Statistical Thinking for Data Science

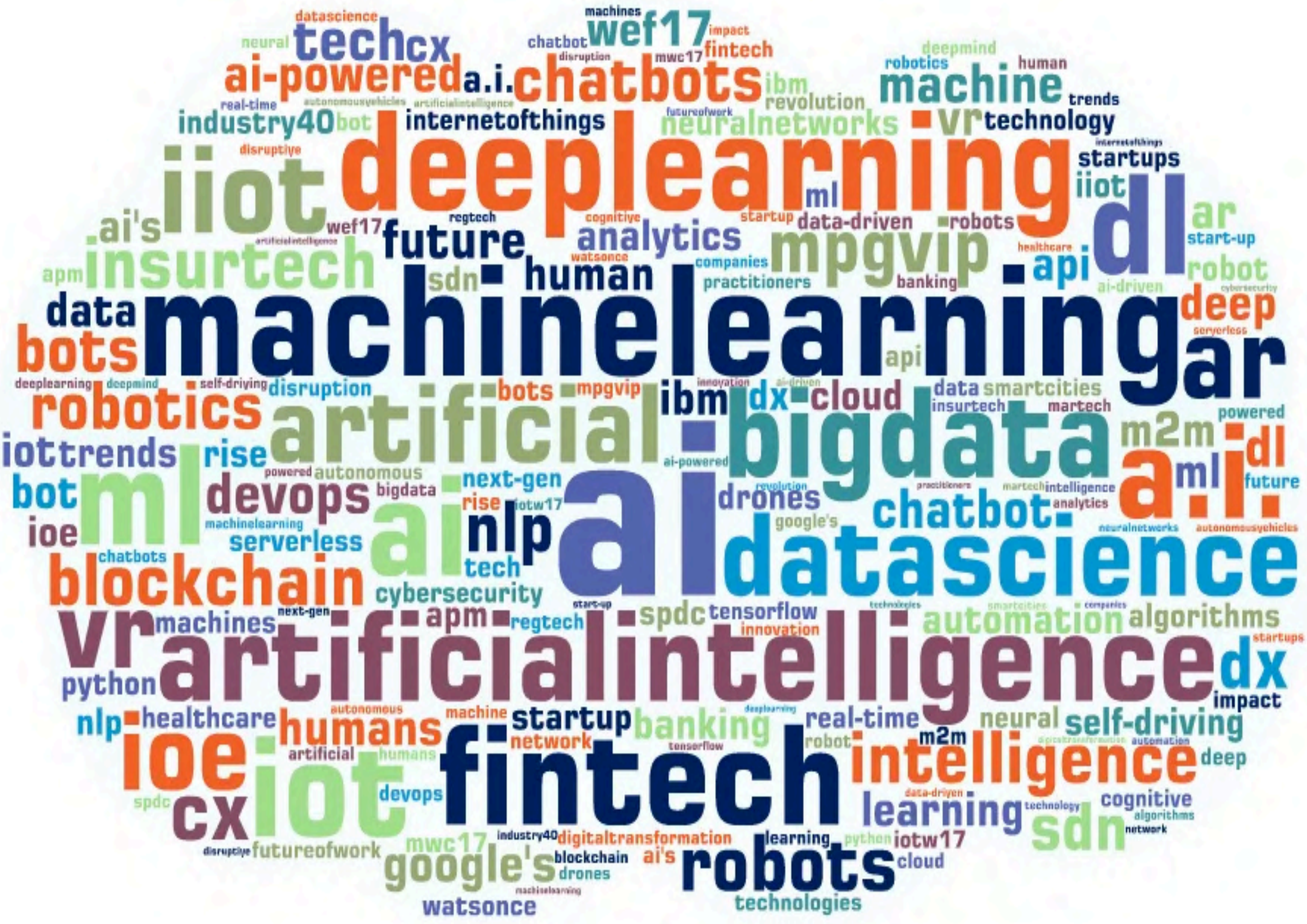
Mehul Motani
Electrical & Computer Engineering
National University of Singapore
Email: motani@nus.edu.sg

