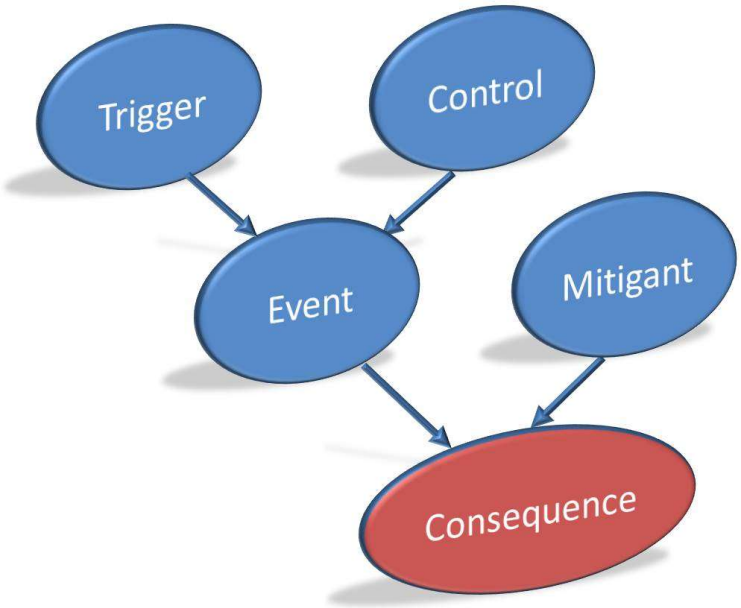


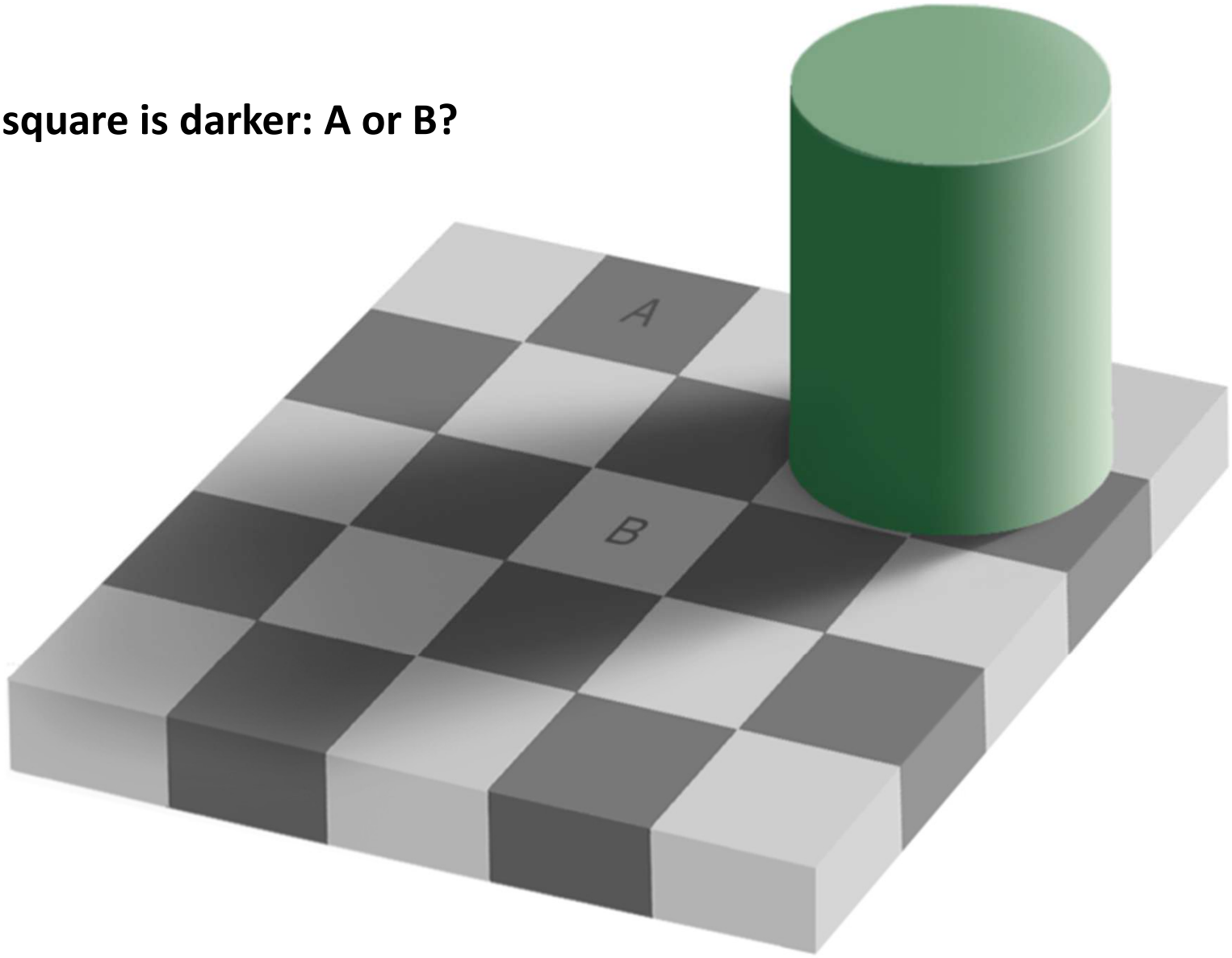
Risk and Decision Making for Data Science and AI

Lesson 1 Risk and Decision Making: Illusions and fallacies

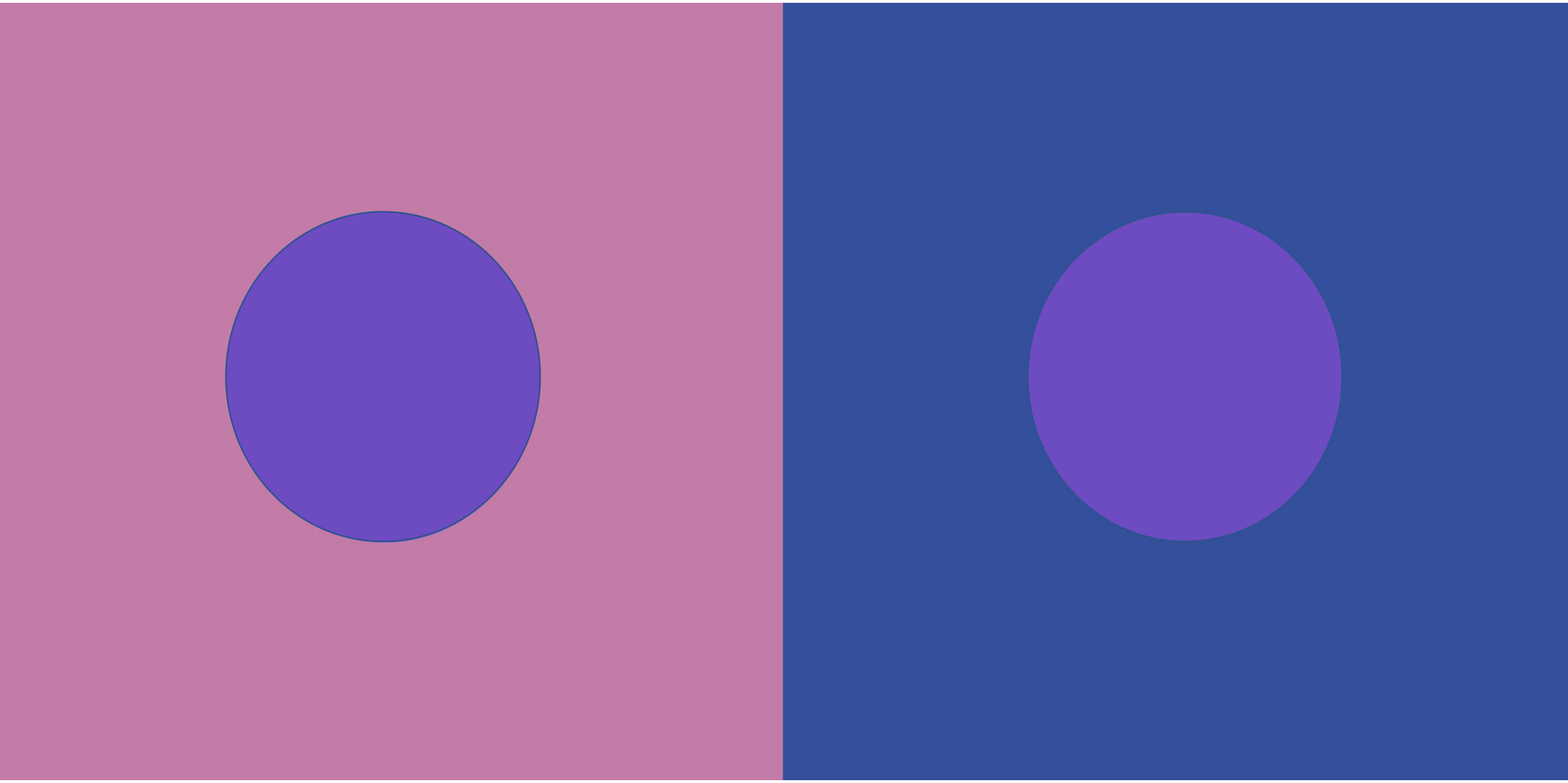
Norman Fenton

@ProfNFenton

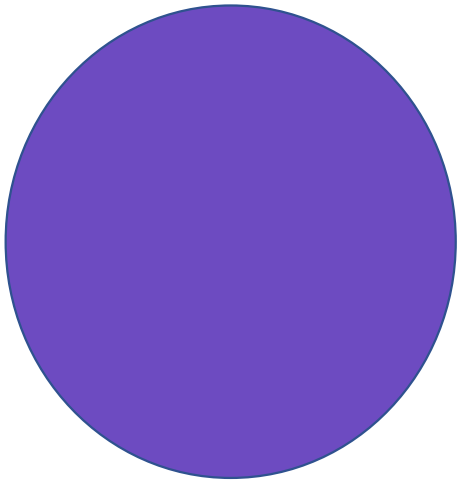
Which square is darker: A or B?

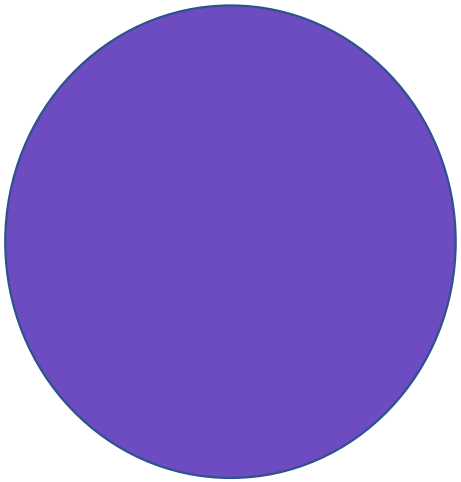


Are the circles different?

