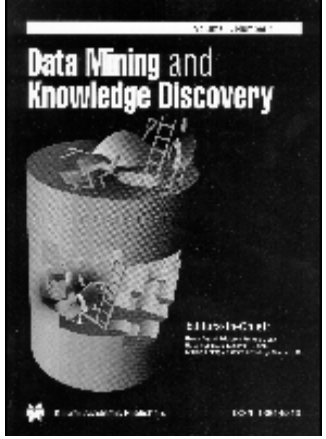
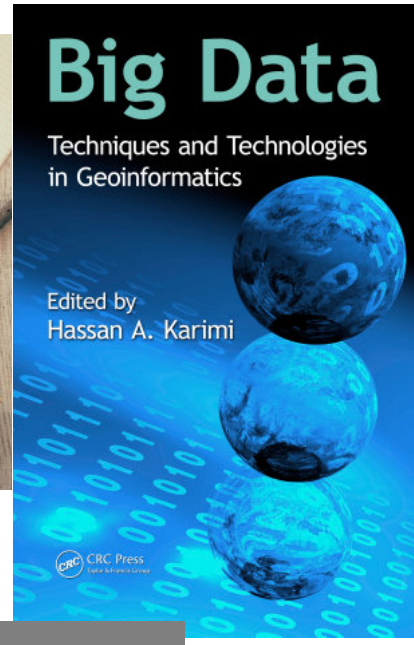
STAT 588 : DATA MINING/MACHINE LEARNING/BIG DATA

Javier Cabrera

Fall 2019





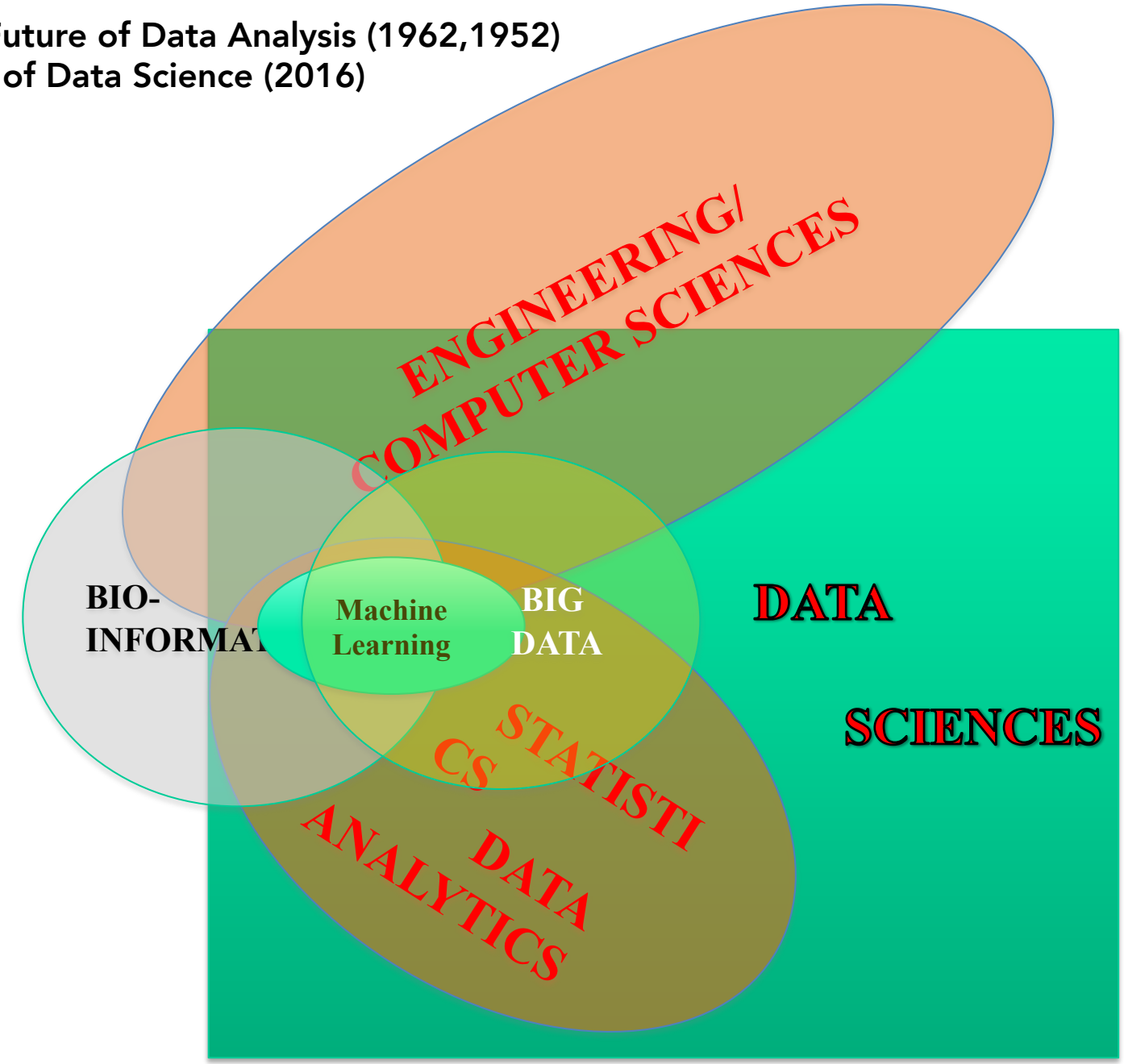
DB2 Intelligent Miner

CRoss Industry Standard Process
for Data Mining

ORACLE DATABASE 10g
Oracle Data Mining

What is Data Science?

John W. Tukey The Future of Data Analysis (1962,1952)
D. Donoho 50 years of Data Science (2016)



Prof. Javier Cabrera's views on Data Sciences(Jan

IBM SPSS Modeler(old Clementine)

Data understanding

- Generate subsets of data automatically from graphs and tables
- Show summary statistics, histograms, and distribution graphics for each data field, and display them in an easy-to-read matrix with the data audit node. This provides you with a comprehensive first look at your data.
- Visually interact with your data
  - Select node or field and view information in a table
  - Create histograms, distributions, line plots, and point plots
  - Display 3-D, panel, and animated graphs
  - Use Web association detection

Modeling

- Prediction and classification
  - Neural networks (multi-layer perceptrons trained using error-back propagation with momentum, radial basis function, and Kohonen network)
  - Decision trees and rule induction [C5.0 and Classification and Regression Trees (C&RT)]
  - Linear regression, logistic regression, and multinomial logistic regression
- Clustering and segmentation
  - Kohonen network, K-means, and TwoStep
    - View summary statistics and distributions for fields between clusters using the Cluster Viewer
- Association detection
  - GRI, apriori, and sequence
- Data reduction
  - Factor analysis and principle components analysis

Stories – Investment Institution

- A study of mailings from the investment institution showed that older people, particularly over 65, do not respond to IRA offers (Individual Retirement Accounts).

The VP who reviewed the work asked why he was paying good money for such obvious discoveries.

The consultant replied that it is the VP's institute that is sending these offers...

Stories – Non-actionable Segment

- A bank discovered a cluster of customers that have left the bank:
  - Older than the average customer.
  - Less likely to have a mortgage.
  - Less likely to have a credit card.

They were also...

Stories – Cash Management

- A large bank asked to find the factors affecting churn of companies for which they manage cash. Specifically, to characterize which companies are likely to leave.

Among the strongest factors was...

If the account relationship manager is called <name>, over 50% of the clients leave!

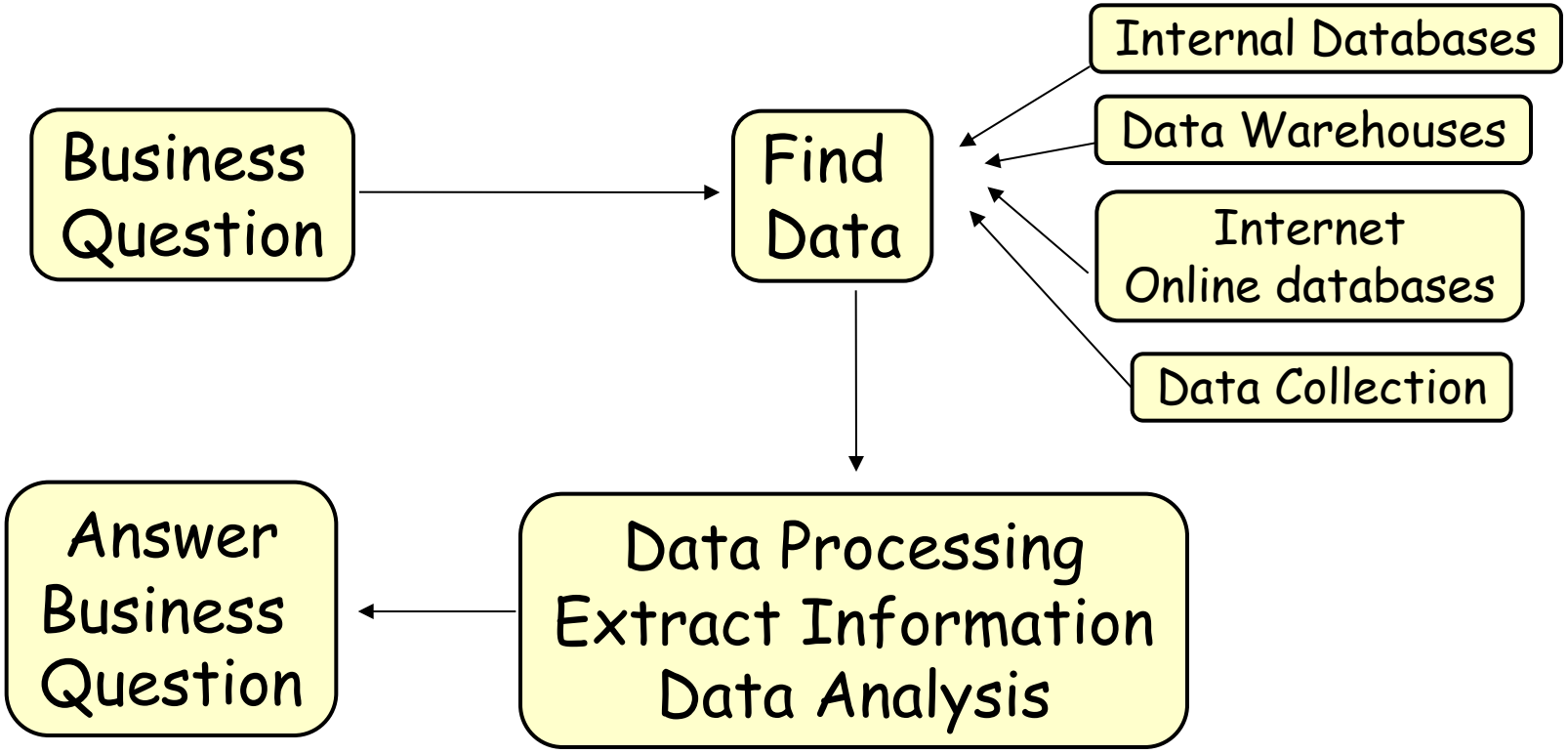
What is Data mining?

Finding interesting structure in data

- *Structure:* refers to statistical patterns, predictive models, hidden relationships

- Examples of tasks addressed by Data Mining
  - Predictive Modeling (classification, regression)
  - Segmentation (Data Clustering)
  - Summarization
  - Visualization

What is Data mining?



Welcome to Data mining

Data collected in large databases:

- Relational databases, Internet, Data Warehouses:
  Large Datasets, Many variables and cases.
- Mostly noisy data: Missing Values, Zeros, Outliers.
- No random samples.

Data mining objective: To extract valuable information.
To identify nuggets, small clusters of observations in these data that contain potentially valuable information.

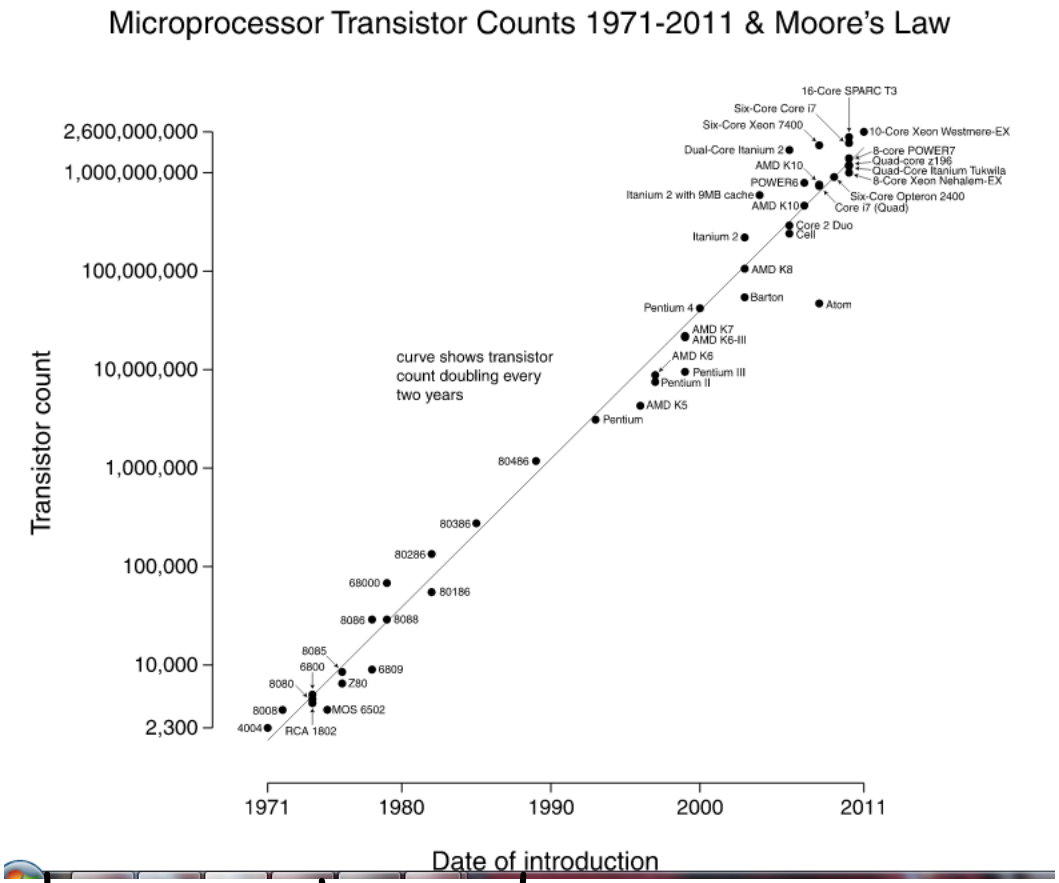
The definition of valuable is generally reflected by a large response value of a specific category of a qualitative response.

