

Master's Thesis Defense:

# Automatic Cancer Diagnostic Decision Support System for Gene Expression Domain

**Alexander Statnikov**

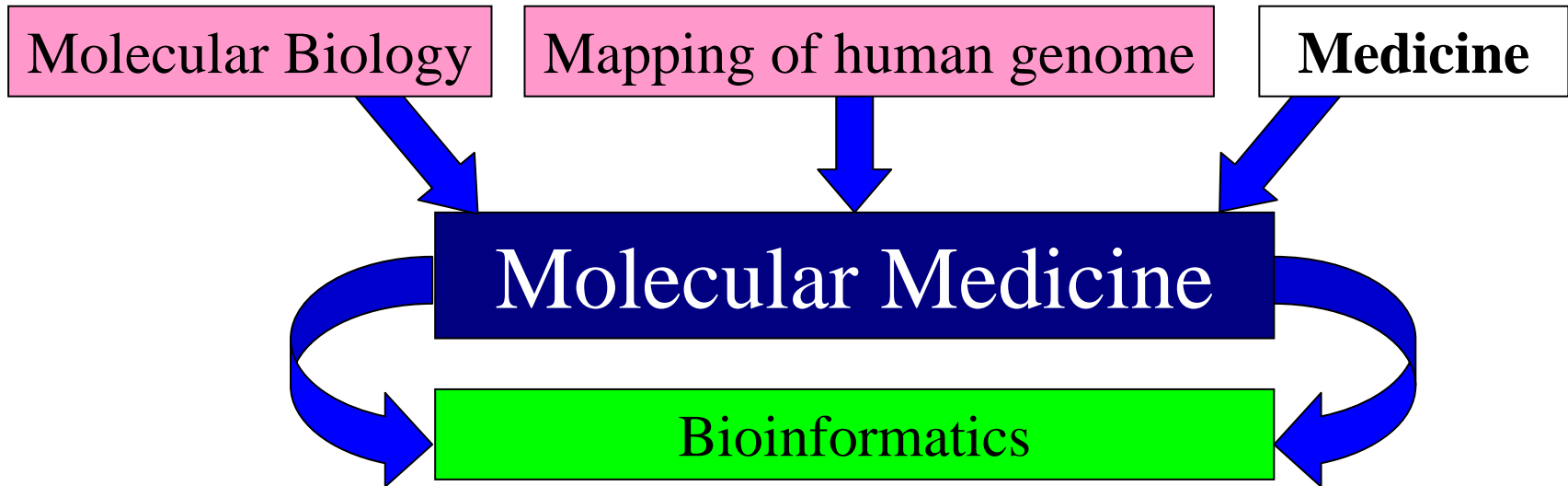
Committee Members:

- Dr. Constantin F. Aliferis (advisor)
- Dr. Douglas P. Hardin
- Dr. Shawn Levy
- Dr. Ioannis Tsamardinos (advisor)

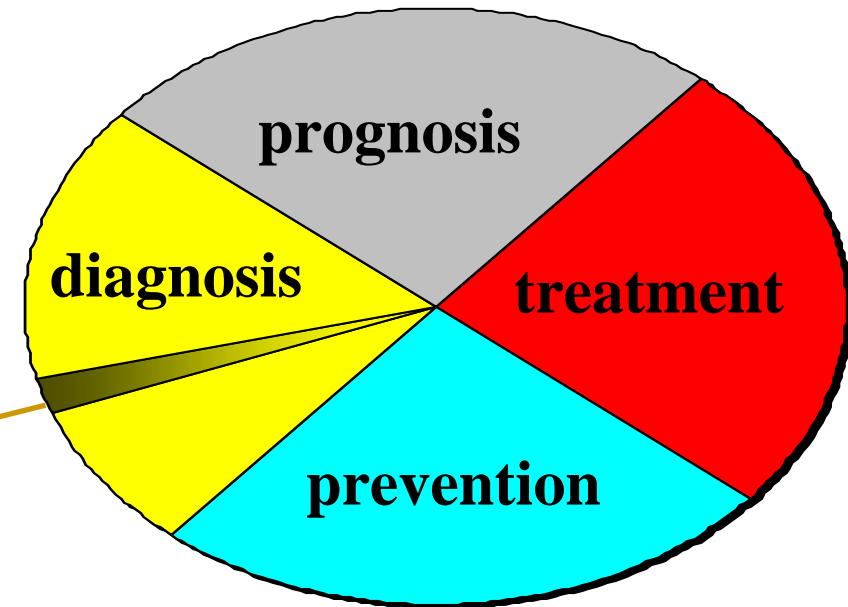
Discovery Systems Laboratory, Department of Biomedical Informatics,  
Vanderbilt University, Nashville, TN, USA

