Statistical evidence 101

D. Blasi

U Zürich & MPI SHH

Ars Statistica

The anathema on P-values

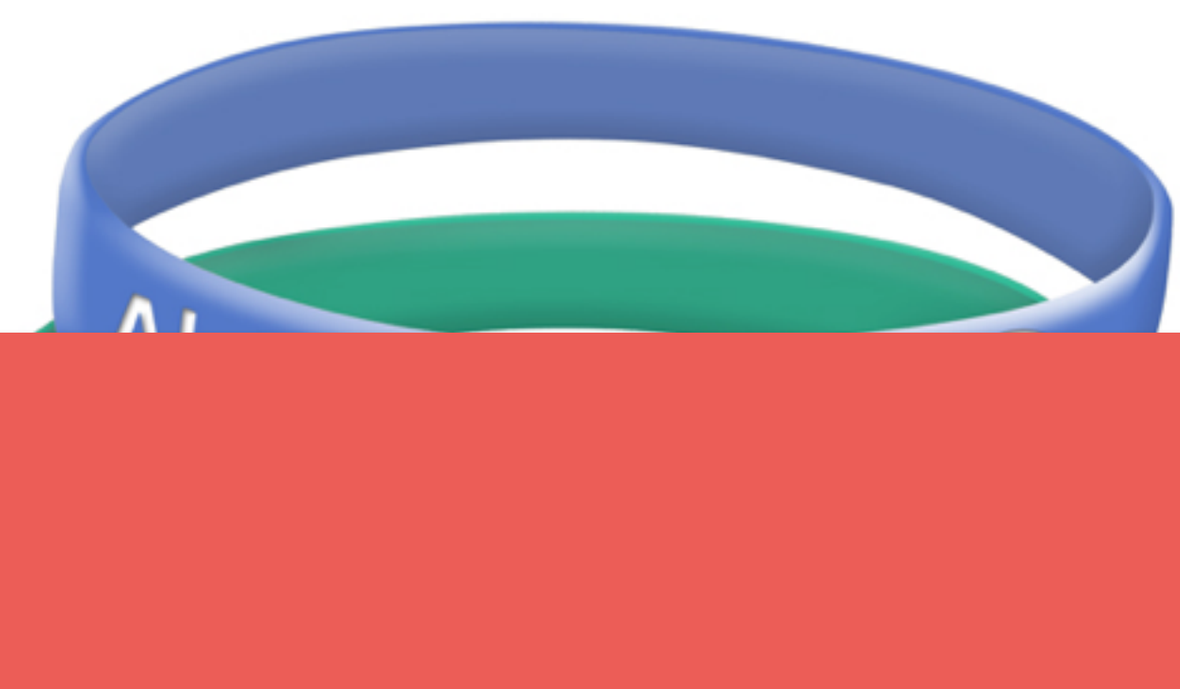
Using your (visual) brain

Learning to learn

On fishing and mining

Ars Statistica

Ars Statistica



These things suck

2

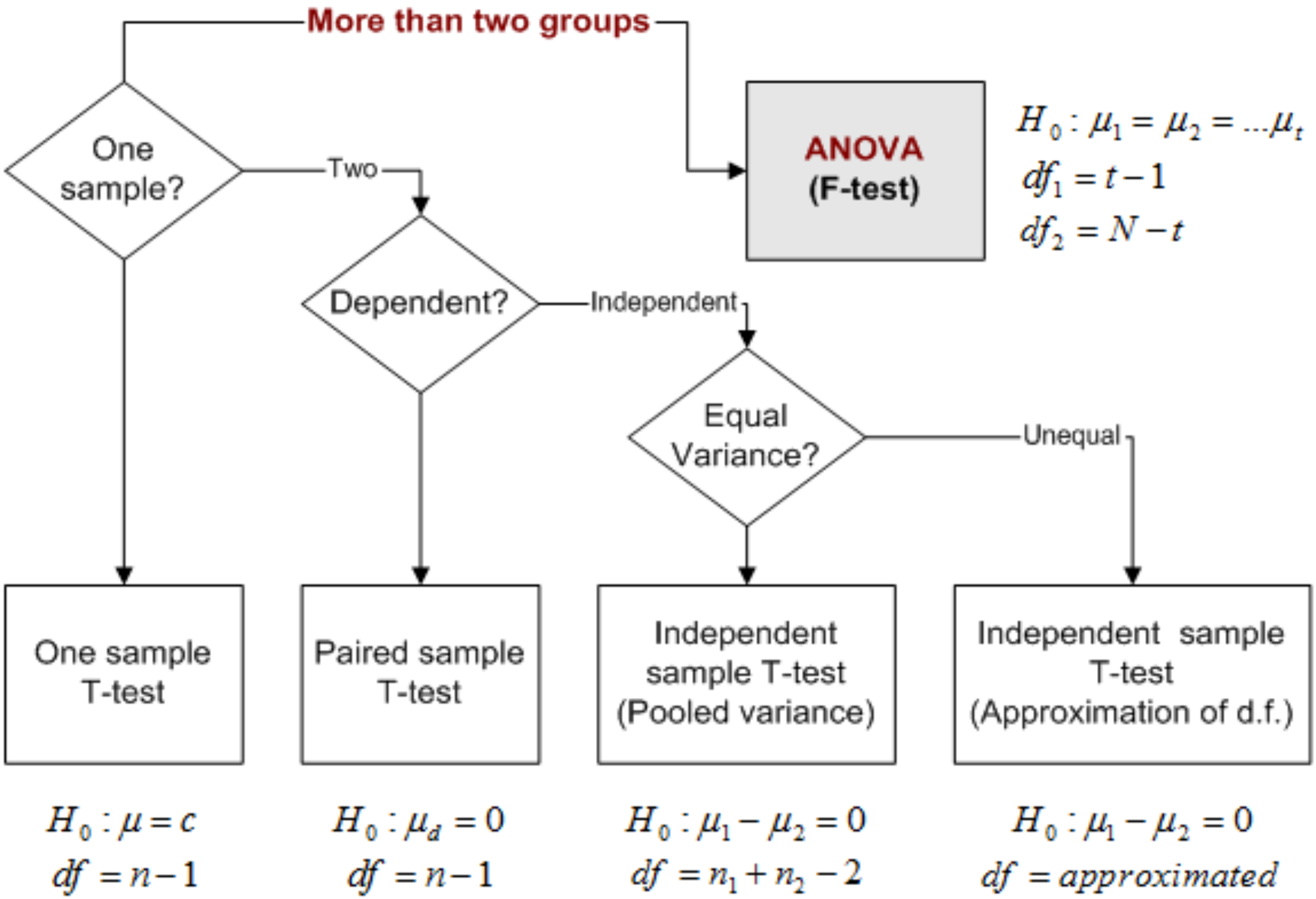
statistics

The study of percentages, bars, graphs, and charts, all in an attempt to make some sort of logical conclusion out of a bunch of numbers so that even more percentages, bars, graphs, and charts can be made.

*Statistics may not be of practical use in everyday life, unless you own a **casino**.*

by AYB July 29, 2003

Ars Statistica



Cancer and Smoking

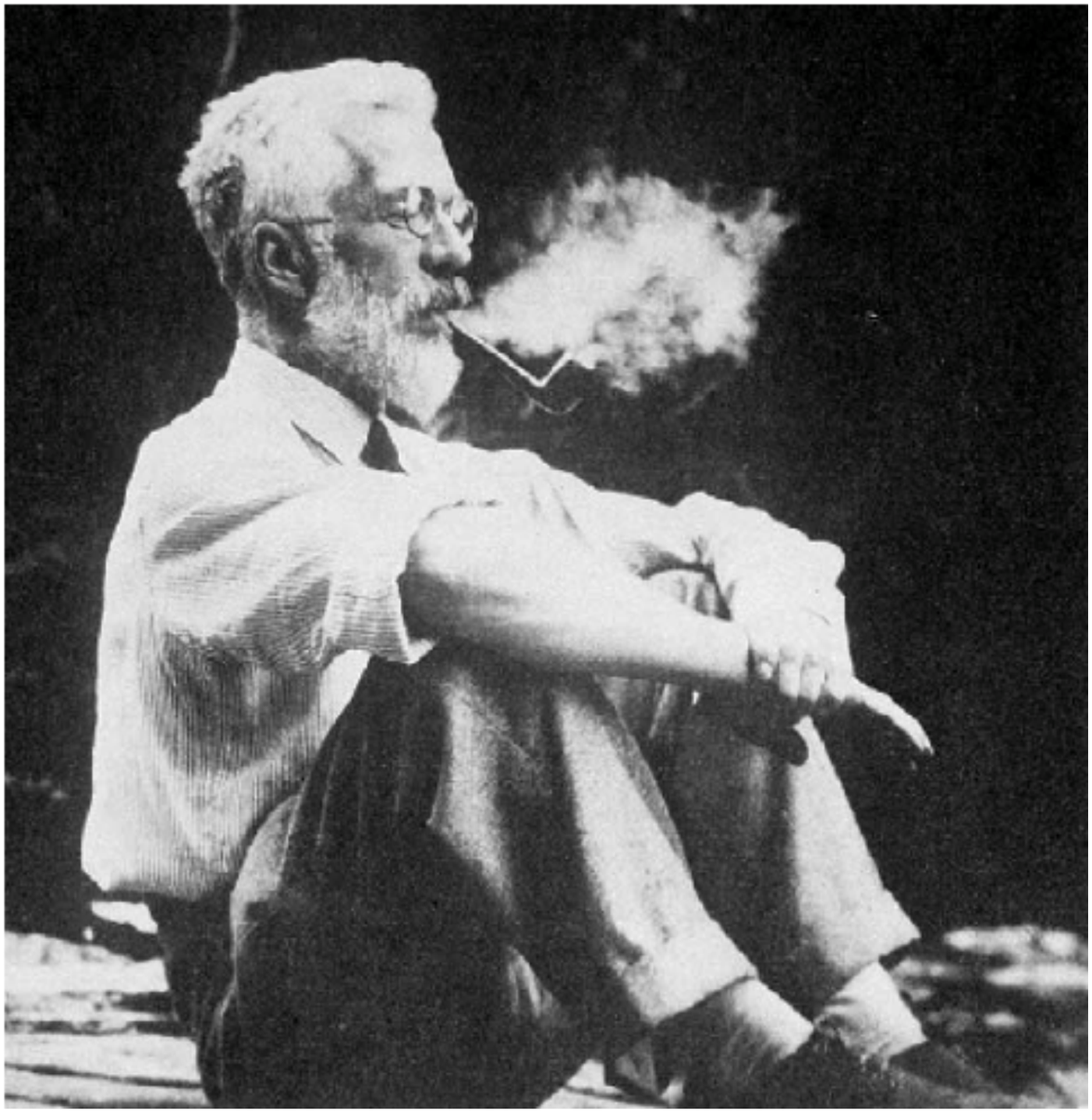
THE curious associations with lung cancer found in relation to smoking habits do not, in the minds of some of us, lend themselves easily to the simple conclusion that the products of combustion reaching the surface of the bronchus induce, though after a long interval, the development of a cancer. If, for example, it were possible to infer that smoking cigarettes is a cause of this disease, it would equally be possible to infer on exactly similar grounds that inhaling cigarette smoke was a practice of considerable prophylactic value in preventing the disease, for the practice of inhaling is rarer among patients with cancer of the lung than with others.

Nature (1950)

Ars Statistica

Ronald Fisher

Ars Statistica



Sometimes the assumptions and the statements of (early) statistical methods look way *too* complex to have anything to do with the real world.

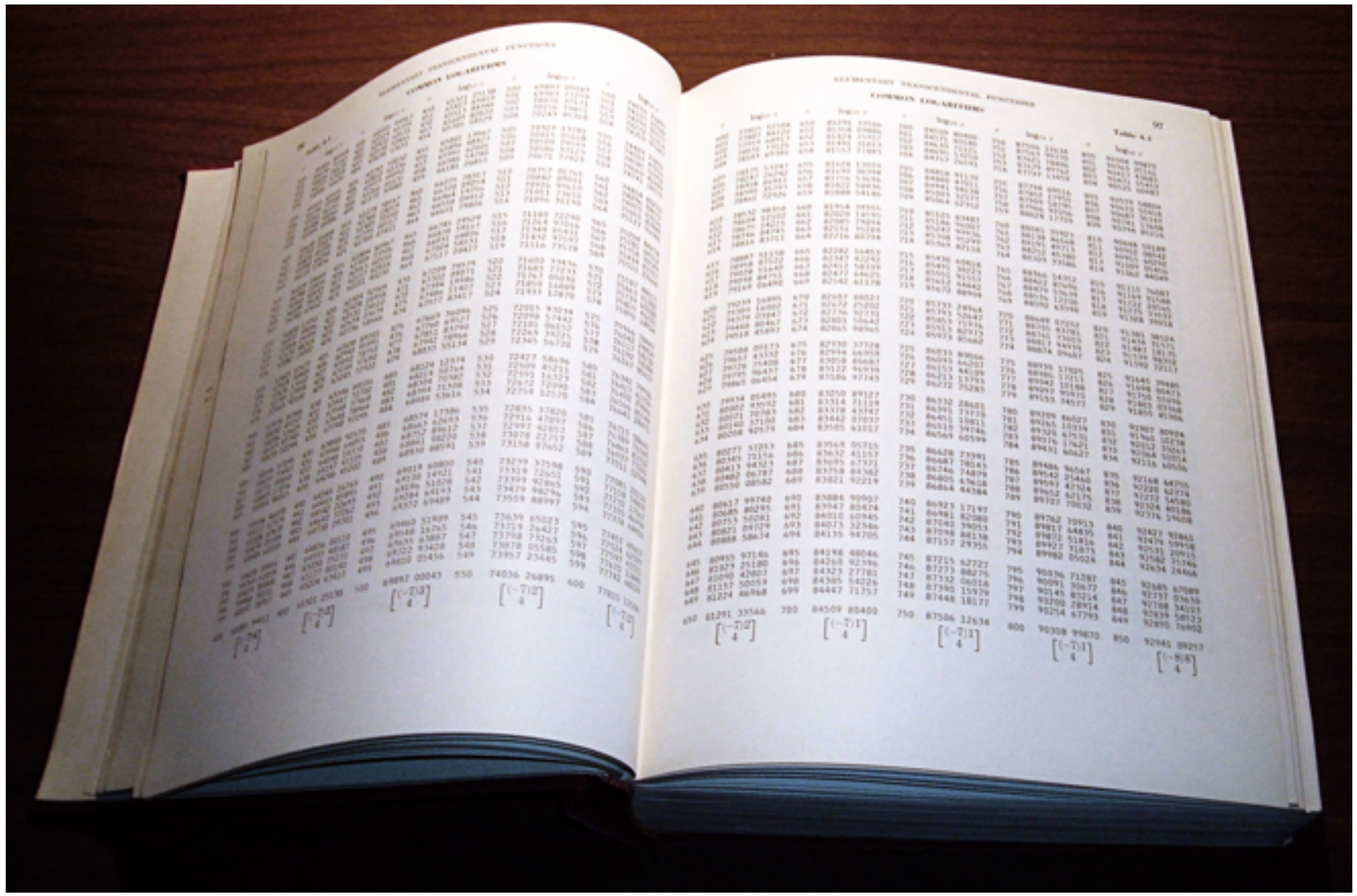
E.g. assumptions behind a basic linear regression:

homoskedastic,
independent,
normally distributed data

No data I know in linguistics/ anthropology/cognitive sciences satisfy this!

Ars Statistica

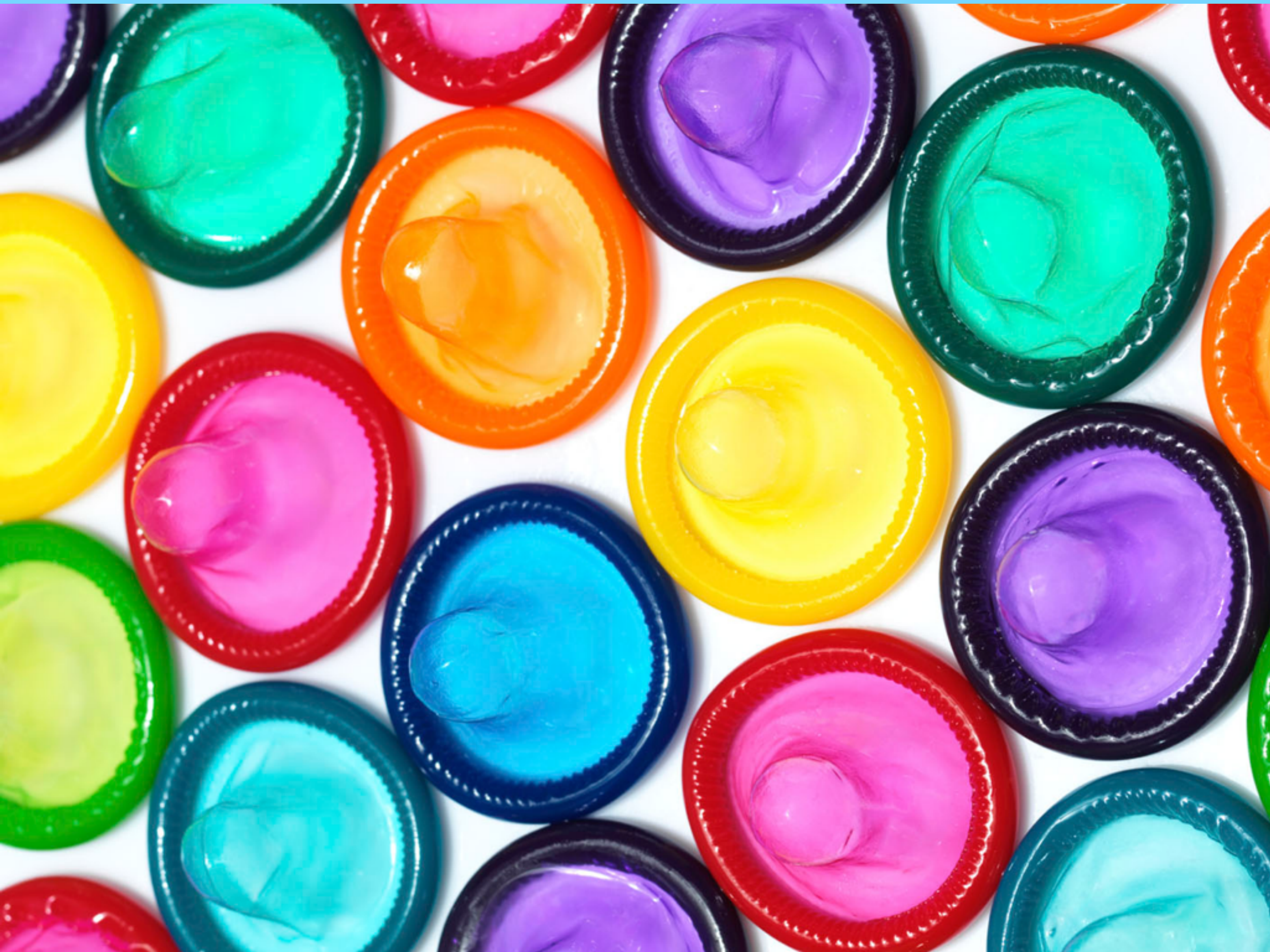
These assumptions were motivated by sheer convenience



