5 Statistical Mistakes and How to Avoid Them

Linda Lapp
SICSA PhD Conference 2018

About me

- BSc Hons in Maths and Stats at University of Strathclyde (2012-2016)
- MPhil in Healthcare Analytics (Predictive Modelling) (2016-2017)
- PhD Student in Healthcare Analytics (2017-…)
  - Patients undergoing heart surgery
  - Working with the NHS
  - Statistical and Machine Learning methods
  - Predictive Modelling
  - Tutor in Maths and Stats

- Royal Statistical Society Fellow

Agenda

Data

Mean vs Median

Outliers

Correlation vs Causation

P-Values

Media Friendly Research Titles

1

Missing Data

Missing Data

Name	Age	BMI	Smoking Status
Marge Simpson	55	32.45	No
Montgomery Burns	82	23.02	Ex-Smoker
Homer Simpson	58		Ex-Smoker
Seymour Skinner	64	19.92	
Edna Krabappel	61	26.45	Yes

Missing Data

Name	Age	BMI	Smoking Status
Marge Simpson	55	32.45	No
Montgomery Burns	82	23.02	Ex-Smoker
Homer Simpson	58	?	Ex-Smoker
Seymour Skinner	64	19.92	?
Edna Krabappel	61	26.45	Yes

Missing Data

Name	Age	BMI	Smoking Status
Marge Simpson	55	32.45	No
Montgomery Burns	82	23.02	Ex-Smoker
Homer Simpson	58	?	Ex-Smoker
Seymour Skinner	64	19.92	?
Edna Krabappel	61	26.45	Yes

Numerical

Categorical

Missing Data

Name	Age	BMI	Smoking Status
Marge Simpson	55	32.45	No
Montgomery Burns	82	23.02	Ex-Smoker
~~Homer Simpson~~	~~58~~		~~Ex-Smoker~~
Seymour Skinner	64	19.92	?
Edna Krabappel	61	26.45	Yes

Missing Data

Name	Age	~~BMI~~	Smoking Status
Marge Simpson	55	~~32.45~~	No
Montgomery Burns	82	~~23.02~~	Ex-Smoker
Homer Simpson	58	~~?~~	Ex-Smoker
Seymour Skinner	64	~~19.92~~	?
Edna Krabappel	61	~~26.45~~	Yes

Missing Data

Name	Age	BMI	Smoking Status
Marge Simpson	55	32.45	No
Montgomery Burns	82	23.02	Ex-Smoker
Homer Simpson	58	Mean BMI: 25.46 Median BMI: 21.47	Ex-Smoker
Seymour Skinner	64	19.92	?
Edna Krabappel	61	26.45	Yes

Missing Data

Name	Age	BMI	Smoking Status
Marge Simpson	55	32.45	No
Montgomery Burns	82	23.02	Ex-Smoker
Homer Simpson	58	?	Ex-Smoker
Seymour Skinner	64	19.92	?
Edna Krabappel	61	26.45	Yes

Numerical

Categorical

Missing Data

Name	Age	BMI	~~Smoking Status~~
Marge Simpson	55	32.45	~~No~~
Montgomery Burns	82	23.02	~~Ex-Smoker~~
Homer Simpson	58	?	~~Ex-Smoker~~
Seymour Skinner	64	19.92	~~?~~
Edna Krabappel	61	26.45	~~Yes~~

Missing Data

Name	Age	BMI	Smoking Status
Marge Simpson	55	32.45	No
Montgomery Burns	82	23.02	Ex-Smoker
Homer Simpson	58	?	Ex-Smoker
~~Seymour Skinner~~	~~64~~	~~19.92~~	
Edna Krabappel	61	26.45	Yes

Missing Data

Name	Age	BMI	Smoking Status
Marge Simpson	55	32.45	No
Montgomery Burns	82	23.02	Ex-Smoker
Homer Simpson	58	?	Ex-Smoker
Seymour Skinner	64	19.92	Unknown
Edna Krabappel	61	26.45	Yes

Missing Data

- How much data?
- Can you afford losing entries due to missing data?
- How important is the variable?
- How many missing values do you have?
- How is your population distributed?
  - Could be a strongly skewed distribution
  - -> Can’t use mean BMI in this case

Missing Data

Name	Age	BMI	Smoking Status
Marge Simpson	55	32.45	No
Montgomery Burns	82	23.02	Ex-Smoker
Homer Simpson	58		Ex-Smoker
Seymour Skinner	64	19.92	
Edna Krabappel	61	26.45	Yes

Missing Data

Name	Age	BMI	Smoking Status
Marge Simpson	55	32.45	No
Montgomery Burns	82	23.02	Ex-Smoker
Homer Simpson	58	?	Ex-Smoker
Seymour Skinner	64	19.92	?
Edna Krabappel	61	26.45	Yes

Missing Data

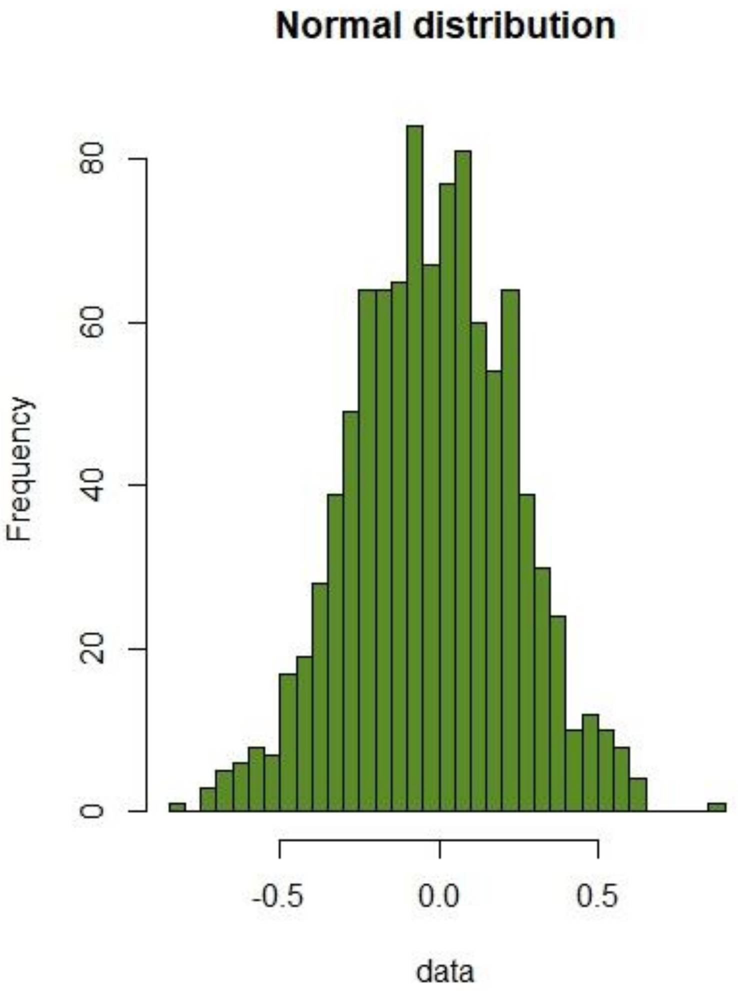
ID	~~Name~~	Age	BMI	Smoking Status
1	~~Marge Simpson~~	55	32.45	No
2	~~Montgomery Burns~~	82	23.02	Ex-Smoker
3	~~Homer Simpson~~	58	?	Ex-Smoker
4	~~Seymour Skinner~~	64	19.92	?
5	~~Edna Krabappel~~	61	26.45	Yes

