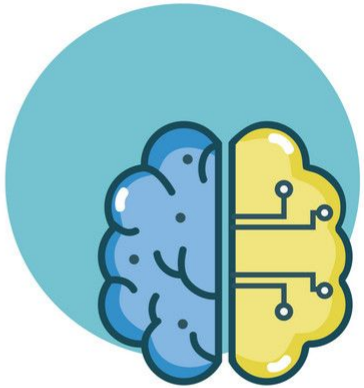
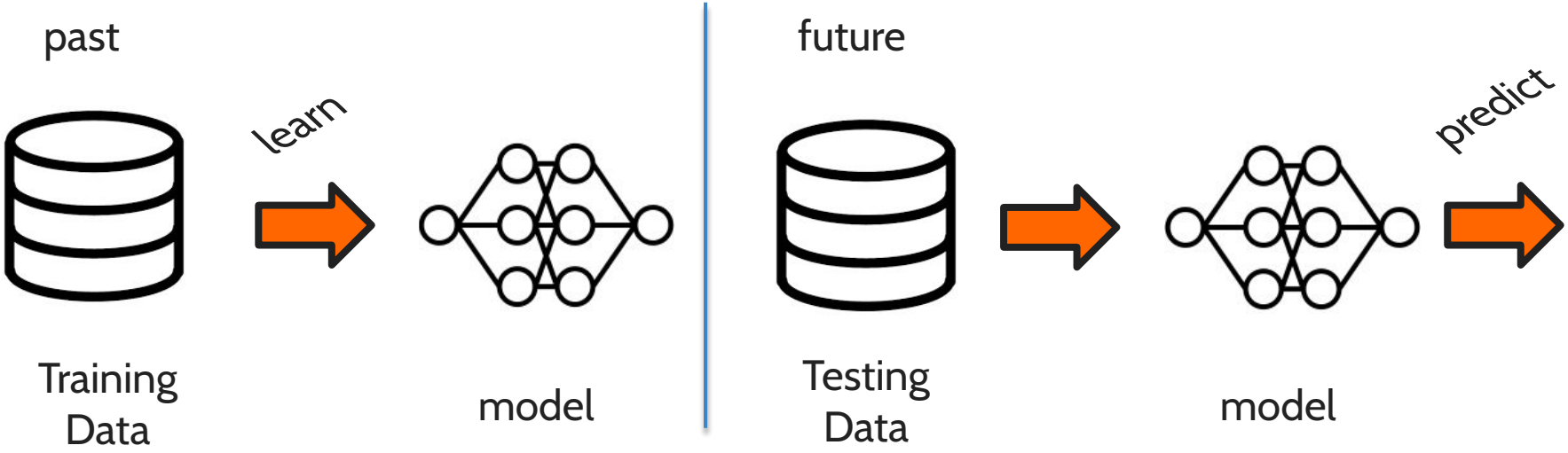
Introduction to Machine Learning

Machine Learning Basics: Data, Features, Models

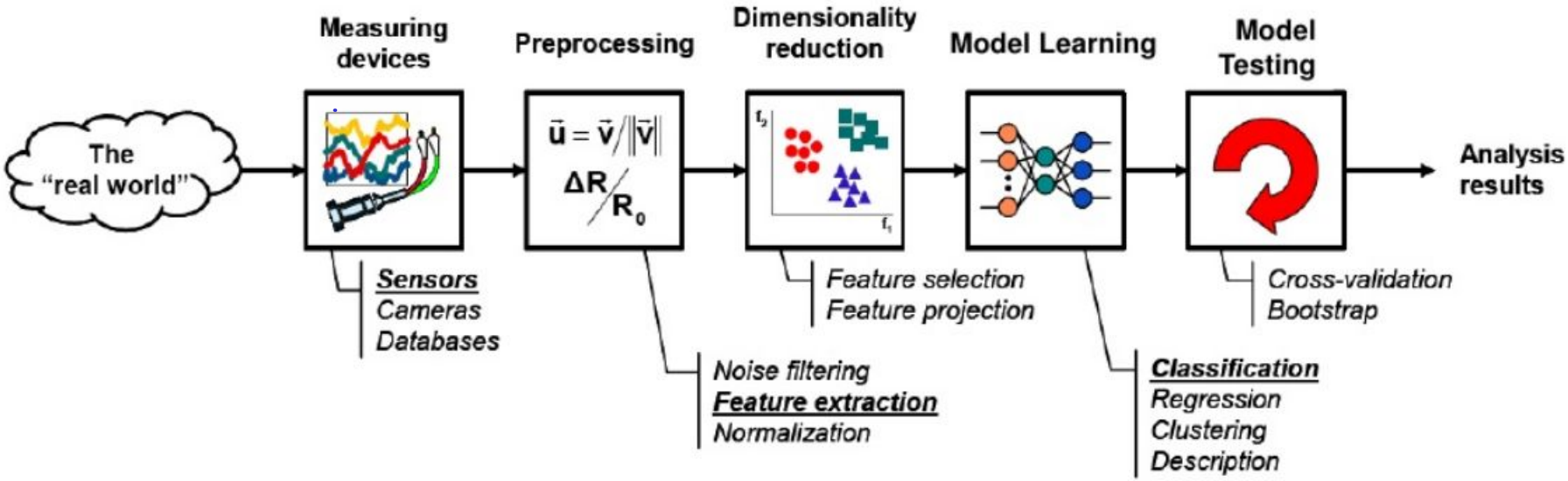
Elisa Ricci



The Learning Process



The Learning Process



The Learning Process

Data Acquisition: This phase involves collecting relevant data for the problem at hand. This data can come from various sources such as databases or sensors. Ensuring the quality and relevance of the data is crucial at this stage.

The Learning Process

Data Preprocessing: Once the data is acquired, it often needs to be cleaned and prepared for analysis. This step involves handling missing values, dealing with outliers, and transforming the data into a suitable format for further analysis. Common preprocessing techniques include normalization, feature scaling, and handling categorical variables.

a. Employees

name	*age*	*salary*
Paul	1978	NULL
Paul	NULL	29,000
Paul	1979	NULL
Melanie	1990	NULL
Bob	NULL	37,000
Bob	1977	NULL
Charlie	1978	32,000

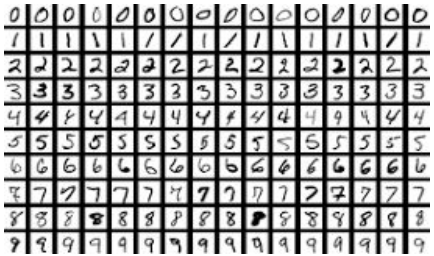
b. Employees

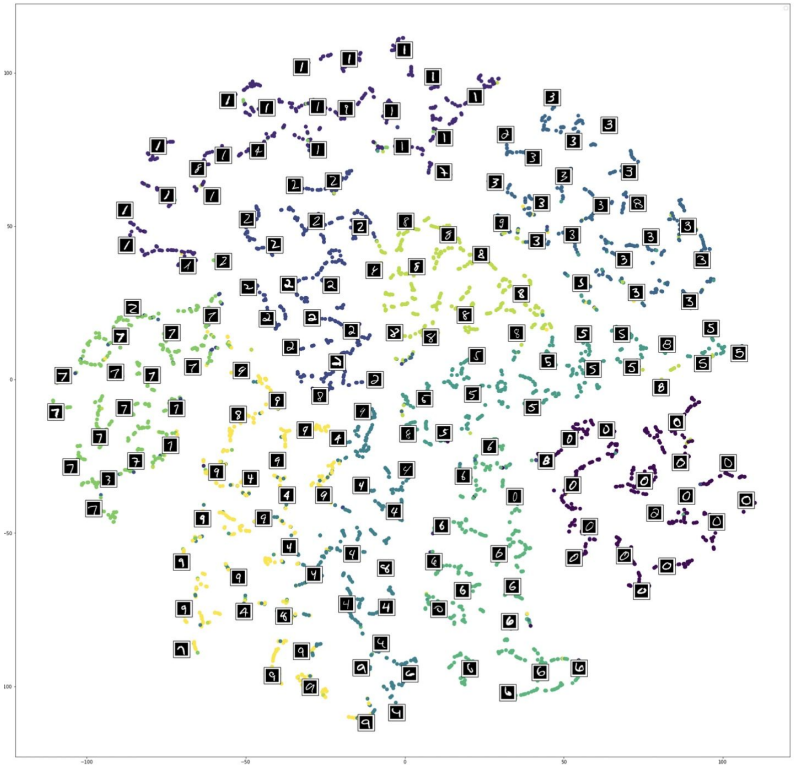
name	*age*	*salary*
Paul	?	29,000
Melanie	1990	NULL
Bob	1977	37,000
Charlie	1978	32,000

Paul.age = 1978 OR 1979 ?

