About me

Toufic ZARAKET

Senior Consultant @ **EURODECISION**

- Data & Business analytics

PhD in Industrial Engineering (Ecole Centrale Paris)

EURODECISION
OPERATIONAL RESEARCH

About me

Hakim IDJIS

Senior Data Scientist @ **CAPGEMINI Invent**

PhD in Industrial Engineering (Ecole Centrale Paris)

Kaggle Master

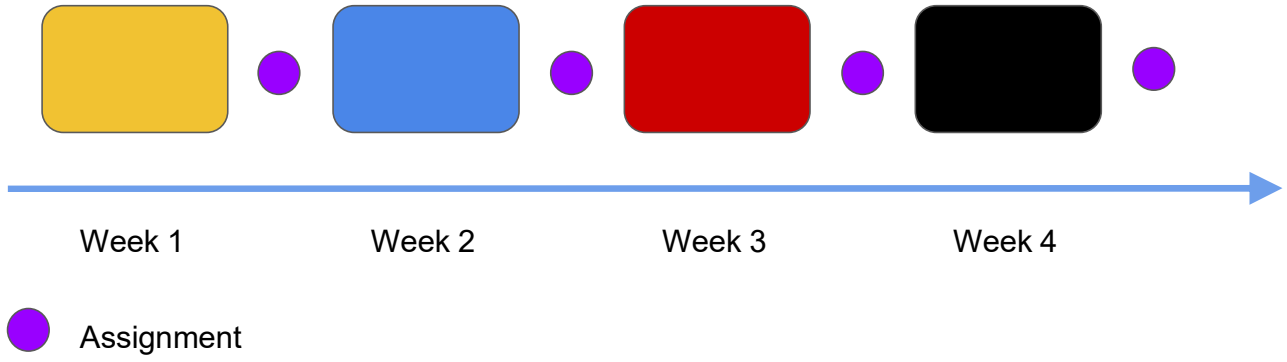
What is this course about

- Learn about Data Science
- Learn about machine learning and its applications
- How to build machine learning systems
- How the algorithms behind them work
- How to use those algorithms

Course planning

A Case study approach:

- Course
- Practical work (case study)



Course overview

1. Week 1: Introduction to Data Science and Machine Learning
    1. Introduction to Data Science
    2. Introduction to Machine Learning
    3. Machine Learning Tools
2. ...

1.1
Introduction to Data Science

The Era of Big Data

- **90%** of the information ever generated was generated in the last two years?

Every minute we send 2014 million emails, generate 1,8 million Facebook likes, send 278,000 Tweets, and upload 200,000 photos to Facebook
Source

Around 100 hours of video are uploaded to Youtube every minute and it would take you around 15 years to watch every video uploaded by users in one day
Source

If you burned all of the data created in just one day onto DVD's, you would stack them on top of each other and reach the moon - twice
Source

- This growing torrent of data + growing storage and computation capacity (cloud) ⇒ **Big Data Era**

What is Data Science ?

- It goes back a little further than 2004, which is where the Google search term history begins
- Data Science is not just limited to tech companies

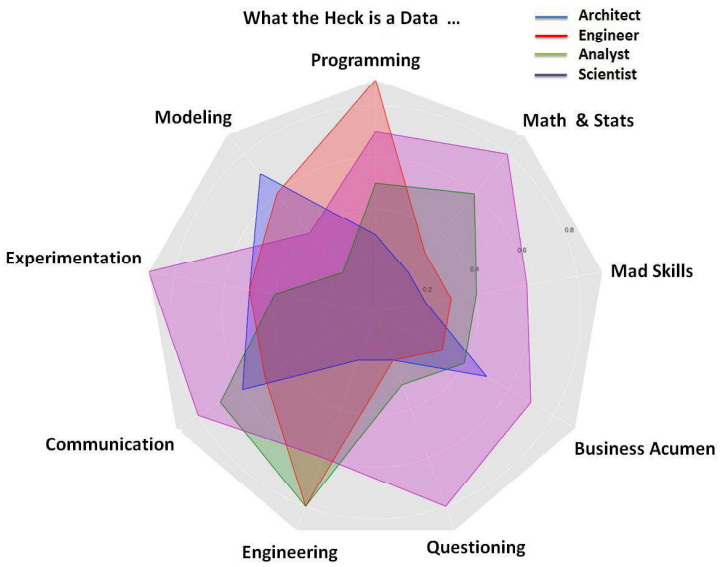
  Almost every company is turning to data science to better understand how to build products, serve customers and leverage new opportunities

- Data Science is used in multiple disciplines: computer science, behavioural sciences, law & business, etc..
- All of these actors need data-driven methodologies to aid in their discovery:

  From statistical analysis, machine learning, & text mining to information visualization

