What is Data Science?

Ernst C. Wit
Università della Svizzera italiana
Institute of Computing

wite@usi.ch

https://www.ci.usi.ch/

15 March 2022

What do you think Data Science is about?

Data Science ...

What do you think Data Science is about?

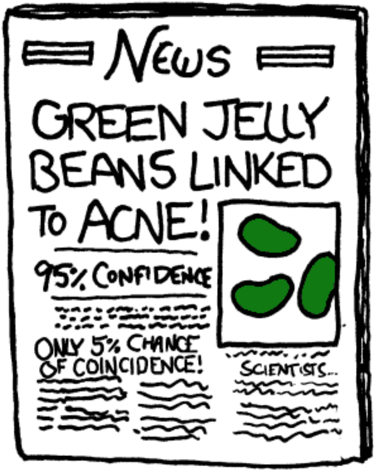
Data Science ...

- ... analyzes loads of data
- ... use artificial intelligence
- ... discovers secret patterns, such as

What do you think Data Science is about?

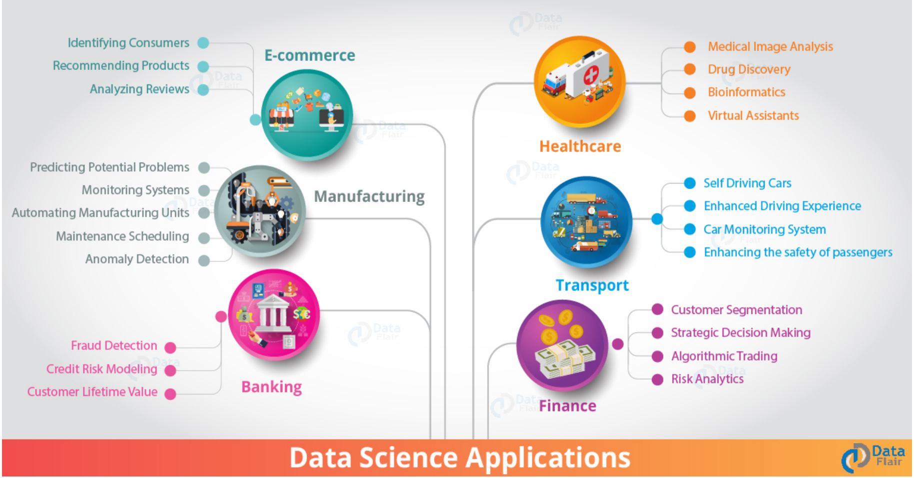
Data Science …

- … analyzes loads of data
- … use artificial intelligence
- … discovers secret patterns, such as


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Università della Svizzera italiana

Where is Data Science used?



How does Data Science work?

It starts out with a **question**:

- What causes Y (e.g. fraud, Covid infection, engineering faults)?
- How to predict Y (e.g. consumers, new drugs, disease)?

It then gathers **data**...

- by designing a study
- by collecting what is available

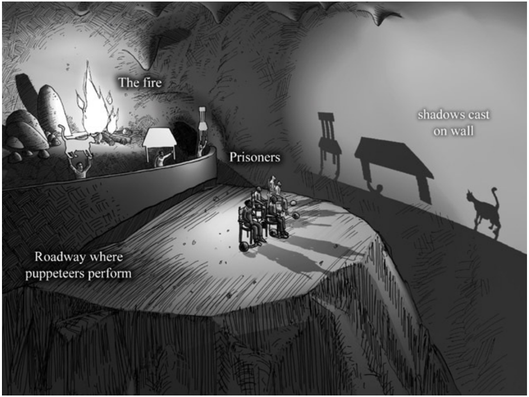
Then use **statistical data science techniques** to analyze data:

- plotting data
- model relationships in data
- formulate conclusions

Data Science? Let's ask Plato!

shadows = data, **cat/table/chair** = true answer

flickering fire = random noise, sampling, ascertainment bias, confounding



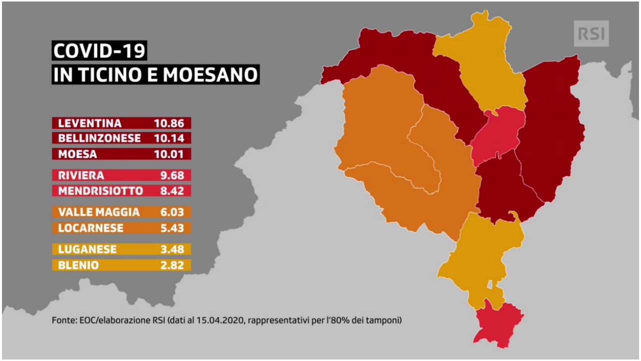
Data Science is controlling **fire** so **prisoners** learn **real** objects only from **shadows**.

