

## Manipulation des données avec Pandas

### Présentation

#### Analyse de données

C'est un ensemble de méthodes descriptives ayant pour objectif de résumer et visualiser l'information pertinente contenue dans une grande quantité de données.

L'analyse des données est considérée comme une méthode d'organisation des données permettant de répondre aux questions de recherche et de tester les hypothèses.



#### I-Outil d'analyse de données

##### Microsoft Excel(Tableur)



- Le logiciel Excel intègre des fonctions de calcul numérique, d'analyse de données
- Impossible de traiter une grande quantité de données (16384 colonnes et un peu plus d'un million de lignes).
- Leur puissance de calcul est très insuffisante.

#### Le language Python



. Python est un **langage de programmation** simple.

Python possède de nombreuses librairies pour analyser et visualiser les données.

Python est soutenu par une large communauté.

### II- Pourquoi Python

#### 1. C'est quoi Python ?

C'est quoi Python ? Le langage de programmation Python a été créé en 1989 par 'Guido van Rossum'

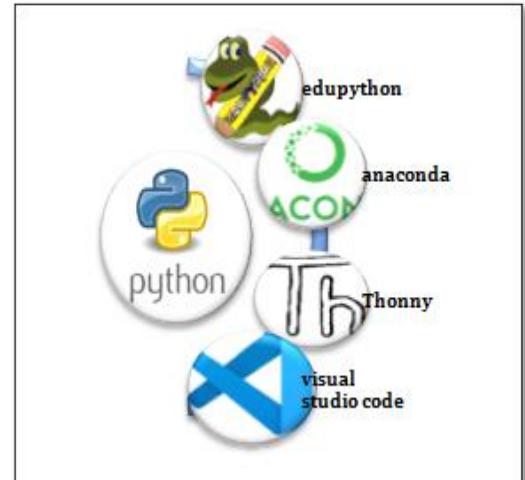


## 2. Éditeur de texte

L'apprentissage d'un langage informatique comme Python va nécessiter d'écrire des lignes de codes à l'aide d'un **éditeur** de texte (exemple : Jupyter Notebook, Thonny, Mu, anaconda, EduPython, , IDLE python ... )

Premier contact avec Python est un langage **interprété**, c'est-à-dire que chaque ligne de code est lue puis interprétée afin d'être exécutée par l'ordinateur.

Comme tout langage informatique, on peut enregistrer ces instructions dans un fichier, que l'on appelle communément **un script** (ou **programme**) Python.



### Remarque :

- 1-L'extension du fichier standard des scripts Python est **.py**
- 2-on peut travailler en ligne ou installer un éditeur python pour notre smartphone.

Notre choix est le logiciel **Jupyter Notebook**

- 3-L'extension d'un fichier **Jupyter Notebook** est **.ipynb**



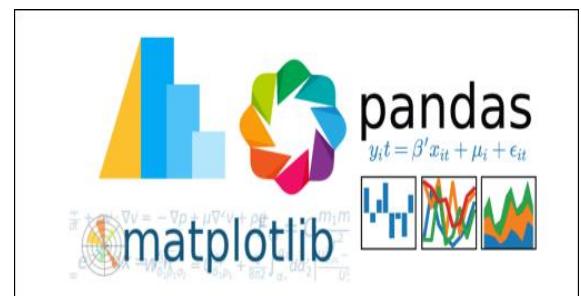
## 3. Introduction à Pandas

Pandas est une librairie ou bibliothèque python qui permet de manipuler facilement des données à analyser : Manipuler des tableaux de données avec des étiquettes de variables (colonnes) et d'individus (lignes). Ces tableaux sont appelés Data Frames.

On peut facilement lire et écrire ces data frames à partir ou vers un fichier tabulé.

On peut facilement tracer des graphes à partir de ces data

Frames grâce à **matplotlib**. Le module pandas a été conçu pour la manipulation et l'analyse de données. Il est particulièrement puissant pour manipuler des données structurées sous forme de tableau.



## 4- Présentation de la bibliothèque Pandas

Columns				
	Name	Score	Attempts	Qualify
0	Anastasia	12.5	1	yes
1	Dima	9.0	3	no
2	Katherine	16.5	2	yes
3	James	Nan	3	no
4	Emily	9.0	2	no

**Pandas** est une librairie python qui permet de manipuler facilement des données à analyser. Elle permet de:

Manipuler des tableaux de données basé sur des étiquettes. Ces tableaux sont appelés Series (données selon une dimension), DataFrames (données selon deux dimension), Panels (3d) et multiIndex ( plus de 3d).

Appliquer rapidement des opérations sur ces structures.

Lire et écrire ces tableaux à partir ou vers un fichier tabulé.

Les tableaux Pandas peut avoir des types différents (numérique, string, boolean ...).