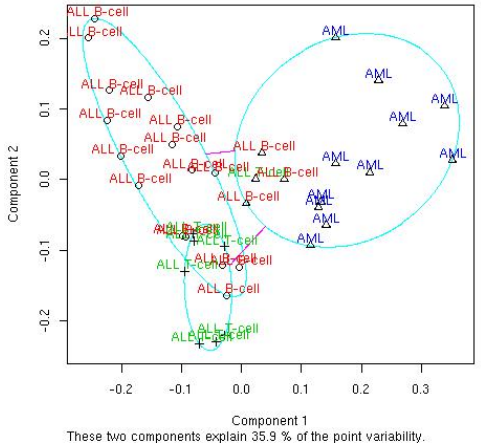
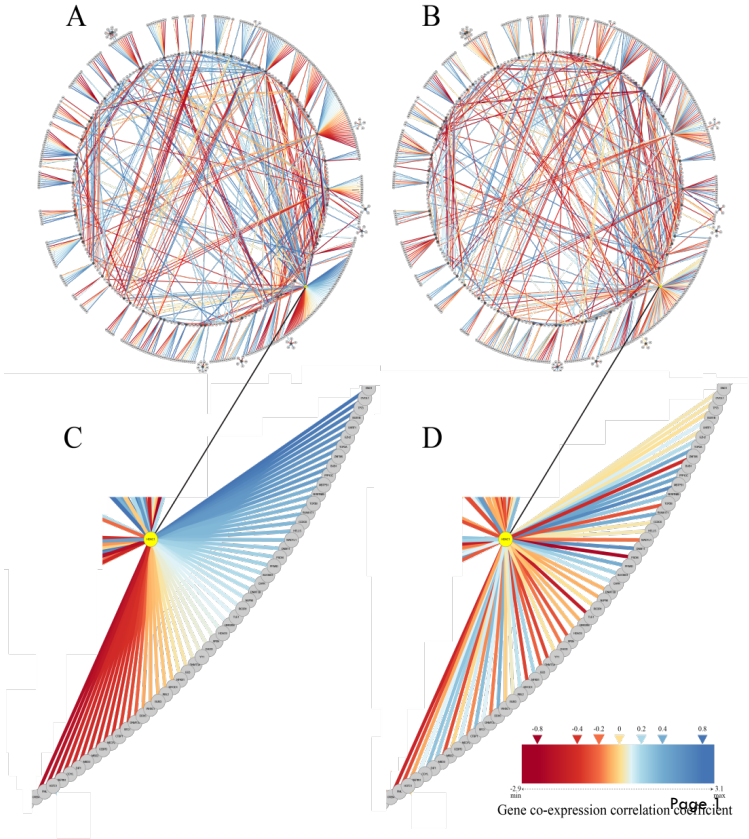
Introduction to Classification

Omics analysis – finding biomarkers

AMED3002





Risk Factors: Medical, Psychological, Experiential, and genome-wide genotyping

Genotypes
Affy 6.0/Illumina Quad
1000G imputation

Genotypes
Whole Exome Sequencing

DNA methylation, histone acetylation

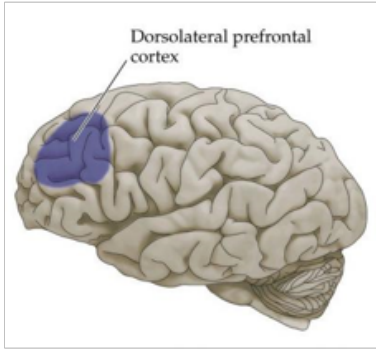
DNA Methylation
Illumina 450K
Histone H3K9Ac
ChIP-Seq

mRNAseq

RNA profile
miRNA & RNAseq

MS Proteomic and metabolomic

LC/MS profile
Lipids, proteins



Quantitative neurobiology

Neuropathology
AD, CVD, LBD, HS, TDP
Gross- and micro- chronic infarctions

Resilience Markers
Synaptic proteins, LC

Structural and function imaging

Flair, MP Rage, DTI, SWI, rsfMRI

post-mortem MRI
DTI, MP Rage, T2

Quantitative clinical phenotype

Cognitive Function
19 tests annually

Motor Function
Disability BMI, Actigraphy

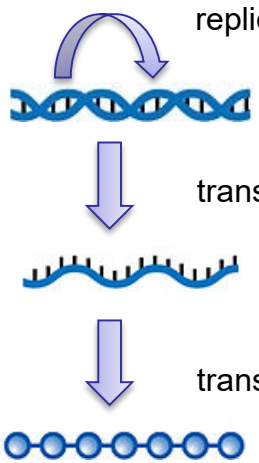
Dynaport – Gait, Sleep, circadian rhythms, Behavioral Economics, Olfaction,

