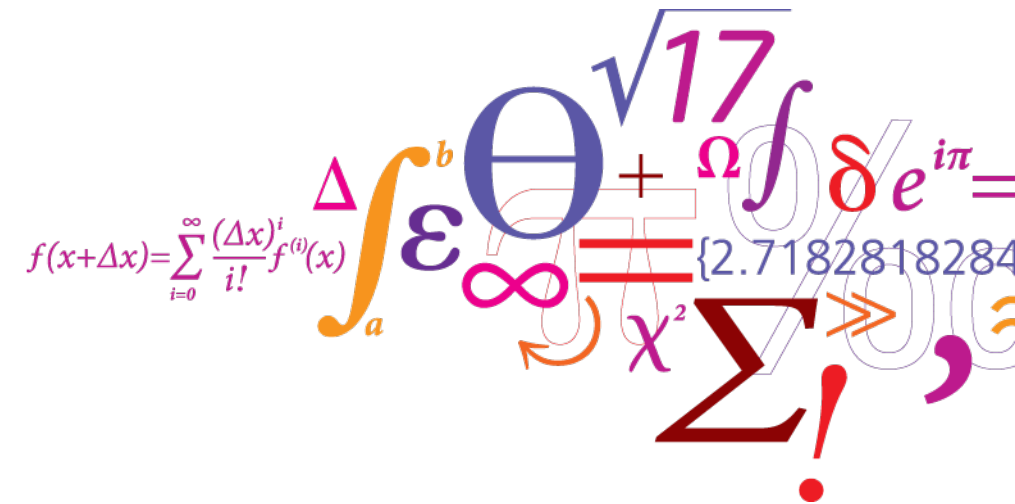# Reproducibility in science

Tommy Sonne Alstrøm

Associate Professor

Section of Cognitive Systems

# Agenda

13.00-13.50: Tommy: intro, reproducibility in science

14.00-17.00: Nick: git lecture/tutorials

# Topics that will be covered

Topic 1 – Reproducibility issues – examples of correlations and causation

- Examples of correlation

- Publication on correlation, causation and p-values


Topic 2 – Which obstacles hinders research to be reproducible and which solutions are offered by literature?
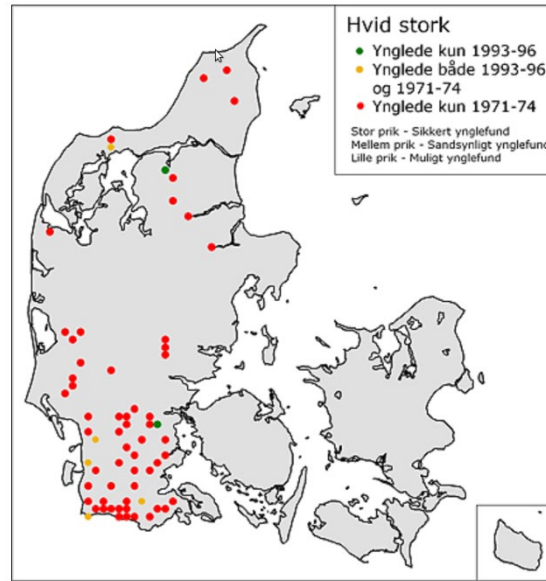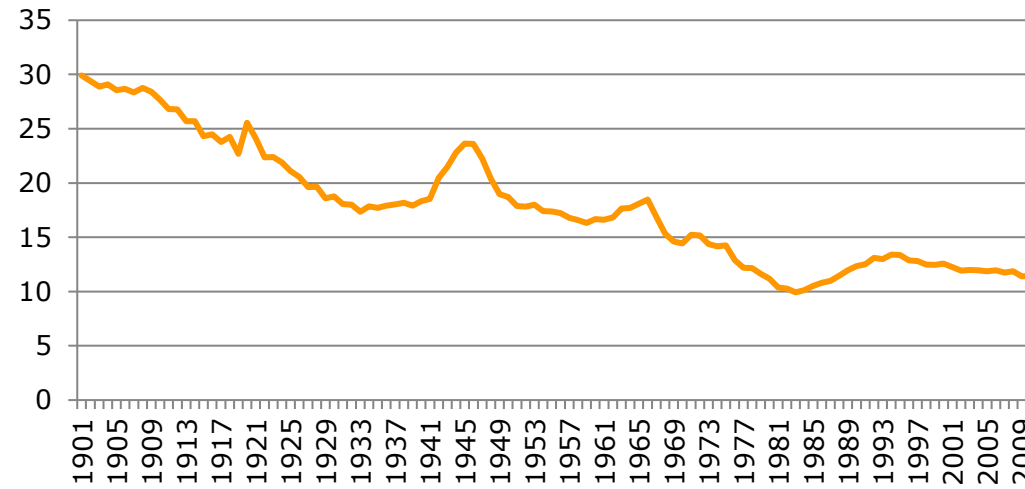
- Nature publishing take on reproducible research

- Survey from ~1500 researchers


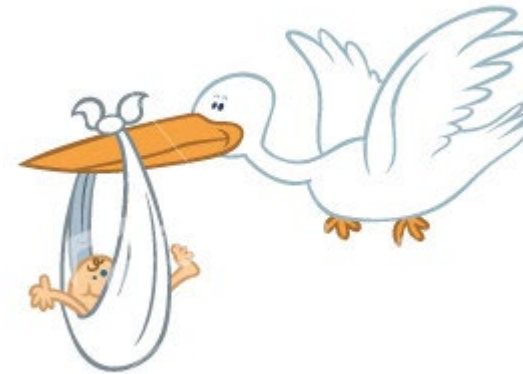Topic 3 – What tools can boost reproducibility?

