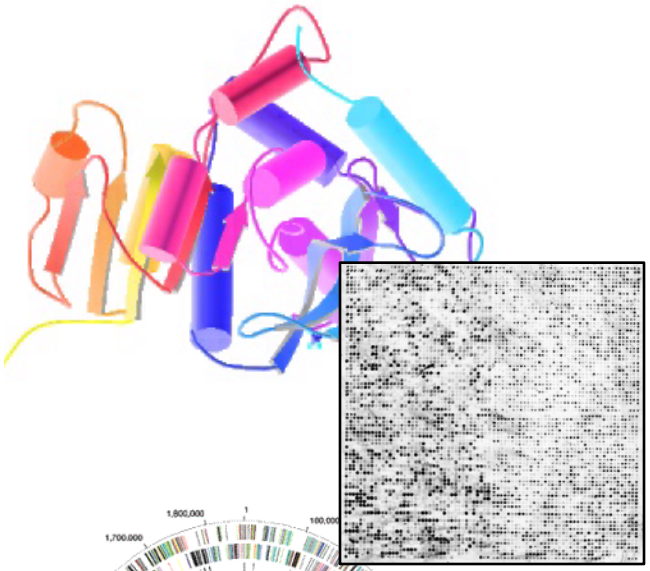
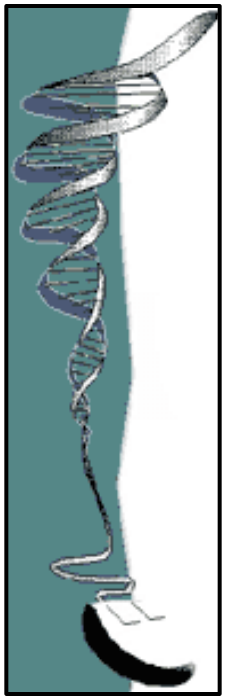
Biomed. Data Science:

Unsupervised Datamining



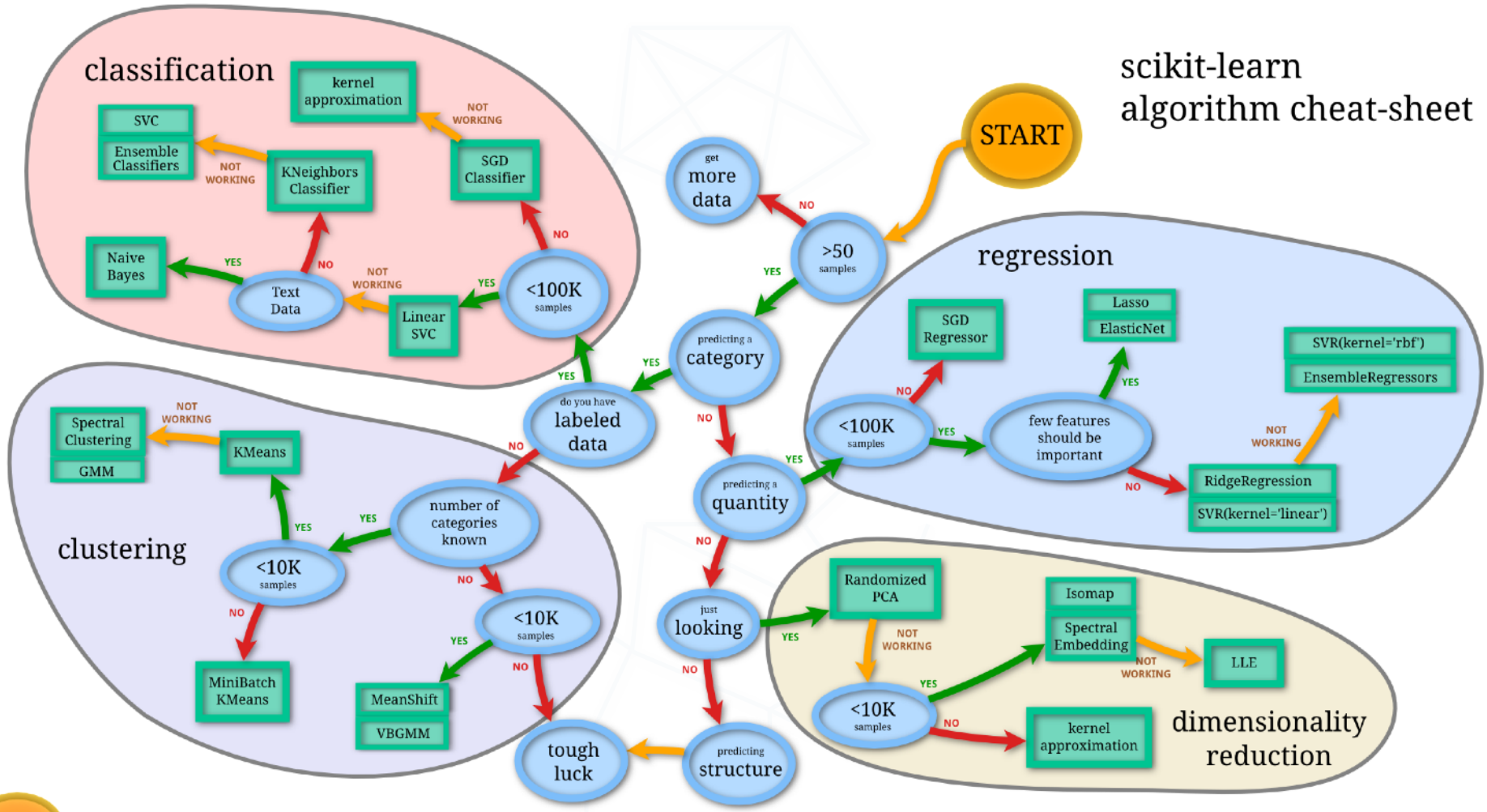


Mark Gerstein, Yale University

gersteinlab.org/courses/452

(last edit in spring ’19, pack #8)

The World of Machine Learning

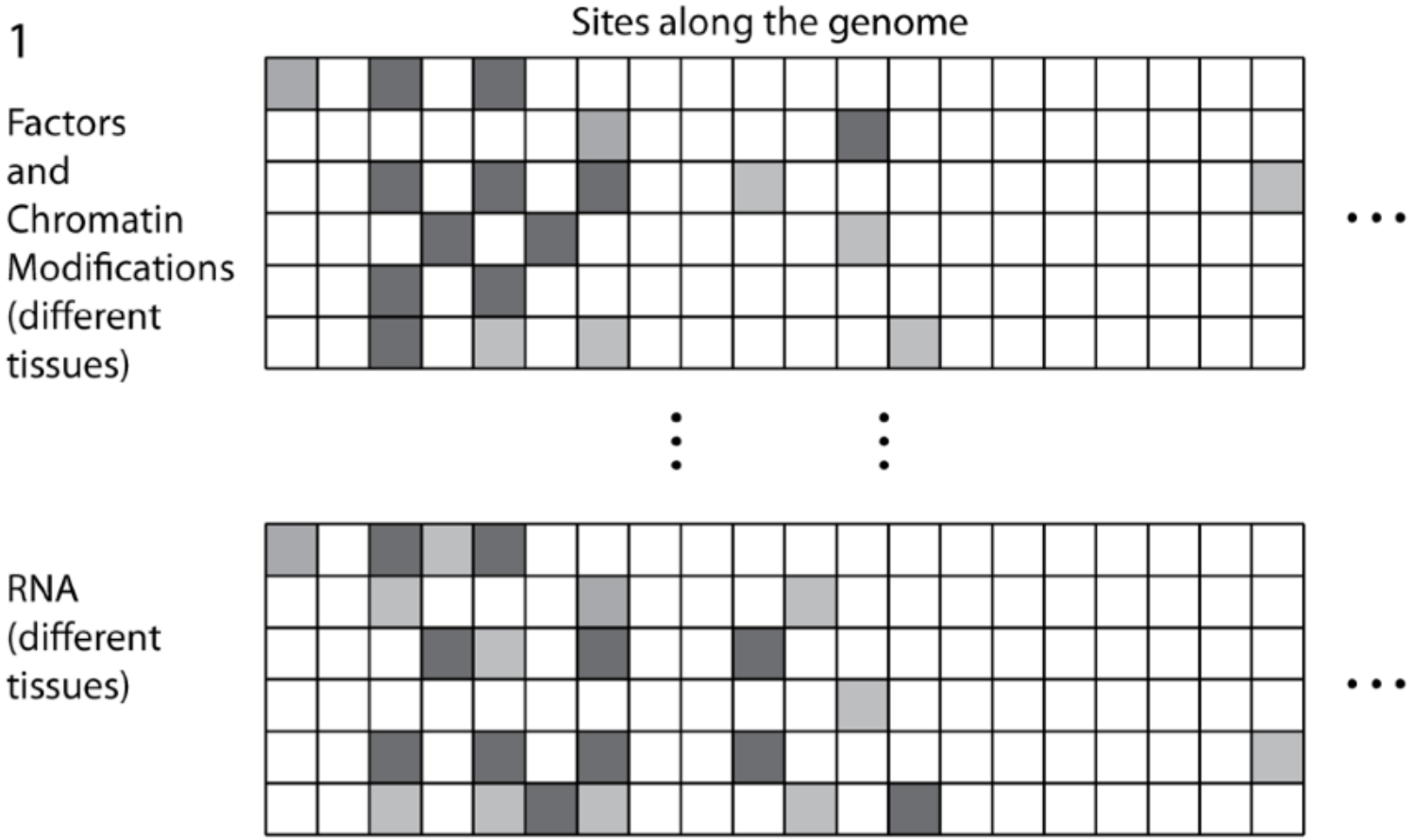


Back

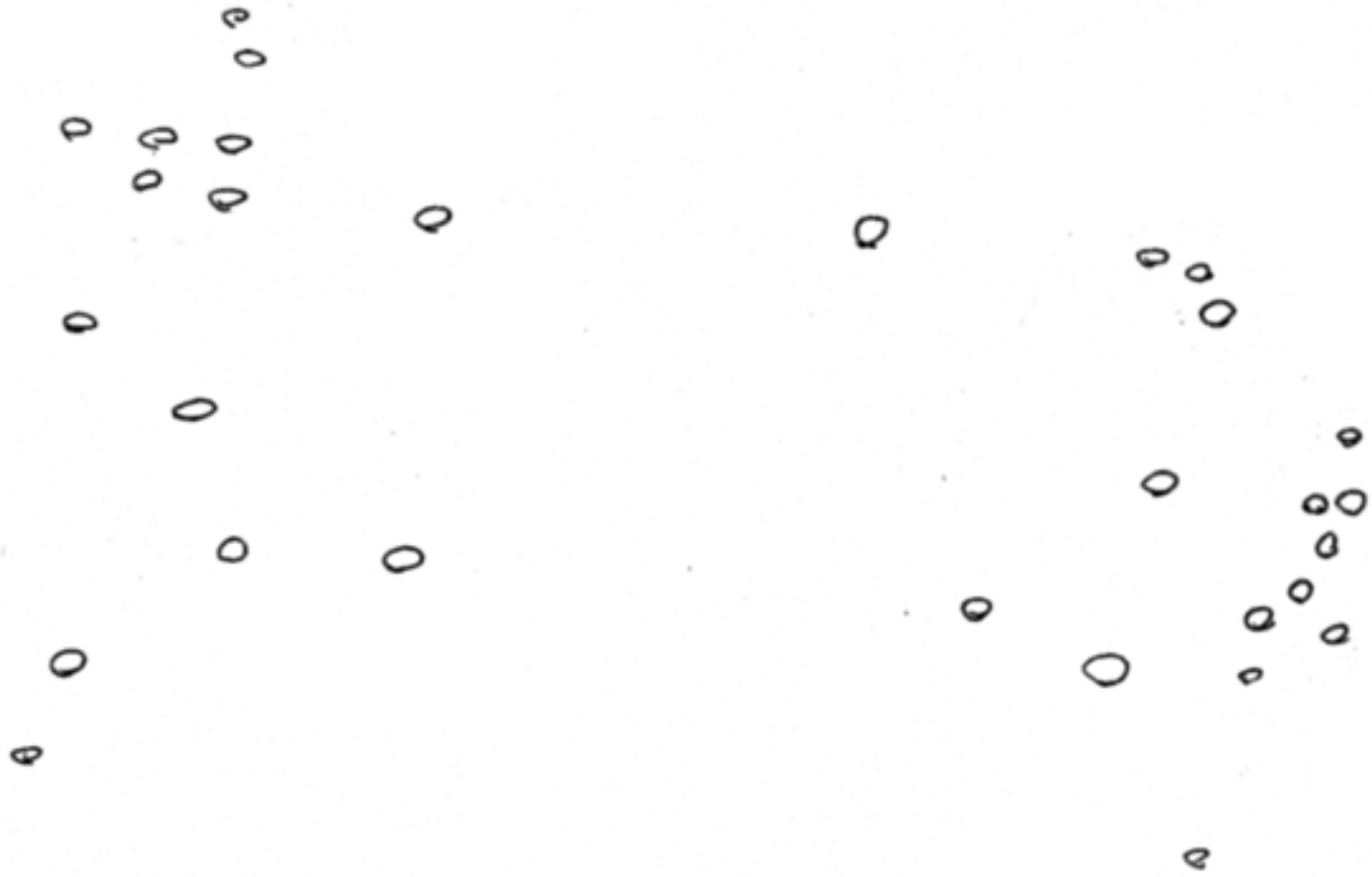
SciKit learn: http://scikit-learn.org/stable/tutorial/machine_learning

Abstract Overview: Supervised vs Unsupervised Mining

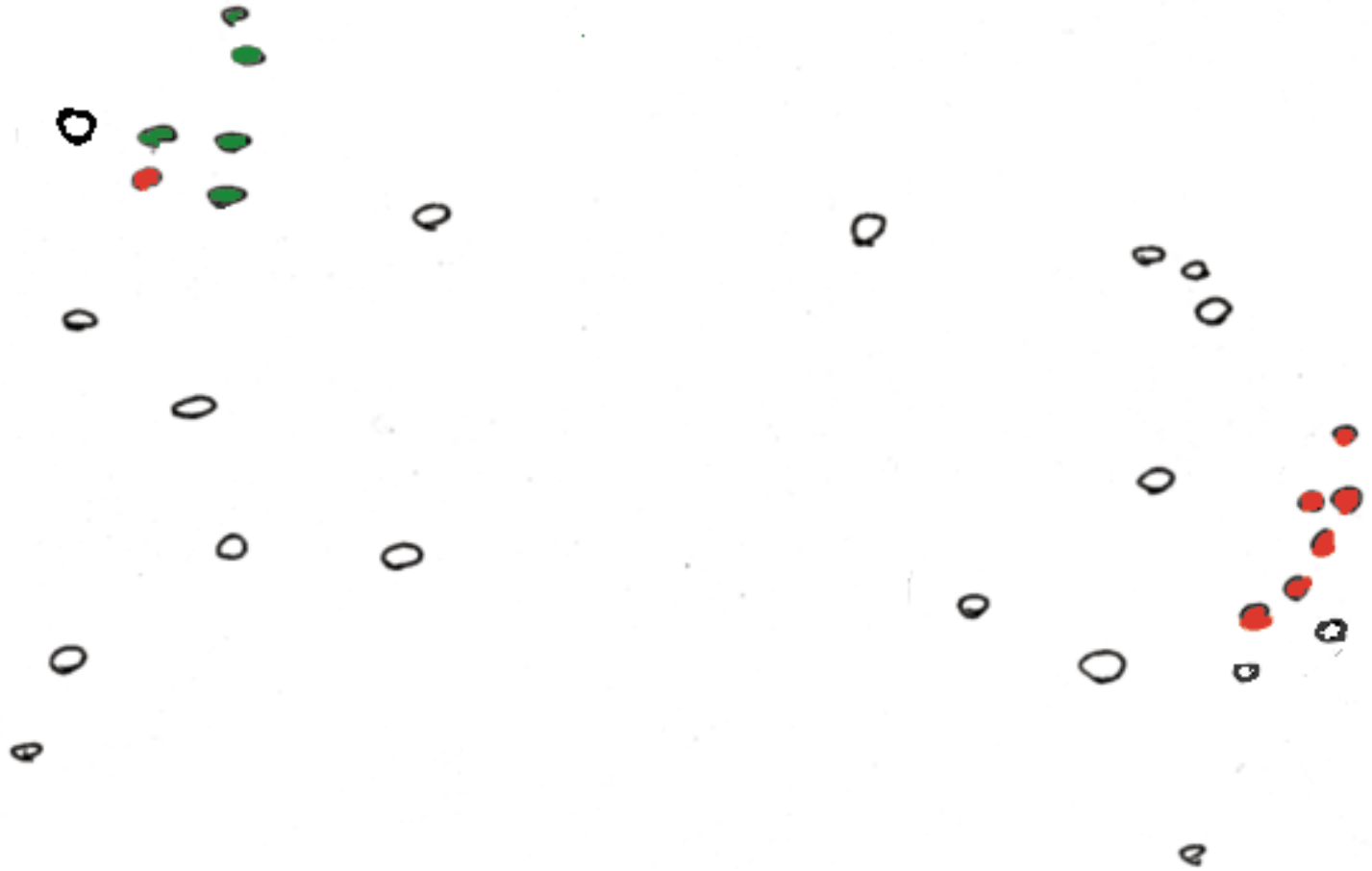
Structure of Genomic Features Matrix



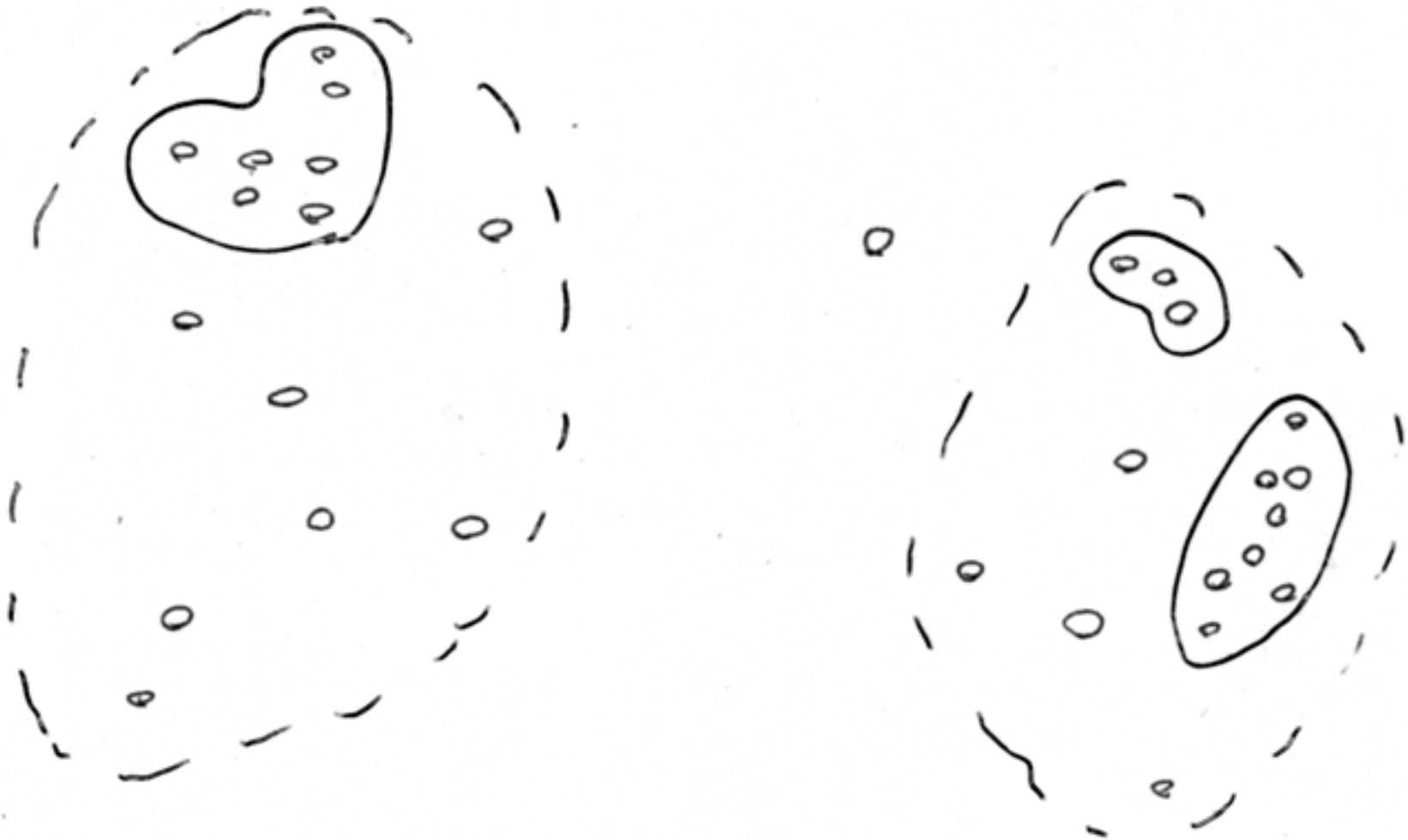
Represent predictors in abstract high dimensional space



