Análise de Dados em Informática

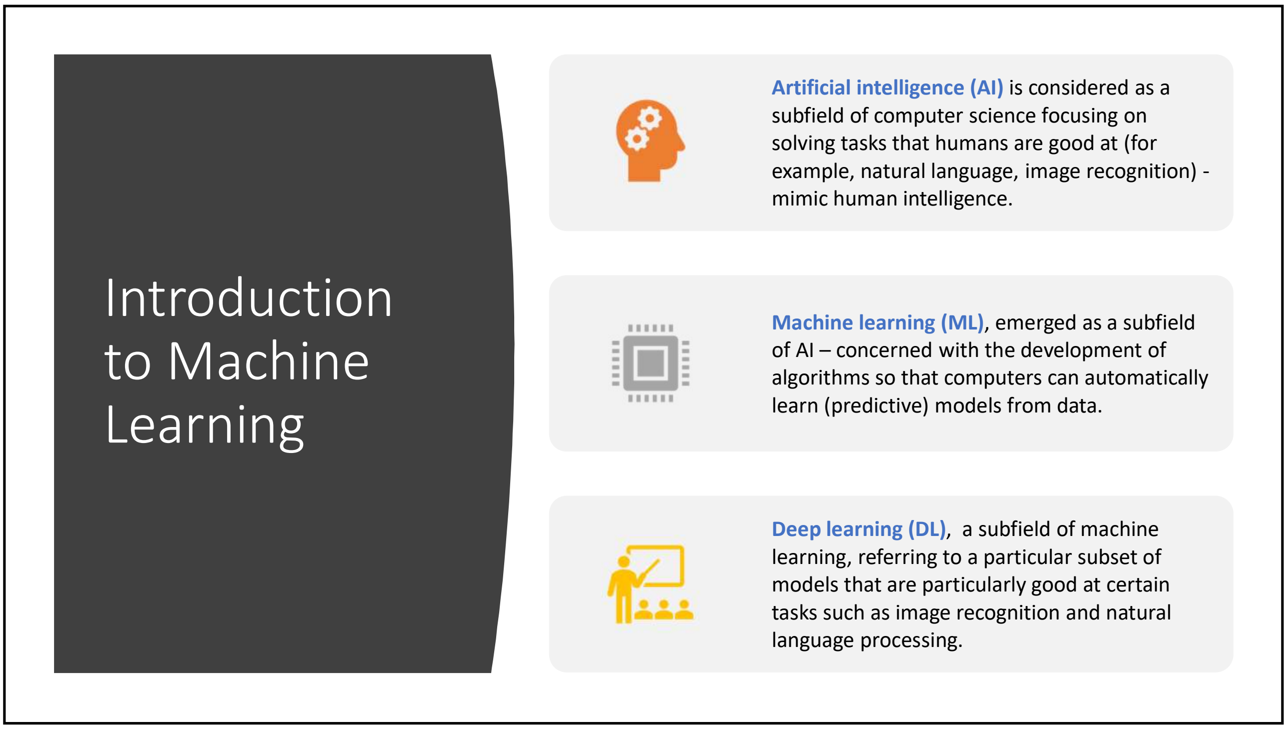
Licenciatura em Engenharia Informática
ISEP/IPP

Ana Maria Madureira
2023/2024

Introduction to Machine Learning

Fontes:

- Tom Mitchell, Machine Learning. McGraw-Hill, 1997.
- Christopher Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
- Sebastian Raschka, STAT479 FS18. L01: Intro to Machine Learning, Fall 2018



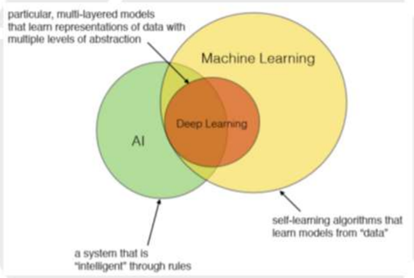
Introduction to Machine Learning

Artificial intelligence (AI) is considered as a subfield of computer science focusing on solving tasks that humans are good at (for example, natural language, image recognition) - mimic human intelligence.

Machine learning (ML), emerged as a subfield of AI – concerned with the development of algorithms so that computers can automatically learn (predictive) models from data.

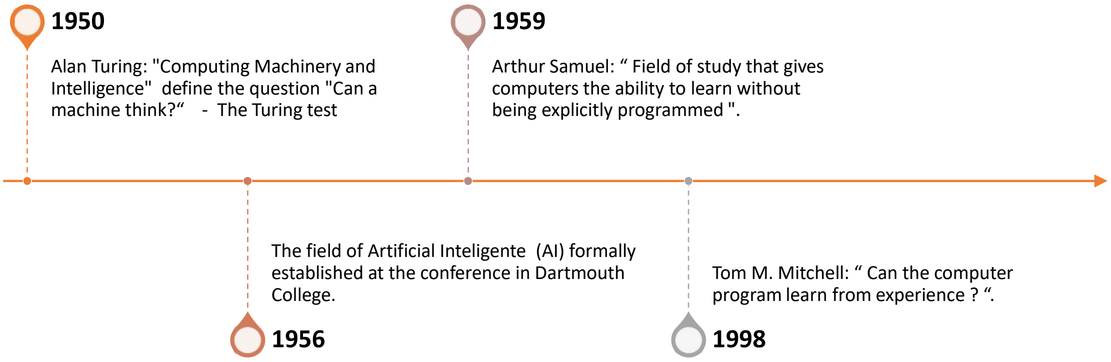
Deep learning (DL), a subfield of machine learning, referring to a particular subset of models that are particularly good at certain tasks such as image recognition and natural language processing.

