# Overview

This directory contains an LAURed (unlimited releasable) power dataset from two experiments. These were conducted by a contractor, Grey Beard Engineering. The data is divided into “original\_data” and “new\_data”. In both sets, there are some “good” and some “bad” files - they refer to these as “nominal” and “off nominal” (sometimes “abnormal”).

The data is CSV time series files with time (in seconds) and voltage and current captured from 3 phase power.

# Data Source (.zip)

https://drive.google.com/file/d/12TdaQgiNQolx\_QFEM3EtLLDO2JtFSmNx/view?usp=sharing

# Tasks

1. Get the data / unzip it, familiarize yourself w/ the file and directory structure
2. Take a look at some of the CSV files
3. I suggest starting with the “original\_data” as it is easier to find anomalies than it is in the new data - be warned that your techniques for finding anomalies may not translate from original to “new” data.
4. Read the READMEs in the zip file to get an idea what you’re looking at. There are some example plots in there.
5. Read about 3 phase power and NILM (non-intrusive load monitoring) so you can understand why we are doing this.
6. Load some data up and make some plots of voltage, current, power (figure the trivial equation out, if it’s not something you know already). Plot the 3 phases - classic exploratory data analytics.
7. The files have a lot of samples (~300k rows / time series events). Consider downsampling techniques such as windowing, moving averages, sliding windows, etc. These are all open to exploration.
8. Test out simple statistical metrics for assisting this - it’s possible this will be helpful.
9. We might need machine learning techniques for this, but I’m not pushing that - open to any ideas.
10. It’s probable that feature engineering of the dataset and then using with ML will be fruitful (such as adding new statistical metrics to the dataset, rolling windows, etc.)
11. I have also attached some matlab code provided by the contractor - you may find some of the calculations in there useful.
12. The ultimate goal of the project is to identify:
    1. Which welds (files) are nominal and which are bad
    2. The off nominal welds are only off nominal for a time period - can we identify *when* the weld goes bad? It might go back to being nominal again at some point (e.g. time 8-10 seconds in a 30 second time slice)
    3. There are also *types* of bad welds - can you identify that the bad welds are “differently bad” than other welds?
    4. As explained below in NOTE, we only want to train on the input values (PHX Voltage and Current, where X is 1,2,3). There are specific reasons for that and it is very likely that using the other columns will make this project too easy. Fine to study initially, but we need to “hide” them in later analysis. The final use of this will not have these columns discussed below - not because it’s a game or a toy problem, but because it is difficult/impossible to capture them in a real manufacturing setting.
13. **NOTE**: In both datasets (original and new) there are “output” and other metadata columns - these are the Arc voltage and current in the original and ArcV/I and WFS (wire feed speed) in the new data. These may provide ground truth “checks” for good/bad but, AND THIS IS IMPORTANT, for the final project these should not be included.