Machine Learning for Economic Research

StatsTalk-Africa

CHRISTOPHE HURLIN - UNIVERSITY OF ORLEANS AND IUF

September 12, 2023

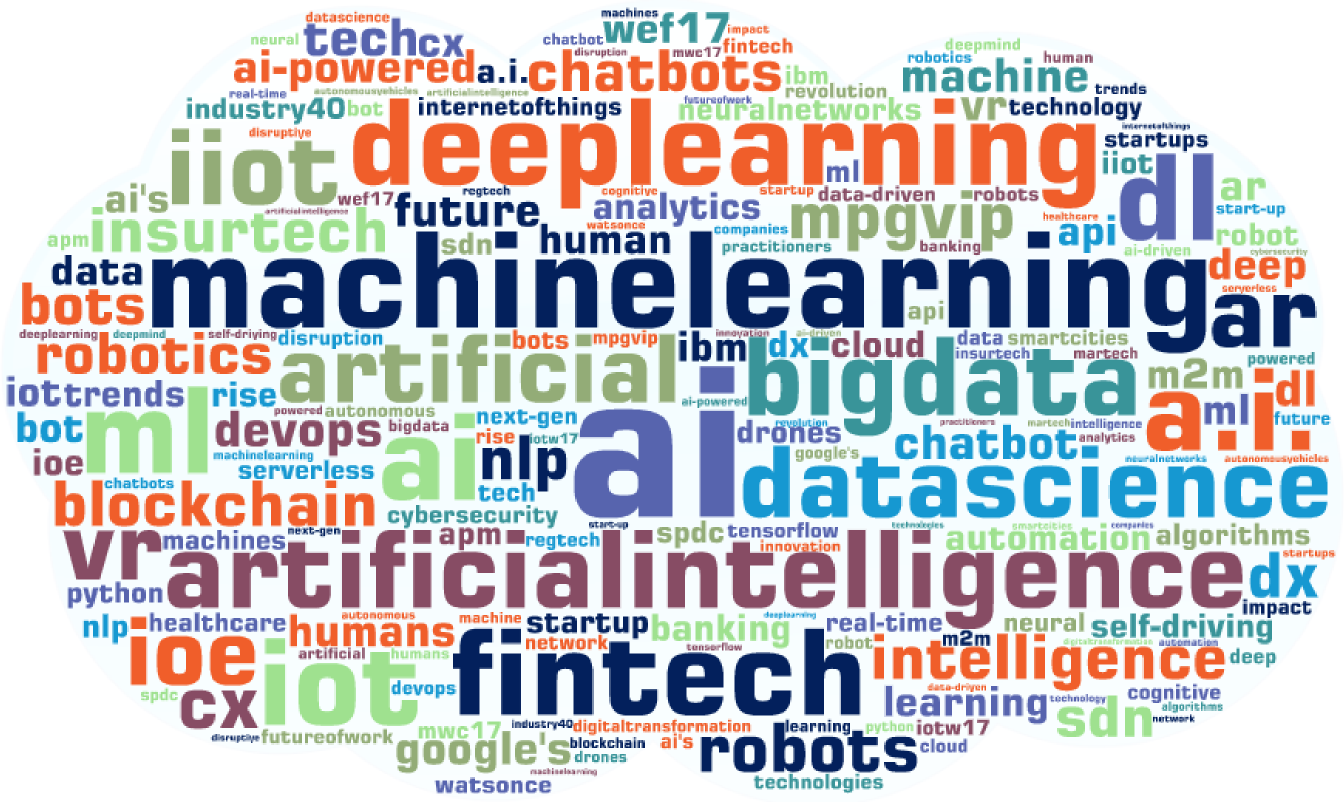
Outline

This short presentation is curated for economists interested in using Machine Learning (ML thereafter) for their research and applications.

It will be focused on 4 questions :

1. What is Machine Learning?
2. What are the main differences /similarities between ML and econometrics?
3. When is ML useful in economics?
4. Which are the ML models generally used in economics?

Many words and many concepts



Machine learning: general definition

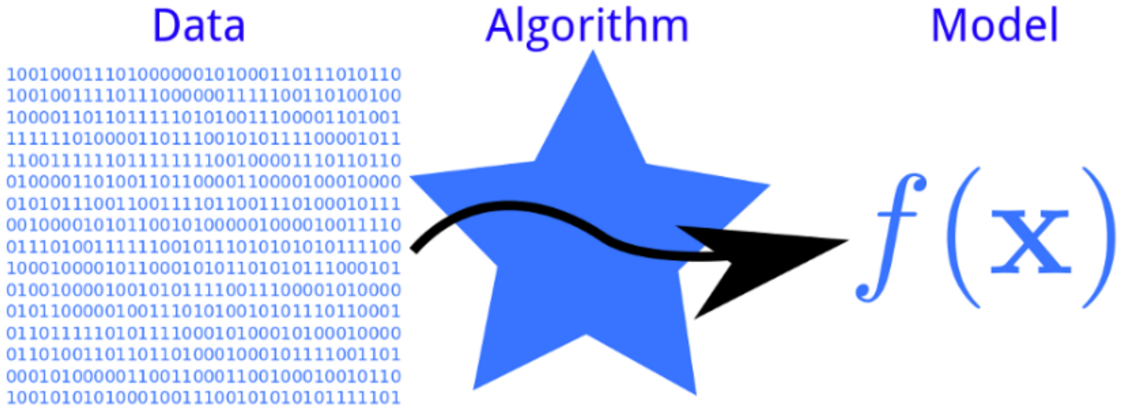
Machine Learning :

« *Machine learning is a field of inquiry devoted to* ***understanding*** *and* ***building*** *methods that "****learn****" – that is, methods that leverage data to improve performance on some set of tasks. It is seen as a part of* ***artificial intelligence****.*

Machine learning ***algorithms*** *build a* ***model*** *based on sample data, known as* ***training data****, in order to make predictions or decisions without being explicitly programmed to do so* ».

Source : Wikipedia.

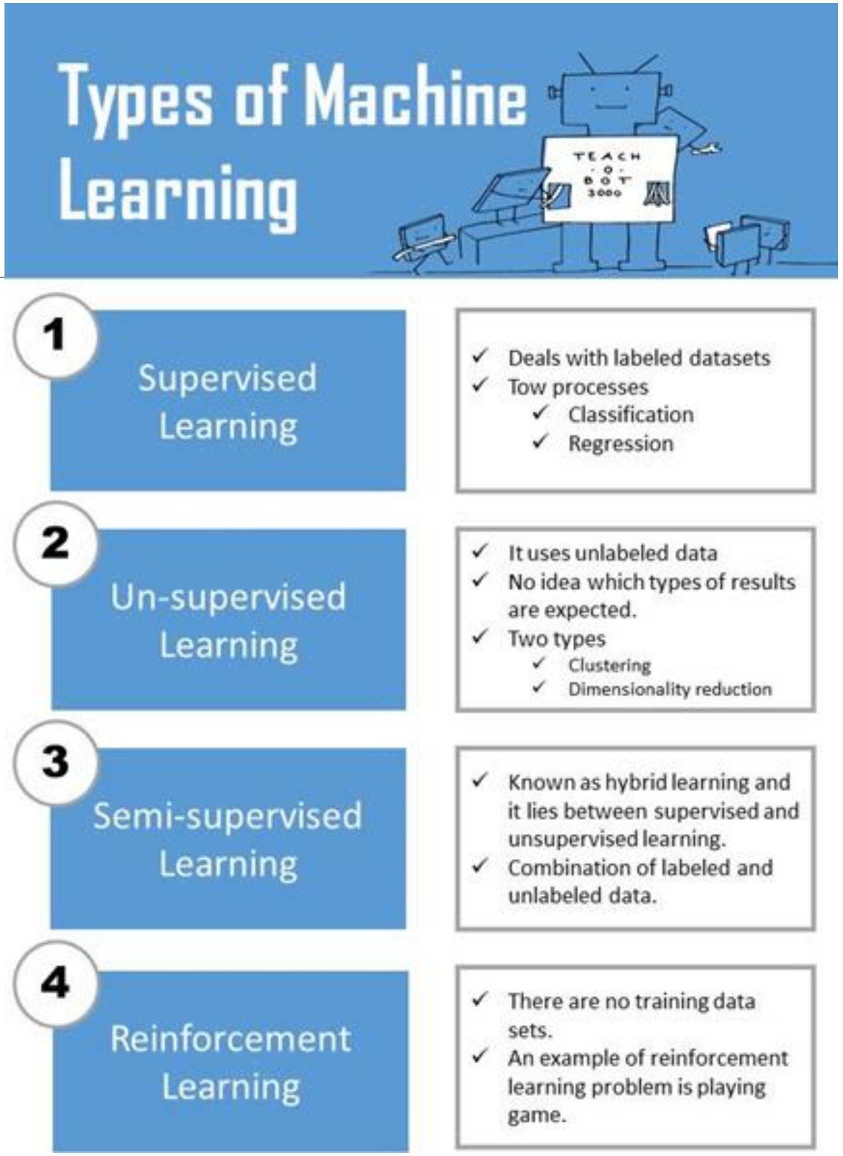
Machine learning: general definition



Type of Machine Learning

There are 4 major machine learning modes:

1. **Supervised learning :** aims to detect an explanatory or predictive model on labeled data.
2. **Unsupervised learning** *:* based on the use of unlabeled data.
3. **Semi-supervised learning** *:* learning based on labeled and unlabeled data.
4. **Reinforcement learning** *:* A behavioral learning model in which the algorithm receives feedback from data analysis and guides the user to the best result



Types of Machine Learning

TEACH-O-BOT 3000

1 Supervised Learning

- ✓ Deals with labeled datasets
- ✓ Tow processes
  - ✓ Classification
  - ✓ Regression

2 Un-supervised Learning

- ✓ It uses unlabeled data
- ✓ No idea which types of results are expected.
- ✓ Two types
  - ✓ Clustering
  - ✓ Dimensionality reduction

3 Semi-supervised Learning

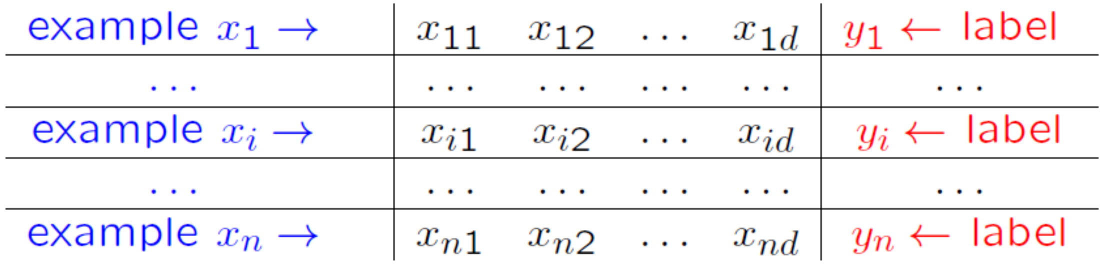
- ✓ Known as hybrid learning and it lies between supervised and unsupervised learning.
- ✓ Combination of labeled and unlabeled data.

4 Reinforcement Learning

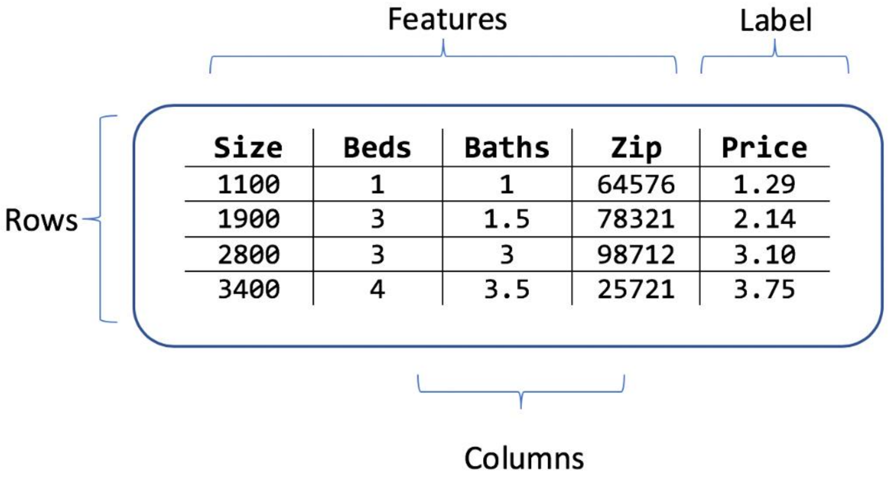
- ✓ There are no training data sets.
- ✓ An example of reinforcement learning problem is playing game.

