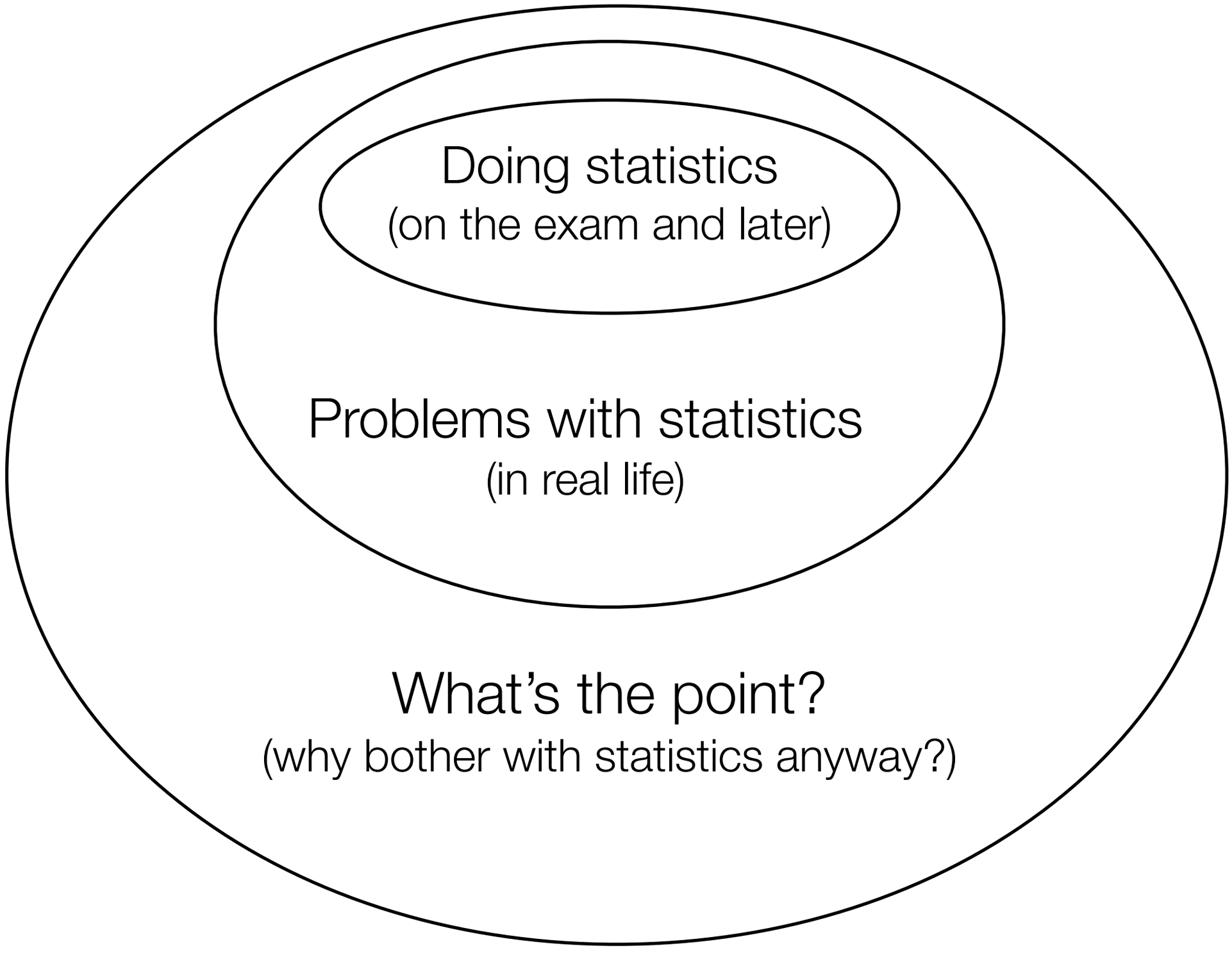
Wrapping it all up

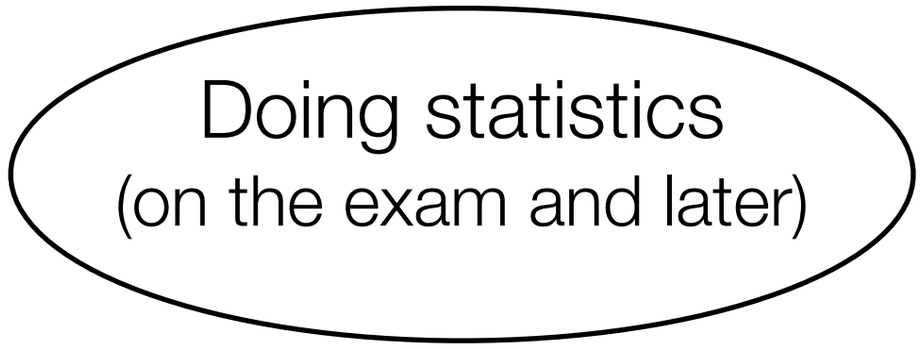
Research Methods for Human Inquiry
Andrew Perfors

For the final video let’s take a step back for more of a “forest” view of all of this, rather than the trees we’ve been mired in lately

What's in the forest?

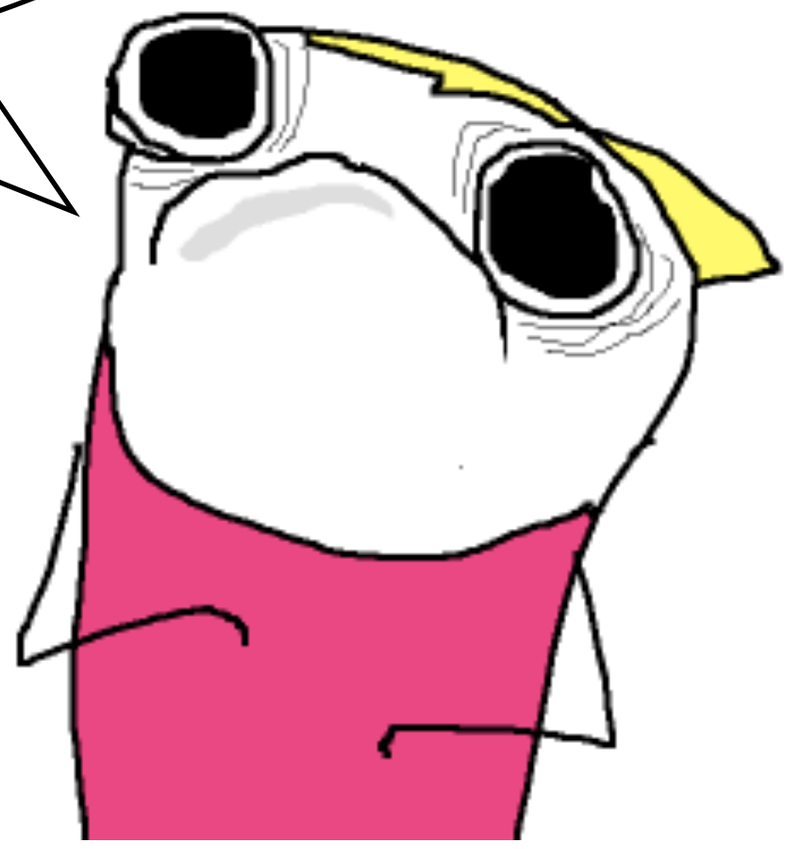


What's in the forest?



Doing statistics

RMHI has *so much stuff* to think about. When I have a dataset, a problem set or a research question, how do I even begin?? It's hugely overwhelming!



Doing statistics: A handy breakdown

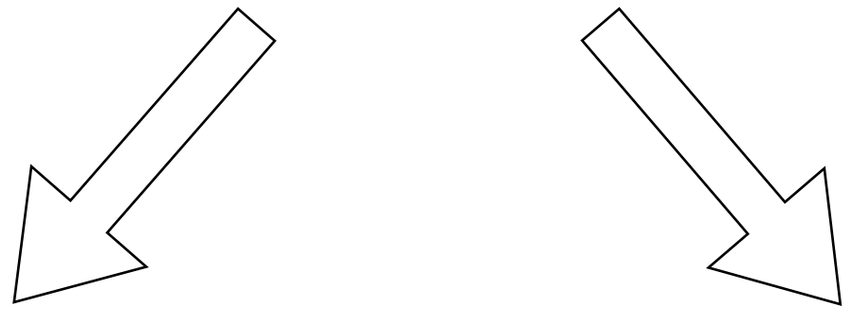
Step	
1	What is your research question?
2	What are your variables? What do they look like?
3	How does the research question determine what you are comparing?
4	How does your comparison map onto a specific statistical test?
5	How can you test the assumptions of your test?
6	How do you do the test in R?
7	How do you interpret the test in light of your research question?

Doing statistics: A handy breakdown

Step	
1	What is your research question?

Suppose I wanted to figure out if this subject helped people in Bunnyland

operationalise
“helped”



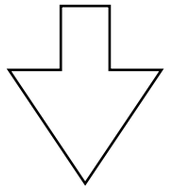
did people
like it?

did people
learn?

Doing statistics: A handy breakdown

Step	
2	What are your variables? What do they look like?

did people like it?

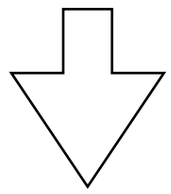


survey? Likert scale?

on a scale of 1-7, how much did you like the subject?

1…2…3…4…5…6…7

did people learn statistics?

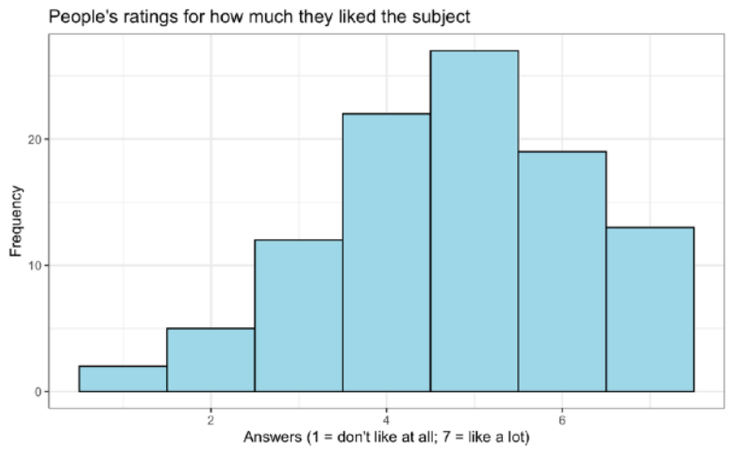


Performance on a stats exam compared to people who weren't enrolled?