What TRUTH do we want to discover?

Covid-19

We want to study three questions in this class:

1. How did pandemic evolve in Ticino?
2. How was pandemic affected by inter-cantonal transmission?
3. Did mortality rate for Covid19 improve?



COVID-19
IN TICINO E MOESANO

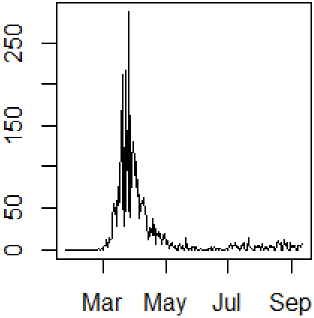
LEVENTINA	10.86
BELLINZONESE	10.14
MOESA	10.01
RIVIERA	9.68
MENDRISIOTTO	8.42
VALLE MAGGIA	6.03
LOCARNESE	5.43
LUGANESE	3.48
BLENIO	2.82

Fonte: EOC/elaborazione RSI (dati al 15.04.2020, rappresentativi per l'80% dei tamponi)

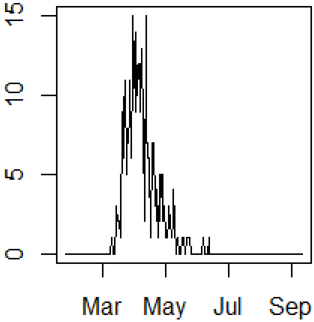
Shadows = Data

shadows cast
on wall

Let’s look at Swiss Covid-19 data (2020)



Daily cases in TI



Daily deaths in TI

Fire = Noise, sampling bias, confounding, ...

Sources of Uncertainty: “Noise”

- “Chance”
  - Measurement uncertainty;
  - Intrinsic system noise;
- “Sampling”
  - Experimental design
- “Confounding”

Sources of Uncertainty: “Noise”

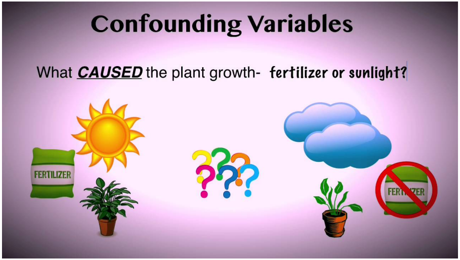
- **“Chance”**
  - Measurement uncertainty;
  - Intrinsic system noise;
- “Sampling”
  - Experimental design
- “Confounding”

Sources of Uncertainty: “Noise”

- “Chance”
  - Measurement uncertainty;
  - Intrinsic system noise;
- **“Sampling”**
  - Experimental design
- “Confounding”

Sources of Uncertainty: “Noise”

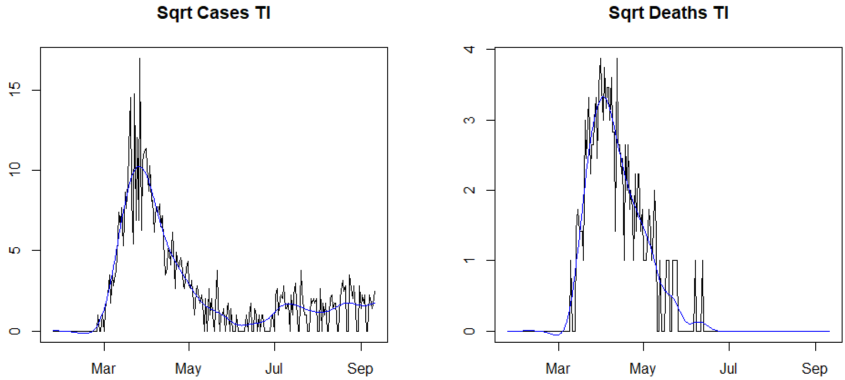
- “Chance”
  - Measurement uncertainty;
  - Intrinsic system noise;
- “Sampling”
  - Experimental design
- **“Confounding”**



Confounding Variables

What ***CAUSED*** the plant growth- **fertilizer or sunlight?**

How to deal with noise?