Vocabulary

- Consider a **training dataset** denoted by $(x_1, y_1), \dots, (x_n, y_n)$ issued from an **experiment**.
- Each element of the sample corresponds to an **example** or an **instance.**
- $x_i \in \mathbb{R}^d$represents the **features vector** of an instance.
- $y_i$is the **target variable** and its value is the **label**.



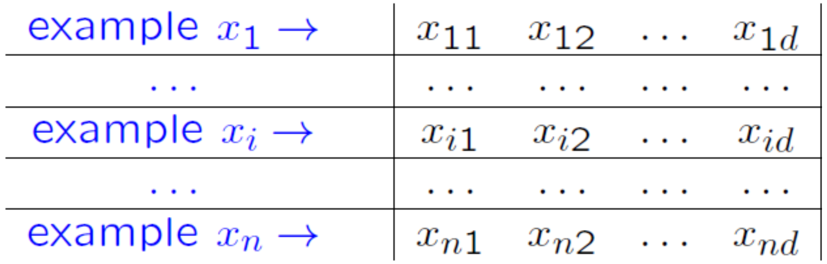
example $x_1 \rightarrow$	x_{11}	x_{12}	$\dots$	x_{1d}	$y_1 \leftarrow$ label
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
example $x_i \rightarrow$	x_{i1}	x_{i2}	$\dots$	x_{id}	$y_i \leftarrow$ label
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
example $x_n \rightarrow$	x_{n1}	x_{n2}	$\dots$	x_{nd}	$y_n \leftarrow$ label


Features
Label
Rows
Columns

Size	Beds	Baths	Zip	Price
1100	1	1	64576	1.29
1900	3	1.5	78321	2.14
2800	3	3	98712	3.10
3400	4	3.5	25721	3.75

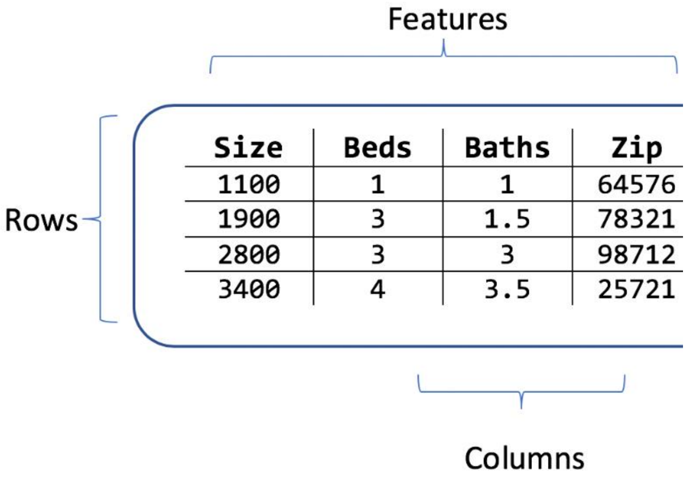
Vocabulary

- Consider a **training dataset** denoted by $(x_1, y_1), \dots, (x_n, y_n)$ issued from an **experiment**.
- Each element of the sample corresponds to an **example** or an **instance.**
- $x_i \in \mathbb{R}^d$ represents the **features vector** of an instance.
- y_i is the **target variable** and its value is the **label**.



example $x_1 \rightarrow$	x_{11}	x_{12}	$\dots$	x_{1d}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
example $x_i \rightarrow$	x_{i1}	x_{i2}	$\dots$	x_{id}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
example $x_n \rightarrow$	x_{n1}	x_{n2}	$\dots$	x_{nd}

unlabeled data



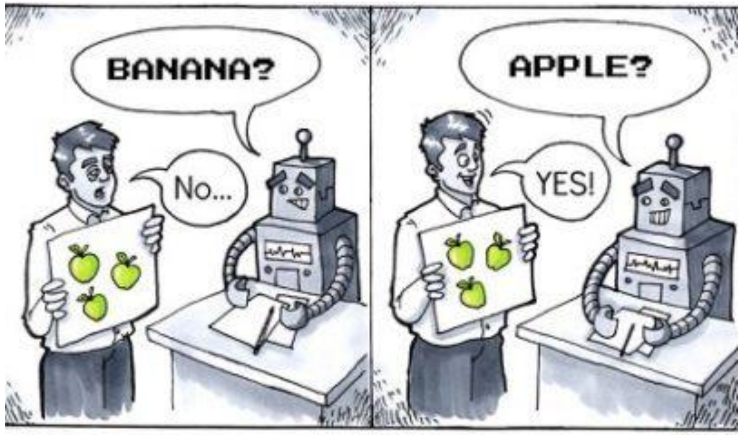
Size	Beds	Baths	Zip
1100	1	1	64576
1900	3	1.5	78321
2800	3	3	98712
3400	4	3.5	25721

Supervised learning

Supervised learning: definition

Supervised learning : the algorithm learns from a set of learning data comprising labeled data, i.e. data already labeled with the correct label.

The algorithm then learns a general **rule (the model)** for classification or regression which will then be used **to predict labels** when new data is analyzed.



Supervised Learning

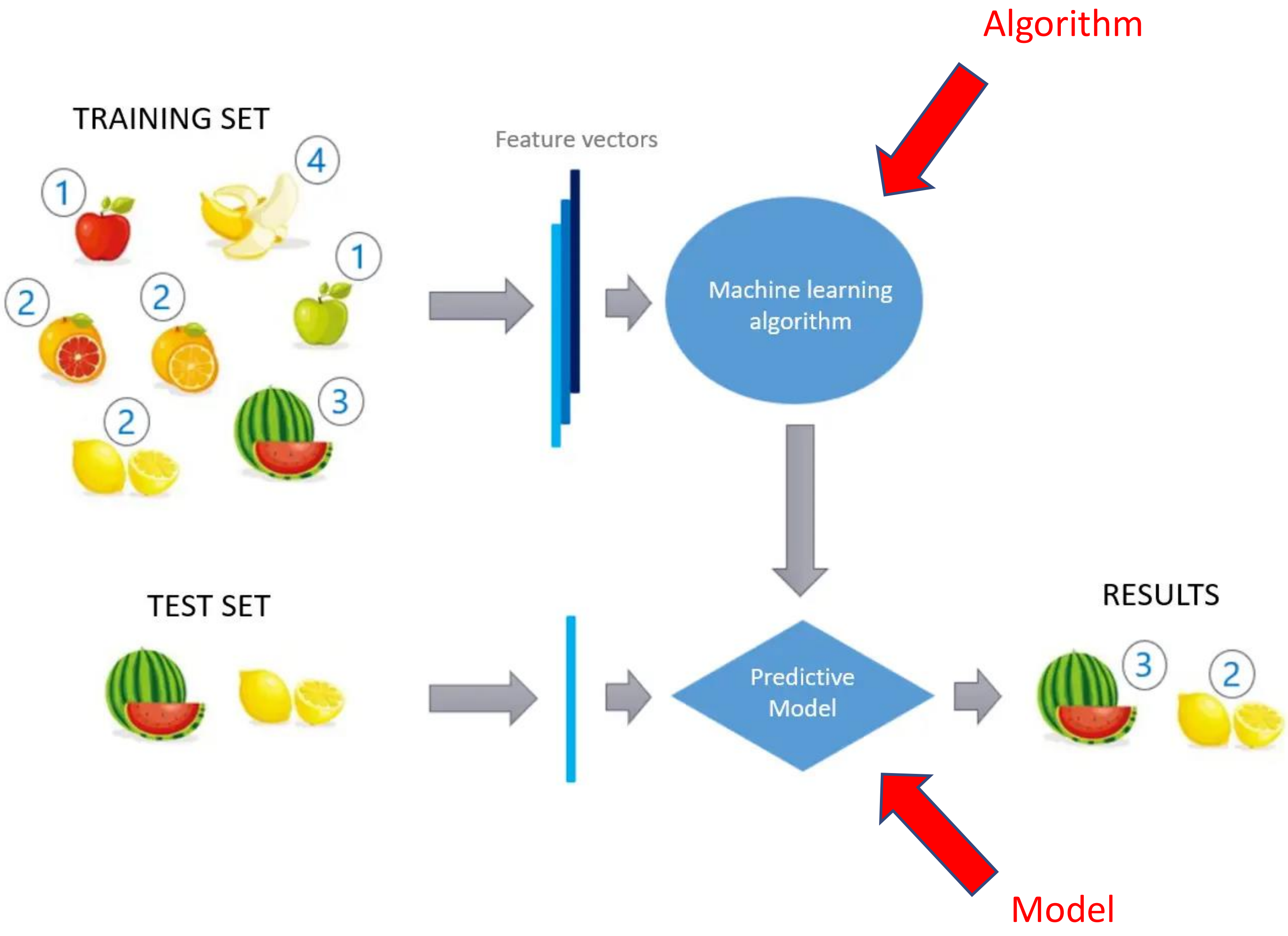
Supervised learning

We have a set of images (data) labeled as 1 if apples, 2 if citrus, 3 if watermelon and 4 if banana.

For each image, we have a set of characteristics (color, shape, etc.) presented in the form of vectors and it is with these that **the supervised learning algorithm is trained** .

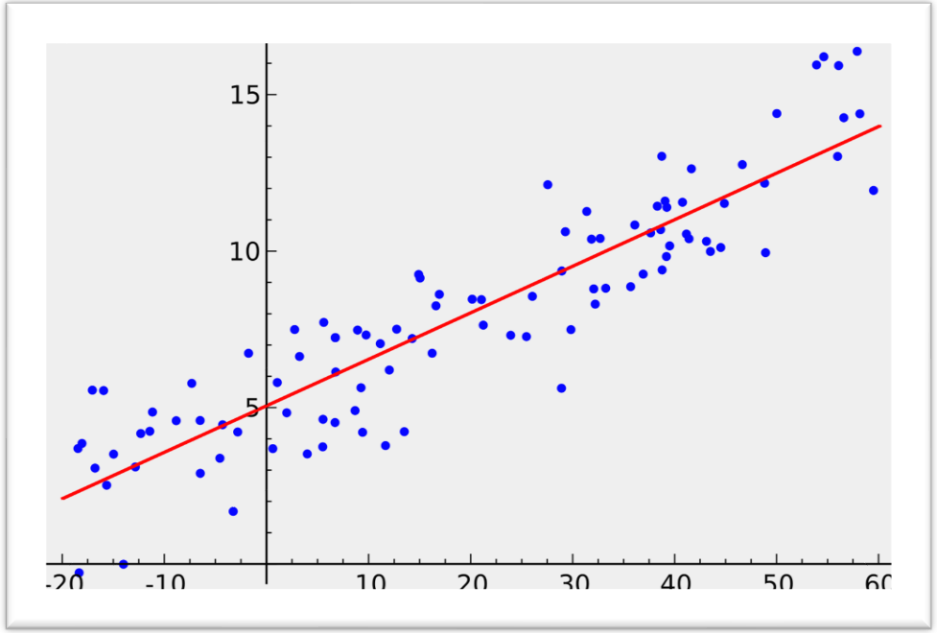
The **generated model** then makes it possible to classify the new elements that we have never seen in the training phase.