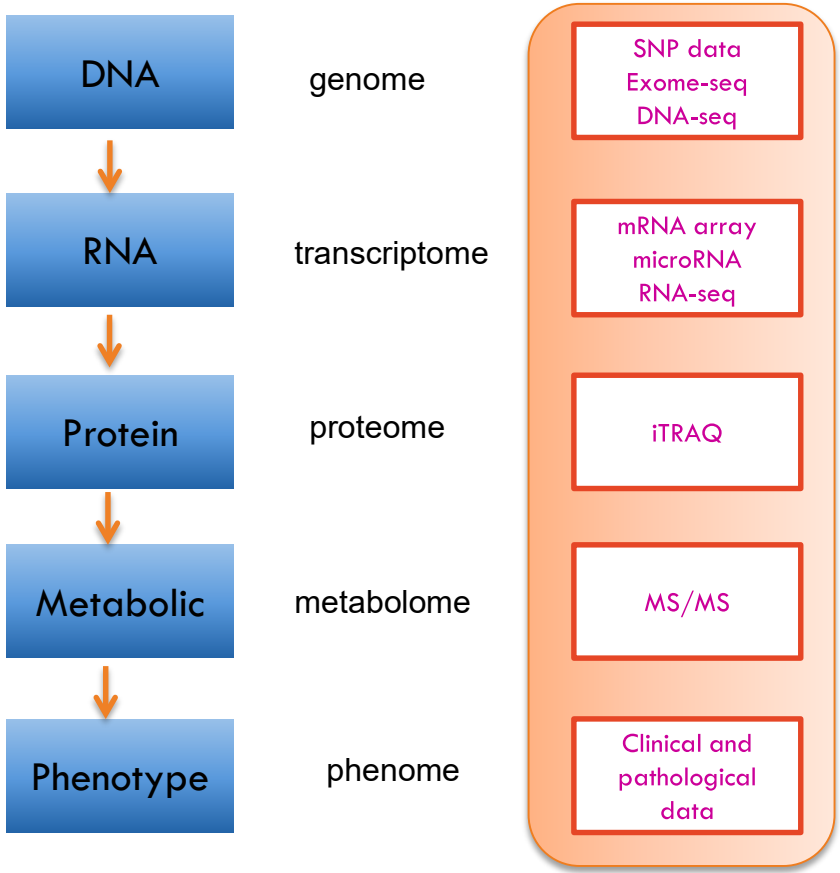
Syndromic phenotype

Clinical Diagnoses
AD, Stroke, PD,

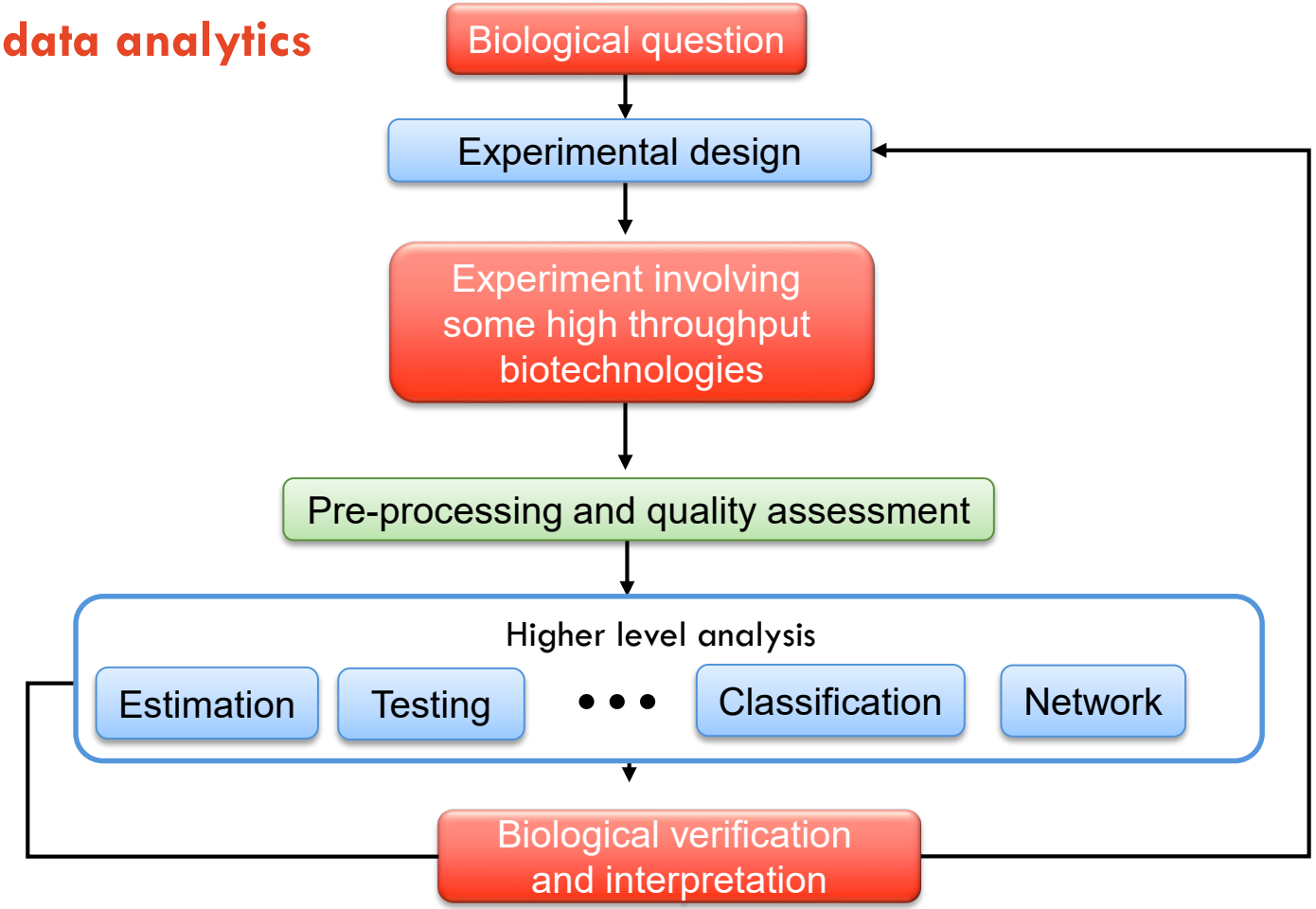
From last week

Types of data

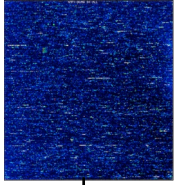




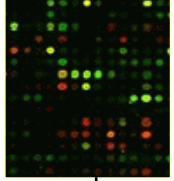
Omics data analytics



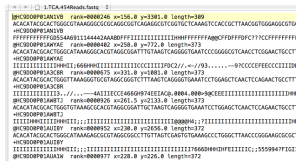
Raw data processing



CEL files



various file types



Fastq files

Quality assessment
Pre-processing

Short-oligonucleotide chip data:
- quality assessment,
- background correction,
- probe-level normalization,
- probe set summary

Two-color array data:
- quality assessment; diagnostic plots,
- background correction,
- array normalization.

RNA-seq data:
- mapping: map to reference,
- annotation
- gene-level summarization
- normalization

Analysis

probes by sample matrix of
- log-ratios or log-intensities (microarray)
- count or RPKM or Total counts (RNA-seq data)

Analysis of expression data:
- Identify D.E. genes, estimation and testing,
- clustering, and
- discrimination.

Expression data - What is your question?

- What are the targets genes for my knock-out gene?
- Look for genes that have different time profiles between different cell types.

  Gene discovery, differential expression

- Is a specified group of genes all up-regulated in a specified conditions?

  Gene set, differential expression

- Are there tumour sub-types not previously identified?
- Are there groups of co-expressed genes?

  Class discovery, clustering

- Can I use the expression profile of cancer patients to predict survival?
- Identification of groups of genes that predictive of a particular class of tumors?

  Class prediction, classification

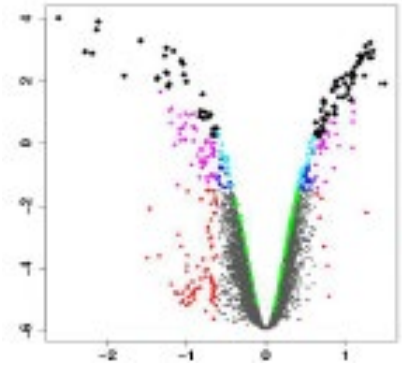
- Detection of gene regulatory mechanisms.
- Do my genes group into previously undiscovered pathways?

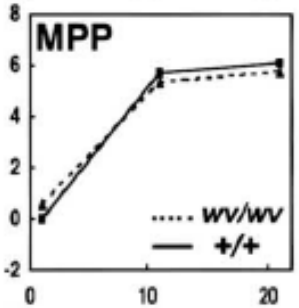
  Clustering. Often expression data alone is not enough, need to incorporate sequence and other information

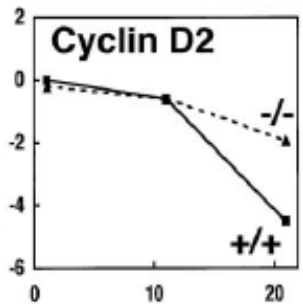
Estimation / Linear Models

Possible questions:

- Identify differential expression genes among two or more tumor subtypes or different cell treatments.
- Is a specified group of genes all up-regulated in a specified conditions?
- Look for genes that have different time profiles between different mutants.
- Looking for genes associated with survival.







Specific methods

- T-tests
- F-tests
- Empirical Bayes
- Moderated t-test
- SAM

Design considerations

- Number of replications and types of replications.
- Main contrasts (comparisons of interest) and how it affect multiple testing.
- Appropriate controls?
- Appropriate time points?

Gene set tests

Questions

Is a specified group of genes all up-regulated in a specified condition?

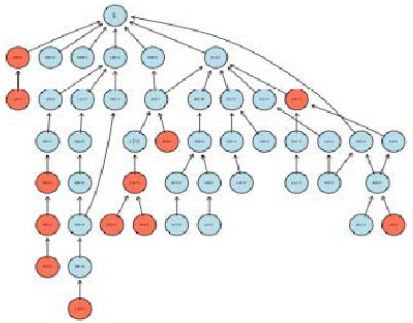
Over- representation of annotations. [E.g. GO, KEGG, TRASFAC].

Specific methods

Fisher exact test – e,g, GOSTAT

Wilcoxon rank-sum test -- GSEA

51	416	467
125	8588	8713
173	9004	9177



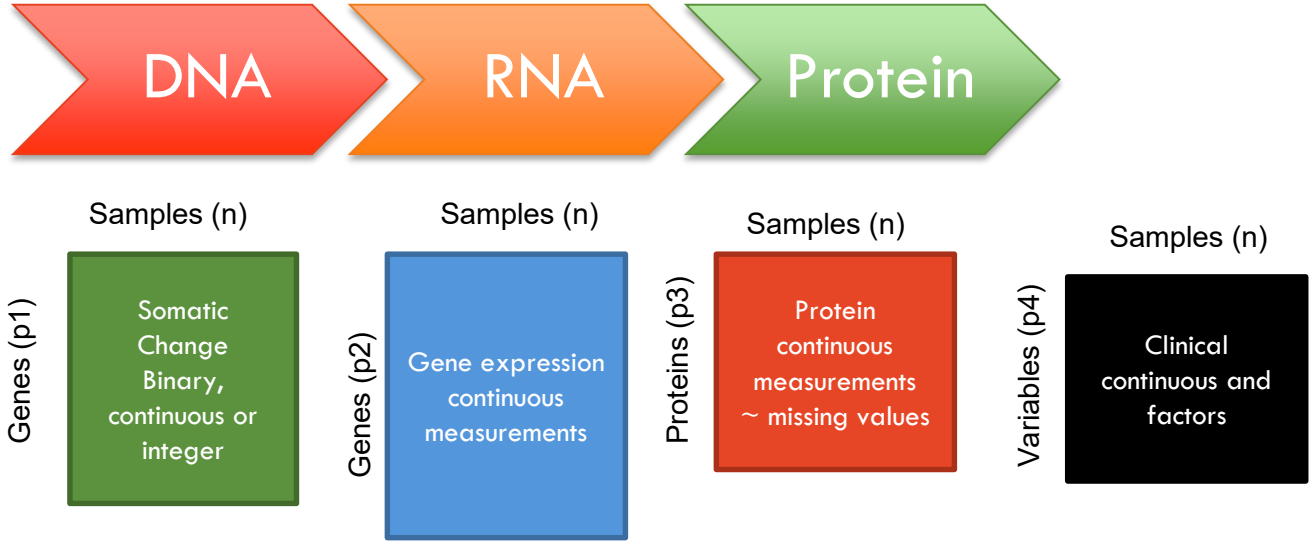
Design considerations

Is gene expression data alone enough ?

What other databases (e.g. GO, sequence, TRASFAC, miRNA) that can be utilized.

Classification

What does biomedical data look like?



Typical questions

How can we find **meaningful biological relationships** between these multiple datasets?

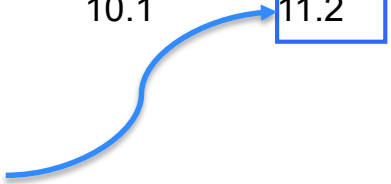
For gene expression data:

Gene expression data on **G** genes for **n** samples

Gene by array data matrix

mRNA samples

Genes	sample1	sample2	sample3	sample4	sample5	...
1	4.63	6.49	2.44	4.55	5.98	...
2	9.05	4.39	12.4	10.4	8.46	...
3	10.4	11.2	10.4	10.3	9.10	...
4	9.33	9.46	8.99	10.9	7.66	...
5	5.33	6.88	10.1	11.2	2.33	...



Gene expression level of gene *i* in mRNA sample *j*

For gene expression data:

Gene expression data on **G** ~~genes~~ for **n** samples.

Gene by array data matrix

samples

Genes		sample1	sample2	sample3	sample4	sample5	...
	1	4.63	6.49	2.44	4.55	5.98	...
	2	9.05	4.39	12.4	10.4	8.46	...
	3	10.4	11.2	10.4	10.3	9.10	...
	4	9.33	9.46	8.99	10.9	7.66	...
	5	5.33	6.88	10.1	11.2	2.33	...

Expression level of ~~gene~~ *i* in mRNA sample *j*

phenotype

Possible input data?

