

Analyse de données

Séance 1 - Comprendre un dataset

Master MIAS - M1
hadriensalem@gmail.com

Introduction

L'importance de la donnée en 2022

Donnée

Valeur qui porte de l'information

Littérale, numérique,
booléenne, etc.

Quantités, faits, statistiques, etc.

⇒ Exploiter des données, c'est utiliser des informations à son avantage

Exploiter des données, c'est utiliser des informations à son avantage

Toute activité génère de la donnée, donc toute activité est sujette à l'exploitation de données

Source de l'image :

[What is Data Science? sur hackr.io](https://hackr.io/what-is-data-science)



Deux grands types d'exploitation

- ❖ Analyse
- ❖ Apprentissage

... Rendus possibles par des avancées technologiques et théoriques

Quelques exemples en santé

L'exploitation des données de santé a de nombreuses applications en recherche et dans l'industrie.

Épidémiologie

Prédiction de maladies

Gestion des plannings

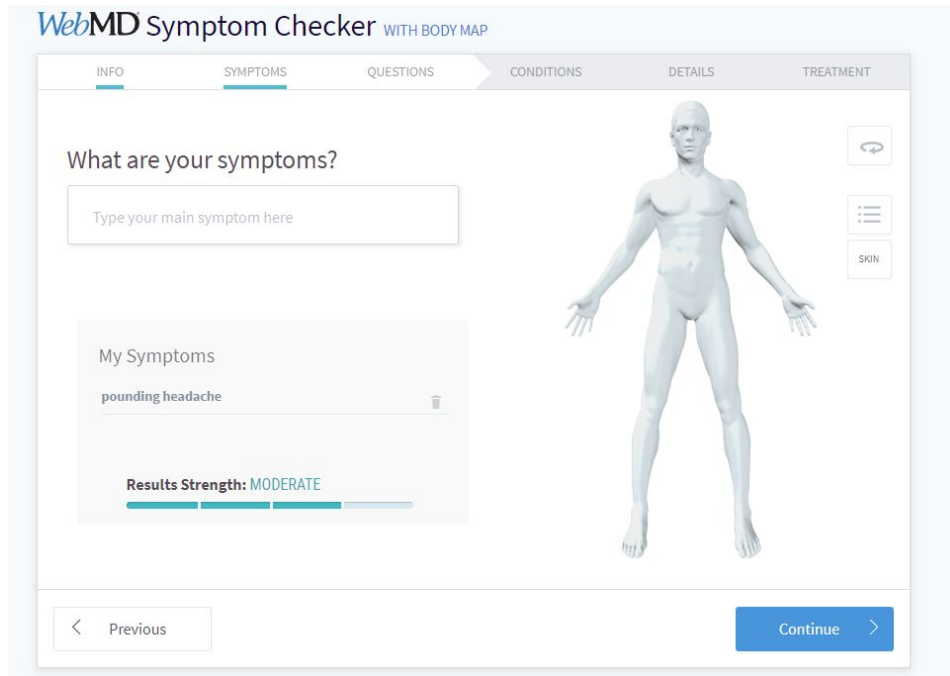
Alertes de santé

... etc.

Quelques exemples en santé

Les **symptom checkers** permettent aux patients d'évaluer leur propre état pour décider s'ils doivent consulter un médecin ou aller aux urgences.

Ils fonctionnent sous la forme d'un chat bot.



The screenshot displays the WebMD Symptom Checker interface, titled "WebMD Symptom Checker WITH BODY MAP". The interface features a navigation bar with tabs: INFO, SYMPTOMS (active), QUESTIONS, CONDITIONS, DETAILS, and TREATMENT. The main content area is divided into two columns. The left column contains a text input field labeled "What are your symptoms?" with the placeholder "Type your main symptom here". Below this is a section titled "My Symptoms" which lists "pounding headache" with a trash icon. At the bottom of this section, it shows "Results Strength: MODERATE" with a corresponding progress bar. The right column features a 3D body map of a human figure. To the right of the body map are three buttons: "HEAR", "SEEK", and "SKIN". At the bottom of the interface, there are two buttons: "Previous" and "Continue".

