Rebecca Conover
Mathematics
Class of 2019

**How does Susan Athey argue that machine learning methods can help take the 'con' out of 'econometrics'?**

Economist have historically looked to theoretical econometric frameworks to make causal inferences on data sets with a relatively small amount of predictors. But, a lot goes on behind the scenes to get to this point that isn't necessarily statistically, or ethically sound. Athey's example of a hypothetical pharmaceutical trail illustrates this well, say you have data on the effectiveness of a drug for one thousand people, and say you have one thousand co-variates for each person. It wouldn't be too hard to find a few characteristics that the people who responded well to the drug had in common, just because there are so many covariates. But this result would be highly specific to the training set and probably could not be replicated in a future study. This p-value hacking happens often in Econometrics When running many variations on a data set, the classical frameworks of hypothesis testing don't hold, but results are (or were) presented as they do.

Machine learning provides some tools to remedy these issues of honesty and improve the predicative power of these models. Regularization methods can help to eliminate spurious results and select necessary covariates, separating out proxy variables from the variables that have potential causal effects. Economist have effective tools for small data causal inference, Data Scientists have effective tools for big data predictions. Athey argues that by bringing together the strengths of each side, using larger and more complex sets of data/variables but constraining the models to give more reliable and interpretable results in order to better understand where causal inference is or is not warranted. Companies like amazon and google lean use black boxes that produce valid results but they are extremely context specific and effective mostly in the short term. Econometric methods and theory have the power to give results context and make causal statements about some of the results of machine learning models without resorting to p-hacking or trying to constrain analysis to traditional hypothesis testing methods and potentially misinterpreting results.