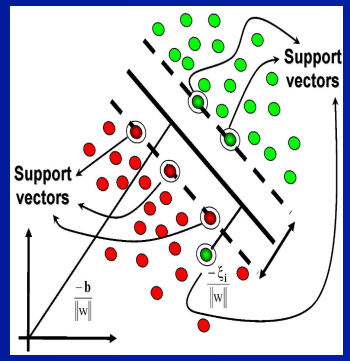
Data Mining and Machine Learning in Bioinformatics

PRINCIPAL METHODS AND SUCCESSFUL APPLICATIONS

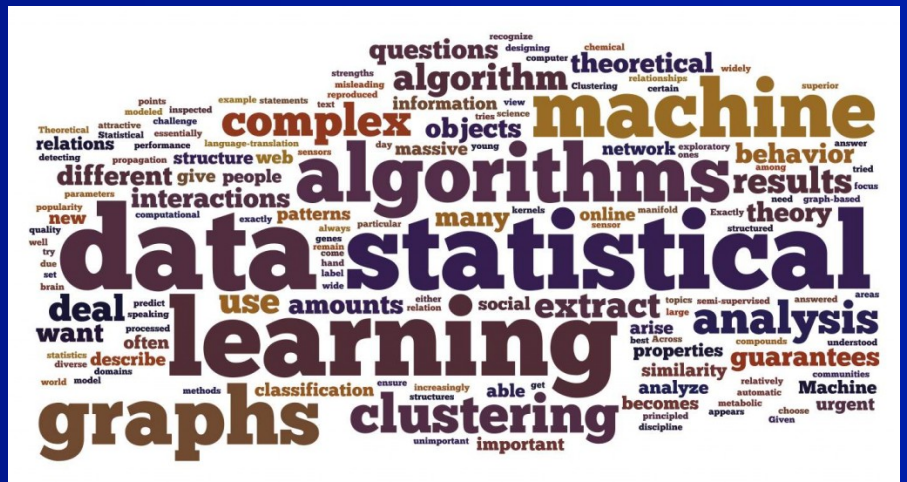
Ruben Armañanzas
http://mason.gmu.edu/~rarmanan

Adapted from Iñaki Inza slides
http://www.sc.ehu.es/isg

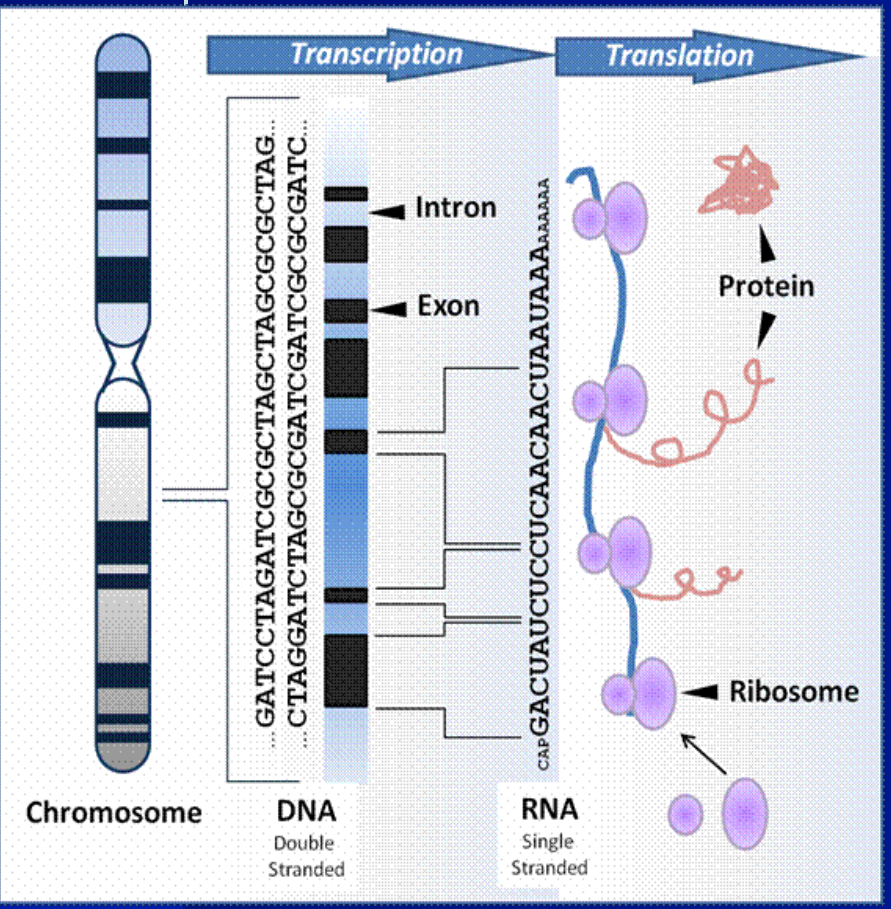


OUTLINE

- Intro to bioinf and data mining
- Basic vocabulary and main techniques in machine learning
  - Bioinformatic applications

