HANDS-ON AI I

Tabular Data, Dimensionality Reduction and Clustering

Andreas Schörgenhumer
Institute for Machine Learning

JOHANNES KEPLER
UNIVERSITY LINZ

Institute for
Machine Learning

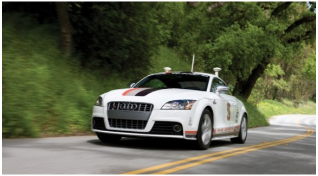
Copyright Statement

Content of Unit 1

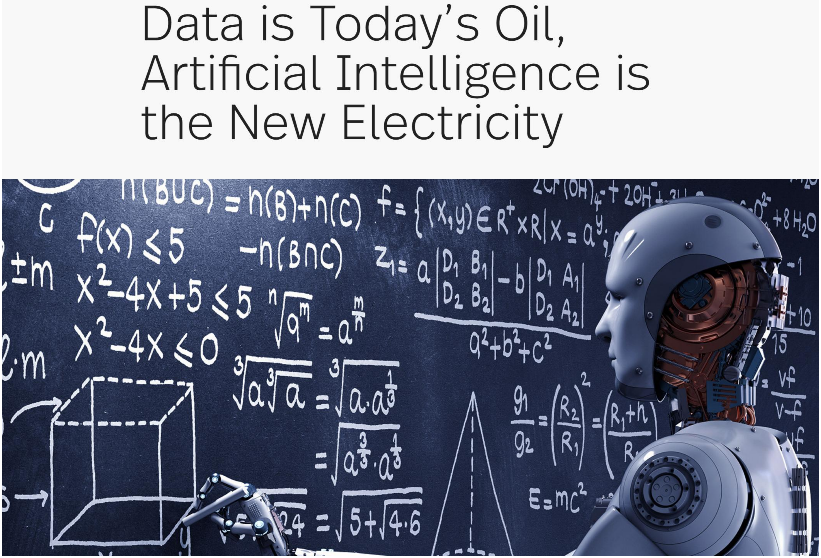
- Short motivation
- First data source: tabular data
- Dimensionality reduction
- Clustering

AI is Ubiquitous

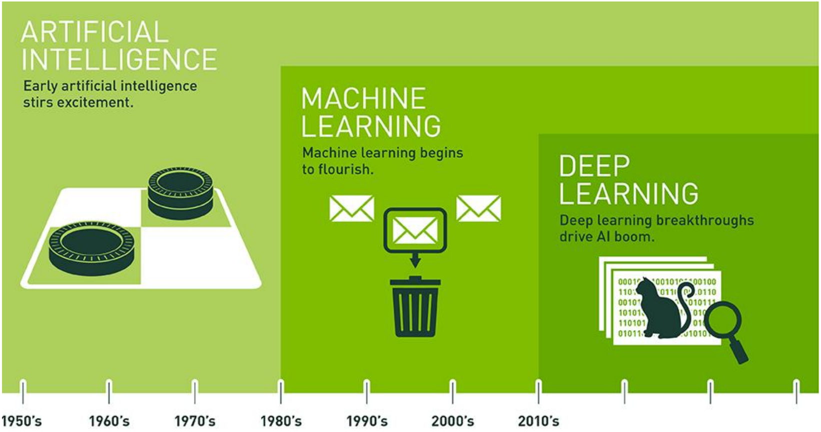
- AI **pervades commercial applications** in an unprecedented manner and is fundamentally changing how businesses operate across **virtually all sectors**:
  - Information technology
  - Manufacturing and supply chains
  - Medicine and healthcare
  - Education
  - Financial, legal and tax services
  - News and publishing
  - Transportation
  - ...
  - Science



Golden Age of AI



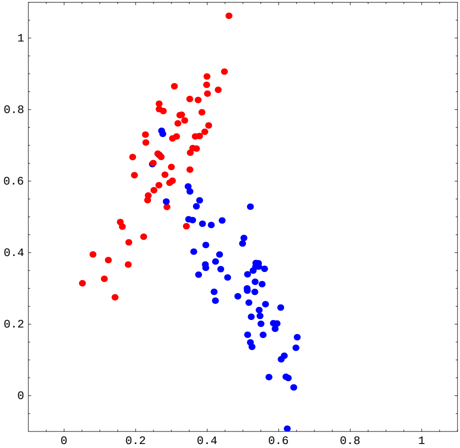
AI is a Broad Field



Data


Geographical
Transport
Cultural
Natural
Scientific
Types of Data
Meteorological
Financial
Statistical