07/06/2005

# Problem

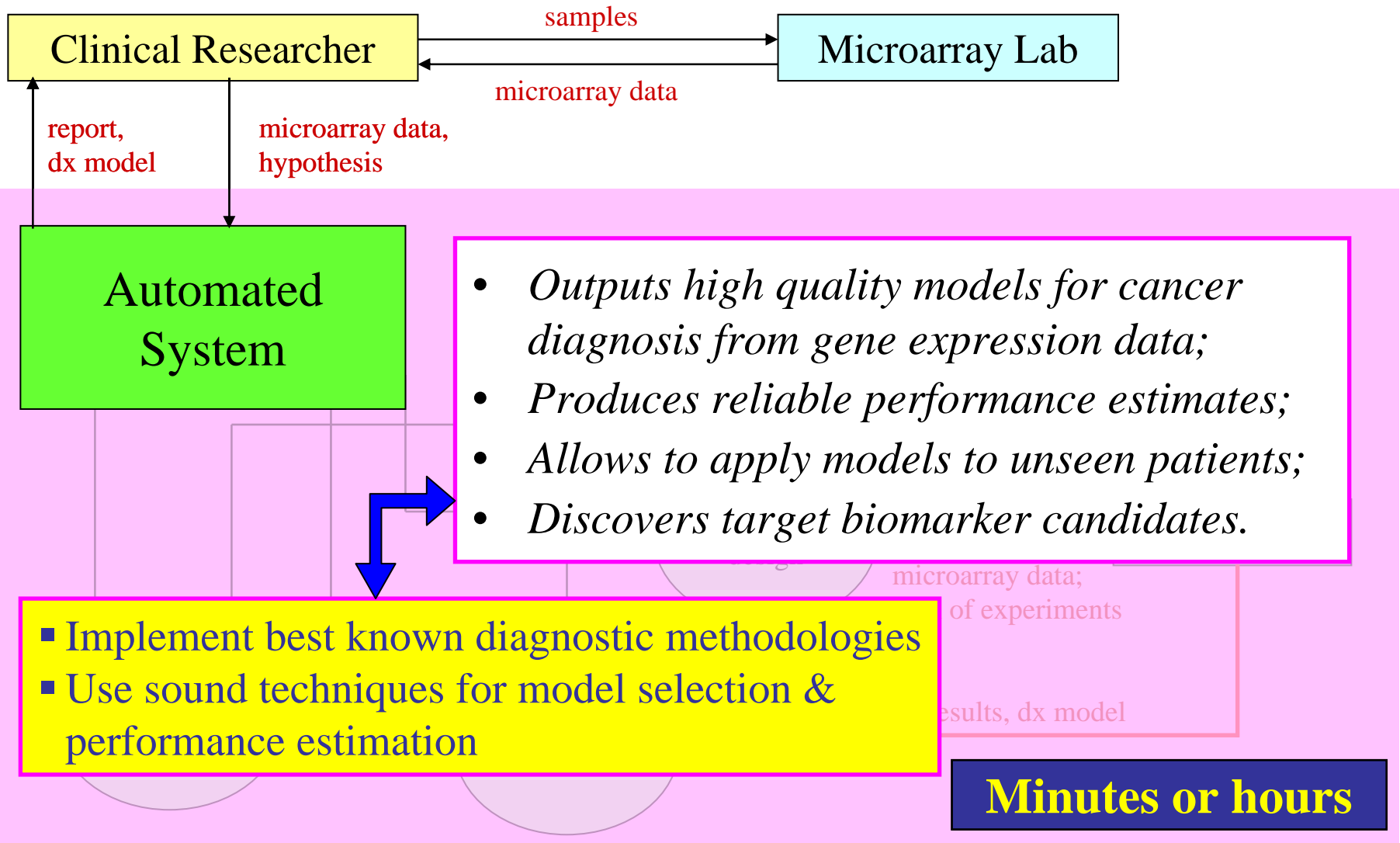


An automated system for development and evaluation of high-quality cancer diagnostic models and biomarker discovery from microarray gene expression data





# Automated Diagnostic System **GEMS**: Gene Expression Model Selector



# Other Systems for Supervised Analysis of Microarray Data

Name	Version	Developer	Automatic model selection for classifier and gene selection methods
<i>ArrayMiner ClassMarker</i>	5.2	Optimal Design, Belgium	No
<i>Avadis Prophetic</i>	3.3	Strand Genomics, USA	No
<i>BRB ArrayTools</i>	3.2 Beta	National Cancer Institute, USA	No
<i>caGEDA</i>	(accessed 10/2004)	University of Pittsburgh and University of Pittsburgh Medical Center, USA	No
<i>Cleaver</i>	1.0	Stanford University, USA	No
<i>GeneCluster2</i>	2.1.7	Broad Institute, Massachusetts Institute of Technology, USA	No
<i>GeneLinker Platinum</i>	4.1	Protein Data Bank, Switzerland	No
<i>GeneMaths XT</i>	1.2	Applied Maths, Belgium	No
<i>GenePattern</i>	1.2.4	Massachusetts Institute of Technology, USA	No
<i>Genesis</i>	1.5.0	Graz University of Technology, Austria	No
<i>GeneSpring</i>	7	Silicon Genetics, USA	No
<i>GEPAS</i>	1.1	National Center for Cancer Research (CNIO), Spain	Limited (number of genes)
<i>MultiExperiment Viewer</i>	1.0	Stanford University, USA	No
<i>PAM</i>	1.21a	Stanford University, USA	No (for a single parameter of the classifier)
