# Correlation, Model Combination

**Jindřich Libovický** (reusing materials by Milan Straka)

November 20, 2025

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

# Today's Lecture Objectives

After this lecture you should be able to

- Explain and implement different ways of measuring correlation: Pearson's correlation, Spearman's correlation, Kendall's $\tau$.
- Decide if correlation is a good metric for your model.
- Measure inter-annotator agreement and draw conclusions for data cleaning and for limits of your models.
- Use correlation with human judgment to validate evaluation metrics.

# Covariance

# Covariance

Given a collection of random variables $\mathrm{x}_1, \ldots, \mathrm{x}_N$, we know that

$$\mathbb{E}\left[\sum\nolimits_i \mathrm{x}_i\right] = \sum_i \mathbb{E}\big[\mathrm{x}_i\big].$$

But how about $\mathrm{Var}\left(\sum_i \mathrm{x}_i\right)$?

$$\begin{aligned}
\mathrm{Var}\left(\sum\nolimits_i \mathrm{x}_i\right) &= \mathbb{E}\left[\left(\sum\nolimits_i \mathrm{x}_i - \sum\nolimits_i \mathbb{E}[\mathrm{x}_i]\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum\nolimits_i \big(\mathrm{x}_i - \mathbb{E}[\mathrm{x}_i]\big)\right)^2\right] \\
&= \mathbb{E}\left[\sum\nolimits_i \sum\nolimits_j \big(\mathrm{x}_i - \mathbb{E}[\mathrm{x}_i]\big)\big(\mathrm{x}_j - \mathbb{E}[\mathrm{x}_j]\big)\right] \\
&= \sum_i \sum_j \mathbb{E}\left[\big(\mathrm{x}_i - \mathbb{E}[\mathrm{x}_i]\big)\big(\mathrm{x}_j - \mathbb{E}[\mathrm{x}_j]\big)\right].
\end{aligned}$$

# Covariance

We define **covariance** of two random variables $\mathrm{x}, \mathrm{y}$ as

$$\operatorname{Cov}(\mathrm{x}, \mathrm{y}) = \mathbb{E}\Big[\big(\mathrm{x} - \mathbb{E}[\mathrm{x}]\big)\big(\mathrm{y} - \mathbb{E}[\mathrm{y}]\big)\Big].$$

Then,

$$\operatorname{Var}\left(\sum\nolimits_i \mathrm{x}_i\right) = \sum_i \sum_j \operatorname{Cov}(\mathrm{x}_i, \mathrm{x}_j).$$

Note that $\operatorname{Cov}(\mathrm{x}, \mathrm{x}) = \operatorname{Var}(\mathrm{x})$ and that we can write covariance as

$$\begin{aligned}
\operatorname{Cov}(\mathrm{x}, \mathrm{y}) &= \mathbb{E}\Big[\big(\mathrm{x} - \mathbb{E}[\mathrm{x}]\big)\big(\mathrm{y} - \mathbb{E}[\mathrm{y}]\big)\Big] \\
&= \mathbb{E}\big[\mathrm{x}\mathrm{y} - \mathrm{x}\mathbb{E}[\mathrm{y}] - \mathbb{E}[\mathrm{x}]\mathrm{y} + \mathbb{E}[\mathrm{x}]\mathbb{E}[\mathrm{y}]\big] \\
&= \mathbb{E}\big[\mathrm{x}\mathrm{y}\big] - \mathbb{E}\big[\mathrm{x}\big]\mathbb{E}\big[\mathrm{y}\big].
\end{aligned}$$

# Correlation

# Correlation

Random variables $\mathrm{x}, \mathrm{y}$ are **uncorrelated** if $\operatorname{Cov}(\mathrm{x}, \mathrm{y}) = 0$; otherwise, they are **correlated**.

