nature genetics

LETTERS

https://doi.org/10.1038/s41588-019-0560-2

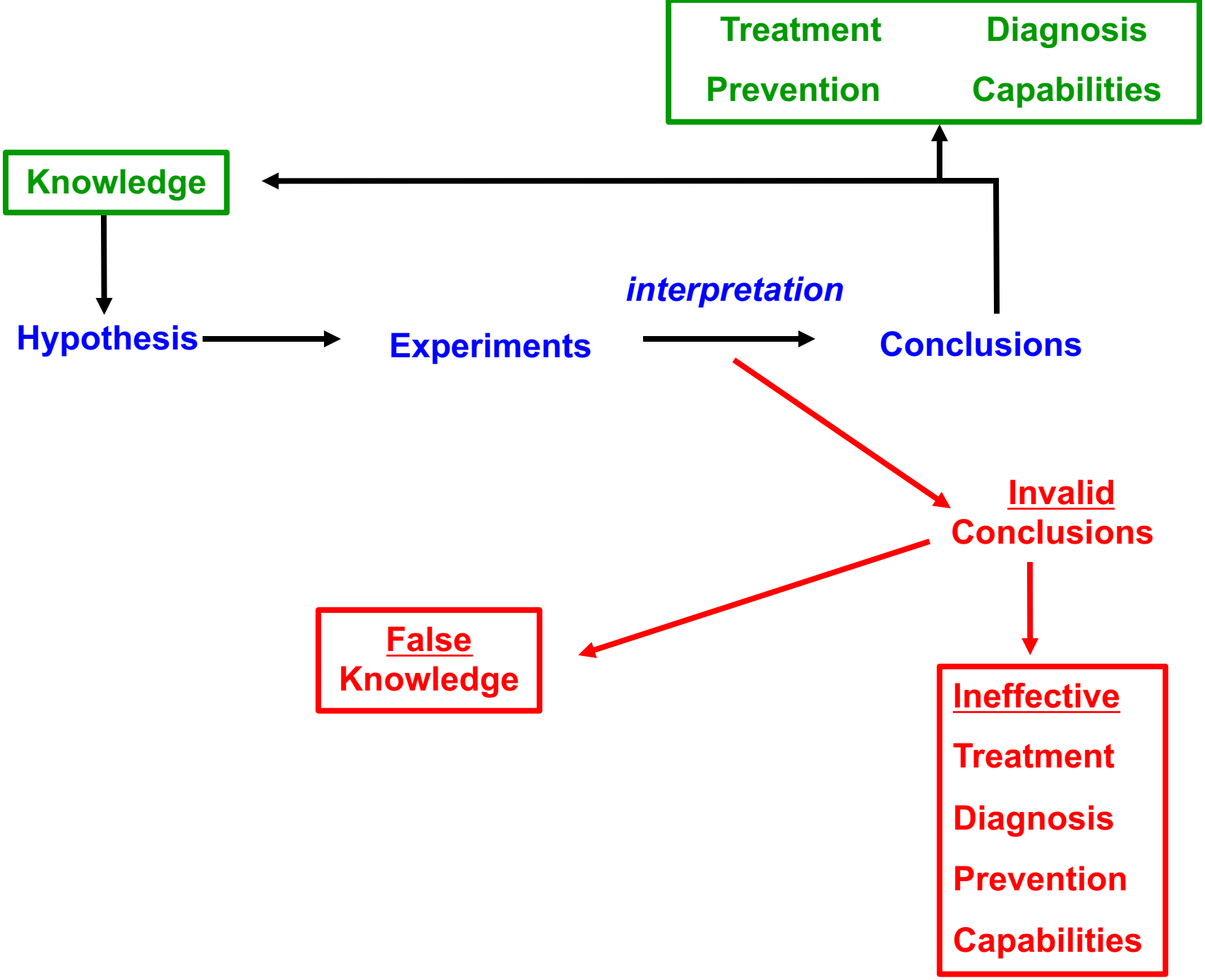
GeVIR is a continuous gene-level metric that uses variant distribution patterns to prioritize disease candidate genes

Nikita Abramovs[1,2], Andrew Brass[1,3] and May Tassabehji[2,4*]

Nature Genetics **52**, 35–39(2020)

You are leading the bioinformatics group at a pharmaceutical company which is developing new therapies for rare genetic disorders

What are your risks in prioritizing candidate genes for drug targeting based on this paper's GeVIR gene variation intolerance ranking method?


Treatment
Diagnosis
Prevention
Capabilities
Knowledge
Hypothesis
Experiments
interpretation
Conclusions
Invalid
Conclusions
False
Knowledge
Ineffective
Treatment
Diagnosis
Prevention
Capabilities

2 pharmaceutical companies examine the validity of published pre-clinical research they hope to use for drug development

Drug development: Raise standards for preclinical cancer research
C. Glenn Begley & Lee M. Ellis [AMGEN]. Nature 483: 531–533, 2012

Believe it or not:
How much can we rely on published data on potential drug targets?
Prinz, Schlange & Asadullah [BAYER]. Nature Reviews Drug Discovery 10: 712, 2011

Published research was confirmed in only 20 of 110 drug development projects

What might have led 90 out of 110 research groups to publish conclusions from their experiments that subsequently couldn't be reproduced by other researchers?

Fields with high proportions of non-reproducible published findings:

Molecular Biology	**Microarrays**	**Genetic Associations**
Bioinformatics	**Biomarkers**	**Physiology**
Epidemiology	**Neuroimaging**	**Psychology**
Economics	**Sociology**	**Astrology**

Etcetera

What are some common pathways to false positives? [validity traps]

How can you reduce your chances of publishing a false positive?

As various validity traps and defenses are mentioned:

-make notes about how these might be relevant to your research plan

- give your input, questions, disputations etc at that point,
or save for end of class

- and/or discuss them with your research group later

- and/or email Rob for discussion, clarification, references etc.

Where we are going today:

1. Common methodological validity traps affecting biomedical research

2. One example of a common validity trap in bioinformatics, hopefully leading to student-led discussion of other validity traps

3. Common motivational and cognitive biases affecting all humans, thus affecting all research

Validity Trap: Publication Bias

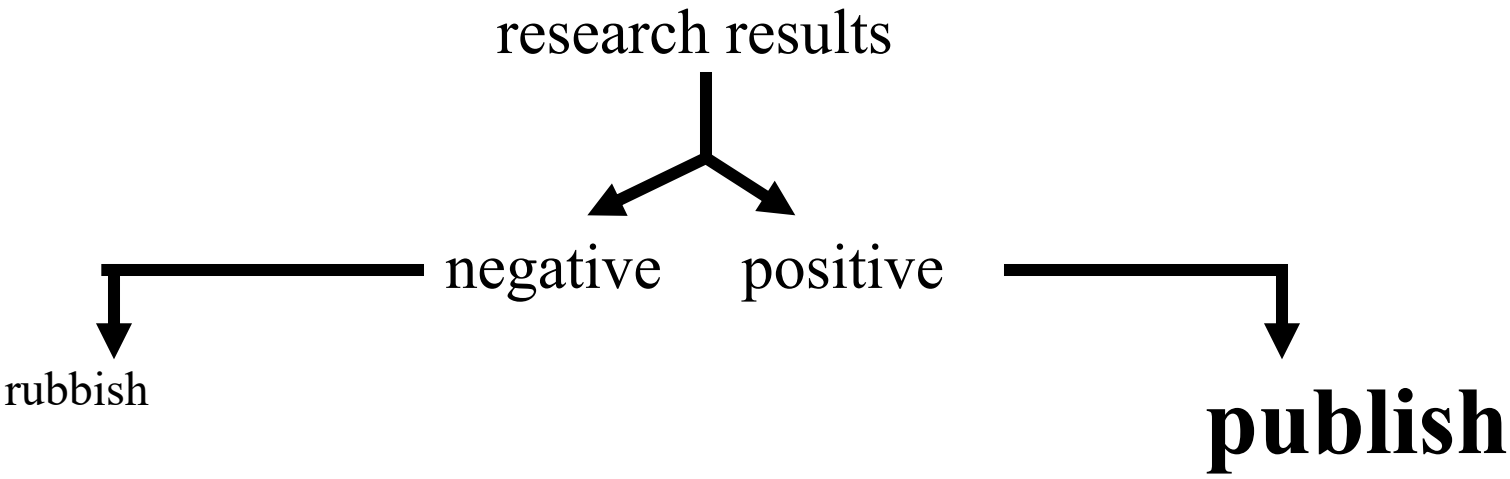
Journals are impressed by [and rewarded by] studies that confirm a plausible hypothesis, especially if the hypothesis is exciting and novel and useful

As a result, it is much easier to publish studies that are positive than negative, even though:

- negative studies are often very useful

- positive studies are often very exaggerated

- in complex systems research, most hypotheses are wrong



Publication Bias can massively skew a research field towards publishing positive results

When this happens,
most of these results are expected to be false positives

Almost all articles on cancer prognostic markers report statistically significant results

Panayiotis A. Kyzas[a], Despina Denaxa-Kyza[a], John P.A. Ioannidis[a,b,c,]*

Of 1915 published papers on cancer biomarkers, only 22 did not claim prognostic value of a biomarker.

