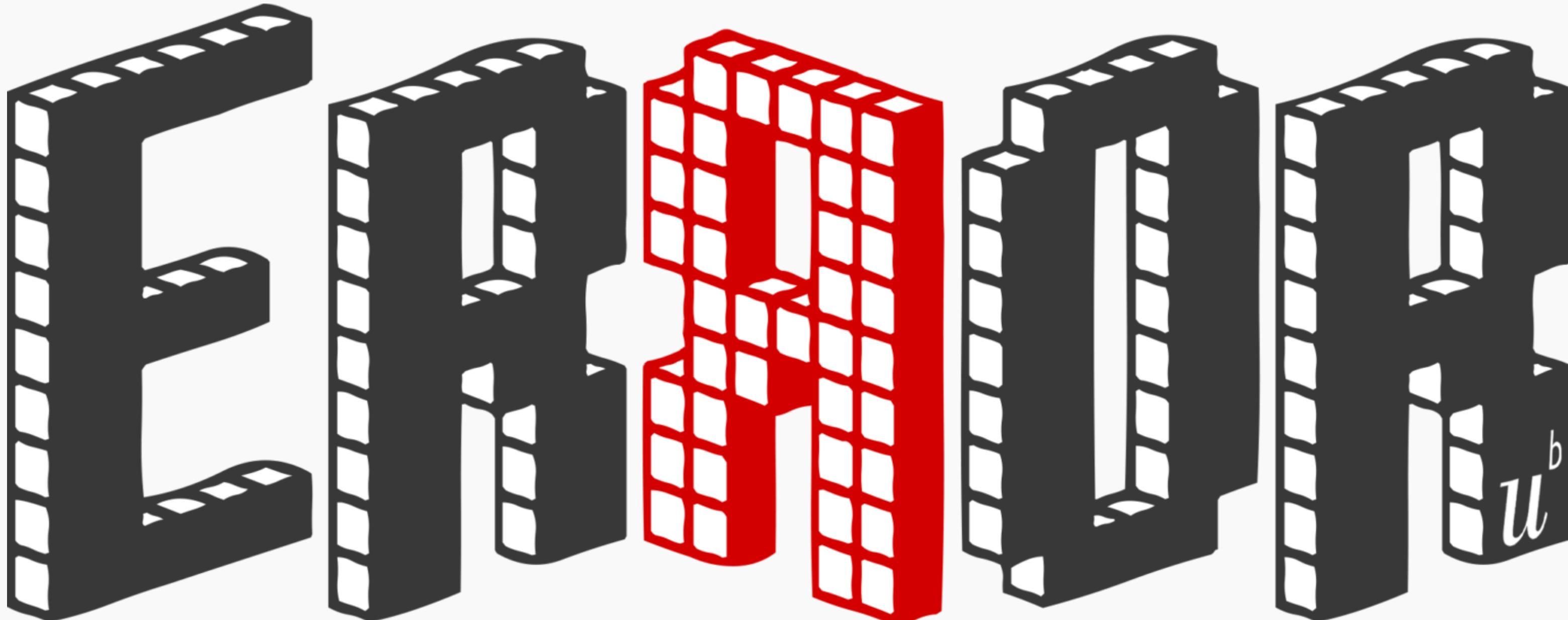


xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx



xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

A three-pronged effort to improve  
post-publication critique & error detection

Ian Hussey

## PRI~~I~~CIPAL INVESTIGATORS



**Malte Elson**

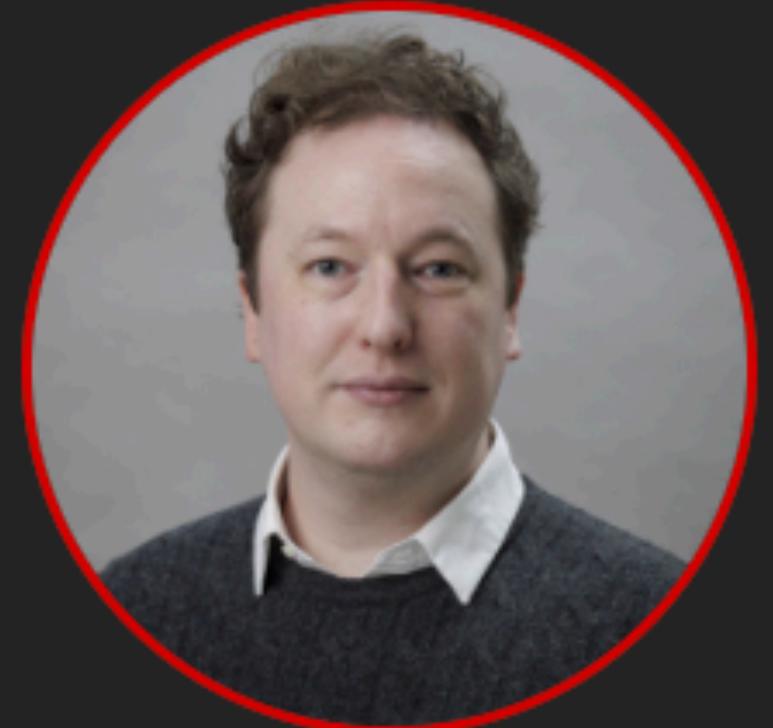
*University of Bern*



**Ruben Arslan**

*University of Leipzig*

## CHIEF RECOMMENDER RECOMMENDER



**Ian Hussey**

*University of Bern*



**Jamie Cummins**

*University of Bern*

# ADVISORY BOARD



**Dorothy Bishop**  
*University of Oxford*



**Nick Brown**  
*Linnaeus University*



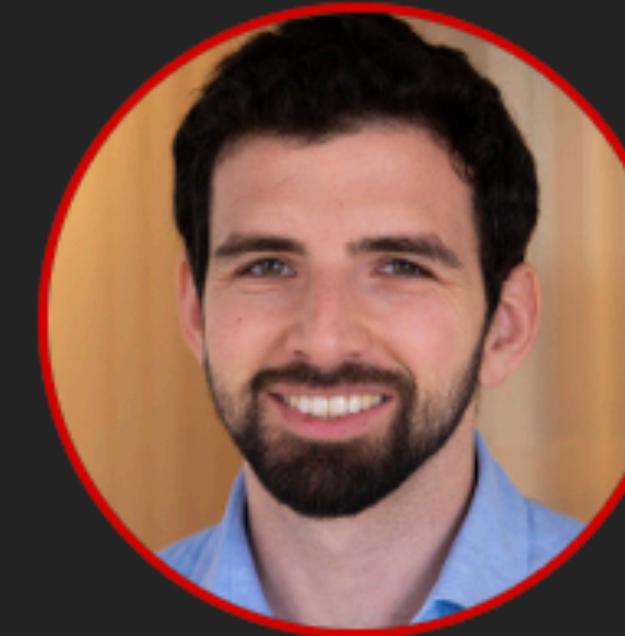
**Matthias Egger**  
*University of Bern*  
*University of Cape Town*  
*University of Bristol*



**Julia Rohrer**  
*University of Leipzig*



**Anne Scheel**  
*Utrecht University*



**Leo Tiokhin**  
*Eindhoven University of Technology*



**Richard McElreath**  
*Max Planck Institute for Evolutionary Anthropology*



**Brian Nosek**  
*Center for Open Science*



**Michèle Nuijten**  
*Tilburg University*



**Simine Vazire**  
*University of Melbourne*



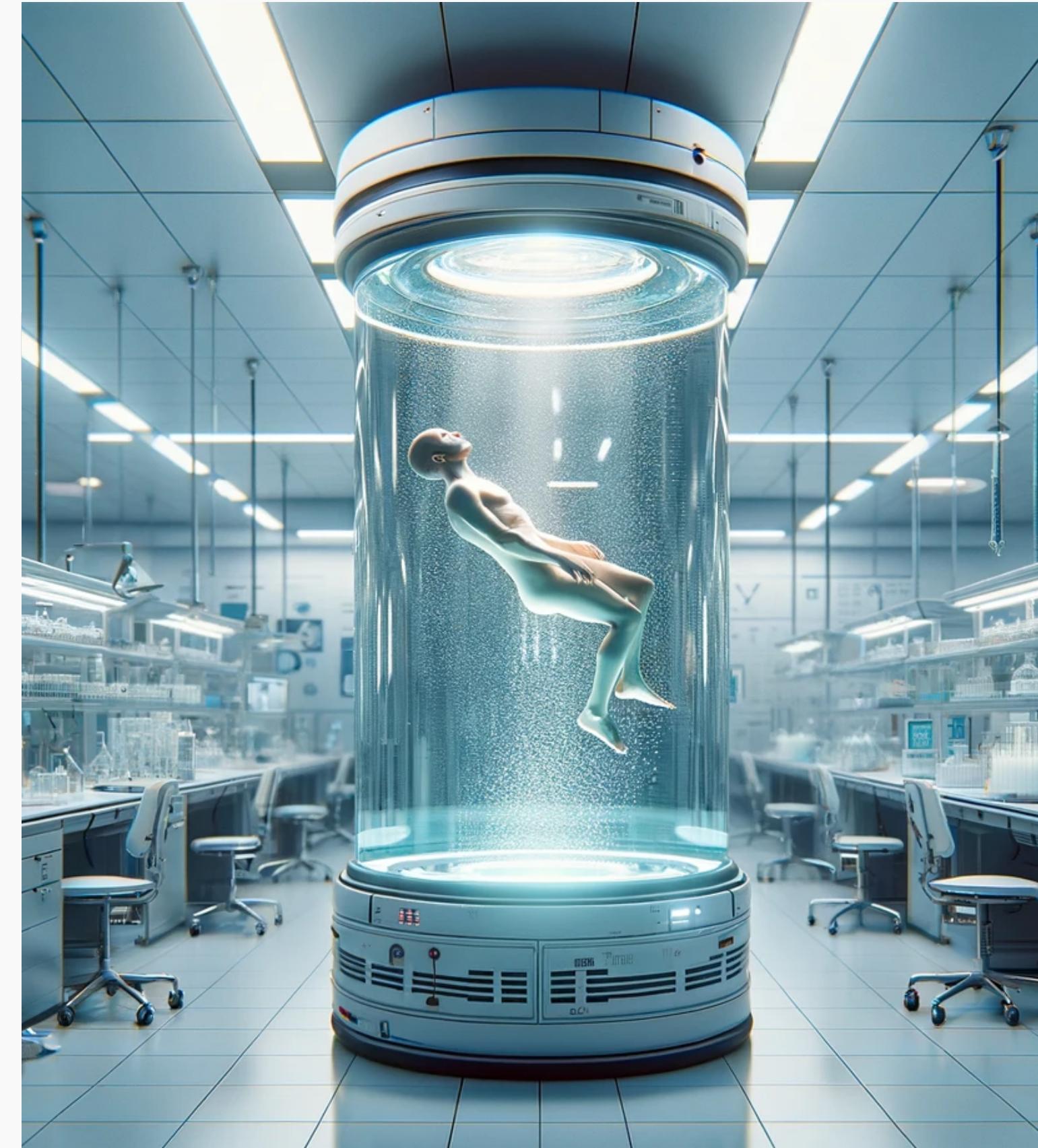
## We ask too much of peer-review

- X One-shot, little redundancy  
Despite imperfect reliability & evolving knowledge
- X Generally non-technical
- X Surprising little **error** checking  
for an industry built on discovering truth

