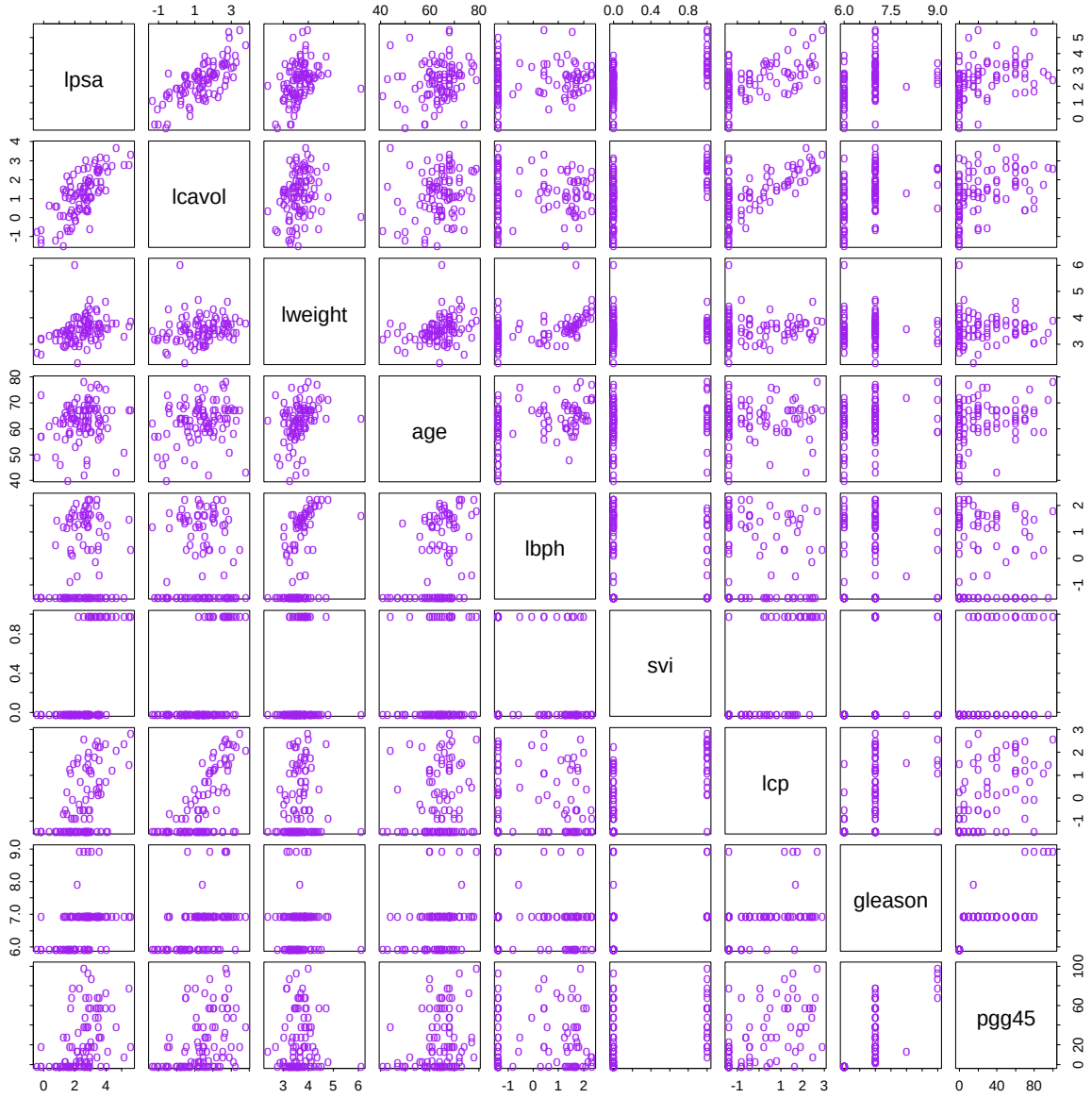
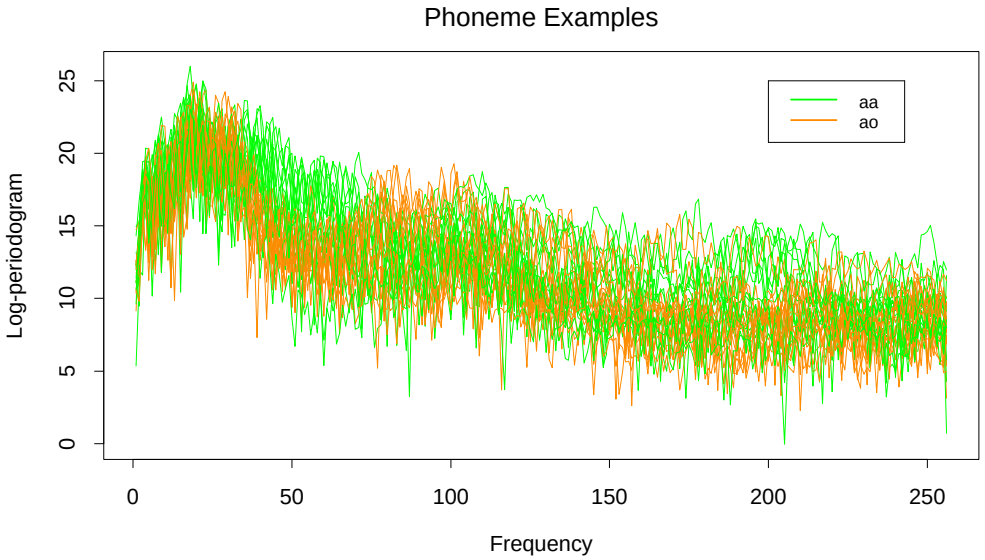