Samples (n)

SNP (p1)

Somatic Change
Binary, continuous or integer

Samples (n)

Proteins (p3)

Protein
continuous measurements
~ missing values

Samples (n)

Genes (p2)

Gene expression
continuous measurements

Samples (n)

Variables (p4)

Clinical
continuous and factors

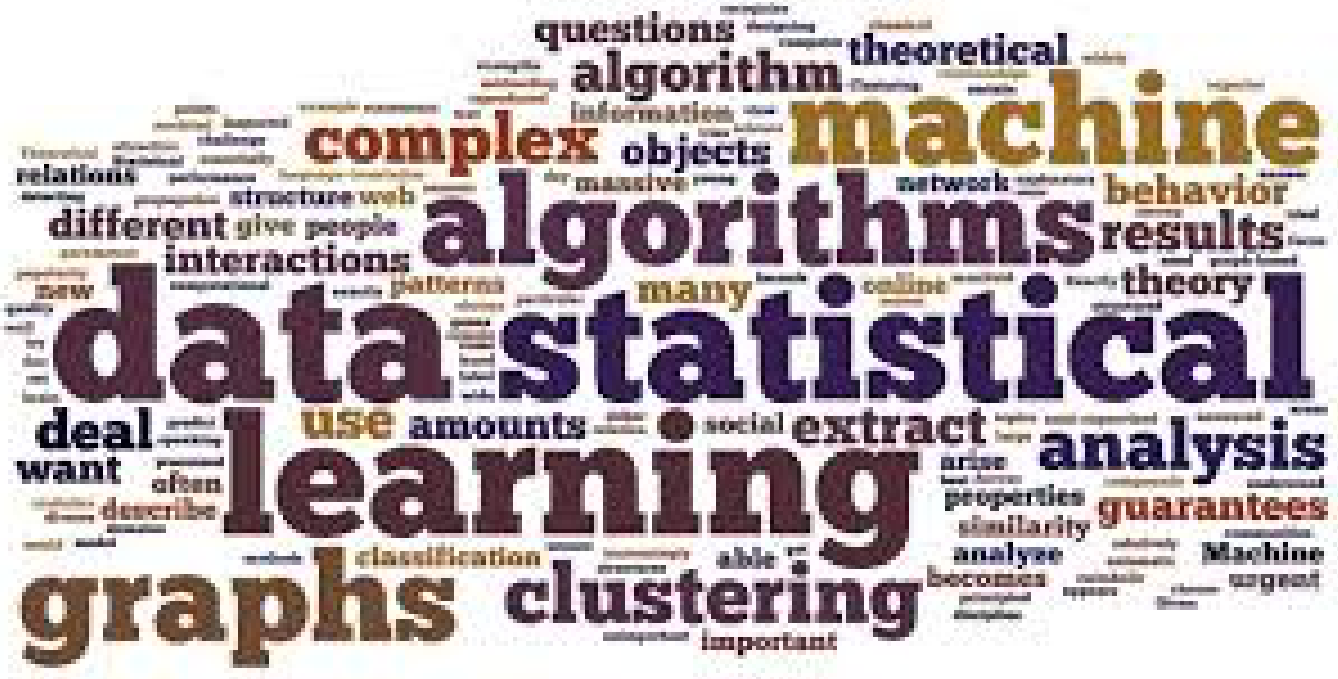
Classification - cluster and discriminant analysis

- Unsupervised: classes unknown, want to discover them from the data (cluster analysis)

- Supervised: classes are predefined, want to use a (training or learning) set of labeled objects to form a classifier for classification of future observations

- Alternative terminology
  - Computer science: unsupervised and supervised learning.
  - Bioinformatics literature: class discovery and class prediction.
  - Statistics: Clustering and Discriminant analysis

Many different terms



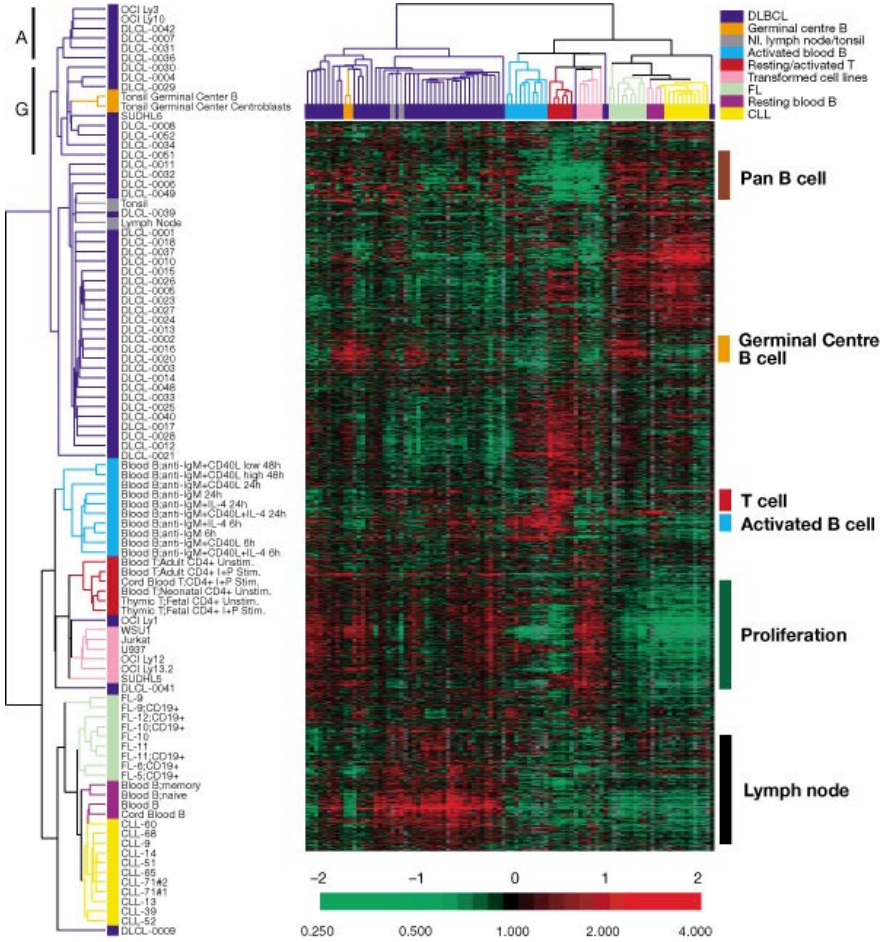
Clustering

Example 1:

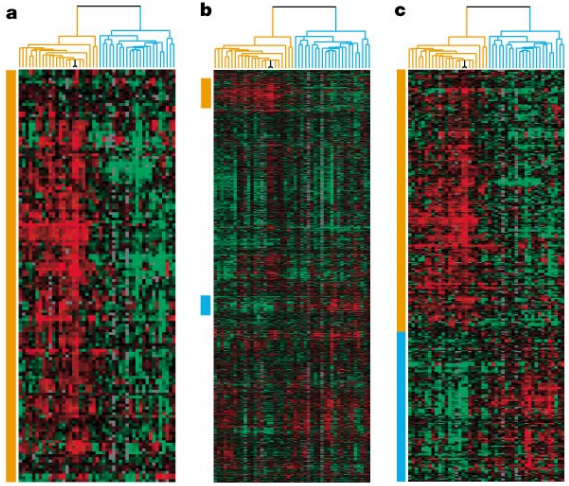
Clustering samples and genes with expression arrays