Source: Diego Calvo

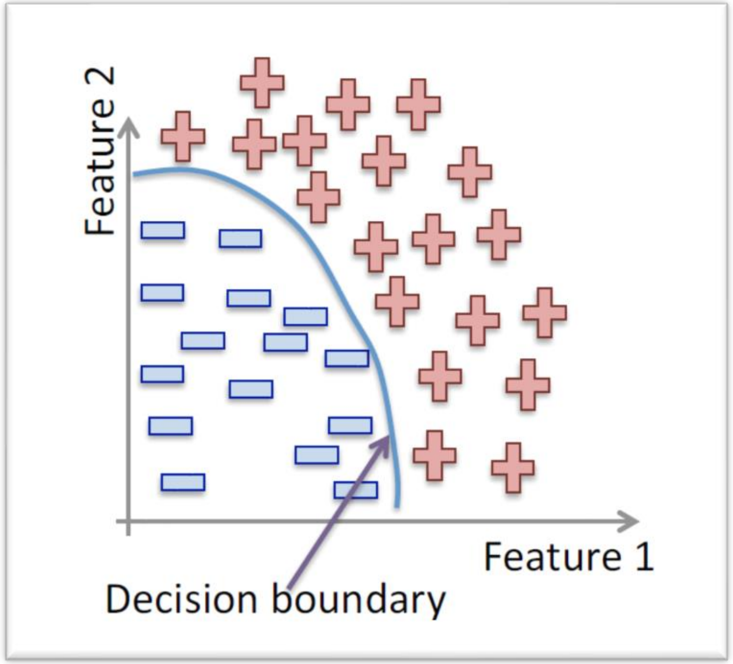


Supervised learning

REGRESSION



CLASSIFICATION



Source : Machine Learning Basic Concepts - edX

Application example for credit risk

Dumitrescu, E.-I., Hué , S. Hurlin, C., and Tokpavi, S. (2022), Machine Learning for Credit Scoring: Improving Logistic Regression with Non-Linear Decision-Tree Effects, European Journal of Operational Research, 297(3), 1178-1192 .

In this article, we propose a powerful and interpretable scoring method called PLTR (Penalised Logistic Regression Model) which uses information from decision trees to improve the performance of logistic regression.

Rules extracted from various shallow decision trees constructed with original predictor variables are used as predictors in a penalized logistic regression model.

Context: Ensemble methods (e.g., Random Forest) offer better classification performance than logistic regression models for credit scoring, but pose the problem of their lack of interpretability.

ELSEVIER
Volume 304, Issue 3, 1 February 2023
ISSN 0377-2217
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
Available online at www.sciencedirect.com
ScienceDirect
www.elsevier.com/locate/ejor
EURO

Unsupervised learning

Unsupervised learning

Unsupervised learning refers to algorithms that learn from a set of **unlabeled data**.

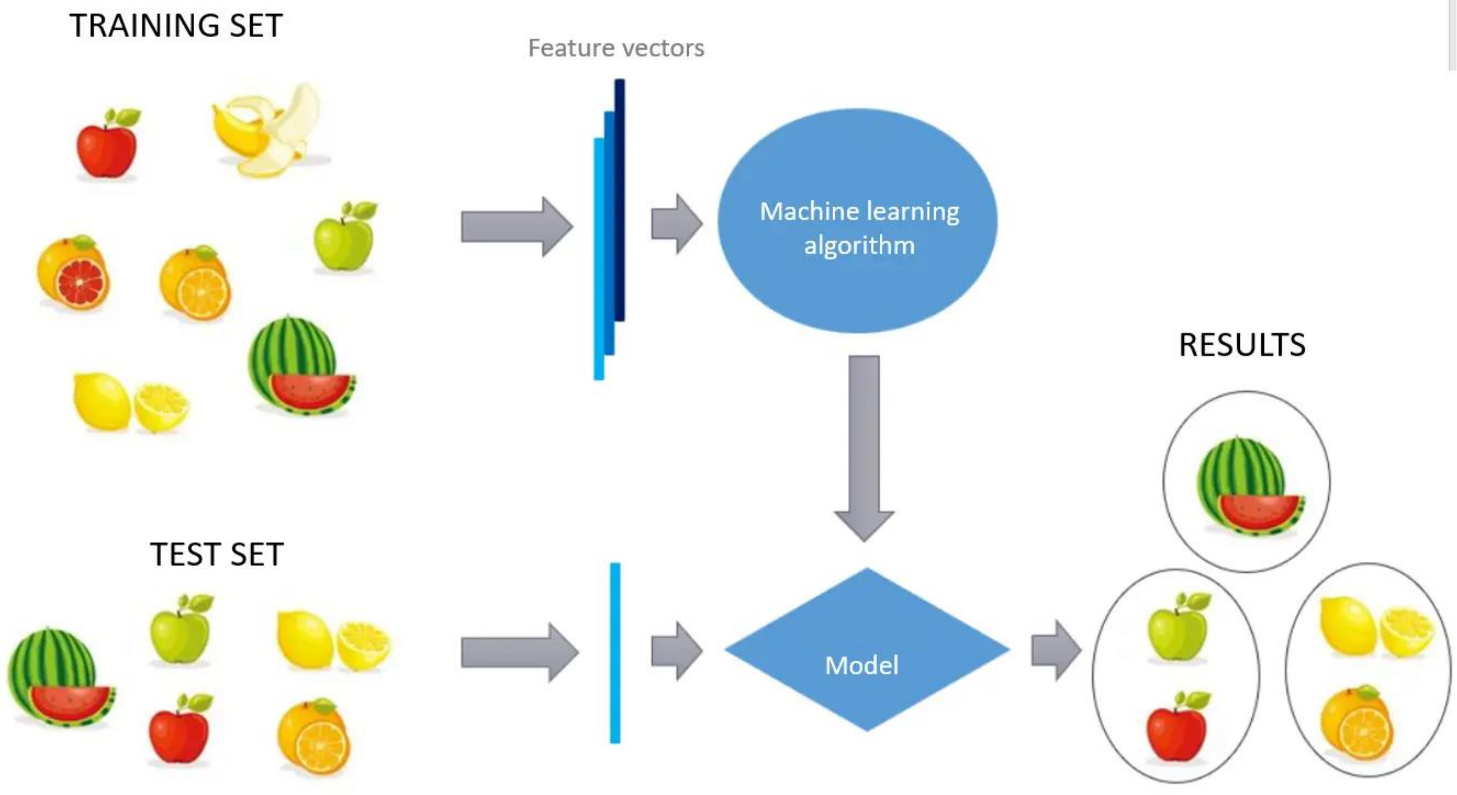
- The objective is no longer to predict the target variable.
- Algorithms look for regularities or association rules in data by identifying **groups of similar observations** (**clusters**).
- The main problems addressed using unsupervised learning algorithms are **clustering** and **dimension reduction** (PCA, etc.).

Unsupervised learning

We have a set of unlabeled images (data).

The model then groups the images (clustering): lemons with lemons, pumpkins with pumpkins, apples with apples, etc.

Source: Diego Calvo



Application example for credit risk

Bakoben M, Bellotti T, Adams N. (2020) Identification of credit risk based on cluster analysis of account behaviors . Journal of the Operational Research Society. 71:775–783.

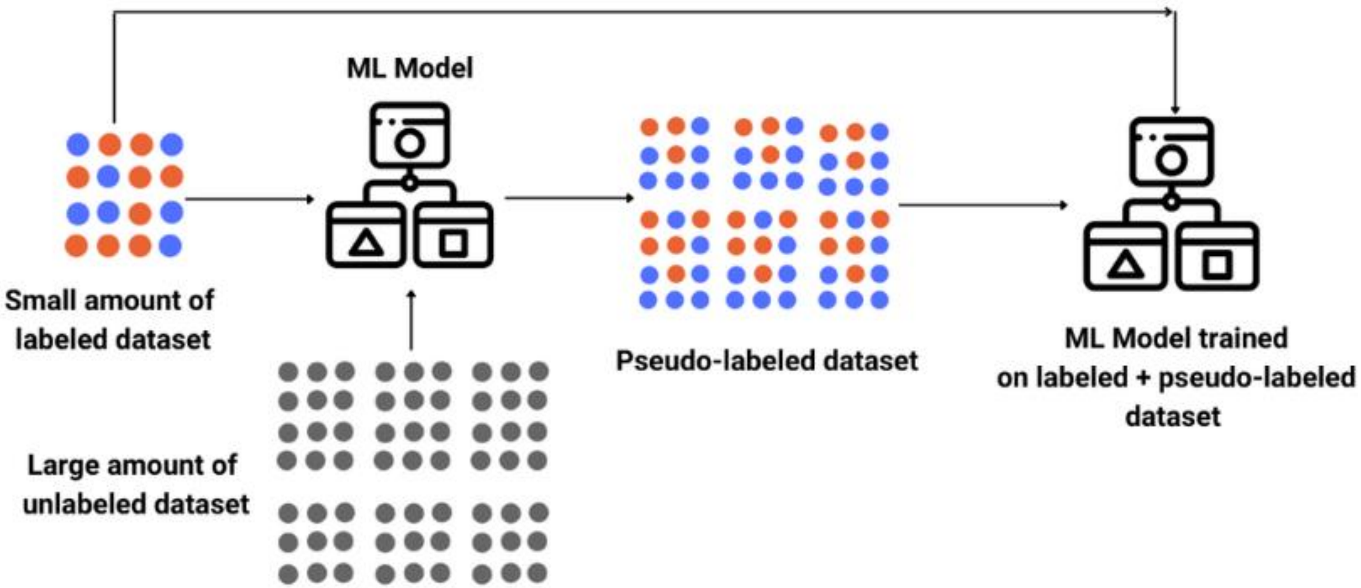
- New clustering method to constitute risk groups relating to the behavior of bank accounts (revolving credit).
- Monthly credit card data for 494 UK bank accounts for up to two years.
- The behavior of the accounts is modeled parametrically by a VAR model (utilization rate, reimbursement rate).
- Behavioral cluster analysis using a dissimilarity measure of statistical model parameters (VAR parameters, here).
- Aggregations of actual credit card behavior data.
- New default prediction model that includes customer assignments in clusters. Gain in performance (AUC).

JORS
JOURNAL OF THE OPERATIONAL RESEARCH SOCIETY
John Boylan
Martin H. Kunc
Said Salhi
Zhe George Zhang
or
THE OPERATIONAL RESEARCH SOCIETY
Taylor & Francis

Semi-supervised learning

Semi-supervised learning

Semi-supervised learning : learning based on both labeled (generally minority) and unlabeled (generally majority) data .



ML Model
Small amount of labeled dataset
Large amount of unlabeled dataset
Pseudo-labeled dataset
ML Model trained on labeled + pseudo-labeled dataset

Application example for credit risk

Credit score with reintegration of refused clients.

Shen F., Yang Z., Zhao X. and Lan D. (2022), Reject inference in credit scoring using a three-way decision and safe semi- supervised support vector machine, Information Sciences 606, 614–627.

When assessing credit risk, reintegrating rejected customers into the learning base is a technique to address sample selection bias.

In credit assessments, the accepted sample is labeled and the rejected sample is unlabeled.

This article proposes a new approach based on:

(i) A method of correcting feature distributions between accepted and rejected,

(ii) A semi-supervised support vector machine (S4VM) type semi- supervised classification method .

INFORMATION SCIENCES

Informatics and Computer Science

Intelligent Systems

Applications

AN INTERNATIONAL JOURNAL

ScienceDirect

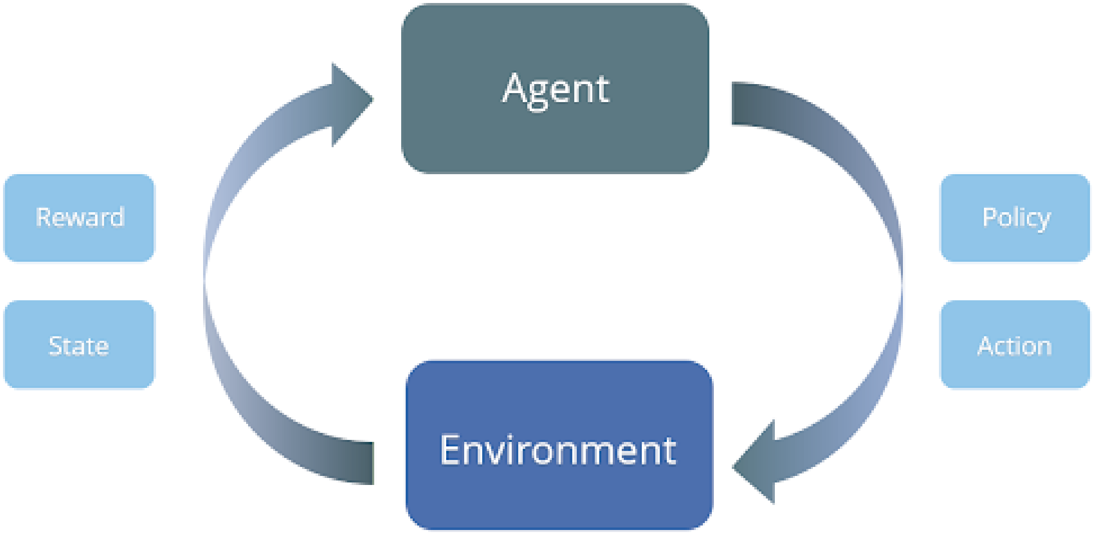
Other forms of learning

Reinforcement learning

Reinforcement learning: behavioral learning model in which the algorithm receives feedback from the agent and guides the user towards the best result.

