Data management (Bloc 2/3 of the Certificate "Fundamentals of Data Science")

ECTS : 3

Description du contenu de l'enseignement :

Data science is an interdisciplinary field that is rapidly evolving. Many companies have widely adopted machine learning and artificial intelligence methods to power many applications that have captured the imagination of society at large. Data systems and data engineering are an inevitable part of all these large-scale data-driven applications and decisions, as ML/AI methods are powered by massive collections of potentially heterogeneous and messy datasets and, as such, should be managed as part of an organization's overall data lifecycle.

This course corresponds to the third block of the Certificate “Fundamentals of Data Science”. This Certificate is designed to train and familiarize professionals with the key technologies in this interdisciplinary field, with the aim of enabling them to take full advantage of the opportunities offered by data science and to become active players in this field within of their organizations. This is an accelerated training focused on the key modules of the profession of data scientist, in particular the management of massive data and machine learning.

Course outline:

Module 1 : Why shall we engage a Data transformation program (1h)*

- Introduction
  - The Role of Data in a company
  - Review of the evolution of Data topics
- Data value chain
- Presentation of the pillars of a Data transformation (challenges / objectives)
  - The Data Strategy
  - Data Management & Governance
  - Analytics
  - IT
  - Project to Product team

Module 2: Data Management (4h)

- General presentation of the main concept of the framework (DAMA)
- Structure & organization (roles & responsibilities)
- Lineage & Metadata: data knowledge
- The importance of data quality
- Privacy / GDPR
- Data types and their characteristics
  - Structured
  - Unstructured data
- Examples of architectures
- Main tools to manipulate Data

Module 3: Case Analysis - From Theory to Practice: data retrieved from both a database and an excel file. (10h)

- From integration to visualization
  - Integration of data from Excel file via Python
    - Representation, cleaning, recoding
    - Aggregates
    - Merge and join

- *Practical work 1*
- Data integration from the database
  - Relational model
  - Introduction to SQL
  - SQL in Python

- *Practical work 2*
- Beyond SQL, other possible cases (NoSQL)
- Problems encountered when reconciling data (duplication, quality, veracity) > Can you put your trust into your data

&