DATA SCIENCE AVEC PYTHON

Année académique 2024-2025

Master: X-MAS-DE-4

Enseignant : David Rhenals

Introduction à la Data Science et au langage Python

Contenu

- Qu'est-ce que la Data Science
- Quelques domaines d'application en Data science
- Installation et configuration de l'environnement Python
  - ✓ Anaconda
  - ✓ Python (installation de librairies – optionnel)
    - ✓ Jupyter Notebook /Jupyter Lab (Environnement de Développement Interactif)
    - ✓ Spyder (Environnement de Développement Intégré)
- Quelques rappels basiques de programmation Python
- Manipulation de l'environnement Anaconda-Python a travers d'exercices simples

Qu'est-ce que la Data Science

- **Data** : Ensemble de mesures, de caractéristiques ou de faits concernant un groupe de variables **(Shea, 2024)**
- **Data science simple définition**: Processus de extraction du sens des données **(Shea, 2024)**
- **Data Science au sens large:** Domaine interdisciplinaire qui utilise des méthodes, des processus, des algorithmes et des systèmes scientifiques pour extraire des connaissances et des informations à partir de données (structurées et non structurées). Cette discipline intègre des concepts issus de plusieurs domaines tels que la statistique, l'informatique, l'intelligence artificielle et la gestion des données pour analyser et interpréter les données dans le but de résoudre des problèmes complexes ou de prendre des décisions éclairées **(Igual & Seguí, 2017; Provost & Fawcett, 2013)**

Quelques domaines d'application en Data science

- **Secteur de la Santé :**
  - ✓ **Diagnostic médical** : Utilisation de modèles de machine learning pour analyser les images médicales (comme les IRM ou les rayons X) et aider au diagnostic de maladies.
  - ✓ **Médecine personnalisée** : Analyse des données génétiques pour proposer des traitements personnalisés en fonction du profil génétique de chaque patient.
- **Finance :**
  - ✓ **Détection des fraudes** : Les institutions financières utilisent des algorithmes de Data Science pour identifier les transactions suspectes et prévenir la fraude.
  - ✓ **Gestion des risques** : Modélisation des risques financiers pour optimiser les portefeuilles d'investissement et évaluer la solvabilité des clients.

Quelques domaines d'application en Data science

- **Marketing :**
  - ✓ **Segmentation de marché** : Analyse des données client pour segmenter le marché et cibler les campagnes marketing de manière plus efficace.
  - ✓ **Prévision de la demande** : Utilisation de modèles prédictifs pour anticiper la demande de produits et optimiser les stratégies d'approvisionnement.
- **Industrie pétrolière :**
  - ✓ Amélioration de la prédiction de la production pétrolière à la aide de l'apprentissage automatique (Machine Learning)
  - ✓ Détection semi-automatique de zones productrices d'hydrocarbures à partir de diagraphies (Enregistrements des puits, neutron, sonique, électrique, etc.)

Installation et configuration de l'environnement Python

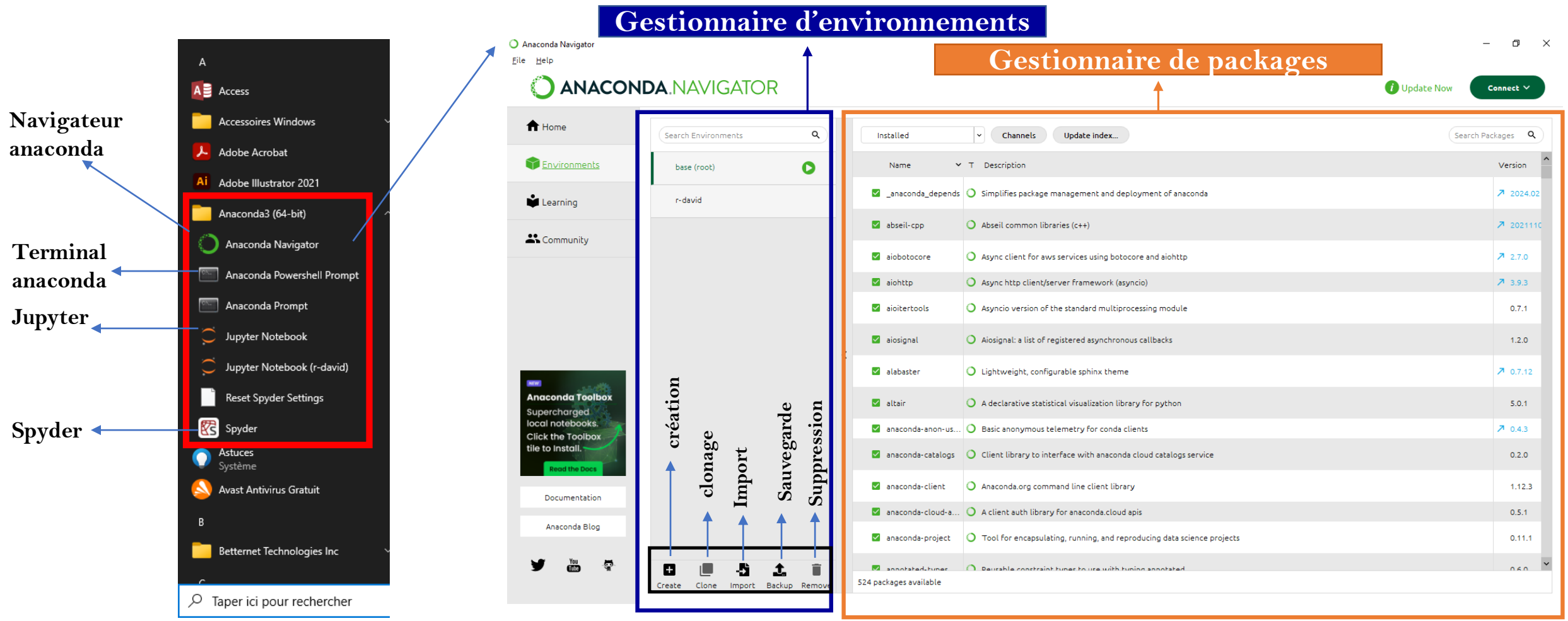
- ✓ Langage de programmation interprété et polyvalent (relativement facile à lire et à écrire – exécution ligne par ligne)
- ✓ Multiplateforme (Windows, macOS et Linux)
- ✓ Large bibliothèque standard pour manipulation de tâches courantes (fichiers, opérations mathématiques, gestion protocoles internet, etc.)
- ✓ Programmation orienté objet (POO) et fonctionnelle
- ✓ Applications diverses (Développement web, Développement logiciel, Administration de systèmes, **Data Science, Analyse de données, intelligence artificielle, machine learning**, etc
- ✓ Entre autres

