Matlab exercise #1: “Wheel of Fortune”

- Each group gets a pair of genes that are known to be correlated.
- Each group also gets a random pair of genes selected by the “Wheel of Fortune”. They may or may not be correlated
- Download (log-transformed) expression_table.mat
- Run command fitlm(x,y) on assigned and random pairs
- Record β_0, β_1, R^2, P-value of the slope β_1 and write them on the blackboard
- Validate Matlab result for R^2 using your own calculations
- Look up gene names (see gene_description in your workspace) and write down a brief description of biological functions of genes. Does their correlation make biological sense?

Correlated pairs

plausible biological connection based

on short description

1, 6 g1=1994; g2=188;

2, g1=2872; g2=1269;

3, g1=1321; g2=10;

4, g1= 886; g2=819;

5, g1=2138; g2=1364;

no obvious biological common function

```
g1=1+floor(rand.*3000); g2=1+floor(rand.*3000);
disp([g1, g2])
```

Random pairs

```
>> g1=floor(3000.*rand)+1; g2=floor(3000.*rand)+1;
disp([g1,g2]);
```

```
>> g1=floor(3000.*rand)+1; g2=floor(3000.*rand)+1;
disp([g1,g2]);
```

```
>> g1=floor(3000.*rand)+1; g2=floor(3000.*rand)+1;
disp([g1,g2]);
```

```
>> g1=floor(3000.*rand)+1; g2=floor(3000.*rand)+1;
disp([g1,g2]);
```

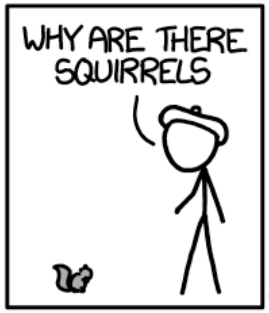
Matlab code

- load expression_table.mat
- g1=2907; g2=288;
- x=exp_t(g1,:)'; y=exp_t(g2,:)';
- figure; plot(x,y,'ko');
- lm=fitlm(x,y)
- y_fit=lm.Fitted;
- hold on; plot(x,lm.Fitted,'r-');
- SST=sum((y-mean(y)).^2);
- SSR=sum((y_fit-mean(y)).^2);
- SSE=sum((y-y_fit).^2);
- R2=SSR./SST
- disp([gene_names(g1), gene_names(g2)]);
- disp(gene_description(g1)); disp(gene_description (g2));

Credit: XKCD comics

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT



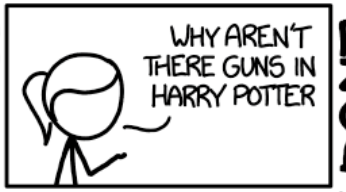
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE LAVA
WHY ARE THERE FEMALE MR MIMES
WHY IS GPS FREE
WHY IS LIFE SO BORING
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY IS YKK ON ALL ZIPPERS
WHY IS THERE HELL IF GOD FORGIVES

Matlab exercise on #2 on MLR

- Every group works with
  g0=2907; g1=1527; g2=2629; g3=2881;
  g4=1144; g5=1066;
- Compute Multiple Linear Regression (MLR):
  where
  y=exp_t (g0); x1= exp_t (g1); x2= exp_t (g2);
- How much better the MLR did compared to the Single Linear Regression (SLR)?
- Continue increasing the number of genes in x until R_adj starts to decrease

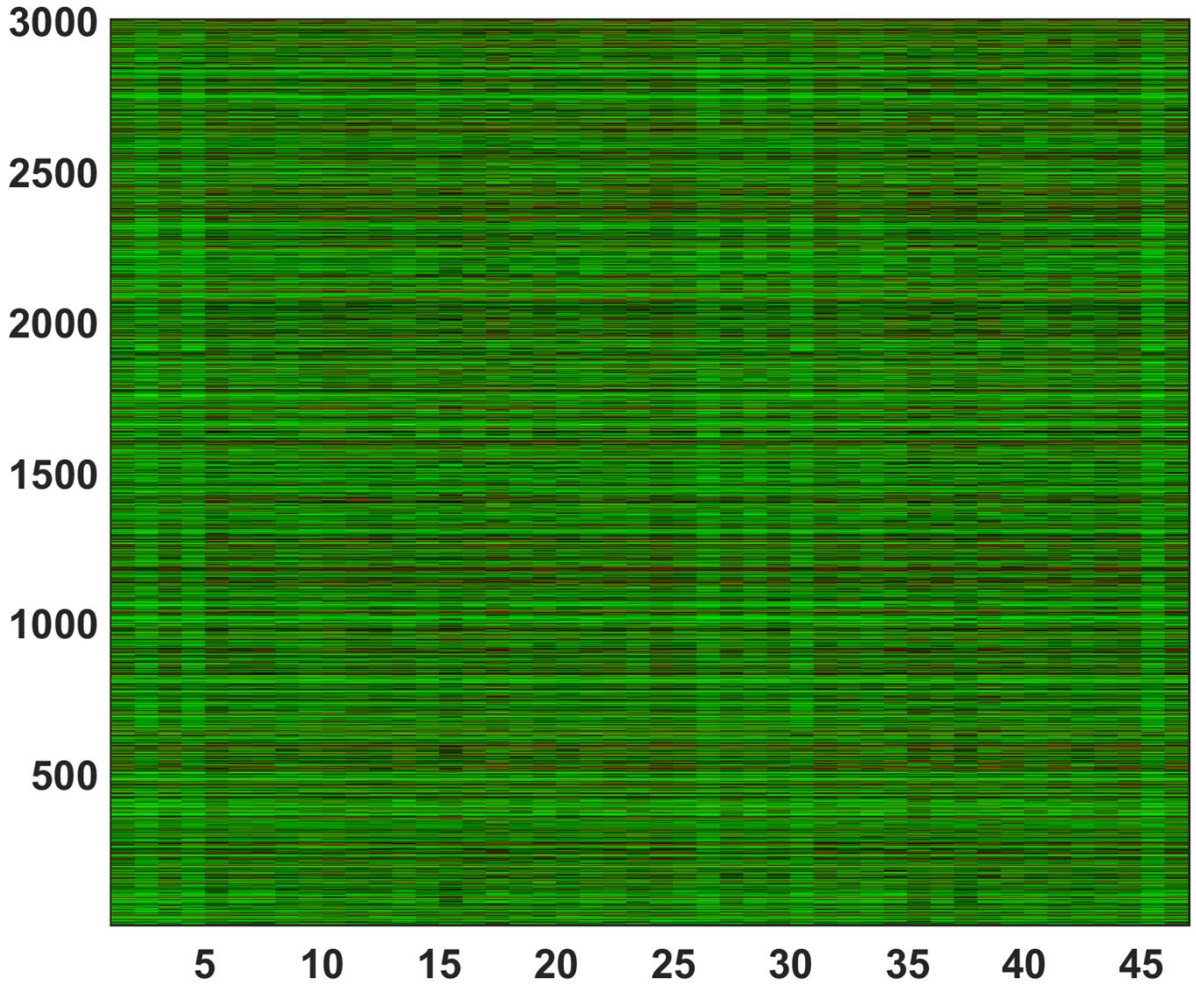
How I did it

```
g0=2907; g1=1527; g2=2629; g3=2881;g4=1144; g5=1066;
y=exp_t(g0,:)';
%% first use one x to predict y
x=exp_t(g1,:)';
figure; plot(x,y,'ko')
lm=fitlm(x,y)
y_fit=lm.Fitted;
hold on;
plot(x,lm.Fitted,'r-');
%% now use 2 x's to predict y
x=[exp_t(g1,:)', exp_t(g2,:)'];
lm2=fitlm(x,y)
y_fit=lm2.Fitted;
hold on; plot(x(:,1),y_fit,'gd');
%% now use m x's to predict y
corr_matrix=corr(exp_t');
g0=2907;
[u v]=sort(corr_matrix(g0,:),'descend');
x=[exp_t(v(2:m+1),:)'];
lm3=fitlm(x,y)
y_fit=lm3.Fitted;
plot(x(:,1),y_fit,'s');
```

Clustering analysis of gene expression data

Chapter 11 in
Jonathan Pevsner,
Bioinformatics and Functional Genomics,
3rd edition
(Chapter 9 in 2nd edition)

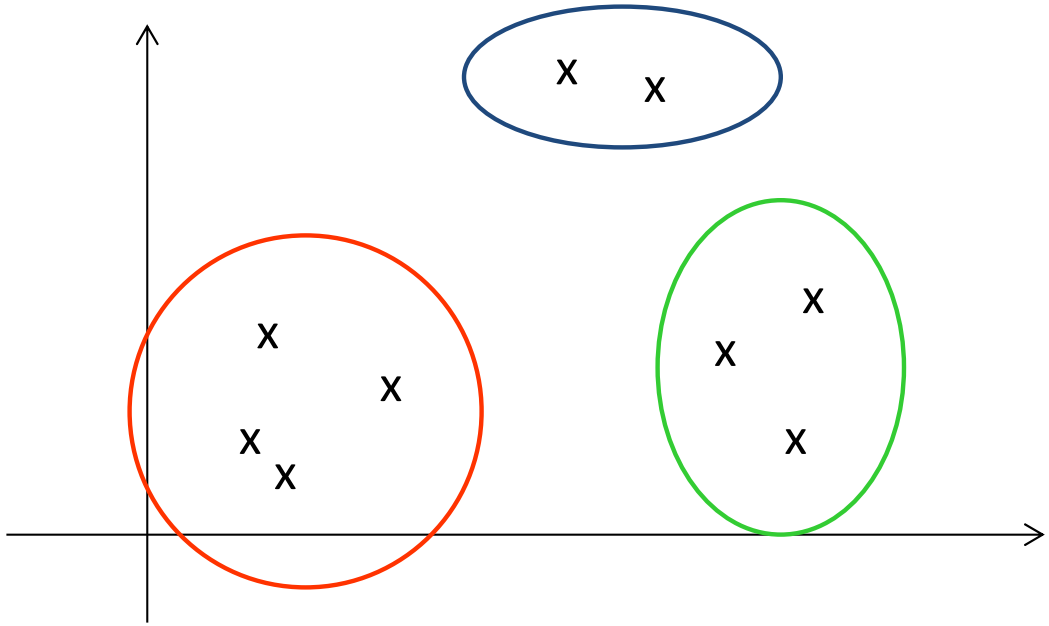
How to find the entire groups of mutually correlated genes if you have many genes and many samples?



