Formation Python pour la Data Science

Objectifs : Apprendre à utiliser le langage Python et ses principales librairies scientifiques pour traiter, visualiser et modéliser les données en Data Science

Compétences visées : - Identifier et utiliser les librairies incontournables de Python pour la data science : La Scipy Stack
- Identifier et utiliser les principales librairies de visualisation de données et notamment orientées cartographie
- Savoir manipuler des données volumineuses ne tenant pas en mémoire
- Avoir une bonne compréhension de l'écosystème scientifique de Python, savoir trouver ses librairies et juger de leur qualité

Durée : 5 jour(s) (35 heures)

Public : Développeurs, chefs de projets, data scientists développant des applications scientifiques requérant d'importantes capacités de calculs

Pré-requis : Pour suivre ce stage dans de bonnes conditions, il est recommandé d'avoir suivi en amont la formation Python – Bases et introduction aux librairies scientifiques

Méthode pédagogique : Nos formations sont majoritairement pratiques (70%), les concepts théoriques sont illustrés d'exemples et exercices. Les supports sont essentiellement construits avec les notebooks Jupyter/Lab et sont suffisamment détaillés pour être repris seul(e) après la formation.
Pédagogie active mêlant exposés, exercices et applications pratiques dans le logiciel Python.

Modalités d'évaluation : Un formulaire d'auto-évaluation proposé en amont de la formation nous permettra d'évaluer votre niveau et de recueillir vos attentes. Ce même formulaire soumis en aval de la formation fournira une appréciation de votre progression.
Des exercices pratiques seront proposés à la fin de chaque séquence pédagogique pour l'évaluation des acquis.
En fin de formation, vous serez amené(e) à renseigner un questionnaire d'évaluation à chaud.
Une attestation de formation vous sera adressée à l'issue de la session.
Trois mois après votre formation, vous recevrez par email un formulaire d'évaluation à froid sur l'utilisation des acquis de la formation.

Accessibilité : Vous souhaitez suivre notre formation Formation par ville et êtes en situation de handicap ? Merci de nous contacter afin que nous puissions envisager les adaptations nécessaires et vous garantir de bonnes conditions d'apprentissage

Tarif : Présentiel : 3250 € HT - Distanciel : 3000 € HT (-10% pour 2 inscrits, -20% dès 3 inscrits)

Nos prochaines sessions

Distance
du 2 au 6 décembre 2024
du 24 au 28 mars 2025

Lyon
du 21 au 25 octobre 2024
du 12 au 16 mai 2025

Paris
du 18 au 22 novembre 2024
du 16 au 20 juin 2025

Toulouse
du 7 au 11 octobre 2024
du 14 au 18 avril 2025

Programme :

- L'écosystème scientifique Python

Il n'est pas facile d'y voir clair dans l'écosystème scientifique de Python tant les librairies sont variées et nombreuses.
Cette présentation vous apportera une vue d'ensemble et les éléments clefs qui vous aideront à choisir vos librairies et outils de travail pour vos projets de data science avec Python.

- Les incontournables: Numpy, Scipy, Pandas, Matplotlib et iPython qui sont le ciment de toutes les autres librairies scientifiques
- Panorama des librairies et logiciels scientifiques par domaine
- Les critères permettant de juger de la qualité d'une librairie

- Calculer avec des nombres réels: comprendre les erreurs de calculs

Les nombres réels, dans la plupart des langages, dont Python, utilisent la norme en virgule flottante. Celle-ci n'est pas précise et peut générer des erreurs de calcul parfois bien gênantes.

- La représentation des nombres réels
- Comprendre les erreurs de calculs et les contourner

- La scipy stack

La librairie Numpy qui signifie Numeric Python est la première que vous devez apprendre. Elle constitue avec Scipy, Matplotlib et Pandas le socle sur lequel s'appuient toutes les autres librairies scientifiques.

- Manipuler des tableaux de nombres: Numpy
    - Différences avec les listes Python
    - Création, sélection, filtres et principales fonctions

- Visualiser ses données: Matplotlib
    - Les concepts de la librairie
    - Principaux graphiques: nuages de points, courbes, histogrammes, boxplot, ...
    - Fonctionnalités avancées: 3D, légendes, colorbar, manipuler les axes, annotations, ...
- Analyse de données: Pandas
    - Les fondements de la librairie: Manipuler des données de type CSV et Excel
    - Séries et Dataframes
    - Index, sélection de données, filtres/recherche, agrégations, jointures et fonctions avancées
    - Manipuler des séries temporelles
- Les fonctions mathématiques avancées: Scipy
    - Statistiques, optimisation, interpolations/régressions, traitement d'images

- Visualisation de données

Bien que Matplotlib constitue la première librairie de visualisation que vous devrez apprendre, elle possède 2 limites majeures: elle ne sait pas gérer les données volumineuses et n'est pas adaptée au Web. Mais Python a su développer un riche écosystème de visualisation de données qui devrait pouvoir répondre à toutes vos attentes.

- Présentation de l'écosystème de visualisation de données de Python
- Les librairies orientées Web: Bokeh, Altair et Plotly
- Les "écosystèmes" PyViz et HoloViz
- La visualisation de données volumineuses/big data avec DataShader
- Les statistiques avec Seaborn

- Visualiser des données géospatiales

Posséder des données disposant de coordonnées géospatiales apporte une toute autre dimension à leur représentation. Python est très bien outillé dans ce domaine.

- Convertir ses données d'un système de coordonnées à l'autre
- Cartographie interactive "à la Open Street Map/Google Maps" avec Folium/iPyleaflet
- Cartographie statique avec Cartopy
- Autres librairies géospatiales

- Manipulation de données volumineuses

Numpy et Pandas sont 2 librairies incroyables, mais elles ont 2 limites majeures: elles ne savent pas traiter des données de très grande volumétrie qui ne tiennent pas en mémoire et ne savent pas toujours paralléliser leurs calculs.

Python a su développer des solutions.

- Les librairies h5py, pytables, netcdf4, xarray, iris, parquet permettant de lire vos fichiers scientifiques
- Paralléliser ses calculs avec Dask
- Paralléliser ses calculs avec CuDF
- Manipuler des dataframes gigantesques avec Dask

- Personnalisation

Sous réserve de contraintes techniques ou de confidentialité, nous vous proposons de personnaliser la formation en réalisant des exercices directement sur vos données métiers.

Date de dernière modification : 6 juin 2024