2

Mean vs Median

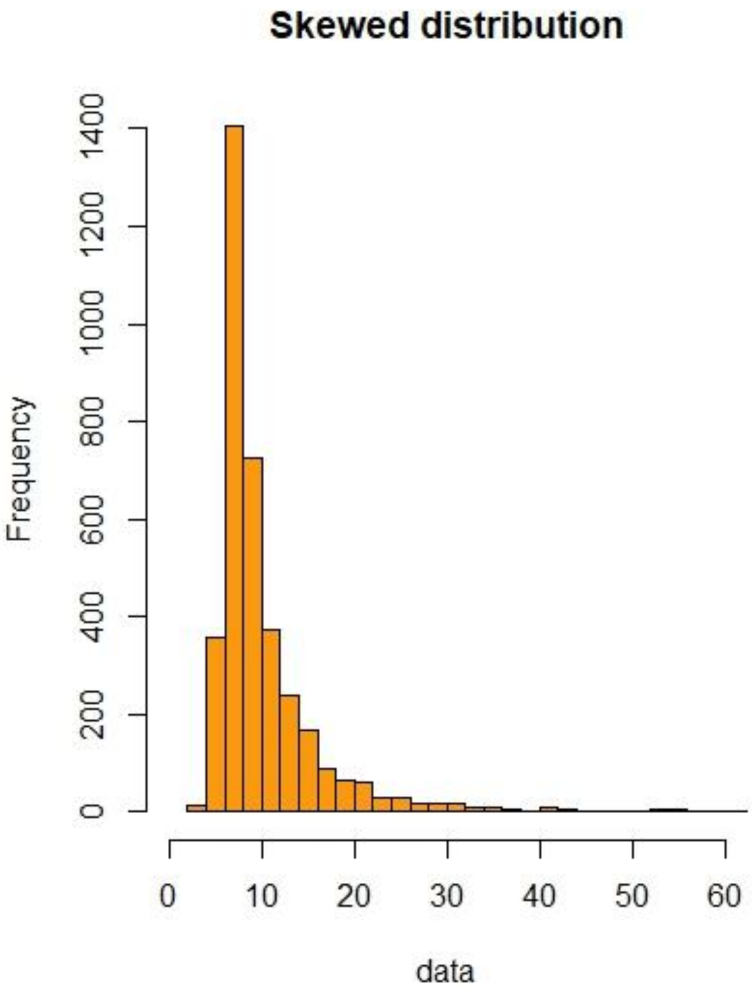
Mean vs Median

- Mean:
- The arithmetic average of a set of numbers or distribution
- Used for normal distributions
- Largely influenced by outliers



Mean vs Median

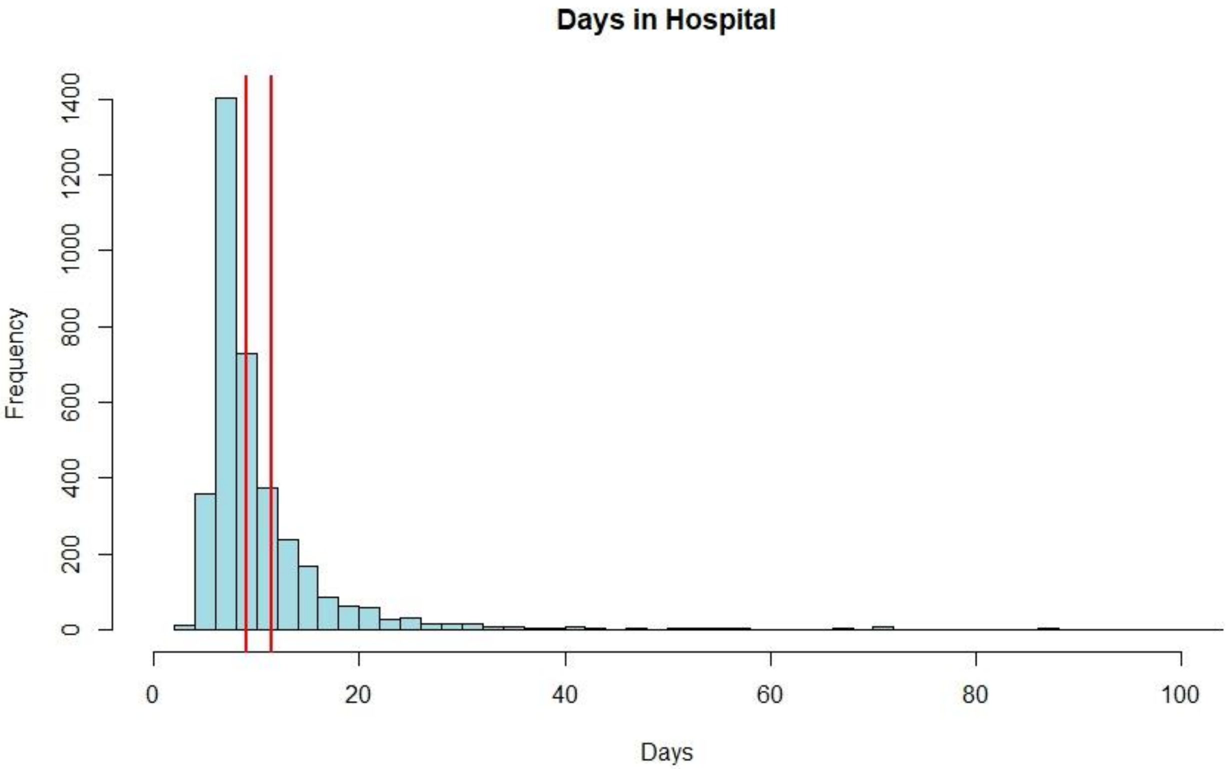
- Median:
- The numeric value separating the higher half of a sample, a population, or a probability distribution, from the lower half
- Used for skewed distributions
- NOTE:
  - Normal dist <- mean≈median



Mean vs Median

- Hospital looks to cover costs of patients staying in hospital after a heart surgery.
- Need to send a report about hospital expenses.
- Based on this report, the government gives funds to the hospital.
- Given the information, which statistics should be used to calculate financial needs?

Mean vs Median



- Days in hospital:
- Min = 2 days
- Max = 183 days
- Mean = 11.42 days
- Median = 9 days

Mean vs Median

- Assuming one night in hospital costs £400
- Hospital carries out 1500 heart surgeries per year

- Mean: Hospital receives £400 x 11.42 x 1500 = *£6,852,000*
- Median: Hospital receives £400 x 9 x 1500 = *£5,400,000*
- That is ~£1.5M difference!

- Which one is correct to report?

3

Outliers – What to do with them?

Outliers – What to do with them?