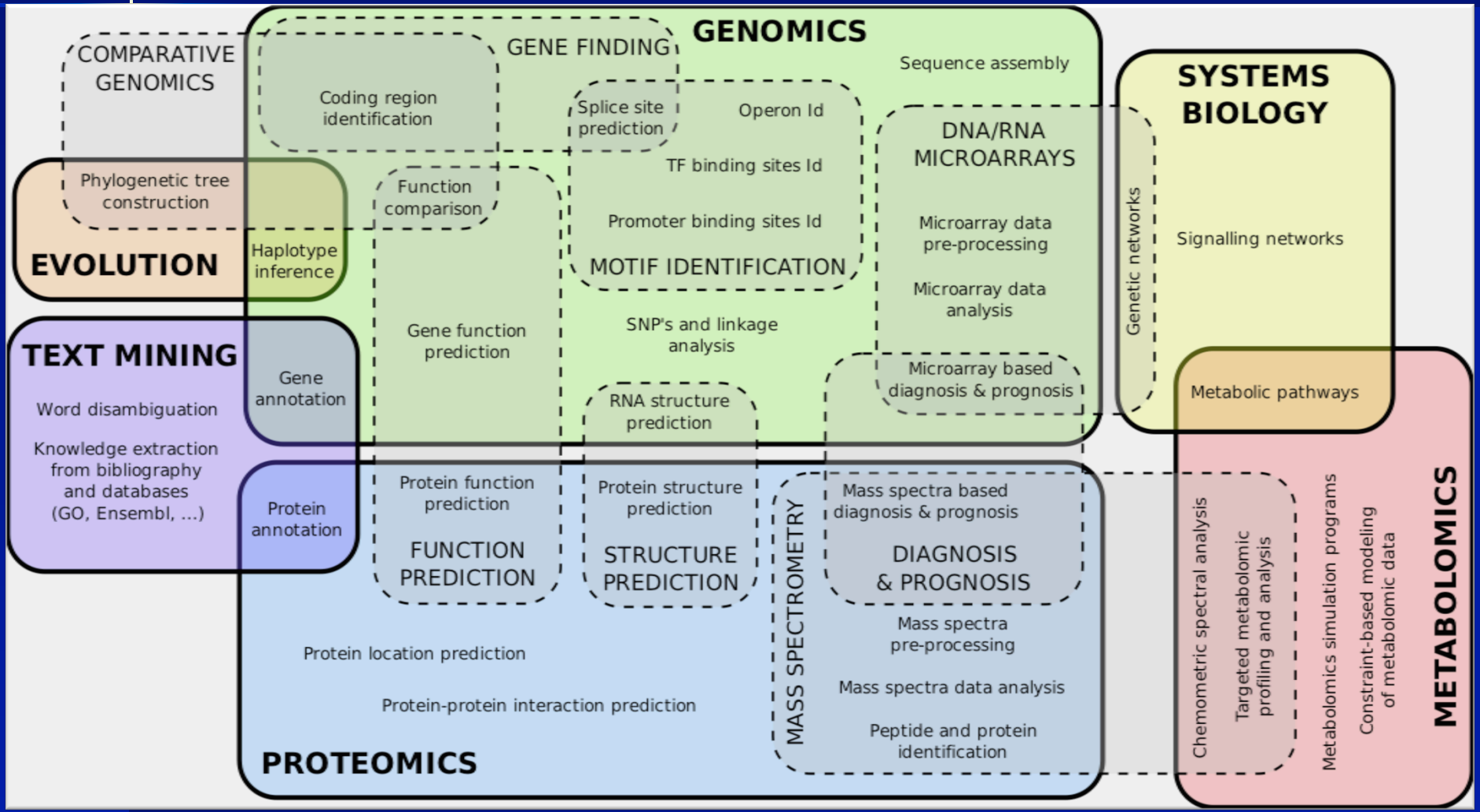


BIOINFORMATICS SUCCINCT INTRO



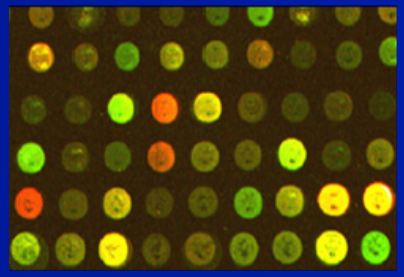
- **Bioinformatics** applies methods of information science for the analysis, modeling, and knowledge discovery of biological processes in living organisms
- It brings together several disciplines – molecular biology, mathematics, chemistry, physics, and informatics, with the aim of understanding life

GENERAL SCHEME OF ML APPS IN BIOINFORMATICS



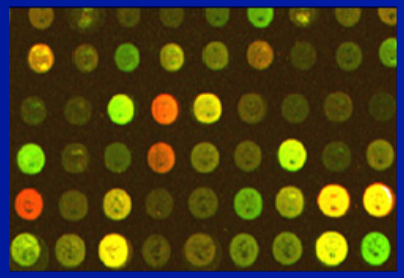
DATA MINING ROOTS

- Data collected and stored at enormous speeds (GB/hour). **Data collections-floods,** which were not envisioned to be analyzed few years ago, are being collected and warehoused:
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
  - electronic purchases and transactions
  - …



DATA MINING ROOTS

- Computers and storage systems have become **cheaper** and more **powerful**
- Since 90's, much more data is being stored **than** analyzed (around 5-10%)
- "Data tsunami": in 2010 enterprises stored 7 exabytes (10^{18}bytes)= 7,000,000,000 GB
- **Traditiona**l data analysis techniques **unfeasible** for raw data



DEFINITION: DATA MINING

Definition (Fayyad et. al): The non-trivial discovery of *novel, valid , comprehensible* and potentially *useful* patterns from data.

What is a pattern? A relationship in the data. E.g.,

Data Mining is Not …

- Data warehousing
- Ad Hoc Query/ Reporting
- Online Analytical Processing (OLAP)
- Data Visualization
- Software Agent

Data mining, also popularly referred to as *knowledge discovery in databases (KDD)*, is the automated or convenient extraction of patterns representing knowledge implicitly stored in large databases, data warehouses, and other massive information repositories.

– *Data Mining: Concepts and Techniques* by Han and Kamber

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data.

– *Data Mining* by Witten and Frank

- Technologies for analysis of data and discovery of (very) hidden patterns
- Uses a combination of statistics, probability analysis and database technologies
- Fairly young (<20 years old) but clever algorithms developed through database research

BIG DATA

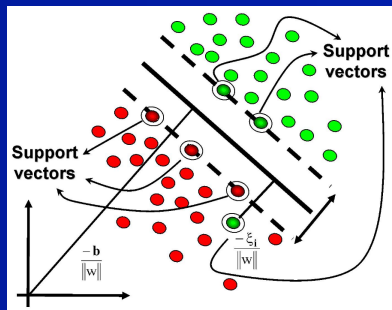
- New technological concept
- Related to the challenges exposed to manipulate massive datasets (petabytes, exabytes):
  - Capture and storage
  - Processing and computing
  - Analysis and mining
- Demands the development of new platforms: MapReduce, Hadoop, ...

DEFINITION: MACHINE LEARNING

- **Machine Learning** refers to the application of induction algorithms, which is one step in the knowledge discovery process
- Training examples are either *externally supplied*, or supplied by a previous stage of the data mining process.
- **Machine Learning** is the field of scientific study that concentrates on induction algorithms and on other algorithms that can be said to **learn**
- Kohavi & Provost: Glossary of ML Terms:
  - http://ai.stanford.edu/~ronnyk/glossary.html