Problem Solving

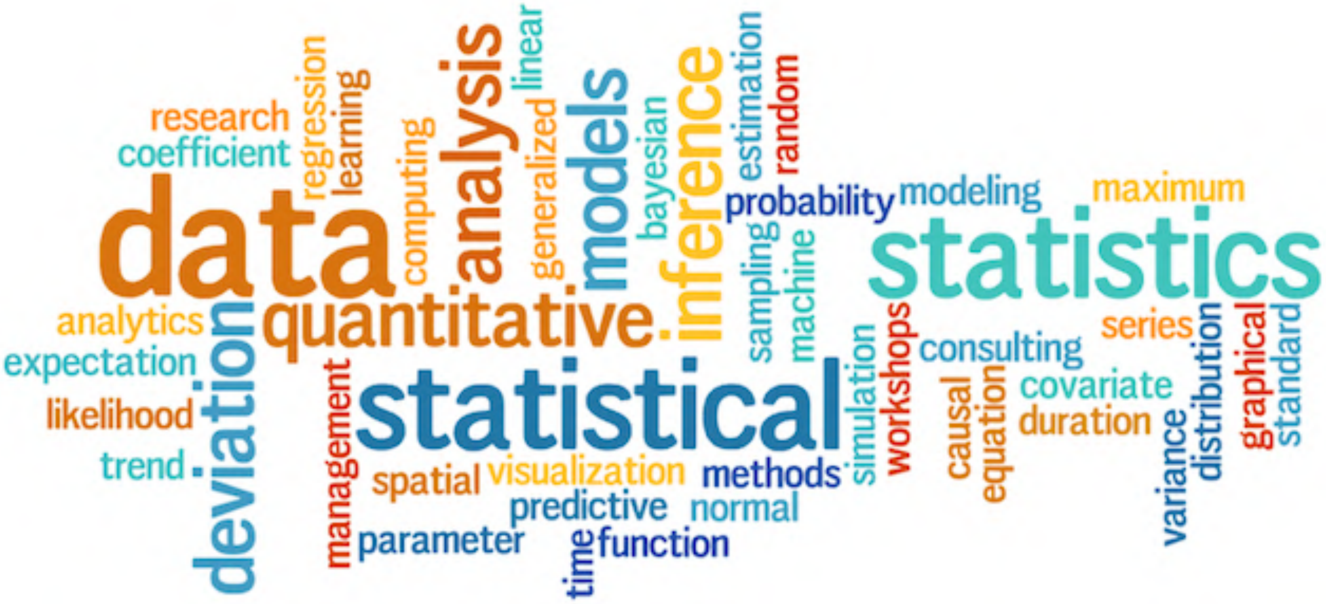
- Throughout this video, I will ask you to work out your own solutions to certain problems.
- What you will see is a message to pause the video and then a countdown timer before the solution is given.
- Please pause the video for the suggested amount of time and work out your own solution.
- Question: There is a patch of lily pads on a lake. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

00

Pause here for five minutes and work out the solution.

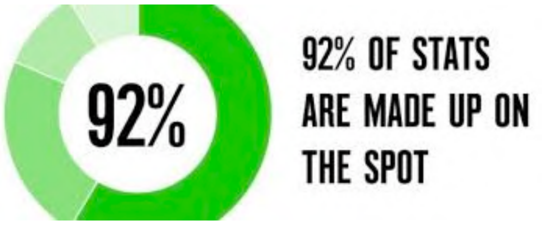


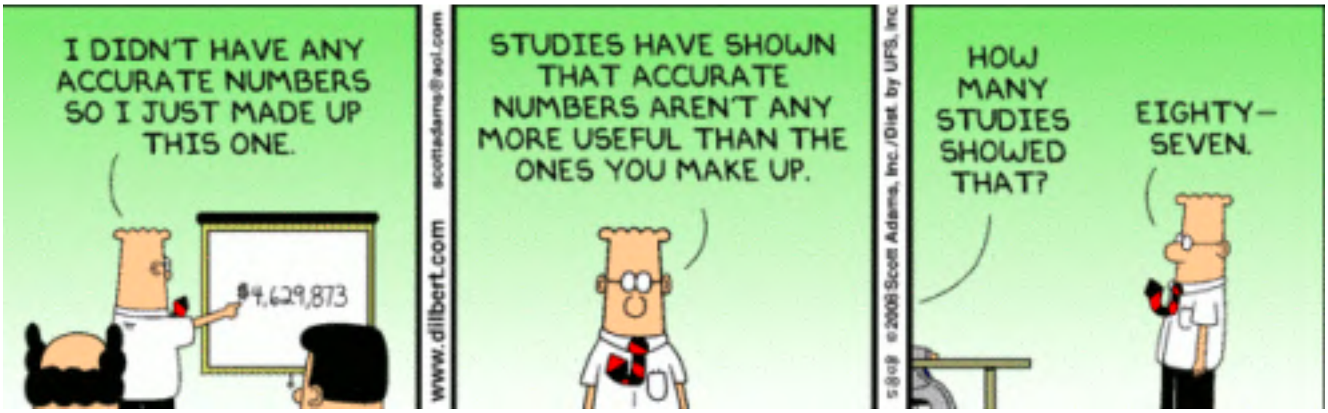
machinelearning
deeplearning
artificialintelligence
ai
bigdata
datascience
fintech
artificial
iiot
dl
ml
a.i.
iot
ioe
robots
blockchain
insurtech
mpgvip
chatbots
robotics
intelligence
tech
cx
ai-powered
a.i.
wef17
machine
technology
industry40
internetofthings
neuralnetworks
vr
startups
future
human
analytics
ai's
data
bots
ar
api
deep
robot
start-up
cloud
ibm
dx
drones
chatbot
nlp
devops
serverless
rise
bot
iottrends
python
healthcare
humans
startup
banking
real-time
self-driving
learning
sdn
cognitive
google's
watsonce
technologies
digitaltransformation
futureofwork
mwc17
iotw17
automation
algorithms
tensorflow
cybersecurity
apm
regtech
spdc
next-gen
m2m
impact
neural
network
autonomous
disruption
powered
practitioners
companies
banking
data-driven
startup
revolution
deepmind
fintech
impact
datascience
neural
chatbot
disruptive



data
statistics
statistical
quantitative
analysis
models
inference
research
coefficient
regression
learning
computing
generalized
linear
bayesian
estimation
random
probability
modeling
maximum
sampling
machine
analytics
expectation
likelihood
trend
deviation
management
spatial
visualization
methods
predictive
normal
parameter
time
function
simulation
workshops
consulting
series
covariate
duration
causal
equation
variance
distribution
graphical
standard

