**D A T A S C I E N C E C O N S U L T I N G S e s s i o n 1**

January 23rd, 2023

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 1

The team with whom you will spend your session

**Olivier AULIARD** Director of the course – Chief Data Scientist 

olivier.auliard@

capgemini.com

**Ines EL KASMI** 

Data Scientist

ines.el-kasmi@

capgemini.com

**Guillaume REMONT** Consultant

guillaume.remont@ capgemini.com

**Fatima Zahra MJERREB** Consultant 

fatima-zahra.mjerreb@ capgemini.com

**Thibault VENET** 

Data Scientist

thibault.venet@

capgemini.com

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 2

Agenda

**1. Who are we?**

2. Course modalities

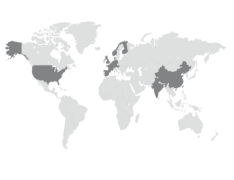
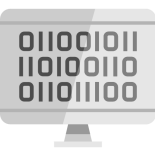
3. Analysis objectives & approach

4. Case presentation

5. Data collection

6. Html presentation & Selectors 7. Scraping with Selenium 8. Summary of the session

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 3

Capgemini Invent, a leader in digital & data transformation 

**Data Driven**

Blending **the power of data** with strategy, creative, and technology expertise to shape the future of our clients

**Our brand** 

**ecosystem**

**Leader on the market**

has been positioned by Gartner Inc as a **“Leader”** in its 2022 Gartner Magic Quadrant for Data and Analytics Service providers



**World class footprint**

30 offices, + 11 000 consultants, coverage of more than 80% companies of CAC 40 and DAX 30 at CxO level



Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 4

We are a hybrid consulting firm orchestrating business, data science, technologies, creative skills from vision to delivery

**We stimulate**

**ideation & disruptive vision**

By helping you finding the

unconventional ideas to hack your

business

**We ensure go-to-market for your best ideas**

From digital assets to business

value



**Digital products & services with high added value**

**We provide end-to end commitment & delivery** 

By involving & orchestrating the different required capabilities

**We innovate at**

**startup speed for a fast delivery**

Through proven agile and lean

methodologies

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 5

**Business expertise from projects with diverse clients**

****Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 6

Agenda

1. Who are we?

**2. Course modalities**

3. Analysis objectives & approach

4. Case presentation

5. Data collection

6. Html presentation & Selectors 7. Scraping with Selenium 8. Summary of the session

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 7

What are your expectations for this course? 

Go to **www.menti.com** and use the code **7147 4670**

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 8

Objectives of the case study

**Handle a business problematic associated to data** Increase both knowledge and skills on these topics 

**Learn how to identify & implement the required analysis** Handle a data project from the beginning to the end 

**Understand the strategic & transformation stakes** Qualify and quantify the associated stakes

**Grasp the consulting aspects** 

Learn how to manage these kinds of projects

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 9

Organization of the course



**Content**

• Each session focused on a **specific topic**

• **Balance** between theoretical explanations, brainstorming and hands-on applications

• **Tasks** to be prepared between sessions and to be presented in the beginning of the next session

**Format**

• **Weekly sessions on Monday afternoons** – from 2pm to 6pm

• **In Ecole Polytechnique & 147** (Capgemini offices) if the health conditions are met (virtual classes on Microsoft Teams otherwise)

• **Animation by Data Scientists & Consultants** from Capgemini Invent

**Practical organization**

• **Teamwork :** divided in groups of 5, each group with an assigned coach

• **Data Science tools** : Jupyter notebook on your local computer

• **The team is available via Teams** :

− **Discuss** among yourselves

− **Ask questions** to coaches

− **Get access to all the questions & answers** – including from the other students & groups

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 10

Planning & key steps of the course

**7 courses**

**Jan. 23 Mar. 13**

**Jan. 30**

**Feb. 06**

**Feb. 13**

**Feb. 20**

**Mar. 06**

****

**Session 1 Kick-off**

**Session 2 Data Prep.**

**Session 3 Embedding**

**Session 4**

**Client meeting**

**Session 5 Embedding**

**Session 6**

**Sentiment Analysis**

**Session 7**

**Final presentation Hackathon**

**Regular sessions:** Synthetic status update on the progress made and foreseen objectives to be prepared 

**Committees:** Complete pres. of the progress made, results; difficulties and foreseen next steps to be prepared 

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 11

Evaluation

**40% 20% 40%**

Homework

At each session you might have homework. We will ask you to send us your results and will evaluate them !