Publication Bias in Reports of Animal Stroke Studies
Emily S. Sena et al.,PLoS Biol. 8(3): e1000344, 2010

In 525 publications on animal models of stroke, only 10 reported no significant effects on stroke

Publication Bias is a <u>Cultural</u> Problem
There is not much you can personally do about it

Drug development: Raise standards for preclinical cancer research

C. Glenn Begley & Lee M. Ellis. Nature 483: 531–533, 2012

Prior scientific findings were confirmed in only 6 of 53 oncology drug development projects at Amgen

Features of the small number of publications that <u>were</u> reproduced:

- Experiments included effective controls
- Reagents were validated
- Complete data sets were shown
- Statistical tests were appropriately used
- Experiments were blinded
- Experiments were repeated

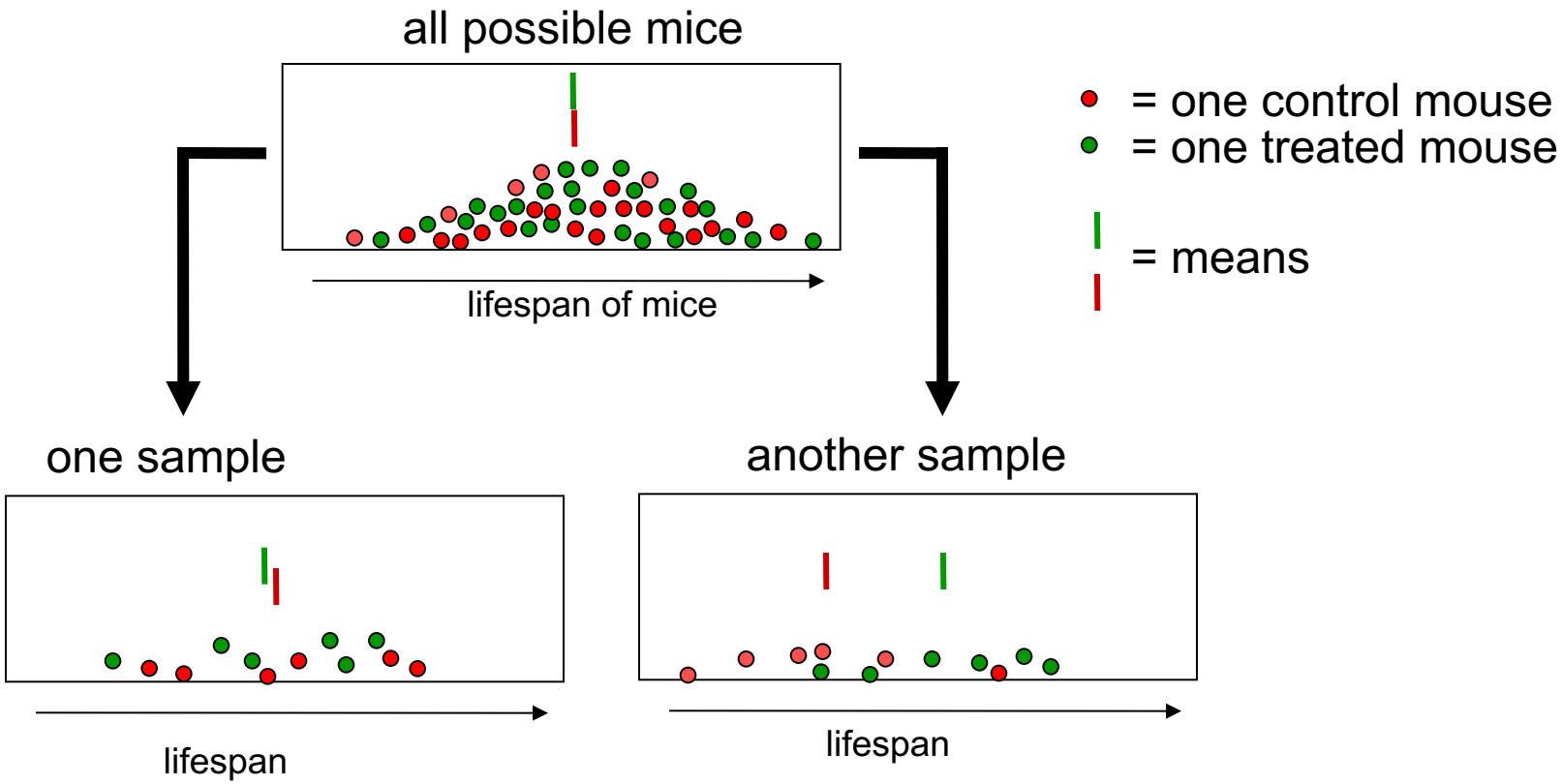
rigorous = research design

USING RIGOROUS RESEARCH METHODS
can greatly reduce the risk of producing false positives

YOU can do something about this !

Validity Trap: **High variation in the sampled population**

By chance, noise sometimes looks like effect when variation is high



no effect of treatment

treatment extends life ! ! ! !

false positive
due to random sampling variation

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

Nature Reviews Neuroscience 14: 365-76, 2013

730 studies in top neuroscience journals: median statistical power = 21%

Statistical POWER:
the expected ability to avoid false negatives, and false positives

- increases with real effect size
- decreases with variation
- increases with sample size

When power is low [~ < 80%] :

- Studies often miss true effects = false negatives

- Published studies are often false positives
[positive results are often just due to ”lucky” sampling]

- When real effects exist, the published effect size is often exaggerated
[only statistically significant when “lucky” sampling gives a misleadingly large effect size]

