Artificial Intelligence and Data Analytics for Engineers (AIDAE)

Lecture 7
June, 19th

Andrés Posada — Today's Lecturer
Anas Abdelrazeq
Marco Kemmerling
Vladimir Samsonov

Artificial Intelligence and Data Analytics for Engineers
Overview Lectures 1 – 4

1 **Introduction to Data Analytics and Artificial Intelligence in Engineering**: Organizational matters (e.g. exam, exercises, dates). Goals, Challenges, Obstacles, and Processes.

2 **Introduction into the primary programming language of the lecture, Python**: Syntax, libraries, IDEs etc. Why is Python the *lingua franca* of the Data Scientist?

3 **Data Preparation**: Cleansing and Transformation. How do real world data sets look like and why is cleaning and transformation an integral part of a Data Scientist's workflow?

4 **Data Integration**: Architectures, Challenges, and Approaches. How can you integrate various data sources into an overarching consolidating schema and why is this important?

Artificial Intelligence and Data Analytics for Engineers
Overview Lectures 5 – 8

✓ 5	**Data Representation**: Feature Extraction and Selection. How to pick relevant features for the task at hand. Manual vs automatic methods. What is the curse of dimensionality?
✓ 6	**Data-Driven Learning**: Supervised (Classification, Regression) methods and algorithms. What is an artificial neural net? What methods are there for evaluation of your model?
7	**Data-Driven Learning**: Unsupervised (Clustering) methods and algorithms. How can machines learn without labels? What methods are there for evaluation of your model?

Today's Lecture

Unsupervised Learning

- What is it?
- Methods
- Applying

Learning Objectives

Learning Objective
w.r.t. Knowledge/Understanding.

After successfully completing this lecture, the students will have achieved the following learning outcomes:

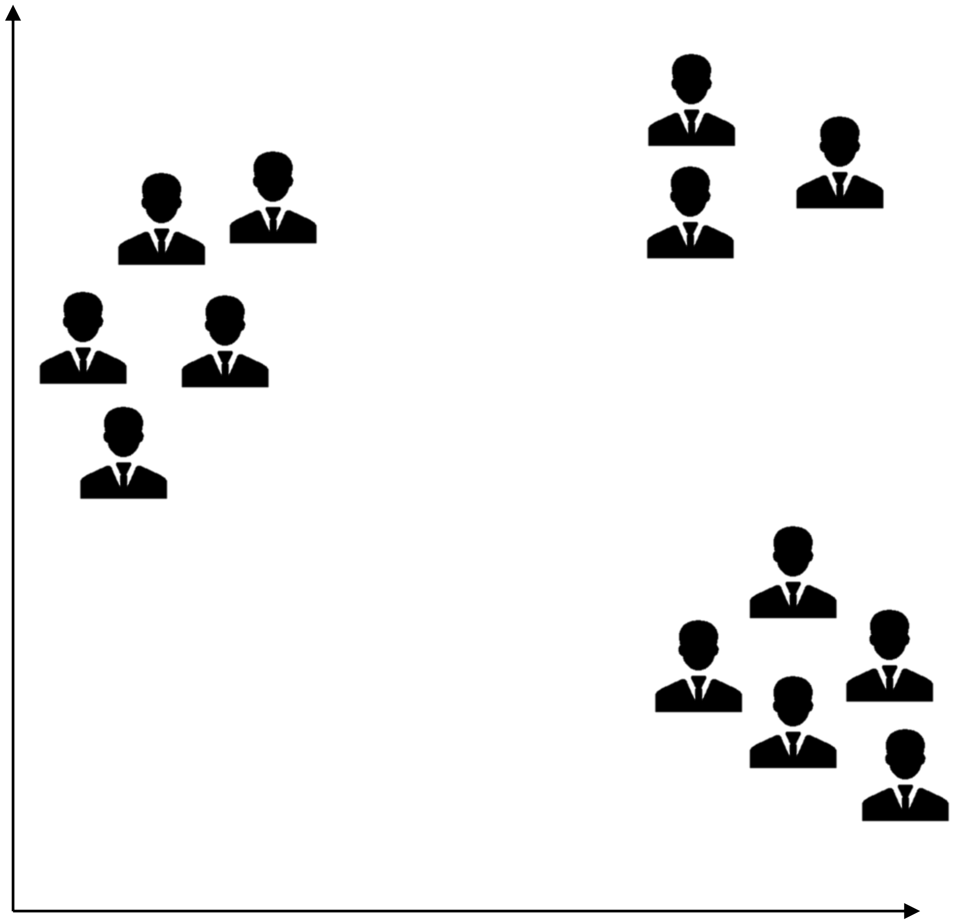
- Have an understanding of what unsupervised learning is.
- Know the different families of unsupervised learning algorithms.
- Learn how to use unsupervised learning algorithms.

Motivation and Introduction

Unsupervised Learning in Different Industries

By using **clustering algorithms** to segment markets, companies can:

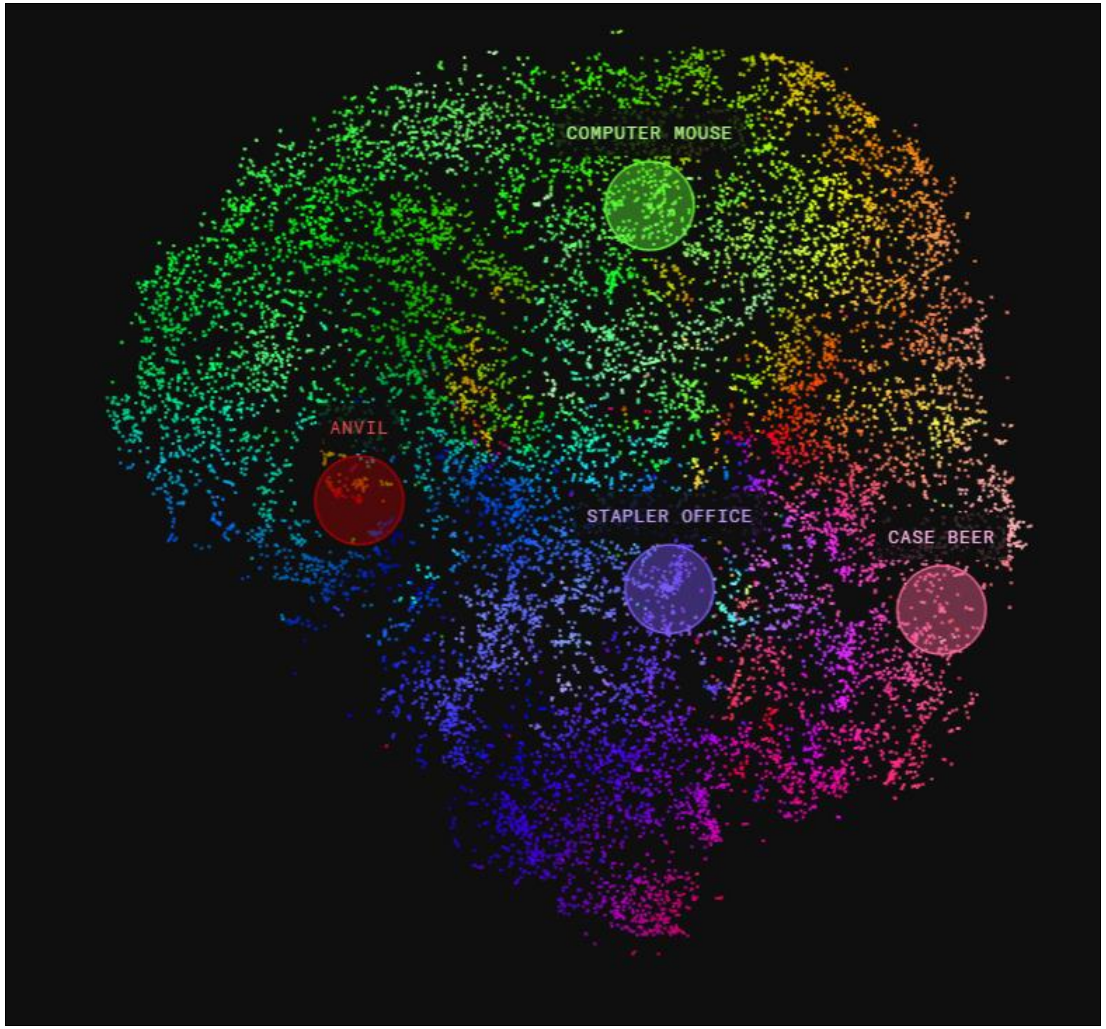
- Identify your most profitable group of customers.
- Focus your marketing on segments most likely to purchase.
- Discover potential niche markets.
- Develop or improve products to meet customer needs of groups of customers.



