Elias Van Sickle, Mathematics, 18'

Econometrics is the branch of economics concerned with the use of mathematical methods (especially statistics) in describing economic systems. While econometrics has yielded information that has certainly resulted in important social science work and policy changes, it is also fraught with certain challenges, chief among them being the potential for "finding stuff that isn't there". As a result of this issue, which has necessitated that large quantities of research be invalidated in certain fields/cases, an effort has developed to take the "con" out of econometrics. If implemented carefully, computers and machine learning techniques have great potential to assist in this effort. To understand the role machine learning can play, it is first necessary to unpack what "finding stuff that isn't there" entails.

The phrase data mining has certain negative connotations associated with it, and for good reason. Say a researcher is analyzing a set of data with a certain hypothesis in mind. With data mining, or more pejoratively "fishing", there is the potential that this researcher will be able to reveal a preconceived insight by trying many many models based on various combinations of control variables or specifications. In this way, the data is not speaking for itself, but rather the researcher is making the data say whatever he or she desires it to say. With each statistical inference comes a standard error measurement known as a p-value, which reports the chance that a certain finding could have occurred by chance. If a low p-value is reported with results derived from fishing, then the assertion is made that the result was statistically significant or that it is unlikely to have occurred by simple chance. The problem is that because the finding was derived by mining the data, such an assertion is likely to be misleading or flat out wrong.

To prevent data mining issues, an alternative "data driven" approach may be taken. In this paradigm, the data is meant to speak for itself and it is here where machine learning comes into play. Instead of the researcher mining the data looking for the best fit, the machine runs a slew of different models on a set of data to find the ones that fit best. These models are then implemented on other similar data sets to make predictions. Sample splitting, or only using a portion of the entire data set to train the machine learning algorithm, is extremely important to ensure that the models are generalizable. The focus of machine learning has long been big data prediction, whereas the focus of econometrics has been small data causal inference. The potential exists to leverage the power of machine learning to fit large sets of data with many variables while maintaining the practice of systematic model selection central to econometrics so as to ensure valid statistical inference. It is the collaboration between man and machine that will yield the better results than when either party works in isolation.