Bioinformatical Considerations for High Dimensional Data Derived from High Throughput Assays

Yu Shyr, Ph.D.

Professor of Biostatistics & Preventive Medicine

Director of Biostatistics Shared Resource
Vanderbilt-Ingram Cancer Center

E-mail:Yu.Shyr@vanderbilt.edu

iSBTc

November 5, 2004

The problem

- The major challenge in high throughput experiments, e.g., microarray data, MALDI-TOF data, or SELDI-TOF data, is that the data is often **high dimensional**.

- When the number of dimensions reaches thousands or more, the **computational time** for the pattern recognition algorithms can become unreasonable. This can be a problem, especially when some of the features are **not** discriminatory.

The problem

- The irrelevant features may cause a reduction in the accuracy of some algorithms. For example (Witten 1999), experiments with a decision tree classifier have shown that adding a random binary feature to standard datasets can deteriorate the classification performance by **5 - 10%.**

- Furthermore, in many pattern recognition tasks, the number of features represents the dimension of a search space - the **larger** the number of features, the **greater** the dimension of the search space, and the **harder** the problem.

Issues in the Analysis of High-Throughput Experiment

- Experiment Design
- Measurement
- Preprocessing
  - Filtering, Baseline Correction, Normalization
  - Profile Alignment, Transformation, Variance correction
- Feature Selection
- Classification

Steps in the Analysis of High-Throughput Experiment

- Computational Validation
  - Estimate the classification error rate
  - bootstrapping, k-fold validation, leave-one-out validation
- Significance Testing of the Achieved Classification Error
- Validation – blind test cohort
- Reporting the result - graphic & table

Experiment Design

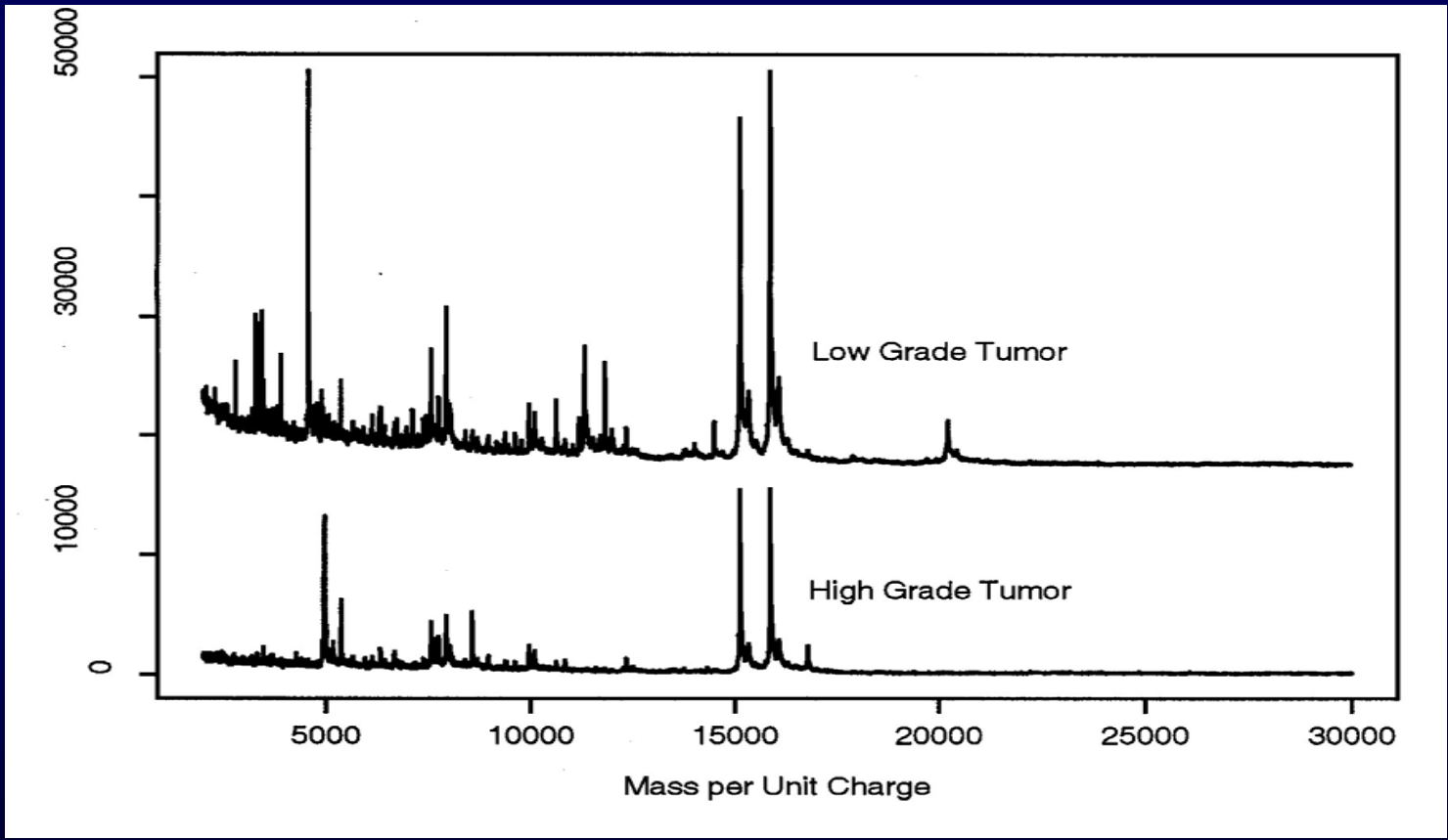
- **Study Objectives:**

  Class Discovery (unsupervised)

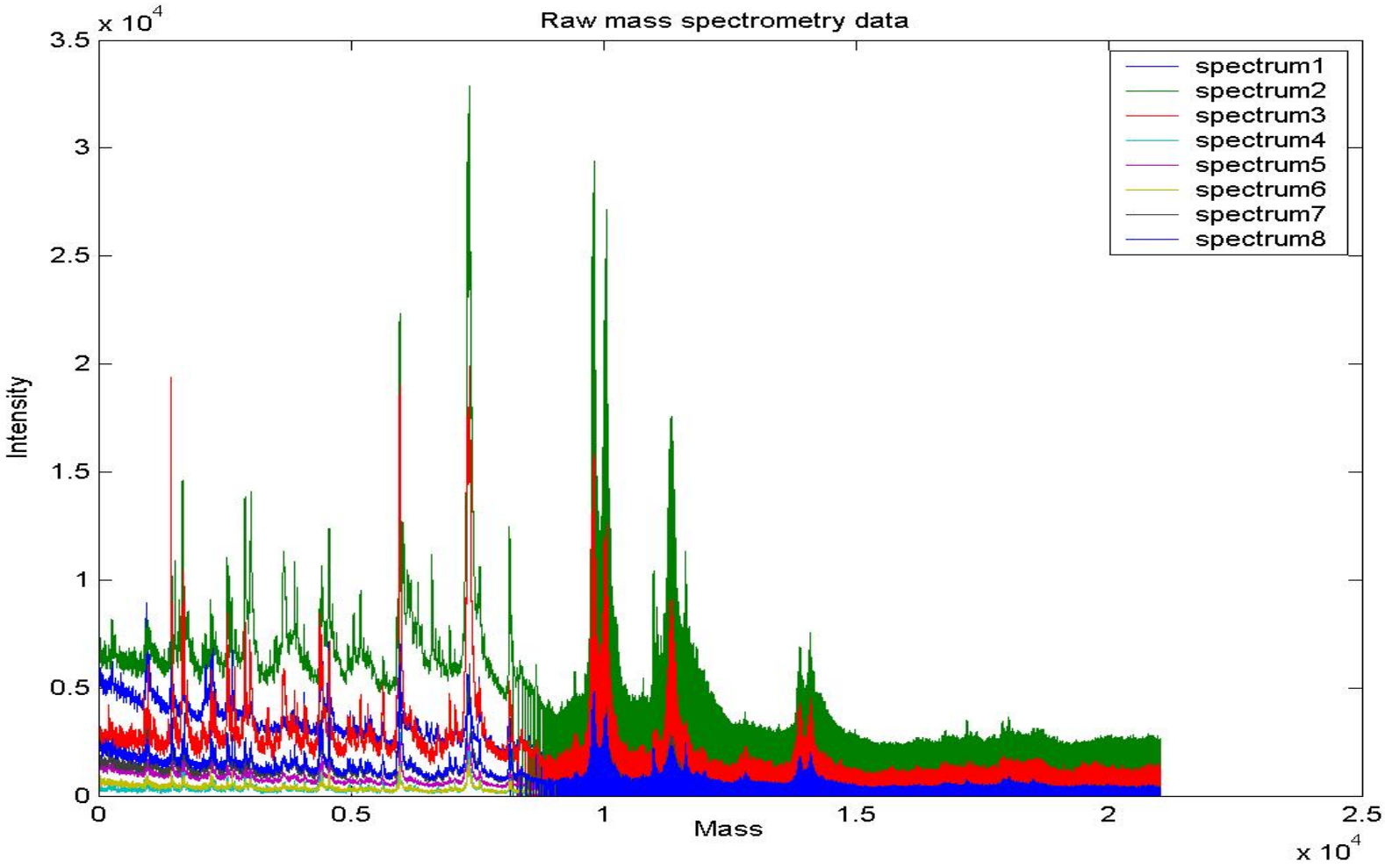
  Class Comparison (supervised)