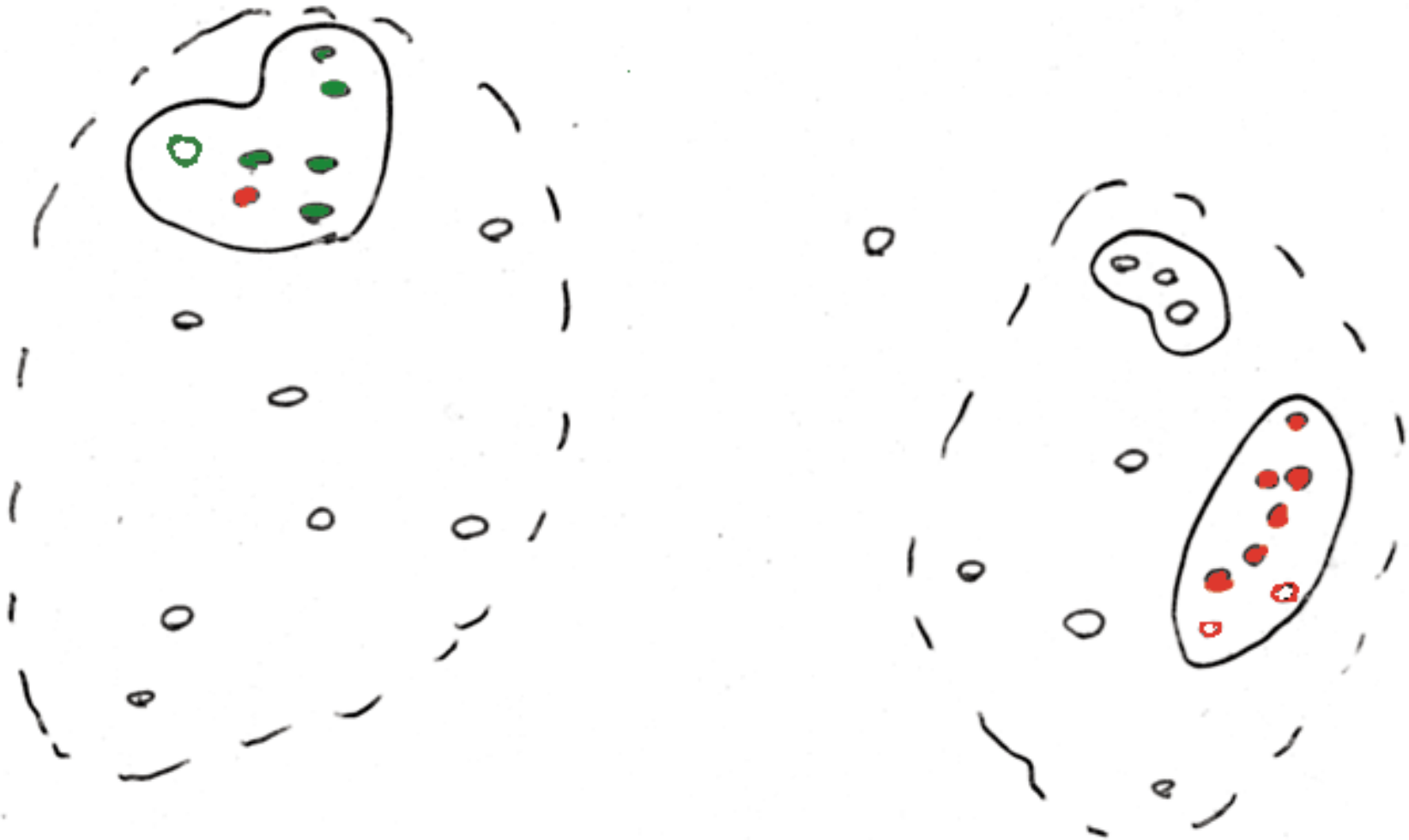
“Label” Certain Points



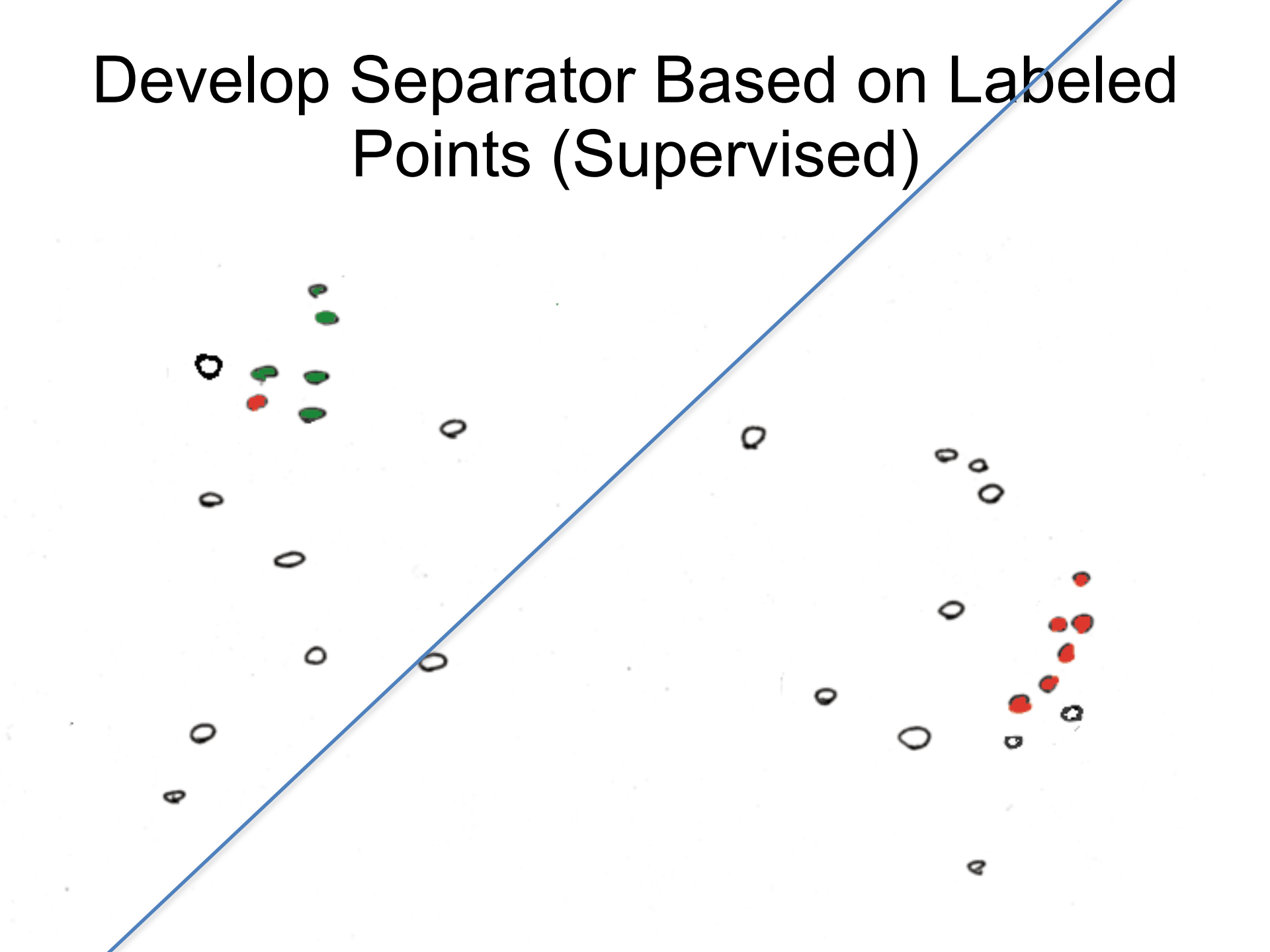
“Cluster” predictors (Unsupervised)



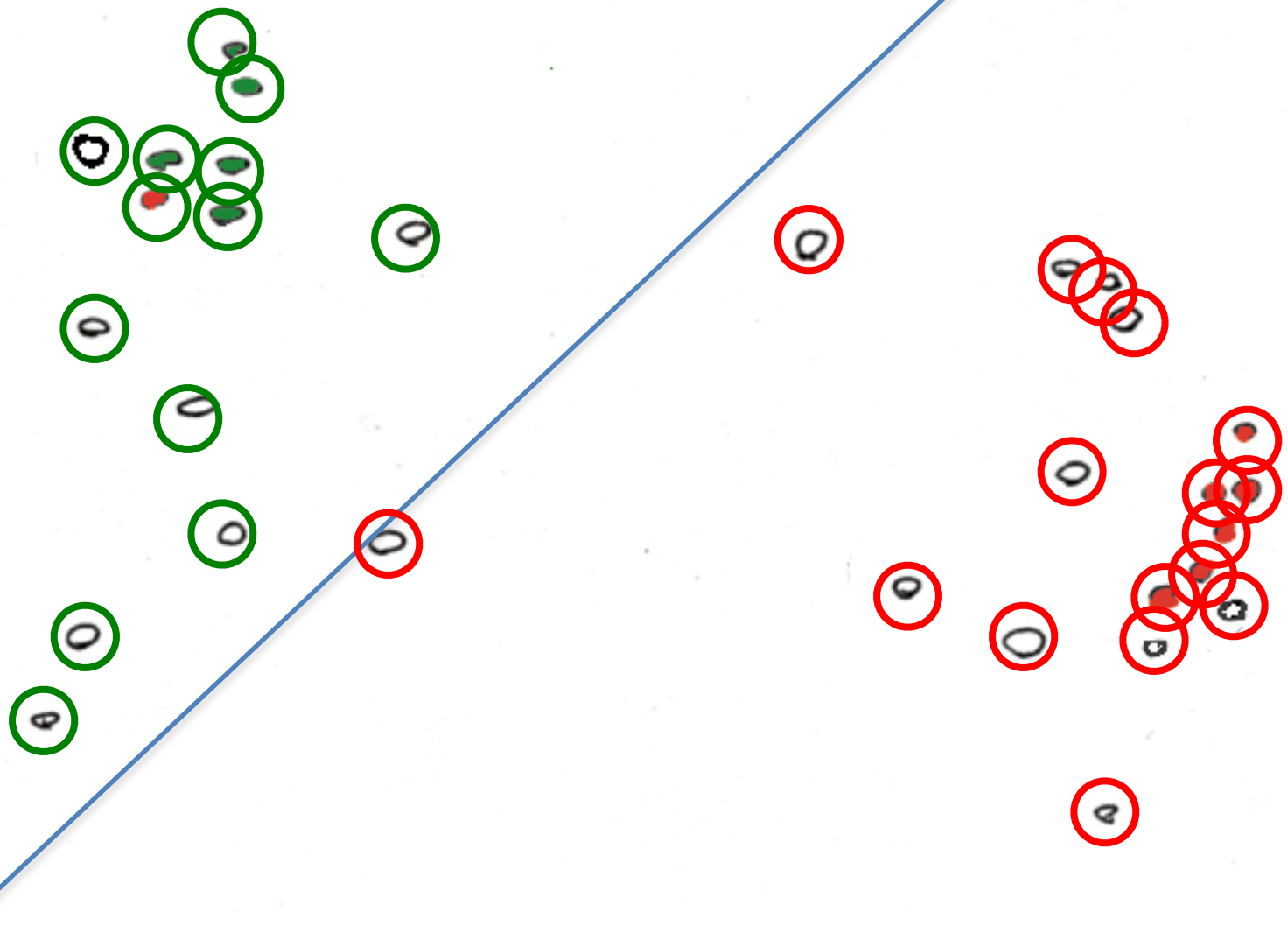
Use Clusters to predict Response
(Unsupervised, guilt-by-association)



Develop Separator Based on Labeled Points (Supervised)



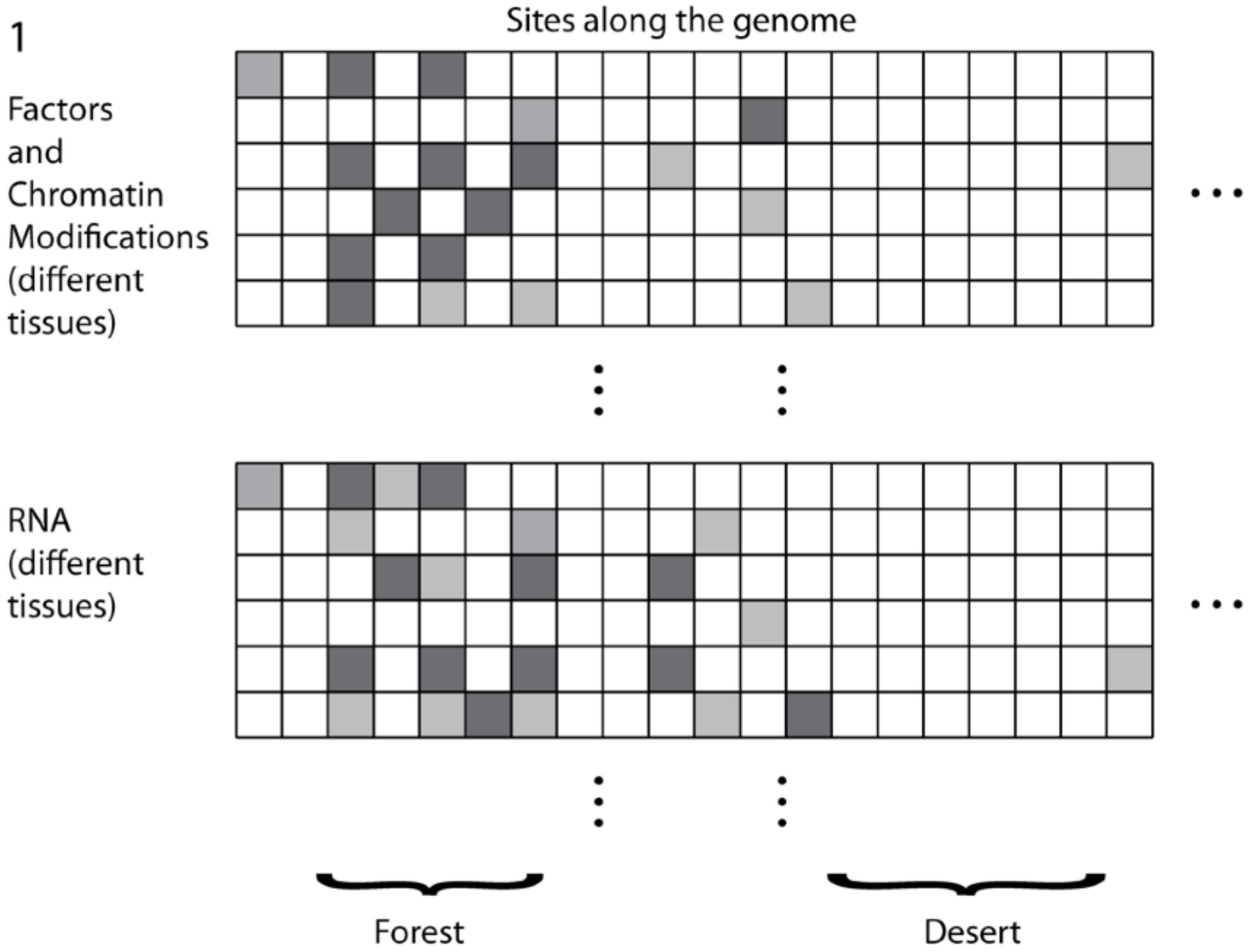
Predict based on Separator (Supervised)



Unsupervised Mining

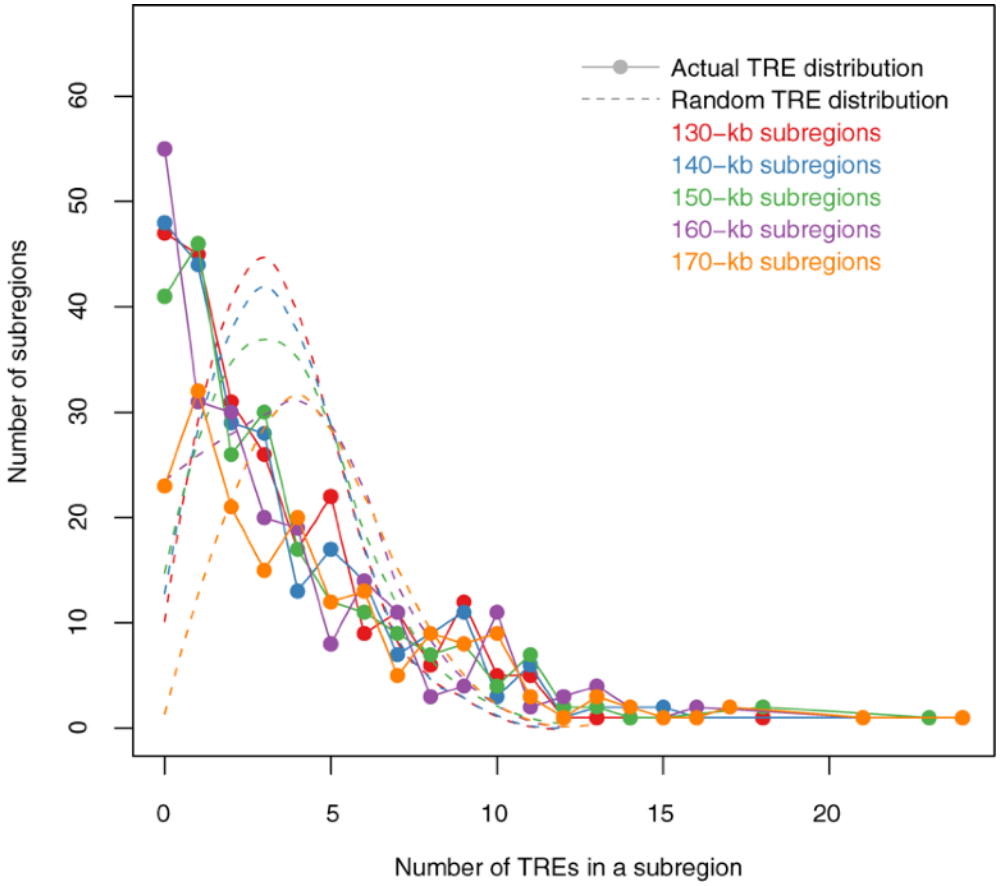
- Simple overlaps & enriched regions
- Clustering rows & columns (networks)
- PCA
- SVD (theory + appl.)
- Weighted Gene Co-Expression Network
- Biplot
- CCA