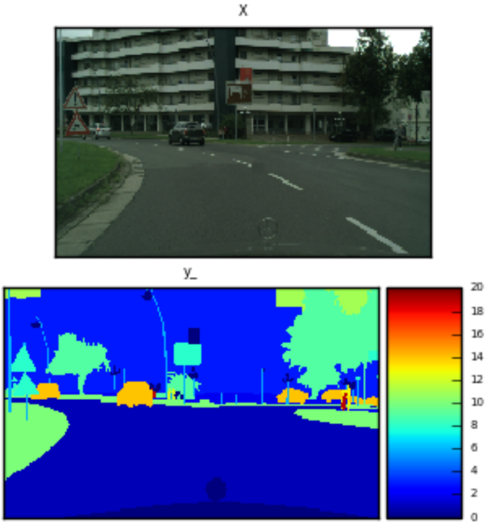
Example Data (1)



Example Data (2)

0.99516	0.890813	0.933726	0.793397	0.826405	0.236946	-1
0.853206	0.611647	0.317486	0.633609	0.411492	0.985231	+1
0.387494	0.459847	0.815049	0.394526	0.678227	0.031886	-1
0.733515	0.640438	1.19068	0.639685	0.0793674	0.160503	+1
0.274817	0.261054	1.20056	0.689895	0.401913	0.277955	-1
0.329943	0.241299	0.848705	0.721673	0.973852	0.795238	-1
0.334784	0.350487	0.315131	0.928277	0.816343	0.558292	-1
0.481578	0.738839	0.0925513	0.294667	0.612725	0.573062	-1
0.0940846	0.278992	0.451819	0.900141	0.220497	0.541176	+1
0.360569	0.638554	1.0307	0.260456	0.00658296	0.380672	+1
0.0857518	0.3775	0.386551	0.570562	0.15437	0.102717	+1
0.755808	0.1362	0.544536	0.848888	0.874862	0.307479	-1
0.421025	0.785714	0.449038	0.920612	0.420418	0.749187	-1
0.939446	0.0468747	0.15846	0.625944	0.198894	0.176125	+1
0.845362	0.767883	0.824993	0.725803	0.808218	0.63495	-1
0.484793	0.129329	0.0783719	0.465347	0.291457	0.254278	+1
0.399041	0.751829	0.763511	0.894785	0.47902	0.15156	-1
0.643232	0.615629	0.430261	0.0458972	0.446513	0.844081	+1
...	...	...	...	...	...	...

Example Data (3)



What is Data?

- Etymologically, data is the plural of datum in Latin, which means “given”.
- Data is typically **generated from a real world process** (e.g., measurements), but **synthetic** data also exists.

One Example

- Our ears measure differences in the pressure from the surrounding air:
    - Our ears transform the differences to signals.
    - Signals are further processed and represented as sound.

One Example

- Our ears measure differences in the pressure from the surrounding air:
  - Our ears transform the differences to signals.
  - Signals are further processed and represented as sound.
- We could also convert the pressure differences into an electrical signal via a microphone:
  - Convert analog signal into a digital signal.
  - Save the resulting binary symbols to a hard disk.

One Example

- Our ears measure differences in the pressure from the surrounding air:
  - Our ears transform the differences to signals.
  - Signals are further processed and represented as sound.
- We could also convert the pressure differences into an electrical signal via a microphone:
  - Convert analog signal into a digital signal.
  - Save the resulting binary symbols to a hard disk.
- We now present the process of varying changes in the air pressure as zeros and ones.

One Example

- Our ears measure differences in the pressure from the surrounding air:
  - Our ears transform the differences to signals.
  - Signals are further processed and represented as sound.
- We could also convert the pressure differences into an electrical signal via a microphone:
  - Convert analog signal into a digital signal.
  - Save the resulting binary symbols to a hard disk.
- We now present the process of varying changes in the air pressure as zeros and ones.
- **Binary representation** of data is the **basis of computerized data processing** at present.

TABULAR DATA

Tabular Data

- Data type of today's lecture/exercise.

Tabular Data

- Data type of today's lecture/exercise.
- Data is structured in a tabular form.

