Introduction to Data Science course

Data Science & Business Analytics

Tiago Cabo

Plan

1. Introduction to Data Science
2. Python core
3. Numpy, Pandas, matplotlib

Full course plan here

What is Data Science?

According to IBM:

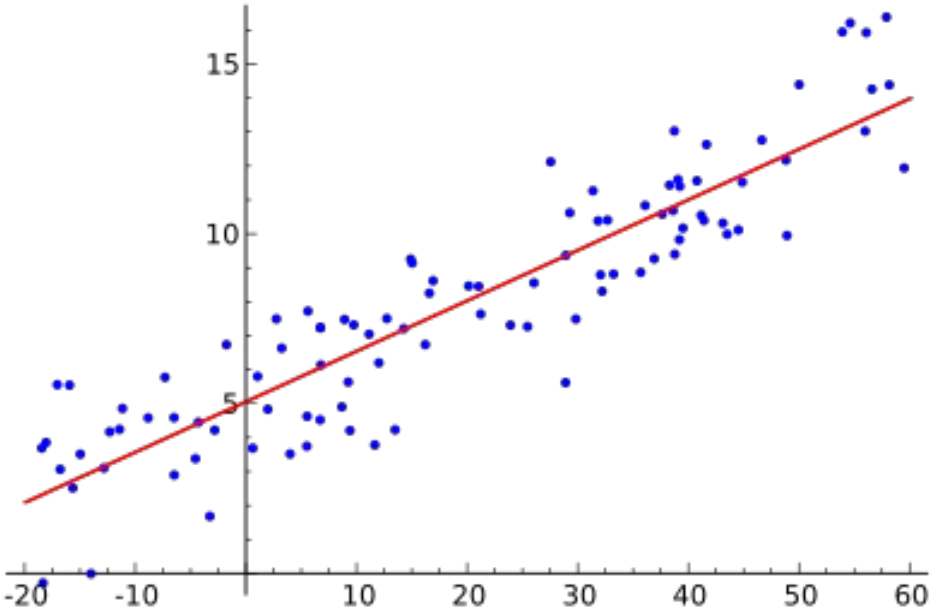
"Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization's data. These insights can be used to guide decision making and strategic planning."

https://www.ibm.com/topics/data-science

What is Machine Learning?

www.menti.com/7yi36vfdoa

Is this Machine Learning?

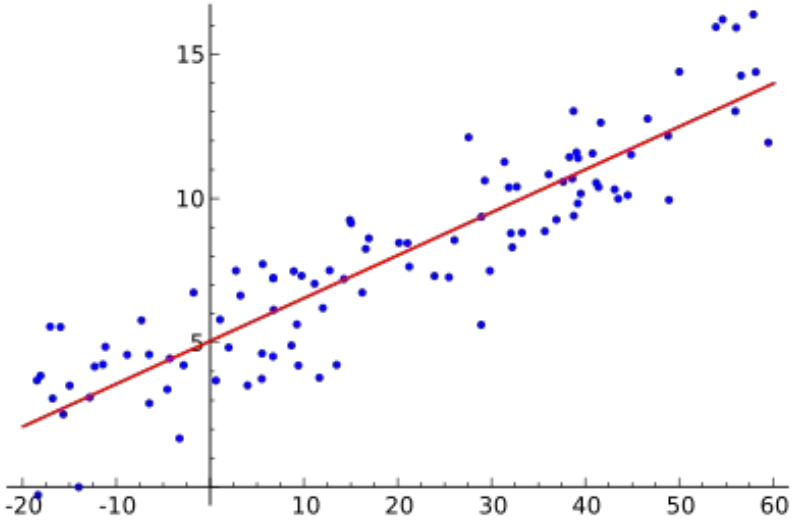


Is this Machine Learning?

Yes. The example on right shows a standard linear regression fitting. The process that takes to use a linear regression in ML problems is the same when compared with deep neural networks.