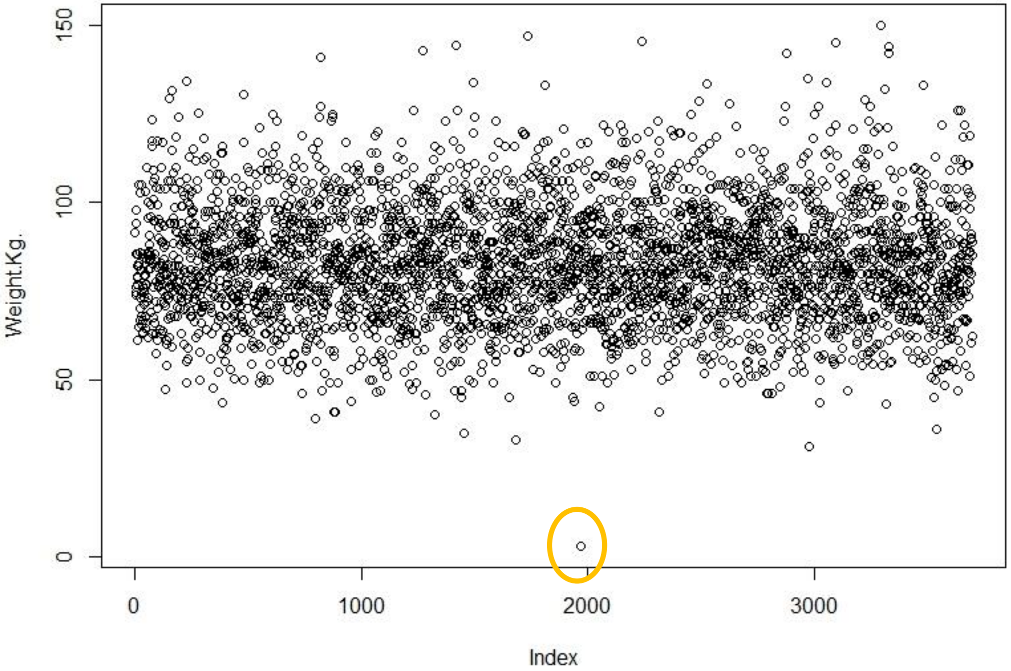
- Means, standard deviations, correlations and every other statistic based on these measures are highly sensitive to outliers.
- It is NOT acceptable to drop an observation just because it is an outlier.
- Outliers can be legitimate observations and are sometimes the most interesting ones.

Outliers – What to do with them?

Case I

- Data: people’s weight reported in a database
- Outlier weight = 3 kg (!)
- Obviously incorrectly entered or measured data

-> Drop the outlier

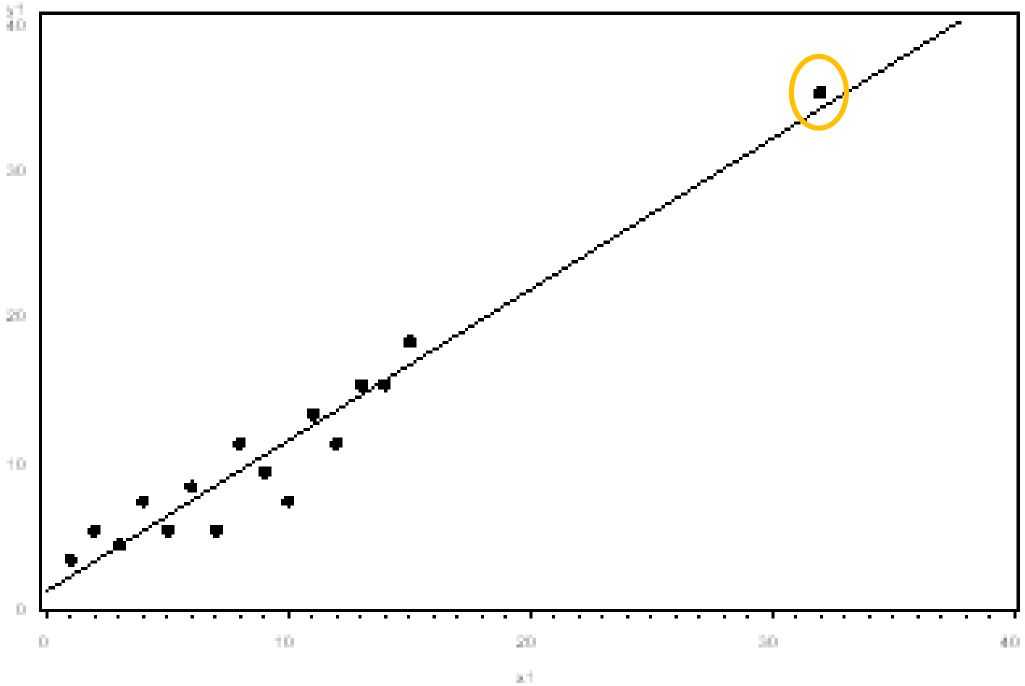


Outliers – What to do with them?

Case II

- Neither presence nor absence of the outlier changes the regression line
- Outlier does not change results, but affects assumptions

-> Drop the outlier, but provide explanation



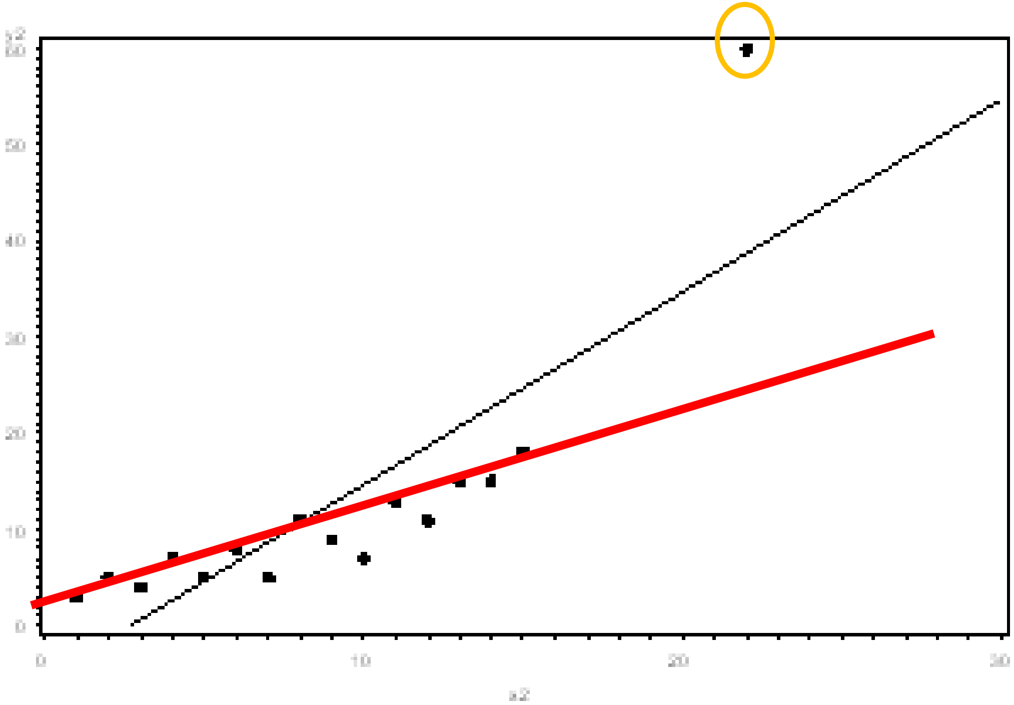
Outliers – What to do with them?