We do 2 things:

- transform data with $\sqrt{\ }$ (takes away large extremes)
- smooth data (spreads deaths/cases over empty weekends)

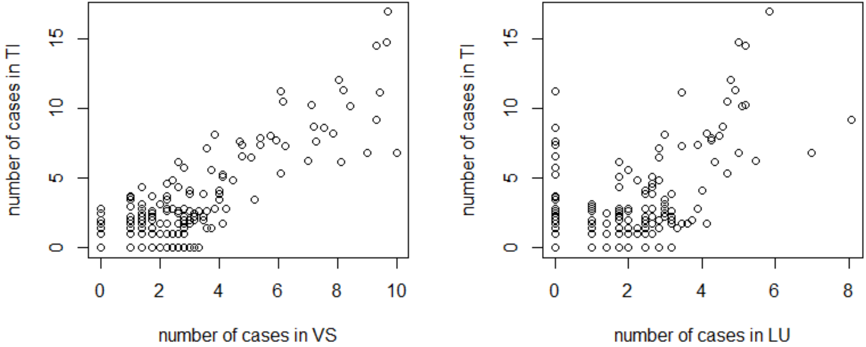
TRUTH

How did infections spread through Switzerland?

Università della Svizzera italiana

Relationship between Cantons: who infected who?

Consider relationship between TI and other cantons:

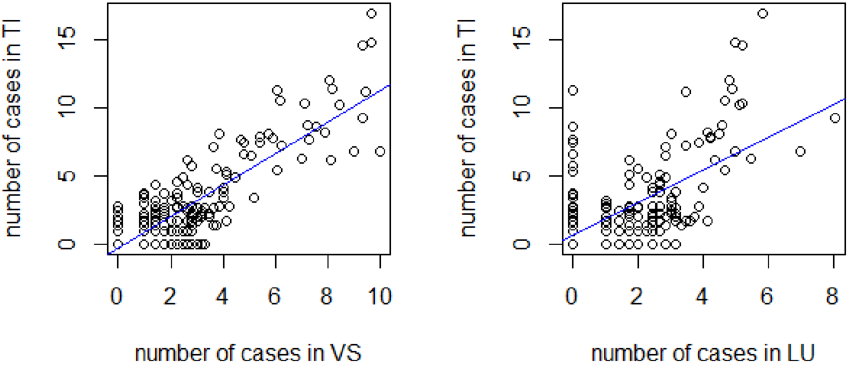


Question: Who affected who?

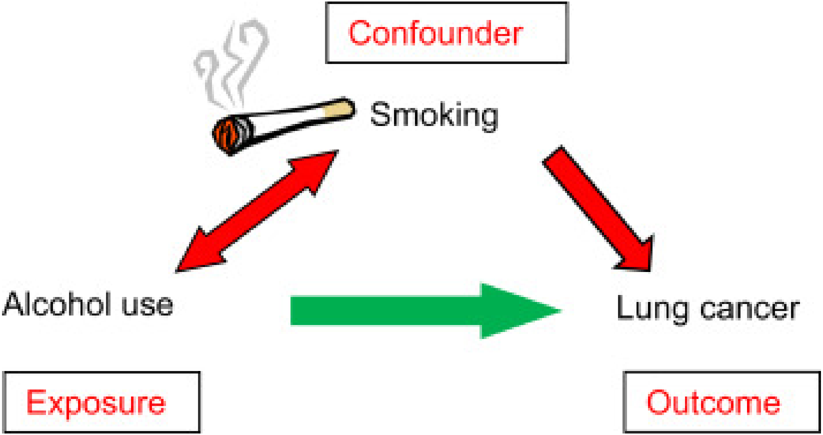
Linear Regression: modelling relationships

