

A hand is shown in the foreground, pointing towards a complex digital interface. The interface features a large circular gauge with multiple concentric rings, some of which are highlighted in green. In the center of the gauge are three interlocking gears. The background is dark blue with various data visualizations, including bar charts and line graphs, some of which are also highlighted in green. The overall aesthetic is futuristic and technological.

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 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

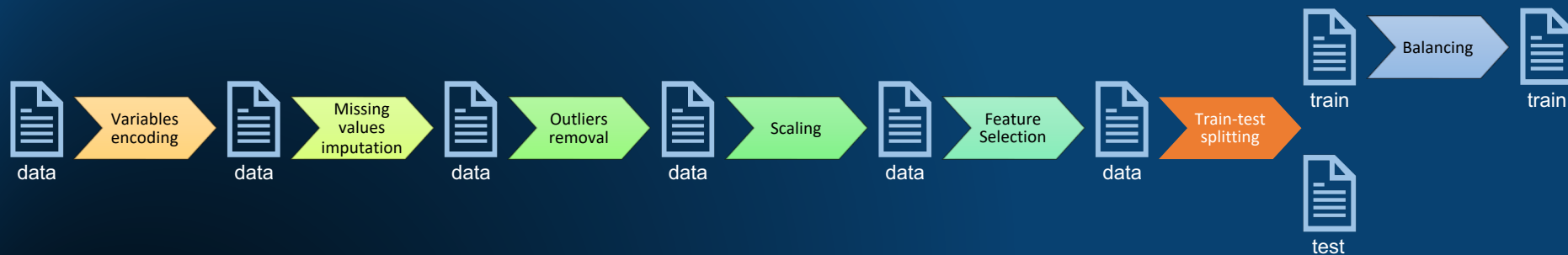
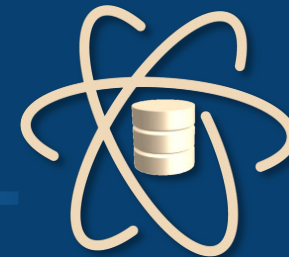
large dimensionality
high complexity
low expressivity

No
information
loss

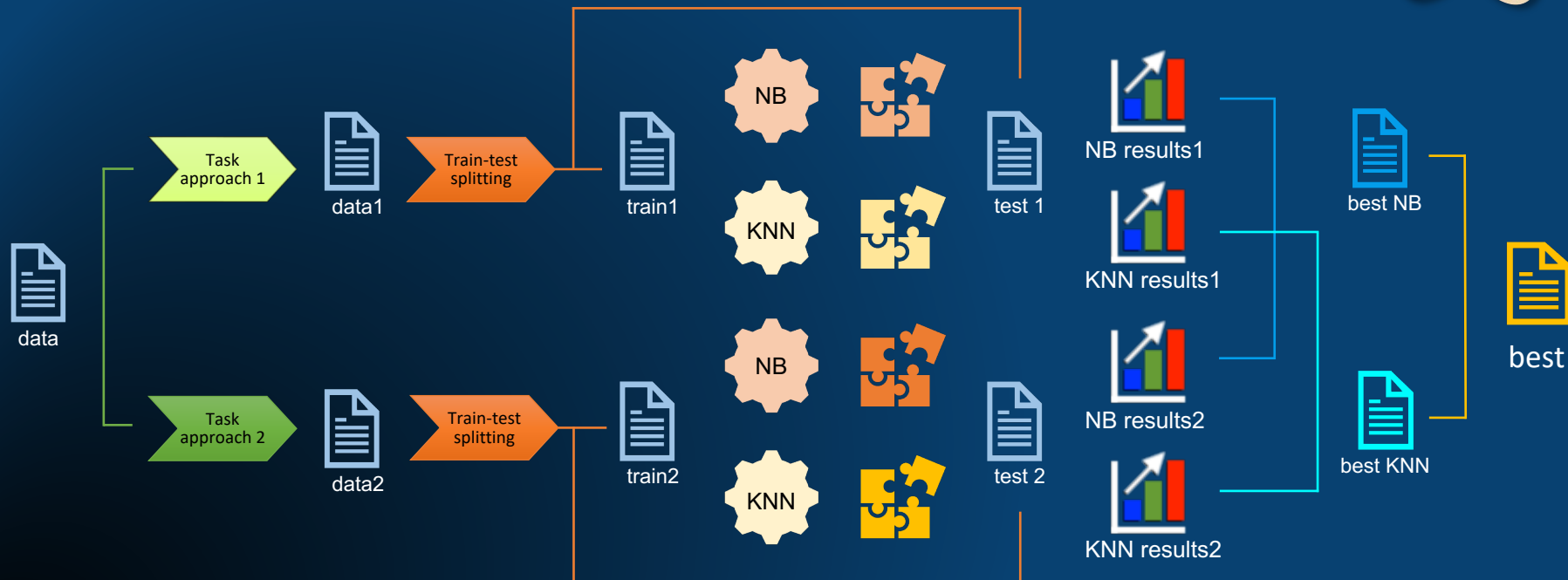
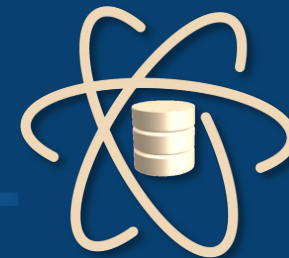
A hand is shown in the foreground, reaching towards a complex, futuristic digital interface. The interface features a large circular gauge with multiple concentric rings, some of which are highlighted in green. In the center of the gauge are three interlocking gears. The background is dark blue with various abstract digital elements, including lines, dots, and faint data visualizations. The overall aesthetic is high-tech and modern.

