

**Toward a System to Support Automated
Development and Evaluation of Cancer
Diagnostic Models from Gene Expression Data**

Alexander Statnikov

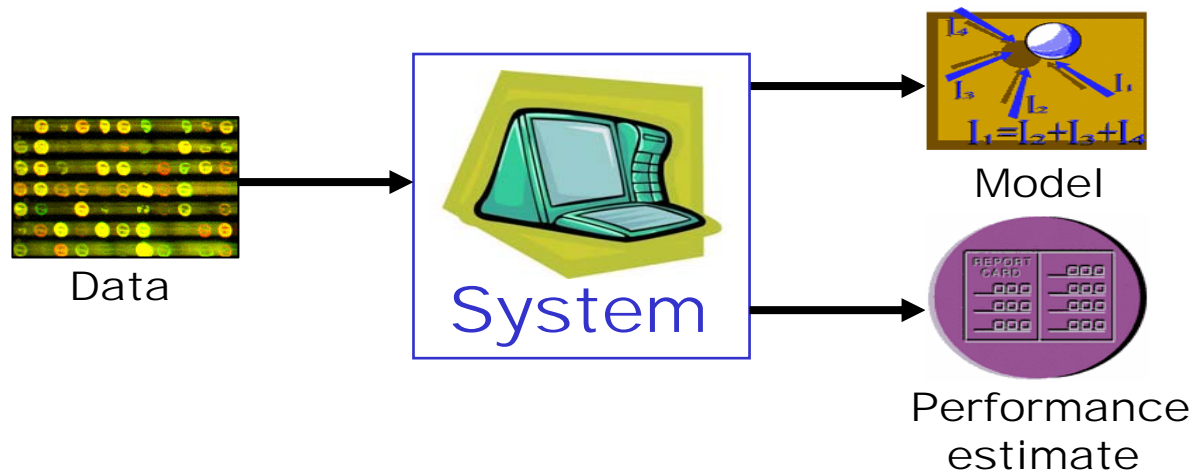
11/19/2003

Department of Biomedical Informatics,
Vanderbilt University

Problem

To build **fully automated** software system which:

1. *Develops optimal classification models* for cancer diagnosis from gene expression data;
2. *Estimates their [models'] performance* using sound experimental procedures.



Why do we need this system?

1. The problem of *overfitting* which results in **classifiers that may generalize poorly to new data despite excellent performance on the training data.**

Recent reports questioning generalization ability of classifiers produced by major studies in the field:

- ✓ **Schwarzer, 2000:** *On the misuses of NN in prognostic and diagnostic classification in oncology.*
- ✓ **Reunanen, 2003:** *Overfitting in making comparison between variable selection methods.*
- ✓ **Guyon, 2003:** *Gene selection experimental methodology flaw.*
- ✓ **Ntzani, 2003:** *Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment*

Why do we need this system?

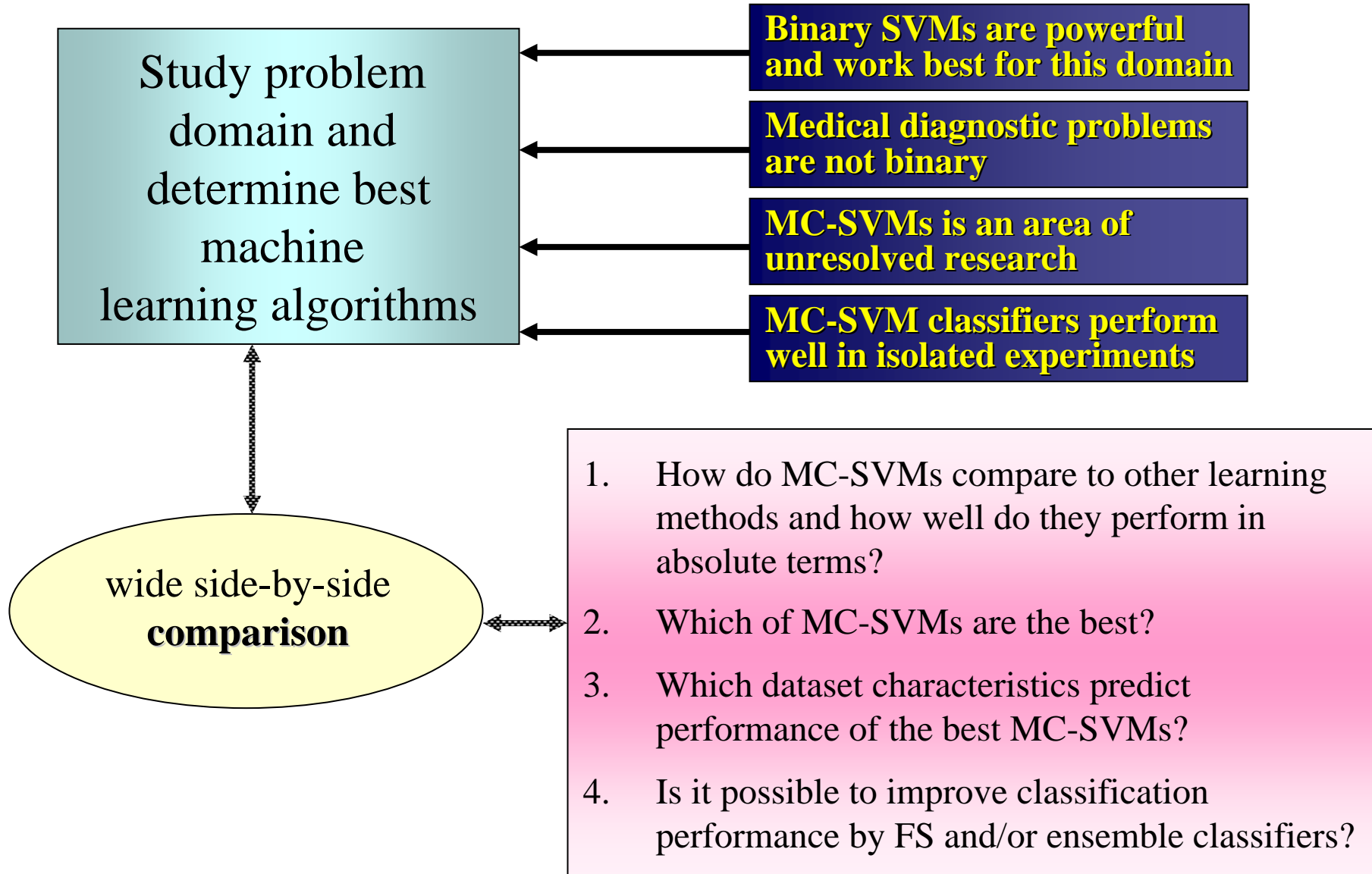
2. The problem of *underfitting* which results in **classifiers that may not be powerful due to limited experimentation (e.g., parameters of algorithms are not optimized, alternative algorithms are not considered, etc).**

Examples of possible underfitting in major studies in the field:

- ✓ **Furey, 2000:** *Support vector machine classification and validation of cancer tissue samples using microarray expression data.*
- ✓ **Guyon, 2002:** *Gene selection for cancer classification using support vector machines.*
- ✓ **Ramaswamy, 2001:** *Multiclass cancer diagnosis using tumor gene expression signatures.*

