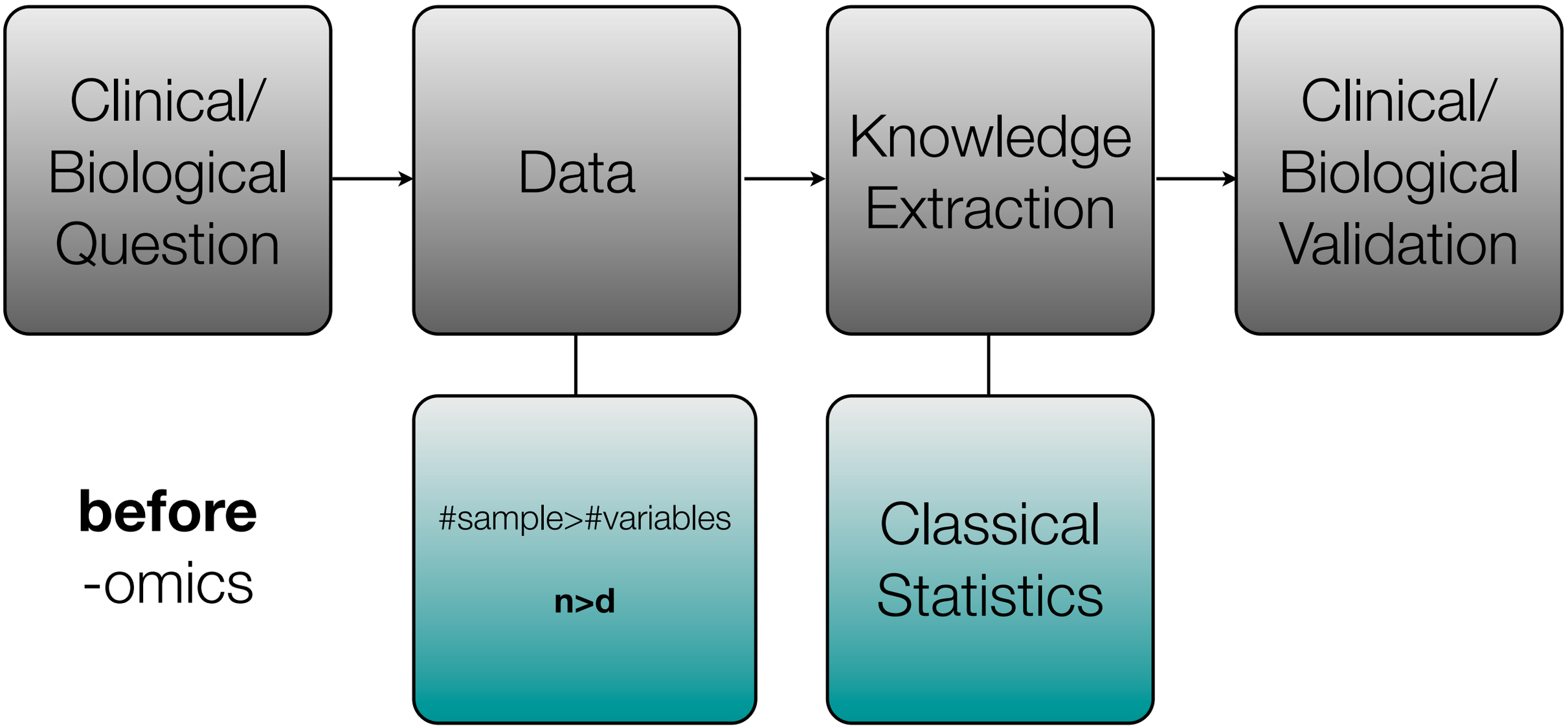
Machine learning approaches for molecular data analysis

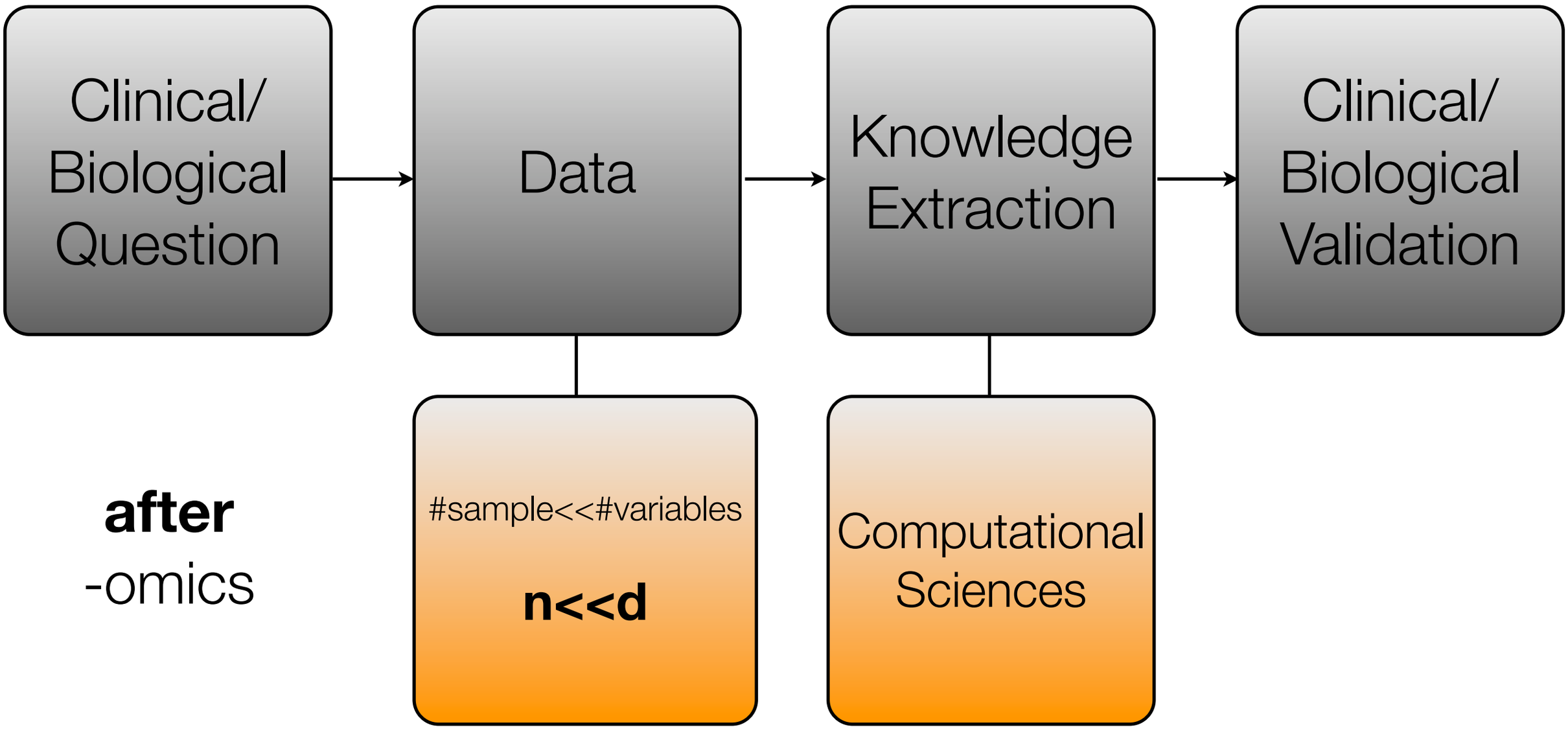
Annalisa Barla

Medicine and Genomic Medicine



before
-omics

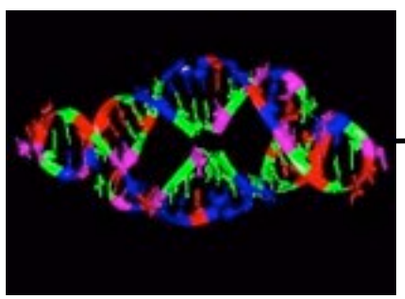
Medicine and Genomic Medicine



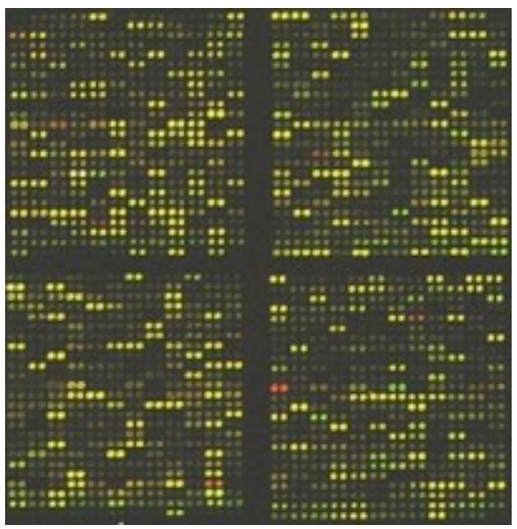
after
-omics

Genomic Medicine: dealing with high-throughput data

- Typical scenario is **n<<d**
- Number of samples is **limited** (e.g. rare diseases and expensive technology)
- (mostly) High-throughput data
  - new technologies (DNA microarrays, CGH, SNP, etc.)
  - possibility to measure the whole genome
  - most of the times the data are noisy (getting better any day now..)



biological samples



microarray gene expression

computational methods

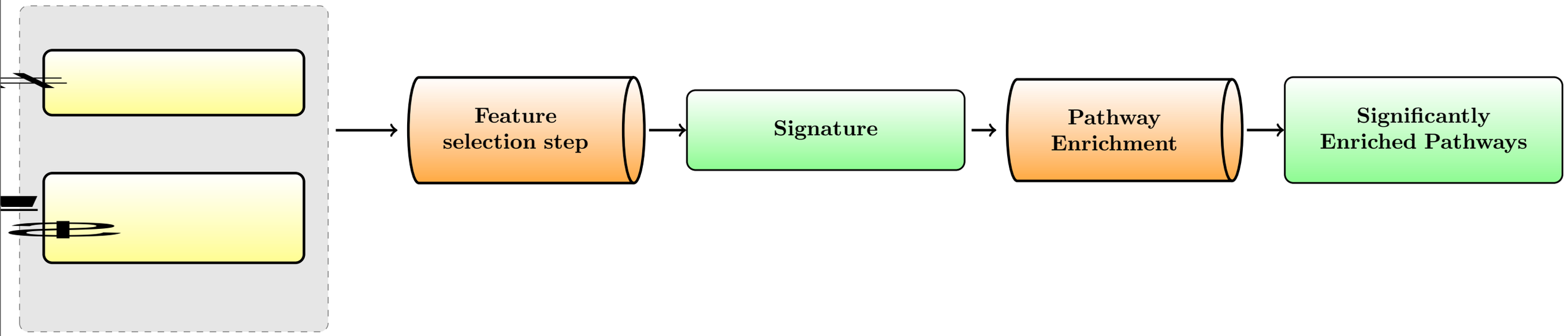
Relevant Gene List

230746_s_at	STC1
230710_at	---
230630_at	AK3L2
228499_at	PFKFB4
228483_s_at	TAF9B
227337_at	ANKRD37
227068_at	PGK1
226632_at	CYGB
226452_at	PDK1
226348_at	---
226347_at	---

