Omics data analysis

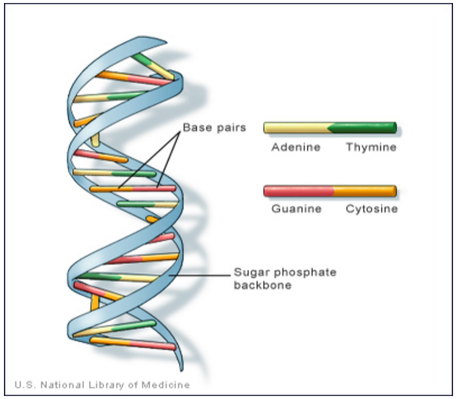
Torgeir R. Hvidsten

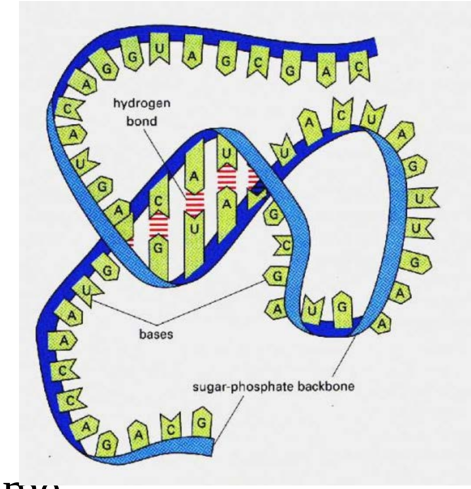
Assistant professor in Bioinformatics

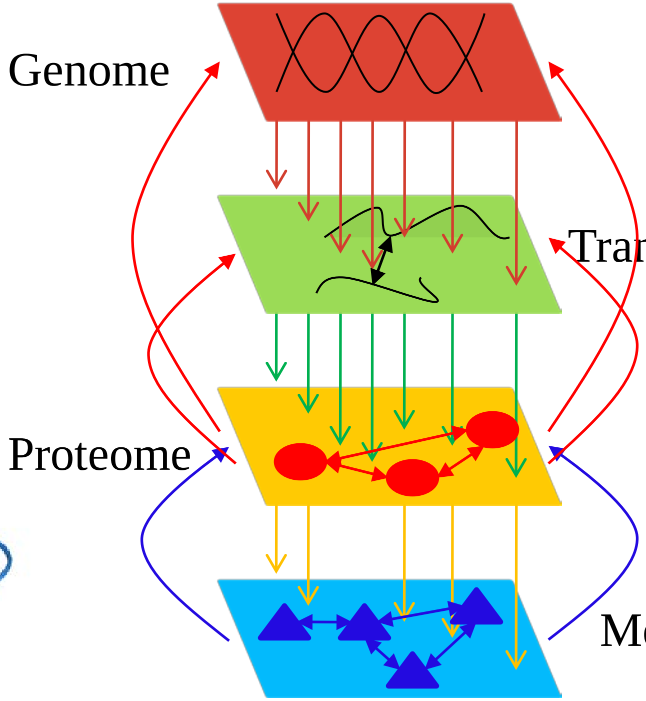
Umeå Plant Science Center (UPSC)

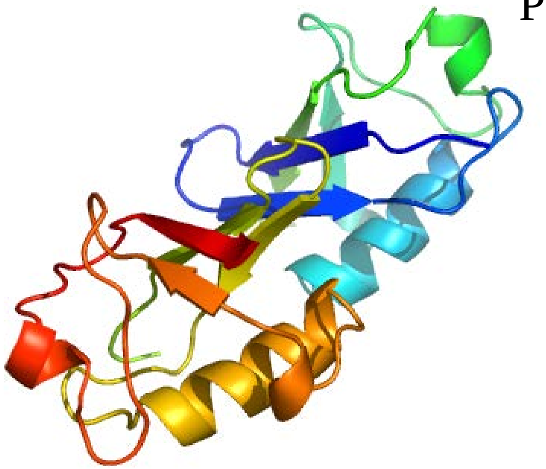
Computational Life Science Centre (CLiC)

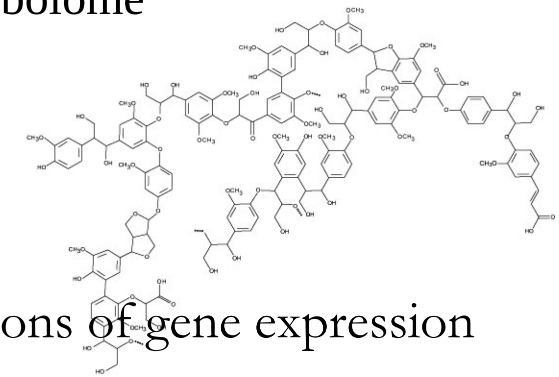
’omics data











- <span>Transcriptomics</span> - quantifications of gene expression
- Proteomics - quantifications of proteins (peptides)
- Metabolomics - quantifications of metabolites

Gene expression data

M < 100

Gene/Expr	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	…	EM
G1	-0.47	-3.32	-0.81	0.11	-0.60	-1.36	-1.03	-1.84	-1.00	-0.60	…	-0.94
G2	0.66	0.07	0.20	0.29	-0.89	-0.45	-0.29	-0.29	-0.15	-0.45	…	-0.42
G3	0.14	-0.04	0.00	-0.15	-0.58	-0.30	-0.18	-0.38	-0.49	-0.81	…	-1.12
G4	-0.04	0.00	-0.23	-0.25	-0.47	-0.60	-0.56	-1.09	-0.71	-0.76	…	-0.62
G5	0.28	0.37	0.11	-0.17	-0.18	-0.60	-0.23	-0.58	-0.79	-0.29	…	-0.74
G6	0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36	-0.49	-0.58	…	-1.47
G7	0.20	0.14	0.00	0.11	-0.34	-0.03	0.04	-0.76	-0.81	-1.12	…	-1.36
G8	0.40	0.43	0.18	0.00	-0.14	0.29	0.07	-0.79	-0.81	-0.92	…	-1.22
G9	0.01	0.46	0.28	-0.34	-0.23	-0.36	-0.45	-0.64	-0.79	-1.22	…	-1.09
…	…	…	…	…	…	…	…	…	…	…	…	…
GN	-0.23	0.04	0.00	-0.30	-0.29	-0.45	-0.97	-2.06	-0.89	-1.22	…	-0.97

N > 10000

Two-channel experiments:	ratio-based intensities (“Red/Green”)
One-channel experiments:	”absolut” intensities
RNA-Seq:	”number” of transcripts expressed

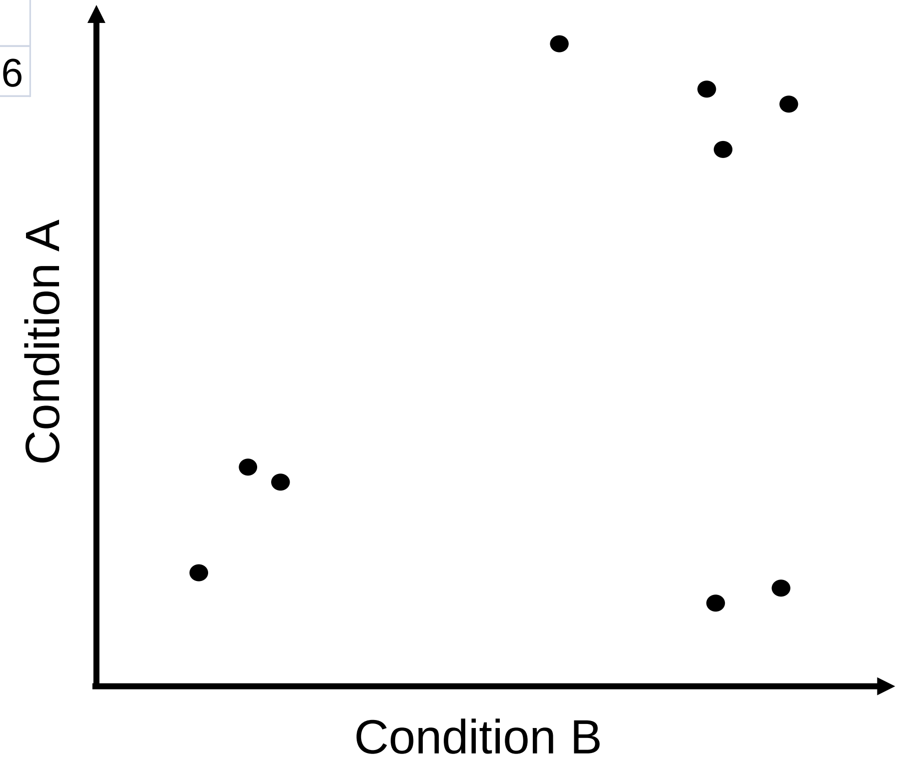
’omics data

Conditions/tissues/time

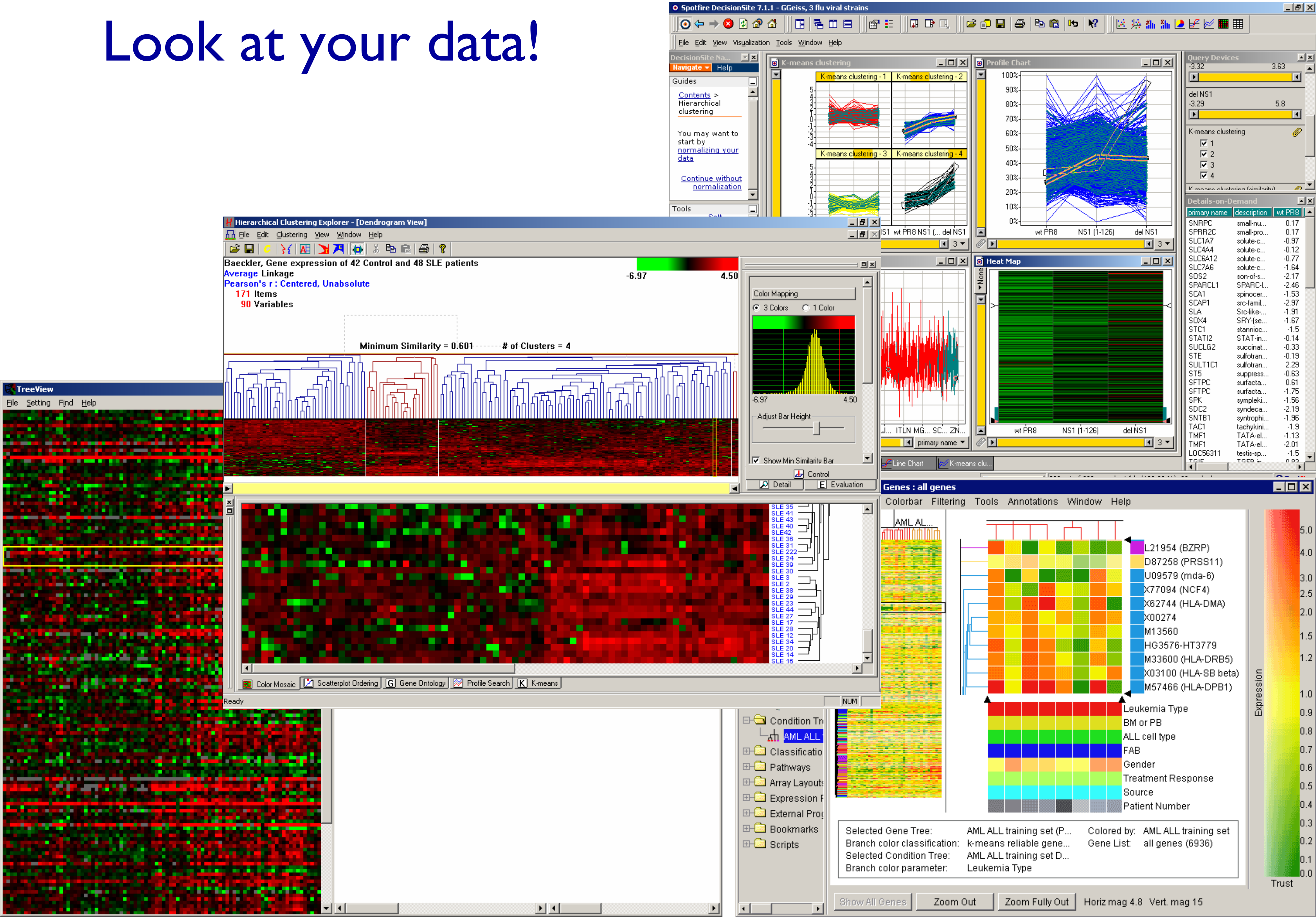
Genes/metabolites/proteins

0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36
-0.47	-3.32	-0.81	0.11	-0.60	-1.36	-1.03	-1.84
0.66	0.07	0.20	0.29	-0.89	-0.45	-0.29	-0.29
0.14	-0.04	0.00	-0.15	-0.58	-0.30	-0.18	-0.38
-0.04	0.00	-0.23	-0.25	-0.47	-0.60	-0.56	-1.09
0.28	0.37	0.11	-0.17	-0.18	-0.60	-0.23	-0.58
0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36
0.20	0.14	0.00	0.11	-0.34	-0.03	0.04	-0.76
0.40	0.43	0.18	0.00	-0.14	0.29	0.07	-0.79
0.01	0.46	0.28	-0.34	-0.23	-0.36	-0.45	-0.64
…	…	…	…	…	…	…	…
-0.23	0.04	0.00	-0.30	-0.29	-0.45	-0.97	-2.06

Time series versus
Feature space



Look at your data!



Spotfire DecisionSite 7.1.1 - GGeiss, 3 flu viral strains

You may want to start by normalizing your data

Continue without normalization

Hierarchical Clustering Explorer - [Dendrogram View]

Baeckler, Gene expression of 42 Control and 48 SLE patients
Average Linkage
Pearson's r : Centered, Unabsolute
171 Items
90 Variables

Minimum Similarity = 0.601 # of Clusters = 4

Details-on-Demand

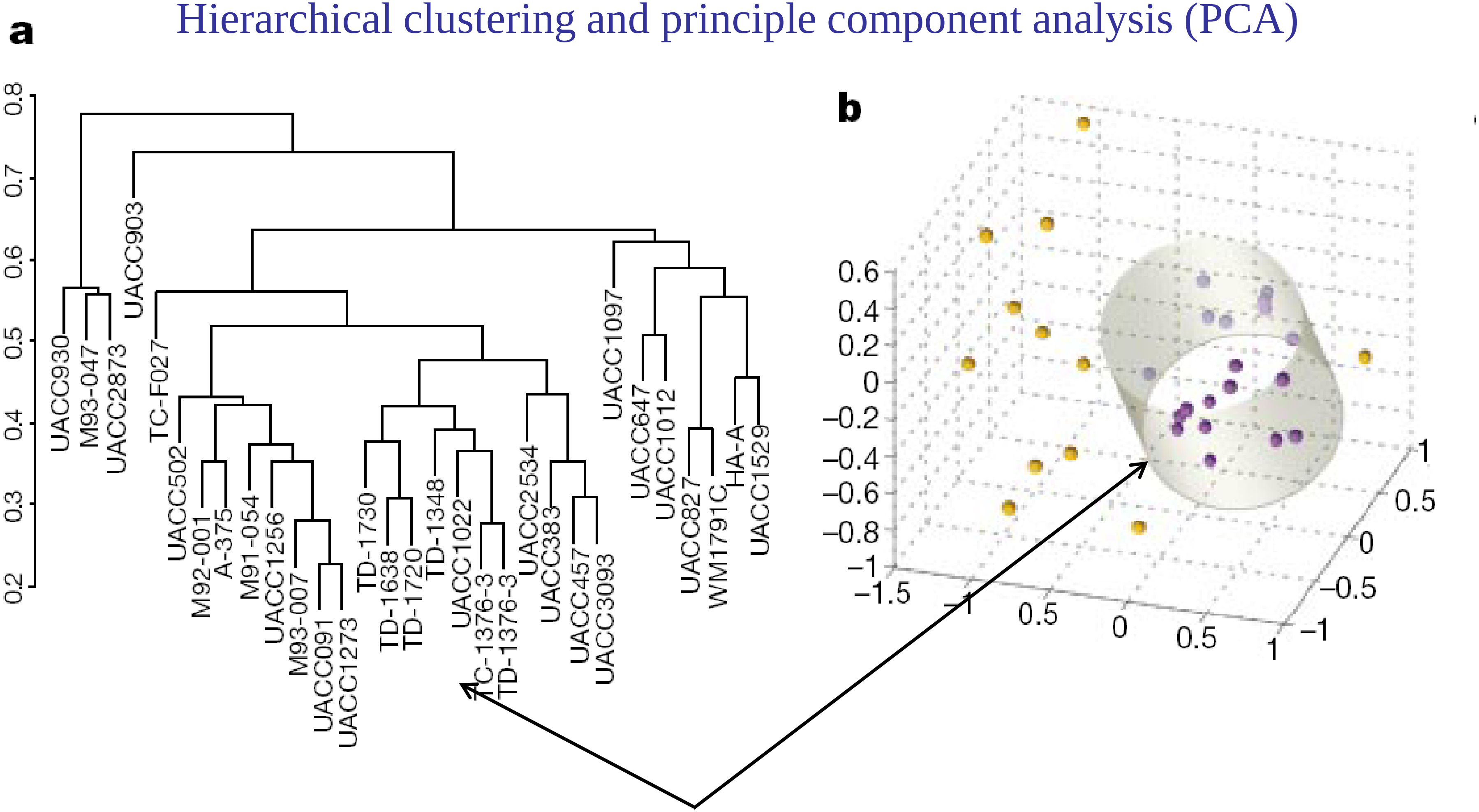
primary name	description	wt PR8
SNRPC	small-nu...	0.17
SPRR2C	small-pro...	0.17
SLC1A7	solute-c...	-0.97
SLC4A4	solute-c...	-0.12
SLC6A12	solute-c...	-0.77
SLC7A6	solute-c...	-1.64
SOS2	son-of-s...	-2.17
SPARCL1	SPARC-l...	-2.46
SCA1	spinocer...	-1.53
SCAP1	src-famil...	-2.97
SLA	Src-like-...	-1.91
SOX4	SRY-(se...	-1.67
STC1	stannioc...	-1.5
STATI2	STAT-in...	-0.14
SUCLG2	succinat...	-0.33
STE	sulfotran...	-0.19
SULT1C1	sulfotran...	2.29
ST5	suppress...	-0.63
SFTPC	surfacta...	0.61
SFTPC	surfacta...	-1.75
SPK	sympleki...	-1.56
SDC2	syndeca...	-2.19
SNTB1	syntrophi...	-1.96
TAC1	tachykini...	-1.9
TMF1	TATA-el...	-1.13
TMF1	TATA-el...	-2.01
LOC56311	testis-sp...	-1.5

Genes : all genes

Selected Gene Tree:	AML ALL training set (P...	Colored by:	AML ALL training set
Branch color classification:	k-means reliable gene...	Gene List:	all genes (6936)
Selected Condition Tree:	AML ALL training set D...		
Branch color parameter:	Leukemia Type		

Looking into more than 3D:

Hierarchical clustering and principle component analysis (PCA)



19 melanomas of all 31 cutaneous melanoma samples
(Bitter et al. *Nature*. 406: 536, 2000)

Machine learning

- Supervised learning; used to learn a model from a set of examples with predefined classes of genes/experiments (training set)

- Unsupervised learning (clustering, class discovery); used to “discover” natural groups of genes/experiments

Training examples

M < 100

Gene/Expr	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**	**E9**	**E10**	**…**	**EM**
G1	-0.47	-3.32	-0.81	0.11	-0.60	-1.36	-1.03	-1.84	-1.00	-0.60	…	-0.94
G2	0.66	0.07	0.20	0.29	-0.89	-0.45	-0.29	-0.29	-0.15	-0.45	…	-0.42
G3	0.14	-0.04	0.00	-0.15	-0.58	-0.30	-0.18	-0.38	-0.49	-0.81	…	-1.12
G4	-0.04	0.00	-0.23	-0.25	-0.47	-0.60	-0.56	-1.09	-0.71	-0.76	…	-0.62
G5	0.28	0.37	0.11	-0.17	-0.18	-0.60	-0.23	-0.58	-0.79	-0.29	…	-0.74
G6	0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36	-0.49	-0.58	…	-1.47
G7	0.20	0.14	0.00	0.11	-0.34	-0.03	0.04	-0.76	-0.81	-1.12	…	-1.36
G8	0.40	0.43	0.18	0.00	-0.14	0.29	0.07	-0.79	-0.81	-0.92	…	-1.22
G9	0.01	0.46	0.28	-0.34	-0.23	-0.36	-0.45	-0.64	-0.79	-1.22	…	-1.09
…	…	…	…	…	…	…	…	…	…	…	…	…
GN	-0.23	0.04	0.00	-0.30	-0.29	-0.45	-0.97	-2.06	-0.89	-1.22	…	-0.97

Cell growth

Transcripton

WT

Transgenic

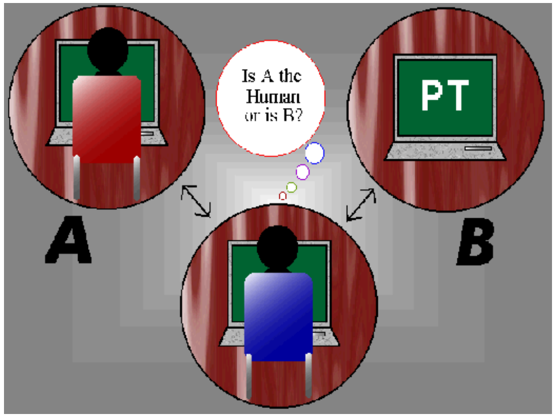
N > 10000

Artificial intelligence: The Turing test

1912-1954

- Turing proposed that a computer program show intelligent behavior if is able to fool a human interrogator

- The Turing test: the computer is interrogated by a human via a teletype, and passes the test if the interrogator cannot tell if there is a computer or a human at the other end
  - natural language processing
  - knowledge representation
  - automated reasoning
  - machine learning



AI techniques

- Logics
- Knowledge representation
- Search
- Machine learning
- Pattern recognition
- Automatic theorem proving
- Planning
- Machine vision
- Natural language processing

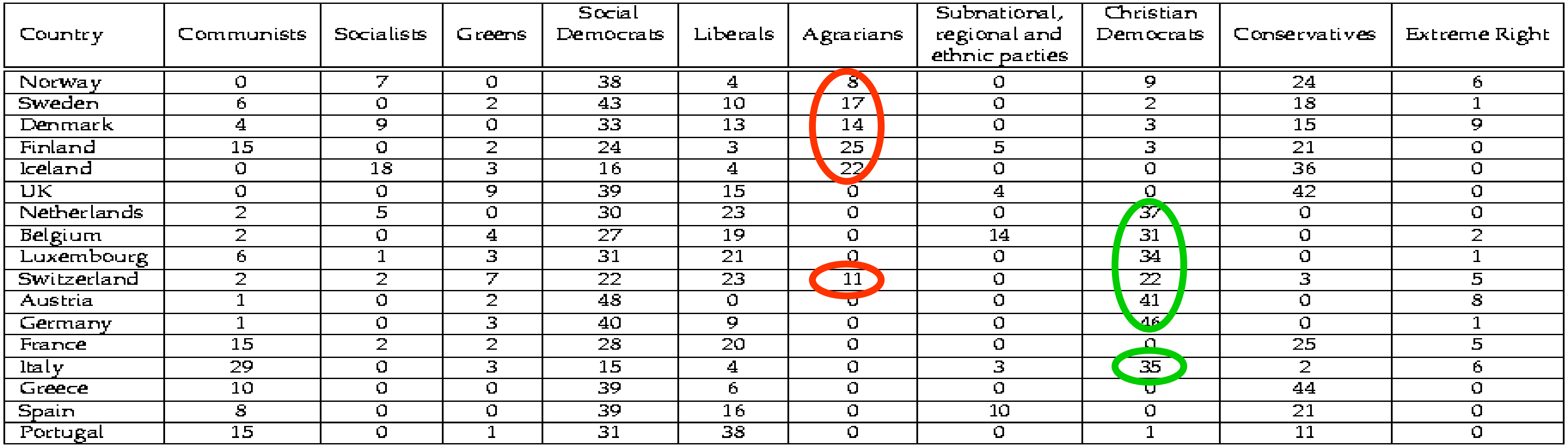
"…making a machine behave in ways that would be called intelligent if a human were so behaving"

- John McCarthy, August 31, 1955

"The subfield of computer science concerned with the concepts and methods of symbolic inference by computer and symbolic knowledge representation for use in making inferences."

