DE 300 Final Project Proposal

Ohmies: Allegra DeCherrie, Brodey Lu, Hannah Wilks

Electricity demand varies significantly over time due to factors such as weather, human activity patterns, and economic conditions. Accurately anticipating periods of peak demand is a critical challenge for power system operators, as underestimation can lead to increased operational costs, grid instability, or even large-scale outages, while overestimation results in inefficient energy generation and unnecessary environmental impact.

Growing electrification of transportation and infrastructure, as well as the integration of renewable energy sources, are making modern energy systems increasingly more complex. As a result, more reliable and accurate predictive models are required to ensure efficient and stable grid operation.

In this project, we propose to design an end-to-end data pipeline that uses historical electricity consumption data from a variety of grid operators to predict future peak demand. Our system will ingest updated load data, preprocess it through feature engineering and cleaning, and train predictive models to forecast upcoming peak periods. In particular, we aim to understand how electrical load impacts infrastructure, identify likely peak hours for dynamic pricing and battery discharge, and analyze demand patterns across different regional systems. We then hope to present the results in a dashboard.

Much of the data we intend to use is published by Independent System Operators (ISOs) or Regional Transmission Organizations (RTOs), which perform similar roles in managing regional electricity markets and coordinating high-voltage transmission systems. ISOs and RTOs are non-profit entities that manage a region's high-voltage electricity transmission grid. They ensure reliable power delivery, fair market access for generators, and balance supply and demand in real-time. Rather than owning physical infrastructure, they act as neutral coordinators that operate wholesale electricity markets and maintain grid stability across large geographic areas, often spanning multiple states.

By leveraging publicly available ISO and RTO data, this project aims to produce a practical, reproducible forecasting pipeline that reflects real operational challenges faced by modern power systems. The resulting system will not only demonstrate core data engineering and modeling techniques, but also highlight how predictive analytics can be directly applied to critical infrastructure domains where accuracy, reliability, and scalability are essential.

Our data will be retrieved from various ISOs/RTOs such as ISO New England's ISO Express, California ISO (CAISO) Today's Outlook, NYISO load data, PJM Data Miner2, and the EIA hourly electric grid monitor. Many of these operators supply historical data in tabular format which can be downloaded as Excel or CSV files. This data is provided on hourly and sometimes sub-hourly resolutions. Other sources, like PJM, offer pre-built tools which can be queried for metered load data by the hour. This data can be downloaded directly from the operator websites, and used to train a predictive model which would take current data as input and return a projected peak.

The load data is useful for forecasting peak load hours because they reflect actual demand over time at a very fine resolution. Historical patterns can reveal useful trends in seasonal peaks, daily cycles, and any potential outliers. For a more holistic prediction model, we can combine load data with weather information, economic conditions, and other relevant factors to build a more robust architecture.

The amount of data depends on the granularity of our data sources. Hourly data would yield ~8,700 rows per zone per year while 15-minute sampling intervals would yield ~35,000 rows per zone per year. If we are sampling multiple zones over 5-10 years, this could easily result in several hundred rows

of data spread across multiple tables (potentially millions if using something like ERCOT that provides 5-minute resolution). This is a fairly large dataset.

With the data we ingest, we have three goals we hope to achieve from our collection.

Goal 1: Create a detailed database of different related data.

We hope to ingest a large amount of load data from a variety of operators and online resources and compile them to understand how different factors may predict large amounts of load. Pulling from different ISO/RTOs and online resources like weather databases, will allow us to collect a large amount of data and relate it to each other. We hope that whatever pipeline of data we build, that it can be scalable across the different ISOs. They may contain different data (location/region level specificity, weather, key date information (like a holiday)), which all contribute to our ability to understand what is happening on the grid at the time they are collected. We can standardize labels like: `time`, `iso`, `zone`, `load_mw`, `temp_f`, `load_forecast_mw`, and `lmp_usd_mwh` to organize data into usable information. For missing or differently chunked data, we can standardize the smallest increment we have between sources. This may require us to average or linearly interpolate missing gaps or median by hour of the day/day of the week. If data ends up being imputed, this can be a flag in the row. The data can be pulled in hourly and then stored in the database so it does not have to be called for again.

Goal 2: Forecast consumption peaks and peak hour classification

We hope to use this curated dataset to predict electricity demand peaks. We aim to forecast the maximum hourly load for the upcoming day and to estimate which hour is most likely to be the peak. These predictions are directly relevant for applications such as battery discharge scheduling, grid reliability planning, and energy market pricing strategies. We will treat this as a supervised learning problem. For each hour in historical data, the target variable is the maximum load observed over the following 24 hours. Input features can include lagged load, rolling statistics (over the past 24 hours), calendar features, seasonality, and weather variables, computable with the PySpark window functions. With the models, we can start with a baseline model like naive persistence (tomorrow's peak is today's peak). We can then try models like linear regression or XGBoost for peak MW and multinomial logistic regression for the peak hour. Because the data is both categorical and numerical, we may consider a neural network model as well. For scalability we can train per-ISO/RTO and cross-ISO/RTO. PySpark will allow us to do this with scalability between different amounts of data or locations of data. With the training/prediction, we can answer questions such as how accurately tomorrow's peak can be predicted, how peak magnitude varies by season, and whether models trained on one ISO generalize to others. However, we cannot predict exact real-time market prices or operational decisions, as those depend on factors not present in the public datasets, such as generator bids, outages, and internal operator constraints.

Goal 3: Make the results ingestible and understandable

We like the idea of displaying the data in a dashboard or making the peak predictions come out of the model in some sort of time-series report. For example, daily peak forecasts could be written to a table with columns such as `iso`, `date`, `predicted_peak_mw`, and `predicted_peak_hour`. These outputs could then be displayed in a dashboard that shows historical load trends, forecast accuracy, and upcoming risk periods. By storing both raw data and processed results in structured databases, users can explore historical predictions without rerunning the entire pipeline. While this goal is not directly related to the required parts, it will make the results easier for people to understand. It relies on the use of the above techniques.

References:

1. PJM Interconnection. (n.d.). *Data Miner 2: List of data feeds*. PJM. Retrieved February 3, 2026, from https://dataminer2.pjm.com/list
2. New York ISO. (n.d.). *Load Data.* NYISO. Retrieved February 3, 2026, from https://www.nyiso.com/load-data
3. ISO New England. (n.d.). *Markets and Operations ISO Express*. ISO-NE. Retrieved February 3, 2026, from https://www.iso-ne.com/markets-operations/iso-express
4. California ISO (n.d.). *Today's Outlook Current and Forecasted Demand.* CAISO. Retrieved February 3, 2026, from https://www.caiso.com/todays-outlook
5. U.S. Energy Information Administration. (n.d.). *Hourly Electric Grid Monitor*. EIA. Retrieved February 3, 2026, from https://www.eia.gov/electricity/gridmonitor/dashboard/electric_overview/US48/US48
6. The Weather Company API Hub. (n.d.). *Weather APIs*. The Weather Company. Retrieved February, 3, 2026m from https://developer.weather.com/data/default/introduction