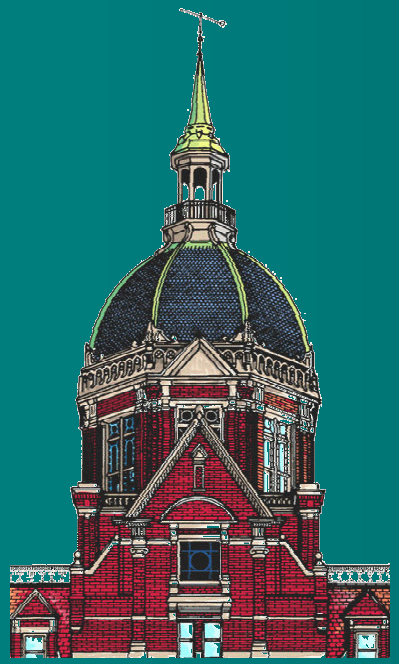
Issues in the Analysis of SELDI-TOF-MS Data



Zhen Zhang, PhD

Associate Professor and Associate Director

Center for Biomarker Discovery
Johns Hopkins University School of Medicine

Issues in the Analysis of SELDI-TOF-MS Data (and Other Protein Expression Profile Data) from Clinical Samples for Biomarker Discovery

The Usage and Abusage of Bioinformatics Tools

Protein Expression Profile

- Extreme dynamic range in expression levels of different proteins (10^{10});
- Dynamic changes of the same proteins over time and varying conditions;
- Biological variability among individuals within the same populations;
- Sample preprocessing also introduces additional analytical variability.

Expression Data from Clinical Samples

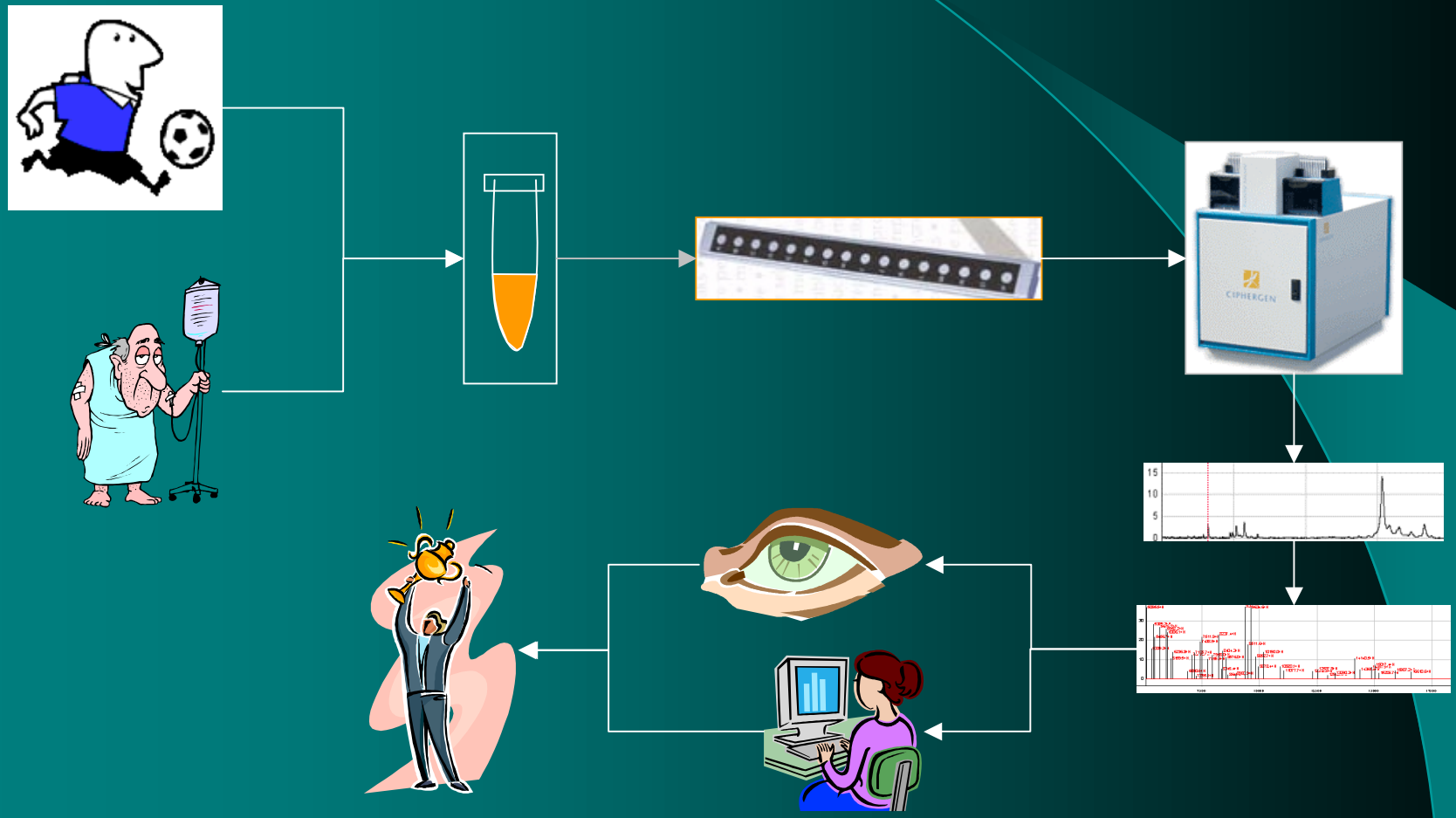
In addition to $p >> n$, we also have:

- Much more significant within-class variability due to biological variability or sub-phenotypes.
- Possible systematic biases due to pre-analytical variables.
- Difference in sample populations.
- Possible mislabeling of clinical samples.

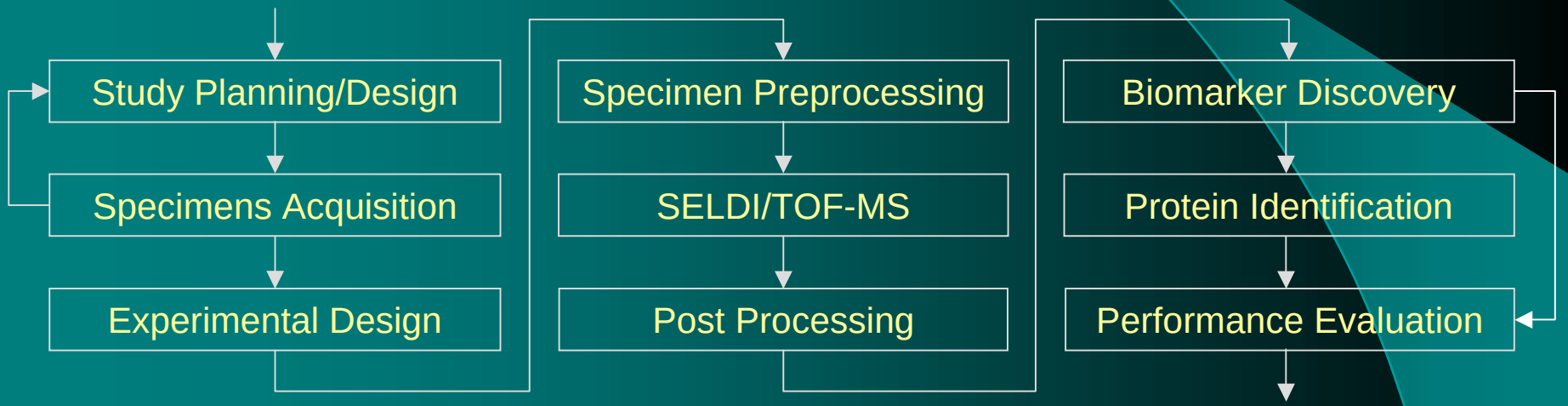
Analysis for Biomarker Discovery

- Most are case-controlled studies;
- Most use supervised approaches;
- Sensitive to systematic biases in data;
- Thousands of candidates does not mean any of them have to be good.

