Dimensionality Reduction of High-throughput Proteomic Data using Partial Discriminative Projections

Michal Valko
University of Pittsburgh

Joint work with
Dr. Milos Hauskrecht
Richard Pelikan
Guang Xiang

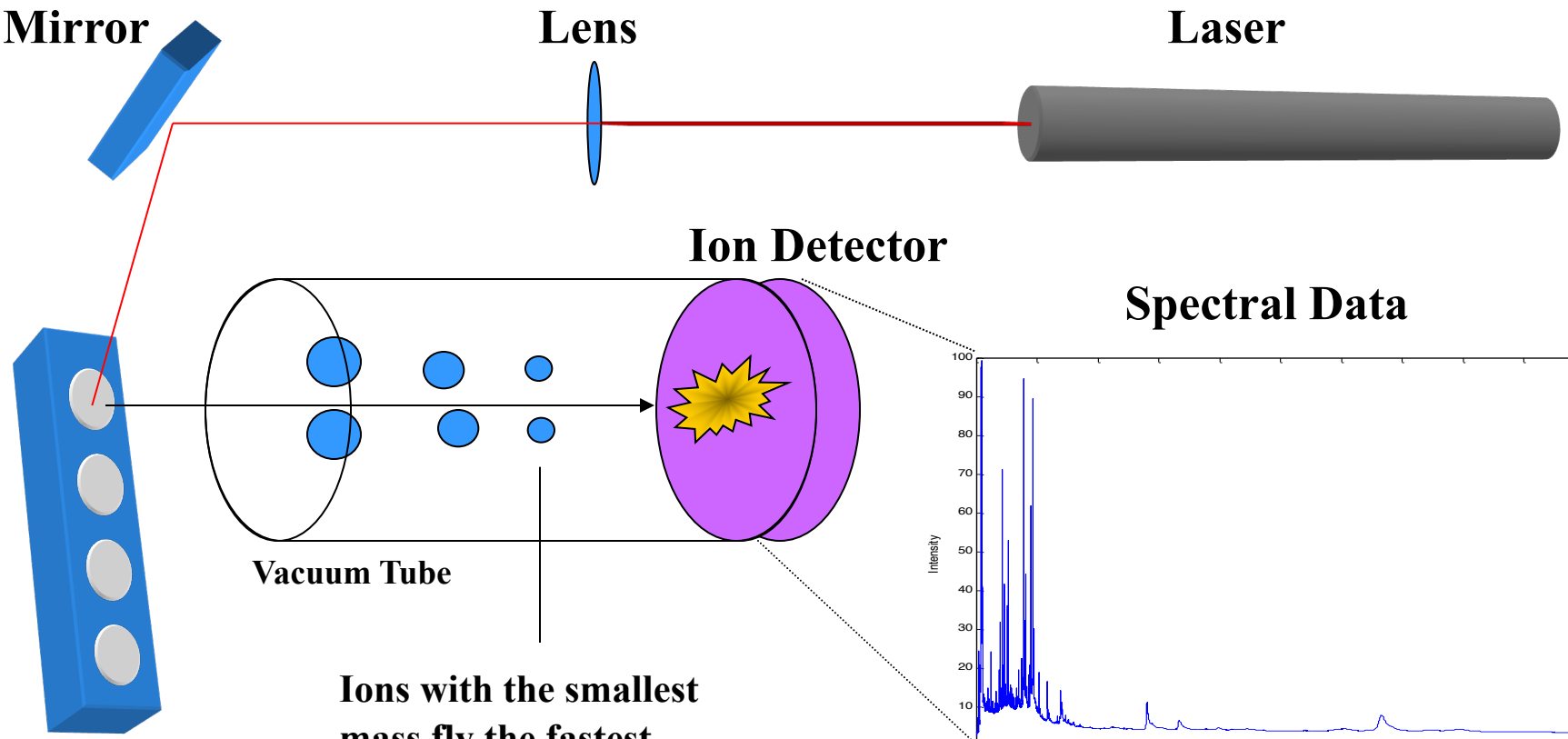
ai seminár, 23. mája 2006, 1. akvárko, fmfi

SELDI-TOF MS proteomic profiling

- **Surface Enhanced Laser Desorption/Ionization Time of Flight Mass Spectrometry**
- A technology developed by Ciphergen Biosystems
- A mass profile of a sample (serum, urine, cell lysates) in the range of 0-200K Daltons
- A profile is believed to provide a rich source of information about a patient. Potentially useful for:
- **Detection of a disease**
  - Many promising results on various types of cancer
- **Discovery of disease biomarkers**
  - Identify species (protein/peptides) responsible for the differences

SELDI-TOF MS

Surface Enhanced Laser Desorption/Ionization Time of Flight Mass Spectrometry



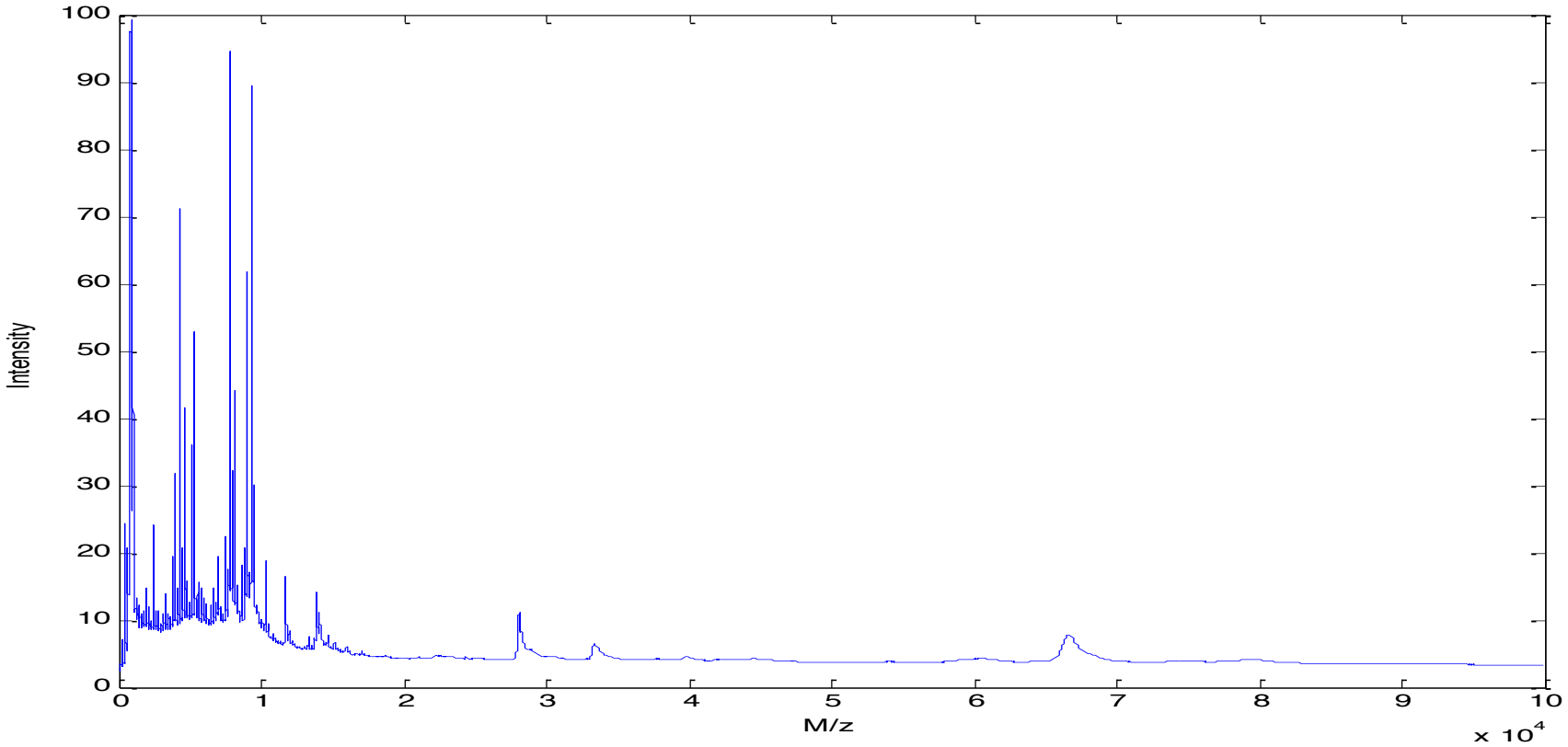
Ions with the smallest mass fly the fastest, and are detected first.

Sample (serum, cell lysates, urine)

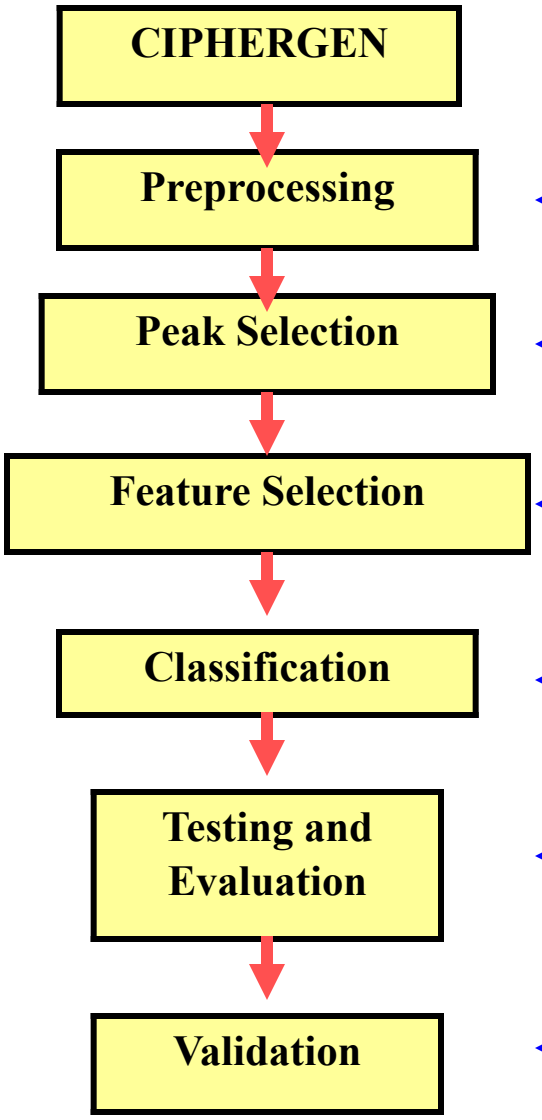
Reflects the mass of proteins, peptides, nucleic acids in the sample

