DASCI

Instituto Andaluz Interuniversitario
en Ciencia de Datos e
Inteligencia Computacional

Ciencia de Datos a través del Big Data

Diego García (djgarcia@ugr.es)
Isaac Triguero (isaaktriguero@ugr.es)

Financiado por
la Unión Europea
NextGenerationEU

GOBIERNO
DE ESPAÑA

MINISTERIO
PARA LA TRANSFORMACIÓN DIGITAL
Y DE LA FUNCIÓN PÚBLICA

SECRETARÍA DE ESTADO
DE DIGITALIZACIÓN
E INTELIGENCIA ARTIFICIAL

Plan de
Recuperación,
Transformación
y Resiliencia

UNIVERSIDAD
DE GRANADA

UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

REPSOL

CAMBRIDGE UNIVERSITY PRESS

Chapter X

Introduction to Machine Learning

Outline

- Introduction to Machine Learning
- ML Paradigms
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning
- Data cleansing
- Evaluating ML algorithms
- Pipelines

Terminology

- Learning
- Instances
- Features / attributes
- Dataset
- Train/Test data
- Model
- Generalization

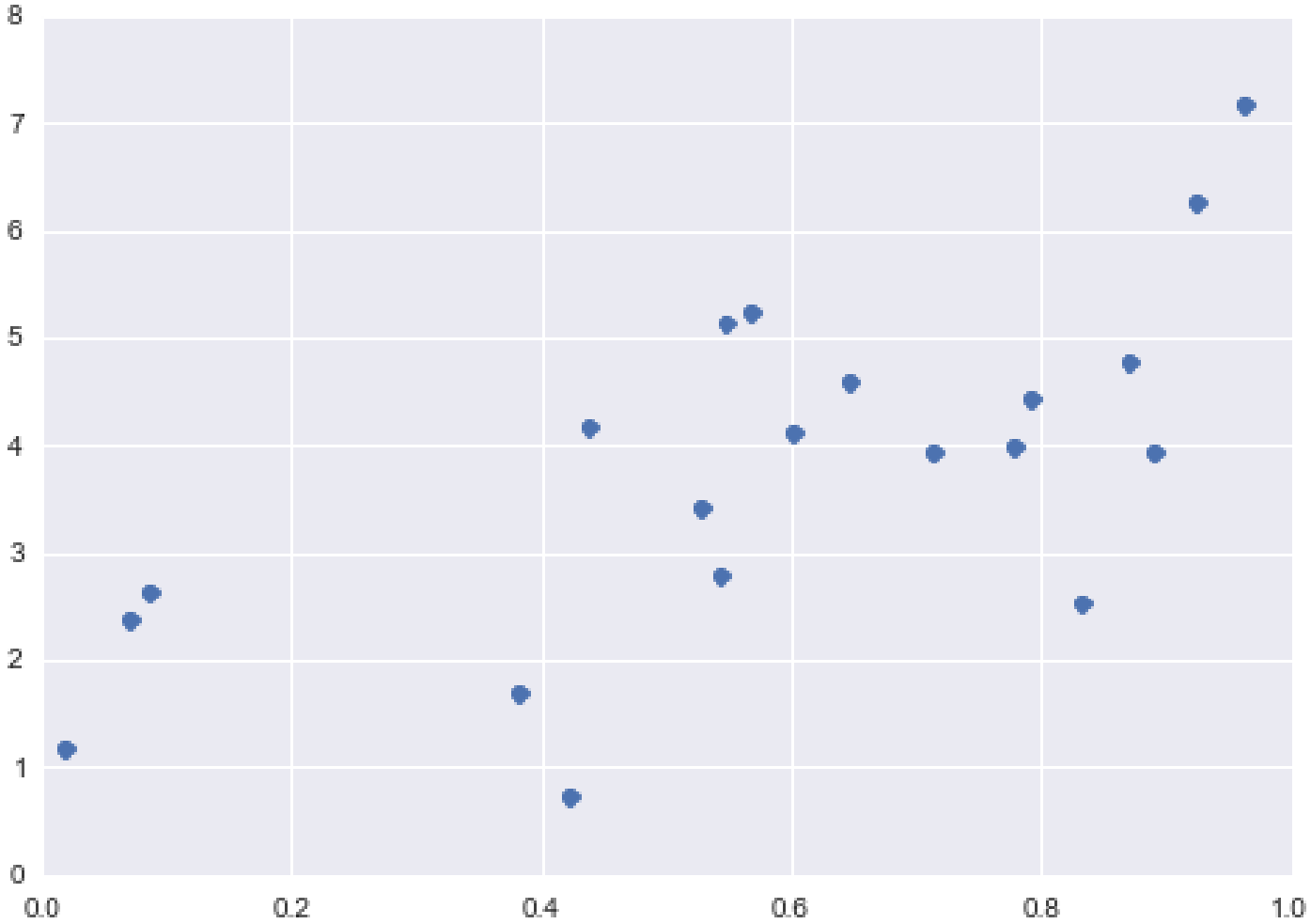
Machine Learning Paradigms

- Supervised learning
  - Labeled examples / instances
- Unsupervised learning
  - We don’t have labels, the desired output is unknown
- Reinforcement learning
  - The machine acts and receives a rewards depending on its actions

