Python Big Data

Olivier Ricou

2025

https://www.lrde.epita.fr/~ricou/

Travailler les données, pour faire quoi ?

- découvrir la vérité cachée dans les données (citoyenneté / journaliste)
- découvrir un potentiel d'économie ou de création de richesse
- préparer les données pour un réseau neuronal ou autres travaux
- trouver un boulot très bien payé

Trouver la vérité

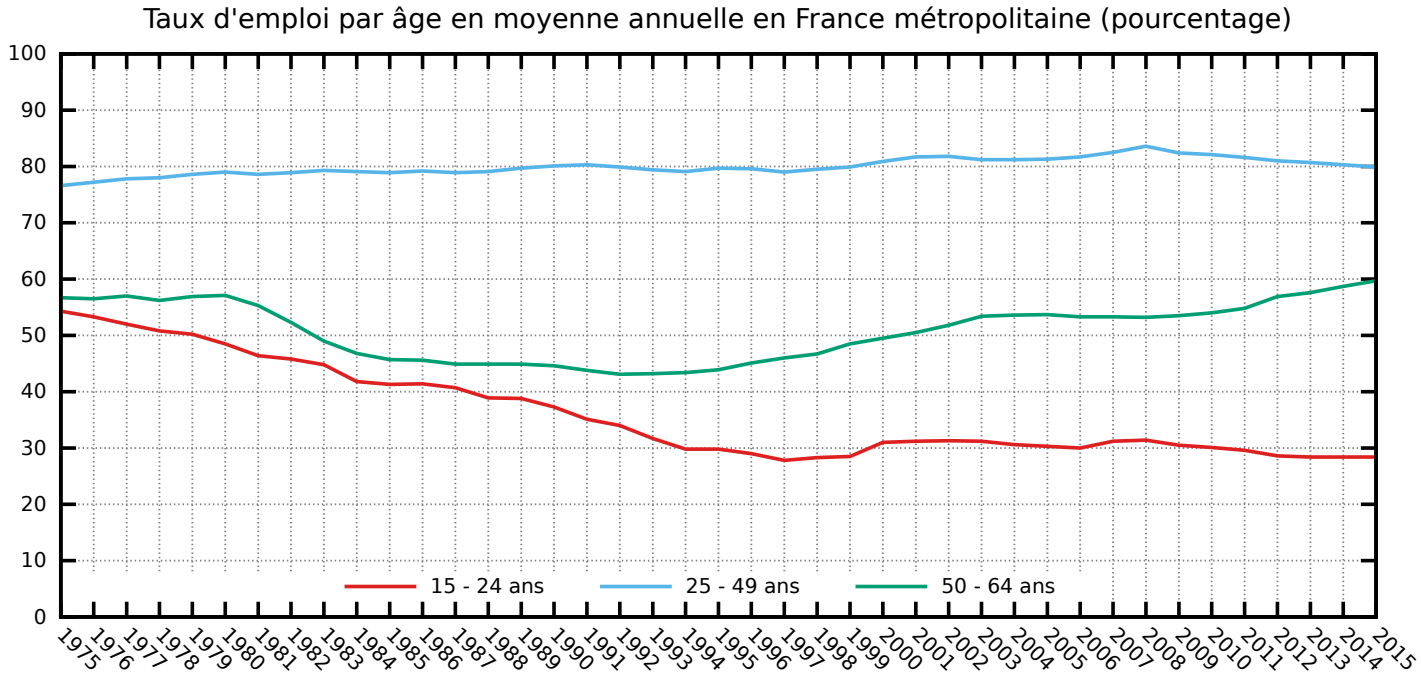
La vérité est souvent cachée.