Case III

- Outlier changes results, and affects assumptions

-> Run analysis with outlier and without

-> Discuss how the results changed



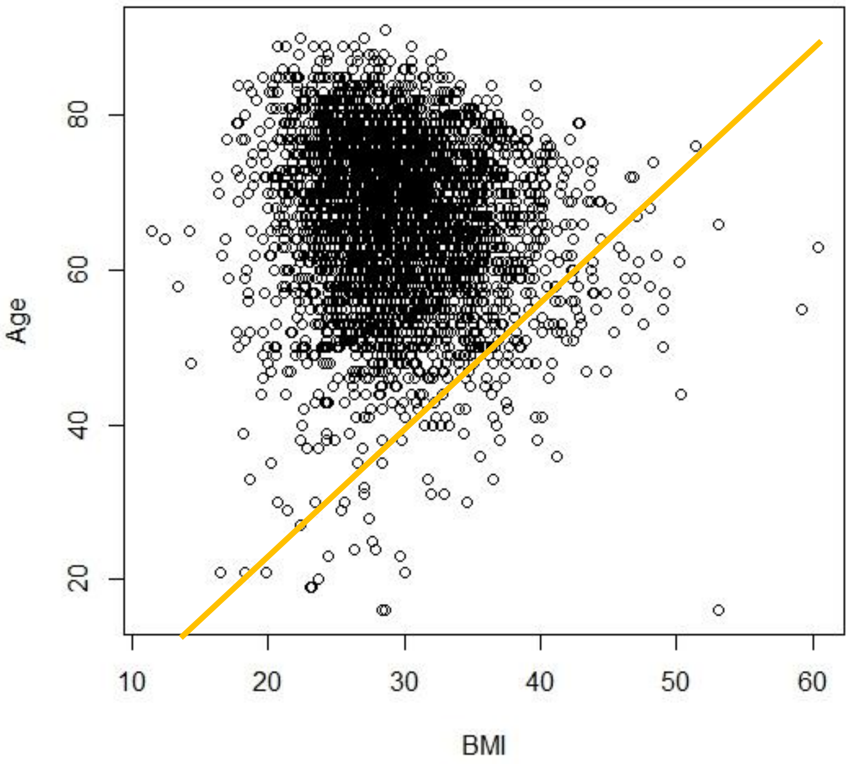
Outliers – What to do with them?

Case IV

- Outlier creates a significant association

-> Drop the outlier

-> Choose a different method of analysis



4

Correlation vs Causation

Correlation vs Causation

Correlation

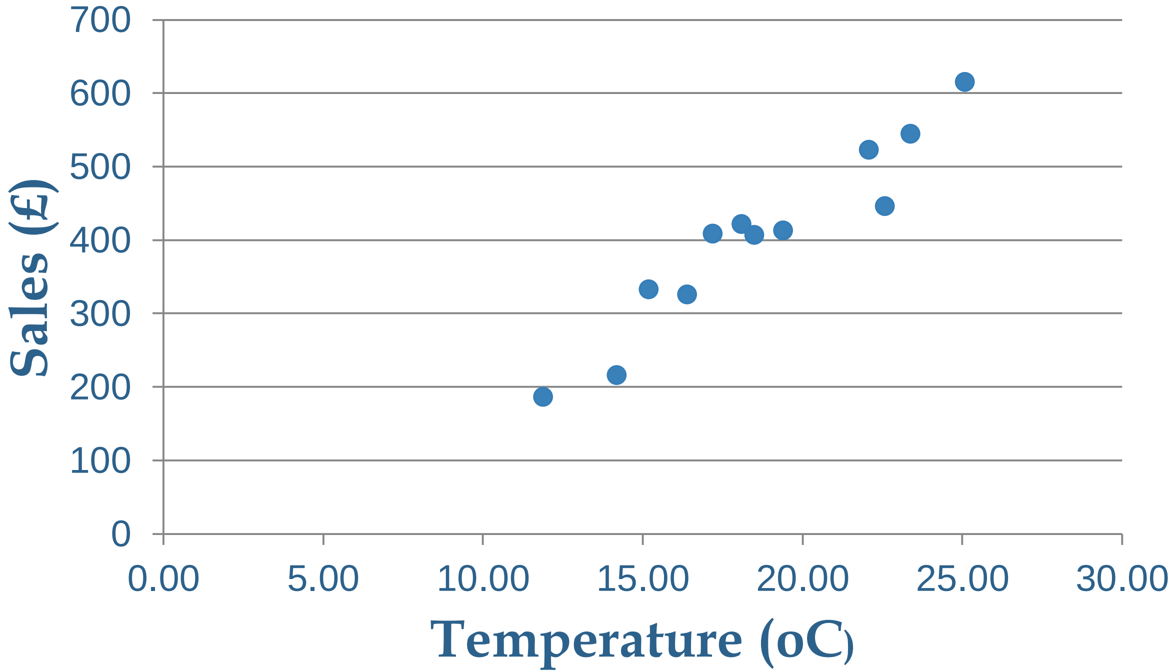
Describes the size and direction of a relationship between two or more variables.

Causation

Indicates that one event is the result of the occurrence of the other event. Shows causal relationship.

Correlation vs Causation

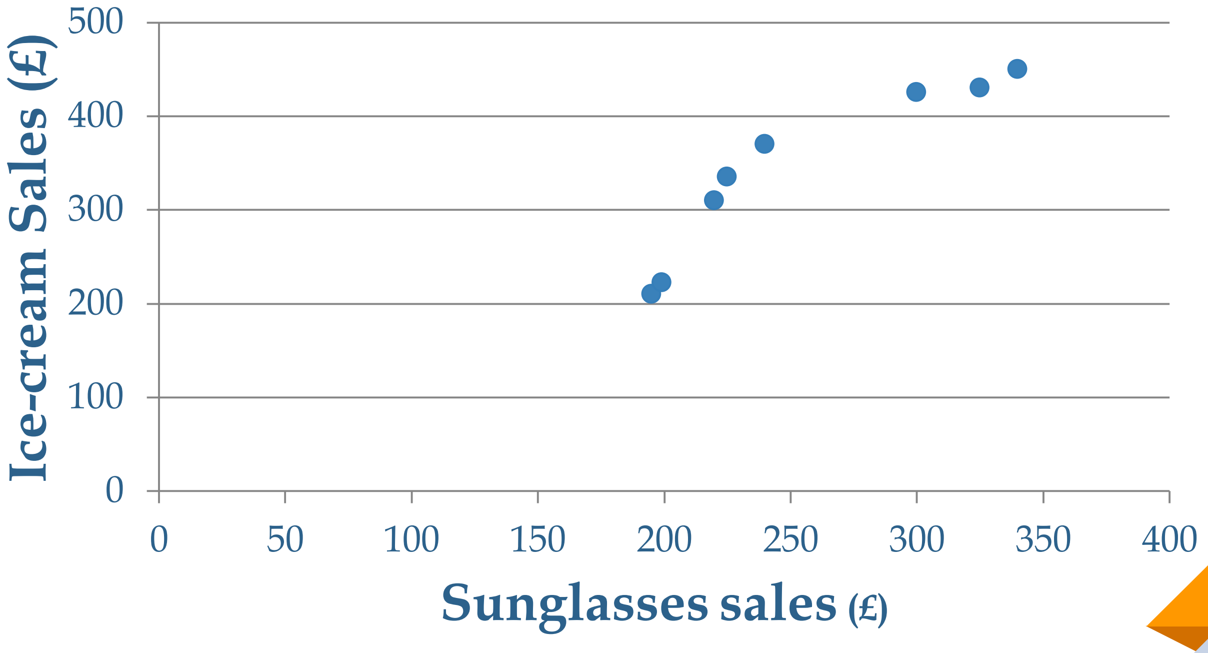
Is there correlation between ice cream sales and temperature?



Correlation vs Causation

Is there correlation between ice cream sales and sunglass sales?

Does buying ice cream cause people buy sunglasses?



Correlation vs Causation

Do storks bring babies?

- Dr Gustav Fischer (1936) looked at records of Copenhagen for 10 years following WW II.
- Positive correlation between
  - X – annual number of storks nesting in city
  - Y – annual number of babies born in city

Does this mean that *storks bring babies*?

Correlation vs Causation

Do storks bring babies?