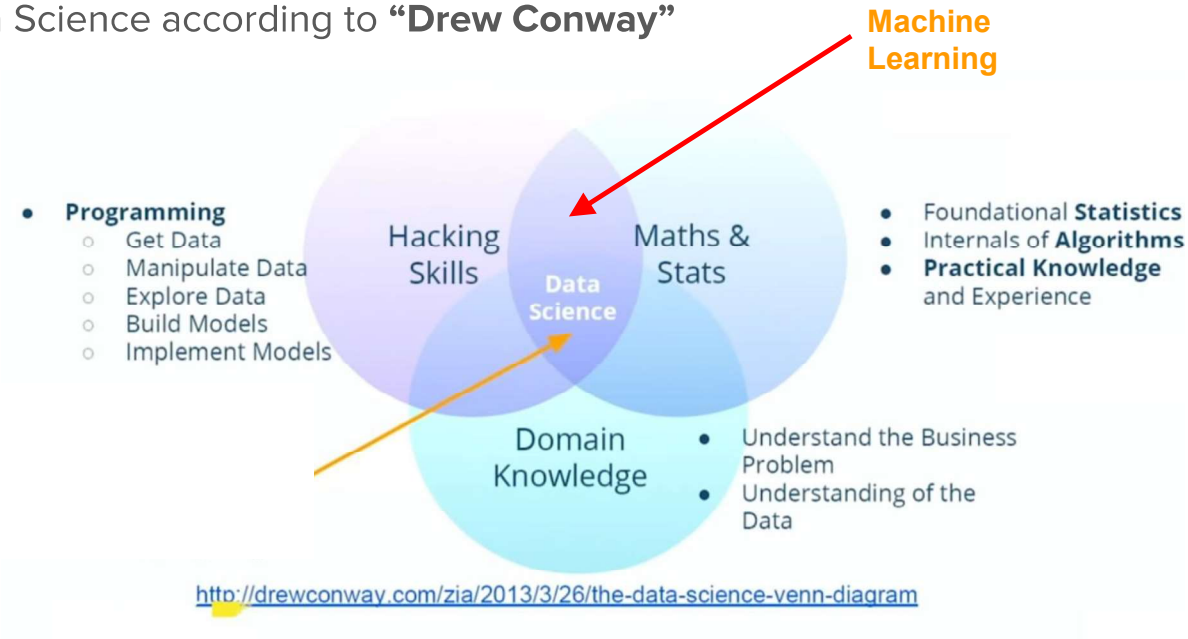
What is Data Science

Data Science is an umbrella term and it's basically the marriage of many different fields.



What is Data Science

Definition of Data Science according to **"Drew Conway"**



What is Data Science

Data Science

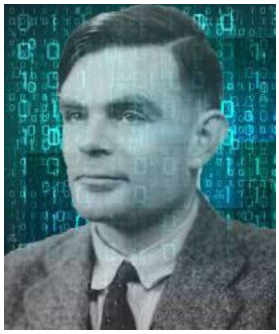
- David Donoho, "50 Years of Data Science"
  1. *Data Exploration and Preparation*
  2. *Data Representation and Transformation*
  3. *Computing with Data*
  4. *Data Modeling*
  5. *Data Visualization and Presentation*

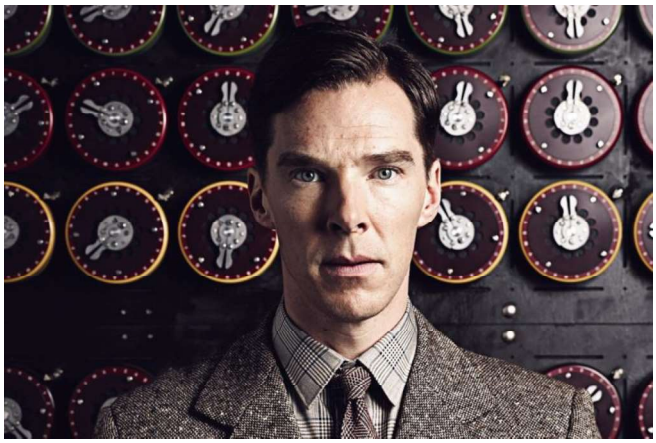
Sources: http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

1.2
Introduction to Machine Learning

What is Machine Learning ?

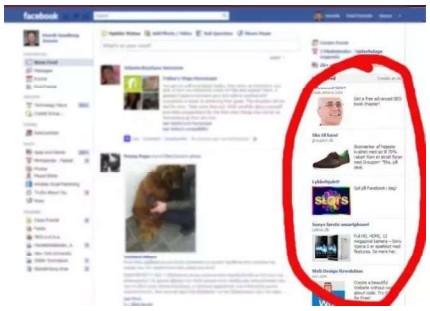
- Researchers interested in artificial intelligence wanted to see if computers could learn from data
- **ML is not a new science**: many machine learning algorithms have been around for a long time





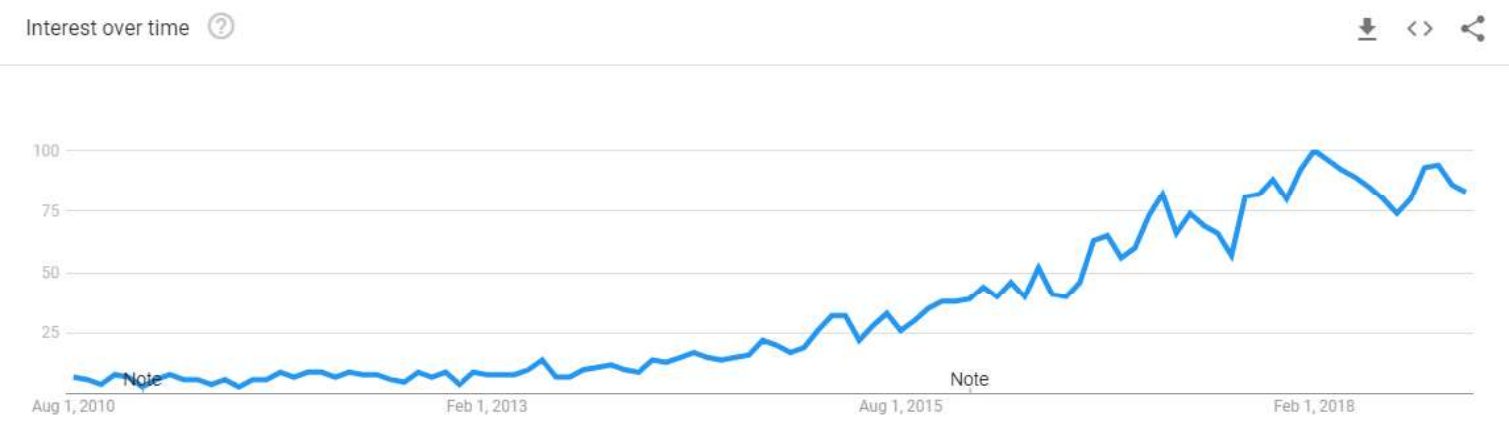
What is Machine Learning ?

- BUT – **it is a science that's gaining fresh momentum**: the ability to automatically apply complex mathematical calculations to **big data – over and over, faster and faster** – is a recent development



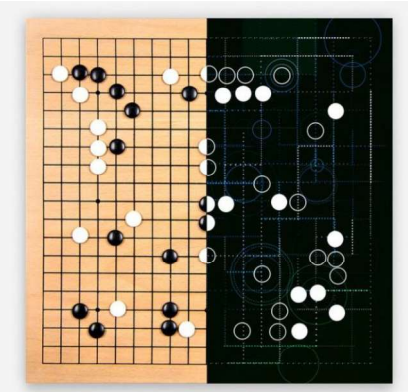
What is Machine Learning ?

Google trends for the term “Machine Learning”



What is Machine Learning ?

