# Data Preprocessing

Sibylle Hess

Technische Universiteit
Eindhoven
University of Technology

# Notation: Assumptions for this Lecture

| ID | $\mathtt{F}_1$ | $\mathtt{F}_2$ | ... | $\mathtt{F}_d$ |
|---|---|---|---|---|
| 1 | 5.1 | 3.5 | ... | 0.2 |
| 2 | 6.4 | 3.5 | ... | 1.2 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |

We assume that we have a dataset with $n$ observations of $d$ features, which is represented by a $D \in \mathbb{R}^{n \times d}$-dimensional data matrix

$\mu_{\mathtt{F}_k} = \frac{1}{n} \sum_{i=1}^{n} D_{ik}$ The vector of all feature sample mean values

$\mu_{\mathtt{F}} = (\mu_{\mathtt{F}_k})_k \in \mathbb{R}^d$ The vector of all feature sample mean values

$\sigma^2_{\mathtt{F}_k} = \frac{1}{n} \sum_{i=1}^{n} (D_{ik} - \mu_{\mathtt{F}_k})^2$ Sample variance of feature $\mathtt{F}_k$

# Basic Data Characteristics

# Raw Data is Usually not "Clean"

## Covid: Man offered vaccine after error lists him as 6.2cm tall

18 February 2021

Coronavirus pandemic


Liam Thorp was wrongly classed as morbidly obese according to his height and weight

# Finding a Good Data Representation

| ID | height[m] | eyes | children | sex |
|---|---|---|---|---|
| 1 | 1.8 | blue | 0 | m |
| 2 | **0.4** | green | 2 | f |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $n$ | 1.65 | brown | **NA** | f |

↓

| ID | height [m] | eyes | | | children | sex |
|---|---|---|---|---|---|---|
| | | bl | gr | br | | |
| 1 | 1.8 | 1 | 0 | 0 | 0 | 0 |
| 2 | 1.6 | 0 | 1 | 0 | 2 | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $n$ | 1.65 | 0 | 0 | 1 | 0 | 1 |

Finding a suitable representation of your data starts with the encoding, for which it's important to identify the feature types first.

Feature types are distinguished into qualitative and quantitative ones.

# Qualitative Features

Nominal Data:

- categories
- no order

| ID | eyecolor | sex | zip code |
|---|---|---|---|
| 1 | blue | m | 5629NZ |
| 2 | green | f | 5381 |
| ⋮ | ⋮ | ⋮ | ⋮ |

Ordinal Data:

- categories
- with order
- subjective scale

| ID | rating | size | grade |
|---|---|---|---|
| 1 | ★★★☆☆ | S | 7.5 |
| 2 | ★★☆☆☆ | XL | 9 |
| ⋮ | ⋮ | ⋮ | ⋮ |

# Quantitative Features

Discrete Data:

- countable amount of values
- distances are meaningful

| ID | age | nr. children | nr. yes votes |
|---|---|---|---|
| 1 | 22 | 2 | 1298 |
| 2 | 24 | 0 | 2780 |
| ⋮ | ⋮ | ⋮ | ⋮ |

Continuous Data:

- measurements
- distances are meaningful

| ID | income[€] | height[m] | temp[°C] |
|---|---|---|---|
| 1 | 2260.35 | 1.75 | 23 |
| 2 | 3502.60 | 1.68 | 26 |
| ⋮ | ⋮ | ⋮ | ⋮ |

It's not always easy to identify the exact feature type.

However, being aware of your feature types is important, because **NOT ALL ENCODINGS AND STATISTICS MAKE SENSE FOR ALL TYPES**.

# Ordinal Encoding

Map ordinal values to integers (be careful keeping the order).

| ID | pic | size |
|---|---|---|
| 1 | dog | S |
| 2 | cat | XL |
| ⋮ | ⋮ | ⋮ |
| $n$ | fish | S |

$\text{cat} \rightarrow 0$
$\text{dog} \rightarrow 1$
$\text{fish} \rightarrow 2$
$\text{bird} \rightarrow 3$

$\text{XS} \rightarrow 0$
$\text{S} \rightarrow 1$
$\text{M} \rightarrow 2$
$\text{L} \rightarrow 3$
$\text{XL} \rightarrow 4$

