SRCASW, University of Delhi

"Fooled" by Statistics & ML

STATISTICS + MACHINE LEARNING + DATA SCIENCE

Dr. Tanujit Chakraborty
Ph.D from Indian Statistical Institute, Kolkata, India.
Assistant Professor of Statistics at Sorbonne University
tanujitisi@gmail.com | https://www.ctanujit.org

Normality is a myth!

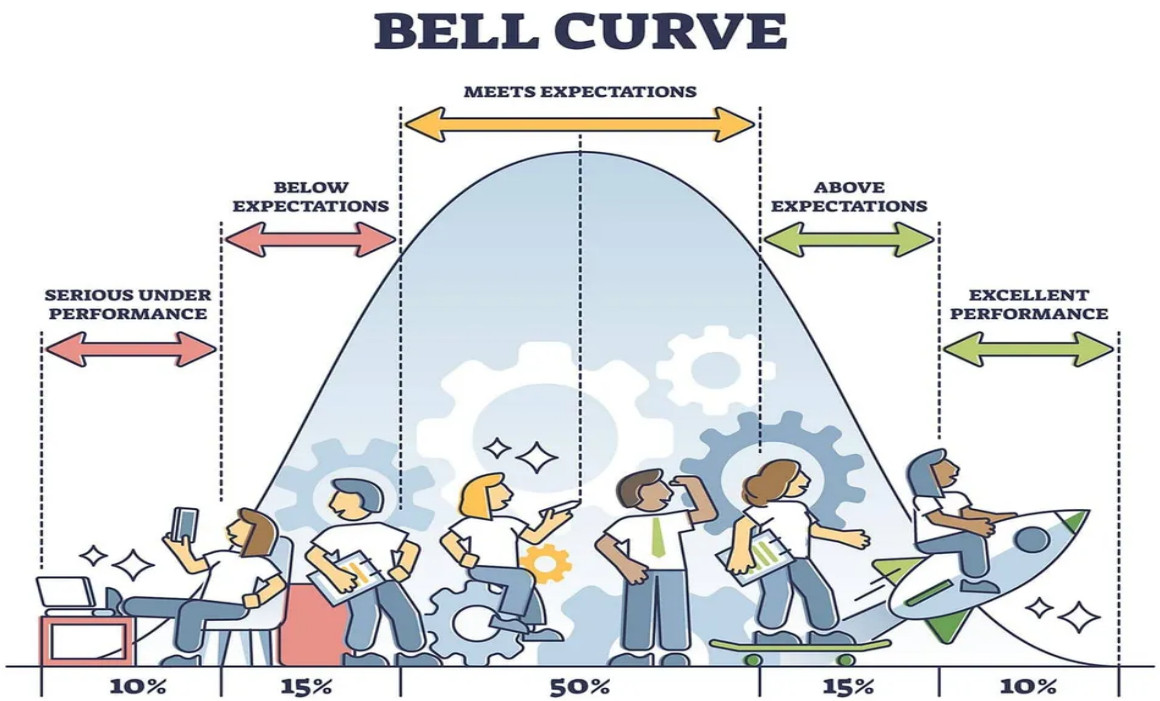
Correlation does not imply causation!

All models are wrong, but some are useful!

Normality is a myth!

Normality & Beyond Normality

Normality is a paved road. It is easy to walk but no flowers grow on it. — Vincent Van Gogh.



By Dr. Saul McLeod (2019)

Few Famous Quotations

Normality is a myth; there never was, and there never will be a normal distribution — Roy C. Geary (1947; Biometrika, vol. 34, 248).

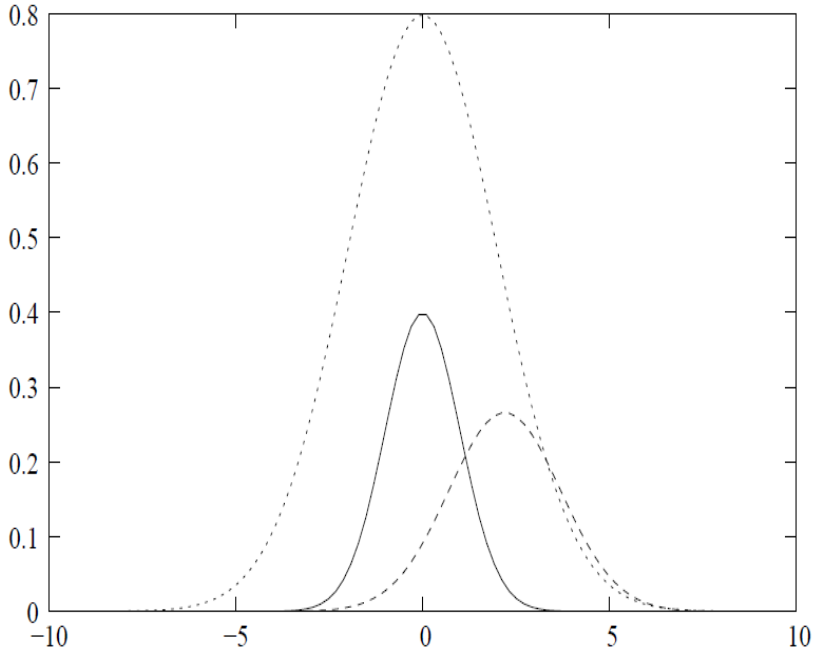
Everybody believes in the exponential law of errors (the normal distribution), the experimenters, because they think it can be proved by mathematicians; and the mathematicians, because they believe that it has been established by observations — E.T. Whittaker and G. Robinson (1967).

... the statisticians knows ... that in nature there never was a normal distribution, there never was a straight line, yet with normal and linear assumptions, known to be false he can often derive results which match to a useful approximation, those found in real world — George W. Box (1976, Journal of American Statistical Association, vol. 71, 791-799).

Normal Distribution

A random variable X is said to be normally distributed with mean μ and variance σ^2, if the probability density function of X is the following (for $-\infty < \mu < \infty$ and $\sigma > 0$)

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}; \ -\infty < x < \infty$$

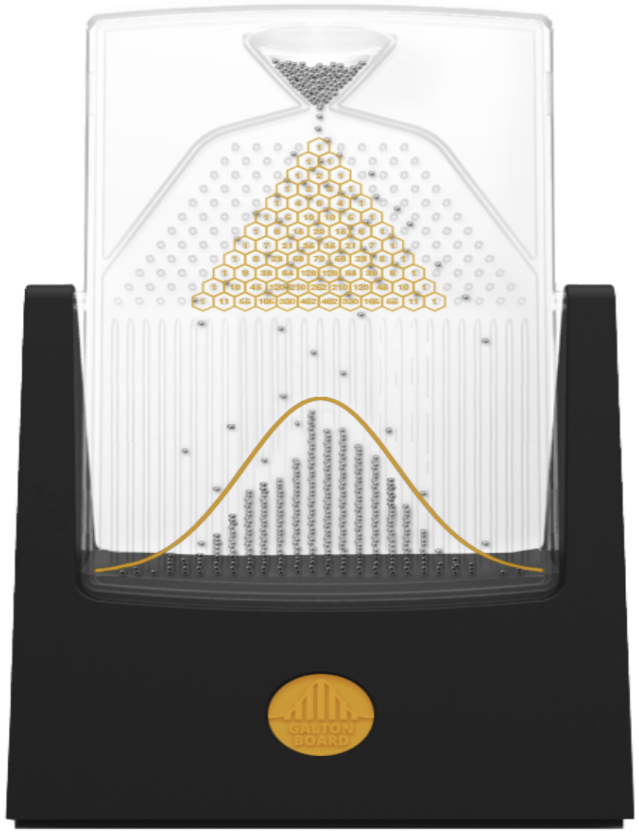


Probability Density Function of Normals

Galton Board

- Sir Francis Galton, Charles Darwin's half-cousin, invented the 'Galton Board' in 1874 to demonstrate that the normal distribution is a natural phenomenon.

- It specifically shows that the binomial distribution approximates a normal distribution with a large enough sample size.



Picture of Galton Board

How it has started?

Gambling Question: *A 17th century gambler, the Chevalier de Mere, asked Pascal for an explanation of his unexpected losses in gambling.*

The famous correspondence between Pascal and Fermat was instigated in 1654, and they were mainly interested to calculate the following binomial sum:

$$\sum_{k=i}^{j} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$$

The problem was not difficult when n is small.

A Brief History

Within few years the following problem arises in a sociological study, where the following computation was necessary:
$n = 11,429, i = 5745, j = 6128$

$$\sum_{k=i}^{j} \binom{n}{k} p^k (1-p)^{n-k}$$

Original Problem: *The problem is to test the hypothesis that male and female births are equally likely against the actual birth in London over 82 years from 1629 - 1710. It is observed that the relative number of male births varies from a low of* $7765/15,448 = 0.5027$ *in 1703 to a high of* $4748/8855 = 0.5362$ *in 1661. Given that 11,429 is the average number of births in London over 82 years, and 5745 and 6128 are two limits.*

Solution

Using the following recurrence relation

$$\binom{n}{x+1} = \binom{n}{x}\binom{n-x}{x+1}$$

and some involved rational approximation it has been obtained

$$P(5747 \leq X \leq 6128 \mid p = 1/2) = \sum_{i=5745}^{6128} \binom{11,429}{i} \left(\frac{1}{2}\right)^i$$
$$\approx 0.292$$

Solution

Using the following recurrence relation

$$\begin{pmatrix} n \\ x+1 \end{pmatrix} = \begin{pmatrix} n \\ x \end{pmatrix} \begin{pmatrix} n-x \\ x+1 \end{pmatrix}$$

and some involved rational approximation it has been obtained

$$\begin{aligned} P(5747 \leq X \leq 6128 \mid p = 1/2) &= \sum_{i=5745}^{6128} \begin{pmatrix} 11,429 \\ i \end{pmatrix} \left(\frac{1}{2}\right)^i \\ &\approx 0.292 \end{aligned}$$

The Breakthrough

De Moivre began the search for this approximation in 1721 , and in 1733 it has been proved that

$$\binom{n}{\frac{n}{2}+x}\left(\frac{1}{2}\right)^n \approx \frac{2}{\sqrt{2\pi n}} e^{-2x^2/n}$$

and

$$\sum_{|x-n/2|\leq a} \binom{n}{x}\left(\frac{1}{2}\right)^n \approx \frac{4}{\sqrt{2\pi}} \int_0^{a/\sqrt{n}} e^{-2y^2} dy.$$

Normal Approximation

Eventually using the second approximation one gets

$$\sum_{k=i}^{j} \binom{n}{k} p^k (1-p)^k \approx \Phi\left(\frac{j-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{i-np}{\sqrt{np(1-p)}}\right)$$

where

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx$$

which is the cumulative distribution function (CDF) of the standard normal distribution.

