Lecture 2:

- Central Limit Theorem
- Properties of Normal Distributions
- Trials and Tribulations!
- Regression to the Mean
- Correlations

As a consequence of the Central Limit Theorem, many (but not all!) physical processes often tend towards the Normal Distribution shape.

However, few achieve this <u>exactly</u> !!

Although calculated probabilities are often couched in terms of an ideal Normal Distribution to give a rough intuition of the scaling

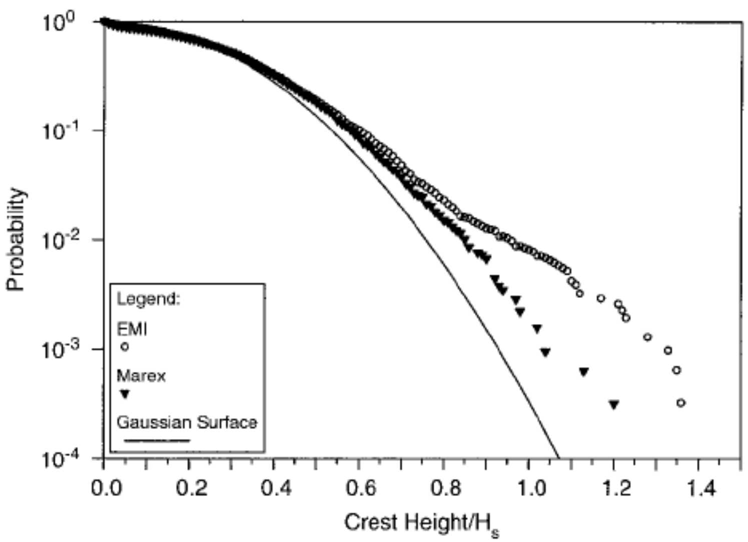
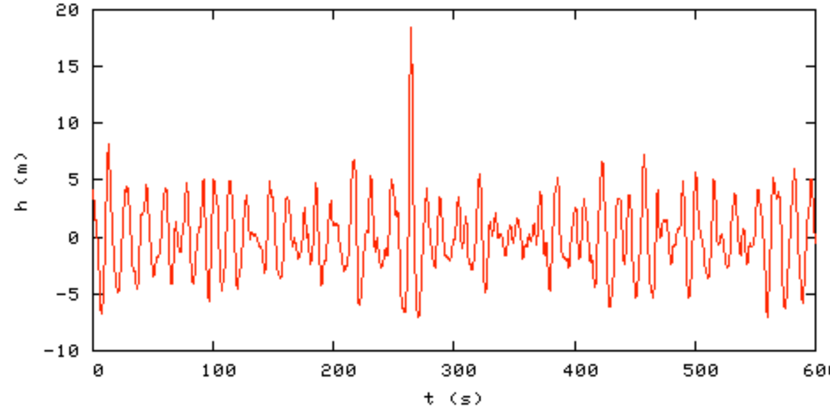
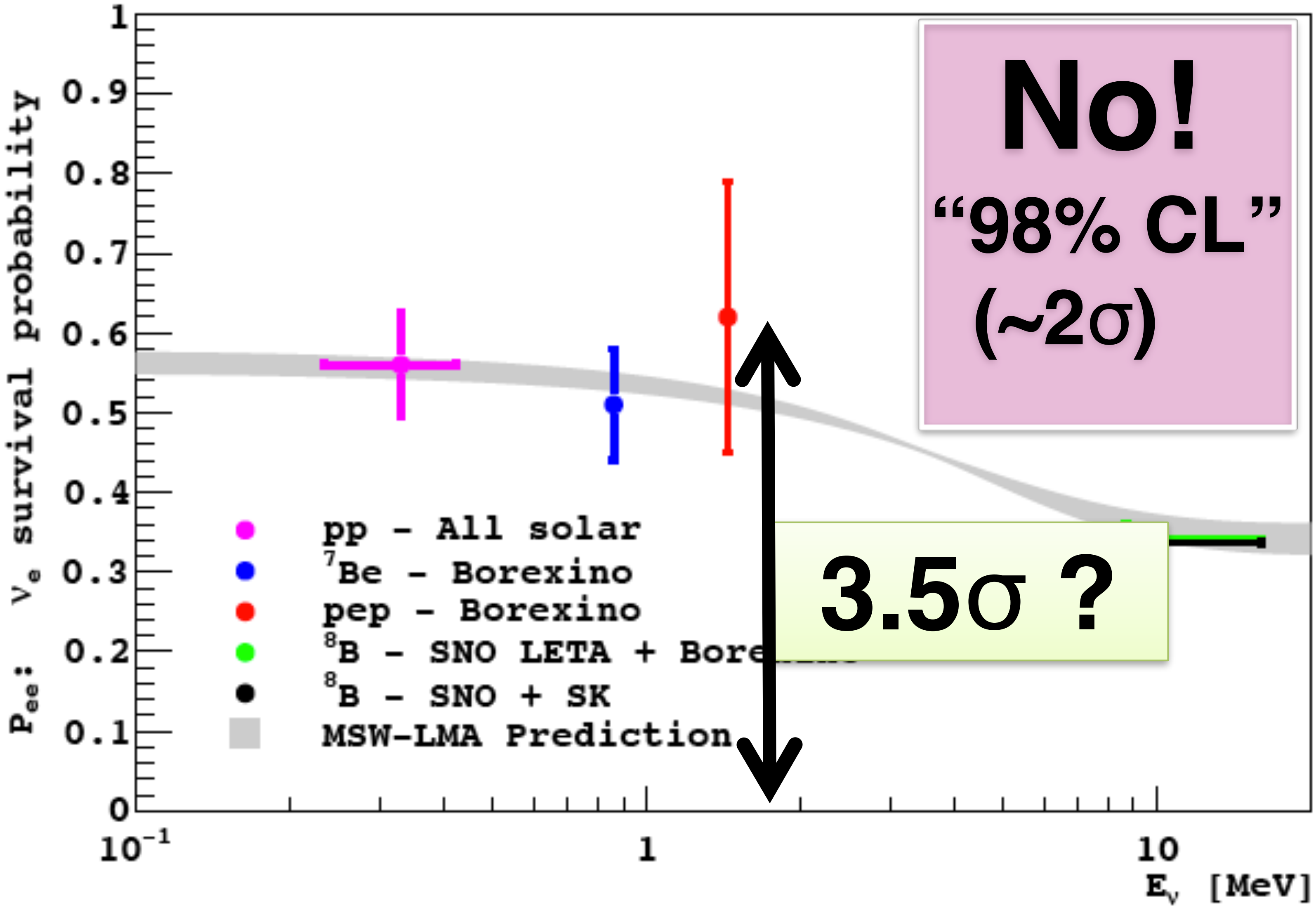