An Outsider's View of Biomarker Discovery Using SELDI MS-TOF



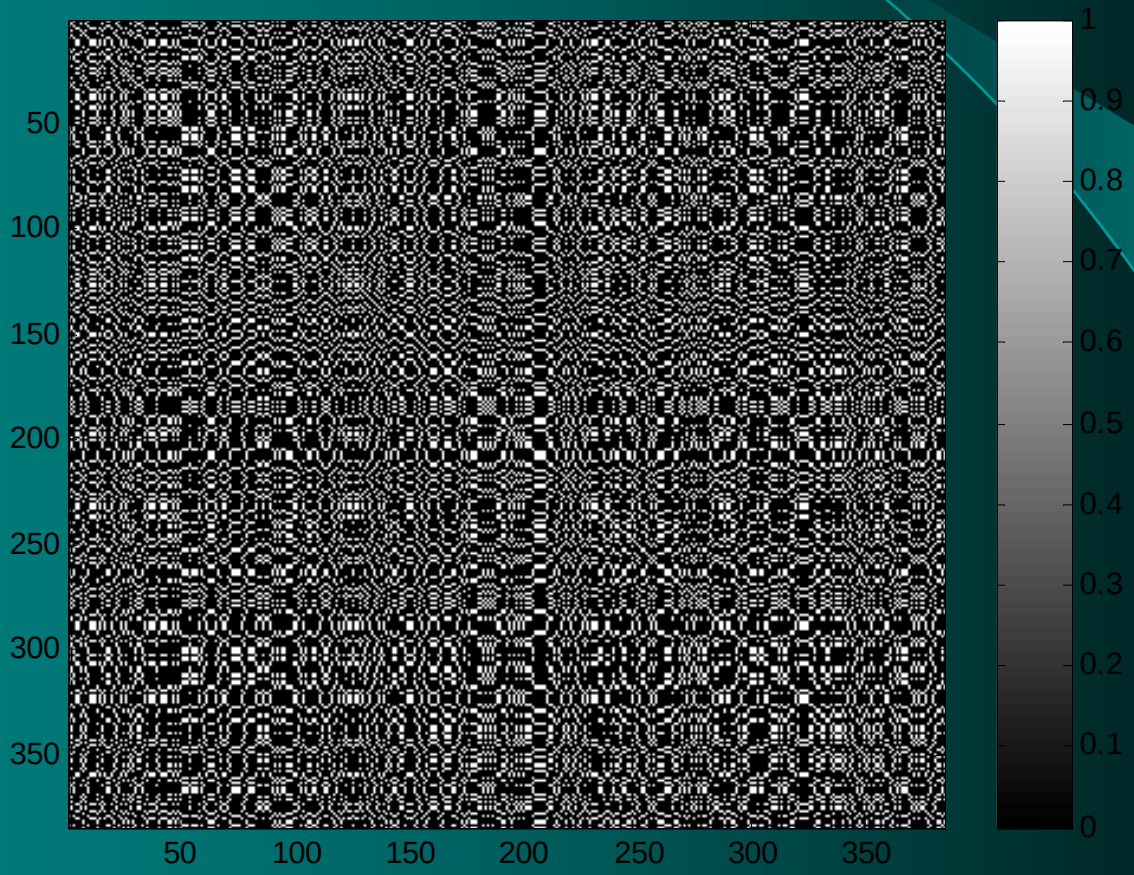
The Insider's (??) View of Biomarker Discovery Using SELDI MS-TOF



Issues w.r.t. Bioinformatics

Experimental Design and Execution

Experimental Design



Issues w.r.t. Bioinformatics

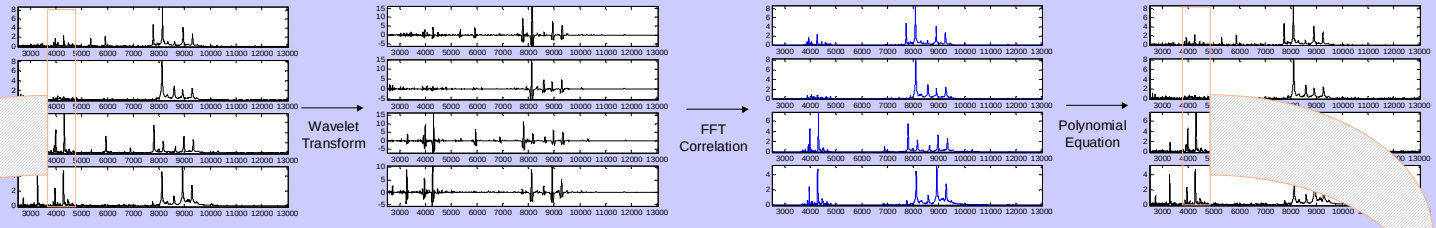
Spectra Processing

Spectrum Alignment

C. Nicole White, Z. Zhang

The Software Tool: Intra-Calibration

The following steps graphically demonstrate the method introduced in the text of this poster.

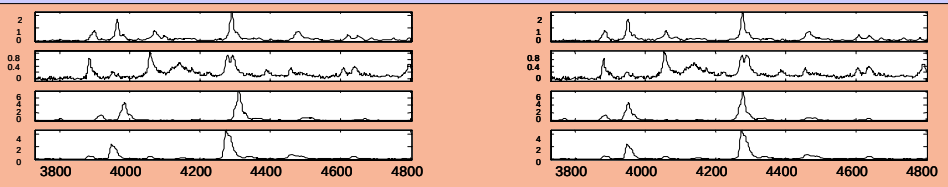


Original Spectra. Similar contextual information is retained among spectra

Spectra after wavelet decomposition

Alignment of several regions selected by algorithm

Intra-calibrated spectra, final result



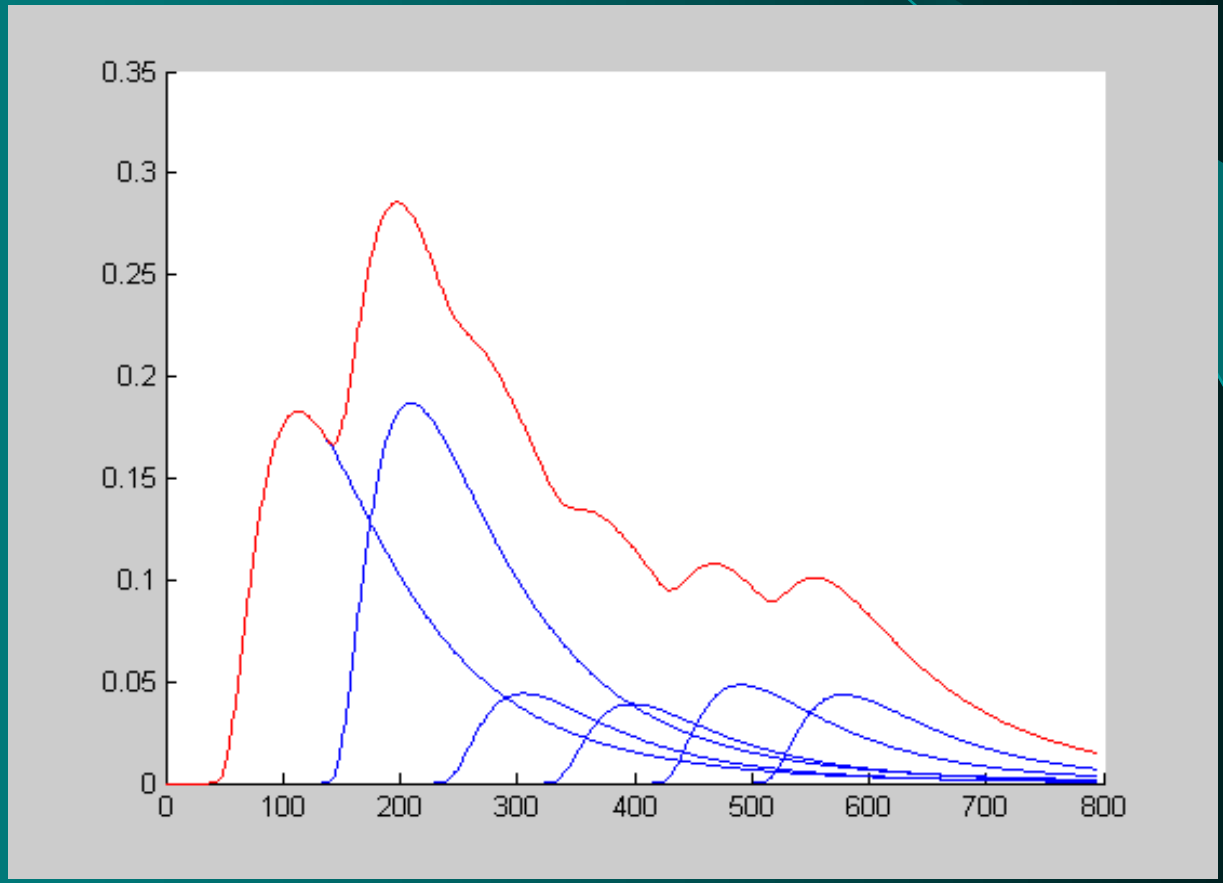
Original Spectra

Intra-Calibrated Spectra

Identical regions of the original spectra and the aligned spectra are blown up to demonstrate the algorithm's improvement.

