Data Preparation

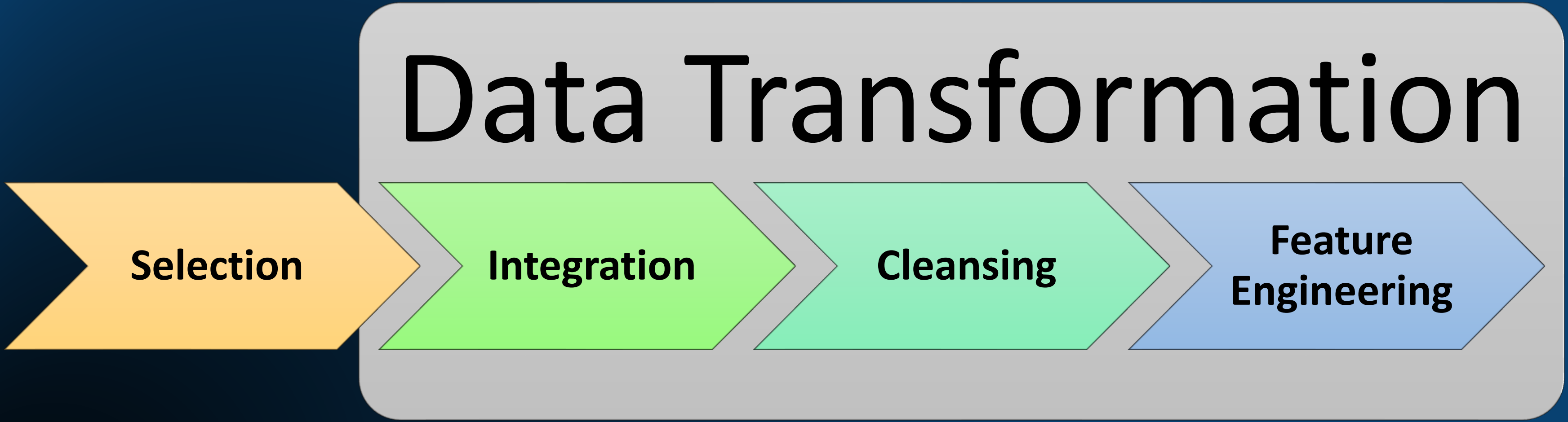
Cláudia Antunes

Instituto Superior Técnico – Universidade de Lisboa

Goals

- To improve the quality of data
- To adjust data to feed learning algorithms

DATA PREPARATION



Data Transformation

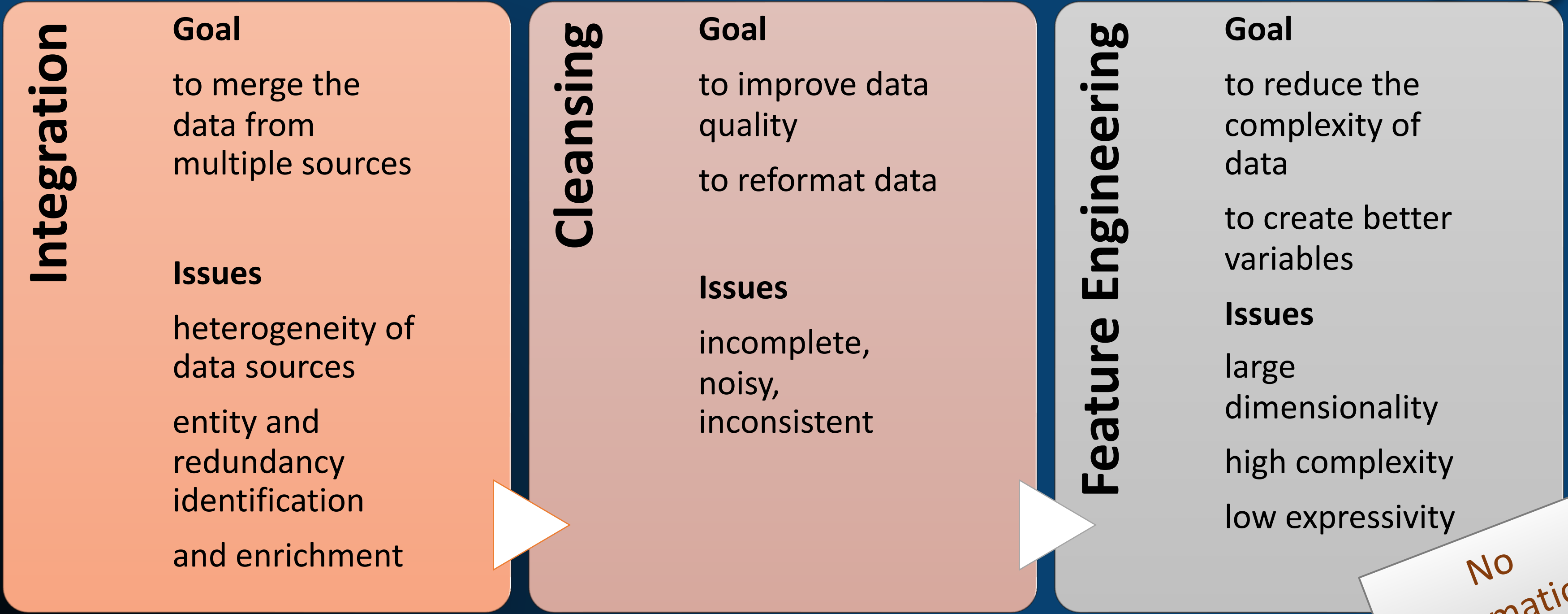
Selection

Integration

Cleansing

Feature Engineering

Data Preparation



Integration

Goal