Sifting through a large volume of data that is noisy, badly behaved, and that may have many missing values, or that may just be irrelevant is the main challenge of data mining.

Moore's law:

+

- Processing "capacity" doubles every 18 months (Exponential)
- Hard Disk storage "capacity" doubles every 12 months (Use to be every 36 months)



−

- Bottle necks are not speed anymore.
- Processing capacity is not growing as fast as data acquisition.

How large is large?

By number of cases:

- Small: N < 30 (No CLT)
- Moderate: 30 < N < 1000 (CLT)
- Moderately large: 1000 < N < 100000 (tolerable N^2)
- Large: 100000+: No N^2 computations.

By the number of variables:

- Small: One variable.
- Moderate: Less than 1000 Variables. Matrix inversion.
- Large: More than 1000 Variables.

By database size:

- Large: Does not fit in memory.
- Big Data: Does not fit in one disk

Data mining Software

Fast computations.

- Economic use of memory.
- Flexible (and user friendly) Graphics Interface.

Software that will used in class:

- R
- winBUGS Weka, Ggobi

Other Software:

- Enterprise Miner, Spottfire, C5, Clementine, SAS

Methods and Techniques

- Data summarization, EDA, Basic Statistics.
- Advanced Data Visualization.
  - Data Reduction: variable and case Subsetting, Sampling.
  - Dimension Reduction: Principal Components, Covariance.
  - Cluster analysis (Segmentation): k-means, hierarchical.
  - Classification techniques (Pattern recognition):
    - LDA, QDA
    - Trees, Random Forest, Boosting
    - Neural nets, Support Vector Machines,
    - Nearest Neighbors.
  - Model based methods: Linear, Non-Linear, logistic, lasso

Data Mining Algorithms

A data mining algorithm is a well-defined procedure that takes data as input and produces output in the form of models or patterns

"well-defined": can be encoded in software

"algorithm": must terminate after some finite number of steps

Algorithm Components

1. The *task* the algorithm is used to address (e.g. classification, clustering, etc.)

2. The *structure* of the model or pattern we are fitting to the data (e.g. a linear regression model)

3. The *score function* used to judge the quality of the fitted models or patterns (e.g. accuracy, BIC, etc.)

4. The *search or optimization method* used to search over parameters and/or structures (e.g. steepest descent, MCMC, etc.)

5. The *data management technique* used for storing, indexing, and retrieving data (critical when data too large to reside in memory)

	CART	Backpropagation	A Priori
Task	Classification and Regression	Regression	Rule Pattern Discovery
Structure	Decision Tree	Neural Network (Nonlinear functions)	Association Rules
Score Function	Cross-validated Loss Function	Squared Error	Support/Accuracy
Search Method	Greedy	Gradient Descent	Breadth-First with Pruning
Data Management Technique	Unspecified	Unspecified	Linear Scans

What is new?

- Improved Methodology and Software.

- Solve business problems:
  Data is from regular businesses.

Objective: Better business decisions.

The Role of visualization

Data visualization methods are attractive tools to use for analyzing such datasets for several reasons:

- Data visualization methods show many features (expected and unexpected) of a dataset at once and, as such, are well equipped to pick up subtle structures of interest and anomalies as well as clear patterns.
- They allow (in fact, encourage) flexible interaction with the data.
- They can be more readily understood by non-statisticians (although their properties may not be).
- Good user-friendly graphics software is becoming more readily available.

Data visualization methods

Large datasets create visualization challenges.

- Scatterplots: Large numbers of points may hide the underlying structure.

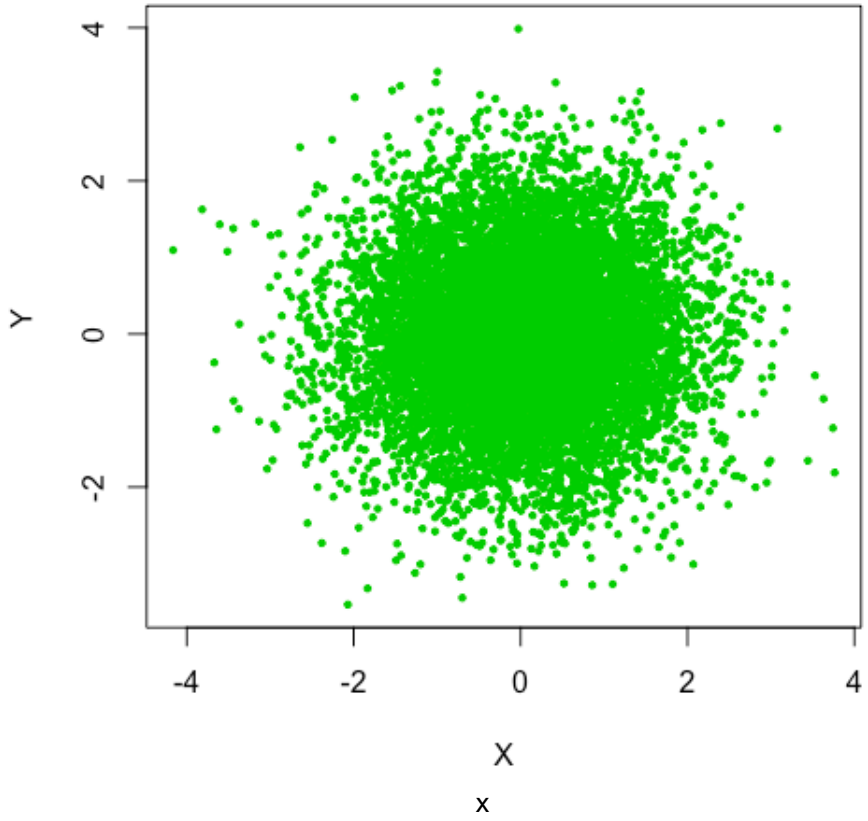
- Apply *Data Binning* and use an image graph.
- Avoid *Masking* by duplicating plots and highlighting subgroups.
- Sometimes is enough to graph a subset selected at random.

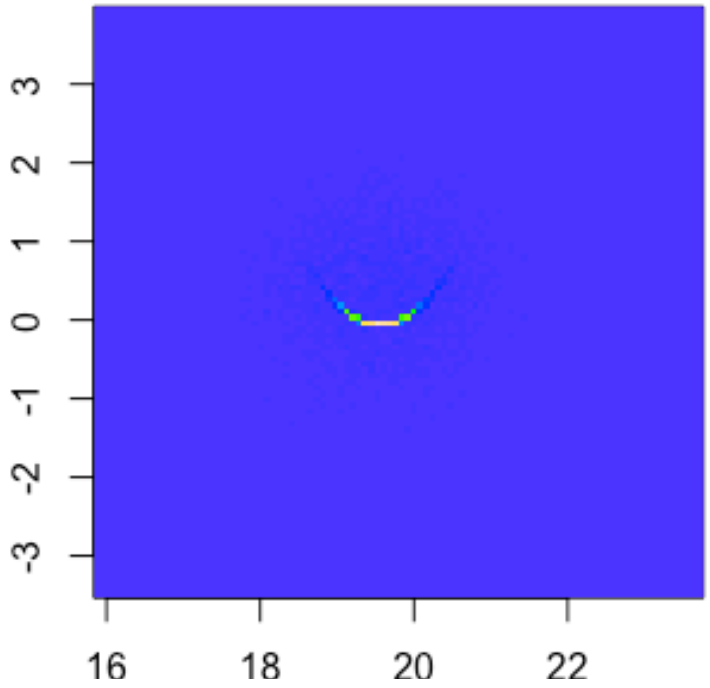
Many variables at once. There are many ingenious tools for this.

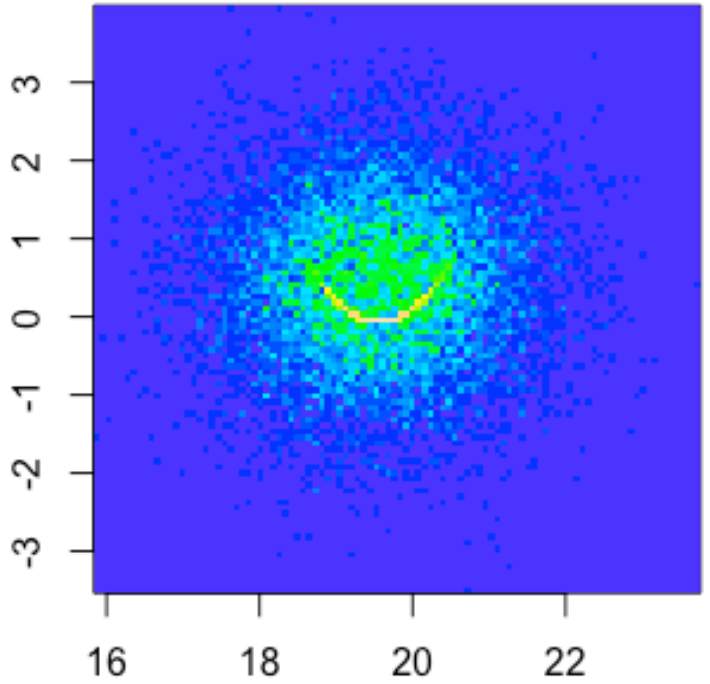
- Scatterplot matrix
  - all variables
  - all descriptor variables with color coding according to one response
  - all response variables with color coding according to one descriptor
- plot selected 2D views to highlight some feature of the data:
  - principal components analysis (spread)
  - projection pursuit (clustering)]
- look at "all" 2D views of the data via a dynamic display
  [rotating 3D display, grand tour]
- conditional plots
- multiple windows with brush and link

Data Binning

Scatter Plot







R example of Binning Plot

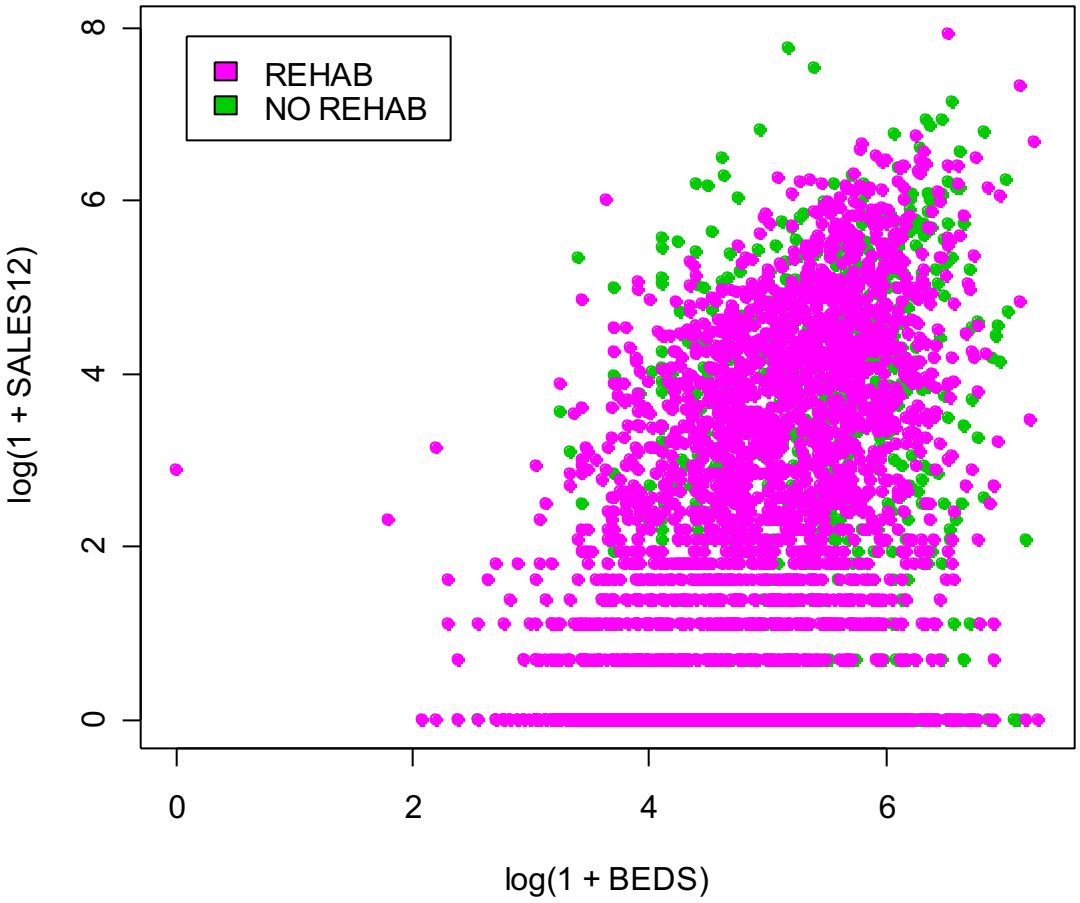
```
binplot <- function(x,y,nr=20,nc=20, scale="raw") {
   zx = c(1:nr,rep(1,nc),1+trunc( nr*(x- min(x))/(max(x)-min(x)) ))
   zx[zx>nr] = nr
   zy = c(rep(1,nr),1:nc,1+trunc( nc*(y- min(y))/(max(y)-min(y)) ))
   zy[zy>nc] = nc
   z = table(zx,zy); z[,1]=z[,1]-1; z[1,]=z[1,]-1;
   if (scale=="l") z= log(1+z)
   image(z=t(z),x=seq(length=nr+1,from=min(x),to=max(x)),
        y= seq(length=nc+1,from=min(y),to=max(y)),
        xlab="",ylab="", col=topo.colors(100))
}
# Run this code line by line
x = rnorm(10000) ; y = rnorm(10000)
plot(x,y)
binplot(x,y,10,10)
binplot(x,y,50,50)
binplot(x,y,100,100)
binplot(x,y,100,100,'l')
binplot(x,y,500,500,'l')
ux = rnorm(5000)/3
uy = ux^2 -0.5
par(mar=c(4,4,1,1))
plot(x=c(y,uy),y=c(x,ux),pch=20,col=3,xlab="X",ylab="Y",cex=0.7)
binplot(c(y,uy)+20,c(x,ux),100,100)
binplot(c(y,uy)+20,c(x,ux),100,100,"l")
```

Case Study: SALES OF ORTHOPEDIC EQUIPMENT

The objective of this study is to find ways to increase sales of orthopedic material from our company to hospitals in the United States.