Practical Methodology

METHODOLOGY



APPLICATION OF ONE PREPARATION TASK





A hand is shown in the foreground, reaching out towards a complex, futuristic digital interface. The interface features a large circular gauge with multiple concentric rings, some of which are highlighted in green. In the center of the gauge are three interlocking gears. The background is dark blue with various data visualizations, including bar charts and line graphs, some of which are also highlighted in green. The overall aesthetic is high-tech and data-driven.

Variable Encoding

VARIABLE ENCODING



Mining Algorithms

Distance
based

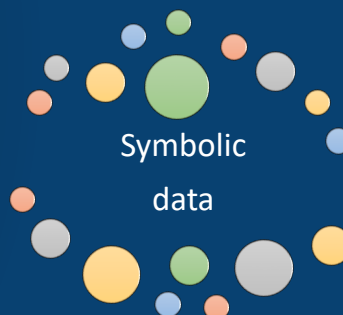
...

Frequency
based



Numeric
data

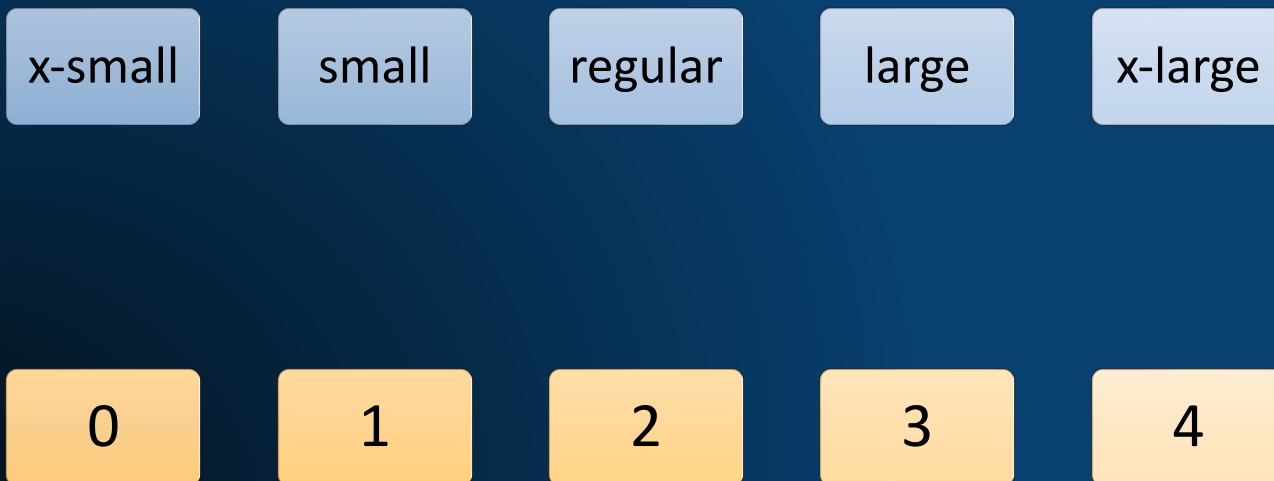
(ordinal variables)