  Class Prediction (supervised)

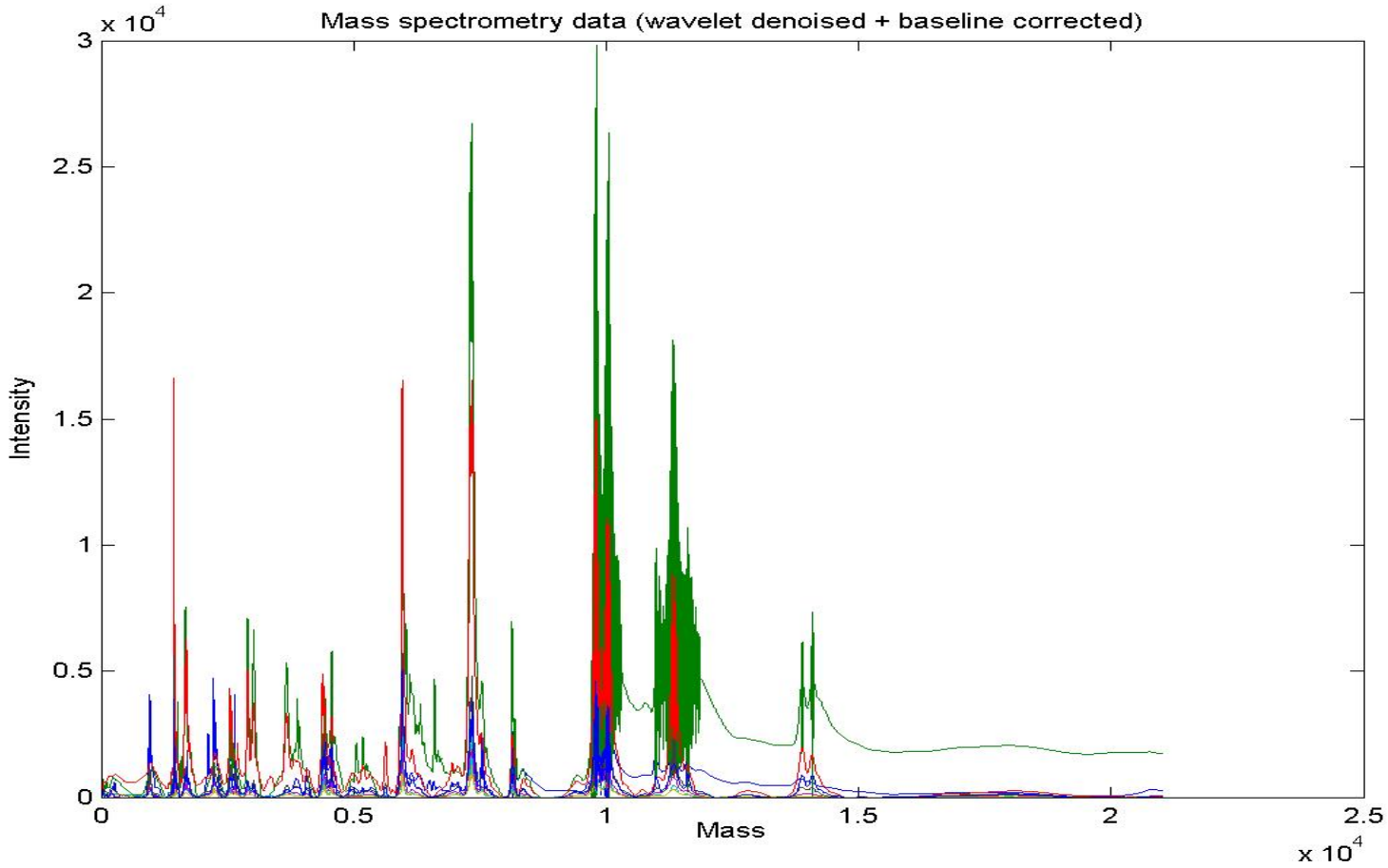
Outcome Measurement



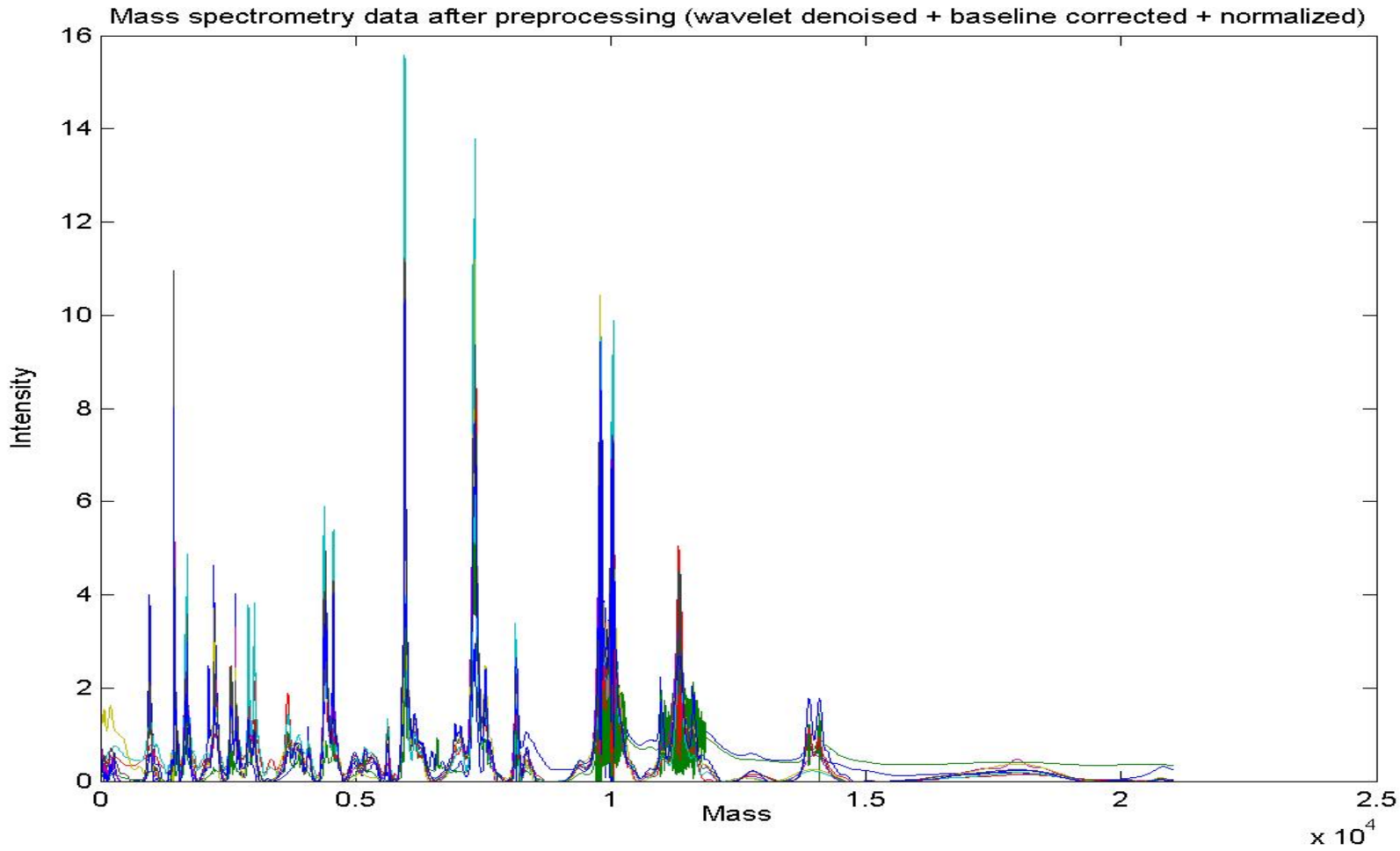
Pre-processing (MALDI-TOF)