<https://symptoms.webmd.com/>

Un peu de vocabulaire

Dataset

Un ensemble de données

Big Data

Un énorme
ensemble de données

Data Analysis

Observer des données
pour les comprendre

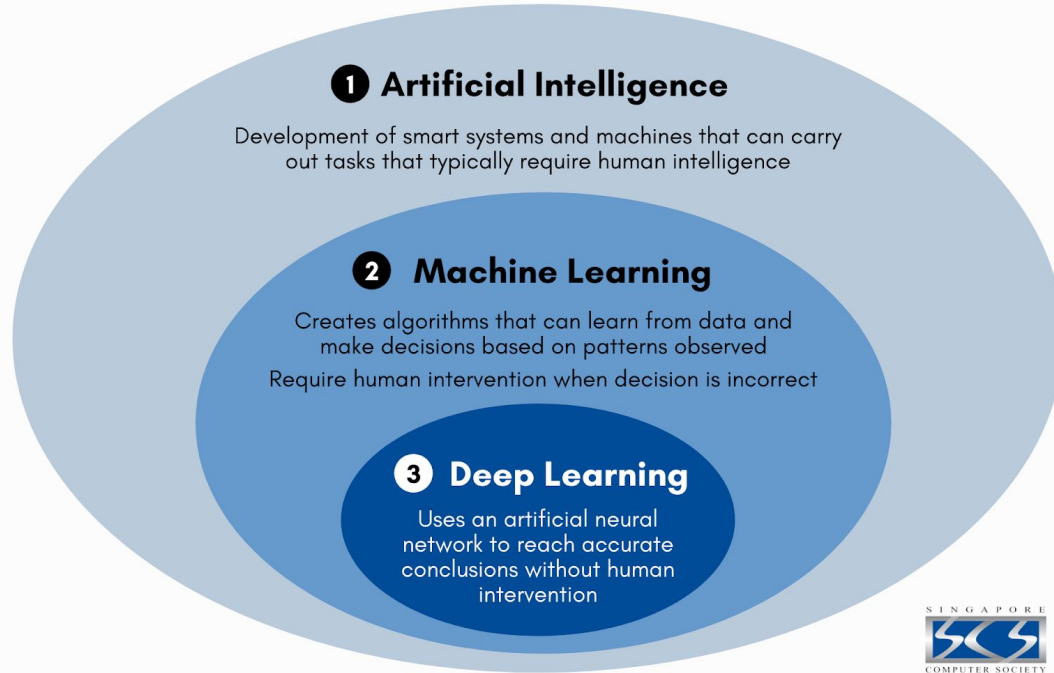
Data Engineering

Préparation des données
pour analyse

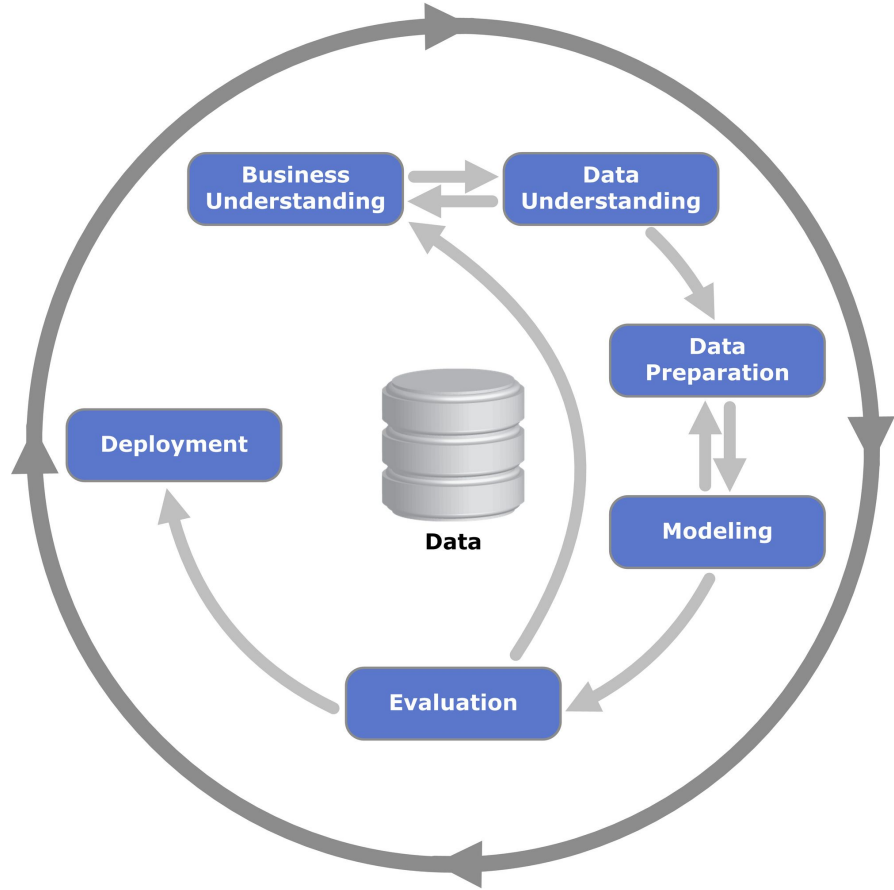
Data Science

Modélisation des données

ARTIFICIAL INTELLIGENCE VS MACHINE LEARNING VS DEEP LEARNING



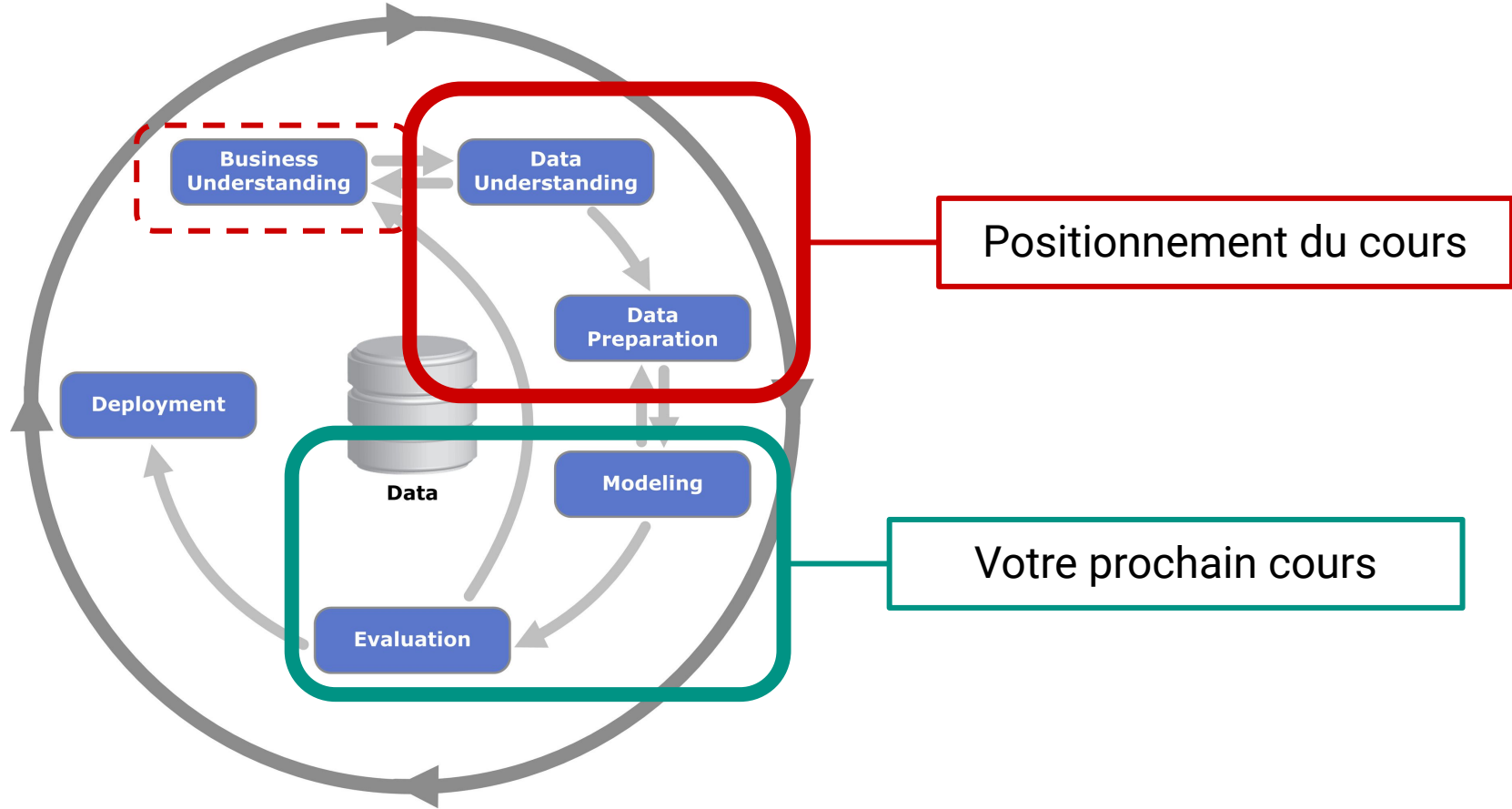
Comment exploite-t-on
des données ?



La méthode CRISP-DM

Cross-Industry Standard Process for Data Mining

- Publiée en 1999
- Méthode suivie dans l'industrie
- Toujours d'actualité



Plan de cours

Séance 1 : Comprendre un dataset

- ❖ Étude exploratoire des données
- ❖ Visualisation des données

Séance 2 : Préparer un dataset (1/2)

- ❖ Overview des types de pré-traitement des data
- ❖ Gestion des valeurs manquantes & absurdes

Séance 3 : Préparer un dataset (2/2)

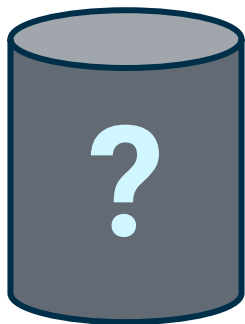
- ❖ Introduction à la réduction de dimension
- ❖ Typologie des algorithmes de machine learning

Étude exploratoire des données

Introduction

Étude exploratoire des données

La première chose à faire avec un dataset
est d'**apprendre à le connaître**



Qu'est-ce qu'on
cherche à apprendre?



Qu'est-ce qu'on cherche à apprendre?

Questions d'ordre général (lire et compter)

- Quelles données le dataset contient-il ?
- Comment ces données sont-elles représentées ?
- De quel type sont ces données ?
- Y a-t-il des "trous" dans les données ?
- Y a-t-il des doublons dans les données ?
- Les données sont-elles équilibrées ?

Questions plus avancées (comprendre)

- Quelle est la distribution statistique des données ?
- Y a-t-il des corrélations entre les colonnes ?
- Si oui, lesquelles ?

⇒ Plus on avance dans l'exploration, plus les questions qui émergent se font nombreuses.

Étude exploratoire des données

Mise en pratique

Quels langages pour l'analyse de données ?

Les plus utilisés sont Python et R, mais il en existe bien d'autres (e.g. Kotlin, Java, etc.).

De nombreux packages sont disponibles dans ces langages pour exploiter, analyser et modéliser les données.



Nous utiliserons le langage Python

Quels logiciels pour l'analyse de données ?

Par souci de simplicité, nous exécuterons notre code sur des Jupyter notebook via un logiciel en ligne.