Validity Trap: Noise via hidden [or ignored] variables

The ALS mouse model is designed to have just 1 independent variable

Lithium treated

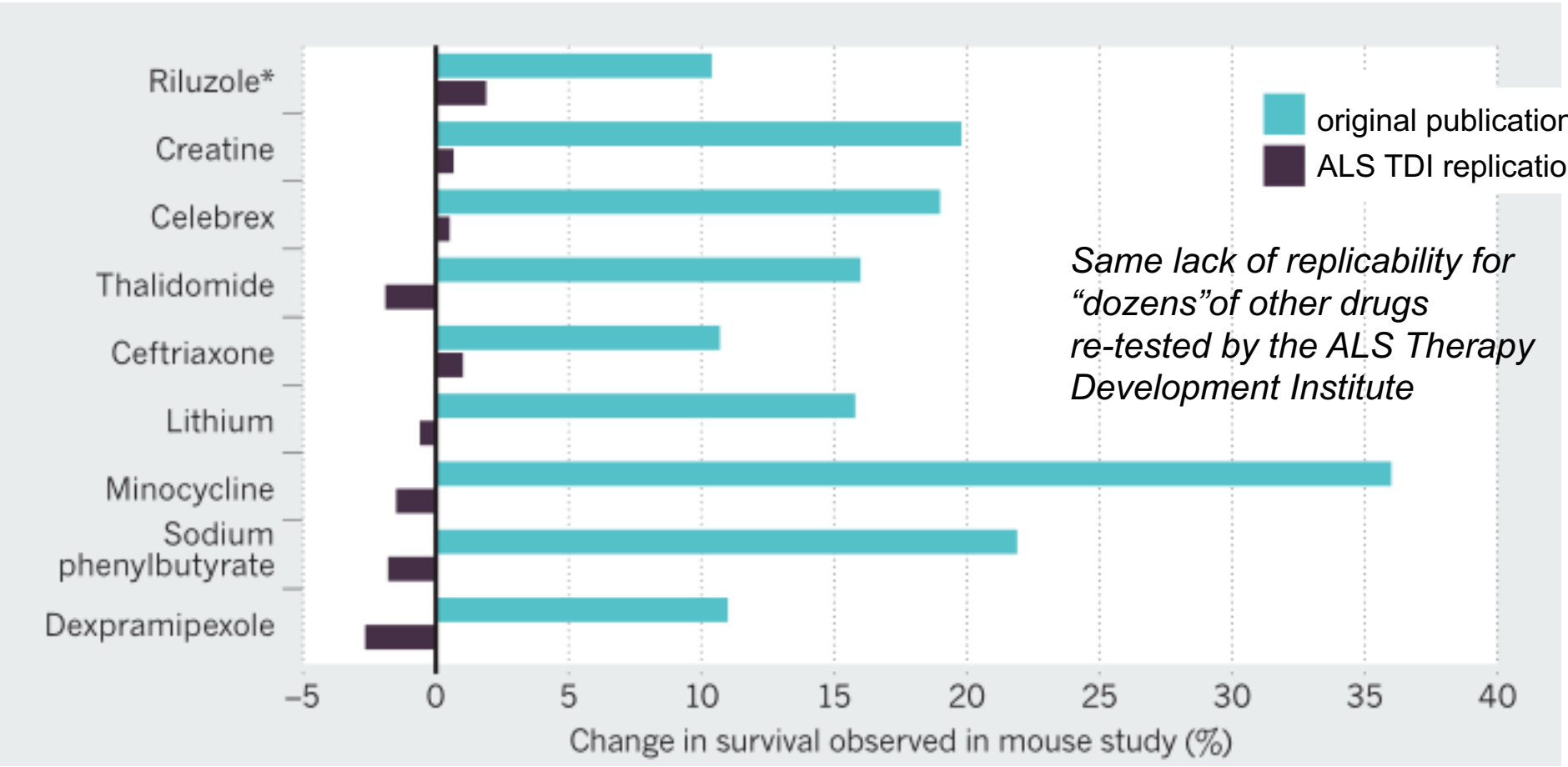
untreated controls

all mice are SOD1 G93A **+**

But what if there is an additional, unrecognized variable
which happens to be unequally distributed between your treated vs control mice,
and affects your measured outcome [e.g. mouse lifetime] ?

Can you see where any additional variables might be hiding?

Published positive results vanish after eliminating hidden variables in the ALS mouse model



Scott S et al., Amyotrophic Lateral Sclerosis 9: 4-15, 2008 and Nature 507: 423-5, 2014

A huge amount of futile effort in clinical trials would have been avoided if the ALS model had been used rigorously, thus weeding out false positives at the pre-clinical stage

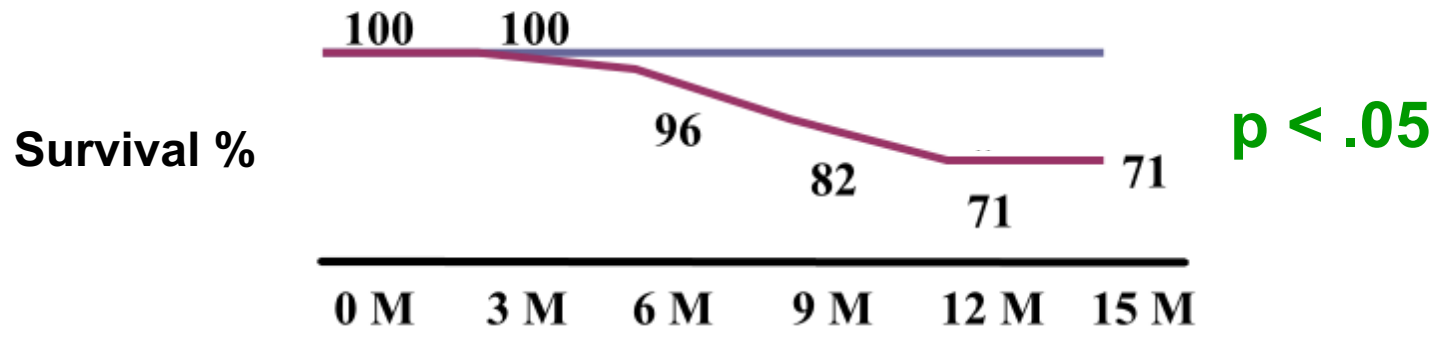
Validity Trap: Over-confidence in NHST p-values as false positive detectors

NHST = null hypothesis statistical testing

How p-values are usually used [but shouldn't be]:

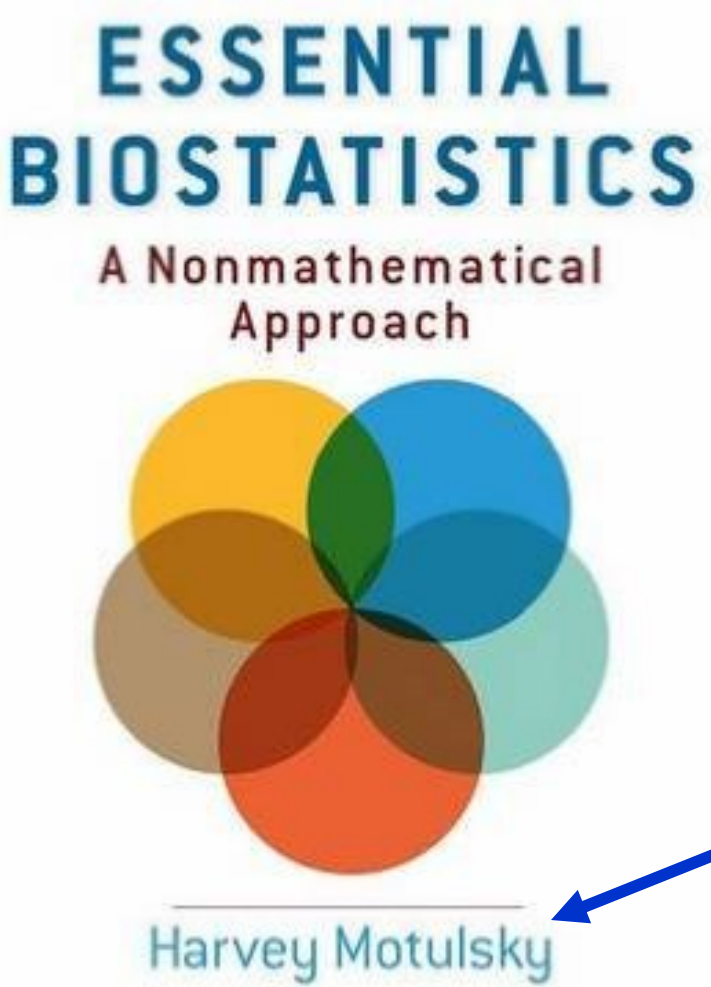
IF p < .05 THEN the observed effect is highly likely to be real and accurate

So there is no need to test for replicability of this experiment



Defenses against over-confidence in NHST p-values, and other aspects of statistical analyses:

READ THIS BOOK !