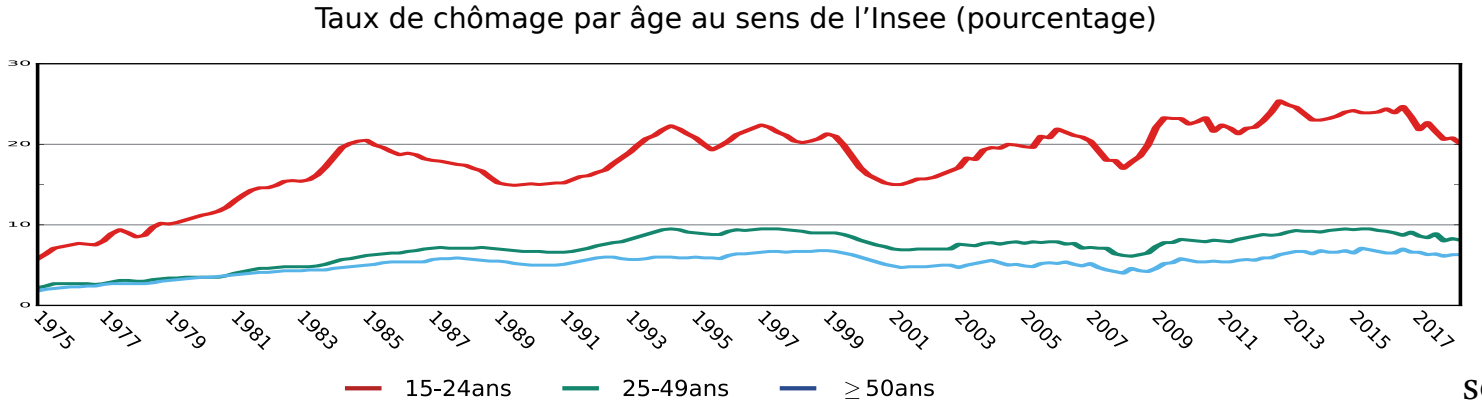
Attention

- à ne pas vouloir trouver **sa** vérité → celle qui nous arrange
- à ne pas se faire piéger (données fausses, définition trompeuse, biais...)
- à la fausse interprétation
  - en parler avec des experts
  - en parler avec les personnes concernées
- à la manipulation des autres, ceux qui n'ont pas respecté le premier point volontairement ou pas et qui nous influencent

De bonnes données peuvent permettre de faire surgir des vérités.

La vérité existe-t-elle ?