0-100%

Doing statistics: A handy breakdown

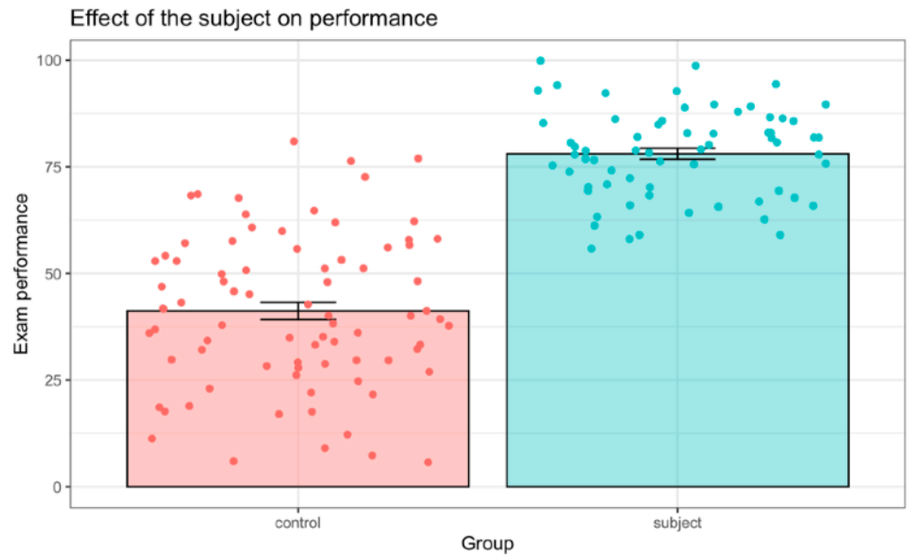
Step	
2	What are your variables? What do they look like?



survey? Likert scale?

on a scale of 1-7, how much did you like the subject?

1…2…3…4…5…6…7



Performance on a stats exam compared to people who weren't enrolled?

0-100%

Doing statistics: A handy breakdown

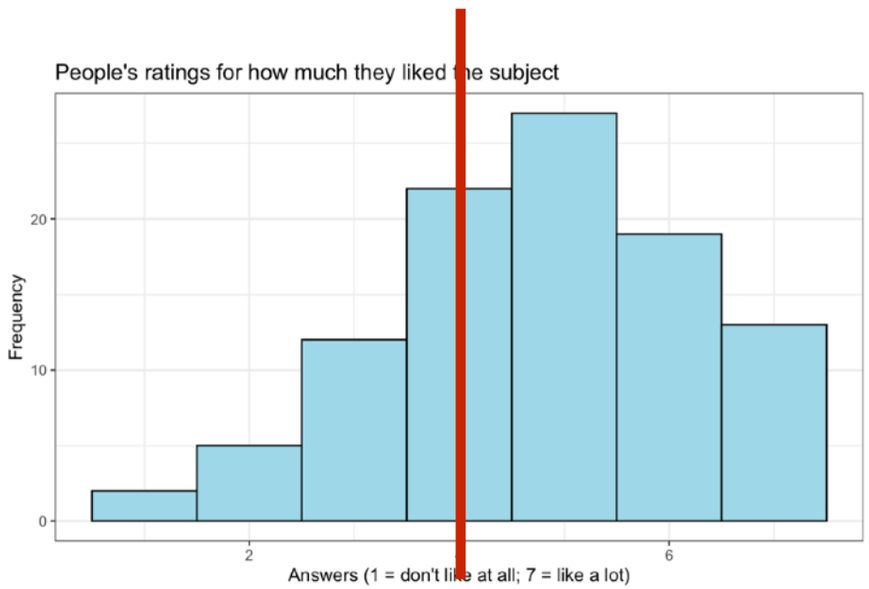
	Step	Example
1	What is your research question?	I want to figure out if this subject helped Bunnyland folks… 1. did they like it? 2. did people learn statistics?
2	What are your variables? What do they look like?	1. Survey question, scale from 1-7 (`liking`) 2. Exam scores between two groups (`control,subject`)

Doing statistics: A handy breakdown

Step	
3	How does the research question determine what you are comparing?

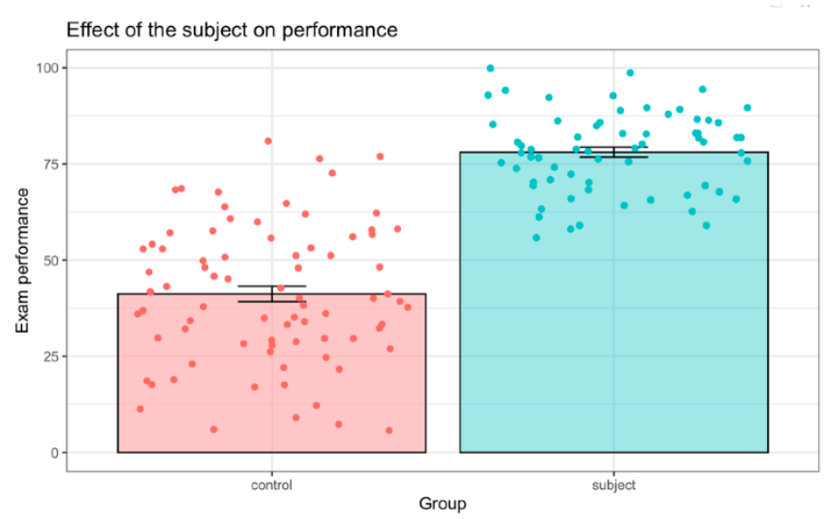
did people like it?

is average `liking` higher than middle of scale?



did people learn statistics?

Is mean `performance` higher in the `subject` group?



Doing statistics: A handy breakdown

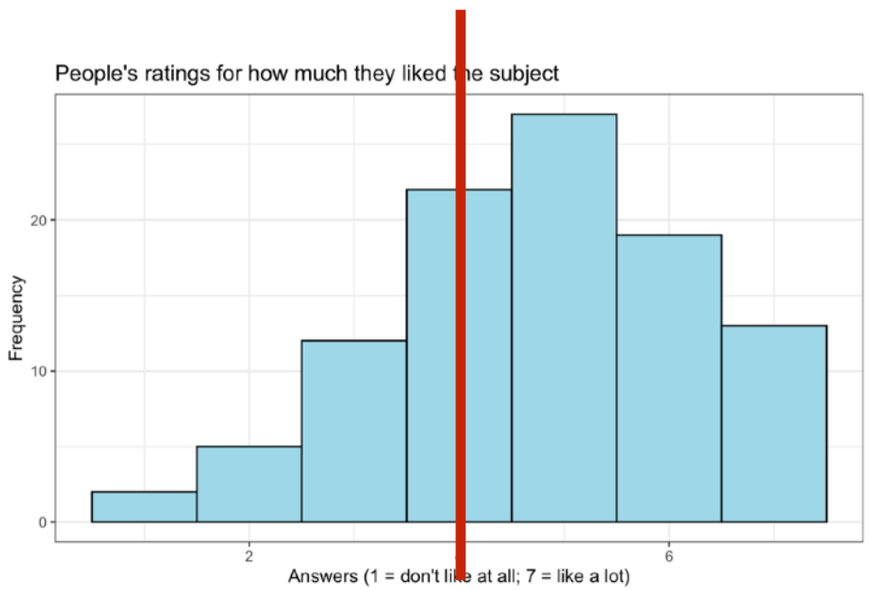
	Step	Example
1	What is your research question?	I want to figure out if this subject helped Bunnyland folks… 1. did they like it? 2. did people learn statistics?
2	What are your variables? What do they look like?	1. Survey question, scale from 1-7 (`liking`) 2. Exam scores between two groups (`control`,`subject`)
3	How does the research question determine what you are comparing?	1. Compare `liking` to a standard (e.g., 50% of scale?) 2. Is `performance` better in the `subject` group?

Doing statistics: A handy breakdown

Step	
4	How does your comparison map onto a specific statistical test?

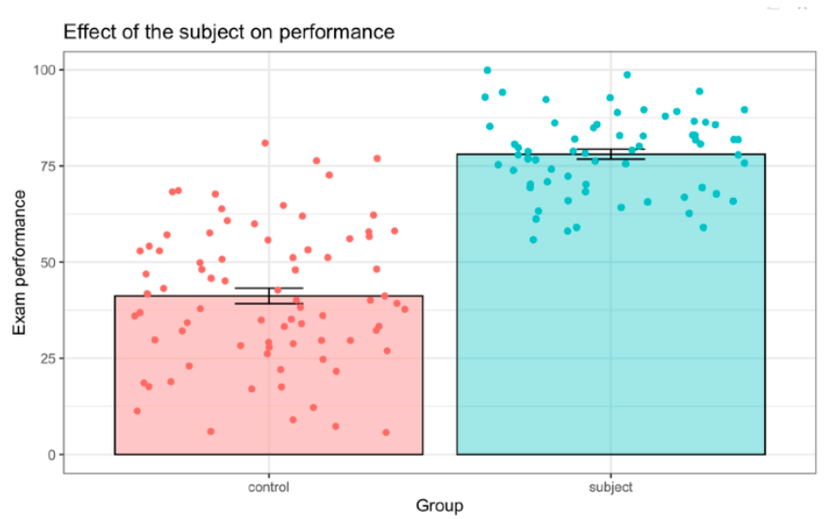
comparing the mean of `liking` to a standard (4): one-sample t-test

is average `liking` higher than middle of scale?



comparing two means (`performance` by `group`): independent-samples t-test

Is mean `performance` higher in the `subject` group?