Would I lie to you?



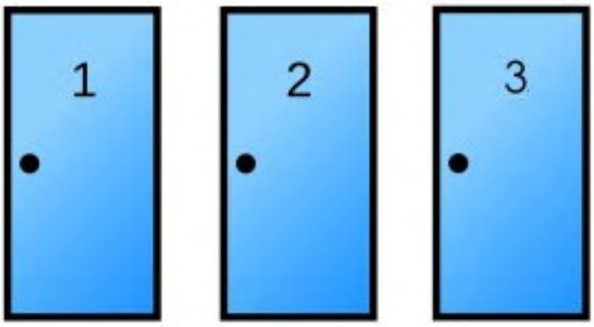

I DIDN'T HAVE ANY ACCURATE NUMBERS SO I JUST MADE UP THIS ONE.
$4,629,873
STUDIES HAVE SHOWN THAT ACCURATE NUMBERS AREN'T ANY MORE USEFUL THAN THE ONES YOU MAKE UP.
HOW MANY STUDIES SHOWED THAT?
EIGHTY-SEVEN.
www.dilbert.com
scottadams@aol.com
©2008 Scott Adams, Inc./Dist. by UFS, Inc.

Source: dilbert.com

A word of caution – Data and Statistics are tricky!

Let's Play a Game

Monty Hall Problem



- Three Doors
- One Prize, Two Empty
- Choose a door
- I show you an empty door
- **Do you switch your choice?**

00

Pause here for five minutes and work out the solution.

Monty Hall Problem

- Solution: You should switch your choice!
- On your first guess, you are right 1/3 of the time and wrong 2/3 of the time.
- You know the host will open an empty door. By switching, you are actually choosing the other two doors.
- So by switching, your chances of choosing the winning door are 2/3.
- https://en.wikipedia.org/wiki/Monty_Hall_problem

Treating Kidney Stones

- **Kidney stones** (nephrolithiasis) are hard deposits made of minerals and salts that form inside your kidneys.

	Success	**Failure**
Treatment A	273	77
Treatment B	289	61

Which treatment is better at removing kidney stones?

1. Treatment A
2. Treatment B

00

R. Charig, D. R. Webb, S. R. Payne, O. E. Wickham (1986). "Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy". *Br Med J (Clin Res Ed)* **292** (6524): 879–882

Treating Kidney Stones

Small stones	Success	Failure
Treatment A	81	6
Treatment B	234	36

Large Stones	Success	Failure
Treatment A	192	71
Treatment B	55	25

Which treatment is better at removing ***small*** kidney stones?

1. Treatment A
2. Treatment B

Which treatment is better at removing ***large*** kidney stones?

1. Treatment A
2. Treatment B

- Treatment A is more effective for all kidney stones, but the data shows Treatment B to be effective overall!
- How is this possible?

00

Kidney Stones – Simpson’s Paradox

- Paradox was described by Edward Simpson in 1951.

ALL STONES	Success	Failure	Success Rate
Treatment A	273	77	78%
Treatment B	289	61	83%

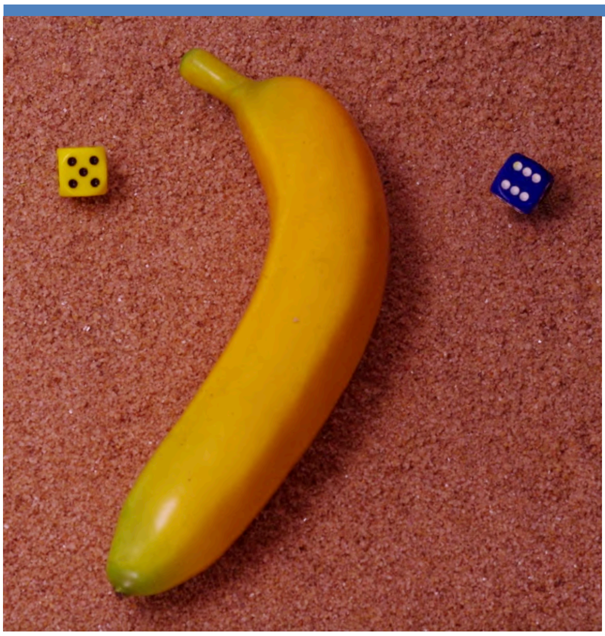
Small Stones	Success	Failure	Success Rate
Treatment A	81	6	93%
Treatment B	234	36	87%