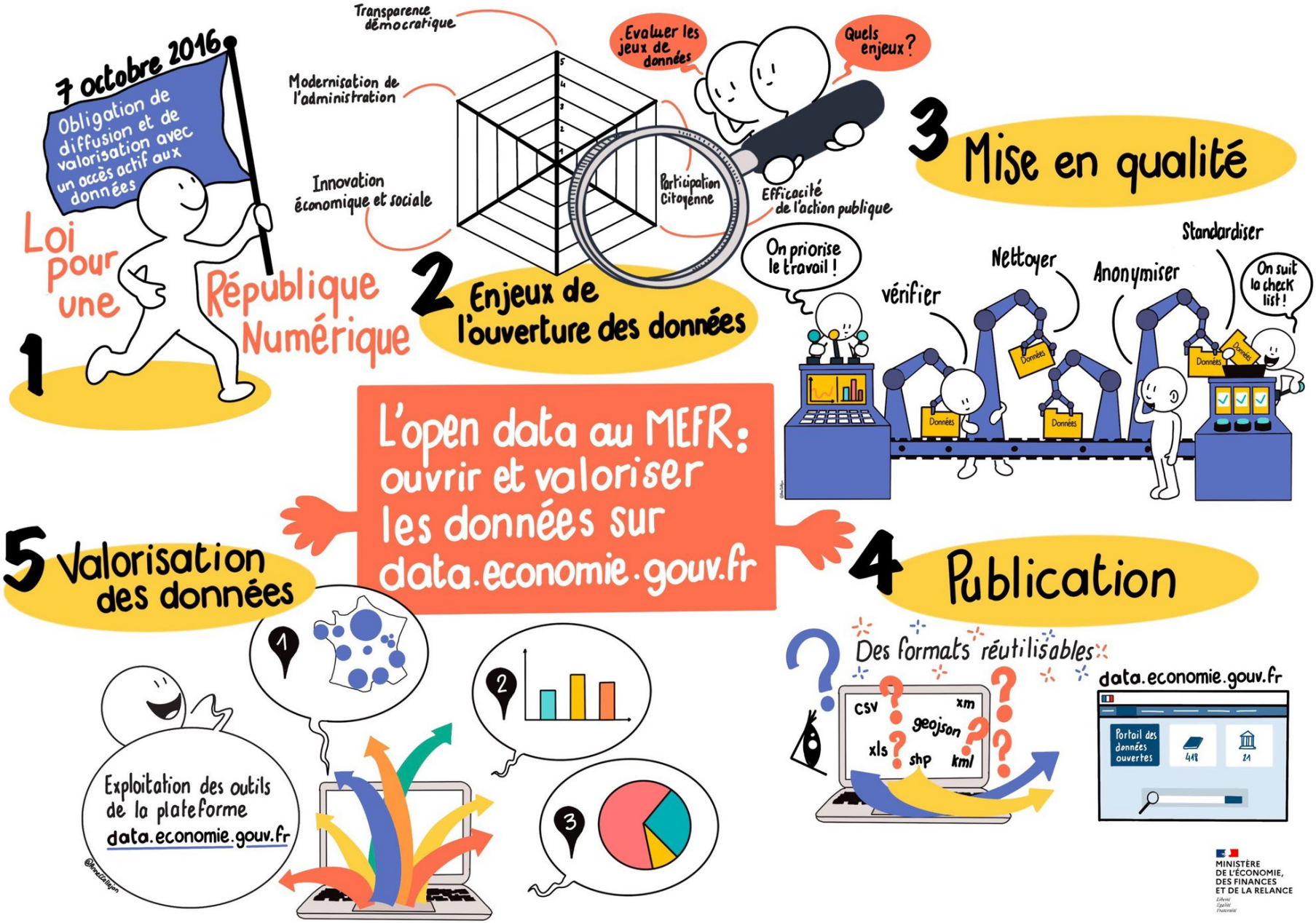




source : Wikipédia

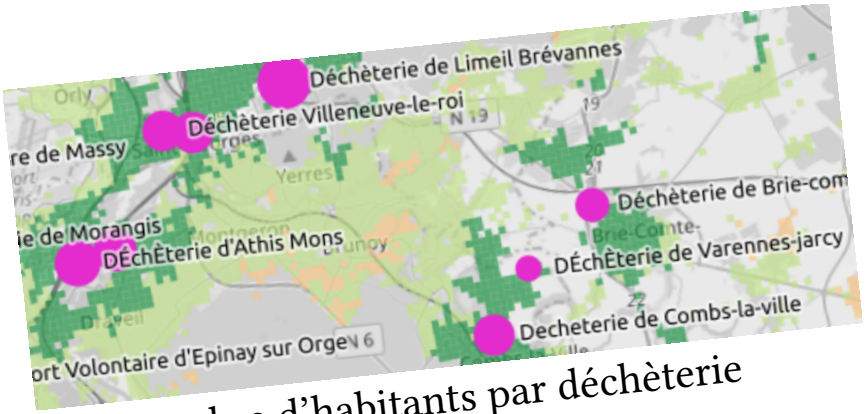
Quelle est la définition de chômage ? Ça veut dire quoi rechercher un travail ?

Open data


7 octobre 2016
Obligation de diffusion et de valorisation avec un accès actif aux données
Loi pour une République Numérique
1
Transparence démocratique
Modernisation de l'administration
Innovation économique et sociale
Participation Citoyenne
Efficacité de l'action publique
Evaluer les jeux de données
Quels enjeux ?
2 Enjeux de l'ouverture des données
3 Mise en qualité
On priorise le travail !
vérifier
Nettoyer
Anonymiser
Standardiser
On suit la check list !
Données
L'open data au MEFR: ouvrir et valoriser les données sur data.economie.gouv.fr
4 Publication
Des formats réutilisables
csv
xls
shp
kml
geojson
xml
data.economie.gouv.fr
Portail des données ouvertes
448
21
5 Valorisation des données
1
2
3
Exploitation des outils de la plateforme data.economie.gouv.fr
MINISTÈRE DE L'ÉCONOMIE, DES FINANCES ET DE LA RELANCE

Les communs numériques

Wikipédia et OpenStreetMap sont des données communes largement utilisées et ré-utilisées.



