Modern data collection techniques have made it possible to find meaningful connections between the severity of a disease in an individual, and the individual’s biology. Biological data collected for these purposes can result in noisy, especially large-dimension data sets. These data sets are typically composed of moderate-to-highly correlated features, where the correlation structure is unknown prior to data collection. Additionally, many of the data features may have no connection to a quantity of interest. Finding meaningful associations between an outcome of interest, such as disease phenotype severity or a disease risk score, and the feature space is challenging in this data scenario.

One of the ways in which statistical researchers working with biological data provide interpretability to these data sets is through the training of predictive statistical models, like ordinary least squares regression (OLS). When the OLS approach is taken, moderate to high correlation in the feature space leads to an unstable, highly variable model.

If some of the features of the data are unrelated to the quantity of interest, they add no information to the model. Thus, a researcher would like a regularized (penalized) modeling procedure that accurately identifies the unrelated features, and sets their model coefficients to 0, resulting in a sparse model. (OLS does not return a sparse model)

The ridge regularization procedure was developed to address the moderate-highly correlated data scenario, yielding stable statistical models, while Lasso regularization was introduced to accurately identify sparse statistical models. The elastic net combines both regularization procedures to yield a stable, sparse statistical model.

However, it has been shown that as the dimension of the feature space grows along with the number of observations, the lasso and the elastic net do not asymptotically converge in probability to the true nonzero coefficients of the underlying model. This non-asymptotic convergence in probability is referred to as *inconsistency*, and thus the lasso and elastic net are inconsistent estimators. It has been demonstrated that when the signal to noise ratio is low, the lasso yields highly erroneous models.

Consistency of estimation, or convergence in probability of an estimator to a parameter, is a desirable property for any statistical modeling procedure, as it ensures the accuracy of the estimates of the model coefficients improves as sample size increases. It has been shown consistency can be introduced to the lasso and the elastic net by:

1. Adopting additional regularization constraints.
2. Weighting the coefficients in the L1 penalty term of the lasso and elastic net with a well-defined data-adaptive weight.

These data adaptive versions of the lasso and the elastic net are referred to as the adaptive lasso and adaptive elastic net.

The consistency results are not only theoretically interesting, but demonstrably useful when modeling large dimensional data-sets with a moderate to high correlation structure. Yet, to date, an easy-to-use computational package has not been developed or implemented in Python. Combining Python, Sklearn and CVXPy, I have been able to develop a set of Python functions that allow users to easily implement a cross-validated, data adaptive lasso and elastic net.

In this presentation I provide a brief introduction to linear statistical models, and demonstrate the issues that arise when applying these models to correlated, noisy data sets. This is followed by a demonstration in Python on how regularization addresses these issues. I also demonstrate the inconsistency of the lasso and the elastic net estimators, and the consistency of the adaptive lasso and adaptive elastic net estimators. These data adaptive regression functions are described and made available for anyone interested in using them for personal or research use. My presentation and aforementioned code live in Github in a Jupyter notebook, which will be sent out in a follow up announcement.