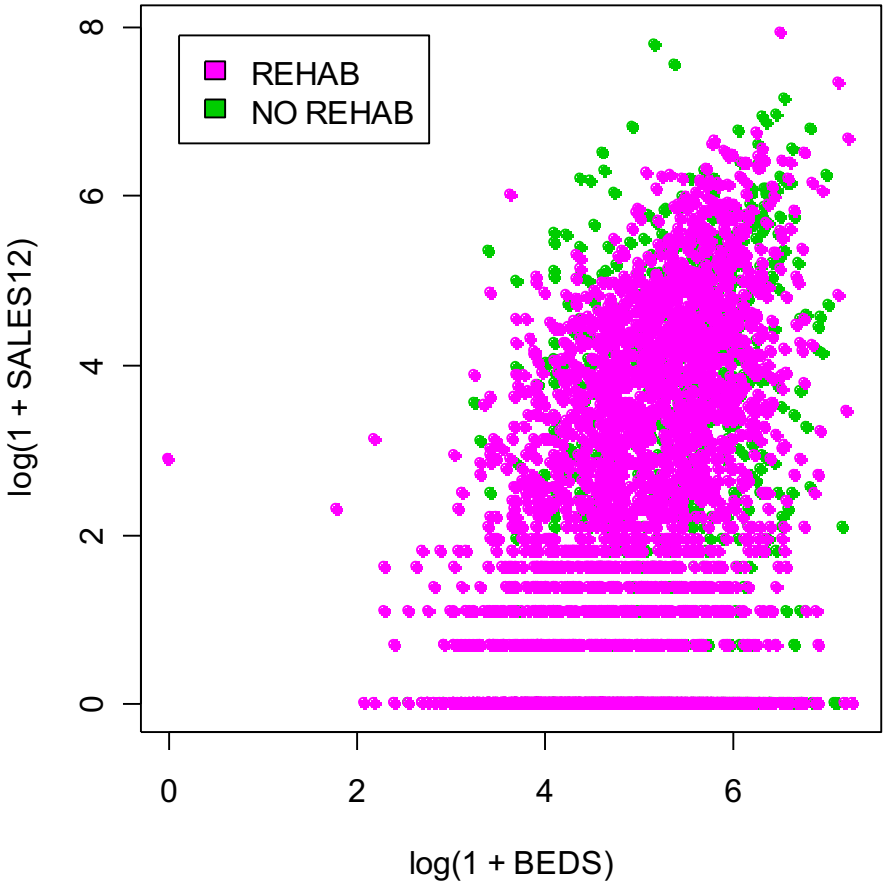
VARIABLES:
BEDS : NUMBER OF HOSPITAL BEDS
RBEDS : NUMBER OF REHAB BEDS
OUT-V : NUMBER OF OUTPATIENT VISITS
ADM : ADMINISTRATIVE COST(In $1000's per year)
SIR : REVENUE FROM INPATIENT
SALES12 : SALES OF REHAB. EQUIP. FOR THE LAST 12 MO
HIP95 : NUMBER OF HIP OPERATIONS FOR 1995
KNEE95 : NUMBER OF KNEE OPERATIONS FOR 1995
TH : TEACHING HOSPITAL? 0, 1
TRAUMA : DO THEY HAVE A TRAUMA UNIT? 0, 1
REHAB : DO THEY HAVE A REHAB UNIT? 0, 1
HIP96 : NUMBER HIP OPERATIONS FOR 1996
KNEE96 : NUMBER KNEE OPERATIONS FOR 1996
FEMUR96 : NUMBER FEMUR OPERATIONS FOR 1996

