Data Science and Machine Learning

For New Enthusiasts

Rafael M Mamede (rafael.c.maia@inesctec.pt)

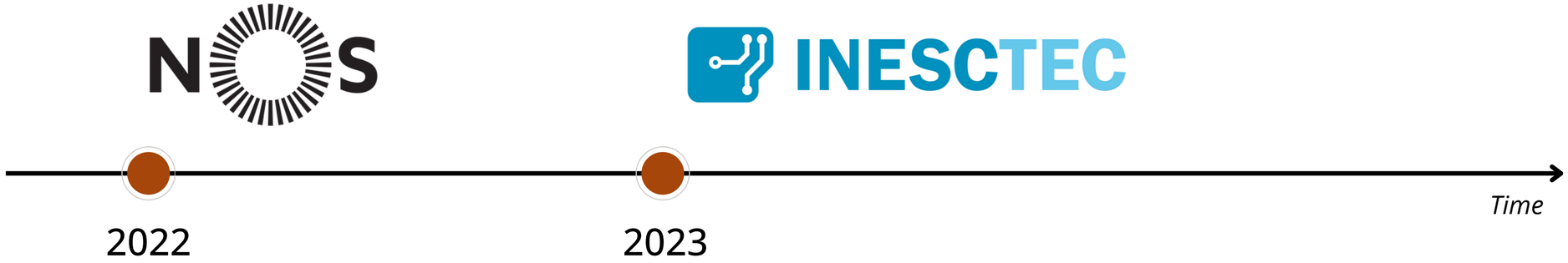
About me...

completed

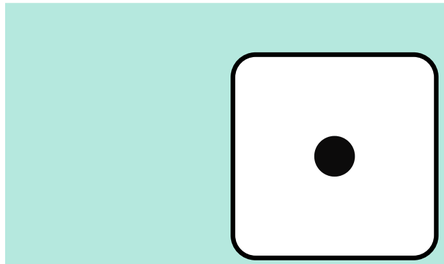
Physics BSc.	FCUP
Data Science MSc.	FCUP

on going

Informatics Engineering PhD	FEUP



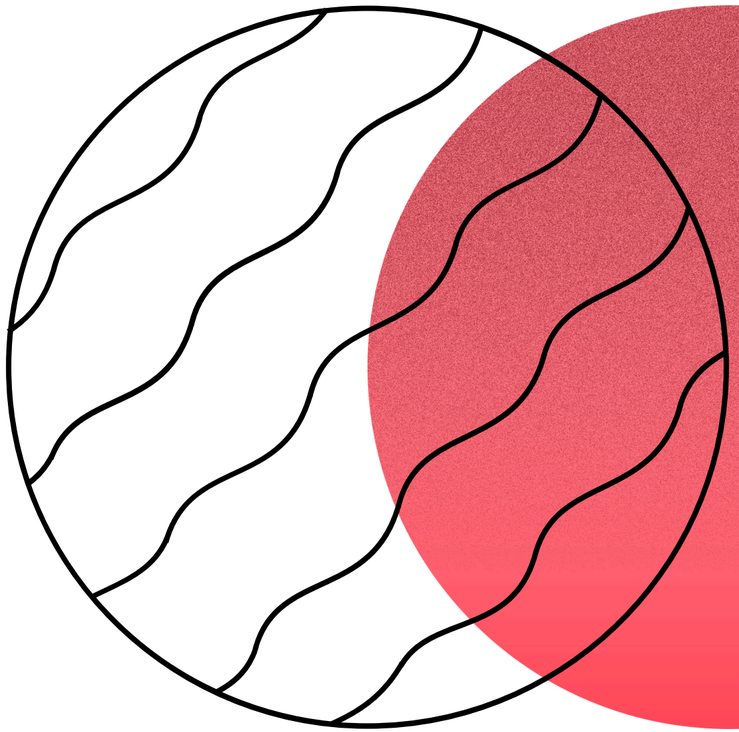
About the course…



Course Introduction
Data Exploration
Data Preprocessing

ML Modeling Approaches

Model Evalutation



Data Science?
Machine Learning?

What is Data Science?

What is Data Science?

"a field of study that uses scientific methods, processes, and systems to **extract knowledge and insights from data.**"
-U.S. Census Bureau

"Data Science is an interdisciplinary field which uses statistics, computer science, programming, and domain knowledge to collect, process, and analyze data for the purpose of **acquiring knowledge or solving a problem**."
-National Library of Medicine

"Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI) and machine learning with specific subject matter expertise to **uncover actionable insights** hidden in an organization's data."
-IBM

What is Data Science?

Data Science is the study of observations
to fulfill a goal

What is Data Science?

Data Science is the study of observations to fulfill a goal

Prove Hypothesis

- Does a new drug reduce symptoms more effectively?
- Is student performance linked to class attendance?

What is Data Science?

Data Science is the study of observations
to fulfill a goal

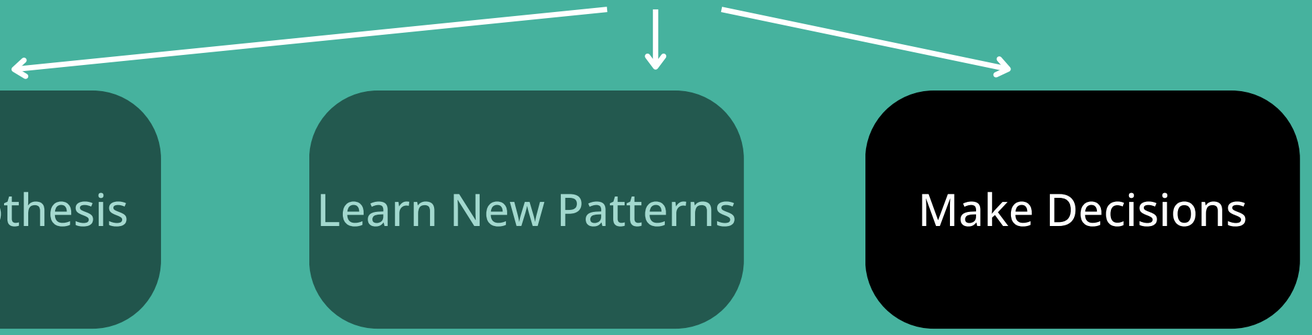
Prove Hypothesis

Learn New Patterns

- Segment customers by shopping behavior
- Discover latent patterns in cognitive ability

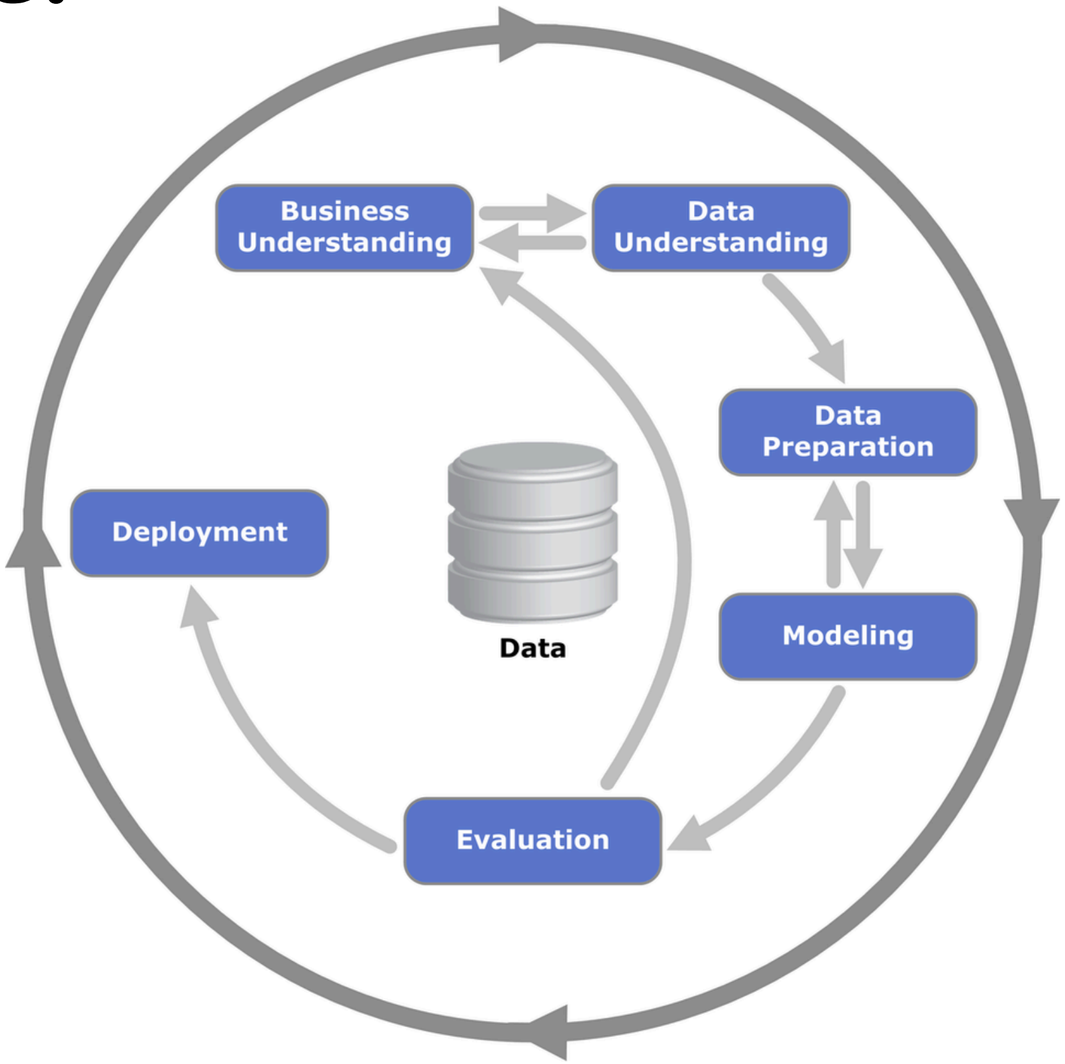
What is Data Science?