We can model relationships between TI and other cantons:

$$\begin{aligned}\sqrt{\text{cases in TI}} &= \alpha_1 + \beta_1\sqrt{\text{cases in VS}} + \text{noise}\\ \sqrt{\text{cases in TI}} &= \alpha_2 + \beta_2\sqrt{\text{cases in LU}} + \text{noise}\end{aligned}$$



Sources of Uncertainty: Confounding



Confounder

Smoking

Alcohol use

Lung cancer

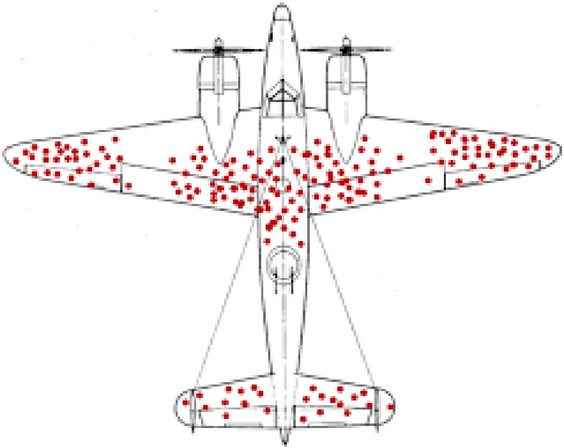
Exposure

Outcome

Confounding: what data do you get to see?

Different types

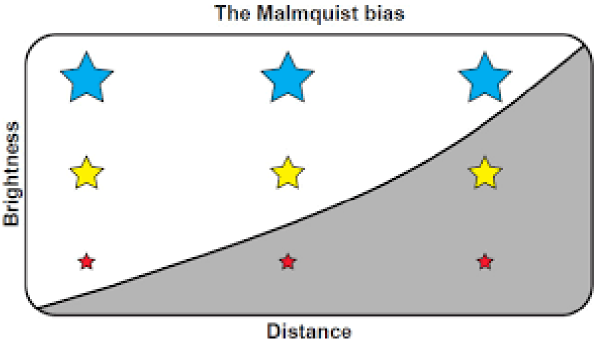
- Non-response bias
- Healthy user bias
- Berkson's fallacy
- Overmatching
- **Survivorship bias**
- Malmquist bias



Confounding: what data do you get to see?

Different types

- Non-response bias
- Healthy user bias
- Berkson's fallacy
- Overmatching
- Survivorship bias
- **Malmquist bias**



Linear Regression: modelling complex relationships

We can model relationships between TI and other cantons TOGETHER:

$$\sqrt{\text{cases in TI}} = \beta_0 + \beta_1 \sqrt{\text{cases in VS}} + \beta_2 \sqrt{\text{cases in LU}} + \text{noise}$$

Linear Regression: modelling complex relationships

We can model relationships between TI and other cantons TOGETHER:

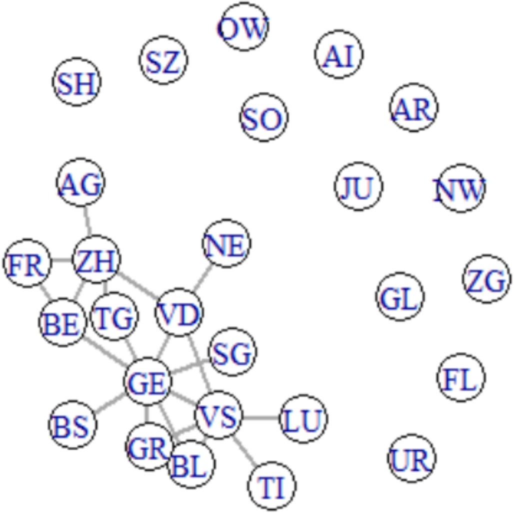
$$\sqrt{\text{cases in TI}} = \beta_0 + \beta_1\sqrt{\text{cases in VS}} + \beta_2\sqrt{\text{cases in LU}} + \text{noise}$$

	Estimate	Std. Error	t value	Pr(>|t|)
β_0	-0.26	0.15	-1.66	0.09
β_1 (VS)	1.11	0.07	16.3	< 0.01
β_2 (LU)	0.07	0.09	0.74	0.48

