Machine Learning

Melissa Sanabria, TU Dresden
melissa.sanabria@tu-dresden.de

Organization

9.5 Introduction to Biostatistics

16.5 Descriptive statistics

23.5 Hypothesis testing

6.6 Introduction to Machine Learning

13.6 Machine Learning

20.6 Neural Networks

27.6 Object Detection

4.7 Dimensionality Reduction

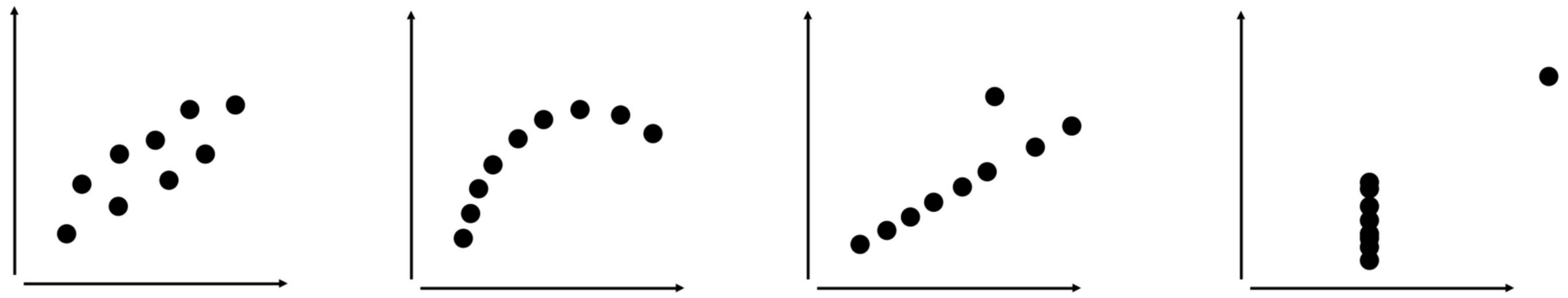
11.7 Summary

Correlations

What for?
To compare paired data in a population.

Correlations are defined by a correlation coefficient (R) and a p-value

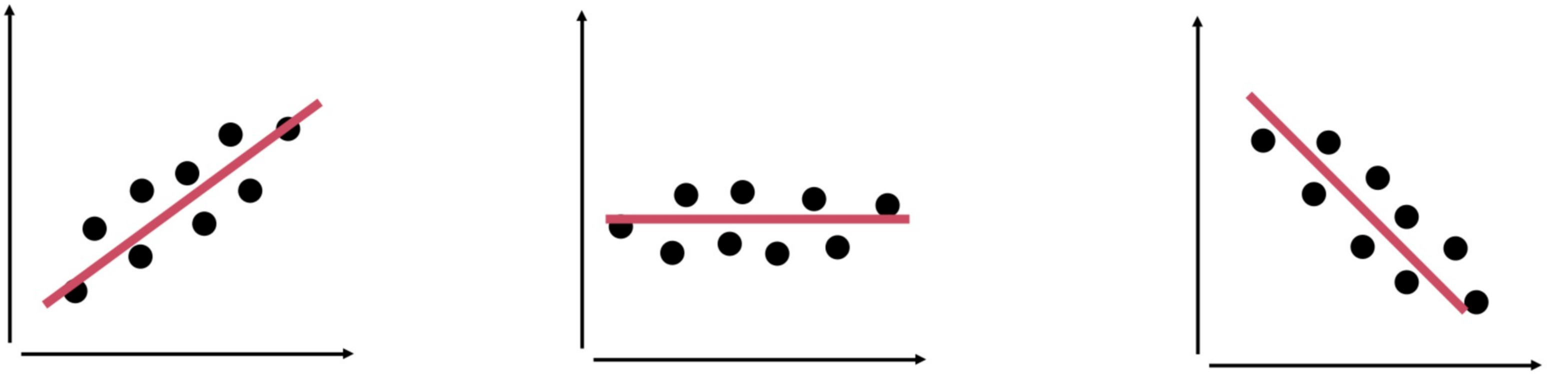
Main rule for any correlation analysis: **Look at your data first!**



These would all roughly have the same correlation coefficient!