- Google AI beats **Go** world champion
  - historic 4-1 series victory
- How?
  - "...The software combines good old-fashioned neural network algorithms and machine-learning techniques with superb software engineering" (scientific american magazine)
  - At the heart of the computations are neural networks



Source: https://www.scientificamerican.com/article/how-the-computer-beat-the-go-master/

Definition of Machine Learning

- Machine learning is the subfield of computer science that "gives computers the ability to learn **without being explicitly programmed**" (Arthur Samuel, 1959)

- A more modern definition by Tom Mitchell: "**A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.**"

Example: playing checkers.

- E = the experience of playing many games of checkers
- T = the task of playing checkers.
- P = the probability that the program will win the next game.

Machine Learning in Practice

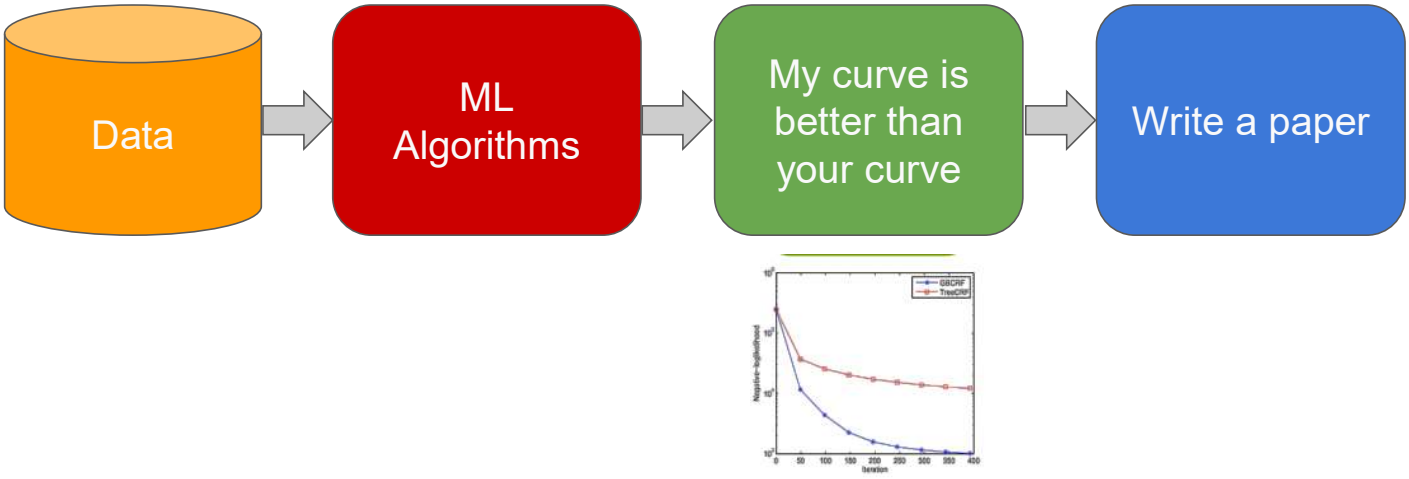
Image recognition : https://www.clarifai.com/demo

https://www.autodraw.com/

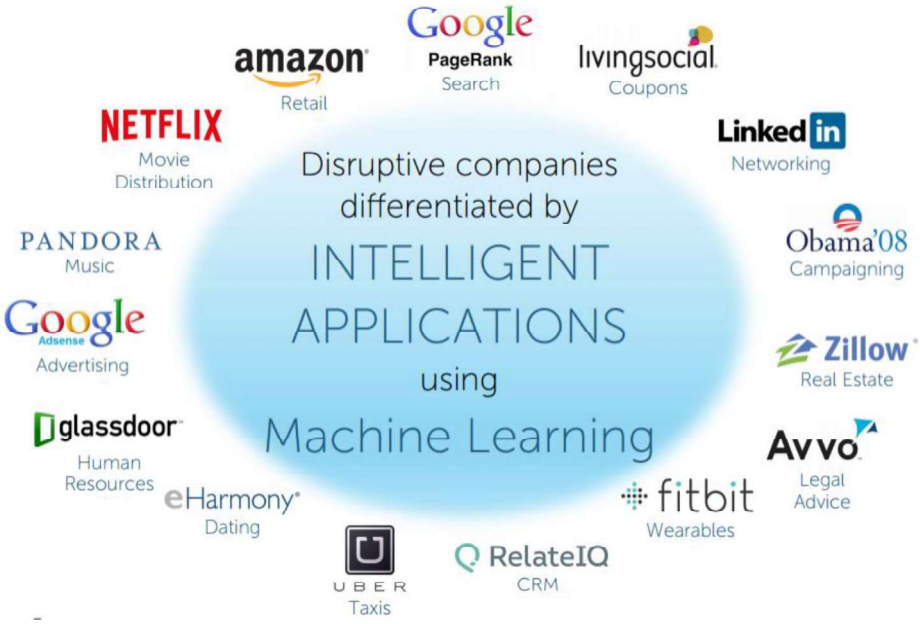
Sentiment analysis: https://azure.microsoft.com/fr-fr/services/cognitive-services/text-analytics/

Check out other demos : https://experiments.withgoogle.com/ai

Old View of Machine Learning



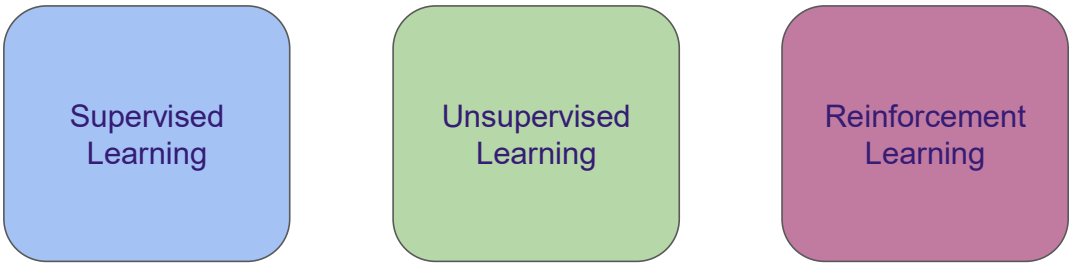
Machine Learning in Intelligent Applications