Symbolic
data

Discretization

SEQUENTIAL VALUES



CYCLIC VARIABLES



Variables having a cyclic nature

- Seasons
- Months
- Weekdays
- Cardinal points

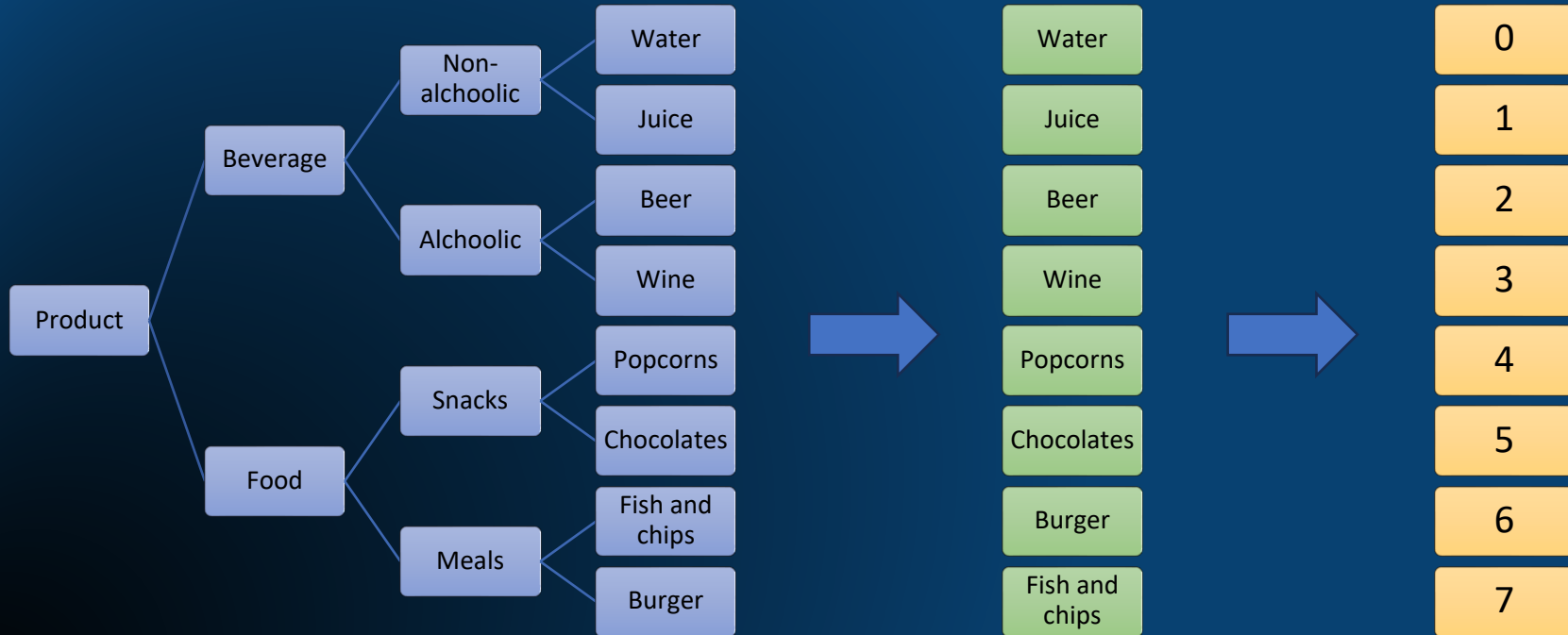
If $x \in [0 : x_{max}]$

→

