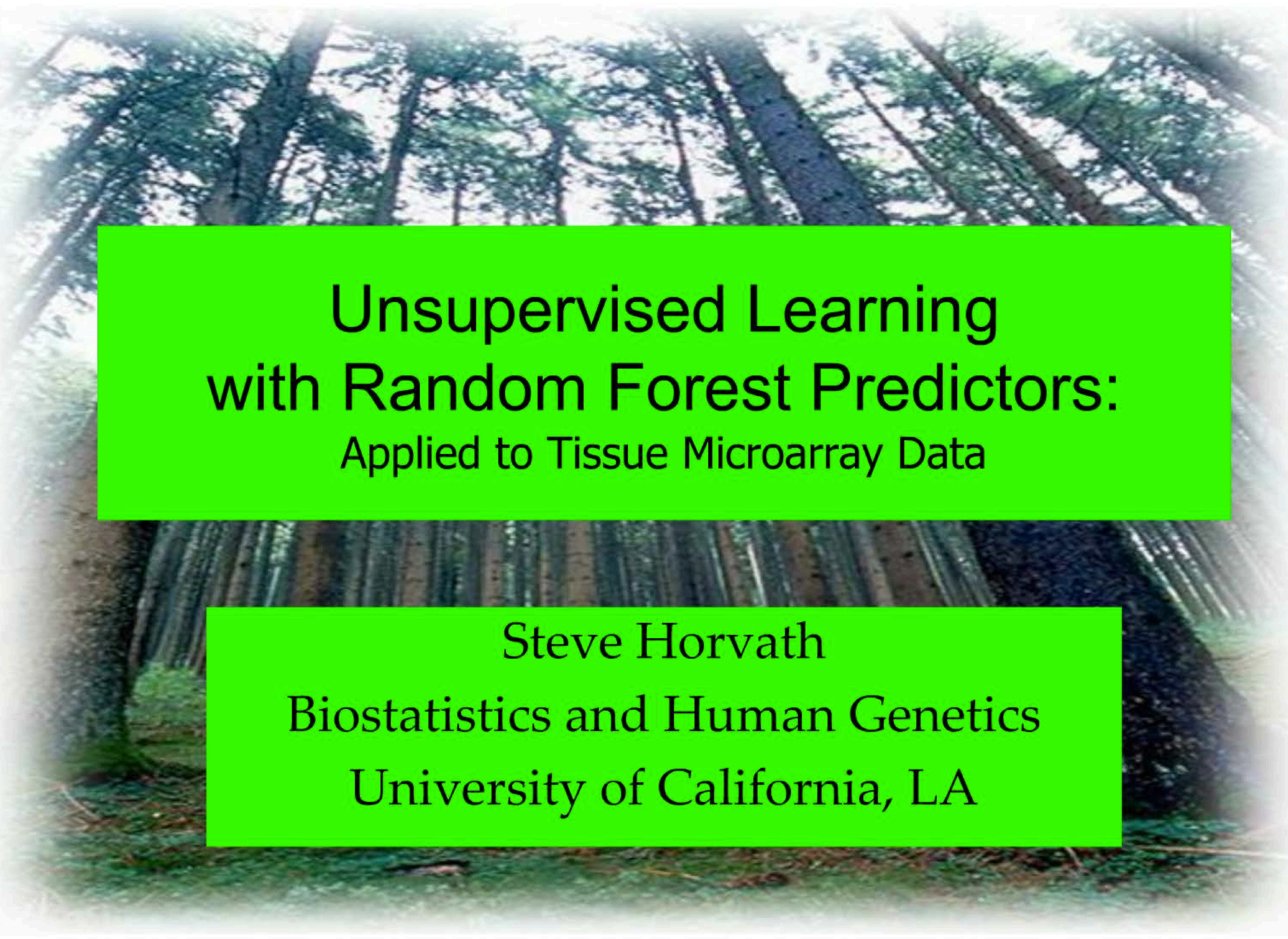


Unsupervised Learning with Random Forest Predictors:

Applied to Tissue Microarray Data

Steve Horvath

Biostatistics and Human Genetics

University of California, LA

Contents

- Tissue Microarray Data
- Random forest (RF) predictors
- Understanding RF clustering
  - Shi, T. and Horvath, S. (2006) "Unsupervised learning using random forest predictors" J. Comp. Graph. Stat.
- Applications to Tissue Microarray Data:
    - Shi et al (2004) "Tumor Profiling of Renal Cell Carcinoma Tissue Microarray Data" Modern Pathology
    - Seligson DB et al (2005) Global histone modification patterns predict risk of prostate cancer recurrence. Nature

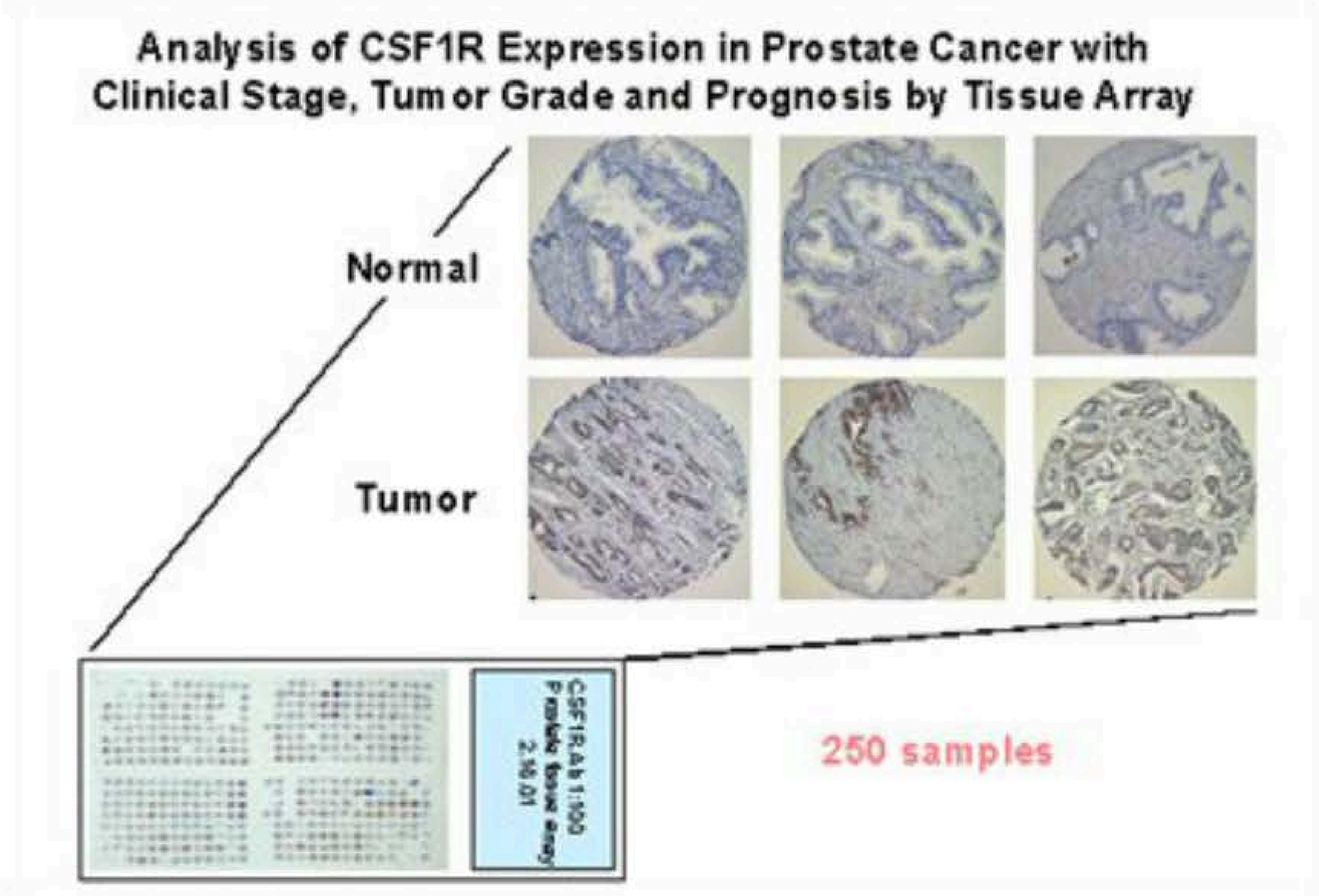
Acknowledgements

- **Former students & Postdocs for TMA**
  - **Tao Shi, PhD**
  - Tuyen Hoang, PhD
  - Yunda Huang, PhD
  - Xueli Liu, PhD

UCLA
Tissue Microarray Core

- **David Seligson, MD**
- Aarno Palotie, MD
- Arie Belldegrun, MD
- Robert Figlin, MD
- Lee Goodglick, MD
- David Chia, MD
- Siavash Kurdistani, MD

Tissue Microarray Data



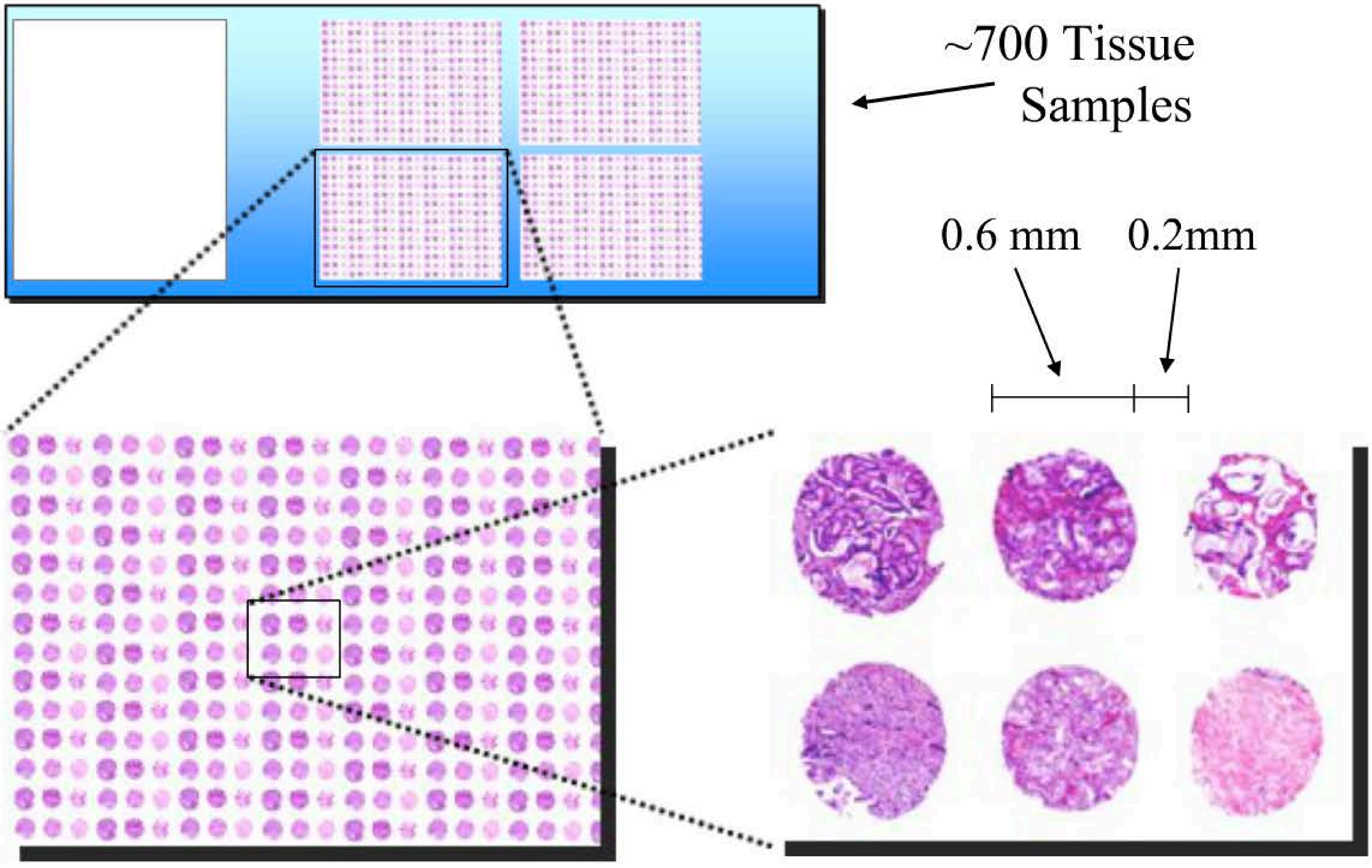
Analysis of CSF1R Expression in Prostate Cancer with Clinical Stage, Tumor Grade and Prognosis by Tissue Array

Tissue Microarray

$\neq$

DNA Microarray

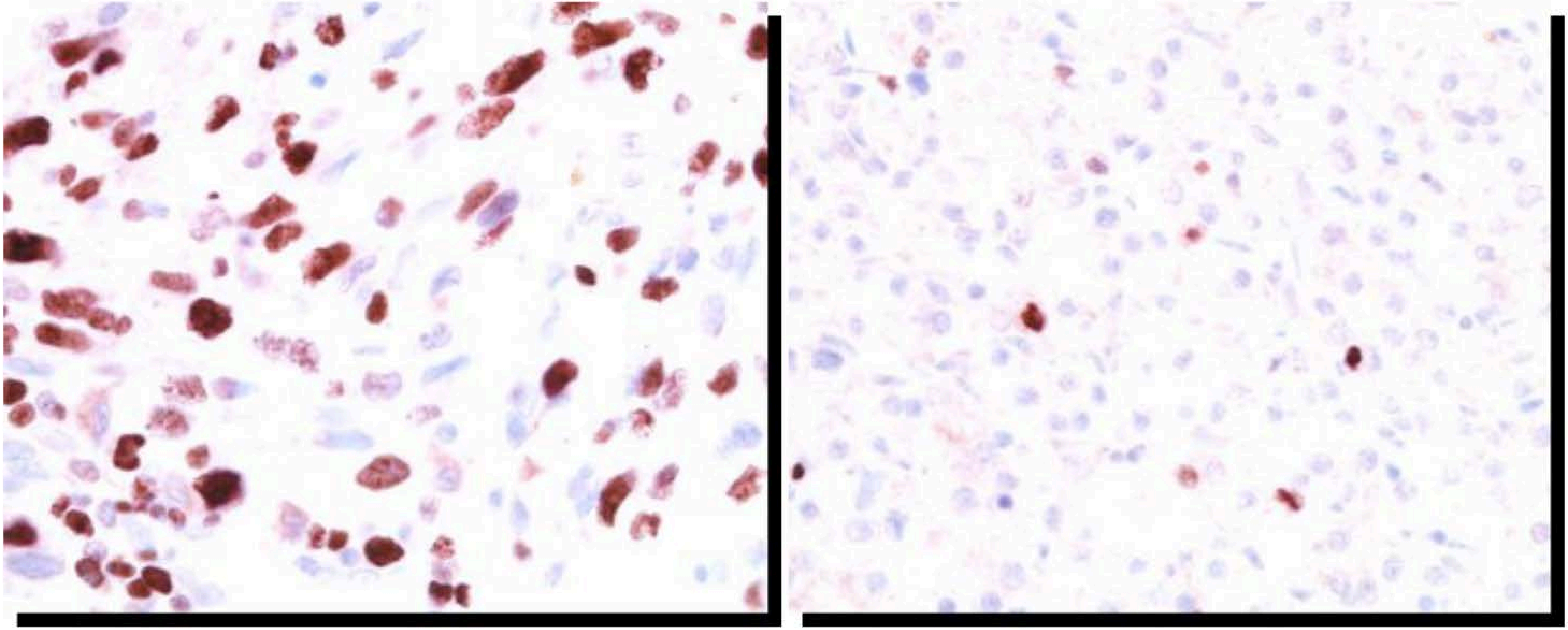
Tissue Array Section



~700 Tissue Samples

0.6 mm 0.2mm

Ki-67 Expression in Kidney Cancer

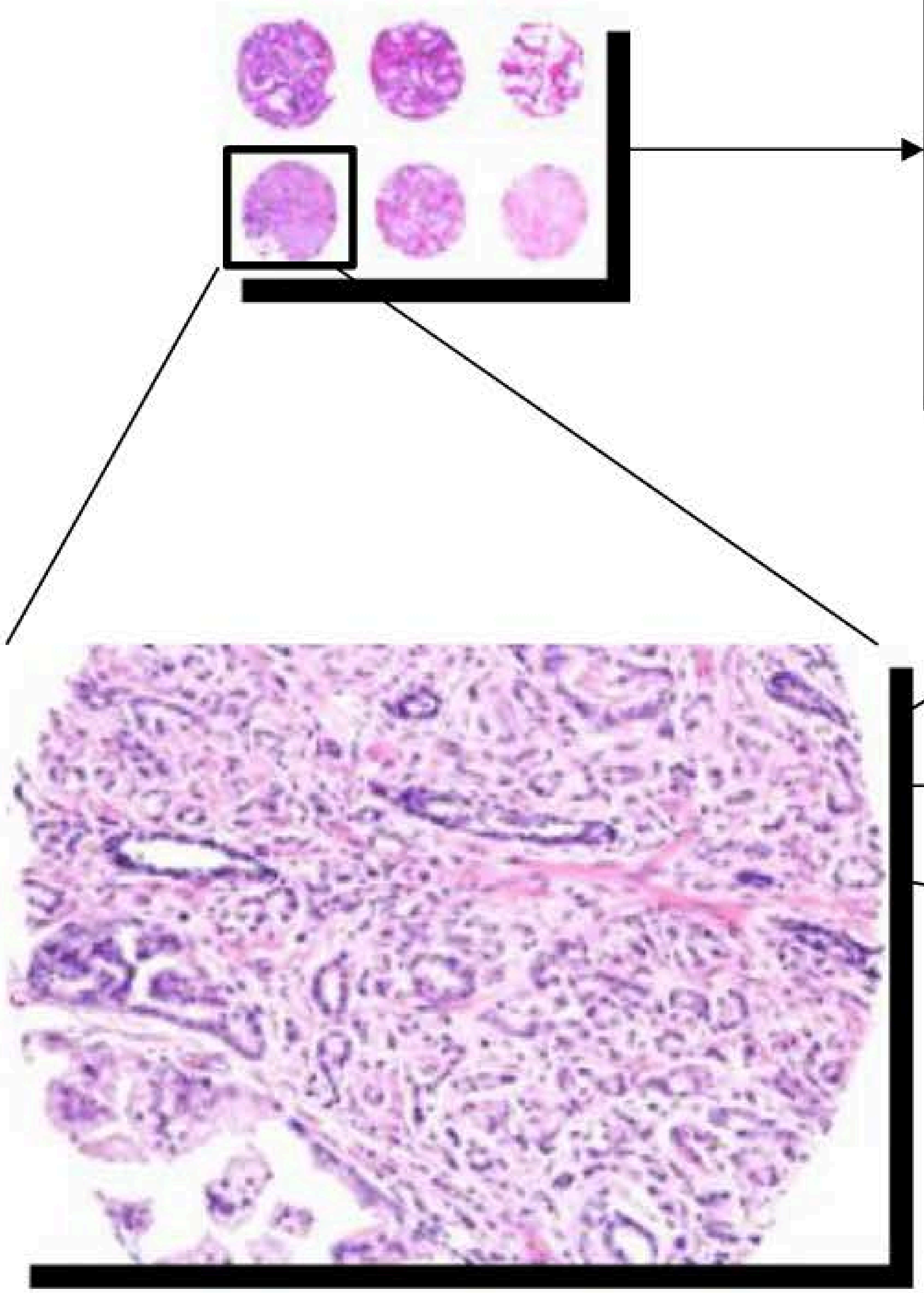


High Grade

Low Grade

Message: brown staining related to tumor grade

Multiple measurements per patient: Several spots per tumor sample and several “scores” per spot



- **Each patients (tumor sample) is usually represented by multiple spots**
  - **3 tumor spots**
  - **1 matched normal spot**

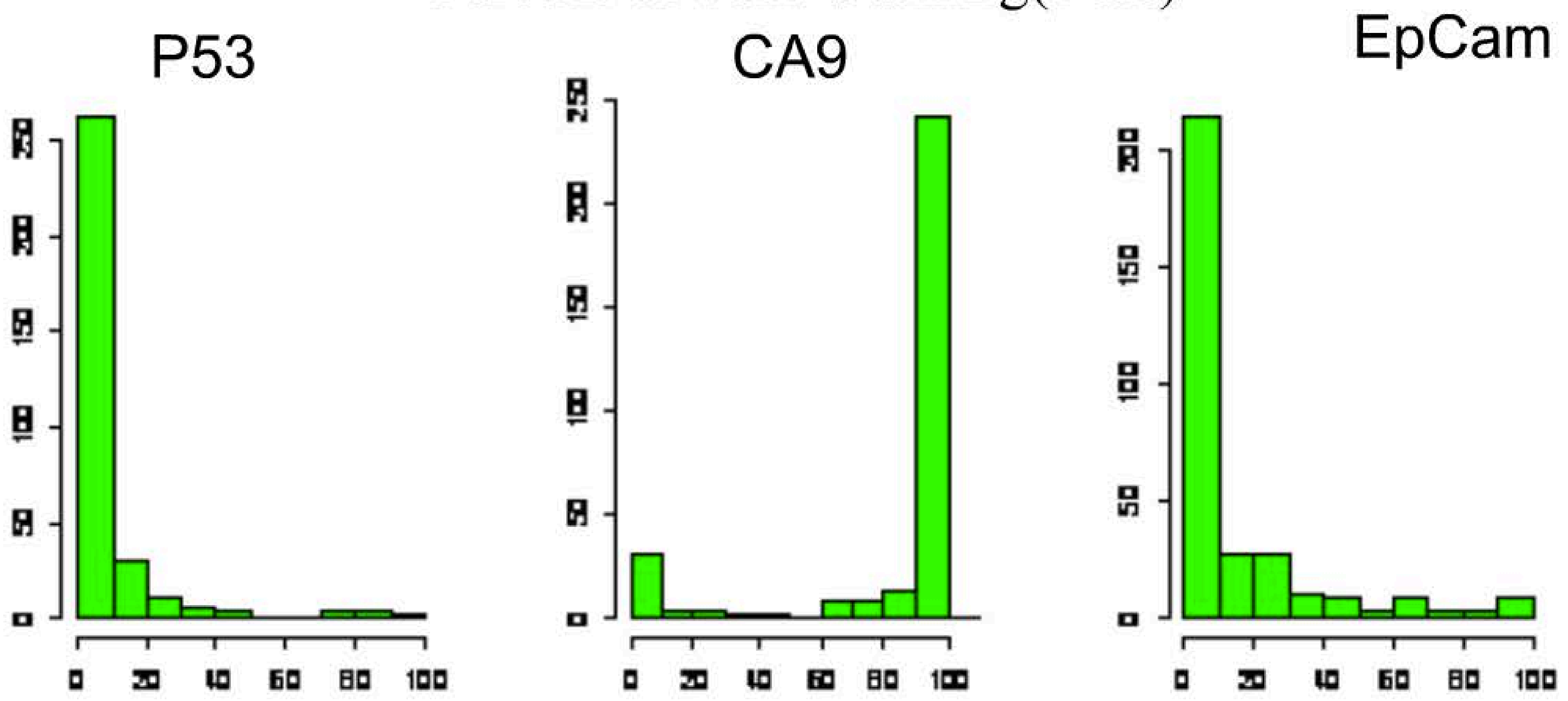
- Maximum intensity = Max
- Percent of cells staining = Pos
- Spots have a spot grade: NL,1,2,.

Properties of TMA Data

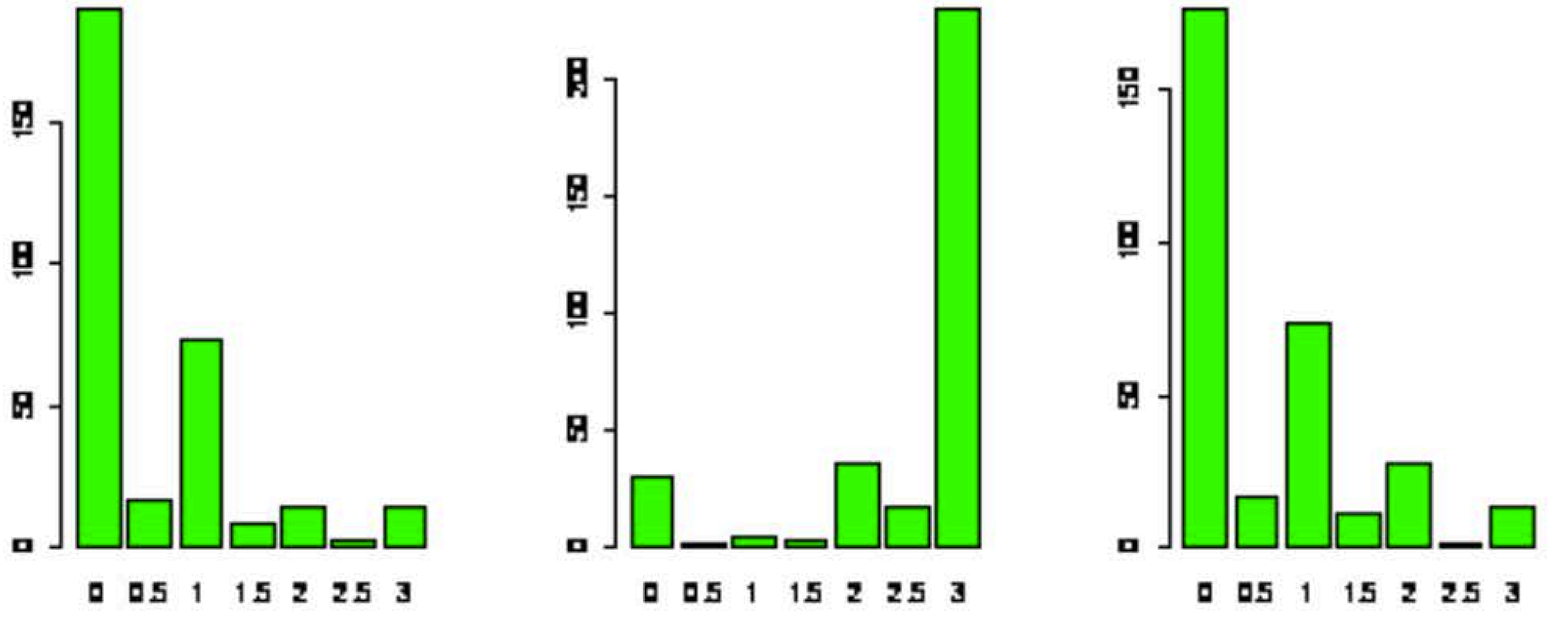
- Highly skewed, non-normal, semi-continuous.
  - Often a good idea to model as ordinal variables with many levels.
- Staining scores of the same markers are highly correlated

Histogram of tumor marker expression scores: POS and MAX

Percent of Cells Staining(POS)



Maximum Intensity (MAX)



Thresholding methods for tumor marker expressions

- Since clinicians and pathologists prefer thresholding tumor marker expressions, it is natural to use statistical methods that are based on thresholding covariates, e.g. regression trees, survival trees, rpart, forest predictors etc.
- Dichotomized marker expressions are often fitted in a Cox (or alternative) regression model
  - Danger: Over-fitting due to optimal cut-off selection.
  - Several thresholding methods and ways for adjusting for multiple comparisons are reviewed in
    - Liu X, Minin V, Huang Y, Seligson DB, Horvath S (2004) Statistical Methods for Analyzing Tissue Microarray Data. J of Biopharmaceutical Statistics. Vol 14(3) 671-685

Tumor class discovery

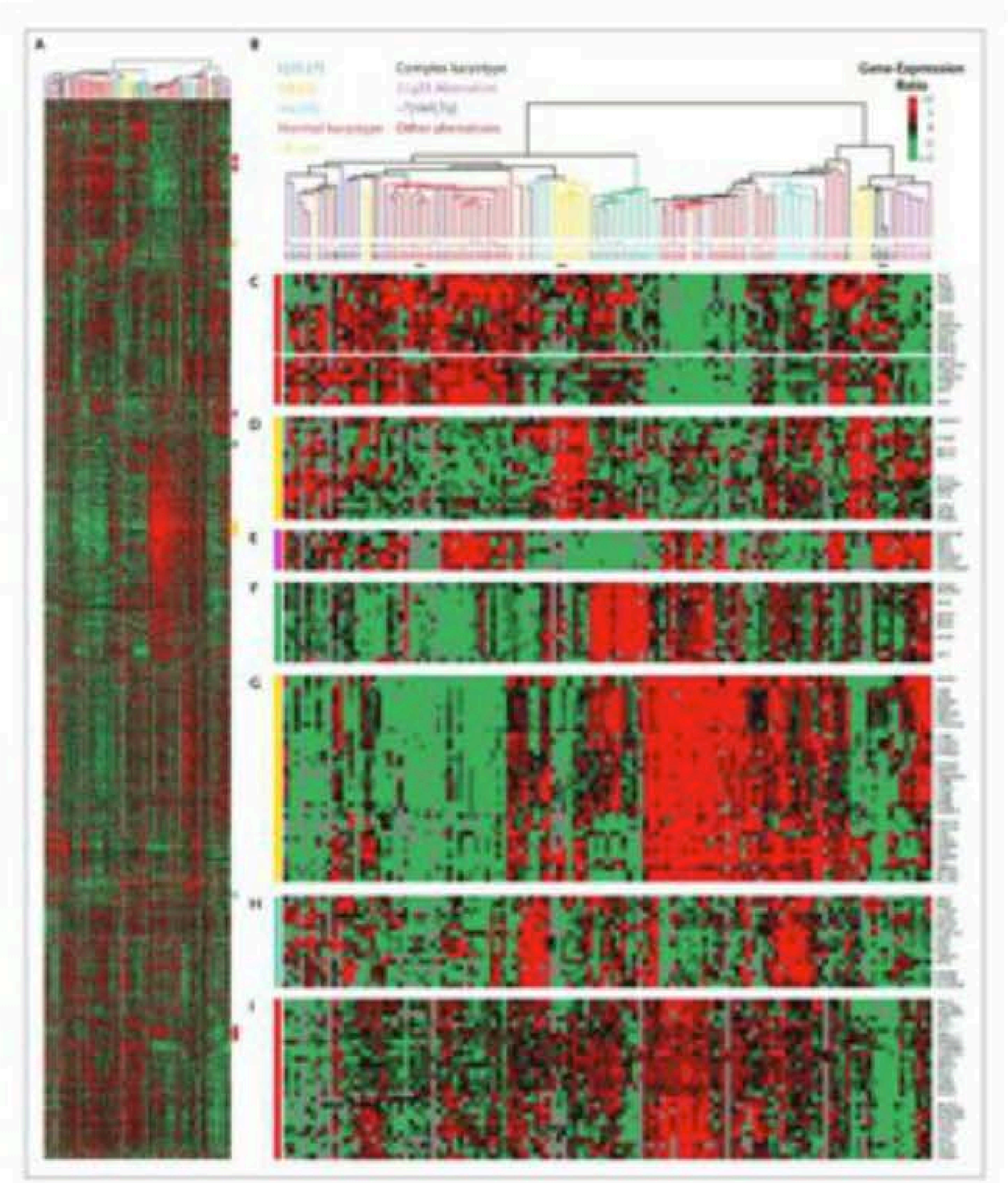
Keywords: unsupervised learning, clustering

Tumor Class Discovery

- Molecular tumor classes=clusters of patients with similar gene expression profiles
- Main road for tumor class discovery
  - DNA microarrays
  - Proteomics etc
  - unsupervised learning: clustering, multi-dimensional scaling plots
- Tissue microarrays have been used for tumor marker validation
  - supervised learning, Cox regression etc
- Challenge: show that tissue microarray data can be used in unsupervised learning to find tumor classes
  - road less travelled

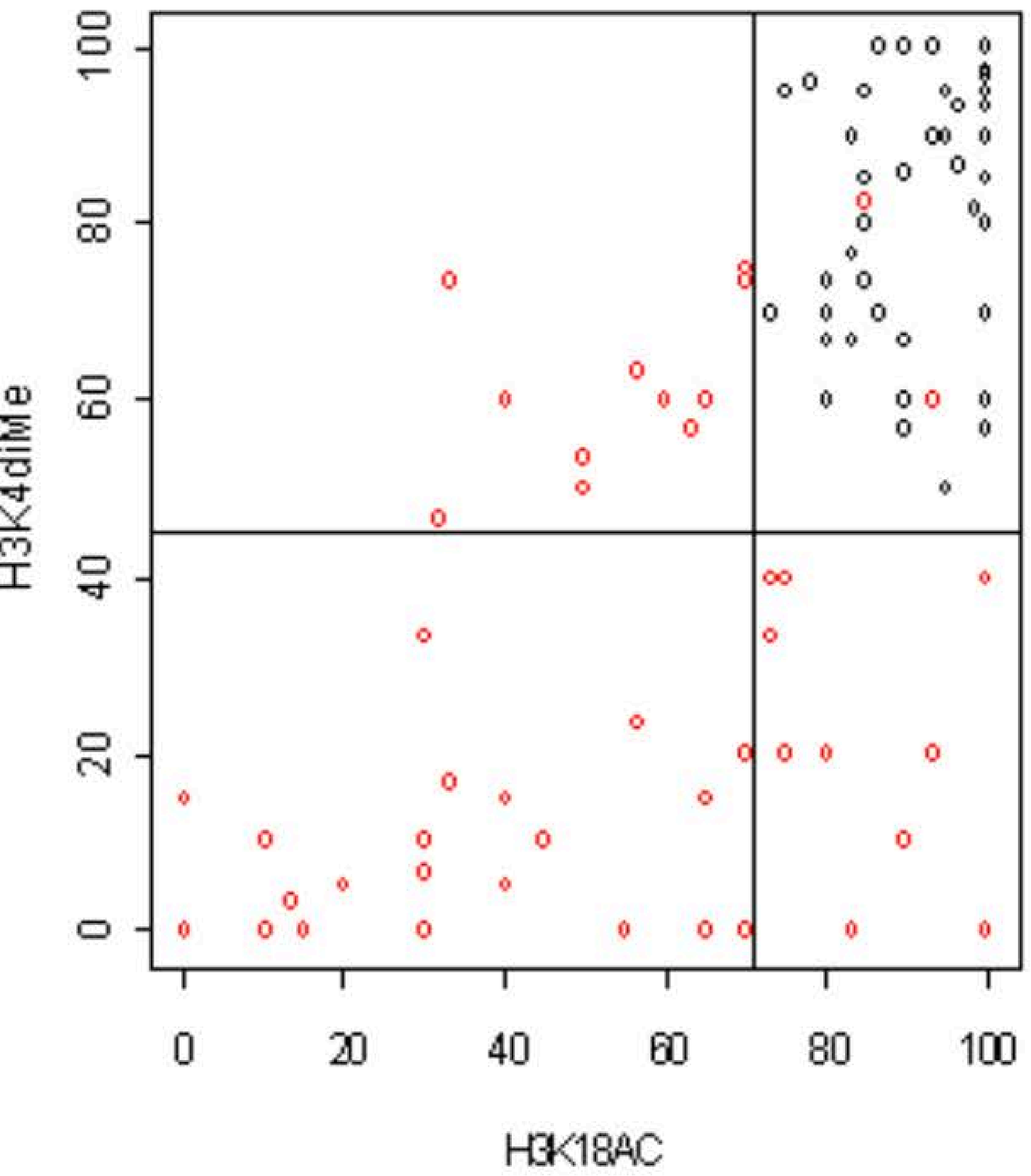
Tumor Class Discovery using DNA Microarray Data

- Tumor class discovery entails using a unsupervised learning algorithm (e.g hierarchical, k-means, clustering etc.) to automatically group tumor samples based on their gene expression pattern.