- Look for a third variable Z influencing both X and Y
- Possible explanation:
- Baby-boom ->

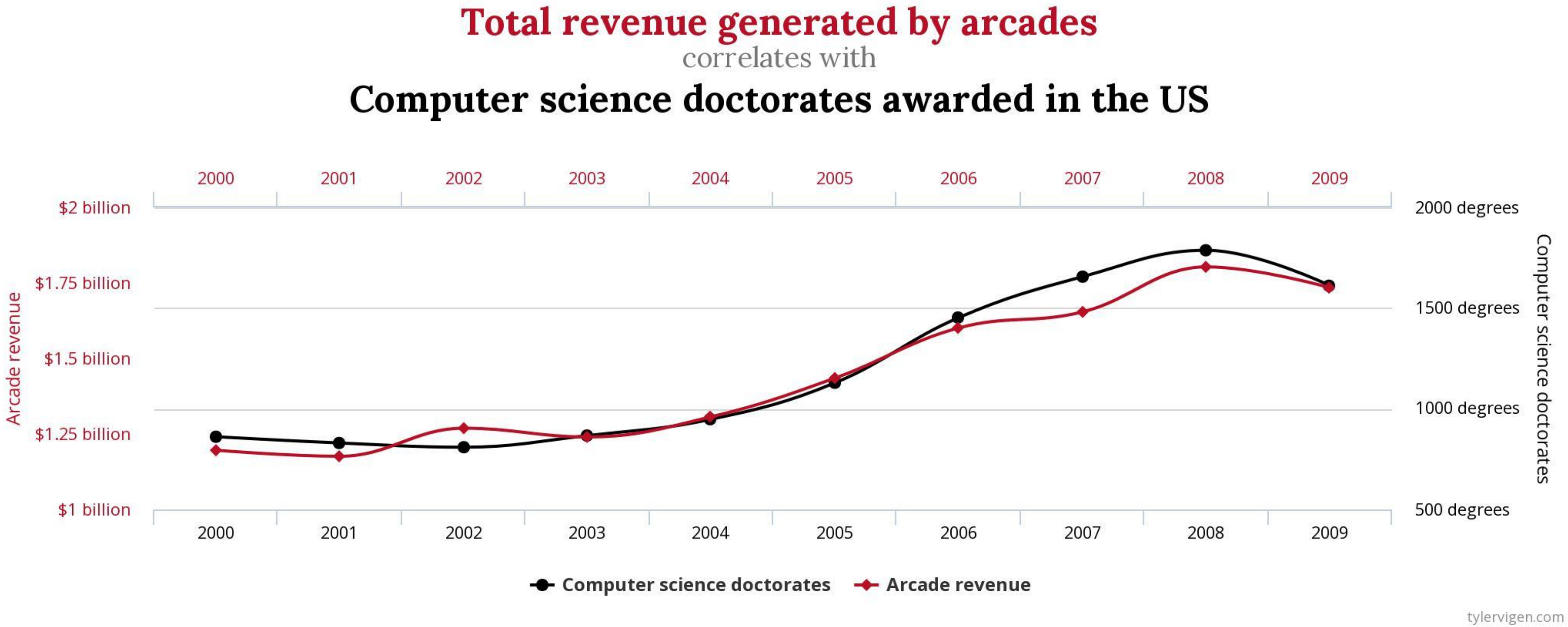
  -> More houses ->

  -> More nesting places for storks ->

  -> More storks!

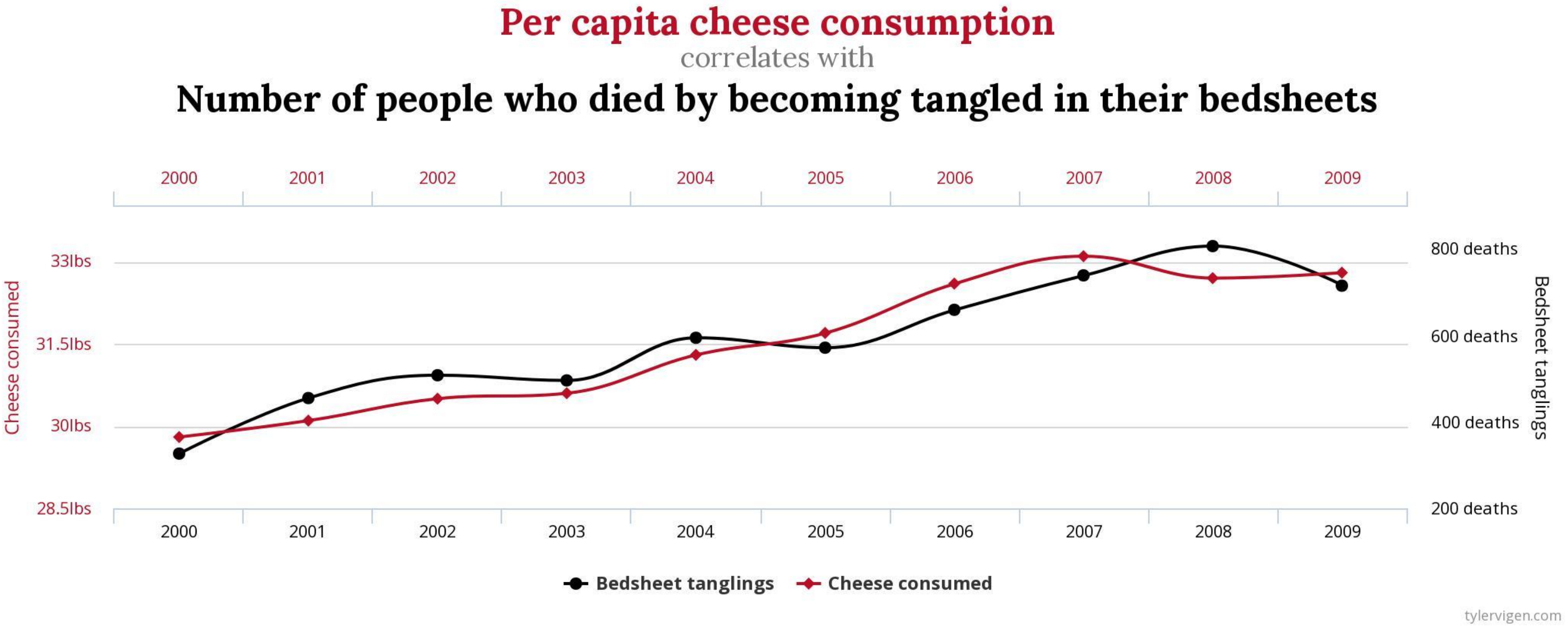
Correlation vs Causation

Spurious correlation – no sensible link between correlated variables



Correlation vs Causation

Spurious correlation – no sensible link between correlated variables



Per capita cheese consumption

correlates with

Number of people who died by becoming tangled in their bedsheets

tylervigen.com

Correlation vs Causation

Spurious correlation – no sensible link between correlated variables

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

5

P-values and significance

What does P-value tell you?

- If the null hypothesis is true, the P-value is the probability of obtaining your sample data.
- Based exclusively on information contained within a sample.

Significance

- Significance means a P-value of less than 0.05 (or 0.01, depending on significance level)

What is P-Hacking?

- Stop collecting data once $p<0.05$
- Analyse many measures, but report only those with $p<0.05$
- Exclude participants to get $p<0.05$
- Transform the data to get $p<0.05$

How to avoid P-hacking?