- The Free On-line Dictionary of Computing (September 27, 2003)

Example: Decision tree learning



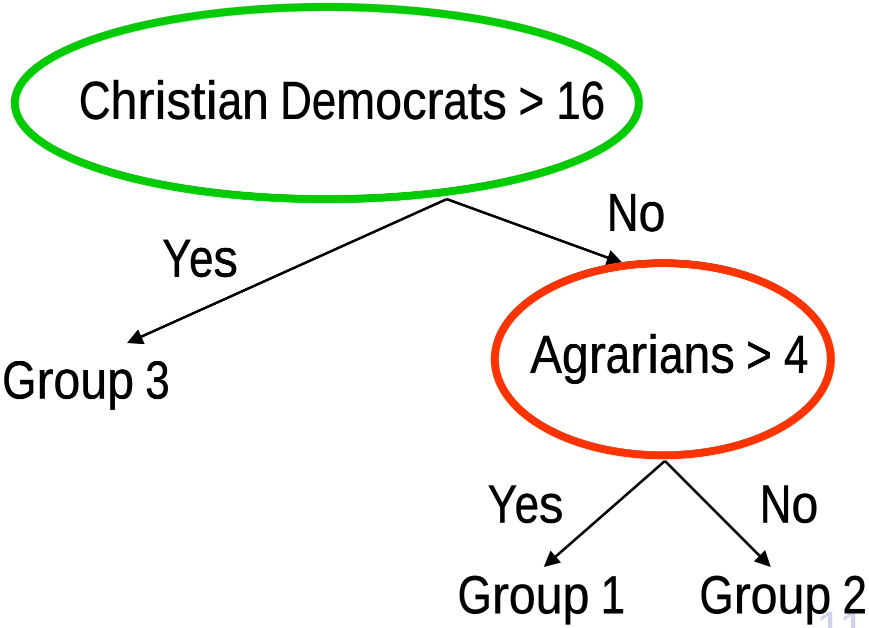
Country	Communists	Socialists	Greens	Social Democrats	Liberals	Agrarians	Subnational, regional and ethnic parties	Christian Democrats	Conservatives	Extreme Right
Norway	0	7	0	38	4	8	0	9	24	6
Sweden	6	0	2	43	10	17	0	2	18	1
Denmark	4	9	0	33	13	14	0	3	15	9
Finland	15	0	2	24	3	25	5	3	21	0
Iceland	0	18	3	16	4	22	0	0	36	0
UK	0	0	9	39	15	0	4	0	42	0
Netherlands	2	5	0	30	23	0	0	37	0	0
Belgium	2	0	4	27	19	0	14	31	0	2
Luxembourg	6	1	3	31	21	0	0	34	0	1
Switzerland	2	2	7	22	23	11	0	22	3	5
Austria	1	0	2	48	0	0	0	41	0	8
Germany	1	0	3	40	9	0	0	46	0	1
France	15	2	2	28	20	0	0	0	25	5
Italy	29	0	3	15	4	0	3	35	2	6
Greece	10	0	0	39	6	0	0	0	44	0
Spain	8	0	0	39	16	0	10	0	21	0
Portugal	15	0	1	31	38	0	0	1	11	0

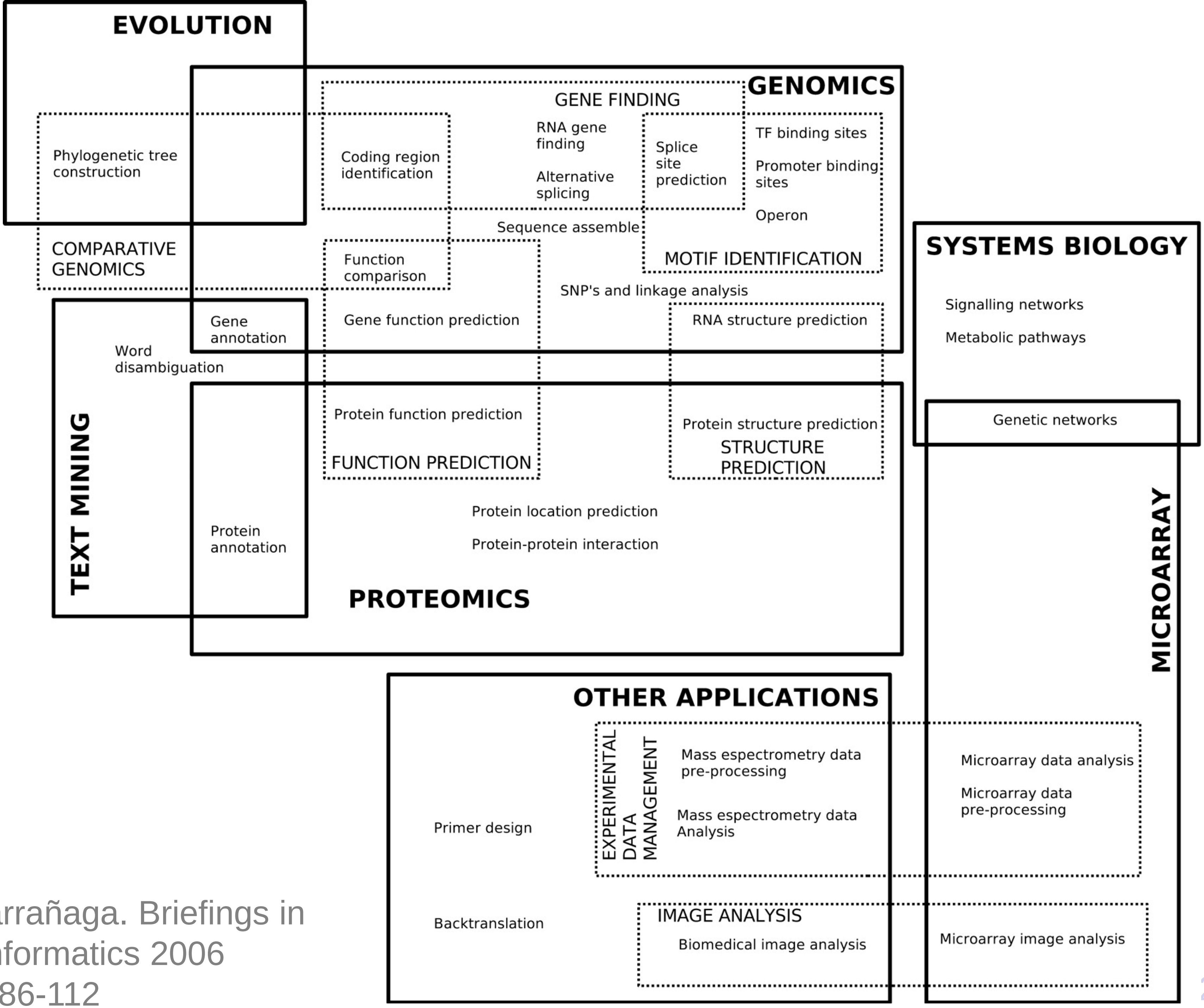
Class knowledge:

Group 1: Nordic countries

Group 2: UK, France, Greece, Spain, Portugal

Group 3: Benelux countries, Switzerland, Austria, Italy, Germany





EVOLUTION

GENOMICS

GENE FINDING

Phylogenetic tree construction

COMPARATIVE GENOMICS

Coding region identification

RNA gene finding

Alternative splicing

Splice site prediction

TF binding sites

Promoter binding sites

Operon

MOTIF IDENTIFICATION

Sequence assemble

Function comparison

SNP's and linkage analysis

Gene function prediction

RNA structure prediction

SYSTEMS BIOLOGY

Signalling networks

Metabolic pathways

Genetic networks

TEXT MINING

Gene annotation

Word disambiguation

Protein annotation

Protein function prediction

FUNCTION PREDICTION

Protein structure prediction

STRUCTURE PREDICTION

Protein location prediction

Protein-protein interaction

PROTEOMICS

MICROARRAY

OTHER APPLICATIONS

EXPERIMENTAL DATA MANAGEMENT

Mass espectrometry data pre-processing

Mass espectrometry data Analysis

Microarray data analysis

Microarray data pre-processing

Primer design

Backtranslation

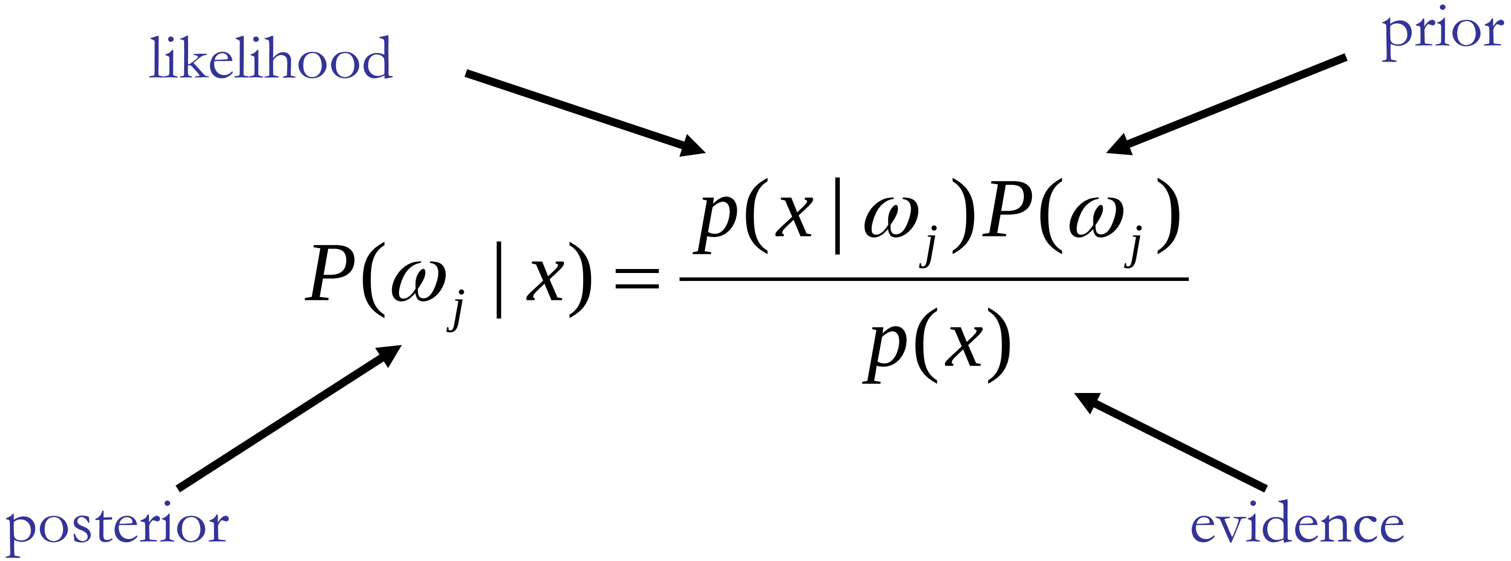
IMAGE ANALYSIS

Biomedical image analysis

Microarray image analysis

P. Larrañaga. Briefings in Bioinformatics 2006 7(1):86-112

Bayes decision rule



likelihood

prior

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j)P(\omega_j)}{p(x)}$$

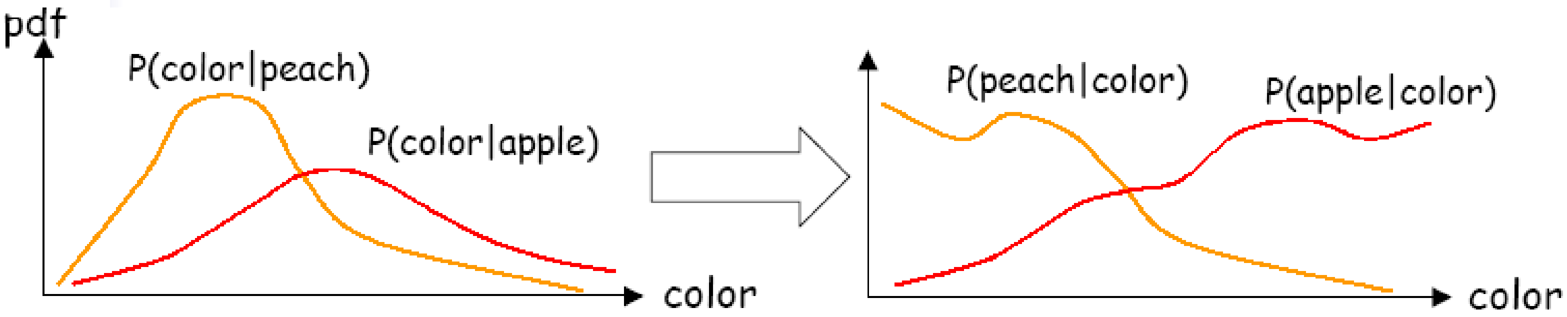
posterior

evidence

Bayes decision rule:

If P(w_1|x) > P(w_2|x) then choose w_1, else choose w_2.

Example

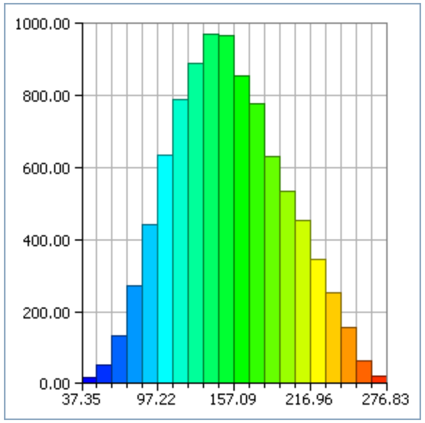


- Bayes Decision Rule
  - If P(apple | color) > P(peach | color) then choose apple

- Note that the evidence p(color) is only necessary for normalization purposes; it does not affect the decision rule

So, what about the data?

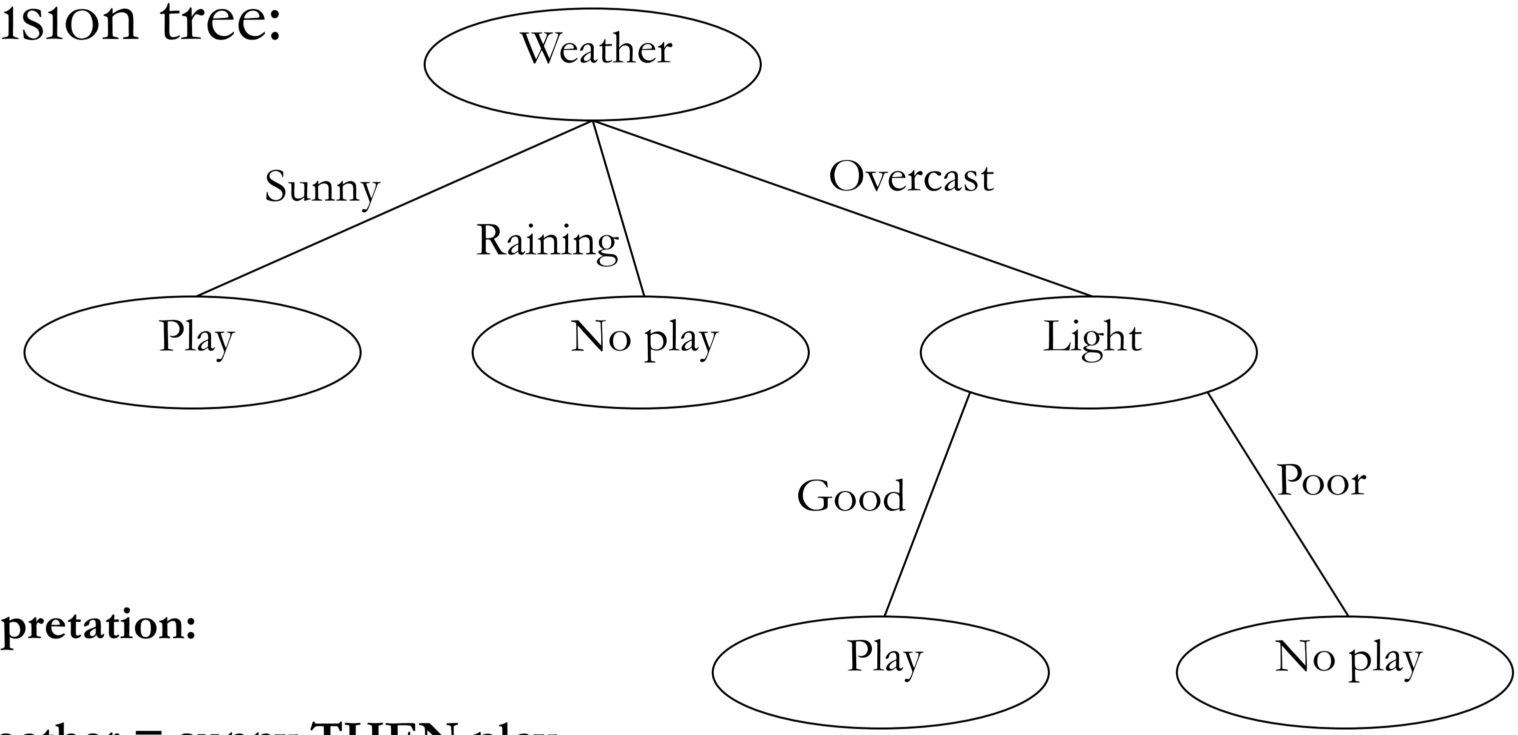
- Use examples to estimate the probability distributions:
  - P(w_j) is easy.
  - p(x|w_j): Histogram!



- One feature: bins are rectangles, Two features: cubes, *n*-features: hyper-cubes.
- More dimensions/features require more training data: Curse of dimensionality!
  - If we need 10 observations when we have one feature (to get a good histogram), then we need 10^n observations when we have *n*-features!
- If the true probability distributions are known, then Bayes decision rule is optimal (minimizes error rate)

Decision trees / Rule learning

Final decision tree:



Interpretation:

IF weather = sunny THEN play
IF weather = raining THEN no play
IF weather = overcast AND light = good THEN play
IF weather = overcast AND light = poor THEN no play

Overfitting

- Overfitting: The method learns the random patterns in the data as well as the underlying process that created the data
  - Occurs because the alg. tries to reduce the classification error

- To identify this phenomenon:
  - Split data into training data (≈75%) and test data (≈25%)
  - Build tree on the training data and test the model on the test data

- A decision tree X is overfitted if there exists a tree Y that do better on an unseen test set, but worse on the training set

- “Solution”: Prune complex branches of the tree

Occam's razor

- William of Occam 14th century: *things should not be multiplied unnecessarily*

- Issac Newton (1687): *we are to admit no more causes of natural things than such as are both true and sufficient to explain their appearance*

- Albert Einstein (20th century): *everything should be made as simple as possible, but not simpler*

The simplest model that explains the data should be chosen

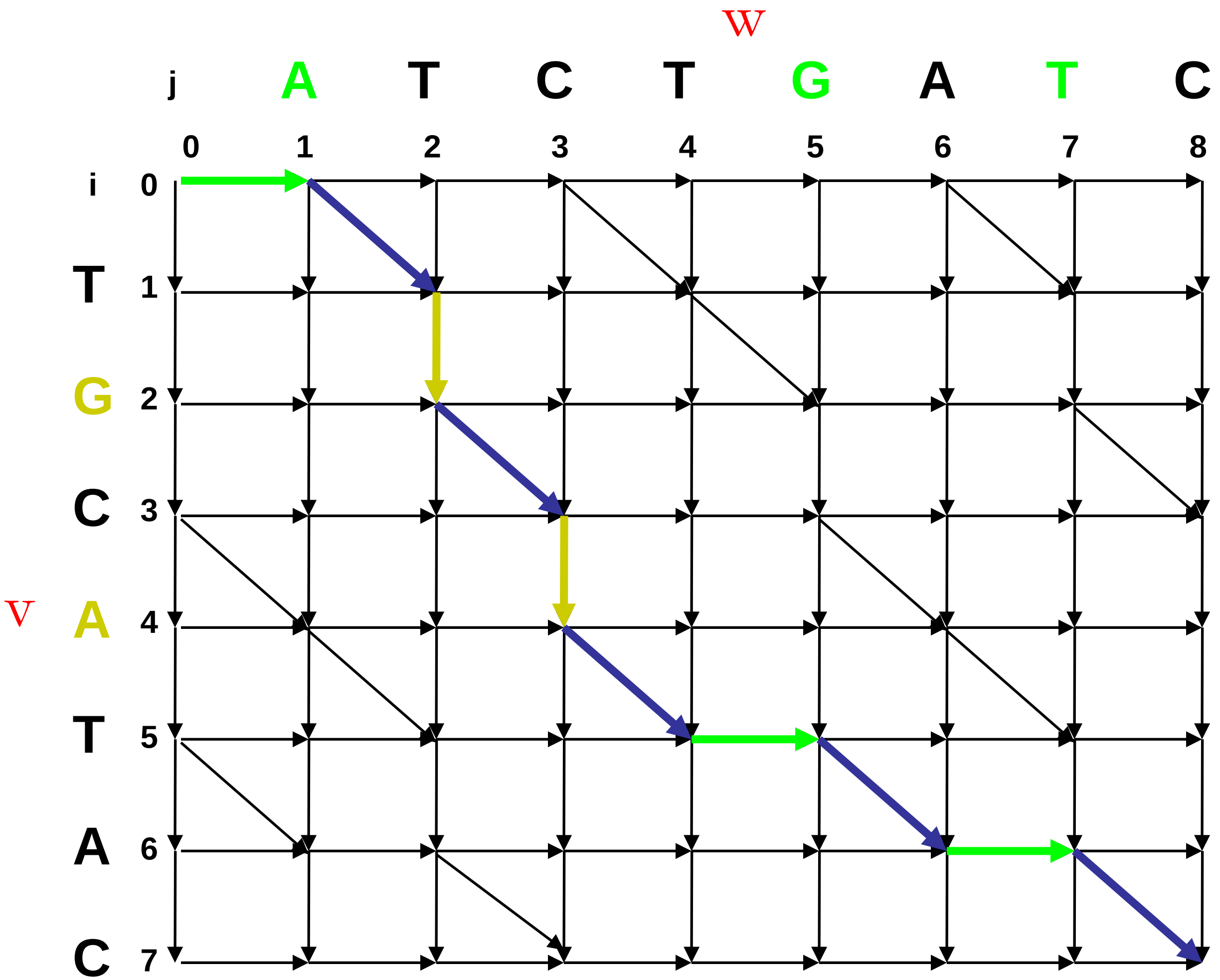
Decision trees: greedy algorithm

- Decision trees are built by iteratively splitting the training examples using the “best” feature: greedy

- Would benefit from some search strategy
  - A split could be evaluated in terms of its current ability to classify the data AND the accuracy of the splits later on in the algorithm run

- All problems are search problems!

Sequence alignment as a search problem



Deletion

Matches

Insertion

```
-TGCAT-A-C
AT-C-TGATC
```

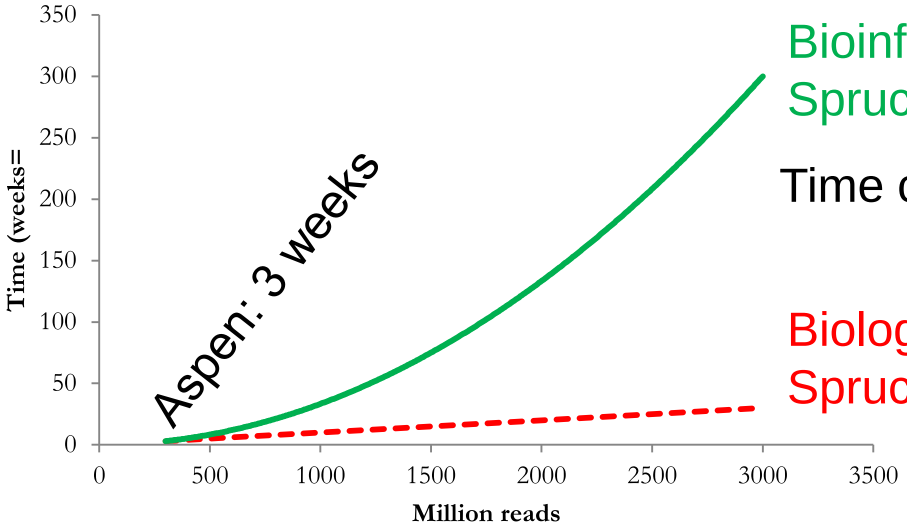
Algorithm design

- Exhaustive algorithms (brute force): examine every possible alterative to find the solution
- Greedy algorithms: find the solution by always choosing the currently ”best” alternative
- Randomized algorithms: finds the solution based on randomized choices

Time complexity

- Genome assembly: pice together a genome from short reads (~200bp)
  - Aspen: 300M reads
  - Spruce: 3000M reads

- Pair-wise all-against all alignment for Aspen takes 3 weeks on 16 processors
- What about spruce?



Aspen: 3 weeks

Bioinformatician:
Spruce: 300 uker

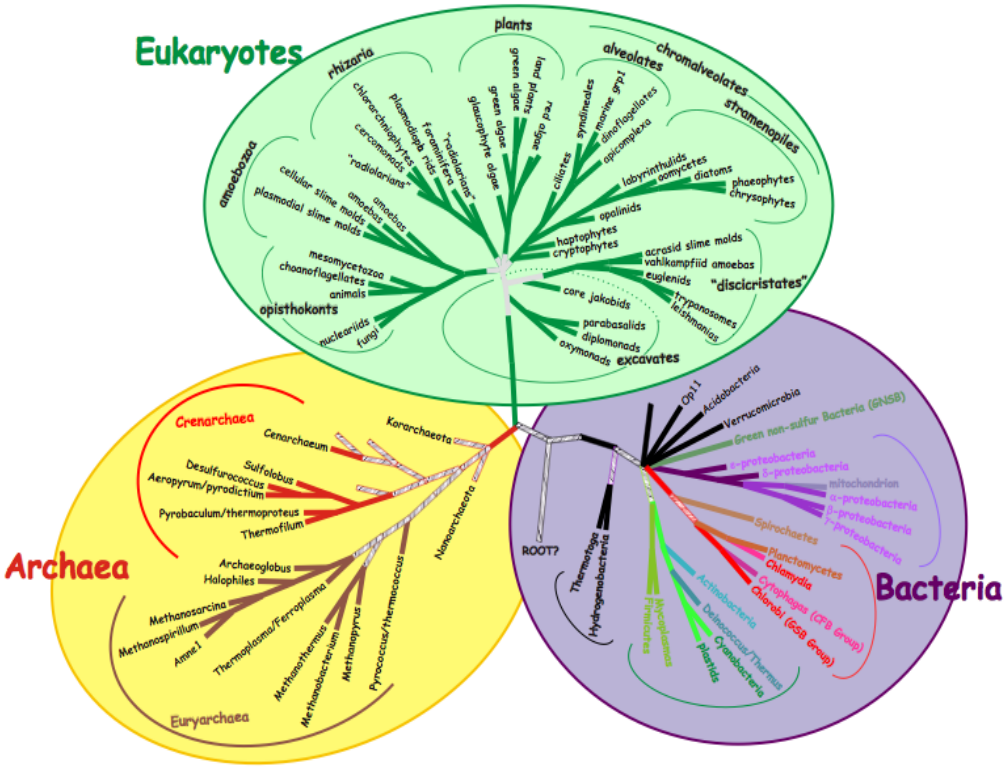
Time complexity: O(n^2)

Biologist:
Spruce: 30 weeks

Tractable versus intractable problems

- Some problems requires polynomial time
  - e.g. sorting a list of integers
  - called tractable problems
- Some problems require exponential time
  - e.g. listing every subset in a list
  - called intractable problems
- Some problems lie in between
  - e.g. the traveling salesman problem
  - called NP-complete problems
  - nobody have proved whether a polynomial time algorithm exists for these problems

Phylogenetic trees/Decision trees



- Number of trees with n leaves: n^{n-2}
  - n=10: 10^{8}
  - n=30: 10^{41}
  - n=50: 10^{81}

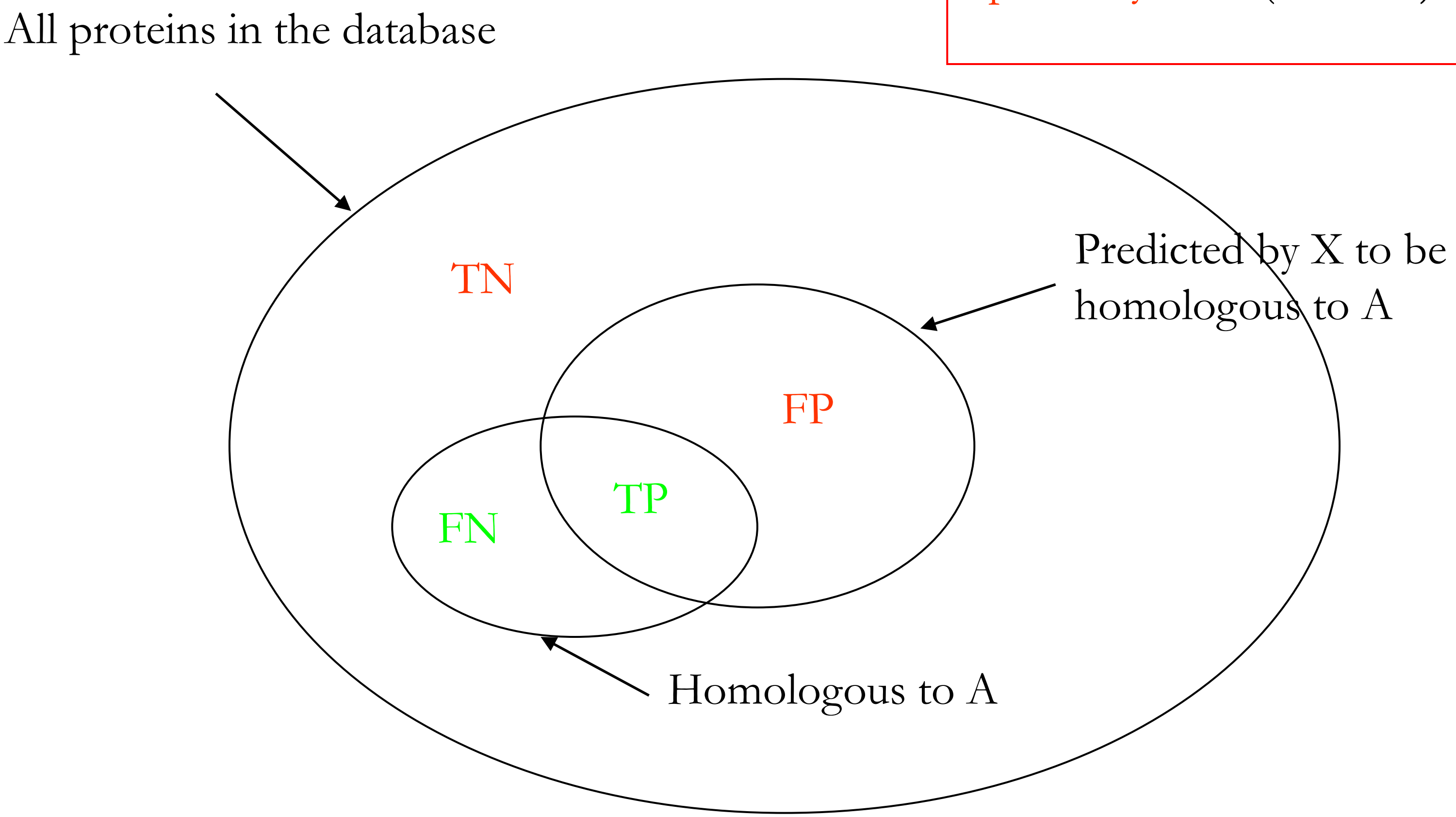
- There are 10^{80} particles in the universe!