Note that two *independent* random variables are uncorrelated, because

$$\begin{aligned}
\operatorname{Cov}(\mathrm{x}, \mathrm{y}) &= \mathbb{E}\Big[\big(\mathrm{x} - \mathbb{E}[\mathrm{x}]\big)\big(\mathrm{y} - \mathbb{E}[\mathrm{y}]\big)\Big] \\
&= \sum_{x,y} P(x, y)\big(x - \mathbb{E}[x]\big)\big(y - \mathbb{E}[y]\big) \\
&= \sum_{x,y} P(x)\big(x - \mathbb{E}[x]\big) P(y)\big(y - \mathbb{E}[y]\big) \\
&= \Big(\sum_{x} P(x)\big(x - \mathbb{E}[x]\big)\Big)\Big(\sum_{y} P(y)\big(y - \mathbb{E}[y]\big)\Big) \\
&= \mathbb{E}_{\mathrm{x}}\big[\mathrm{x} - \mathbb{E}[\mathrm{x}]\big]\mathbb{E}_{\mathrm{y}}\big[\mathrm{y} - \mathbb{E}[\mathrm{y}]\big] = 0.
\end{aligned}$$

However, dependent random variables can be uncorrelated – random uniform $\mathrm{x}$ on $[-1, 1]$ and $\mathrm{y} = |\mathrm{x}|$ are not independent ($\mathrm{y}$ is completely determined by $\mathrm{x}$), but they are uncorrelated.

# Pearson correlation coefficient

There are several ways to measure correlation of random variables $\mathrm{x}, \mathrm{y}$.

**Pearson correlation coefficient**, denoted as $\rho$ or $r$, is defined as

$$\rho \stackrel{\text{def}}{=} \frac{\operatorname{Cov}(\mathrm{x}, \mathrm{y})}{\sqrt{\operatorname{Var}(\mathrm{x})}\sqrt{\operatorname{Var}(\mathrm{y})}}$$

$$r \stackrel{\text{def}}{=} \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}},$$

where:

- $\rho$ is used when the full expectation is computed (population Pearson correlation coefficient);
- $r$ is used when estimating the coefficient from data (sample Pearson correlation coefficient);
  - $\bar{x}$ and $\bar{y}$ are sample estimates of the respective means.

# Pearson correlation coefficient

The value of Pearson correlation coefficient is in fact normalized covariance, because its value is always bounded by $-1 \leq \rho \leq 1$ (and the same holds for $r$).