to merge the data from multiple sources

Issues

heterogeneity of data sources

entity and redundancy identification

and enrichment

Cleansing

Goal

to improve data quality

to reformat data

Issues

incomplete, noisy, inconsistent

Feature Engineering

Goal

to reduce the complexity of data

to create better variables

Issues

large dimensionality

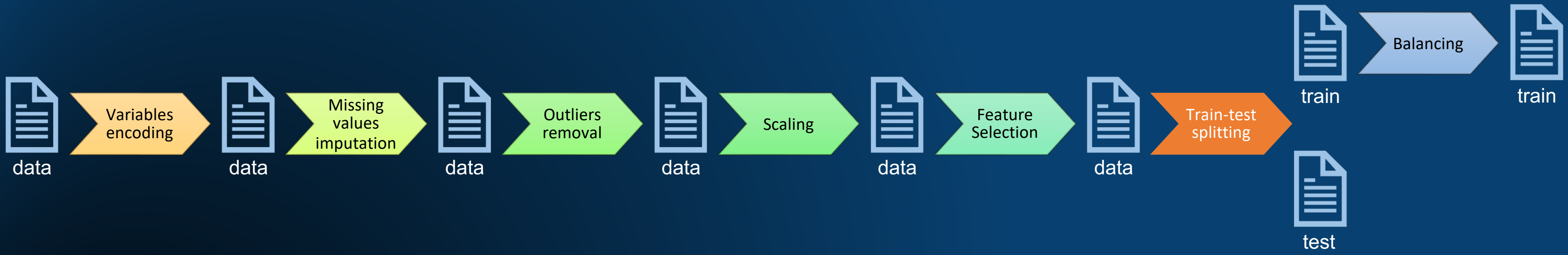
high complexity