selected genes

Potential biomarker Identification

Functional characterization

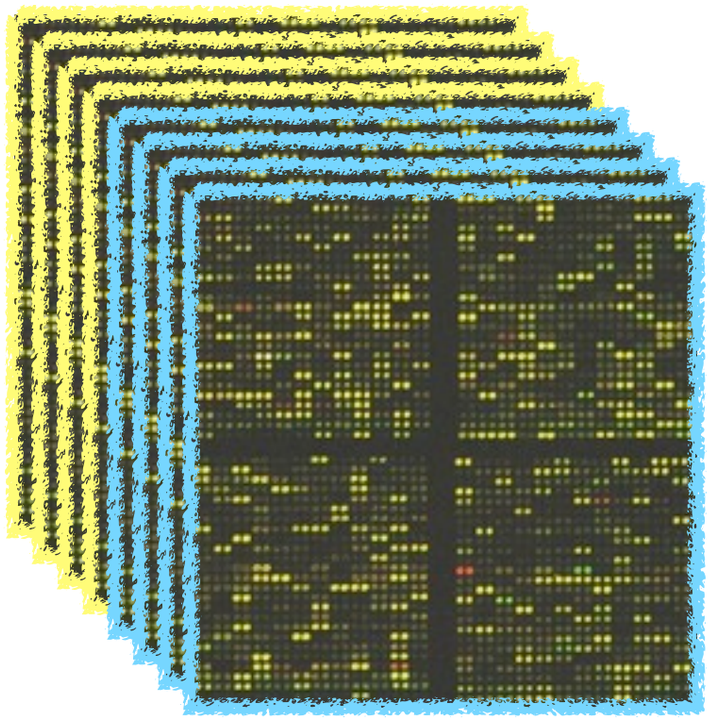


Analysis Workflow

Feature Selection Step

Statistical Analysis

n measures



n phenoytpes

healthy
healthy
healthy
diseased
diseased
diseased

data matrix

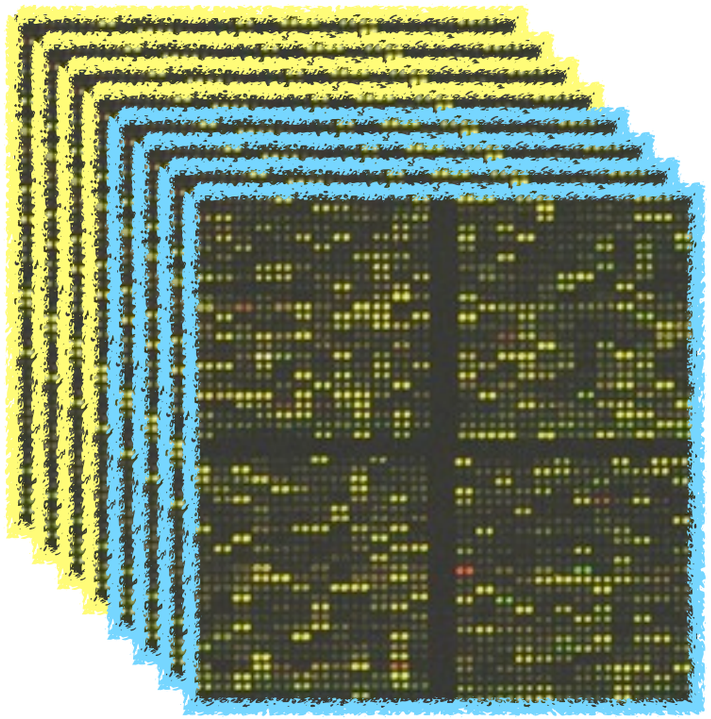
0.5779	0.5035	0.0938	0.4064	0.8184	0.7848	+1
0.3457	0.4131	0.5515	0.0046	0.6842	0.6159	+1
0.8035	0.6612	0.0870	0.1205	0.7088	0.5677	+1
0.8568	0.0304	0.9938	0.6638	0.0162	0.5096	+1
0.4309	0.9815	0.4585	0.5874	0.2894	0.7539	+1
0.0616	0.0028	0.7594	0.9018	0.0610	0.8240	+1
0.0356	0.5365	0.7559	0.1312	0.6758	0.8992	-1
0.5348	0.8569	0.4018	0.7751	0.8999	0.9637	-1
0.8493	0.6705	0.3569	0.4694	0.3314	0.3283	-1
0.0924	0.5027	0.6009	0.9251	0.1063	0.8210	-1

labels

knowledge extraction

Statistical Analysis

n measures



n phenoytpes

healthy
healthy
healthy
diseased
diseased
diseased

data matrix

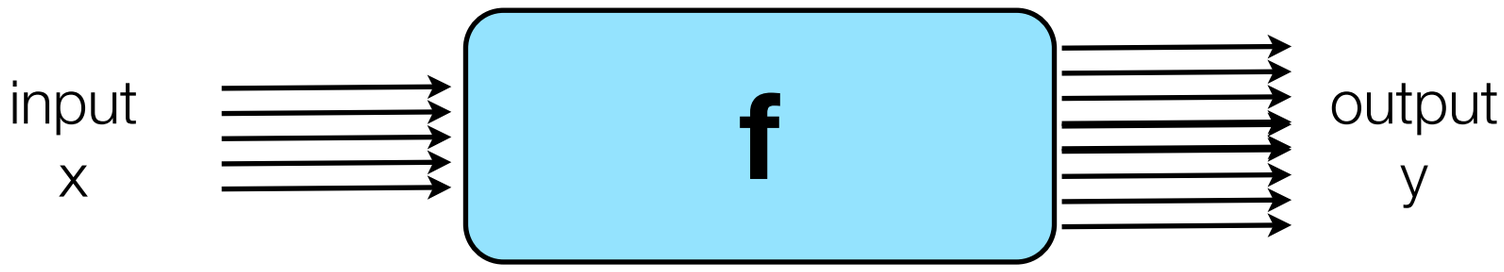
0.5779	0.5035	0.0938	0.4064	0.8184	0.7848	+1
0.3457	0.4131	0.5515	0.0046	0.6842	0.6159	+1
0.8035	0.6612	0.0870	0.1205	0.7088	0.5677	+1
0.8568	0.0304	0.9938	0.6638	0.0162	0.5096	+1
0.4309	0.9815	0.4585	0.5874	0.2894	0.7539	+1
0.0616	0.0028	0.7594	0.9018	0.0610	0.8240	+1
0.0356	0.5365	0.7559	0.1312	0.6758	0.8992	-1
0.5348	0.8569	0.4018	0.7751	0.8999	0.9637	-1
0.8493	0.6705	0.3569	0.4694	0.3314	0.3283	-1
0.0924	0.5027	0.6009	0.9251	0.1063	0.8210	-1

labels

knowledge extraction

Learning from examples paradigm

the GOAL is not to memorize but to GENERALIZE, e.g. predict



given a set of **examples:**

$$\{(x_1,y_1), (x_2,y_2),...., (x_n,y_n)\}$$

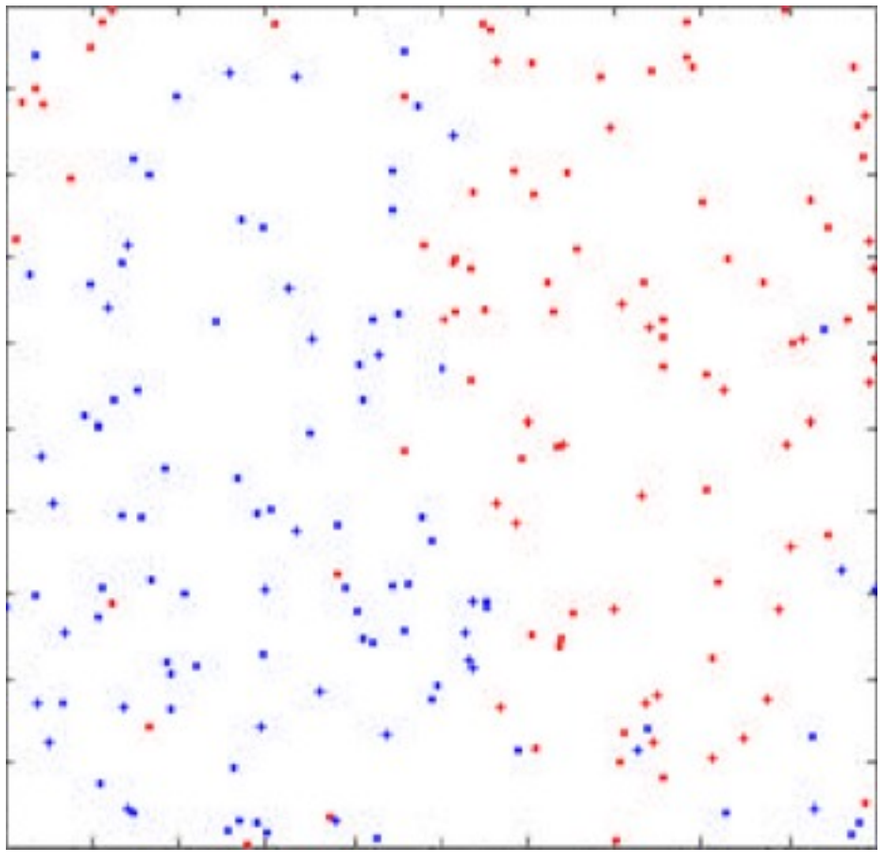
find a function:

$$f(x) \sim y$$

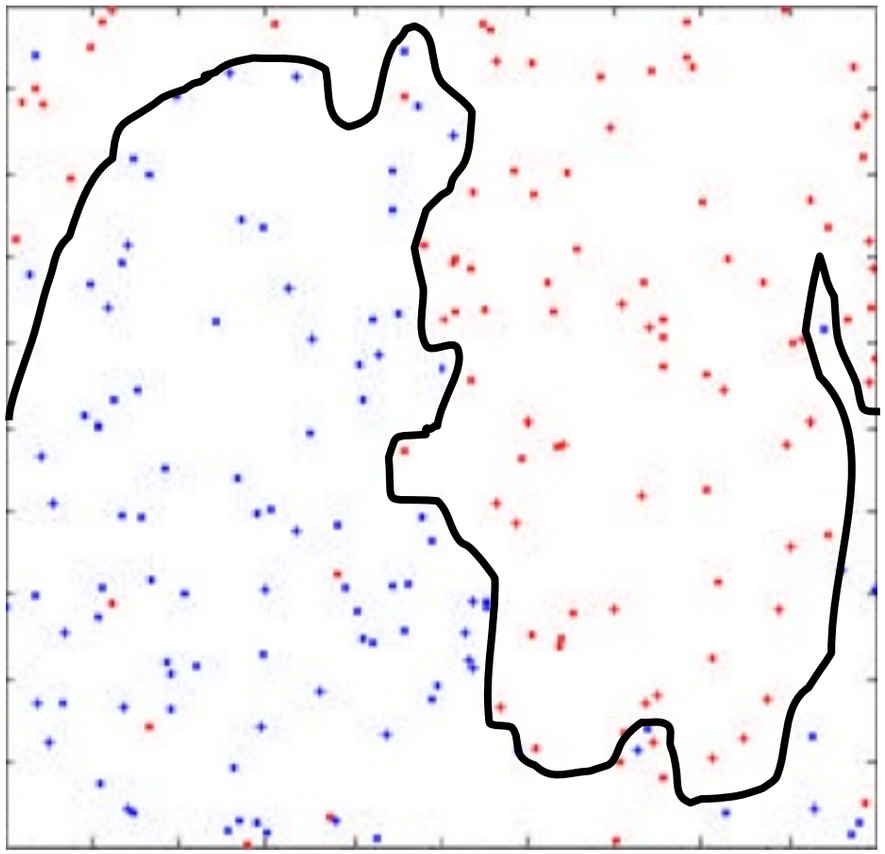
such that **f** is a **good predictor on new data** as well as on the given dataset