Genomic Features Matrix: Deserts & Forests



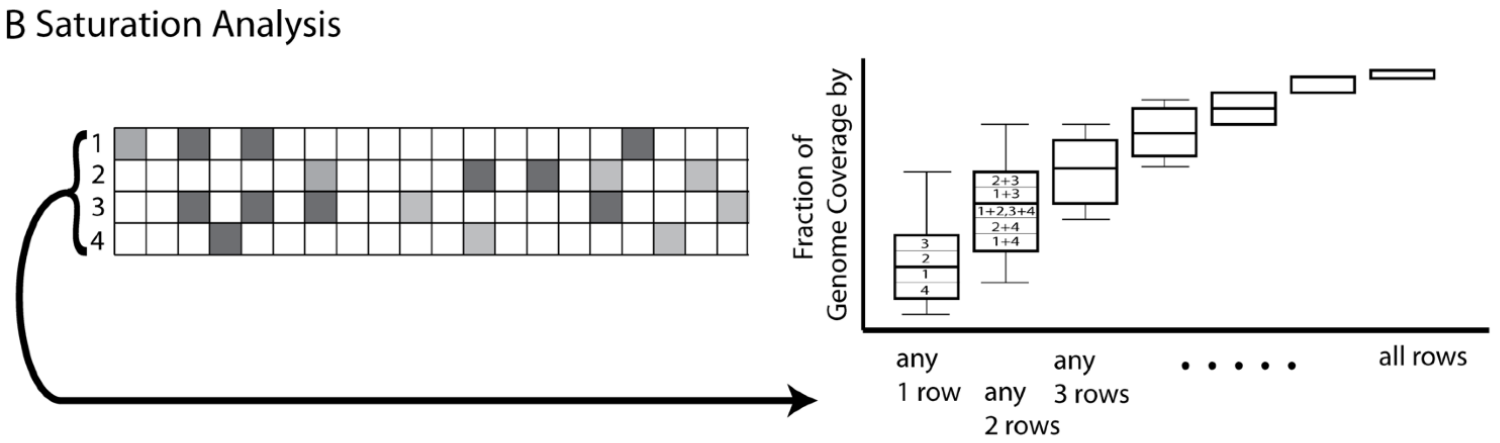
Non-random distribution of TREs

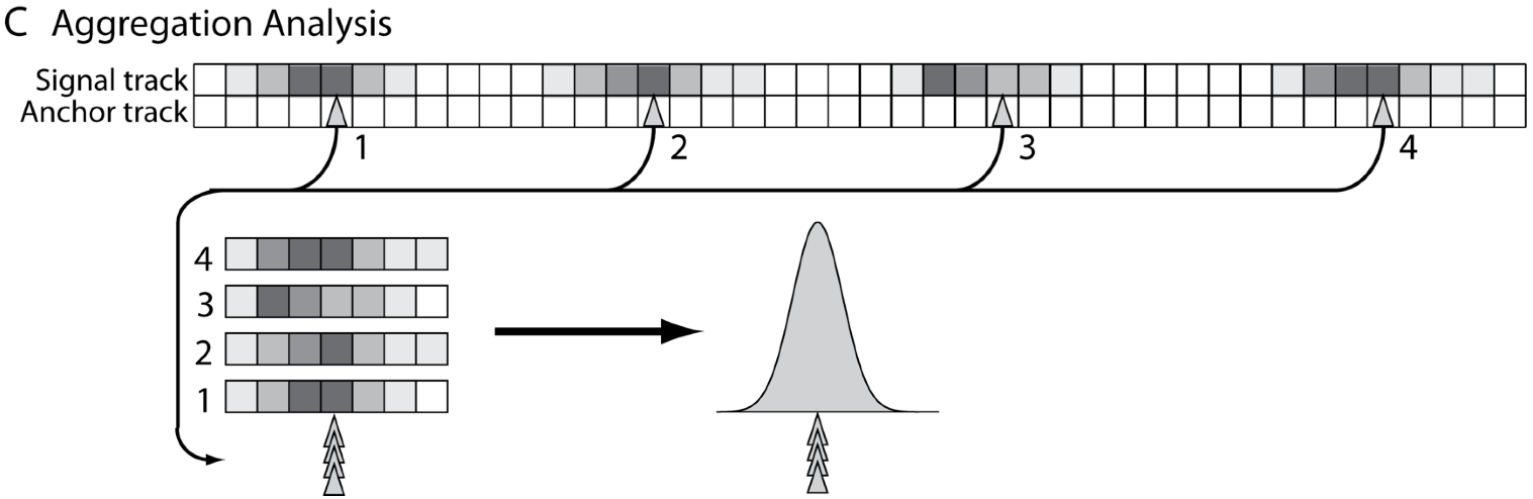
- TREs are not evenly distributed throughout the encode regions ($P < 2.2\times10^{-16}$).
- The actual TRE distribution is power-law.
- The null distribution is ‘Poissonesque.’
- Many genomic subregions with extreme numbers of TREs.



Zhang et al. (2007) Gen. Res.

Aggregation & Saturation



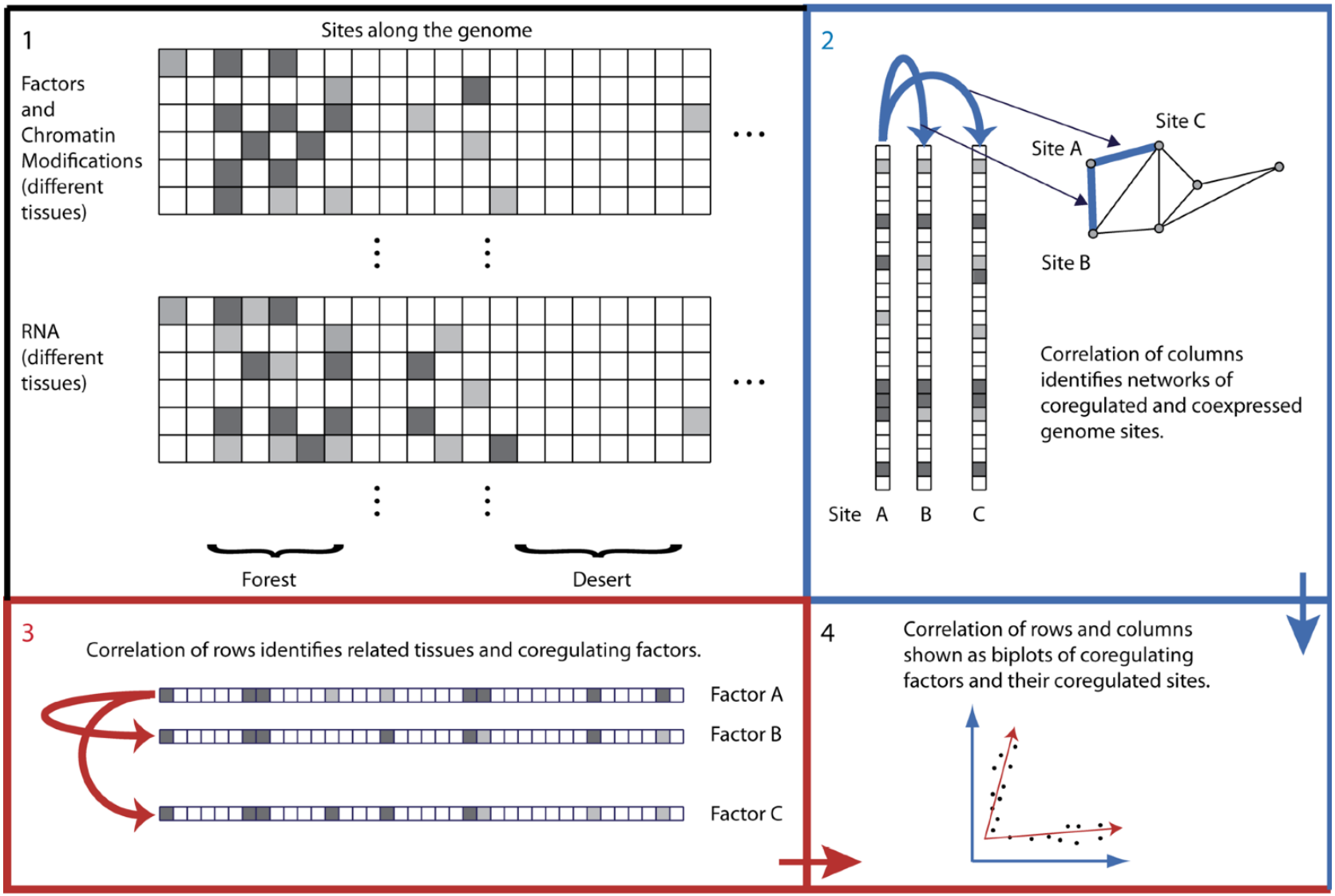


[*Nat. Rev. Genet.* (2010) 11: 559]

Unsupervised Mining

Clustering Columns & Rows of the Data Matrix

Correlating Rows & Columns



[*Nat. Rev. Genet.* (2010) 11: 559]

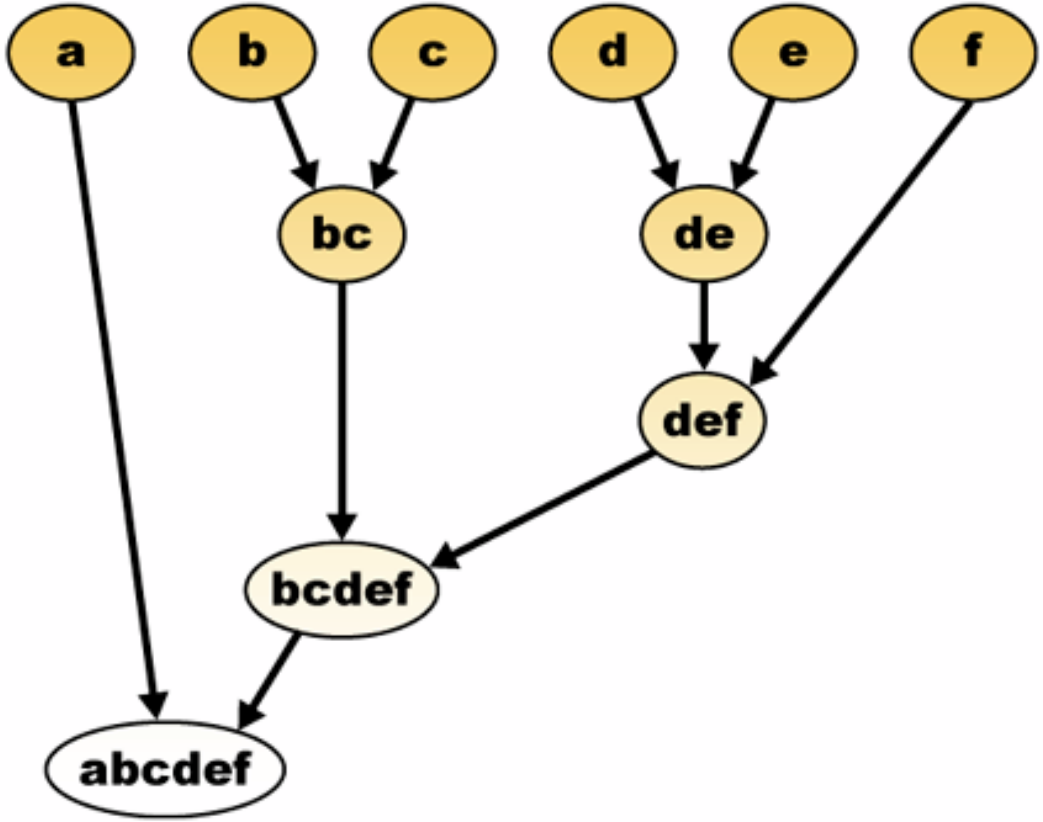
Spectral Methods
Outline & Papers

- Simple background on PCA (emphasizing lingo)
- Expression Clustering
- More abstract run through on SVD
- Application to
  - O Alter et al. (2000). "Singular value decomposition for genome-wide expression data processing and modeling." PNAS 97: 10101
  - Langfelder P, Horvath S (2007) Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology 2007, 1:54
  - Z Zhang et al. (2007) "Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions." Genome Res 17: 787
  - TA Gianoulis et al. (2009) "Quantifying environmental adaptation of metabolic pathways in metagenomics." PNAS 106: 1374.

Expression Clustering

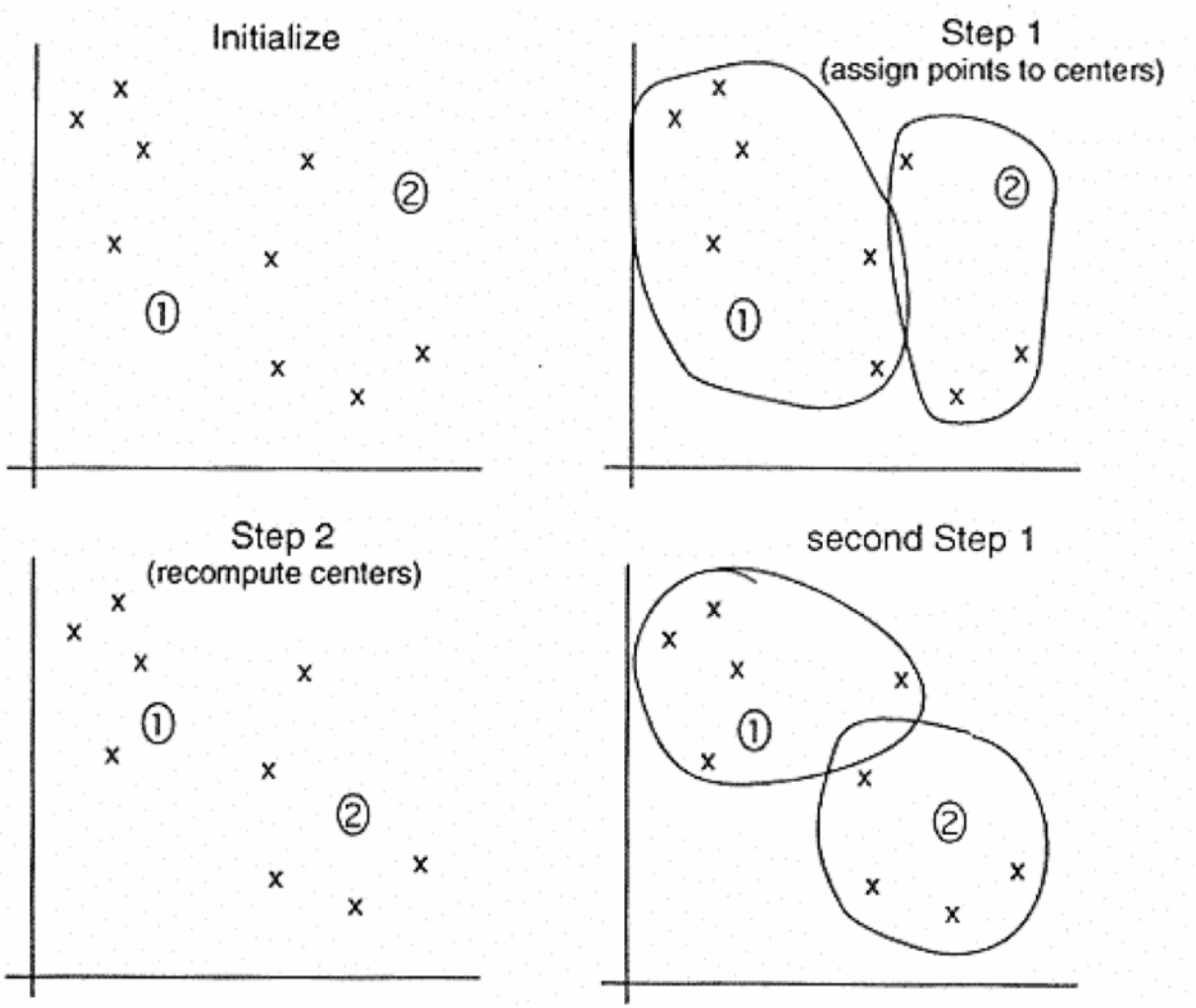
Agglomerative Clustering

- Bottom up v top down (K-means, know how many centers)
- Single or multi-link
  - threshold for connection?



http://commons.wikimedia.org/wiki/File:Hierarchical_clustering_diagram.png

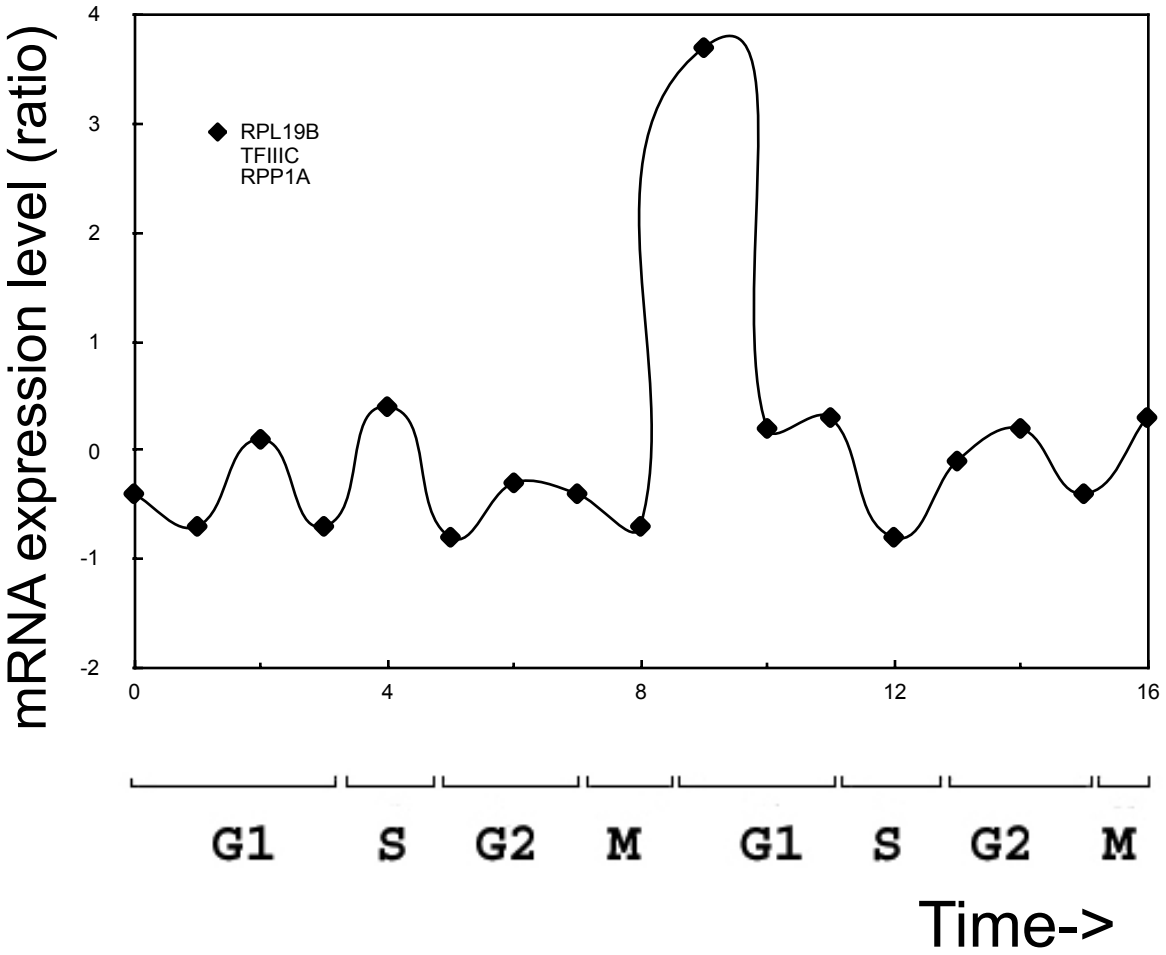
K-means

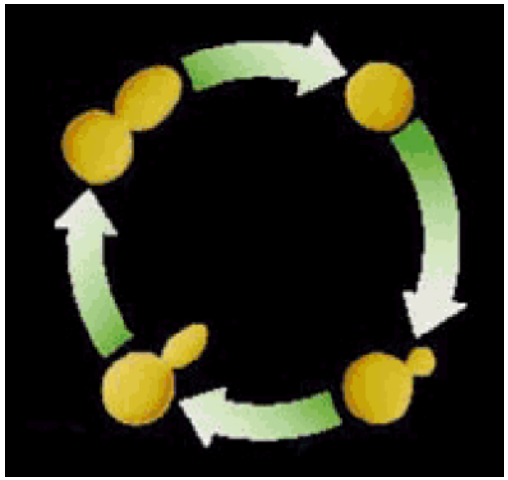


1) Pick ten (i.e. k?) random points as putative cluster centers.
2) Group the points to be clustered by the center to which they are closest.
3) Then take the mean of each group and repeat, with the means now at the cluster center.
4)Stop when the centers stop moving.