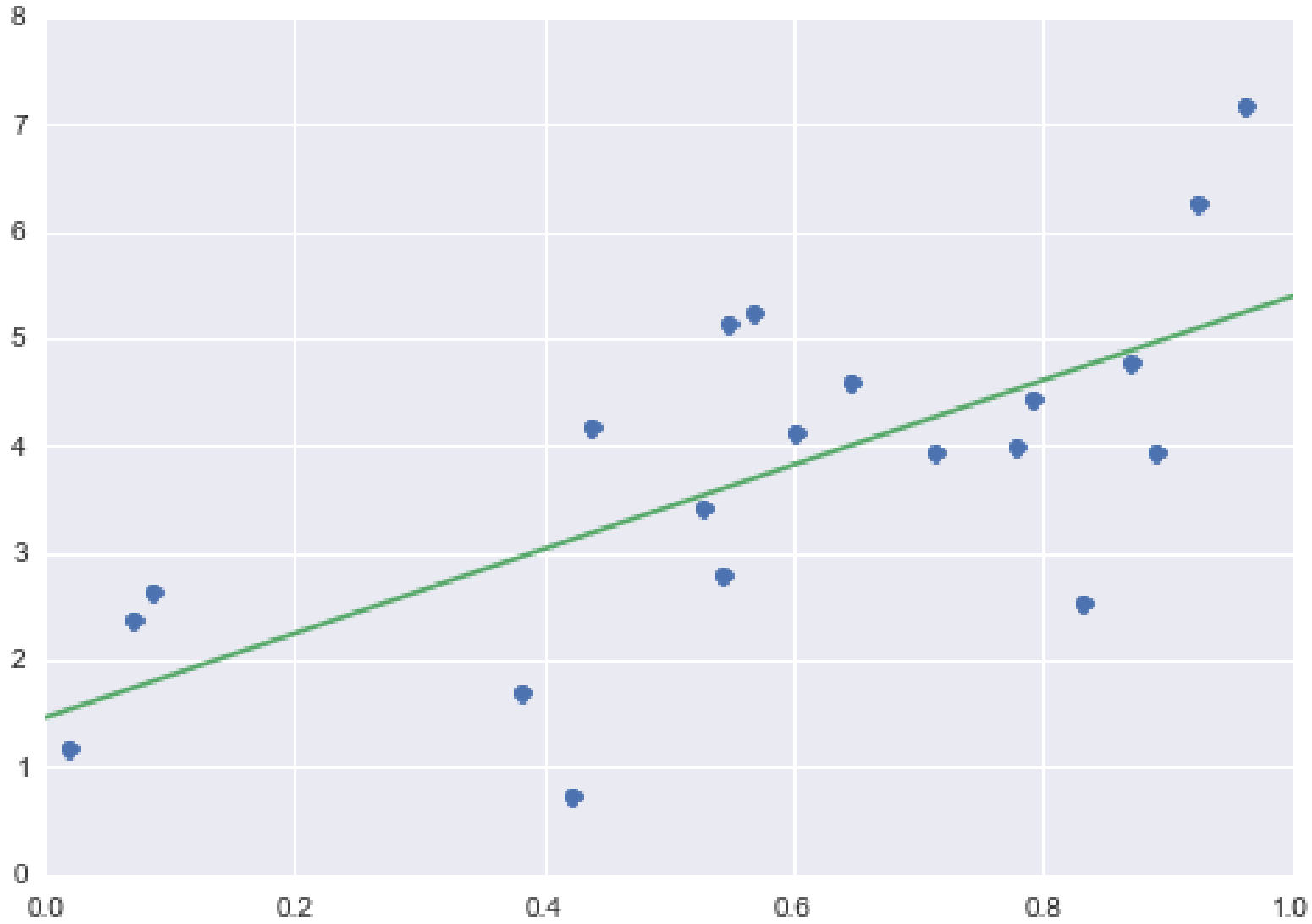
Supervised Learning

- We aim to find the relationship between the training features, and their labels
- Depending on the nature of the output label:
  - Classification
    - Discrete value (faulty/non-faulty, cat/dogs, true/false, etc.)
    - Multi-class
  - Regression
    - Continuous value (next day's temperature, gas prices, etc.)
- Linear and logistic regression, Naïve Bayes, decision trees, k-Nearest Neighbors (kNN), support vector machines (SVMs), neural networks, and ensembles