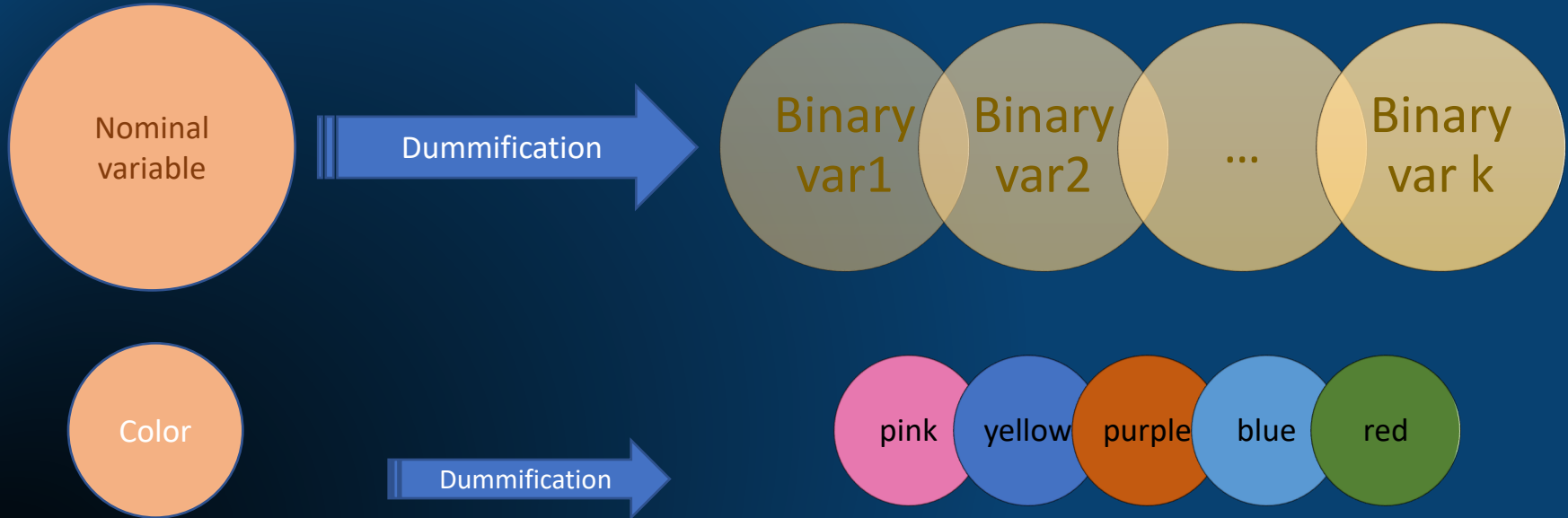
$$x_{sin} = \sin \frac{2\pi x}{x_{max}}$$

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

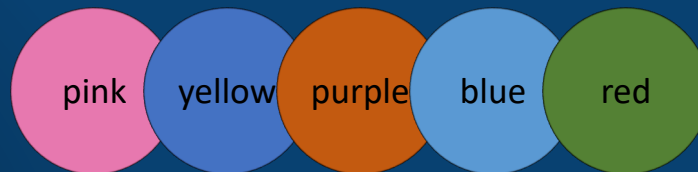
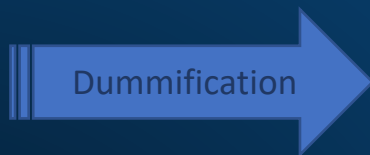
HIERARCHICAL VALUES