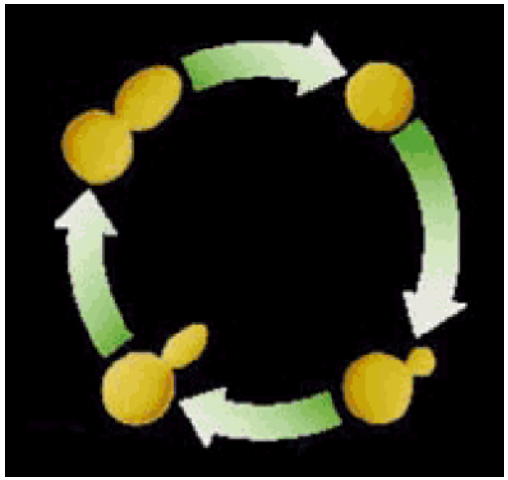
Clustering the yeast cell cycle to uncover interacting proteins

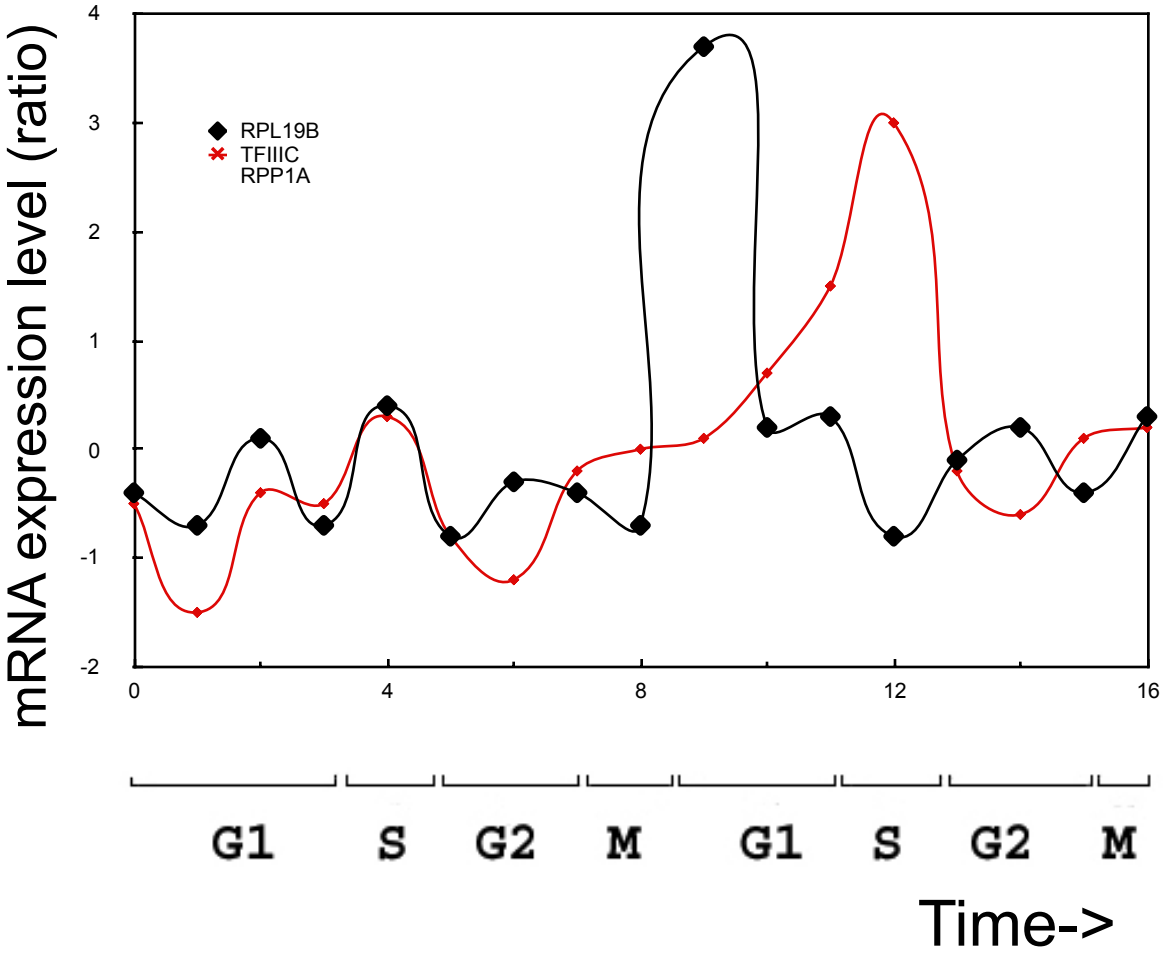
[Brown, Davis]





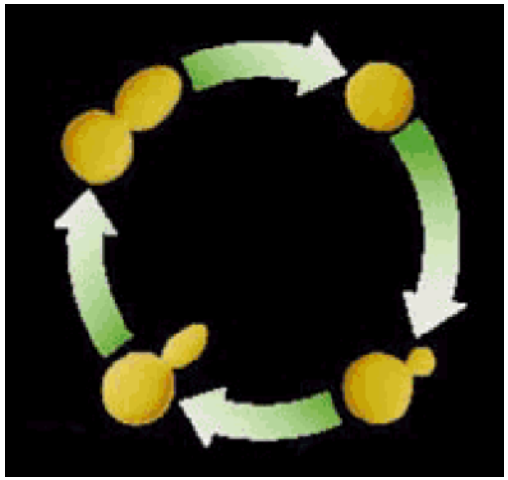
Clustering the yeast cell cycle to uncover interacting proteins

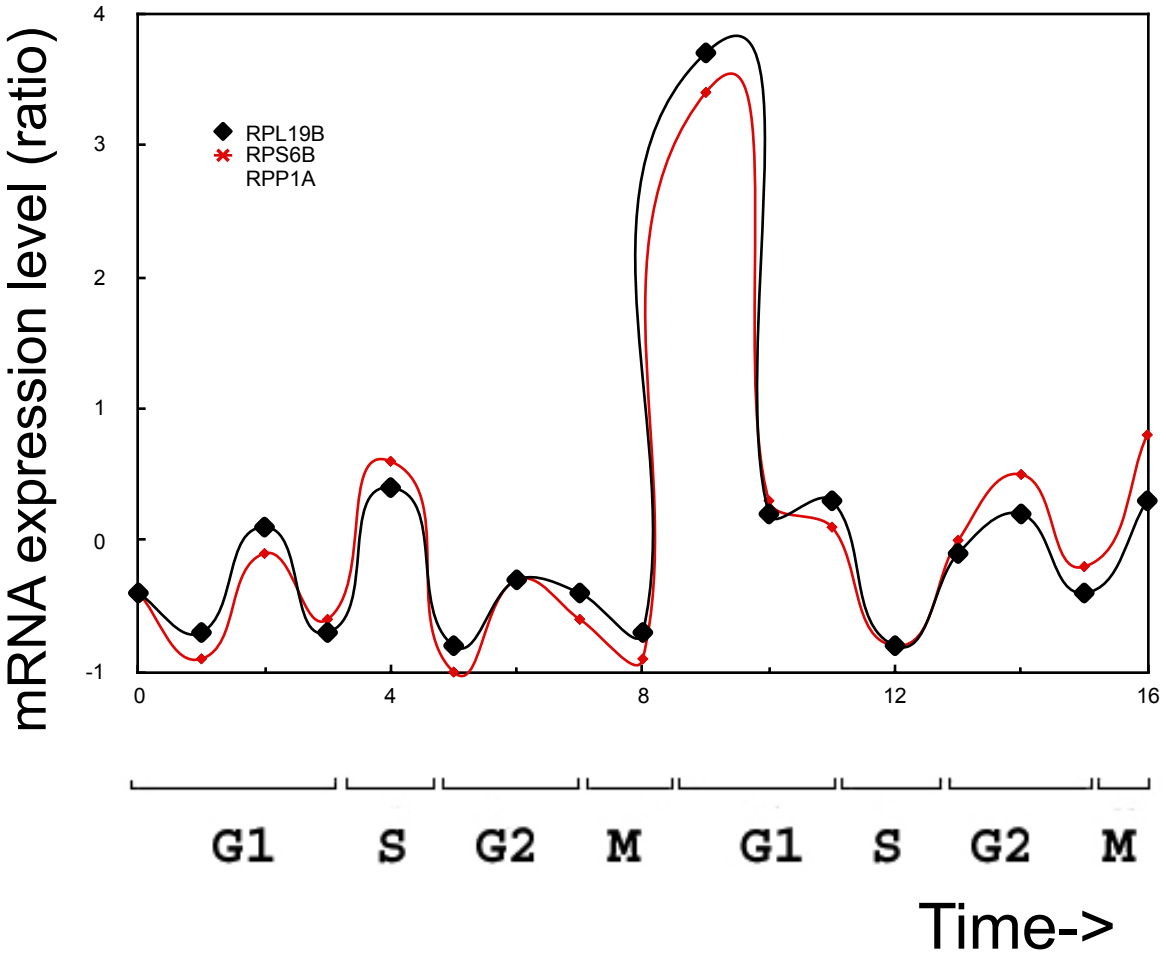




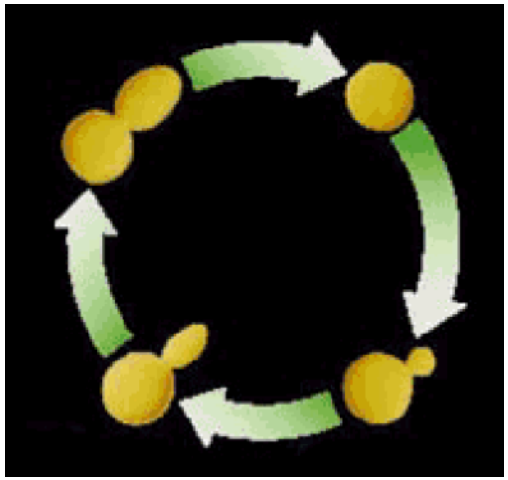
Clustering the yeast cell cycle to uncover interacting proteins

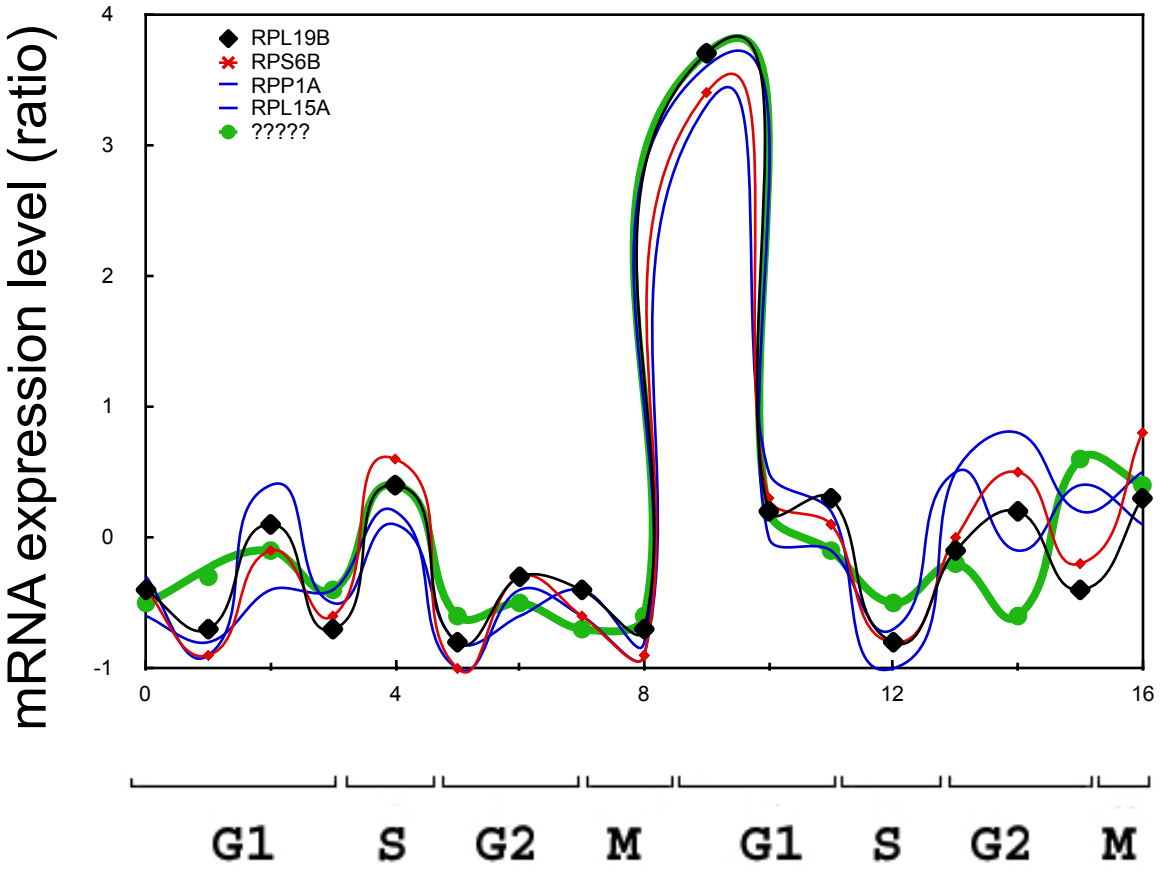
[Botstein; Church, Vidal]



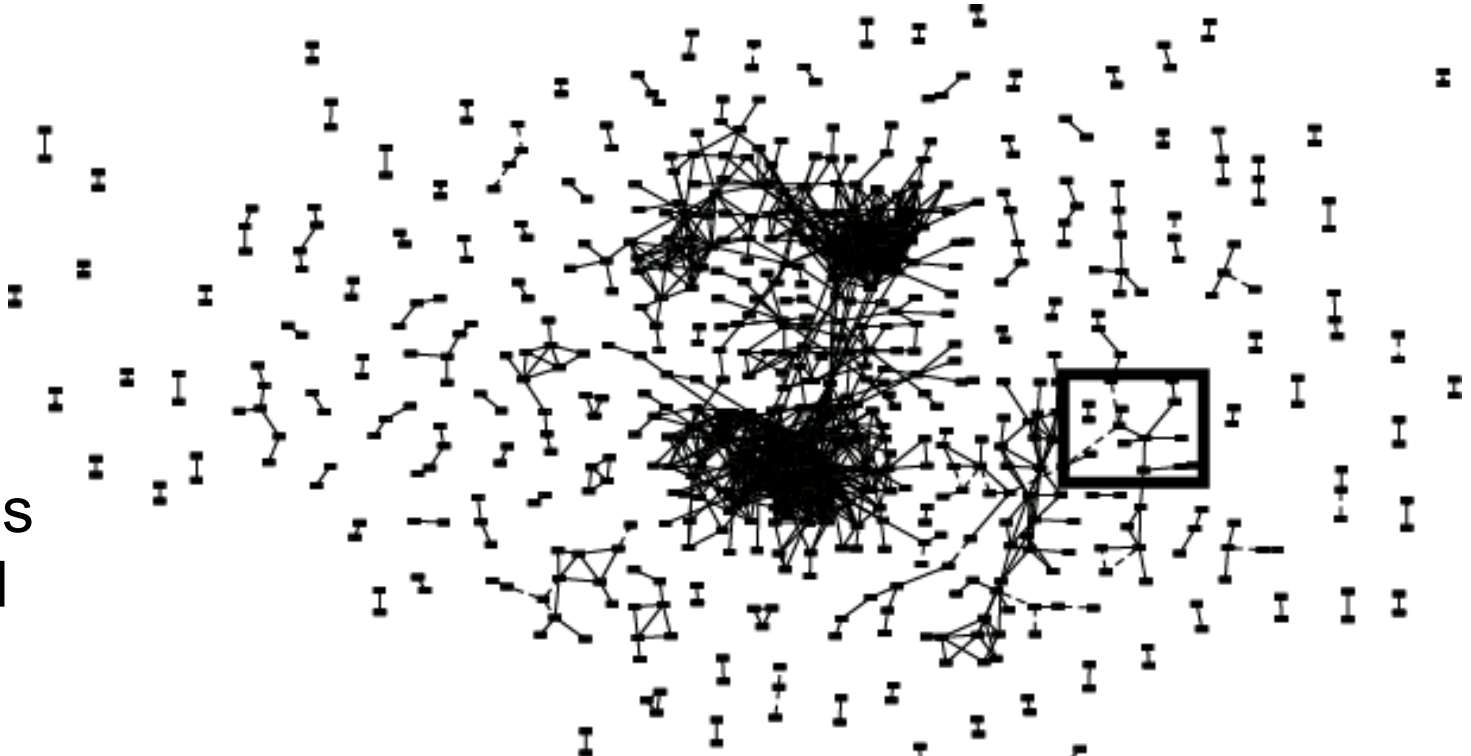


Clustering the yeast cell cycle to uncover interacting proteins





Global Network of Relationships

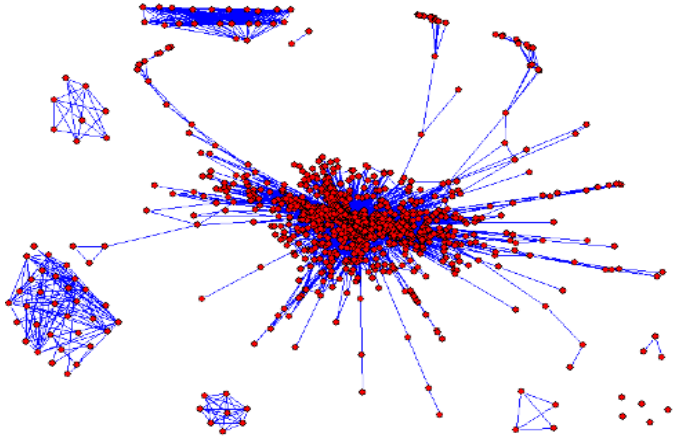


~470K significant relationships from **~18M** possible

Network = Adjacency Matrix

- Adjacency matrix $A=[a_{ij}]$ encodes whether/how a pair of nodes is connected.
- For unweighted networks: entries are 1 (connected) or 0 (disconnected)
- For weighted networks: adjacency matrix reports connection strength between gene pairs