and possibly identify the most discriminating variables
(**gene signature)**

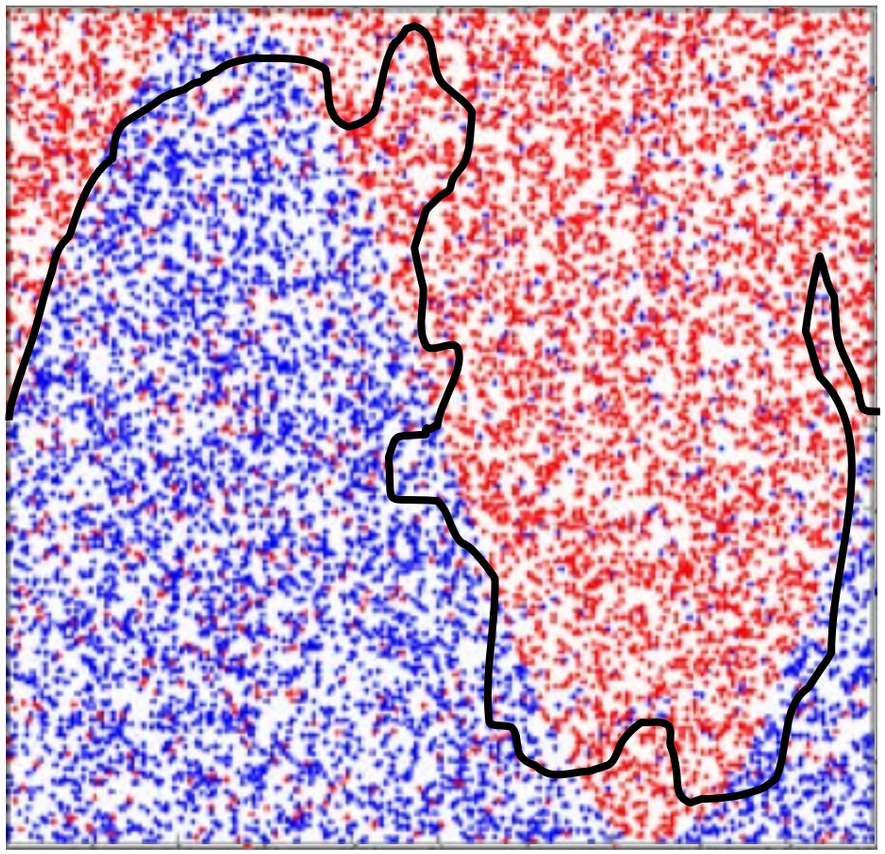
Probabilistic Nature of our problem



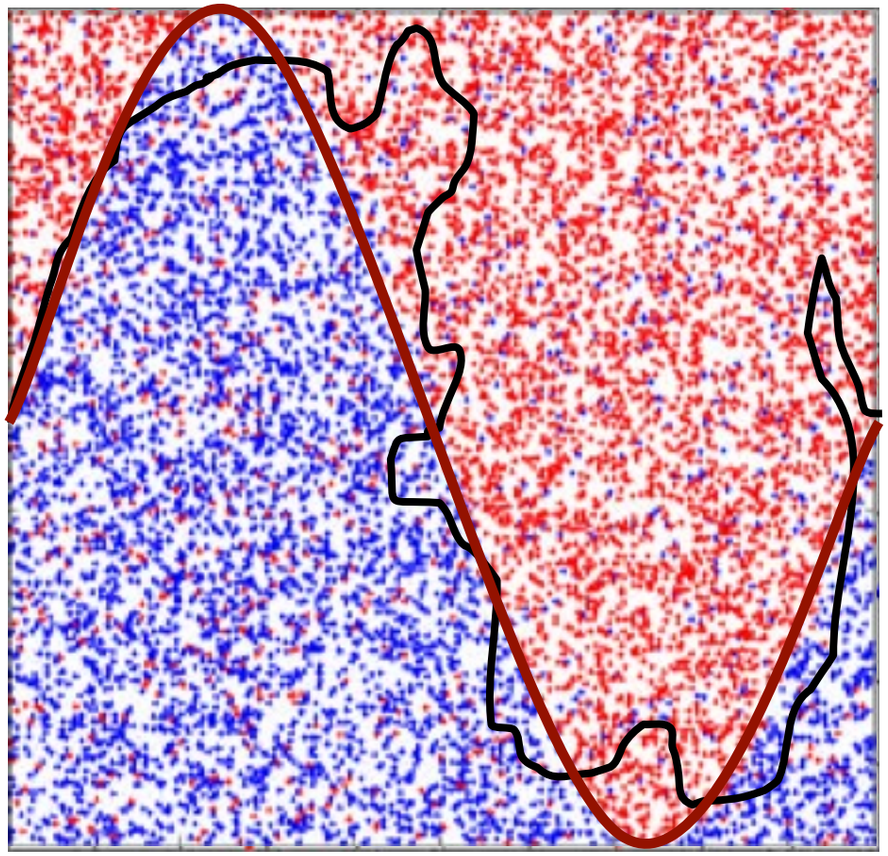
Probabilistic Nature of our problem



Probabilistic Nature of our problem



Probabilistic Nature of our problem



Feature Selection

- Search problem in a space of feature subsets
- Alleviating the effect of the **curse of dimensionality**.
- Enhancing **generalization capability**.
- **Speeding up** learning process.
- Improving model **interpretability**.

see also I.Guyon lectures available online

Feature Selection Methods

see also I.Guyon lectures available online

Feature Selection Methods

Filter

select subsets of variables as a pre-processing step

see also I.Guyon lectures available online

Feature Selection Methods

Filter	Wrapper
↓	↓
select subsets of variables as a pre-processing step	**assess the relevance of a feature subset according to the prediction performance of a learning machine**

see also I.Guyon lectures available online

Feature Selection Methods

Filter	Wrapper	Embedded
select subsets of variables as a pre-processing step	**assess the relevance of a feature subset according to the prediction performance of a learning machine**	**incorporate variable selection as part of the training process**

see also I.Guyon lectures available online

Feature Selection Methods

Filter	Wrapper	Embedded
Statistical scores (Differentially expressed) t-test ANOVA Wilcoxon Pearson **Mutual Information** distribution entropy **Single variable classifiers**	**Forward elimination** **Backward elimination** SVM-RFE SVM-ERFE ...	**FS incorporated into classification** decision trees **l1l2** LASSO/LARS Adaboost

Feature Selection Methods

Journal of Machine Learning Research 3 (2003) 1157-1182 Submitted 11/02; Published 3/03

An Introduction to Variable and Feature Selection

Isabelle Guyon ISABELLE@CLOPINET.COM
Clopinet
955 Creston Road
Berkeley, CA 94708-1501, USA

André Elisseeff ANDRE@TUEBINGEN.MPG.DE
Empirical Inference for Machine Learning and Perception Department
Max Planck Institute for Biological Cybernetics
Spemannstrasse 38
72076 Tübingen, Germany

Editor: Leslie Pack Kaelbling

BIOINFORMATICS **REVIEW** Vol. 23 no. 19 2007, pages 2507–2517
doi:10.1093/bioinformatics/btm344

Gene expression

A review of feature selection techniques in bioinformatics

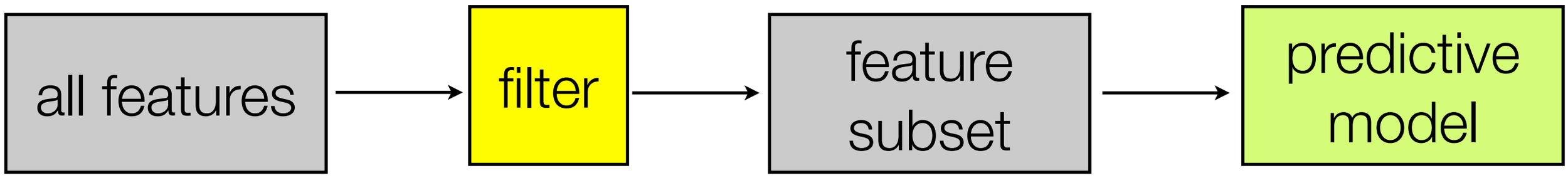
Yvan Saeys[1,*], Iñaki Inza[2] and Pedro Larrañaga[2]

Filter Approaches

- Filter methods do not incorporate learning: they are based on an evaluation function that relies solely on **properties of the data**, thus is independent on any particular algorithm

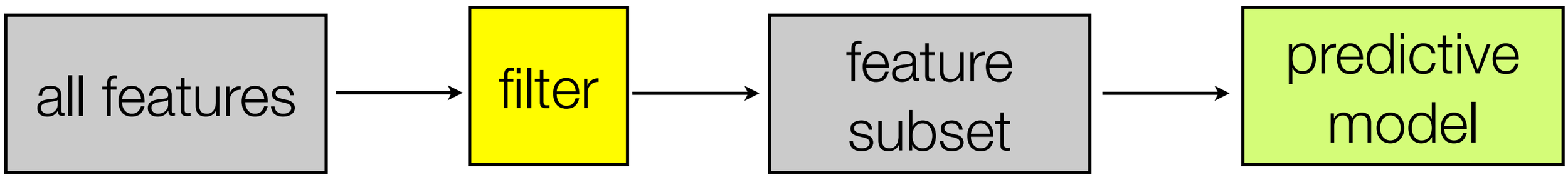
- Filter methods are **fast**

- Usually based on classical statistical techniques and often univariate



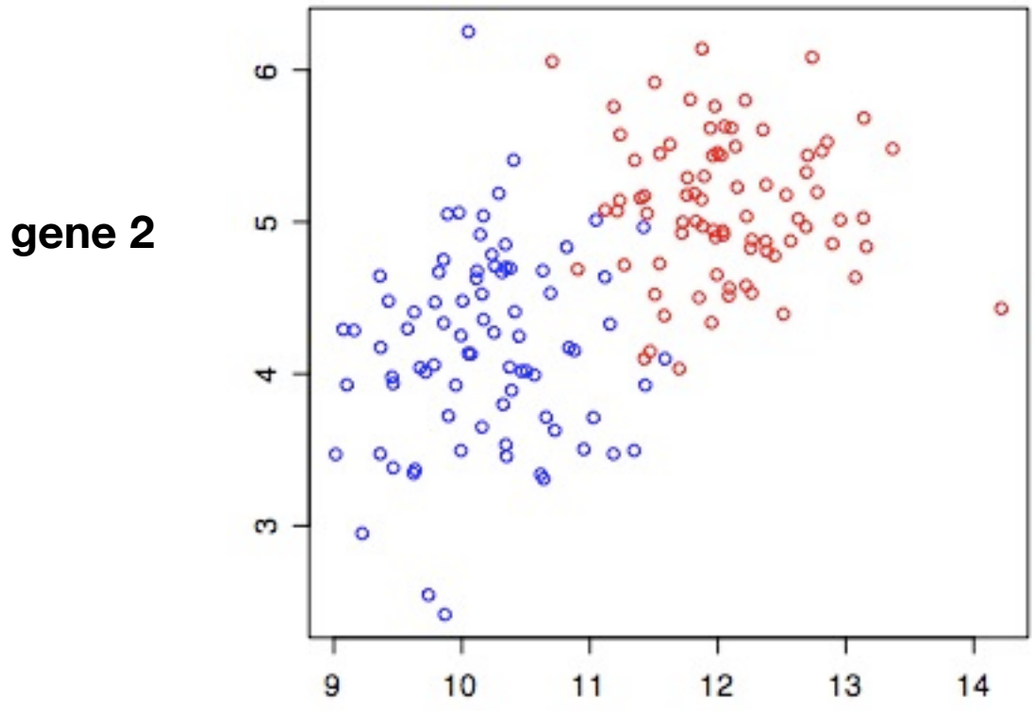
Filter Approaches

- **Criterion**: Measure feature/feature subset *relevance*
- **Search:** Usually sort features (individual feature ranking or nested subsets of features)
- **Assessment:** By means of statistical tests
- **PRO:** Are (relatively) robust against overfitting
- **CON:** May fail to select the most meaningful features



