

Methods for Multi-Category Cancer Diagnosis from Gene Expression Data:

*A Comprehensive Evaluation to Inform
Decision Support System Development*

Alexander Statnikov M.S.,
Constantin F. Aliferis M.D., Ph.D.,
Ioannis Tsamardinos Ph.D.

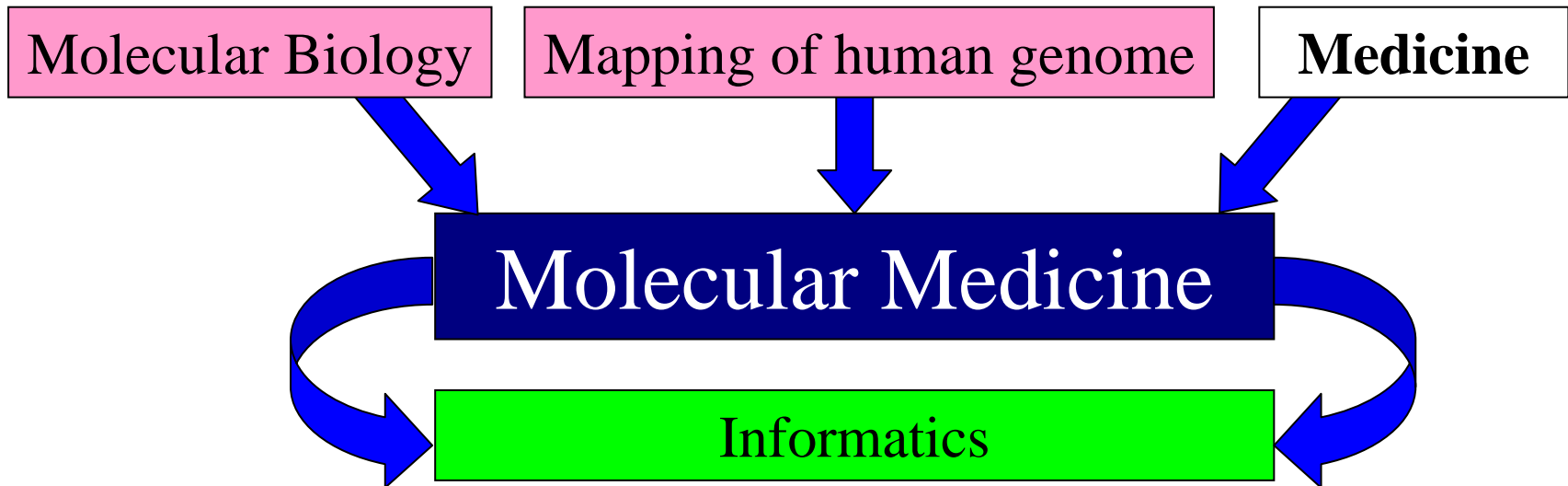
Discovery Systems Laboratory,
Department of Biomedical Informatics,
Vanderbilt University,
MEDINFO 2004



VANDERBILT
UNIVERSITY

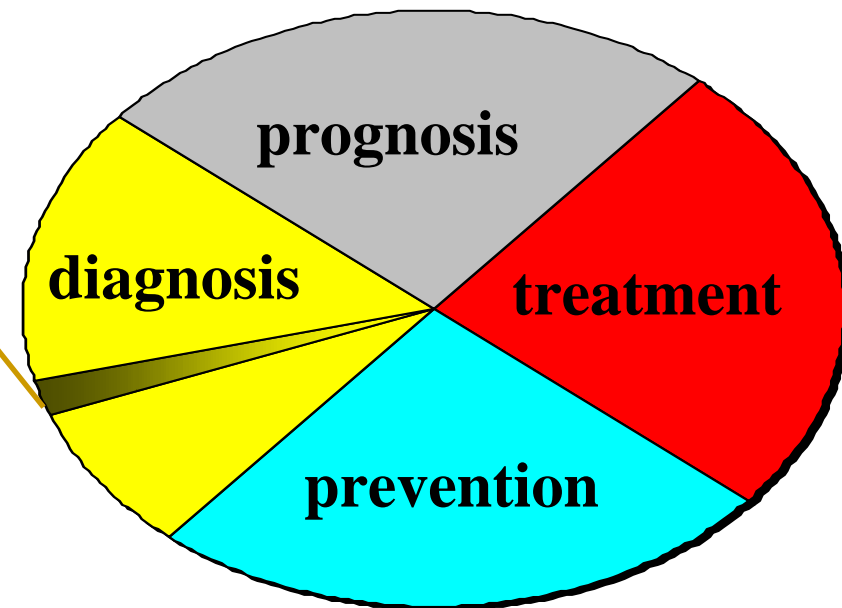


Research Problem

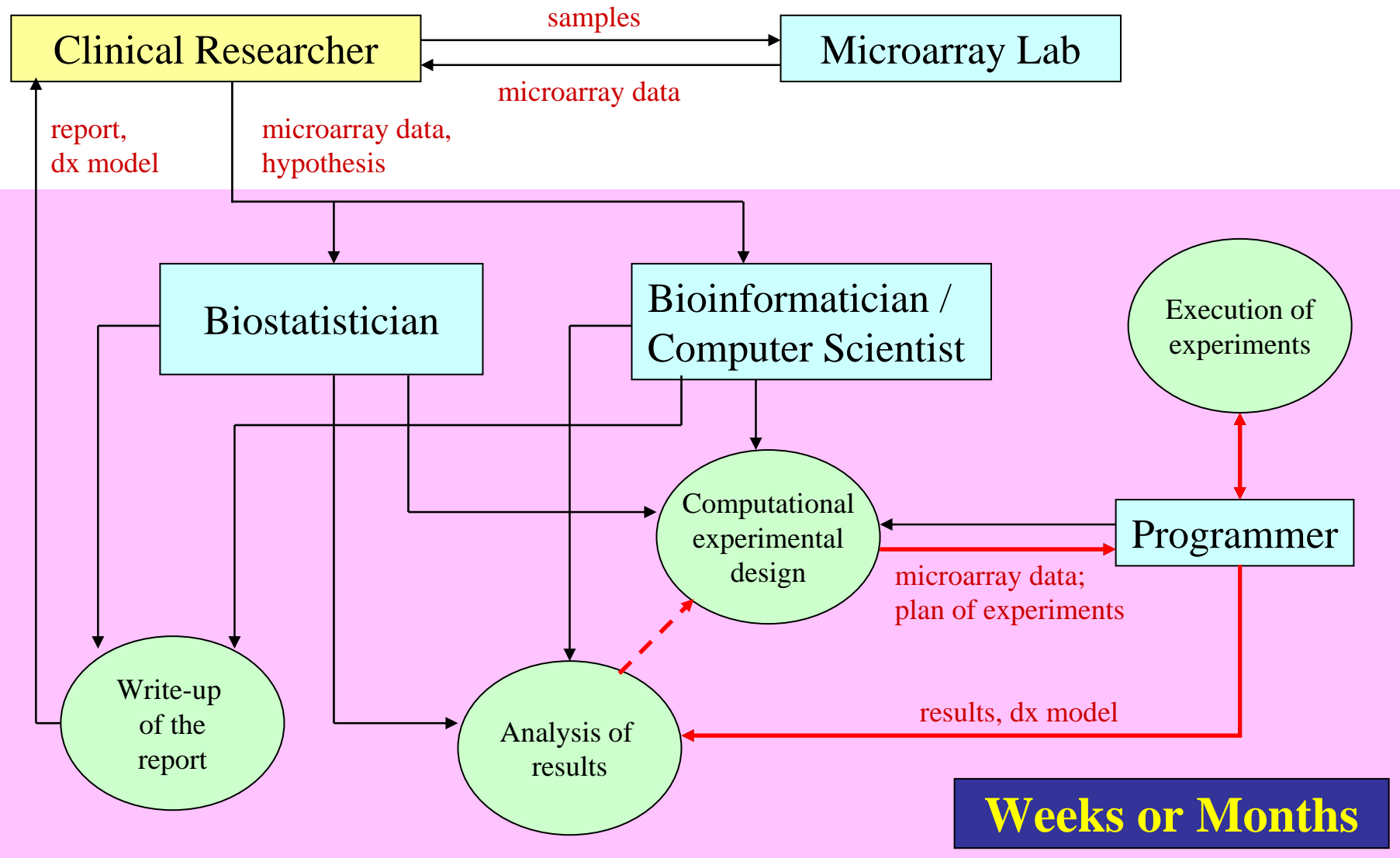


Development of an automated system for high-quality cancer diagnosis from microarray gene expression data