Relationship between Machine Learning, Deep Learning, and Artificial Intelligence


particular, multi-layered models that learn representations of data with multiple levels of abstraction
Machine Learning
Deep Learning
AI
self-learning algorithms that learn models from "data"
a system that is "intelligent" through rules

3

Machine Learning Historic Perspective



1950

Alan Turing: "Computing Machinery and Intelligence" define the question "Can a machine think?" - The Turing test

1956

The field of Artificial Inteligente (AI) formally established at the conference in Dartmouth College.

1959

Arthur Samuel: " Field of study that gives computers the ability to learn without being explicitly programmed ".

1998

Tom M. Mitchell: " Can the computer program learn from experience ? ".

4

What is Machine Learning? An Overview

- “Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.”- Arthur L. Samuel, AI pioneer, 1959
- The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience. — Tom Mitchell, Professor Machine Learning at Carnegie Mellon University and author of the popular “Machine Learning” textbook
- Machine learning is the hot new thing. — John L. Hennessy, President of Stanford (2000–2016)
- A breakthrough in machine learning would be worth ten Microsofts. — Bill Gates, Microsoft Co-Founder

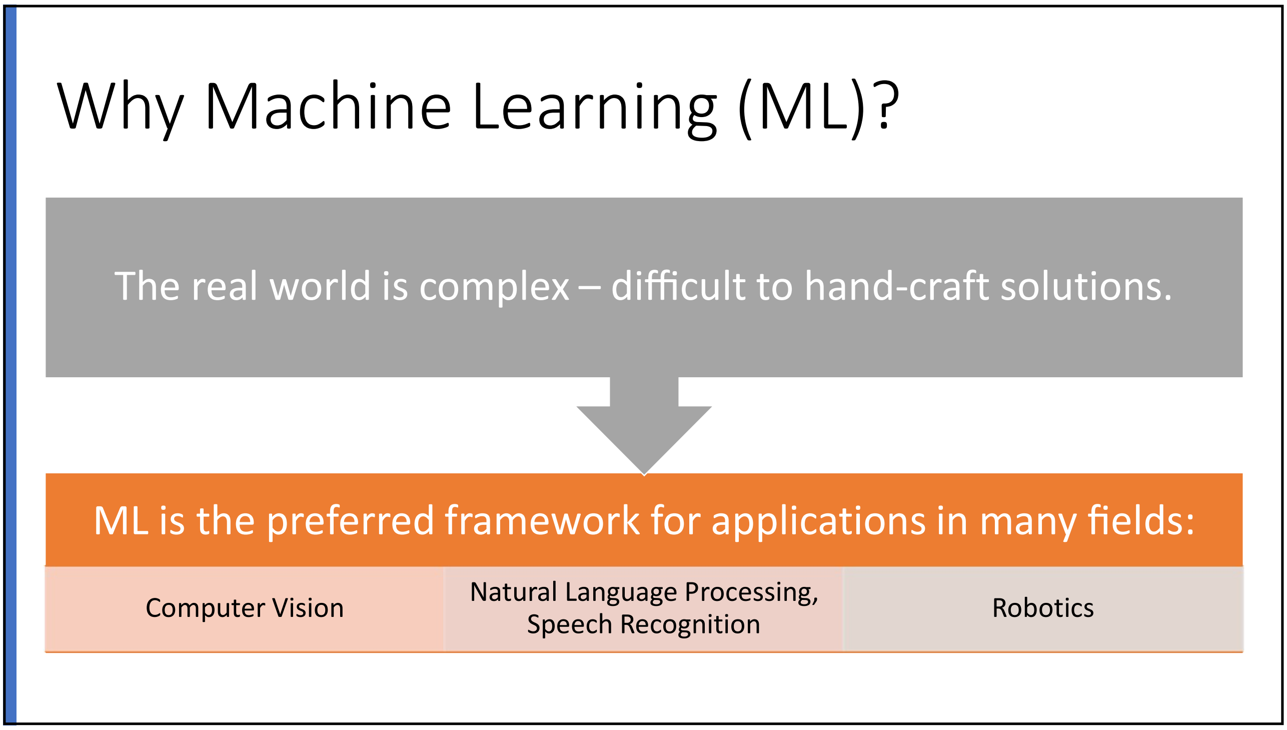
5

What is Machine Learning (ML)?