Tabular Data

- Data type of today's lecture/exercise.
- Data is structured in a tabular form.
- Data elements are arranged in vertical columns and horizontal rows.

Tabular Data

- Data type of today's lecture/exercise.
- Data is structured in a tabular form.
- Data elements are arranged in vertical columns and horizontal rows.
- Each column and row is uniquely numbered.

Tabular Data

- Data type of today’s lecture/exercise.
- Data is structured in a tabular form.
- Data elements are arranged in vertical columns and horizontal rows.
- Each column and row is uniquely numbered.
- Tabular data has a virtually infinite range for mass data storage (can always add rows).

Tabular Data

- Data type of today's lecture/exercise.
- Data is structured in a tabular form.
- Data elements are arranged in vertical columns and horizontal rows.
- Each column and row is uniquely numbered.
- Tabular data has a virtually infinite range for mass data storage (can always add rows).
- Tabular databases include the following key properties:
  - Share the same set of properties per record, i.e., every row has the same column titles.
  - Each column is (usually) assigned with a header title (metadata).
  - Access through identifiers, i.e., each object can be retrieved by a query through key values.

Example: Iris Data Set

- **Iris flower data set** of **Fisher's Iris data set** is a famous data set introduced by British statistician Ronald Fisher.
- It is also sometimes called **Anderson's Iris data set** since biologist Edgar Anderson collected the data.

Example: Iris Data Set

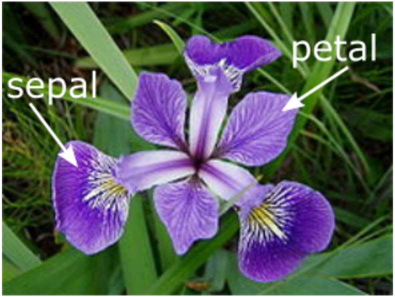
- **Iris flower data set** of **Fisher's Iris data set** is a famous data set introduced by British statistician Ronald Fisher.
- It is also sometimes called **Anderson's Iris data set** since biologist Edgar Anderson collected the data.
- The data set consists of 50 samples from each of three species of the **Iris flower**:
  - Iris setosa
  - Iris virginica
  - Iris versicolor

Example: Iris Data Set

We have the following $d = 4$ **features**:

- Sepal length in cm
- Sepal width in cm
- Petal length in cm
- Petal width in cm



Terminology

sep-len	sep-width	pet-len	pet-width	species
6.7	3.1	4.7	1.5	versicolor
6.7	3.1	4.4	1.4	versicolor
6.5	3.2	5.1	2.0	virginica
5.0	3.6	1.4	0.2	setosa
6.5	3.0	5.8	2.2	virginica
...	...	...	...	...

Terminology

sep-len	sep-width	pet-len	pet-width	species
6.7	3.1	4.7	1.5	versicolor
6.7	3.1	4.4	1.4	versicolor
6.5	3.2	5.1	2.0	virginica
5.0	3.6	1.4	0.2	setosa
6.5	3.0	5.8	2.2	virginica
...	...	...	...	...

- Every flower entry is referred to as a **sample**.

Terminology

sep-len	sep-width	pet-len	pet-width	species
6.7	3.1	4.7	1.5	versicolor
6.7	3.1	4.4	1.4	versicolor
6.5	3.2	5.1	2.0	virginica
5.0	3.6	1.4	0.2	setosa
6.5	3.0	5.8	2.2	virginica
...	...	...	...	...

- Every flower entry is referred to as a **sample**.
- Every sample is described by 4 **features** (sep-len, sep-width, pet-len, pet-width), which can be represented as a **feature vector**, e.g.: $\boldsymbol{x} = (6.7, 3.1, 4.7, 1.5)$.

Terminology

sep-len	sep-width	pet-len	pet-width	species
6.7	3.1	4.7	1.5	versicolor
6.7	3.1	4.4	1.4	versicolor
6.5	3.2	5.1	2.0	virginica
5.0	3.6	1.4	0.2	setosa
6.5	3.0	5.8	2.2	virginica
...	...	...	...	...

- Every flower entry is referred to as a **sample**.
- Every sample is described by 4 **features** (sep-len, sep-width, pet-len, pet-width), which can be represented as a **feature vector**, e.g.: $\boldsymbol{x} = (6.7, 3.1, 4.7, 1.5)$.
- There are 3 species (setosa, virginica, versicolor), which means that there are 3 **classes**.

