DATA ENGINEERING PROJECT

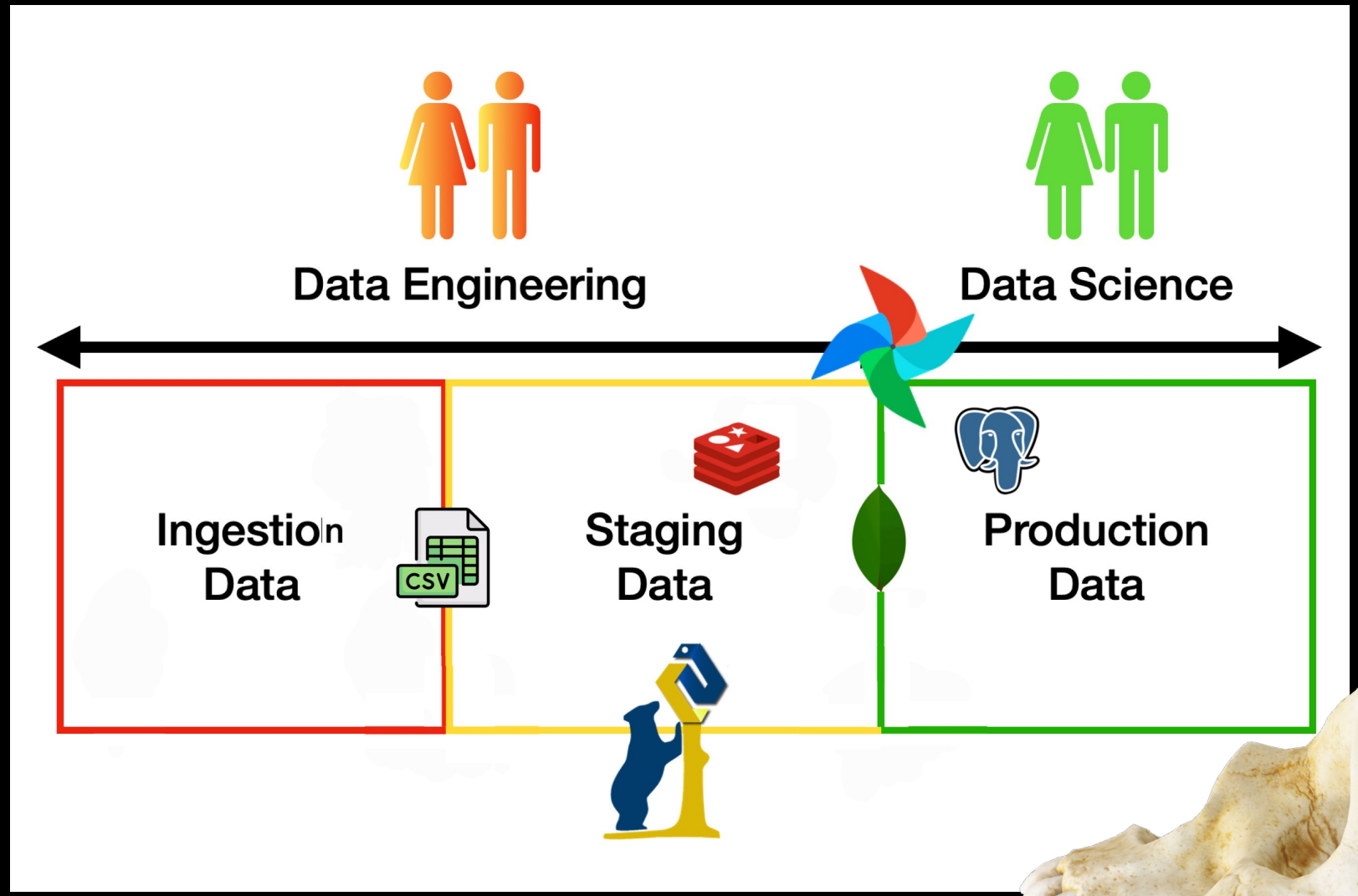
CHAILLAN Ewen

ROUX Matthieu

INTRODUCTION

- We thought it would be intersting to see where most people die in France.
- To spice things up, we thought it could be interesting to compare this data with location of power plants
- Eventually we would like to visualize our data on a death heatmap of france, with locations of power plants