The pipeline of Machine Learning

Types of Machine Learning

- Machine learning tasks are typically classified into **three broad categories**



- Depending on the nature of the learning "signal" or "feedback" available to a learning system

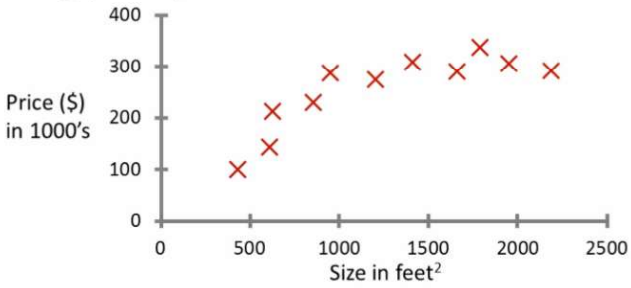
Supervised Learning

- The program is given a data set and already **know what our correct output** should look like
  - Having the idea that there is a relationship between the input and the output
- The goal is to learn a general rule that maps inputs to outputs.

Regression

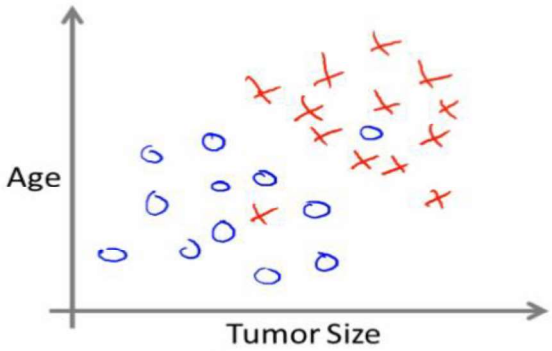
- ❑ **Predict results within a continuous output,**
- ❑ **⇒ map input variables to some continuous function**

Housing price prediction.



Classification

- ❑ **predict results in a discrete output.**
- ❑ **⇒ map input variables into discrete categories**



Unsupervised Learning

- **No labels** are given to the learning algorithm, leaving it on its own to find structure in its input.
- Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning)

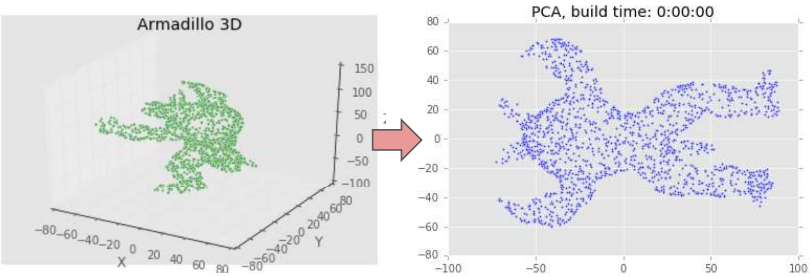
Clustering

- ❑ **Group similar samples into sets.**
- ❑ **⇒ Find structure within the data**

Customer segmentation

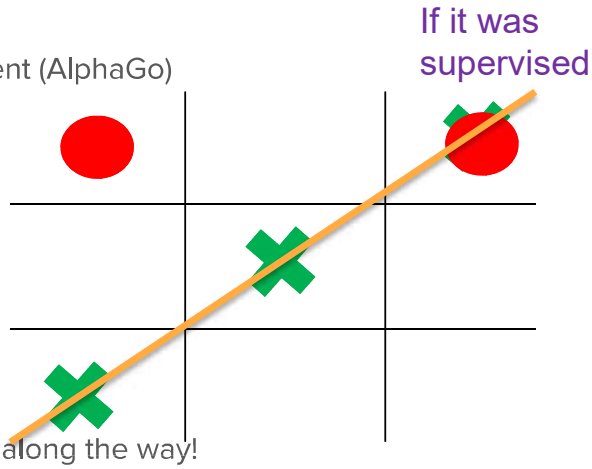
Dimensionality Reduction

- ❑ **Intelligently reduce the number of features considered**
- ❑ **⇒ Data compression, or Data visualization**



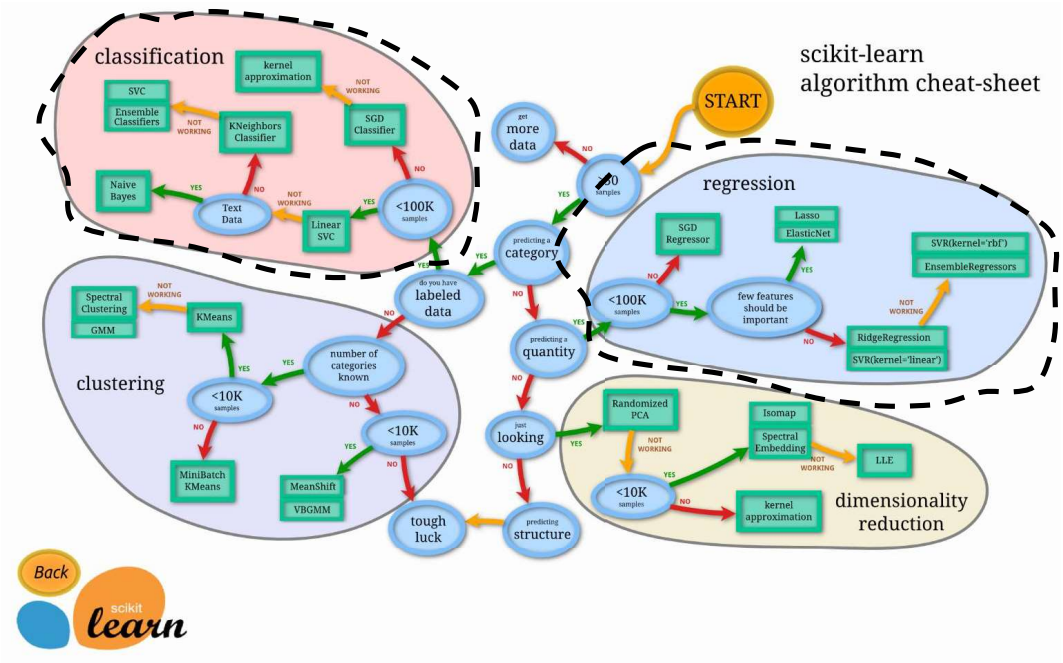
Reinforcement Learning

- A computer program interacts with a dynamic environment in which it must perform a certain goal, without a teacher explicitly telling it whether it has come close to its goal.
    - Learning to drive a car (Google Car)
    - Learning to play a game by playing against an opponent (AlphaGo)

