MA8701
General Statistical Methods

Spring 2011

Bo Lindqvist

Statistical Learning Problems

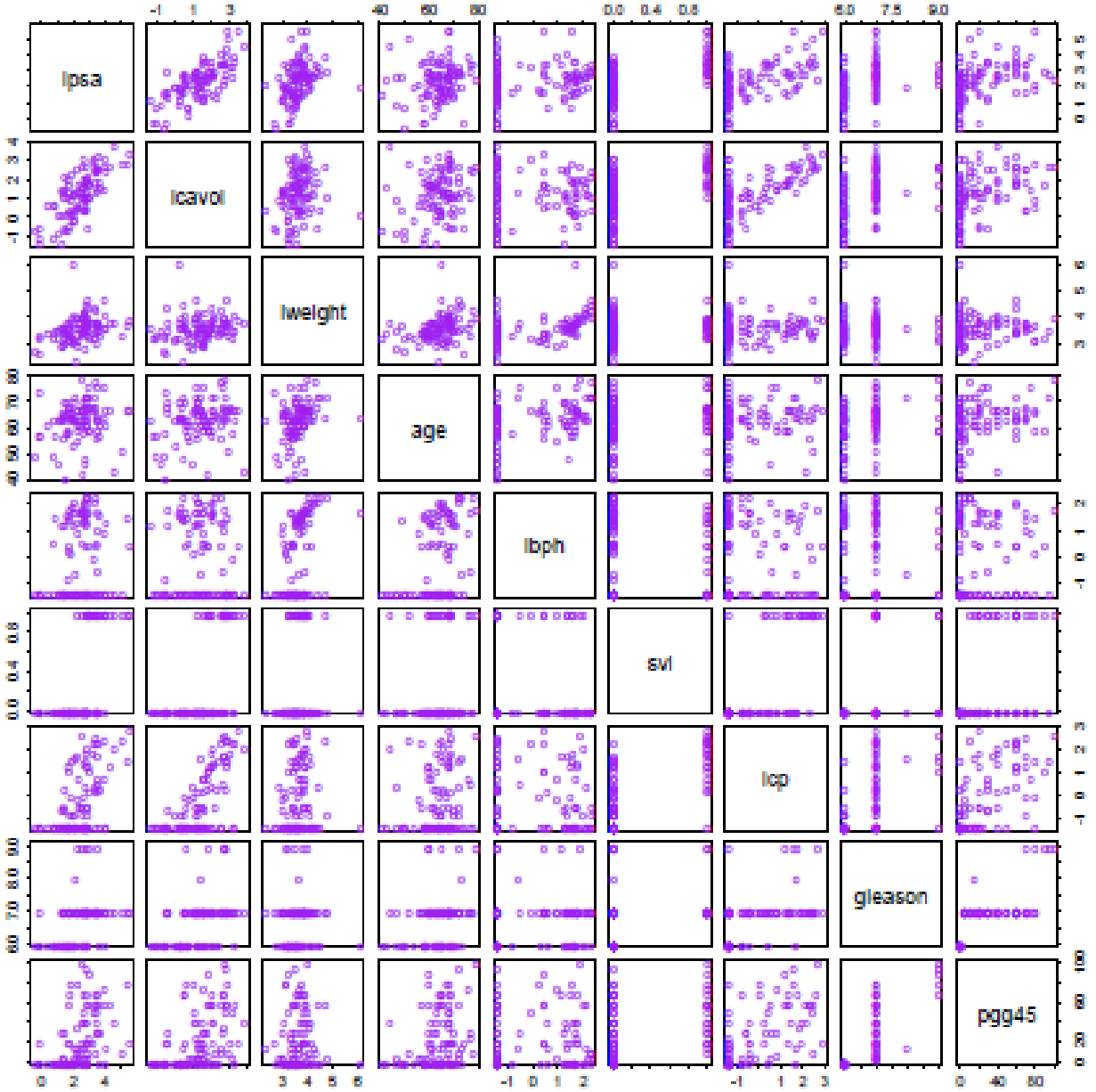
- Identify the risk factors for prostate cancer
- Classify a recorded phoneme (Fig 5.5) based on a log-periodogram.
- Predict whether someone will have a heart attack (Fig 4-12) on the basis of demographic, diet and clinical measurements
- Customize an email spam (Tab 1.1) detection system.
- Identify 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:

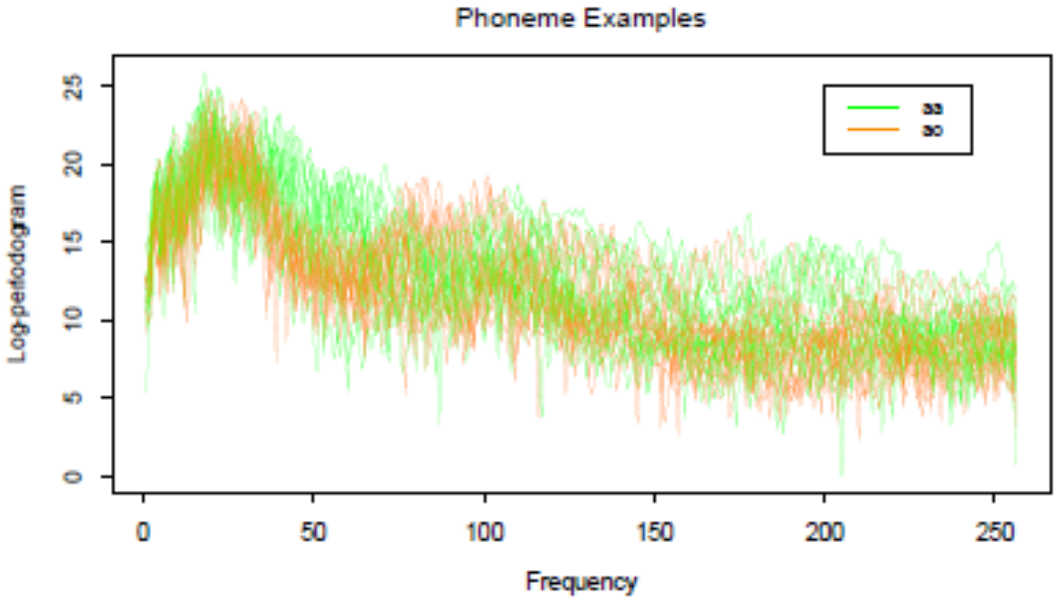
  { *red soil, cotton, vegetation stubble, mixture, gray soil, damp gray soil*}