Clustering to the rescue!

What is clustering?

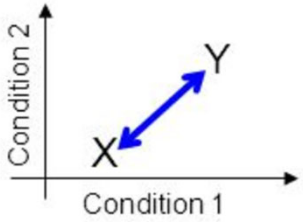
- The goal of clustering is to
  - group data points that are close (or **similar**) to each other
  - Usually, one needs to identify such groups (or clusters) in an **unsupervised** manner
  - Sometimes one takes into account **prior information** (Bayesian methods)
- Need to define some distance d_{ij} between objects i and j
- Clustering is easy in 2 dimensions but hard in 3000 dimensions -> need to somehow reduce dimensionality



How to define the distance?

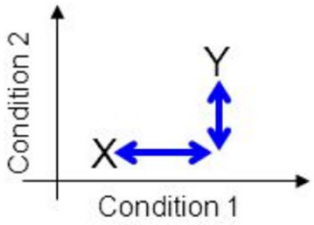
- Euclidean distance:
  - Most commonly used distance
  - Sphere shaped cluster
  - Corresponds to the geometric distance into the multidimensional space

$$d(X,Y) = \sqrt{\sum_i (x_i - y_i)^2}$$



- City Block (Manhattan) distance:
  - Sum of differences across dimensions
  - Less sensitive to outliers
  - Diamond shaped clusters

$$d(X,Y) = \sum_i |x_i - y_i|$$



The Canberra distance metric is calculated in R by

$$\sum \left(\frac{|x_i - y_i|}{|x_i + y_i|} \right).$$

Correlation coefficient distance

$$d(X,Y) = 1 - \rho(X,Y) = 1 - \frac{Cov(X,Y)}{\sqrt{(Var(X) \cdot Var(Y)}}$$

Common types of clustering algorithms

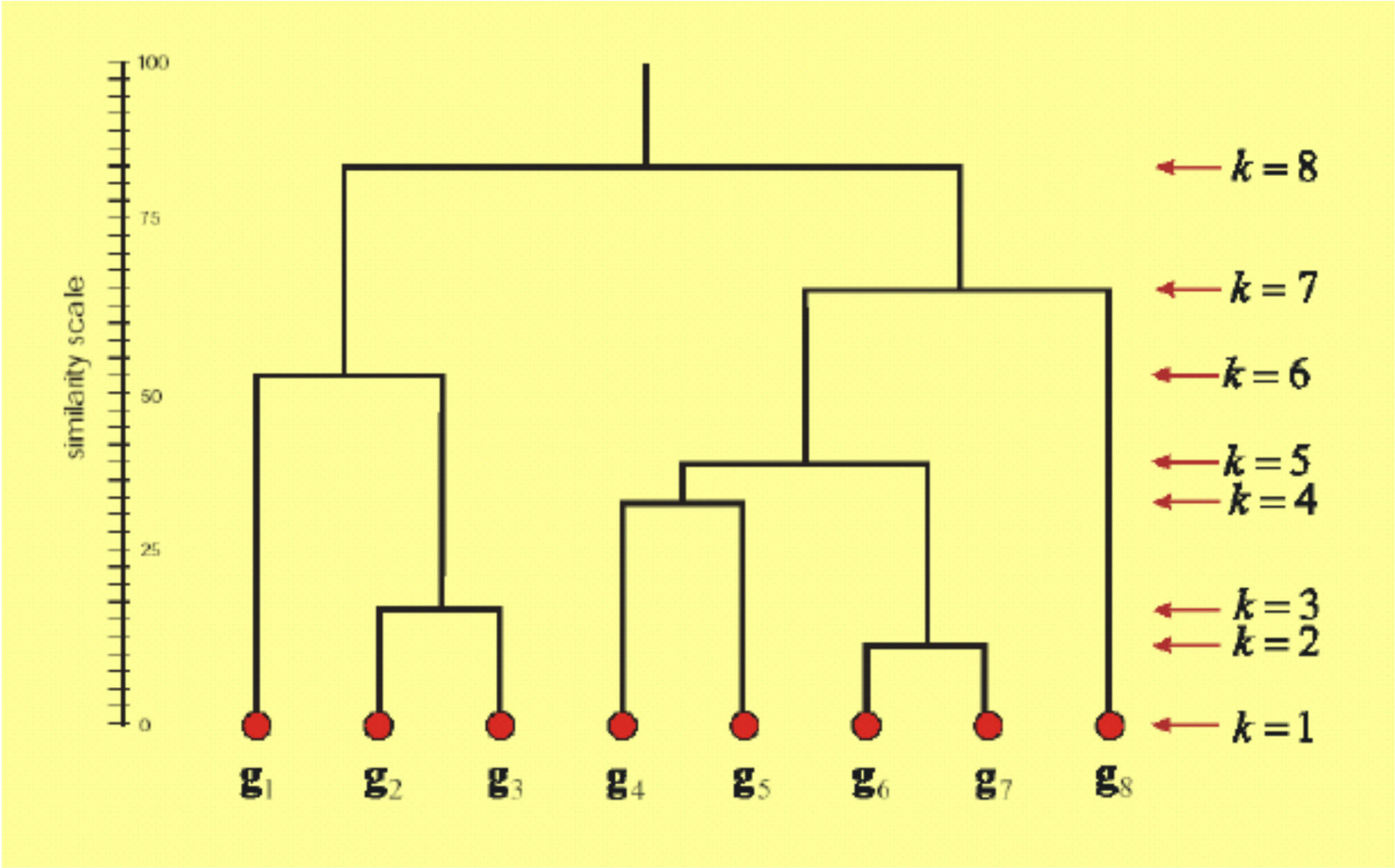
- Hierarchical if one doesn't know in advance the # of clusters
  - Agglomerative: start with N clusters and gradually merge them into 1 cluster
  - Divisive: start with 1 cluster and gradually break it up into N clusters
- Non-hierarchical algorithms
  - K-means clustering:
    - Iteratively apply the following two steps:
    - Calculate the centroid (center of mass) of each cluster
    - Assign each to the cluster to the nearest centroid
  - Principal Component Analysis (PCA)
    - plot pairs of top eigenvectors of the covariance matrix $Cov(X_i, X_j)$ and uses visual information to group

Hierarchical clustering

UPGMA algorithm

- Hierarchical agglomerative clustering algorithm
- **UPGMA** = Unweighted Pair Group Method with Arithmetic mean
- Iterative algorithm:
- Start with a pair with the smallest d(X,Y)
- Cluster these two together and replace it with their arithmetic mean (X+Y)/2
- Recalculate all distances to this new “cluster node”
- Repeat until all nodes are merged

Output of UPGMA algorithm



Clustering and network analysis of gene expression data

Chapter 11 in
Jonathan Pevsner,
Bioinformatics and Functional Genomics,
3rd edition
(Chapter 9 in 2nd edition)

Clustering in Matlab

Choices of distance metrics in clustergram(… 'RowPDistValue' …, 'ColumnPDistValue' …,)

Metric	Description
'euclidean'	Euclidean distance (default).
'seuclidean'	Standardized Euclidean distance. Each coordinate difference between rows in X is scaled by dividing by the corresponding element of the standard deviation S=nanstd(X). To specify another value for S, use D=pdist(X,'seuclidean',S).
'cityblock'	City block metric.
'minkowski'	Minkowski distance. The default exponent is 2. To specify a different exponent, use D = pdist(X,'minkowski',P), where P is a scalar positive value of the exponent.
'chebychev'	Chebychev distance (maximum coordinate difference).
'mahalanobis'	Mahalanobis distance, using the sample covariance of X as computed by nancov. To compute the distance with a different covariance, use D = pdist(X,'mahalanobis',C), where the matrix C is symmetric and positive definite.
'cosine'	One minus the cosine of the included angle between points (treated as vectors).
'correlation'	One minus the sample correlation between points (treated as sequences of values).
'spearman'	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).
'hamming'	Hamming distance, which is the percentage of coordinates that differ.
'jaccard'	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ.
custom distance function	A distance function specified using @: D = pdist(X,@distfun) A distance function must be of form d2 = distfun(XI,XJ) taking as arguments a 1-by-*n* vector XI, corresponding to a single row of X, and an *m*2-by-*n* matrix XJ, corresponding to multiple rows of X. distfun must accept a matrix XJ with an arbitrary number of rows. distfun must return an *m*2-by-1 vector of distances d2, whose *k*th element is the distance between XI and XJ(k,:).