# **Who can help carry this burden?**

**Need for parallel systems  
of scientific verification**

# Who can help carry this burden?



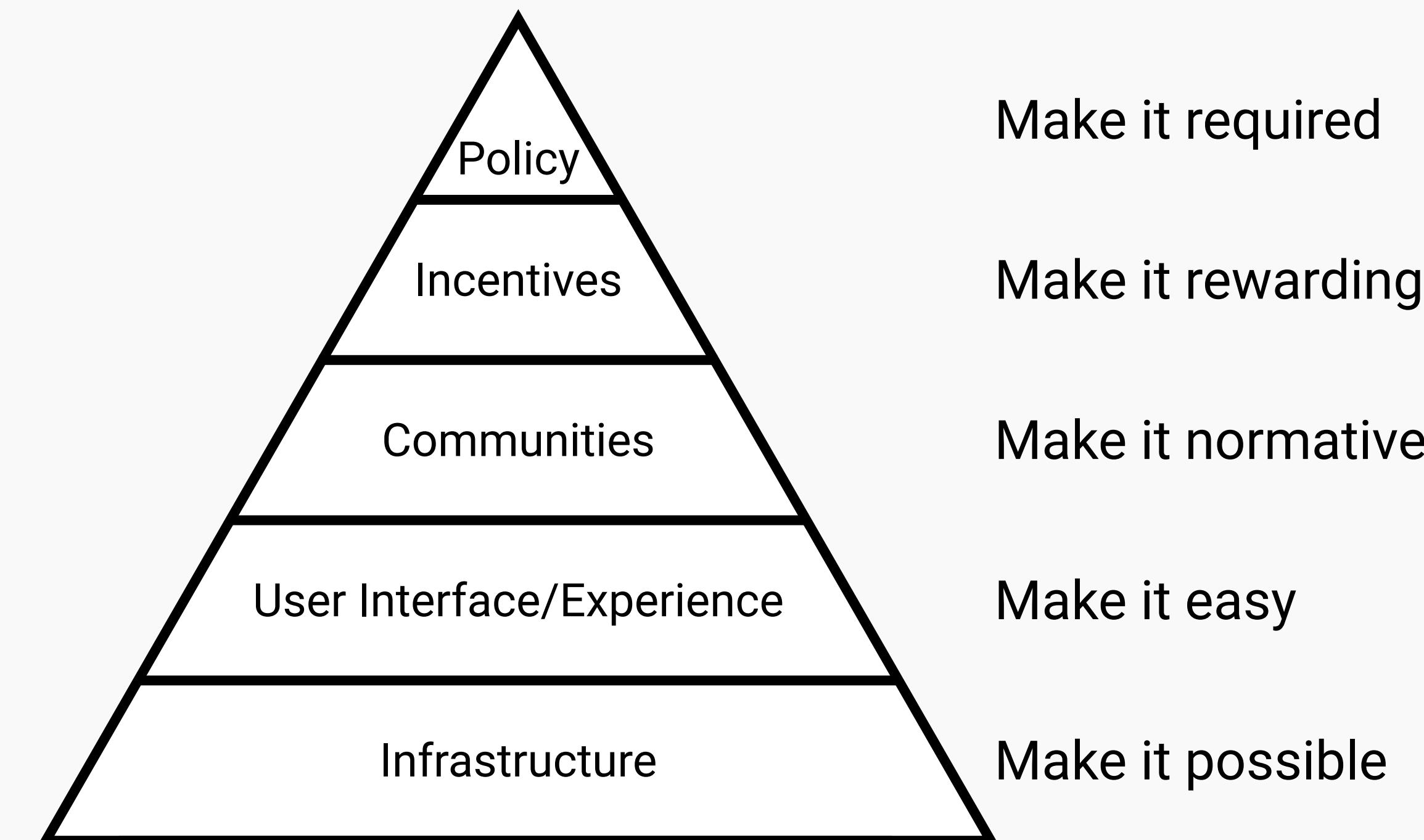
A clone-army of Nick Browns, Dorothy Bishops, & Elisabeth Bik

# Who can help carry this burden?

“A few outsiders and weirdos saw the giant lie at the heart of [science], and they saw it by doing something the rest never thought to do:  
They looked.”

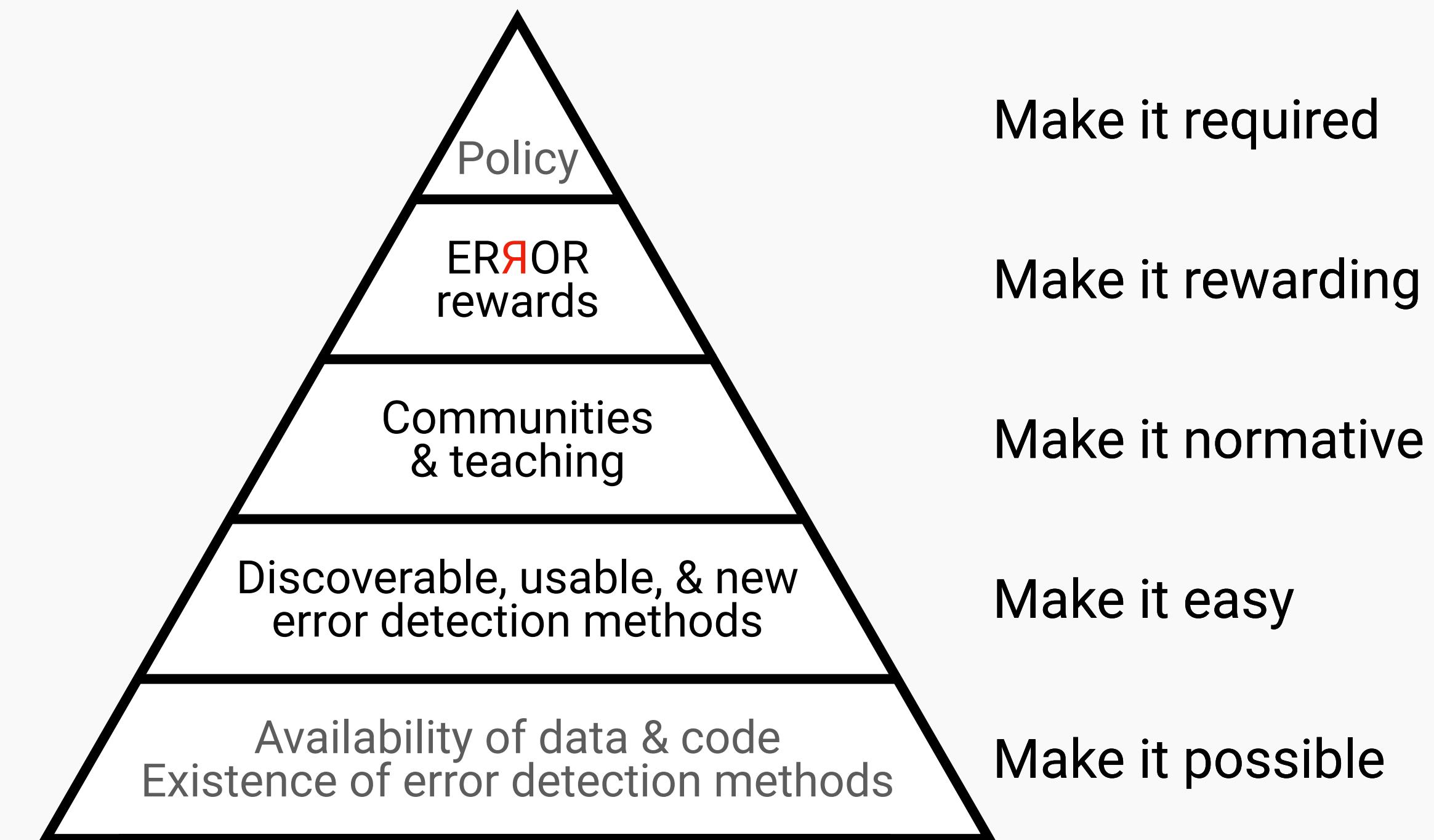


# Centre for Open Science's strategy to increase preregistration & data sharing



Nosek (2019)

# Our strategy to increase error checking, reporting, & correction





### **Make it easy**

- Increasing the discoverability & usability of error detection methods
- New error detection methods

### **Make it normative**

A master's degree course in error detection

### **Make it rewarding**

ERROR: A bug bounty program for science

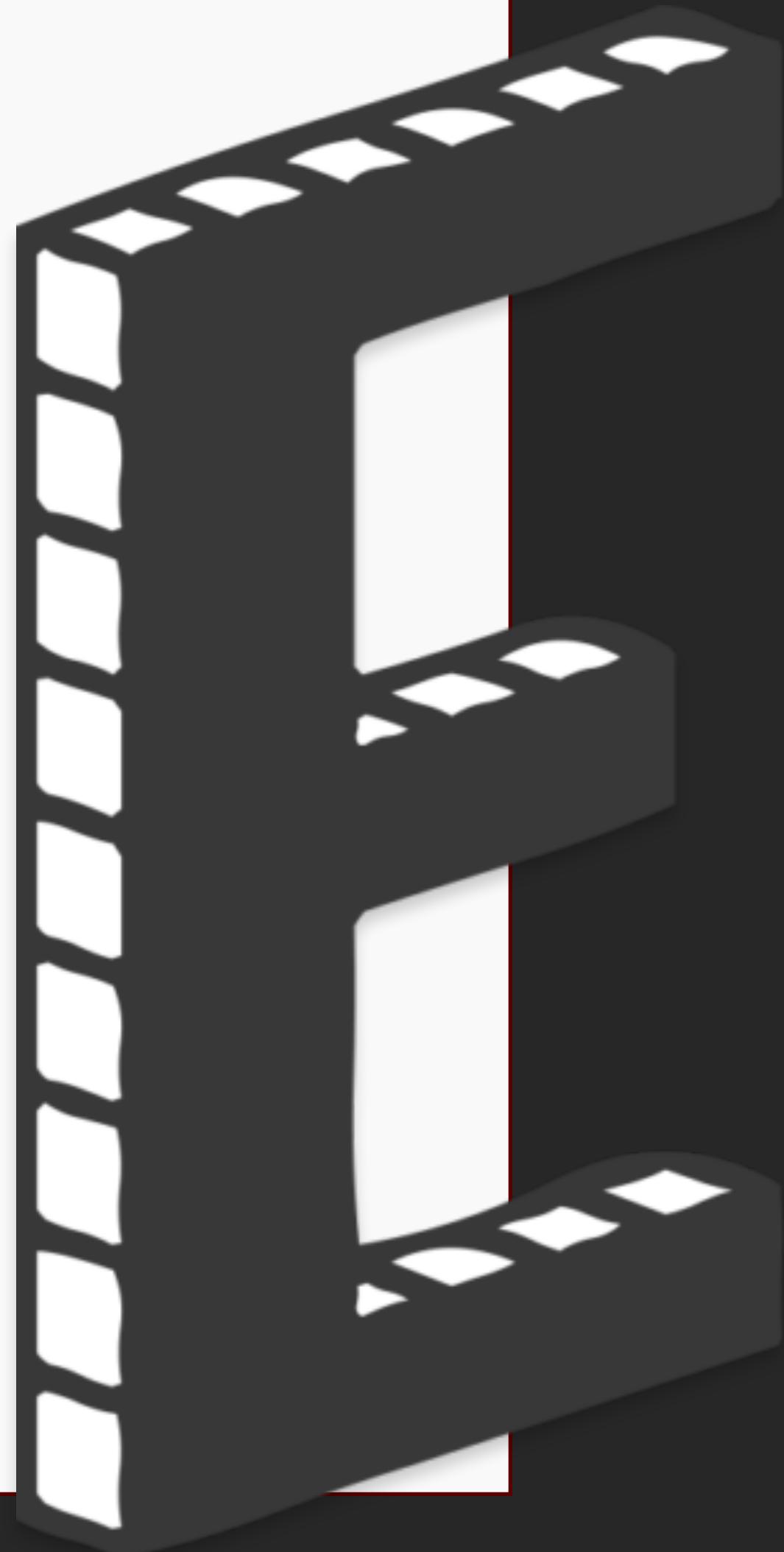
# Increasing the discoverability & usability of error detection methods

xxxxxxxxxxxxxxxxxxxxxxxxxxxx

01.