- Proper statistical thinking

- Proper research data management

- Recommendations for further reading

Topic 1 – Reproducibility issues – examples of correlations and causation
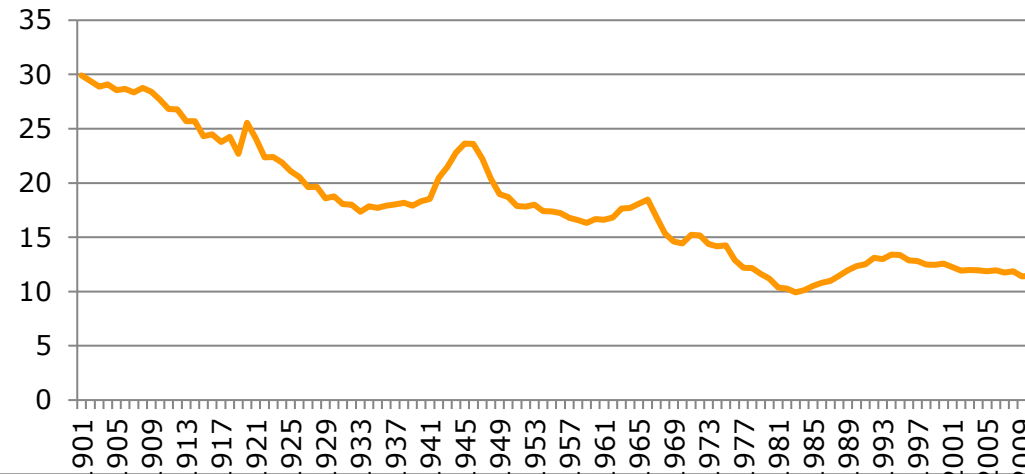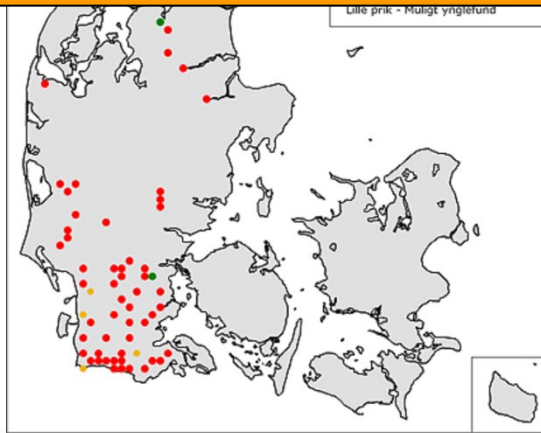
# Children born in total /population 1.jan (in 1000)

**Children born in total /population 1.jan (in 1000)**

Causality is **not** the same as correlation / connection / association
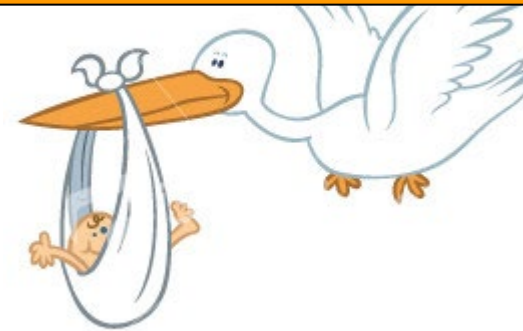
# Professor Franz H. Messerli: Chocolate Consumption, Cognitive Function and Nobel Laureates

- Science documents that chocolate intake improves cognitive function

- Number of Nobel prize winners is a surrogate to reflect a country's overall cognitive function.

- Extract data from Wikipedia and various other resources, e.g. chocolate consumption data is based on the years 20xx.

*Chocolate Consumption, Cognitive Function, and Nobel Laureates. Messerli, Franz H. New England Journal of Medicine 367, 1562-1564, 2012*