Faire tourner du code en local nécessite d'installer Python et ses packages soi-même.

colab
kaggle



ANACONDA®



Quels packages pour l'analyse de données ?

De nombreuses librairies (ensembles d'objets et de fonctions) existent en Python pour différents aspects de l'exploitation de données.



Mathématiques



Manipulation de datasets



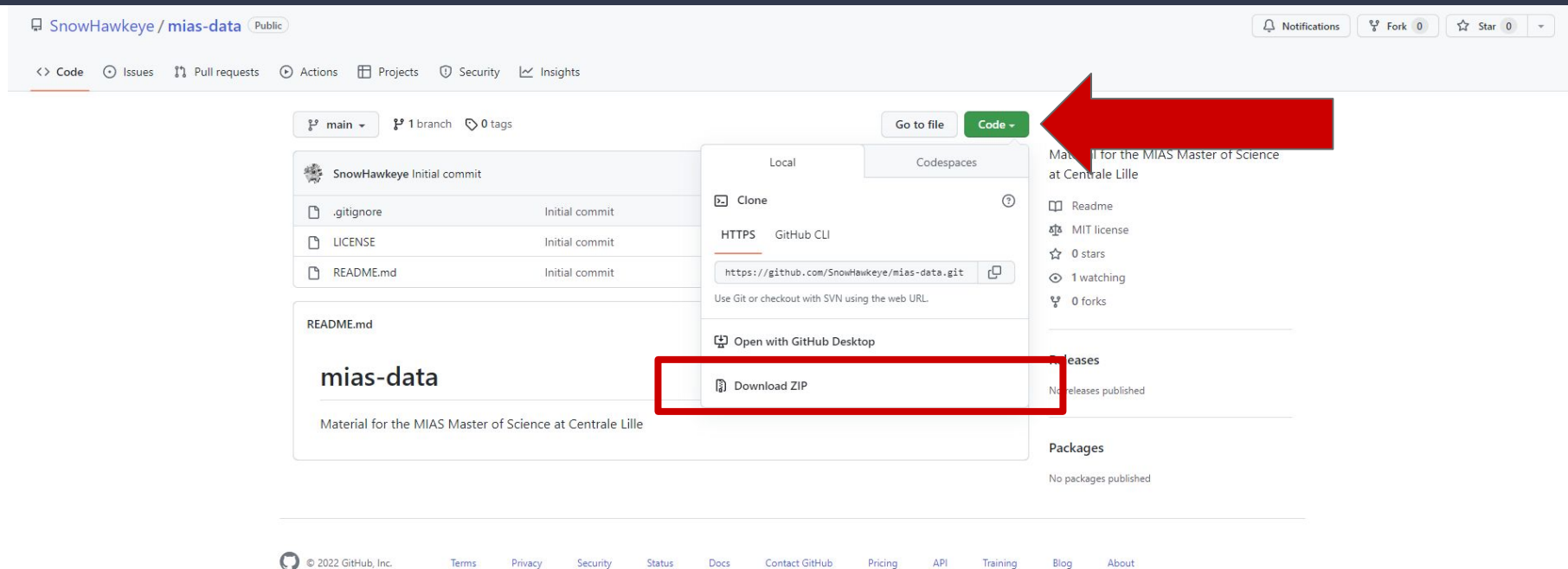
**Machine Learning
(hors Deep learning)**



Affichages



Récupération du notebook



The screenshot shows the GitHub repository page for `SnowHawkeye/mias-data`. The repository is public and has 0 forks, 0 stars, and 1 watching. The repository contains a single commit on the `main` branch. The commit message is `SnowHawkeye Initial commit`. The files listed are `.gitignore`, `LICENSE`, and `README.md`, all marked as `Initial commit`. The `README.md` file is expanded, showing the title `mias-data` and the description `Material for the MIAS Master of Science at Centrale Lille`. A red arrow points to the `Code` button, and a red box highlights the `Download ZIP` option in the dropdown menu.

© 2022 GitHub, Inc. Terms Privacy Security Status Docs Contact GitHub Pricing API Training Blog About

À l'adresse : <https://github.com/SnowHawkeye/mias-data>

Ouvrir le notebook

Il suffit de l'importer sur le logiciel de son choix.

Datalore permet une édition simultanée entre plusieurs collaborateurs.
(Share > Manage invitations)

Manage invitations



< Go back

Note that all users with edit permissions **will use the computational resources of the notebook owner** when they open it. To end machines run by such users, revoke their access using this dialog. The owner can also use the Running machines dialog to terminate any running computation.

Nobody except users invited by email have access.



No access link



Share this notebook with someone by email

mybestfriend@gmail.com X

Invite users

can view



Send invitation

can view

can edit

Au travail !

Le notebook contient une mise en pratique et des questions à traiter.