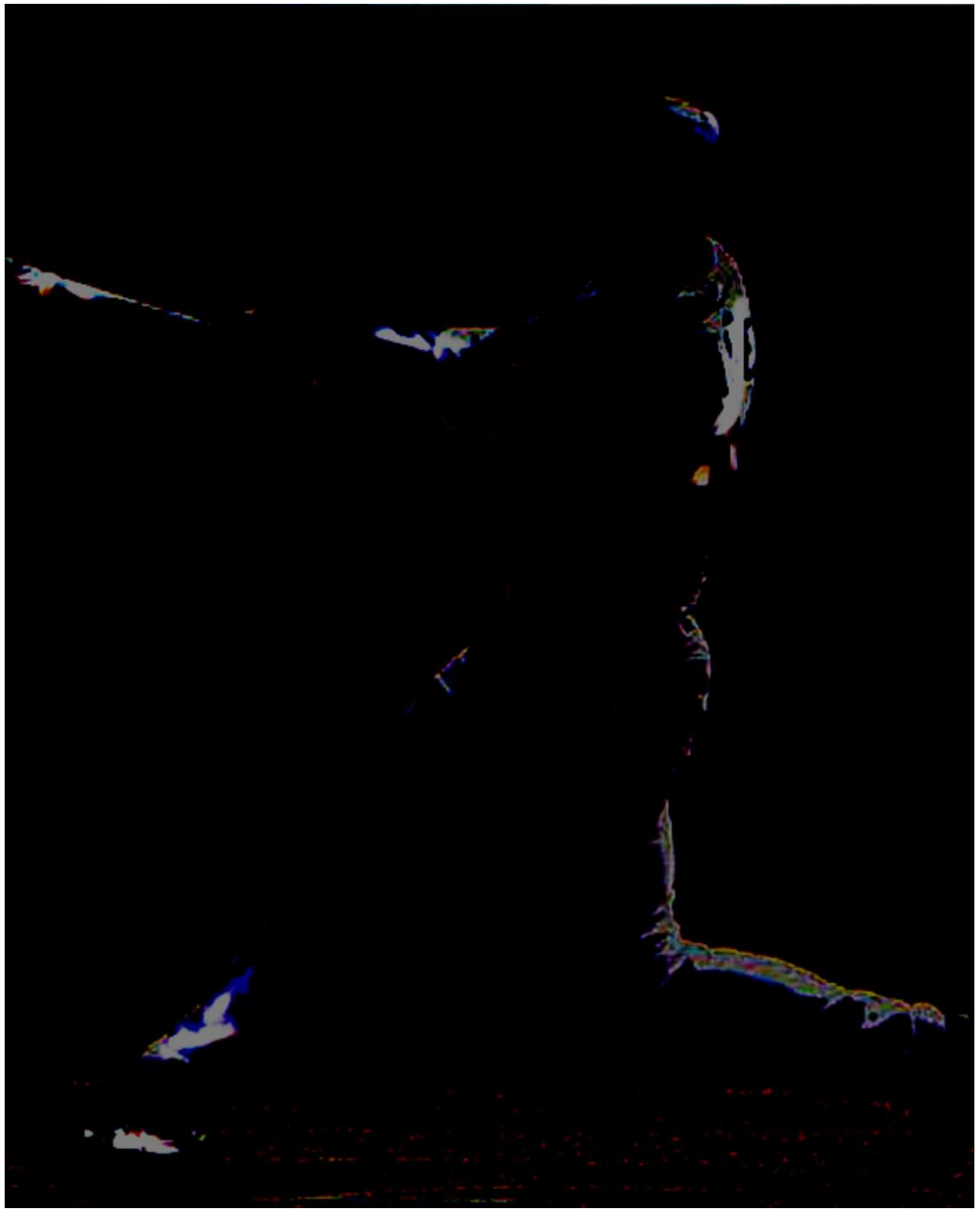


Are the circles different?

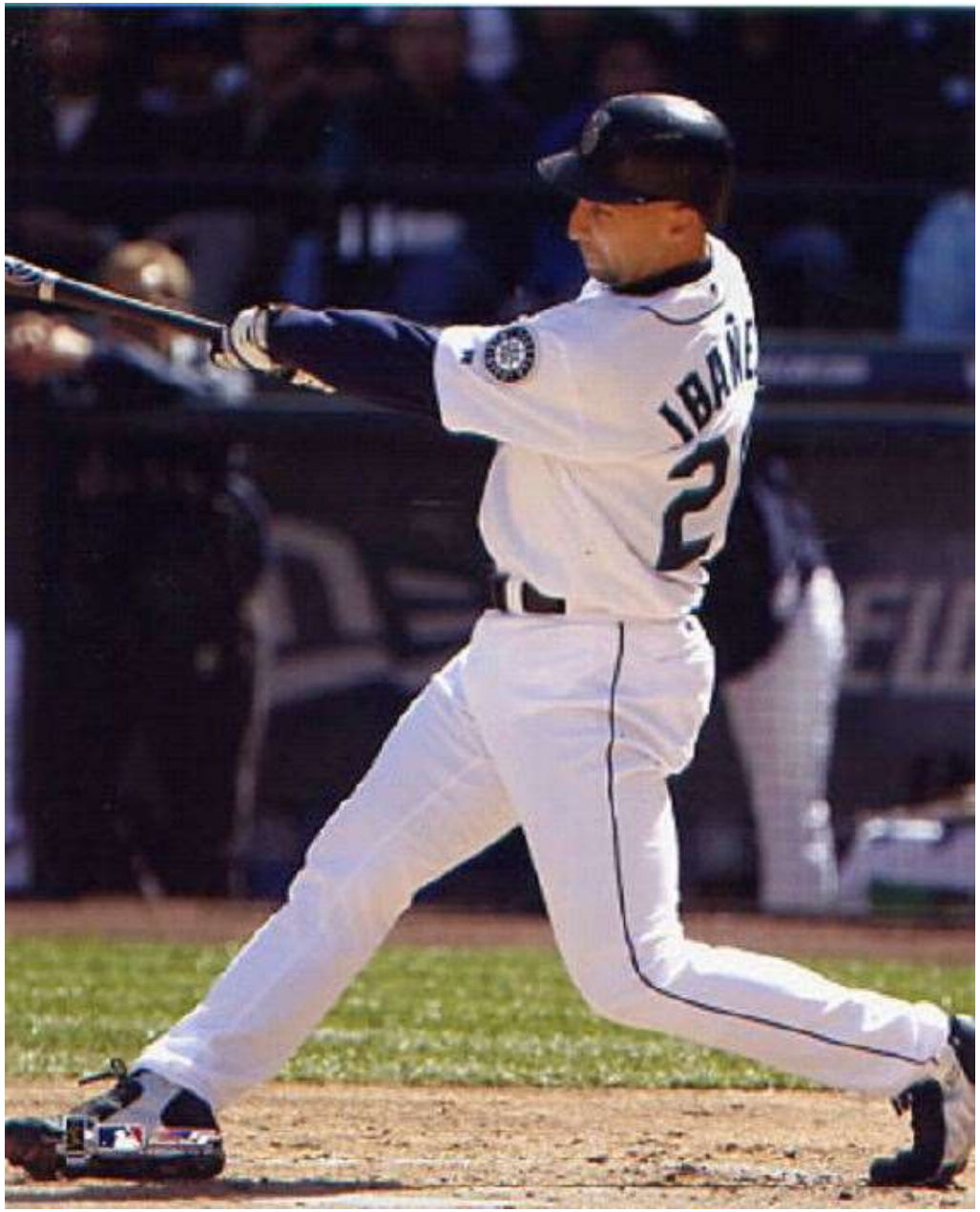




What do you see?



What do you see?



What do you see?

What do you see?

Confirmation Bias

No matter how much you try to resist knowledge of the target image, once you have seen it, it becomes impossible for it not to affect your judgement about the source image. This is a well-known cognitive illusion called confirmation bias.

Confirmation bias appears in many different guises and has serious implications in all walks of life, but especially for decision-making and risk assessment in science, the law, and medicine.

Which of these is the most likely sequence from tossing a fair coin 16 times?

A

T T T T T T T T T T T T T T T T

B

T H H T H T T H T H H T T H T H

Most mundane events are actually miraculous

Imagine shuffling a deck of cards and then revealing that they were arranged in this perfect order

The probability of that sequence is

$$\frac{1}{52} \times \frac{1}{51} \times \frac{1}{50} \times \frac{1}{49} \ldots\ldots \frac{1}{2} \times \frac{1}{1} = \frac{1}{52!}$$

The denominator is greater than the number of particles in the universe

But whatever sequence is revealed has the same probability – and is just as much an unpredictable miracle

Conversely many events that people believe are incredible miracles are actually completely mundane and predictable.

How incredible was this, and was a fraud enquiry really required?

One in 42,375,200 chance of this particular set of numbers being drawn on that day.
But that is the same as any particular set of numbers.
And nothing unusual about the relatively large number of winners (many people pick sequential sequences)

In fact there is about 12% that in any 10 year period there would be a set of numbers just like this drawn in a lottery somewhere in the world

https://probabilityandlaw.blogspot.com/2020/12/no-there-is-nothing-especially-unusual.html

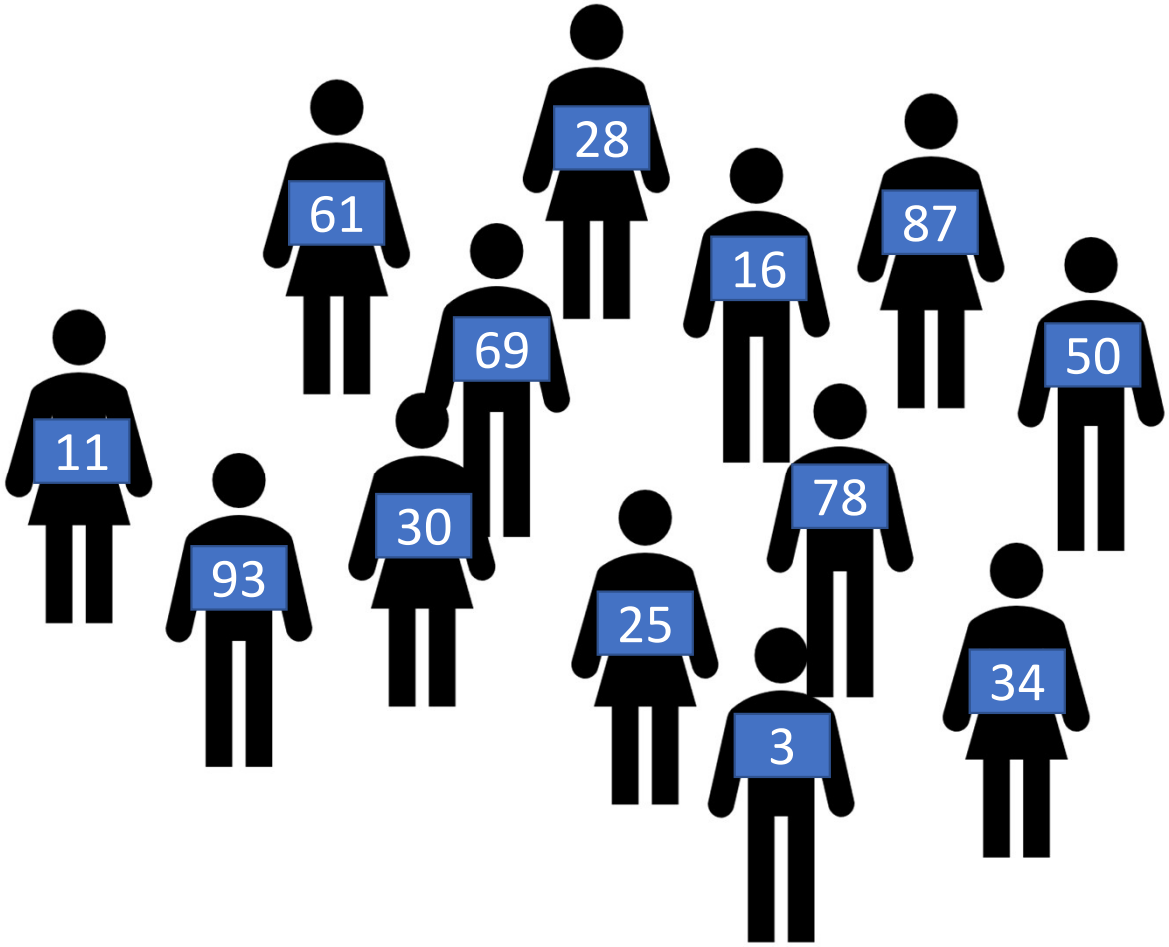
How incredible was this?