## **‘The scientific article’ is a well-defined genre**

Yet “no handbooks exist for writing replies”  
(Hyman, 1995)





# An ecosystem of error checking methods

Methods to scrutinise the **article, data, code**, or the **relationships among them**

**Smoke** vs **fire** methods

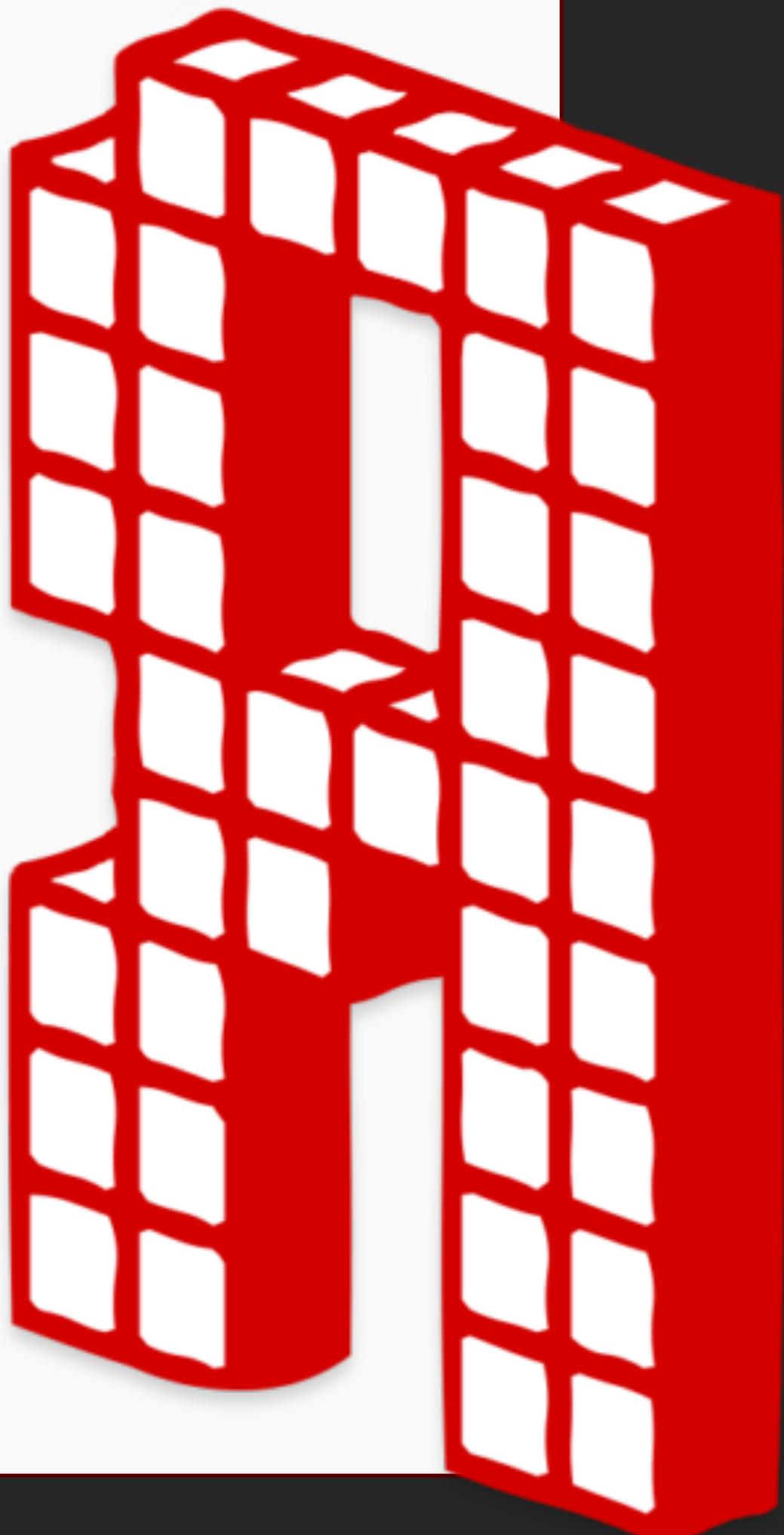
*Often*

- Very simple math
- Exploiting overlooked details & repetition
- Established principles redeployed for error detection

## Repetition & recalculation methods

- StatCheck

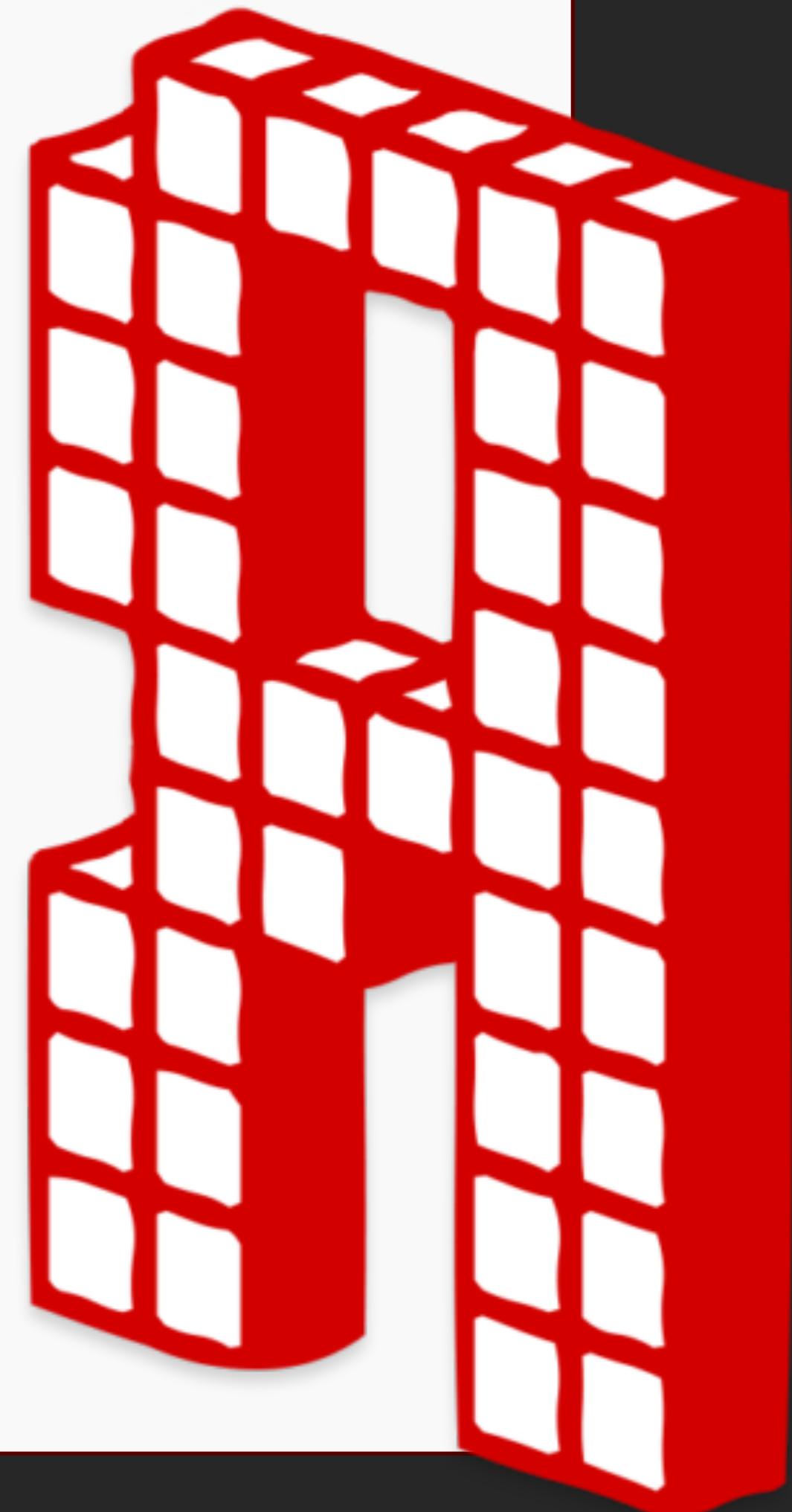
“ $t(58) = 1.46, p = .03$ ” →  $p = .19$