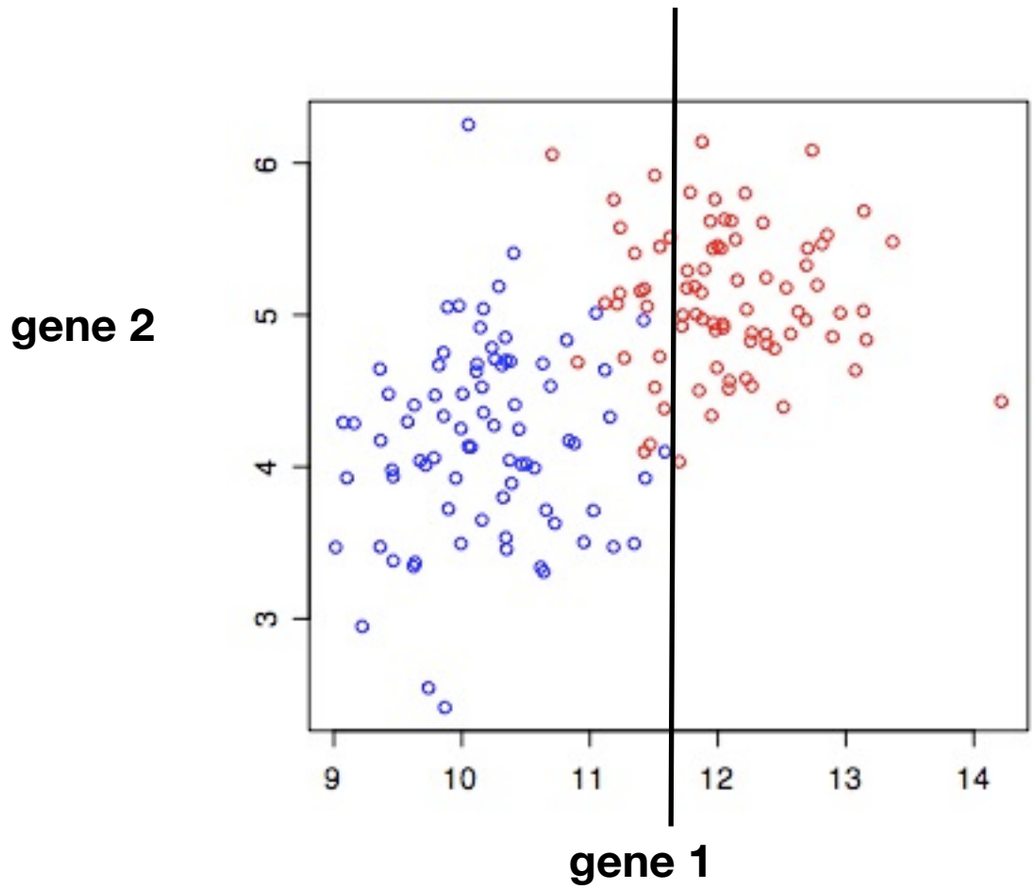
Why going multivariate?

search for **DIFFERENTIALLY EXPRESSED GENES** is not always sufficient!
univariate approaches may not be flexible enough...



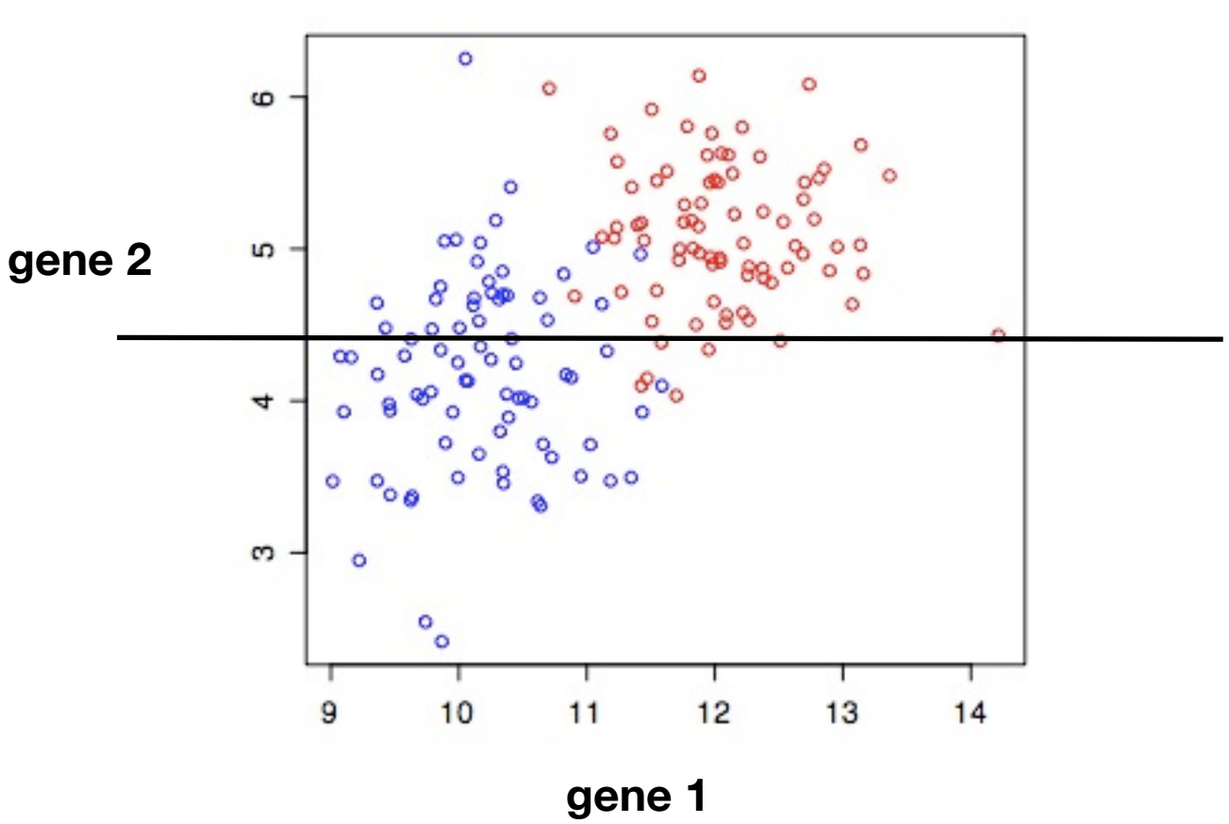
Why going multivariate?

search for **DIFFERENTIALLY EXPRESSED GENES** is not always sufficient!
univariate approaches may not be flexible enough...



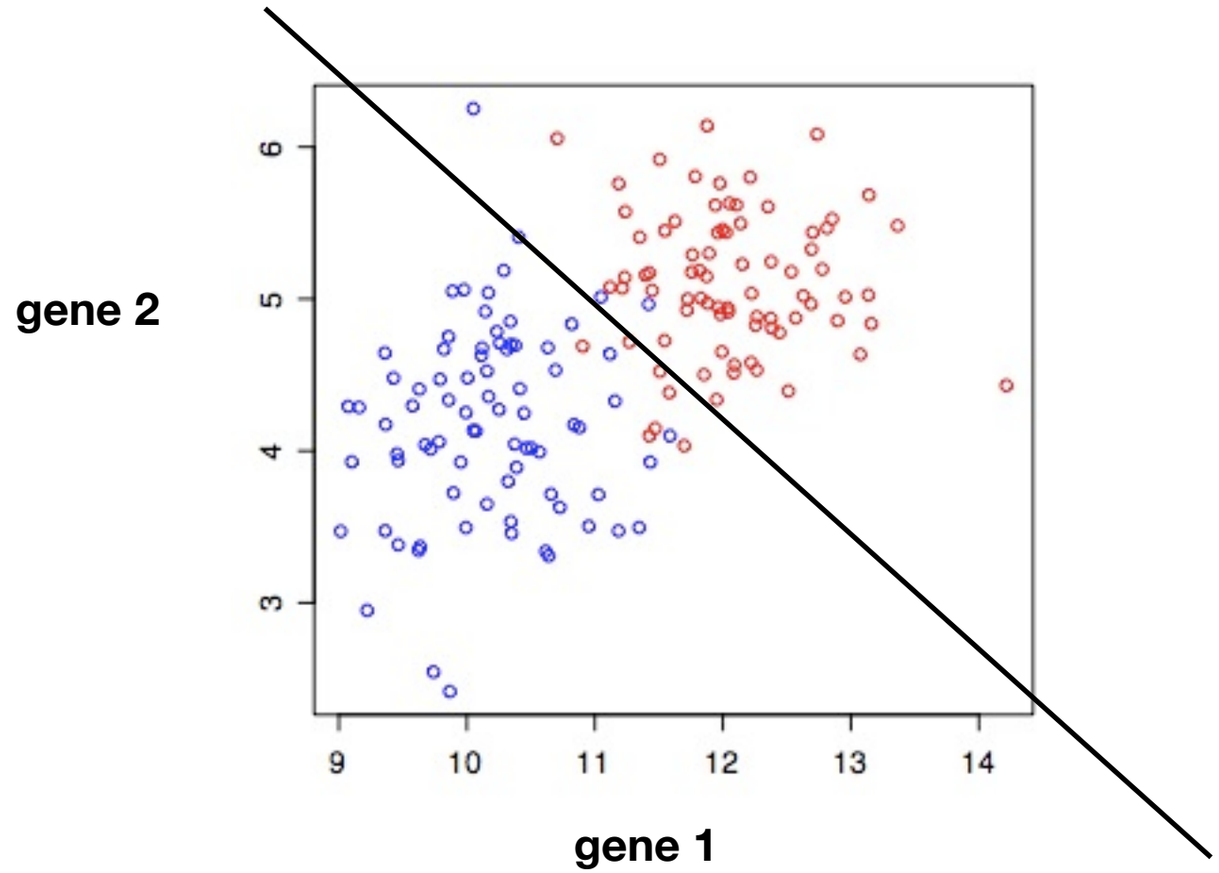
Why going multivariate?

search for **DIFFERENTIALLY EXPRESSED GENES** is not always sufficient!
univariate approaches may not be flexible enough...



Why going multivariate?

search for **DIFFERENTIALLY EXPRESSED GENES** is not always sufficient!
univariate approaches may not be flexible enough...



Why going multivariate?

Why going multivariate?

"You cannot be serious!"

(J.McEnroe, Wimbledon 1981)

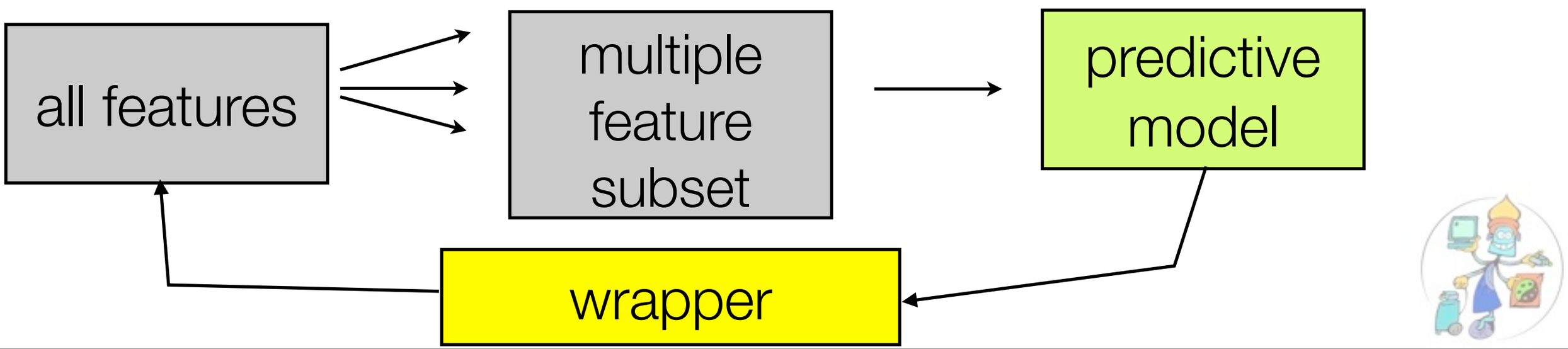
Why going multivariate?