Method power

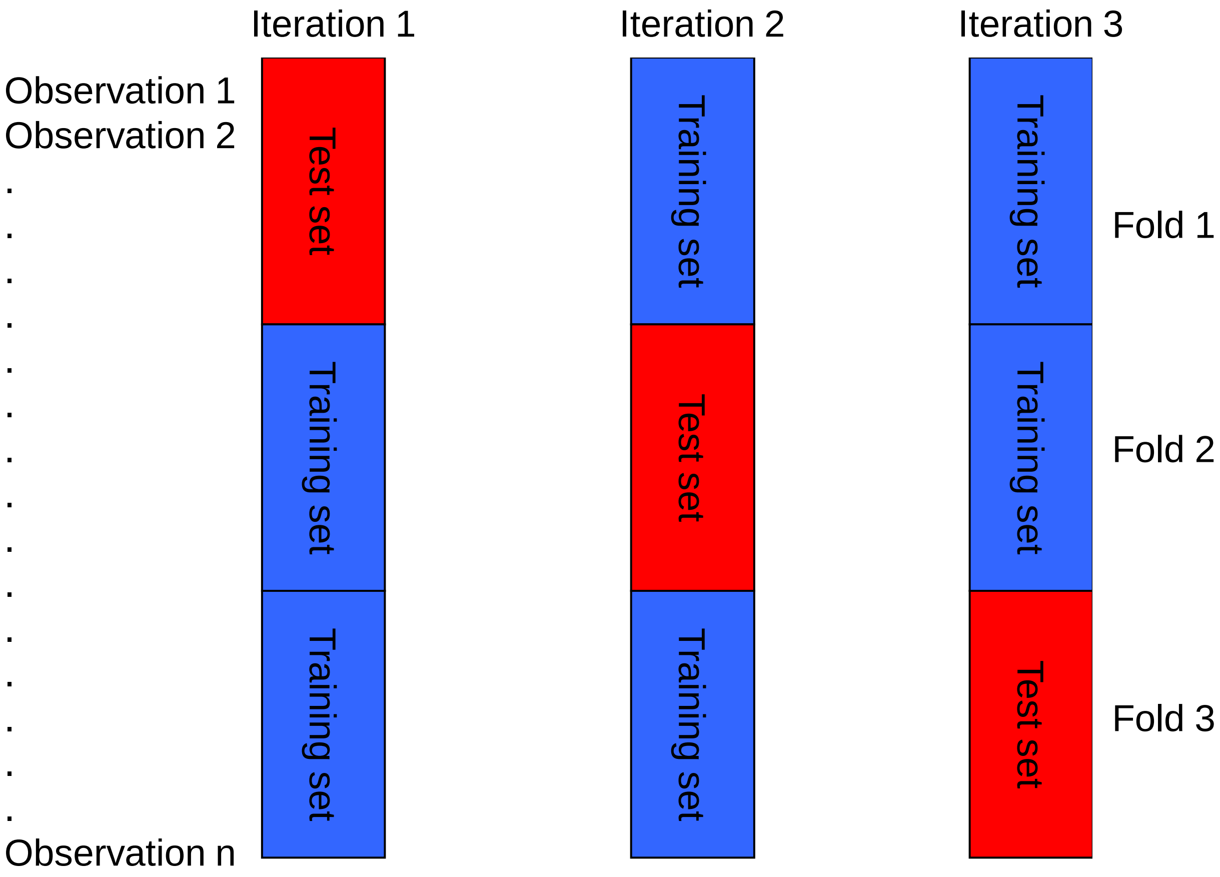
You want to find homologous proteins to a specific protein A using some computational method X:

Sensitivity: TP/(TP+FN)

Specificity: TN/(TN+FP)


All proteins in the database
TN
Predicted by X to be homologous to A
FP
TP
FN
Homologous to A

Cross validation



- *k*-fold cross validation: *k* iterations
- Leave-one out cross validation: *n* iterations

Evaluation

- Classifications can be
  - True positives (TP)
  - False negatives (FN)
  - True positives (TP)
  - False positives (FP)
- Evaluation measures:
  - accuracy = (TP+TN)/(TP+FN+TN+FP)
  - sensitivity = TP/(TP+FN)
  - specificity = TN/(TN+FP)
- Confusion matrix:

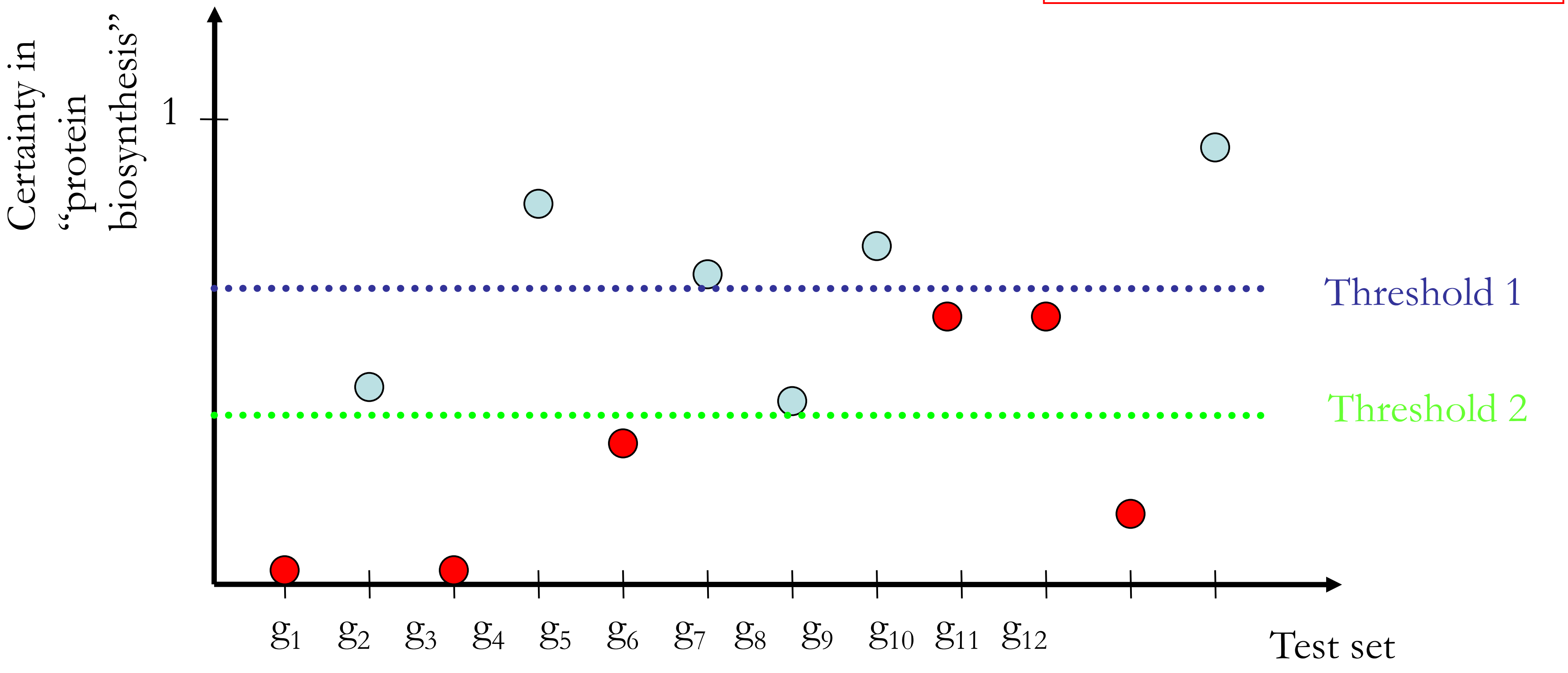
		Predicted	
		Class 0	Class 1
Actual	Class 0	TN	FP
	Class 1	FN	TP

Threshold selection

○ Gene with function “protein biosynthesis”

● Gene with a different function

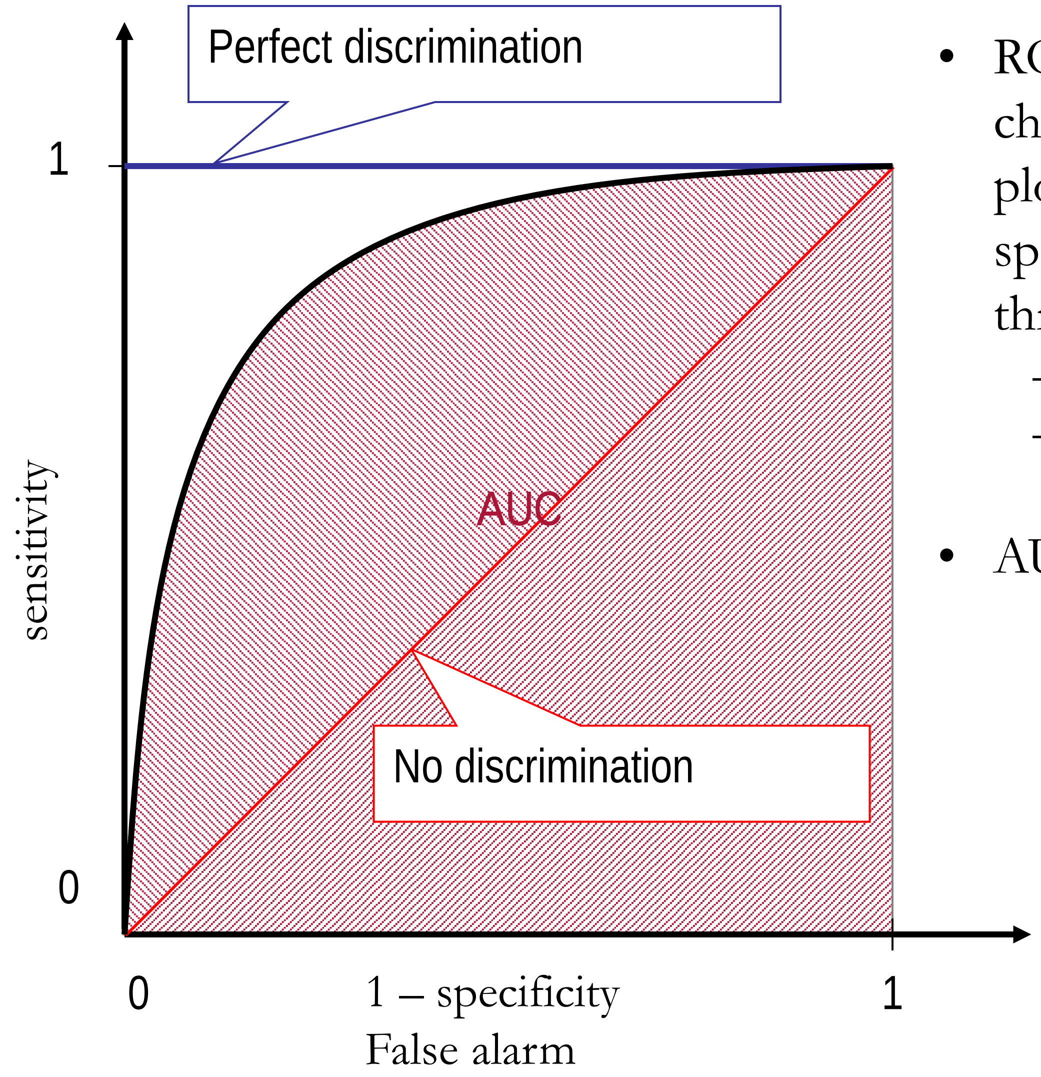
sensitivity:
TP/(TP+FN)
specificity:
TN/(TN+FP)



Sensitivity = 2/3, Specificity=1

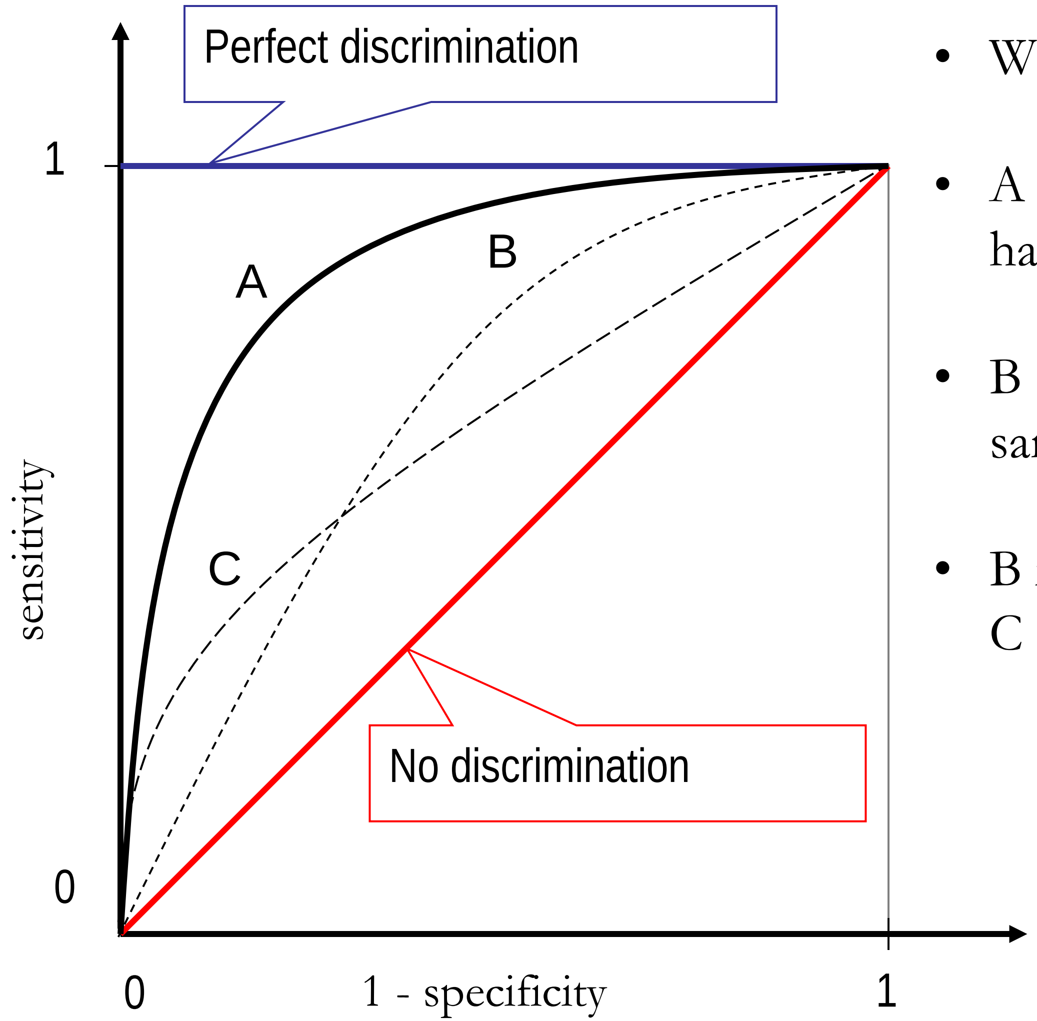
Sensitivity = 1, Specificity=2/3

ROC analysis and classifier evaluation



- ROC: Receiver operating characteristics curve results from plotting sensitivity against specificity for all possible thresholds
  - sensitivity: TP/(TP+FN)
  - specificity: TN/(TN+FP)
- AUC: Area under the ROC curve

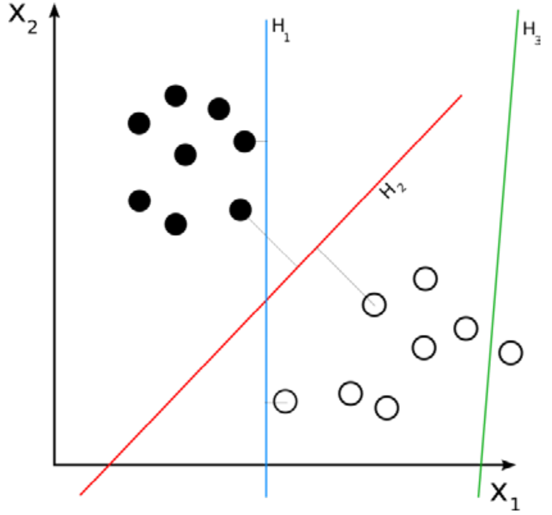
ROC analysis and classifier evaluation



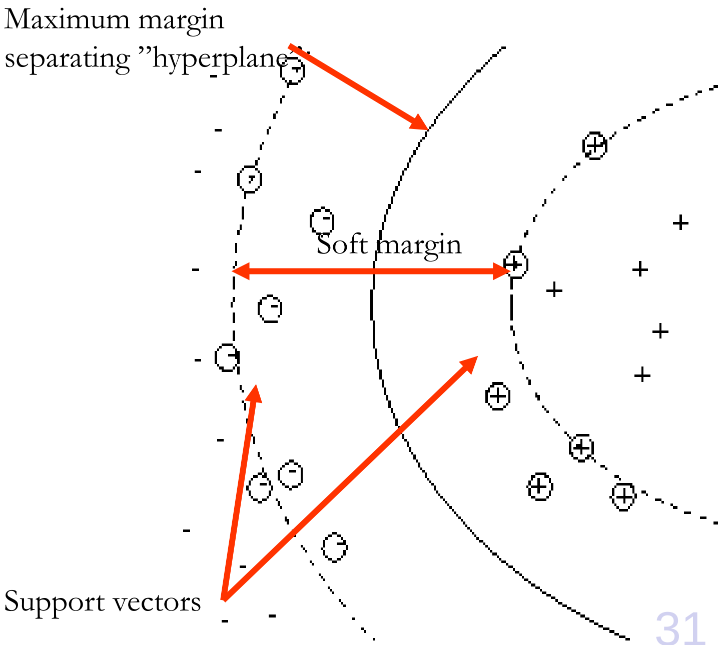
- Which ROC curve is better?
- A dominants B and C and clearly has a higher AUC
- B and C have approximately the same AUC
- B is better for some thresholds, C for others

Linear versus non-linear classifiers

- Linear: Finds a hyperplane that separates the classes
  - In two dimensions: $w_0 + w_1 \cdot x_1 + w_2 \cdot x_2$
  - Use the examples $\boldsymbol{x}$ to estimate $\boldsymbol{w}$



- Non-linear:
  - Support vector machines uses the kernel trick: The kernel maps the observations into a higher dimensional space where the problem is linearly separable
  - Artifical neural networks

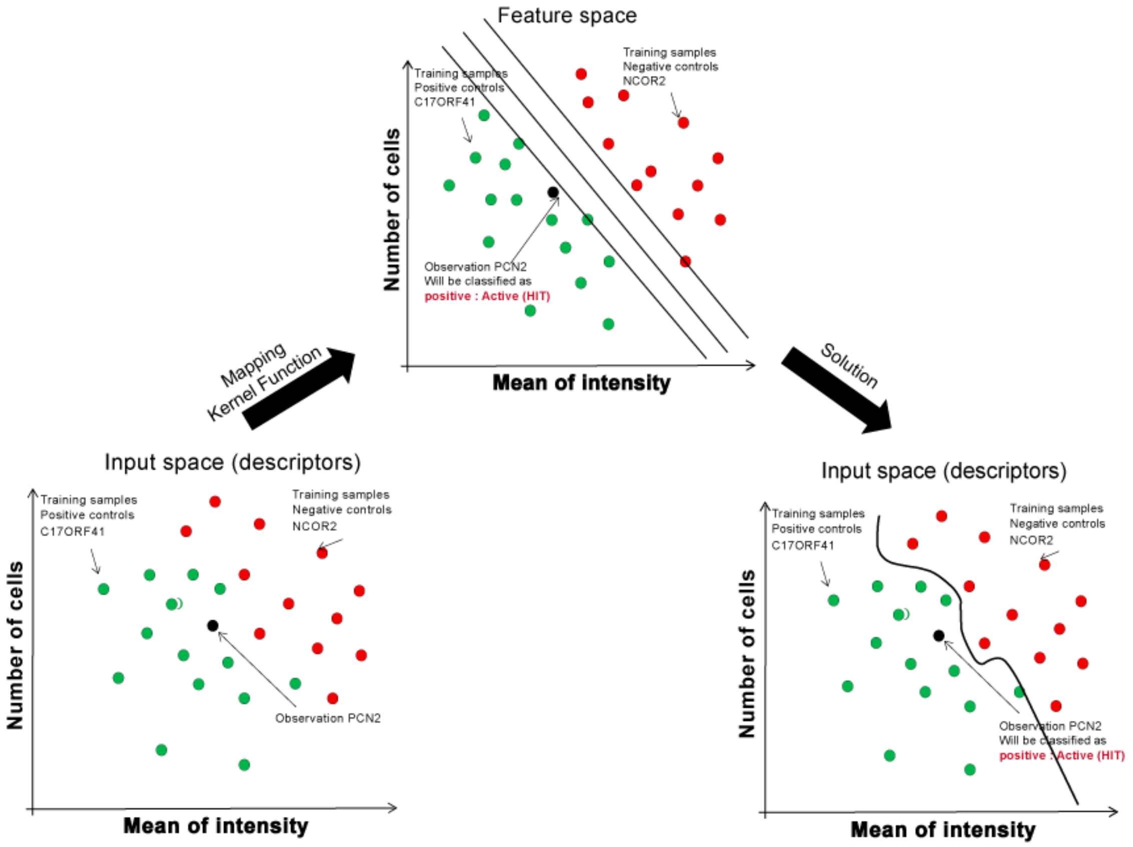

Maximum margin separating ”hyperplane”

Soft margin

Support vectors

The kernel trick

siRNA classification



Artificial neural networks

- Inspired by how the brain works – a mathematical model for the operation of the brain
- Learning in an ANN is reduced to the process of using the training data to tune the weights so that the network represents the desired function