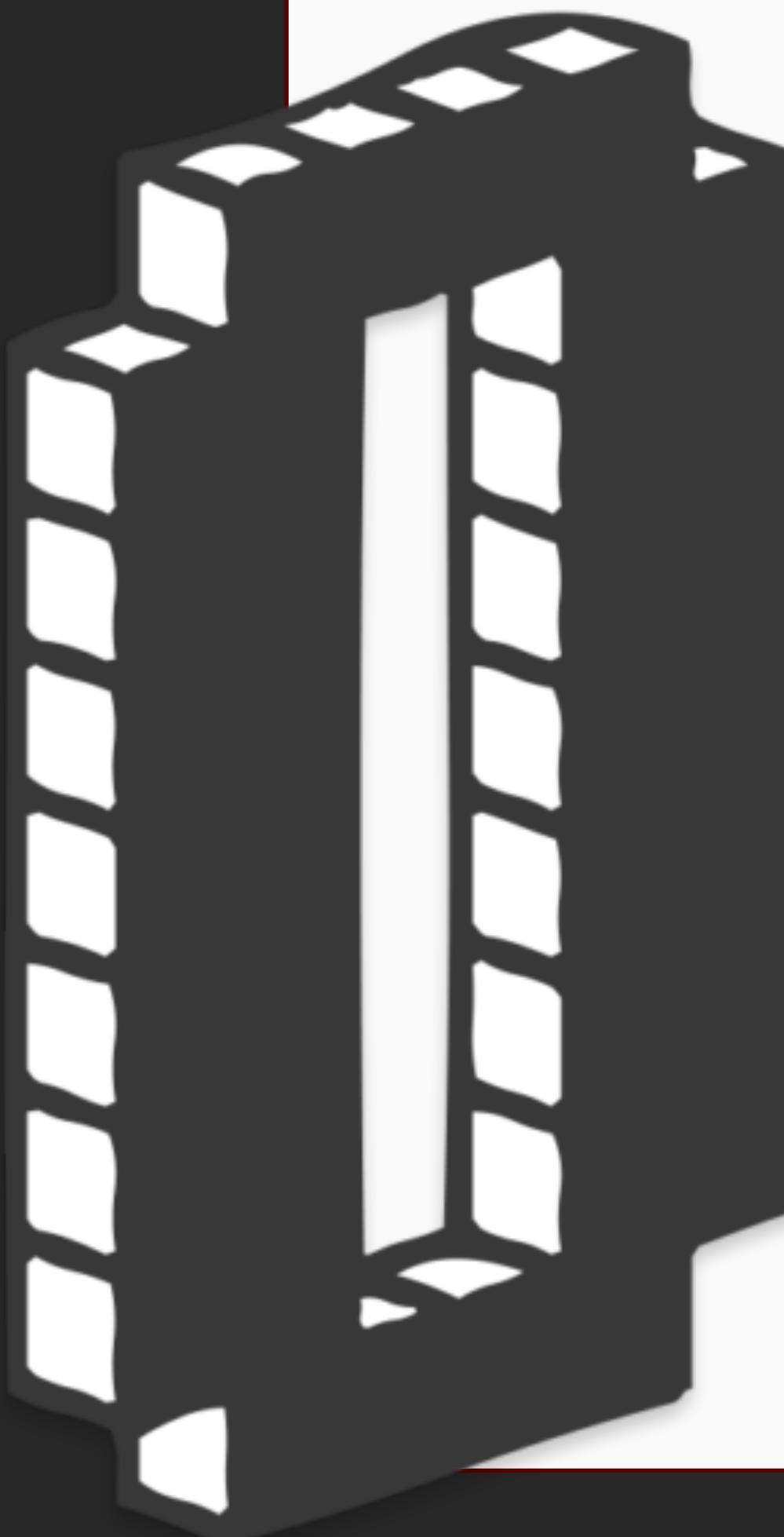


## Repetition & recalculation methods

- StatCheck
- GRIM / GRIMMER / GRIMMEST
- SPRITE
- RIVETS
- DEBIT
- Reproduce effect-sizes & test statistics from summary statistics
- ...

*R packages, vignettes, how-to guides, blogs, & examples needed!*





## ‘Shoe-Leather’ error checks

- Full sample - exclusions = analytic sample size?
- Point estimates inside the confidence intervals?

## Expert-knowledge based checks

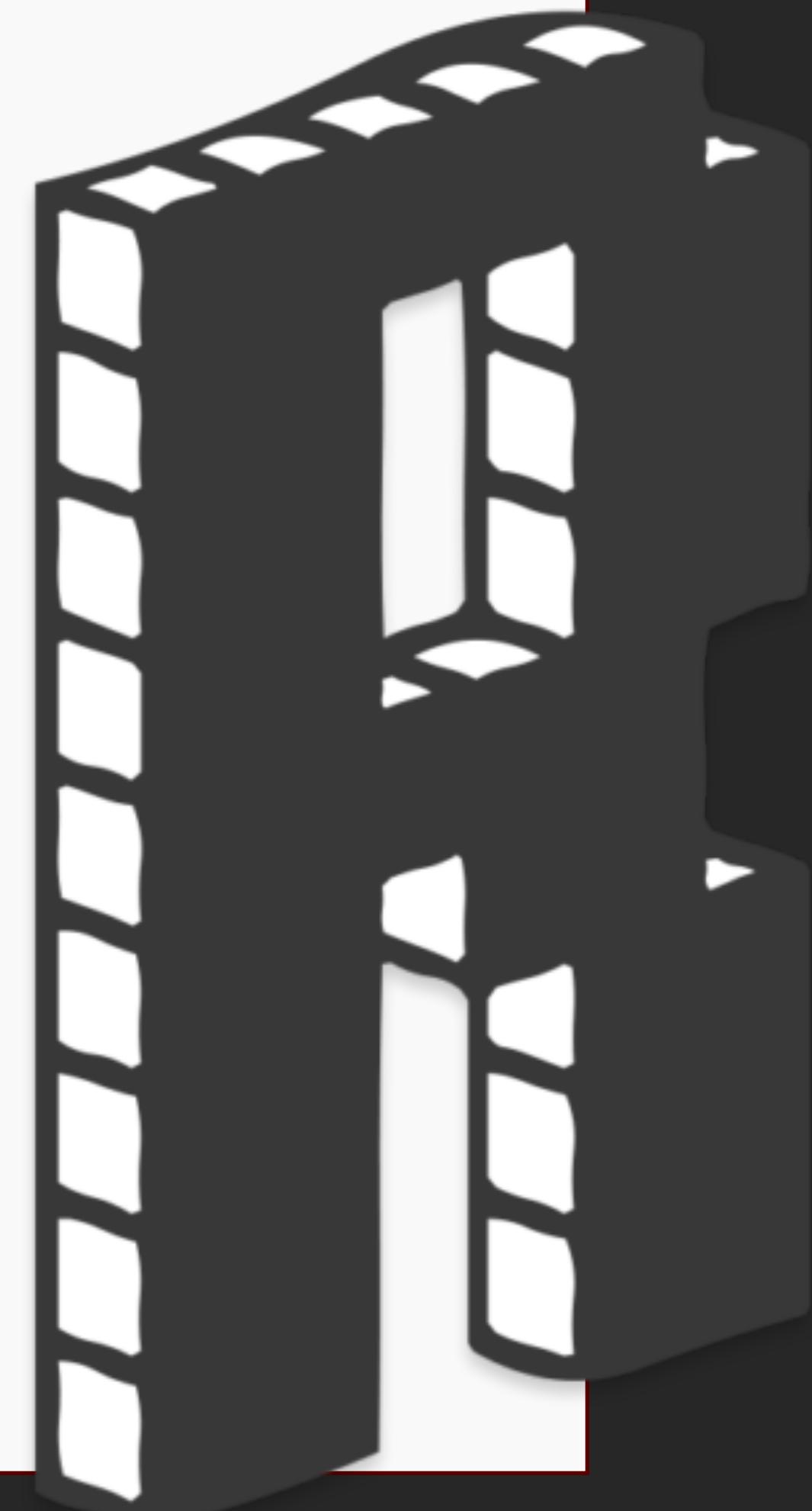
- Beck Depression Inventory II: “SD = 3.40” is too low

*Currently few training materials*

A foreseeable output of ERROR:

## An open handbook for error detection

- What do you have?
  - Article, measures, data, code, the original authors' good-will?
- How can you get more?
  - Extracting additional data (plots, tables, summary statistics, asking)
- What could you check and how?
- How to prioritise?
  - What are the authors' claims? (Scheel, 2022)
- How to ensure reproducibility of checks?
- How to effectively communicate errors





**Where do you want to look?**

In the paper only

**What resources do you have available?**

- The paper
- The code
- The raw data
- The first author's home address

**What do you care about?**

- Statistical errors
- Conceptual errors
- Reporting errors
- Integrity errors
- Failing-to-cite-me errors

**How much time do you have?**

< 2 hours

**Error-checking tools for you:**

- GRIM
- GRIMMER
- SPRITE
- TIDES
- PORT
- STALT

# New error detection methods

xxxxxxxxxxxxxxxxxxxxxxxxxxxx

02.

## What's wrong with these summary stats?

(Can't use GRIM/MER because N is too high)

“145 participants responded using a 5-item Likert scale  
(response options 1-7, sum scored),  $M = 38.1$ ,  $SD = 4.2$ ”

“200 participants responded using a single-item Likert scale  
(response options 1-7),  $M = 2.1$ ,  $SD = 2.4$ ”

## What's wrong with these summary stats?

(Can't use GRIM/MER because N is too high)

“145 participants responded using a 5-item Likert scale (response options 1-7, sum scored),  $M = 38.1$ ,  $SD = 4.2$ ”

→ Mean must be within [5, 35]

“200 participants responded using a single-item Likert scale (response options 1-7),  $M = 2.1$ ,  $SD = 2.4$ ”

→ SD must be within [0.31, 2.31]



## Truncation-Induced Dependency in Summary Statistics

A method for checking the compatibility of reported means, SDs, and Ns given the min and max of the scale

Mean:

2.1

Standard Deviation:

2.4

Sample Size:

30

Minimum Value:

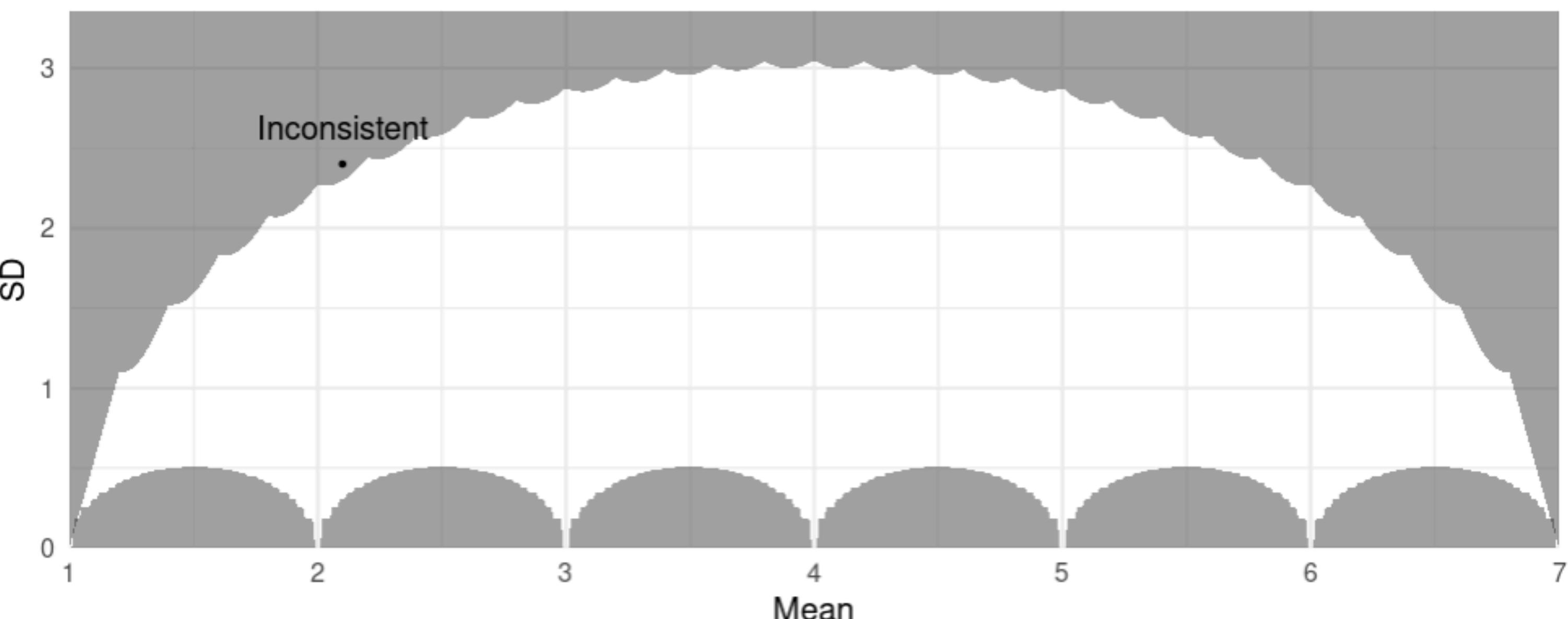
1

Maximum Value:

