Lies, Damned Lies and Statistics

EuroPython 2018
Edinburgh, UK
July 2018

@MarcoBonzanini

In the Vatican City there are 5.88 popes per square mile

This talk is about:

- The misuse of statistics in everyday life
- How (not) to lie with statistics

This talk is not about:

- Python
- Advanced Statistical Models

The audience (you!):

- Good citizens
- An interest in statistical literacy
  (without an advanced Math degree?)

LIES, DAMNED LIES AND CORRELATION

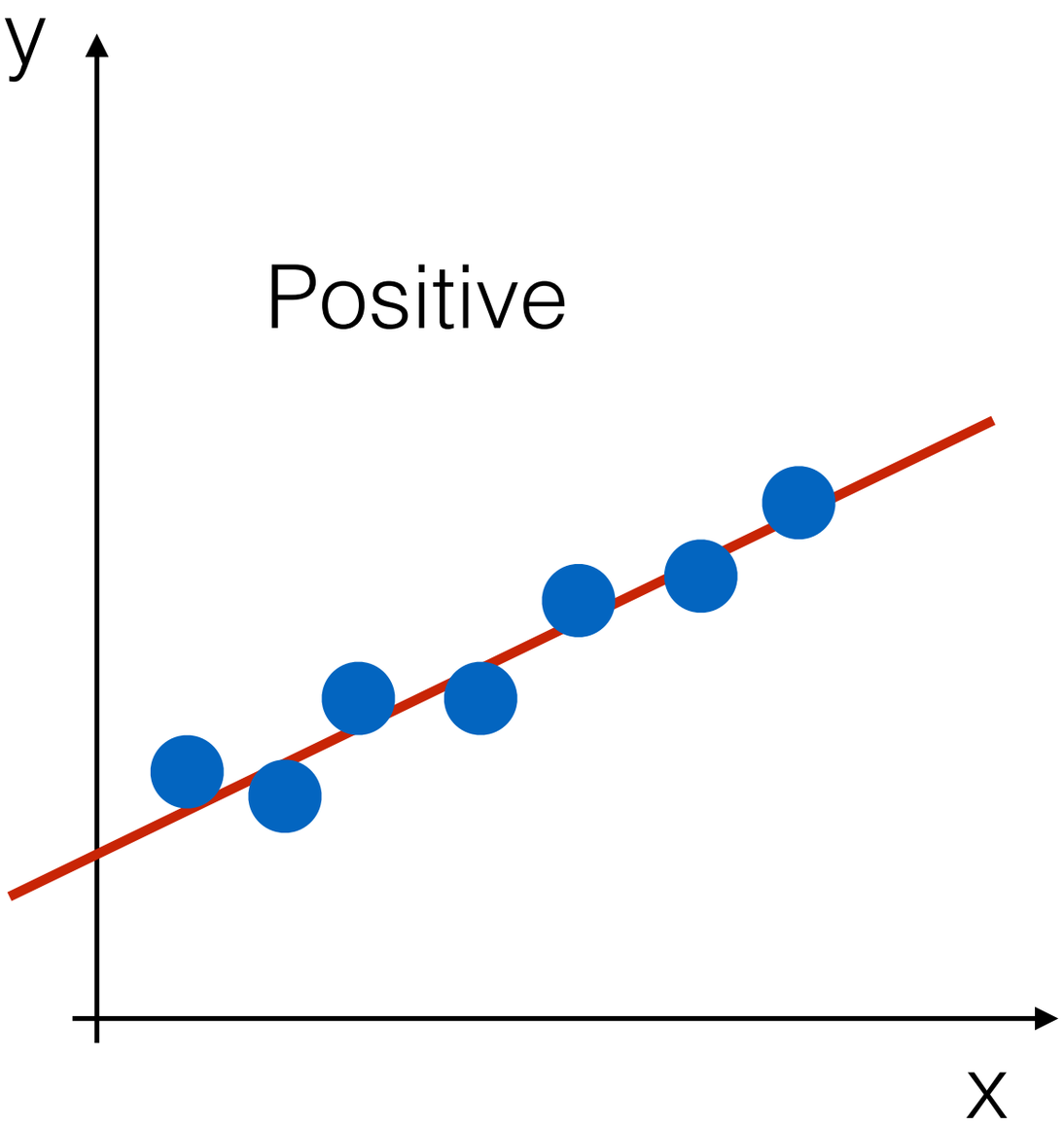
Correlation

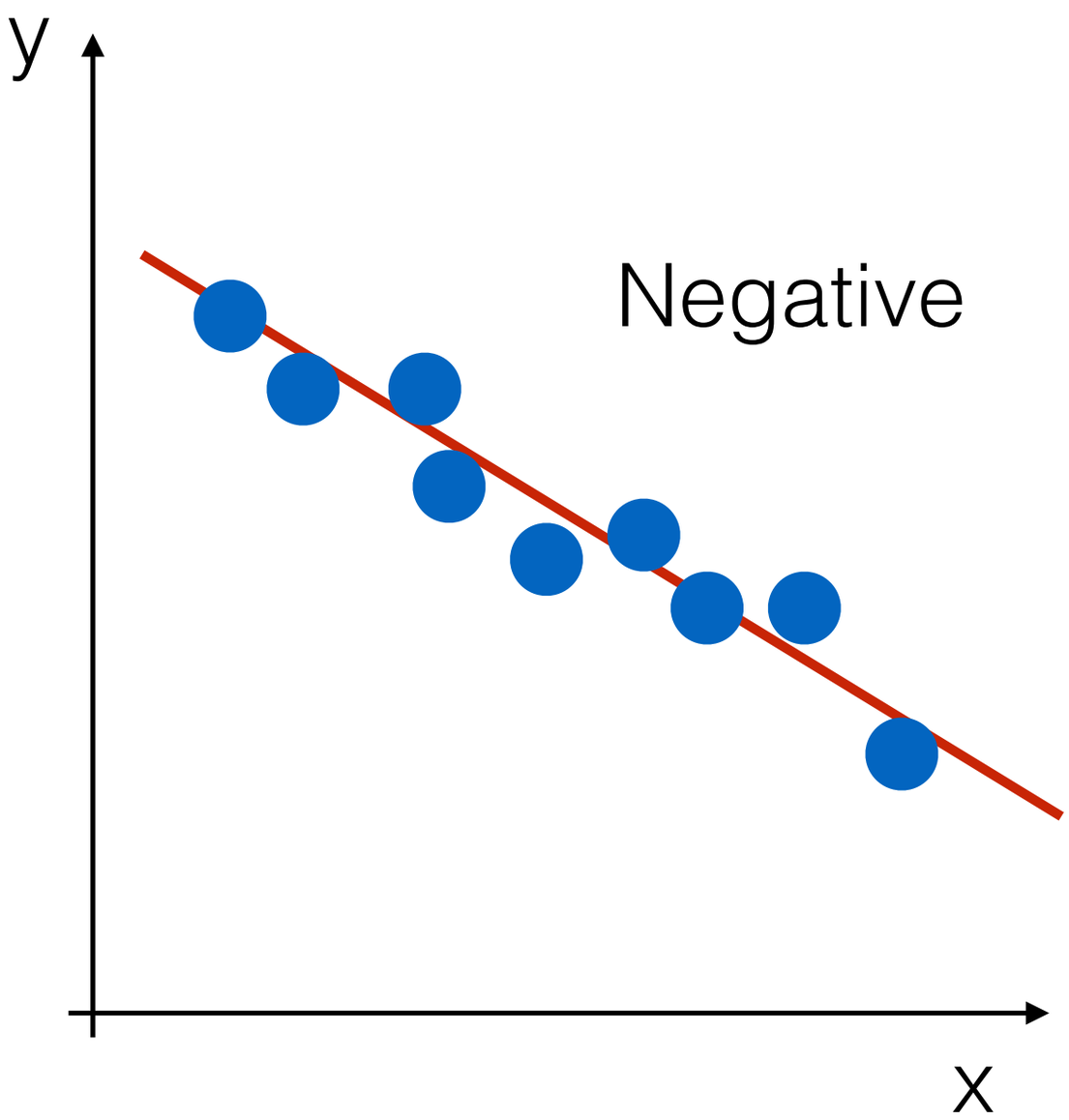
Correlation

- Informal: a connection between two things
- Measure the strength of the association between two variables