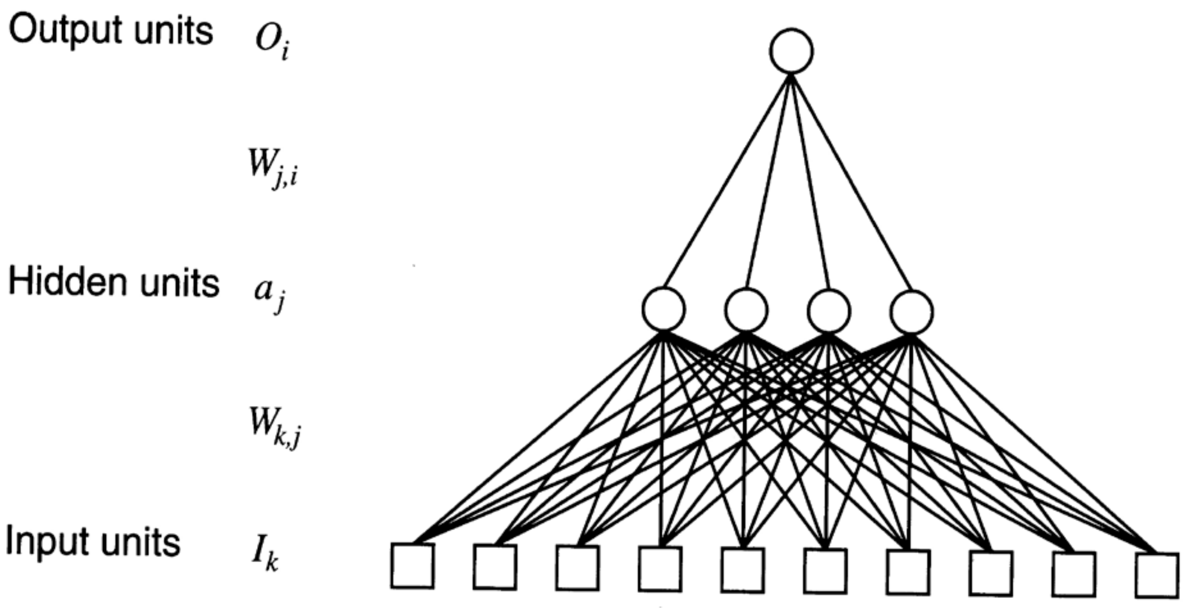
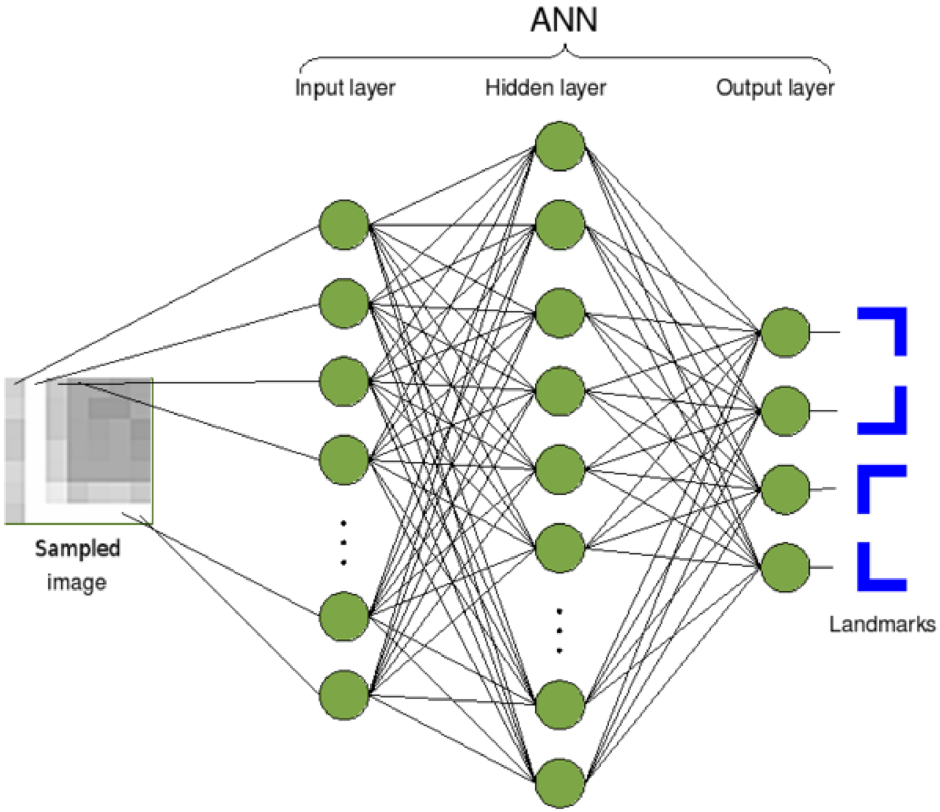


Image recognition



Clustering analysis

Need to define;

- measure of similarity
- algorithm for using the measure of similarity to discover natural groups in the data

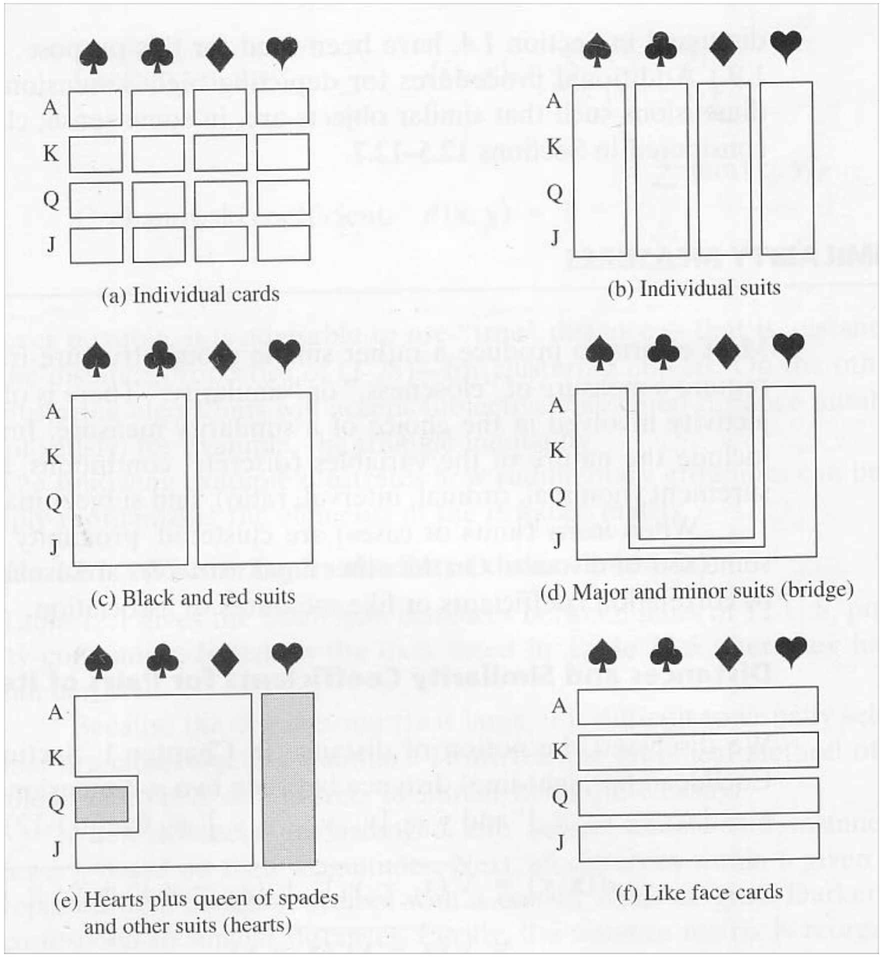
The number of ways to divide n items into k clusters:

$$k^n / k!$$

Example: $10^{500}/10! = 2.756 \times 10^{493}$

Measure of similarity

What is similar?



(a) Individual cards

(b) Individual suits

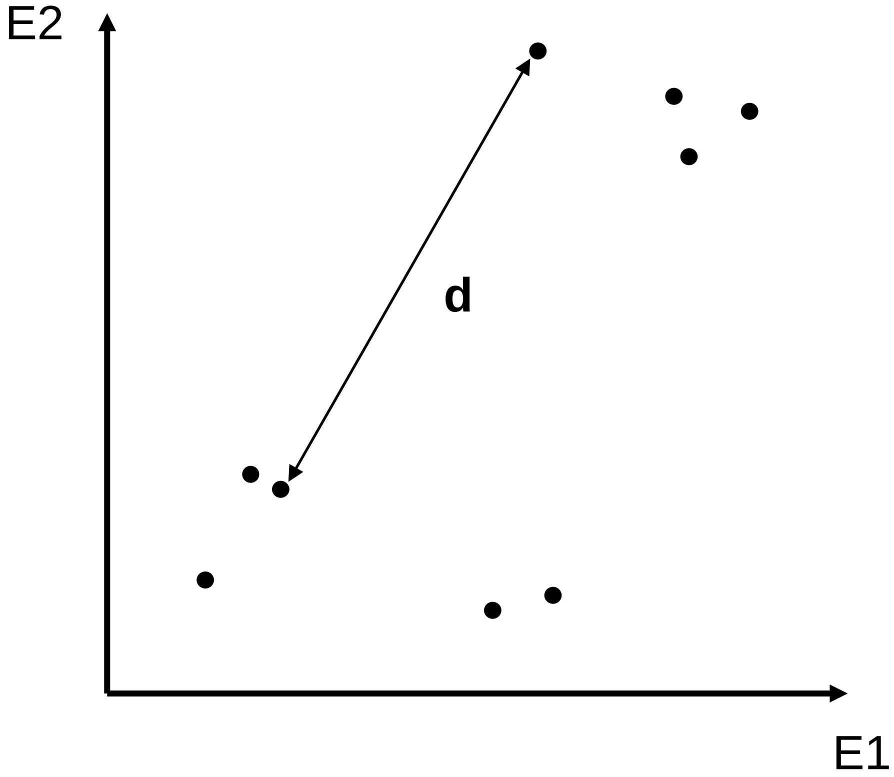
(c) Black and red suits

(d) Major and minor suits (bridge)

(e) Hearts plus queen of spades and other suits (hearts)

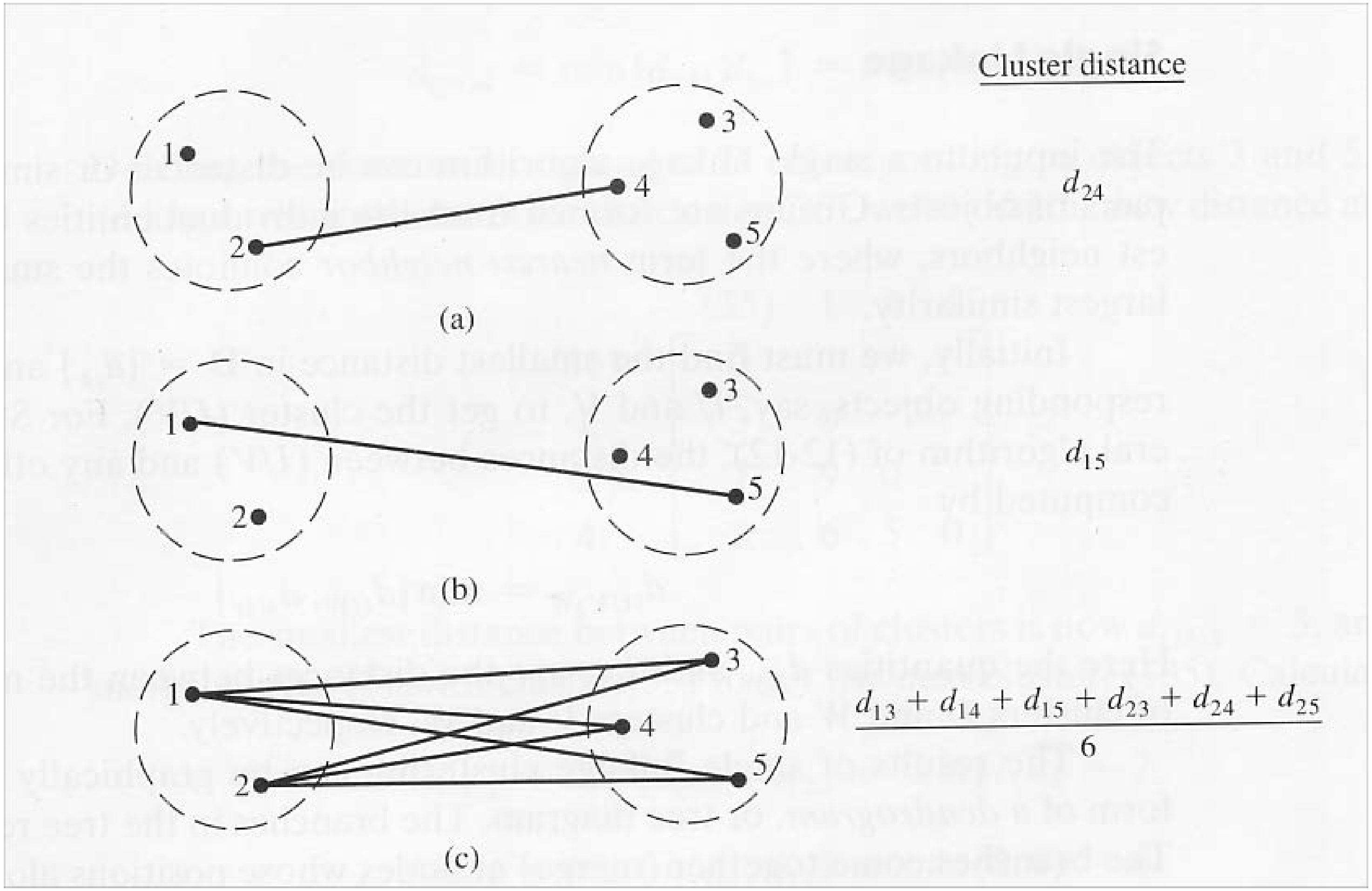
(f) Like face cards

Euclidean distance



Hierarchical clustering

Inter-cluster similarity measures: (a) single linkage, (b) complete linkage and (c) average linkage


Cluster distance
1
2
3
4
5
(a)
d_{24}
1
2
3
4
5
(b)
d_{15}
1
2
3
4
5
(c)
$\frac{d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25}}{6}$

Example of hierarchical clustering: languages of Europe

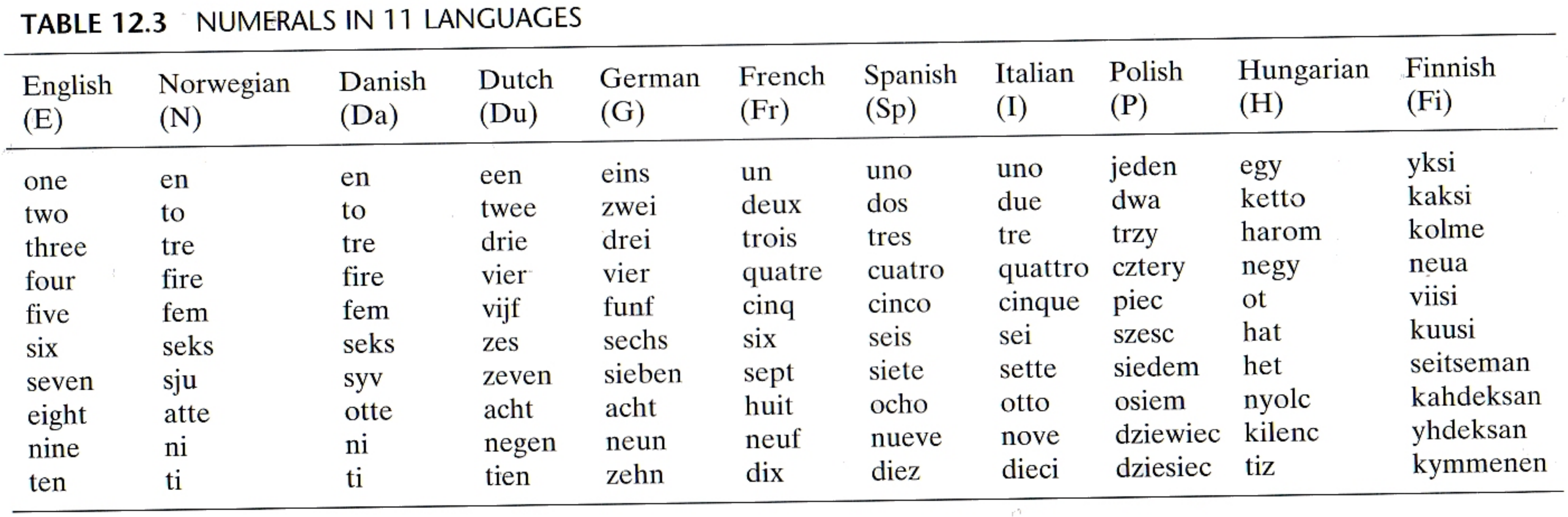


TABLE 12.3 NUMERALS IN 11 LANGUAGES

English (E)	Norwegian (N)	Danish (Da)	Dutch (Du)	German (G)	French (Fr)	Spanish (Sp)	Italian (I)	Polish (P)	Hungarian (H)	Finnish (Fi)
one	en	en	een	eins	un	uno	uno	jeden	egy	yksi
two	to	to	twee	zwei	deux	dos	due	dwa	ketto	kaksi
three	tre	tre	drie	drei	trois	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	neua
five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot	viisi
six	seks	seks	zes	sechs	six	seis	sei	szesc	hat	kuusi
seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyolc	kahdeksan
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen

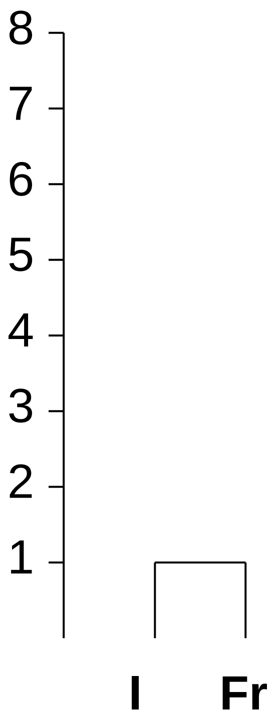
Distance: Frequency of numbers with different first letter e.g.

$d_{EN} = 2 \quad d_{EDu} = 7 \quad d_{SpI} = 1$

Inter-cluster strategy: SINGEL LINKAGE

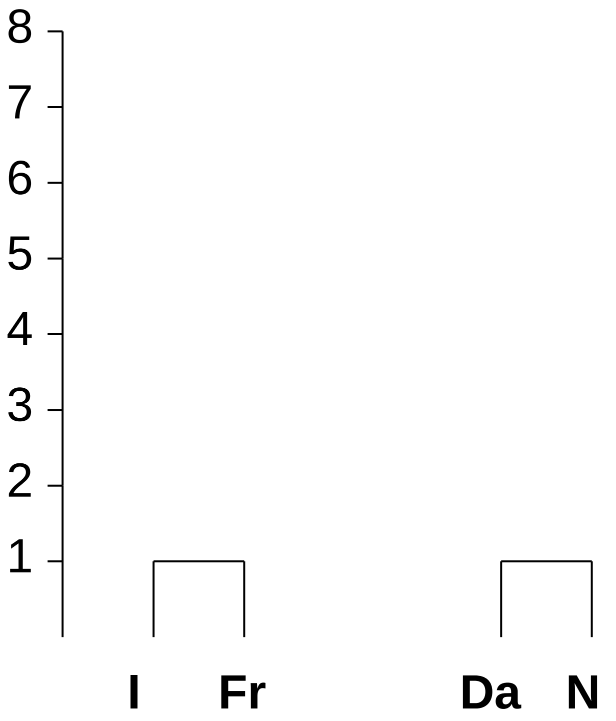
Iteration 1

	E	N	Da	Du	G	Fr	Sp	I	P	H	Fi
E	0										
N	2	0									
Da	2	1	0								
Du	7	5	6	0							
G	6	4	5	5	0						
Fr	6	6	6	9	7	0					
Sp	6	6	5	9	7	2	0				
I	6	6	5	9	7	1	1	0			
P	7	7	6	10	8	5	3	4	0		
H	9	8	8	8	9	10	10	10	10	0	
Fi	9	9	9	9	9	9	9	9	9	8	0



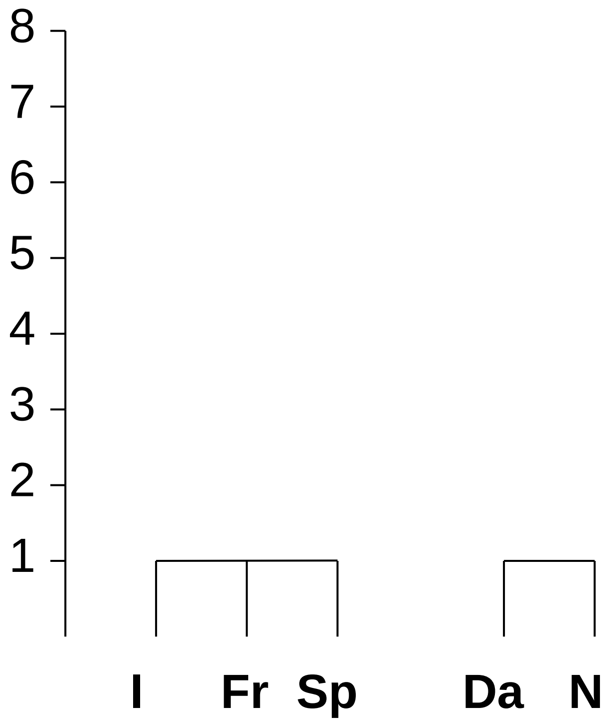
Iteration 2

	I Fr	E	N	Da	Du	G	Sp	P	H	Fi
I Fr	0									
E	6	0								
N	6	2	0							
Da	5	2	1	0						
Du	9	7	5	6	0					
G	7	6	4	5	5	0				
Sp	1	6	6	5	9	7	0			
P	4	7	7	6	10	8	3	0		
H	10	9	8	8	8	9	10	10	0	
Fi	9	9	9	9	9	9	9	9	8	0



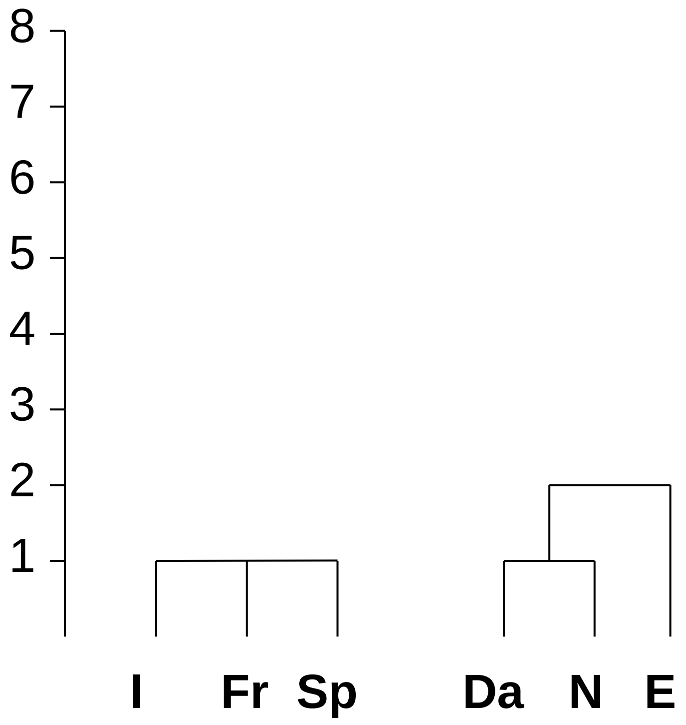
Iteration 3

	Da N	I Fr	E	Du	G	Sp	P	H	Fi
Da N	0								
I Fr	5	0							
E	2	6	0						
Du	5	9	7	0					
G	4	7	6	5	0				
Sp	5	1	6	9	7	0			
P	6	4	7	10	8	3	0		
H	8	10	9	8	9	10	10	0	
Fi	9	9	9	9	9	9	9	8	0



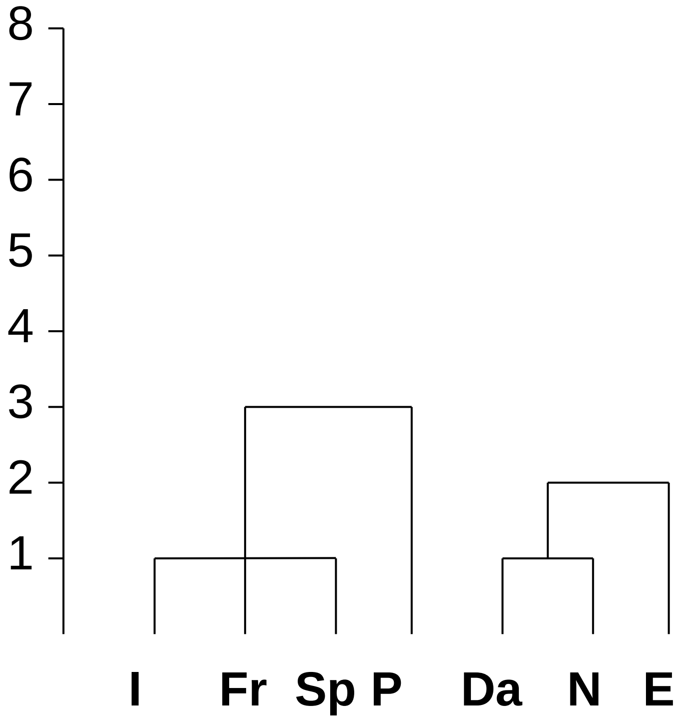
Iteration 4

	Sp I Fr	Da N	E	Du	G	P	H	Fi
Sp I Fr	0							
Da N	5	0						
E	6	2	0					
Du	9	5	7	0				
G	7	4	6	5	0			
P	3	6	7	10	8	0		
H	10	8	9	8	9	10	0	
Fi	9	9	9	9	9	9	8	0



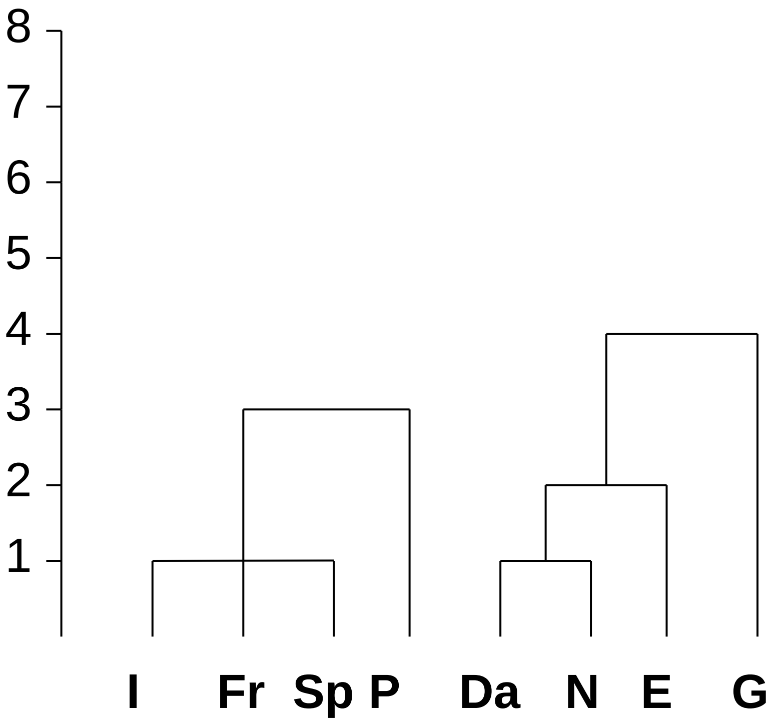
Iteration 5

	E Da N	Sp I Fr	Du	G	P	H	Fi
E Da N	0						
Sp I Fr	5	0					
Du	5	9	0				
G	4	7	5	0			
P	6	3	10	8	0		
H	8	10	8	9	10	0	
Fi	9	9	9	9	9	8	0



