Organisation

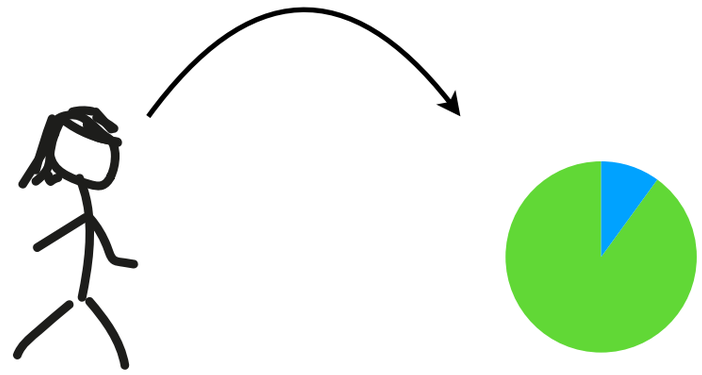
- 9.5. Introduction to biostatistics (Anna)
- 16.5. Descriptive statistics (Anna)
- 23.5. Hypothesis testing (Anna)
- 6.6. Introduction machine learning (Robert)
- 13.6. Machine learning (Melissa)
- 20.6. Deep learning (Robert)
- 21.6. Dimensionality Reduction (Melissa)
- 11.7. Summary (all)

Introduction into Biostatistics

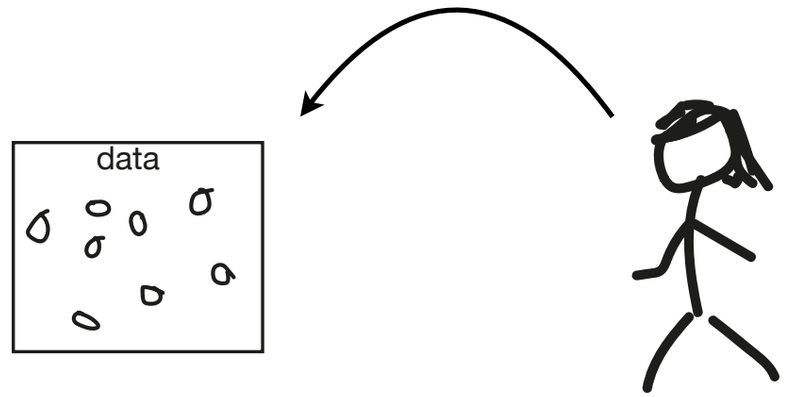
Anna Poetsch, Biotechnology Center, TU Dresden

Recap on probability

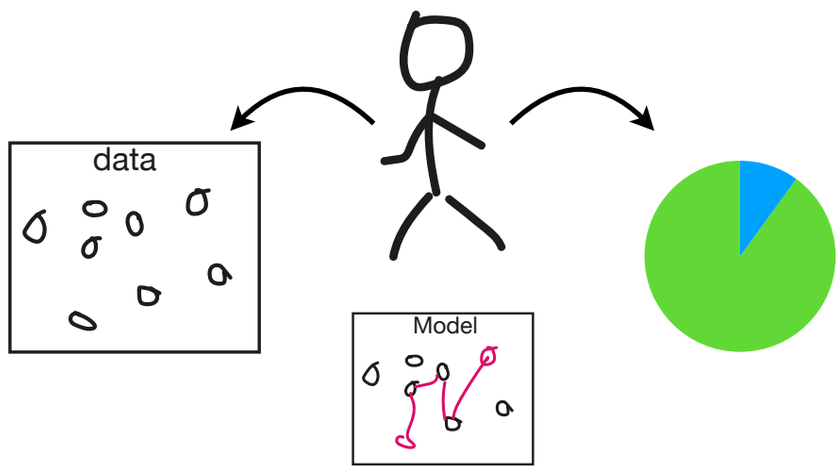
A model



Data

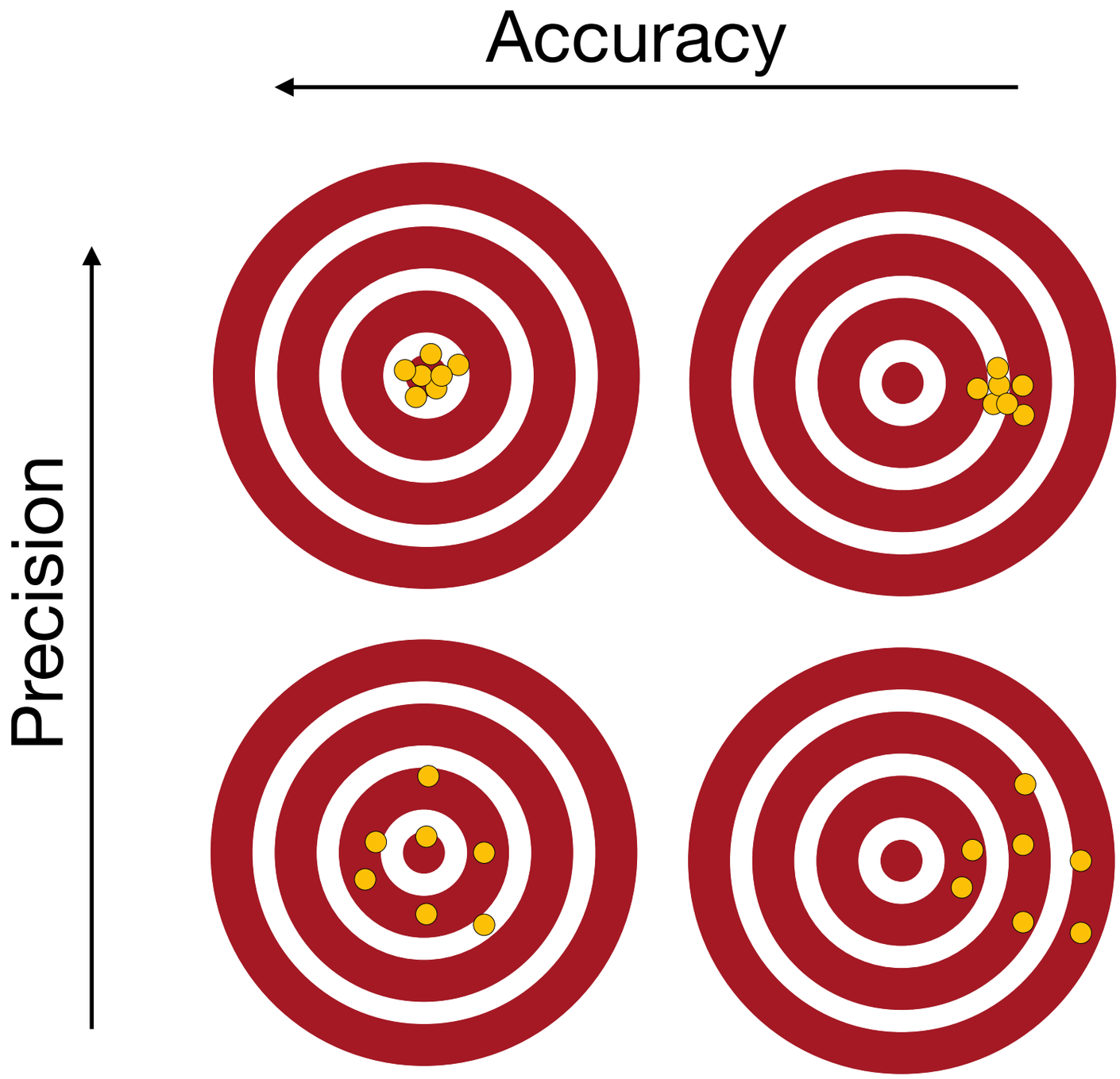


A model based on data



! **Estimating probabilities can only be as good as your assumptions/ data**

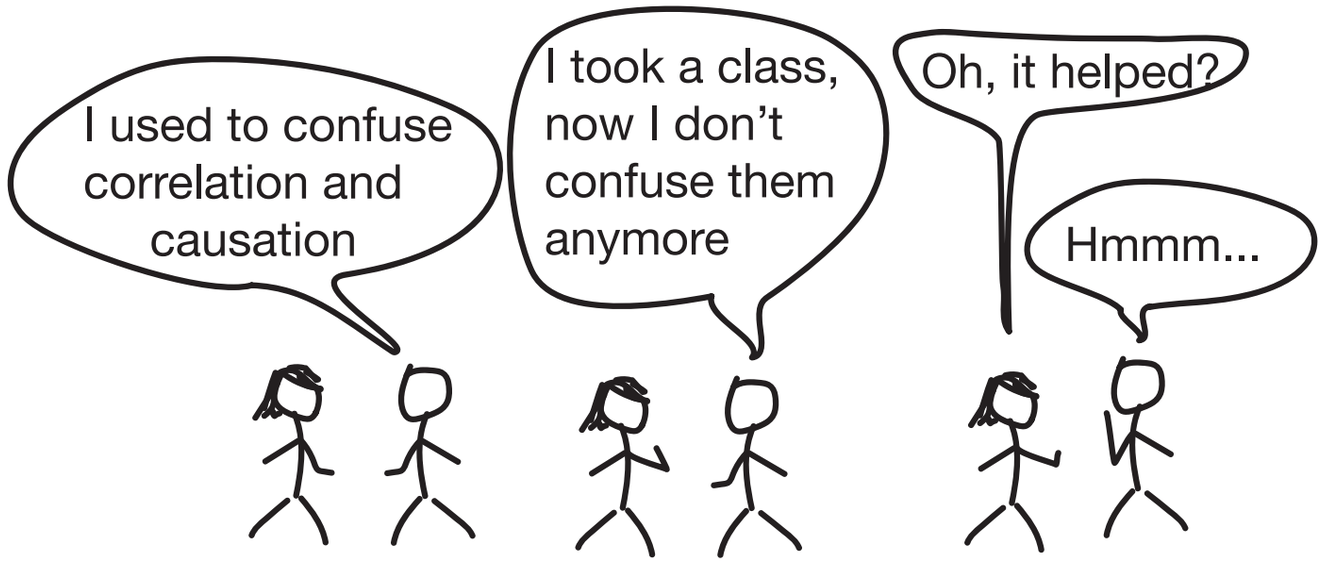
Recap on accuracy and precision



How do these relate to confidence intervals?

Does the confidence interval get bigger, if you increase n?

