CLASSIFICATION OF GENE EXPRESSION DATA

Mario.Guarracino@cnr.it
National Research Council, Italy

Outline

□ Microarrays: when, what, why, how, …

□ Classification: SVM, ReGEC, RBF NN

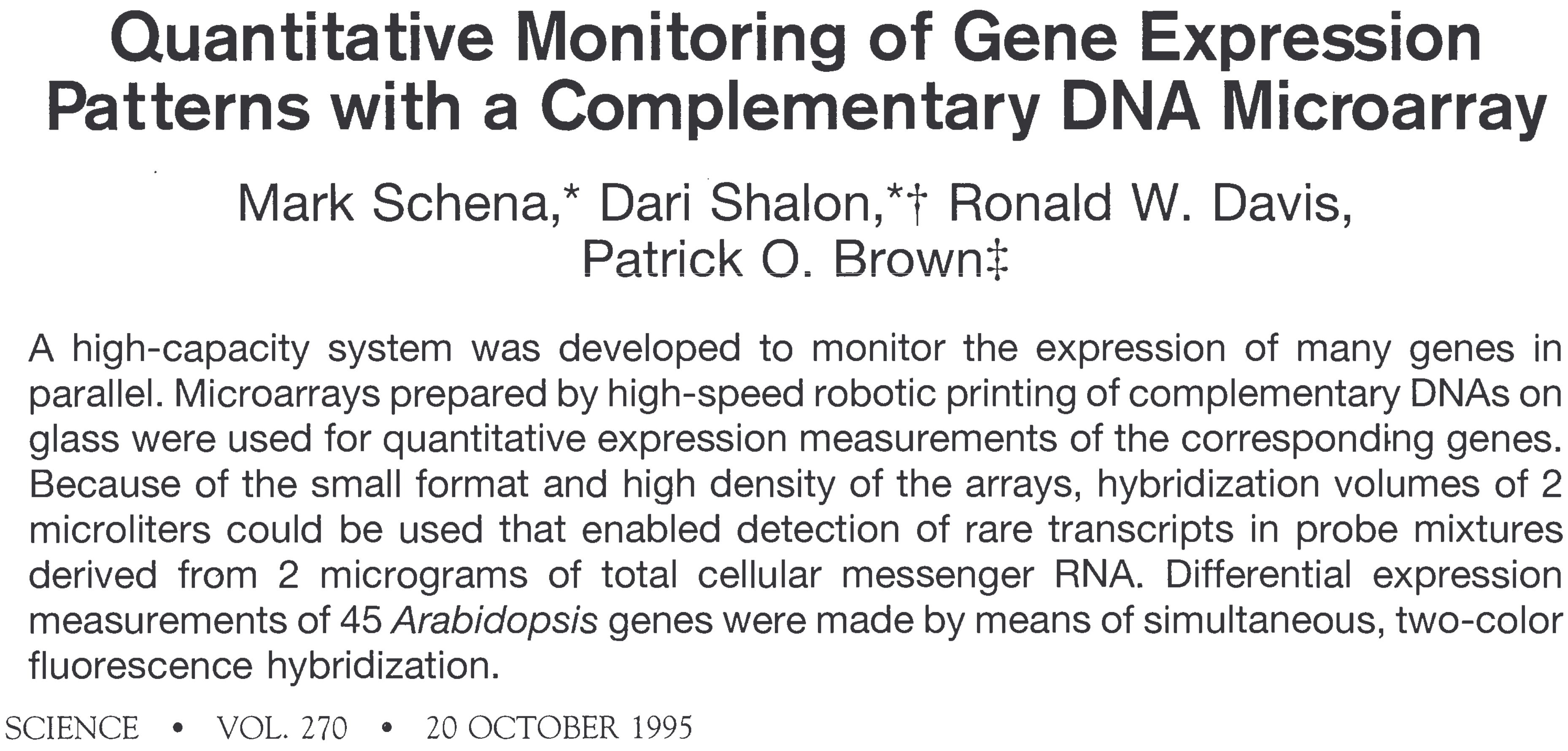
□ A priori knowledge in classification models

□ Knowledge as a mining task

□ A case study

□ Conclusions

When did it all begin?



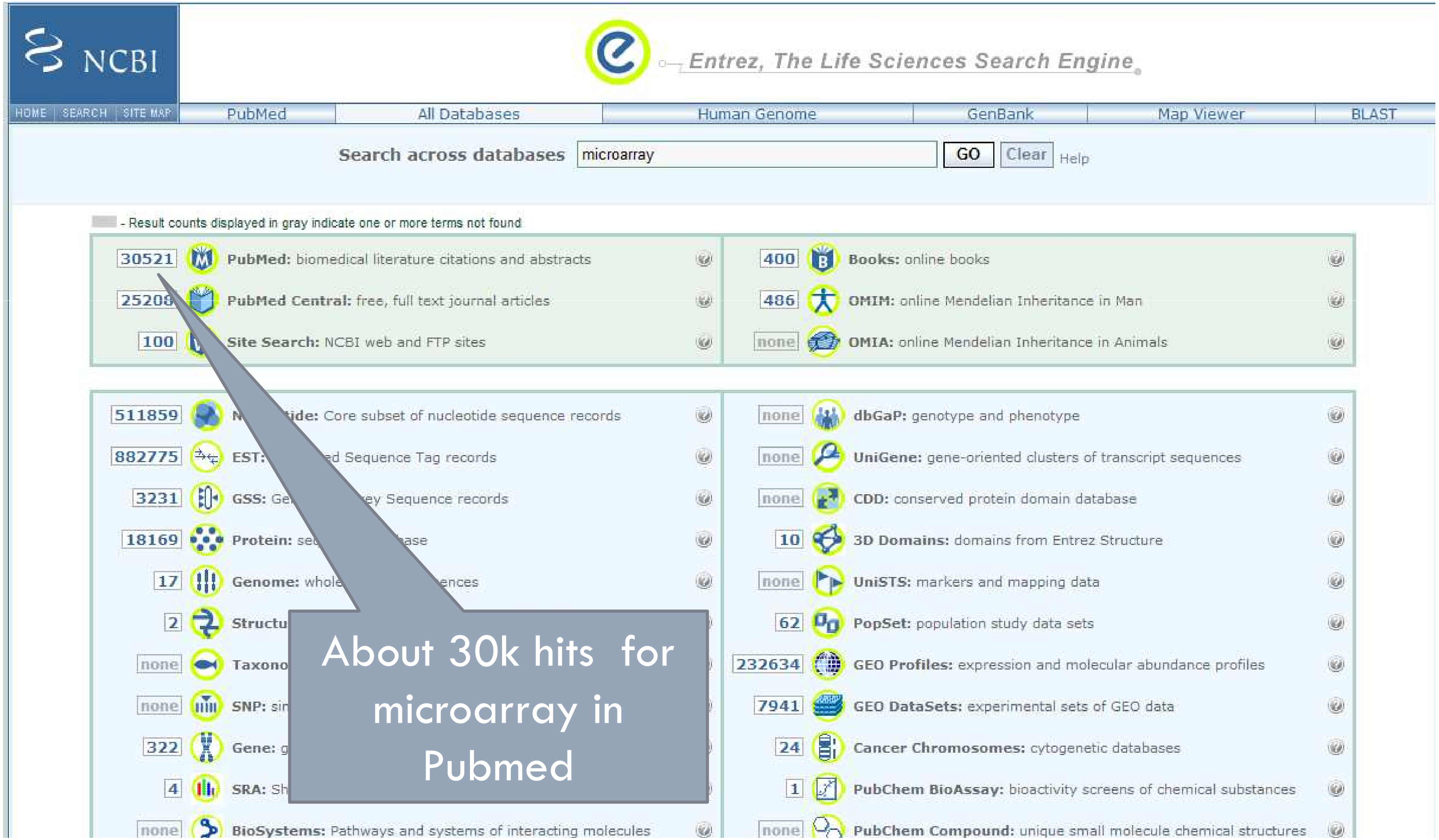
Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray

Mark Schena,* Dari Shalon,*† Ronald W. Davis,
Patrick O. Brown‡

A high-capacity system was developed to monitor the expression of many genes in parallel. Microarrays prepared by high-speed robotic printing of complementary DNAs on glass were used for quantitative expression measurements of the corresponding genes. Because of the small format and high density of the arrays, hybridization volumes of 2 microliters could be used that enabled detection of rare transcripts in probe mixtures derived from 2 micrograms of total cellular messenger RNA. Differential expression measurements of 45 *Arabidopsis* genes were made by means of simultaneous, two-color fluorescence hybridization.

SCIENCE • VOL. 270 • 20 OCTOBER 1995

Where are we now?!