**Figure 1.** Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.
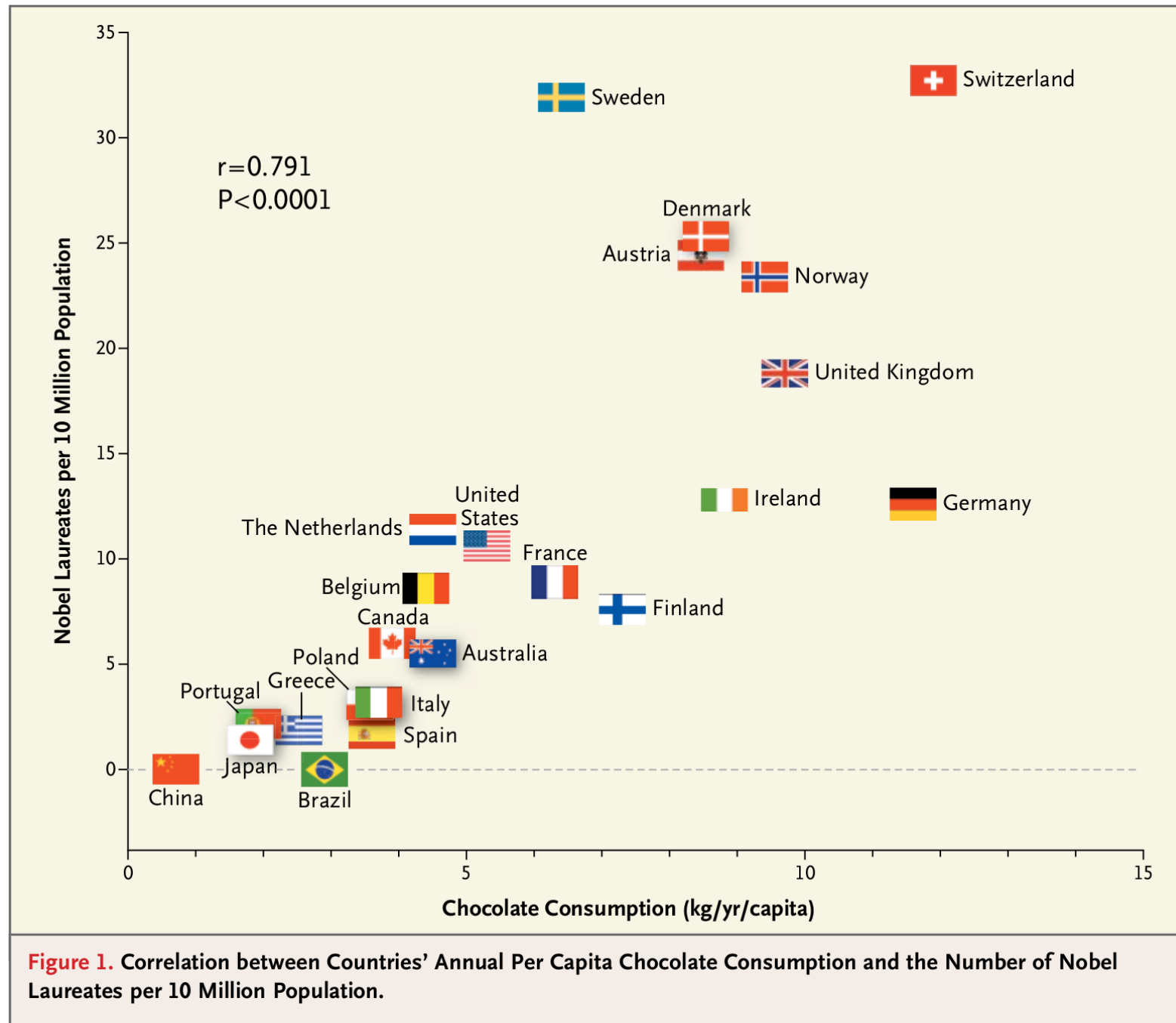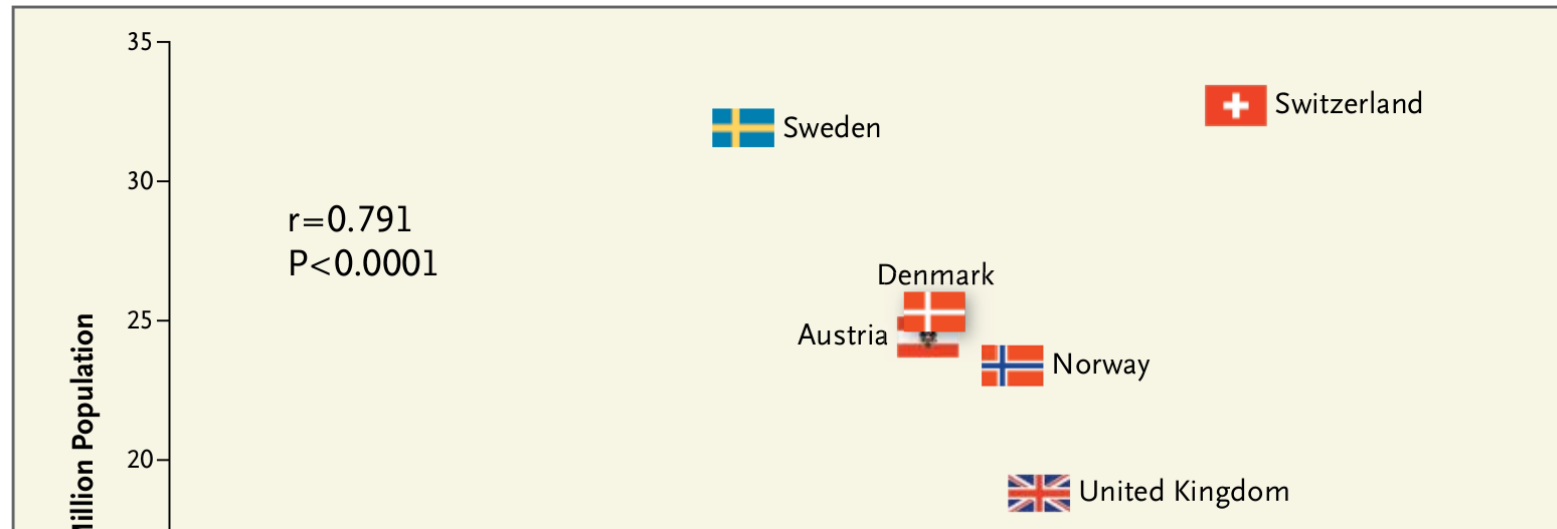
Messerli said **the whole idea is absurd**, although the data are legitimate **and contain a few lessons about the fallibility of science.**
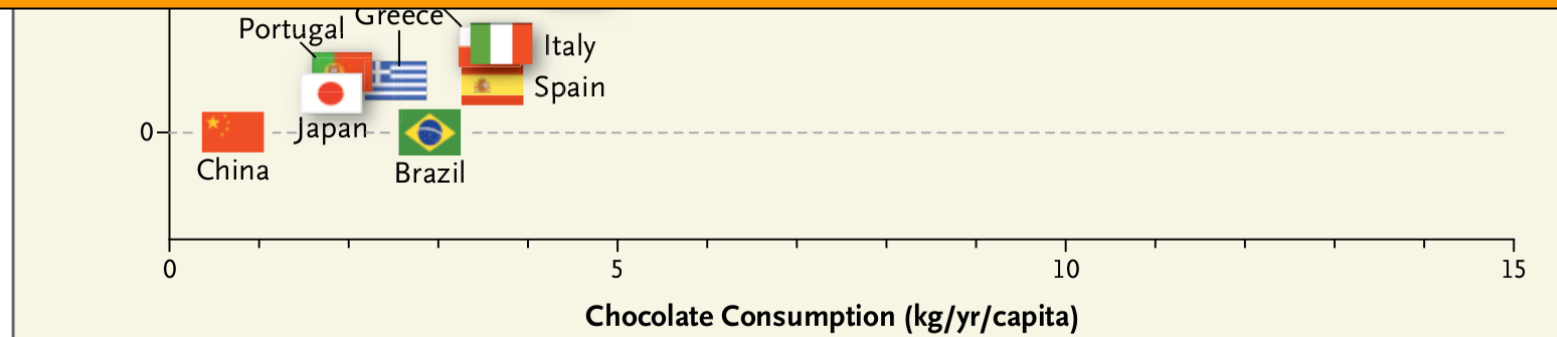
https://www.reuters.com/article/us-eat-chocolate-win-the-nobel-prize/eat-chocolate-win-the-nobel-prize-idUSBRE8991MS20121010

# Reproducibility – what does Wikipedia say

- Reproducibility is the ability of **an entire experiment** or study to be duplicated, either by the same researcher or **by someone else working independently**.

- Reproducible data - **repeatability** which is the degree of agreement of tests or measurements on replicate specimens by the same observer in the same laboratory.

- Reproducible research - the idea that the ultimate product of **academic research** is the paper along with the **full computational environment** used to produce the results in the paper such as the code, data, etc. that can be used to reproduce the results and create new work based on the research.

https://en.wikipedia.org/wiki/Reproducibility

# Nature Publishing Group has declared war on irreproducible research

- There is growing alarm about results that cannot be reproduced. Explanations include increased levels of scrutiny, complexity of experiments and statistics, and pressures on researchers. Journals, scientists, institutions and funders all have a part in tackling reproducibility.

- *Nature* has taken substantive steps to improve the transparency and robustness in what we publish, and to promote awareness within the scientific community.



http://go.nature.com/huhbyr

# Unreliable research - Trouble at the lab

The biotech company Amgen had a team of about 100 scientists trying to reproduce the findings (preclinical) of 53 "landmark" articles in cancer research published by reputable labs in top journals.
Only 6 of the 53 studies were reproduced (about 10%).

## REPRODUCIBILITY OF RESEARCH FINDINGS
Preclinical research generates many secondary publications, even when results cannot be reproduced.

| Journal impact factor | Number of articles | Mean number of citations of non-reproduced articles* | Mean number of citations of reproduced articles |
|---|---|---|---|
| >20 | 21 | 248 (range 3–800) | 231 (range 82–519) |
| 5–19 | 32 | 169 (range 6–1,909) | 13 (range 3–24) |

Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.
*Source of citations: Google Scholar, May 2011.