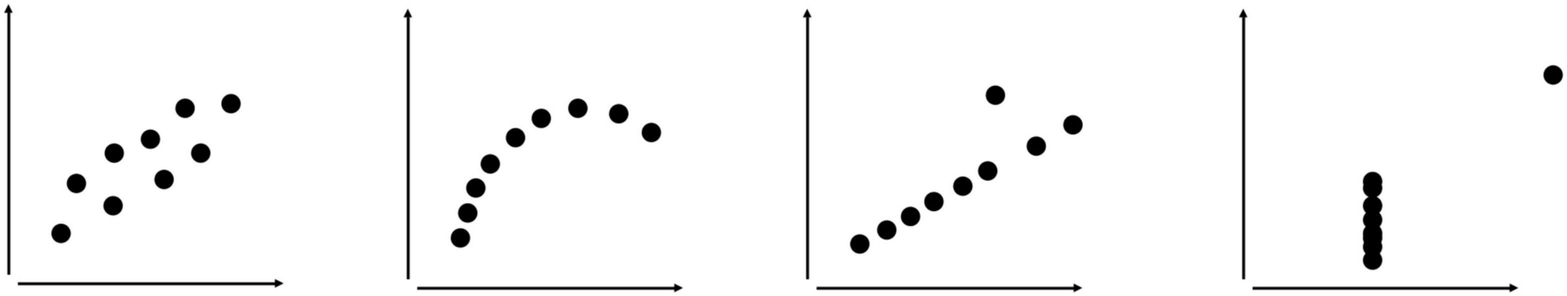
Correlations

Positive	None	Negative



R = 0.7	R = 0.05	R = -0.7
p = 0.01	p = 0.01	p = 0.01

Assumptions



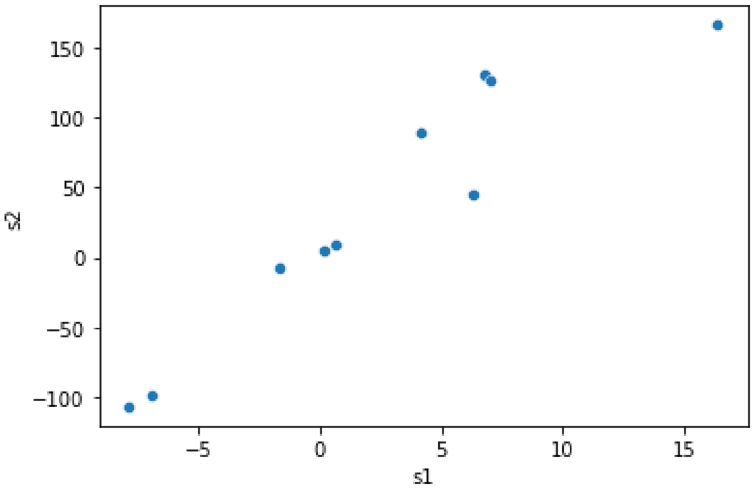
- Random sample
- Paired samples
- Sampled from one populations
- Independent observations
- X-values are not used to compute y-values
- Values are not experimentally controlled

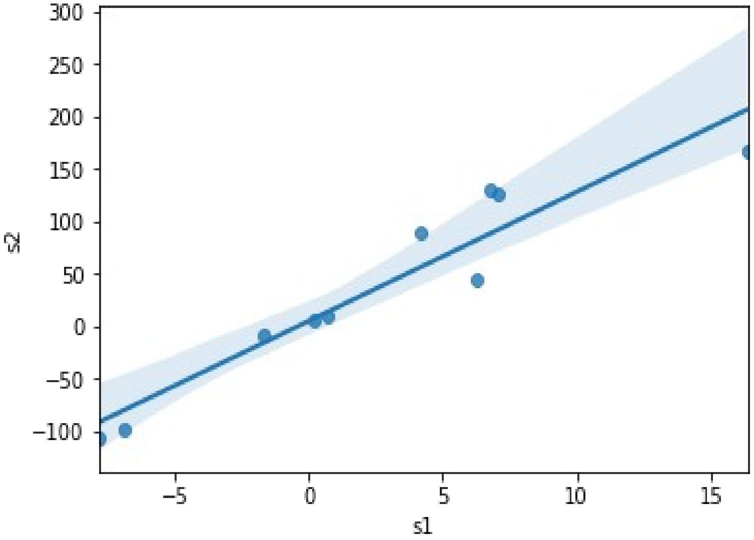
Specifically for parametric:

- Approximate normal distribution
- All covariation is linear
- No outliers !!!!