Choices of hierarchical clustering algorithm in clustergram(…'linkage',…)

`X`	Matrix with two or more rows. The rows represent observations, the columns represent categories or dimensions.
`method`	Algorithm for computing distance between clusters.

Method	Description
`'average'`	Unweighted average distance (UPGMA)
`'centroid'`	Centroid distance (UPGMC), appropriate for Euclidean distances only
`'complete'`	Furthest distance
`'median'`	Weighted center of mass distance (WPGMC), appropriate for Euclidean distances only
`'single'`	Shortest distance
`'ward'`	Inner squared distance (minimum variance algorithm), appropriate for Euclidean distances only
`'weighted'`	Weighted average distance (WPGMA)

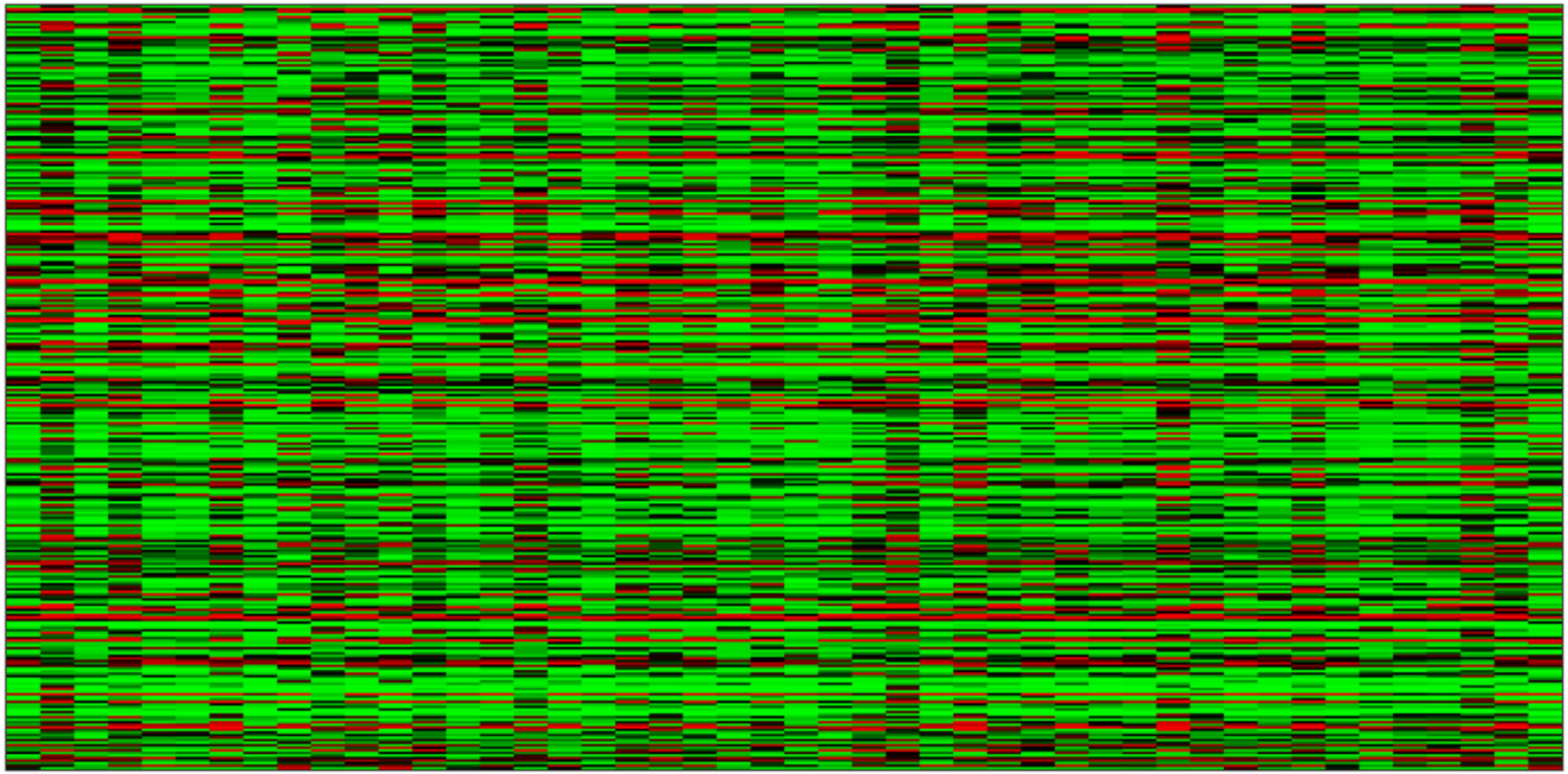
Default: `'single'`

Clustering group exercise

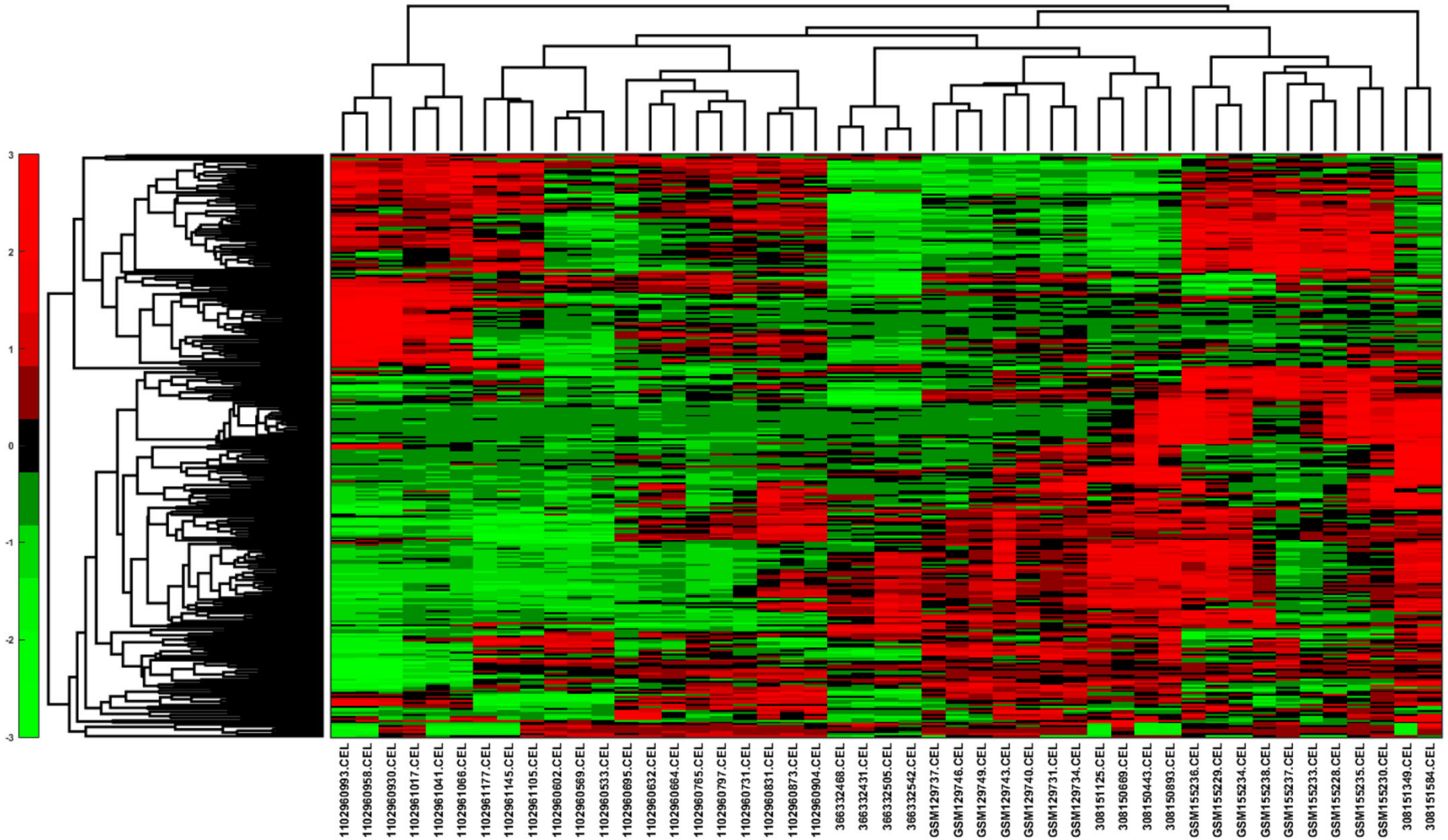
- Each group will analyze a cluster of genes identified in the T cell expression table
- Analyze the table of top 100 genes **by variance** in 47 samples
- Cluster them using:
  - Group 1: 'linkage', 'average', 'RowPDistValue', 'euclidean',
  - Group 2: 'linkage', 'average', 'RowPDistValue', 'cityblock',
  - Group 3: 'linkage', 'average', 'RowPDistValue', 'correlation',
  - Group 4: 'linkage', 'single', 'RowPDistValue', 'euclidean',
  - Group 5: 'linkage', 'single', 'RowPDistValue', 'cityblock',
  - Group 6: 'linkage', 'single', 'RowPDistValue', 'correlation',
- Use clustergram(…, 'Standardize','Row', 'linkage', ***as specified for your group,*** 'RowPDistValue' ***as specified for your group,*** 'RowLabels',gene_names1,'ColumnLabels', array_names)

```
load expression_table.mat
gene_variation=std(exp_t')';
[a,b]=sort(gene_variation,'descend');
ngenes=100;
exp_t1=exp_t(b(1:ngenes),:);
gene_names1=gene_names(b(1:ngenes));
%%% for group 1
CGobj1 = clustergram(exp_t1,
'Standardize','Row',...
'RowLabels',
gene_names1,'ColumnLabels',array_names)
set(CGobj1,'RowLabels',gene_names1,'ColumnLab
els',array_names,'linkage',
'average','RowPDist','euclidean');
set(CGobj1,'RowLabels',gene_names1,'ColumnLab
els',array_names,'linkage',
'average','RowPDist','correlation');
```

