UiO : Department of informatics
University of Oslo

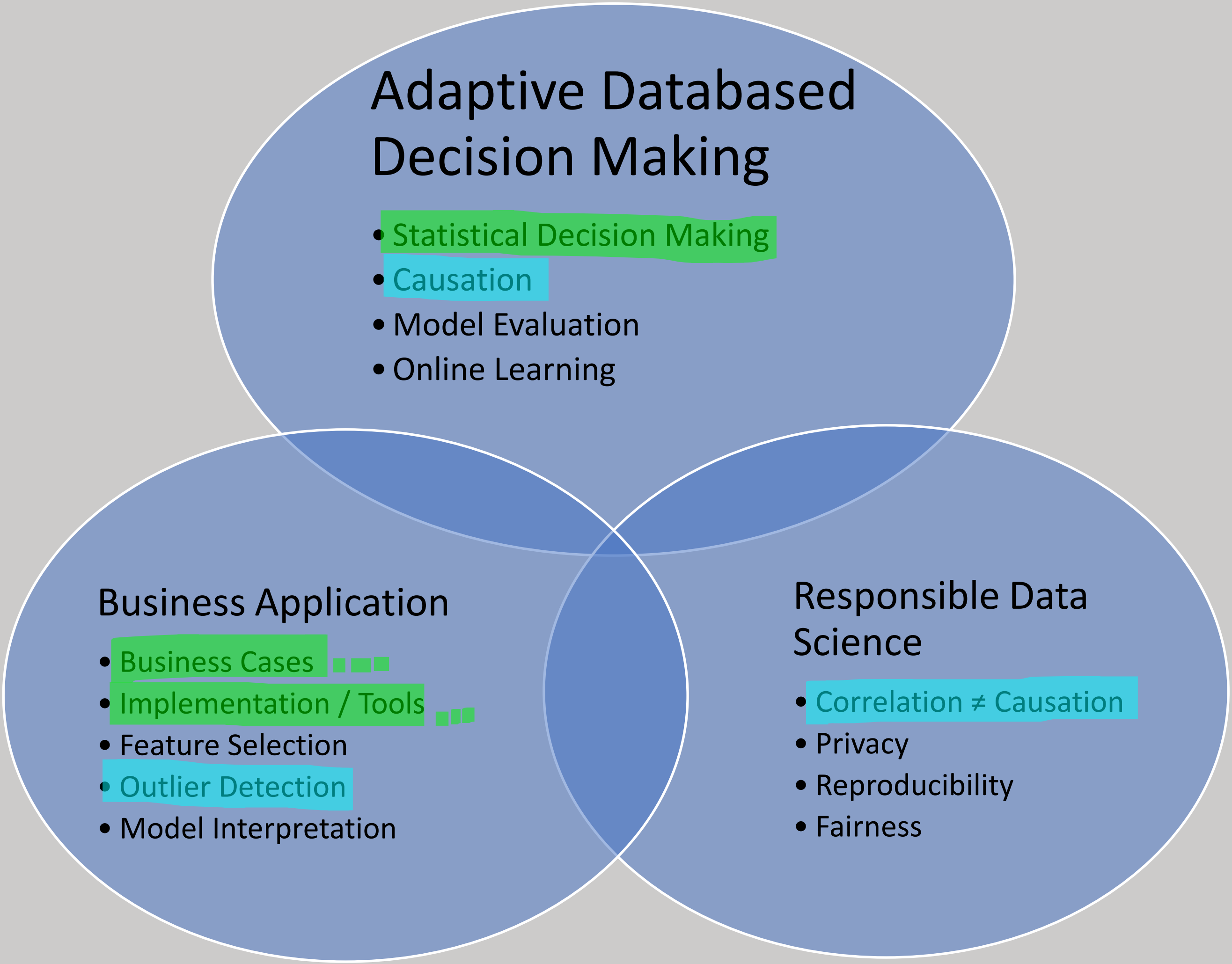
Adaptive Methods for Data-based Decision Making *Lecture 4*

IN-STK 5000 / 9000

Autumn 2022

Slides by Dr. Anne-Marie George, UiO

Course Overview



What we talk about today

- Correlation Coefficients
- Outlier / Anomaly Detection
- Clustering
- Causation
- Hypothesis Testing

The Ice Cream Van

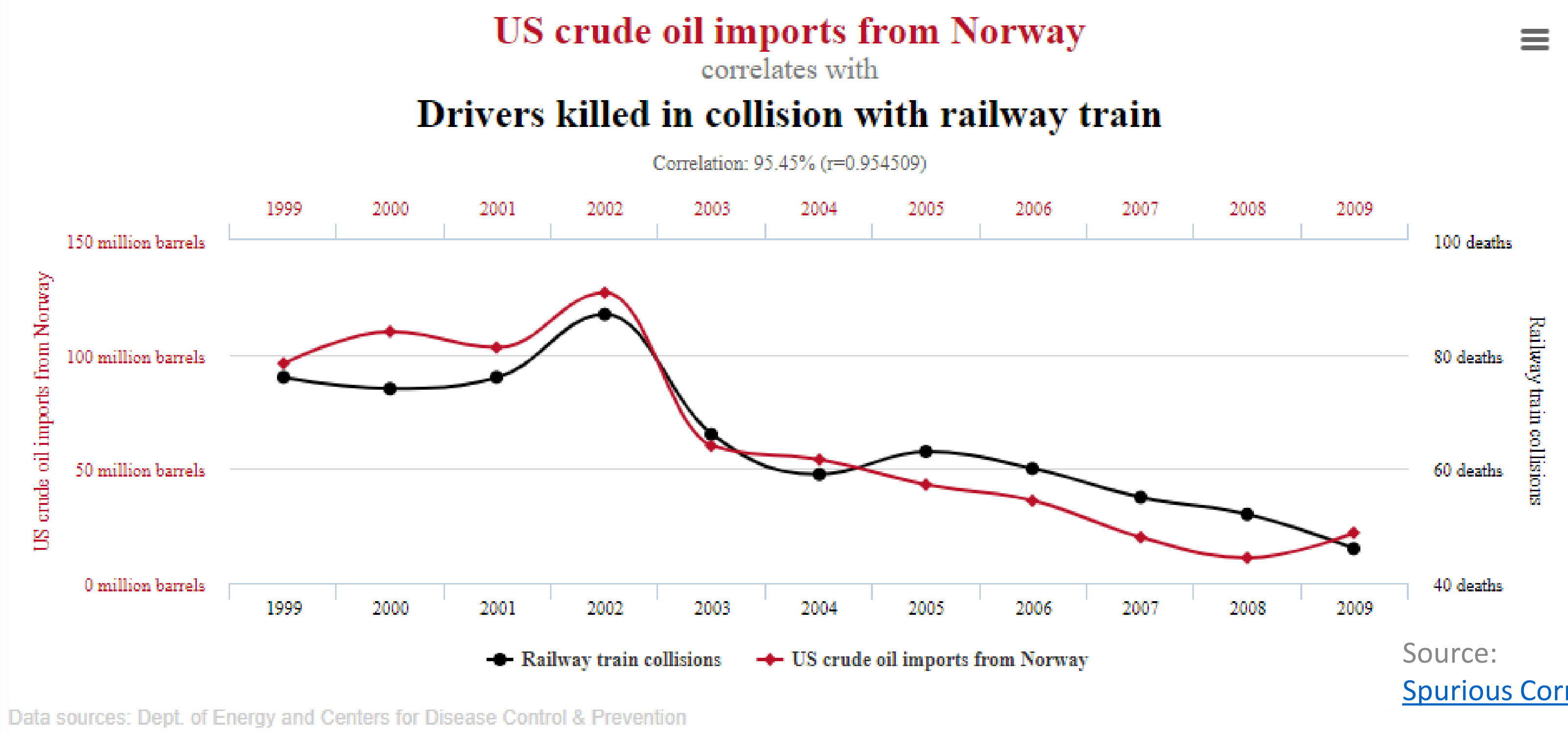
<u>Data on the last few ice cream sales shows</u>:
When you play a jingle, you sell a lot of ice cream!