Pearson Correlation

With regression line and confidence interval



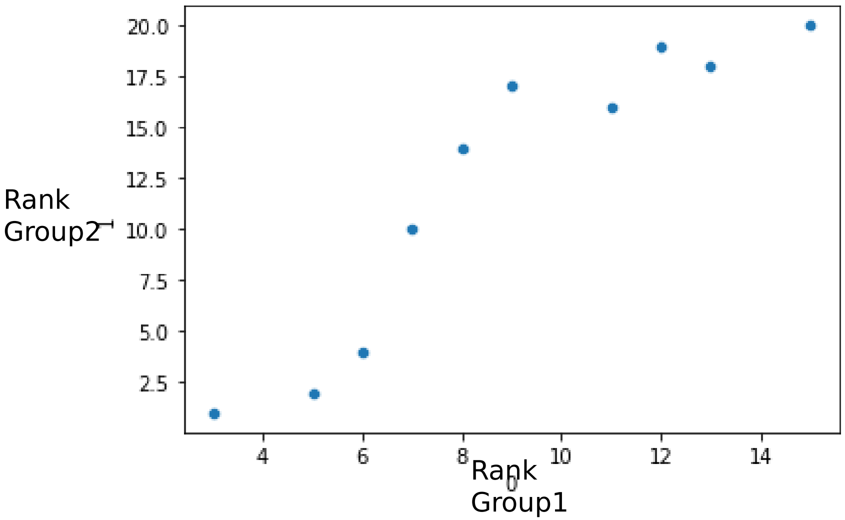


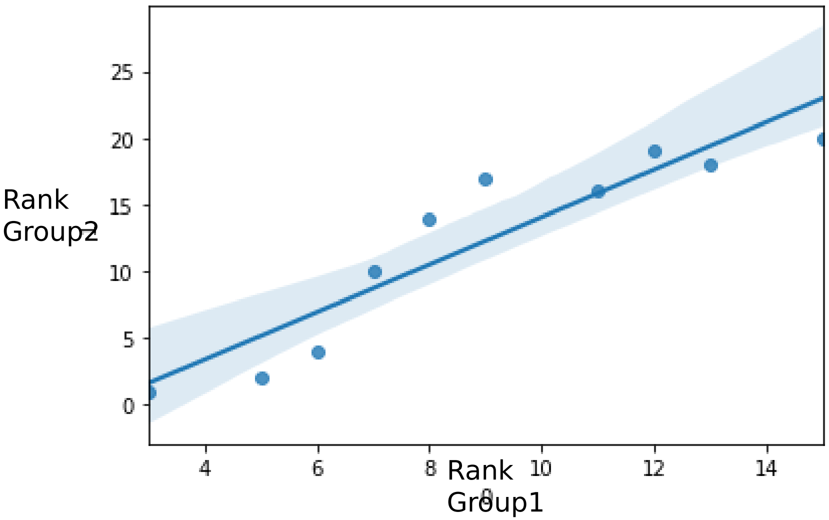
Parametric correlation statistics

R = 0.95
p = 2.6e-05

Spearman Correlation

With regression line and confidence interval





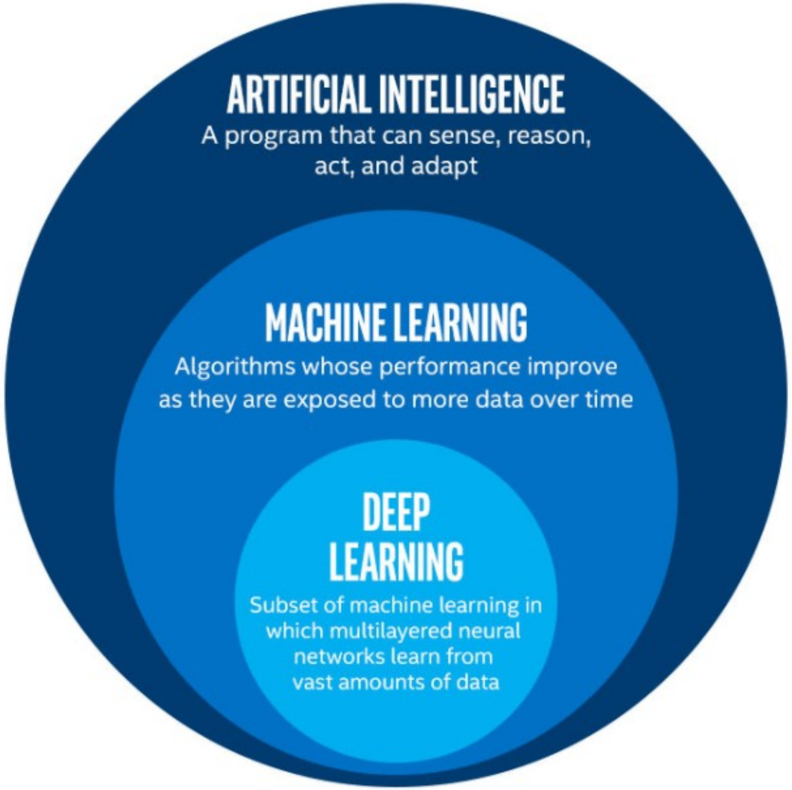
Parametric correlation statistics

R = 0.97
p = 1.5e-06

Correlation statistics

Correlation does not mean causation!
Beware your data structure and outliers!

Machine Learning



Why Artificial intelligence is so difficult to grasp?

Frequently, when a technique reaches mainstream use, it is no longer considered as artificial intelligence; this phenomenon is described as the **AI effect**:
“AI is whatever hasn't been done yet.” (*Larry Tesler*)

e.g. GPS, Alpha Go, Face detection in our phones

AI is continuously evolving and so very difficult to grasp.

Machine Learning

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

Training
Learn how to solve the task

Validation
Verify if you are actually **learning** and **not** just **remembering**.
Modify parameters

Test
Unseen data
Real-life score

Real-life example

Task
Get your driving license

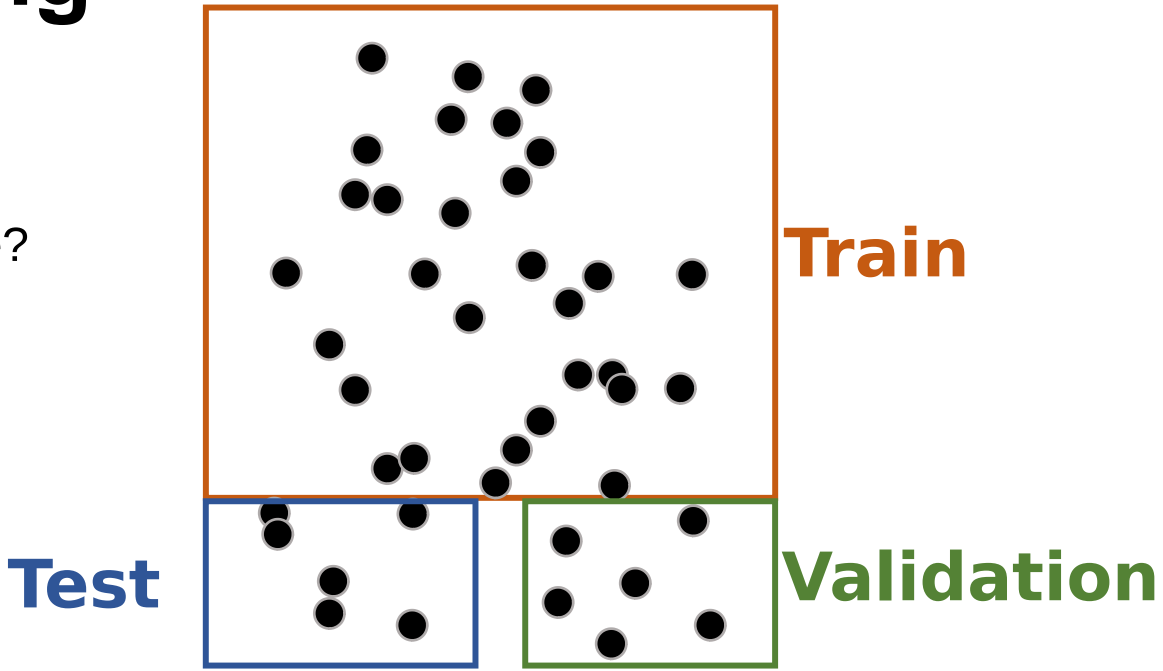
Training
Driving lessons

Validation
Practice in some roads.
Apply certain techniques to see how they actually are in the road.