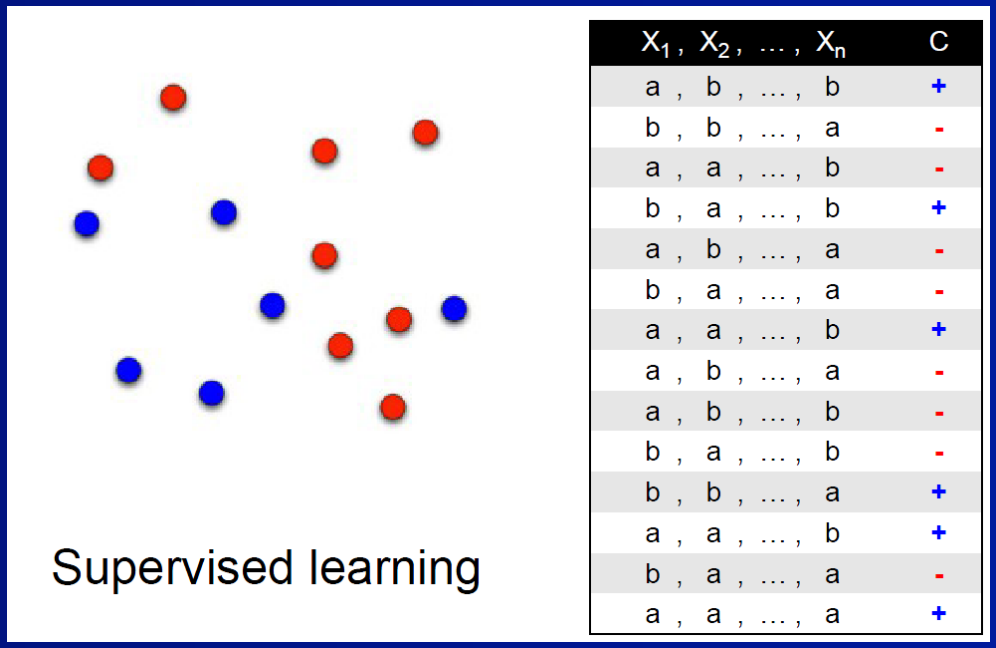
DM + ML: MAIN TASKS

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables
    - *Supervised classification: nominal variable to be predicted*
    - *Regression: ordinal variable to be predicted*
- Description Methods
  - Find human-interpretable patterns that describe the data
    - *Clustering – unsupervised classification*
    - *Association rule discovery*
    - *Feature selection: discover the key predictive features*
    - *Outlier detection*



SUPERVISED CLASSIFICATION

- Given a collection of records-samples (*training set*)
  - Each record contains a set of *attributes-features-predictors*
  - Each record belongs to a *class, our variable of interest (variable to be predicted)*

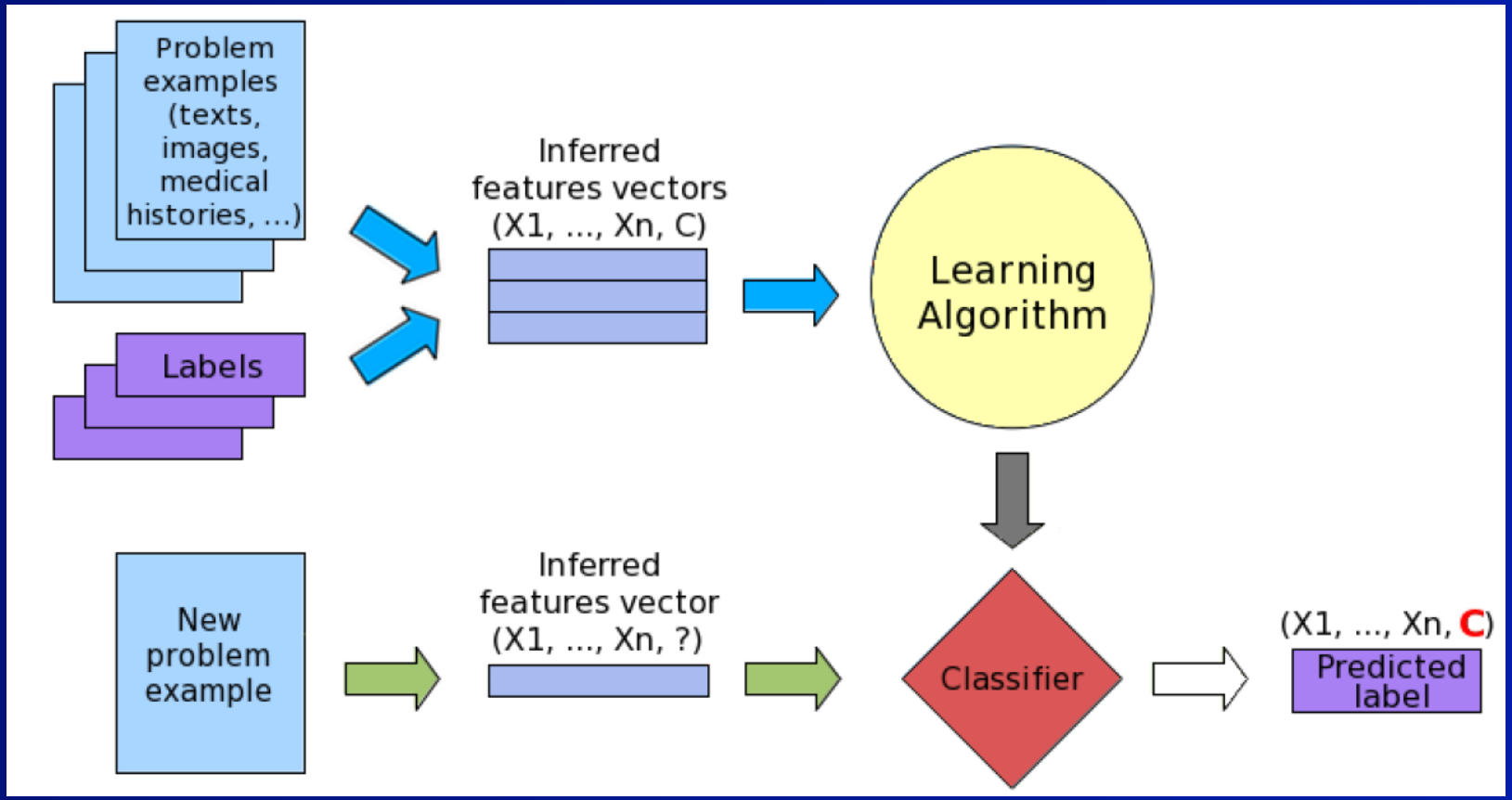


X_1 , X_2 , ... , X_n	C
a , b , ... , b	+
b , b , ... , a	-
a , a , ... , b	-
b , a , ... , b	+
a , b , ... , a	-
b , a , ... , a	-
a , a , ... , b	+
a , b , ... , a	-
a , b , ... , b	-
b , a , ... , b	-
b , b , ... , a	+
a , a , ... , b	+
b , a , ... , a	-
a , a , ... , a	+

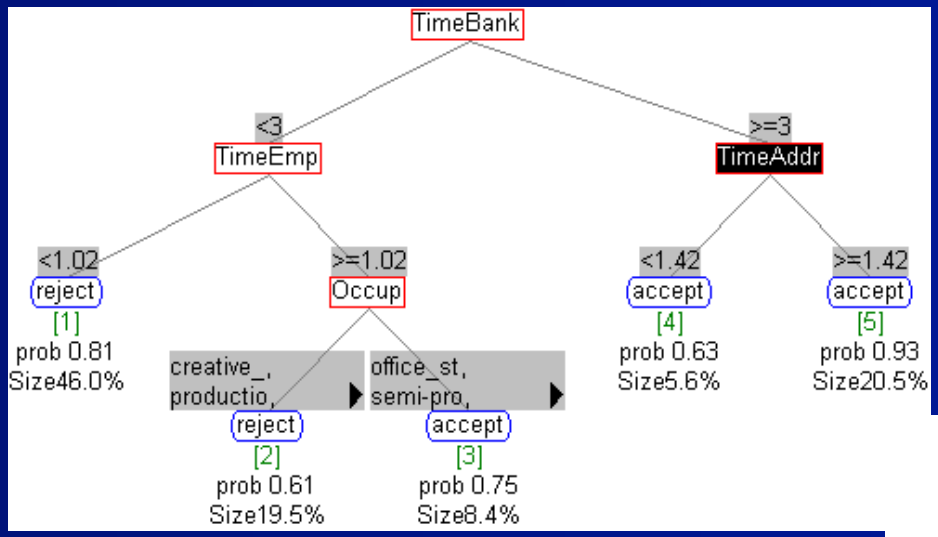
SUPERVISED CLASSIFICATION

- Find a *model* for class attribute as a function of the values of other attributes. There is a broad range of model types:
  - Decision trees, Bayesian networks, neural networks…
- Goal: previously unseen records should be assigned a class as accurately as possible
  - A *test set* is used to *estimate the accuracy* of the model. There is a broad range of techniques for accuracy estimation: cross-validation, hold-out, bootstrap, …