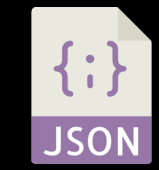
OUR ARCHITECTURE


Data Engineering
Data Science
Ingestion Data
CSV
Staging Data
Production Data

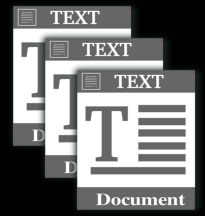
DATA INGESTION

- Data sources: from https://www.data.gouv.fr/fr/datasets/

  - Deaths and location of french citizens
    - Stored as multiple .txt files
    - All files are downloaded every time
    - We put a number limit for testing



Get list get resources



Load all resources

  - Thermal power plants
    - Stored as .csv
  - Nuclear power plants
    - Stored as .csv

Get URL of CSV resource

Load CSV

  - Geolocalisation of french cities
    - Stored as csv

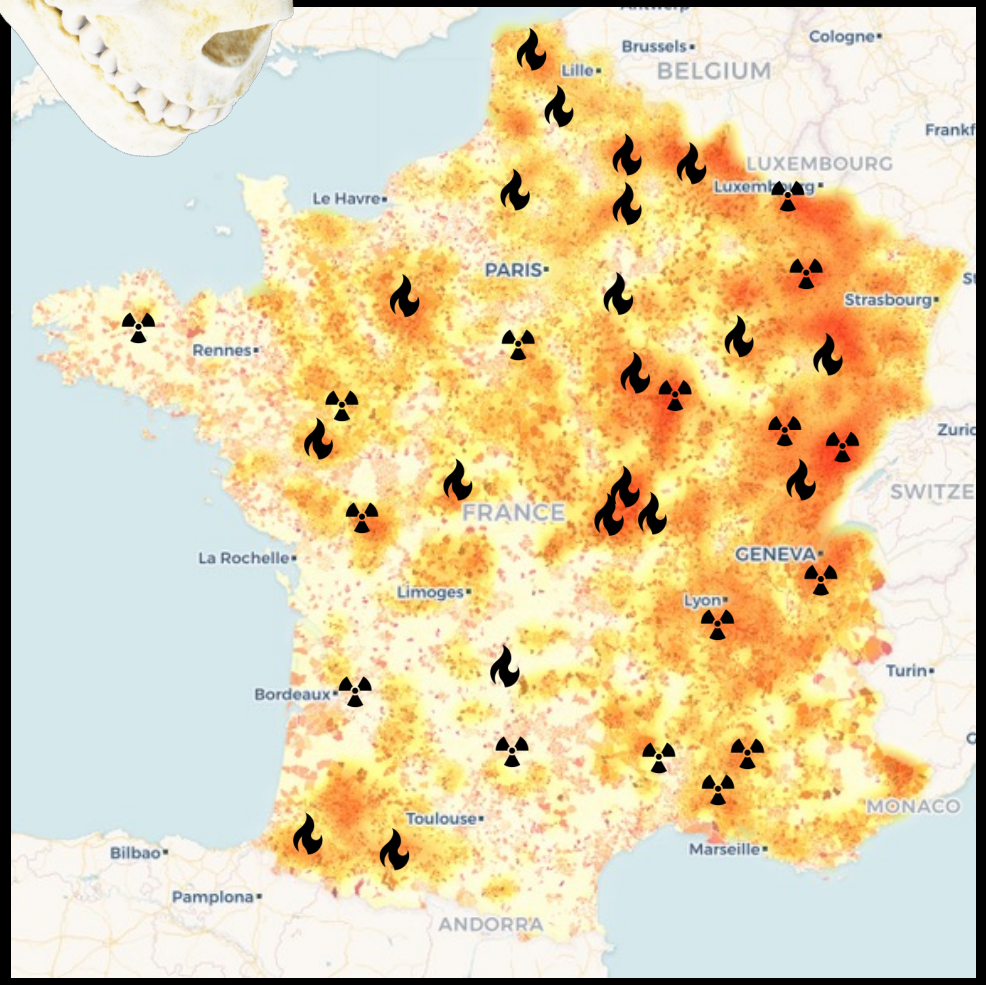
Load CSV directly

DATA CLEANSING

- Nuclear and Thermal power plants:
  - The data was already quite steady and usable. We had to drop some columns and rename others but it was not a hard task.
  - We used pandas to modify the csv. For now, the curated data is save as csv but we wish to save it in a mongoDB database.