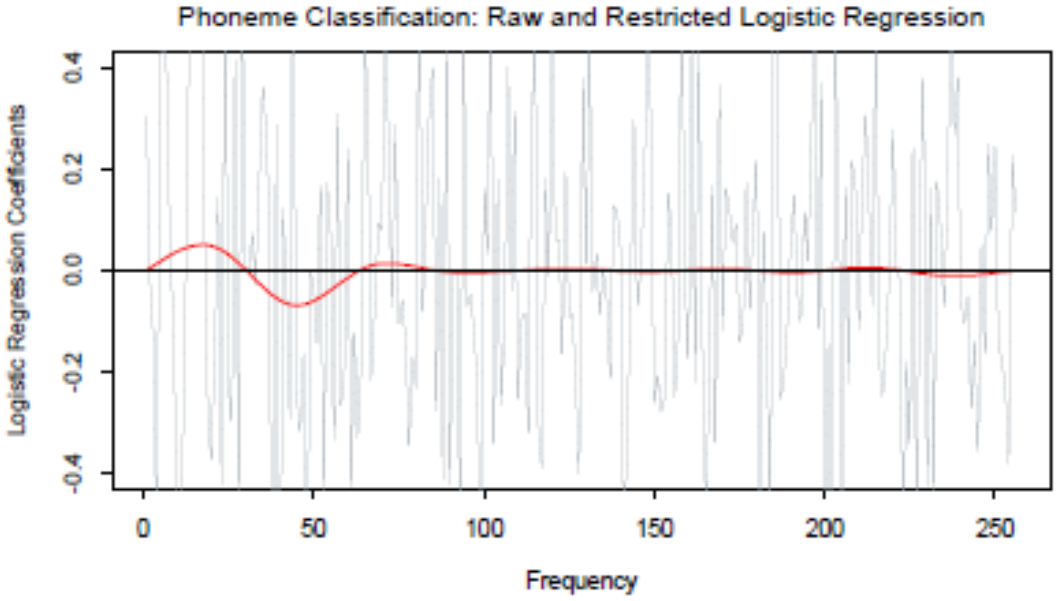
Example - Prostate

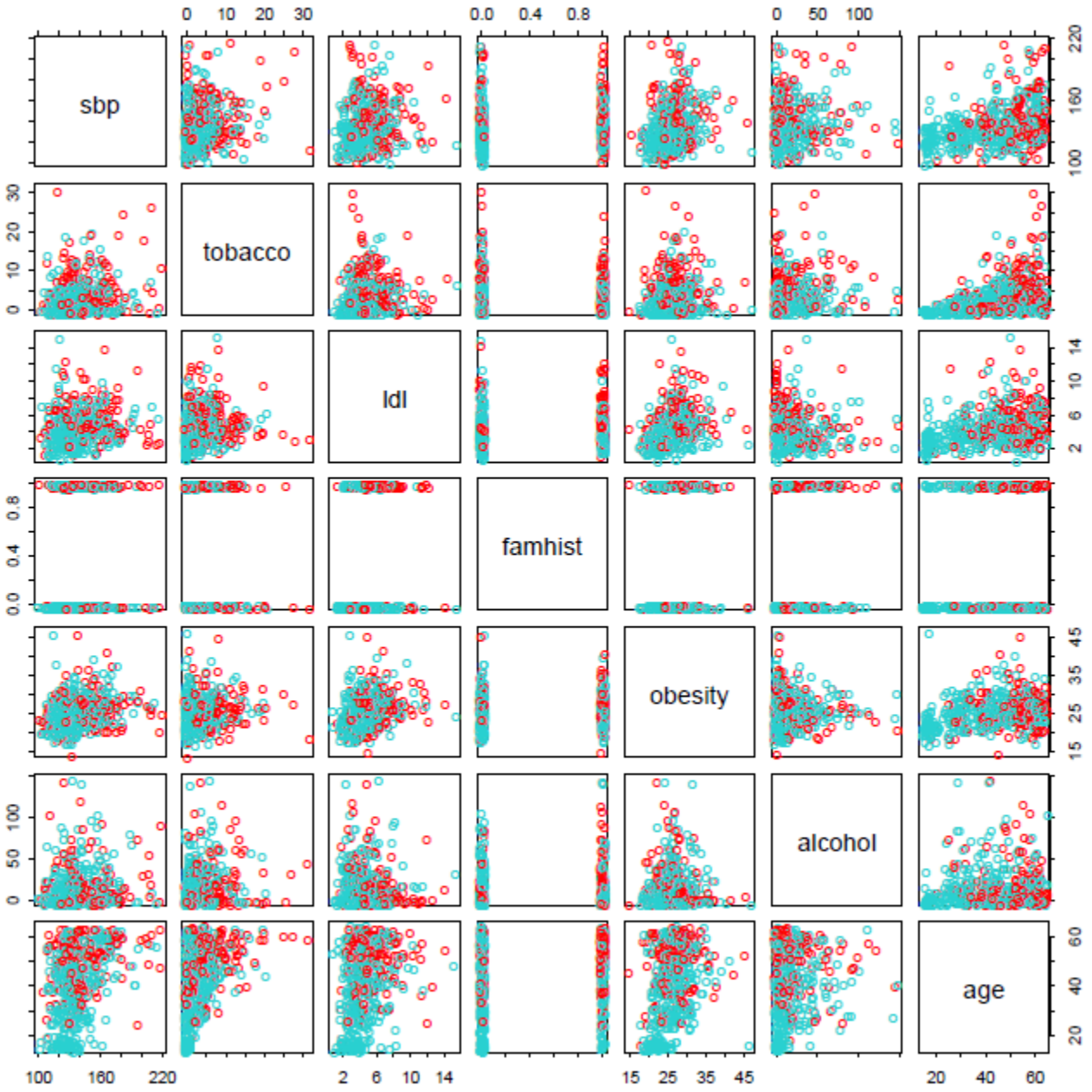
Figure 1: Scatter matrix of prostate cancer data

- Response: Prostate specific antigen (PSA)
- Covariates: clinical measurements:
- Regression problem.


