Introduction; class 1

1. **Data Science**

How to define **big data**?

Growing torrent of data + growing storage and computation capacity (i.e reading and manipulating the data)

Media + social media data produced every day are stored.

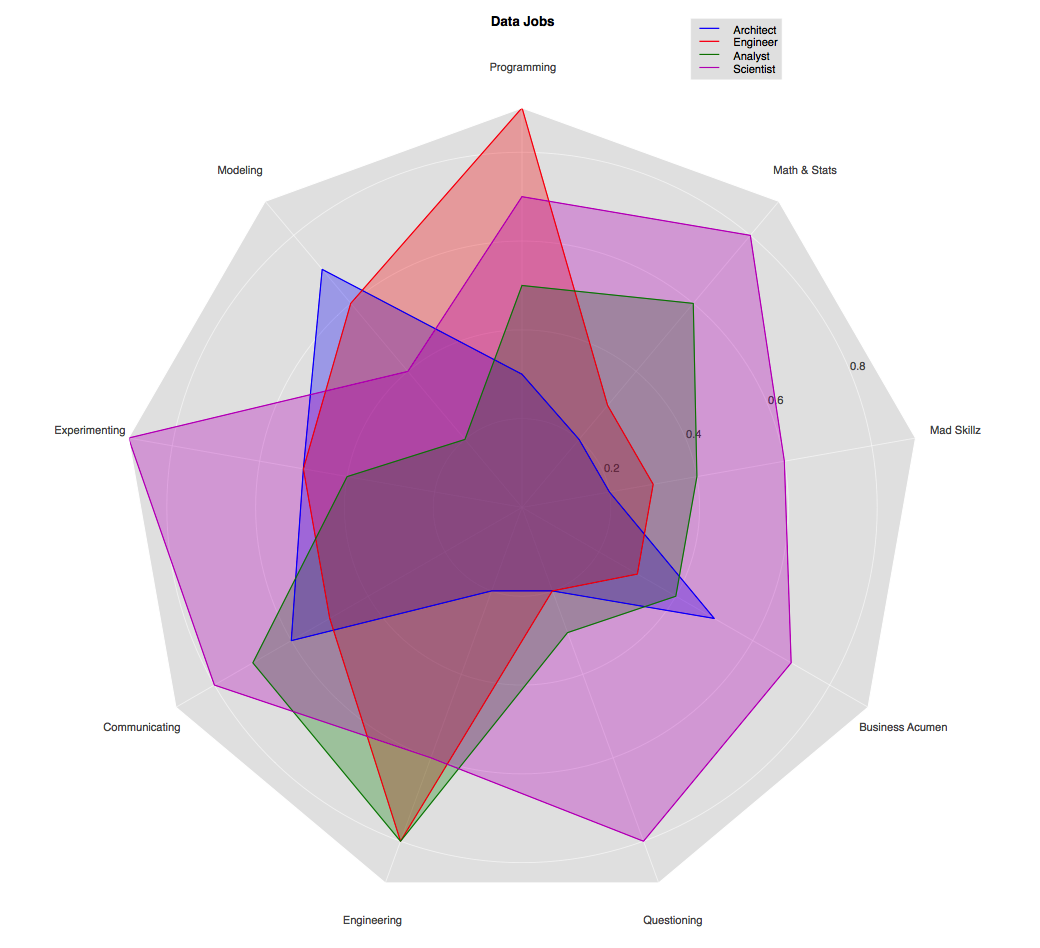
What is **data science**?

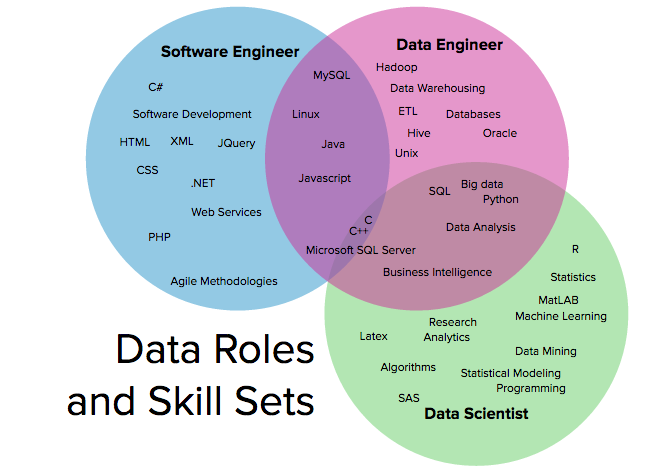
Applicable to multiple fields, even outside of tech companies and computer science (ex: behavioural sciences, law & business, and more)

It combines : Programming, Math & Stats, Skills, Business, Questioning, Engineering, Communication, Experimentation & Modeling

1. Data exploration and preparation
2. Data representation / Transformation
3. Computing with Data
4. Data Modeling
5. Data Visualization and Presentation

<https://towardsdatascience.com/the-dynamics-of-data-roles-teams-6c450b27e59e>





Data **Architect**: someone who designs / develops databases and the data systems’ environments. (low level architecture) ; does not deal with algorithms or data analysis

Data **Engineer** : someone who constructs the data flows.