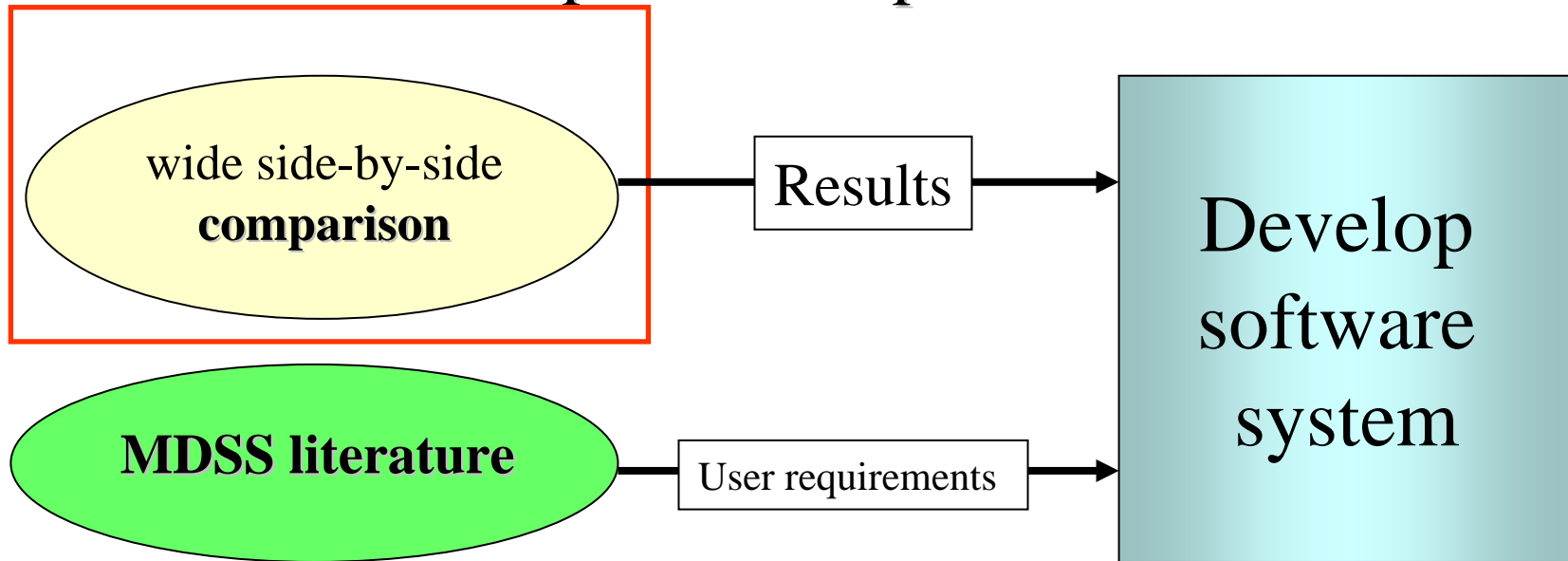
How will we build this system?

Step 1: Study problem domain



How will we build this system?

Step 2: Develop software



1. Based on the **best classification algorithms**
2. **Encapsulates most of experimental procedures**
3. Provides **intuitive wizard-like user interface**
(does not require expertise in data analysis)
4. Developed using convenient software architecture (**client-server**)

Prior research

MC-SVM research in clinical bioinformatics

Study	Number of datasets	MC-SVM methods	Optimization of SVM parameters	Evidence of selection bias*	Experiments without dimensionality reduction	Comparison with other ML algorithms
<i>Ramanwamy, 2001; Yeang, 2001</i>	1	one-vs-rest; one-vs-one	No	Yes	Yes	Yes
<i>Su, 2001</i>	1	one-vs-rest	No	No	No	No
<i>Yeo, 2001</i>	1	one-vs-rest	No	No	Yes	Yes
<i>Lee, 2003</i>	1	MSVM	Yes, but not in a nested fashion	No	Yes	Yes, based on literature only

“...results demonstrate the feasibility of accurate multiclass molecular cancer classification [by use of MC-SVMs] and suggest a strategy for future clinical implementation of molecular cancer diagnosis”

“This study demonstrates the feasibility of predicting tissue origin of a carcinoma in a context of multiple cancer classes [by use of MC-SVMs]”

“Small round, blue cells tumors can be easily classified [perfectly] into their classes by classical methods for classification, such as ... linear [multiclass] SVM classifiers...”

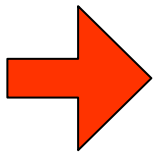
“We demonstrated that MSVM’s can classify cancer types accurately based on gene expression profiles”

All studies conclude that MC-SVM algorithms are very promising and perform very well.

Prior evaluations of MC-SVM algorithms

The **only** existing evaluation is done by Hsu and Lin (Hsu, 2001):

- Compared all available MC-SVM algorithms:
 - ▶ One-vs-rest;
 - ▶ One-vs-one;
 - ▶ DAGSVM;
 - ▶ Method by Weston and Watkins;
 - ▶ Method by Crammer and Singer.
- Employed **datasets were of non-medical nature** (e.g. wine recognition, letter recognition, shuttle control);
- **Number of variables 4 – 180 (avg. = 32) is not representative of gene expression datasets.**



This evaluation does not apply to biomedicine, especially to gene expression domain.

Methods and Materials

Classification algorithms

1. MC-SVM algorithms:

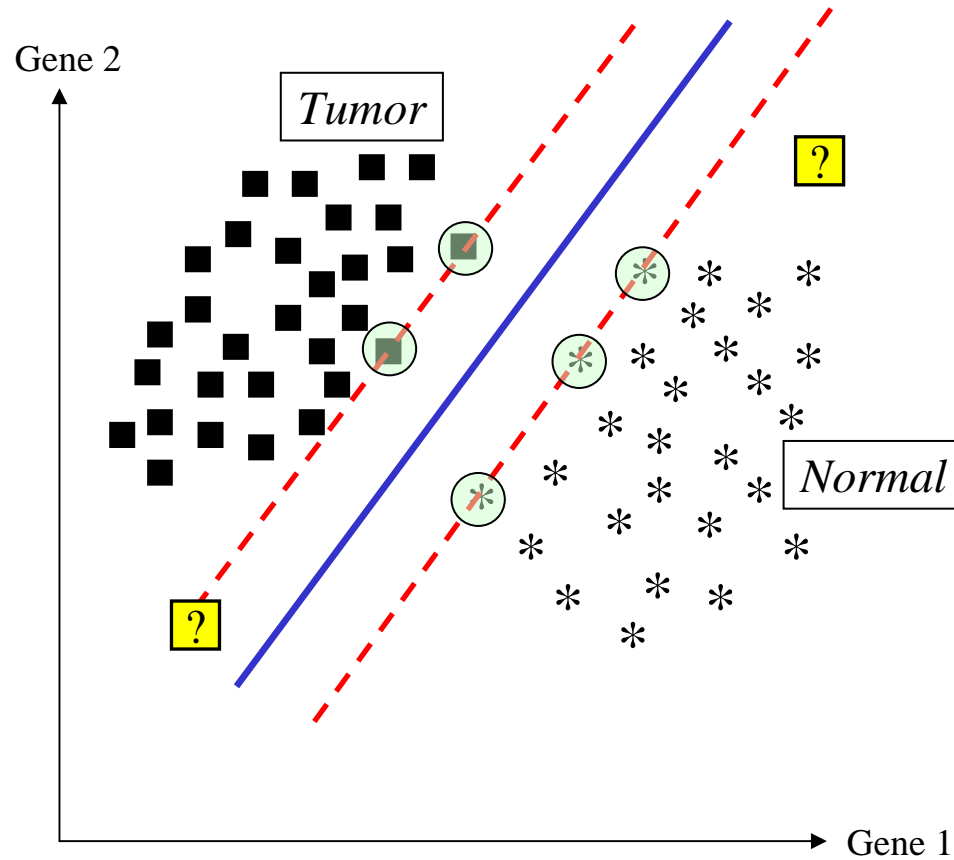
- One-vs-rest (OVR)
- One-vs-one (OVO)
- DAGSVM
- Method by Weston and Watkins (WW)
- Method by Crammer and Singer (CS)