<i>Partek Predict</i>	6.0	Partek, USA	Limited (does not allow optimization of the choice of gene selection algorithms)
<i>Weka Explorer</i>	3.4.3	University of Waikato, New Zealand	No

**There exist many good software packages for supervised analysis of microarray data, but...**

- **Neither system provides a protocol for data analysis that precludes overfitting.**

- **A typical software either offers an overabundance of algorithms or algorithms with unknown performance. Thus is it not clear to the user how to choose an optimal algorithm for a given data analysis task.**

- **The software packages address needs of experienced analysts. However, there is a need to use this software (and still achieve good results) by users who know little about data analysis (e.g., biologists and clinicians).**

# What Does Prior Research Suggest About the Best Performing Methods?

193 primary studies

2 meta-analyses

- 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



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

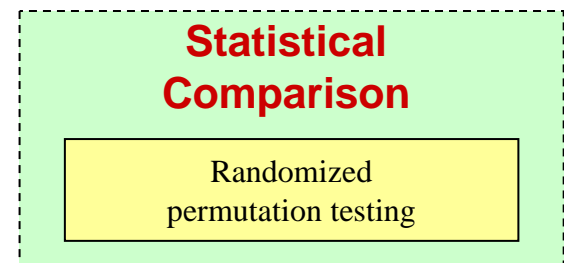
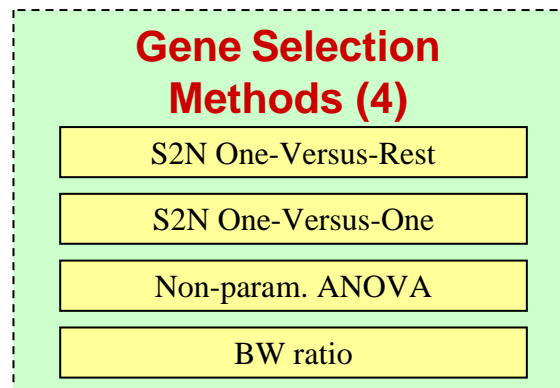
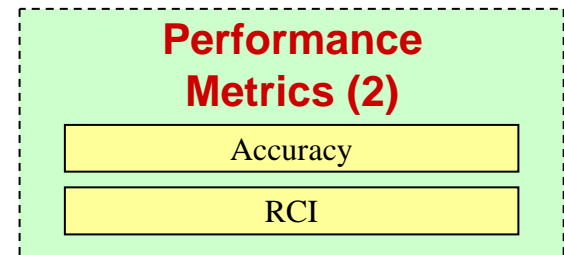
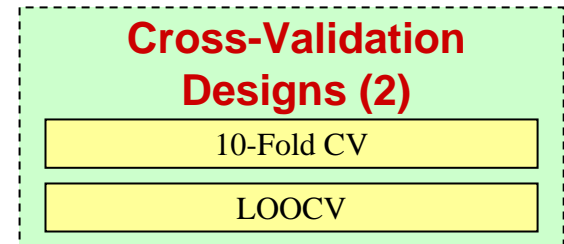
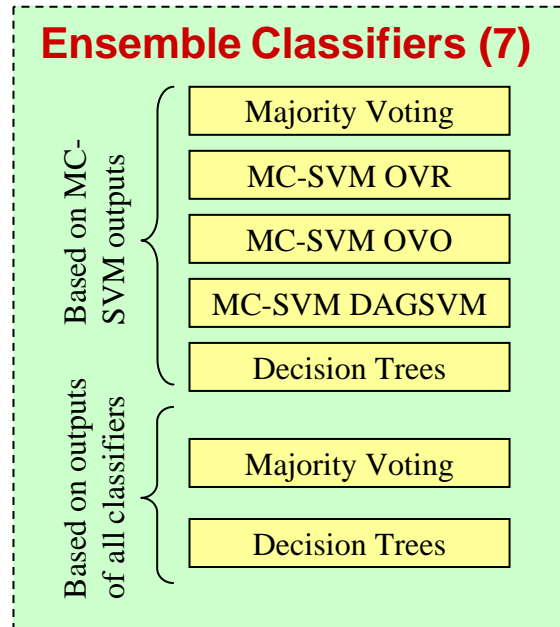
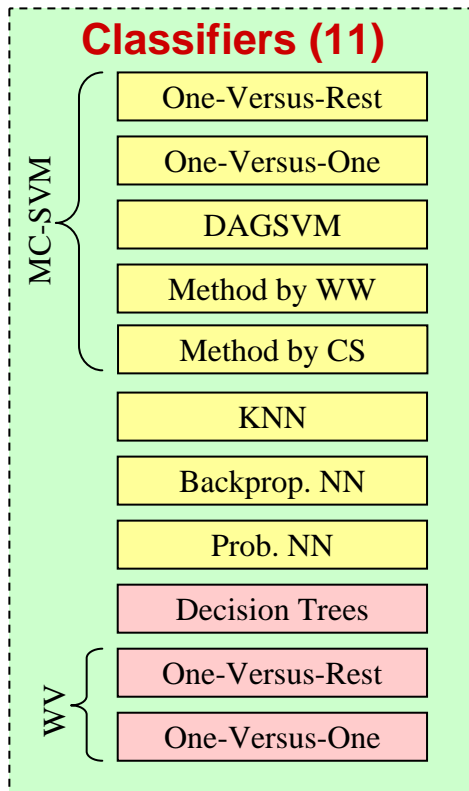
# Algorithmic Evaluations to Inform Development of the System