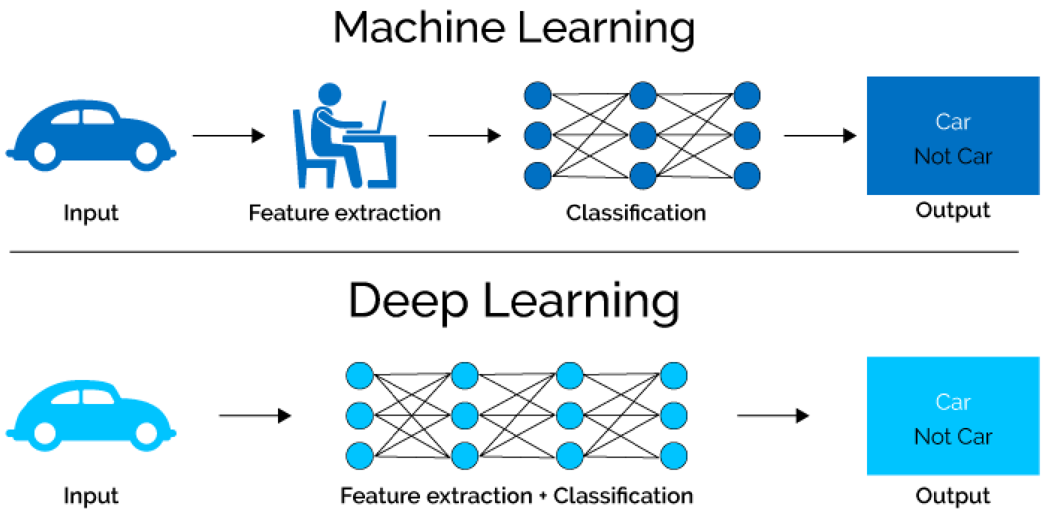
A priori, no (or few) uses in economics.

Main applications: gaming, autonomous cars, etc.



Deep Learning

Deep Learning : A subcategory of ML that refers to specific methods that integrate neural networks in successive layers in order to learn data in an iterative way.

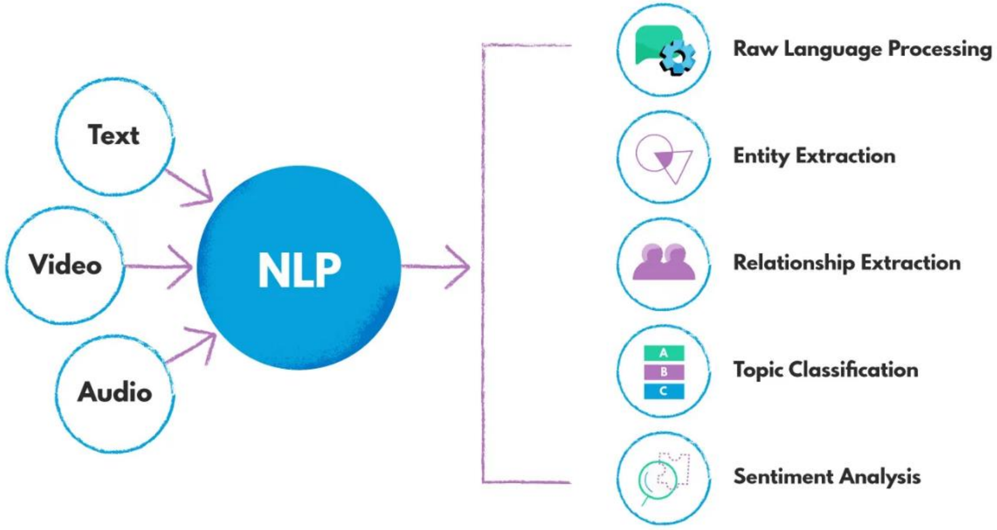


Natural Language Processing

Natural Language Processing (NLP) refers to the branch of computer science concerned with giving computers the ability to understand **text** and **spoken words** in much the same way human beings can.

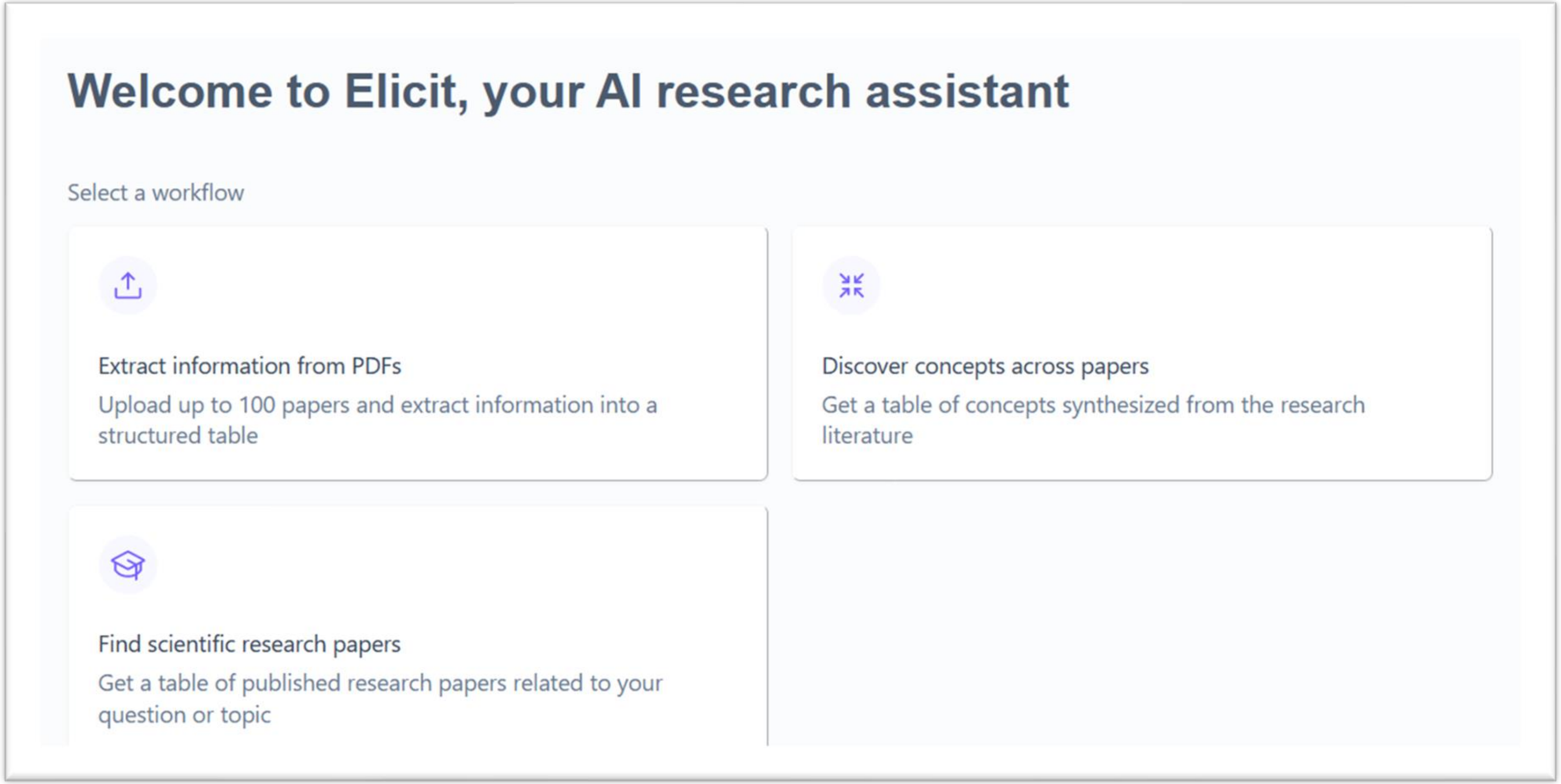
NLP combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models.

Large Language Models (LLMs) are Deep Learning models trained to produce text. With this impressive ability, LLMs have become the backbone of modern NLP.