Accessible,
non-mathematical,
and very useful
explanation of statistical concepts
and applications, including pitfalls

0.9 cm thick

By the founder of GraphPad

Validity Trap: **Using the p-value from 1 experiment to predict the reproducibility of that experiment**

G Cummings Persp. Psychol. Sci. 3: 286-300, 2008

Computer-simulated repeat experiments

Comparing 2 groups of n =32
effect size = 0.5 of standard deviation [real effect of small-ish size]
predicted statistical power = 0.5

10 repeat experiments gave these p-values:

p-value	effect size
.016	0.6
.353	0.2
.001	0.9
.759	0.1
.203	0.3
.609	0.15
.003	0.7
.008	0.65
.153	0.35
.053	0.45

EXACTLY REPLICATED experiments can have wildly differing p-values therefore the p-value of 1 experiment is useless as a predictor of reproducibility

Validity Trap: Ignoring prior probability

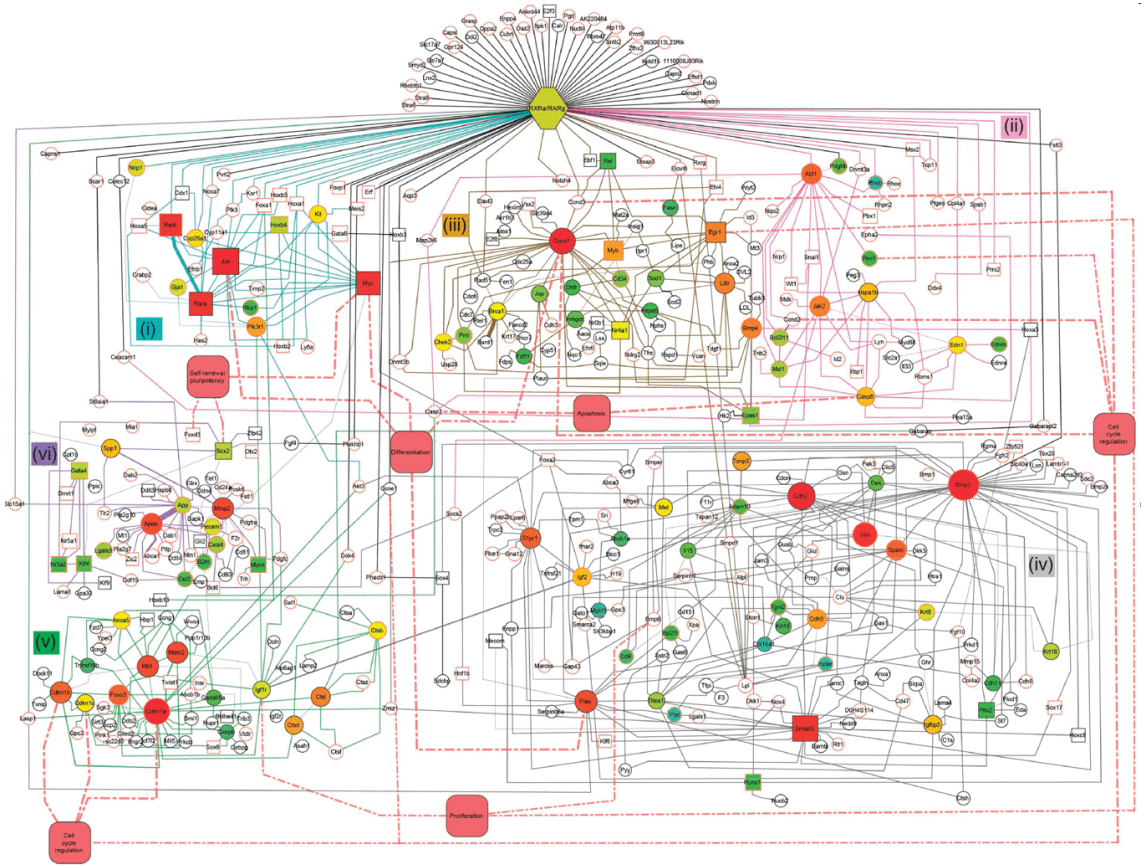
If the prior probability of your hypothesis being true is:	And you do an experiment which gives:	Then the probability of your hypothesis being true increases to:
50%	p = 0.05	71%
5%		11%
50%	p = 0.01	89%
5%		30%

Taking prior probability into consideration is often most useful as a humility exercise.

This process can remind you of how much you don't know about:

- **The biology you are examining**
- **Your experimental system**
- **The hidden uncertainties of using p-values to assess the credibility of experimental results**

Validity Trap: System Complexity



High complexity: Lots of plausible hypotheses, most of which will be wrong

Lots of unknown and unexpected components, mechanisms etc

Lots of evolutionary opportunism= appearances can be deceiving

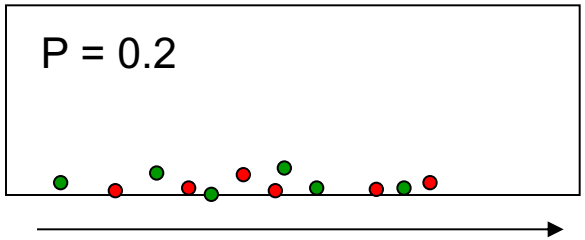
Validity Trap: Cherry Picking

Searching through:
noisy data,
multiple experiments,
statistical tests,
outlier exclusion rules,

etcetera

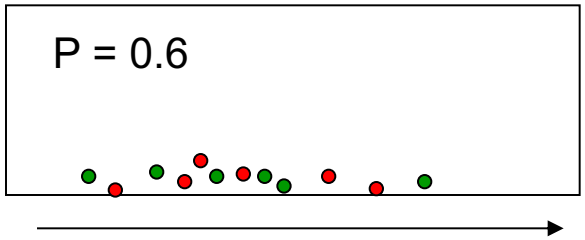
to find something positive

1st experiment



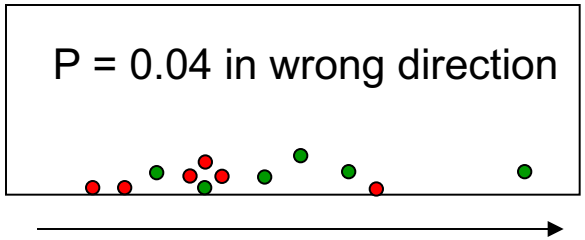
An effect ! and almost significant ! Try again

2nd experiment



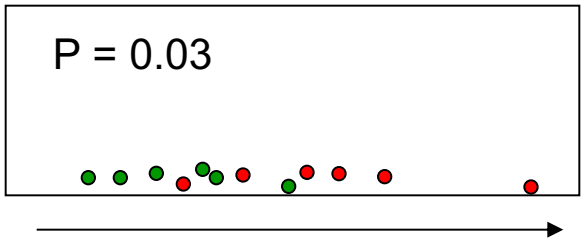
Hmm... Try again

3rd experiment



Obviously a screwy experiment ! So try again

4th experiment



Positive ! Publish !

Scanning the horizon: towards transparent and reproducible neuroimaging research

Russell A. Poldrack[1], Chris I. Baker[2], Joke Durnez[1,3], Krzysztof J. Gorgolewski[1], Paul M. Matthews[4], Marcus R. Munafò[5,6], Thomas E. Nichols[7], Jean-Baptiste Poline[8], Edward Vul[9] and Tal Yarkoni[10]