# 1<sup>st</sup> Algorithmic Evaluation Study

**Main Goal:** *Investigate which ones among the many powerful classifiers currently available for gene expression diagnosis perform the best across many datasets and cancer types.*

Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. *A comprehensive evaluation of multiclassification methods for microarray gene expression cancer diagnosis.* Bioinformatics, 2005, 21: 631-643.

# Methods at a Glance



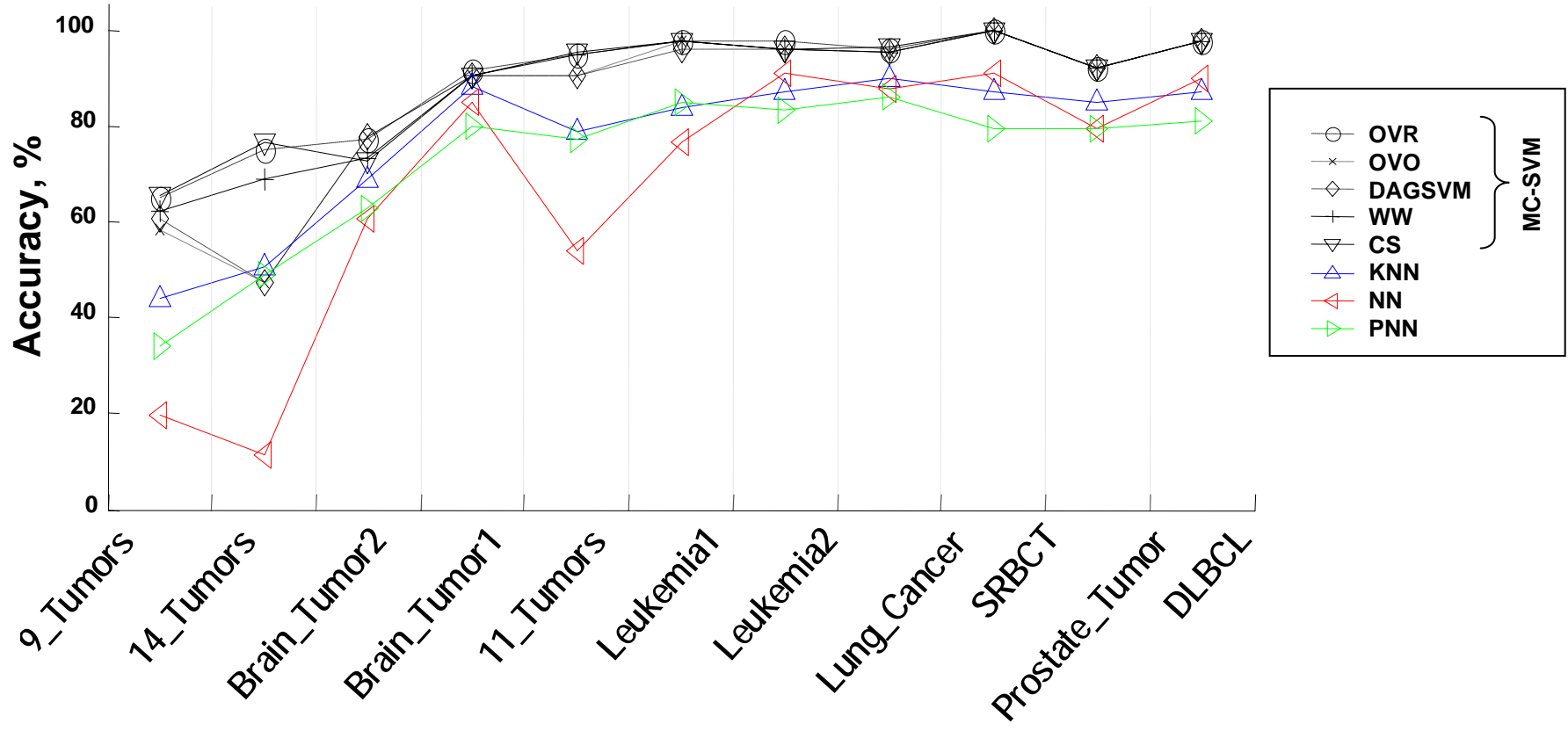
# Datasets Used in Evaluation

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

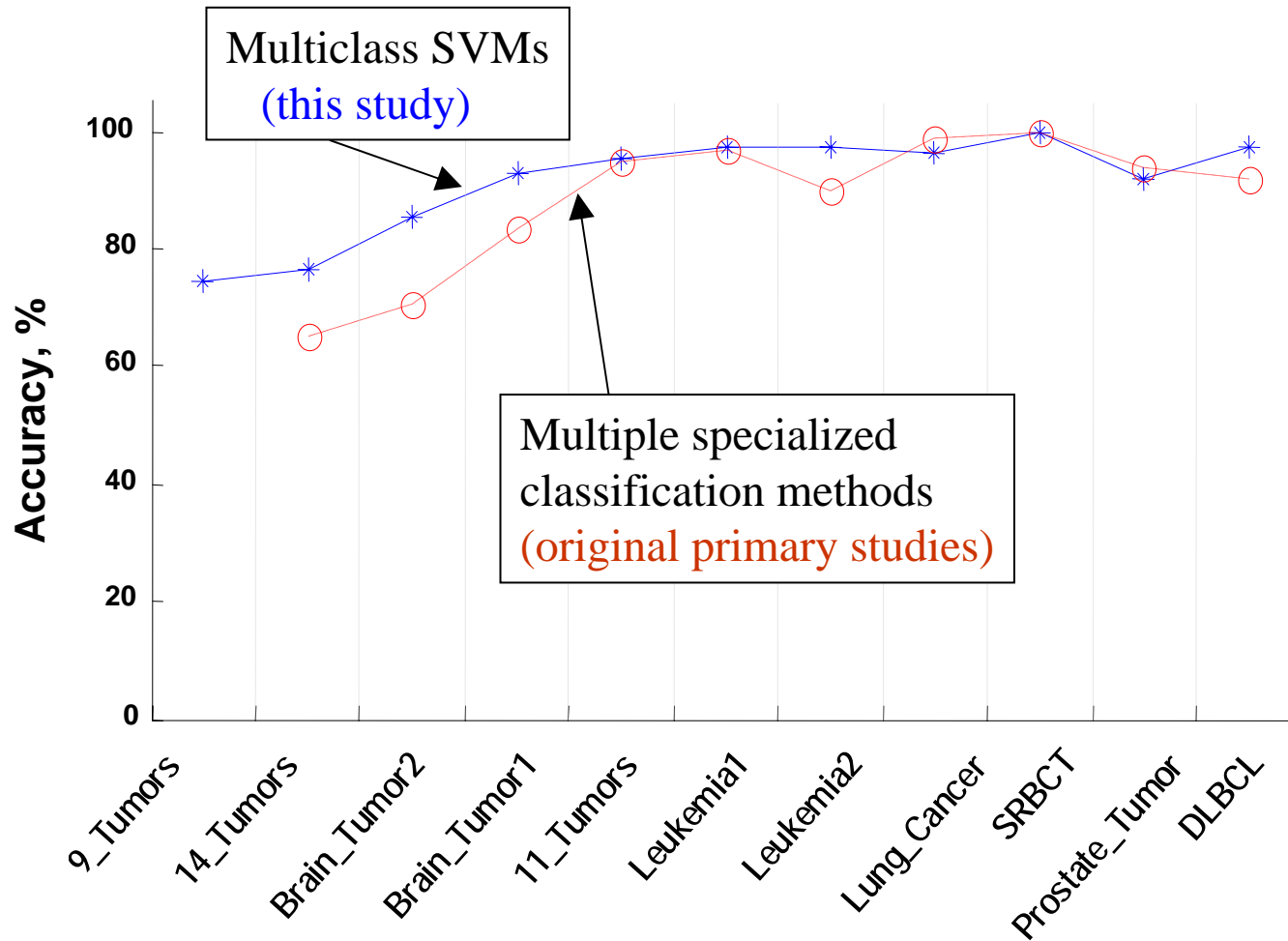
## Total:

- 1291 samples
- 74 diagnostic categories
- 41 cancer types and 12 normal tissue types

# Performance of Algorithms



# Comparison with Literature



# Conclusions of Evaluation

- 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.

# 2<sup>nd</sup> Algorithmic Evaluation Study

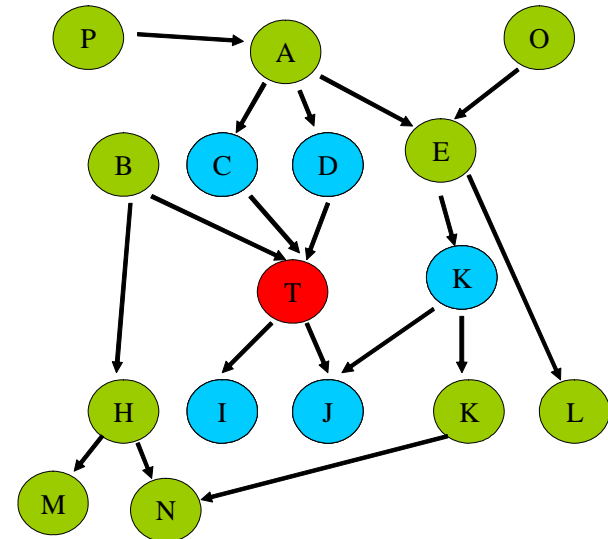
(Ongoing DSL research, not a part of MS project)

**Main Goal:** *Determine feature selection algorithms (applicable to high-dimensional microarray gene expression or mass-spectrometry data) that significantly reduce the number of predictors, maintaining optimal classification performance.*

Aliferis CF, Tsamardinos I, Statnikov A. *HITON: A novel Markov Blanket algorithm for optimal variable selection.* AMIA Symposium, 2003.

# Conclusion of Evaluation

Markov Blanket techniques (e.g., HITON) provide the smallest subsets of predictors that achieve optimal classification performance.



# Algorithms Implemented in *GEMS*

## Classifiers

MC-SVM

One-Versus-Rest

One-Versus-One

DAGSVM

Method by WW

Method by CS

## Cross-Validation Designs

N-Fold Cross-Validation

LOOCV

Nested N-Fold Cross-Validation

Nested LOOCV

## Normalization Techniques

[a, b]

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

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

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

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

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

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

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

ABS(x)