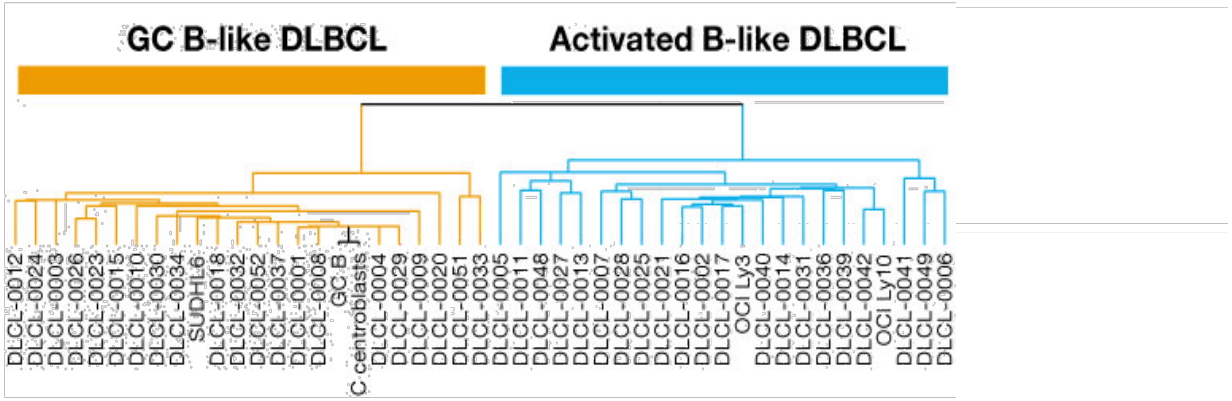
A Alizadeh *et al* *Nature 2000.*

Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling

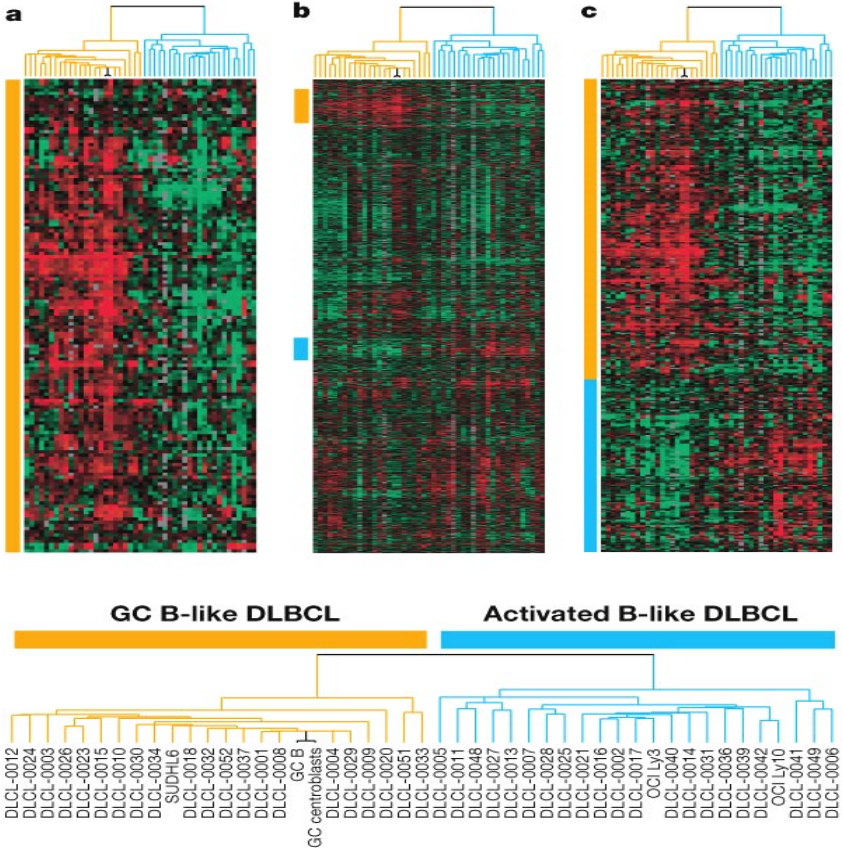


Subtype discovery Example: B-cell lymphoma (Alizadeh et al 2000)





Clustering cell samples
Discovering sub-groups



Taken from Alizadeh *et al* *(Nature, 2000)*

Example 2:

Clustering samples to discover cell types

Cell Stem Cell

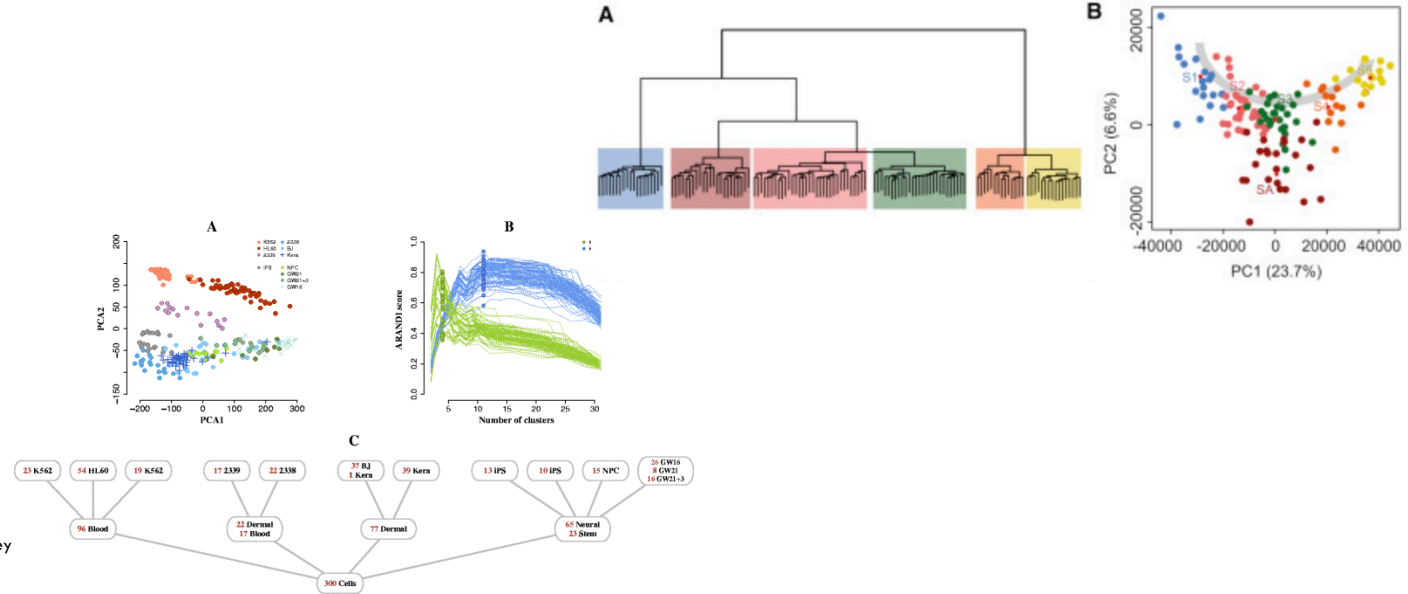
Volume 17, Issue 3, 3 September 2015, Pages 360–372

Resource

Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis

Jaehoon Shin[1, 2], Daniel A. Berg[2, 3, 7], Yunhua Zhu[2, 3], Joseph Y. Shin[4], Juan Song[2, 3], Michael A. Bonaguidi[2, 3], Grigori Enikolopov[8, 9], David W. Nauen[5], Kimberly M. Christian[2, 3], Guo-li Ming[1, 2, 3, 4, 6], Hongjun Song[1, 2, 3, 4],

Show more



Clustering: quality assessment

Caution with removal of samples.

Question:

How good is my replication?

Design:

1. Identify genes with high **F-statistic** among multiple cell types.
2. Use these genes to cluster the **samples**.
3. Remove replicates that did not cluster together.

Issues:

Further investigation is required.

Clustering: quality assessment

Caution with genes selection

Question: Do expression data cluster according to the survival status.

Design:

1. Identify genes with high **t-statistic** between **short** and **long** survivors.
2. Use these genes to cluster **samples**.
3. Get excited that samples cluster according to survival status.

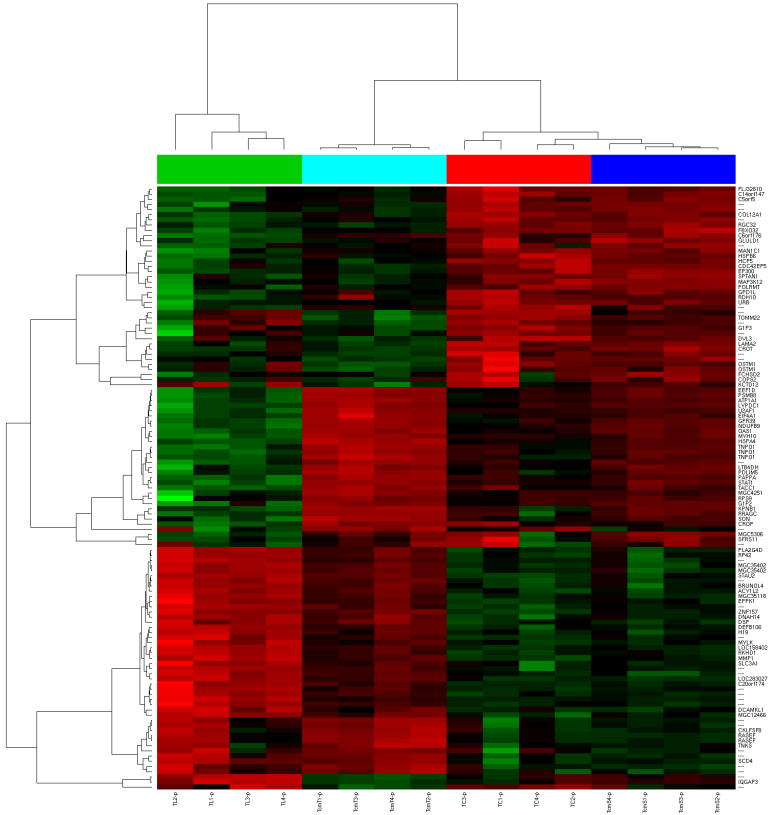
Issues:

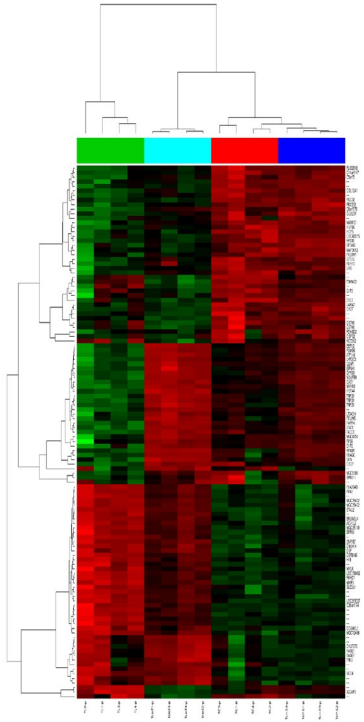
The genes were already selected based on the survival status. Therefore, it would rather be surprising if samples did **NOT** cluster according to their survival.

Valid conclusion are not possible as variable selection was driven by class distinction.