- We tend to be overconfident
- We tend to jump to conclusions
- We see patterns in random data
- We don’t realise that coincidences are common
- We don’t expect variability to depend on sample size
- We are fooled by multiple comparisons
- We intuitively follow logic that is in fact dictated by regression to the mean
- We are biased
- We confuse correlation with causation



We crave easy explanations,

follow intuitions,

and aim for certainty.

Statistics offers probabilities!

Be aware of reversing probabilities!

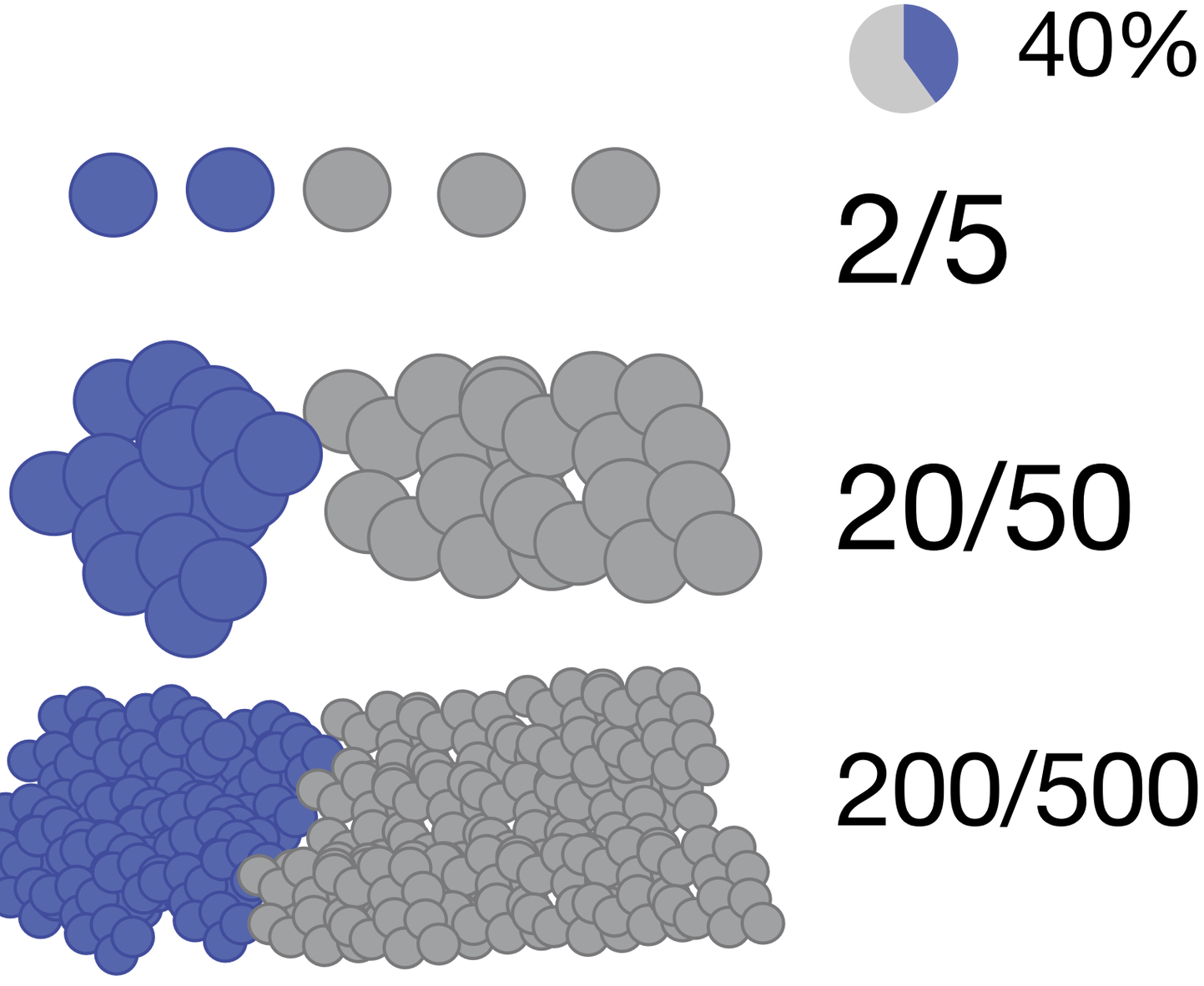
Most COVID19 patients have a cough, it is still unlikely that someone who coughs in a tram will have COVID19

If you find that statistics books are boring, it does not mean that every boring book is about statistics

When you find a good cancer drug, you can kill your cell culture with it.
If your cells die, you might want to think for a moment before you ring Stockholm.

Know your problem, know your distribution!

Binomial



40%

2/5

20/50

200/500

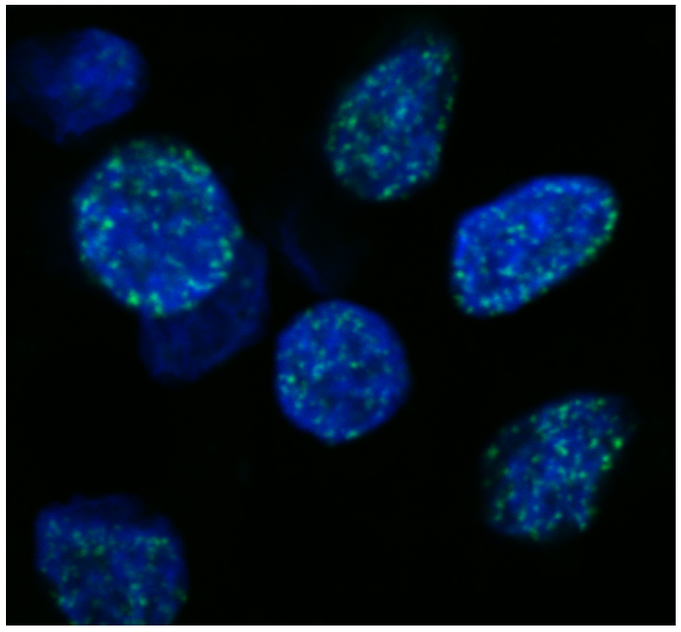
Confidence increases with n

Normal

Confidence does not increase with n

Counted data/ Poisson distribution

- Radioactive decay
- Raisins in a Dresdner Stollen
- Mutations in a genome
- Imaging foci in a cell



Discrete variables

Ordinal variables

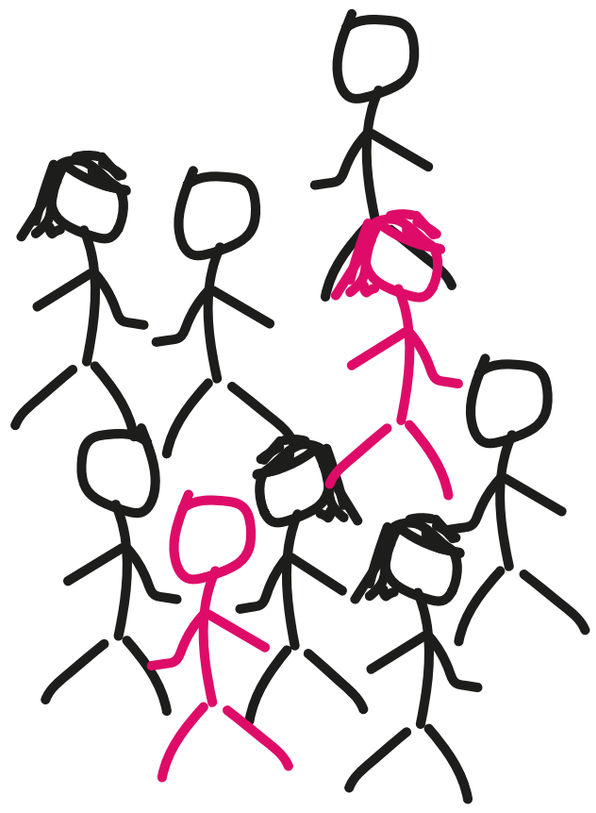
- limited set of discrete values with order