*Drug development: Raise standards for preclinical cancer research. C. Glenn Begley and Lee M. Ellis. Nature 483, 531–533, 2012.*

# Estimating the Reproducibility of Psychological Science
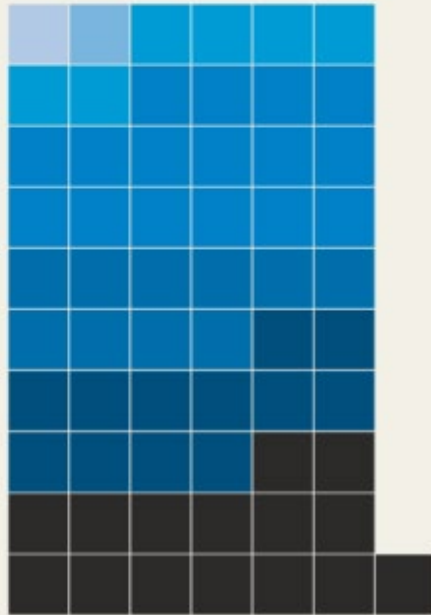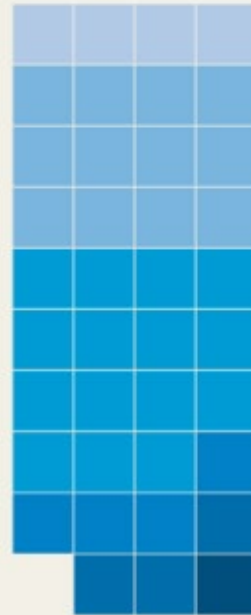
Topic 2 – Which obstacles hinders research to be reproducible and which solutions are offered by literature?

IS THERE A REPRODUCIBILITY CRISIS?

7% Don't know

52% Yes, a significant crisis

3% No, there is no crisis

1,576 researchers surveyed

38% Yes, a slight crisis

©nature

*1,500 scientists lift the lid on reproducibility. Monya Baker. Nature 533, 452–454, May 2016.*
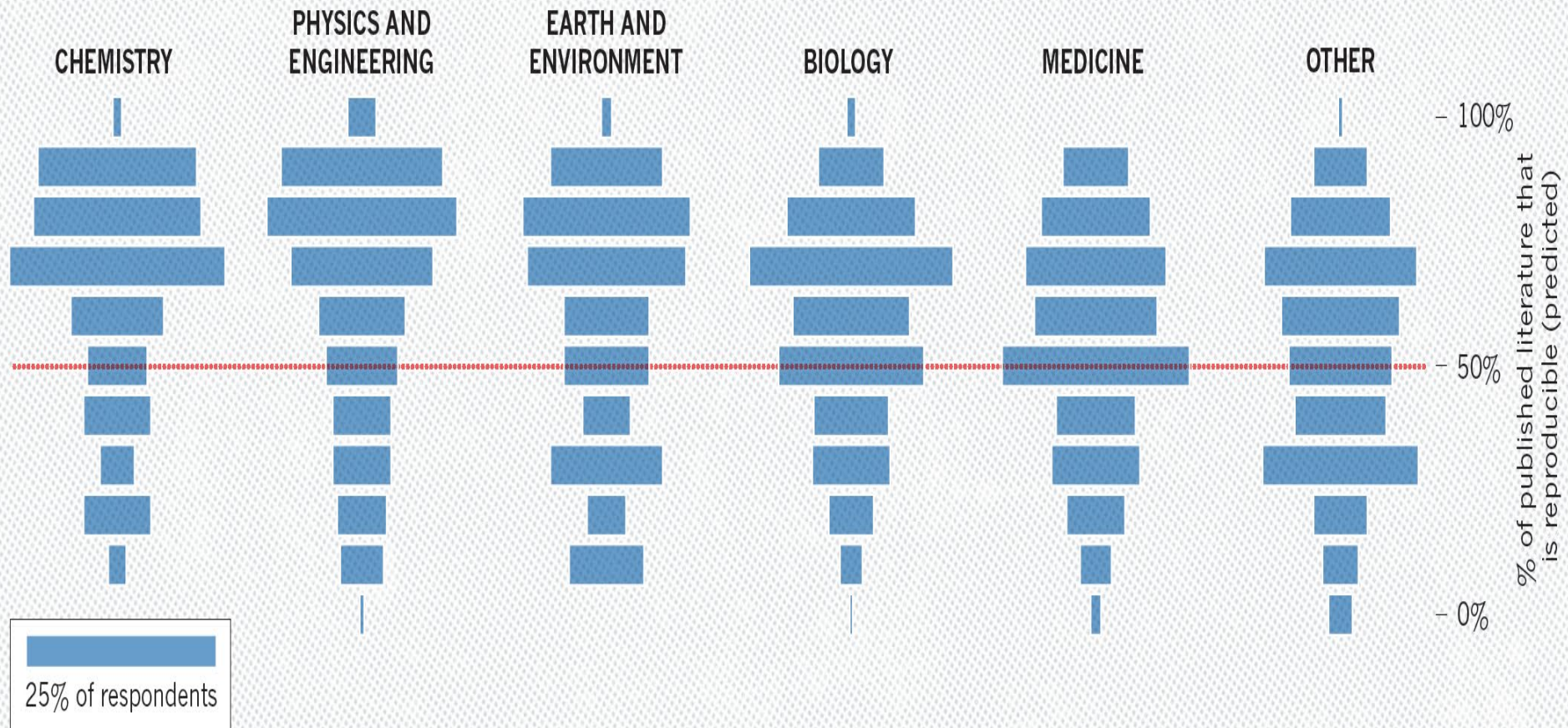
## WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.