2. Non-SVM algorithms:

- ▶ KNN
- ▶ NN

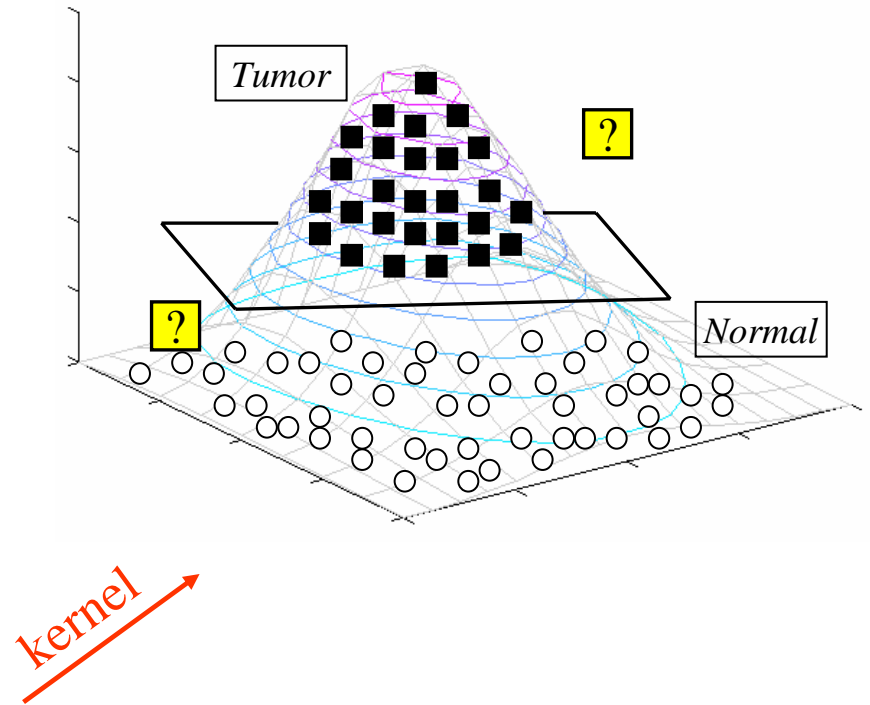
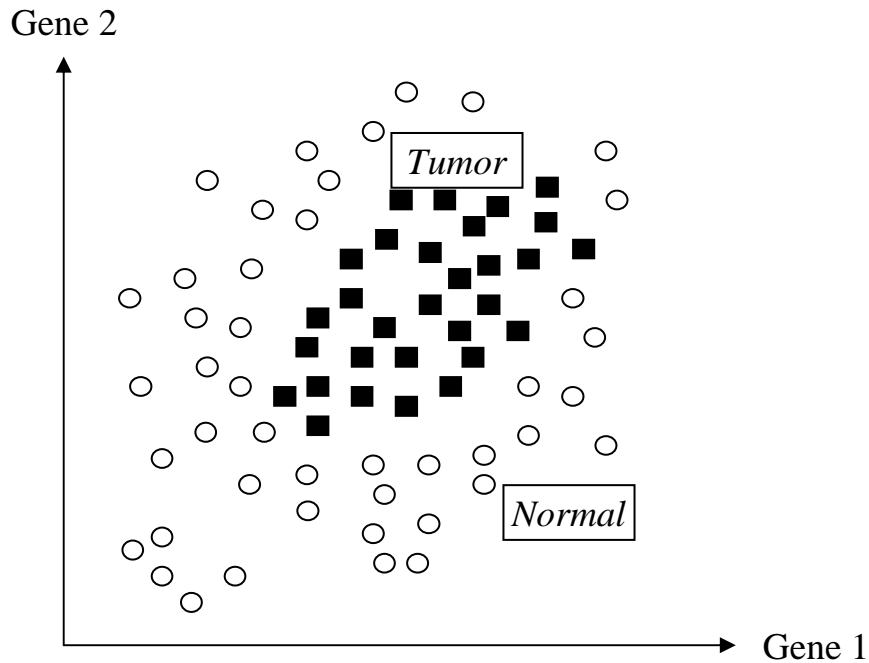
Support Vector Machines 101