Nature Reviews Neuroscience 18: 115-126, 2017

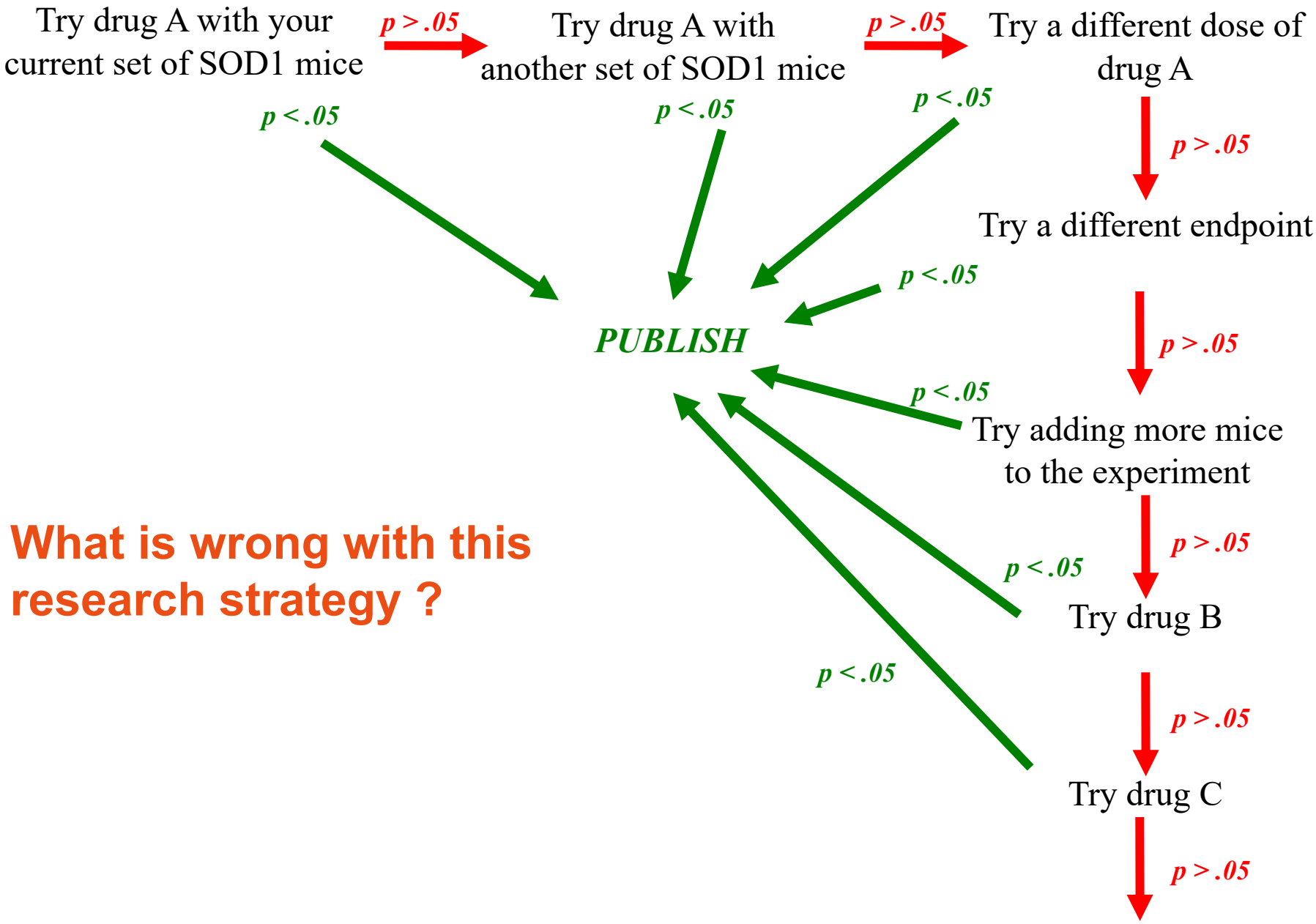
cherry picking opportunities

High data variation and low effect sizes and expensive scans
= studies are usually underpowered

Many brain regions, several patient groupings etc

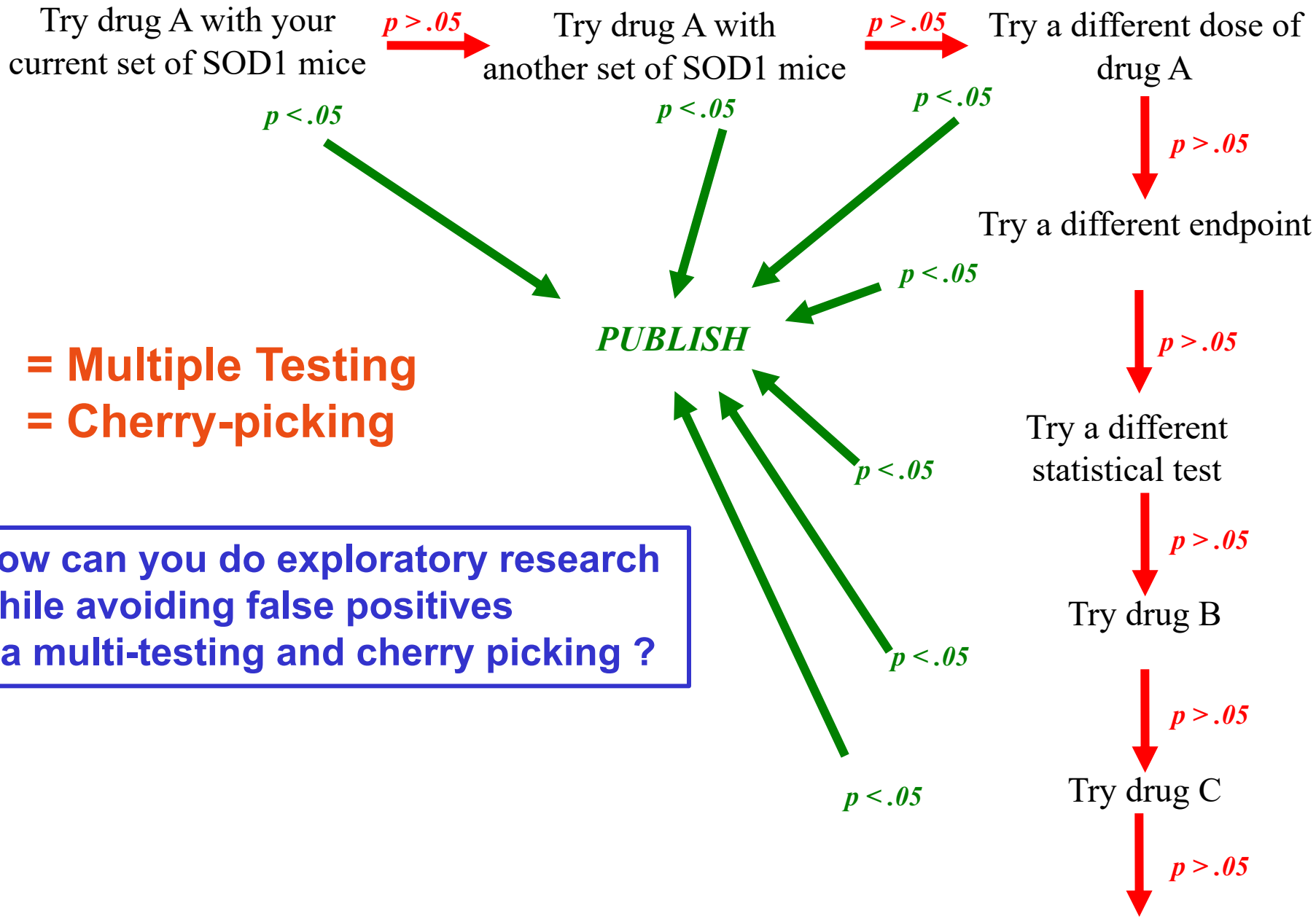
Many options in data analysis
= potential over-fitting of analysis to the data
= finding the analysis that produces a positive

Hypothesizing after results known = cherry-picking the hypothesis that fits the data



What is wrong with this research strategy ?