$\rightarrow$

| ID | pic | size |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 0 | 4 |
| ⋮ | ⋮ | ⋮ |
| $n$ | 2 | 1 |

**PROBLEMATIC:** suggests that the feature can be treated like a discrete one when the feature is qualitative.

Usually applied for class labels (nominal!) and ordinal data

# One-Hot Encoding

Create for every value a new binary feature.

| ID | pic |
|---|---|
| 1 | dog |
| 2 | cat |
| ⋮ | ⋮ |
| $n$ | fish |

$\rightarrow$

| ID | pic | | | |
|---|---|---|---|---|
| | cat | dog | fish | bird |
| 1 | 0 | 1 | 0 | 0 |
| 2 | 1 | 0 | 0 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $n$ | 0 | 0 | 1 | 0 |

**PROBLEMATIC:** increases the dimensionality of the feature space

Usually applied for nominal data

# Statistics Depending on Feature Types

Let $x_1, \ldots, x_n$ be a set of feature values and let $x_{(1)}, \ldots, x_{(n)}$ be the ordered sequence of feature values ($x_{(i)} \leq x_{(i+1)}$)

| statistic | | permissible for | |
|---|---|---|---|
| | | nominal | ordinal |
| mode | most frequent value | yes | yes |
| median | $\begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ \frac{x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)}}{2} & \text{otherwise} \end{cases}$ | no | yes |
| mean | $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ | no | no |
| variance | $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2$ | no | no |

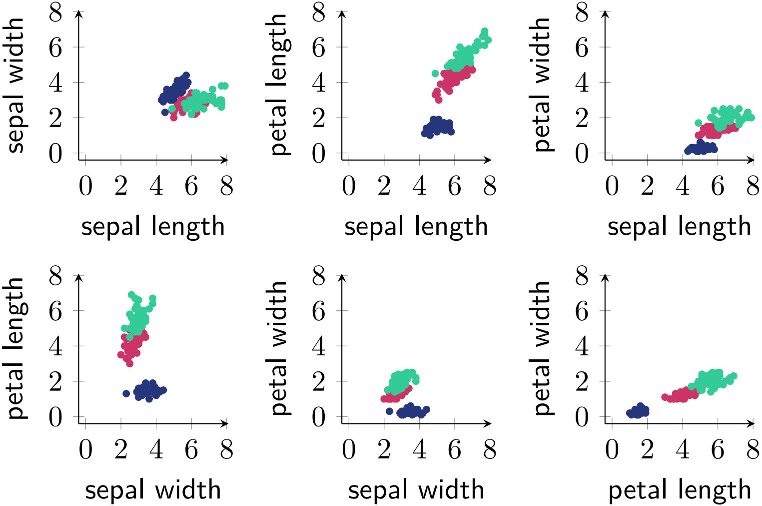
In addition to looking at statistics, it's always good to plot your data.

For this, we have a look at the Iris dataset.

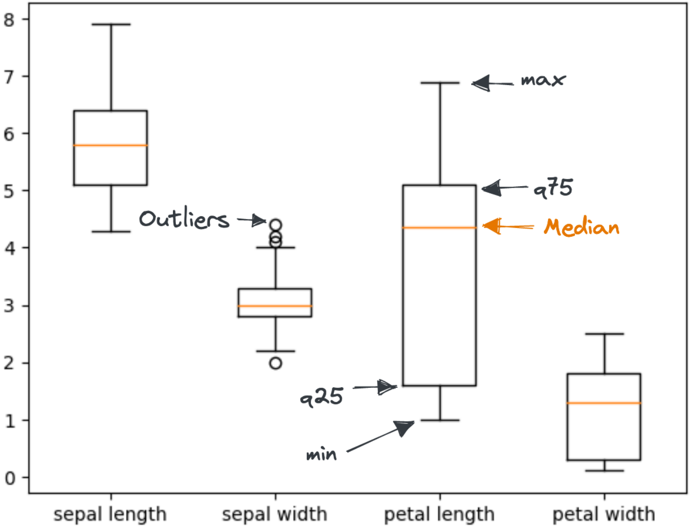
# Exploring the Iris Dataset


| sepal length | sepal width | petal length | petal width | class |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 6.4 | 3.5 | 4.5 | 1.2 | versicolor |
| 5.9 | 3.0 | 5.0 | 1.8 | virginica |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

