**Presentation Notes:**

* Dataset:
  + Dataset was post secondary institutions, concentrated of 4 year public institutions.
  + Over 400 variables detailing institution characteristics e.g. revenues, expenditures, student demographics, degree completions, staff, financial aid levels etc
  + Split by testing into testing and training data set first off all. 80% train and 20% testing. Stratified by region – not by state. Clustered by institution i.e. all years for institution included if institution selected.
  + Excluded around 30 institutions with missing grad rates.
* Approach 2 Summary:
  + More of a prediction problem.
* We had a lot of variables so we were susceptible to overfitting. We tried to predict our outcomes, using regularized learning methods.
* Kept the original data intact, i.e. did not carry out dimension reduction through PCA, so as to make interpretations with our model.
* Same schools as in approach 1
* Removed variables colinear to outcome e.g. fte\_retention rate, total\_completions per fte.
* Model selection through Lasso - kept 10 largest absolute value coefficients – kind of arbitrary.
* Variable Selection
  + Federal Grants show up
  + Part time variable show up
  + Also get some expenditure related variables.
* Models:
  + Tried Lasso, Ridge, Elastic net on the data, Lasso performs the best
  + With selected variables, adding interaction terms just leads to overfitting
  + Gaussian Kernalised regression performs the best, perhaps suggesting non linearity
  + Bachelor degrees generally more noisey.
* Average marginal effects Kernalised model:
  + Seems reasonable, signs are what we expect, and some coefficients with surprisingly large or strange signed lasso coefficients have dropped out.
* Overall
  + Automatic variable selection gave good predictors, in correspondence with the variables of interest, which gives credibility to the variables selected.