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

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

Example: Spam mail

Data: Information from 4601 email

$$Y = \begin{cases} \text{email} & \text{if true email} \\ \text{spam} & \text{if junk} \end{cases}$$

$X =$frequencies of 57 most common words

Classification problem
Table 1.1: X_i's with largest differences in frequencies
Possible rule:

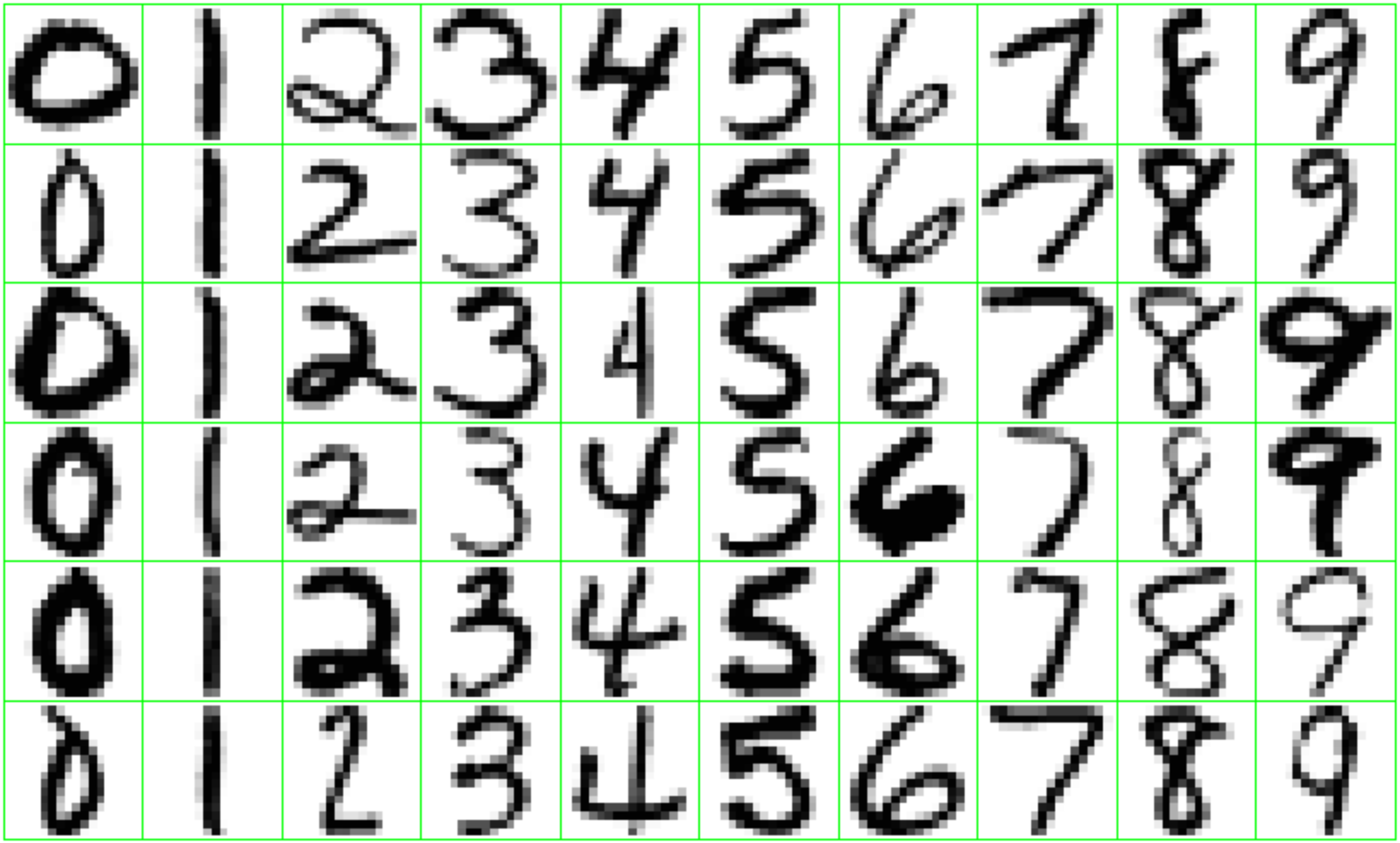
$$\hat{y} = \begin{cases} \text{spam} & \text{if (\%george} < 0.6) \text{ \& (\%you} > 1.5) \\ \text{email} & \text{otherwise} \end{cases}$$

Note: More important that no true emails are discarded than some extra spam!