Error Modeling

Gauss (1809) made the following assumptions and deduce the normal distribution as an error distribution:

1. Small errors are more likely than large errors.
2. For any real numbers ϵ, the likelihood of errors of magnitudes ϵ and $-\epsilon$ are equal.
3. In the presence of several measurements of the same quantity, the most likely value of the quantity being measured is their average.

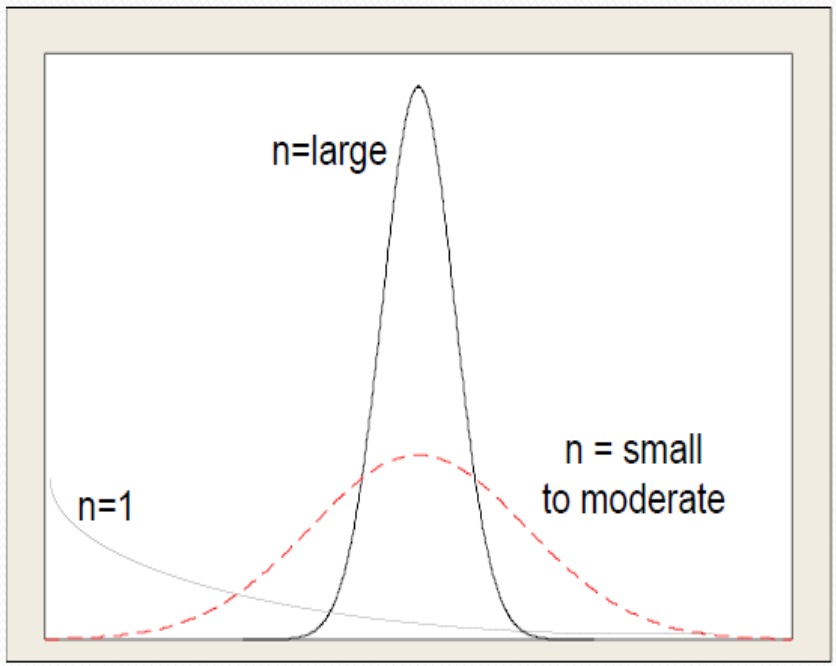
To read more about the evolution of normal distribution: Saul Stahl (2006), "The evolution of normal distribution", Mathematics Magazine, vol. 79, no. 2, 96 - 113.

Central Limit Theorem

Lindeberg-Levy CLT:

Suppose $\{X_1, X_2, \cdots\}$ is a sequence of independent identically distributed random variables with mean μ and variance $\sigma^2 < \infty$, then as $n \to \infty$

$$\frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \to N(0,1)$$



CLT in Practice

Some Drawbacks

What will happen if the data indicate that the parent distribution

1. is not symmetric?
2. is heavy tail?
3. is not unimodal?

What will happen if error distribution is not normal during regression modeling?

Possible Remedy

In Distribution Theory:

1. Skew Normal Distribution (A Azzalini, Scandinavian Journal of Statistics 1985)
2. Power Normal Distribution (RD Gupta, Test 2008)
3. Geometric Skew-Normal Distribution (D Kundu, Sankhya 2014), etc.

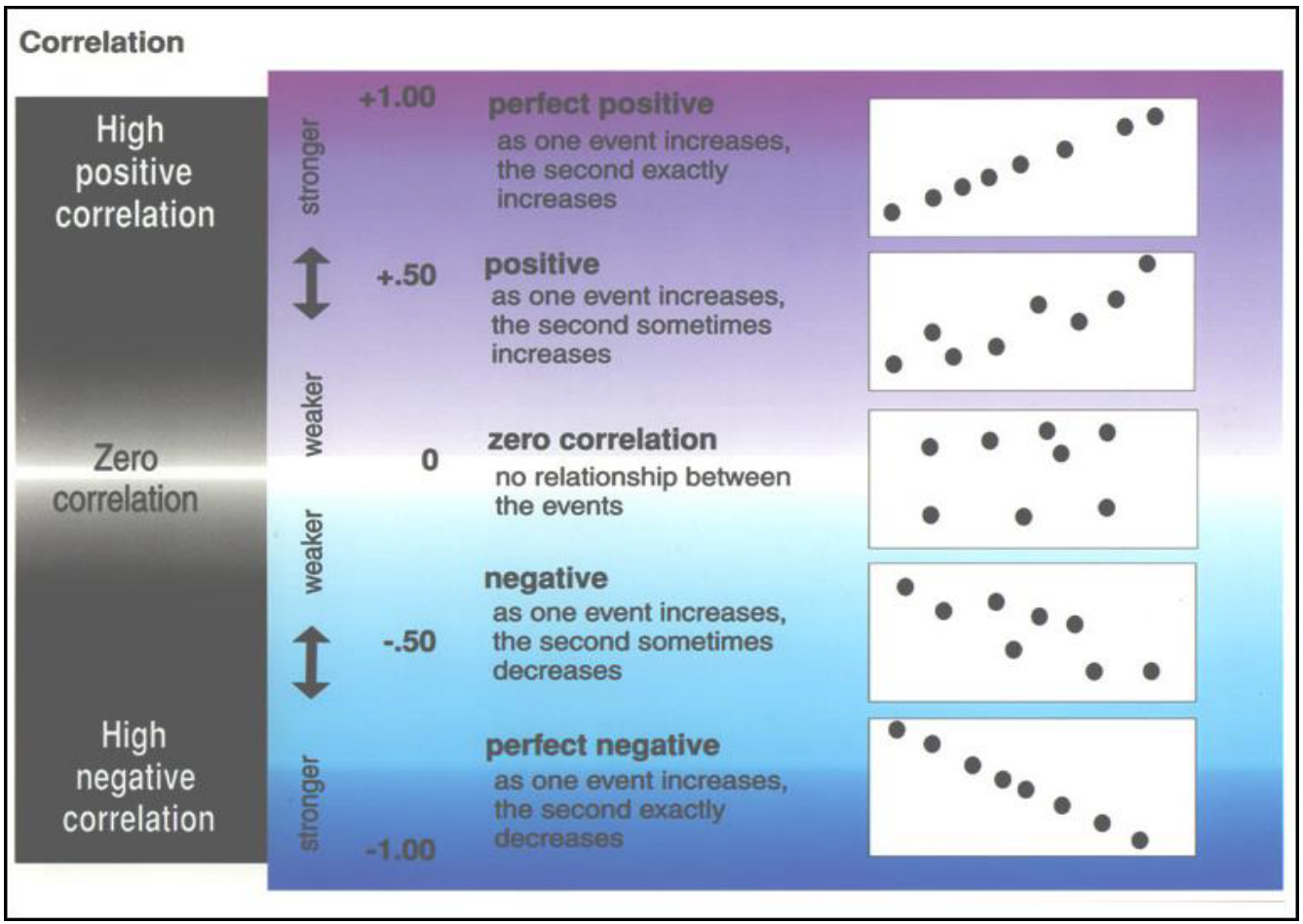
In Regression Theory:

1. Box-Cox Transformation (Box, Cox, JRSS Series-B 1964)
2. Generalized linear model (Nelder, Wedderburn, JRSS Series-A 1972)
3. Semiparametric and Nonparametric Approaches (see ESLR/ISLR Book), etc.

CORRELATION DOES NOT IMPLY CAUSATION!

Correlation

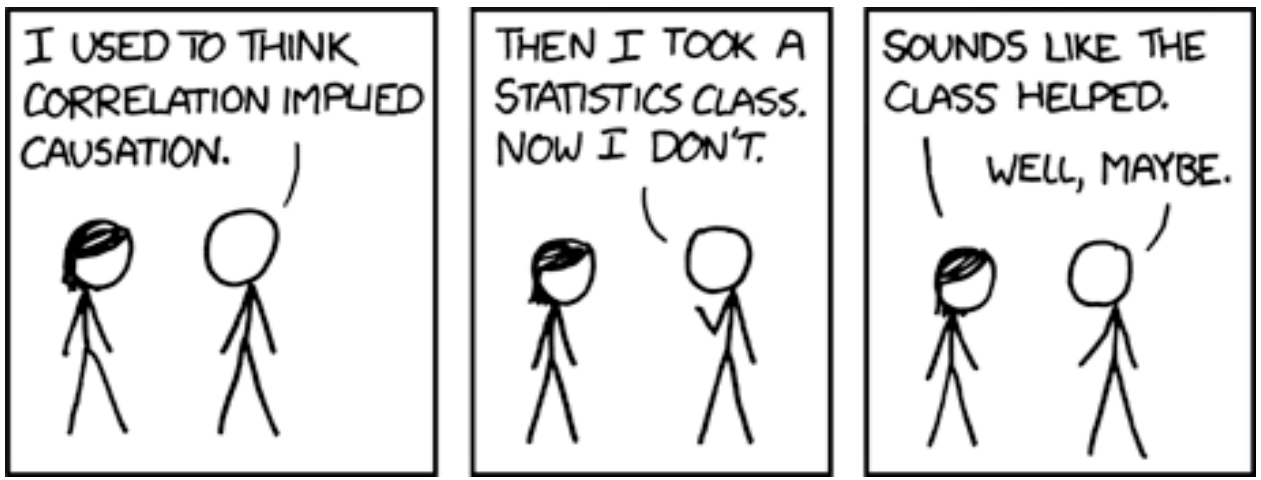
Correlation may indicate any type of association. Correlation implies association, but not causation. Conversely, causation implies association, but not correlation[1]



[1] Altman, Krzywinski, "Association, correlation and causation", Nature Methods (2015)

Causality: What is it?