"The chances of this happening are approximately one in 200 billion"

One in 200 billion is the chance of a specific person winning a '6 from 49' ball lottery twice in two attempts

But what are the chances over a 20-year period that at least one person in the USA will win the lottery jackpot at least twice?

Imagine a room with 13 people….



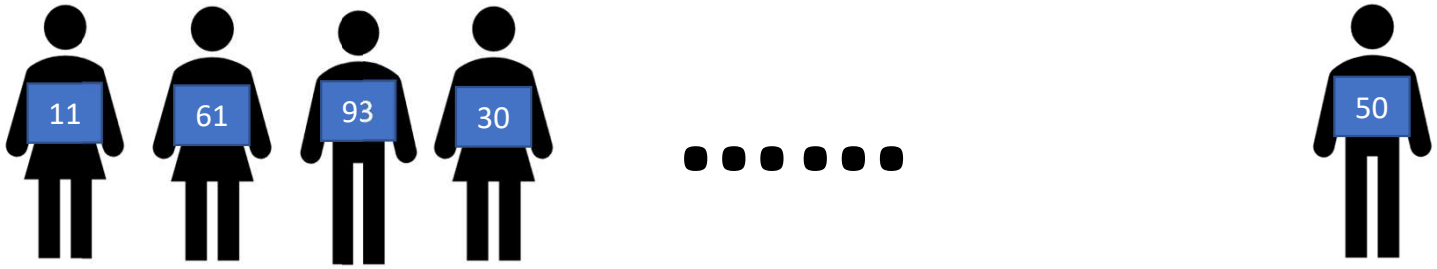
Each is assigned a random number from 1 to 100

What is the approximate probability at least 2 of them have the same number?

Calculating probability at least 2 have same number

If *A* is “at least two of 13 have same number” then *not A* is “all 13 have DIFFERENT numbers”

It’s easier to calculate the probability of *not A*. Then we know P(*A*) = 1 – P(*not A*)



Prob person 2 different to person 1 is 99/100

Prob person 3 different to person 1 and 2 is 98/100

Prob person 4 different to person 1, 2 and 3 is 97/100

Prob person 13 different to person 1,2,3,..and 12 is 88/100

So, probability of *not A* is the product of these 12 probabilities:

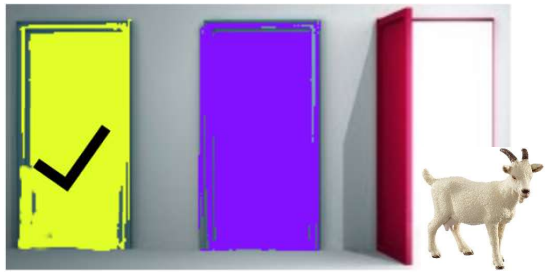
$$\frac{99}{100} \times \frac{98}{100} \times \frac{97}{100} \times \frac{96}{100} \times \frac{95}{100} \times \frac{94}{100} \times \frac{93}{100} \times \frac{92}{100} \times \frac{91}{100} \times \frac{90}{100} \times \frac{89}{100} \times \frac{88}{100} = 0.47$$

So, probability of *A* is 0.53, i.e. 53%

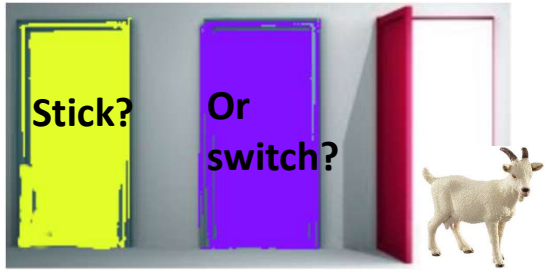
The Monty Hall Problem

There is a goat behind each of two doors but a sports car behind the other.

You choose one door – say the yellow one

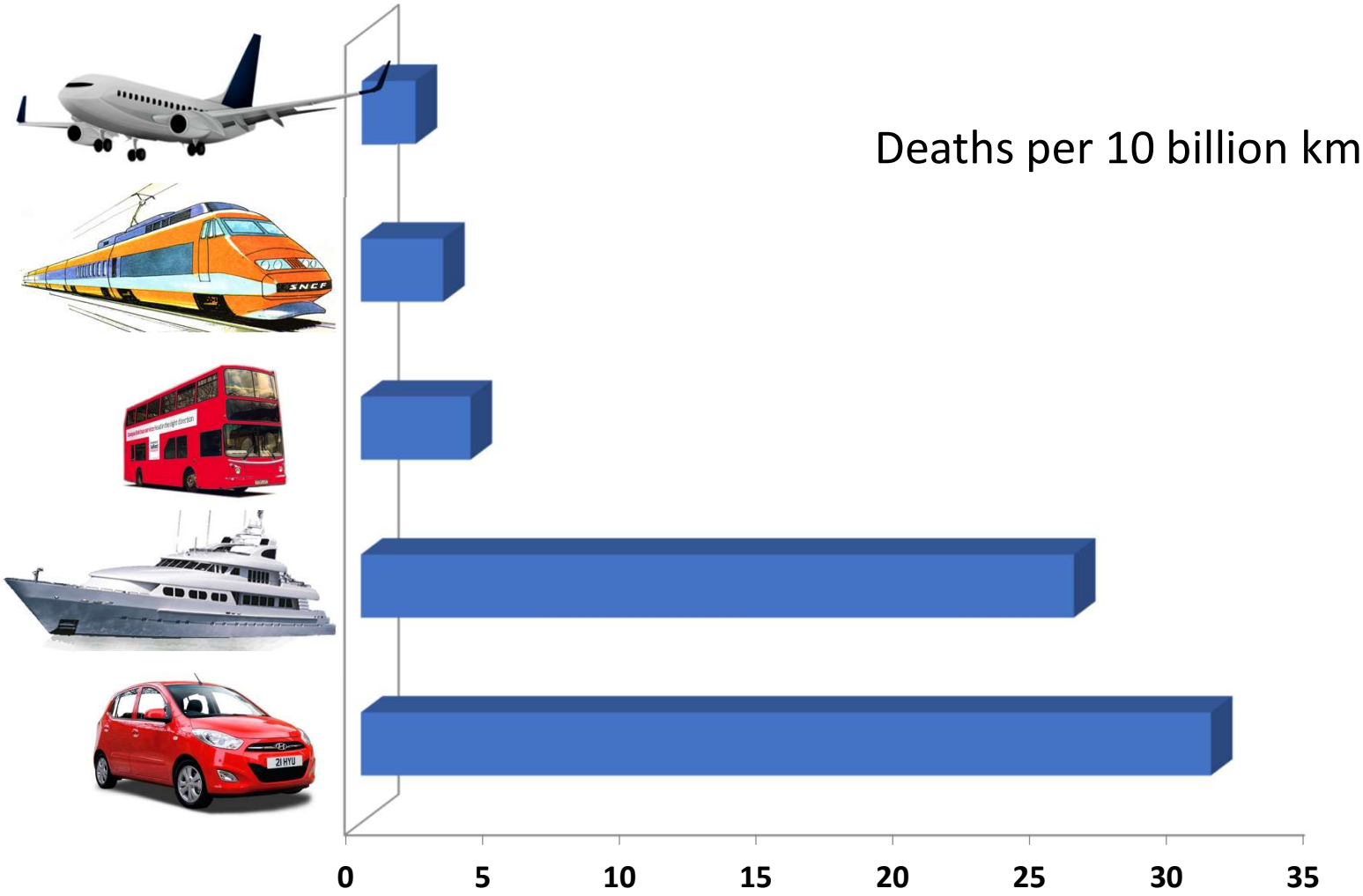


Monty Hall – who knows what's behind each door, must then open another door that has a goat behind it.