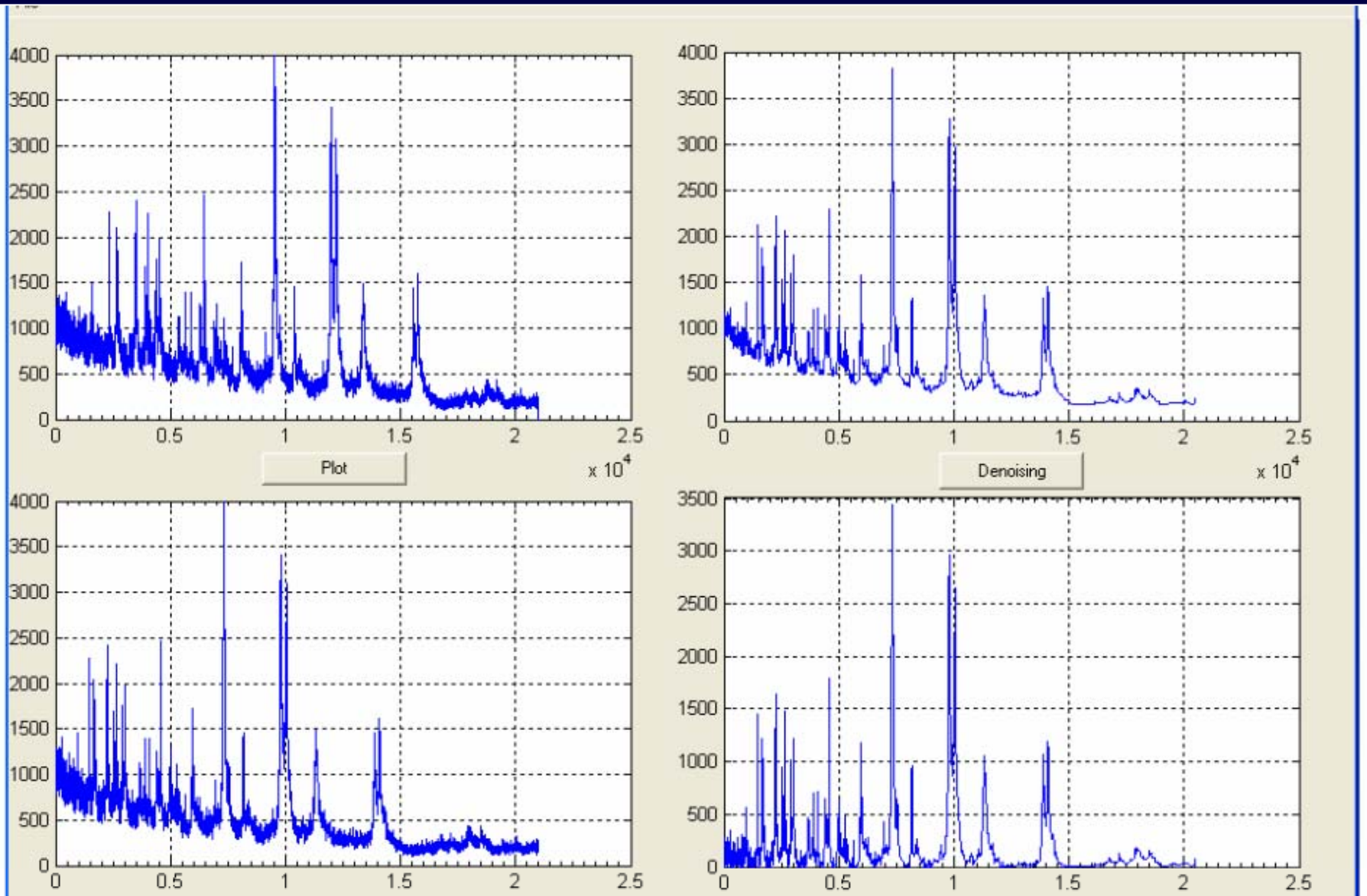
Pre-processing
Wavelet denoise + baseline correction



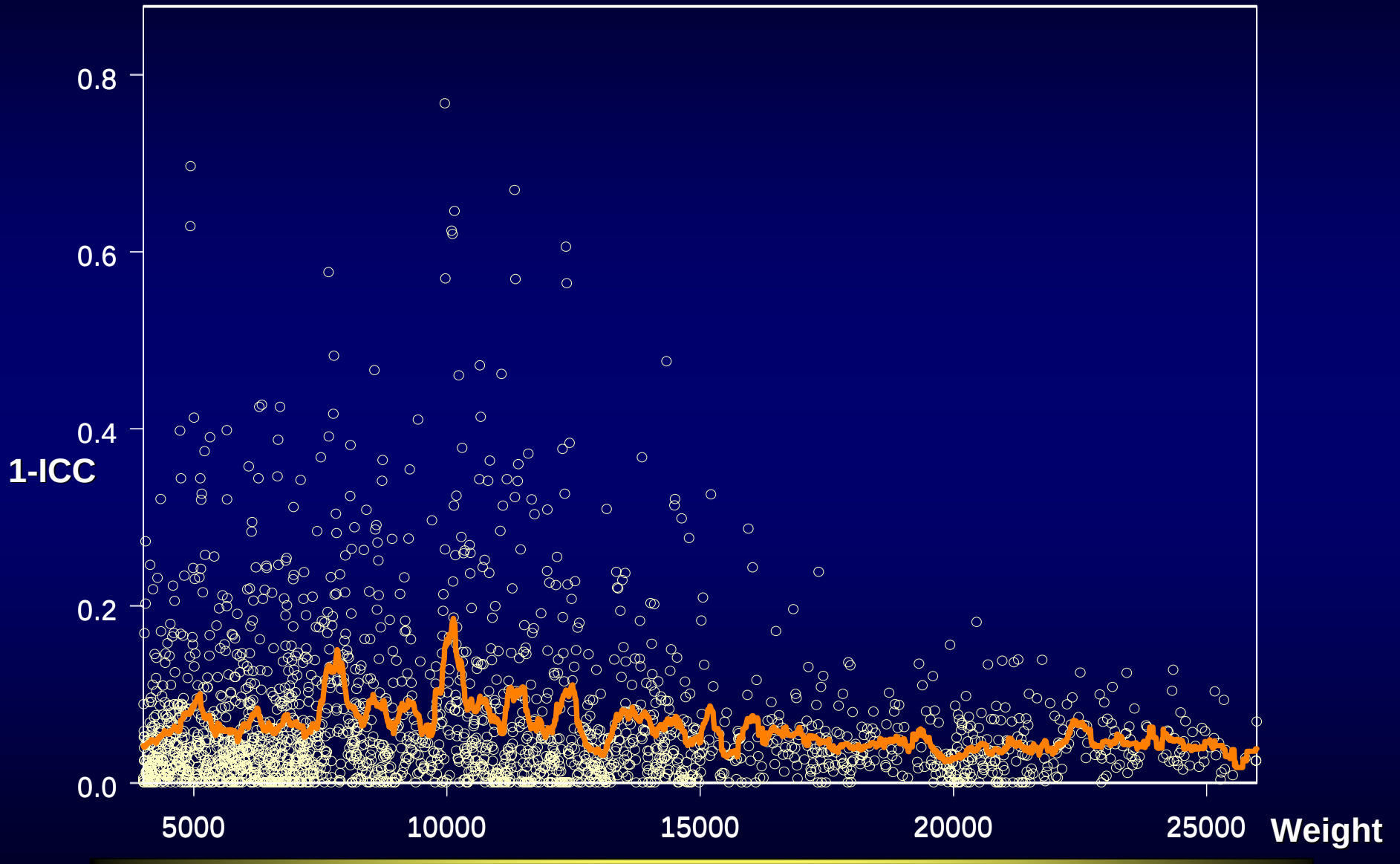
Pre-processing
Wavelet denoise + baseline correction + normalization