Visualisation des données

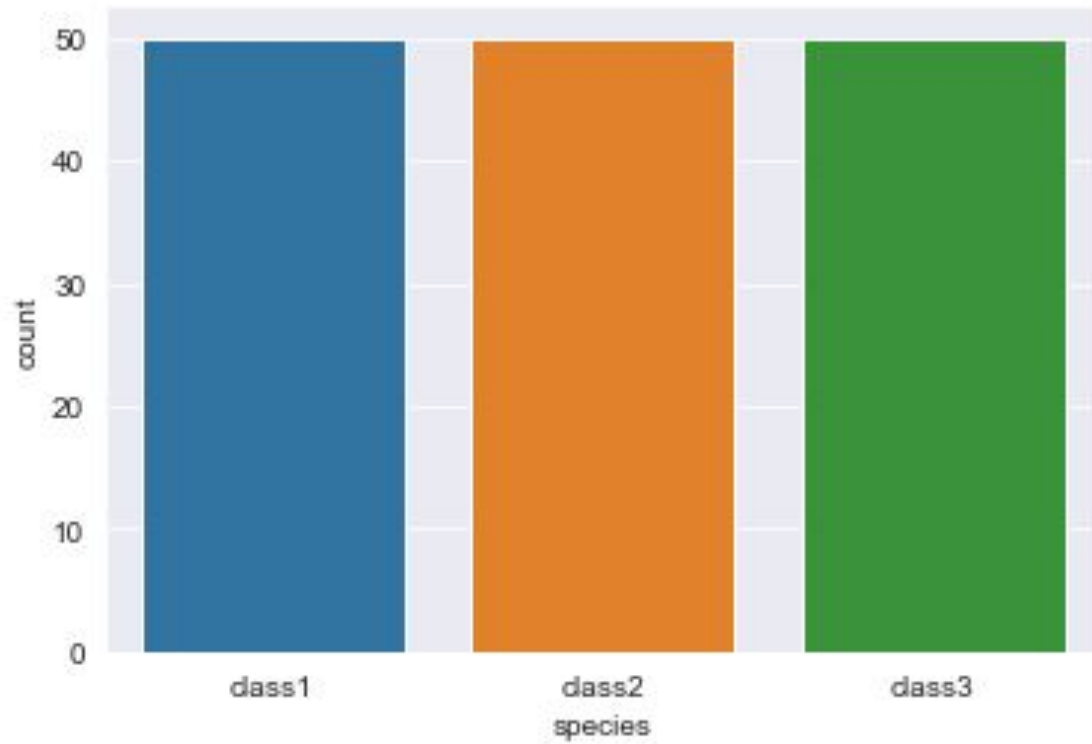
Quels sont les intérêts
de la visualisation des
données ?