e.g. scale from 1-10

Nominal, binomial variables

- limited set of discrete values without order

e.g. responder <-> non responder



Continuous variables

Interval variables

- continuous value, for which intervals make sense, but no ratios

e.g. °C

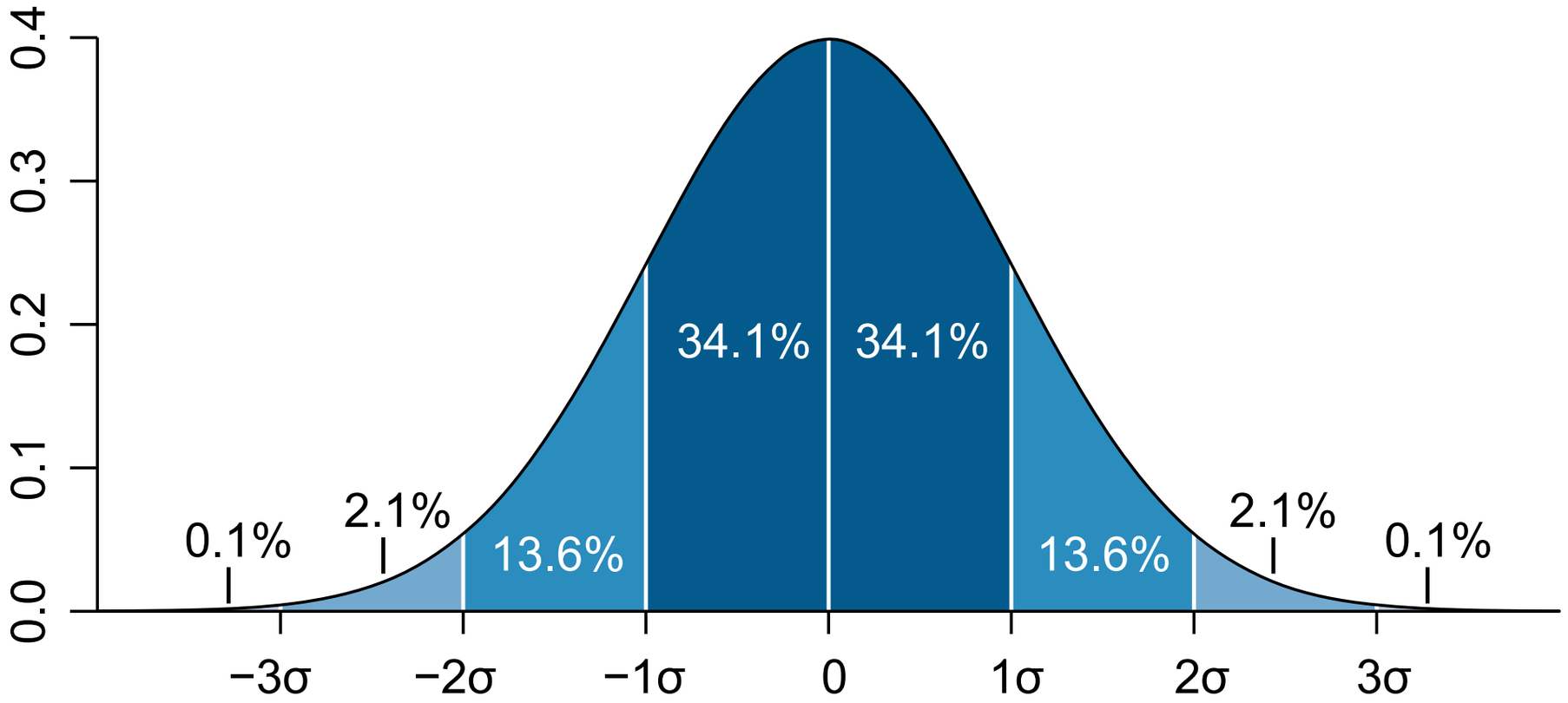
Ratio variables

- continuous value, for which ratios make sense

e.g. height, weight, enzyme activity, Kelvin

Normal distribution

Gaussian distribution, bell-shaped distribution



The result of general imprecision: weighing, pipetting, randomness

Therefore also: height, weight

Density defined by mean and standard deviation

Hypotheses in the statistical sense

Innocent until proven guilty!

-> at first sight counterintuitive…

0-Hypothesis:

“The Astra Zeneca vaccine does not protect from COVID-19 in > 65 yo”

Test: Can we reject it?

A few months ago: No

Does it mean that it is not protective? No - we just don’t know!

A few months later, H_0 can be rejected

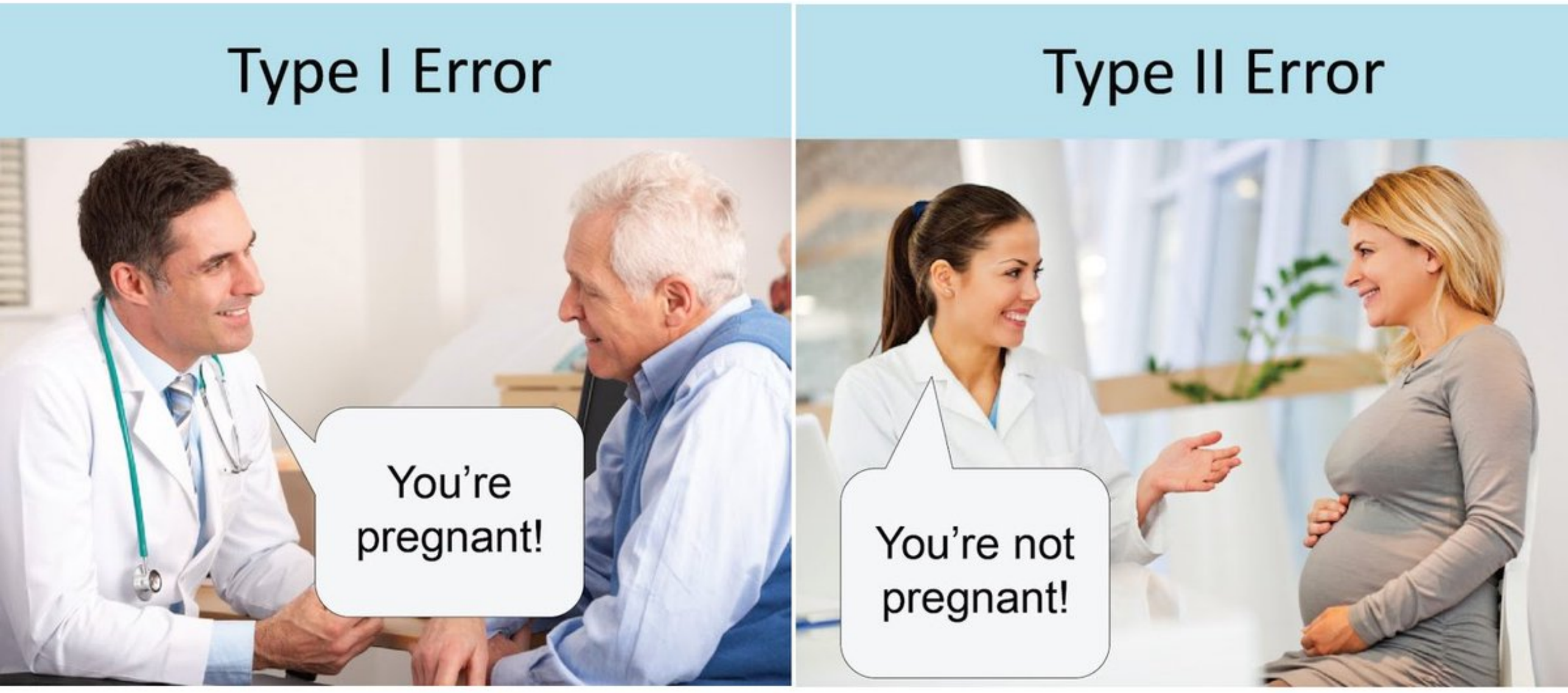
How to reject H_0

How much probability do you allow yourself to be wrong?

- last line chemotherapy treatment: Every bit of hope counts

- vaccination side-effects: Even rare events can be too much

What can go wrong?



False positive

False negative

P-values

The probability that you reject H_0 by chance.

Other ways to phrase it:

The probability that two samples are declared different although they belong to the same population.

The probability of observing a difference as large as you see it (or larger), if the samples are indeed from the same population.

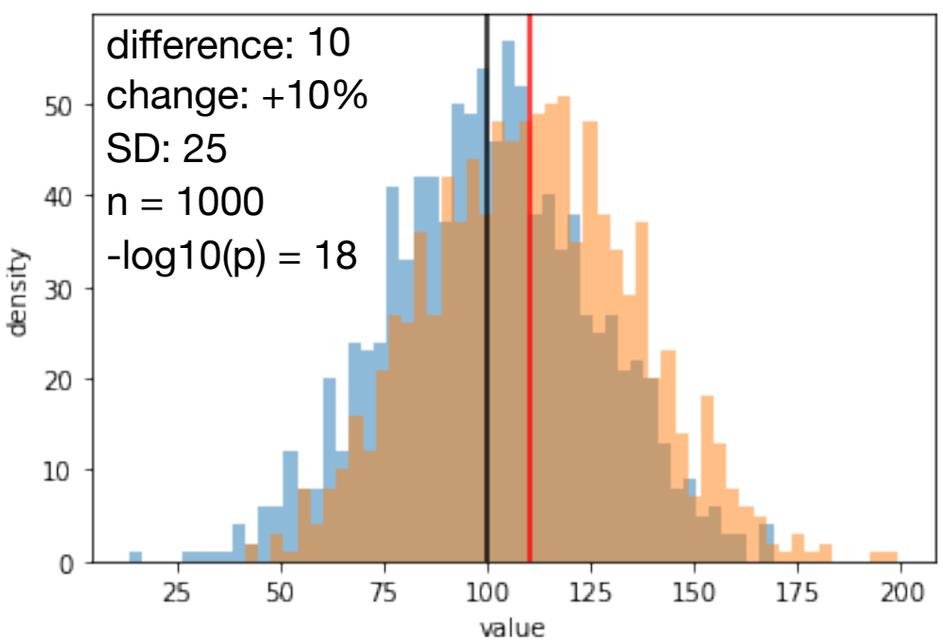
Misconceptions about P-values

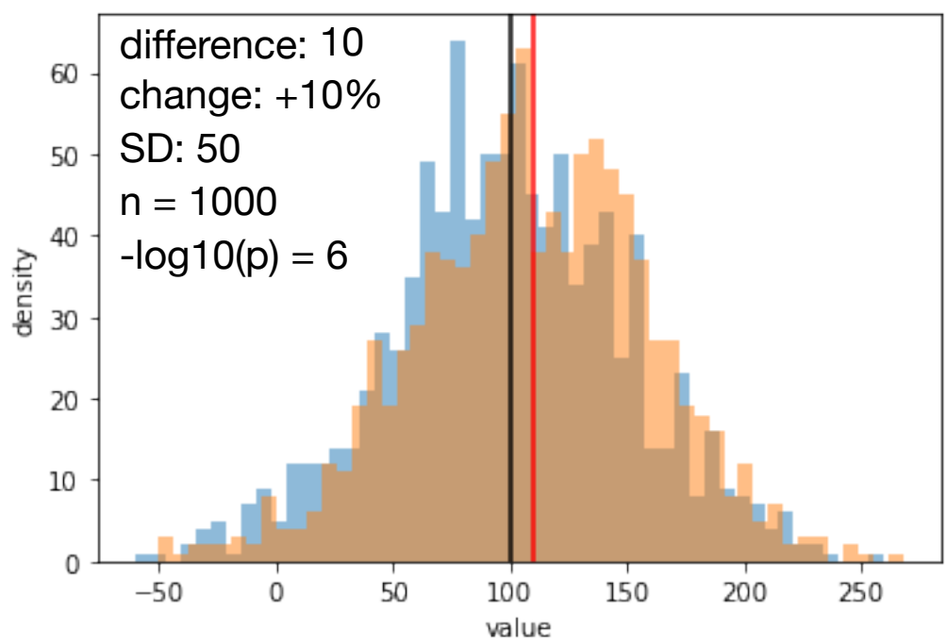
The logic is not reversible:

A p-value of 0.05 means a 5% chance concluding on a difference by chance. Don’t try to interpret the 95%!

You cannot determine whether H_0 is true.

A p-value is not appropriate to conclude about the magnitude of a difference





Misconceptions about P-values

Reproducibility of p-values is inherently very poor.

For measures of reproducibility of an effect the appropriate measure is the effect size, e.g. the actual difference or ratio.