Unsupervised Learning in Different Industries

- Unsupervised learning algorithms such as manifold learning are often used to get a more understandable representation of a high dimensional dataset.
- Examples of this are applications that create sound maps of
  - Instruments
  - bird songs
  - Images
  - machine noises

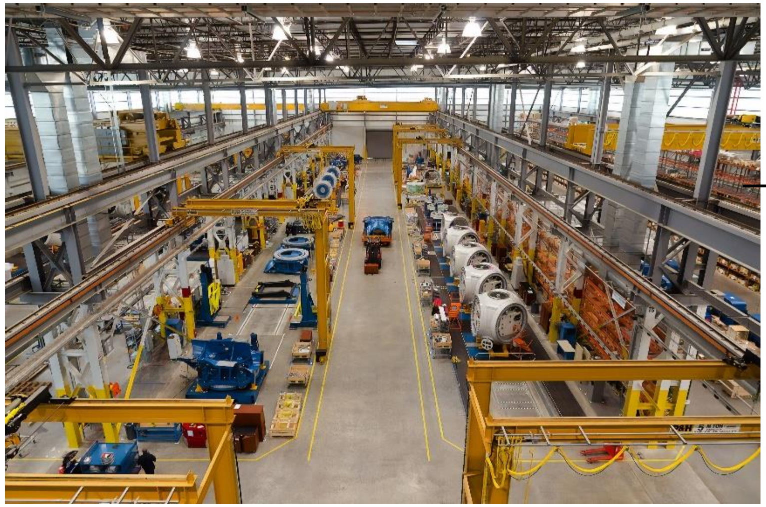
  in order to find out more about the data.



Source: https://experiments.withgoogle.com/ai/drum-machine/view

Unsupervised Learning in Different Industries

In the manufacturing industry, **Association rule discovery** is used over SCADA logs of events and alerts to better understand the relations between the failures in multiple machines.



Info P1

Info P2

Info P3

Association rule discovery

Rule	S	C
Fan ID A67 is damaged or missing & Material Life AB 69 and Oil pump is damaged or missing => Concentrator ID 1 is missing	0.31	1
Immersion Pump P1015 is damaged or missing & Fan ID A67 & Hydrauliuc Oil pump is missing => Concentrator ID 2 is missing	0.17	0.75
Fan ID A89 is damaged & Fan ID A89 is damaged & Lift Workshop is damaged or missing => Concentrator ID 21 is missing	0.31	1

Introduction

Introduction

- Brief view into Artificial intelligence and Machine Learning

Artificial Intelligence:
Any technique which enables a computer to mimic human behaviour.

General AI (GAI)
Transfer knowledge across domains.

Narrow AI
Performs a single task extremely well

Machine Learning
Algorithms whose performance improve as they are exposed to more data.

Supervised Learning
(Regression, Classification, etc.)

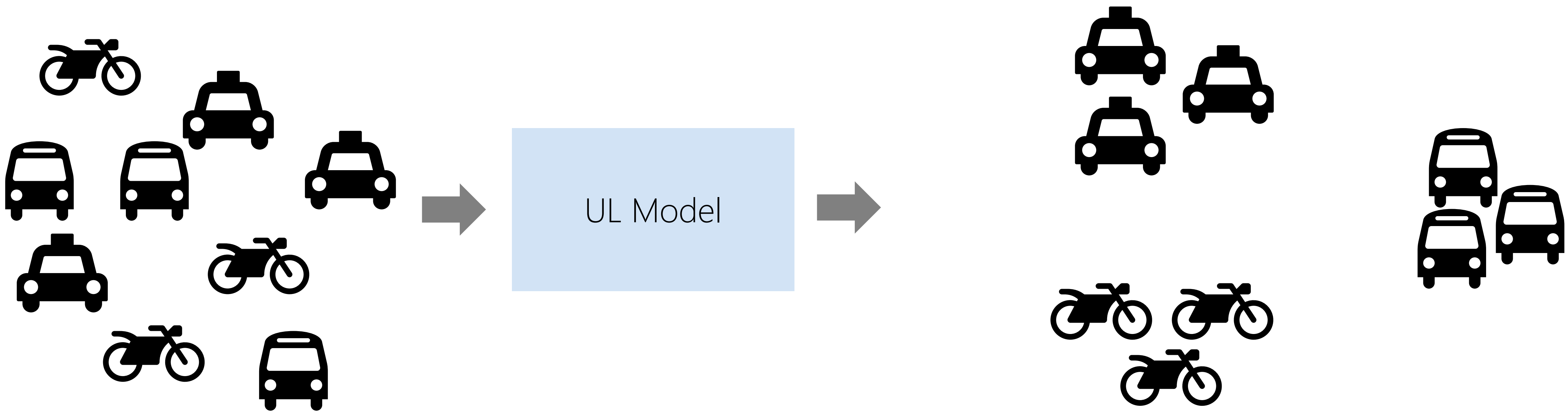
Unsupervised Learning
(Clustering, Dimensionality, Reduction.Z, Rule discovery, etc.)

Reinforced Learning

What is supervised learning?