The Learning Process

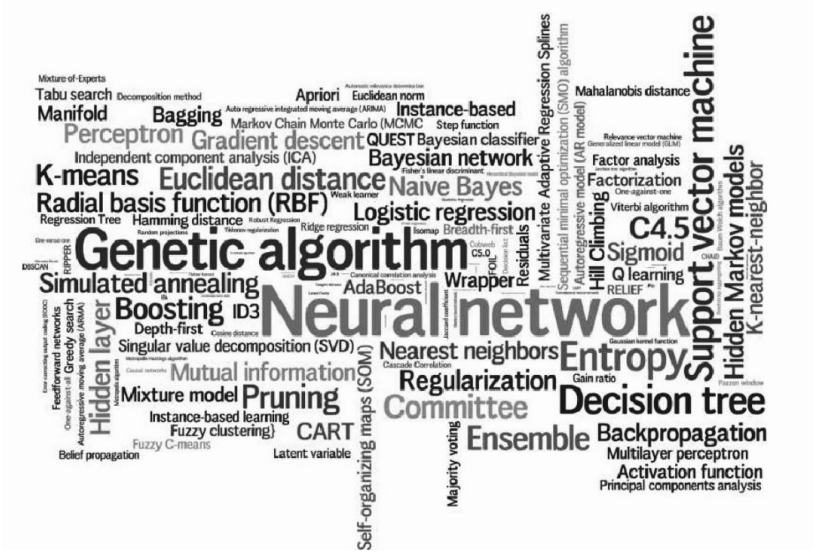
Dimensionality Reduction: Datasets may contain a large number of features, which can lead to overfitting and increased computational complexity. Dimensionality reduction techniques or feature selection methods can be applied to reduce the number of features while preserving the most important information.





The Learning Process

Model Learning: This is the core of the machine learning pipeline, where a model is trained on the preprocessed data to learn patterns and relationships. The choice of model depends on the nature of the problem (e.g., classification, regression) and the characteristics of the data. Common machine learning algorithms include linear regression, decision trees, support vector machines, and neural networks.



The Learning Process

Model Testing: Once the model is trained, it needs to be evaluated to assess its performance and **generalization capabilities.** This is typically done using a separate dataset called the test set, which was not used during the training phase. **Performance metrics** such as accuracy, precision, recall, F1-score, or mean squared error are computed to quantify the model's performance. Additionally, techniques like cross-validation can be used to assess the model's robustness.

		Predicted	
		Has Cancer	Doesn't Have Cancer
Ground Truth	Has Cancer	TP	FN
	Doesn't Have Cancer	FP	TN

The Learning Process

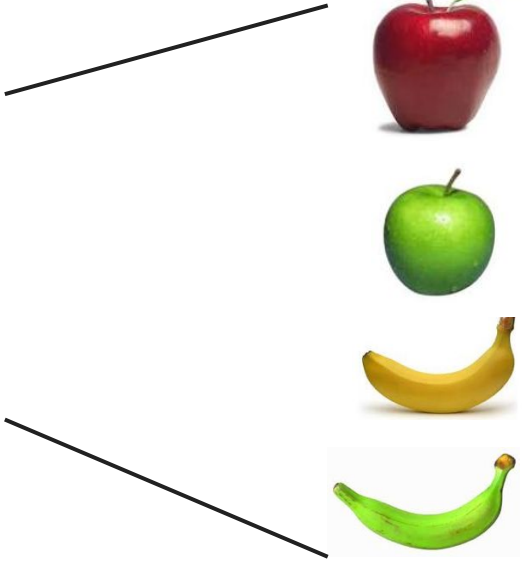
Throughout the pipeline, it's important to **iterate and refine** each step based on the insights gained from the evaluation phase. This iterative process helps improve the performance of the model and ensure its effectiveness in real-world applications.

Let's start with Data

Data

examples

Data



Features

Rappresentazione tramite features

examples

features

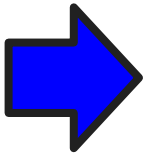
$$x^1_1, x^1_2, x^1_3, \ldots, x^1_n$$
$$x^2_1, x^2_2, x^2_3, \ldots, x^2_n$$
$$x^3_1, x^3_2, x^3_3, \ldots, x^3_n$$
$$x^4_1, x^4_2, x^4_3, \ldots, x^4_n$$

How our algorithms actually "view" the data

Features

examples

perdita di informazione



features

red, round, leaf, 3oz, ...

green, round, no leaf, 4oz, ...

yellow, curved, no leaf, 8oz, ...

green, curved, no leaf, 7oz, ...

How our algorithms actually "view" the data

Features are the questions we can ask about the examples

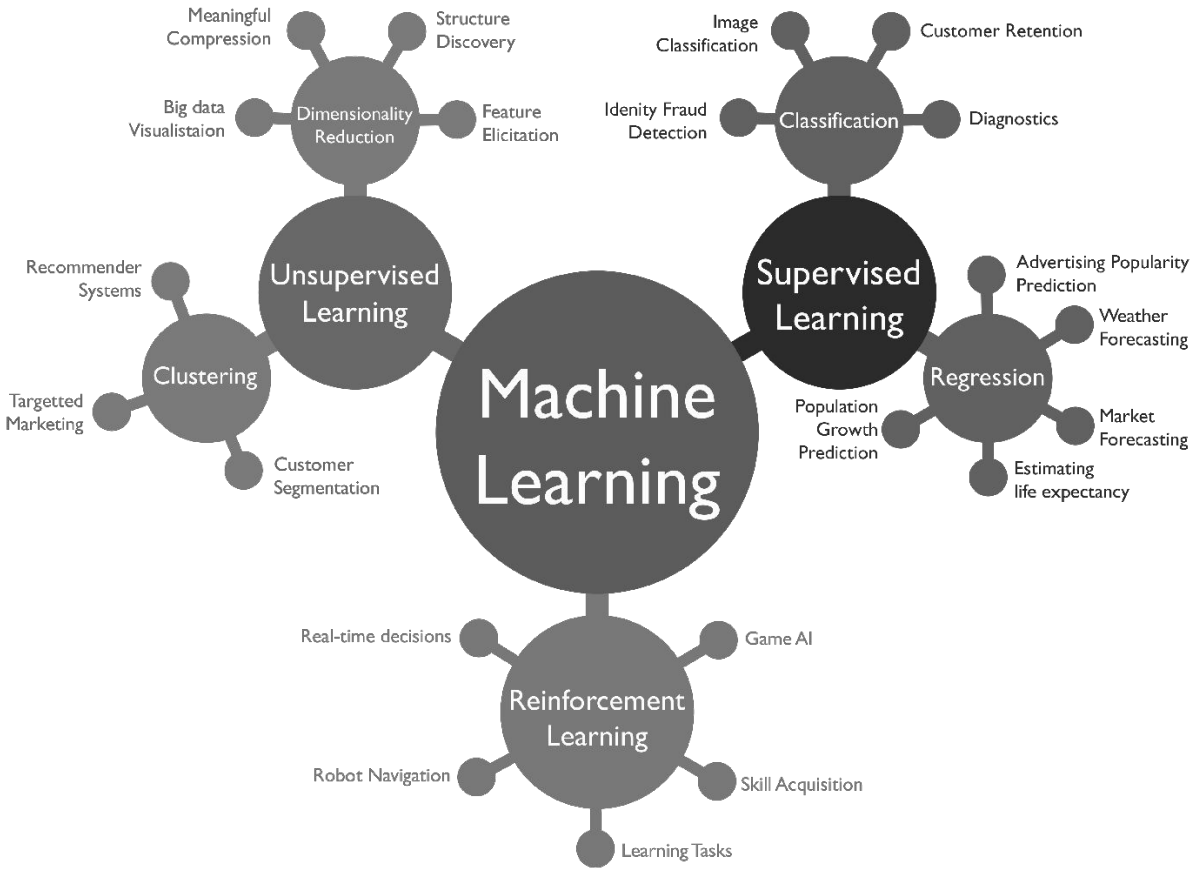
Features in generally are represented with vectors

Take Home Message

Mapping data to features often implies a **loss of information** because features are typically derived or selected representations of the original data that capture certain aspects deemed relevant for a particular task or analysis.