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

## Gene Selection Methods

S2N One-Versus-Rest

S2N One-Versus-One

Non-param. ANOVA

BW ratio

HITON\_MB

HITON\_PC

## Performance Metrics

Accuracy

RCI

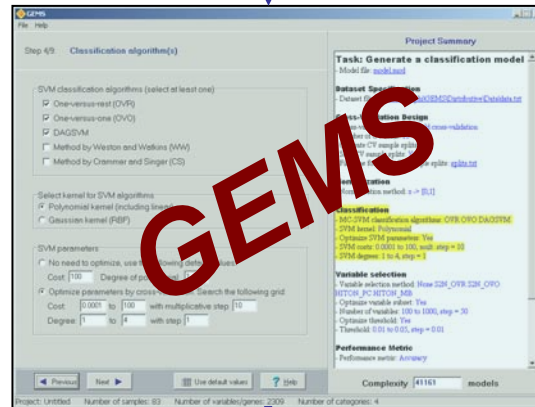
AUC ROC

# System Description

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

# Inputs & Outputs

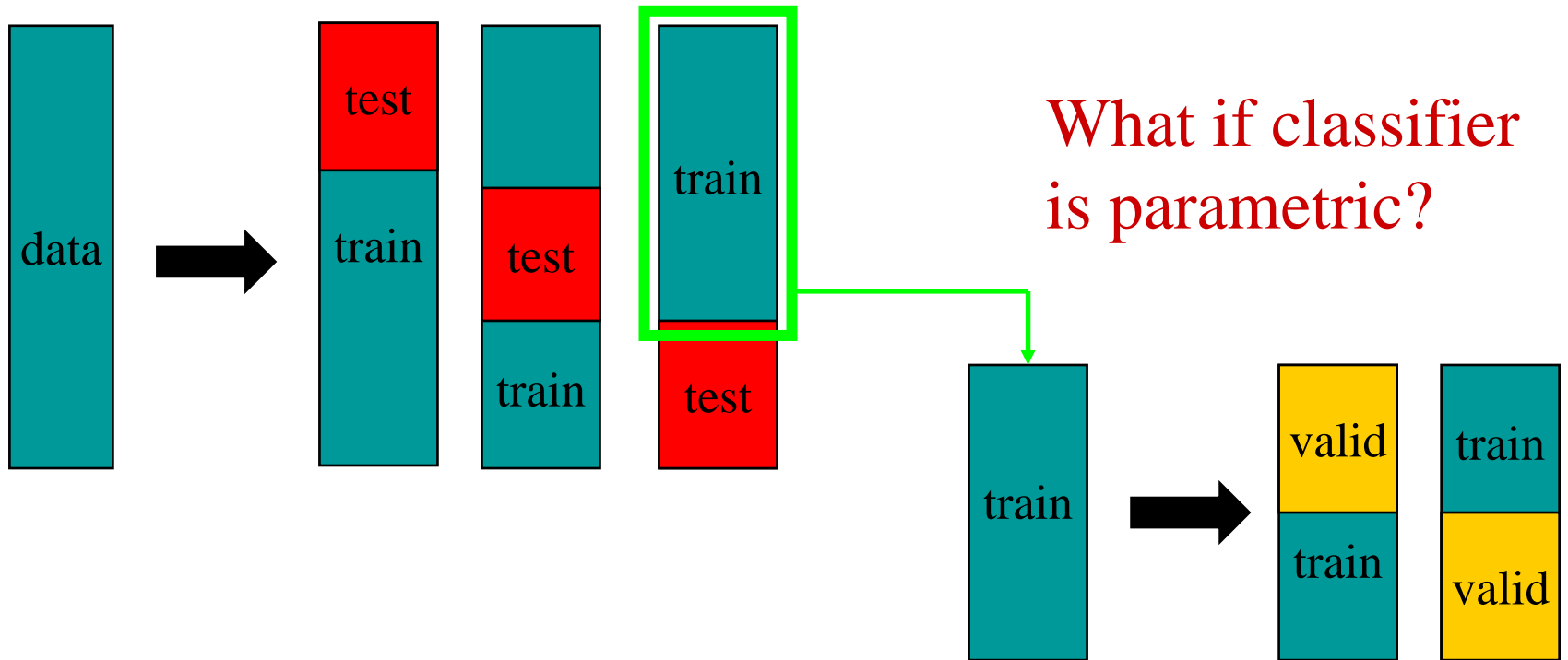
1. Dataset & outcome or diagnostic labels
2. *Optional*: gene names and/or accession numbers
3. Various choices of parameters for the analysis (defaults)
4. *In application mode*: previously saved model



1. Classification model
2. Performance estimate
3. *In application mode*: the model's diagnoses/predictions & overall performance
4. A reduced set of genes
5. Links from the genes to literature and other resources.

# Cross-Validation Design

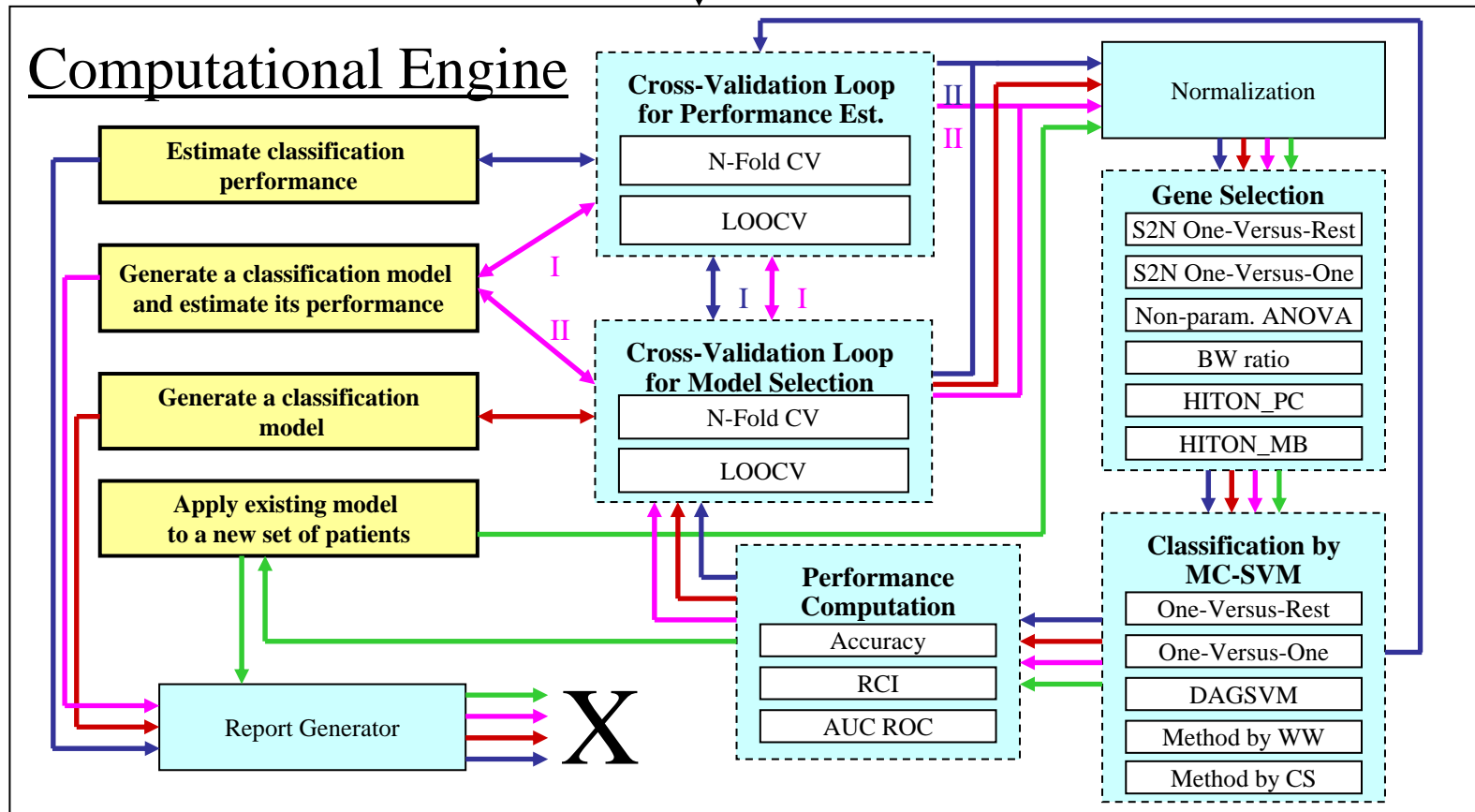
Performance estimation by Cross-Validation



Model selection / parameter optimization by (nested) cross-validation

# Software Architecture

Client (Wizard-Like User Interface)



# User Interface

The screenshot displays the GEMS software interface, specifically the 'Classification algorithm(s)' step (Step 4/3). The interface is divided into two main panels: the left panel for configuration and the right panel for a 'Project Summary'.

