

Ressources **Python** pour s'autoformer sur le **SSPCloud**

Onyxia ScaleUp

Lino Galiana

Insee

11 mars 2022

Présentation catalogue

Catalogue (1/2)

1. Modules d'autoformation sur sspcloud.fr/documentation;

- Notebooks pédagogiques, différents niveaux accessibles
- Environnements normalisés: pas de problèmes d'installation de *packages*
- Formation aux bonnes pratiques: modularité, contrôle de version, stockage séparé du code

2. Ressources **Python** 🐍 (et **Git** 💎):



- Initiation à Python sur sspcloud.fr/documentation ;
- Perfectionnement avec le cours de [2A ENSAE "Python pour data scientists et économistes"](#)



Présentation catalogue

Catalogue (2/2)

3. Ressources **R** (et **Git**

- Formations sur des points plus spécifiques. ex: [Travail collaboratif avec R](#)
- **utilitR**: environnement normalisés pour tester les exemples de la documentation
- Autres ressources **R**  ou **Python** 

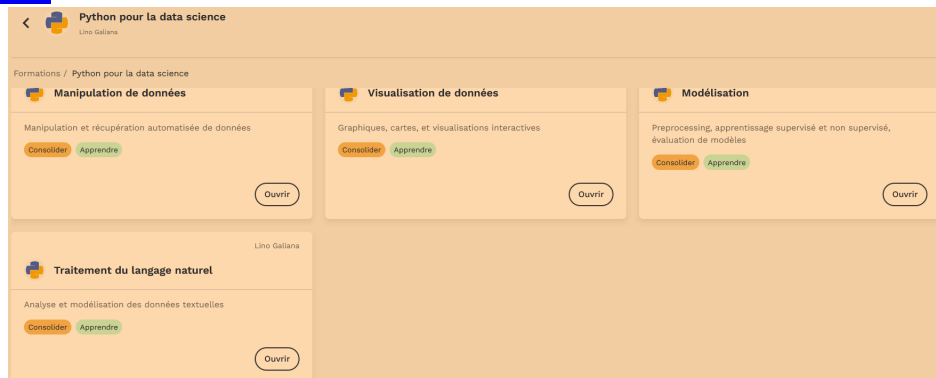
Ressources : cours de l'ENSAE

<https://linogaliana-teaching.netlify.app/> (dépôt 🔄)

Un parcours complet pour la *data-science* :

- Manipulation et structuration de données ;
- Visualisation ;
- Modélisation (économétrie et *machine learning*) ;
- Données textuelles (NLP) ;
- `Git` 📦 et `GitHub` 🔄.

[Les notebooks sont ici](#)



Ressources : cours de l'ENSAE

<https://linogaliana-teaching.netlify.app/> ([dépôt](#) )

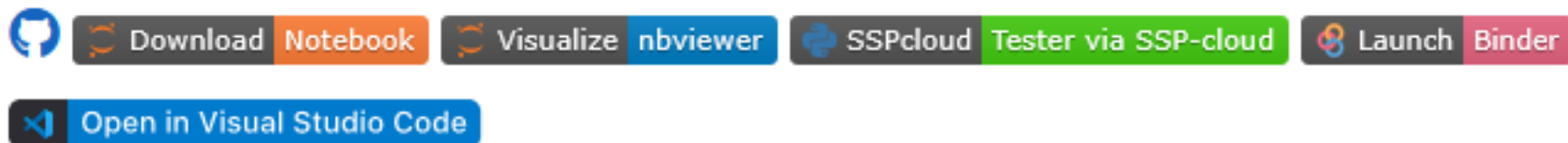
Introduire à la philosophie de la data-science :

- Reproductibilité et pérennité des projets
- Versioning et partage des codes
- Réduction des coûts de mise en prod

Mise à disposition tutoriels et exercices facilitée:


[Home](#) / [Homepage](#) / [Partie 1: manipuler des données](#) / [Webscraping avec python](#)

Webscraping avec python



Comment déployer sur le SSPCloud ?

Pourquoi ?

- Des notebooks sont disponibles sur [Github](#) 
([exemple](#))
- Objectifs:
 - permettre aux personnes intéressées de tester les exemples/exercices dans un environnement Jupyter
 - contrôler l'environnement d'exécution pour éviter les problèmes d'installation
- Solution: déploiement automatique sur le SSPCloud

Comment déployer sur le SSPCloud

Avantages

- Pour les formés:
 - Pas d'installation d'Anaconda, Jupyter, modules Python sur Windows, etc.
 - Très simple d'usage : un click et on a un notebook opérationnel
- Pour les formateurs:
 - Plus à gérer plusieurs configurations ou des spécificités propres à chaque OS ou poste de travail
 - Possibilité d'associer ça à l'intégration continue pour tester la reproductibilité
- Challenge:
 - Mise en place assez exigeante initialement...
 - ... mais gains de confort rapidement

Comment déployer sur le **SSPCloud** ?

Contrôler la portabilité de sa formation (1/2)

- L'idéal est que les formés aient un environnement prêt à l'emploi:
 - Un environnement de développement avec des modules adéquats (auto-complétion, Gitlens...)
 - Pas de dépendance à installer..
 - ... ou installation facile de celles-ci
- Pour les formés, possibilité de sauvegarder les codes avec **Git** ♦ (formation aux bonnes pratiques par l'usage)

Comment déployer sur le SSPCloud ?

Contrôler la portabilité de sa formation (2/2)

- Construire une image **Docker** 🐳 qui contient les éléments minimaux pour pouvoir démarrer :
 - Exemple **Python** 🐍: un environnement **Anaconda** prêt à l'emploi ;
 - Exemple **R** 📊: dépendances listées (ou préinstallées) dans un fichier **Description**
- Chaque **commit** sur **Github** 🏠 permet de mettre à jour les notebooks disponibles dans le catalogue

Comment déployer sur le SSPCloud ?

Mettre à disposition sa formation

- Créer un script `init.sh` pour, quand l'image est active, automatiquement :

1. Clôner le dépôt / télécharger les fichiers d'intérêt

2. Définir les fichiers à ouvrir automatiquement

- Exemples :

- `Python` avec le cours ENSAE

- `R` avec `utilitR`

- Créer l'URL à partager qui combine tous ces éléments.

- <https://datalab.sspcloud.fr/launcher/insee-frlab-helm-charts-datascience/jupyter?onyxia.friendlyName=%C2%ABpython-datascience%C2%BB&init.personalInit=%C2%ABhttps%3A%2F%2Fraw.githubusercontent.com%2Fmaster%2Fsspcloud%2Finit-jupyter.sh%C2%BB&init.personalInitArgs=%C2%ABvisualisation%20maps%C2%BB&security.a>

Retour d'expérience de l'ENSAE

- Plus à galérer entre les étudiants utilisant les postes de l'ENSAE et ceux ayant leur propre ordinateur ;
- Installation des modules très pratique (merci Linux 🐧 !)
- Beaucoup de ressources RAM et CPU sur SSPCloud (vs réseau ENSAE) ;
- Formation à Git 💎 par l'usage