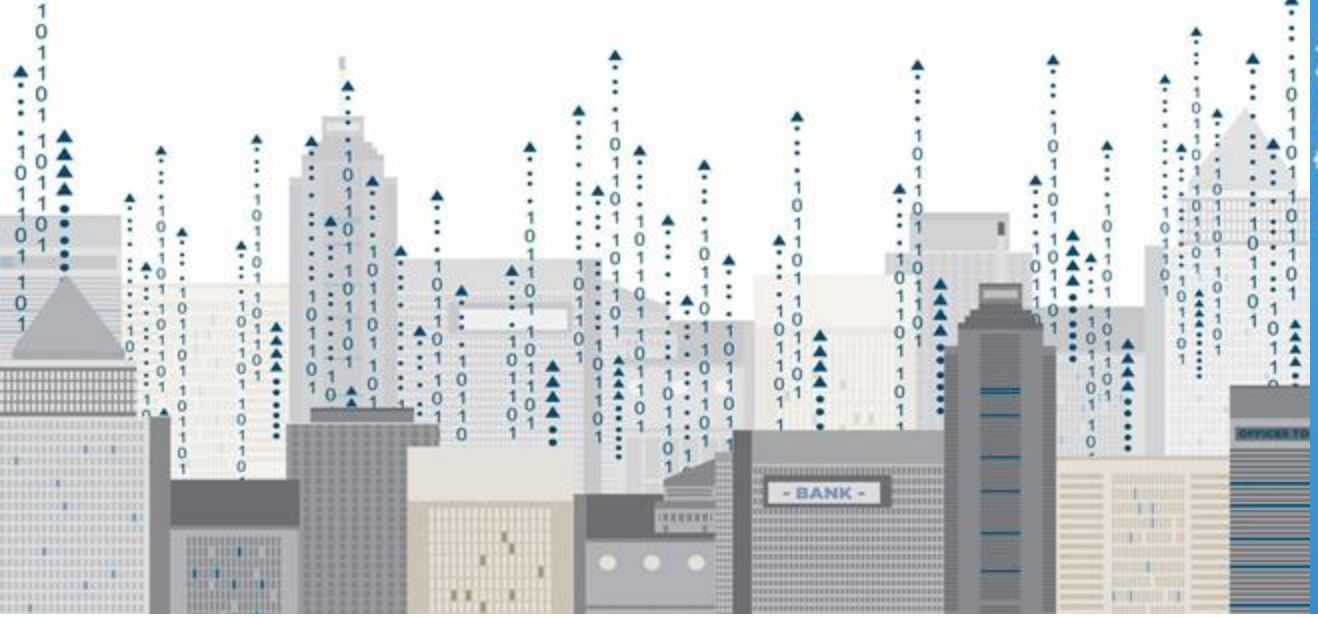


DATA AND PROJECTS FOR YOUR EXAM

...a brief guide to the datasets and the practical objectives of this course

ANGELO FURNO, TRISTAN LEMOALLE
LICIT-ECO7 (ENTPE – Université Gustave Eiffel)

Lecture of the Course:
Data Science: Principles and Applications

Exam: Project & Presentation(s)

1. **Self-organize into groups of ~3 persons**
2. **Each group should work independently on one project**
3. **Define** the project you want to work on
   - To get ideas of possible projects you can :
     - check the projects subject given last year (and that will be presented today)
     - check the TPs
     - develop your own ideas based on the data provided to you (the data will presented today as well)
4. **Analyze the data** we provide with respect to the task goals
   - The main goal of this course is to develop a data science project **with ML algorithms**
   - All efforts on attempts to go further than what is seen on the course / practical session will be very appreciated
     - code/speed optimization analysis, distributed calculation, machine learning approaches not seen in the lessons
     - IT JUST MAKES ME HAPPY!
5. Prepare your **project presentation** for the exam

Data Description

- **This data will be soon available online**

- **Multiple Data sources avalaible:**

- **Floating Car Data (FCD)**
  - One week march 2020
  - Three weeks november 2019
  - On A10 Paris highway for sept / oct 2024