Data Science is the study of observations to fulfill a goal



- Approve or reject loans based on risk models.
- Forecast product demand to manage stock.

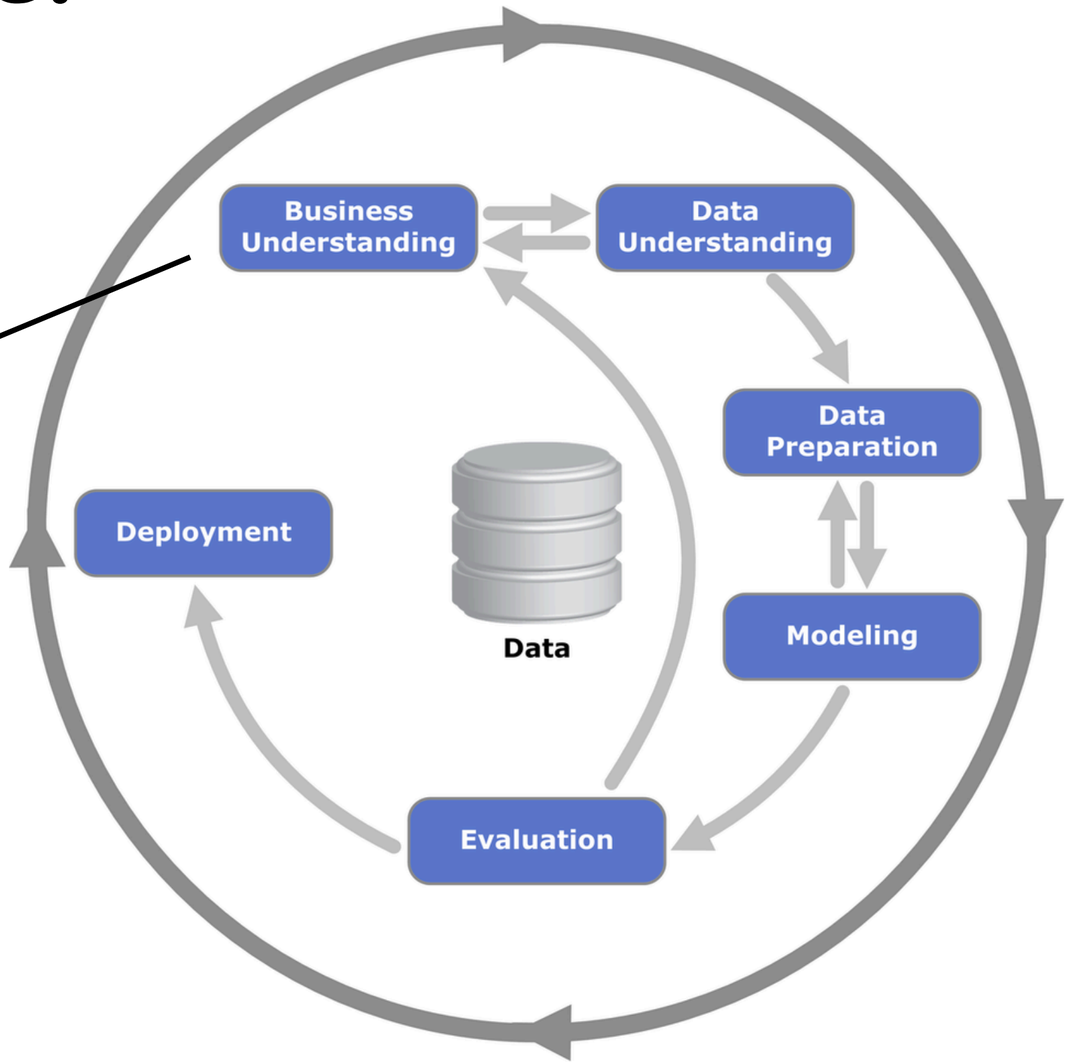
What does a Data Science project look like?



What does a Data Science project look like?

What is the real problem?

Project goals from a non technical perspective
Understand constraints, success criteria, priorities



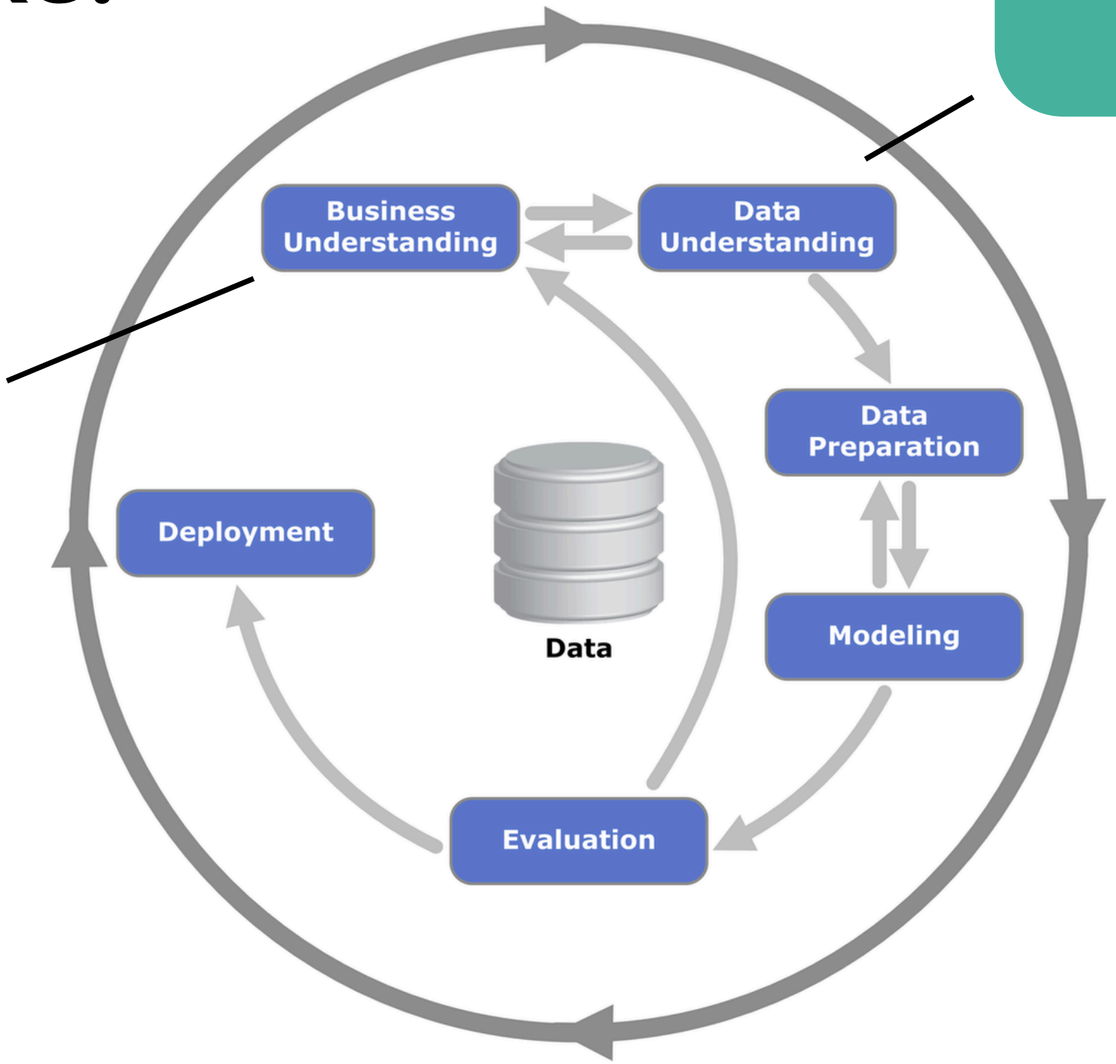
What does a Data Science project look like?

What data do we have?

Exploratory analysis to find patterns and issues
Identify missing, noisy, or biased data

What is the real problem?