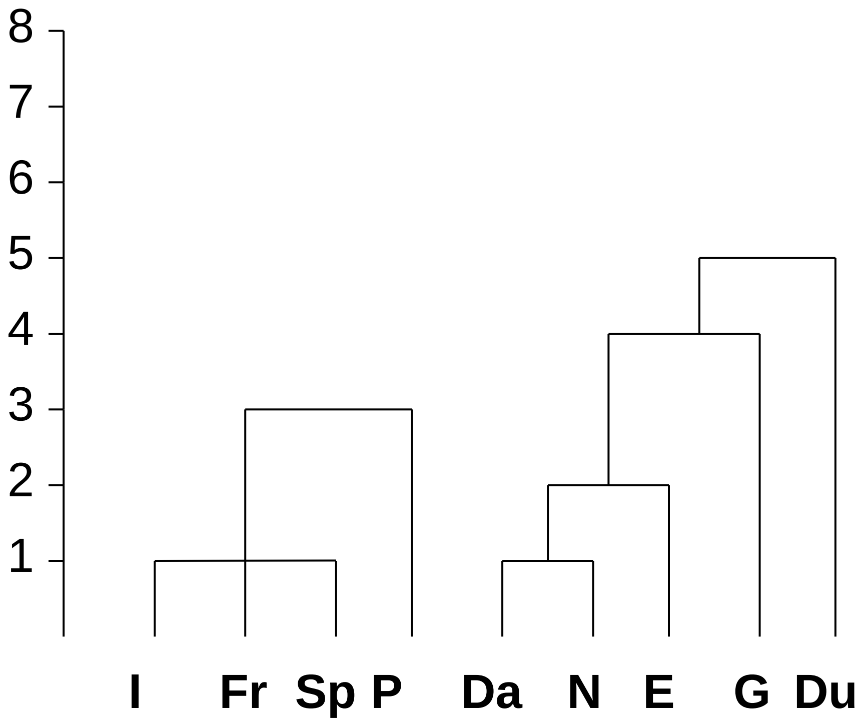
Iteration 6

	P Sp I Fr	E Da N	Du	G	H	Fi
P Sp I Fr	0					
E Da N	5	0				
Du	9	5	0			
G	7	4	5	0		
H	10	8	8	9	0	
Fi	9	9	9	9	8	0



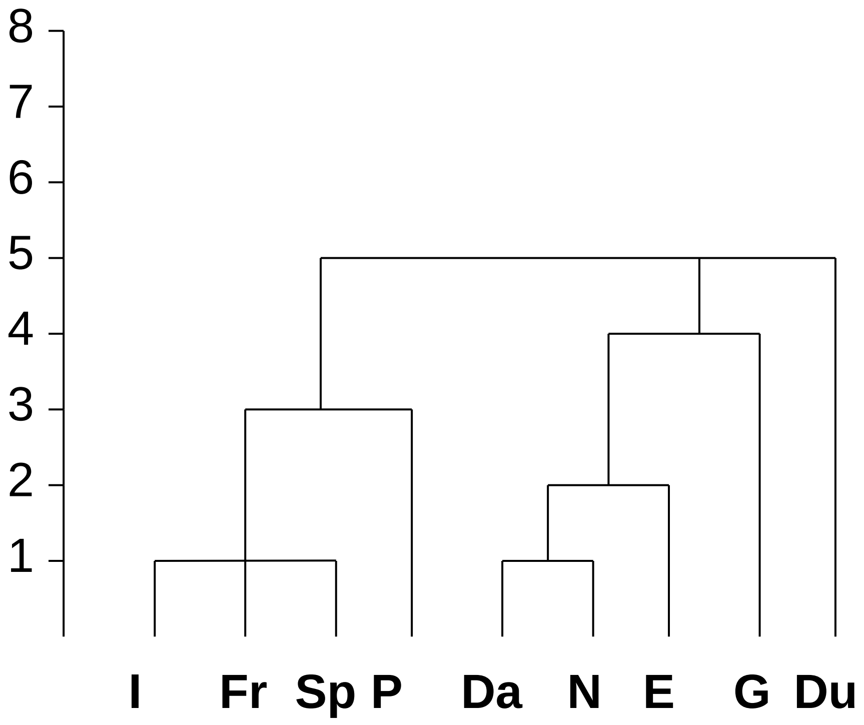
Iteration 7

	G E Da N	P Sp I Fr	Du	H	Fi
G E Da N	0				
P Sp I Fr	5	0			
Du	5	9	0		
H	8	10	8	0	
Fi	9	9	9	8	0



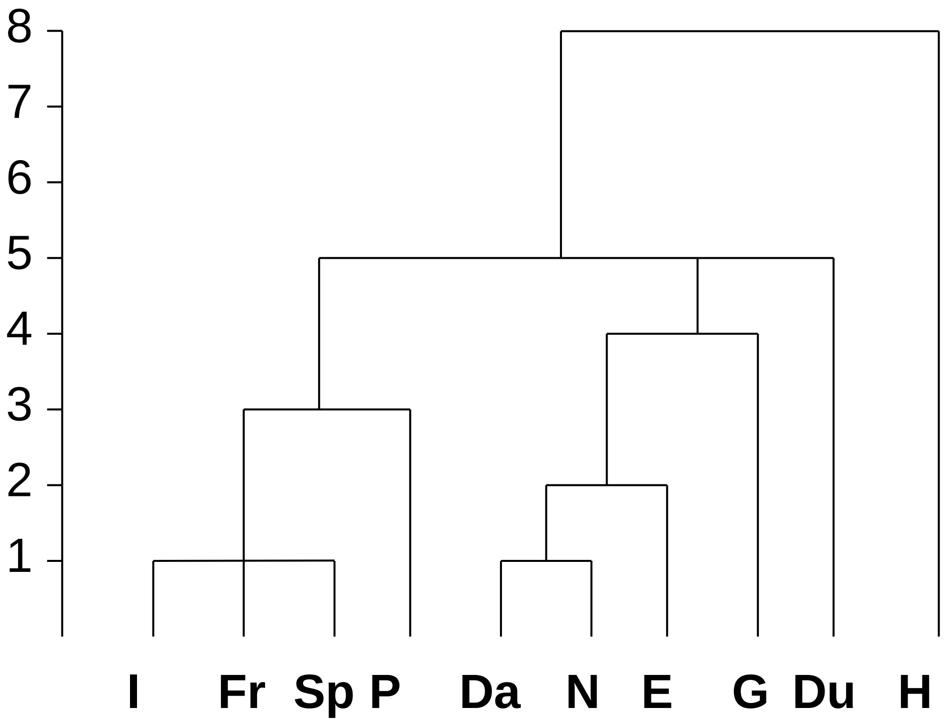
Iteration 8

	Du G E Da N	P Sp I Fr	H	Fi
Du G E Da N	0			
P Sp I Fr	5	0		
H	8	10	0	
Fi	9	9	8	0



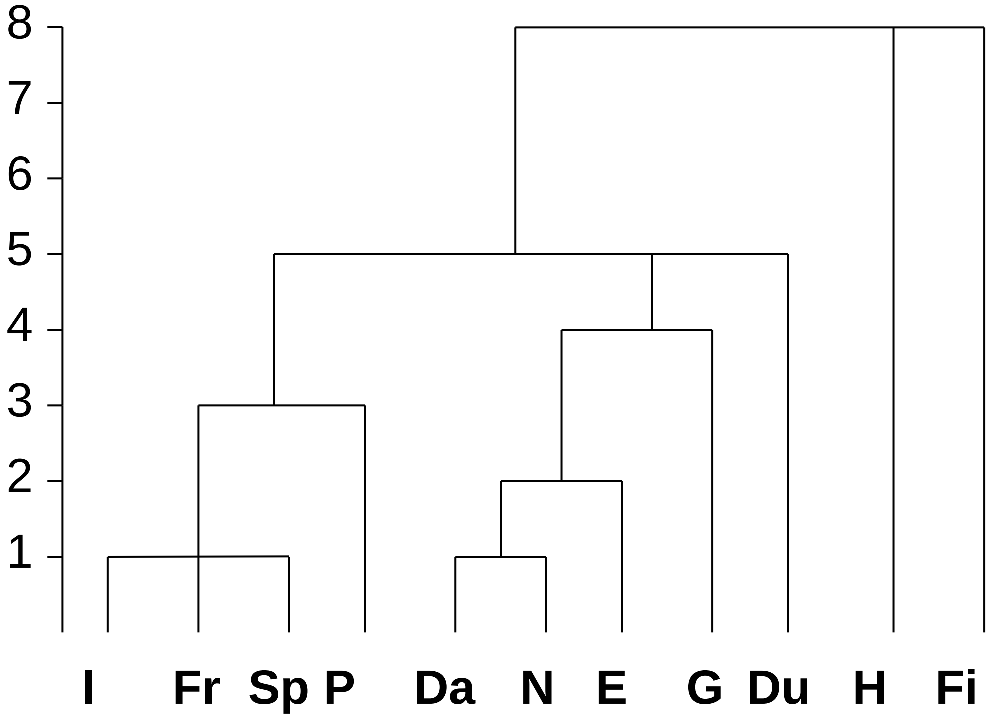
Iteration 9

	P Sp I Fr Du G E Da N	H	Fi
P Sp I Fr Du G E Da N	0		
H	8	0	
Fi	9	8	0

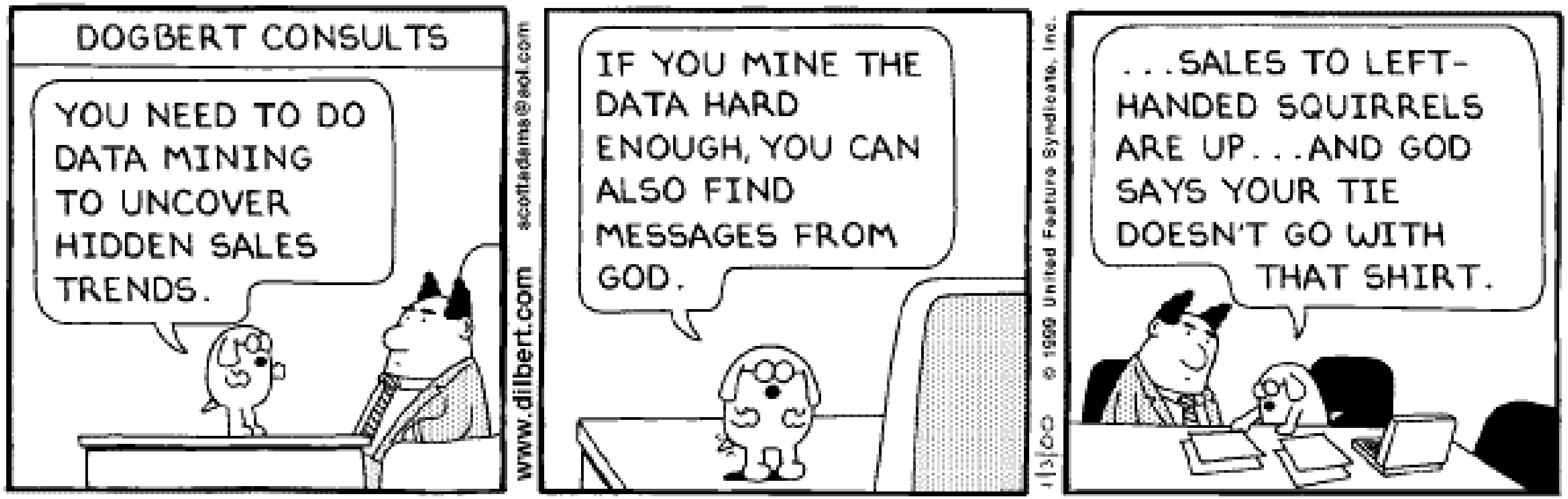


Iteration 10

	Fi H	P Sp I Fr Du G E Da N
Fi H	0	
P Sp I Fr Du G E Da N	8	0

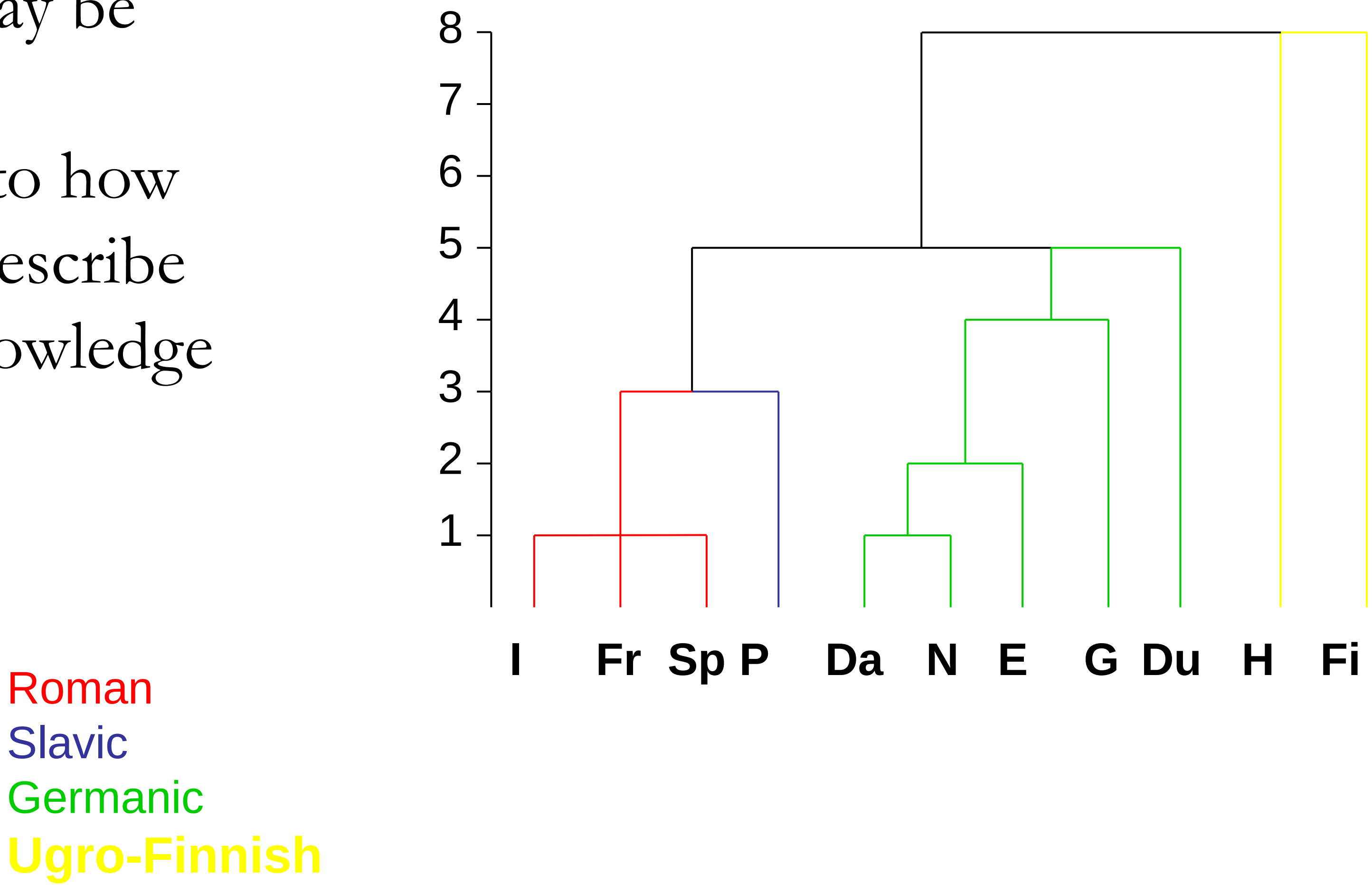


Any data mining result needs to be consistent BOTH with the data and current knowledge!

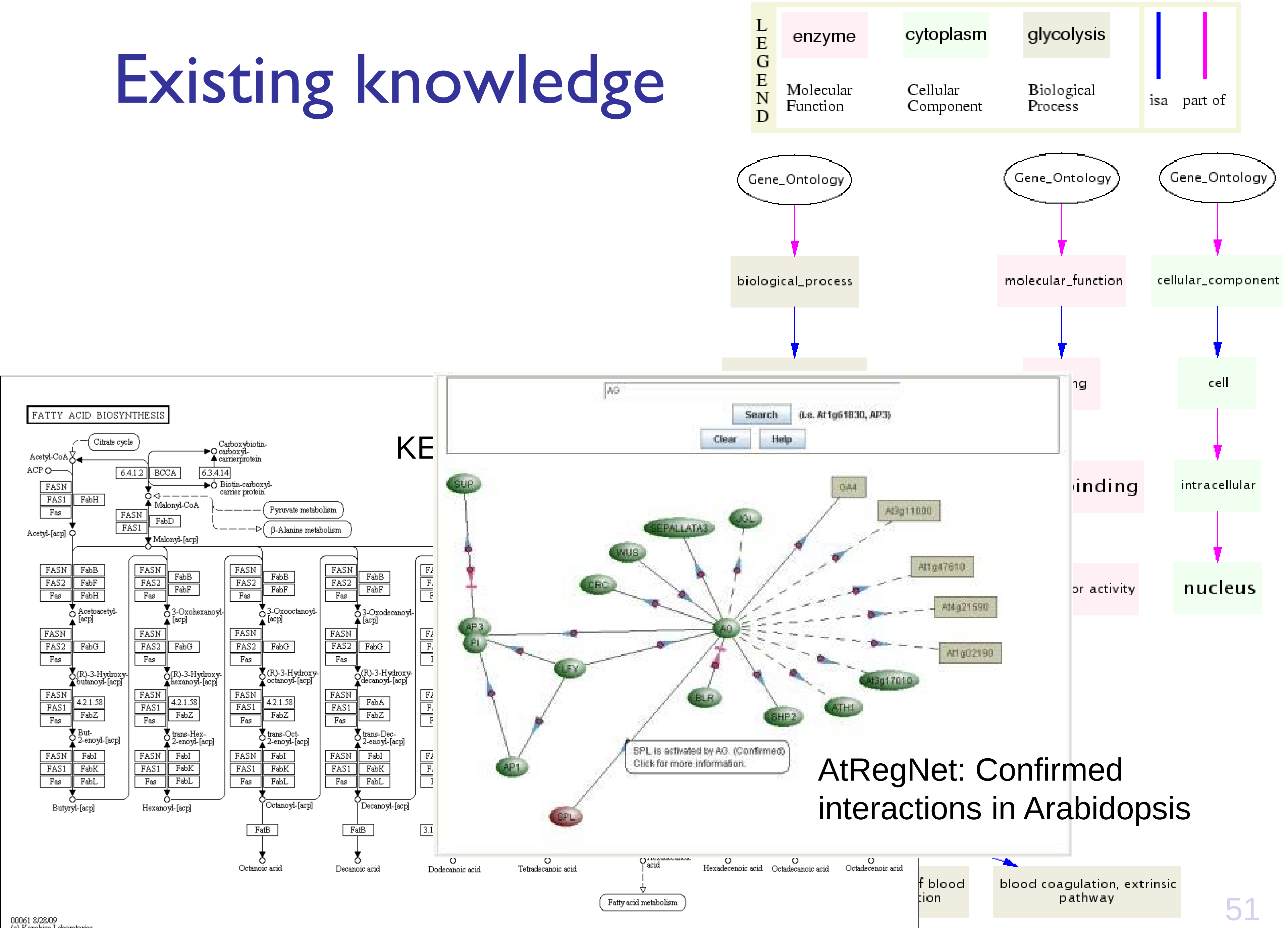

DOGBERT CONSULTS
YOU NEED TO DO DATA MINING TO UNCOVER HIDDEN SALES TRENDS.
scottadams@aol.com
www.dilbert.com
IF YOU MINE THE DATA HARD ENOUGH, YOU CAN ALSO FIND MESSAGES FROM GOD.
© 1999 United Feature Syndicate, Inc.
1/3/00
...SALES TO LEFT-HANDED SQUIRRELS ARE UP...AND GOD SAYS YOUR TIE DOESN'T GO WITH THAT SHIRT.

Evaluation of clusters

Clusters may be evaluated according to how well they describe current knowledge



Existing knowledge



SPL is activated by AG (Confirmed)
Click for more information.

AtRegNet: Confirmed interactions in Arabidopsis

Randomization experiments

- Randomize the input data

- P-values: fraction of randomized datasets resulting in ”better” models than the real data

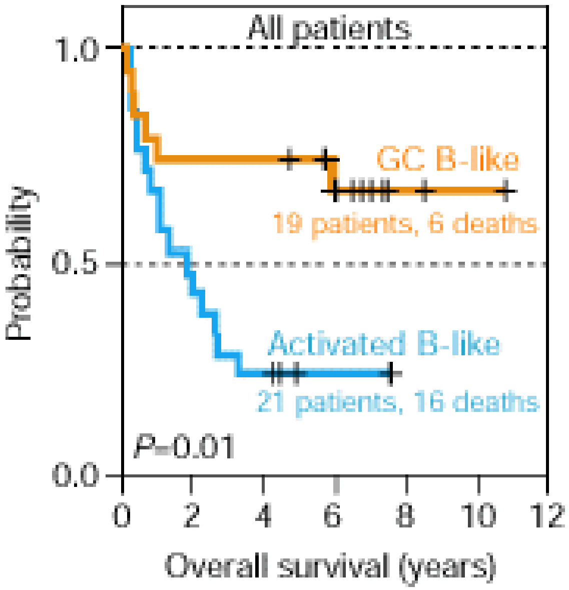
- Better?
  - Cross validation
  - Existing knowledge
  - Other model properties

Example: Hierarchical clustering

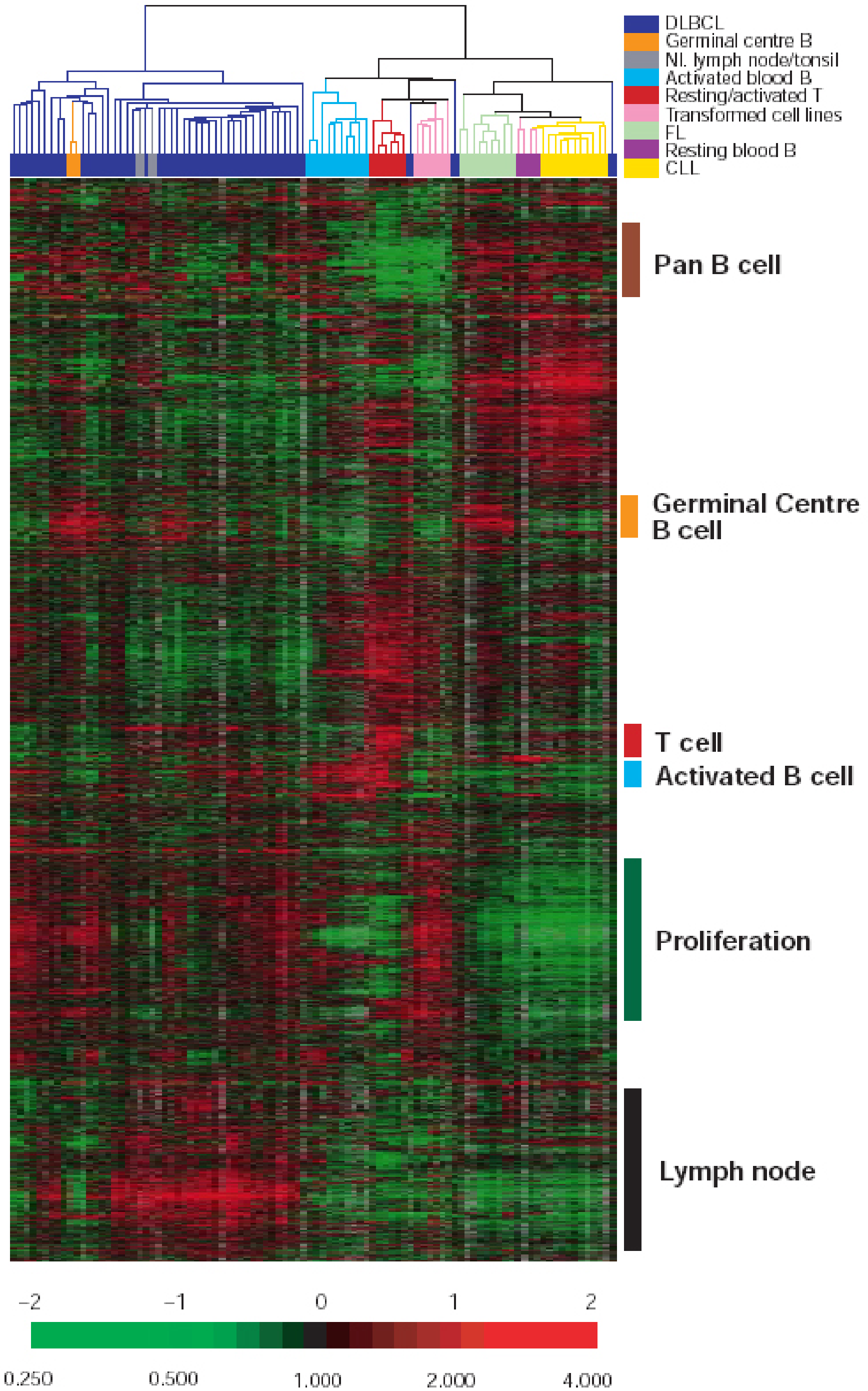
96 normal and malignant lymphocyte samples

Almost 20 000 cDNA clones

Two sub-clusters of DLBCL were shown to include patients with significantly different expected survival time!



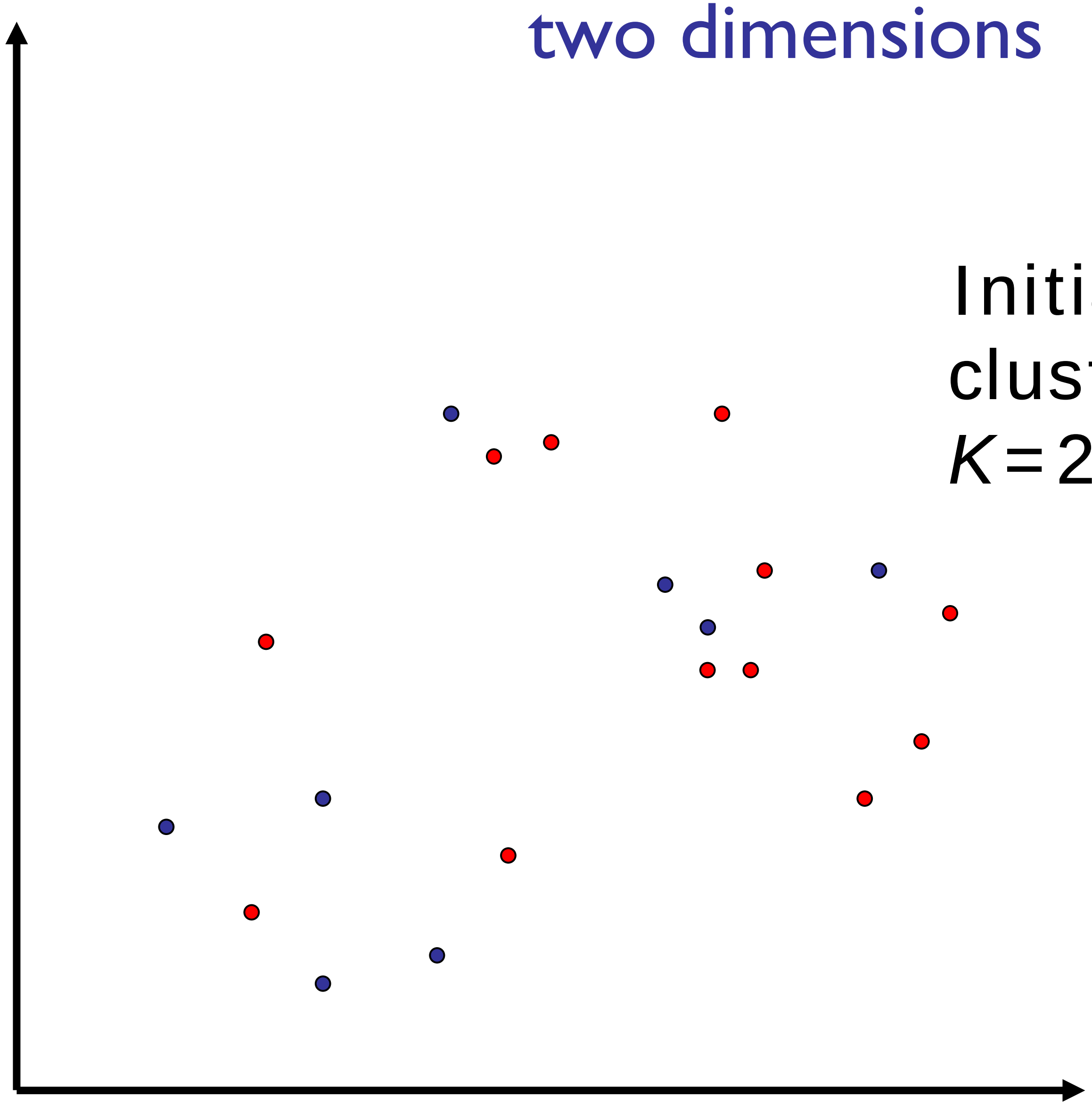
Alizadeh et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, 403:503-511, 2000.