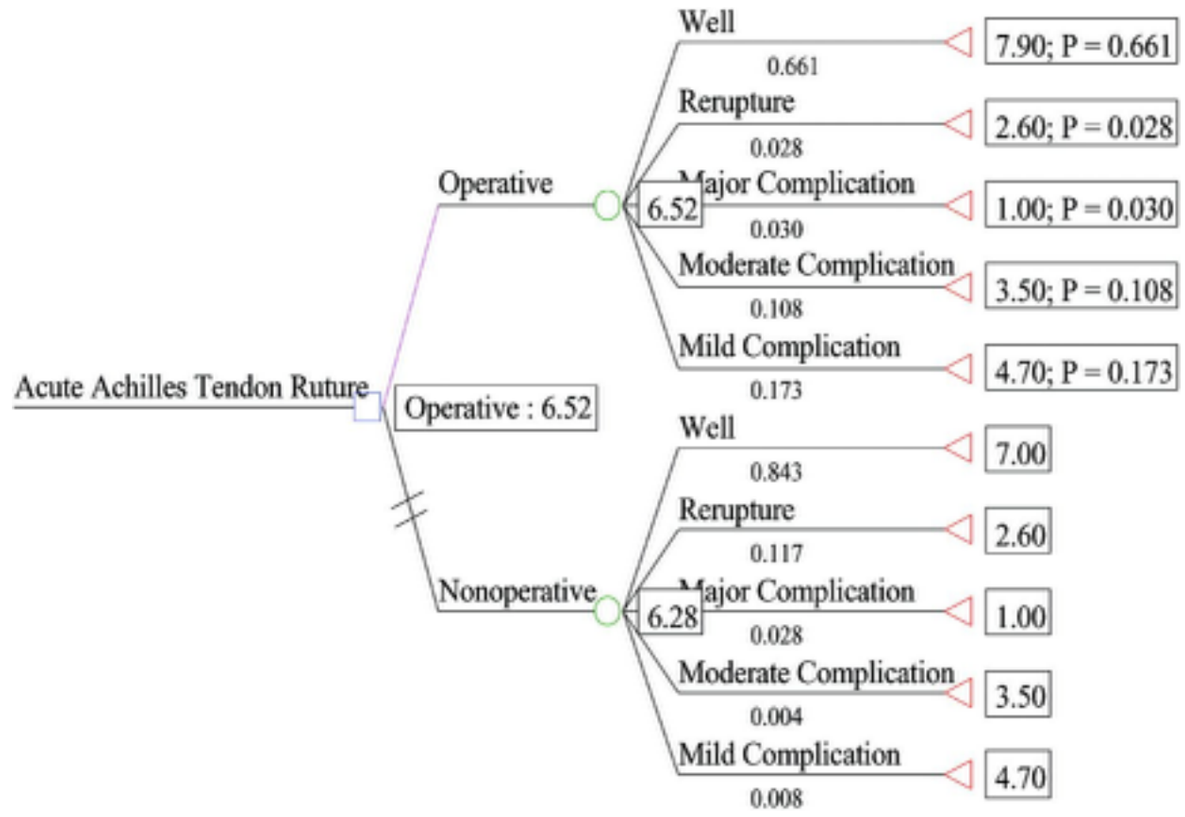
Ars Statistica

George Barnard

The development of early methods was **very** practically oriented



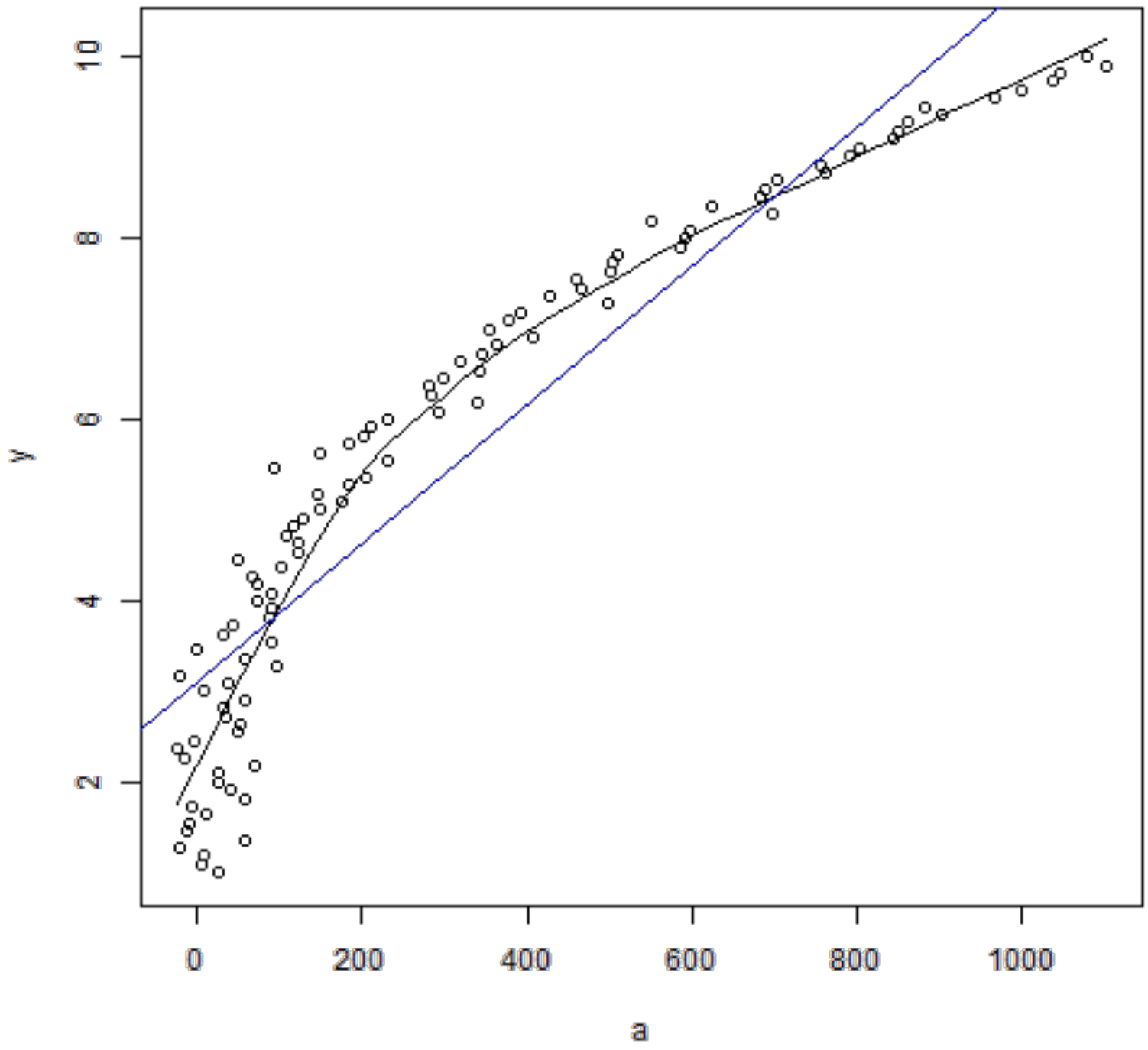
The usual statistics course in the social and health sciences is a result of the needs of **decision taking**



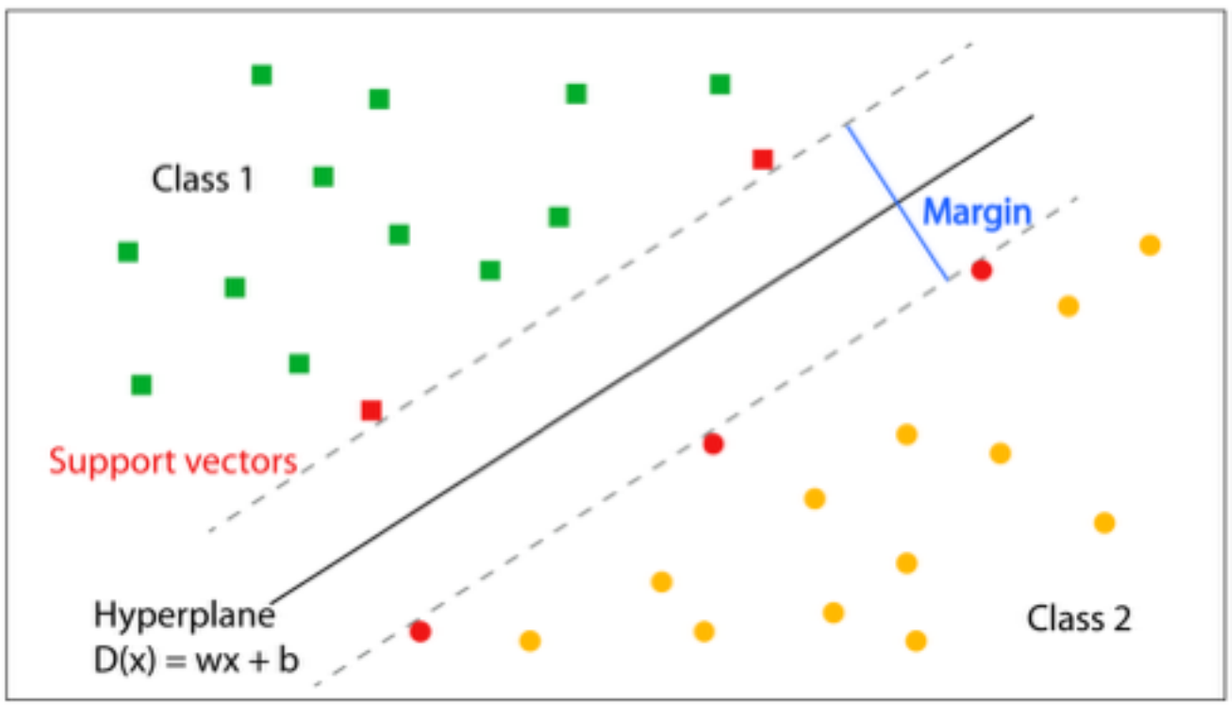
Things have changed quite a lot thanks to computers.

Now we can attack very complex problems without using Procrustean techniques.

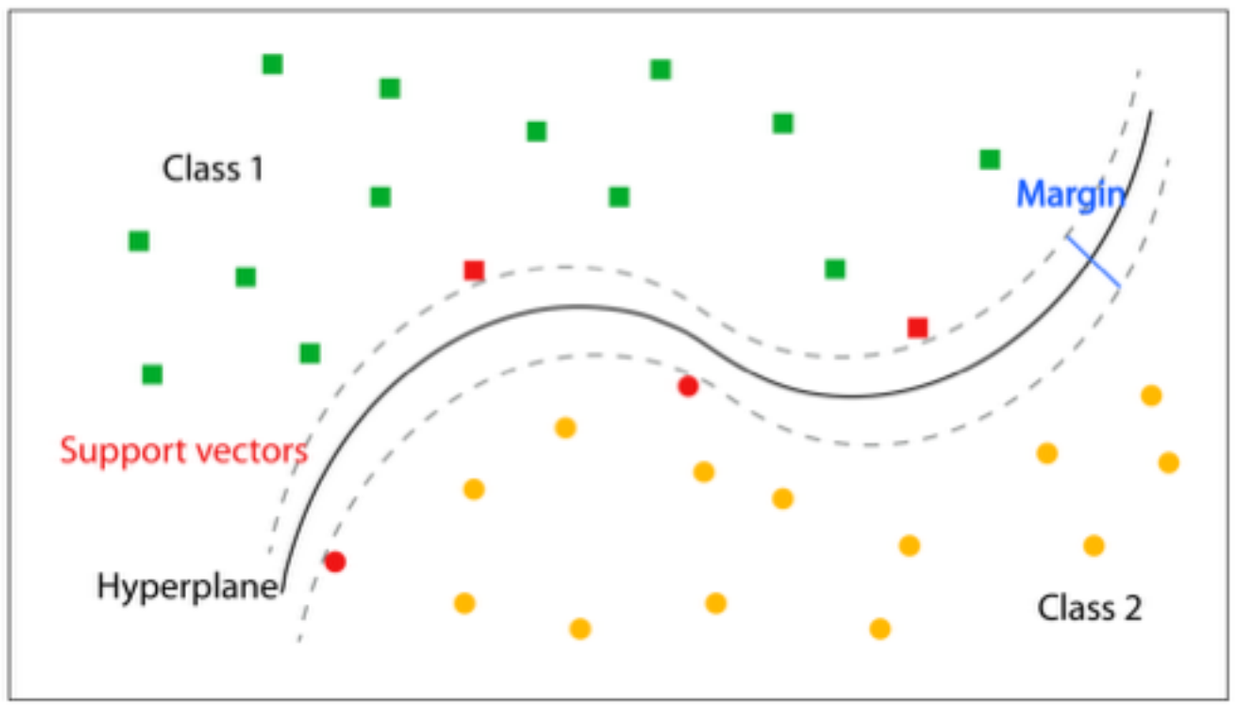
Ars Statistica



A. Linear separation



B. Non-linear separation



Why do we still use the old stuff?

Convenience

Laziness / ignorance

The anathema on P-values

The anathema on P-values

Since you have all read the tutorial (or you know about stats) you know that the P-value is the probability of finding an effect as large (or small) as the observed one **provided that** the null hypothesis is true

The anathema on P-values

Everyone Uses P-Values, But No One Knows What They Are

—By ***Kevin Drum*** | Mon Mar. 7, 2016 1:01 PM EST

Experts issue warning on problems with P values

Misunderstandings about common statistical test damage science and society

BY **TOM SIEGFRIED** 10:30AM, MARCH 11, 2016

NATURE | **NEWS**

Statisticians issue warning over misuse of *P* values

Policy statement aims to halt missteps in the quest for certainty.

Monya Baker

The anathema on P-values

null hypothesis statistical testing

or

statistical hypothesis inference testing ?

The anathema on P-values

null hypothesis statistical testing

or

Statistical **H**ypothesis **I**nference **T**esting ?

The anathema on P-values

Is the P-value useful at all?

It depends. The first condition is that you need to have a useful/probable/ reasonable null hypothesis

The anathema on P-values

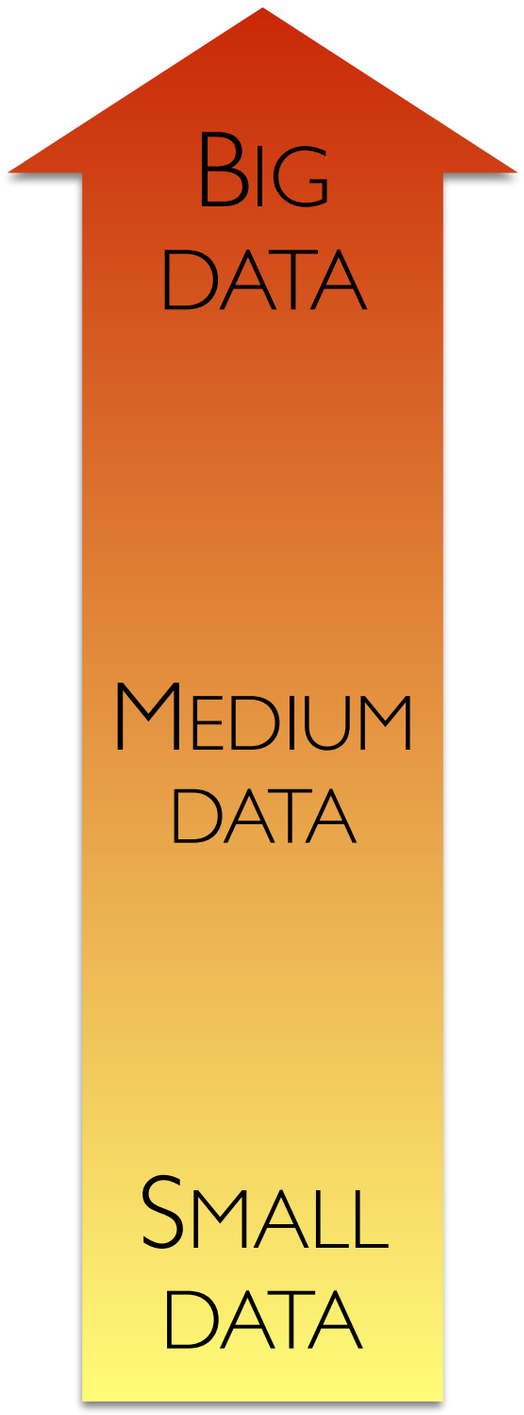
Are All Economic Hypotheses False?*

J. Bradford De Long
Department of Economics
Harvard University
Cambridge, MA 02138
and
National Bureau of Economic Research

and

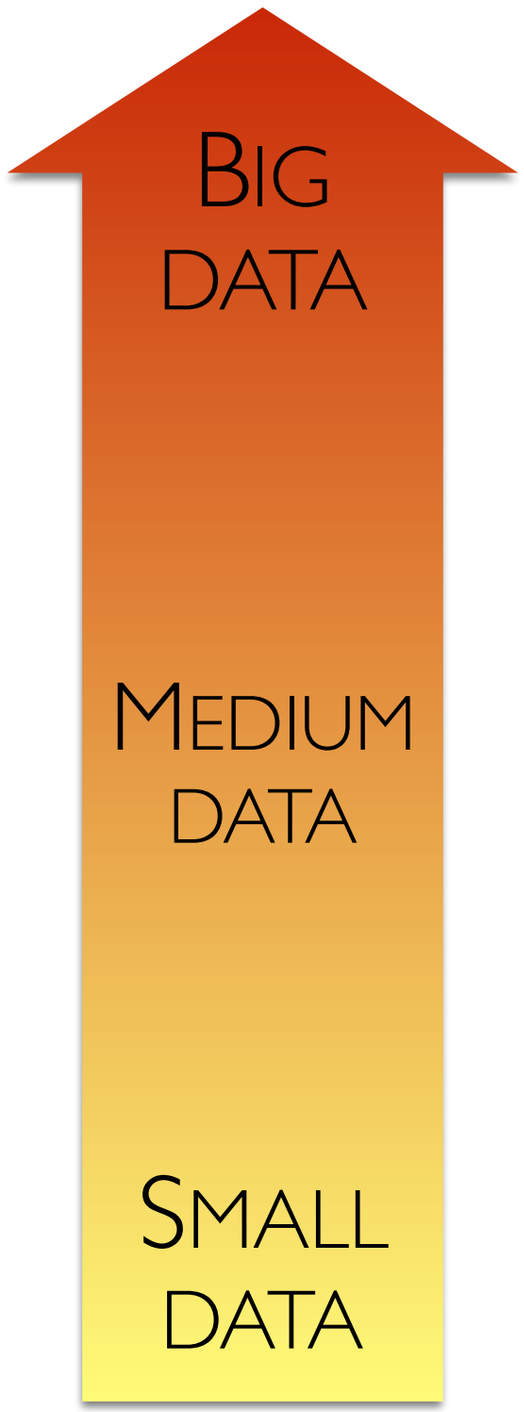
Kevin Lang
Department of Economics
Boston University
270 Bay State Road
Boston, MA 02215
and
National Bureau of Economic Research

The anathema on P-values



Crucially, the practical interpretation of P-values changes with the amount of data

The anathema on P-values



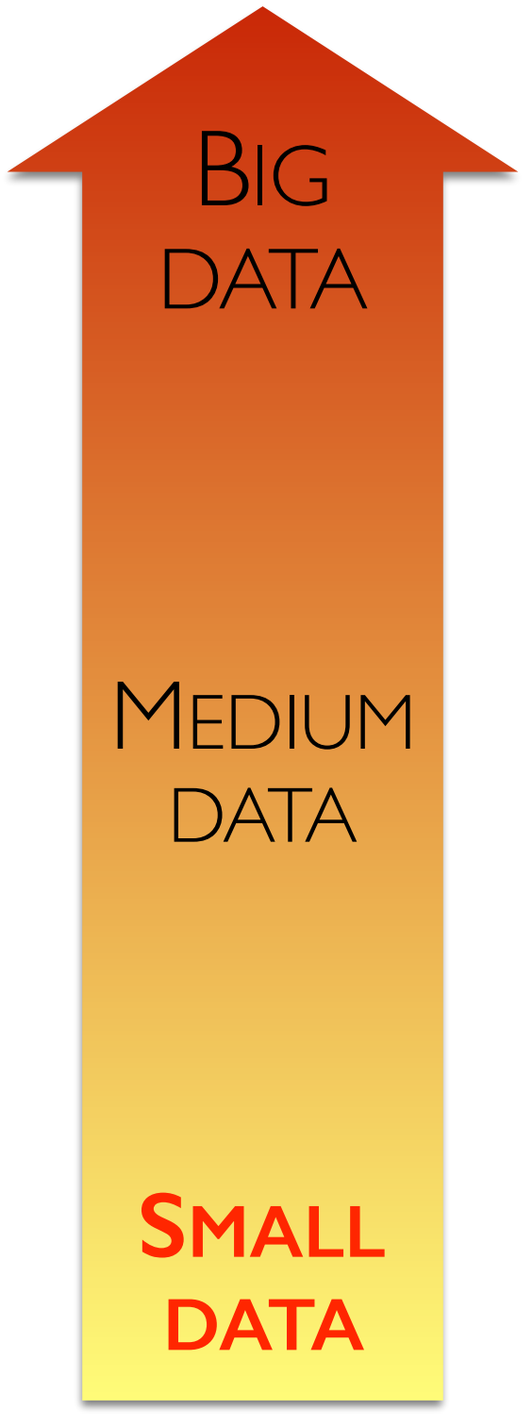
How large your data are depends on the models and the effects you want to test. A **CRUDE AND DEFINITIVELY NOT GENERAL** rule of thumb would be:

Small data: less than 20 datapoints per parameter

Medium data: between 20 and 3000 datapoints per parameter

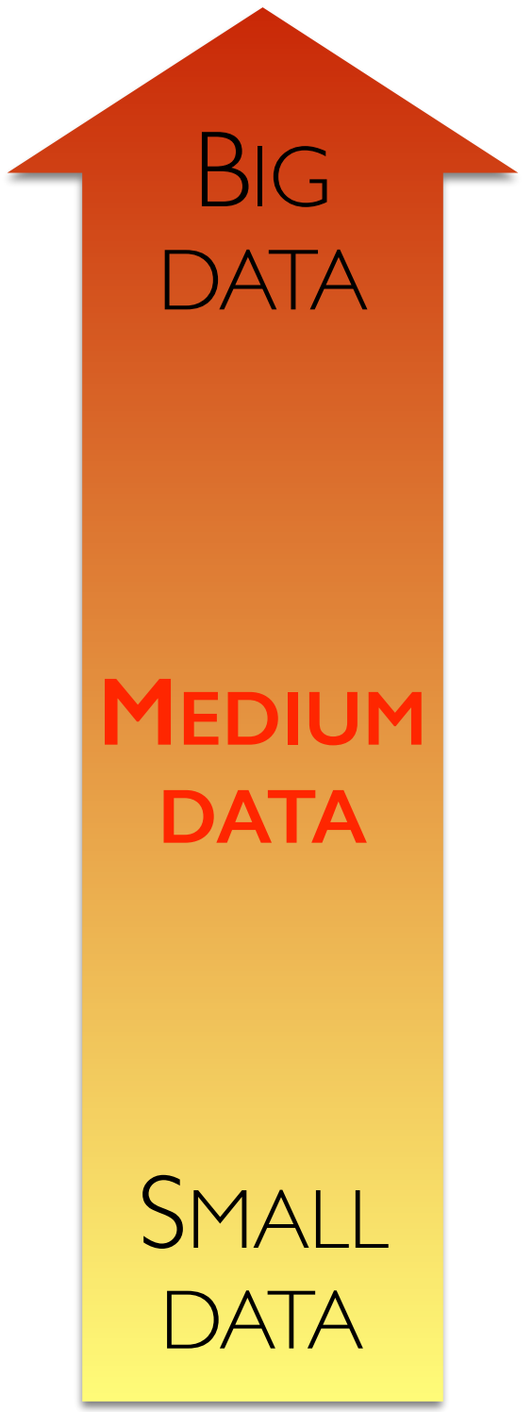
Big data: more than 3000 datapoints per parameter

The anathema on P-values



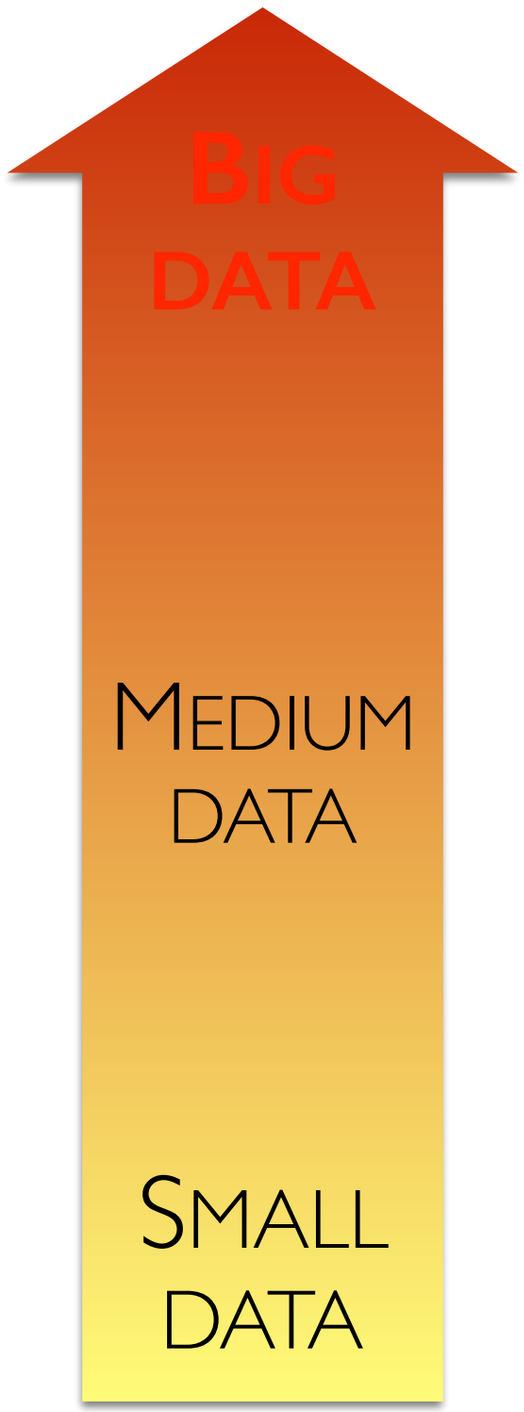
With small amounts of data, the P-value might be unable to distinguish the null hypothesis from alternative interesting models (TYPE II ERROR). So, potentially informative if small but proceed with caution when large.