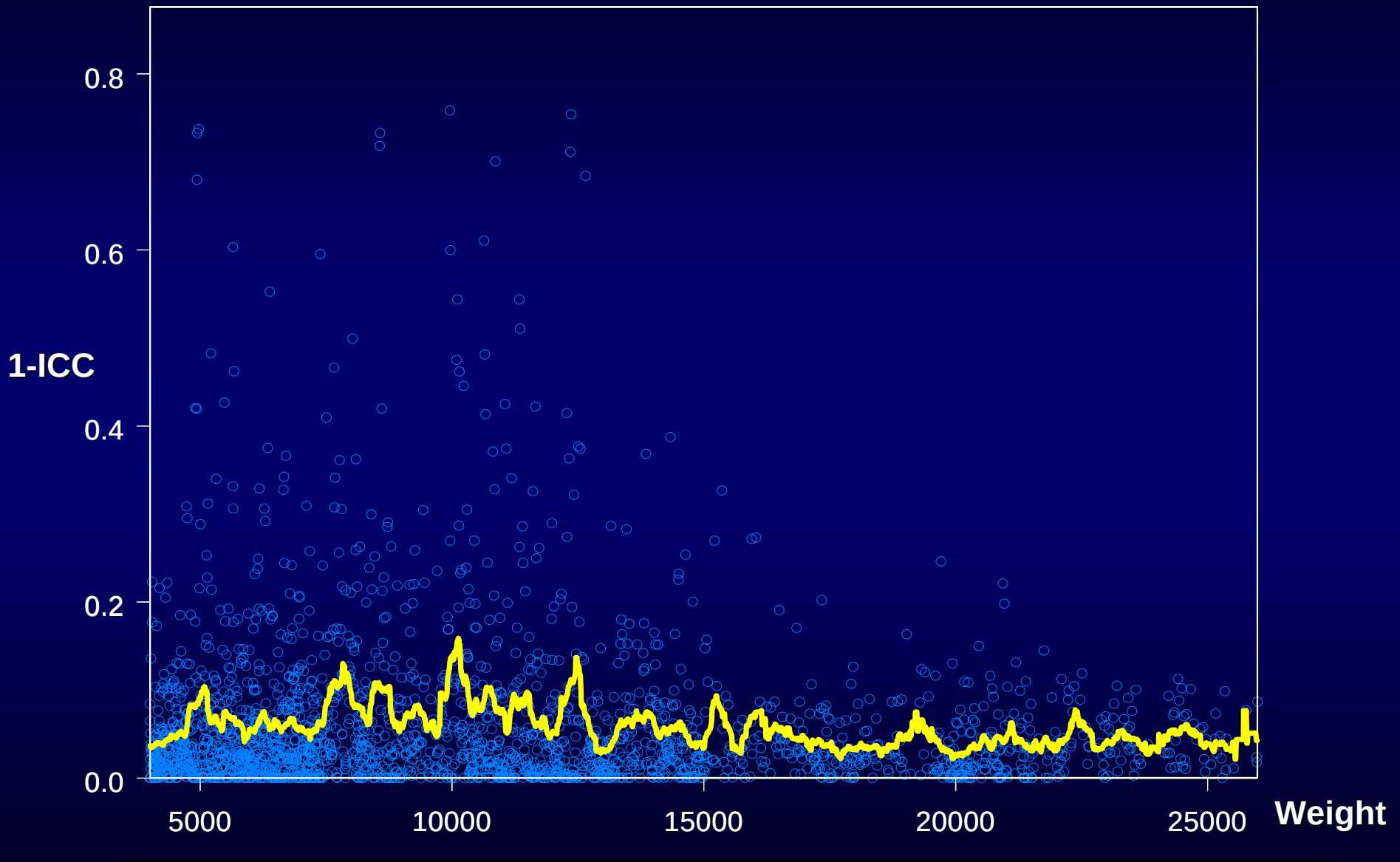
MALDI-TOF MS Data Preprocessing Tool



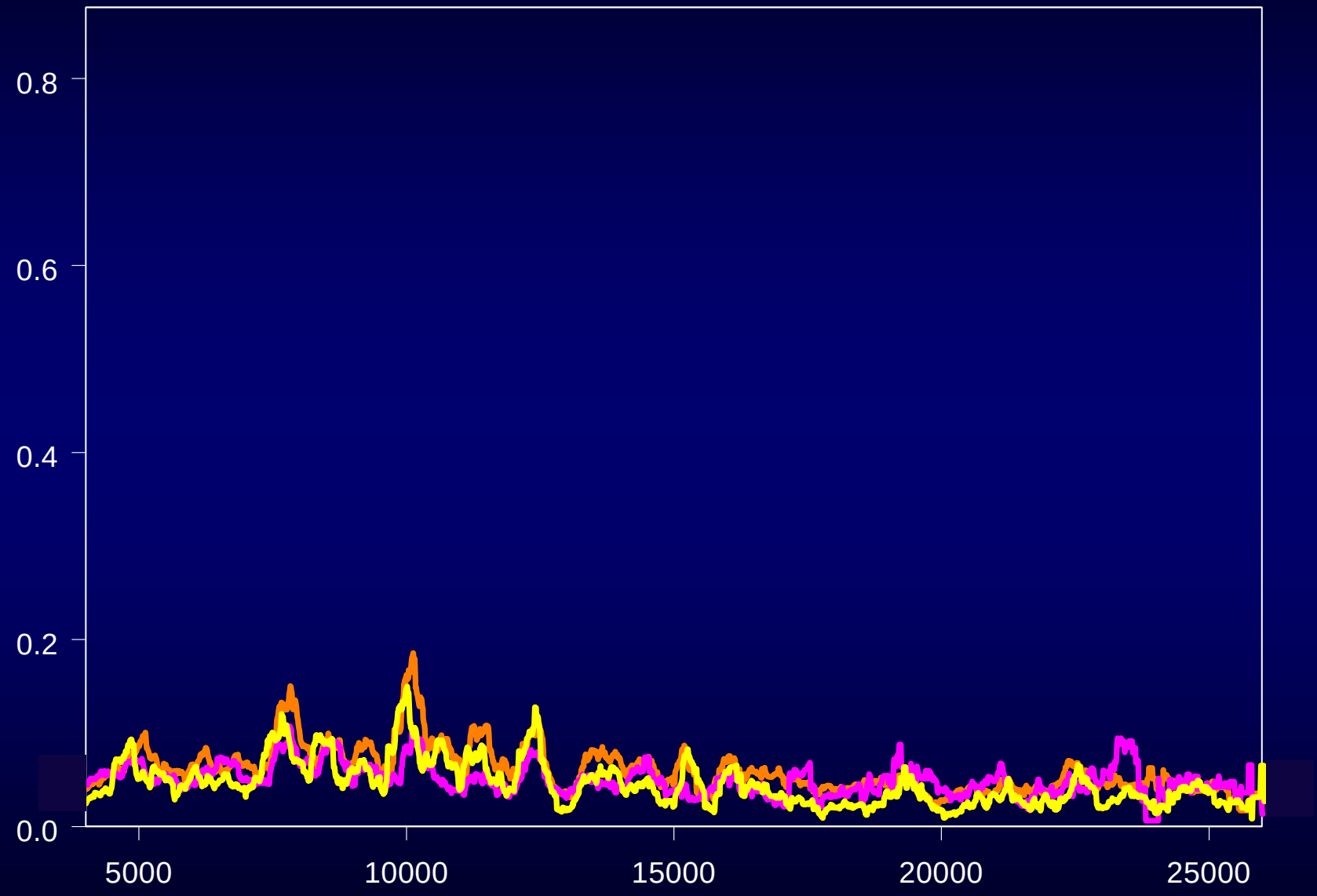
Intra Class Correlation: Training



Intra Class Correlation: Training and Testing Combined



Intra Class Correlation: Training, Testing, and Combined



Dimension Reduction

Possible approaches:

- **Principal Component Analysis (PCA)**
- **Multidimensional scaling (MDS)**
- **Self-Organizing Map (SOM)**

Feature Selection - Class Comparison

- t-test, permutation t-test, permutation F test.
- Weighted Gene Analysis
- Threshold Number of Misclassification Score (*TNoM*)
- Mutual-information Scoring (Info Score)
- Significance Analysis of Microarray (SAM)
- REML based Mixed effect model
- The P-values for Identifying Differentially Expressed genes (PIDEX)

Feature Selection - Class Comparison

- Tree Algorithms: CART, Quest, Slip, CHAID
- Projection Pursuit Regression (PPR)
- Partially Lease Square Method (PSS)
- Smoothing Spline
- Knowledge Extraction Engines (KXEN)
- Multivariate Adaptive Regression Splines (MARS)
- TreeNet: Stochastic Gradient Boosting (MART)

Classification - Compound Covariate Method

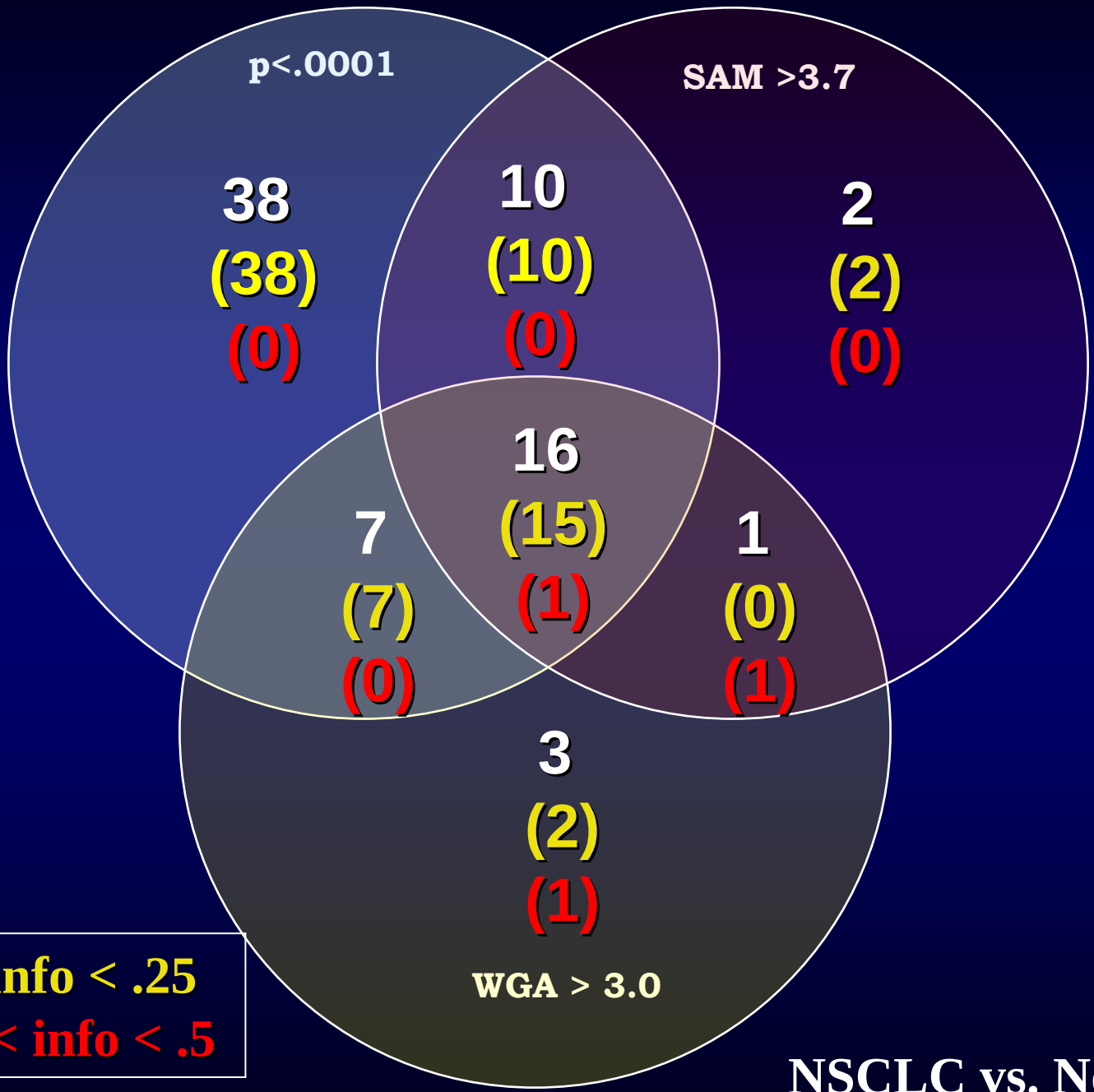
- The compound covariate method was proposed by Tukey (1993). Hedenfalk et al. (2001) successfully applied this method to class prediction analysis for BRCA1+ vs. BRCA1-.