- Deaths and location of french citizens, Geolocalisation of french cities:
  - Text files that have not been imported are imported and added to the list of imported files (cached in redis)
  - The imported deaths are then mapped to geo coordinates based on the death's INSEE code
  - Dates are parsed, names are hashed into Ids to increase privacy
  - Result is stored to CSV, will store in database next (Mongo)

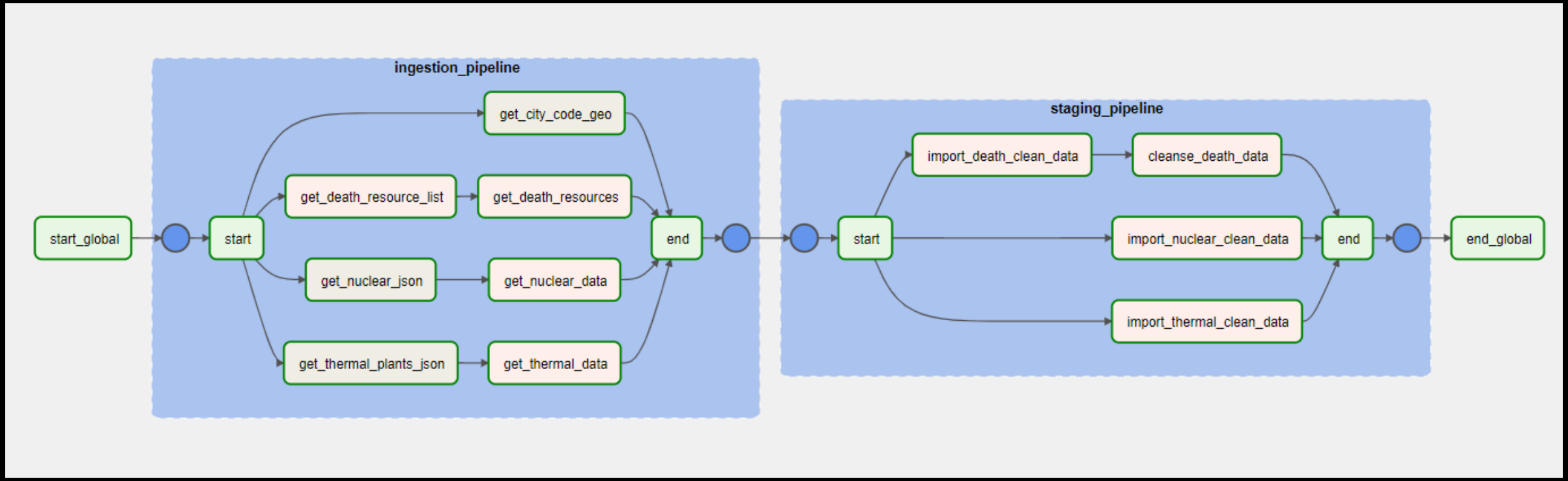
PRODUCTION (COMING SOON)



- Our fist step would be to have a table of deaths in a 50km radius of the power plants
- The second step would be to display those on a map
- Then we would display a heatmap of deaths on this map

The data viz could be pushed even further:
- Use a slider that filters the year and lets us see the data based on that year. Seeing factories opening and closing, and death heatmaps moving as we move the slider

PROGRESS SO FAR



THANK YOU!

Questions?