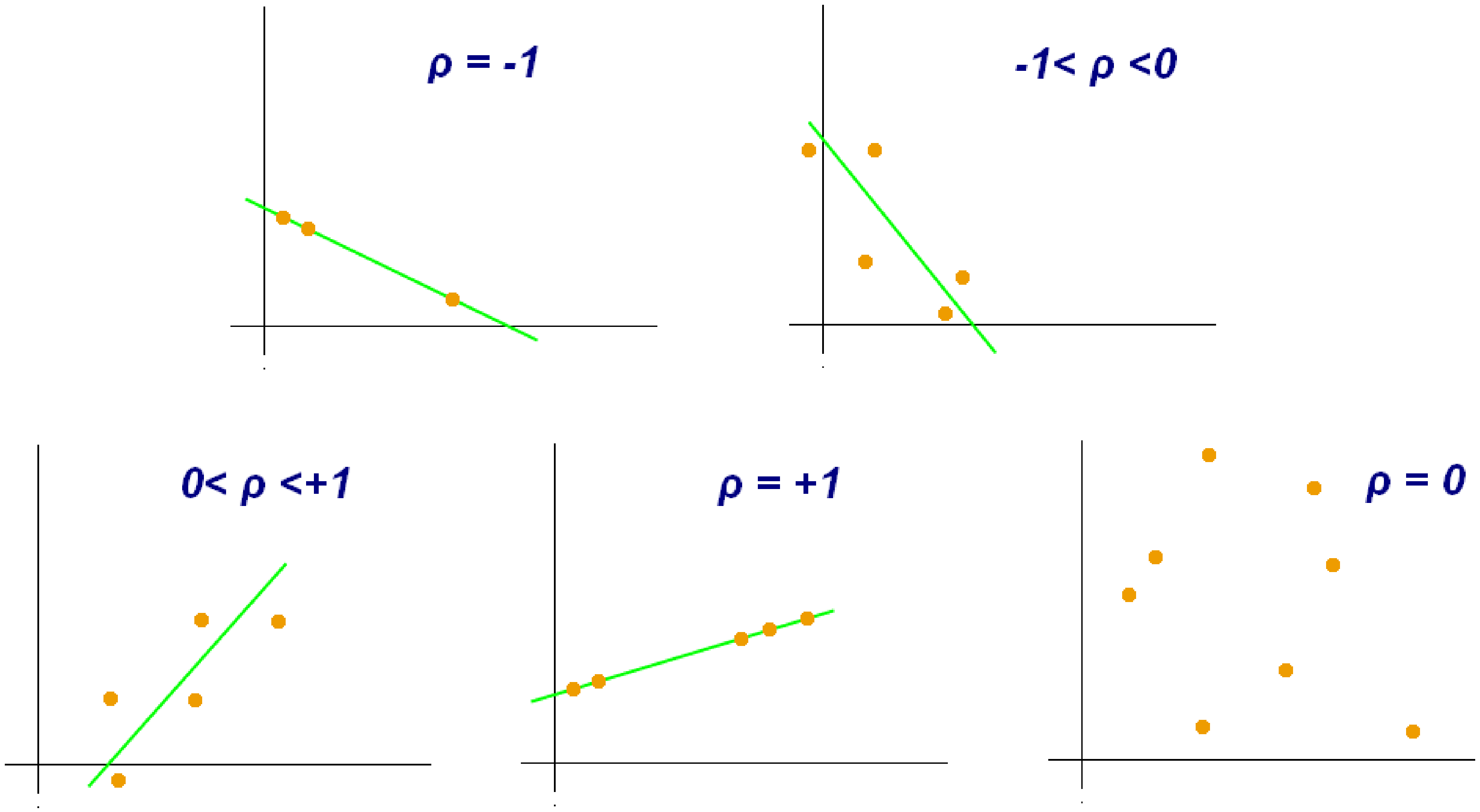
The bound can be derived from

$$
\begin{aligned}
0 &\leq \mathbb{E}\left[\left(\frac{(\mathrm{x}-\mathbb{E}[\mathrm{x}])}{\sqrt{\mathrm{Var}(\mathrm{x})}} - \rho\frac{(\mathrm{y}-\mathbb{E}[\mathrm{y}])}{\sqrt{\mathrm{Var}(\mathrm{y})}}\right)^2\right] \\
&= \mathbb{E}\left[\frac{(\mathrm{x}-\mathbb{E}[\mathrm{x}])^2}{\mathrm{Var}(\mathrm{x})}\right] - 2\rho\mathbb{E}\left[\frac{(\mathrm{x}-\mathbb{E}[\mathrm{x}])}{\sqrt{\mathrm{Var}(\mathrm{x})}}\frac{(\mathrm{y}-\mathbb{E}[\mathrm{y}])}{\sqrt{\mathrm{Var}(\mathrm{y})}}\right] + \rho^2\mathbb{E}\left[\frac{(\mathrm{y}-\mathbb{E}[\mathrm{y}])^2}{\mathrm{Var}(\mathrm{y})}\right] \\
&= \frac{\mathrm{Var}(\mathrm{x})}{\mathrm{Var}(\mathrm{x})} - 2\rho\cdot\rho + \rho^2\frac{\mathrm{Var}(\mathrm{y})}{\mathrm{Var}(\mathrm{y})} = 1-\rho^2,
\end{aligned}
$$

which yields $\rho^2 \leq 1$.