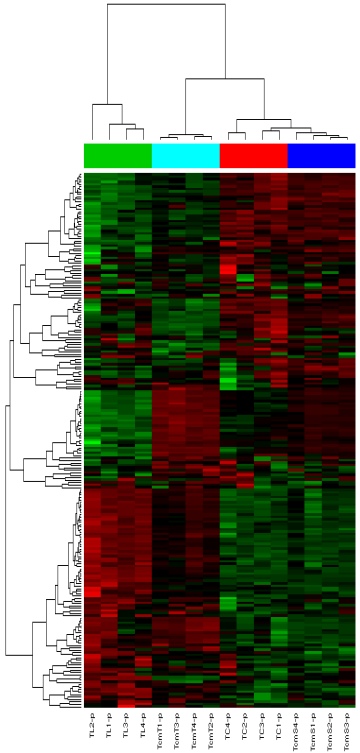
Quality assessment example:

Use clustering to check within/between experiment group variability and potential confounding factors (batch effect etc)

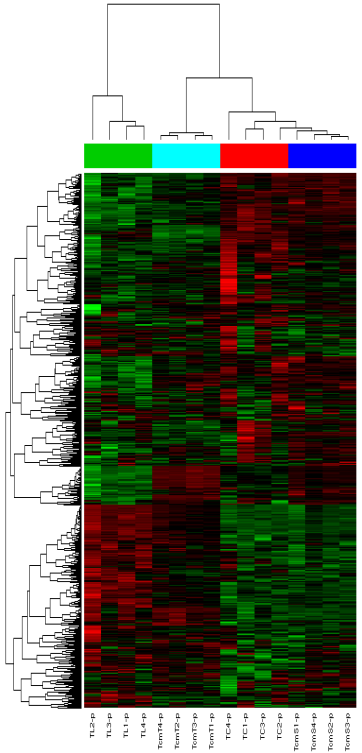




Top 100 most variable genes



+ 100 random genes



+ 500 random genes

Clustering - Summary

- Useful as **exploratory/visualization** tools
- Usually outside the normal framework of statistical inference
- Choice of metric, methods and parameters usually guided by prior knowledge about the question… **The result is guided by what you are looking for!**
- Be aware…
  - Clustering cannot NOT work. Always produce some clusters!
  - Test for covariate-cluster association is valid only when cluster definition is independent of the covariate.

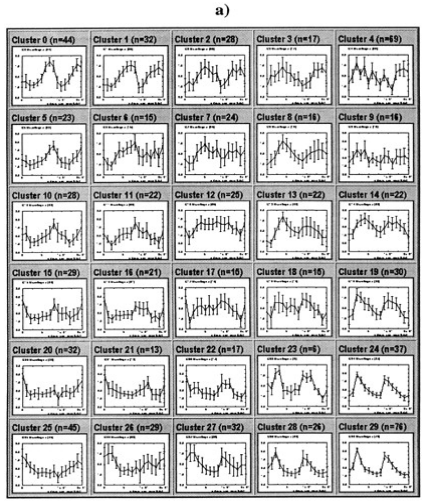
Clustering - class discovery

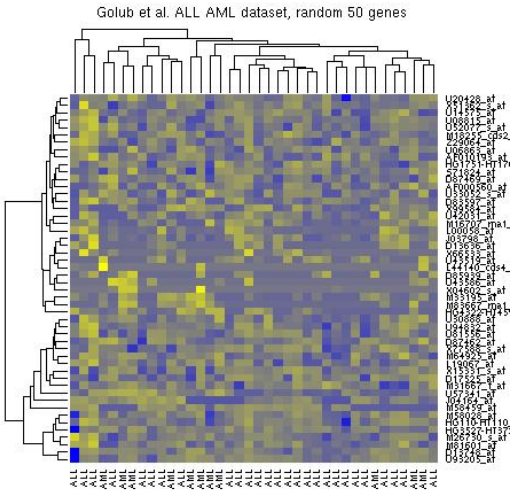
Questions

We can cluster cell samples (cols),
the identification of new / unknown tumor sub classes or cell sub types using gene expression profiles.
We can cluster genes (rows) ,
to identify groups of co-expressed genes.

Algorithms

- Hierarchical clustering
- Self-organizing maps
- Partition around medoids (pam)





Design considerations

1) Clutsering cell samples

- Large number of samples.
- How to sample from the population?

2) Clutsering cell genes

- Number of conditions.
- Replications.

Discrimination

THE UNIVERSITY OF SYDNEY

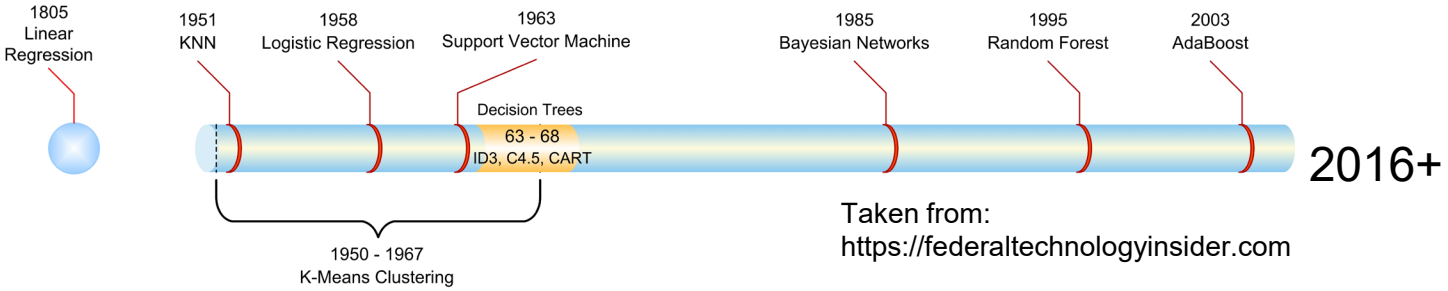
Clustering and discrimination

- Unsupervised: classes unknown, want to discover them from the data (cluster analysis)

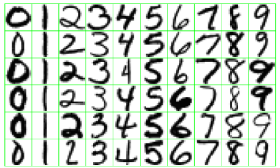
- Supervised: classes are predefined, want to use a (training or learning) set of labeled objects to form a classifier for classification of future observations

History

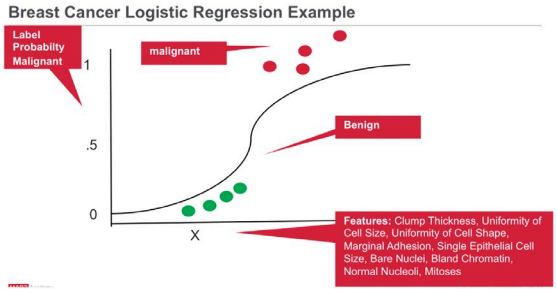
Time Line of Machine Learning Algorithms



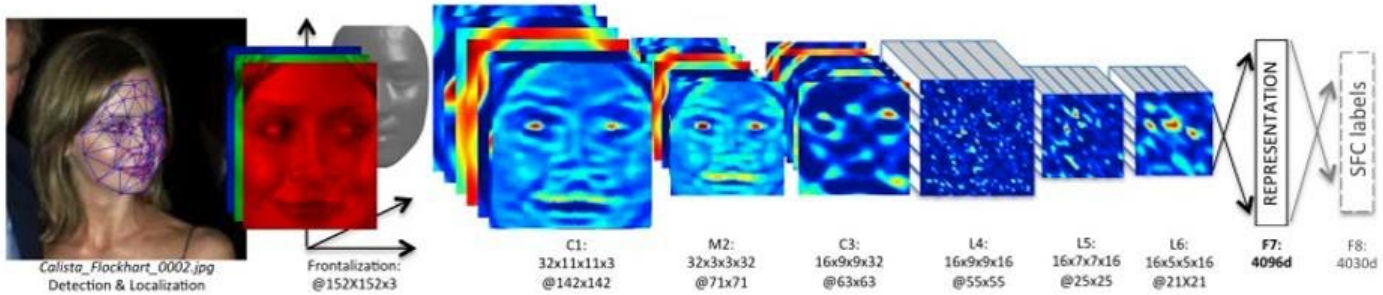
Taken from:
https://federaltechnologyinsider.com



Deep learning



Application of deep learning in our lives



MAR 18, 2014 @ 05:58 PM

Facebook's DeepFace Software Can Match Faces With 97.25% Accuracy

Amit Chowdhry, CONTRIBUTOR

I cover noteworthy technology, startups and gadgets **FULL BIO**

Opinions expressed by Forbes Contributors are their own.

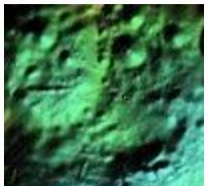
Forbes / Tech

https://www.forbes.com/sites/amitchowdhry/2014/03/18/facebooks-deepface-software-can-match-faces-with-97-25-accuracy/#5408054127e5

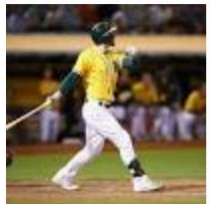
Deep learning in the news …

MIT Uses **Deep Learning** to Create ICU, EHR Predictive Analytics
Health IT Analytics - 3 hours ago
Deep learning and natural language processing are contributing to two new approaches to EHR predictive analytics and clinical decision ...

MIT projects explore machine **learning** applications to improve EHRs
Healthcare IT News - 4 hours ago



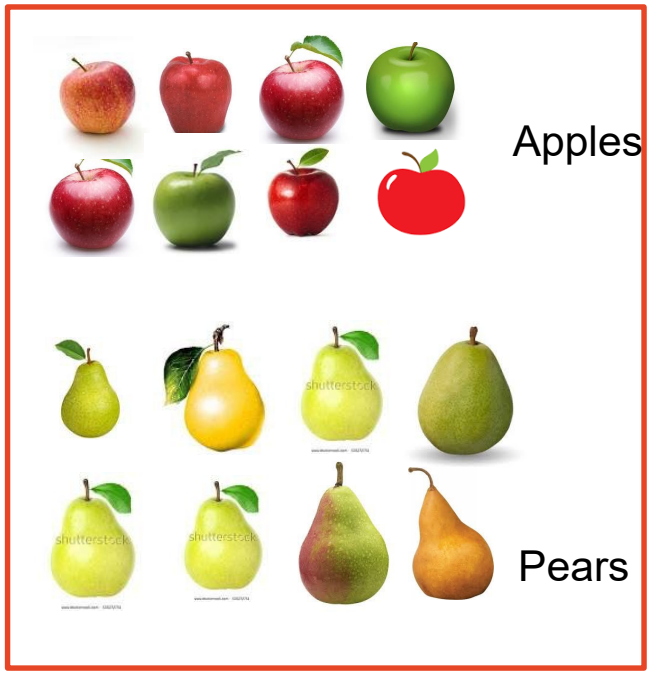
NASA is using Intel's **deep learning** to find better landing sites on the ...
TechCrunch - 18 Aug. 2017
That partnership includes Nervana, the **deep learning** startup the cl acquired back in 2016, which it's leveraging to translate those ...