No ORDER → DUMMIFICATION



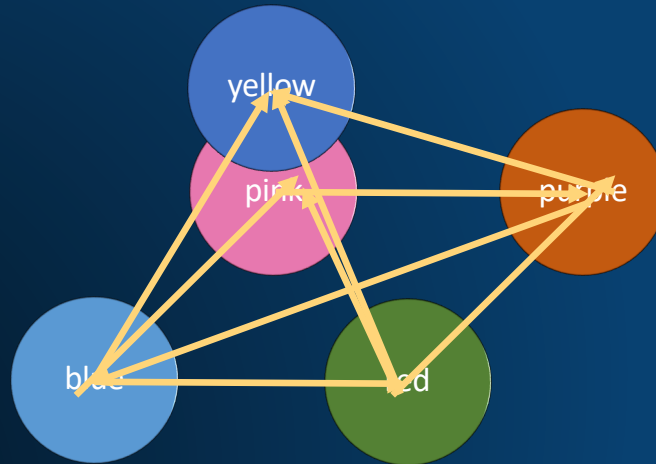
DUMMIFICATION



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

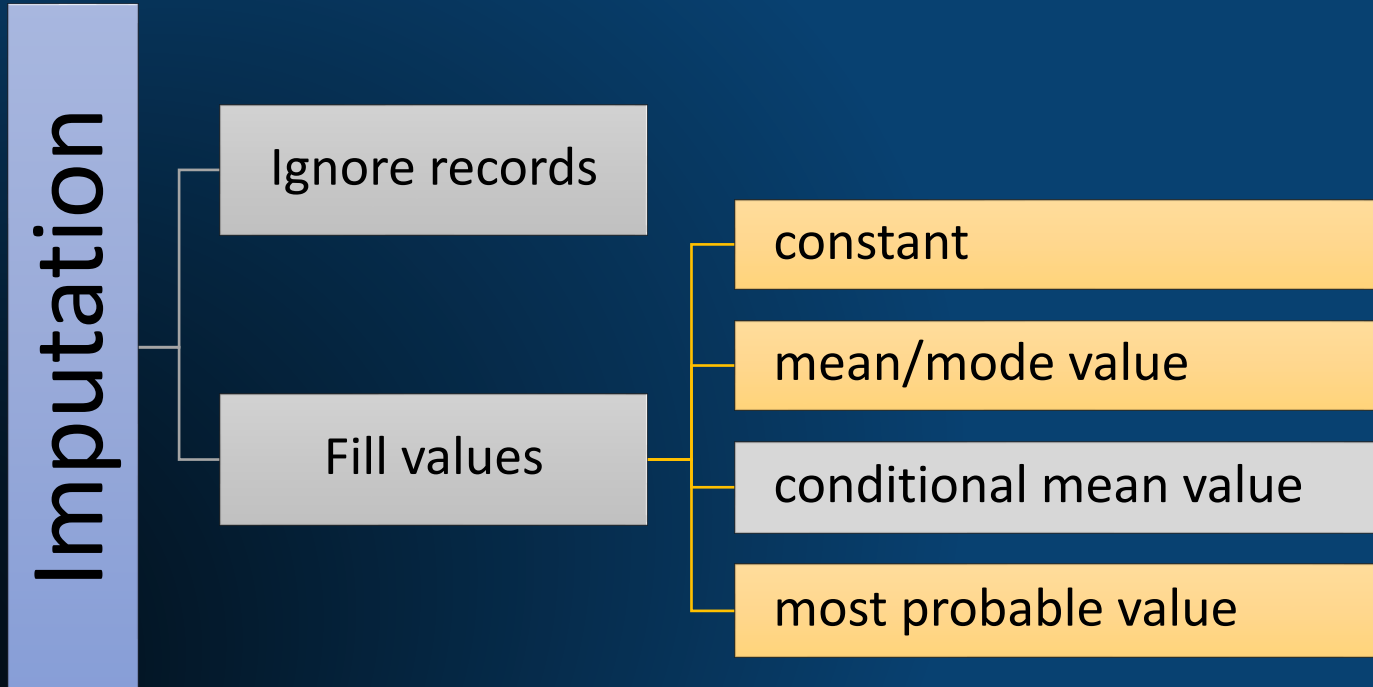




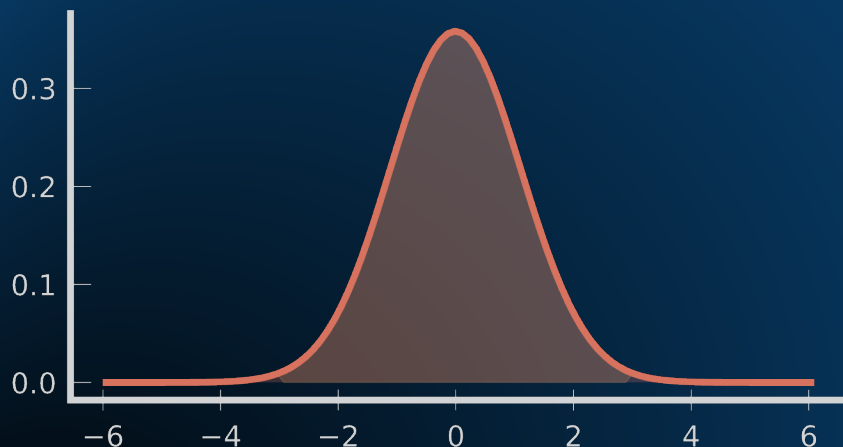
A hand is shown in the foreground, reaching out towards a complex, futuristic digital interface. The interface features a large circular gauge with multiple concentric rings, some of which are highlighted in green. Inside the gauge, there are several interlocking gears. The background is dark blue with various data visualizations, including bar charts and line graphs, some of which are also highlighted in green. The overall aesthetic is high-tech and data-driven.