Terminology

sep-len	sep-width	pet-len	pet-width	species
6.7	3.1	4.7	1.5	versicolor
6.7	3.1	4.4	1.4	versicolor
6.5	3.2	5.1	2.0	virginica
5.0	3.6	1.4	0.2	setosa
6.5	3.0	5.8	2.2	virginica
...	...	...	...	...

- Every flower entry is referred to as a **sample**.
- Every sample is described by 4 **features** (sep-len, sep-width, pet-len, pet-width), which can be represented as a **feature vector**, e.g.: $\boldsymbol{x} = (6.7, 3.1, 4.7, 1.5)$.
- There are 3 species (setosa, virginica, versicolor), which means that there are 3 **classes**.
- Every sample lists the species/class via its **label**, e.g.: $y =$ versicolor.

Example: Wine Data Set

- The **wine data set** comprises the results of a chemical analysis of wines grown in the same region in Italy, but derived from three different cultivars/cultivators (3 classes).

Example: Wine Data Set

- The **wine data set** comprises the results of a chemical analysis of wines grown in the same region in Italy, but derived from three different cultivars/cultivators (3 classes).
- The analysis determined the quantities of 13 constituents found in each of the three types of wines.

Example: Wine Data Set

- The **wine data set** comprises the results of a chemical analysis of wines grown in the same region in Italy, but derived from three different cultivars/cultivators (3 classes).
- The analysis determined the quantities of 13 constituents found in each of the three types of wines.
- The data set consists of 178 samples with 13 features (13 constituents).

Example: Wine Data Set

We have the following $d = 13$ **features**:

- Alcohol
- Malic acid
- Ash
- Alcalinity of ash
- Magnesium
- Total phenols
- Flavanoids
- Nonflavanoid phenols
- Proanthocyanins
- Color intensity
- Hue
- OD280/OD315 of diluted wines
- Proline

VISUALIZATION

Visualization

- Gaining **insights** into your data is essential and is often one of the first steps.

Visualization

- Gaining **insights** into your data is essential and is often one of the first steps.
- Looking at the raw data is often infeasible or does not help (too much data, too many features).

Visualization

- Gaining **insights** into your data is essential and is often one of the first steps.
- Looking at the raw data is often infeasible or does not help (too much data, too many features).
- **Visualization** can be a powerful tool in this regard.

Visualization

- Gaining **insights** into your data is essential and is often one of the first steps.
- Looking at the raw data is often infeasible or does not help (too much data, too many features).
- **Visualization** can be a powerful tool in this regard.
- Visualization is highly dependent on the data you are dealing with:
  - Individual features: histograms

Visualization

- Gaining **insights** into your data is essential and is often one of the first steps.
- Looking at the raw data is often infeasible or does not help (too much data, too many features).
- **Visualization** can be a powerful tool in this regard.
- Visualization is highly dependent on the data you are dealing with:
    - Individual features: histograms
    - 2-dimensional data: scatter plots

Visualization

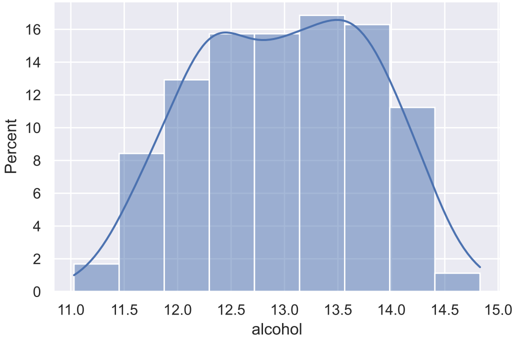
- Gaining **insights** into your data is essential and is often one of the first steps.
- Looking at the raw data is often infeasible or does not help (too much data, too many features).
- **Visualization** can be a powerful tool in this regard.
- Visualization is highly dependent on the data you are dealing with:
  - Individual features: histograms
  - 2-dimensional data: scatter plots
  - Time series data: line plots

Visualization

- Gaining **insights** into your data is essential and is often one of the first steps.
- Looking at the raw data is often infeasible or does not help (too much data, too many features).
- **Visualization** can be a powerful tool in this regard.
- Visualization is highly dependent on the data you are dealing with:
  - Individual features: histograms
  - 2-dimensional data: scatter plots
  - Time series data: line plots
  - Labeled data $\rightarrow$ separation into classes: combined plots with class-color encoding
  - etc.

