What to submit:

* Jupyter notebook file (Christain, Jason, Ziheng)
* Pipeline (workflow) diagram (PoHsien)
* 500-word distribution (Liyun)

Questions for AJ:

1. PCA and data-extraction
2. Suggestions on preprocessing
3. Do we need to do background removal when perform PCA.

Feedback from AJ:

1. Depends on how the baseline work to determine next step (Use intuitive features instead of data driven curve, e.g. PCA) [Not required for the coming due]
2. Treat each point of the entire curve as input features and do PCA (I.e. 1000\*41 matrix), extract 3-4 new features [may try GridSearchCV + pipeline, or just cross-validations]. Compare the PCA projection model with our original data.
3. On baseline model: better to do rescaling; can use GridSearchCV to find parameters. For the future, if results are accurate, see if new features change, if not, see how to improve.
4. More than one performance matric (accuracy, recall, precision). Not necessary to deal with class imbalance for baseline model. May use generative model in Module 5.
5. On cross-validation: 40-fold CV (good for small dataset). Best way is to make a prediction and do the test. 90-10 train/test split -> 10-fold CV
6. On pipeline chart: raw data-> data cleaning/preprocessing -> feature extraction (physical feature)->
7. For data imbalance, we would have to come up with a performance metric