Adapted from : http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork



Weighted Gene Co-Expression Network Analysis

Adapted from : http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork

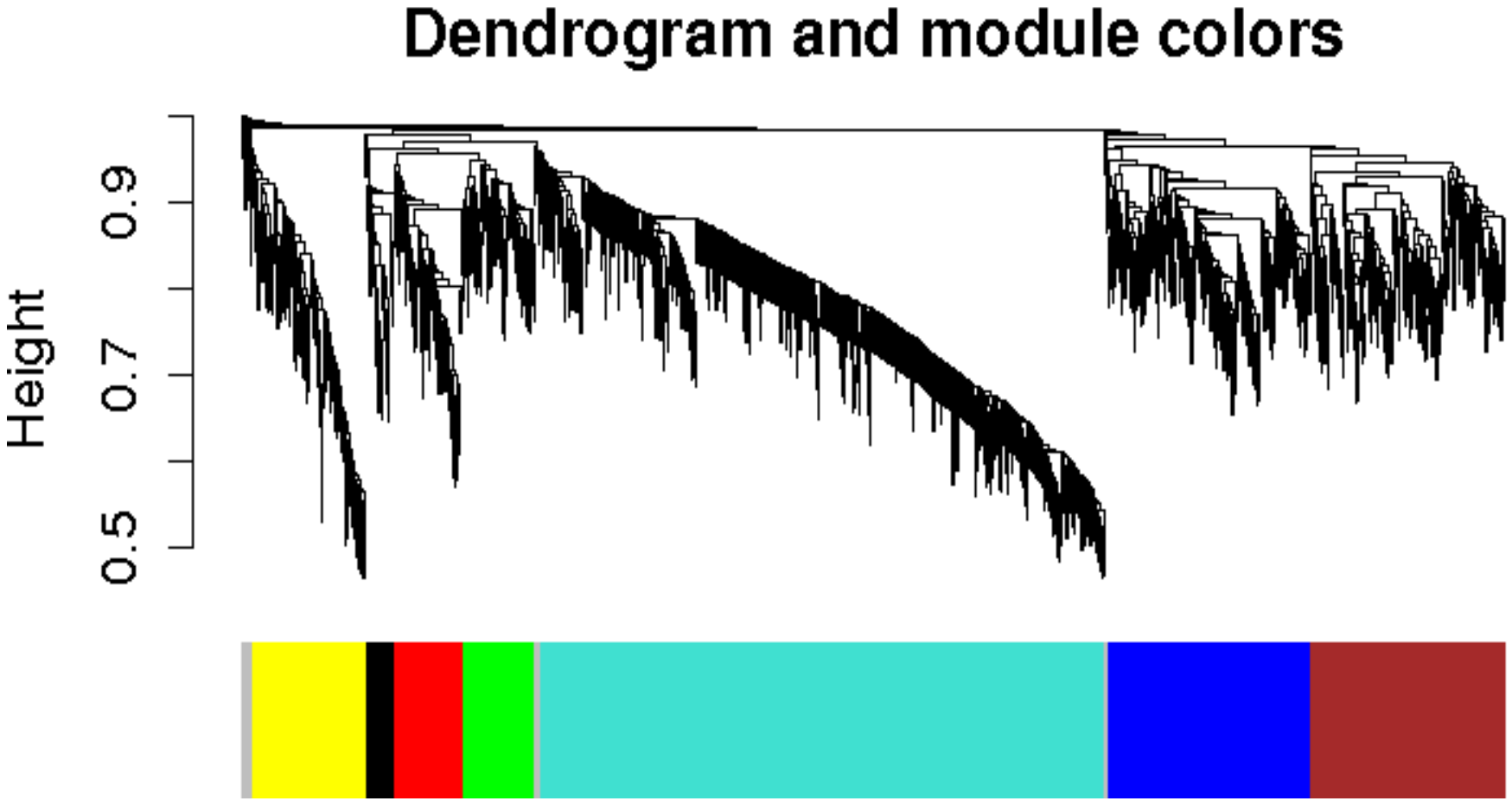
Module Detection

- Numerous methods exist
- Many methods define a suitable gene-gene dissimilarity measure and use clustering.
- In our case: dissimilarity based on topological overlap
- Clustering method: Average linkage hierarchical clustering
  - branches of the dendrogram are modules

Adapted from : http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork

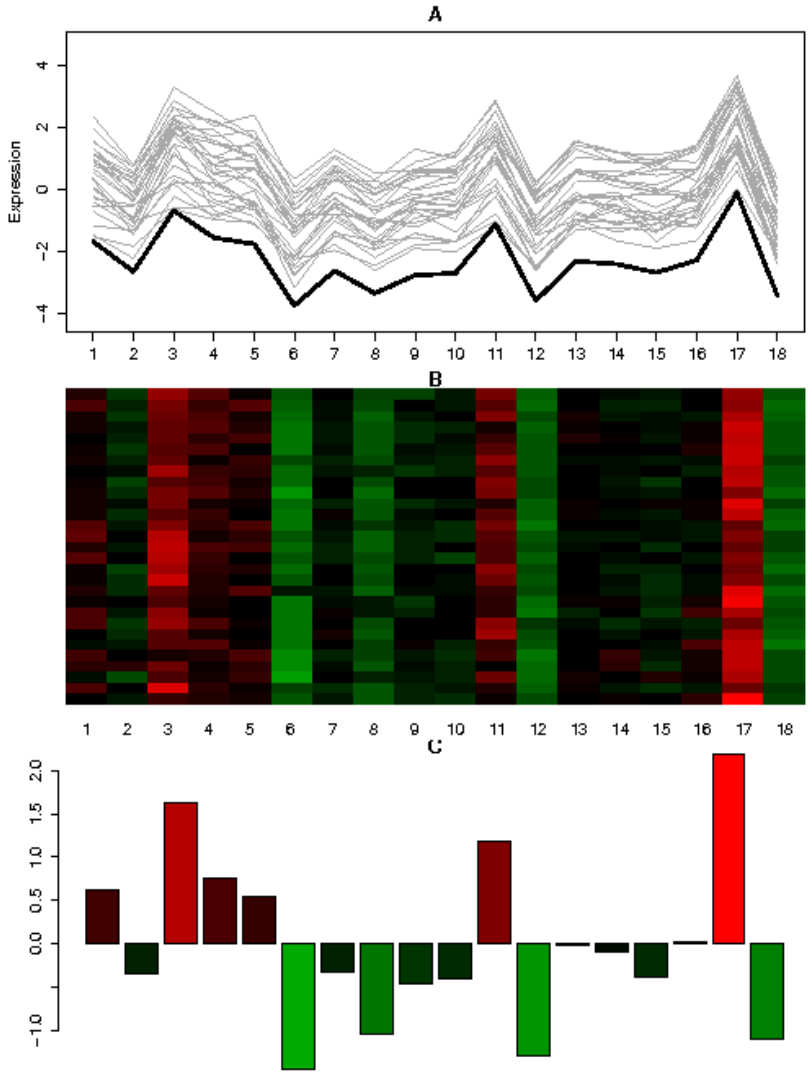
Example of module detection via hierarchical clustering

- Expression data from human brains, 18 samples.



Adapted from : http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork

Module eigengenes



- Often: Would like to treat modules as single units
  - Biologically motivated data reduction
- Our choice: module eigengene = 1st principal component of the module expression matrix
- Intuitively: a kind of average expression profile

Human brain expression data, 18 samples

Module consisting of 50 genes

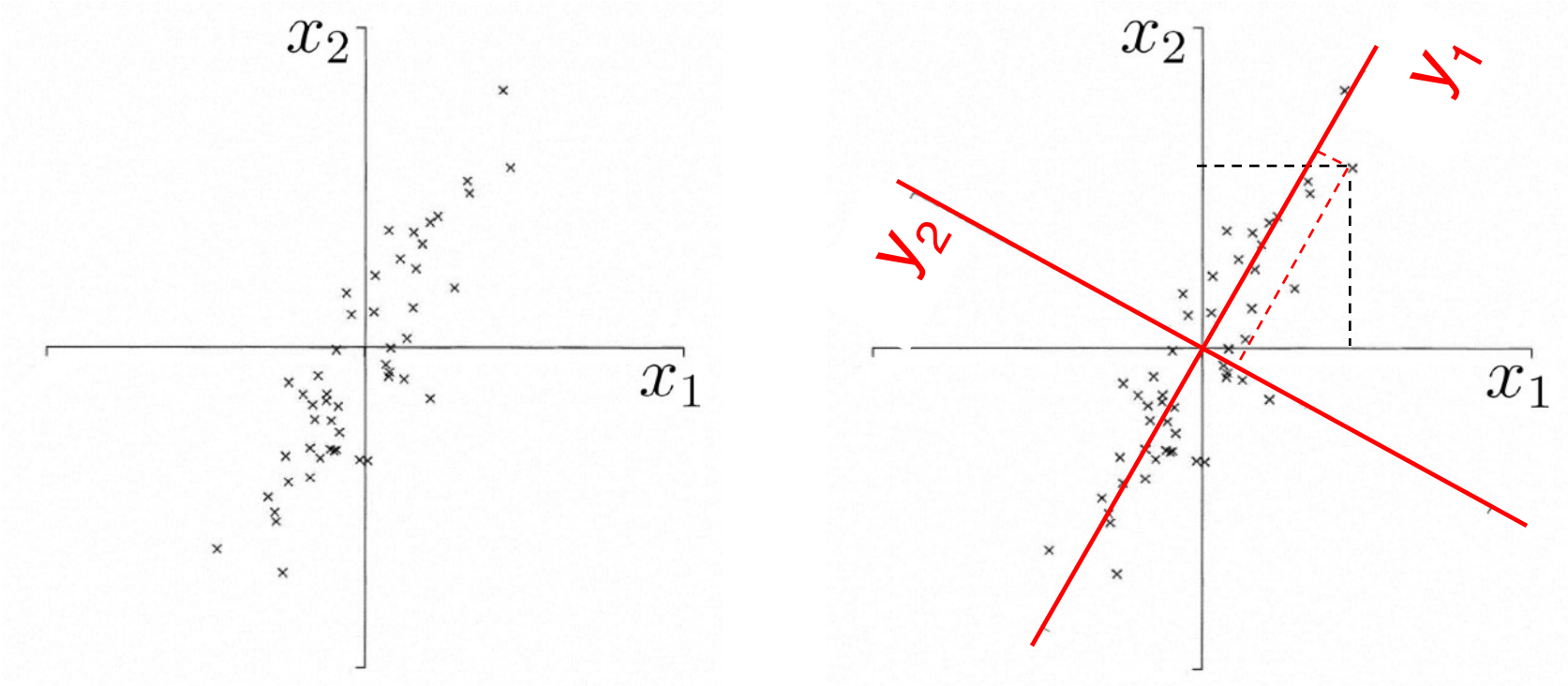
Langfelder P, Horvath S (2007) Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology 2007, 1:54

Adapted from : http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork

Quick Refresher on PCA/Matrices

What is PCA?

- A technique used to reduce the dimensionality of a data set by finding directions of maximum variability
- Projection (typically a rotation) into new axes
- But still retains the dataset's variation



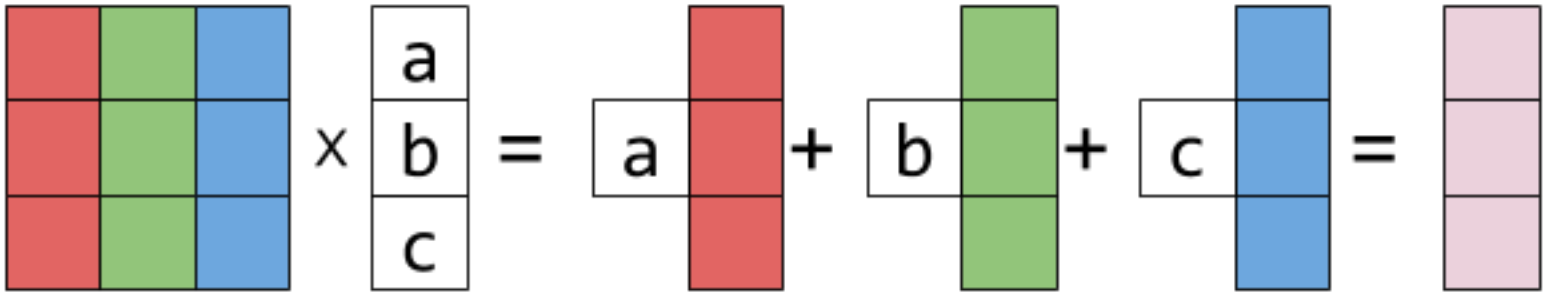
Adapted from http://www.astro.princeton.edu/~gk/A542/PCA.ppt

Quick Refresher on Matrices

$$\begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{pmatrix} * \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} ax_1 + by_1 + cz_1 \\ ax_2 + by_2 + cz_2 \\ ax_3 + by_3 + cz_3 \end{pmatrix}$$

$$\underset{\text{Matrix A is } 3x4}{\begin{bmatrix} 8 & 3 & 0 & 1 \\ . & . & . & . \\ . & . & . & . \end{bmatrix}} \underset{\text{Matrix B is } 4x4}{\begin{bmatrix} 5 & . & . & . \\ 4 & . & . & . \\ 3 & . & . & . \\ 1 & . & . & . \end{bmatrix}} = \underset{\text{Matrix C is } 3x4}{\begin{bmatrix} 53 & . & . & . \\ . & . & . & . \\ . & . & . & . \end{bmatrix}}$$

$$\text{because } c_{11} = \sum_{k=1}^{4} a_{1k} b_{k1} = 8 \cdot 5 + 3 \cdot 4 + 0 \cdot 3 + 1 \cdot 1 = 53$$



http://eli.thegreenplace.net/2015/visualizing-matrix-multiplication-as-a-linear-combination/

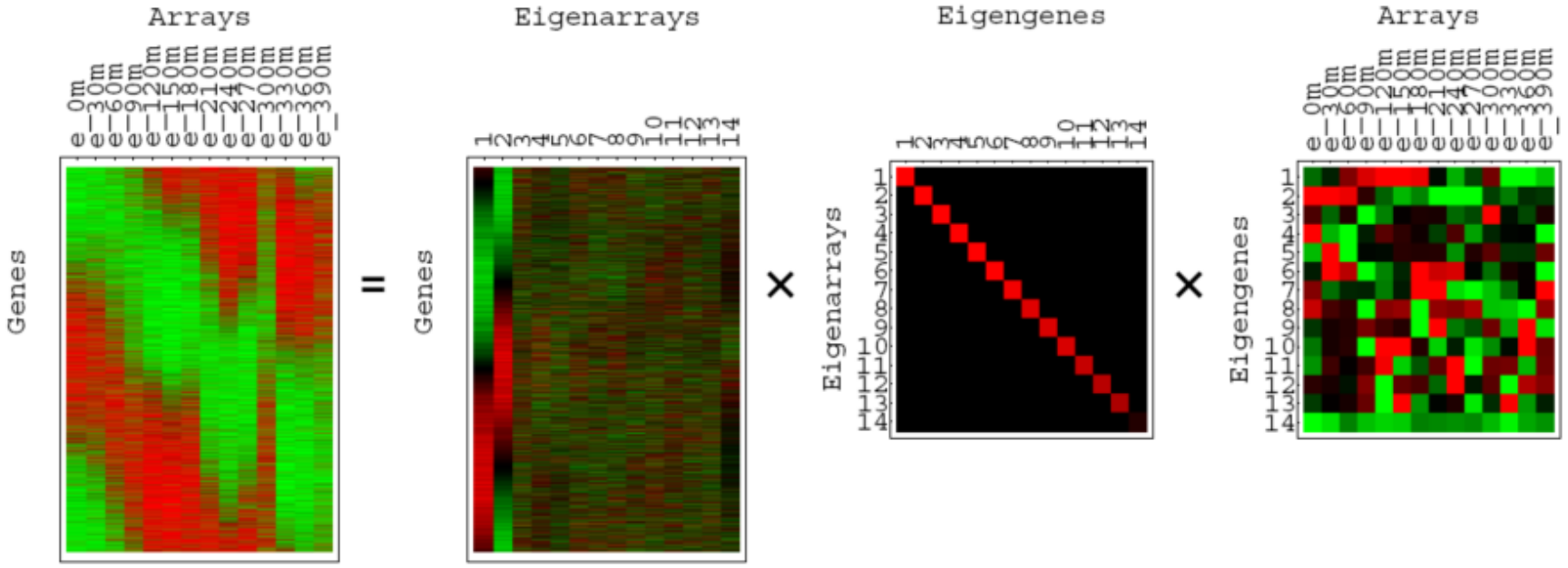
http://www.catonmat.net/blog/mit-linear-algebra-part-three/

Unsupervised Mining

SVD

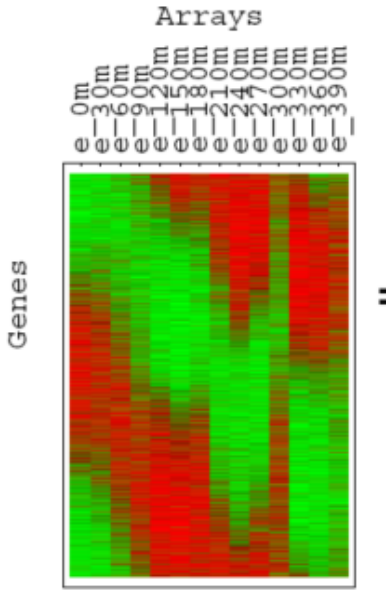
Puts together slides prepared by
Brandon Xia with images from
Alter et al. papers

SVD for microarray data (Alter et al, PNAS 2000)



$$A = USV^T$$

- A is any rectangular matrix (m ≥ n)
- Row space: vector subspace generated by the row vectors of A
- Column space: vector subspace generated by the column vectors of A
  - The dimension of the row & column space is the rank of the matrix A: r (≤ n)
- A is a linear transformation that maps vector x in row space into vector Ax in column space

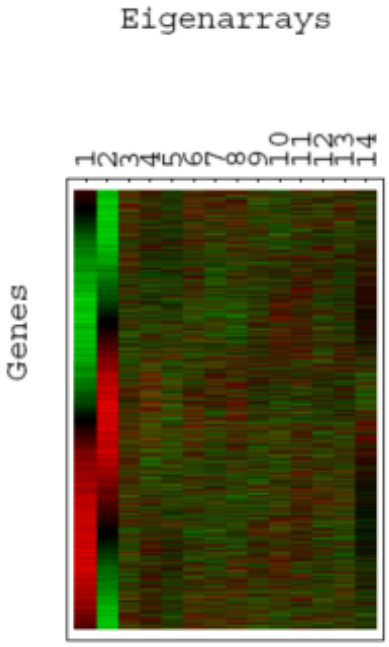


$$A = USV^T$$

- U is an “orthogonal” matrix (m ≥ n)
- Column vectors of U form an orthonormal basis for the column space of A: $U^T U = I$

$$U = \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & ? & \mathbf{u}_n \\ | & | & & | \end{pmatrix}$$

- $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ in U are eigenvectors of AA^T
  - $AA^T = USV^T\, VSU^T = US^2\, U^T$
  - “Left singular vectors”

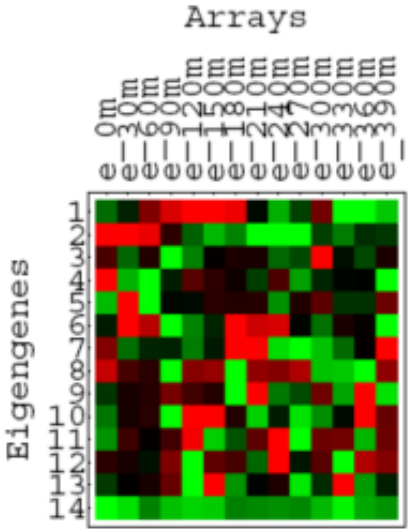


$$A = USV^T$$

- V is an orthogonal matrix (n by n)
- Column vectors of V form an orthonormal basis for the row space of A: $V^TV=VV^T=I$

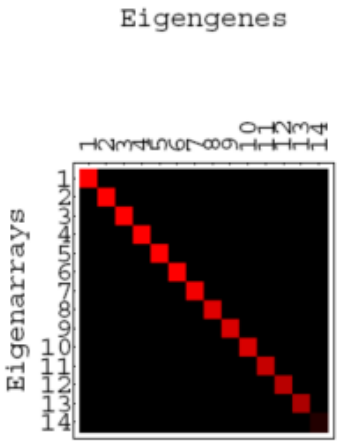
$$V = \begin{pmatrix} | & | & & | \\ \mathbf{v}_1 & \mathbf{v}_2 & ? & \mathbf{v}_n \\ | & | & & | \end{pmatrix}$$

- $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ in V are eigenvectors of A^TA
  - $A^TA = VSU^T\,USV^T = VS^2\,V^T$
  - "Right singular vectors"



$$A = USV^T$$

- S is a diagonal matrix (n by n) of non-negative singular values
- Typically sorted from largest to smallest
- Singular values are the non-negative square root of corresponding eigenvalues of A^TA and AA^T



$$AV = US$$

- Means each $A\boldsymbol{v}_i = s_i\boldsymbol{u}_i$
- Remember A is a linear map from row space to column space
- Here, A maps an orthonormal basis $\{\boldsymbol{v}_i\}$ in row space into an orthonormal basis $\{\boldsymbol{u}_i\}$ in column space
- Each component of u_i is the projection of a row of the data matrix A onto the vector v_i

SVD as sum of rank-1 matrices

- $A = USV^T$
- $A = s_1\boldsymbol{u}_1\boldsymbol{v}_1^T + s_2\boldsymbol{u}_2\boldsymbol{v}_2^T + \ldots + s_n\boldsymbol{u}_n\boldsymbol{v}_n^T$
- $s_1 \geq s_2 \geq \ldots \geq s_n \geq 0$
- What is the rank-r matrix $\hat{A}$ that best approximates A ?
  - Minimize $\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\hat{A}_{ij} - A_{ij}\right)^2$
- $\hat{A} = s_1\boldsymbol{u}_1\boldsymbol{v}_1^T + s_2\boldsymbol{u}_2\boldsymbol{v}_2^T + \ldots + s_r\boldsymbol{u}_r\boldsymbol{v}_r^T$
- Very useful for matrix approximation

an outer product (uv^T) giving a matrix rather than the scalar of the inner product

LSQ approx. If r=1, this amounts to a line fit.