- Most of the known diseases are of system nature
- Univariate methods may neglect the interplay among biologically related variables
- The final aim is the understanding of the **molecular pathways** (from the transcription to the signaling inside the cells).

Wrapper Approaches

- **Wrapper methods** use a learning machine to measure the quality of subsets of features
- They do not incorporate knowledge about the specific structure of the classification or regression function, and **can therefore be combined with any learning machine**:

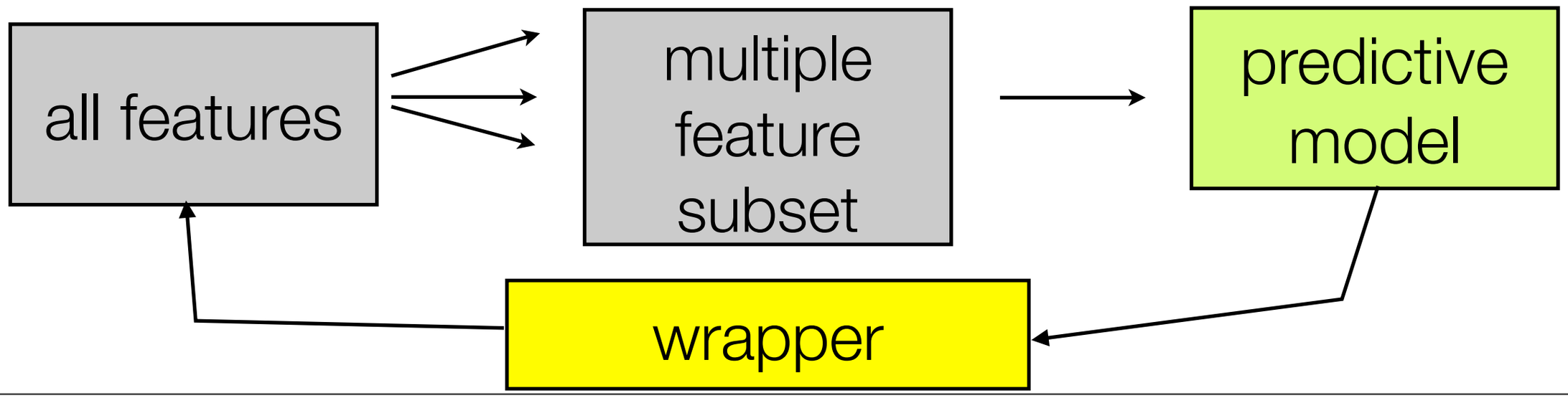
1. a classifier is trained
2. it obtains an estimation of the accuracy in predicting a class label that is known
3. if the accuracy is **good** then the subset of features is retained



Wrapper Approaches

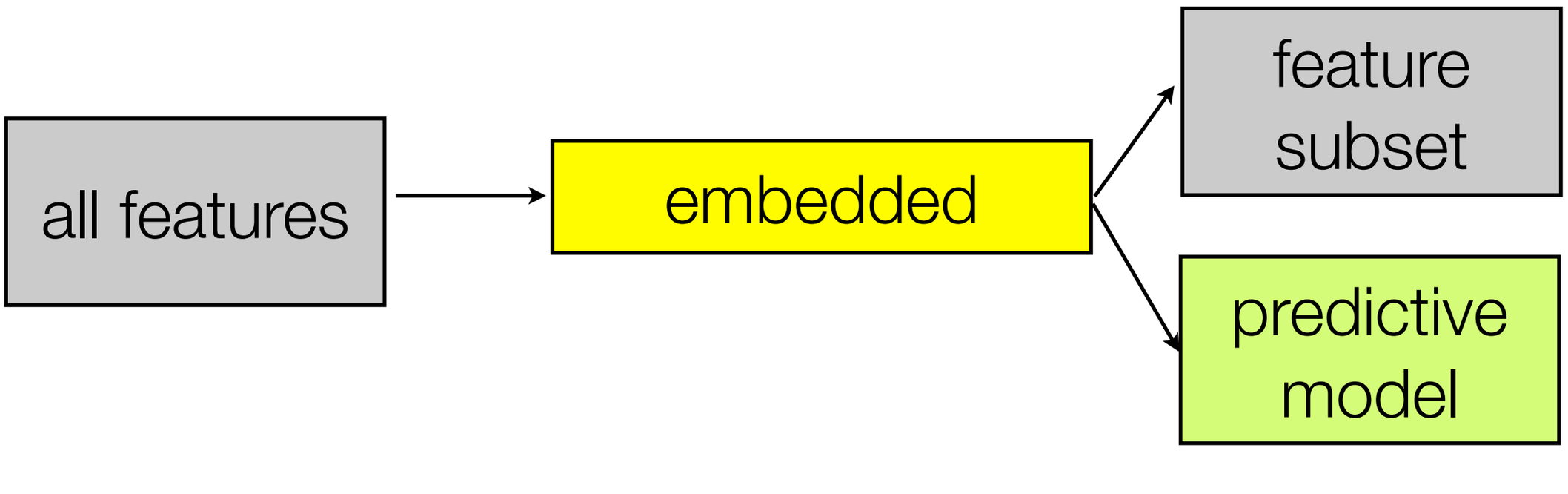
- **Criterion:** Measure feature subset **prediction ability** *(usefulness)*
- **Search:** Search the space of all feature subsets
- **Assessment:** Use cross-validation

- **PRO:** Can in principle find the most meaningful features
- **CON:** Are prone to overfitting



Embedded Approaches

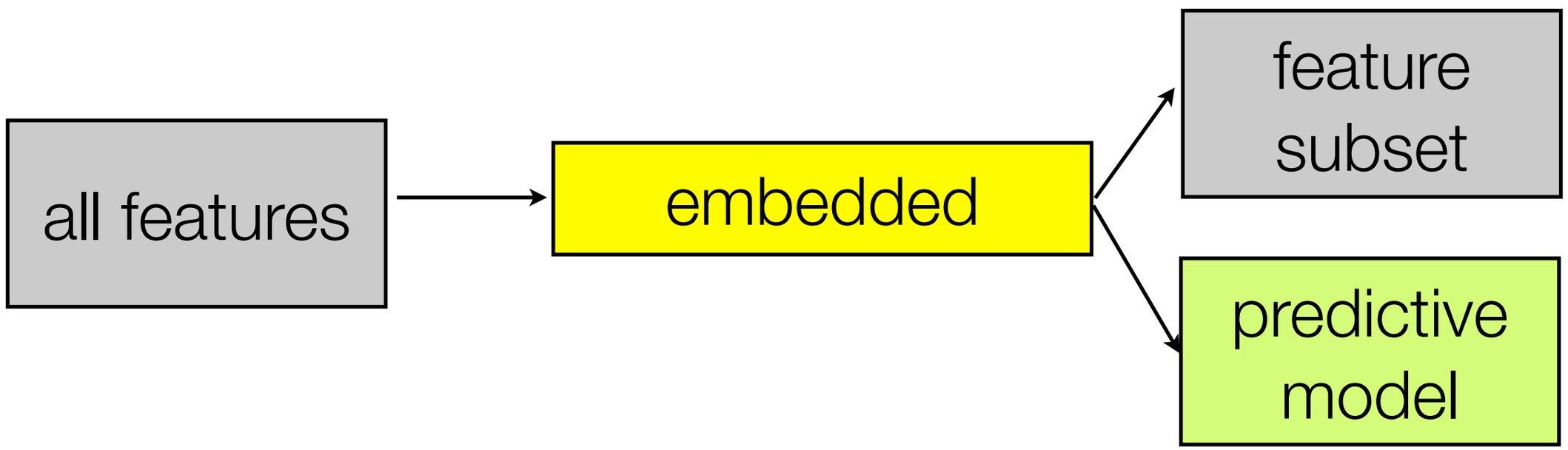
- The learning part and the feature selection part can not be separated



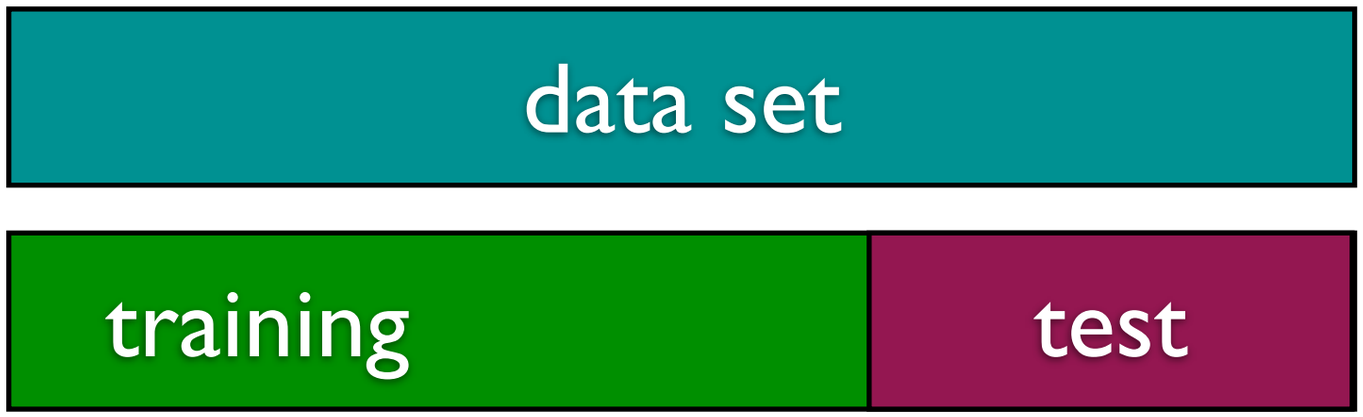
Embedded Approaches

- **Criterion:** Measure feature subset "usefulness"
- **Search:** Search guided by the learning process
- **Assessment:** Use cross-validation

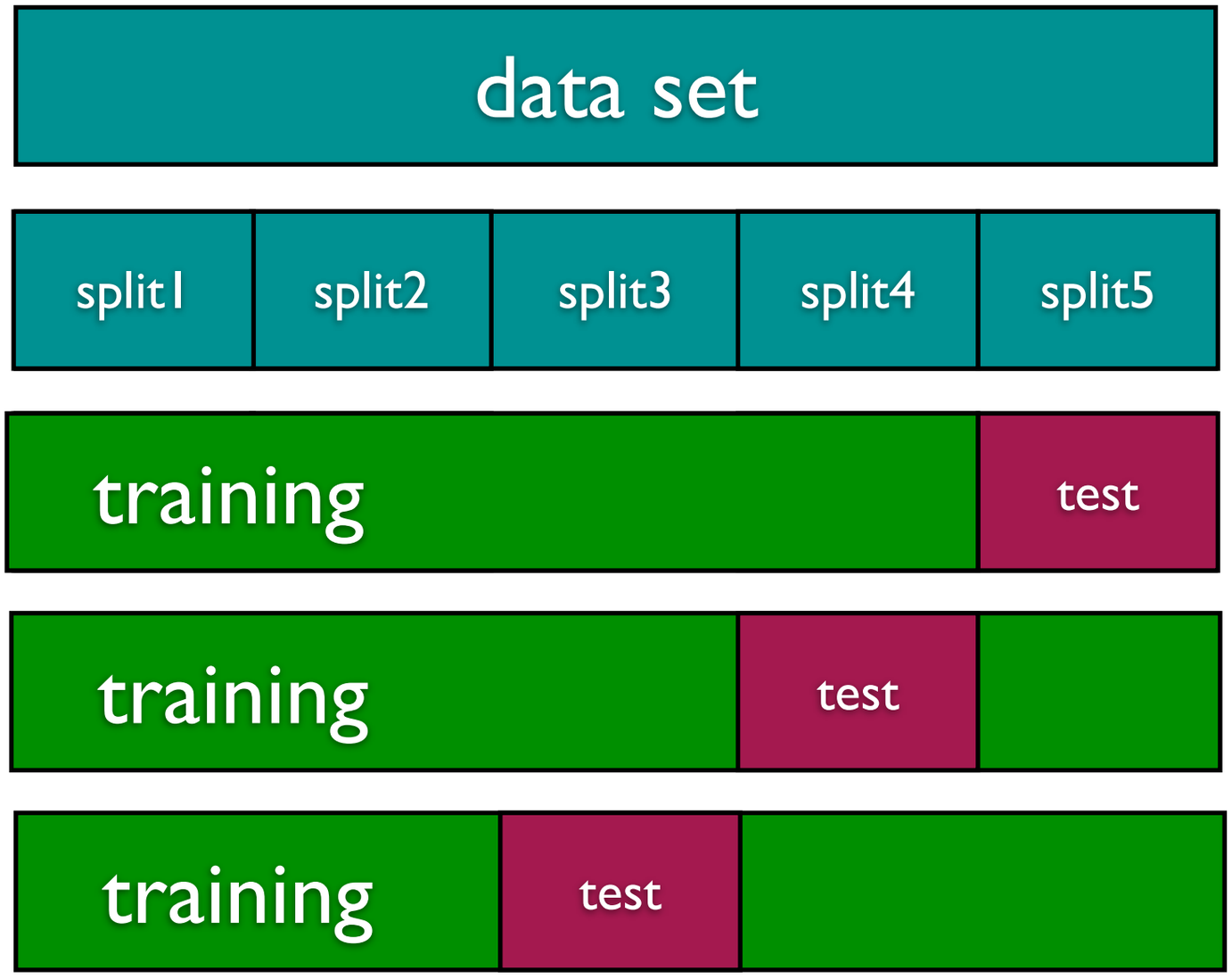
- **PRO:** Less prone to overfitting than wrappers
- **CON:** Need many training data