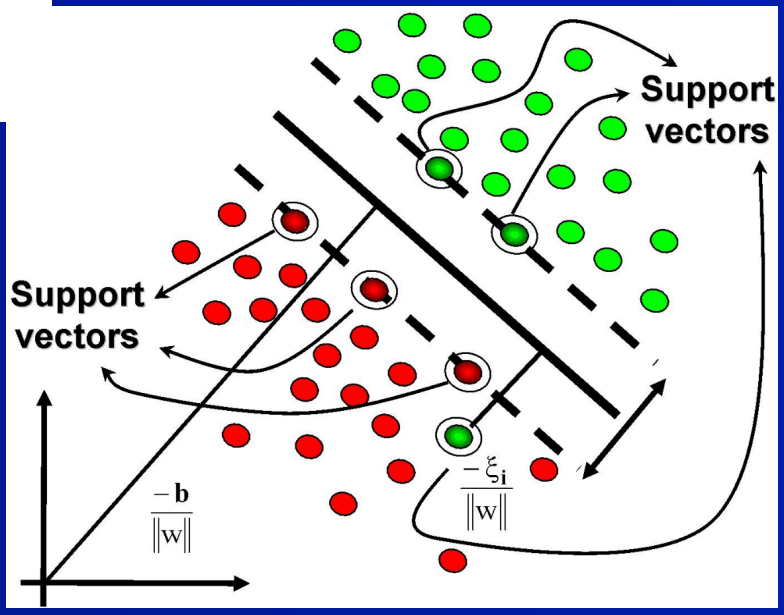
SUPERVISED CLASSIFICATION: the standard scenario



SUPERVISED CLASSIFICATION: models

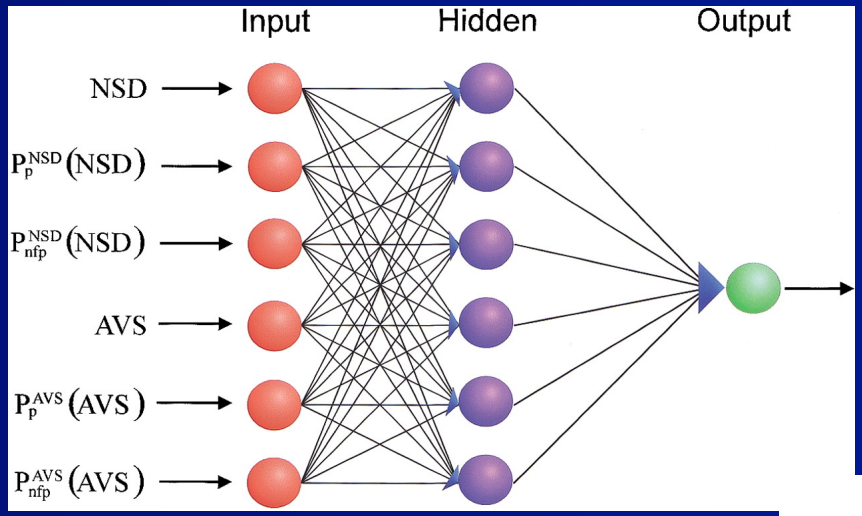


Decision trees

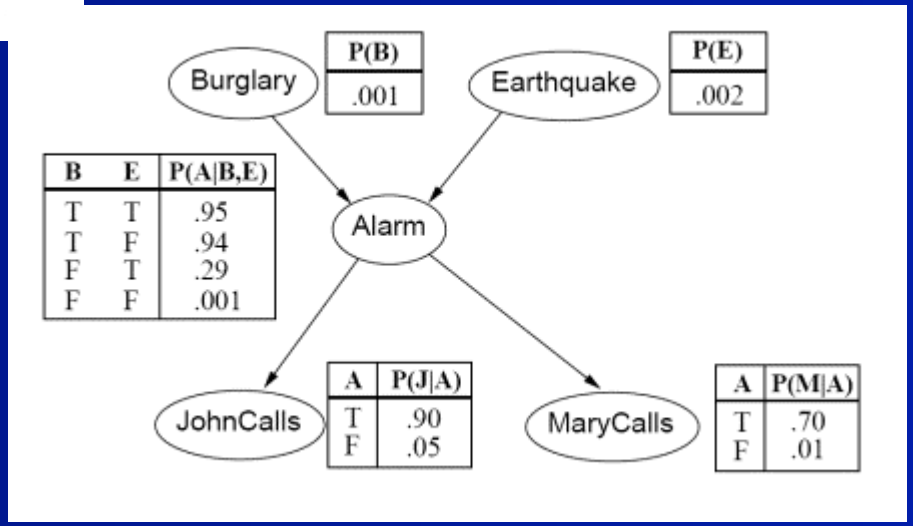


Support vector machines

SUPERVISED CLASSIFICATION: models

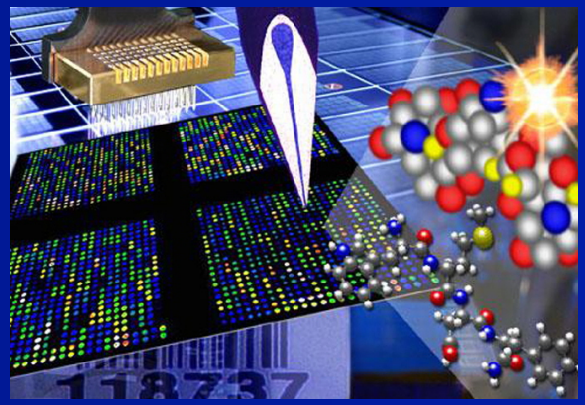


Neural networks



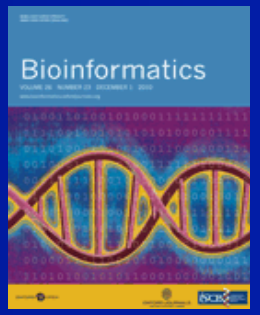
Bayesian networks

BIOMEDICAL INFORMATICS - BIOINFORMATICS
DIAGNOSIS AND PROGNOSIS OF DISEASES
BIOMARKER DISCOVERY



BIOMEDICAL INFORMATICS - BIOINFORMATICS
DIAGNOSIS AND PROGNOSIS OF DISEASES
BIOMARKER DISCOVERY

Differential Micro RNA Expression in PBMC from Multiple Sclerosis Patients



Oncology
International Journal for Cancer Research and Treatment

Random Forest for Gene Expression Based Cancer Classification: Overlooked Issues

Ensemble machine learning on gene expression data for cancer classification

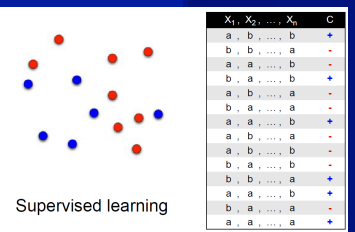
nature

Genome Biology

Classification of Alzheimer's Disease and Parkinson's Disease by Using Machine Learning and Neural Network Methods