- ✓ Distribution de python (et R) avec des packages ou librairies préinstallées. Elle est préconçue pour la science de données (cela ne veut pas dire que celle-ci ne peut pas être étendue à d'autres tâches).
- ✓ Inclut non seulement l'interpréteur python, mais aussi des outils comme Jupyter, Spyder (IDE pour python), anaconda prompt (terminal) et des nombreux packages pour la science de données (Numpy, Pandas, Matplotlib, Scipy, Scikit-learn, entre autres)
- ✓ Permet de démarrer plus facilement des projets de science de données grâce à sa configuration initiale par rapport à python où il serait nécessaire l'installation de toutes les librairies

Installation et configuration de l'environnement Python (Anaconda)

- Télécharger la distribution Anaconda depuis le site anaconda. Les instructions d'installation pour les différents systèmes d'exploitation (Windows, macOS et Linux).
- Fournir l'adresse email afin d'avoir accès au fichier exécutable .exe (windows) ou .pkg (mac)
- Exécuter le fichier .exe ou .pkg avec les options par défaut
- Finalisée l'installation, le navigateur Anaconda devrais être disponible dans le menu windows et mac.

Installation et configuration de l'environnement Python



Création d'environnement virtuel (Anaconda-Python) par ligne de commandes

- Clonant anaconda environnement de base

  conda create --name <myenv> -- clone base

- Création d'environnement pyhon depuis anaconda

  conda create --name <myenv> python=3.8

- Activation de l'environnement python

  conda activate myenv

- Désactivation de l'environnement python

  conda deactivate

- Installation avec un fichier requirements.txt

  pip install -r requirements.txt

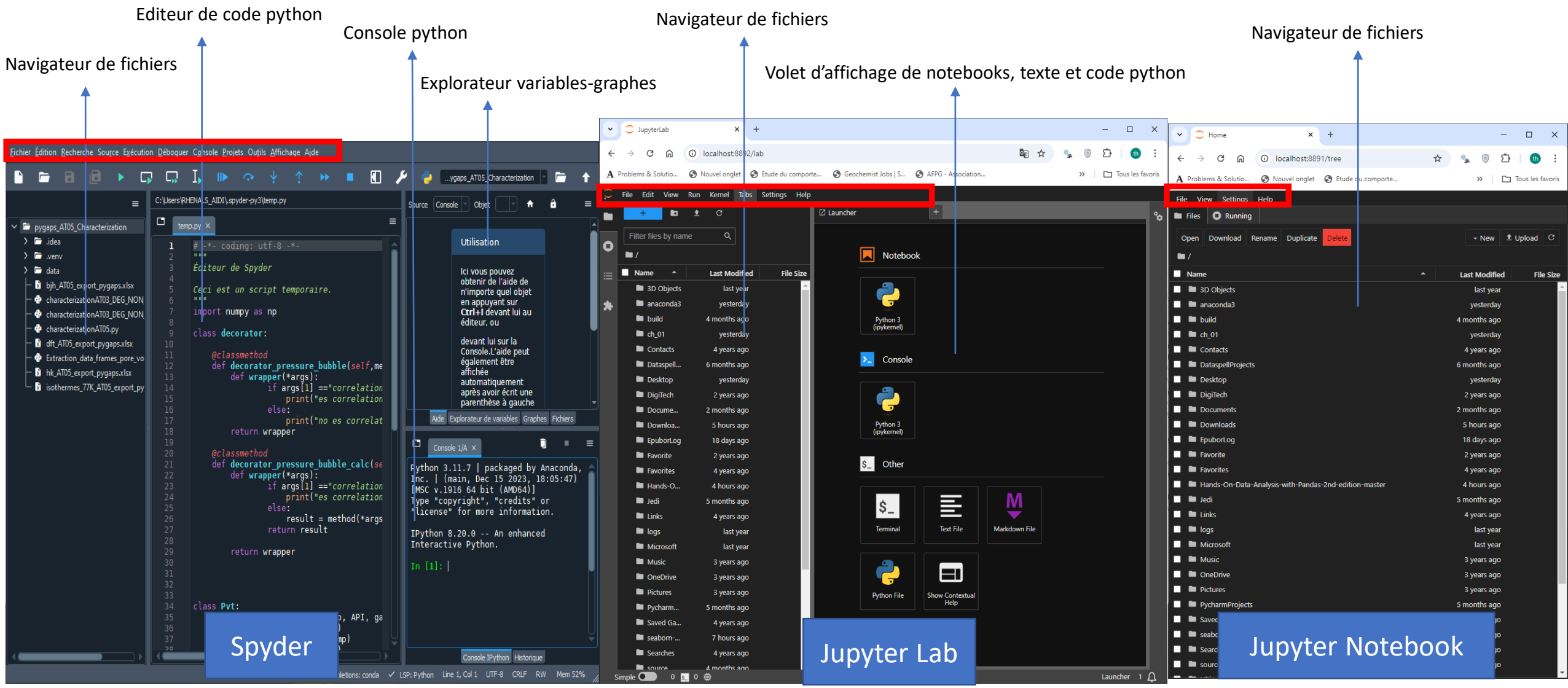
- Définition du workspace jupyter lab et jupyter notebook

  jupyter notebook --notebook-dir <workdir>
  jupyter lab --notebook-dir <workdir>

Tableau comparatif Jupyter Notebook/Lab/Spyder

Caractéristique	Jupyter Notebook	Jupyter Lab	Spyder
Type	Environnement de notebook interactif	Environnement de développement web interactif	IDE scientifique pour Python
Interface	Basée sur des cellules, simple et intuitive	Basée sur des onglets, personnalisable, interface web moderne	Interface classique d'un IDE, avec éditeur, console, explorateur de variables
Flexibilité	Moins flexible, mais suffisante pour de nombreux cas d'utilisation	Très flexible, hautement extensible grâce aux extensions	Moins flexible que JupyterLab, mais plus structuré que Jupyter Notebook