Data is separable

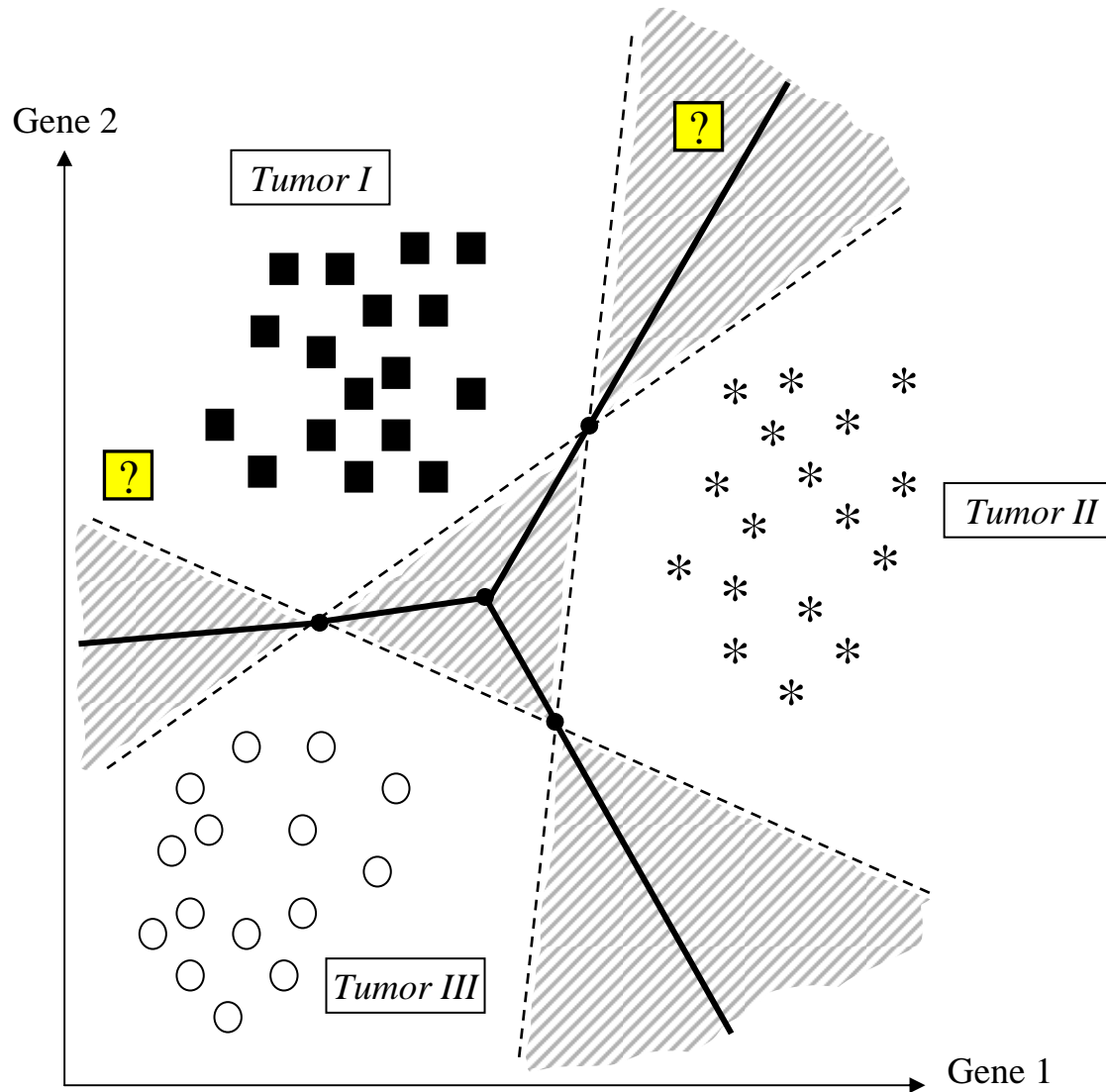


Support Vector Machines 101

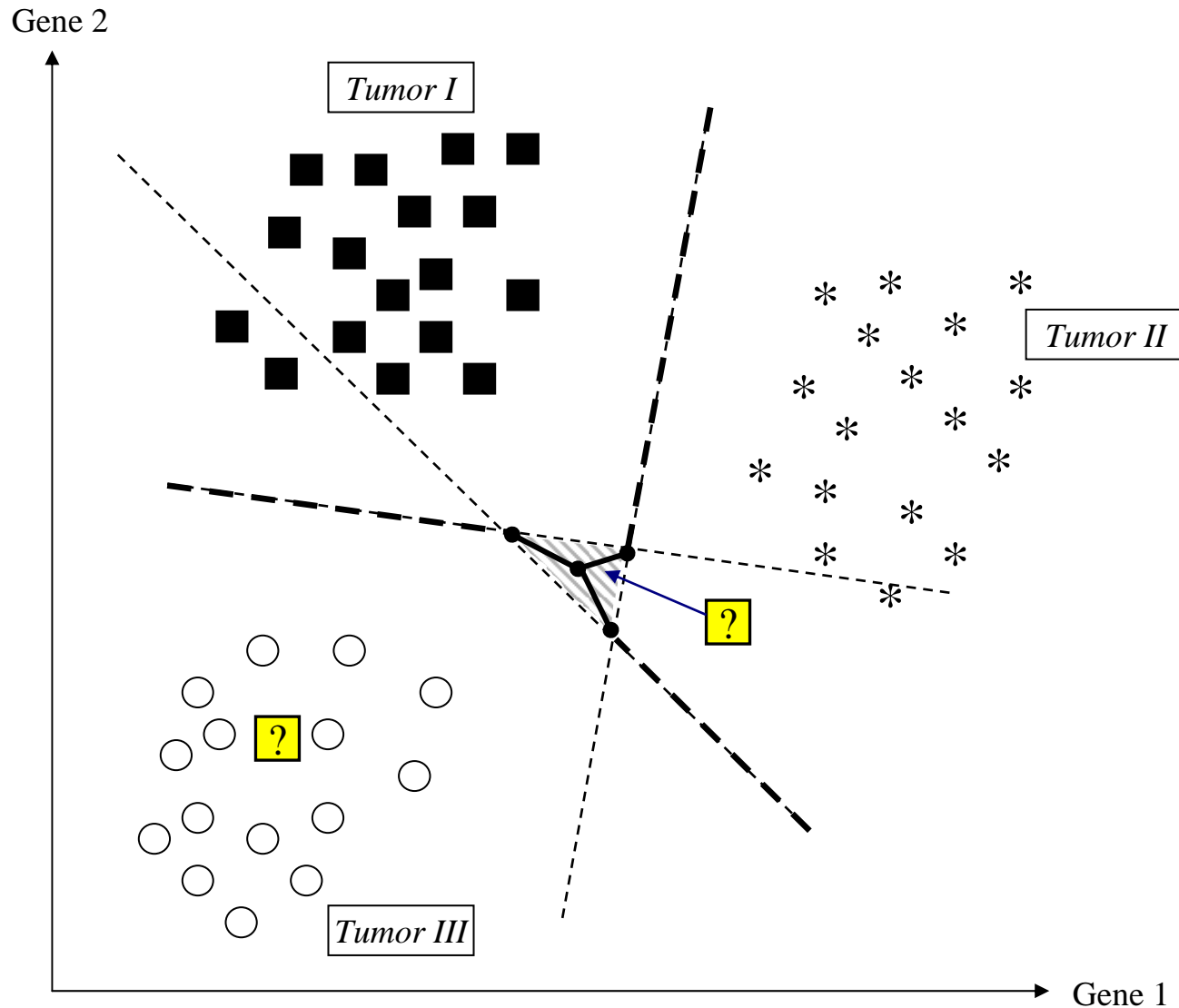
Data is not separable



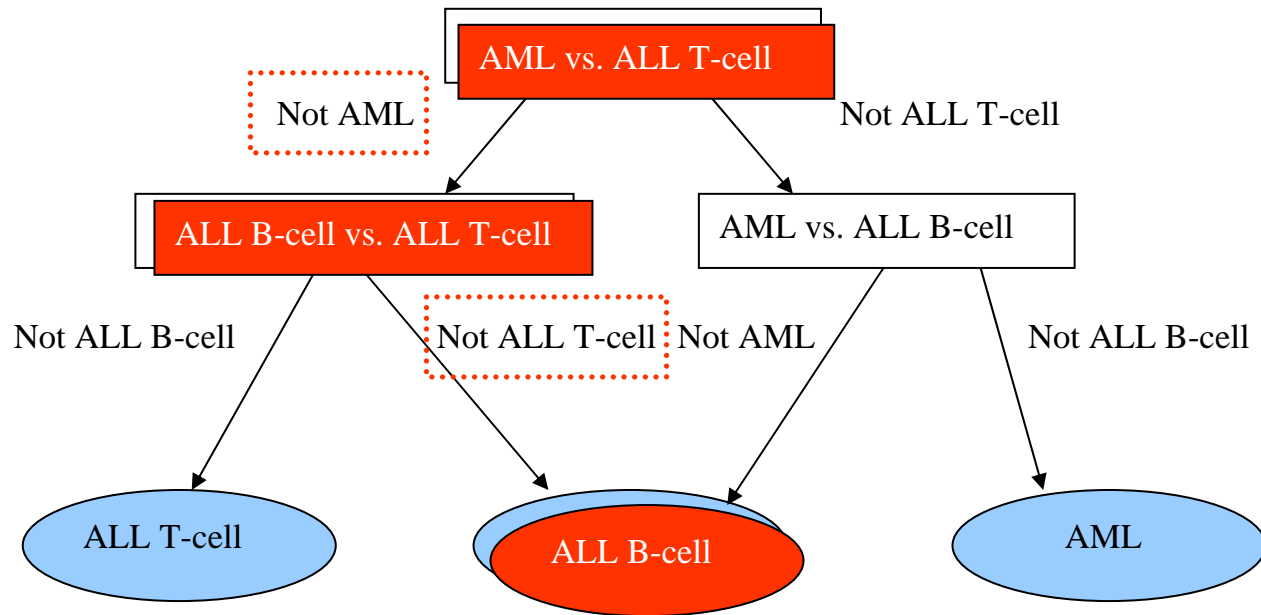
MC-SVM: One-versus-rest (OVR)