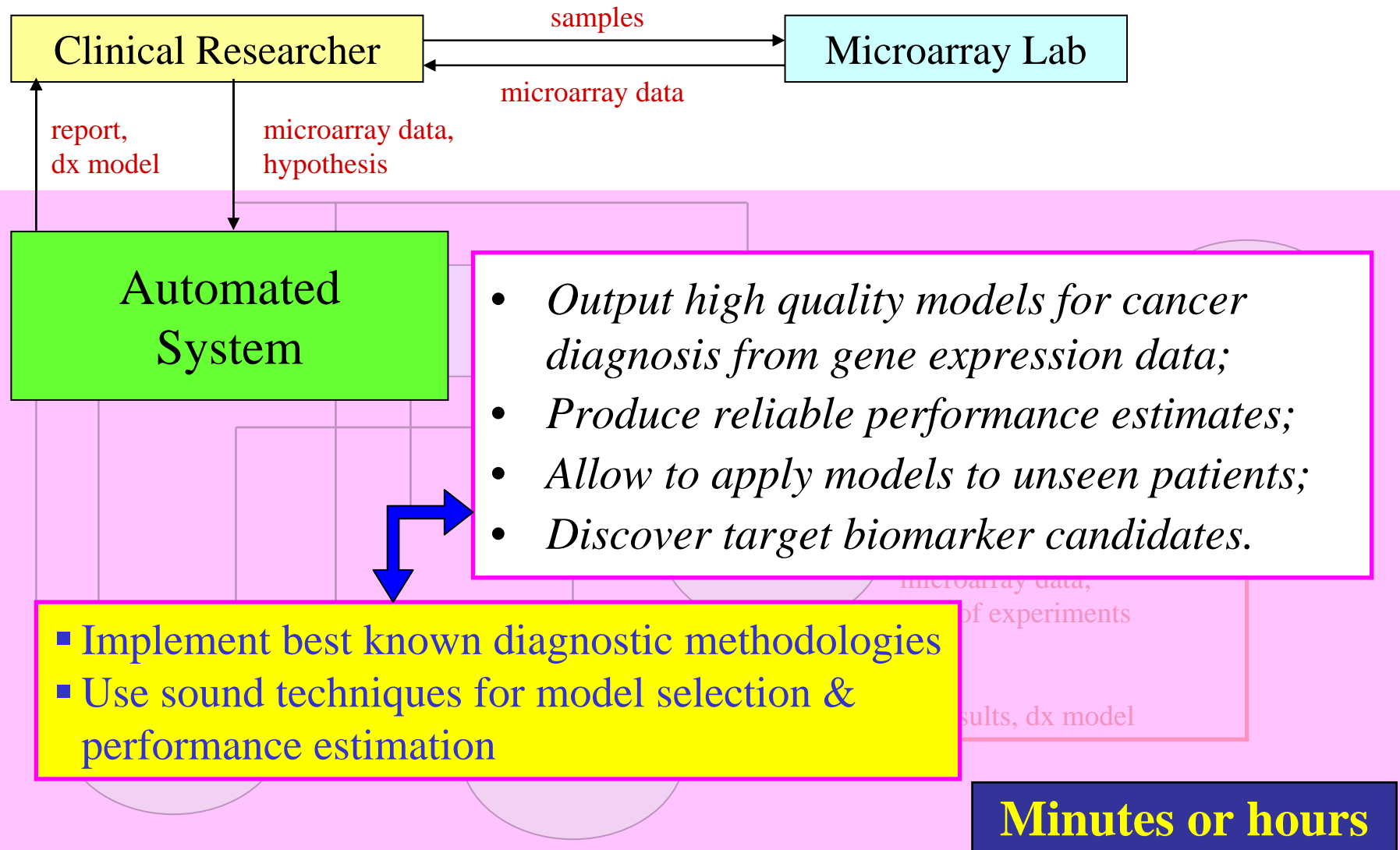
- *How the efforts to develop this system necessitated a large-scale evaluation of learning methods*
- *How evaluation findings informed the resulting system*



Building Cancer Diagnostic Models from Microarray Data



Automated Diagnostic System **GEMS**: Gene **E**xpression **M**odel **S**elector



Prior Research

193 primary studies

2 meta-analyses

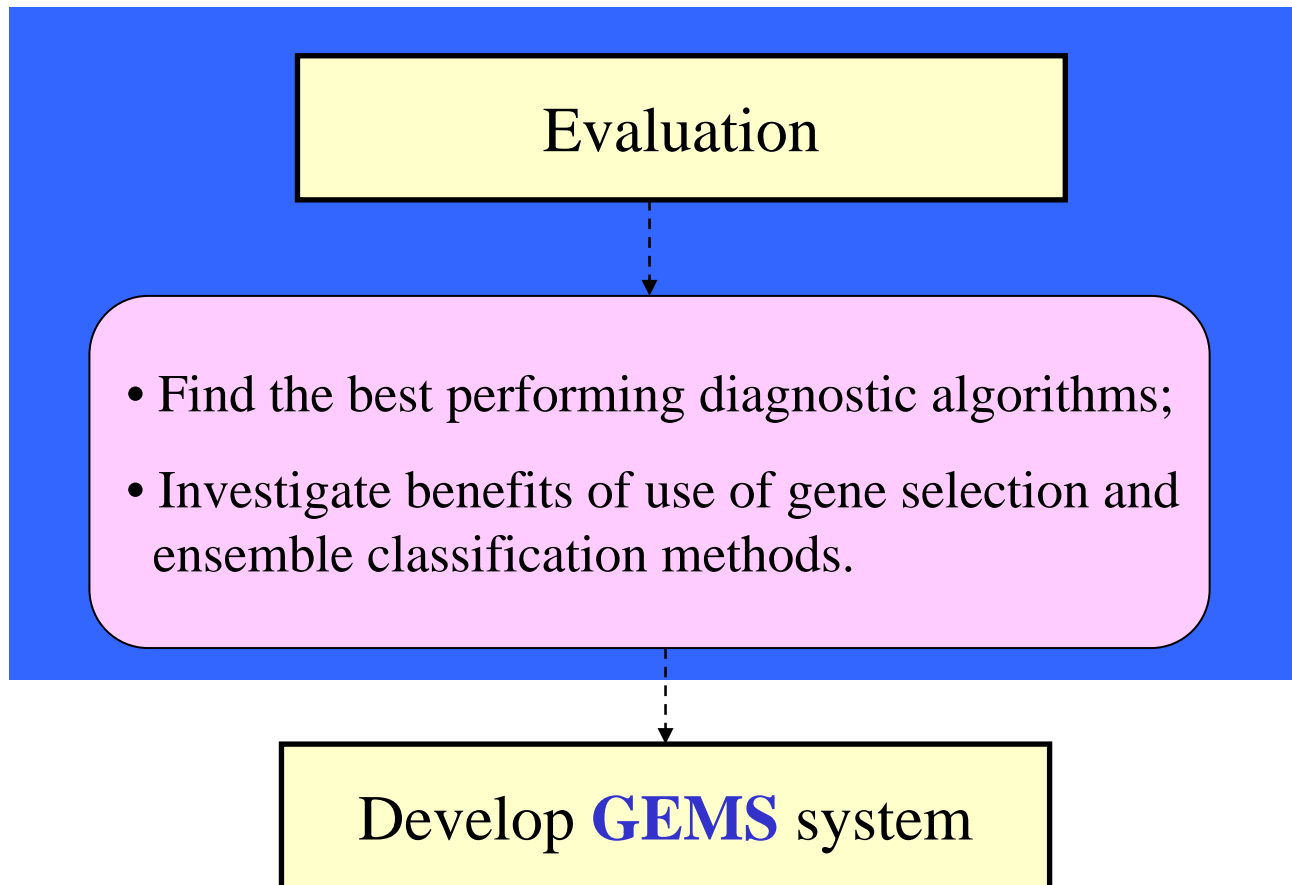
20 software systems

- Limited range of methods & datasets per study
- No description of parameter optimization of learners
- Different experimental designs are employed
- Overfitting [*Ntzani et al.*, Lancet 2003]:
 - 74% no validation
 - 13% incomplete cross-validation
 - 13% implemented cross-validation correctly
- The available meta-analyses are not aimed at identification of best performing methodologies
- Existing systems do not justify choices of diagnostic methods