- **Shared Bikes Data**
  - Data from Lyon's Velo'V
  - **Available on-demand for Lyon and Toulouse in 2019 and 2020**

- **Public Transport Data**
  - Ticketing data from TCL for Lyon's PT

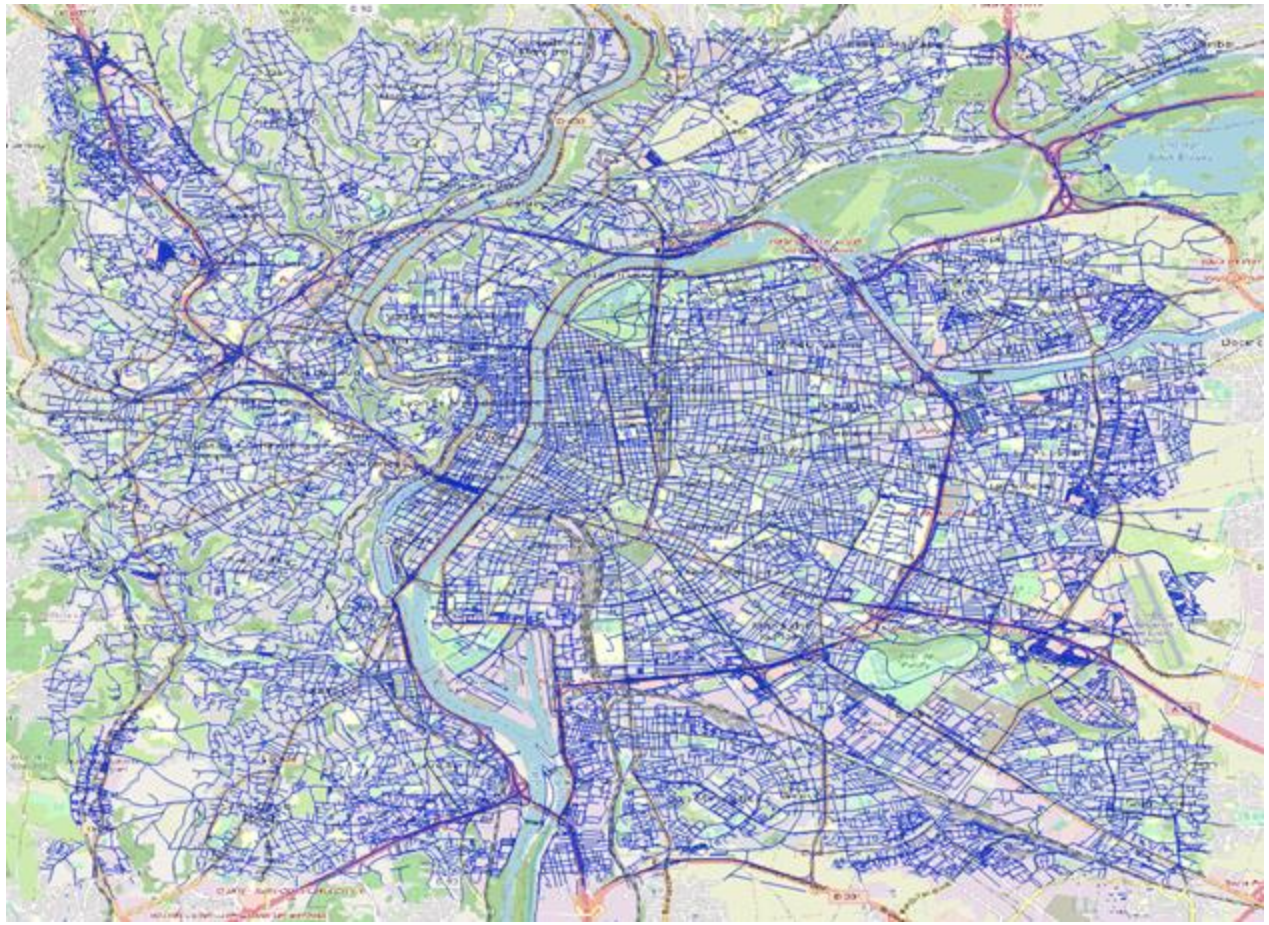
In the context of your project, do not hesitate to ask us for additional data if needed !