Working Definition

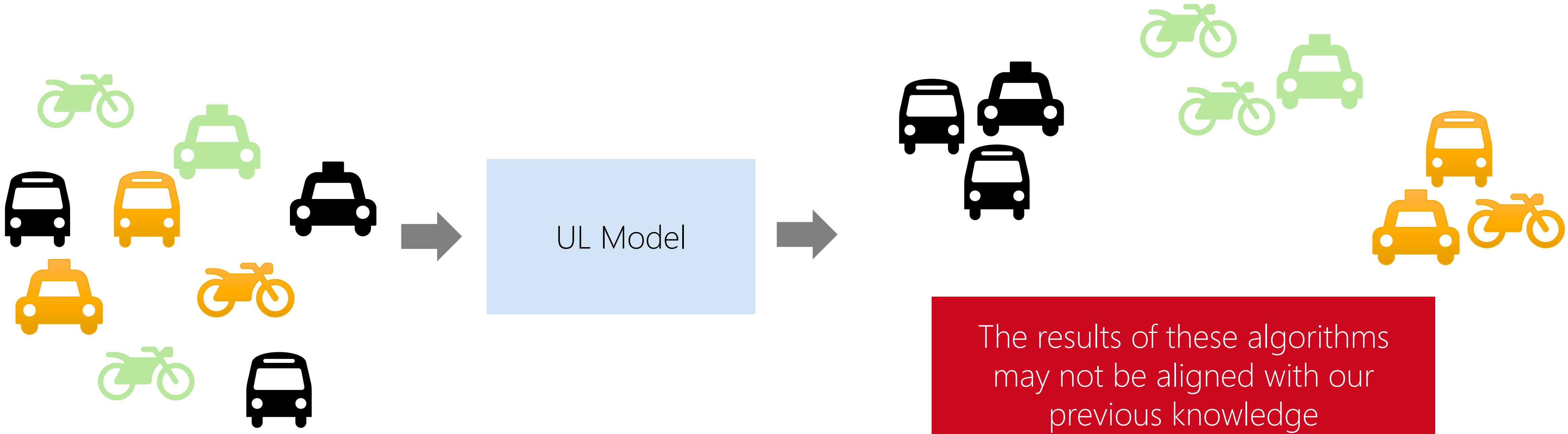
Unsupervised learning deals with problems in which your dataset **doesn't have labels**. Instead, the model is allowed to discover **relations** in the data on its own.



What is supervised learning?

Working Definition

Unsupervised learning deals with problems in which your dataset **doesn't have labels**. Instead, the model is allowed to discover **relations** in the data on its own.



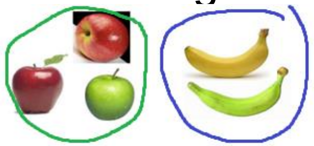
The results of these algorithms may not be aligned with our previous knowledge

What is supervised learning?

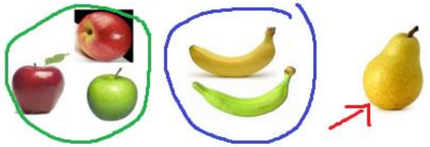
What Kind of Unsupervised learning algorithms exist?

Finding clusters

- Clustering

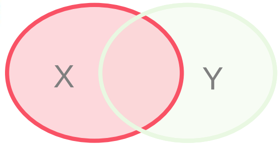


- Anomaly Detection



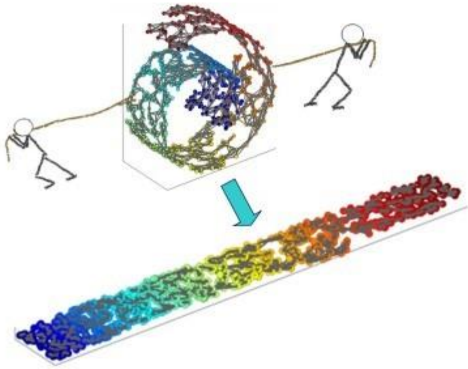
Association rule mining

Association:
„If product X is bought, it‘s likely product Y is bought as well.“



Dimensionality Reduction

Reduce data to fewer Dimensions.

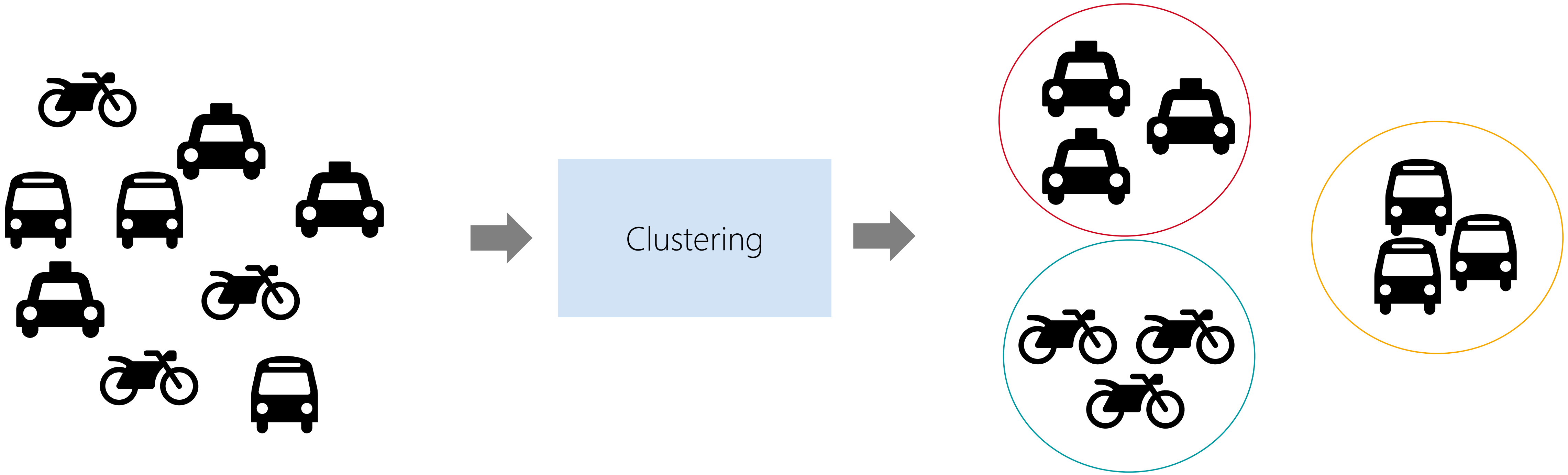


Clustering

What clustering?

Working Definition

Clustering is the task of **grouping** sets of objects (**clusters**), so that more **similar** items are in the same group and less **similar** items are in separate groups (**clusters**).



What clustering?

3 main families of techniques for clustering

Hierarchical clustering

Finding a cluster-hierarchy, e.g. beginning with as many clusters as data-objects.

- Single-Linkage
- Wards method

Partitioning Clustering

Optimizing cluster centers to minimize the distance to the data-objects to a cluster.