FIG. 4. Probability distribution of normalized crest heights measured at Tern during the storm on 4 Jan 1993. The crest heights are normalized by the significant wave height during each hour of the measurements. Nine hours of measurements with an average significant wave height of about 12 m were combined to produce the observed distribution.



The Draupner wave, a single giant wave measured on New Year's Day 1995, finally confirmed the existence of freak waves, which had previously been considered near-mythical

Borexino "1 σ error" on solar pep flux (2011)



Example: Search for Episodic X-Ray Emission

Over the course of a year, 36000 x-rays are observed to come from a particular astrophysical object. However, on one particular day, 130 events are observed. What is the statistical significance of this observed burst?

$$\langle x \rangle = \frac{36000}{365} = 98.6 \qquad \mu \simeq \langle x \rangle \qquad \sigma = \sqrt{\mu}$$

$$s \cong \frac{(130 - 98.6)}{\sqrt{98.6}} = 3.16\sigma$$

odds of getting at least this many events by a chance fluctuation from the average rate of emission

$$P = 8 \times 10^{-4}$$

Is this sufficient to claim the observation of a burst from this object?

Correct question:

What is the chance of seeing at least one burst with an excess at least as large given the number of independent tests I've done ?

Binomial !!

N Bernoulli trials where the chance of each success is **P**

$$\sum_{i=1}^{\infty} \binom{N}{i} P^i (1-P)^{N-i} = 1 - \binom{N}{0} P^0 (1-P)^{N-0}$$

$$\mathrm{P}_{\text{post-trial}} = 1 - (1-P)^N \qquad (\sim NP \quad \text{for} \quad NP << 1)$$

P = 8 x 10^{-4}, N = 365 → $\mathrm{P}_{\text{post-trial}} = 25\%$

How many timescales were considered? How many objects examined?

An appreciation of trials factors (“look elsewhere effect”) is hugely important... an improper handling of this can lead to incorrect conclusions and opens the door to biased analyses!

This is not trivial ! **A full accounting for this can be tricky:**

- How many hypotheses have you actually tested?
- How many different ways have you tested each hypothesis?
- How many other things would have caught your eye?
- In general, how many ways have you looked at the data?

At the same time, the data needs to be thoroughly checked to look for possible problems and confirm how well it’s understood

This is why physicists set the bar high in terms of significance level in order to claim a discovery

But it’s easy to get carried away...