ALGORITHMS THAT AUTOMATICALLY IMPROVE PERFORMANCE THROUGH EXPERIENCE

OFTEN THIS MEANS DEFINE A MODEL BY HAND, AND USE DATA TO FIT ITS PARAMETERS

6



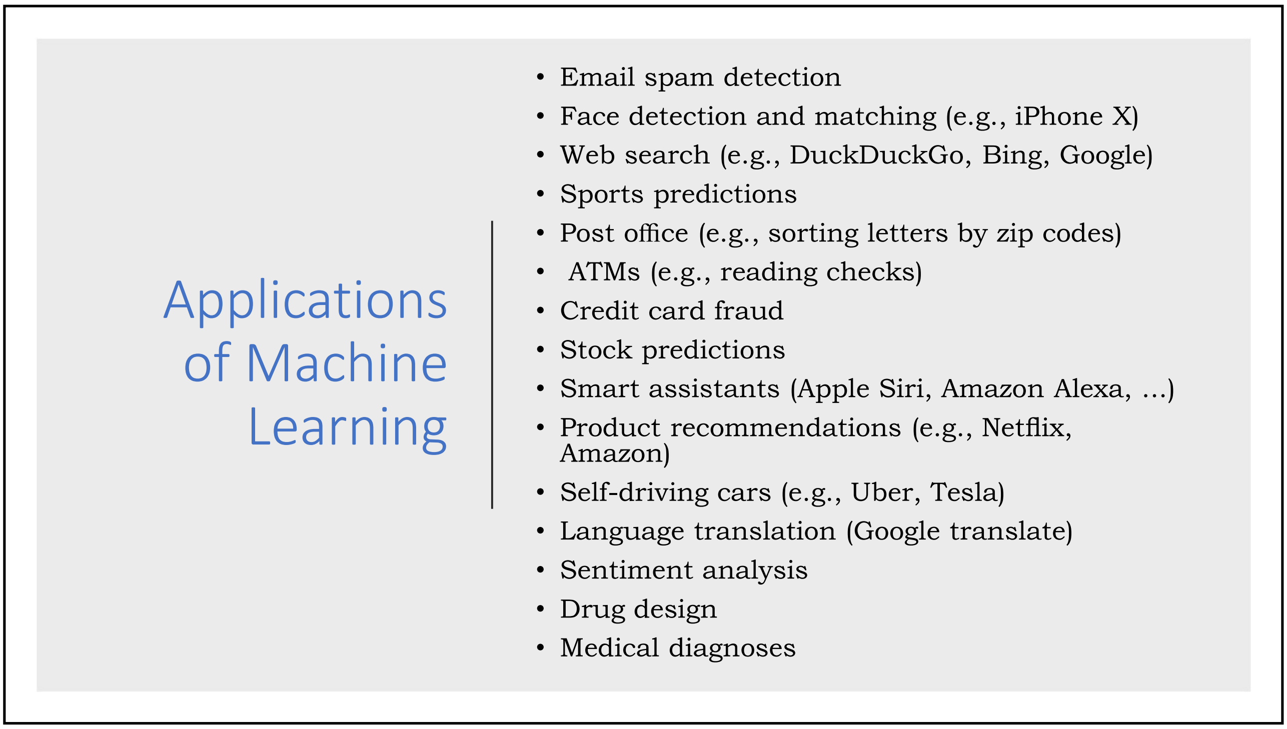
Why Machine Learning (ML)?

The real world is complex – difficult to hand-craft solutions.

ML is the preferred framework for applications in many fields:

Computer Vision	Natural Language Processing, Speech Recognition	Robotics

7



Applications of Machine Learning

- Email spam detection
- Face detection and matching (e.g., iPhone X)
- Web search (e.g., DuckDuckGo, Bing, Google)
- Sports predictions
- Post office (e.g., sorting letters by zip codes)
- ATMs (e.g., reading checks)
- Credit card fraud
- Stock predictions
- Smart assistants (Apple Siri, Amazon Alexa, …)
- Product recommendations (e.g., Netflix, Amazon)
- Self-driving cars (e.g., Uber, Tesla)
- Language translation (Google translate)
- Sentiment analysis
- Drug design
- Medical diagnoses

8

The availability of massive amounts of good data and fast computation changes everything.....

New data is complex and unstructured.

Structured data

a flat file with a fixed number of measurements. Eg. response of a patient to a drug, and 10 measurements— age, sex (not gender), lab measurements.

Unstructured Data

doctor's notes, Twitter feeds, broker

9

Data Types

Structured

- Examples: Database
- Fixed struct
- Each field has a format well defined
- The format is a standard accepted by the field