Linear Correlation

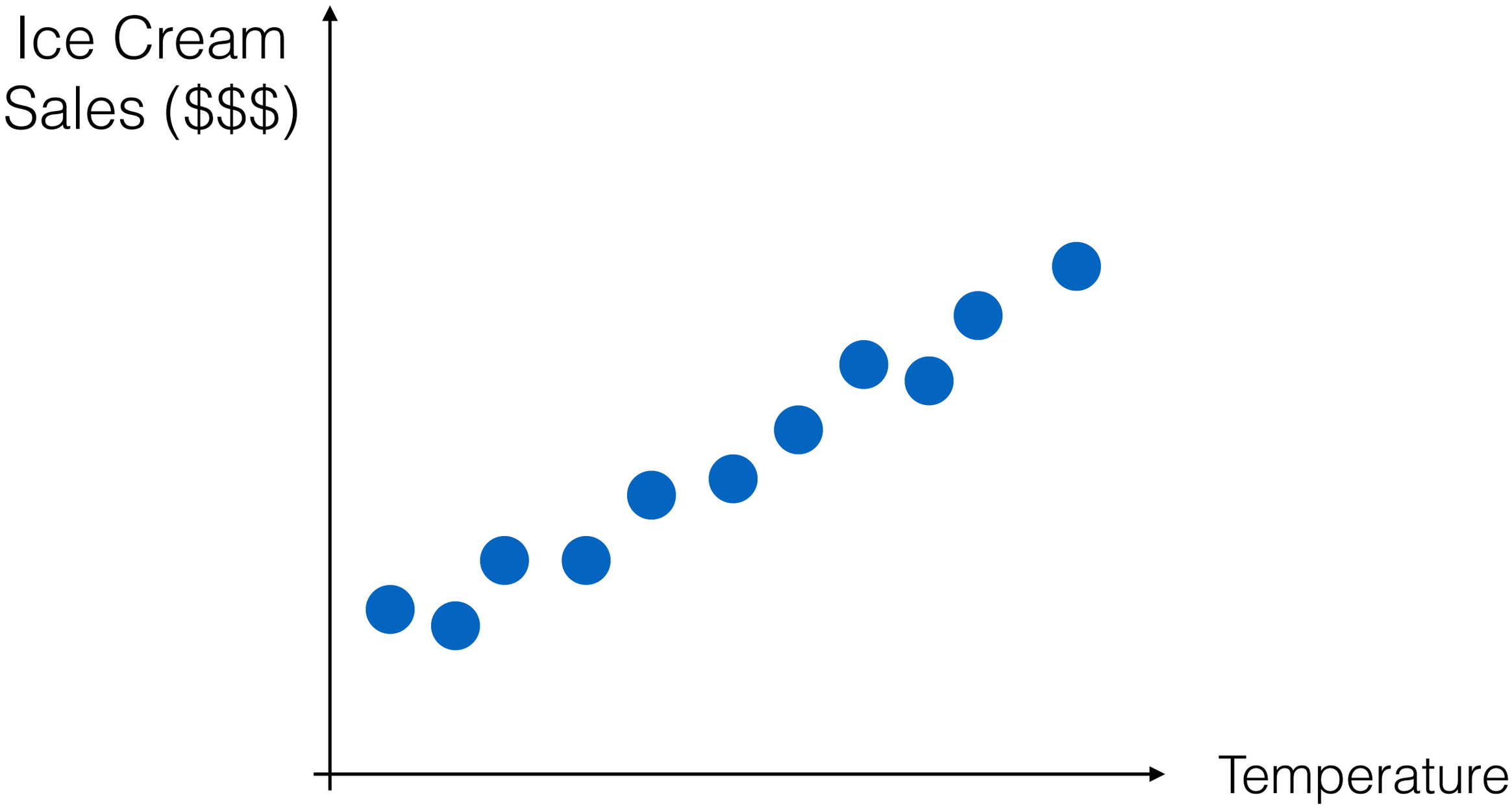
Linear Correlation



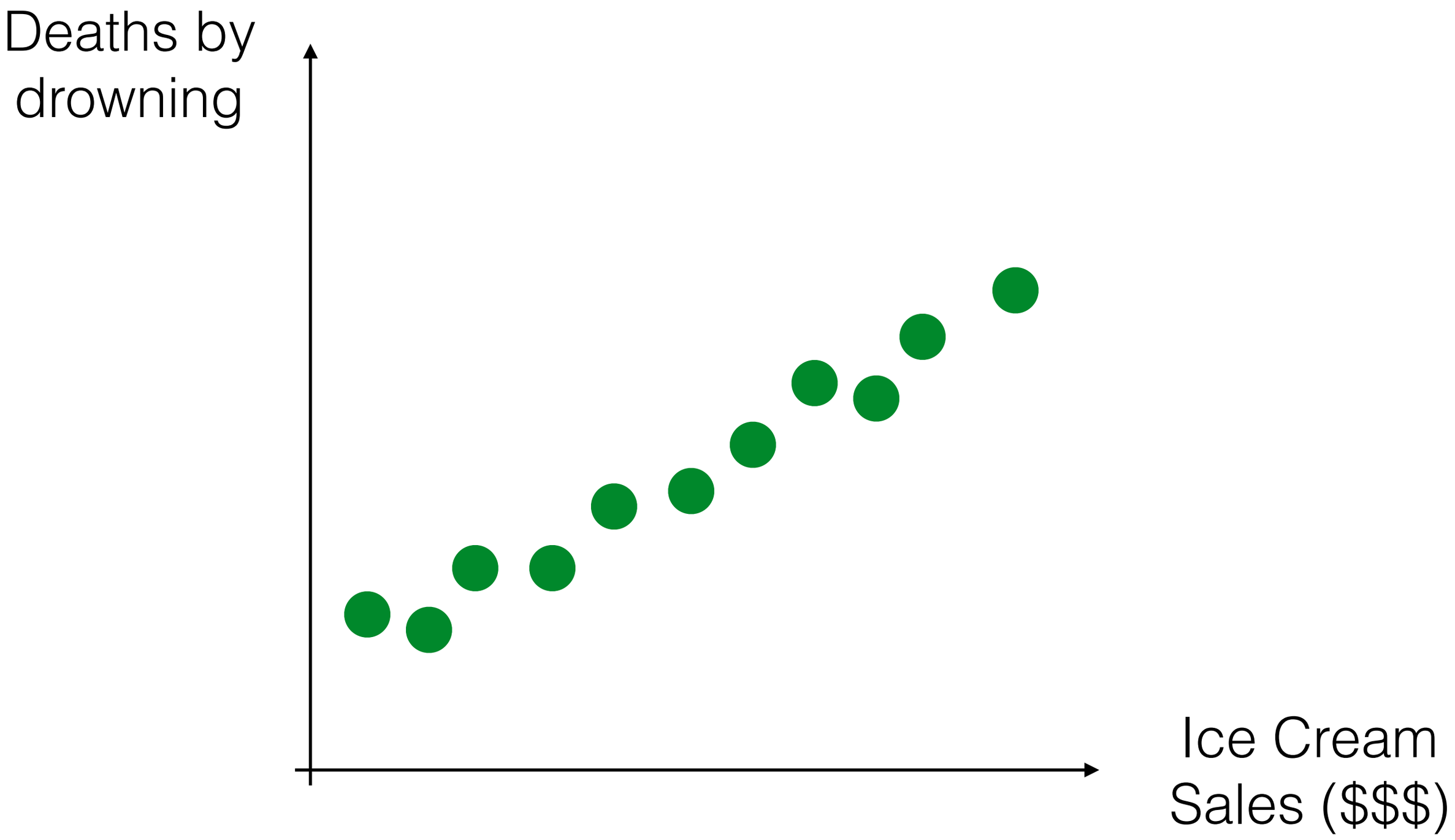


Correlation Example

Correlation Example

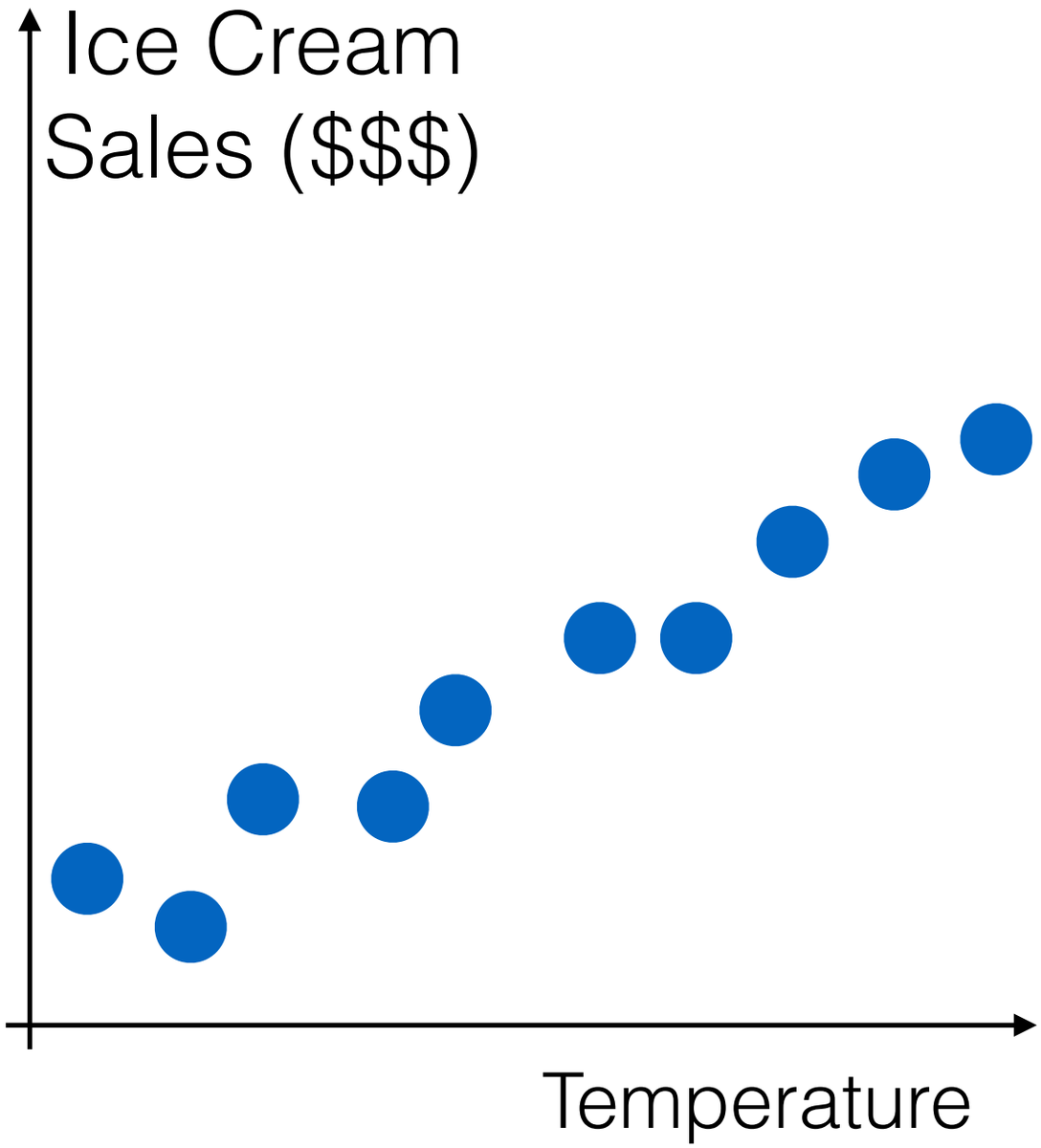


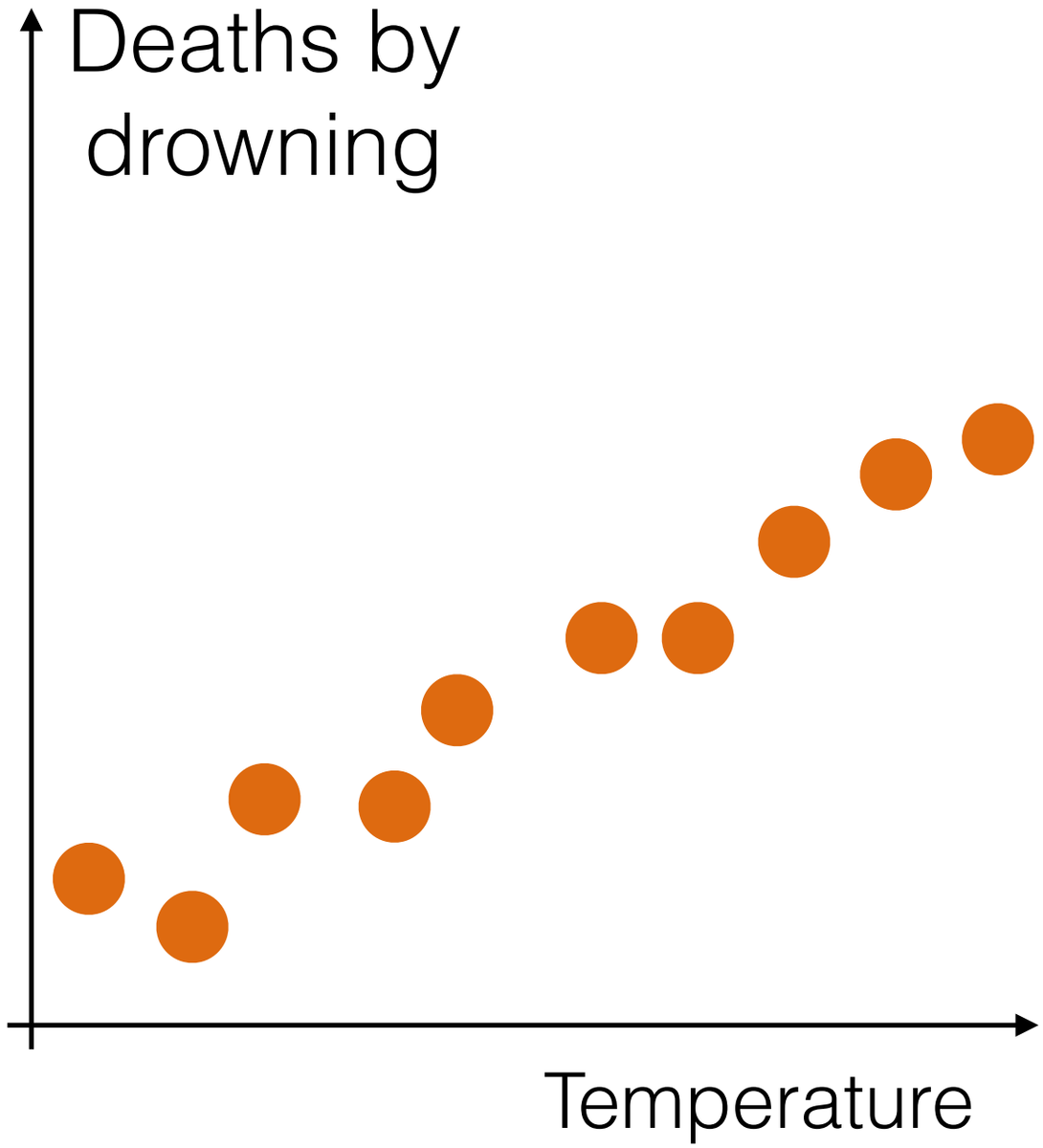
“Correlation does not imply causation”


Deaths by drowning
Ice Cream Sales ($$$)