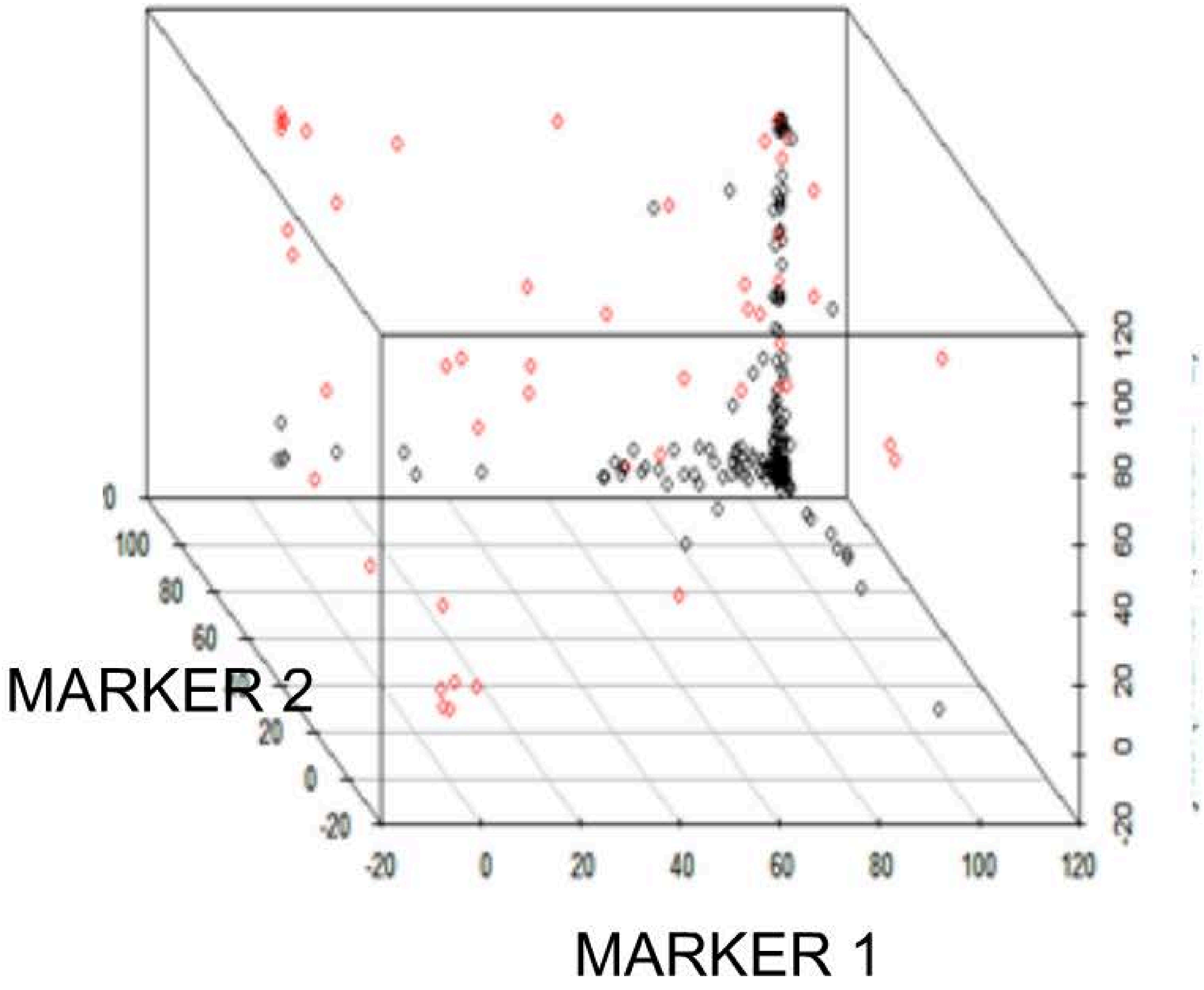
Bullinger et al. *N Engl J Med.* 2004

Clusters involving TMA data may have unconventional shapes: Low risk prostate cancer patients are colored in black.



- Scatter plot involving 2 `dependent' tumor markers. The remaining, less dependent markers are not shown.
- Low risk cluster can be described using the following rule Marker H3K4 > 45% and H3K18 > 70%.
- The intuition is quite different from that of Euclidean distance based clusters.

Unconventional shape of a clinically meaningful patient cluster



- 3 dimensional scatter plot along tumor markers
- Low risk patients are colored in black

How to cluster patients on the basis of Tissue Microarray Data?

A dissimilarity measure is an essential input for tumor class discovery

- Dissimilarities between tumor samples are used in clustering and other unsupervised learning techniques

- Commonly used dissimilarity measures include Euclidean distance, 1 - correlation

Challenge

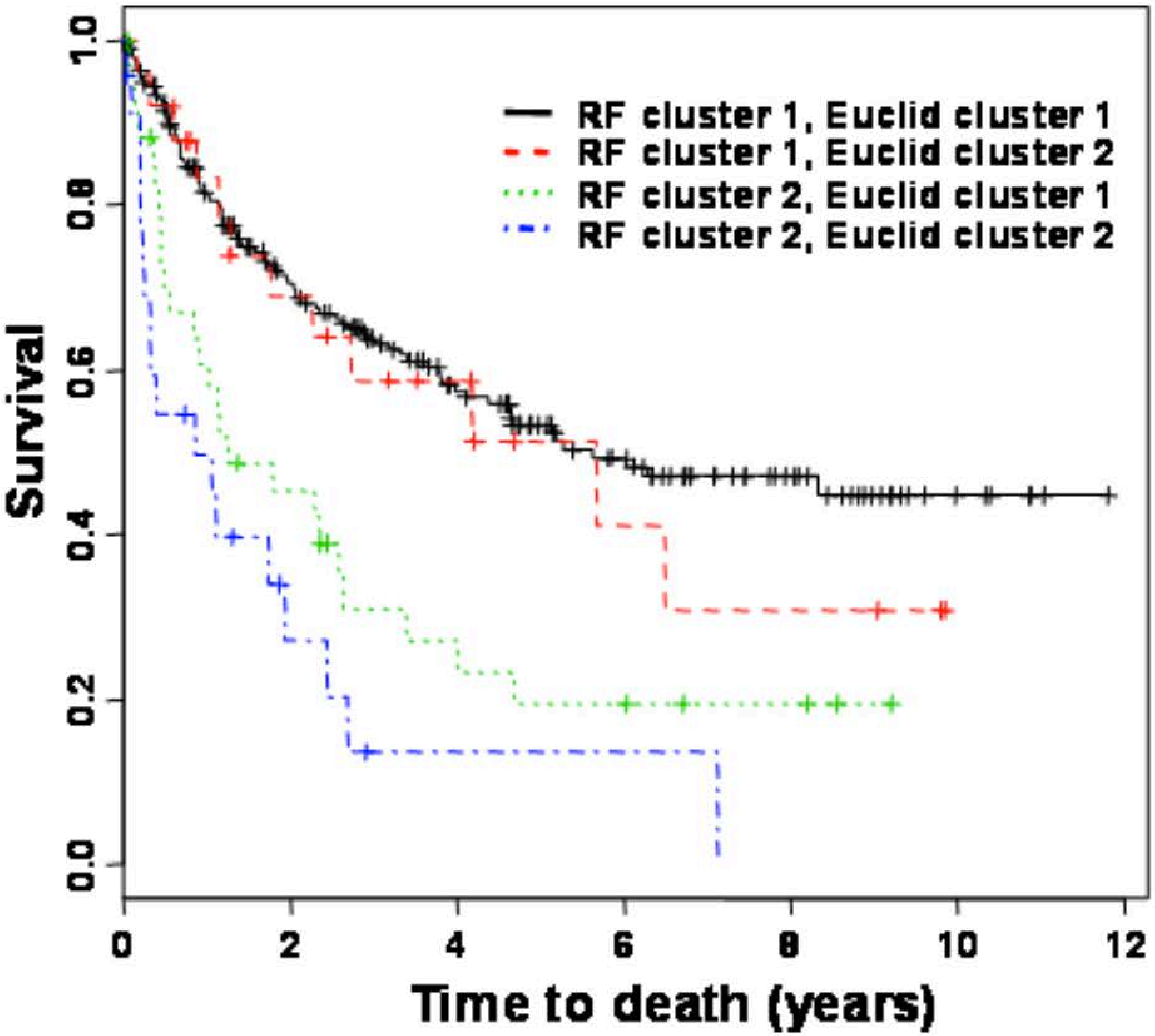
- Conventional dissimilarity measures that work for DNA microarray data may not be optimal for TMA data.
  - Dissimilarity measure that are based on the intuition of multivariate normal distributions (clusters have elliptical shapes) may not be optimal
  - For tumor marker data, one may want to use a different intuition: clusters are described using thresholding rules involving dependent markers.
  - It may be desirable to have a dissimilarity that is invariant under monotonic transformations of the tumor marker expressions.

We have found that a random forest (Breiman 2001) dissimilarity can work well in the unsupervised analysis of TMA data.

Shi et al 2004, Seligson et al 2005.

http://www.genetics.ucla.edu/labs/horvath/RFclustering/RFclustering.htm

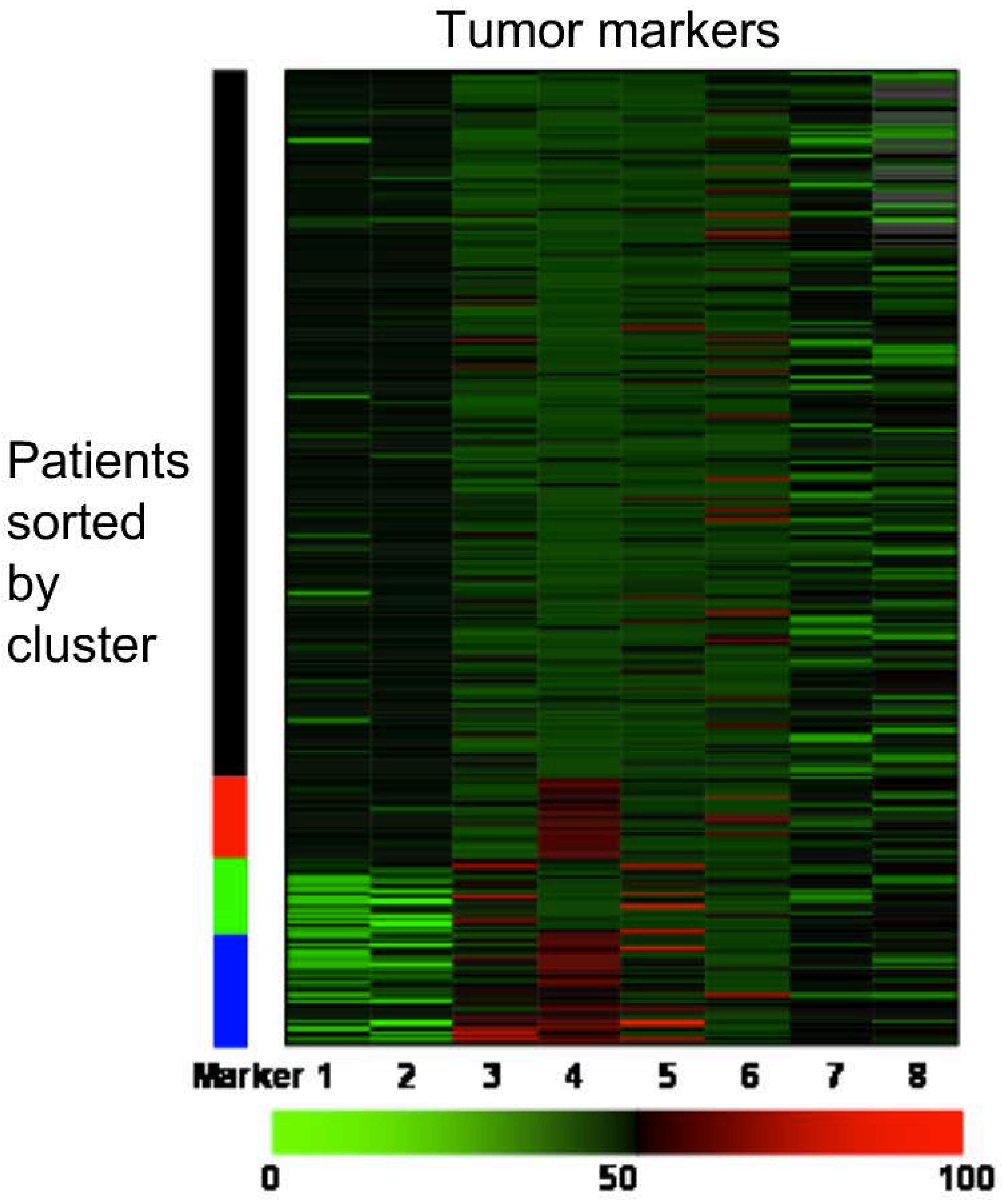
Kidney cancer: Comparing PAM clusters that result from using the RF dissimilarity vs the Euclidean distance



Kaplan Meier plots for groups defined by cross tabulating patients according to their RF and Euclidean distance cluster memberships.

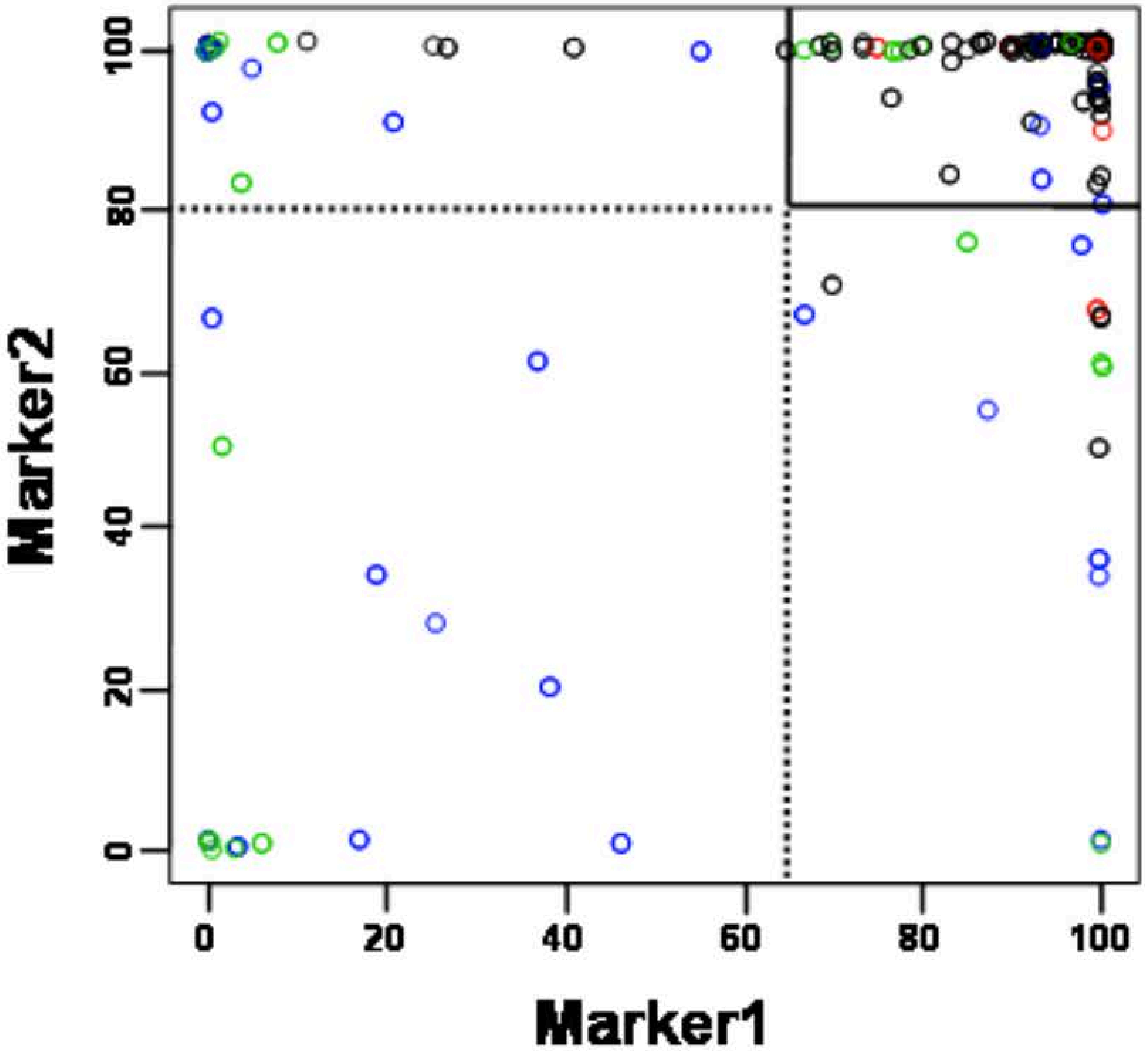
Message:
In this application, RF clusters are more meaningful regarding survival time

The RF dissimilarity is determined by dependent tumor markers



- The RF dissimilarity focuses on the most dependent markers (1,2).
- In some applications, it is good to focus on markers that are dependent since they may constitute a disease pathway.
- The Euclidean distance focuses on the most varying marker (4)

The RF cluster can be described using a thresholding rule involving the most dependent markers



- Low risk patient if marker1>cut1 & marker2> cut2
- This kind of thresholding rule can be used to make predictions on independent data sets.
- Validation on independent data set

Random Forest Predictors

Breiman L. Random forests. Machine Learning 2001;45(1):5-32
http://stat-www.berkeley.edu/users/breiman/RandomForests/

Tree predictors are the basic unit of random forest predictors

Classification and Regression Trees (CART)

by

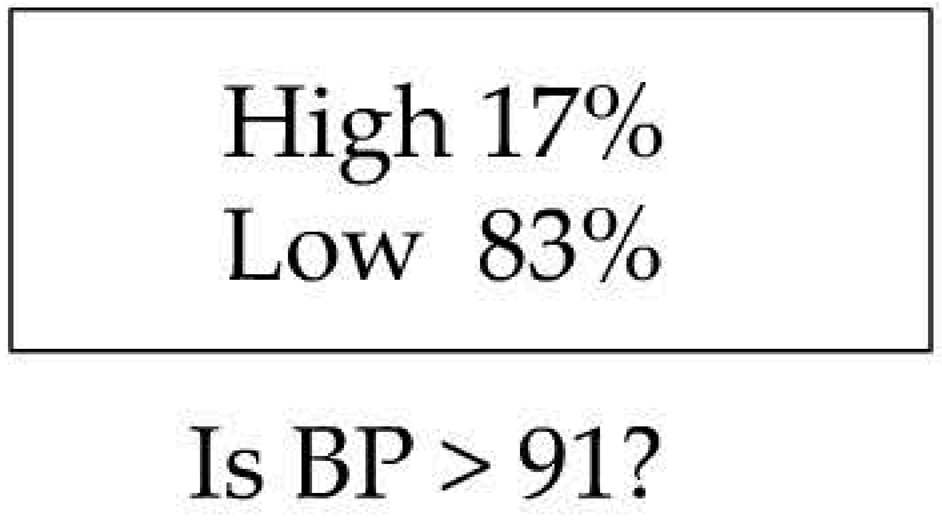
- Leo Breiman
- Jerry Friedman
- Charles J. Stone
- Richard Olshen

- RPART library in R software
  Therneau TM, et al.

An example of CART

- Goal: For the patients admitted into ER, to predict who is at higher risk of heart attack
- Training data set:
  - No. of subjects = 215
  - Outcome variable = High/Low Risk determined
  - 19 noninvasive clinical and lab variables were used as the predictors

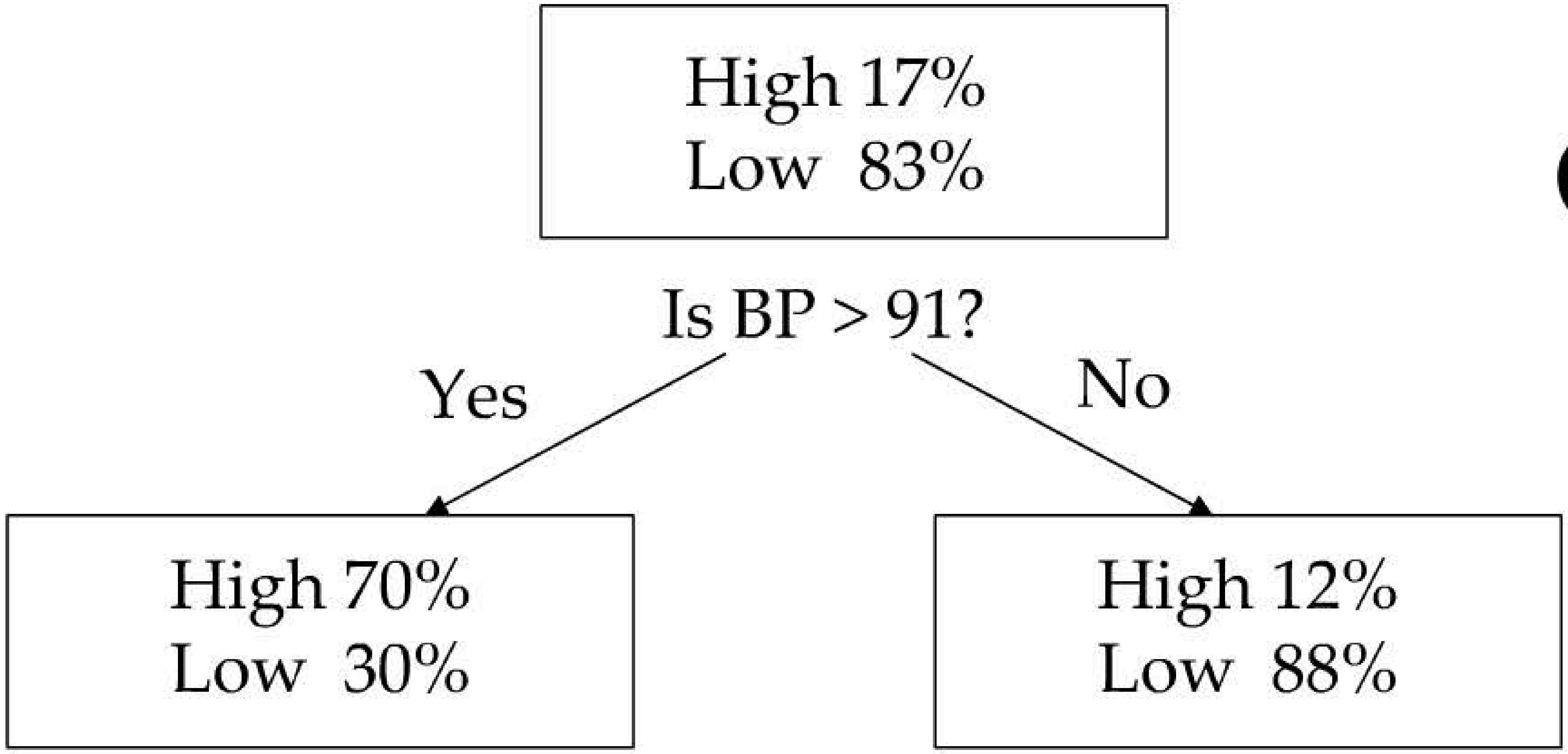
CART Construction



High 17%
Low 83%

Is BP > 91?

CART Construction



High 17%
Low 83%

Is BP > 91?

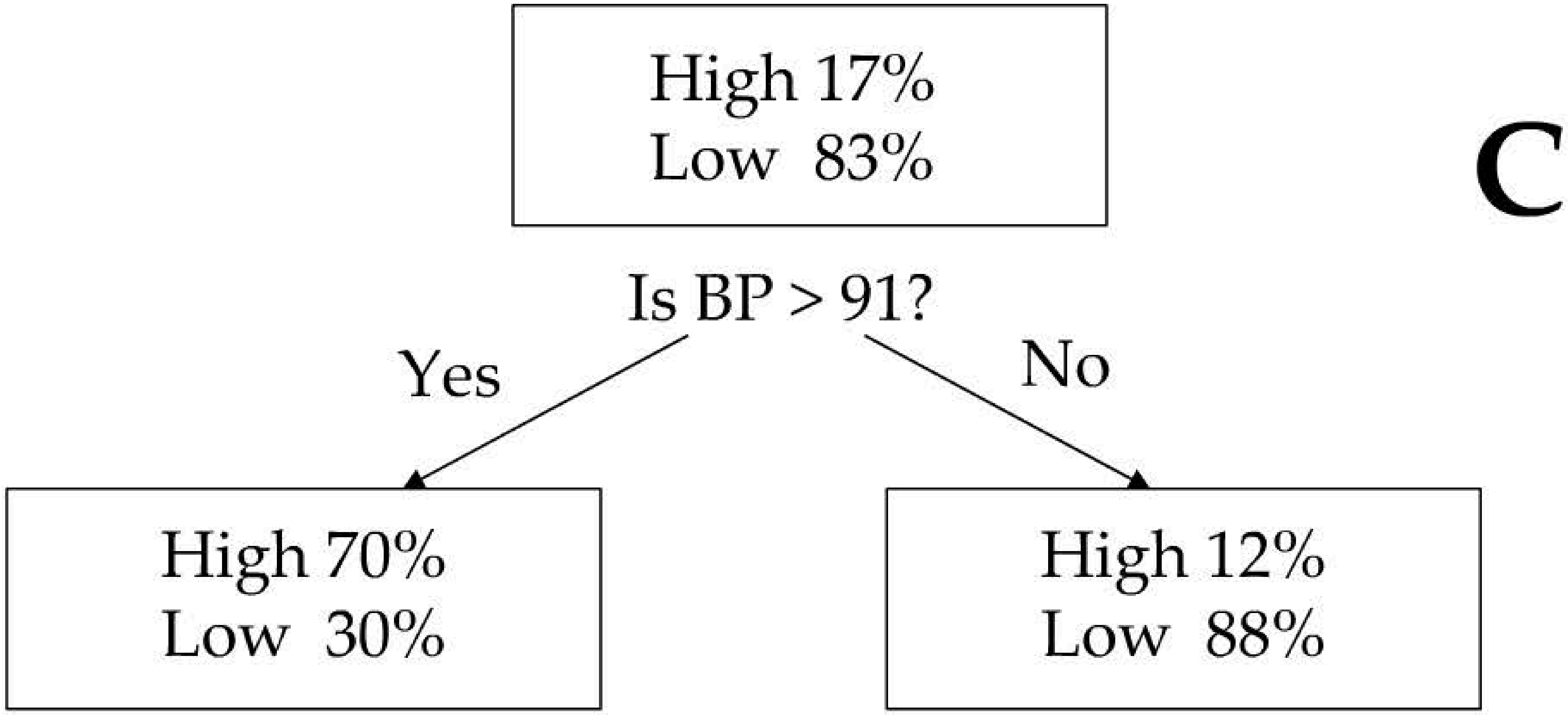
Yes

No

High 70%
Low 30%

High 12%
Low 88%

CART Construction



High 17%
Low 83%

Is BP > 91?

Yes

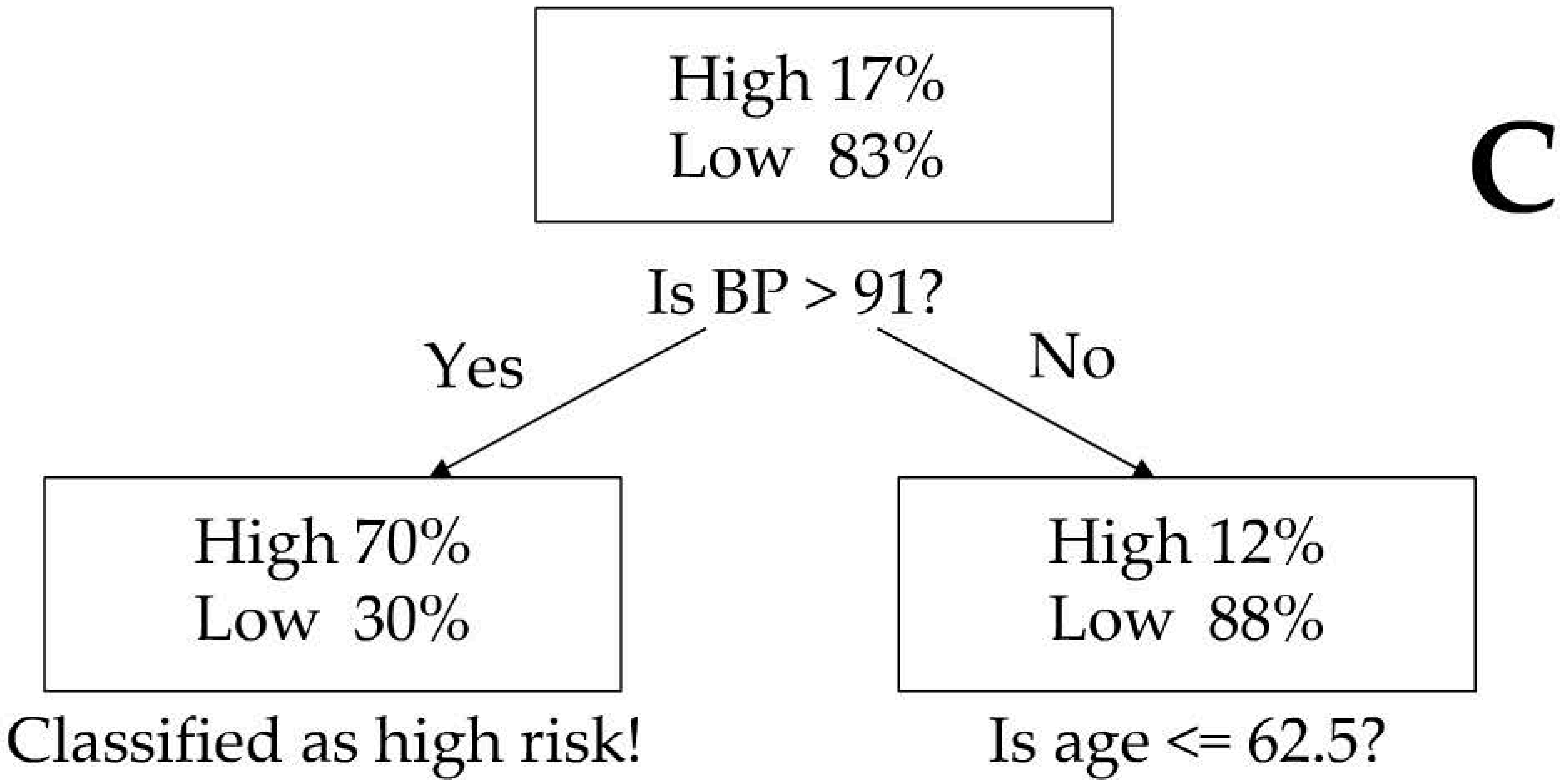
No

High 70%
Low 30%

High 12%
Low 88%

Classified as high risk!