# Scatter Matrix Plot



# Boxplot


8
7
6
5
4
3
2
1
0
max
q75
Median
Outliers
q25
min
sepal length
sepal width
petal length
petal width

# Feature Transforma- tions

# Scaling

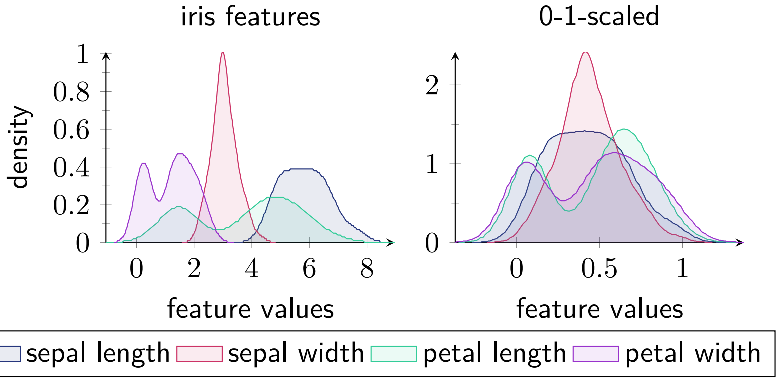
Scaling is used to make features comparable.

- If one feature is measured in kg and the other in g, differences in the feature in g might seem more important
- Most ML methods learn based on similarities/distances between data points. These similarities/distances might be distorted when some of the features have much bigger values than others

# Min-Max Scaling a.k.a 0-1 Scaling

Feature $\mathtt{F}_k$ is scaled to the range $[0, 1]$.

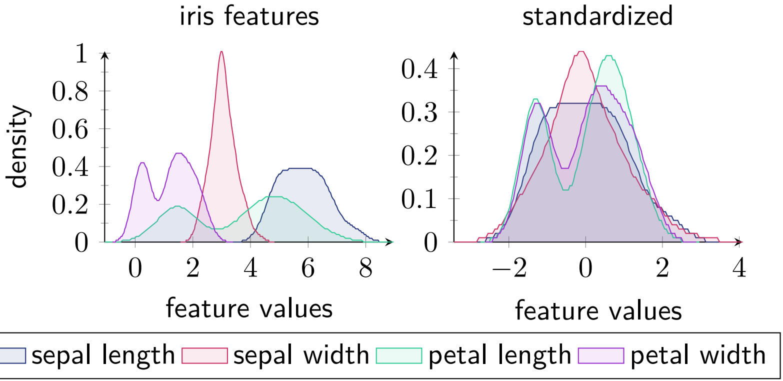
$$\hat{D}_{ik} = \frac{D_{ik} - \min(D_{\cdot k})}{\max(D_{\cdot k}) - \min(D_{\cdot k})}$$



# Standardization

Feature $\mathtt{F}_k$ gets sample mean-centered and scaled to a sample standard deviation of one.

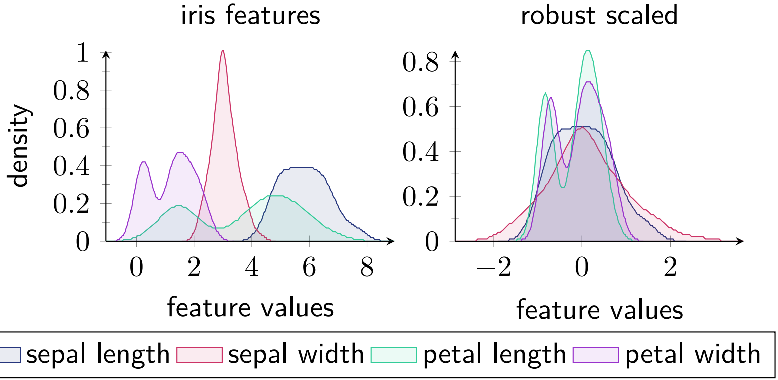
$$\hat{D}_{ik} = \frac{D_{ik} - \mu_{\mathtt{F}_k}}{\sigma_{\mathtt{F}_k}}$$



# Robust Scaling

Feature $\mathtt{F}_k$ gets median-centered and scaled by the inter quantile range

$$\hat{D}_{ik} = \frac{D_{ik} - \mathrm{median}(D_{\cdot k})}{q_{75}(D_{\cdot k}) - q_{25}(D_{\cdot k})}$$