The anathema on P-values



With a reasonable amount of data, the P-value gets its usual interpretation: it helps deciding whether to keep or discard the null hypothesis.

The anathema on P-values



Unless you strongly believe data might follow the exact null hypothesis stipulated, the P-value is not very useful (and most likely it will be ~0)

The anathema on P-values

Assessing statistical significance is simply the first, often uninformative or trivial step.

The essence of any thorough statistical analysis is the discussion of **effect sizes**, or how strong/weak your inferred statistics are and how do they refer to the data .

The anathema on P-values

A good effect size has

Interpretable scale

It comes with some measure of uncertainty

It allows you, in principle, to make predictions about unobserved/hypothetical cases

The anathema on P-values

More general advise:
if you *really* care about
comparing your data against
the null, then you should
probably go Bayesian
(tomorrow you'll learn how!)

Using your (visual) brain

Using your (visual) brain

Visualize your data.

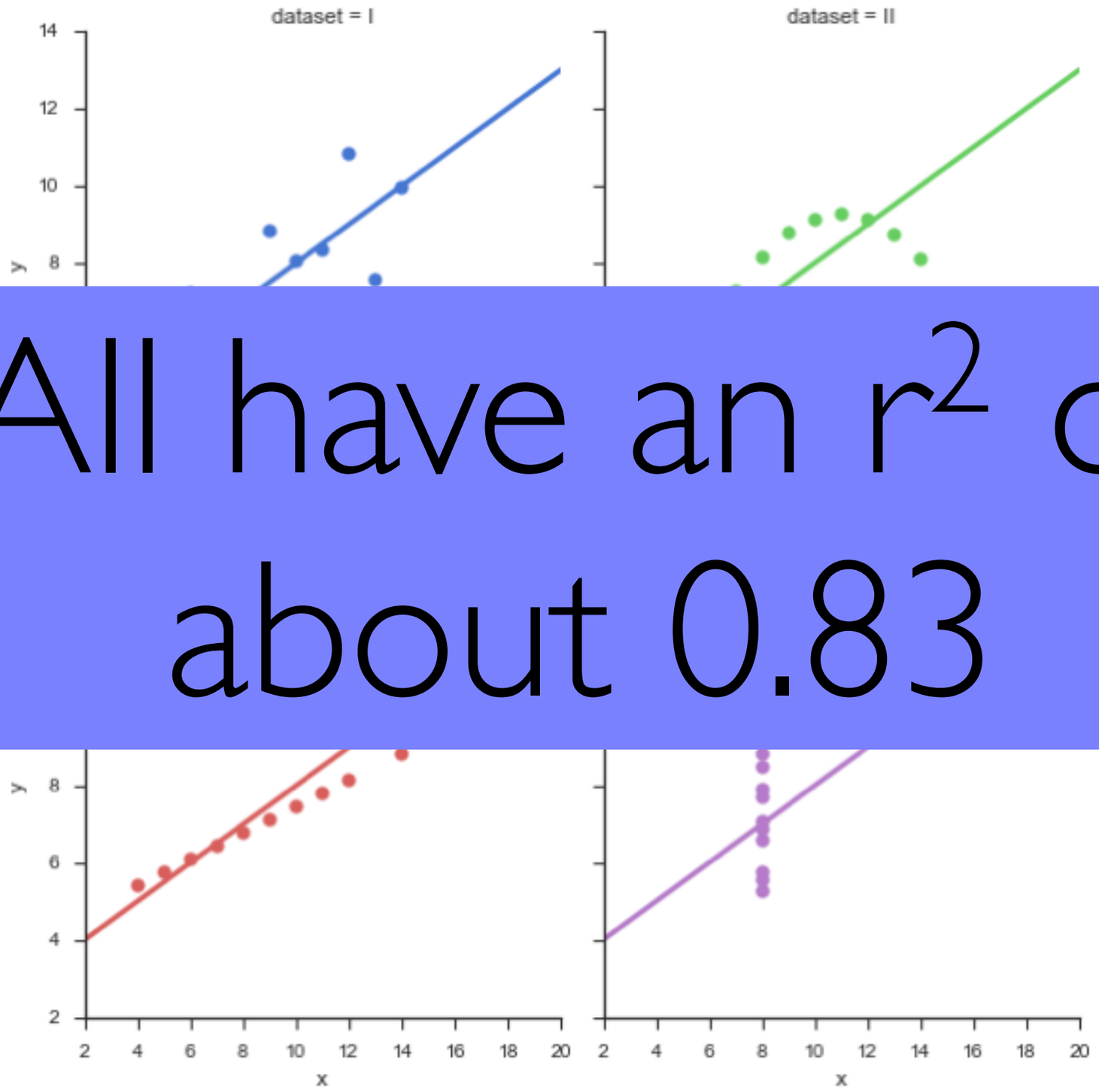
Your brain is at the same time a bag of biases but also a gigantic heuristic toolkit.

Using your (visual) brain

We rely a lot on bunches of numbers - the statistics - but they all imply some loss of information with respect to the raw data.

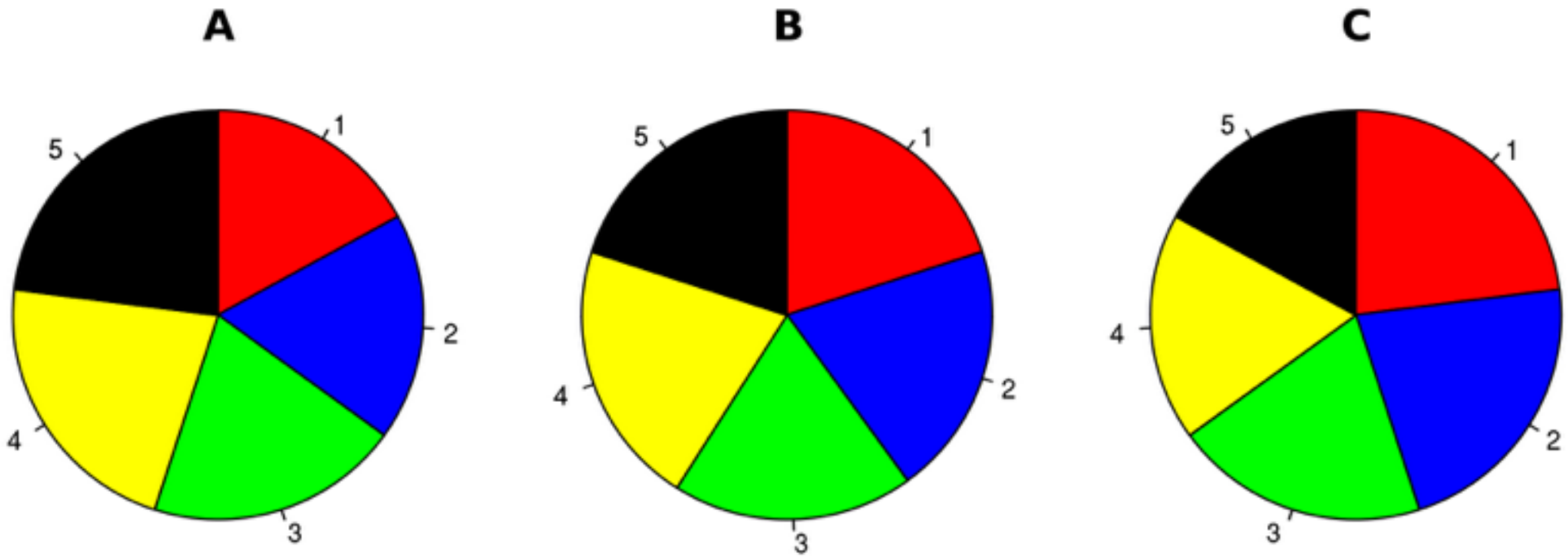
Most of the time this is a good thing

Using your (visual) brain



All have an r^2 of about 0.83

Using your (visual) brain

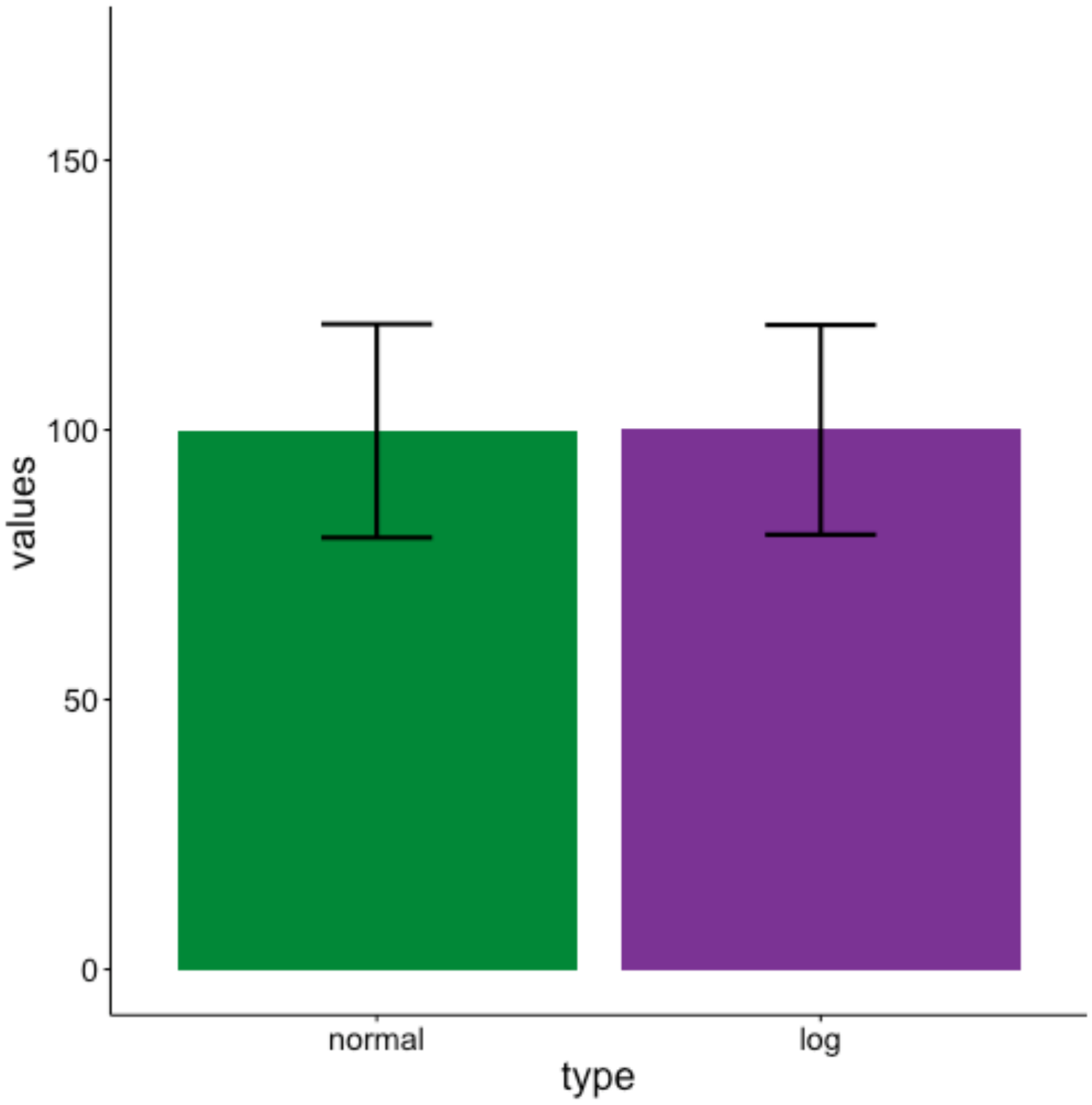


Using your (visual) brain

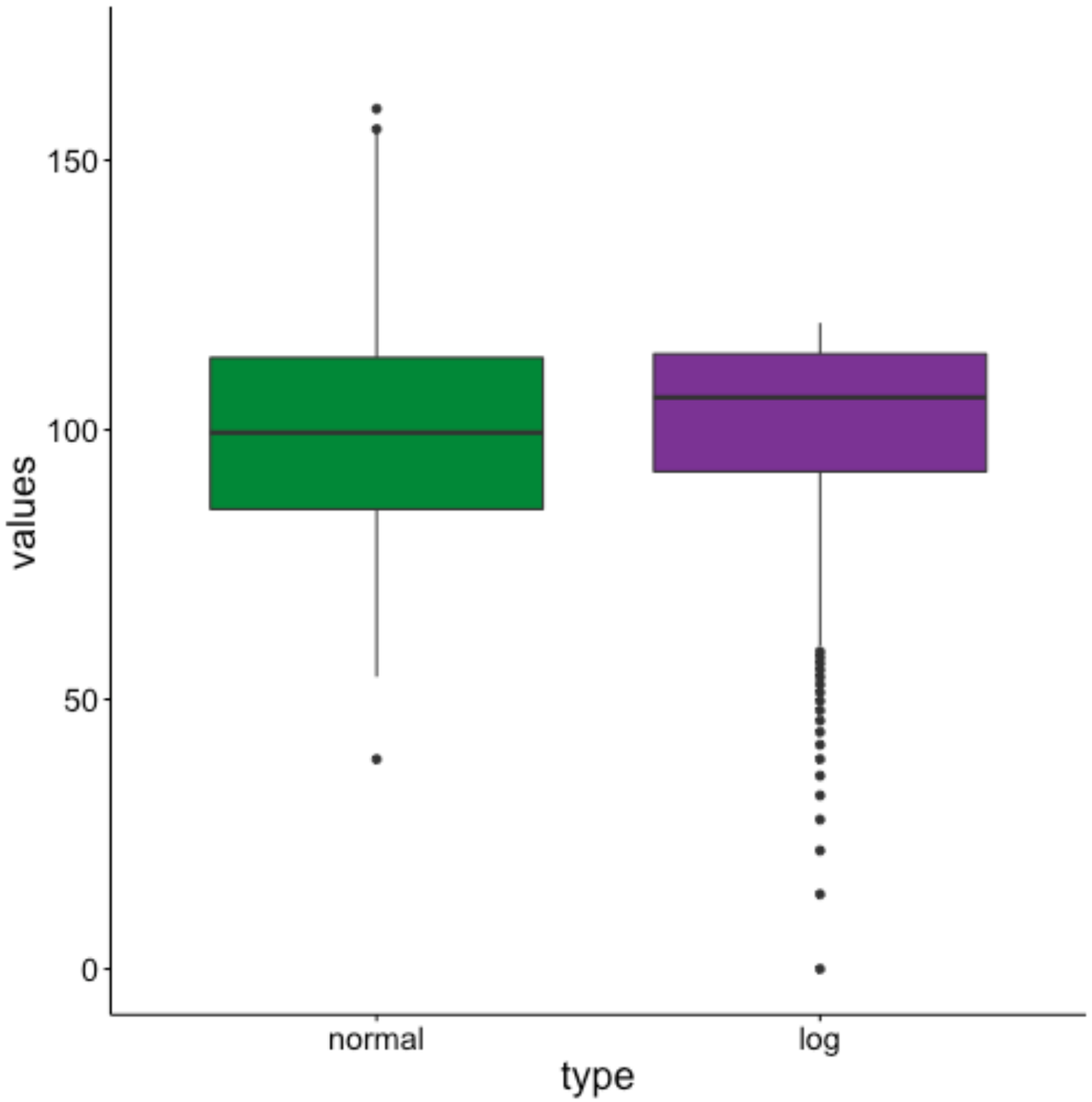
The pie chart is easily the worst way to convey information ever developed in the history of data visualization.