low expressivity

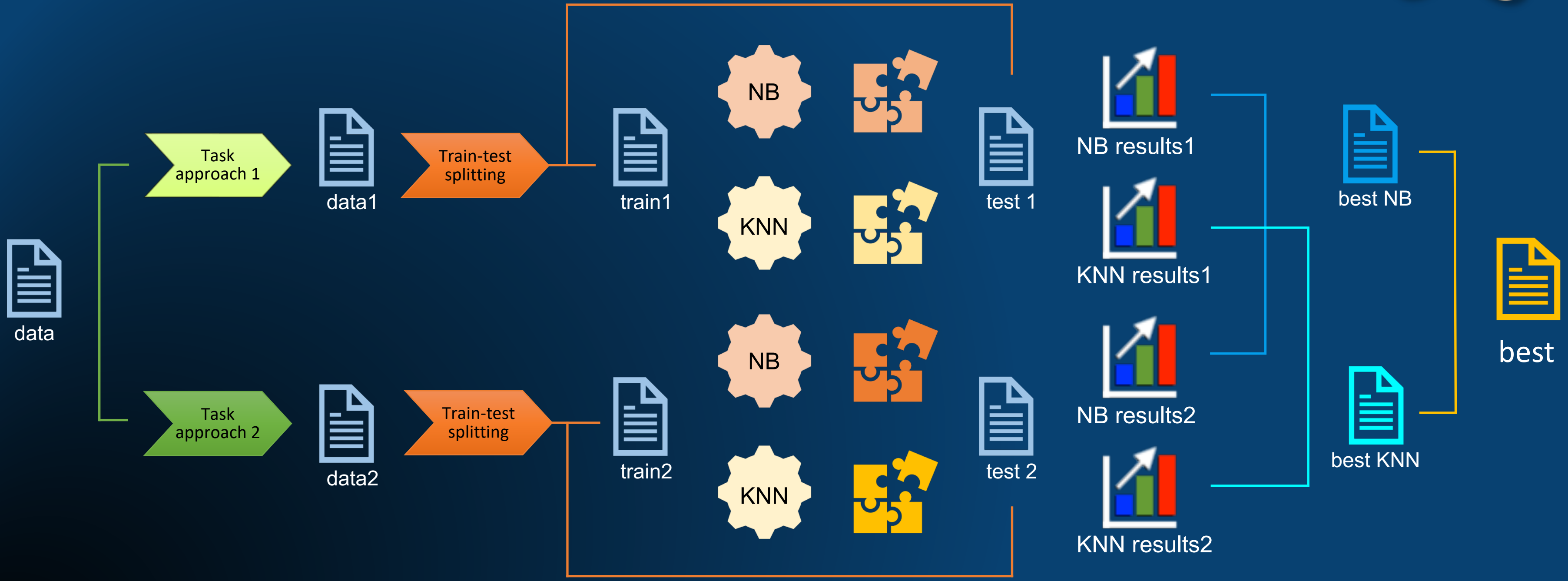
No information loss

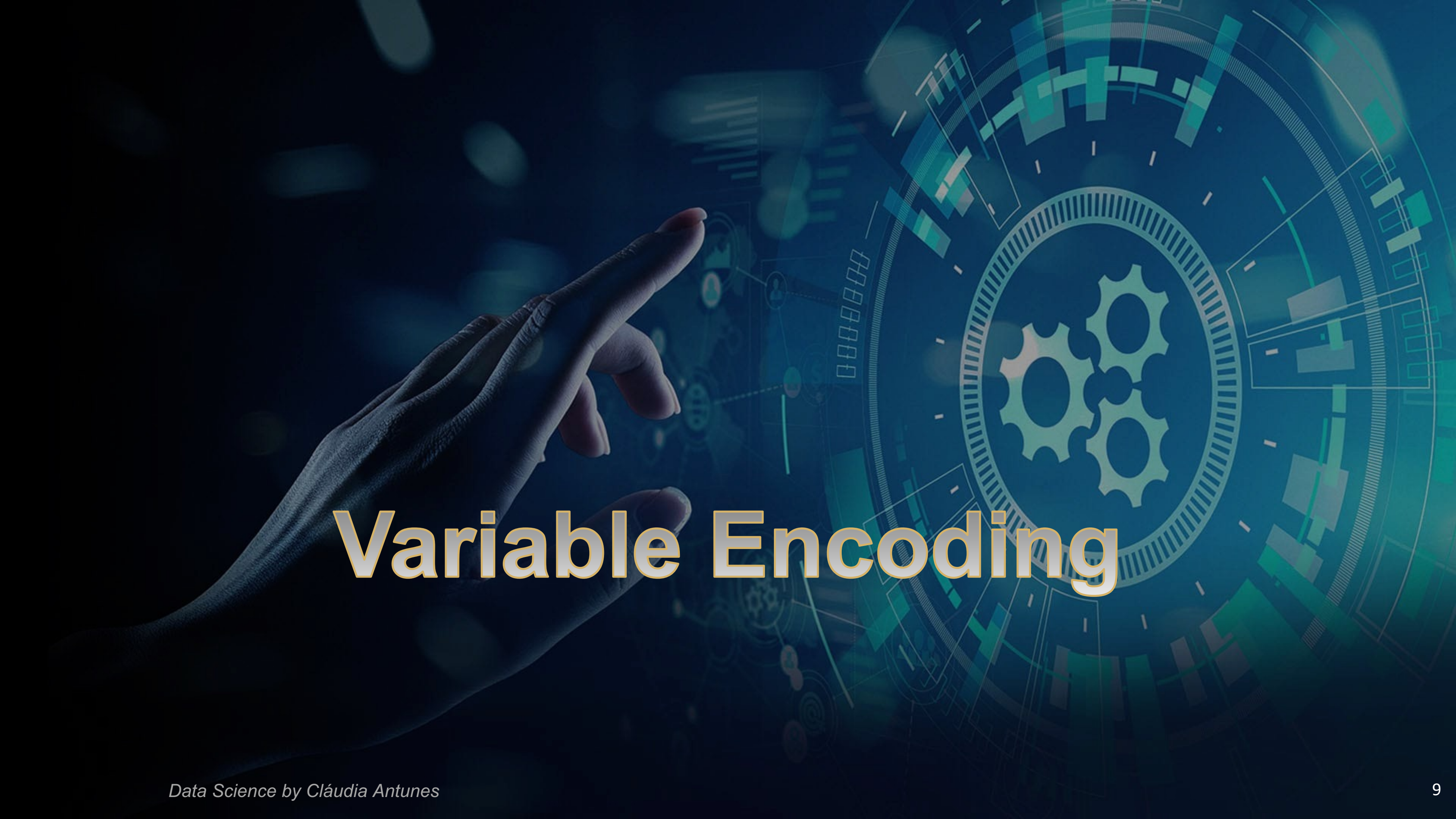
Practical Methodology

Methodology



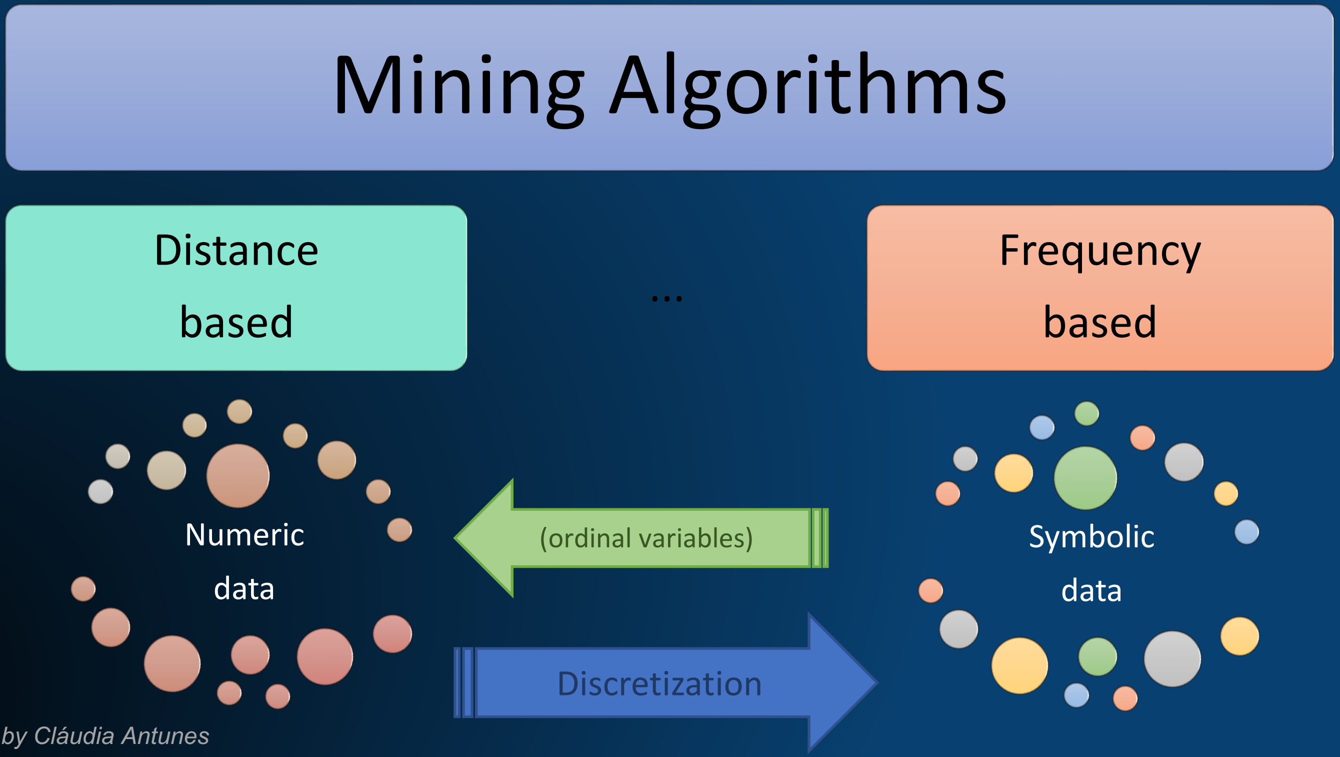
Application of One Preparation Task





Variable Encoding

Variable Encoding



Mining Algorithms

Distance based

…

Frequency based

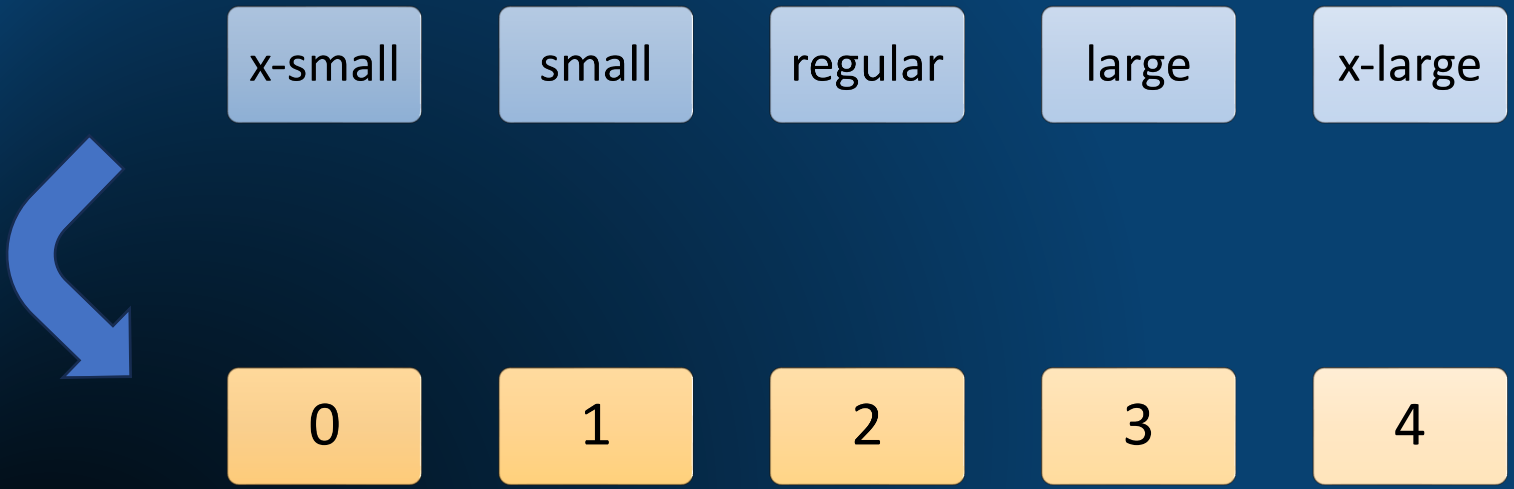
Numeric data

(ordinal variables)