- Avoid selection or tweaks after seeing the data
- Register your study to show reliability
- Publish negative results
- Use also other measures: odds ratios, absolute risk, relative risk, and confidence intervals

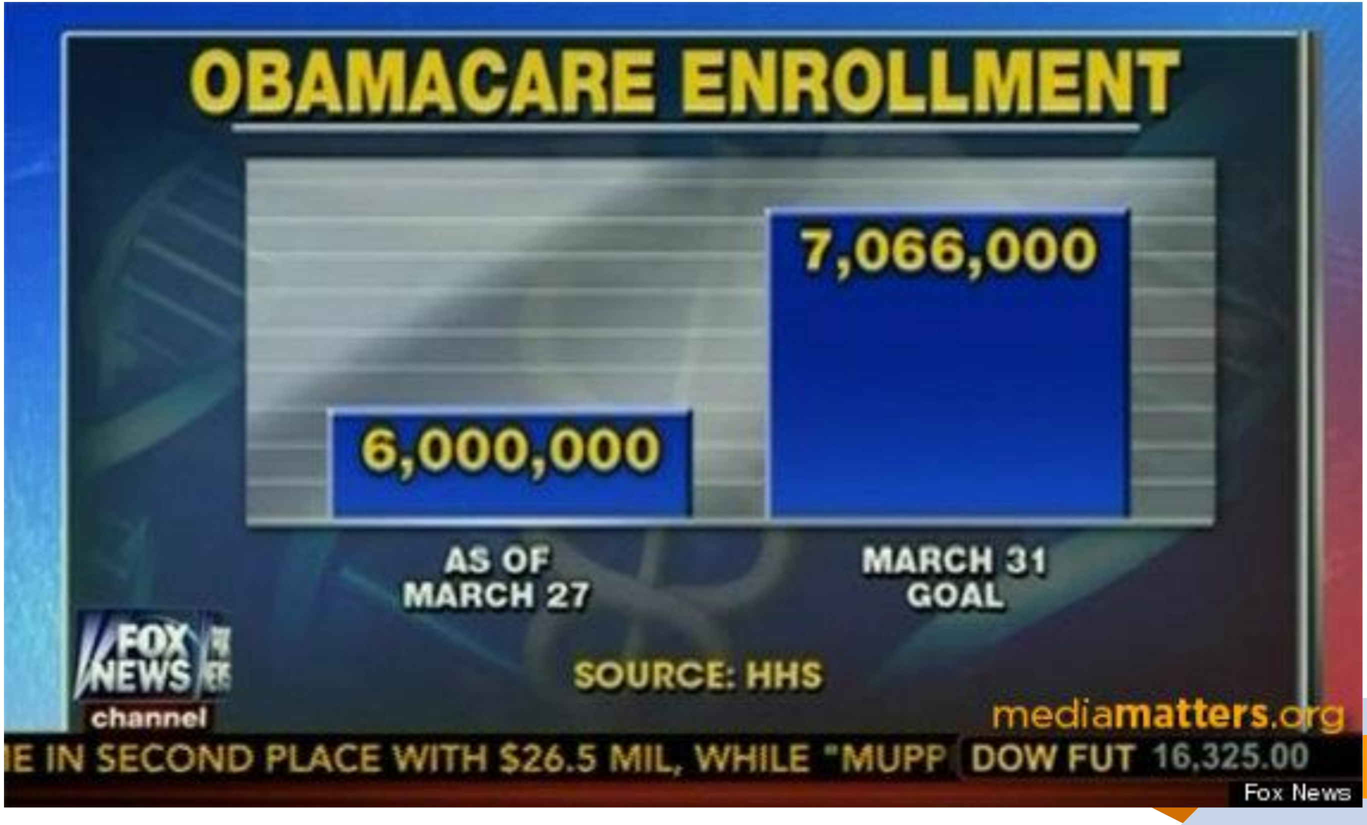
Concluding thoughts

- Think carefully about the missing data and how to approach it.
- Understand your data:
  - How is it distributed?
  - Which variables do you need?
- Do NOT just remove outliers – sometimes these are the most interesting parts!
- Correlation ≠ causation.
- P-value does not necessarily answer your question.
- Always question statistical evidence!