# Scaler Comparison

**MIN-MAX SCALING**: might make sense for ordinal data, but is not robust against outliers

$$\hat{D}_{ik} = \frac{D_{ik} - \min(D_{\cdot k})}{\max(D_{\cdot k}) - \min(D_{\cdot k})}$$

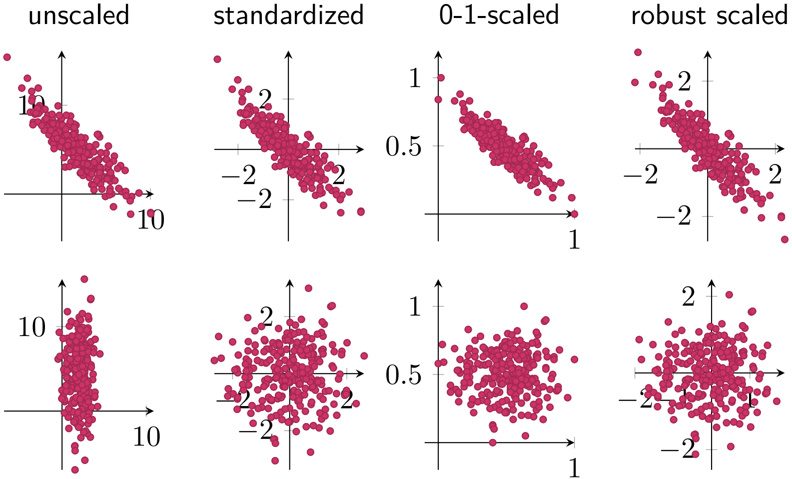
**STANDARDIZATION**: makes mostly sense for normal distributed data, a bit robust to outliers

$$\hat{D}_{ik} = \frac{D_{ik} - \mu_{\mathtt{F}_k}}{\sigma_{\mathtt{F}_k}}$$

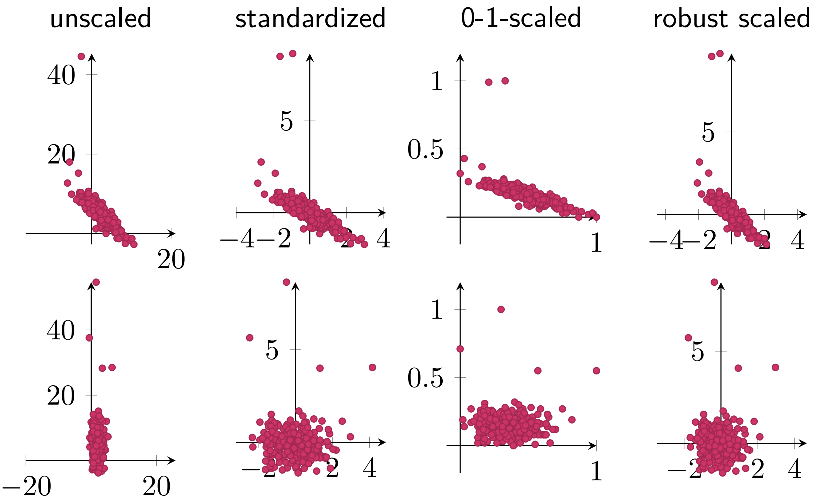
**ROBUST SCALING**: might make sense for ordinal data, and is robust against outliers

$$\hat{D}_{ik} = \frac{D_{ik} - \text{median}(D_{\cdot k})}{q_{75}(D_{\cdot k}) - q_{25}(D_{\cdot k})}$$

# Comparison of Scalers without Outliers



# Comparison of Scalers with Outliers



Scaling sounds fair, every feature is treated the same, e.g., every feature gets the same variance.

However, those differences in, e.g., the variance can also tell you about the usefulness of the features.

$\rightarrow$ **Feature Selection**

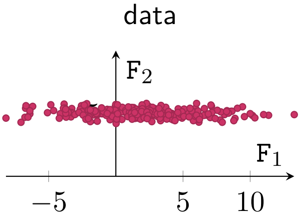
# Feature Selection

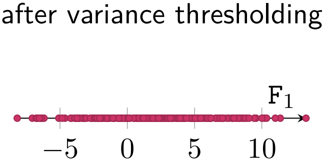
Feature selection removes features which are probably not needed. This is useful, because:

- every feature adds a dimensionality to your data points,
- in high dimensional data, data points tend to be equidistant to each other, so all observations seem to be alike $\rightarrow$**CURSE OF DIMENSIONALITY**,
- having fewer features improves the interpretability of results: in many tasks we want to know which features are important to get a good model.