Floating Car Data in real world

- **Issue**
  - Getting information on traffic state
    - vehicle flow, concentration, presence of accidents, …
- **Vehicles act as probes**
- **Different types of vehicles**
  - Taxis
  - Public transport
  - Delivery service vehicles
- **Rely on satellite-based positioning (GPS)**
- **Deployed in Vienna, Lyon, Düsseldorf, Berlin, Beijing quality depends on**
  - Total fleet mileage and area covered
  - Sample size with regard to overall traffic (penetration rate)
  - Parameters collected
  - Sampling rate of GPS positions

Our FCD data

- **FCD** – GPS points of floating vehicles
  - **deviceID**: id of the vehicle
  - **latitude**: latitude of the gps point
  - **longitude**: longitude of the gps point
  - **speed**: estimated speed in km/h
  - **heading**: orientation angle of the vehicle
  - **timestamp**
  - **linkId** : id of the link (route segment)



Shared Bikes Data

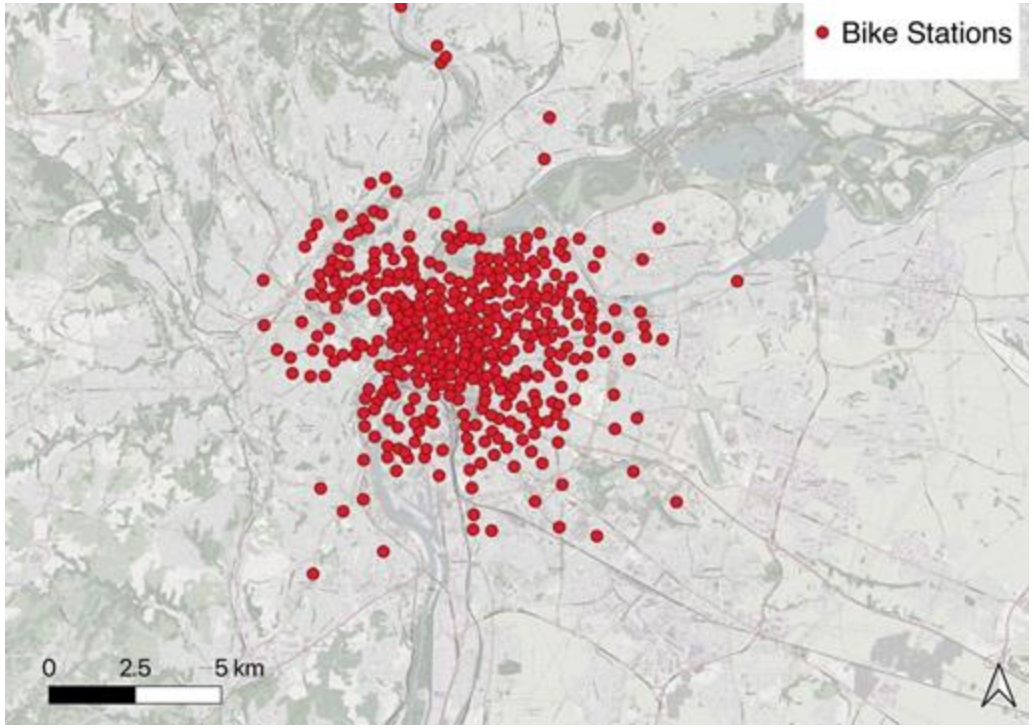
- **Origin**
  Data from Lyon's Velo'V

- **How it is recorded**
  Trips are represented with the times and stations of the bike retractions/returns

- **Containing information**
  - Retraction and return timestamp
  - Retraction and return station