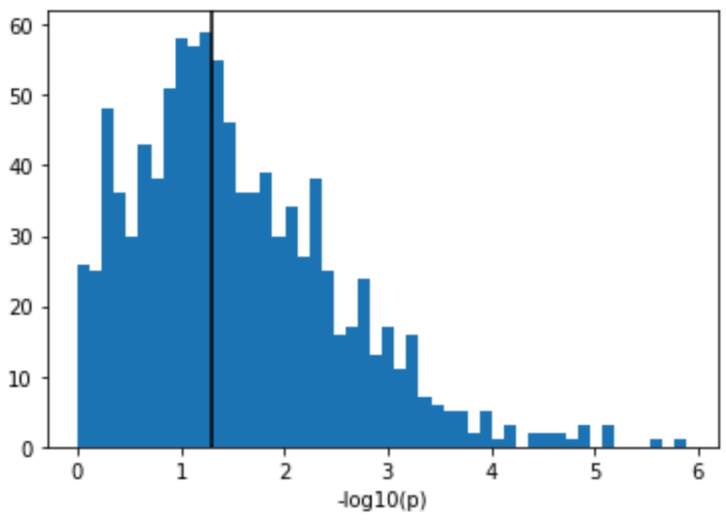
In [3]:
```
pvals = []

for _ in range (1000):
    np.random.seed()
    samp1 = np.random.normal(mean1, sd, n)
    samp2 = np.random.normal(mean2, sd, n)
    new_test = st.ttest_ind(samp1,samp2)
    pvals.append(new_test[1])

#Can you simplify the loop?

#Lets plot this
plt.hist(-np.log10(pvals),bins=50)
plt.axvline(-np.log10(0.05), color="black")
plt.xlabel("-log10(p)")
```

Out[3]: Text(0.5, 0, '-log10(p)')



How to perform the actual hypothesis testing

Do we know the distribution?

-> Parametric testing

-> Fit a distribution to our data and compare whether

the two groups are sufficiently different

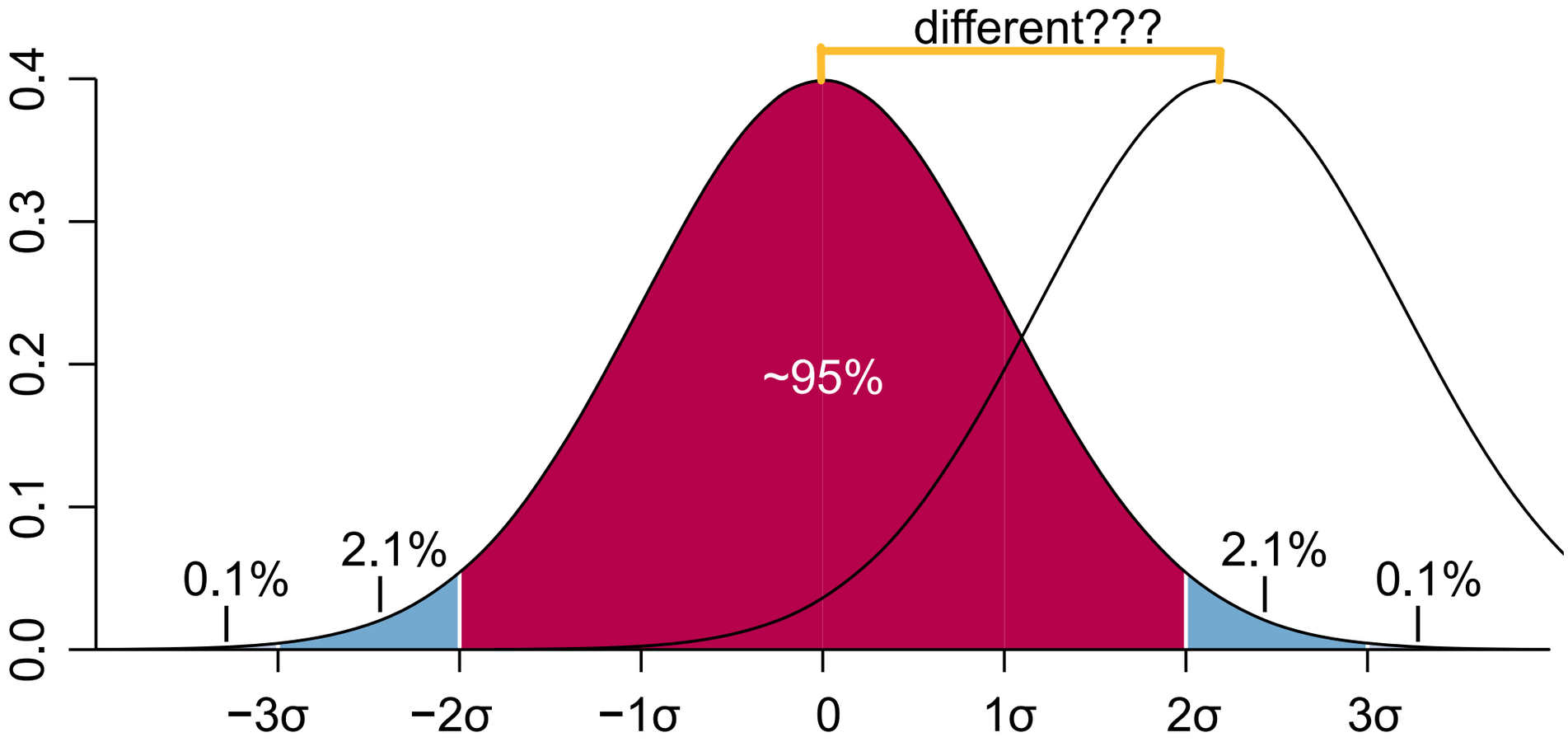
Do we not know the distribution?

-> Non-parametric testing

-> Determine the ranks of our datapoint and look

whether the ranks are sufficiently unbalanced

Assumptions for unpaired parametric statistical testing



Assumptions:

- Our data follow a certain distribution
- They are representative samples
- Independent observations
- Accurate data

The graph is adapted from: M. W. Toews - Own work, based (in concept) on figure by Jeremy Kemp, on 2005-02-09, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=1903871

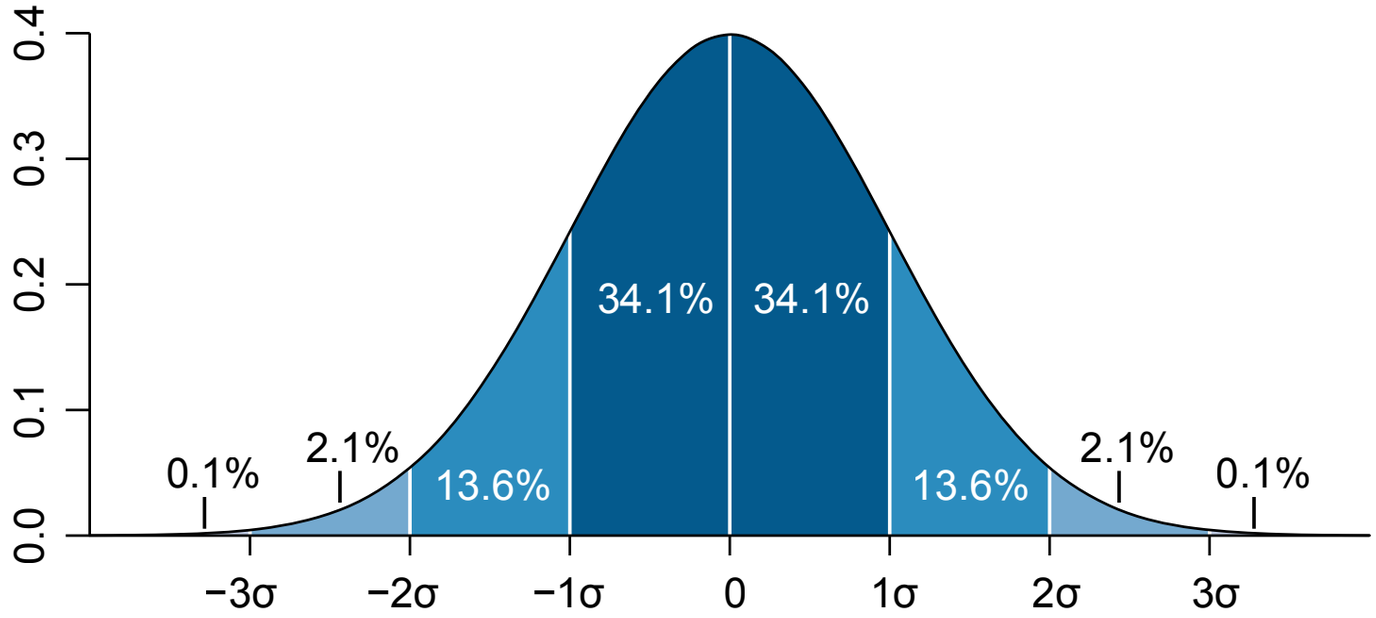
Comparing Two Means

Or: The t-test

Assumptions:

- Our data follow a distribution that can be approximated by the mean
- Equal standard deviation between samples
- They are representative samples
- Independent observations
- Accurate data

The Standard Error of the Mean (SEM)



SEM = SD/square_root(n)

The t-test calculates the standard error of the difference between two means

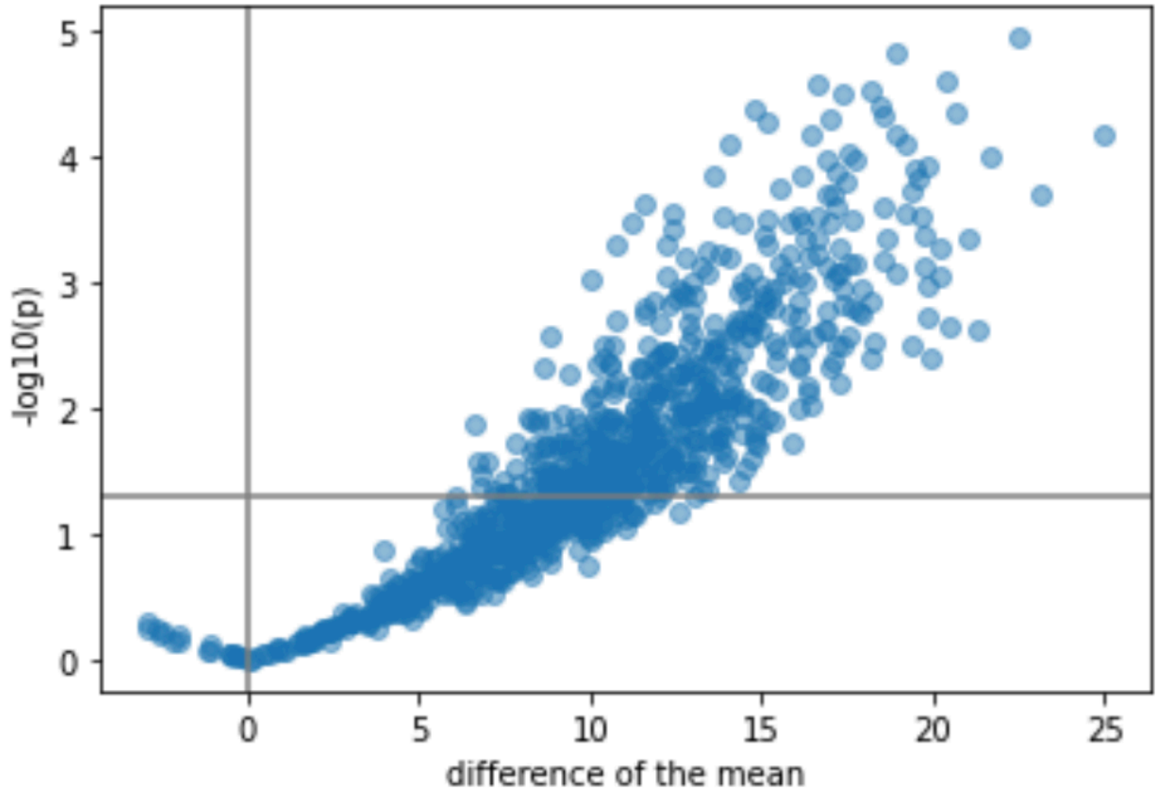
From this, the t-ratio is generated, the difference of the means divided by the standard error of that difference

The p-value is computed from this t-ratio and total sample size.