Large Stones	Success	Failure	Success Rate
Treatment A	192	71	73%
Treatment B	55	25	69%

Kidney Stones – Simpson's Paradox

- Treatment A is used more often on large stones, which are harder to treat.
- Treatment B is used to more often on small stones which are easier to treat.
- When the less effective treatment (B) is applied more frequently to less severe cases, it can appear to be a more effective treatment.
- This is an example of ***Simpson's Paradox***: an observed relationship between two variables can change (or even reverse!) when a third variable is considered
- Key problem: Sample sizes are imbalanced
- Solution: Work with equal sample sizes
- Resolving the paradox once and for all: http://ftp.cs.ucla.edu/pub/stat_ser/r414.pdf

Another game: The last banana



- Imagine a game played with two players and two dice
  - If the bigger number rolled is one, two, three, or four, player 1 wins.
  - If the bigger number rolled is five or six, player 2 wins.
  - The winner gets the last banana!
  - Do you want to be player 1 or player 2?

00

Pause here for five minutes and work out the solution.

The Last Banana

- Solution: You want to be Player 2!
- There are a total of 36 outcomes of the roll of two dice.
- Player 1 wins if the the highest number is 1/2/3/4
- Hence Player 1 wins in 16 of the 36 outcomes.
- Hence Player 2 wins in 20 of the 36 outcomes.
- So Player 2 has a 5/9 chance of winning
- While Player 1 has a 4/9 chance of winning.
- https://youtu.be/Kgudt4PXs28

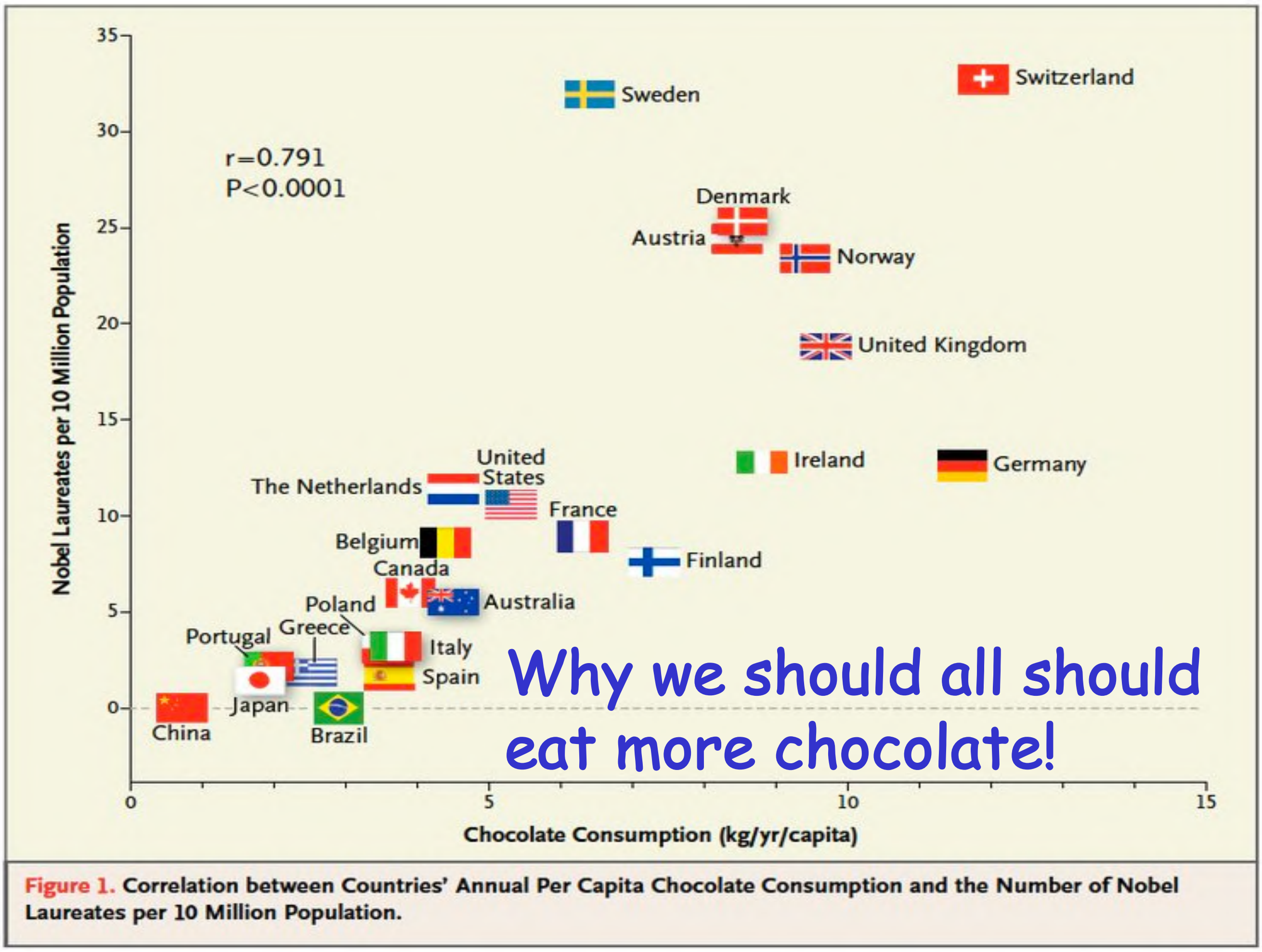
Trivia Question

Ice-cream sales are strongly correlated with death from drowning rates.