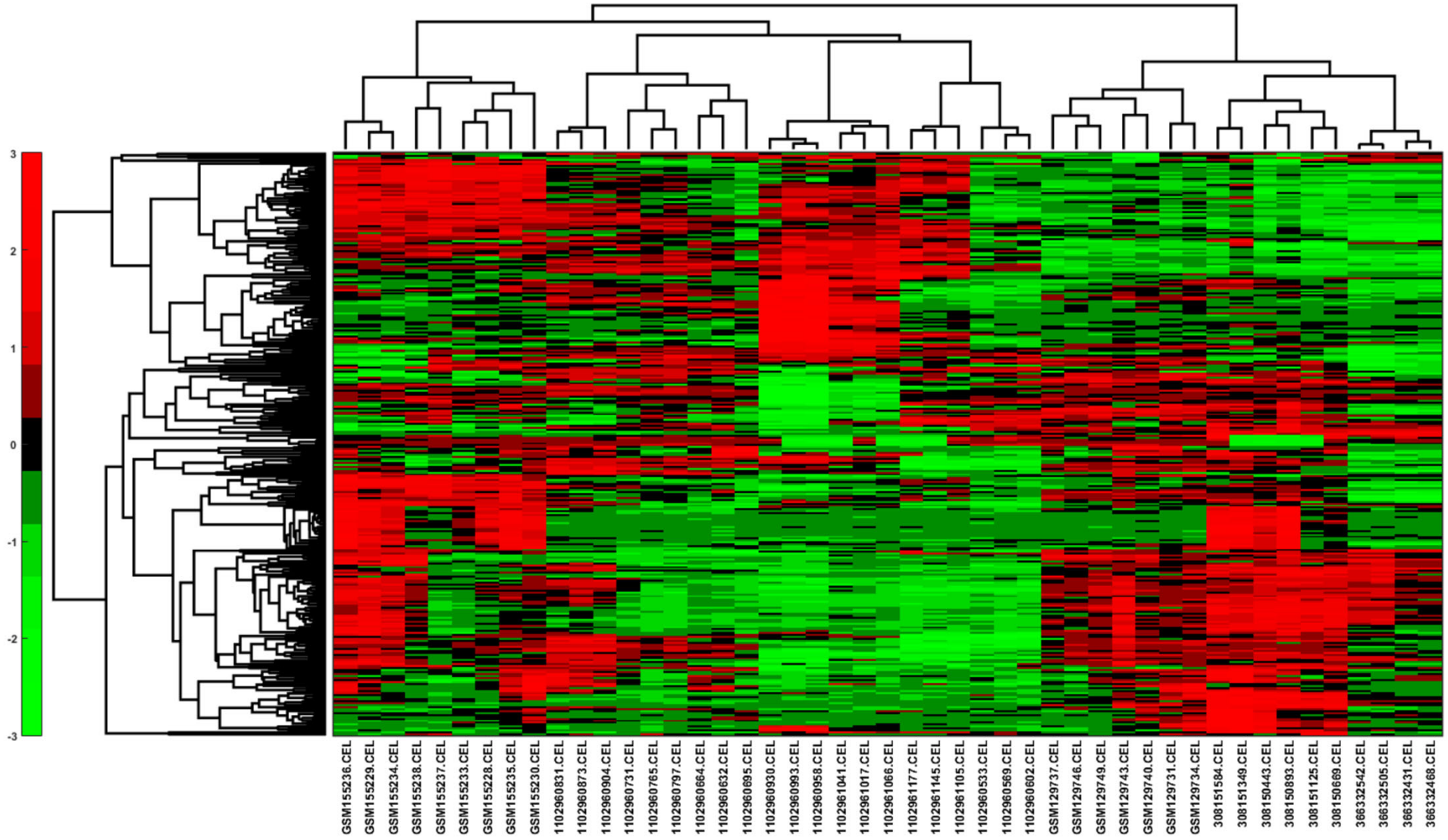
Before clustering



UPGMA hierarchical clustering, Euclidian distance



UPGMA hierarchical clustering, correlation distance



Search for shared biological functions

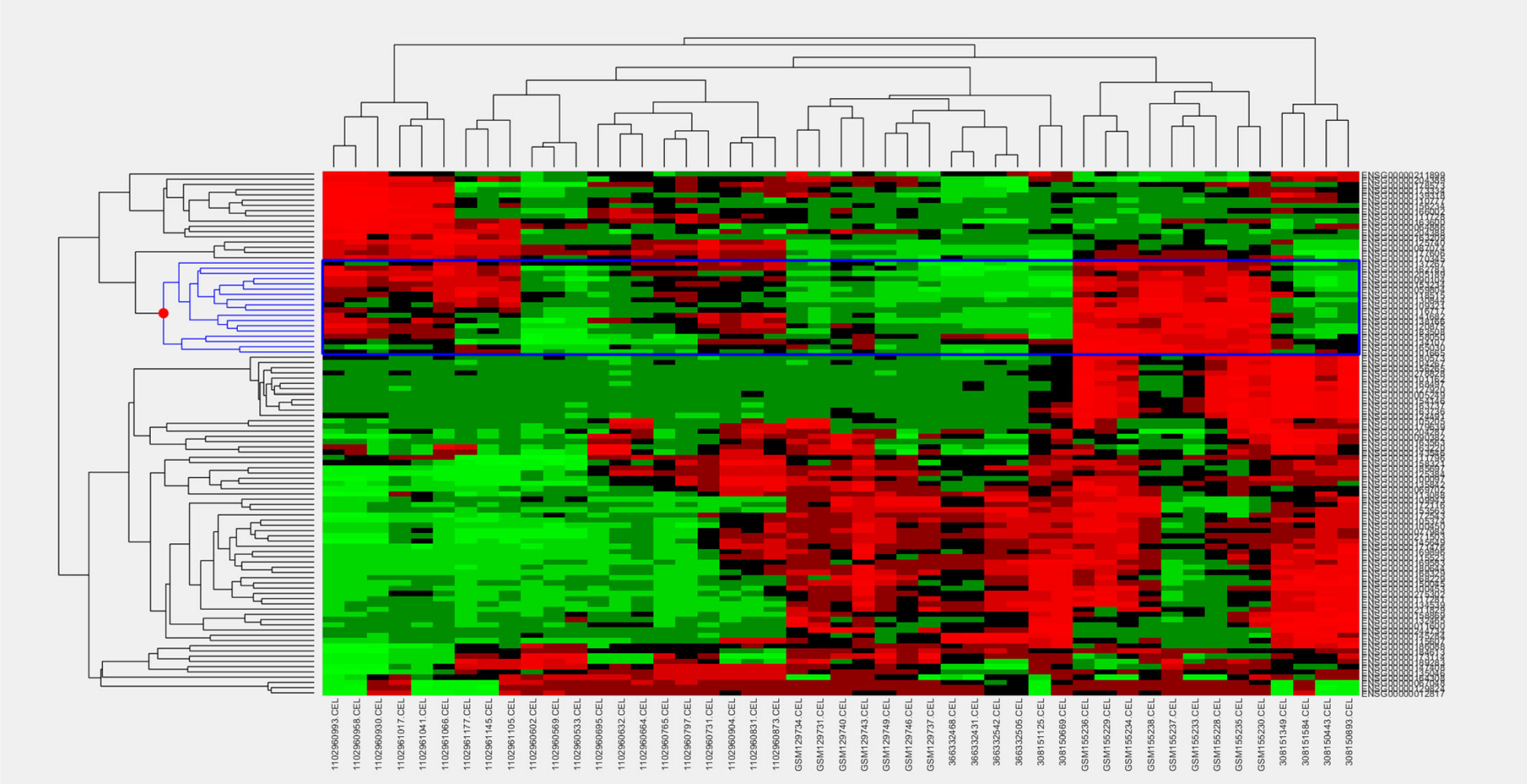
- copy the list of displayed genes
- go to "Start Analysis" on https://davidbioinformatics.nih.gov/summary_new.jsp
  Paste genes from gene list displayed by Matlab into the box in the left panel of the website
- select ENSEMBL_GENE_ID and "gene list" radio button
- Click "Functional Annotation Clustering"
- Select groups in "Annotation Summary Results" which have many genes from your list. Definitely select "PUBMED_ID" and interaction databases like "Biogrid"
- First look at "Functional Annotation Chart" rectangular button below to display all overrepresented terms. Sort by "Benjamini" correction for multiple hypotheses testing
- Select "Functional Annotation Clustering" rectangular button below to display annotation results for gene list broken into multiple groups (clusters) each with related biological functions
- Write down the # of genes in the cluster and the top functions in two most interesting clusters

```
%%%
%Which biological functions are
overrepresented in different clusters?
%1) Pick a cluster:
%2) Select a node on the tree of rows,
%3) Right click
%4) Choose “export group info” into
the workspace
%5) Name it gene_list
%Run the following two Matlab
commands to display genes
g1=gene_list.RowNodeNames;
for m=1:length(g1);
disp(g1{m});
end;
```