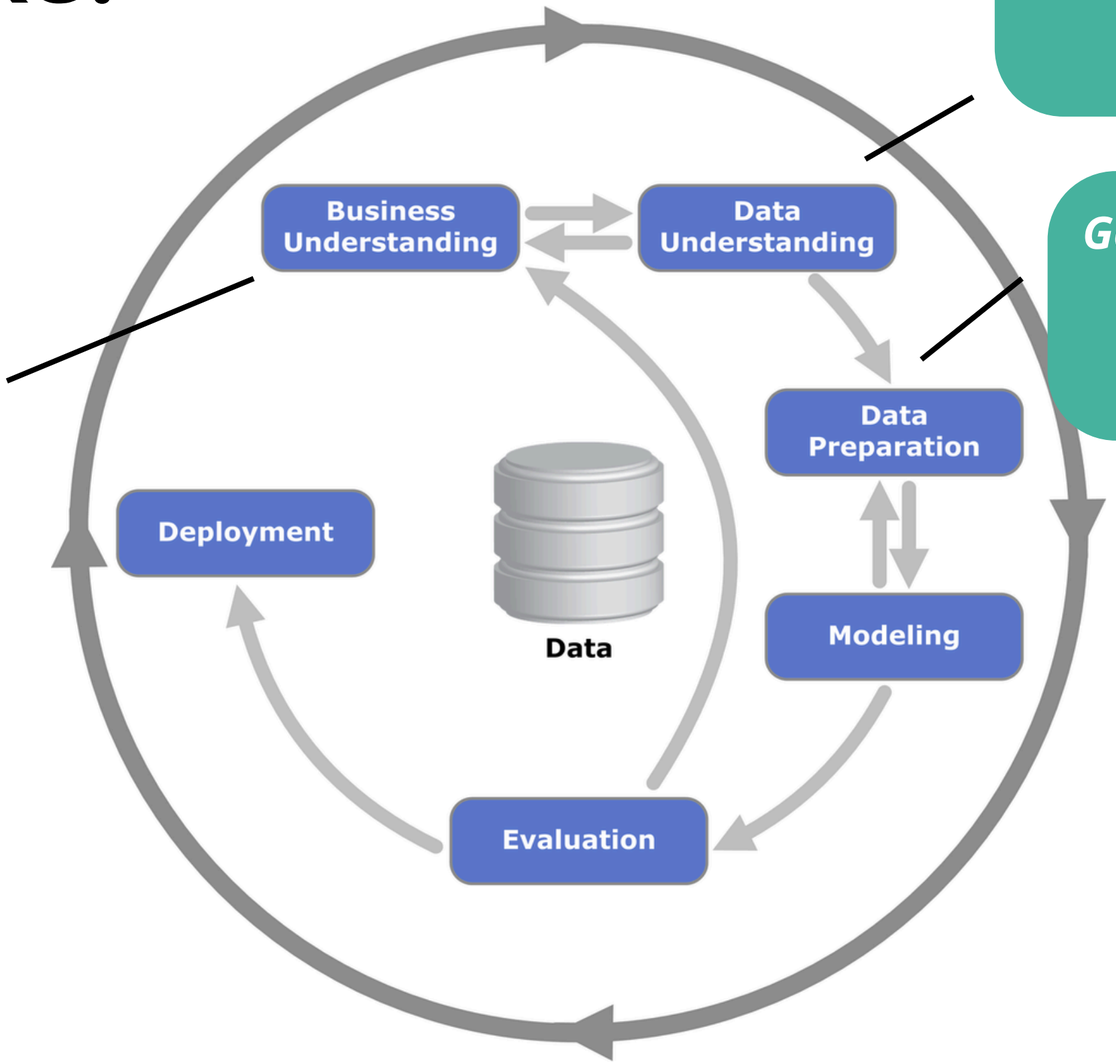
Project goals from a non technical perspective
Understand constraints, success criteria, priorities



What does a Data Science project look like?

What is the real problem?

Project goals from a non technical perspective
Understand constraints, success criteria, priorities



What data do we have?

Exploratory analysis to find patterns and issues
Identify missing, noisy, or biased data

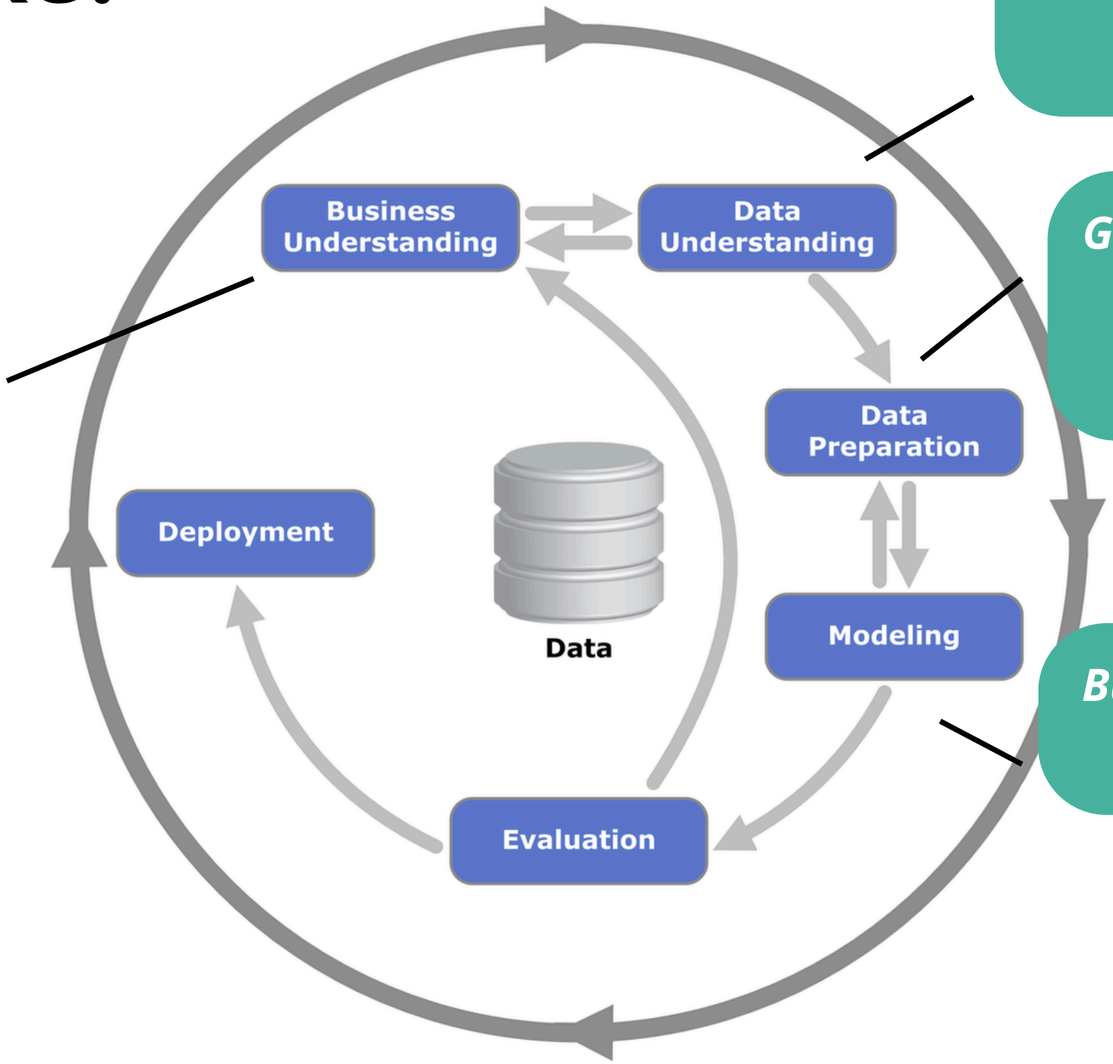
Get the data ready for modeling.

Clean data (missing values, outliers)
Data Transformations

What does a Data Science project look like?

What is the real problem?

Project goals from a non technical perspective
Understand constraints, success criteria, priorities



What data do we have?

Exploratory analysis to find patterns and issues
Identify missing, noisy, or biased data

Get the data ready for modeling.