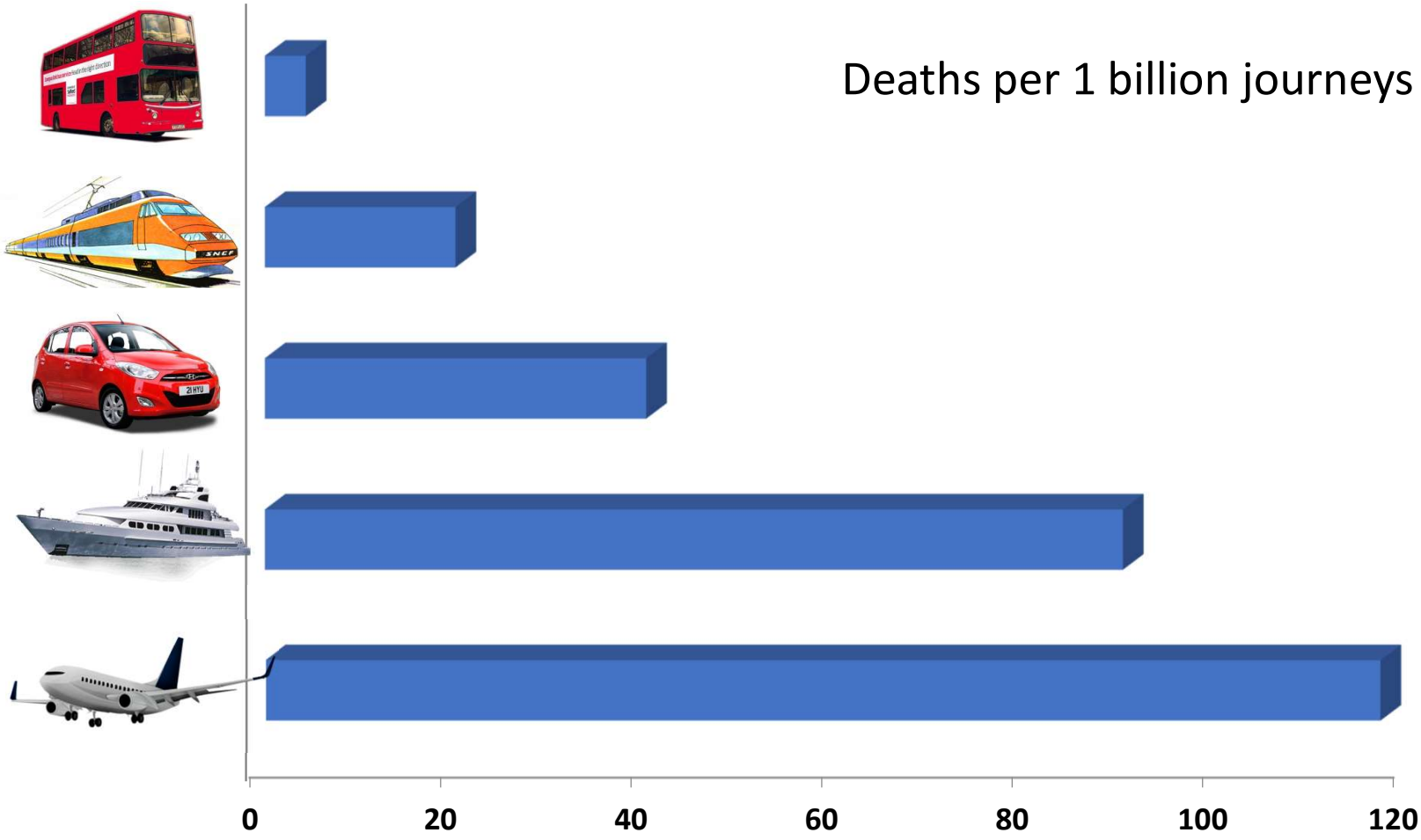


Should you stick with the yellow door or switch to the blue one?

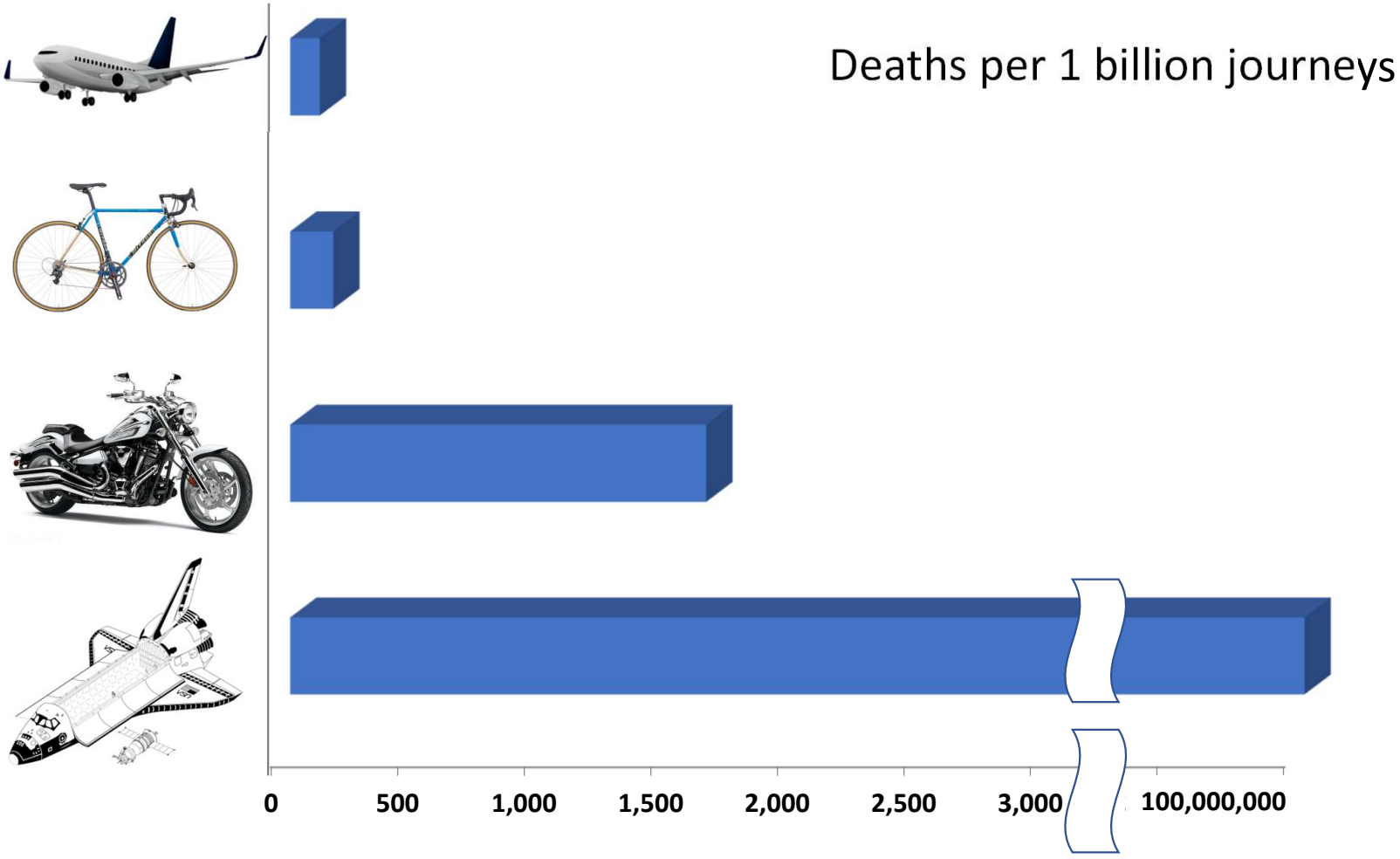
Safest form of travel?

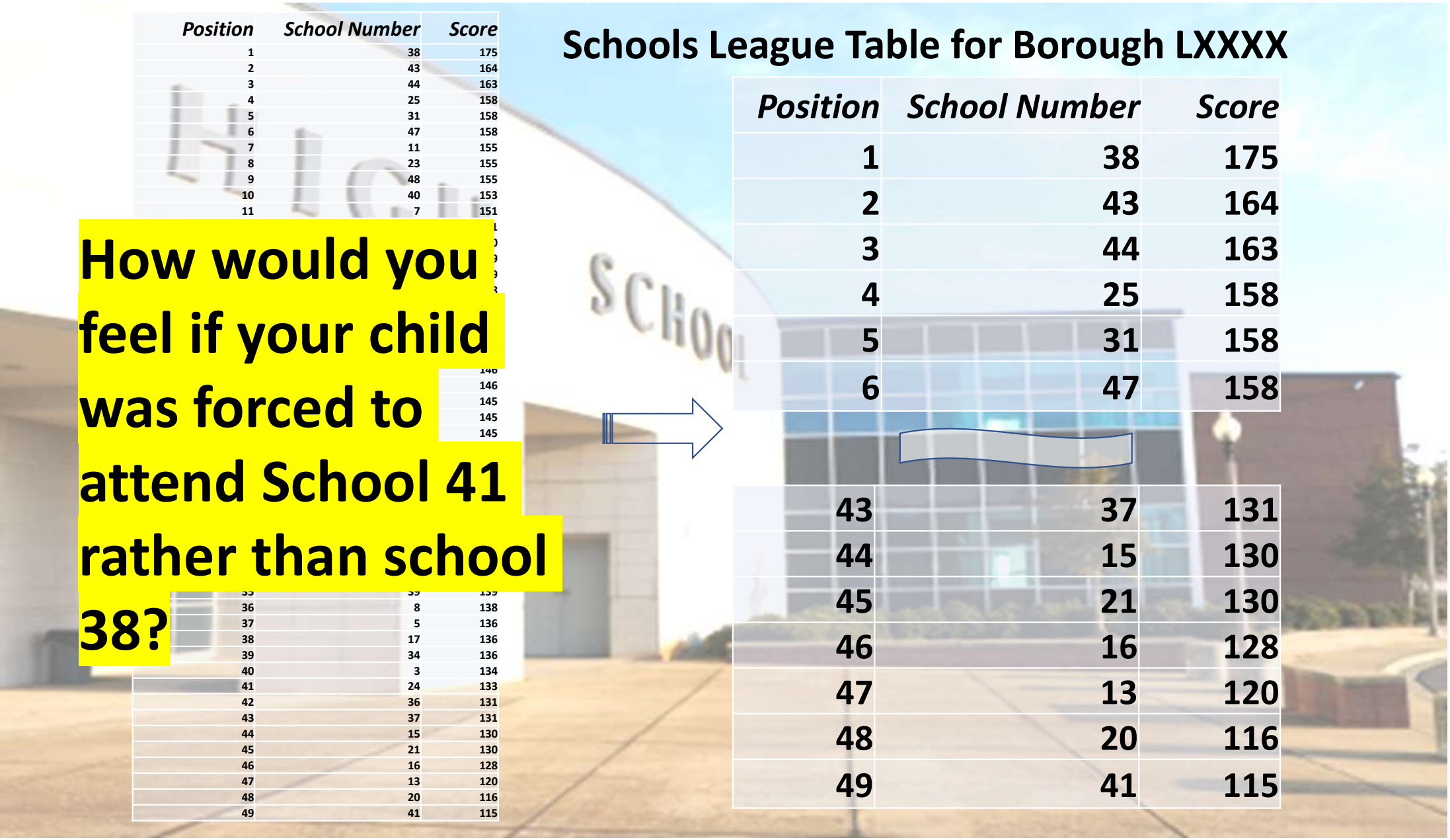


Safest form of travel?



Safest form of travel?





Position	School Number	Score
1	38	175
2	43	164
3	44	163
4	25	158
5	31	158
6	47	158
7	11	155
8	23	155
9	48	155
10	40	153
11	7	151
36	8	138
37	5	136
38	17	136
39	34	136
40	3	134
41	24	133
42	36	131
43	37	131
44	15	130
45	21	130
46	16	128
47	13	120
48	20	116
49	41	115

How would you feel if your child was forced to attend School 41 rather than school 38?