Causality *is central notion in science, decision-taking and daily life.*
Causal inference $\approx$ *Causal language/model + Statistical inference.*

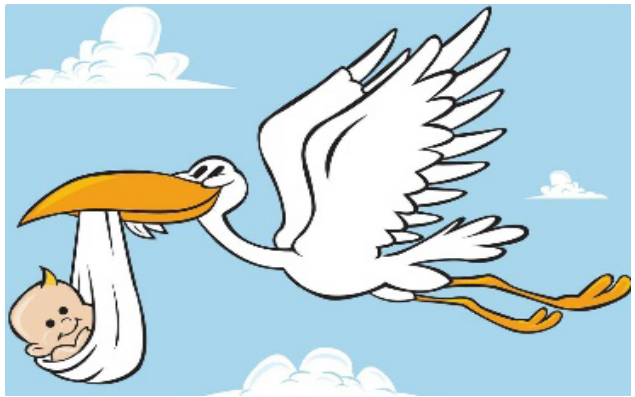


Question: How do you define cause and effect?

Causality in Philosophy

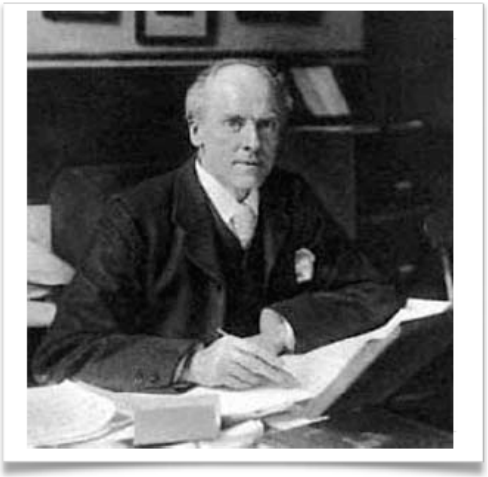
"...Thus we remember to have seen that species of object we call flame, and to have felt that species of sensation we call heat. We likewise call to mind their constant conjunction in all past instances. Without any farther ceremony, we call the one cause and the other effect, and infer the existence of the one from that of the other."

*- **David Hume**, A Treatise of Human Nature (1738).*



But: Does the stork really bring babies?

Causality in Statistics



"Beyond such discarded fundamentals as `matter' and `force' lies still another fetish amidst the inscrutable arcana of even modern science, namely, the category of cause and effect."

Karl Pearson (1857-1936)

Correlation does not imply causation.

Since then, many statisticians tried to avoid causal reasoning

- *"Considerations of causality should be treated as they have always been in statistics: preferably not at all."* (Terry Speed, 1990)
- *"It would be very healthy if more researchers abandon thinking of and using terms such as cause and effect."* (Bengt Muthen, 1987)

Causality in Statistics

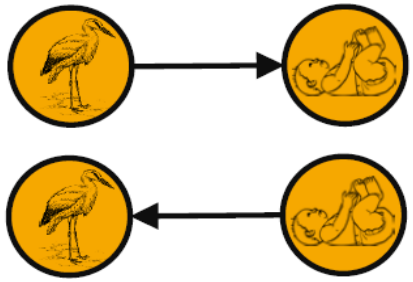
But dependence says us something about causation:

"Common Cause Principle"
Hans Reichenbach (1891-1953)

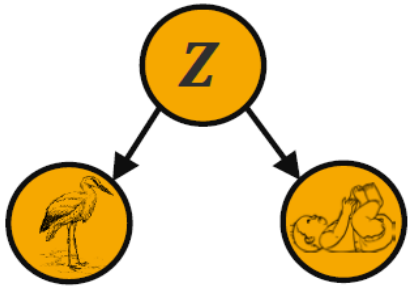
If there is a statistical dependence between variables X and Y, e.g.,

then either

- X causally influences Y (or vise versa), e.g.,



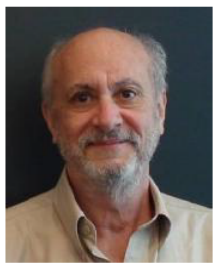
- or there exists Z causally influencing both, e.g.,



A Paradigm Shift: Basic Contributions

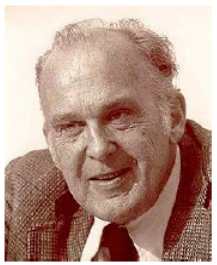
- The modeling of the underlying structures provides a language to encode causal relationships – the basis of a causality theory.
- Causality theory helps to decide when, and how, causation can be inferred from domain knowledge and data.

Some people who contributed to causality theories:



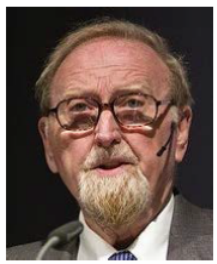
Donald Rubin (*1943)

Judea Pearl (*1936)



Donald Campbell (1916-1996)

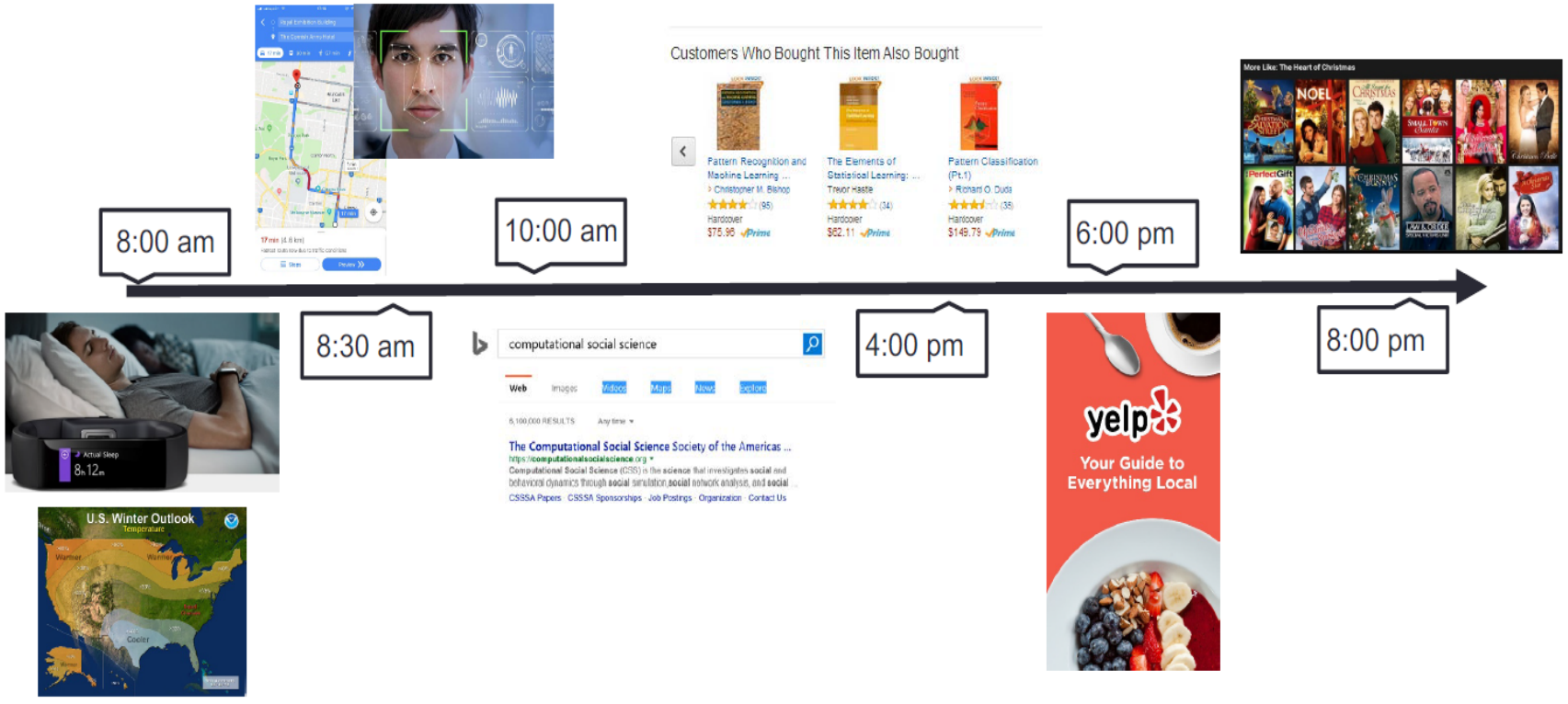
Dawid Philip (*1946)



Clive Granger (1934-2009)

ML techniques are impacting our life

A day in our life with Machine Learning techniques...



Now we are stepping into risk-sensitive

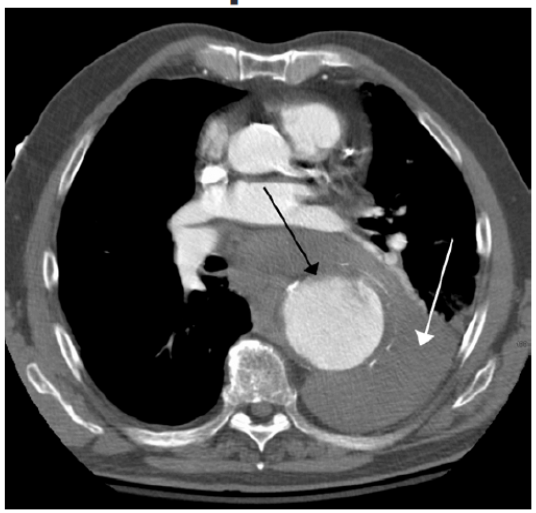
Shifting from Performance Driven to Risk Sensitive...