W. Hickey

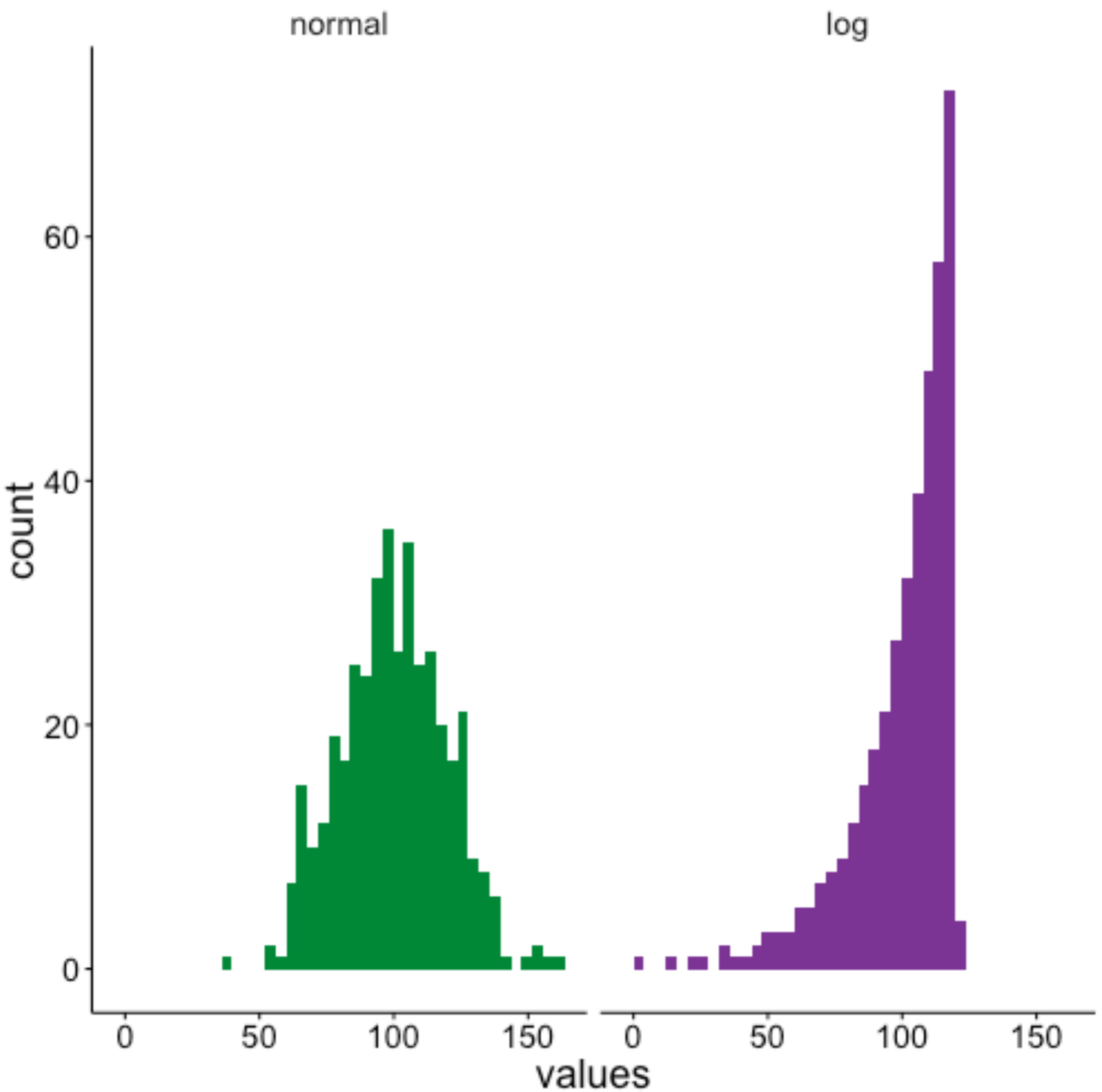
Using your (visual) brain



Using your (visual) brain



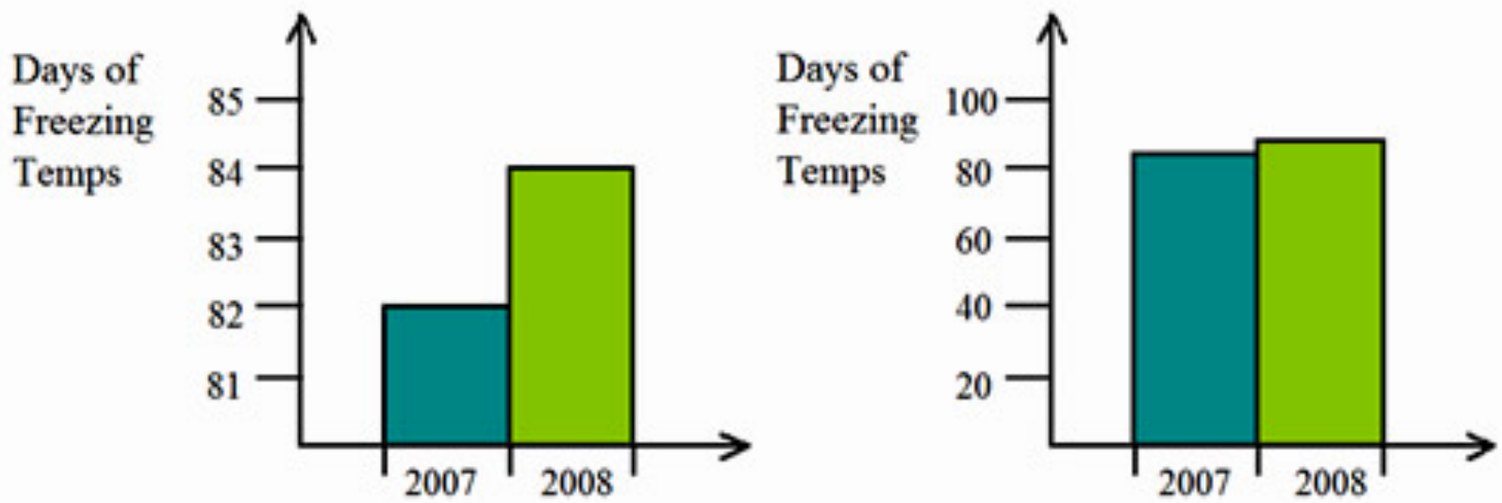
Using your (visual) brain



Using your (visual) brain

Misleading Graphs

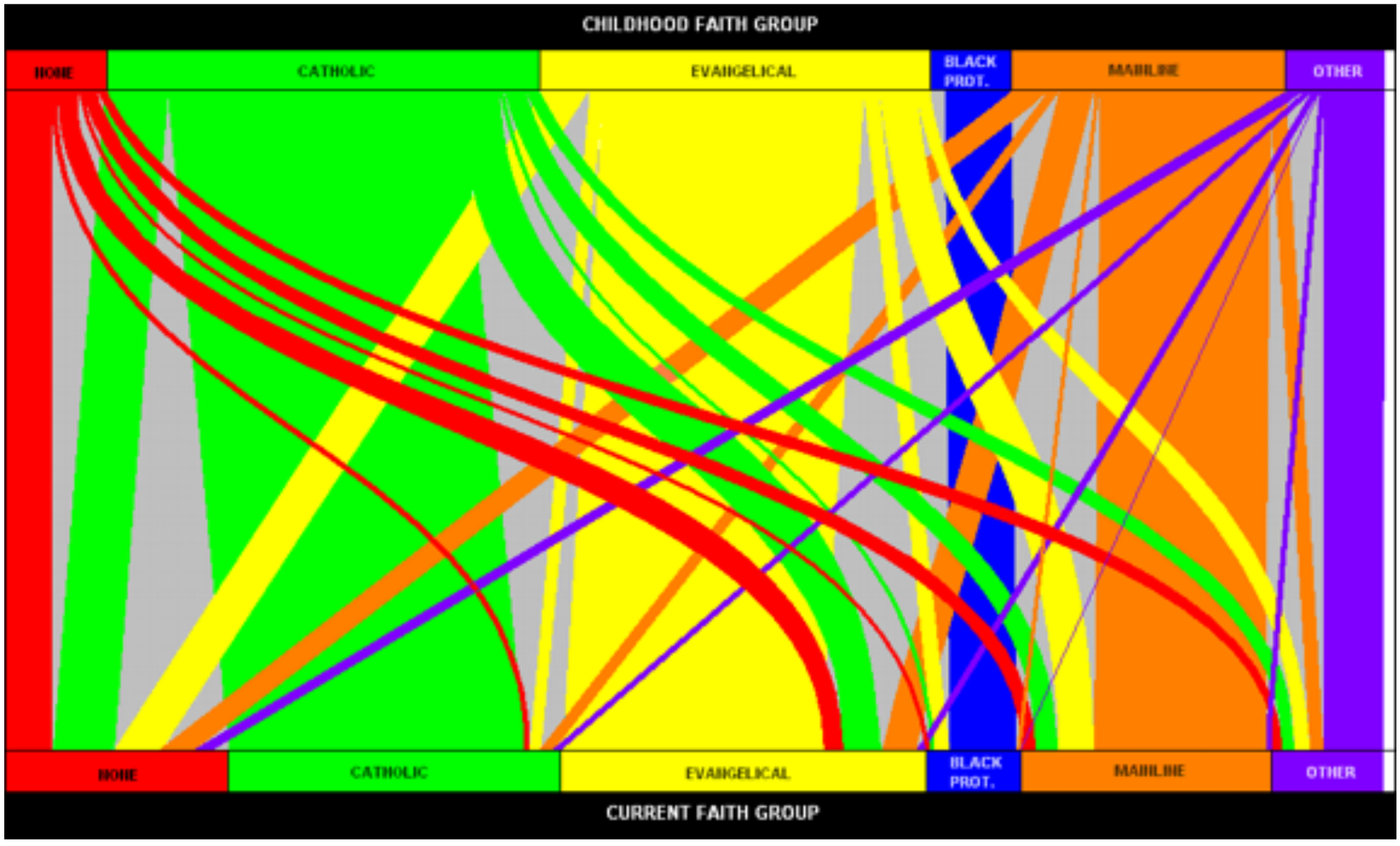
Compare the two graphs



Both show exactly the same data. However, the graph on the left makes the change appear to be much larger than it really is because the numbers on the vertical axis do not start at 0. Each vertical mark on the left graph represents 1 and each mark on the right represents 20 (the scale changes).

Image Source: http://www.his.washk12.org/

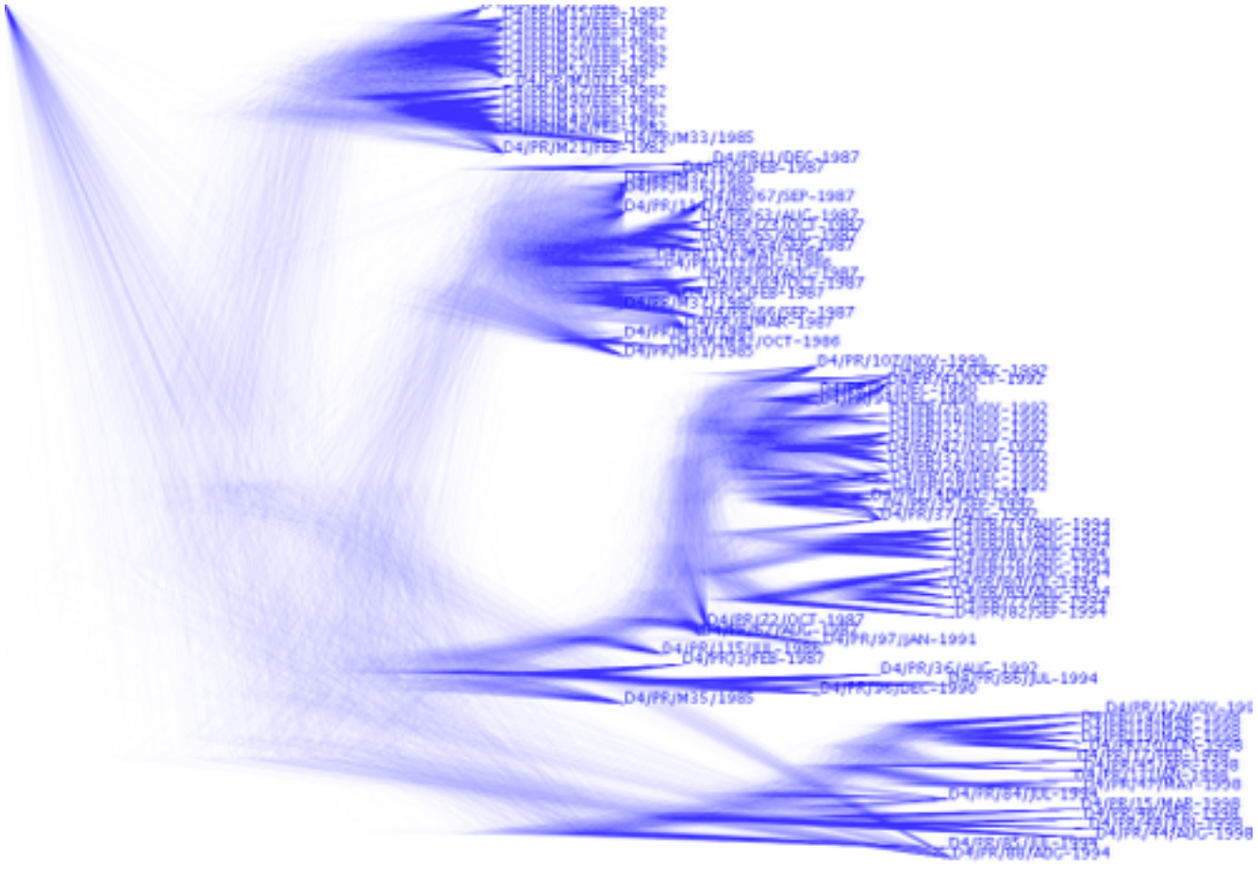
Using your (visual) brain

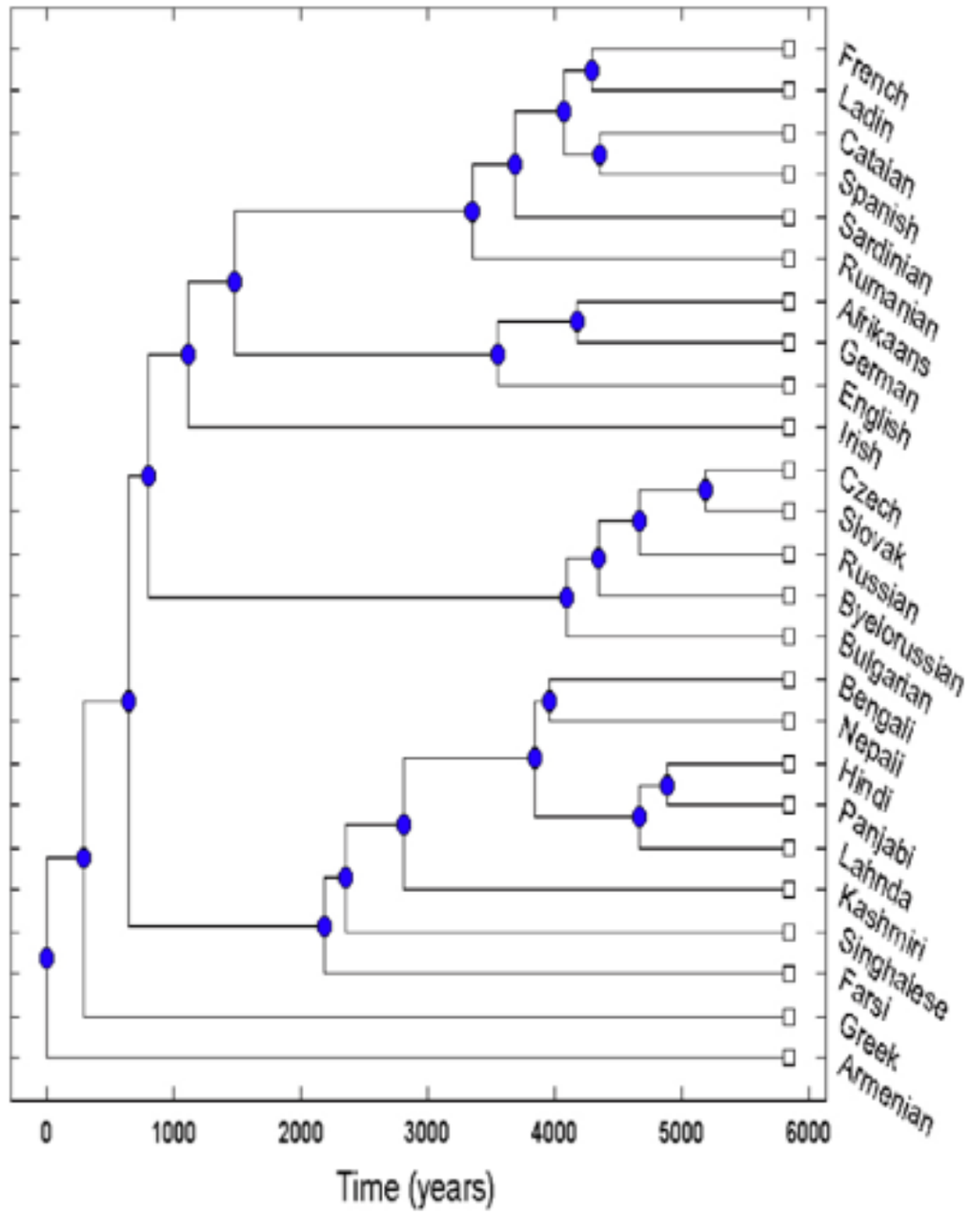


Using your (visual) brain

Visualization is particularly relevant for understanding change in phylogenies

Using your (visual) brain





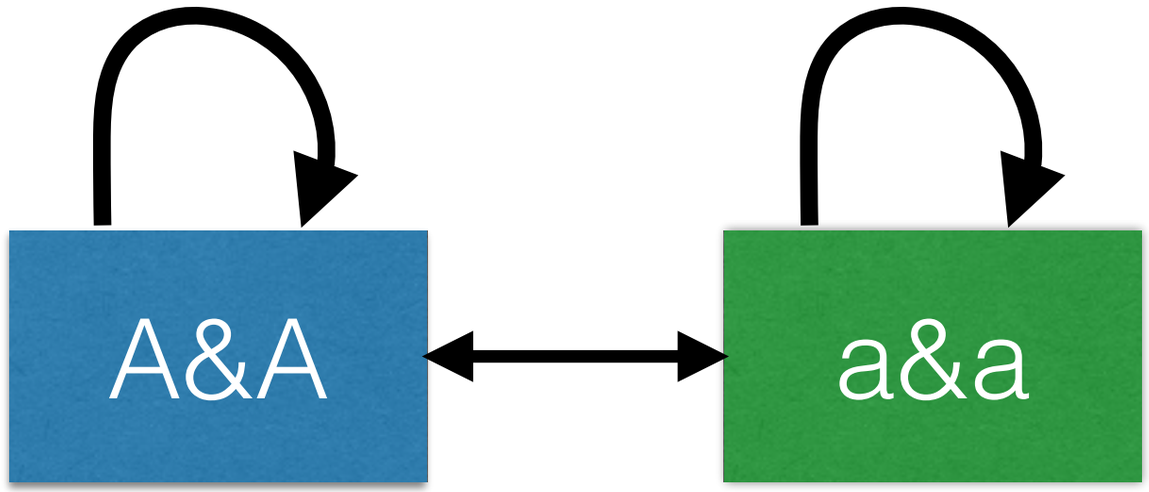
Using your (visual) brain

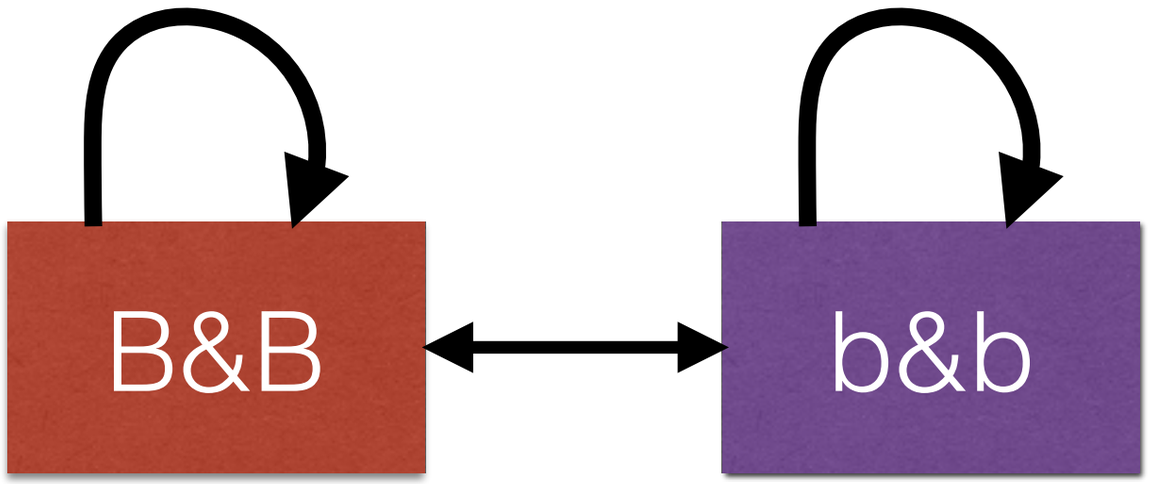
Suppose we infer from cultural or linguistic data the following transition matrix

	A&A	a&a	B&B	b&b
A&A	1-e	e	0	0
a&a	e	1-e	0	0
B&B	0	0	1-a	a
b&b	0	0	a	1-a

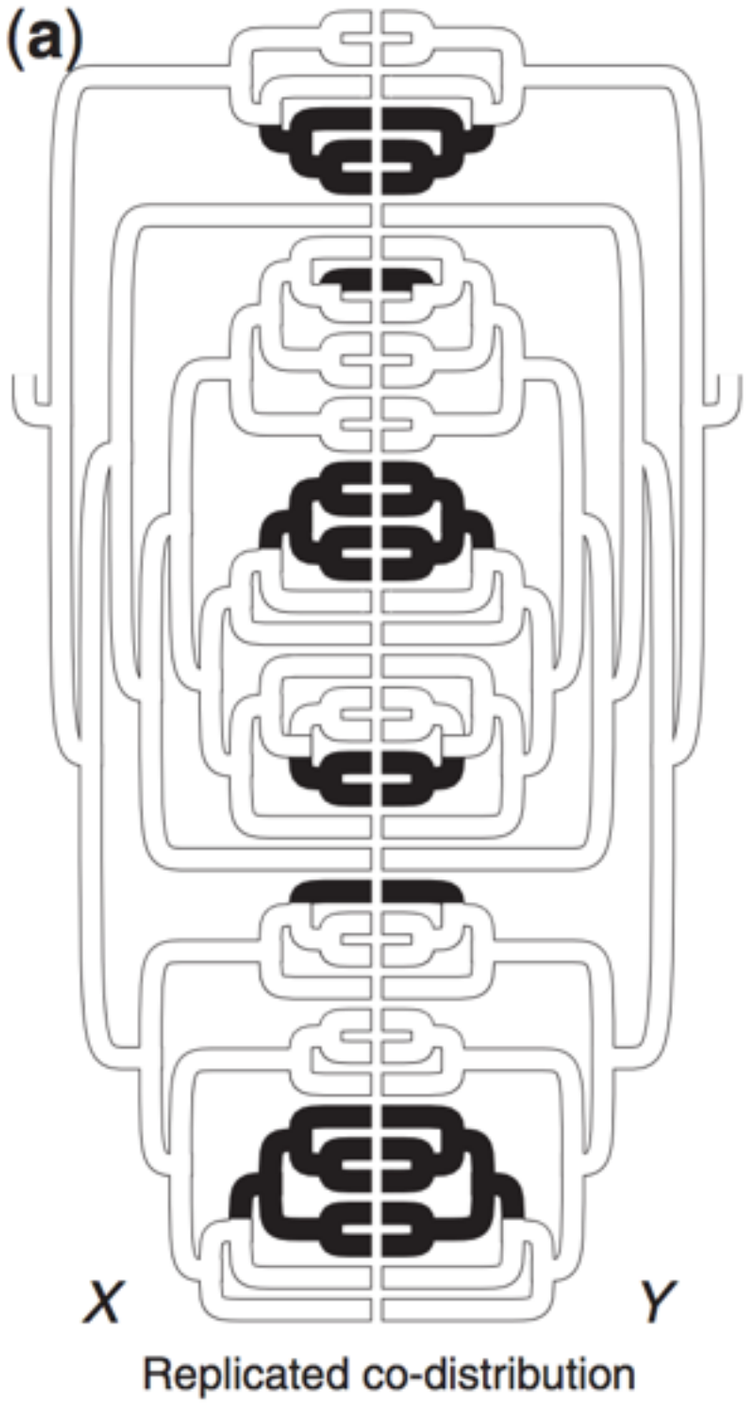
Using your (visual) brain

	A&A	a&a	B&B	b&b
A&A	1-e	e	0	0
a&a	e	1-e	0	0
B&B	0	0	1-a	a
b&b	0	0	a	1-a



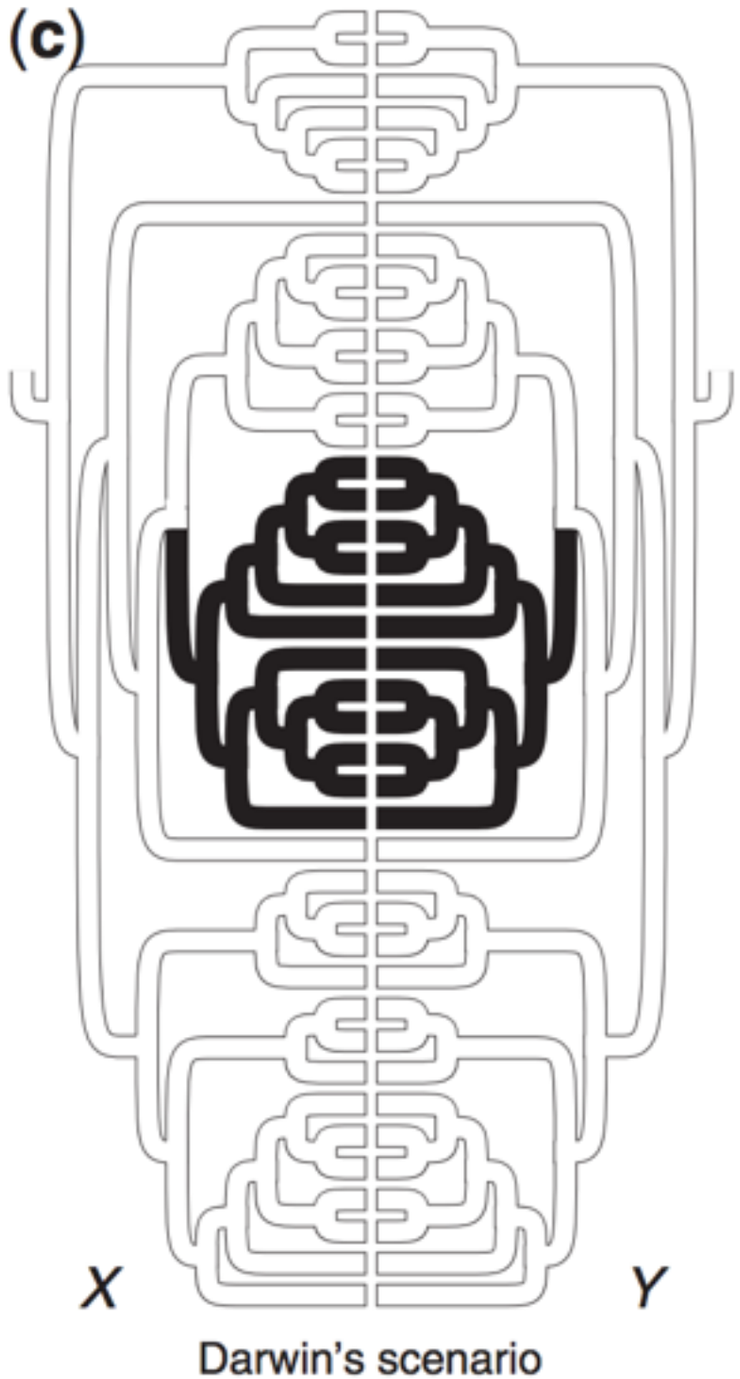


Using your (visual) brain



In this case, the previously inferred matrix suggests a general preference for some states changing into other states

Using your (visual) brain



Darwin's scenario

Here the inference depends on a single event - hardly generalizable from this data only

Learning to learn

Learning to learn

The development of new data analysis and visualization tools has **accelerated**, and the delay between publication and practical implementation has reduced **dramatically**