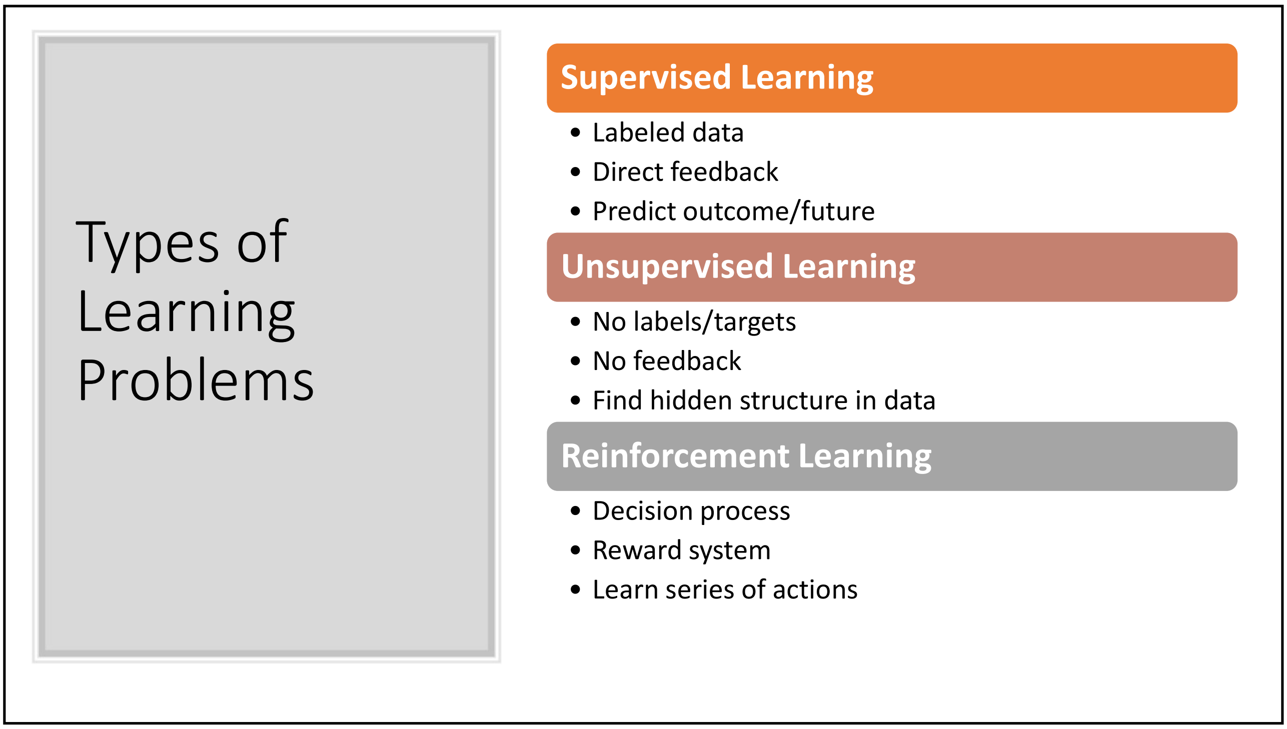
Semi-Structured

- Examples:XML, JSON, RDF, OWL
- Flexible Structure
- Each field has a structure but no format imposition

Non-Structured

- Examples: Text, documents, images, video, audio, social networks
- No structure
- Represents more than 80% of global data

10



Types of Learning Problems

Supervised Learning

- Labeled data
- Direct feedback
- Predict outcome/future

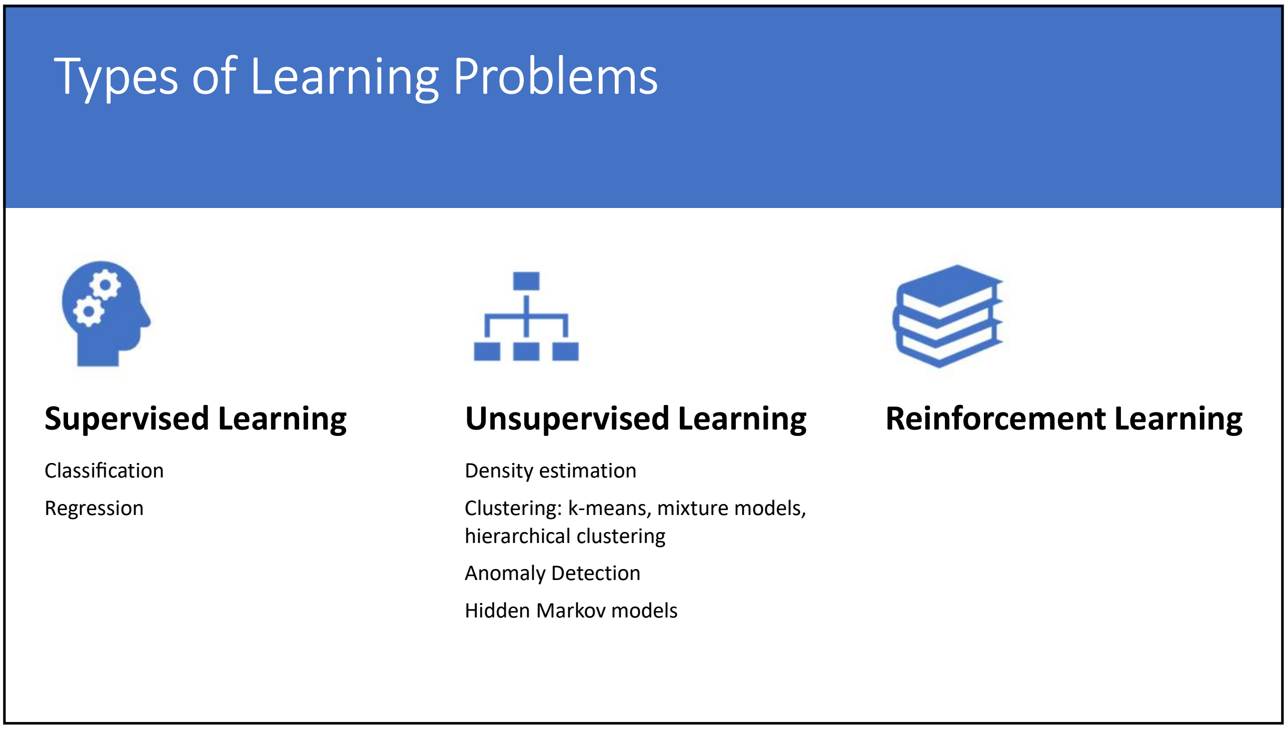
Unsupervised Learning

- No labels/targets
- No feedback
- Find hidden structure in data

Reinforcement Learning

- Decision process
- Reward system
- Learn series of actions

11



Types of Learning Problems

Supervised Learning

Classification

Regression

Unsupervised Learning

Density estimation

Clustering: k-means, mixture models, hierarchical clustering

Anomaly Detection

Hidden Markov models

Reinforcement Learning

12

Frameworks de Data Science

- **SEMMA -** The SAS Institute developed SEMMA as the process of data mining. It has five steps (Sample, Explore, Modify, Model, and Assess), earning the acronym of SEMMA. You can use the SEMMA data mining methodology to solve a wide range of business problems, including fraud identification, customer retention and turnover, database marketing, customer loyalty, bankruptcy forecasting, market segmentation, as well as risk, affinity, and portfolio analysis.

- **Knowledge Discovery in Database (KDD) -** back to 1989, KDD represents the overall process of collecting data and methodically refining it. The KDD Process is a classic data science life cycle that aspires to purge the 'noise' (useless, tangential outliers) while establishing a phased approach to derive patterns and trends that add important knowledge.