Human

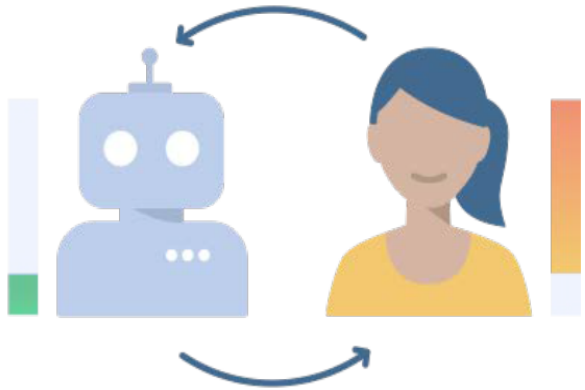
Problems of today's ML - Explainability

Most machine learning models are black-box models...

Unexplainable



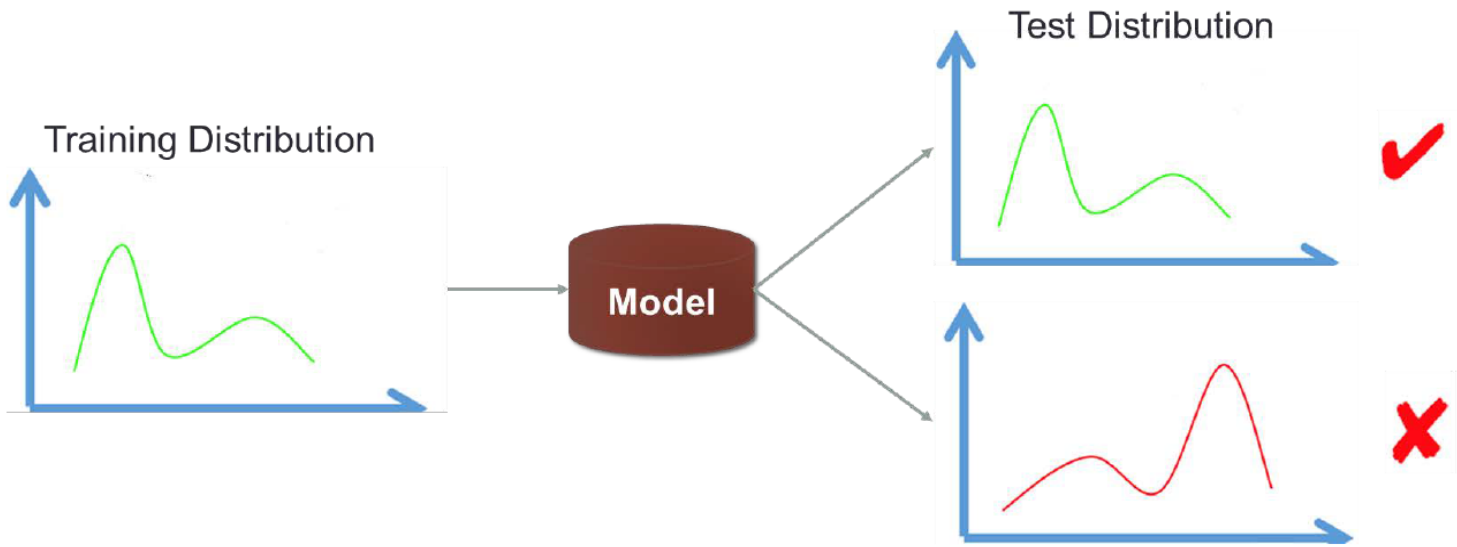
Human in the loop



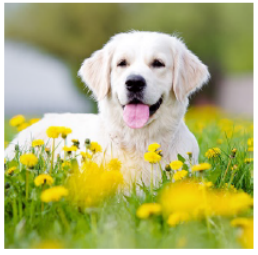
Health Military Finance Industry

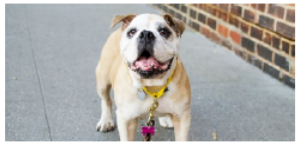
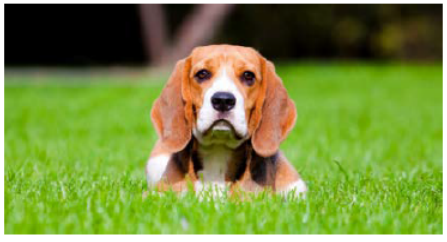
Problems of today's ML - Stability

Most ML methods are developed under I.I.D hypothesis...