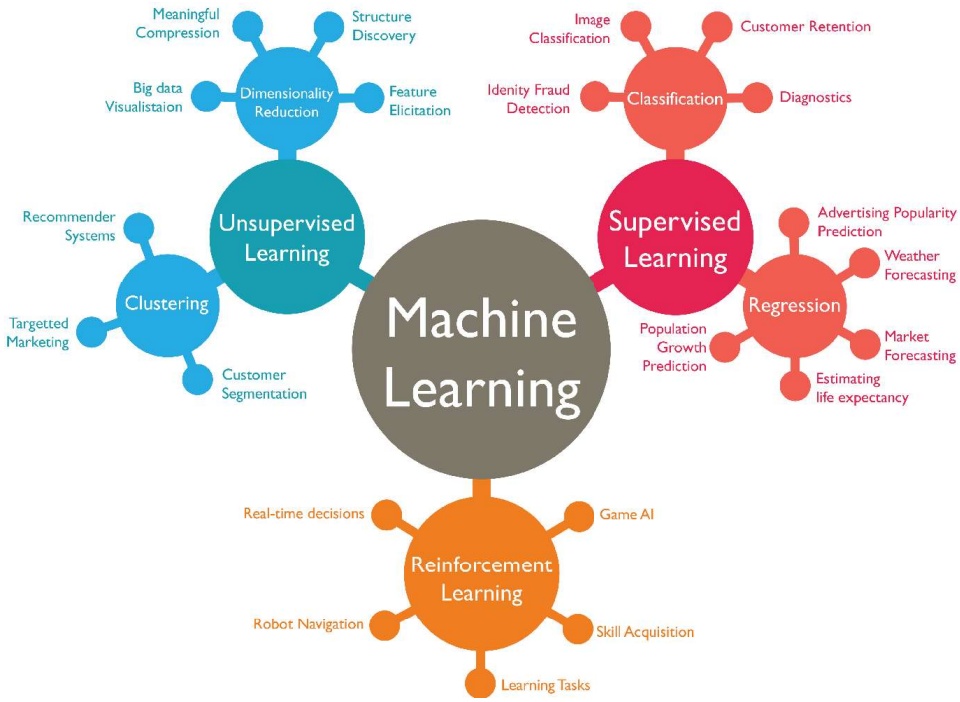


- Player  learns that he made a mistake somewhere along the way!