This process involves **simplification or abstraction**, which can lead to a reduction in the amount of information available compared to the raw data.

Types of Learning



Supervised learning

examples

label

y_1

y_2

y_3

y_4

un umano fornisce dei label

LABELED EXAMPLES

Given labeled examples...

Supervised learning

examples

label

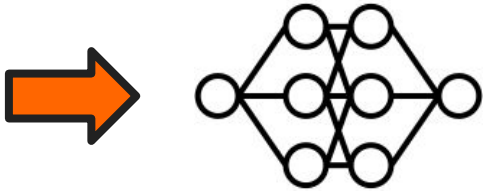
y_1

y_2

y_3

y_4

model

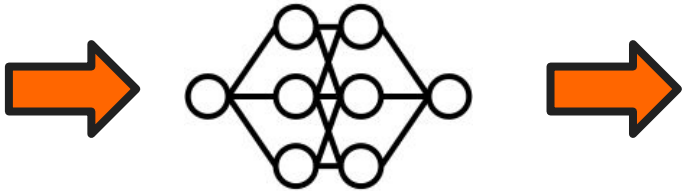


… build a model and…

Supervised learning

test example

produrre un label quando gli viene
fornito un nuovo sample dopo aver imparato
da una lista di sample e label assegnati da un umano



predicted label

y

... learn to predict the label associated to a new example

Supervised learning: classification

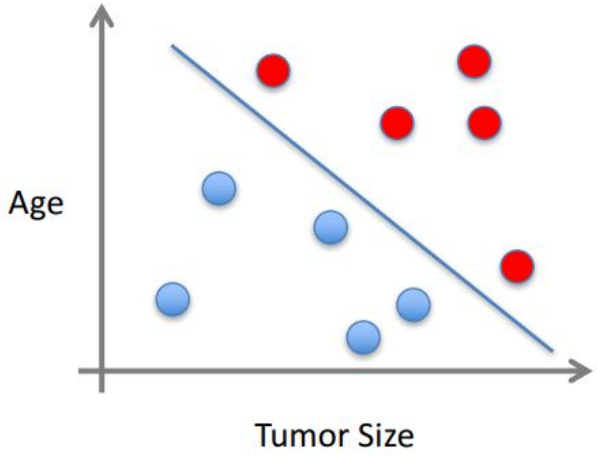
	label
	*apple*
	*apple*
	*banana*
	*banana*

Classification: a finite set of labels

Classification

- Given $\mathcal{T} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_m, y_m)\}$ set - training set
- Learn a function to predict y given $\mathbf{x}$
- $\mathbf{x}$ is generally multidimensional (multiple features) feratures =~ caratteristiche

$$f : \mathbb{R}^d \rightarrow \{1, 2, \ldots, k\}$$



Classification Applications

- Face recognition
- Character recognition
- Spam detection
- Medical diagnosis: from symptoms to illnesses
- Biometrics: Recognition/authentication using physical and/or behavioral characteristics: Face, iris, signature, etc

…

Supervised learning: regression

label

-4.5

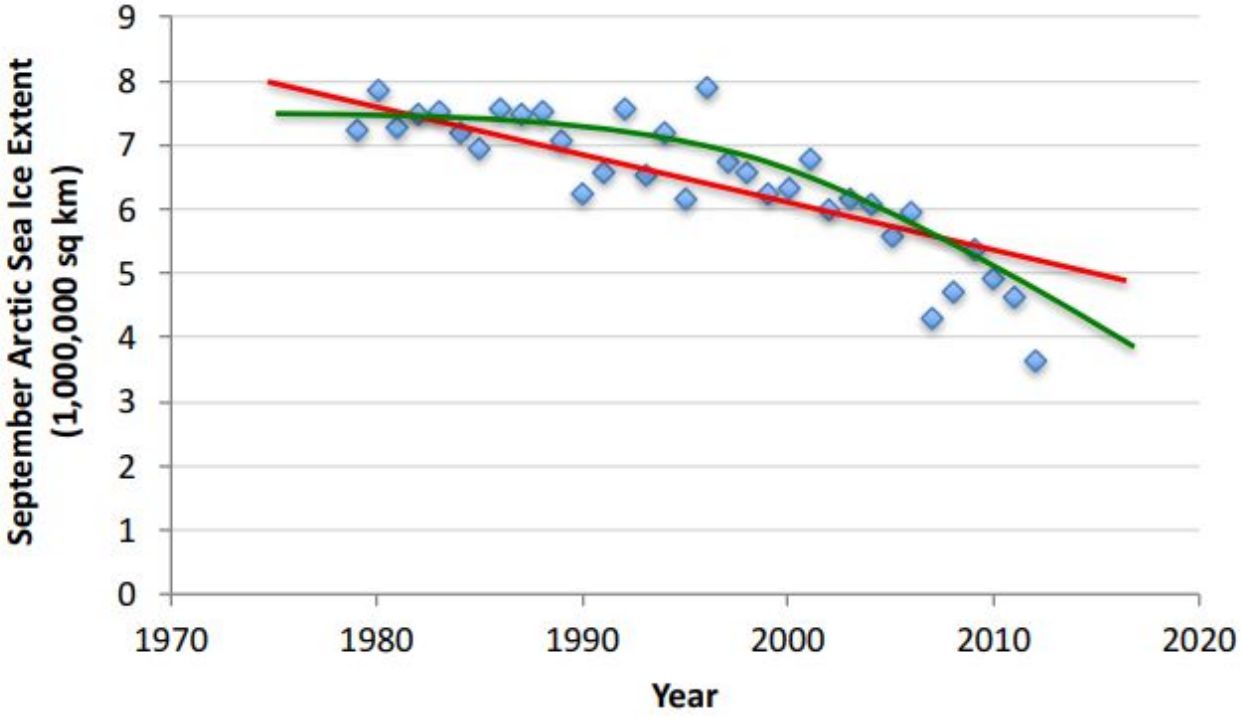
10.1

label is real-valued

3.2

4.3

Regression Example

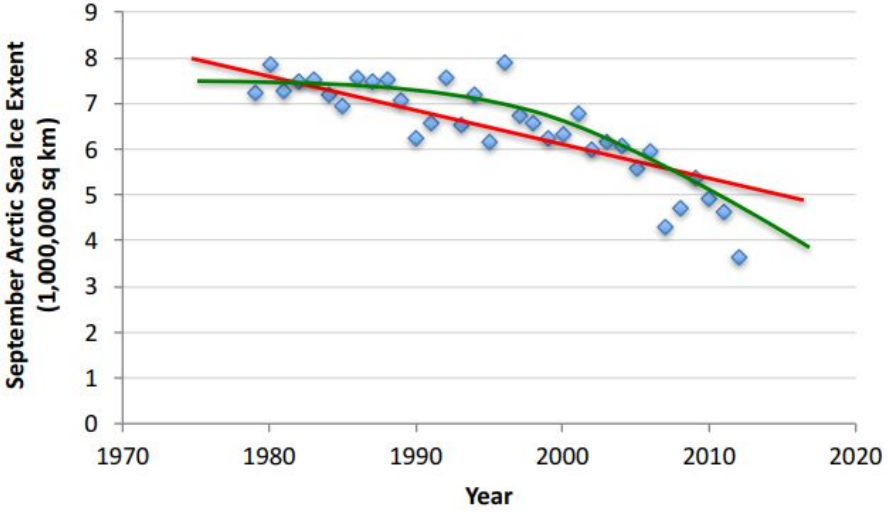


x : year y : ice-extent

Regression

- Given $\mathcal{T} = \{(x_1, y_1), \cdots, (x_m, y_m)\}$
- Learn a function to predict y given x
- y is real-valued, *d=1*.

