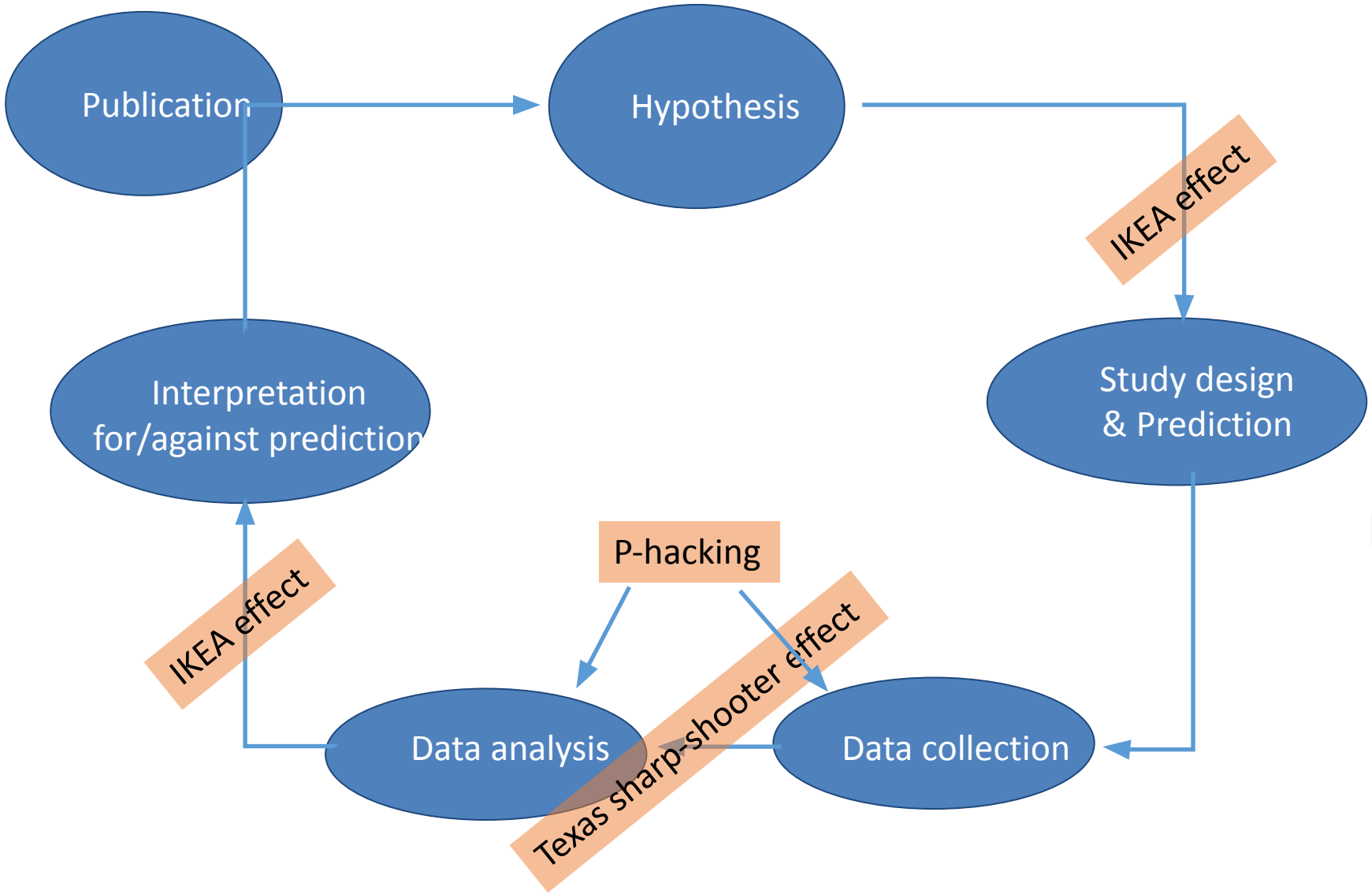
The scientific method (hack 1)



IKEA effect: consumers place artificially high value on products that they have built themselves → your own design and analysis script/method might not be has good as what you think

- **Discuss Design and analysis plan**
- **Have you coded yourself some method? How do you now this is correct**