Doing statistics: A handy breakdown

	Step	Example
1	What is your research question?	I want to figure out if this subject helped Bunnyland folks… 1. did they like it? 2. did people learn statistics?
2	What are your variables? What do they look like?	1. Survey question, scale from 1-7 (`liking`) 2. Exam scores between two groups (`control`,`subject`)
3	How does the research question determine what you are comparing?	1. Compare `liking` to a standard (e.g., 50% of scale?) 2. Is `performance` better in the `subject` group?
4	How does your comparison map onto a specific statistical test?	1. One-sample t-test `liking` vs 4 (mid on 1-7 scale) 2. Independent-sample t-test, `performance` by `group`

Doing statistics: A handy breakdown

Step	
5	How can you test the assumptions of your test?

One-sample t-test: Normality

```
> shapiro.test(d$liking)

     Shapiro-Wilk normality test

data:  dl$liking
W = 0.94006, p-value = 0.0001945
```

Not normal: Need to do a one-sample Wilcoxon

Independent-samples t-test: Normality of each variable

```
> shapiro.test(dp$performance[dp$group=="subject"])

      Shapiro-Wilk normality test

data:  dp$performance[dp$group == "subject"]
W = 0.98308, p-value = 0.5165

> shapiro.test(dp$performance[dp$group=="control"])

      Shapiro-Wilk normality test

data:  dp$performance[dp$group == "control"]
W = 0.98763, p-value = 0.6403
```

Both groups are normal; can do Welch independent-samples t-test

Doing statistics: A handy breakdown

	Step	Example
1	What is your research question?	I want to figure out if this subject helped Bunnyland folks… 1. did they like it? 2. did people learn statistics?
2	What are your variables? What do they look like?	1. Survey question, scale from 1-7 (`liking`) 2. Exam scores between two groups (`control`,`subject`)
3	How does the research question determine what you are comparing?	1. Compare `liking` to a standard (e.g., 50% of scale?) 2. Is `performance` better in the `subject` group?
4	How does your comparison map onto a specific statistical test?	1. One-sample t-test `liking` vs 4 (mid on 1-7 scale) 2. Independent-sample t-test, `performance` by `group`
5	How can you test the assumptions of your test?	1. Shapiro-Wilk shows not normal: Wilcoxon 2. Shapiro-Wilk shows normal: Welch t-test

Doing statistics: A handy breakdown

Step	
6	How do you do the test in R?

comparing the mean of `liking` to a standard (4) when the data are not normal: Wilcoxon

```
> wilcox.test(dl$liking,mu=4)

    Wilcoxon signed rank test with continuity
correction

data:  dl$liking
V = 2441.5, p-value = 4.729e-06
alternative hypothesis: true location is not
equal to 4
```

comparing two means (`performance` by `group`): Welch independent-samples t-test

```
> t.test(performance~group,dp)

    Welch Two Sample t-test

data:  performance by group
t = -15.418, df = 130.97, p-value < 2.2e-16
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 -41.60840 -32.14545
sample estimates:
mean in group control mean in group subject
             41.20000              78.07692
```

Doing statistics: A handy breakdown

Step		Example
1	What is your research question?	I want to figure out if this subject helped Bunnyland folks… 1. did they like it? 2. did people learn statistics?
2	What are your variables? What do they look like?	1. Survey question, scale from 1-7 (`liking`) 2. Exam scores between two groups (`control,subject`)
3	How does the research question determine what you are comparing?	1. Compare `liking` to a standard (e.g., 50% of scale?) 2. Is `performance` better in the `subject` group?
4	How does your comparison map onto a specific statistical test?	1. One-sample t-test `liking` vs 4 (mid on 1-7 scale) 2. Independent-sample t-test, `performance` by `group`
5	How can you test the assumptions of your test?	1. Shapiro-Wilk shows not normal: Wilcoxon 2. Shapiro-Wilk shows normal: Welch t-test
6	How do you do the test in R?	1. `wilcox.test(dl$liking,mu=4)` 2. Independent-sample t.test(`performance~group,dp`)

Doing statistics: A handy breakdown

Step	
7	How do you interpret the test in light of your research question?

$p<0.05$,
CI/mean far from 4,
extreme test statistic:
people liked it!

$p<0.05$,
CI doesn't contain zero:
extreme t-statistic
people learned statistics!

```
> wilcox.test(dl$liking,mu=4)

    Wilcoxon signed rank test with continuity
correction

data:  dl$liking
V = 2441.5, p-value = 4.729e-06
alternative hypothesis: true location is not
equal to 4
```

```
> t.test(performance~group,dp)

    Welch Two Sample t-test

data:  performance by group
t = -15.418, df = 130.97, p-value < 2.2e-16
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 -41.60840 -32.14545
sample estimates:
mean in group control mean in group subject
             41.20000              78.07692
```

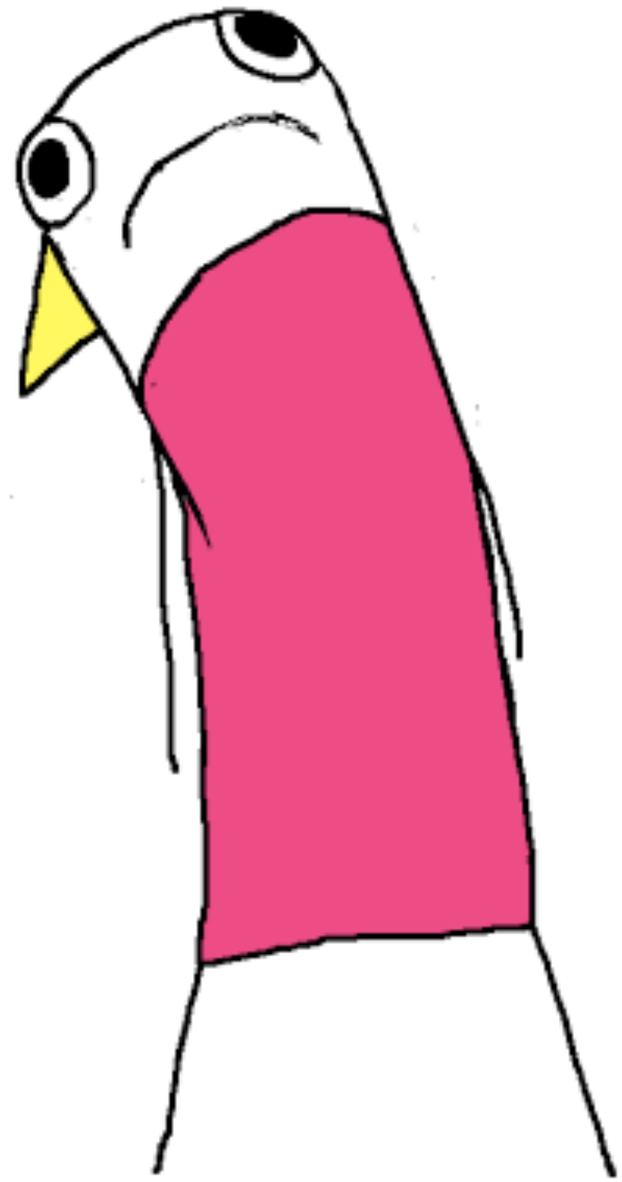
Doing statistics: A handy breakdown

Step		Example
1	What is your research question?	I want to figure out if this subject helped Bunnyland folks… 1. did they like it? 2. did people learn statistics?
2	What are your variables? What do they look like?	1. Survey question, scale from 1-7 (`liking`) 2. Exam scores between two groups (`control,subject`)
3	How does the research question determine what you are comparing?	1. Compare `liking` to a standard (e.g., 50% of scale?) 2. Is `performance` better in the `subject` group?
4	How does your comparison map onto a specific statistical test?	1. One-sample t-test `liking` vs 4 (mid on 1-7 scale) 2. Independent-sample t-test, `performance` by `group`
5	How can you test the assumptions of your test?	1. Shapiro-Wilk shows not normal: Wilcoxon 2. Shapiro-Wilk shows normal: Welch t-test
6	How do you do the test in R?	1. `wilcox.test(dl$liking,mu=4)` 2. Independent-sample t.test(`performance~group,dp`)
7	How do you interpret the test in light of your research question?	1. $p<0.05$, people liked it! 2. $p<0.05$, people learned statistics!

What's in the forest?