Test
Go to real trips.
Drive in streets you have never been.

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

Is the scenario where the Machine Learning model is **already aware** of some part of test data during training.

Feature Leakage
A prediction target is inadvertently used in the training process

Training example Leakage
When you aren't careful to distinguish training data from testing data.

Data Leakage

Feature Leakage

JOURNAL OF MEDICAL INTERNET RESEARCH Ye et al

Original Paper

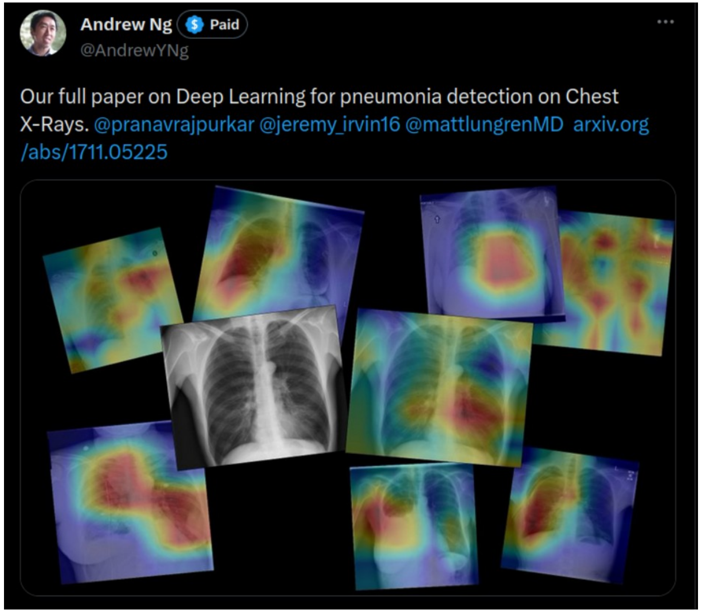
Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning

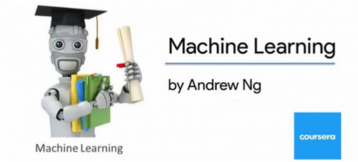
Of the six most important variables, five were: lisinopril, hydrochlorothiazide, enalapril maleate, amlodipine besylate, and losartan potassium. All of these are popular **antihypertensive drugs**.

Just one variable (**the use of a hypertension drug**) is sufficient for physicians to infer the presence of hypertension.

Data Leakage

Training Example Leakage



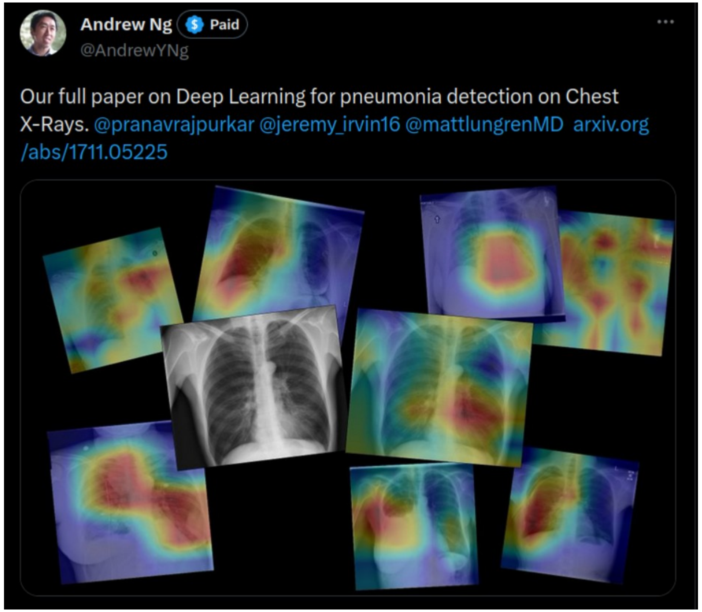


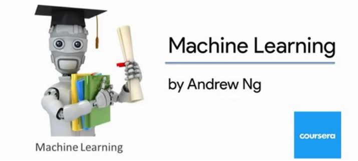
3.1. Training

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.

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.

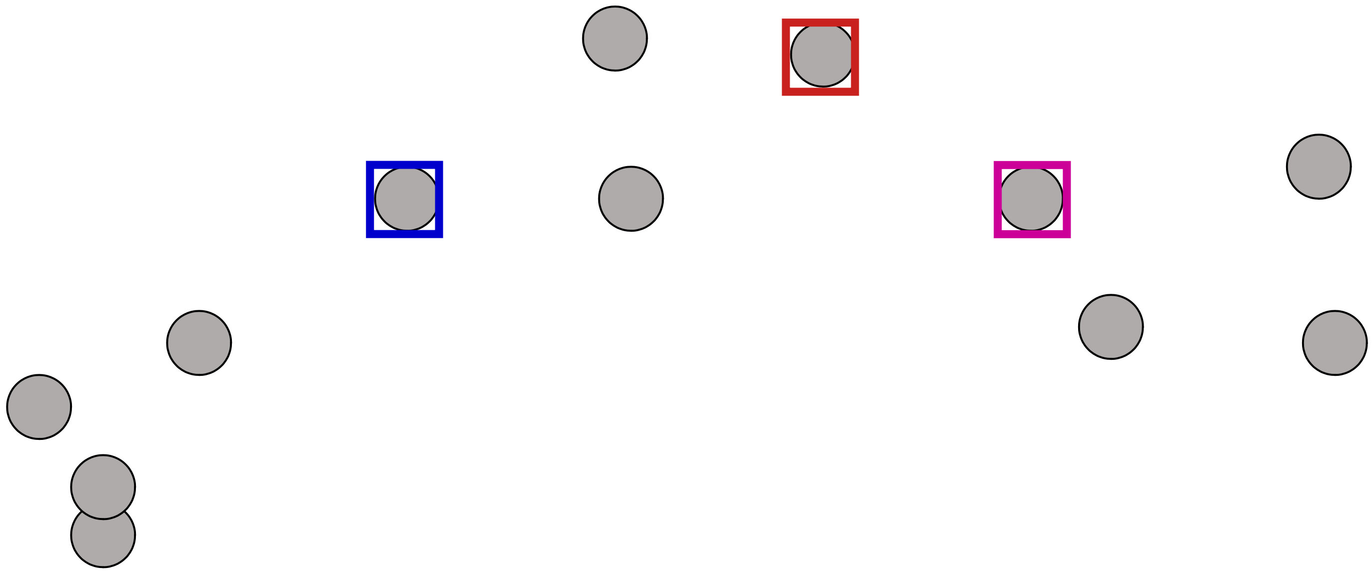
Initialization: set **k** centroids (randomly)