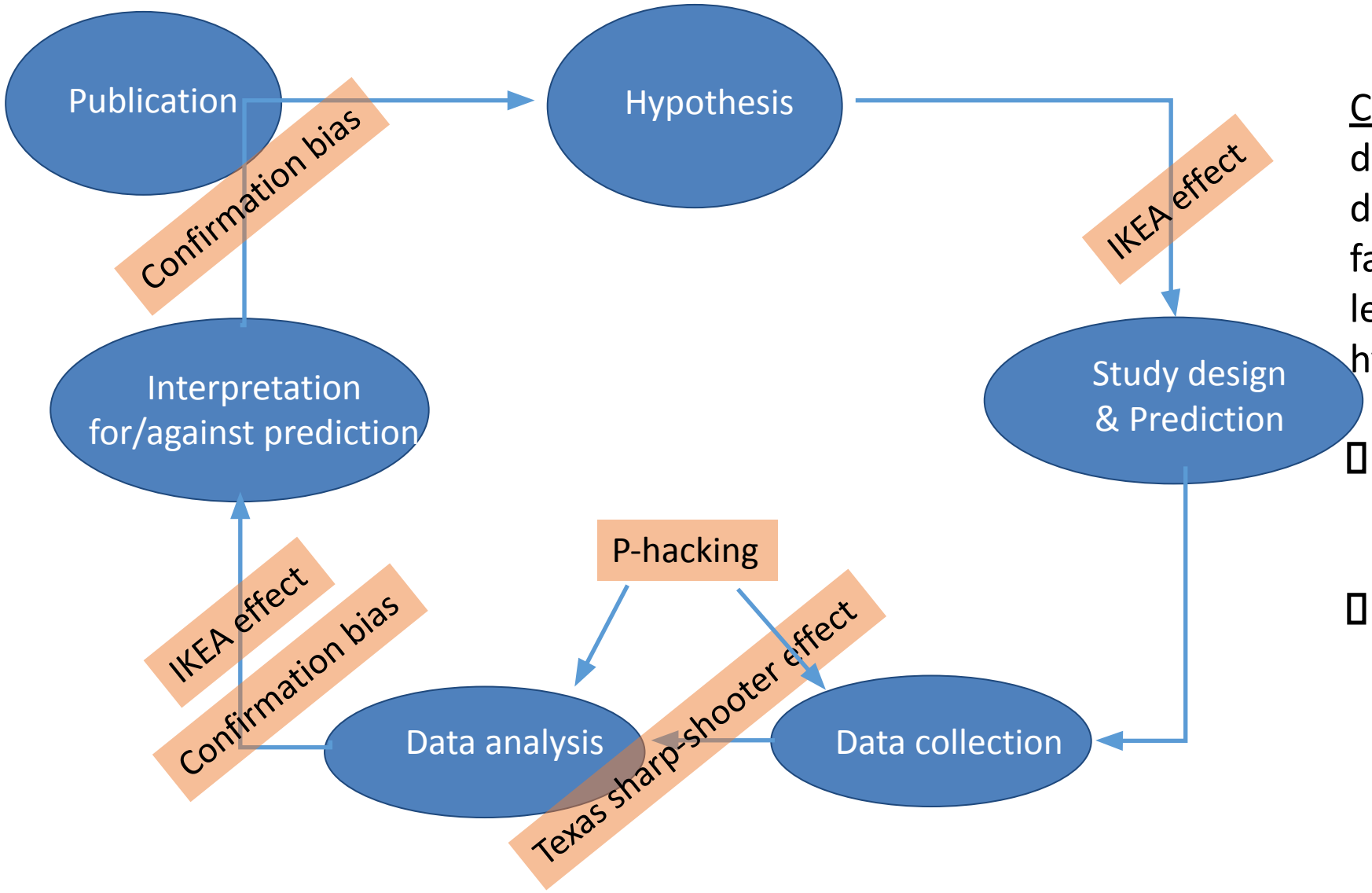
The scientific method (hack 2)



Texas sharp-shooter effect: firing off a few rounds and then drawing a bull's eye around the bullet holes

- trying many analyses techniques until is start showing some results.
- **Plan analyses based on hypotheses**