Peak Decomposition

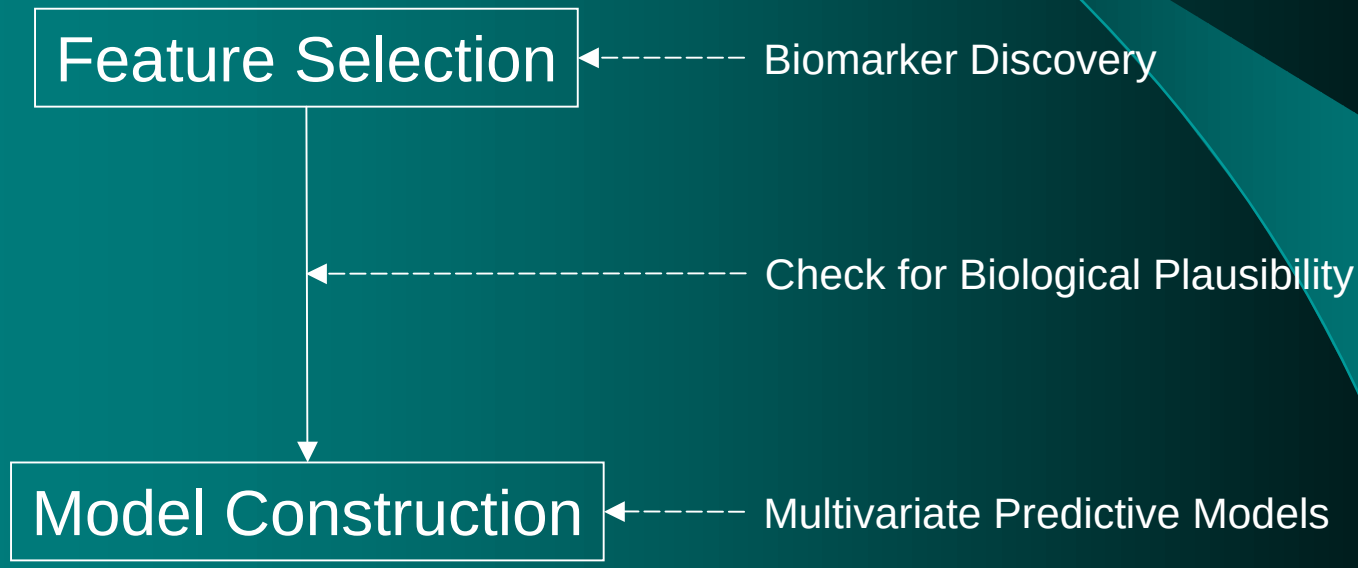
H. Zhang, C. Nicole White, Z. Zhang



Issues w.r.t. Bioinformatics

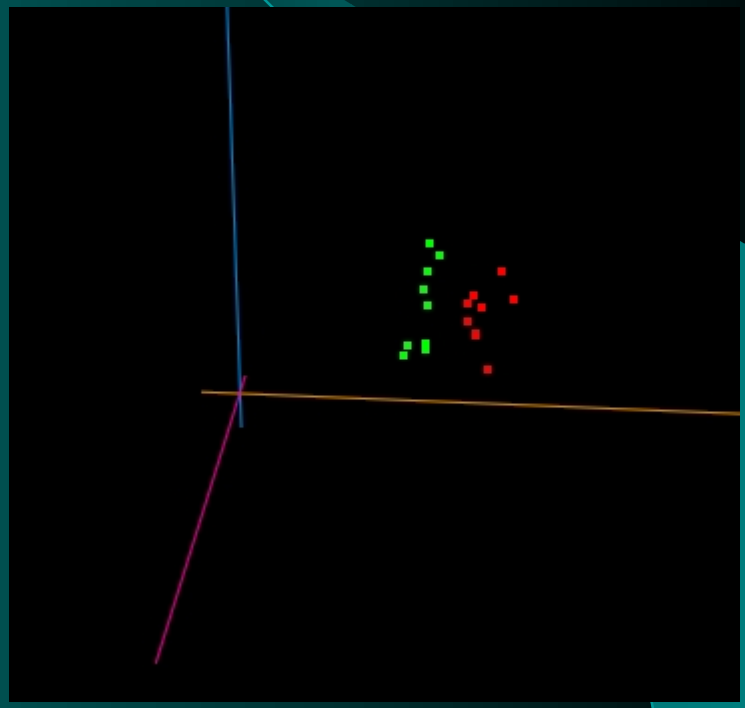
Variable (biomarker) Selection

A One-Step or Two-Step Process ?

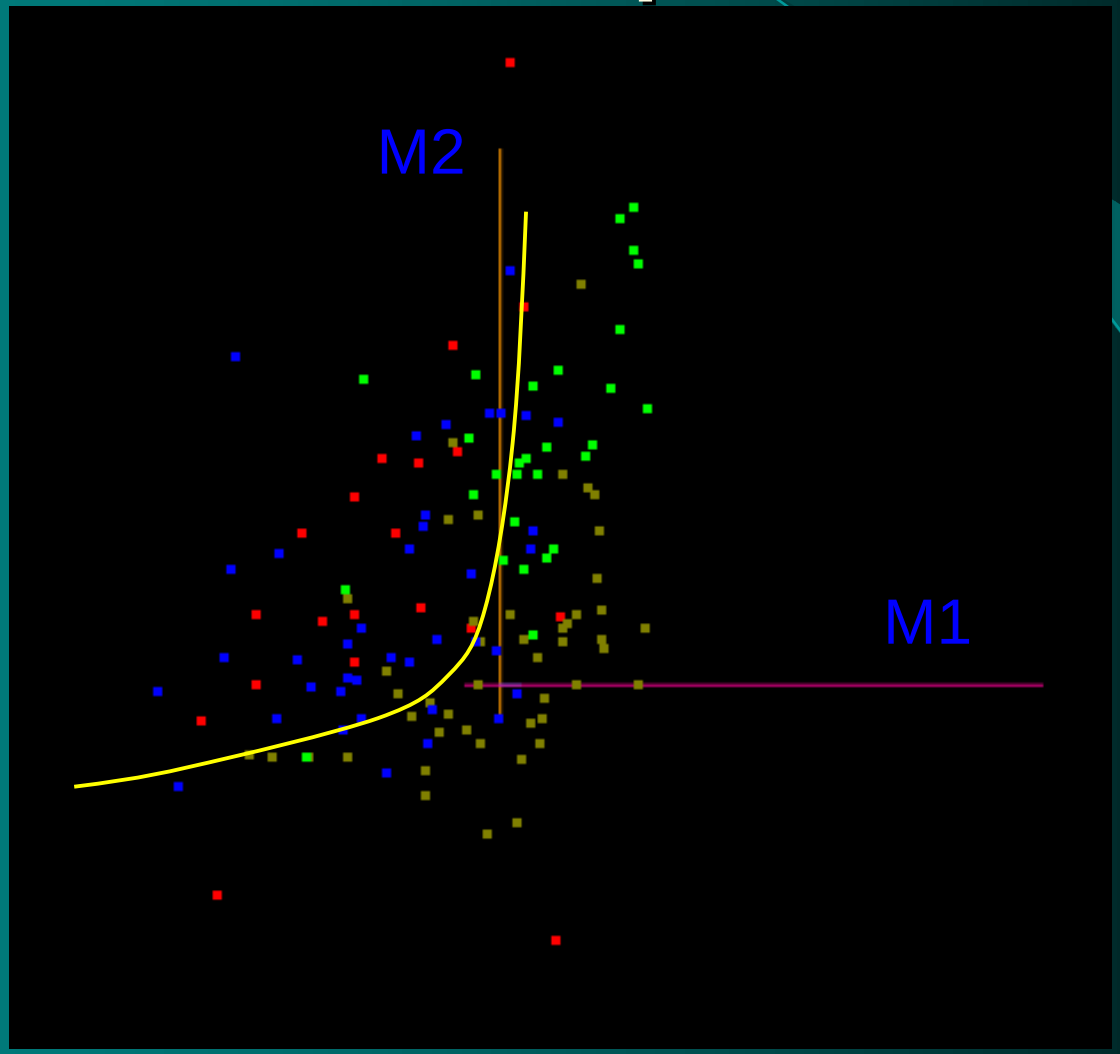


Signature of Diseases?

- Nonlinear combination of variables from a large number of peaks (10^2 – 10^4) could result in an astronomically large number of “signatures.” By chance, some of them could be uniquely linked to groups of samples of small sizes.
- 20 simulated “samples” each with 150 “peaks”, all data generated with random numbers.
- It’s very easy to find a subset of peaks that in combination perfectly separate two arbitrarily labeled groups.