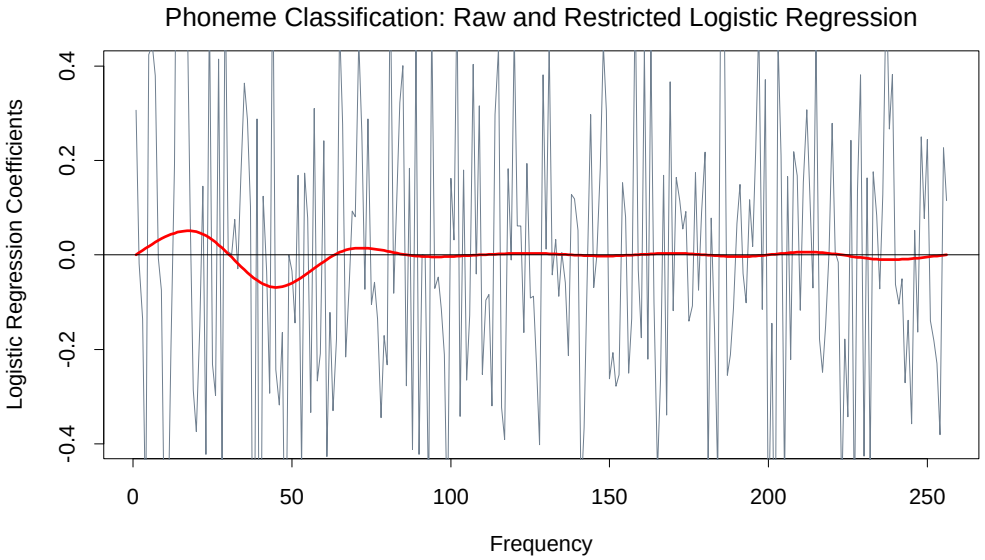
Statistical Learning Problems