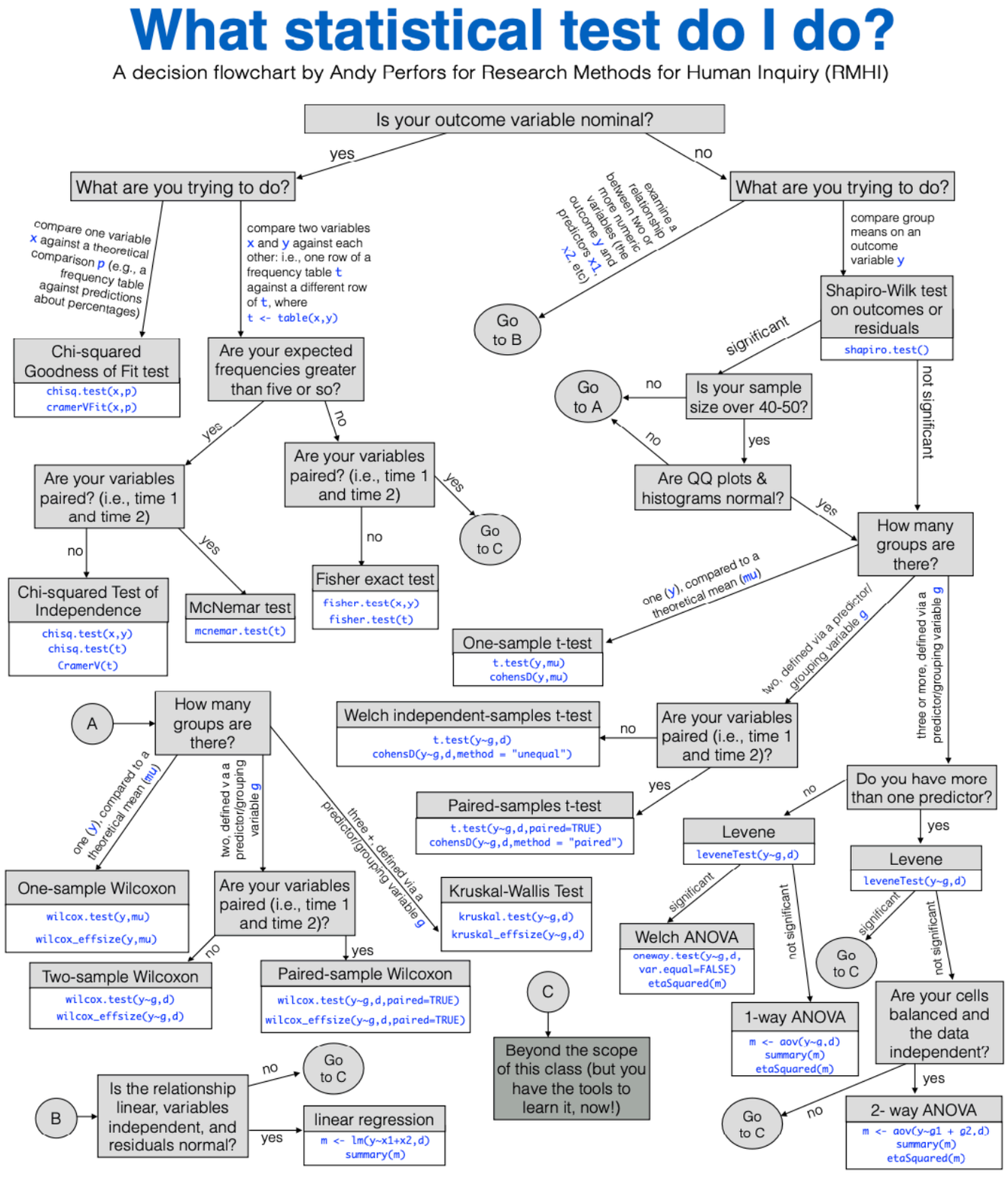
You glossed over one of the hardest parts: figuring out which test to use

Boo, hiss.

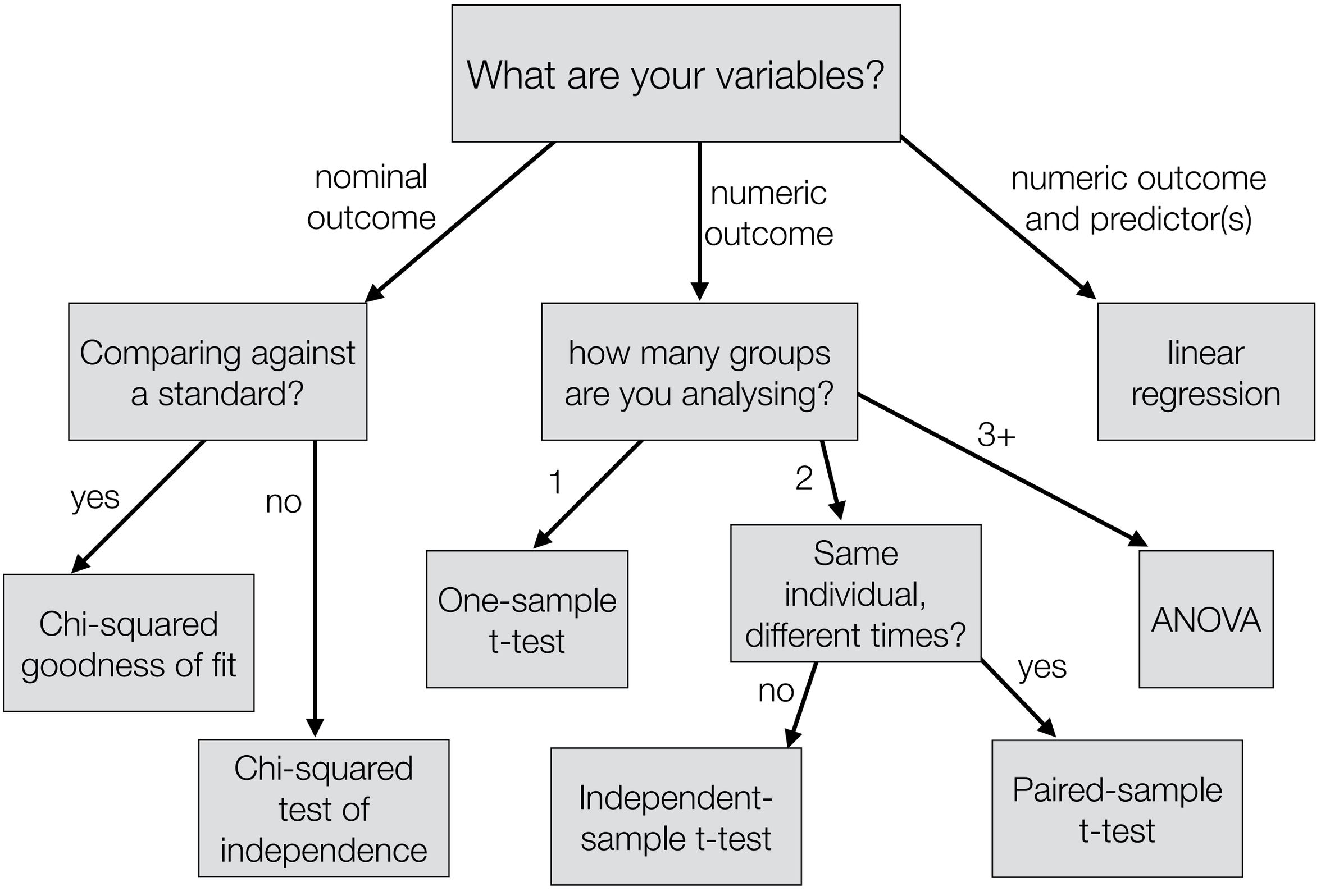


What test do I use?

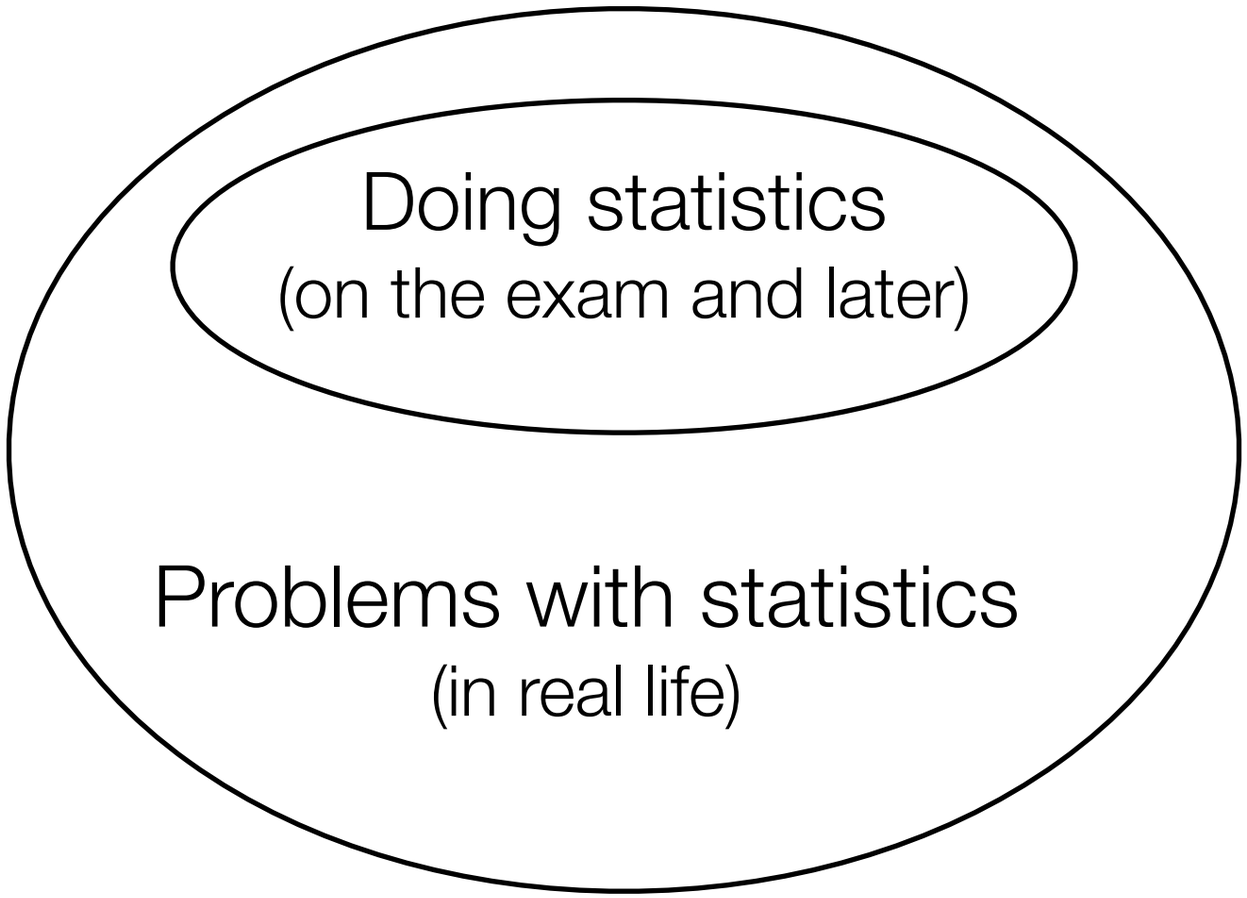
Check out the full flowchart on the LMS



What test should you use?