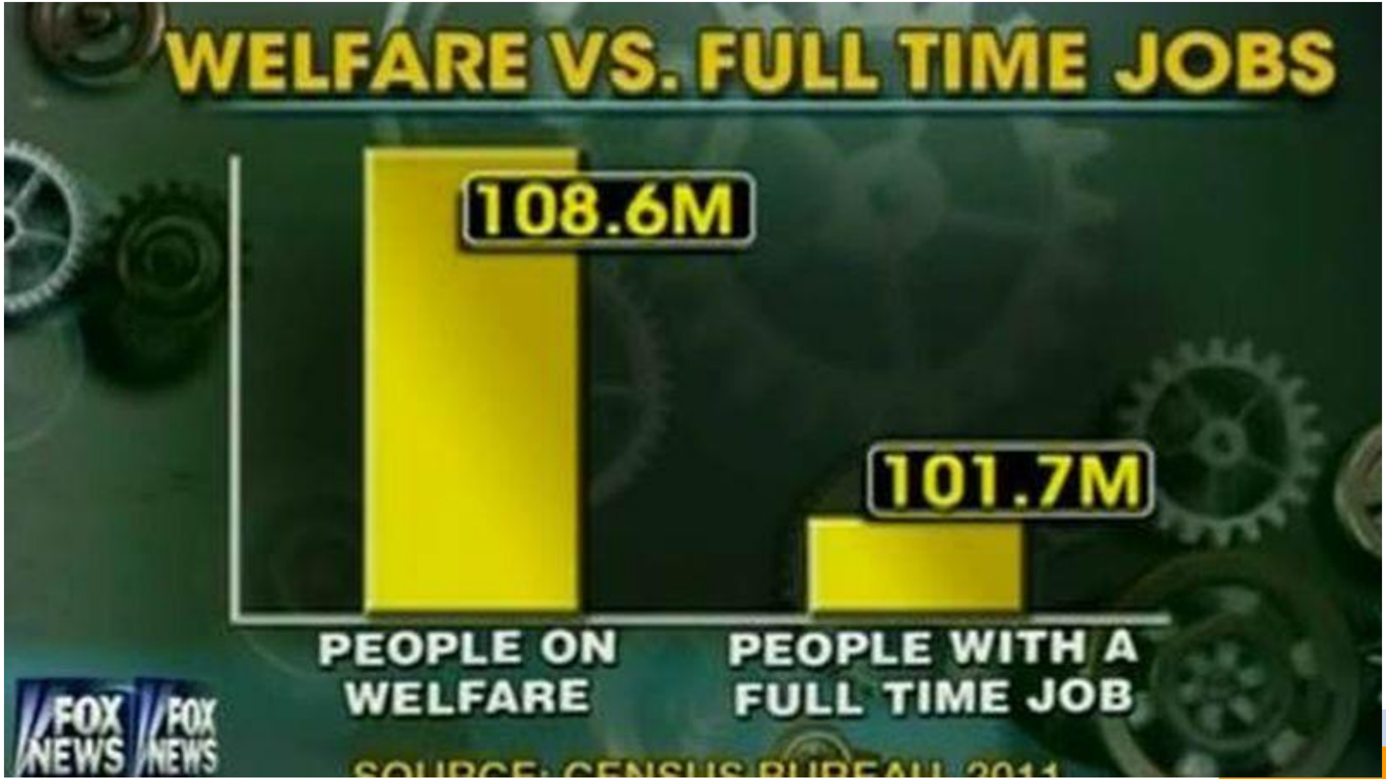
6

Abstract to News

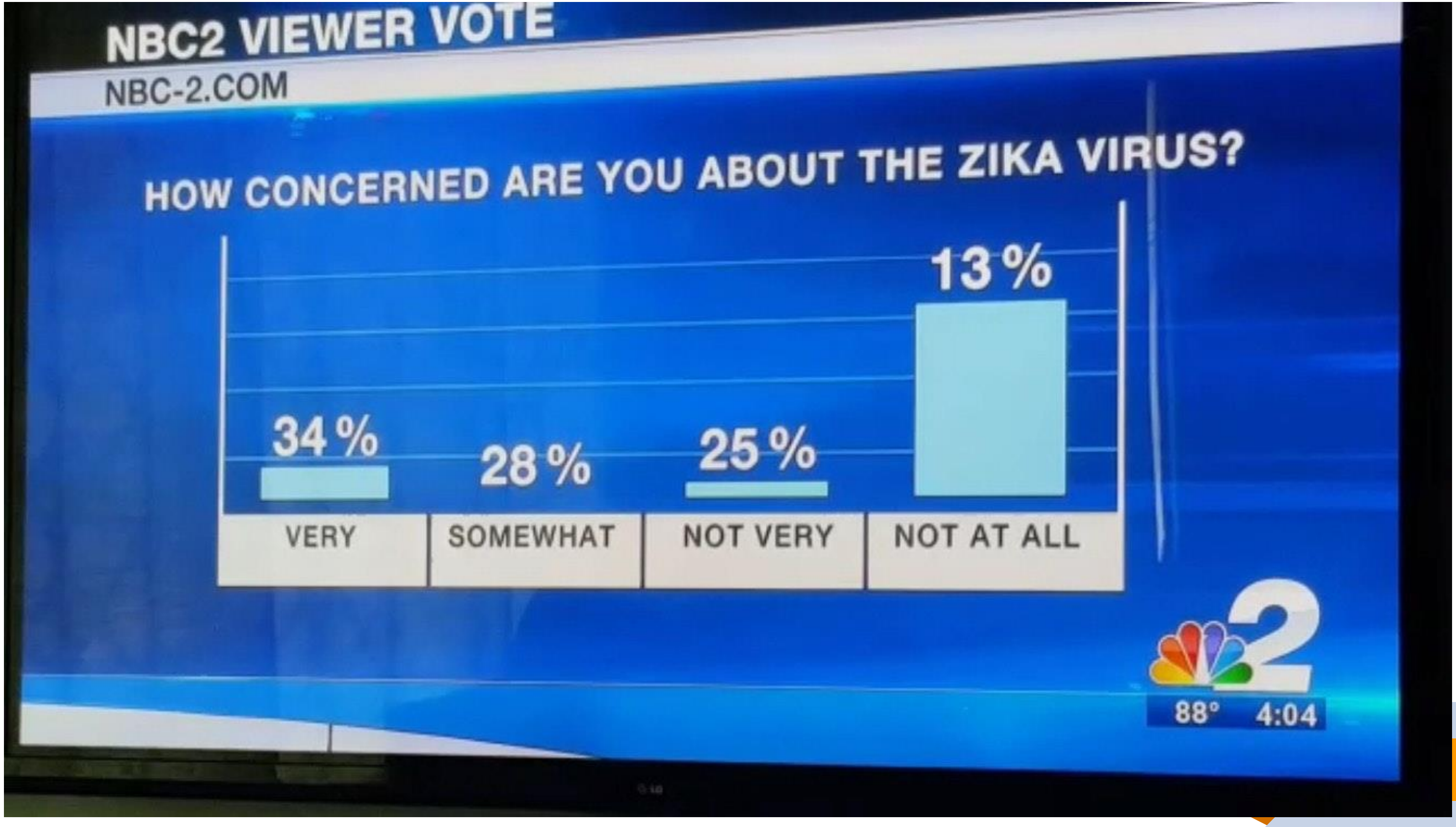
News and Statistics



News and Statistics



News and Statistics



Abstract to News

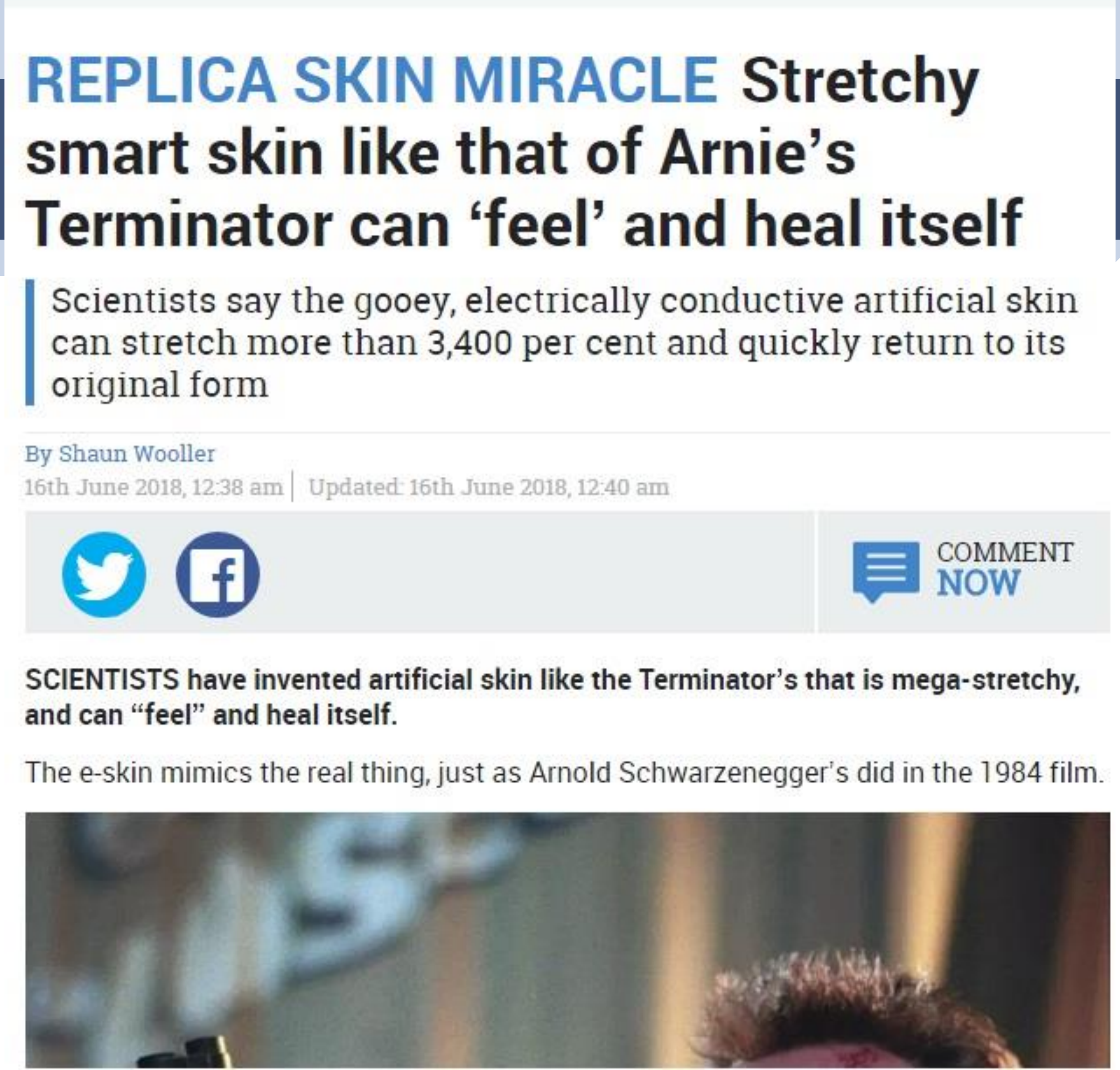
Media often does not understand statistical evidence. Our duty as scientists is to communicate our results in an understandable way to wider audiences.