Intermediate restitution

At the 4th session, we will simulate a client meeting with intermediary results. It will be a presentation where you need to summarize what you’ve

seen/done until then.

Final restitution

On the last day, you will make a final presentation that will enable you to present the business insights produced throughout from your data Project.

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 12

Agenda

1. Who are we?

2. Course modalities

**3. Analysis objectives & approach**

4. Case presentation

5. Data collection

6. Html presentation & Selectors 7. Scraping with Selenium 8. Summary of the session

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 13

First of all … what is a data use case?

**1** 

**2**

**3**

**1**

**3**

**5**

**4**

**4**

**2**

**5**

**Motor**

Robust technical platforms

**Fuel**

Comprehensive data sources

**Tires, wheel**

Digital solutions & touchpoints

**Technical team**

Advanced analytics capability

**Driver**

Identified end-users

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 14

Typical approach for use case delivery projects **The operating model is based on a progressive approach to ensure the correct product delivery**

2 – 3 weeks Low investment

**Go / No go**

2 – 4 weeks

High investment

**Go / No go**

X \* 3 – 4-week sprints High investment

**Scoping / Qualification Sprint 0**

**Build **

**… …**

Does the project bring sufficient **value** ? Is it **feasible** ?

Is **everything ready** to launch

the project? **Delivery** of industrialized data products

***PoC / PoV***

If necessary, launch of a PoC / PoV to prove the project feasibility / value

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 15

❶ **Business needs**

• Elaborate a description of needs and matching features to respond

• Present the breakdown of features into user stories

• Design the prototype of the solution (if applicable)

• Process data and evaluate the algorithms & logic to be implemented

Zoom on Sprint 0

❷ **Data integration**

• Confirm required & available data • Identify data sources

• Define data to be ingested

• Build and initate data ingestion process

❸ **IT architecture**

• Identify required elements for solution setup • Adapt existing architecture to the target solution

• Define data ingestion architecture • Ensure availability of necessary architecture models

❹ **Team**

• Define competences to gather

• Ensure teams’ availability (including Product Owner & Support)

• Define delivery model

❺ **Planification & project organisation**

• Set up steering and delivery model : injection planning, sprint 1 planning

• Evaluate and prioritize features (business value & complexity)

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 16

Group constitution

5’ 

**Divide the class into 7 balanced groups**

The group should include different school 

backgrounds

The group should include a gender mix

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 17

Let’s create the groups!

**7 Groups of 5 students : please, send your groups to fatima-zahra.mjerreb@capgemini.com**

| **Group 1**  ▪ Xxx  ▪ Xxx  ▪ Xxx  ▪ Xxx | **Group 2**  ▪ Xxx  ▪ Xxx  ▪ Xxx  ▪ Xxx  ▪ Xxx | **Group 3**  ▪ Xxx  ▪ Xxx  ▪ Xxx  ▪ Xxx  ▪ Xxx | **Group 4**  ▪ Xxx  ▪ Xxx  ▪ Xxx  ▪ Xxx  ▪ Xxx |
| --- | --- | --- | --- |
| **Group 5**  ▪ Xxx  ▪ Xxx  ▪ Xxx  ▪ Xxx  ▪ Xxx | **Group 6**  ▪ Xxx  ▪ Xxx  ▪ Xxx  ▪ Xxx  ▪ Xxx | **Group 7**  ▪ Xxx  ▪ Xxx  ▪ Xxx  ▪ Xxx  ▪ Xxx |  |

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 18

Agenda

1. Who are we?

2. Course modalities

3. Analysis objectives & approach

**4. Case presentation**

5. Data collection

6. Html presentation & Selectors 7. Scraping with Selenium 8. Summary of the session

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 19

Focus of the case study : voice of customer

▪ Our client, a **French energy supplier**, is expecting an 

intensifying competition to retain customers in 2023

given the tense energy market, including but not

limited to soaring energy prices.