Learning to learn

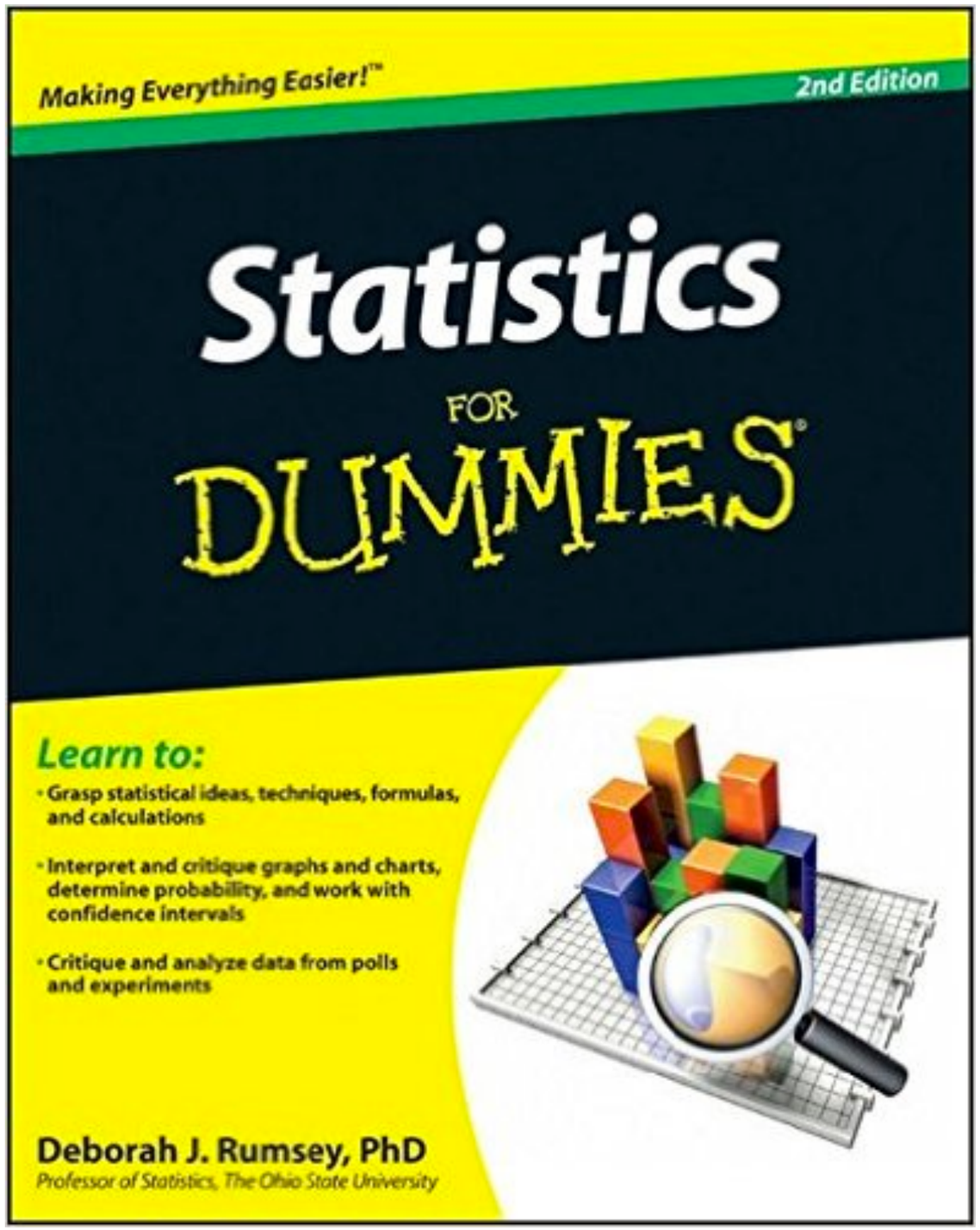
Actually, you can check right now how many up-to-date packages there are:

```
dim(available.packages())
```

Learning to learn

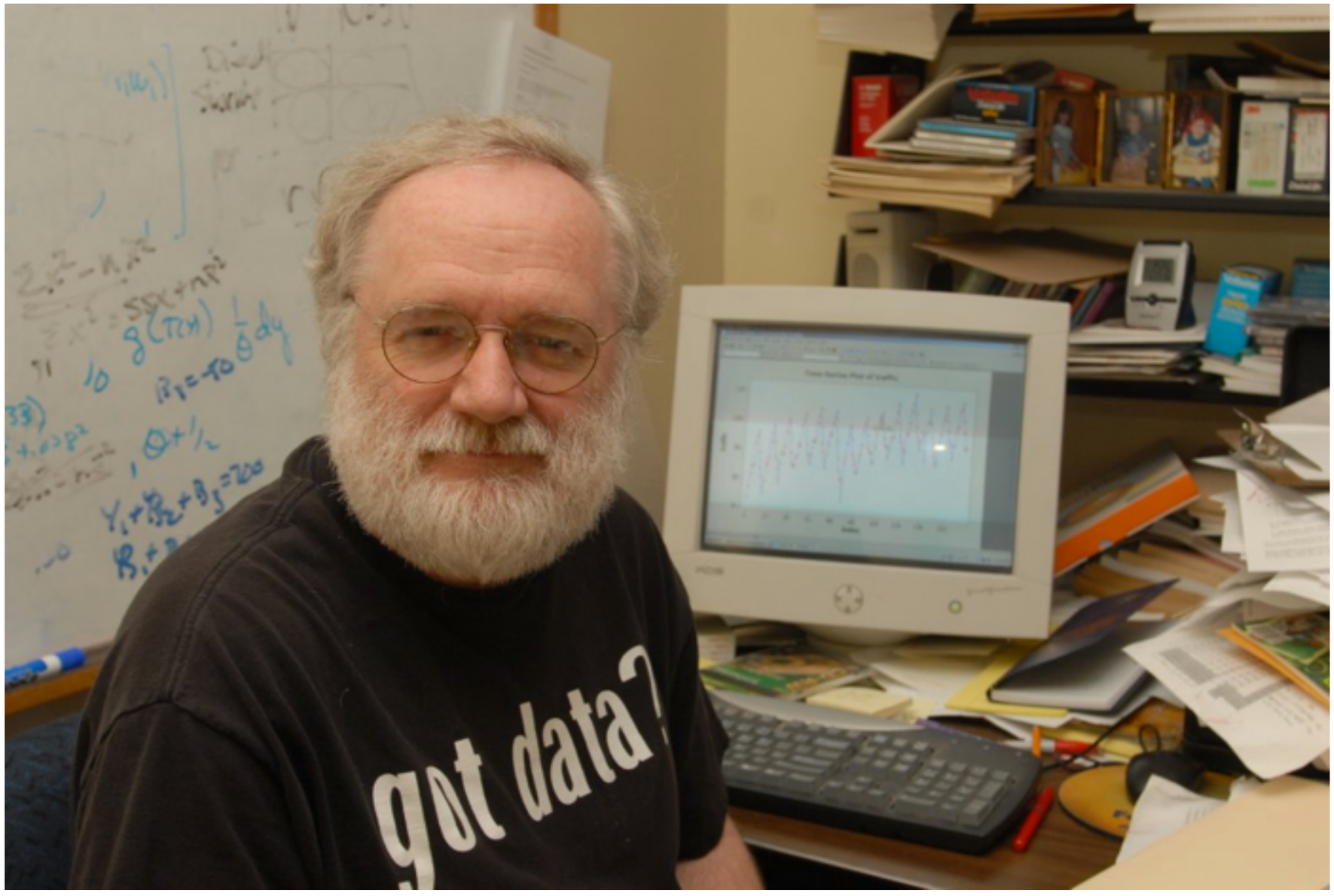
We are getting better and better at solving problems, which entails that the technique you learned a decade ago is probably not the most efficient solution today

Learning to learn



?

Learning to learn



?

Learning to learn

The efficiency of many new methods is assessed through benchmarks.

How much more efficient/ adequate is a method outside of the benchmark can vary a lot

Learning to learn

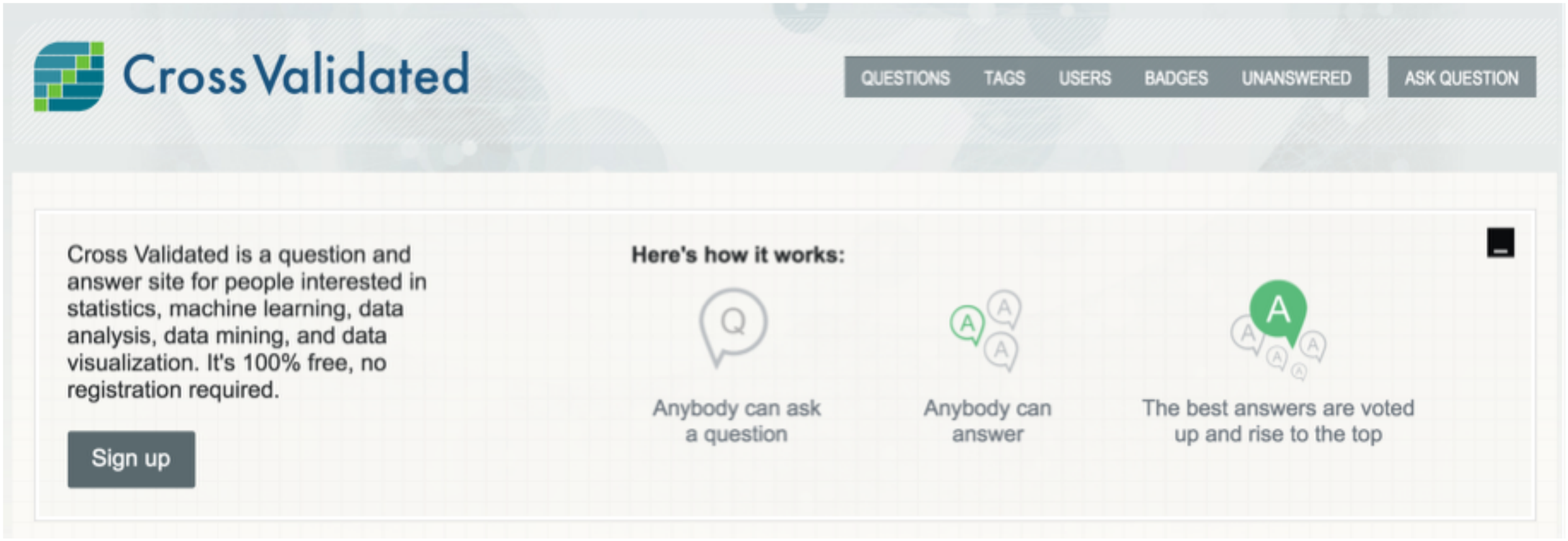
Most people that develop methods are academics. They also need

citations

Even the most well-intentioned and clever among them cannot predict all possible outcomes or situations people might be applying the methods on

Excellent strategy:

learn from the crowd



On fishing and mining

On fishing and mining

Phrasebook for the stats-skeptical

"You can prove anything with statistics"

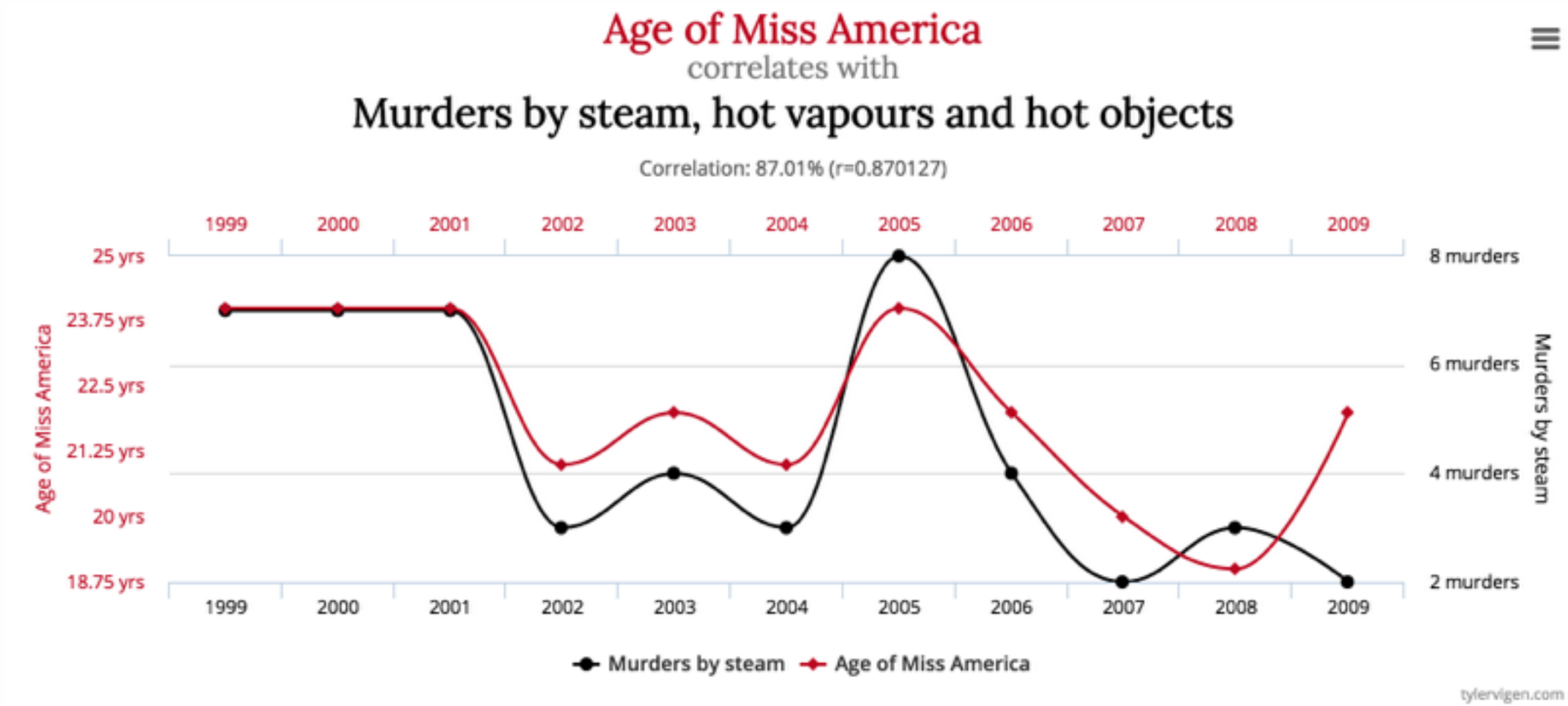
"Correlation doesn't imply causation"

"Haha, the number of movies in which Nicholas Cage appeared in is correlated with number of suicides in Northern California. Ergo stats is bollocks"

"Oh, but we already knew that"

"You know, that's like, your opinion dude"

On fishing and mining

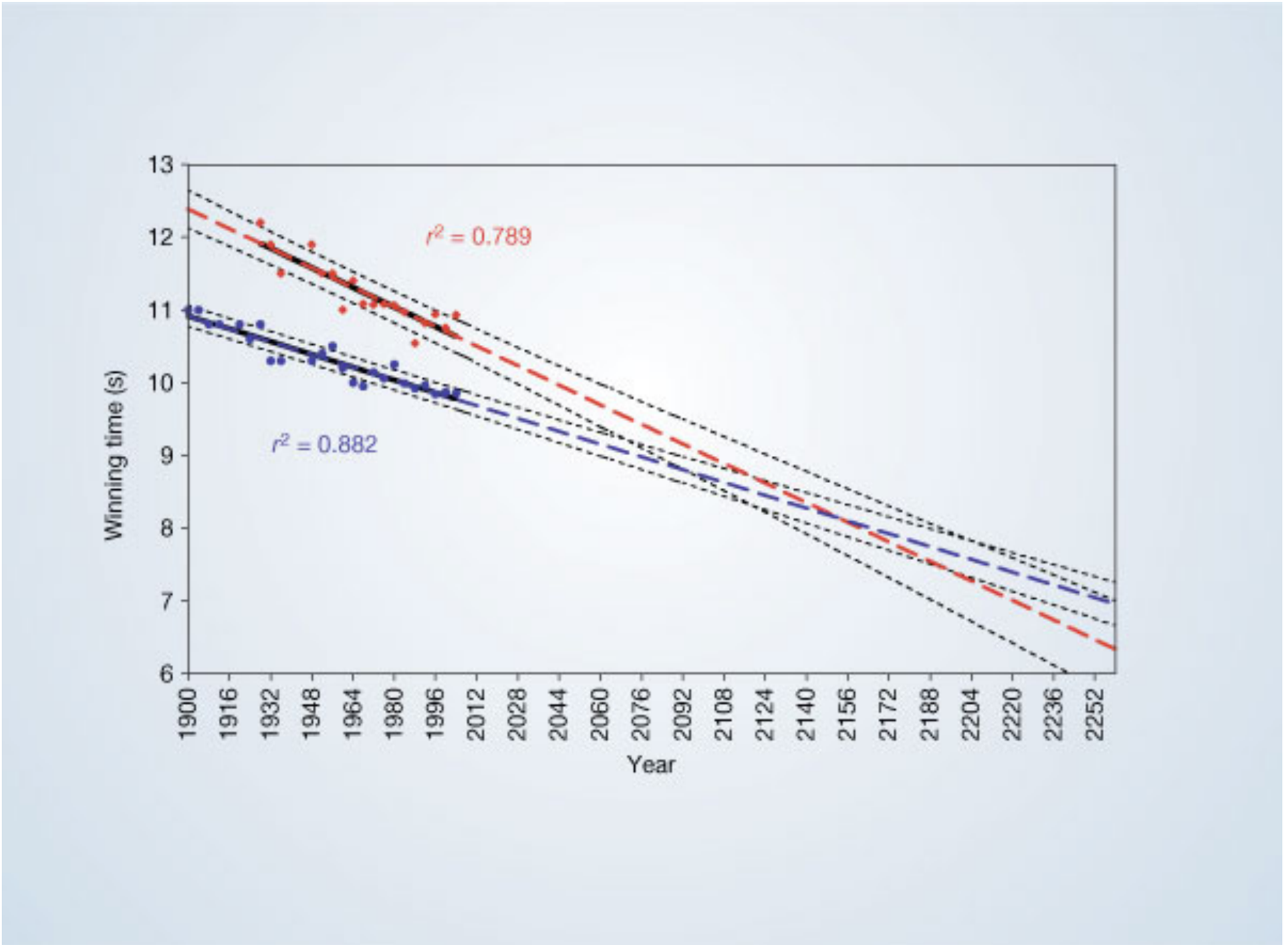


To be fair, you can get pretty stupid results if you have big (or very little) data and lots of imagination…

On fishing and mining

…or sometimes even when you have the proper data and skills. Getting right a complex model might be an extremely challenging task.

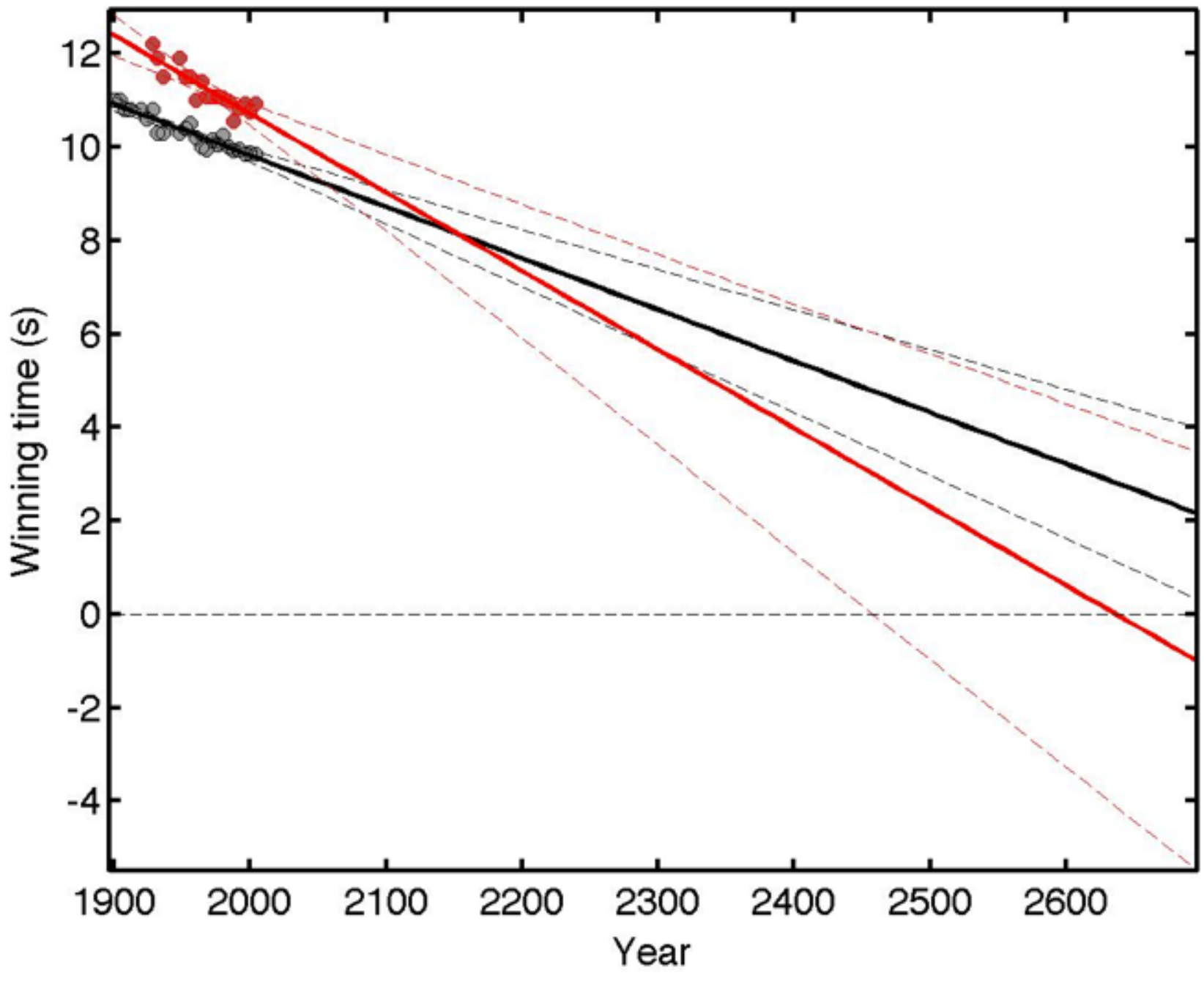
On fishing and mining



On fishing and mining

(…) Should these trends continue, the projections will intersect at the 2156 Olympics, when — for the first time ever — ***the winning women's 100-metre sprint time of 8.079 seconds will be lower than that of the men's winning time of 8.098 seconds***