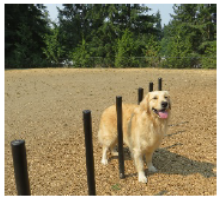


Problems of today's ML - Stability

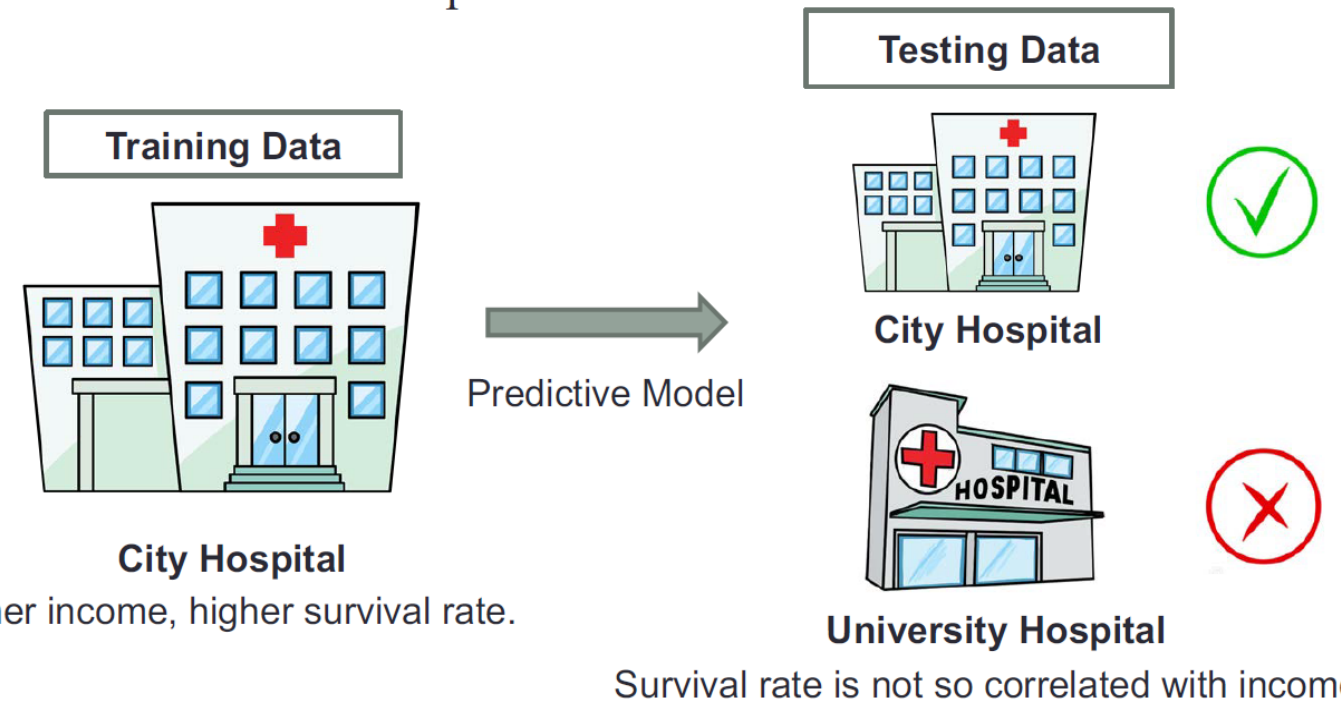

Yes


Maybe

No

Problems of today's ML - Stability

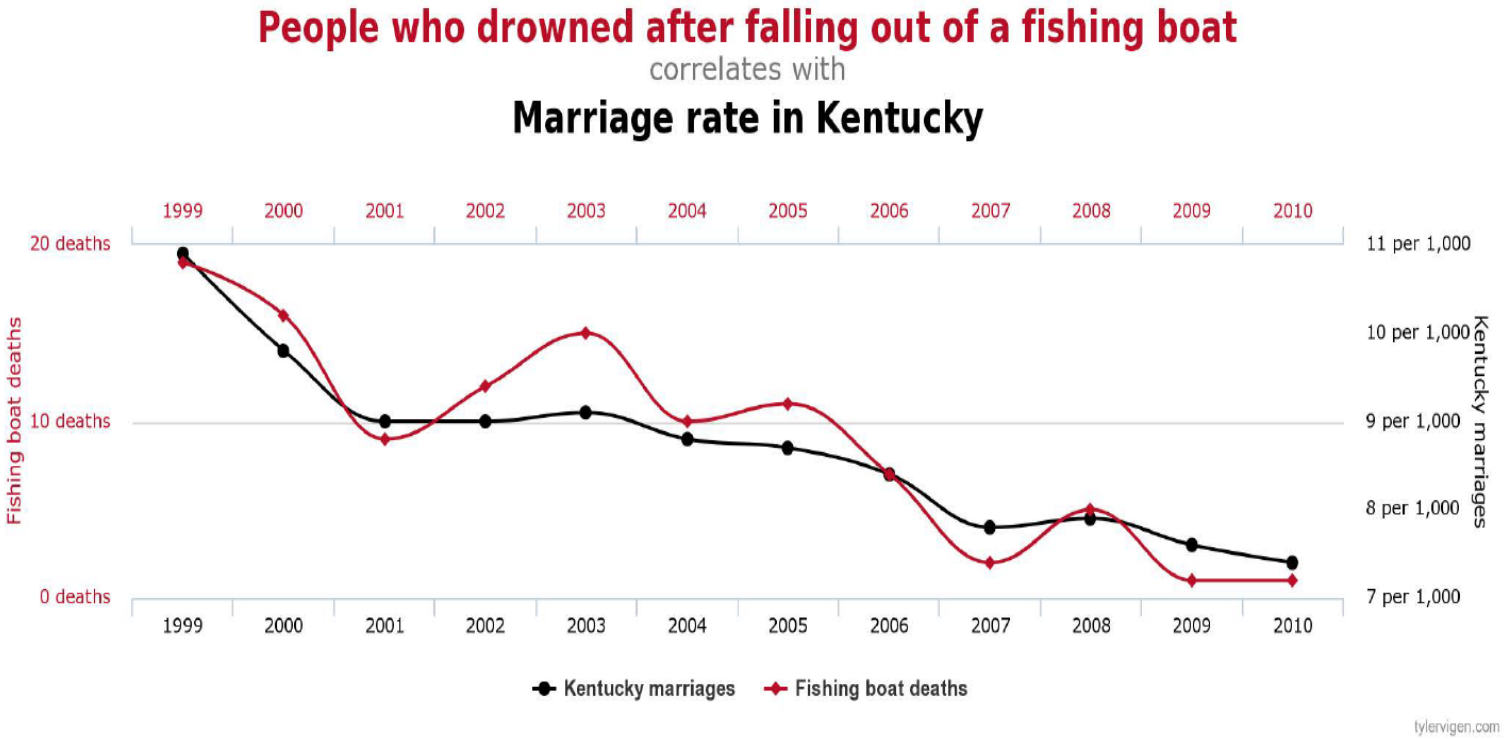
- Cancer survival rate prediction



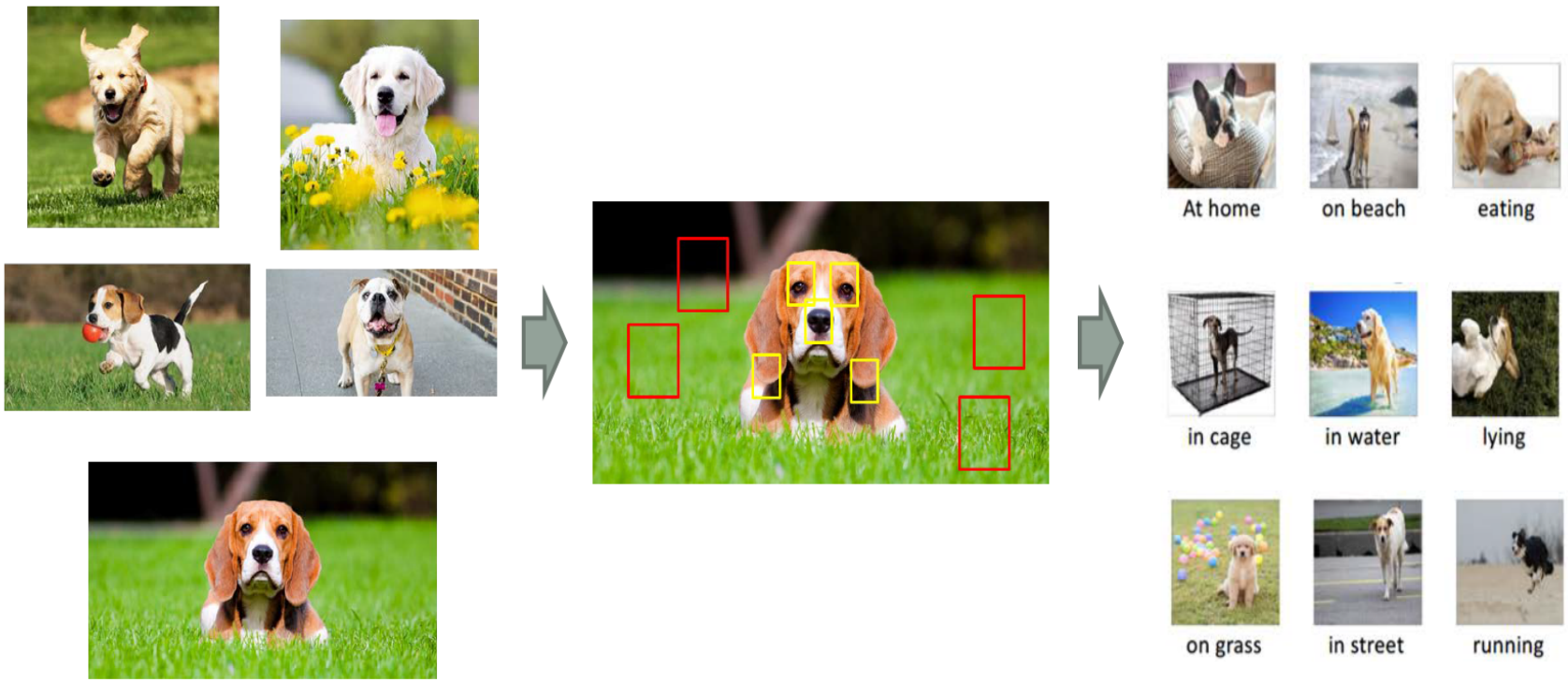
A plausible reason: Correlation

Correlation is the very basics of machine learning....

Correlation is not explainable



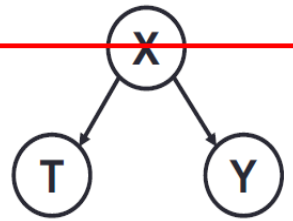
Correlation is "unstable"

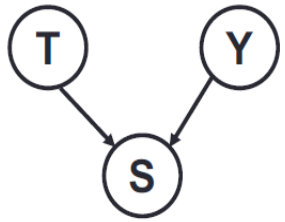


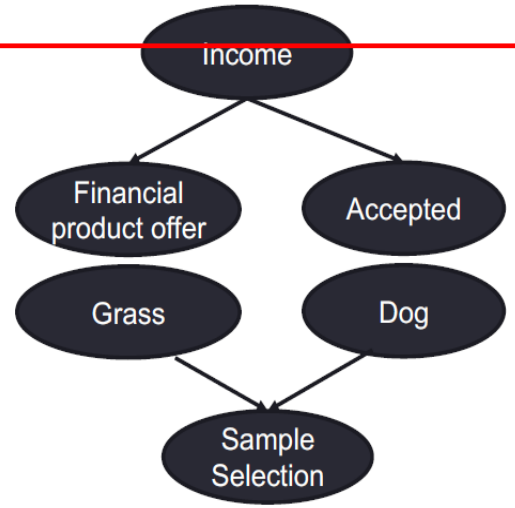
Correlation Vs. Causation

It's not the fault of correlation, but the way we use it...

- Three sources of correlation:
  - Causation
    - Causal mechanism
    - Stable and explainable
  - Confounding
    - Ignoring X
    - Spurious Correlation
  - Sample Selection Bias
    - Conditional on S
    - Spurious Correlation





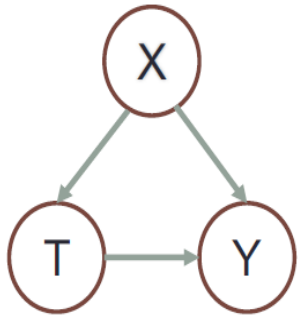


A Practical Definition of Causality

Definition: T causes Y if and only if

changing T leads to a change in Y,

while keeping everything else constant.



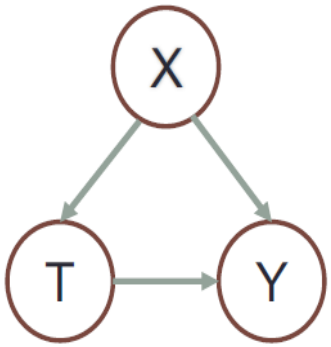
Causal effect is defined as the magnitude by which Y is changed by a unit change in T.

Called the “interventionist” interpretation of causality.

Benefits of bringing causality into learning

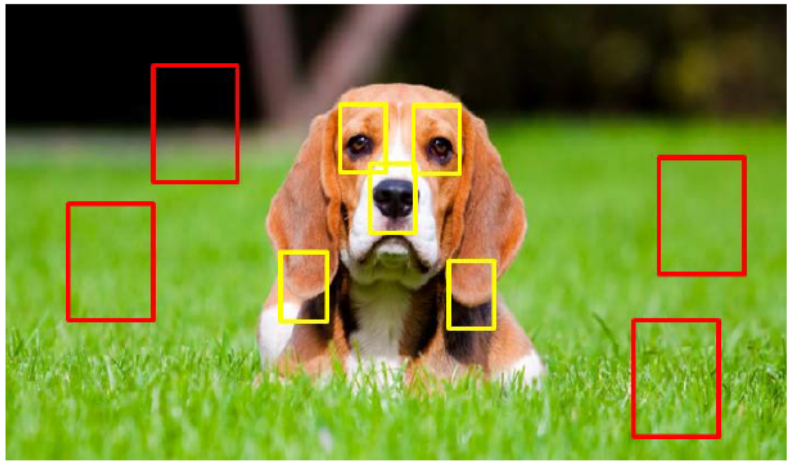
More Explainable and More Stable...