Source: nexocode

Natural Language Processing



Welcome to Elicit, your AI research assistant

Select a workflow

Extract information from PDFs

Upload up to 100 papers and extract information into a structured table

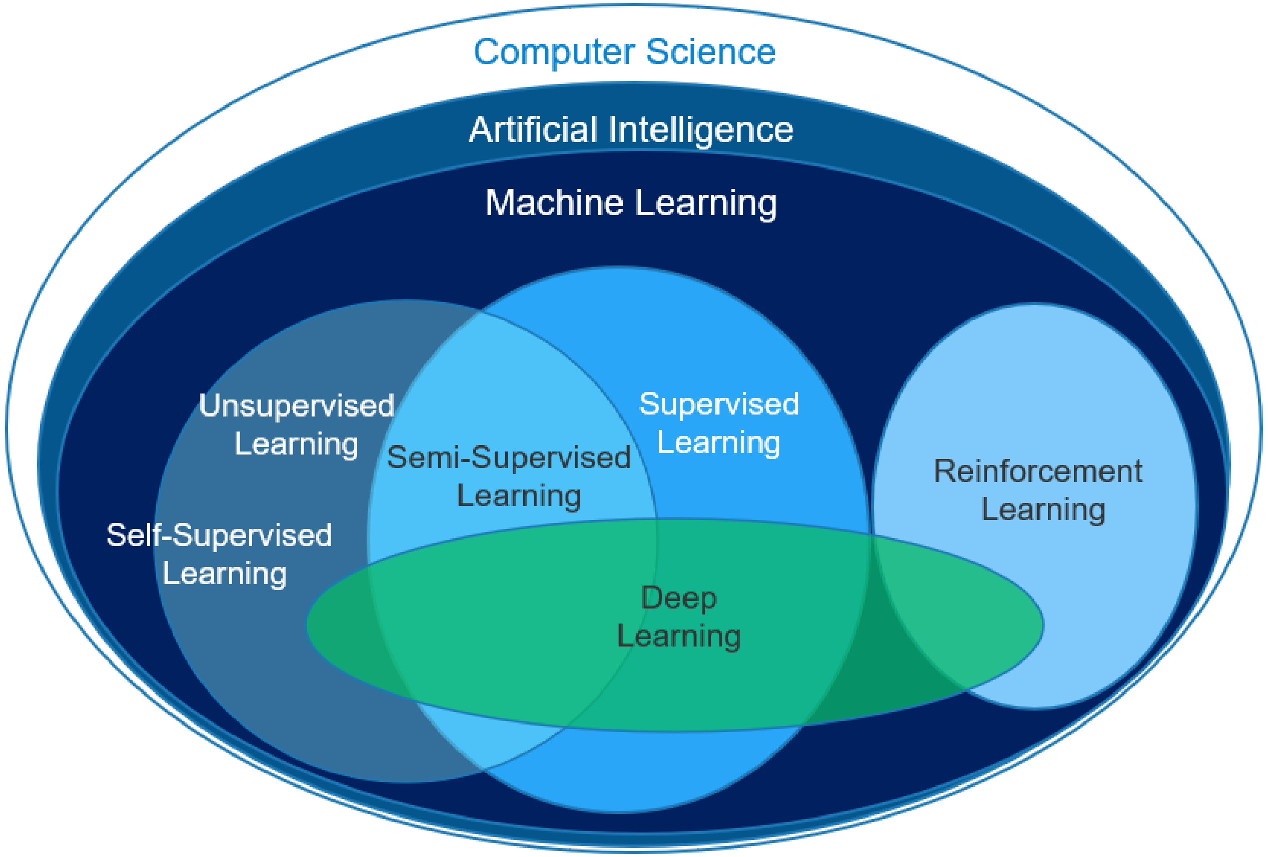
Discover concepts across papers

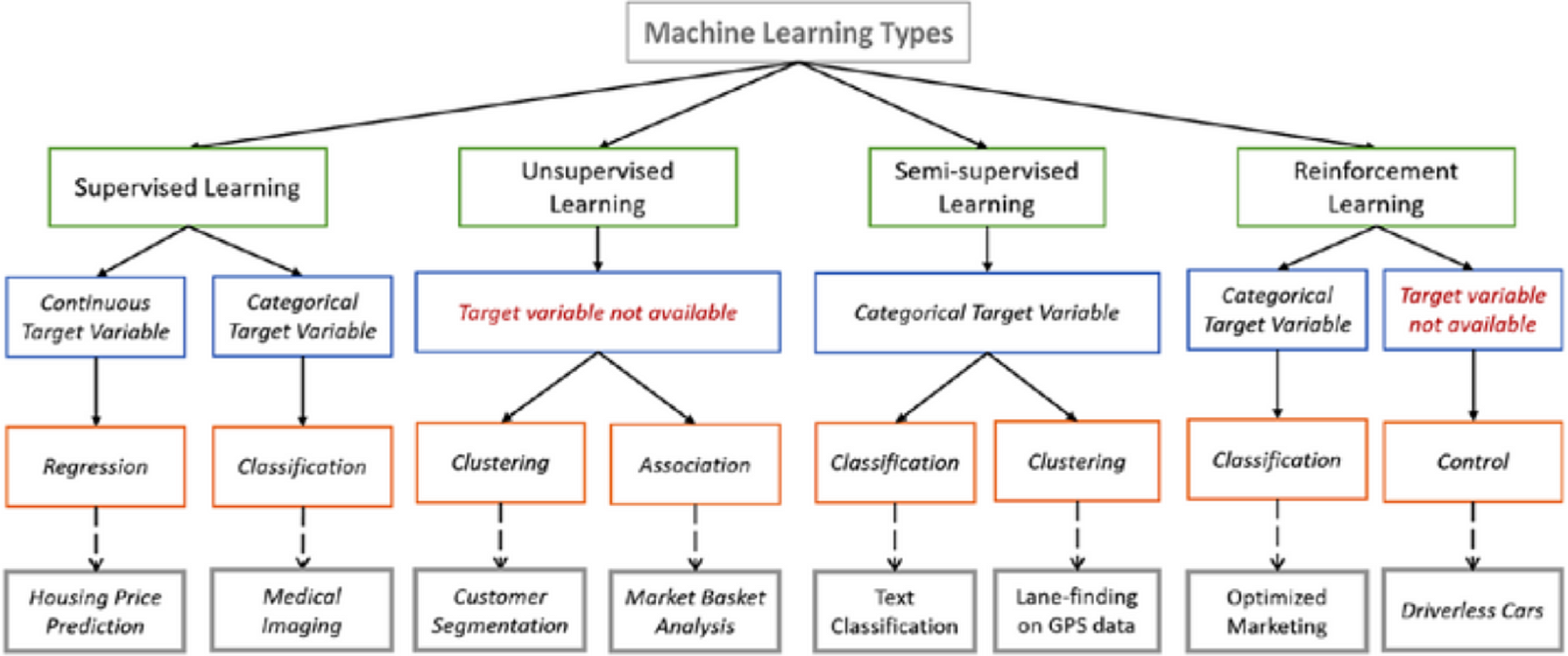
Get a table of concepts synthesized from the research literature

Find scientific research papers

Get a table of published research papers related to your question or topic

Types of Machine Learning




Machine Learning Types
Supervised Learning
Unsupervised Learning
Semi-supervised Learning
Reinforcement Learning
Continuous Target Variable
Categorical Target Variable
Target variable not available
Categorical Target Variable
Categorical Target Variable
Target variable not available
Regression
Classification
Clustering
Association
Classification
Clustering
Classification
Control
Housing Price Prediction
Medical Imaging
Customer Segmentation
Market Basket Analysis
Text Classification
Lane-finding on GPS data
Optimized Marketing
Driverless Cars

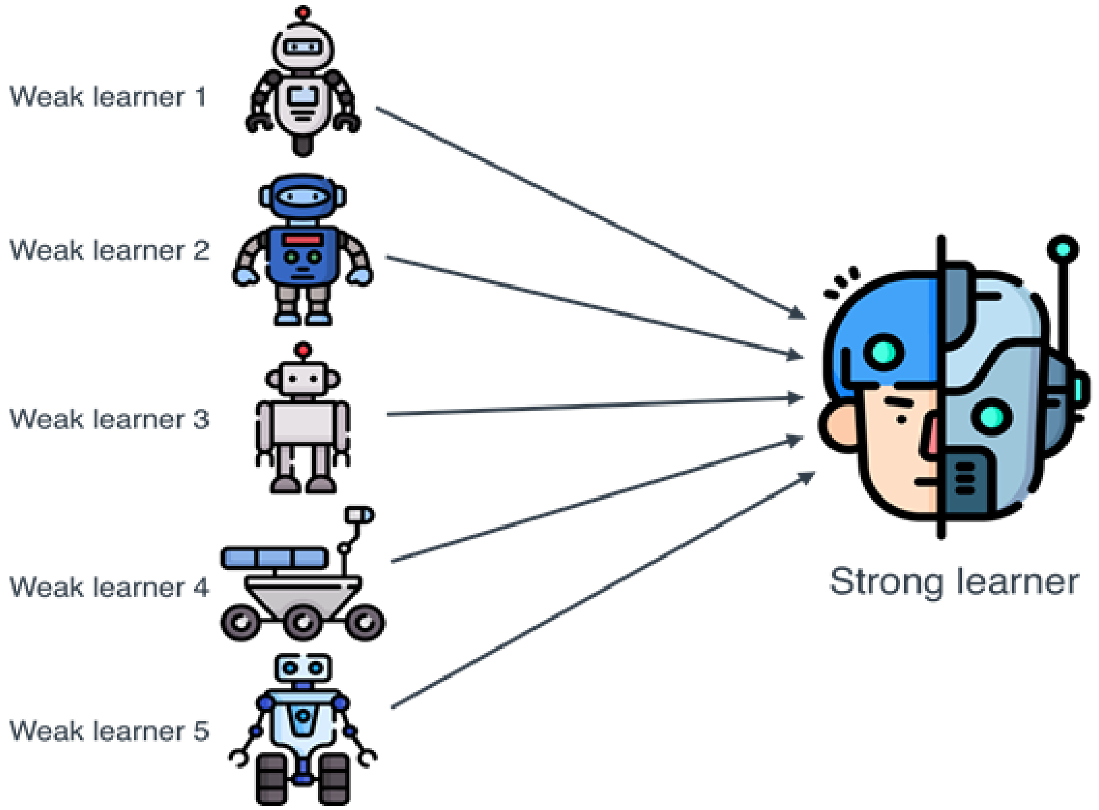
Ensemble methods

Ensemble methods

Aggregation methods or ensemble methods use an aggregation or a combination of a large number of learning models or algorithms.

It is empirically demonstrated that these methods significantly improve the predictive quality of individual nonlinear and unstable models or algorithms.

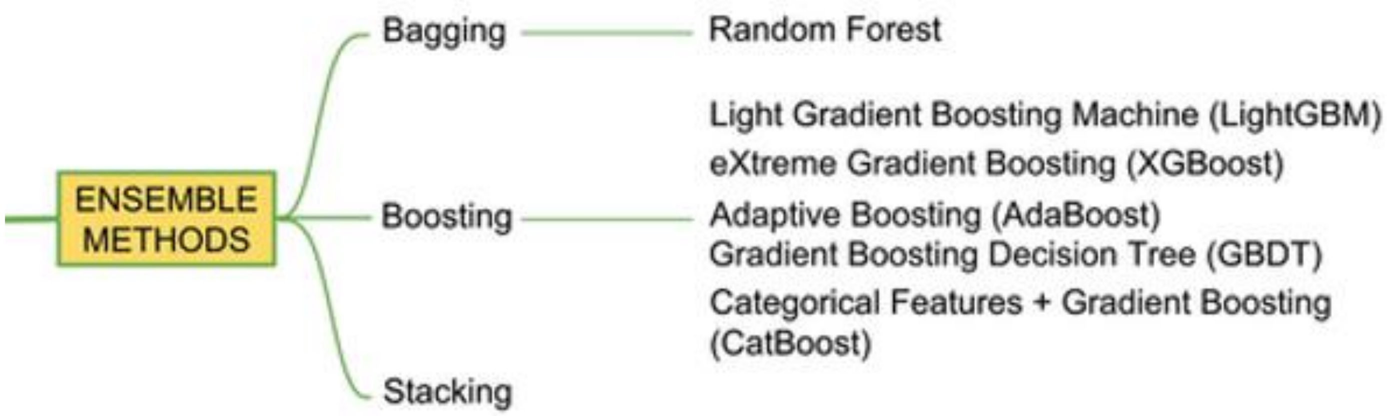
Ensemble methods generaly reduce the variance of the forecasts on the test set.



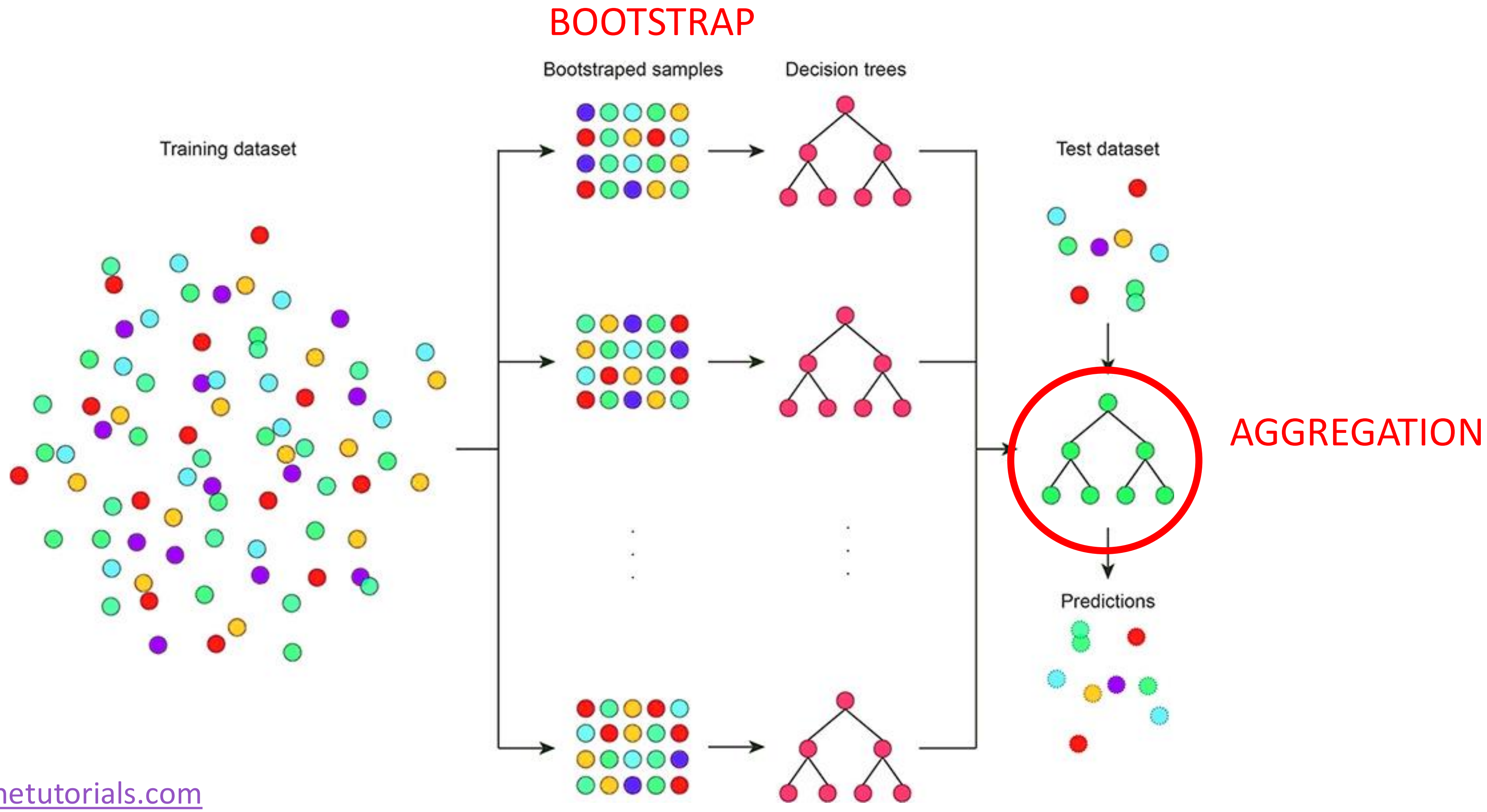
Set methods

There are three main classes of ensemble methods (or aggregation methods):

1. Boosting methods which rely on an adaptive strategy to "boost" predictive performance.
2. Bagging (Bootstrap Aggregation) methods , which use Boostrap sampling and aggregation to improve the predictive quality of models or algorithms with low generalization power, such as decision trees.