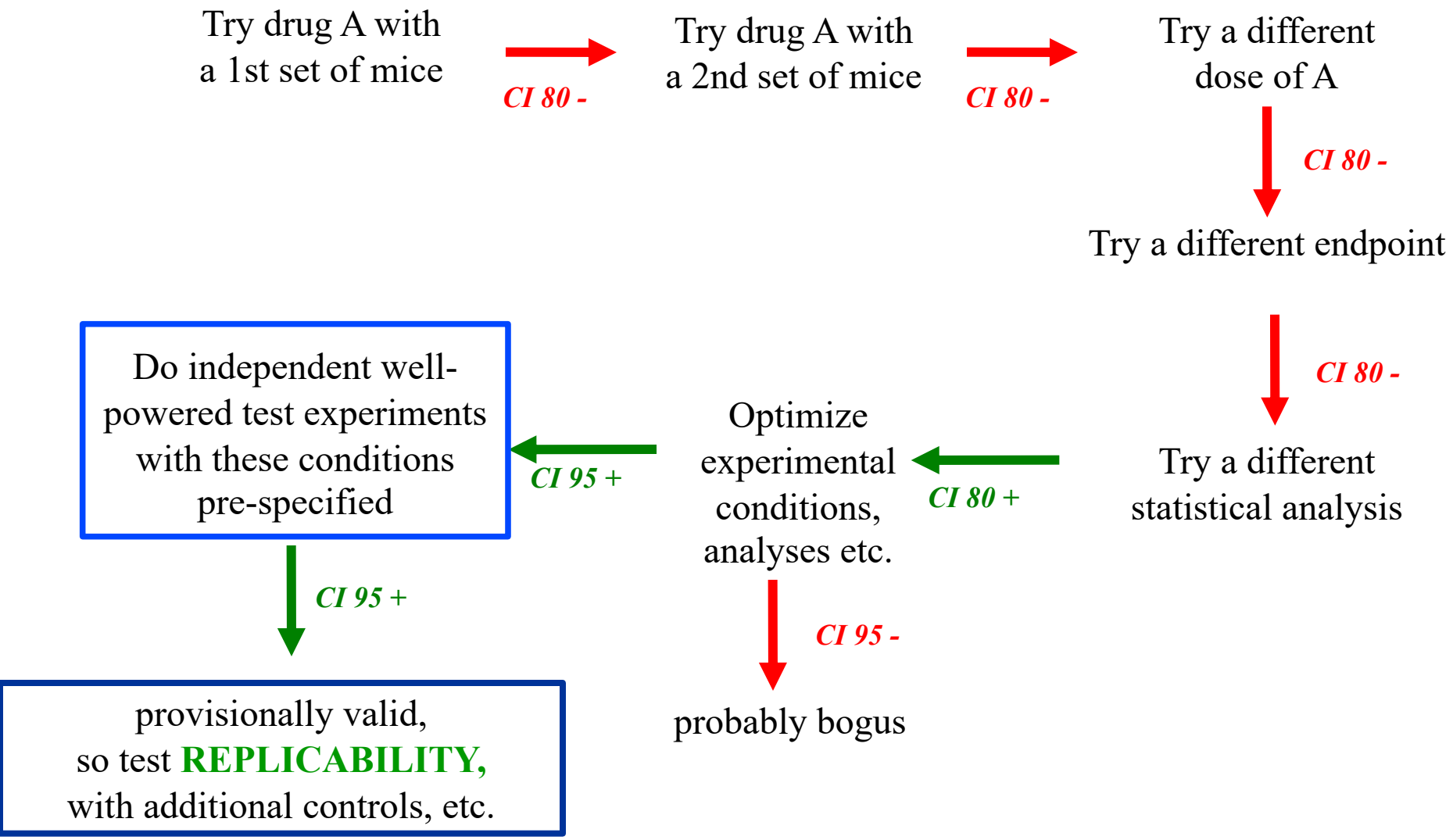
Validity Trap: **Research Programs that Pursue until Positive, then Publish**



= Multiple Testing
= Cherry-picking

How can you do exploratory research while avoiding false positives via multi-testing and cherry picking ?

Explore, Optimize, Pre-specify, Test, Replicate :
The effective defense against hidden variables, underpowering, weak statistical test thresholds, multiple testing and a lot more



Validity trap: Failure to rigorously and sceptically test replication

Small scale, unspecified and incompletely reported single clinical trial of lithium for ALS

Hypothesis -generating exploration

Low Rigour

Done once

Large scale, pre-specified, well-described clinical trials of lithium for ALS

	Study design	Study size	Outcome
Aggarwal et al (2010)[14]	Sequential, time-to-event, futility	88	Stopped early (mean duration 5·4 months) because futility boundary (p=0·68) was crossed
Chio et al (2010)[15]	Single blind	171	Stopped early by data monitoring and ethics committee because 117 patients discontinued
Miller et al (2011)[17]	Historical controls, unmasked	107	No benefit of lithium therapy
Wicks et al (2011)[18]	Observational using self-reported patient data	447	No benefit of lithium therapy
Verstraete et al (2012)[16]	Randomised sequential	133	No benefit of lithium therapy

Hypothesis-testing

High Rigour

Reproducibility tested collectively

PLOS | COMPUTATIONAL BIOLOGY 2015

EDITORIAL

Ten Simple Rules for Reducing Overoptimistic Reporting in Methodological Computational Research

Anne-Laure Boulesteix*

Emphasis on avoiding False Positives

1 Assess rigorously

2 Compare rigorously

3 Consider enough datasets

4 No dataset fishing

5 No free lunch

6 Consider several criteria

7 Validate with independent data

8 Use demanding simulations

9 Provide all information

10 Learn about what can go wrong and how to avoid it

Validity Trap: Over-fitting

Search for and optimize only on the dataset that works best with your method

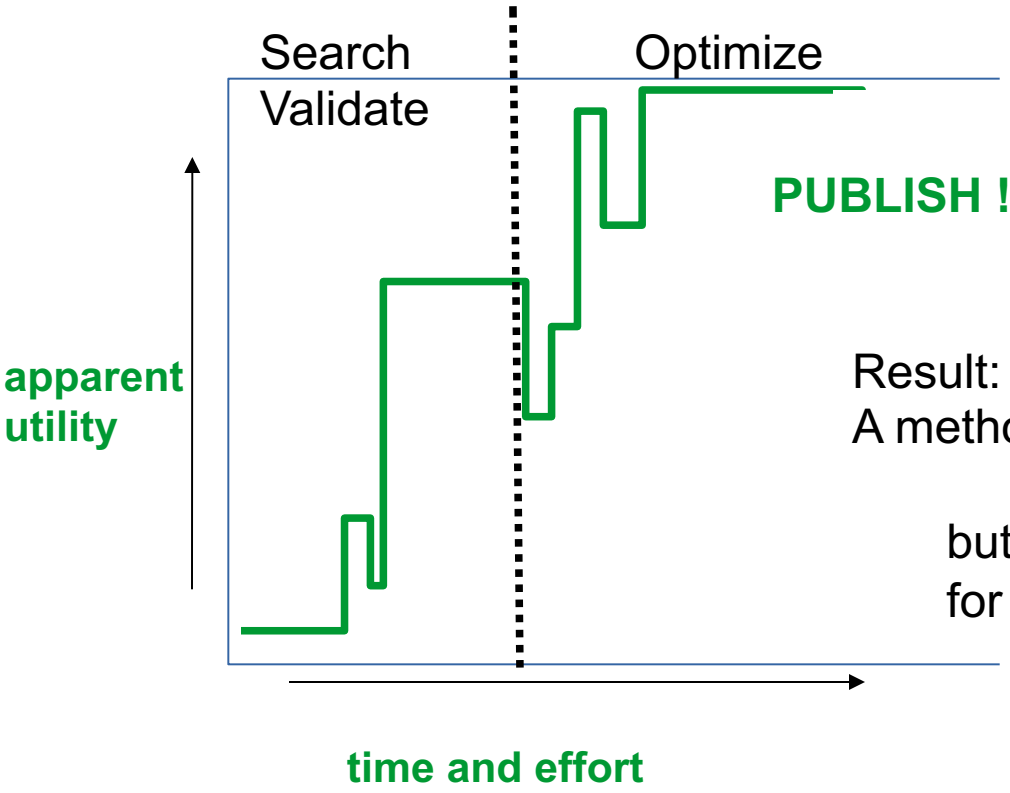
or

Search for and optimize on settings within your method/database etc. chosen because they make your method work well

or

Search for a benchmark, e.g. an existing alternative method, that is chosen because it is weak compared to your method

= CHERRY-PICKING



See:
Jelizarow M et al., Over-optimism in bioinformatics: an illustration, Bioinformatics 26: 1990-1998, 2010

Result:
A method that appears to be highly effective,

but may be ineffective
for most potential applications

Validity trap in investing:
Advisors who overfit their technical analyses

Predictive Markers: West Fraser Timber **decline**d in 2015 and then remained below a falling trend-line (dotted line) and its falling 40-week moving average (40wMA) for another year. The stock then settled into a horizontal trading range (dashed lines). The recent rise to $49.57 (C) brought the stock above the falling trend-line, above its 40wMA and above the large trading range.