CART Construction



High 17%
Low 83%

Is BP > 91?

Yes

No

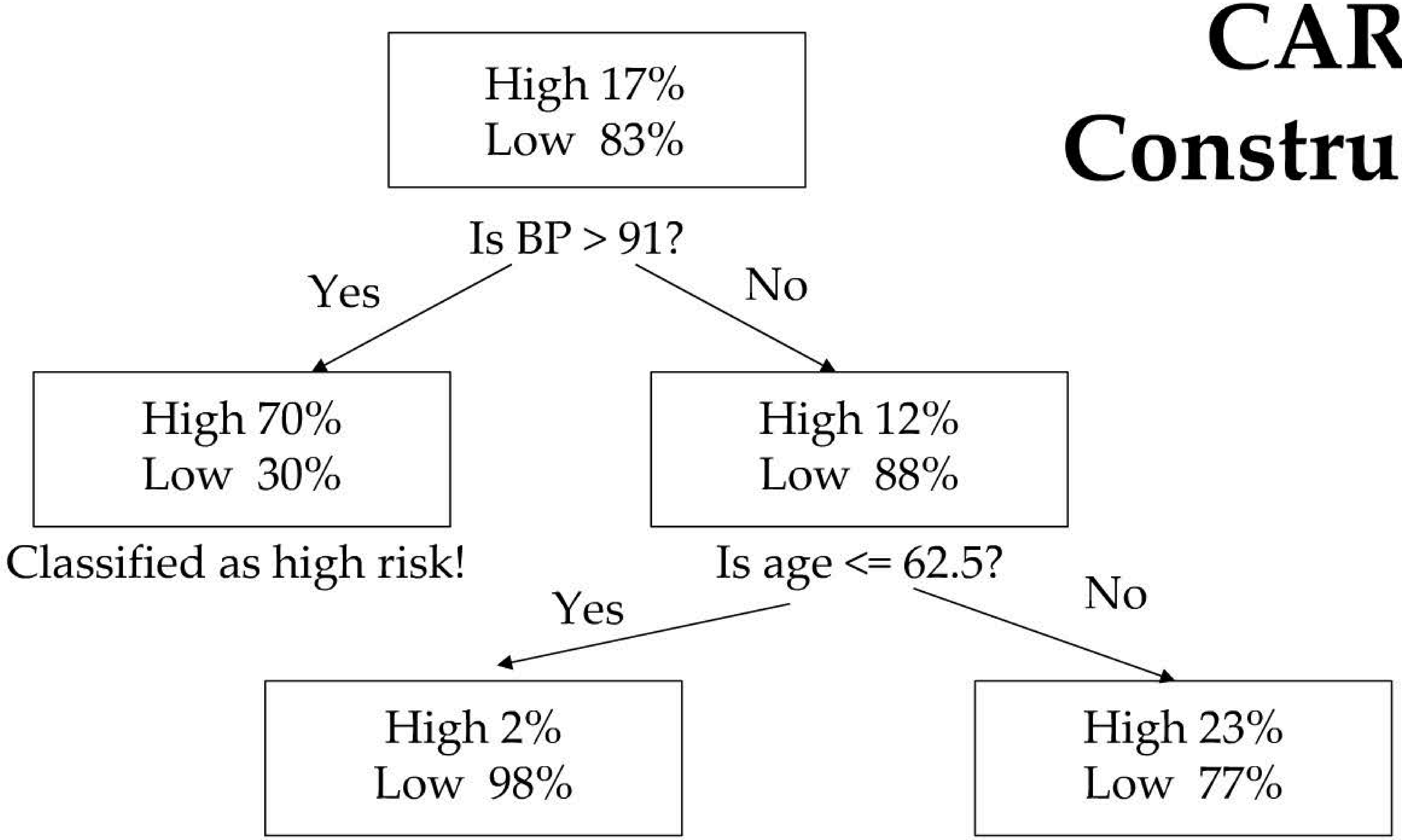
High 70%
Low 30%

Classified as high risk!

High 12%
Low 88%

Is age <= 62.5?

CART Construction



High 17%
Low 83%

Is BP > 91?

Yes

No

High 70%
Low 30%

Classified as high risk!

High 12%
Low 88%

Is age <= 62.5?

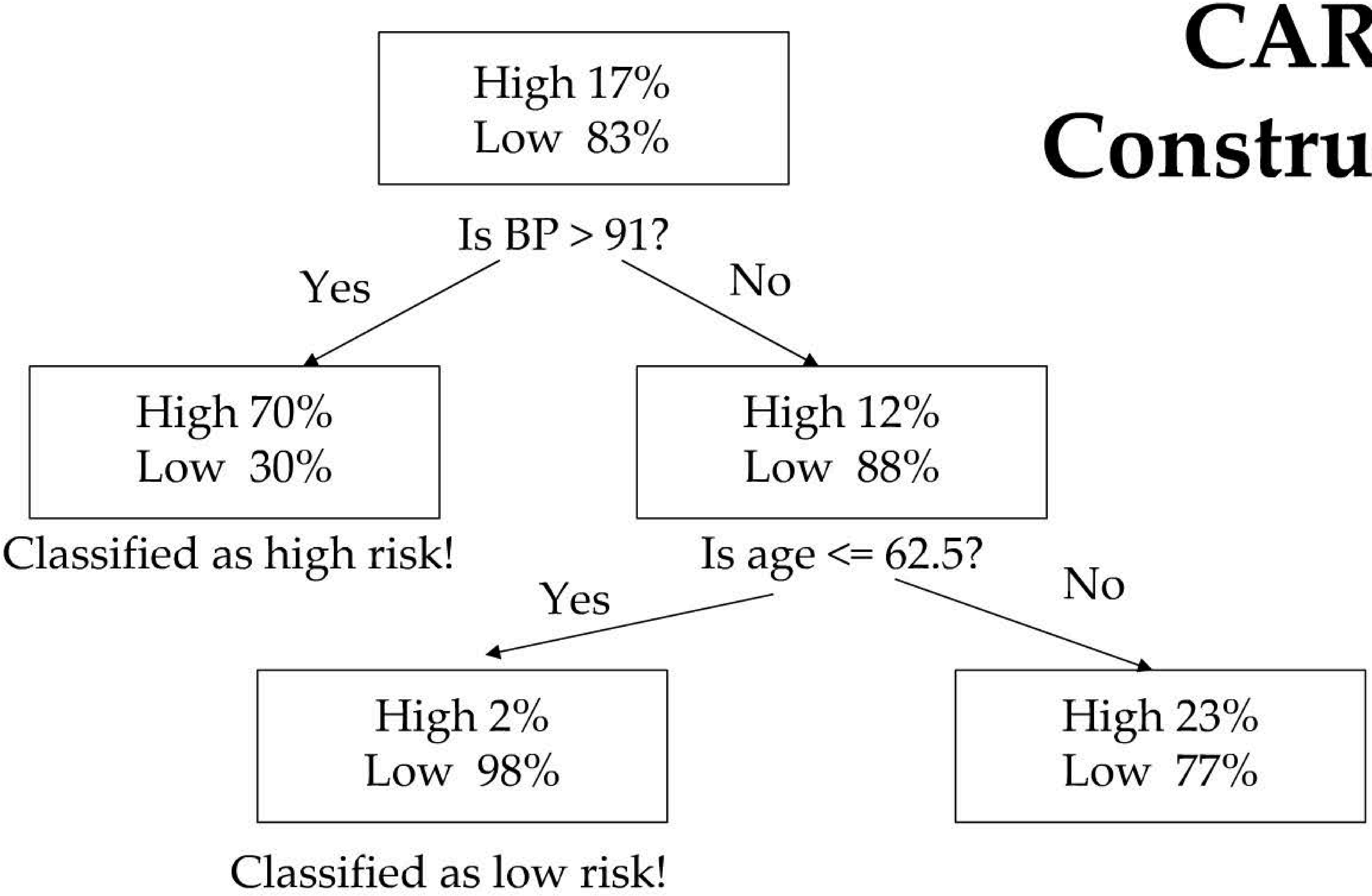
Yes

No

High 2%
Low 98%

High 23%
Low 77%

CART Construction



High 17%
Low 83%

Is BP > 91?

Yes

No

High 70%
Low 30%

Classified as high risk!

High 12%
Low 88%

Is age <= 62.5?

Yes

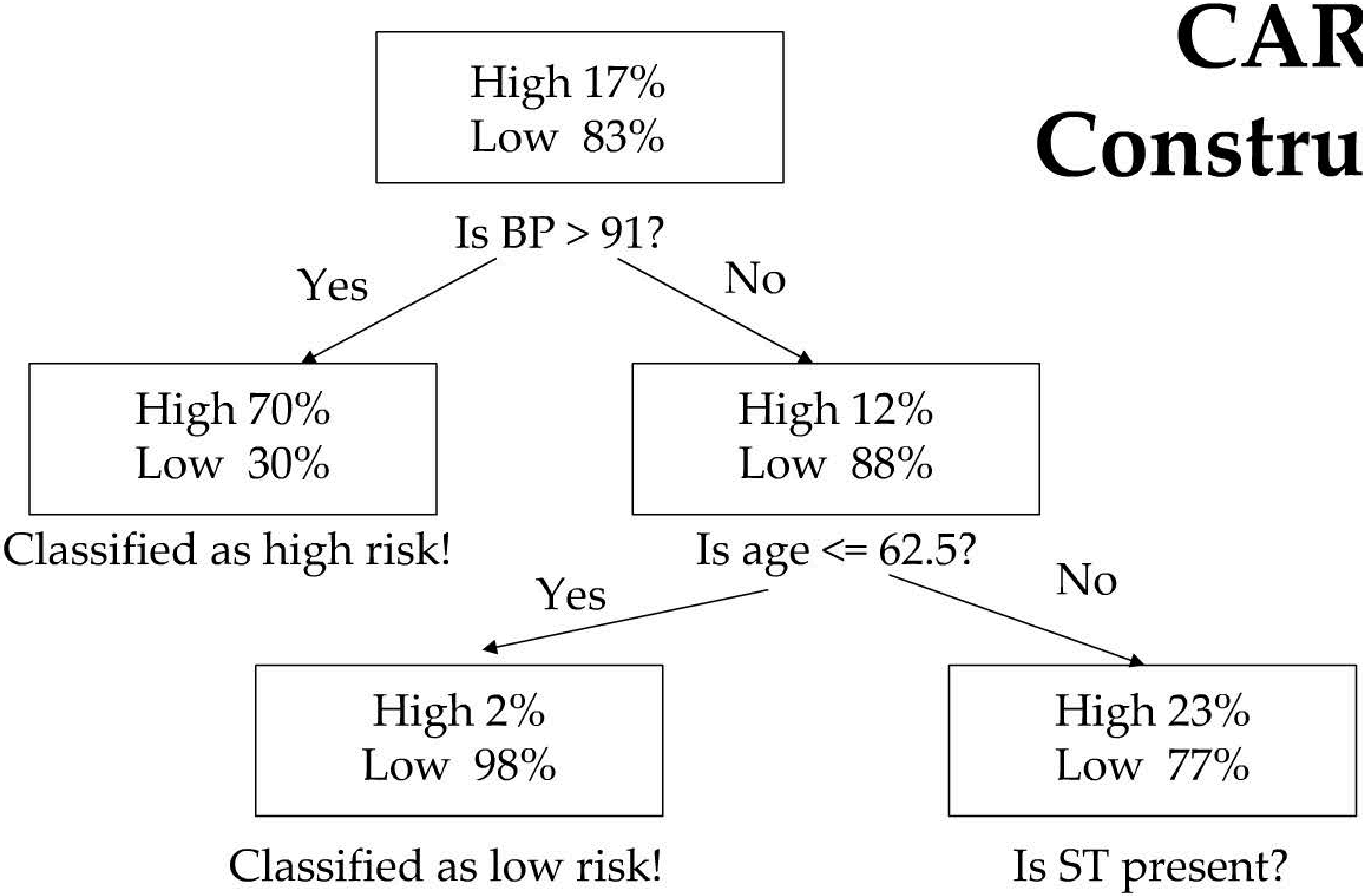
No

High 2%
Low 98%

Classified as low risk!

High 23%
Low 77%

CART Construction



High 17%
Low 83%

Is BP > 91?

Yes

No

High 70%
Low 30%

Classified as high risk!

High 12%
Low 88%

Is age <= 62.5?

Yes

No

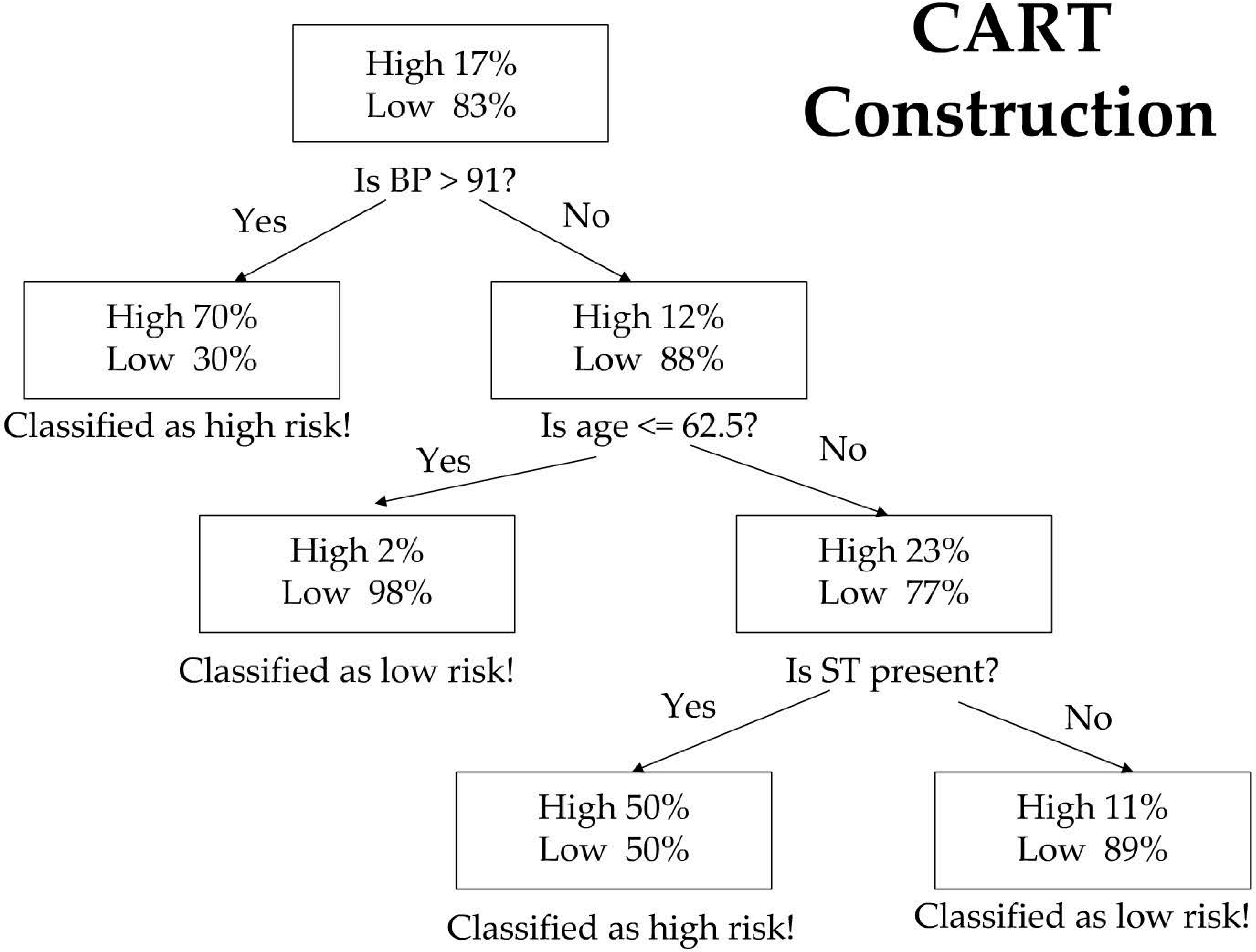
High 2%
Low 98%

Classified as low risk!

High 23%
Low 77%

Is ST present?

CART Construction



High 17%
Low 83%

Is BP > 91?

Yes

No

High 70%
Low 30%

Classified as high risk!

High 12%
Low 88%

Is age <= 62.5?

Yes

No

High 2%
Low 98%

Classified as low risk!

High 23%
Low 77%

Is ST present?

Yes

No

High 50%
Low 50%

Classified as high risk!

High 11%
Low 89%

Classified as low risk!

CART Construction

- Binary

  -- split parent node into two child nodes

- Recursive

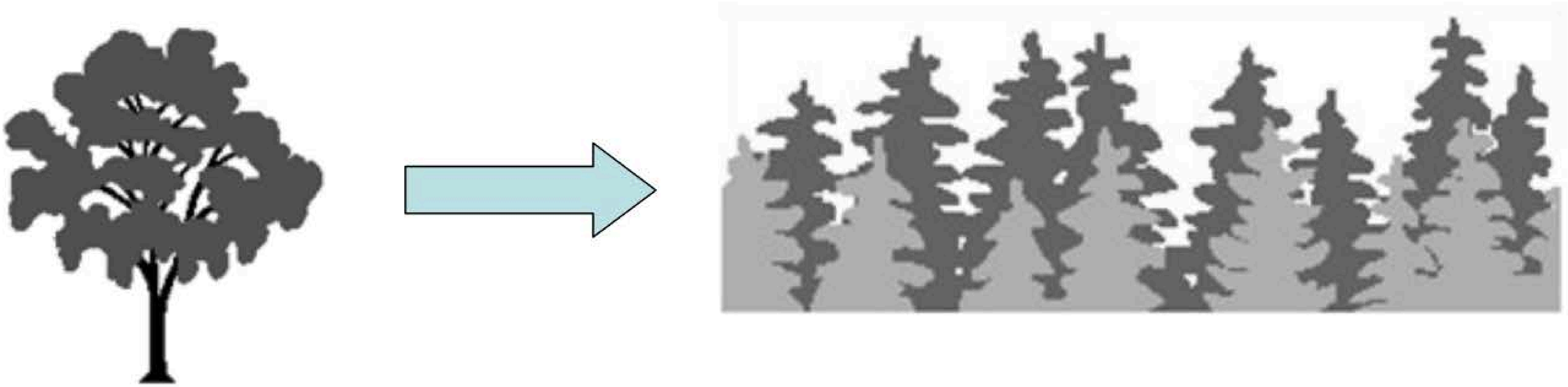
  -- each child node can be treated as parent node

- Partitioning

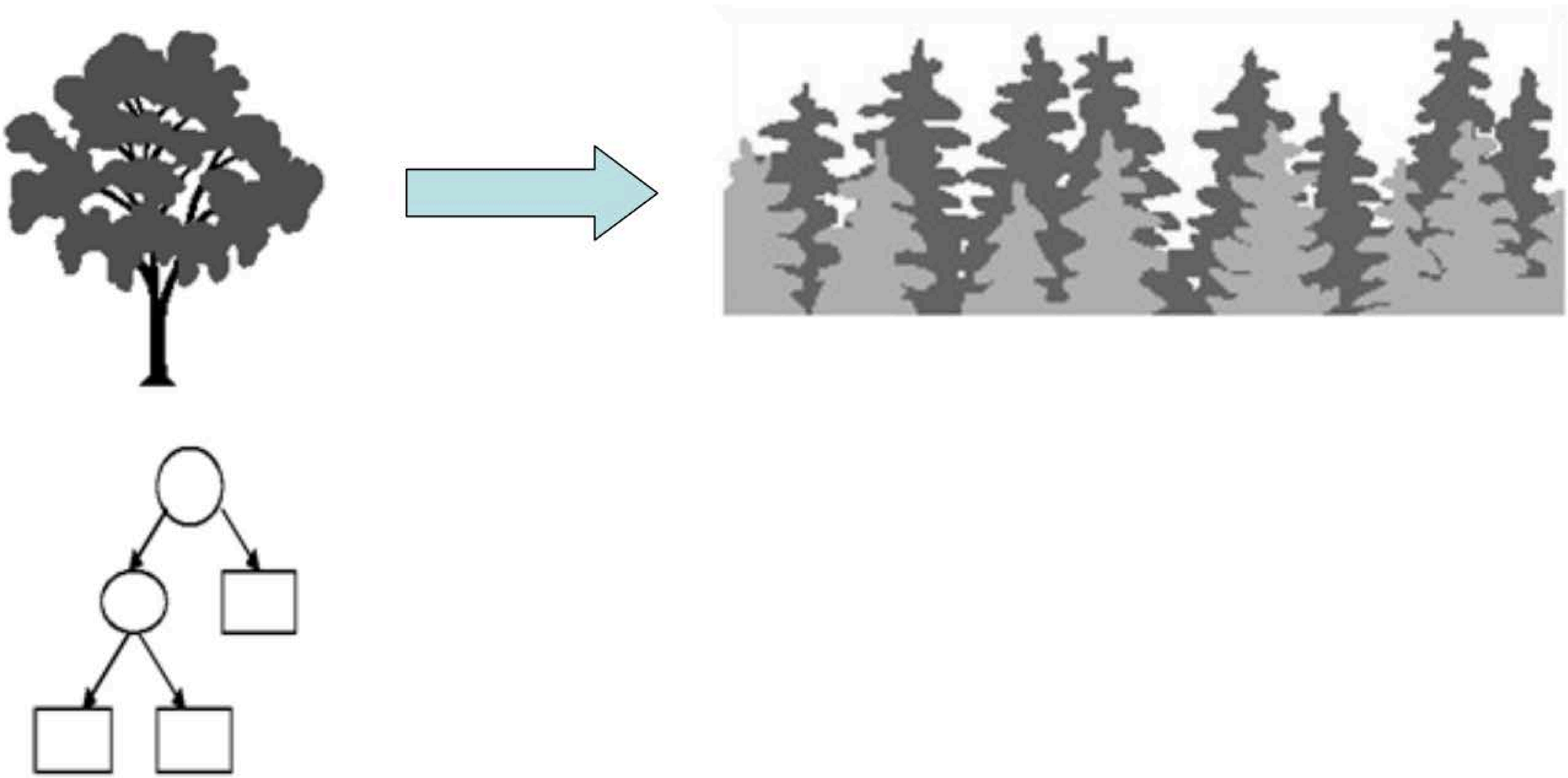
  -- data set is partitioned into mutually exclusive subsets in each split

RF Construction

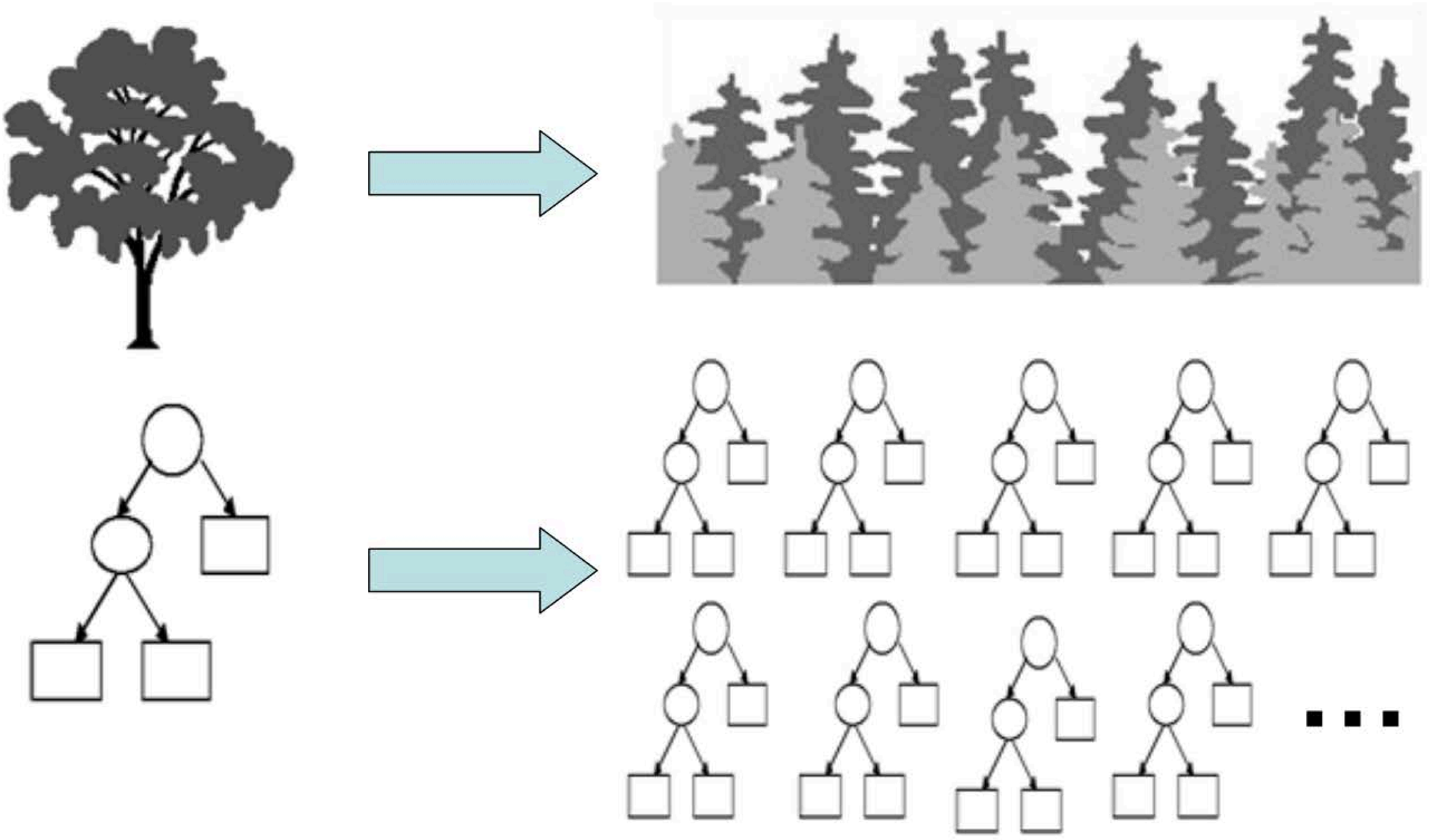
RF Construction



RF Construction

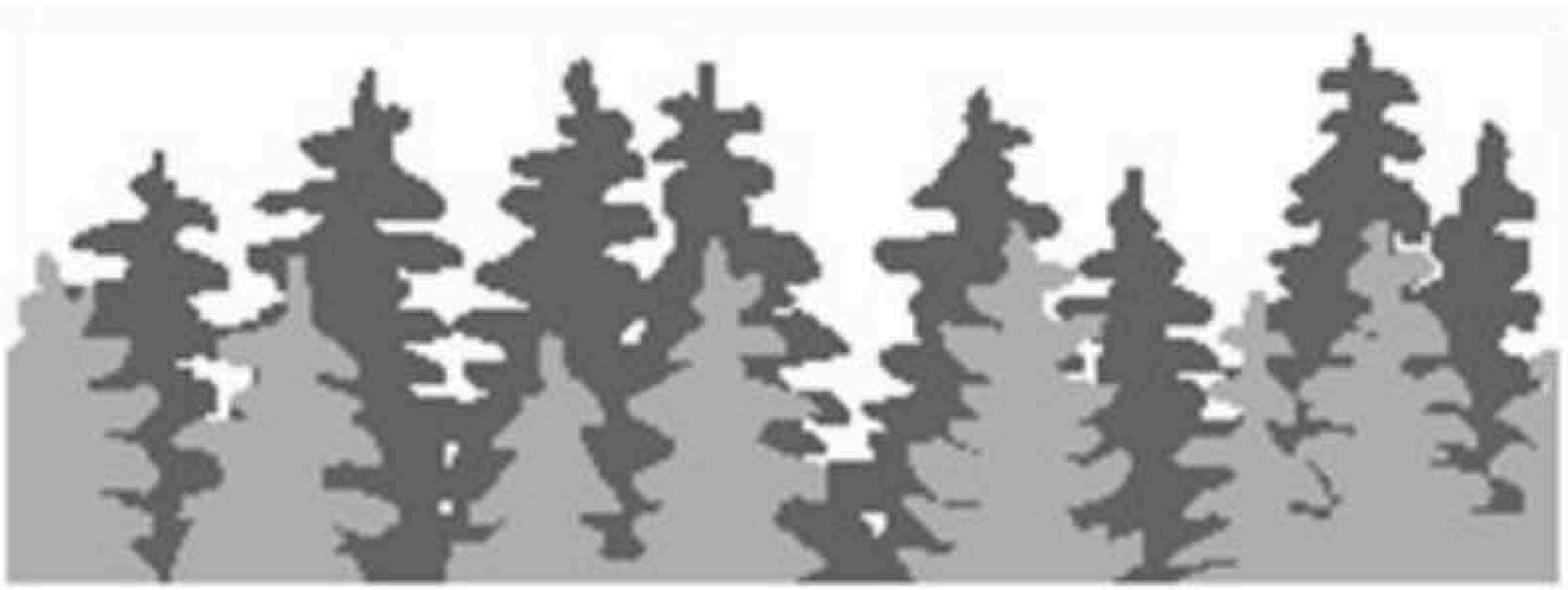


RF Construction



Random Forest (RF)

- An RF is a collection of tree predictors such that each tree depends on the values of an independently sampled random vector.