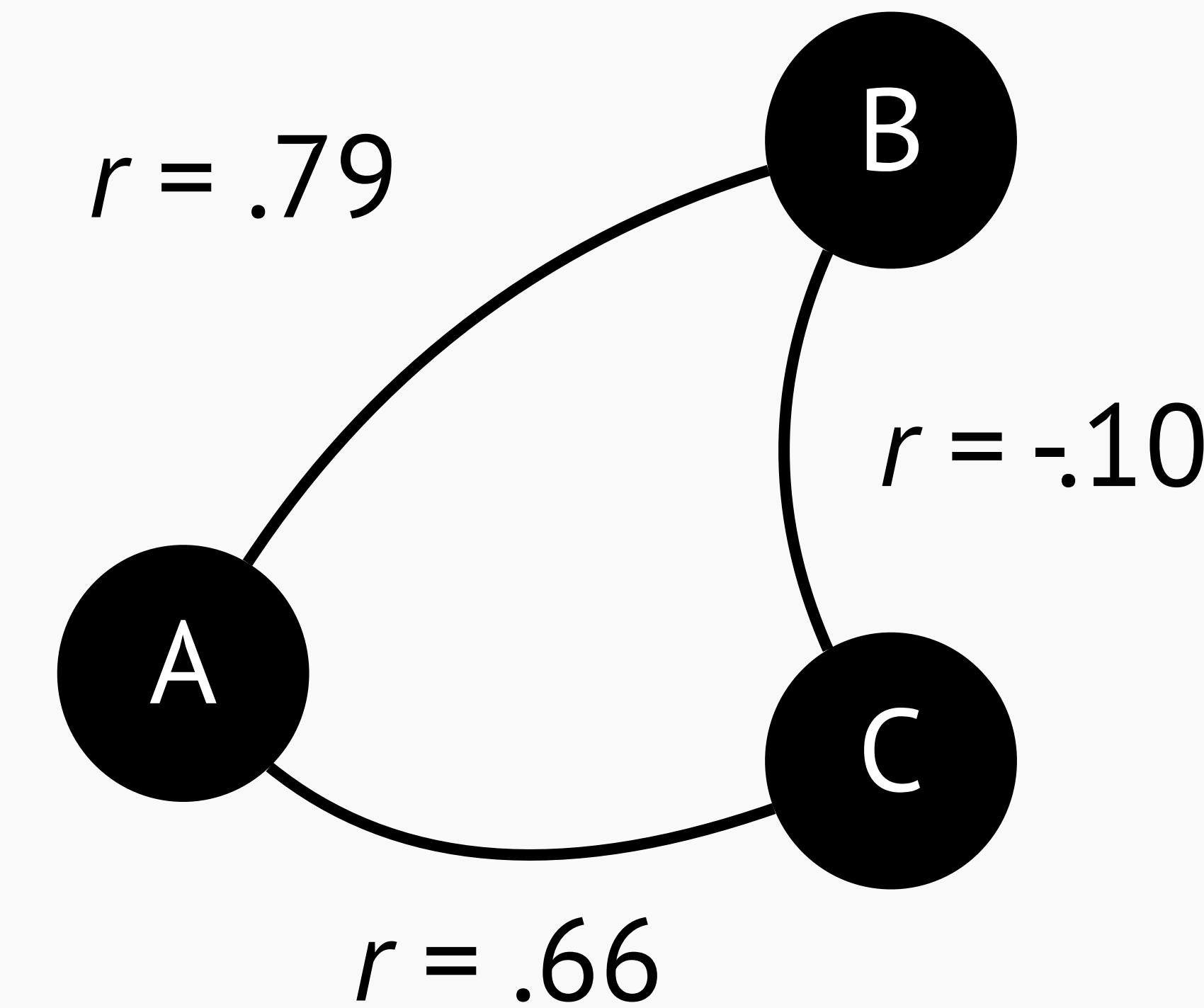
7

Scale is interval:

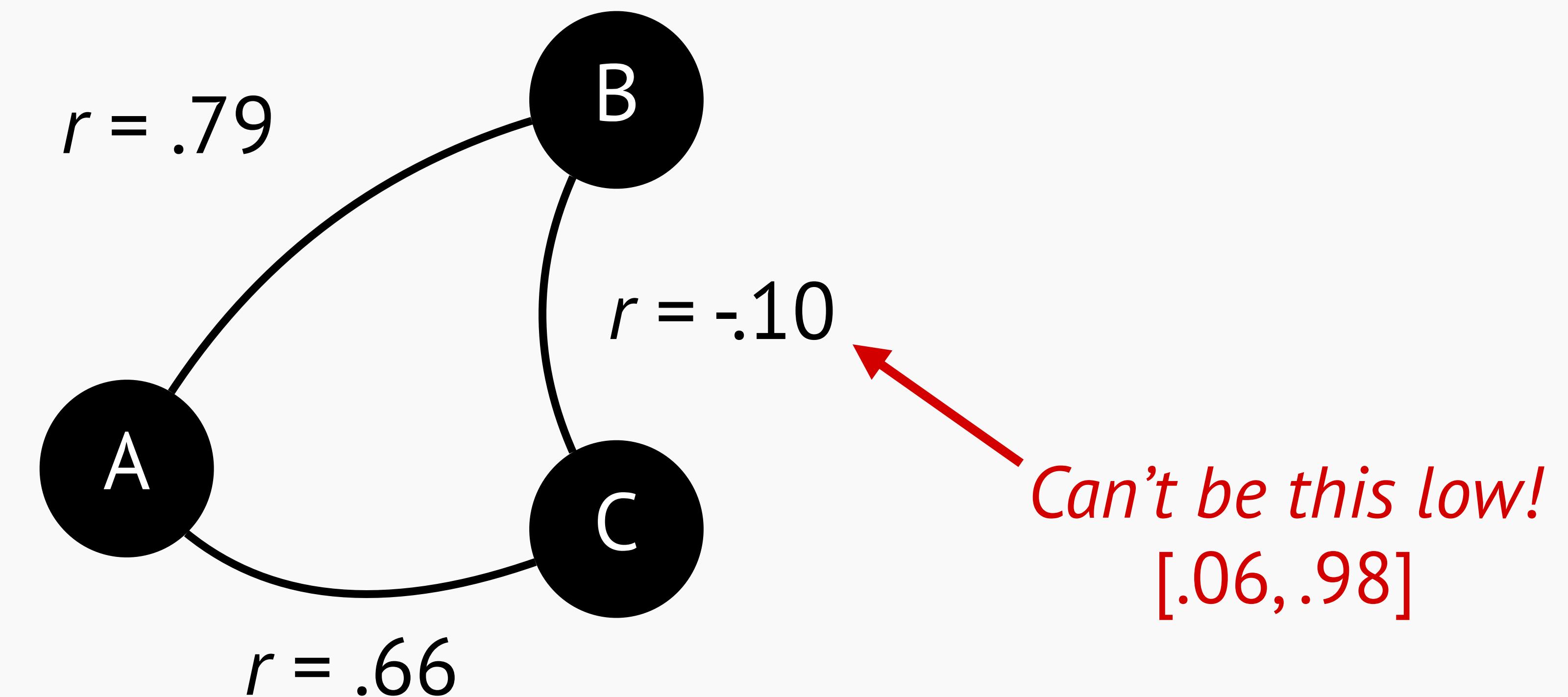
Yes, also calculate minimum SD



## What's wrong with these correlations?



## What's wrong with these correlations?



02.

# Does this correlation table contain errors?

Hard to spot by eye; easier with tools



# Positive-definiteness Of (Pearson's) $r$ Tables

# A method for assessing inconsistencies among reported correlations

 Download Example CSV

### Choose CSV or Excel File

Browse... example correlation matrix.csv

Upload complete

 Download Results

## Test Results

## Correlation Imputation

This table displays the reported correlations. The difference between each reported correlation and its imputed value is shown in square brackets. E.g., if imputation suggests that a given correlation may be 0.30 larger than the reported value, this is displayed as [+0.30]. The three correlations with the largest absolute differences are annotated \*\*\*, \*\*, and \* in descending order.



## Compare preregistrations with papers. Instantly.

RegCheck leverages large-language models and AI to automatically compare preregistration plans with scientific publications.

This lets researchers effortlessly identify if and how executed studies deviated from the initial plan.

[Try it now](#)

Jamie Cummins, [Psychology of Digitalisation](#), University of Bern, 2024

Cover template for [Bootstrap](#), by [@mdo](#).

# Compare preregistrations with papers. Instantly.

This app is an **alpha version**. It has not yet been extensively tested.

Due to token limits, Llama 3 will only work with short papers.

ChatGPT should work in most cases.

Processing files may take a minute or two.

Choose your model:

ChatGPT-4o

Does the paper have multiple experiments?