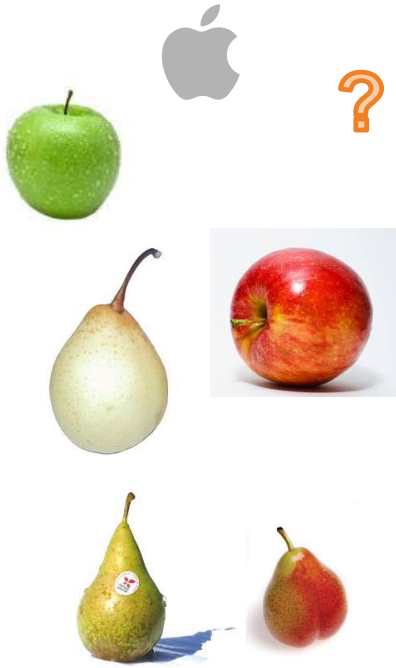
How Is **Deep Learning** Changing The World Of Sports?
Forbes - 16 Aug. 2017
How is **deep learning** affecting sports? originally appeared on Quora: the place ...
Deep learning has the potential to use simple data like video ...

Why **deep learning** won't replace its human counterparts anytime soon
TechRepublic - 21 Aug. 2017
Even so, **deep learning**, based on artificial neural networks, offers real promise. It's just that it doesn't promise to replace humans.

Concept of machine learning



Build a model



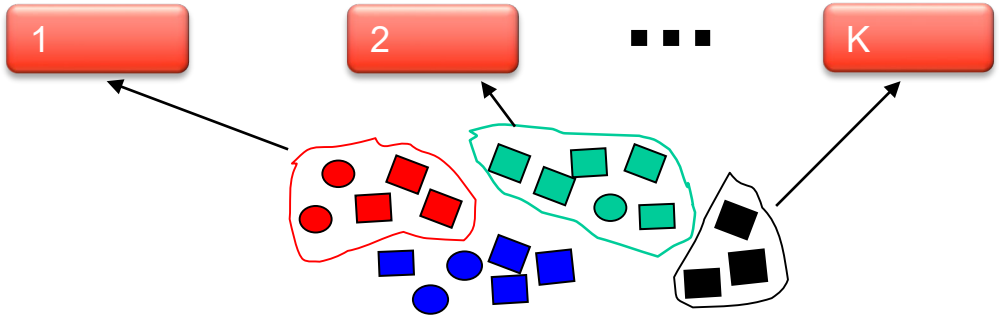
Training Data

Predicting our new data – apple or pear ?

Basic principles of discrimination

•Each object associated with a class label (or response) $Y \in \{1, 2, \ldots, K\}$ and a feature vector (vector of predictor variables) of G measurements: $\mathbf{X} = (X_1, \ldots, X_G)$

Aim: predict Y from **X**.



Predefined class {1,2,…K}

Objects

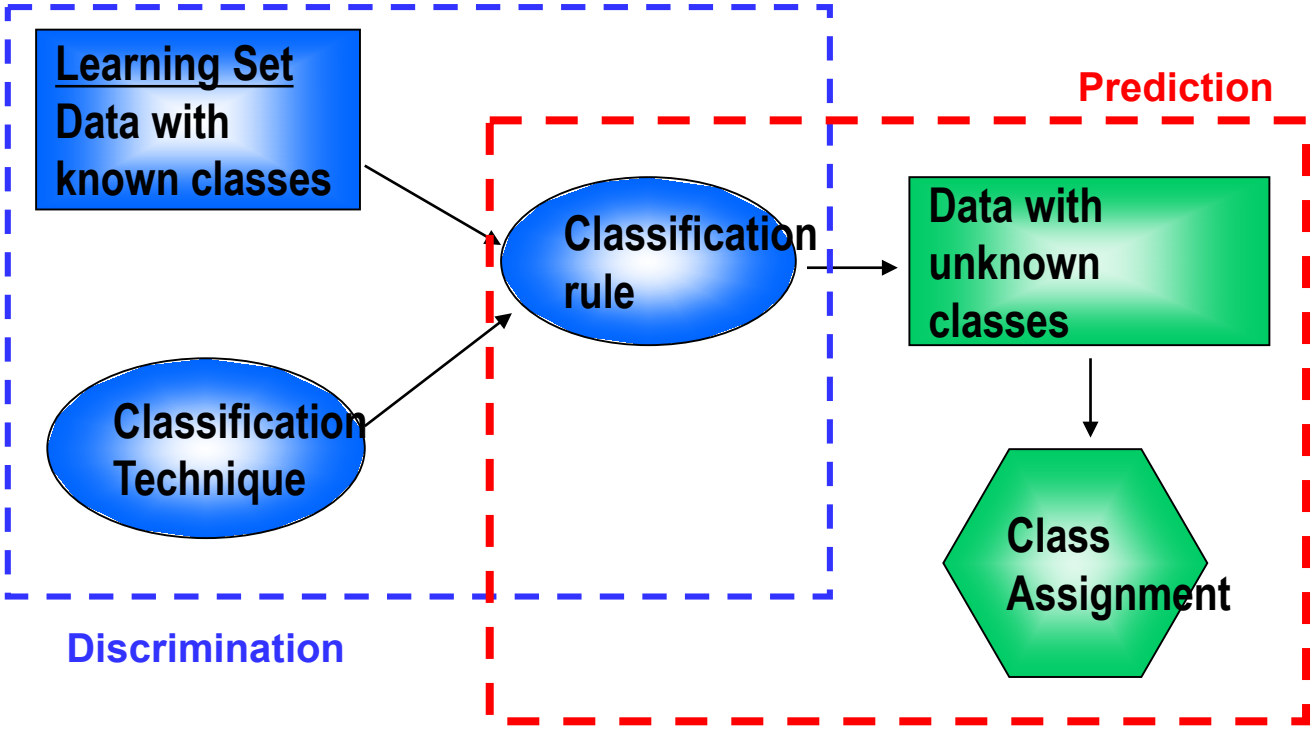
Y = Class Label = 2

X = Feature vector
{colour, shape}

Classification rule ?

X = {red, square}
Y = ?

Classification



Case Study:
Van't Veer breast cancer study study

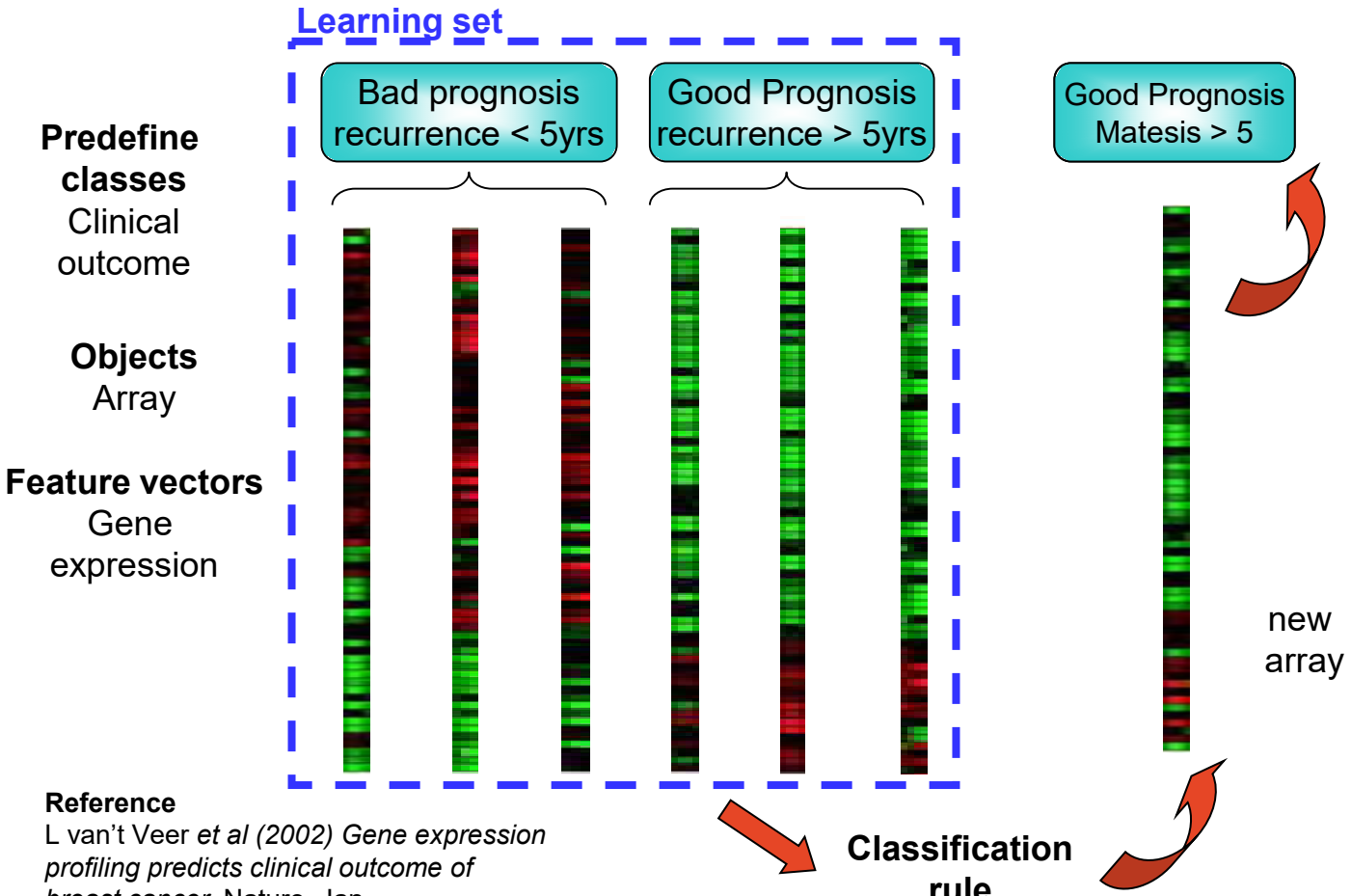
- Investigate whether tumor ability for metastasis is obtained later in development or inherent in the initial gene expression signature.