Prediction by plurality voting

- The forest consists of N trees

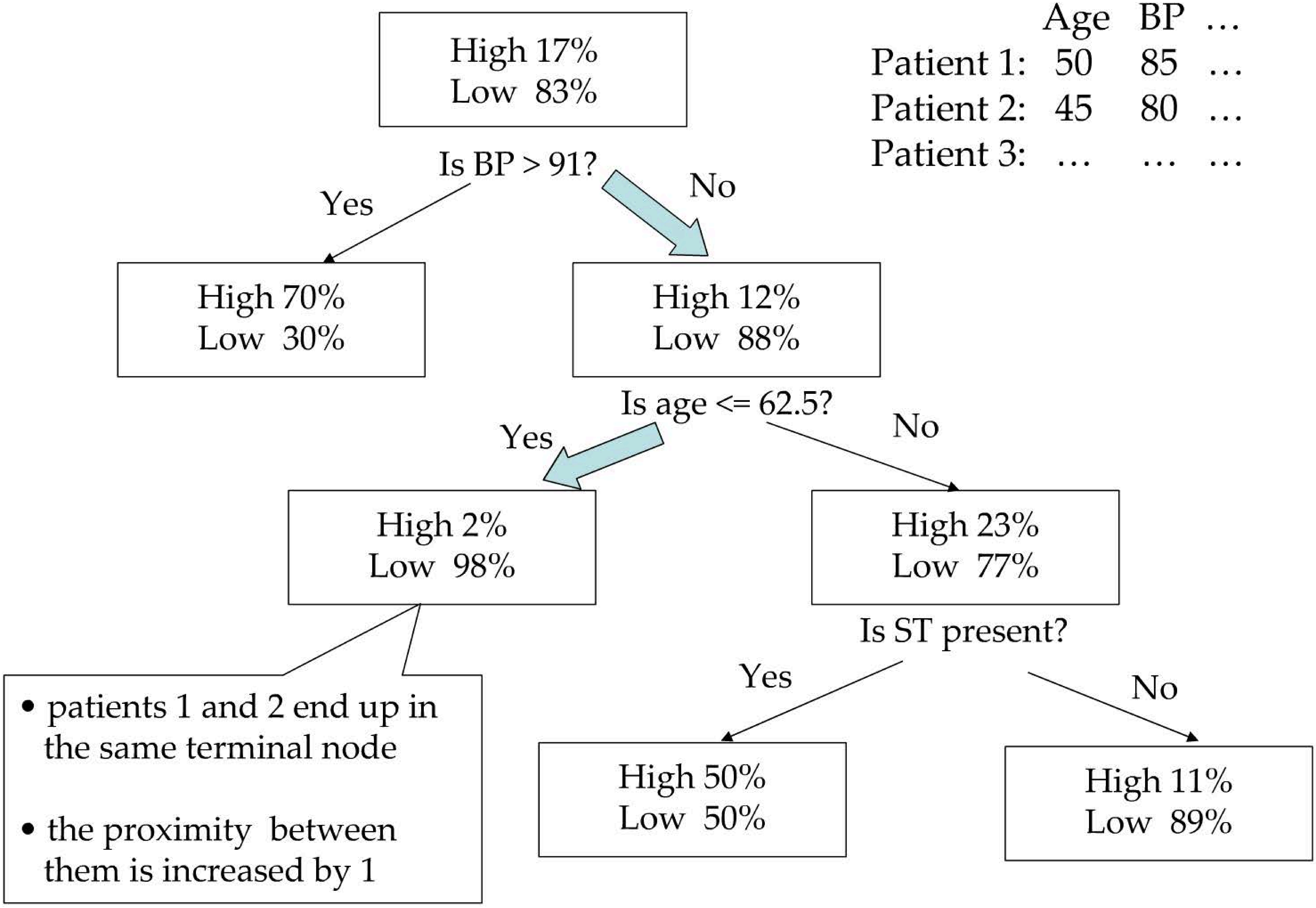
- Class prediction:
  - Each tree votes for a class; the predicted class C for an observation is the plurality, $\max_C \Sigma_k [f_k(\boldsymbol{x},\boldsymbol{T}) = C]$

Random forest predictors give rise to a dissimilarity measure

Intrinsic Similarity Measure

- Terminal tree nodes contain few observations
- If case *i* and case *j* both land in the same terminal node, increase the similarity between *i* and *j* by 1.
- At the end of the run divide by 2 x no. of trees.
- Dissimilarity = sqrt(1-Similarity)

	Age	BP	…
Patient 1:	50	85	…
Patient 2:	45	80	…
Patient 3:	…	…	…



- patients 1 and 2 end up in the same terminal node
- the proximity between them is increased by 1

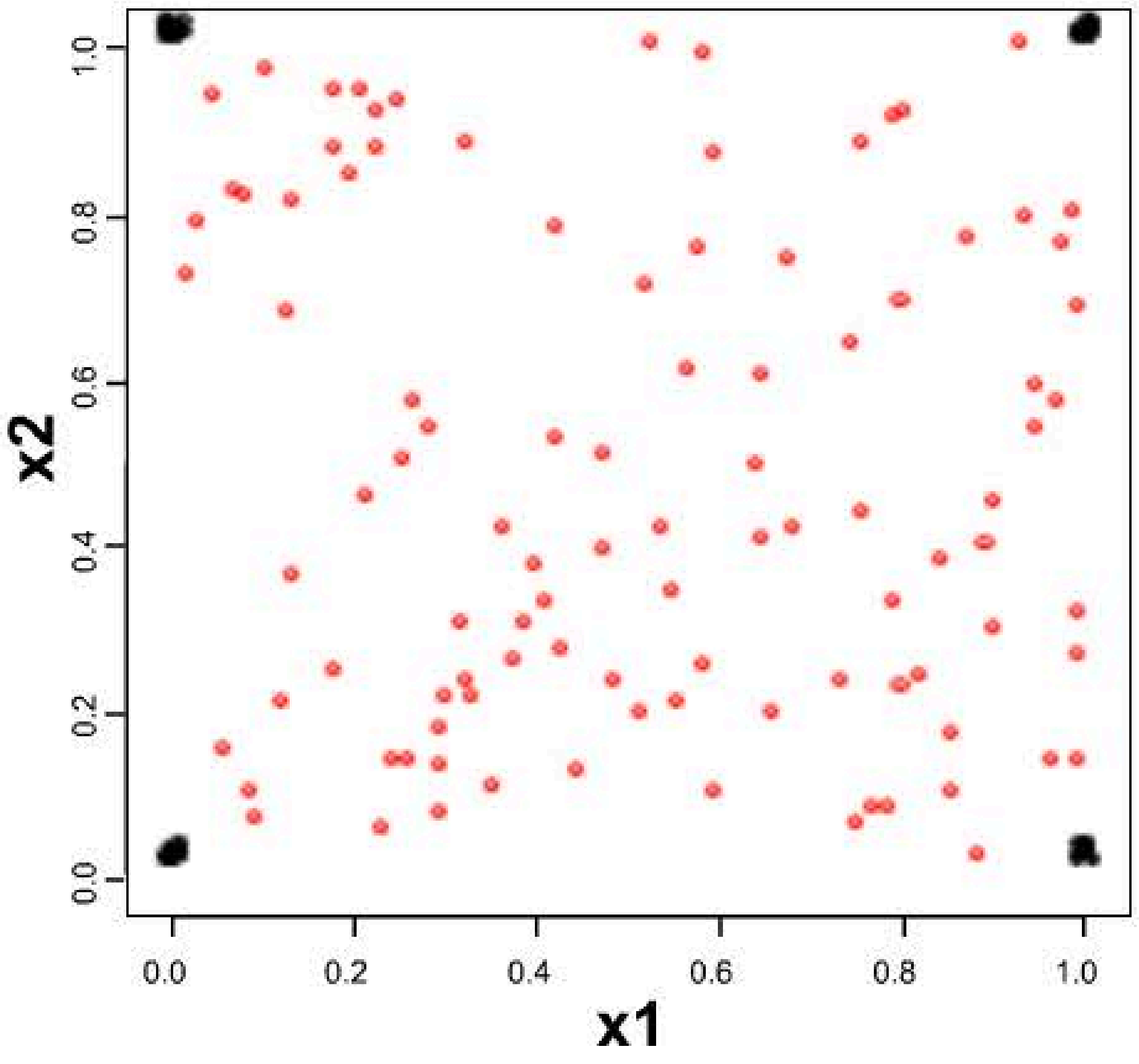
Unsupervised problem as a Supervised problem (RF implementation)

- Key Idea (Breiman 2003)
  - Label observed data as class 1
  - Generate synthetic observations and label them as class 2
  - Construct a RF predictor to distinguish class 1 from class 2
  - Use the resulting dissimilarity measure in unsupervised analysis

Two standard ways of generating synthetic covariates

- independent sampling from each of the univariate distributions of the variables (**Addcl1 =independent marginals**).
- independent sampling from uniforms such that each uniform has range equal to the range of the corresponding variable (**Addcl2**).

The scatter plot of original (black) and synthetic (red) data based on Addcl2 sampling.



RF clustering

- Compute distance matrix from RF
  - distance matrix = sqrt(1-similarity matrix)
- Conduct partitioning around medoid (PAM) clustering analysis
  - input parameter = no. of clusters k

Understanding RF Clustering (Theoretical Studies)

Shi, T. and Horvath, S. (2005) "Unsupervised learning using random forest predictors" J. Comp. Graph. Stat

Abstract: Random forest dissimilarity

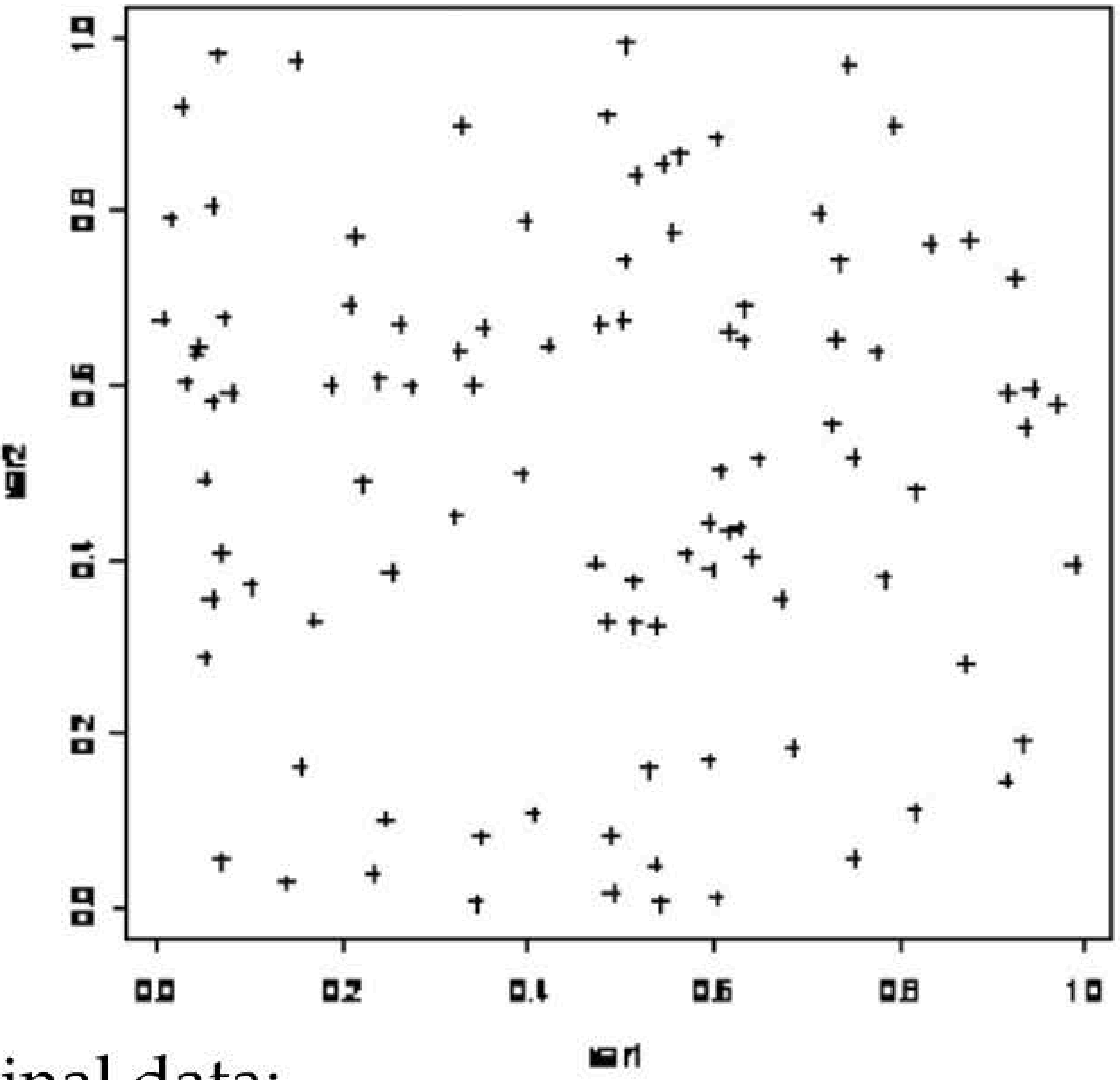
- Intrinsic variable selection focuses on dependent variables
  - Depending on the application, this can be attractive
- Resulting clusters can often be described using thresholding rules→attractive for TMA data.
- RF dissimilarity invariant to monotonic transformations of variables
- In some cases, the RF dissimilarity can be approximated using a Euclidean distance of ranked and scaled features.

- RF clustering was originally suggested by L. Breiman (RF manual). Theoretical properties are studied as part of the dissertation work of Tao Shi. Technical report and R code can be found at

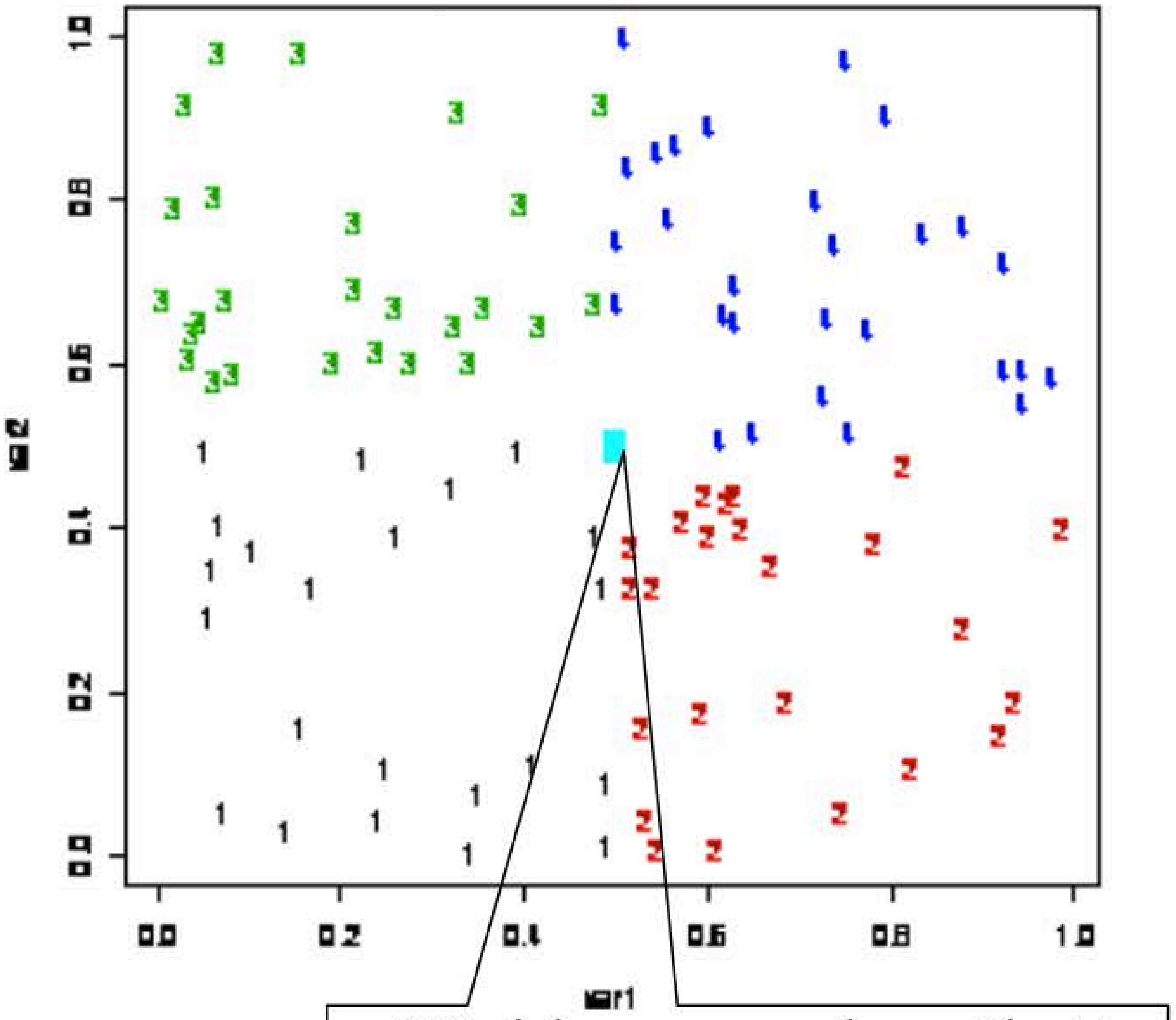
  www.genetics.ucla.edu/labs/horvath/RFclustering/RFclustering.htm
  www.genetics.ucla.edu/labs/horvath/kidneypaper/RCC.htm

Geometric interpretation of RF clusters

- RF cuts along the feature axes that isolate synthetic from observed observations will lead to clusters.



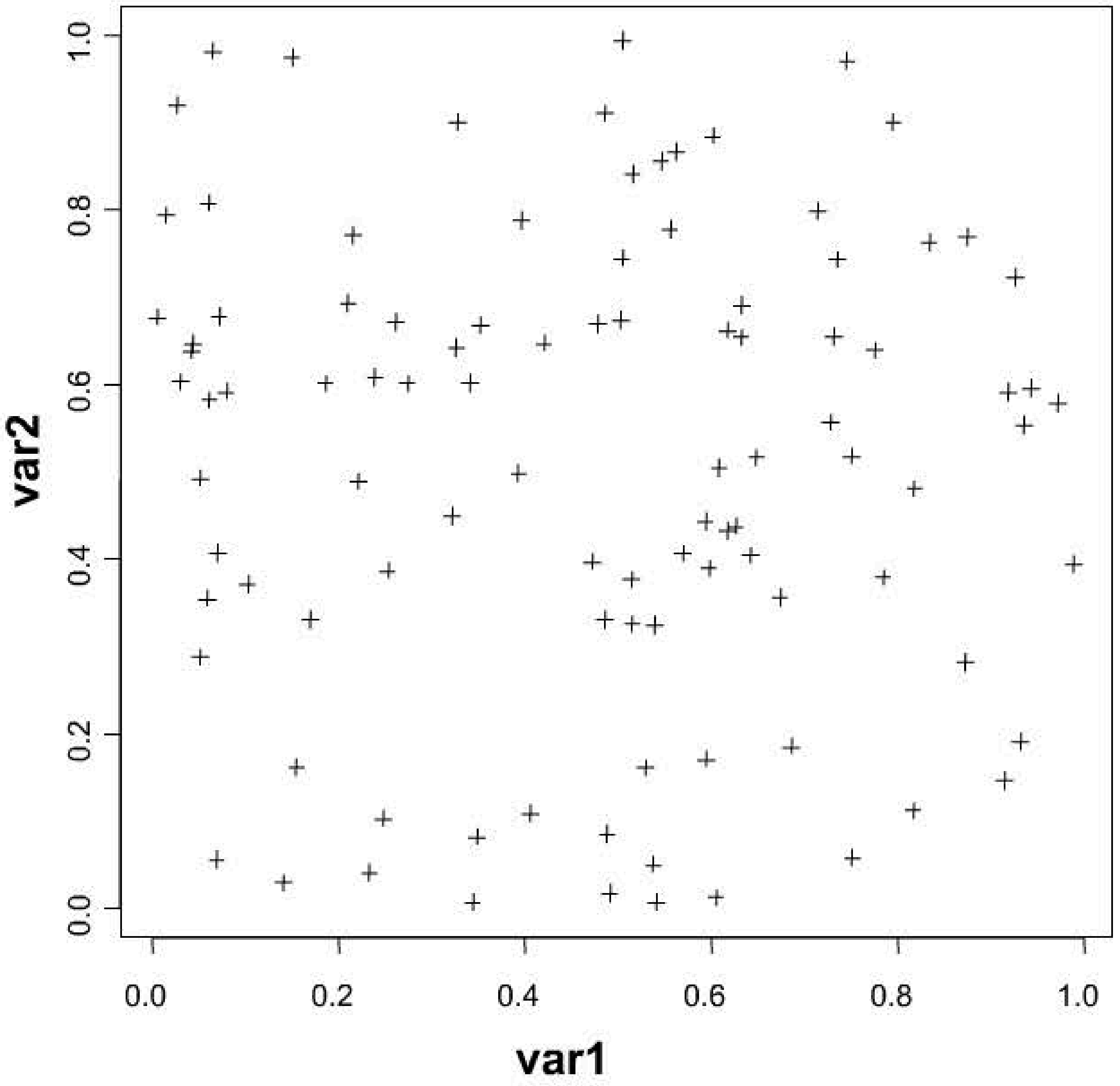
Original data:
no cluster structure according to Euclidean distance



Highly unusual synthetic data lead to 4 clusters

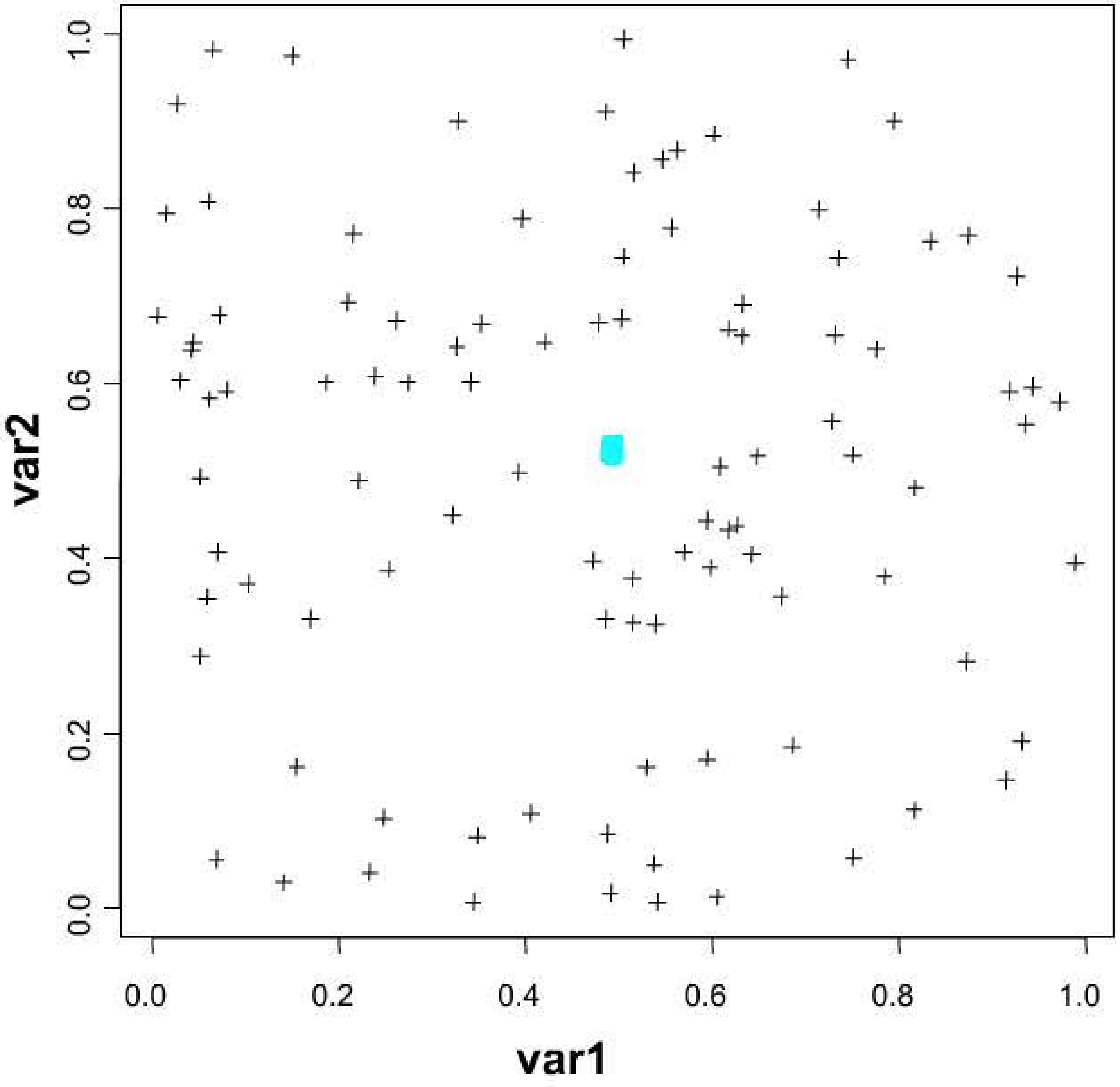
Geometric interpretation of RF clusters

- RF cuts along the feature axes that isolate synthetic from observed observations will lead to clusters.



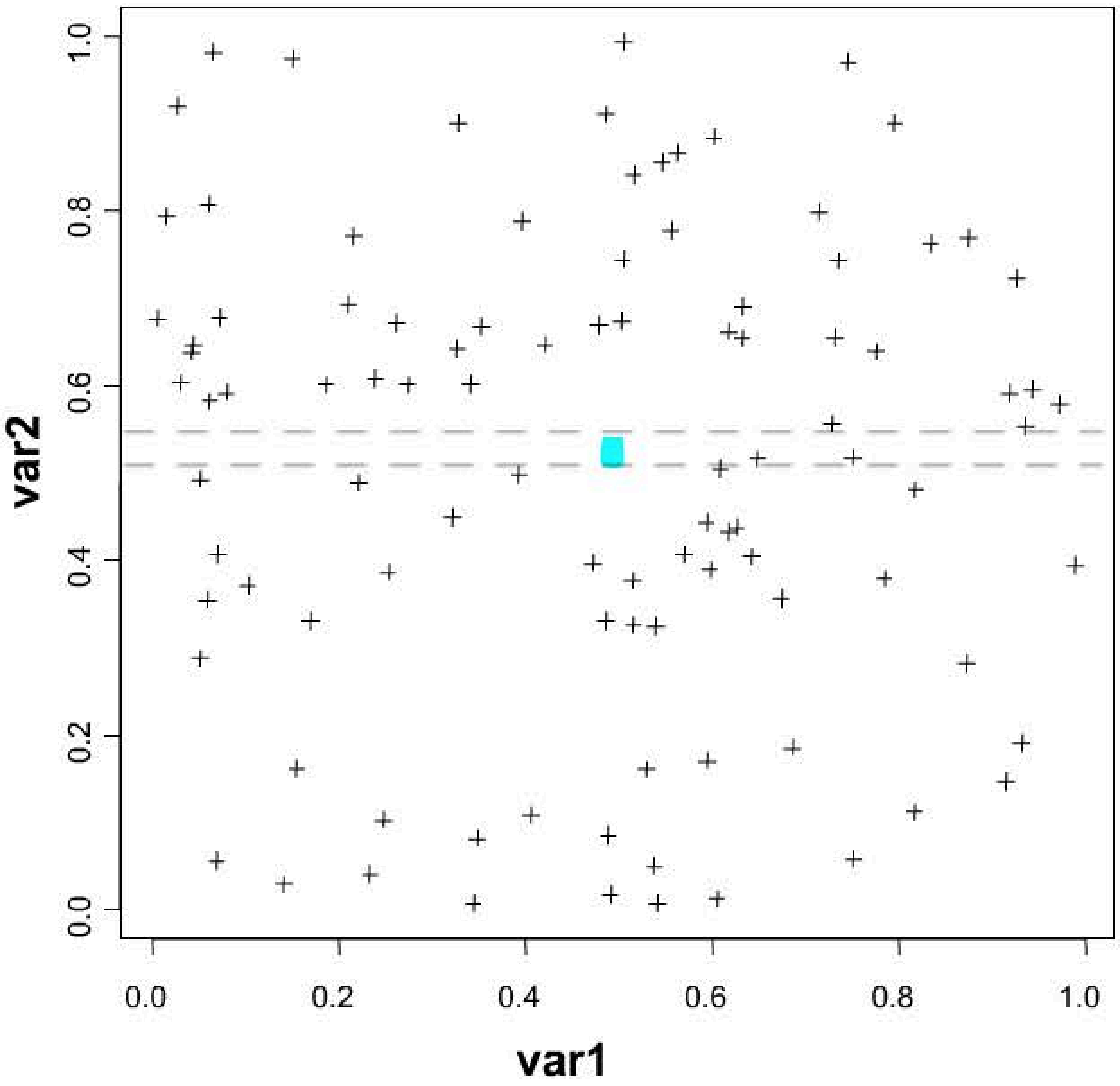
Geometric interpretation of RF clusters

- RF cuts along the feature axes that isolate synthetic from observed observations will lead to clusters.



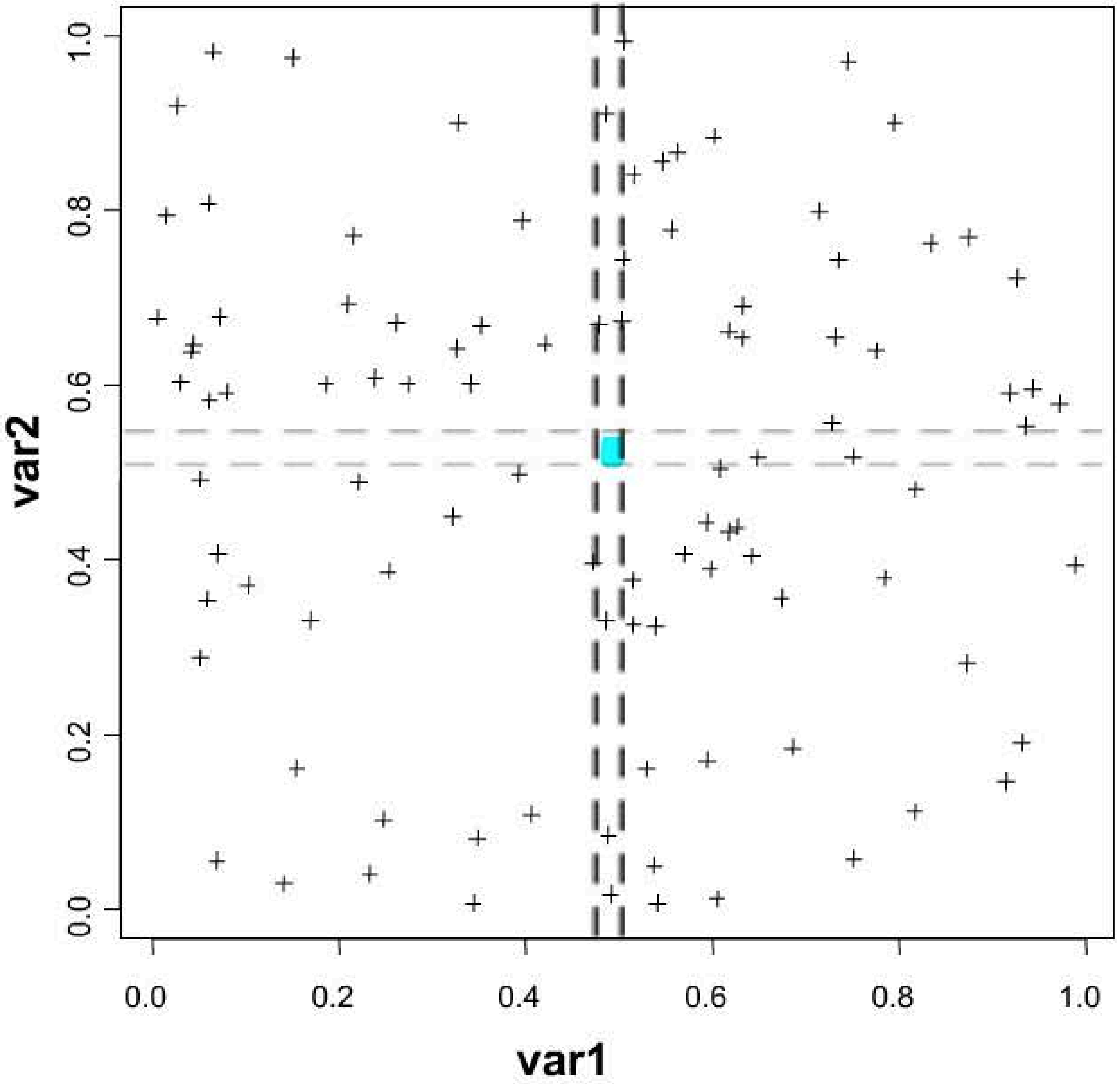
Geometric interpretation of RF clusters

- RF cuts along the feature axes that isolate synthetic from observed observations will lead to clusters.