On fishing and mining



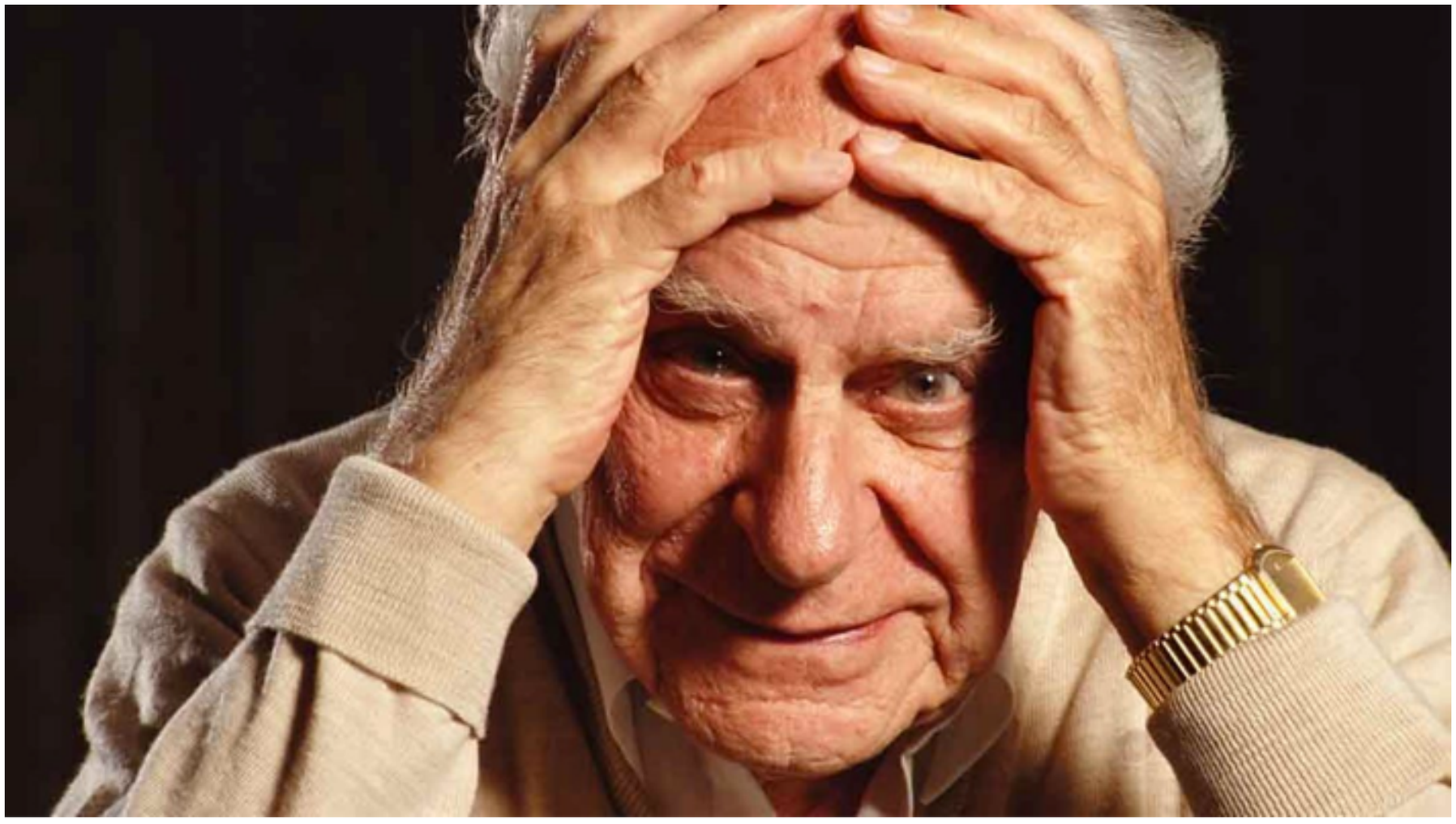
taken from https://sirogers.wordpress.com/

On fishing and mining

On fishing and mining

In 2620, Mary Jones completed the 100m race in -5 seconds.

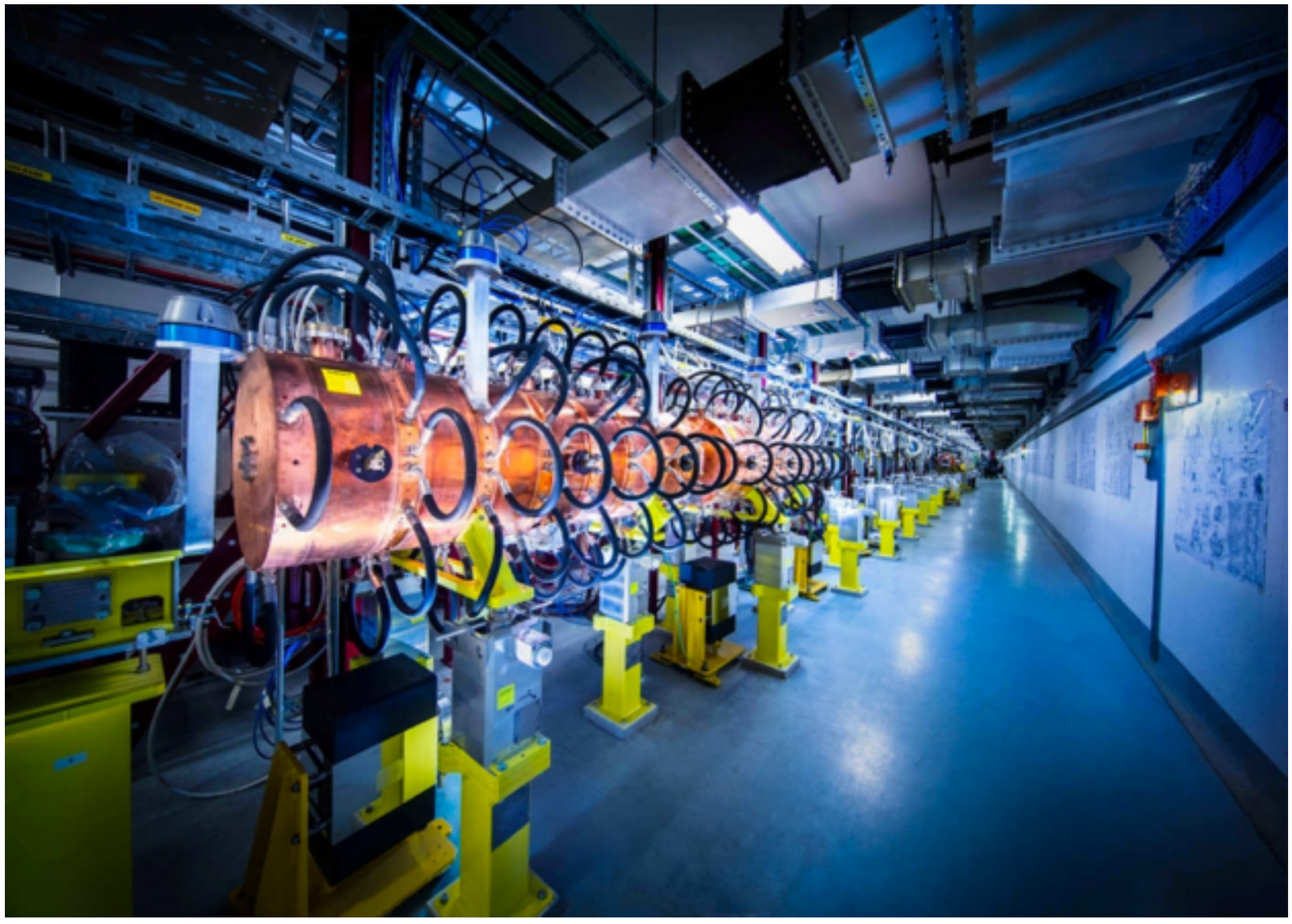
On fishing and mining



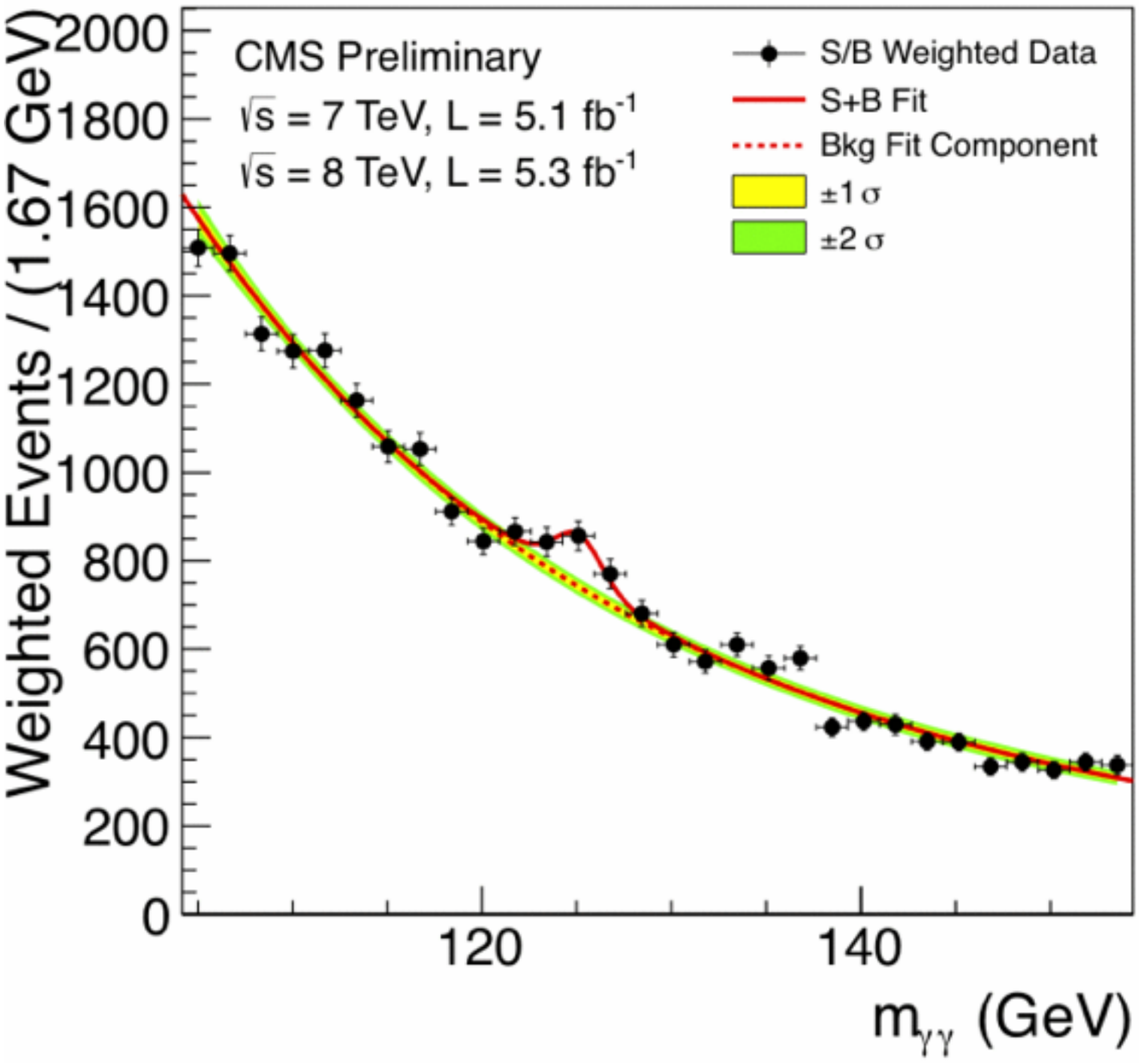
After Popper, it became standard to stress the distinction between theorizing and testing

On fishing and mining

Sometimes we have clear hypotheses that can be readily projected into a statistical evaluation



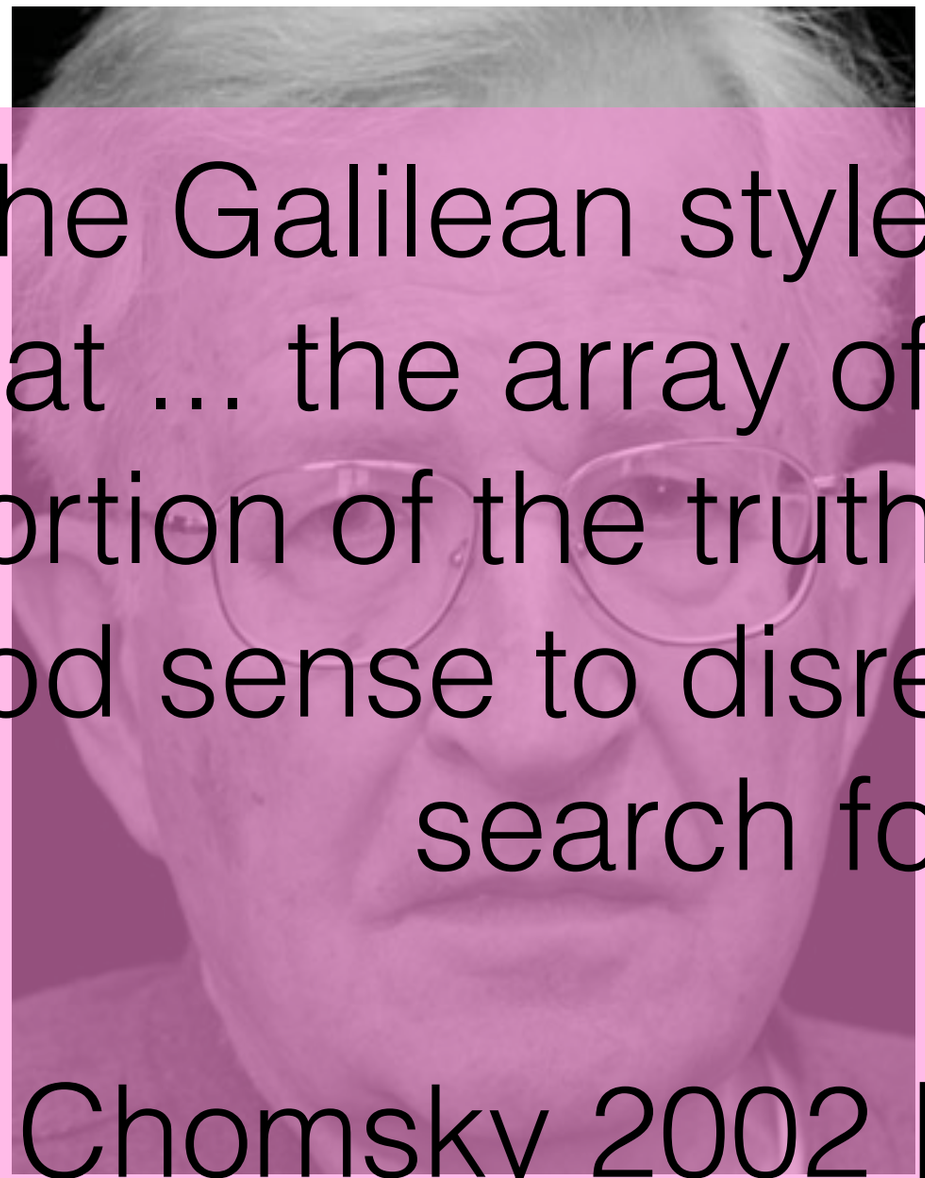
On fishing and mining

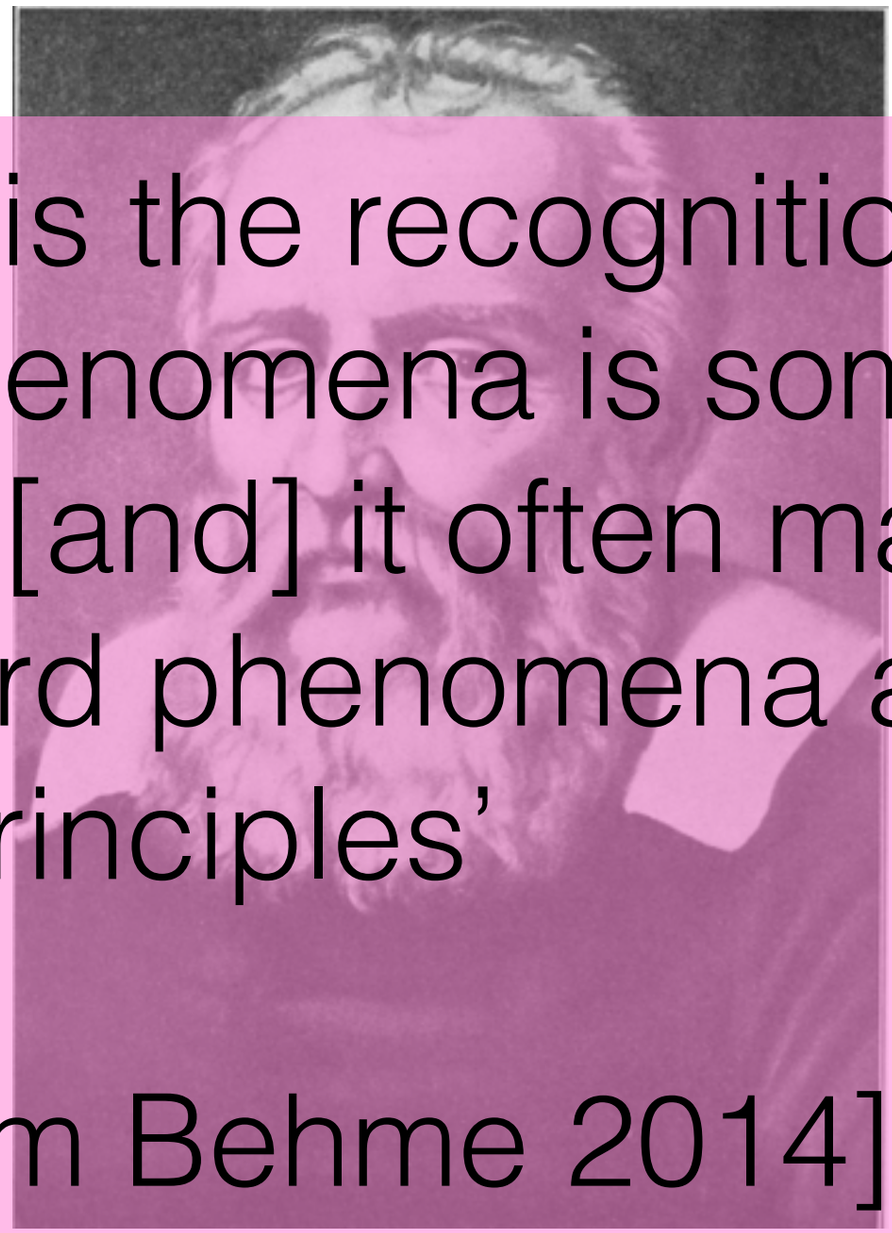


On fishing and mining

…but not always. Exploratory data analysis is not only a **totally legitimate** thing to do, but sometimes also the best and most rational thing to do.

On fishing and mining





‘the Galilean style … is the recognition that … the array of phenomena is some distortion of the truth … [and] it often makes good sense to disregard phenomena and search for principles’

Chomsky 2002 [from Behme 2014]