Nombre d'habitants par déchèterie

Post-Disaster Building Database Updating Using Automated Deep Learning: An Integration of Pre-Disaster OpenStreetMap and Multi-Temporal Satellite Data

by Saman Ghaffarian [1,*], Norman Kerle [1], Edoardo Pasolli [2] and Jamal Jokar Arsanjani [3]

Learning to Interpret Satellite Images using Wikipedia

Burak Uzkent, Evan Sheehan, Chenlin Meng, Zhongyi Tang, Marshall Burke, David Lobell, Stefano Ermon

Exploration of OpenStreetMap missing built-up areas using twitter hierarchical clustering and deep learning in Mozambique

Hao Li [a], Benjamin Herfort [a], Wei Huang [b, c], Mohammed Zia [a], Alexander Zipf [a]

Taking advantage of Wikipedia in Natural Language Processing

T. Yano, Moonyoung Kang • Published 2008

Wikipedia is an online encyclopedia created on the web by various participants. Although it is not cr
purpose of helping studies in language processing, its size and well-formed structure is attracting m

Carte touristique

Voir aussi `https://www.ign.fr/la-demarche-geocommuns`

Potentiel économique

C'est la raison principale du succès des données massives[1].

Exemple de Total

- Total utilise beaucoup de centrifugeuses
- elles ont des consignes d'utilisation dont le nombre de tours/mn max
- les centrifugeuses ont pleins de capteurs → données
- Total décide de faire tourner à fond ses centrifugeuses
- elles meurent mais on a plein de données sur leurs façons de mourir
- aujourd'hui elles tournent à fond et sont arrêtées juste avant la panne
- on répare ou remplace la pièce qui va casser et c'est reparti

Total gagne de l'argent en ayant ses centrifugeuses qui tournent à fond.

Les constructeurs des centrifugeuses aimeraient avoir l'info qui permet cela.

[1] traduction de *big data*

Captation de données

C'est un modèle économique

- potentiellement très rentable
- parfois amoral voire dangereux pour la société

Les bracelets connectés

- Les données vont sur les serveurs du fabriquant.
- Le plus souvent on ne peut pas l'empêcher.
- Quelles garanties sur l'exploitation des données ?
- Des personnes sont intéressées (médecin, assurance, employeur, famille)

Le droit à la portabilité des données (RGPD) permet de récupérer ses données. Qui le fait ?

Préparer les données

Rendre des données propres et exploitable est un travail de grande valeur.