Missing Values and Outliers

MISSING VALUES IMPUTATION



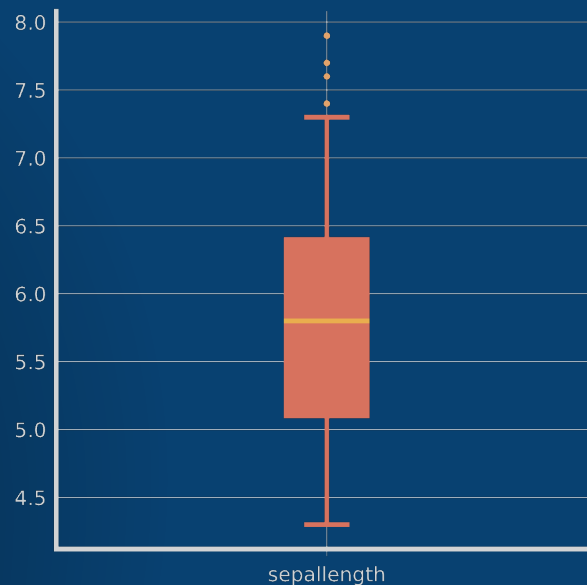
OUTLIERS IDENTIFICATION



$$X < \mu - n\sigma$$

or

$$X > \mu + n\sigma$$



$$X < Q1 - 1.5 \times \text{IQR}$$

or

$$X > Q3 + 1.5 \times \text{IQR}$$

OUTLIERS IMPUTATION STRATEGIES



	Truncate	<ul style="list-style-type: none">• new max and min
	New value	<ul style="list-style-type: none">• Outlier identifier• Mean / Median
	Discard	



The background of the image is a dark blue, almost black, space. On the left, a human hand is visible, with the index finger pointing towards the right. The hand is rendered in a realistic style with some lighting on the skin. In the center and right, there is a complex, futuristic digital interface. It features several concentric circles, some of which are filled with teal and green data-like patterns. There are also gear-like shapes, some of which are white and some are teal. The overall aesthetic is high-tech and digital. The text 'Scale Transformation' is overlaid on the image, centered horizontally and slightly below the vertical center. It is written in a bold, sans-serif font. The word 'Scale' is on the top line and 'Transformation' is on the bottom line. The text is white with a yellow-orange glow or outline, making it stand out against the dark background.

Scale Transformation

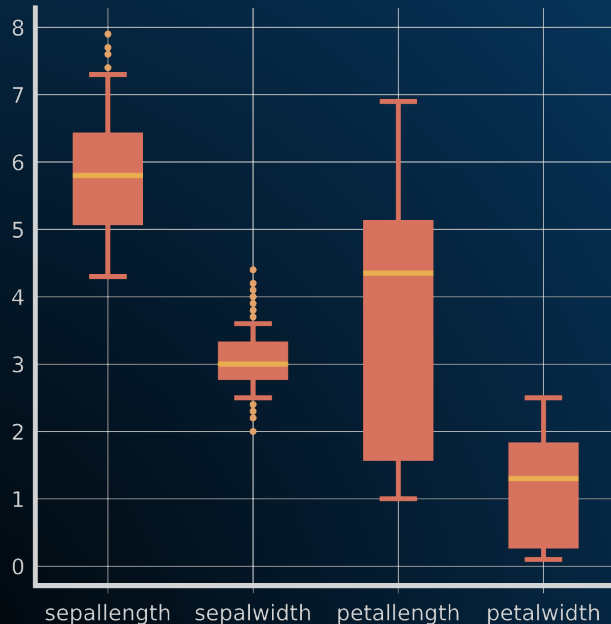
NORMALIZATION



$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max} - \text{new_min}) + \text{new_min}$$

values $\in [\text{new_max}, \text{new_min}]$

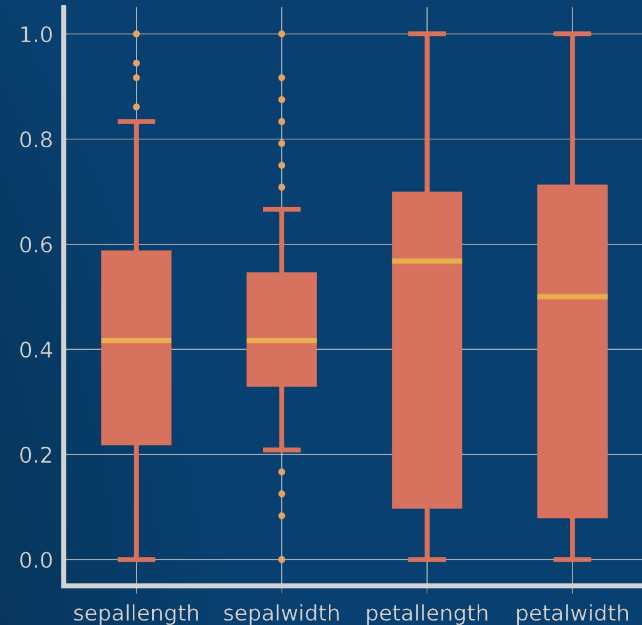
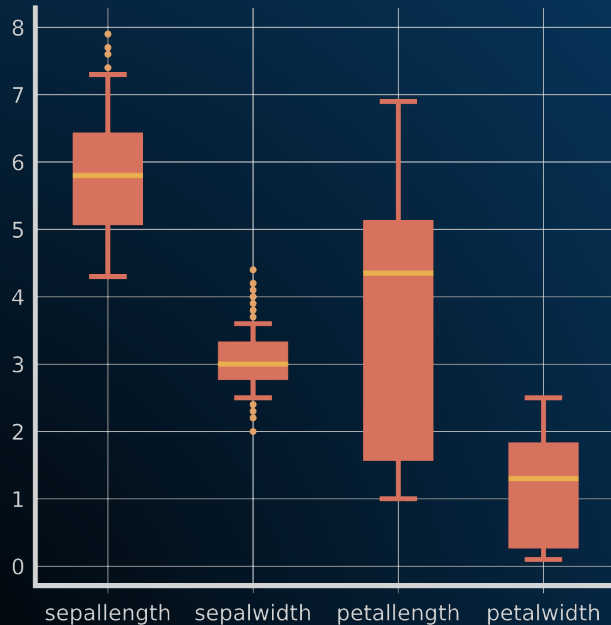
Normalization