Cannot specify a small set of best performing diagnostic algorithms;
Have to perform evaluation *de novo*

Step I: Evaluation



Methods & Datasets

Cross-Validation Designs (2)

10-Fold CV

LOOCV

Gene Selection Methods (4)

S2N One-Versus-Rest

S2N One-Versus-One

Non-param. ANOVA

BW ratio

Performance Metrics (2)

Accuracy

RCI

Statistical Comparison

Randomized permutation testing

Classifiers (11)

MC-SVM

One-Versus-Rest

One-Versus-One

DAGSVM

Method by WW

Method by CS

KNN

Backprop. NN

Prob. NN

Decision Trees

WV

One-Versus-Rest

One-Versus-One

Ensemble Classifiers (7)

Based on MC-SVM outputs

Majority Voting

MC-SVM OVR

MC-SVM OVO

MC-SVM DAGSVM

Decision Trees

Based on outputs of all classifiers

Majority Voting

Decision Trees

Gene Expression Datasets (11)

Multicategory Dx

11_Tumors

14_Tumors

9_Tumors

Brain Tumor1

Brain_Tumor2

Leukemia1

Leukemia2

Lung_Cancer

SRBCT

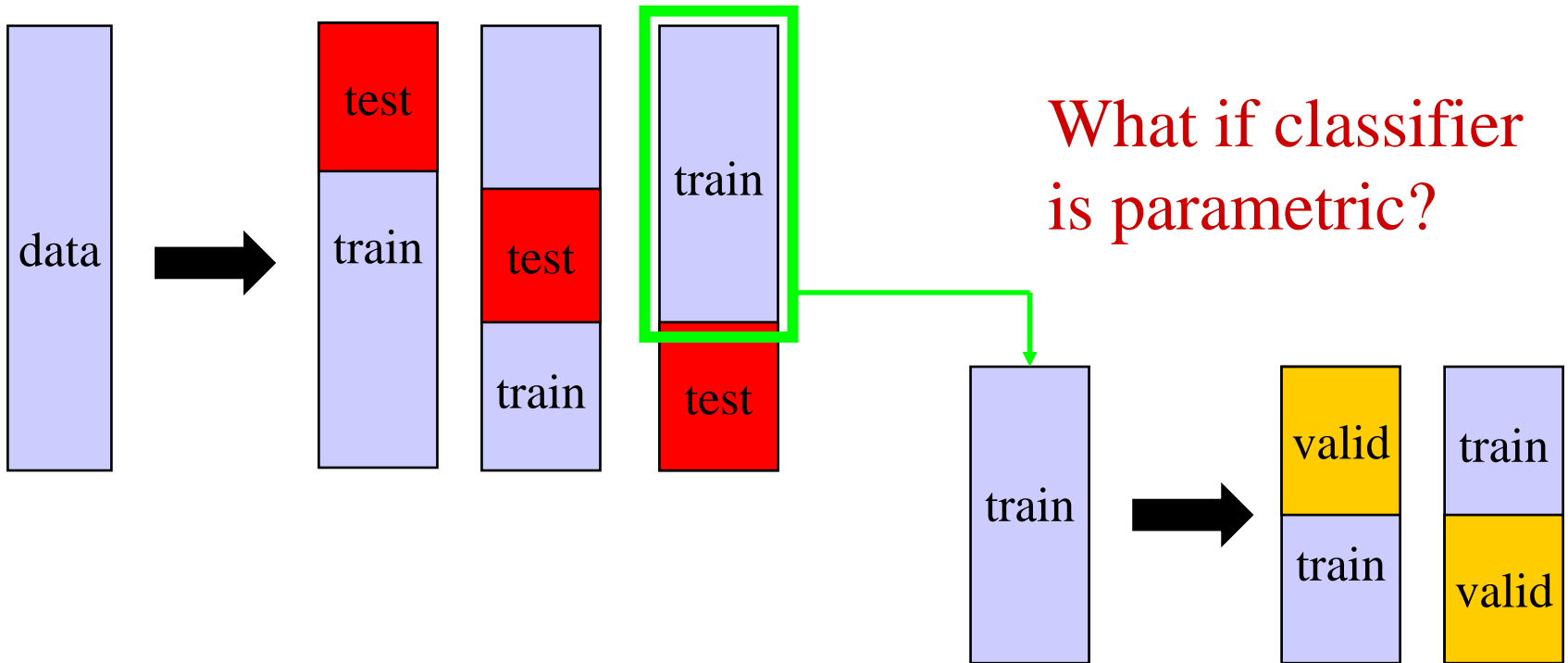
Binary Dx

Prostate_Tumors

DLBCL

Cross-Validation Design

Performance estimation by 10-fold Cross-Validation (**CV**) and Leave-One-Out Cross-Validation (**LOOCV**)



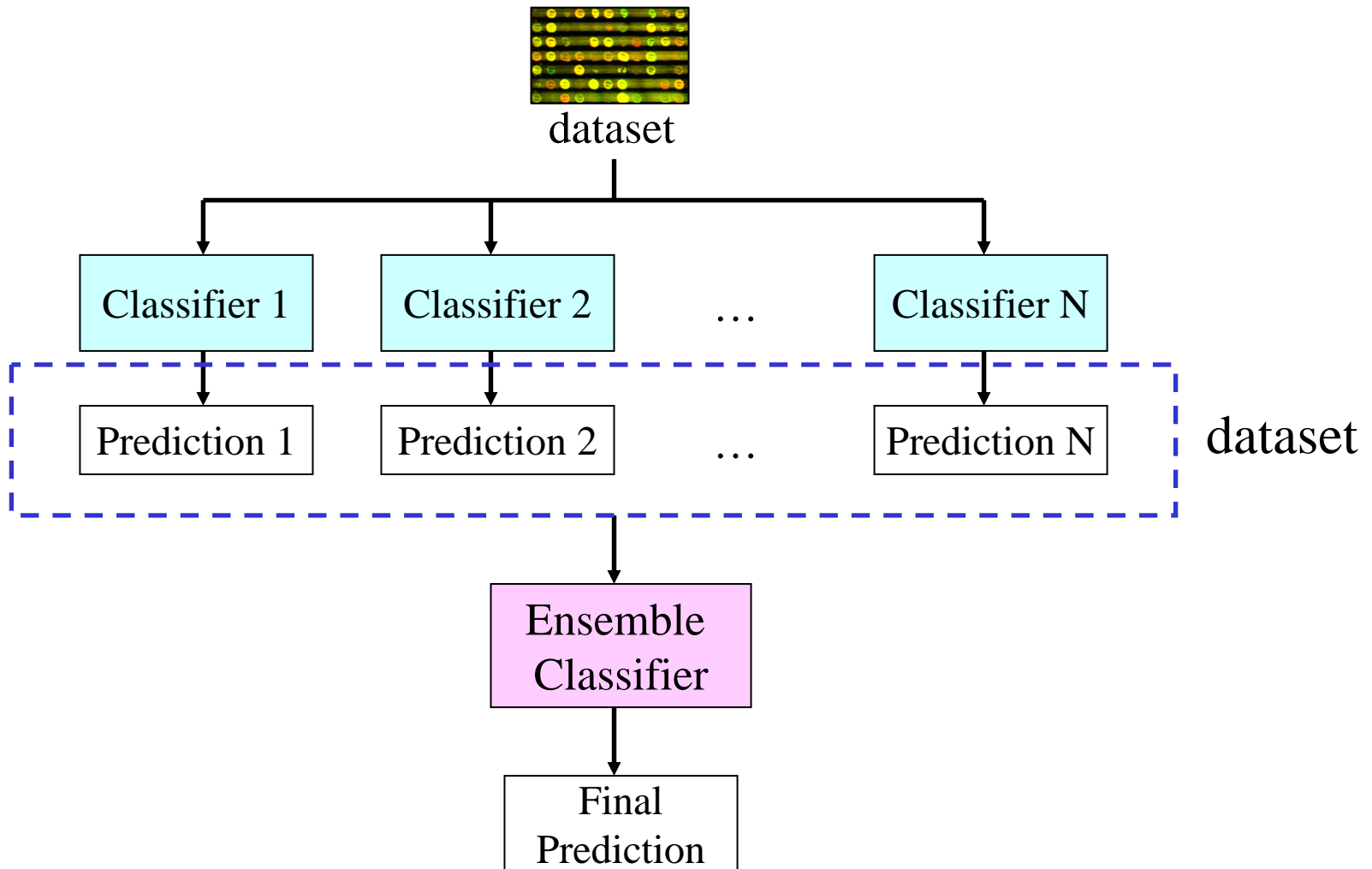
What if classifier is parametric?