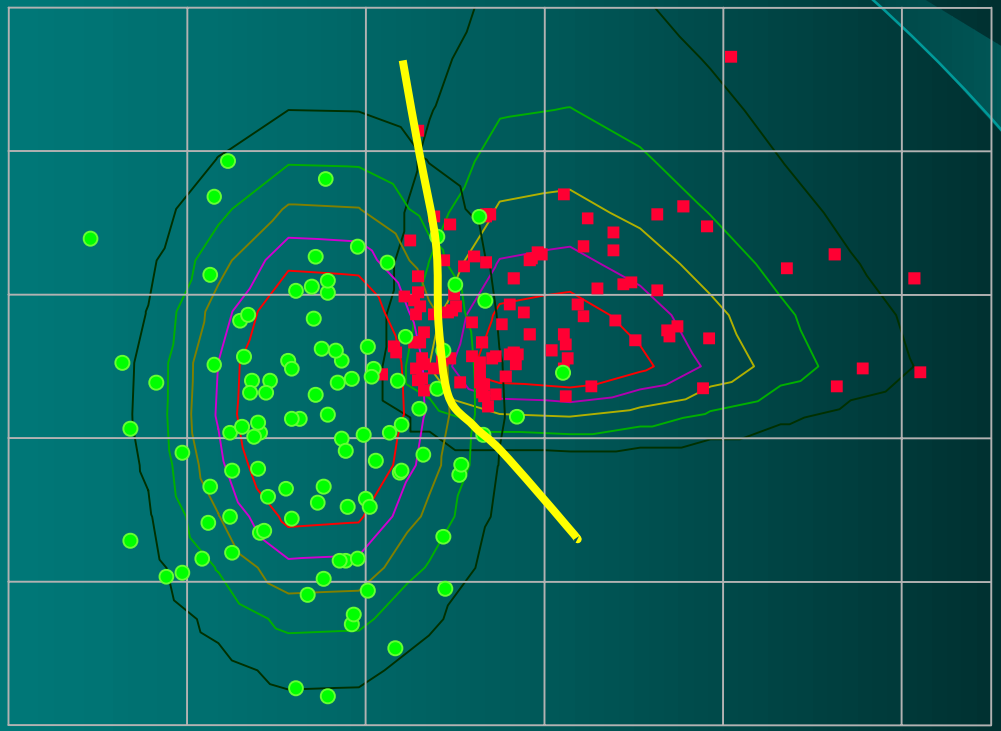
Check Carefully: A Real Data Example



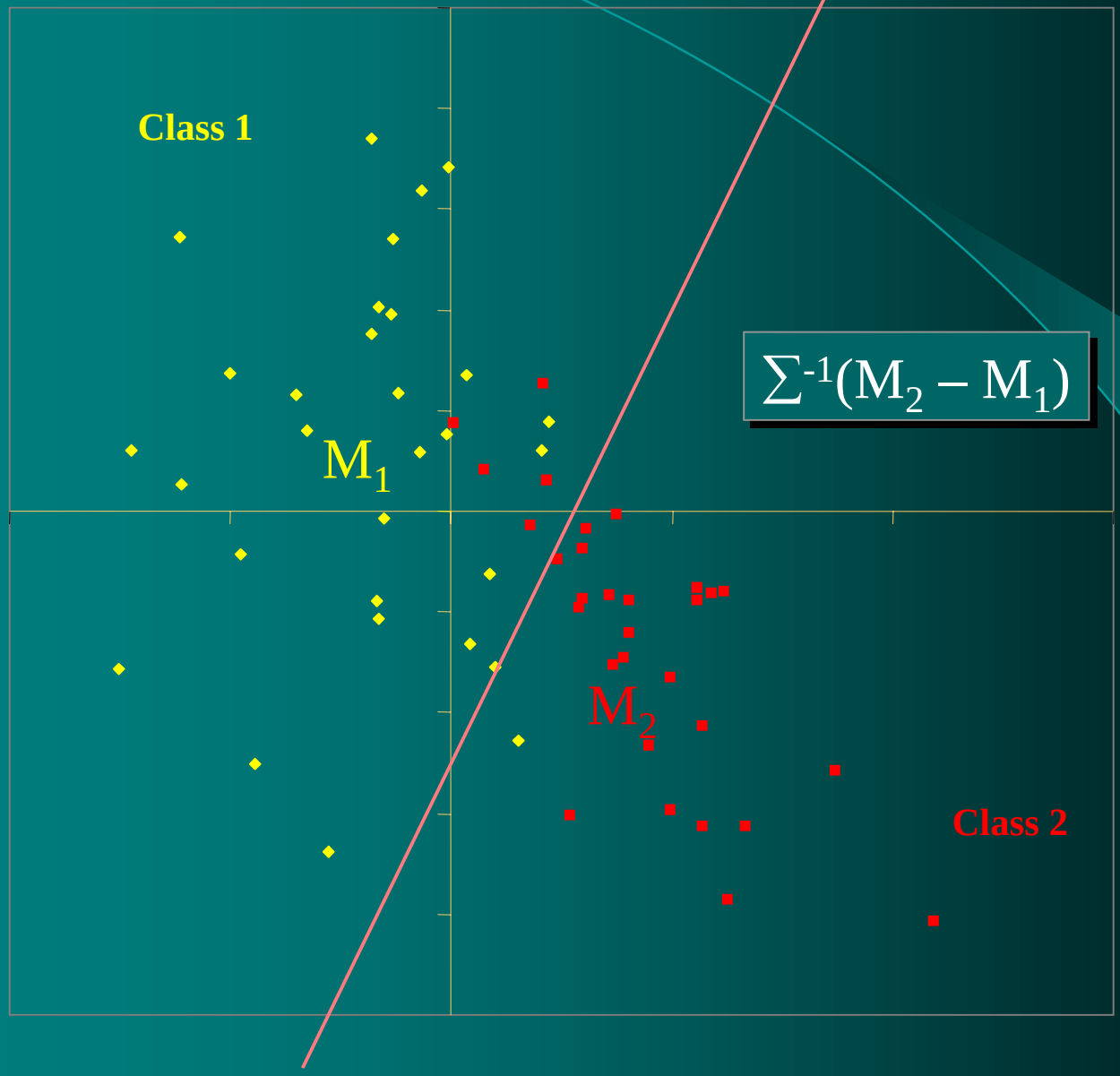
A Basic Supervised Approach for Biomarker Discovery

- Derive a classifier that best separate the groups of samples;
- Determine the contributions of individual variables;
- Select a subset of most informative variables;
- Evaluate the performance of the selected variables.

Classifiers Based on Estimated Conditional Distributions



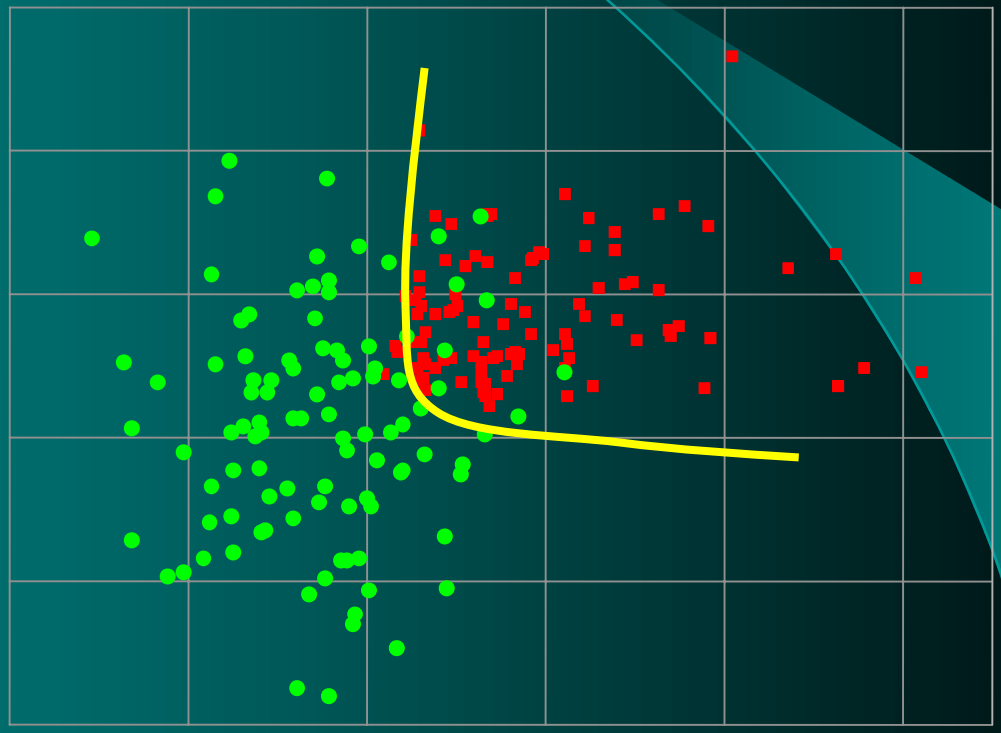
Fisher's LDA



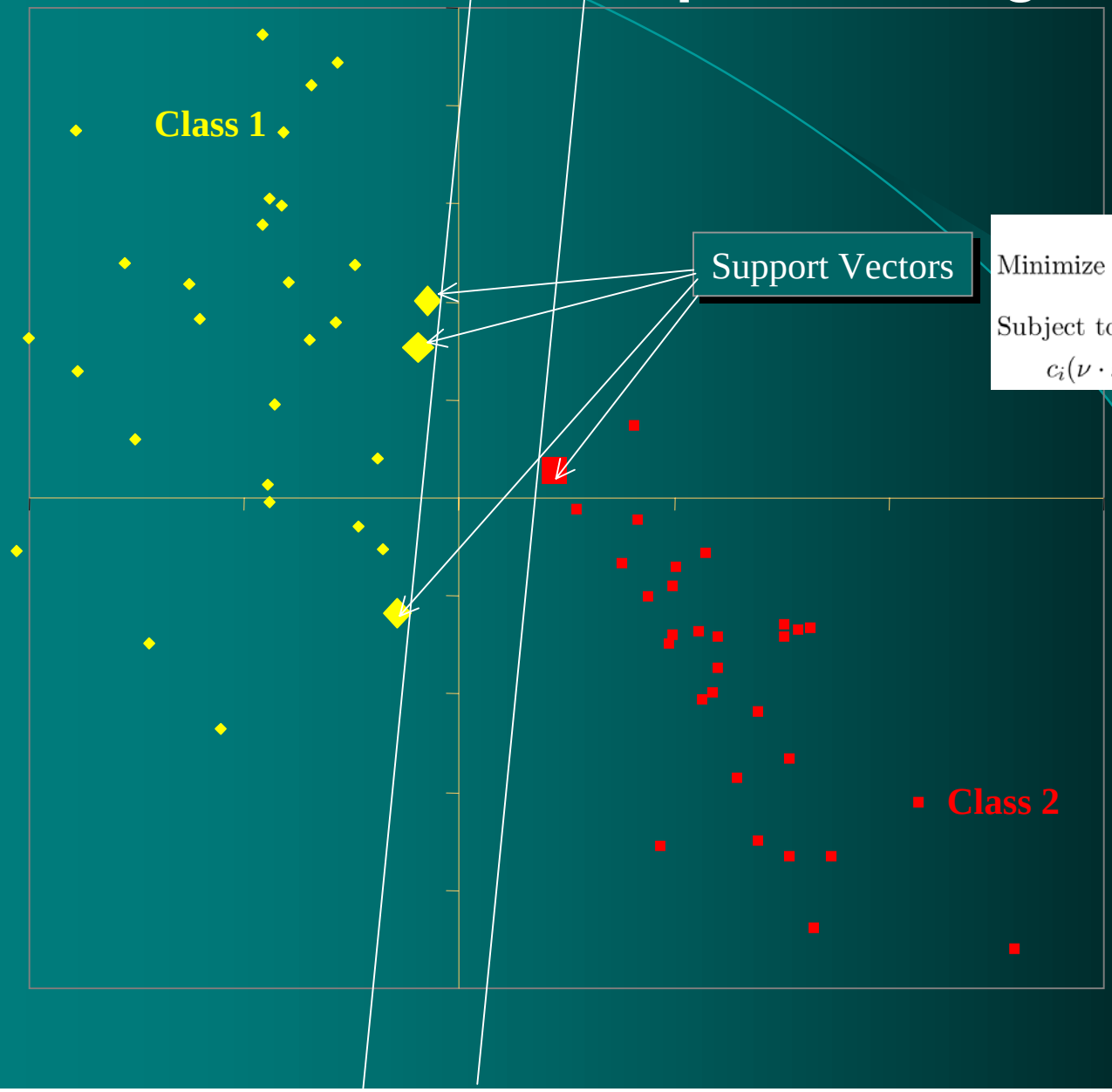
Based on data distribution information.