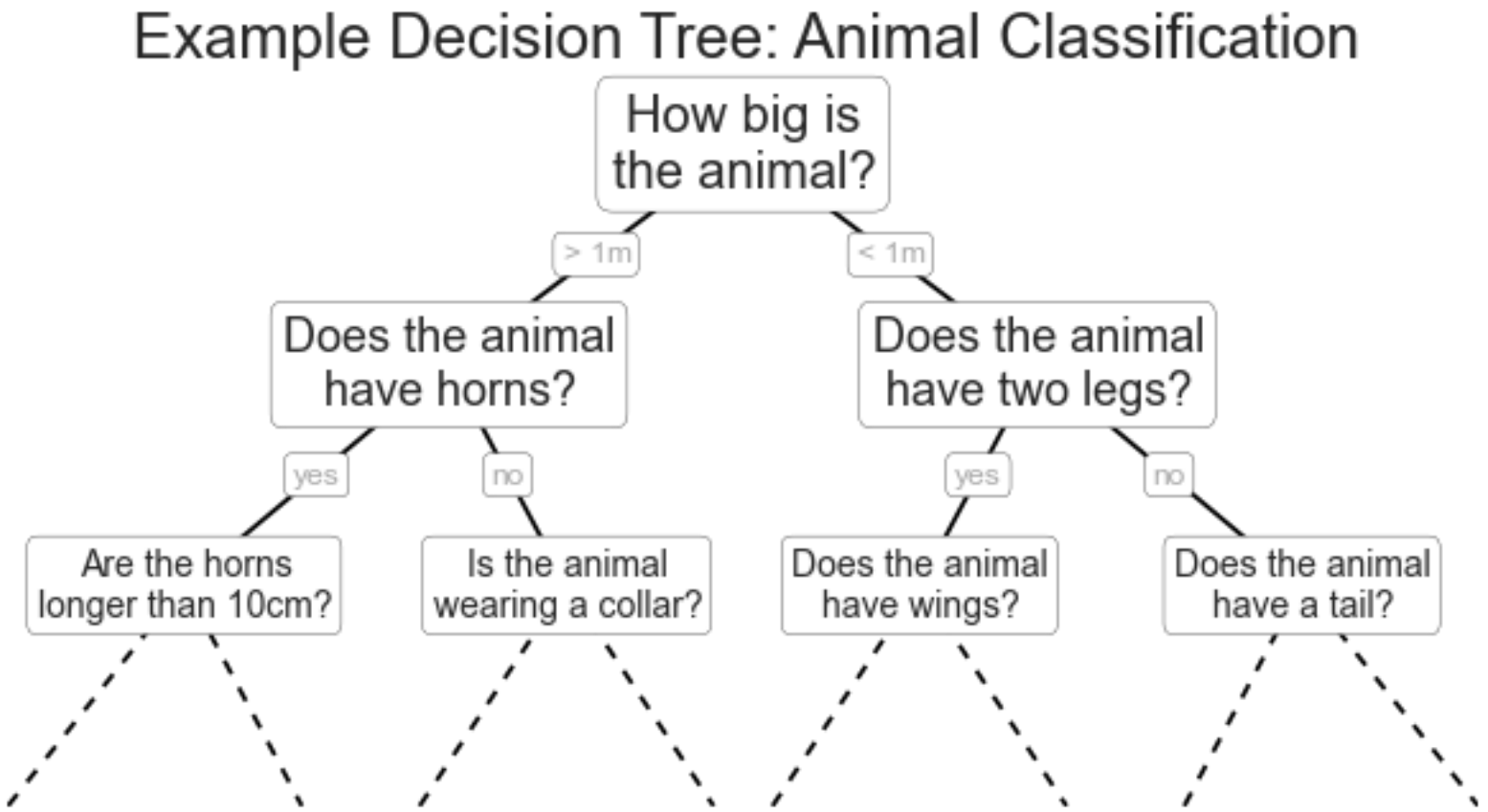
Linear Regression



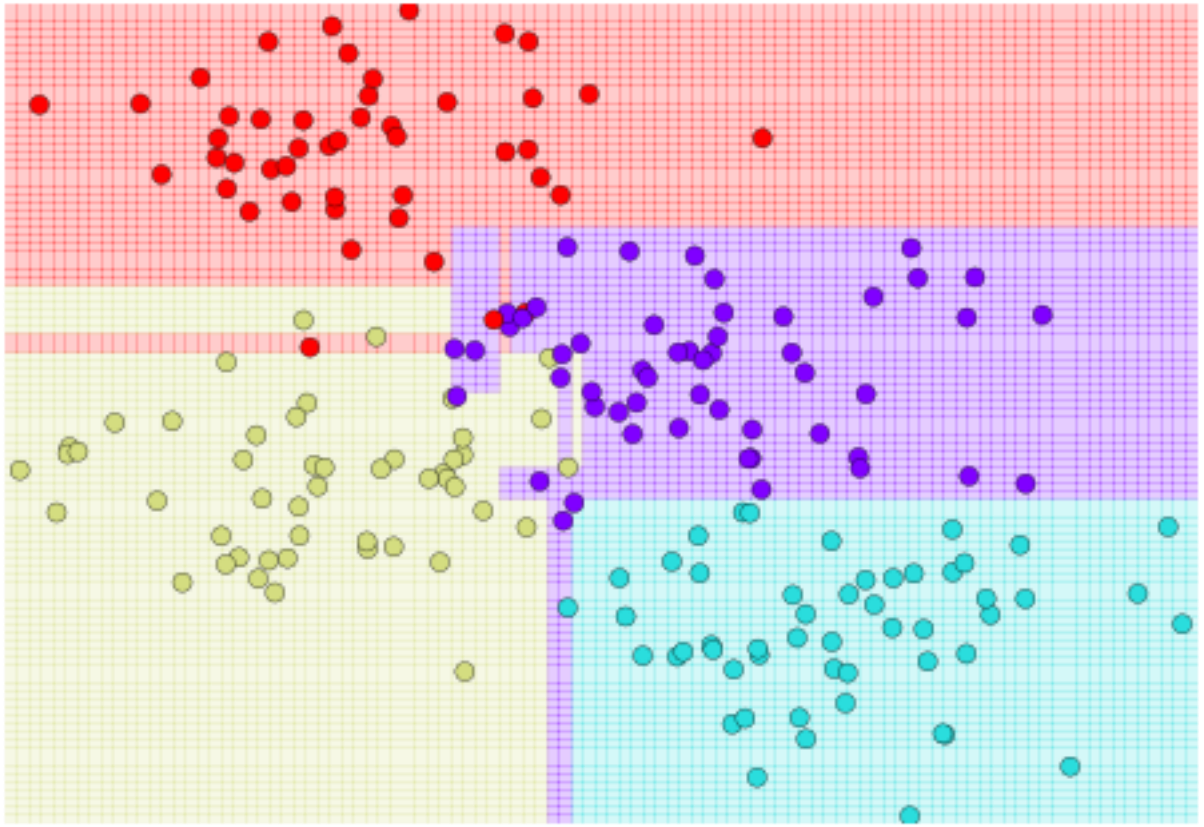
Linear Regression



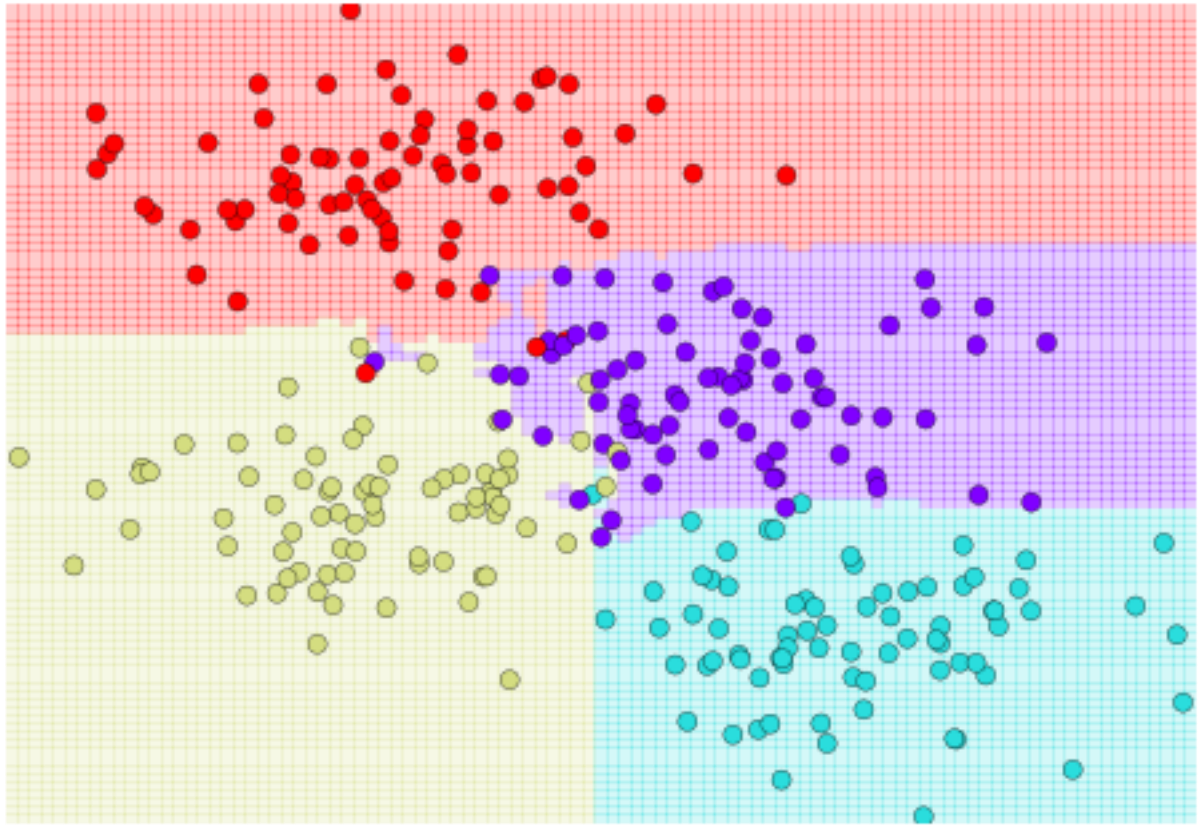
Decision Trees



Decision Trees



Random Forest (Ensemble)

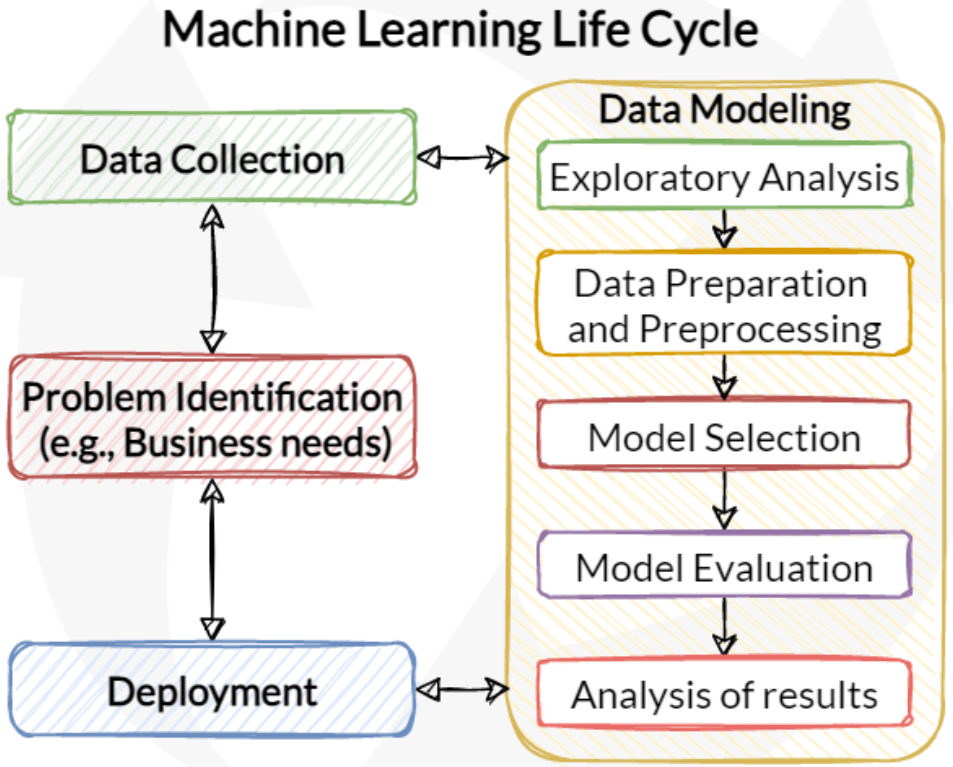


Unsupervised Learning

- We don’t aim to find a relationship between features and labels.
- The goal is to describe and understand the structure and patterns within the data
- Depending on the objective:
  - Dimensionality reduction
    - Summarizes the data by reducing the number of dimensions (features), preserving the largest amount of information.
    - Principal components analysis (PCA), singular value decomposition (SVD)
  - Clustering
    - Groups similar examples into “clusters”
    - k-Means, density-based clustering