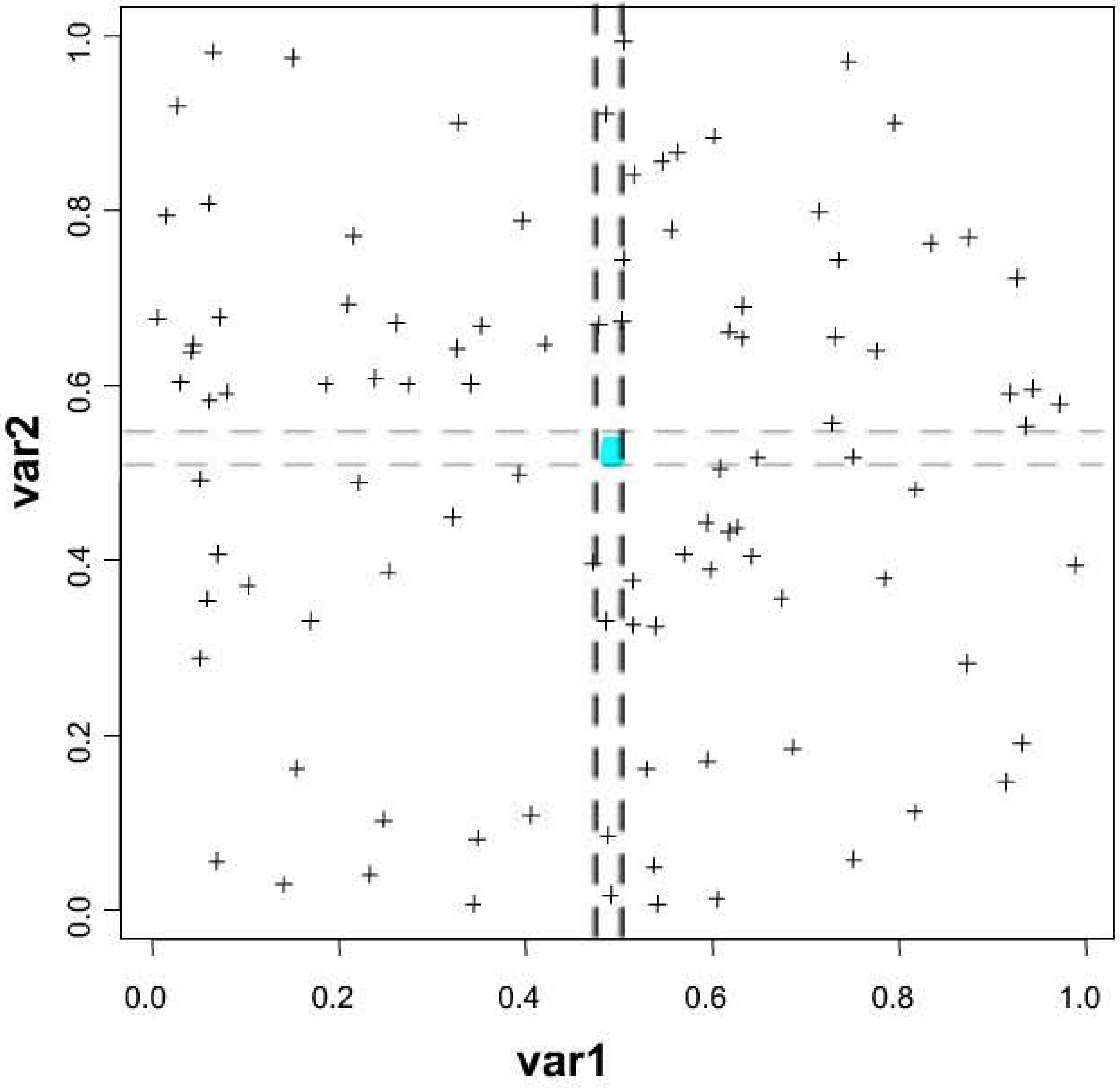
Geometric interpretation of RF clusters

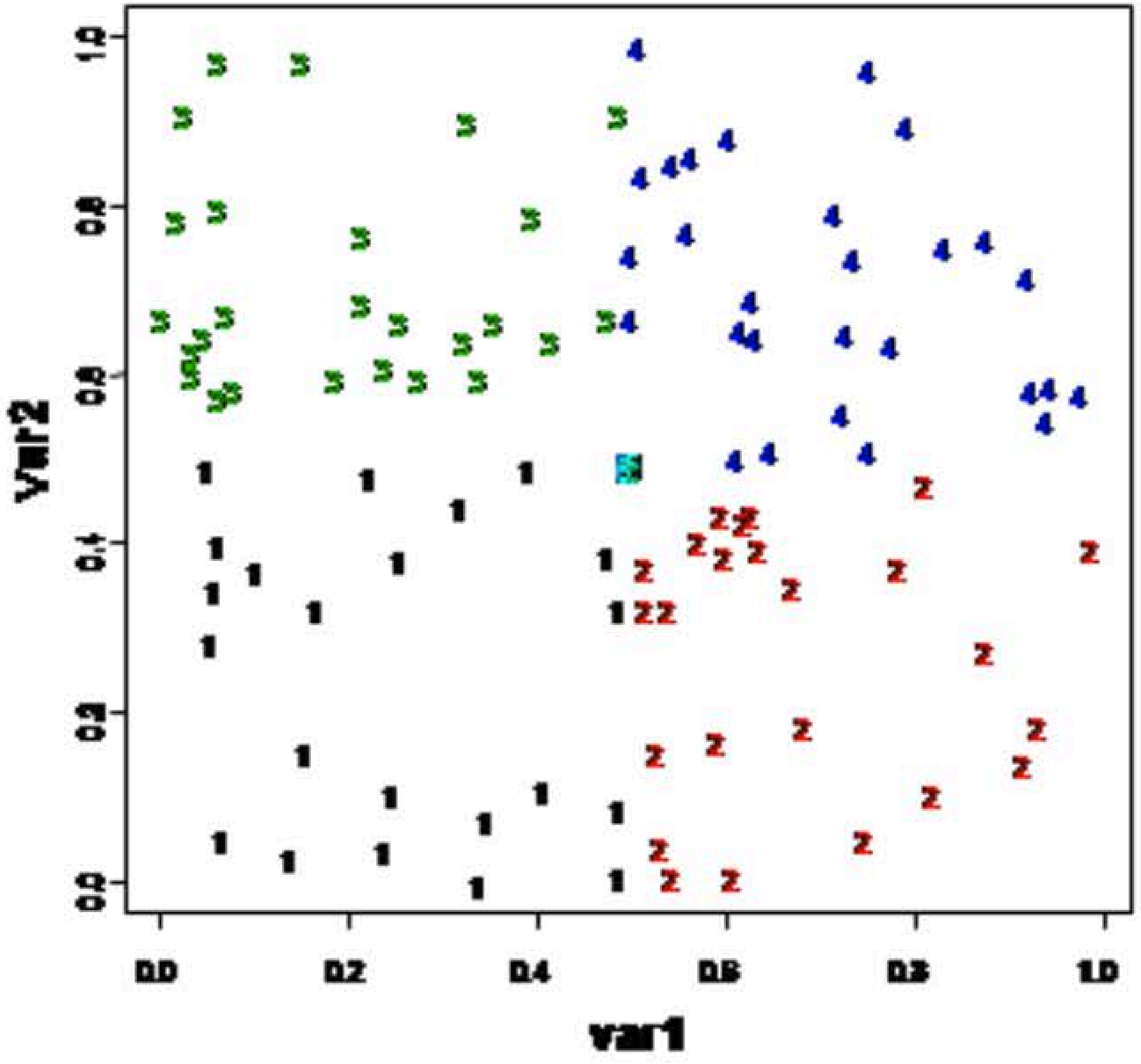
- RF cuts along the feature axes that isolate synthetic from observed observations will lead to clusters.



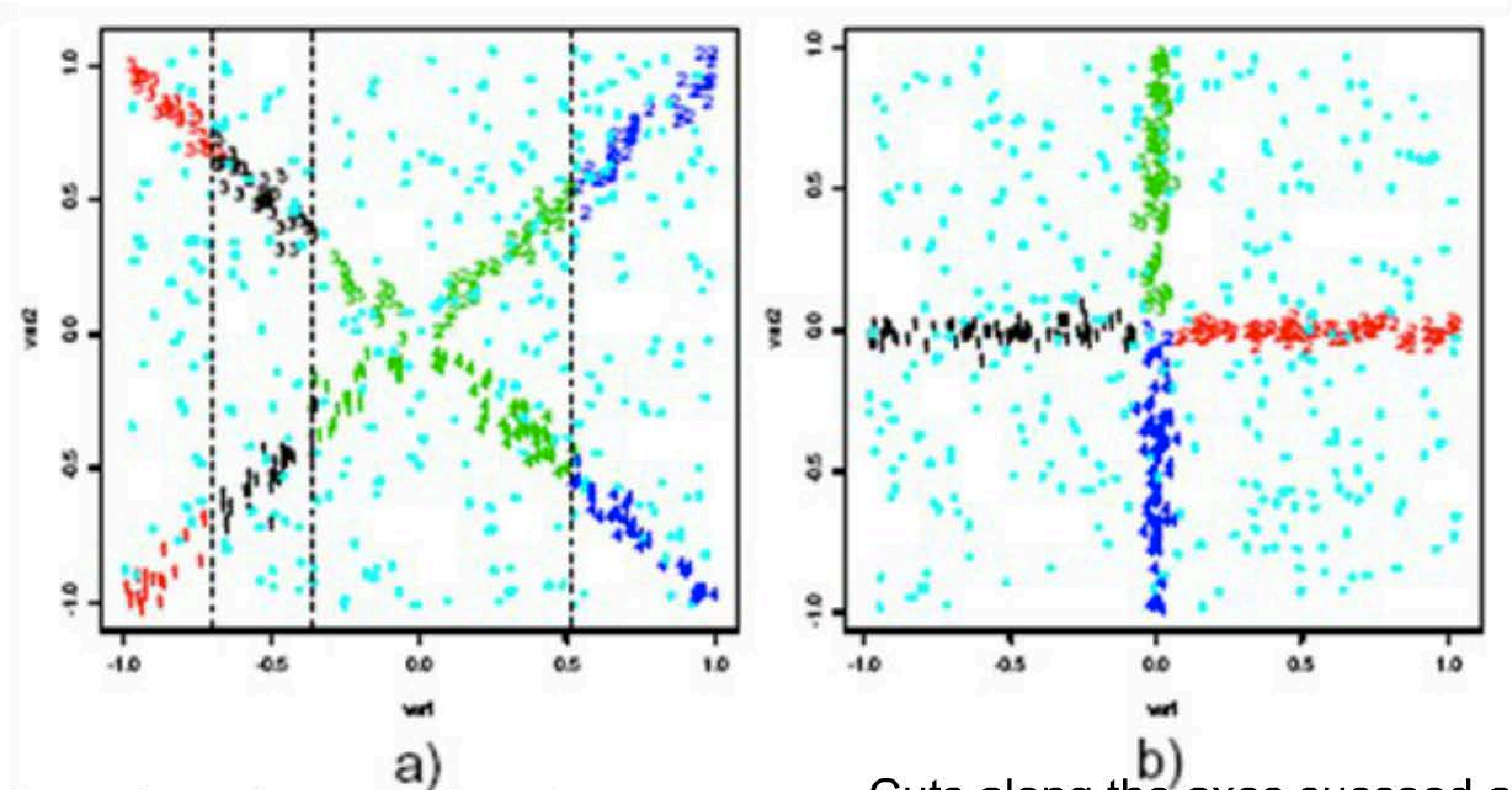
Geometric interpretation of RF clusters

- RF cuts along the feature axes that isolate synthetic from observed observations will lead to clusters.





RF clustering is not rotationally invariant



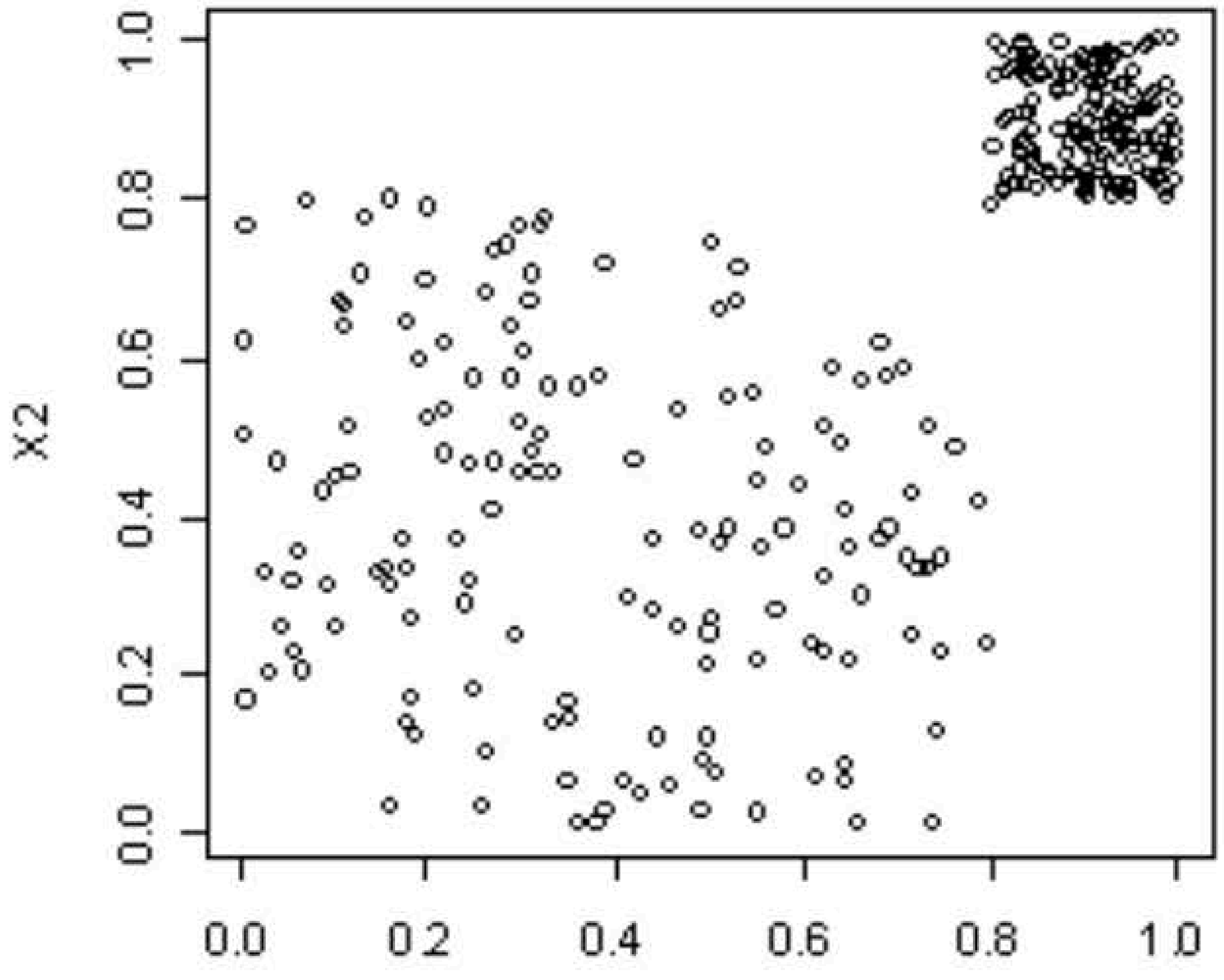
Cuts along the axes do not separate observed from synthetic (turquoise) data.

Cuts along the axes succeed at separating observed data from Synthetic data.

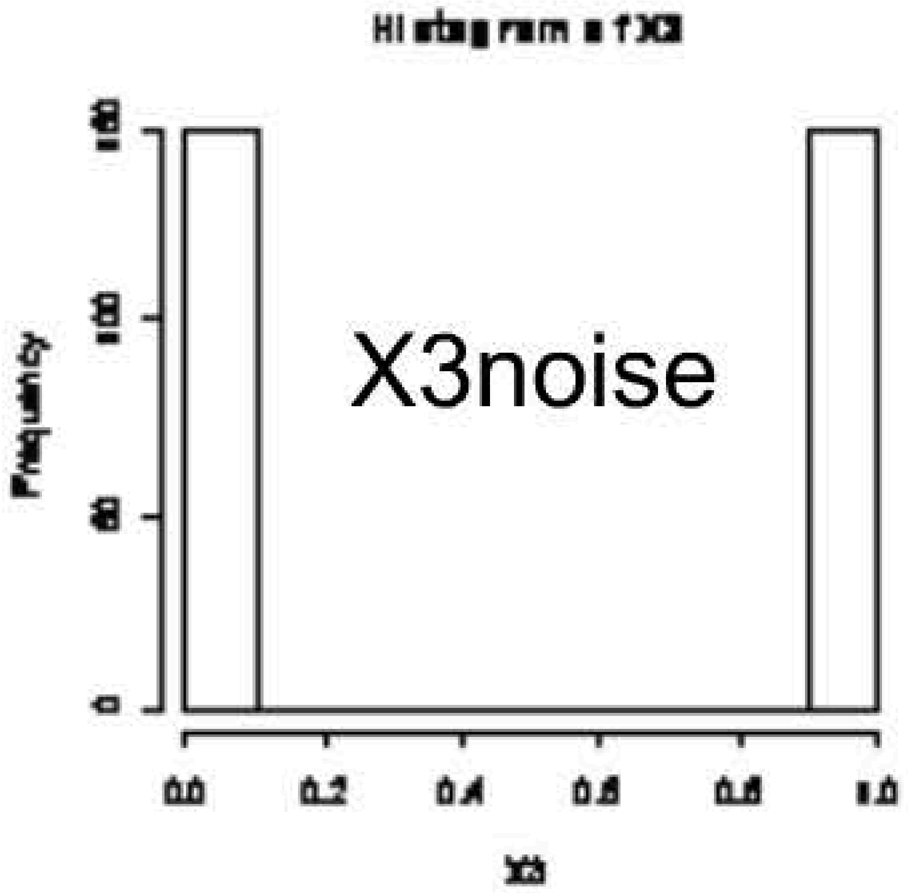
Simulated Example ExRule: contrast RF dissimilarity with Euclidean distance

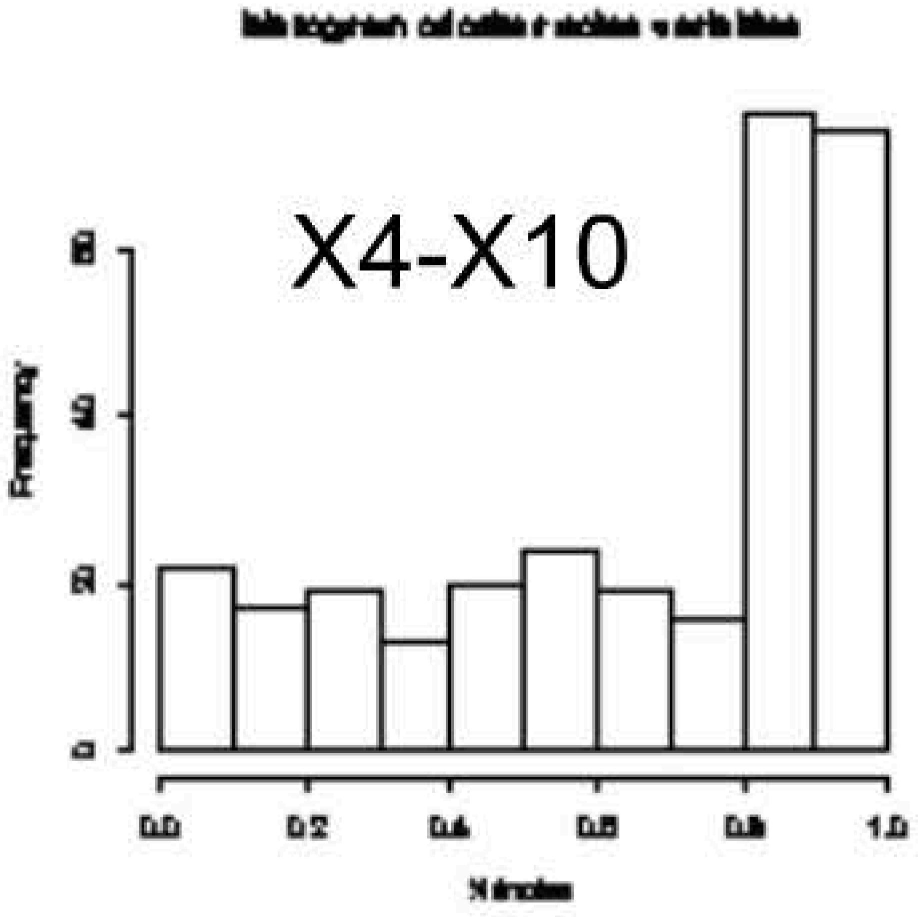
Simulated Cluster structure

Scatter plot of 2 signal variables



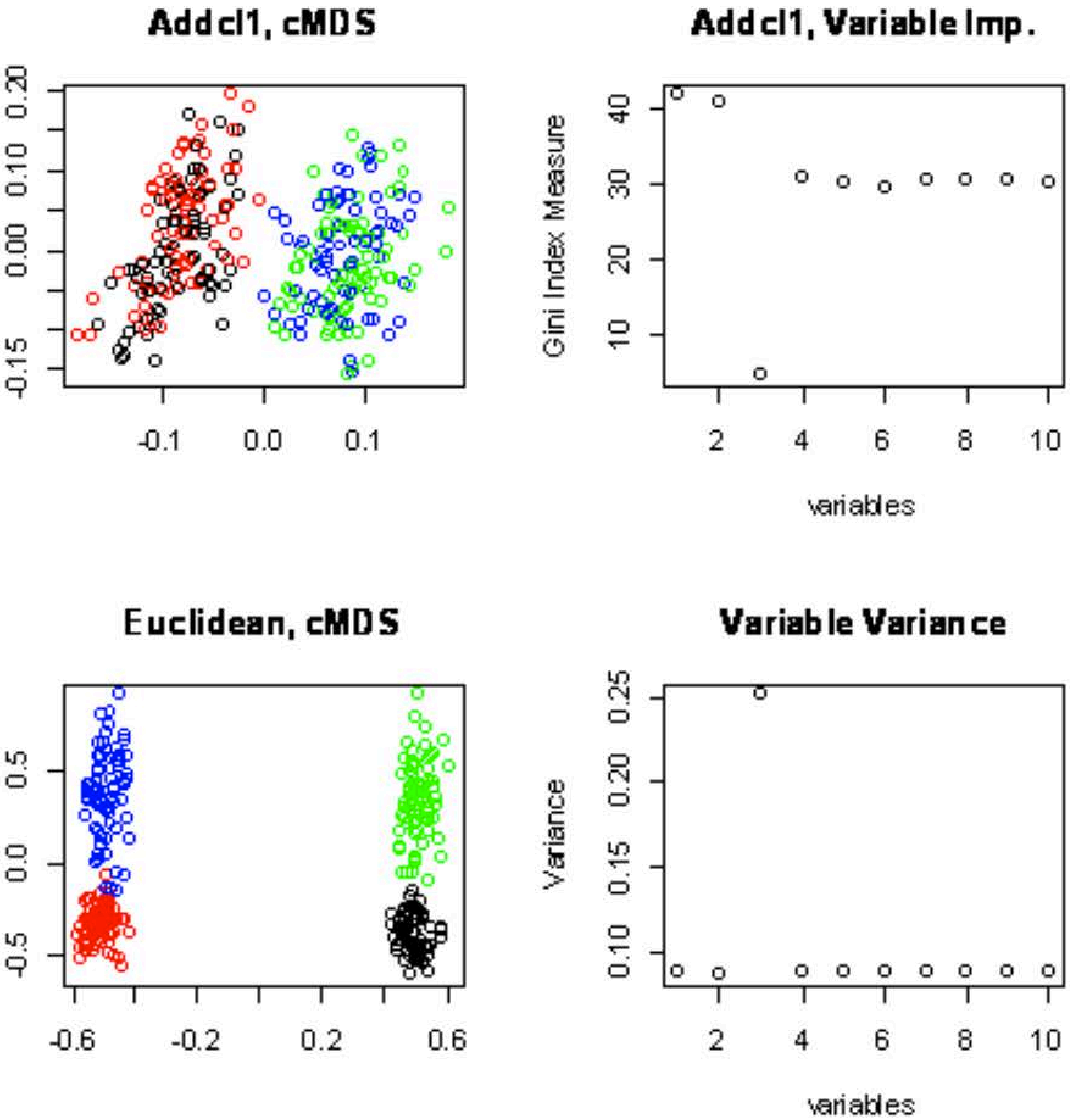
Histogram of noise variables





Cluster can be described by threshold rules.
150 observations in each cluster.

Example ExRule



Black	if X1>0.8 & X3=0
Red	if X1>0.8 & X3=1
Green	if X1<0.8 & X3=0
Blue	if X1<0.8 & X3=1

Message: RF clusters correspond to variable X1 while Euclidean clusters correspond to X3.

The clustering results for example ExRule

- Addcl1 dissimilarity focuses on most dependent variables→ clusters are determined by cuts along variables X1 and X2.
- Resulting clusters can be described using a simple thresholding rule.
- Euclidean distance focuses on most varying variable X3 → PAM clusters and MDS point clouds are driven by X3.

Typical Addcl2 Example

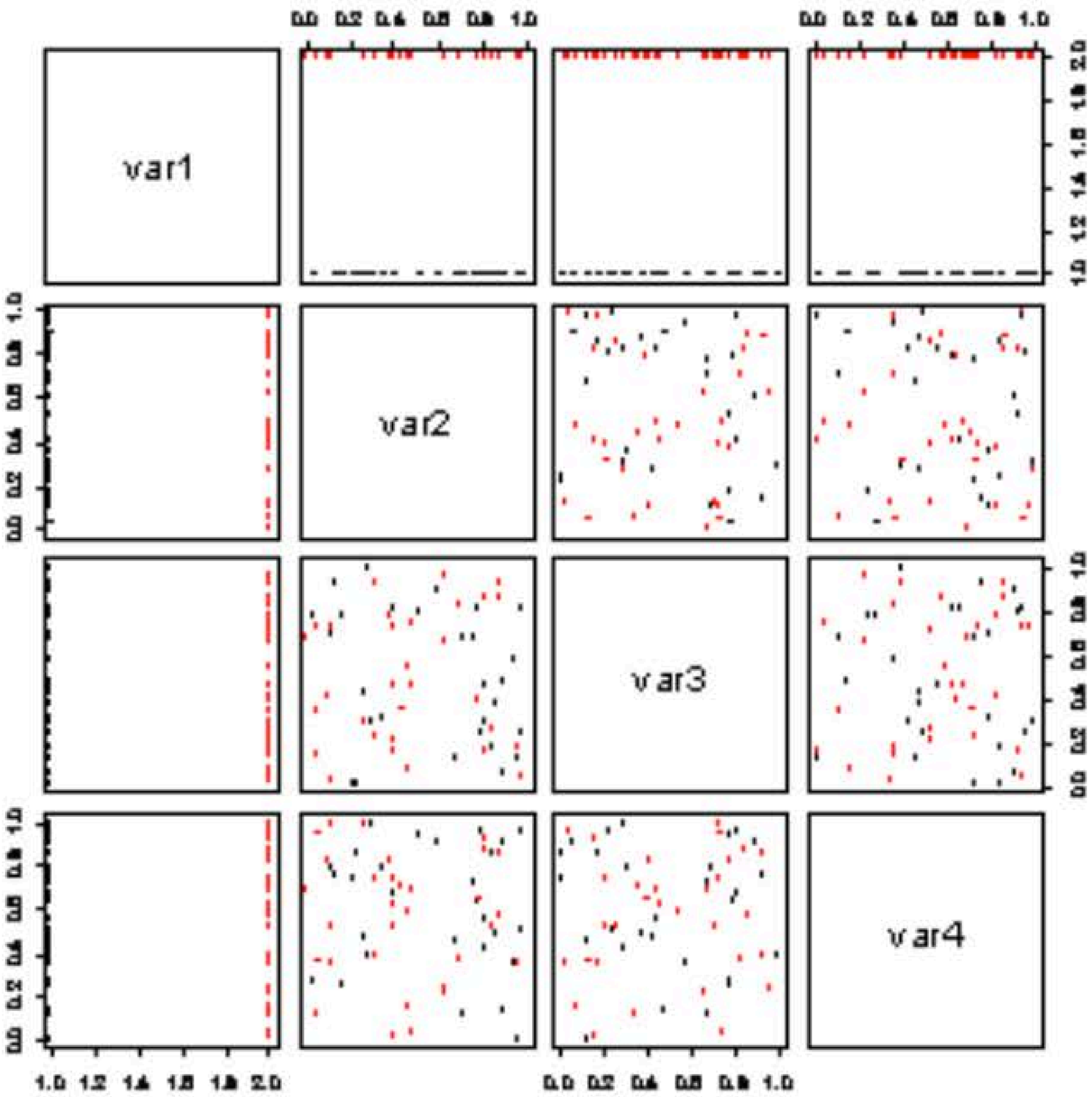
Few independent covariates contains cluster info (binary signal), rest are noise

Pairwise scatter plot

Example:

One binary variable

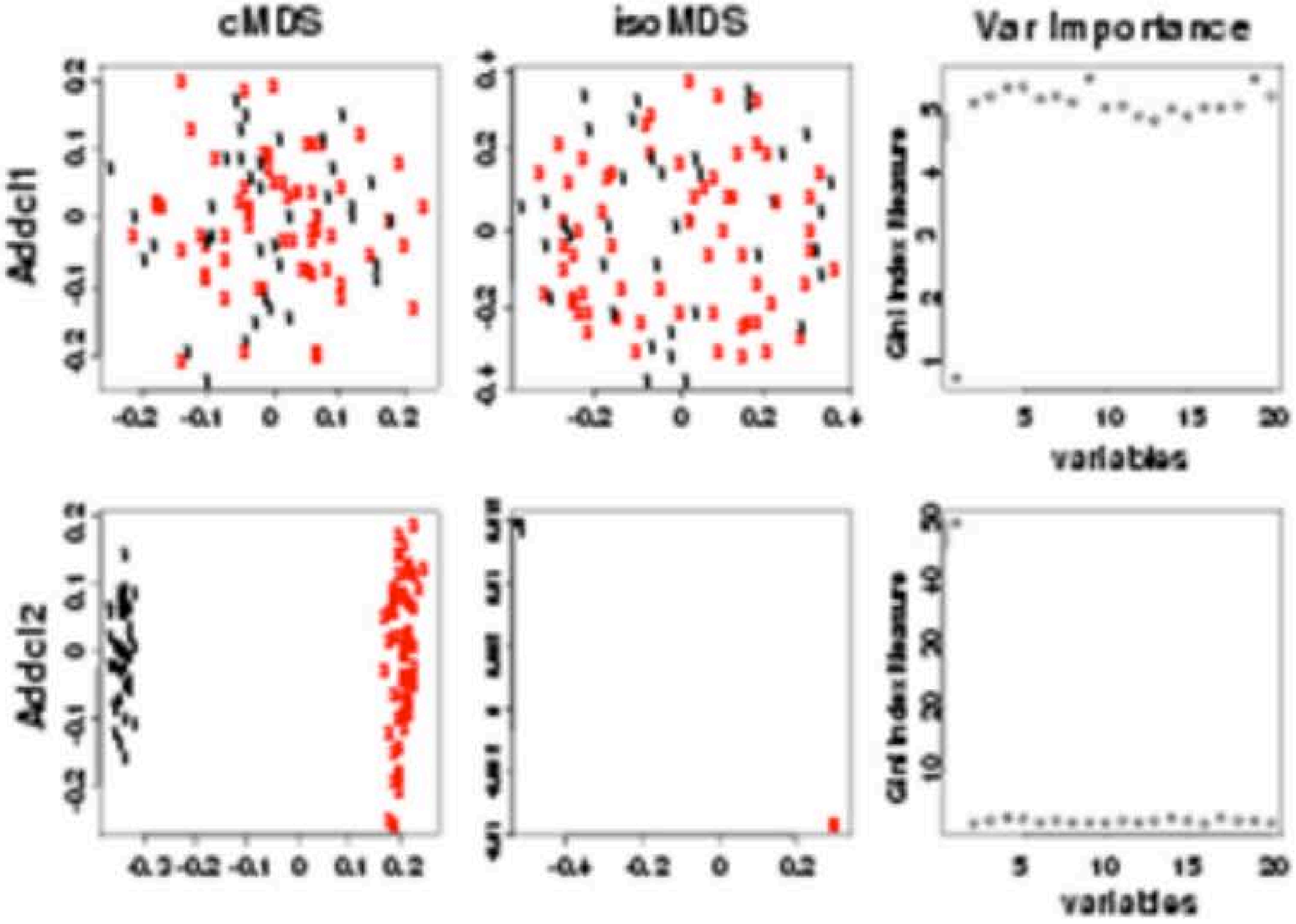
Rest random uniform



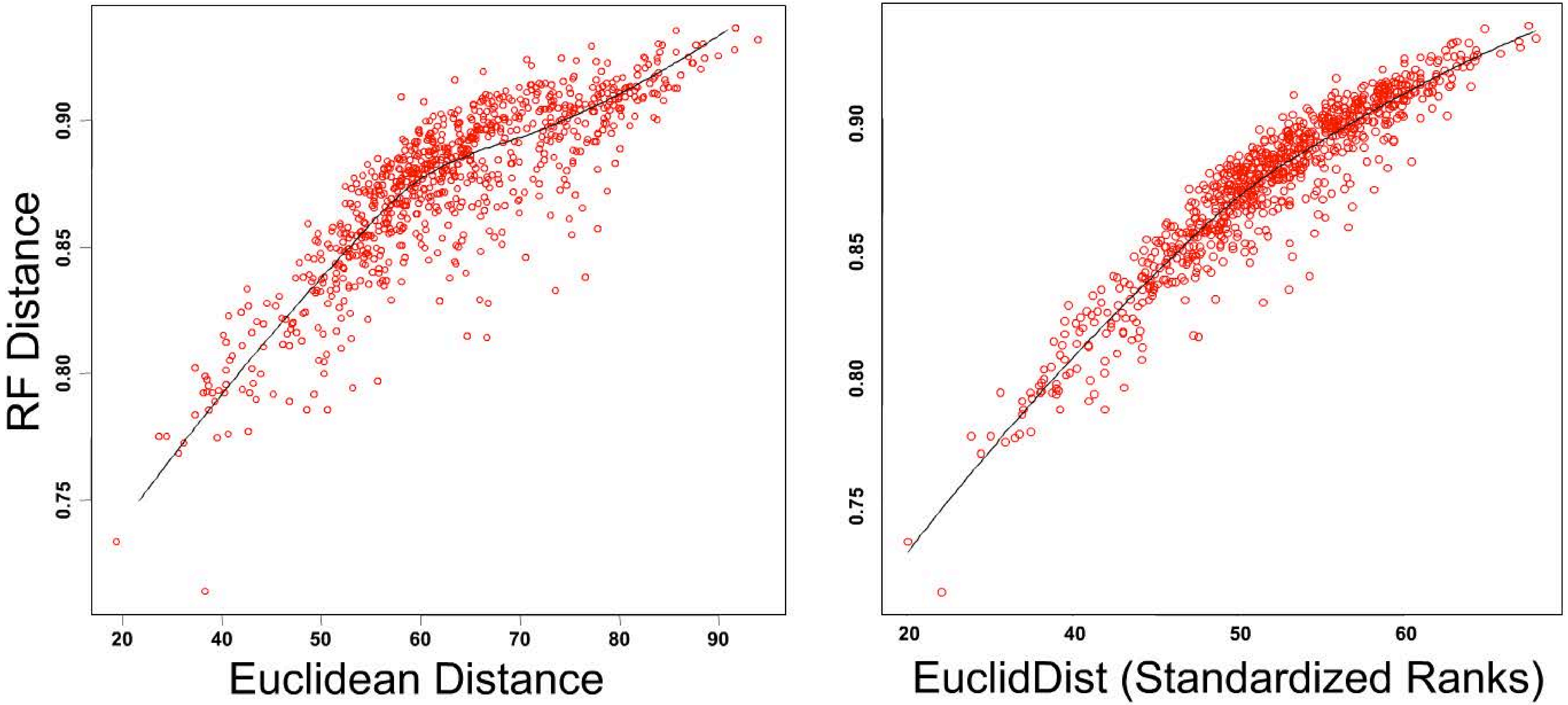
Nature of Addcl2 RF clustering

- Addcl1 completely fails.