What does the p-value from the t-test tell us?

The probability that we are wrong, if we consider the two distributions to be different.

The effect size, i.e. the difference of the mean is another important parameter.



When is a t-test inappropriate?

When any of the assumptions is violated, especially the assumption about that the mean needs to be a good approximation of the distribution.

Why?

When using t-tests inappropriately, outliers become very powerful and misleading!!!!

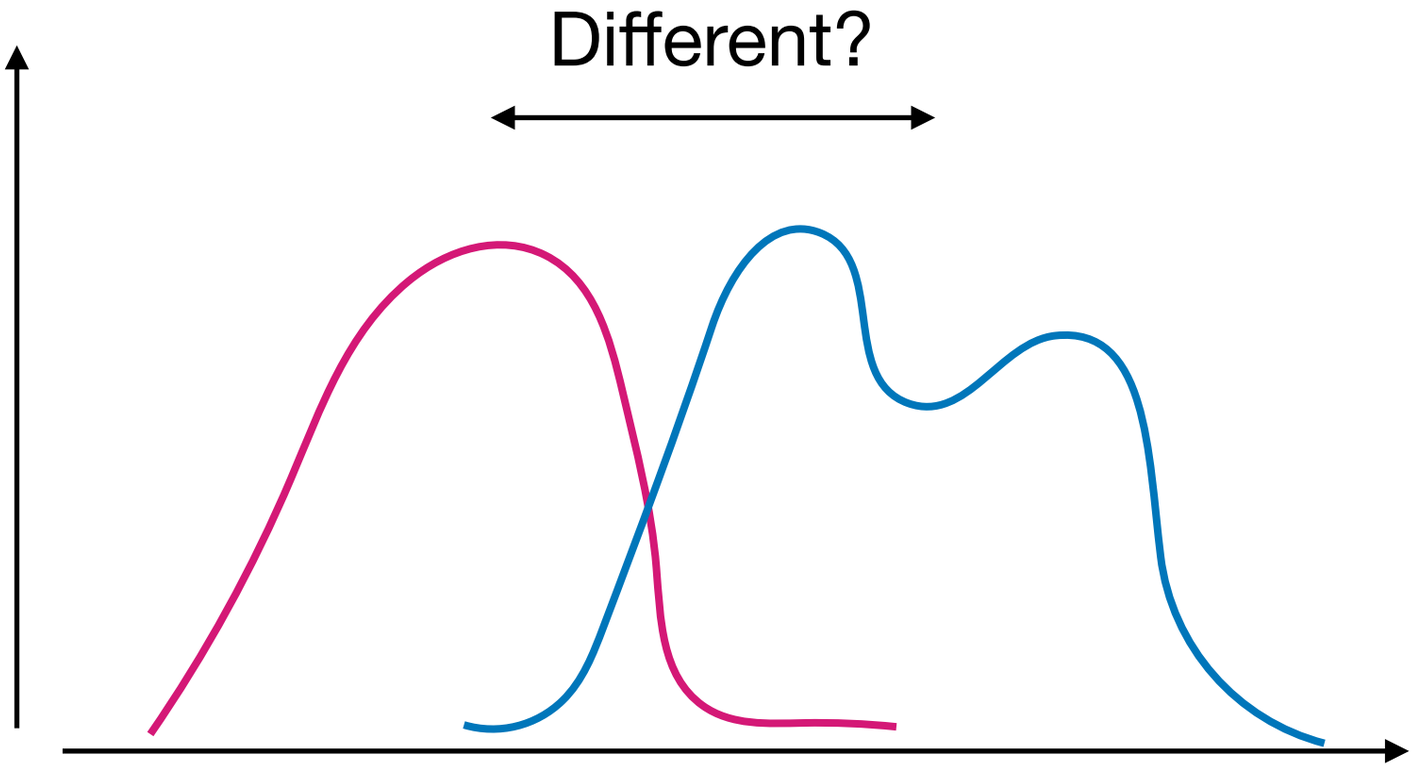
What are the alternatives?

- If data are supposed to meet the criteria theoretically, find the source of your issues

- Assume a different distribution

- Change to non-parametric testing

Non-parametric testing



Choosing between tests

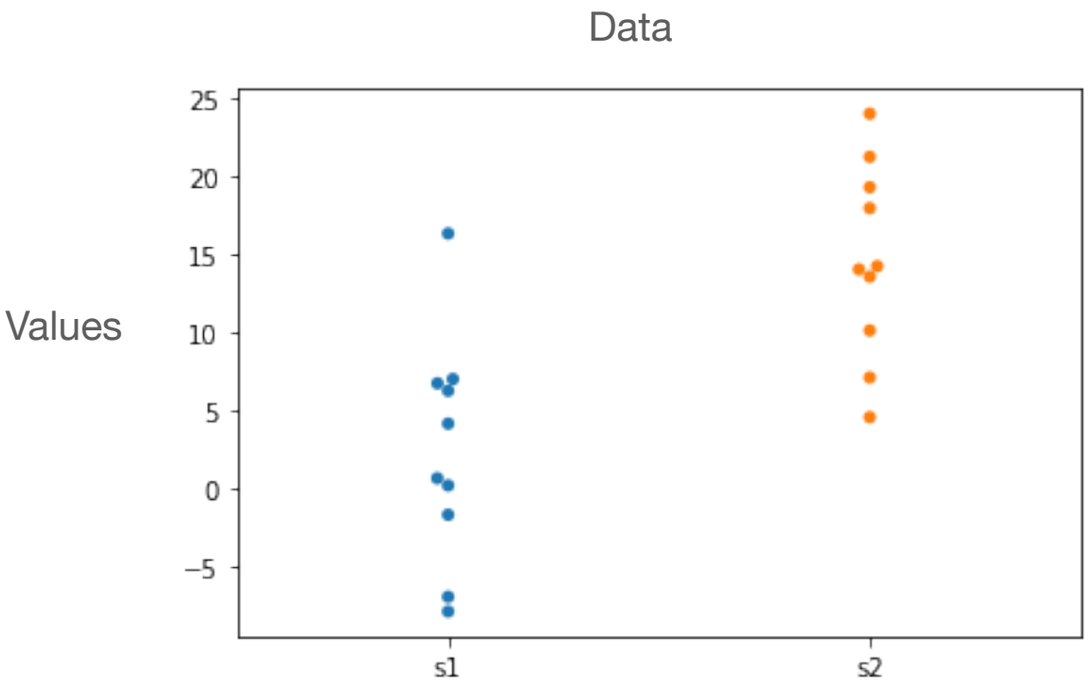
- Whenever you know your distributions and none of the assumptions are violated, go with parametric tests
- Outliers are the most important issue in this regard!
- With lower numbers you will always have more power with a parametric test
- Bootstrapping is a good alternative to rank based non-parametric tests, but it can get computationally very intense and they are not really custom in molecular biology (yet)

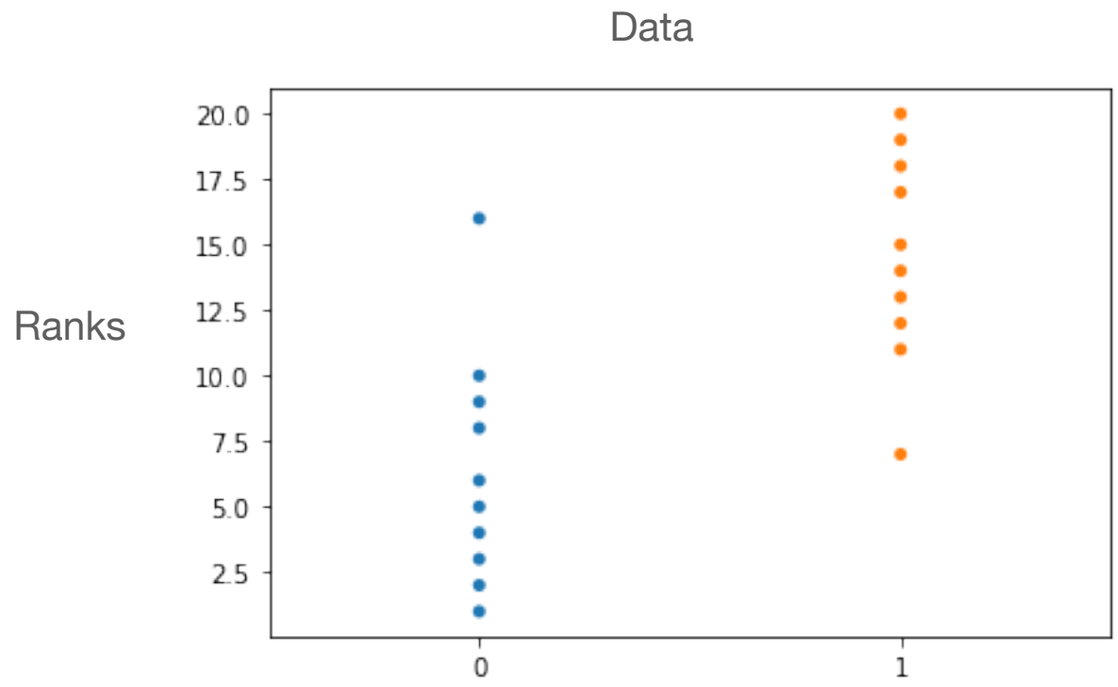
The test's names

- There tends to be a bit of confusion on how to call them….
- Comparing two unpaired groups: Mann-Whitney test
- Comparing two paired groups: Wilcoxon matched-pairs signed-rank test
- Comparing multiple samples (i.e. the non-parametric version of ANOVA): Kruskal-Wallis test

The names are frequently interchanged, e.g. Mann-Whitney is frequently called "unpaired Wilcoxon"!