Prognosis: This signals targets of $57 and $62. **Buy**

Irreproducibility in technical analyses

West Fraser Timber **Jan 7, 2017**

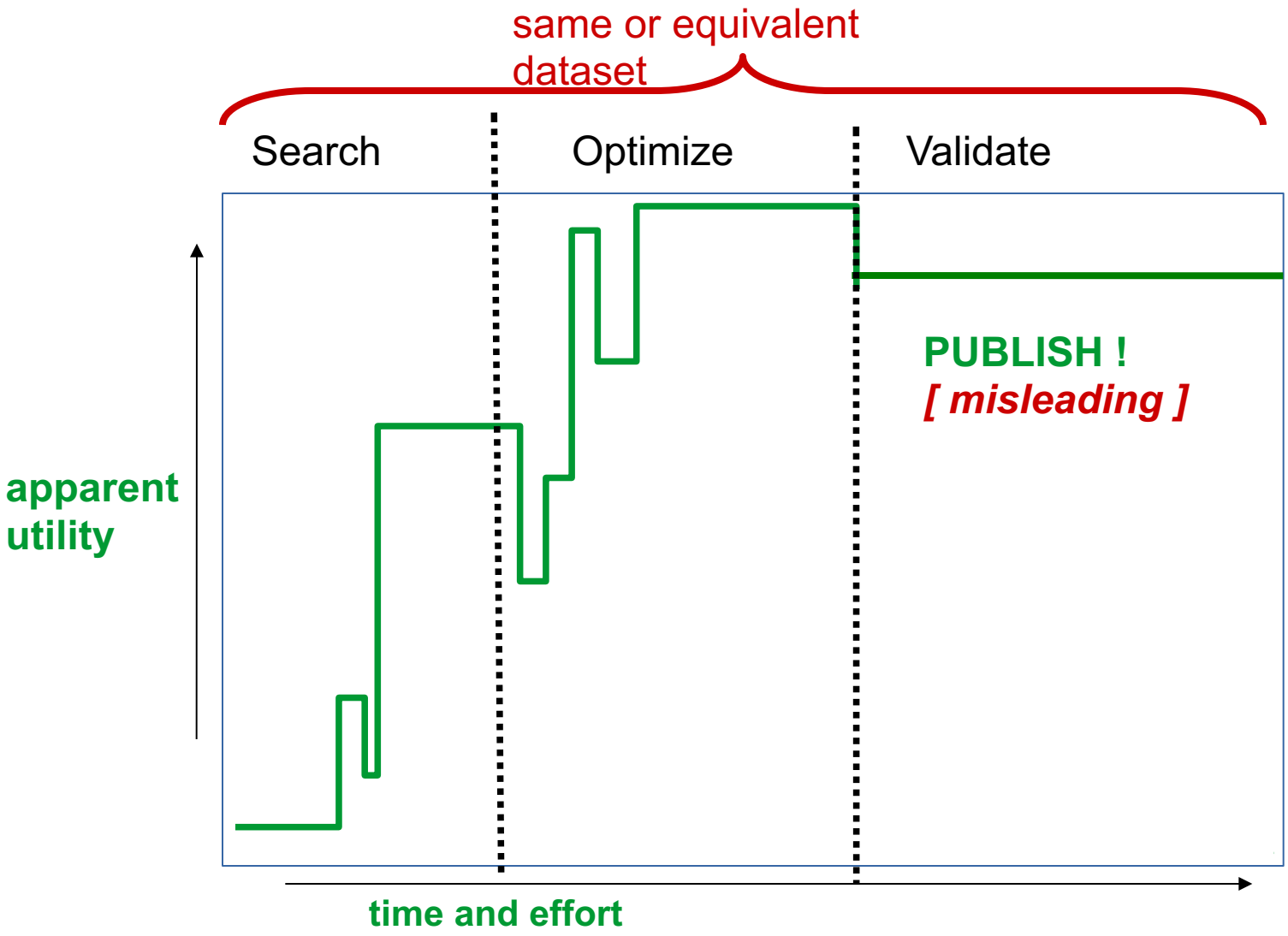
Technical Indicators »

Name	Value	Action
RSI(14)	33.694	Sell
STOCH(9,6)	10.043	Oversold
STOCHRSI(14)	17.161	Oversold
MACD(12,26)	-0.510	Sell
ADX(14)	17.788	Neutral
Williams %R	-92.949	Oversold
CCI(14)	-109.6730	Sell
ATR(14)	0.2914	Less Volatility
Highs/Lows(14)	-0.2486	Sell
Ultimate Oscillator	28.833	Oversold
ROC	-1.303	Sell
Bull/Bear Power(13)	-0.8940	Sell

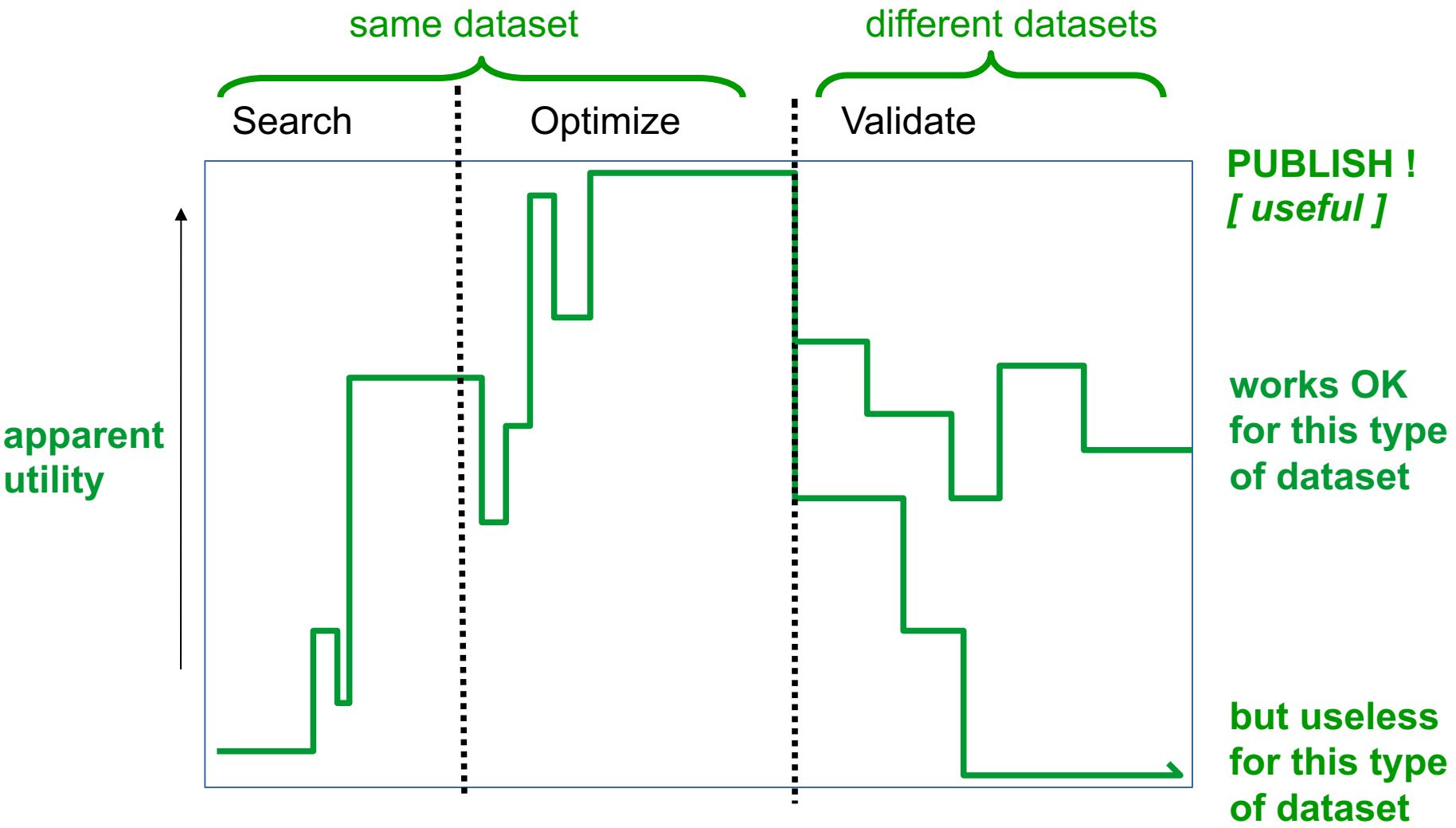
Buy: 0 Sell: 6 Neutral: 5

Summary: STRONG SELL

Validity Trap: Straw-man Validation



If the method is validated on a dataset that is equivalent to the dataset used to search and optimize,
the method may not be as real-world useful as it appears