No

Preregistration:

Choose file

No file chosen

 or supply OSF link



Supported files types:  
.txt and .pdf

Paper:

Choose file

No file chosen

 or supply DOI

Compare

Study Feature	Information in Paper	Preregistered Protocol	Match
Final sample size	147 participants	150 participants	FALSE
Power Analysis Basis	95% power for medium effect size ( $f^2 = 0.15$ )	No power analysis for first analysis; 95% power for medium effect size ( $f^2 = 0.15$ ) in second analysis	TRUE
Sampling Strategy	Recruit 150 and exclude those with incomplete data until sufficient sample size is achieved	Recruit 150 and apply exclusion criteria, then add in batches of 10 until at least 150 participants meet criteria	TRUE
Within-subjects factor	Prime Valence: positive vs. negative	Prime Valence: positive vs. negative	TRUE
Dependent Variables	Influence-awareness ratings, Target stimulus evaluations	Evaluations within the AMP as pleasant or unpleasant	FALSE
Exclusion criteria	Incomplete data	Completion time < 3 minutes, Partial data on demographics/AMP	TRUE
IA-AMP Task Trials	10 practice trials, 120 critical trials	10 practice trials, 120 main trials	TRUE
Exploratory	Three exploratory	Five exploratory questions	FALSE

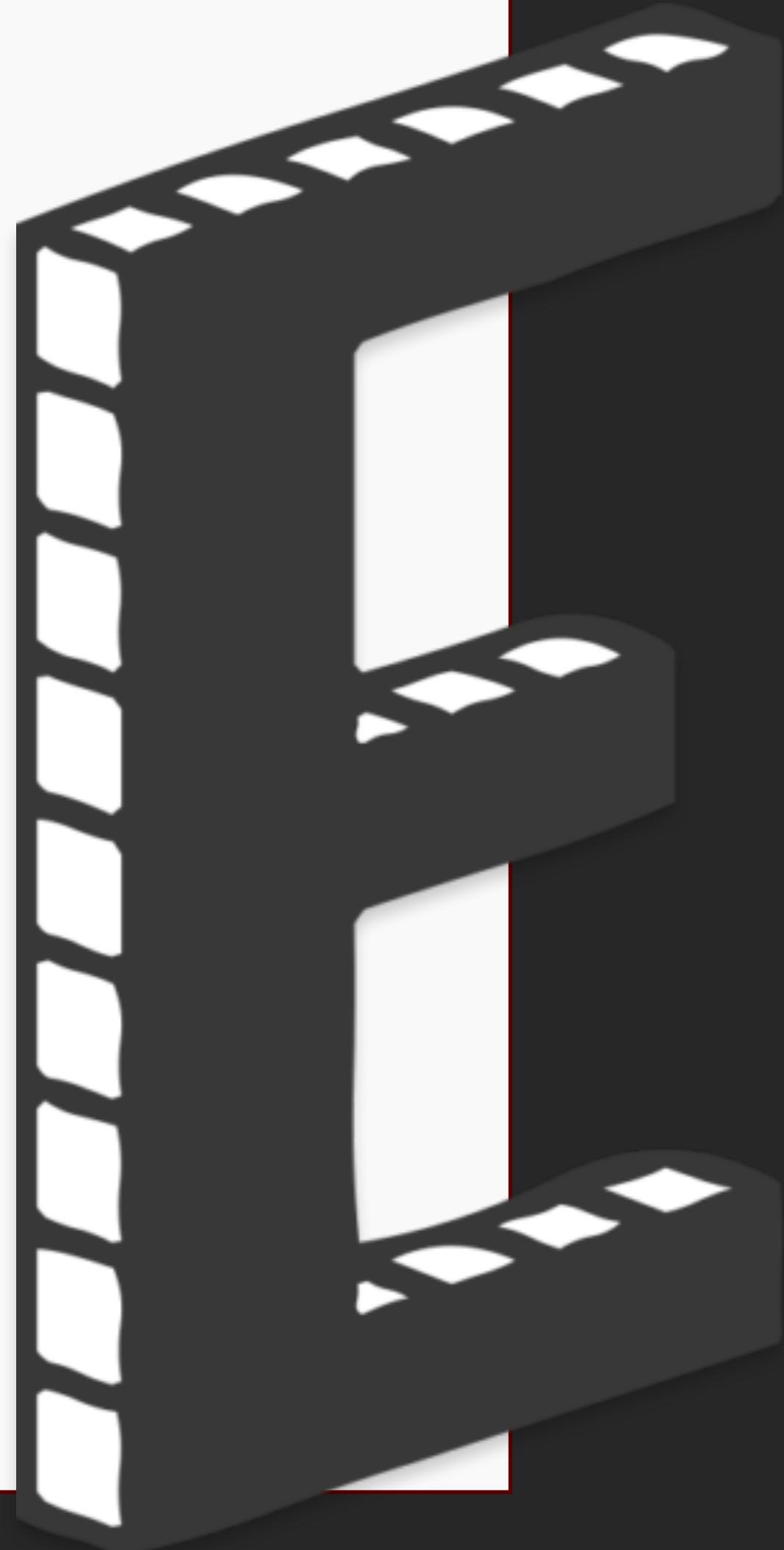
# A master's degree course in error detection

xxxxxxxxxxxxxxxxxxxxxxxxxxxx

03.

# **Conspicuous absence of error detection in the curriculum**

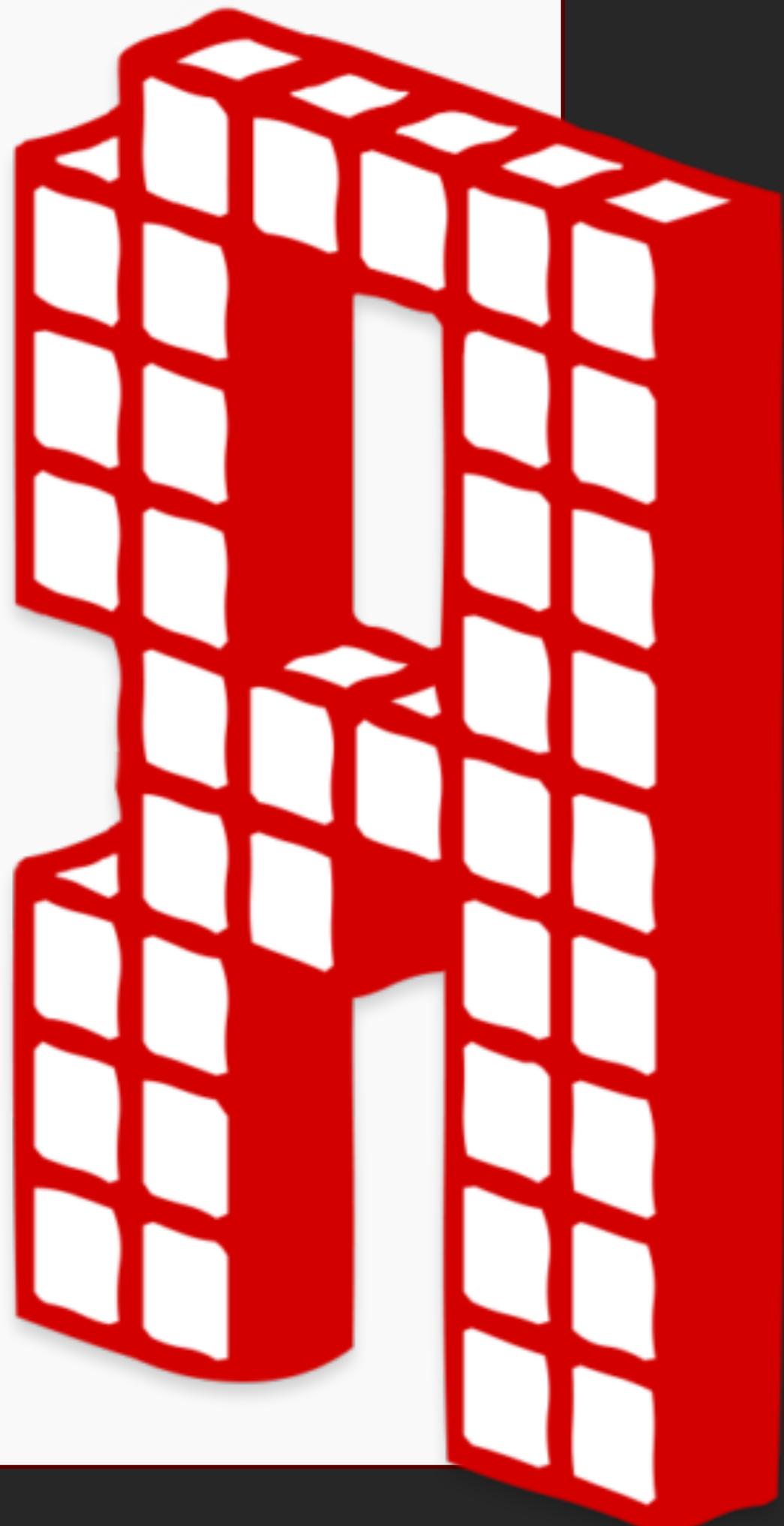
Defence against the dark arts:  
a proposal for a new MSc course  
(Bishop, 2023)



Masters course:  
**“Estimating the credibility of past research”**

(@UniBe since Fall '23)

Science as an intensely human & fallible activity  
Organised skepticism

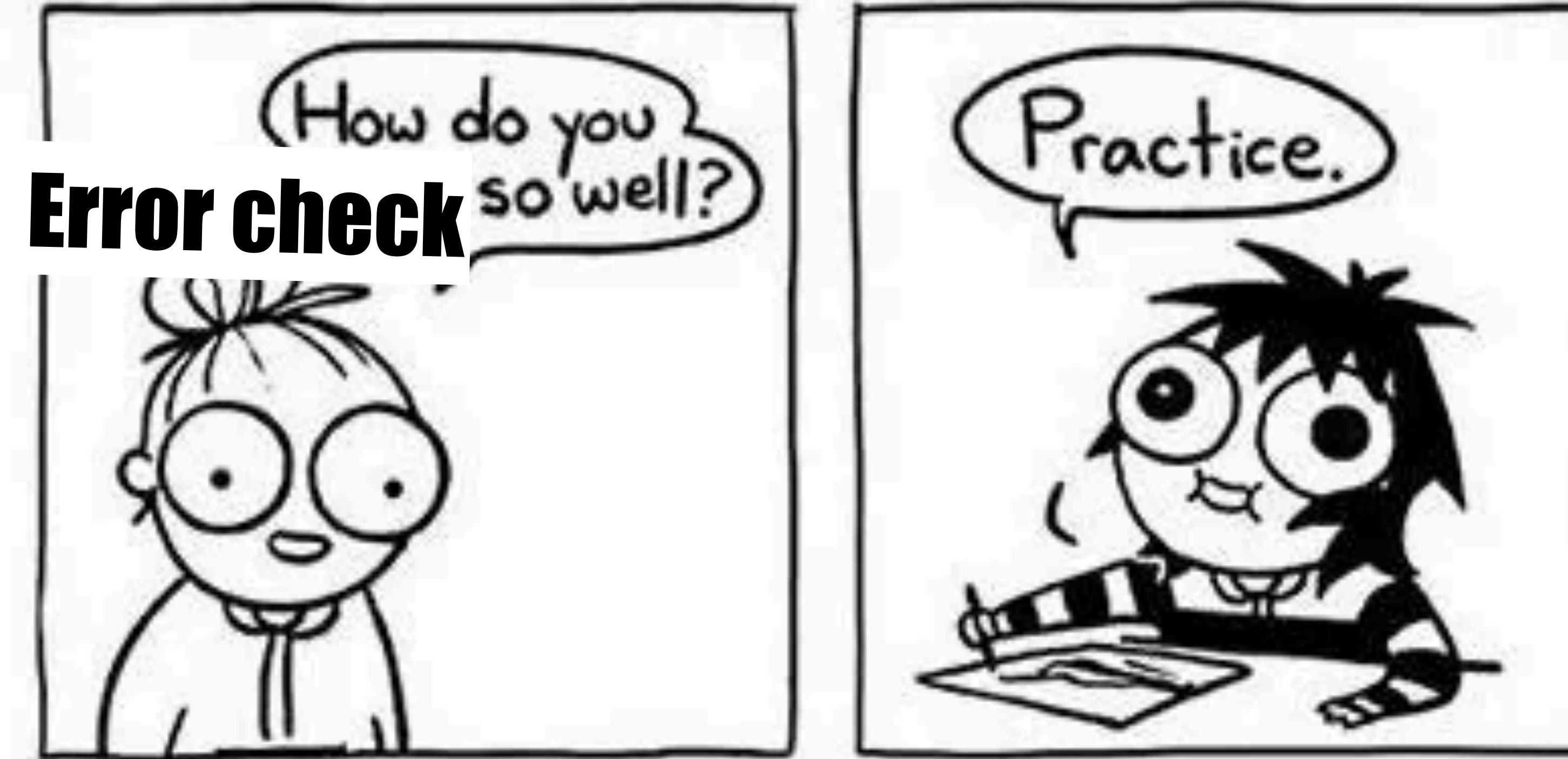




## Syllabus

- Mentality: Functions of science other than truth discovery
- *p*-curve, StatCheck, GRIM, GRIMMER, SPRITE, TIDES, PORT, ...
- Recomputing results from summary statistics
- Effect size plausibility
  - Reliability corrections
- Meta-analysis bias detection & its failures
- (Missing) causal language
- Spotting jingle/jangle fallacies
- Obtaining & extracting data
- Communicating errors + examples of critique

*Not merely information, practice!*



‘Normalise’ is a verb

Problems encountered

Permission to critique

xxxxxxxxxx

Language of critique

*Exposure & Practice!*



# ERROR

## A bug bounty program for science

xxxxxxxxxxxxxxxxxxxxxxxxxxxx

04.