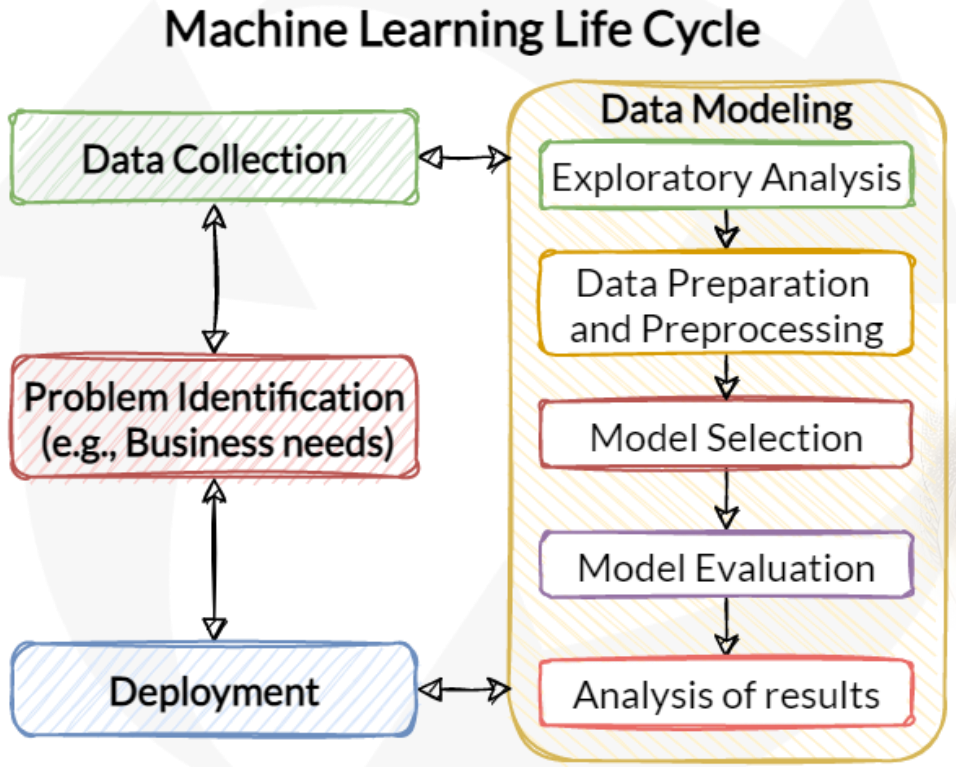
Machine Learning Life Cycle

- For an ML application to be successful, we will go through various stages
- The Machine Learning Life Cycle
- Iterative process
  - We may go back and forth through the steps



Machine Learning Life Cycle

- All of the stages **may** (**or may not**) be affected by Big Data
- Shouldn't preprocessing help? (Feature Selection in a 5TB dataset?)
- You may not encounter any problem to train a model (e.g. you do it with GPUs), but the problem lies in deploying it



Data Preparation and Preprocessing

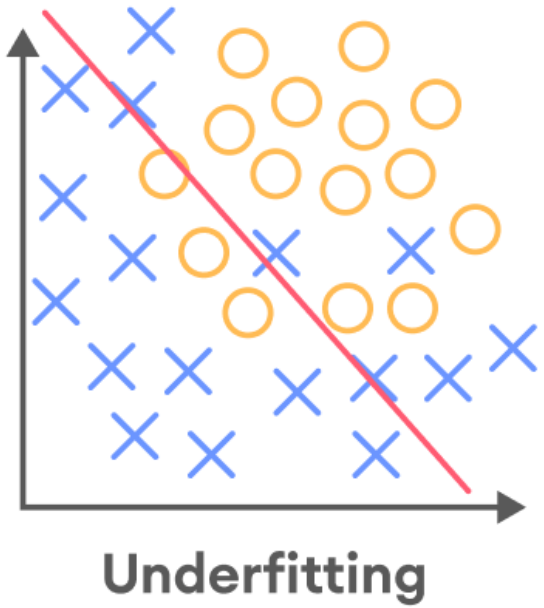
- Data Reduction
  - Feature Selection / Generation
  - Instance Selection / Generation
  - Discretization
  - Normalization
- Imperfect Data
  - Missing Values Imputation
  - Noise Treatment
- Imbalanced Datasets
- Most of them are problem-dependent. The “random” use can simply end with a model that is not useful!