Discretization

Symbolic data

Sequential Values



x-small	small	regular	large	x-large
0	1	2	3	4

Cyclic Variables

Variables having a cyclic nature

- Seasons
- Months
- Weekdays
- Cardinal points

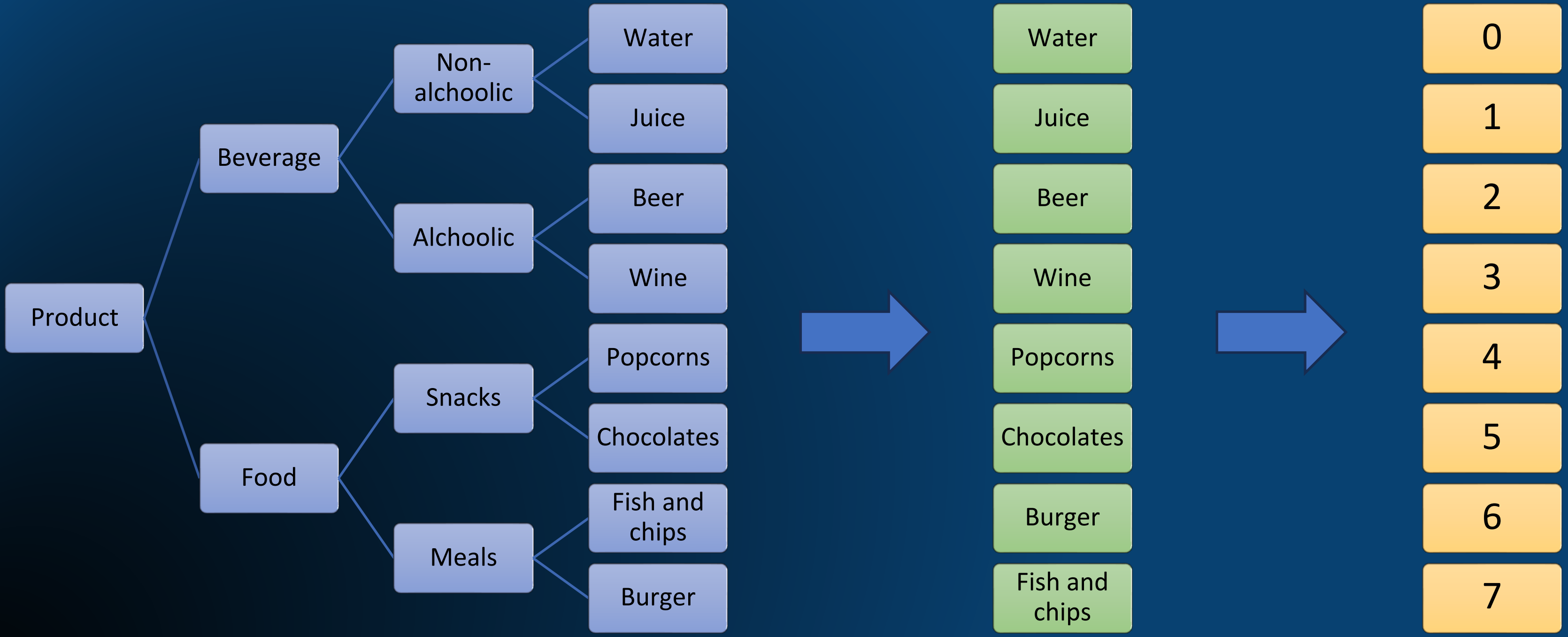
If $x \in [0 : x_{max}]$

→

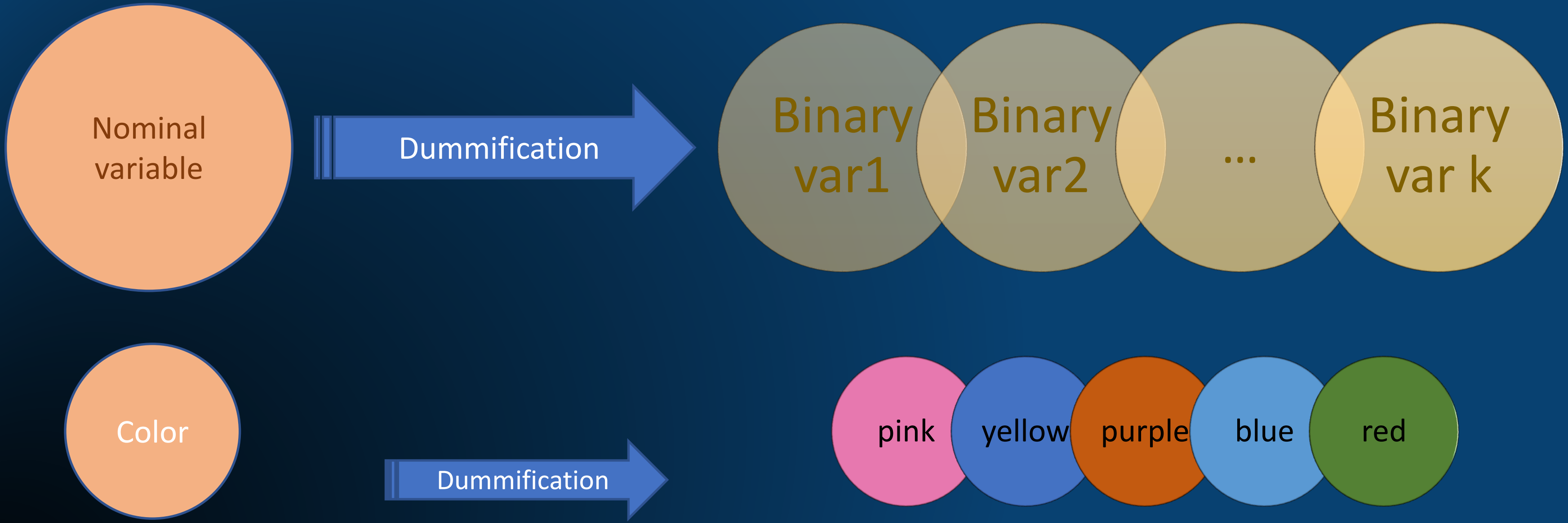
$$x_{sin} = \sin \frac{2\pi x}{x_{max}}$$