K-means clustering

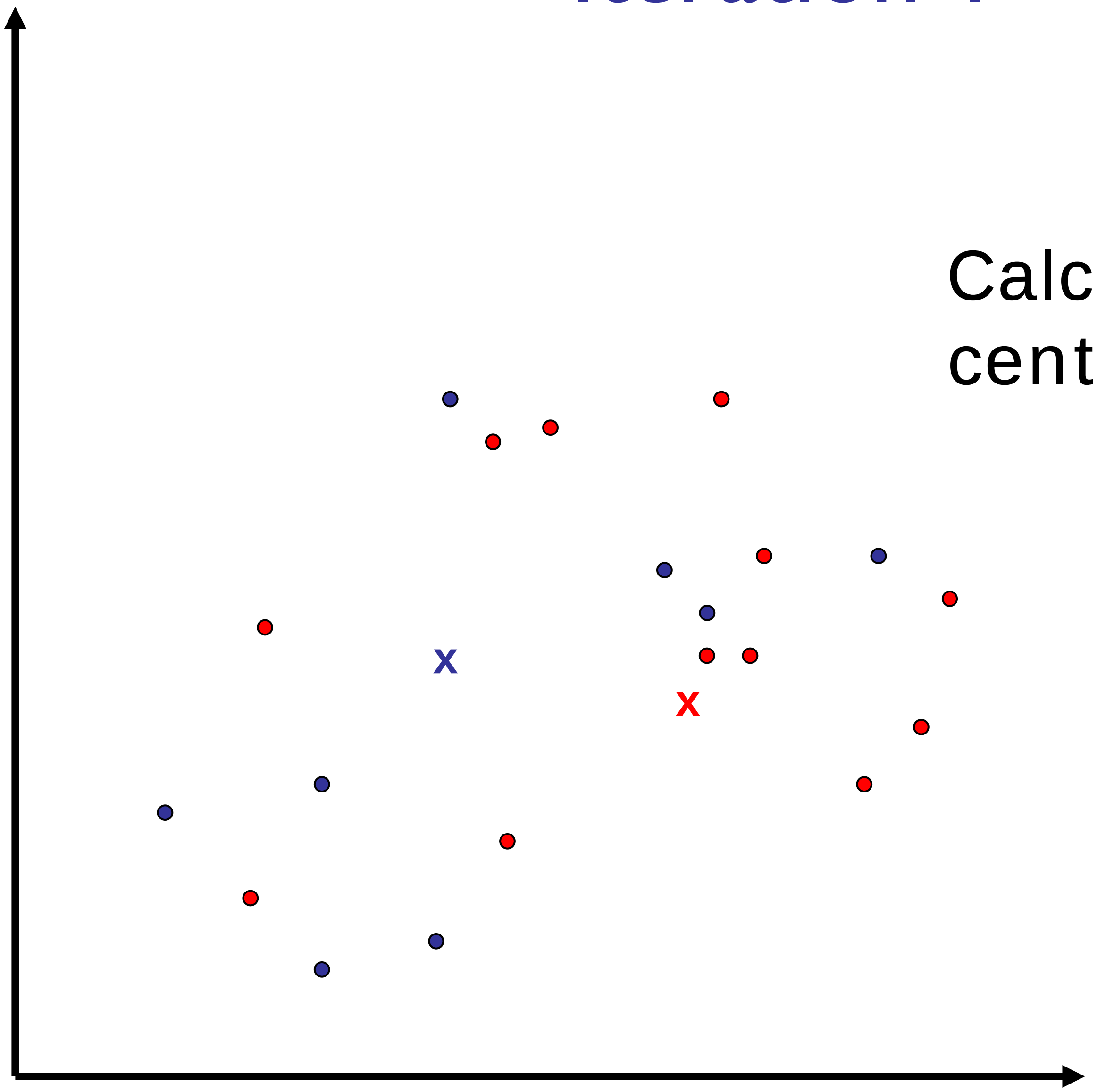
- Split the data into *k* random clusters
- Repeat
  - calculate the centroid of each cluster
  - (re-)assign each gene/experiment to the closest centroid
  - stop if no new assignments are made

Example of K-means: two dimensions



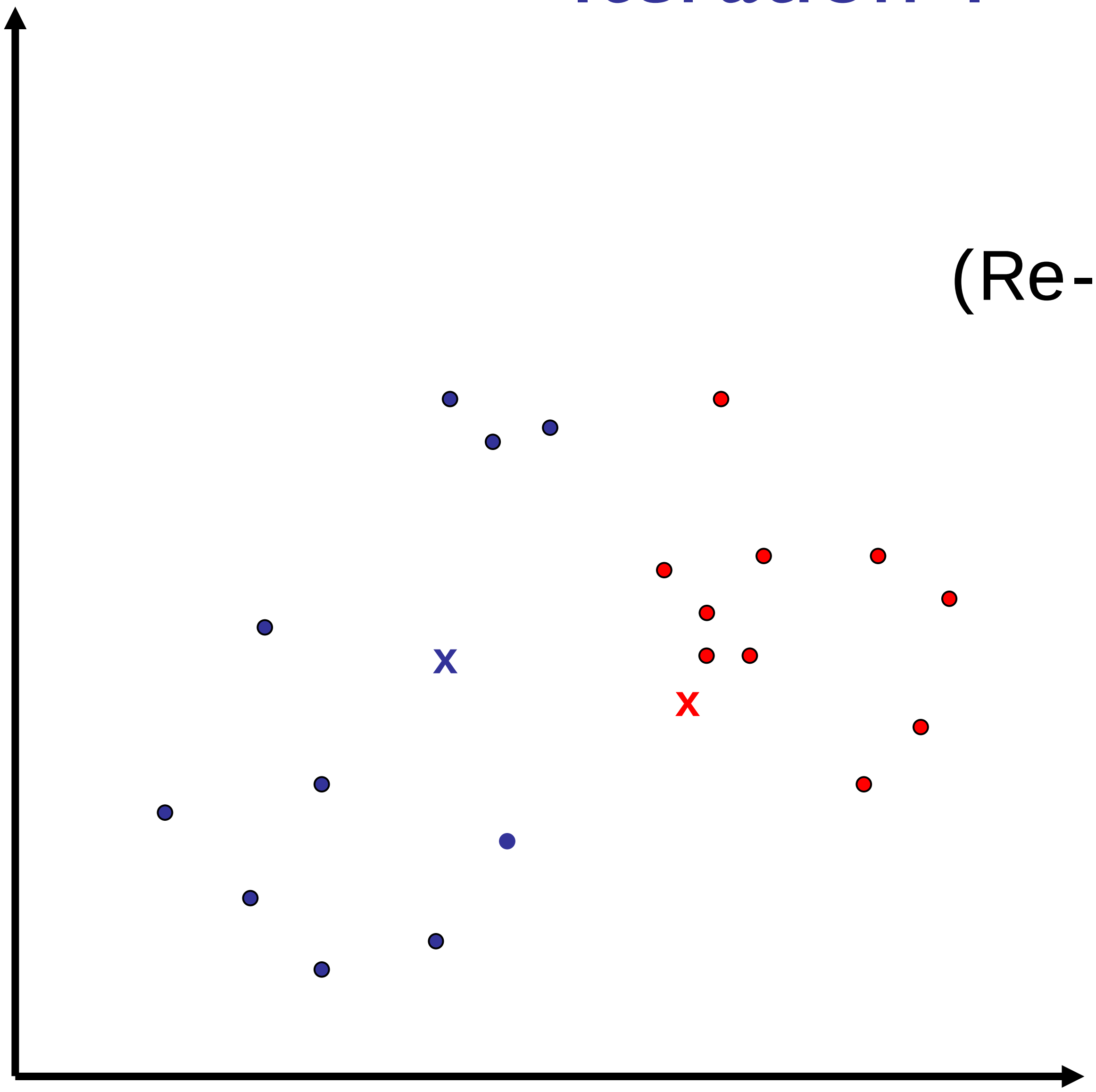
Initial clusters $K=2$

Iteration 1



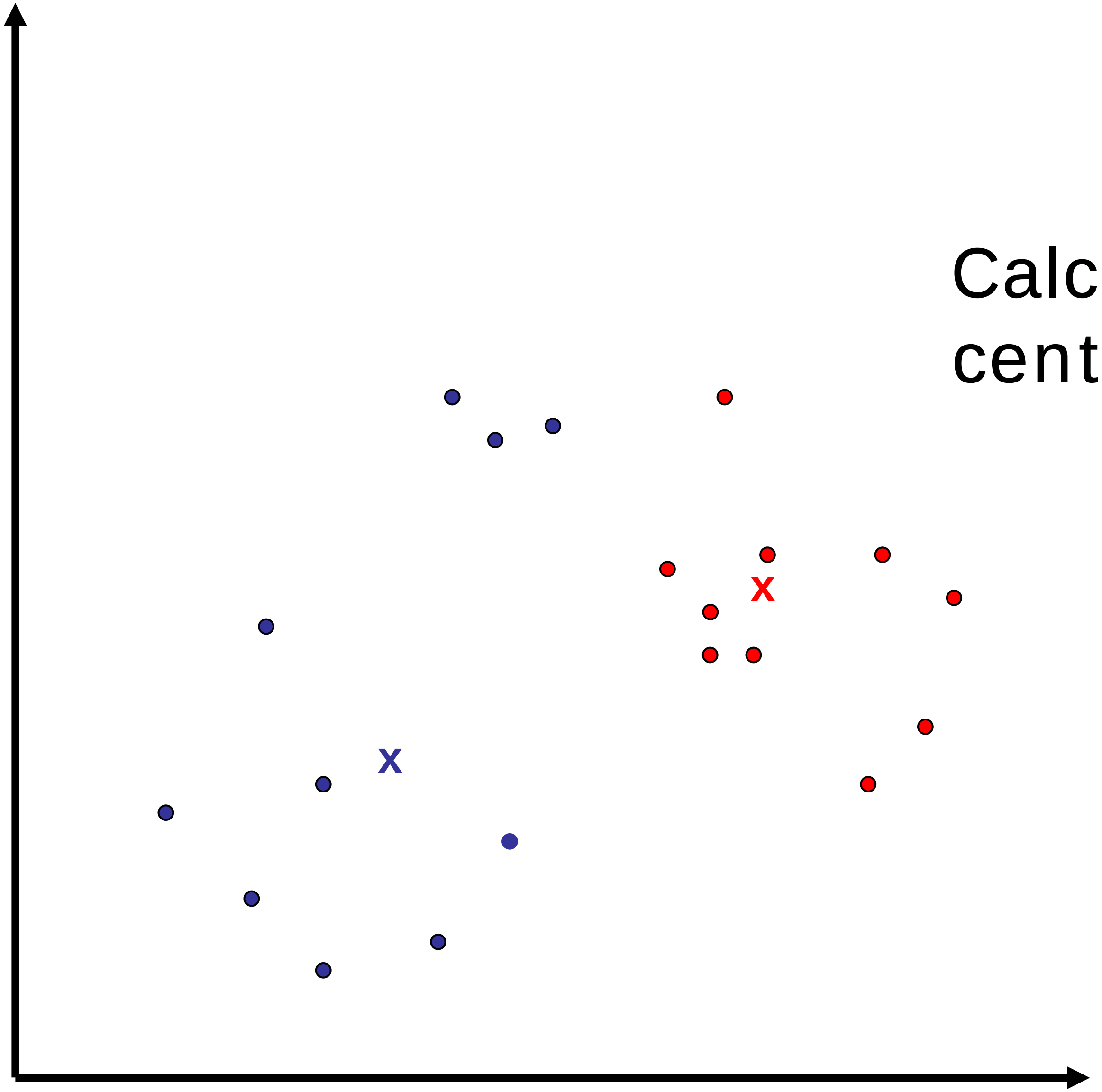
Calculate centroids

Iteration 1



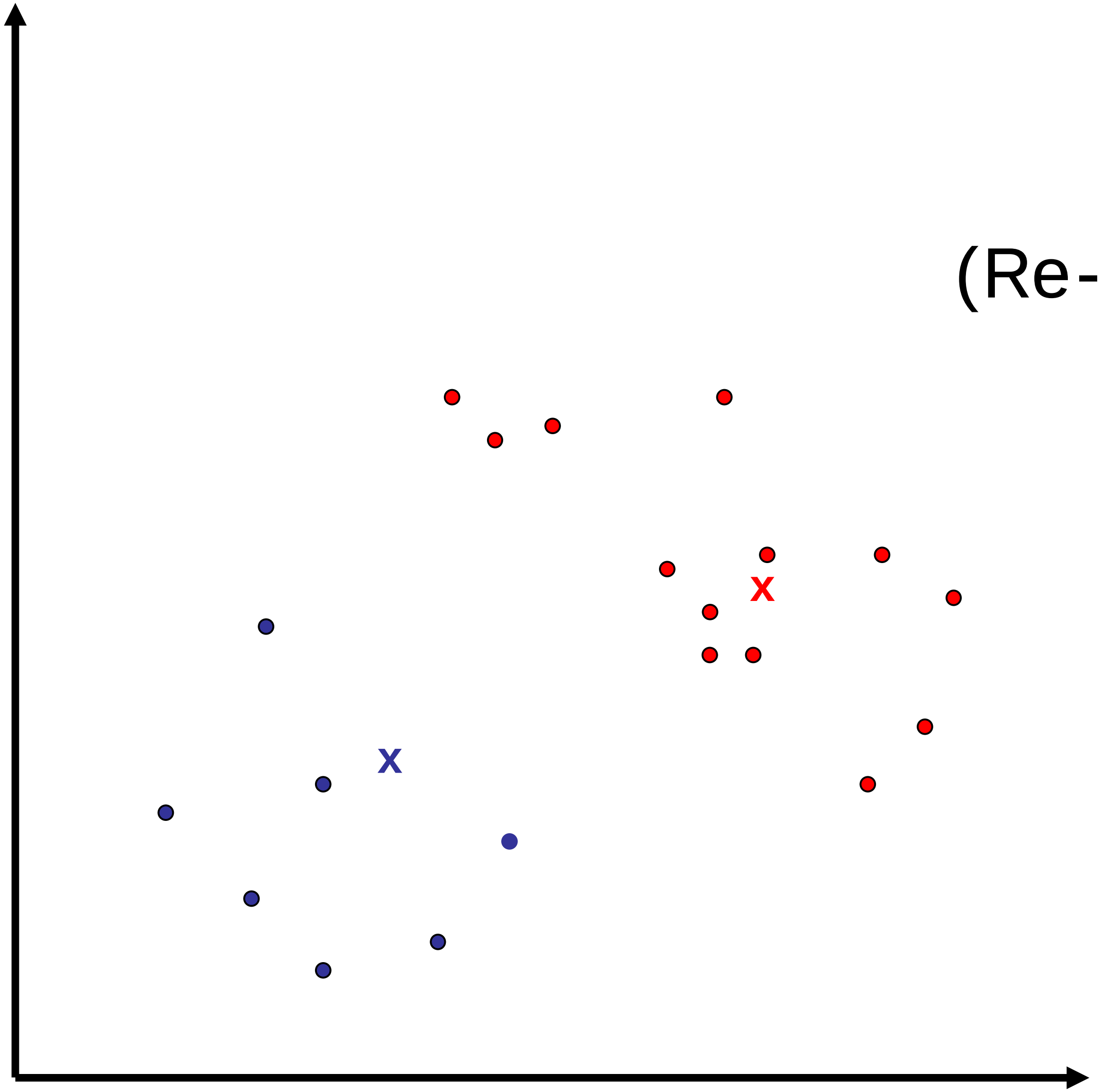
(Re-)assign

Iteration 2



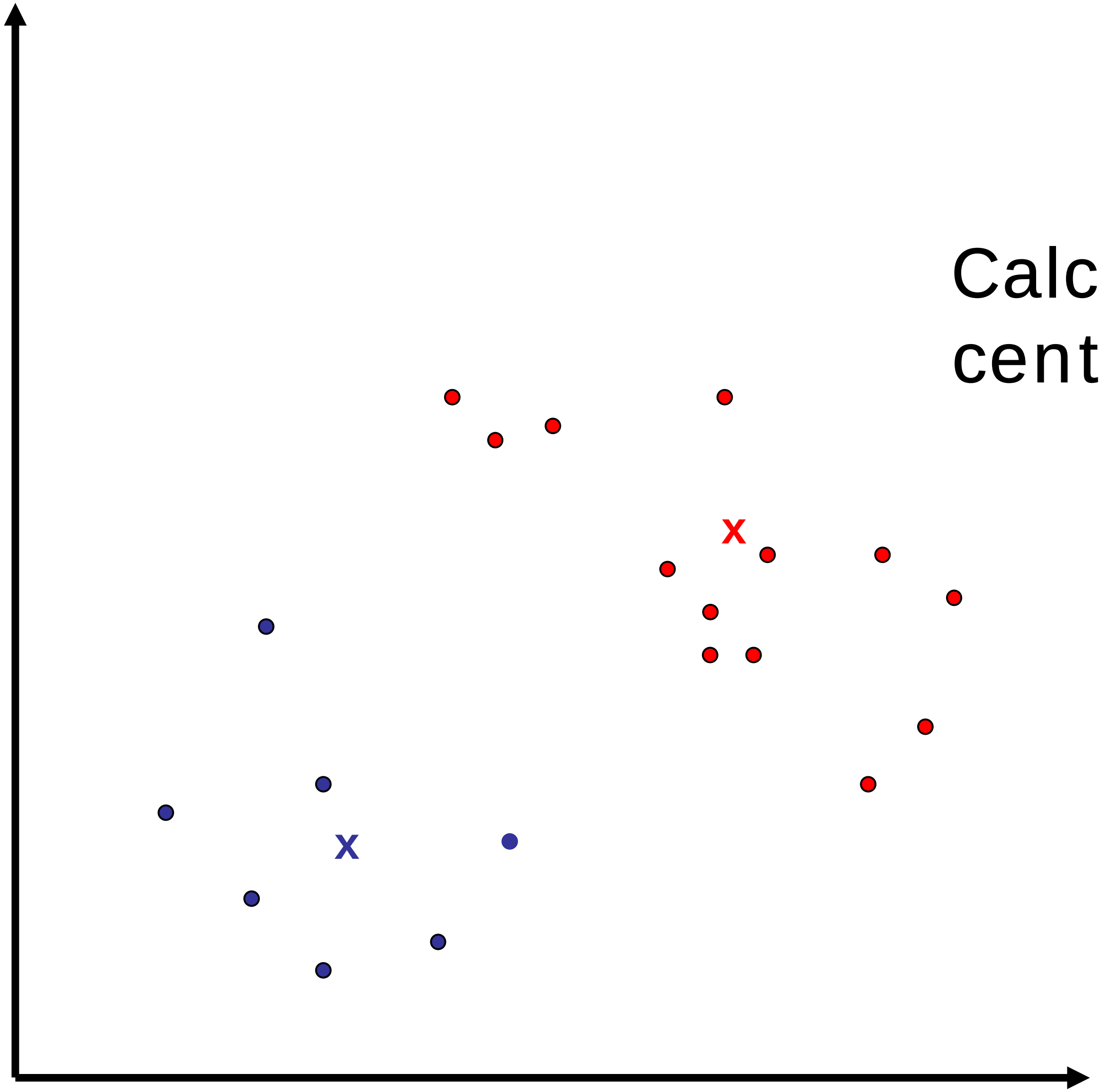
Calculate centroids

Iteration 2



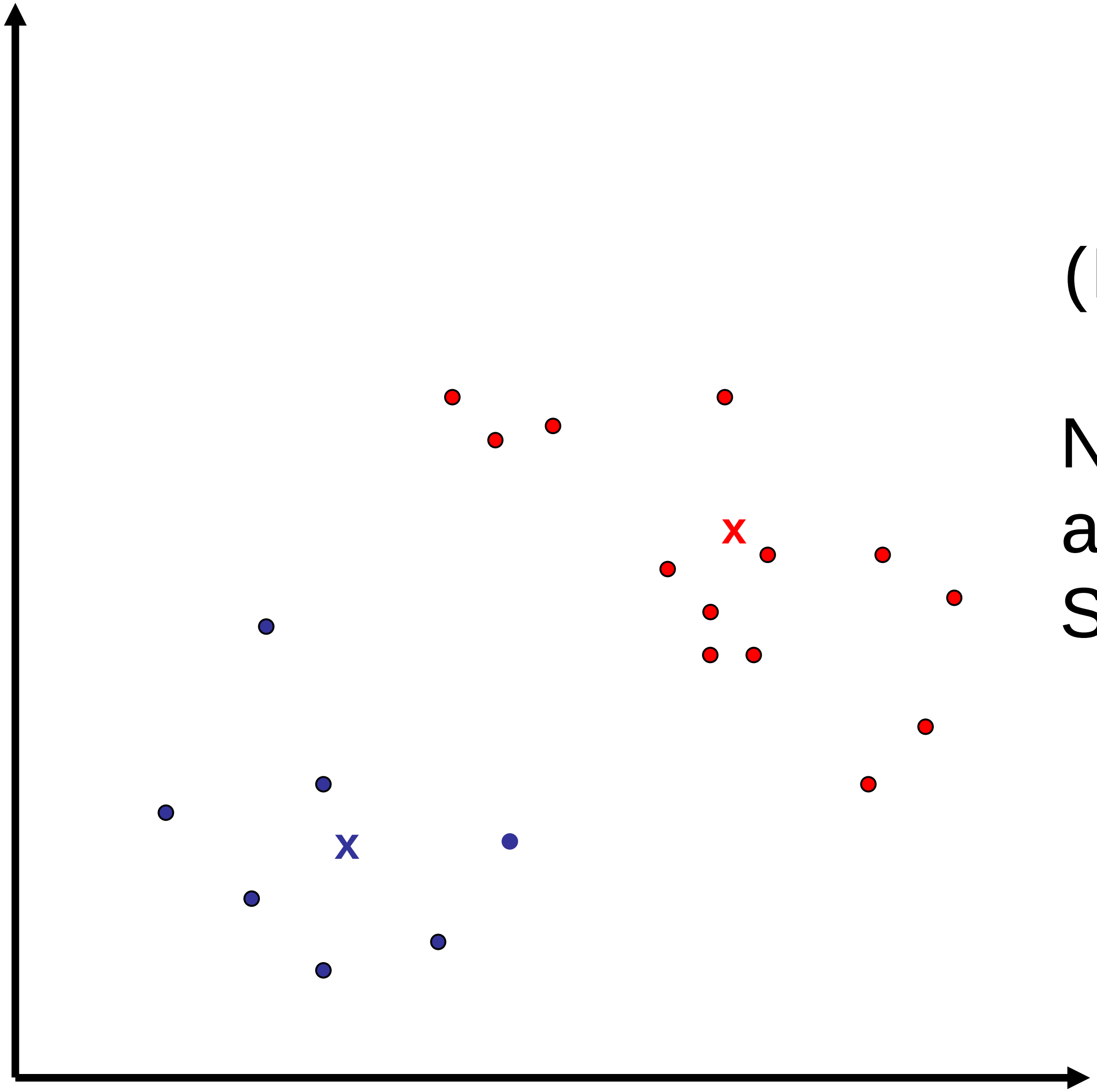
(Re-)assign

Iteration 3



Calculate centroid

Iteration 3



(Re-)assign

No new assignments! STOP

Hierarchical vs. k-means

- Hierarchical clustering:
  - Huge memory requirements: stores the $n \times n$ matrix
  - Running time: O(n^3)
  - Nice visualization: dendrogram
  - Deterministic
  - Cannot correct early ”mistakes” (greedy alg.)