SELDI-TOF MS

Data reflect the expression
level of proteins/peptides in the sample



Analysis of SELDI-TOF-MS



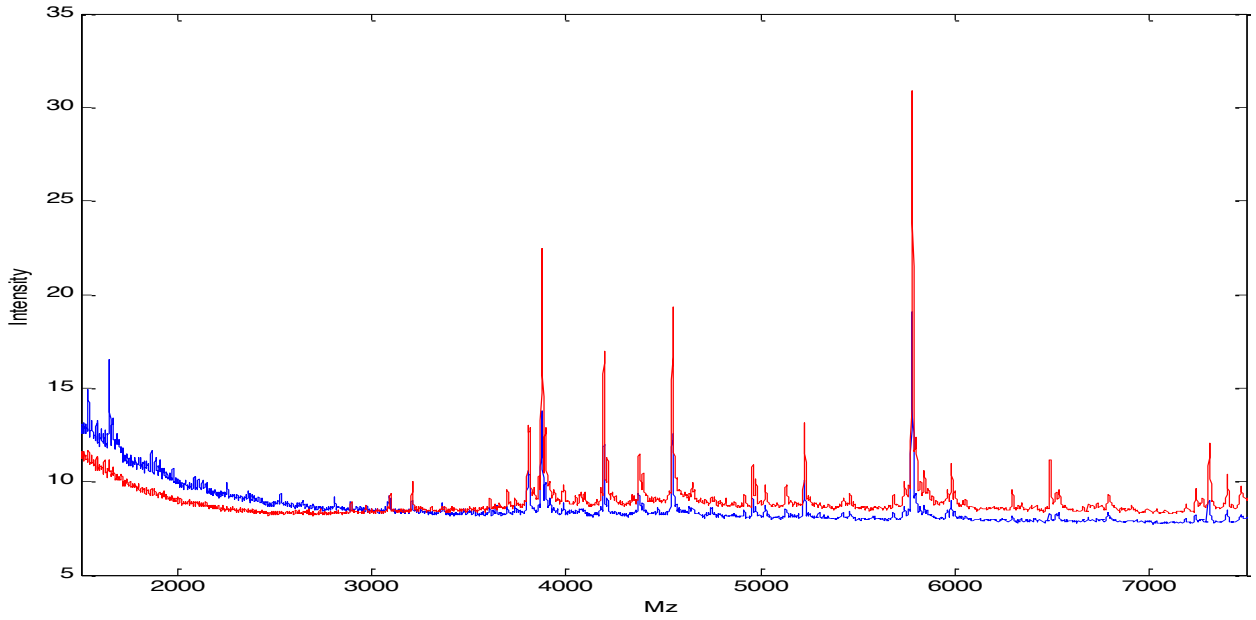
- Removal of systematic (instrument) biases and noise
- Identify and quantify spikes in profiles
- Identify profile positions or regions relevant for building a classification model
- **learn** a classification model
- Determine the quality (accuracy) of the classification model
- Check the reproducibility of detection results on independent datasets

Challenges in data analysis

Many sources of variation Sample collection, sample storage & sample processing Instrument conditions + natural variation in protein expression levels among individuals

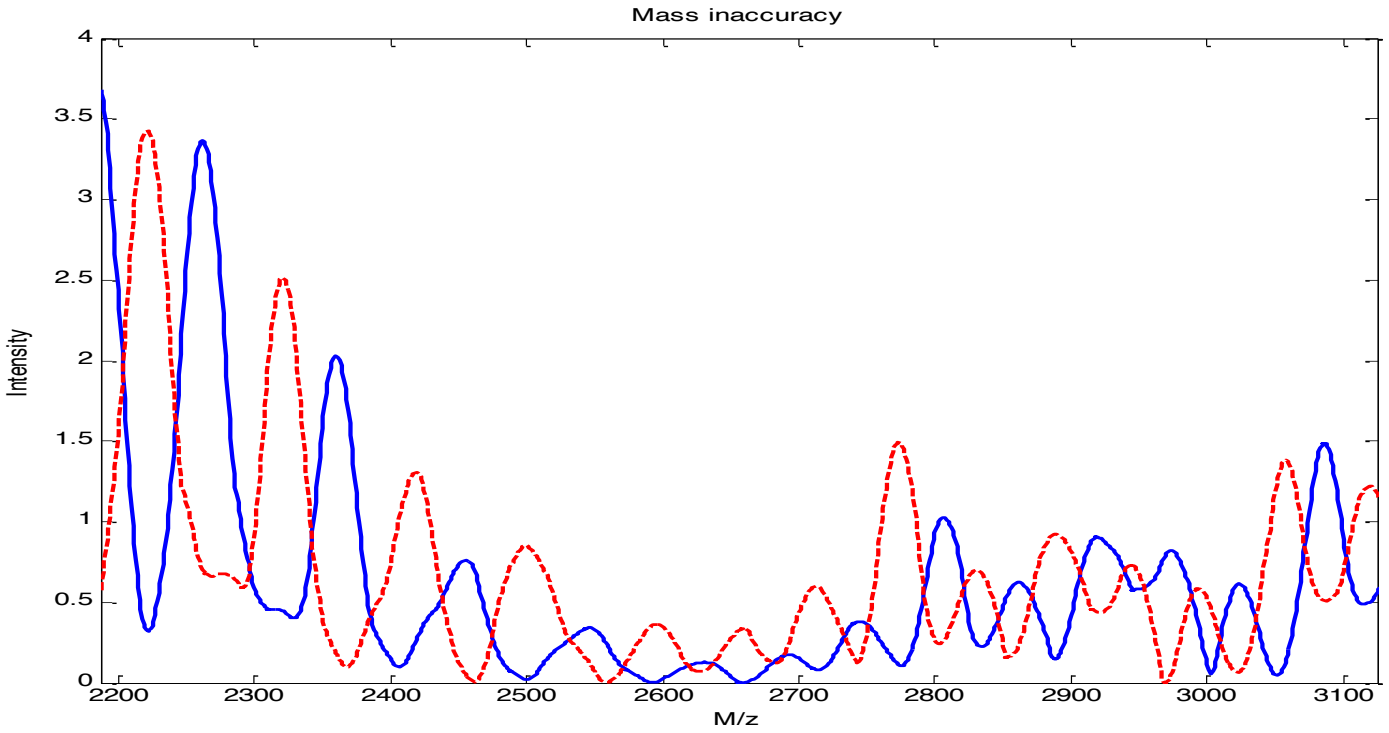
Three types of systematic instrument errors Baseline shift, Intensity, measurement error, Mass inaccuracy

Example: two profiles for the same reference serum (EDRN)



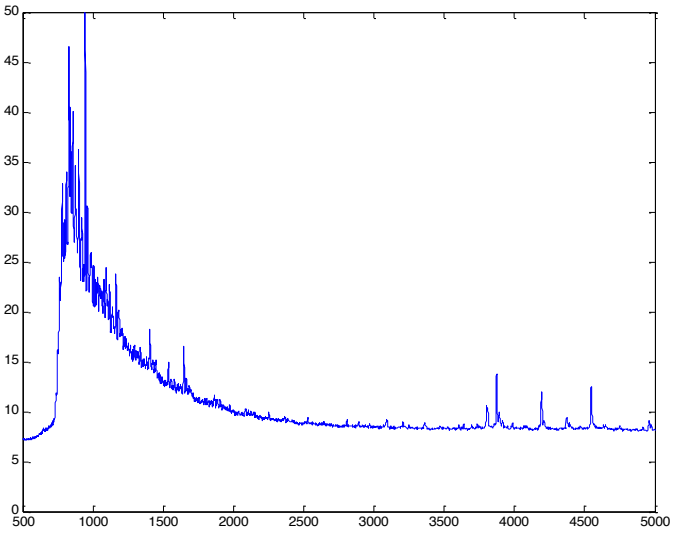
Mass inaccuracy

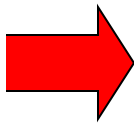
- Misalignment of readings for different *m/z* values.
- **Example:**
  - 2 profiles of the same pooled reference (QA/QC) serum

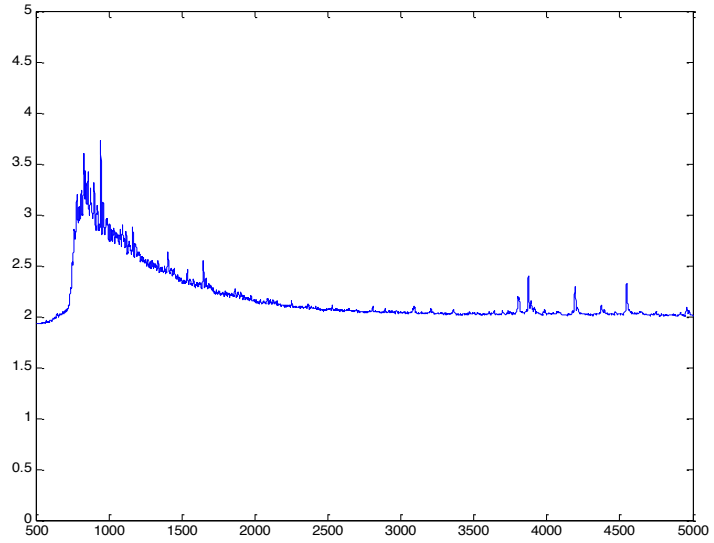


Preprocessing: variance stabilization

- **Aims to eliminate** the dependency of the noise on the intensity of the signal;
- **Example transforms used in variance stabilization:**
  - Log transform
  - Square-root transform
  - **Cube-root transform** - appears to work the best (suggested by J. Morris from MD Anderson).

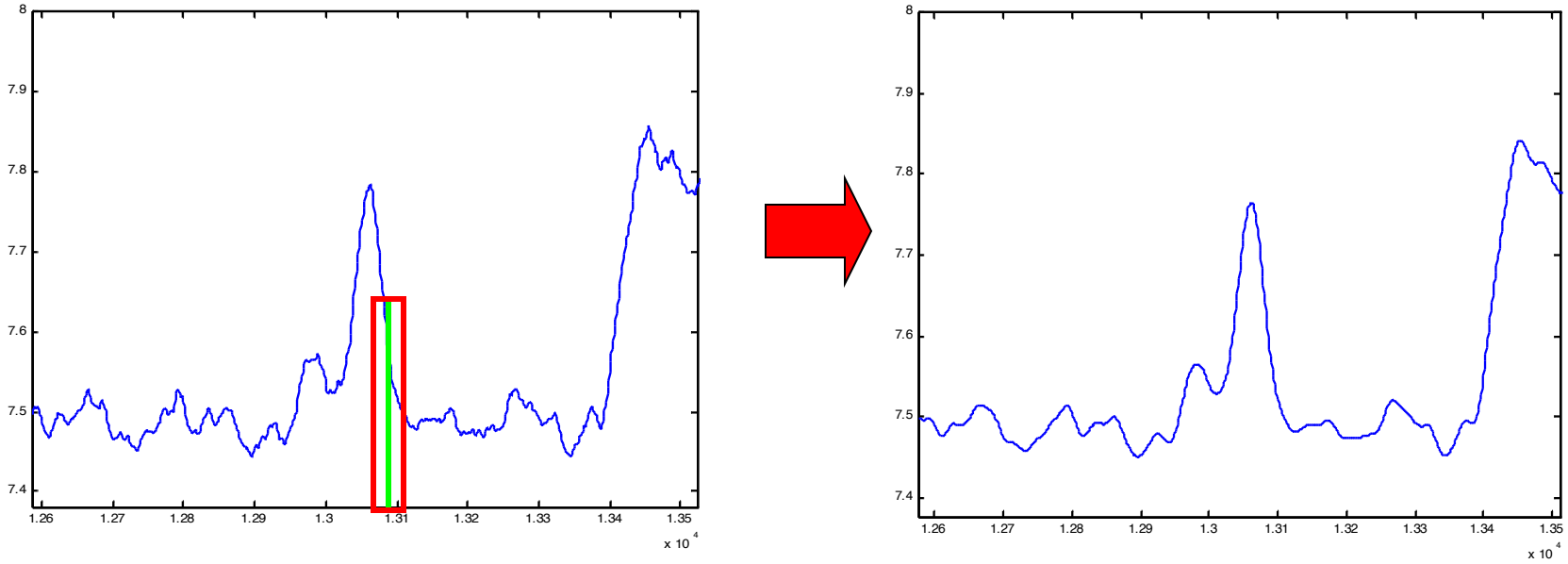






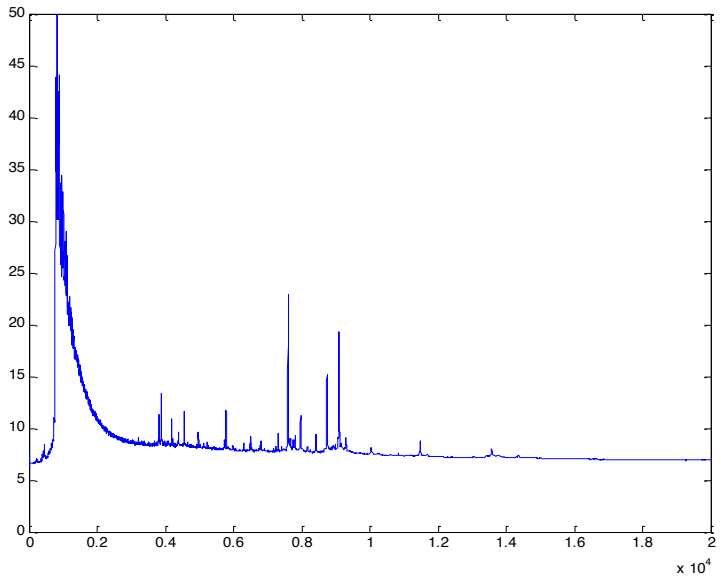
Preprocessing: Smoothing

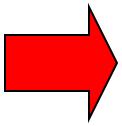
- **Aims to** eliminate a high-frequency noise in the signal
- **Threat:** a loss of information in the high frequency signal
- **Gaussian kernel smoothing:**
  - Weighted average of neighbors of a point
  - weights determined according to the Gaussian distribution