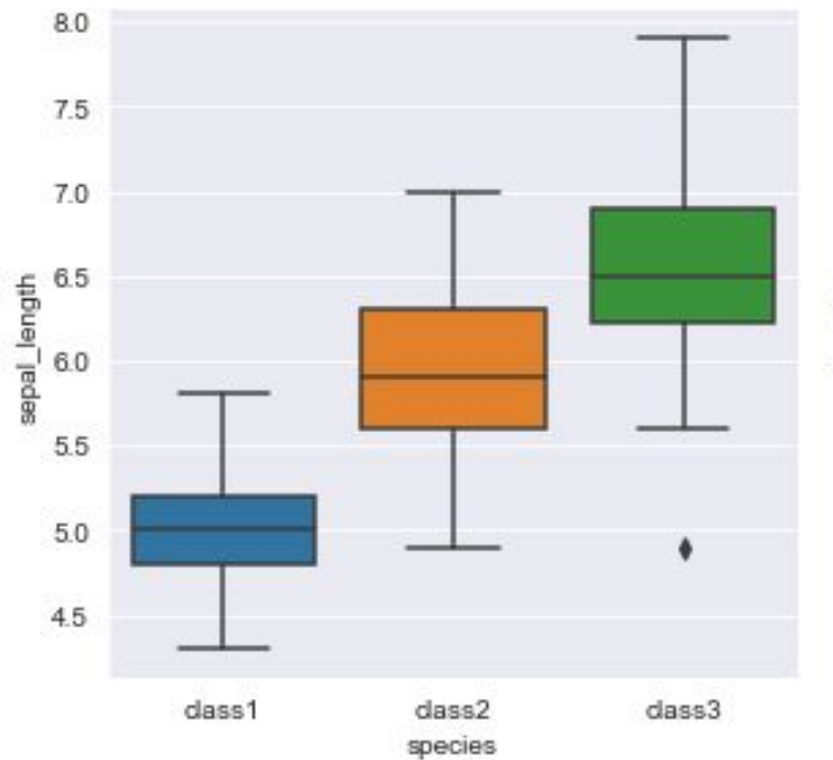


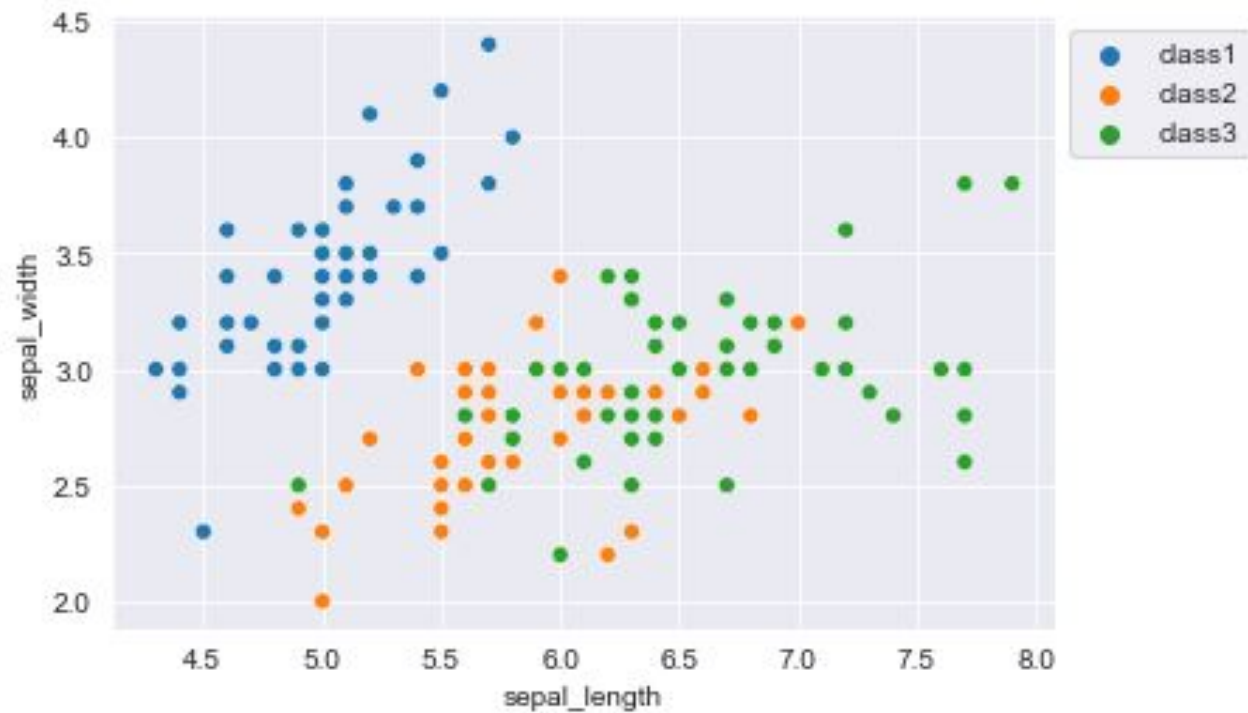
Quels sont les intérêts de la visualisation des données ?

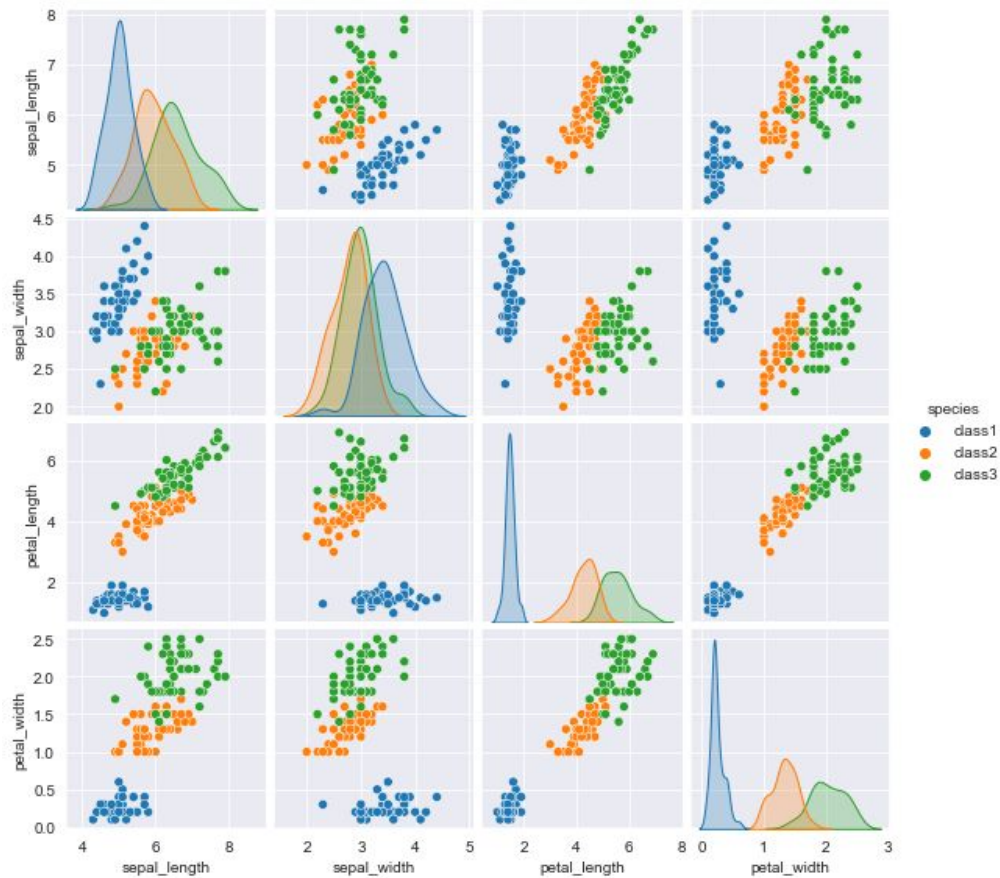
Les intérêts de la visualisation

- La visualisation peut servir à **comprendre des données** : détecter les *outliers*, visualiser la distribution d'une variable, le nombre d'éléments d'une classe, la corrélation entre des variables, l'importance des différentes features, etc.
- Elle peut permettre de **choisir un algorithme** (par exemple dans le cas de données linéairement séparables)
- Les graphes sont également un **outil de communication essentiel** dans un contexte professionnel (pour convaincre ou expliquer, en particulier à des gens qui ne sont pas experts techniques)