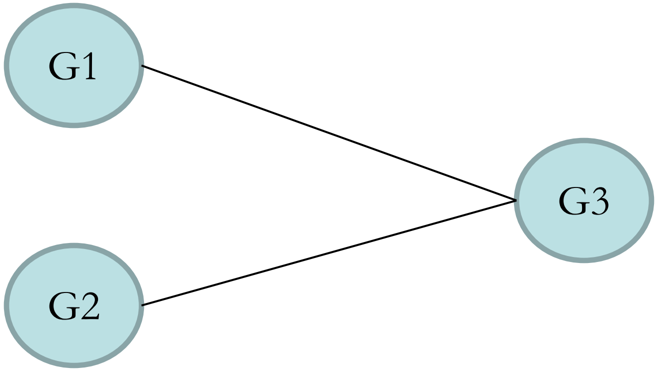
- K-means:
  - Low memory usage
  - Running time: O(kn), where k is the number of iterations
  - Improves iteratively: not trapped in previous mistakes (randomization alg.)
  - Non-deterministic: will produce different clusters with different initializations
  - Number of clusters must be decided in advance

Network representations

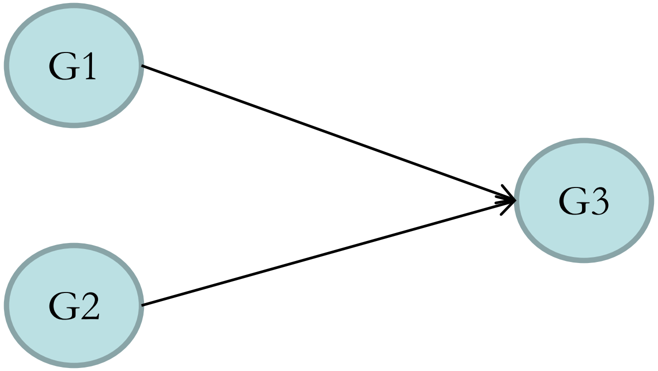
- Network: nodes connected by edges
- Nodes represent genes, proteins, metabolites
- Edges represent relationships
  - Co-expression networks: expression correlation
  - Gene networks: genes affect the expression of other genes
  - Regulatory network: transcription factors regulate genes by binding DNA motifs in the promoter region

- Network representations are flexible and allow integration of heterogeneous data

Co-expression networks versus gene networks

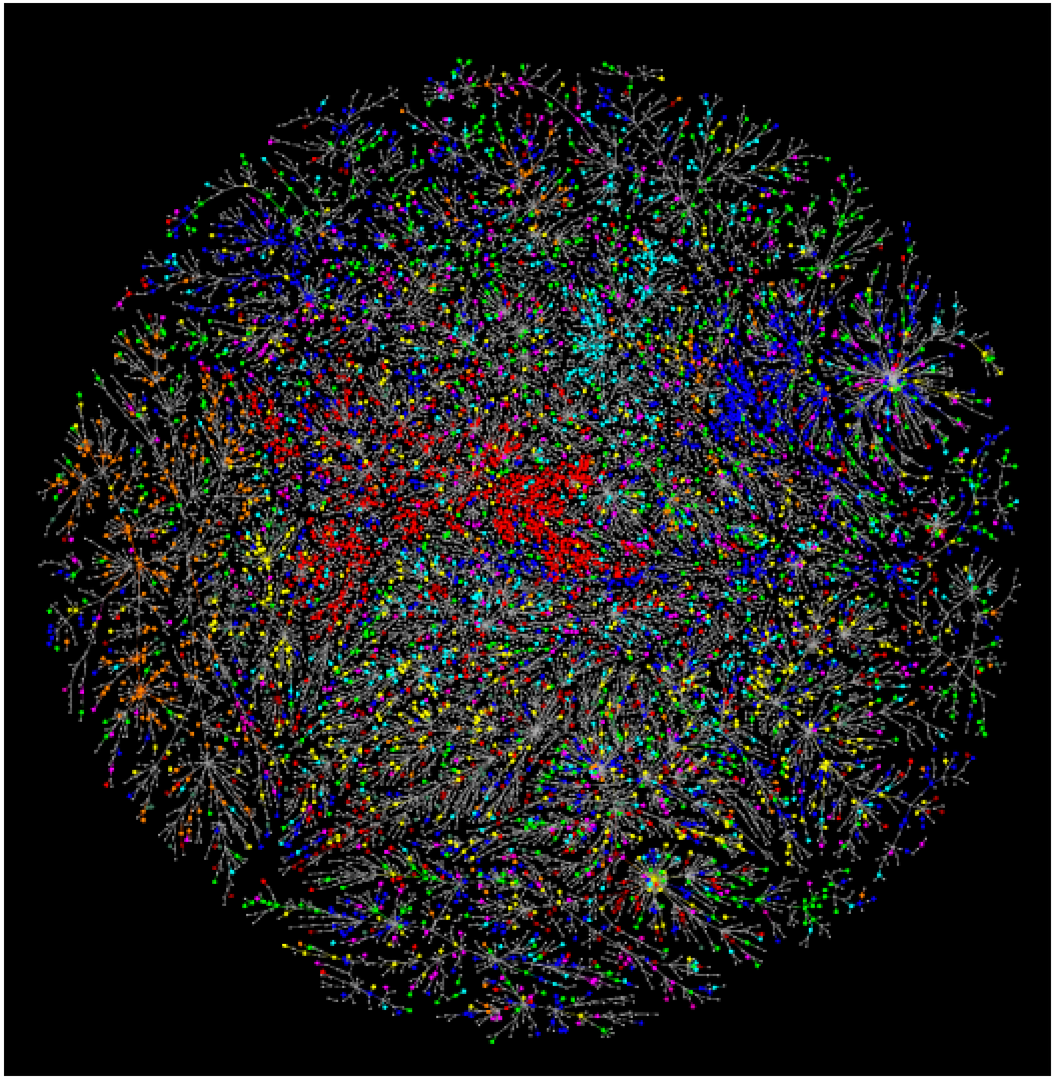


Co-expression network:
Expression of G1 correlates with that of G3
Expression of G2 correlates with that of G3



Gene network:
The expression of G3 can be predicted from that of G1 *and* G2

Co-expression network in aspen trees



Based on a UPSC collection of over 1000 cDNS microarrays

A Grönlund, RP Bhalerao, J Karlsson. Modular gene expression in Poplar: a multilayer network approach. New Phytologist, 2009.

Regulatory network in Arabidopsis

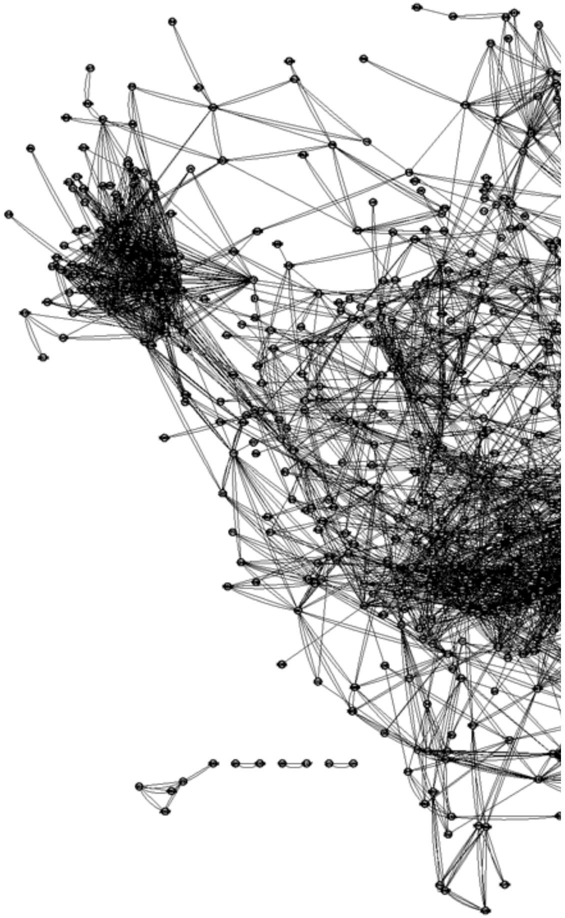


Figure 1

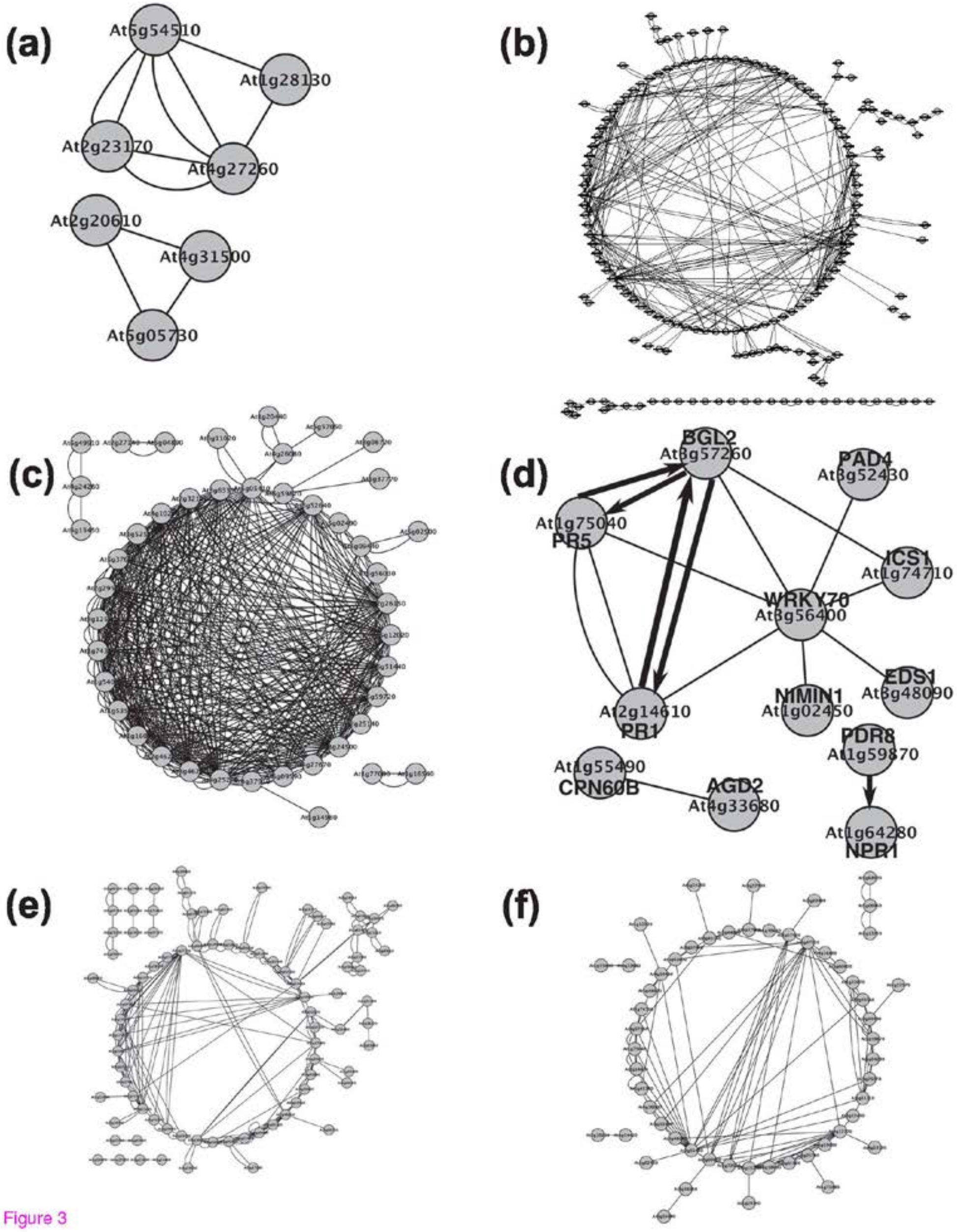
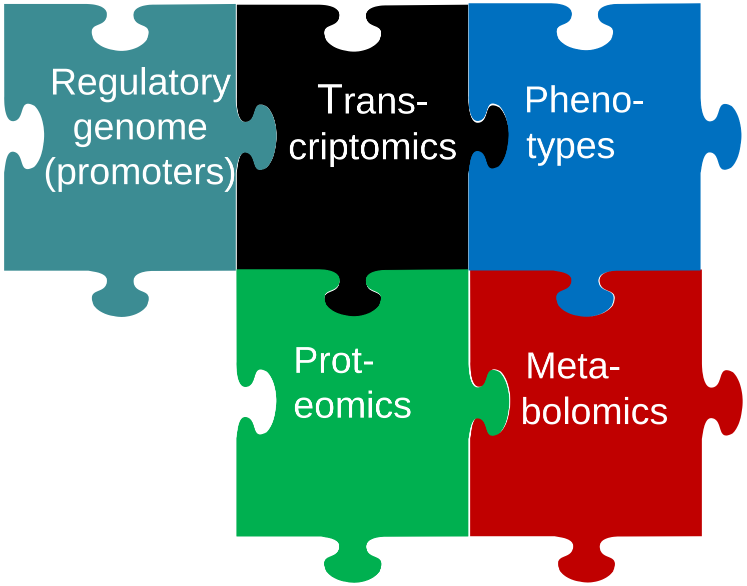


Figure 3

J. Carrera , G. Rodrigo , A. Jaramillo and S. F Elena. Rev
network under changing environmental conditions. Genor

Systems biology



Regulatory genome (promoters)

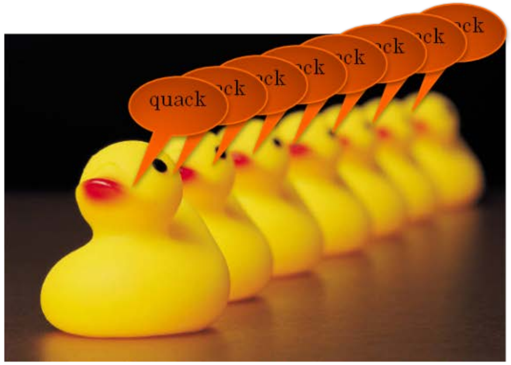
Transcriptomics

Phenotypes

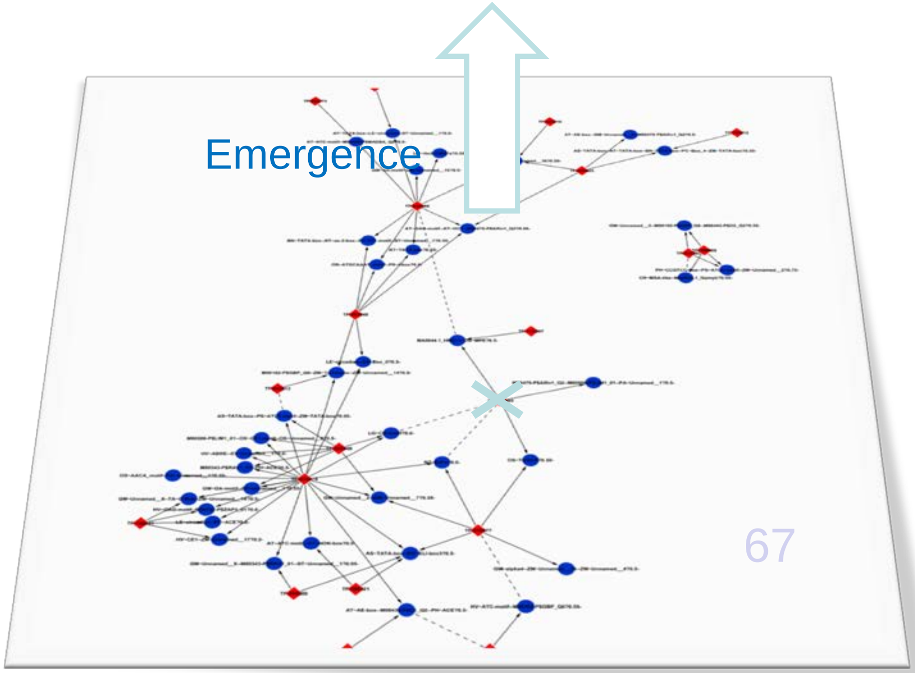
Proteomics

Metabolomics

Synergy from integration



Phenotypes



Emergence

Interacting genes/protein/metabolites

67

Holistic versus reductionistic

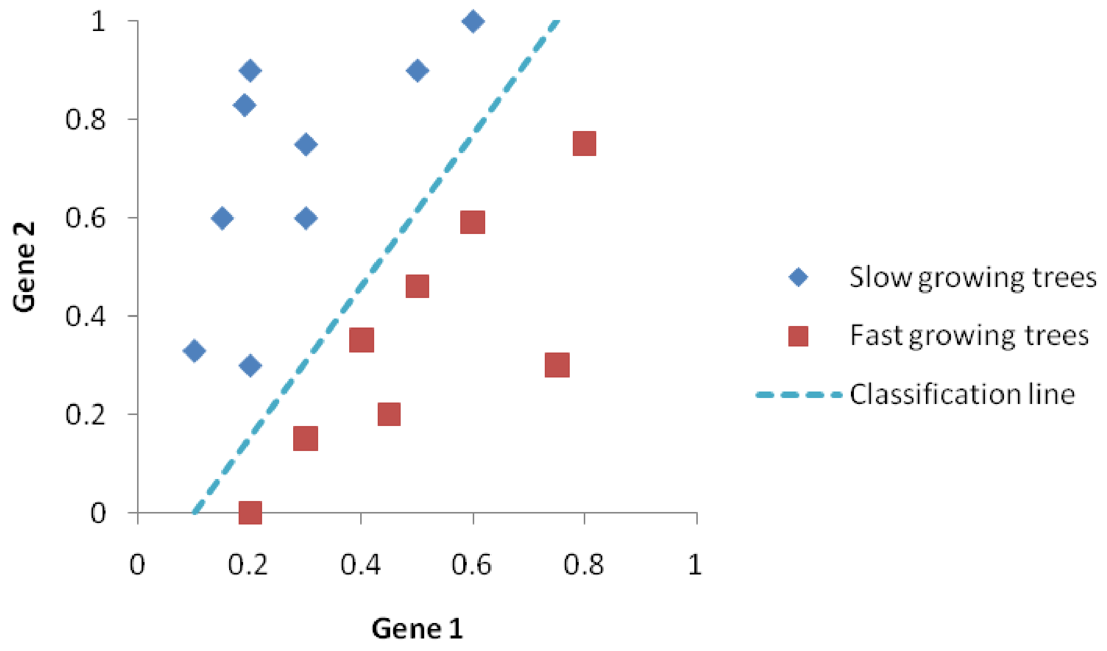
- Emergent properties:
  - Can biology be reduced to chemistry?
  - Can chemistry be reduced to physics?

  - Ernest Rutherford : "Physics is the only real science. The rest are just stamp collecting.“

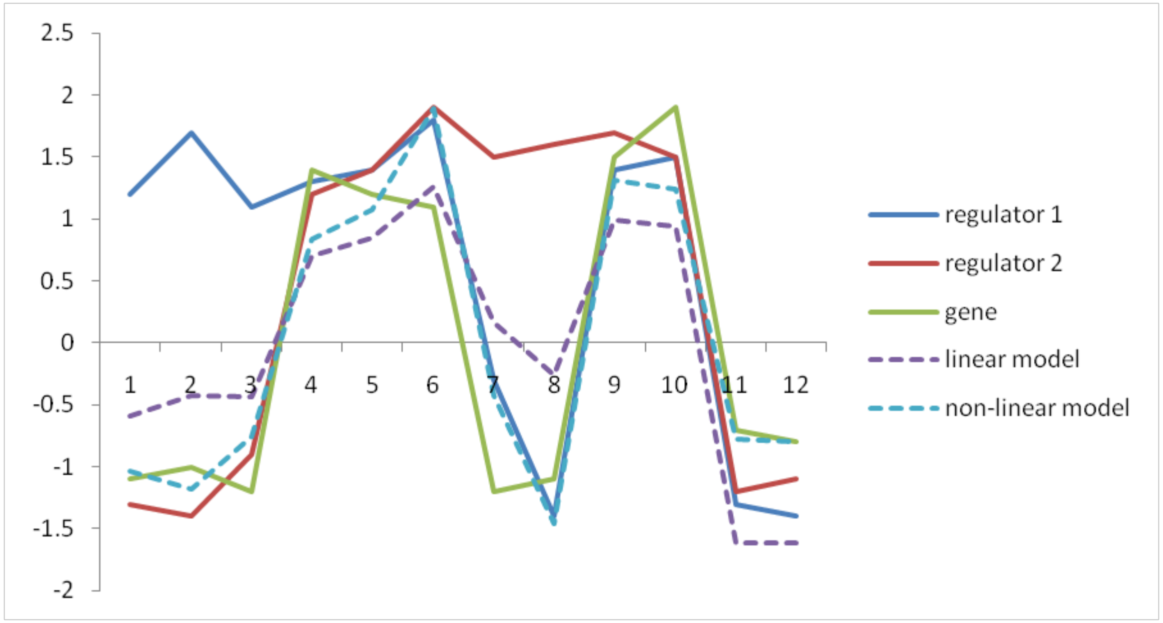
- Can biological systems be reduced to individual genes, proteins and metabolites?

Derek Gatherer . So what do we really mean when we say that systems biology is holistic? *BMC Systems Biology* 4: 22, 2010.

Emergent properties: differential expression

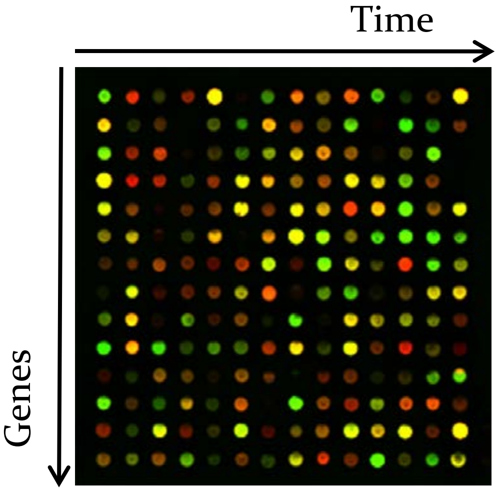


Emergent properties: AND logics in regulation



Inferring regulatory mechanism

Time series data



For each gene *i* :

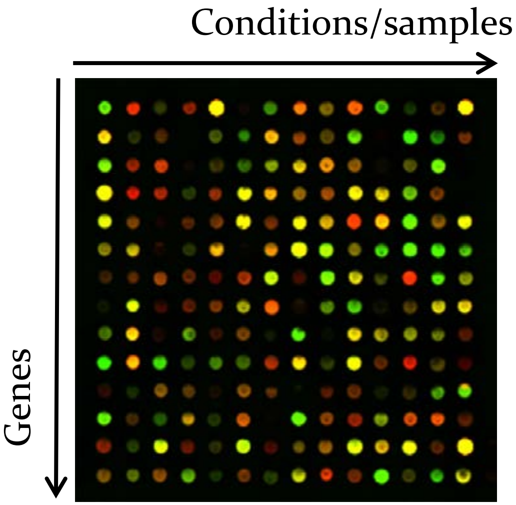
$$\frac{dy_i}{dt} = \alpha_i - \partial_i y_i + \sum_j \beta_{ij} y_j$$

where α_i is its transcription rate,

∂_{i} the degradation coefficient,

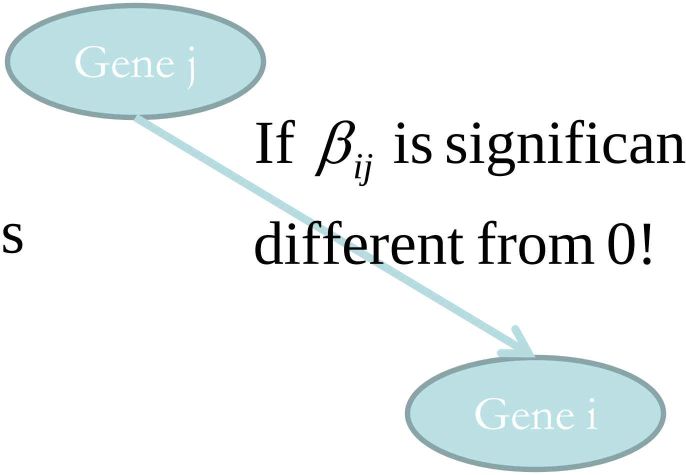
and β_{ij} is the regulatory effect that gene *j* has on gene *i*.

Steady state data



$$\frac{dy_i}{dt} = 0 \text{ and } \partial_{\mathrm{i}} = 1\text{, thus}$$

$$\boxed{y_i = \alpha_i + \sum_j \beta_{ij} y_j}$$



If β_{ij} is significantly different from 0!

Example: Three genes

α = -0.46
β_{12}= 0.43
β_{13}=0.50

$$y_1 = \alpha + \beta_{12} y_2 + \beta_{13} y_3$$

Expr	y_2	y_3	y_1	y_1 predicted	
Cond. A	1.2	-1.3	-1.1	$a + \beta_{12} \cdot 1.2 - \beta_{13} \cdot 1.3$	-0.594
Cond. B	1.7	-1.4	-1	$a + \beta_{12} \cdot 1.7 - \beta_{13} \cdot 1.4$	-0.429
Cond. C	1.1	-0.9	-1.2	$a + \beta_{12} \cdot 1.2 - \beta_{13} \cdot 0.9$	-0.437
Cond. D	1.3	1.2	1.4	$a + \beta_{12} \cdot 1.3 + \beta_{13} \cdot 1.2$	0.699
Cond. E	1.4	1.4	1.2	$a + \beta_{12} \cdot 1.4 + \beta_{13} \cdot 1.4$	0.842
Cond. F	1.8	1.9	1.1	$a + \beta_{12} \cdot 1.8 + \beta_{13} \cdot 1.9$	1.264
…	…	…	…	…	…

Correlation: 0.78

Choose α, $\beta_{12}$and β_{13} so that the correlation between observed (y_1) and predicted (y_1 predicted) expression is maximized!