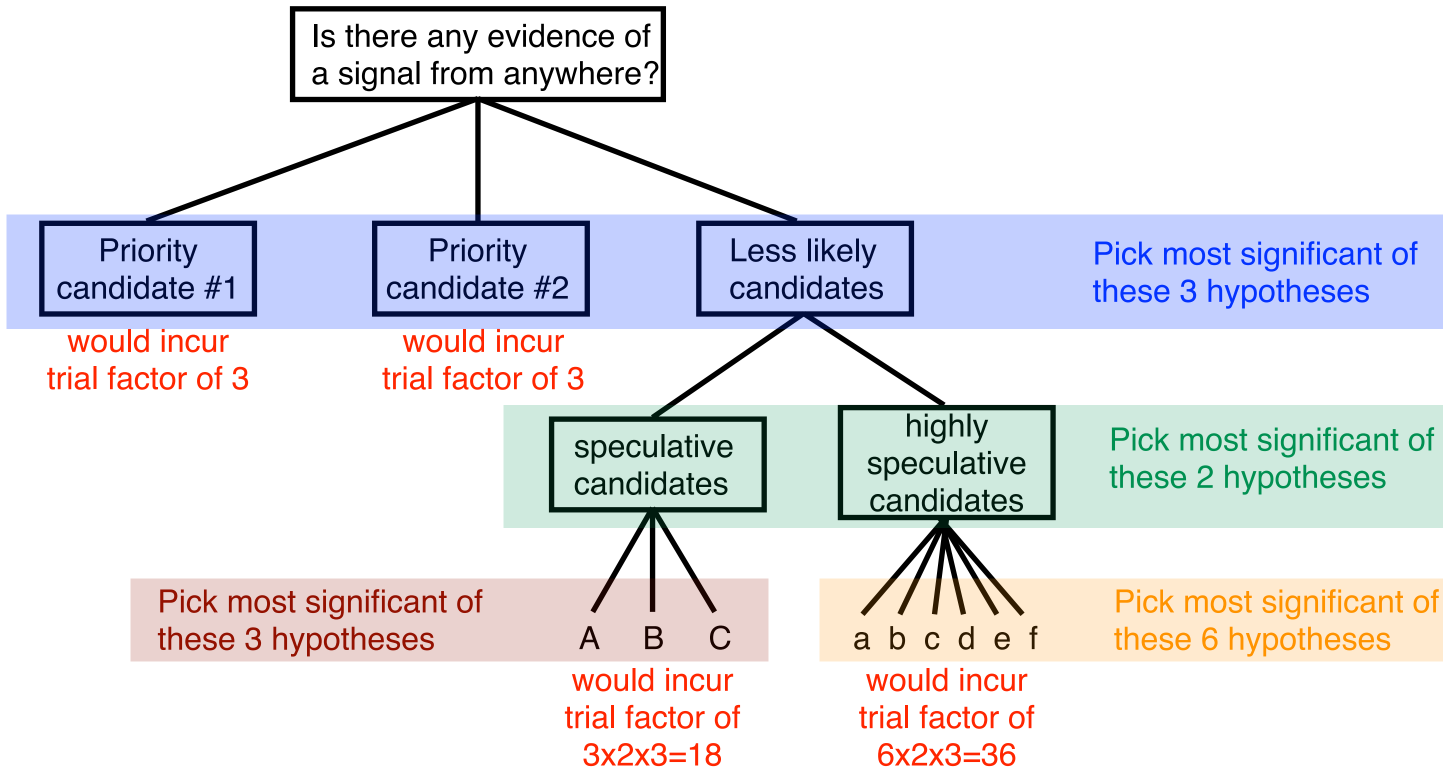
KEEP
CALM
AND
Look
Elsewhere

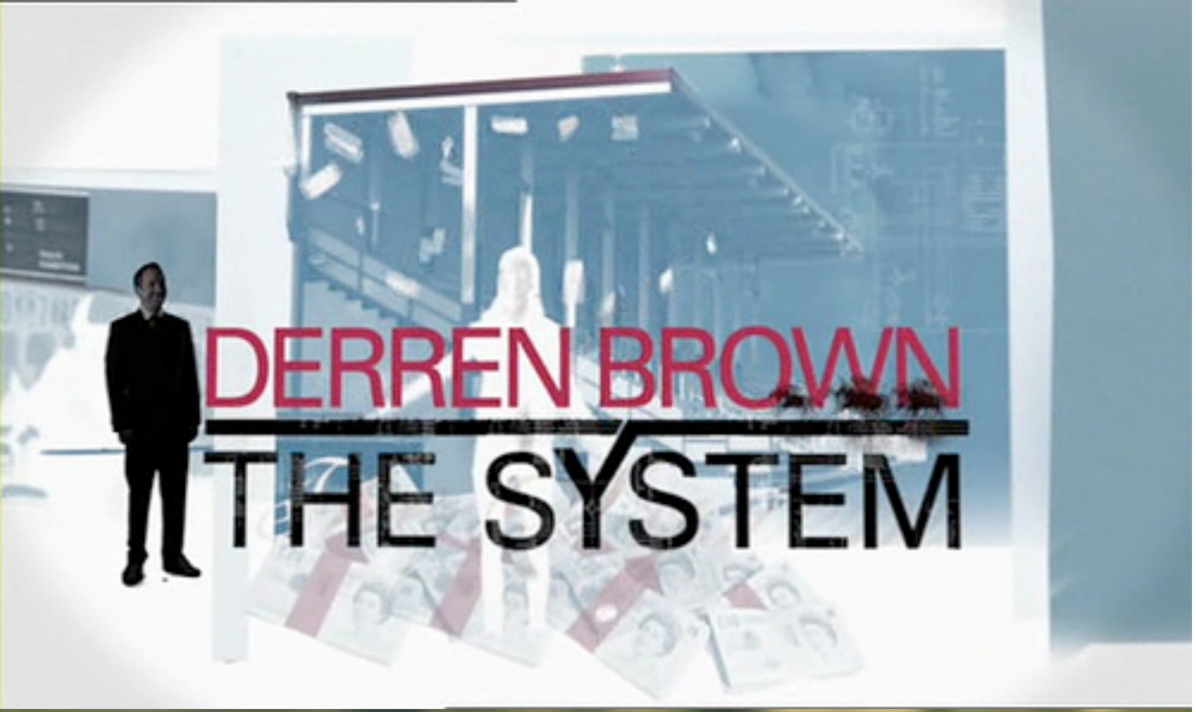
1) Trials factors apply to observations that would potentially lead to making a meaningful claim.