% Go to https://davidbioinformatics.nih.gov/summary_new.jsp
% select ENSEMBL_GENE_ID and “gene list” radio button
% Click "Functional Annotation Clustering“
% Select groups in “Annotation Summary Results”
% which have many genes from your list.
% Definitely select interaction databases such as “Biogrid”
% First look at "Functional Annotation Chart" rectangular button below
% to display all overrepresented terms.
% Sort by “Benjamini” correction for multiple hypotheses testing
% Select "Functional Annotation Clustering" rectangular button below
% to display annotation results for gene list broken into multiple groups
% (clusters) each with related biological functions
% Write down the # of genes in the cluster and the top functions
% in two most interesting clusters

Using options:

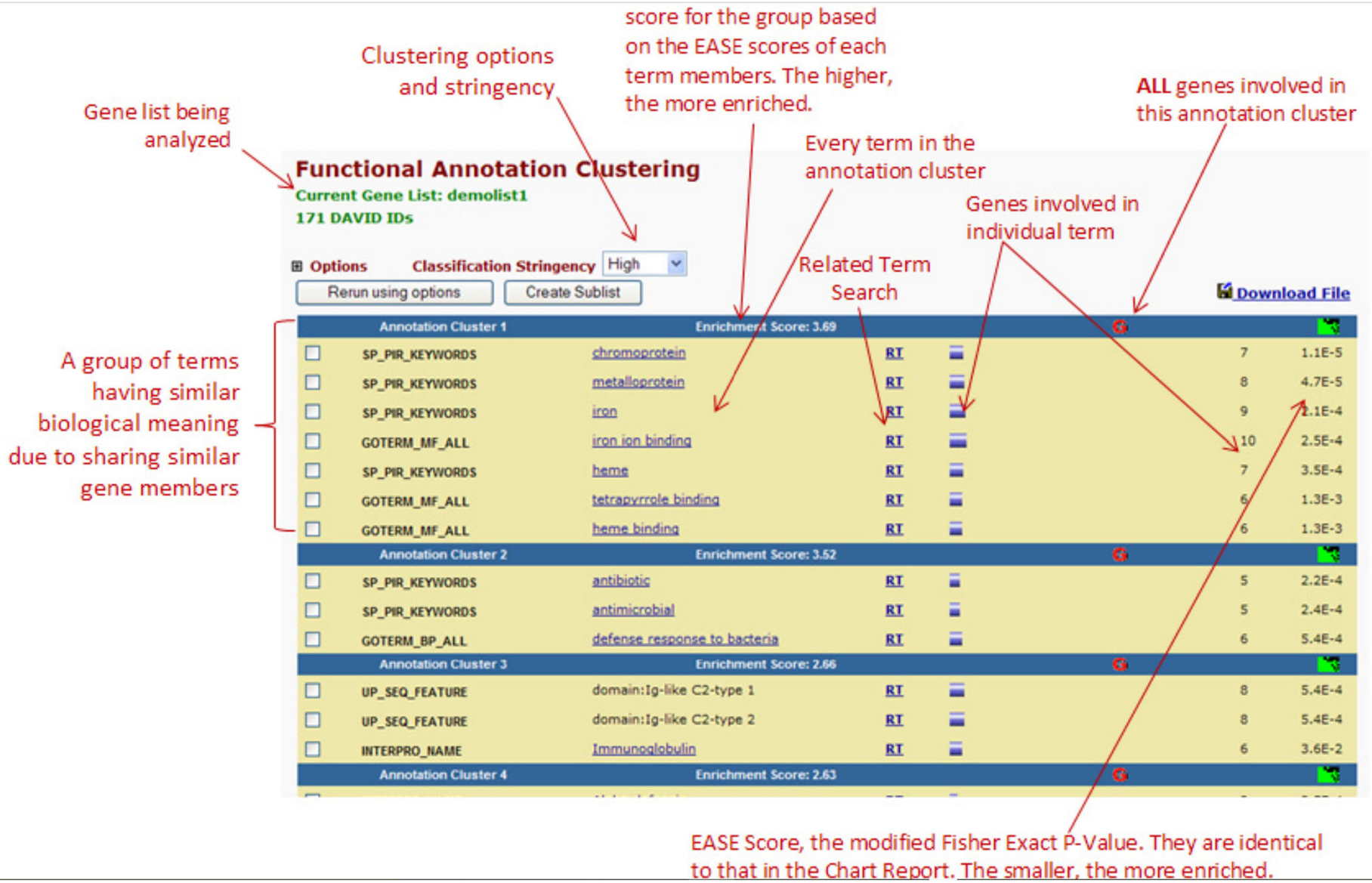
‘linkage’, ‘average’, ‘RowPDistValue’, ’euclidean’,



54 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_CC_DIRECT	nucleus	RT		16	88.9	8.1E-7	3.7E-5
	PIR_SUPERFAMILY	dual specificity protein phosphatase (MAP kinase phosphatase)	RT		3	16.7	4.0E-5	8.0E-5
	GOTERM_MF_DIRECT	protein tyrosine/threonine phosphatase activity	RT		3	16.7	3.4E-5	1.3E-3
	GOTERM_MF_DIRECT	MAP kinase tyrosine phosphatase activity	RT		3	16.7	3.4E-5	1.3E-3
	GOTERM_MF_DIRECT	MAP kinase tyrosine/serine/threonine phosphatase activity	RT		3	16.7	5.9E-5	1.5E-3
	INTERPRO	Mitogen-activated protein (MAP) kinase phosphatase	RT		3	16.7	3.3E-5	1.9E-3
	SMART	RHOD	RT		3	16.7	2.5E-4	4.8E-3
	INTERPRO	Rhodanese-like domain	RT		3	16.7	2.2E-4	6.2E-3
	SMART	DSPc	RT		3	16.7	8.4E-4	8.0E-3
	INTERPRO	Dual specificity phosphatase, catalytic domain	RT		3	16.7	6.0E-4	9.2E-3
	INTERPRO	Dual specificity phosphatase, subgroup, catalytic domain	RT		3	16.7	6.6E-4	9.2E-3
	GOTERM_BP_DIRECT	endoderm formation	RT		3	16.7	5.6E-5	1.1E-2
	UP_KW_CELLULAR_COMPONENT	Nucleus	RT		13	72.2	1.5E-3	1.3E-2
	SMART	PTPc_motif	RT		3	16.7	2.3E-3	1.5E-2
	GOTERM_MF_DIRECT	phosphoprotein phosphatase activity	RT		3	16.7	8.0E-4	1.5E-2
	INTERPRO	Protein-tyrosine phosphatase, catalytic	RT		3	16.7	1.4E-3	1.6E-2
	UP_KW_PTM	Ubl conjugation	RT		7	38.9	4.5E-3	1.9E-2
	UP_KW_PTM	Isopeptide bond	RT		6	33.3	5.4E-3	1.9E-2
	INTERPRO	Protein-tyrosine phosphatase, active site	RT		3	16.7	2.1E-3	2.0E-2
	INTERPRO	Protein-tyrosine/Dual specificity phosphatase	RT		3	16.7	2.8E-3	2.3E-2
	UP_SEQ_FEATURE	DOMAIN:Rhodanese	RT		3	16.7	1.9E-4	2.4E-2
	KEGG_PATHWAY	MAPK signaling pathway	RT		5	27.8	5.9E-4	2.8E-2
	GOTERM_MF_DIRECT	myosin phosphatase activity	RT		3	16.7	2.4E-3	3.6E-2
	GOTERM_MF_DIRECT	protein tyrosine phosphatase activity	RT		3	16.7	4.2E-3	5.3E-2
	GOTERM_CC_DIRECT	nucleoplasm	RT		10	55.6	2.3E-3	5.4E-2
	GOTERM_BP_DIRECT	negative regulation of MAPK cascade	RT		3	16.7	7.0E-4	6.8E-2