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

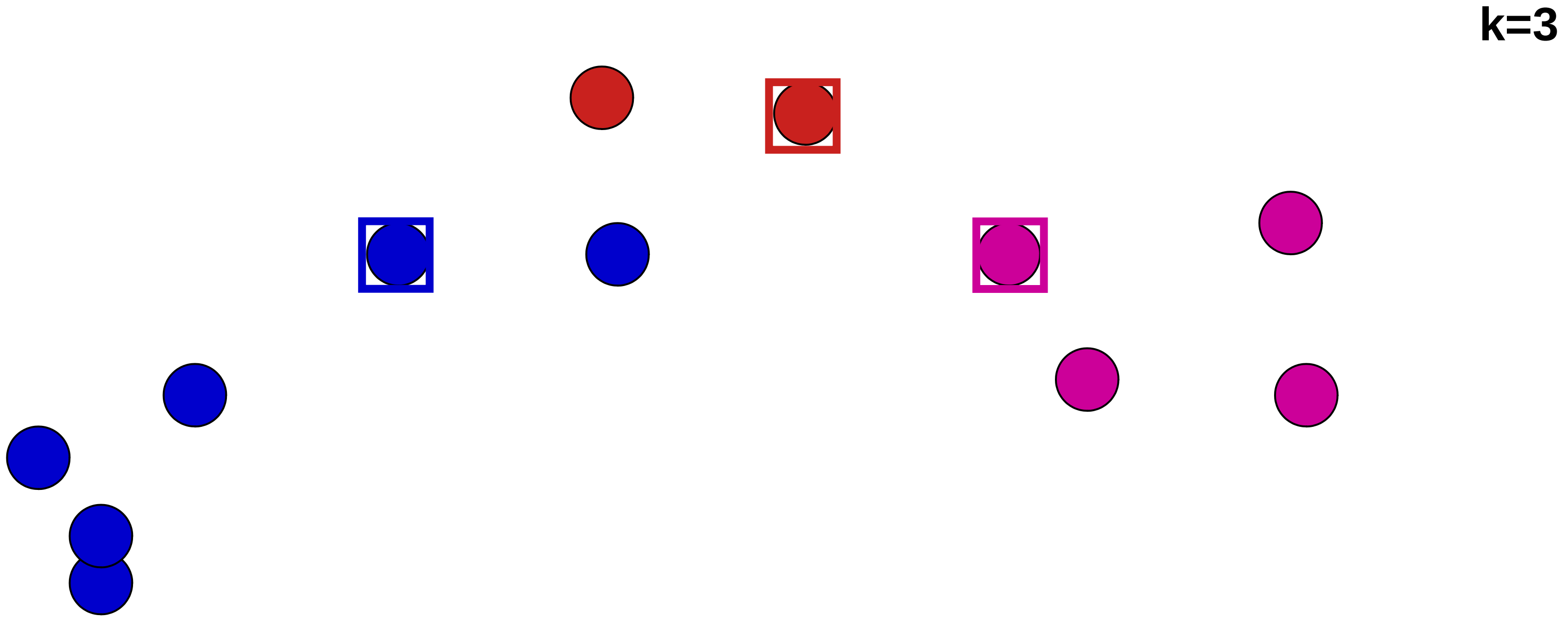
Initialization: set **k** centroids (randomly)