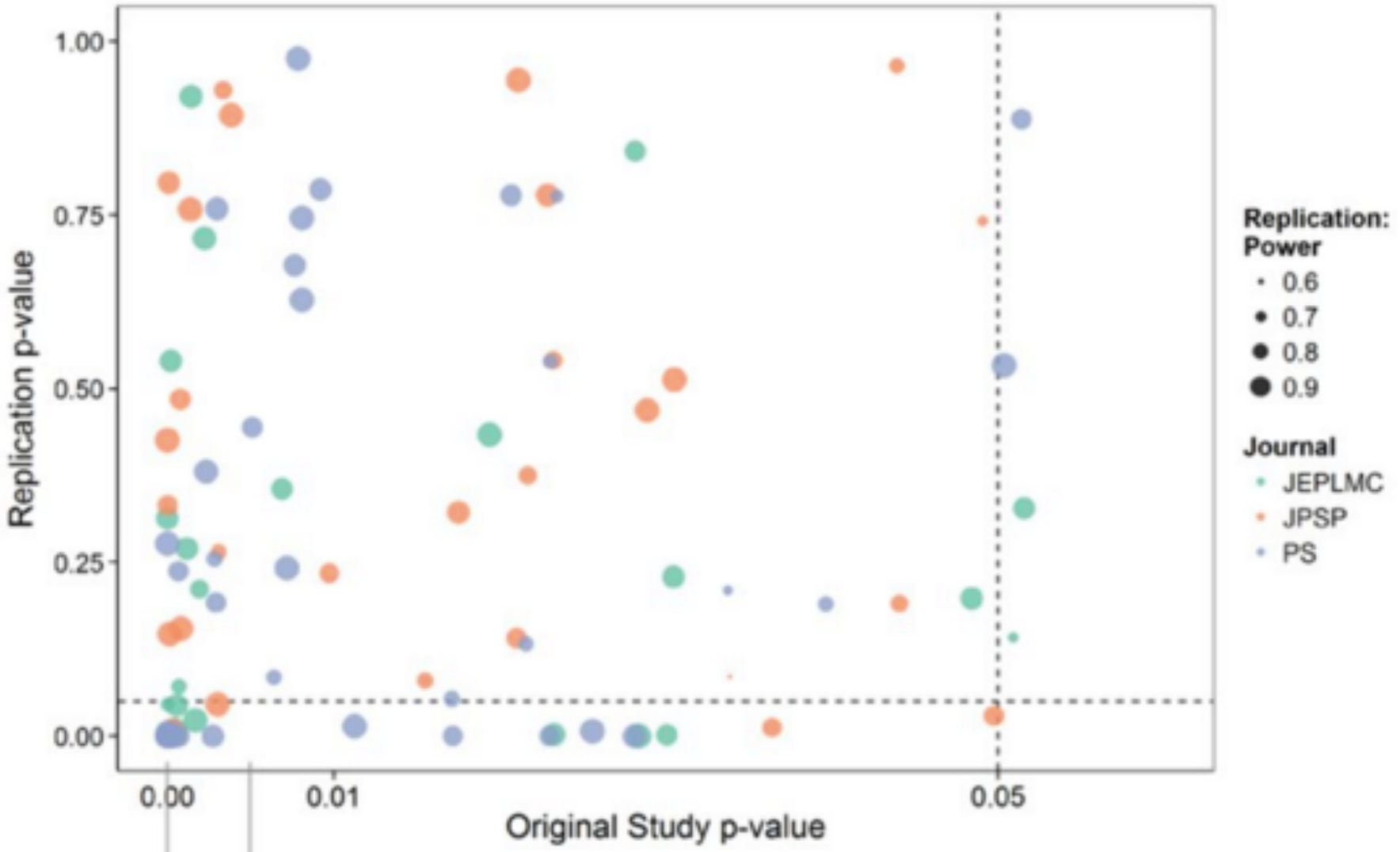


What's in the forest?



The reproducibility crisis

Conducted replications of 100 original studies using the same materials and procedure (as much as possible).

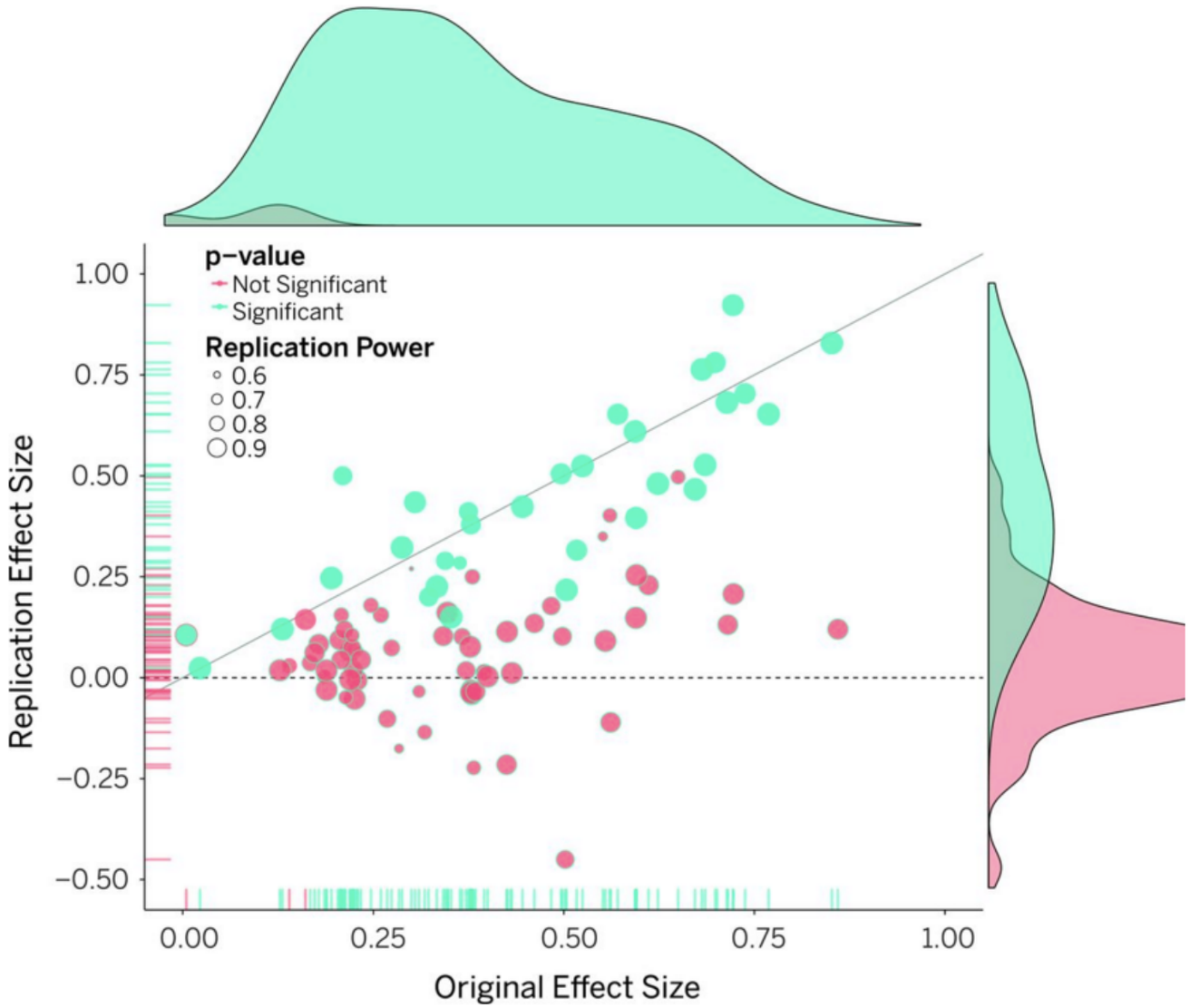


Many originally-significant findings did not replicate!

Nosek et al., (2015). Estimating the reproducibility of psychological science. Science

The reproducibility crisis

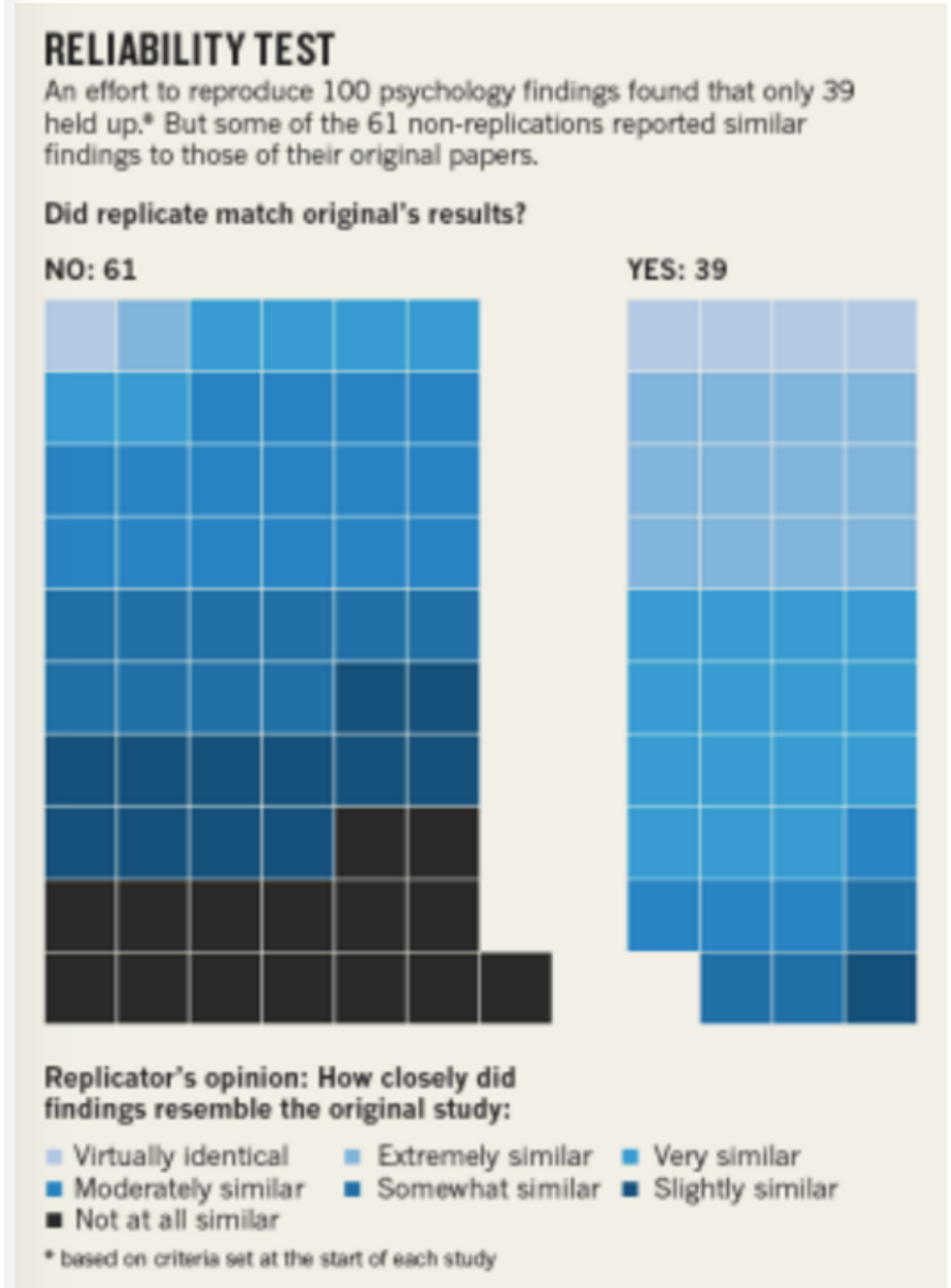
Conducted replications of 100 original studies using the same materials and procedure (as much as possible).