What do you make out of that?

Spurious Correlations

Should Norway stop exporting oil?

- Mathematical relationship between events / variables that are associated but not *causally* related (coincidence, unseen factor, …).



Source:
[Spurious Correlations (tylervigen.com)]

Correlation $\neq$ Causation

- Definition:
  Two (or more) variables have a relation to one another.
- Measure:
  Pearson Coefficient, Spearman Rank, Kendall Tau Distance, …

Let's start with this! …

But first some reminders / basics.

→ Improve predictions

→ First step towards identifying causation

- Definition:
  Some variable(s) causes the behavior of another variable.
- Measure:
  Hypothesis testing, …

→ Influence future events

→ Improve decisions

Expectation, Variance and Standard Deviation

Random variable x with **probability distribution** p over **outcomes** $R \subseteq \mathbb{R}$.

Notation: $x \sim p$

Expectation / Mean: $\mu_x = \mathbb{E}_p[x] = \int_{r \in R} r \, dp(r)$ or $\sum_{r \in \mathrm{R}} p(r) \cdot r$

Variance: $\mathbb{V}[x] = var(x) = \mathbb{E}_p[x^2] - \left(\mathbb{E}_p[x]\right)^2 = \sigma_x^2$

Standard Deviation: $\sigma_x = \sqrt{\mathbb{V}[x]}$

Gaussian Distribution aka. Normal Distribution

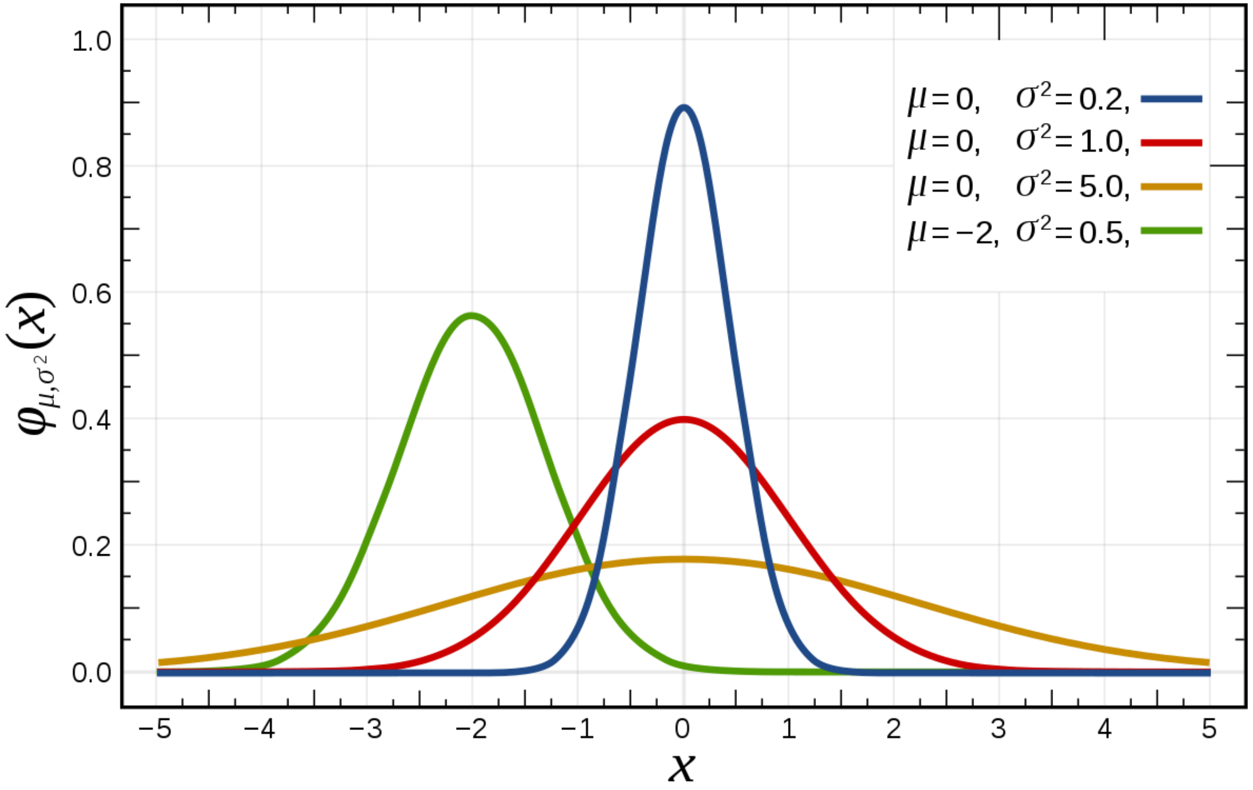
Normal distribution: $\mathcal{N}(\mu, \sigma)$ with mean μ and standard deviation σ (variance σ^2).

PDF: $$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- The higher the mean μ …

  … the more the peak moves to the right.
- The higher the standard deviation σ …

  … the flatter the curve.