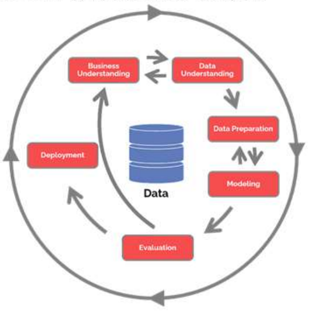
- **CRoss Industry Standard Process for Data Mining (CRISP-DM)** is a process model that serves as the base for a data science process.

13

What is CRISP DM?

The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (*CRISP-DM*) is a process model that serves as the base for a data science process.

1. Business understanding – What does the business need?
2. Data understanding – What data do we have / need? Is it clean?
3. Data preparation – How do we organize the data for modeling?
4. Modeling – What modeling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How do stakeholders access the results?



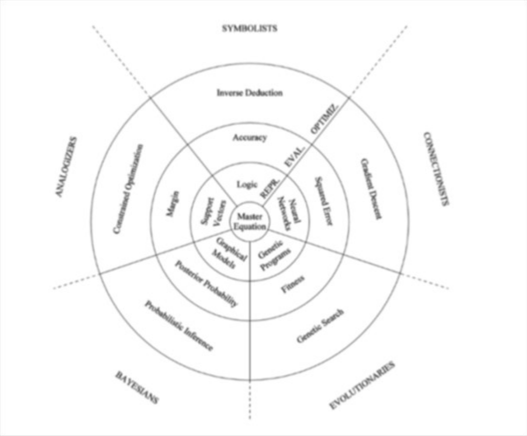
Published in 1999 to standardize data mining processes across industries, it has since become the most common methodology for data mining, analytics, and data science projects.

Data science teams that combine a loose implementation of CRISP-DM with overarching team-based agile project management approaches will likely see the best results.

14

Components of Machine Learning Algorithms

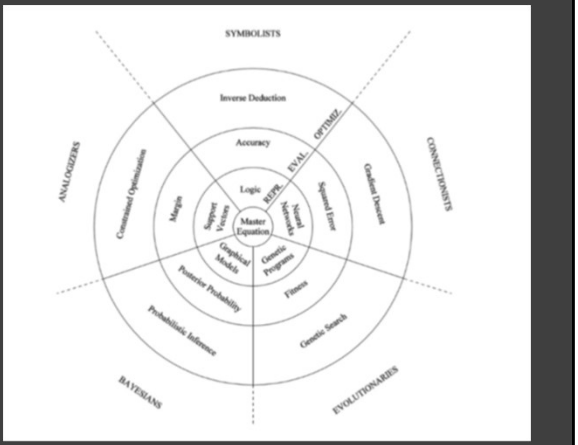
- **Representation** - which hypotheses we can represent given a certain algorithm class.
- **Evaluation** - the evaluation component is the step where we evaluate the performance of the model after model fitting.
- **Optimization** - the second component is the optimization metric that we use to fit the model.



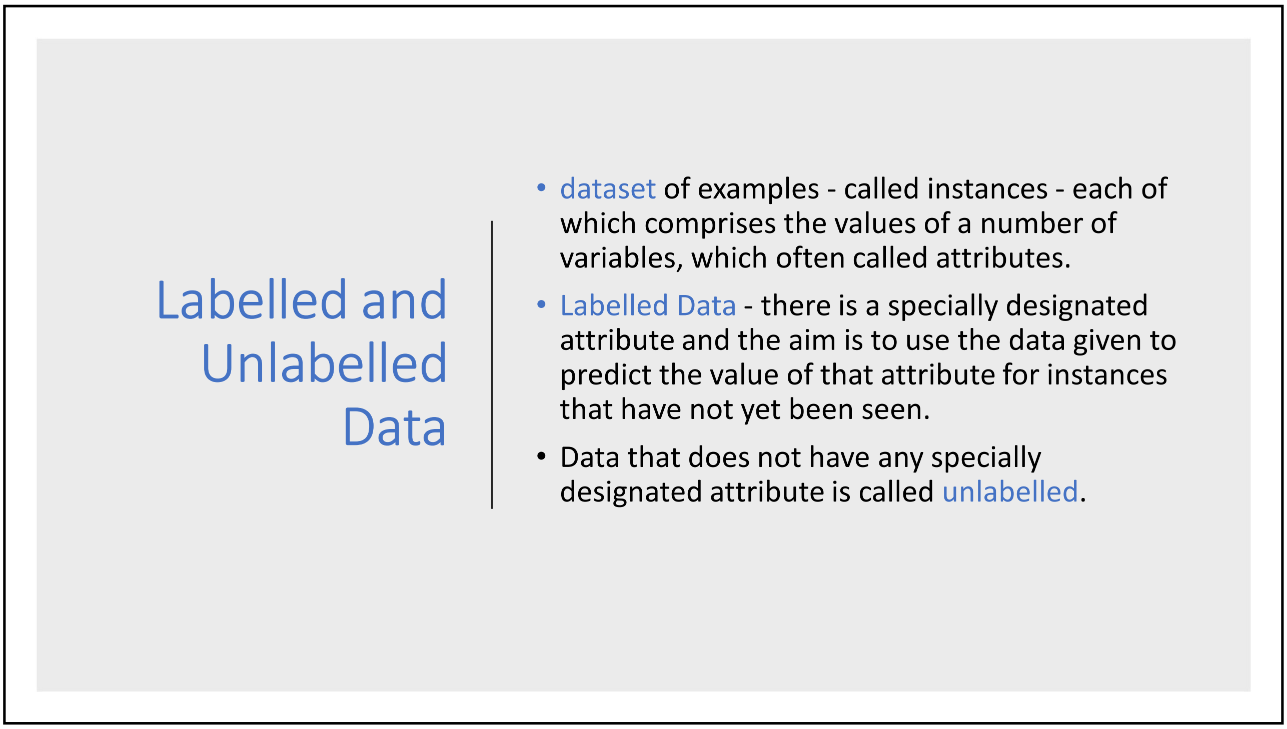
15

Machine learning application developing

- Define the problem to be solved.
- Collect (labeled) data.
- Choose an algorithm class.
- Choose an optimization metric for learning the model.
- Choose a metric for evaluating the model.