- K-means
- Fuzzy C-means
- Affinity-Propagation
- EM-Clustering (GMM)

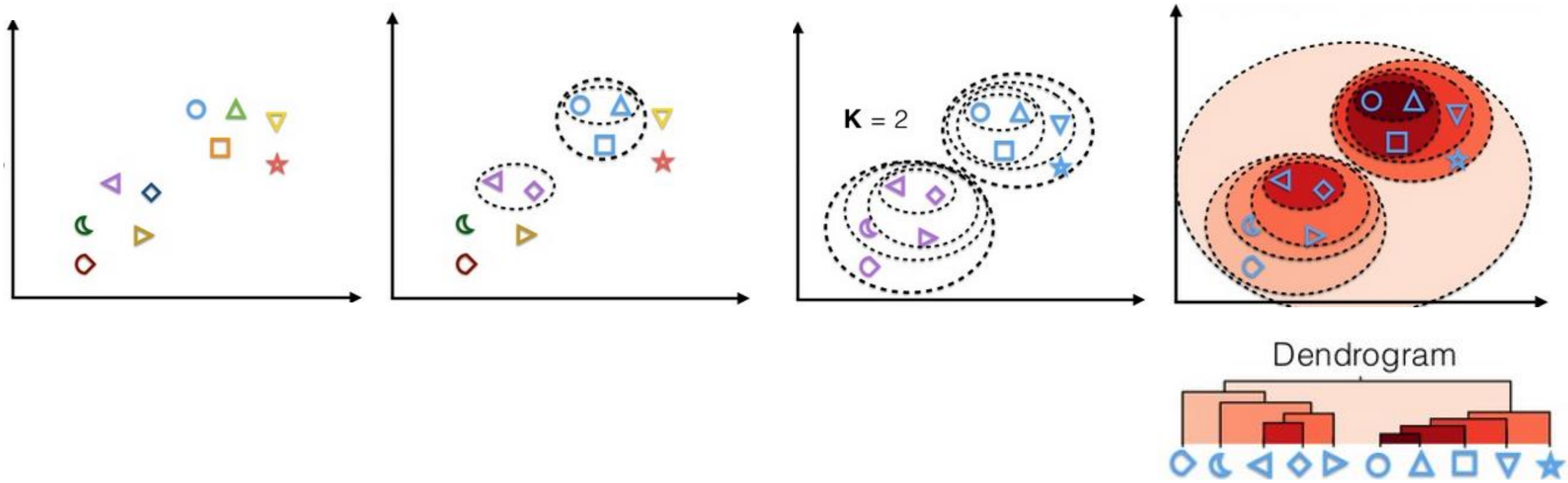
Density-based Clustering

Locates high-density regions separated form one another by regions of low density.

- DBSCAN
- Mean-Shift
- OPTICS

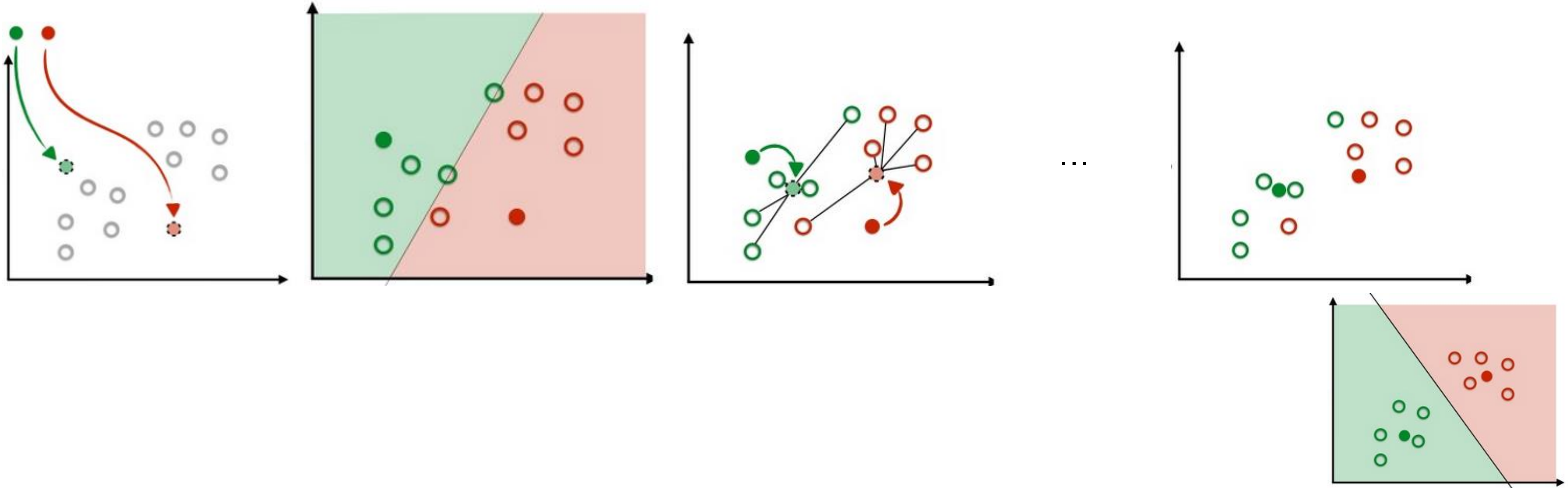
What clustering?

Hierarchical clustering
Finding a cluster-hierarchy, e.g. beginning with as many clusters as data-objects.



What clustering?

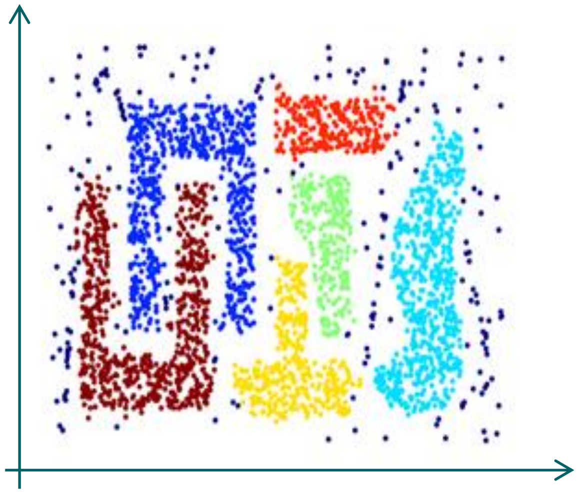
Partitioning Clustering
Optimizing cluster centers to minimize the distance to the data-objects to a cluster.