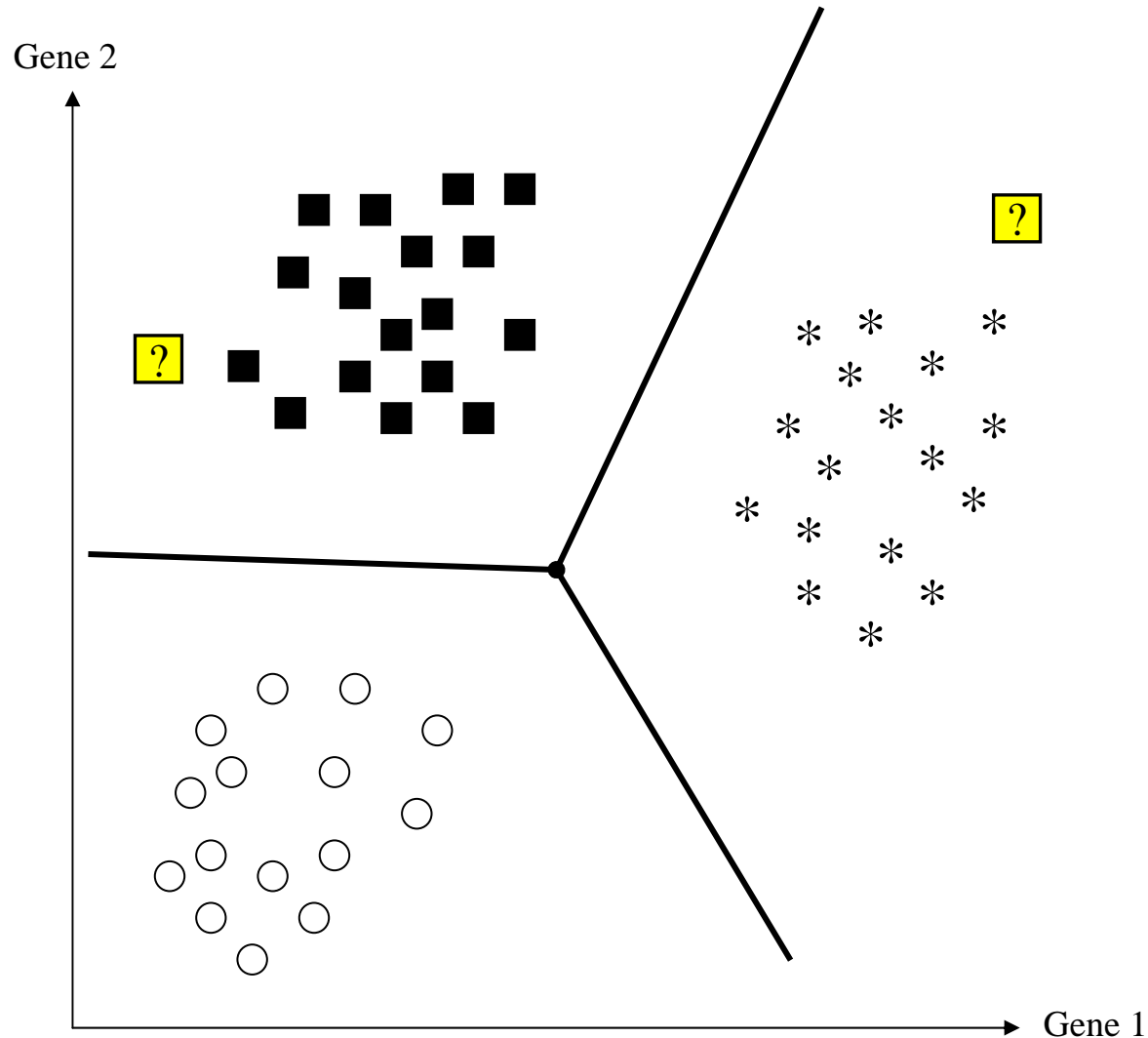
MC-SVM: One-versus-one (OVO)



MC-SVM: DAGSVM



MC-SVM: Method by Weston and Watkins (WW), and by Crammer and Singer (CS).



Gene Selection: Why do we need it?

Gene expression datasets contain measurements of thousands of genes.

■ **Cancer biomarker discovery:**

- To understand pathophysiology of cancer;
- To help with early disease detection and surrogate endpoints in clinical trials.

■ **Improving accuracy of diagnostic models:**

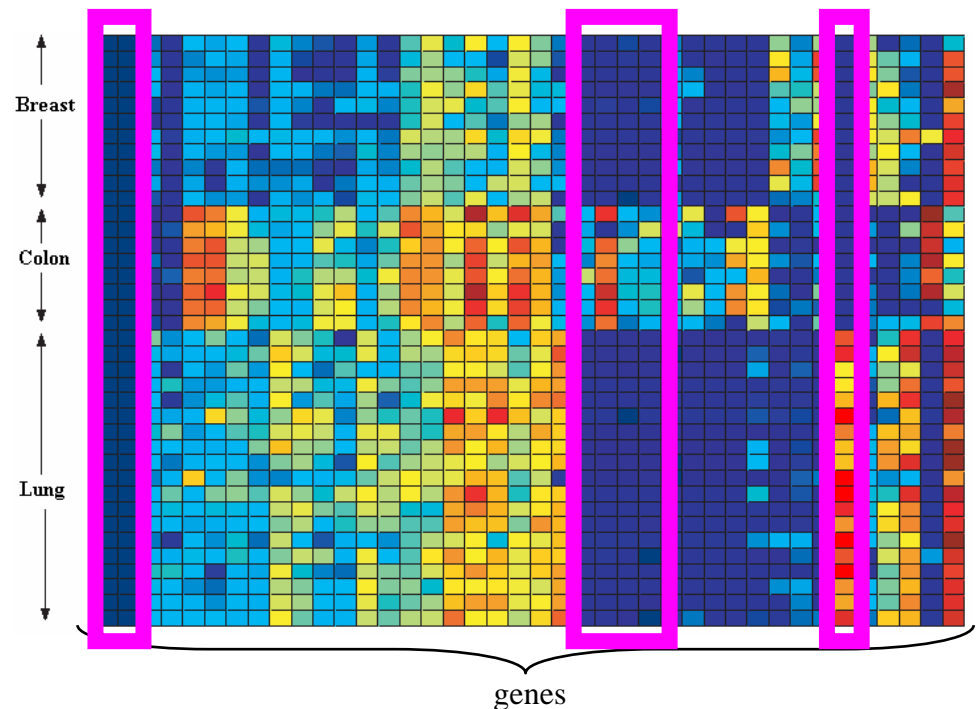
- Classification performance of many learning algorithms can be improved by reduction of dimensionality.

Application of Feature Selection

Ideally one would like to cross-validate number of features (and possibly, FS algorithm).

However, due to computational complexity of the factorial experiment, classifiers were developed with subsets of **{25, 50, 100, 500, 1000}** top-ranked genes according to the following metrics:

- ✚ Ratio of features between-categories to within-category sum of squares;
- ✚ Signal-to-noise ratio in a one-versus-rest fashion;
- ✚ Signal-to-noise ratio in a one-versus-one fashion;
- ✚ Kruskal-Wallis nonparametric one-way ANOVA



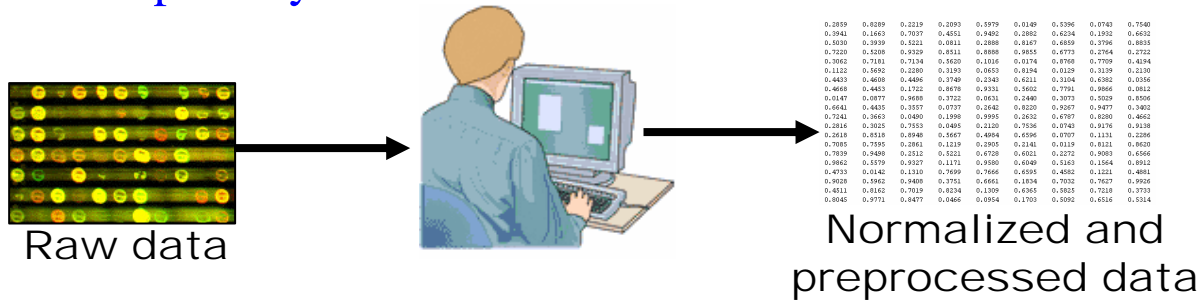
Datasets and data preparatory steps

11 datasets:

- 10 oligonucleotide-based (Affymetrix)
- 1 cDNA (Research Genetics)

- ◆ 2-26 distinct diagnoses (75 cancer types total)
- ◆ 50-308 samples (~1300 patients total)
- ◆ 2308-15009 genes

We adapted standard normalization and data preparatory steps performed by the authors of the primary studies.