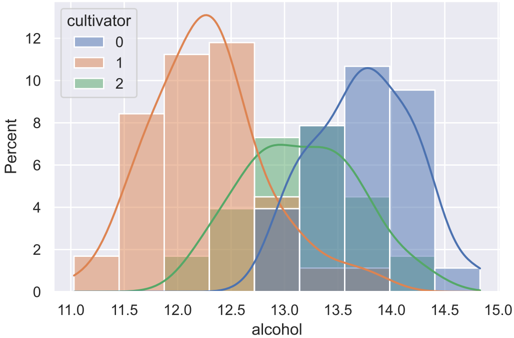
Example: Wine Data Set

- Visualize the feature `alcohol` via a histogram:



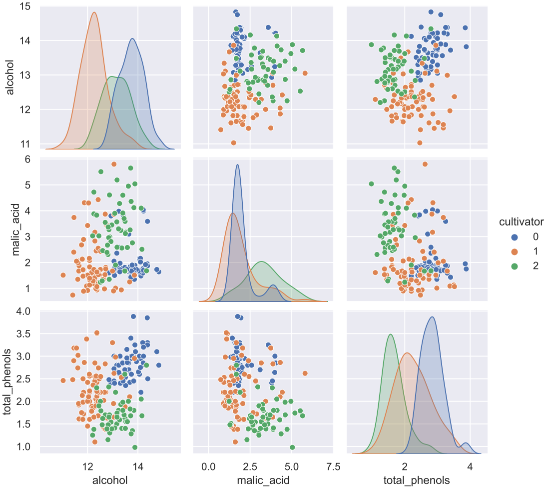
Example: Wine Data Set

- Visualize the feature `alcohol` via a histogram and separate the three cultivators (classes):



Example: Wine Data Set

- Visualize multiple features simultaneously by always comparing pairs of features (including class separation):



Visualization Problems

- The wine data set has 13 features: We can still look at all of them manually and make comparisons with each other.

Visualization Problems

- The wine data set has 13 features: We can still look at all of them manually and make comparisons with each other.
- What about bigger data sets? Let's say 500 features?
  - Of course, we could still look at all features individually or compare features pair-wise.

Visualization Problems

- The wine data set has 13 features: We can still look at all of them manually and make comparisons with each other.
- What about bigger data sets? Let's say 500 features?
  - Of course, we could still look at all features individually or compare features pair-wise.
  - Would probably take a "couple" of minutes ...

DIMENSIONALITY REDUCTION

Dimensionality Reduction

- Problem: Too many features to see anything in the data.
- Often, data is described with hundreds (or thousands) of features $\rightarrow$ visualization is a common problem.
- Idea: **Reduce dimensionality** of the data set, while still preserving as much information as possible.

Dimensionality Reduction

- Problem: Too many features to see anything in the data.
- Often, data is described with hundreds (or thousands) of features $\rightarrow$ visualization is a common problem.
- Idea: **Reduce dimensionality** of the data set, while still preserving as much information as possible.
- Popular algorithms are **PCA** (principal component analysis) or **t-SNE** (t-distributed stochastic neighbor embedding).

Dimensionality Reduction

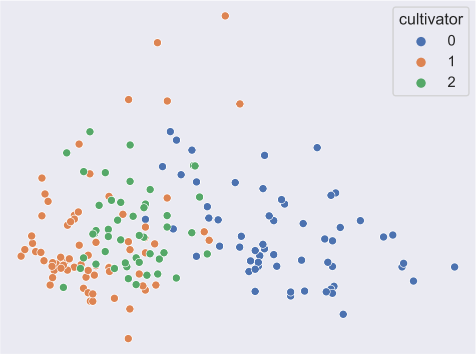
- Problem: Too many features to see anything in the data.
- Often, data is described with hundreds (or thousands) of features $\rightarrow$ visualization is a common problem.
- Idea: **Reduce dimensionality** of the data set, while still preserving as much information as possible.
- Popular algorithms are **PCA** (principal component analysis) or **t-SNE** (t-distributed stochastic neighbor embedding).
- Can reduce n-dimensional data to, e.g., 2-dimensional data $\rightarrow$ can be easily visualized.

Example: Wine Data Set

- Reduce the 13 features down to a 2-dimensional space.