Lurking Variable

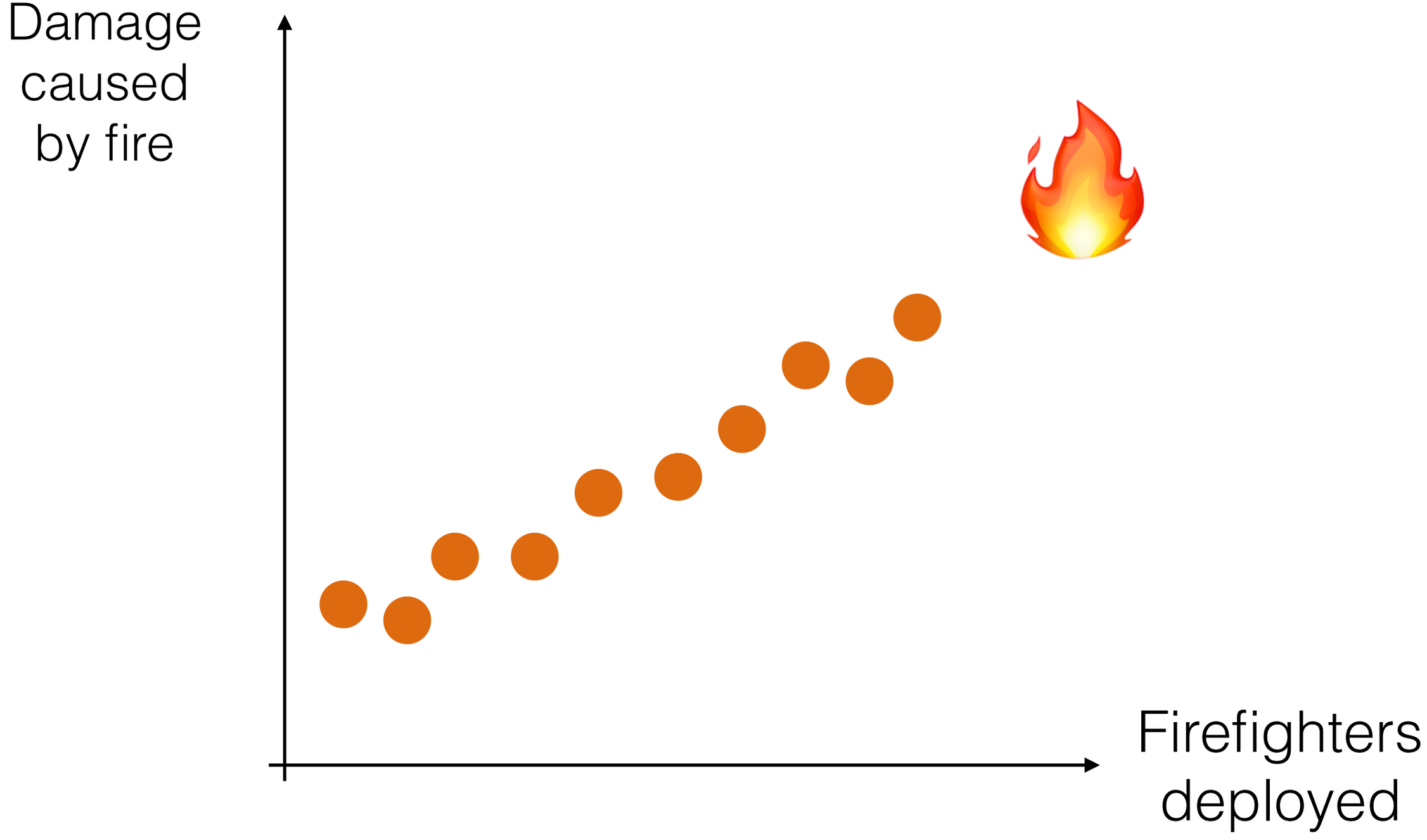
Lurking Variable



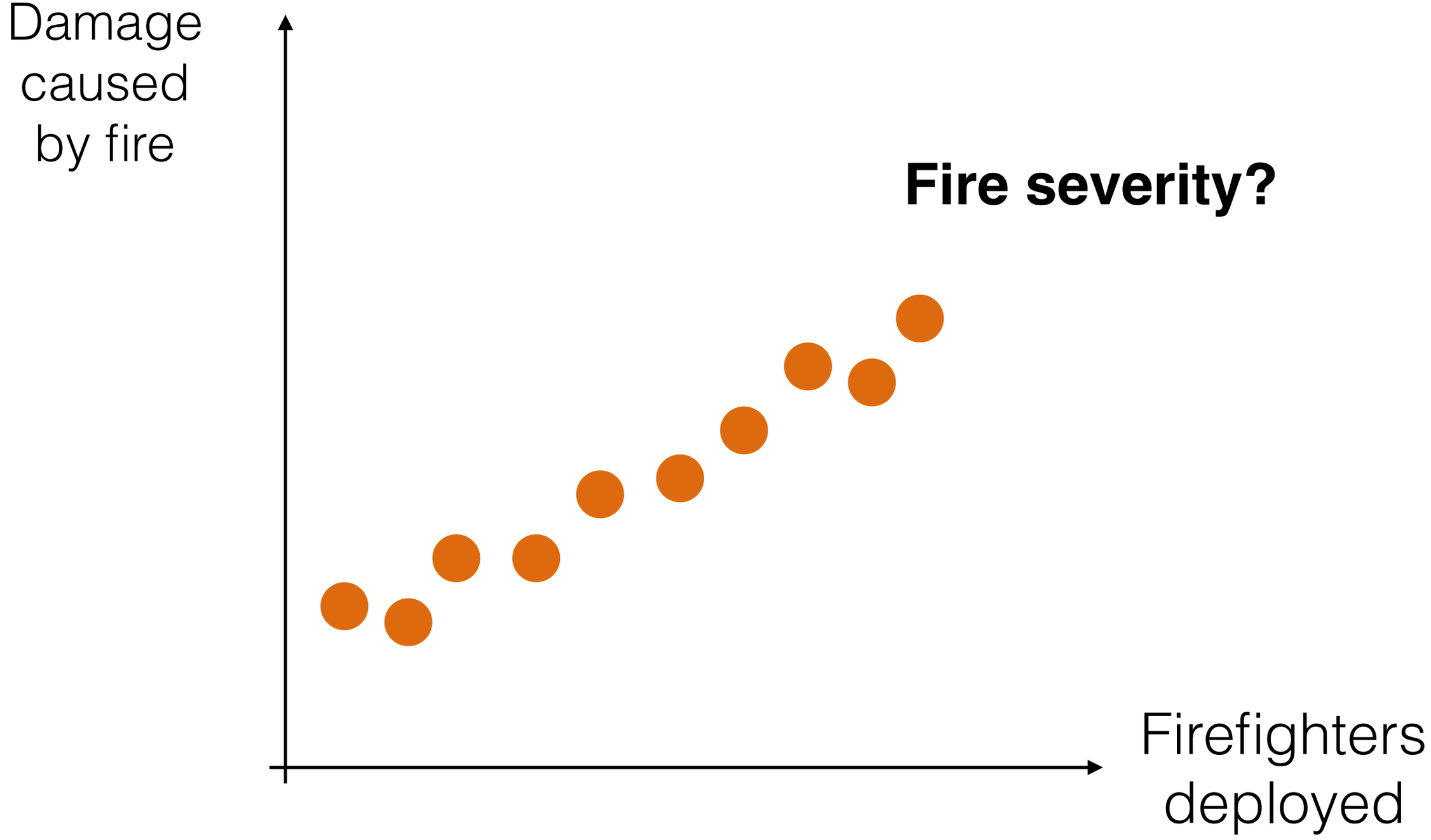


More Lurking Variables

More Lurking Variables



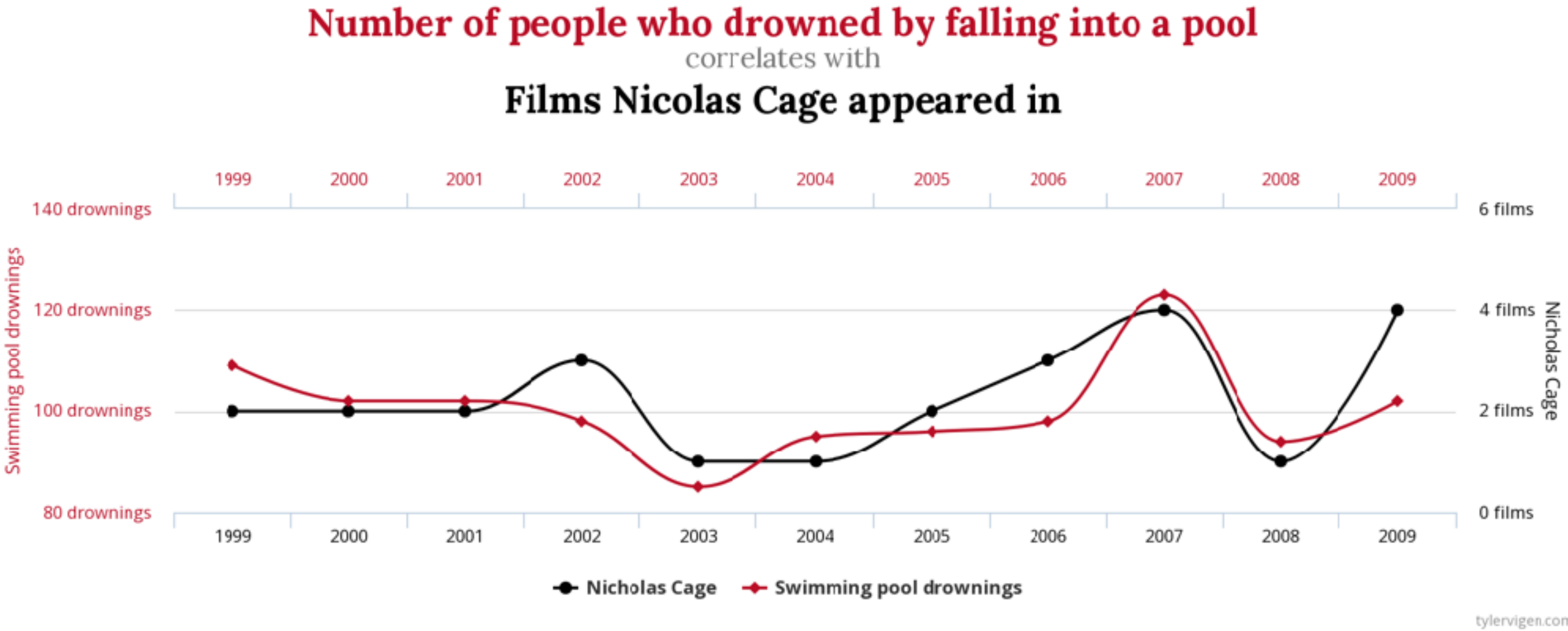
More Lurking Variables



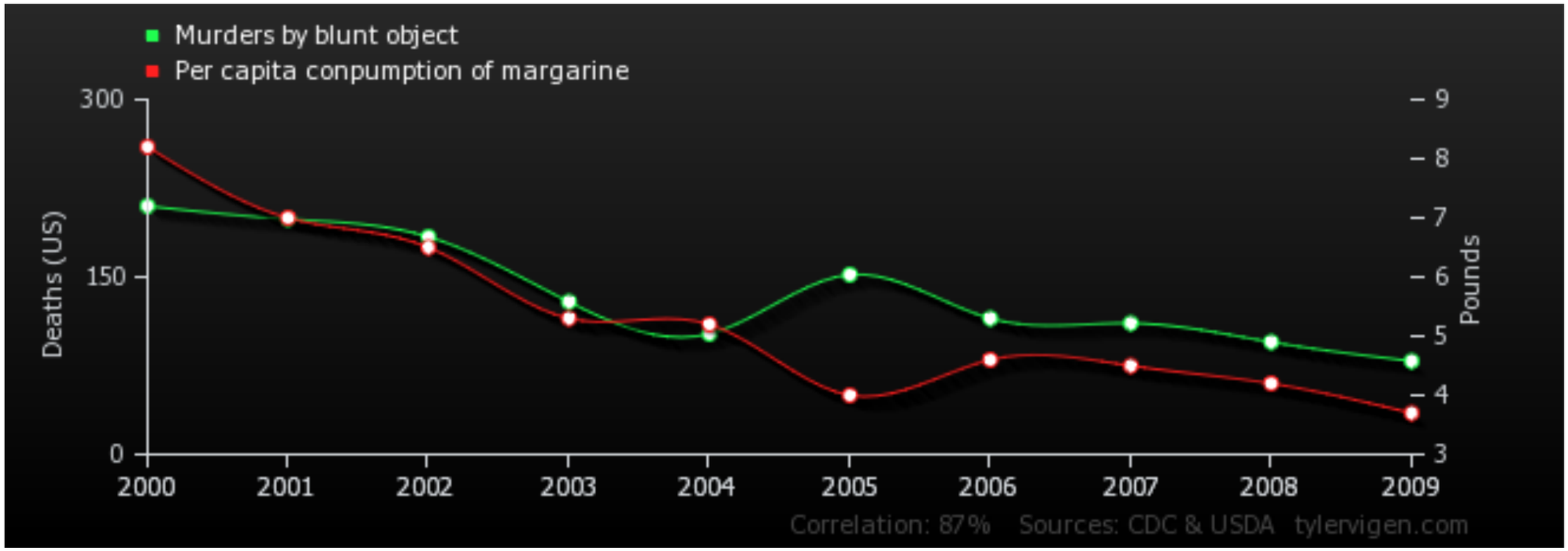
Correlation and causation

Correlation and causation

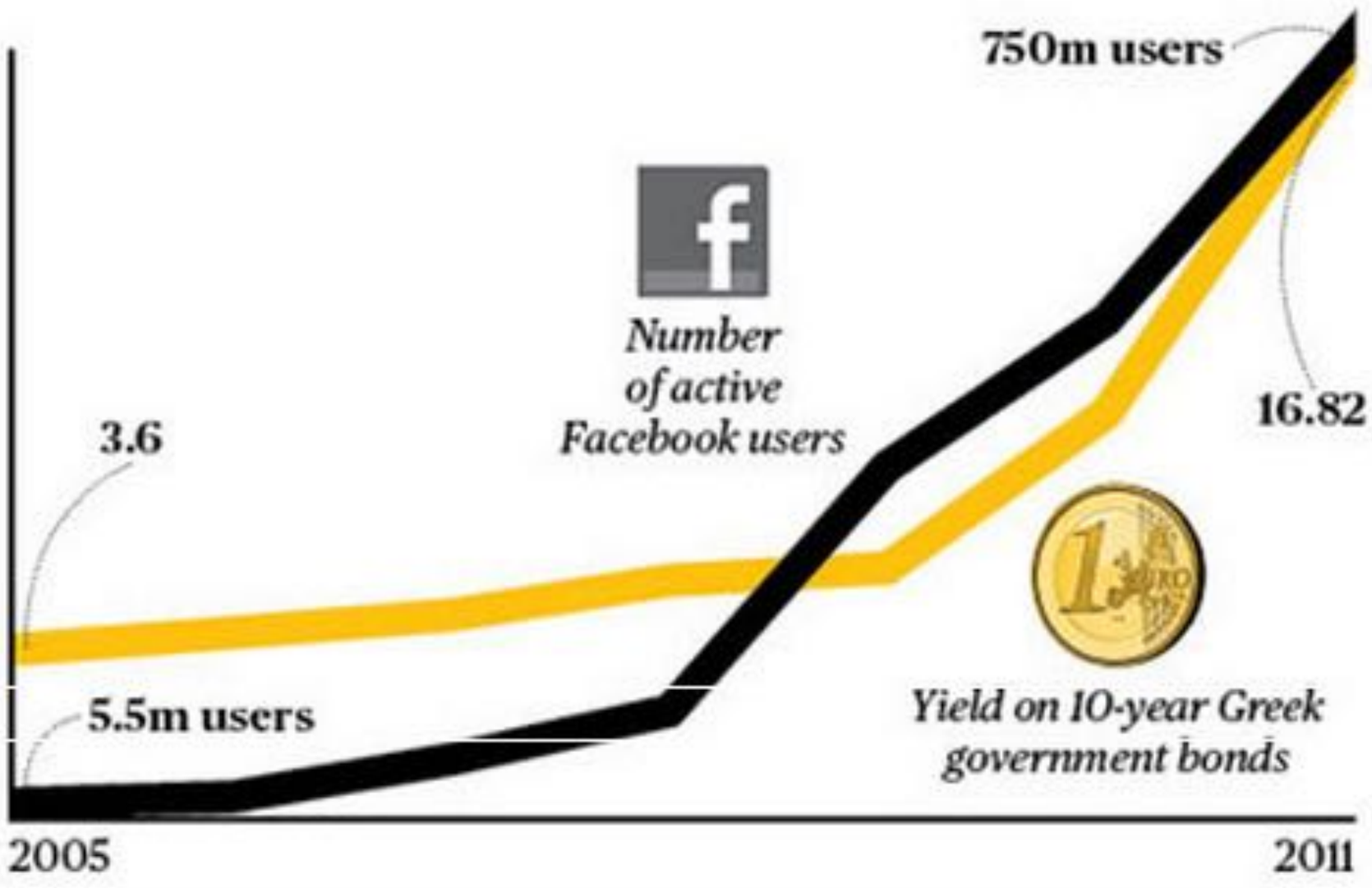
- A causes B, or B causes A
- A and B both cause C
- C causes A and B
- A causes C, and C causes B
- No connection between A and B



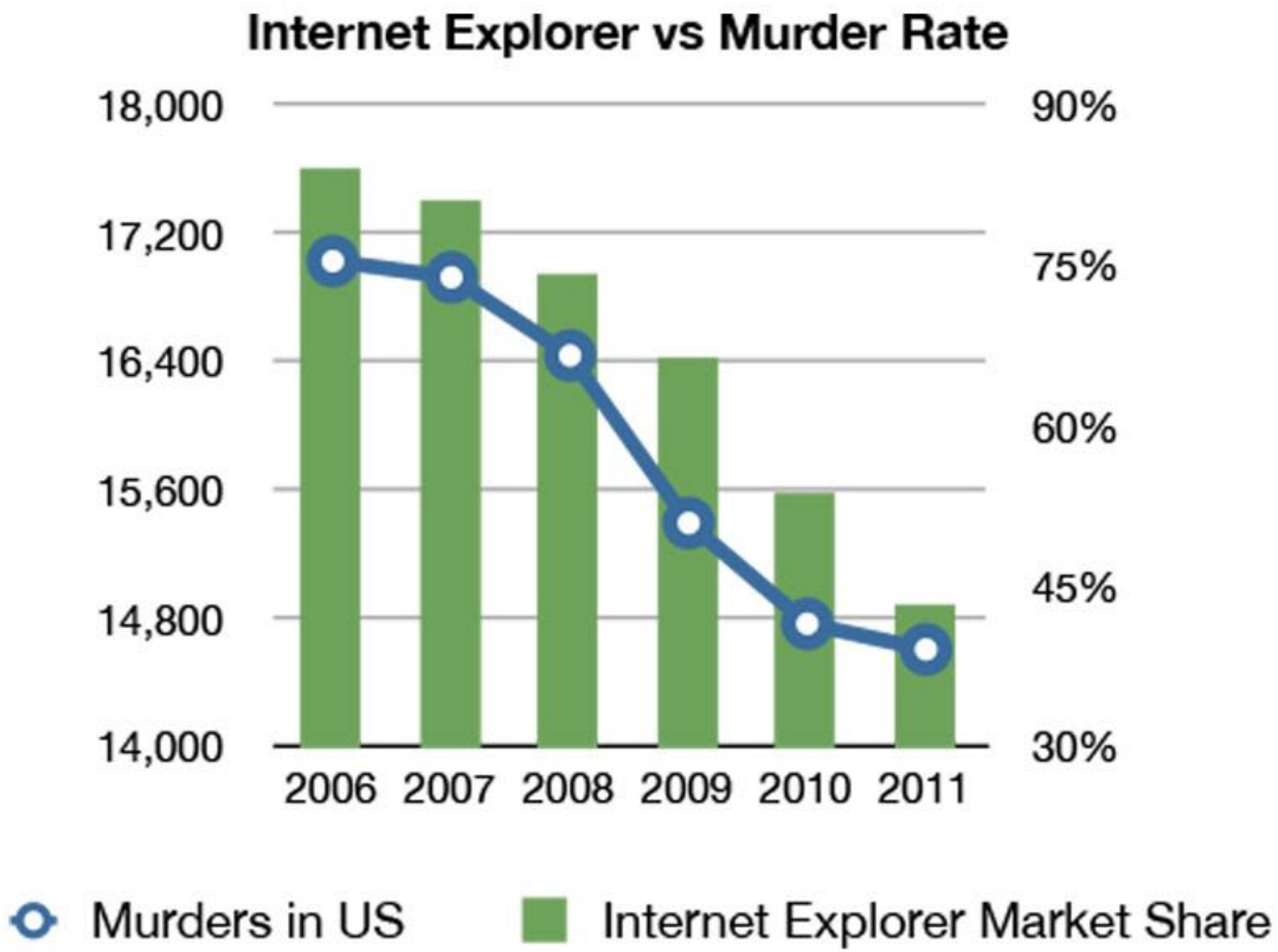
http://www.tylervigen.com/spurious-correlations



http://www.tylervigen.com/spurious-correlations



https://www.buzzfeed.com/kjh2110/the-10-most-bizarre-correlations



https://www.buzzfeed.com/kjh2110/the-10-most-bizarre-correlations

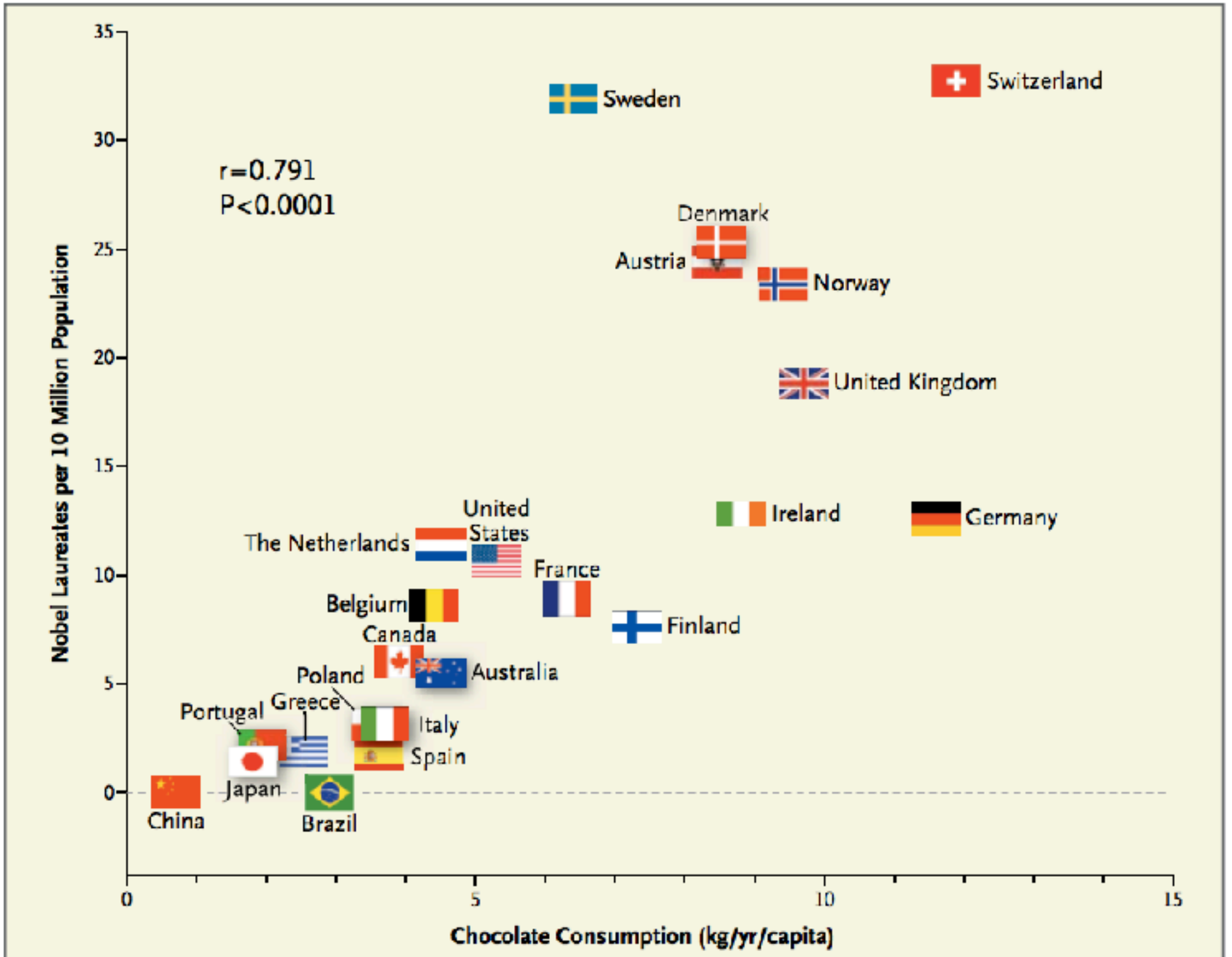


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

http://www.nejm.org/doi/full/10.1056/NEJMon1211064

LIES, DAMNED LIES, SLICING AND DICING YOUR DATA

Simpson's Paradox

University of California, Berkeley Graduate school admissions in 1973

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

https://en.wikipedia.org/wiki/Simpson%27s_paradox

University of California, Berkeley Graduate school admissions in 1973

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

Gender bias?

https://en.wikipedia.org/wiki/Simpson%27s_paradox

University of California, Berkeley Graduate school admissions in 1973

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

https://en.wikipedia.org/wiki/Simpson%27s_paradox

University of California, Berkeley Graduate school admissions in 1973

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

https://en.wikipedia.org/wiki/Simpson%27s_paradox

University of California, Berkeley Graduate school admissions in 1973

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

https://en.wikipedia.org/wiki/Simpson%27s_paradox

University of California, Berkeley Graduate school admissions in 1973

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

https://en.wikipedia.org/wiki/Simpson%27s_paradox

LIES, DAMNED LIES AND SAMPLING BIAS

Sampling

Sampling

- A selection of a subset of individuals
- Purpose: estimate about the whole population
- Hello Big Data!