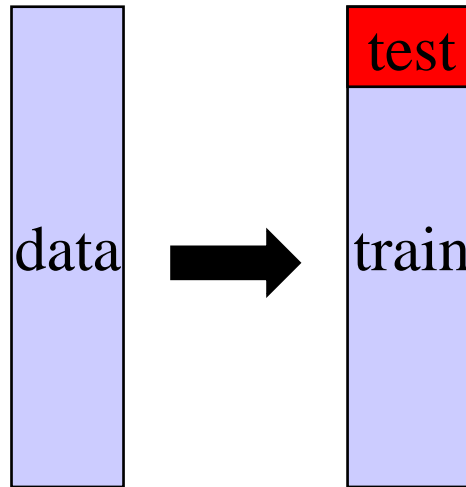


In addition, we performed rescaling of gene expression values to speed-up training of SVMs.

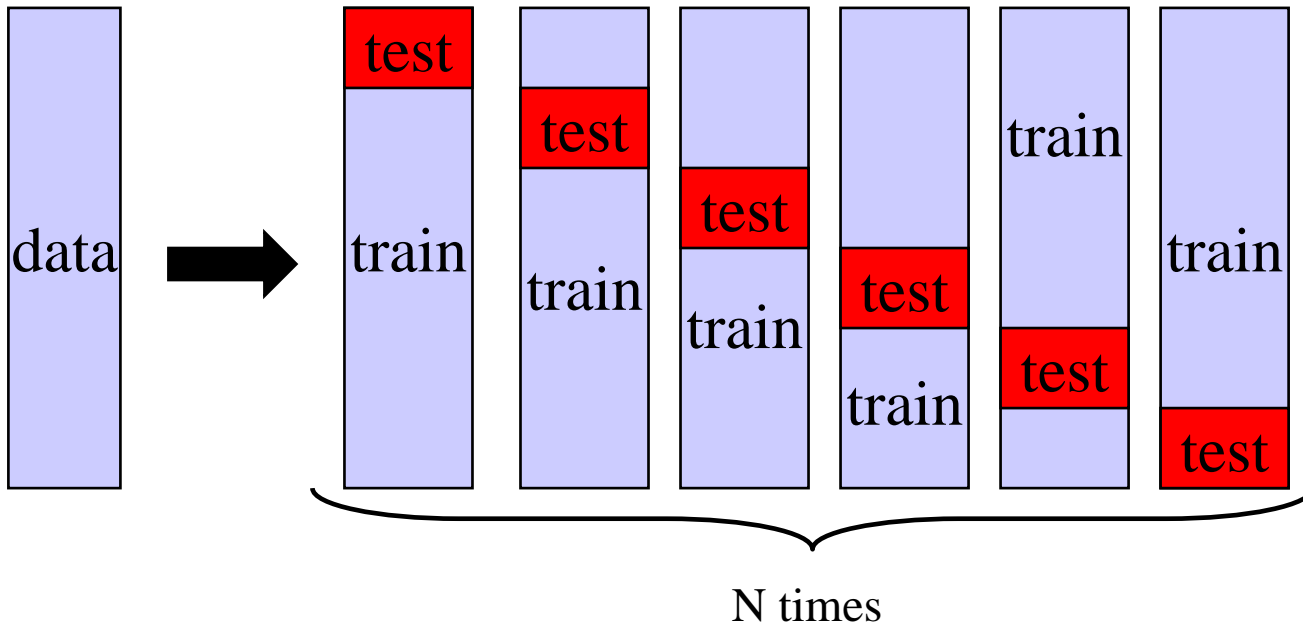
Cross-Validation 101

Hold-out cross-validation

Learn on **train** and
estimate error on **test**:

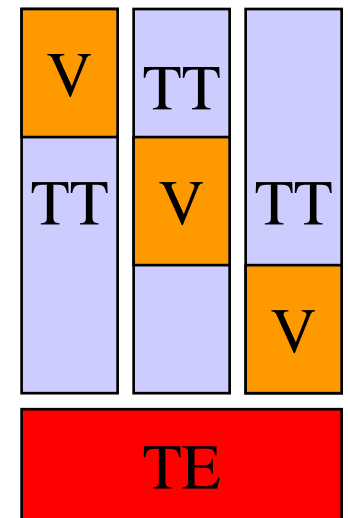
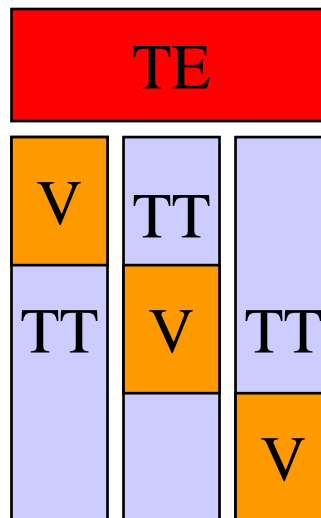
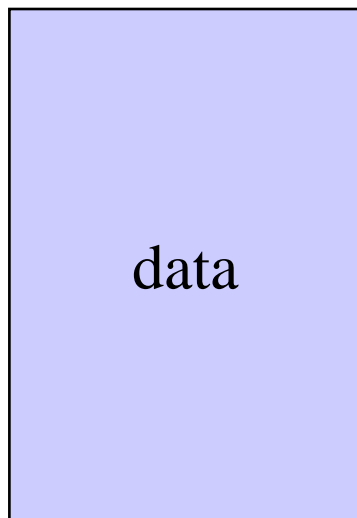
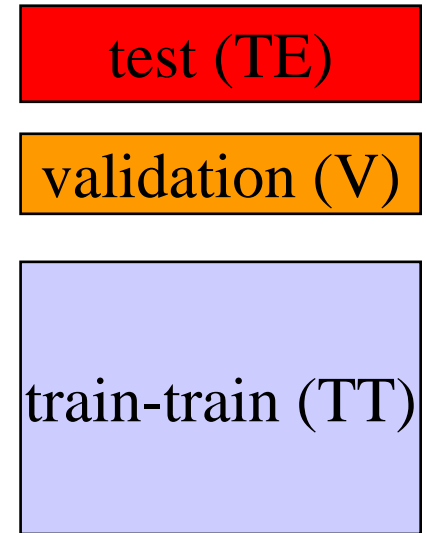
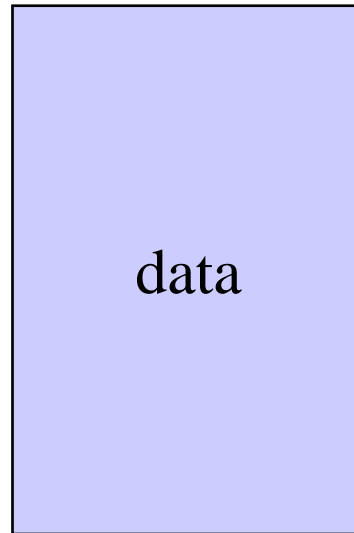


N-fold cross-validation:



Cross-Validation 101

What if classifier is parametric?



Experimental design

Two experimental designs were employed to estimate performance:

- **Design I:** stratified N-fold (N=10) CV in the outer loop and N-1-fold CV in the inner loop;
- **Design II:** LOOCV in the outer loop and N-fold (N=10) CV in the inner loop

LOOCV and N-fold (N=10) CV was used to select the best model.

Performance Metric and Statistical Comparison

We chose to use **accuracy as performance metric** for the following reasons:

- Generally accepted misclassification functions do not currently exist for the studied domain;
- Possible alternatives to area under ROC curve are not applicable because of
 - 1) *multicategory tasks*
 - 2) *different number of classes in datasets*

To test that performance differences between the best method and remaining methods are non-random, **random permutation testing** was used.