Not just p-values: effect sizes were smaller this time too!

Nosek et al., (2015). Estimating the reproducibility of psychological science. Science

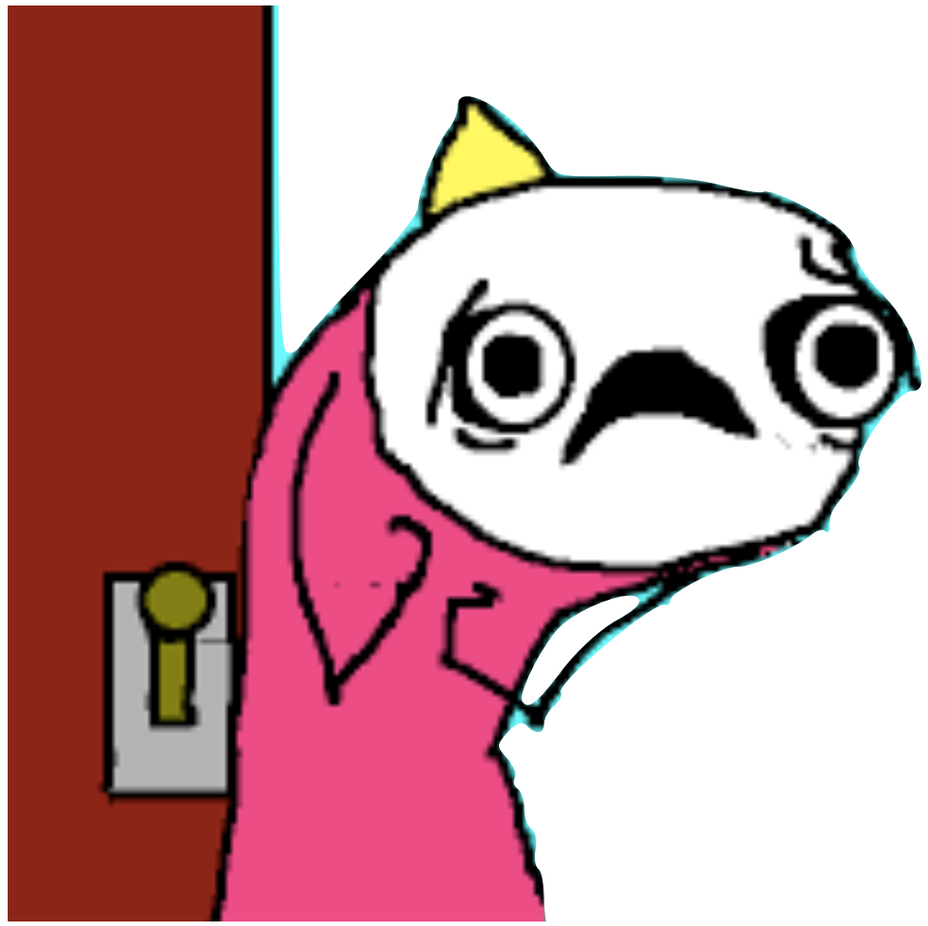
The reproducibility crisis



Qualitatively: only 39% showed "similar" findings

The reproducibility crisis

Uh oh… what's going on?



The reproducibility issue

1. different samples of people behave differently (this is genuinely interesting)

Lowest replication rate was in social psychology (26%)
Highest was in cognitive psychology (51%)

probably because people from different locations genuinely vary more in social context and assumptions

The reproducibility issue

2. publication bias: the file-drawer effect

journals / conferences reject if findings aren't significant findings

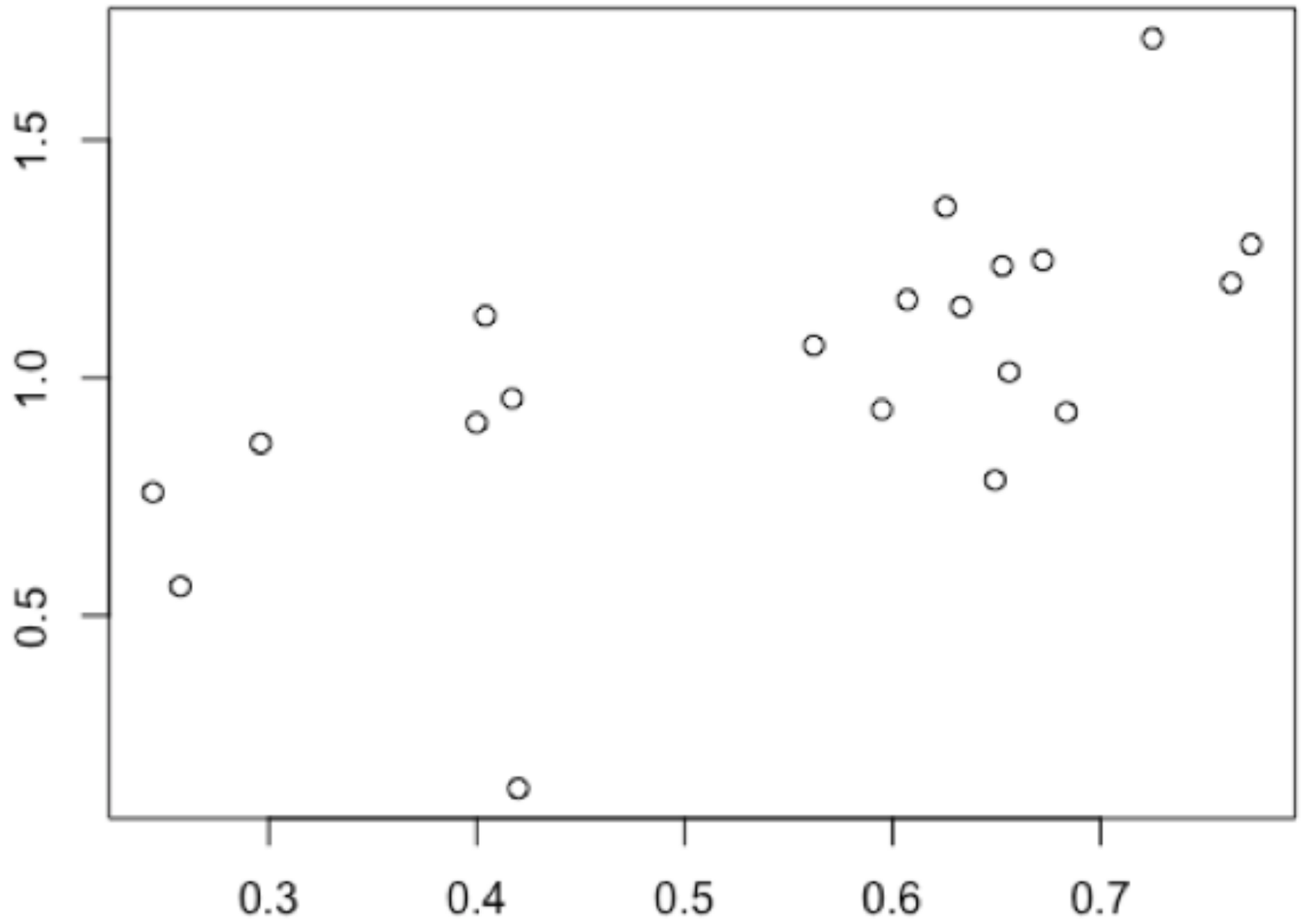
people don't even try to publish if findings aren't significant

leads to biased sampling of outcomes!