Schools League Table for Borough LXXXX

Position	School Number	Score
1	38	175
2	43	164
3	44	163
4	25	158
5	31	158
6	47	158
43	37	131
44	15	130
45	21	130
46	16	128
47	13	120
48	20	116
49	41	115

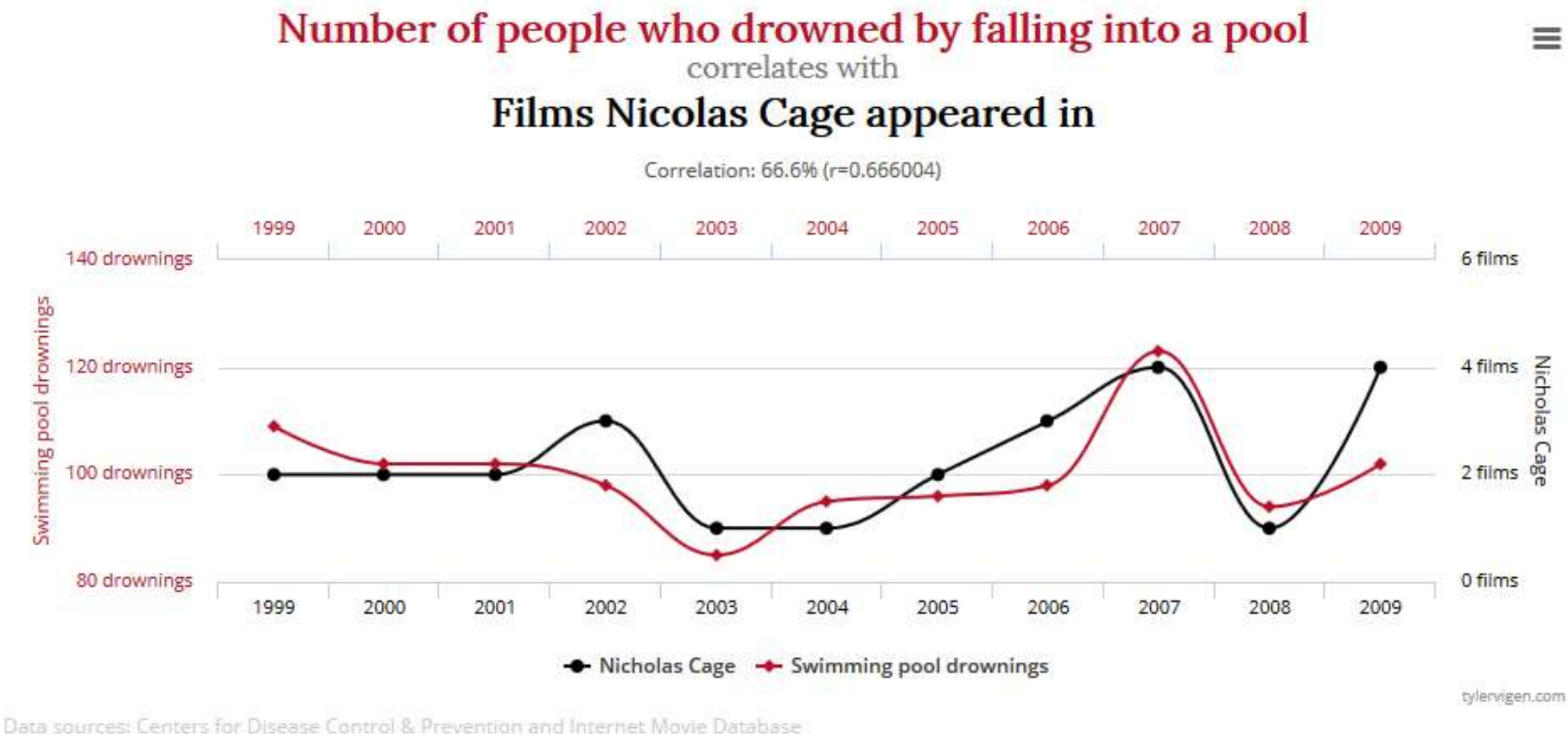
Position	Ball Number	Time drawn
1	38	175
2	43	164
3	44	163
4	25	158
5	31	158
6	47	158
7	11	155
8	23	155
9	48	155
10	40	153
11	7	151
12	30	151
13	6	150
14	9	149
15	33	149
16	19	148
17	10	147
18	12	147
19	32	147
20	2	146
21	27	146
22	42	146
23	28	145
24	35	145
25	49	145
26	45	144
27	46	143
28	1	142
29	18	142
30	22	141
31	26	141
32	4	140
33	14	140
34	29	140
35	39	139
36	8	138
37	5	136
38	17	136
39	34	136
40	3	134
41	24	133
42	36	131
43	37	131
44	15	130
45	21	130
46	16	128
47	13	120
48	20	116
49	41	115

The National Lottery®

These are simply the number of times the particular balls were drawn in the first 1172 games

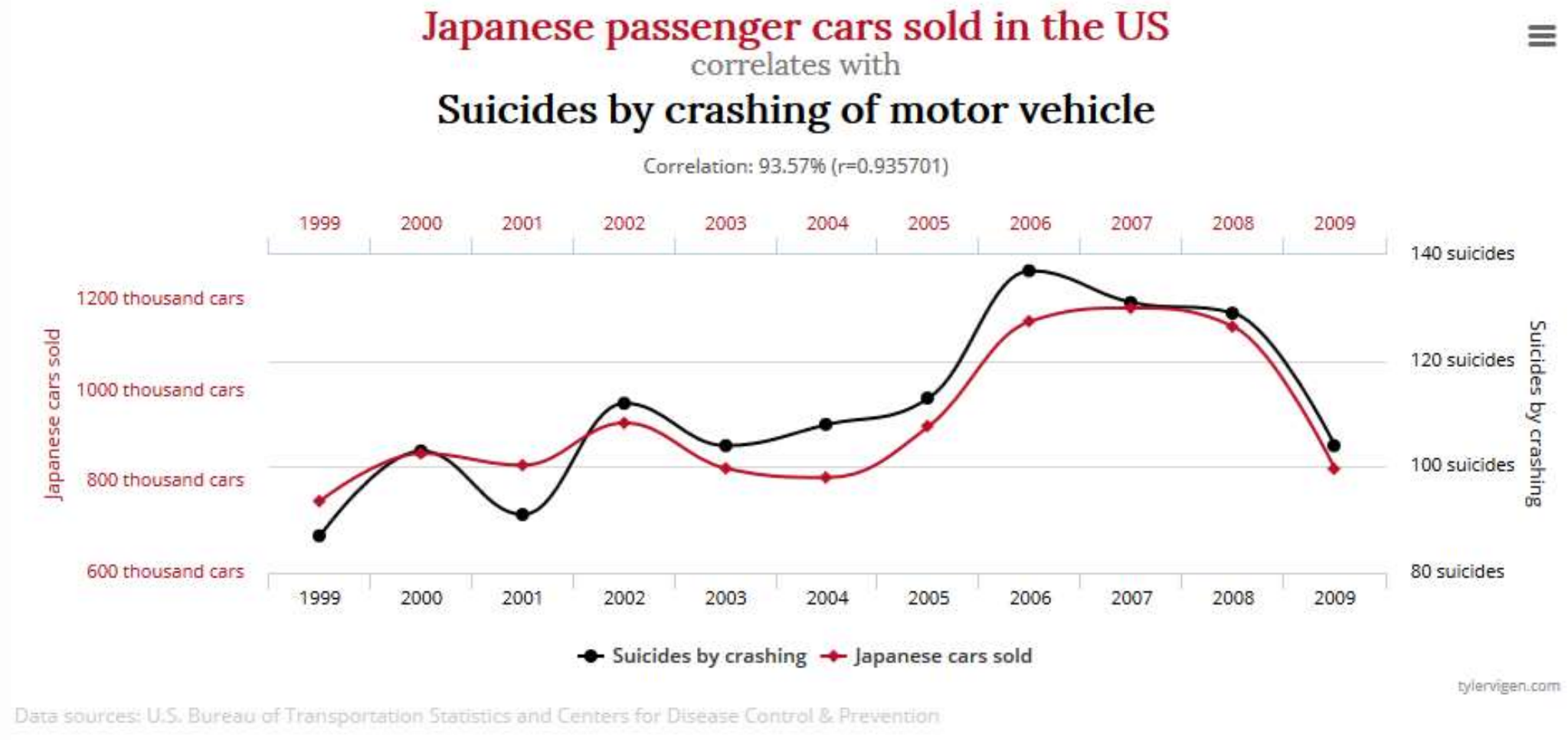
So is ball number 38 'better' than ball number 41?

Spurious correlations?



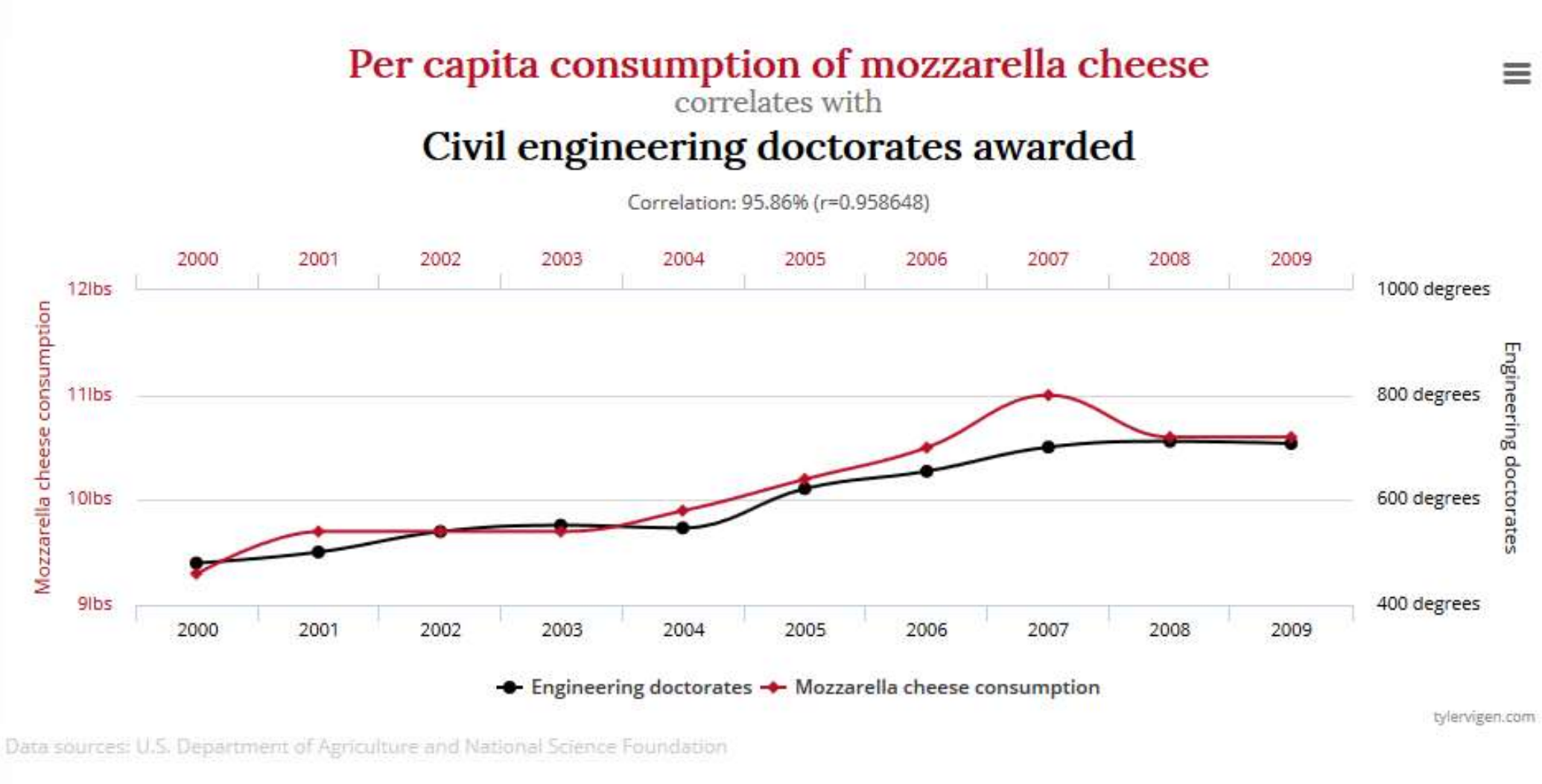
Reproduced from tylervigen.com

Spurious correlations?