k=3



K-means: An example

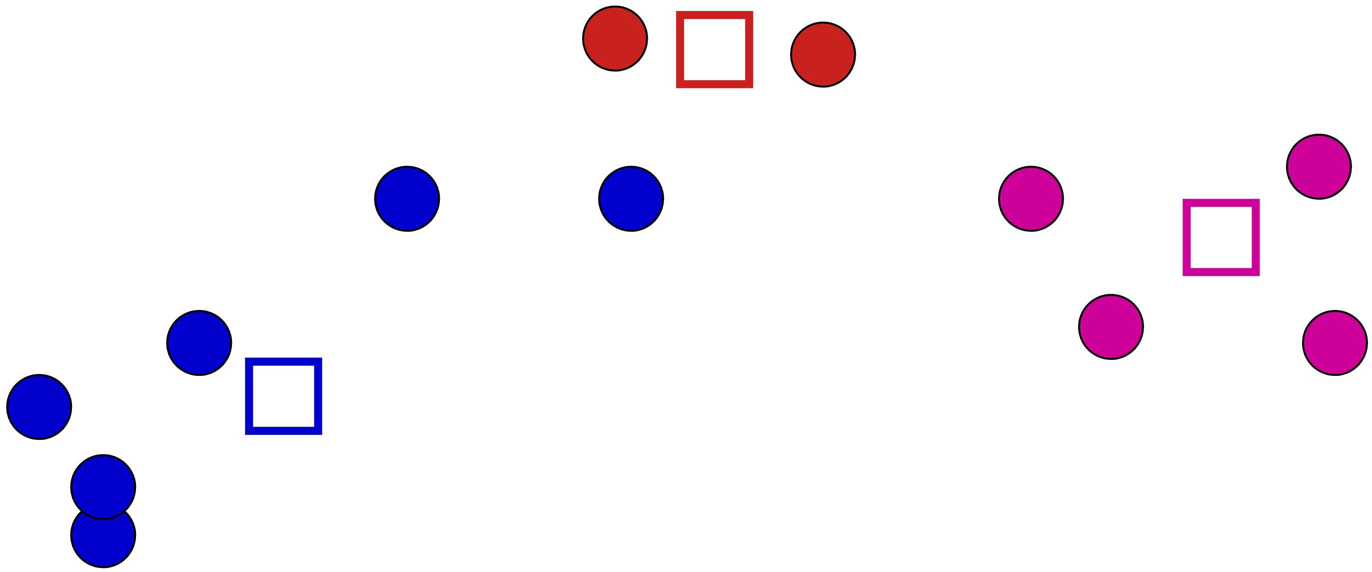
Assign points to nearest centroid



K-means: An example

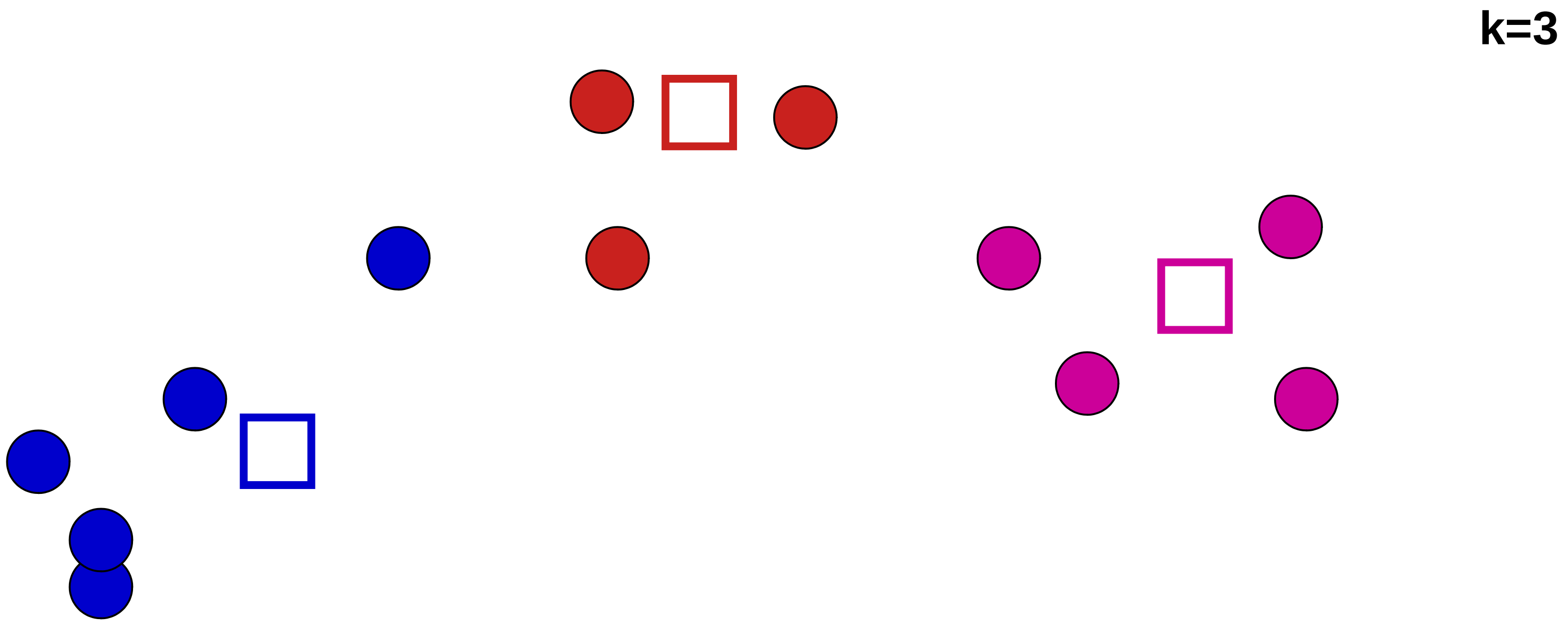
Compute new centroid points

k=3



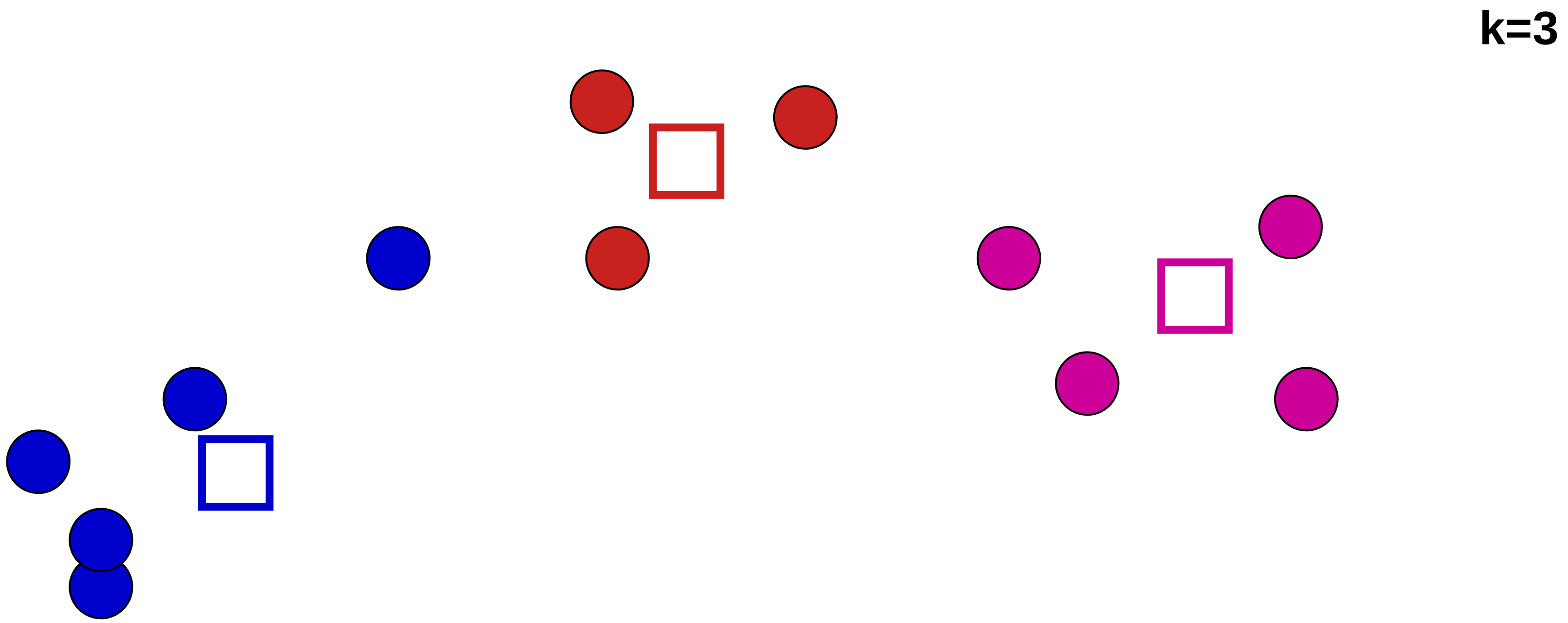
K-means: An example