Source: Wikipedia

Covariance and Pearson's Correlation

Real random variables x, y with **probability distribution** p over **outcomes** R.
Notation: $(x, y) \sim p$ where $p_x(y) = p(x, y)$ and $p_y(x) = p(x, y)$

Covariance:

$$cov(x, y) = \mathbb{E}_p\left[\left(x - \mathbb{E}_{p_y}[x]\right)\left(y - \mathbb{E}_{p_x}[y]\right)\right]$$
$$= \mathbb{E}_p[x \cdot y] - \mathbb{E}_{p_y}[x] \cdot \mathbb{E}_{p_x}[y]$$

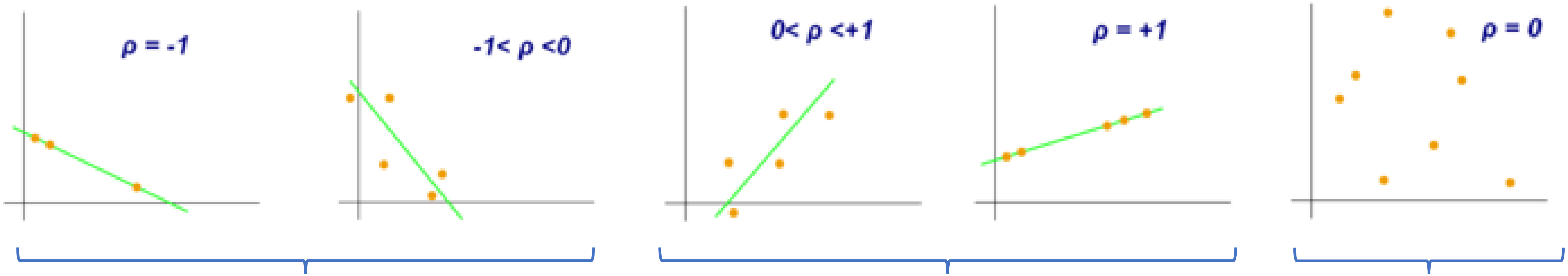
Pearson's correlation coefficient:

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y}, \quad \text{if } \sigma_x \cdot \sigma_y > 0$$ (non-zero st. dev.s)

Normalized covariance → Values in $[-1, +1]$

Pearson’s Correlation Coefficient

Measure of <u>linear</u> correlation
→ Other types of correlations are ignored



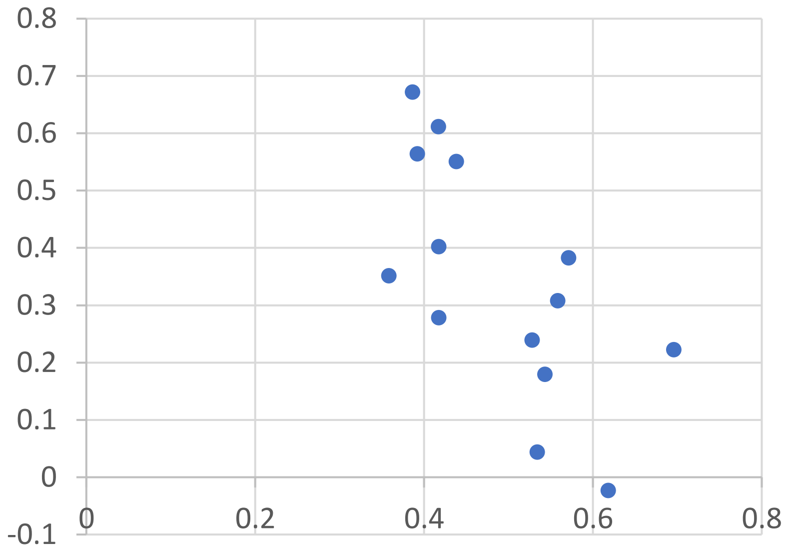
Negative (linear) correlation | Positive (linear) correlation | No (linear) correlation

Source: [Wikipedia]

Sample Versions

But …
How do I calculate

- the mean
- the covariance
- the variance
- the standard deviation
- the correlation



… for my **data**?!?

Sample Versions of Standard Notions

Unbiased version. (the biased version uses $1/N$)

	General: $x \sim p,\ p: R \longrightarrow [0,1],\ R \subseteq \mathbb{R}$	**Samples** $x_1, \ldots, x_N \in \mathbb{R}$
Expectation / Mean	$\mu_x = \mathbb{E}_p[x] = \int_{r \in R} r\, dp(r)$ (R continuous) or $= \sum_{r \in R} p(r) \cdot r$ (R discrete)	$\bar{x} = \frac{1}{N} \sum_{i=1,\ldots,N} x_i$
Variance	$\mathbb{V}[x] = var(x) = \mathbb{E}_p[x^2] - \left(\mathbb{E}_p[x]\right)^2 = \sigma^2$	$\bar{\sigma}^2 = \frac{1}{N-1} \sum_{i=1,\ldots,N} (x_i - \bar{x})^2$
Standard Deviation	$\sigma = \sqrt{\mathbb{V}[x]}$	$\bar{\sigma} = \sqrt{sample\ variance}$

	General: $(x, y) \sim p,\ p: R^2 \longrightarrow [0,1],\ R \subseteq \mathbb{R}$	**Samples** $(x_1, y_1), \ldots, (x_N, y_N) \in \mathbb{R}$
Covariance	$cov(x,y) = \mathbb{E}_p[x \cdot y] - \mathbb{E}_{p_y}[x] \cdot \mathbb{E}_{p_x}[y]$	$\overline{cov}(x,y) = \frac{1}{N-1} \sum_{i=1,\ldots,N} (x_i - \bar{x})(y_i - \bar{y})$
Pearson's Correlation Coeff.	$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y}$, if $\sigma_x \cdot \sigma_y > 0$	$\bar{\rho}_{x,y} = \frac{\overline{cov}(x,y)}{\overline{\sigma_x} \cdot \overline{\sigma_y}}$, if $\overline{\sigma_x} \cdot \overline{\sigma_y} > 0$

Sample Covariance

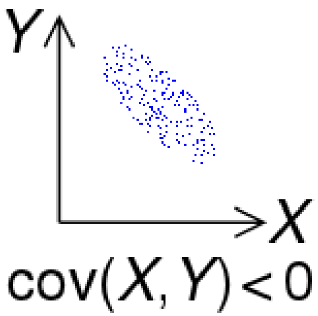
Samples $(x_1, y_1), \dots, (x_N, y_N) \in \mathbb{R}^2$.

Covariance:

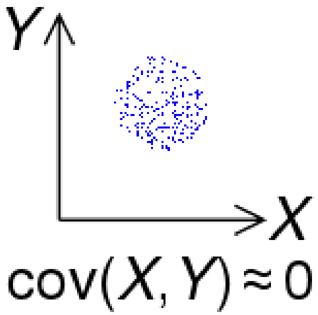
$$\overline{cov}(x, y) = \frac{1}{N-1} \sum_{i=1,\dots,N} (x_i - \bar{x})(y_i - \bar{y})$$

→ For data with K features:
The covariance is a $K \times K$ matrix.

