

SAE 303

Data-visualisations

TP#3 S3

Danielo **JEAN-LOUIS**

Données - Le nouvel or noir

- Priorité des entreprises
 - Source de revenus potentielle conséquentes
- Priorité au sein de l'État – E. Macron 2018
- Secteur très porteur dans l'industrie
 - Data-engineer / Data-analyste / Gestionnaire de serveurs...

Source(s) :

- <https://www.lebigdata.fr/emmanuel-macron-big-data-priorite-europe>

Données

- Indispensables pour la data-visualisation
- Origines multiples :
 - Open data
 - Messages / Photos / vidéos
 - e-mail
 - [...]
 - VOUS

Numpy

- Bibliothèque externe à Python
- Optimisé pour les tableaux jusqu'à que votre ordinateur n'en puisse plus
- ~50x plus rapide que Python pour la gestion des tableaux
- Intégrée par défaut dans Anaconda / Google Colab → pas besoin de pip

pip

- Signifie Pip Installs Python
- Gestionnaire de bibliothèques externes

Numpy

- Rajoute de nouvelles fonctionnalités aux tableaux en Python
- Possibilité d'utiliser certaines syntaxes de Python sur les tableaux numpy
- Spécialisé pour gérer les données non labellisées. De préférence des chiffres.

Pandas

- Bibliothèque Python
- Contraction de Panel Data
 - ou Python Data Analysis
- Intègre des fonctionnalités de numpy

Pandas

- Permet de gérer les données labellisées (dites tabulaires)
 - Super-excel en programmation
- Gère sans problème de très gros fichiers
 - Plus d'un million de lignes
- Intégrée par défaut dans Anaconda / Google Colab → pas besoin de pip

Pratiquons ! - Pandas

Pré-requis :

- Avoir le notebook ressources/pandas/initiation-pandas-partie-1.ipynb

A télécharger ici :

<https://downgit.github.io/#/home?url=https://github.com/DanYellow/cours/tree/main/big-data-s4/travaux-pratiques/numero-5/ressources/pandas/pratique>

Questions ?