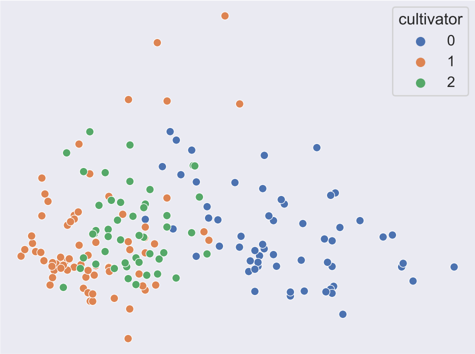
Example: Wine Data Set

- Reduce the 13 features down to a 2-dimensional space.
- Resulting 2D data can be visualized with a scatter plot:



Example: Wine Data Set

- Reduce the 13 features down to a 2-dimensional space.
- Resulting 2D data can be visualized with a scatter plot:



- While we lose some information, we quickly gain interesting insights: Samples from the same cultivar form a so called **cluster** (close to each other in space).

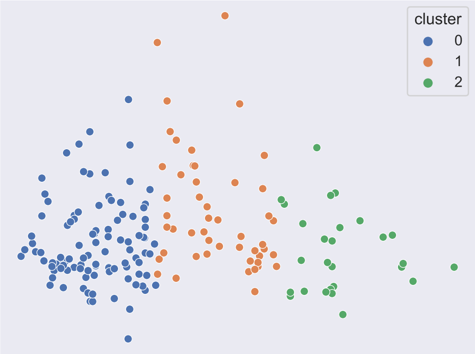
CLUSTERING

Clustering Algorithms

- So far, all our data was labeled.
- Imagine now that the data is unlabeled and we still want to find out which data belongs together.

Clustering Algorithms

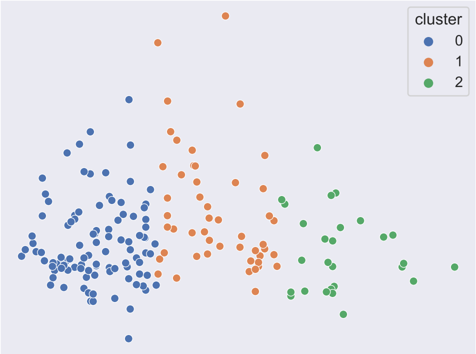
- So far, all our data was labeled.
- Imagine now that the data is unlabeled and we still want to find out which data belongs together.
- We can now use so-called **clustering algorithms** that try to group samples into "similar" and "dissimilar" samples.[1]



[1] What is considered "similar" highly depends on the algorithm.

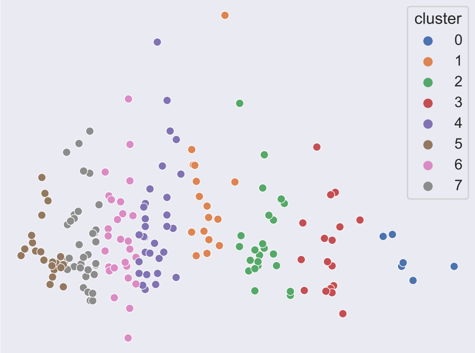
k-means

- In the k-means clustering algorithm, k is the most important parameter.
- k determines how many clusters the algorithm should search for.
- k is set by the user.



Affinity Propagation

- For affinity propagation, the number of clusters does not have to be specified.
- For the wine data set, affinity propagation determines 8 cluster centers and assigns points to them.



Notes on Clustering

- Clustering is **not** classification (which we will discuss in later units).
- On the contrary, the clustering algorithms we pass our data to do not have any knowledge about the classes/labels, they just receive the raw features of the samples (unlabeled data).

Notes on Clustering

- Clustering is **not** classification (which we will discuss in later units).
- On the contrary, the clustering algorithms we pass our data to do not have any knowledge about the classes/labels, they just receive the raw features of the samples (unlabeled data).
- This also means that we often do not really know whether the identified clusters are “correct” $\rightarrow$ must inspect again (e.g., with the help of visualization) to see if they make sense.

SUMMARY

Summary

- Tabular data is very common.
- Data is structured in a tabular form.
- Data elements are arranged in columns (features, labels) and rows (samples).
- Visualization is a powerful tool to gain insights into the data.
- High-dimensional data can be handled with dimensionality reduction techniques.
- Clustering allows to find samples “close” to each other.
- Note: The described methods like dimensionality reduction or clustering can also be applied to other forms of data.