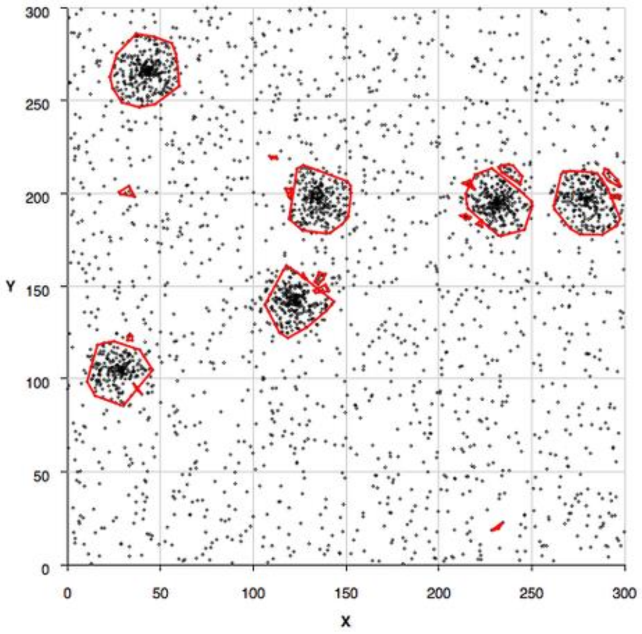


What clustering?

Density-based Clustering
Locates high-density regions separated form one another by regions of low density.

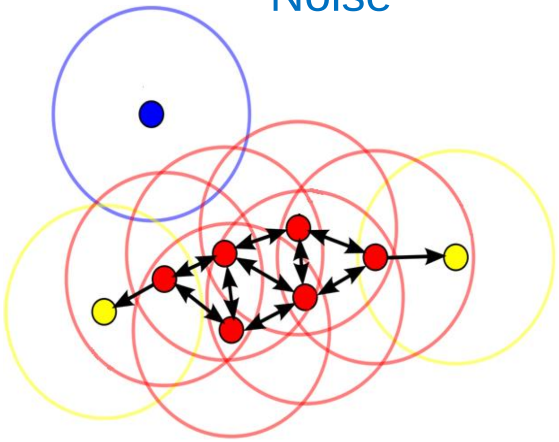
How do they look?





Belonging/borders and noise

Core objects
Border objects
Noise

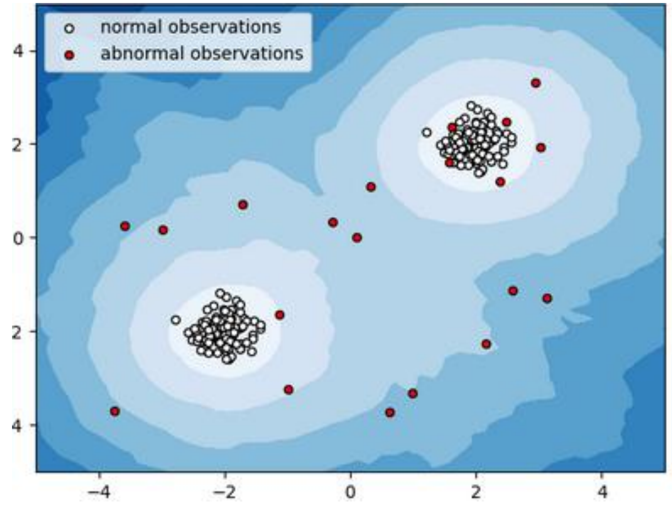


How do we use clustering for anomaly detection?

Using clustering algorithms

→ Objects outside clusters

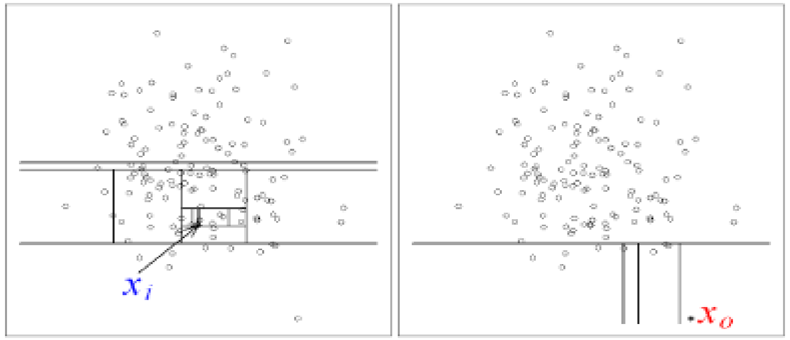
Example:
One-Class Support Vector Machine:



Anomaly Detection algorithm

Example:
Isolation Forest Algorithm

- Objects are isolated by dividing the data space
- Small number of isolation steps
- → High chance for an anomaly



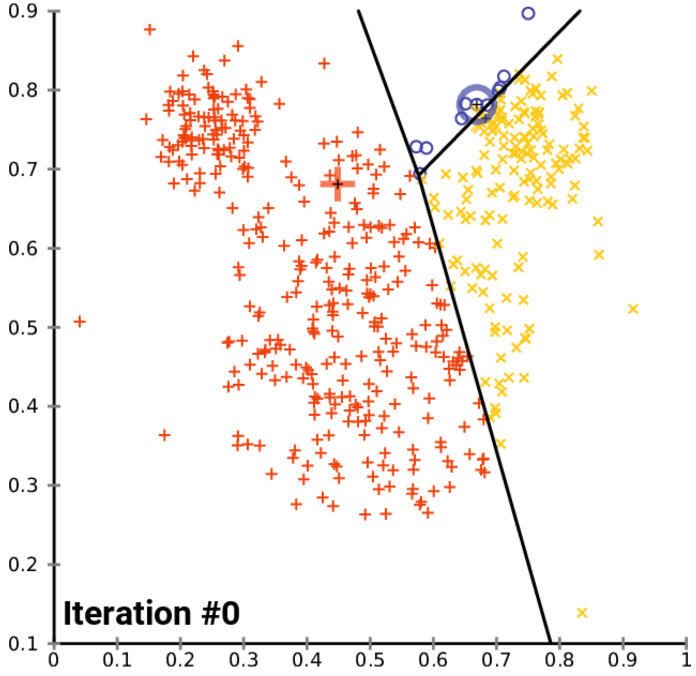
What clustering?

K-means
Optimizing cluster centers to minimize the distance to the data-objects to a cluster.

1) Randomly initialize k data points (means or centroids).

2) Associate each item to its closest mean (based on the squared Euclidean distance)

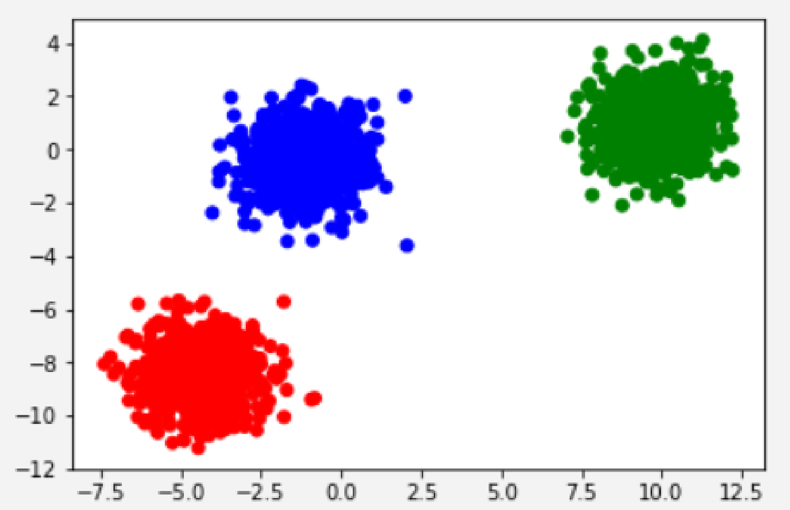
3) Recompute the centroids: Calculate the mean of all items currently associated with the centroid



What clustering?