Classifiers by Empirical Risk Minimization

$$
\begin{aligned}
R_{\mathrm{emp}}(\alpha) &= \frac{1}{l}\sum_{i=1}^{l} Q(z_i, \alpha) \\
&= \frac{1}{l}\sum_{i=1}^{l} (y_i - f(x_i, \alpha))^2
\end{aligned}
$$



Optimal Margin Classifier



Empirical Risk Minimization.

$$\text{Minimize } \frac{1}{2}\nu \cdot \nu + C\sum_{i=1}^{m}\xi_i$$

Subject to

$$c_i(\nu \cdot x_i + b) \geq 1 - \xi_i,\ i = 1, 2, \ldots, m$$

The Unified Maximum Separability Analysis (UMSA) Algorithm

Minimize $\frac{1}{2}\nu \cdot \nu + \sum_{i=1}^{m} p_i \xi_i$

Individualized Softness

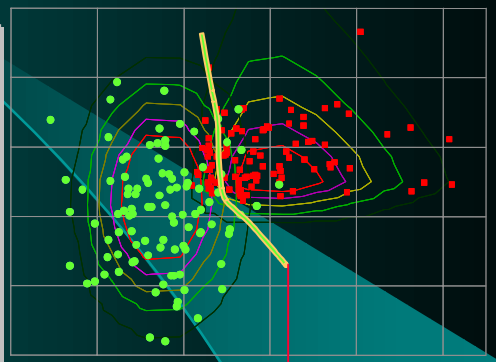
Subject to

$$c_i(\nu \cdot x_i + b) \geq 1 - \xi_i, \ i = 1, 2, \ldots, m,$$

$$p_i = K\phi(\delta_i),$$

A typical choice for the function $\phi(\cdot)$ would be

$$\phi(x) = e^{-x^2/\sigma^2}$$



UMSA Component Analysis

- Find a projection vector *d* along which two classes of data are optimally separated for a given set of UMSA parameters.
- Project the data onto a subspace perpendicular to *d*.
- Iteratively, apply UMSA to compute a new projection vector within this subspace, until a desired number of components have been reached.

Procedure: UMSA component analysis for a two-class dataset with *m* variables and *n* samples

inputs:

UMSA parameters C and σ;

number of components $q \le \min(m, n)$;

data $X = (x_1, x_2, \ldots, x_n)$; and

class labels $L = (l_1, l_2, \ldots, l_n)$, $l_i \in \{-1,+1\}$.

initialization:

component set $D \leftarrow \{\}$;

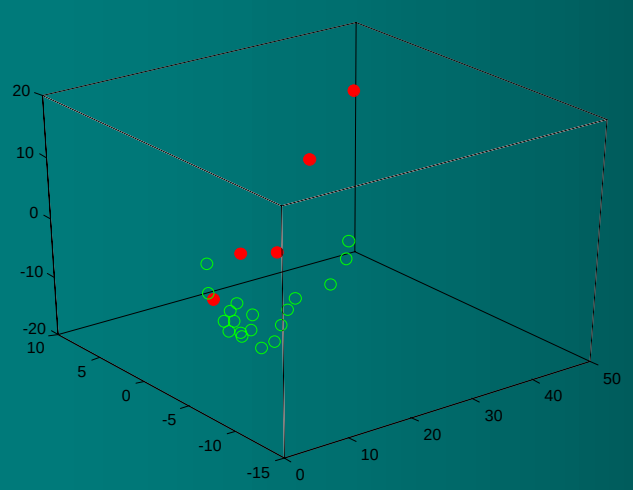
$k \leftarrow 1$.

while $k \le q$

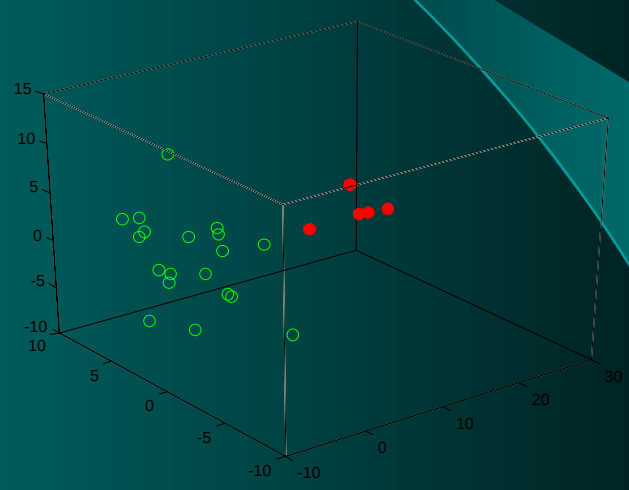
1. applying UMSA(σ , C) on $X = (x_1, x_2, \ldots, x_n)$ and L;
2. $d_k \leftarrow v / \|v\|$; $D \leftarrow D \cup \{ d_k \}$;
3. $x_i \leftarrow x_i - (x_i^T d_k) d_k$, $i = 1, 2, \ldots, n$;
4. $k \leftarrow k + 1$.

return D.

UMSA Component Analysis vs. PCA/SVD

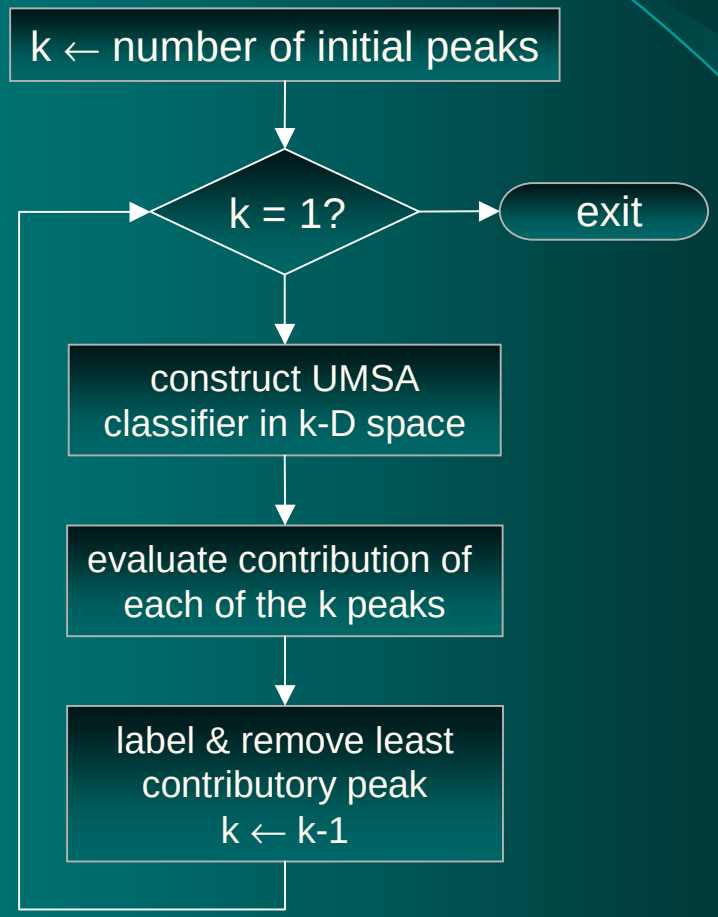


SVD



UMSA CA

Backward Stepwise Variable Ranking/Selection



Procedure: Stepwise backward UMSA variable selection for a two-class dataset with *m* variables and *n* samples

inputs:

UMSA parameters C and σ;

data $e = \{e_{ji} \mid j = 1,2,\ldots,m; i = 1,2,\ldots,n\}$; and

class labels $L = (l_1, l_2, \ldots, l_n)$, $l_i \in \{-1,+1\}$.

initialization:

$G_k \leftarrow G_m = \{g_j = (e_{j1}, e_{j2}, \ldots, e_{jn})^T, j = 1,2,\ldots,m\}$;

score vector $w = (w^1, w^2, \ldots, w^m)^T \leftarrow (0,0,\ldots,0)^T$.