NCBI
HOME | SEARCH | SITE MAP
Entrez, The Life Sciences Search Engine
PubMed | All Databases | Human Genome | GenBank | Map Viewer | BLAST

Search across databases: microarray

▨ - Result counts displayed in gray indicate one or more terms not found

30521	**PubMed:** biomedical literature citations and abstracts	400	**Books:** online books
25208	**PubMed Central:** free, full text journal articles	486	**OMIM:** online Mendelian Inheritance in Man
100	**Site Search:** NCBI web and FTP sites	none	**OMIA:** online Mendelian Inheritance in Animals

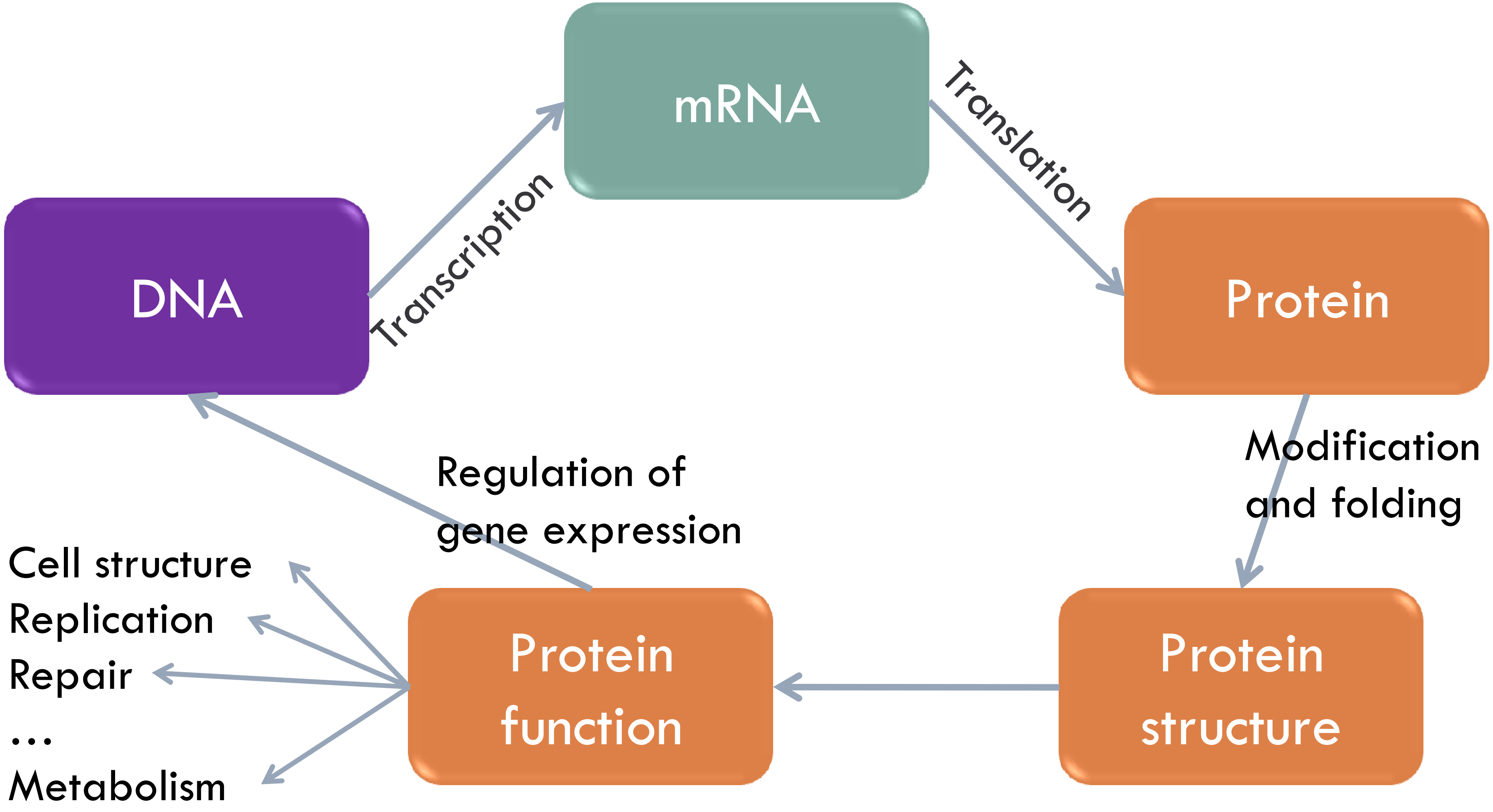
511859	**N tide:** Core subset of nucleotide sequence records	none	**dbGaP:** genotype and phenotype
882775	**EST:** ed Sequence Tag records	none	**UniGene:** gene-oriented clusters of transcript sequences
3231	**GSS:** Ge ey Sequence records	none	**CDD:** conserved protein domain database
18169	**Protein:** se ase	10	**3D Domains:** domains from Entrez Structure
17	**Genome:** whole ences	none	**UniSTS:** markers and mapping data
2	**Structu**	62	**PopSet:** population study data sets
none	**Taxono**	232634	**GEO Profiles:** expression and molecular abundance profiles
none	**SNP:** sin	7941	**GEO DataSets:** experimental sets of GEO data
322	**Gene:** g	24	**Cancer Chromosomes:** cytogenetic databases
4	**SRA:** Sh	1	**PubChem BioAssay:** bioactivity screens of chemical substances
none	**BioSystems:** Pathways and systems of interacting molecules	none	**PubChem Compound:** unique small molecule chemical structures

About 30k hits for microarray in Pubmed

Gene expression process

- The genetic information of an organism is stored in a string composed of 4 letters (nucleotides).
- These strings form the DNA molecules that compose the genome of an organism.
- The genome contains segments of DNA that encode genes.
- Genes are transcribed in messenger RNA and translated to form proteins.

Gene expression process



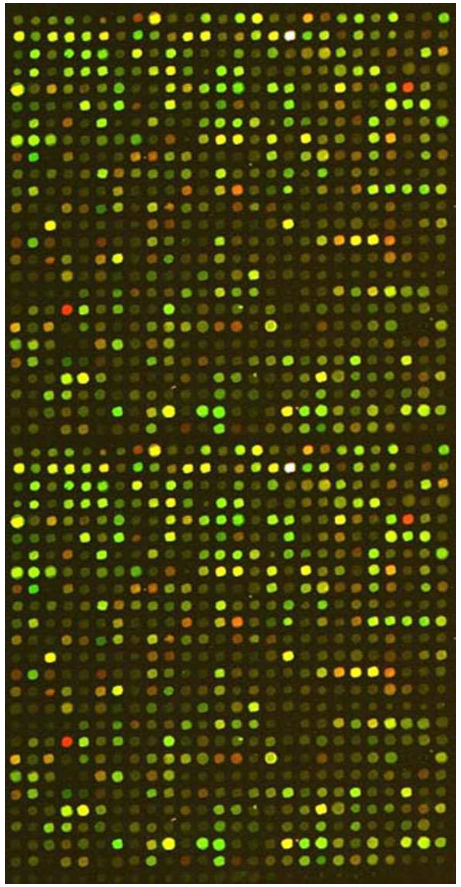
H. Causton, J. Quackenbush, and A. Brazma. Microarray gene expression data analysis, 2003

Microarrays

- DNA is present in nearly all cells of an organism, but these are not all the same.
- Many differences are due to the different subset of genes that are expressed in the different cell types.
- Microarrays permit the detection of abundance of various mRNA molecules in a cell.
- The abundance of each mRNA can provide information on the corresponding protein.

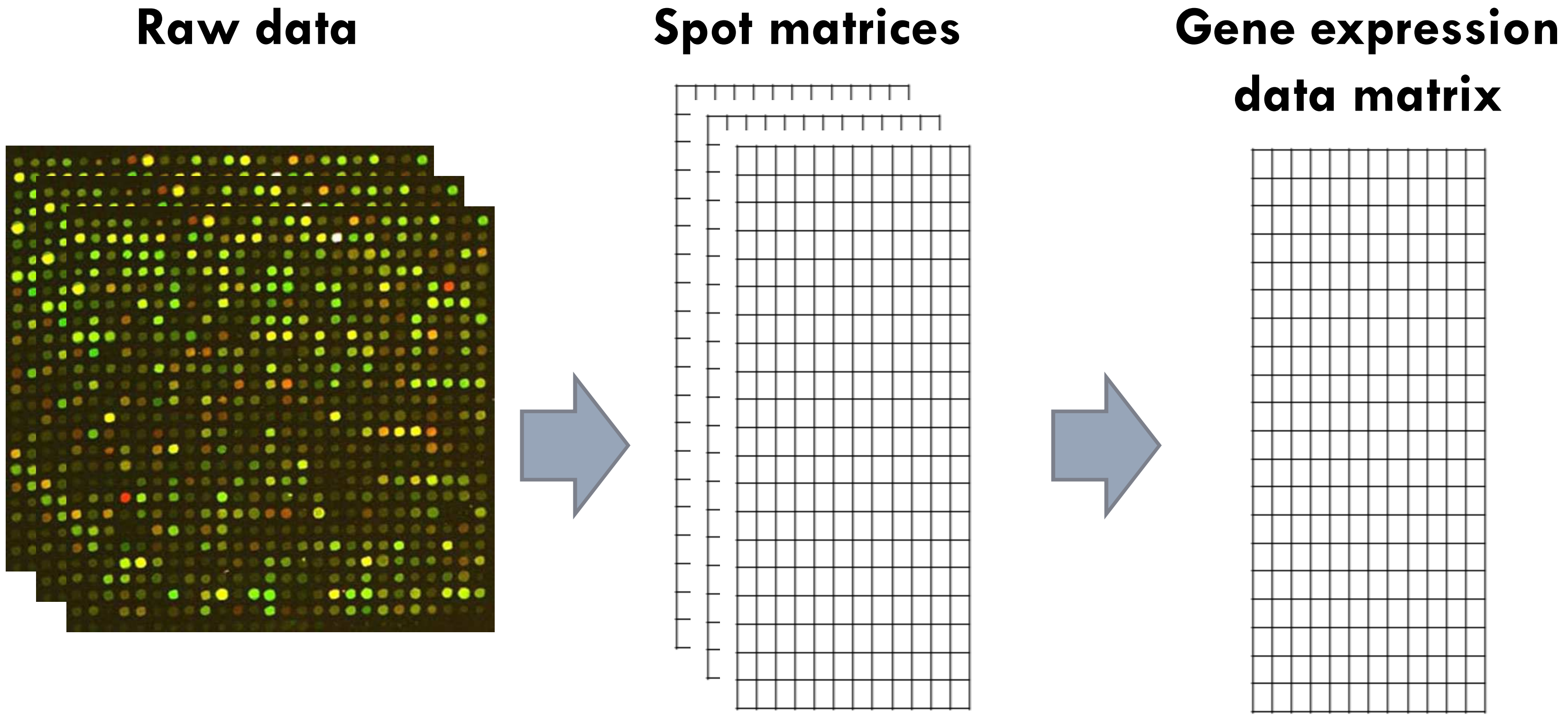
How do microarrays work?