- Addcl2 clustering works well



RF dissimilarity vs. Euclidean distance (DNA Microarray Data)



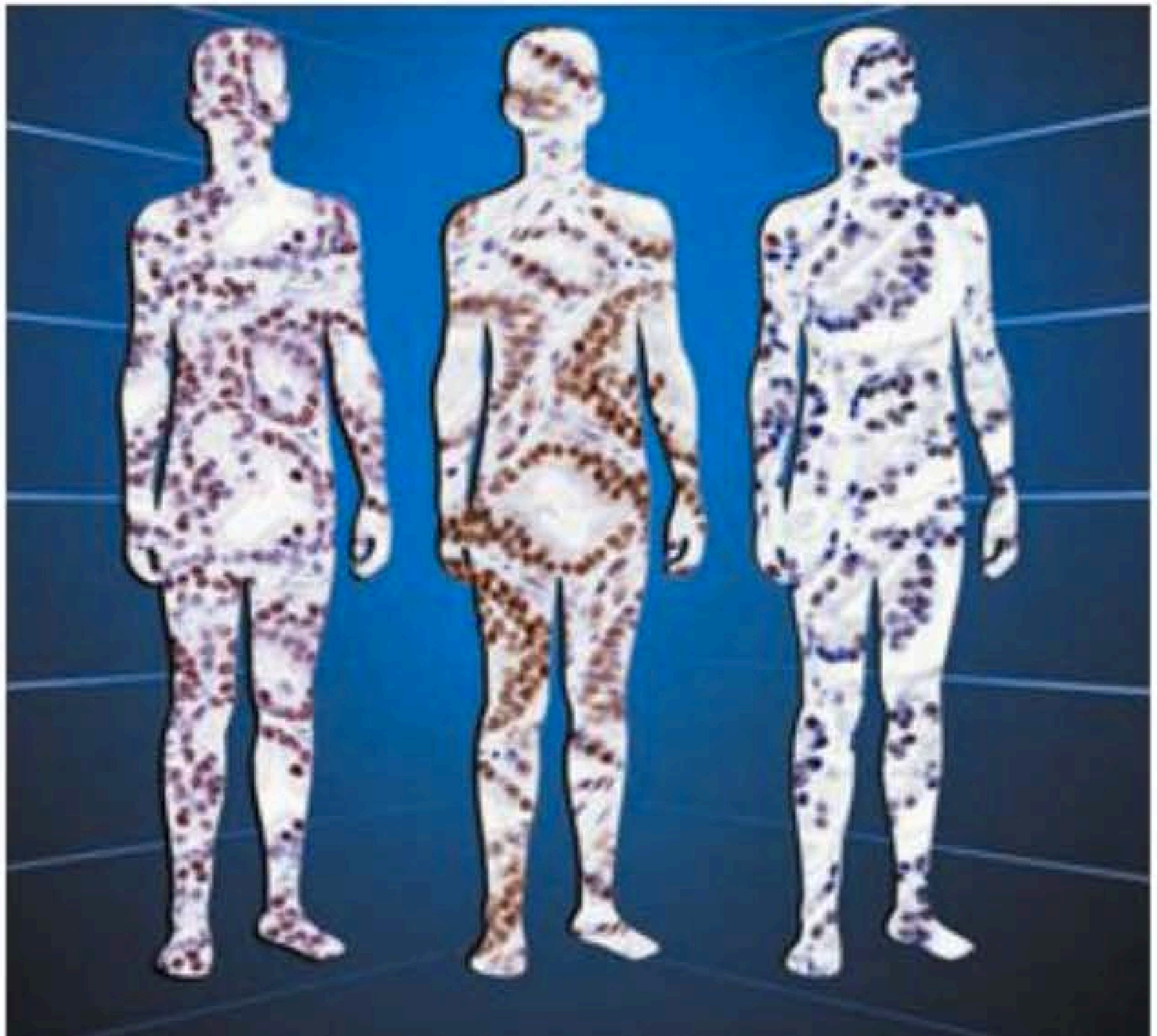
Theoretical reasons for using an RF dissimilarity for TMA data

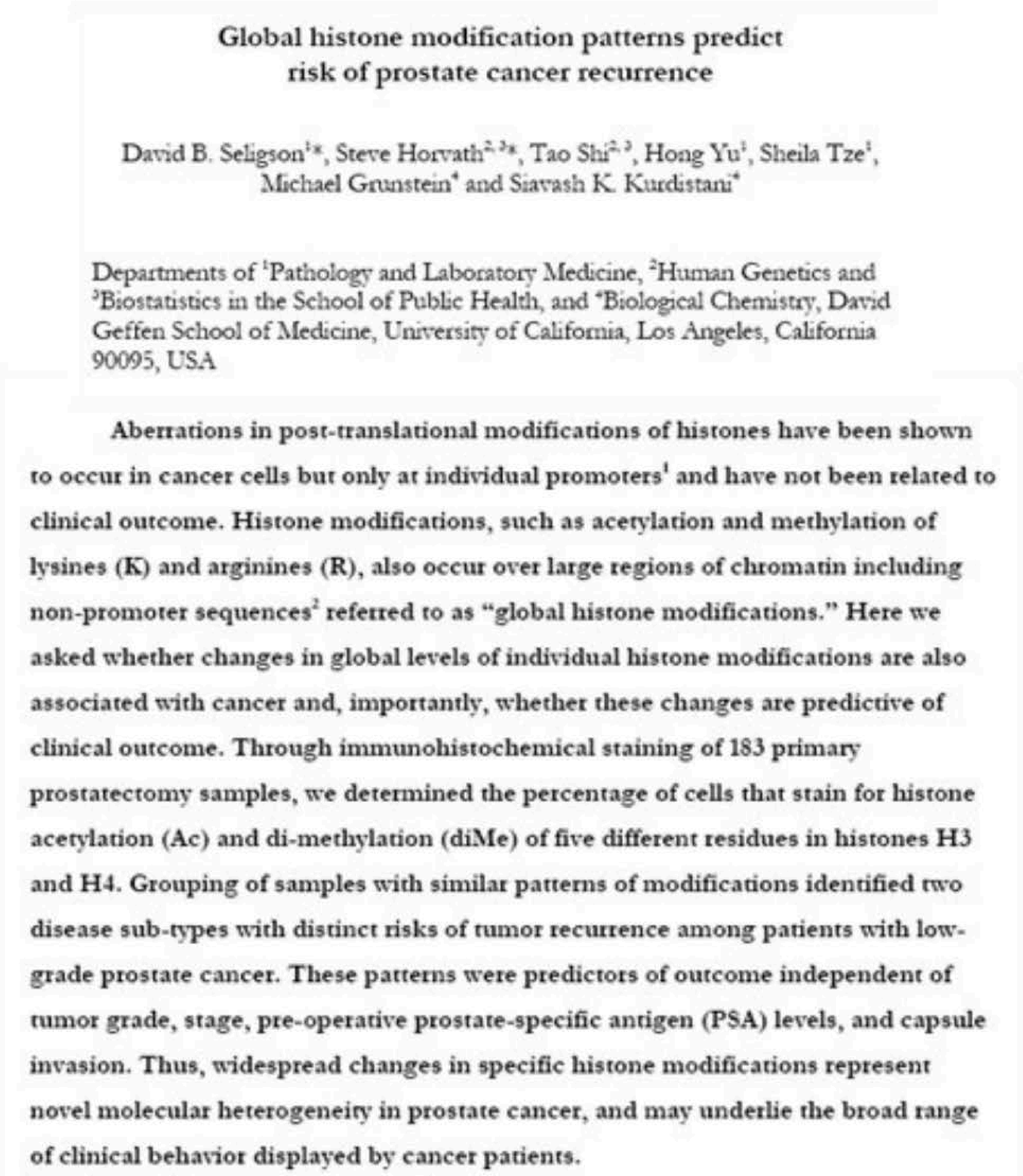
- Main reasons
  - natural way of weighing tumor marker contributions to the dissimilarity
    - The more related a tumor marker is to other tumor markers the more it contributes to the definition of the dissimilarity
  - no need to transform the often highly skewed features
    - based feature ranks
  - Chooses cut-off values automatically
  - resulting clusters can often be described using simple thresholding rules
- Other reasons
  - elegant way to deal with missing covariates
  - intrinsic proximity matrix handles mixed variable types well

- CAVEAT: The choice of the dissimilarity should be determined by the kind of patterns one hopes to find. There will be situations when other dissimilarities are preferrable.

Applications to prostate tissue microarray data

Seligson DB, Horvath S, Shi T, Yu H, Tze S, Grunstein M, Kurdistani SK (2005) Global histone modification patterns predict risk of prostate recurrence. Nature





Global histone modification patterns predict risk of prostate cancer recurrence

David B. Seligson[1*], Steve Horvath[2,3*], Tao Shi[2,3], Hong Yu[1], Sheila Tze[1],
Michael Grunstein[4] and Siavash K. Kurdistani[4]

Departments of [1]Pathology and Laboratory Medicine, [2]Human Genetics and [3]Biostatistics in the School of Public Health, and [4]Biological Chemistry, David Geffen School of Medicine, University of California, Los Angeles, California 90095, USA

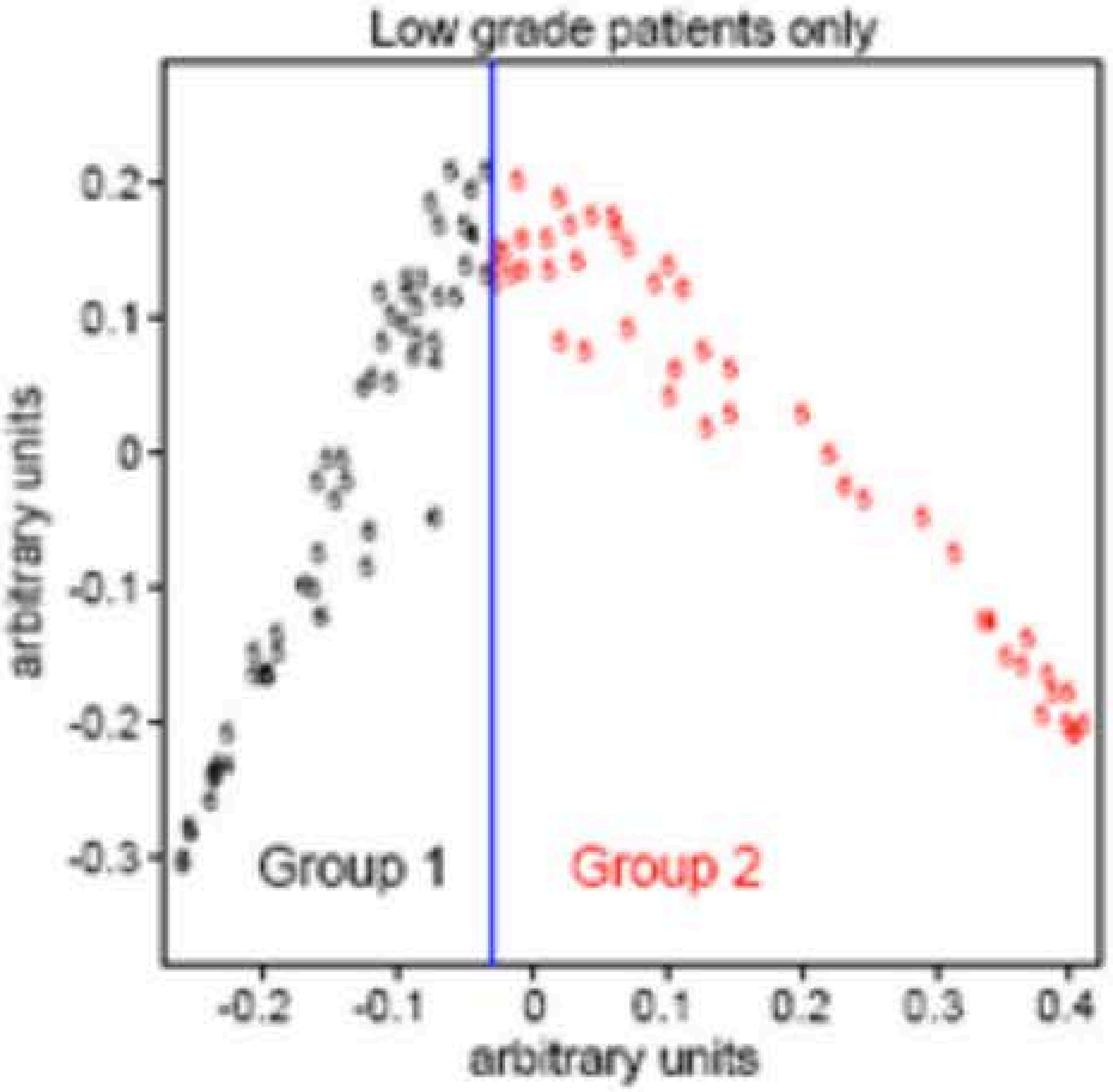
Aberrations in post-translational modifications of histones have been shown to occur in cancer cells but only at individual promoters[1] and have not been related to clinical outcome. Histone modifications, such as acetylation and methylation of lysines (K) and arginines (R), also occur over large regions of chromatin including non-promoter sequences[2] referred to as "global histone modifications." Here we asked whether changes in global levels of individual histone modifications are also associated with cancer and, importantly, whether these changes are predictive of clinical outcome. Through immunohistochemical staining of 183 primary prostatectomy samples, we determined the percentage of cells that stain for histone acetylation (Ac) and di-methylation (diMe) of five different residues in histones H3 and H4. Grouping of samples with similar patterns of modifications identified two disease sub-types with distinct risks of tumor recurrence among patients with low-grade prostate cancer. These patterns were predictors of outcome independent of tumor grade, stage, pre-operative prostate-specific antigen (PSA) levels, and capsule invasion. Thus, widespread changes in specific histone modifications represent novel molecular heterogeneity in prostate cancer, and may underlie the broad range of clinical behavior displayed by cancer patients.

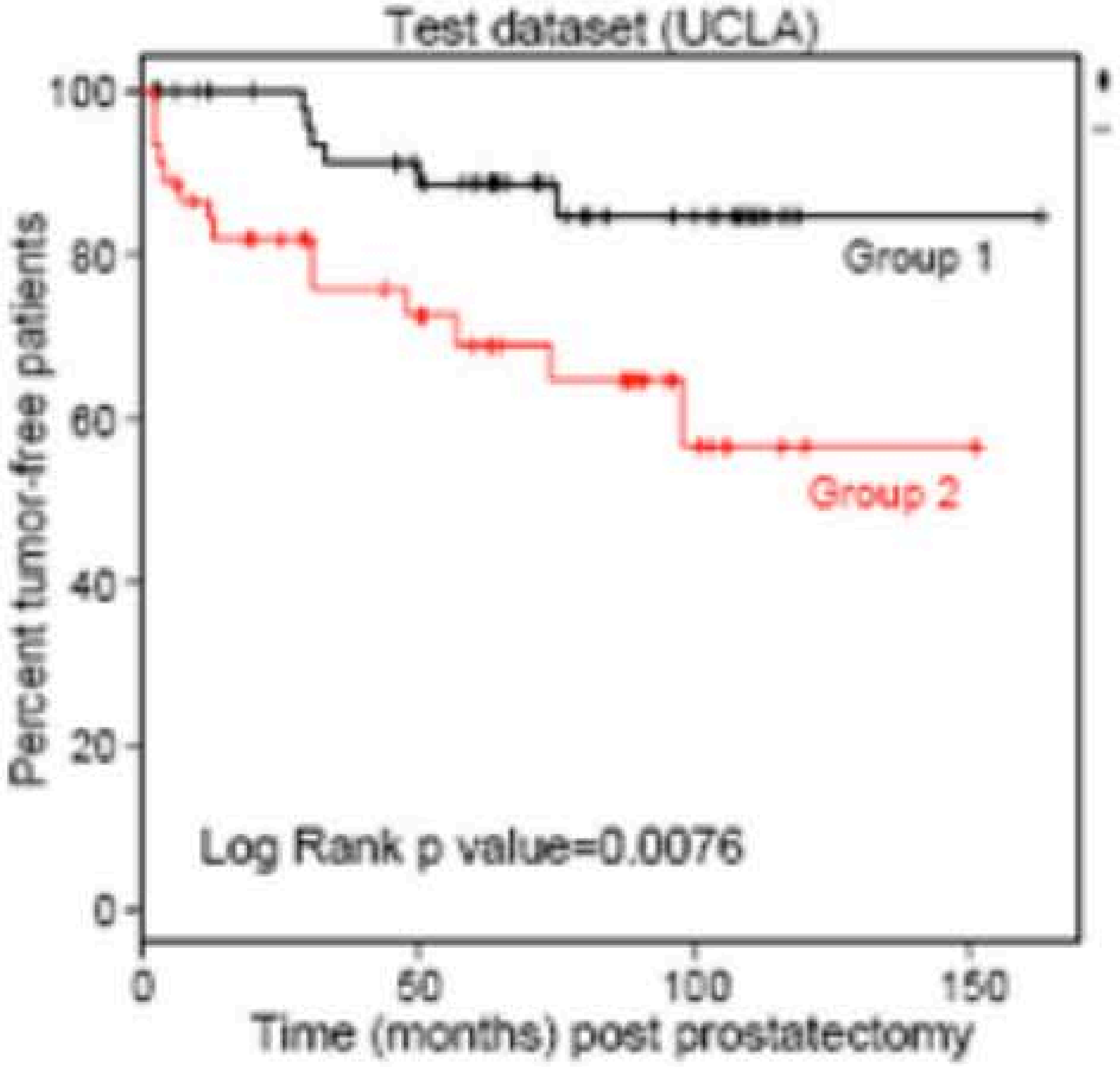
Analysis Outline

- Used RF clustering to find distinct patient clusters without regard to outcome
- Relating the clusters to clinical information showed that patient clusters have distinct PSA recurrence profiles
- Constructed a rule for predicting cluster membership
- Applied this rule to an independent validation data set to show that the rule predicts PSA recurrence

Cluster Analysis of Low Gleason Score Prostate Samples

(UCLA data)

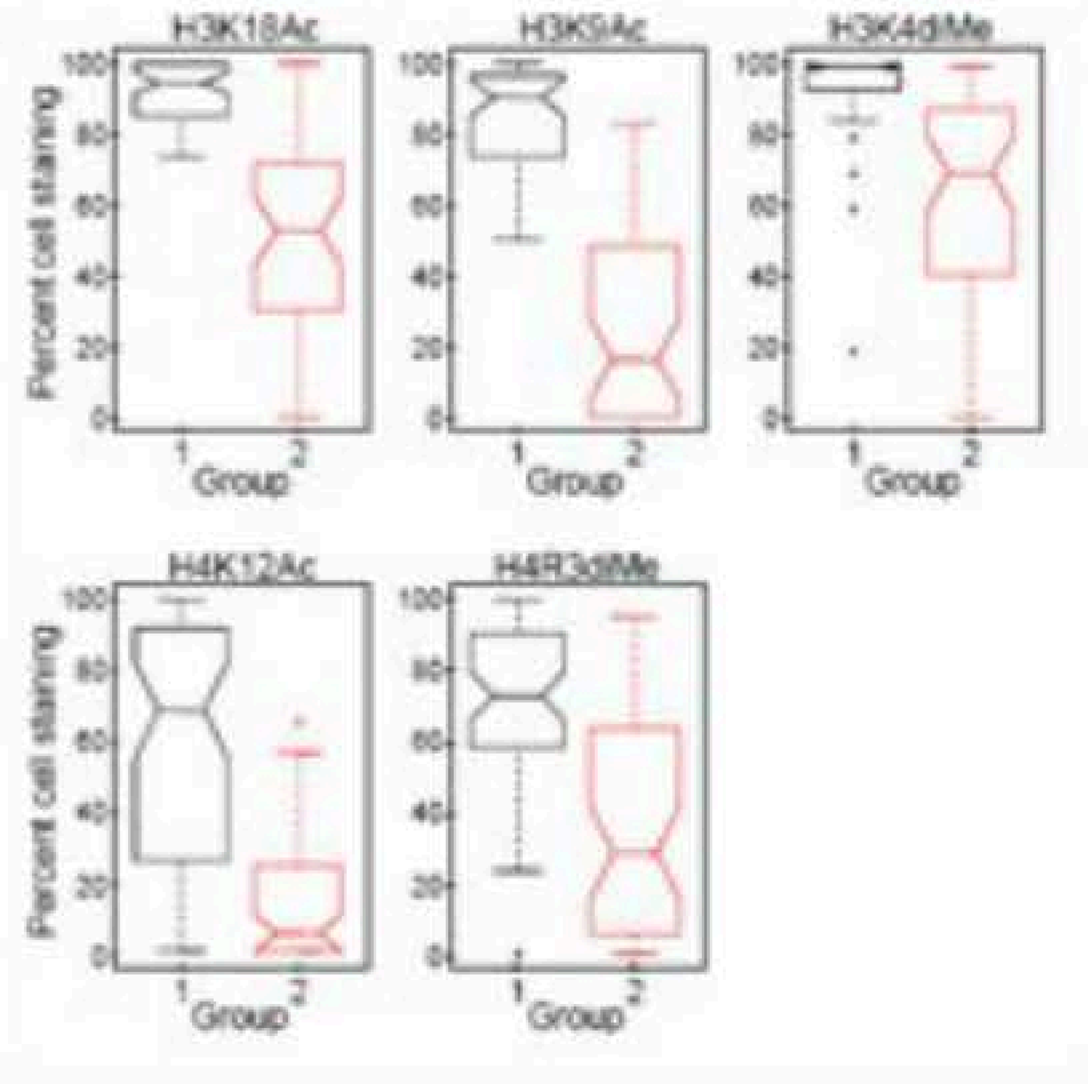


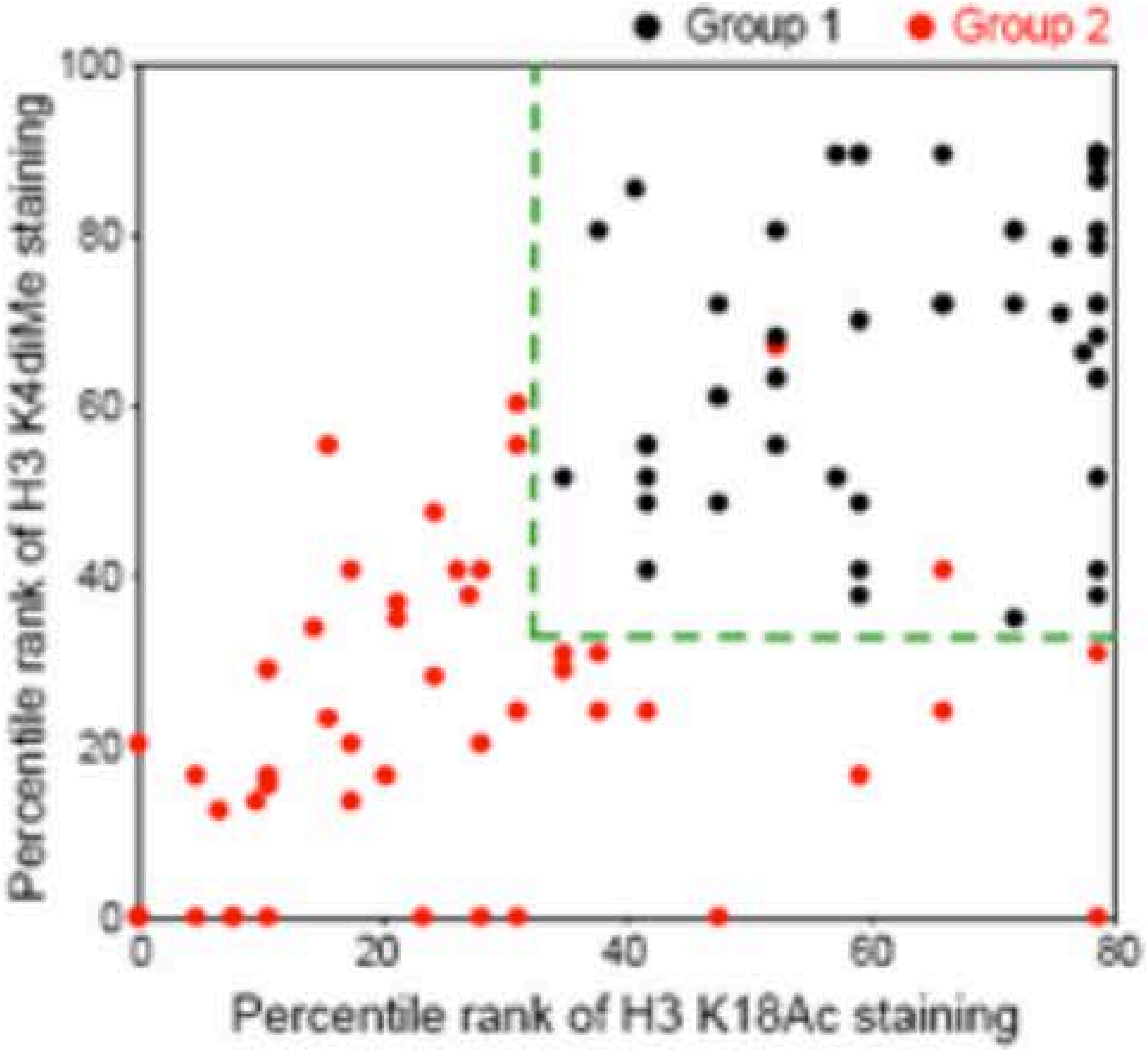


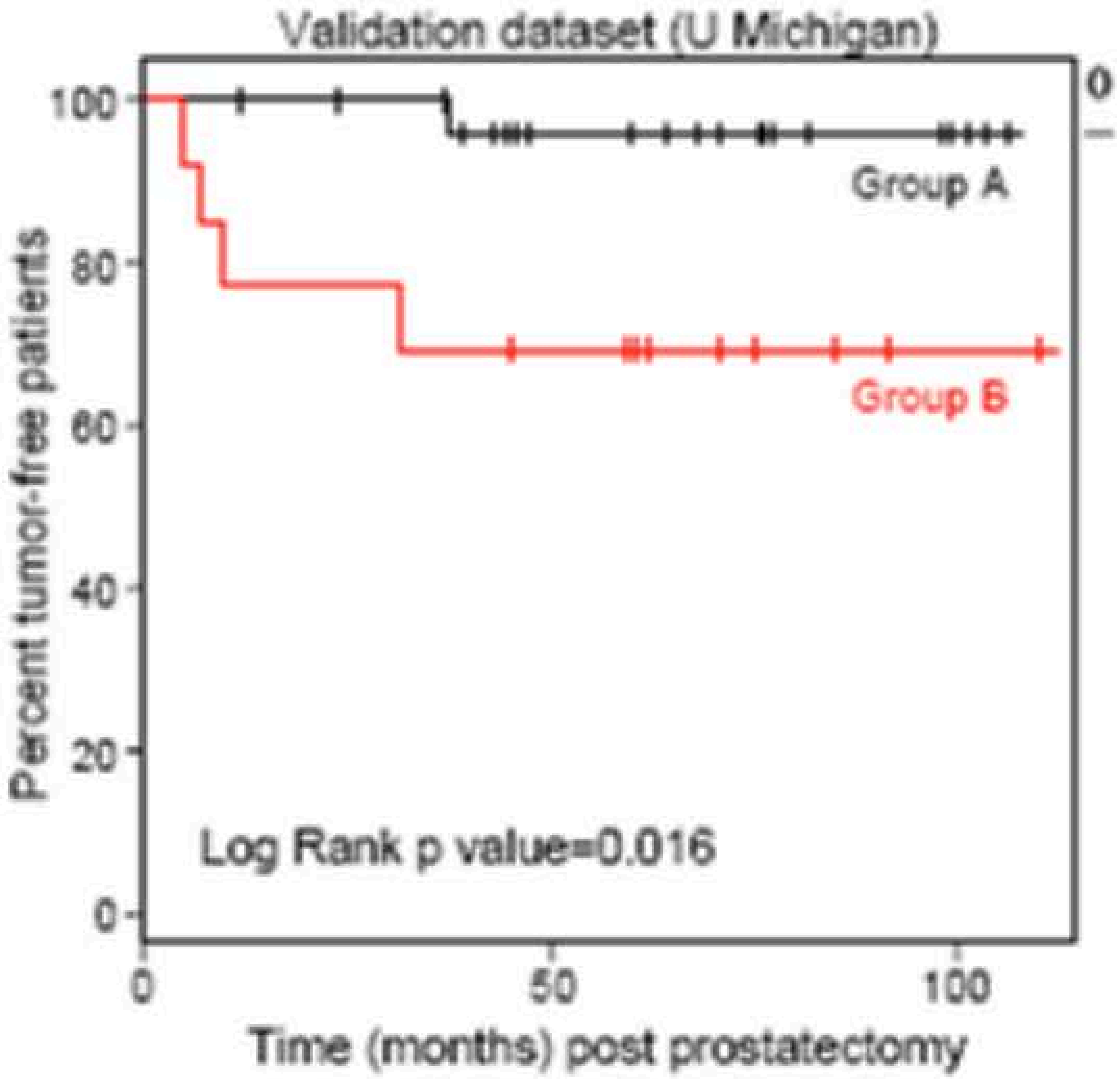
1) Construct a tumor marker rule for predicting RF cluster membership. 2) Validate the rule predictions in an independent data set

Threshold Rule

Validation







Discussion Prostate TMA Data

- Very weak evidence that individual markers predict PSA recurrence
- None of the markers validated individually
- However, cluster membership was highly predictive, i.e the rule could be validated in an independent data set.

Summary

- We have been motivated by the special features of TMA data and explored the use of RF dissimilarity in clustering analysis.
- We have carried out theoretical studies to gain more insights into RF clustering.
- We have applied RF clustering to different types of genomic data such as TMA, DNA microarray, genomic sequence (Allen et al. 2003) and SAGE (unpublished) data.