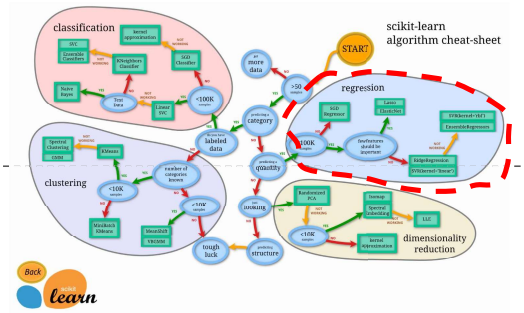
Machine Learning Algorithms



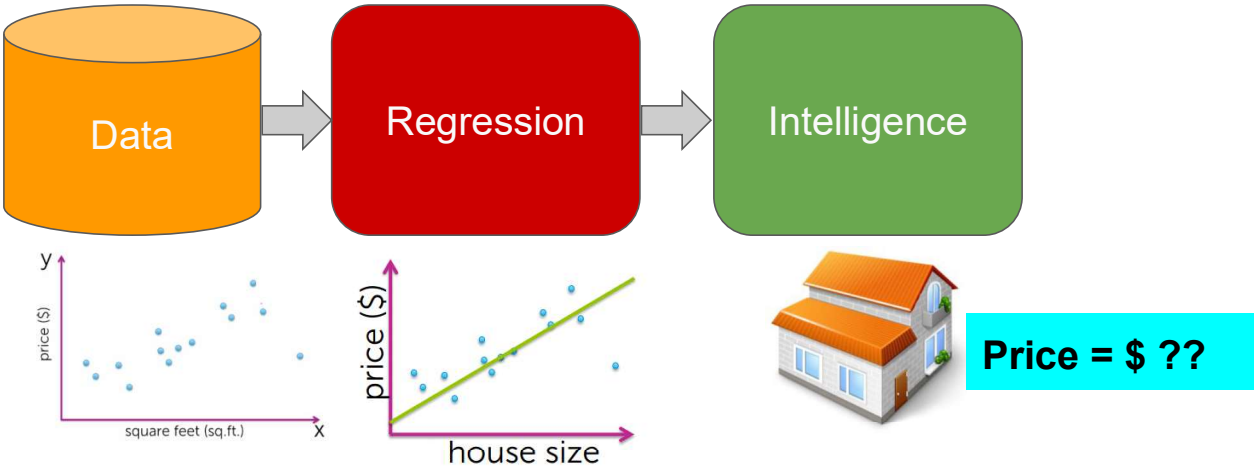
Machine Learning Applications



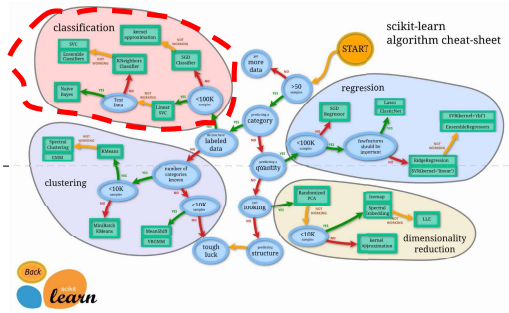
Machine Learning in this course



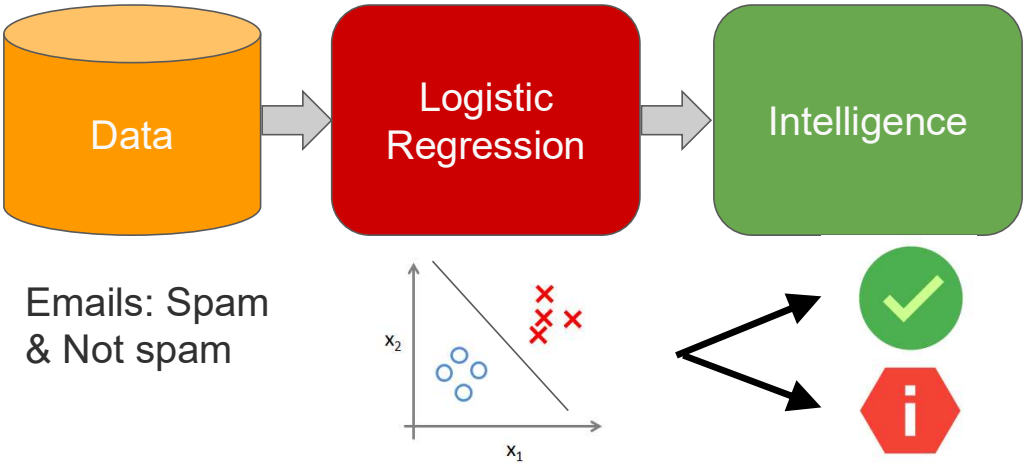
- Regression



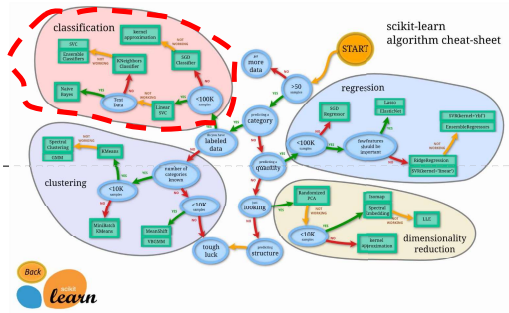
Machine Learning in this course



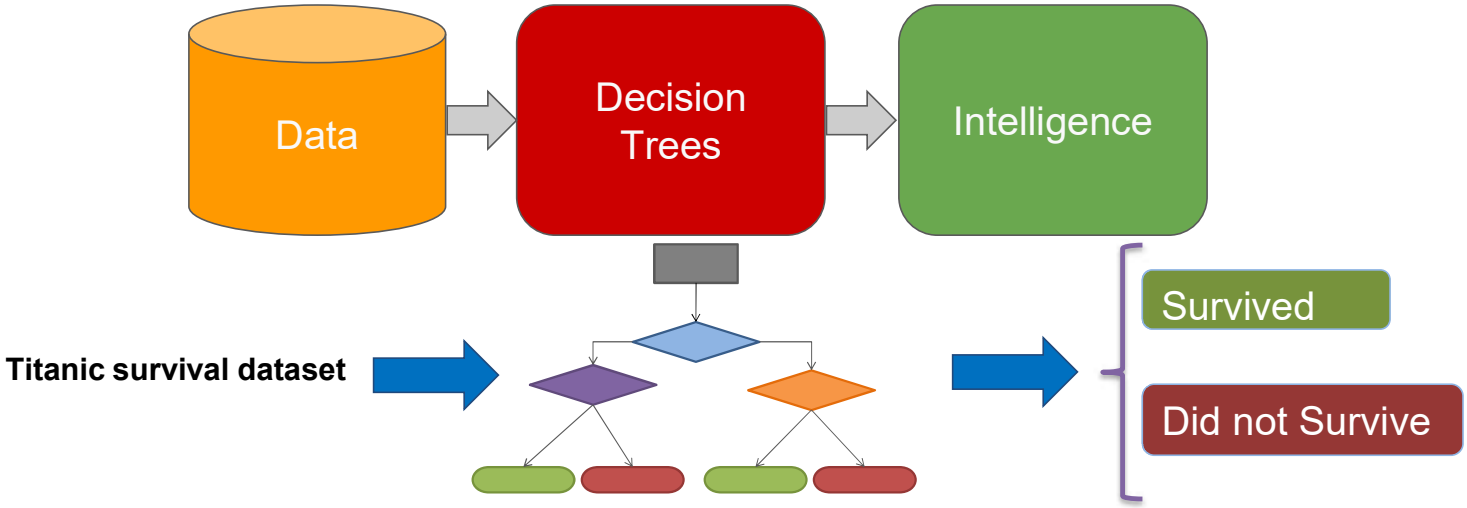
- ➢ **Classification (Logistic Regression)**



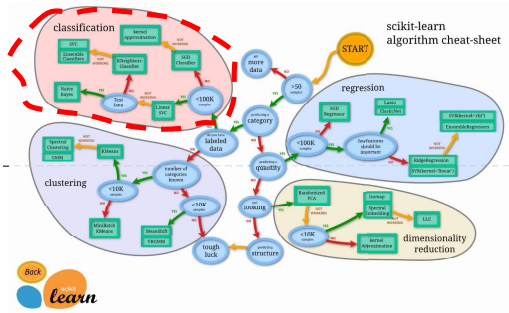
Machine Learning in this course



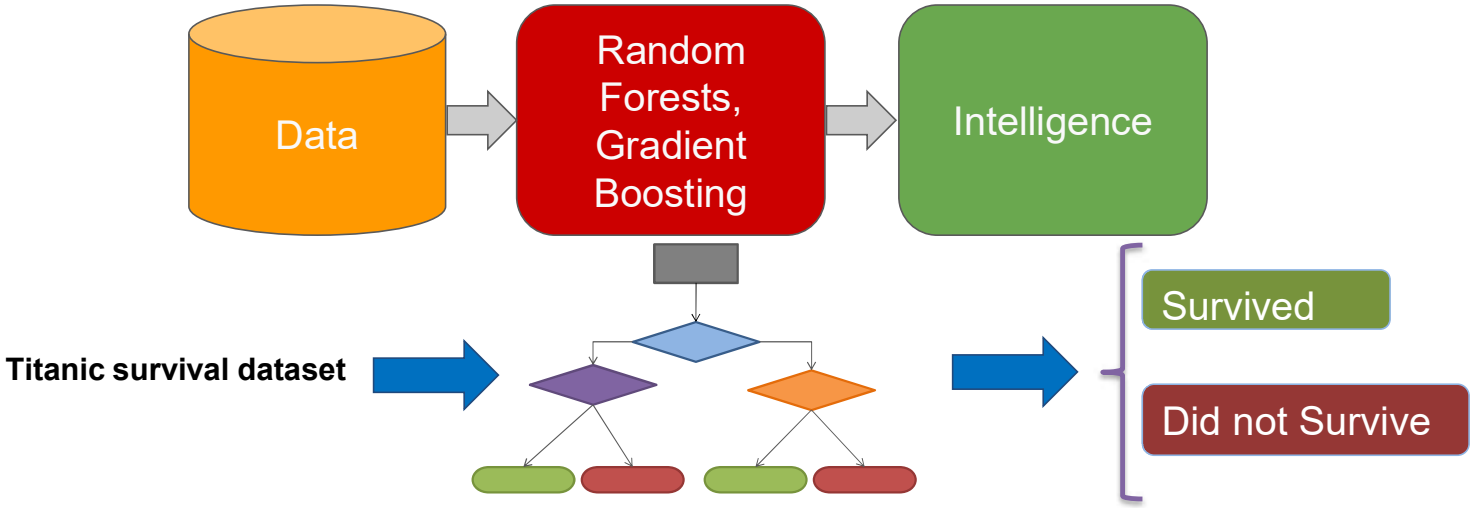
- Classification (Decision Trees)