Bias

Bias

- Prejudice? Intuition?
- Cultural context?
- In science: a systematic error

“Dewey defeats Truman”

“Dewey defeats Truman”

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

“Dewey defeats Truman”

- The Chicago Tribune printed the wrong headline on election night
- The editor trusted the results of the phone survey
- … in 1948, a sample of phone users was not representative of the general population

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

Survivorship Bias

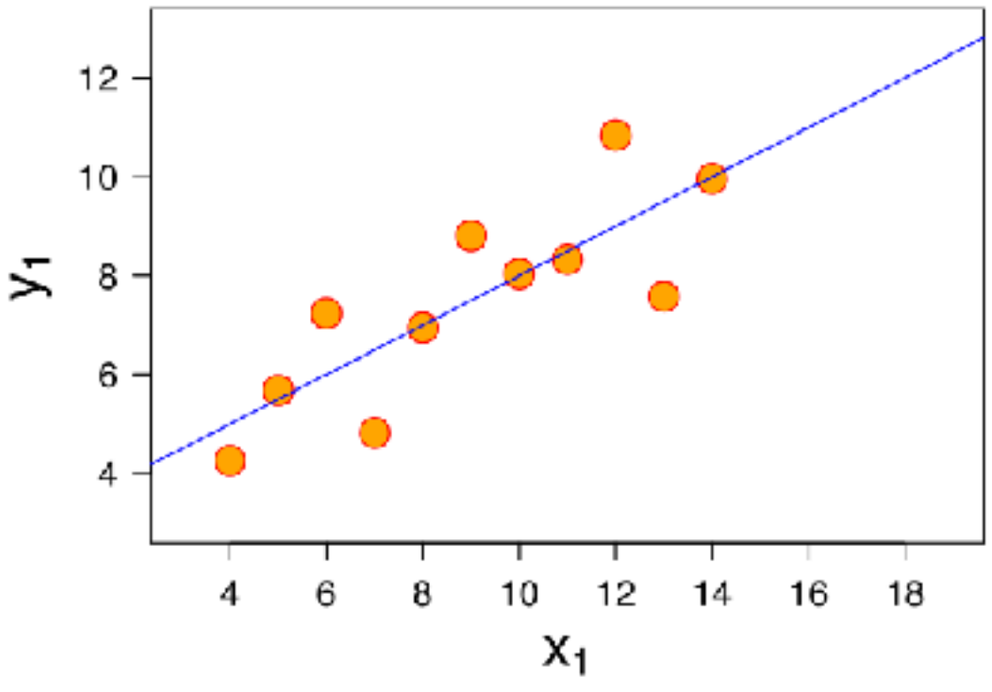
Survivorship Bias

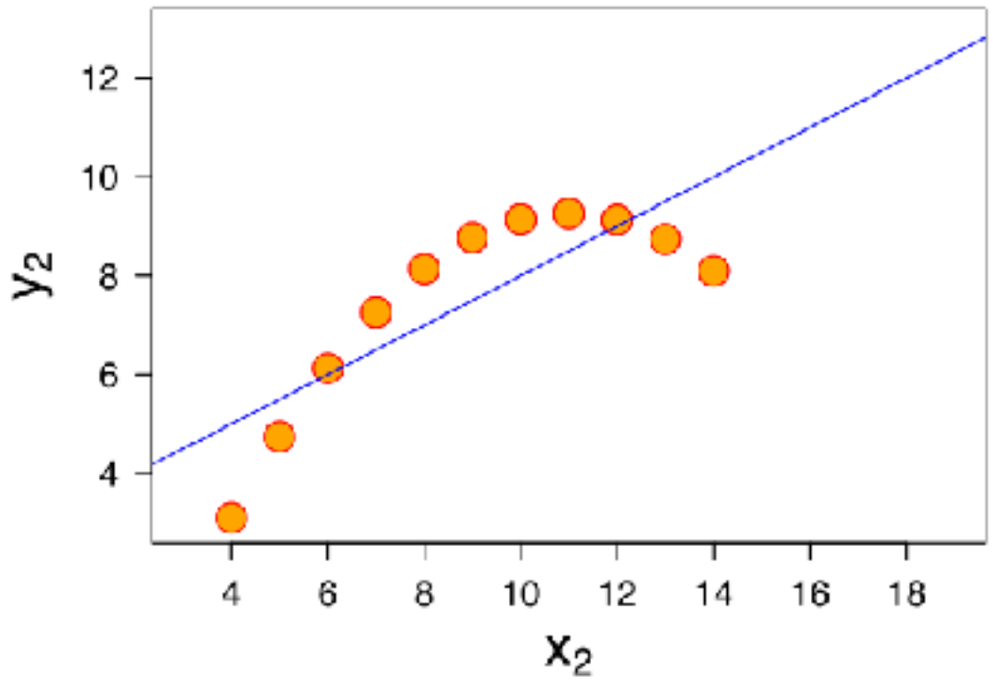
- Bill Gates, Steve Jobs, Mark Zuckerberg are all college drop-outs

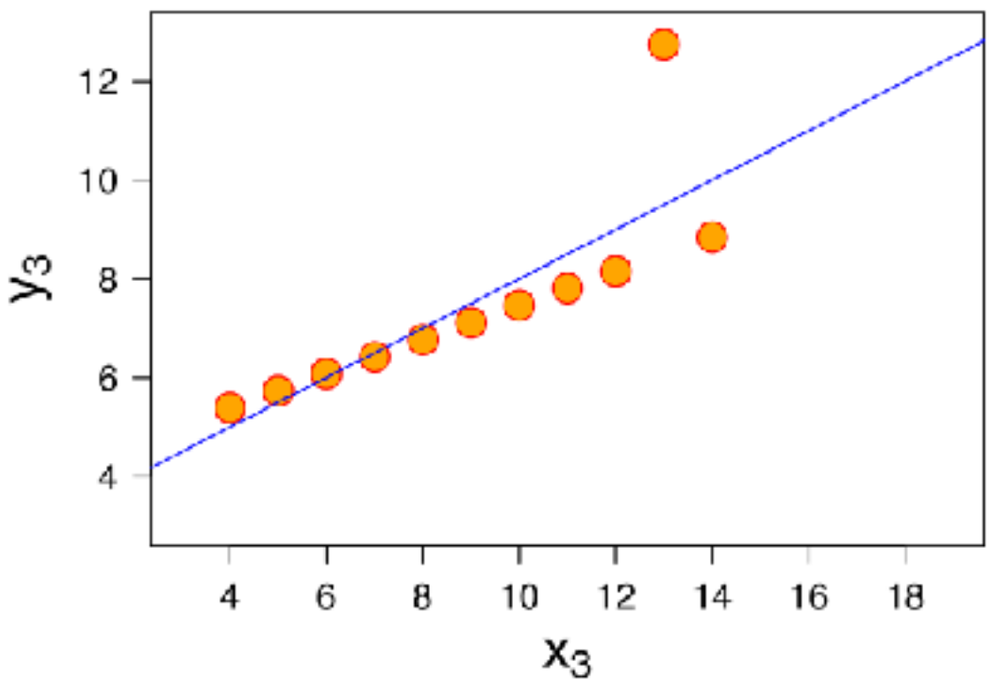
- … should you quit studying?

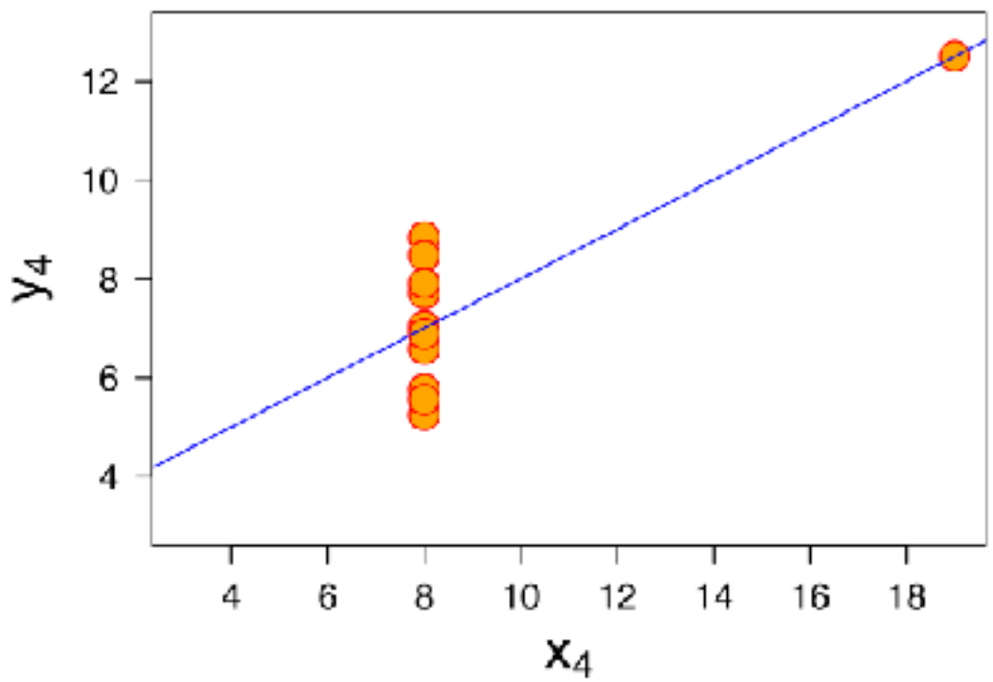
LIES, DAMNED LIES AND DATAVIZ

“A picture is worth a thousand words”