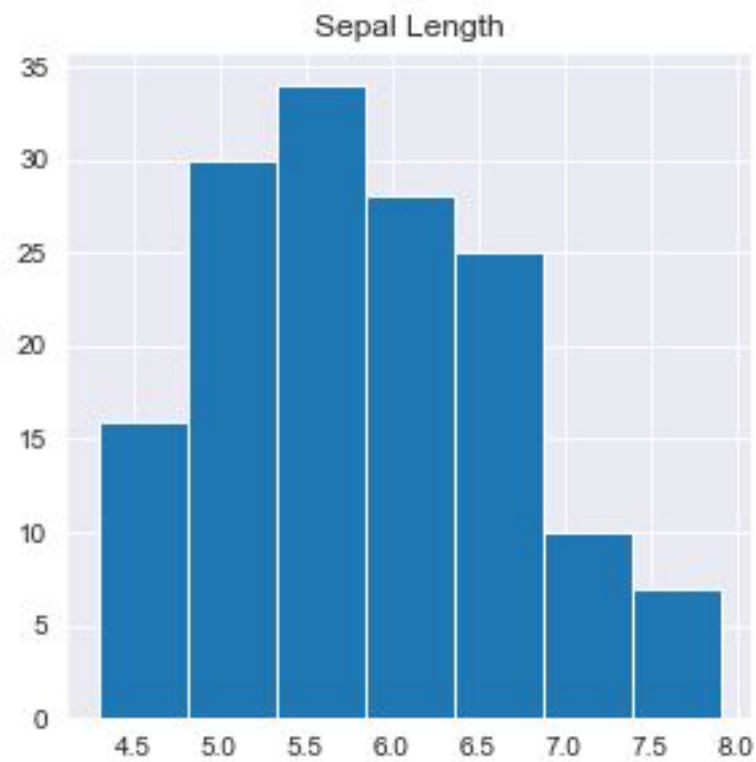


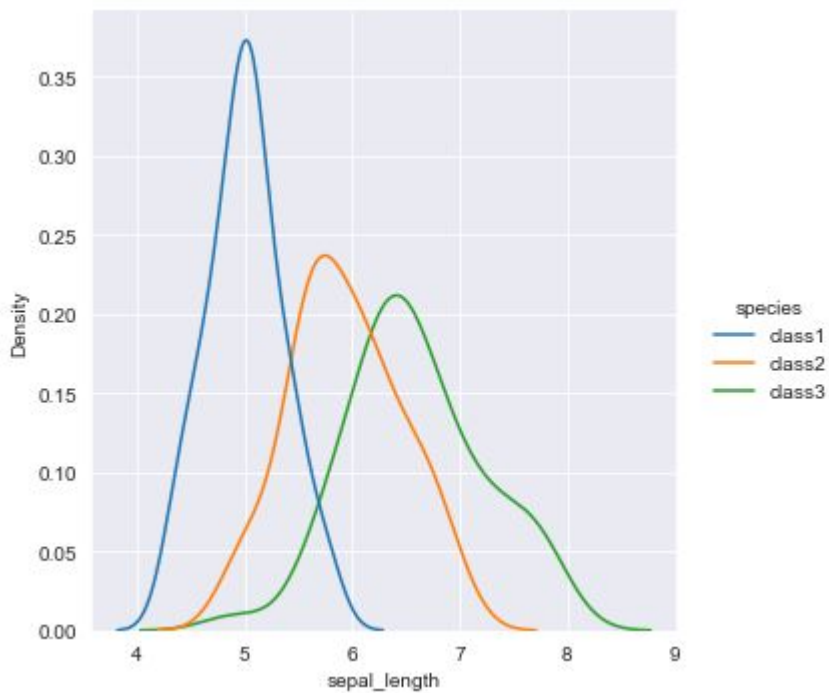
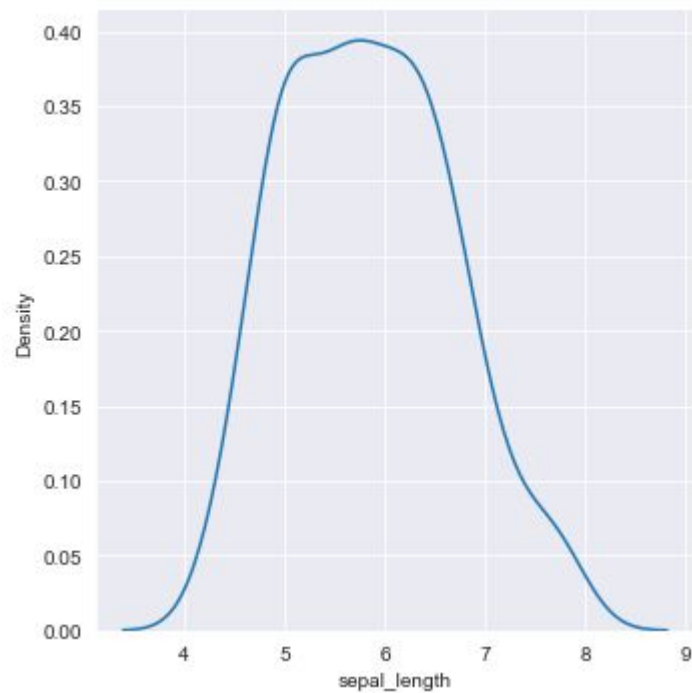