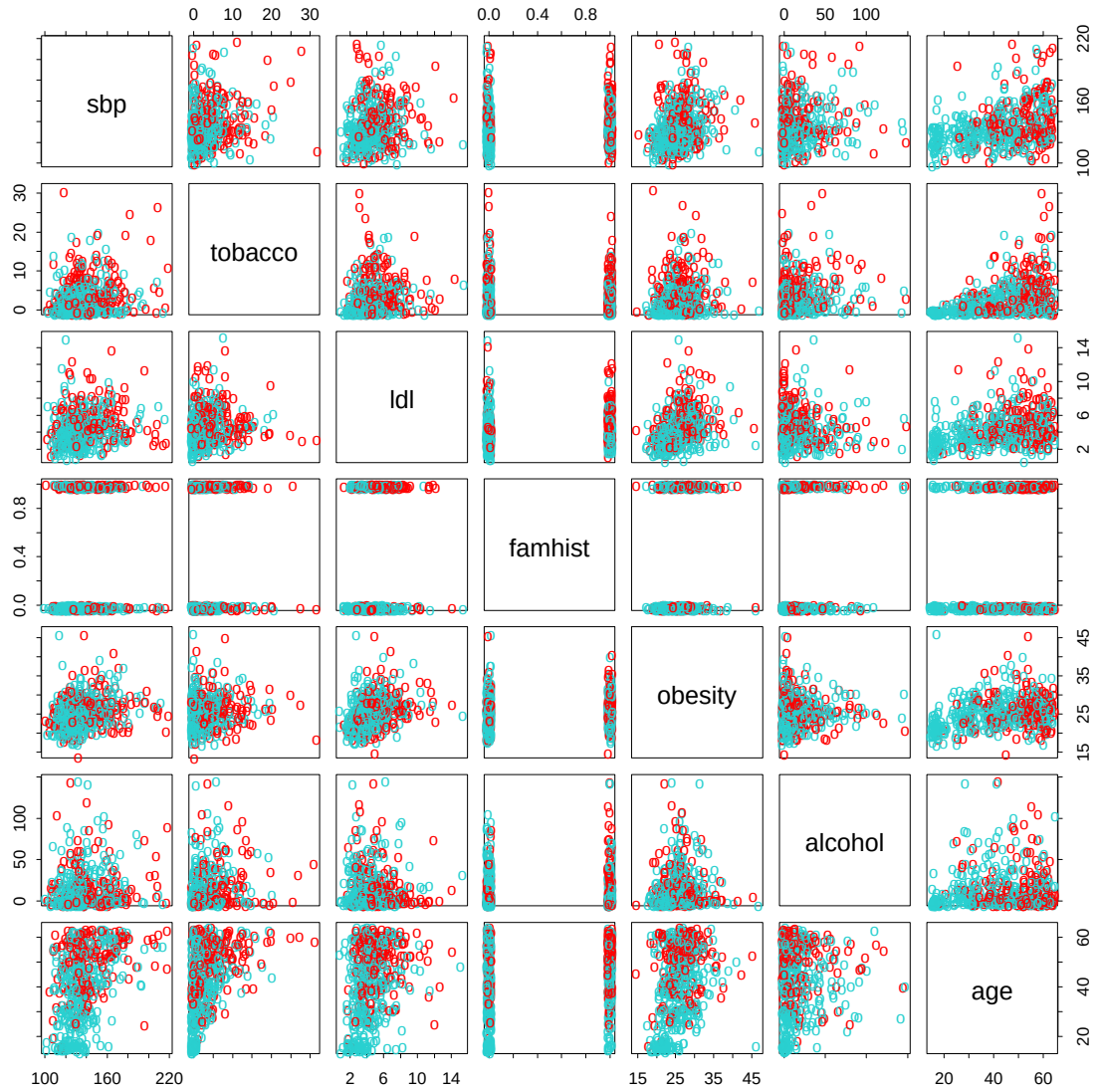
- Identify the risk factors for prostate cancer
- Classify a recorded phoneme (Fig 5.5) based on a log-periodogram.
- Predict whether someone will have a heart attack (Fig 4-12) on the basis of demographic, diet and clinical measurements
- Customize an email spam (Tab 1.1) detection system.
- Identify the numbers in a handwritten zip code (Fig 1.2), from a digitized image
- Classify a tissue sample into one of several cancer classes, based on a gene expression (Fig 1.3) profile
- Classify the pixels in a LANDSAT (Fig 13.6) image, by usage:

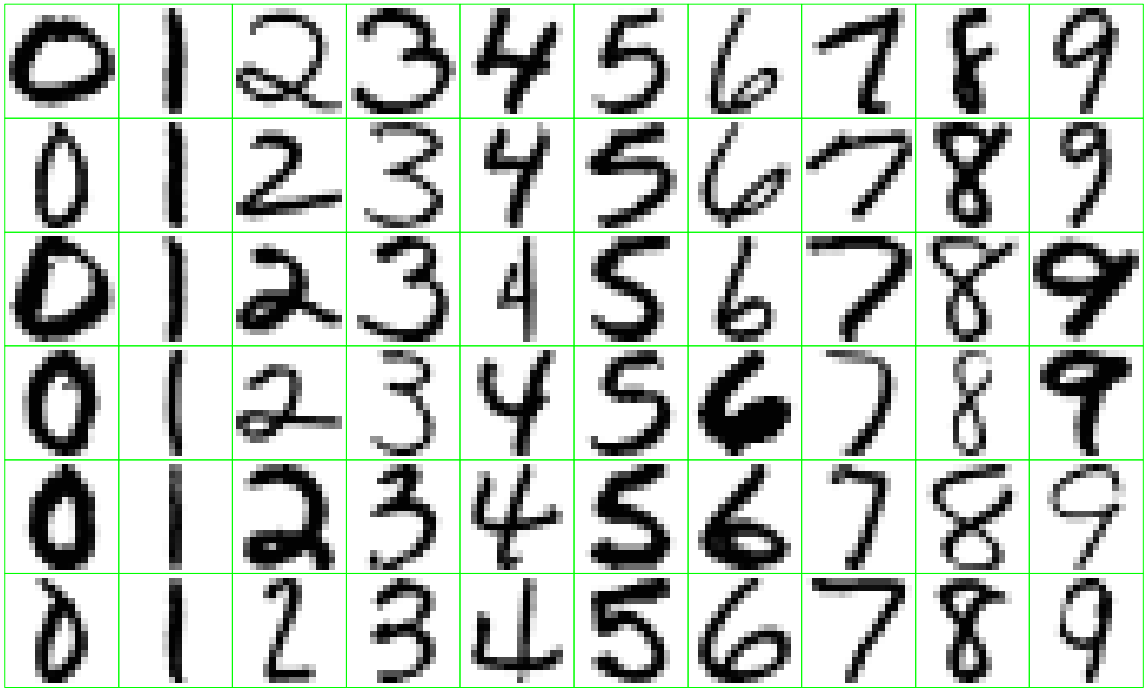
  { *red soil, cotton, vegetation stubble, mixture, gray soil, damp gray soil*}