Main Steps:

- Choose goal to your business problem
- Extract data about your problem
- Map your data with your goal



Example

Example

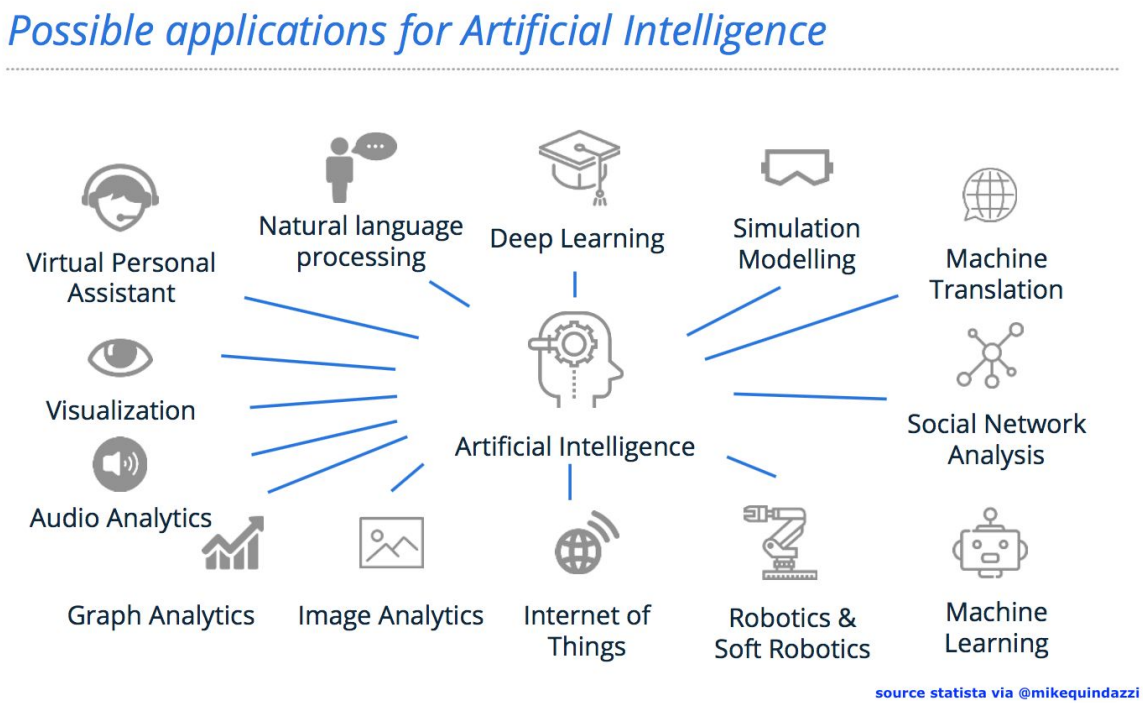
Some differences:

- Petal Length
- Petal width
- Color

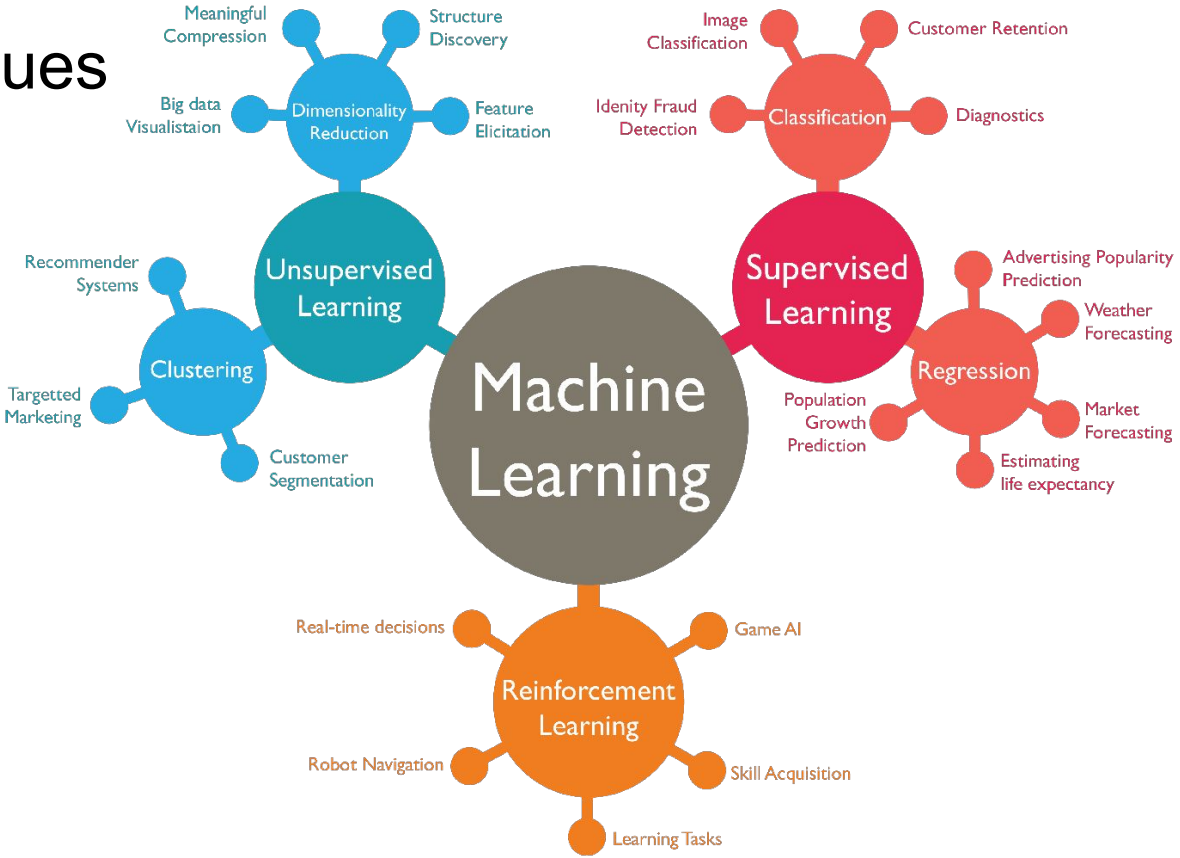
Possible Algorithm:

If Petal Length < 5 & Petal Width > 10 & color == “red” then rose

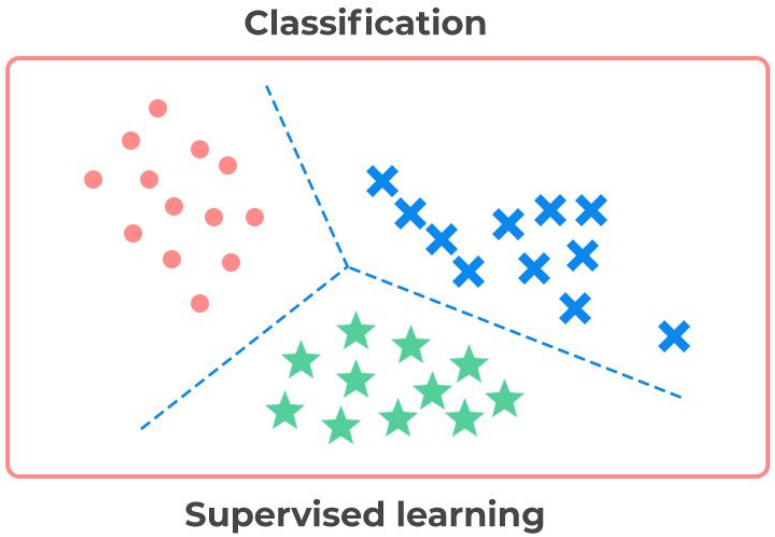
Data Science Applications

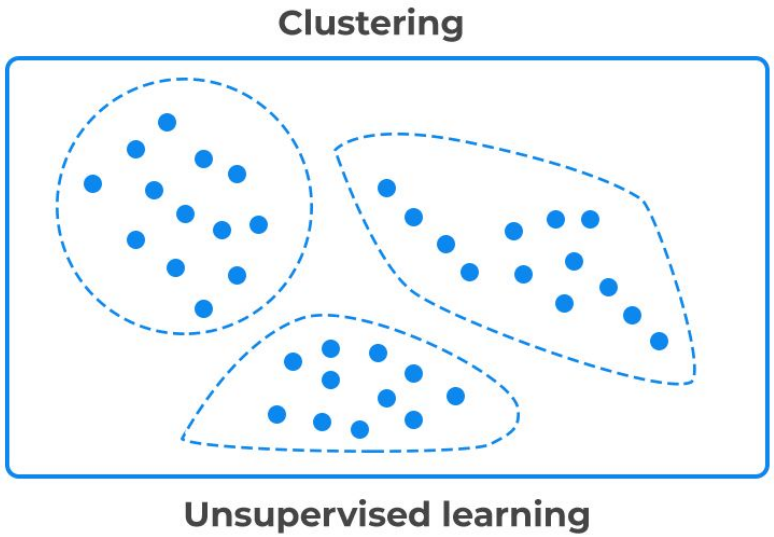


Main techniques



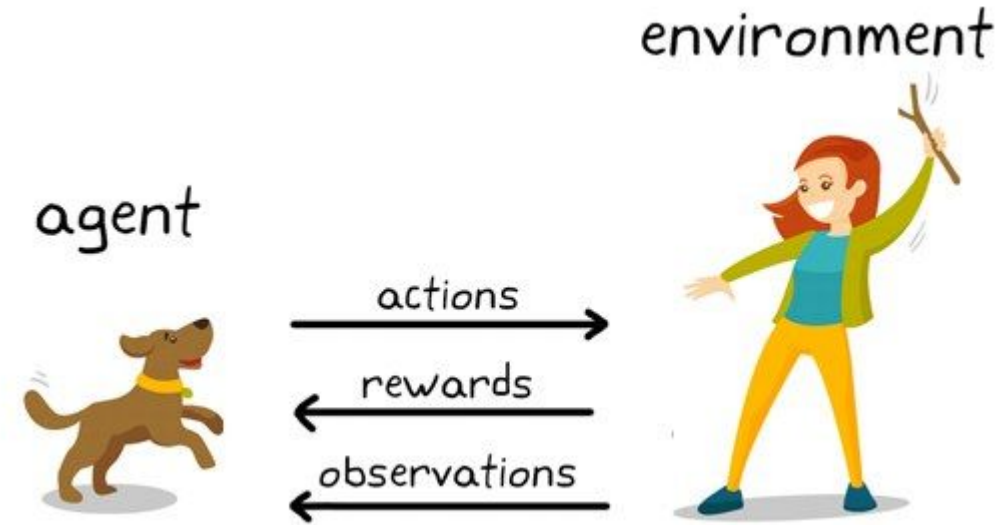
Supervised vs. Unsupervised Learning





Reinforcement Learning

In this types of models, the algorithm learns by trial and error i.e performs an actions and based on the outcome receives a reward which can stimulate simulate behavior or forces a change in the next iteration.



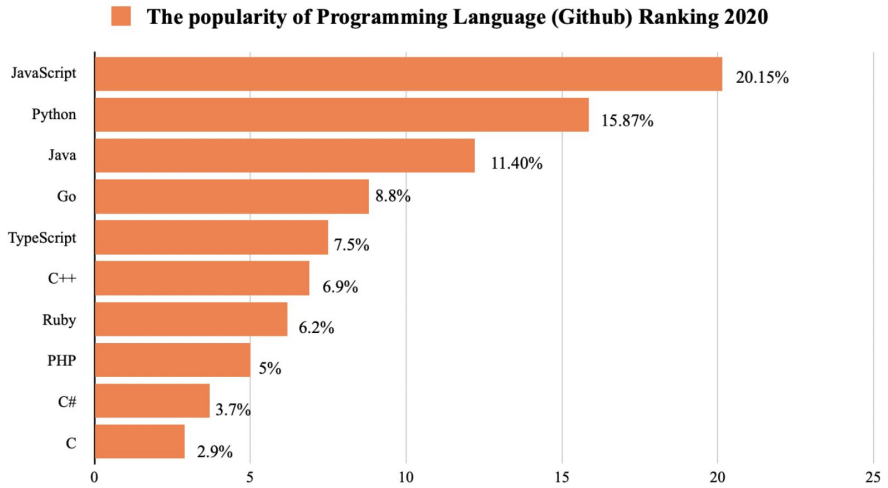
Python Fundamentals

Python was created by Guido Van Rossum in late 1980s in the Netherlands.