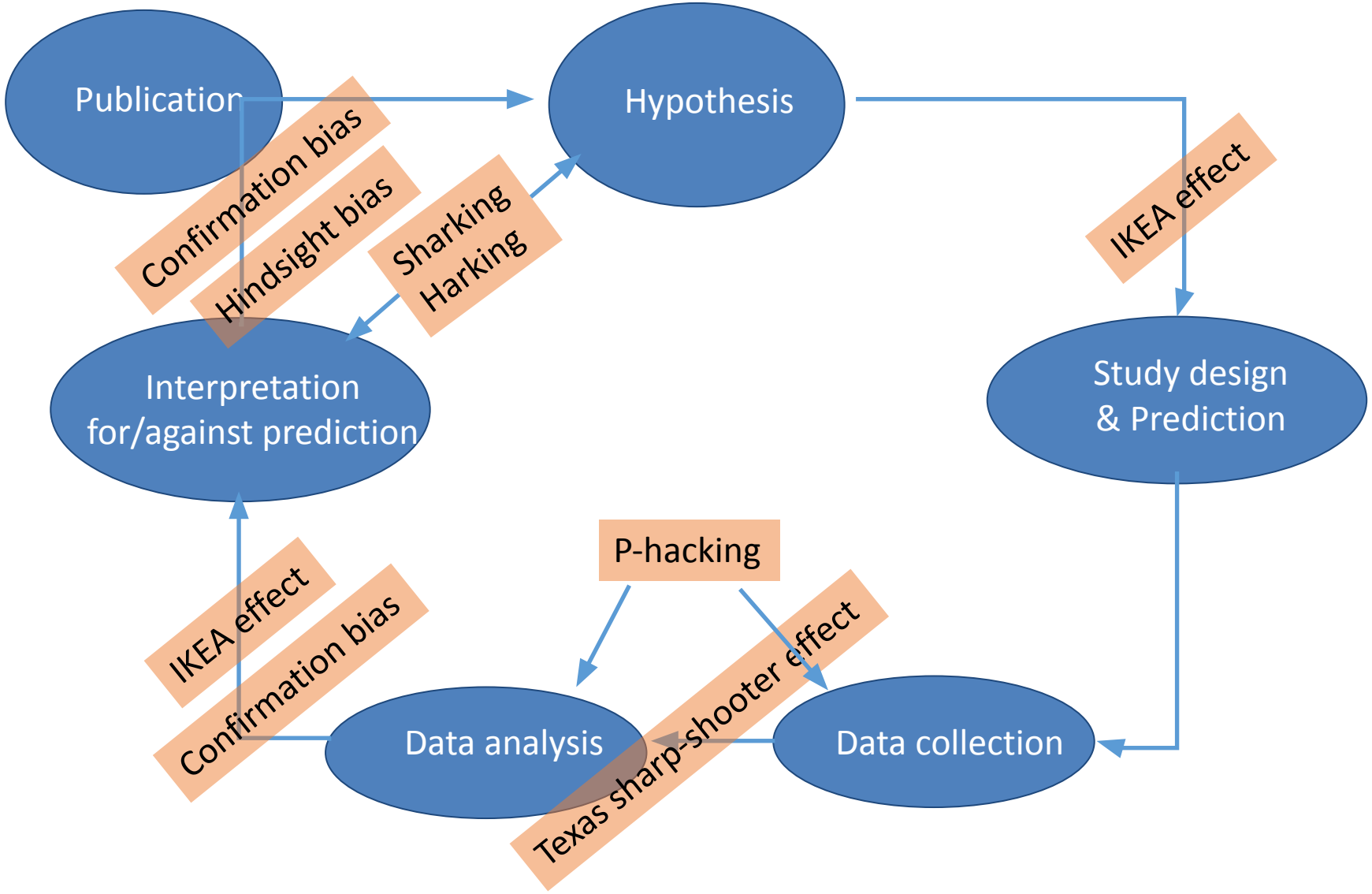
The scientific method (hack 3)



Confirmation bias: carefully debugging analyses and debunking data that counter a favoured hypothesis, while letting evidence in favour of the hypothesis slide by unexamined

- **Use data blinding** (removing 'outliers' that do not fit the hypothesis)
- **Discuss your code and get it reviewed – how do you know it is correct? Because it gives results you expect !?**

The scientific method (hack 4)



Hindsight bias: inclination, after an event has occurred, to see the event as having been predictable, despite there having been little or no objective basis for predicting it.

Hypothesizing **A**fter the **R**esults are **K**nowns (Kerr, 1998) and **S**electing **H**ypothesized **A**reas after **R**esults are **K**nown (Poldrack et al 2017)

- **Hypotheses, hypotheses, hypotheses !**
- **Independent ROI** (related to circularity analyses too)

Again a sarcastic tone is implied!

How to p-hack: TAKE HOME MESSAGE

- Don’t fight your biases ! Sure it’s bad for science, but you are only human and want to publish flashy articles

✔ Don’t plan, discuss or get reviewed your study, design, and analyses

✔ Try many analysis strategies to make the most sense (understand the most significant) of the data

✔ Trust your own code as long as it gives you the expected results

✔ Change hypotheses if unexpected variables appear to have nice effects

✔ Define ROI in space or time after looking at where the signal is, and run analyses on those regions, reducing the too stringent control for type 1 error

Intended Learning Outcomes

- Being able to explain what is and what is not a p-value
- List various procedures that lead to p-hacking

Some references

- http://journal.frontiersin.org/article/10.3389/fpsyg.2016.01444/full
- http://freakonometrics.hypotheses.org/19817
- https://www.youtube.com/watch?v=A0vEGuOMTyA
- https://fivethirtyeight.com/features/science-isnt-broken/#part1
- http://www.nature.com/neuro/journal/v20/n6/full/nn.4550.html