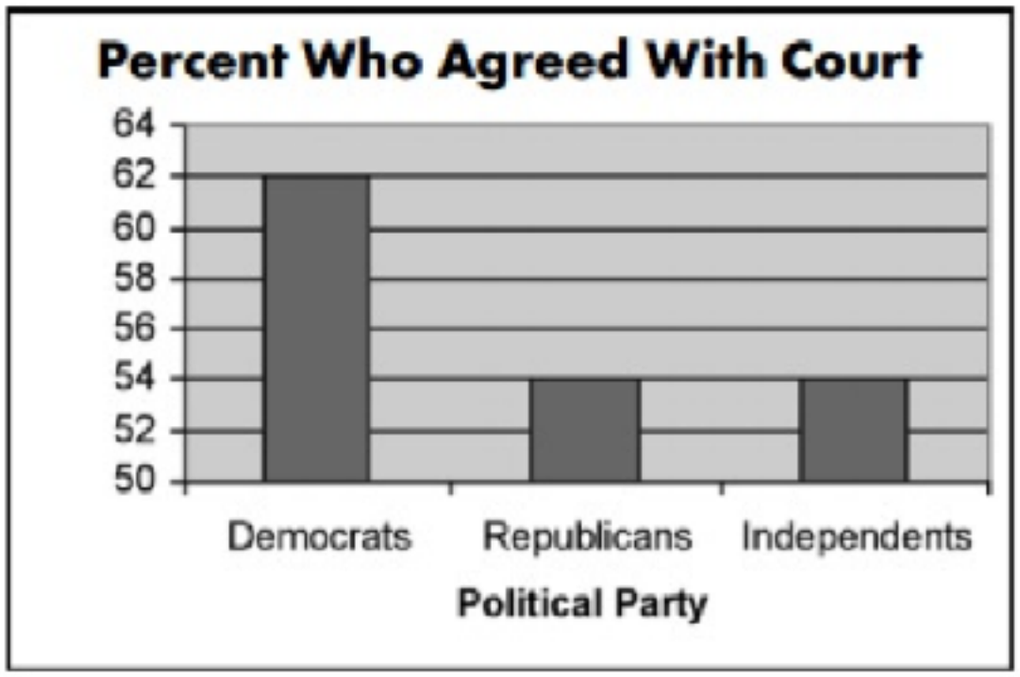




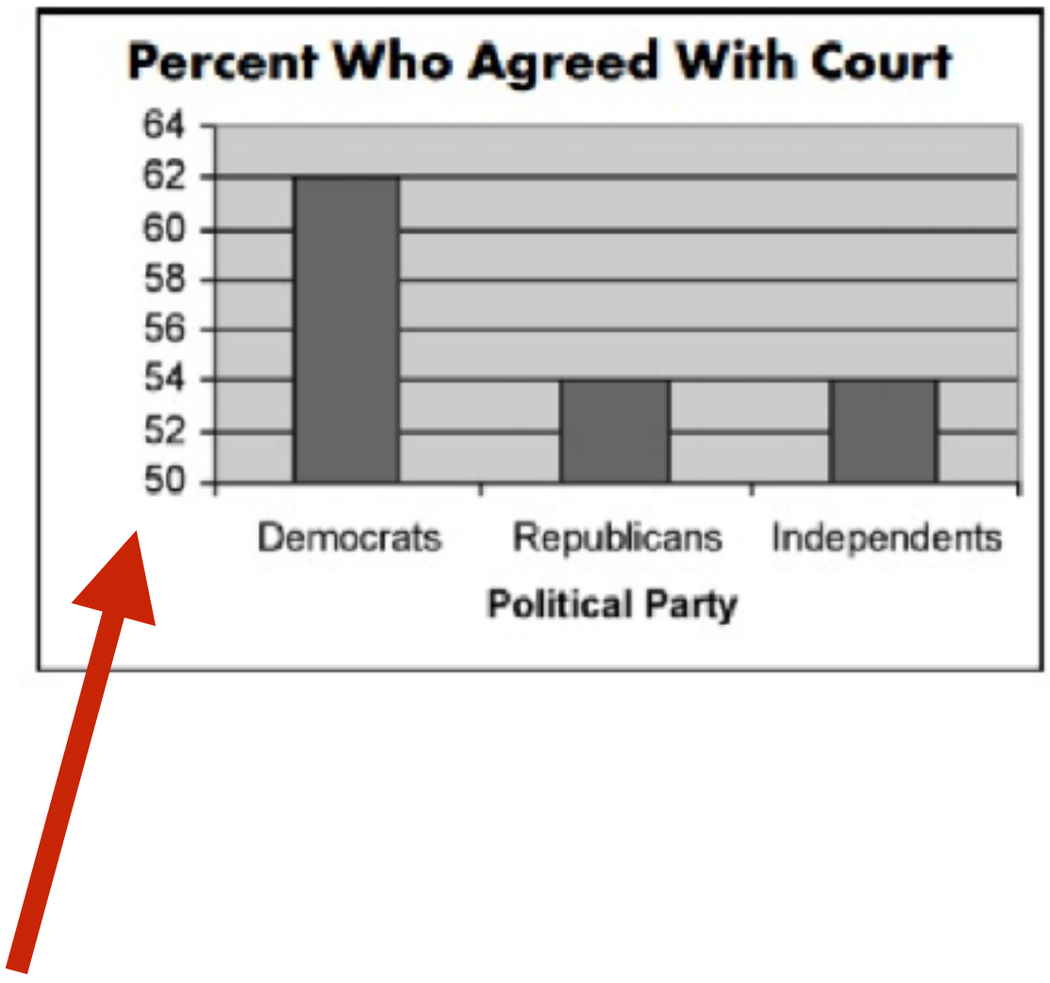




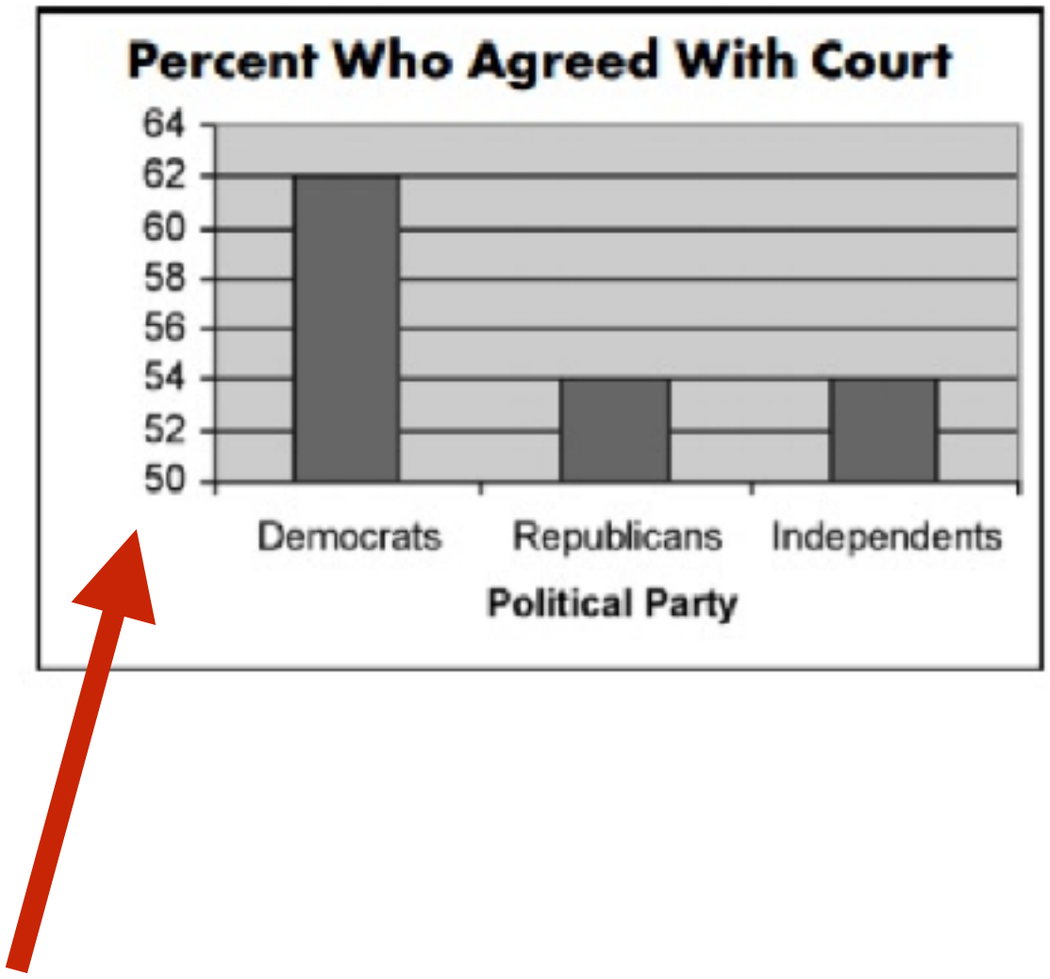
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