**Left Panel: Classification algorithm(s)**

- 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

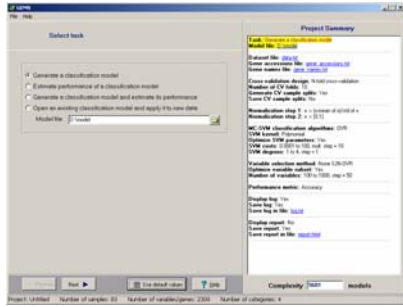
**Right Panel: Project Summary**

- 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**
  - MC-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](#)

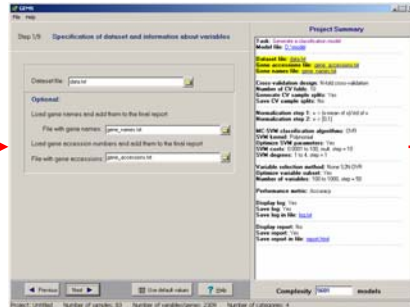
**Bottom Panel:**

- Navigation:
- Buttons:
- Complexity:  models
- Project: [Untitled](#) Number of samples: [83](#) Number of variables/genes: [2309](#) Number of categories: [4](#)

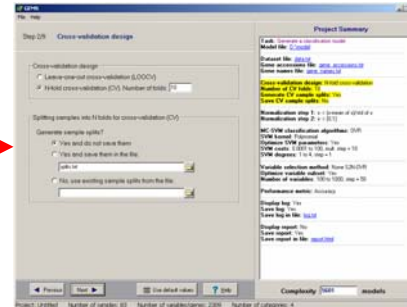
# Steps in User Interface



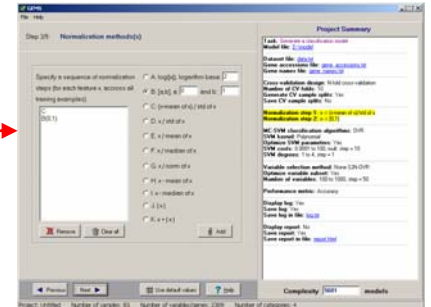
Task selection



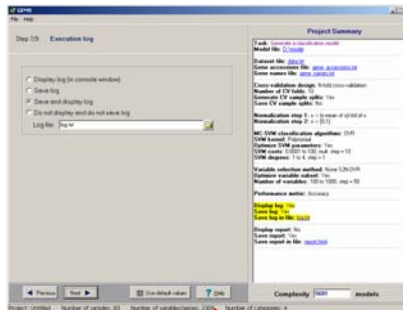
Dataset specification



Cross-validation design



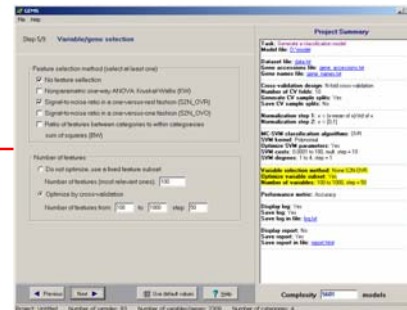
Normalization



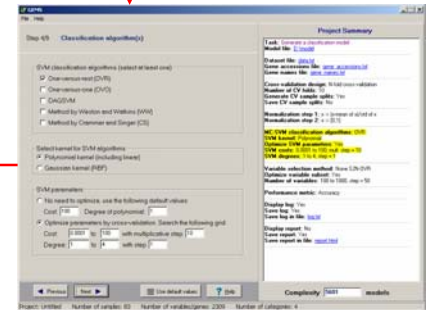
Logging



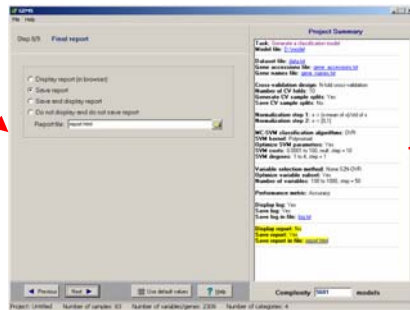
Performance metric



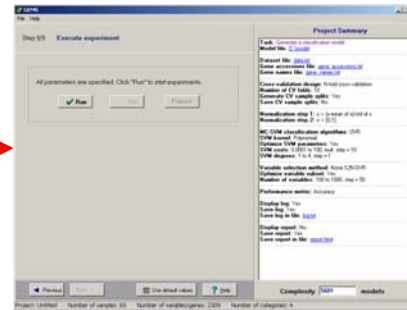
Gene selection



Classification



Report generation



Analysis execution

# An Evaluation of the System:

- Apply *GEMS* to datasets not involved in algorithmic evaluation and compare results with ones obtained by human analysts and published in the literature;
- Verify generalizability of models produced by *GEMS* in cross-dataset applications.

Statnikov A, Tsamardinos I, Aliferis CF. *GEMS: A system for decision support and discovery from array gene expression data*. International Journal of Medical Informatics, 2005.

# Evaluation Using New Datasets

## Datasets

Dataset name	Number of			Reference
	Sam- ples	Vari- ables (genes)	Cate- gories	
<i>6_Tumors</i>	353	7069	6	Shedden, 2003
<i>Leukemia3</i>	248	12135	6	Yeoh, 2002
<i>Lung_Cancer2</i>	96	7129	2	Beer, 2002
<i>Lung_Cancer3</i>	181	12533	2	Gordon, 2003
<i>DLBCL2</i>	210	32404	2	Savage, 2003

## Comparison with literature

Dataset name	GEMS classification accuracy	Published classification accuracy
<i>6_Tumors</i>	97.2%	96.0%
<i>Leukemia3</i>	98.4%	98.4%
<i>Lung_Cancer2</i>	100.0%	100.0%
<i>Lung_Cancer3</i>	99.4%	99.3%
<i>DLBCL2</i>	87.1%	83.9%

Analyzes were completed within 10-30 minutes with *GEMS*.

# Verify Generalizability of Models in Cross-Dataset Applications

Dataset used for construction of a classification model		Performance estimate of the model* (AUC, %)	Dataset used for independent validation of the classification model		Performance on the independent dataset (AUC, %)
Author	Distribution of samples		Author	Distribution of samples	
Lung cancer	<b>Bhattacharjee</b> 186 tumors 17 normals	100.00%	<b>Beer</b> 86 tumors 10 normals	100.00%	
Leukemia	<b>Armstrong</b> 24 ALL 28 AML	100.00%	<b>Golub</b> 47 ALL 25 AML	99.15%	

\* This performance estimate was obtained by nested cross-validation on the dataset used for construction of the model.

