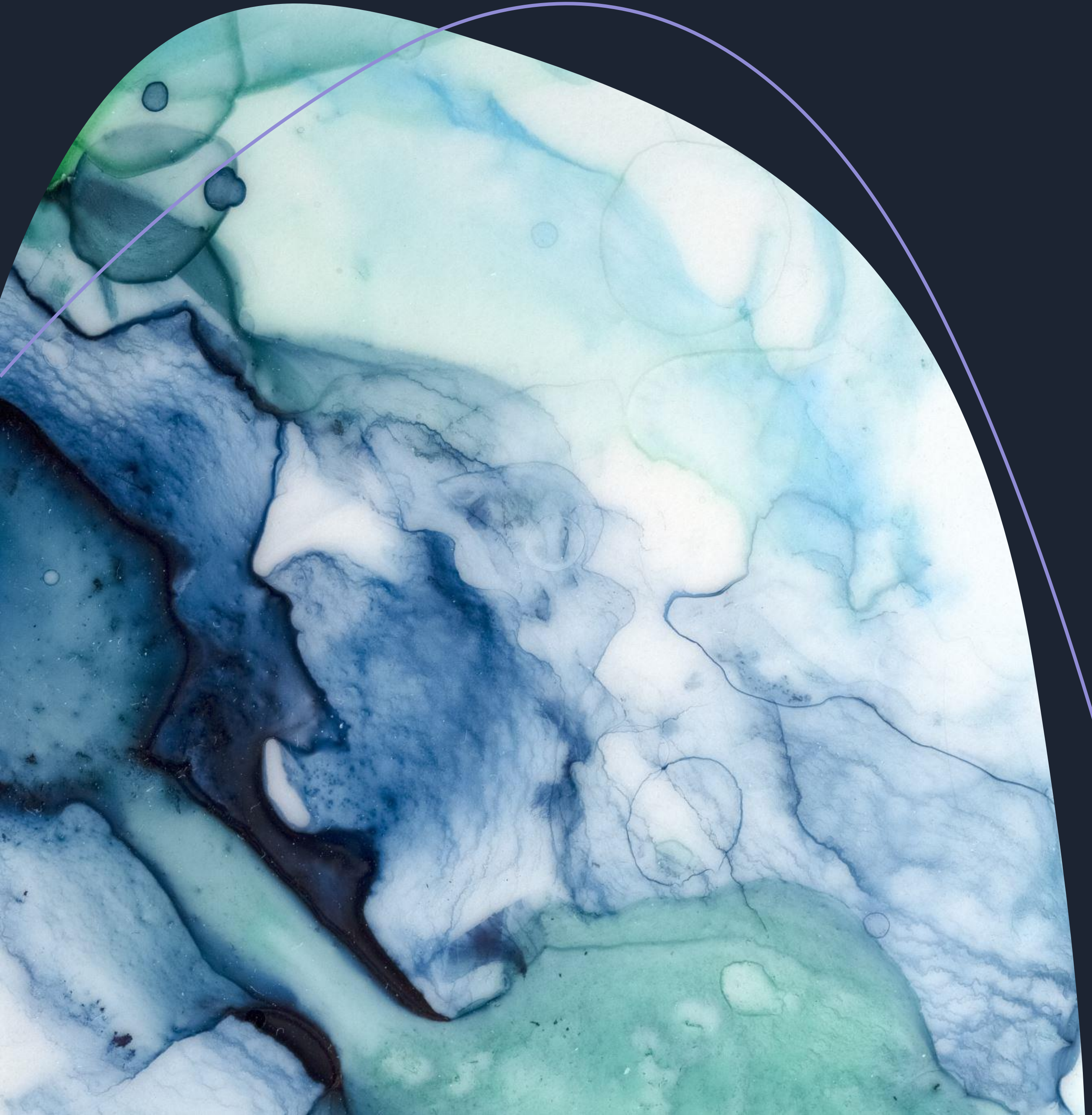


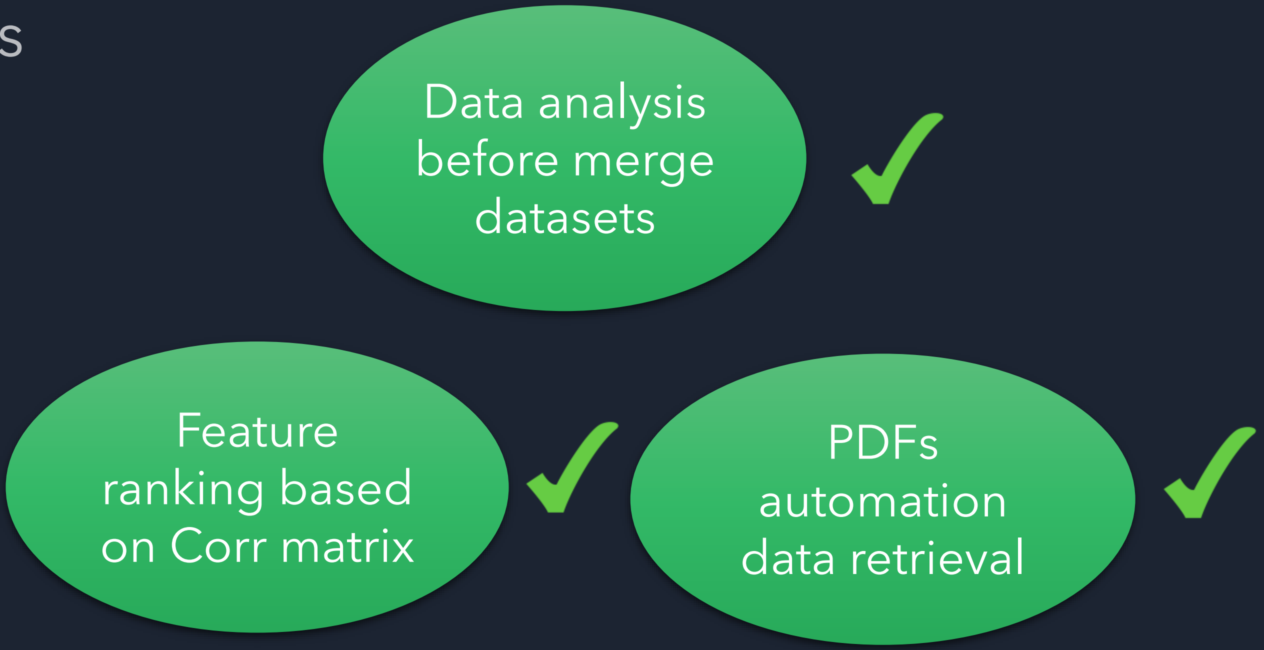
Schneider Electric Hackathon

Data Science Challenge

Luis Carranza

EDA – How complex this problem is?

- 6 data sources from 3 different types
- +65k Rows
- +4.5k Cities
- +24K Reporter names! (noise)
- Important variables:
  - EPRTRAnnexIMainActivityLabel
  - eprtrSectorName
  - countryName



Data analysis before merge datasets ✓

Feature ranking based on Corr matrix ✓

PDFs automation data retrieval ✓

ML Experiments – LightGBM

train1.csv + train2.csv	train1.csv + train2.csv + json1 + json2 + json3
f1_macro: 0.648 (0.007)	f1_macro: 0.470 (0.004)

CSV + JSON decision chosen

↓ Less precise

↑ More robust

Stack

- Anaconda for the Jupiter Notebook and DS libraries
- LightGBM for fast ML experimentation
- Pymupdf for PDF text extraction