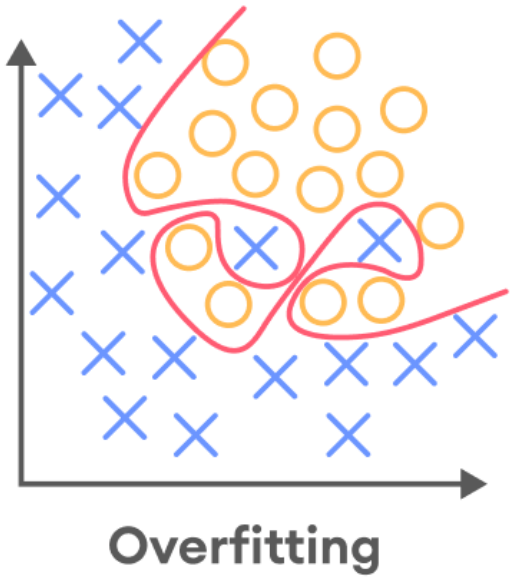
Model Evaluation and Selection

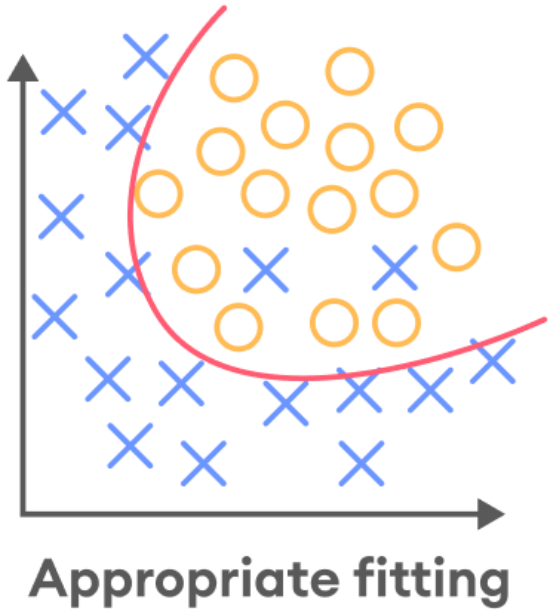
- ML methods have several hyperparameters that influence and govern their behavior
- Parameters vs hyperparameters
- How to evaluate the success in learning?
- Validation strategy:
  - Training set: used to teach a ML model with a given set of hyperparameters
  - Validation set: used to check the error / success rate of the taught ML model
  - Test set: used to evaluate the behavior of the chosen model and its configuration. Gives a hint of how the model will perform in production

Model Evaluation and Selection

- Bias & variance / underfitting & overfitting
- Model selection or hyperparameter tuning

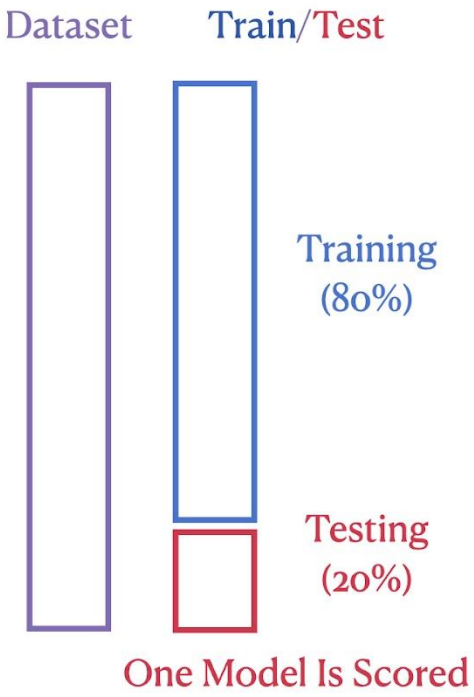


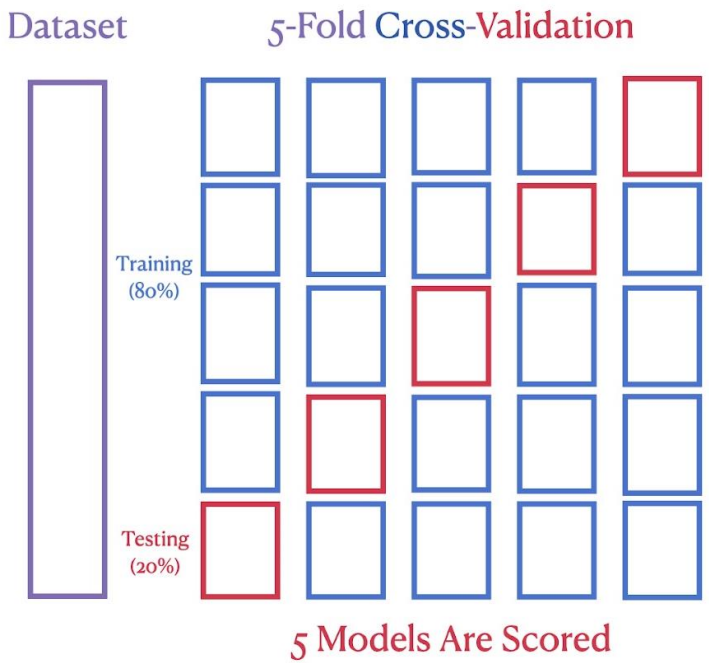




Model Evaluation and Selection

- K-fold cross validation
- Problem dependent validation strategies: stratified partitioning, leave-one-out, time series