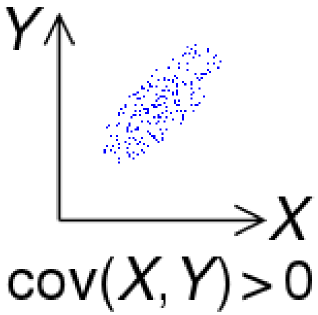
"When x increases,
y decreases"



"No trend between
x and y"



"When x increases,
y increases"



Source: [Wikipedia]

Rank Correlations

x	R(x)	y	R(y)
0.4	1	0.1	1
0.7	2	1.6	3
1.9	3	1.3	2

Samples $(x_1, y_1), \dots, (x_N, y_N) \in \mathbb{R}$.

Rankings of values $R(x)$ and $R(y)$.

- Spearman's Rank Correlation Coefficient: $\rho_{R(x),R(y)} = \frac{cov(R(x),R(y))}{\sigma_{R(x)} \cdot \sigma_{R(y)}}$
- Kendall's Rank Correlation Coefficient:

$$\tau_{x,y} = \frac{\#\, concordant\ pairs - \#\, discordant\ pairs}{\#\, all\ pairs}$$

→ Can be applied for "ranking data" from e.g. user ratings.

Example 1: Differences in Rank Correlations

		x	R(x)	y	R(y)
Data		0.4	1	0.1	1
		0.7	2	1.6	3
		1.9	3	1.3	2
Mean	$\bar{x} = \frac{1}{N}\sum_{i=1,\ldots,N} x_i$	1.0	2	1.0	2
Variance	$\sigma^2 = \frac{1}{N-1}\sum_{i=1,\ldots,N}(x_i - \bar{x})^2$	0.63	1	0.63	1
Standard dev.	σ	~0.8	1	~0.8	1
Covariance	$\overline{cov}(x,y) = \frac{1}{N-1}\sum_{i=1,\ldots,N}(x_i - \bar{x})(y_i - \bar{y})$	0.315			
Pearson's Coeff.	$\bar{\rho}_{x,y} = \frac{\overline{cov}(x,y)}{\bar{\sigma}_x \cdot \bar{\sigma}_y}$	~0.5			
Rank Covar.	$\overline{cov}\big(R(x), R(y)\big)$	0.5			
Spearman's Coeff.	$\bar{\rho}_{R(x),R(y)}$	0.5			
Kendall's Coeff.	$\tau_{x,y}$	1/3			

Exercise:

Calculate the Values!

(~5 Minutes,

with your neighbor)

Example 2: Pearson's vs. Rank Correlation

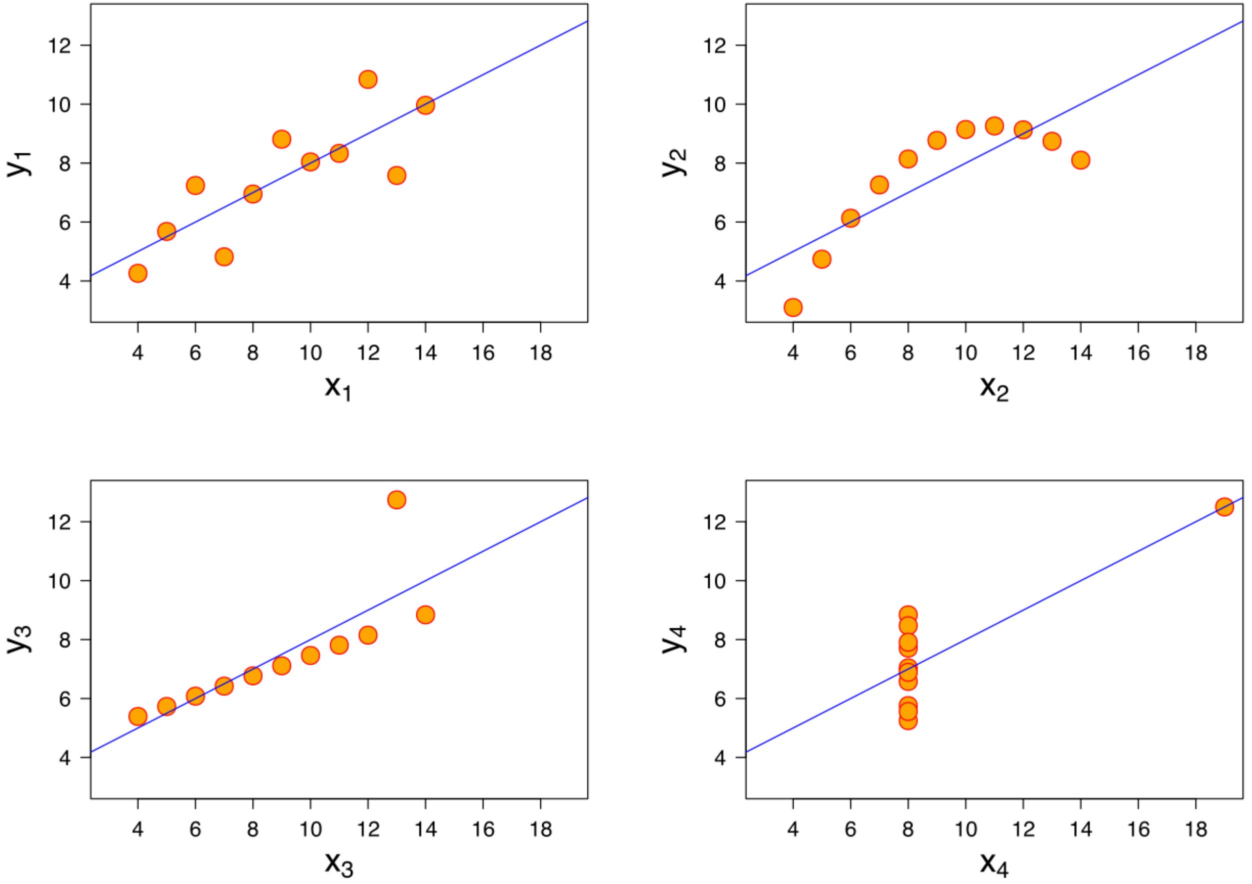
		x	R(x)	y	R(y)
Data		0.4	1	0.1	1
		0.7	2	1.3	2
		1.9	3	1.6	3
Mean	$\bar{x} = \frac{1}{N} \sum_{i=1,\ldots,N} x_i$	1.0	2	1.0	2
Variance	$\bar{\sigma}^2 = \frac{1}{N-1} \sum_{i=1,\ldots,N} (x_i - \bar{x})^2$	0.63	1	0.63	1
Standard dev.	$\bar{\sigma}$	~0.8	1	~0.8	1
Covariance	$\overline{cov}(x, y) = \frac{1}{N-1} \sum_{i=1,\ldots,N} (x_i - \bar{x})(y_i - \bar{y})$		0.495		
Pearson's Coeff.	$\bar{\rho}_{x,y} = \frac{\overline{cov}(x,y)}{\bar{\sigma}_x \cdot \bar{\sigma}_y}$		~0.77		
Rank Covar.	$\overline{cov}\big(R(x), R(y)\big)$		1		
Spearman's Coeff.	$\bar{\rho}_{R(x),R(y)}$		1		
Kendall's Coeff.	$\tau_{x,y}$		1		