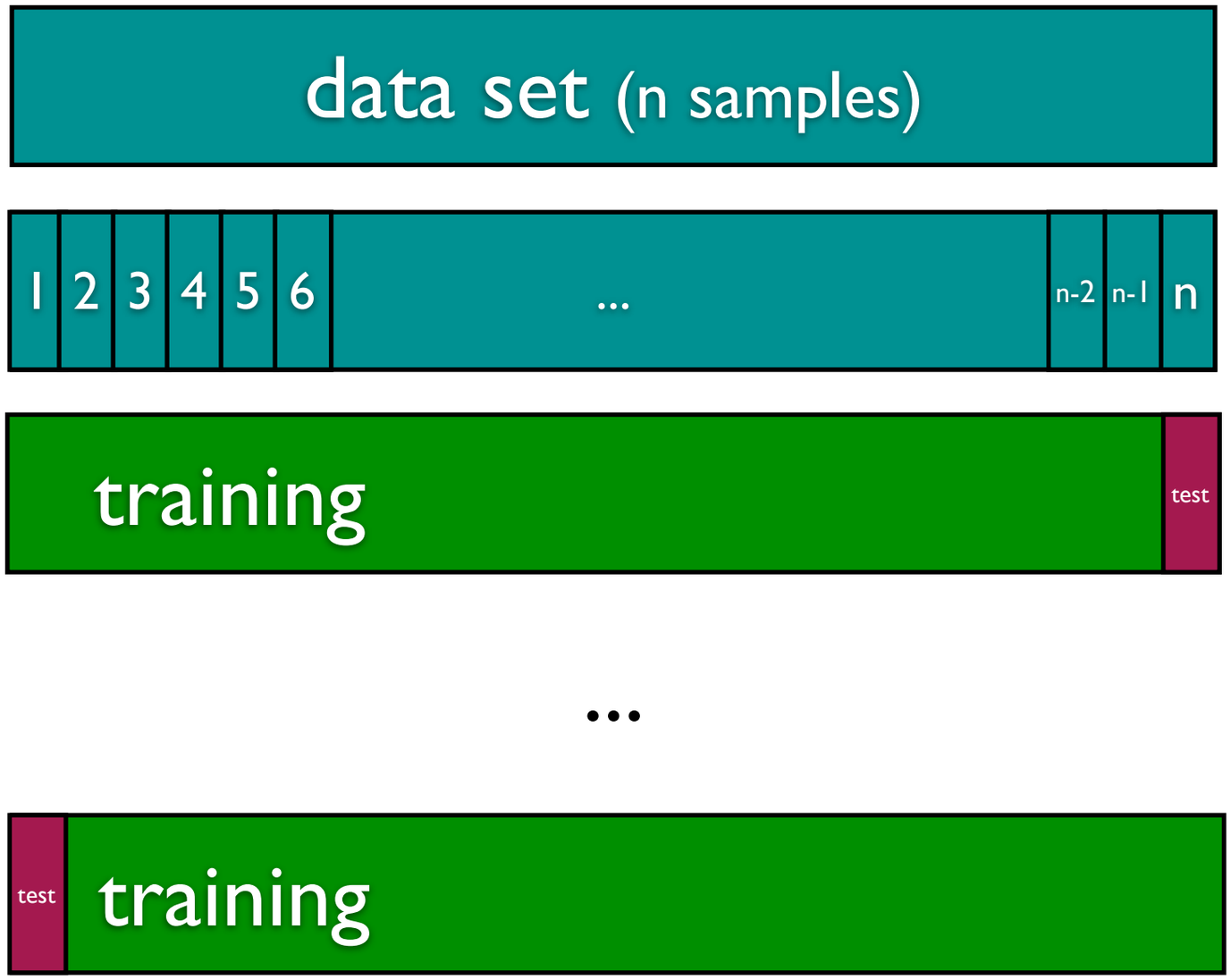
Result Assessment: Validation



Result Assessment: K-fold Cross Validation



Result Assessment: Leave One Out Cross Validation



Zou, H, Hastie, T.
Regularization and variable selection via the elastic net.
Journal of the Royal Statistical Society, 2005.

De Mol, C. Devito, E., Rosasco, L.
Elastic-net regularization in learning theory
Journal of Complexity, 2009

l1l2 variable selection method

Empirical Risk minimization combined with a mixed penalty:

- l1 norm (sum of absolute values of β) enforcing **sparsity**
- l2 norm (sum of squared values of β) preserving **correlation**

error term — **l1 norm** — **l2 norm**

$$\phi_{\tau,\mu} = ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 + \tau||\boldsymbol{\beta}||_1 + \mu||\boldsymbol{\beta}||_2^2$$

Consistency guaranteed (the more samples available the better the estimator)

Not univariate: takes into account behavior of many genes at once.

Zou, H, Hastie, T.
Regularization and variable selection via the elastic net.
Journal of the Royal Statistical Society, 2005.

De Mol, C. Devito, E., Rosasco, L.
Elastic-net regularization in learning theory
Journal of Complexity, 2009

l1l2 variable selection method

Empirical Risk minimization combined with a mixed penalty:

- l1 norm (sum of absolute values of β) enforcing **sparsity**
- l2 norm (sum of squared values of β) preserving **correlation**

regularization parameter

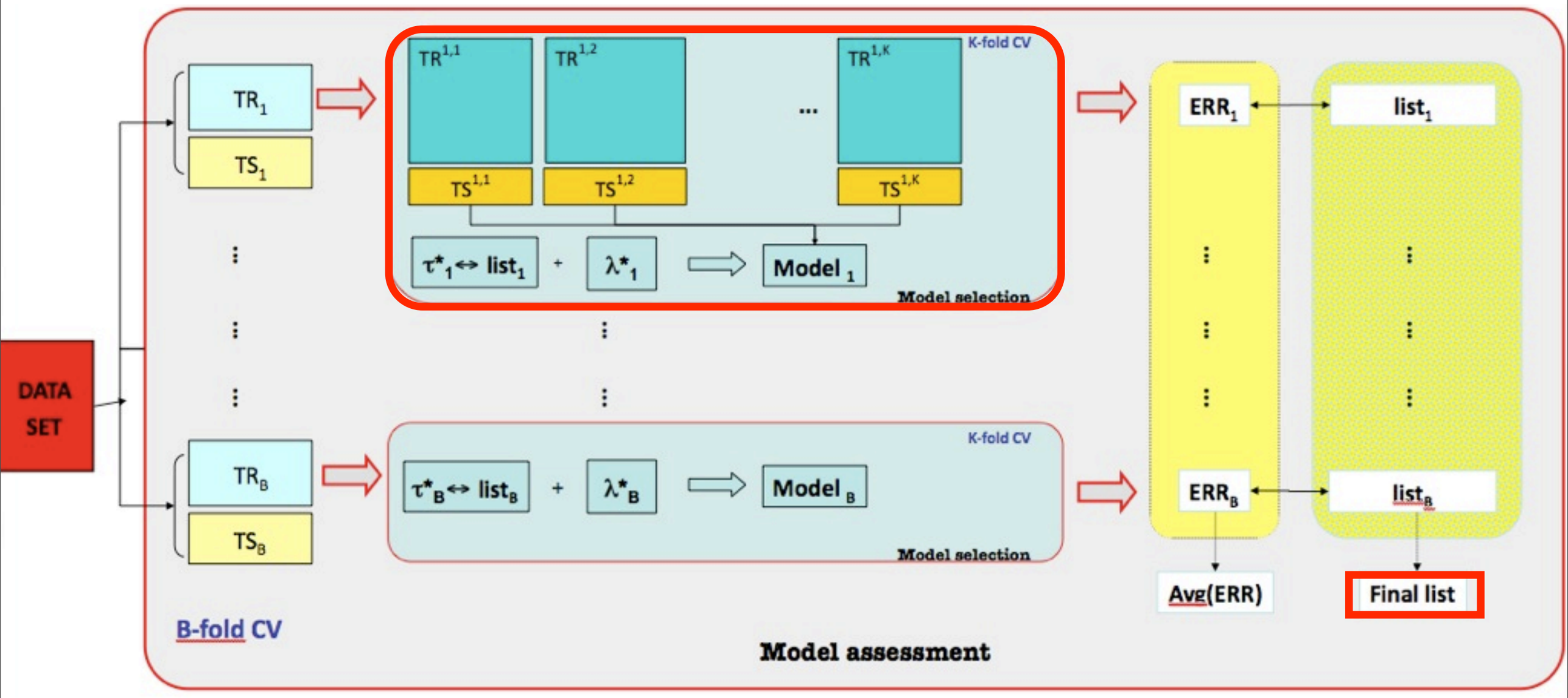
correlation parameter

$$\phi_{\tau,\mu} = ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 + \tau||\boldsymbol{\beta}||_1 + \mu||\boldsymbol{\beta}||_2^2$$

Consistency guaranteed (the more samples available the better the estimator)

Not univariate: takes into account behavior of many genes at once.