- □ DNA microarrays are typically glass slides on which is printed a series of spots (tens of thousands) of DNA.
- □ Each spot corresponds to some portion of a known gene or predicted open reading frame.
- □ Each spot should identify the expression level of mRNA transcript by a gene.



From images to data

- □ The raw data are digital images.
- □ To obtain information about expression levels, each spot is identified and its intensity measured.



Raw data

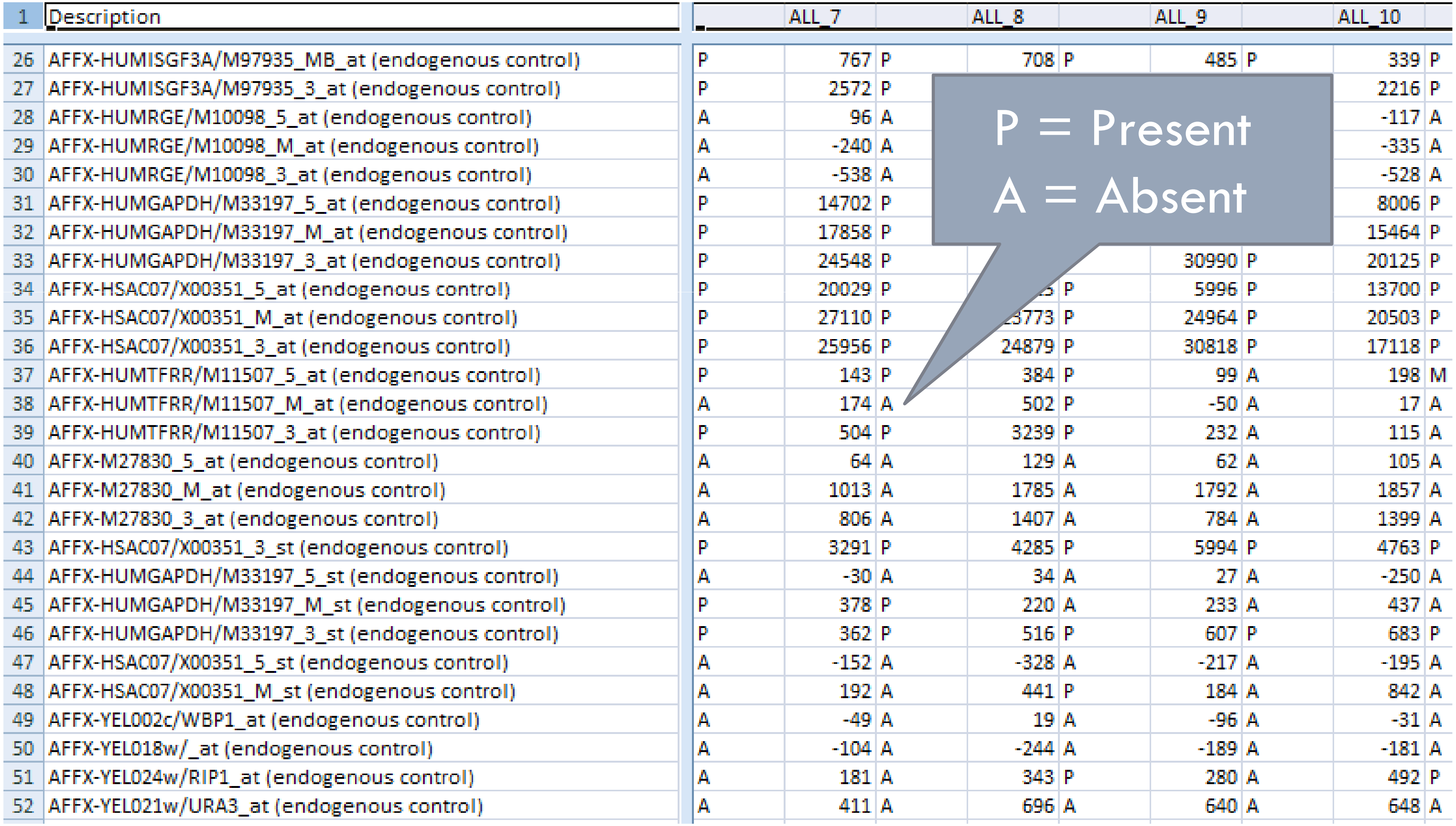
Spot matrices

Gene expression data matrix

Missing and noisy data

- In the process of extracting one intensity level from each spot, many values are missing or affected by an error.
- Solutions adopted: ignore sample, estimate or impute a value.
- Due to the cost of each experiment, missing values are estimated.

An example of data file



1	Description		ALL_7		ALL_8		ALL_9		ALL_10	
26	AFFX-HUMISGF3A/M97935_MB_at (endogenous control)	P	767	P	708	P	485	P	339	P
27	AFFX-HUMISGF3A/M97935_3_at (endogenous control)	P	2572	P	[illegible]	[illegible]	[illegible]	[illegible]	2216	P
28	AFFX-HUMRGE/M10098_5_at (endogenous control)	A	96	A	[illegible]	[illegible]	[illegible]	[illegible]	-117	A
29	AFFX-HUMRGE/M10098_M_at (endogenous control)	A	-240	A	[illegible]	[illegible]	[illegible]	[illegible]	-335	A
30	AFFX-HUMRGE/M10098_3_at (endogenous control)	A	-538	A	[illegible]	[illegible]	[illegible]	[illegible]	-528	A
31	AFFX-HUMGAPDH/M33197_5_at (endogenous control)	P	14702	P	[illegible]	[illegible]	[illegible]	[illegible]	8006	P
32	AFFX-HUMGAPDH/M33197_M_at (endogenous control)	P	17858	P	[illegible]	[illegible]	[illegible]	[illegible]	15464	P
33	AFFX-HUMGAPDH/M33197_3_at (endogenous control)	P	24548	P	[illegible]	[illegible]	30990	P	20125	P
34	AFFX-HSAC07/X00351_5_at (endogenous control)	P	20029	P	[illegible]	P	5996	P	13700	P
35	AFFX-HSAC07/X00351_M_at (endogenous control)	P	27110	P	[illegible]3773	P	24964	P	20503	P
36	AFFX-HSAC07/X00351_3_at (endogenous control)	P	25956	P	24879	P	30818	P	17118	P
37	AFFX-HUMTFRR/M11507_5_at (endogenous control)	P	143	P	384	P	99	A	198	M
38	AFFX-HUMTFRR/M11507_M_at (endogenous control)	A	174	A	502	P	-50	A	17	A
39	AFFX-HUMTFRR/M11507_3_at (endogenous control)	P	504	P	3239	P	232	A	115	A
40	AFFX-M27830_5_at (endogenous control)	A	64	A	129	A	62	A	105	A
41	AFFX-M27830_M_at (endogenous control)	A	1013	A	1785	A	1792	A	1857	A
42	AFFX-M27830_3_at (endogenous control)	A	806	A	1407	A	784	A	1399	A
43	AFFX-HSAC07/X00351_3_st (endogenous control)	P	3291	P	4285	P	5994	P	4763	P
44	AFFX-HUMGAPDH/M33197_5_st (endogenous control)	A	-30	A	34	A	27	A	-250	A
45	AFFX-HUMGAPDH/M33197_M_st (endogenous control)	P	378	P	220	A	233	A	437	A
46	AFFX-HUMGAPDH/M33197_3_st (endogenous control)	P	362	P	516	P	607	P	683	P
47	AFFX-HSAC07/X00351_5_st (endogenous control)	A	-152	A	-328	A	-217	A	-195	A
48	AFFX-HSAC07/X00351_M_st (endogenous control)	A	192	A	441	P	184	A	842	A
49	AFFX-YEL002c/WBP1_at (endogenous control)	A	-49	A	19	A	-96	A	-31	A
50	AFFX-YEL018w/_at (endogenous control)	A	-104	A	-244	A	-189	A	-181	A
51	AFFX-YEL024w/RIP1_at (endogenous control)	A	181	A	343	P	280	A	492	P
52	AFFX-YEL021w/URA3_at (endogenous control)	A	411	A	696	A	640	A	648	A