# Pearson correlation coefficient

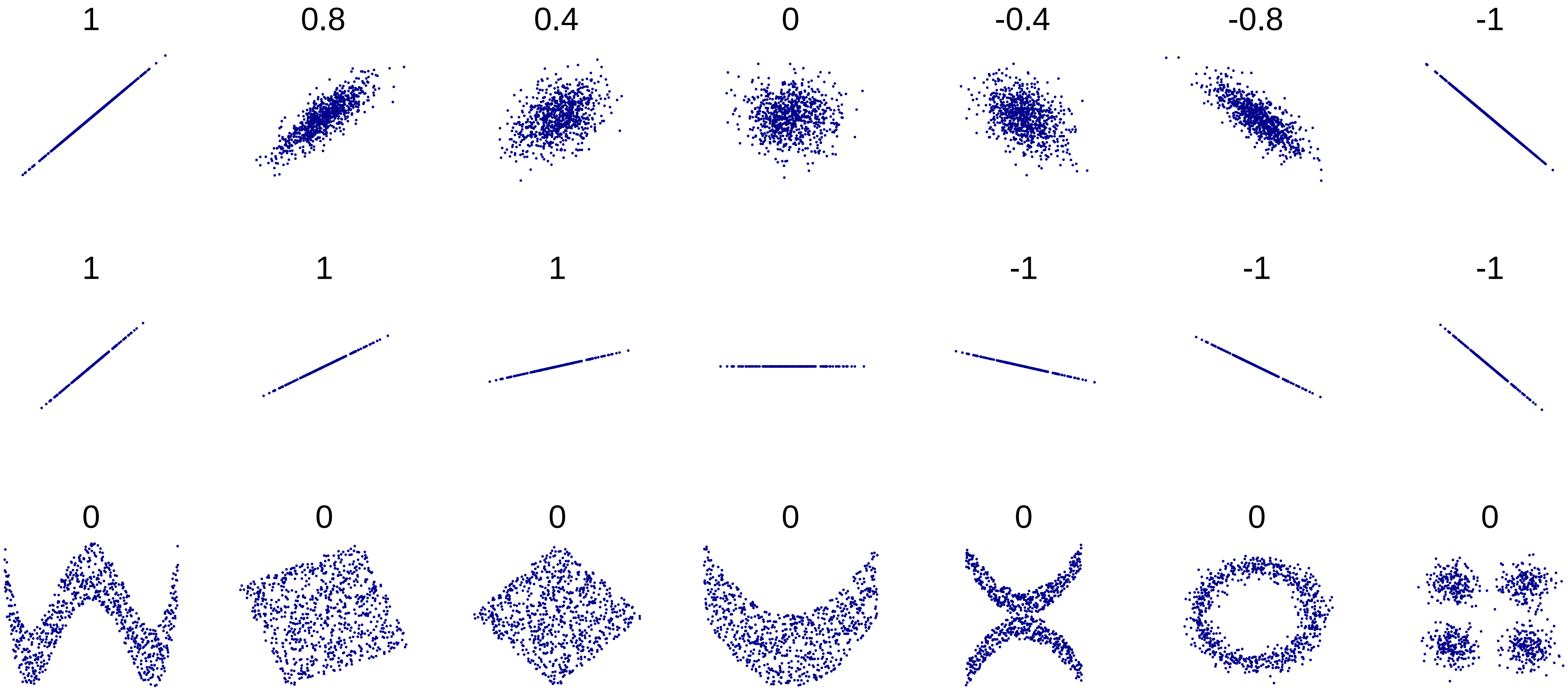
Pearson correlation coefficient quantifies **linear dependence** of the two random variables.




*https://upload.wikimedia.org/wikipedia/commons/3/34/Correlation_coefficient.png*

# Pearson correlation coefficient

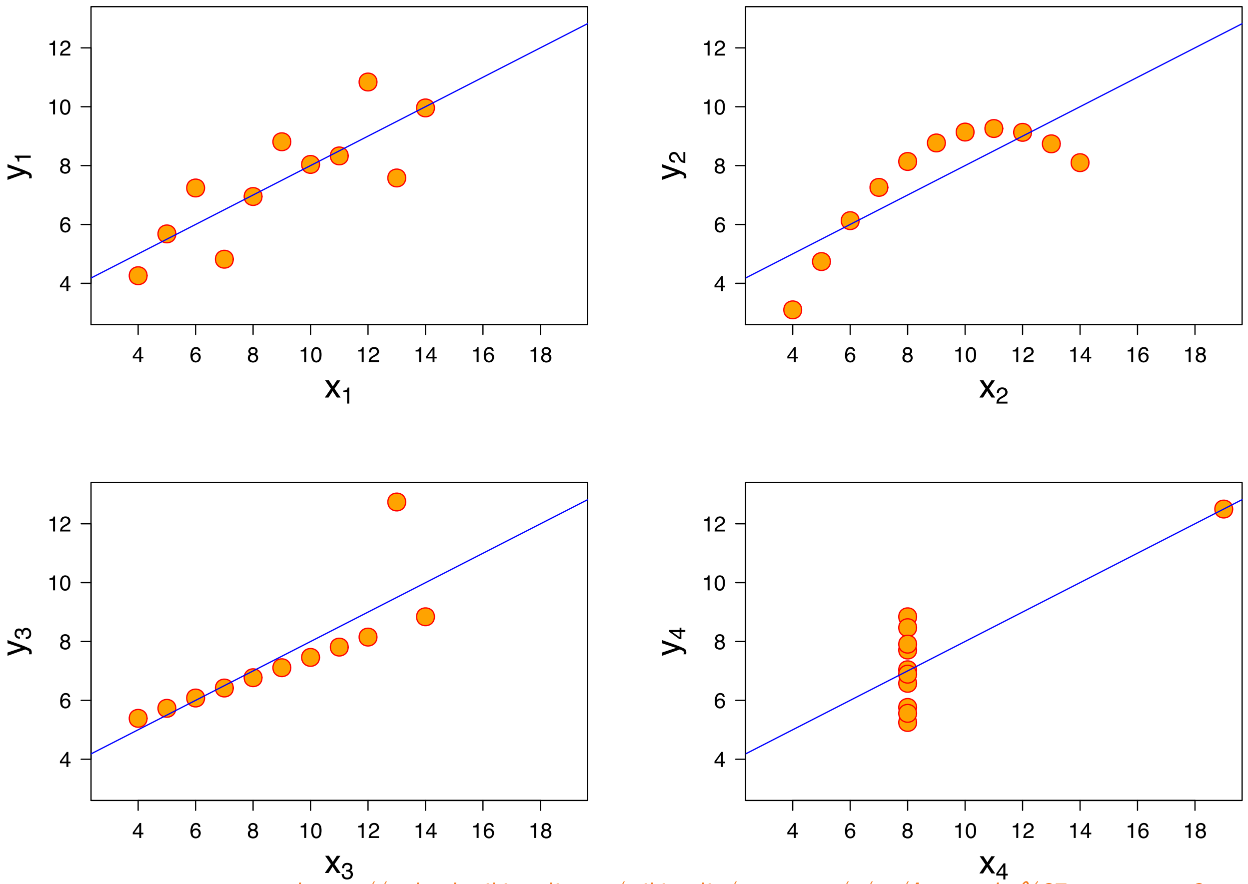
Pearson correlation coefficient quantifies **linear dependence** of the two random variables.