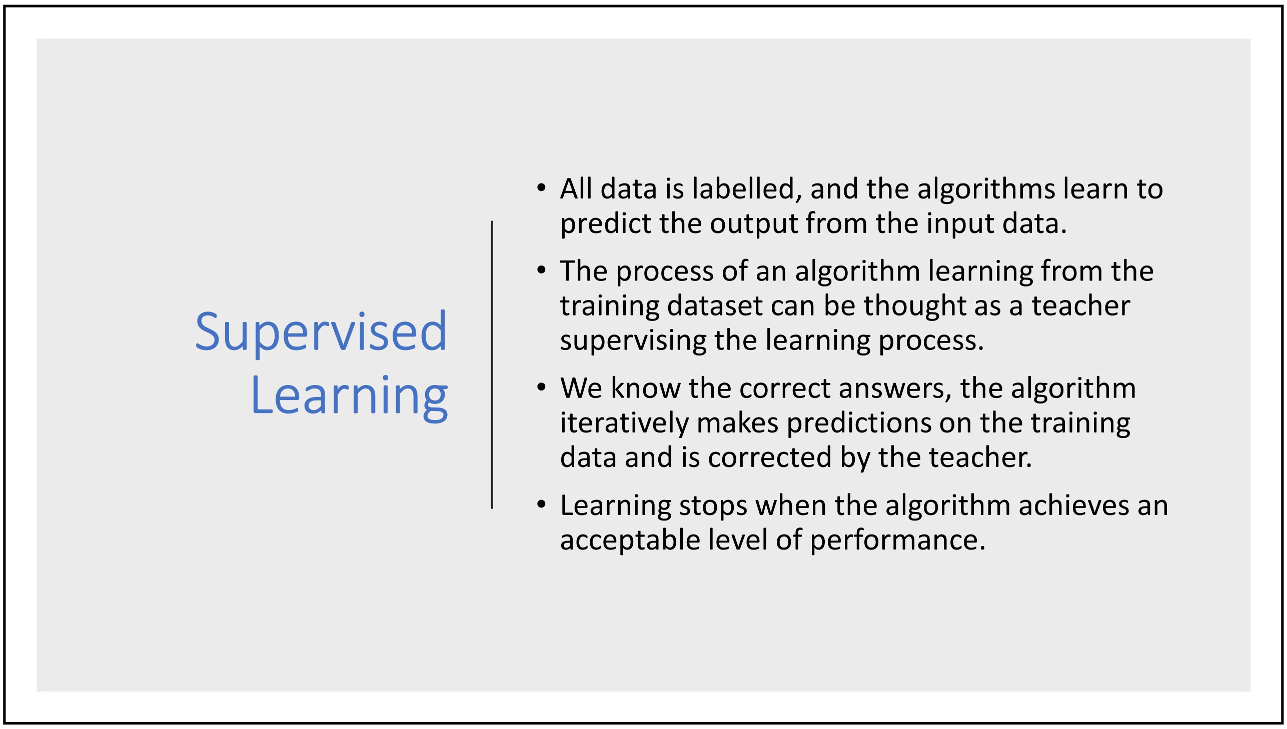
16



Labelled and Unlabelled Data

- dataset of examples - called instances - each of which comprises the values of a number of variables, which often called attributes.
- Labelled Data - there is a specially designated attribute and the aim is to use the data given to predict the value of that attribute for instances that have not yet been seen.
- Data that does not have any specially designated attribute is called unlabelled.

17



Supervised Learning

- All data is labelled, and the algorithms learn to predict the output from the input data.
- The process of an algorithm learning from the training dataset can be thought as a teacher supervising the learning process.
- We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher.
- Learning stops when the algorithm achieves an acceptable level of performance.

18

What Is an Model in Machine Learning?

- A “model” in ML is the output of a machine learning algorithm run on data.
- A model represents what was learned by a machine learning algorithm.
- The model is the “thing” that is saved after running a machine learning algorithm on training data and represents the rules, numbers, and any other algorithm-specific data structures required to make predictions.
- Some examples :
    - The linear regression algorithm results in a model comprised of a vector of coefficients with specific values.
    - The decision tree algorithm results in a model comprised of a tree of if-then statements with specific values.
    - The neural network / backpropagation / gradient descent algorithms together result in a model comprised of a graph structure with vectors or matrices of weights with specific values.