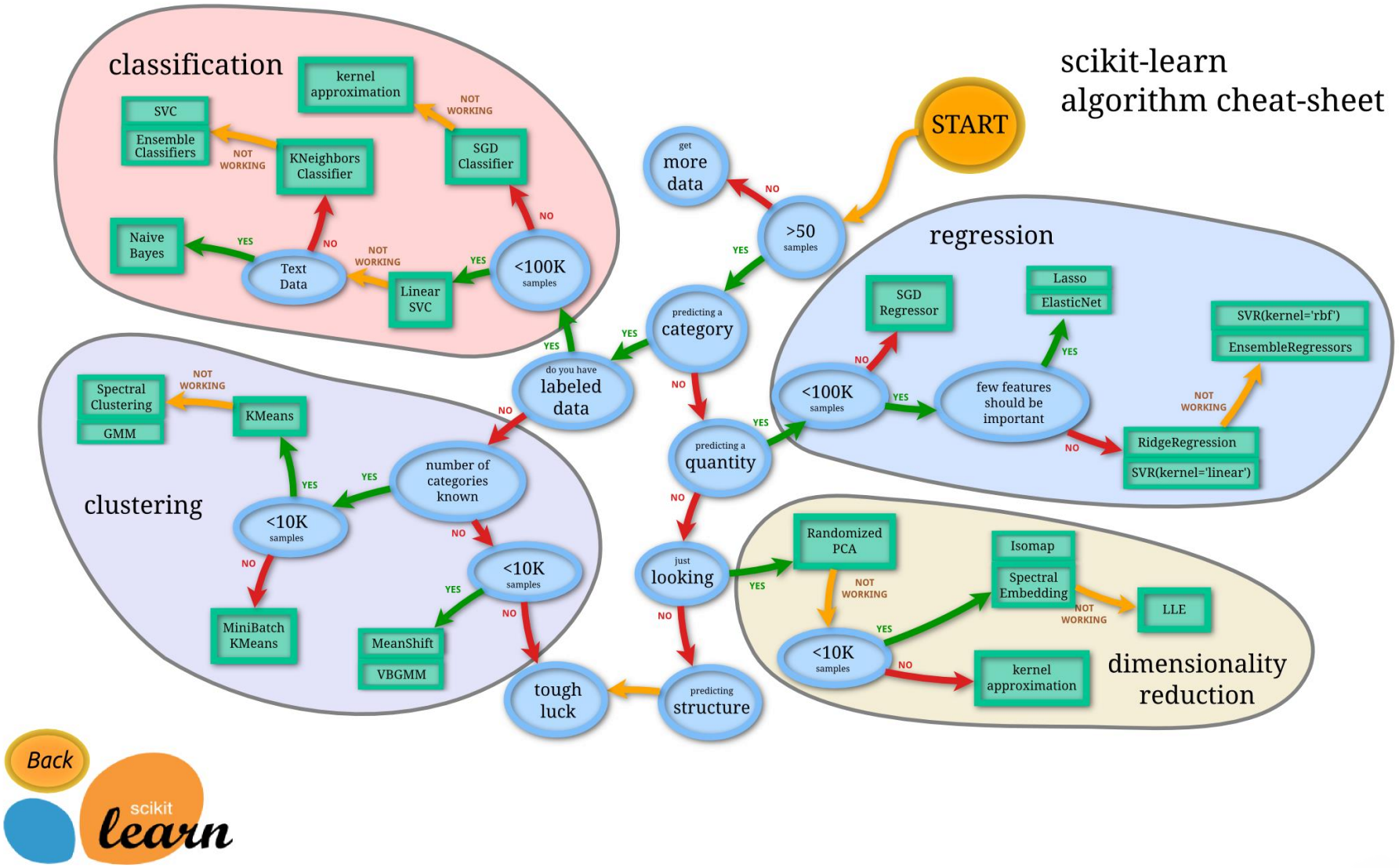
Machine Learning Pipelines

- The ML life cycle follows an iterative workflow from data capture to model deployment
- Automating part of this workflow is known as *pipeline*
- Not a step of ML life cycle, but a useful MLOps tool
- Avoids transformations of training data not being applied to test data
- Pipelines will help group a set of preprocessing and transformation techniques needed prior to teaching and evaluating models
- Already provided in scikit-learn, and the Spark MLlib library

Machine Learning Libraries

- Machine learning end-users are usually not experts in machine learning
- They usually rely on open-source libraries that allow them to perform data analytics - **Off-the-shelf techniques**
- **Minimum knowledge still required**:
  - Understanding the behavior of the techniques
  - Appropriateness for the problem at hand
  - Hyperparameters' optimization
  - Experiment validation/performance metrics
- In the most extreme scenario, complete newbies could use Auto-ML (not much for Big data though! Have a look at *H2O*)

ML Basic Cheat-Sheet



To Learn More

- **If you are not very familiar with Machine Learning:**
  - Classic books: (Bishop, 2006), (Hastie, 2001), (Ng, 2017)
  - Data pre-processing books: (Garcia, 2015), (Luengo, 2020)
  - On-line courses: Machine Learning Specialization by Coursera, Machine Learning in Python with Scikit-learn MOOC
  - Hands on with Python: (Garreta, 2017), (Geron, 2022)
- **To deal with Big Data and Machine Learning**
  - MLlib: (Xiangrui, 2016)
  - Spark Packages
  - Other libraries: FlinkML, *DaskML* (Daniel, 2019).

CAMBRIDGE UNIVERSITY PRESS

Chapter X

Introduction to Machine Learning