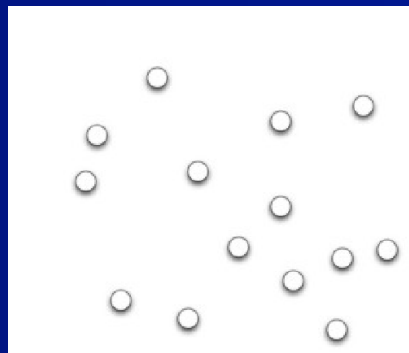
UNSUPERVISED CLASSIFICATION
CLUSTERING

- Given a collection of records-samples (*training set*)
  - Each record contains a set of *attributes-features-predictors*
  - No "target feature" (class) which supervises the learning process
- Find groups of cases with:
  - Large intra-group homogeneity
  - Large inter-groups heterogeneity
- Difficult evaluation-measure of these properties → no recognition rate
- Number of groups…



Supervised learning

X_1 , X_2 , … , X_n	C
a , b , … , b	+
b , b , … , a	-
a , a , … , b	-
b , a , … , b	+
a , b , … , a	-
b , a , … , a	-
a , a , … , b	+
a , b , … , a	-
a , b , … , b	-
b , a , … , b	-
b , b , … , a	+
a , a , … , b	+
b , a , … , a	-
a , a , … , a	+

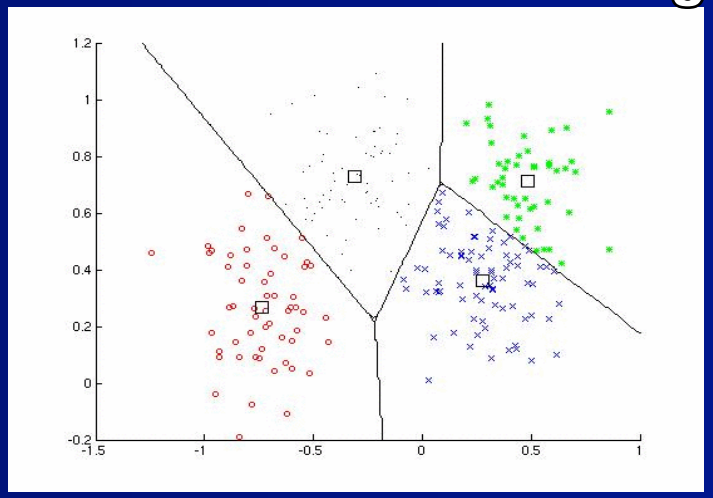


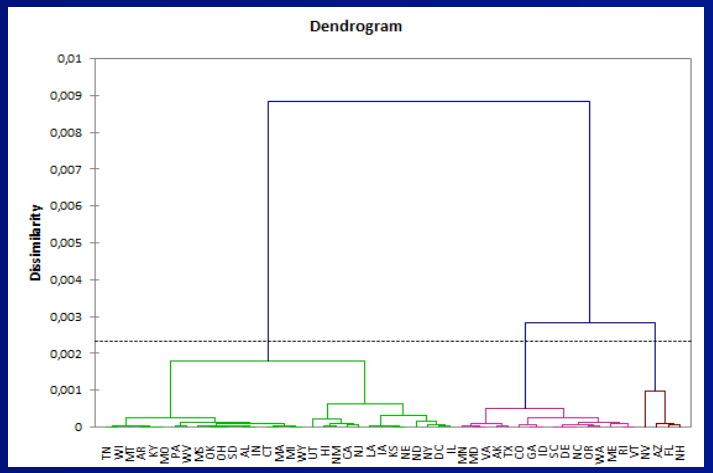
Unsupervised learning

X_1 , X_2 , … , X_n	C
a , b , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , b	?
a , b , … , a	?
b , a , … , a	?
a , a , … , b	?
a , b , … , a	?
a , b , … , b	?
b , a , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , a	?
a , a , … , a	?