- Retrospective sampling of node-negative women: 44 non-recurrences within 5 years of surgery and 34 recurrences. Additionally, 19 test sample (12 recur. and 7 non-recur).

- Want to demonstrate that gene expression profile is independently predictive of the recurrence.

Reference
L van't Veer *et al (2002)*
Gene expression profiling predicts clinical outcome of breast cancer. Nature, Jan.

.



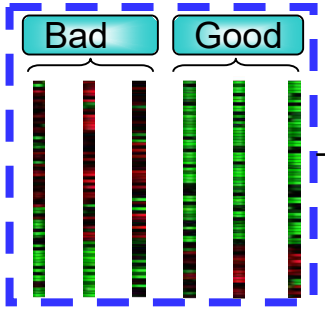
Reference

L van't Veer *et al (2002) Gene expression profiling predicts clinical outcome of breast cancer.* Nature, Jan.

.

Case studies

Learning set



Classification Rule

Feature selection.
Correlation with class labels, very similar to t-test.

Using cross validation to select 70 genes

295 samples selected
from Netherland Cancer Institute tissue bank (1984 – 1995).

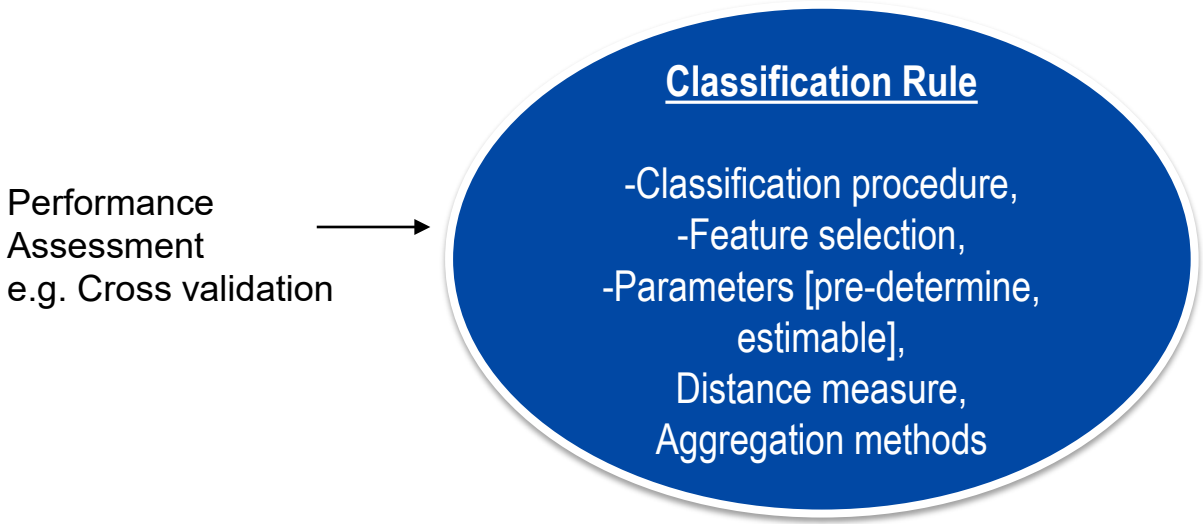
Results" Gene expression profile is a more powerful predictor then standard systems based on clinical and histologic criteria

Agendia (formed by reseachers from the Netherlands Cancer Institute)
Plan to start in Oct, 2003
1) 3000 subjects [Health Council of the Netherlands]
2) 5000 subjects New York based Avon Foundation.
Custorm arrays are made by Aglient including
70 genes + 1000 controls

Reference 1
Retrospective study
L van't Veer *et al Gene expression profiling predicts clinical outcome of breast cancer.* Nature, Jan 2002.
.

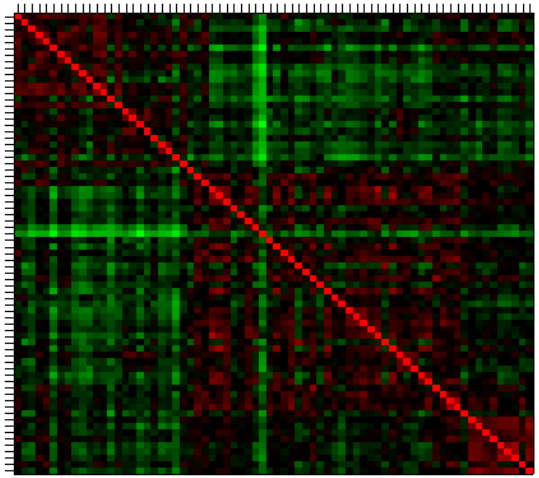
Reference 2
Retrospective study
M Van de Vijver *et al*. A gene expression signature as a predictor of survival in breast cancer. The New England Jouranl of Medicine, Dec 2002.

Reference 3
Prospective trials.
Aug 2003
Clinical trials
http://www.agendia.com/

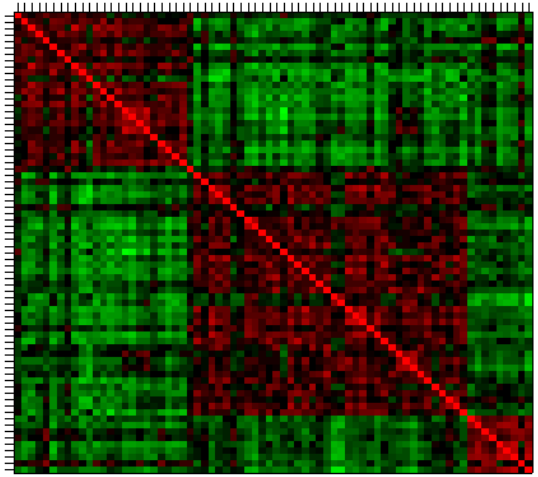


- One can think of the classification rule as a black box, some methods provides more insight into the box.
- Performance assessment needs to be looked at for all classification rule.

Why select features?



No feature selection

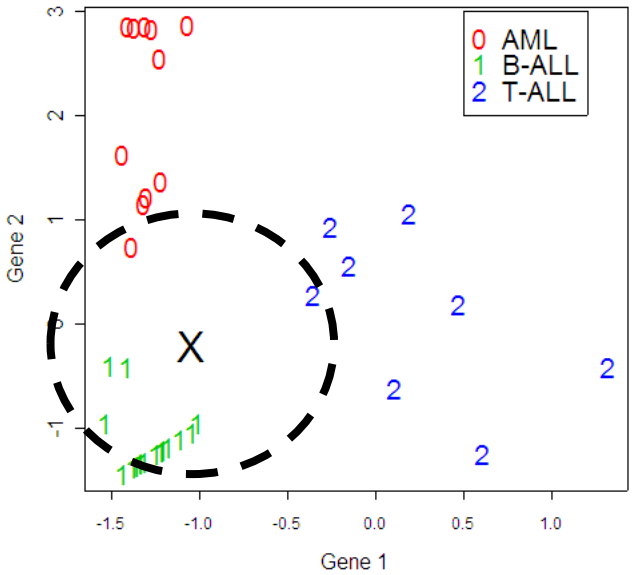


Top 100 feature selection
Selection based on variance

Correlation plot
Data: Leukemia, 3 class

Nearest neighbor classification

- Based on a measure of distance between observations (e.g. Euclidean distance or one minus correlation).
- k-nearest neighbor rule (Fix and Hodges (1951)) classifies an observation X as follows:
  - find the k observations in the learning set closest to X
  - predict the class of X by majority vote, i.e., choose the class that is most common among those k observations.
- The number of neighbors k can be chosen by cross-validation (more on this later).