Example: K-means

```
# %% load data
from sklearn import datasets
import numpy as np
X, y = datasets.make_blobs(n_samples=2000, random_state=45)

# cluster data
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=3, random_state=0).fit(X)
predicted_clusters = kmeans.labels_

# plot
c = [{0:"b",1:"g",2:"r"}[c] for c in predicted_clusters]
plt.scatter(X[:,0],X[:,1],c=c)
```



Association rule mining

What is Association rule mining?

Working Definition

Association rule mining methods try to **discover** interesting **associations** (relationships or dependencies) between **variables** hidden in large datasets of **items**.

Customer 1 milk, bread
Customer 2 bread, butter
Customer 3 beer
Customer 4 milk, bread, butter
Customer 5 bread, butter
…

→ Association rule mining → Rule: bread → butter

What is Association rule mining?

A dataset for association rule mining is composed by **transactions**, which have a unique id and contain a subset of **items.**

$$D = \{t_1, t_2, \ldots, t_m\} \longrightarrow$$

Customer 1 milk, bread
Customer 2 bread, butter
Customer 3 beer
Customer 4 milk, bread, butter
Customer 5 bread, butter
…

$\downarrow$

$$I = \{i_1, i_2, \ldots, i_n\}$$ = {Milk, bread, butter, beer,…}

What is Association rule mining?

An association rule is an implication expression of the form “X → Y” where X and Y are itemsets

Rules evaluation metrics:

- Support (s):
  - is the fraction of transactions that contain both X and Y.
- Confidence (c):
  - measures how often items in Y appear in transactions that contain X.

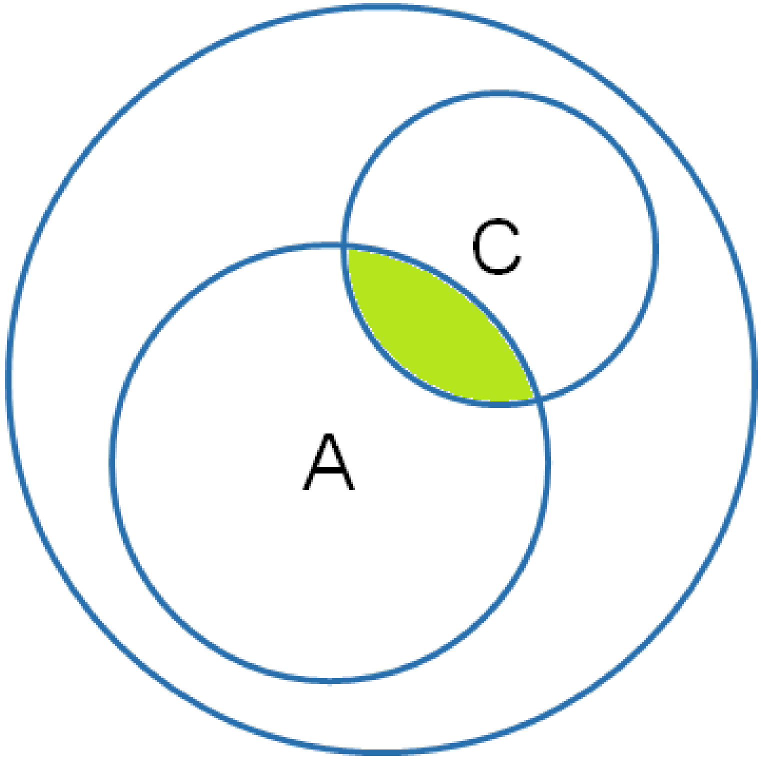
What is Association rule mining?

How does it work?

1) Find rules which have the highest support.

Support

Percentage of instances which match antecedent „A“ and consequent „C“.



What is Association rule mining?

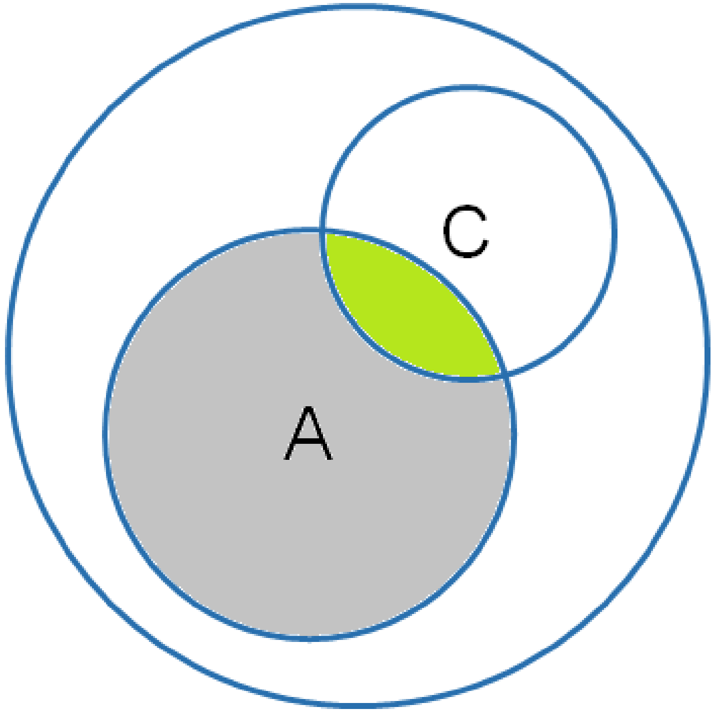
How does it work?

2) Select the relevant rules above a confidence threshold

Confidence

Percentage of instances in the antecedent which also contain the consequent.

Support
Coverage



0% Confidence   100% Confidence

What is Association rule mining?

Brute force approach?

1) List all possible association rules
2) Compute the support and confidence for each rule
3) Prune rules that fail the minimum thresholds

But….. For medium or large datasets it's computationally expensive (seriously, don't do it this way)

What is Association rule mining?

Other approaches: Apriori algorithm

1) Let k=1
2) Generate frequent itemsets of length 1
3) Repeat until no new frequent itemsets are identified
   - Generate length k+1 candidate itemsets from length k that are frequent
   - Prune candidate itemsets containing subsets of length k that are infrequent
   - Compute support
   - Filter infrequent candidates

Other alternatives such as the eclat algorithm also exist, but most are computationally expensive.

What clustering?