Using Color

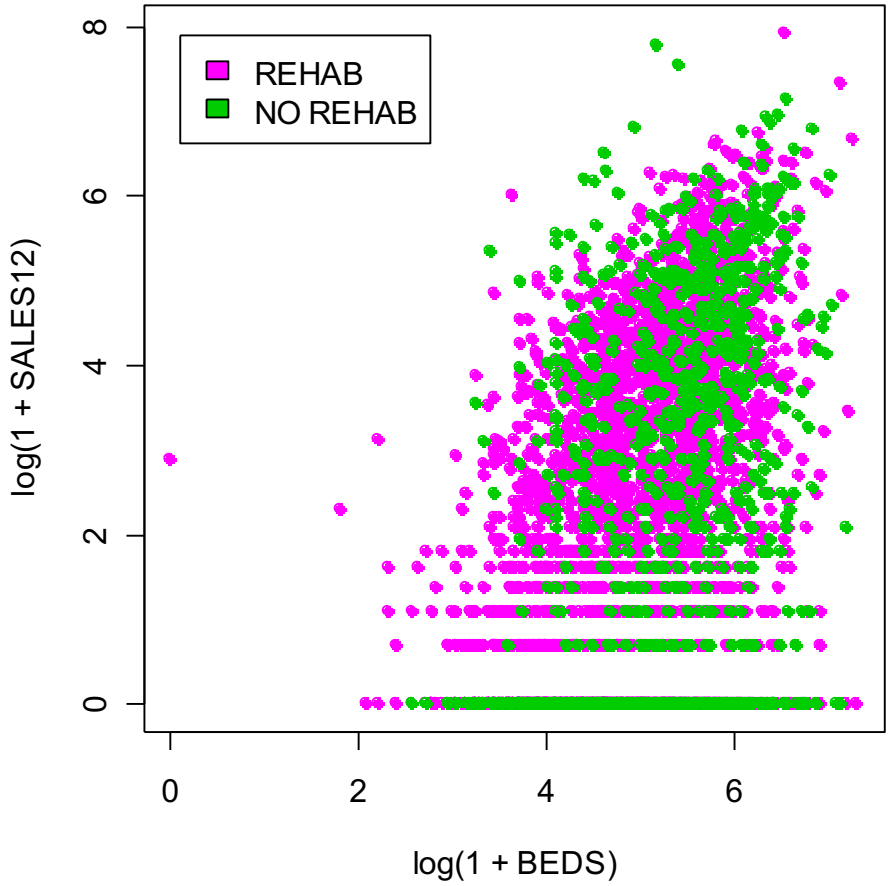


Masking effect

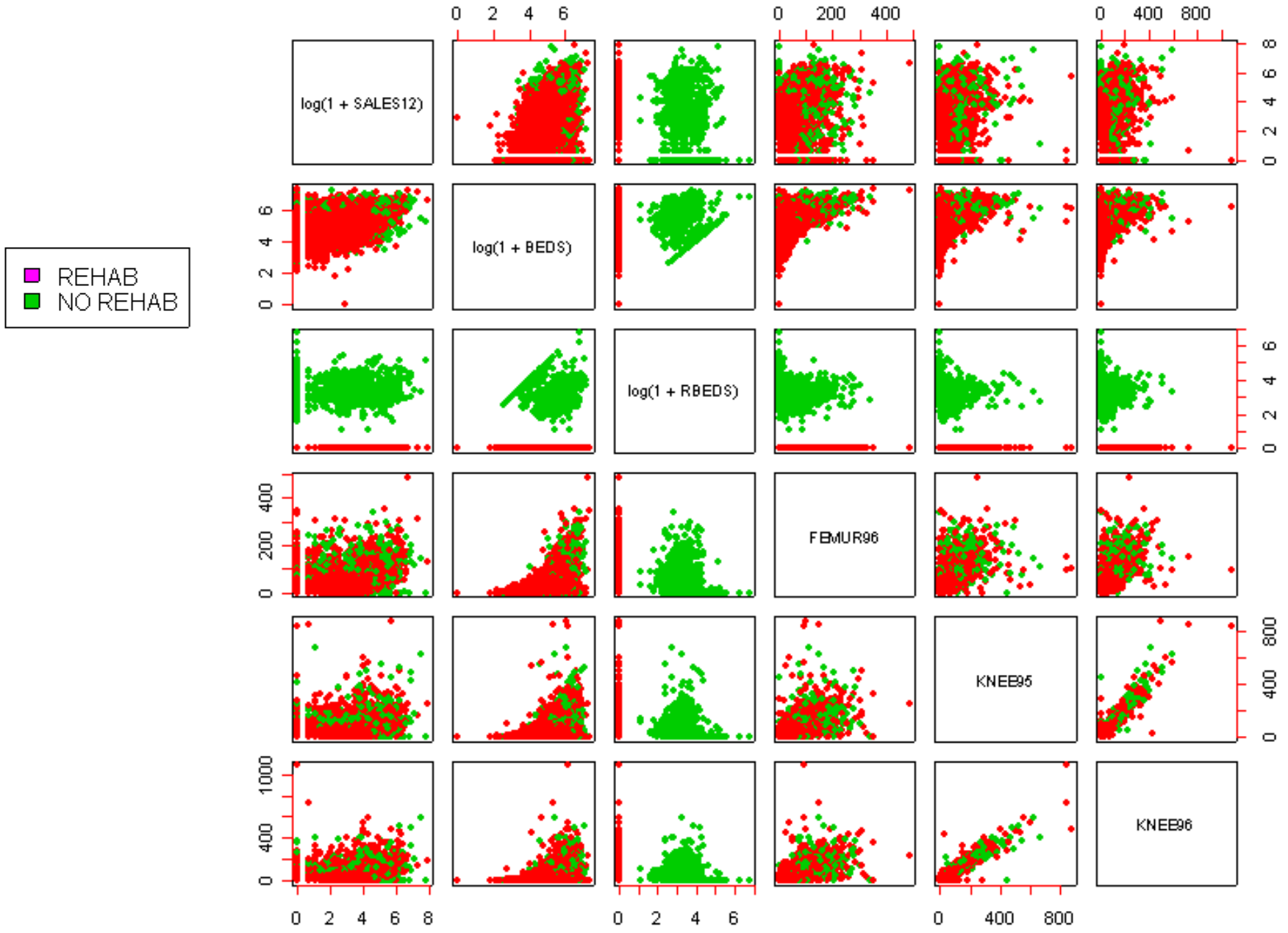
Drawing green dots first



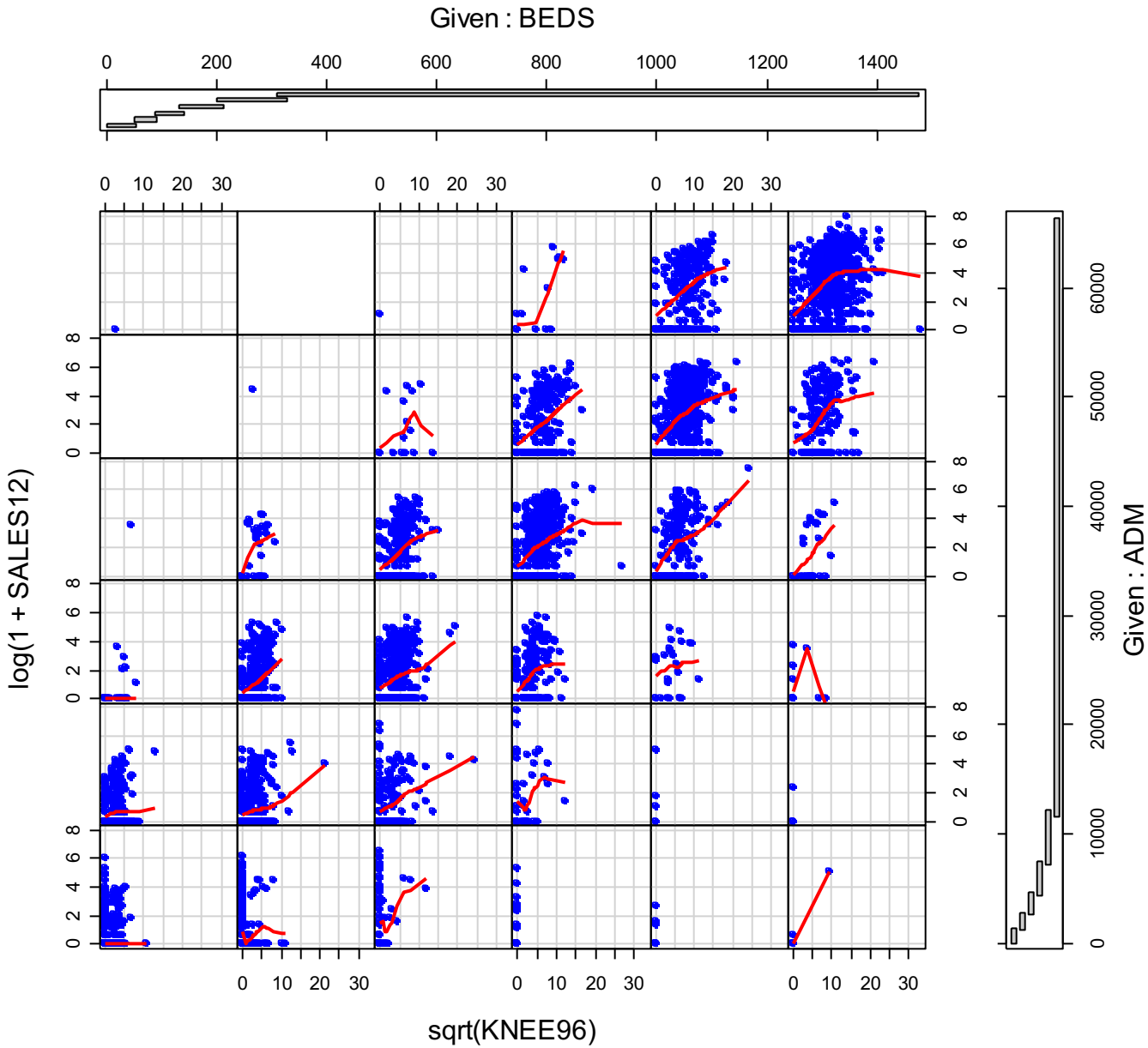
Drawing purple dots first



Pairwise Scatter Plot



Conditional Plot example



Feature recognition/prediction methods

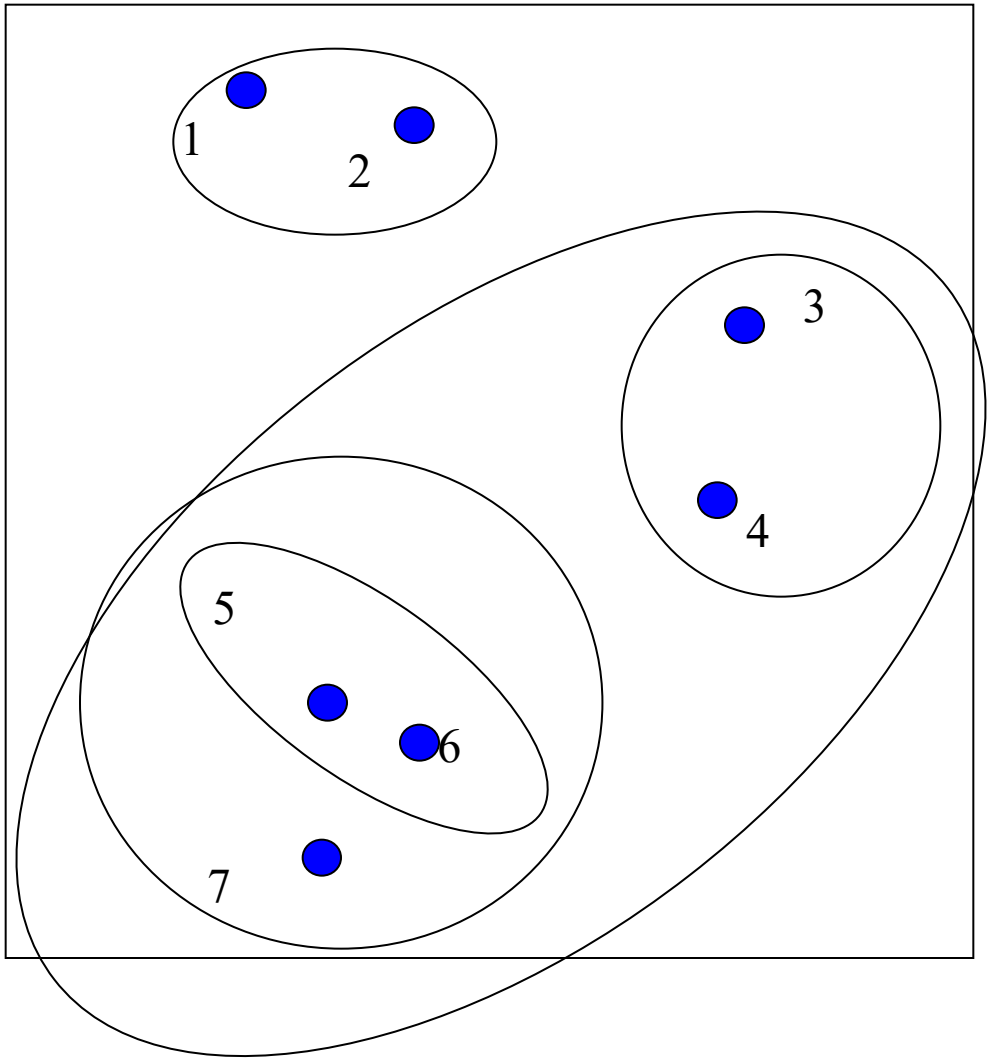
- variable and case selection

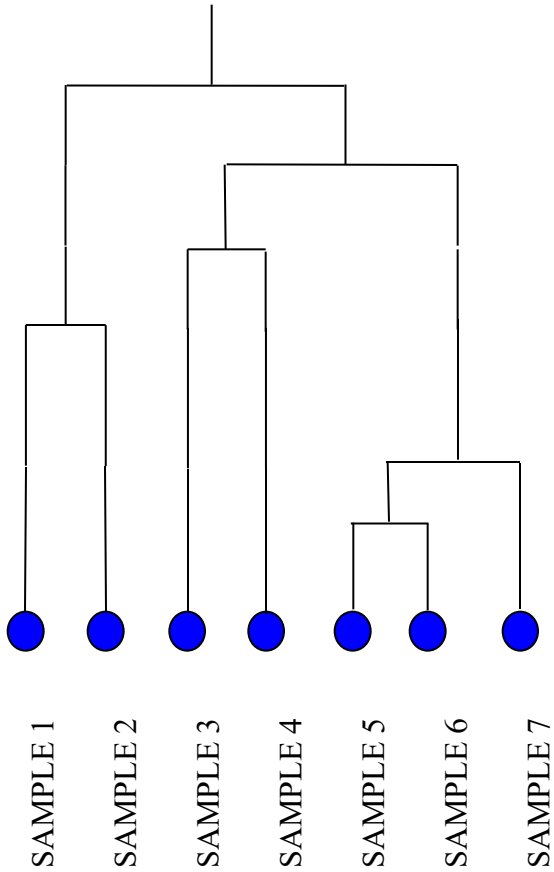
- cluster analysis (unsupervised pattern recognition)
- partitioning methods (e.g., *k*-means, *k*-medioids)
- hierarchical methods (e.g., agglomerative nesting)
- two-way clustering and biclustering

- classification (supervised pattern recognition, discriminant analysis)
- trees (e.g., CART, C5, Firm, Tree, ARF)
- model-based methods (e.g., logistic regression)
- SVM, boosting
- artificial neural networks