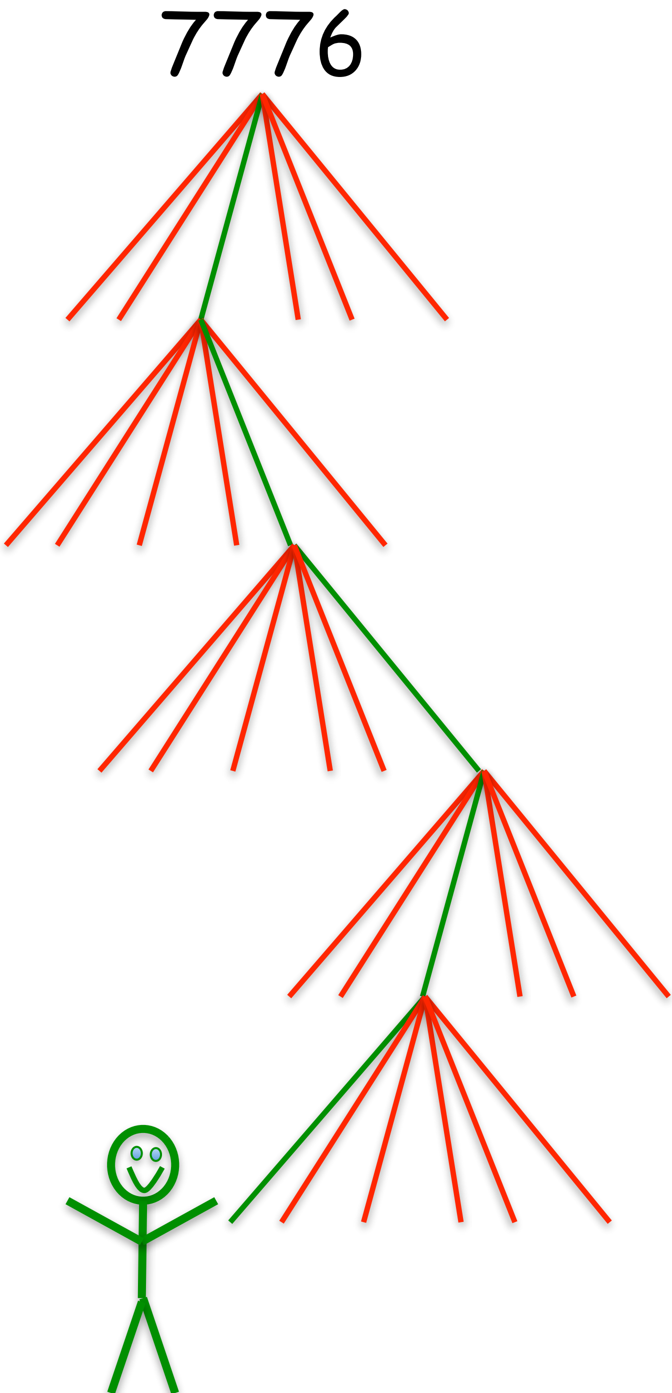
2) Verification based on applying the same analysis to an independent set of data is a good way to avoid misinterpretation of statistical fluctuations.

How do you deal with trial factors in the context of an open-ended search when an independent data set may not be available?

It's possible to structure trial factors based on an a priori ranking of hypothesis plausibility:



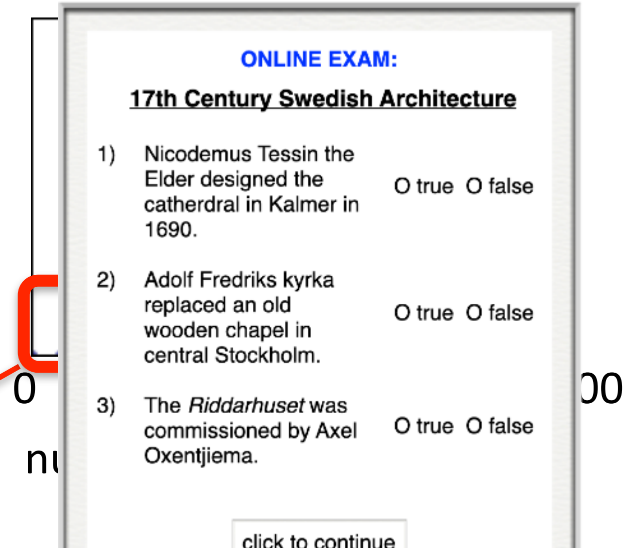

DERREN BROWN
THE SYSTEM


7776

“Regression to the Mean”

Pop Quiz:

100 true/false questions on details of 17th century Swedish architecture.



ONLINE EXAM:

17th Century Swedish Architecture

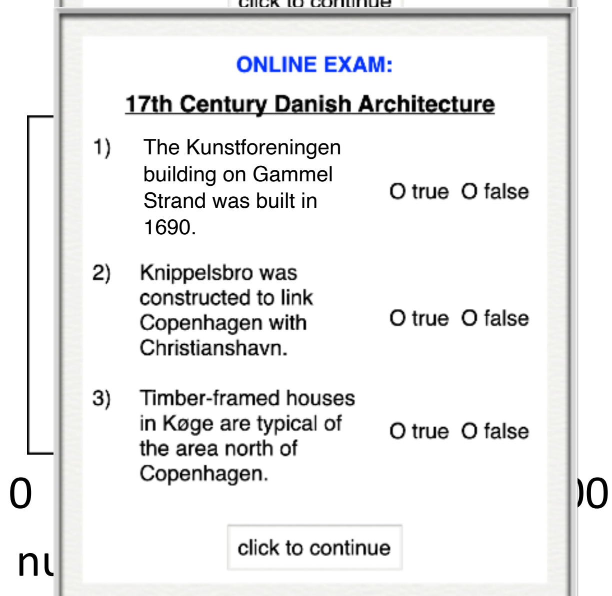
1) Nicodemus Tessin the Elder designed the catherdral in Kalmer in 1690. O true O false

2) Adolf Fredriks kyrka replaced an old wooden chapel in central Stockholm. O true O false

3) The *Riddarhuset* was commissioned by Axel Oxentjiema. O true O false

click to continue

100 true/false questions on details of 17th century Danish architecture.



ONLINE EXAM:

17th Century Danish Architecture

1) The Kunstforeningen building on Gammel Strand was built in 1690. O true O false

2) Knippelsbro was constructed to link Copenhagen with Christianshavn. O true O false