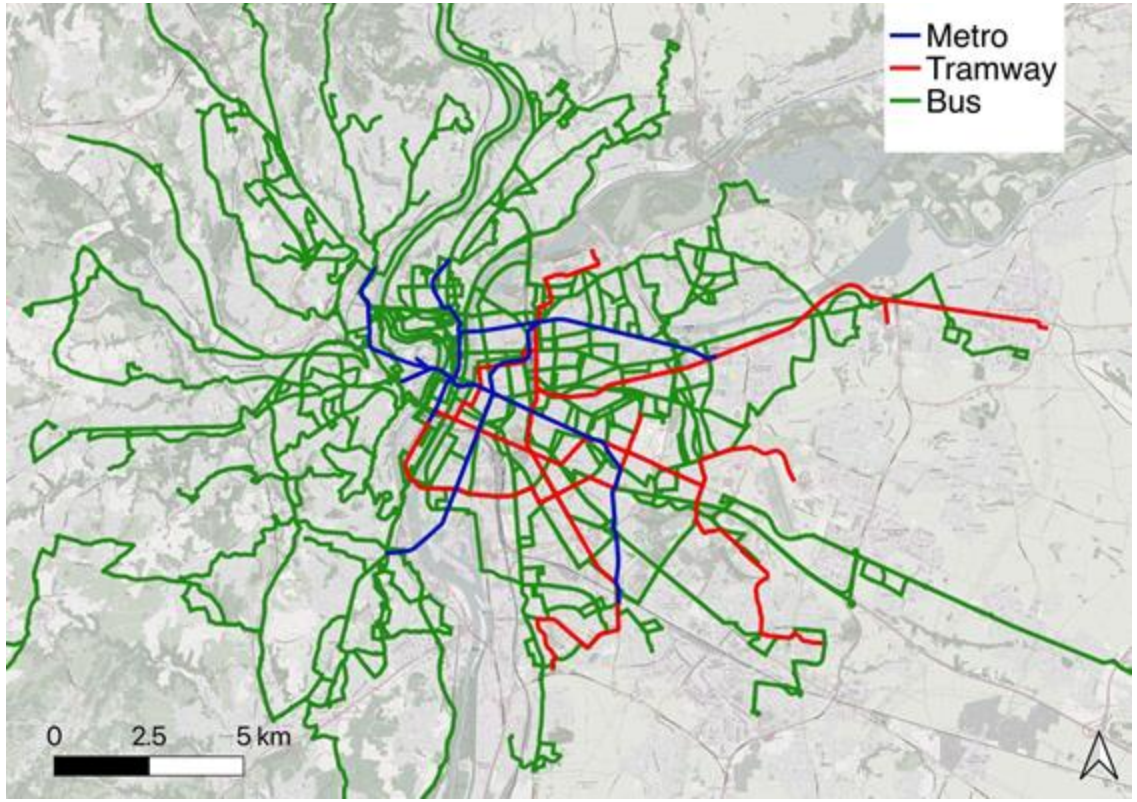
Shared Bikes Data

- **Data structure**
  - **id_sortie**: The ID of the retraction station
  - **Borne sortie**: The name of the retraction station
  - **date_sortie**: Time of retraction
  - **id_retour**: The ID of the return station
  - **Borne retour**: The name of the retraction station
  - **date_retour**: Time of return



Public Transport Data

- **Origin**
  Ticketing data from TCL for Lyon's PT
- **Modes**
  Metro, Funicular, Tram, Bus
- **How it is recorded**
  Validations while boarding the vehicles
- **Containing information**
  Validation timestamp, station, PT mode



Our Public Transport Data

- **Data processing**
  - Aggregation
  - Pre-processing

- **Data structure**

  - **station**: the station code for each aggregate validation count
  - **C_x** & **C_y**: the coordinates of the station
  - **MODE**: the PT mode for each aggregate validation count
  - **VAL_DATE**: the validation date (by hour)
  - **count**: the hourly number of validations recorded for the station/mode of interest

Am I obliged to work on transport-related data?

