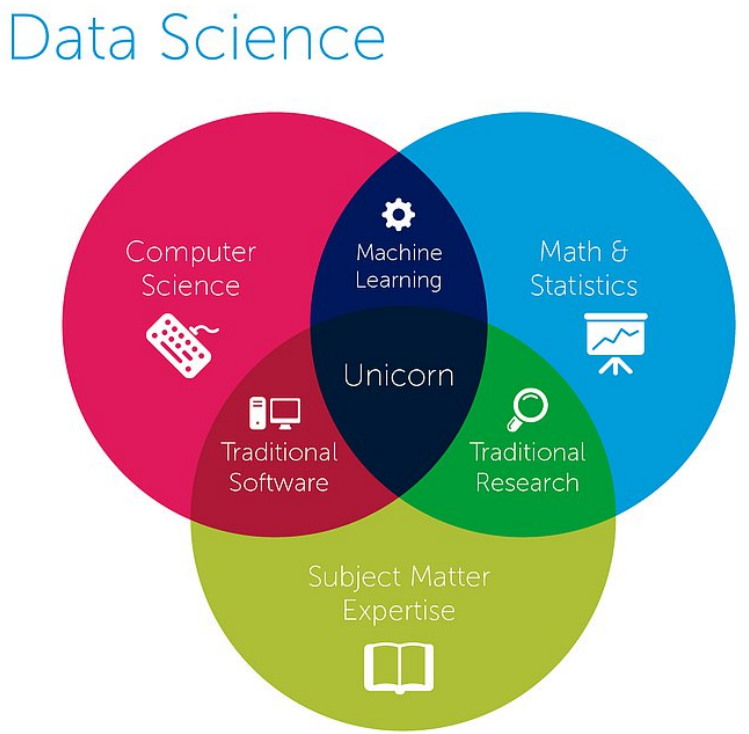
Statistical Learning: Introduction

36-290 – Introduction to Statistical Research Methodology

Week 1 – Fall 2021

Data Science: The Search for the Unicorn



(Credit: Computer Science Department, Luther College)

What is Statistical Learning?

Statistical learning is the process of ascertaining *associations* between groups of variables.

Conventionally, we attempt to uncover associations between a set of *predictor* (or *independent* or *explanatory*) variables and a single *response* (or *dependent*) variable. (In the terminology of machine learning, this is dubbed *supervised* learning. We'll discuss unsupervised learning elsewhere.)

Example: we measure the brightness of a galaxy at a set of p wavelengths. Can we uncover a meaningful relationship between these measurements (which are the predictors) and the galaxy's distance (which is the response)?

Example: we are given receipts from purchases at Lowes and Home Depot. Can we apply text analysis techniques that would allow us to confidently predict which store a new receipt came from? (Let's assume the new receipt does not actually have the name of the store!)

Examples of Statistical Learning Algorithms

You are probably already familiar with statistical learning, even if you did not know exactly what the phrase meant before now. Examples of statistical learning algorithms include:

- Linear regression and its variants (e.g., variable selection methods)
- Logistic regression
- Generalized additive models
- Trees and its variants (e.g., random forest, boosting)
- Support vector machines
- Deep learning (e.g., convolutional neural networks)

Some of these are examples of *regression* models, in which we estimate the average value of the response variable *given* values for the predictor variables.

And some are examples of *classification* models, in which we attempt to determine which of a set of discrete response variable classes is most likely, *given* values for the predictor variables.

Which Algorithms Should I Apply in an Analysis?

That depends somewhat on what you want to do. Let's start with the big picture: inference vs. prediction.

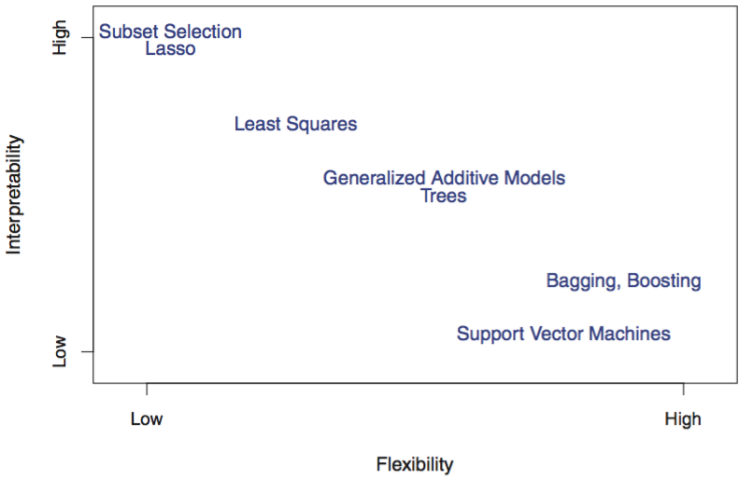
Let Y represent our response variable, and X represent our predictors. Then our learned model will take the form

$$\hat{Y} = \hat{f}(X)$$

- (The hat represents an estimate, as opposed to the unknown true value or true function.)
- If you care about the details of $\hat{f}(X)$, then you want to perform statistical inference.
- If you treat $\hat{f}(X)$ as a black-box, then your interest is in prediction.

One can perform prediction with any algorithm. For inference, though, the choices are more limited.

Model Flexibility vs. Interpretability



(Figure 2.7, *Introduction to Statistical Learning* by James et al.)

In general, there is a tradeoff between the "flexibility" of a model (i.e., how "curvy" it is) and how interpretable it is: the simpler the parametric form of the model, the simpler to interpret. (This in large part motivates the use of linear regression in practice.)

So-called *parametric* models, for which we can write down a mathematical expression for $f(X)$ *a priori* (as we can for linear regression), are inherently less flexible. Parametric models may be contrasted with *nonparametric* models, in which $f(X)$ is estimated from the data themselves.

If prediction is your goal, then your model can be as arbitrarily flexible as it needs to be. We'll discuss how one estimates the optimal amount of flexibility in another set of notes.