David Valentin
Computer Science
Sophomore '19
Statistical Learning
Professor Kim

How can machine learning methods can help take the 'con' out of 'econometrics'?

First lets address the issue that Russ Roberts and Susan Athey address which is that data can be mined and shaped till the desired results of the desired hypothesis are shown, and that underscores 'con' in "econometrics.' With modern data collection, Athey describes the issue of data now having more variables or predictors than actual observations in which she describes how the variables begin to generate this "data-driven" hypothesis where hypothesis are created from small samples sizes that have spurious correlated results due to the nature of having so many variables or predictors. The main ways that Athey describes several ways to minimize this which is too utilize the causal variable that you are investigating in your models, and treat it separately while making sure you treat the control variables differently. The important rule that Athey notes is that in econometrics is too avoid simply dumping it into the regression model. In addition, Athey notes how machine learning utilizes shrinkage to prioritize certain variables and shrink others that might be a proxy for important variables. Despite the benefits of machine learning utilization in econometrics, Athey also highlights the dangers of machine learning simply being utilized to draw connections between variables over large data sets, and emphasizes the need for sample splitting to compensate for overfitting.