P = Present

A = Absent

Microarray applications

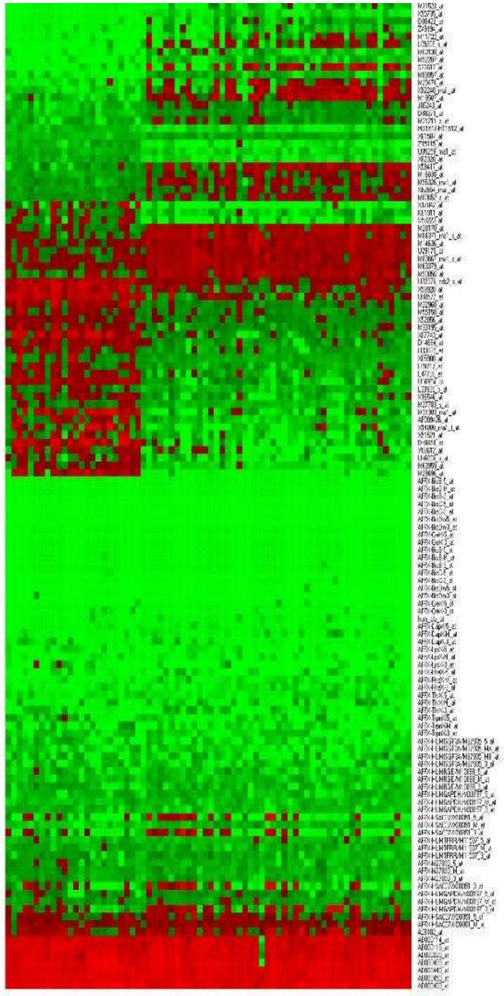
- Gene expression data have proven to be highly informative of disease state.
- In the area of oncology, accurate diagnosis and appropriate treatment are critical.
- Studies on clinical samples have shown gene expression data can be used to classify tumor types, detect subtypes, and to predict prognostic outcomes.

Classification

- □ Classification has become an important tool for microarray data analysis.
- □ Extracting information and knowledge from large amount of data is important to understand the underlying motivations of complex phenomena.
- □ Binary classification is among the most successful methods for microarray data analysis.

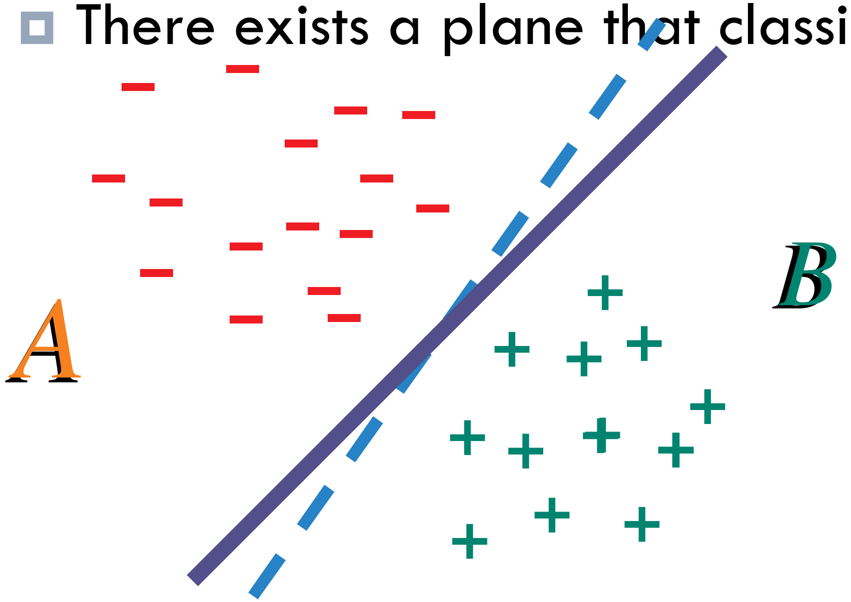
Challenges in microarrays

- Data produced by microarrays are exponentially increasing.
- Publicly available datasets contain gene expression data with tens of thousands characteristics.
- Data are incomplete and noisy.
- Current classification methods can over-fit the problem, providing models that do not generalize well.



Linear discriminant planes

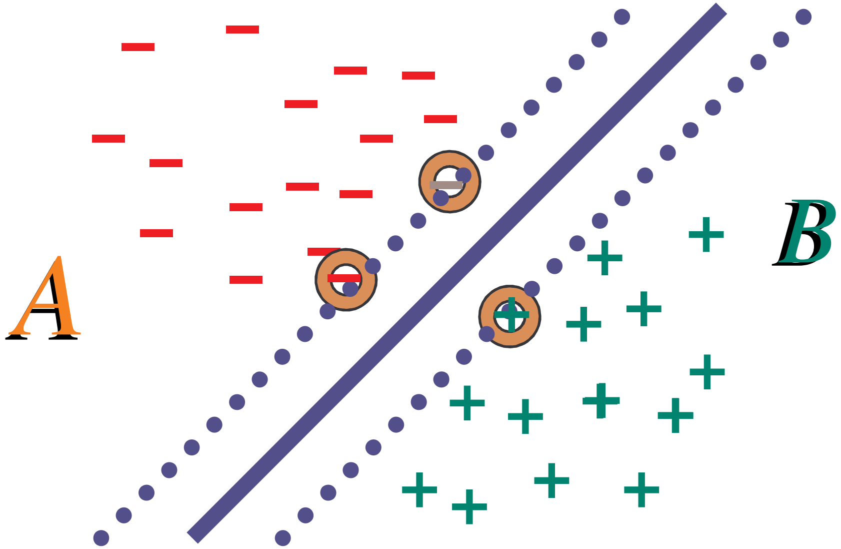
- Consider a binary classification task with points in two linearly separable sets.
  - There exists a plane that classifies all points in the two sets



- There are infinitely many planes that correctly classify the training data.

Support Vector Machines

- Find the plane $x'\omega\text{-}b=0$ which maximizes the margin between the two classes



$$\min_{\omega \neq 0} \frac{\|\omega\|^2}{2}$$

$$\begin{aligned} s.t. \quad & A\omega + b \geq e \\ & B\omega + b < -e \end{aligned}$$

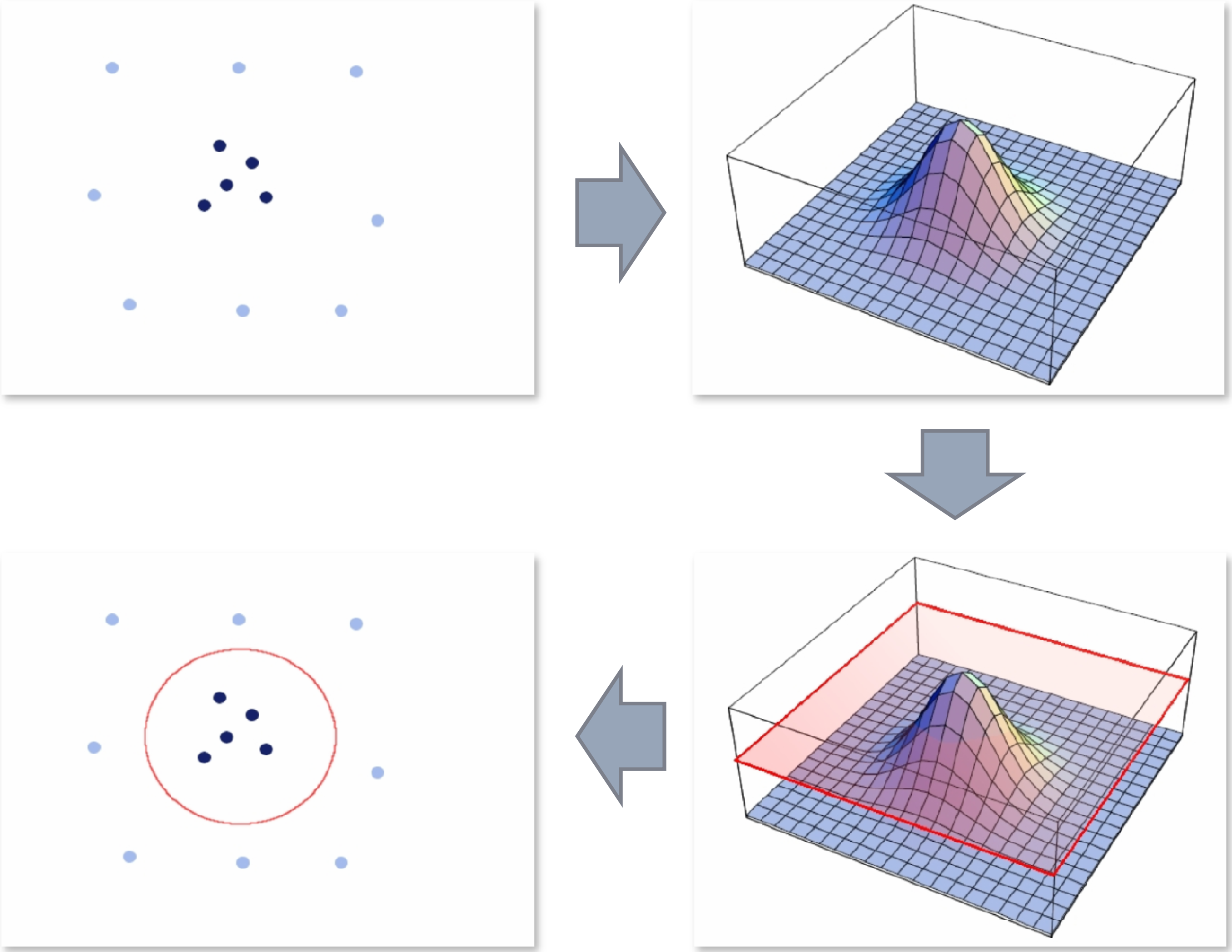
- Only few points are needed to compute the plane (*support vectors*).