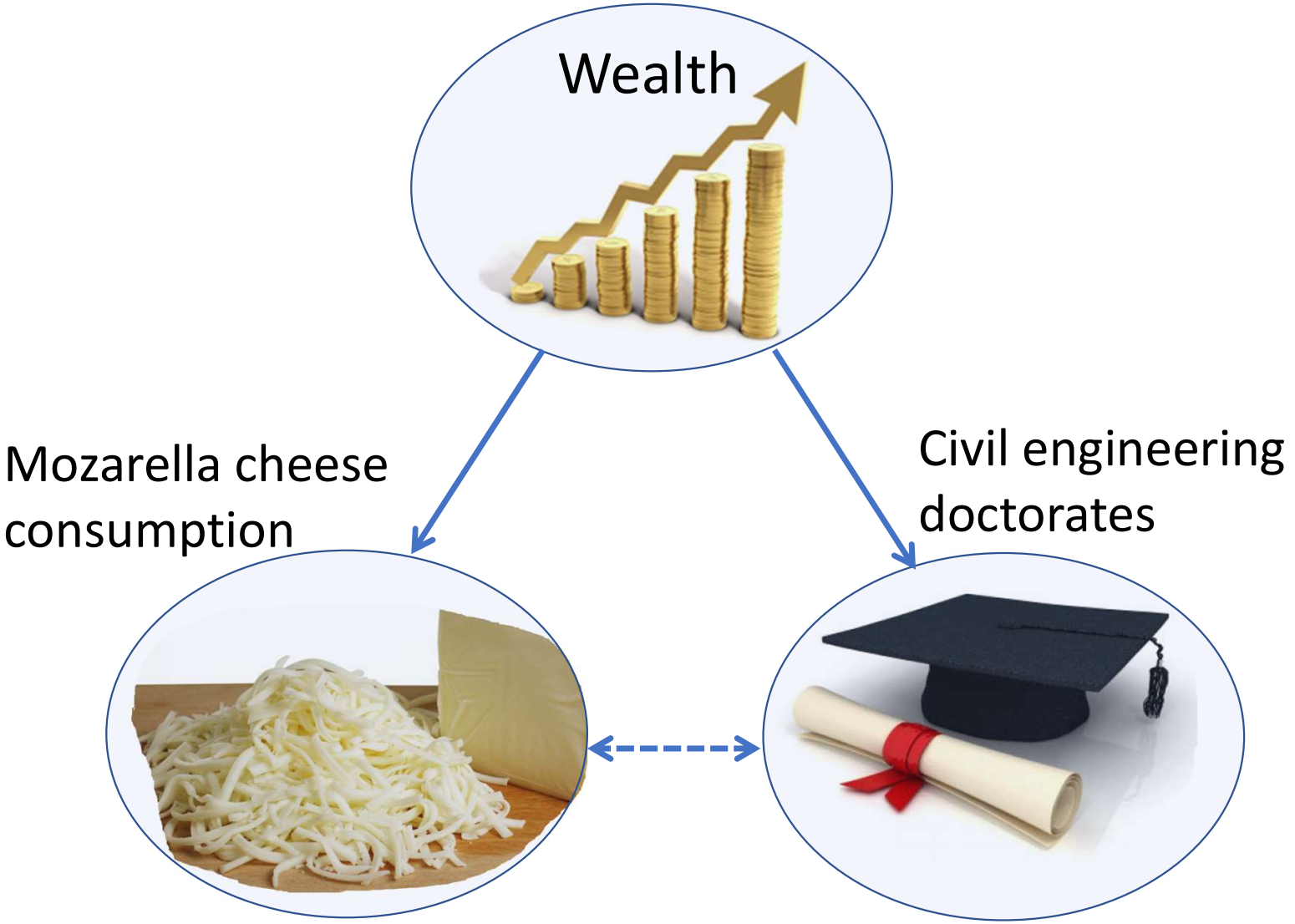
Reproduced from tylervigen.com

Spurious correlations?



Reproduced from tylervigen.com

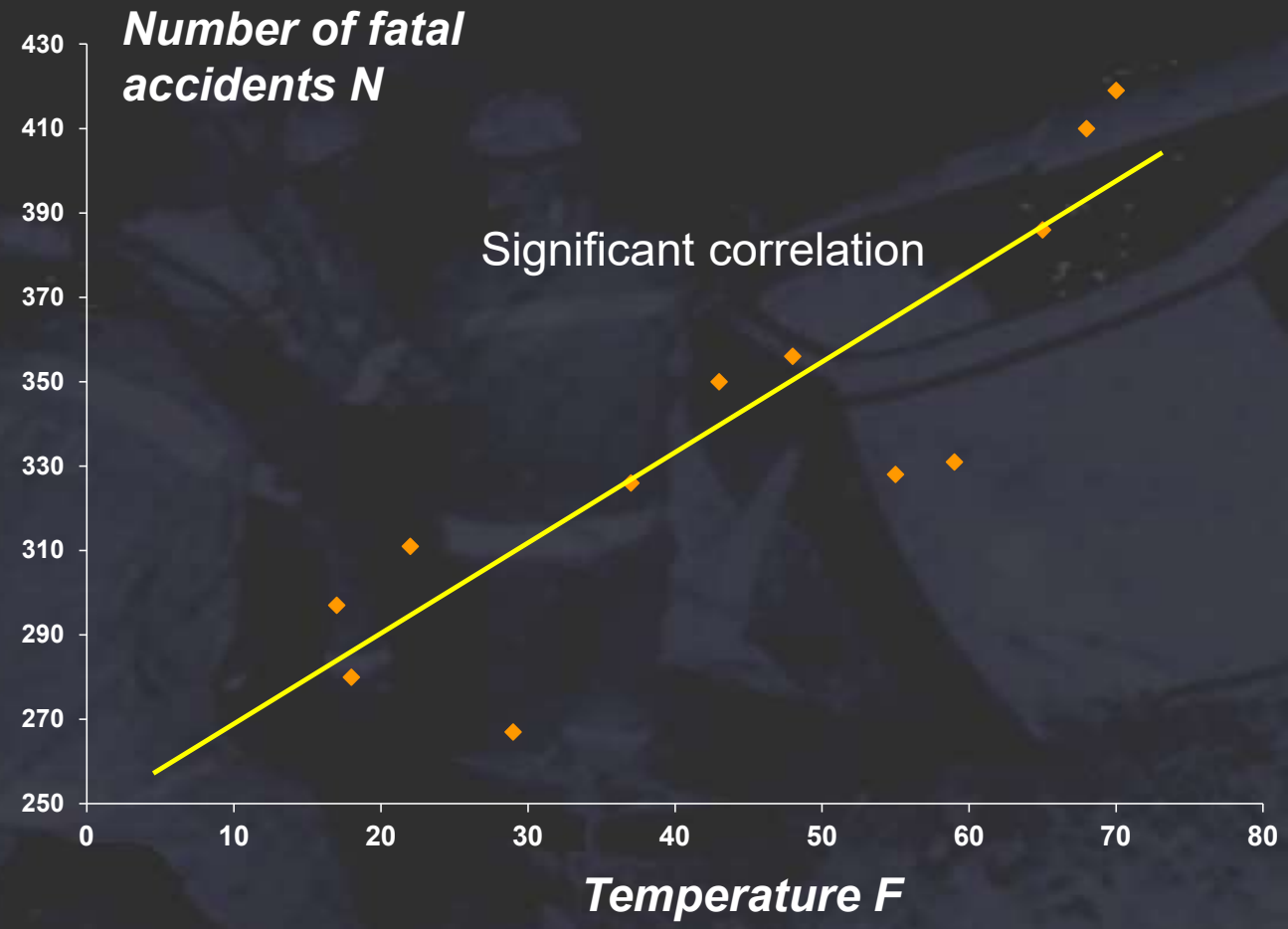
Hidden common causal explanation

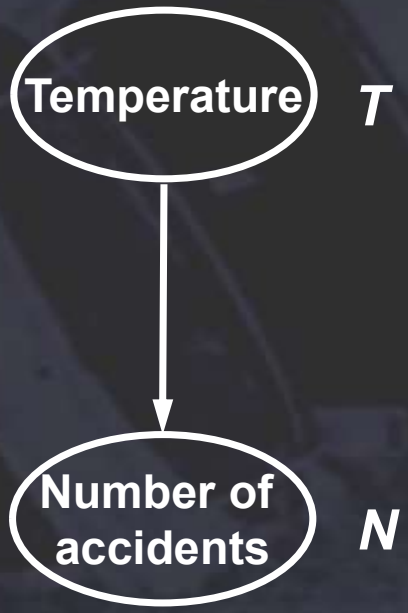


When do most road fatalities occur in Europe and USA?

Month	Average temperature	Total fatal crashes
January	17.0	297
February	18.0	280
March	29.0	267
April	43.0	350
May	55.0	328
June	65.0	386
July	70.0	419
August	68.0	410
September	59.0	331
October	48.0	356
November	37.0	326
December	22.0	311

Assessing Risk of Road Accidents





$N = 2.1 \times T + 243$

Regression equation

HEADLINE: Safest to drive when roads are more dangerous

A paradox?

Fred and Jane both take 10 modules in total over 2 years. If Fred's average grades are better than Jane's every year he must end up with a better overall average grade than Jane, mustn't he?

	Year 1 average	Year 2 average	Overall Average ?
Fred	50	70	60
Jane	40	62	51

Fred took 7 modules in Year 1 and 3 modules in Year 2
Jane took 2 modules in Year 1 and 8 modules in Year 2.

	Year 1 total	Year 2 total	Overall total	Real Overall average
Fred	350 (7 x 50)	210 (3 x 70)	560	56.0
Jane	80 (2 x 40)	496 (8 x 62)	576	57.6

The Telegraph

Two glasses of wine a night triples risk of mouth cancer, Government warns

Drinking two large glasses of wine a day triples the risk of developing mouth cancer, a government campaign will warn.

Two glasses of wine a night triples the chance of mouth cancer, Government will warn Photo: ALAMY

By Laura Donnelly, Health Correspondent
9:45AM GMT 05 Feb 2012

What does the increased risk mean here?

Nine hours of sleep and long naps raise your stroke risk, says study

By **Stephen Beech**

IF you sleep more than nine hours a night or grab a lengthy afternoon nap your risk of suffering a stroke rises by up to a quarter, a study has warned.

Researchers found people who slept for long periods were 23 per cent more likely to have a stroke than people who slept less than eight hours per night.