Examples of (almost) rank-1 matrices

- Steady states with fluctuations

$$\begin{pmatrix} 101 & 103 & 102 \\ 302 & 300 & 301 \\ 203 & 204 & 203 \\ 401 & 402 & 404 \end{pmatrix}$$

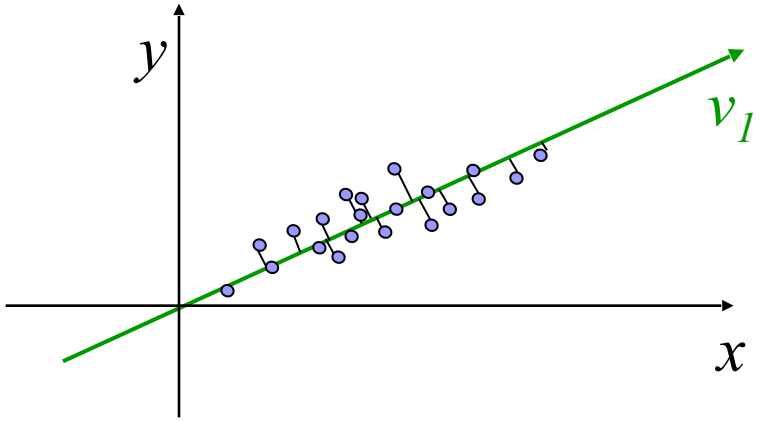
- Array artifacts?

$$\begin{pmatrix} 101 & 303 & 202 \\ 102 & 300 & 201 \\ 103 & 304 & 203 \\ 101 & 302 & 204 \end{pmatrix}$$

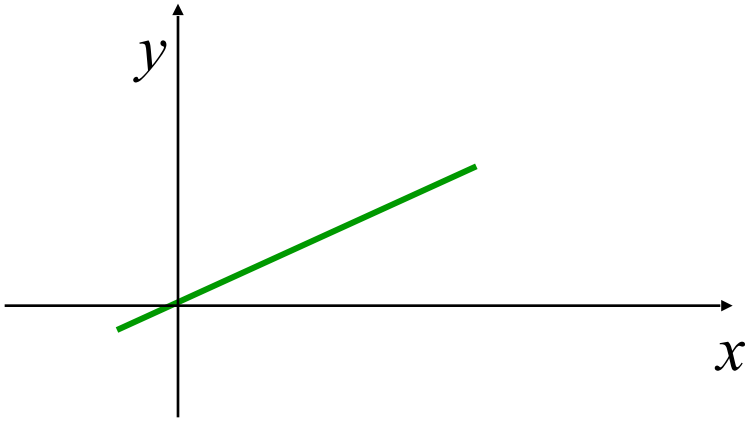
- Signals?

$$\begin{pmatrix} 1 & 2 & -1 \\ 2 & 4 & -2 \\ -1 & -2 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

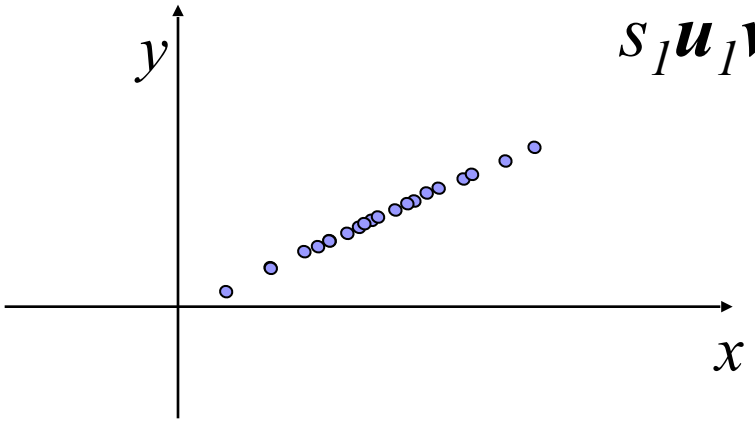
Geometry of SVD in row space



A



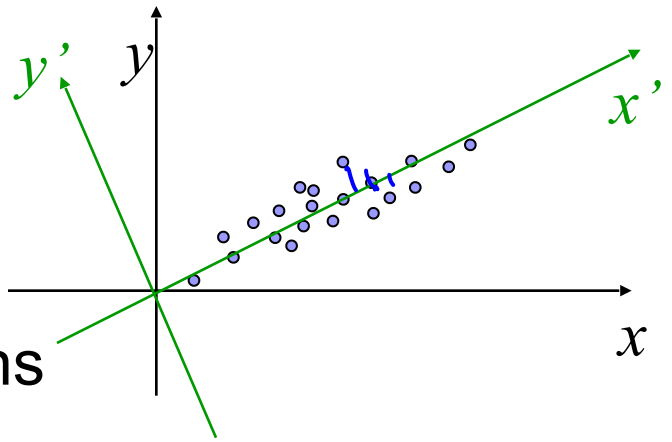
This line segment that goes through origin approximates the original data set



The projected data set approximates the original data set

Geometry of SVD in row space

- A as a collection of m row vectors (points) in the row space of A
- $s_1\boldsymbol{u}_1\boldsymbol{v}_1^T + s_2\boldsymbol{u}_2\boldsymbol{v}_2^T$ is the best rank-2 matrix approximation for A
- Geometrically: $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are the directions of the best approximating rank-2 subspace that goes through origin
- $s_1\boldsymbol{u}_1$ and $s_2\boldsymbol{u}_2$ gives coordinates for row vectors in rank-2 subspace
- $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ gives coordinates for row space basis vectors in rank-2 subspace



$$A\,\mathbf{v_i} = s_i\mathbf{u_i}$$

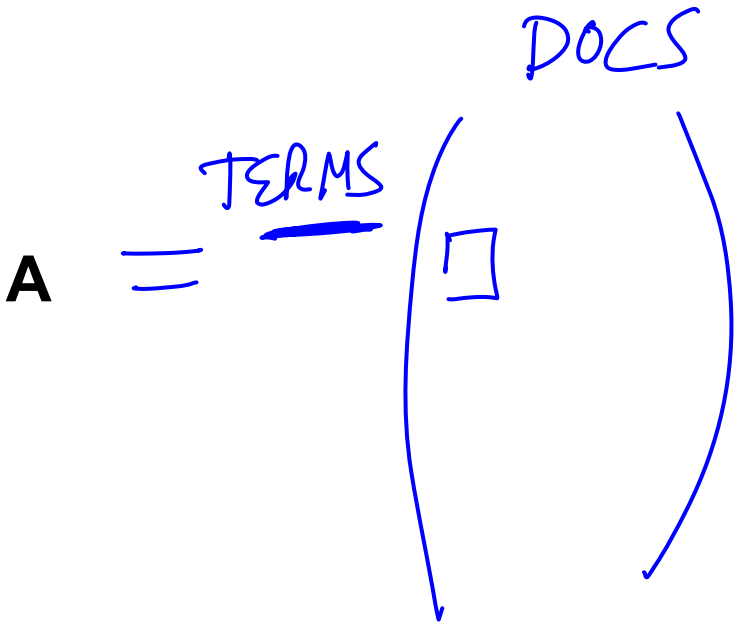
$$I\,\mathbf{v_i} = \mathbf{v_i}$$

What about geometry of SVD in column space?

- $A = USV^T$
- $A^T = VSU^T$
- The column space of A becomes the row space of A^T
- The same as before, except that U and V are switched

Additional Points

- Time Complexity (Cubic)
- Application to text mining
  - Latent semantic indexing
  - sparse



Potential problems of SVD/PCA

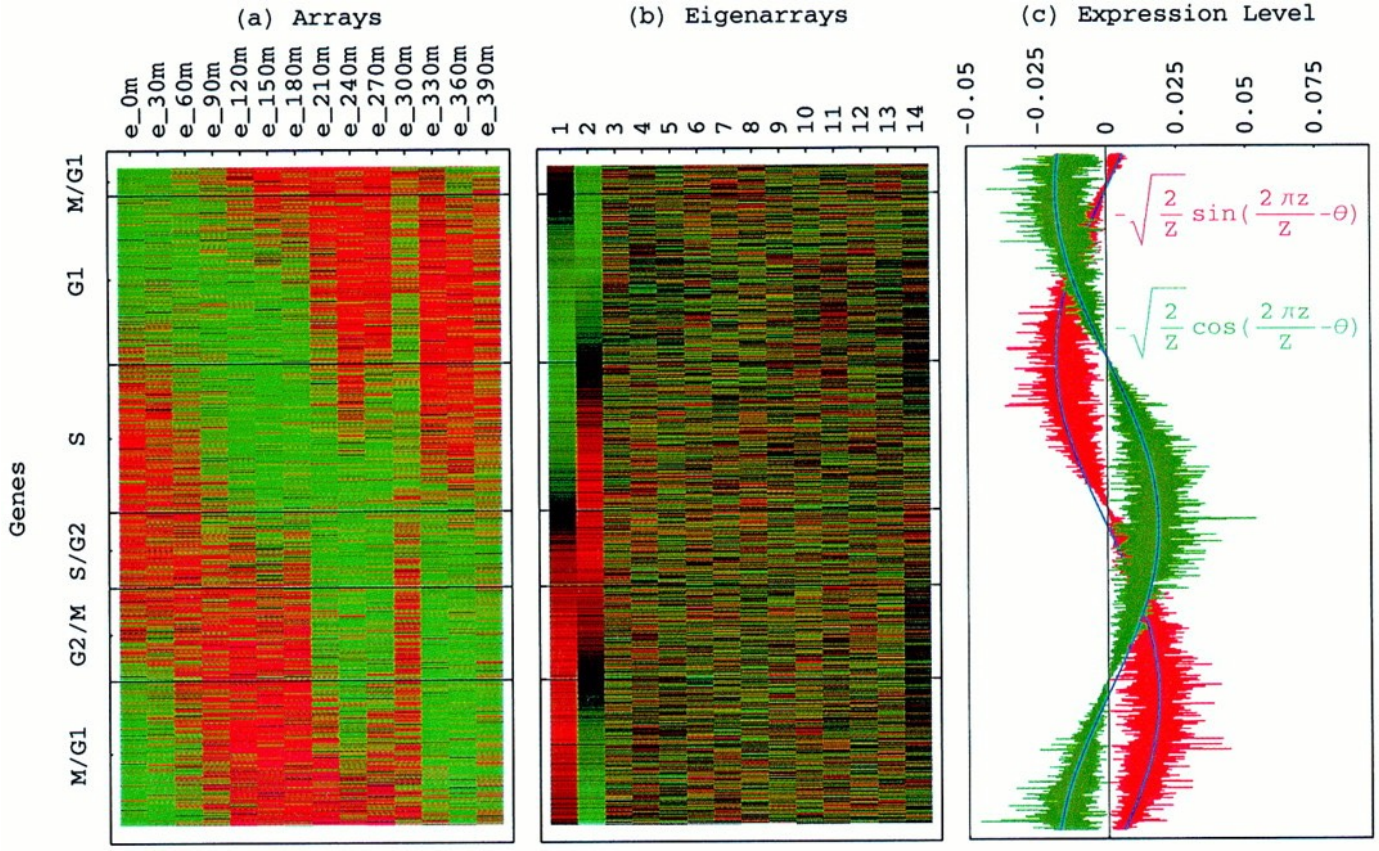
If the dataset…

- Lacks Independence
  - **NO PROBLEM**
- Lacks Normality
  - Normality desirable but not essential
- Lacks Precision
  - Precision desirable but not essential
- Lacks Linearity
  - **Problem**: Use other non-linear (kernel) methods
- Many Zeroes in Data Matrix (Sparse)
  - **Problem**: Use Correspondence Analysis

Unsupervised Mining

Intuition on interpretation of SVD in terms of genes and conditions

Genes sorted by correlation with top 2 eigengenes



Alter, Orly et al. (2000) Proc. Natl. Acad. Sci. USA 97, 10101-10106

Fig. 3. Genes sorted by relative correlation with $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ of normalized elutriation. (*a*) Normalized elutriation expression of the sorted 5,981 genes in the 14 arrays, showing traveling wave of expression. (*b*) Eigenarrays expression; the expression of $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$, the eigenarrays corresponding to $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$, displays the sorting. (*c*) Expression levels of $|\alpha_1\rangle_N$ (red) and $|\alpha_2\rangle_N$ (green) fit normalized sine and cosine functions of period $Z \equiv N - 1 = 5{,}980$ and phase $\theta \approx 2\pi/13$ (blue), respectively.

Normalized elutriation expression in the subspace associated with the cell cycle

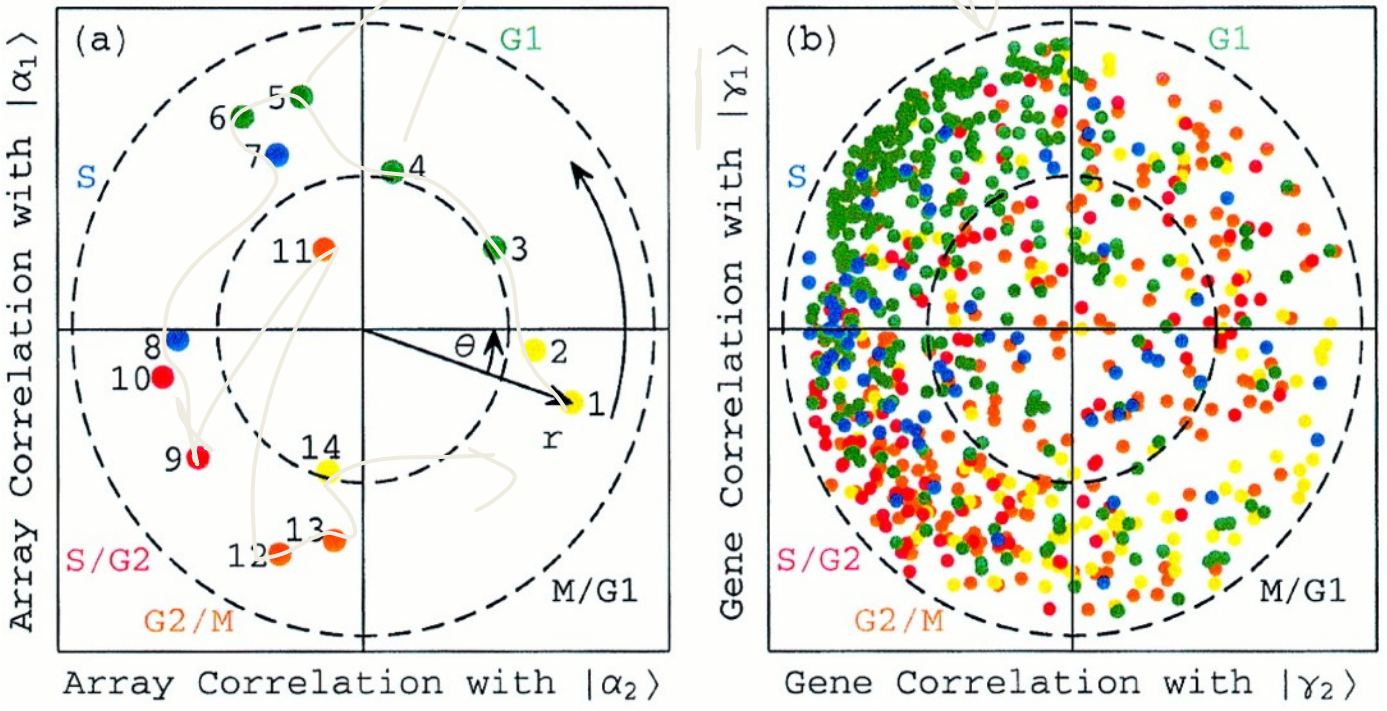


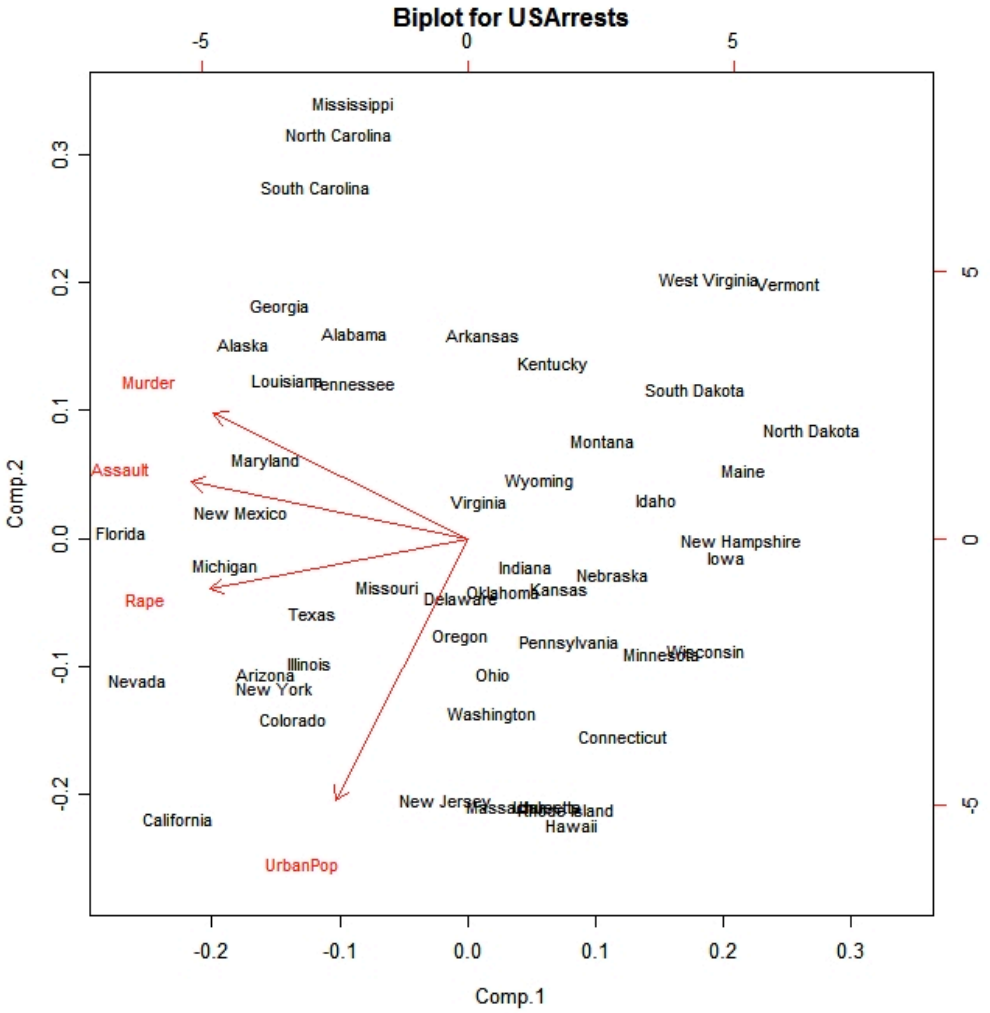
Fig. 2. Normalized elutriation expression in the subspace associated with the cell cycle. (*a*) Array correlation with $|\alpha_1\rangle_N$ along the *y*-axis vs. that with $|\alpha_2\rangle_N$ along the *x*-axis, color-coded according to the classification of the arrays into the five cell cycle stages, M/G_1 (yellow), G_1 (green), S (blue), S/G_2 (red), and G_2/M (orange). The dashed unit and half-unit circles outline 100% and 25% of overall normalized array expression in the $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ subspace. (*b*) Correlation of each gene with $|\gamma_1\rangle_N$ vs. that with $|\gamma_2\rangle_N$, for 784 cell cycle regulated genes, color-coded according to the classification by Spellman *et al.* (3).

