# Anomaly Detection in the U.S. Electricity Grid data

ORIE 5741

Jiaqi Ding (jd2269) Lewis Tian (zt89) Prabhat Koutha (pk454)

## Overview

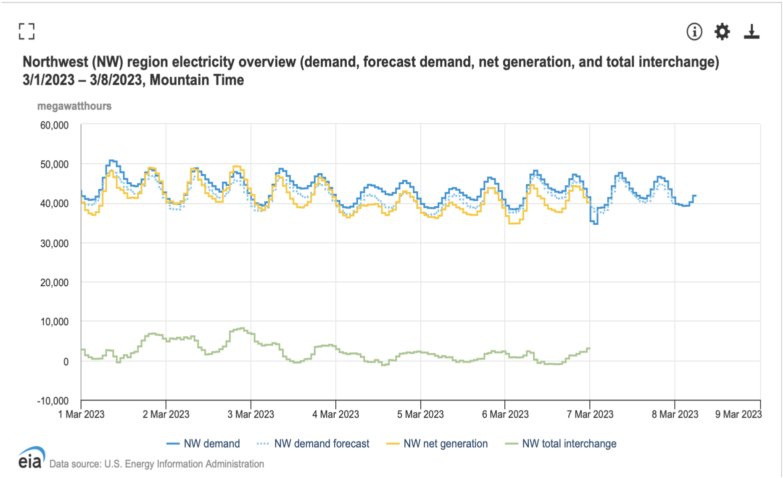
This project aims to identify anomalies in the U.S. electricity grid data to help mitigate the impact of data quality issues for electricity traders, among others. The project uses a dataset from EIA that contains hourly electric grid data for the NW region with over 66000 hourly observations. We will implement various semi-supervised/supervised algorithms and compare their performance. Additionally, we will perform time series analysis to replace the anomalous values.

## Background

Electricity trading involves buying and selling electricity on various markets, and it is facilitated by financial instruments such as futures contracts, options, and swaps. These financial instruments enable traders to manage risks and make profitable trades in the dynamic electricity markets. Electricity is traded by a variety of market participants, including energy producers, retailers, traders, brokers, industrial consumers, government entities, banks, hedge funds, and proprietary trading firms. Electricity grid data helps traders make informed decisions on when to buy and sell electricity by providing insights into market trends, price forecasting, risk management, portfolio optimization, and regulatory compliance. Anomalies in EIA electric grid data can have significant consequences for traders, and it is crucial that traders stay informed about potential data quality issues and take steps to mitigate their impact on their trading strategies. Hence, we aim to solve this problem by identifying anomalies and replacing the anomalies with appropriate values through time series analysis.

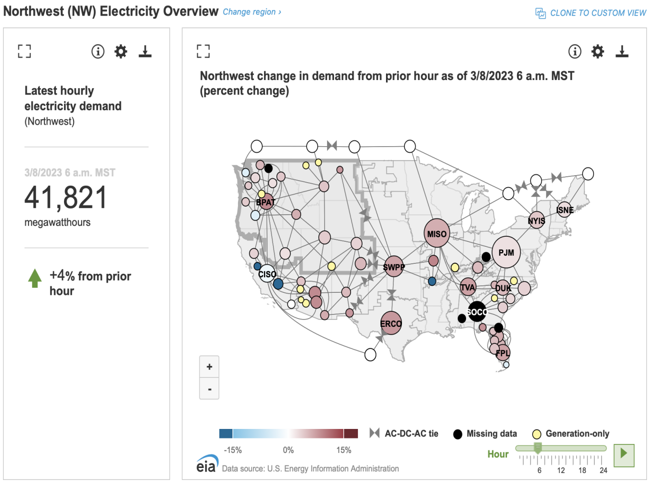
## Data description

The dataset used in this project contains the hourly electric grid data for the northwest region (NW), downloaded from the Energy Information Administration (EIA) website. We choose to focus on the NW region because, unlike regions with large and centralized data sources, the NW region is composed of many smaller data sources, making it more likely to have anomalies. The data spans from July 1, 2015, to Feb. 10, 2023, and has over 66000 hourly observations with a size of 22 MB. Features in the data include demand, demand forecast, aggregated net generation, net generation from various energy sources, and other features like CO2 emissions, etc.  With these features and a well-maintained 8-year dataset, we are able to algorithmically detect the anomalies in the data, especially demand.



Northwest (NW) region electricity overview (demand, forecast demand, net generation, and total interchange)
3/1/2023 – 3/8/2023, Mountain Time

Data source: U.S. Energy Information Administration



Northwest (NW) Electricity Overview *Change region ›*

*CLONE TO CUSTOM VIEW*

Latest hourly electricity demand (Northwest)

3/8/2023 6 a.m. MST

41,821

megawatthours

+4% from prior hour

Northwest change in demand from prior hour as of 3/8/2023 6 a.m. MST (percent change)

Data source: U.S. Energy Information Administration

## Analysis plan

First, clean and prepare the data by handling missing values, like KNN imputation, normalizing the data, and transforming it into a format that is suitable for modeling. Then, Identify relevant features and generate new ones if needed. Choose an appropriate model for anomaly detection such as Isolation Forest, Gaussian mixture model, One-Class SVM, or Autoencoder Neural Network. Train the selected model on the preprocessed data. We will evaluate the performance of the models and compare them by using metrics such as precision, recall, F1 score, and ROC curve. Finally, we will perform time series analysis to replace these anomaly values.