*https://upload.wikimedia.org/wikipedia/commons/d/d4/Correlation_examples2.svg*

# Pearson correlation coefficient

The four displayed variables have the same mean 7.5, variance 4.12, Pearson correlation coefficient 0.816 and regression line $3 + \frac{1}{2}x$.

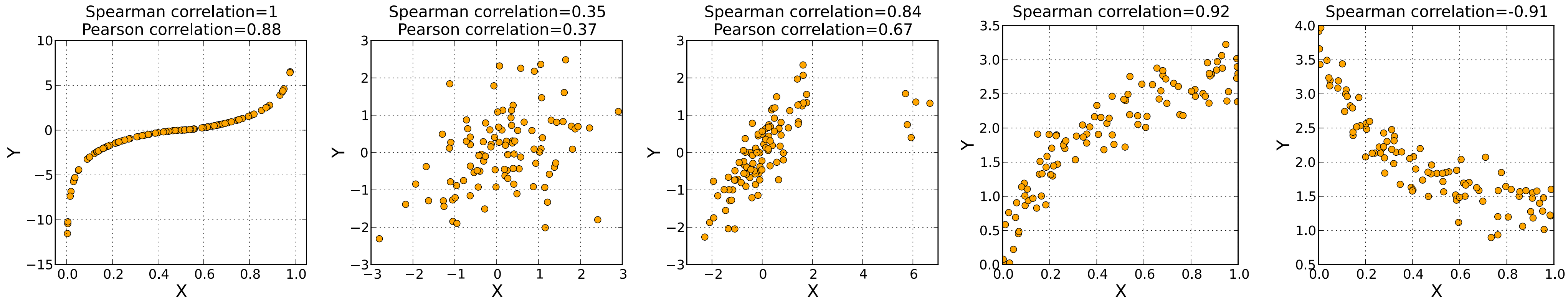



*https://upload.wikimedia.org/wikipedia/commons/e/ec/Anscombe%27s_quartet_3.svg*

To measure also nonlinear correlation, two coefficients are commonly used.

## Spearman's rank correlation coefficient $\rho$

Spearman's $\rho$ is Pearson correlation coefficient measured on **ranks** of the original data, where a rank of an element is its index in sorted ascending order.