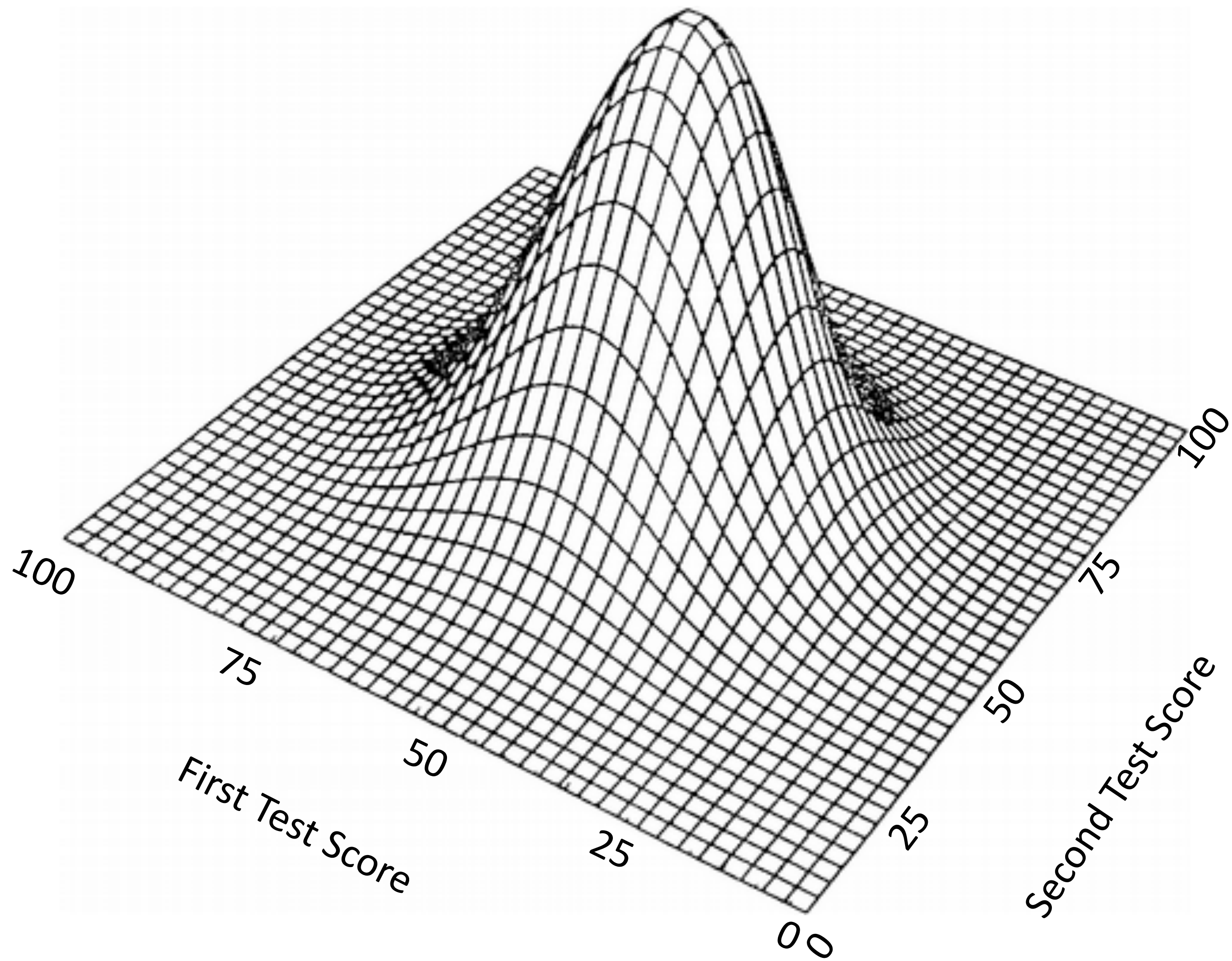
3) Timber-framed houses in Køge are typical of the area north of Copenhagen. O true O false

click to continue

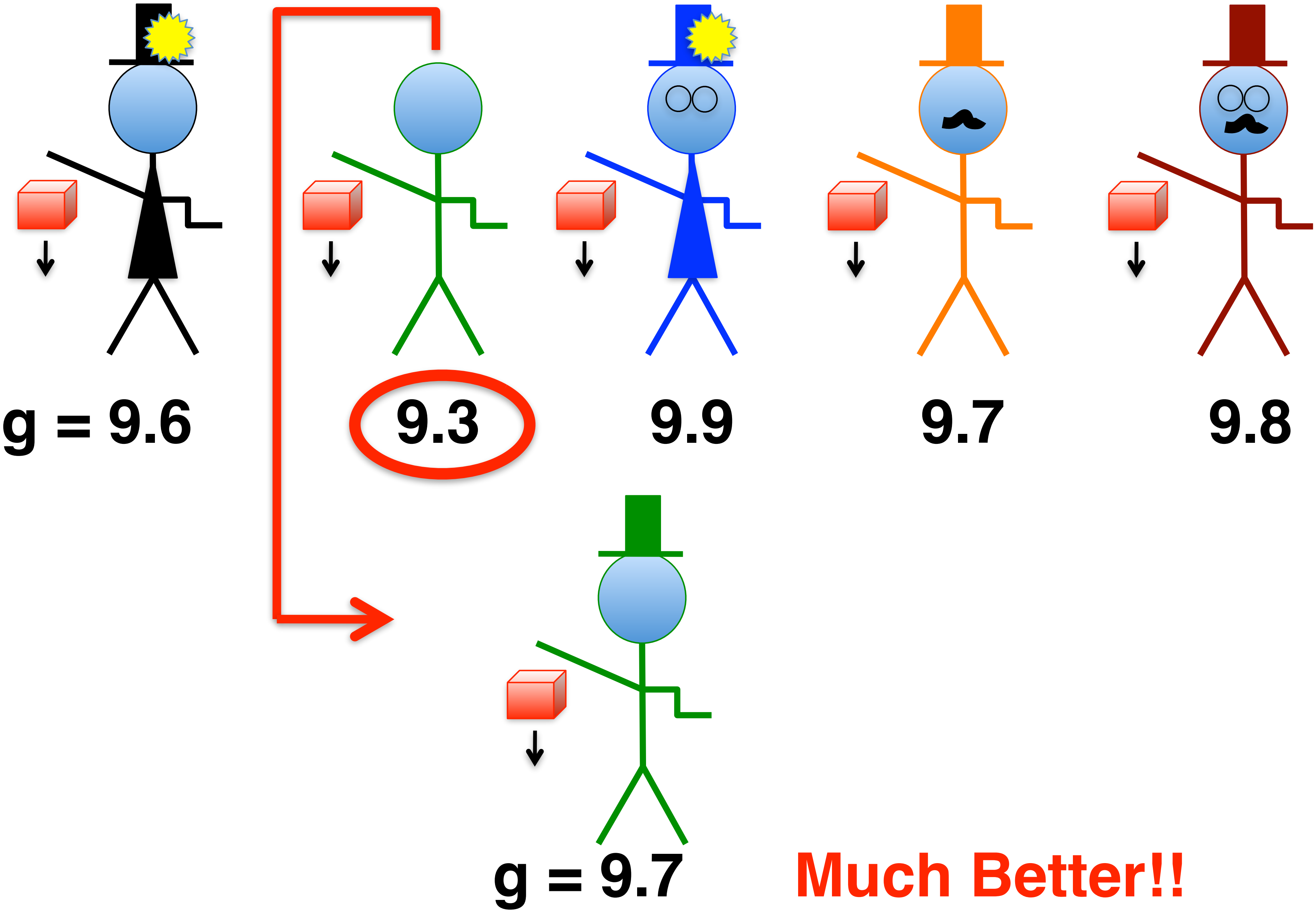
What an improvement! This particular group of students must know much more about Danish architecture!!