● Always/often contribute   ● Sometimes contribute
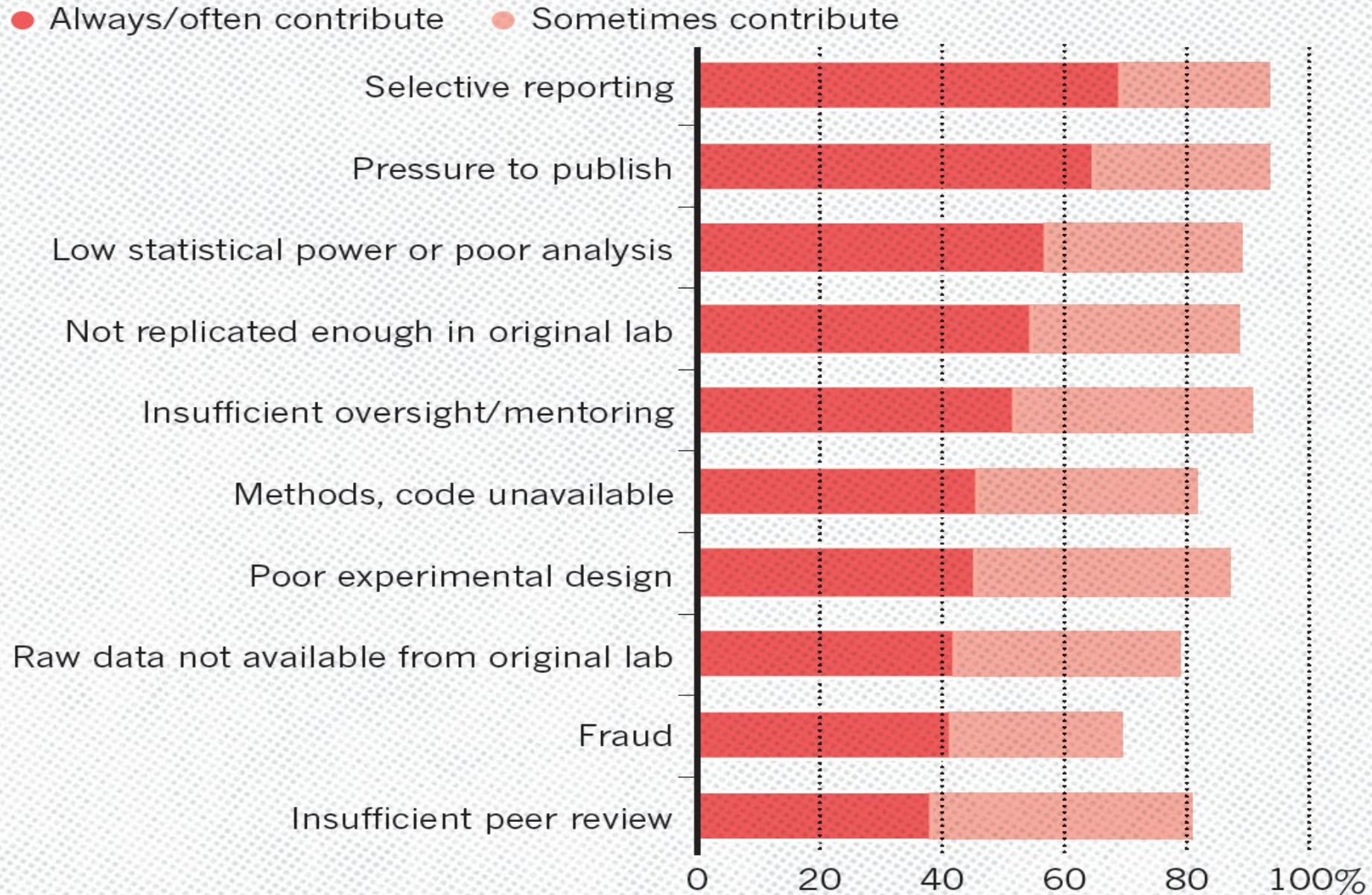
**WHAT FACTORS COULD BOOST REPRODUCIBILITY?**

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

● Very likely ● Likely

Better understanding of statistics
Better mentoring/supervision
More robust design
Better teaching
More within-lab validation
Incentives for better practice
Incentives for formal reproduction
More external-lab validation
More time for mentoring
Journals enforcing standards
More time checking notebooks

0    20    40    60    80    100%

©nature

# Study design

Results from observational studies start to become believable when:

(i) the results are replicated in many studies,

(ii) each of the studies controlled for plausible confounding variables,

(iii) there is a plausible scientific explanation for the existence of a causal relationship.

"A good example is smoking and cancer. Numerous studies have shown a relationship between smoking and cancer even after adjusting for many **confounding variables**. Moreover, in **laboratory studies**, smoking has been shown to **damage lung cells**. Finally, a **causal link between** smoking and cancer has been found **in randomized animal studies**. It is this collection of evidence over many years that makes this a convincing case. One single observational study is not, by itself, strong evidence. Remember that when you read the newspaper."

*Wasserman, All of Statistics, sec 16.3*

# Study design

Results from observational studies start to become believable when:

(i) the results are replicated in many studies,

(ii) each of the studies controlled for plausible confounding variables,

(iii) there is a plausible scientific explanation for the existence of a causal relationship.

"A good example is smoking and cancer. Numerous studies have shown a relationship between smoking and cancer, even after ad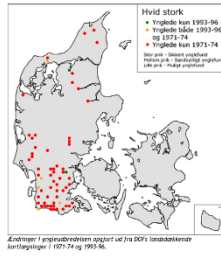justing for many **confounding variables**. Moreover, in **laboratory studies**, smoking has been shown to **damage lung cells**. Finally, a **causal link between** smoking and cancer has been found **in randomized animal studies**. It is this collection of evidence over many years that makes this a convincing case. One single observational study is not, by itself, strong evidence. Remember that when you read the newspaper."

**Theorem**

In general, correlation is **not** causation, however, when the data is randomly collected, correlation estimates causation.