Example - Zip code

Figure 1: Digitized numbers

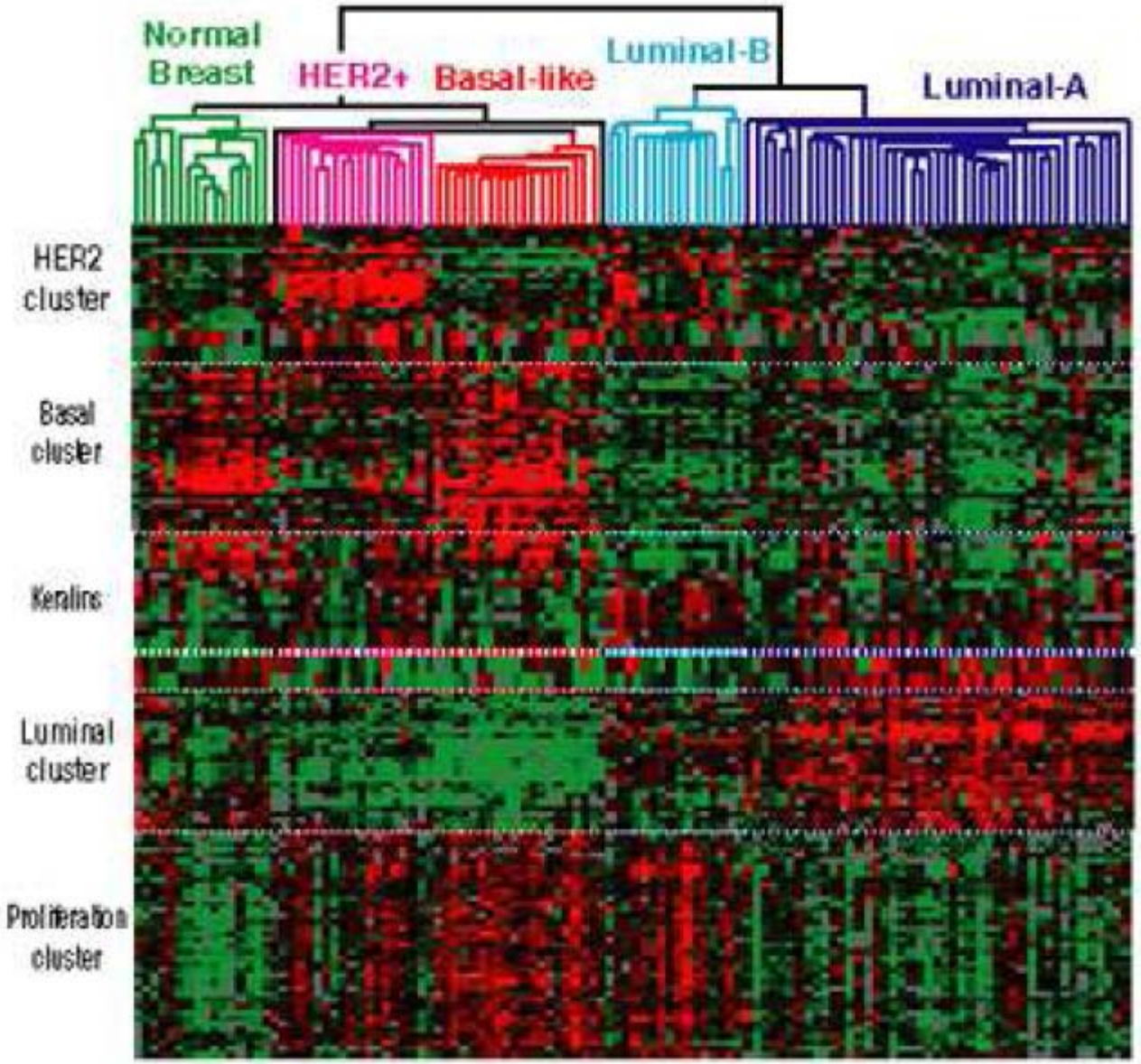
- Problem: Predict numbers
- Classification problem.
- Numbers are categories.
- Note: Ordering between numbers not of interest.