Therefore, ice-cream causes drowning.

Do you agree?

00



Why we should all should eat more chocolate!

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

Chocolate Consumption, Cognitive Function, and Nobel Laureates by Franz H. Messerli, M.D., New England Journal of Medicine, Oct 2012

- Since chocolate consumption could hypothetically improve cognitive function not only in individuals but also in whole populations, I wondered whether there would be a correlation between a country's level of chocolate consumption and its population's cognitive function.
- To my knowledge, no data on overall national cognitive function are publicly available. Conceivably, however, the total number of Nobel laureates per capita could serve as a surrogate end point reflecting the proportion with superior cognitive function and thereby give us some measure of the overall cognitive function of a given country.
- The principal finding of this study is a surprisingly powerful correlation between chocolate intake per capita and the number of Nobel laureates in various countries.
- Of course, a correlation between X and Y does not prove causation but indicates that either X influences Y, Y influences X, or X and Y are influenced by a common underlying mechanism.
- However, since chocolate consumption has been documented to improve cognitive function, it seems most likely that in a dose-dependent way, chocolate intake provides the abundant fertile ground needed for the sprouting of Nobel laureates. Obviously, these findings are hypothesis-generating only and will have to be tested in a prospective, randomized trial.

Does correlation imply causation?

In a Gallup poll, surveyors asked, “Do you believe correlation implies causation?’”

- 64% of Americans answered “Yes” .
- 38% replied “No”.
- The other 8% were undecided.

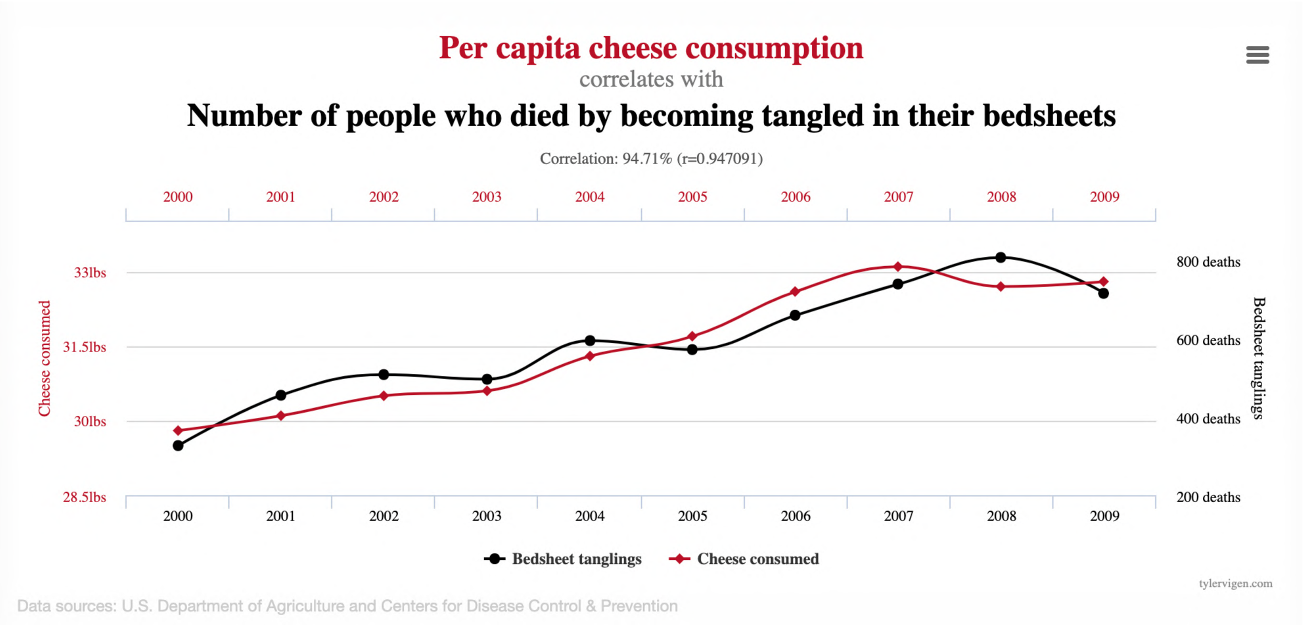
Source: dilbert.com

SOMEONE SENT ME ANOTHER ANONYMOUS EMAIL WITH A LINK TO AN ARTICLE ABOUT THE WORLD'S WORST BOSSES.
Dilbert.com DilbertCartoonist@gmail.com
I GET ONE OF THOSE EMAILS EVERY TIME I LEAVE YOUR CUBICLE. DID YOU THINK I WOULDN'T NOTICE THE CORRELATION?
11-28-11 ©2011 Scott Adams, Inc./Dist. by Universal Uclick
CORRELATION DOES NOT IMPLY CAUSATION.
SEND