Results and Conclusions

Classification without gene selection

		<i>Multicategory classification</i>				
Method		9_Tumors	11_Tumors	14_Tumors	Brain_Tumor1	Brain_Tumor2
MC-SVM	OVR	65.10%	94.68%	74.98%	91.67%	77.00%
	OVO	58.57%	90.36%	47.07%	90.56%	77.83%
	DAGSVM	60.24%	90.36%	47.35%	90.56%	77.83%
	WW	62.24%	94.68%	69.07%	90.56%	73.33%
	CS	65.33%	95.30%	76.60%	90.56%	72.83%
non-SVM	KNN	43.90%	78.51%	50.40%	87.94%	68.67%
	NN	19.38%	54.14%	11.12%	84.72%	60.33%

❖ MC-SVM significantly outperform KNN and NN

❖ MC-SVM diagnose with accuracy >90% in 8/11 datasets

❖ MC-SVM methods OVR, WW, and CS work best

		<i>Multicategory classification</i>				<i>Binary classification</i>	
Method		Leukemia1	Leukemia2	Lung_Cancer	SRBCT	Prostate_Tumor	DLBCL
MC-SVM	OVR	97.50%	97.32%	96.05%	100.00%	92.00%	97.50%
	OVO	97.32%	95.89%	95.59%	100.00%	92.00%	97.50%
	DAGSVM	96.07%	95.89%	95.59%	100.00%	92.00%	97.50%
	WW	97.50%	95.89%	95.55%	100.00%	92.00%	97.50%
	CS	97.50%	95.89%	96.55%	100.00%	92.00%	97.50%
non-SVM	KNN	83.57%	87.14%	89.64%	86.90%	85.09%	86.96%
	NN	76.61%	91.03%	87.80%	91.03%	79.18%	89.64%

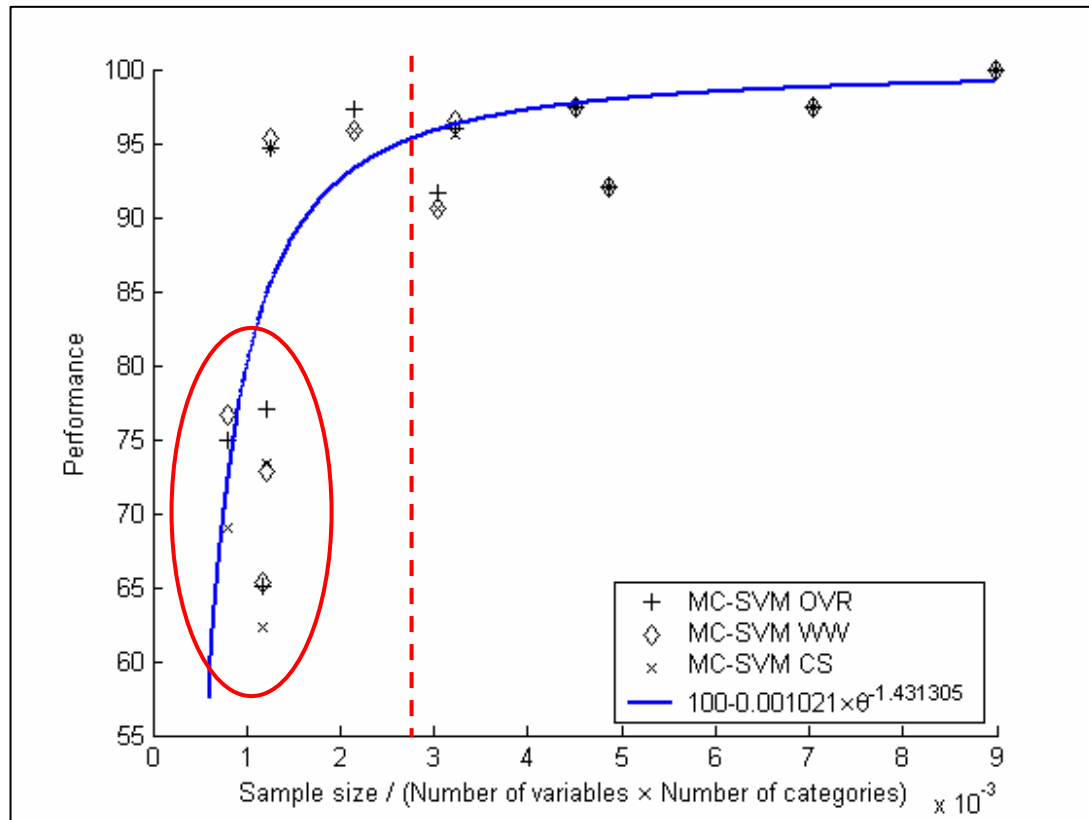
Why best MC-SVM have accuracy <90% in 3/11 datasets?

Given a number of possible performance predictors, Θ :

- ❖ Number of samples
- ❖ Number of categories
- ❖ Number of variables
- ❖ Number of samples divided by number of categories
- ❖ Number of samples divided by number of variables
- ❖ Number of samples divided by product of number of categories and variables.

We fitted inverse power-law curves of the type:

$$\text{Performance}^* = 100\% - a \Theta^{-b}$$



Time results*

Method		Time in hours	
		Design I	Design II
MC-SVM	OVR	19.28	772.43
	OVO	9.86	388.11
	DAGSVM	9.93	390.97
	WW	7.95	290.77
	CS	7.88	289.01
non-SVM	KNN	3.40	109.60
	NN	195.68	N/A

❖ CS and WW are the fastest MC-SVM methods

* All experiments were executed on Intel Xeon 2.4GHz dual-CPU workstations.

Improving classification performance with gene selection.

Focus on most “difficult” datasets. *Preliminary results.*

Method		9_Tumors		14_Tumors		Brain_Tumor1		Brain_Tumor2	
		w/o FS	with FS	w/o FS	with FS	w/o FS	with FS	w/o FS	with FS
MC-SVM	OVR	65.10%	69.62%	74.98%	74.98%	91.67%	92.67%	77.00%	85.67%
	OVO	58.57%	68.76%	47.07%	52.66%	90.56%	90.56%	77.83%	81.50%
	DAGSVM	60.24%	68.76%	47.35%	53.61%	90.56%	90.56%	77.83%	81.50%
	WW	62.24%	68.67%	69.07%	69.07%	90.56%	91.67%	73.33%	82.33%
	CS	65.33%	74.86%	76.60%	76.60%	90.56%	90.56%	72.83%	77.50%
non-SVM	KNN	43.90%	56.43%	50.40%	65.37%	87.94%	89.31%	68.67%	80.17%
	NN	19.38%	67.00%	11.12%	70.90%	84.72%	90.67%	60.33%	78.33%

- ❖ FS improves classification performance of MC-SVMs (up to 9.5%)
- ❖ FS improves classification performance of KNN and NN significantly (up to 59.8%)

Software Demonstration

DSL MC-SVM System File Task

variables: 12601 observations: 203 The first variable (column) of the dataset should be a target variable.