Example: apriori

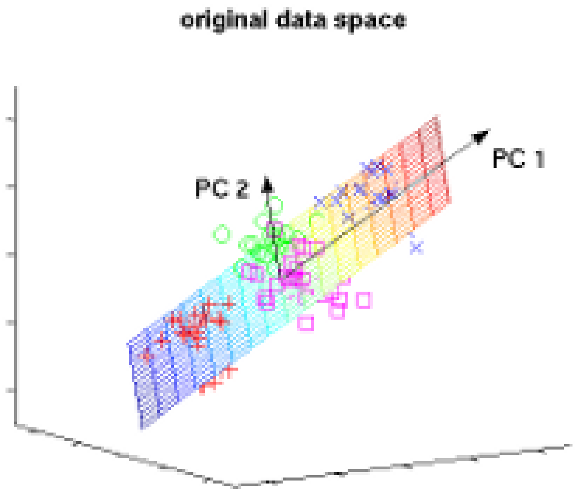
```
from apyori import apriori

transactions = [
['beer', 'nuts'],
['beer', 'cheese'],
['nuts', 'cheese'],
['nuts'],
['nuts'],
['cheese', 'banana'],
['cheese', 'beer'],
['cheese', 'beer']
]

association_rules = list(apriori(transactions))
```

	confidence	rule_if	rule_then	support
0	0.125	banana	-	0.125
1	1.000	banana	cheese	0.125
2	0.250	nuts	beer	0.125
3	0.200	nuts	cheese	0.125
4	0.750	beer	cheese	0.375
5	0.500	beer	-	0.500
6	0.500	nuts	-	0.500
7	0.625	cheese	-	0.625

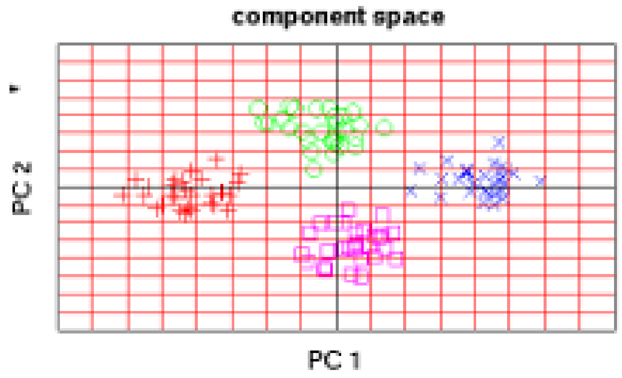
Dimensionality Reduction

What dimensionality reduction?

Working Definition
Dimensionality reduction techniques are meant to **reduce** the number of **dimensions** of a featureset.

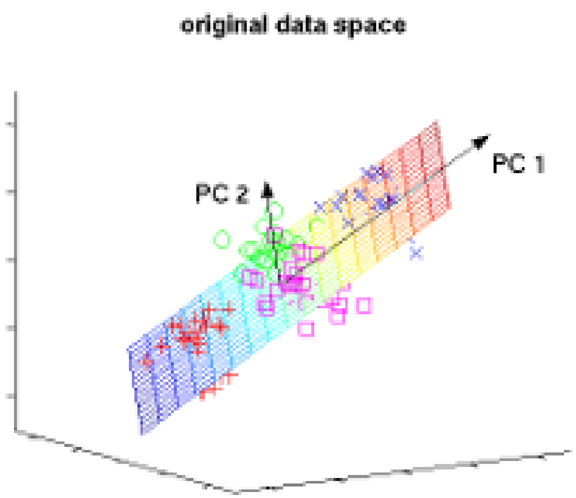


Dimensionality reduction

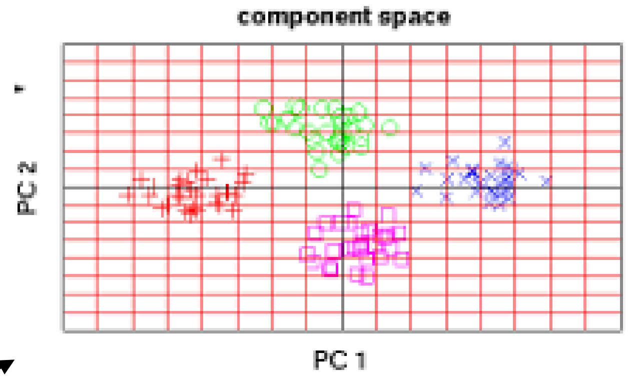


What dimensionality reduction?

Working Definition
Dimensionality reduction techniques are meant to **reduce** the number of **dimensions** of a featureset.



Dimensionality reduction



$$P : \mathbb{R}^n \rightarrow \mathbb{R}^q$$

What dimensionality reduction?

Why?

- Storage space
- Model complexity / computational cost
- Curse of dimensionality can affect some models
- Helps visualize data (humans are not good at understanding many dimensions)

What dimensionality reduction?

How?

- Feature selection
- Linear projections (normally component based)
- Non linear projections (Manifold learning)

What dimensionality reduction?

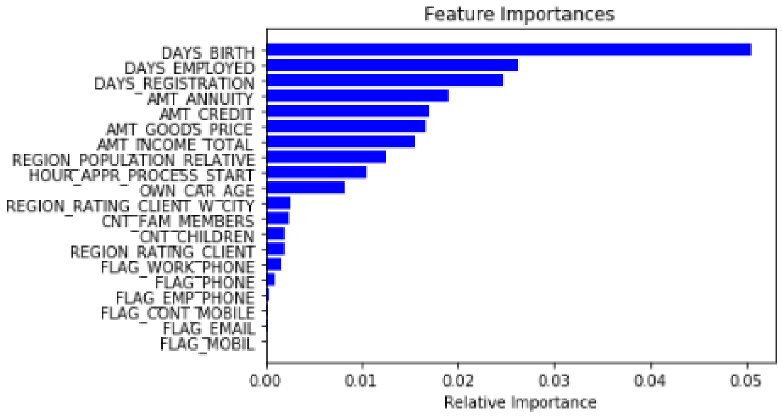
Feature selection
Techniques that focus on only keeping the most relevant variables from the original featureset

Feature filtering

- Missing values
- Low variance
- Correlation filter

Random forest

Gini importance (Mean Decrease in Impurity)

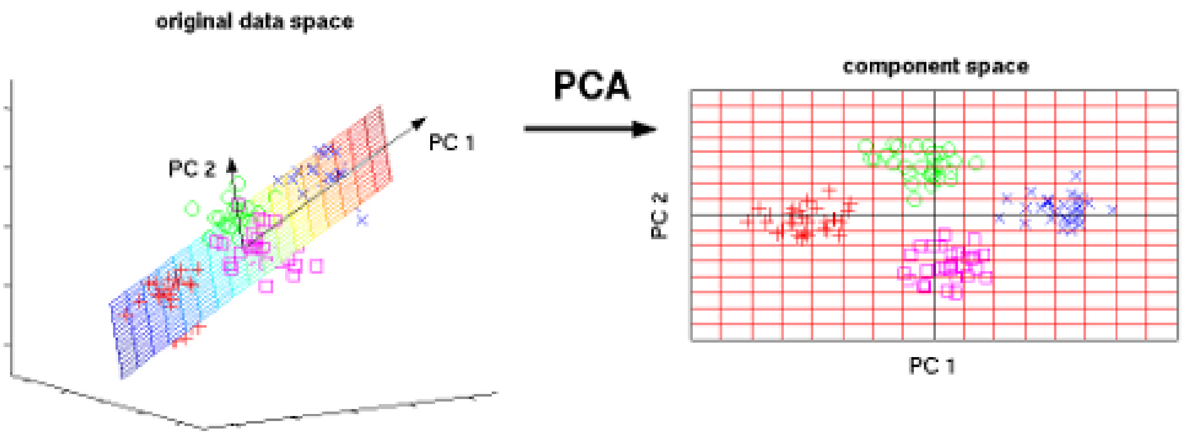


What dimensionality reduction?