## Goals



Explore and test **practical challenges** in the implementation of an error checking system



Estimating a **benefit-cost ratio** of an error detection system relative to **not detecting** these errors



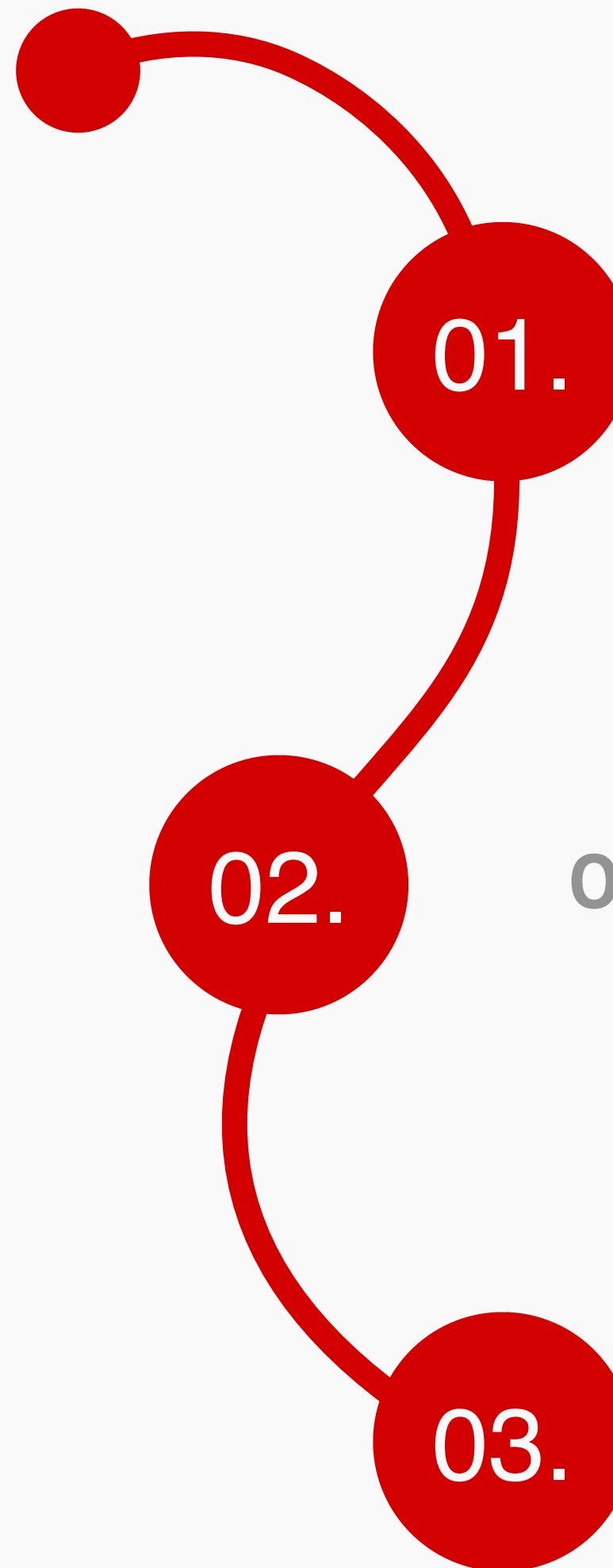
Obtaining empirical estimates of the **types** of errors & their **prevalence**

£ 220,000  
(250k CHF, €260k, \$285k)

fund

Check  
100  
published articles  
for errors

Pay authors & reviewers  
+ **bonus reward**  
contingent on errors found

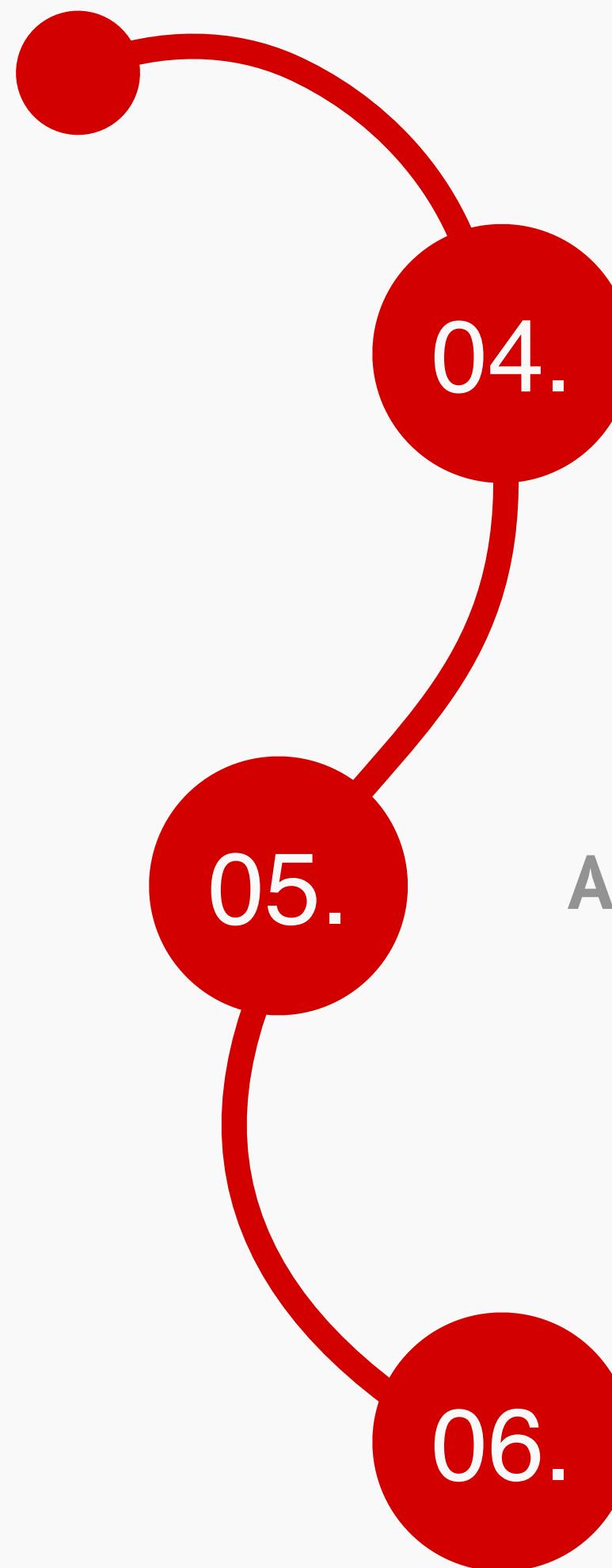


## Candidate articles

- Selected for 'importance':
  - Would matter if it contained errors
  - 30+ citations per year
- Recent (<10 years)
- Attention paid to representation & power
  - Gender, precarious employment, etc.

## Obtain author consent

## Match with reviewer



Error report

Author response

Decision, recommendation,  
& reward payments

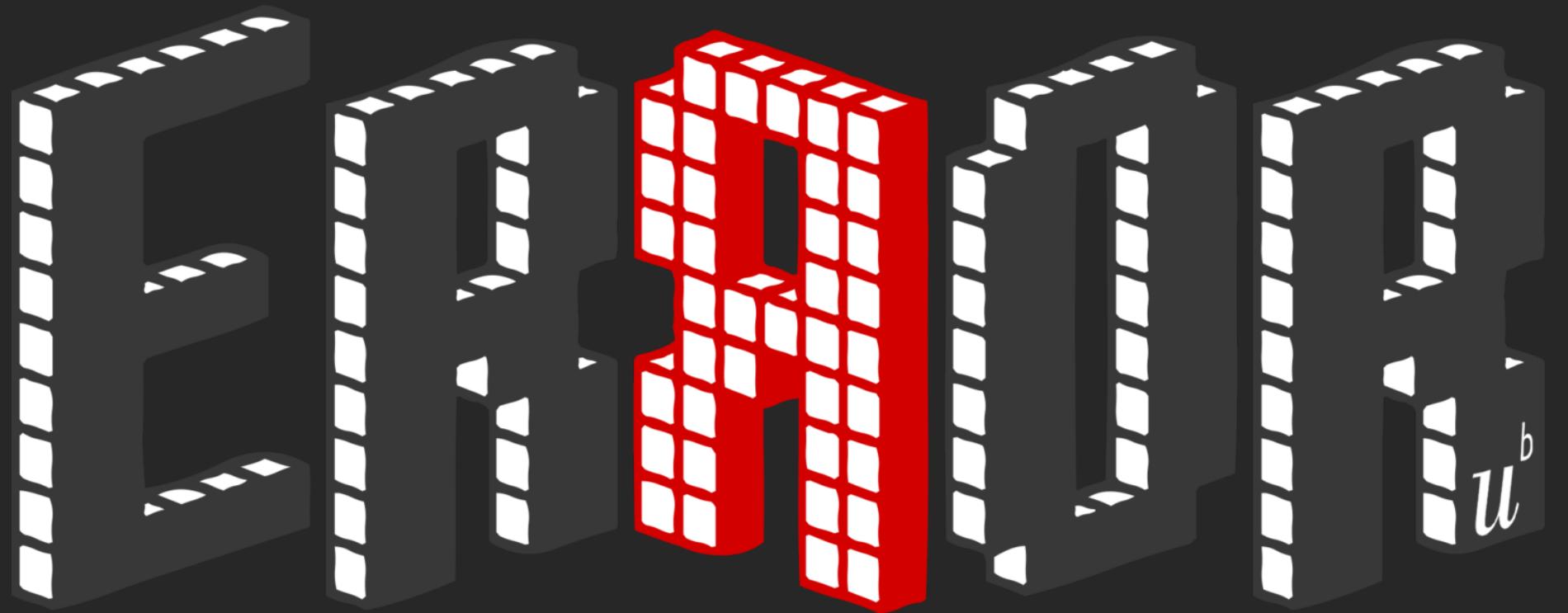
- Recommender makes decision & recommendation
  - All published on error.reviews website
  - CC-By 4.0 licence
- Possibly in conjunction with the advisory board

