The book introduces Statistical Learning by providing three different examples.

1 Email Spam

The data used was downloaded from Spam dataset as part of UCI reporsitory. Its a classification problem where you want to filter out spam without losing any relevant emails. The data has features like certain *WORD* frequency, *Special* character frequency, word length etc. This is used to classify whether email is spam or not.

2 Prostate Cancer

The goal is to predict the *lpsa* log of PSA antigen in prostate cacncer using different clinical features. This was more of a regression problem.

3 Digit Recognition

Handwritten ZIP code database is used to classify different digits for USPS. The images have been normalized to 16x16 image. This is a classification problem. Here the error-rate needs to be low, so above a certain threshold the mail get classified into "Dont Know" category. These are then classified by hand.

4 Gene Expression Microarrays