Linear versus non-linear models

- Linear model:

$$y_1 = \alpha + \beta_{12} y_2 + \beta_{13} y_3$$

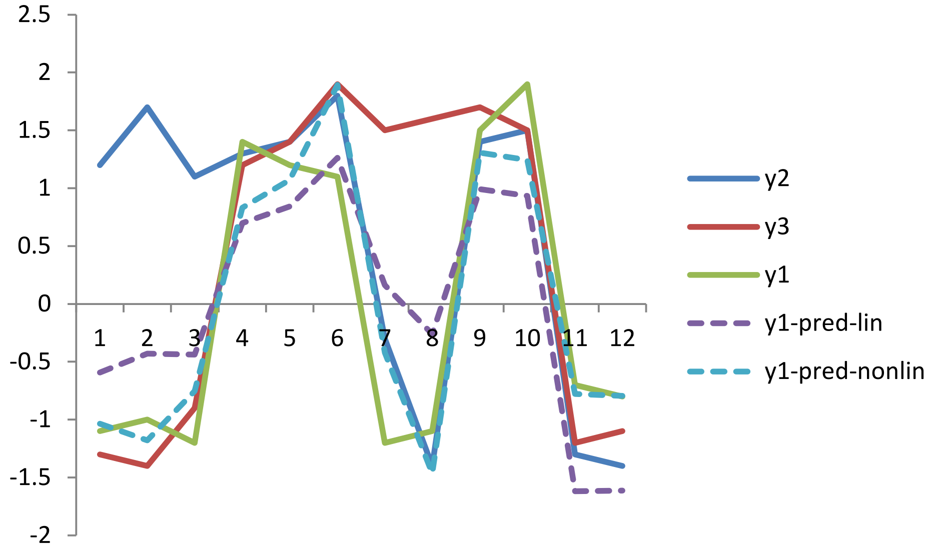
- Non-linear model:

$$y_1 = \alpha + \beta_{12} y_2 + \beta_{13} y_3 + \beta_{123} y_2 y_3$$

$$\beta_{123} > 0 : \text{synergistic interactions}$$

$$\beta_{123} < 0 : \text{competitive relationship}$$

AND - logic



Linear model:

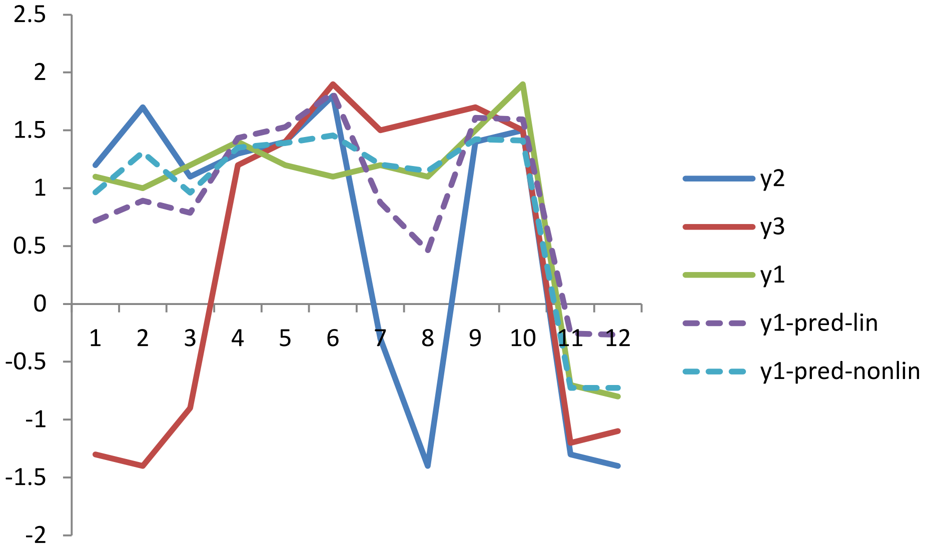
α = -0.46
β_{12}= 0.43
β_{13}=0.50

Non-linear model:

α = -0.55
β_{12}= 0.37
β_{13}=0.27
β_{123}=0.37

Correlation between observed and predicted:	
Linear model:	0.77
Non-linear model:	0.91
Correlation between gene 1 and	
gene 2:	0.55
gene 3:	0.65

OR - logic



Linear model:

$\alpha = 0.59$

$\beta_{12}= 0.40$

$\beta_{13}=0.27$

Non-linear model:

$\alpha = 0.64$

$\beta_{12}= 0.43$

$\beta_{13}=0.40$

$\beta_{123}=-0.21$

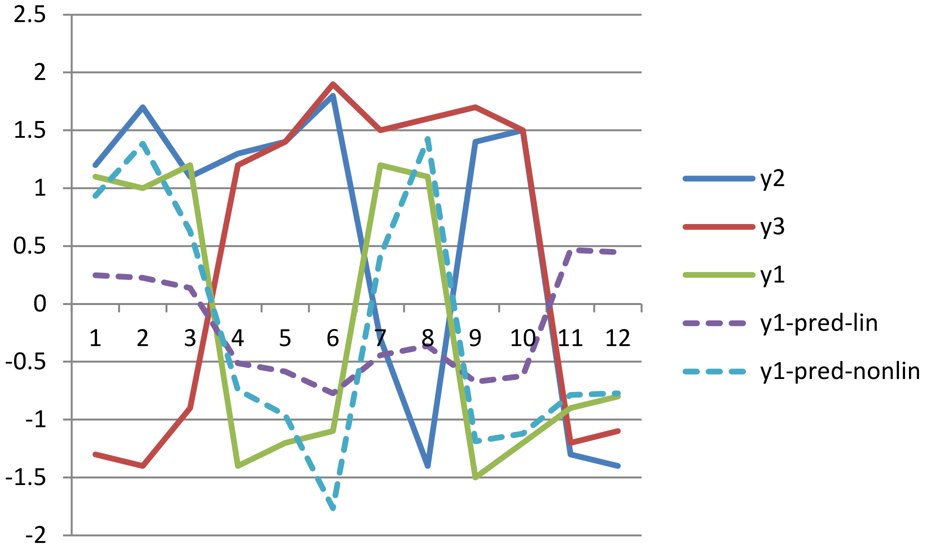
Correlation between observed and predicted:

Linear model:	0.85
Non-linear model:	0.96

Correlation between gene 1 and

gene 2:	0.72
gene 3:	0.60

XOR - logic



Linear model:

α = -0.02
β_{12}= -0.10
β_{13}=-0.30

Non-linear model:

α = 0.11
β_{12}= -0.01
β_{13}=0.03
β_{123}=-0.56

Correlation between observed and predicted:

Linear model:	0.40
Non-linear model:	0.92

Correlation between gene 1 and

gene 2:	-0.19
gene 3:	-0.39

Overfitting and the course of dimensionality

$$x = 7y$$
$$y = 3 + x$$

Has a unique solution: x=-3.5, y=-0.5

$$x = 7y$$
$$y = z + x$$

Has many solutions: z=3, x=-3.5, y=-0.5
z=6, x=-7, y=-1
…

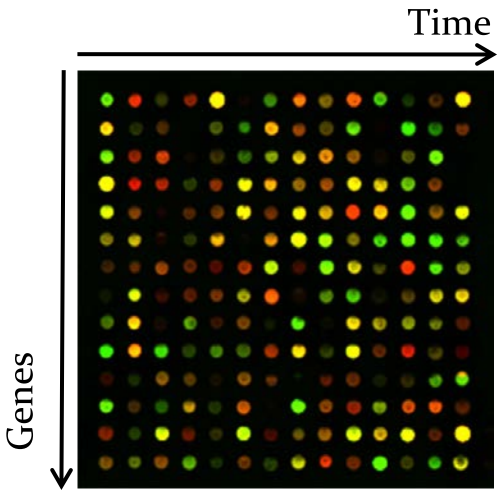
i.e. we need more samples than genes in order to solve:

$$\boxed{y_i = \alpha_i + \sum_j \beta_{ij} y_j}$$

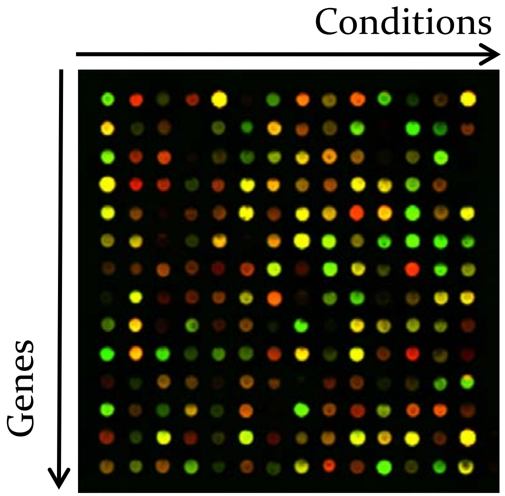
there are ~45 000 genes in *Populus* …
and even ~2500 transcription factors …

Data dimentionality:
How many samples do I need?

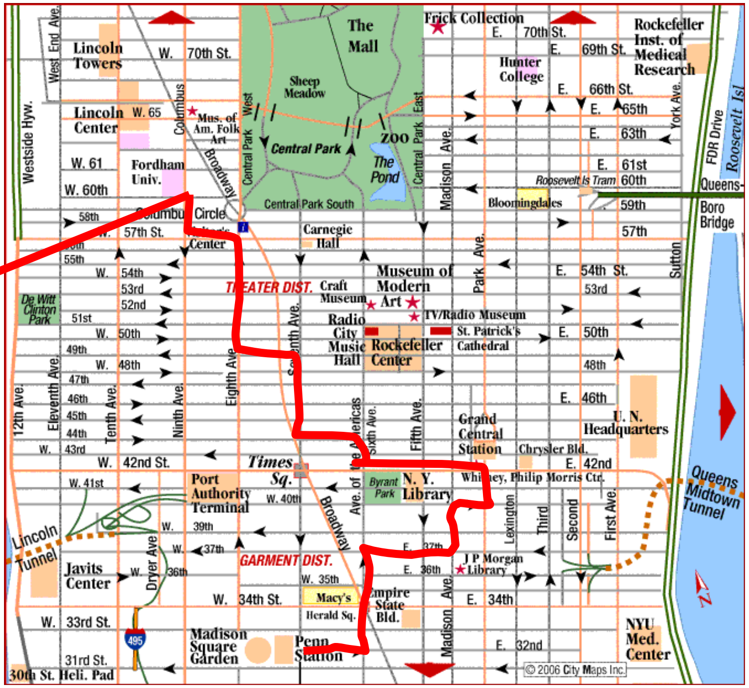
Time series data



Steady state data



- Predicting relationships between genes require high quality data observed over many different conditions
- Co-expression: Analogous to establishing whether you are being followed by the car behind you



- Gene networks: Analogous to establishing whether you are being followed by many collaborating car behind you

Complexity of data analysis

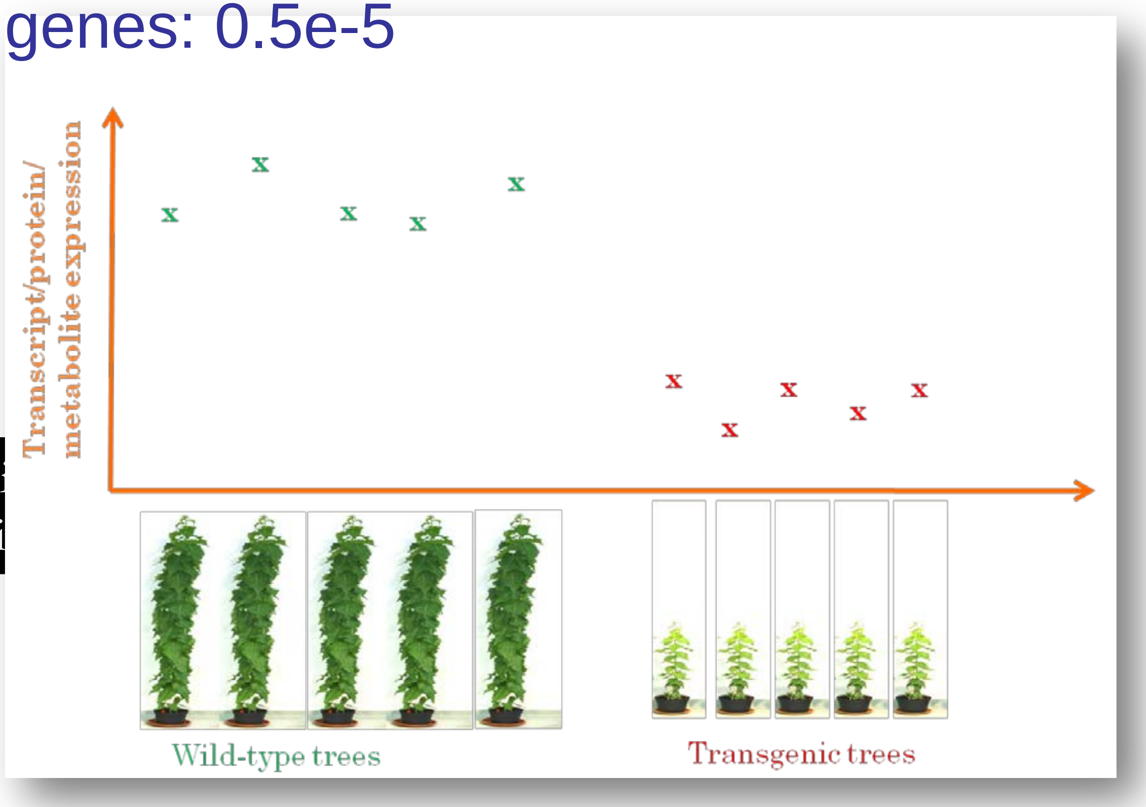
To do e.g. a t-test you need at least three biological replicates from each class

Bonferroni for 10k genes: 0.5e-5

Complexity

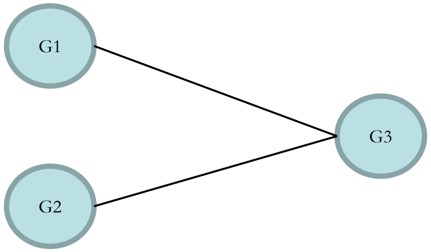
Expressi
similari

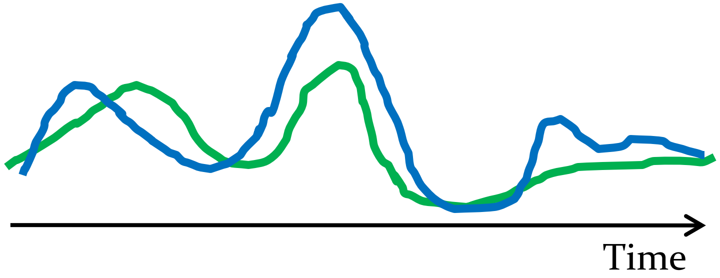
Differential expression



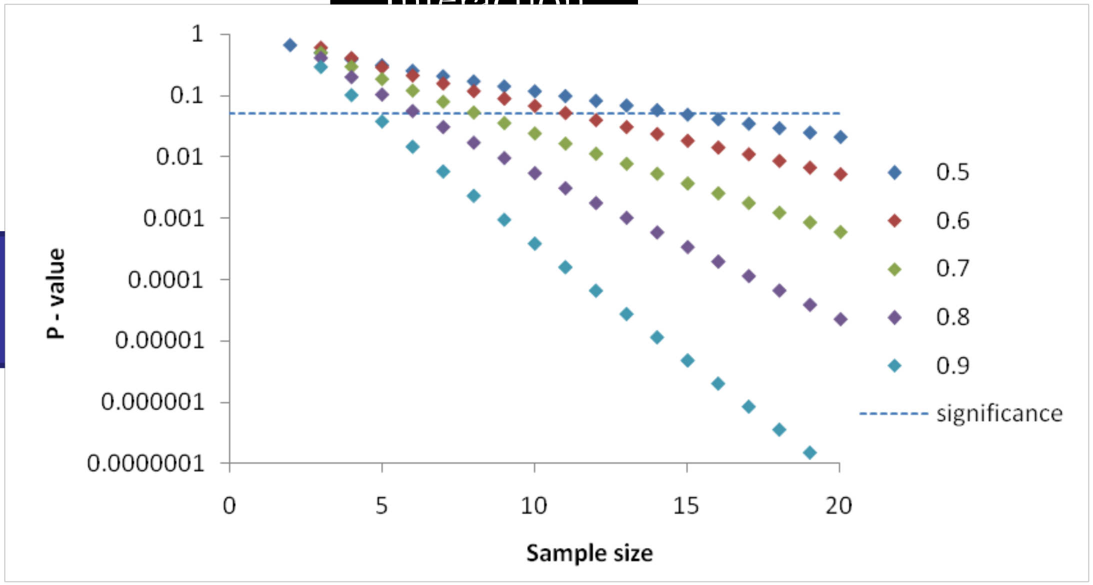
Complexity of data analysis

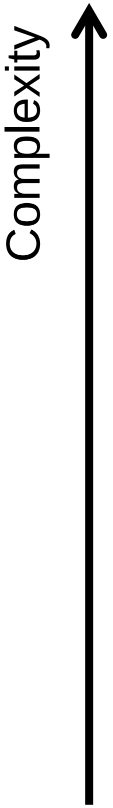
Co-expression network:





Similarity can happen by chance (e.g. Pearson correlation).





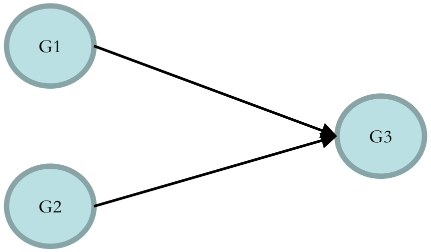
Differential expression

Expression similarity

Interaction

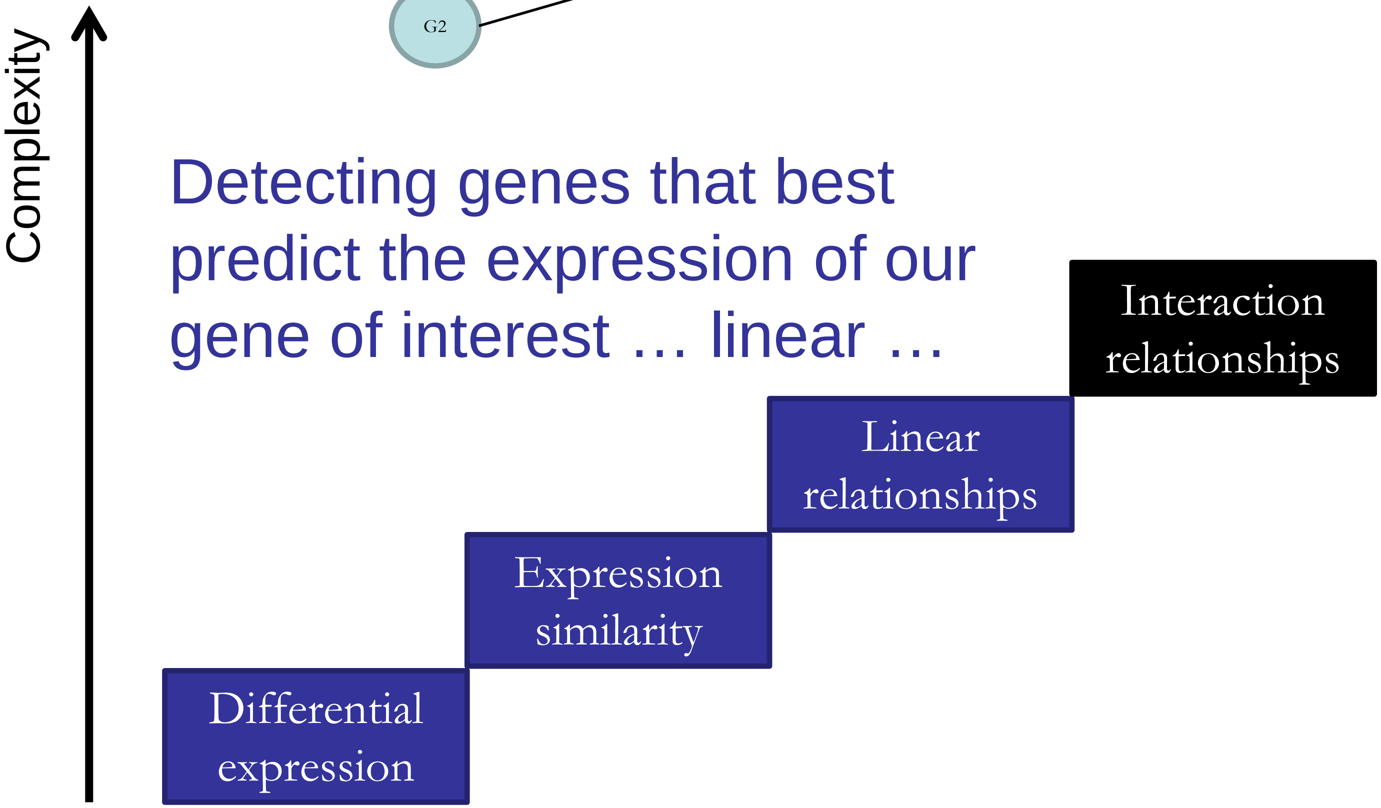
Complexity of data analysis

Gene network:



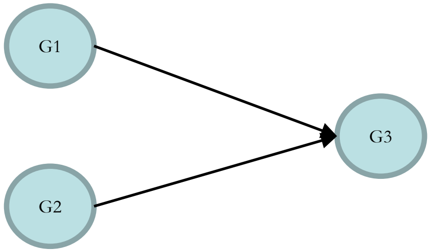
$$y_1 = \alpha + \beta_{12} y_2 + \beta_{13} y_3$$

Detecting genes that best predict the expression of our gene of interest … linear …



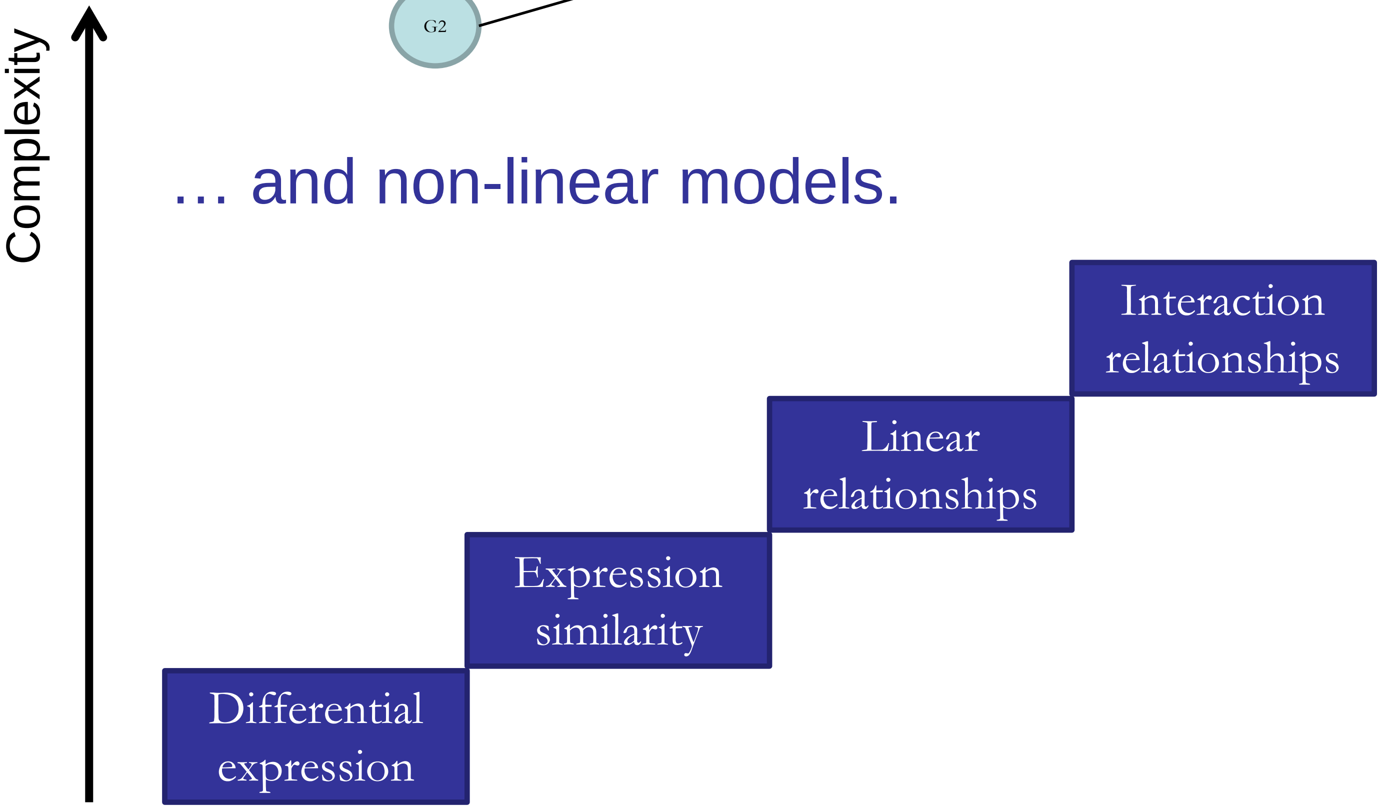
Complexity of data analysis

Gene network:



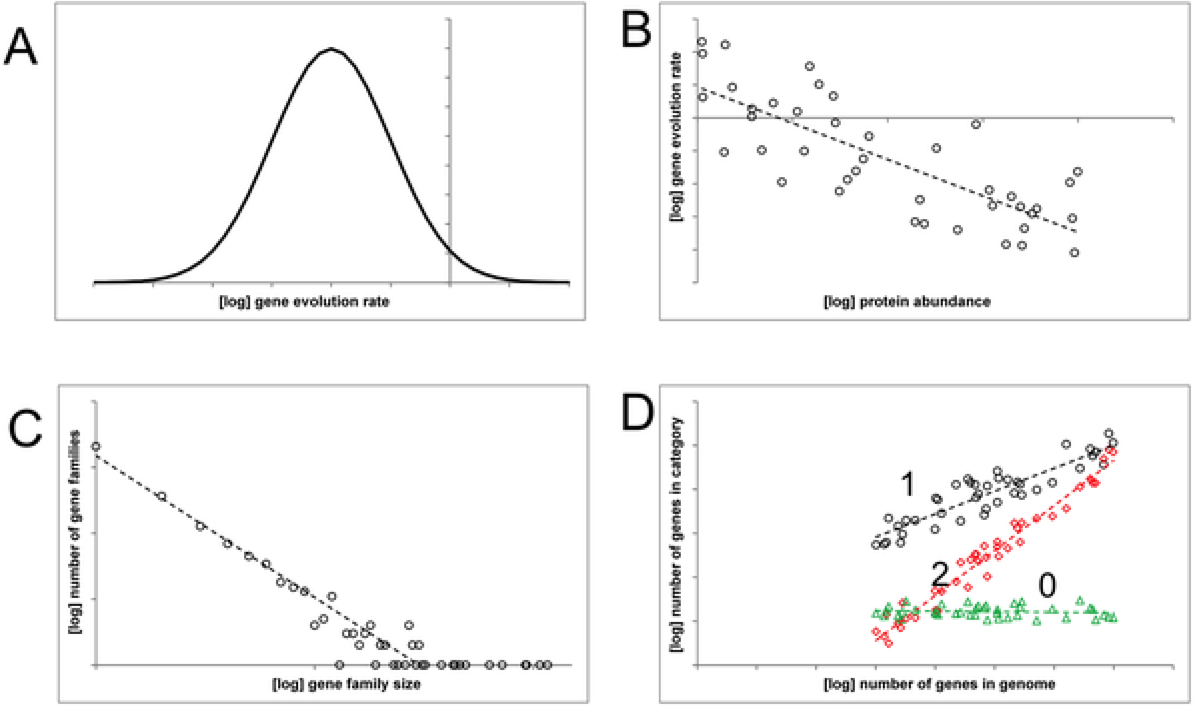
$$y_1 = \alpha + \beta_{12} y_2 + \beta_{13} y_3 + \beta_{123} y_2 y_3$$

… and non-linear models.



Are there laws of genome evolution?

(or Is biology more than stamp collecting?)



(A) Log-normal distribution of evolutionary rates of orthologous genes.

(B) Anticorrelation between gene expression level (protein abundance) and sequence evolution rate.

(C) Power law–like distribution of paralogous family size and out-degrees in networks.

(D) Differential scaling of functional classes of genes with the total number of genes in a genome: 0 – no dependence, typical of translation system component; 1 – linear dependence, characteristic of metabolic enzymes; 2 – quadratic dependence, characteristic of regulatory and signal transduction system components.

Koonin EV (2011) Are There Laws of Genome Evolution? PLoS Comput Biol 7(8): e1002173.

Some freely available tools

- R contains packages for most methods discussed here
- Machine learning: RapidMiner
- Networks: Cytoscape