*https://upload.wikimedia.org/wikipedia/commons/4/4e/Spearman_fig{1,2,3,5,4}.svg*

### Kendall rank correlation coefficient $\tau$

Kendall's $\tau$ measures the amount of *concordant pairs* (pairs where $y$ increases/decreases when $x$ does), minus the *discordant pairs* (where $y$ increases/decreases when $x$ does the opposite):

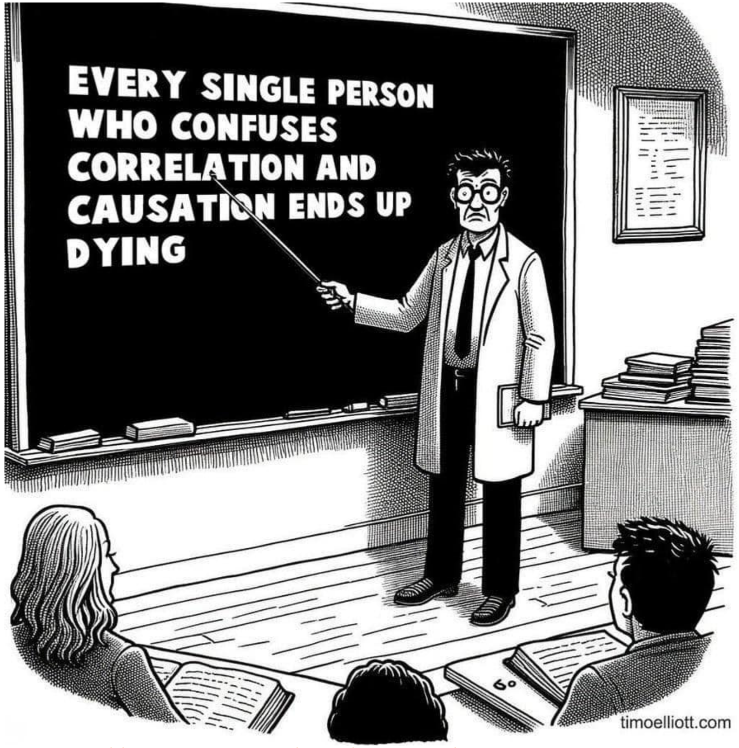
$$\tau \stackrel{\text{def}}{=} \frac{\left|\{\text{pairs } i \neq j : x_j > x_i, y_j > y_i\}\right| - \left|\{\text{pairs } i \neq j : x_j > x_i, y_j < y_i\}\right|}{\binom{n}{2}}$$

$$= \frac{\sum_{i<j} \operatorname{sign}(x_j - x_i) \operatorname{sign}(y_j - y_i)}{\binom{n}{2}}.$$

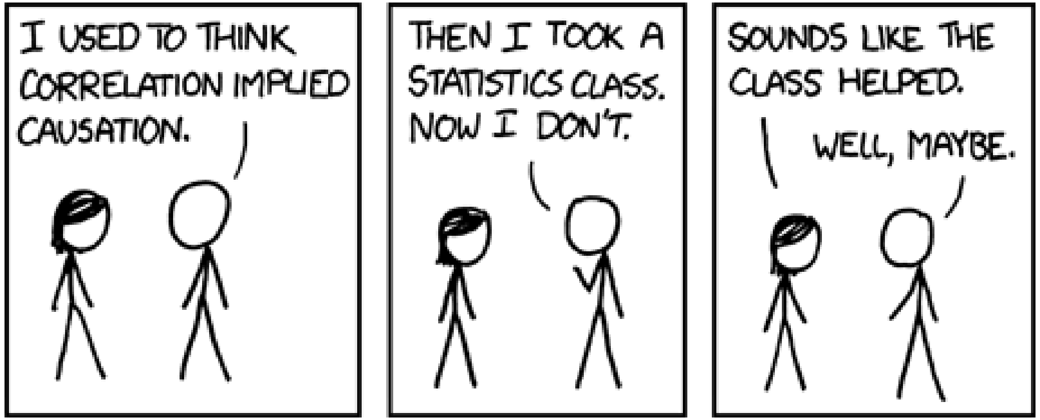
There is no clear consensus on whether to use Spearman's $\rho$ or Kendall's $\tau$. When there are no/few ties in the data, Kendall's $\tau$ offers two minor advantages – $\frac{1+\tau}{2}$ can be interpreted as a probability of a concordant pair, and Kendall's $\tau$ converges to a normal distribution faster.

As defined, the range of Kendall's $\tau \in [-1, 1]$. However, if there are ties, its range is smaller – therefore, several corrections (not discussed here) exist to adjust its value in case of ties.

# Correlation is not causation




https://timoelliott.com/blog/cartoons/yet-more-analytics-cartoons




https://xkcd.com/552/

# Correlation in Machine Learning

# Use of Correlation in Machine Learning

In ML, correlation is commonly used as

- Evaluation metric for some tasks;
- Measuring data annotation quality;
- Assessing the quality of automatic metrics by comparing them to human judgment.