Dataset:

Use gene names for output report:

Use gene accession numbers for output report:

Experimental design: N-fold cross-validation (CV). Number of folds:
 Leave-one-out cross-validation (LOOCV)

Number of folds for parameter optimization (inner loop) of LOOCV:

Generate sample splits: Yes, and do not save splits
 Yes, save splits into file:

 No, use existing sample splits:

MC-SVM classification methods: DVR OVO DAGSVM WW CS

Sequence of normalization steps (for each feature x , across all observations):

<input type="radio"/> A. $\log(x)$, logarithm base: <input type="text"/>	<input type="radio"/> E. $x / \text{mean of } x$
<input checked="" type="radio"/> B. $[a, b]$, a: <input type="text" value="0"/> and b: <input type="text" value="1"/>	<input type="radio"/> F. $x / \text{median of } x$
<input type="radio"/> C. $(x - \text{mean of } x) / \text{std of } x$	<input type="radio"/> G. $x / \text{norm of } x$
<input type="radio"/> D. $x / \text{std of } x$	<input type="radio"/> H. $x - \text{mean}(x)$
	<input type="radio"/> I. $x - \text{median}(x)$
	<input type="radio"/> J. $ x $

Feature selection: None
 Nonparametric one-way ANOVA (Kruskal-Wallis)
 Signal-to-noise ratio in a one-versus-rest fashion
 Signal-to-noise ratio in a one-versus-one fashion
 Ratio of features between categories to within-category sum of squares

Number of features: Optimized. Try from to features, step
 Specific:

Kernel for SVM algorithm: Polynomial (including linear)
 Radial base functions

Optimize parameters of SVM: Yes
 No, use cost:
and degree:
and gamma: Default value: 0.0049261

Optimization grid for parameters of SVM:

Cost: to multiplicative step
Degree: to step
Gamma: to multiplicative step

Output log: Yes, log into file:

 No, output log on the screen

Task: Estimate performance.
 Generate best model. Output:

Save report in:

Performance estimation options: Use parameters specified above
 Use previously generated best model:

and a set of independent samples:

MC-SVM Tool: Experimental Report

Task:	Generate best model
Experiment execution time:	46 seconds
Number of samples:	203
Number of variables:	12601
Number of categories:	5
Validation accuracy:	96.5517%
Dataset filename:	D:\Sasha\Matlab\MC-SVM\Toolbox_Development\distributive\data\Lung_Cancer\data.txt
Gene names filename:	D:\Sasha\Matlab\MC-SVM\Toolbox_Development\distributive\data\Lung_Cancer\gene_names.nam
Gene accession numbers filename:	D:\Sasha\Matlab\MC-SVM\Toolbox_Development\distributive\data\Lung_Cancer\gene_accessions.acc
Model filename:	model.mod

Description of the best model for the current data-split:

- SVM method: OVR
- SVM cost: 100
- SVM kernel: poly
- SVM kernel parameter (degree): 1

Feature selection method: **Signal-to-noise ratio in a one-versus-rest fashion**

Optimal number of features: **100**

Ranking (1 is 'best')	Column index of features (in dataset file)	Gene names	Accession numbers
1	8485	Cluster Incl U81561:Human protein tyrosine phosphatase receptor pi (PTPRP) mRNA, complete cds /cds=(42,3038) /gb=U81561 /gi=2351575 /ug=Hs.74624 /len=4699	U81561
2	5850	RaP2 interacting protein 8	AF055026
3	3074	piccolo (presynaptic cytomatrix protein)	AB011131
4	8473	Cluster Incl U48437:Human amyloid precursor-like protein 1 mRNA, complete cds /cds=(41,1993) /gb=U48437 /gi=1709300 /ug=Hs.74565 /len=2336	U48437
5	4854	Cluster Incl N90862:zb11b06.s1 Homo sapiens cDNA, 3 end /clone=IMAGE-301715 /clone_end=3" /gb=N90862 /gi=1444189 /ug=Hs.172684 /len=605"	L76703
6	3876	cadherin, EGF LAG seven-pass G-type receptor 3, flamingo (Drosophila) homolog	AB011536
7	3192	S100 calcium-binding protein A11 (calgizzarin)	D38583

Entrez-Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide&cmd=search&term=U81561&tool=gquery>

NCBI

CGCTCAGGATAGACTTCCGCTGCTAGATGATCGGATCCCCGGGCTATTATATAGTCGATCGATCT
 TTCTCTATATGACCGGATGSGGATATATACACACAGATGCGGATAGCATGCTGATCTA
 CACAGACTACCGCTCTACTTACTTAC TAAC CAAT TGGGAGAGGGGAGGAGATGGGCGGAG

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for U81561 Go Clear

Limits Preview/Index History Clipboard Details

Display Summary Show: 20 Send to Text

About Entrez

Search for Genes
 LocusLink provides curated information for human, fruit fly, mouse, rat, and

1: [U81561](#) Links

Human protein tyrosine phosphatase receptor pi (PTPRP) mRNA, complete cds
 gi|2351575|gb|U81561.1|HSU81561[2351575]

Display default Show: 20 Send to File Get Subsequence Features

1: [U81561](#). Human protein tyr...[gi:2351575] Links

LOCUS HSU81561 4699 bp mRNA linear PRI 03-SEP-1997

DEFINITION Human protein tyrosine phosphatase receptor pi (PTPRP) mRNA, complete cds.

ACCESSION U81561

VERSION U81561.1 GI:2351575

KEYWORDS .

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 4699)

AUTHORS Jiang,S., Tulloch,G., Fu,Y., London,R., Hummel,G.S., White,R.A., Avraham,H. and Avraham,S.

TITLE Characterization and chromosomal localization of PTPRP, a receptor protein tyrosine phosphatase predominantly expressed in brain and pancreas

JOURNAL Unpublished

REFERENCE 2 (bases 1 to 4699)

AUTHORS Jiang,S., Tulloch,G., Fu,Y., London,R., Hummel,G.S., White,R.A., Avraham,H. and Avraham,S.

TITLE Direct Submission

JOURNAL Submitted (10-DEC-1996) Genetics, The Children's Mercy Hospital, 2801 Wyandotte Ave., Kansas City, MO 64108, USA

FEATURES Location/Qualifiers

Conclusions

Step 1: Evaluation of algorithms

- ❖ To solve the cancer diagnostic problem from GE data, MC-SVMs is the preferred family of algorithms (outperforming NN and KNN);
- ❖ Overall classification performance achieved by MC-SVMs is excellent;
- ❖ Among tested algorithms, MC-SVMs CS, WW, and OVR lead to most accurate diagnostic models;
- ❖ We identified specific data characteristics which predict performance of the best MC-SVM models;
- ❖ Gene selection further improves classification performance.

Step 2: Development of system

- ❖ We created a preliminary version of a system that supports development of optimal classification models and estimates their performance using sound experimental procedures;
- ❖ **The results obtained by the system in an labor efficient manner are on par or better than previously published results in the literature on the same datasets.**

Ongoing work

We will be working with other Vanderbilt researchers, so that **our system can be run from various laboratories and best matches all needs of local scientific community.**

Priority ↑

- Make the system usable by researchers without expertise in data analysis by designing and implementing wizard-like GUI;
- Organize gene output so that it can be easier used for discovery (i.e., extend the links to existing knowledge in literature);
- Enhance Biomarker Discovery options by incorporating our own causal discovery techniques (Markov blanket & local neighborhood algorithms).

- Ensemble classification:
 - Use majority voting, decision trees, MC-SVMs to ensemble classifiers;
 - Experiment with boosting and bagging.
- Performance metrics:
 - Experiment with weighted accuracy;
 - Design and apply analogues to AUC ROC.

Acknowledgements

Members of my Master's committee:

- ❑ Constantin F. Aliferis, M.D., Ph.D. (*co-author*)
- ❑ Douglas P. Hardin, Ph.D.
- ❑ Shawn Levy, Ph.D.
- ❑ Ioannis Tsamardinos, Ph.D. (*co-author*)

The system development effort is in part supported by BISTI planning grant (Co-PI's: [Dr. Stead](#) and [Dr. Magnuson](#)) pilot project: “*Computational Models of Lung Cancer: Connecting Classification, Gene Selection, and Molecular Sub-typing*”
(Co-PI's: [Dr. Constantin F. Aliferis](#) and [Dr. Pierre Massion](#))

Other acknowledgements:

- ❑ Terry Ni, Ph.D.