Si le jeu de données a des erreurs ou des trous

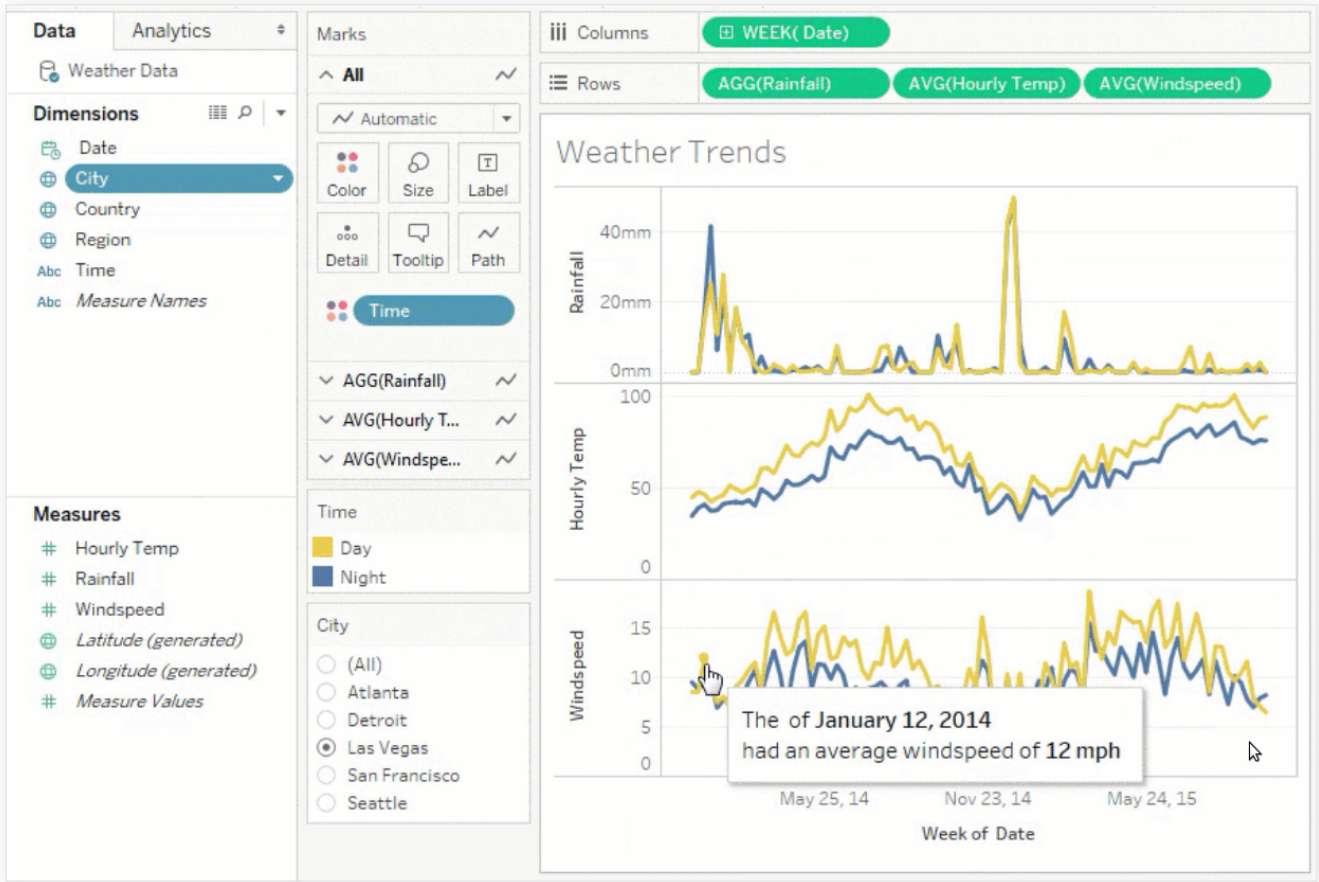
- les sciences expérimentales auront du mal à progresser
- les réseaux neuronnaux convergeront moins bien voire pas du tout
- la vérité devient une erreur
- le gain financier espéré n'arrive pas

Les données sont rarement propres.

Les outils

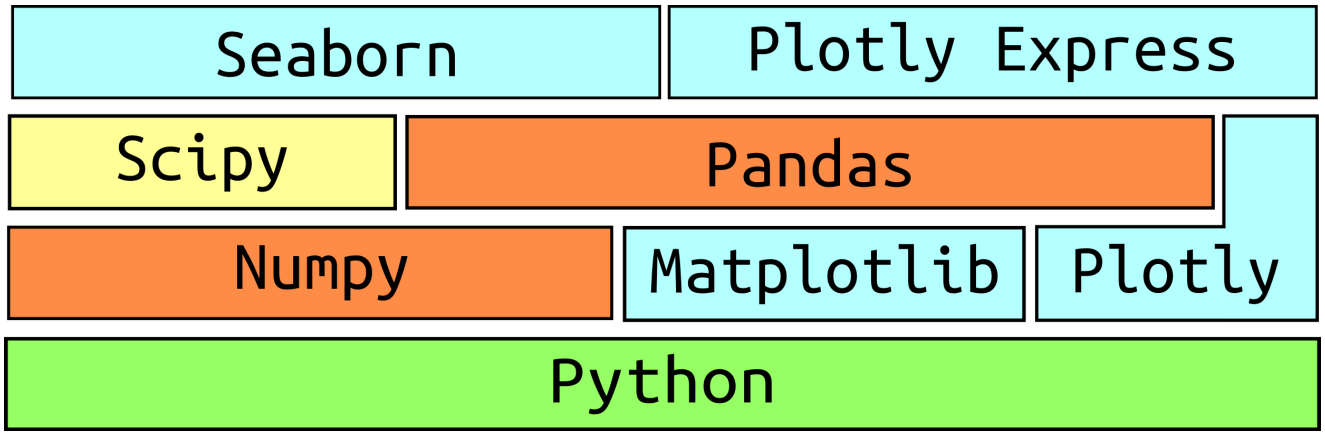
Les principaux dans l'ordre de puissance[2] :