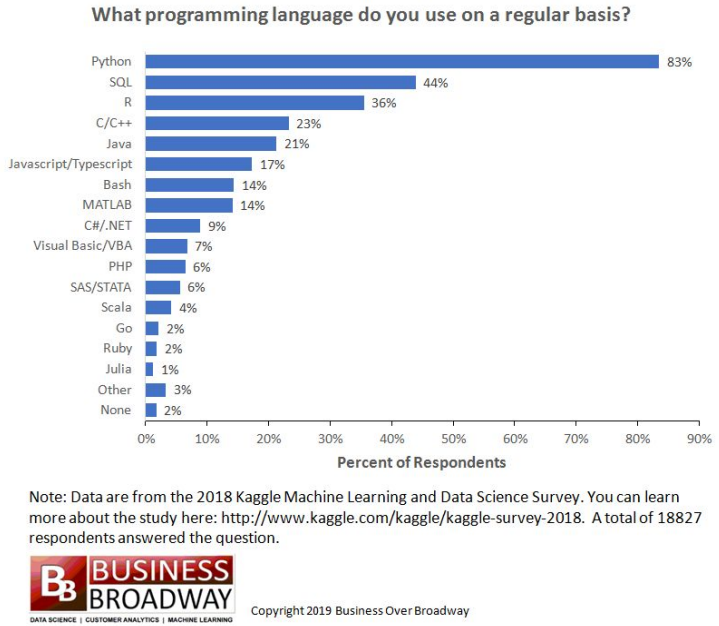
The main idea for the language was to emphasize code readability with the use of significant indentation.

Also includes several complex functions in its base library.

Why python for data science



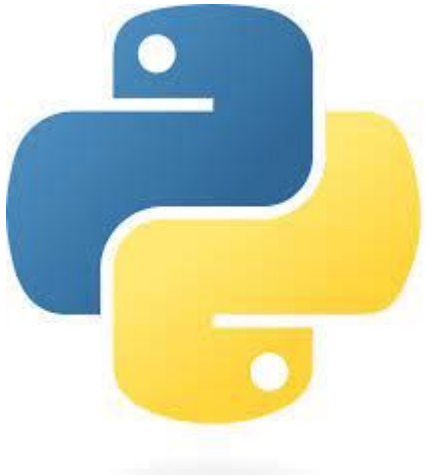
https://darly.solutions/the-most-popular-programming-languages-in-2021/



https://businessoverbroadway.com/2019/01/13/programming-languages-most-used-and-recommended-by-data-scientists/

Why python?

- open-source object-oriented programming language
- Simple and clean syntax (easier for people without prior coding skills)
- Majority of deep learning research uses Python
- Big community with people from various backgrounds
- Extensive Support Libraries
- Not as good as R for statistical modelling
- Not as good as R for data visualizations



Main Components

- **Built-in Functions**
- **Built-in Constants**
- **Built-in Types**
- **Built-in Exceptions**

- **Text Processing Services**
- **Data types**
- **Numeric and mathematic modules**
- **Functional Programming Modules**

Libraries:

- NumPy - Numerical computation on efficient arrays.
- Pandas - to work with structured (tabular, multidimensional) and time series data
- matplotlib - python 2D plotting library
- ggplot - is a plotting system for Python based on R's ggplot2 and the Grammar of Graphics.
- SciPy - SciPy (pronounced “Sigh Pie”) is open-source software for mathematics, science, and engineering
- SciKit-Learn
- Keras / TensorFlow / Theano – Deep Learning

Python environment

- Jupyter Notebook
- Python + 3.10
- IDE:
  - JupyterLab
  - PyCharm (recommended)
  - Spyder
- Install Pip https://www.liquidweb.com/kb/install-pip-windows/
- Anaconda https://docs.anaconda.com/anaconda/install/index.html
- Python Virtualenv https://www.pythoncentral.io/how-to-install-virtualenv-python/

How to create a virtual environment using conda

-> conda create -n <env_name>

-> conda activate <env_name>

-> conda deactivate

How to create a virtual environment using virtualvenv

```
-> python -m venv ./venv
```