Model selection / parameter optimization by “nested” cross-validation

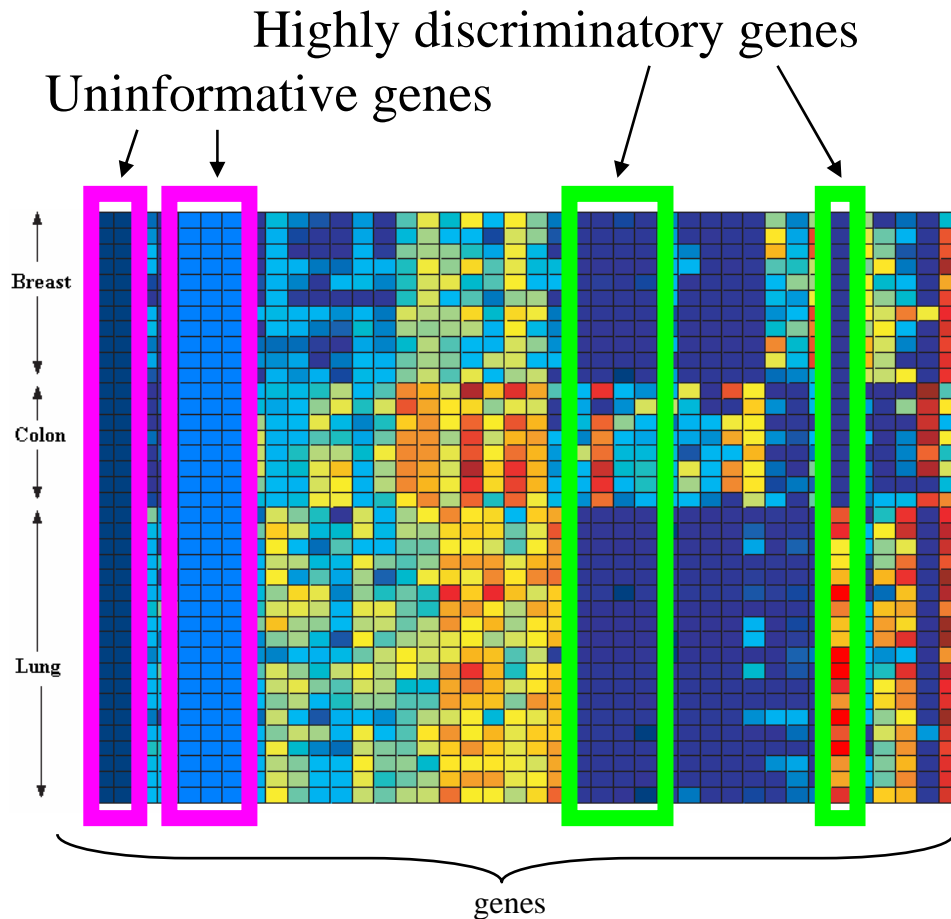
Classifiers

- K-Nearest Neighbors (**KNN**)
 - Backpropagation Neural Networks (**NN**)
 - Probabilistic Neural Networks (**PNN**)
 - Multi-Class SVM: One-Versus-Rest (**OVR**)
 - Multi-Class SVM: One-Versus-One (**OVO**)
 - Multi-Class SVM: **DAGSVM**
 - Multi-Class SVM by Weston & Watkins (**WW**)
 - Multi-Class SVM by Crammer & Singer (**CS**)
 - Weighted Voting: One-Versus-Rest
 - Weighted Voting: One-Versus-One
 - Decision Trees: CART
- instance-based
- neural networks
- kernel-based
- voting
- decision trees

Ensemble Classifiers



Gene Selection Methods



1. Signal-to-noise (**S2N**) ratio in one-versus-rest (OVR) fashion;
2. Signal-to-noise (**S2N**) ratio in one-versus-one (OVO) fashion;
3. Kruskal-Wallis nonparametric one-way ANOVA (**KW**);
4. Ratio of genes between-categories to within-category sum of squares (**BW**).

Performance Metrics & Statistical Comparison

1. Accuracy

- + can compare to previous studies
- + easy to interpret & simplifies statistical comparison

2. Relative classifier information (RCI)

- + easy to interpret & simplifies statistical comparison
- + not sensitive to distribution of classes
- + accounts for difficulty of a decision problem

- Randomized permutation testing to compare accuracies of the classifiers ($\alpha=0.05$)