Causal Framework

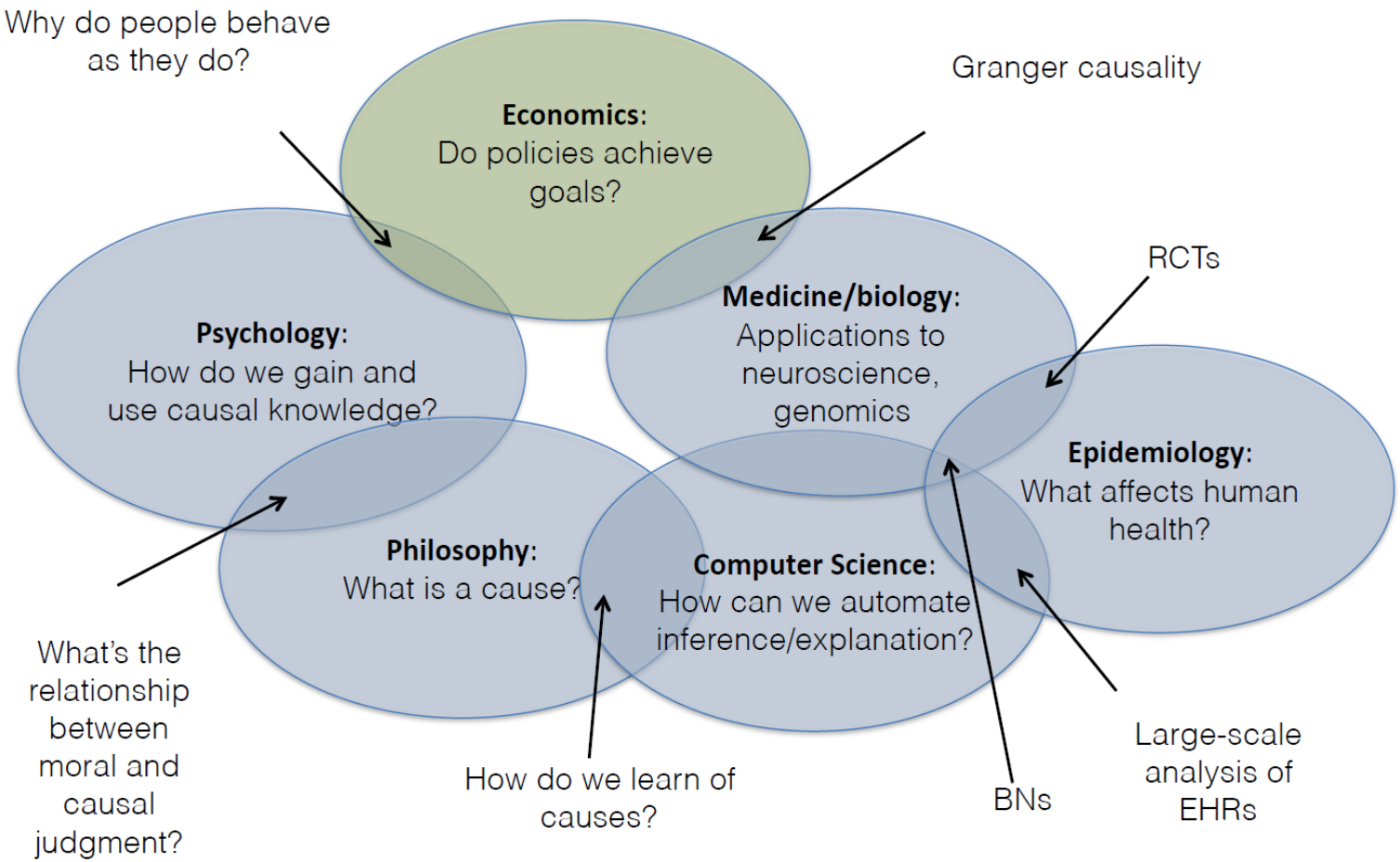


T: grass
X: dog nose
Y: label

Grass—Label: Strong correlation
Weak causation
Dog nose—Label: Strong correlation
Strong causation