$$x_{cos} = \cos \frac{2\pi x}{x_{max}}$$

Hierarchical Values



No order → Dummification



Nominal variable

Dummification

Binary var1

Binary var2

…

Binary var k

Color

Dummification

pink

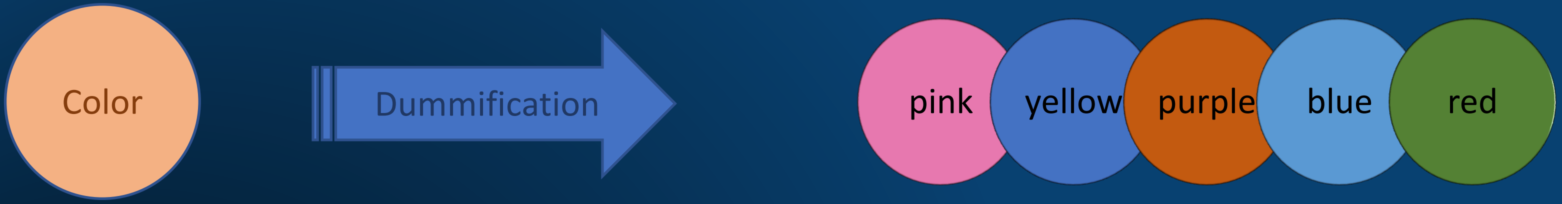
yellow

purple

blue

red

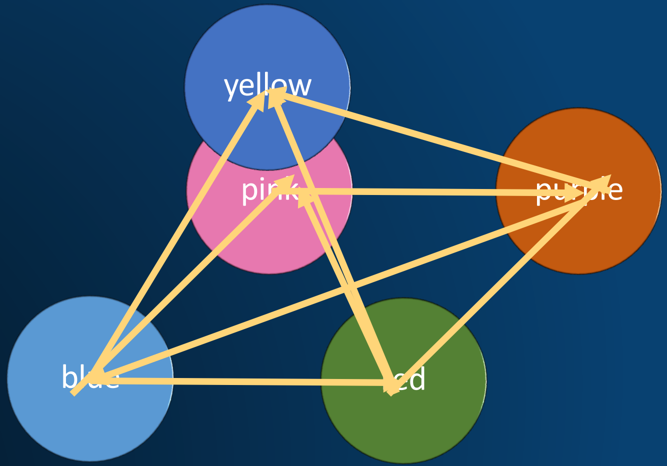
Dummification



Color
pink
red
blue
yellow
purple
blue
pink

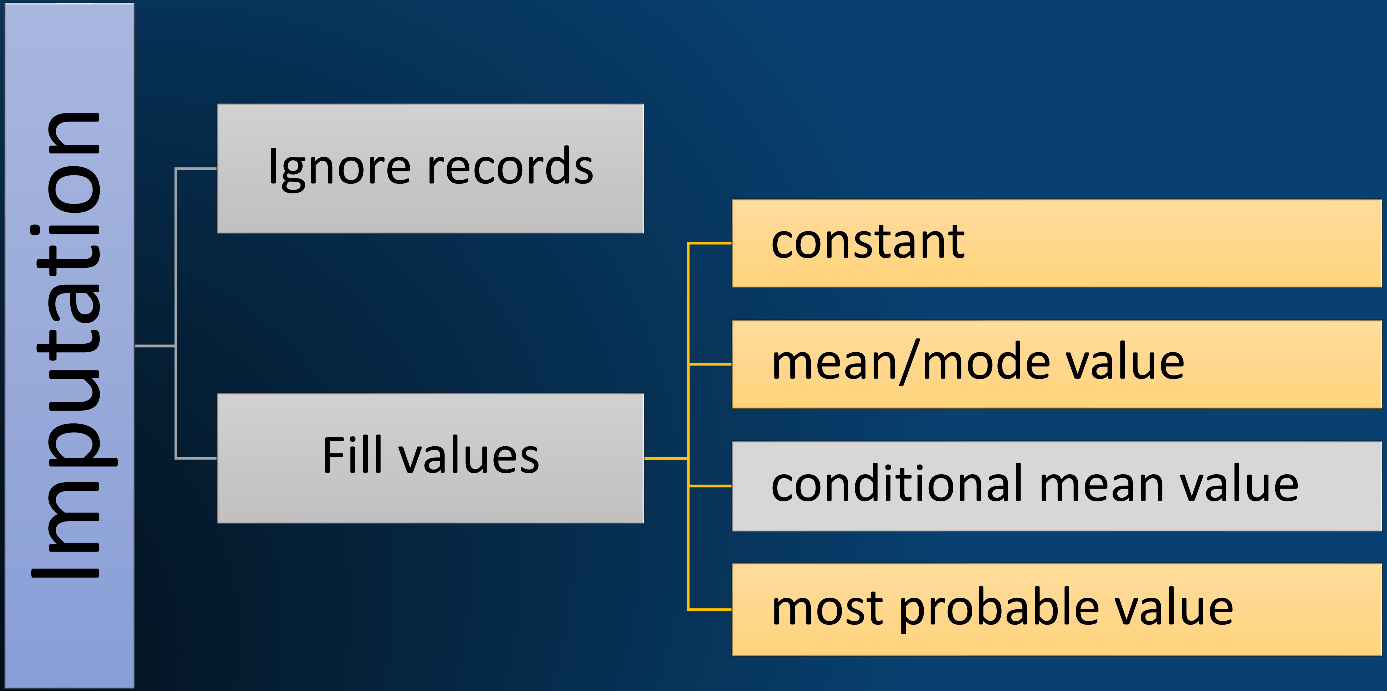
pink	yellow	purple	blue	red
1	0	0	0	0
0	0	0	0	1
0	0	0	1	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
1	0	0	0	0