*Wasserman, All of Statistics, sec 16.4*

*Wasserman, All of Statistics, sec 16.3*

## 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

78.7*
11.5*

Twice as likely

Equally likely

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

**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

Twice as likely

Equally likely

78.7*
11.5*

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

Any single team's results are **strongly influenced by subjective choices** during the analysis phase. Had any one of these 29 analyses come out as a single peer-reviewed publication, **the conclusion could have ranged from no skin color bias in referee decisions to a huge bias**.
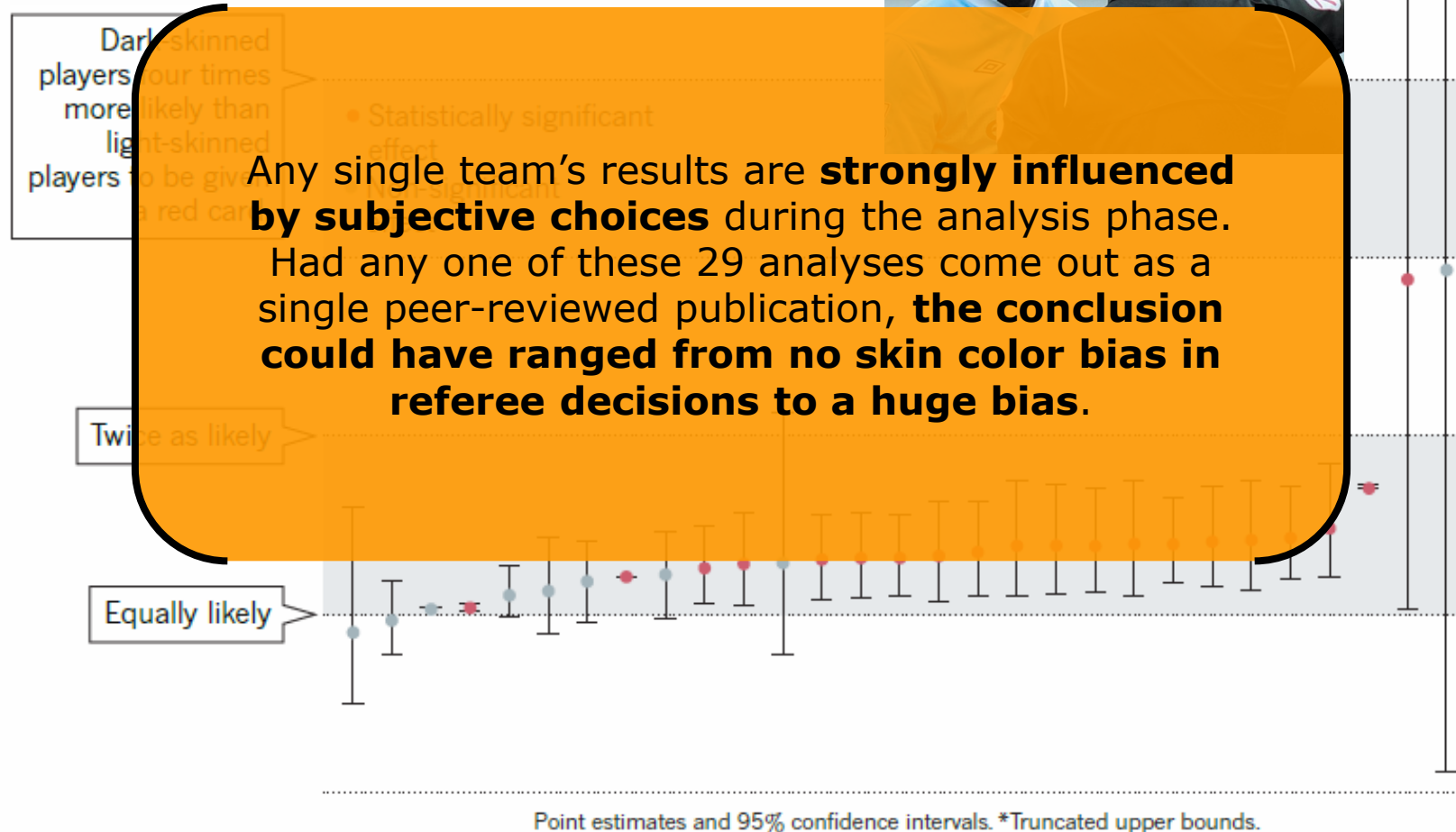
# Cognitive Bias



How scientists fool themselves – and how they can stop, *Nature* **526**, 182-185 (08 October 2015)

# COGNITIVE FALLACIES IN RESEARCH

**HYPOTHESIS MYOPIA**
Collecting evidence to support a hypothesis, not looking for evidence against it, and ignoring other explanations.

**TEXAS SHARPSHOOTER**
Seizing on random patterns in the data and mistaking them for interesting findings.

**ASYMMETRIC ATTENTION**
Rigorously checking unexpected results, but giving expected ones a free pass.

**JUST-SO STORYTELLING**
Finding stories after the fact to rationalize whatever the results turn out to be.

# DEBIASING TECHNIQUES

**DEVIL'S ADVOCACY**
Explicitly consider alternative hypotheses — then test them out head-to-head.

**PRE-COMMITMENT**
Publicly declare a data collection and analysis plan before starting the study.

**TEAM OF RIVALS**
Invite your academic adversaries to collaborate with you on a study.

**BLIND DATA ANALYSIS**
Analyse data that look real but are not exactly what you collected — and then lift the blind.

go.nature.com/nqyohl

© Nature

The Texas Sharpshooter Procedure
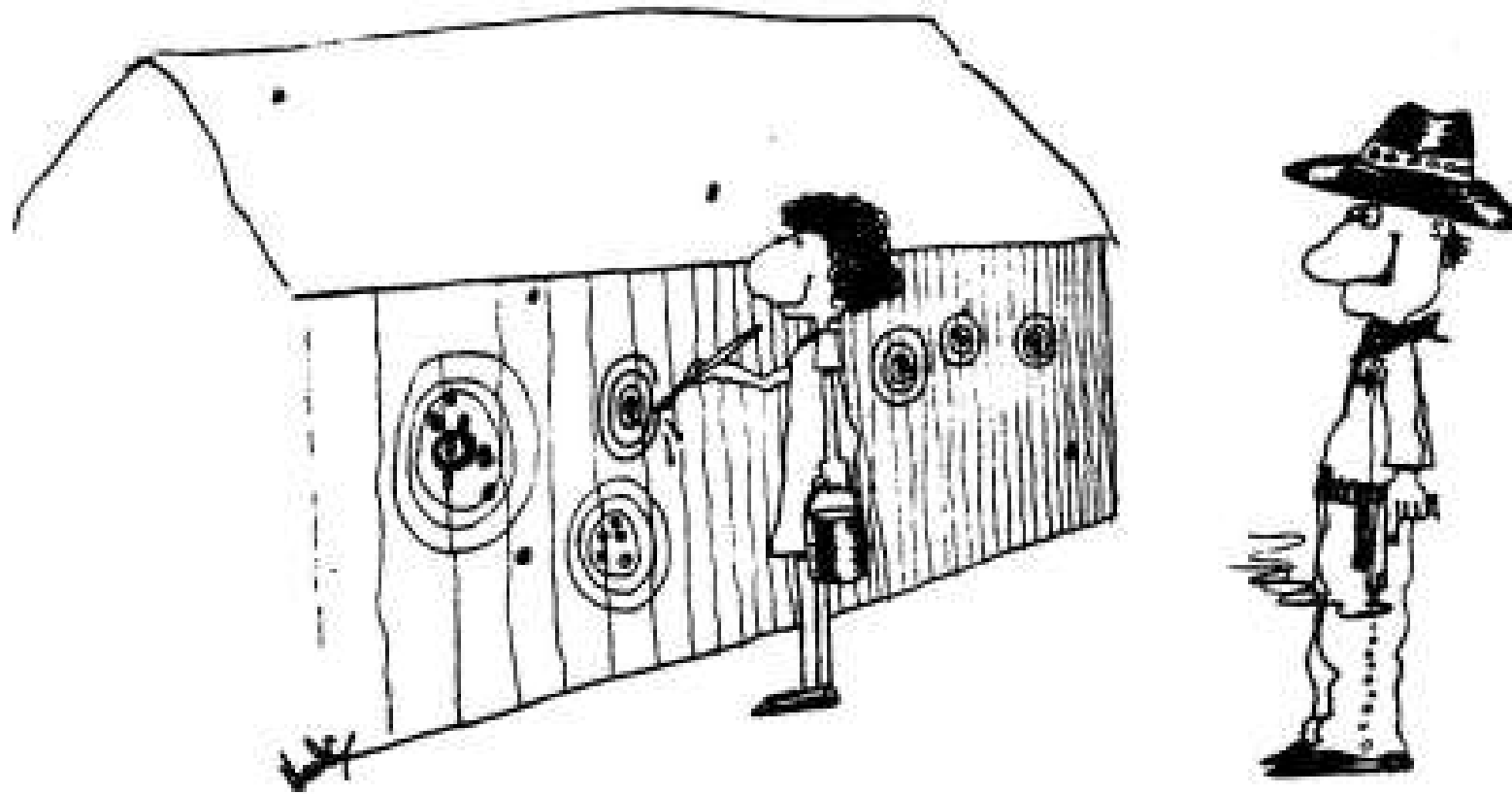
Topic 3 – What tools can boost reproducibility?