Assign point to nearest centroid



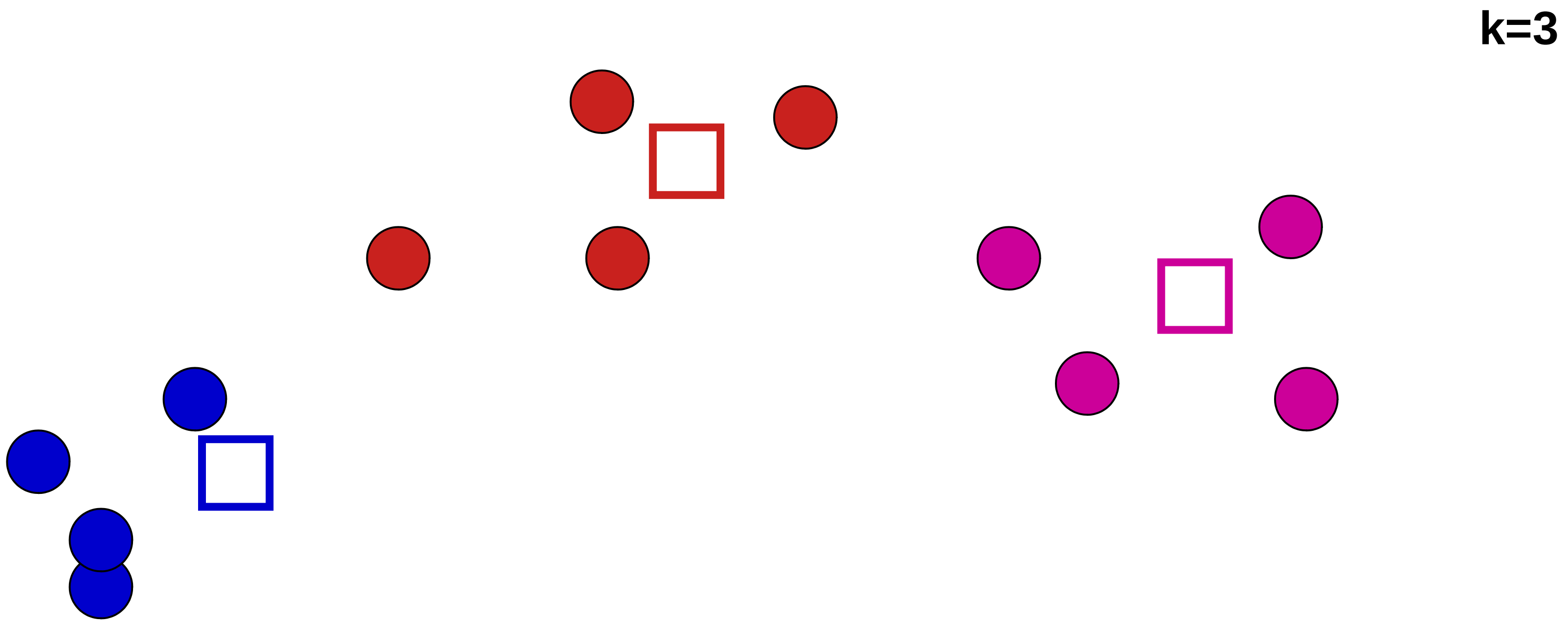
K-means: An example

Compute new centroid points



K-means: An example

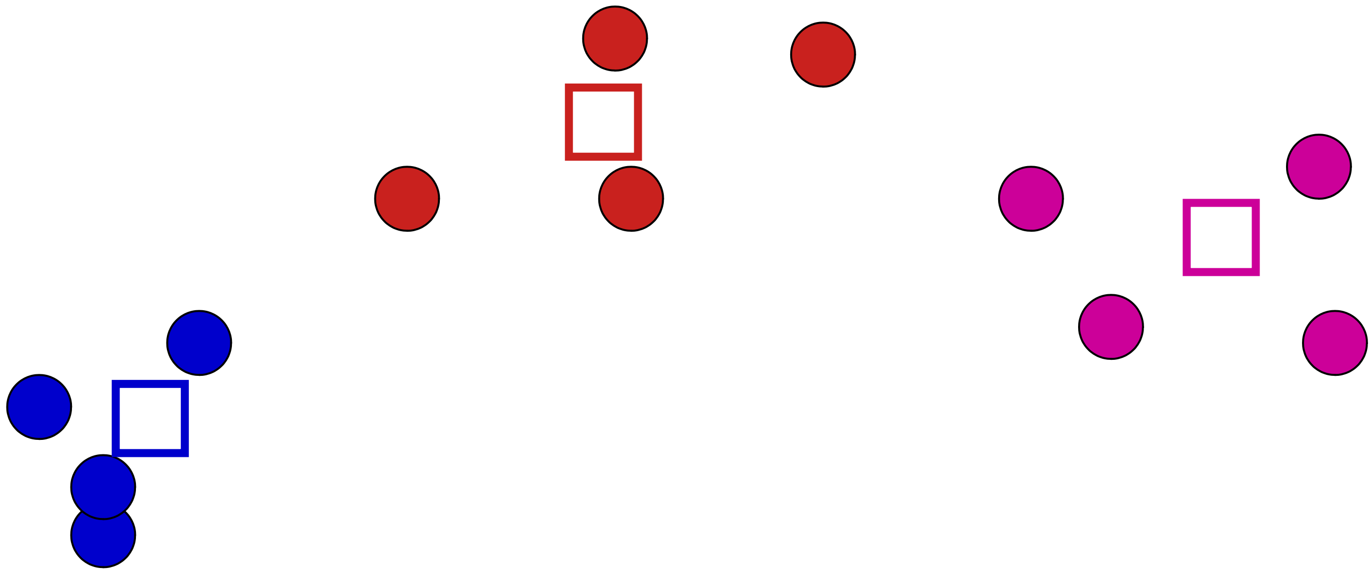
Assign point to nearest centroid



K-means: An example