$$f : \mathbb{R}^d \rightarrow \mathbb{R}$$



Regression Applications

- Economics/Finance: predict the value of a stock
- Car navigation: angle of the steering wheel, acceleration, …
- Temporal trends: weather over time

…

Supervised learning: ranking

la vediamo poco nel corso

	label
	1
	4
	2
	3

è supervised perchè impara dai ranking dati da umani

Ranking: label is a ranking

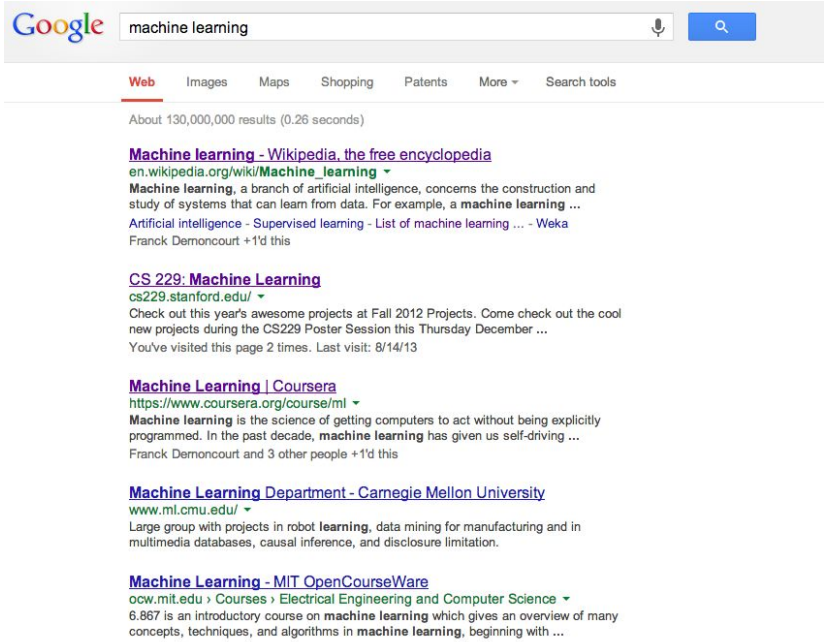
Ranking example

Given a query and

a set of web pages,

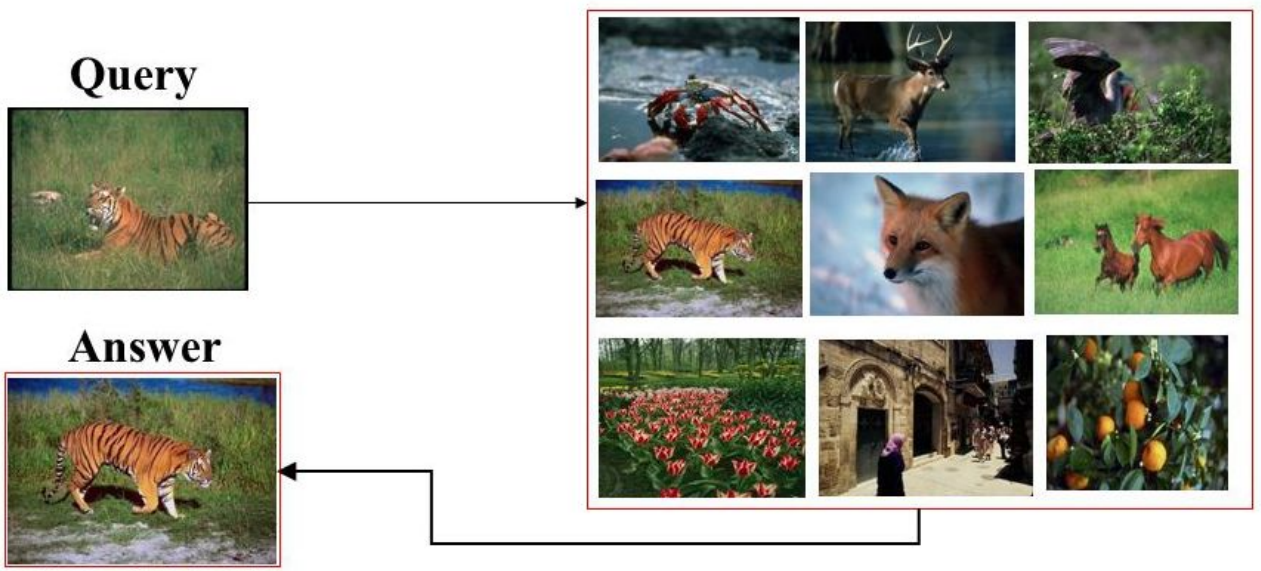
rank them according

to relevance



Ranking example

Given a query image, find the most visually similar images (with an order) in the database.



Ranking Applications

- User preference, e.g. Netflix movie ranking
- Image retrieval
- Flight search (search in general)

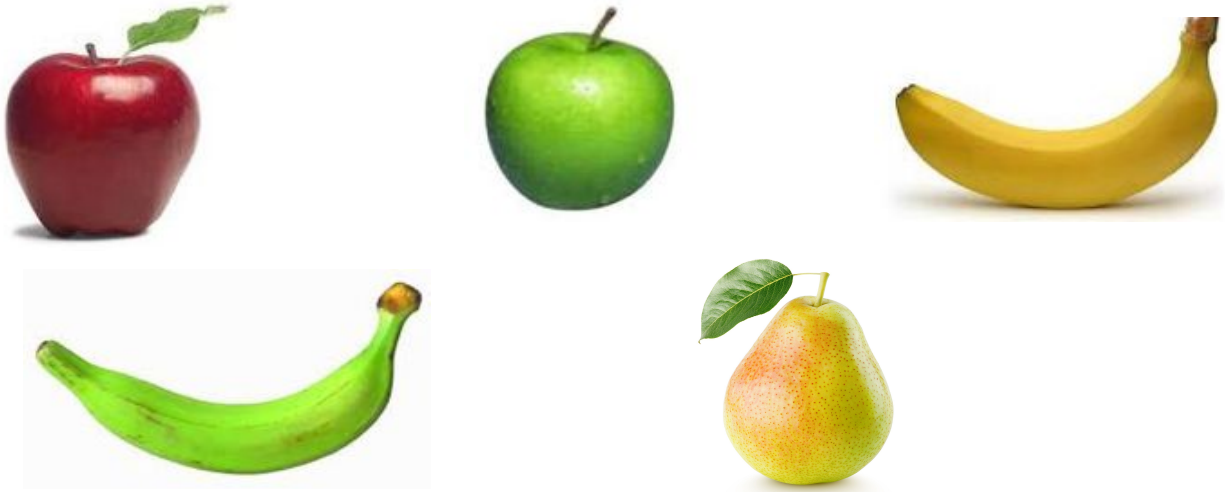
...

Supervised Learning Summary

- **Classification:**
  - Classification categorizes input data into predefined classes or categories.
- **Regression:**
  - Regression predicts continuous numerical values based on input features.
- **Ranking:**
  - Ranking orders a set of items based on their relevance or preference.