Dummification



Missing Values and Outliers

Missing Values Imputation



Imputation

Ignore records

Fill values

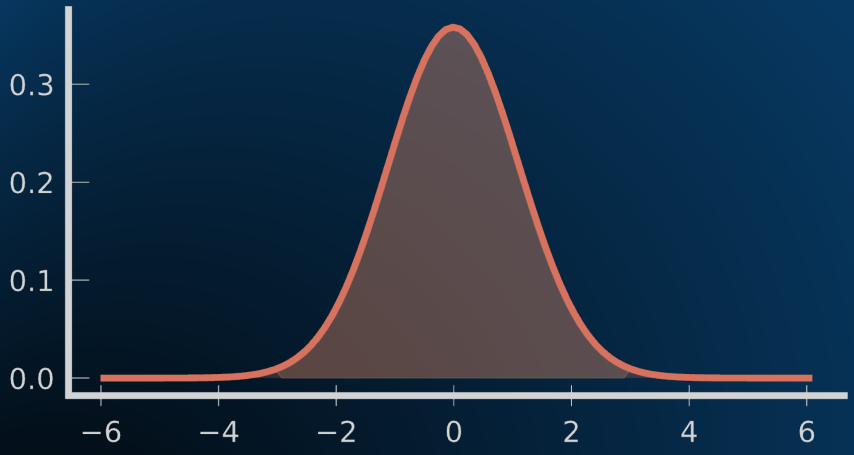
constant

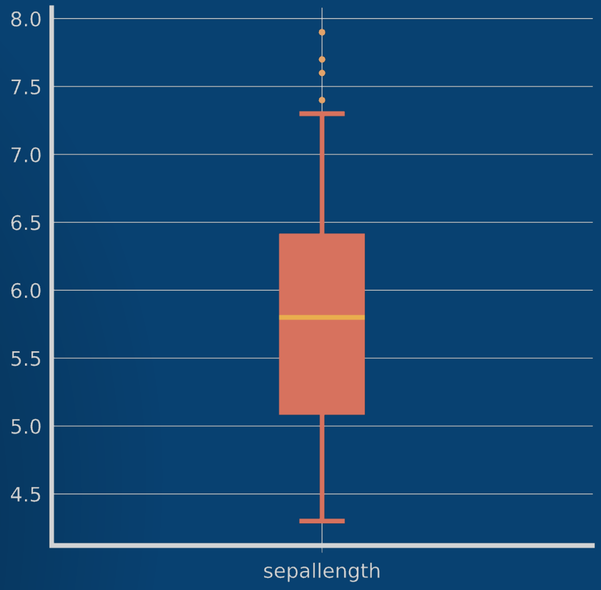
mean/mode value

conditional mean value

most probable value

Outliers Identification





X < μ−*n*σ

or

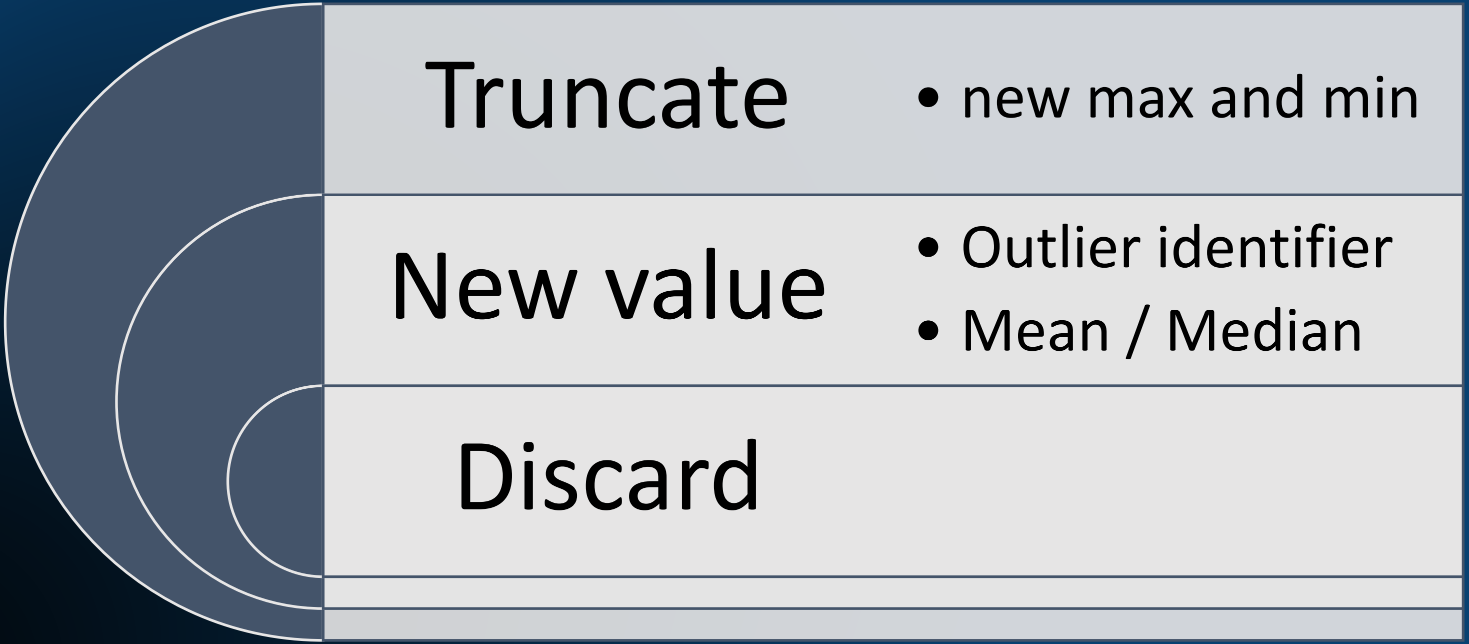
X > μ+nσ

X < Q1 - 1.5 x IQR

or

X > Q3 + 1.5 x IQR

Outliers Imputation Strategies



Truncate

- new max and min

New value

- Outlier identifier
- Mean / Median

Discard

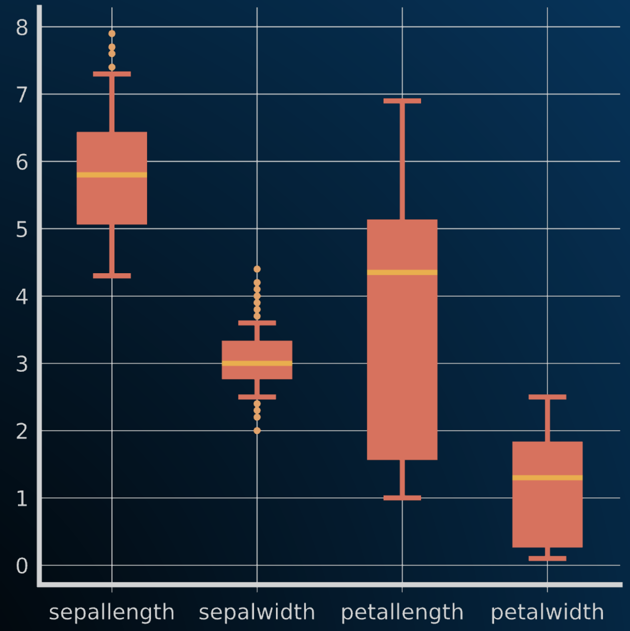
Scale Transformation

Normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_\max - new\_\min) + new\_min$$

values ∈ [new_max, new_min]

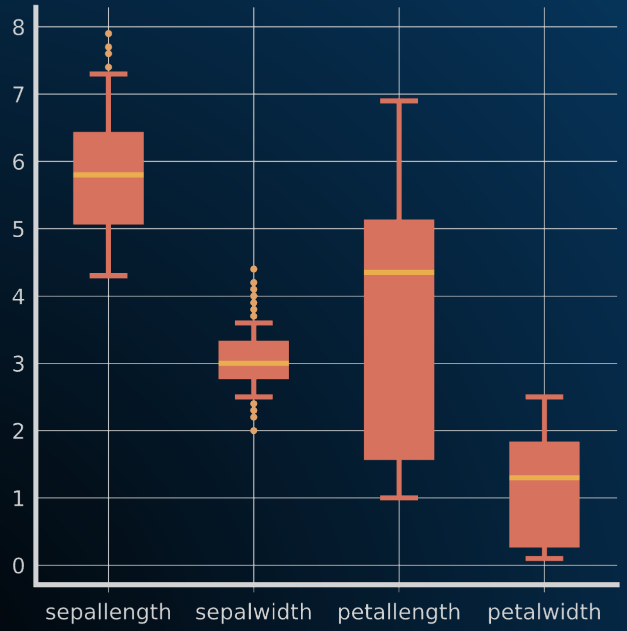
Normalization

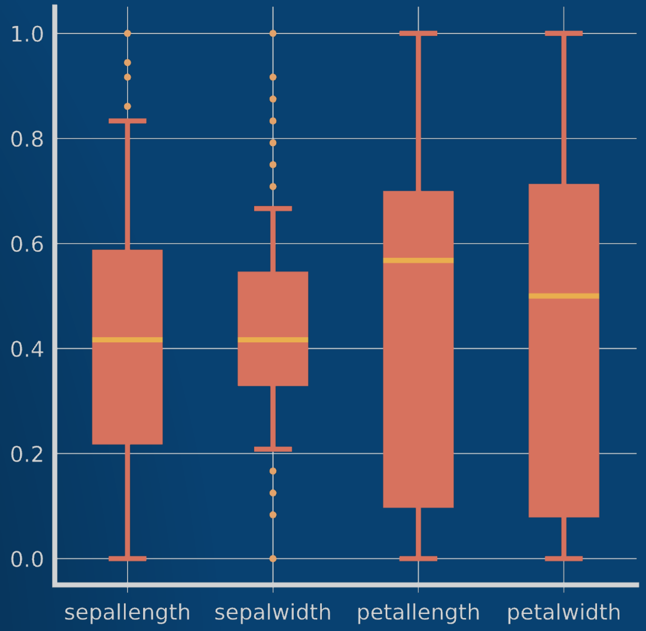


$$v' = \frac{5.9 - 4.3}{7.9 - 4.3}(1 - 0) + 0 = \frac{1.6}{3.6} = 0.44$$

	Sepal Length	Sepal Width	Petal Length	Petal Width
count	150	150	150	150
mean	5.84	3.05	3.76	1.20
std	0.83	0.43	1.76	0.76
min	**4.3**	**2.0**	**1.0**	**0.1**
Q1	5.1	2.8	1.6	0.3
median	5.8	3.0	4.4	1.3
Q3	6.4	3.3	5.1	1.8
max	**7.9**	**4.4**	**6.9**	**2.5**