Abstract to News

MXene electrochemical microsupercapacitor integrated with triboelectric nanogenerator as a wearable self-charging power unit

QiuJiang[a1]ChangshengWu[b1]ZhengjunWang[b]Aurelia ChiWang[b]Jr-HauHe[c]Zhong LinWang[b]Husam N.Alshareef[a]

The development of miniaturized, wearable, and implantable electronics has increased the demand for small stand-alone power modules that have steady output and long life-time. Given the limited capacity of energy storage devices, one promising solution is to integrate energy harvesting and storage materials to efficiently convert ambient mechanical energy to electricity for direct use or to store the harvested energy by electrochemical means. Here, a highly compact self-charging power unit is proposed by integrating triboelectric nanogenerator with MXene-based microsupercapacitors in a wearable and flexible harvester-storage module. The device can utilize and store the random energy from human activities in a standby mode and provide power to electronics when active. As a result, our microsupercapacitor delivers a capacitance of 23 mF/cm^2 with 95% capacitance retention after 10,000 charge-discharge cycles, while the triboelectric nanogenerator exhibits a maximum output power of 7.8 μW/cm^2. Given the simplicity and compact nature, our device can be integrated with a variety of electronic devices and sensors.



REPLICA SKIN MIRACLE Stretchy smart skin like that of Arnie's Terminator can 'feel' and heal itself

> Scientists say the gooey, electrically conductive artificial skin can stretch more than 3,400 per cent and quickly return to its original form

By Shaun Wooller

16th June 2018, 12:38 am | Updated: 16th June 2018, 12:40 am

SCIENTISTS have invented artificial skin like the Terminator's that is mega-stretchy, and can "feel" and heal itself.

The e-skin mimics the real thing, just as Arnold Schwarzenegger's did in the 1984 film.

Abstract to News

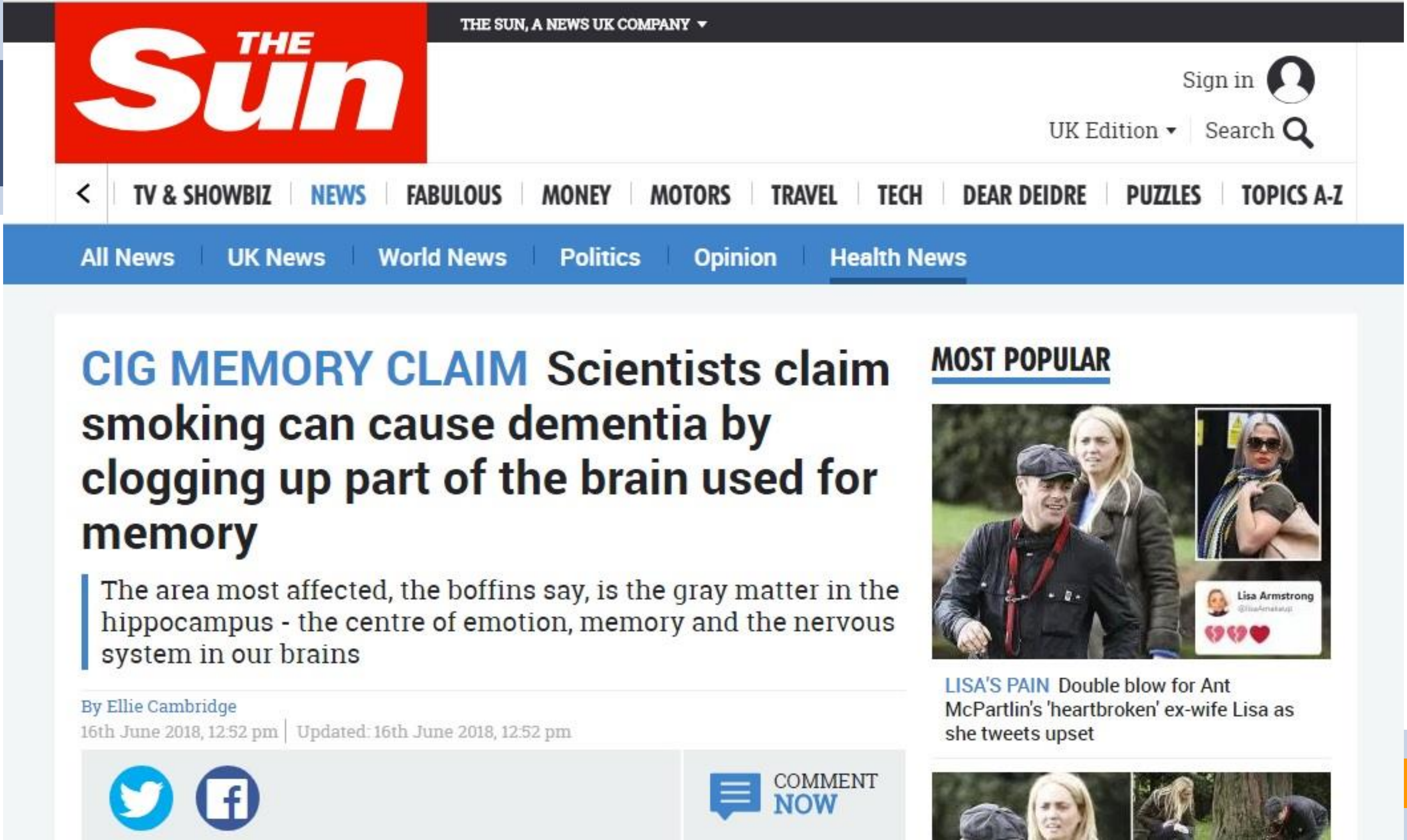
Hippocampal Calcifications: Risk Factors and Association with Cognitive Function

Esther J. M. de Brouwer , Remko Kockelkoren, Jules J. Claus, Annemarieke de Jonghe, Mirjam I. Geerlings, Thomas E. F. Jongsma, Willem P. T. M. Mali, Jeroen Hendrikse, Pim A. de Jong, Huiberdina L. Koek

Purpose: To identify risk factors for hippocampal calcifications and to investigate the association between hippocampal calcifications and cognitive function.

Results: A total of 1991 patients (mean age, 78 years; range, 45–96 years) were included. The mean age of women was 79 years (range, 47–96 years), and the mean age of men was 77 years (range, 45–95 years). Of the 1991 patients, 380 (19.1%) had hippocampal calcifications. Older age (odds ratio [OR] per year, 1.05; 95% confidence interval [CI]: 1.03, 1.06), diabetes mellitus (OR, 1.50; 95% CI: 1.12, 2.00), and smoking (OR, 1.49; 95% CI: 1.05, 2.10) were associated with the presence of hippocampal calcifications. No associations were found between presence and severity of hippocampal calcifications and cognitive function.

Conclusion: Older age, diabetes mellitus, and smoking were associated with an increased risk of hippocampal calcifications. A greater degree of hippocampal calcifications was not associated with lower cognitive function in patients with memory complaints.



CIG MEMORY CLAIM Scientists claim smoking can cause dementia by clogging up part of the brain used for memory

> The area most affected, the boffins say, is the gray matter in the hippocampus - the centre of emotion, memory and the nervous system in our brains

By Ellie Cambridge

16th June 2018, 12:52 pm | Updated: 16th June 2018, 12:52 pm

LISA'S PAIN Double blow for Ant McPartlin's 'heartbroken' ex-wife Lisa as she tweets upset

Abstract to News

- Describe your research in one sentence.
- Come up with an accurate, but exciting title.
- Come up with “The Sun” title for your research.

Fun and Educational:

Books:

- “The Norm Chronicles” by Michael Blastland and David Spiegelhalter
- “How to Lie with Statistics” by Darrell Huff

Podcasts:

- “More or Less: Behind the Statistics” by BBC Radio 4
- “The Infinite Monkey Cage” by BBC Radio 4

Linda Lapp linda.lapp@strath.ac.uk

SICSA PhD Conference 2018