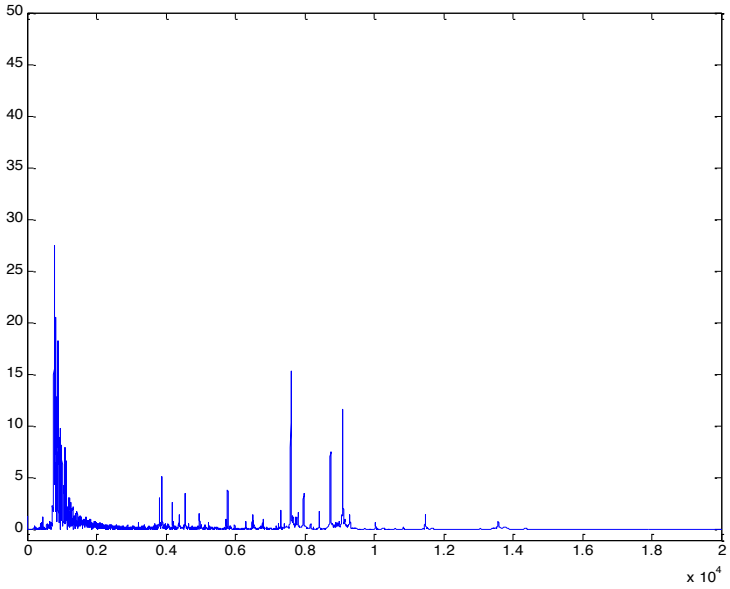


Preprocessing: Baseline correction

- **Aims is to** eliminate a systematic intensity bias by which the profile readings differ from 0

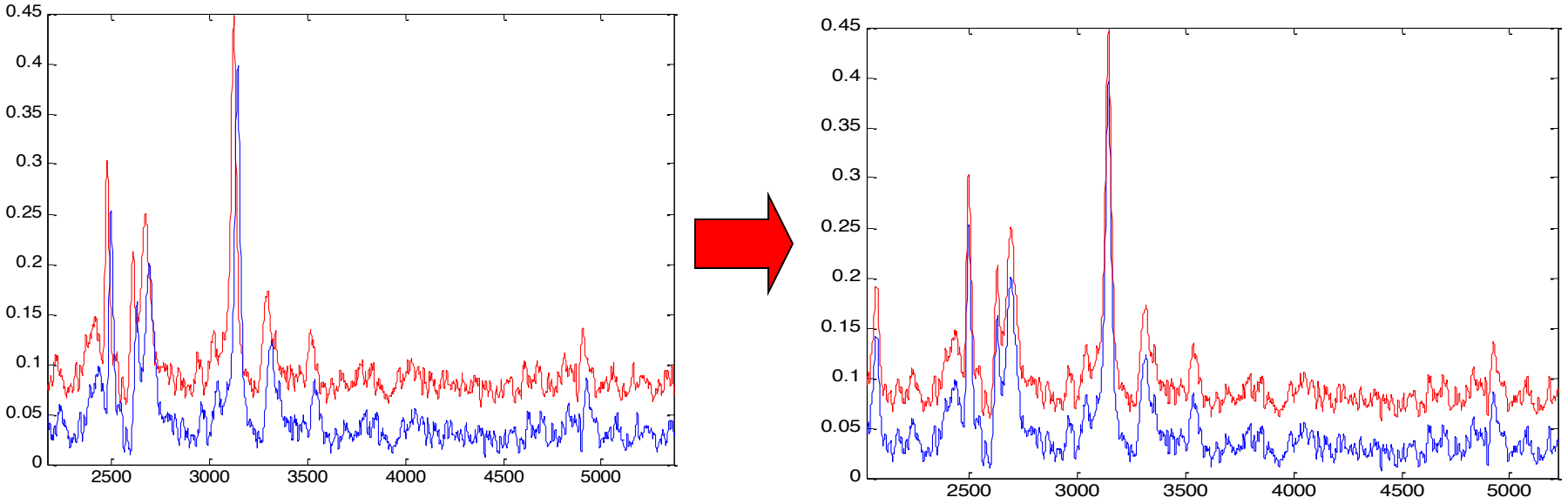






Preprocessing: Alignment

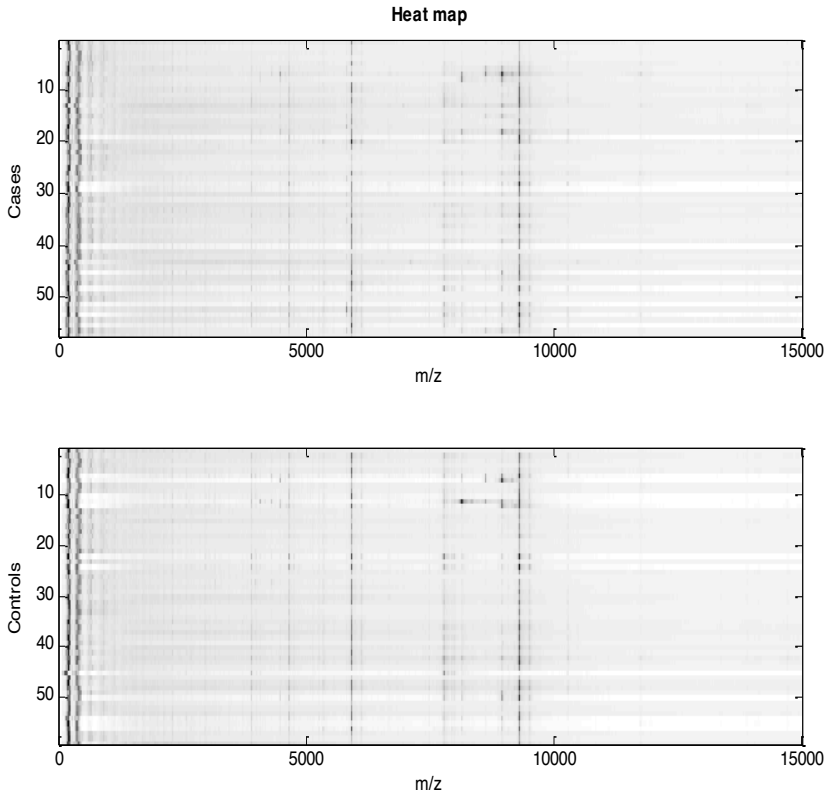
- **Aims to reduce** the noise (shifts) on x axis
- **Alignment procedures:**
  - Dynamic time-warp (DP based)
  - Parametric and semi-parametric warps
  - **Peak-based alignment** (identify peaks in profiles and align them to peaks in the mean profile)



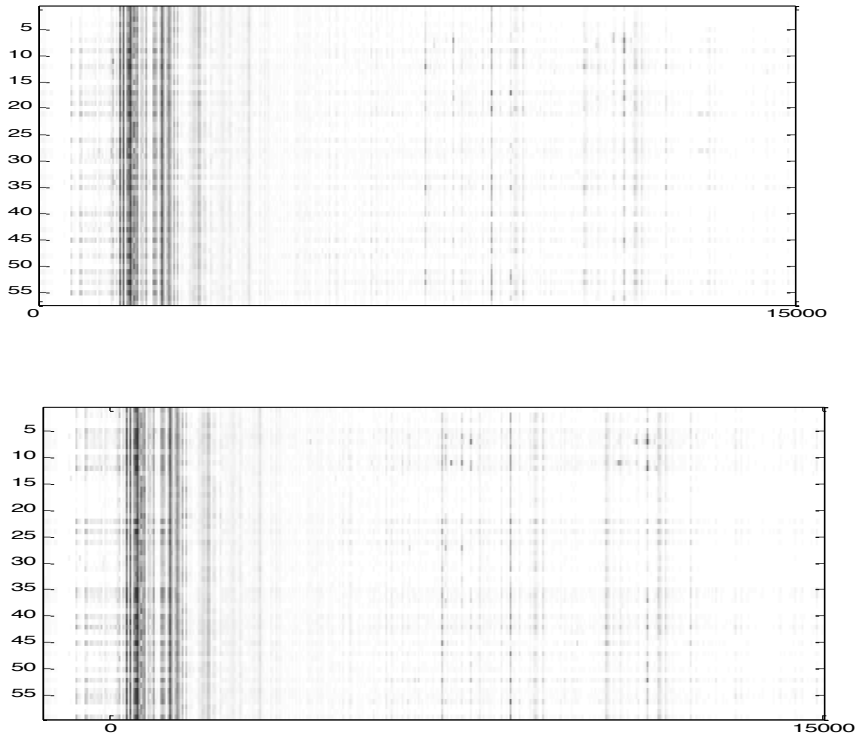
Preprocessing: Pancreatic cancer data

57 cases, 59 controls (courtesy of Drs. Whitcomb, Zeh and Bigbee)

Before preprocessing

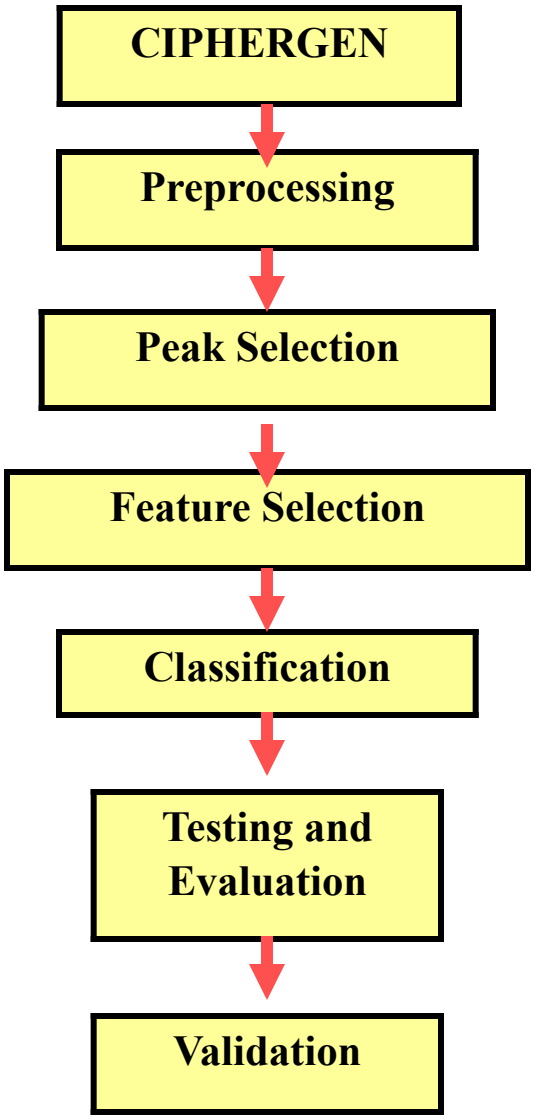


After preprocessing



The profiles' structure becomes crisp and more visible

Analysis of SELDI-TOF-MS



Detection of a disease

Objective:

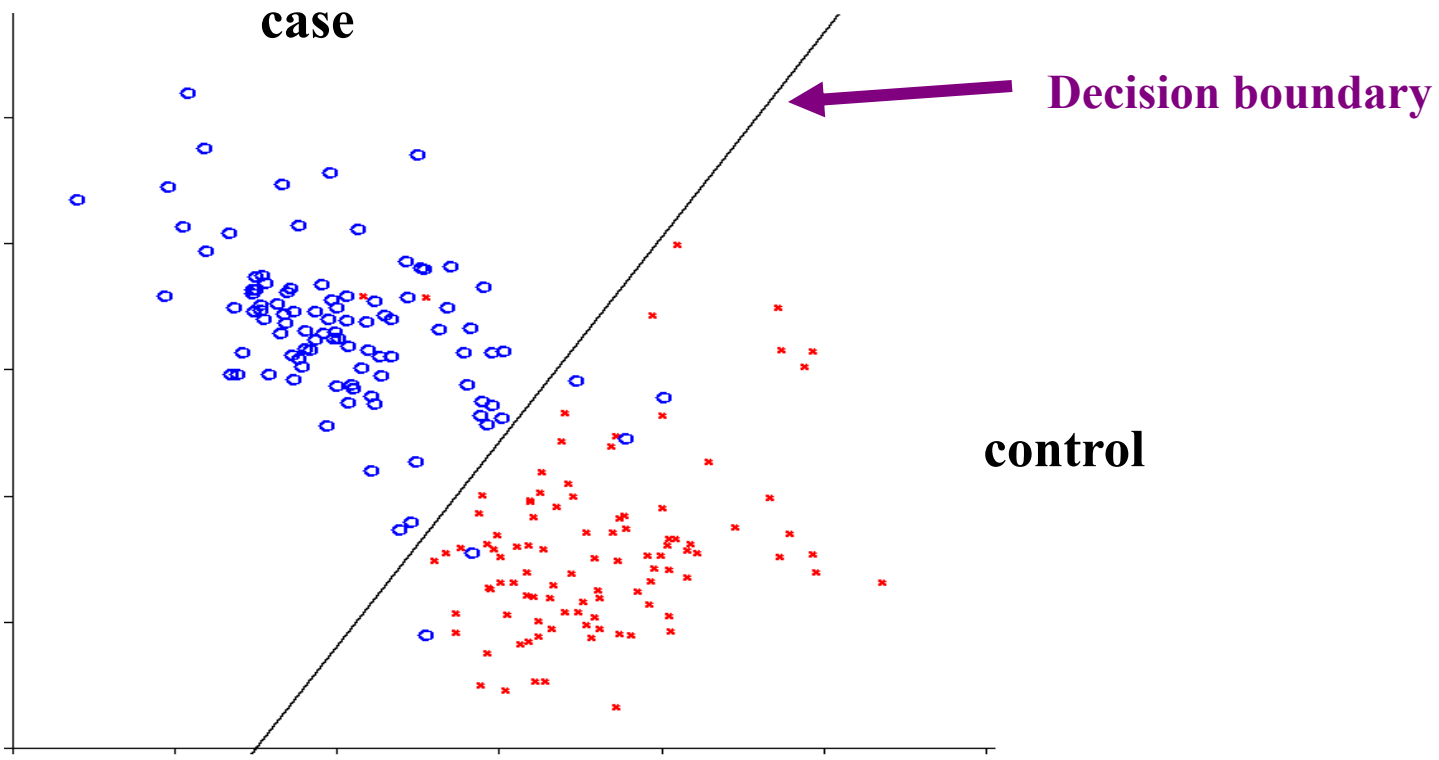
- Find a classification model that assigns with a high accuracy correct cancer/non-cancer label to SELDI-TOF profiles

Profile → **Classifier model** → **Class label (case or control)**

- **Classification models:**
  - Logistic regression
  - Linear and quadratic discriminant analysis
  - Support vector machines
  - Classification and regression trees (CART)
  + Bagging and boosting ensembles of classifiers
  - Random Forests

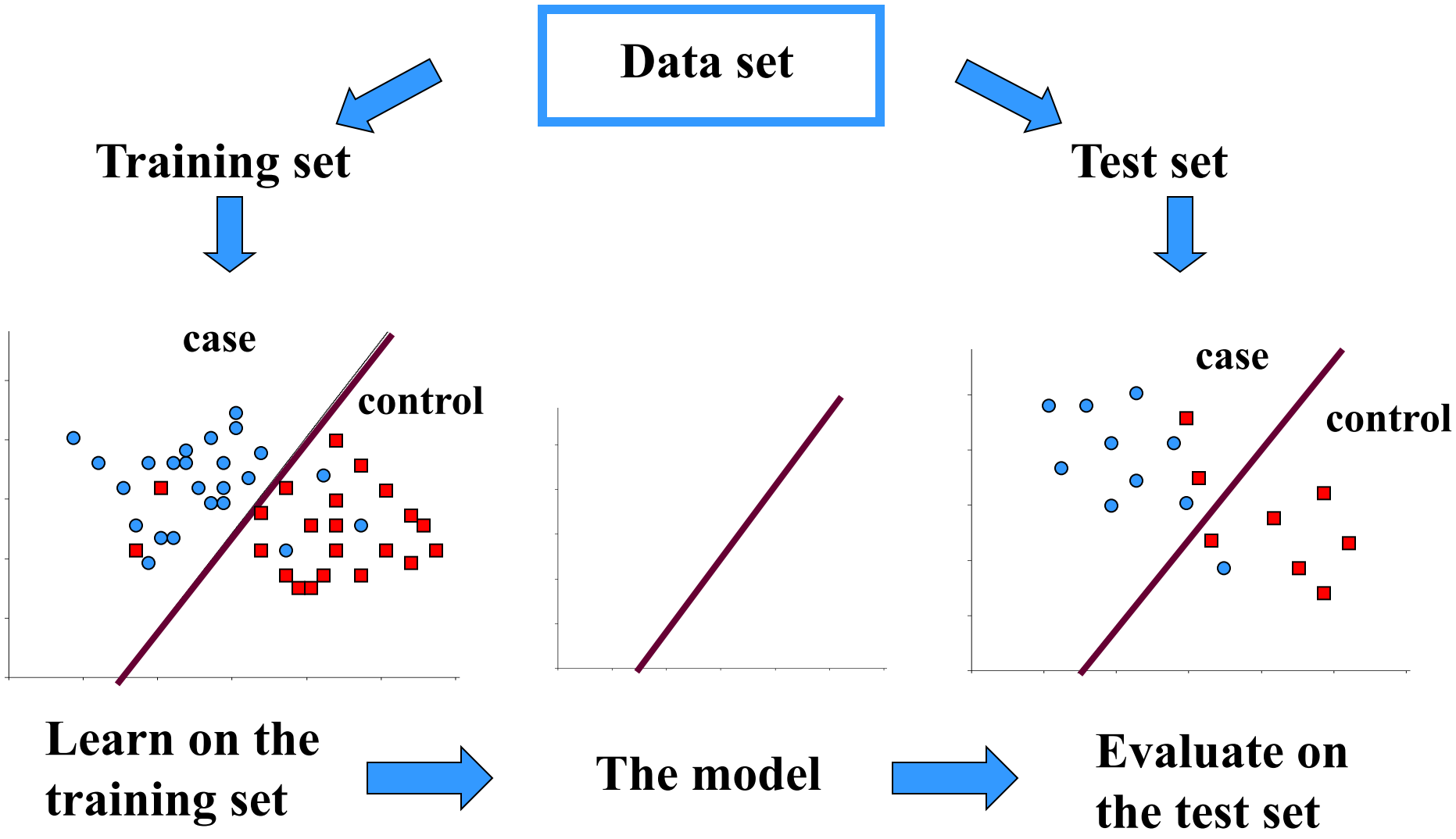
Learning a classification model

- **Example: a dataset with 2 features**



We learn **the decision rule** that achieves the best separation between case and control samples

Evaluation framework

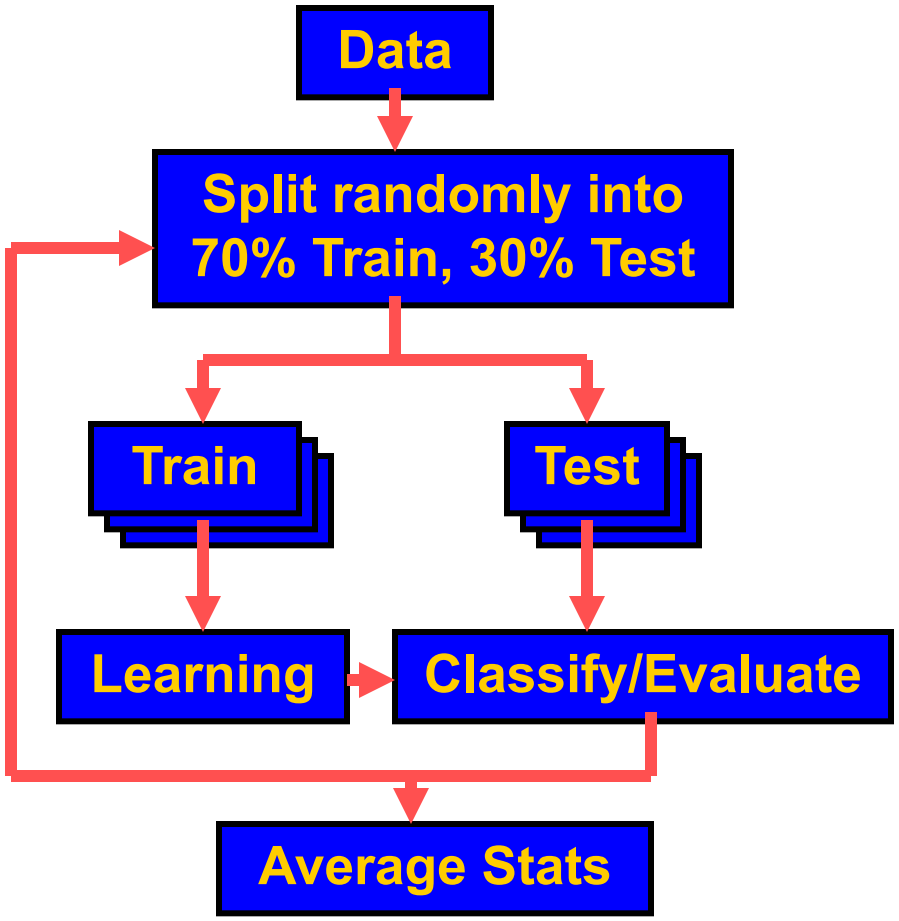


Evaluation

- To limit the effect of one lucky or unlucky train/test split average results over multiple train/test splits:
  - **bootstrap**
  - **random sub-sampling**
  - **k-fold cross-validation**

Example:

- random sub-sampling



Dimensionality problem

One important thing before we can learn a classifier

- **The dimension of the profile for proteomic data is very large** (~60,000 M/z measurements)

- **A large number of model parameters** must be to learned (estimated)

- **The total number of cases in the study is relatively small** (typically < 100)

- Leads to a **large variance of parameter estimates (**overfit)

Dimensionality reduction

Goal: identify a small subset of features that achieve a good discriminatory performance