Gene list being analyzed
Clustering options and stringency
score for the group based on the EASE scores of each term members. The higher, the more enriched.
ALL genes involved in this annotation cluster
Every term in the annotation cluster
Genes involved in individual term
Related Term Search
A group of terms having similar biological meaning due to sharing similar gene members
Functional Annotation Clustering
Current Gene List: demolist1
171 DAVID IDs
Options
Classification Stringency
High
Rerun using options
Create Sublist
Download File
Annotation Cluster 1 Enrichment Score: 3.69
SP_PIR_KEYWORDS chromoprotein RT 7 1.1E-5
SP_PIR_KEYWORDS metalloprotein RT 8 4.7E-5
SP_PIR_KEYWORDS iron RT 9 2.1E-4
GOTERM_MF_ALL iron ion binding RT 10 2.5E-4
SP_PIR_KEYWORDS heme RT 7 3.5E-4
GOTERM_MF_ALL tetrapyrrole binding RT 6 1.3E-3
GOTERM_MF_ALL heme binding RT 6 1.3E-3
Annotation Cluster 2 Enrichment Score: 3.52
SP_PIR_KEYWORDS antibiotic RT 5 2.2E-4
SP_PIR_KEYWORDS antimicrobial RT 5 2.4E-4
GOTERM_BP_ALL defense response to bacteria RT 6 5.4E-4
Annotation Cluster 3 Enrichment Score: 2.66
UP_SEQ_FEATURE domain:Ig-like C2-type 1 RT 8 5.4E-4
UP_SEQ_FEATURE domain:Ig-like C2-type 2 RT 8 5.4E-4
INTERPRO_NAME Immunoglobulin RT 6 3.6E-2
Annotation Cluster 4 Enrichment Score: 2.63
EASE Score, the modified Fisher Exact P-Value. They are identical to that in the Chart Report. The smaller, the more enriched.

Functional Annotation Clustering

Help and Manual

Current Gene List: List_3
Current Background: Homo sapiens
18 DAVID IDs

Options **Classification Stringency** Medium

Rerun using options | Create Sublist

25 Cluster(s)

Download File

Annotation Cluster 1	Enrichment Score: 5.2			Count	P_Value	Benjamini
DISGENET	Juvenile arthritis	RT		7	1.5E-8	4.7E-7
DISGENET	Juvenile psoriatic arthritis	RT		7	1.5E-8	4.7E-7
DISGENET	Polyarthritis, Juvenile, Rheumatoid Factor Negative	RT		7	1.5E-8	4.7E-7
DISGENET	Polyarthritis, Juvenile, Rheumatoid Factor Positive	RT		7	1.5E-8	4.7E-7
DISGENET	Juvenile-Onset Still Disease	RT		7	1.8E-8	4.7E-7
KEGG_PATHWAY	MAPK signaling pathway	RT		5	5.9E-4	2.8E-2
BIOGRID_INTERACTION	mitogen-activated protein kinase 1(MAPK1)	RT		4	3.8E-3	1.0E0
WIKIPATHWAYS	MAPK signaling pathway	RT		3	5.8E-2	6.9E-1
GAD_DISEASE_CLASS	UNKNOWN	RT		5	1.5E-1	9.9E-1
Annotation Cluster 2	**Enrichment Score: 2.83**			**Count**	**P_Value**	**Benjamini**
INTERPRO	Mitogen-activated protein (MAP) kinase phosphatase	RT		3	3.3E-5	1.9E-3
GOTERM_MF_DIRECT	protein tyrosine/threonine phosphatase activity	RT		3	3.4E-5	1.3E-3
GOTERM_MF_DIRECT	MAP kinase tyrosine phosphatase activity	RT		3	3.4E-5	1.3E-3
PIR_SUPERFAMILY	dual specificity protein phosphatase (MAP kinase phosphatase)	RT		3	4.0E-5	8.0E-5
GOTERM_BP_DIRECT	endoderm formation	RT		3	5.6E-5	1.1E-2
GOTERM_MF_DIRECT	MAP kinase tyrosine/serine/threonine phosphatase activity	RT		3	5.9E-5	1.5E-3
PUBMED_ID	27880917	RT		4	1.7E-4	2.5E-2
UP_SEQ_FEATURE	DOMAIN:Rhodanese	RT		3	1.9E-4	2.4E-2
INTERPRO	Rhodanese-like domain	RT		3	2.2E-4	6.2E-3
SMART	RHOD	RT		3	2.5E-4	4.8E-3

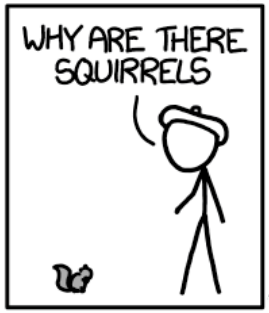
Annotation Cluster 3	Enrichment Score: 2.43	G		Count	P_Value	Benjamini
DISGENET	Arsenic Poisoning, Inorganic	RT		3	3.5E-3	4.6E-2
DISGENET	Nervous System, Organic Arsenic Poisoning	RT		3	3.5E-3	4.6E-2
DISGENET	Arsenic Poisoning	RT		3	3.5E-3	4.6E-2
DISGENET	Arsenic Encephalopathy	RT		3	3.5E-3	4.6E-2
DISGENET	Arsenic Induced Polyneuropathy	RT		3	3.5E-3	4.6E-2
DISGENET	Dermatologic disorders	RT		3	5.1E-3	5.6E-2
Annotation Cluster 4	**Enrichment Score: 2.26**	G		**Count**	**P_Value**	**Benjamini**
PUBMED_ID	19322201	RT		7	1.3E-8	5.9E-6
BIOGRID_INTERACTION	ELAV like RNA binding protein 1(ELAVL1)	RT		7	4.4E-3	1.0E0
UCSC_TFBS	CEBPA	RT		7	1.8E-1	1.0E0
UCSC_TFBS	CDPCR3HD	RT		7	6.5E-1	1.0E0
UCSC_TFBS	FOXD3	RT		5	7.4E-1	1.0E0
Annotation Cluster 5	**Enrichment Score: 2.14**	G		**Count**	**P_Value**	**Benjamini**
GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		6	1.4E-3	9.1E-2
BIOGRID_INTERACTION	retinoid X receptor alpha(RXRA)	RT		3	6.1E-3	1.0E0
GOTERM_MF_DIRECT	protein heterodimerization activity	RT		3	4.5E-2	3.7E-1
Annotation Cluster 6	**Enrichment Score: 1.95**	G		**Count**	**P_Value**	**Benjamini**
REACTOME_PATHWAY	Generic Transcription Pathway	RT		7	2.8E-3	1.7E-1
REACTOME_PATHWAY	RNA Polymerase II Transcription	RT		7	4.6E-3	1.7E-1
REACTOME_PATHWAY	Gene expression (Transcription)	RT		7	8.2E-3	2.0E-1
GAD_DISEASE_CLASS	UNKNOWN	RT		5	1.5E-1	9.9E-1
Annotation Cluster 7	**Enrichment Score: 1.76**	G		**Count**	**P_Value**	**Benjamini**
PUBMED_ID	18029348	RT		6	1.8E-5	3.4E-3
UP_KW_PTM	Isopeptide bond	RT		6	5.4E-3	1.9E-2
PUBMED_ID	15342556	RT		3	7.9E-3	4.8E-1
PUBMED_ID	26496610	RT		3	1.0E-1	1.0E0
GOTERM_MF_DIRECT	metal ion binding	RT		4	4.5E-1	1.0E0
UCSC_TFBS	TAL1ALPHAE47	RT		3	7.9E-1	1.0E0

Annotation Cluster 3	Enrichment Score: 2.43	G		Count	P_Value	Benjamini
DISGENET	Arsenic Poisoning, Inorganic	RT		3	3.5E-3	4.6E-2
DISGENET	Nervous System, Organic Arsenic Poisoning	RT		3	3.5E-3	4.6E-2
DISGENET	Arsenic Poisoning	RT		3	3.5E-3	4.6E-2
DISGENET	Arsenic Encephalopathy	RT		3	3.5E-3	4.6E-2
DISGENET	Arsenic Induced Polyneuropathy	RT		3	3.5E-3	4.6E-2
DISGENET	Dermatologic disorders	RT		3	5.1E-3	5.6E-2
Annotation Cluster 4	**Enrichment Score: 2.26**	G		**Count**	**P_Value**	**Benjamini**
PUBMED_ID	19322201	RT		7	1.3E-8	5.9E-6
BIOGRID_INTERACTION	ELAV like RNA binding protein 1(ELAVL1)	RT		7	4.4E-3	1.0E0
UCSC_TFBS	CEBPA	RT		7	1.8E-1	1.0E0
UCSC_TFBS	CDPCR3HD	RT		7	6.5E-1	1.0E0
UCSC_TFBS	FOXD3	RT		5	7.4E-1	1.0E0
Annotation Cluster 5	**Enrichment Score: 2.14**	G		**Count**	**P_Value**	**Benjamini**
GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		6	1.4E-3	9.1E-2
BIOGRID_INTERACTION	retinoid X receptor alpha(RXRA)	RT		3	6.1E-3	1.0E0
GOTERM_MF_DIRECT	protein heterodimerization activity	RT		3	4.5E-2	3.7E-1
Annotation Cluster 6	**Enrichment Score: 1.95**	G		**Count**	**P_Value**	**Benjamini**
REACTOME_PATHWAY	Generic Transcription Pathway	RT		7	2.8E-3	1.7E-1
REACTOME_PATHWAY	RNA Polymerase II Transcription	RT		7	4.6E-3	1.7E-1
REACTOME_PATHWAY	Gene expression (Transcription)	RT		7	8.2E-3	2.0E-1
GAD_DISEASE_CLASS	UNKNOWN	RT		5	1.5E-1	9.9E-1
Annotation Cluster 7	**Enrichment Score: 1.76**	G		**Count**	**P_Value**	**Benjamini**
PUBMED_ID	18029348	RT		6	1.8E-5	3.4E-3
UP_KW_PTM	Isopeptide bond	RT		6	5.4E-3	1.9E-2
PUBMED_ID	15342556	RT		3	7.9E-3	4.8E-1
PUBMED_ID	26496610	RT		3	1.0E-1	1.0E0
GOTERM_MF_DIRECT	metal ion binding	RT		4	4.5E-1	1.0E0
UCSC_TFBS	TAL1ALPHAE47	RT		3	7.9E-1	1.0E0