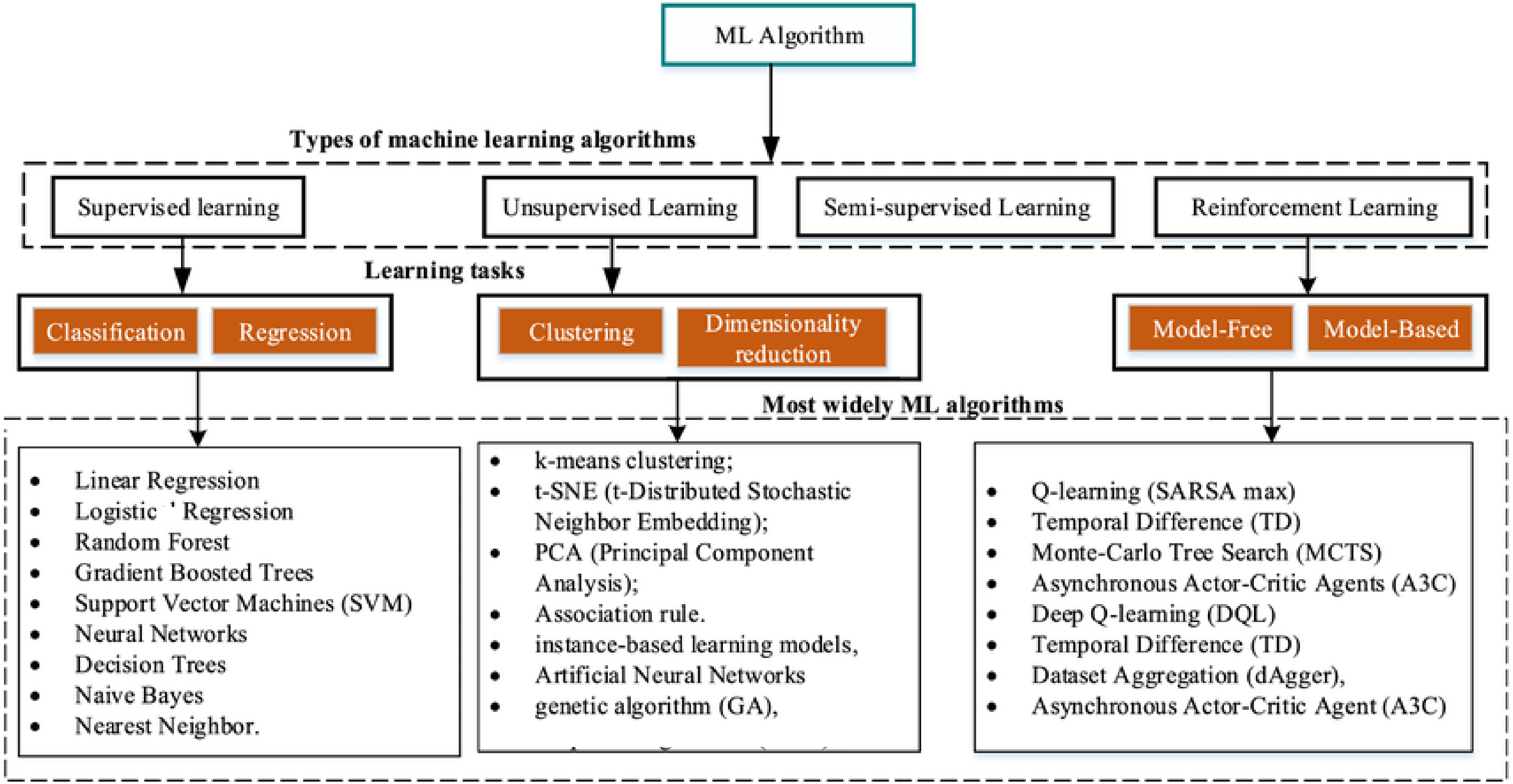

ENSEMBLE METHODS
Bagging
Random Forest
Boosting
Light Gradient Boosting Machine (LightGBM)
eXtreme Gradient Boosting (XGBoost)
Adaptive Boosting (AdaBoost)
Gradient Boosting Decision Tree (GBDT)
Categorical Features + Gradient Boosting (CatBoost)
Stacking

Bagging: general principle



Source: https://learnetutorials.com

Which algorithms do economists use?



ML Algorithm

Types of machine learning algorithms

Supervised learning

Unsupervised Learning

Semi-supervised Learning

Reinforcement Learning

Learning tasks

Classification

Regression

Clustering

Dimensionality reduction

Model-Free

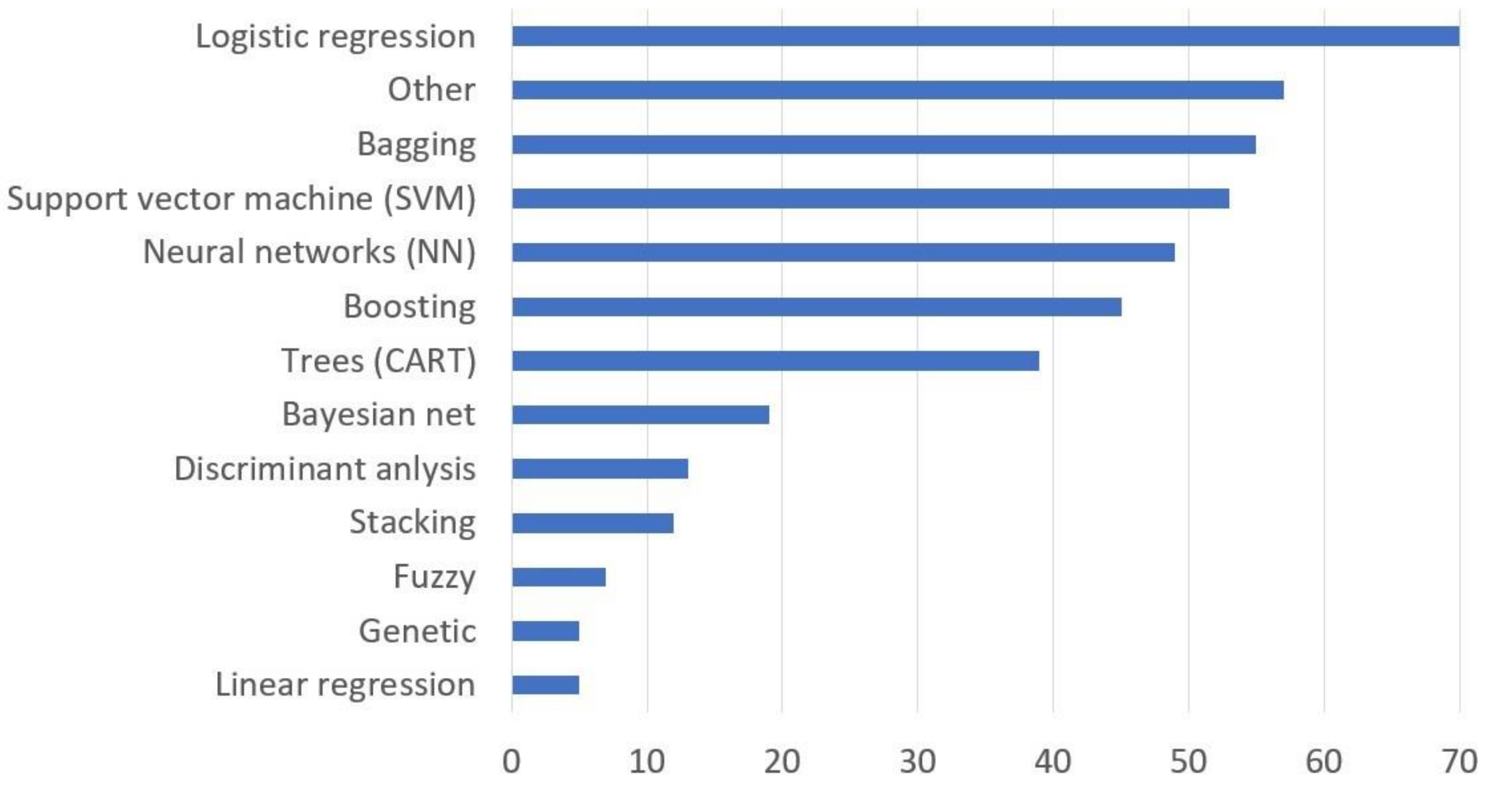
Model-Based

Most widely ML algorithms

- Linear Regression
- Logistic ' Regression
- Random Forest
- Gradient Boosted Trees
- Support Vector Machines (SVM)
- Neural Networks
- Decision Trees
- Naive Bayes
- Nearest Neighbor.

- k-means clustering;
- t-SNE (t-Distributed Stochastic Neighbor Embedding);
- PCA (Principal Component Analysis);
- Association rule.
- instance-based learning models,
- Artificial Neural Networks
- genetic algorithm (GA),

- Q-learning (SARSA max)
- Temporal Difference (TD)
- Monte-Carlo Tree Search (MCTS)
- Asynchronous Actor-Critic Agents (A3C)
- Deep Q-learning (DQL)
- Temporal Difference (TD)
- Dataset Aggregation (dAgger),
- Asynchronous Actor-Critic Agent (A3C)



(a) Note: This figure displays the number of times a given ML algorithm has been used in the 110 articles considered in the survey conducted by Markov et al. (2022). Each article can use more than one algorithm. Source: Markov, A., Seleznyova, Z., and Lapshin, V. (2022). Credit scoring methods: Latest trends and points to consider. Journal of Finance and Data Science, 8:180–201.

References for an introduction to ML

Books:

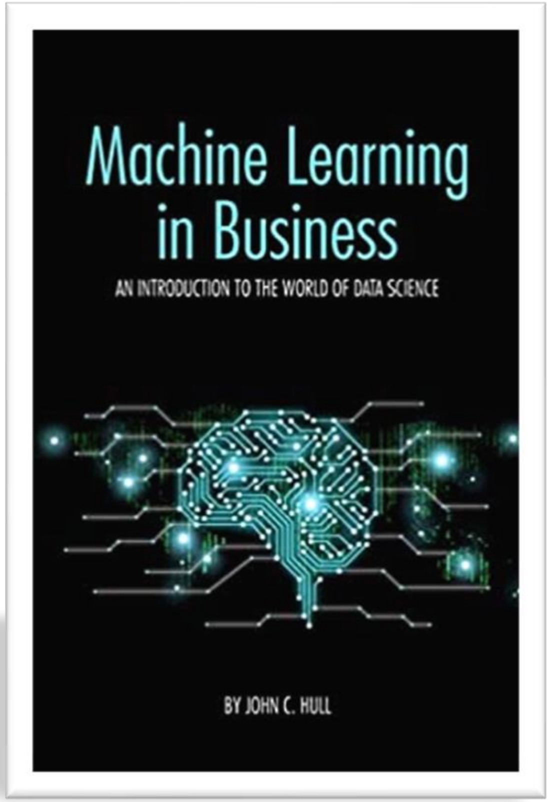
John C. Hull (2021), Machine Learning in Business: An Introduction to the World of Data Science

Surveys:

Sendhil Mullainathan and Jann Spiess (2017) , Machine Learning: An Applied Economtric Approach, *Journal of Economic Perspectives,* Vol. 31, No. 2, Spring 2017 , pp. 87–106.