- This predictor is built in two steps.

First, a statistical test is performed to identify genes with significant differences (at level α, Hedenfalk et al. picked $\alpha = 0.0001$) between the two tissue classes.

Second, the ratios of differentially expressed genes are combined into a single compound covariate for each tissue sample; the compound covariate is used as the basis for class prediction.

$$C_i = \Sigma_j M_j X_{ij}$$



Yellow: info < .25
Red: .25 < info < .5

NSCLC vs. Normal: 77

Classification
Weighted Flexible Compound Covariate method

- We have proposed a more flexible compound covariate method (**Weighted Flexible Compound Covariate method**) based on the mutual-information scoring (*Info Score*) , significance analysis of microarrays (SAM), Weighted gene analysis, Fisher's exact test, Mixed effect model and permutation *t*-test.

WFCCM: WFCCM is an extension of the compound covariate method which allows considering more than one statistical analysis methods into the compound covariate.

The WFCCM for tumor sample *i* is defined as

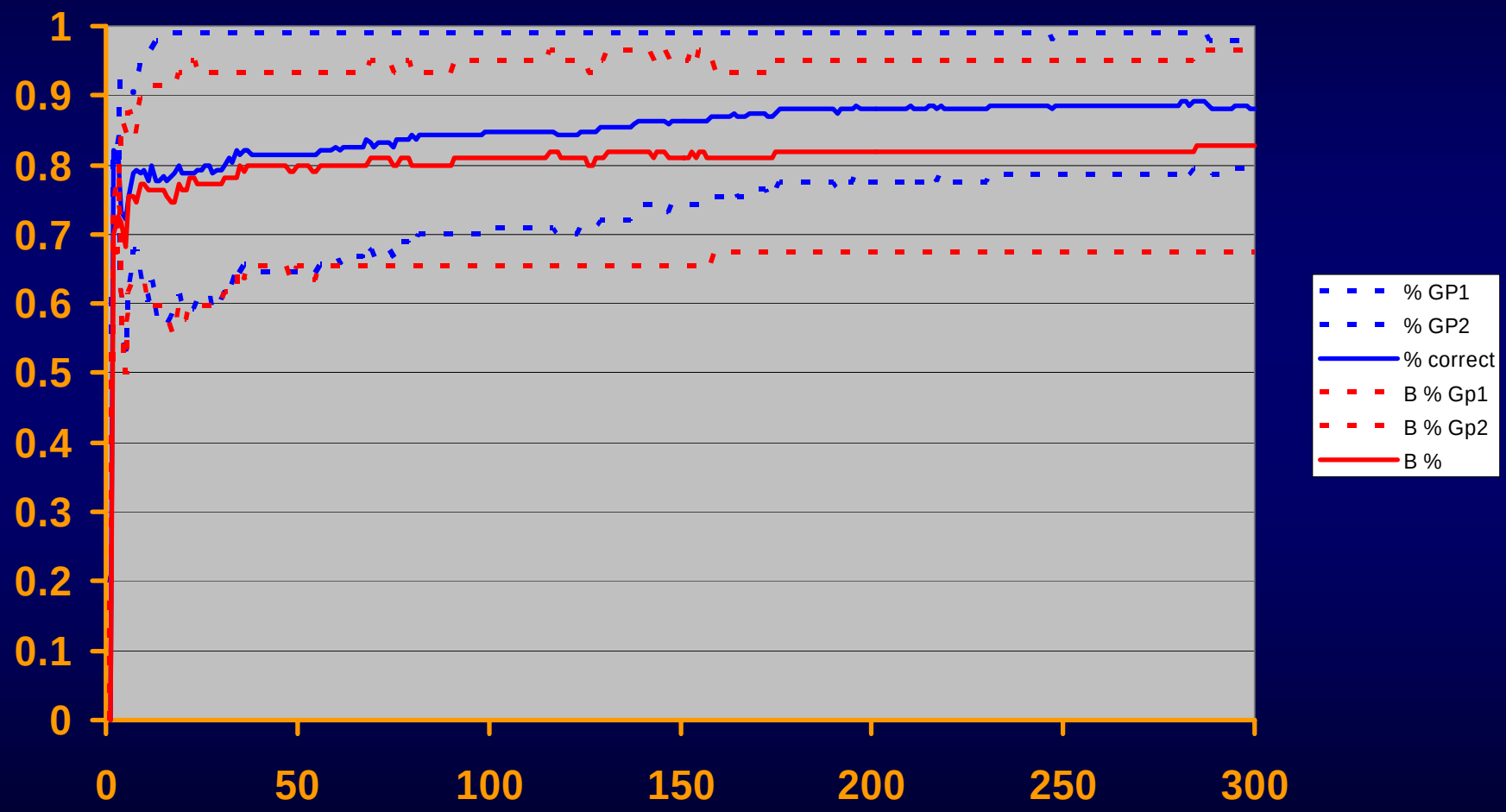
$$WFCCM(i) = \Sigma_j [\Sigma_k (ST_{jk})] [W_j] X_{ij},$$

where j represents statistically significant gene j. ST_{jk} is the standardized statistic, e.g., t-statistic, for statistical analysis method k.

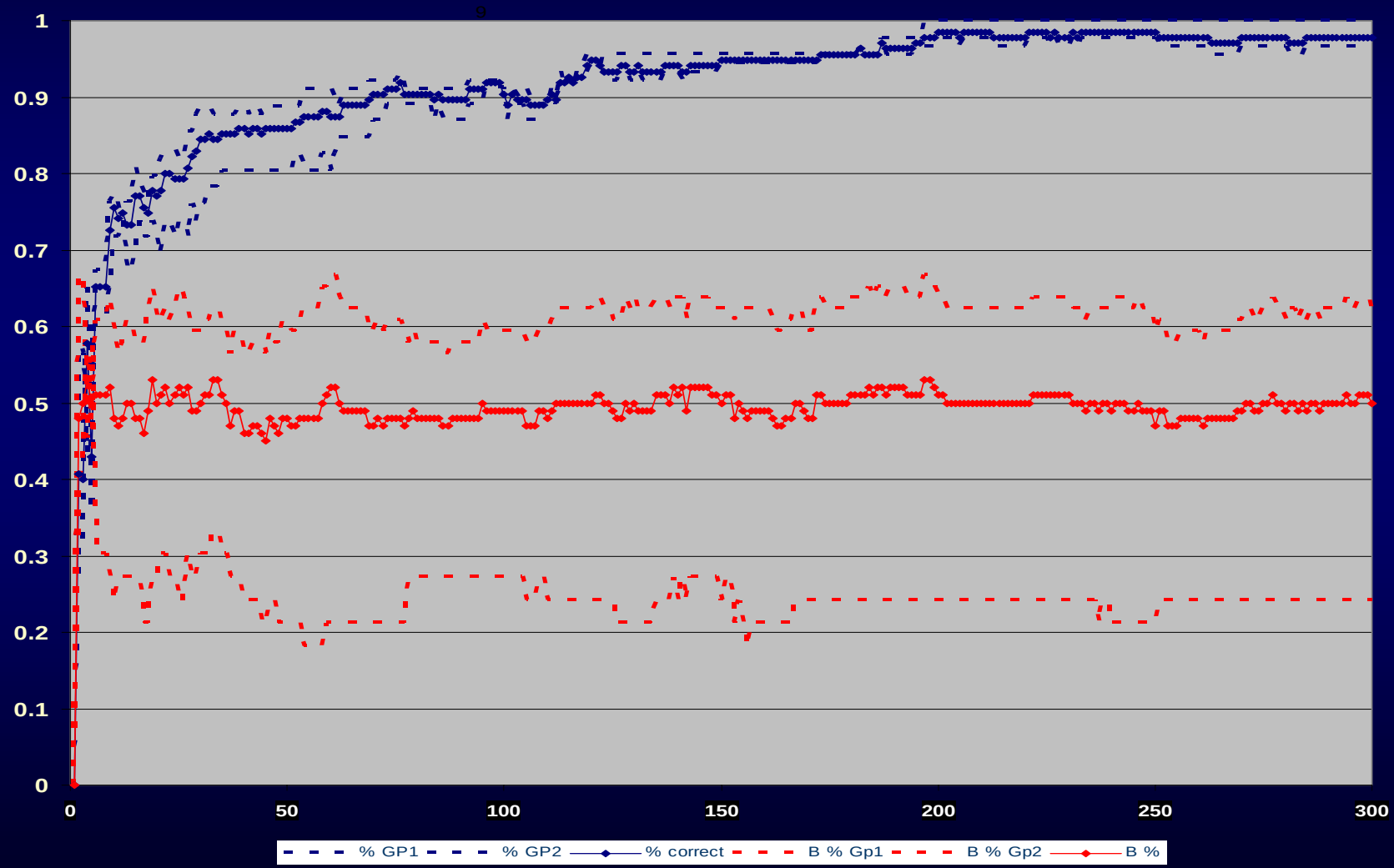
Validation

Classification (sample size)	No. of diff. Expressed genes	No. of Misclassified sample	% of random permutations with misclassification
Normal (3) vs. Tumor (26)	54	0 (Normal, 0) (Tumor, 0)	0.7% With ≤ 0 misclassification
Lung (24) vs. Non-lung (5)	62	0 (Lung, 0) (N-Lung,0)	0.5% With ≤ 0 misclassification