- **NO! …** You are not obliged to use the data presented here
- You can use any dataset you want on the condition that you are capable of satisfying your duties (see next slide)
  - The only difference is we will not provide support in the data collection stage
    - **Means "You are on your own to find the data"**
- My advice is to **focus on time series and spatial** data or any relevant data with a mix of numerical features
  - Avoid data with a lot of text (as we do not address this in the course)
- Other ideas of dataset (see next slide)

Other data (Some Ideas)

- **Environment & Climate**
  - Météo-France – open climate data (temperatures, rainfall, wind, etc.)
  - European Environment Agency – Air quality
- **Energy & Utilities**
  - RTE ÉCO2mix – Electricity consumption & production
  - https://www.rte-france.com/en/eco2mix/download-indicators
  - ENTSO-E Transparency Platform (European electricity grid data)
- **Finance & Economy**
  - OECD Data (economic indicators: GDP, unemployment, inflation)
- **Health & Demographics**
  - Our World in Data – COVID-19 dataset (cases, vaccinations, testing)
  - INSEE – Demographic data (population, density, age distribution)
- **Sports & Culture**
  - Football-Data.co.uk (match stats, odds, results for European leagues)
  - Kaggle – Spotify dataset (music listening statistics & metadata)
  - https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset , https://www.kaggle.com/datasets/atharvasoundankar/global-music-streaming-trends-and-listener-insights)
- **Geospatial & Open Data**
  - OpenStreetMap via Geofabrik (POIs, road networks by region)
  - Copernicus – Land Monitoring Service (satellite indicators, NDVI, land cover/use)

Your duties for the exam

- **Short report (10 pages max) on your project**
  - Context, Problem and Goals (1 pages)
  - Assumptions and difficulties if any (1 page)
  - Pre-processing steps and feature engineering (2 pages)
  - Data analysis and ML methods (2-3 pages)
  - Results and interpretation (3-4 pages)
    - Work on finding good visual representations (plots, maps, tables, etc.) for results/conclusions (qgis, matplotlib, etc.) to support your statements and give useful insights
  - The goal of the report is mainly to help you with the presentation

- **Your code on a Git repository**
  - **Including separate and commented files for functions related to:**
    - Pre-processing & feature engineering code
    - Preliminary data analysis & statistics
    - Machine learning task
    - …
  - **A jupyter notebook where the functions above are used and the results analyzed with plots, tables, markdown descriptions**

Deadlines

Date	Course Title	Course Hours	Time Slot (include 15 min. pause)	Teachers
Sept. 9	DS-PA 1: Intro to Data Science, ML Concepts, Anaconda Installation + environment prep.	3.5	8:45 - 12:30	Angelo FURNO
Sept. 16	DS-PA 2: Guidelines to Programming with Python, Ethics in Data and AI + Python recap for data science,	3.5	8:45 - 12:30	Angelo FURNO + Rim SLAMA SALMI
Sept.22	DS-PA 3: Networks, Time Series and Spatial Data Analysis + Project Data Presentation	3.5	8:45 - 12:30	Angelo FURNO + Tristan LEMOALLE
Sept. 23	DS-PA 4: Regression and Forecasting	3.5	8:45 - 12:30	Manon SEPPECHER + Tristan LEMOALLE
Sept. 30	DS-PA 5: Clustering and PCA	3.5	8:45 - 12:30	Angelo FURNO + Tristan LEMOALLE
Oct. 7	DS-PA 6: Classification (Arbre, SVM)	3.5	8:45 - 12:30	Angelo FURNO + Tristan LEMOALLE
Oct. 21	DS-PA 7: Project Work Session 1	3	9:15 - 12:30	Angelo FURNO + Bahman MADADI
Nov. 4	DS-PA 8: Project Work Session 2	3	9:15 - 12:30	Angelo FURNO + Bahman MADADI
Nov. 10	DS-PA 9: Evaluation Test, Project Presentation	3	9:15 - 12:30	Angelo FURNO + Bahman MADADI

- In-between two follow up meetings, but you can interact with us in each of the classes (TP)
- **Due date of the report is on Saturday 8th November (23:59)**
  - teamwork is the key: don't let one person do the whole work
  - split the work in sub-tasks if possible, test multiple solutions in parallel, ...