Unsupervised learning

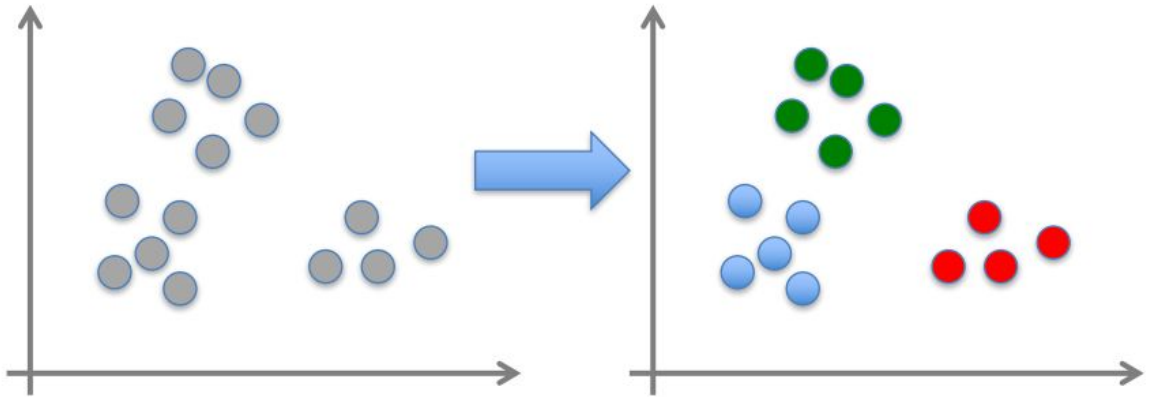
no training ?



Unsupervised learning: given data, i.e. examples, but no labels

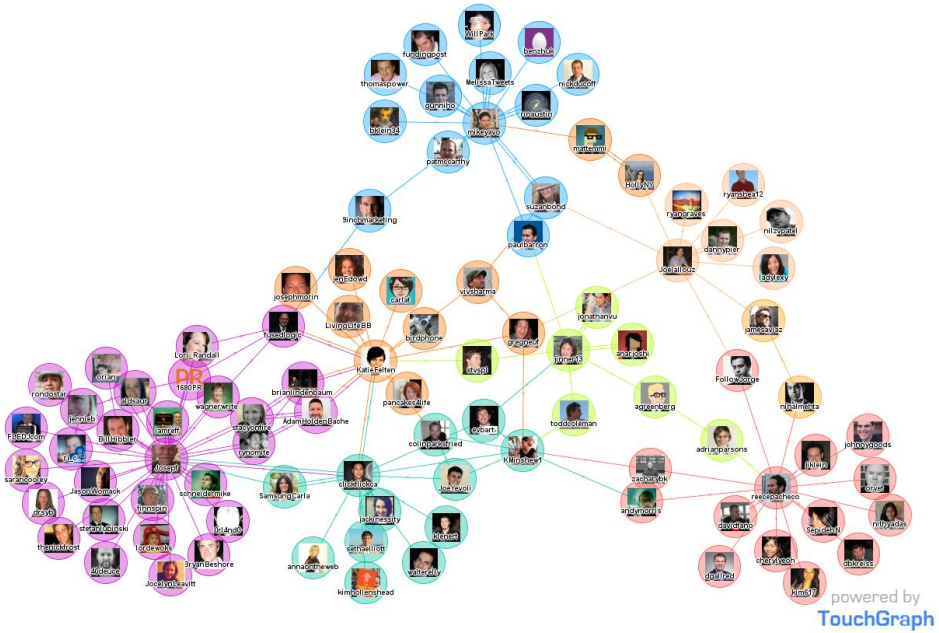
Clustering

Given $\mathcal{T} = \{\mathbf{x}_1, \cdots, \mathbf{x}_m\}$ (without labels) output hidden structure behind the x's, that is the clusters



Clustering applications

Social Network Analysis

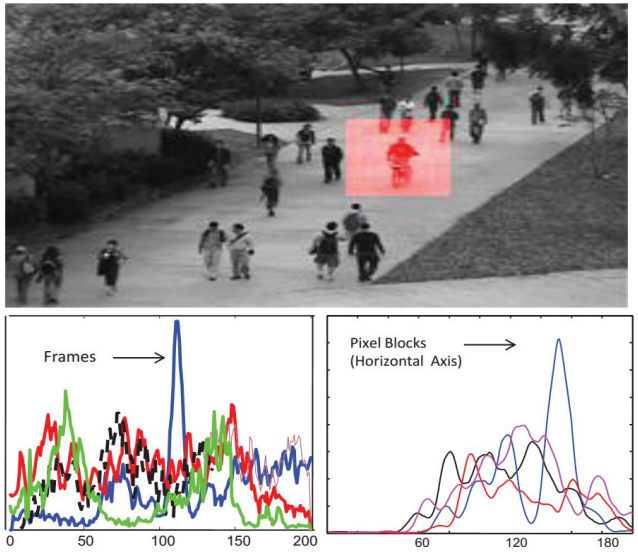


Clustering applications

Image segmentation

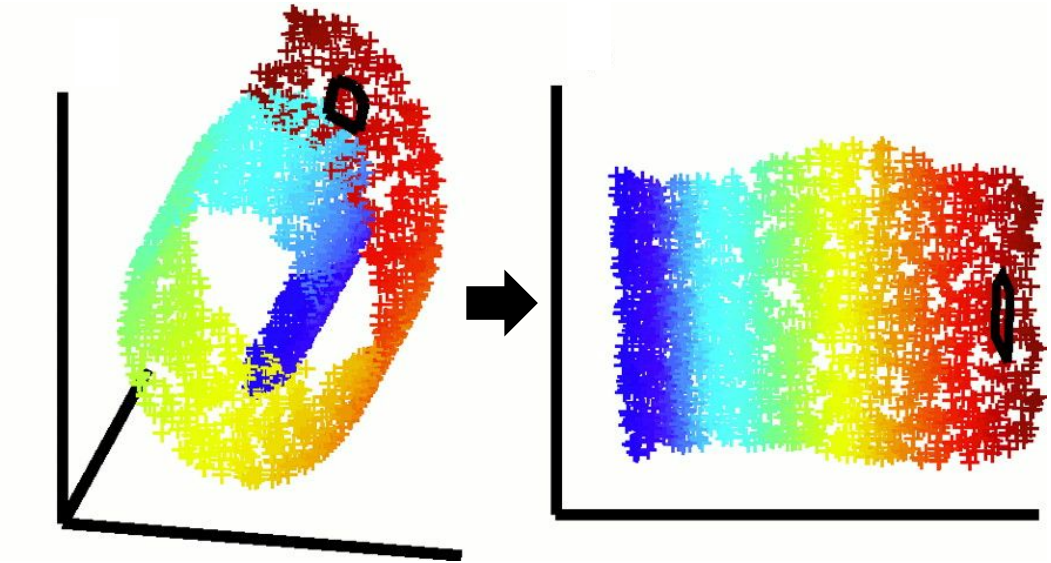
Anomaly detection

- Analyze a set of events or objects and flags some of them as being unusual or atypical.
- Example: credit card fraud detection, video surveillance.



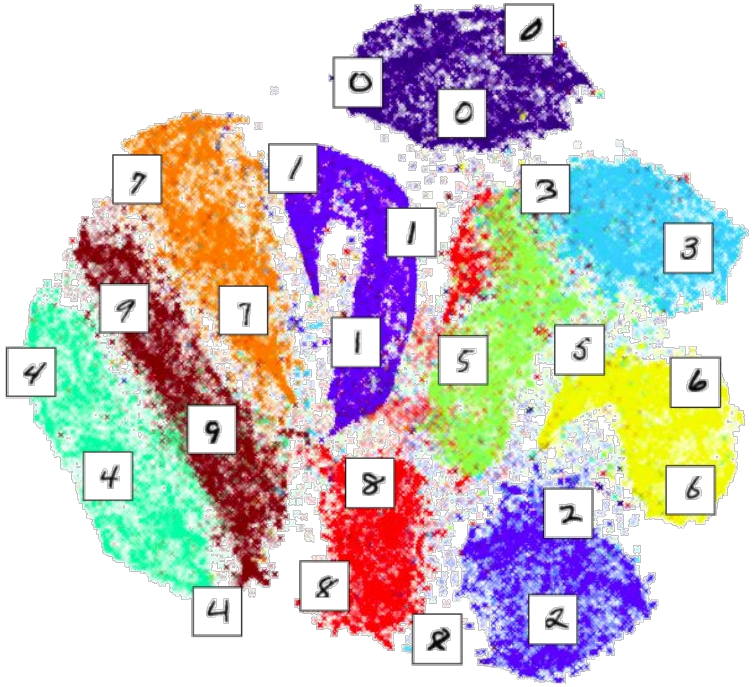
Dimensionality Reduction

Reduce the number of features under consideration by mapping data into another low dimensional space.



