Machine Learning Pipeline *(find and work with a partner)*

1. **Cross-validation**

   (a) What are the costs of k-fold cross validation?

   (b) What are the pros and cons of no longer having one model?

   (c) How would you choose k and would it depend on the amount of data?

2. **Communities and Crime dataset**

   In Lab 2 we are analyzing the Communities and Crime dataset. A description of the predictive variable (i.e. y) is included below. Discuss the following questions:

   (a) The name of the predictive variable is `ViolentCrimesPerPop` in the file `communities.names`. Given the more detailed description below, what sources of error are in a generic machine leaning pipeline if we train a model to predict `ViolentCrimesPerPop`?

   (b) What would you list as the intended vs. unintended uses of this model?

The predictive variable in the Communities and Crime dataset, violent crimes per population, is determined by using the FBI Uniform Crime Reporting Program (UCR) data from 1995, normalized by the population data from the 1990 U.S. census. The FBI UCR relies on voluntary reporting by local law enforcement agencies and the violent crimes count used to calculate the predictive variable in this dataset includes counts of murder, rape, robbery, and assault. The dataset information, available at `https://archive.ics.uci.edu/dataset/183/communities+and+crime`, notes that: "There was apparently some controversy in some states concerning the counting of rapes. These resulted in missing values for rape, which resulted in incorrect values for per capita violent crime. These cities are not included in the dataset. Many of these omitted communities were from the midwestern USA."