# Feature Selection by Variance Thresholding

**IDEA**: features with very low variance will not have a big impact on the result and can be removed
$\rightarrow$ keep only features $\mathtt{F}_k$ with a variance $\sigma^2_{\mathtt{F}_k} > \tau$.

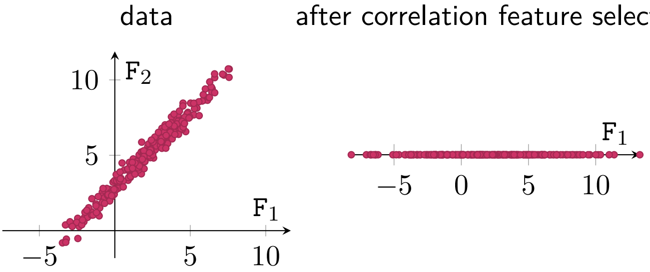



**PROBLEMATIC**: how do we scale the data to allow for fair variance comparisons? How to set $\tau$?
**OPTION**: scale by dividing each feature by its mean

# Feature Selection by Correlation

**IDEA**: correlation is a statistical measure of similarity of features and similar features are assumed to convey redundant information $\rightarrow$ remove $\mathtt{F}_k$ or $\mathtt{F}_l$ when $|\mathrm{corr}(D_{\cdot k}, D_{\cdot l})| > \tau$.



# Intermission: Covariance and Correlation

# Covariance as Inner Product Similarity

We compute the sample covariance of two feature vectors by

$$\mathrm{cov}(D_{\cdot k}, D_{\cdot l}) = \frac{1}{n}\sum_{i=1}^{n}(D_{ik} - \mu_{\mathrm{F}_k})(D_{il} - \mu_{\mathrm{F}_l}),$$

If we consider the centered data matrix $C = D - \mathbf{1}\mu_{\mathrm{F}}$, then the covariance can be written as an inner product:

$$\mathrm{cov}(D_{\cdot k}, D_{\cdot l}) = \frac{1}{n}\sum_{i=1}^{n} C_{ik}C_{il} = \frac{1}{n}C_{\cdot k}^{\top}C_{\cdot l}$$

That is, the **COVARIANCE** can be seen as **THE INNER PRODUCT SIMILARITY** of the centered feature values.

# Covariance and Variance

The **VARIANCE** is the covariance of a feature with itself and thus equal to the squared norm:

$$\sigma^2_{\mathrm{F}_k} = \operatorname{cov}(D_{\cdot k}, D_{\cdot k}) = \frac{1}{n} C_{\cdot k}^\top C_{\cdot k} = \frac{1}{n} \| C_{\cdot k} \|^2$$

The covariance between two features is large when the angle between the feature vectors is small and the variance of the features is high

$$\begin{aligned} \operatorname{cov}(D_{\cdot k}, D_{\cdot l}) = \frac{1}{n} C_{\cdot k}^\top C_{\cdot l} &= \frac{1}{n} \cos \sphericalangle (C_{\cdot k}, C_{\cdot l}) \| C_{\cdot k} \| \| C_{\cdot l} \| \\ &= \cos \sphericalangle (C_{\cdot k}, C_{\cdot l}) \sigma_{\mathrm{F}_k} \sigma_{\mathrm{F}_l} \end{aligned}$$

# The Covariance Matrix

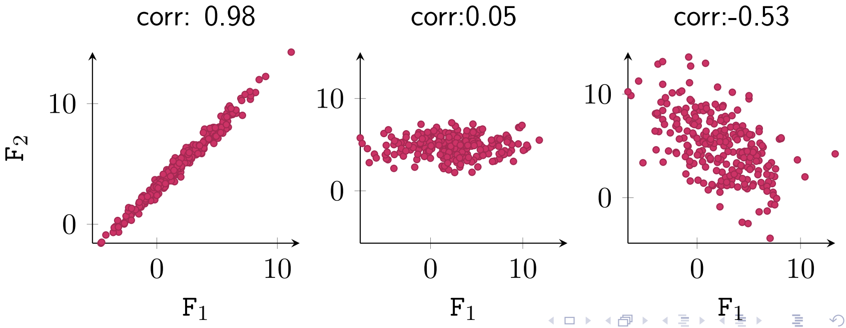
With $C$ being the centered data matrix, the covariance matrix is given by