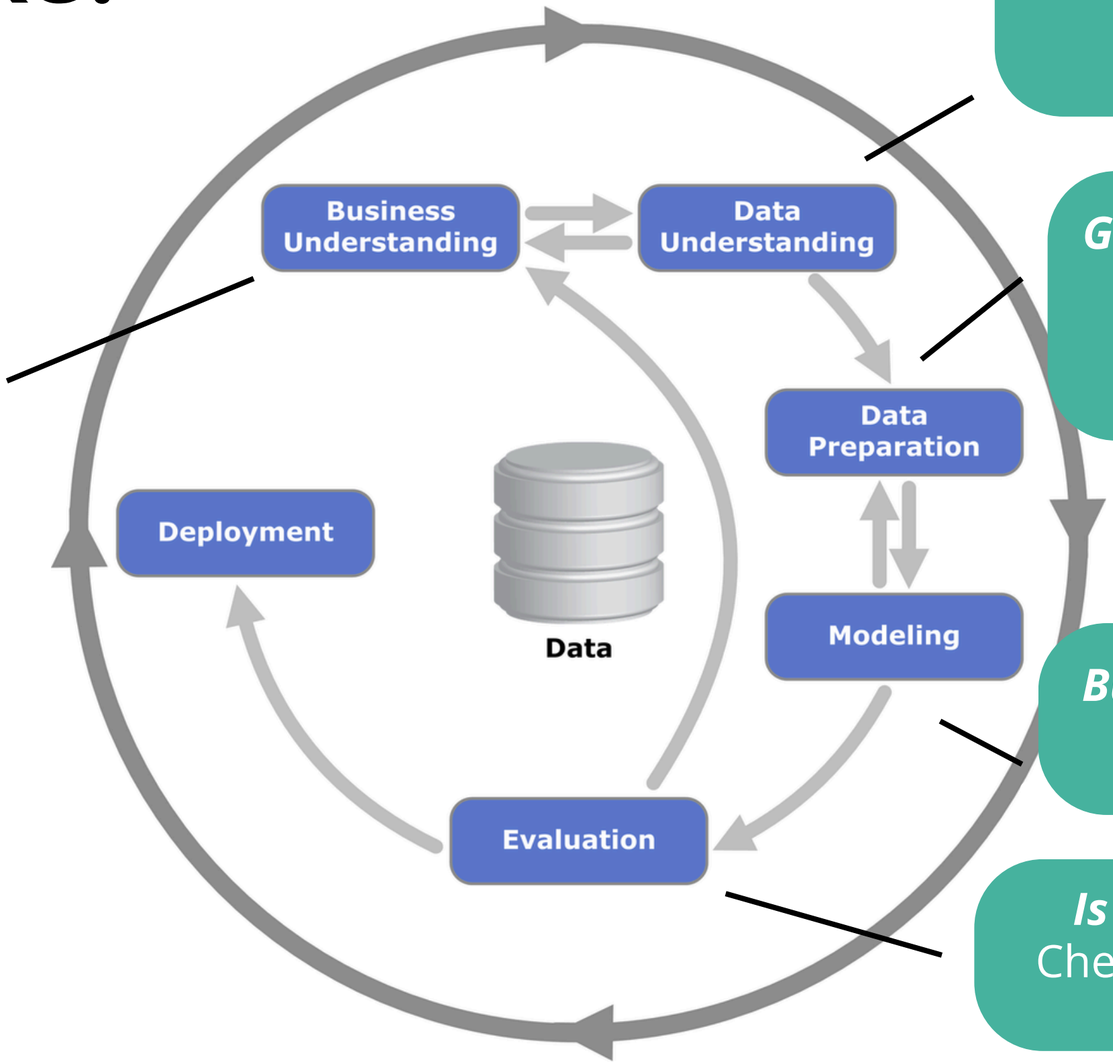
Clean data (missing values, outliers)
Data Transformations

Build models to find patterns or make predictions.

What does a Data Science project look like?

What is the real problem?

Project goals from a non technical perspective
Understand constraints, success criteria, priorities



What data do we have?

Exploratory analysis to find patterns and issues
Identify missing, noisy, or biased data

Get the data ready for modeling.

Clean data (missing values, outliers)
Data Transformations

Build models to find patterns or make predictions.

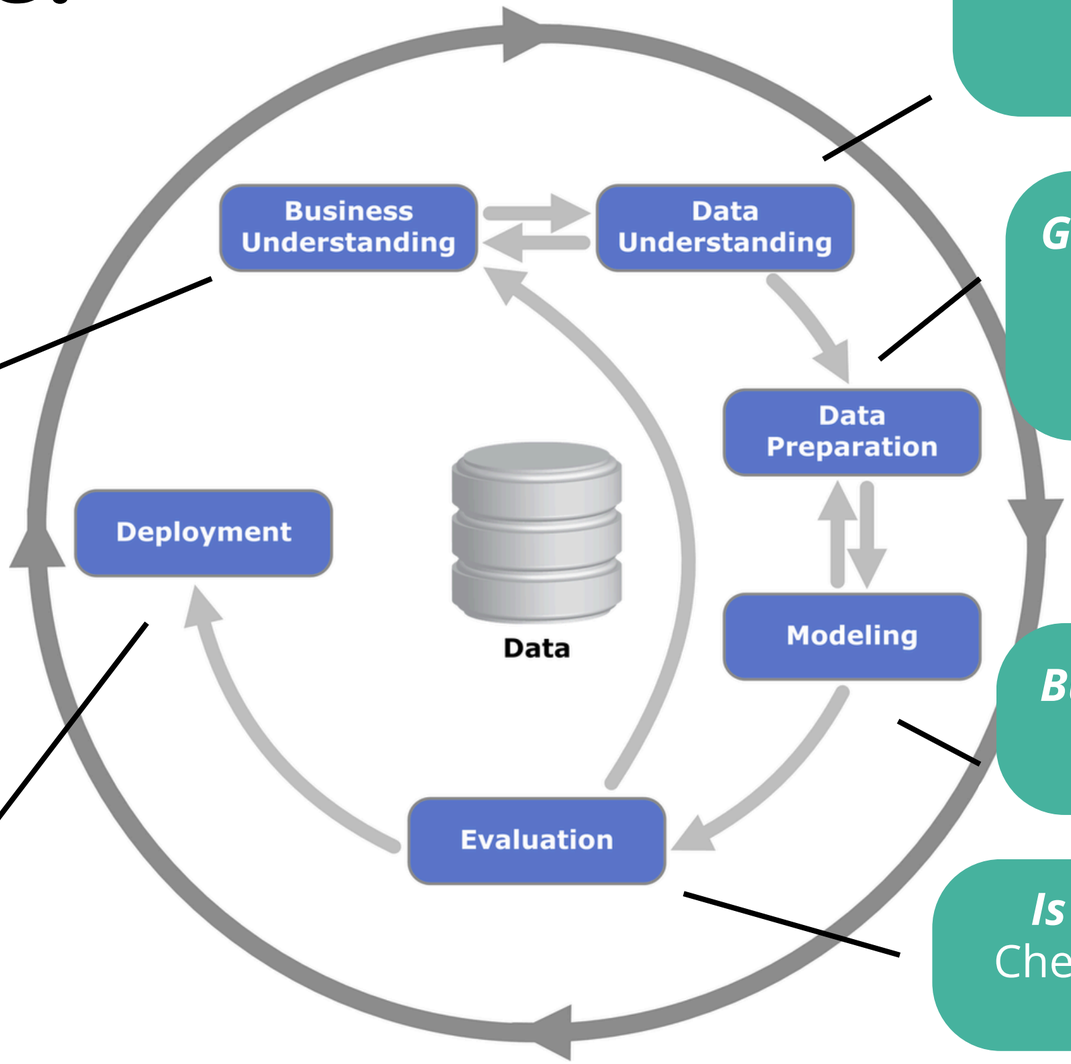
Is the model good enough?

Check performance (and other relevant) metrics

What does a Data Science project look like?

What is the real problem?

Project goals from a non technical perspective
Understand constraints, success criteria, priorities



What data do we have?

Exploratory analysis to find patterns and issues
Identify missing, noisy, or biased data

Get the data ready for modeling.

Clean data (missing values, outliers)
Data Transformations

Build models to find patterns or make predictions.

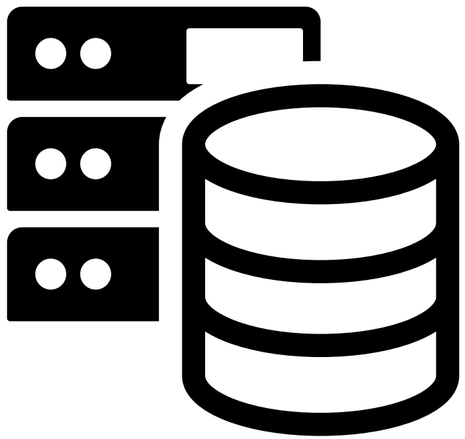
Is the model good enough?

Check performance (and other relevant) metrics

Put the model into real use.

What is Machine Learning?

What is Machine Learning?