Linear projections (normally component based)

These family of techniques look for the main components of the data. Common working principles are to look for uncorrelated variables, to group correlated variables or to search eigenvectors of latent variables.

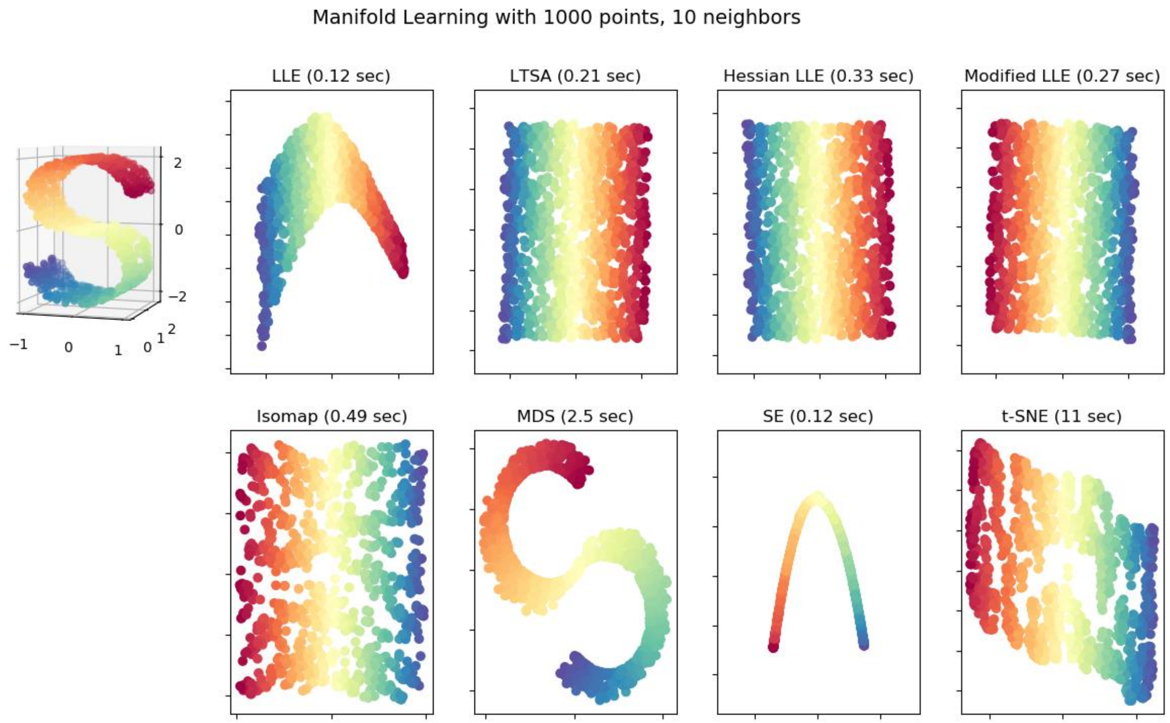
- Principal Component Analysis(PCA)
- Singular Value Decomposition(SVD)
- Independent Component Analysis(ICA)



What dimensionality reduction?

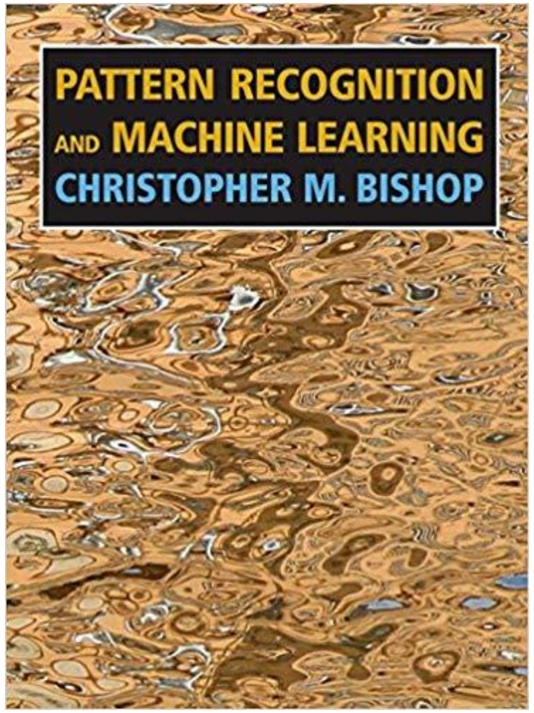
Non linear projections (Manifold learning)

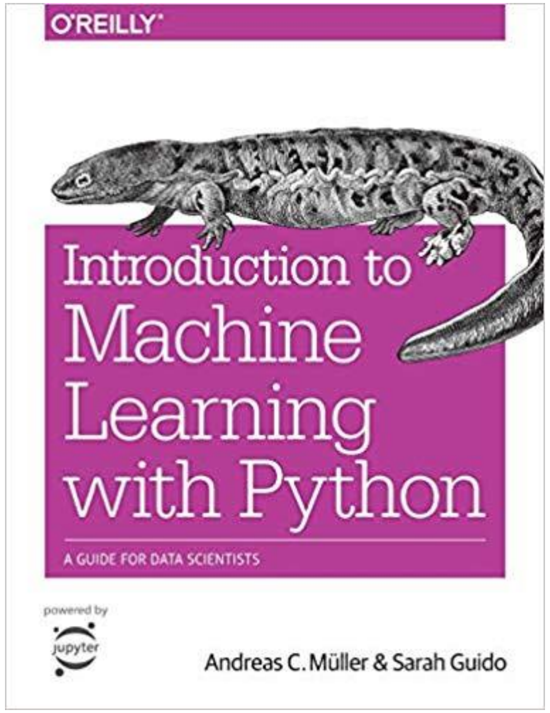
Algorithms on this family are based on the idea that the dimensionality of many datasets is artificially high. They try to maintain local or global distance metrics while transforming the feature space.



Further Reading Material

- https://www.youtube.com/watch?v=bQI5uDxrFfA [Introduction Supervised Learning, Andrew Ng]
- https://scikit-learn.org/stable/supervised_learning.html [Short Explanation and Code Snippets]





Thank you for your attention!

Lecture Team AIDAE
aidae@ima-ifu.rwth-aachen.de