SVM classification

- The robustness of SVM relies in the strong fundamentals of statistical learning theory.
- The training relies on optimization of a quadratic convex cost function, for which many methods are available.
  - Available packages for R, Matlab, Weka include SVM-Lite and LIBSVM.
- These techniques can be extended to the nonlinear discrimination, embedding the data in a nonlinear space using *kernel functions*.

The kernel trick

- □ To obtain greater separability between classes, nonlinearly embed points into a higher dimensional space



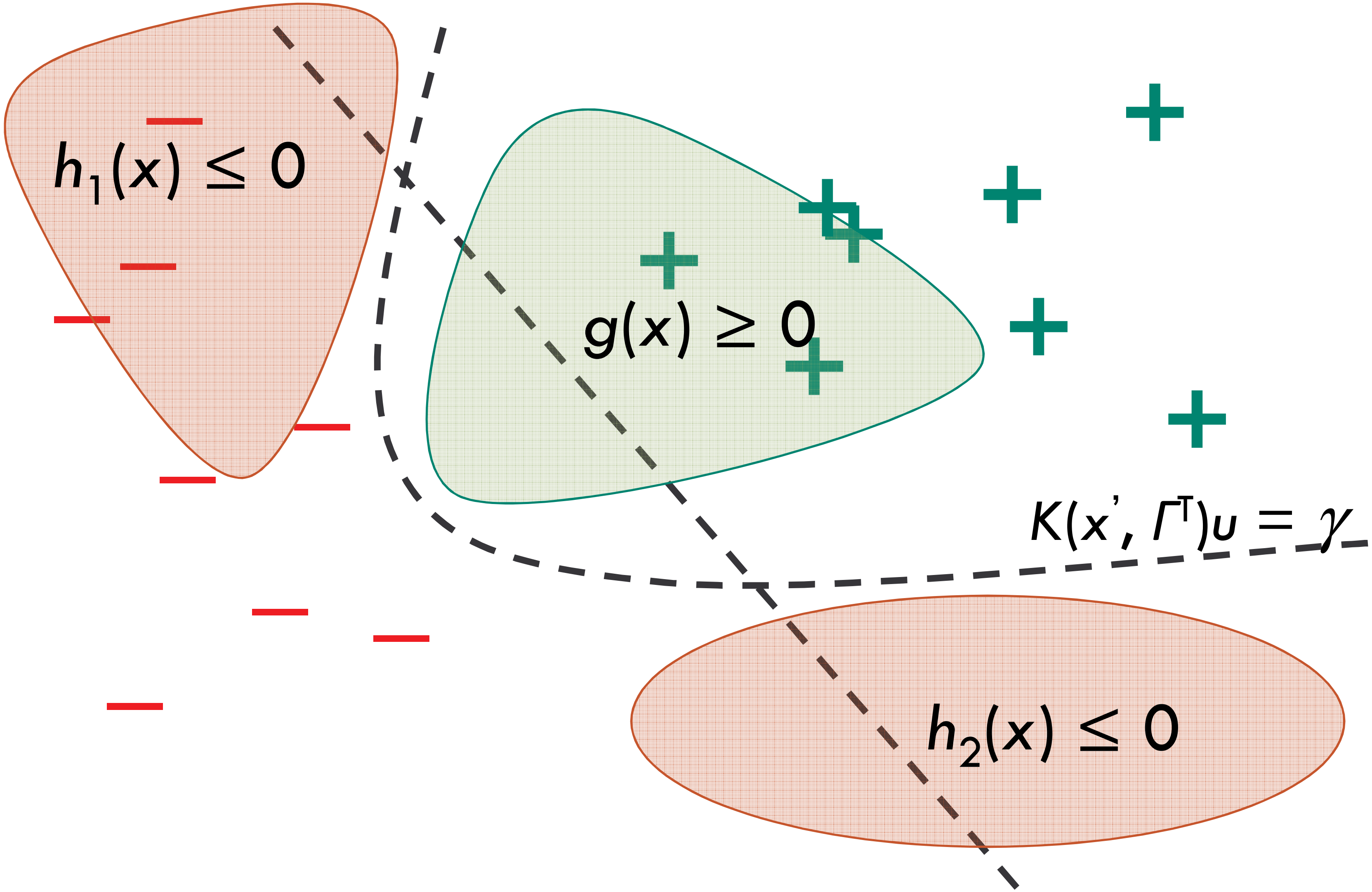
Prior knowledge

- It is possible to integrate external or prior knowledge in a classification model.
- A natural approach is to plug such knowledge in a classifier adding directly more points to the training set.
- This results in higher computational complexity, and in a tendency to overfitting.
- Different strategies need to be devised to take advantage of prior knowledge.

Prior knowledge

- An interesting approach is to analytically express knowledge as additional constraints to the optimization problem defining a standard SVM.
- This solution has the advantage
  - not to increase the dimension of the training set,
  - to avoid overfitting and poor generalization of the classification model.
- An analytical expression of knowledge is needed.

Prior knowledge incorporation



Prior knowledge in SVM

- Maximize the margin between the two classes, constraining the classification model to leave one positive region in the corresponding halfspace:

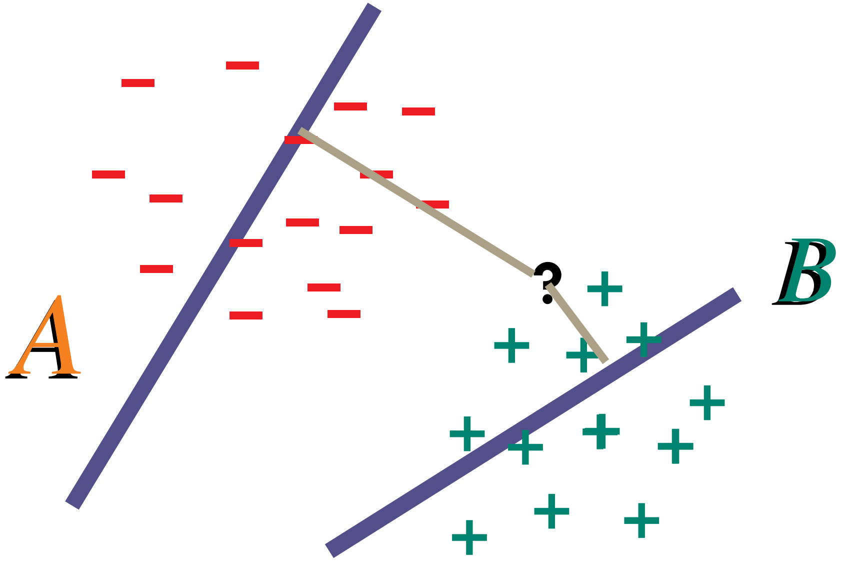
$$
\begin{array}{lll}
\underset{u,\gamma,y,s,v,z_1,\ldots,z_l}{min} & \nu e'y + e's + \sigma \sum_{i=1}^{l} z_i & \\
s.t. & D(K(\Gamma,\Gamma^T)u - \gamma e) + y & \geq e, \\
 & -s \leq u \leq s,\ y & \geq 0, \\
 & K(x_i',\Gamma^T)u - \gamma - \alpha + v'g(x_i) + z_i & \geq 0, \\
 & v \geq 0,\ z_i & \geq 0, \\
 & i = 1, \ldots, l. &
\end{array}
$$

- Simple extension to multiple knowledge regions.