# Topic 3 – Reproducibility of computational results.
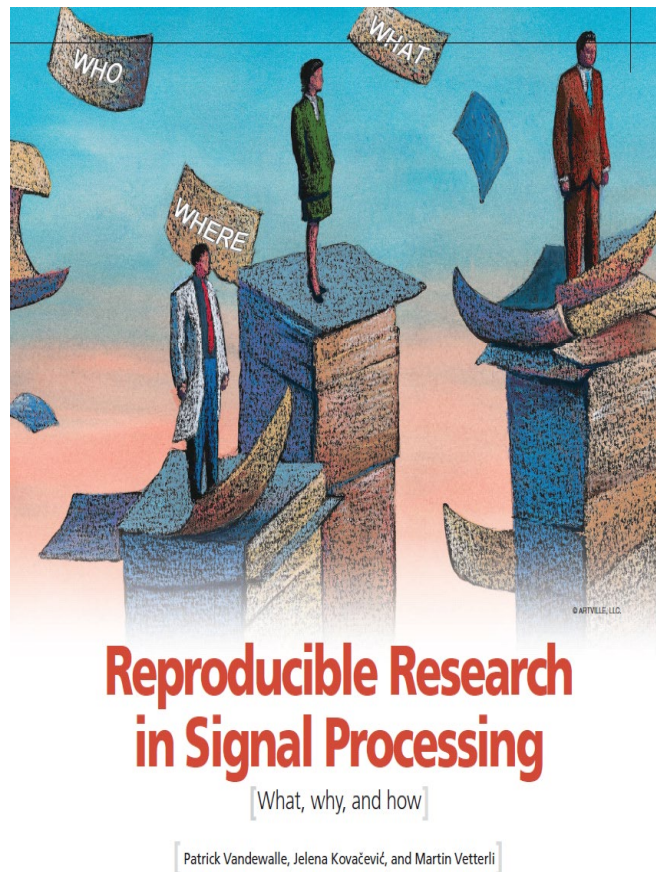
**DEFINITION: REPRODUCIBLE RESEARCH IN COMPUTATIONAL SCIENCE**

A research work is called reproducible if all information relevant to the work, including, but not limited to, text, data and code, is made available, such that an independent researcher can reproduce the results.

# The basic principle of computational reproducible research
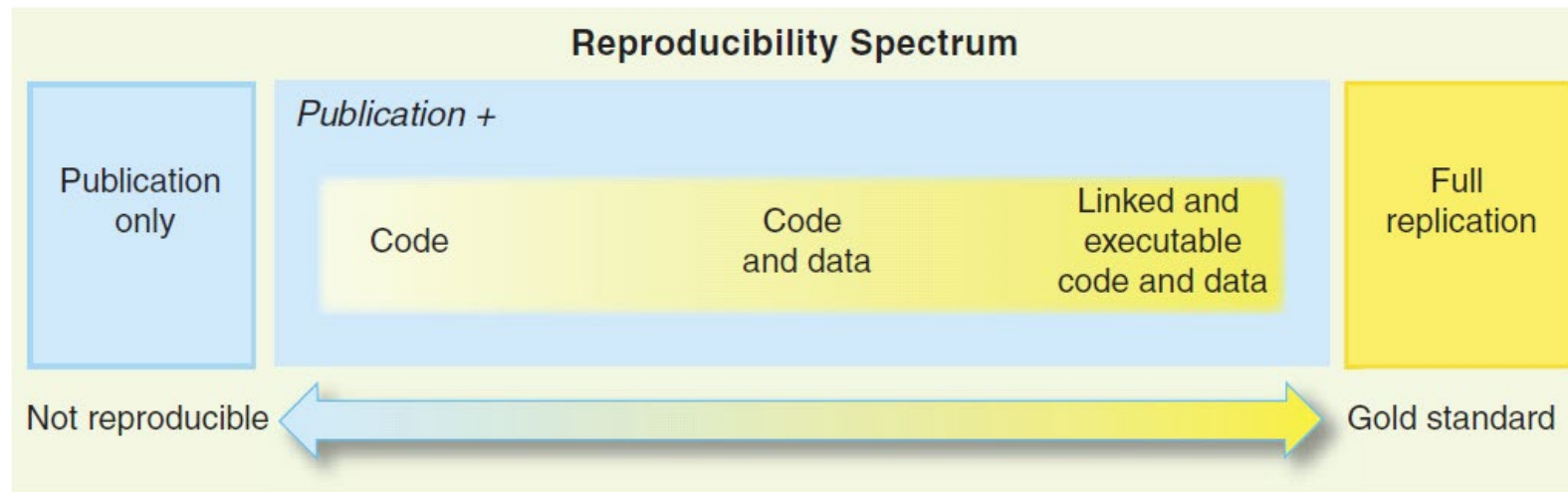


**Fig. 1.** The spectrum of reproducibility.

# Checklist of reproducibility

➢ The paper
  ➢ Is the algorithm described in sufficient detail?
  ➢ Are exact parameter values given?
  ➢ Is there a block diagram?
  ➢ Is there a pseudocode?
  ➢ Are there proofs for all the theorems?
  ➢ Is the algorithm compared to other algorithms?

➢ The code
  ➢ Are implementation details (programming language, platform, compiler flags, etc.) given?
  ➢ Is the code available online?

➢ The data
  ➢ Is there an explanation of what the data represents?
  ➢ Is the size of the data set acceptable?
  ➢ Is the data set available online?

PROBABLE CAUSE
A *P* value measures whether an observed result can be attributed to chance. But it cannot answer a researcher's real question: what are the odds that a hypothesis is correct? Those odds depend on how strong the result was and, most importantly, on how plausibile the hypothesis is in the first place.