MODEL REPRESENTATION

f(x;Θ): Predictors → Target

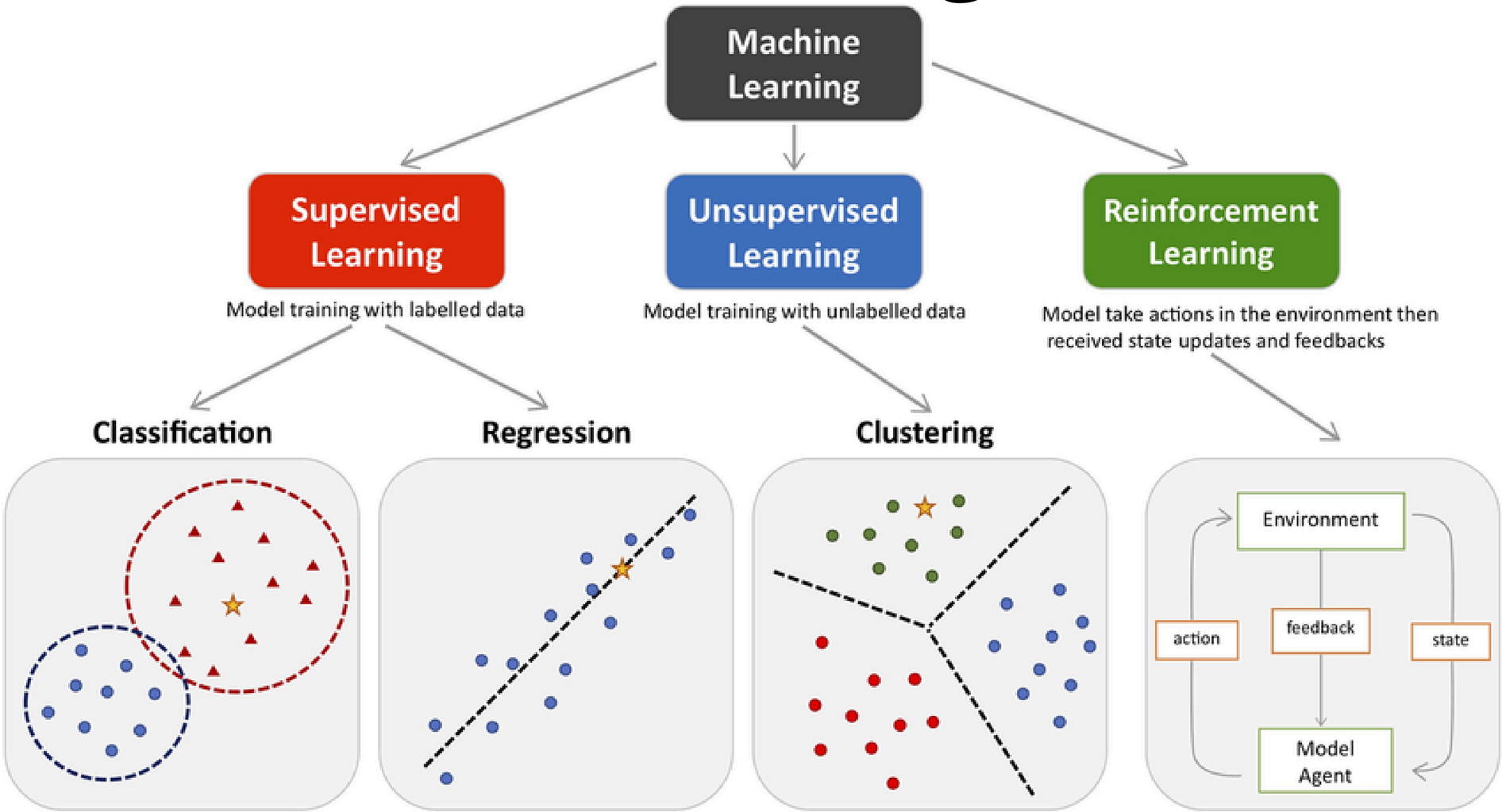
Goal

Find Best Θ

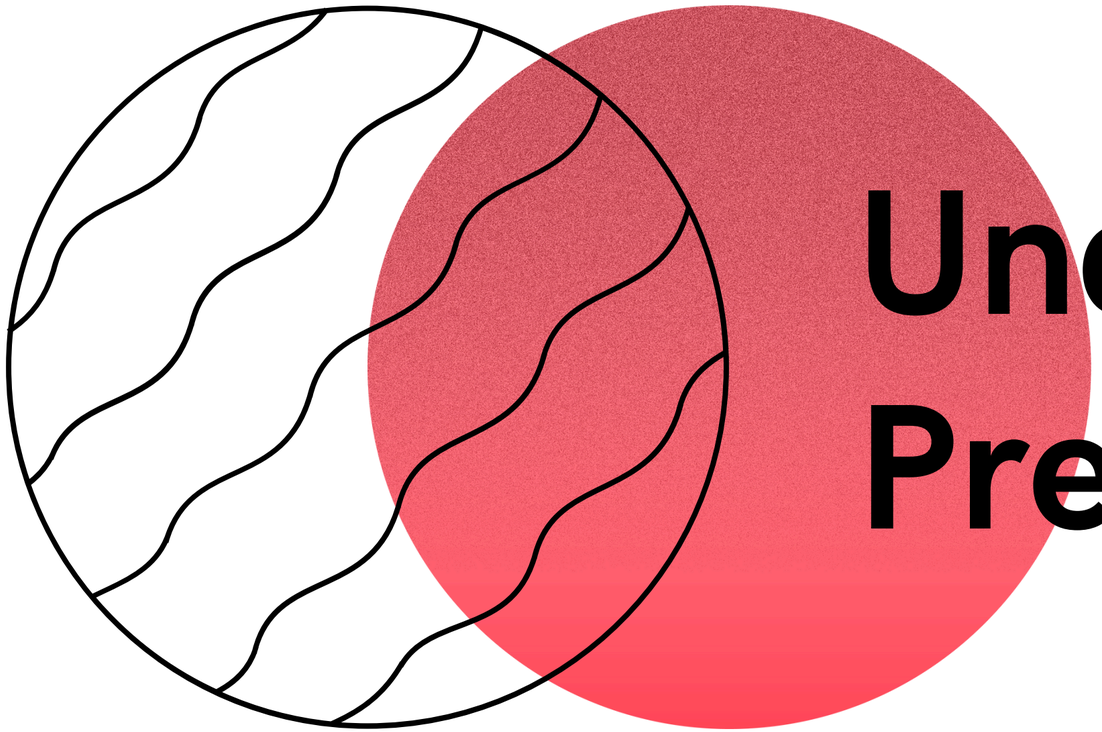
→ **Define optimizing function (either a Loss or a Goal)**

→ **Optimization Procedure**

What is Machine Learning?



FROM: Peng, Junjie & Jury, Elizabeth & Dönnes, Pierre & Ciurtin, Coziana. (2021). Machine Learning Techniques for Personalised Medicine Approaches in Immune-Mediated Chronic Inflammatory Diseases: Applications and Challenges. Frontiers in Pharmacology. 12. 10.3389/fphar.2021.720694.



Understanding and Pre-Procesing Data

Data Understanding

Types of Data

Data Understanding

Types of Data

Data Understanding

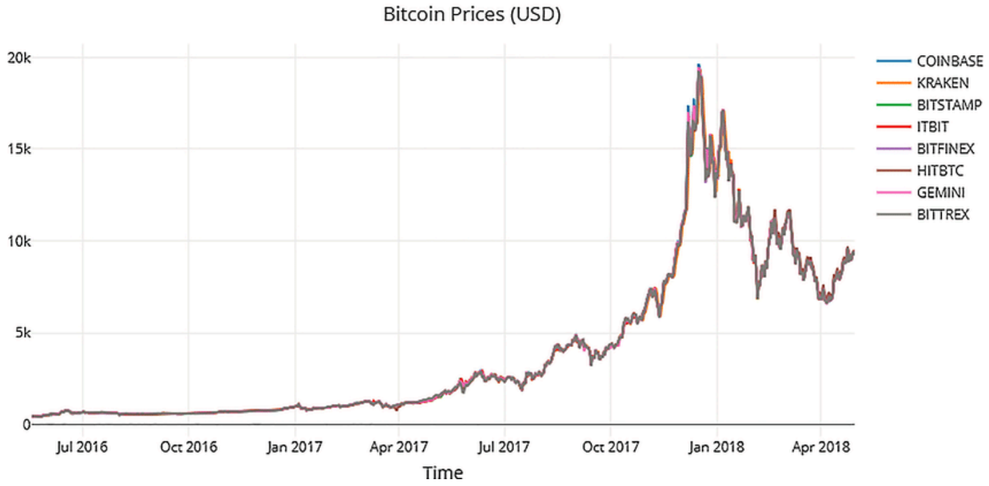
Types of Data

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3	1.4	0.1	Iris-setosa
14	4.3	3	1.1	0.1	Iris-setosa
15	5.8	4	1.2	0.2	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa
17	5.4	3.9	1.3	0.4	Iris-setosa
18	5.1	3.5	1.4	0.3	Iris-setosa
19	5.7	3.8	1.7	0.3	Iris-setosa

Data Understanding

Types of Data

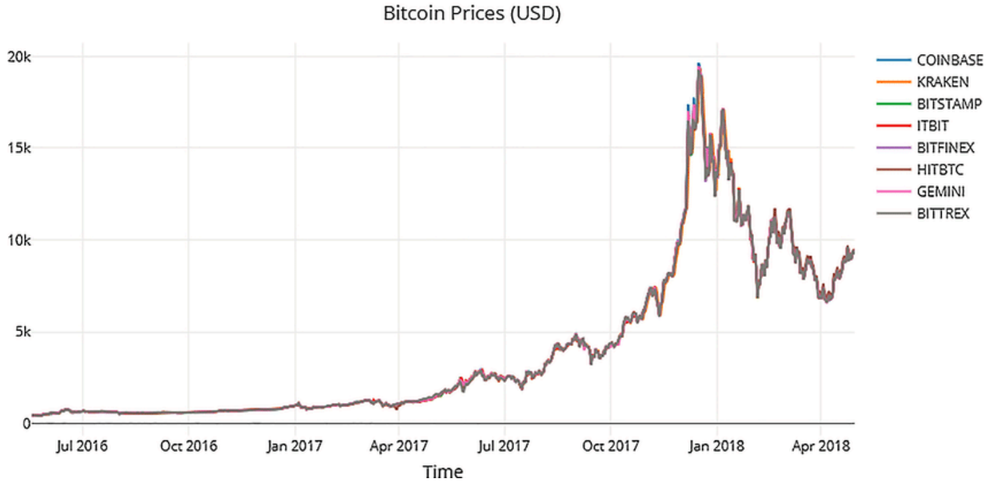
Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3	1.4	0.1	Iris-setosa
14	4.3	3	1.1	0.1	Iris-setosa
15	5.8	4	1.2	0.2	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa
17	5.4	3.9	1.3	0.4	Iris-setosa
18	5.1	3.5	1.4	0.3	Iris-setosa
19	5.7	3.8	1.7	0.3	Iris-setosa