Anscombe's quartet

All data sets have the same:

- Mean
- Variance
- Pearson's Correlation
- Regression Line

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



Source: Wikipedia

Let’s take a break…

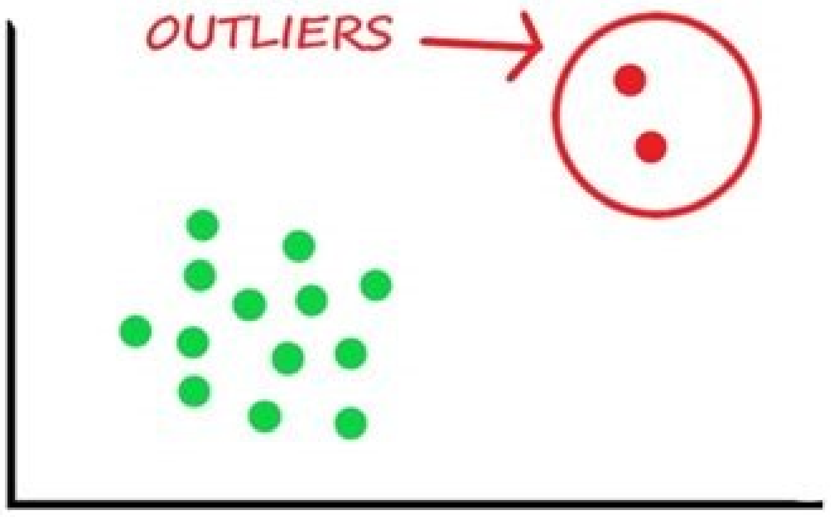
Back on in 5 min!

Takeaway: Q: What have we learned so far?

1. Correlation coefficients show relation between variables / features.
2. Correlation $\neq$ Causation (but Causation $\Longrightarrow$ Correlation)
3. Identifying correlations helps to make better predictions!
4. The correlation coefficients can give different results:
   - Pearson's Coeff. considers *linear* correlations
   - Rank coeff. disrespect the actual values (consider only their *ordinal relations*)
5. It is important to plot your data!

… Identifying (and eliminating) outliers can help finding correlations!

Outliers (Anomalies)



Outlier = data point that looks different than the rest of the data

Causes:

- Faulty sensor, mistakes in entering data.
- Special events (holiday, extreme weather, …)
- Purposeful manipulations (fraud, strategic behaviour, …)

Effect:

- Model fitting / training is skewed
- Error metrics are inflated

Anomaly Detection

Application Examples:

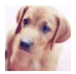
- Fraud Detection:
  Identify unusual behavior of users.

- Predictive Maintenance:
  Identify machines that might break soon.

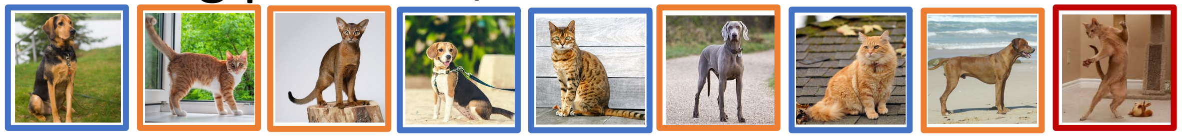
- Monitoring:
  Recognise changes in behavior.

Machine Learning - Overview

- Supervised Learning
  - Learning a function from *labeled* training data: Classification & Regression

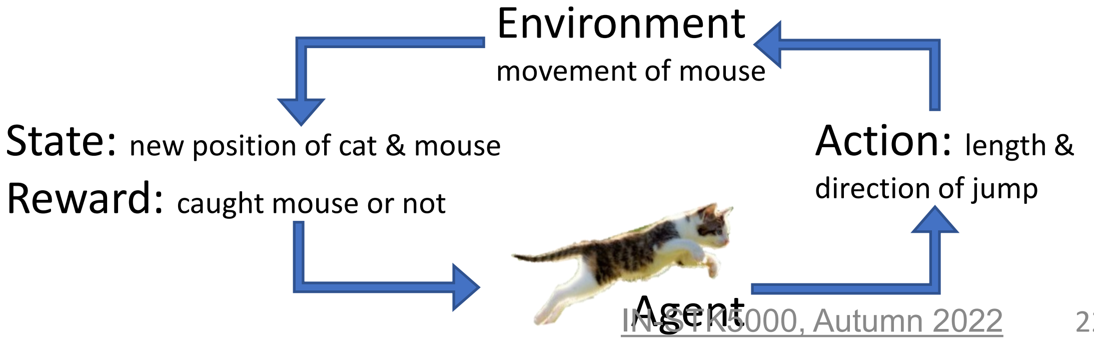
→ Function: F()= DOG

- Unsupervised Learning
  - Learning patterns / structure from *unlabeled* data



→ Clusters: STANDING SITTING ?