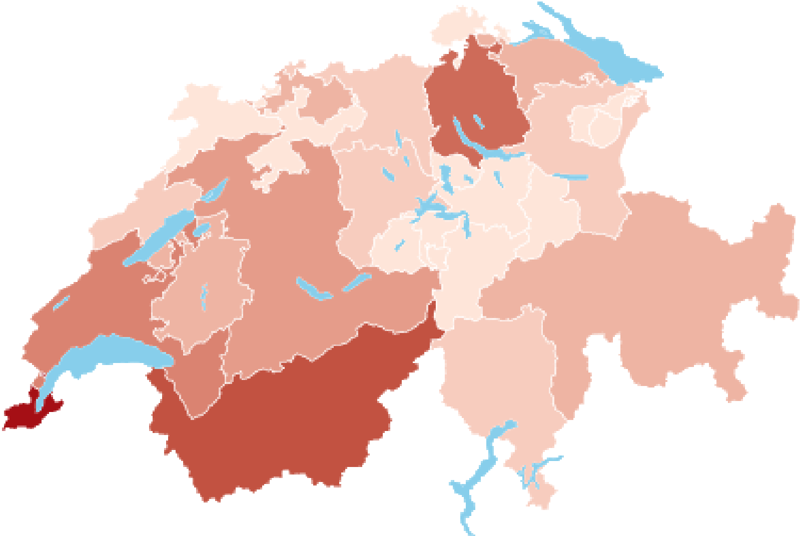
So, VS affected TI, but LU didn't!

Università della Svizzera italiana

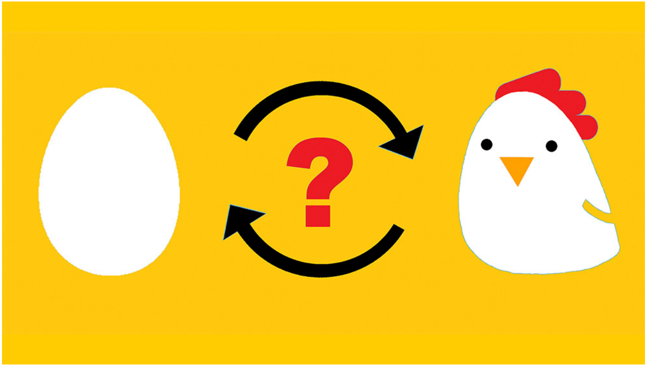
… Now do this for all cantons simultaneously!



Which canton has the most connections?

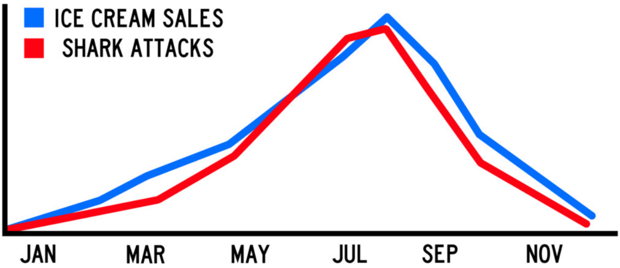


Direction of arrow?



Correlation $\neq$ causation

CORRELATION IS NOT CAUSATION!

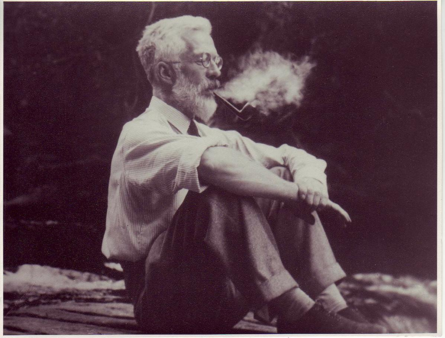


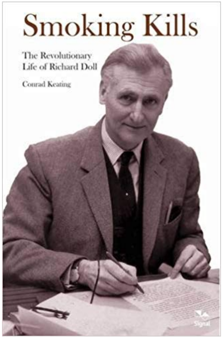
Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

Events are connected by a **common cause**: confounding

R.A. Fisher vs. Richard Doll

R.A. Fisher (geneticist and statistician) was a fervent smoker.