Credit: XKCD comics

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT



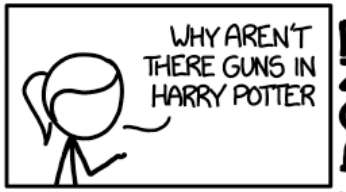
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE



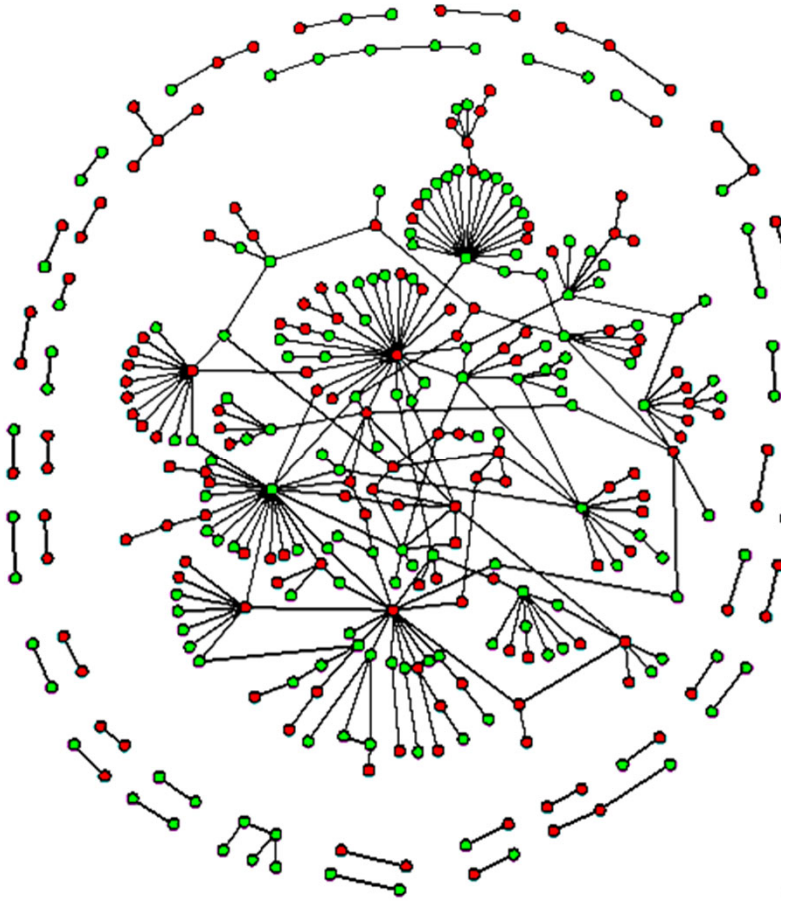
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

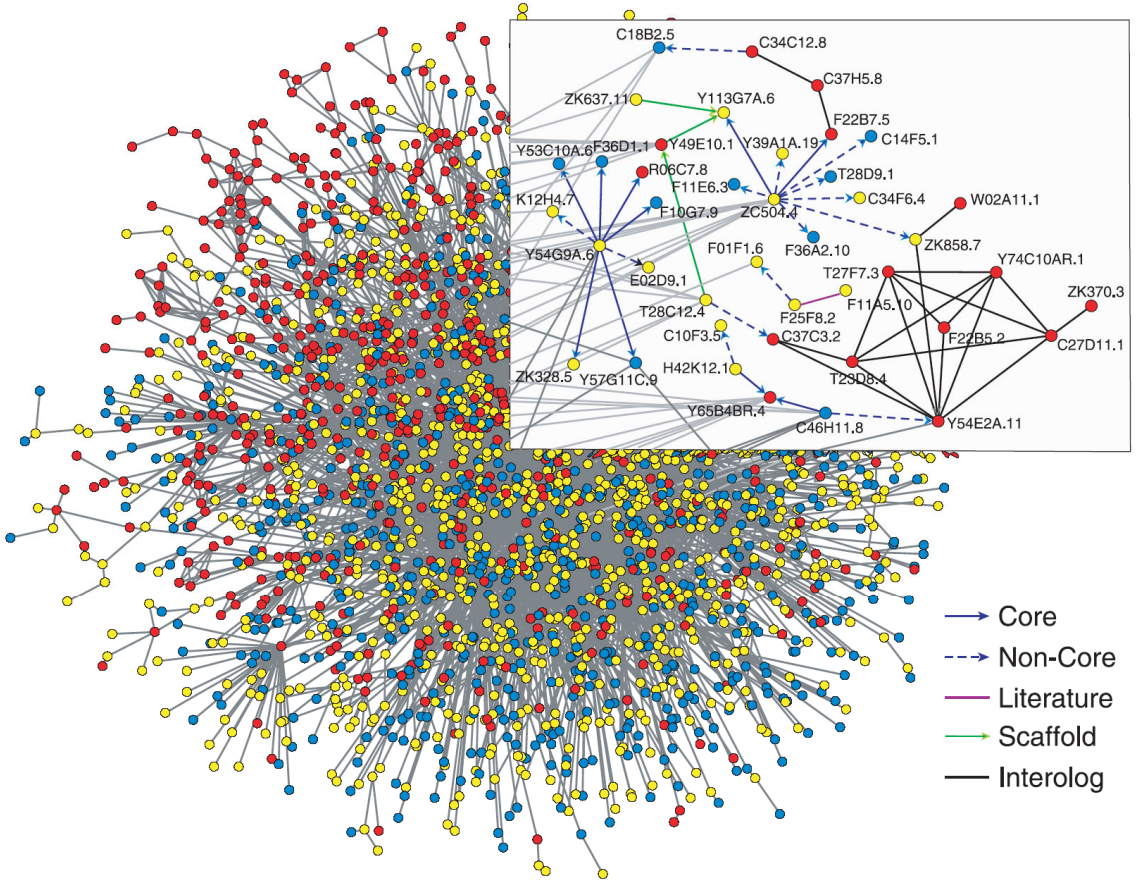
WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE LAVA
WHY ARE THERE FEMALE MR MIMES
WHY IS GPS FREE
WHY IS LIFE SO BORING
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY IS YKK ON ALL ZIPPERS
WHY IS THERE HELL IF GOD FORGIVES

Basic concepts of network analysis

Reminder from the first lecture

Protein-Protein binding
IntAct Database (Dec 2015)
Interactions: 577,297 Proteins: 89,716