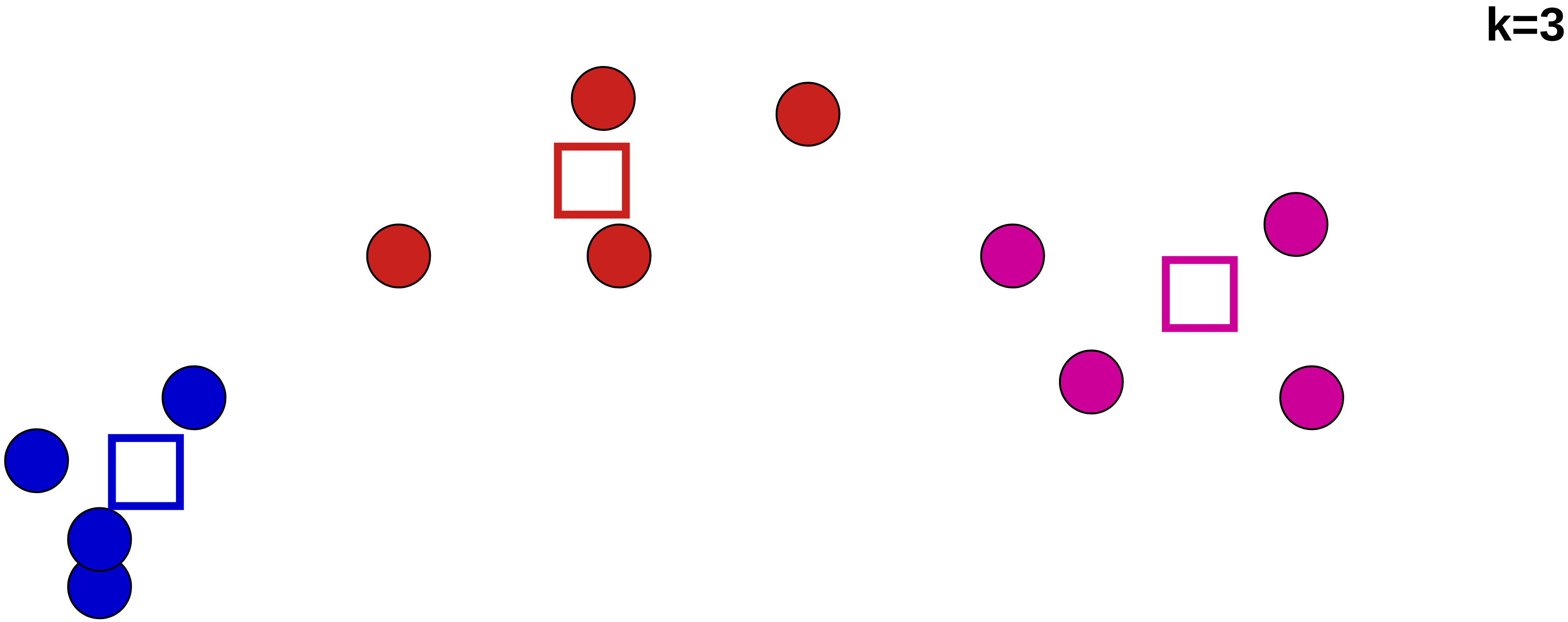
Compute new centroid points

k=3



K-means: An example

Assign point to nearest centroid



No changes: **Done!**

Supervised Learning

Support Vector Machine (SVM)

Random Forest

Boosting

Naive Bayes

….

Let’s practice