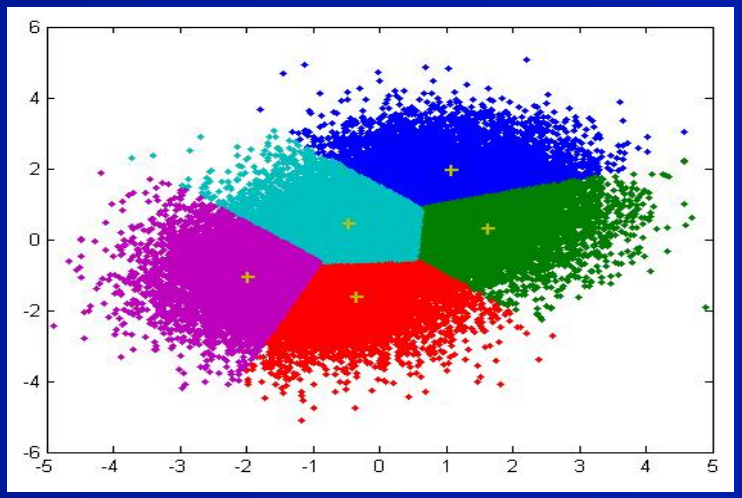
CLUSTERING: MODELS

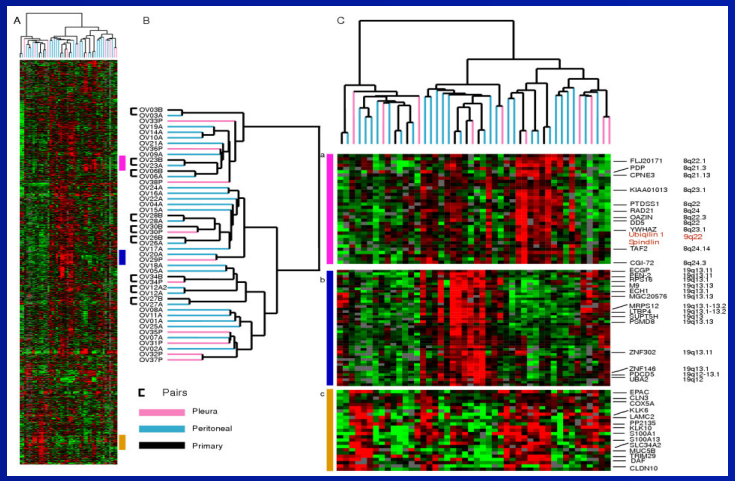
Hierarchical clustering





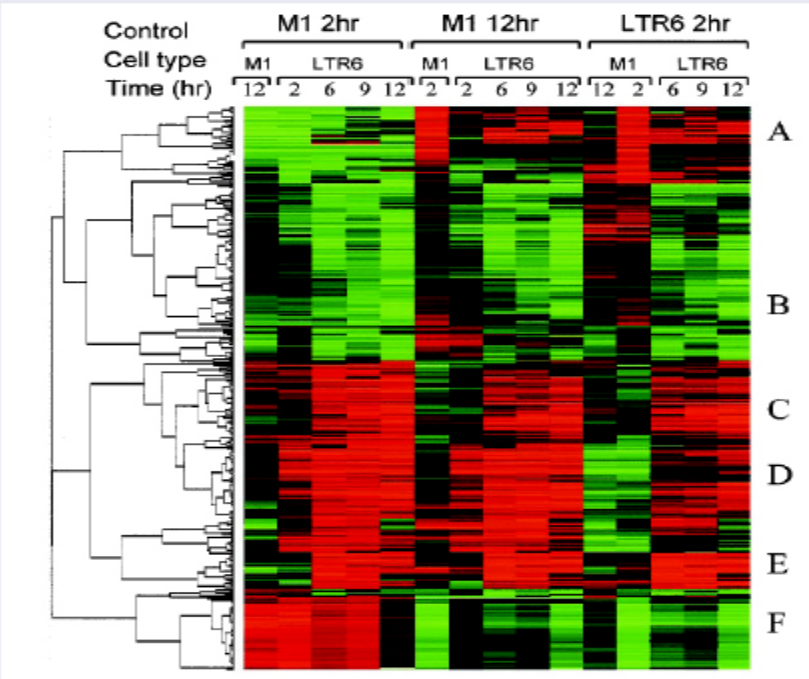
Partitional clustering (k-means)

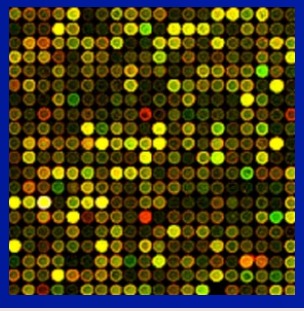


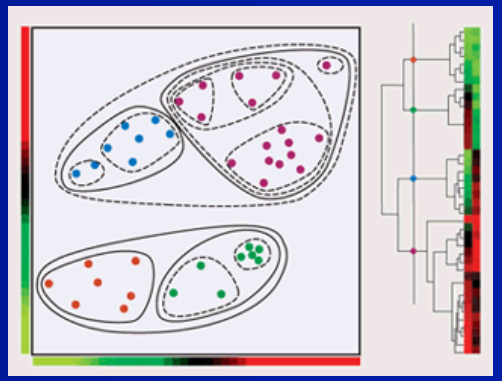


DNA MICROARRAY CLUSTERING

- Find genes with similar expression profiles → a way to infer the function of genes whose function is unknown
- Biclustering… a classic concept in fashion again: Hartigan JA (1972). "Direct clustering of a data matrix". *Journal of the American Statistical Association* **67** (337)

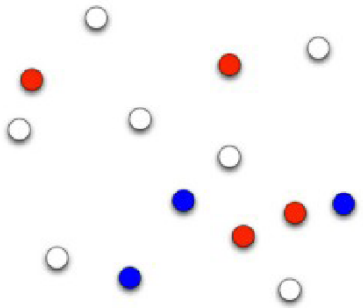






SEMI-SUPERVISED CLASSIFICATION

- Given a collection of records-samples (*training set*)
  - Each record contains a set of *attributes-features-predictors*
  - A small subset of the samples is categorized (known class value)
  - Most of the samples do not show a class value. Why?
    - Categorization: human-time consuming task
    - No knowledge to categorize the samples
  - Can a learning process which takes advantage of unlabeled samples, construct a better supervised classification model?

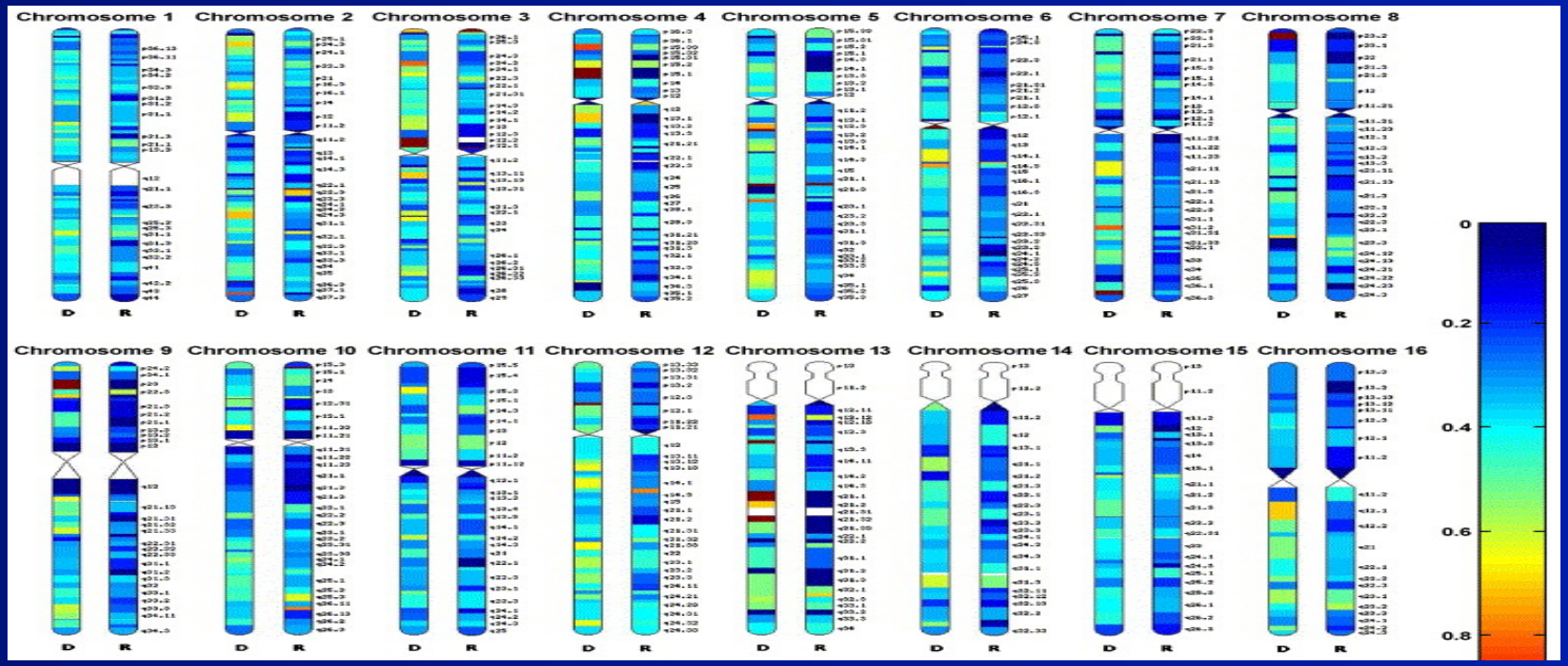


Semi-supervised learning

X_1 , X_2 , … , X_n	C
a , b , … , b	+
b , b , … , a	-
a , a , … , b	-
b , a , … , b	+
a , b , … , a	-
b , a , … , a	-
a , a , … , b	+
a , b , … , a	?
a , b , … , b	?
b , a , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , a	?
a , a , … , a	?