If the method is validated on datasets
- that are independent of the dataset used to search and optimize,
- and are representative of real-world uses,

THEN the effectiveness and limitations of the method will be evident

Publish all of the informative validation results
so the utility of the method can be assessed by readers of your paper

PLOS | COMPUTATIONAL BIOLOGY

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, Eivind Hovig

Published: October 24, 2013 • https://doi.org/10.1371/journal.pcbi.1003285

Rule 1: For Every Result, Keep Track of How It Was Produced

Rule 2: Avoid Manual Data Manipulation Steps

Rule 3: Archive the Exact Versions of All External Programs Used

Rule 4: Version Control All Custom Scripts

Rule 5: Record All Intermediate Results, When Possible in Standardized Formats

Emphasis on reproducibility of output in the hands of other users

etcetera

Validity Trap: Motivational Biases

Motivational Biases via pursuit of truth and benevolence

I want to get positive, high impact results because:

This will expand the boundaries of knowledge

This may help others

Motivational Biases via personal benefit

I want to get positive, high impact results because:

I will be famous and admired

I will get promoted

I can get lucrative and prestigious consultancies

I will survive in the grants/tenure market
and thus can continue to do excellent research

Defense against motivational biases

= rational planning around a deliberate goal

My primary goal: ***to make important discoveries*** **that are TRUE**

To achieve this I will:

- **Use rigorous methods**
- **Carefully search for potential validity traps**
- **Suppress motivational biases that conflict with my goal**

Why might a scientist pursue this rational plan

and yet still produce an avoidable false positive ?

THINKING, FAST AND SLOW

DANIEL KAHNEMAN

Cognitive System 2

Slow

Deliberative

Uses reasoning

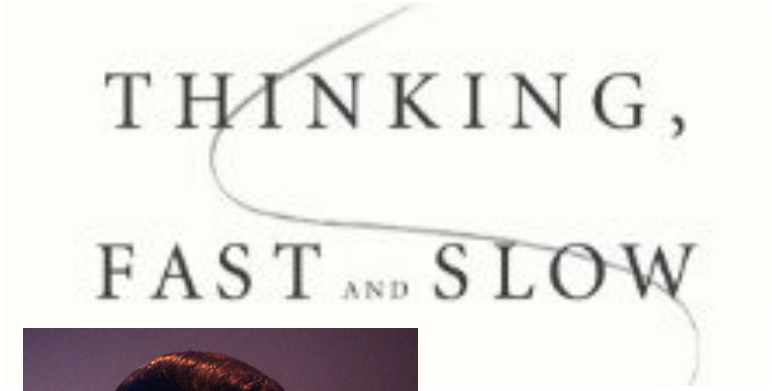
Conscious

Requires activation

Experiment

Interpretation

Conclusions



DANIEL
KAHNEMAN

Cognitive System 2

RATIONAL

Conscious

Deliberative

Requires activation and effort

Experiment

Interpretation

Conclusions

Cognitive System 1

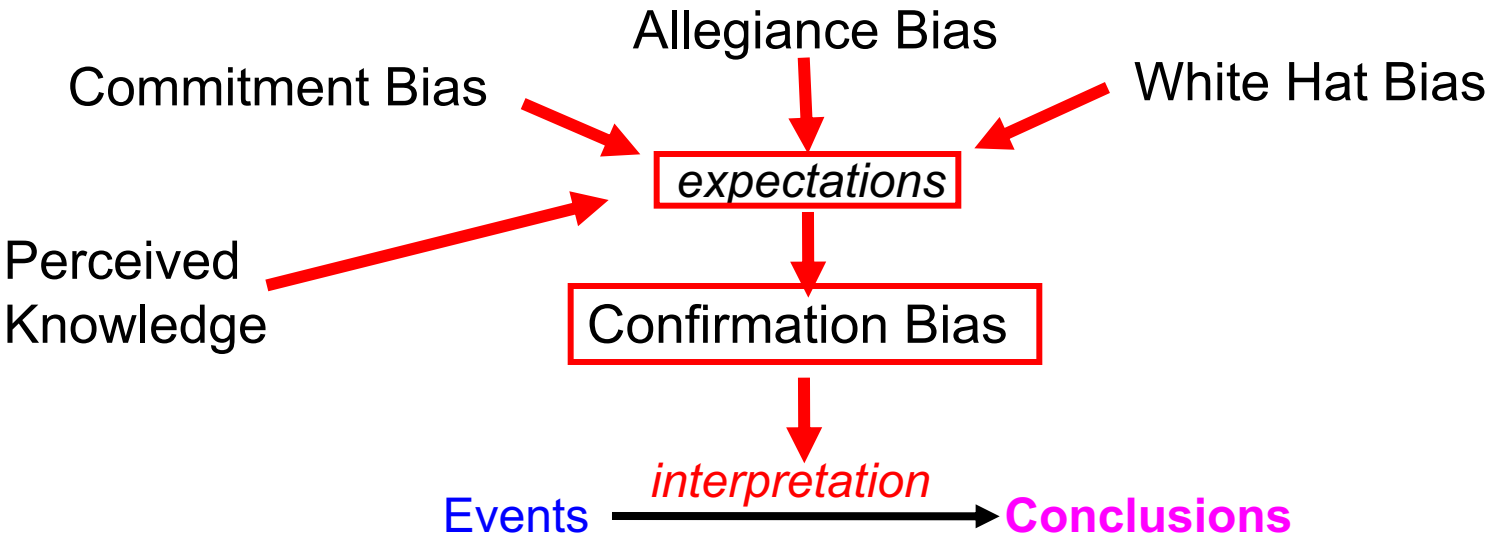
RATIONALIZING

Sub-conscious

Uses heuristics and biases

Automatic, easy

Validity Trap: Sub-Conscious Cognitive Biases



If the conclusion meets my expectations, I'm content

- Accept it

If the conclusion contradicts my expectations, I'm dissatisfied

- Reject it and keep searching

Defending your Research against your Confirmation Bias

YOU WON'T

unless you're FORCED TO

- **Make yourself part of a team**
- **Discuss your research as it proceeds**
- **Seek constructive criticism**
- **Understand the methods you use**
- **Put up obstacles to cherry-picking**

To prevent sub-conscious cherry-picking:

<u>Blind experiments and analyses</u>

unless…

There is minimal risk of researcher bias affecting the experiment or analysis

or

Blinding will seriously impede the research

To tie your cherry-picking hands:

Pre-specify

- hypothesis being tested
- experimental methods
- data acceptance criteria
- data analyses
- etcetera

You might have to adjust your specifications
as you learn about and optimize your experimental system……

But when it comes time to seriously test your hypothesis - pre-specify if you can

Transparent Pre-specification:
- Tell everyone in your lab that you're pre-specifying
- Clearly state in your paper exactly what and how and when you pre-specified

Reducing the risk of false positives
by pre-empting your confirmation and commitment biases:

Pre-mortem your proposed research design !

***Imagine* you have already done a proposed experiment and published it but it subsequently turns out to be a false positive:**

- What are the various ways this might have occurred?

- Which of these are most likely, given your research design?

- How could the research design be changed to protect against this?