Causality everywhere



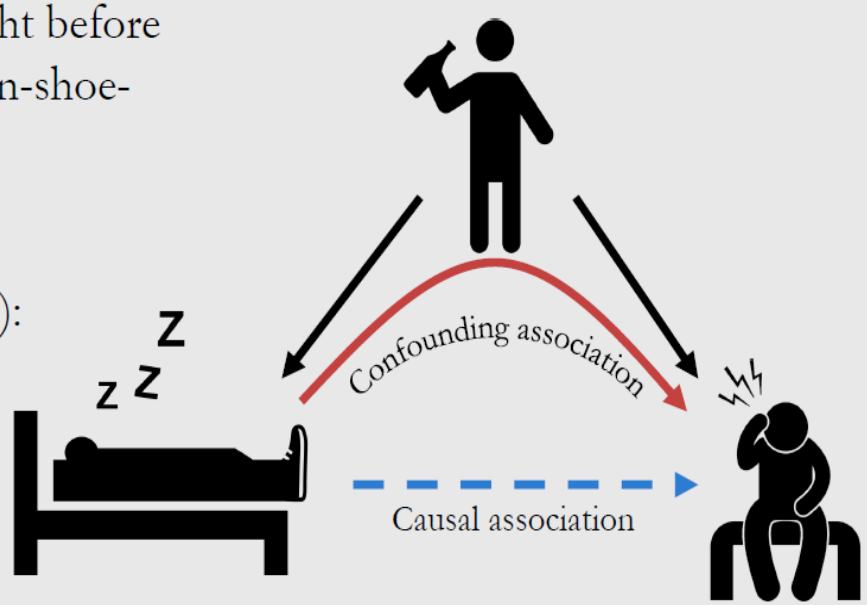
Correlation does not imply causation

Sleeping with shoes on is strongly correlated with waking up with a headache

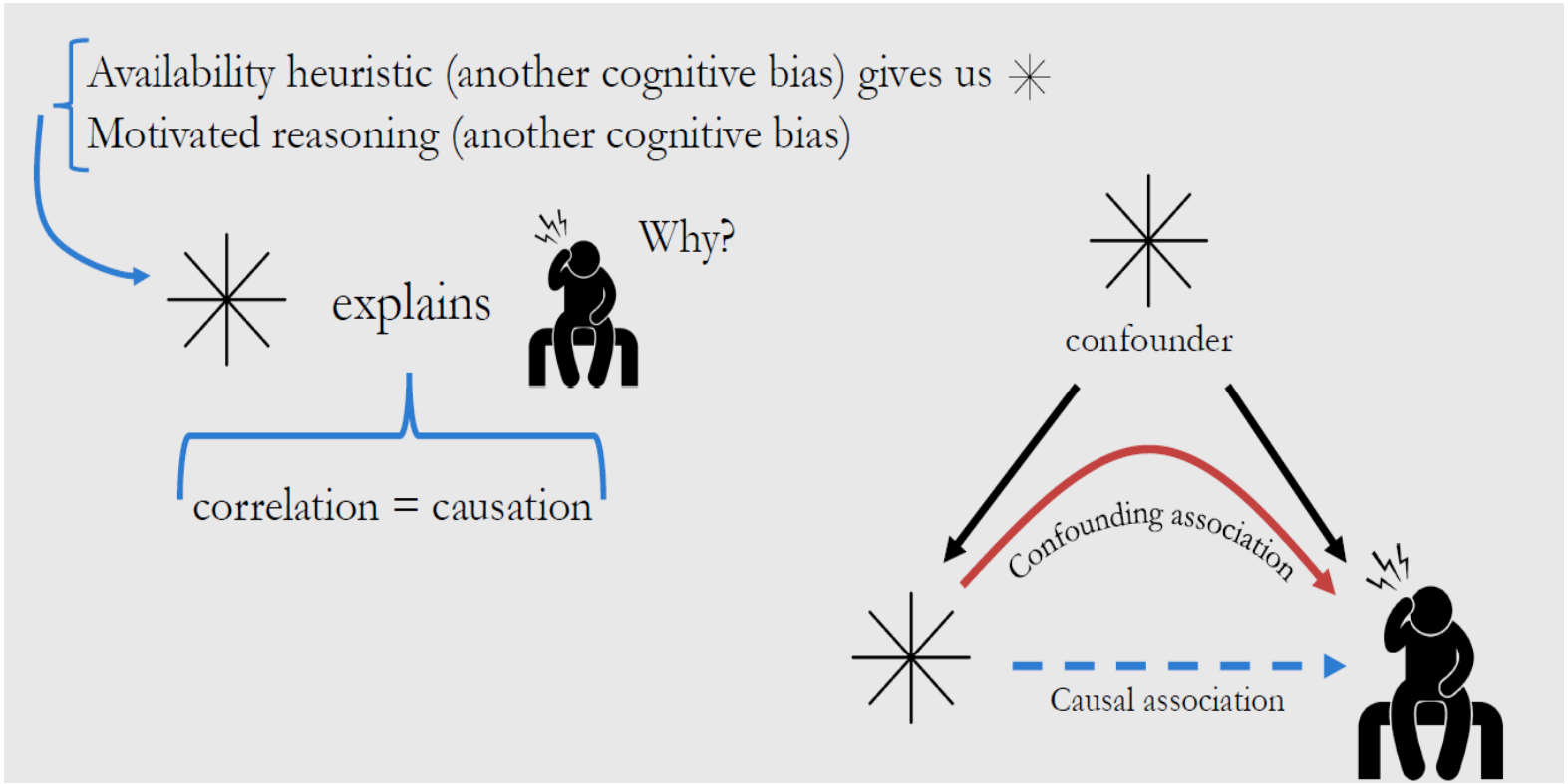
Common cause: drinking the night before

1. Shoe-sleepers differ from non-shoe-sleepers in a key way
2. Confounding

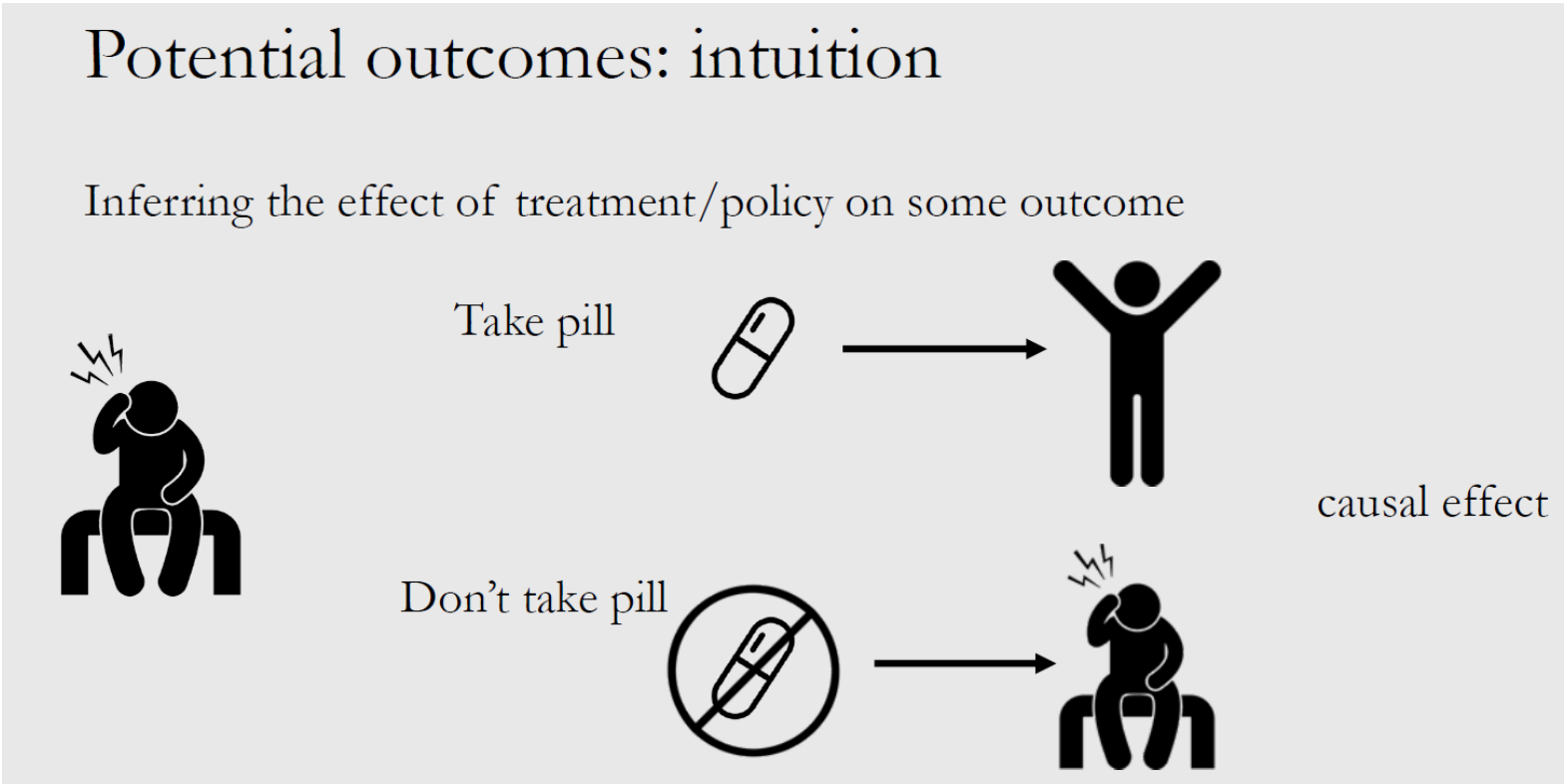
Total association (e.g. correlation): mixture of causal and confounding association



“Correlation = Causation” is a cognitive bias

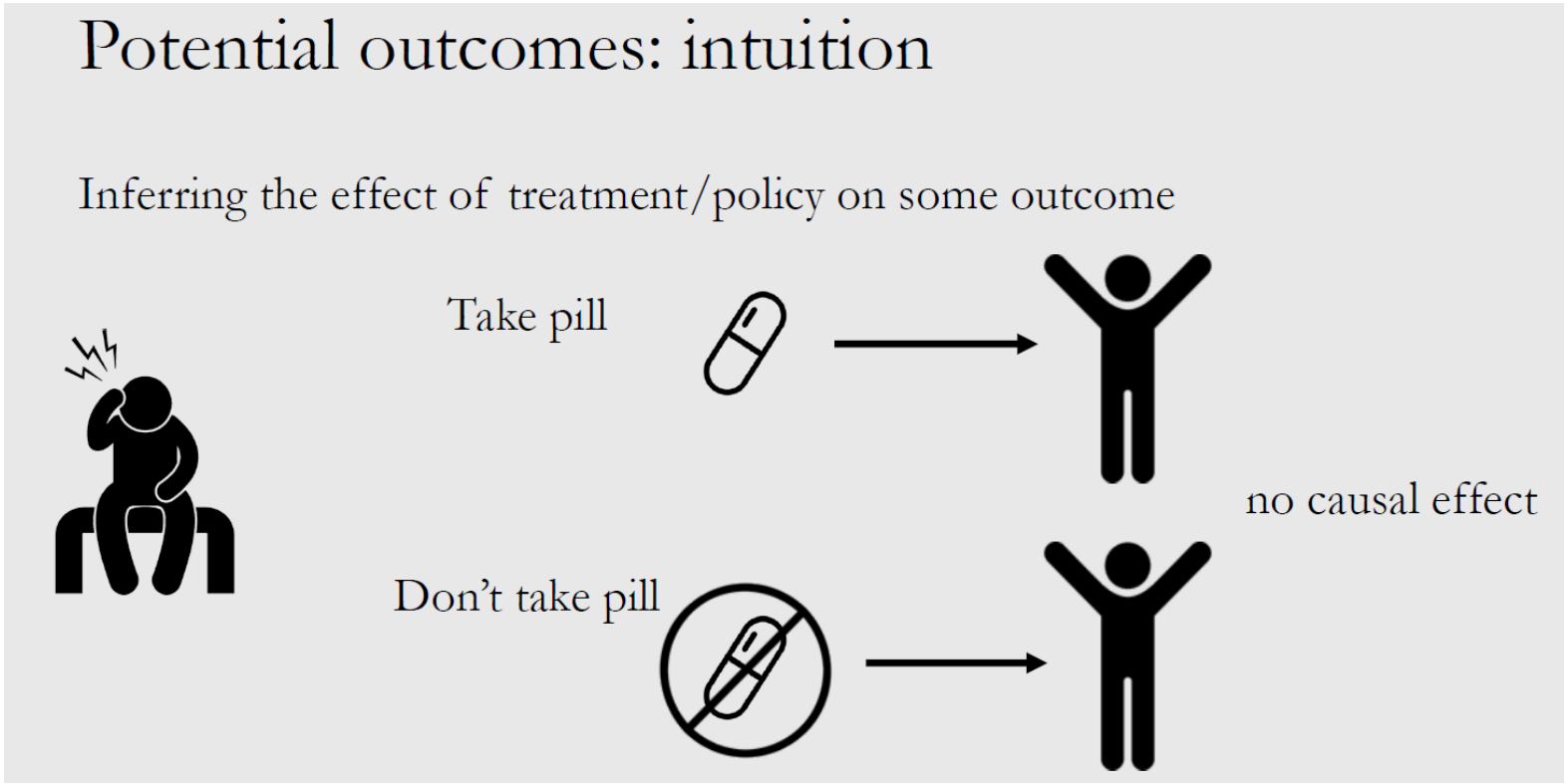


Then, what does imply causation?



Source: https://www.bradyneal.com/causal-inference-course

Then, what does imply causation?



Source: https://www.bradyneal.com/causal-inference-course

Languages for Causality

Using graphs:

- 1921 Wright (genetics);
- 1988 Pearl (computer science "AI");
- 1993 Spirtes, Glymour, Scheines (philosophy).

Using structural equations:

- 1921 Wright (genetics);
- 1943 Haavelmo (econometrics);
- 1975 Duncan (social sciences);
- 2000 Pearl (computer science).

Using potential outcomes / counterfactuals:

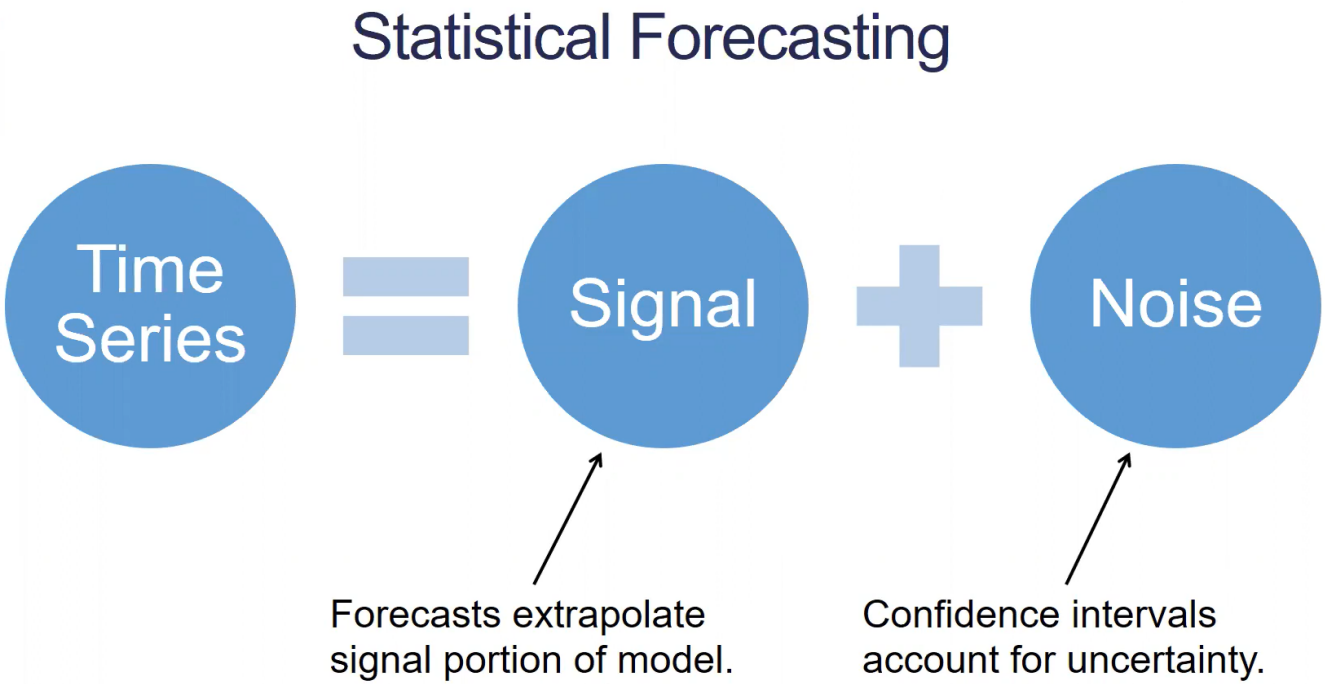
- 1923 Neyman (statistics);
- 1973 Lewis (philosophy);
- 1974 Rubin (statistics);
- 1986 Robins (epidemiology);

Reference: The Book of Why: The New Science of Cause and Effect by Judea Pearl and Dana Mackenzie (2019).