Dimensionality Reduction applications

Inspect your classification algorithm

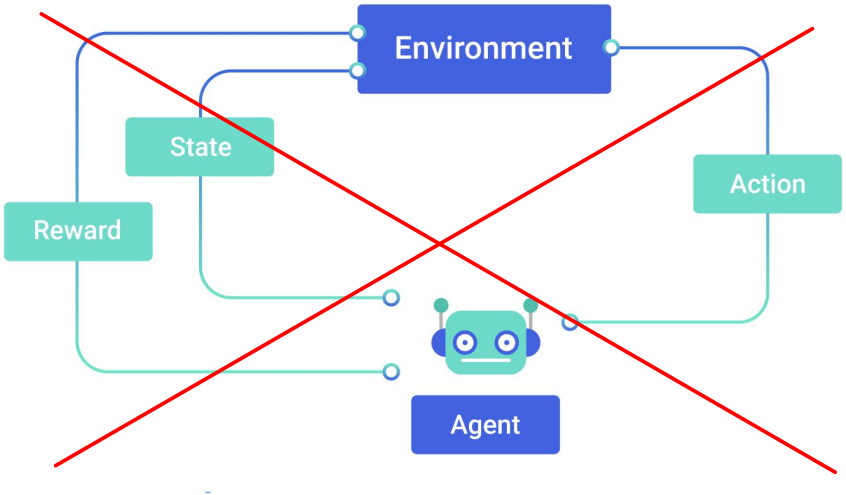


Unsupervised Learning Summary

- **Clustering:**
  - Clustering groups similar data points into clusters or groups based on their inherent characteristics or features.
- **Dimensionality Reduction:**
  - Dimensionality reduction reduces the number of input variables or features in a dataset while preserving the most important information.

Reinforcement learning

Idea: an **agent** learns from the **environment** by interacting with it and receiving **rewards** for performing **actions.**



BOH

Reinforcement learning example

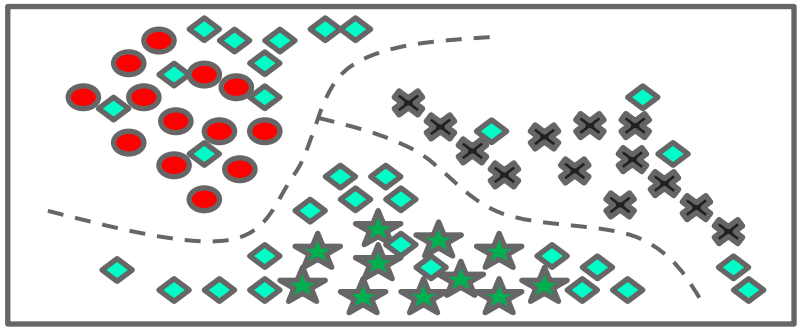
Backgammon

Given sequences of moves and whether or not the player won at the end, learn to make good moves

Other learning variations

non penso che sia da sapere
solo cultura personale

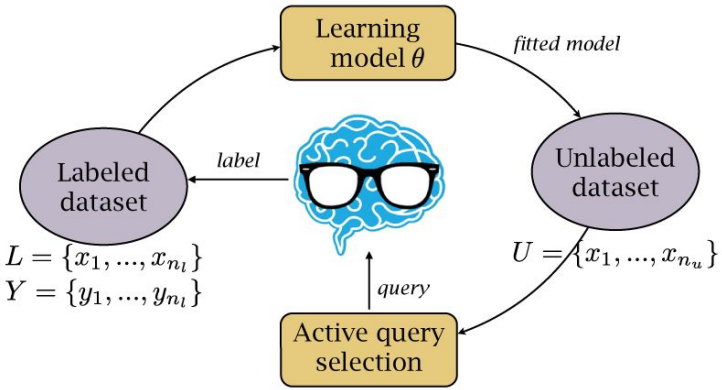
- What data is available:
  - We saw supervised, unsupervised, reinforcement learning
  - Other settings: **semi-supervised**, active learning, …



Semi-supervised learning

Other learning variations

- What data is available:
  - We saw supervised, unsupervised, reinforcement learning
  - Other settings: semi-supervised, **active learning**, …



Other learning variations

How are we getting the data: **online vs. offline (batch) learning**

The difference between online and batch (offline) learning lies in how the learning process occurs and **how the model is updated with new data.**

Other learning variations

Batch (Offline) Learning:

- The model learns from the entire dataset in one go and is updated only after processing all the data.
- Once the model is trained, it remains static and does not change unless retrained with the entire dataset.
- Typically used when the dataset fits into memory and can be processed efficiently as a whole.

Online Learning: stream di data

- The model learns incrementally from each new data point or small batches of data.
- Allows the model to adapt to changes in the data distribution over time.
- Suitable for scenarios where data arrives sequentially and needs to be processed in real-time or where computational resources are limited.

Back to Data and Features

examples

features

red, round, leaf, 3oz, …

green, round, no leaf, 4oz, …

yellow, curved, no leaf, 8oz, …

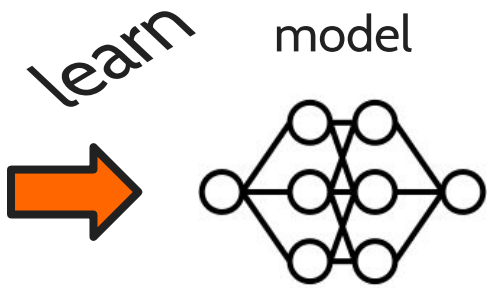
green, curved, no leaf, 7oz, …

How our algorithms actually "view" the data

Features are the questions we can ask about the examples

Classification revisited

examples	label
red, round, leaf, 3oz, …	*apple*
green, round, no leaf, 4oz, …	*apple*
yellow, curved, no leaf, 8oz, …	*banana*
green, curved, no leaf, 7oz, …	*banana*

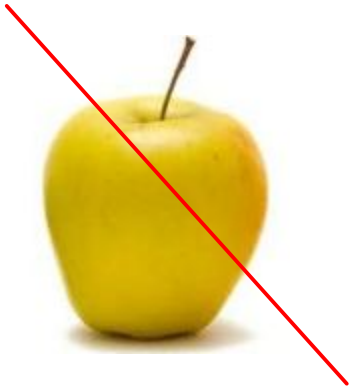


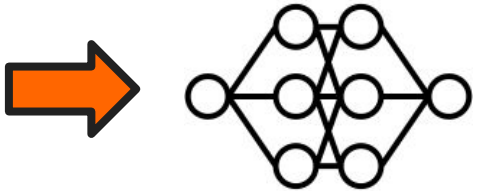
During training, learn a model of what distinguishes apples and bananas ***based on the features***

il modello vede solo le features non più i raw data

Supervised learning

test example





predict

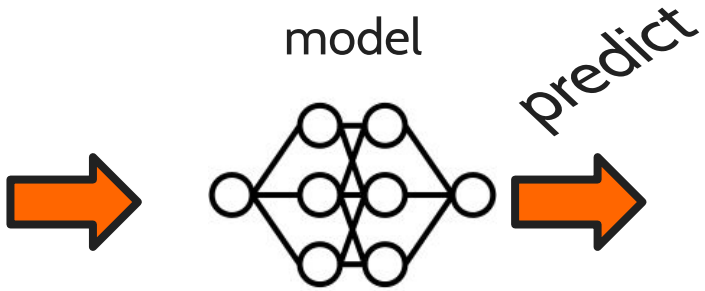
Apple or banana?