Correlation vs. Causation

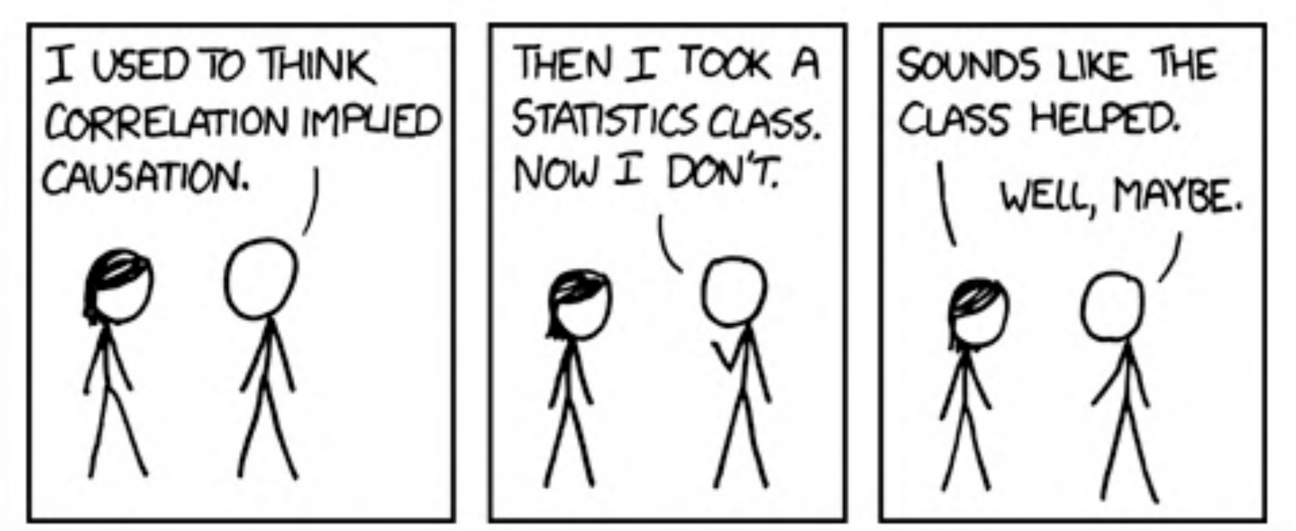
- Correlation tells us two variables are related
- Types of relationship reflected in correlation
  - X causes Y or Y causes X (causal relationship)
  - X and Y are caused by a third variable Z (spurious relationship)
- **Correlation does not imply causation!**
- In order to imply causation, a scientific experiment must be performed where subjects are randomly assigned to different conditions.
- Spurious correlations: http://www.tylervigen.com/spurious-correlations

Spurious Correlations



Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

Correlation does not imply Causation


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Source: xkcd.com

00

The failure of predictive models

What happened?

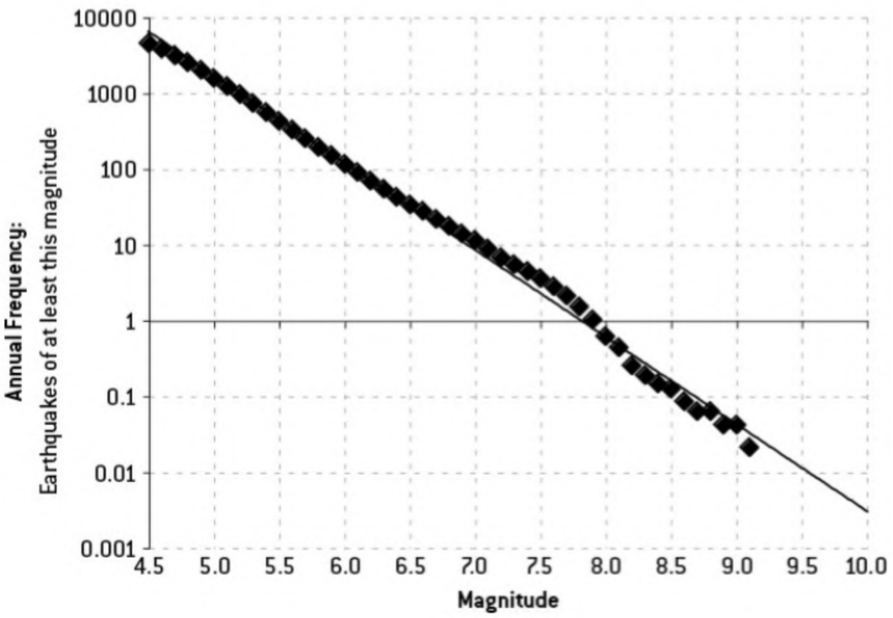
- On 11 March 2011, a magnitude 9 earthquake in the Tohoku region in Japan caused a 15 meter high tsunami.
- A magnitude 9.0 earthquake = 99 million tons of TNT explosives
- A magnitude 9.0 earthquake = 25000 nuclear bombs
- 15 meters = 3 times as tall as a Giraffe
- The tsunami destroyed most of the electrical generators (including the backups) which were used for cooling.
- Unable to cool itself, the nuclear reactors melted down.
- The damage costs of this disaster are estimated to be $150 billion USD.