Bi-variate Distribution with Identical Marginal Distributions

(i.e. uncorrelated)

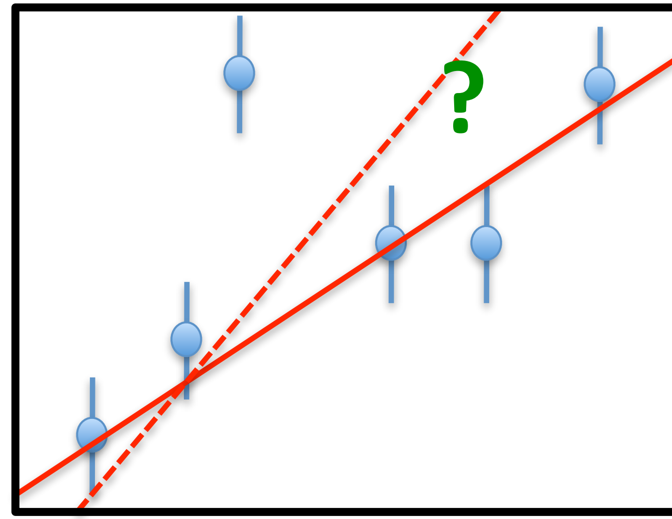


“The Effect of Hats on the Measurement of Gravity”



So How Do You Handle Outliers?

No clear rules!



Rules of Thumb:

- Look for possible systematic biases in the data;
- However, only reject outliers based on clear statistical/scientific criteria;
- Explicitly point out the issue and discuss the details;
- Be aware of any potential bias that could result and review the robustness of your final conclusions.

The total number of known species is ~1.5 million

The number of known species that can fly is ~500,000

$P(\text{flying}) > 5\text{x}10^5/1.5\text{x}10^6 = 0.33$

The number of plant species ~400,000

$P(\text{plant}) = 4\text{x}10^5/1.5\text{x}10^6 = 0.27$

Thus, probability of finding a flying plant is

$P(\text{plant}) \text{ x } P(\text{flying}) = 0.089$

And the expected number of flying plant species is

$(0.089)(1.5\text{x}10^6) = 133{,}500$

Correlations

	Insects	Birds	Mammals	Plants	Reptiles
Flying	500k **0.35**	10k **0.007**	400 **2.8×10^{-4}**	**0**	**0**
Non-Flying	500k **0.35**	54 **3.8×10^{-5}**	6k **0.004**	400k **0.28**	10k **0.007**

Joint PDF

Insects	Birds	Mammals	Plants	Reptiles
500k **0.54**	54 **5.9×10^{-5}**	6k **0.006**	400k **0.44**	10k **0.011**

PDF for Non-Flying Species

Insects	Birds	Mammals	Plants	Reptiles
500k **0.70**	54 **0.00704**	6k **0.00428**	400k **0.28**	10k **0.007**

“Marginalised” PDF for All Species

Just to be clear:

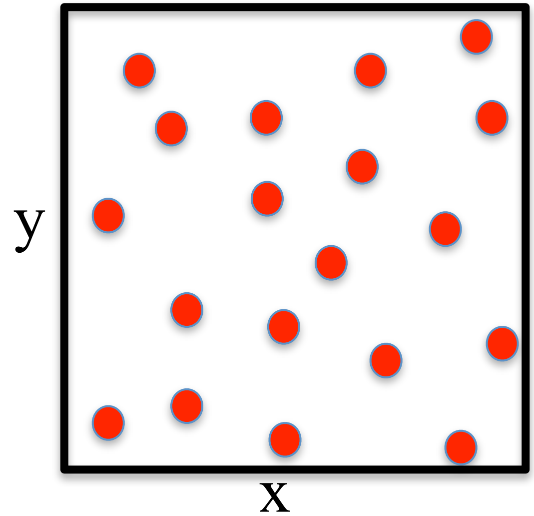
For example, if we have 2 dependent variables, x & y:

$$\int P(x,y)dxdy = 1$$

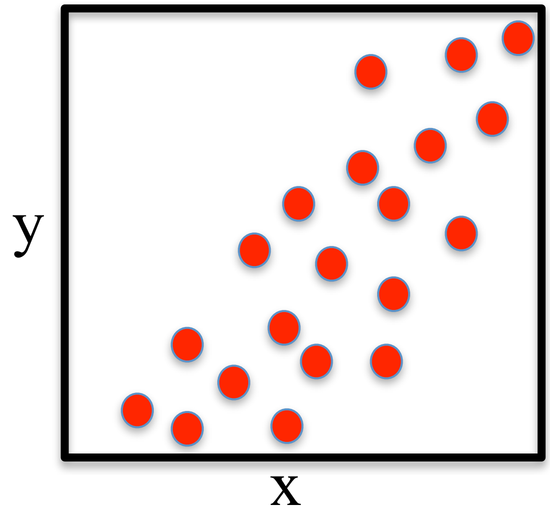
and

$$\langle f(x,y)\rangle = \int f(x,y)P(x,y)dxdy$$

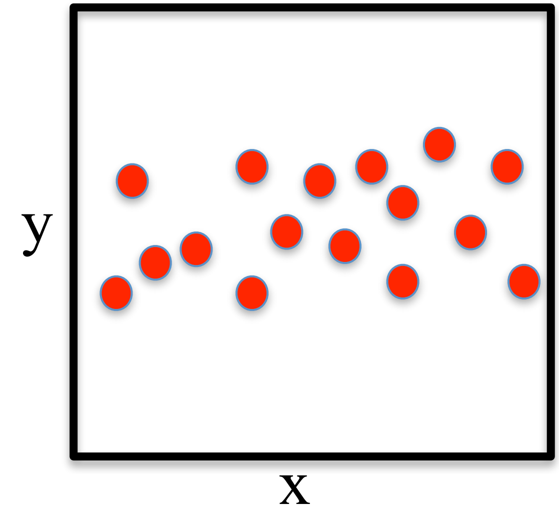
Correlated or Uncorrelated?



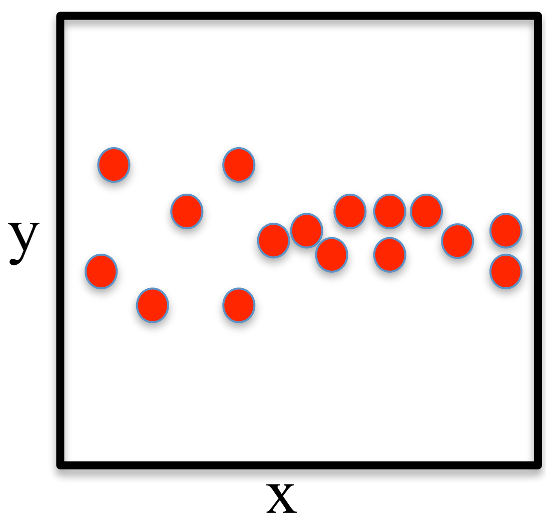
relatively uncorrelated



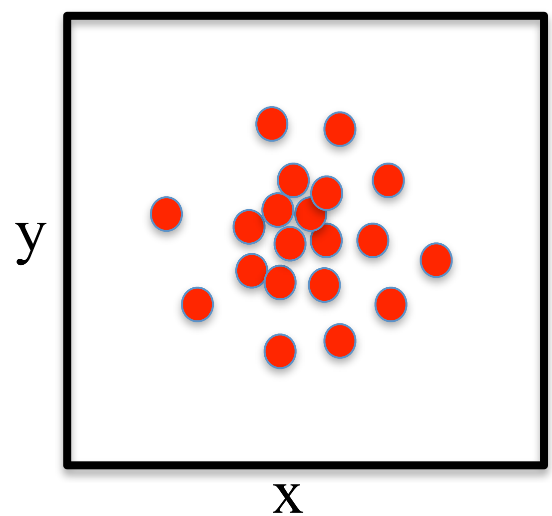
correlated



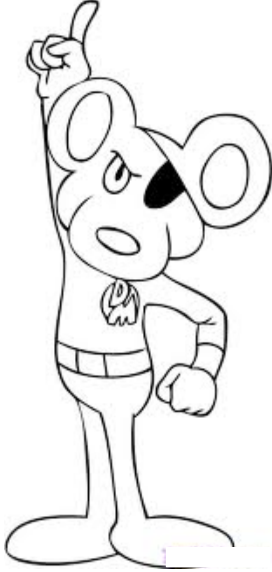
relatively uncorrelated



relatively uncorrelated mean,
but correlated variance



relatively uncorrelated
(symmetric with similar marginal distributions)