O. Mangasarian, E. Wild Nonlinear Knowledge-Based Classification. IEEE TNN, 2008.

A different religion: ReGEC

- A binary classification problem can be formulated as a generalized eigenvalue problem (ReGEC).
  - Find $x'w_1=\gamma_1$ the closer to A and the farther from B:

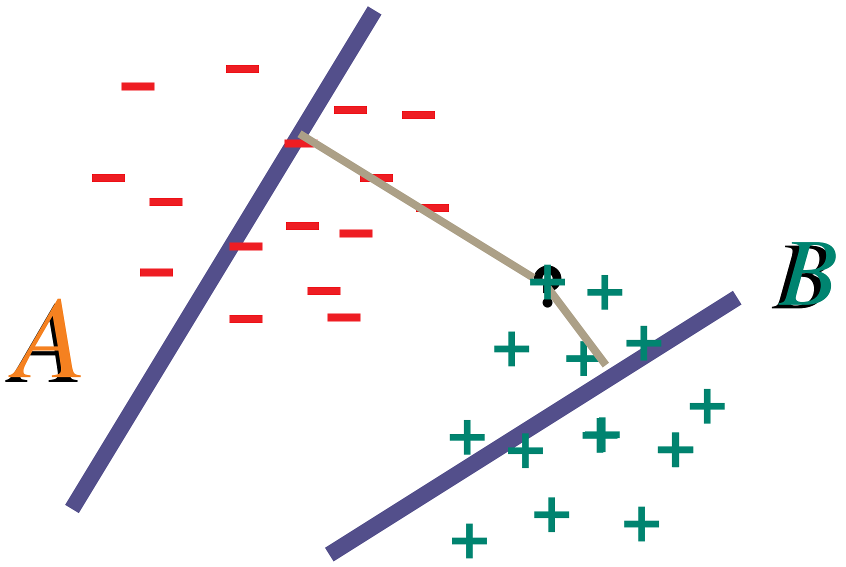


$$\min_{\omega,\gamma\neq 0}\frac{\|A\omega-e\gamma\|^2}{\|B\omega-e\gamma\|^2}$$

O. Mangasarian, E. Wild Multisurface Proximal Support Vector Classification via Generalized Eigenvalues. IEEE PAMI 2006.

A different religion: ReGEC

- A binary classification problem can be formulated as a generalized eigenvalue problem (ReGEC).
  - Find $x'w_1 = \gamma_1$ the closer to A and the farther from B:

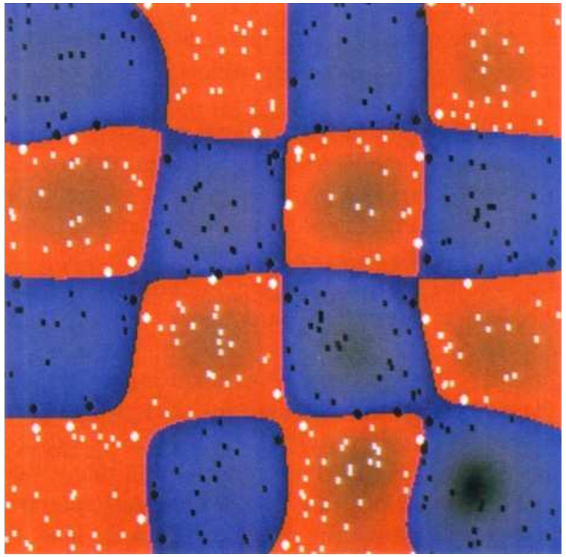


$$\min_{\omega,\gamma\neq 0} \frac{\| A\omega - e\gamma \|^2}{\| B\omega - e\gamma \|^2}$$

O. Mangasarian, E. Wild Multisurface Proximal Support Vector Classification via Generalized Eigenvalues. IEEE PAMI 2006.

The kernel trick for ReGEC

- The nonlinear embedding is obtained with a *RBF kernel function*:



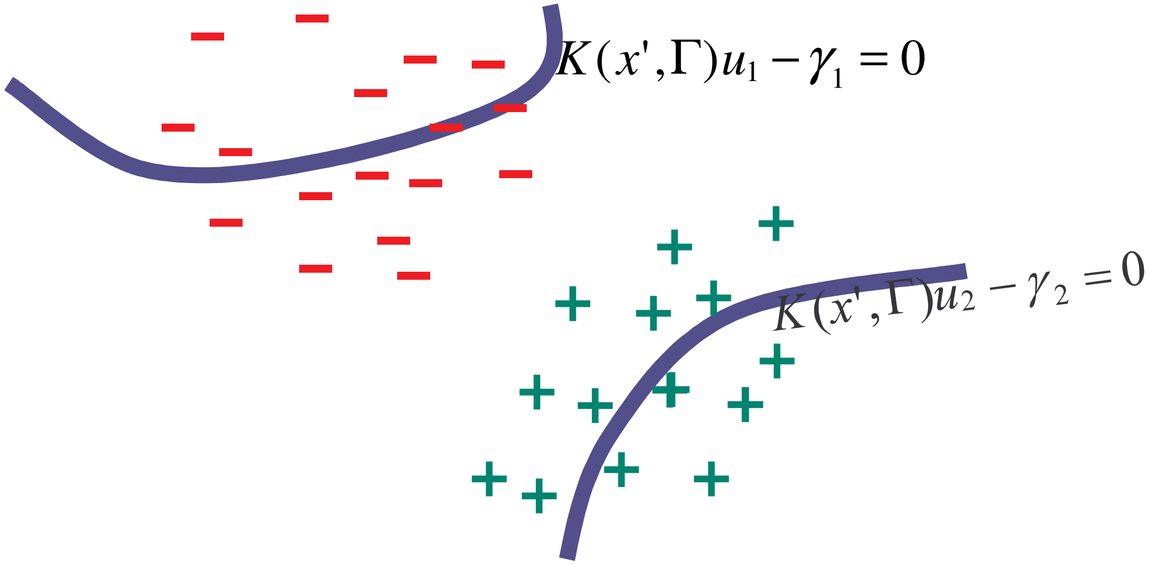
$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{\sigma}}$$

- Each element of kernel matrix is:

$$K(A, \Gamma)_{ij} = e^{-\frac{\|A_i - \Gamma_j\|^2}{\sigma}}$$

A different religion: ReGEC

- The problem can be restated as: find two hyperplanes (in the feature space), each the closest to one set and the furthest from the other.



$$\min_{u,\gamma\neq 0}\frac{\|K(A,\Gamma)u-e\gamma\|^2}{\|K(B,\Gamma)u-e\gamma\|^2}$$

$$\Gamma=\begin{bmatrix}A\\B\end{bmatrix}$$

- The binary classification problem can be solved as a generalized eigenvalue problem.

ReGEC

$$\min_{u,\gamma\neq 0} \frac{\| K(A,\Gamma)u - e\gamma \|^2}{\| K(B,\Gamma)u - e\gamma \|^2} = \min_{u,\gamma\neq 0} \frac{\| [K(A,\Gamma) \quad -e]^T [u' \quad \gamma]' \|^2}{\| [K(B,\Gamma) \quad -e]^T [u' \quad \gamma]' \|^2}$$

- □ Let

$$\begin{aligned} G &= [K(A,\Gamma) \quad -e]^T [K(A,\Gamma) \quad -e], \\ H &= [K(B,\Gamma) \quad -e]^T [K(B,\Gamma) \quad -e], \\ z &= [u' \quad \gamma]'. \end{aligned}$$

- □ the equation becomes:

$$\min_{z\in R^{n+1}} \frac{z'Gz}{z'Hz}$$

- □ Rayleigh quotients of $Gz = \lambda Hz$.

Regularization of ReGEC

- □ To regularize the problem, generate the two proximal surfaces:

$$K(x',\Gamma)u_1-\gamma_1=0 \qquad K(x',\Gamma)u_2-\gamma_2=0$$

- □ solving

$$\min_{u,\gamma\neq 0}\frac{\|K(A,\Gamma)u-e\gamma\|^2+\delta\|\tilde{K}_A u-e\gamma\|^2}{\|K(B,\Gamma)u-e\gamma\|^2}$$

$$\min_{u,\gamma\neq 0}\frac{\|K(B,\Gamma)u-e\gamma\|^2+\delta\|\tilde{K}_B u-e\gamma\|^2}{\|K(A,\Gamma)u-e\gamma\|^2}$$

- □ $\tilde{K}_A$ and $\tilde{K}_B$ main diagonals of K(A, Γ) and K(B, Γ)