The Selection Bias Problem



A Barla, S Mosci, L Rosasco, A Verri.
A method for robust variable selection with significance assessment.
Proc. of ESANN, 2008.

the optimal pair (λ^*, τ^*) is one of the $A \cdot B$ possible pairs $(\lambda, \tau)_{ij}$

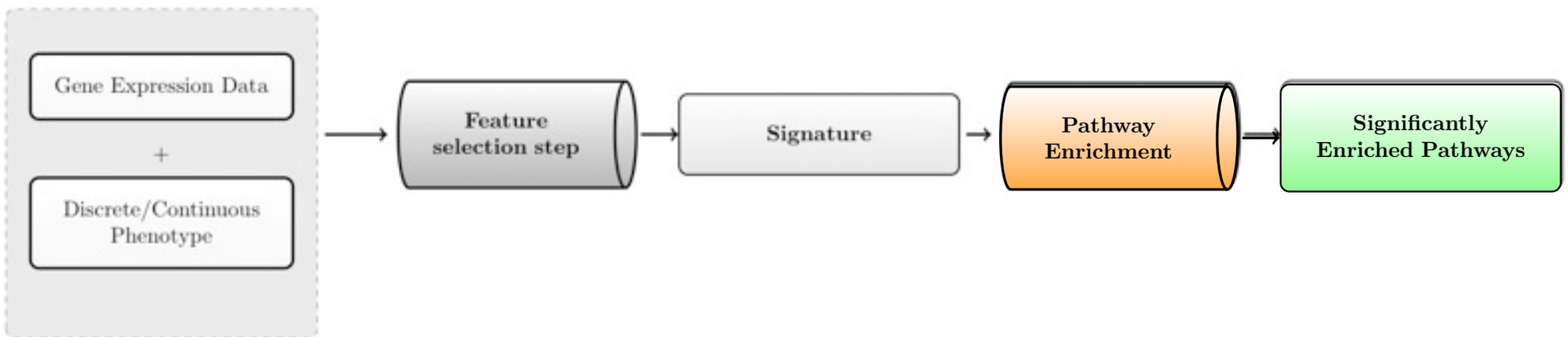
$\lambda \rightarrow (\lambda_1, ..., \lambda_A)$
$\tau \rightarrow (\tau_1, ..., \tau_B)$

computational time in the LOO case (for one task):

$time_{1\text{-optim}} = (2.5s \div 25s)$ depending on the correlation parameter

Total Time $= A \cdot B \cdot N.samples \cdot time_{1\text{-optim.}} \sim 20 \cdot 20 \cdot 30 \cdot time_{1\text{-optim}} \sim 2 \cdot 10^4 s \div 2 \cdot 10^5$

Pathway Enrichment Step



Pathway Enrichment
(functional characterization of the signature)

Published online 25 November 2008

Nucleic Acids Research, 2009, Vol. 37, No. 1 1–13
doi:10.1093/nar/gkn923

SURVEY AND SUMMARY

Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists

Da Wei Huang, Brad T. Sherman and Richard A. Lempicki*

BMC Bioinformatics

BioMed Central

Research article

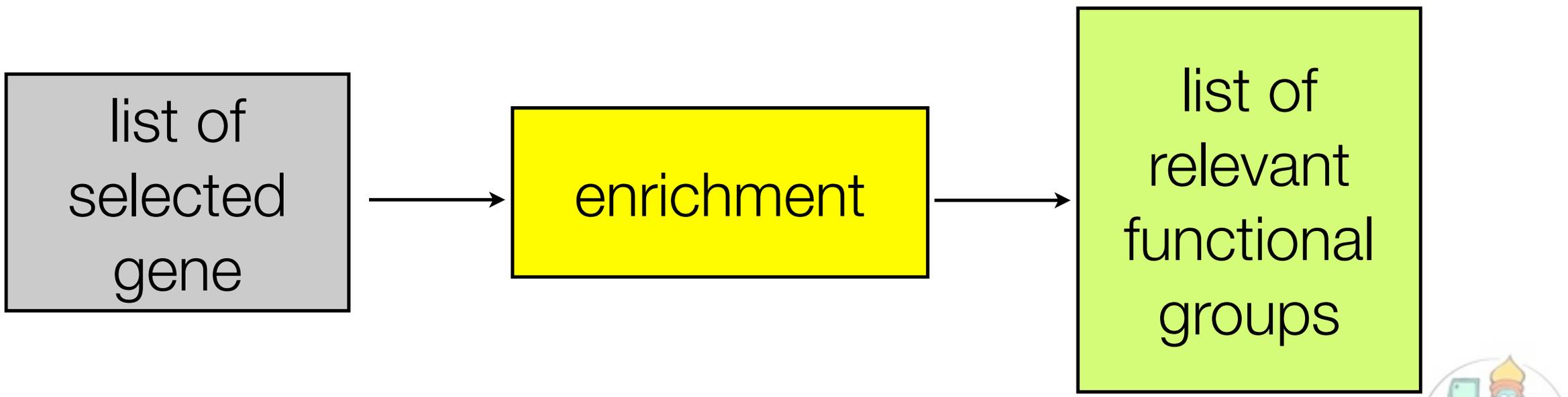
Open Access

Comparative study of gene set enrichment methods

Luca Abatangelo[1], Rosalia Maglietta[1], Angela Distaso[1], Annarita D'Addabbo[1], Teresa Maria Creanza[1], Sayan Mukherjee[2] and Nicola Ancona*[1]

Pathway Enrichment
(functional characterization of the signature)

- The **biological interpretation** of selected genes (ranging in size from hundreds to thousands of genes) is still a **challenging task**
- Lots of biological knowledge was accumulated in **public databases** in the last decade (Gene Ontology, KEGG, UniProt, ...)
- Bioinformatics **enrichment tools** have played a very important and successful role contributing to the **gene functional analysis** of large gene lists

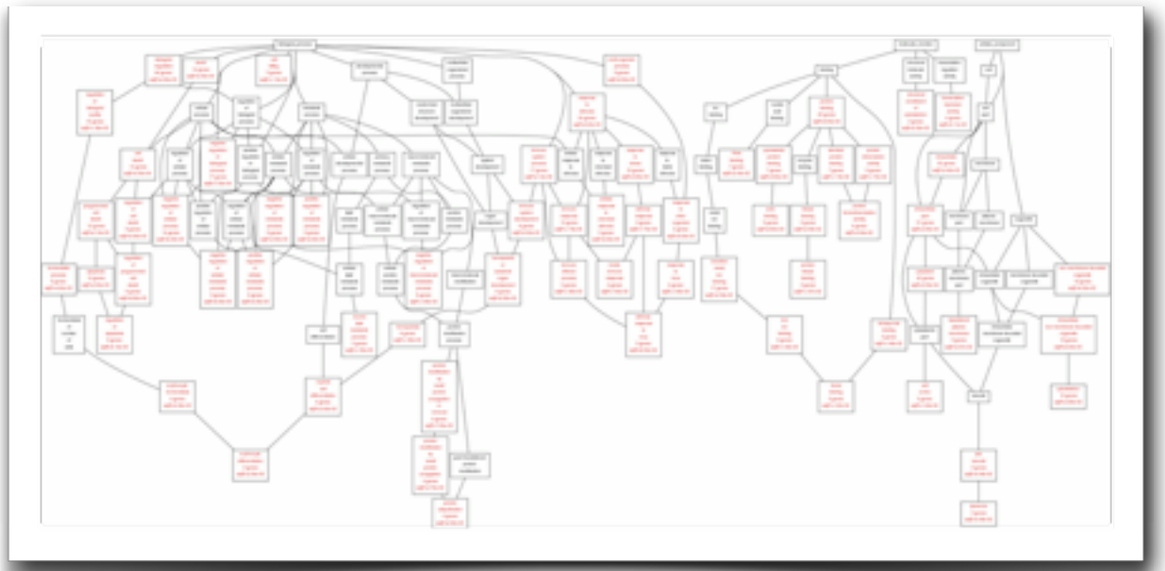


WebGestalt

- WebGestalt is a "WEB-based GEne SeT AnaLysis Toolkit".
  The tool is available at: http://bioinfo.vanderbilt.edu/webgestalt/

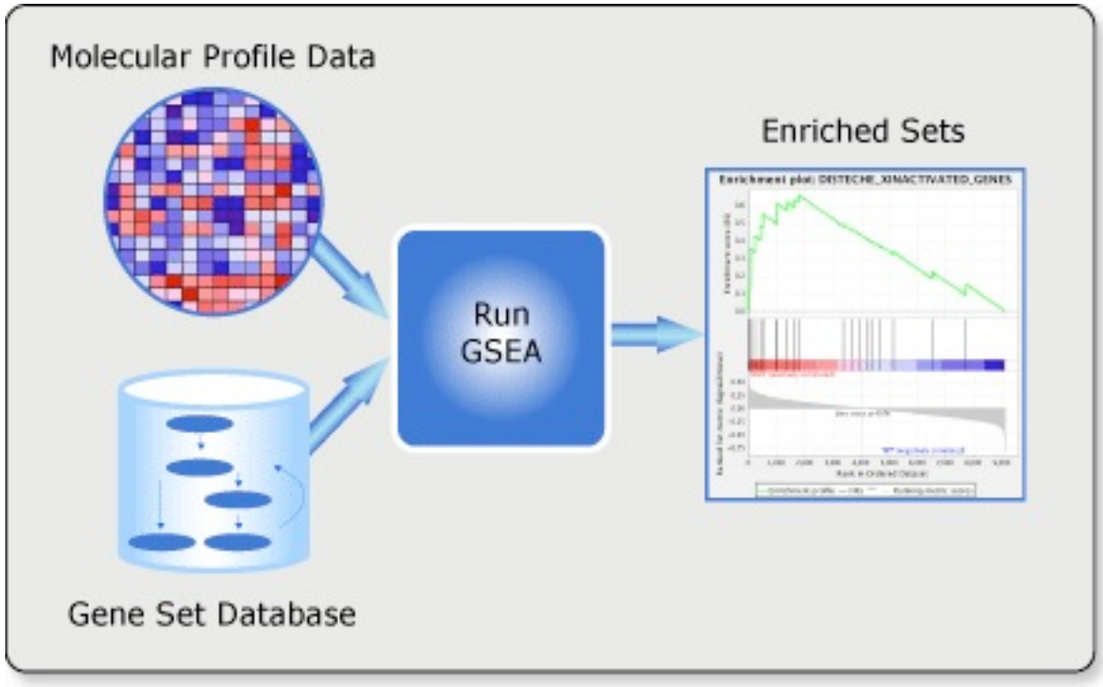
- The analysis consists in performing a Gene Set **Enrichment Analysis** on **Gene Ontology** and/or **KEGG**, provided the gene signature obtained in the Feature Selection step.

- The result is a **set of relevant GO nodes/KEGG pathways**



1. Zhang, B., Kirov, S.A., Snoddy, J.R.
WebGestalt: an integrated system for exploring gene sets in various biological contexts.
Nucleic Acids Res, 33(Web Server issue), W741-748. 2005
2. Duncan, D.T., Prodduturi, N., Zhang, B.
WebGestalt2: an updated and expanded version of the Web-based Gene Set Analysis Toolkit.
BMC Bioinformatics, 11(Suppl 4):P10. 2010

GSEA



- GSEA is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states

http://www.broadinstitute.org/gsea/

Alzheimer's as a case study

M Squillario and A Barla,
BMC Med Gen 2011

Alzheimer's disease (AD) as a case study

	controls	**cases**	**technology**	**notes**
Proteo	90	85	ELISA	2 separate test sets
GSE1297	9	22	Affymetrix HG-U133 A	various stages
GSE5281	62	68	Affymetrix HG-U133 Plus 2.0	late stage