Classification revisited

The new example is described ***by the features***

il modello vede le cose attraverso features

red, round, no leaf, 4oz, …

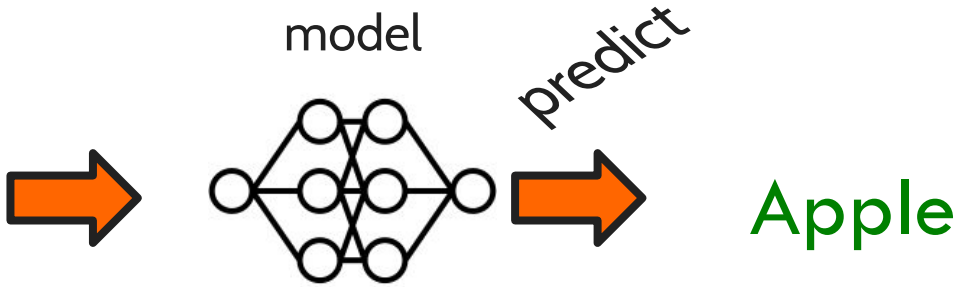


Apple or banana?

Classification revisited

The model can then classify a new example ***based on the features***

red, round, no leaf, 4oz, …



Apple

Why? Come ?

Classification revisited

Training data: examples	label
red, round, leaf, 3oz, …	*apple*
green, round, no leaf, 4oz, …	*apple*
yellow, curved, no leaf, 4oz, …	*banana*
green, curved, no leaf, 5oz, …	*banana*

Test set

come facciamo a capire che questa sia una mela ?

red, round, no leaf, 4oz, … ?

Learning is about ***generalizing*** from the training data

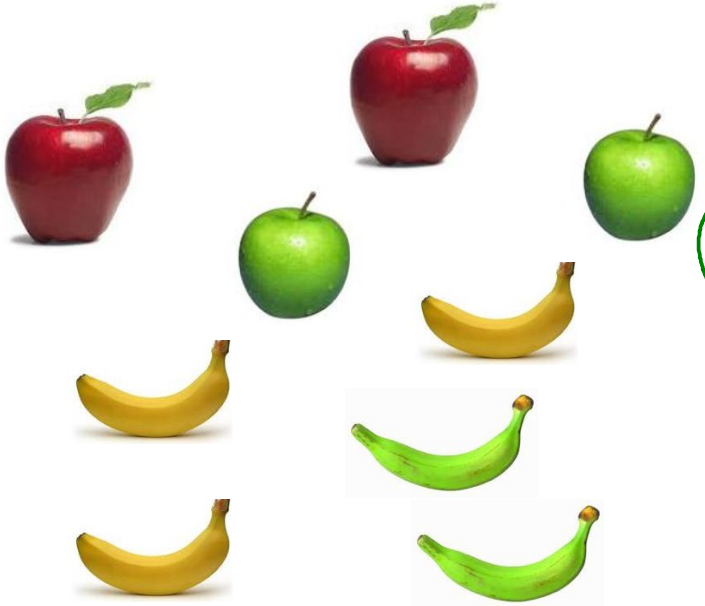
Generalization

- **Generalization** in machine learning refers to the ability of a trained model to perform well on new, unseen data that was not used during the training process.

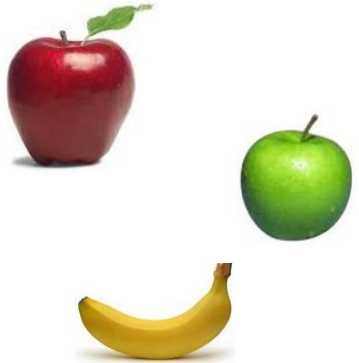
- A model generalizes well when it can accurately make predictions or produce outputs for data it has not encountered before.

Training & Test Set

Training data

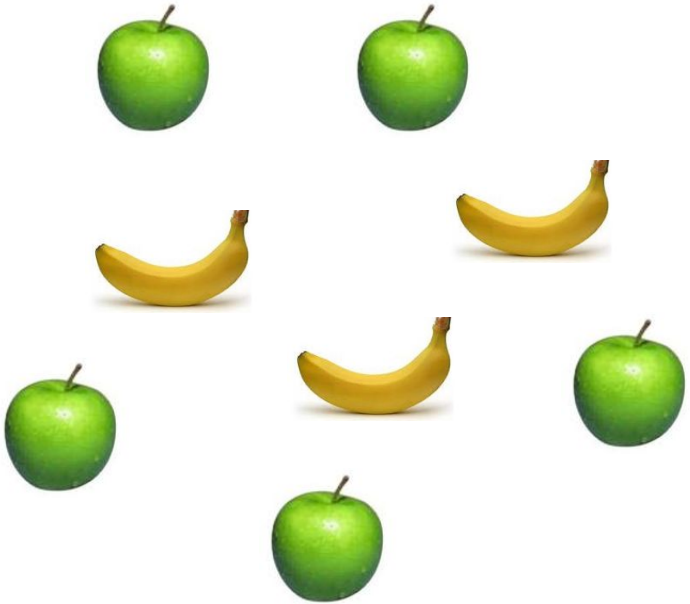


Test set

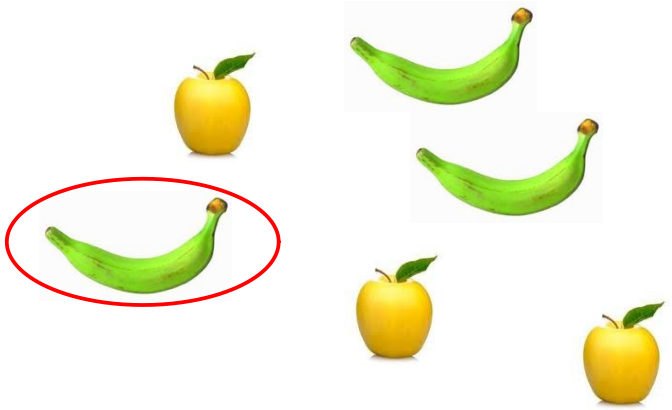


Training & Test Set

Training data



Test set



Not always the case, then learning is not possible!

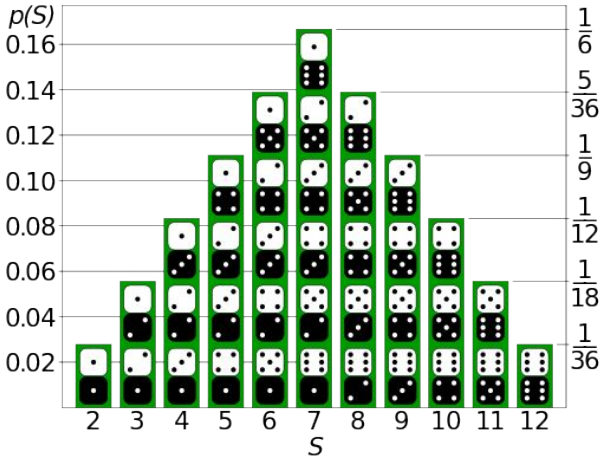
More technically...

- We are going to use the ***probabilistic model*** of learning
- There is some probability distribution over example/label pairs called the ***data generating distribution***
- **Both** the training data **and** the test set are generated based on this distribution

What is a probability distribution?