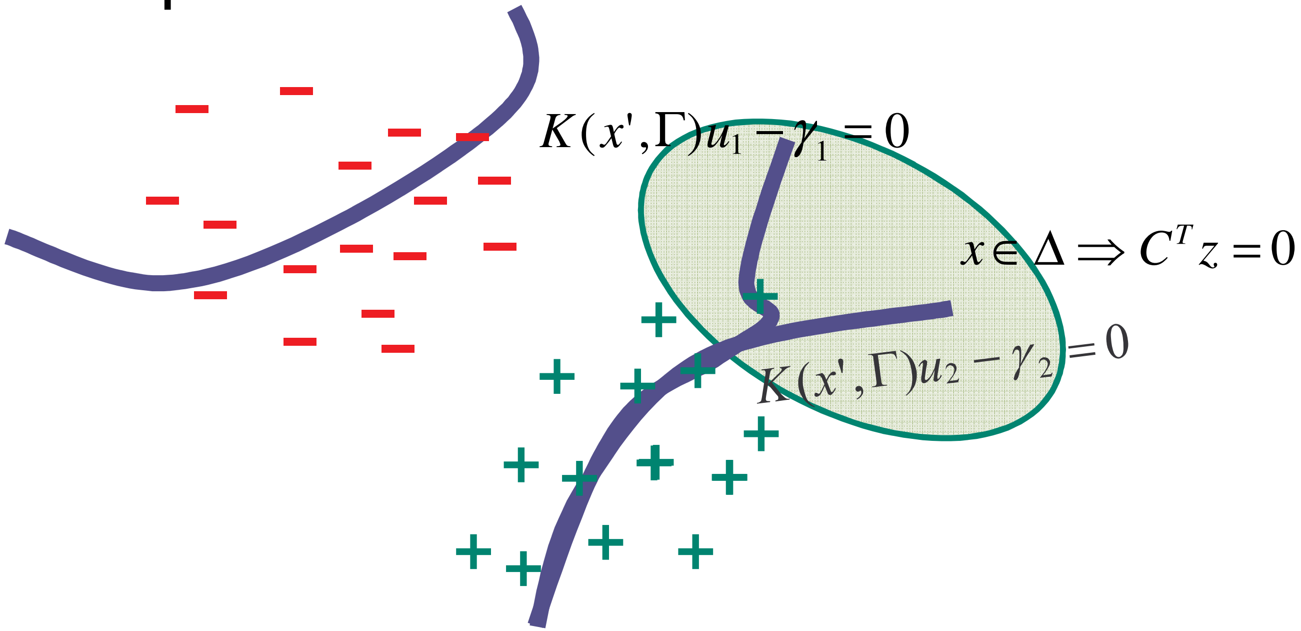
M.R. Guarracino, C. Cifarelli, O. Seref, P. Pardalos. A Classification Method Based on Generalized Eigenvalue Problems, OMS, 2007.

Prior knowledge in ReGEC

- □ It is possible to extend prior knowledge to Regularized Generalized Eigenvalue Classifier (ReGEC).
- □ The new algorithm halves the missclassification error of the original method.
- □ The idea of increasing the information contained in the training set with additional knowledge is appealing for biomedical data.
- □ The experience of field experts or previous results can be readily transferred to new problems.

Prior knowledge in ReGEC

- Let Δ be the set of points in B describing a priori knowledge, constraint matrix C represents knowledge imposed on class B :


$K(x',\Gamma)u_1-\gamma_1=0$
$x\in\Delta\Rightarrow C^Tz=0$
$K(x',\Gamma)u_2-\gamma_2=0$

- Constraint imposes all points in Δ to have zero distance from the plane => to belong to B

Prior knowledge in ReGEC

- □ Prior knowledge can be expressed in terms of orthogonality of the solution to a chosen subspace:

$$C^T z = 0$$

where C is a n × p matrix of rank r, with r < p < n

- □ The constrained eigenvalue problem with prior knowledge for points in class B is:

$$\min_{z \neq 0} \frac{z' G z}{z' H z},$$

$$s.t. \quad C^T z = 0$$

Radial Basis Function Neural Networks

- A RBF network is divided into two operative blocks: an inner hidden layer, and the output layer.
- The hidden layer, as it is based on neurons with a radial basis activation function, creates a response localized on the input vector x; the binary output will then be calculated as a weighted sum of these localized responses.
- Training a RBF network is a procedure divided into two phases:
  1. With an unsupervised learning technique, the parameters of the radial basis function are calculated.
  2. Values of the weights w, which determine the binary output y, are then computed.

RBF network parameter estimation

- Traditionally there are two strategies for this first phase of unsupervised learning.
- The classic strategy calculates these parameters through different clustering techniques.
- These aim to divide the training set into a fixed amount of homogeneous groups, organized according to the distance of the points in the training set.
- Besides clustering, it is possible to have an incremental approach.
- In this way, one seeks to reduce the mean quadratic error under a threshold ϵ by adding nodes to the hidden layer.

RBF network weights estimation

- In the second part of the training, we search for values of the weights w which determine the binary output y.
- Such weights are calculated by minimizing the following error function:

$$E = \frac{1}{2}\sum_{i=1}^{m}(y(X_{i.}) - c_i)^2$$

which tells the distance of the actual solution from the desired one.

- Prior knowledge is added by a modification to this phase.

Prior knowledge in RBF NN

- Prior knowledge is then added as a set of constraints to obtain the following minimization problem:

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{i=1}^{m}(y(X_{i.}) - c_i)^2 \qquad (9) \\ s.t. \quad & Bx \geq 0. \end{aligned}$$

- The constraints of this problem force the hyperplane solution of the equation (9) to pass through the m points represented by the matrix $B \in \mathbb{R}^{m \times n}$.
- Algebraically, this means the solution has to be searched in the subspace generated by prior knowledge points.

Knowledge as a mining task

- Is it possible to choose a method to discover knowledge in the training data, using a learning method consistently different from SVM?
- Logic mining method *Lsquare,* combined with a feature selection based on integer programming, has been used to extract logic formulas from the data.
- The most meaningful portions of such formulas represent prior knowledge for ReGEC.

Knowledge discovery for ReGEC

- Results exhibit an increase in the recognition capability of the system
- We propose a combination of two very different learning methods:
  - ReGEC, that operates in a multidimensional Euclidean space, with highly nonlinear data transformation, and
  - Logic Learning, that operates in a discretized space with models based on propositional logic
- The former constitutes the master learning algorithm, while the latter provides the additional knowledge

Logic formulas

- The additional knowledge for ReGEC is extracted from training data with a logic mining technique
- Such choice is motivated by two main considerations:
  1. the nature of the method is intrinsically different from the ReGEC adopted as primary classifier;
  2. the logic formulas are, semantically, the form of ``knowledge" closest to human reasoning and therefore resemble at best contextual information.
- The logic mining system consists of two main components, each characterized by the use of integer programming models

The Logic Formulas Miner

- Builds logic separations in *Disjunctive Normal Form* (DNF)
- Identifies iteratively the clauses of the DNF that separates the largest part of object in one class from all the objects of the other class
- Clause identification is based on the solution of a *Minimum Cost Satisfiability Problem* (MINSAT), computationally hard

$$s_i = \begin{cases} 1 & p_i = True\,,\ q_i = False \\ -1 & p_i = False,\ q_i = True \\ 0 & p_i = q_i = False \end{cases}$$

	S_1	S_2	S_3	S_4
A	T	T	F	?
A	T	F	F	T
A	T	F	F	F
I	T	T	T	?
I	F	?	F	T

$q_1 \vee q_2 \vee p_3 \vee p_4 \vee q_4$

$q_1 \vee p_2 \vee p_3 \vee q_4$

$q_1 \vee p_2 \vee p_3 \vee p_4$

$\neg q_1 \vee d_1,\ \neg q_2 \vee d_1,\ \neg q_3 \vee d_1, \neg q_4 \vee d_1, \neg p_4 \vee d_1$

$\neg p_1 \vee d_2,\ \neg p_2 \vee d_2, \neg q_2 \vee d_2, \neg p_3 \vee d_1, \neg q_4 \vee d_1$

$\neg p_i \vee \neg q_i$

Satisfying solution:

p_1 *True*, q_1 *False*

p_2 *False*, q_2 *False*

p_3 *False*, q_3 *True*

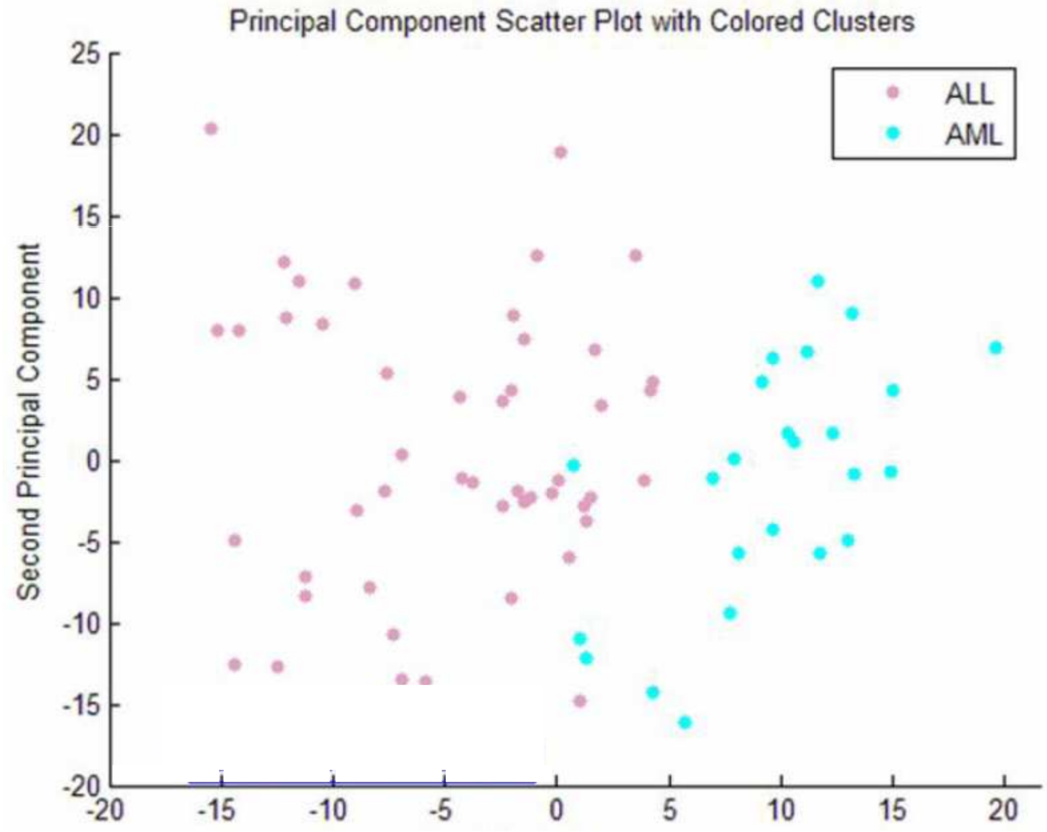
p_4 *False*, q_4 *False*

$$S_1 \wedge \neg S_3$$

P. Bertolazzi, G. Felici, P. Festa, G. Lancia. Logic classification and feature selection for biomedical data, Computer and Mathematics, 2008.