Acknowledgements

- **Former students & Postdocs for TMA**
  - **Tao Shi, PhD**
  - Tuyen Hoang, PhD
  - Yunda Huang, PhD
  - Xueli Liu, PhD
- Special Consultant
  - Panda Bamboo, PhD

UCLA

Tissue Microarray Core

- **David Seligson, MD**
- Aarno Palotie, MD
- Arie Belldegrun, MD
- Robert Figlin, MD
- Lee Goodglick, MD
- David Chia, MD
- Siavash Kurdistani, MD
- ETC

References RF clustering

- Unsupervised learning tasks in TMA data analysis
  - Review random forest predictors (introduced by L. Breiman)
    - Shi, T. and Horvath, S. (2005) "Unsupervised learning using random forest predictors" Journal of Computational and Graphical Statistics
  - www.genetics.ucla.edu/labs/horvath/RFclustering/RFclustering.htm
- Application to Tissue Array Data
  - Shi, T., Seligson, D., Belldegrun, A. S., Palotie, A., Horvath, S. (2004) Tumor Profiling of Renal Cell Carcinoma Tissue Microarray Data. Modern Pathology
  - Seligson DB, Horvath S, Shi T, Yu H, Tze S, Grunstein M, Kurdistani S (2005) Global histone modification patterns predict risk of prostate cancer recurrence. Nature

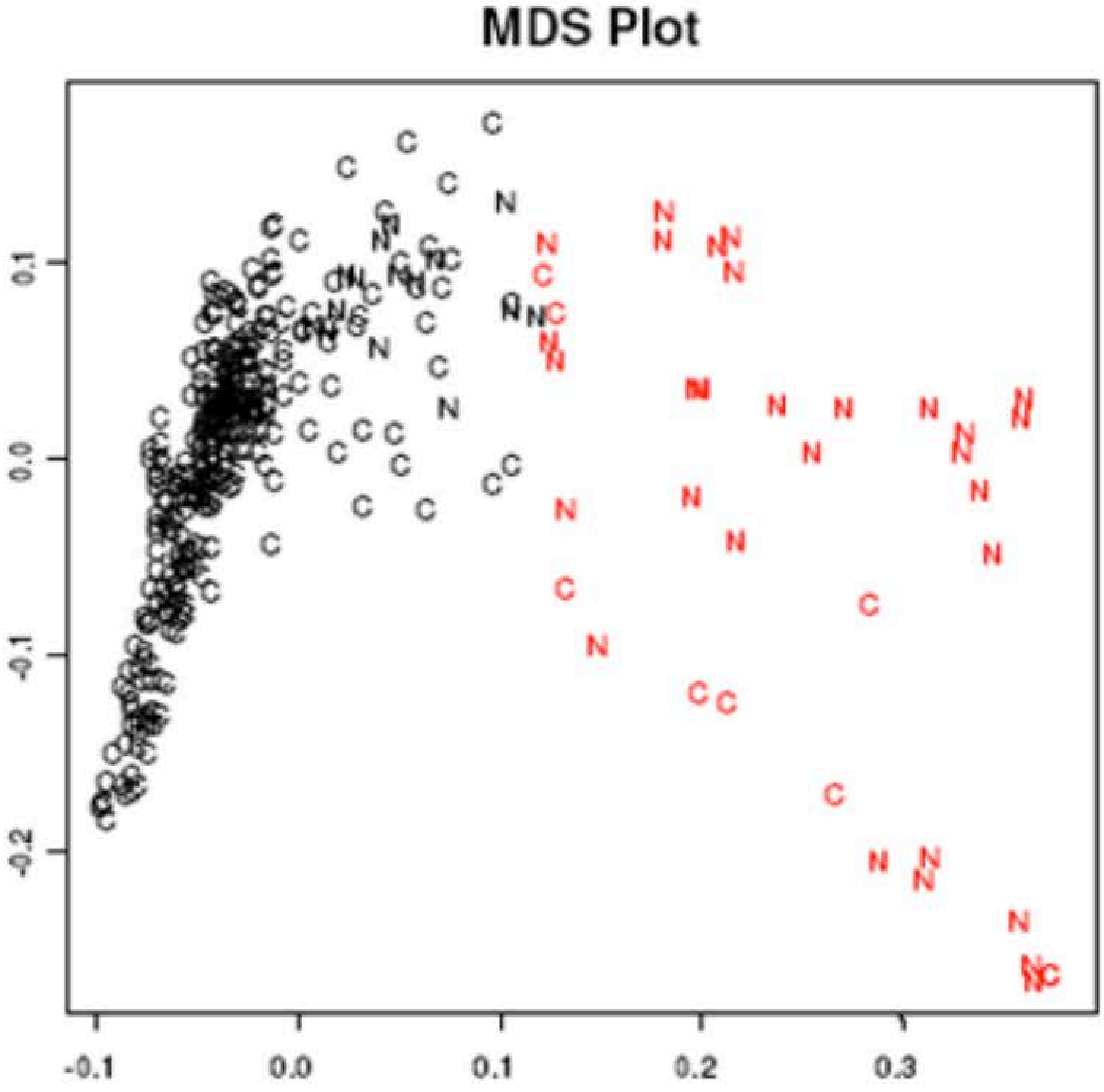
Applications to renal cell carcinoma tissue microarray data

Shi T, Seligson D, Belldegrun AS, Palotie A, Horvath S (2005) Tumor Classification by Tissue Microarray Profiling: Random Forest Clustering Applied to Renal Cell Carcinoma. Mod Pathol. 2005 Apr;18(4):547-57.

TMA Data

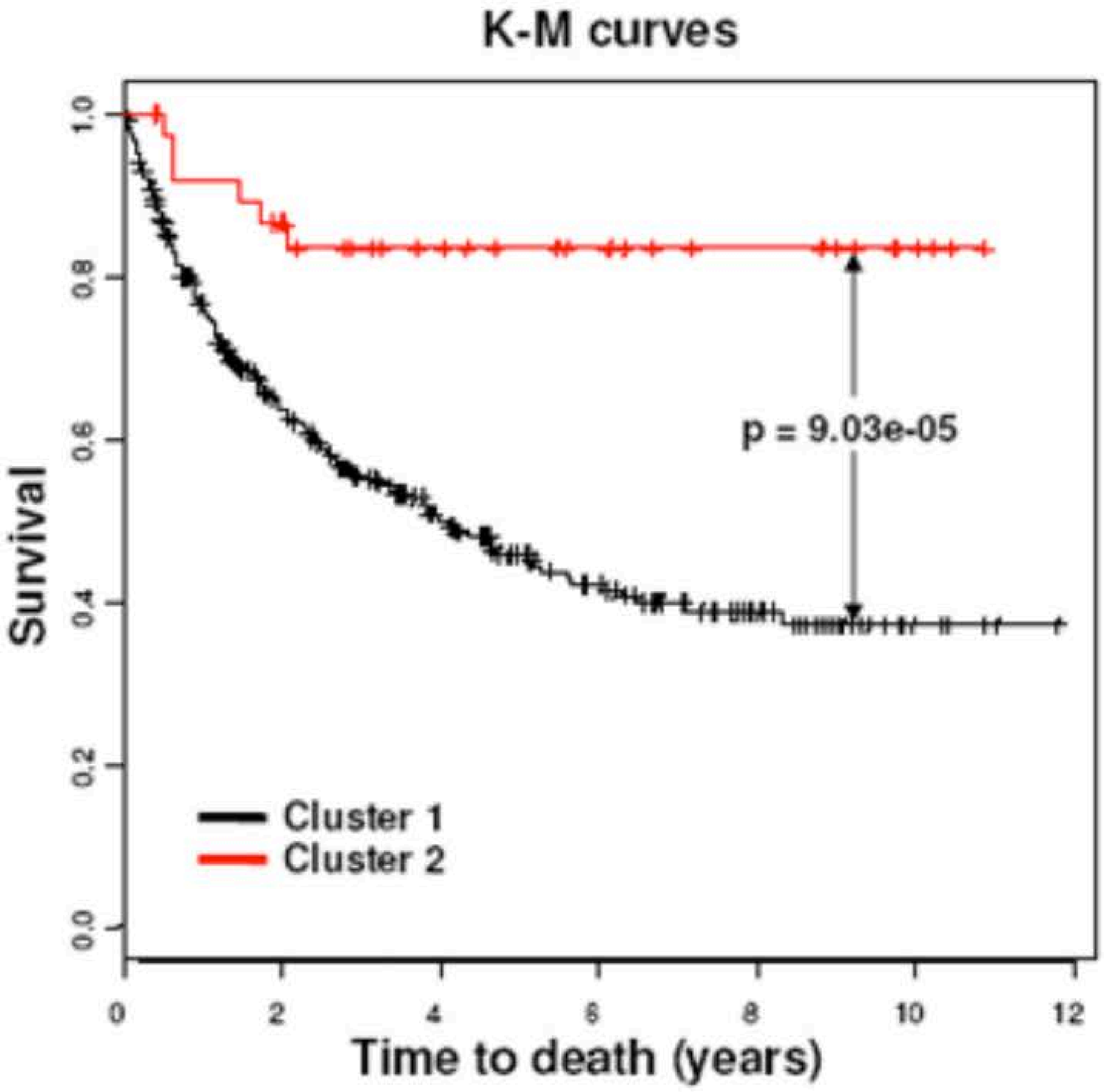
- **366 patients** with Renal Cell Carcinoma (RCC) admitted to UCLA between 1989 and 2000.
- Immuno-histological measures of **8 tumor markers** were obtained from tissue microarrays constructed from the tumor samples of these patients.

MDS Plot of All the RCC Patients



- Colored by their RF cluster and labeled by tumor subtypes.

Interpreting the clusters in terms of survival

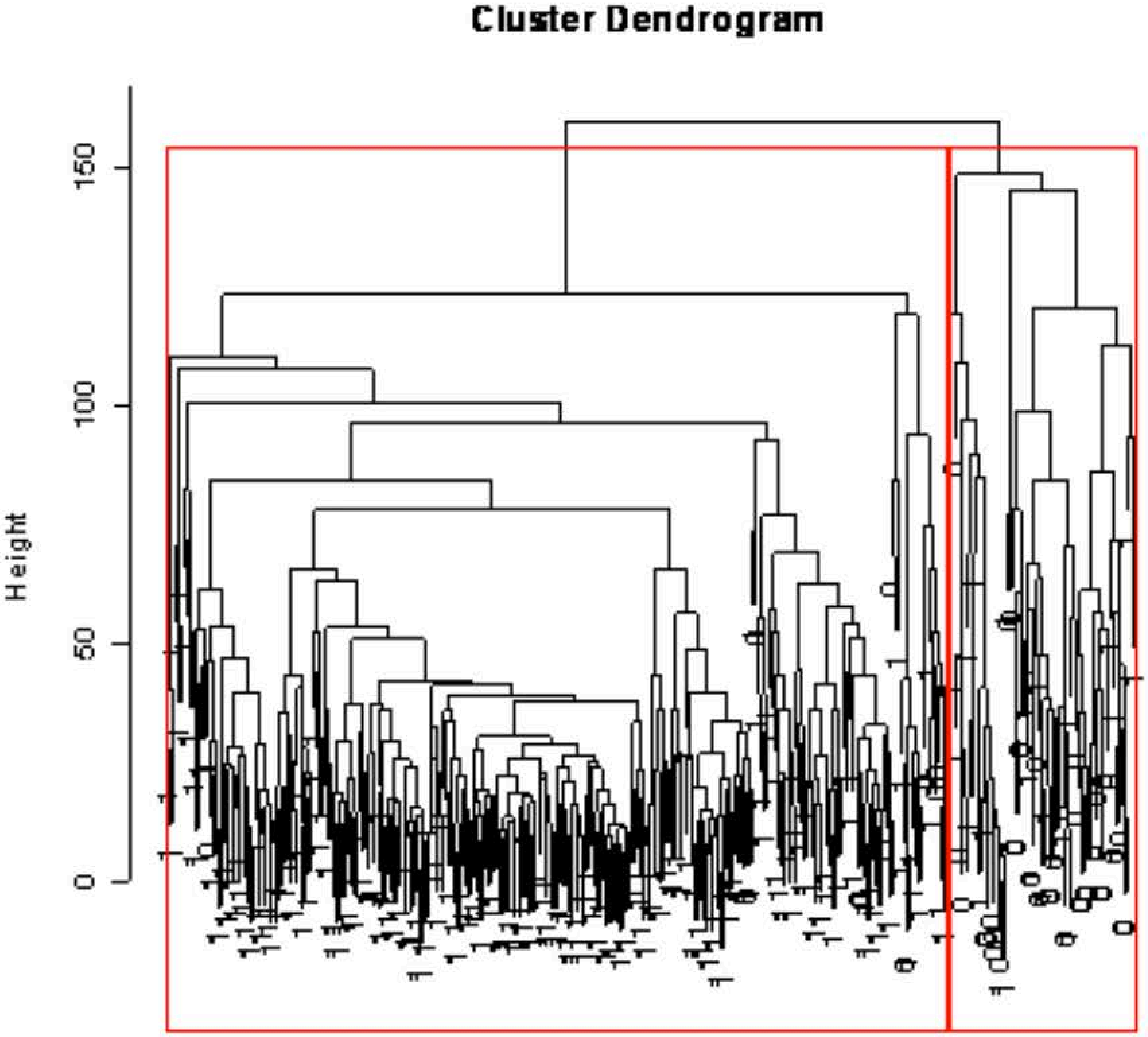


Clustering label	Non clear Cell patients	Clear cell patients
1	20	307
2	30	9

Hierarchical clustering with Euclidean distance leads to less satisfactory results

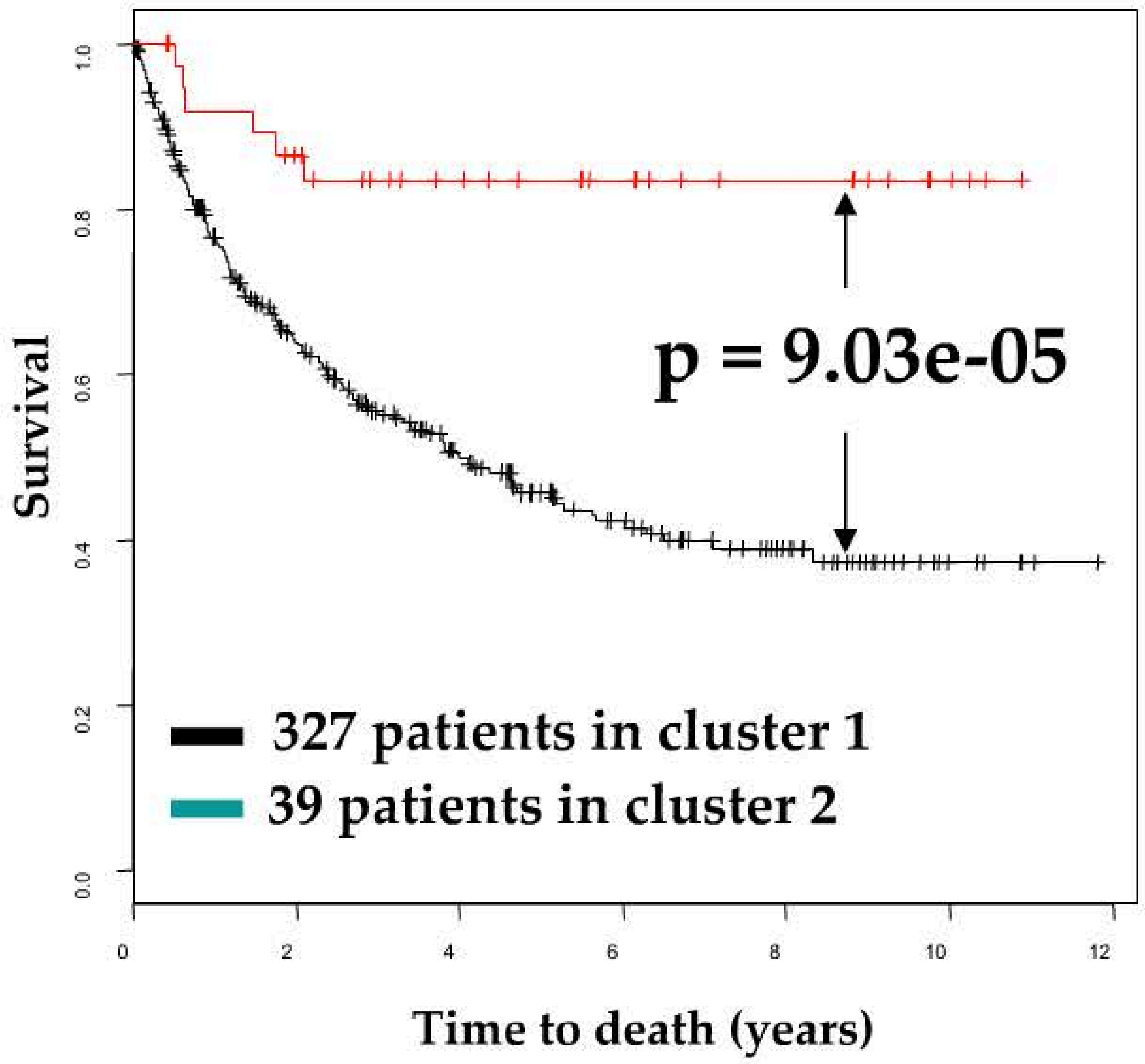
Cluster-ing label	Non clear Cell patients	Clear cell patients
1	9 (20)	286 (307)
2	41 (30)	30 (9)

* RF clustering grouping in red

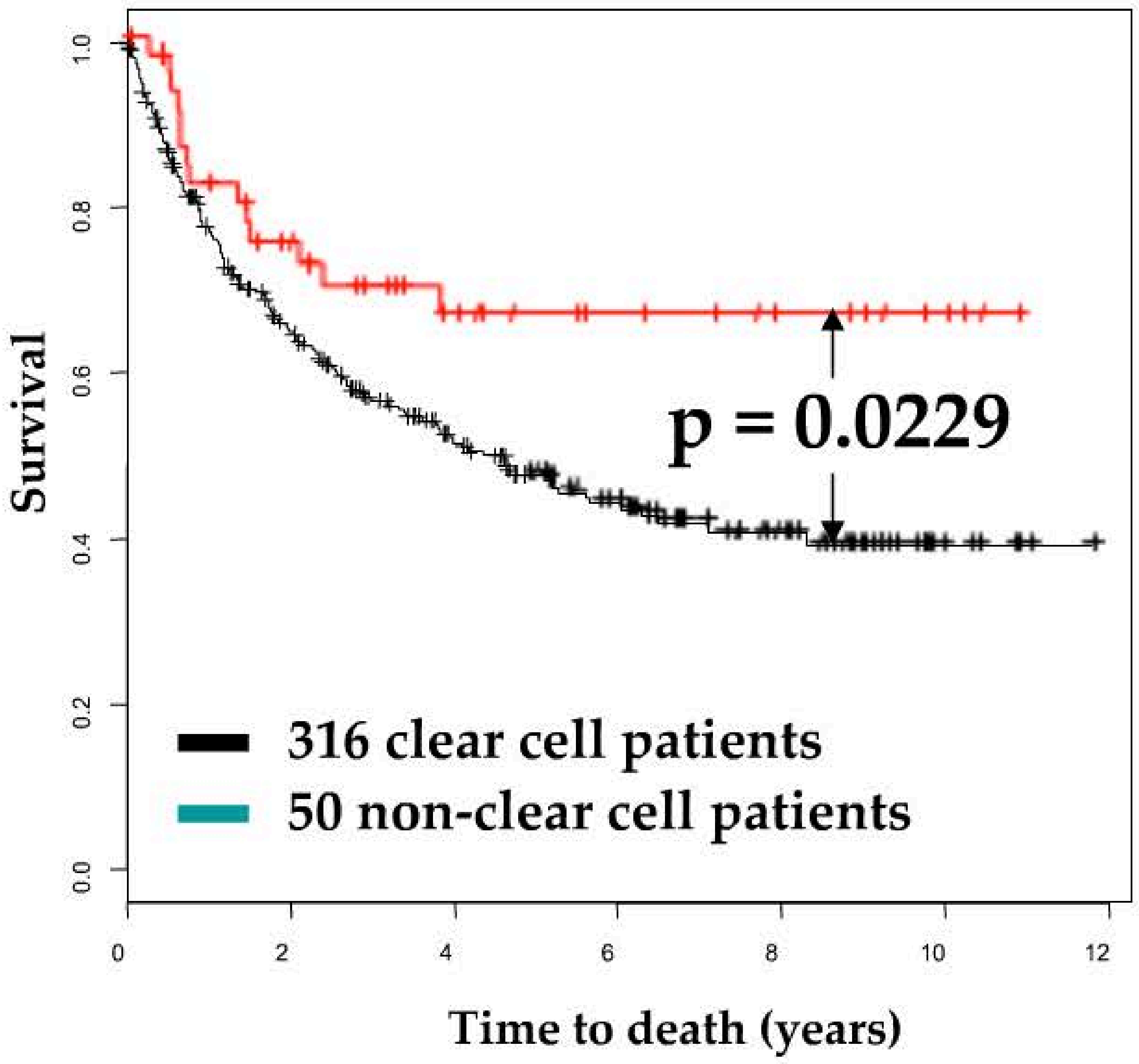


Molecular grouping is superior to pathological grouping

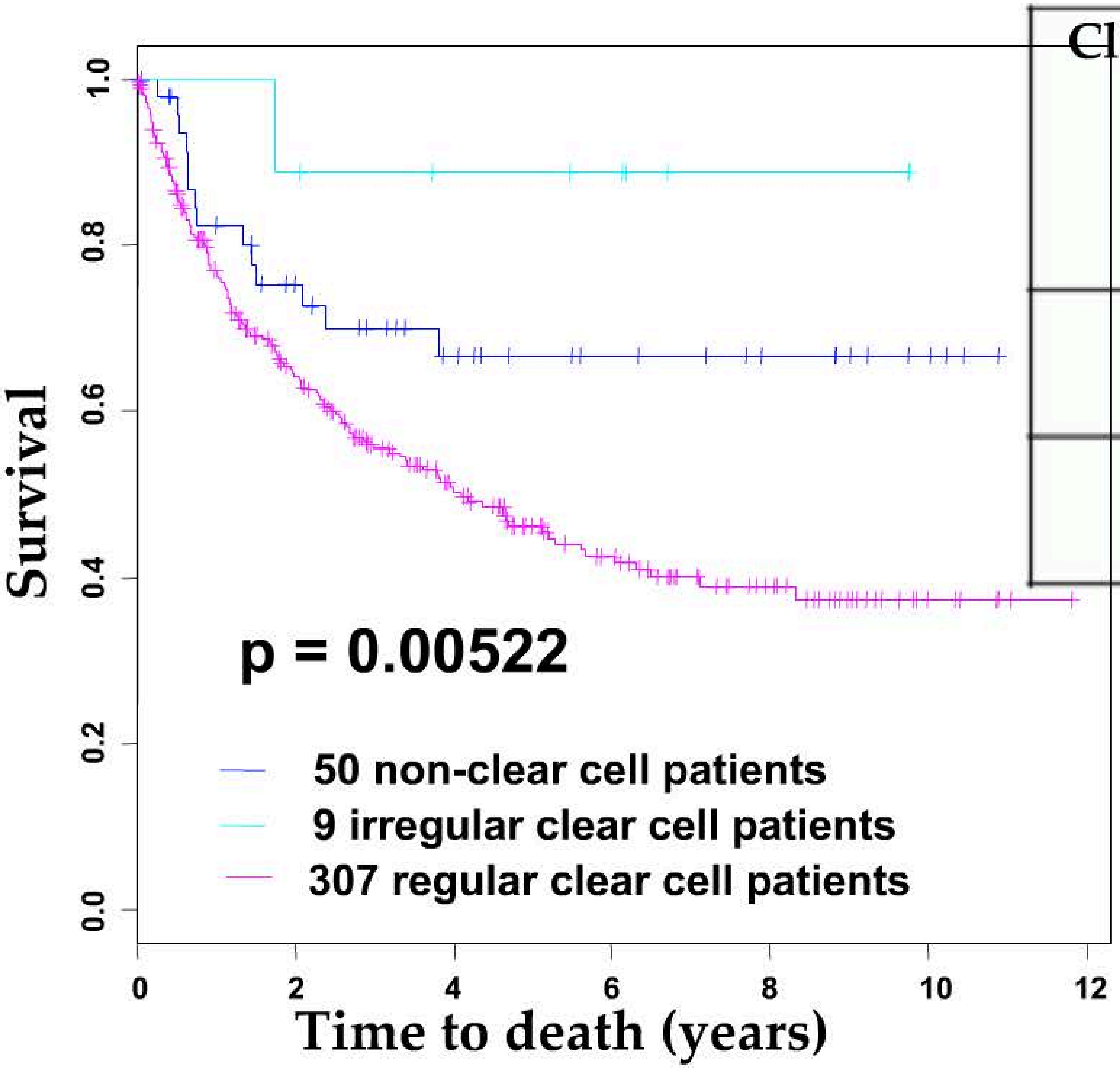
Molecular Grouping



Pathological Grouping

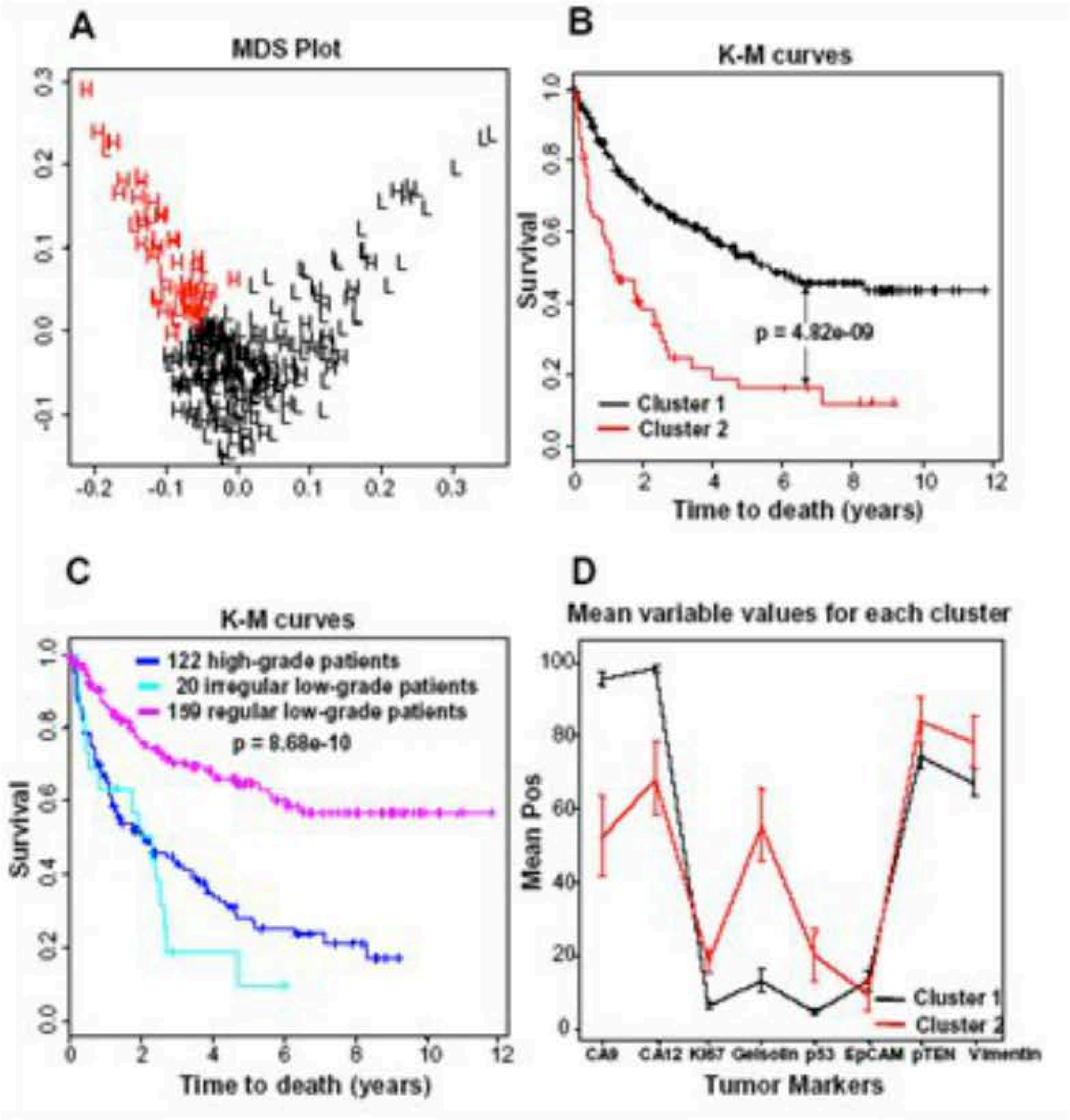


Identify “irregular” patients

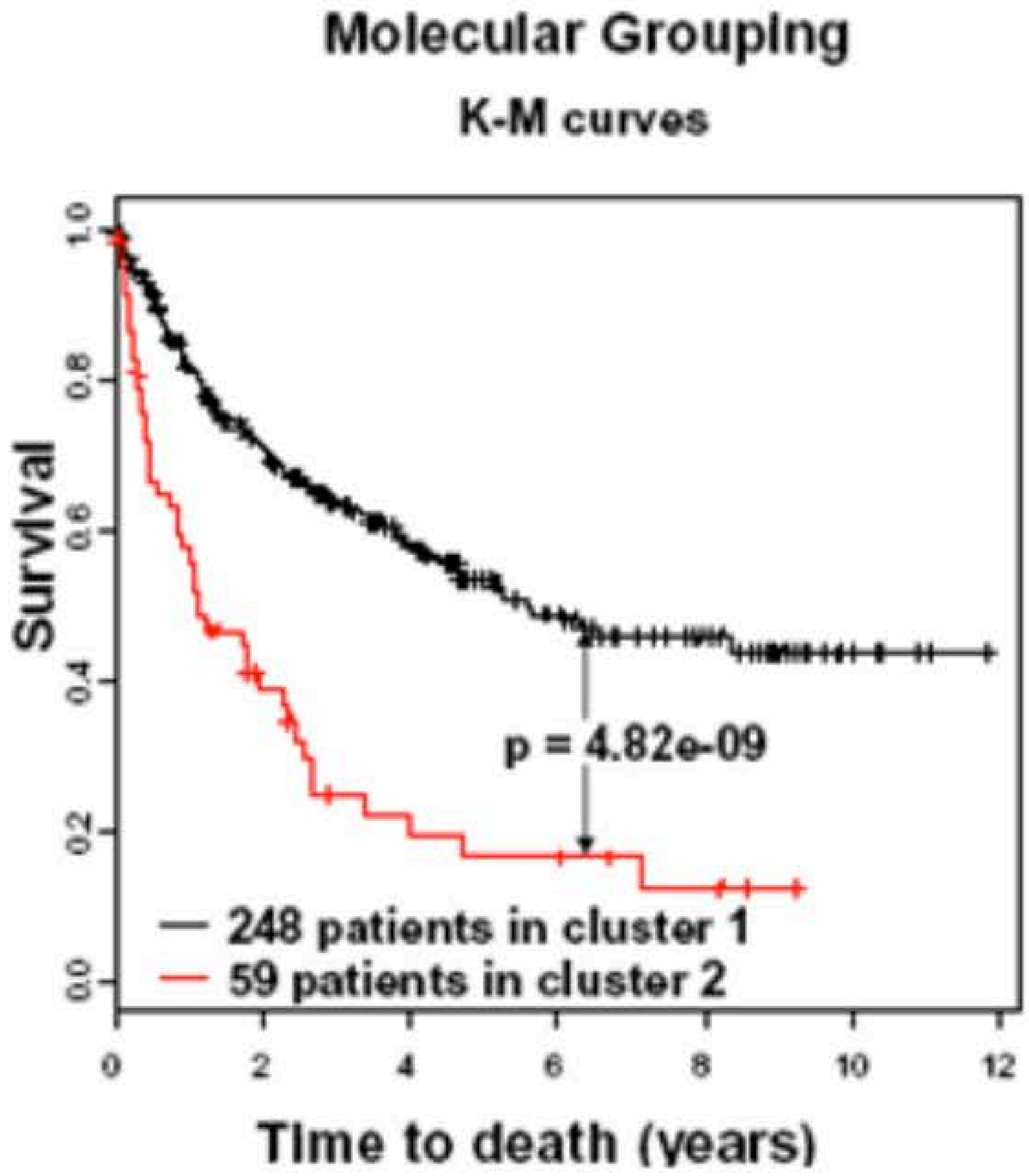


Clustering label	Non clear Cell patients	Clear cell patients
1	20	307
2	30	9

`Regular' Clear Cell Patients

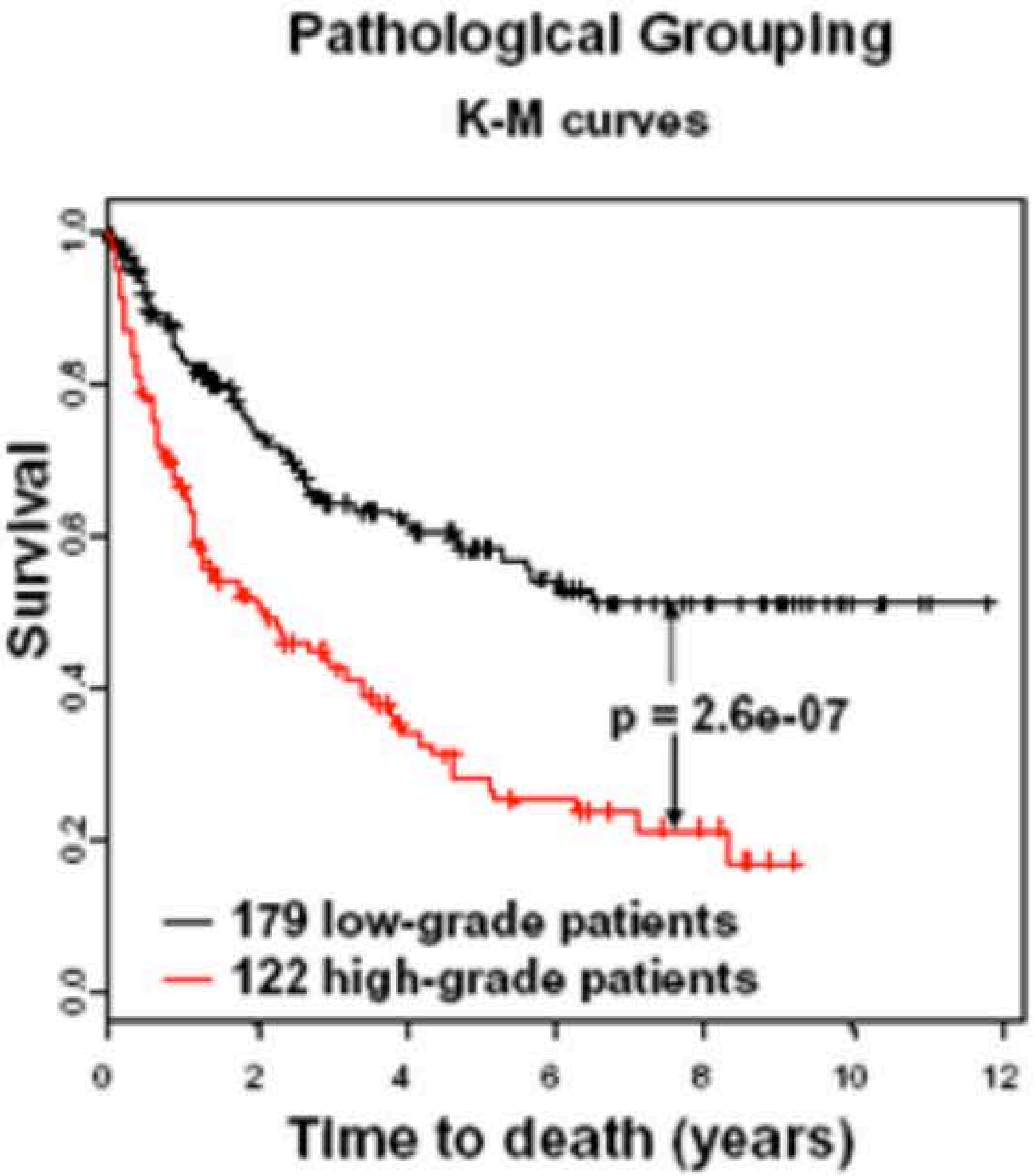


`Regular' Clear Cell Patients (cont.)