Hal Varian (2014), "Big Data: New Tricks for Econometrics", Journal of Economic Perspectives, Spring, 3-28

Jon Kleinberg, H. Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan (2018), "Human Decisions and Machine Predictions", Quarterly Journal of Economics, 237-293.


Machine Learning in Business
AN INTRODUCTION TO THE WORLD OF DATA SCIENCE
BY JOHN C. HULL

Machine Learning advantages and limits for economists

Machine Learning and econometrics

ECONOMETRICS

Econometrics is focused on causal inference. Econometricians are interested in understanding the causal relationships between economic variables.

- ○The specification of the model (linear, etc.) is chosen by the economist.
- ○ The explicative variables of the model are generally selected by the economist.
- ○ The parameters can be interpreted economically.
- ○ There is a possibility to make inference on the parameters (confidence interval, standard error, Significance tests , etc.).

MACHINE LEARNING

ML is focused on predictive accuracy. ML algorithms are designed to make accurate predictions, without any regard for causal inference.

- ○ The specification of the model is chosen by the algorithm and the data => data driven approach.
- ○ The specification of the model is highly flexible and provides high forecasting accuracy.
- ○ The model features are selected by the algorithm among a potential huge number of variables.
- ○ There is no parameter (non-parametric approach).
- ○ There is no inference on the parameters => conformal prediction (NEW)

References for causal inference

Economic researchers have also proposed causal ML approaches (for instance causal random forest) to estimate the individual treatment effects, and so on:

For more details, see (among others):

Susan Athey, 2019: The Impact of Machine Learning on Economics, In: *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, May 2019, pp. 507–547

Susan Athey and Guido Imbens (2019), Machine Learning Methods Economists Should Know About, Annual Review of Economics 2019 11:1, 685-725

Alex Belloni, Victor Chernozhukov and Christian Hansen (2014), High-dimensional methods and inference on structural and treatment effects, Journal of Economic Perspectives, Spring, 29-50.

Advantages for economists

1. ML allows to consider large datasets
   - With a large number of variables (fat small dataset).
   - With a large number of instances (thin large dataset).
2. ML automates the selection of variables (features) in the model.
3. ML automatically chooses very flexible functional forms that capture the richness of the relationships between the data
   - Non-linearities
   - Interaction between the features.
4. ML can be applied to classification or regression problems.
5. ML is easy to implement (Python, R, etc.)

Limits

1. ML captures correlations and not causality relationships

   BUT causal inference with machine learning exists.

2. ML models may be not interpretable => **Blackbox models**

   BUT interpretable machine learning methods exist.

3. There is no parameter inference in ML model

   BUT conformal prediction methods give a confidence interval on the true value of the target

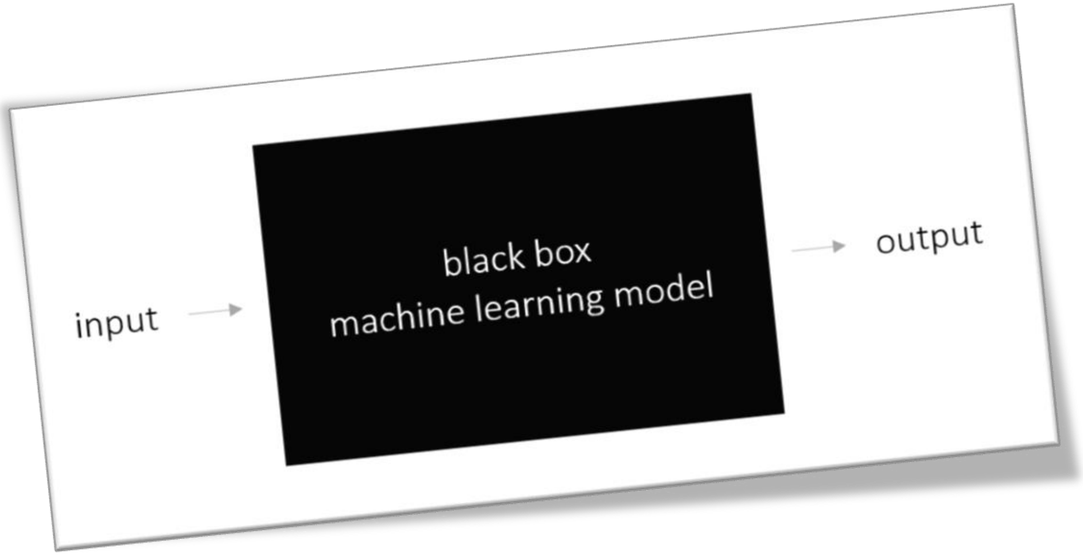
Blackbox models - interpretability

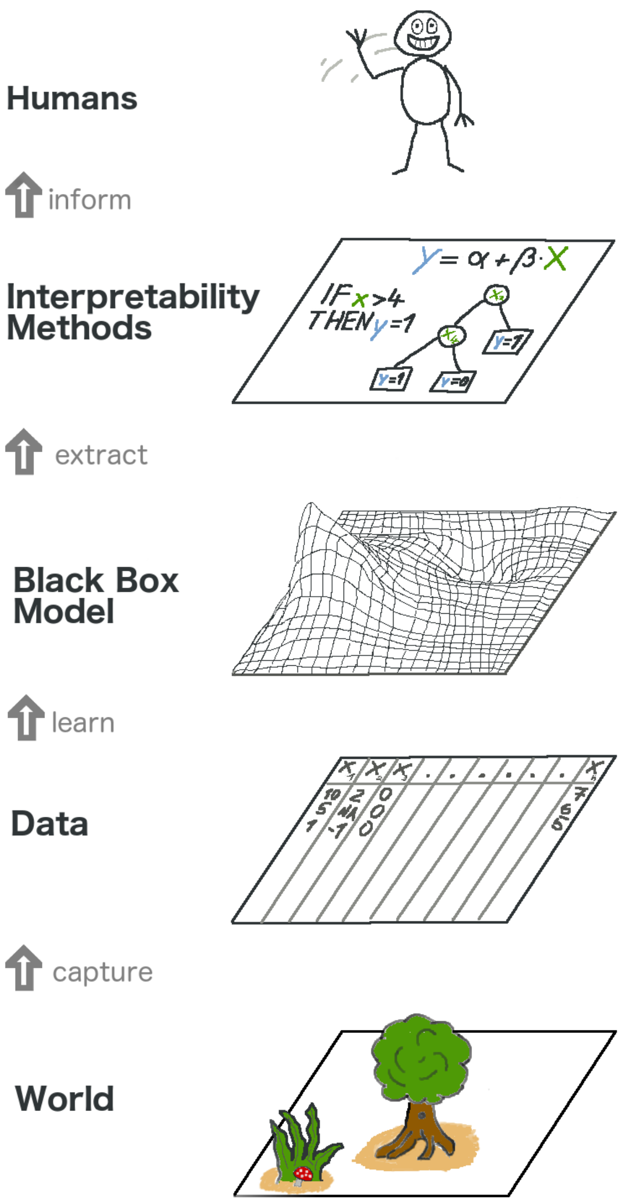
Black box: a model that is not transparent to the user. This means that the user cannot understand how the model makes its predictions

Interpretability => **overall** understanding of the model.

Explicability => **local** understanding of individual decisions made by the model.

Interpretable Machine Learning refers to methods and models that make the behavior and predictions of machine learning systems understandable to humans.





Source: Molnar (2019)

Interpretable ML methods

- Graphical tools (PDP, ALE, ICE, etc.) show the effect of an explanatory variable on the model.
- Feature importance measures reveal the relevance of each explanatory variable in the overall model.
- Shapley values is a solution concept in cooperative game theory, which is used to quantify the contribution of each feature to the model's prediction.
- LIMEs provide simple approximations of the model in the vicinity of an observation.
- Counterfactual explanations indicate how a specific prediction of the model could be modified by changing the values of the explanatory variables as little as possible.

Overfitting

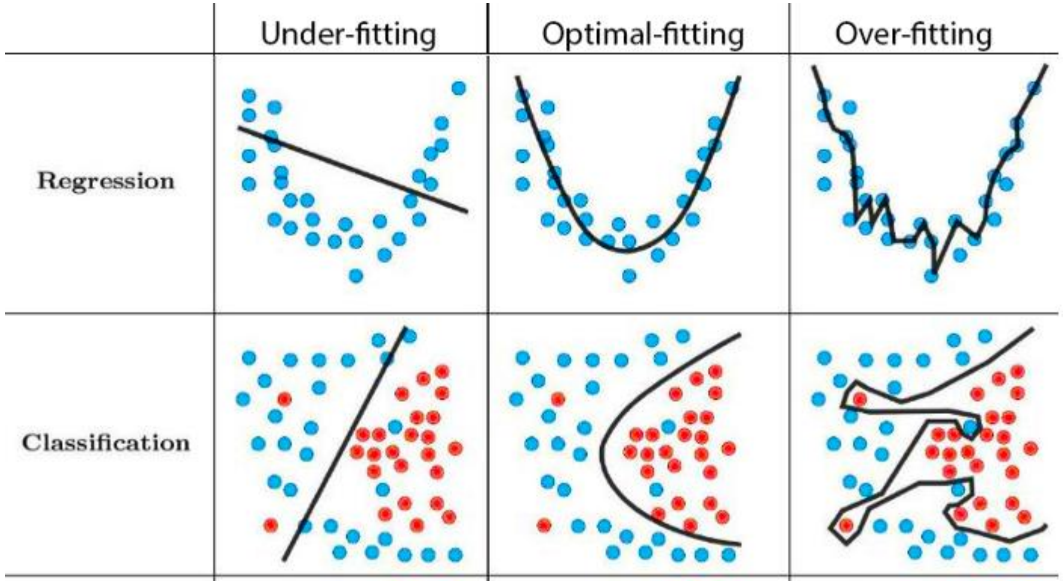
In statistics, overfitting denotes a statistical model that fits too closely to a particular collection of a data set.

Thus, this model may not correspond to additional data or reliably predict future observations.

=> **If a model has been trained too well on training data, it will be unable to generalize on the test set.**

=> Arbitrage between bias and variance of the predictions.

Overfitting occurs when the model adapts to noises in the training data.



Hyperparameters

Hyperparameters are the tuning parameters of machine learning algorithms .

The structure and predictive capabilities of an ML model are largely determined by the values of these hyperparameters.

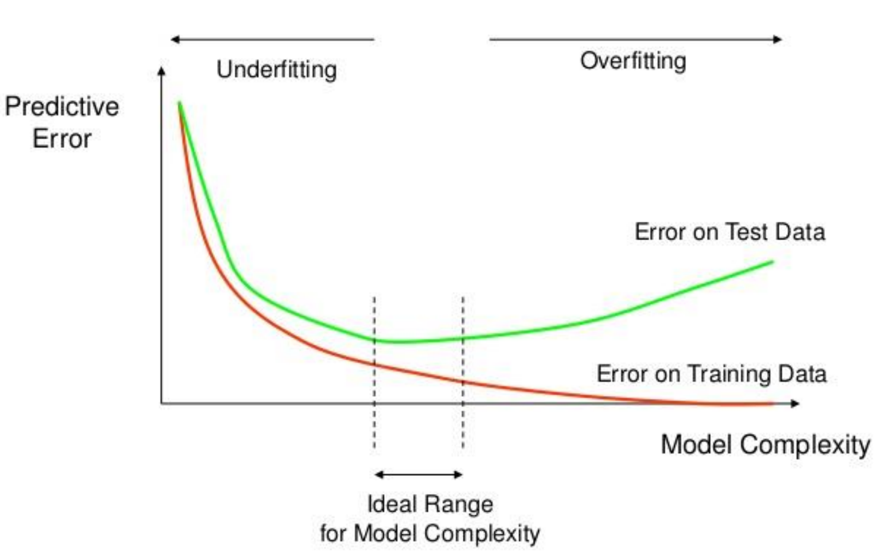
Small changes in these hyperparameters or in the way of determining them can cause large changes in the behavior of the model.