role of robust methods / diagnostics

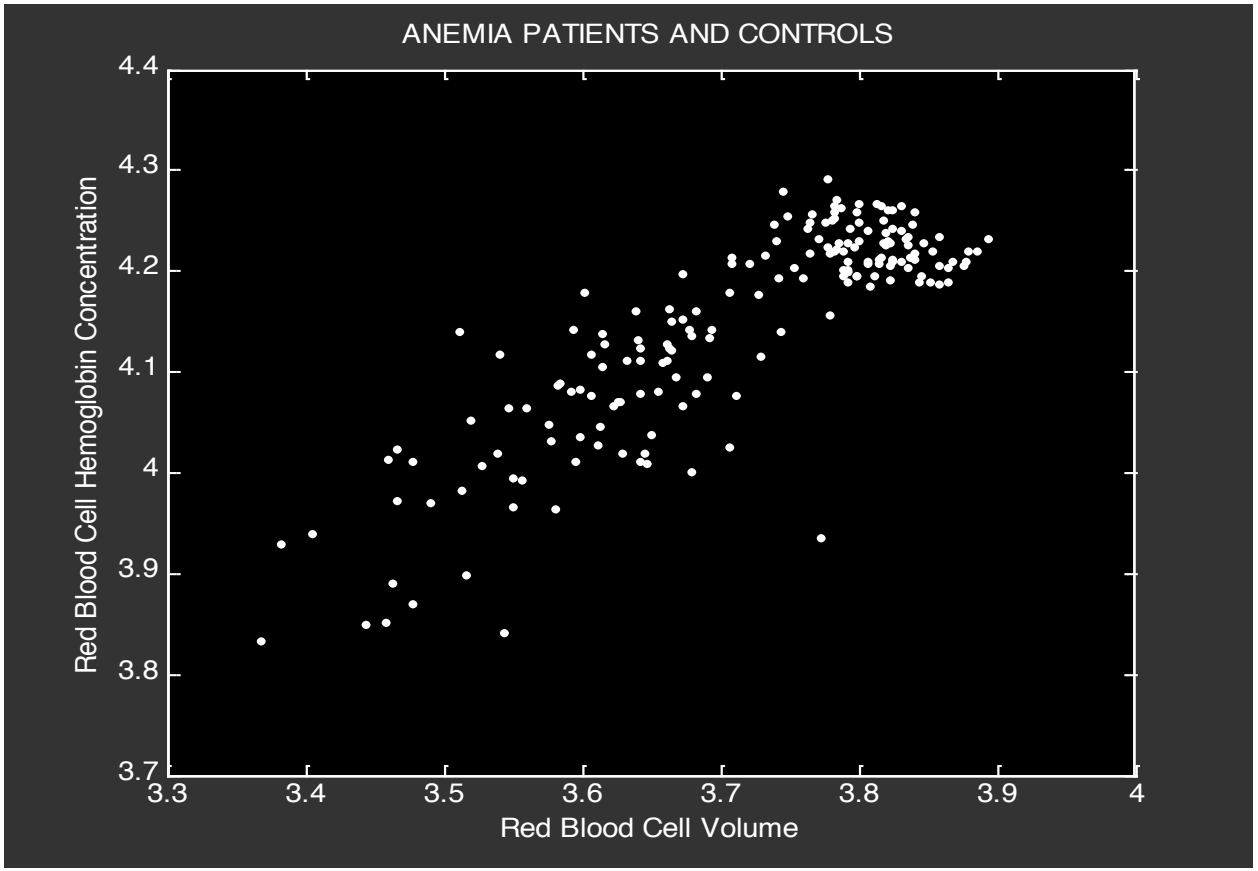
Hierarchical Cluster Example



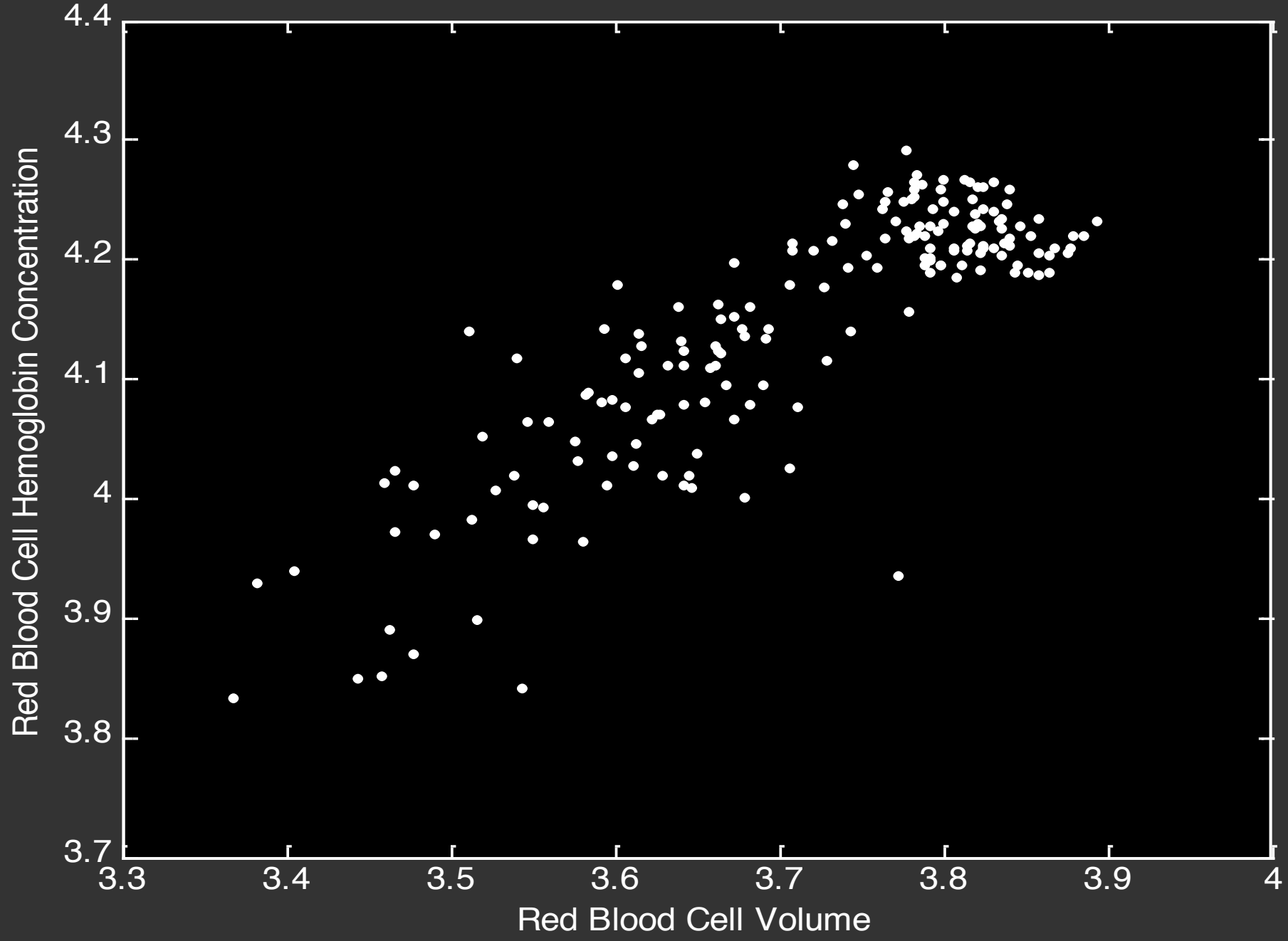


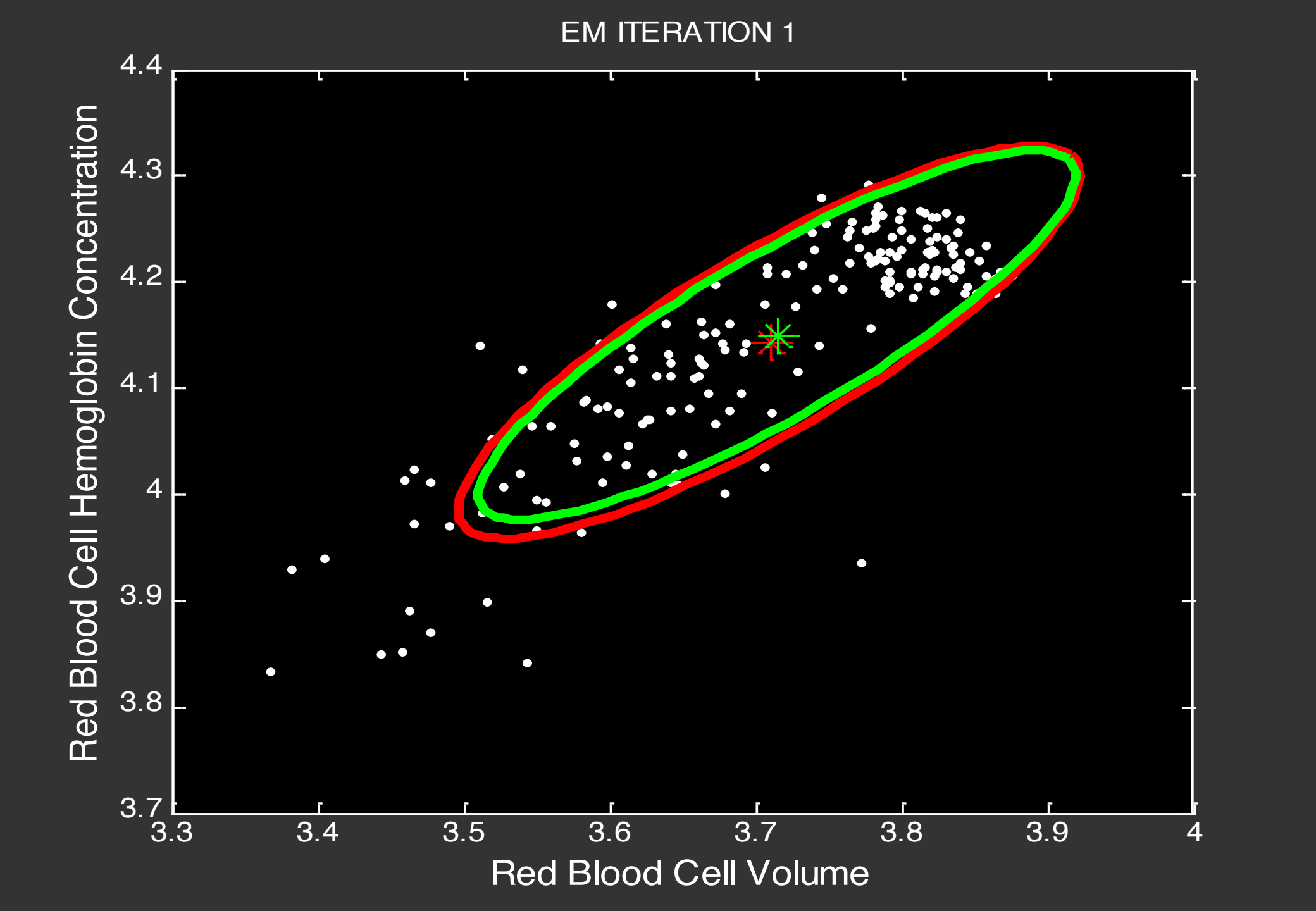
Model-based Clustering

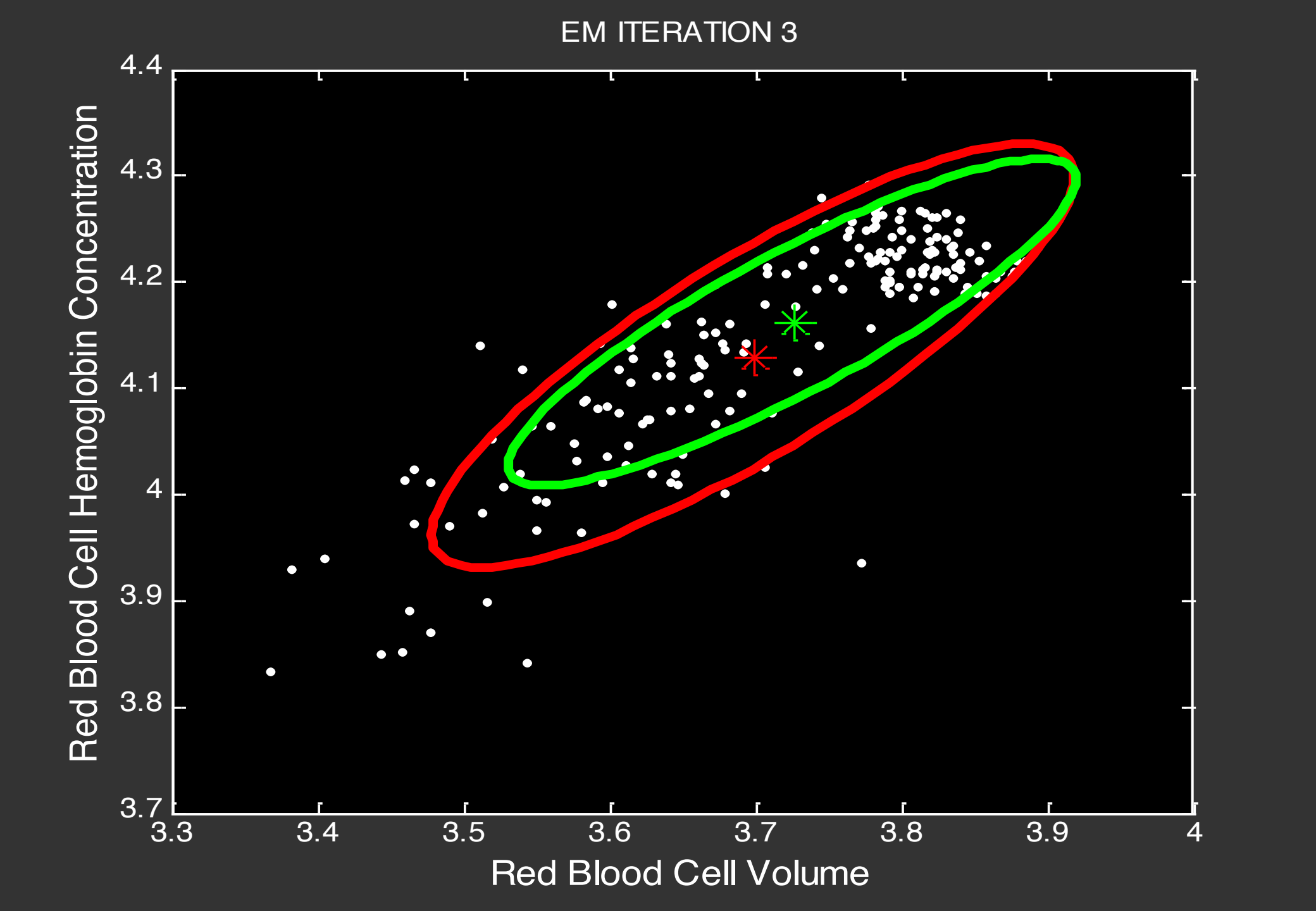
$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x;\theta_k)$$