Normalization





Standardization

$$z = \frac{x - \mu}{\sigma}$$

Z-score

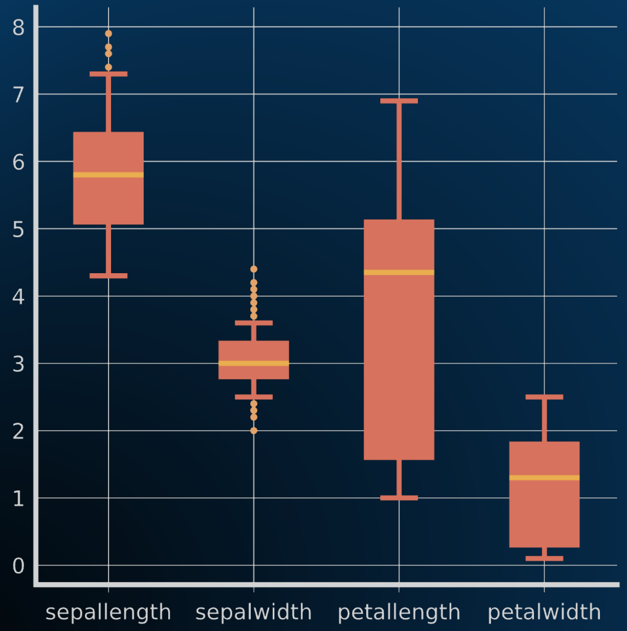
negative

if z < mean

positive

if z > mean

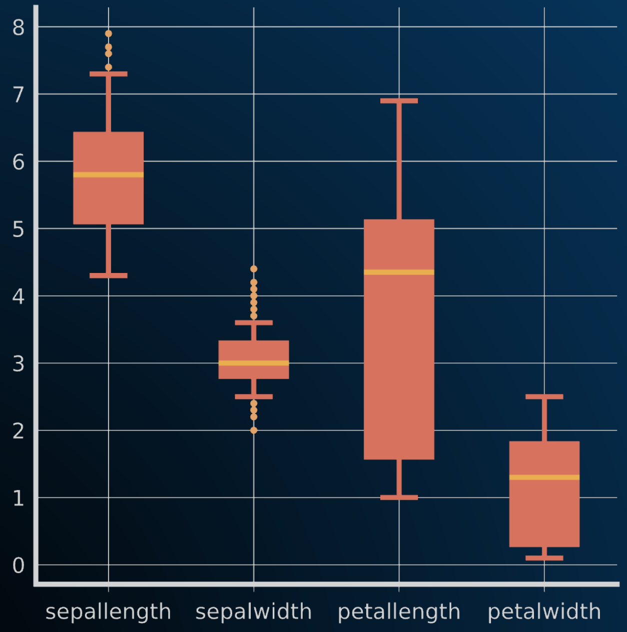
Standardization

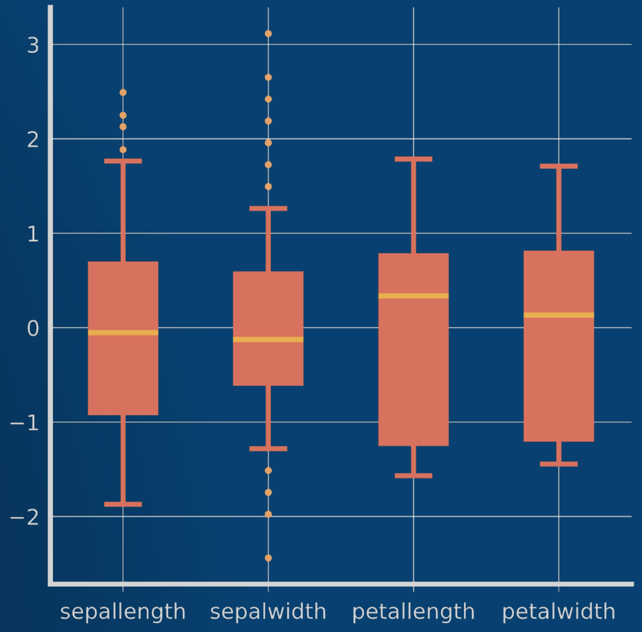


$$z = \frac{5.9 - 5.84}{0.83} = \frac{0.06}{0.83} = 0.07$$

	Sepal Length	Sepal Width	Petal Length	Petal Width
count	150	150	150	150
mean	**5.84**	**3.05**	**3.76**	**1.20**
std	0.83	0.43	1.76	0.76
min	4.3	2.0	1.0	0.1
Q1	5.1	2.8	1.6	0.3
median	5.8	3.0	4.4	1.3
Q3	6.4	3.3	5.1	1.8
max	7.9	4.4	6.9	2.5