19

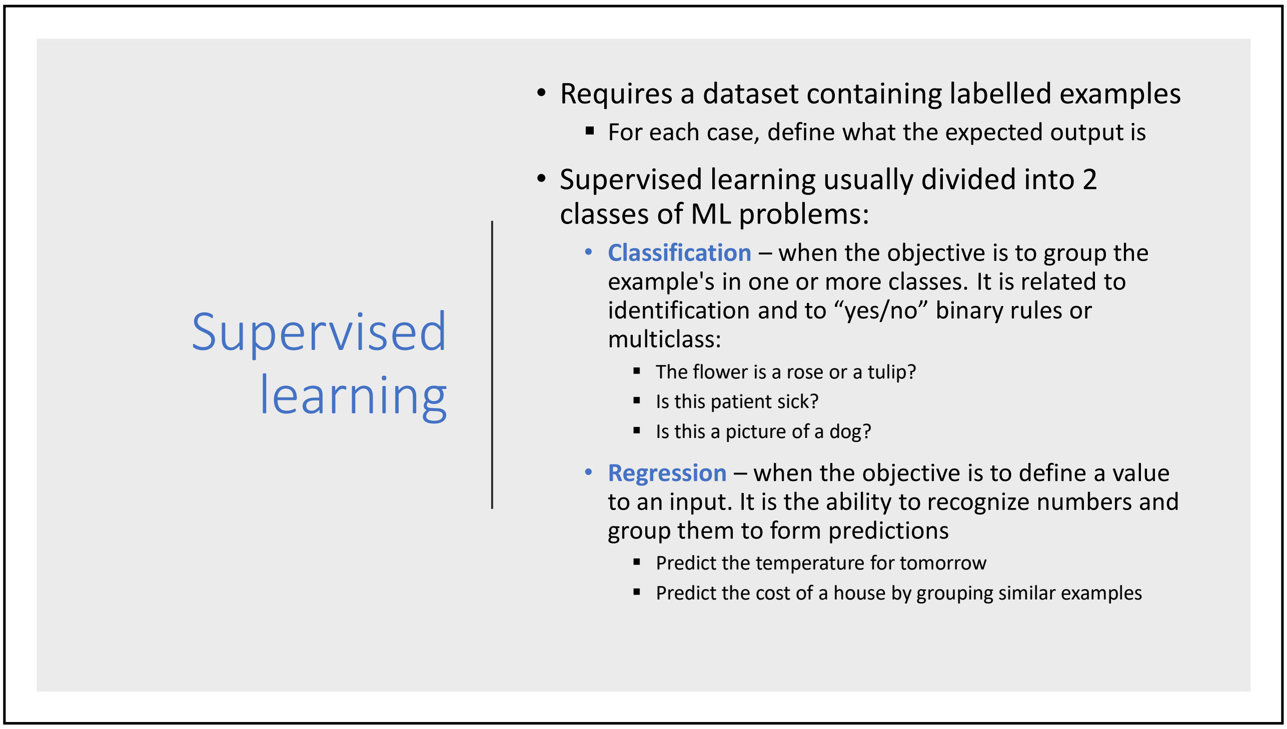
Machine Learning

Machine Learning is the science of programming computers so they can learn from data.

Machine learning can be divided in three areas:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

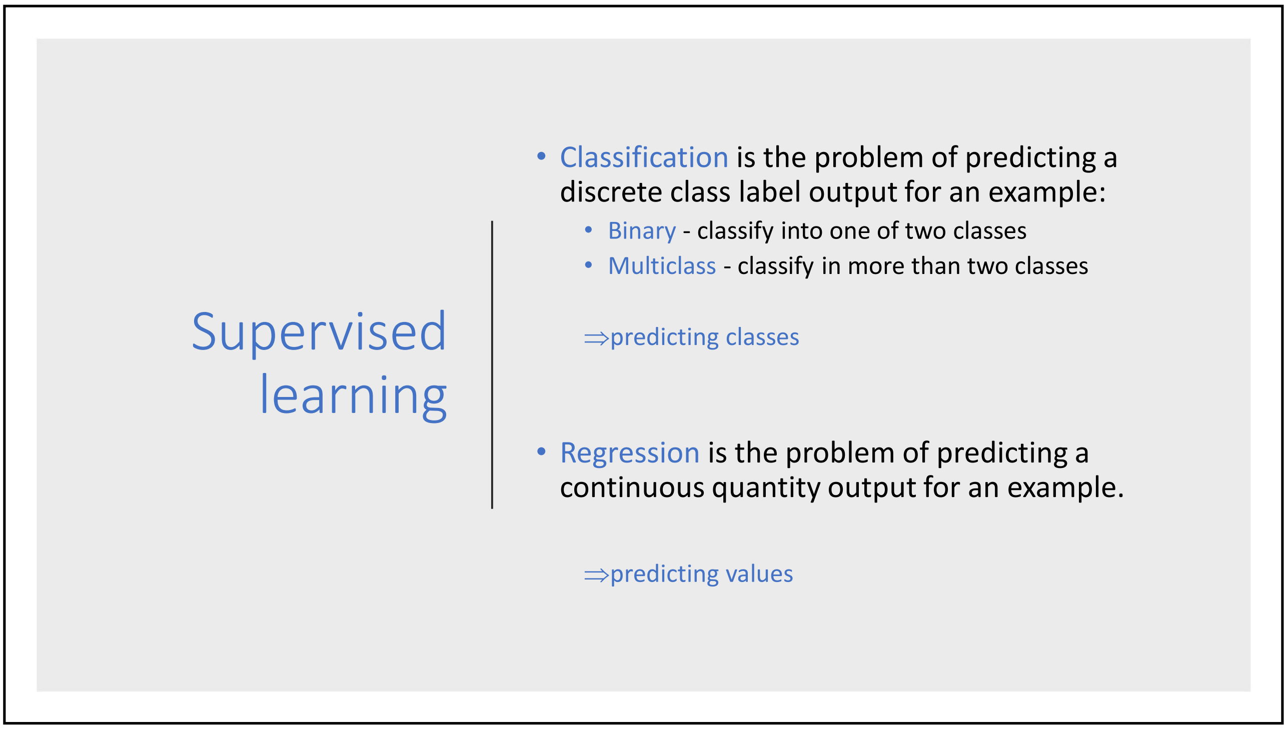
20



Supervised learning

- Requires a dataset containing labelled examples
  - For each case, define what the expected output is
- Supervised learning usually divided into 2 classes of ML problems:
  - **Classification** – when the objective is to group the example's in one or more classes. It is related to identification and to “yes/no” binary rules or multiclass:
    - The flower is a rose or a tulip?
    - Is this patient sick?
    - Is this a picture of a dog?
  - **Regression** – when the objective is to define a value to an input. It is the ability to recognize numbers and group them to form predictions
    - Predict the temperature for tomorrow
    - Predict the cost of a house by grouping similar examples