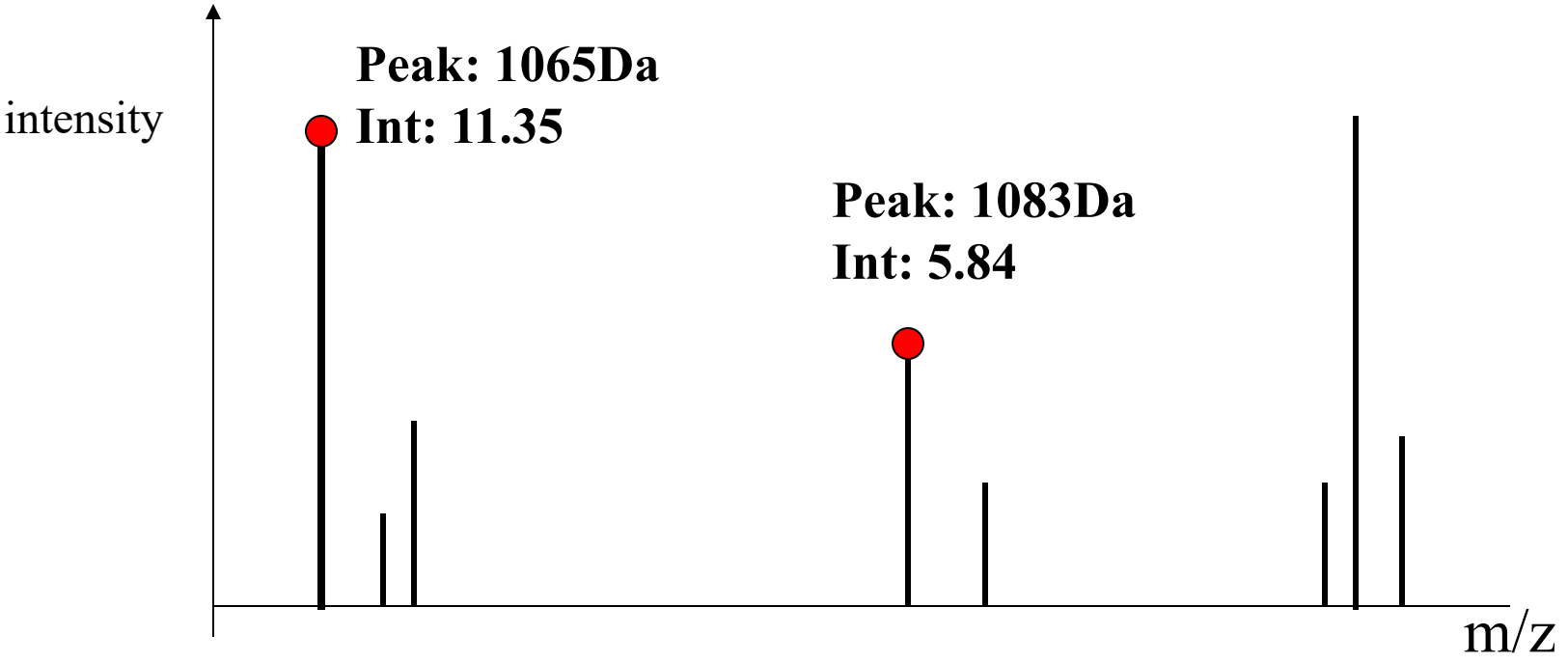
Two steps to reduce the dimensionality of data:

1. Restrict the analysis to the information in profile peaks

2. Apply ML dimensionality reduction methods and their combinations

Peak selection

- ~60,000 measurements reduced to hundreds of peaks
- Typically ~700 peaks in 1.5K-16 KDaltons region
- **Peak analysis:**
  - **Identify a peak position**
  - **Calculate peak intensity**



Dimensionality Reduction

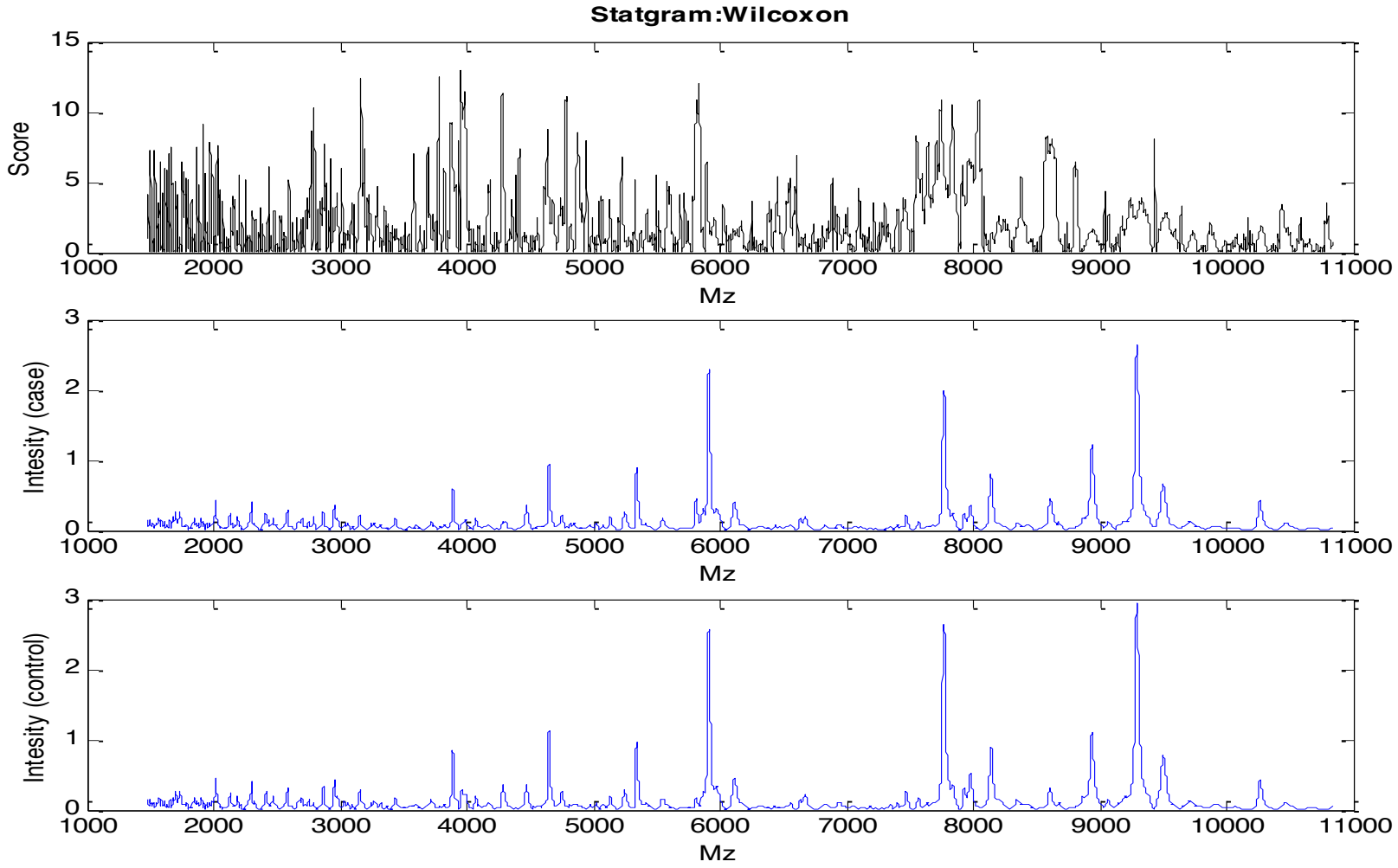
Goal: select a small subset of features that achieve a good discriminatory performance

- Feature selection
  - Filter methods
  - Wrapper methods
  - Embedded methods + heuristics
- Feature construction
  - Unsupervised methods: PCA, ICA
  - Discriminative projections:
    - Fisher linear discriminant
    - Linear discriminant analysis