The study also showed that over-60s who took a regular midday nap of more than 90 minutes were 25 per cent more likely to later suffer a stroke than those who had a regular nap lasting up to an hour or no nap.

And people who said they slept poorly were 29 per cent more likely to have a stroke than those who had good sleep.

The study, published in the journal Neurology, involved 31,750 people in China with an average age of 62. They did not have any history of stroke or other major health problems at the start of the study. Participants were followed for an average of six years, during which time there were 1,557 stroke cases.

Author Dr Xiaomin Zhang, of Huazhong University of Science and Technology in China, added: "These results highlight the importance of moderate napping and sleeping, and maintaining good sleep quality, especially in middle-age and older adults.

"Long napping and sleeping may suggest an overall inactive lifestyle, which is also related to increased risk of stroke."

Suppose a screening test for COVID-19 is:

99% accurate for those with the virus

so the true positive rate – sensitivity - is 99%, meaning 99% of those with the disease will test positive, 1% will test negative

95% accurate for those without the virus

so the true negative rate – specificity - is 95%, meaning 95% of those without the virus will test negative and 5% will test positive

It is estimated the current population infection rate for the virus is 1 in a 1000

Sarah tests positive.

What is the probability Sarah has the virus?

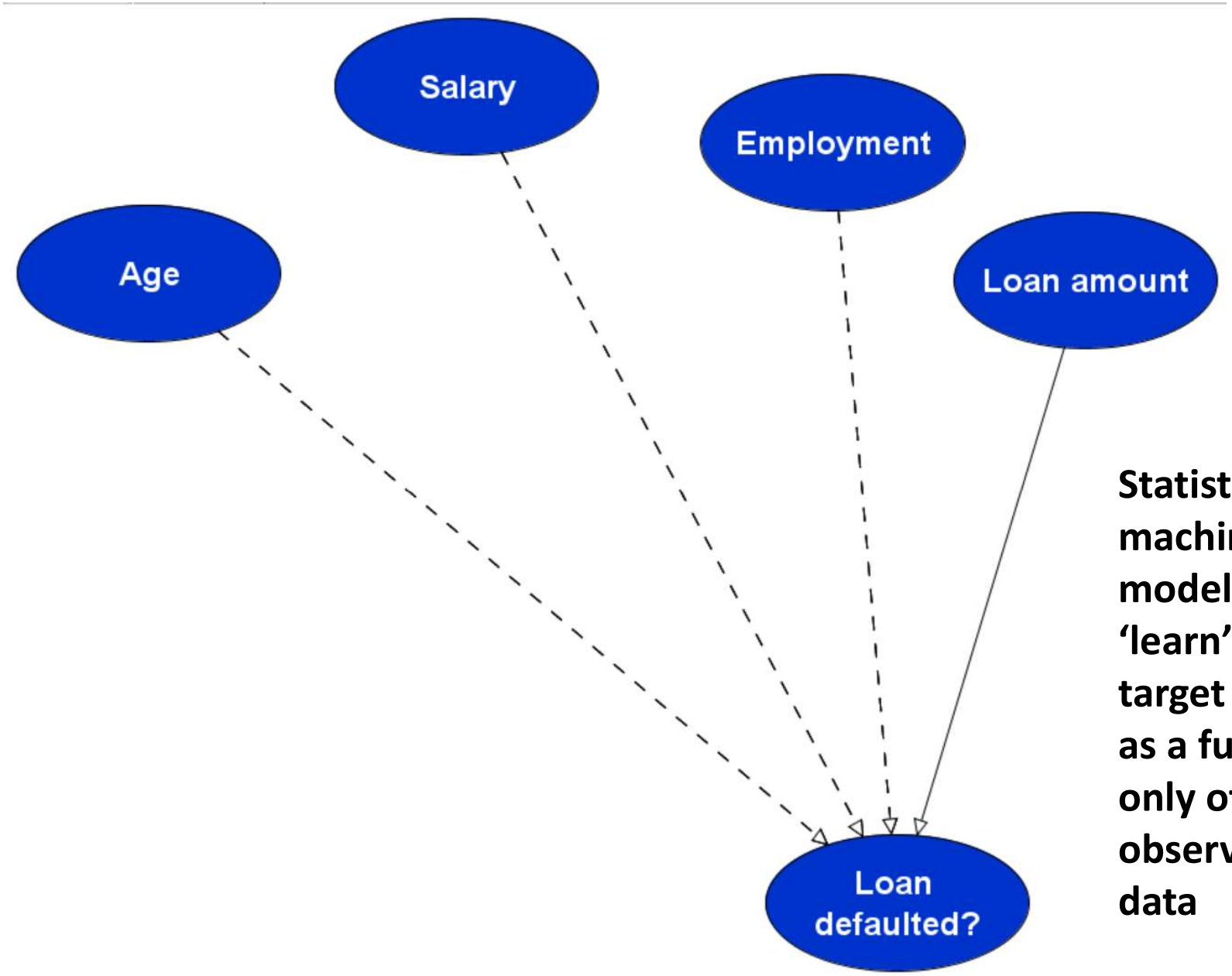
A typical “Big Data”-driven problem: Bank Loan decisions

Customer	Age	Marital status	Employment status	Home owner	Salary	Loan	…	Defaulted
1	37	M	Employed	Y	50000	10000	…	N
2	45	M	Self-employed	Y	60000	5000	…	N
3	26	M	Self-employed	Y	30000	20000	…	Y
4	29	S	Employed	N	50000	15000	…	N
5	26	M	Employed	Y	90000	20000	…	N
6	35	S	Self-employed	N	70000	20000	…	Y
7	32	M	Self-employed	Y	40000	5000	…	N
8	37	M	Employed	Y	25000		…	Y
9	18	S	Unemployed	N	0	50000	…	N
10	40	M	Employed	Y	65000	45000	…	N
11	21	S	Employed	N	20000	10000	…	Y
12	30	S	Employed	N	40000	5000	…	N
13	22	M	Self-employed	N	30000	10000	…	Y
14	35	M	Unemployed	Y	0	3000	…	Y
15	19	S	Unemployed	N	0	100000	…	N
…	…	…	…	…	…	…	…	…
100001	34	M	Employed	Y	45000	1000	…	N
100002	28	S	Self-employed	N	25000	2000	…	N
100003	19	S	Unemployed	N	0	25000	…	N
…	…	…	…	…	…	…	…	…

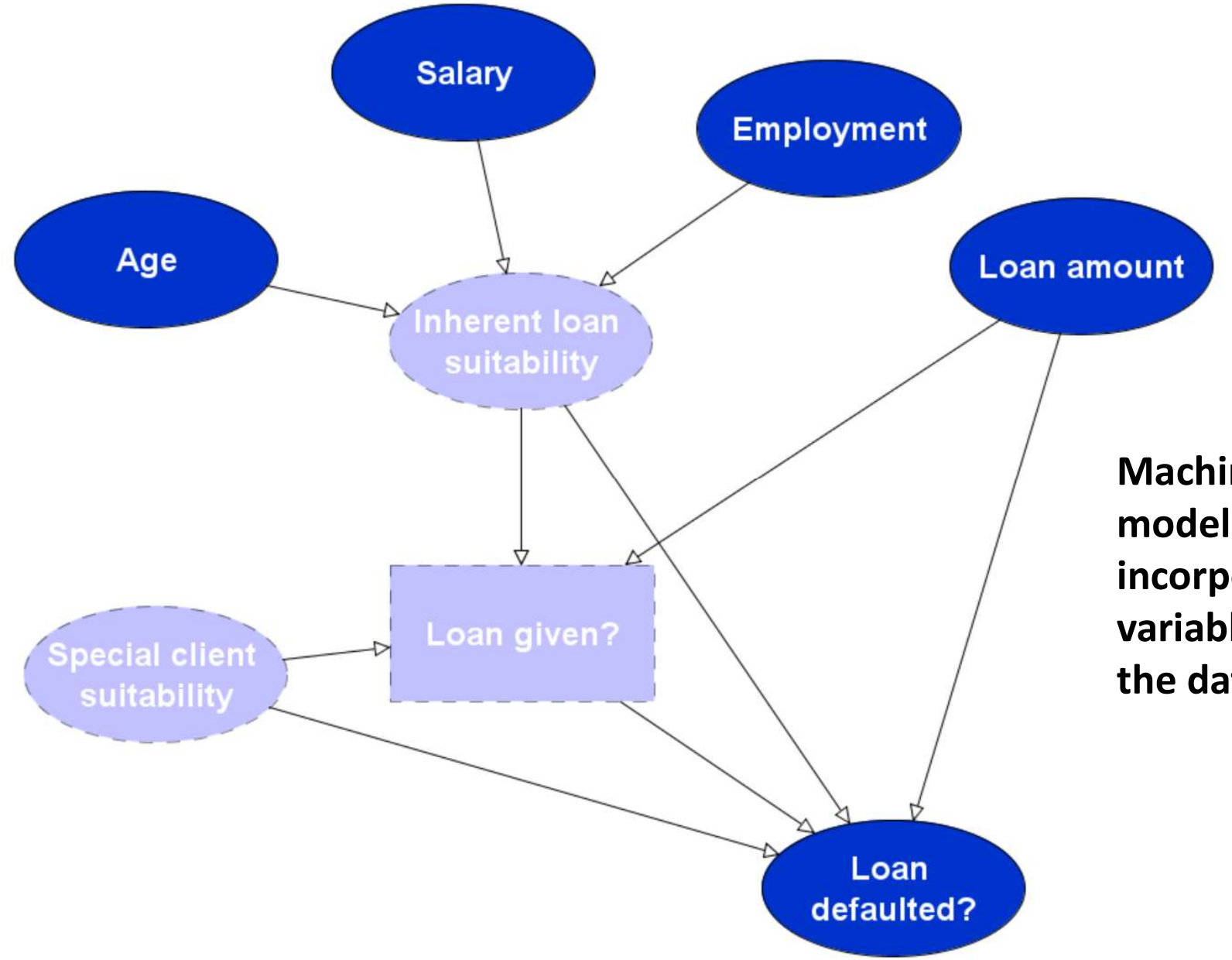
Big data – but it is ‘censored’ Restricted to those *given* loans