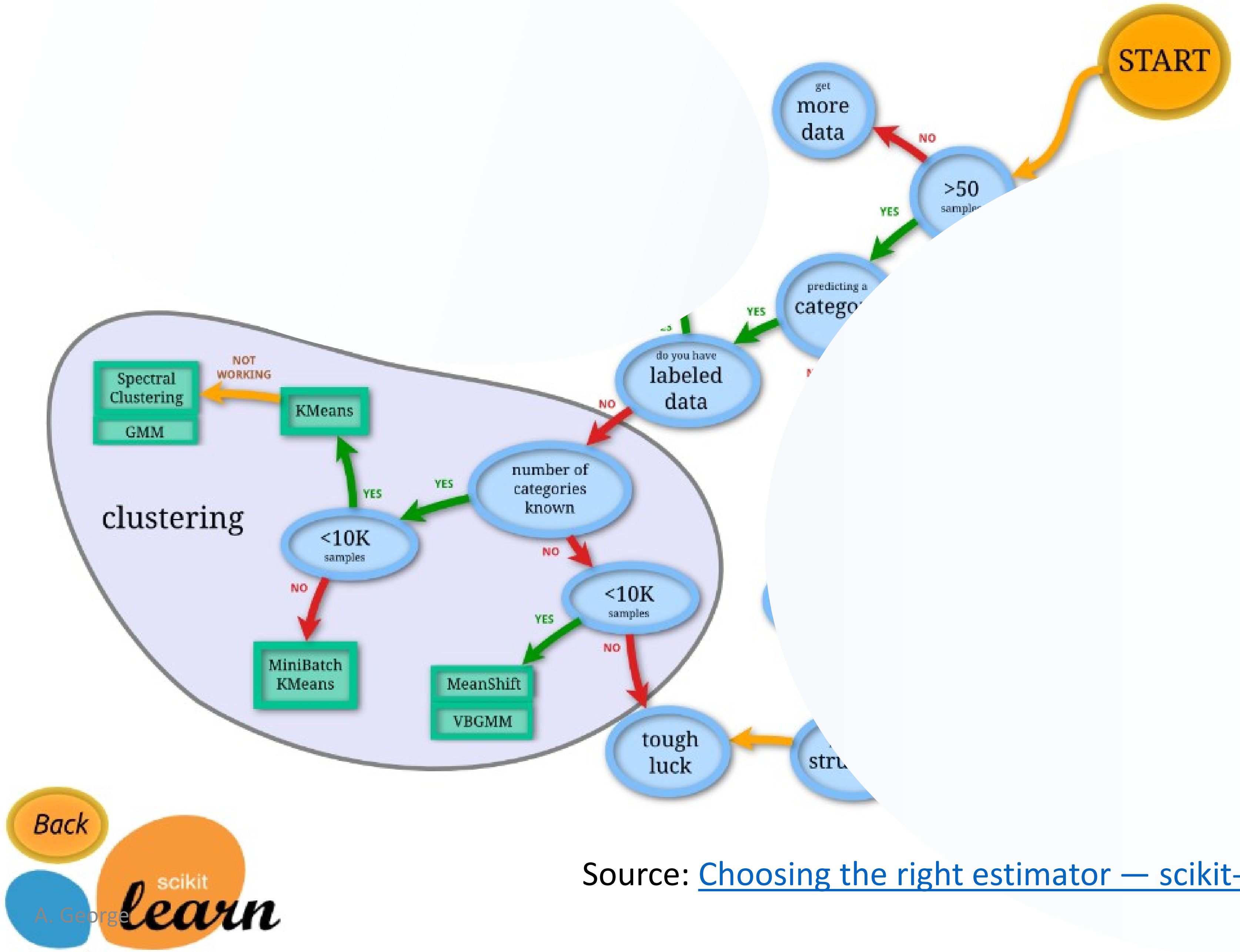
- Reinforcement Learning
  - Learning good actions from *feedback* [interactive!]



Clustering

Standing **Sitting** **???**

scikit-learn algorithm cheat-sheet



Source: [Choosing the right estimator — scikit-learn 1.1.1 documentation]

Density-based method: Clustering → k-Means

<u>Tipps</u>:
- Run k-means several times to avoid local optima.
- Try out different k and compare outcomes.

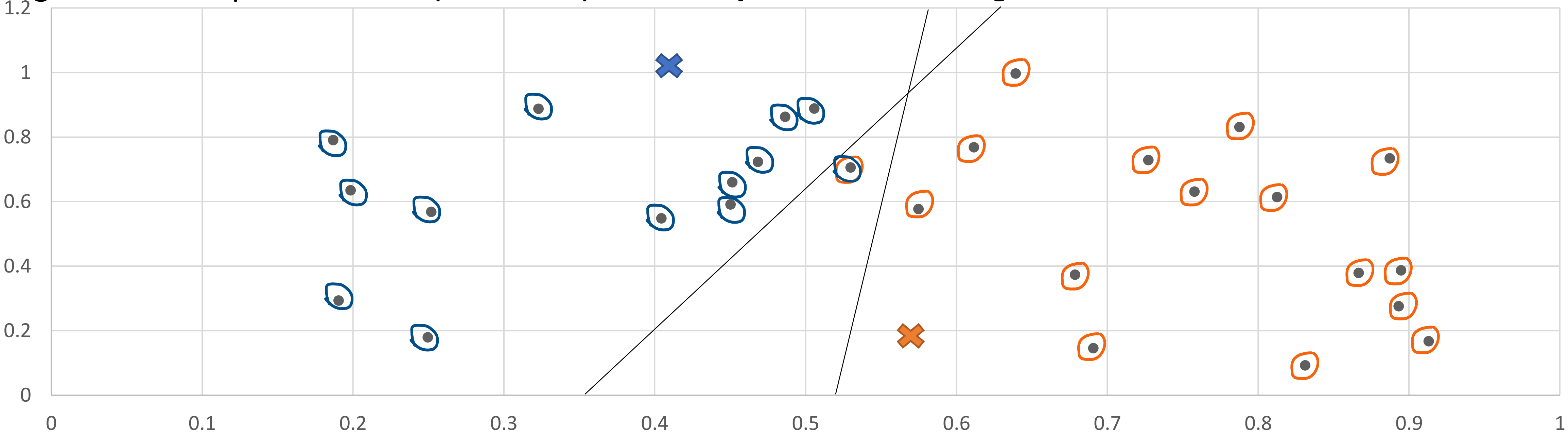
Outlier: Has a "high" distance to its cluster center

In practice: See, e.g. [Scikit Learn documentation]

Algorithm Example: $\boldsymbol{k}$-Means (for $\boldsymbol{k=2}$)

0. Insert k random cluster *centroids* **(e.g. on k data points)**

1. Repeat: Cluster Assignment + Move Centroid to cluster average



Statistical Method: Z-Score (also Standard Score)

- Data: Samples $x_1, \dots, x_N \in \mathbb{R}$

Restriction:
Data has only one feature!

SAMPLE-		
	Mean	$\bar{x} = \frac{1}{N} \sum_{i=1,\dots,N} x_i$
	Variance	$\bar{\sigma}^2 = \frac{1}{N-1} \sum_{i=1,\dots,N} (x_i - \bar{x})^2$
	Standard dev.	$\bar{\sigma}$

- Z-Score: $z_i = \frac{x_i - \bar{x}}{\bar{\sigma}}$

→ $z_1, \dots, z_N$ have mean 0 and standard deviation 1
→ Usually used when x follows a Normal distribution.

- Bounds: User-defined lower bound l and upper bound u

$$x_i \text{ is outlier} \iff z_i < l \quad \text{or} \quad z_i > u$$

Statistical Method: Distance to the Mean

- Data: Samples $x_1, \dots, x_N \in \mathbb{R}^d$

Sample-Mean	$\bar{x} = \frac{1}{N} \sum_{i=1,\dots,N} x_i \quad \in \mathbb{R}^d$

- Distances: $d_i = \sqrt{\sum_{j=1,\dots,N} \left(x_{i,j} - \bar{x_j}\right)^2}$ → Euclidean dist. to mean

  $\bar{d}$ = mean, $\bar{\sigma}$ = st. dev.

- Z-Score: $z_i = \frac{d_i - \bar{d}}{\bar{\sigma}}$
- Bounds: User-defined upper bound u

  x_i is outlier $\Leftrightarrow \quad z_i > u$

Other Methods / Tools

- "Looking at the data": Box Plot
- Support Vector Machines
- Neural Networks
- ... *(see Wikipedia for a more comprehensive list,*
  or Scikit Learn for some methods used in practice)

Outliers!

