

Python pour la data science

séance 1

Axel Morin

axel.morin@polytechnique.edu

CREST (4ème étage de l'ENSAE)

September 16, 2025

Résumé des pages d'introduction

Pourquoi Python

- Python est un langage informatique avec une grande communauté très active et des intérêts très variés.
 - de nombreux tutoriels et documentations.
 - des solutions partagées par la communauté *~700,000 packages python sur pip [source](#).*
 - “second meilleur langage” pour toute tâches
- Language facile à apprendre, de haut niveau¹, rapide à mettre en place.

Pourquoi s'embêter quand on a Copilot ?

- Stratégie de fond et logique générale nécessaire au data scientist
- Pour formuler un problème il faut connaître le domaine
- Aiguiser son regard critique
- Retard structurel
- Travail avec des données sensibles
- *considérations éthiques et environnementales*

Objectifs, évaluation et plan du cours

Lecture en Autonomie

Démarche à adopter face à un jeu de données

- Où trouver des données ?
 - Insee (sondages, données individuelles from fichiers administratif, données agrégées par des groupes privés) ou l'IGN
 - **Site de la ville de Paris, data.gouv**
 - OpenStreetMap, Wikidata, OpenFoodFacts
 - Secteur privé

Avant d'utiliser un jeu de données

- D'où proviennent mes données ? Sont elles fiables ?
- Quelle informations puis-je récupérer ? Dans quelle mesure est ce que mes données répondent à ma problématique ?
- Mes données sont-elles libres ?

- Nettoyer mes données
 - *Sujet des prochains TDs*

Pendant l'analyse descriptive

Challenge the dataset!

- Est ce que mes données sont représentatives de mon sujet d'étude ?
- Est ce que j'ai les bons ordres de grandeur ?
- Est ce que je comprends le sens de mes variables ?
- Est ce que j'ai des valeurs aberrantes (*outliers*) ? Qu'est ce que j'en fait ?
- Est ce que mes résultats sont surprenants ? Est ce que j'ai moi-même pourri mes données ?

Pendant la modélisation

La modélisation ne peut que confirmer mes hypothèses.

Rookie moves :

- Se lancer directement dans la modélisation sans avoir fait un travail préliminaire approfondi.
- Choisir la méthode avant d'avoir compris mon sujet

Partager son code et ses résultats

Partage du code

- Github, Gitlab, Git... pour le versionnement code

Partage des résultats

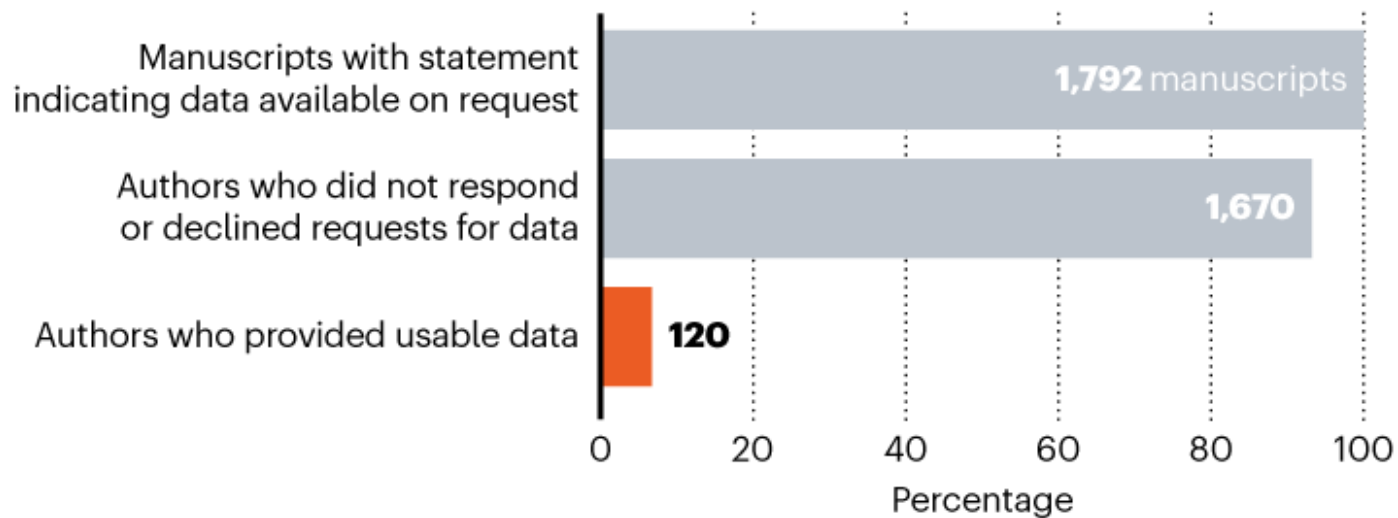
- Notebook clean pour montrer les résultats clefs, toutes les fonctions et classes dans des fichiers à part.
- Visualisation de données

Partage de code + partage des données = Projet reproductible

Reprodu quoi?

DATA-SHARING BEHAVIOUR

Of almost 1,800 manuscripts for which the authors stated they were willing to share their data, more than 90% of corresponding authors either declined or did not respond to requests for data. Only about 7% of authors actually handed over data.



©nature

Éthique et responsabilité

- *En fonction de votre position*, votre travail et votre analyse a un impact sur certaines vies.
 - Politiques d'austérité basé sur des données erronées
 - Mauvais compte de cas de Covid à cause d'une troncation hasardeuse (UK)
- Crise de la reproductibility (*encore pire avec les LLMs*)
 - Il y a une volonté commune de faire face à ses défis et de créer des espaces de travail qui facilitent la reproductibilité, mais c'est pas encore ça.
- Attention aux biais cognitifs dans vos analyses (biais de confirmation) **ET** dans vos visualisations (colormap, pie charts ...)

Approche écologique

- Technologies digitales représente actuellement 4 % des émissions de CO2 mondiales. On s'attend à une augmentation.
 - Utilisation de jeux de données massifs
 - Entraînement de modèles qui prennent des semaines
 - Maintenance des serveurs.
- Conscience en tant que data scientist (codecarbon)



Talk is cheap

Avoir un environnement Python fonctionnel pour la data science

Installation locale

- Nécessite d'installer python avec conda² et un IDE comme VSCode
 - Demande plus de temps pour comprendre ce qu'on fait.³
 - Donne plus de contrôle sur les environnements de travail

Utilisation d'un environnement en ligne

- Pour ce cours il est conseillé d'utiliser les services de SSPCloud ou Google Colab
 - *ready-to-go* solution
 - pas de gestion des environnements

Dans tous les cas vous rencontrerez des problèmes de packages et de version

Découverte des notebooks

>>> Notebook

Introduction à Python

>>> Notebook

Footnotes

1. haut niveau = très proche du langage humain `print("Hello World")` vs `std::cout<<"Hello World"<<<std::endl;`
2. `(base) >>>`
3. Si vous voulez vous y tenter.