- le tableur (Excel…)
- la base de donnée relationelle (SQL)
- les applications à la Tableau
- les langages Python et R



[2]SQL est quand même un peu à part.

Notre programme



On se concentre sur

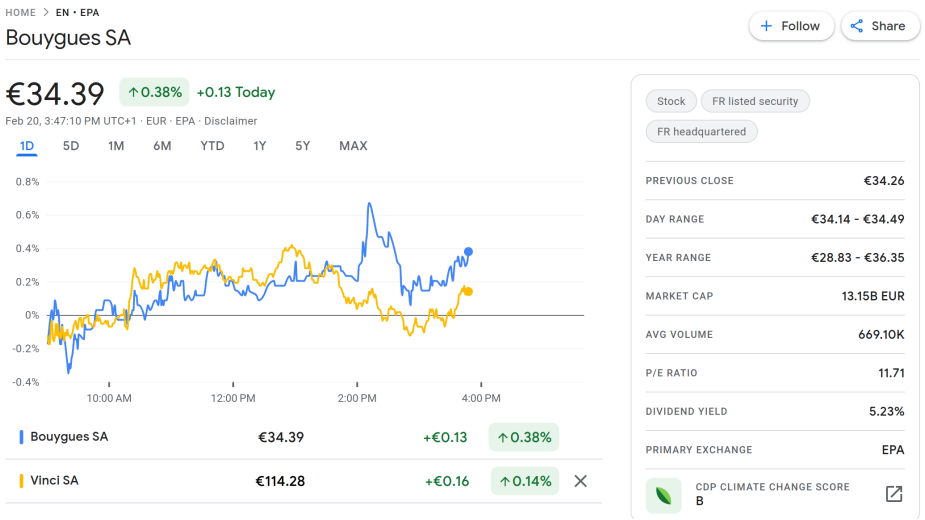
- Pandas
- Matplotlib
- Seaborn (rapidement)
- Plotly + Plotly Express
  & Dash (partie de Plotly pour faire des applications)

Python et Numpy étant censés être déjà connus.

Le projet – Données financières

Il s'agit de faire un tableau de bord qui permette

- d'avoir un rendu global de l'entreprise choisie (comment on la choisie ?)
- de visualiser des cours de bourse, un ou plusieurs, sous forme de courbe (logarithmique)
- de faire des requêtes avec filtres qui produisent un tableau dynamqiue (entreprises belges ayant progressé de plus de 20 % en 2021 par exemple)
- votre idée



Projet = travail, même à 3

Vous êtes noté sur le projet par vos pairs (15 pts projet, 5 pts notation).

Il est à faire en binôme.

Produire un résultat de qualité demande beaucoup de temps. Voici le temps que passe un pro pour produire des graphiques qui seront utilisés dans la presse :

- Concept (au tableau) : 3 h
- Recherche des données : 4 h
- Traitement des données : 5 h
- 1ère version de graphique : 6 h
- 2e version : 2 h
- 3e version : 2 h

Total : 22 heures !

Emploi du temps

4 cours de 3 heures.

1. Présentation + Pandas
2. Pandas + Graphisme
3. Graphisme
4. Projet (début)

Bonus

Petit essai sur les relations entre

- l'ouverture des données
- la transparence
- la démocratie

Site :
`opendata.ricou.eu.org`

Sur Amazon :
papier : 5 €, numérique : 3 €