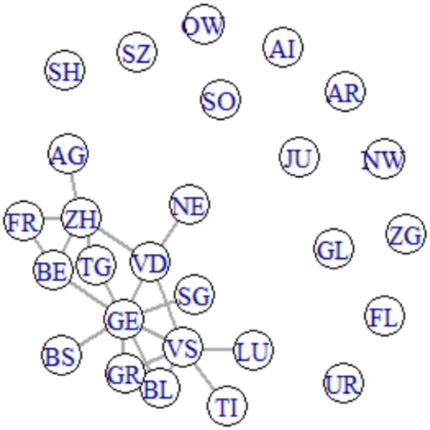
"Smoking and lung cancer are confounded"

"Control for *all* possible confounders"

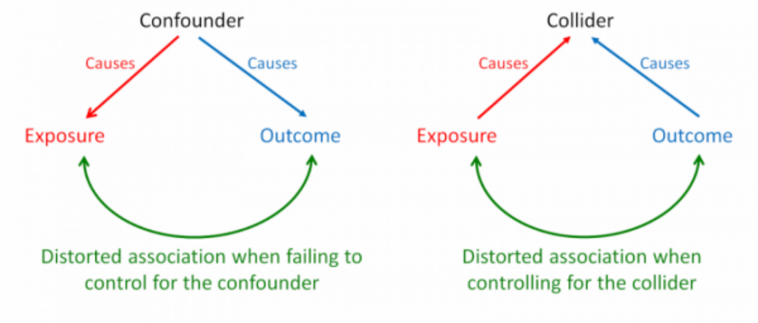
Sir Richard Doll conducted:

- 1950. Lung cancer study in 20 London hospitals.
- 1954–2001 British Doctor Study to eliminate confounders.

Prediction graph



Inferring Causality


Confounder
Causes
Causes
Exposure
Outcome
Distorted association when failing to control for the confounder
Collider
Causes
Causes
Exposure
Outcome
Distorted association when controlling for the collider

COLLIDER RULE: If

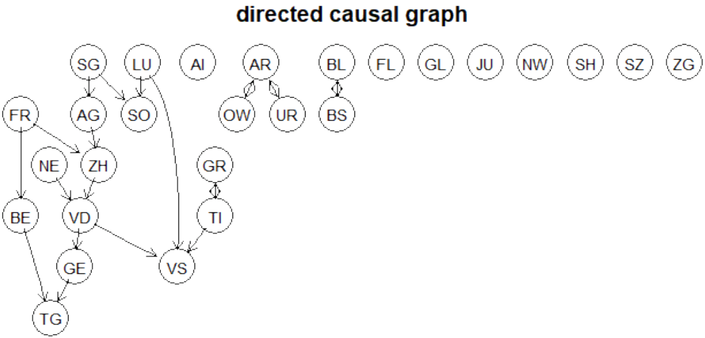
- **conditional** dependence between A-C and B-C and A-B
- no dependence between A-B

Then

- A and B cause C.

Causal network of infections: Who infected who?

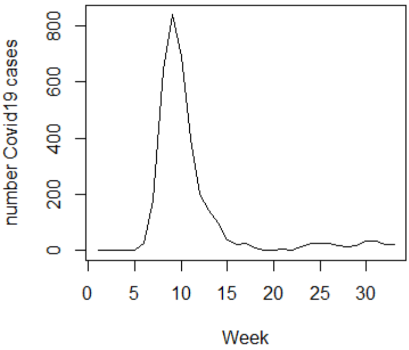
Applying PC algorithm to Covid network in CH, we find:

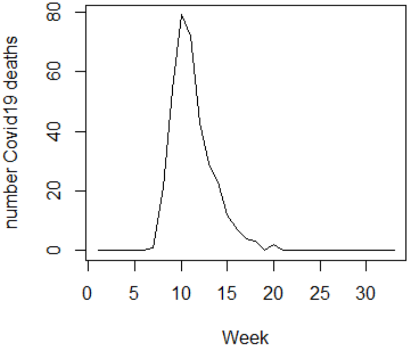


BONUS

Did mortality rates improve during Covid-19 pandemic?