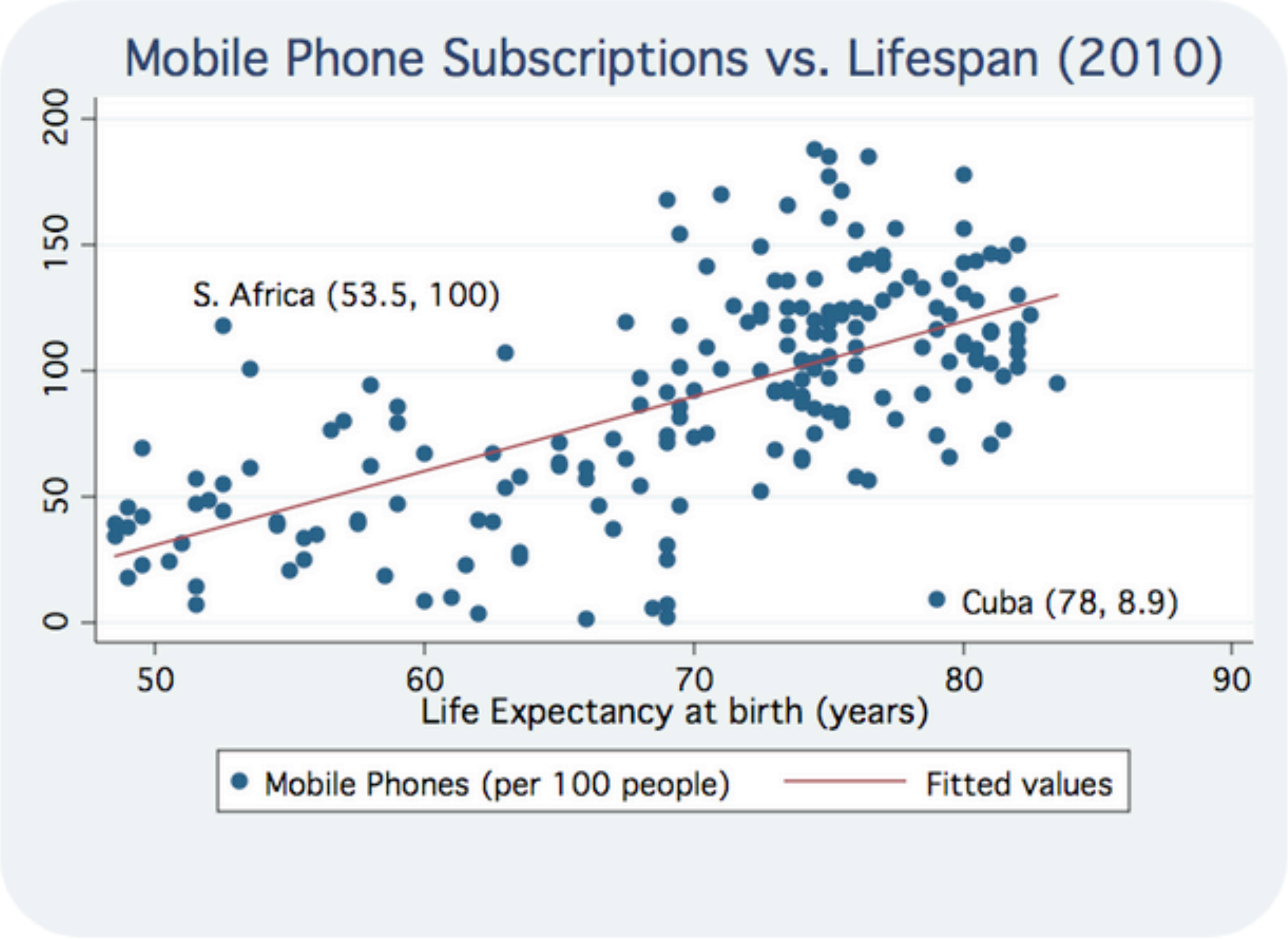
Beware of “hidden” correlations between **ANY** parameters that distinguish elements of your data set

IQ ⟷ Shoe Size

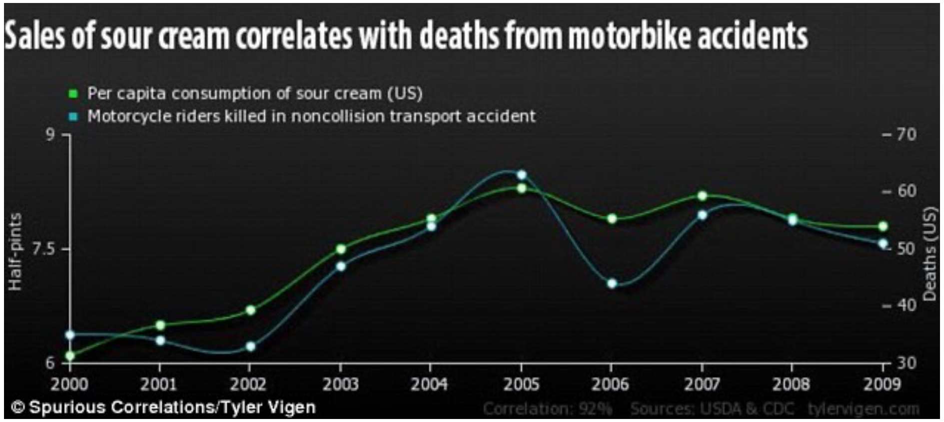
⇕

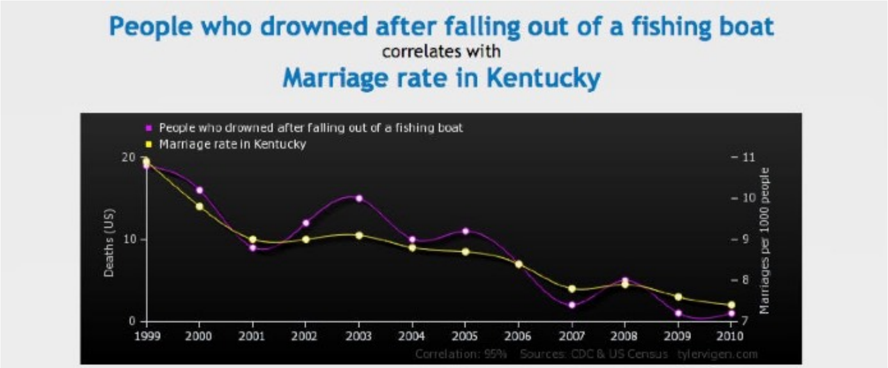
IQ ⇔ Age

Beware of jumping to conclusion about cause and effect



Beware of spurious correlations





	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
People who drowned after falling out of a fishing boat Deaths (US) (CDC)	19	16	9	12	15	10	11	7	2	5	1	1
Marriage rate in Kentucky Marriages per 1000 people (US Census)	10.9	9.8	9	9	9.1	8.8	8.7	8.4	7.8	7.9	7.6	7.4

Correlation: 0.952407