Data **Analyst**: uses the available data and analyses it (make graphs out of available data for example)

Data **Scientist**: creates and train models that can be used to make predictions or identify data

Buzzword:

Data lake: unstructured + structured data put together

Data lab: separate environment built to allow your analysts and data scientists to figure out the value hidden in your data

Data warehouse = large database

**Structured data** = columns/rows, you have a skeleton.

=/= **unstructured data :** does not have a pre-defined data model or is not organized in a pre-defined manner

<https://blogs.oracle.com/bigdata/what-is-a-data-lab> (ca explique plutot bien les termes dont parle le prof ca)

1. **Machine Learning**

Origin: Alan Turing. First person to give information to a machine so it can learn.

Teach computers to learn (new) data; not a new science, but it is now gaining momentum thanks to new technology that allows complex algorithms to be applied over and over on massive amounts of data.

2 main definitions:

* **Subfield of computer science that gives computers the ability to learn without being explicitly programmed** (definition from 1959)
* **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 of applications: driverless cars, personalized advertisements on social media, product recommendations on Amazon

Machine Learning is used in a lot of **intelligent** applications and modern services.

3 broad categories of machine learning tasks:

<https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/>

1. **Supervised Learning**

**Goal:** learn a general rule that maps inputs to outputs

The program is given a data set and already knows what the correct output should be

Uses data frames (from which it extracts data **features**/**inputs**) to predict a set of information (the **labels/outputs**).

From the inputs it deduces an output for a theoretical input value. We have data to prove that the prediction is correct.

Two models:

* **Linear Regression**

Predicts values within a continuous output → map to a continuous function

(multi) linear or non-linear regressions are possible.

* **Classification**

Predict results in a discrete output →

map input variables into discrete categories (or classes)

Example: predict whether a person has cancer or not depending on their age and tumor size. A person either has or doesn’t have cancer in this example; the result, therefore, is a discrete category (has/ doesn’t have cancer).

The model can be multi-class.

1. **Unsupervised Learning**

You give the machine data, and you want to predict something, without a specific desired outcome or correct answer; only features are given, no labels. It is on its own.

2 models:

<https://courses.cs.washington.edu/courses/cse546/10wi/slides/cdr.pdf>

* **Clustering** (k-means clustering)

Group similar samples into sets → find structure within the data

Makes use of a forecast model and the concept of segmentation (group together a set of data according to a set of parameters)

Clustering is more commonly used than dimensionality reduction, and thus is more important to learn in unsupervised learning.

* **Dimensionality Reduction** (PCA: Principal Component Analysis)

Intelligently reduce the number of features considered ⇒ Data compression / data visualization

* Neural networks

1. **Reinforcement Learning**

This category is still under development even now.

No input nor output to work with

The AI attempts to find the optimal way to accomplish a certain goal in a dynamic environment, or improve performance on a specific task. As it gets closer to its goal, it receives a reward; the overall aim is to predict the best next step to take to earn the biggest final reward.

It does not have a guide explicitly telling it whether it has come close to its goal.

Eg : google self-driving car or Alpha GO

**Keywords**:

data frame (rows and columns of data)

data segmentation / data clustering

neural networks

Example websites:

* <https://clarifai.com/demo>
* <https://www.autodraw.com/>
* <https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/>
* <https://experiments.withgoogle.com/collection/ai>

1. **Machine Learning Tools**

[**https://blog.finxter.com/wp-content/uploads/2018/06/CheatSheet-Python-1-Keywords-1.pdf**](https://blog.finxter.com/wp-content/uploads/2018/06/CheatSheet-Python-1-Keywords-1.pdf)

[**https://scikit-learn.org/stable/**](https://scikit-learn.org/stable/)

Challenge & Data & Codes : <https://www.kaggle.com/>

Pandas

Numpy

Scipy

Scikit-learn

Spyder

Jupyter Notebook

Docs : <https://devdocs.io/python~3.7/>

Cheat Sheet : <https://devhints.io/python>

<https://blog.finxter.com/wp-content/uploads/2018/06/CheatSheet-Python-1-Keywords-1.pdf>