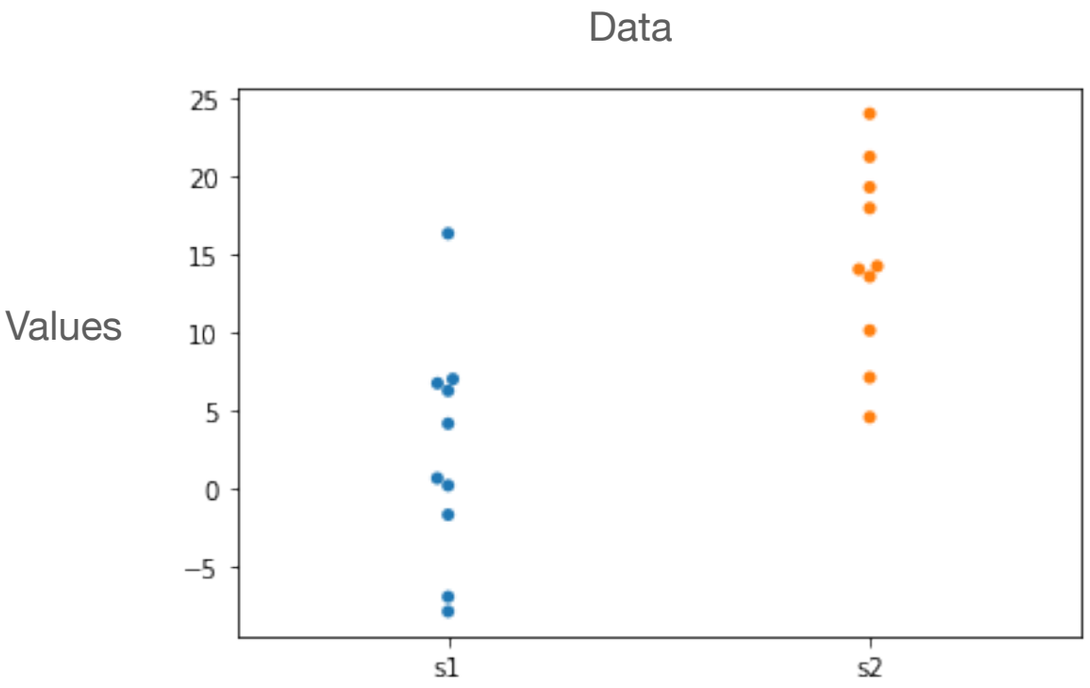
Assumptions

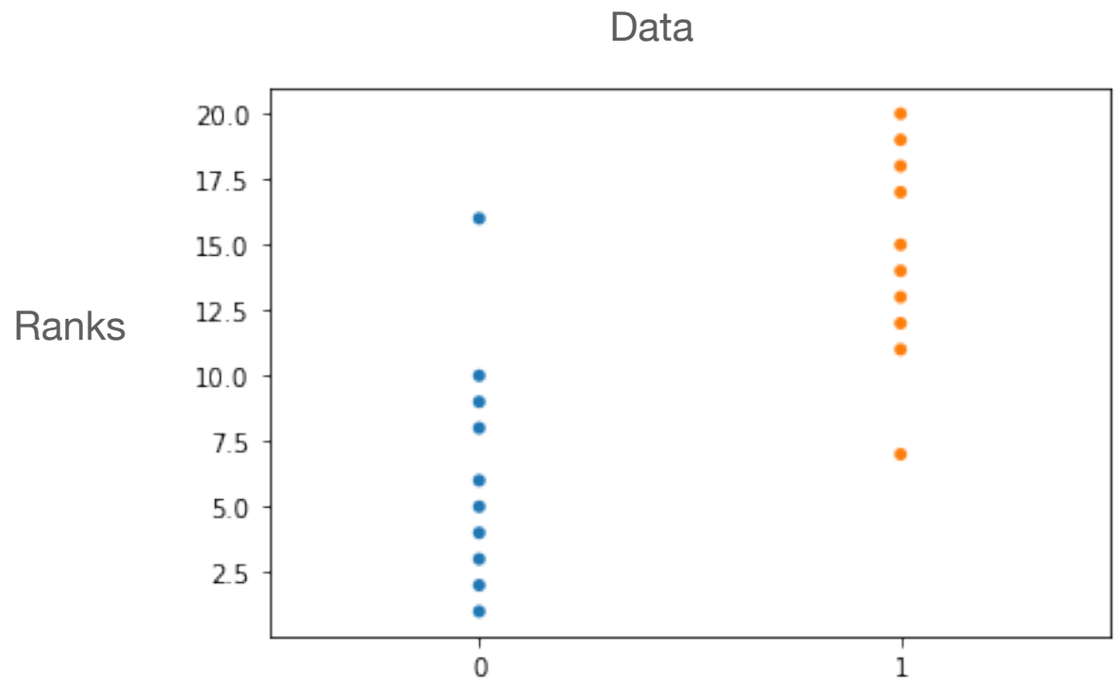




- Random sampling
- Each value is obtained independently

How a rank based test works

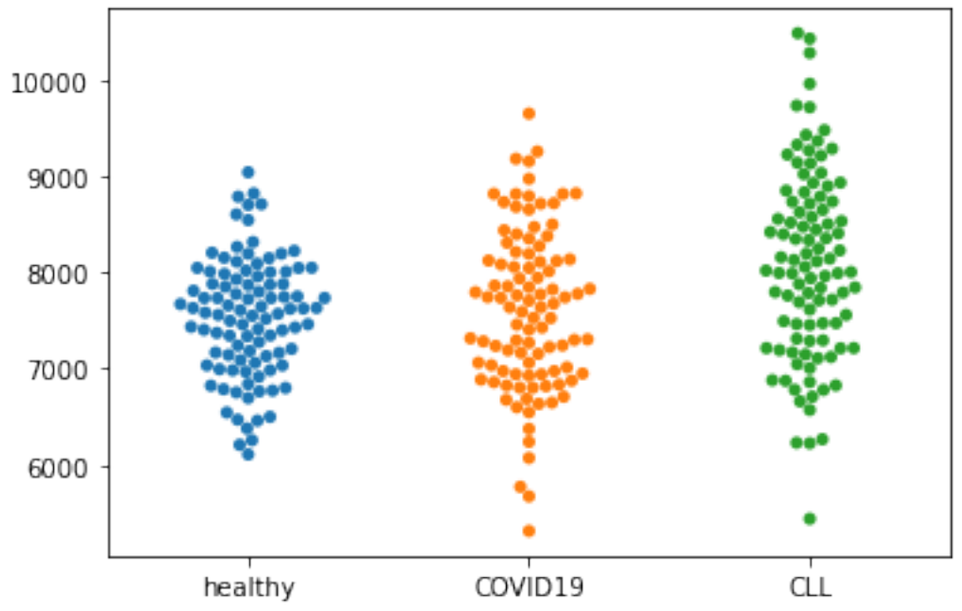




The absolute information is lost and only the ranks are compared.

The p-value describes the probability that the test considers the ranks non-random although they are.

What do you do, if you want to do multiple comparisons?



Do the assumptions for “comparison of means” (t-test) apply?

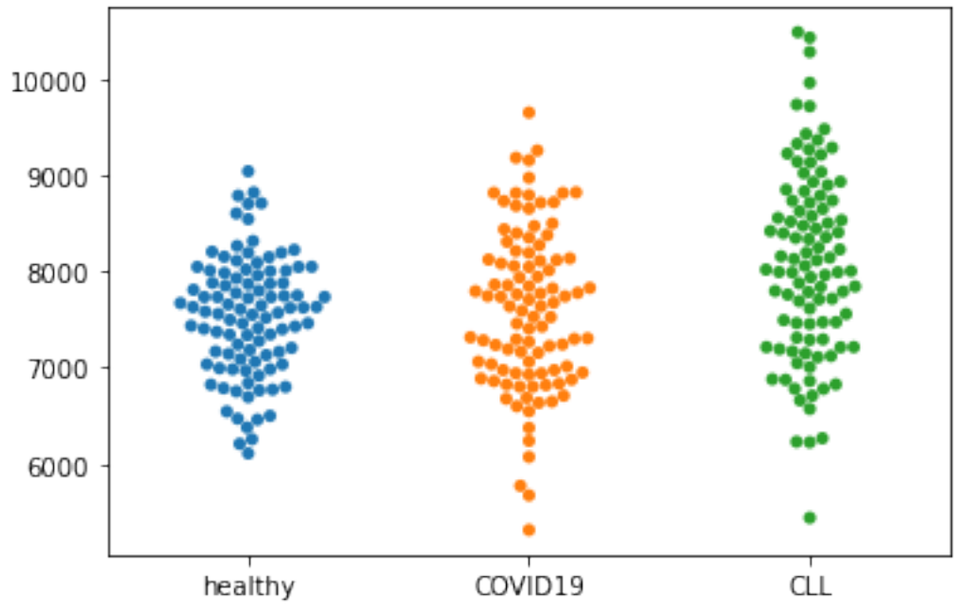
-> Analysis of Variance, one-way ANOVA (= multi-sample-t-test)

-> repeated-samples ANOVA (= multi-sample-paired-t-test)

0-Hypothesis: The mean is identical in all three samples

-> one p-value as output!

But we want to know which one is different!



To extract the p-values for multiple comparisons with corrections, we can take Tukey’s Multiple comparisons test, which takes the differences of the means for each comparing pair and corrects for the number of comparisons.