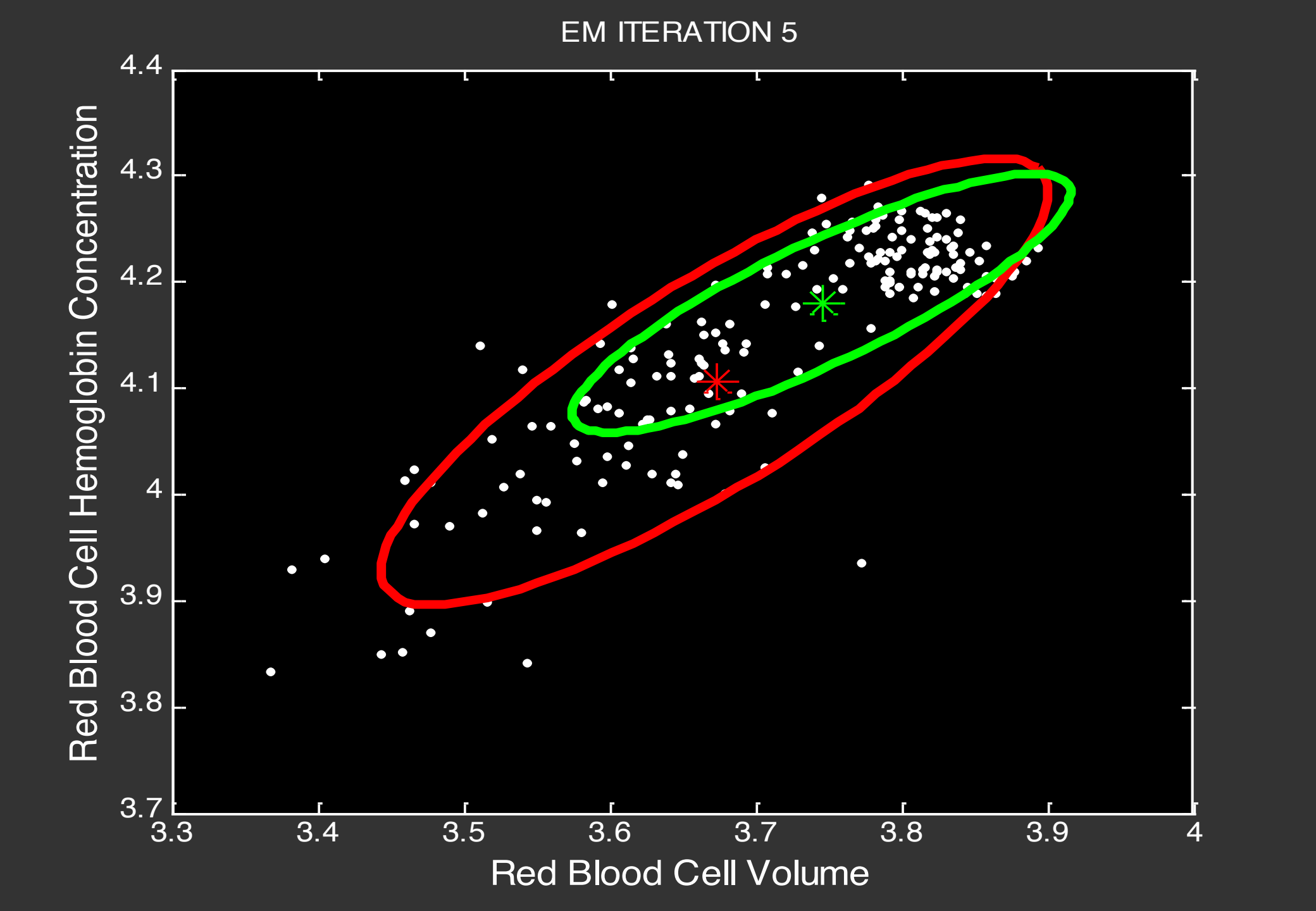


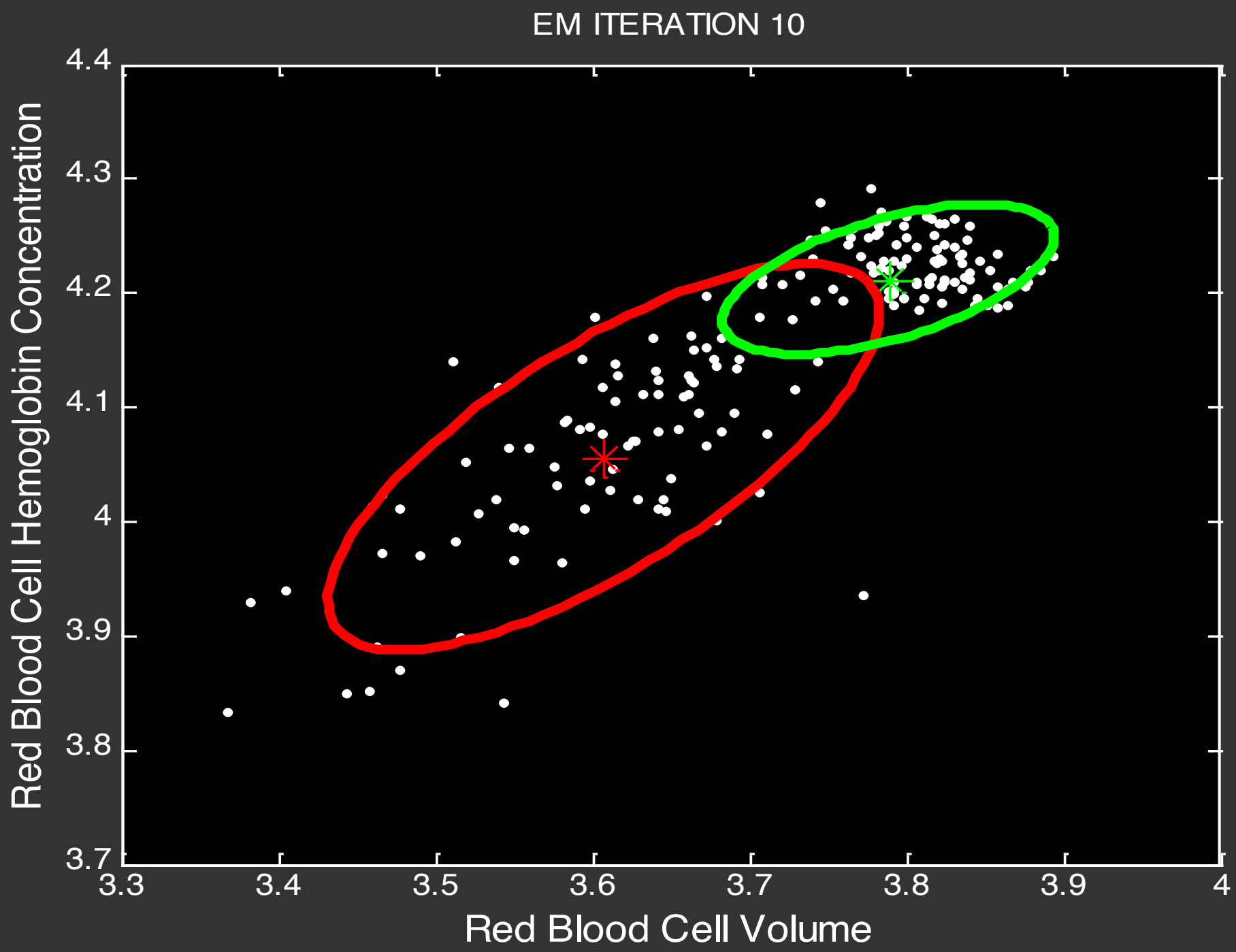
Padhraic Smyth, UCI

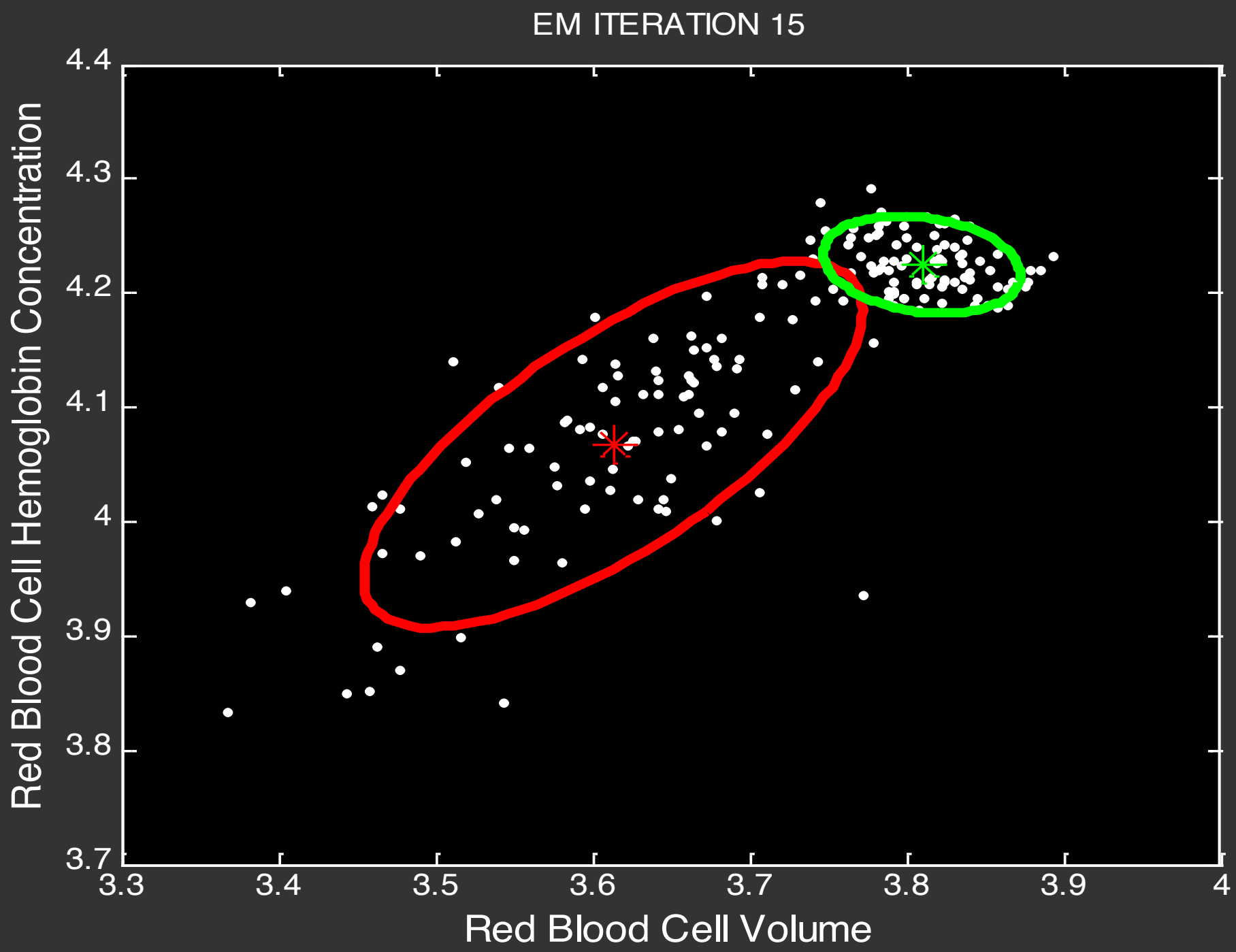

ANEMIA PATIENTS AND CONTROLS
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4

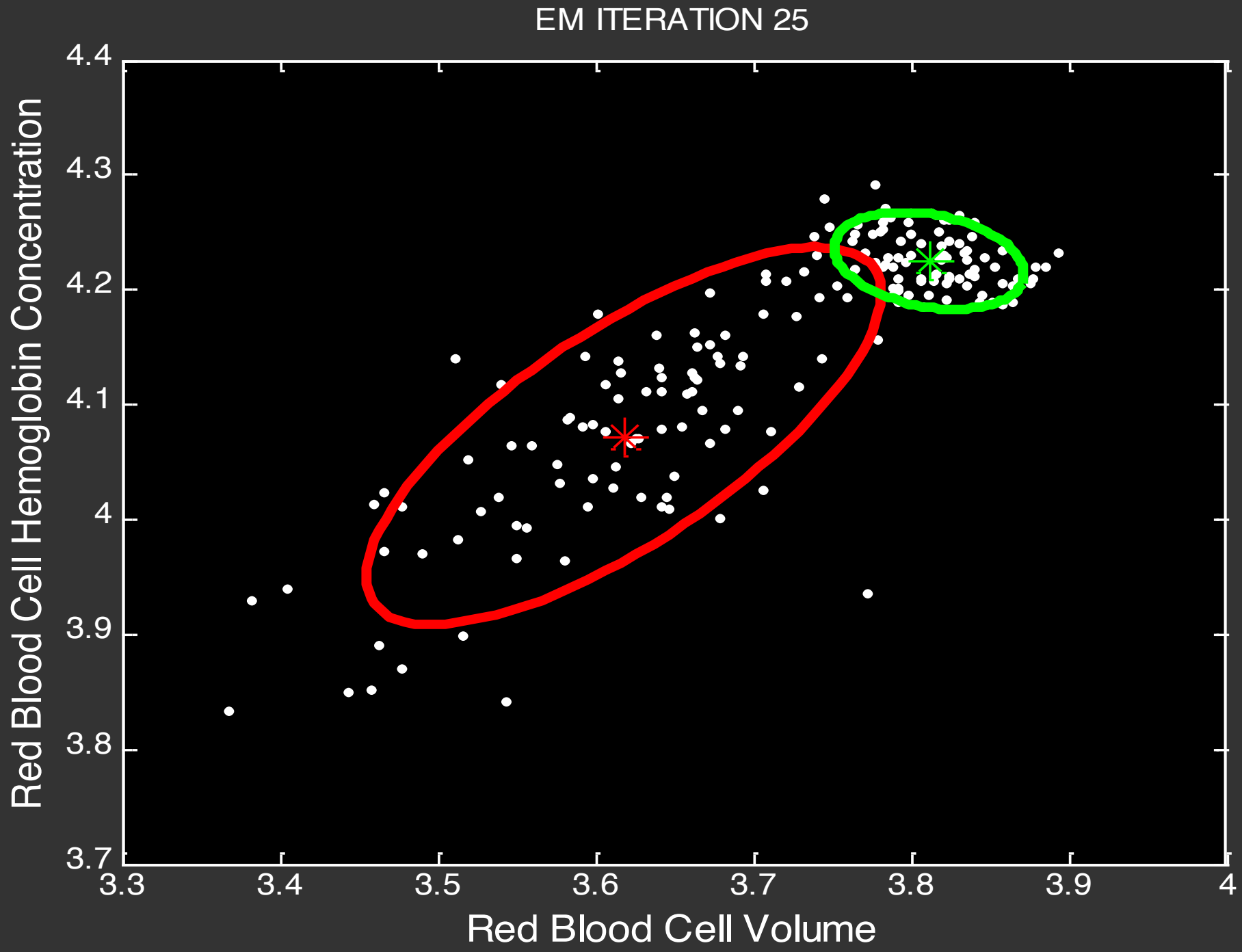

EM ITERATION 1
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4


EM ITERATION 3
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4


EM ITERATION 5
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4


EM ITERATION 10
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4


EM ITERATION 15
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4

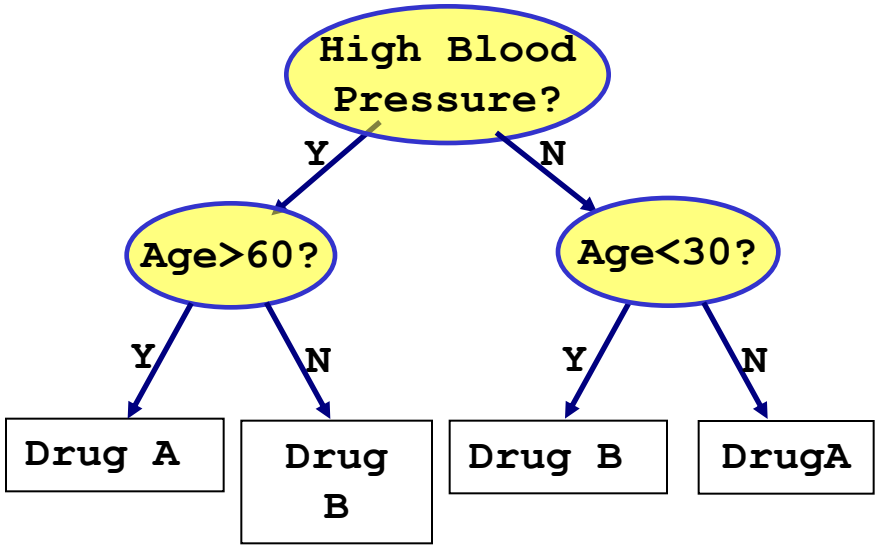

EM ITERATION 25
Red Blood Cell Hemoglobin Concentration
Red Blood Cell Volume
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4

Recursive Partition (CART)

Partition the space into regions of similar response

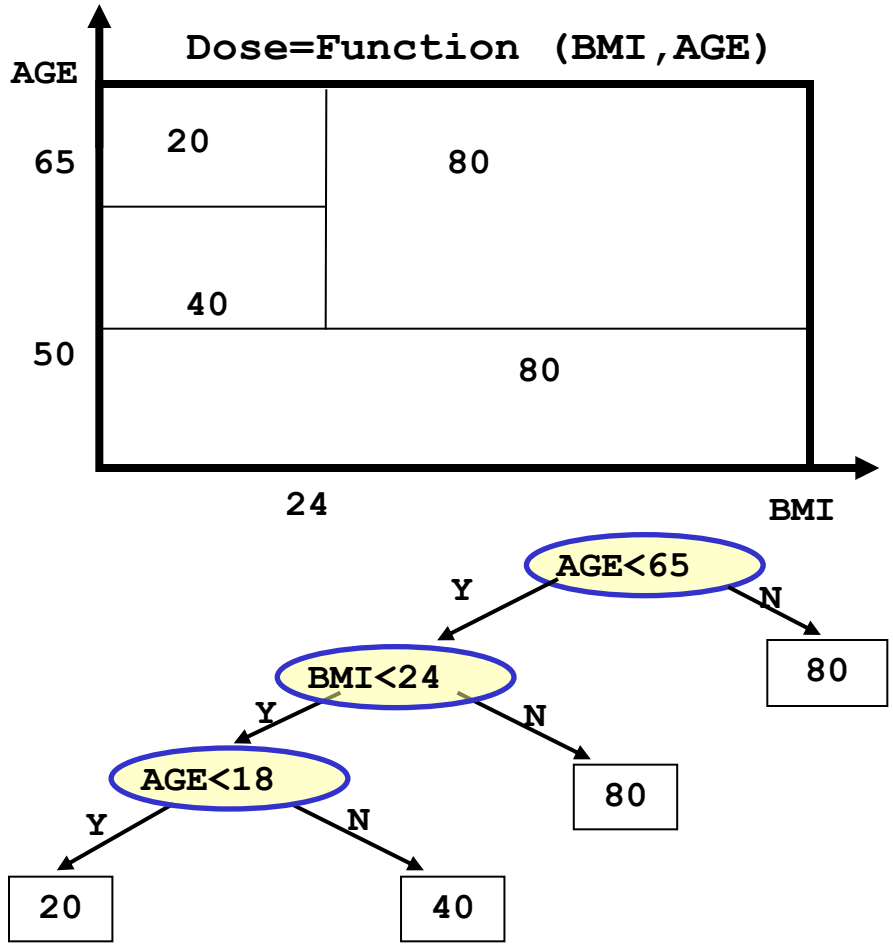
I. Dependent variable is categorical

•Classification Trees, Decision Trees
Example: A doctor might have a rule for choosing which drug to prescribe to high cholesterol patients.