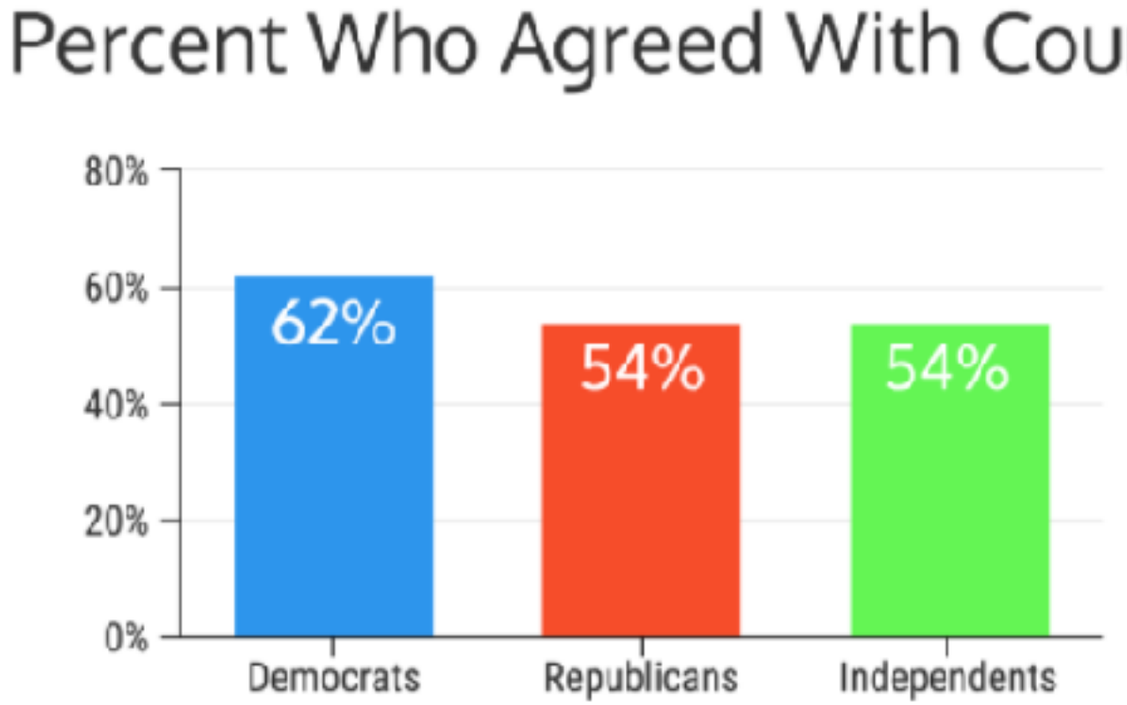


https://venngage.com/blog/misleading-graphs/

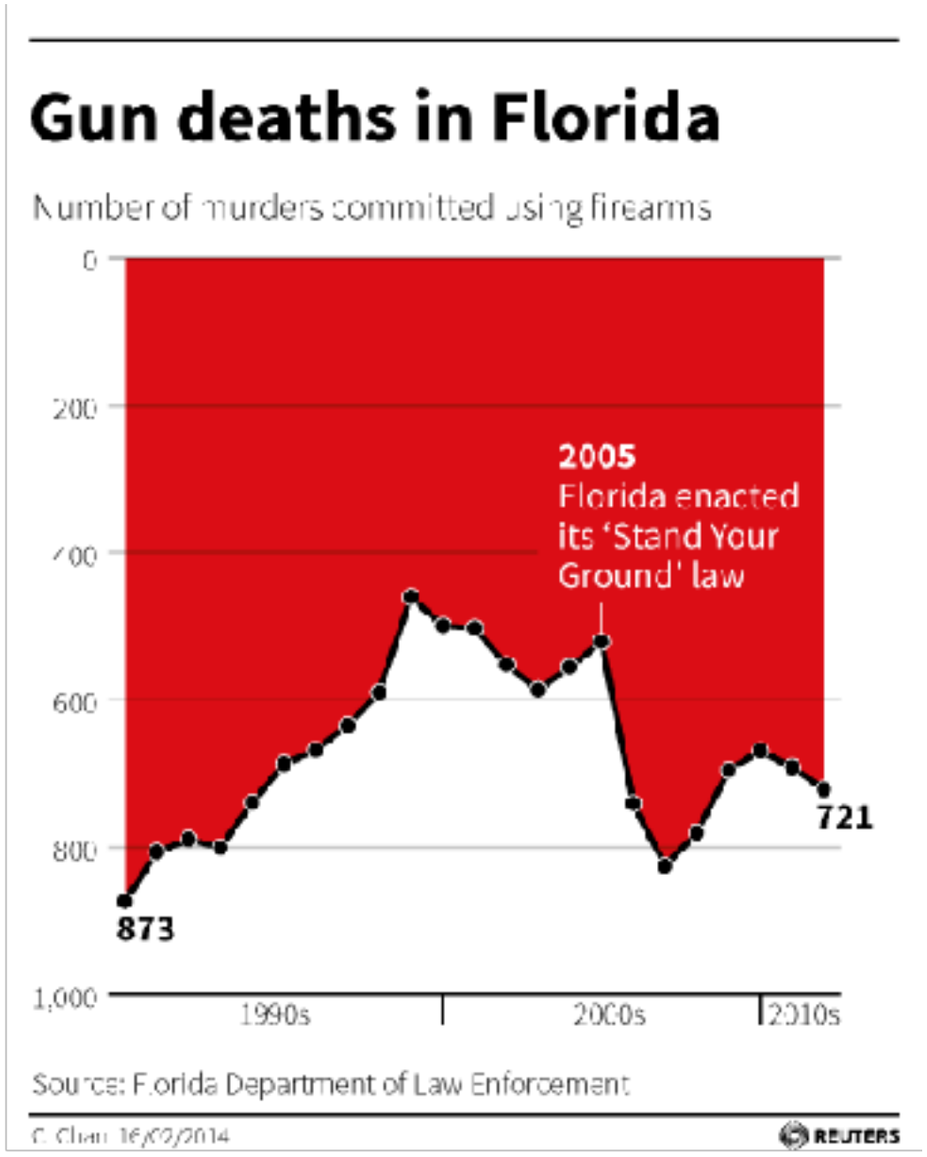


https://venngage.com/blog/misleading-graphs/

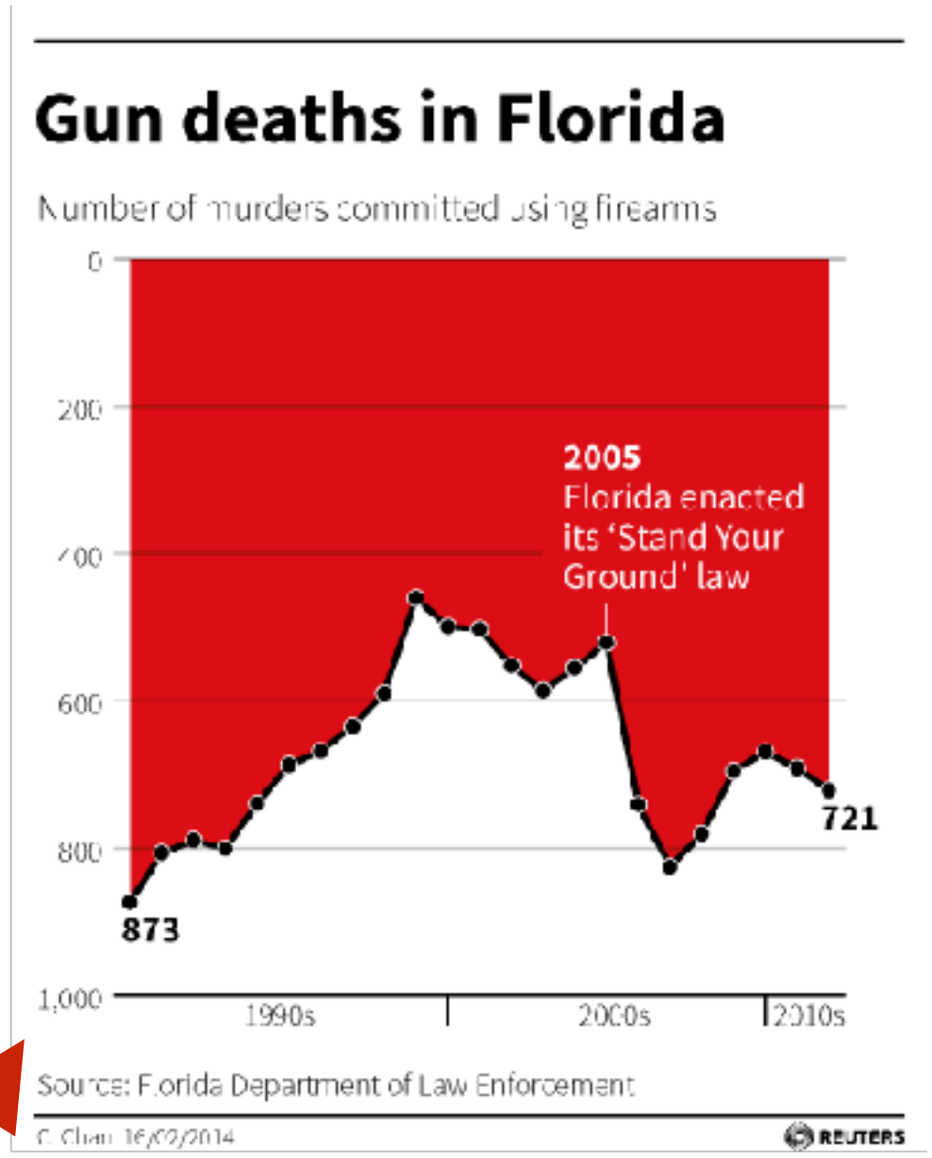




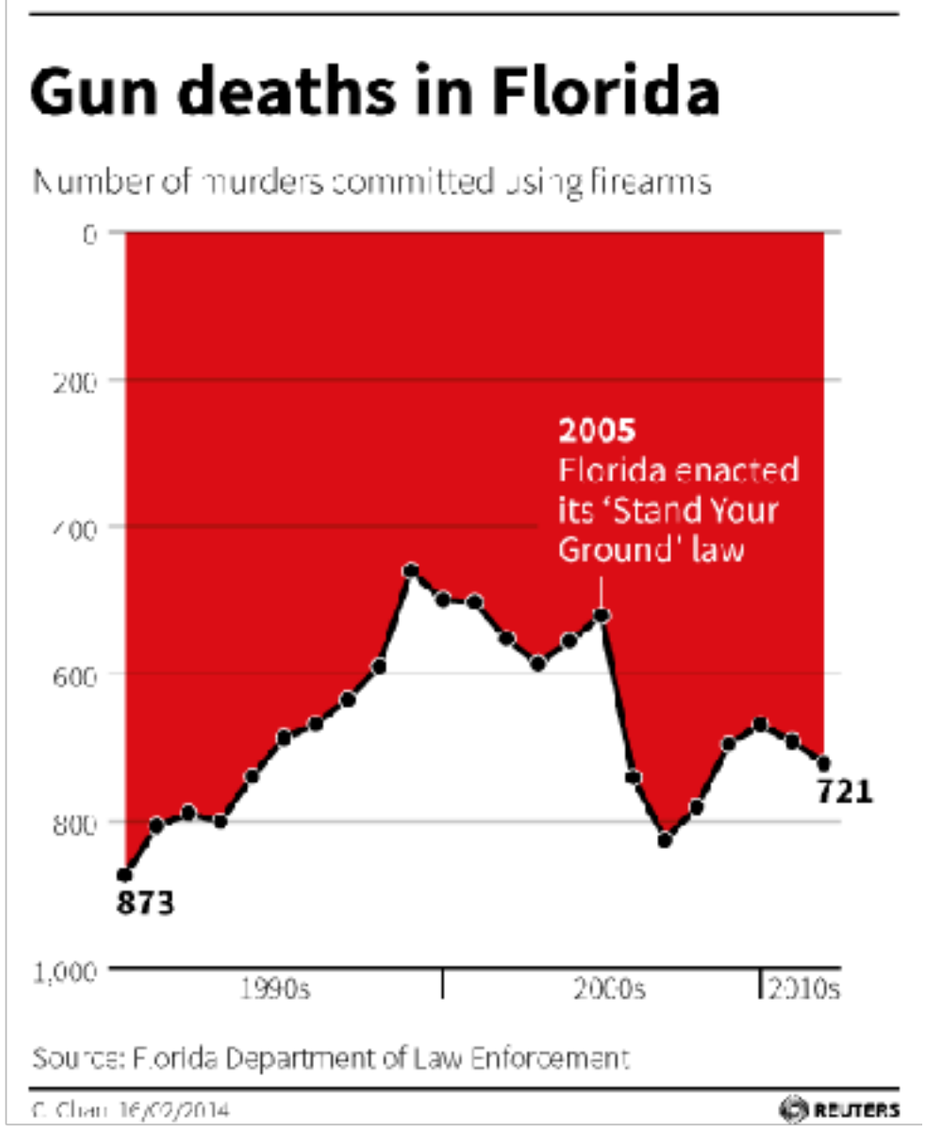
https://venngage.com/blog/misleading-graphs/

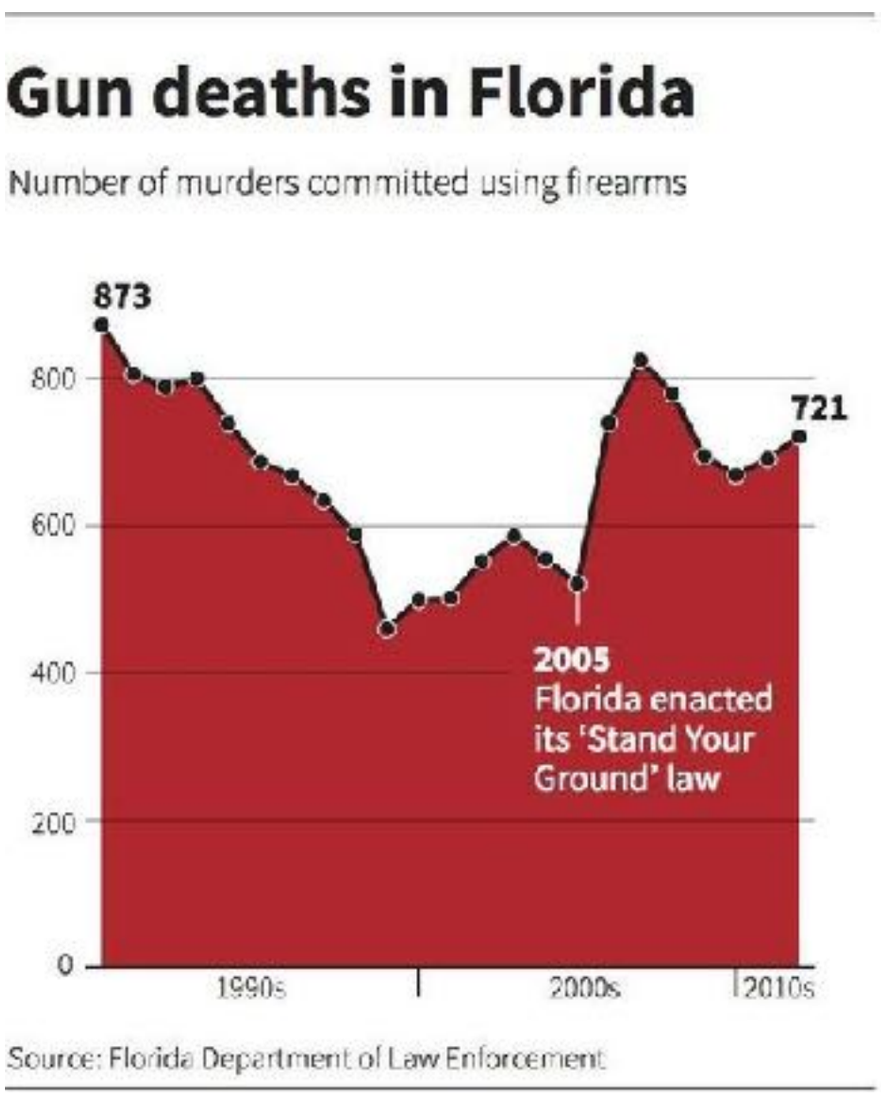


http://www.businessinsider.com/gun-deaths-in-florida-increased-with-stand-your-ground-2014-2?IR=T



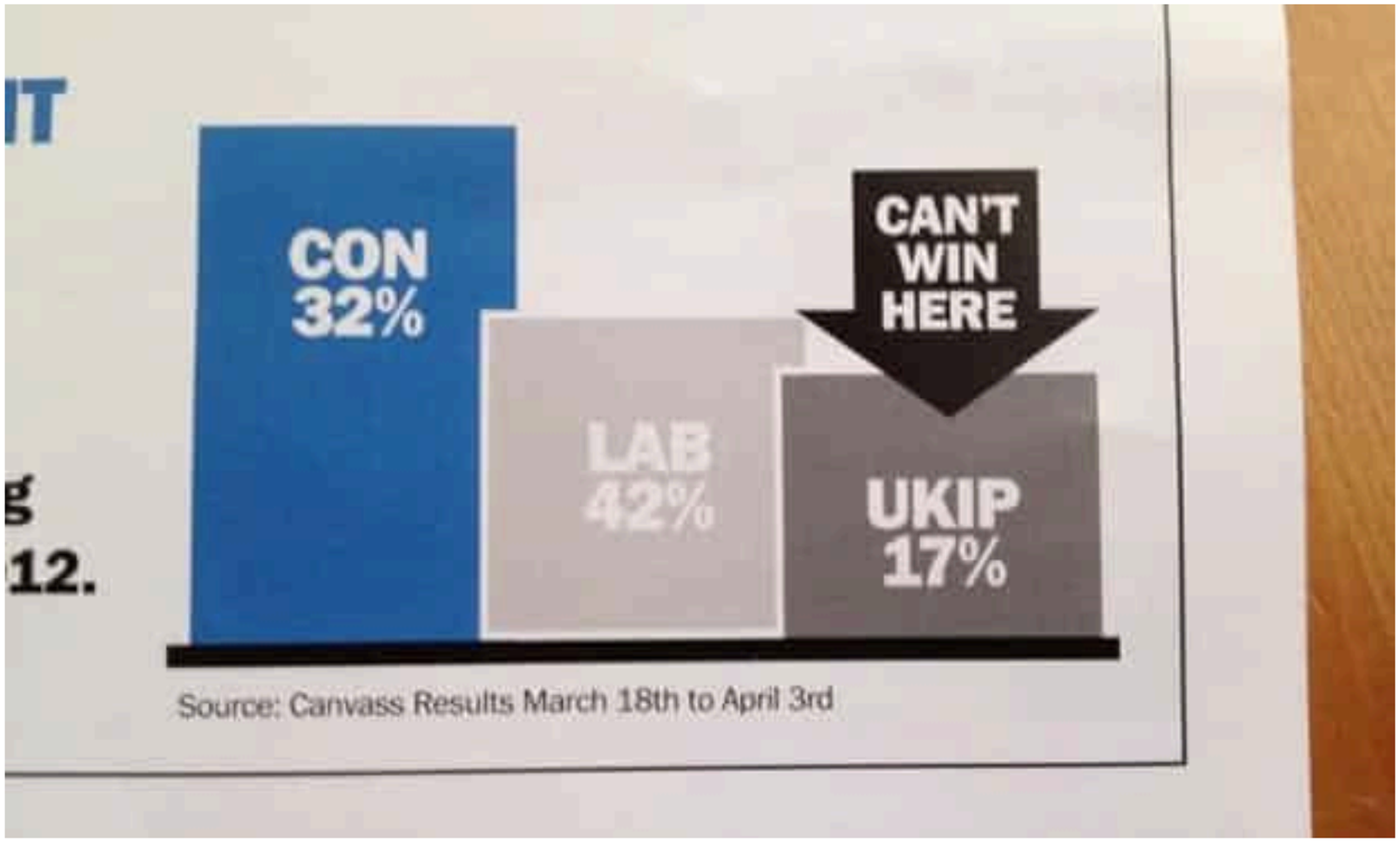
http://www.businessinsider.com/gun-deaths-in-florida-increased-with-stand-your-ground-2014-2?IR=T