Probability distribution

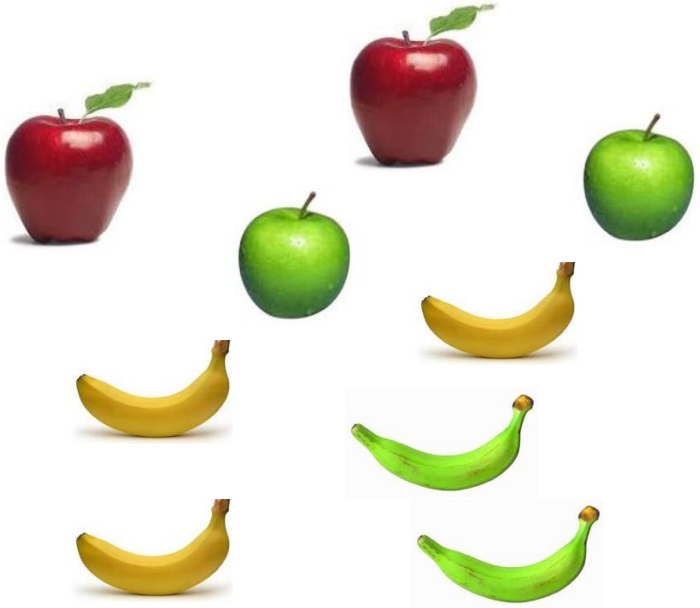
Describes how likely (i.e. probable) certain events are



- Considers probabilities for all possible events
- Probabilities are between 0 and 1 (inclusive)
- Sum of probabilities over all events is 1

Probability distribution

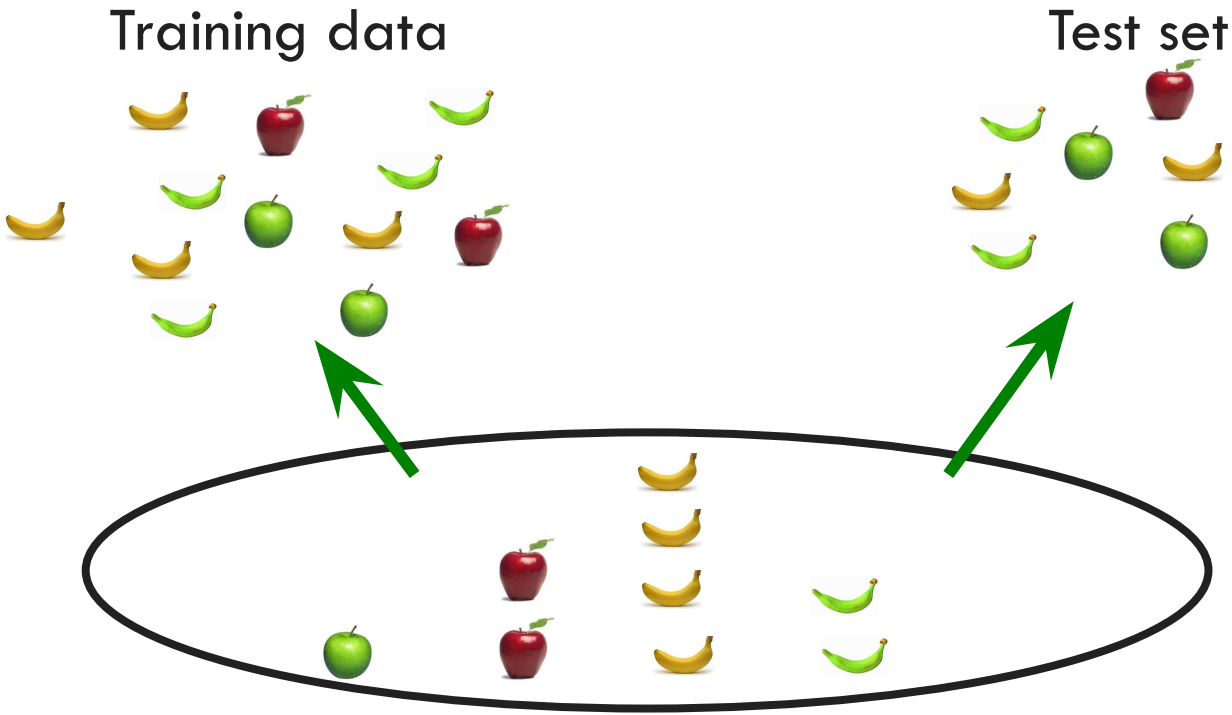
Training data



High probability	Low probability
round apples	curved apples
curved bananas	red bananas
apples with leaves	yellow apples
…	…

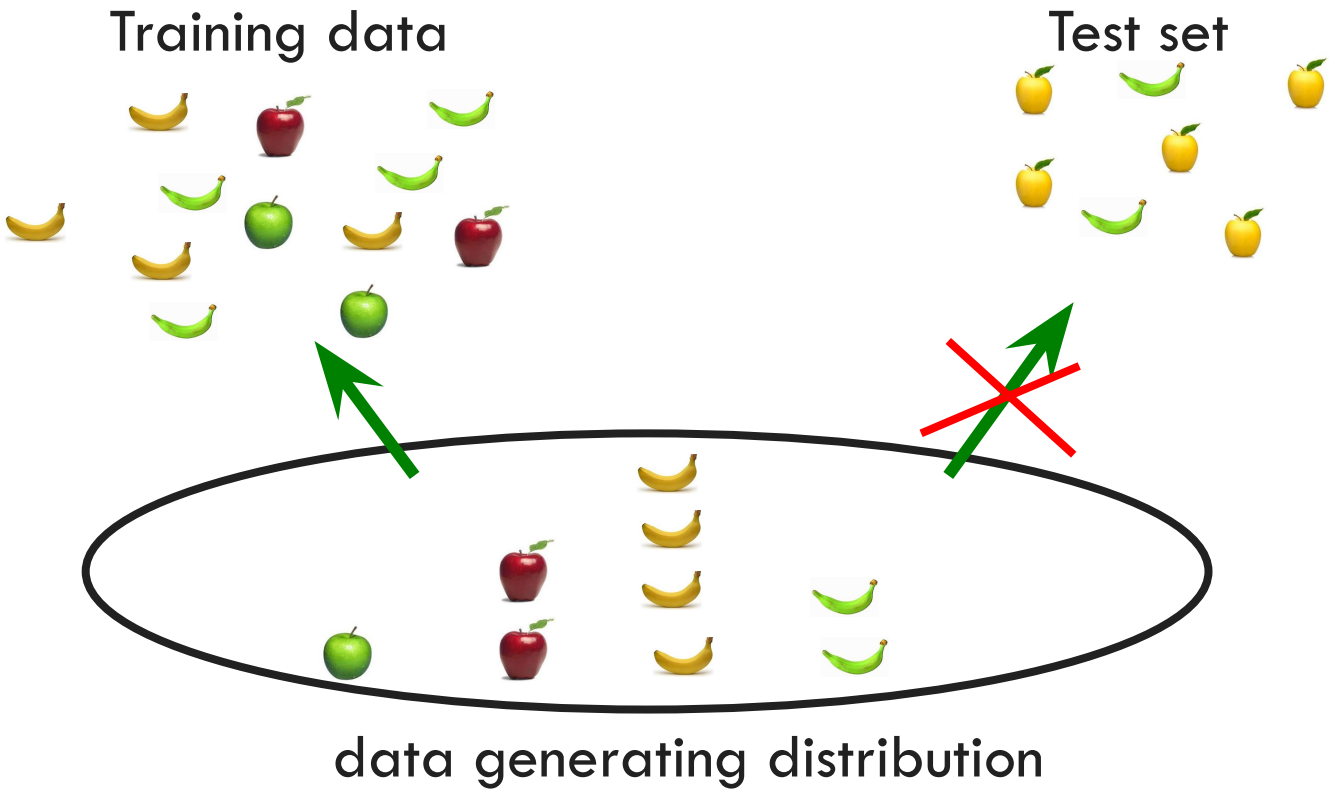
DATA GENERATING DISTRIBUTION

funziona perchè otteniamo entrambi i set dalla stessa distribuzione



Non sappiamo quale sia la vera distribuzione. Abbiamo solo un proxy.

DATA GENERATING DISTRIBUTION



DATA GENERATING DISTRIBUTION

A **data generating distribution** refers to the underlying probability distribution that generates the observed data points in a dataset.

The data generating distribution captures the inherent patterns, structure, and noise present in the data.

Understanding this distribution is crucial for building accurate and generalizable machine learning models because **it enables us to make informed assumptions about the data and to make predictions** or decisions based on probabilistic reasoning.

Summary

- The learning process
- Data, features, models
- Supervised, unsupervised and reinforcement learning
- Training and test data, generalization, data generating distribution

QUESTIONS?