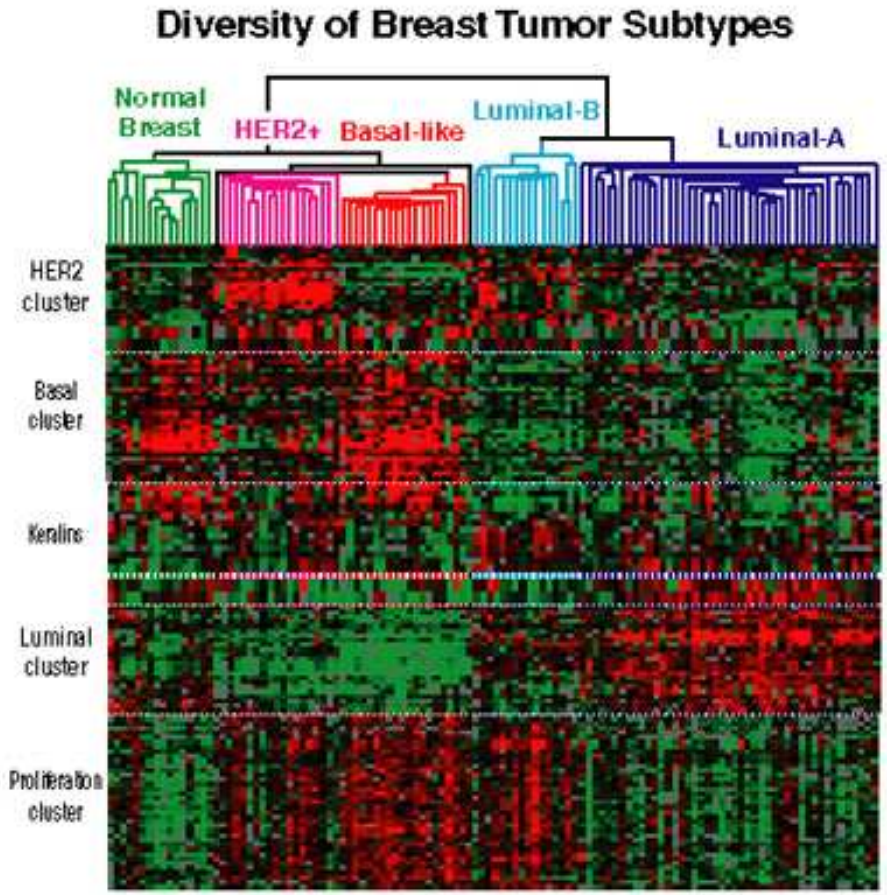

lpsa
lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45


Phoneme Examples
aa
ao
Log-periodogram
0
5
10
15
20
25
0
50
100
150
200
250
Frequency

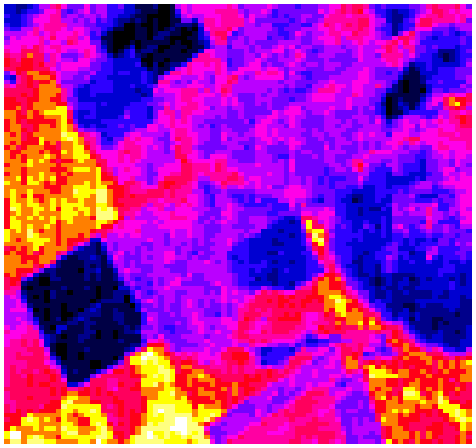

Phoneme Classification: Raw and Restricted Logistic Regression
Logistic Regression Coefficients
-0.4
-0.2
0.0
0.2
0.4
0
50
100
150
200
250
Frequency


sbp
tobacco
ldl
famhist
obesity
alcohol
age

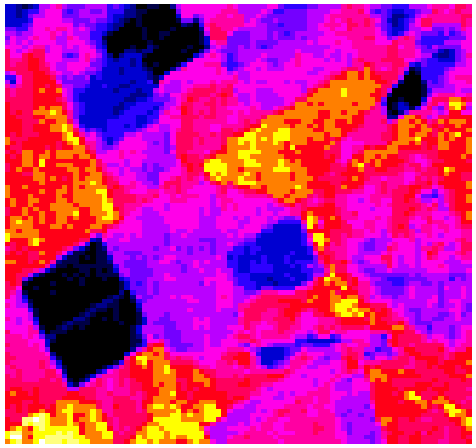



Diversity of Breast Tumor Subtypes
Normal Breast
HER2+
Basal-like
Luminal-B
Luminal-A
HER2 cluster
Basal cluster
Keratins
Luminal cluster
Proliferation cluster