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

	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



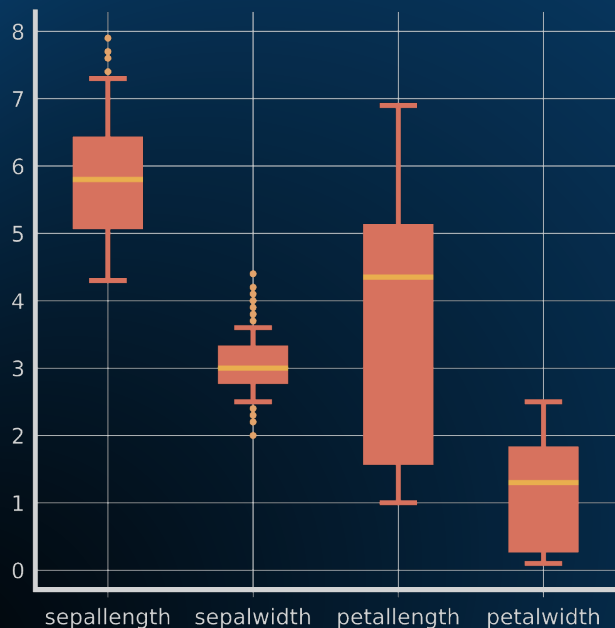
STANDARDIZATION



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

Z-score
negative
if $z < \text{mean}$
positive
if $z > \text{mean}$

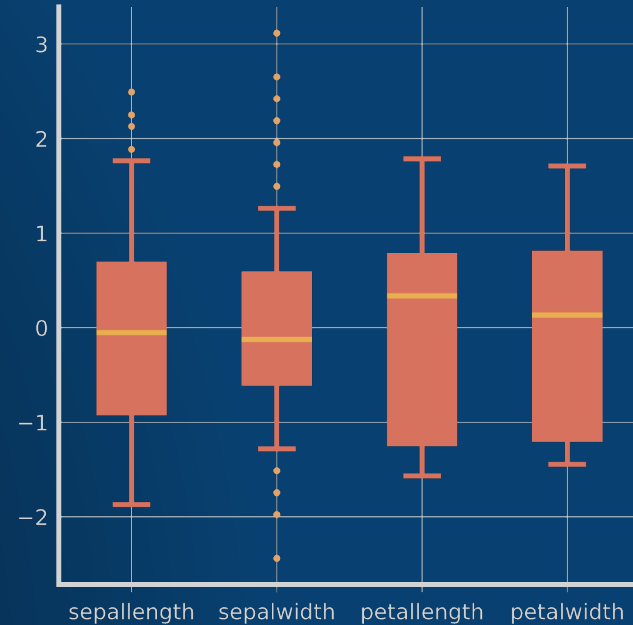
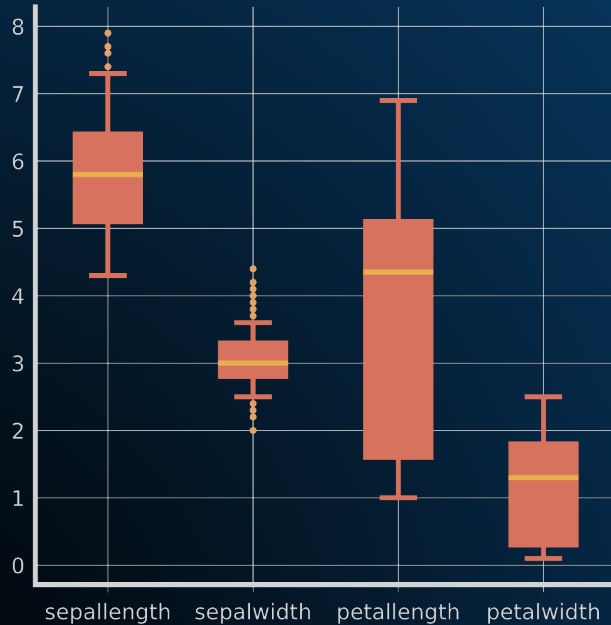
STANDARDIZATION