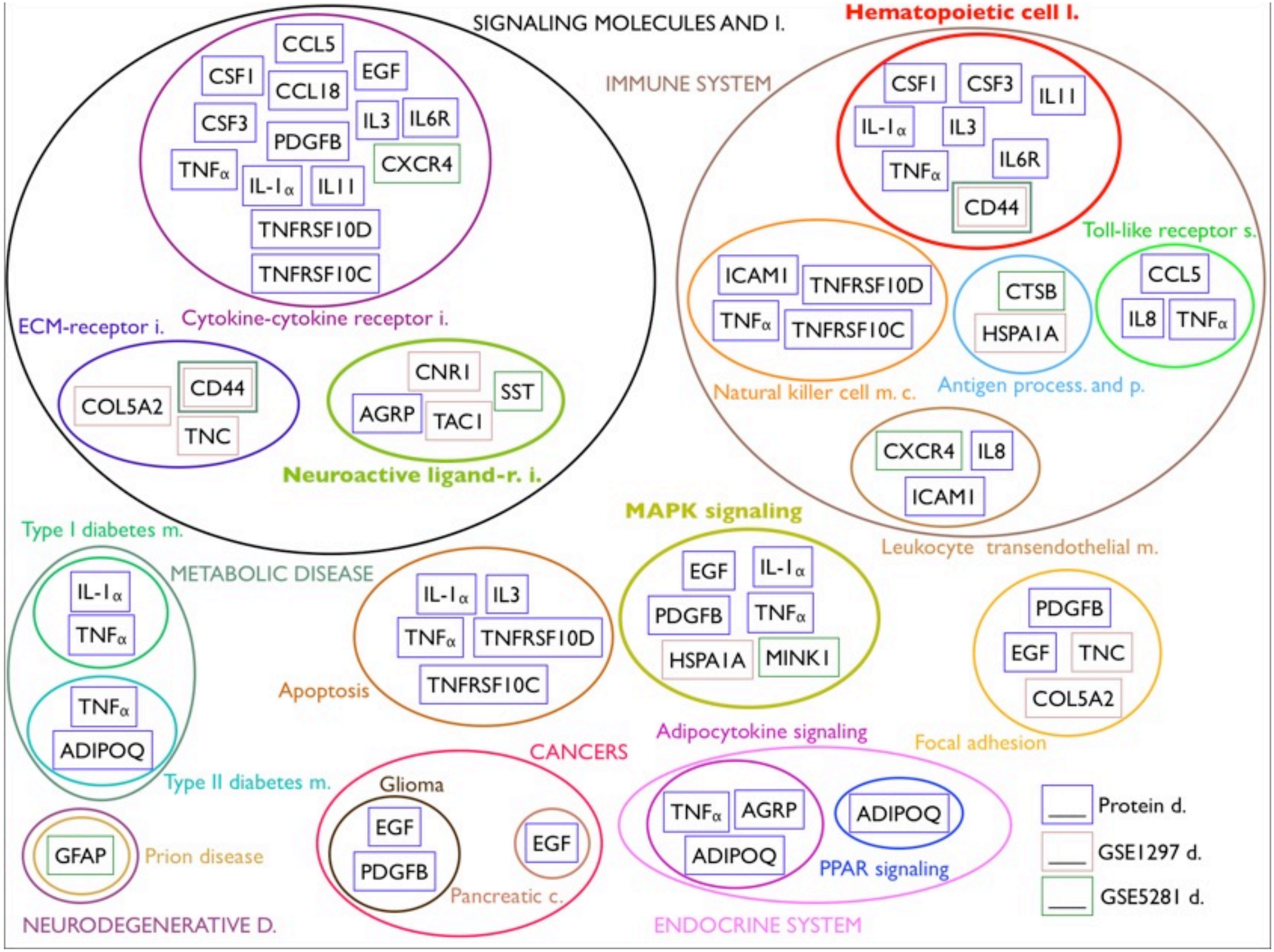
Results: accuracy, selected genes and pathways

	#genes	CV accuracy (%)			#KEGG pathways
			test sets		
Proteo	21	81	92	79	23
GSE1297	11	83			6
GSE5281	39	95			13

Functional Analysis: common characteristics

Despite the small (4) number of common genes across datasets, we have a consensus at the functional level

KEGG pathway	KEGG Category	Protein	GSE1297	GSE5281
Cytokine-cytokine receptor interaction		●		o
Neuroactive ligand-receptor migration	Signaling Molecules and Interaction	o	o	o
ECM-receptor interaction			●	o
Antigen processing and presentation			o	o
Hematopoietic cell lineage	Immune System	●	o	o
Leukocyte transendothelial migration		o		o
MAPK signaling pathway	Signal Transduction	●	o	o
Focal adhesion	Cell Communication	o	o	o



Functional Analysis: common characteristics

Some comments on Microarray and what's on next..

Microarrays: a success story

- Better understanding of response to drug
- Discover different phenotypes of a disease
- Classify the patients based on more or less aggressive phenotypes

Nature Reviews Neuroscience (Oct 2004)

"DNA-microarray-based technologies have already begun to uncover previously unrecognized patient subsets that differ in their survival."

DNA-MICROARRAY ANALYSIS OF BRAIN CANCER: MOLECULAR CLASSIFICATION FOR THERAPY

Paul S. Mischel, Timothy F. Cloughesy‡ and Stanley F. Nelson§*

Abstract | Primary brain tumours are among the most lethal of all cancers, largely as a result of their lack of responsiveness to current therapy. Numerous new therapies hold great promise for the treatment of patients with brain cancer, but the main challenge is to determine which treatment is most likely to benefit an individual patient. DNA-microarray-based technologies, which allow simultaneous analysis of expression of thousands of genes, have already begun to uncover previously unrecognized patient subsets that differ in their survival. Here, we review the progress made so far in using DNA microarrays to optimize brain cancer therapy.

Microarray workflow

Experimental design

- Frame a biological question
- Choose a microarray platform
- Decide on biological and technical replicates
- Design the series of hybridization

Technical performance

- Obtain the samples
- Isolate total RNA
- Label cDNA or mRNA
- Perform the hybridizations
- Scan the slides or chips

Statistical analysis

- Extract fluorescence intensities
- Normalize data to remove biases
- t-tests for pairwise comparisons
- ANOVA for multifactorial designs

Data mining

- Cluster analysis and pattern recognition
- Study lists of genome ontologies
- Search for regulatory motifs
- Design validation and follow-up experiments

This step determines:
the structure of microarray data,
the possible types of analyses,
the quality of the results

} low-level analysis (data cleaning)

} high-level analysis

Microarrays: a success story?? Issues...

BIOINFORMATICS **ORIGINAL PAPER**

Vol. 22 no. 7 2006, pages 789–794
doi:10.1093/bioinformatics/btk046

Gene expression

Comparison of Affymetrix GeneChip expression measures

Rafael A. Irizarry[1,*], Zhijin Wu[2] and Harris A. Jaffee[1]

timore, MD 21205, USA and
sity, 167 Angell Street,

nature genetics

Repeatability of published microarray gene expression analyses

John P A Ioannidis[1–3], David B Allison[4],
Mario Falchi[8,9], Cesare Furlanello[10], Lau
Michael Nitzberg[5], Grier P Page[4,12], Enri

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome

David Venet[1], Jacques E. Dumont[2], Vincent Detours[2,3]*

1 IRIDIA-CoDE, Université Libre de Bruxelles (U.L.B.), Brussels, Belgium, **2** IRIBHM, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium, **3** WELBIO, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium

Microarrays: a success story?? Issues...

Reproducibility of results depend on:

- sample collection (n of sample, characteristics of the biological samples)
- production of the data due to the person that actually does the experiment
- data preprocessing (normalization)
- method used to get the results (univariate/multivariate)
- methodological protocol used to analyze the data (selection bias)

Lesson learned

Google groups

« Groups Home

Scientists for Reproducible Research

Home

Discussions 7 of 199 messages view all »

Duke Saga - Patient Lawsuits, the Economist, Retraction
By Keith Baggerly - Sep 14 - 1 author - 0 replies

IOM Meeting -- Duke's Institutional Response
By Keith Baggerly - Aug 24 - 1 author - 0 replies

www.reproducibleresearch.net
By Thompson,Paul - Jul 27 - 3 authors - 4 replies

Files from IOM Meeting Jun 30
By Mauro Delorenzi - Jul 11 - 3 authors - 2 replies

Duke Saga on front page of NY Times; NCI Workshop; IOM Meeting #3
By Keith Baggerly - Jul 7 - 1 author - 0 replies

Notes from the Council of Science Editors (CSE 2011)
By Victoria Stodden - May 8 - 3 authors - 3 replies

ENAR session update -- sound files!
By Keith Baggerly - May 7 - 1 author - 0 replies

Report this group XML Send email to this group: reproducible-research@googlegroups.com

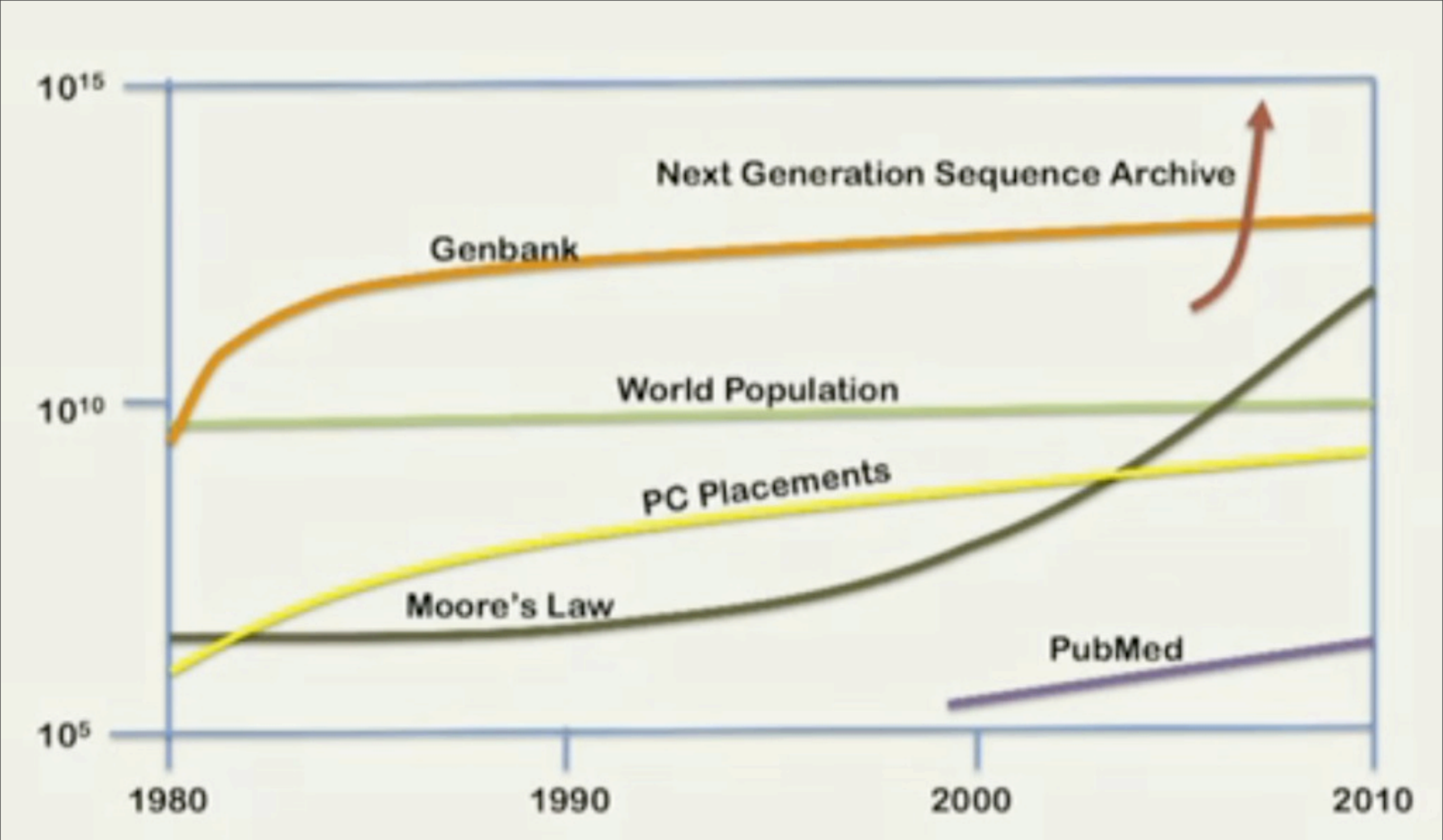
Foreseeing the future: NextGen Sequencing

- NGS experiment **allows for (possibly) whole DNA/RNA sequencing** and is not limited as in the microarray

- Efficacy of the **NGS experiment does not depend on the hybridization phase** as in the microarray experiment

- **More experiments can be performed at once** (i.e. combine DNA, SNP, Chip on Chip microarrays)

- **Cost of NGS machines is decreasing** therefore in the near future they will become much affordable



Sequencing rate

Richard Resnick: Welcome to the genomic revolution - TEDxBoston 2011