$$C^\top C = \begin{pmatrix} \sigma^2_{\mathtt{F}_1} & \operatorname{cov}(D_{\cdot 1}, D_{\cdot 2}) & \dots & \operatorname{cov}(D_{\cdot 1}, D_{\cdot d}) \\ \operatorname{cov}(D_{\cdot 2}, D_{\cdot 1}) & \sigma^2_{\mathtt{F}_1} & \dots & \operatorname{cov}(D_{\cdot 2}, D_{\cdot d}) \\ \vdots & \vdots & \vdots & \vdots \\ \operatorname{cov}(D_{\cdot d}, D_{\cdot 1}) & \operatorname{cov}(D_{\cdot d}, D_{\cdot 2}) & \dots & \sigma^2_{\mathtt{F}_d} \end{pmatrix}$$

# Pearson's Correlation Coefficient

$$\mathrm{corr}(D_{\cdot k}, D_{\cdot l}) = \frac{\mathrm{cov}(D_{\cdot k}, D_{\cdot l})}{\sigma_{\mathtt{F}_k}\sigma_{\mathtt{F}_l}} = \cos\sphericalangle(C_{\cdot k}, C_{\cdot l}) \in [-1, 1]$$

The correlation between two features is the normalized version of the covariance and it's large when the angle between the centered feature vectors is small.



# End of Intermission

# Transformations of the Feature Space

In some cases, it makes sense to transform the features or add transformed features, e.g., transforming the data can transfer nonlinear relations into linear ones.

| ID | $F_1$ | $F_2$ |
|---|---|---|
| 1 | 1.1 | 3.2 |
| 2 | 2.0 | 2.7 |
| ⋮ | ⋮ | ⋮ |

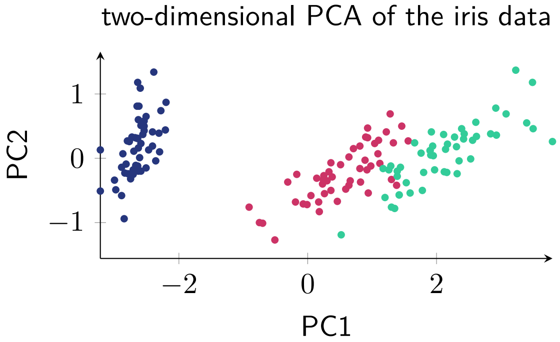
$\rightarrow$

| ID | $F_1^2$ | $F_1F_2$ | $F_2^2$ | $\log(F_2)$ |
|---|---|---|---|---|
| 1 | 1.21 | 3.52 | 10.24 | 1.16 |
| 2 | 4.00 | 5.40 | 7.29 | 1.69 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

The task of dimensionality reduction is to find a **TRANSFORMATION INTO A LOW-DIMENSIONAL FEATURE SPACE** which keeps characteristics of the data

# Principal Component Analysis (PCA)

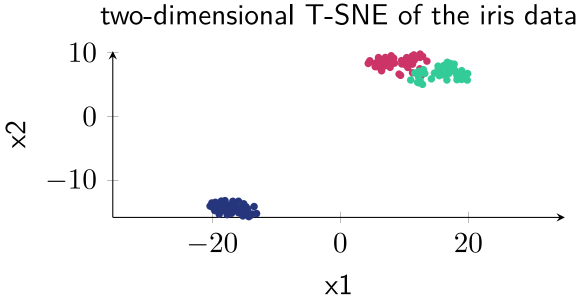
The goal of PCA is to **MAINTAIN MOST OF THE VARIANCE** in the low-dimensional view.




We will discuss PCA and how it works later in the course.

# t-Distributed Stochastic Neighbor Embedding (t-SNE)

The goal of t-SNE is to **MAINTAIN LOCAL SIMILARITIES** in the low-dimensional view




We will not discuss how t-SNE works in detail in the course. If you want to get a good idea about it, have a look here: `https://distill.pub/2016/misread-tsne/`