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

	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

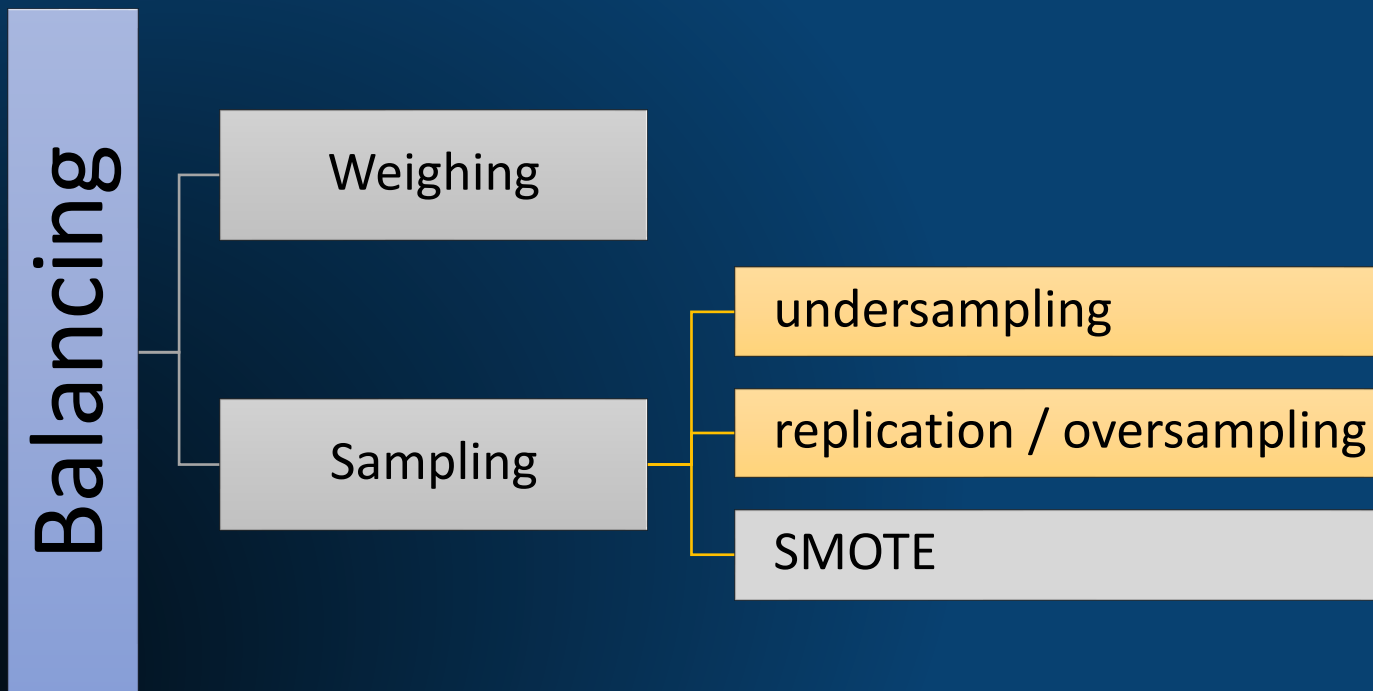




A hand is shown in the foreground, pointing towards a complex, futuristic digital interface. The interface features a large circular gauge with multiple concentric rings, some of which are highlighted in green. In the center of the gauge are three interlocking gears. The background is dark blue with various abstract digital elements, including lines, dots, and faint data visualizations. The overall aesthetic is high-tech and data-driven.

Data Balancing

DATA BALANCING



UNDERSAMPLING



Undersampling

REPLICATION / OVERSAMPLING



Oversampling

SMOTE







*Thank
you!*