Can learn nothing about those not given loans

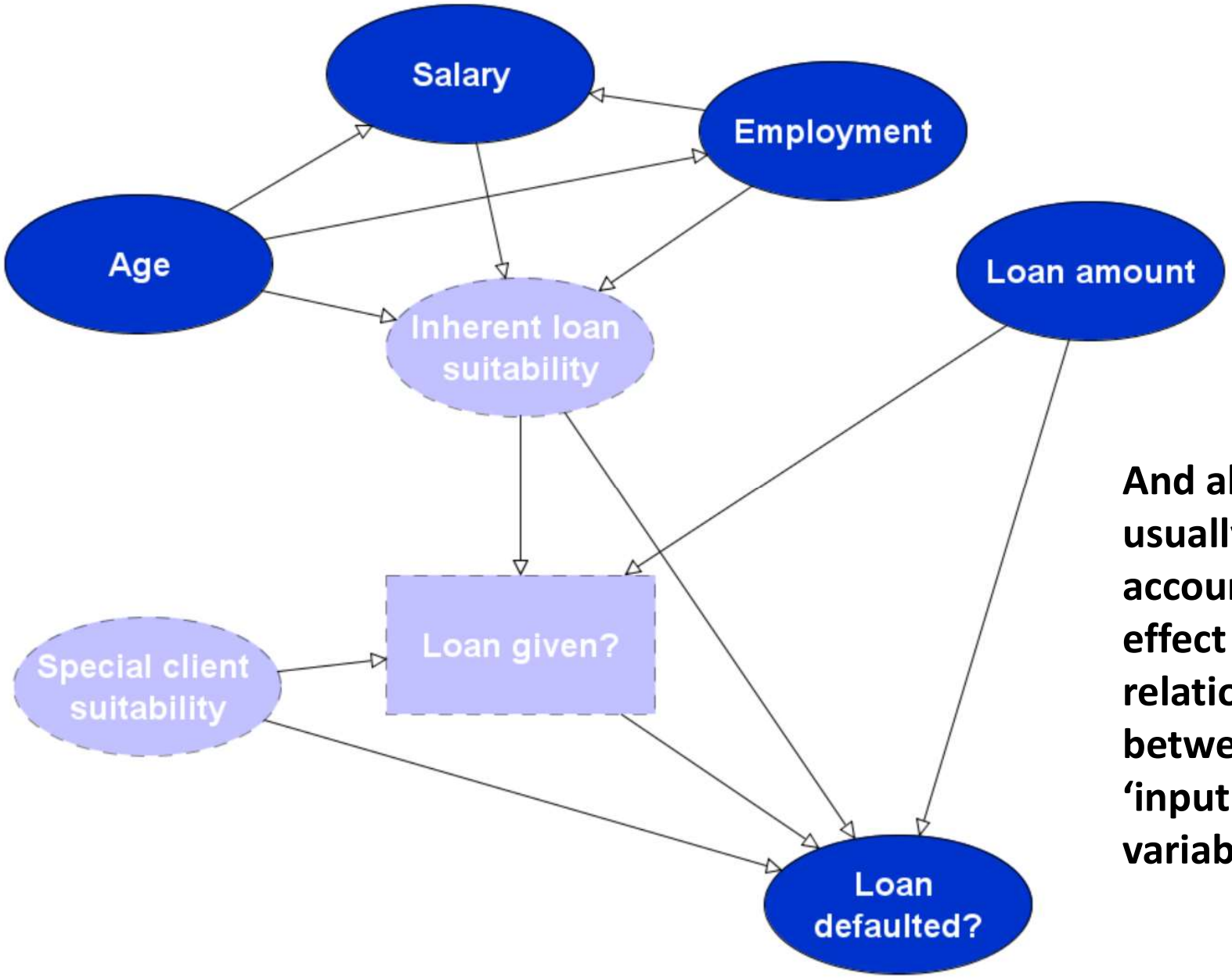
Because of special case interventions algorithms will ‘learn’ really dumb things about limited data cases



Statistical machine learnt model will 'learn' the target outcome as a function only of observable data

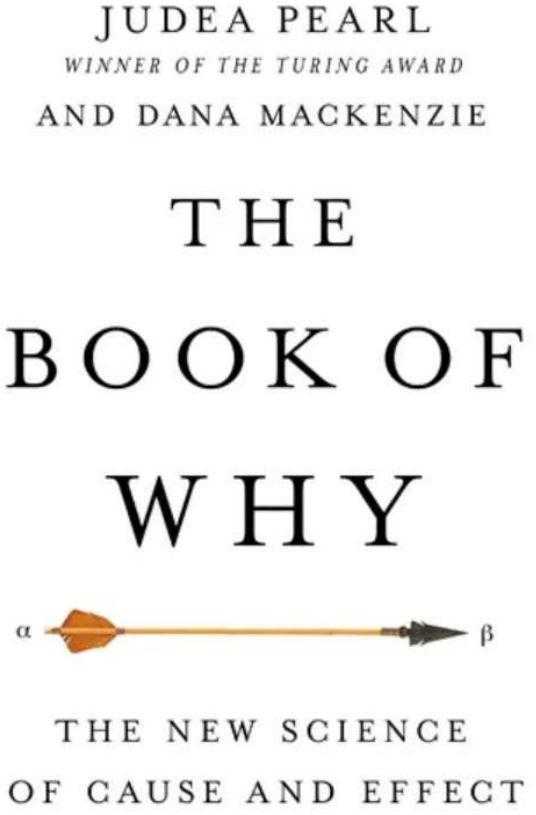


Machine learnt models fail to incorporate variables not in the data



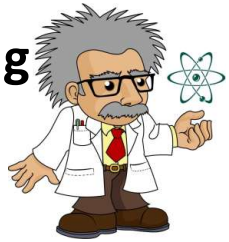
And also fails usually to account for the effect of relationships between the 'input' variables

Pearl's ladder of causation



Pearl's ladder of causation

Imagining



Doing

Seeing

<u>Counterfactuals</u>: "What if I had done…"
If I hadn't taken this drug would my headache still have stopped?

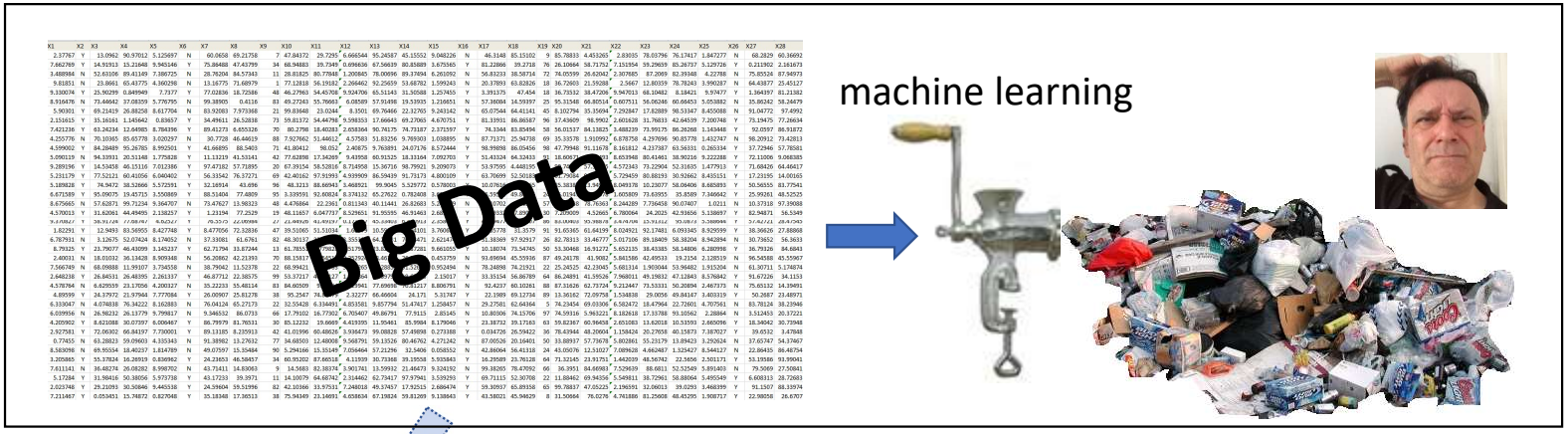
<u>Intervention</u>: "What if I do…"
If I take this drug will it stop my headache?

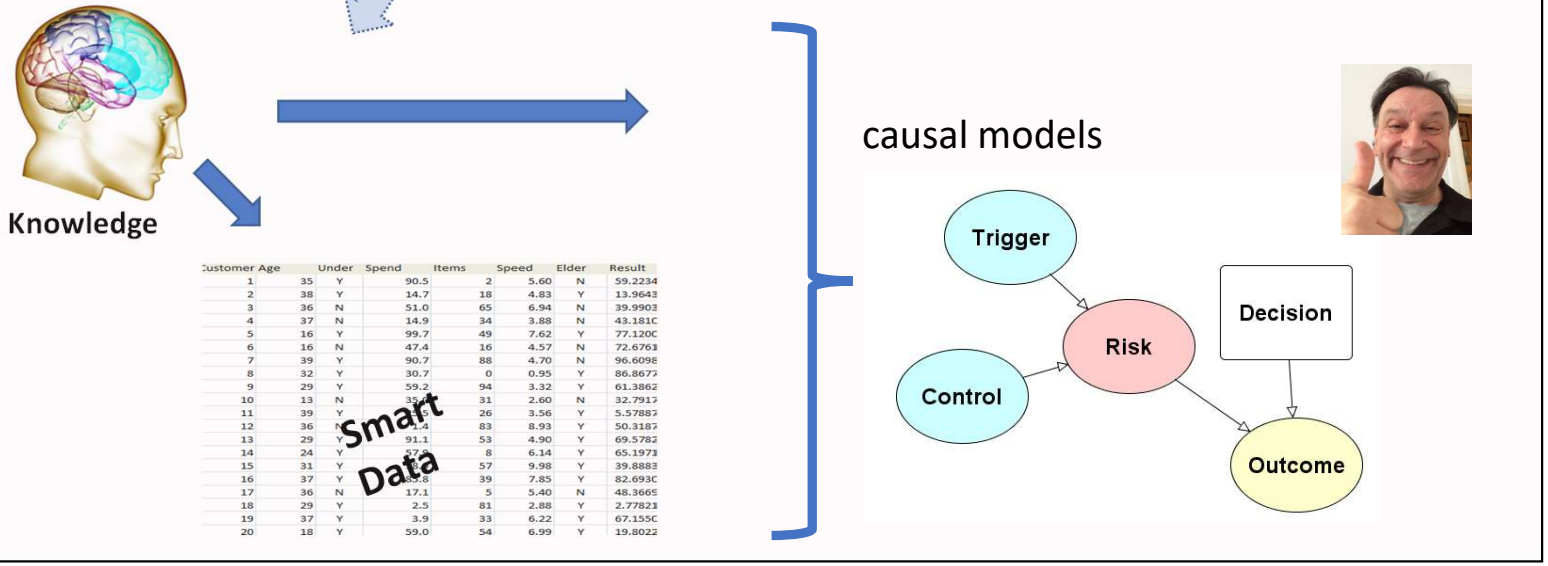
<u>Association</u>: "What if I see…"
From trials data is this drug effective at stopping headaches?

'Standard' statistical methods and machine learning can ONLY really answer questions of association

In this module you will learn how to get to levels 2 and 3 using causal BNs.

Big Data … or Smart Data?





Workshop questions

1. What is the probability that in a class of 23 people at least 2 will share the same birthday? What about in a class of 40?
2. If you select a Facebook friend at random, the probability that you have at least as many friends as that person is at about 50%. True or False (and why)?
3. Calculate the exact answer to the COVID test problem. What assumptions are you making?
4. For the study on sleep and stroke risk draw a ‘causal diagram’ that ‘explains’ the results.