Differentially expressed features

- **A score based on Wilcoxon rank-sum test**

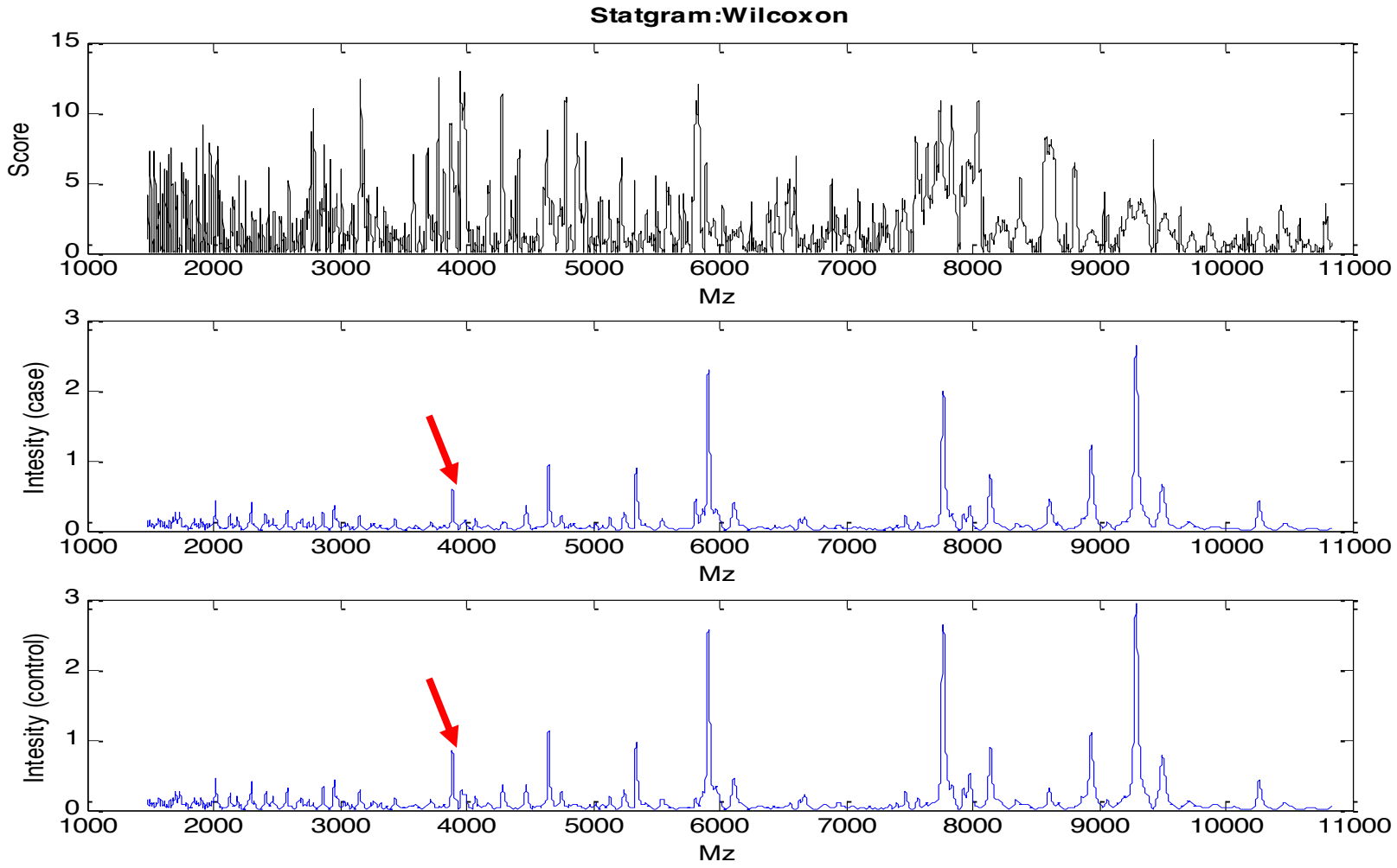
Pancreatic cancer data



Differentially expressed features

- **A score based on Wilcoxon rank-sum test**

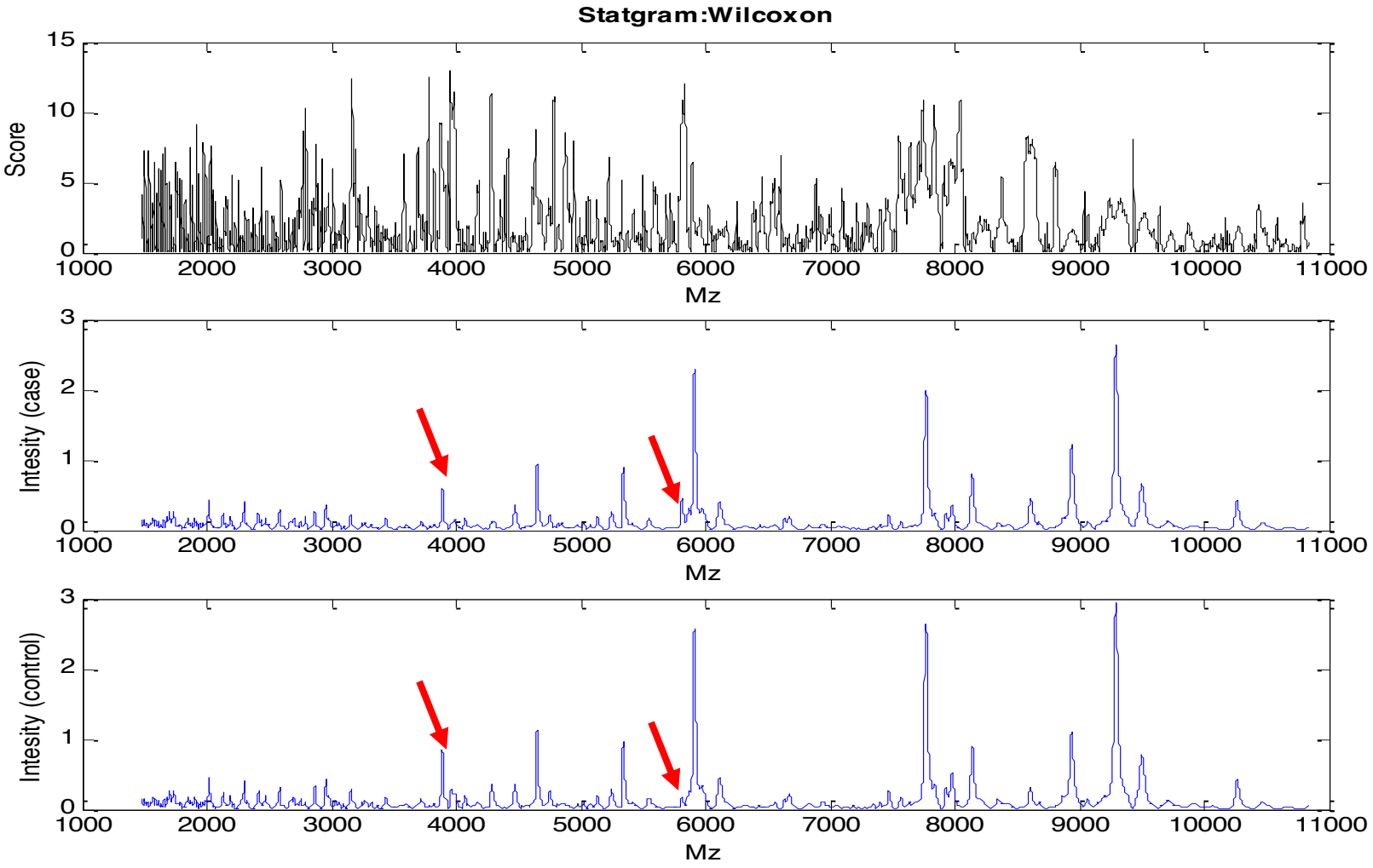
Pancreatic cancer data



Differentially expressed features

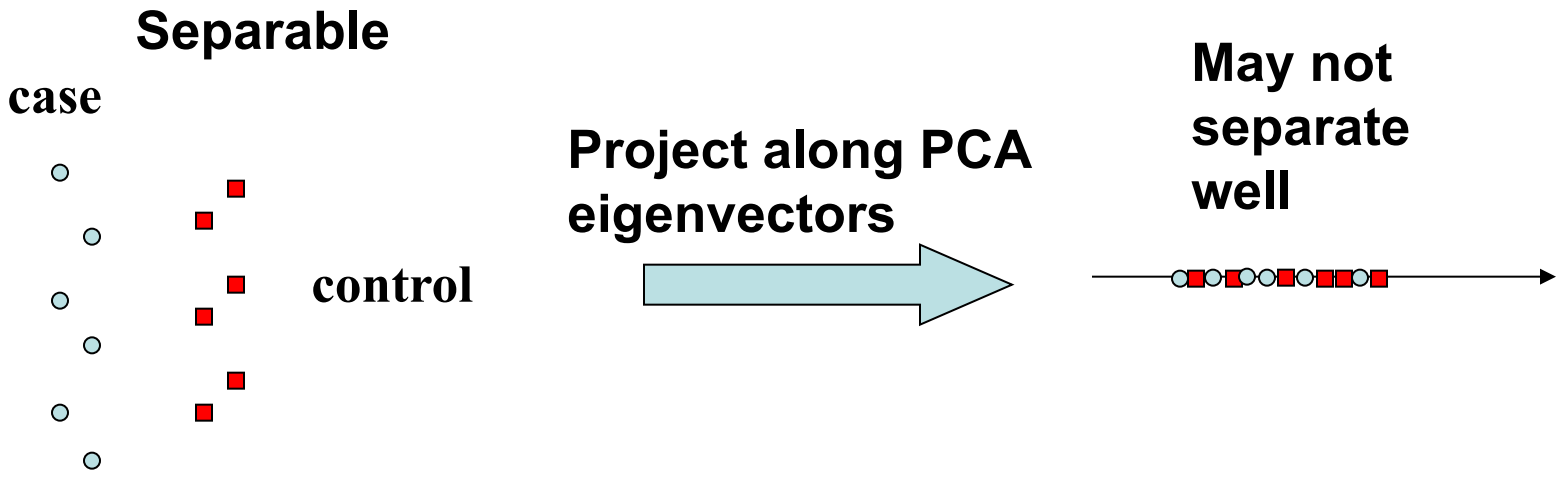
- **A score based on Wilcoxon rank-sum test**

Pancreatic cancer data



Unsupervised Projection Methods

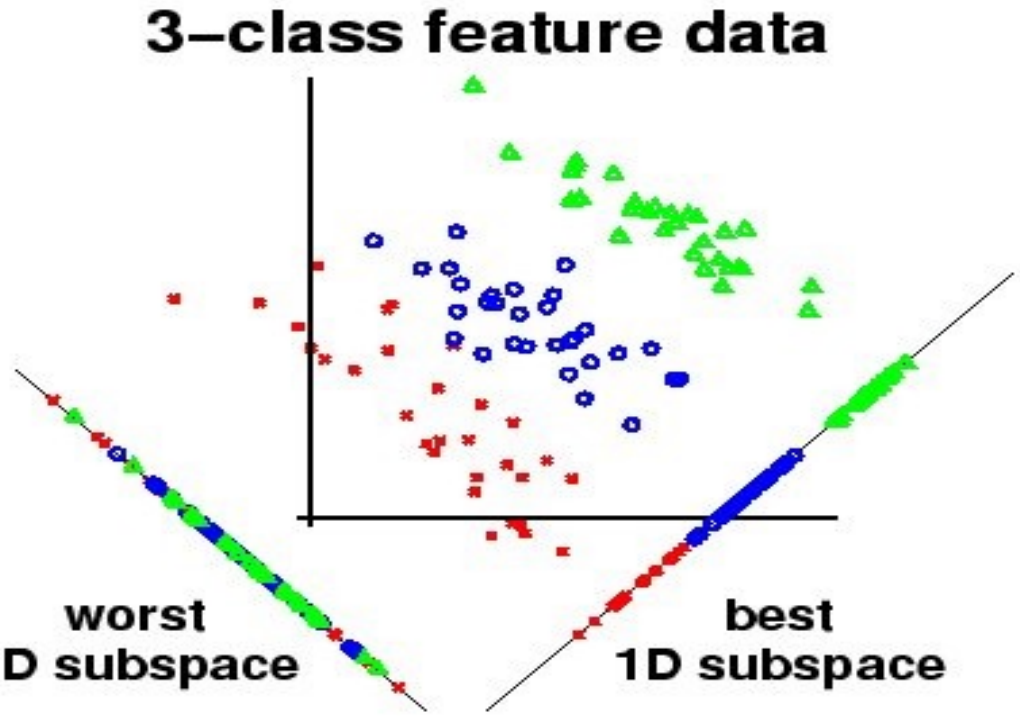
- **Example:** PCA (Principal Component Analysis)
- PCA computes the most meaningful basis to re-express the data with respect to data variance, but ignore the class labels



Discriminative Projection Methods

Examples:

- **Fisher linear discriminant**
  - Calculate single linear data projection that separates the classes the best

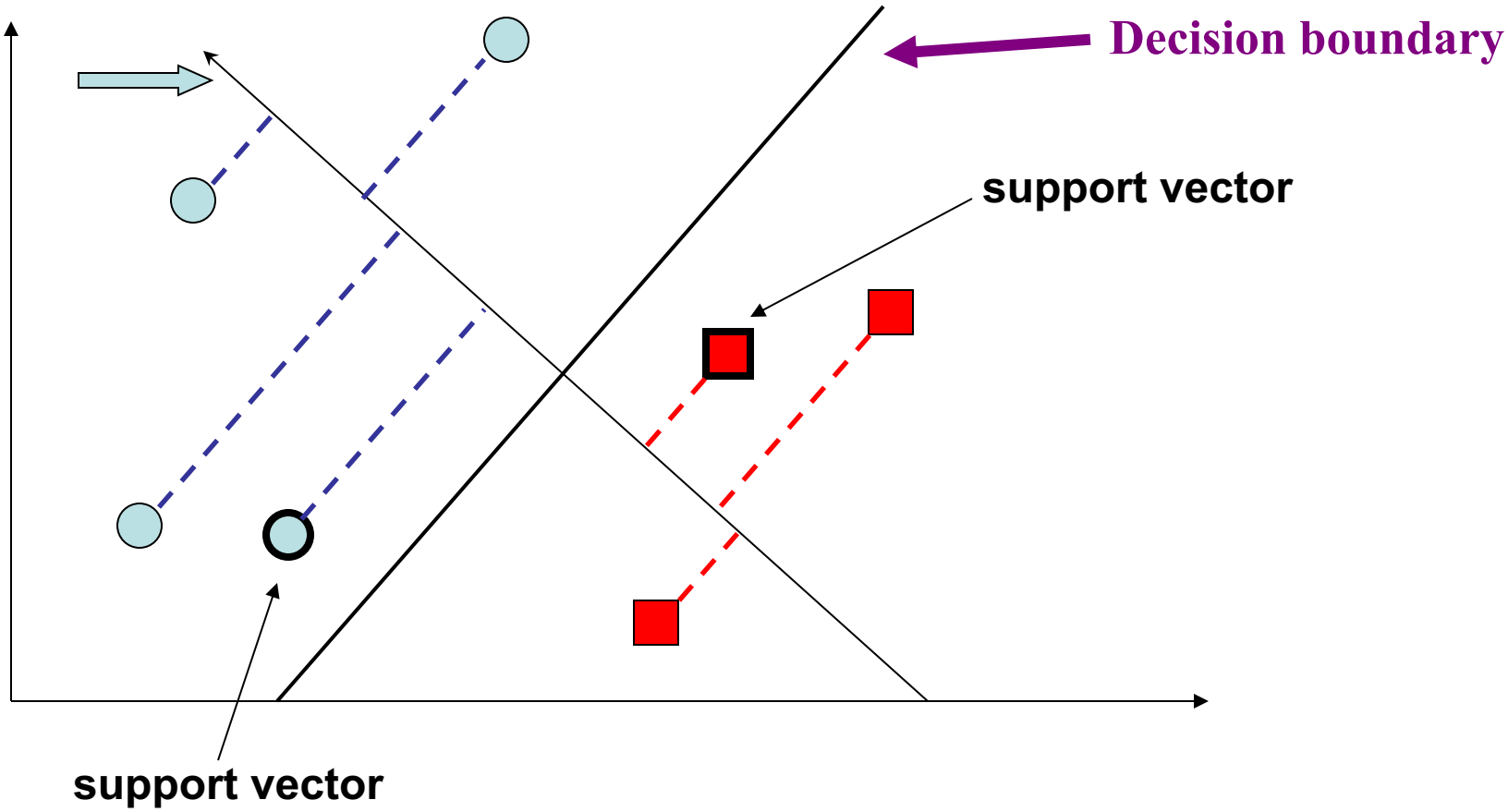


Multiple Discriminative Projections

- Identify a smaller set of the most promising original features (using Wilcoxon sum-rank test)
- Find a linear projection of these features that discriminates the data best (using SVM)
- Project data according to the linear projection, yielding a new aggregate feature
- Repeat the process by increasing the weight of data points not classified properly in the previous step. This yields a set of aggregate features
- Train and test a classifier on the new lower-dimensional feature space