Collaboration	Facile à partager, mais moins collaboratif en temps réel	Très collaboratif grâce aux fonctionnalités de partage et de co-édition	Moins axé sur la collaboration en temps réel, mais peut être utilisé en réseau
Visualisation	Intègre bien les bibliothèques de visualisation (Matplotlib, Seaborn, etc.)	Intègre bien les bibliothèques de visualisation, plus de possibilités de personnalisation	Intègre bien les bibliothèques de visualisation, interface dédiée à l'inspection des données
Débogage	Débogage cellulaire, mais moins complet qu'un IDE classique	Débogage plus avancé, avec points d'arrêt, inspection de variables, etc.	Débogage complet avec toutes les fonctionnalités d'un IDE classique
Profiling	Possibilité de profiler le code pour optimiser les performances	Possibilité de profiler le code	Possibilité de profiler le code
Extensions	Moins d'extensions disponibles	Nombreuses extensions disponibles pour personnaliser l'interface et ajouter des fonctionnalités	Moins d'extensions disponibles que JupyterLab, mais une communauté active
Intégration avec d'autres outils	S'intègre bien avec d'autres outils de la stack data science (Git, Docker, etc.)	S'intègre bien avec d'autres outils de la stack data science	S'intègre bien avec d'autres outils de la stack data science
Utilisation typique	Exploration de données, prototypage rapide, création de rapports interactifs	Développement de projets data science plus complexes, collaboration en équipe, enseignement	Développement scientifique en Python, analyse de données, prototypage

Composantes basiques – Jupyter NB-LAB-Spyder



Quelques rappels basiques de programmation Python

- Quelques notions de base en programmation sur python sont nécessaires pour continuer le cours. ces notions incluent les concepts:
  - Type de données
  - Structures de control
  - Built-in structures
  - modules
  - fonctions
- Pour rappel, voici quelques exemples simples rappelant ces concepts (cliquer ici)

Références

- Shea, J. M. (2024). *Foundations of Data Science with Python*: CRC Press.
- Igual, L., & Seguí, S. (2017). *Introduction to data science : a Python approach to concepts, techniques and applications.*
- Provost, F., & Fawcett, T. (2013). *Data science for business : what you need to know about data mining and data-analytic thinking.* Sebastopol: O'Reilly.