Introduction & Data Preparation

Dataset window: 2022-01-01 00:00 → 2024-06-12 23:00 (UTC)

The data are consolidated into a single **hourly** table. All timestamps use **UTC**. An hourly time grid defines the index.

Each source is reindexed to this grid: day-ahead price, load, generation by technology, cross-border flows, commodity drivers (natural gas, coal, CO_2), and installed wind/solar capacity.

Lower-frequency or irregular series are upsampled with (first) forward fill and (then) back fill. Then, the sources are combined with outer joins.

Column names are cleaned. Units are standardized to **EUR/MWh** (power and gas), **EUR/ton** (coal and CO_2), and **MWh per hour** (load, generation, flows).

For trade flows, a positive `trade_balance` means **net imports**. Exports are negative.

The final dataset is `clean_data.csv`. It is used in Tasks 1–4. The code lives in `original_data/preprocess.ipynb`.

Consistency note. In real systems the hourly energy balance $$\text{Generation}+\text{Imports}-\text{Exports}-\text{Load}\approx 0$$ should be near zero. In this dataset it does not close. The likely cause is that the load series is a **day-ahead forecast**, plus small reporting noise. This was taken into consideration for task 4 specifically.