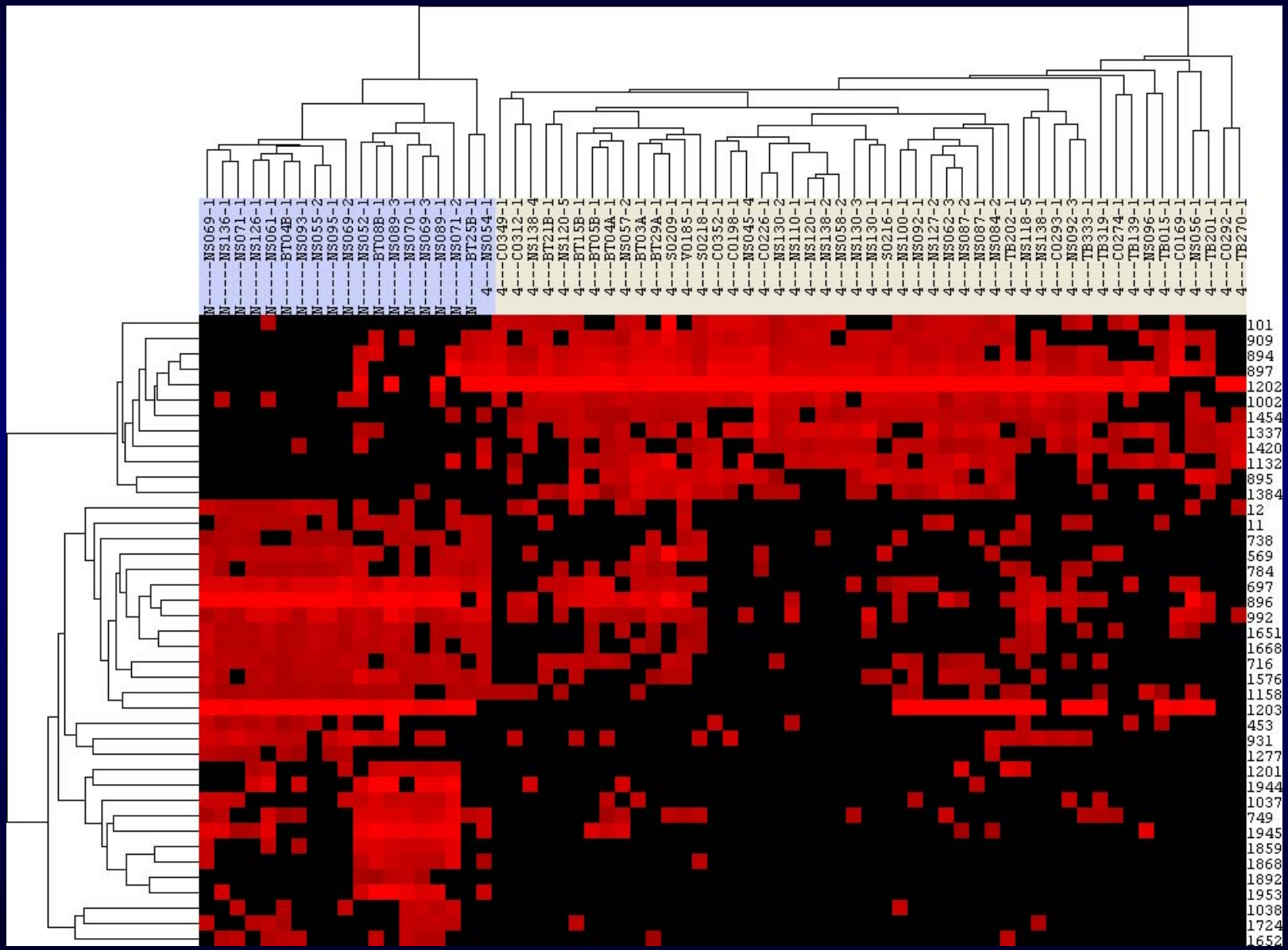
Testing Cohort
Lung SPORE Serum Proteomic Study