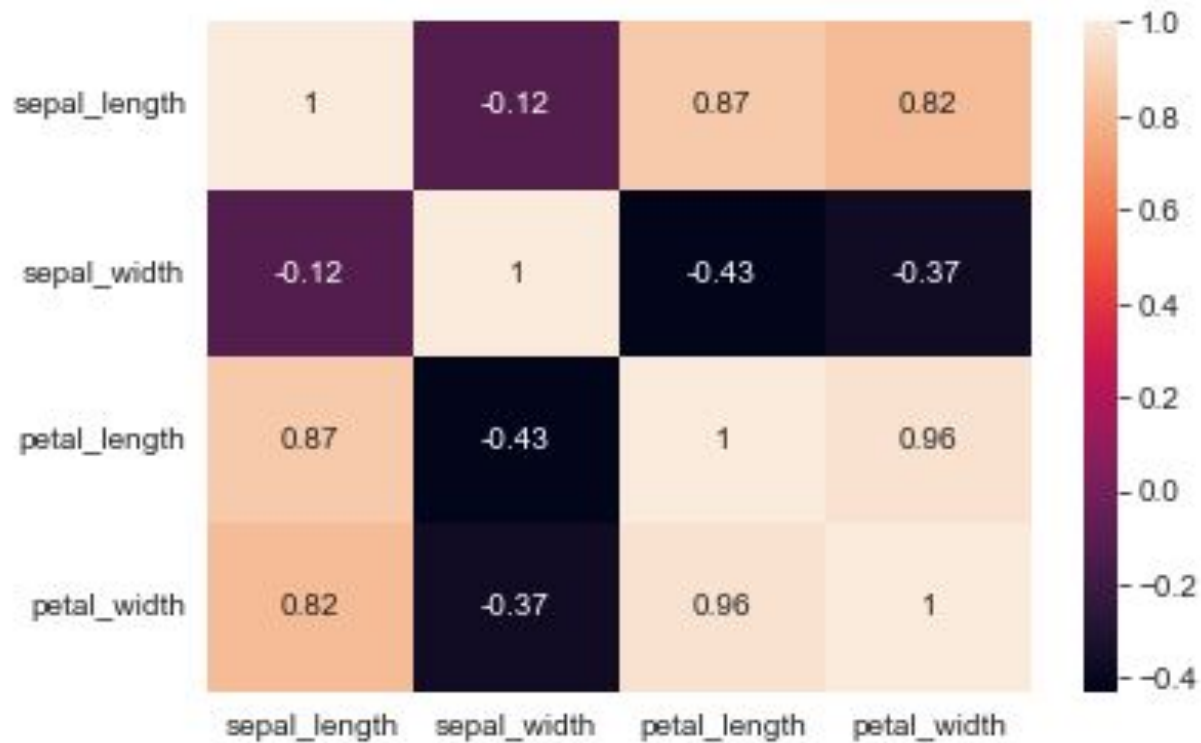


Différents types de graphes : pairplot





Différents types de graphes : `displot`

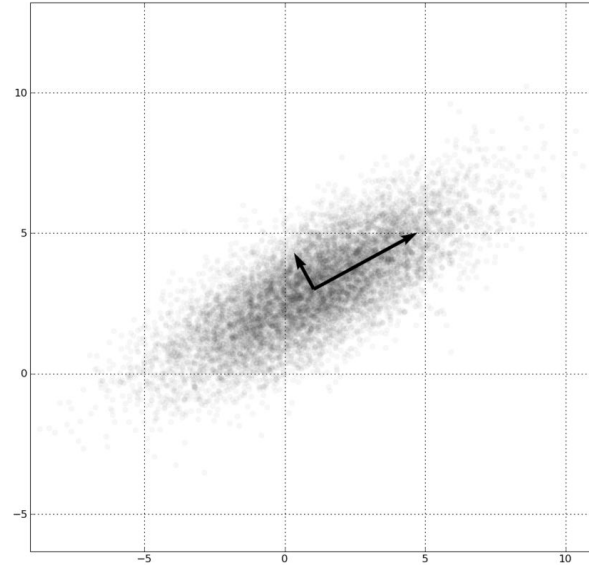


$$\text{Cov}(X, Y) \equiv \mathbb{E}[(X - \mathbb{E}[X]) (Y - \mathbb{E}[Y])]$$

Covariance de deux variables aléatoires

Quantifie à quel point un changement
dans une variable implique un
changement dans l'autre variable

Typiquement en machine learning, on
aime les (co)variances élevées



À représenter avec une heatmap : la covariance

$$r = \frac{\sum (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Coefficient de corrélation de Pearson

Quantifie à quel point les variables évoluent de façon similaire

À représenter avec une `heatmap` : la corrélation