■ Chance of real effect
■ Chance of no real effect

THE LONG SHOT
19-to-1 odds against

THE TOSS-UP
1-to-1 odds

THE GOOD BET
9-to-1 odds in favour

**Before the experiment**
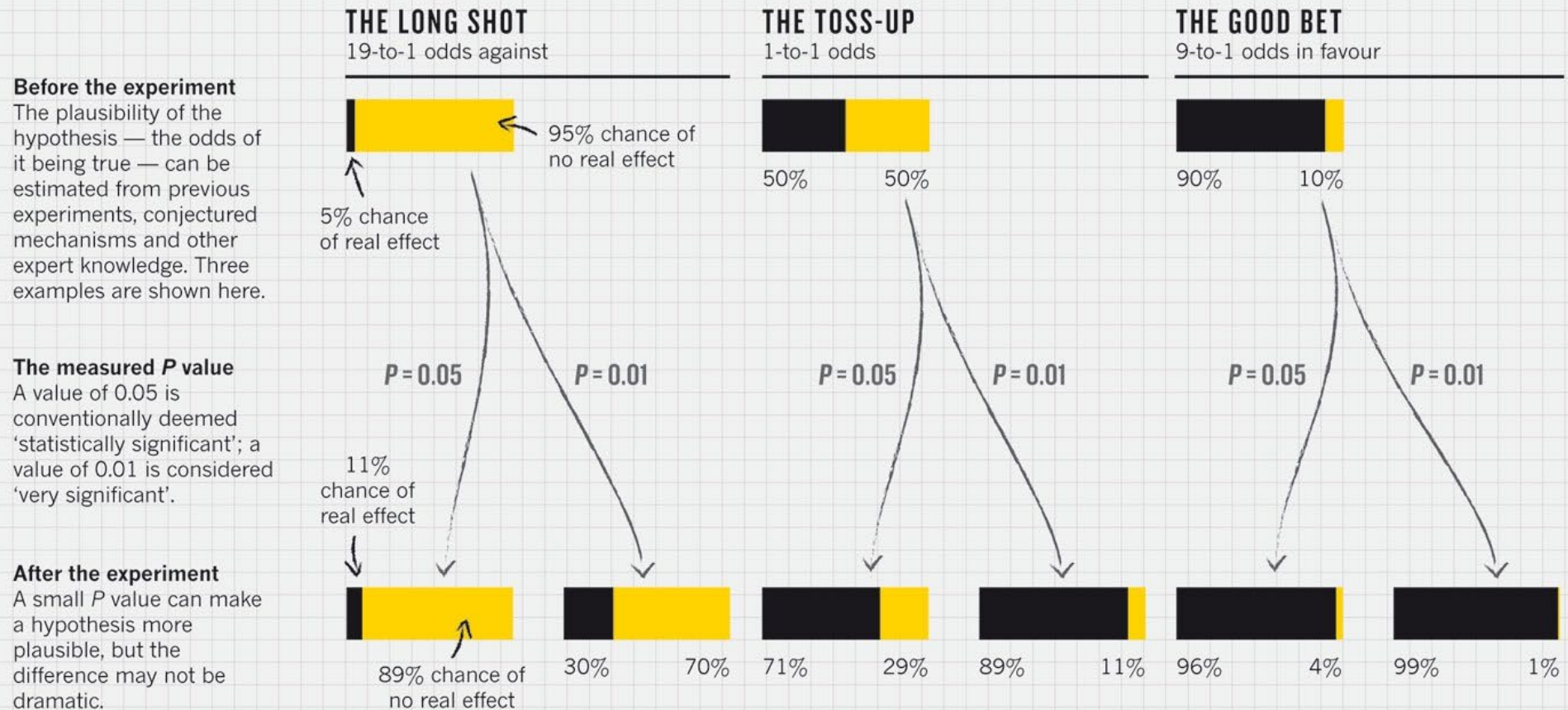The plausibility of the hypothesis — the odds of it being true — can be estimated from previous experiments, conjectured mechanisms and other expert knowledge. Three examples are shown here.

95% chance of no real effect

5% chance of real effect

50% 50%

90% 10%

**The measured *P* value**
A value of 0.05 is conventionally deemed 'statistically significant'; a value of 0.01 is considered 'very significant'.

P = 0.05    P = 0.01

P = 0.05    P = 0.01

P = 0.05    P = 0.01

11% chance of real effect

**After the experiment**
A small *P* value can make a hypothesis more plausible, but the difference may not be dramatic.

89% chance of no real effect

30%    70%    71%    29%    89%    11%    96%    4%    99%    1%

*Scientific method: Statistical errors. Regina Nuzzo. Nature 506, 150–152, 2014*

*Why Most Published Research Findings Are False. John P. A. Ioannidis. PLOS Medicine, 2005*

# Use version control

## How to do research

1. Do awesome work
2. Write it down
3. Submit paper
4. Fame and glory
5. Move on to the next project (step 1)

# Use version control

## How research actually works

1. Have an idea
2. Collect data
3. Experiment
4. Fail
5. (Go to step 1)
6. Impending deadline
7. Submit paper

# Use version control

## How research actually works

1. Have an idea
2. Collect data
3. Experiment
4. Fail
5. (Go to step 1)
6. Impending deadline
7. Submit paper
8. Keep refining (1-5)
9. Paper accepted (months later)
10. Final draft
11. Support it for the rest of your life
12. Keep refining...

# Mindset that ensures your research becomes "reproducible computational research"

1. Think about how to make it easy for others.

1. Think about how to make it easy for others.

1. Think about how to make it easy for others.

Personal Gain:

If the research is easy to reproduce for others, it is **easy to reproduce for you**.

# Steps

- Have a proper code and folder structure.

- Automate as much as possible.

- Manual steps, if any, must be well described.

- Figures should be generated by code.

- Figure scripts should only create figures, or very light data processing.

- Store random seeds. On the cluster, the seeds on the nodes needs to be stored as well.

# Further reading

**Talks from a pioneer in Data Analysis - Andrew Gelman, Professor of statistics and political science and director of the Applied Statistics Center at Columbia University**

The Statistical Crisis in Science and How to Move Forward
https://youtu.be/KS3yPw91iC0

Bayes, statistics, and reproducibility (Rutgers, Foundations of Probability)
https://youtu.be/xgUBdi2wcDI

Crimes against data
https://youtu.be/fc1hkFC2c1E

Mød forskeren, der trak tæppet væk under videnskaben
https://www.zetland.dk/historie/sekdyb4n-a8dQKjjz-258ff