The reproducibility issue

3. p-hacking

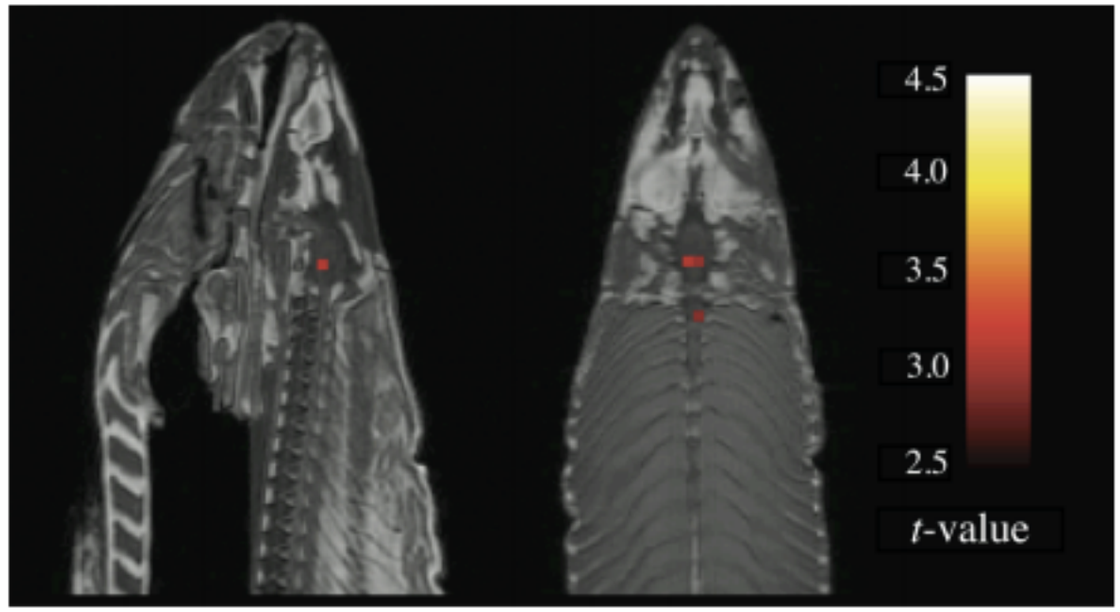
(a) removing outliers in an unprincipled manner



The reproducibility issue

3. p-hacking

(b) data mining: looking at lots of comparisons or tests until you find something, *anything*



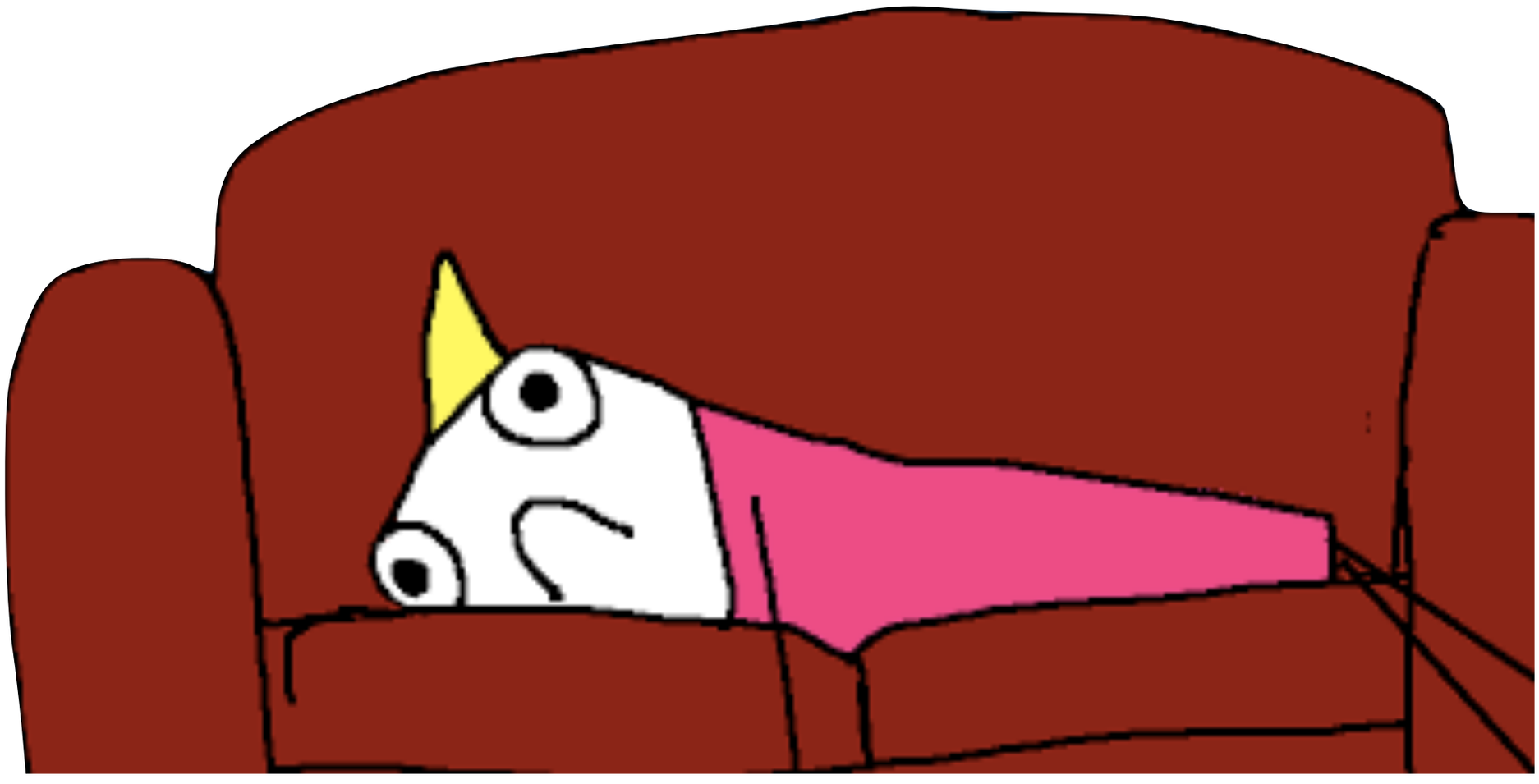
A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

Identical *t*-contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds ($p = 0.25$).

The reproducibility issue

does this mean all is lost?



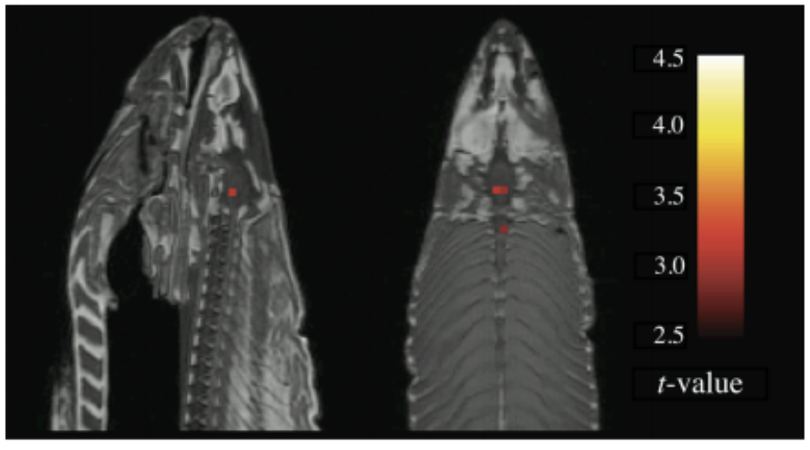
The reproducibility issue

All is not lost!

file-drawer effect

Journal of Negative Results
pre-registering studies
More awareness

p-hacking



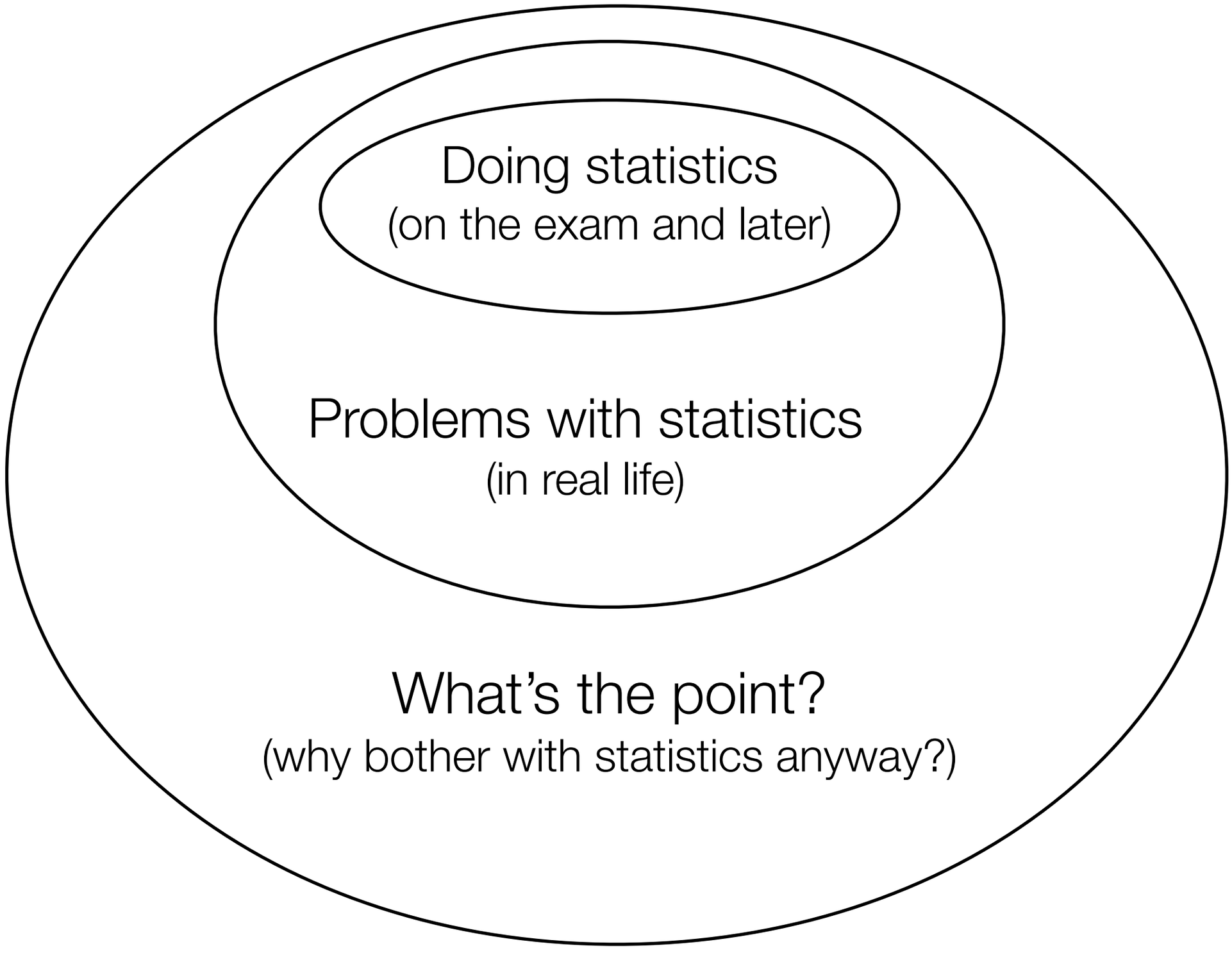
publication of entire datasets
awareness of problems
Bayesian analyses

The reproducibility issue

ultimately we have to remember:

> statistics is *helpful*, and necessary. but it's not a panacea. ultimately any tool is only as good as the person who wields it

What's in the forest?