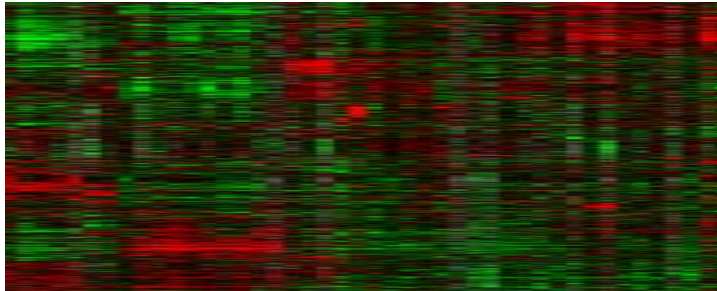
# Live Demonstration of *GEMS*

# Scenario 1:

Binary classification model development and evaluation using a lung cancer microarray gene expression dataset.

# Live Demo of *GEMS* (Scenario 1)

## Binary classification model development and evaluation

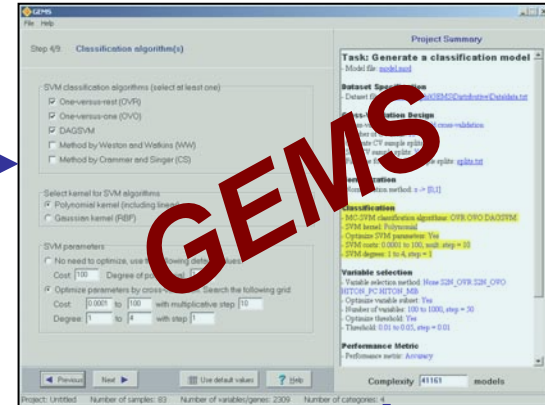


Lung cancer dataset from *Bhattacharjee, 2001*:

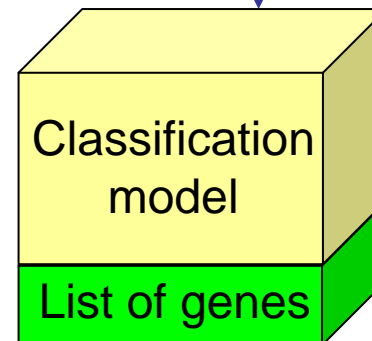
- Diagnostic task:  
Lung cancer vs normal tissues
- Microarray platform:  
Affymetrix U95A
- Number of oligonucleotides:  
12,600\*
- Number of patients:  
203

Various parameters

User



**GEMS**



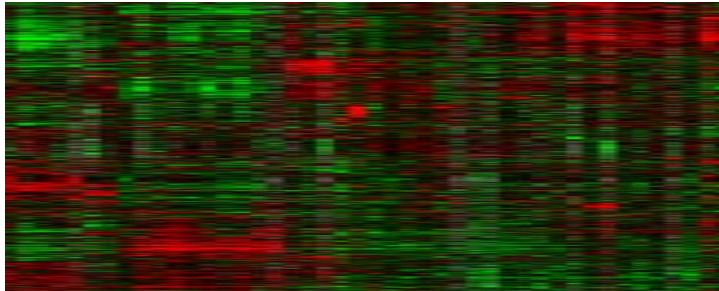
Cross-validation performance estimate (AUC)

## Scenario 2:

Multicategory classification model development and evaluation using a small round blood cell tumor microarray gene expression dataset.

# Live Demo of GEMS (Scenario 2)

## Multicategory classification model development and evaluation

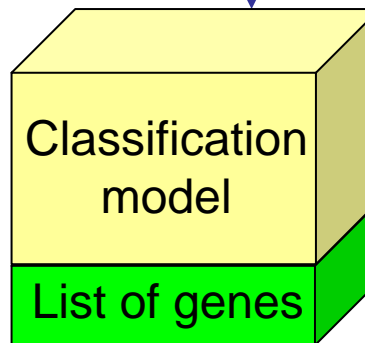
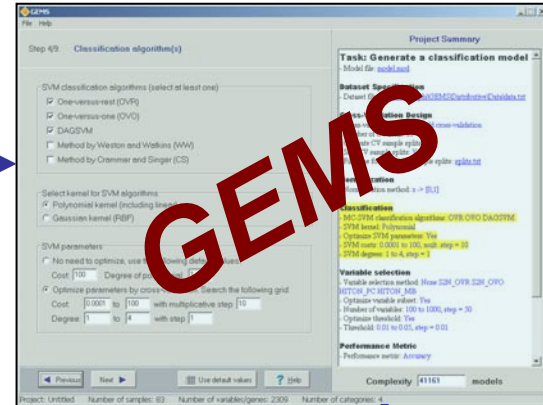


Lung cancer dataset from Khan, 2001:

- Diagnostic task:  
Ewing Sarcoma vs rhabdomyosarcoma vs Burkitt Lymphoma vs neuroblastoma
- Microarray platform:  
cDNA
- Number of probes:  
2,308
- Number of patients:  
63

Various parameters

User



Cross-validation performance estimate (AUC)

## Scenario 3:

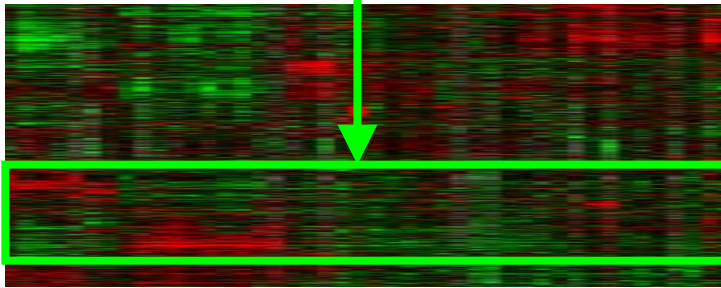
Validating the reproducibility of genes selected in Scenario 1 using another lung cancer microarray gene expression dataset.

# Live Demo of *GEMS* (Scenario 3)

Are selected genes reproducible in another dataset?

