# High frequency data analysis

| | |
|---|---|
| Course title – Intitulé du cours | High frequency data analysis |
| Level / Semester – Niveau /semestre | M2 / second semester |
| School – Composante | Ecole d'Economie de Toulouse |
| Teacher – Enseignant responsable | Max Halford |
| Other teacher(s) – Autre(s) enseignant(s) | |
| Other teacher(s) – Autre(s) enseignant(s) | |
| Other teacher(s) – Autre(s) enseignant(s) | |
| Other teacher(s) – Autre(s) enseignant(s) | |
| Other teacher(s) – Autre(s) enseignant(s) | |
| Lecture Hours – Volume Horaire CM | 12 |
| TA Hours – Volume horaire TD | |
| TP Hours – Volume horaire TP | |
| Course Language – Langue du cours | English / Anglais |
| TA and/or TP Language – Langue des TD et/ou TP | English / Anglais |

**Teaching staff contacts – Coordonnées de l'équipe pédagogique :**

Email : maxhalford25@gmail.com

Office number: /

Office Hours: /

Preferred means of interaction: Email, or after class

**Course's Objectives - Objectifs du cours :**

There are cases where data analysis can be difficult because of high volume of data to process. Whether it be the IoT industry analysing sensors, or in finance with stock-market data, or in marketing when looking at web activity. At the very least, high frequency data can impose technical difficulties that get in the way of data analysis. This can be the case for reporting (looking at the past) as well as inference (forecasting ahead).

The goal of this case if to get you comfortable with the techniques that are used nowadays for analysing high volumes of data. We will distinguish batch methods will process large chunks of data in one go, from online methods which process data in a streaming fashion. In both cases, we will start by covering standard descriptive statistics, before moving on to the machine learning side of things.

During the first part of the lectures, we will be covering batch techniques for processing large amounts of data. We will talk about columnar databases, OLAP vs. OLTP, mini-batch machine

learning, hardware concepts, typical setups in data science teams, how to work with data that doesn't fit on disk, and more. We will also be taking some historical perspective, and examine why recent technologies such as MapReduce and Spark are (almost) already considered old. We will be covering some libraries in this space, such as DuckDB, Polars, and Vaex.

In the second part, we will be looking at streaming techniques for analysing data. We will see how reporting can be done online, with technologies such as Kafka, FlinkML, and Materialize. We will see what algorithms are used, how to approach this new paradigm, and what benefits this can yield. We will also be covering online machine learning, in particular with the River library in Python.

**Course outline :**

- 1 hour to cover general concepts
- 4-5 hours to cover modern batch processing techniques
- 4-5 hours to cover online data analysis and online machine learning
- 1 hour to review, and discuss the projects

Each lecture will consist in two parts. During the first part, students will interact with pre-made Python exercices. The second part will be a lecture covering the topics seen in each exercice. In other words, students will first learn by doing.

**Prerequisites - Pré requis**

We will be covering some advanced concepts in machine learning and programming. Therefore it is expected that students have some experience with Python programming and are somewhat at each with the basics of machine learning. Some knowledge of SQL would help too.

**Grading system - Modalités d'évaluation :**

Students will pair in groups of 2 and will have to work on a project of their choice, covering one or more of the topics seen during the class. A small accompanying report is expected.

**Bibliography/references - Bibliographie/références :**

- Feel free to Google the keywords in the paragraph above
- What your data team is using: the analytics stack gives a good idea of the technologies that are in use today
- Take a look at mlcourse.ai if you want to practice your ML skills
- Take a look at SQLBolt and Select Star SQL If you want to refresh your SQL knowledge