Alter, Orly et al. (2000) Proc. Natl. Acad. Sci. USA 97, 10101-10106

Unsupervised Mining

Biplot

Introduction



- A biplot is a low-dimensional (usually 2D) representation of a data matrix **A**.
  - A point for each of the *m* observation vectors (rows of **A**)
  - A line (or arrow) for each of the *n* variables (columns of **A**)

PCA

TFs: a, b, c...

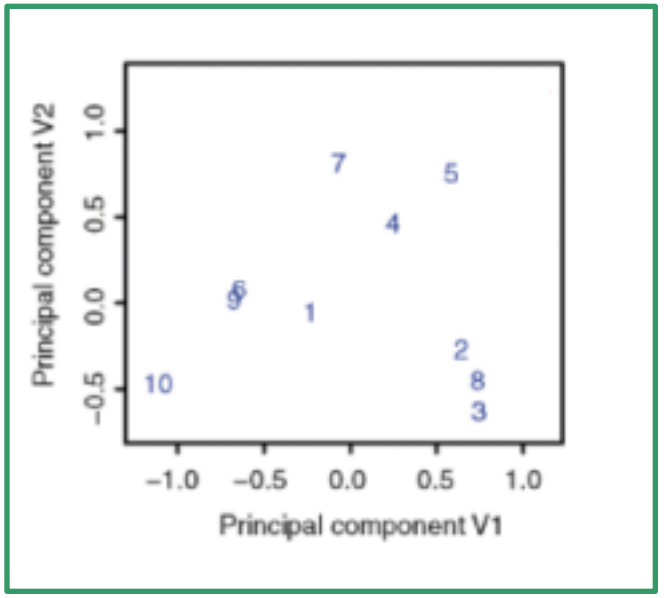
Genomic Sites: 1,2,3...

A

	a	b	c
1	21	16	28
2	14	18	25
3	14	17	22
4	14	19	33
5	17	23	28
6	20	14	34
7	22	21	30
8	15	18	22
9	18	13	36
10	24	10	32

	a	b	c
a	1.00	-0.44	0.48
b	-0.44	1.00	-0.40
c	0.48	-0.40	1.00

A^TA (TF-TF corr.)

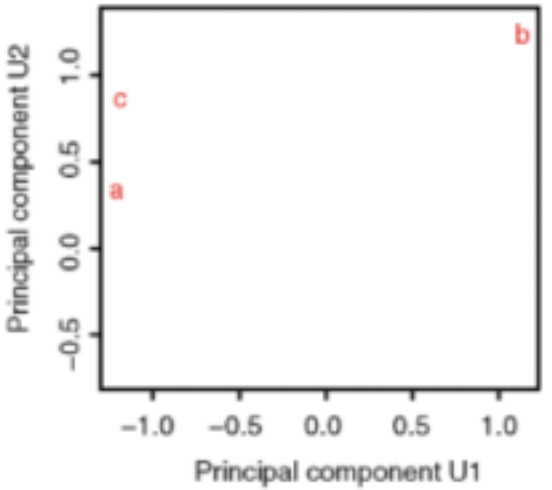


A^T

	1	2	3	4	5	6	7	8	9	10
a	21	14	14	14	17	20	22	15	18	24
b	16	18	17	19	23	14	21	18	13	10
c	28	25	22	33	28	34	30	22	36	32

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.70	0.69	0.77	0.54	0.99	0.95	0.65	0.98	0.97
2	0.70	1.00	1.00	0.99	0.98	0.79	0.89	1.00	0.84	0.50
3	0.69	1.00	1.00	0.99	0.98	0.78	0.89	1.00	0.83	0.49
4	0.77	0.99	0.99	1.00	0.95	0.85	0.94	0.98	0.89	0.59
5	0.54	0.98	0.98	0.95	1.00	0.64	0.78	0.99	0.71	0.31
6	0.99	0.79	0.78	0.85	0.64	1.00	0.98	0.74	1.00	0.93
7	0.95	0.89	0.89	0.94	0.78	0.98	1.00	0.86	0.99	0.84
8	0.65	1.00	1.00	0.98	0.99	0.74	0.86	1.00	0.80	0.43
9	0.98	0.84	0.83	0.89	0.71	1.00	0.99	0.80	1.00	0.89
10	0.97	0.50	0.49	0.59	0.31	0.93	0.84	0.43	0.89	1.00

AA^T (site-site correlation)



Biplot to Show Overall Relationship of TFs & Sites

TFs: a, b, c...

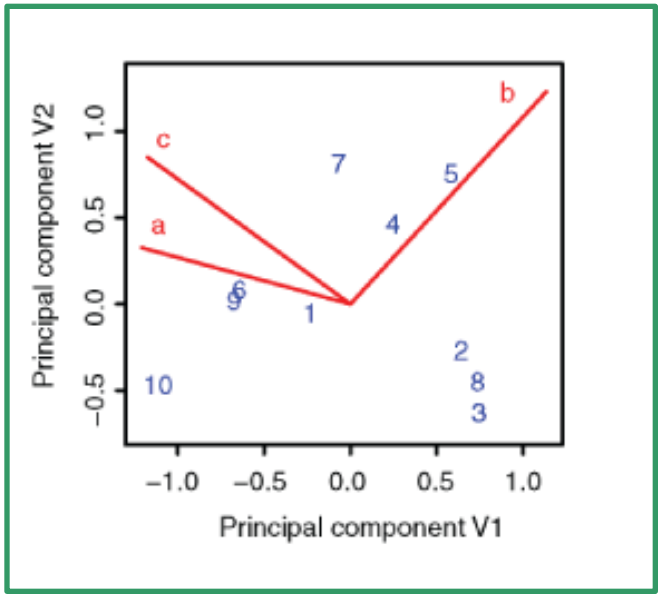
Genomic Sites: 1,2,3...

$A=USV^T$

	a	b	c
1	21	16	28
2	14	18	25
3	14	17	22
4	14	19	33
5	17	23	28
6	20	14	34
7	22	21	30
8	15	18	22
9	18	13	36
10	24	10	32

	a	b	c
a	1.00	-0.44	0.48
b	-0.44	1.00	-0.40
c	0.48	-0.40	1.00

A^TA (TF-TF corr.)

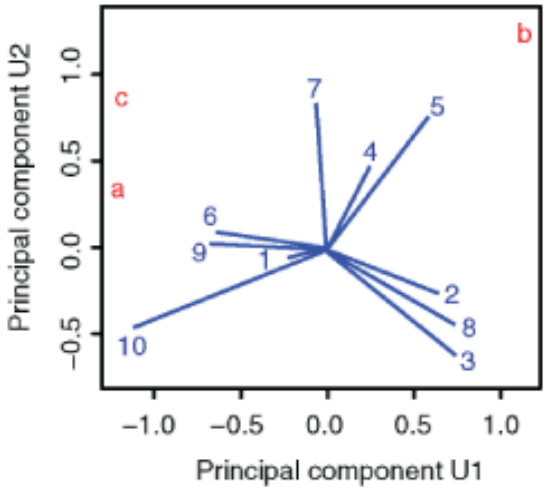


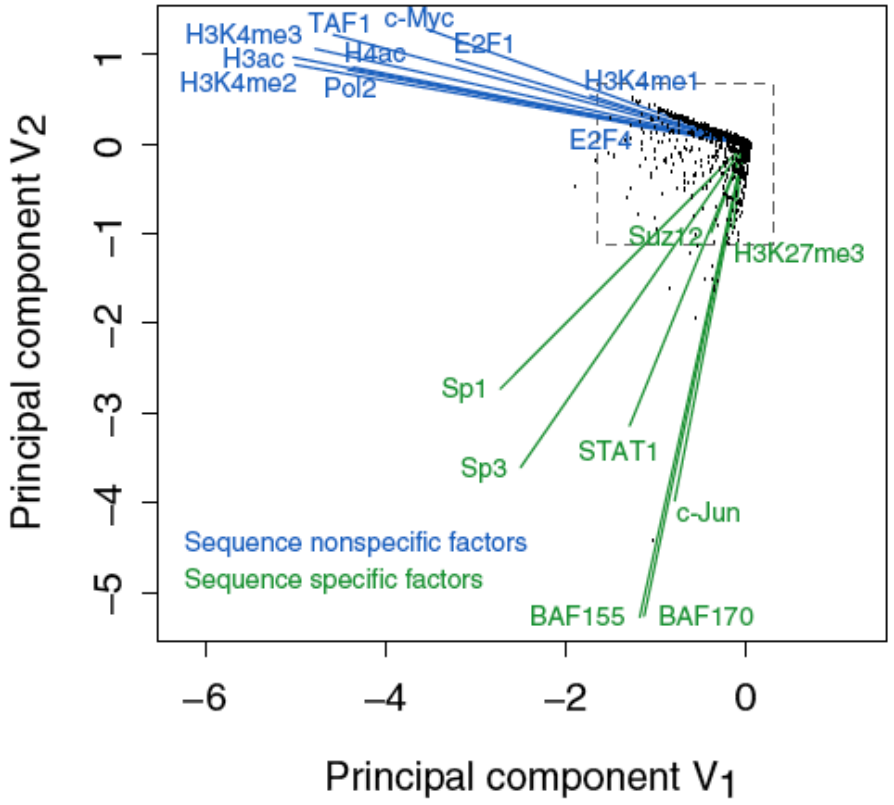
A^T

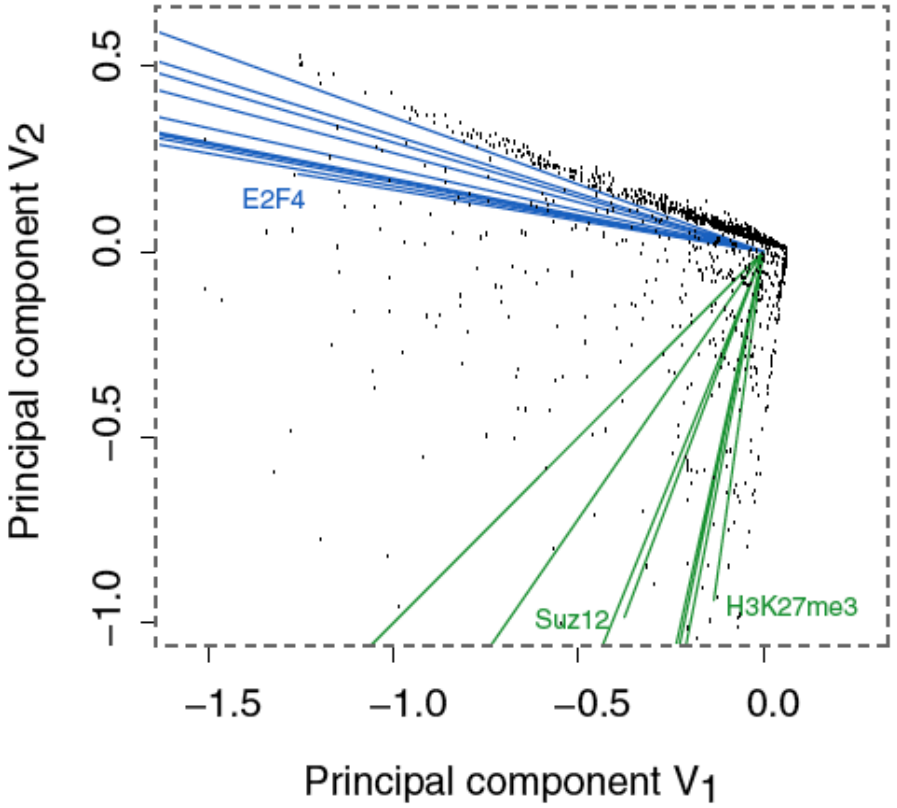
	1	2	3	4	5	6	7	8	9	10
a	21	14	14	14	17	20	22	15	18	24
b	16	18	17	19	23	14	21	18	13	10
c	28	25	22	33	28	34	30	22	36	32

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.70	0.69	0.77	0.54	0.99	0.95	0.65	0.98	0.97
2	0.70	1.00	1.00	0.99	0.98	0.79	0.89	1.00	0.84	0.50
3	0.69	1.00	1.00	0.99	0.98	0.78	0.89	1.00	0.83	0.49
4	0.77	0.99	0.99	1.00	0.95	0.85	0.94	0.98	0.89	0.59
5	0.54	0.98	0.98	0.95	1.00	0.64	0.78	0.99	0.71	0.31
6	0.99	0.79	0.78	0.85	0.64	1.00	0.98	0.74	1.00	0.93
7	0.95	0.89	0.89	0.94	0.78	0.98	1.00	0.86	0.99	0.84
8	0.65	1.00	1.00	0.98	0.99	0.74	0.86	1.00	0.80	0.43
9	0.98	0.84	0.83	0.89	0.71	1.00	0.99	0.80	1.00	0.89
10	0.97	0.50	0.49	0.59	0.31	0.93	0.84	0.43	0.89	1.00

AA^T (site-site correlation)



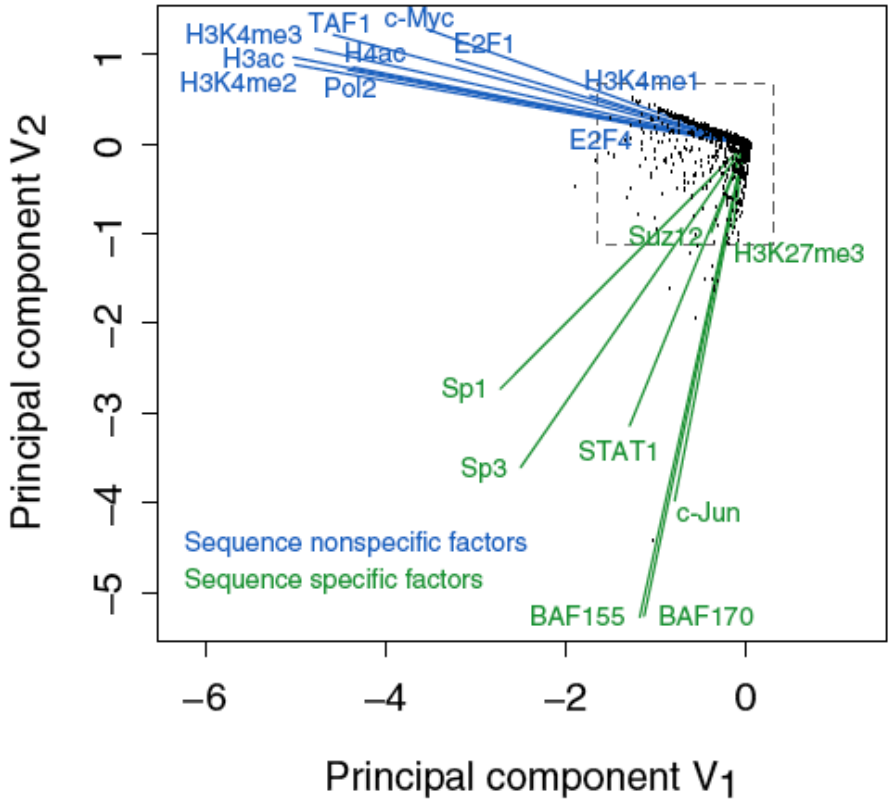


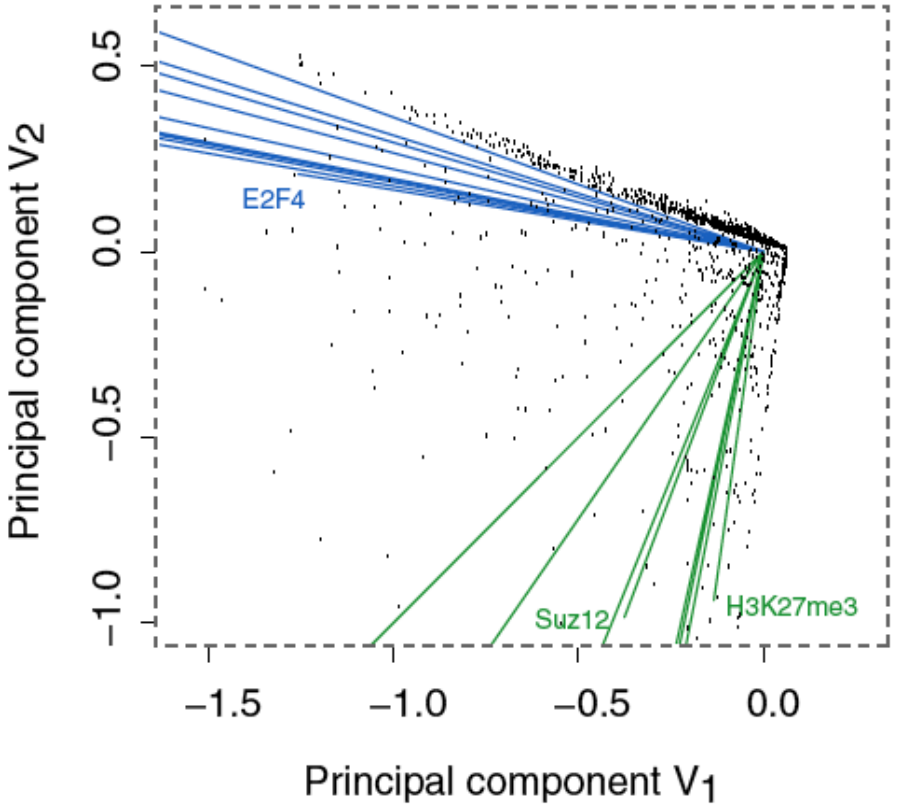


Results of Biplot

- Pilot ENCODE (1% genome): 5996 10 kb genomic bins (adding all hits) + 105 TF experiments → biplot
- Angle between TF vectors shows relation b/w factors
- Closeness of points gives clustering of "sites"
- Projection of site onto vector gives degree to which site is assoc. with a particular factor

Zhang et al. (2007) Gen. Res.





Results of Biplot

- Biplot groups TFs into sequence-specific and sequence-nonspecific clusters.
  - c-Myc may behave more like a sequence-nonspecific TF.
  - H3K27me3 functions in a transcriptional regulatory process in a rather sequence-specific manner.
- Genomic Bins are associated with different TFs and in this fashion each bin is "annotated" by closest TF cluster

Zhang et al. (2007) Gen. Res.