	<b>Base</b>	<b>Bonus reward</b>
Author	£ 220	Up to £220
Reviewer	£220 – £1,100 Depending on expected effort	Up to £2,200

\* *Indicative ranges; will be scaled for regional differences*

# Bonus rewards

<b>Decision</b>	<b>Recommendation</b>	<b>Bonus reward (indicative)</b>
No errors	No additional action beyond publication of error report on the error.reviews website.	£220 to the author
Minor errors	Authors to appropriately recognise these errors in future discussions of the article. <i>Probably most research!</i>	£220 to <b>both</b> the author & reviewer
Indeterminable errors	No determination could be made re the presence or absence of important potential errors. <i>Less desirable than verifiably minor errors!</i> Authors to appropriately recognise this lack of verifiability in future discussions of the article.	£220 to the reviewer
Moderate errors	Correction notice (minor)	£440 to the reviewer
Major errors	Correction notice (major) / may warrant an expression of concern	£880 to the reviewer
Severe errors	Retraction	£2,200 to the reviewer



1 / 14

Reviews Completed / Pending

134

Invited  
Papers

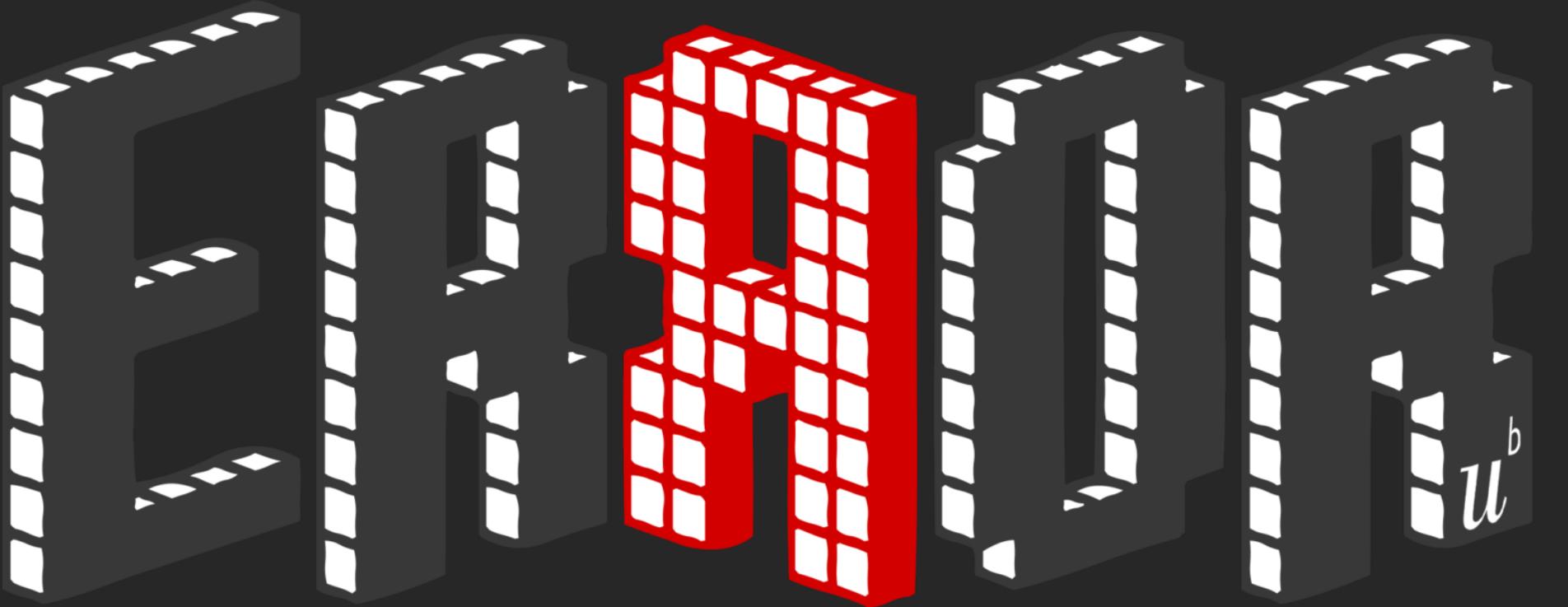
1

Papers  
with Errors

1'250

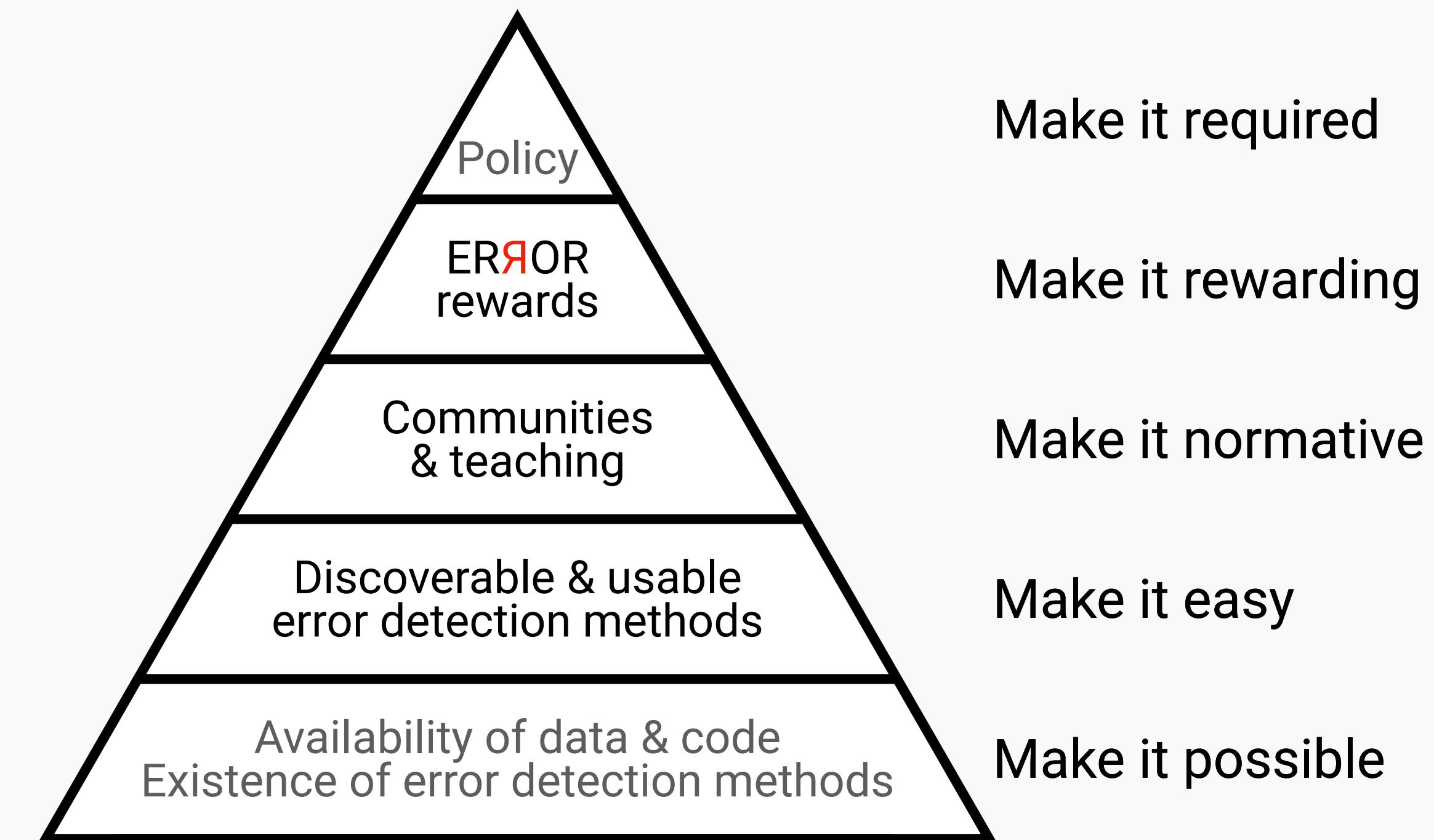
/ 250'000 CHF

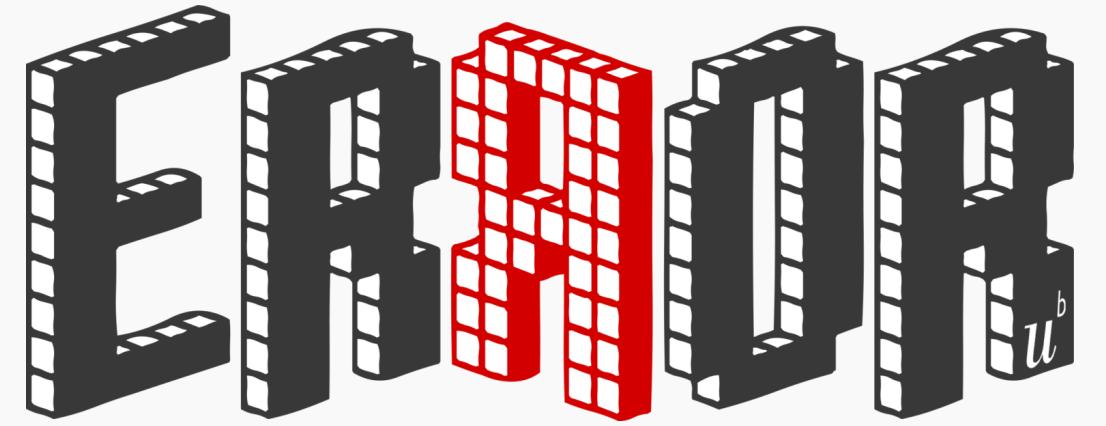
Total  
Payouts



error.reviews

# Our strategy to increase error checking, reporting, & correction





## Contact

ian.hussey@unibe.ch

Twitter: @ianhussey

Bluesky: @ianhussey.bsky.social

error-reviews.psy@unibe.ch

Twitter: @error\_reviews

Bluesky: @error.reviews

$u^b$

XXXXXXXXXXXXXXXXXXXX

## Collaborators

- Malte Elson
- Ruben Arslan
- Jamie Cummins
- Sabrina Norwood
- Frank Bosco
- Imran Kadolkar

## Links

- ERROR reviews [error.reviews](#)
- PORT [mmmdata.shinyapps.io/PORT](#)
- TIDES [mmmdata.shinyapps.io/TIDES](#)
- Finder [drjamiecummins.com/error-selector-demo](#)
- RegCheck [regcheck.app](#)

XXXXXXXXXXXXXXXXXXXX

Other normalisation efforts:  
**“Commentaries don’t get cited”**

	Citations of target article	Citations of commentary	Within-pair proportion
Mean	106	5	25%
Median	80	10	15%
Minimum	2	1	1%
Maximum	490	53	150%

$N = 36$  commentaries published in Psychological Science  
between 2007-2021