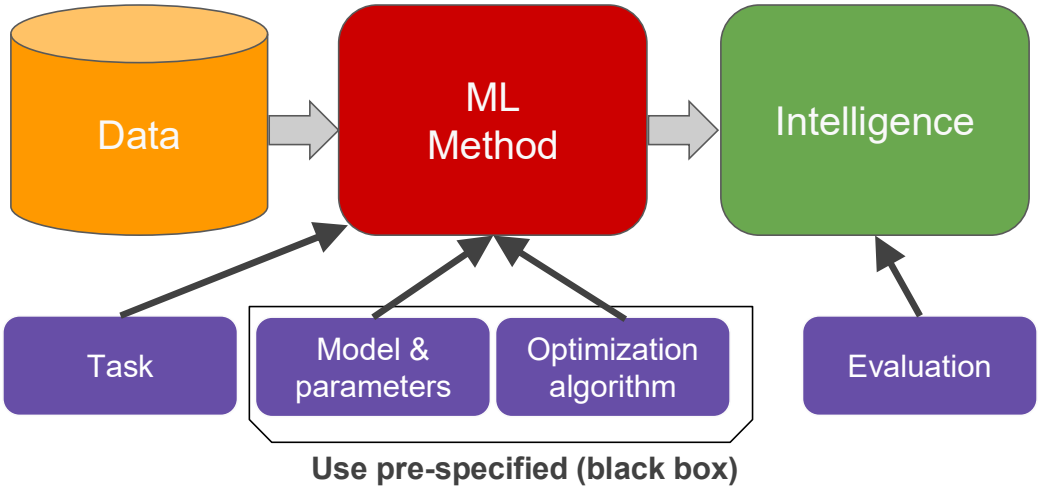
Machine Learning in this course



- Ensemble methods



Machine Learning in this course



Course overview .....

1. Week 1: Introduction to Data Science and Machine Learning
2. Week 2: Univariate & Multivariate Linear **Regression**
3. Week 3: Logistic Regression (**Classification**)
4. Week 4: Decision Trees (**Regression & Classification)**

1.3
Machine Learning Tools

Machine Learning Tools

- Machine Learning programming languages
- Machine Learning Libraries
- Machine Learning Tool Interfaces
  - Graphical User Interfaces
  - Command Line Interfaces

Python

- Python is a **high level** language
  - It is optimized for reading by people instead of machines
- Python is also an **interpreted language** which means it is not compiled into machine code
- It is commonly used in an interactive fashion
  - Java & C: write code, **compile** and run, and then watch the output
  - Python: write and run line by line with the interpreter
- This is very useful for tasks that require a lot of investigations (data cleaning) versus those that require a lot of design !
- Different from C++ and java, Python is **dynamically typed** language (like javascript) : you declare the variable and assign a value to it directly !
  - This enables to quickly set the variable type and content

Why Python for Machine Learning ?

- Python is easy to learn
  - Now the language of choice for 8 of 10 top US computer science programs (Philip Guo, CACM)
- Full featured
  - Not just a statistics language, but has full capabilities for data acquisition, cleaning, databases, high performance computing, and more
- Strong Data Science Libraries
  - The SciPy Ecosystem

Tools to be used in this Course

- Programming language to be used in this course: Python
- Libraries:
  - Pandas
  - Numpy
  - Scipy
  - Scikit-Learn
- Interactive tools:
  - Spyder: IDE for python
  - Jupyter Notebook: A a web application that allows to:
    - create and share documents that contain live code, equations, visualizations and explanatory text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, machine learning and much more.

Tools to be used in this Course

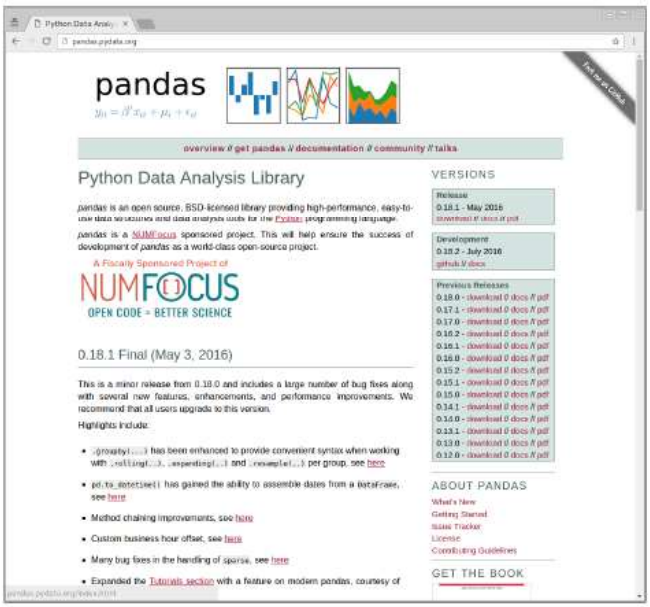
- Anaconda:

  - Anaconda is the leading open data science platform powered by Python. The open source version of Anaconda is a high performance distribution of Python and R and includes over 100 of the most popular Python, R and Scala packages for data science.