Molecular Grouping

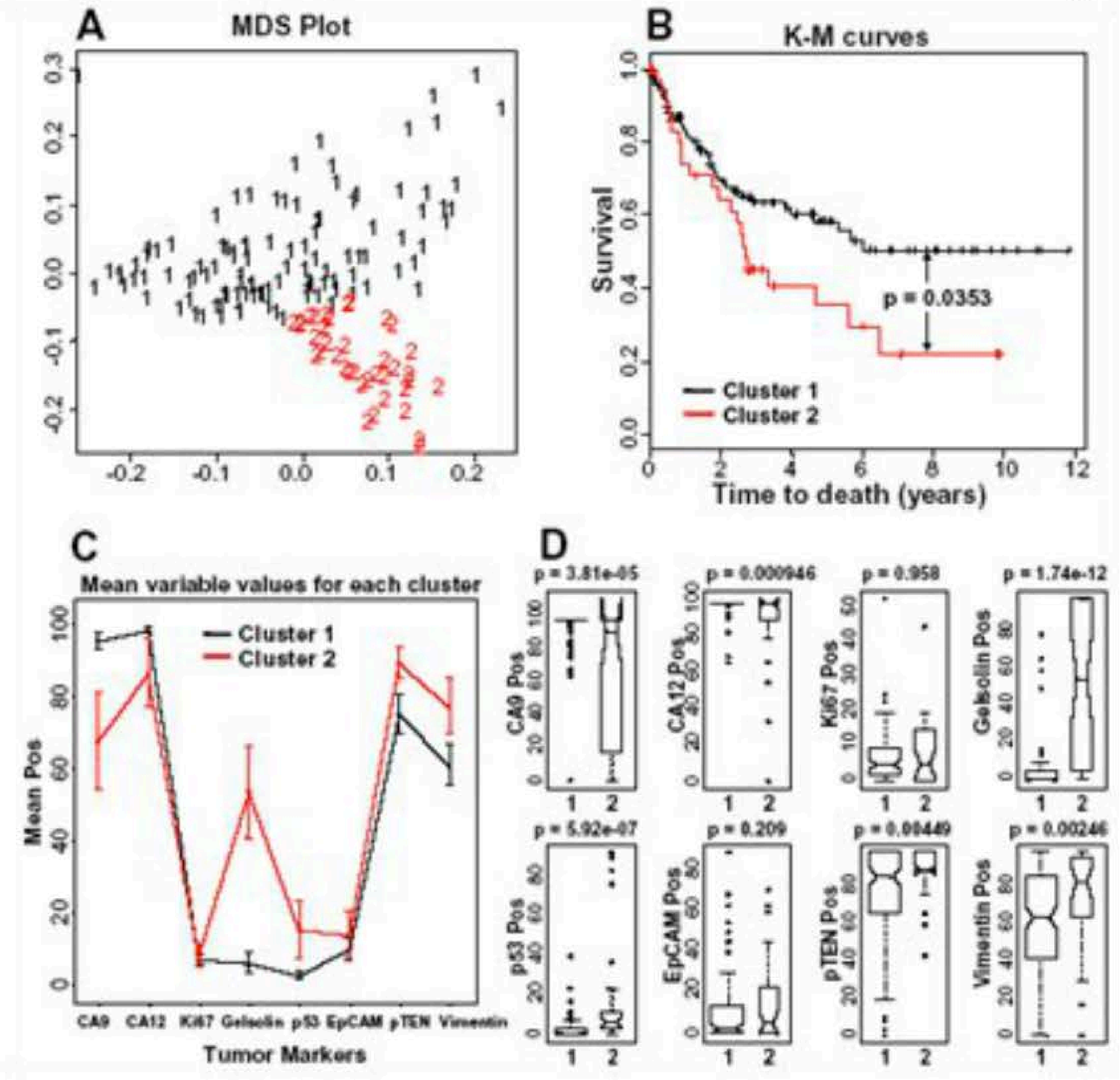
K-M curves



Pathological Grouping

K-M curves

Detect novel cancer subtypes



- Group clear cell grade 2 patients into two clusters with significantly different survival.

Results TMA clustering

- Clusters reproduce well known clinical subgroups
  - Example: global expression differences between clear cell and non-clear cell patients
- RF clustering allows one to identify "outlying" tumor samples.
- Can detect previously unknown sub-groups
- Empirical evidence suggests that RF clustering is better than standard clustering in this setting (prostate data, unpublished)

Acknowledgements

- **Former students & Postdocs for TMA**
  - **Tao Shi, PhD**
  - Tuyen Hoang, PhD
  - Yunda Huang, PhD
  - Xueli Liu, PhD

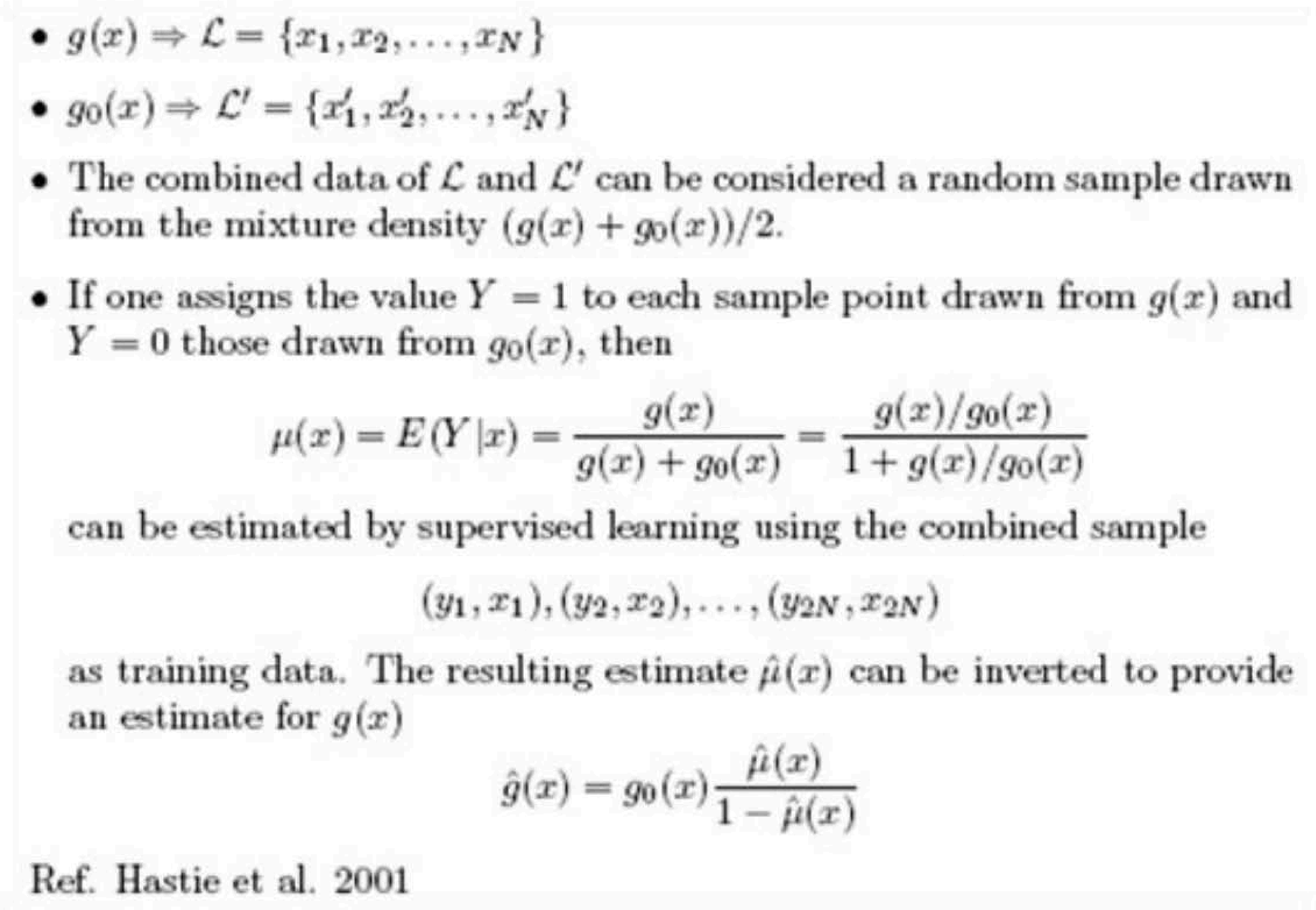
UCLA
Tissue Microarray Core

- **David Seligson, MD**
- Aarno Palotie, MD
- Arie Belldegrun, MD
- Robert Figlin, MD
- Lee Goodglick, MD
- David Chia, MD
- Siavash Kurdistani, MD

THE END

Appendix

Casting an unsupervised problem into a supervised problem



- $g(x) \Rightarrow \mathcal{L} = \{x_1, x_2, \ldots, x_N\}$
- $g_0(x) \Rightarrow \mathcal{L}' = \{x'_1, x'_2, \ldots, x'_N\}$
- The combined data of $\mathcal{L}$ and $\mathcal{L}'$ can be considered a random sample drawn from the mixture density $(g(x) + g_0(x))/2$.
- If one assigns the value $Y = 1$ to each sample point drawn from $g(x)$ and $Y = 0$ those drawn from $g_0(x)$, then

$$\mu(x) = E(Y\,|x) = \frac{g(x)}{g(x) + g_0(x)} = \frac{g(x)/g_0(x)}{1 + g(x)/g_0(x)}$$

can be estimated by supervised learning using the combined sample

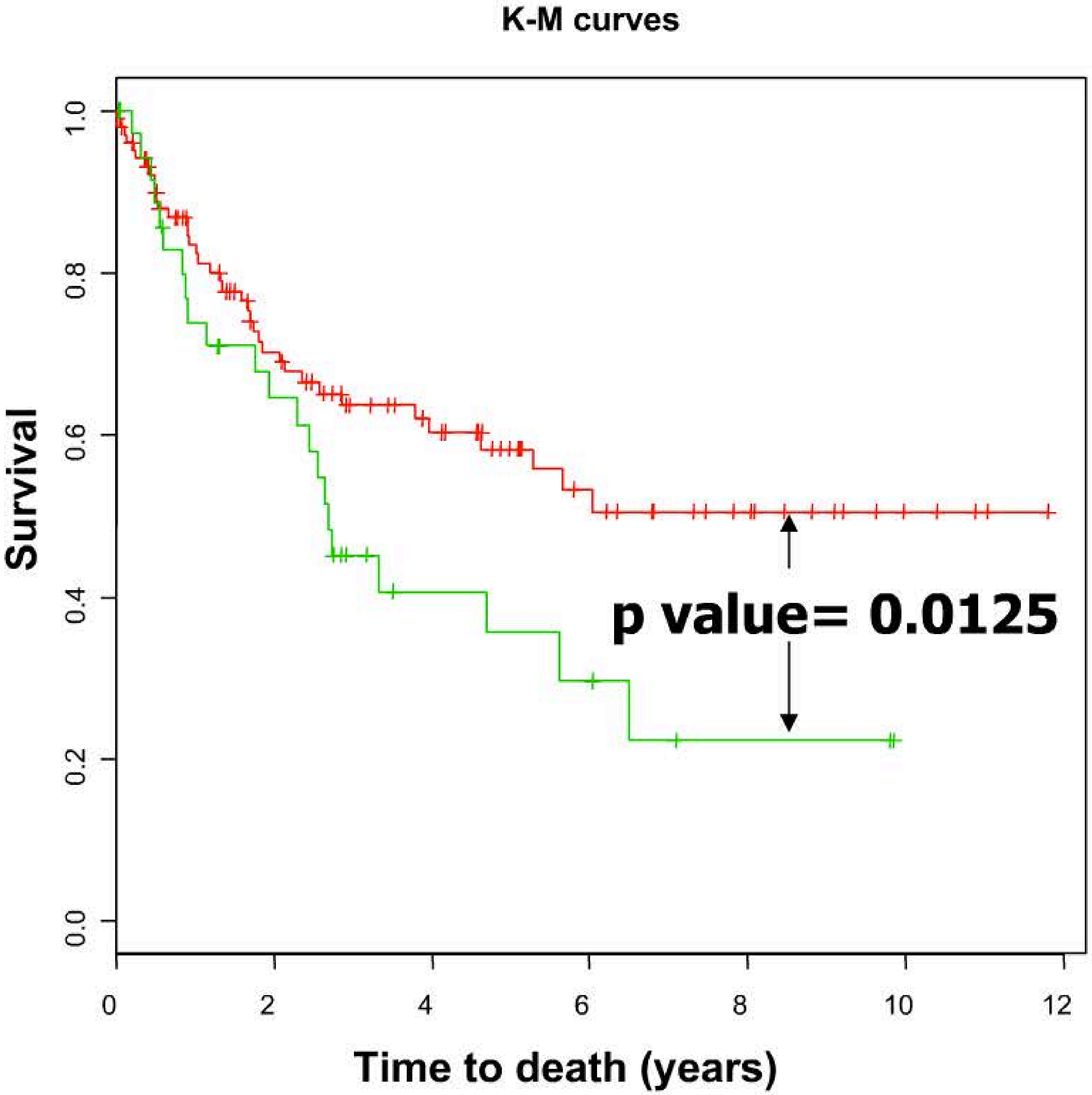
$$(y_1, x_1), (y_2, x_2), \ldots, (y_{2N}, x_{2N})$$

as training data. The resulting estimate $\hat{\mu}(x)$ can be inverted to provide an estimate for $g(x)$

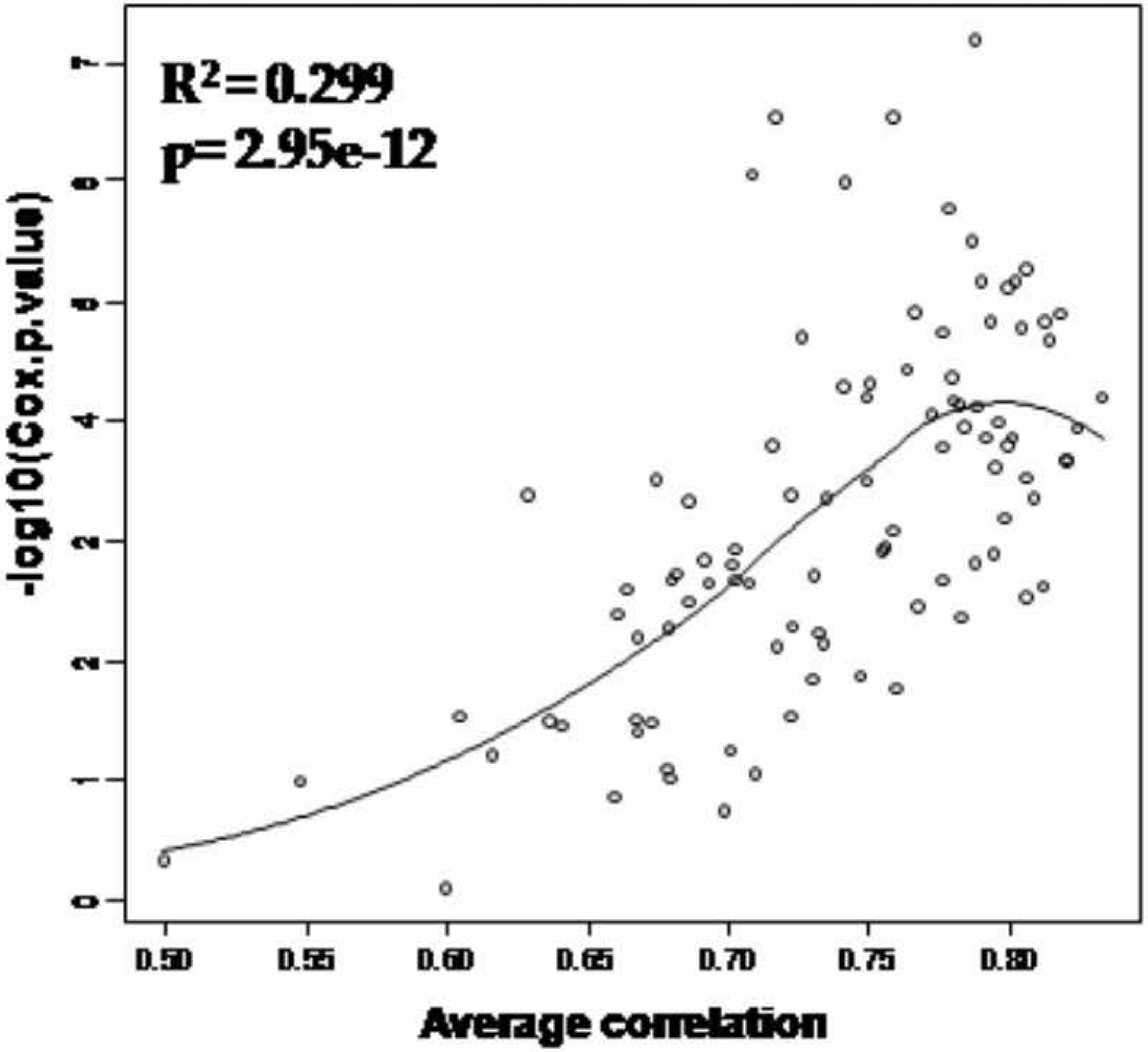
$$\hat{g}(x) = g_0(x)\frac{\hat{\mu}(x)}{1 - \hat{\mu}(x)}$$

Ref. Hastie et al. 2001

Detect novel cancer subtypes



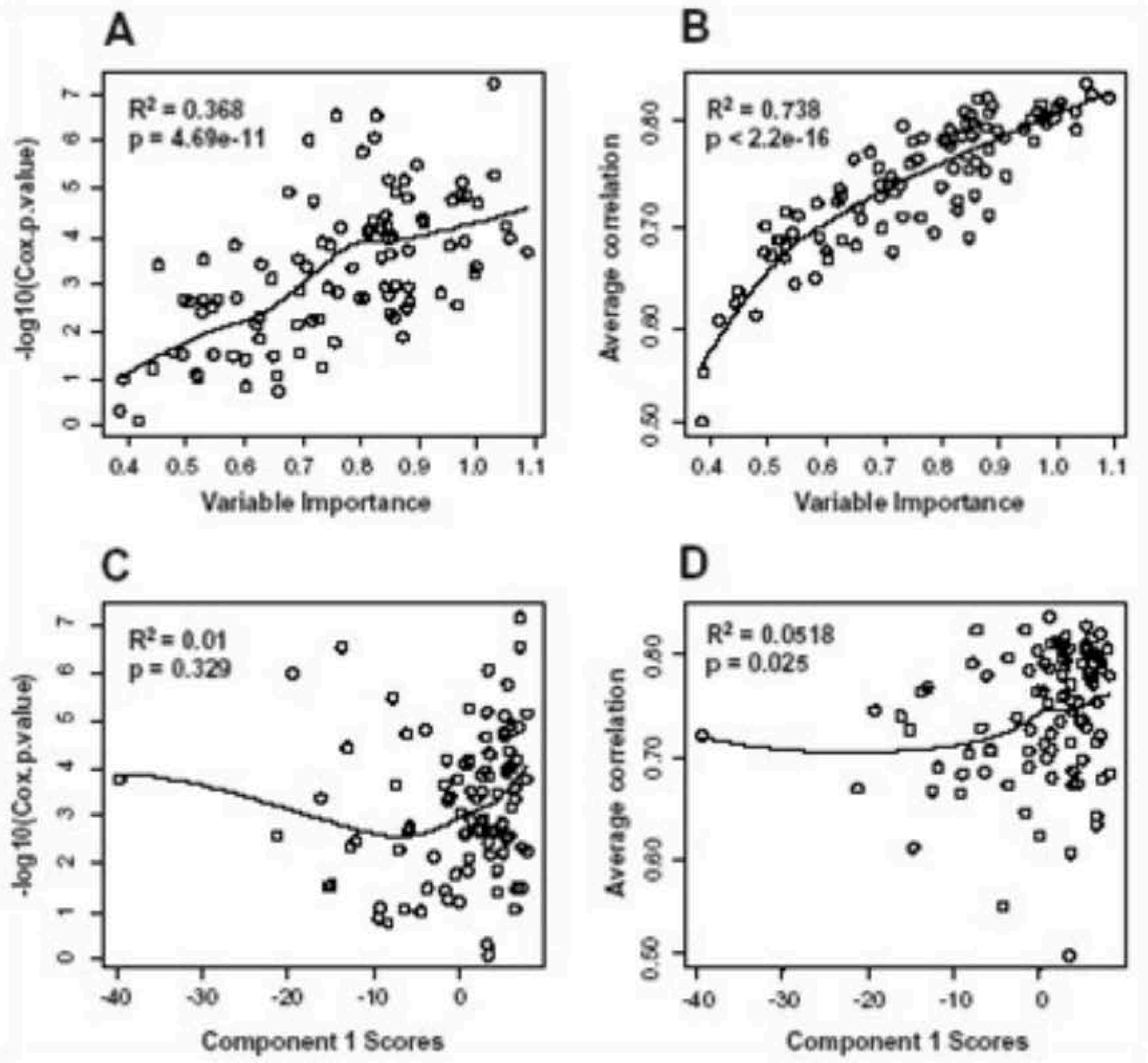
- Group clear cell grade 2 patients into two clusters with significantly different survival.


R²=0.299
p=2.95e-12
-log10(Cox.p.value)
0
1
2
3
4
5
6
7
0.50
0.55
0.60
0.65
0.70
0.75
0.80
Average correlation

RF variable importance vs. Average Corr and Cox p value

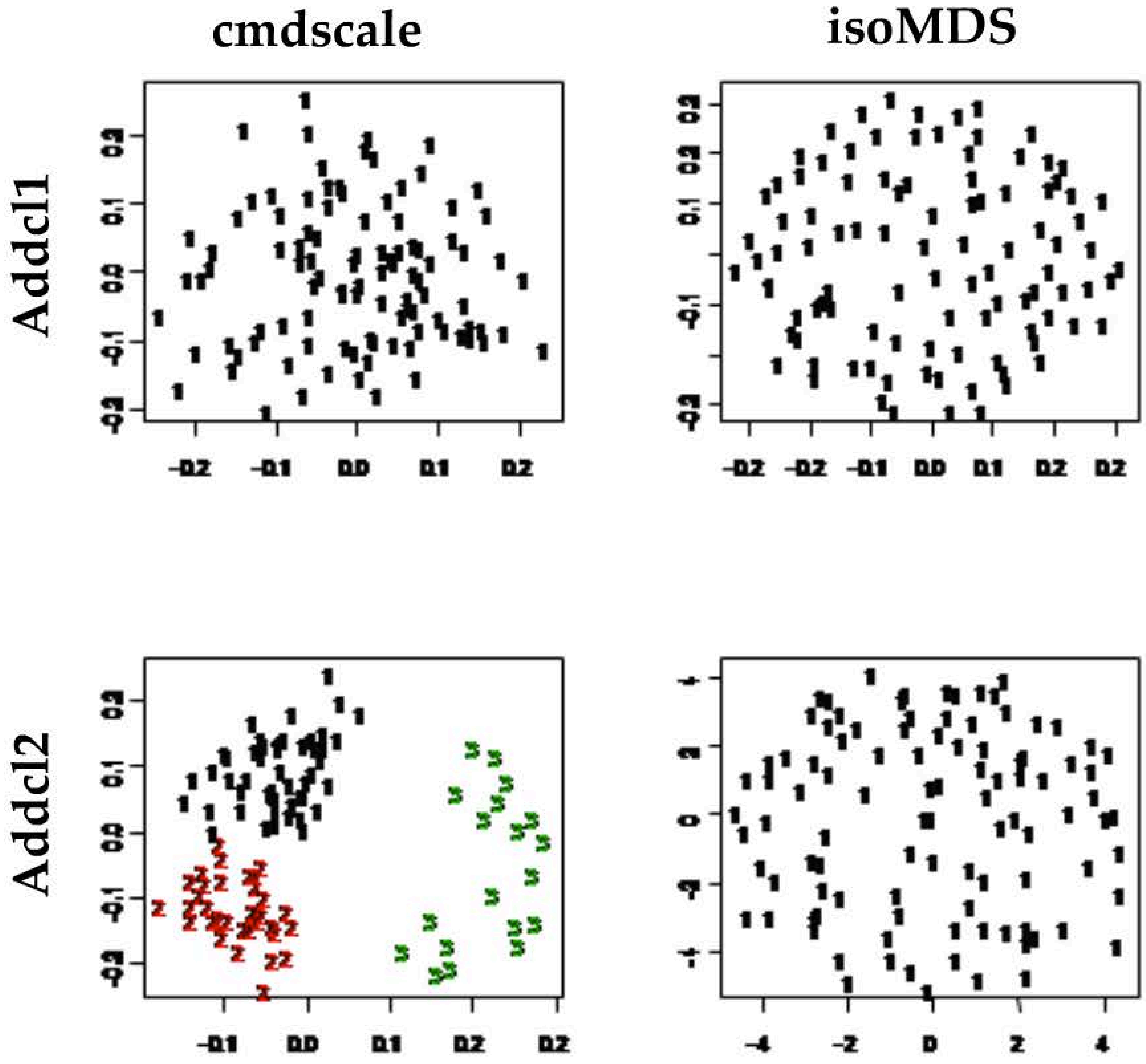
The more important a gene is according to RF, the more important it is for survival prediction

Message:
The more correlated a gene is
With other genes the more
Important it is for the Def



Which multi-dimensional scaling method to use?

- cmdscale usually works well with Addcl1 but not with Addcl2 because it may lead to spurious clusters.
- However isoMDS works well with Addcl2!



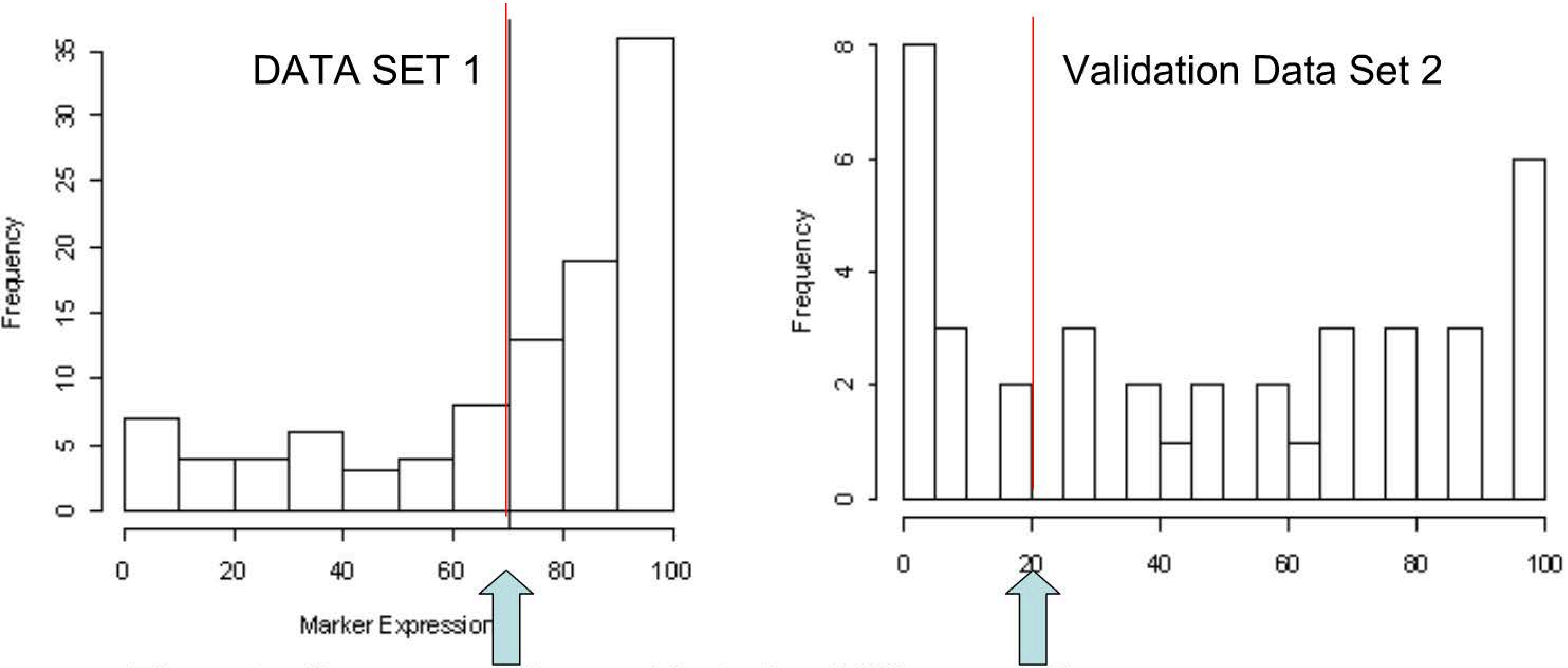
The random forest dissimilarity

L. Breiman: RF manual

Technical Report: Shi and Horvath 2005

http://www.genetics.ucla.edu/labs/horvath/RFclustering/RFclustering.htm

Frequency plot of the same tumor marker in 2 independent data sets



The cut-off corresponds roughly to the 66% percentile.
Thresholding this tumor marker allows one to stratify the cancer patients into high risk and low risk patients. Although the distribution looks very different the percentile threshold can be validated and is clinically relevant.