On fishing and mining

On fishing and mining

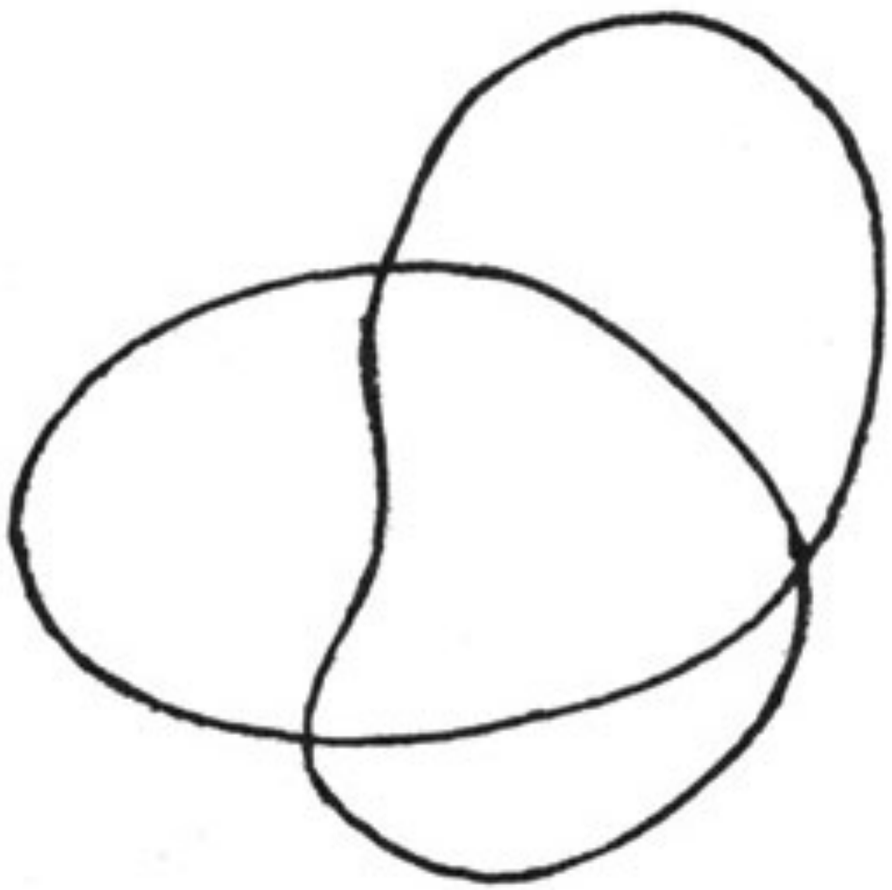


FIG. 18

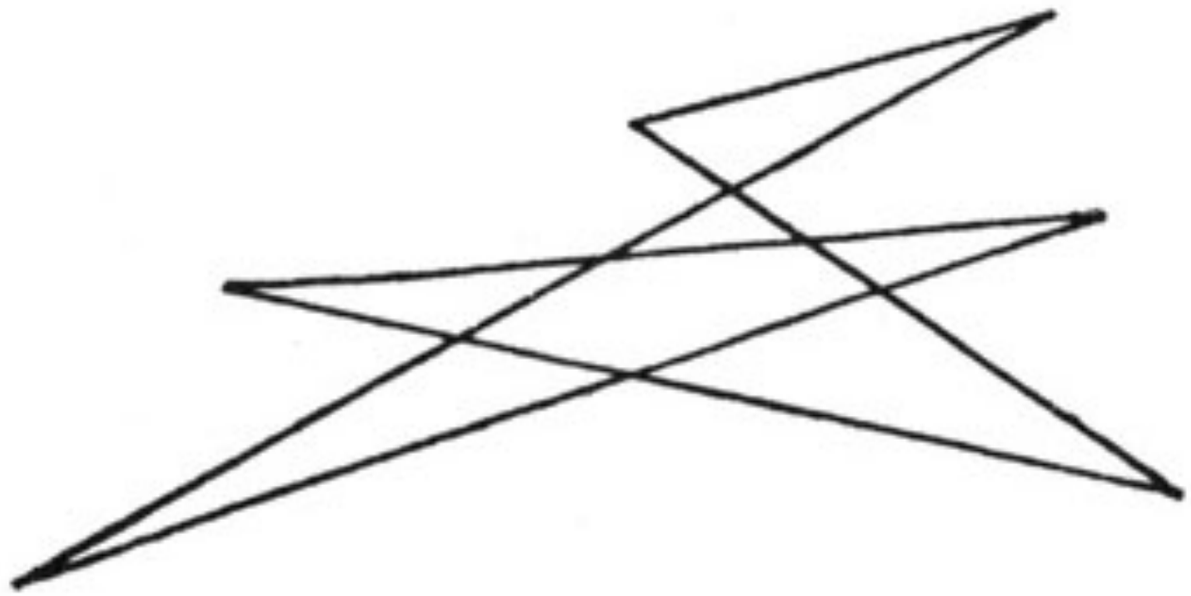


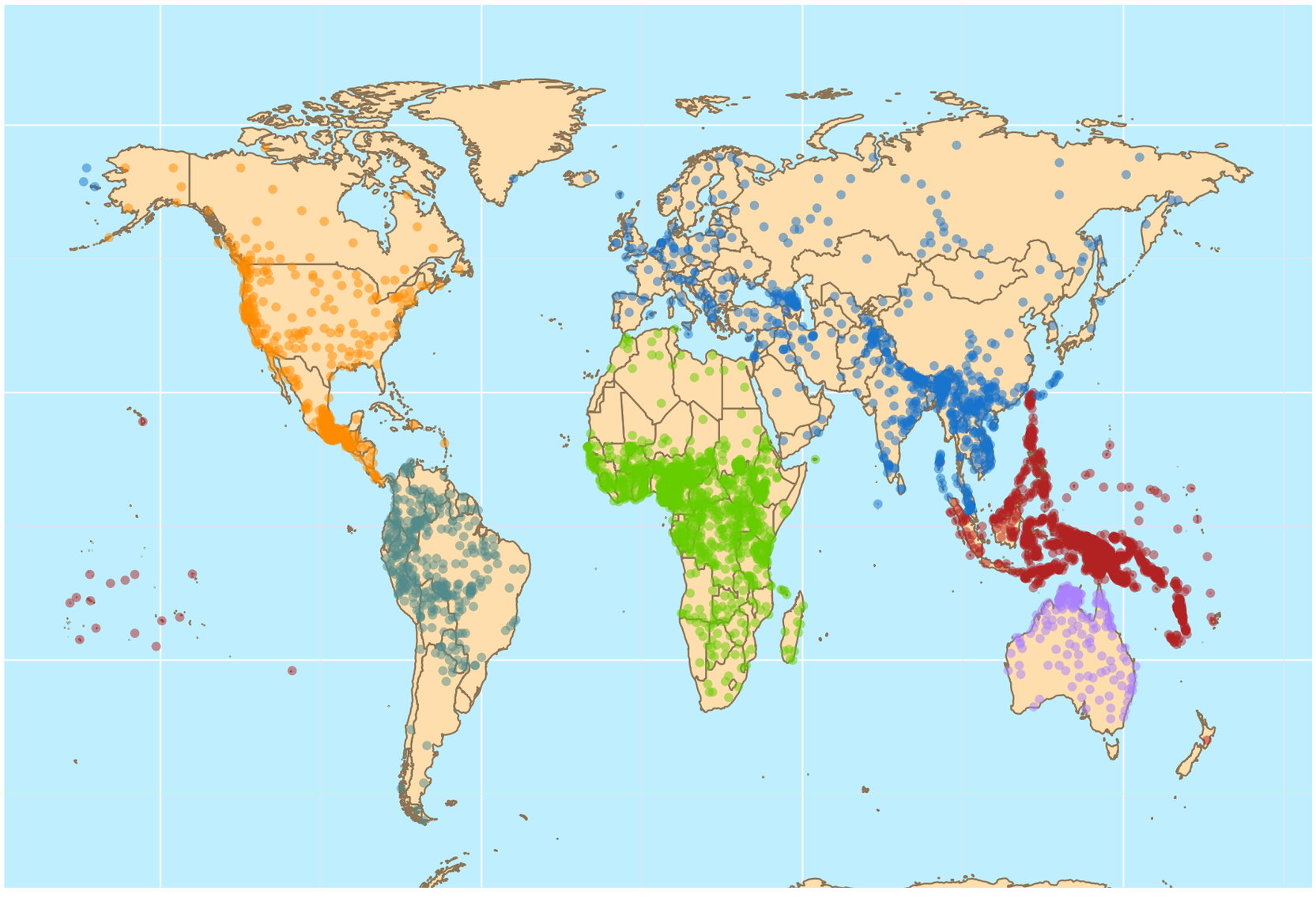
FIG. 19

On fishing and mining

People have proposed all clases of theories that "explain" why apparently some concepts tend to be named through words that carry specific phones:

correlation between F0 and body size
"pointing" gesture of the tongue
proto-human residual
iconicity of mouth openness
…

On fishing and mining



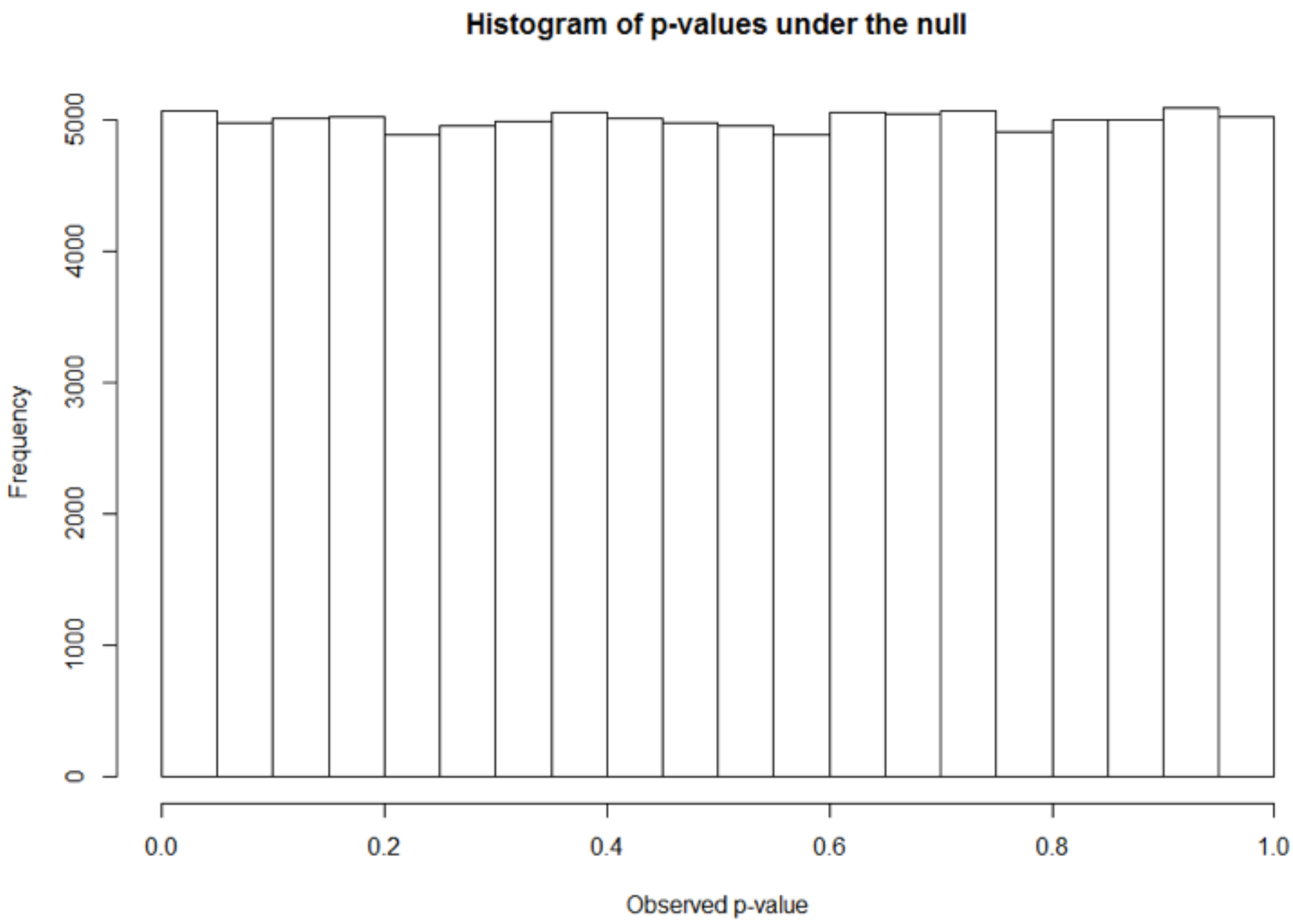
On fishing and mining

dog
red
rock
man
one
sand
water
…

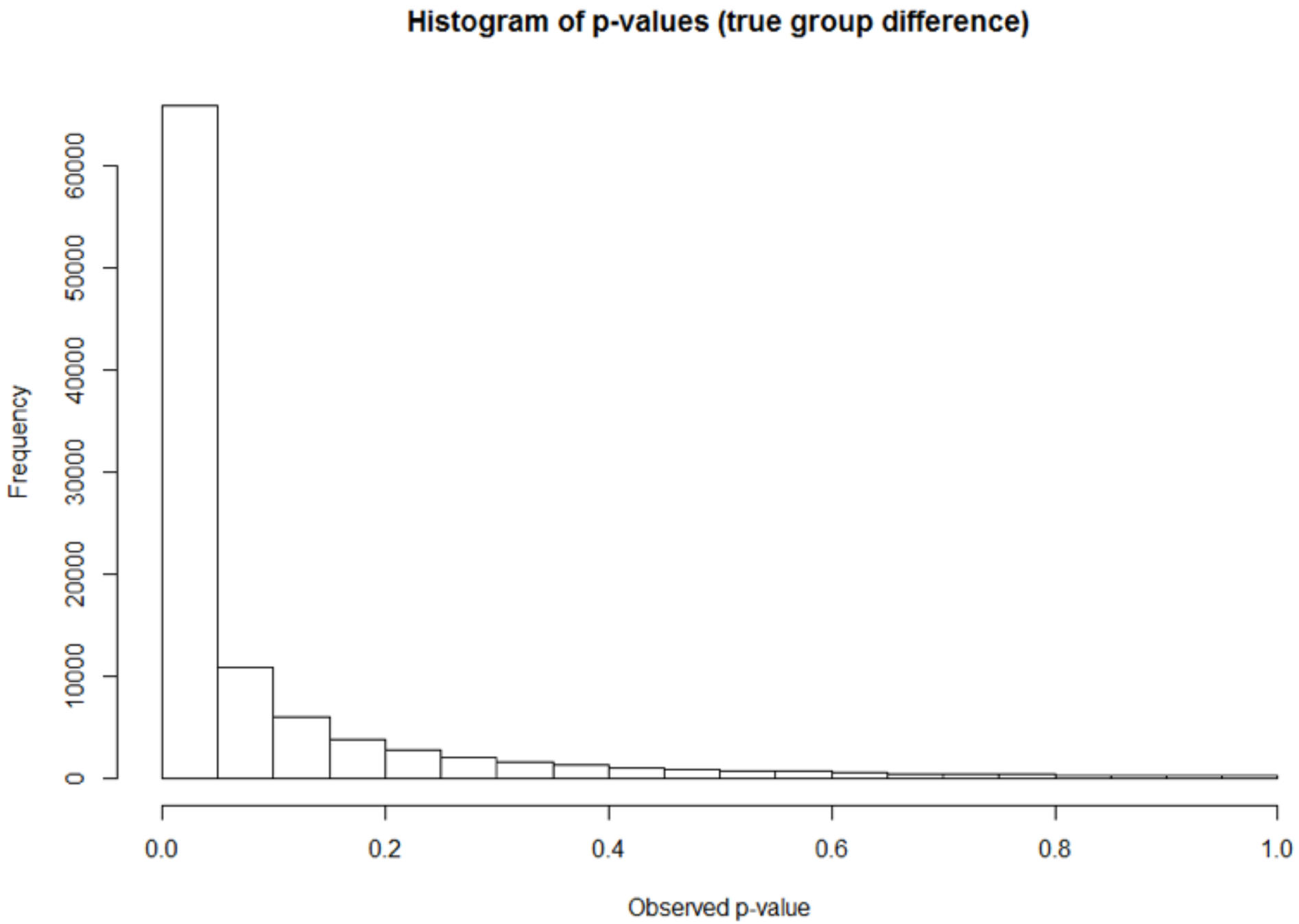
r all kinds of ‘r’ sounds
i high front vowels
l laterals
! clicks
…

100 items x 41 symbols =
4100 tests (!)

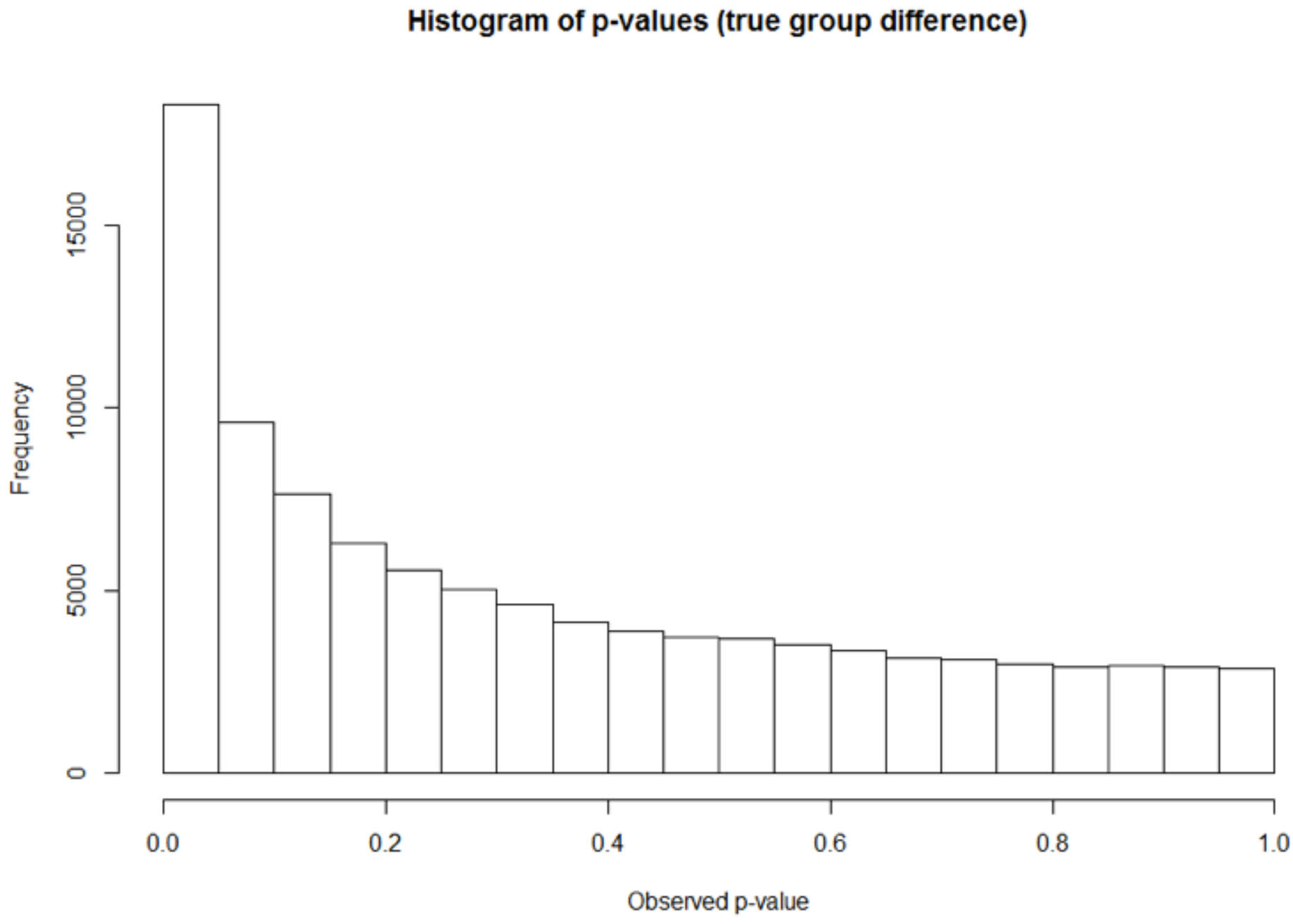
On fishing and mining



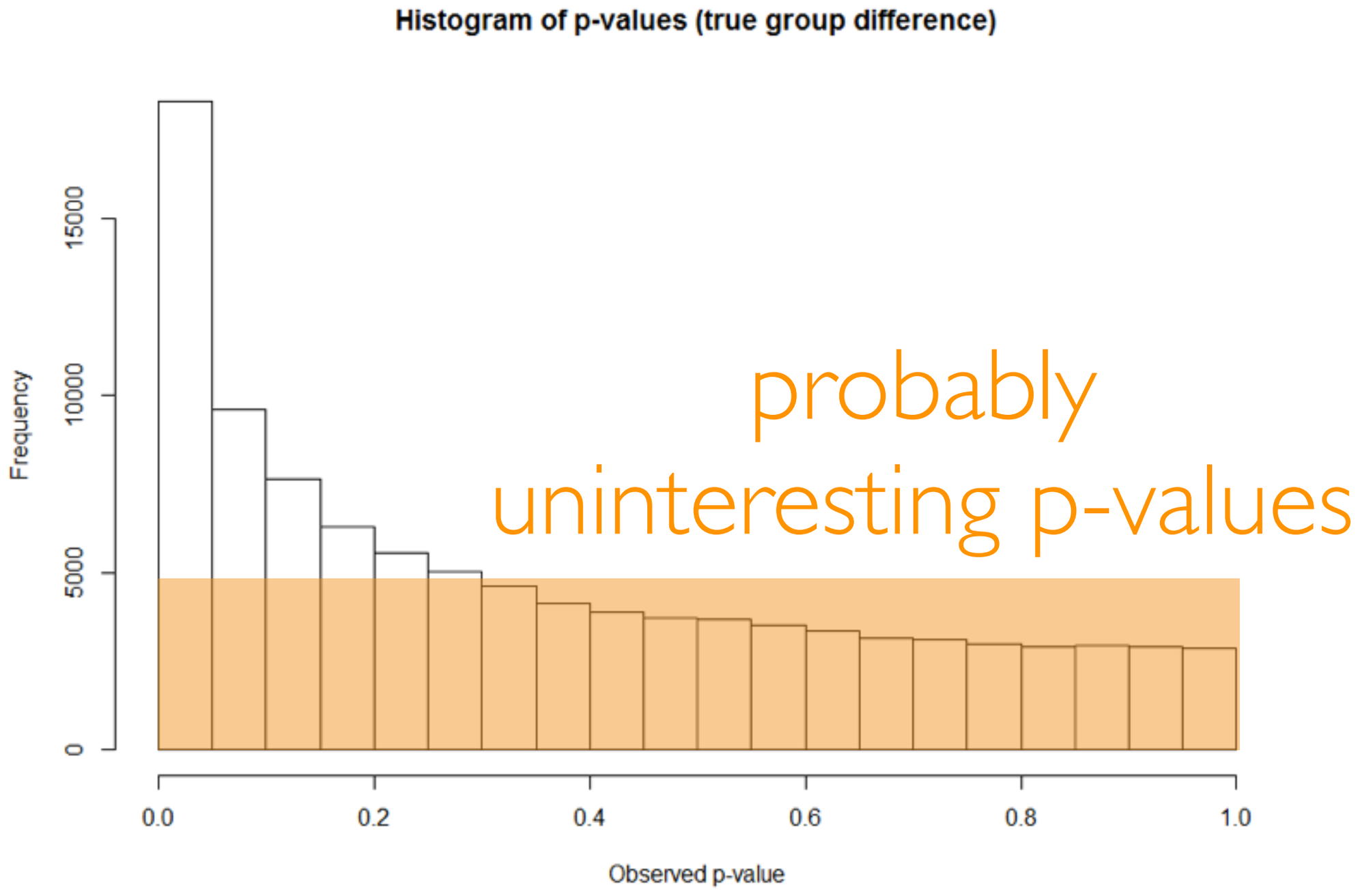
On fishing and mining



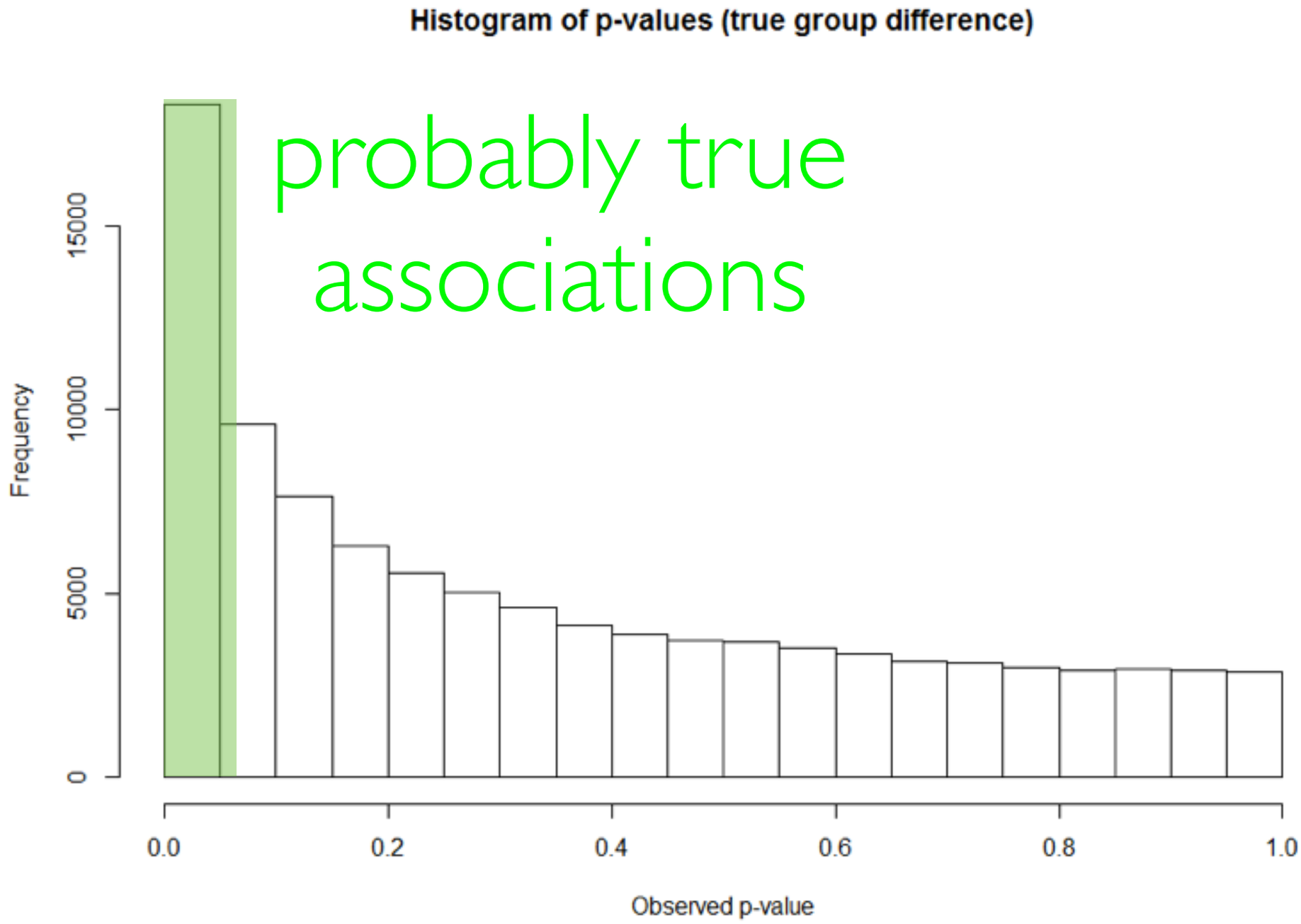
On fishing and mining



On fishing and mining



On fishing and mining



probably true associations

On fishing and mining

small and [i]

tongue and [l]

round and [r]

sand and [s]