▪ To differentiate itself from its competitors, it wants

to **listen to customer feedback** to better understand

and meet customer needs, expectations and identify

the priority pain points to solve across their

customer journey

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 20

Focus of the case study : voice of customer

As a **Data Science consultant**, you will have to assess the customer relationship strategy of a **French energy supplier** and provide recommendations on how to improve each stage of the **customer journey** 

Use Natural Language Processing (NLP) 

approaches to listen to the Voice of the

customer (VoC) and identify pain points

across the entire customer journey

Analyse these results in light of the current 

tense energy market context, then propose

data-driven opportunities that could benefit

to both your client and its customers

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 21

Exercice 1

10’

What could be the main drivers for customers to switch to another energy providers ? 

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 22

What could be the main drivers for customers to switch to another energy providers ? Driver 1 Driver 2 Driver 3

WELCOME DAY\_Plénière Company Confidential © Capgemini Invent 2021. All rights reserved | 23

Homework (due January 27th)

In group, conduct a preliminary analysis of the energy market trends and the main French B2C energy offers The main objectives of this homework are to :

- Identify energy market trends that might lead to a higher churn rate (1 slide)

- Conduct a benchmark on B2C energy offers, to highlight the key characteristics a consumer might look for in an offer (4 slides max).

Do not aim to be completely exhaustive in your benchmark. The target analysis will focus on French B2C energy offers only, both on gas and electricity, of the top 5 energy suppliers in France. Your benchmark should be limited to a maximum of 5 differentiators and should highlight your methodology (what were your assumptions ? which characteristics did you choose for the comparison ? why ?) and your sources.

Please send the results of your benchmark in a PowerPoint File by **Friday 27th of Jan.** 

**evening** to *fatima-zahra.mjerreb@capgemini.com* and *guillaume.remont@capgemini.com*

If you have any questions, feel free to contact us by email.

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 24

Agenda

1. Who are we?

2. Course modalities

3. Analysis objectives & approach

4. Case presentation

**5. Data collection**

6. Html presentation & Selectors 7. Scraping with Selenium 8. Summary of the session

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 25

Data pipeline of a Nature Language Processing (NLP) project

***Subject of today’s session***

******

Data Collection

How to collect data automatically from the web?

Data Cleaning

How to clean and process textual data and clean the noise?

Word Embedding

How to encode text into meaningful

numerical vectors?

Topic Extraction

How to extract the most representative topics in the data

Sentiment Analysis

How to detect and extract the sentiments expressed in textual data?

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 26

Application Programming Interfaces (API)

APIs enable interactions with applications

Several websites provide APIs for the purpose of data sharing

Few lines of code are needed to use them

The main advantage of using APIs is that they require less programming than scraping and are in general well documented

APIs provides structured data output

The output of an API is in general a JSON file which can be easily turned into a database



Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 27

Data can be collected through several data channels

**Databases APIs Web scraping**

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 28

Collecting data from databases

**Databases**

****

Extracting data directly from databases

Writing queries Using database connectors

A user can retrieve data directly from certain type of databases by selecting and copying it through a user interface (*e.g. Excel files*)

Queries are written statements that describes and executes read and write operations on databases (*e.g. SQL queries*). In order to write such queries, the user needs to know how the database is structured

Some data platforms provide connectors that facilitates extraction from databases stored into the platform (*e.g. Power BI provides connectors Azure DBs and Amazon DBs*)

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 29

Social networks have different policies concerning APIs

▪ No direct scraping unless authorized (fill in authorization form – 2 weeks for FB answer) : click here to see the terms