Data Understanding

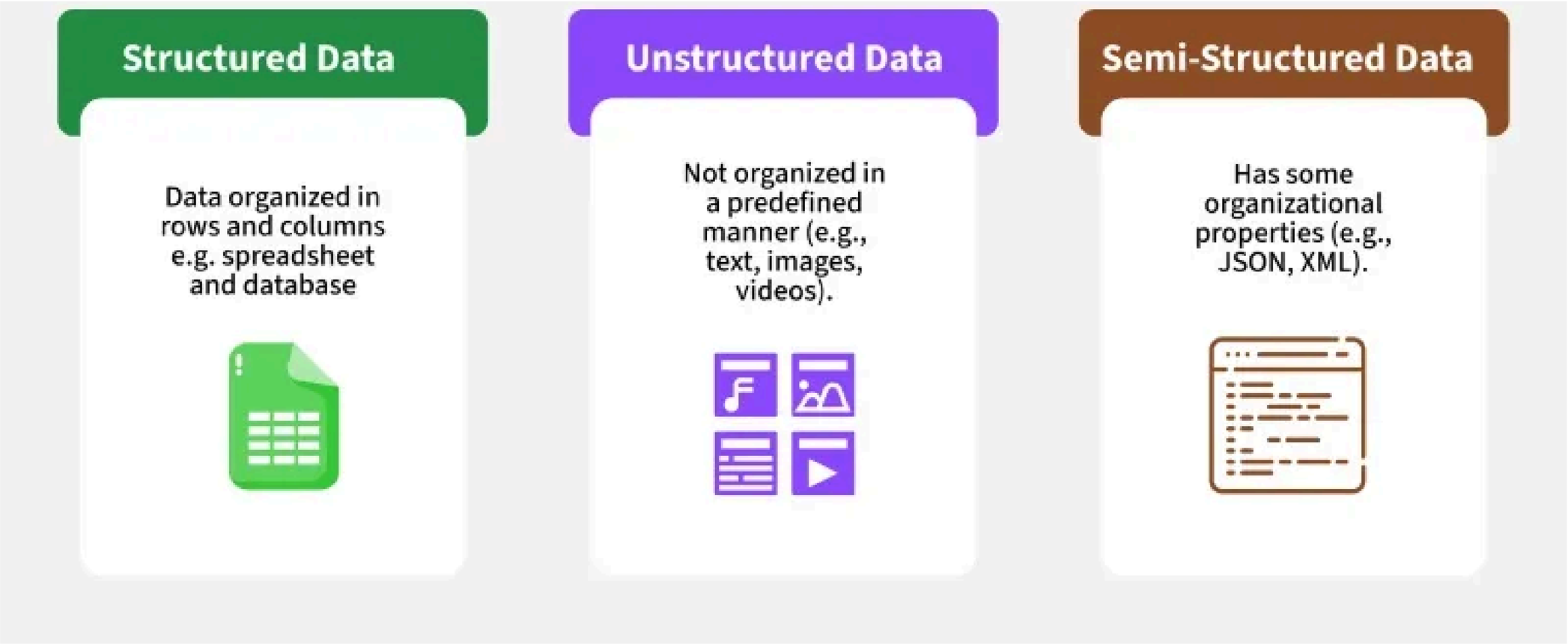
Types of Data

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3	1.4	0.1	Iris-setosa
14	4.3	3	1.1	0.1	Iris-setosa
15	5.8	4	1.2	0.2	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa
17	5.4	3.9	1.3	0.4	Iris-setosa
18	5.1	3.5	1.4	0.3	Iris-setosa
19	5.7	3.8	1.7	0.3	Iris-setosa



Data Understanding

Types of Data



FROM: https://www.geeksforgeeks.org/ml-introduction-data-machine-learning/

Data Understanding

Exploratory Data Analysis

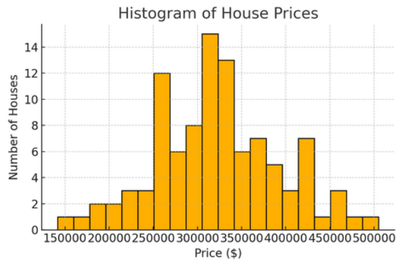
Exploratory Data Analysis (EDA) is the process of **exploring, summarizing, and visualizing data** to uncover patterns, spot problems, and understand key features before building models.

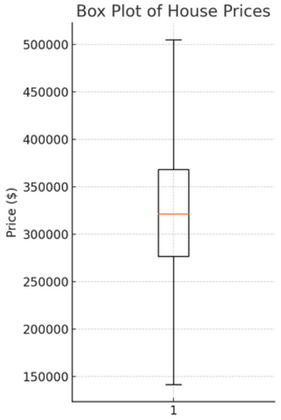
Data Understanding

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is the process of **exploring, summarizing, and visualizing data** to uncover patterns, spot problems, and understand key features before building models.

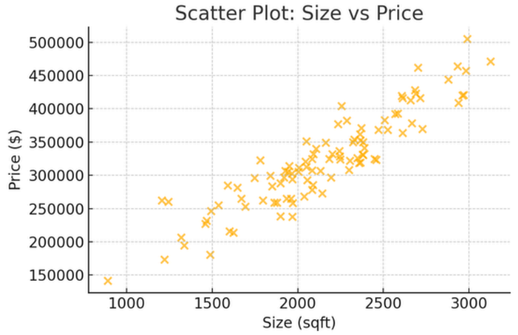
Univariate analysis

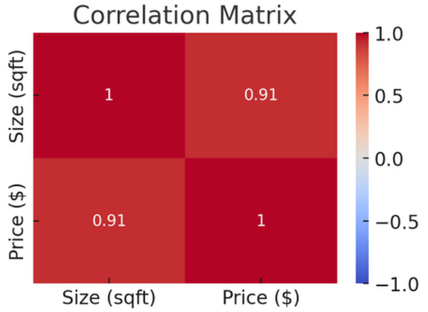




Centrality measures (mean, median)
Dispersion measures (STD, IQR)

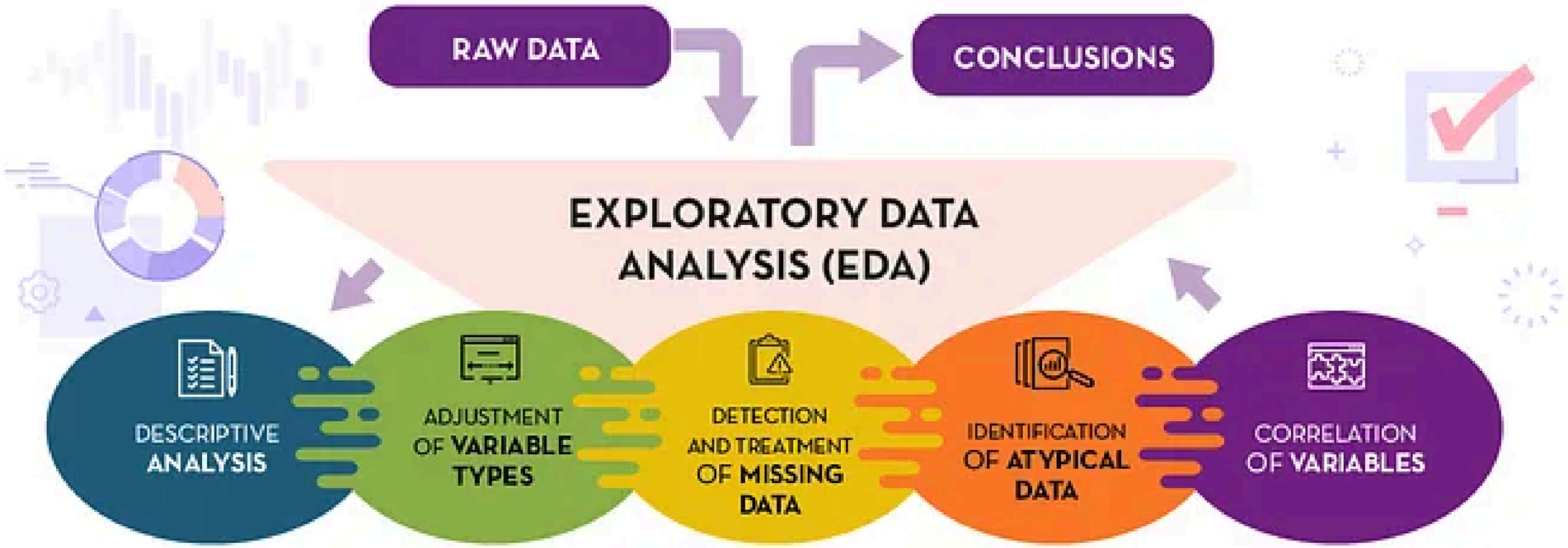
Multivariate Analysis





Data Understanding

Exploratory Data Analysis (EDA)



FROM: https://www.geeksforgeeks.org/ml-introduction-data-machine-learning/

Data Preparation

Data in the real world is dirty:

- *Incomplete*: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
- *Noisy*: containing errors or outliers
- *Inconsistent*: containing discrepancies in codes or names

No quality in the data, no quality in the model!