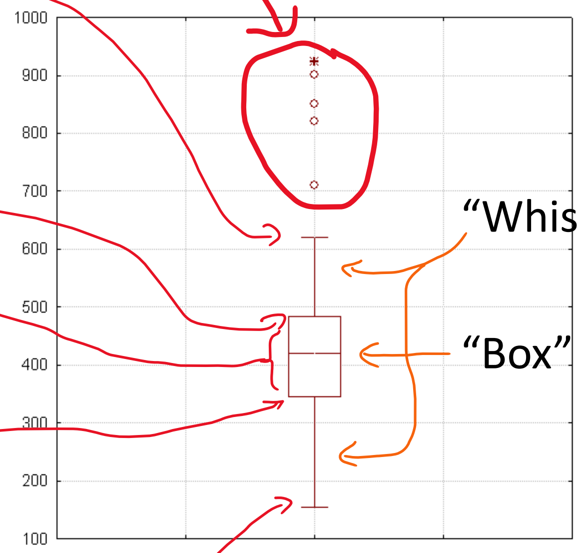
First value ≤ Q3+1.5 IRQ

Q3 = 3rd quartile:
75% of the data has
a value ≤ than this

IRQ (interquartile length)

Q1 = 1st quartile:
25% of the data has
a value ≤ than this

First value ≥ Q1-1.5 IRQ



"Whiskers"

"Box"

The Ice Cream Van

Data on the last few ice cream sales shows:
When you play a jingle, you sell a lot of ice cream!

… so how do we check causation???

Causation

<u>Definition</u>: A discussion in metaphysics / philosophy… A “causes” B if:

→ A and B are correlated

→Time dependency: A appears before B (in time)

→ Counterfactual notion: B occurs if and only if A has occurred.

“You pass you MSc *iff* you have passed your defense.”

→<u>Probabilistic</u> notion: If A occurs, B’s is likelier to occur.

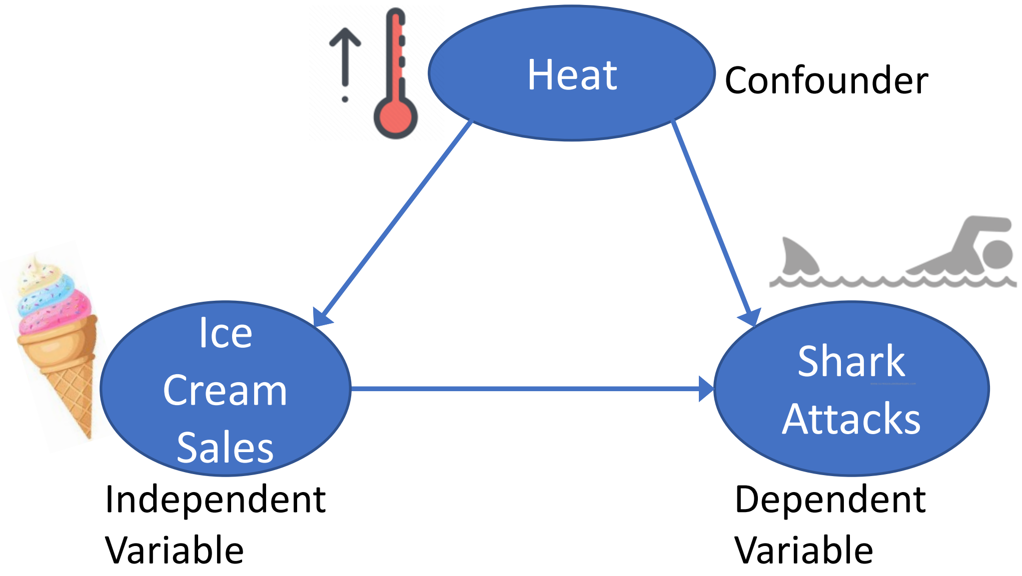
“If you smoke you are at *increased risk* of getting cancer.”

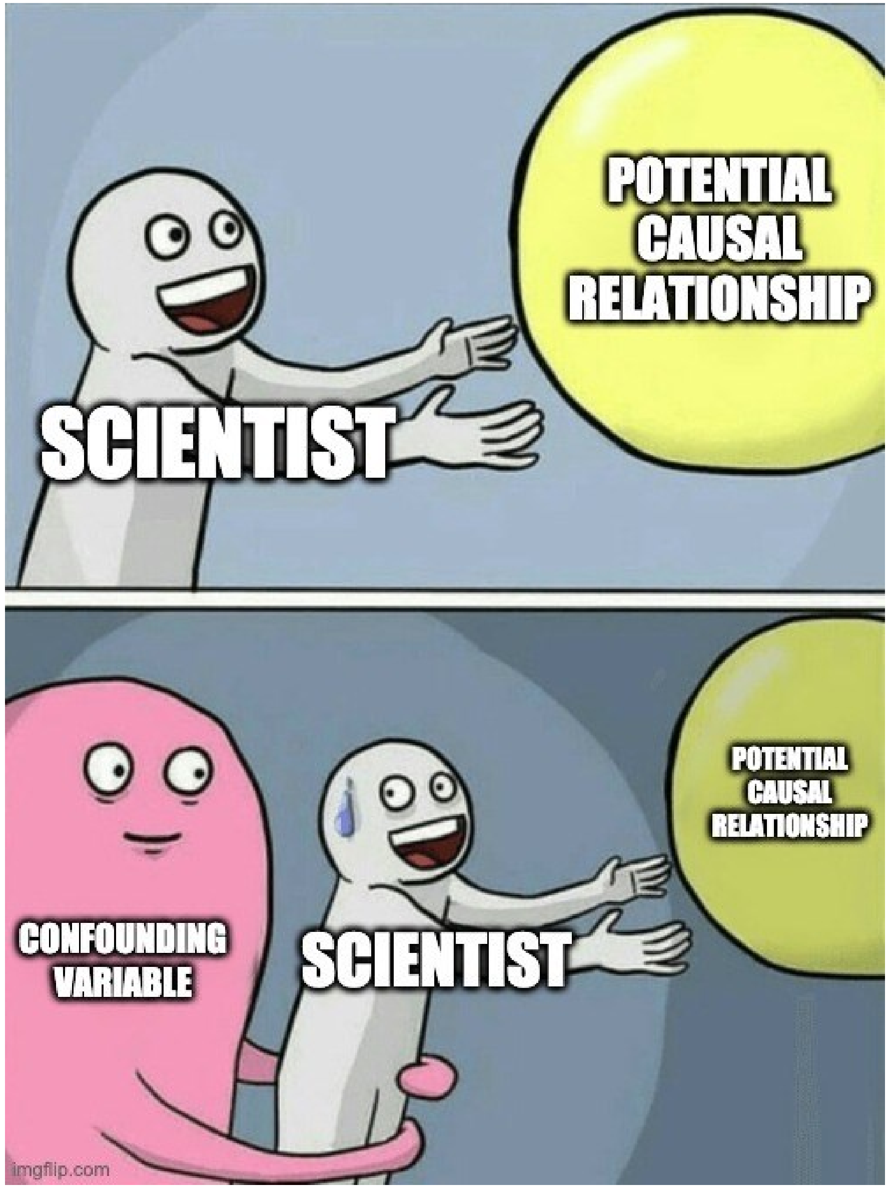
$$P(B|A) > P(B)$$

→ Check this by <u>Hypothesis Testing</u> methods!