Why did it happen?

- The safety analysis for the Fukushima Daiichi nuclear power plant was based on historical data dating back over 400 years.
- The plant was designed to withstand a maximum earthquake of 8.6 magnitude, and a tsunami as high as 5.7 meters.
- The earthquake on March 11, 2011 measured 9.0 and resulted in a 15 meter high tsunami.
- The design basis was developed from a mistake in the regression analysis of the historical earthquake data.

Predicting Earthquakes Worldwide: Gutenberg-Richter Model

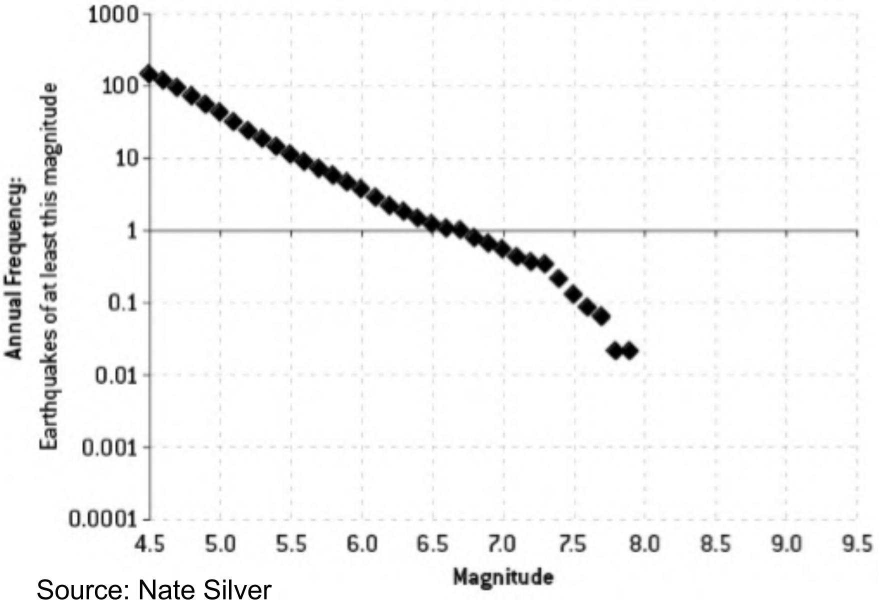
FIGURE 5-3B: WORLDWIDE EARTHQUAKE FREQUENCIES, JANUARY 1964–MARCH 2012, LOGARITHMIC SCALE



Source: Nate Silver

Earthquakes in Japan – The Data

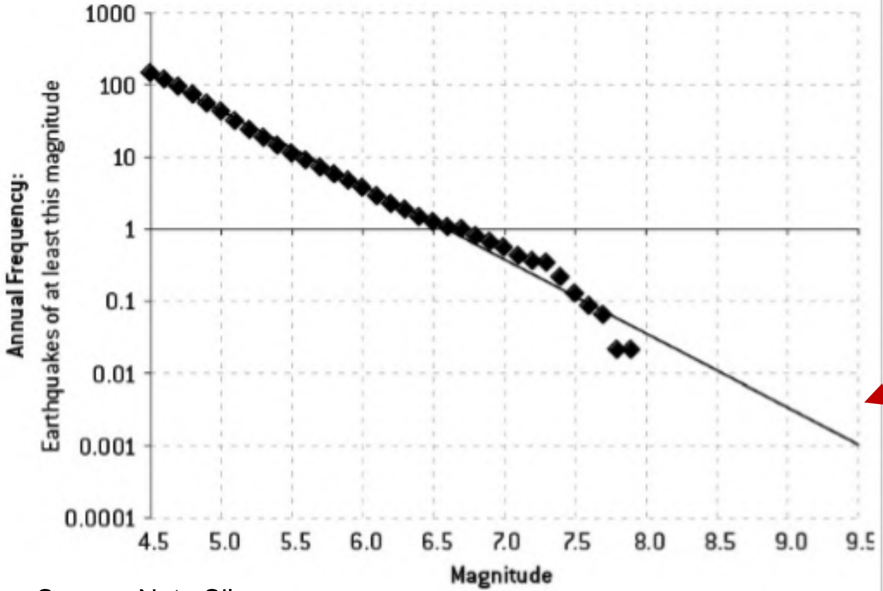
FIGURE 5-7A: TŌHOKU, JAPAN EARTHQUAKE FREQUENCIES
JANUARY 1, 1964–MARCH 10, 2011