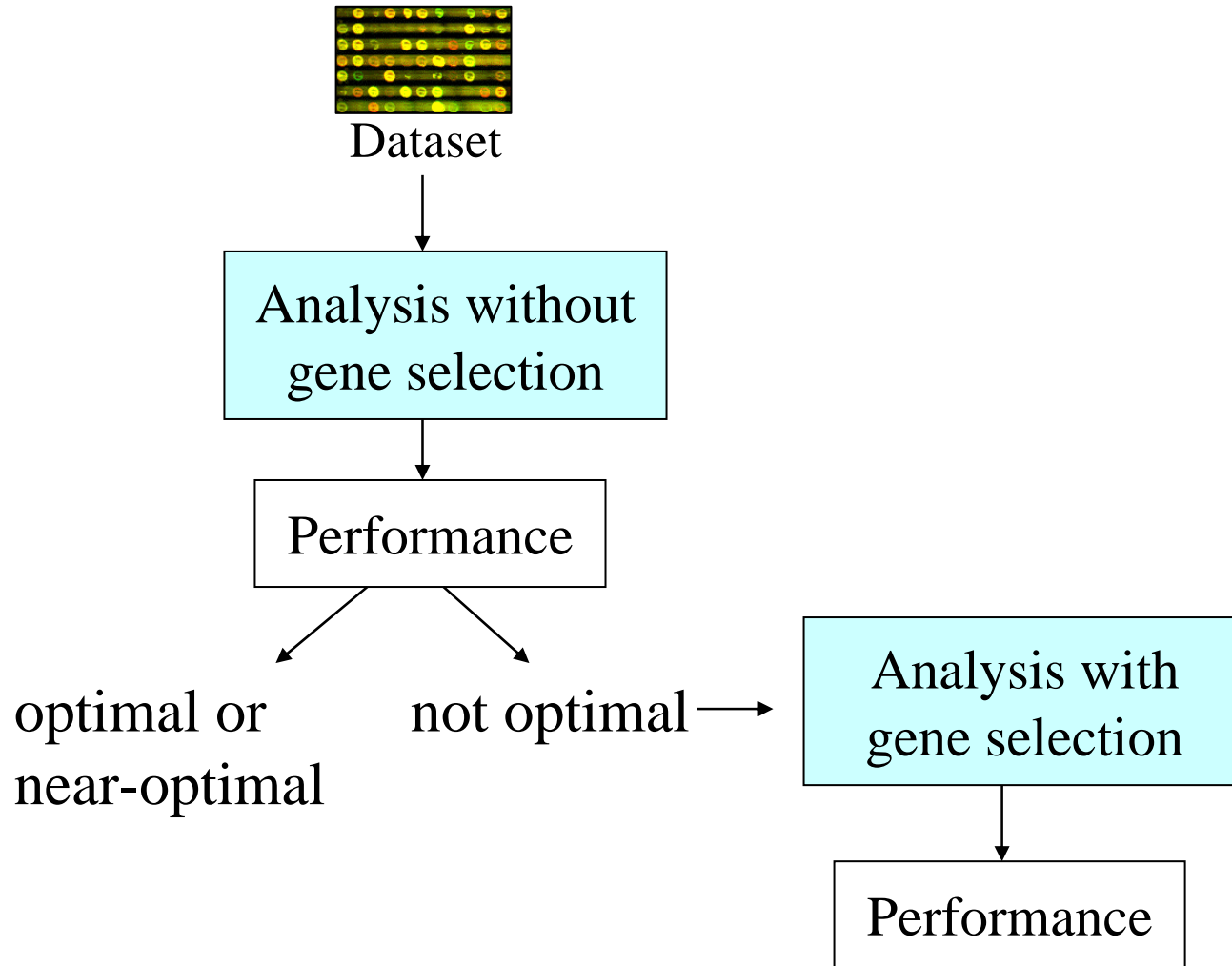
Microarray Datasets

Dataset name	Number of			Reference
	Sam- ples	Variables (genes)	Cate- gories	
<i>11_Tumors</i>	174	12533	11	Su, 2001
<i>14_Tumors</i>	308	15009	26	Ramaswamy, 2001
<i>9_Tumors</i>	60	5726	9	Staunton, 2001
<i>Brain_Tumor1</i>	90	5920	5	Pomeroy, 2002
<i>Brain_Tumor2</i>	50	10367	4	Nutt, 2003
<i>Leukemia1</i>	72	5327	3	Golub, 1999
<i>Leukemia2</i>	72	11225	3	Armstrong, 2002
<i>Lung_Cancer</i>	203	12600	5	Bhattacharjee, 2001
<i>SRBCT</i>	83	2308	4	Khan, 2001
<i>Prostate_Tumor</i>	102	10509	2	Singh, 2002
<i>DLBCL</i>	77	5469	2	Shipp, 2002

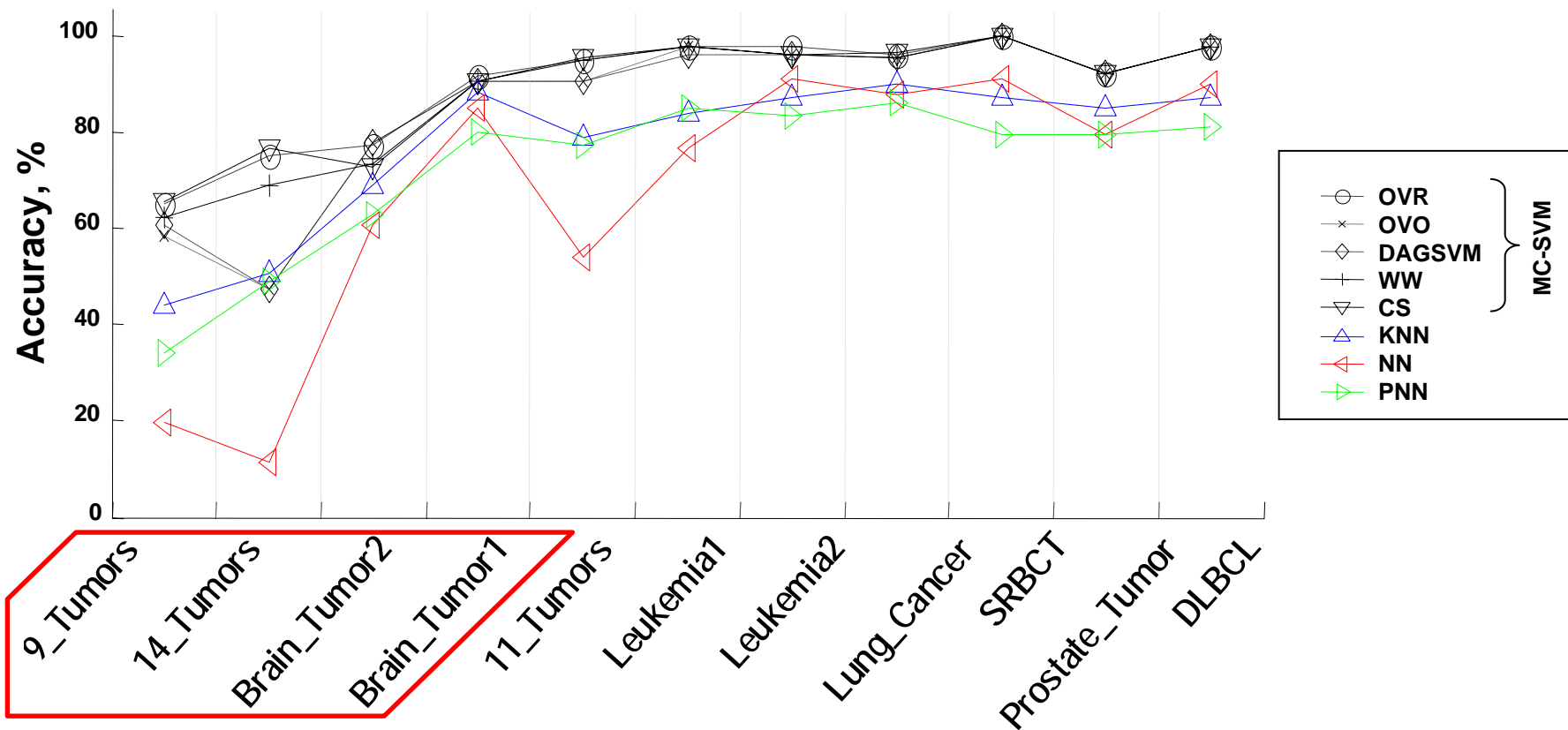
Total:

- ~1300 samples
- 74 diagnostic categories
- 41 cancer types and 12 normal tissue types