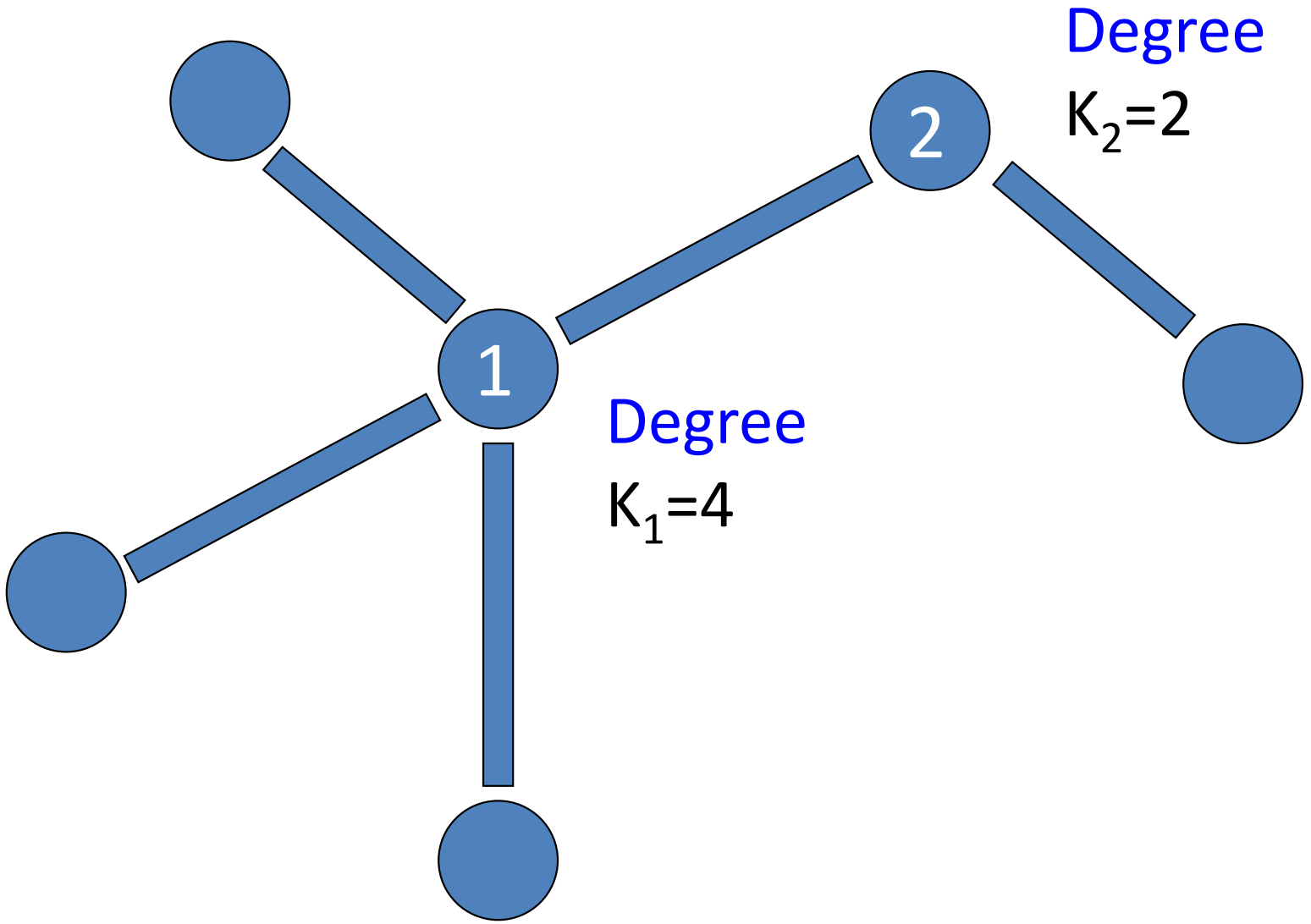




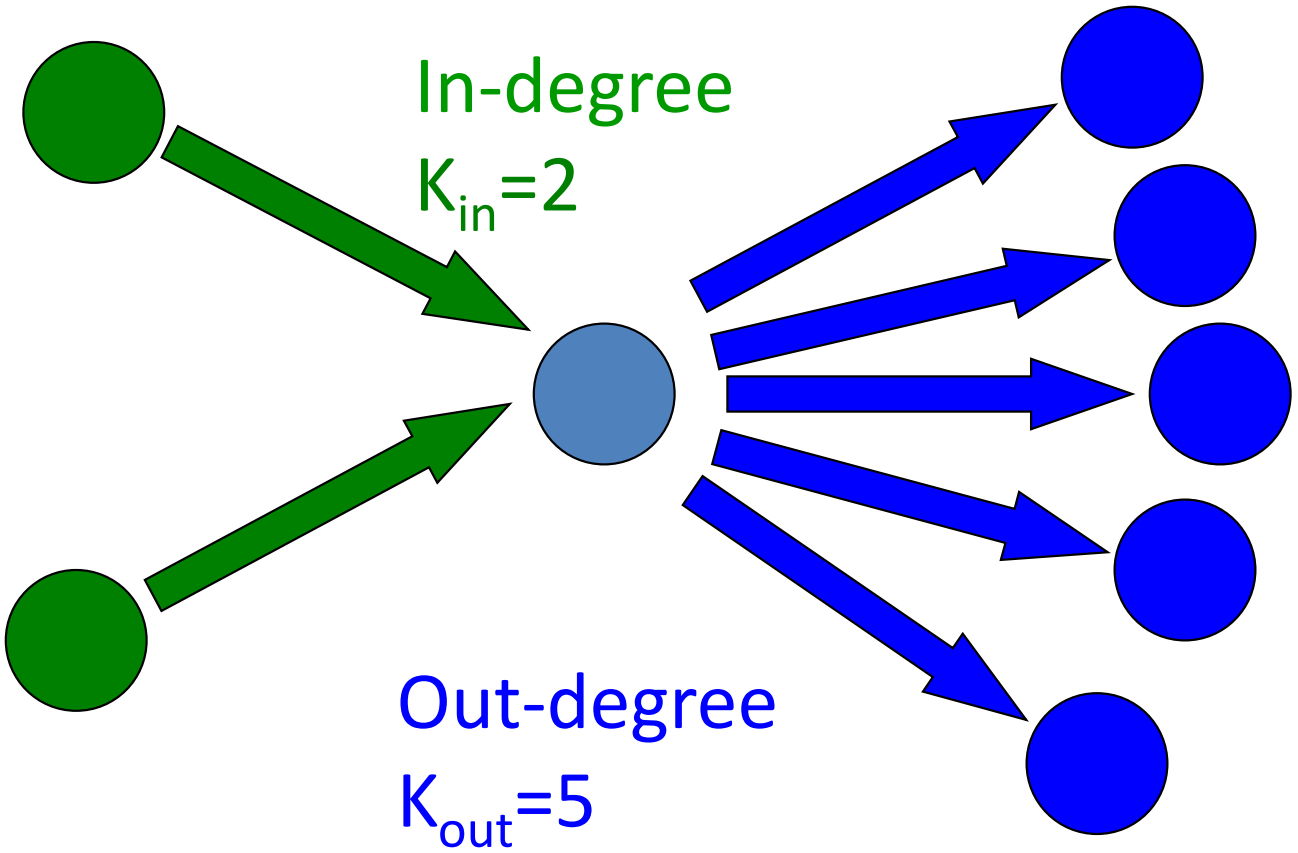
Baker's yeast *S. cerevisiae* (only nuclear proteins shown)
From S. Maslov, K. Sneppen, Science 2002

Worm *C. elegans*
From S. Lee et al , Science 2004

Degree of a node – its # of neighbors

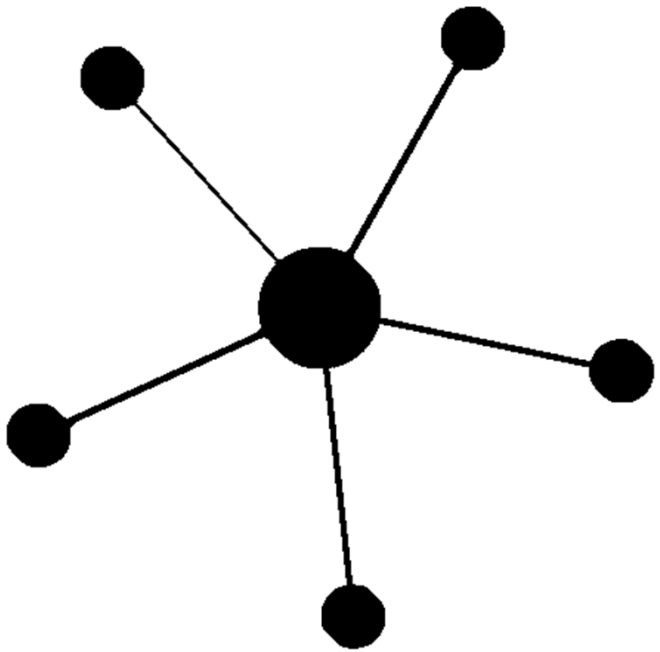


Directed networks have in- and out- degrees



How to find “important” nodes?

- By their degree
- Hubs = important
- Example: Google’s PageRank



How Google PageRank algorithm works?

- Google was solving the following problem in mid-1990s: too many websites match a typical search query: need to rank websites.
- Other popular search engines (e.g. Altavista) count the # of times a query word appears in website's text. Websites respond by putting lots of invisible words
- One could rank the importance of webpages by number of hyperlinks pointing to it (in-degree K_{in}) but:
  - Too democratic: It doesn't take into account the importance of webpages sending hyperlinks
  - it's easy to trick and artificially boost the rank
- Google's solution: simulate the behavior of many "random surfers" and then count the number of times they visited each webpage = it's PageRank
  - Popular pages send more surfers your way → the PageRank weight is proportional to K_{in} but weighted by popularity

PageRank algorithm is Google's $2.8T idea

- PageRank assigns to every webpage an importance score G_i
- The meaning of G_i – how often random surfers visit this website
- To determine solves a self-consistent Eq.:

  $G_i \sim \sum_j T_{ij} G_j$. Here
  $T_{ij} = A_{ij}/K_{out}$ (j) is the normalized adjacency matrix
- It finds the principal eigenvector (the one with the largest eigenvalue).

Problem with PageRank algorithm and how Google solved it

- Problem: surfers can be trapped in infinite loops with one or more entrances and no exits
- Model with random jumps mimicking surfers getting bored when following a chain of links

  $G_i \sim (1-\alpha)\sum_j T_{ij} G_j + \alpha \sum_j G_j$
- α=0.15 meaning that an average web surfer (circa 1995) on average jumped around $1/\alpha \approx 6$ webpages before going somewhere else