CSC4524: Streaming algorithms

Introduction

Professor

- **Mohamed Amine ZGHAL**

- **Data Scientist at** Veepee

- **E-mail : mazghal@vente-privee.com**

  Subject : [CSC4524] Group i, Project k, …

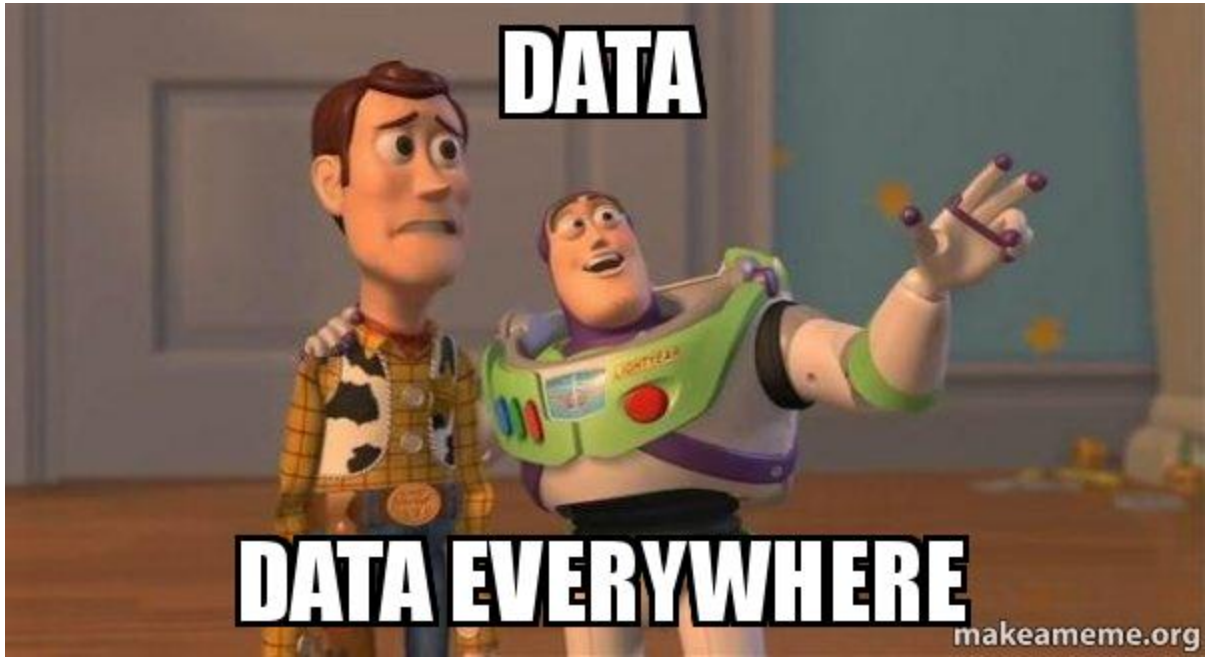
Content

- **4 mini-projects**
  - Work in pairs
  - Format: Jupyter notebook
  - Duration: 1 week per mini-project

- **Paper lecture**
  - Work in pairs
  - Format: slides
  - Duration: 1 week

Definition

- **A data flow is an infinite sequence of elements generated continuously at a fast pace.**



Fields of application

- Stock indices
- List malicious sites
- Website logs
- network traffic logs
- Sensor data

Principles

- **Extract information that characterize the flow**
  - Event detection
  - Cardinality representation
  - Live metrics
  - Minimal size flow representation

    ...
- **Live computations**
- **Low time and memory complexity**

Flow simulation

- **Machine with low resources (CPU, memory, storage)**
- **Too much data to fit in memory at once**
- **Low complexity algorithms**
  - One pass over the data
  - Low memory impact

Configuration

Mini-project 1

- **Be able to know if a given Wikipedia page was visited the 01-08-2016**
  - Input: domain code + page title
  - Output: Yes | No
  - Work due to 17-11- 2019 23:59