Staged Experimental Design

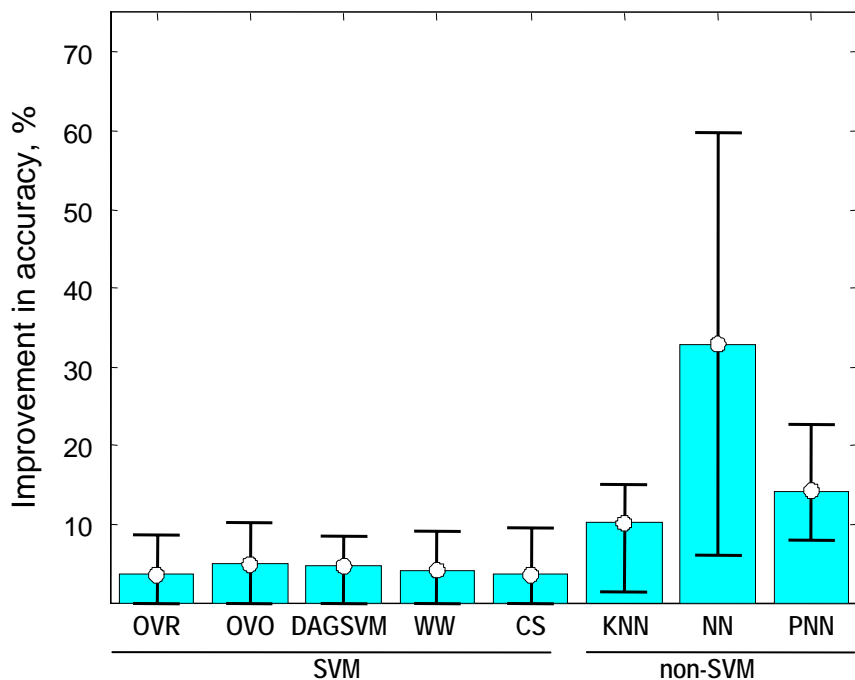


Results Without Gene Selection

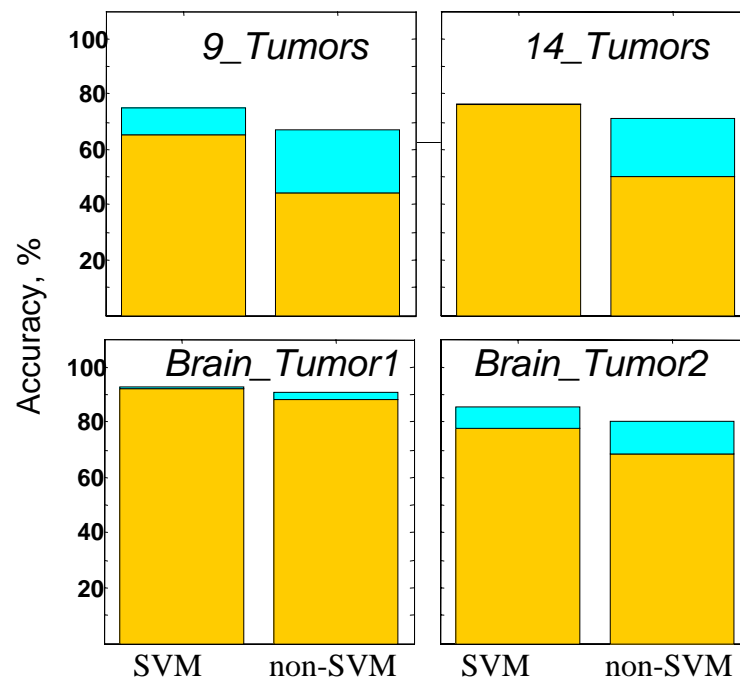


Results With Gene Selection

Improvement of diagnostic performance by gene selection
(averages for the four datasets)

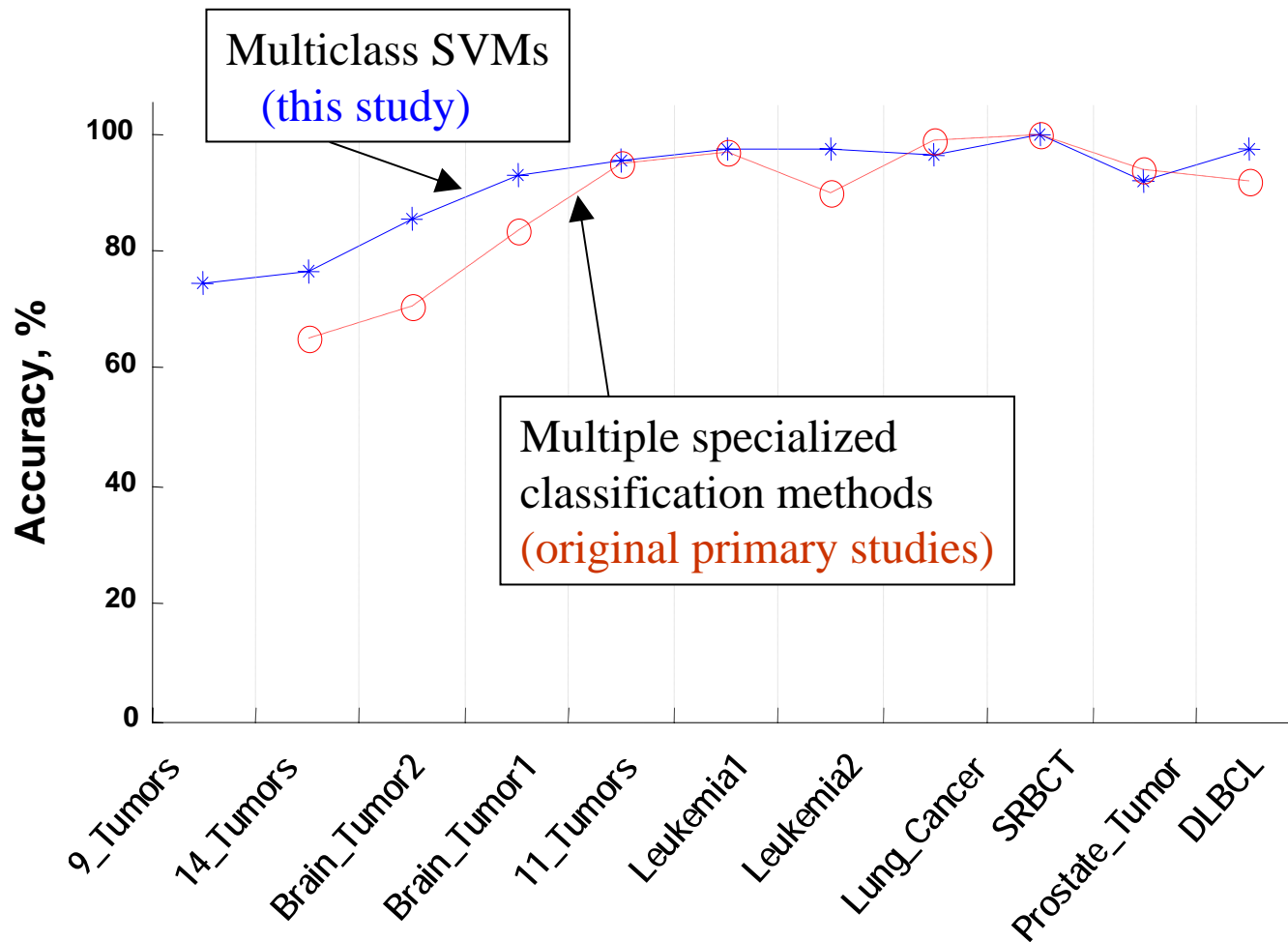


Diagnostic performance before and after gene selection



Average reduction of genes is 10-30 times

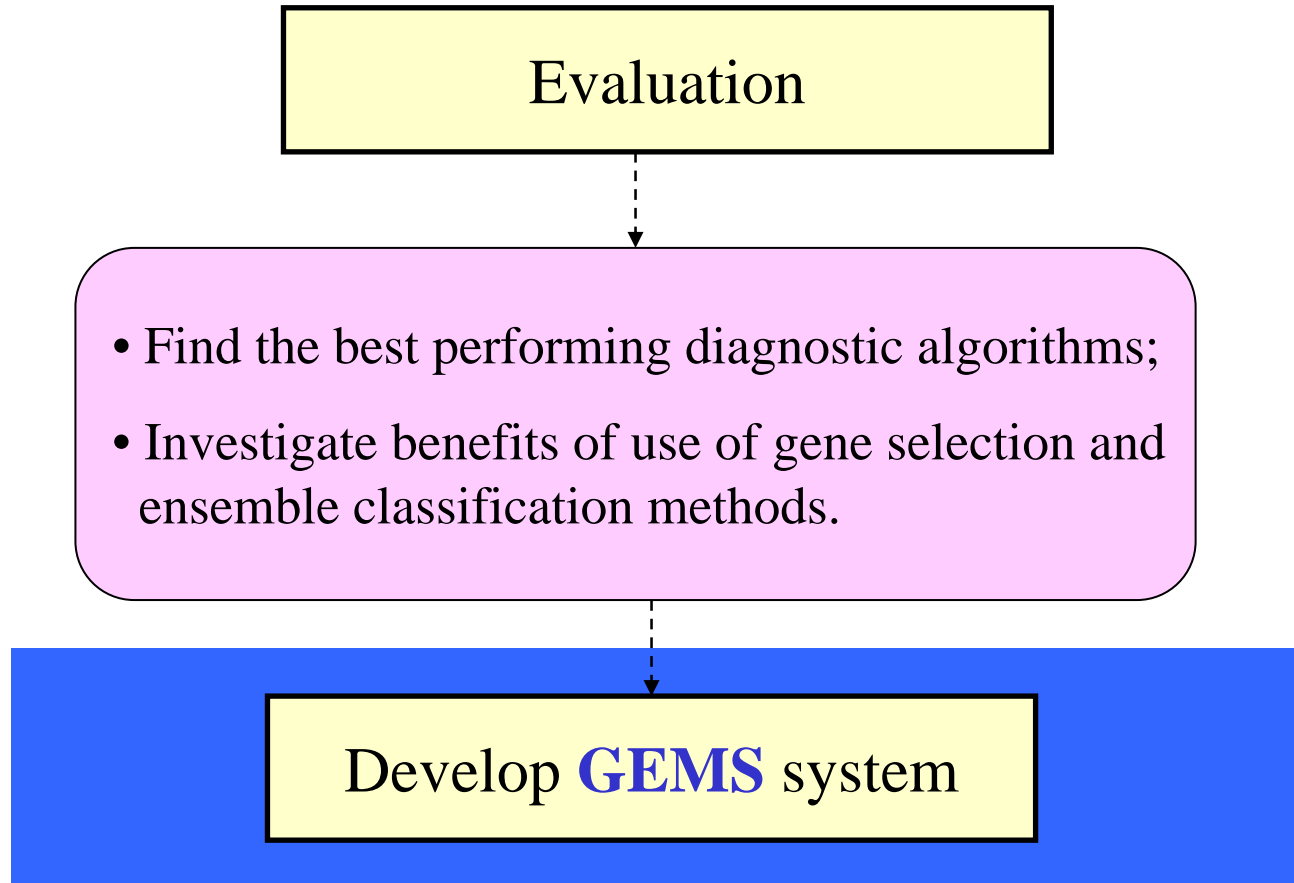
Comparison with Previously Published Results