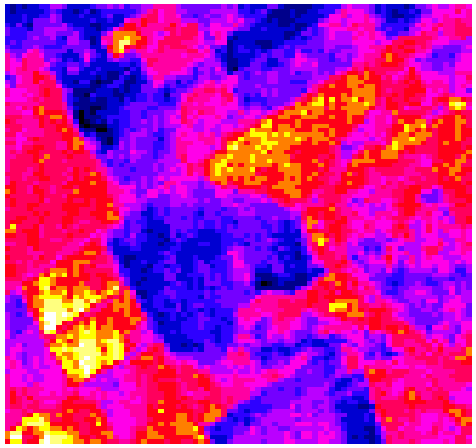
Spectral Band 1



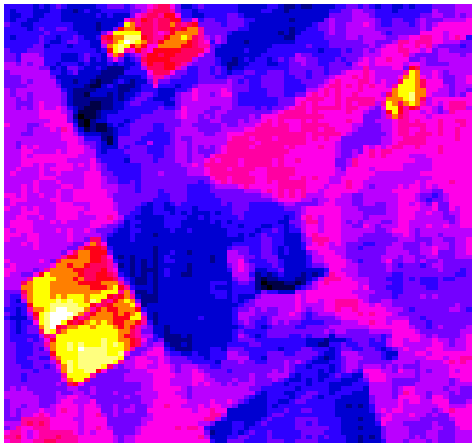
Spectral Band 2



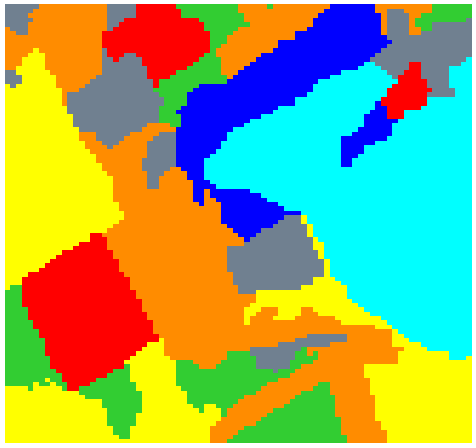
Spectral Band 3



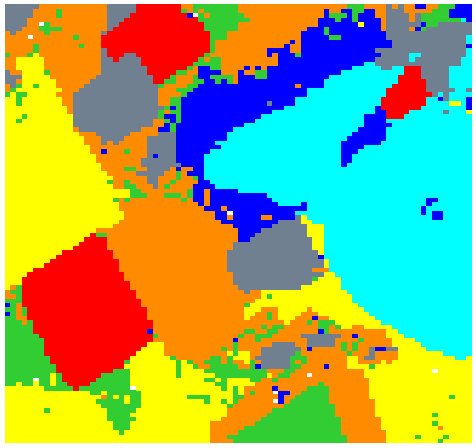
Spectral Band 4



Land Usage



Predicted Land Usage



Spam detection

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

The Supervised Learning Problem

Starting point:

- Outcome measurement Y (also called dependent variable, response, target)
- Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables)
- In the *regression problem*, Y is quantitative (e.g price, blood pressure)
- In the *classification problem*, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample)
- We have training data $(x_1, y_1), \ldots, (x_N, y_N)$. These are observations (examples, instances) of these measurements.

Objectives

On the basis of the training data we would like to:

- Accurately predict unseen test cases
- Understand which inputs affect the outcome, and how
- Assess the quality of our predictions and inferences

Philosophy

- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- One has to understand the simpler methods first, in order to grasp the more sophisticated ones.
- It is important to accurately assess the performance of a method, to know how well or how badly it is working [simpler methods often perform as well as fancier ones!]
- This is an exciting research area, having important applications in science, industry and finance.

Unsupervised learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- objective is more fuzzy- find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- difficult to know how well your are doing
- different from supervised learning, but can be useful as a pre-processing step for supervised learning

The Netflix prize

- competition started in October 2006. Training data is ratings for $18,000$ movies by $400,000$ Netflix customers, each rating between 1 and 5
- training data is very sparse- about 98% missing
- objective is to predict the rating for a set of 1 million customer-movie pairs that are missing in the training data
- Netflix's current algorithm achieves a root MSE of .95. The first team to get achieve a RMSE of .85 wins 1 million dollars.
- is this a supervised or unsupervised problem?