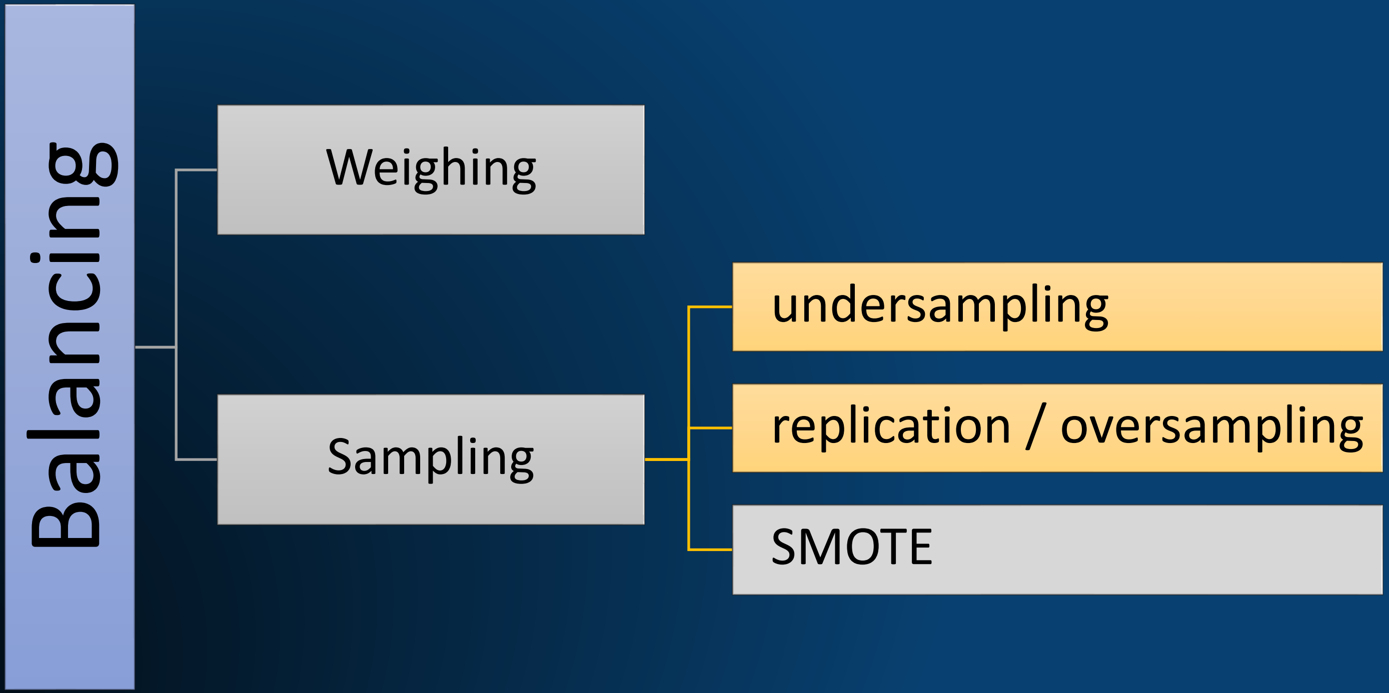
Standardization



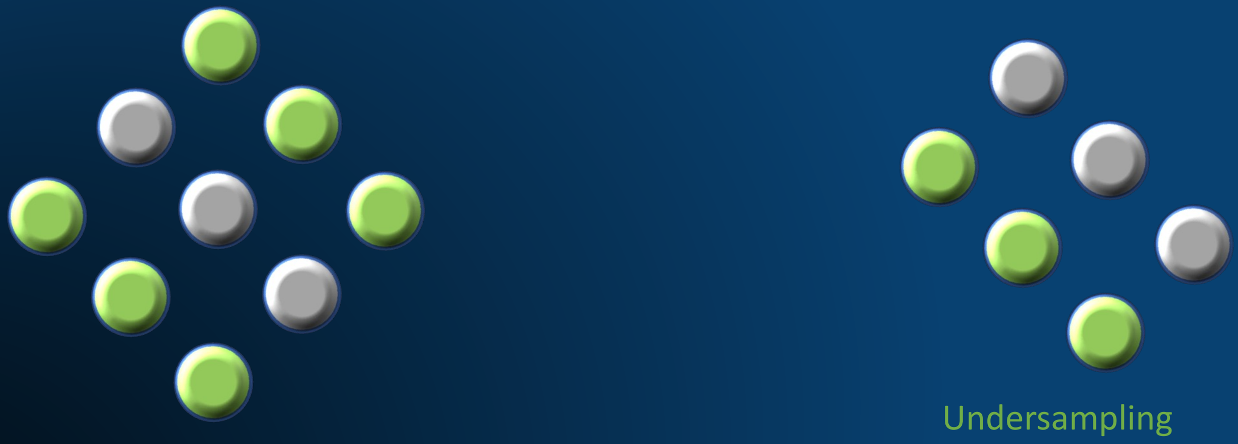


Data Balancing

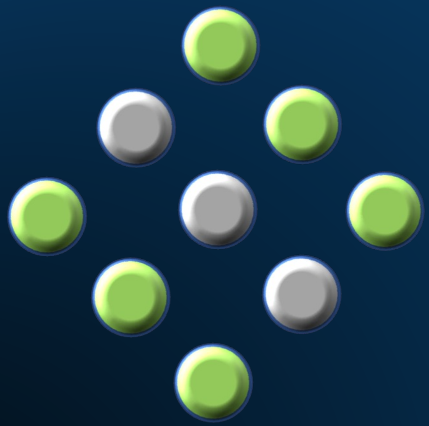
Data Balancing

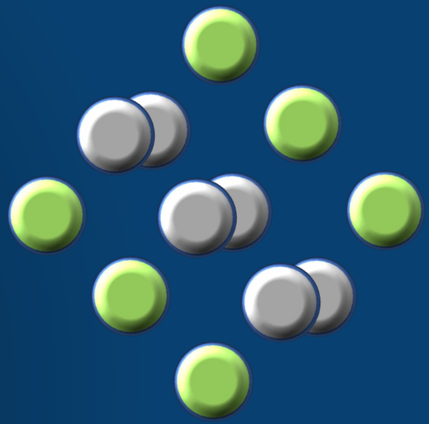


Undersampling



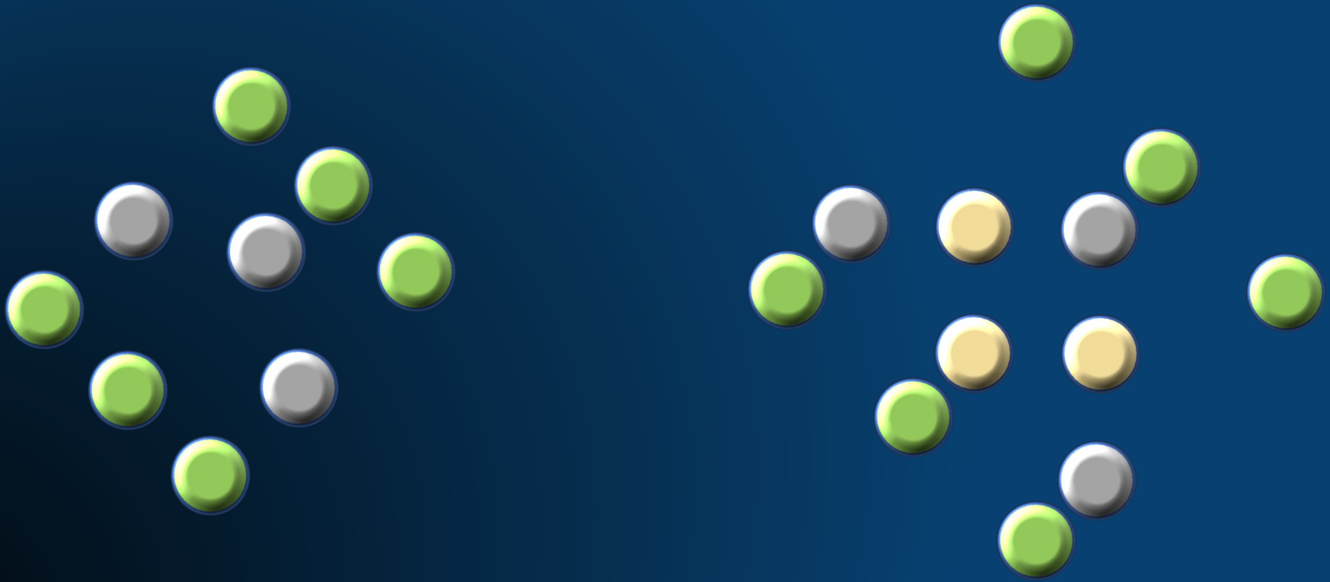
Replication / Oversampling





Oversampling

SMOTE



Thank you!

claudia.antunes@tecnico.ulisboa.pt