…

On fishing and mining

Distinguishing uninteresting from relevant tests in this manner becomes unfeasible with small data

On fishing and mining

Fishing occurs when you check across samples or variables until you find something that looks interesting **but** you do not disclose nor account for this search

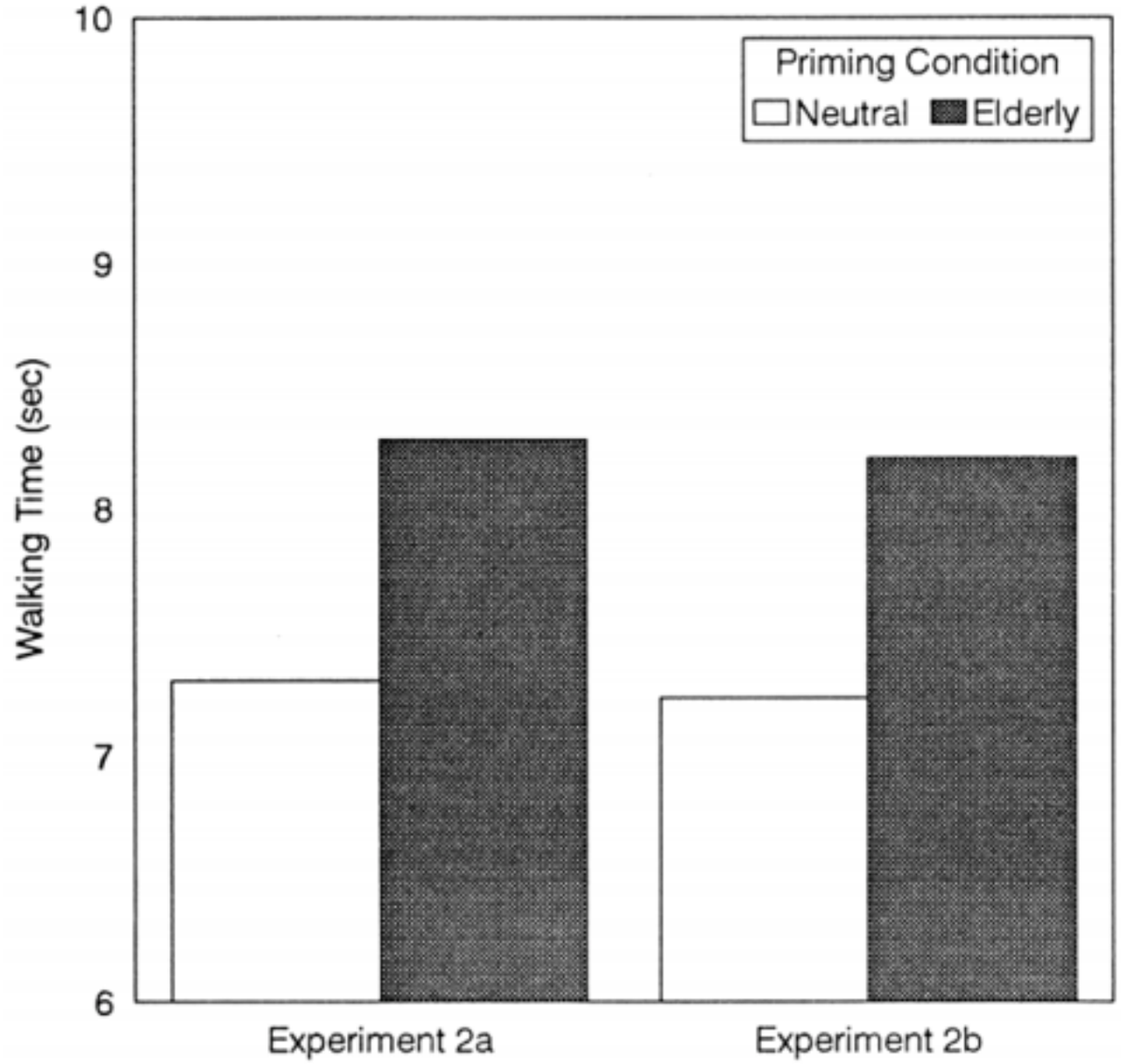
On fishing and mining

Most of the examples of unreplicable research are the result of a large unobserved number of extra tests or unknown or ignored parameters

On fishing and mining

One source of spurious associations is **optional stopping** or **data peeking**: you check whether your results are significant during data collection, and you stop when that is the case. Related to that are the **researcher degrees of freedom**, modelling or testing decisions partially informed by their partial results.

On fishing and mining

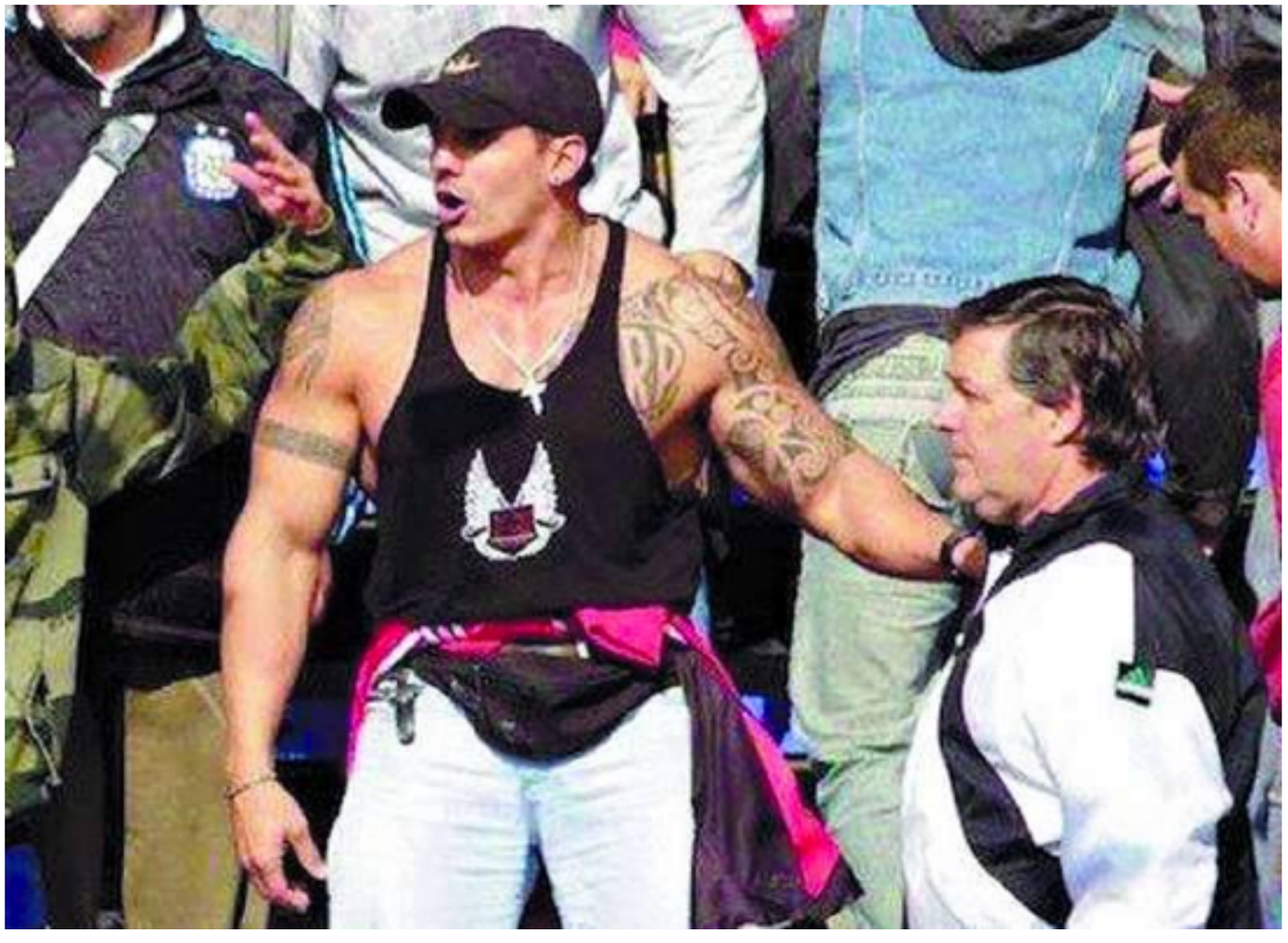


Bargh et al. 1996

Figure 2. Mean time (in seconds) to walk down the hallway after the conclusion of the experiment, by stereotype priming condition, separately for participants in Experiment 2a and 2b.

On fishing and mining

Dijksterhuis and van Knippenberg 1998, Shanks et al. 2013



On fishing and mining

NATURE | NEWS

Over half of psychology test

Largest replication study to date casts dou

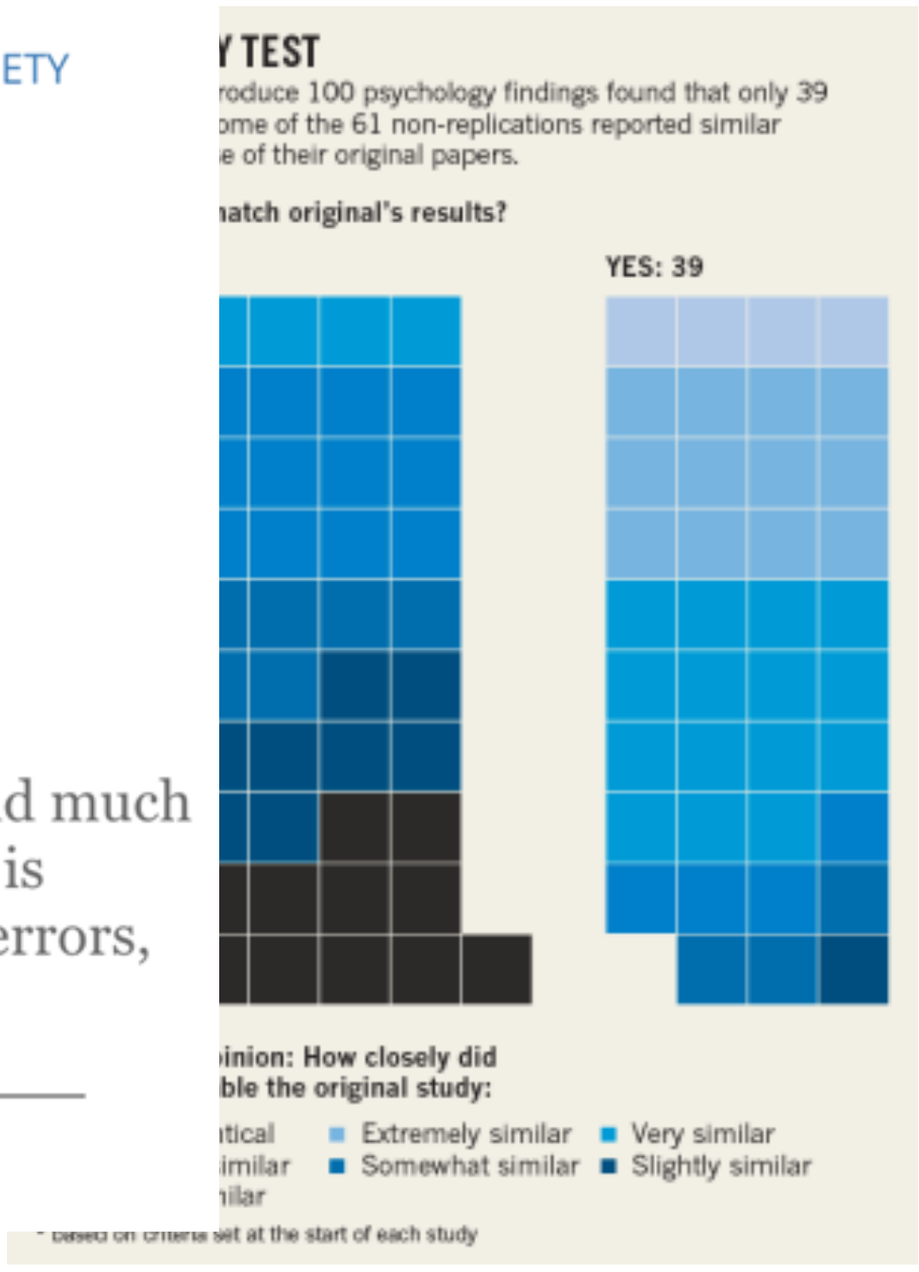
Monya Baker

SCIENCE & HEALTH > CULTURE & SOCIETY

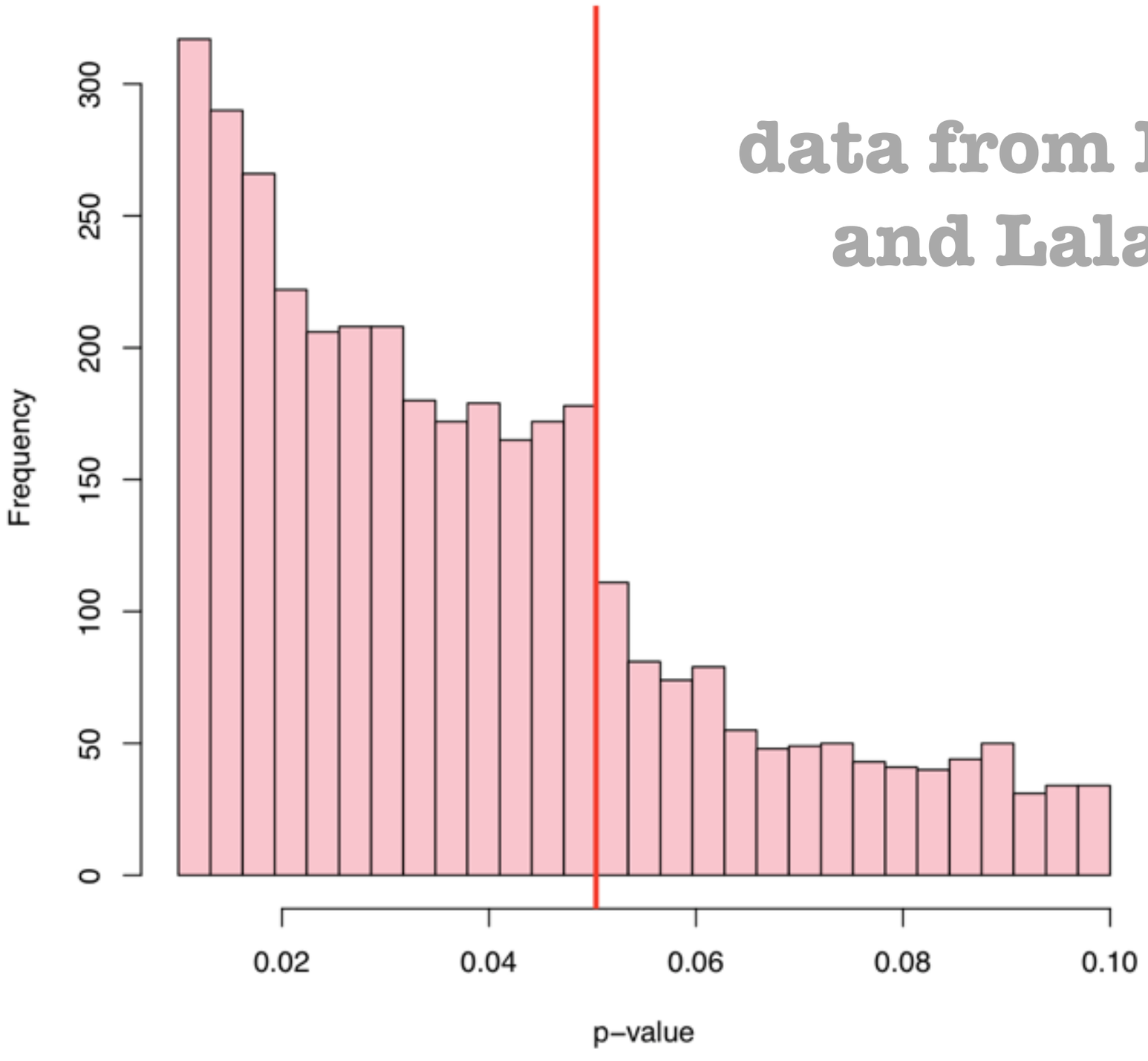
Study that undercut psych research got it wrong

Widely reported analysis that said much research couldn't be reproduced is riddled with its own replication errors, researchers say

March 3, 2016 |



On fishing and mining



data from Massicampo and Lalande 2012

On fishing and mining

What does any of these things have to do with knowing a topic or methods?

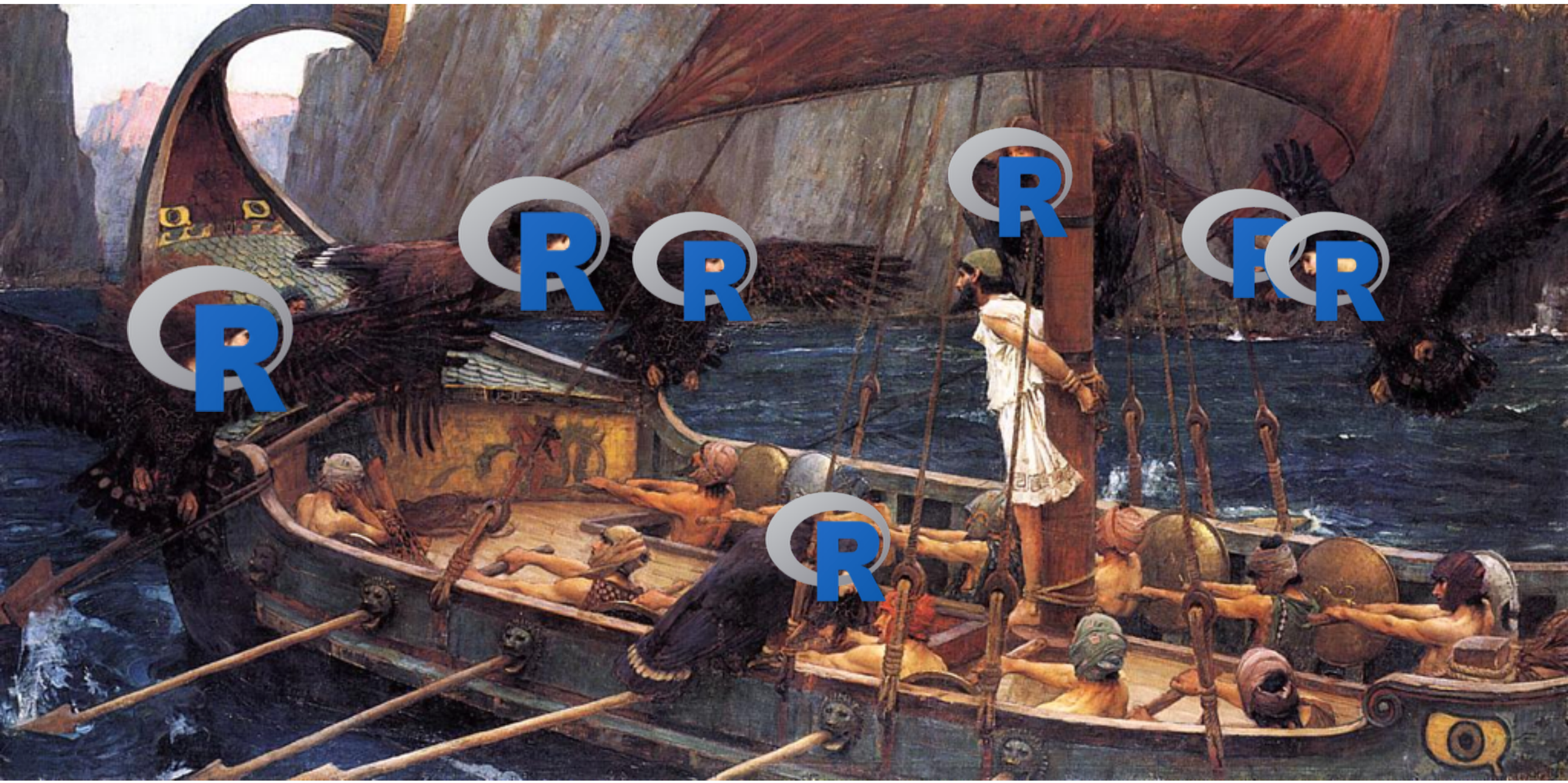
On fishing and mining

The more you know about a topic, **the fewer the number of tests/decisions you will do**, which reduces the inflation of potential false positives / negatives due to multiple testing

On fishing and mining

The more you know about data science, **the fewer the number of models/methods you will use**, which reduces the inflation of potential false positives / negatives due to multiple testing

On fishing and mining



Gracias!