PREDICTION OF GENES RELATED TO CANCER

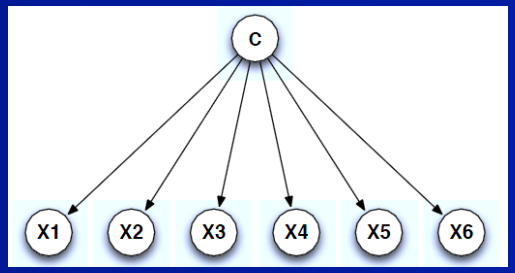
- It is already known that certain genes are related to cancer
- For the rest of the genes it can not be stated that they are not related to cancer
- Helpful to prioritize, for oncogenic experts, the depth-study of specific genes
- More difficult than semi-supervised classification: one-class (partially supervised)

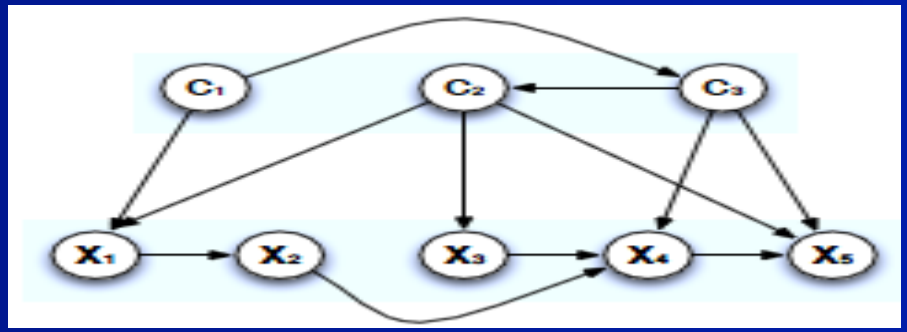


OTHER TYPES OF CLASSIFICATION PROBLEMS
MULTIDIMENSIONAL CLASSIFICATION

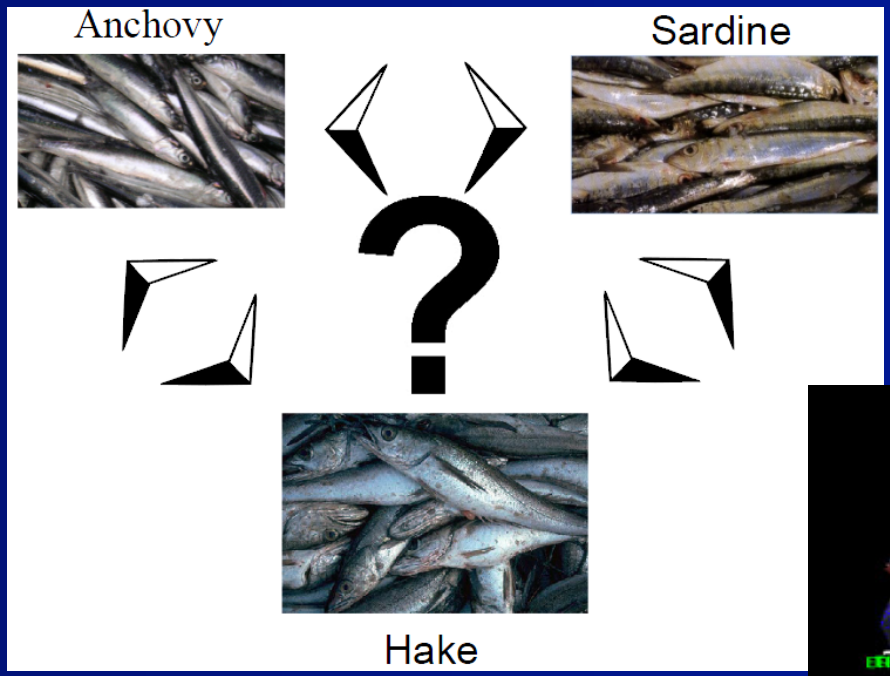
- Several class variables to be jointly predicted
- Learn relationships between class variables
- New term: Joint accuracy

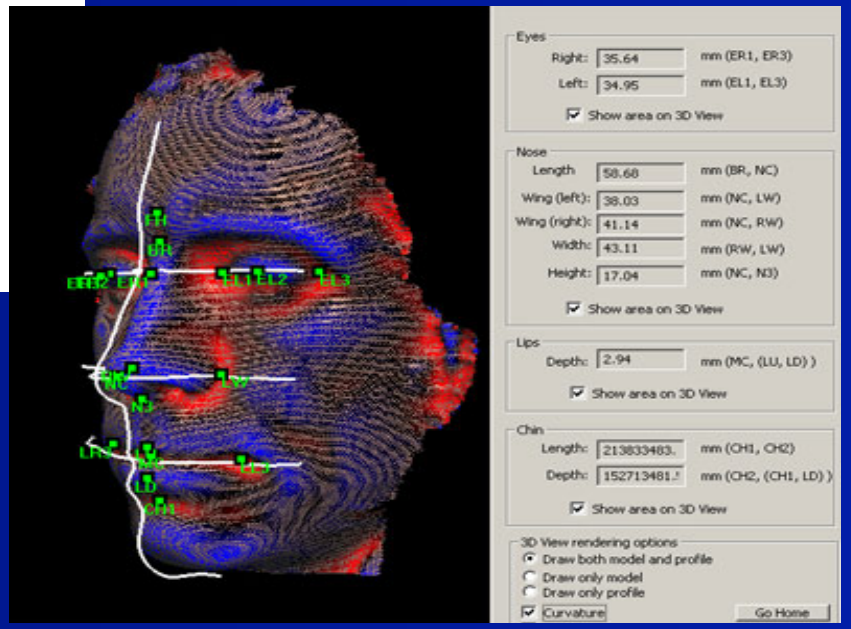
X_1	X_2	...	X_n	C_1	C_2	...	C_m
$x_1^{(1)}$	$x_2^{(1)}$	...	$x_n^{(1)}$	$c_1^{(1)}$	$c_2^{(1)}$	...	$c_m^{(1)}$
$x_1^{(2)}$	$x_2^{(2)}$	...	$x_n^{(2)}$	$c_1^{(2)}$	$c_2^{(2)}$	...	$c_m^{(2)}$
...	...	...	...	...	...	...	...
$x_1^{(N)}$	$x_2^{(N)}$	...	$x_n^{(N)}$	$c_1^{(N)}$	$c_2^{(N)}$	...	$c_m^{(N)}$





MULTIDIMENSIONAL CLASSIFICATION APPLICATIONS





MULTILABEL CLASSIFICATION

X1	X2	...	Xn	C
0	1	...	0	a,c
1	0	...	0	b
1	0	...	1	b,c
0	0	...	1	a,b
1	1	...	0	a,b,c
0	1	...	1	a,b
0	0	...	0	b,c

X1	X2	...	Xn	C
1	1	...	1	?