while $|G_k| > 1$

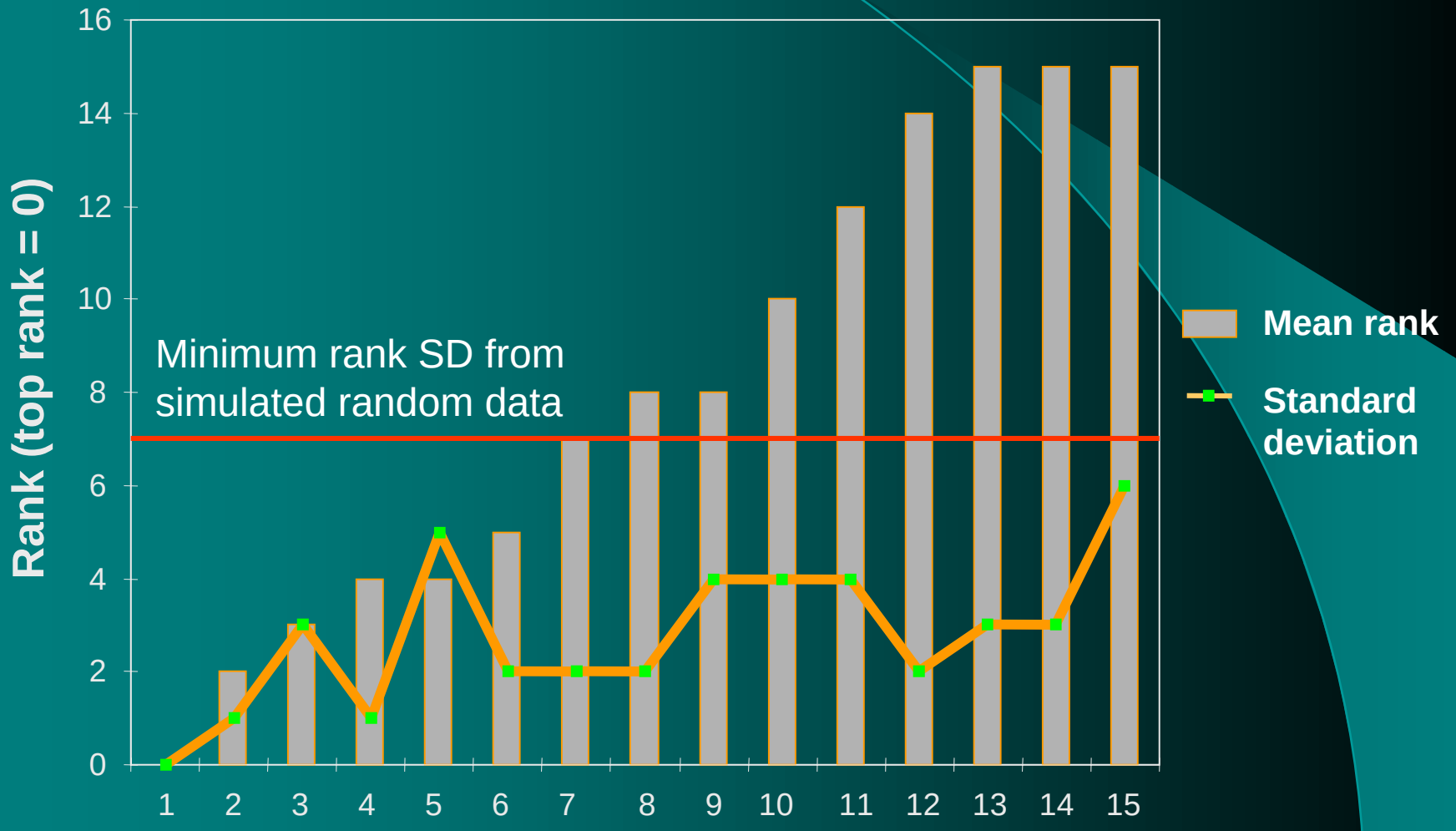
1. forming $X = (x_1, x_2, \ldots, x_n) \leftarrow (g_1, g_2, \ldots, g_k)^T$.
2. applying UMSA(σ, C) on X and L;

   $s_k \leftarrow 2/\|v\|$ and $d_k \leftarrow v/\|v\|$.
3. for all $g_j \in G_k$, if $s_k |d_k^j| > w^j$, $w^j \leftarrow s_k |d_k^j|$.
4. $G_{k-1} \leftarrow G_k - \{g_r\}$, where r is determined from $w^r = \min_{g_j \in G_k} \{w^j\}$.

return w.

Note: $w^{k-1} \leq w^k$, for all k

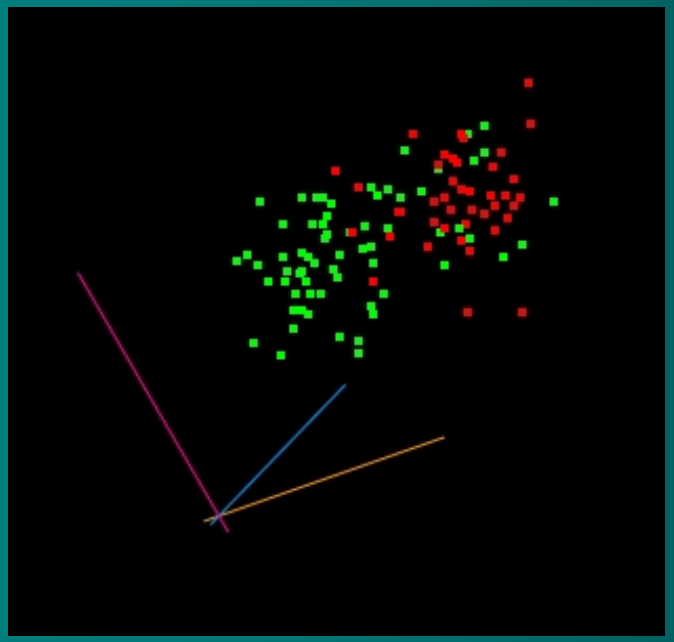
Alleviating Impact of Biological Variability Using Statistical Re-sampling



