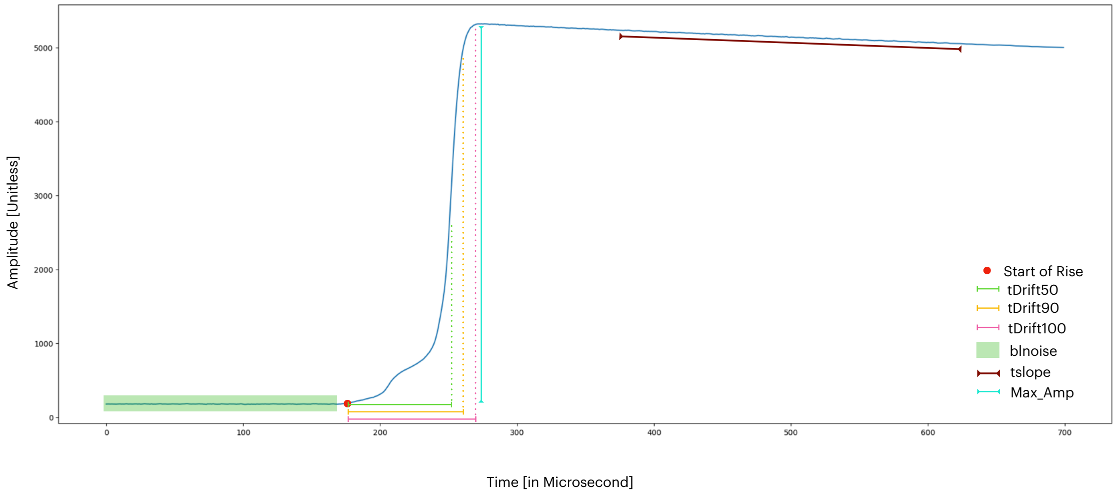
We are hosting a class-wide competition to see who can make the best predictions! Top predictions can earn extra credit on Midterm 1!

High-Purity Germanium (HPGe) detector is one of the most sensitive detectors human beings have ever manufactured. It is sensitive in a sense that it measures the energy of elementary particles (electron, photon, etc) very accurately. Because of this, HPGe detectors have a wide range of applications, including the search for neutrinos and dark matters, medical imaging, as well as nuclear non-proliferation.

When a particle comes into HPGe detector, it produces a waveform, or time series data, as shown in the picture below. A time series is a sequence of data points that occur in successive order over some period of time. More formally, we can define time series this way: for each data point, a time series contains n pairs of $t_i, a_i$ where $t_i$ is the $i^{th}$ time sample and $a_i$ is the value at $i^{th}$ time sample. To simplify this problem, we will not ask you to directly analyze the time series, but we extract certain features from the HPGe time series for you to build model.

In the supplementary Jupyter notebook (linked), you are given access to a CSV containing training data with information about 400 elementary particles which deposits their energy in a HPGe detector. We read this in as a DataFrame where the columns are different parameters. The columns are:

- **Max_Amp:** Maximum amplitude of the waveform, or the largest number among all $a_i$s
- **tDrift50:** Period from the Start of Rise ($t_{SR}$) to when the waveform reaches 50% of Max_Amp ($t_{50}$), can also be written as $t_{50} - t_{SR}$
- **tDrift90:** Period from the Start of Rise ($t_{SR}$) to when the waveform reaches 90% of Max_Amp ($t_{90}$), can also be written as $t_{90} - t_{SR}$
- **tDrift100:** Period from the Start of Rise ($t_{SR}$) to when the waveform reaches Max_Amp, can also be written as $t_{Max_Amp} - t_{SR}$
- **blnoise:** The standard deviation of amplitude values $a_i$ in the green-colored region.
- **tslope:** The slope of the waveform tail.



Your task for this problem is to find the best prediction rule using regression to estimate the energy of each HPGe detector waveforms, given the listed parameters above. The requirements are as follows:

1. You must use regression.

2. The function used for regression is your choice (linear, polynomial, exponential, ...)
3. You may use **up to three variables**. You decide which ones.
4. Your design matrix may have **up to five columns**. You decide what the design matrix looks like.

We've provided you with a function `calculate_MSE` to calculate the mean squared error of your predictions on each waveform in the training data. Your job is to fill in the body of the `predict` function. This function should take as input one row of the DataFrame (corresponding to one particular waveform) and return the predicted energy corresponding to this waveform. How you make this prediction is up to you, subject to the rules above. Feel free to add more cells and functions, and to change the provided `predict` function, but do not change the provided `calculate_MSE` function.

When we grade this question, we will run your prediction function on a hidden test dataset as well, so be mindful of *overfitting* the training data. You'll earn full credit on this homework problem by finding a prediction function whose MSE on the hidden test data is below a certain threshold, which we hope most students will achieve. Additionally, the ten best prediction functions (as determined by the MSE on the hidden test data) will earn some extra credit on the upcoming midterm exam according to the following scheme: for $n \leq 10$, the $n^{th}$ ranked prediction function in the class earns $11 - n$ percentage points as extra credit on Midterm 1.