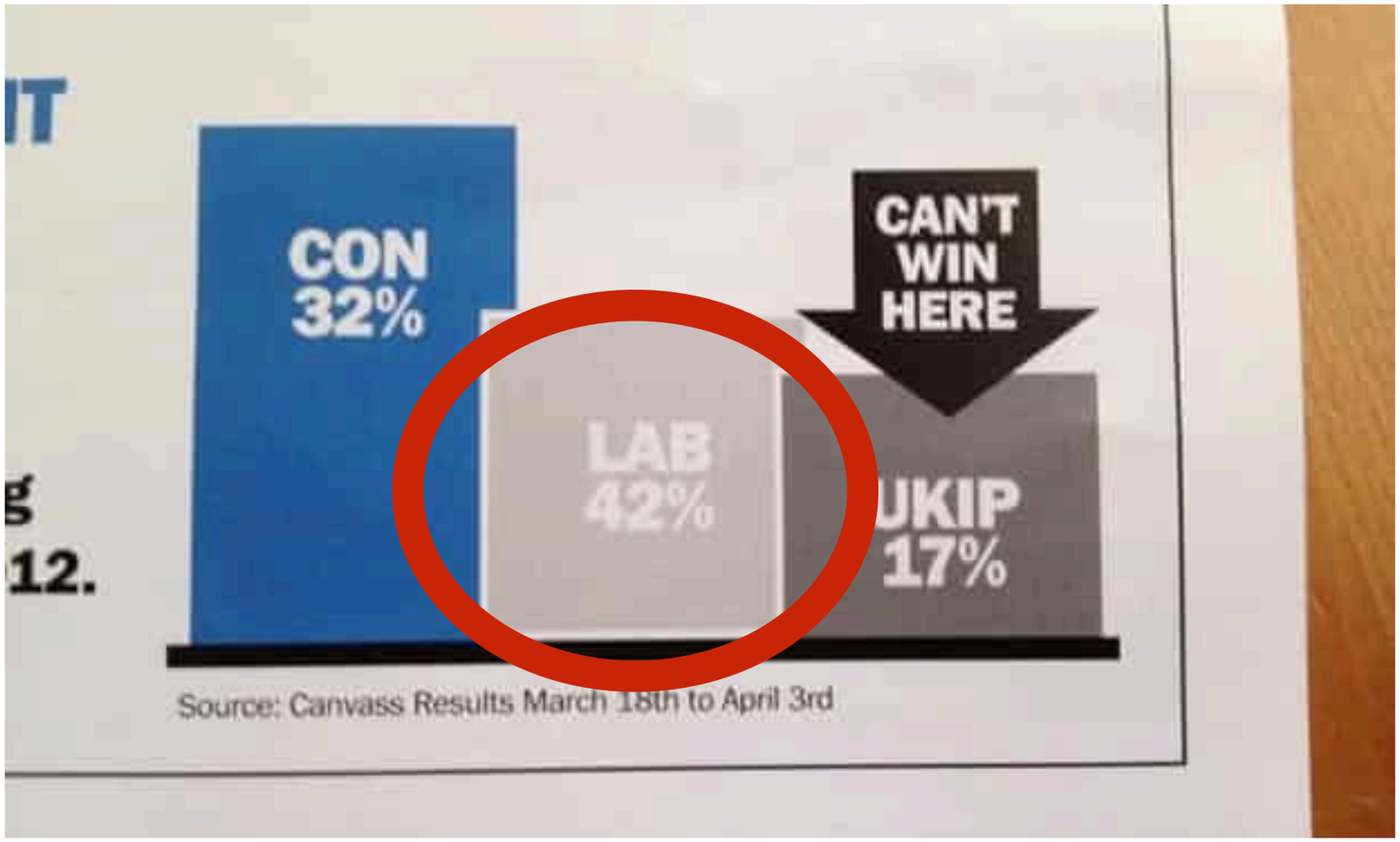


http://www.businessinsider.com/gun-deaths-in-florida-increased-with-stand-your-ground-2014-2?IR=T

https://www.raiplay.it/video/2016/04/Agor224-del-08042016-4d84cebb-472c-442c-82e0-df25c7e4d0ce.html

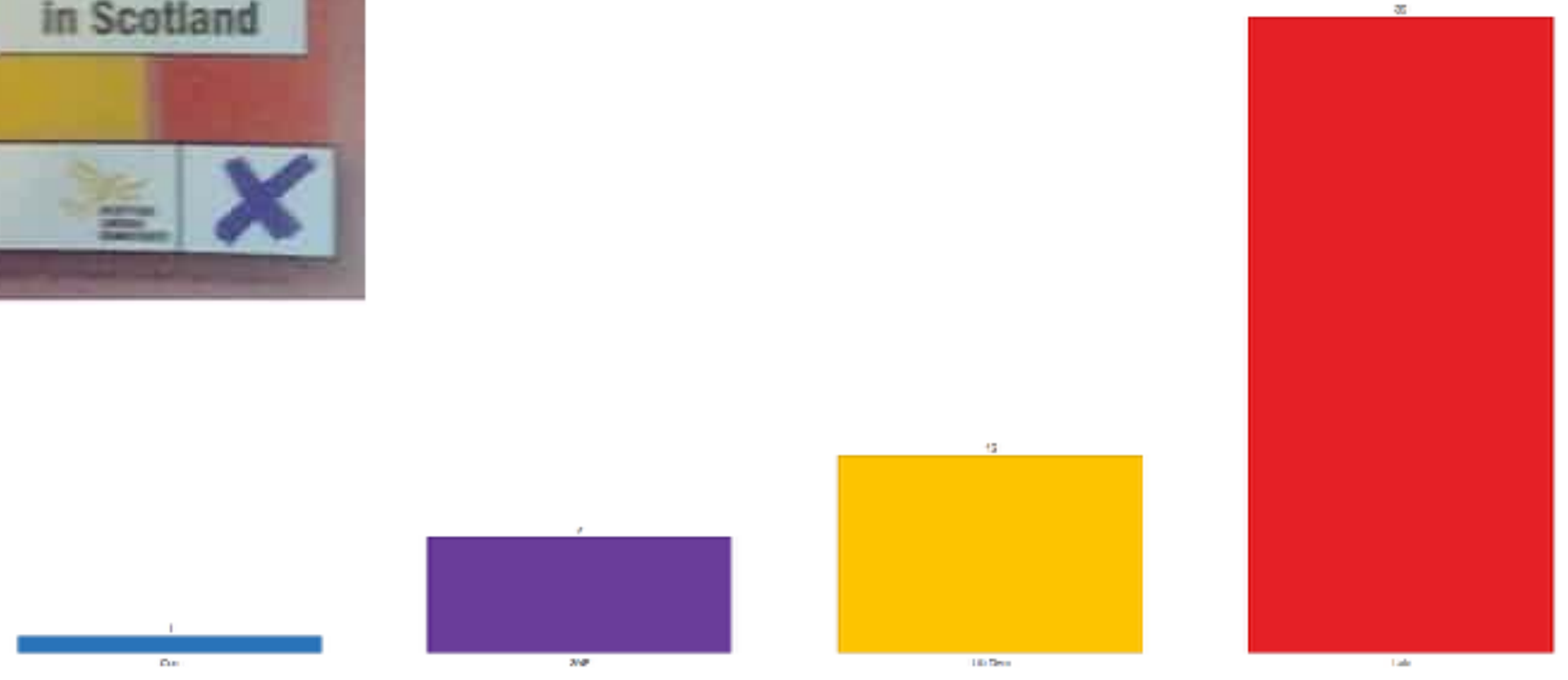


https://www.theguardian.com/news/datablog/2014/may/12/lies-election-leaflets-five-tricks-european-elections



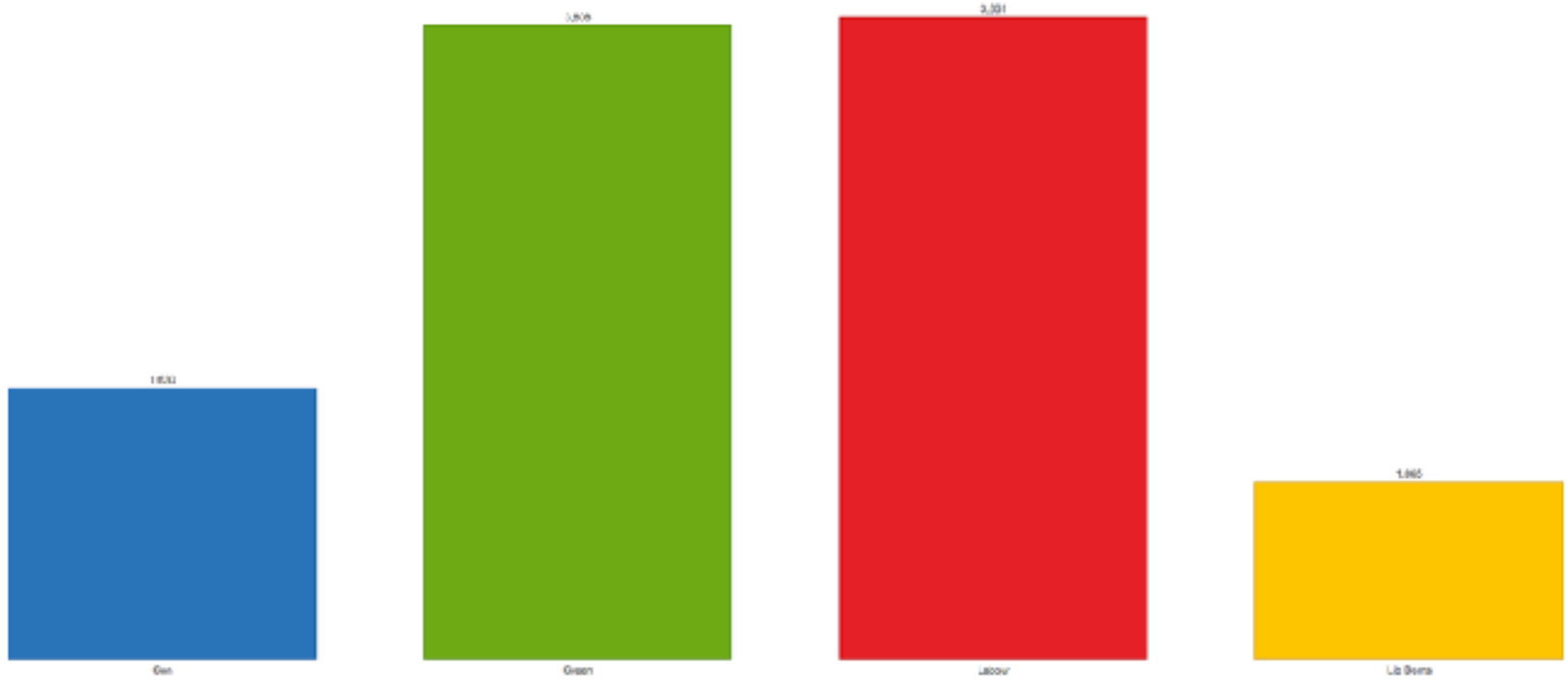
https://www.theguardian.com/news/datablog/2014/may/12/lies-election-leaflets-five-tricks-european-elections

https://www.theguardian.com/news/datablog/2014/may/12/lies-election-leaflets-five-tricks-european-elections



https://www.theguardian.com/news/datablog/2014/may/12/lies-election-leaflets-five-tricks-european-elections

https://www.theguardian.com/news/datablog/2014/may/12/lies-election-leaflets-five-tricks-european-elections



https://www.theguardian.com/news/datablog/2014/may/12/lies-election-leaflets-five-tricks-european-elections

LIES, DAMNED LIES AND SIGNIFICANCE

?

Significant = Important

Statistically Significant Results

Statistically Significant Results

- We are quite sure they are reliable (not by chance)
- Maybe they’re not “big”
- Maybe they’re not important
- Maybe they’re not useful for decision making

p-values

WIKIPEDIA
The Free Encyclopedia

Misunderstandings of *p*-values

From Wikipedia, the free encyclopedia

Misunderstandings of *p*-values are an important problem in scientific research and scientific education. *p*-value to a significance level will yield one of two results: either the null hypothesis is rejected (which how hypothesis is *true*). From a Fisherian statistical testing approach to statistical inferences, a low *p*-value me

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

https://en.wikipedia.org/wiki/Misunderstandings_of_p-values

p-values

- Probability of observing our results (or more extreme) when the null hypothesis is true

- Probability, not certainty

- Often $p < 0.05$ (arbitrary)

- Can we afford to be fooled by randomness every 1 time out of 20?

Data dredging

Data dredging

- a.k.a. Data fishing or p-hacking
- Convention: formulate hypothesis, collect data, prove/disprove hypothesis
- Data dredging: look for patterns until something statistically significant comes up
- Looking for patterns is ok
  Testing the hypothesis on the same data set is not

SUMMARY

“Everybody lies”

— Dr. House

- Good Science ™ vs. Big headlines
- Nobody is immune
- Ask questions: What is the context? Who's paying? What's missing?
- … "so what?"

THANK YOU

@MarcoBonzanini

speakerdeck.com/marcobonzanini

GitHub.com/bonzanini

marcobonzanini.com

PyData London

NUMFOCUS

OPEN CODE = BETTER SCIENCE