List of genes

From Scenario 1  
(*Bhattacharjee's data*)

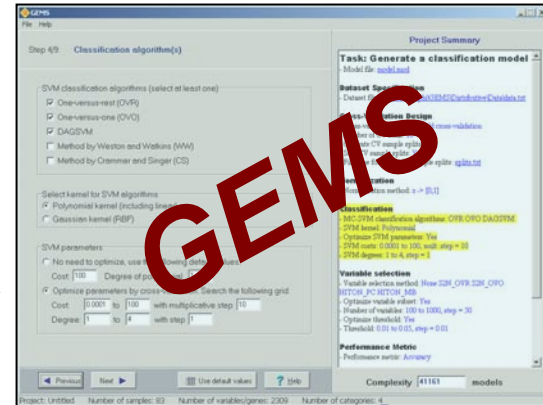


Lung cancer dataset from  
*Beer, 2002*:

- Diagnostic task:  
Lung cancer vs normal tissues
- Microarray platform:  
Affymetrix HuGeneFL
- Number of oligonucleotides:  
7,129\*
- Number of patients:  
96

Various  
parameters

User



Classification  
model

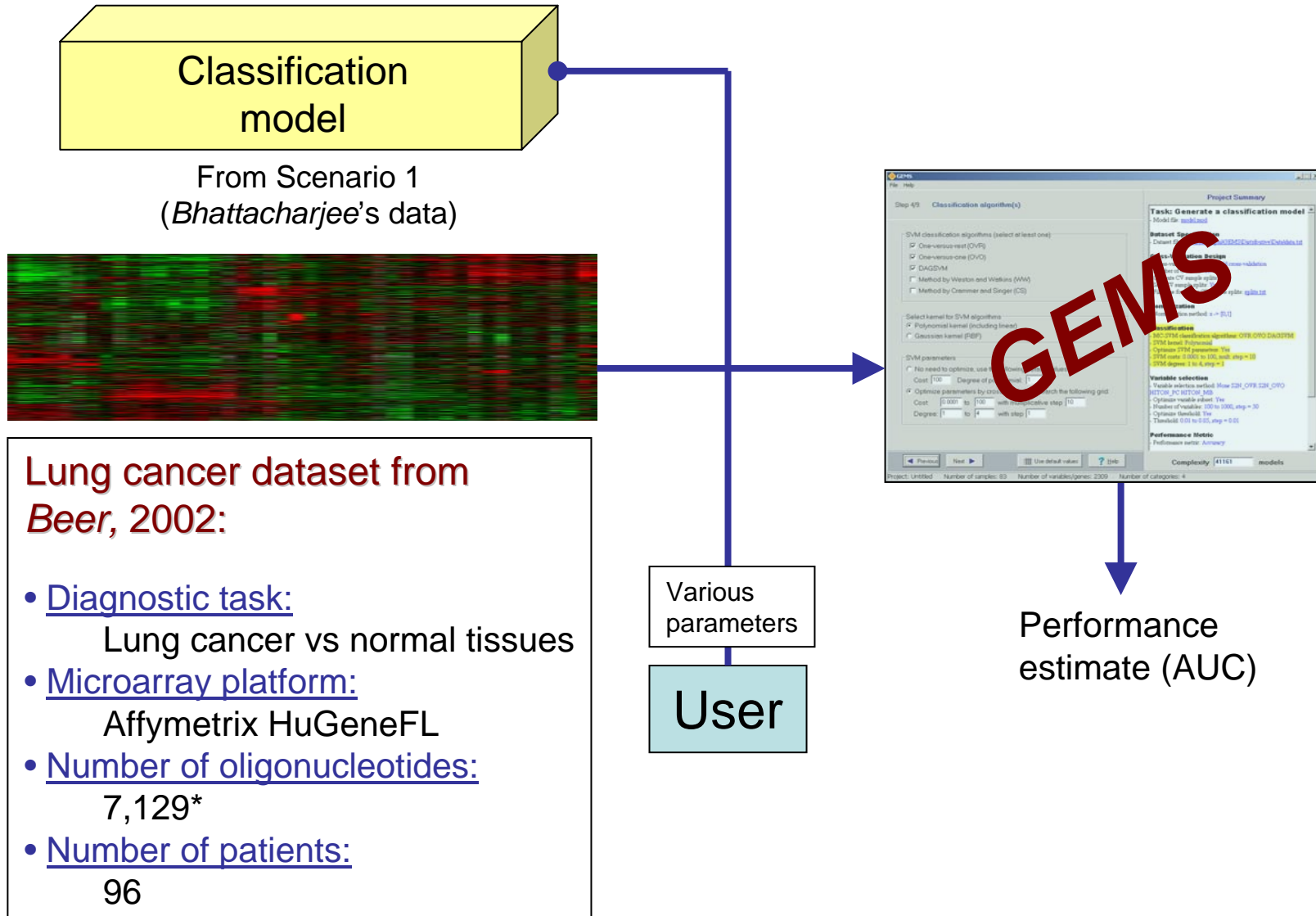
Cross-validation  
performance  
estimate (AUC)

## Scenario 4:

Verifying generalizability of the classification model produced in Scenario 1 using another lung cancer microarray gene expression dataset.

# Live Demo of *GEMS* (Scenario 4)

Is constructed classification model generalizable in another microarray dataset?



# *GEMS* in a Nutshell

1. The system is fully automated, yet provides many optional features for the seasoned analyst.
2. The system is based on a nested cross-validation design that avoids overfitting.
3. *GEMS*'s algorithms were chosen after the two extensive algorithmic evaluations.
4. After the system was built, it was validated in cross-dataset applications and also using new datasets.
5. *GEMS* has an intuitive wizard-like user interface which abstracts data analysis process.
6. *GEMS* possesses a convenient client-server architecture.

# Acknowledgements

Members of my MS Committee:

- Dr. Constantin F. Aliferis (advisor)
- Dr. Douglas P. Hardin
- Dr. Shawn Levy
- Dr. Ioannis Tsamardinou (advisor)
  
- Yerbolat Dosbayev
- Vanderbilt University Department of Biomedical Informatics faculty and students

NIH grants for funding of this project:

- R01 LM007948-01
- P20 LM007613-01

# References

## Journal Papers:

- **Statnikov A**, Tsamardinos I, Dosbayev Y, Aliferis CF. GEMS: A System for Automated Cancer Diagnosis and Biomarker Discovery from Microarray Gene Expression Data. *Int J Med Inform*. 2005.
- **Statnikov A**, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A Comprehensive Evaluation of Multicategory Classification Methods for Microarray Gene Expression Cancer Diagnosis. *Bioinformatics*. 2005 Mar 1;21(5):631-43.

## Papers in Conference Proceedings (peer-reviewed):

- **Statnikov A**, Aliferis CF, Tsamardinos I. Methods for Multi-category Cancer Diagnosis from Gene Expression Data: A Comprehensive Evaluation to Inform Decision Support System Development. *Medinfo*, 2004.

## Posters in Conference Proceedings (peer-reviewed):

- **Statnikov A**, Tsamardinos I, Aliferis CF. Using GEMS for Cancer Diagnosis and Biomarker Discovery from Microarray Gene Expression Data. *13th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 2005.

## Software Demonstrations in Conference Proceedings (peer-reviewed):

- **Statnikov A**, Tsamardinos I, Aliferis CF. GEMS: A System for Cancer Diagnosis and Biomarker Discovery from Microarray Gene Expression Data. *AMIA Annual Symposium*, 2005.
- **Statnikov A**, Tsamardinos I, Aliferis CF. Using the GEMS System for Cancer Diagnosis and Biomarker Discovery from Microarray Gene Expression Data. *12th National Conference on Artificial Intelligence (AAAI)*, 2005.
- **Statnikov A**, Tsamardinos I, Aliferis CF. Using GEMS for Cancer Diagnosis and Biomarker Discovery from Microarray Gene Expression Data. *13th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 2005.