N.	Film	Year	Genre
1	Cadena perpetua	1994	Crime, Drama
2	El padrino	1972	Crime, Drama
3	El padrino. Parte II	1974	Crime, Drama
4	El bueno, el feo y el malo	1966	Adventure, Western
5	Pulp Fiction	1994	Crime, Thriller
6	12 hombres sin piedad	1957	Drama
7	La lista de Schindler	1993	Biography, Drama, History, War
8	El caballero oscuro	2008	Action, Crime, Drama, Thriller
9	El señor de los anillos: El ret...	2003	Action, Adventure, Drama, Fantasy
10	El club de la lucha	1999	Drama

MULTIPLE INSTANCE LEARNING

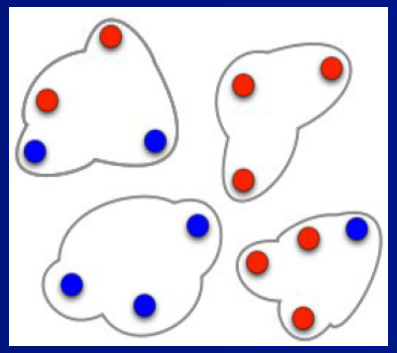
X_1 , X_2 , ... , X_n	C
a , b , ... , b	+
b , b , ... , a	-
a , a , ... , b	-
b , a , ... , b	+
a , b , ... , a	-
b , a , ... , a	-
a , b , ... , a	-
a , a , ... , b	+
a , b , ... , b	-
b , a , ... , b	-
b , a , ... , a	-
a , a , ... , b	+
b , b , ... , a	+
a , a , ... , a	+

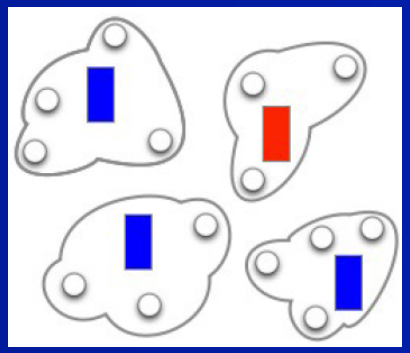
Bag label:

+ At least one instance in the bag is positive.

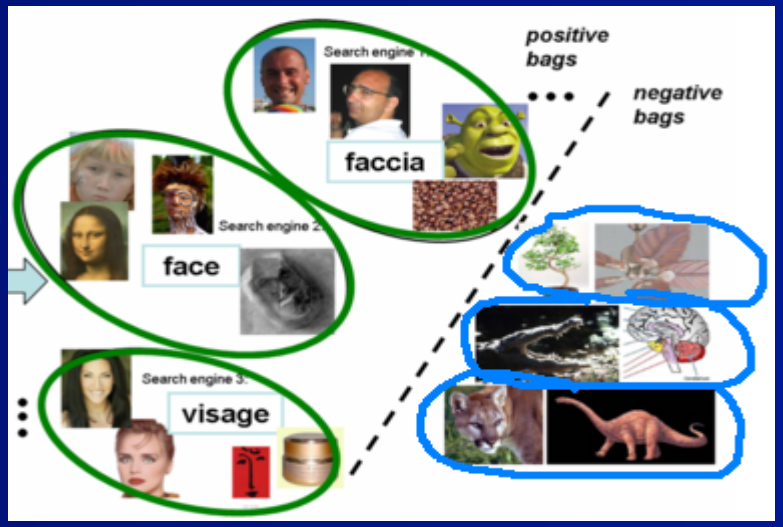
- Otherwise

X_1 , X_2 , ... , X_n	C
a , b , ... , b	
b , b , ... , a	+
a , a , ... , b	
b , a , ... , b	
a , b , ... , a	
b , a , ... , a	-
a , b , ... , a	
a , a , ... , b	
a , b , ... , b	+
b , a , ... , b	
b , a , ... , a	
a , a , ... , b	
b , b , ... , a	+
a , a , ... , a	

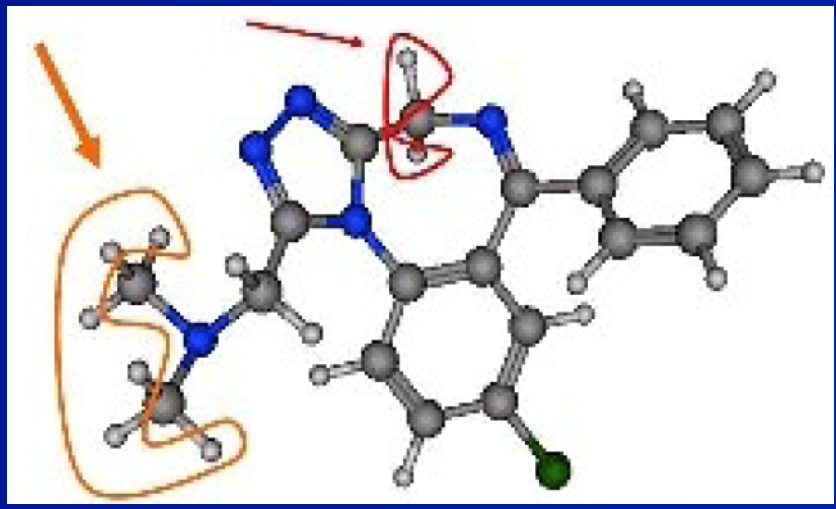




MULTIPLE INSTANCE LEARNING



Are all the images of the bag "faces"?



Are all foldings of a protein of the same type?

ASSOCIATION RULES

- Given a set of records each of which contain some number of items from a given collection;
  - Depedendency rules which will predict occurrence of an item based on occurrences of other items.
  - Rules are composed of "antecedent" and "consequence" parts: IF-THEN form
  - No "class" concept: any item can be in the "antecedent" or "consequence" part
  - **"Support"** and **"Confidence"** concepts

t1: {bread, cheese, fluidmilk}
t2: {apple, eggs, salt, yogurt}
t3: {bananas, eggs, saladvegetable}
.....
tn: {biscuit, eggs, fluidmilk}

1	ANTECEDENT	==>>	CONSEQUENCE	Support (%)	Confidence (%)
2	Pizza & Tomato	==>>	Grated cheese	5%	82%
3	Pizza & "Man"	==>>	Beer	3%	75%
4	SaladVegetable & Meat	==>>	Wine	10%	68%
5	Milk & Bread	==>>	Jam	18%	61%
6	Diaper & "Man"	==>>	Beer	4%	44%
7	Coke & Nachos	==>>	Paper serviette	2%	40%