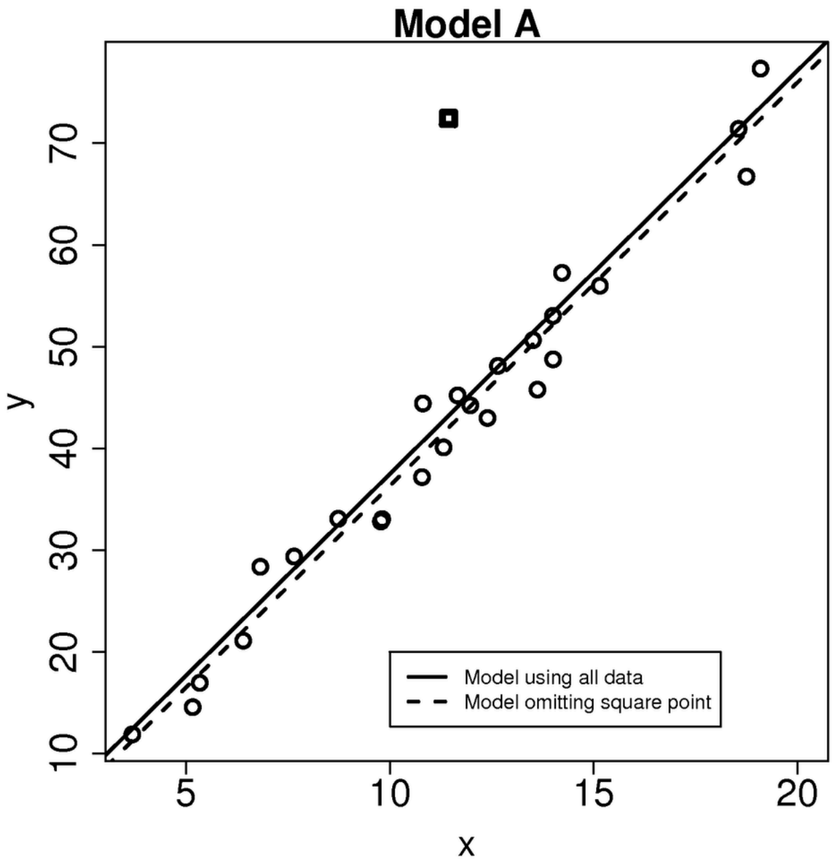
Data Preparation

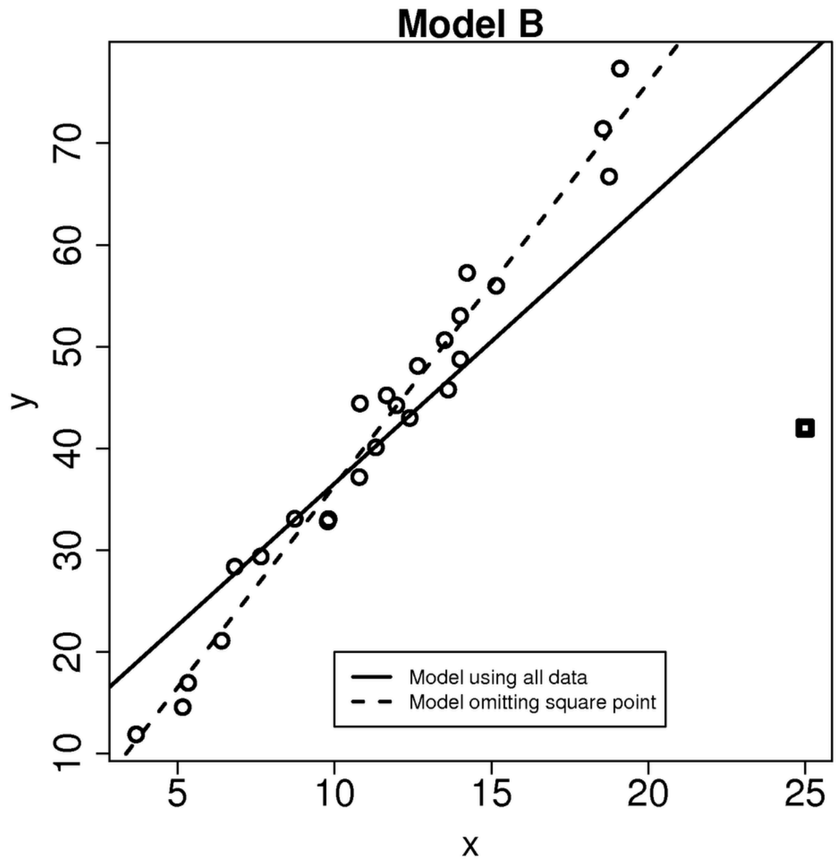
Data Cleaning

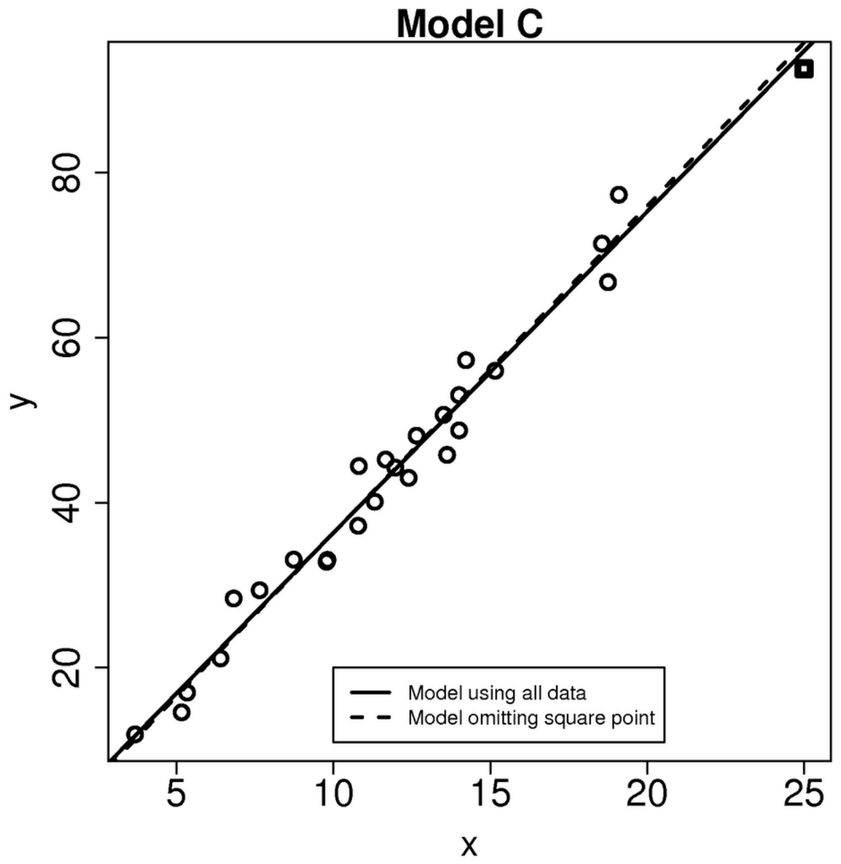
Data Transformation

Dimensionality Reduction

Data Preparation
Dealing with outliers (cleaning)







FROM: https://www.geeksforgeeks.org/ml-introduction-data-machine-learning/

Data Preparation

Missing data (cleaning)

F1	F2	F3	F4	F5	F6
0.0	26.4	NaN	0.0	sunny	NaN
0.0	NaN	96.0	0.0	rainy	0.21
0.0	24.1	68.0	0.0	overcast	0.68
NaN	24.7	98.0	0.0	rainy	0.20
0.0	26.5	98.0	0.0	NaN	0.32
0.0	27.6	78.0	0.0	rainy	0.72
0.0	28.2	NaN	0.0	rainy	0.61
0.0	27.1	70.0	0.0	overcast	NaN
1.0	26.7	75.0	NaN	sunny	0.54
0.0	NaN	NaN	0.0	NaN	NaN
NaN	24.3	77.0	0.0	overcast	0.67
0.0	23.1	77.0	1.0	sunny	0.66
0.0	22.4	89.0	1.0	rainy	0.38
0.0	NaN	80.0	1.0	sunny	0.46
0.0	26.5	88.0	1.0	rainy	NaN
0.0	28.6	76.0	0.0	NaN	0.52
0.0	NaN	NaN	NaN	rainy	NaN

Data Preparation

Missing data (cleaning)