Computing the Projection Direction

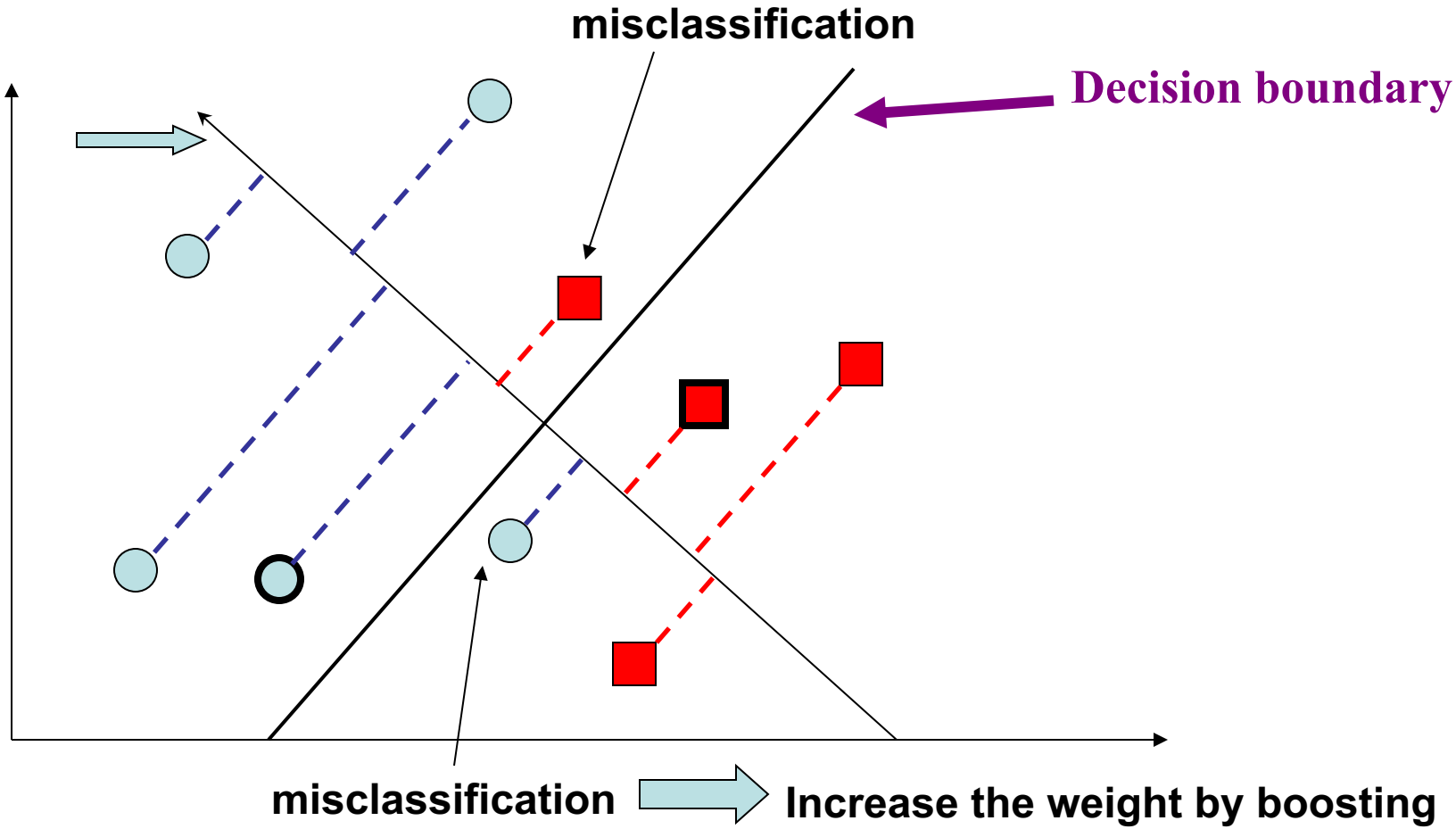
- Apply the support vector machine (SVM)



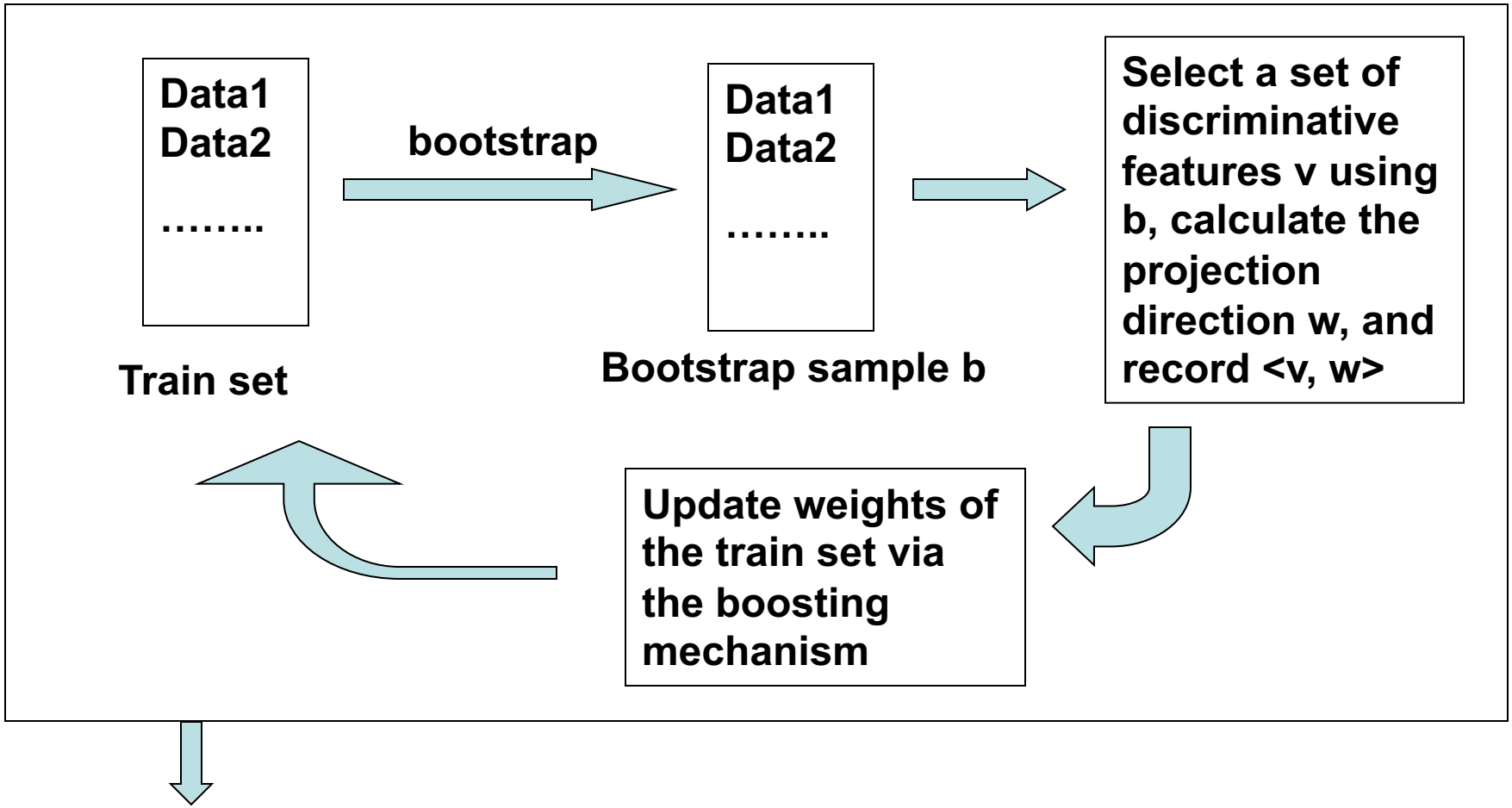
Dealing with Non-separable Cases

- Bootstrap from the data source
- Compute the projection direction on the bootstrap samples
- Update the weights of the original data set via a boosting mechanism – increasing the chance of seeing the hard instances in future bootstrap samples

Computing the Projection Direction



Flowchart



New feature set: <v1, w1>, <v2, w2>, … → Model training and testing

Data and Evaluation metrics

- UPCI pancreatic cancer data
- 602 peaks, 116 patients (57 cases, 59 controls)
- Confusion matrix

		Actual: Case	Actual: Control
Prediction	Case	TP	FP
	Control	FN	TN

Error:

$$E = \frac{FP + FN}{\#total}$$

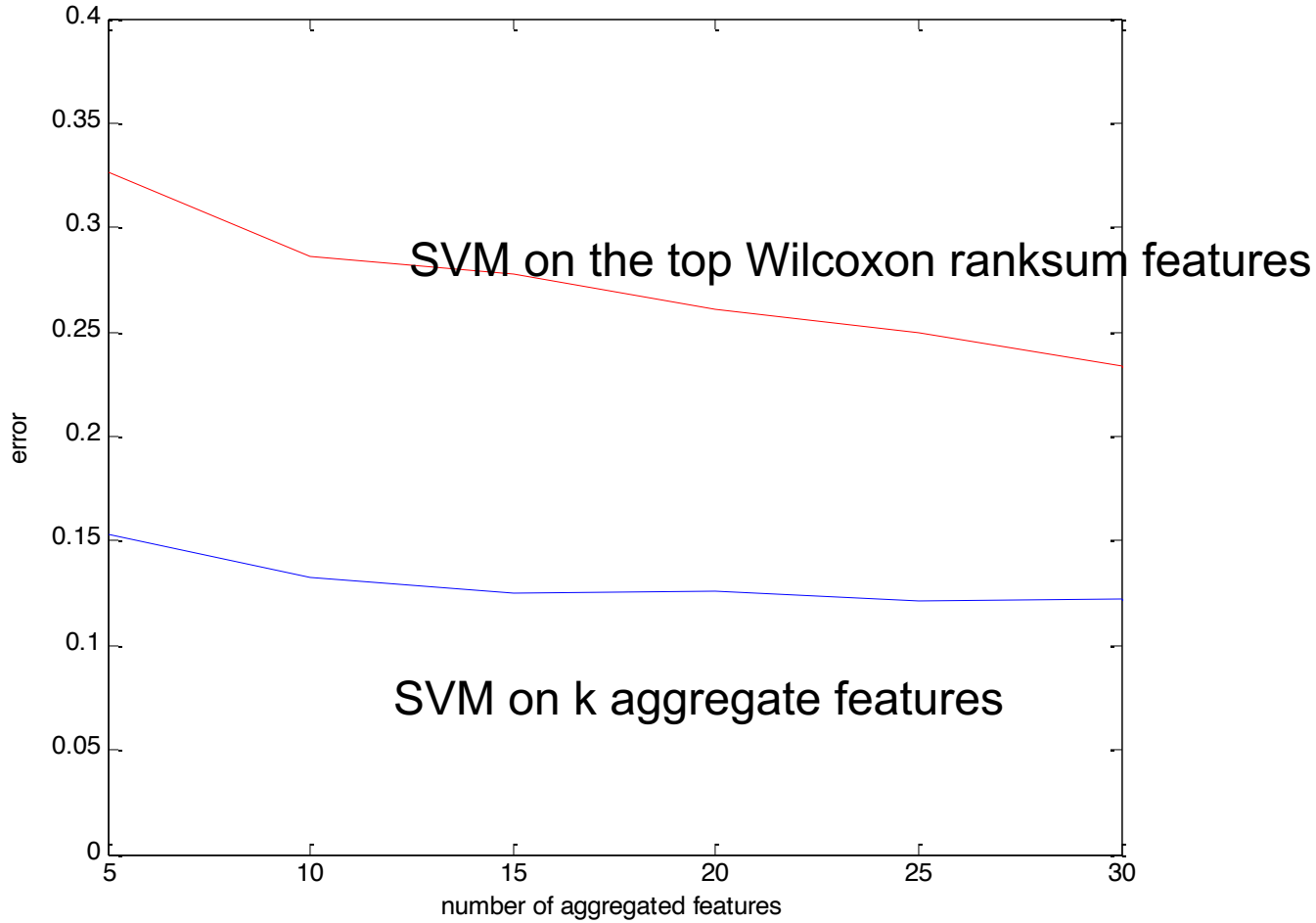
Sensitivity:

$$SN = \frac{TP}{TP + FN}$$

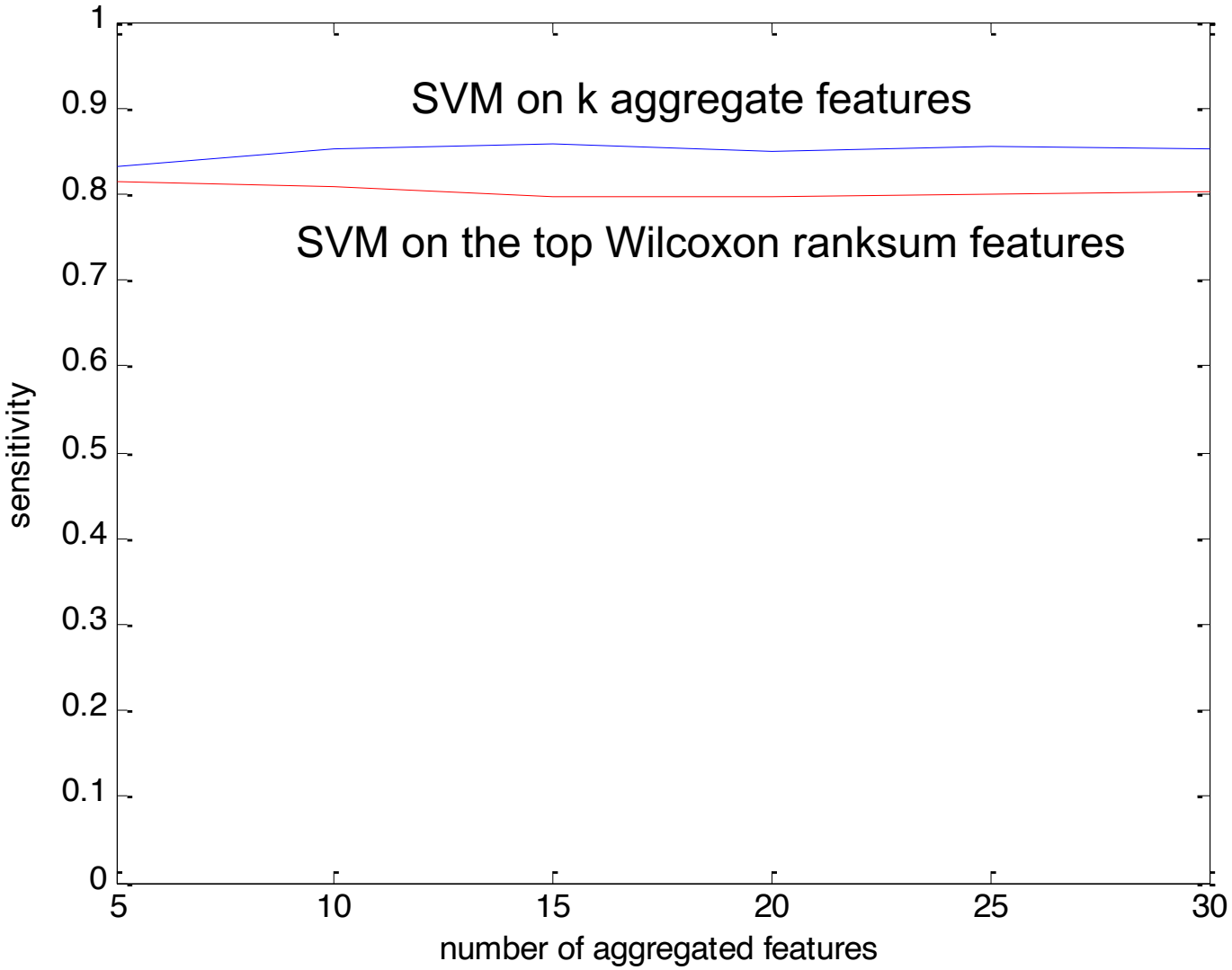
Specificity:

$$SP = \frac{TN}{TN + FP}$$

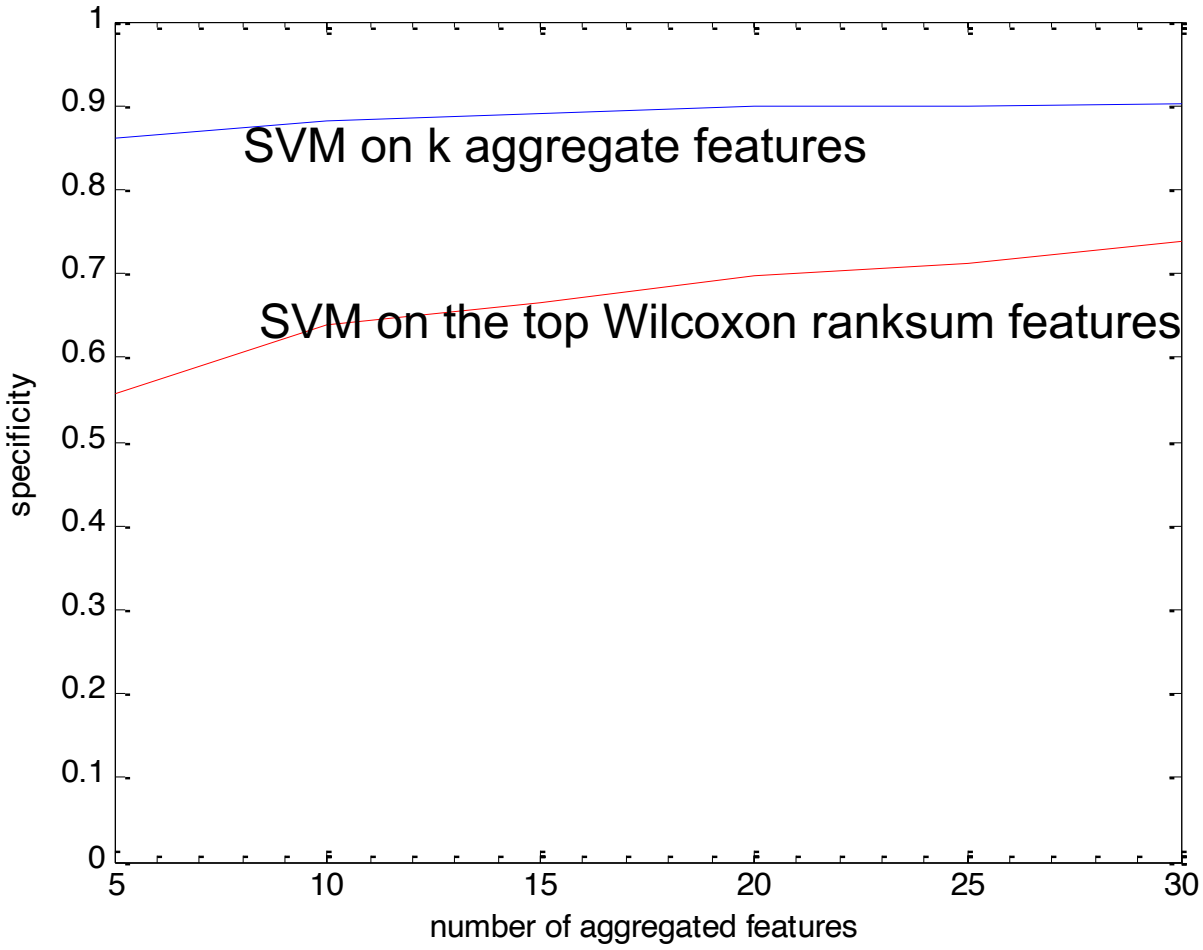
Experiment Result – Error Rate



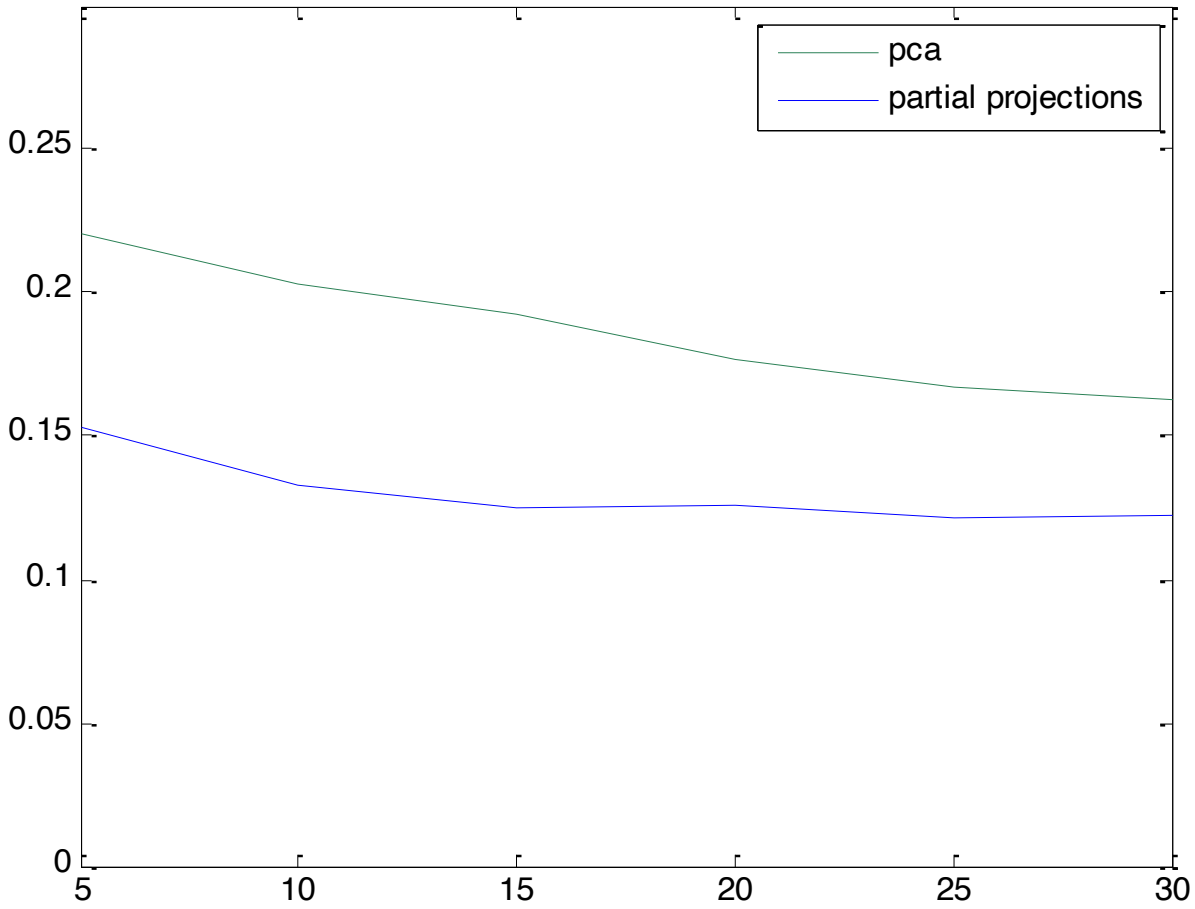
Experiment Result – Sensitivity



Experiment Result – Specificity



PCA features



SVML+PCA **Test error: 16.25%**

Conclusions

- SELDI-TOF MS data are potentially useful for disease detection
- Feature selection is an important step that can significantly improve the performance of the machine learning algorithms
- By performing multiple discriminative projections, learning algorithms like SVM achieve considerable improvement under our hybrid feature selection framework
- Method gives stable results for different scorings