This is a significant source of operational risk in the implementation of ML.

Hyperparameters control the risk of overfiiting the model.

Hyperparameters => complexity => overfitting

How Overfitting affects Prediction



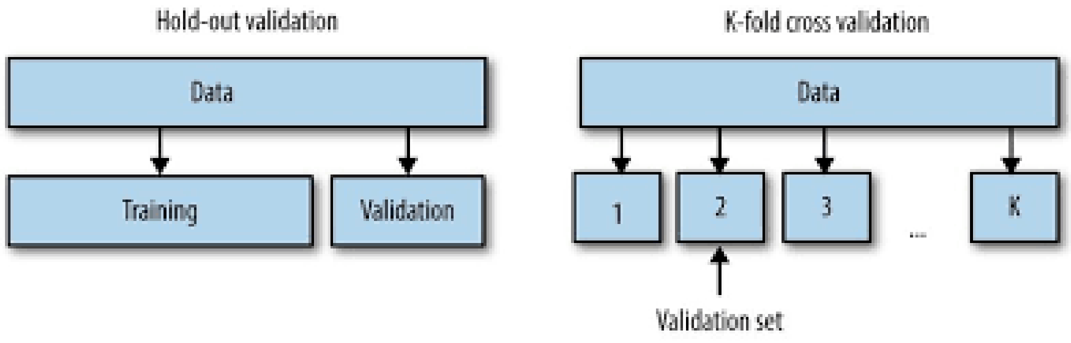
Cross-validation

To fix the value of the hyperparameters so as to control the risk of over-fitting , a method called cross-validation is generally used.

The cross-validation is a statistical method used to evaluate the performance of a machine learning model. It works by dividing the data into two or more parts: a training set and a test set. The training set is used to train the model, and the test set is used to evaluate the model's performance.

There are several variations:

- ○ Hold -out type cross-validation .
- ○ K- fold type cross-validation .
- ○ leave -one-out type cross- validation .



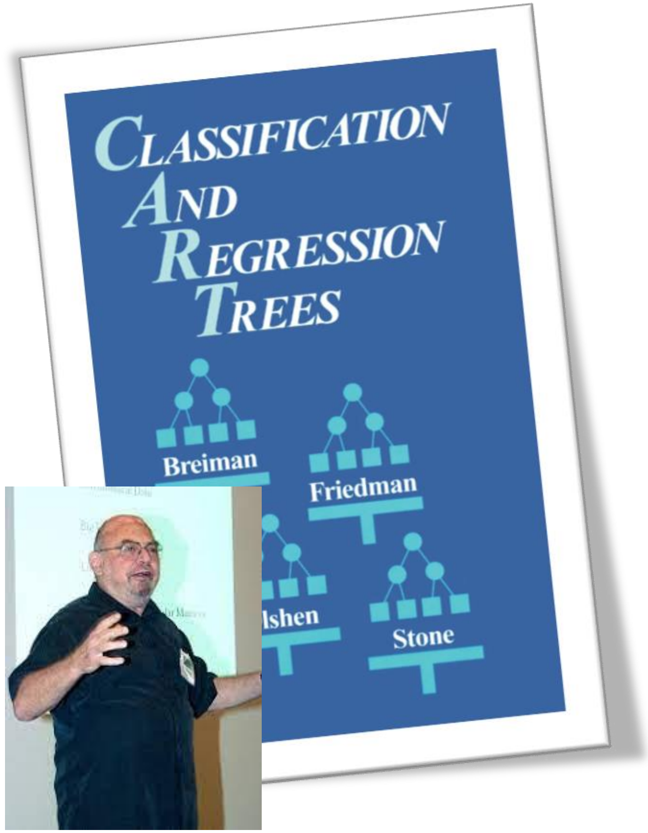
What should be a ML formation for economists?

ML background of economists

Most of the ML methods banks were developed between the mid-1980s and the early 2000s

- CART algorithm: Breiman et al (1984)
- Bagging methods: Breiman (1996).
- Random forests: Breiman (2001).
- XGBoost: Chen and Guestrin (2016) but the Boosting technique has its roots in an early publication by Freund and Schapire (1997).

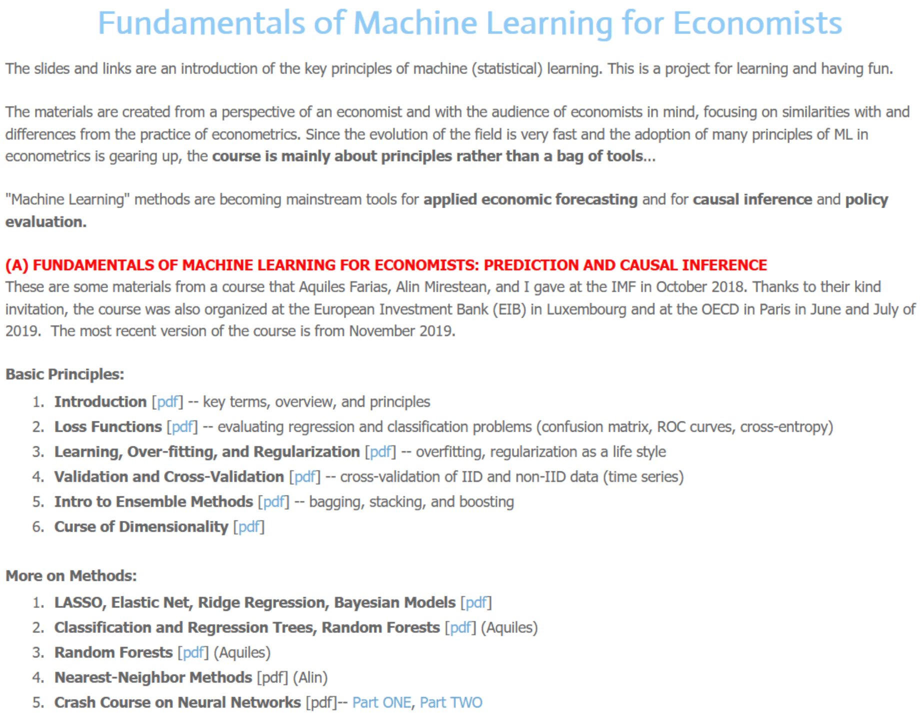
But ML was only introduced into economics and econometrics master's programs in the mid-2010s.



Source: Breiman, L., Friedman, J., Olshen, R. and C. Stone (1984), Classification and Regression Trees, Wadsworth, Int Group.

Two examples

IMF: FUNDAMENTAL OF ML FOR ECONOMISTS



Fundamentals of Machine Learning for Economists

The slides and links are an introduction of the key principles of machine (statistical) learning. This is a project for learning and having fun.

The materials are created from a perspective of an economist and with the audience of economists in mind, focusing on similarities with and differences from the practice of econometrics. Since the evolution of the field is very fast and the adoption of many principles of ML in econometrics is gearing up, the **course is mainly about principles rather than a bag of tools...**

"Machine Learning" methods are becoming mainstream tools for **applied economic forecasting** and for **causal inference** and **policy evaluation.**

(A) FUNDAMENTALS OF MACHINE LEARNING FOR ECONOMISTS: PREDICTION AND CAUSAL INFERENCE
These are some materials from a course that Aquiles Farias, Alin Mirestean, and I gave at the IMF in October 2018. Thanks to their kind invitation, the course was also organized at the European Investment Bank (EIB) in Luxembourg and at the OECD in Paris in June and July of 2019. The most recent version of the course is from November 2019.

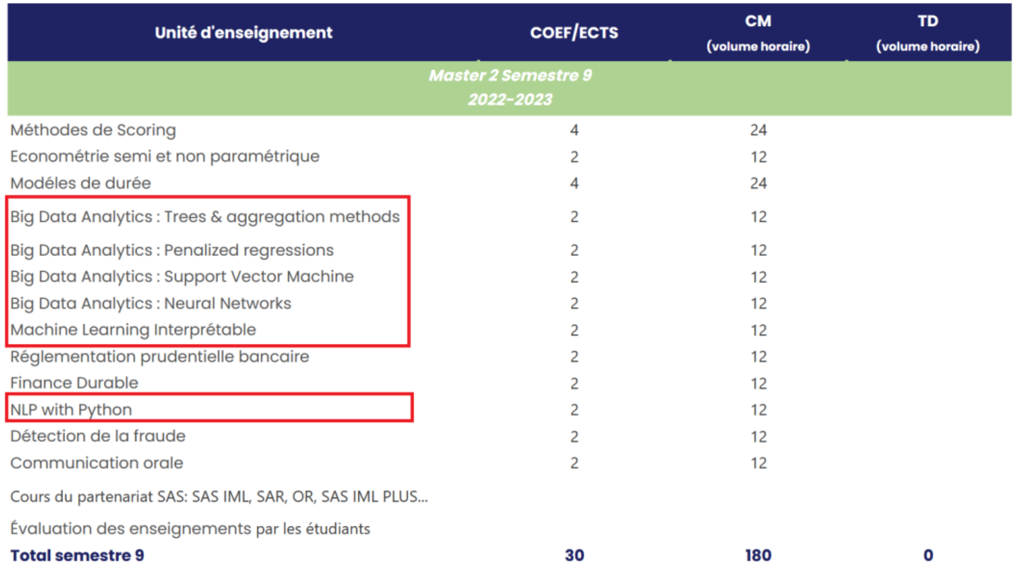
Basic Principles:

1. **Introduction** [pdf] -- key terms, overview, and principles
2. **Loss Functions** [pdf] -- evaluating regression and classification problems (confusion matrix, ROC curves, cross-entropy)
3. **Learning, Over-fitting, and Regularization** [pdf] -- overfitting, regularization as a life style
4. **Validation and Cross-Validation** [pdf] -- cross-validation of IID and non-IID data (time series)
5. **Intro to Ensemble Methods** [pdf] -- bagging, stacking, and boosting
6. **Curse of Dimensionality** [pdf]

More on Methods:

1. **LASSO, Elastic Net, Ridge Regression, Bayesian Models** [pdf]
2. **Classification and Regression Trees, Random Forests** [pdf] (Aquiles)
3. **Random Forests** [pdf] (Aquiles)
4. **Nearest-Neighbor Methods** [pdf] (Alin)
5. **Crash Course on Neural Networks** [pdf]-- Part ONE, Part TWO

MASTER PROGRAM IN ECONOMETRICS (UNIV. ORLEANS)



Unité d'enseignement	COEF/ECTS	CM (volume horaire)	TD (volume horaire)
Master 2 Semestre 9 *2022-2023*			
Méthodes de Scoring	4	24	
Econométrie semi et non paramétrique	2	12	
Modéles de durée	4	24	
Big Data Analytics : Trees & aggregation methods	2	12	
Big Data Analytics : Penalized regressions	2	12	
Big Data Analytics : Support Vector Machine	2	12	
Big Data Analytics : Neural Networks	2	12	
Machine Learning Interprétable	2	12	
Réglementation prudentielle bancaire	2	12	
Finance Durable	2	12	
NLP with Python	2	12	
Détection de la fraude	2	12	
Communication orale	2	12	
Cours du partenariat SAS: SAS IML, SAR, OR, SAS IML PLUS...			
Évaluation des enseignements par les étudiants			
Total semestre 9	**30**	**180**	**0**

Contact

Christophe HURLIN
University of Orleans
University Institute of France
christophe.hurlin@univ-orleans.fr