II. Dependent variable is numerical

• Regression Tree



Recursive Partition Criteria

- For regression trees two criteria functions are:

$$\textbf{Equal variances}(CART)\textbf{: h} = \frac{N_L\hat{\sigma}_L^2 + N_R\hat{\sigma}_R^2}{N_L + N_R}$$

$$\textbf{Non equal variances : h} = \frac{N_L \log\hat{\sigma}_L^2 + N_R \log\hat{\sigma}_R^2}{N_L + N_R}$$

- For classification trees: criteria functions

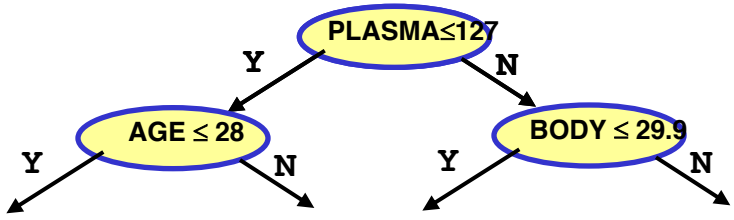
$$\mathbf{h} = p_L \min(p_L^0, p_L^1) + p_R \min(p_R^0, p_R^1)$$

$$\mathbf{h} = p_L(-p_L^0 \log p_L^0 - p_L^1 \log p_L^1) + p_R(-p_R^0 \log p_R^0 - p_R^1 \log p_R^1) \text{ (C5)}$$

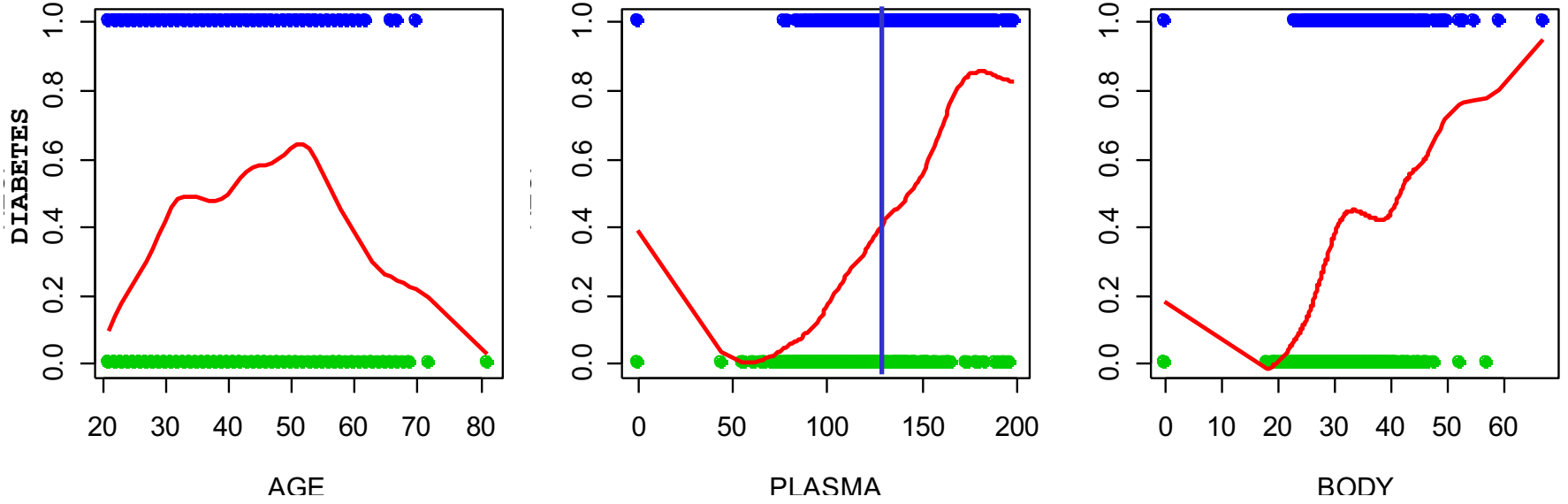
$$\mathbf{h} = p_L p_L^0 p_L^1 + p_R p_R^0 p_R^1 \text{ (CART)}$$

Classic Example of CART
Pima Indians Diabetes

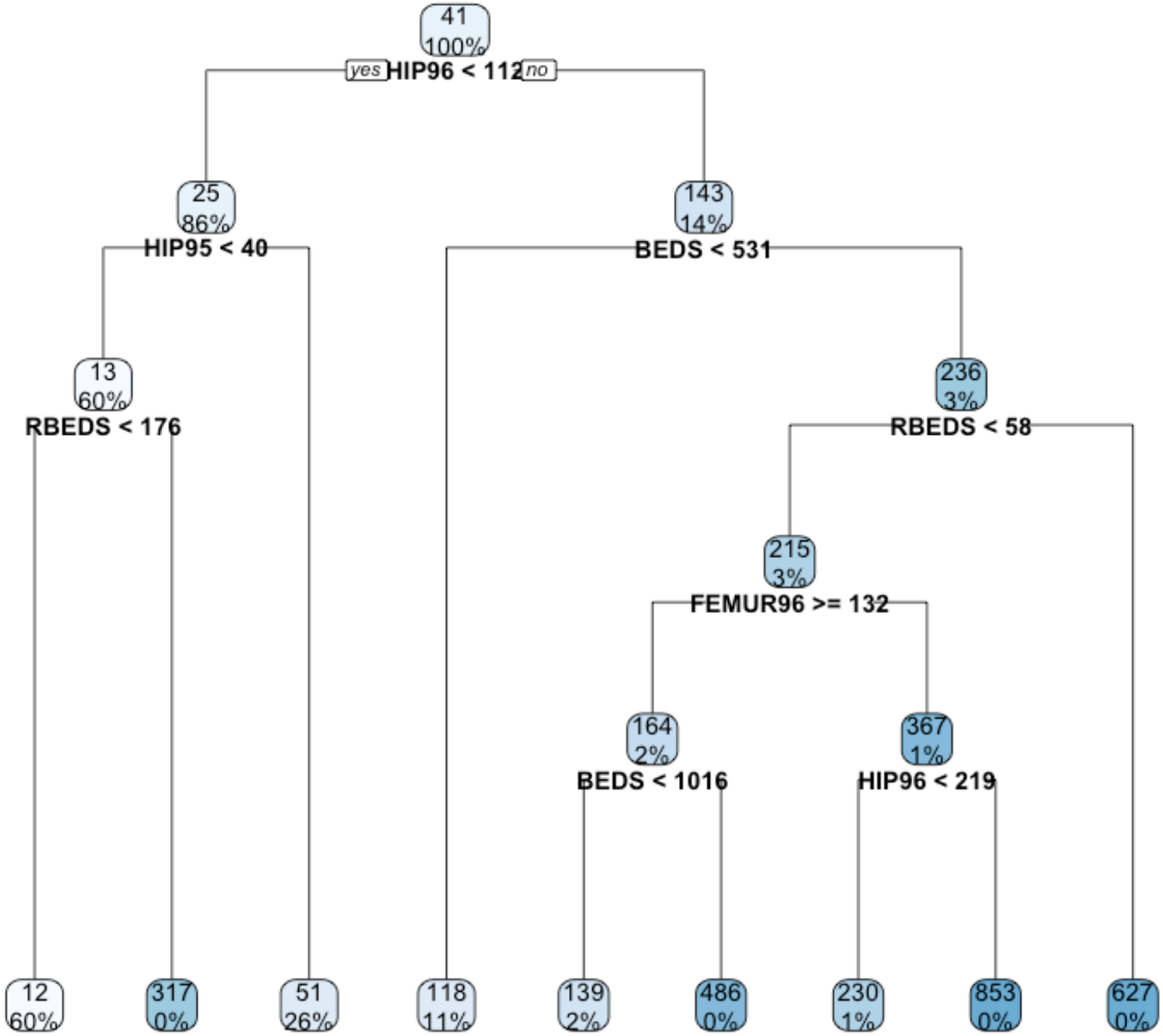
- DATASET: 768 Pima Indian females, 21+ years old ; 268 tested positive to diabetes
- 8 PREDICTORS: PRG, PLASMA, BP, THICK, INSULIN, BODY, PEDIGREE, AGE
- OBJECTIVE: PREDICT DIABETES



Node	CART	N	P(Diabetes)
Combined	993.5	768	35%
PLASMA<=127	854.3	485	19%
PLASMA>127		283	61%
AGE<=28	916.3	367	19%
AGE>28		401	49%
BODY<=27.8	913.7	222	12%
BODY>27.8		546	44%



Regression Tree for Sales



Linear Models

Linear model: $Y = X\beta + \varepsilon$

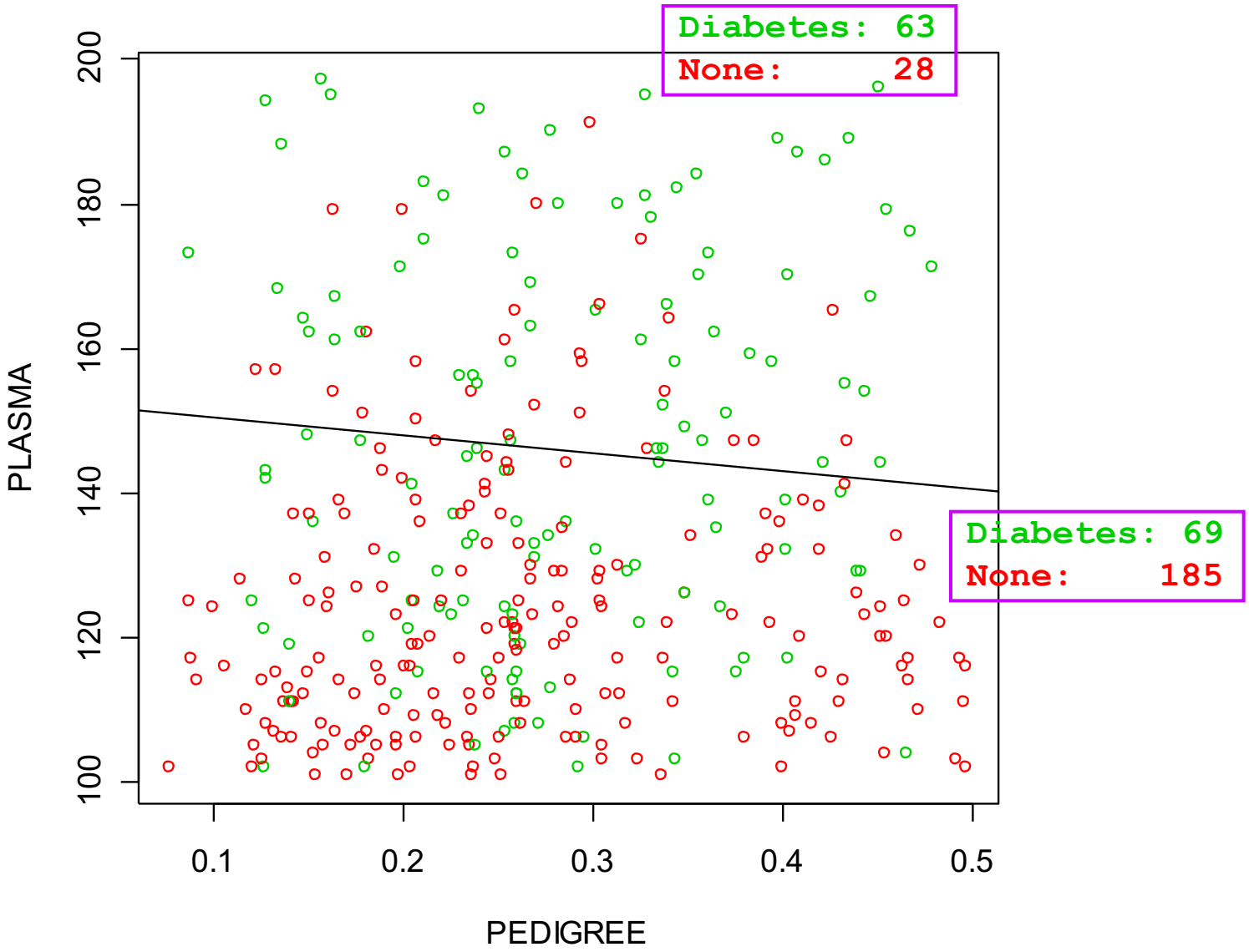
Least Squares Estimator: $b = (X^T X)^{-1} X^T Y$

Linear Discriminants

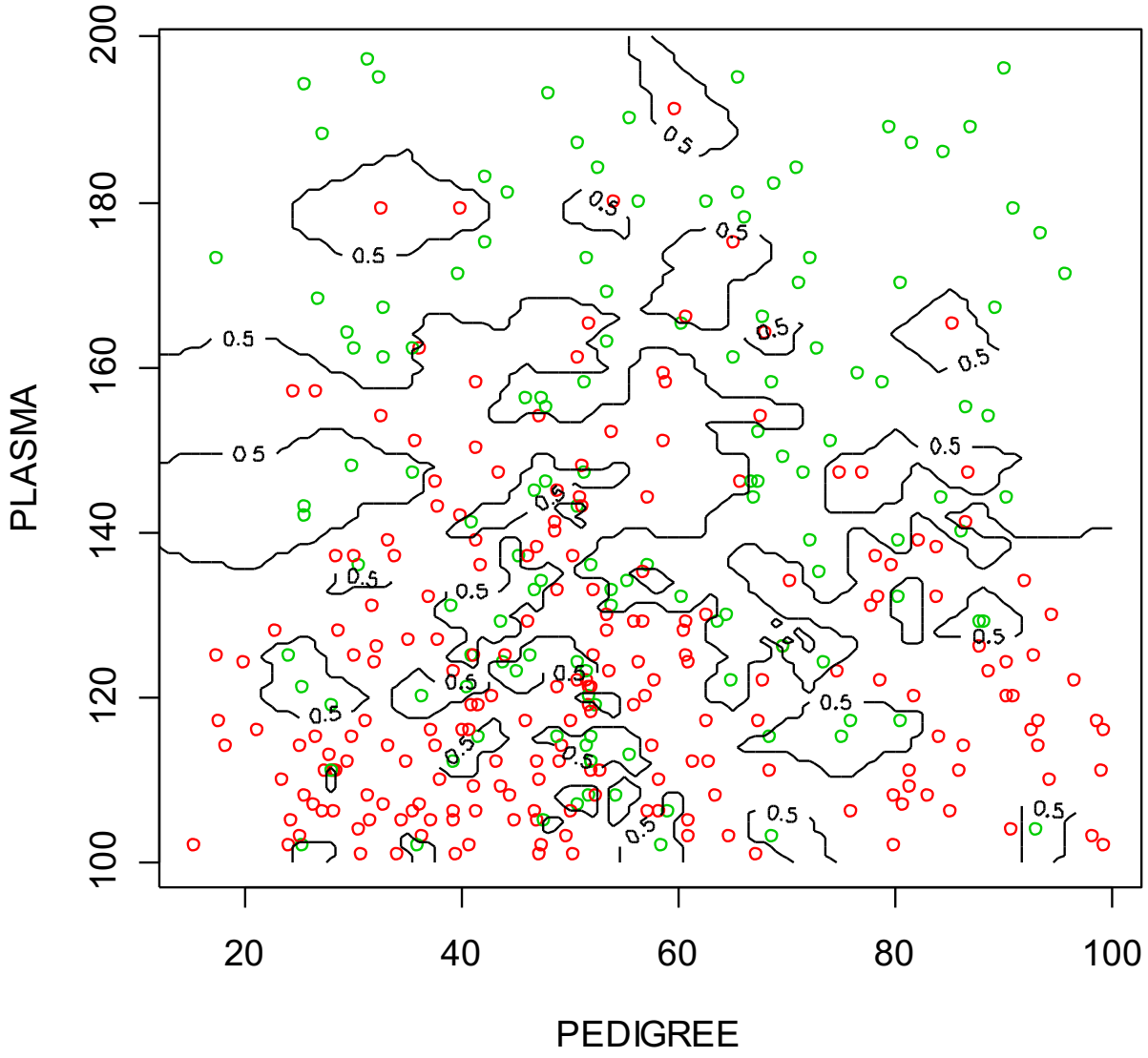
Linear Discriminant: Y = 0 or 1.