Conclusions of Evaluation Stage

- Multi-class SVMs are the best family among the tested algorithms outperforming KNN, NN, PNN, DT, and WV.
- Gene selection in some cases improves classification performance of all classifiers, especially of non-SVM algorithms;
- Ensemble classification does not improve performance;
- Obtained results favorably compare with literature.

Step II: System Development Informed by Evaluation



Methods Implemented in GEMS

Cross-Validation Designs

N-Fold CV

LOOCV

Normalization Techniques

[a, b]

$(x - \text{MEAN}(x)) / \text{STD}(x)$

$x / \text{STD}(x)$

$x / \text{MEAN}(x)$

$x / \text{MEDIAN}(x)$

$x / \text{NORM}(x)$

$x - \text{MEAN}(x)$

$x - \text{MEDIAN}(x)$

$\text{ABS}(x)$

$x + \text{ABS}(x)$

Classifiers

One-Versus-Rest

One-Versus-One

DAGSVM

Method by WW

Method by CS

MC-SVM

Performance Metrics

Accuracy

RCI

AUC ROC

Gene Selection Methods

S2N One-Versus-Rest

S2N One-Versus-One

Non-param. ANOVA

BW ratio

HITON_MB

HITON_PC

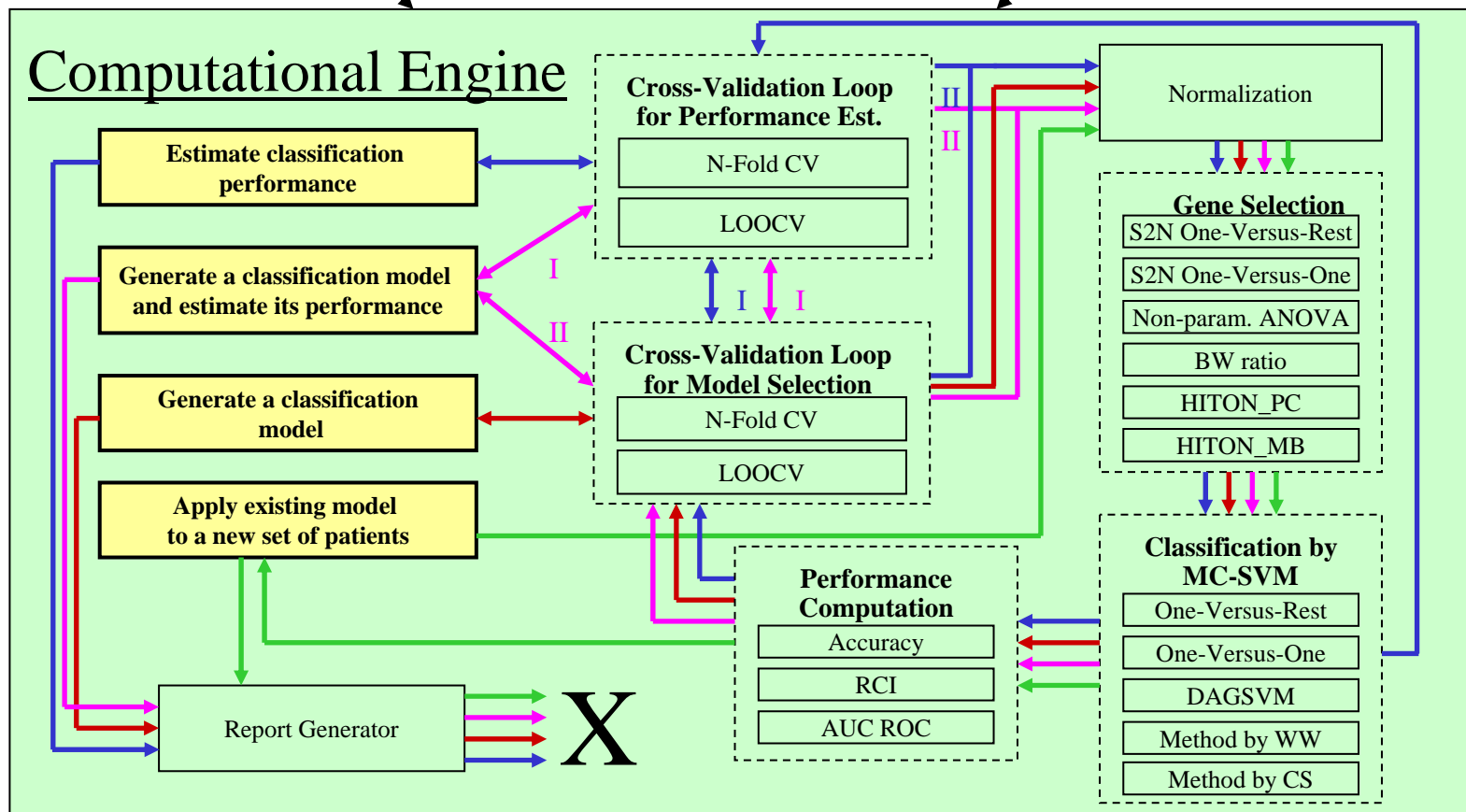
Software Architecture of GEMS

GEMS 1.0

GEMS 2.0

Power-User Interface

Wizard-Like User Interface



GEMS 1.0: Power-User Interface

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: OVR OVO DAGSVM WW CS

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

A. $\log(x)$, logarithm base:

B. $[a, b]$, a : and b :

C. $(x - \text{mean of } x) / \text{std of } x$

D. $x / \text{std of } x$

E. $x / \text{mean of } x$

F. $x / \text{median of } x$

G. $x / \text{norm of } x$

H. $x - \text{mean}(x)$

I. $x - \text{median}(x)$

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:

GEMS 2.0: Wizard-Like Interface

The screenshot displays the GEMS 2.0 software interface. The main window is titled 'GEMS' and has a menu bar with 'File' and 'Help'. The current step is 'Step 4/9: Classification algorithm(s)'. The interface is divided into two main panels: a configuration panel on the left and a 'Project Summary' panel on the right.