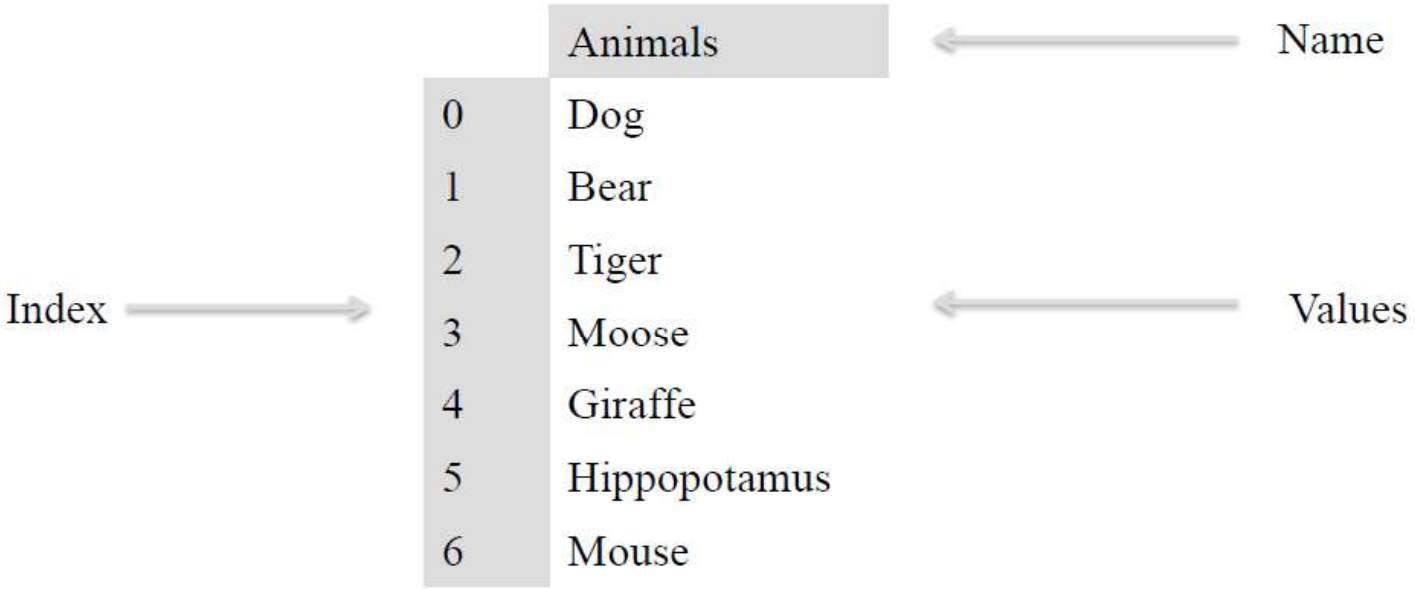
  - https://www.anaconda.com/download

Pandas

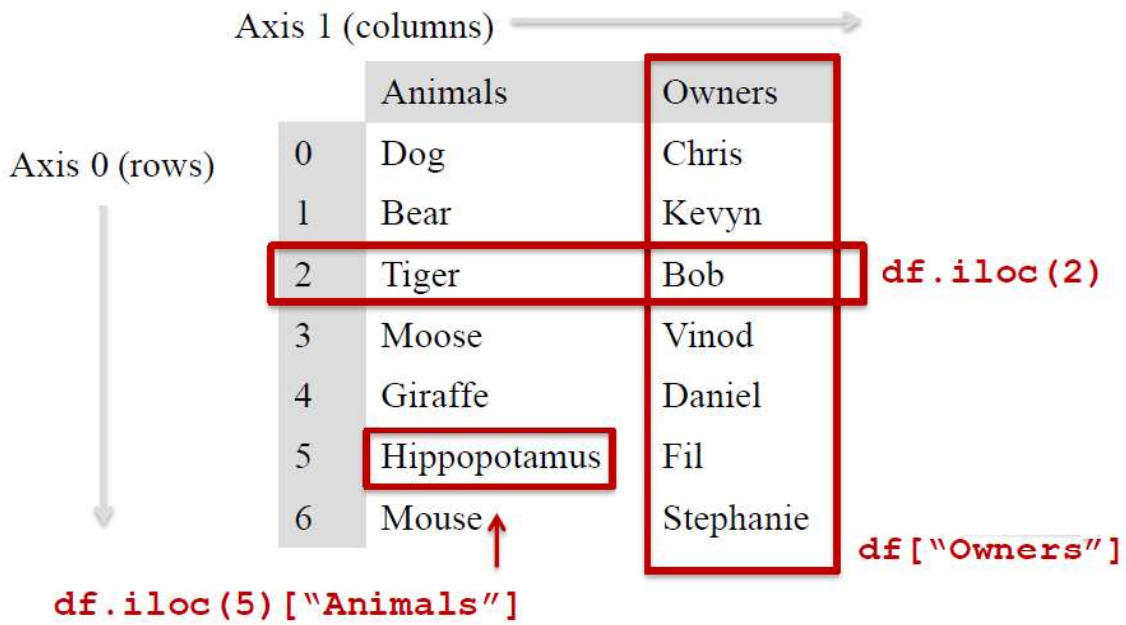
- Created in 2008 by Wes McKinney
- Open source New BSD license
- 100 different contributors
- [Documentation]



Pandas Series



Pandas DataFrame



Pandas DataFrame

df

	Animals	Owners
0	Dog	Chris
1	Bear	Kevyn
2	Tiger	Bob
3	Moose	Vinod
4	Giraffe	Daniel
5	Hippo	Fil
6	Mouse	Stephanie

+

Boolean mask

True	True
True	True
False	False
True	True
False	False
False	False
False	False

=

result

	Animals	Owners
0	Dog	Chris
1	Bear	Kevyn
3	Moose	Vinod

Dive deeper

On Supervised, Unsupervised, and Reinforcement Learning

What Machine Learning Can and Can't Do

What Kind of Problems Can Machine Learning Solve

Python Docs (for general Python documentation)

Python Classes Docs

Scipy (for IPython, Numpy, Pandas, and Matplotlib)

Introduction to pandas: http://nikgrozev.com/2015/12/27/pandas-in-jupyter-quickstart-and-useful-snippets/

Don't forget to check Stack Overflow!

http://planetpython.org/

http://dataskeptic.com/

Thank you for your attention