- Estimate b by L.S.
- Predict 1 if $Xb > 0.5$
  0 otherwise

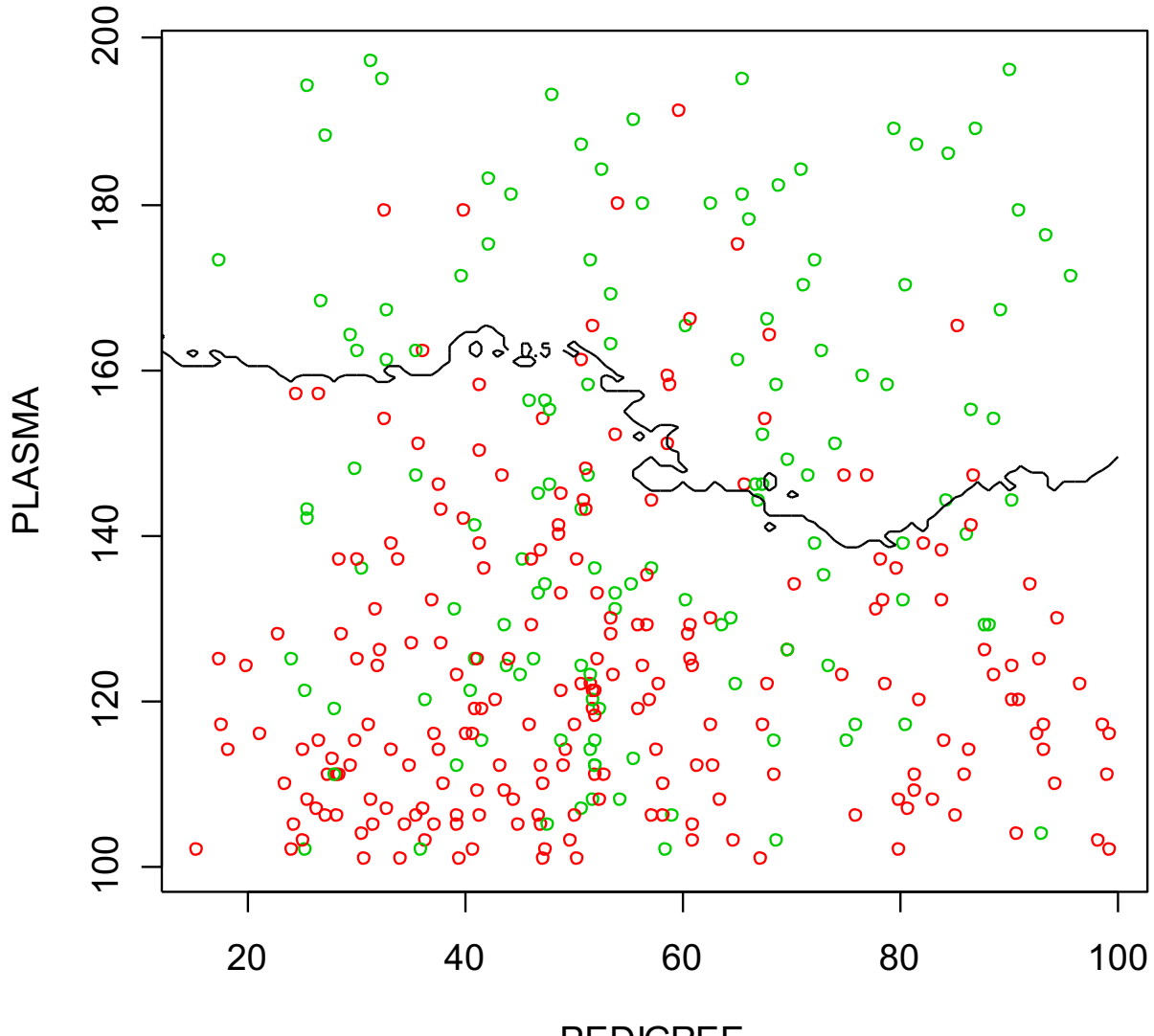
Example: Pima Indians



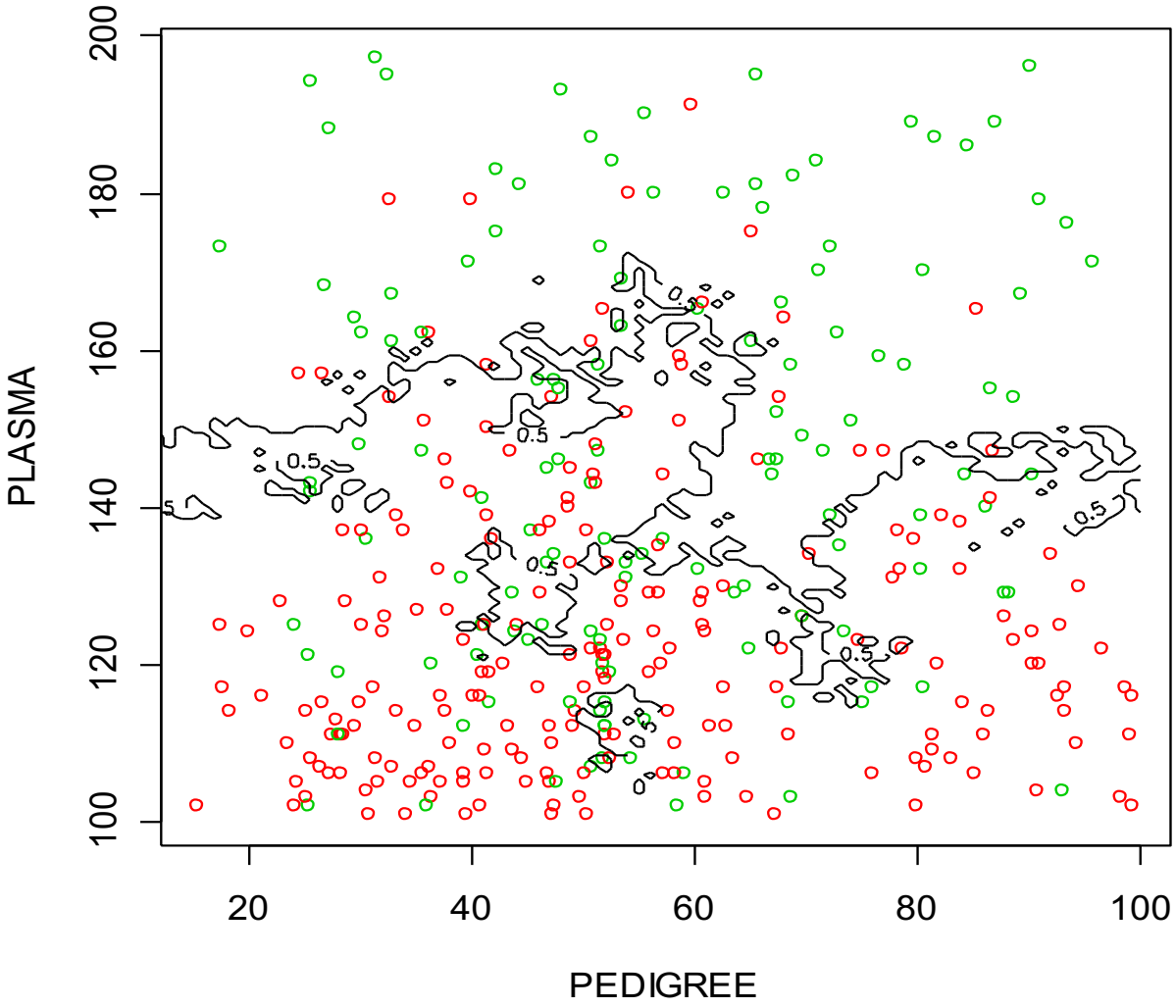
Example: Pima KNN with $k = 1$



Example: Pima KNN with k = 50



Example: Pima Indians K =10



Machine Learning

Pattern recognition

Data Mining Techniques:

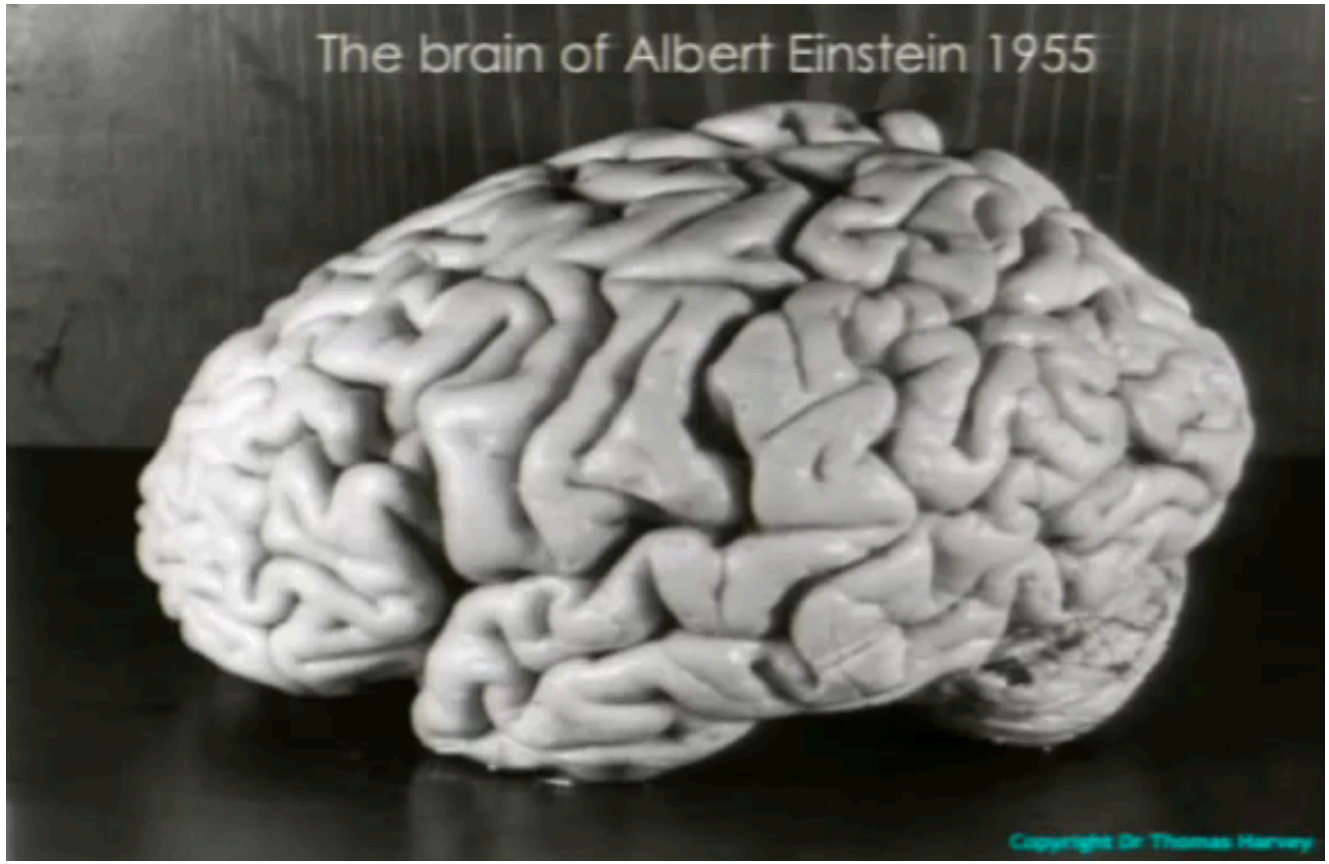
- Artificial Neural Nets
- Support Vector Machines
- Boosting

Objective:

Try to emulate the way the brain works (???)

Hoax:

The mechanisms underlying the functioning of the brain are not yet understood. Any relation with Artificial Neural Nets is purely anecdotal. (Einstein's Brain)



Einstein's brain has a standard count of neurons but extraordinary counts of glial cells known as astrocytes and oligodendrocyteshttp://www.npr.org/templates/story/story.php?storyId=126229305