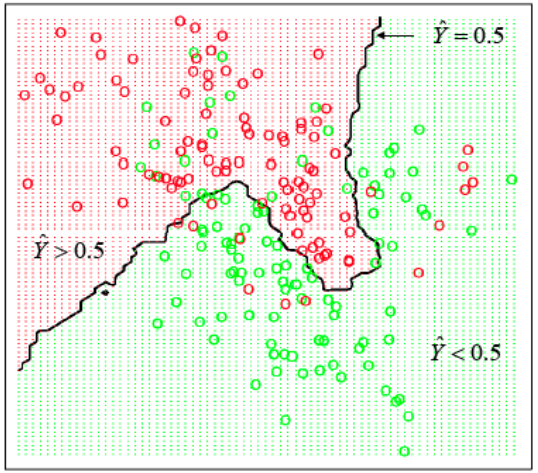


Classification - k Nearest Neighbor

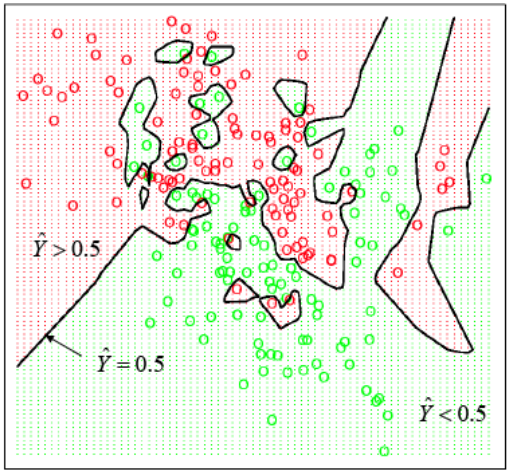
$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

$N_k(x)$: the k closest points to x

15-nearest neighbor averaging

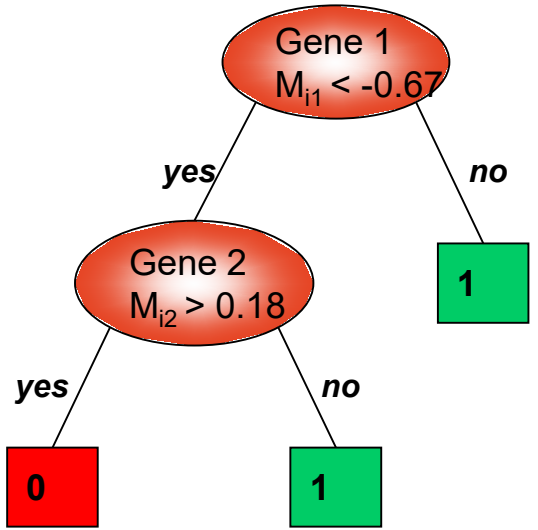


1-nearest neighbor averaging

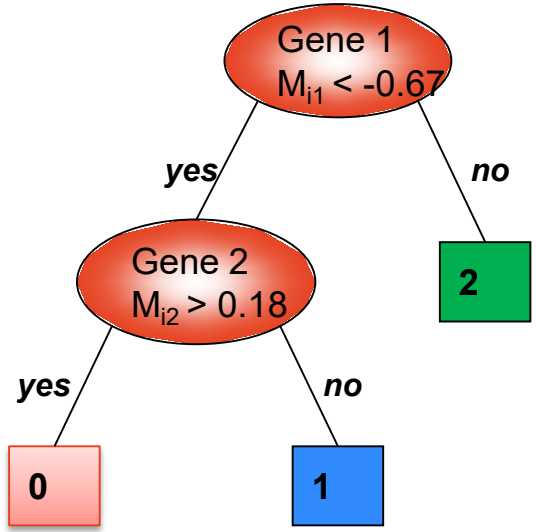


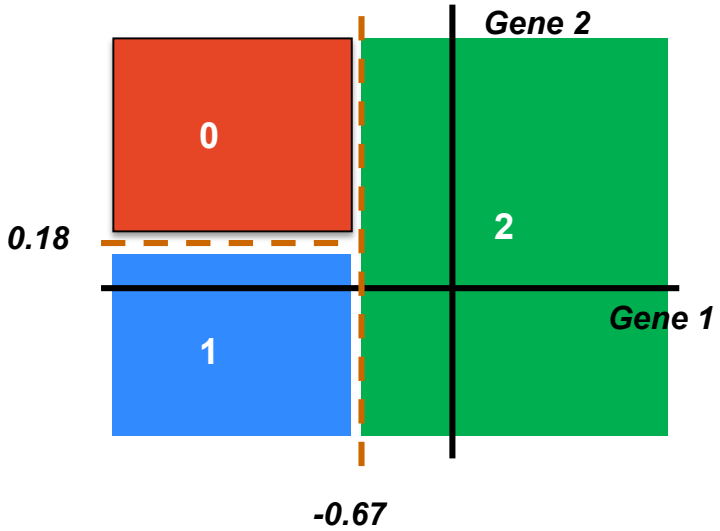
Classification tree

- Partition the feature space into a set of rectangles, then fit a simple model in each one
- Binary tree structured classifiers are constructed by repeated splits of subsets (nodes) of the measurement space X into two descendant subsets (starting with X itself)
- Each terminal subset is assigned a class label; the resulting partition of X corresponds to the classifier



Classification tree





A quick summary

- Classical Maximum Likelihood classifiers:
  - Linear Discriminant Analysis (LDA)
  - DLDA
  - DQDA
  - K-Nearest Neighbour Classifiers
- Modern LDA Derivatives:
  - PAMR (http://www-stat.stanford.edu/~tibs/PAM/)
  - SCRDA
- Support Vector Machines (SVM)
- Aggregated Trees (CART)
- Other classifiers:
  - Neural networks (NN)
  - Bayesian belief networks

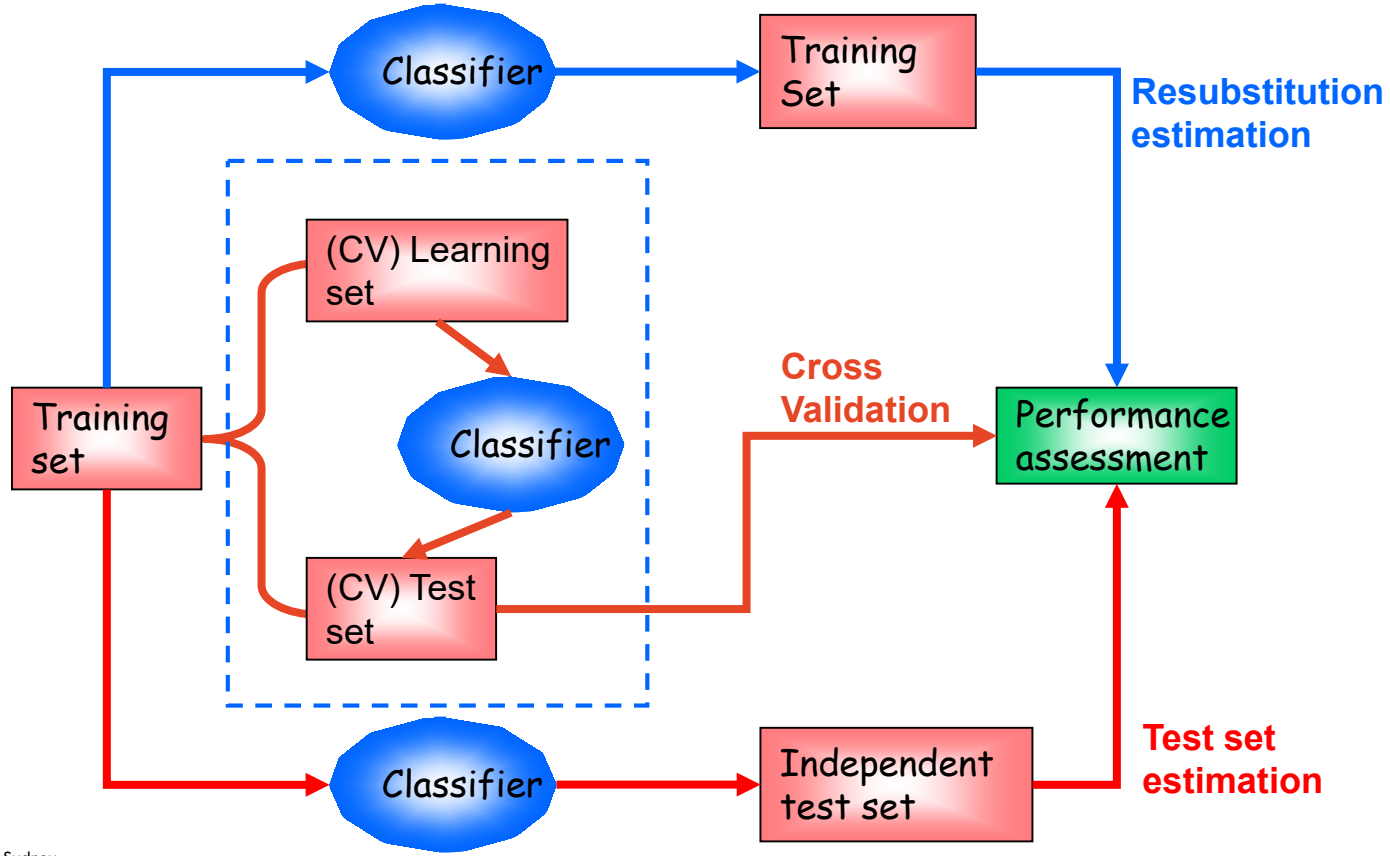
Performance assessment

- Any **classification rule** needs to be **evaluated** for its performance on the future samples. It is almost never the case in omics studies that a large independent population-based collection of samples is available at the time of initial classifier-building phase.

- One needs to estimate future performance based on what is available: often the same set that is used to build the classifier.

- Assessing performance of the classifier based on
  - Cross-validation.
  - Test set.
  - Independent testing on future dataset.
  - Independent testing on existing dataset (integrative analysis).

Diagram of performance assessment



Performance assessment

- [1] **Resubstitution estimation:** error rate on the learning set.
  - Problem: downward bias

- [2] **Test set estimation**:
  1. Divide learning set into two sub-sets, L and T; Build the classifier on L and compute the error rate on T.
  2. Build the classifier on the training set (L) and compute the error rate on an independent test set (T).
     1. L and T must be independent and identically distributed (i.i.d).
     2. Problem: reduced effective sample size

Performance assessment

- [3] **V-fold cross-validation (CV) estimation:** Cases in learning set randomly divided into V subsets of (nearly) equal size. Build classifiers by leaving one set out; compute test set error rates on the left out set and averaged.
  - Bias-variance tradeoff: smaller V can give larger bias but smaller variance
  - Computationally intensive.

  [3a] **Leave-one-out cross validation (LOOCV).**

  (Special case for V=n). Works well for stable classifiers (k-NN, LOOCV)