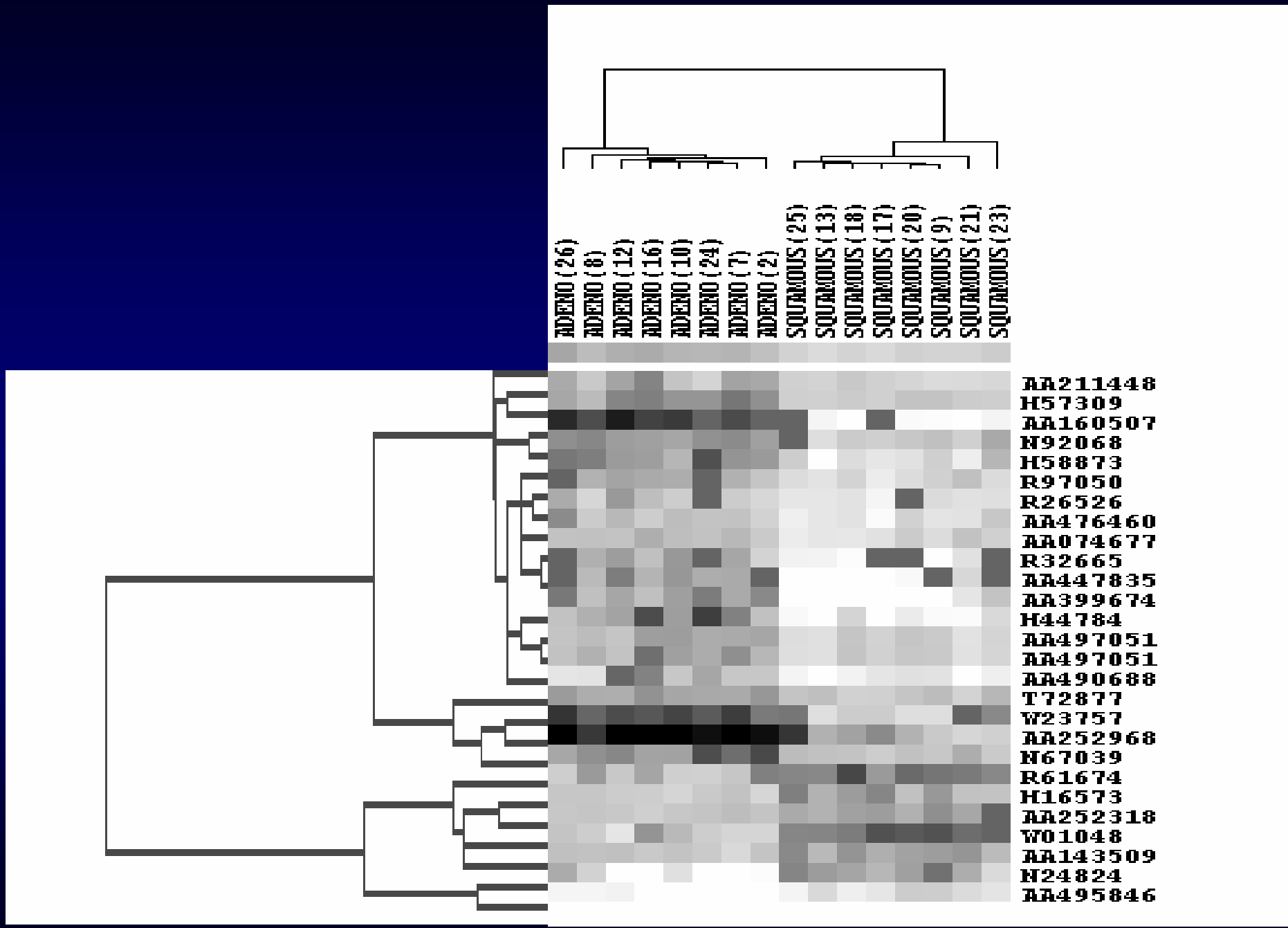


WFCCM – Class Prediction Model

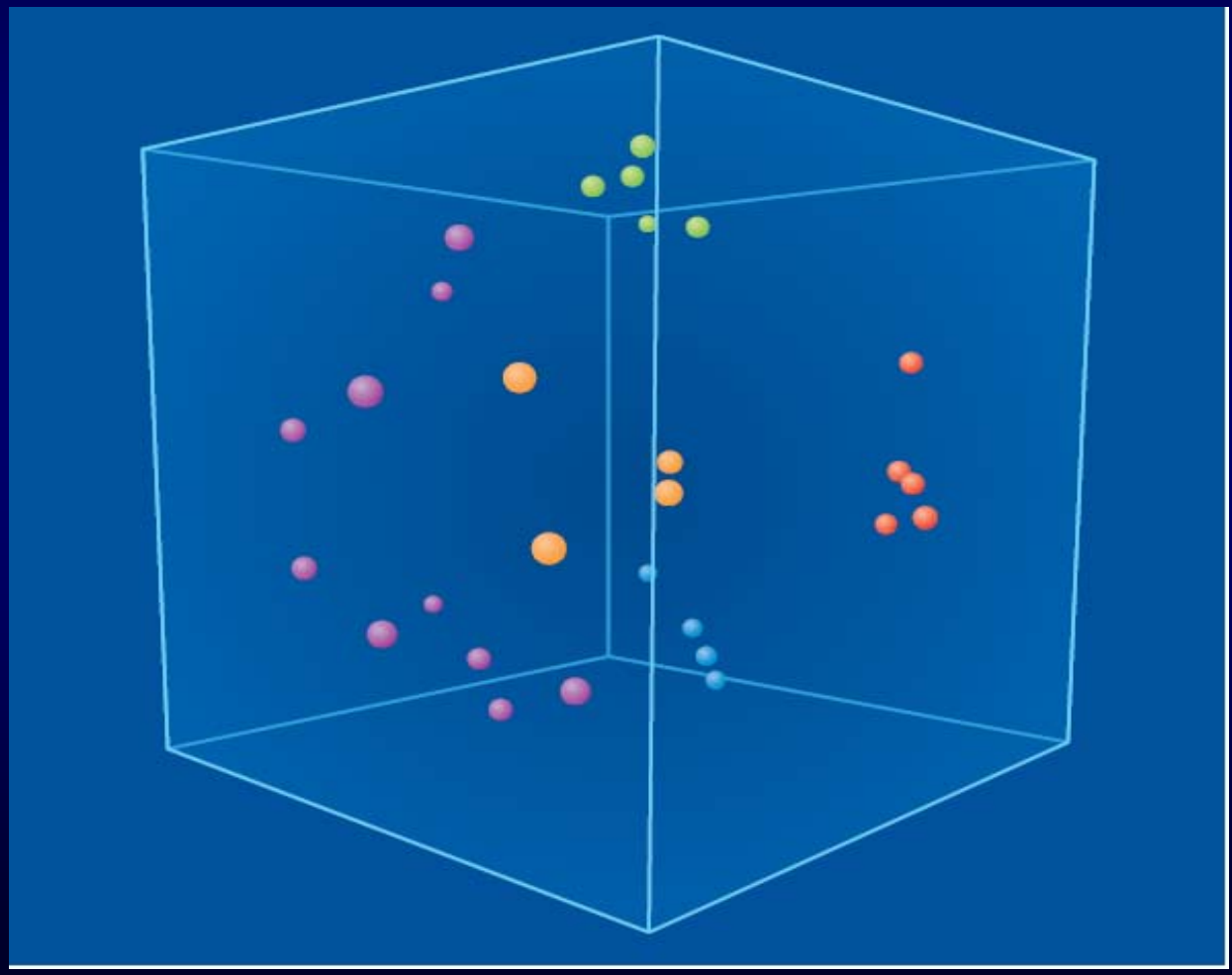


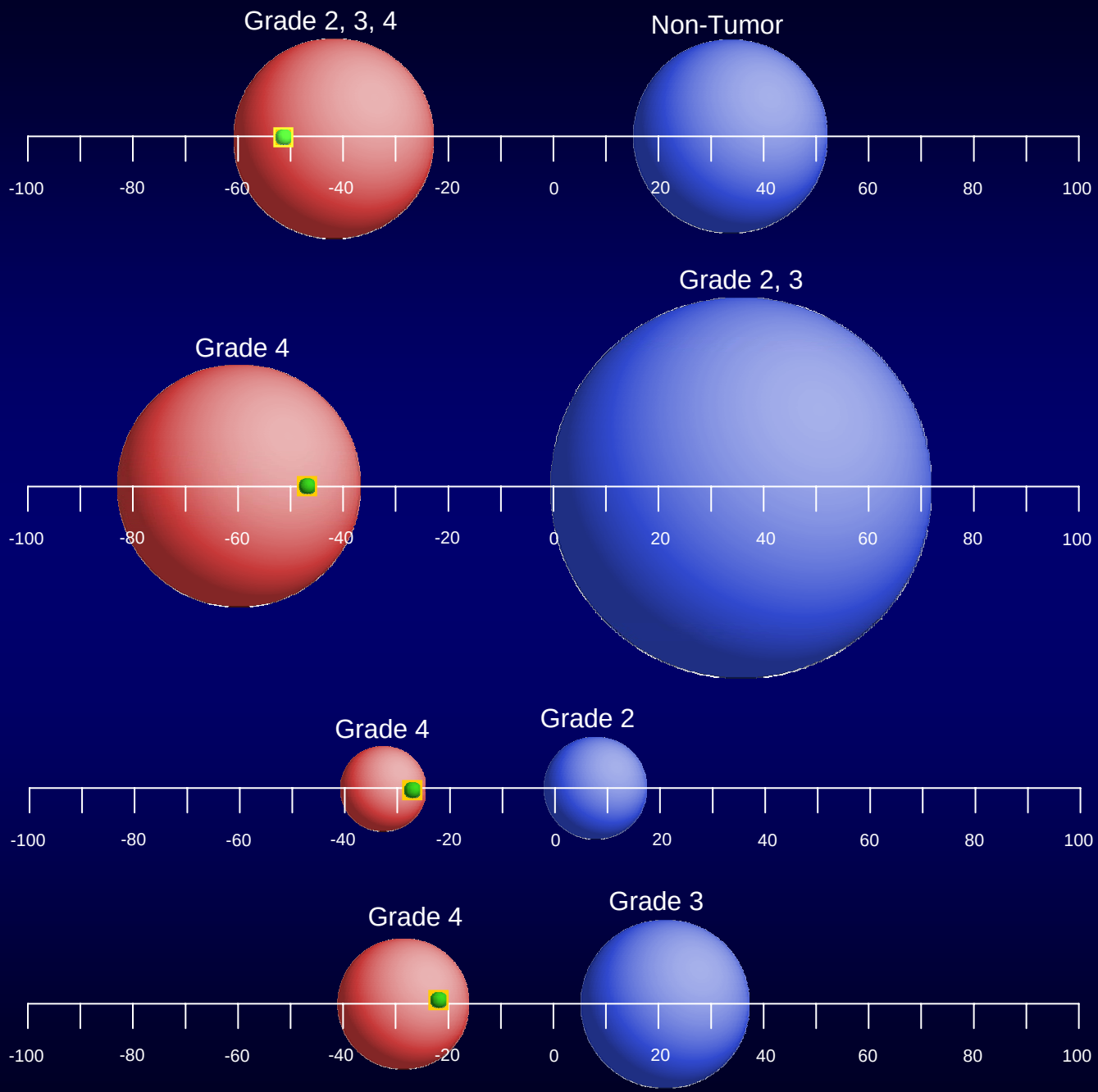
Non-Tumor T_{IV}

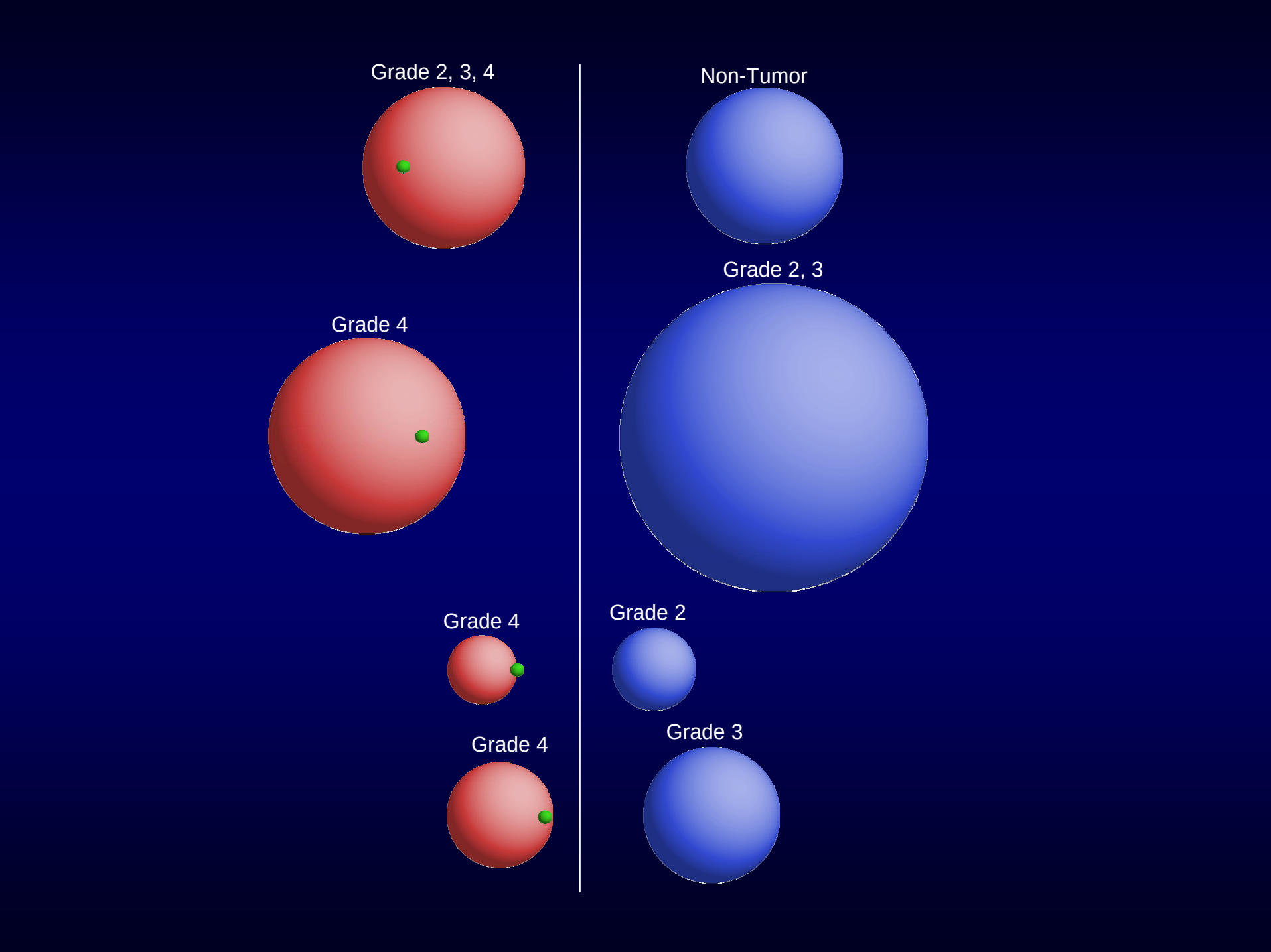



ADENO(26)
ADENO(8)
ADENO(12)
ADENO(16)
ADENO(10)
ADENO(24)
ADENO(7)
ADENO(2)
SQUAMOUS(25)
SQUAMOUS(13)
SQUAMOUS(18)
SQUAMOUS(17)
SQUAMOUS(20)
SQUAMOUS(9)
SQUAMOUS(21)
SQUAMOUS(23)
AA211448
H57309
AA160507
N92068
H58873
R97050
R26526
AA476460
AA074677
R32665
AA447835
AA399674
H44784
AA497051
AA497051
AA490688
T72877
W23757
AA252968
N67039
R61674
H16573
AA252318
W01048
AA143509
N24824
AA495846

Multidimensional scaling (MDS)

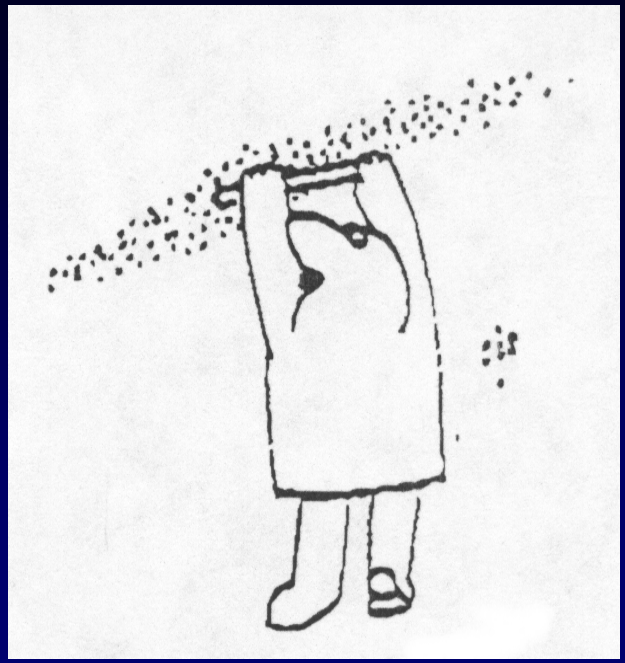


