Download the lymphoma and colon data from the homework page. The lymphoma data is the result of measuring gene expression of patients with two types of lymphoma. The data measures the transcript level (amount of mRNA) of 4026 genes. The colon data measures the transcript level of 2000 genes for patients with or without colon cancer. Feature selection can potentially provide some insight on what differentiates tissues that come from different tumor types, or what differentiates tumor from non-tumor tissue. A small subset of features can also form the basis for a cheaper diagnostic than one that requires measurement of thousands of genes. In the latter case correct evaluation of the selected set of genes is crucial.

As discussed in class, when assessing the performance of a feature selection method, the process of feature selection should be considered a part of the training of the classifier with which it is used, i.e. it is a mistake to first perform feature selection, and then to perform cross-validation or leave-one-out. Illustrate this bias in the case of the RFE feature selection method using the lymphoma and colon datasets. You can use the implementation of RFE in PyML, which is in the featsel module. The PyML tutorial will give you helpful information on how to perform this experiment. Be careful in your analysis — these are small datasets (small number of examples). What are the potential pitfalls of a small number of examples?

The type of bias that occurs when doing feature selection outside of your cross-validation loop can occur whenever doing model selection among a large number of potential classifiers. Please explain!

Reference