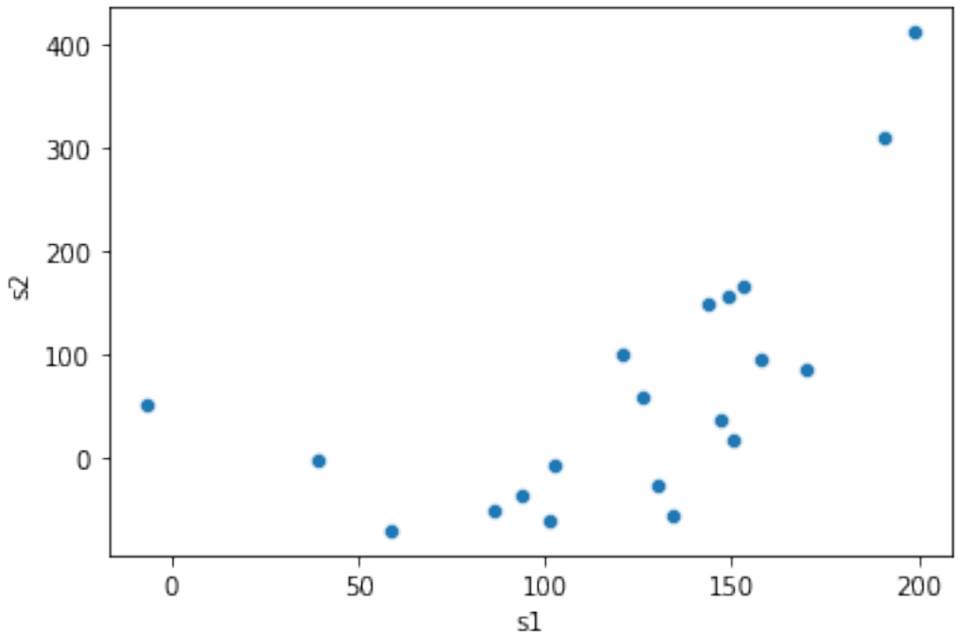
Is Tukey always the best choice?

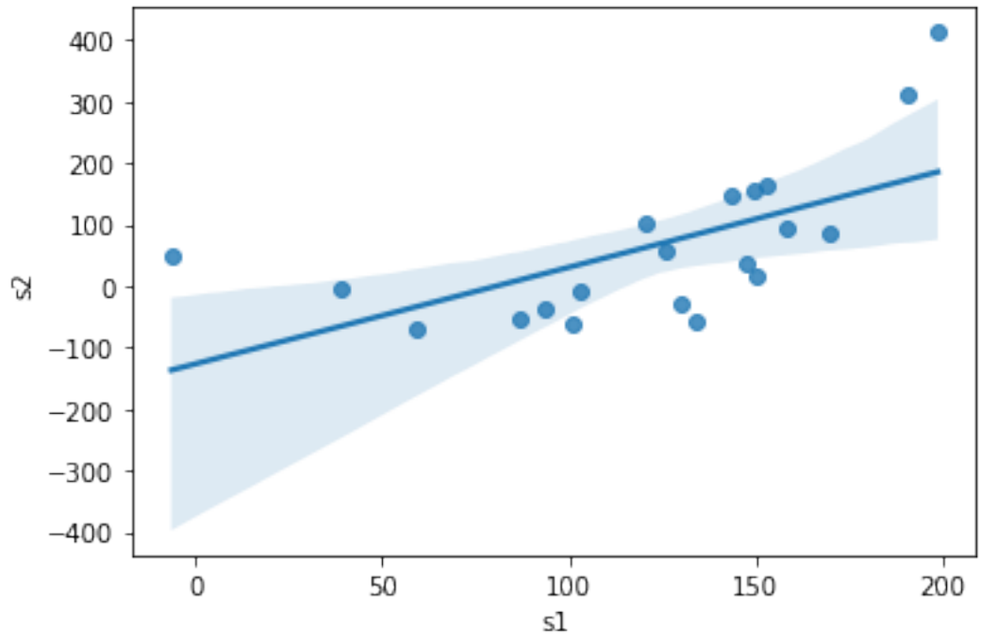
- No, it is the best choice after an ANOVA, because it takes the other comparisons into account, which makes it very powerful
- Alternatives for any other situation are:
  - Bonferroni, which is used a lot in genetics, i.e. divide the p-value by the numbers of comparisons
  - Benjamini-Hochberg: Controlling the false-discovery rate (FDR)

Pearson Correlation

With regression line and

confidence interval

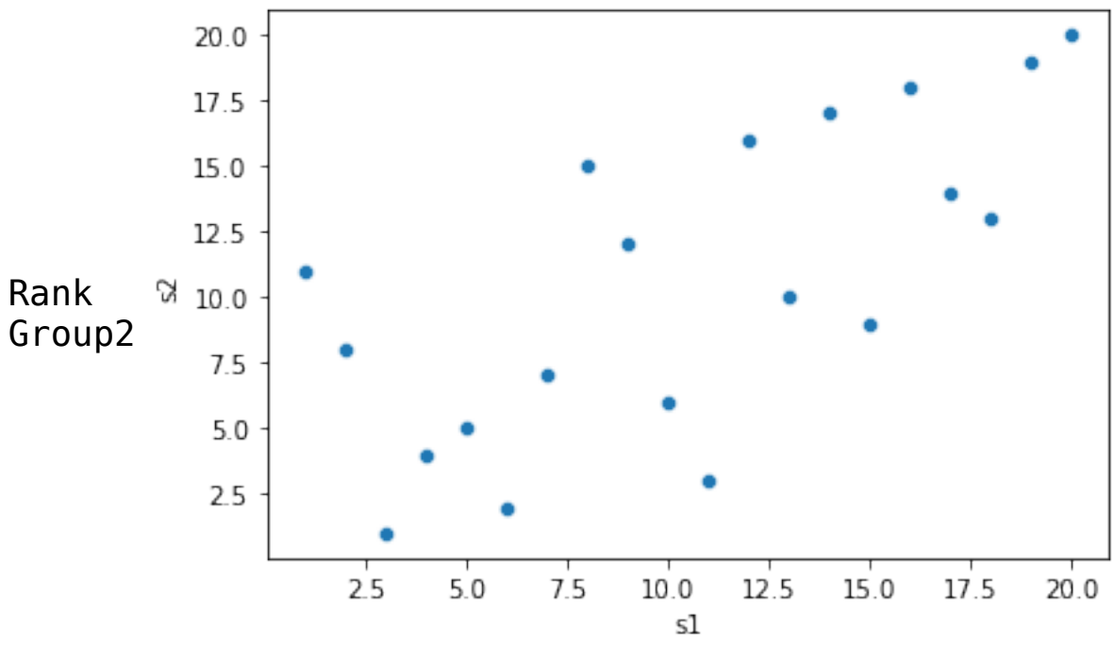




Parametric correlation statistics

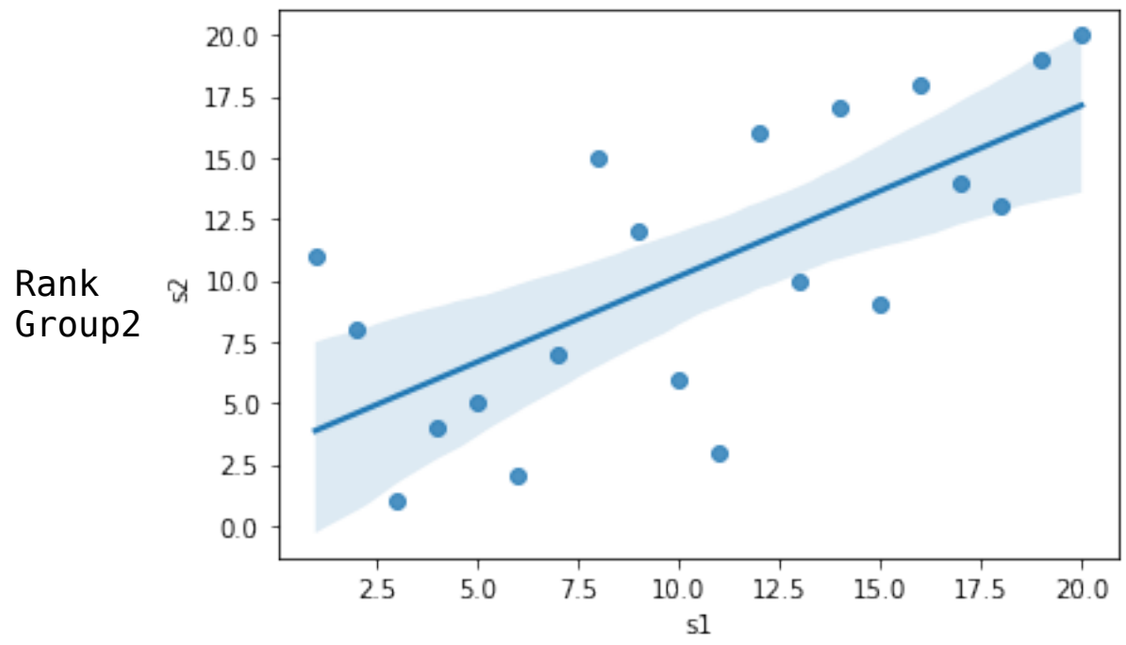
```
R = 0.62
p = 0.003
```

Spearman Correlation



With regression line and

confidence interval



Non-parametric correlation statistics

```
R = 0.70
p = 0.0006
```

Principle Component Analysis (PCA)

- PCA builds linear projections of data into a new coordinate system

- The coordinate system is chosen and ranked by the variance it explains in the data

- Usually the first principle components, the ones with the highest variance explained are shown

- How much variance they explain is indicative of how well the dimensionality reduction has worked

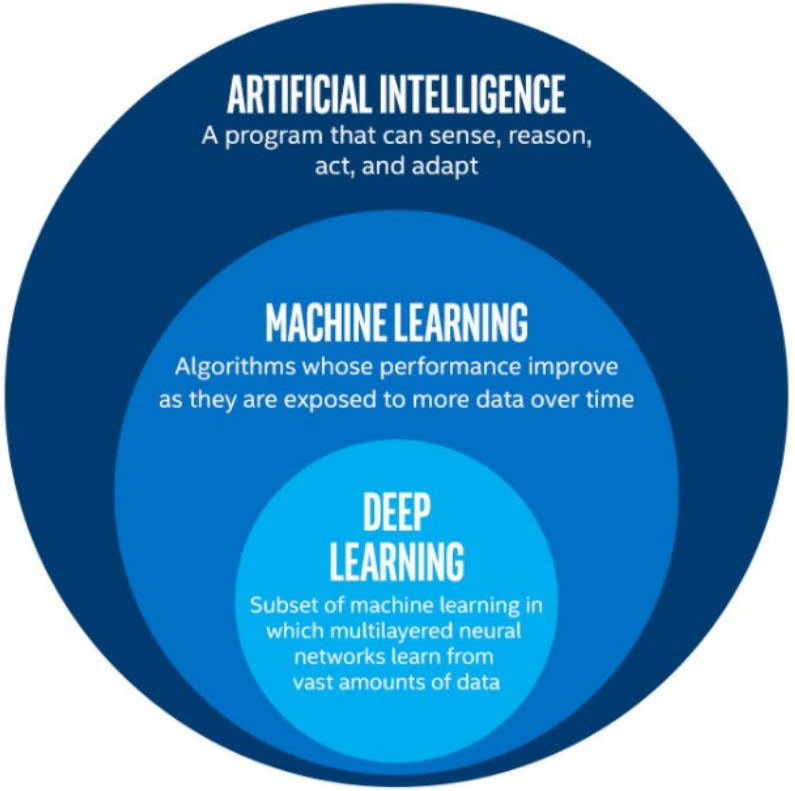
https://setosa.io/ev/principal-component-analysis/

Uniform Manifold Approximation and Projection (UMAP)

UMAP builds something called a **fuzzy simplicial complex**. This is really just a representation of a **weighted graph**, with edge weights representing the l**ikelihood that two points are connected**.