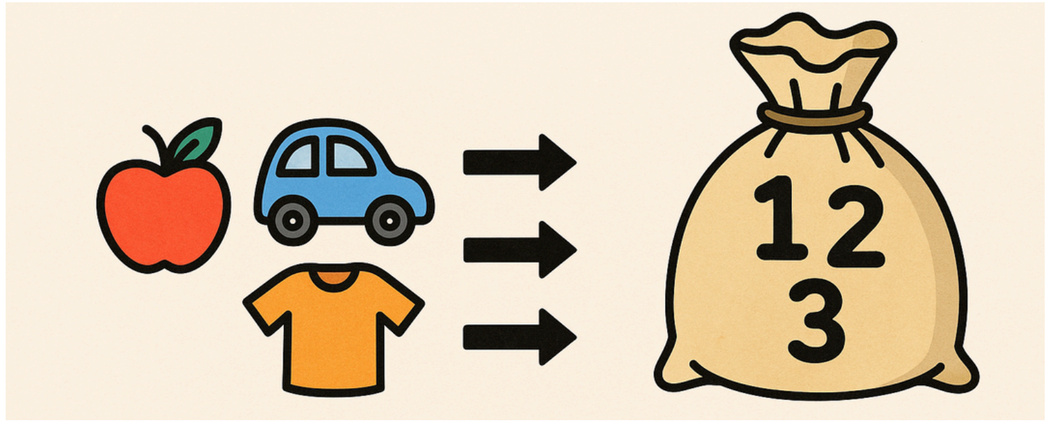
F1	F2	F3	F4	F5	F6
0.0	26.4	NaN	0.0	sunny	NaN
0.0	NaN	96.0	0.0	rainy	0.21
0.0	24.1	68.0	0.0	overcast	0.68
NaN	24.7	98.0	0.0	rainy	0.20
0.0	26.5	98.0	0.0	NaN	0.32
0.0	27.6	78.0	0.0	rainy	0.72
0.0	28.2	NaN	0.0	rainy	0.61
0.0	27.1	70.0	0.0	overcast	NaN
1.0	26.7	75.0	NaN	sunny	0.54
0.0	NaN	NaN	0.0	NaN	NaN
NaN	24.3	77.0	0.0	overcast	0.67
0.0	23.1	77.0	1.0	sunny	0.66
0.0	22.4	89.0	1.0	rainy	0.38
0.0	NaN	80.0	1.0	sunny	0.46
0.0	26.5	88.0	1.0	rainy	NaN
0.0	28.6	76.0	0.0	NaN	0.52
0.0	NaN	NaN	NaN	rainy	NaN

Strategies?

- Remove observations with missing values
- Remove features with missing values
- Impute values (mean, median, mode, fixed value)
- Impute values with auxiliary model

Data Preparation

Data Transformation - Encoding



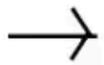
Machine Learning Models Work with Numbers

- Algorithms like linear regression, decision trees, and neural networks require numerical input.
- Text labels (like "Red", "Blue") are not mathematically meaningful to models.

Data Preparation
Data Transformation - Encoding

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50



One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Data Preparation

Data Transformation - Scaling

Feature scaling is the process of adjusting the range of feature values so that they are on a similar scale.

Some models (e.g. Distance-based models) are **sensitive to feature magnitudes.**

Without scaling, features with larger values can dominate the model.

Scaling improves convergence during optimization.

Data Preparation

Data Transformation - Scaling

Normalization (Min-Max Scaling)

$$x_{\text{scaled}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Preserves the original distribution shape (no distortion).
- Easy to interpret (values between 0 and 1).

- Sensitive to outliers (extreme values compress the scale).
- New unseen data outside the original min-max range can cause problems.

Standardization (Z-Score Scaling)

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

- Handles outliers better (compared to Min-Max).
- Centers data (mean = 0), useful for algorithm optimization procedure.

- Doesn't bound data to a specific range (values could be very large or very small).
- Harder to interpret scaled values.

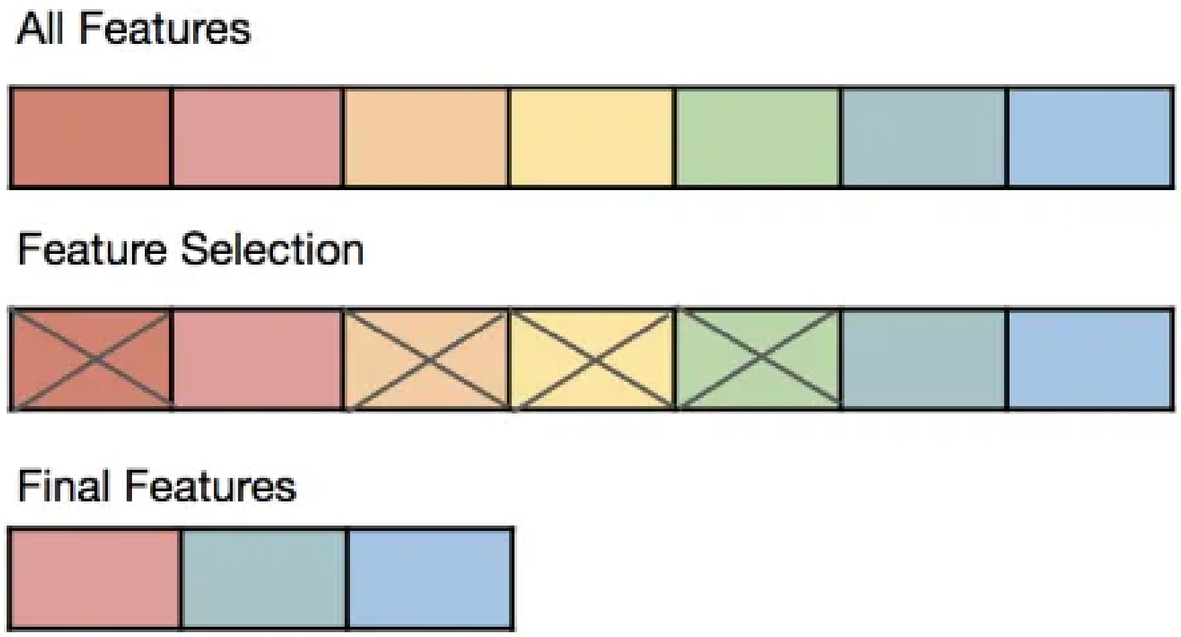
Data Preparation
Dimensionality reduction

Dimensionality reduction is the process of reducing the number of input features in a dataset while preserving important information.

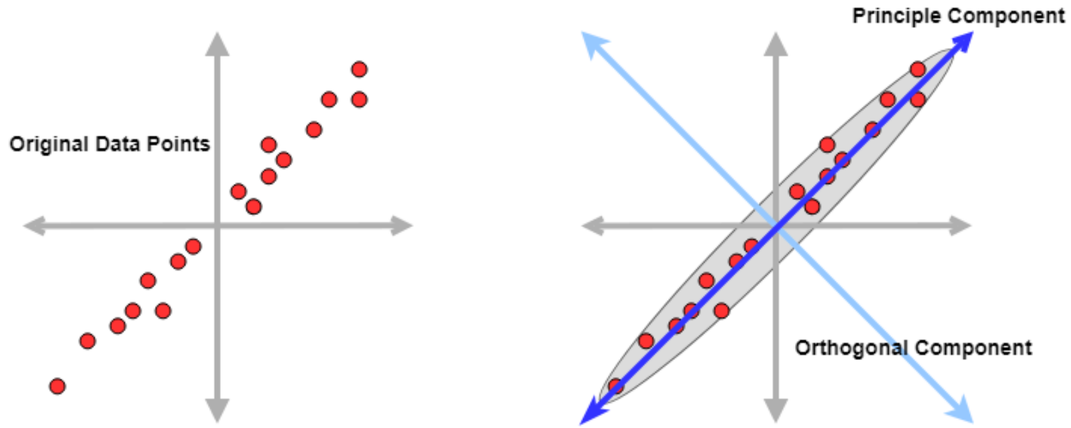
Reduces complexity → easier to understand and visualize.
Speeds up models → faster training and prediction.

Data Preparation
Dimensionality reduction

Feature Selection

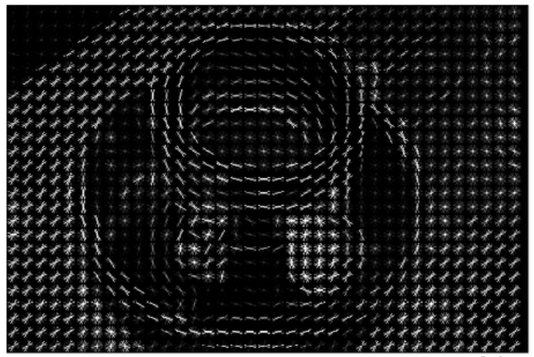


Feature Extraction



Test Image

Histogram of Oriented Gradients



FROM: https://medium.com/@nirajan_DataAnalyst/understanding-feature-selection-techniques-in-machine-learning-02e2642ef63e
https://www.baeldung.com/cs/kernel-principal-component-analysis
https://www.ml-science.com/histogram-of-oriented-gradients

Key Takeaways

Feedback here

- What Data Science is and what it aims to do
- The CRISP-DM structure of a Data Science project
- Importance of Data Understanding and Data Preparation
- Common Steps of Data Preparation (Cleaning, Encoding, Scaling, Dimensionality Reduction)

Building Machine Learning Models