Configuration Panel:

- SVM classification algorithms (select at least one):**
 - One-versus-rest (OVR)
 - One-versus-one (OVO)
 - DAGSVM
 - Method by Weston and Watkins (WW)
 - Method by Crammer and Singer (CS)
- Select kernel for SVM algorithms:**
 - Polynomial kernel (including linear)
 - Gaussian kernel (RBF)
- SVM parameters:**
 - No need to optimize, use the following default values:
 - Cost: Degree of polynomial:
 - Optimize parameters by cross-validation. Search the following grid:
 - Cost: to with multiplicative step
 - Degree: to with step

Project Summary Panel:

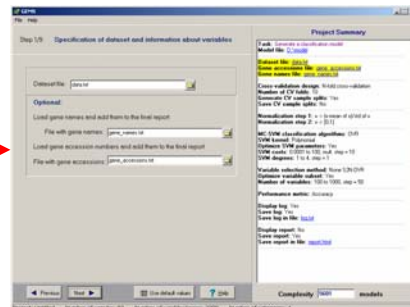
- Task: Generate a classification model**
 - Model file: [model.mod](#)
- Dataset Specification**
 - Dataset file: [D:\Sasha\Delphi\GEMS\Distributive\Data\data.txt](#)
- Cross-Validation Design**
 - Cross-validation design: [N-fold cross-validation](#)
 - Number of CV folds: [10](#)
 - Generate CV sample splits: [Yes](#)
 - Save CV sample splits: [Yes](#)
 - Filename for saving CV sample splits: [splits.txt](#)
- Normalization**
 - Normalization method: [x -> \[0,1\]](#)
- Classification**
 - MIC-SVM classification algorithms: [OVR OVO DAGSVM](#)
 - SVM kernel: [Polynomial](#)
 - Optimize SVM parameters: [Yes](#)
 - SVM costs: [0.0001 to 100, mult. step = 10](#)
 - SVM degrees: [1 to 4, step = 1](#)
- Variable selection**
 - Variable selection method: [None S2N_OVR S2N_OVO HITON_PC HITON_MB](#)
 - Optimize variable subset: [Yes](#)
 - Number of variables: [100 to 1000, step = 50](#)
 - Optimize threshold: [Yes](#)
 - Threshold: [0.01 to 0.05, step = 0.01](#)
- Performance Metric**
 - Performance metric: [Accuracy](#)

At the bottom of the configuration panel, there are navigation buttons: 'Previous', 'Next', 'Use default values', and 'Help'. The 'Project Summary' panel has a 'Complexity' field showing '41161 models'. The status bar at the bottom indicates: 'Project: Untitled', 'Number of samples: 83', 'Number of variables/genes: 2309', and 'Number of categories: 4'.

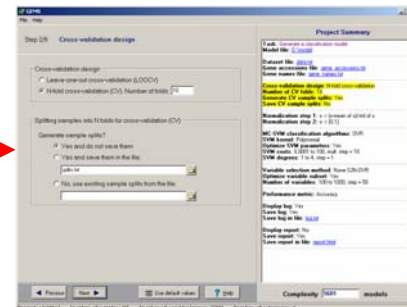
GEMS 2.0: Wizard-Like Interface



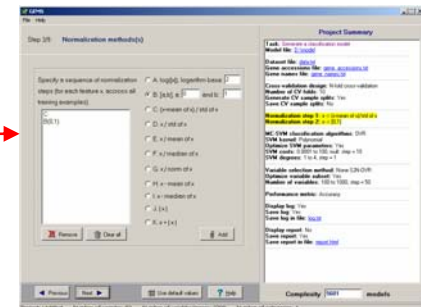
Task selection



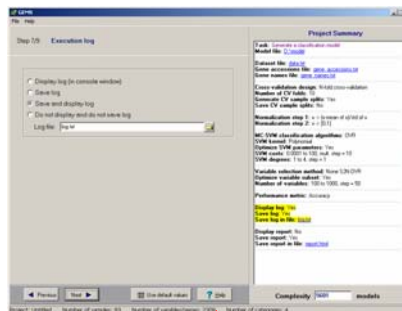
Dataset specification



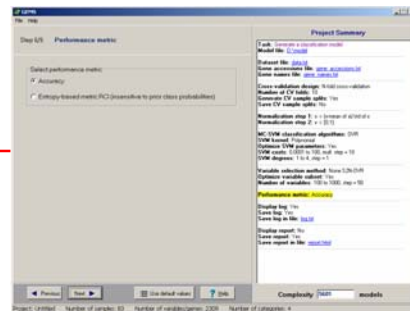
Cross-validation design



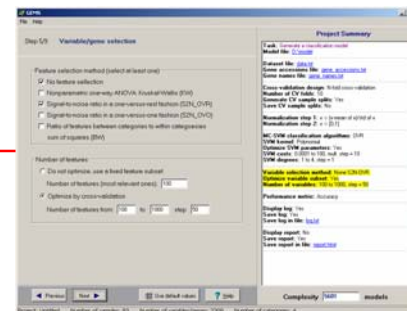
Normalization



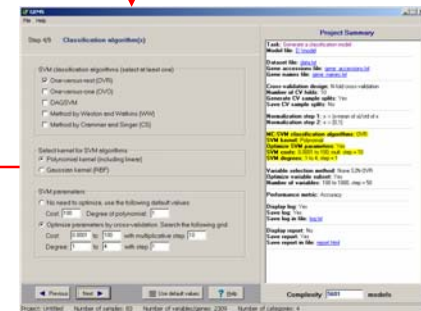
Logging



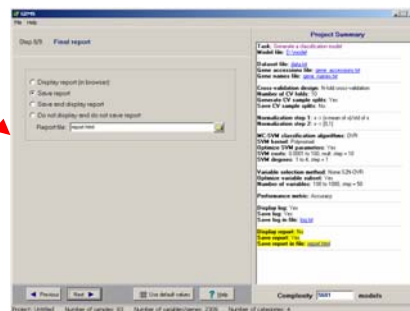
Performance metric



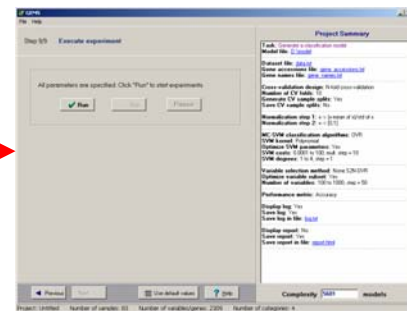
Gene selection



Classification



Report generation



Analysis execution

GEMS 2.0: Wizard-Like Interface

Input microarray
gene expression dataset

File with
gene names

File with gene
accession numbers

Output model

Address: D:\Sasha\Matlab\MC-SVM\Toolbox_Development\distributive\report.htm

GEMS : 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	AF05026
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

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

NCBI Nucleotide

Search Nucleotide for U81561

Display Summary Show: 20 Send to Text

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

Additional Validation of GEMS (post implementation)

	GEMS	Published results	Reference
<i>6_Tumors</i>	97.2%	96.0%	Shedden, 2003
<i>Leukemia3</i>	98.4%	98.4%	Yeoh, 2002
<i>Lung_Cancer2</i>	100%	100%	Beer, 2002

- Analysis in GEMS: 15-30 minutes per dataset
- GEMS is performing as good or better than published studies

First reported use of GEMS: *Classification and Biomarker Discovery from Head & Neck Cancer Microarray Data* (presented at 6th ICHNC)

Limitations & Ongoing Work

Molecular medicine is moving very fast and GEMS is evolving accordingly:

- Periodical literature review to ensure that GEMS implements the best methodologies & algorithms
- Large-scale evaluation of GEMS with various user types
- Evaluation of computational causal discovery capabilities
- New functionality for prognosis & response to treatment

Final Conclusions

- We created a system to support automatic development of high-quality cancer classification models & gene selection for biomarker discovery from gene expression data
- An extensive study of several methods in this domain informed system development
- The evaluation revealed the importance of multi-category SVM classifiers in this domain and task
- Results obtained by GEMS are as good or better than previously published results by human experts and require only a few minutes of user's time to complete
- GEMS is available for download from <http://www.gems-system.org>

For more details, see paper in proceedings, online supplement, and:

A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, S. Levy. “[A Comprehensive Evaluation of Multicategory Classification Methods for Microarray Gene Expression Cancer Diagnosis](#)”, (to appear in *Bioinformatics*)

Acknowledgements

- Academic advisors:
 - Dr. Constantin Aliferis
 - Dr. Ioannis Tsamardinos
- Other members of my committee:
 - Dr. Shawn Levy
 - Dr. Douglas Hardin

<http://www.gems-system.org>