Acute Leukemia data

- Golub microarray dataset (Science, 1999)
- The microarray data have 72 samples with 7129 gene expression values
- Data contain 25 Acute Myeloid Leukemia and 47 Acute Lymphoblastic Leukemia samples



Logic Formulas

- The dataset has been discretized and the logic formulas have been evaluated. Those formulas are in the form:

  $IF\ p(4196) > 3.435\ AND\ p(6041) > 3.004\ THEN\ class1,$

  $IF\ p(6573) < 2.059\ AND\ p(6685) > 2.794\ THEN\ class1,$

  $IF\ p(1144) > 2.385\ AND\ p(4373) < 3.190\ THEN\ class - 1,$

  $IF\ p(4847) < 3.006\ AND\ p(6376) < 2.492\ THEN\ class - 1,$

  where p(i) represents the i-th probe.

- The knowledge region for each class, are those given by the intersection of all chosen formulas.

Classification accuracy

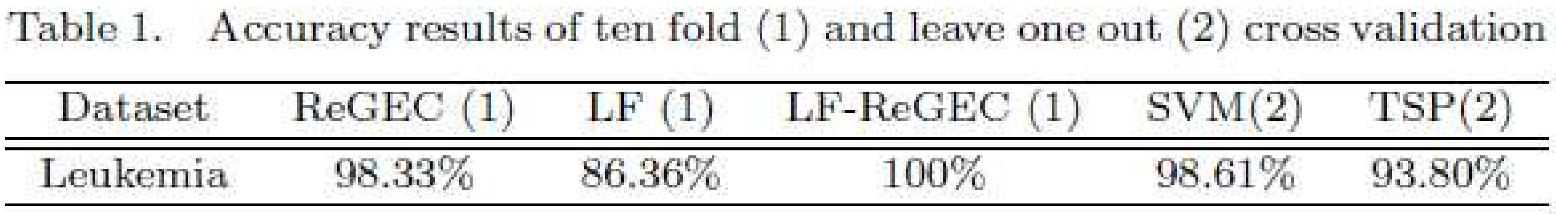


Table 1. Accuracy results of ten fold (1) and leave one out (2) cross validation

Dataset	ReGEC (1)	LF (1)	LF-ReGEC (1)	SVM(2)	TSP(2)
Leukemia	98.33%	86.36%	100%	98.61%	93.80%

- Leave one out cross validation used for ReGEC.
- The ReGEC method with prior knowledge found with LF becomes fully accurate on the dataset.

Microarray experiments

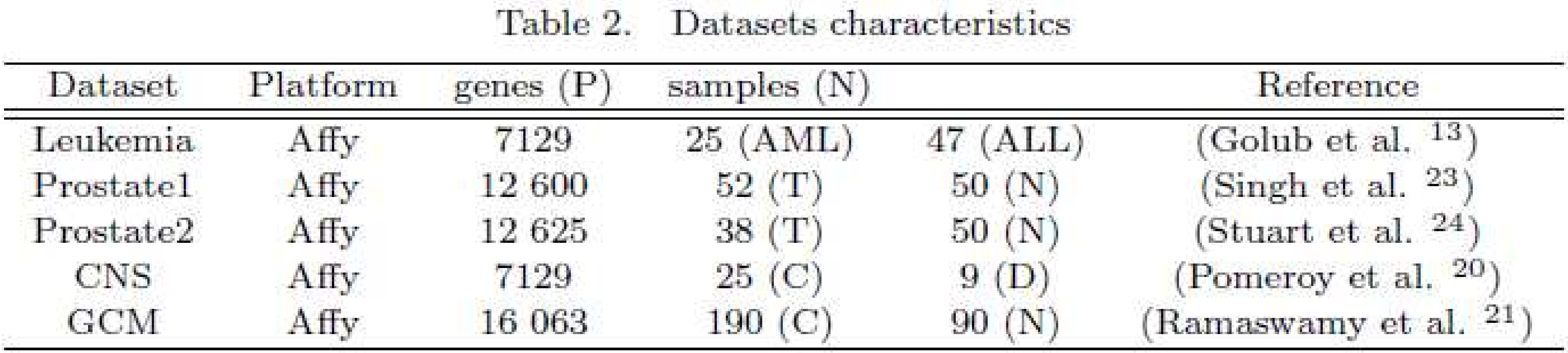


Table 2. Datasets characteristics

Dataset	Platform	genes (P)	samples (N)		Reference
Leukemia	Affy	7129	25 (AML)	47 (ALL)	(Golub et al. [13])
Prostate1	Affy	12 600	52 (T)	50 (N)	(Singh et al. [23])
Prostate2	Affy	12 625	38 (T)	50 (N)	(Stuart et al. [24])
CNS	Affy	7129	25 (C)	9 (D)	(Pomeroy et al. [20])
GCM	Affy	16 063	190 (C)	90 (N)	(Ramaswamy et al. [21])

- Results regard its performance in terms of classification accuracy.

Accuracy results

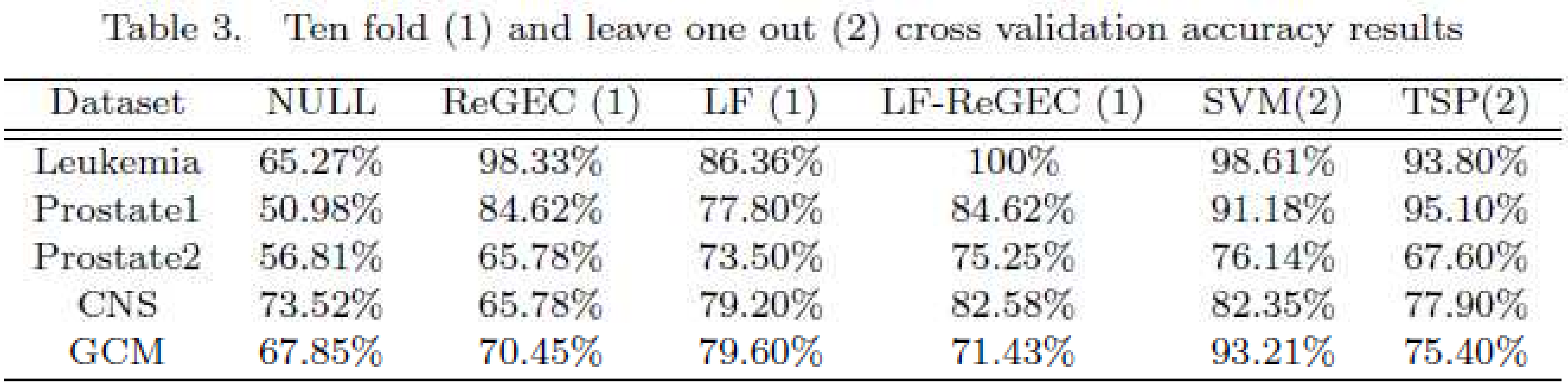


Table 3. Ten fold (1) and leave one out (2) cross validation accuracy results

Dataset	NULL	ReGEC (1)	LF (1)	LF-ReGEC (1)	SVM(2)	TSP(2)
Leukemia	65.27%	98.33%	86.36%	100%	98.61%	93.80%
Prostate1	50.98%	84.62%	77.80%	84.62%	91.18%	95.10%
Prostate2	56.81%	65.78%	73.50%	75.25%	76.14%	67.60%
CNS	73.52%	65.78%	79.20%	82.58%	82.35%	77.90%
GCM	67.85%	70.45%	79.60%	71.43%	93.21%	75.40%

- □ LF method is more accurate than TSP in three cases out of five.
- □ In all cases, LF-ReGEC, produces equal or higher accuracy results.

Conclusion

□ Microarrays experiments produce challenging datasets.

□ Available classification methods provide results affected by noisy and incomplete data.

□ Omics science problems require decisions based on incomplete and uncertain data.