▪ APIs exist for app developers

▪ Limited collection of the data (speed/volume is compared to what « human can reasonably produce »), it allows read & write operations on videos with a limited quota 

▪ No personal data

▪ No direct scraping

▪ API with limited number of call by 15 min window

▪ Prohibition of scraping software

▪ APIs are for app development

▪ The free API is for “non-automated” apps, user authorization needed, Python/Ruby versions exists, 5000 calls per hour : it is being depreciated in favour of the new “Business version” (Instagram Graph API) ▪ Sensitive to user content/media (owned by users) 

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 30

Web scraping enables information retrieval from the web

**Web Scraping**

The art of extracting information on a

specific topic from the Internet using

automating requests

**Why should we use web scraping?**

▪ Market Price Analysis (*e.g. real-time competitiveness*)

▪ Market Intelligence (*e.g. competitive benchmarks*)

▪ Sentiment Analysis (*e.g. social listening*)

**Which data are we looking for?**

▪ Textual data available on websites (*e.g. articles, reviews, prices*)

▪ Metadata (*e.g. number of connections*) ▪ Social Media data (*e.g. tweets*)

**What is the typical process?**

▪ Crawl the web looking for needed information

▪ Centralize collected data and make it structured

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 31

Main tools used in web scraping

BeautifulSoup

HTML parser in Python, which transforms HTML code into a tree object usable with Python to extract content like text, images’ URLs…

Scrapy

Python package combining a crawler AND a parser.

Selenium

Tool allowing to automate web browsers, mainly used in scraping to get data which is written in

JavaScript

through AJAX requests.

RVest

HTML parser like BeautifulSoup, but for R.

Great

documentation online.

Selenium

In this course we will focus on **Selenium** mainly

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 32

Agenda

5. Data collection

1. Who are we?

2. Course modalities

3. Analysis objectives & approach 4. Case presentation

**6. Html presentation & Selectors** 7. Scraping with Selenium 8. Summary of the session

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 33

Scraping enables information retrieval from HTML

▪ Scraping packages enable the user to extract information 

from HTML pages and to structure it into databases

▪ Some browsers (Chrome, Firefox…) provide access to an

interactive “Inspect mode”, which enables the user to 

navigate within the output code of a web page :

CTRL + SHIFT + I

OPTION + CMD + I

▪ The user chooses the information he wants to extract from a

web page by exploring its code and by adding to his scraping

code the tags he wants to extract

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 34



Structure

▪ A web page is structured by two main elements:

▪ HTML : backbone of the web page, it contains text arranged

into blocks, which have attributes

▪ CSS : describes the style of the webpage

How is a web page structured ?

<!DOCTYPE html>

<html>

<head>

<meta charset="utf-8" />

<title>Titre</title>

</head> 

<body>

</body>

</html>

Sequence

▪ It is a sequence of HTML tags which can

be seen as a tree

<tag attribute="valeur"> … </tag> *E.g. Text tags*

Tags

▪ Each tag has a specific format ▪ There are several attributes per tag: class, href, etc.

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 35

XPath notation to navigate HTML 

**What is XPath ?**

▪ Xpath is a string which selects nodes in an HTML tree 

▪ It can also be seen as the linear representation of the requested

element

xpath=’/html/body/h1’

▪ XPath allows you to select:

− The content of a markup

− The content of its attributes (hypertext links for example)

| **Requested Element**  Second division of the body | **Corresponding Xpath**  '/html/body/div[2]' (index start at 1) |
| --- | --- |
| All tables | '//table' #all tables |
| All tables descendants of the 2nd division All paragraphs directly bellow the body | '/html/body/div[2]//table'  '//p' |
| Conditional division | '//div[@id="uid"]' |
| Wildcard | '/html/body/\*' |
| All elements with a condition on the class | '//\*[contains(@class, "class-1")]' |
| Selection of the attribute href cond.  paragraph | '//p[@id="p2"]/a/@href' |