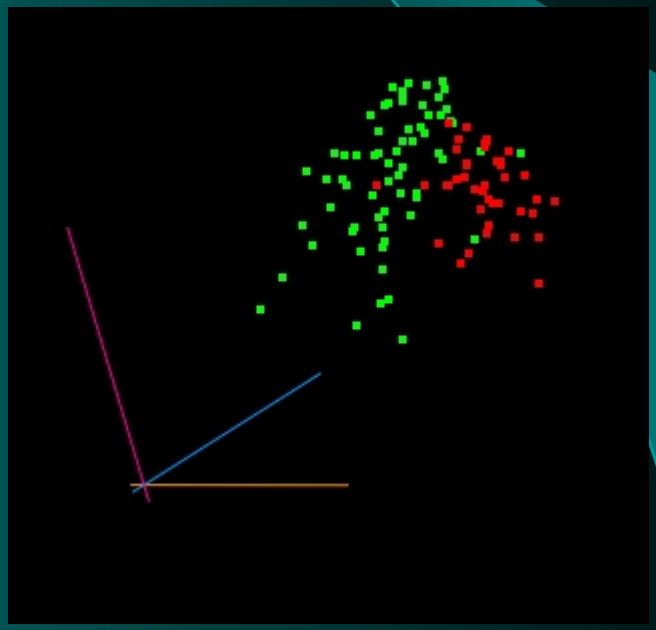
Example: Breast Cancer

Jinong Li, et al

A. All peaks



B. Three peaks



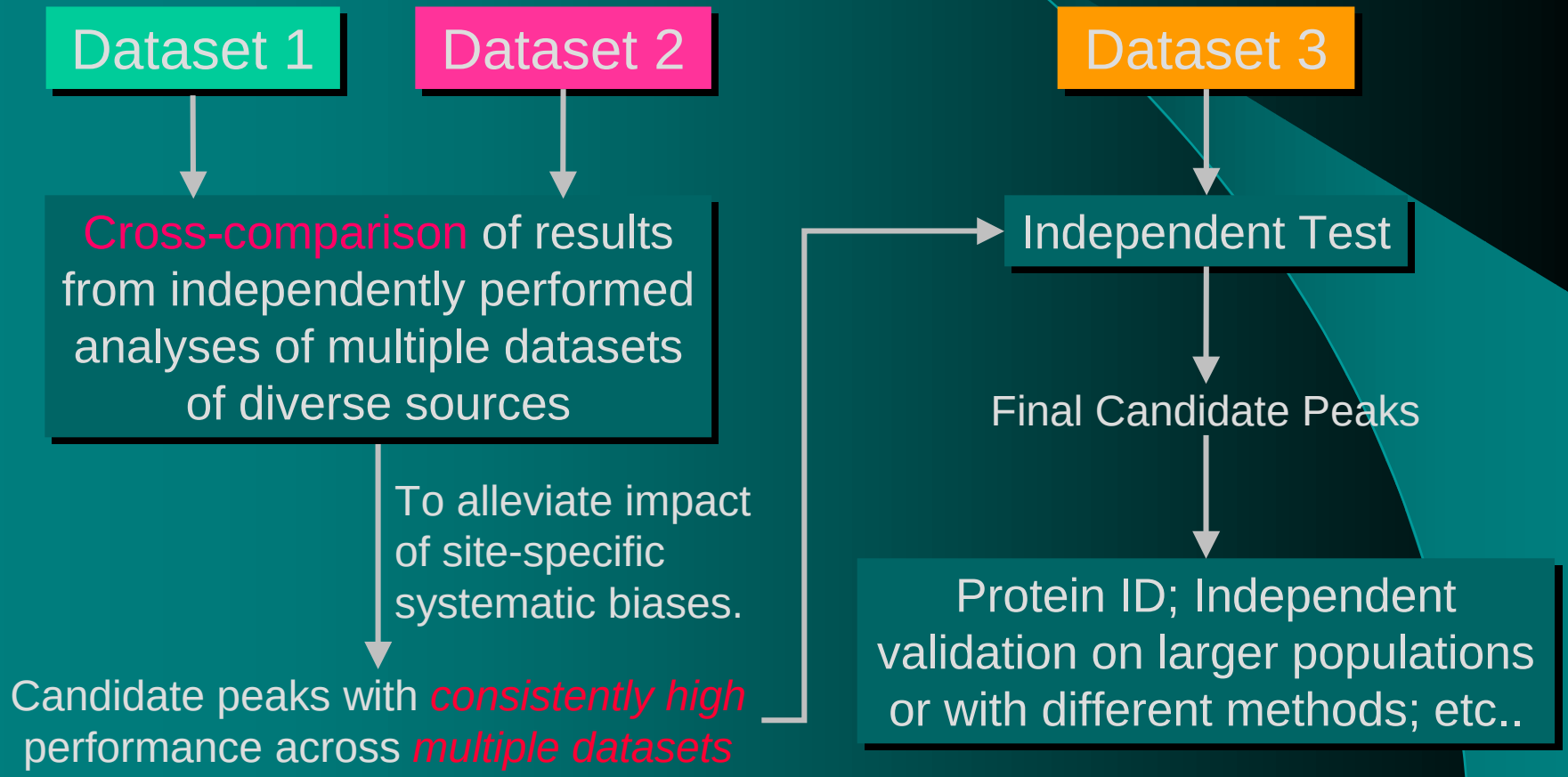
Issues w.r.t. Bioinformatics

Study Design

Considerations

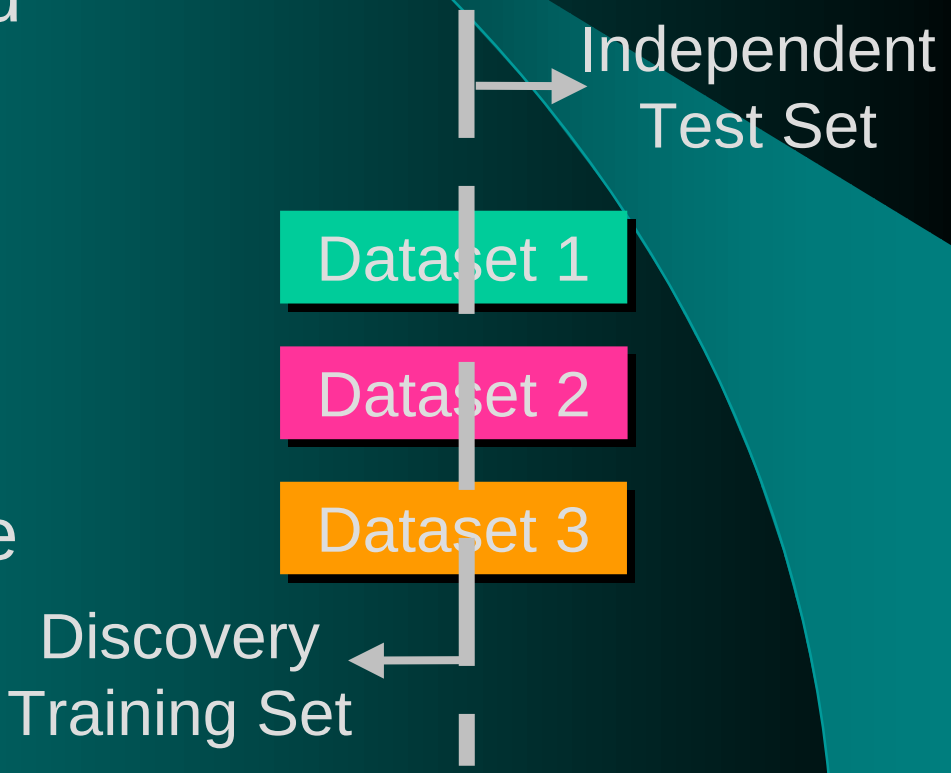
- $V_{obs} = V_d + V_p + V_a + V_b + e.$
- Many variables are not independently and identically distributed (i.i.d.) across different sites
- Hopefully, the real biomarkers are i.i.d..

Analysis of Data from Multiple Sites

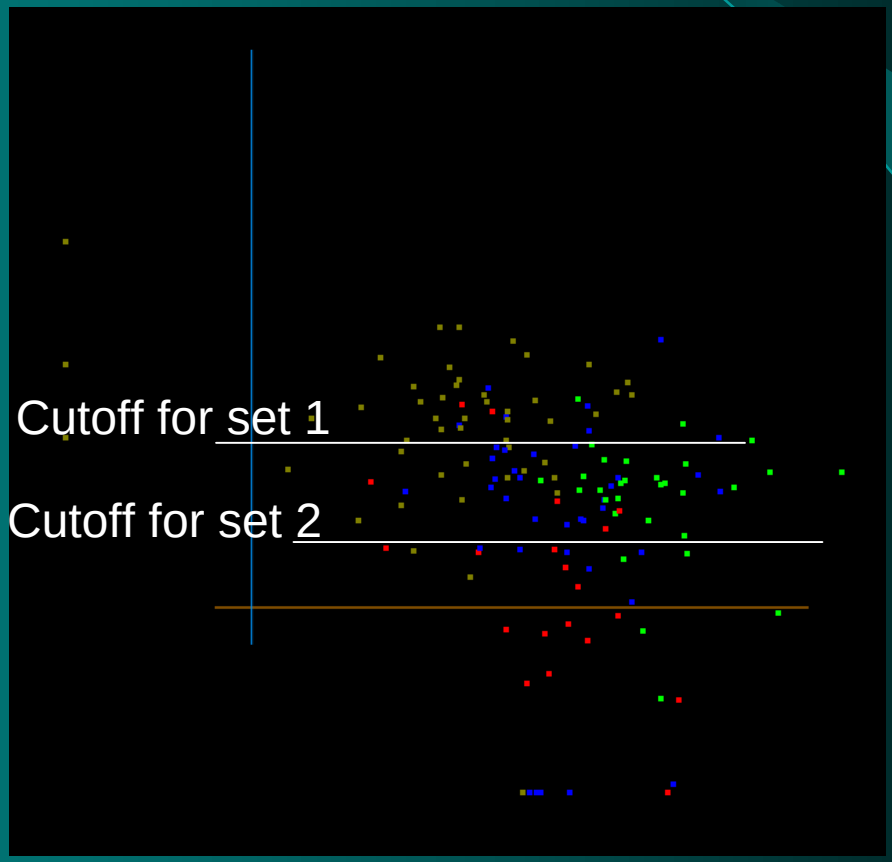


An Alternative (and Common) Approach

Pros: A more diversified dataset for biomarker discovery.
Cons: The discovery/ training set is artificially guaranteed to have the same distribution as the independent test set (i.i.d. condition).