Source: Nate Silver

Earthquakes in Japan – Gutenberg-Richter Model

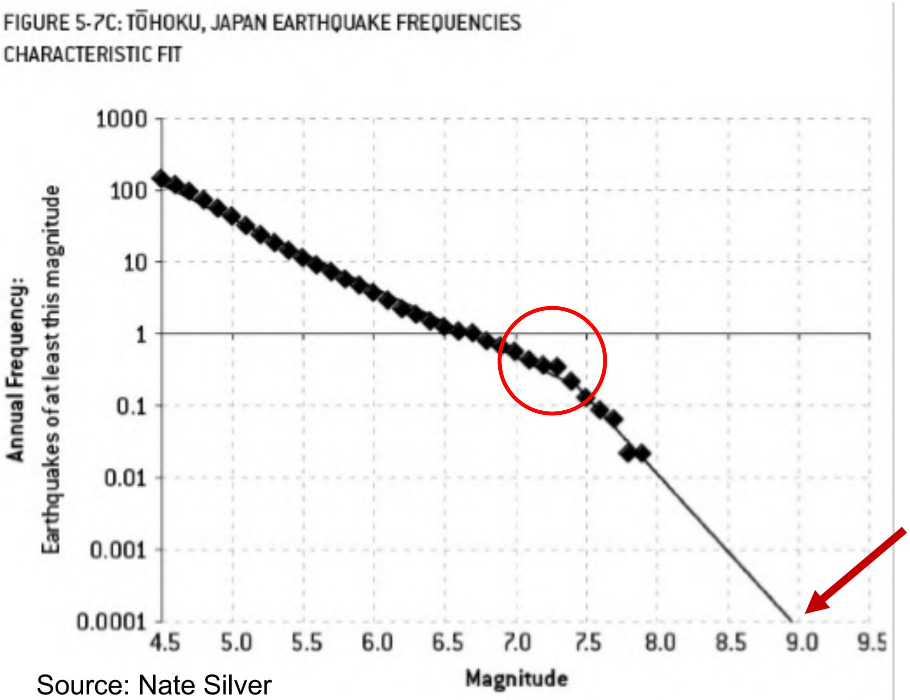
FIGURE 5-7B: TŌHOKU, JAPAN EARTHQUAKE FREQUENCIES
GUTENBERG-RICHTER FIT



7 in 1000 chance

Source: Nate Silver

Earthquakes in Japan – Characteristic Fit

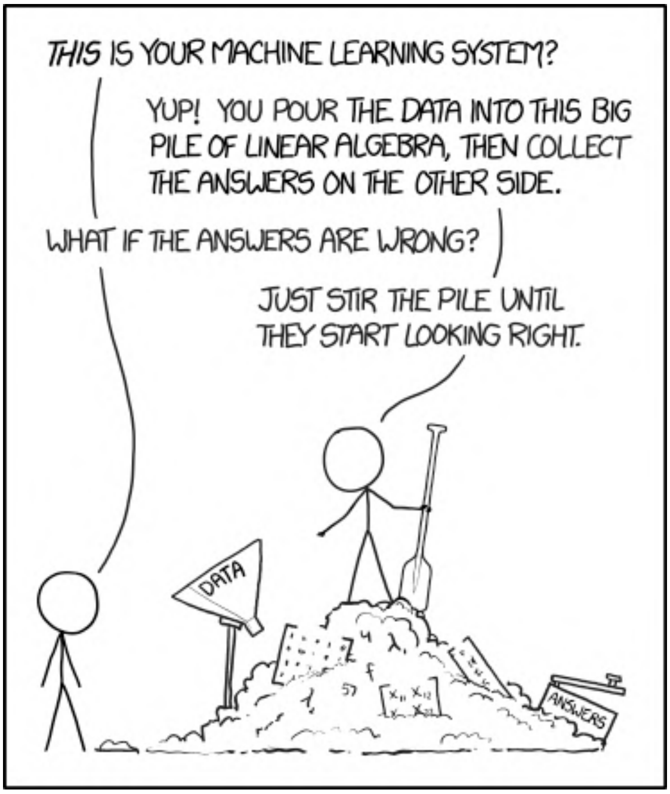


Source: Nate Silver

A Magnitude 9.0 earthquake is about 100 times more likely that the model that was used!

Less than 1 in 10000 chance

A picture is worth a thousand words!


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Source: xkcd.com

00

Puzzle: To bet or not to bet

- I choose two integers at random and hold them in my left and right hands.
- You choose one hand at random, and I show you the number.
- Now you must decide if you want to keep the number or change hands.
- I keep the number you don't keep.
- The person with the larger number wins. The loser pays the winner $100.
- **Would you take this bet?**

- It turns out this is a bet worth taking.
- There is a strategy with which you can win more than 50% of the time.
- Can you find such a strategy?

Thank you!

- Please send me your feedback and any questions you may have.
- The best way to contact me is via email:

  mehul.motani@gmail.com
- Thanks for listening!