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 36

Break

15’

**Feel free to help yourself ! See you at 16h15 !**

****

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 37

Setup

Downloads and installations

▪ Download Python by following the instructions for your operating system ▪ It is also possible to download anaconda and install Selenium in the conda environment 

To download an IDE (Integrated Development Environment), you have several options, among which : ▪ PyCharm 

▪ VS code

▪ Download a driver for your favorite web browser: Chrome, Firefox, Safari

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 38

Web scraping

▪ Introductory notions of HTML structured websites

▪ Getting started with selenium (installations, libraries)

▪ Basic manipulations on a reviews websites (Trustpilot and Avis Vérifiés) ▪ Extracting a first comment and its metadata 

▪ Extracting data from the first then the second page of the website ▪ Building a database of all the comments, their ratings and dates (in 2 different ways) ▪ Exploiting the database by detecting the negative comments

Bonus exercise 

▪ Parsing using BeautifulSoup

▪ Combining BeautifulSoup with Selenium

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 39



Hands-on 1

40’

**Use the notebook 1 to discover Selenium and get some information from the web** If you have any 

question about Python set-up, feel free to contact us about that !

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 40

Restitution

5’

**Could you extract data ? What was challenging for you ?**

****

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 41

Agenda

5. Data collection

1. Who are we?

2. Course modalities

3. Analysis objectives & approach 4. Case presentation

6. Html presentation & Selectors **7. Scraping with Selenium** 8. Summary of the session

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 42

Hands-on 2

**Try to automate the scraping of Trustiplot **

****30’

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 43

Restitution

5’

**Could you extract data ? What was challenging for you ?**

****

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 44

Agenda

5. Data collection

1. Who are we?

2. Course modalities

3. Analysis objectives & approach 4. Case presentation

6. Html presentation & Selectors 7. Scraping with Selenium **8. Summary of the session**

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 45

Summary of the session – To remember



**Approach & organization of a data science consulting project Scraping as a tool to retrieve information from the web**

**Typical approach of this type of project**

• Data science workstream: Data scraping, cleansing & feature engineering, running of the different analyses, restitutions (visualizations etc.)

• Business workstream: Diagnosis of the current situation & transformation stakes (as-is/to-be analysis) with a quantification of the impacts

**Organization & governance :**

• Several dedicated meetings all along the project (e.g. weekly status updates, steering committees) to track progress and escalate potential issues

**Scraping as a data collection tool:**

• Data can be collected through several channels: databases, APIs, web scraping

**Zoom on web scraping:**

• Scraping tools (Python: Scrapy, BeautifulSoup, Selenium ; R: Rvest)

• Key steps and tools used to perform them :

• Parsing : Xpath and CSS locators

• Crawling : Spider classes

• Storing : JSON file

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 46

Work for next week

**Instructions**

To practice what we learnt today, for next week, you’ll have to :

▪ Prepare a competition benchmark of the French energy B2C offers to highlight main offer differentiators for customers ▪ Create a script able to scrape trustpilot website for the following energy suppliers, making sure that you have the comments of the first 10 pages : 

▪ https://fr.trustpilot.com/review/eni.fr

▪ https://fr.trustpilot.com/review/totalenergies.fr ▪ Make sure that for each comment you have the date, the body, the rate and (bonus) the answer of the supplier.

We expect you to send your code and scraped file by Friday 27th evening to ines.el-kasmi@capgemini.com, fatima

zahra.mjerreb@capgemini.com,

guillaume.remont@capgemini.com and

thibault.venet@capgemini.com

If you have any questions, feel free to contact us by email.

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 47

Course evaluation

**Did you like that first course ? It’s time to share your feedbacks !**

****

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 48

**T h a n k y o u f o r y o u r a t t e n t i o n S e e y o u n e x t o n c a m p u s !**

**G O O D B Y E !**

January 23rd , 2023

Master X-HEC – 23.01.2023 © Capgemini 2023. All rights reserved | 49