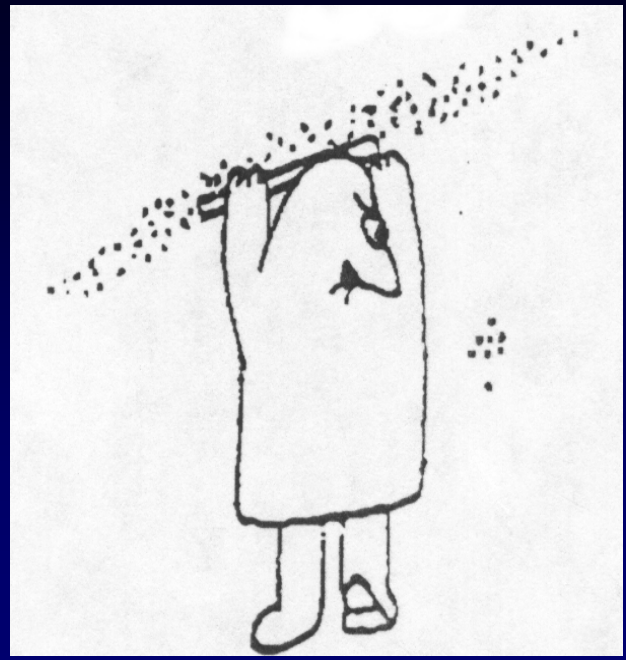

Grade 2, 3, 4
Non-Tumor
-100 -80 -60 -40 -20 0 20 40 60 80 100
Grade 2, 3
Grade 4
-100 -80 -60 -40 -20 0 20 40 60 80 100
Grade 4
Grade 2
-100 -80 -60 -40 -20 0 20 40 60 80 100
Grade 4
Grade 3
-100 -80 -60 -40 -20 0 20 40 60 80 100

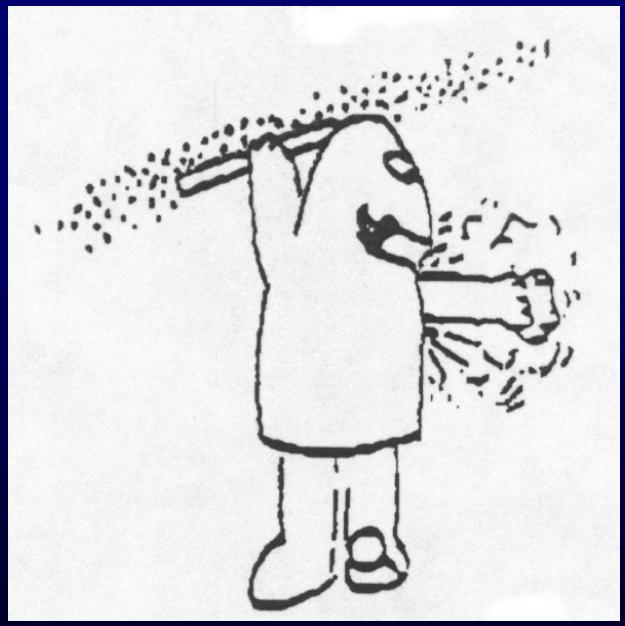

Grade 2, 3, 4
Grade 4
Grade 4
Grade 4
Non-Tumor
Grade 2, 3
Grade 2
Grade 3

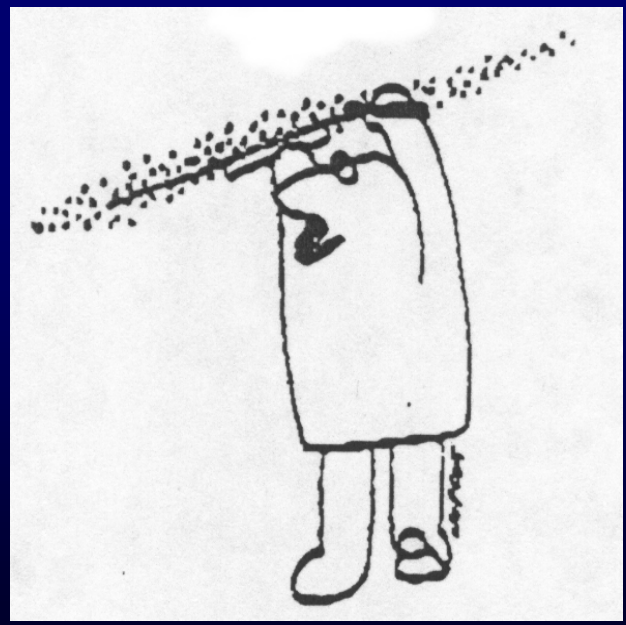
Things DON'T DO

- Fold-change for feature selection
- Cluster analysis for class comparison
- Cluster analysis for class prediction
- Extremely small sample size for the Independent test cohort
- Only report the good news









END