Why do we learn statistics?
Is it because there will be an exam??

Why do we even do science? Why not do psychology by thinking *really hard* about ourselves? Why collect data at all?

we want to understand the world, and ourselves

science is a drunkard looking for his keys under a light post: not because they are there, but because that's where the light is

not irrational: if they are there, he'll find them. if not, he wouldn't have found them anyway.

If we do psychology by just *thinking* about ourselves, if there's no objective data or reasoning that we can call on to arbitrate or judge which is more "correct"…

then anything goes

We use statistics for a similar reason. Why not just get lots of data and then look at it? Why do “significance testing” at all, especially since it’s not infallible and can be misused?

That way lies madness! Without statistics, it’s too easy to see what we want to see in the data.

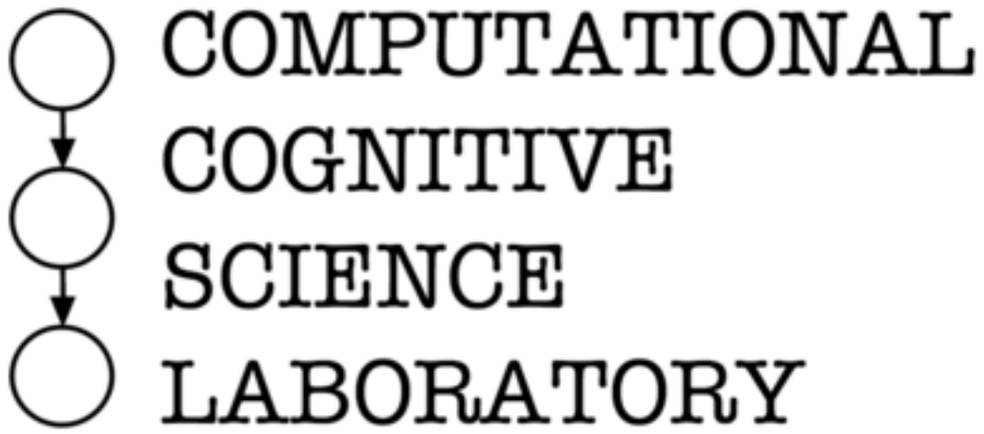
Data mining becomes a huge problem

It’s not a coincidence that psychology didn’t become a science — and we didn’t start actually making progress — until statistics was developed

So, no place for exploratory studies? Or non-quantitative stuff?

There is a place! It is just that they are not the only thing. They are mostly *the beginning.* To get to the point that we know what we know, and we agree on what we know, we need statistics.

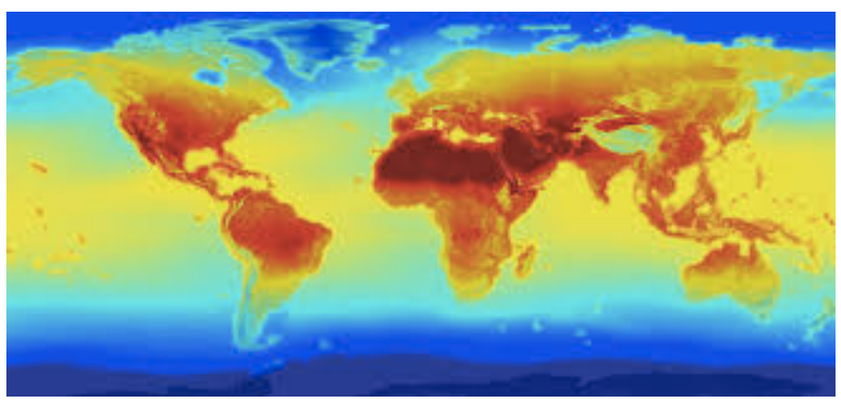
Note: we can use mathematical techniques to study how *humans* know what they know, too. In my lab we try to explain human reasoning using statistical models, and it's surprisingly effective

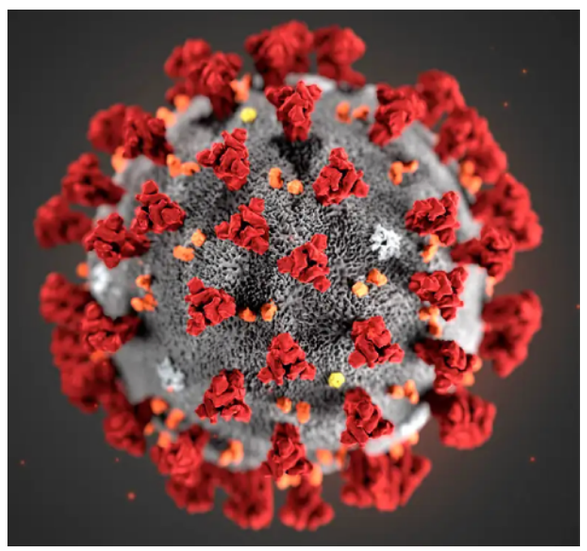


If you like this, check out PSYC30023: Computational Behavioural Science (Sem2)

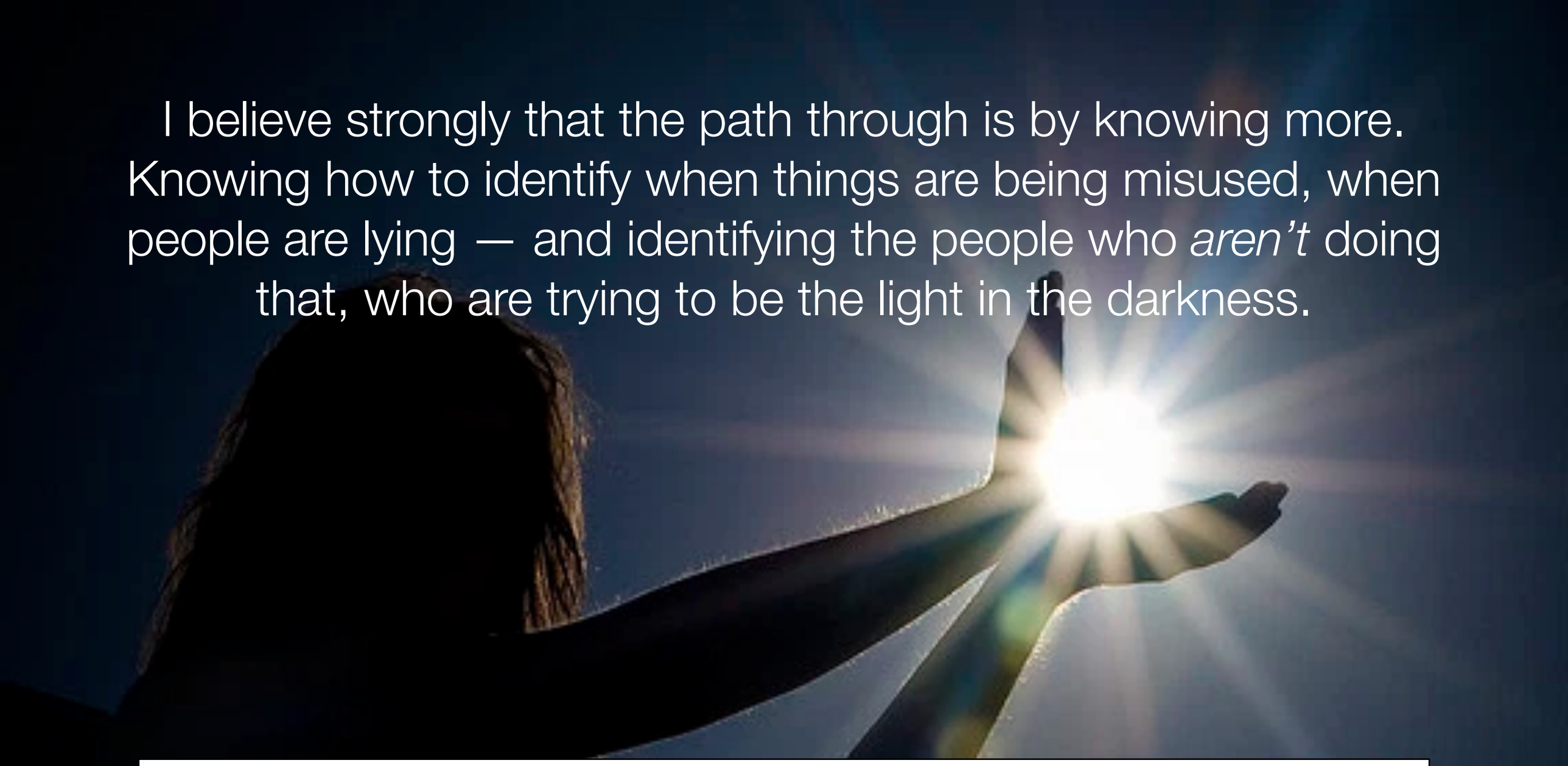
Finally, and personally.

This day and age we have a lot of (justified) reasons for cynicism. Institutions fail us. People game systems. Misinformation is rampant. Even science can be misused.





I believe strongly that the path through is by knowing more. Knowing how to identify when things are being misused, when people are lying — and identifying the people who *aren't* doing that, who are trying to be the light in the darkness.



Statistics and research methods is ultimately about that: about having a question, knowing how to find an answer, and having tools for determining whether that answer is "real" and whether it matters. It's the foundation of what makes psychology a science and what makes us able to learn about ourselves in a scientific way.


Good luck. It’s been a pleasure.