Pros & Cons



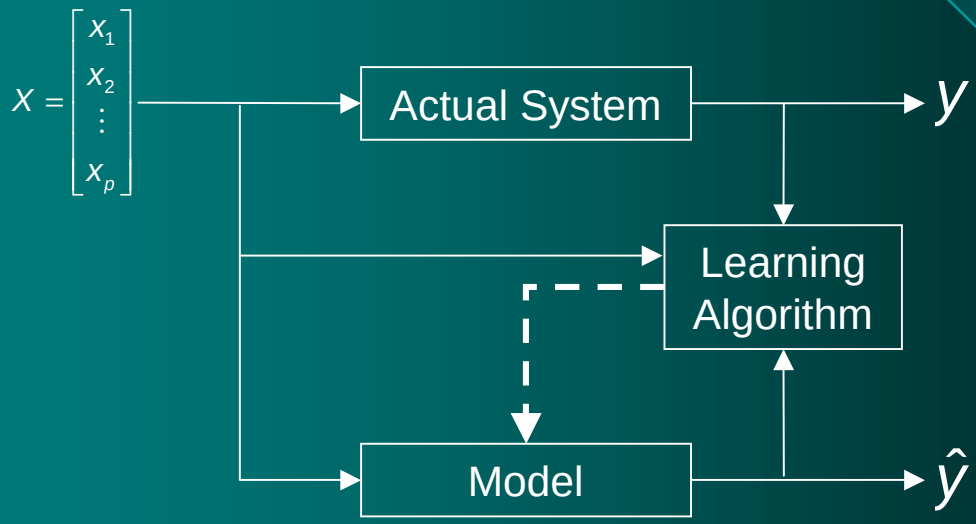
Issues w.r.t. Bioinformatics

Construction of Multivariate Models

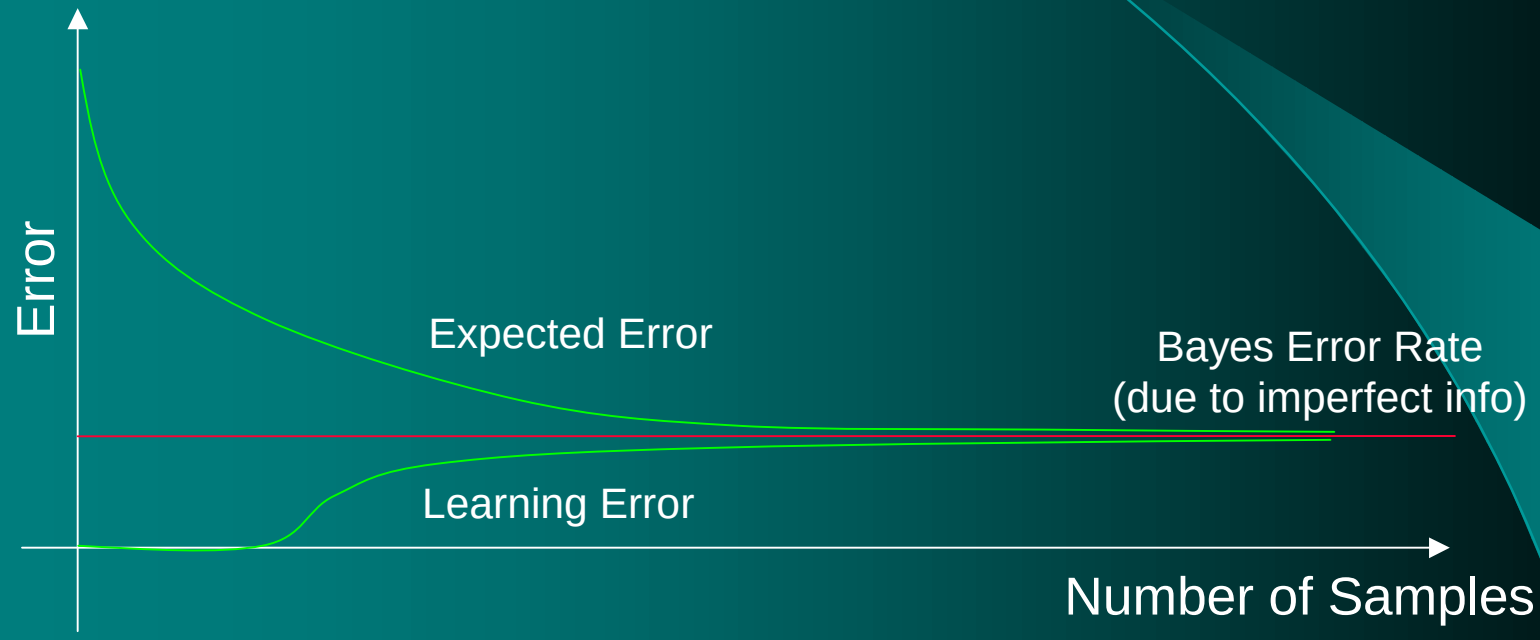
Two Separate Aspects

- Model's capacity to match complexity of problem.
- Learning algorithm's ability to use information in training data.

Model + Learning Algorithm

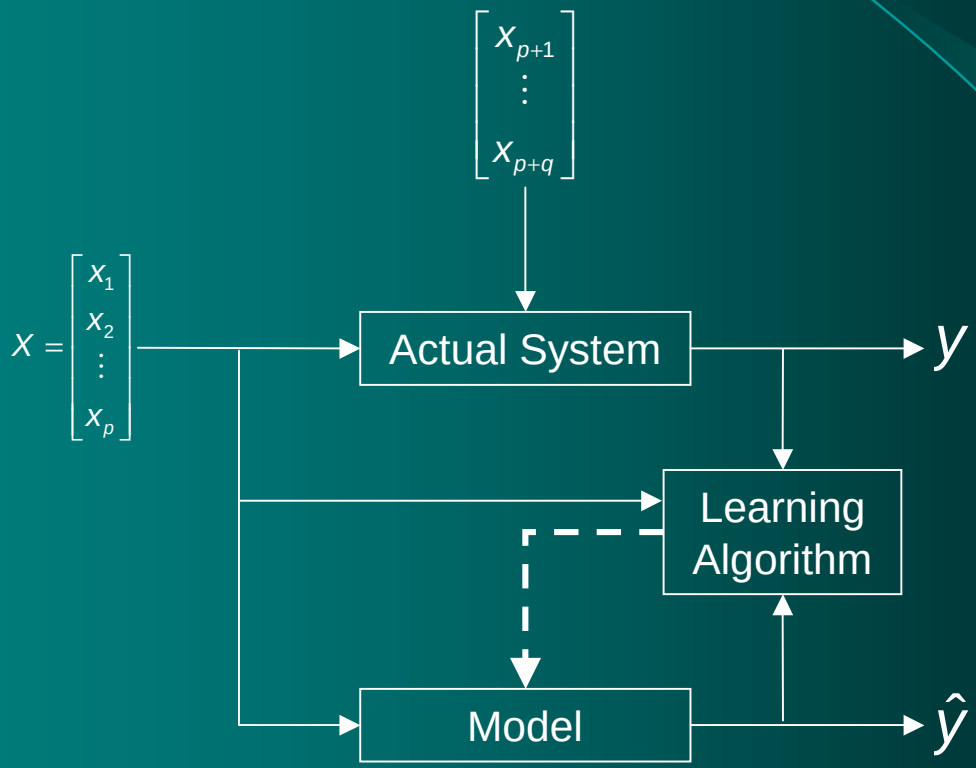


Efficiency in Information Use



Adopted from “The Nature of Statistical Learning Theory”

Imperfect Information



Easy vs. Hard vs. Impossible problems

Check Information in Input Variables

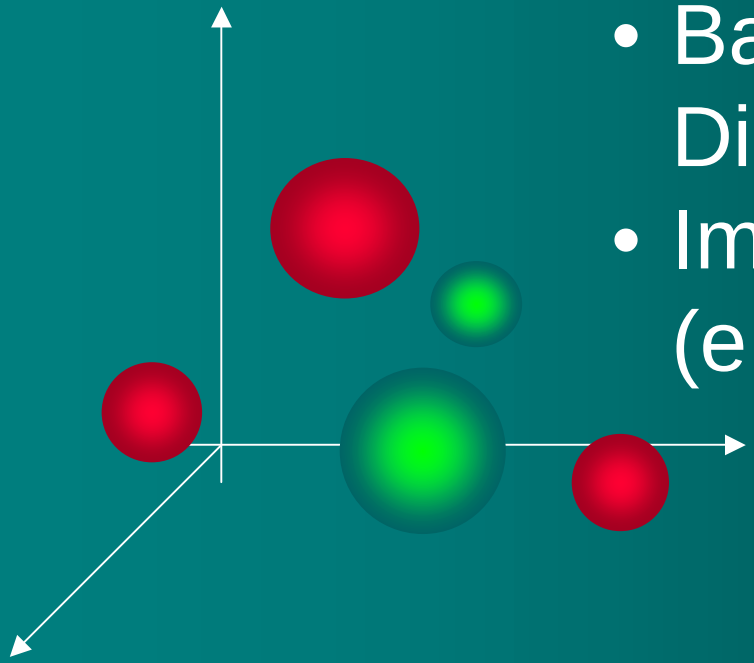
- Easy problem, almost anything works;
- Impossible problem, people still try.
- Hard problem, what really matters.

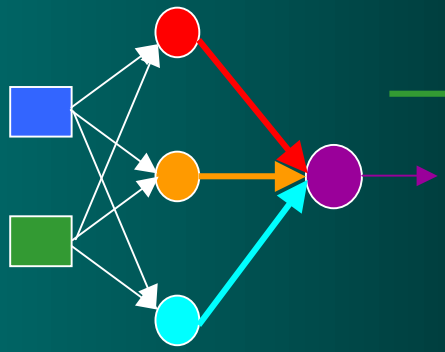
How to find out?

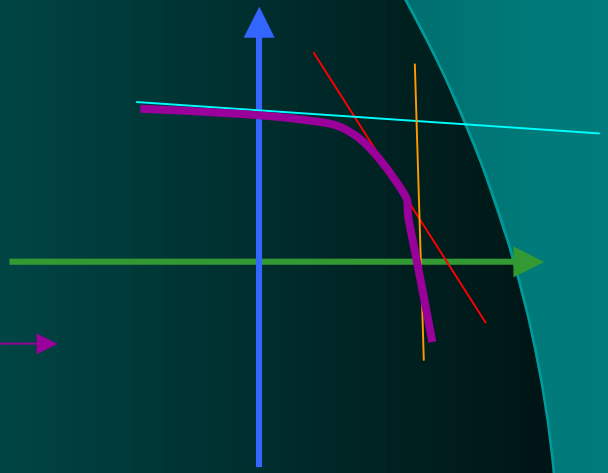
- For nonlinear models, there is no close form analytical solutions.
- Experimentally, the flip-flop phenomenon in learning/test (assuming the learning is done appropriately).

Biological Consideration

- Basis for the clusters in N-Dim space;
- Imposition of monotonicity (e.g. in ANN)







Lessens learned

- Bring "BIO" back into bioinformatics.
- It's an imperfect world;
- Study design/protocol and experimental design first. Bioinformatics cannot fix a faulty study;
- Knowledge of clinical and biological reality keeps us grounded. It takes knowledge to discover knowledge; and
- If it is too good to be true, …