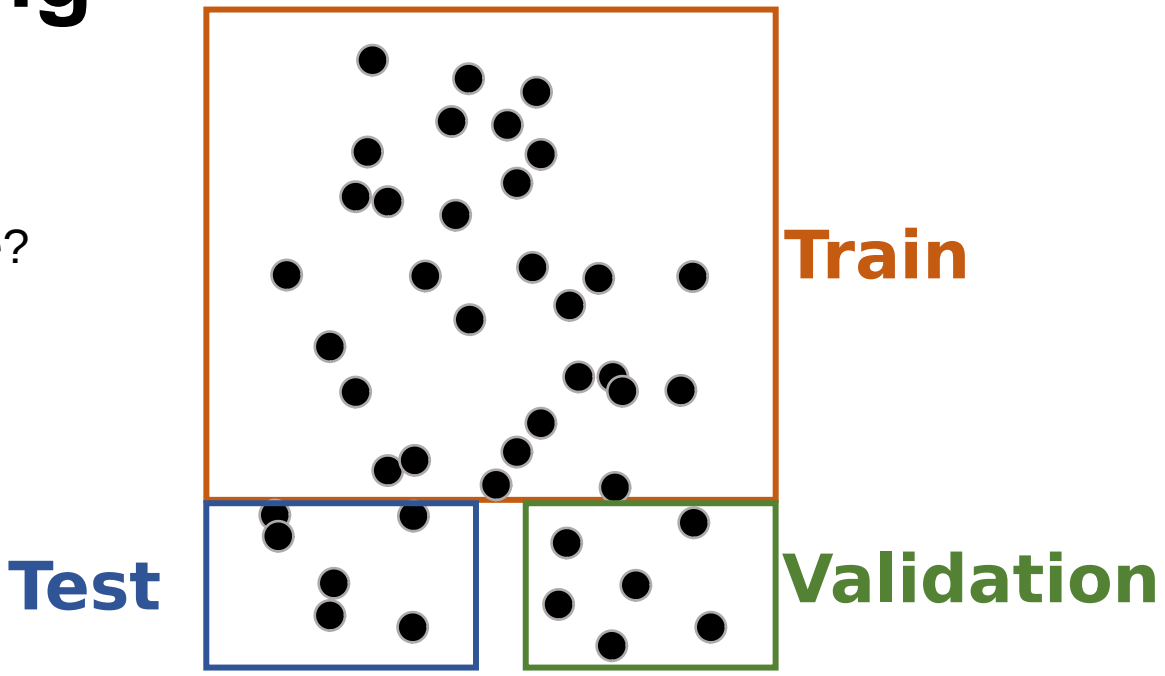
https://pair-code.github.io/understanding-umap/

Machine Learning



Machine Learning

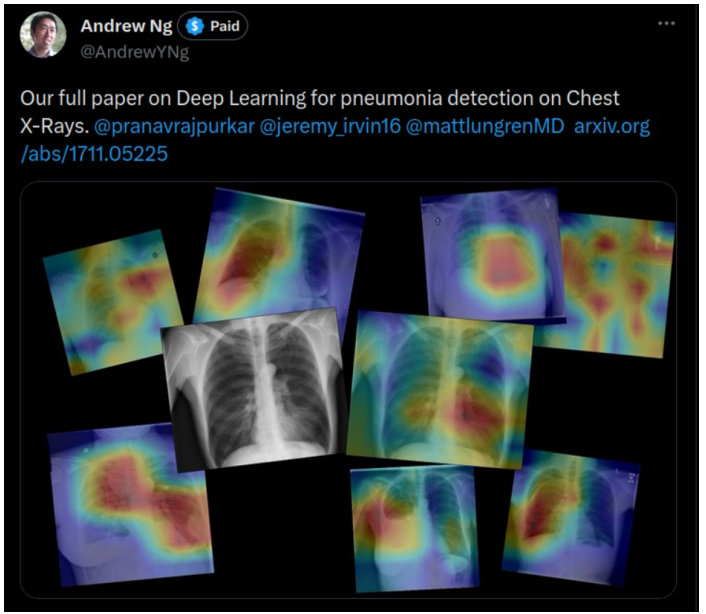
Task
Is it a healthy sample?
Where are the cells in the image?
Is this gene expressed?
….



All the sets are independent of each other and **do not overlap**!

Data Leakage

Training Example Leakage



3.1. Training

We use the ChestX-ray14 dataset released by Wang et al. (2017) which contains 112,120 frontal-view X-ray images of 30,805 unique patients. Wang et al. (2017) annotate each image with up to 14 different thoracic pathology labels using automatic extraction methods on radiology reports. We label images that have pneumonia as one of the annotated pathologies as positive examples and label all other images as negative examples for the pneumonia detection task. We randomly split the entire dataset into 80% training, and 20% validation.

3.1. Training

We use the ChestX-ray14 dataset released by Wang et al. (2017) which contains 112,120 frontal-view X-ray images of 30,805 unique patients. Wang et al. (2017) annotate each image with up to 14 different thoracic pathology labels using automatic extraction methods on radiology reports. We label images that have pneumonia as one of the annotated pathologies as positive examples and label all other images as negative examples. For the pneumonia detection task, we randomly split the dataset into training (28744 patients, 98637 images), validation (1672 patients, 6351 images), and test (389 patients, 420 images). There is no patient overlap between the sets.

Unsupervised vs. Supervised

Unsupervised learning

- **Does not** require labeled data.
- The algorithm must discover by itself hidden/underlying data structure.
- The number of classes and their nature **have not been** predetermined.
- Often used to:
  - Identify patterns and trends
  - Cluster similar data into a specific number of groups

Supervised learning

Require labels.

Requires human oversight.

Unsupervised Learning

K-means

It is an iterative algorithm that divides the unlabeled dataset into **k** different clusters in such a way that each sample belongs only to one group that has similar properties.

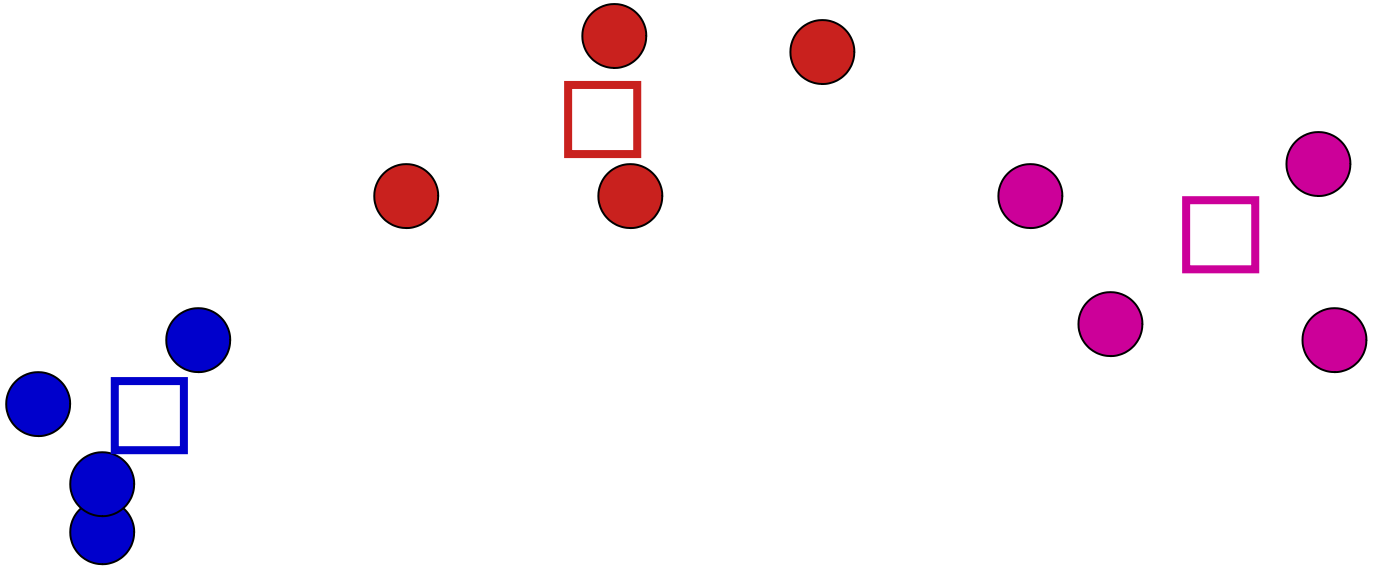
Initialization: set **k** centroids (randomly)

1) Assign each point to the cluster of the nearest centroid measured with a specific distance metric
2) Compute new centroid points (the centroid is the center, i.e., *mean point*, of the cluster)
3) Go back to Step 1), stop when no more new assignment (i.e., membership in each cluster no longer changes)

K-means: An example

Assign point to nearest centroid

k=3



No changes: **Done!**

Supervised Learning

Support Vector Machine (SVM)

Random Forest

Boosting

Naive Bayes

….

https://micro-poll.biotec.tu-dresden.de/micro-poll/public/poll/389ae86a-9d8a-45f0-ad03-aee40ac0a265