Confounders

→ Heat *confounds* the relation between Ice Cream Sales and Shark Attacks, since it causally influences both!





Source: [Twitter]

Null Hypothesis Testing

- Null Hypothesis: H_0 The hypothesis we want to test.

  "A does not have a causal effect on B", i.e., $P(B|A) = P(B)$

- Alternate Hypothesis: Negation of null-hypothesis

  "A has a causal effect on B", i.e., $P(B|A) \neq P(B)$

Desired value for $P(rejecting\ H_0 \mid H_0\ true)$. Typically, $\alpha = 0.05$.

- Perform a Hypothesis Test: t-test, Z-test, Chi-sq. … → get p-value

  Probability of obtaining data at least as extreme, given that the null hypothesis is true.

  - If $p < \alpha$: Reject the null hypothesis!

    (result significant) (Enough evidence to say that "A has a causal effect on B"!)

  - If $p \geq \alpha$: Cannot reject the null hypothesis!

    (result not significant) (Not enough evidence to say whether A has a causal effect on B, or not!)

Correlation of Categorical Features

- Samples $(x_1, y_1), \dots, (x_N, y_N)$ with categorical domains $\mathcal{X}, \mathcal{Y}$, e.g., Y/N
- Null-Hypothesis H_0: “x and y are NOT correlated.“ (variables are independent)
- <u>Chi-Square Test</u>:
  1. Calculate observations $O_{ab} = \#(both\ a\ \&b)$ and expectations $E_{ab} = \frac{\#a \cdot \#b}{Total}$
  2. Calculate $X^2 = \sum_{a \in \mathcal{X}, b \in \mathcal{Y}} \frac{(O_{ab} - E_{ab})^2}{E_{ab}}$ and p-value (→accept/reject H_0).

Data = O / E / $\frac{(O-E)^2}{E}$	Sweden Democrats	Social Dem. Party	Other	Total
Male	27.5 % 18.75 4.08	22.5 % 28.75 1.36	50.0 % 52.50 0.12	100 %
Female	10.0 % 18.75 4.08	35.0 % 28.75 1.36	55.0 % 52.50 0.12	100 %
Total	37.5 %	57.5 %	105 %	200 %

Calculate X^2, look up the p-value … or just use, e.g. **chi2_contingency()** from **scipy.stats**!

Alternative Approaches

- Testing correlation between two (numerical / categorical) features:
  - If there is a correlation, then we could predict one from the other
  - Fit a classifier!
  - If it “works well” then features are correlated

  → Careful: You need to know how to evaluate your classifier!
  → Next week: Evaluation metrics, fairness & privacy + guest lecture on GDPR

- Good Reads: [1, 2, 3, 4, 5, 6, 7, …]
- Other Material: Christos Dimitrakakis’ lectures from last year!

Summary

Correlation

Covariance:

$$cov(x,y) = \mathbb{E}_p[x \cdot y] - \mathbb{E}_{p_y}[x] \cdot \mathbb{E}_{p_x}[y]$$

Pearson's correlation:

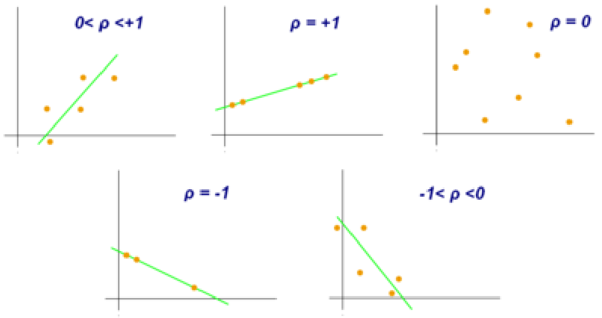
$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y}, \text{ if } \sigma_x \cdot \sigma_y > 0$$

Spearman's Rank Correlation Coefficient:

$$\rho_{R(x),R(y)} = \frac{cov(R(x),R(y))}{\sigma_{R(x)} \cdot \sigma_{R(y)}}$$

Kendall's Rank Correlation Coefficient:

$$\tau_{x,y} = \frac{\#\,concordant\ pairs - \#\,discordant\ pairs}{\#\,all\ pairs}$$



Basics

Variance:

$$\mathbb{V}[x] = var(x) = \mathbb{E}_p[x^2] - \left(\mathbb{E}_p[x]\right)^2$$

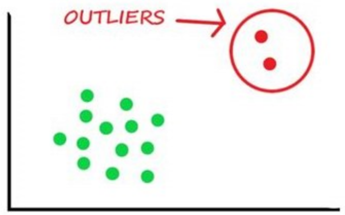
Standard Deviation: $\sigma_x = \sqrt{\mathbb{V}[x]}$

Outliers / Anomalies

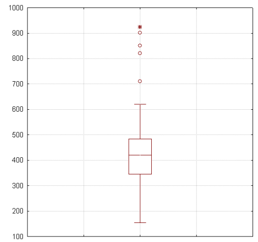
Z-Score: $z_i = \frac{x_i - \bar{x}}{\bar{\sigma}}$, with mean $\bar{x}$, var. $\bar{\sigma}$

x_i outlier $\Leftrightarrow z_i \notin [l,u]$

Clustering:



Box-Plot:



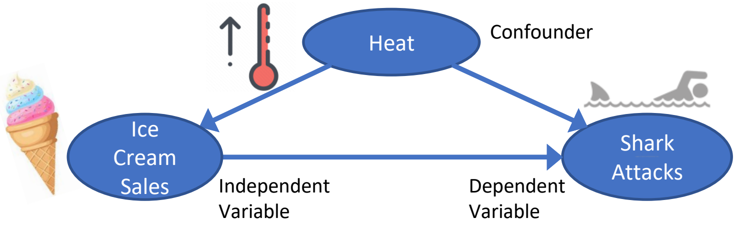
Causation

Necessary Conditions:

Time: a occurs before b

Correlation: a and b are correlated

Non-Confounded: $\nexists c$ that causes a and b



Hypothesis Testing:

Null-Hypothesis H_0: "no correlation"

Alt. Hypothesis H_1: "exists correlation"

1. Do desired statistic, e.g. Chi-Square
2. Determine significance (p-value < α)
3. Accept / reject the hypothesis.

→ Applicable for determining correlations