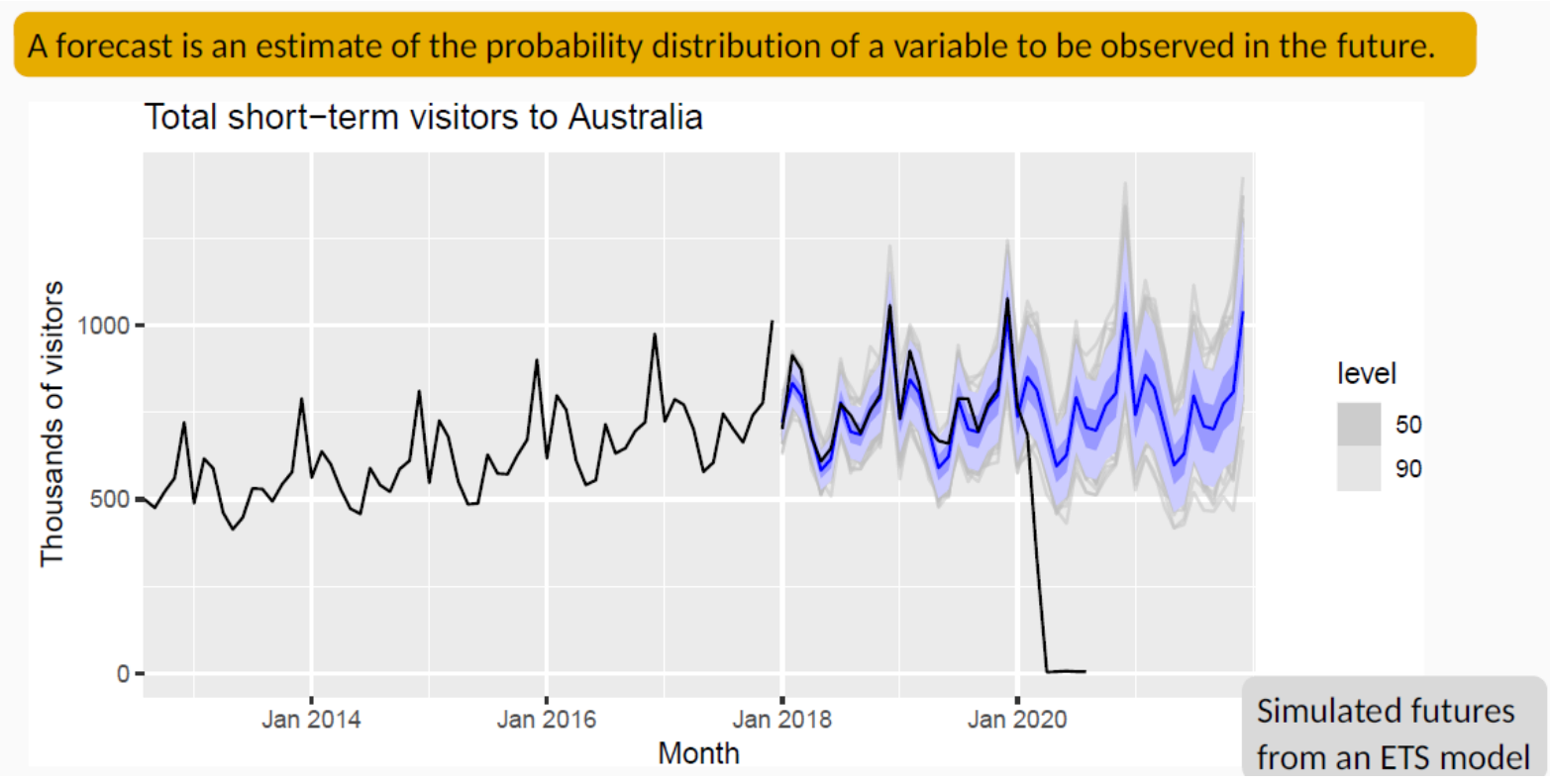
ALL MODELS ARE WRONG, BUT SOME ARE USEFUL!

The Science of Forecasting

Forecasting is estimating how the sequence of observations will continue into the future. Whether it is the rise/fall in exchange rates, the outcome of elections, or winners at the Oscars, there is sure to be something you want to know.



Random futures



Mathematical/Statistical models are simplifications of reality – and life is sometimes too complex to model accurately.

Which is easiest to forecast? (Easy to Tough)

1. Time of sunrise this day next year.
2. Maximum temperature tomorrow.
3. Daily electricity demand in 3 days time.
4. Google stock price tomorrow.
5. Exchange rate of USD/INR next week.

How do we measure "easiest"?

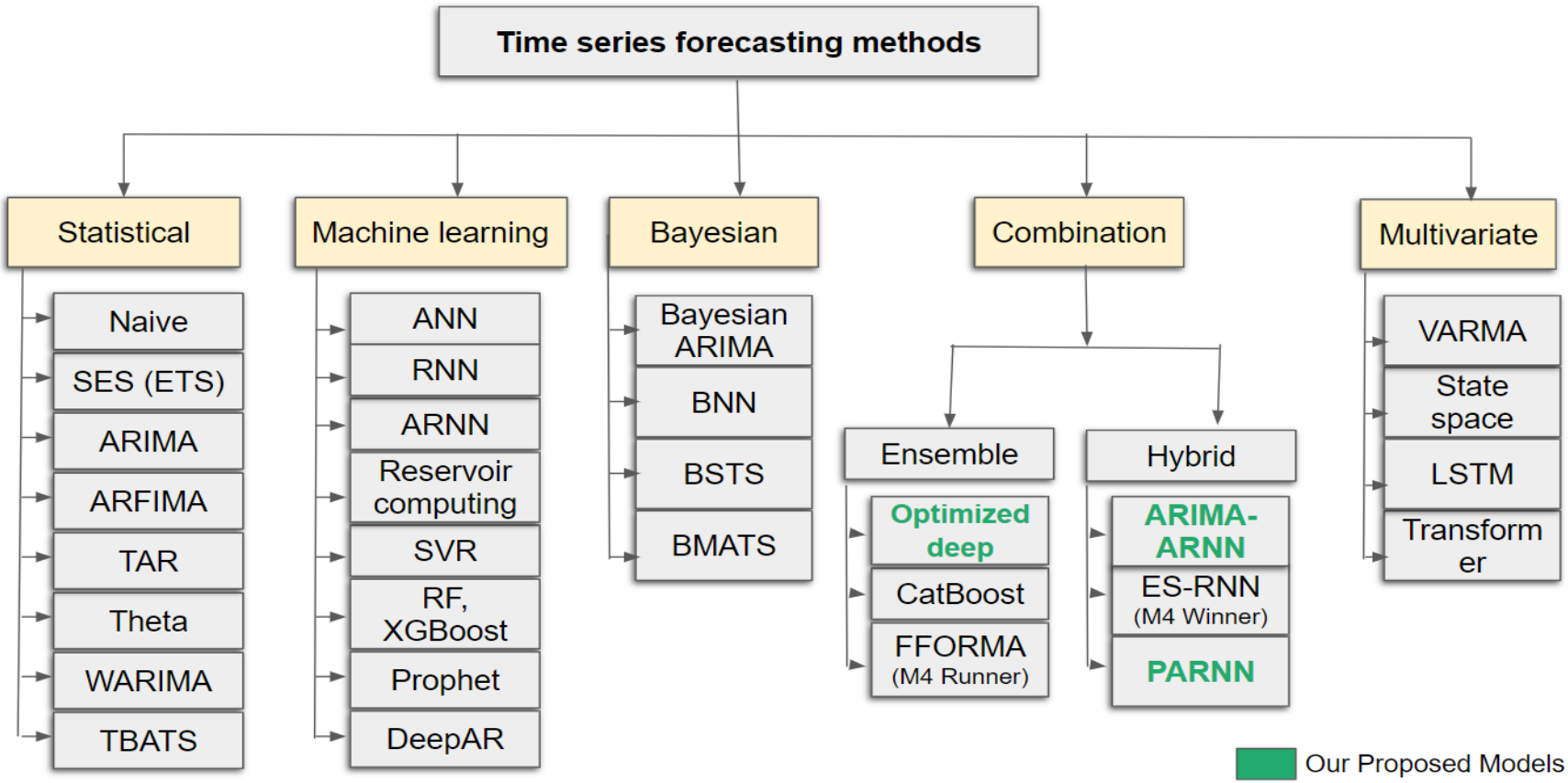
What makes something easy/difficult to forecast?

Forecastability factors

Something is easier to forecast if:

- We have a good understanding of the factors that contribute to it, and can measure them (for stock price and exchange rates causes are mostly unknown).
- There is lots of data available.
- The future is somewhat similar to the past.
- The forecasts cannot affect the thing we are trying to forecast (say, Warren Buffett, CEO of Berkshire Hathaway, make some comment that stock price may change!).
- When should we give up? When there is insufficient data? When the models give implausible forecasts?.

Various Forecasting Models



A recently published survey paper: Nowcasting of COVID-19 confirmed cases: Foundations, trends, and challenges (Chakraborty et al., Modelling, Control and Drug Development for COVID-19 Outbreak Prevention, 2021)

Forecasts can go very wrong

"I think there is a world market for maybe five computers."

(Chairman of IBM, 1943)

"There is no reason anyone would want a computer in their home."

(President, DEC, 1977)

"There's no chance that the iPhone is going to get any significant market share. No chance." (Steve Ballmer, CEO Microsoft, April 2007)

"We're going to be opening relatively soon . . . The virus . . . will go away in April." (Donald Trump, February 2020)

"Prediction is very difficult, especially if it's about the future!"
- Niels Bohr, Danish Physicist & Nobel laureate in Physics.

Textbook and References