21



Supervised learning

- Classification is the problem of predicting a discrete class label output for an example:
  - Binary - classify into one of two classes
  - Multiclass - classify in more than two classes

  ⇒predicting classes

- Regression is the problem of predicting a continuous quantity output for an example.

  ⇒predicting values

22

Overfitting vs
Underfitting vs Generalization

- **Overfitting**
  - when the algorithm is not able to learn a model that represents the concepts under training examples. Is not able to generalize.
  - refers to a model that models the training data too well.
  - happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data
  - **good performance on the training data, poor generalization to other data**
- **Underfitting**
  - a failure to learn the relationships in the training data
  - an underfitted model results in problematic or erroneous outcomes on new data, or data that it wasn't trained on, and often performs poorly even on training data.
  - **Underfitting refers to a model that can neither model the training data nor generalize to new data.**
  - Poor performance on the training data and poor generalization to other data
  - Underfitting is often not discussed as it is easy to detect given a good performance metric. The solution is to move on and try another alternate ML algorithm.
- **Generalization**
  - usually refers to the ability of an algorithm to be effective across a range of inputs and applications.
  - refers to the model's ability to react to new data.
  - a ML algorithm must generalize from training data to help make accurate predictions while using the model.

Overfitting and **underfitting** cause poor **generalization** on the test set

23

Steps to Machine Learning

1. **Data collection** - Machine learning requires training data, a lot of it (either labelled, meaning supervised learning or not labelled, meaning unsupervised learning).

2. **Data preparation** - Raw data alone is not very useful. The data needs to be prepared, normalized, de-duplicated and errors and bias need to be removed. Visualization of the data can be used to look for patterns and outliers to see if the right data has been collected or if data is missing.

3. **Choosing a model** - There are many models that can be used for many different purposes. Upon selecting the model, you need to make sure that the model meets the business goal. In addition, you should know how much preparation the model requires, how accurate it is and how scalable the model is. A more complex model does not always constitute a better model. Commonly used machine learning algorithms include linear regression, logistic regression, decision trees, K-means, principal component analysis (PCA), Support Vector Machines (SVM), Naïve Bayes, Random Forest and Neural Networks.

4. **Training** - Training your model is the main part of ML. The objective is to use your training data and incrementally improve the predictions of the model. Each cycle of updating the weights and biases is one training step. In supervised ML, the model is built using labelled sample data, while unsupervised machine learning tries to draw inferences from non-labelled data .

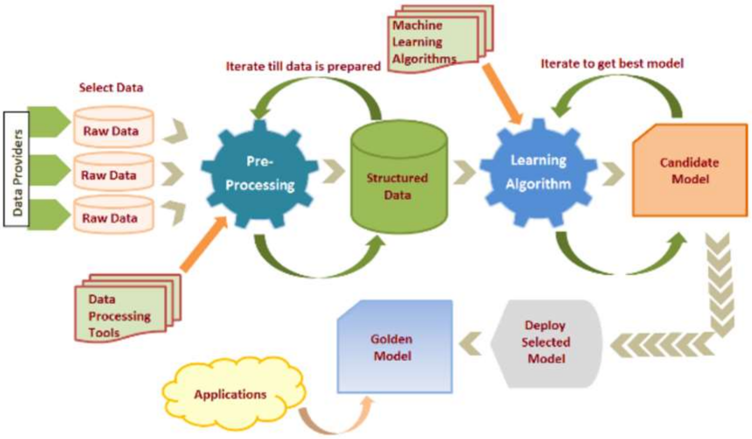
5. **Evaluation** - After training the model comes evaluating the model. This involves testing the ML against an unused control dataset to see how it performs. This might be representative of how the model works in the real world, but this does not have to be the case. The larger the number of variables in the real world, the bigger to training and test data should be.

6. **Parameter tuning** - After evaluating your model, you should test the originally set parameters to improve the AI. Increasing the number of training cycles can lead to more accurate results. However, you should define when a model is good enough as otherwise, you will continue to tweak the model. This is an experimental process.

7. **Prediction** - Once you have gone through the process of collecting data, preparing the data, selecting the model, training and evaluating the model and tuning the parameters, it is time to answer questions using predictions. These can be all kinds of predictions, ranging from image recognition to semantics to predictive analytics.

24

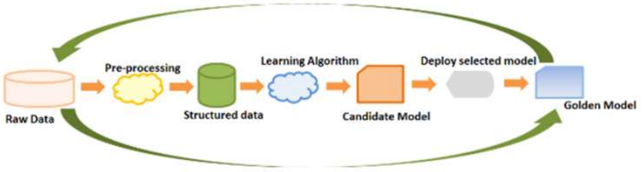
Machine Learning Process And Scenarios



https://elearningindustry.com/machine-learning-process-and-scenarios

25

Machine Learning Process And Scenarios



- (Re)training may be a constant task, and all the process must be repeated frequently to catch the dynamic chan ges of the problem to which ML is being applied.

https://elearningindustry.com/machine-learning-process-and-scenarios

26