Weekly data





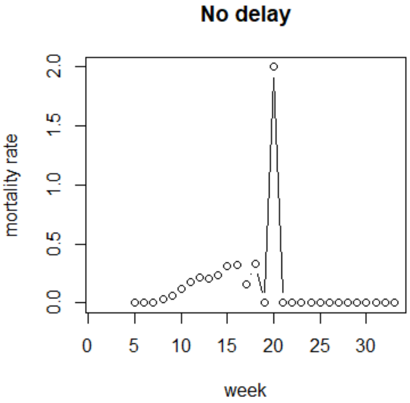
Ratio of deaths over cases should tell us something about death rate....

Did mortality rate improve?

$$\text{mortality}_0(t) = \frac{\text{deaths}(t)}{\text{cases}(t)}$$

Did mortality rate improve?

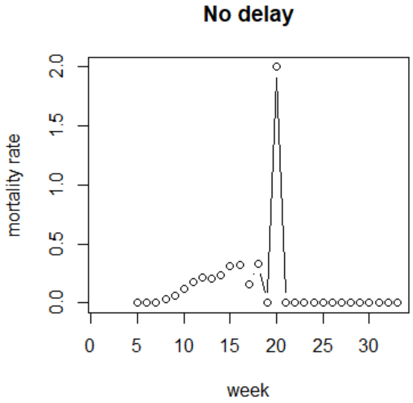
$$\text{mortality}_0(t) = \frac{\text{deaths}(t)}{\text{cases}(t)}$$



Did mortality rate improve?

$$\text{mortality}_0(t) = \frac{\text{deaths}(t)}{\text{cases}(t)}$$

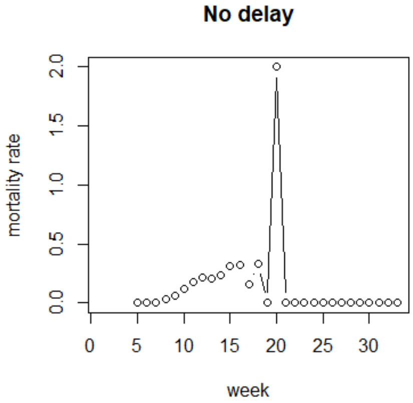
$$\text{mortality}_2(t) = \frac{\text{deaths}(t)}{\text{cases}(t-2)}$$

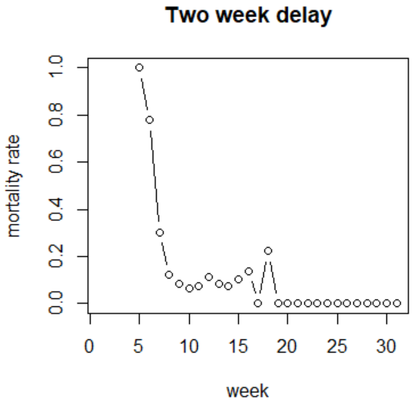


Did mortality rate improve?

$$\text{mortality}_0(t) = \frac{\text{deaths}(t)}{\text{cases}(t)}$$

$$\text{mortality}_2(t) = \frac{\text{deaths}(t)}{\text{cases}(t-2)}$$



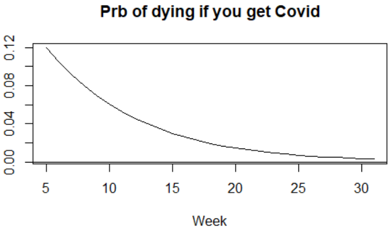


Logistic regression model

We fit mortality regression with time:

$$\text{odds of dying of covid-19 (in week } t) = e^{\alpha+\beta t}$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.262	0.073	-17.35	0.0000
week	-0.147	0.007	-21.51	0.0000



Probability of dying of Covid-19 reduced approx 14% each week:

$$e^{-0.147} = 0.86$$

Conclusions

- **Never take data for granted!**
  (There may be all kind of errors!)
- **Value of data lies entirely in its collection!**
  (Randomized designs are more valuable than observational ones.)
- **Modelling needs to capture underlying process, but also design!**
  (The weaker the design, the stronger the modelling needs to be...)