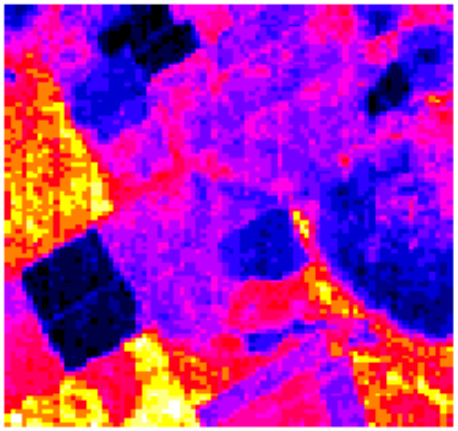
Example - Microarrays

- DNA microarrays measure the expression of a gene in a cell by measuring the amount of mRNA present for that gene.
- Measured in tousands of genes
- Micromatrix:
  - Rows: Individual cells (several tousands)
  - Columns: Individuals (tens)
- Figure 1.3: Example
- Green: Negative Red: positive.
- Different types of problems
  - Identify similar genes (clustering/unsupervised problem)
  - Predict outcome of new patient (Classification)
  - Identify important genes
- Problem: Many more covariates than individuals

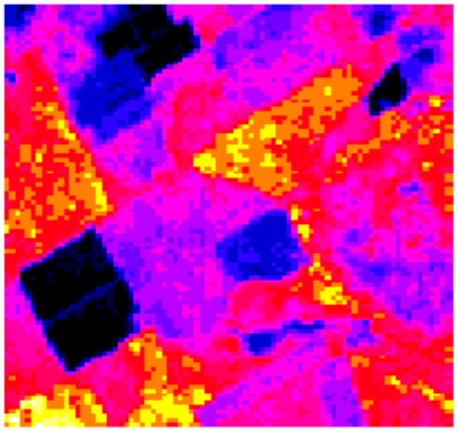
Diversity of Breast Tumor Subtypes



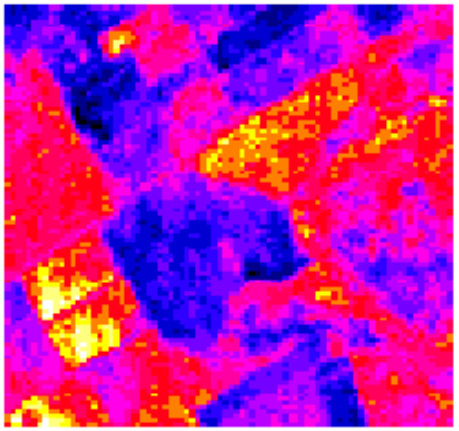
Spectral Band 1



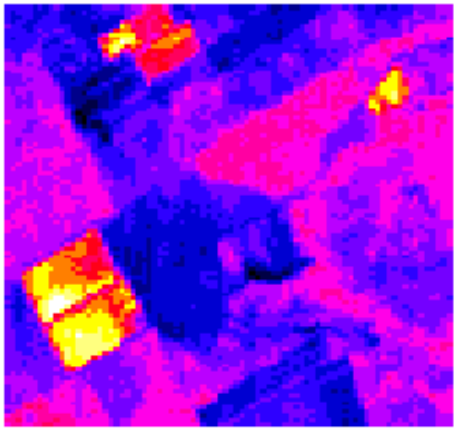
Spectral Band 2



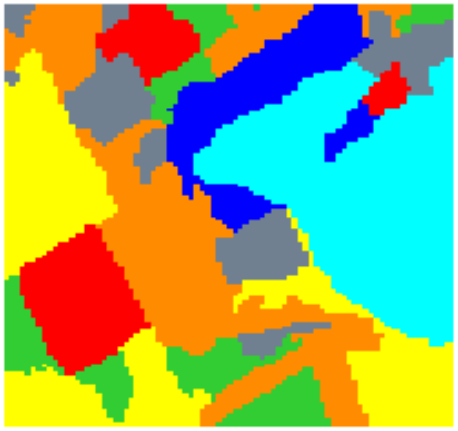
Spectral Band 3



Spectral Band 4



Land Usage



Predicted Land Usage