# Correlation as evaluation metric

- Learning to rank (e.g., document retrieval): we do not care about the actual values
  - Kendall's $\tau$, Spearman's correlation
  - When we want the correct items to rank before incorrect ones: precision (assuming fixed top-$k$, typically at 5, 10), recall (often ill-defined), mean reciprocal rank

$$\mathrm{MRR} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\text{rank of the first relevant item}}$$

- Evaluating pair similarity: word embeddings, sentence embeddings
  - Similarity estimates from psycholinguistic experiments: scores for word/sentence pairs
  - Measure Pearson/Spearman correlation between embedding distances and similarity scores

# Inter-annotator agreement (1)

- Inter-annotator agreement can tell us
  - How well defined the task is
  - How reliable annotators/user ratings are
  - What data items are suspicious / difficult

- For continuous target values: Pearson's/Spearman's correlation
- For classification tasks: Cohen's $\kappa$

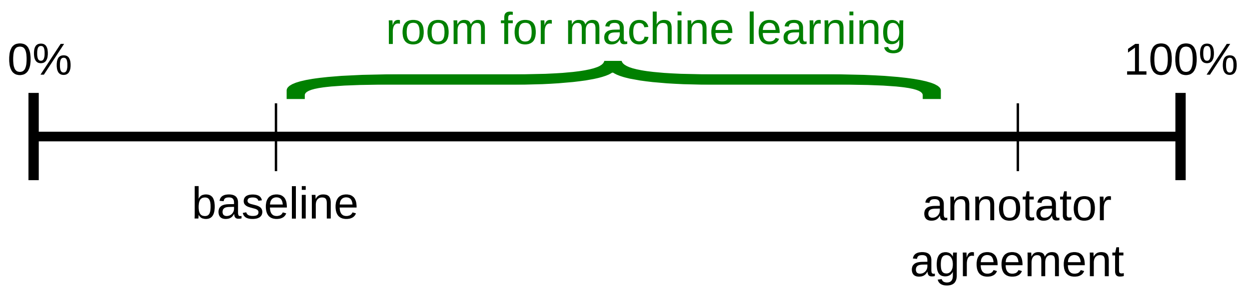
  $p_O$ is observed agreement, $p_E$ expected agreement by chance

$$\kappa = \frac{p_O - p_E}{1 - p_E}$$


*https://www.surgehq.ai/blog/the-pitfalls-of-inter-rater-reliability-in-data-labeling-and-machine-learning*

- Can be used to filter out confusing data points and unreliable annotators
- Not all outliers are noise! Low IAA can reveal cultural differences.

IAA sets natural upper boundary for ML performance. Performance over IAA is suspicious!

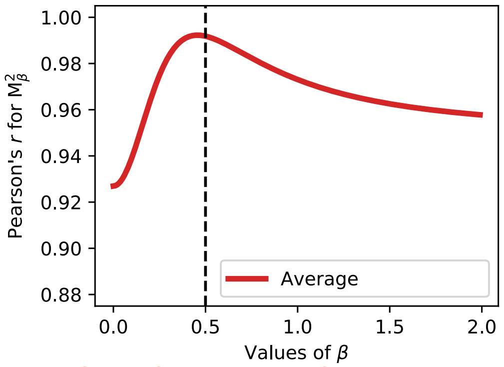



- Trivial baseline for classification: majority class, for regression average, or something based on simple rules
- Performance over IAA is more likely overfitting for the way the data is curated than super-human performance.

# Correlation with human judgment

For some tasks, it might not be clear how to measure the model performance:

**Grammar checking**: the $\beta$ parameter




*J. Náplava, M. Straka, J. Straková, and A. Rosen. 2022. Czech Grammar Error Correction with a Large and Diverse Corpus. In TAACL, 10:452–467.*

**Machine translation**: evaluation is subjective by definition, we design metrics to correlate with human judgment.

- SoTA machine translation metrics are typically machine-learned.
- Different metrics might be suitable for different tiers of translation quality.
- There is an annual competition in MT quality and MT metric quality.

# Today's Lecture Objectives

After this lecture you should be able to

- Explain and implement different ways of measuring correlation: Pearson's correlation, Spearman's correlation, Kendall's $\tau$
- Decide if correlation is a good metric for your model
- Measure inter-annotator agreement and draw conclusions for data cleaning and for limits of your models
- Use correlation with human judgment to validate evaluation metrics