- **Evaluation Monday, 10th November**
  - **15-minutes group presentation** on the data project
    - **Introduce the datasets you used**, the **methodologies** and the **results.**
  - **10-minutes questions** on the project

Project Main Ideas

- **The objective of the project is to perform data analysis/machine learning tasks on large-scale datasets using big data tools**

- **We will achieve this objective gradually, by also showing the limitations of traditional Python data science tools**
  - You will work with Pandas, NetworkX, GeoPandas, Numpy and Scikit-learn
    - For your data science machine learning tasks
    - By selecting smaller samples of your data
      - filtering/aggregating your data to reduce your problem to small parts of the data/network/periods of time
    - You will have to measure the performance (as shown during the TP) of your code
      - E.g, use the time_usage function (or something equivalent)

Project 1: Identification of spatial/temporal speed patterns

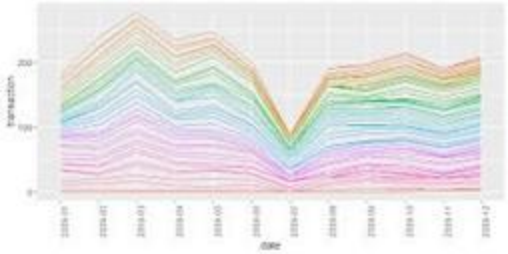
- **Identify the typical patterns based on the speed variable**

1. Define and use a proper spatial/temporal aggregation based on speed

- **Temporal binning**: Choose a proper granularity (e.g., 1 hour, 30 minutes, 30 minutes)
- **Spatial aggregation:** Define a spatial aggregation for the time series (e.g., per link, per area, per IRIS sectors)
- Apply filtering to interpolate/smooth the time series
- Your clustering object is the geospatial data you have built, e.g., 24 hours for a typical day, 24*7 for a typical week, etc.

1. Identify spatial clusters based on speed

- Apply spatial clustering with the time series data you have built with a proper method
- Identify spatial speed clusters and explain them



1. Identify temporal clusters based on speed

- Select one (or more) spatial cluster you have detected, and do a temporal clustering on that zone/link/IRIS
- Identify temporal speed clusters and explain them
- Compare the typical/atypical hourly speed profiles before/during COVID-19 lockdown

Project 2: Route planner

- **Create a framework able to infer the shortest path and estimate travel time for a given trip (origin and destination)**
  1. **Build road transportation network**
     - Build set of nodes and edges from the link data and create a directed graph using Networkx
     - Verify the connectivity of the constructed graph
  2. **Estimate the travel time of the road network edges**
     - Choose the set of links where speed can be estimated accurately from the FCD
     - Choose an appropriate travel time approximation for the links with missing speed estimation
     - Add weight (travel time) to the road network
     - The travel estimation can be considered dynamically (depends on the day, hour of the day …)
  3. **Finalize the approach to propose the shortest path and estimate the travel time given two locations**
     - Visualize the result on a map (contextily library can be used to plot OSM background)
     - Compare the result of your approach with existing route planner such as Google Maps

Project 3: Travel time prediction through regression

- **Predict the average OD travel time using regression methods**

1. **Prepare the data base**
   - Using trip extractions, aggregate the data to build your response depending on the predictive variables you consider relevant.
   - Analyze and provide statistical insights into the resulting dataset.
2. **Single OD model**
   - Start simple, by focusing on the OD couple of your choice.
   - Try out different regression models, that you will conflexify by increasing the number of predictive variables. Conclude on the most relevant(s) model(s).
3. **Model complexification**
   - Try to generalize the model to a multiple origin-destination model.
   - And/or consider adding exogenous and open data such as weather conditions.