Unsupervised Mining

CCA

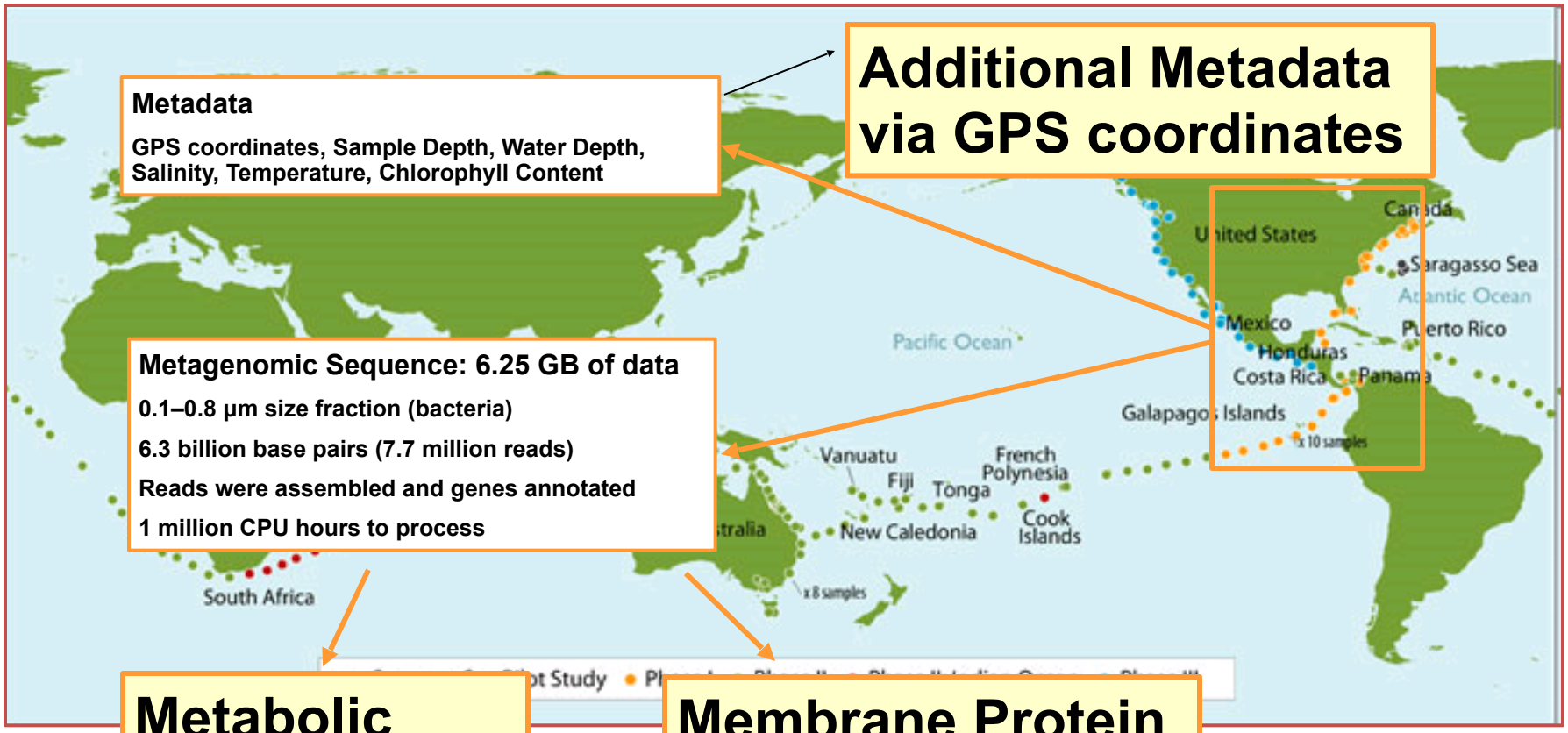
Sorcerer II Global Ocean Survey

***Sorcerer II* journey August 2003- January 2006**

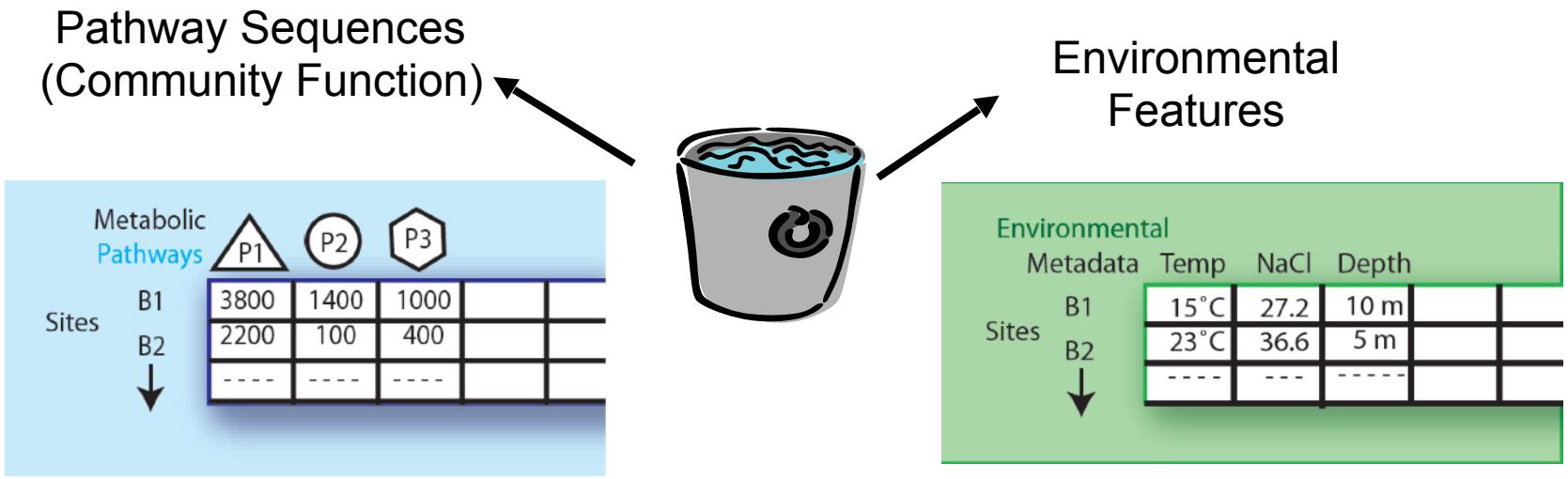
Sample approximately every 200 miles

Rusch, et al., *PLOS Biology* 2007

Sorcerer II Global Ocean Survey



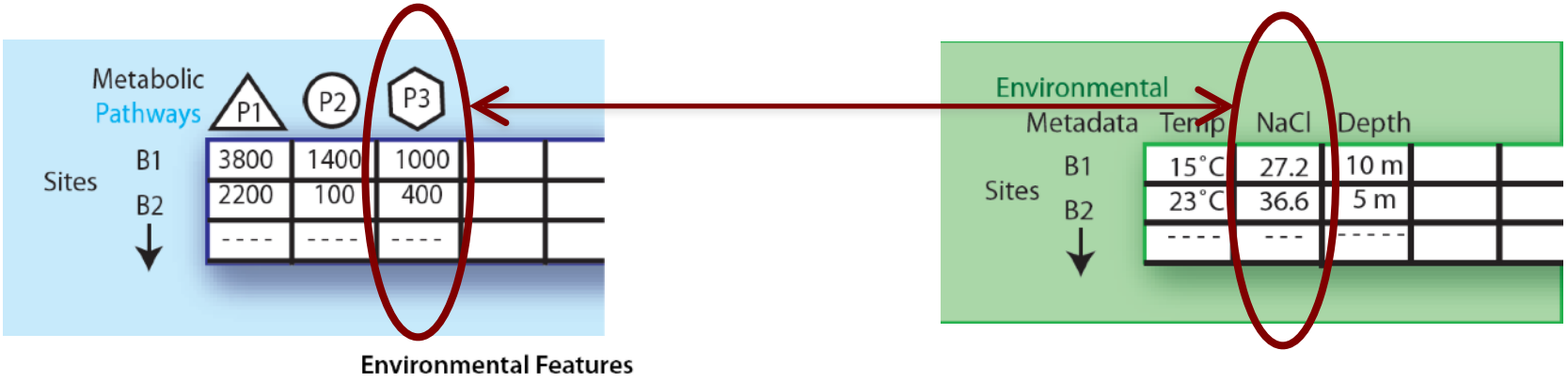
Rusch, et al., *PLOS Biology* 2007



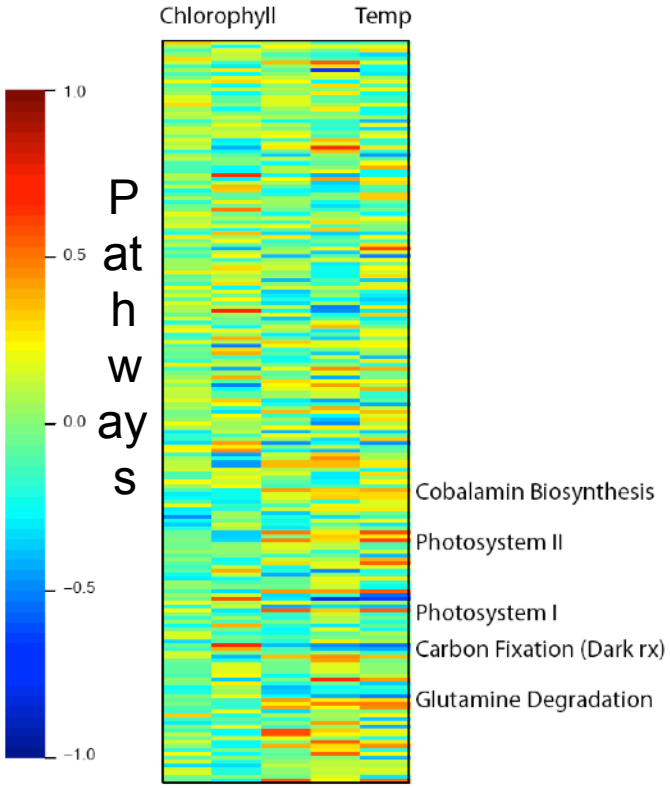
Expressing data as matrices indexed by site, env. var., and pathway usage

[Rusch et. al., (2007) PLOS Biology;
Gianoulis et al., PNAS (in press, 2009]

Simple Relationships: Pairwise Correlations

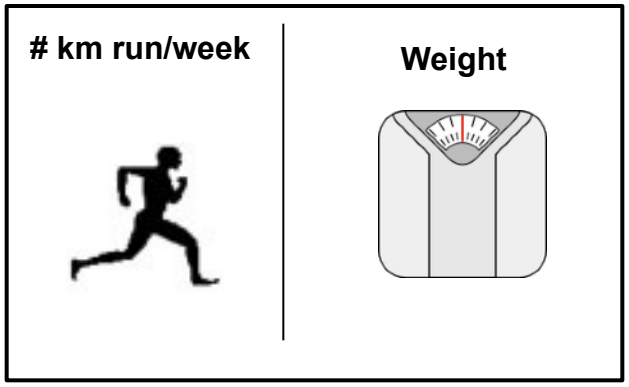


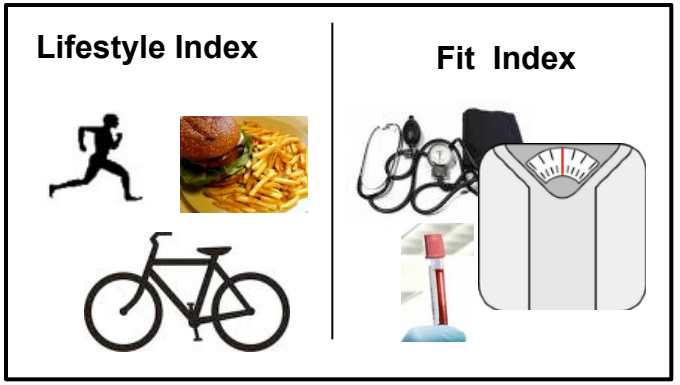
Environmental Features



[Gianoulis et al., PNAS (in press, 2009)]

Canonical Correlation Analysis: Simultaneous weighting

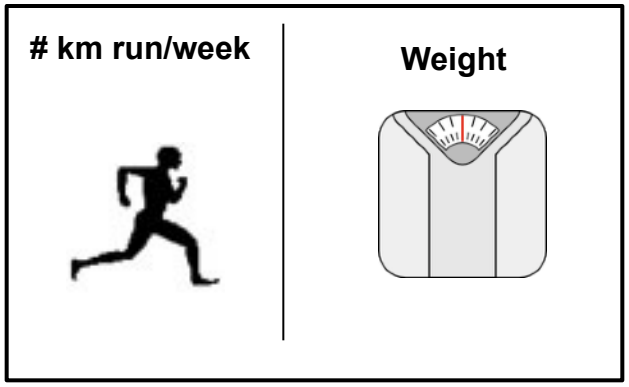


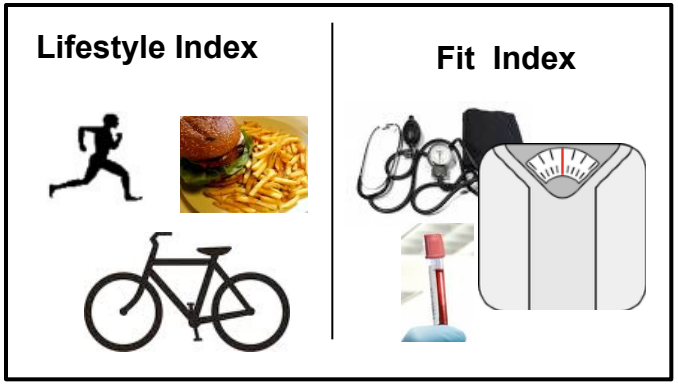


Lifestyle Index = a  +b  + c 

Fit Index = a  + b  +c 

Canonical Correlation Analysis: Simultaneous weighting



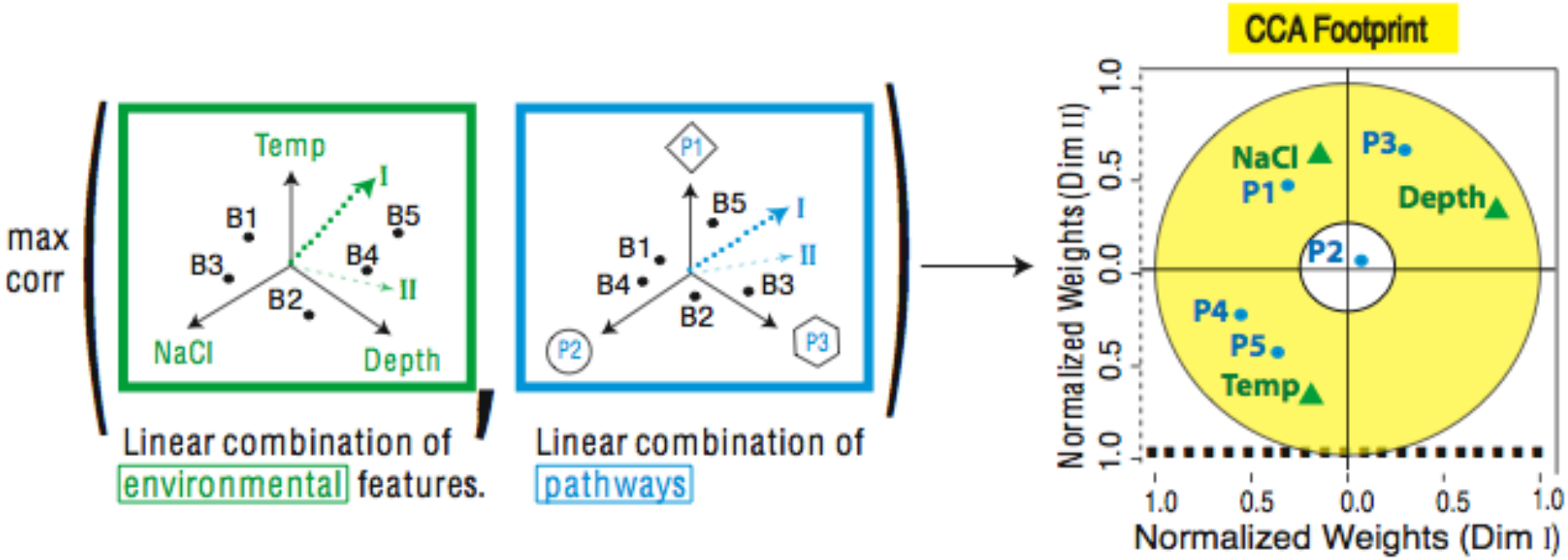


Life

Fit

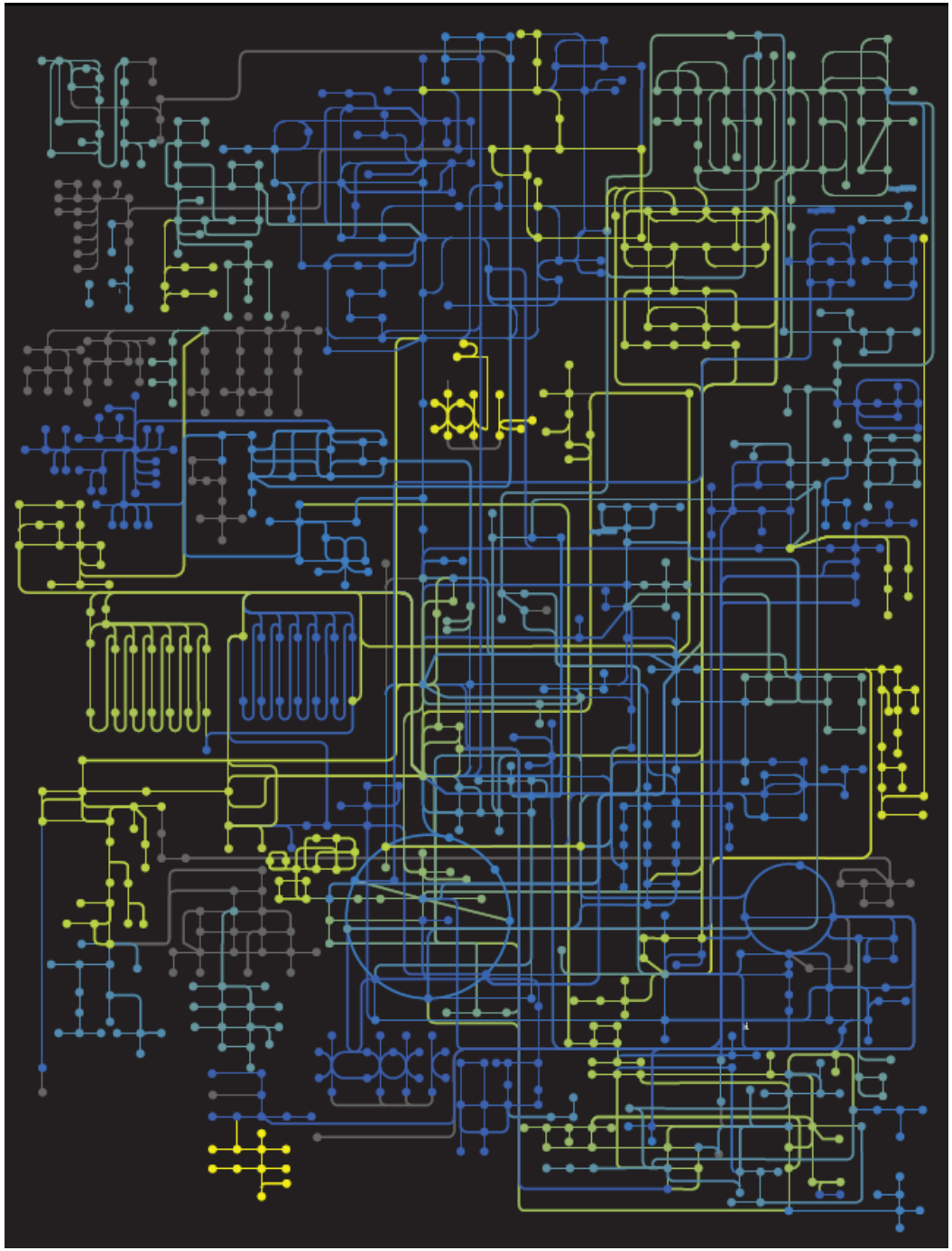
Environmental Features	Metabolic Pathways/ Protein Families
Temp etc	Photosynthesis etc
Chlorophyll	Lipid Metabolism

CCA: Finding Variables with Large Projections in "Correlation Circle"

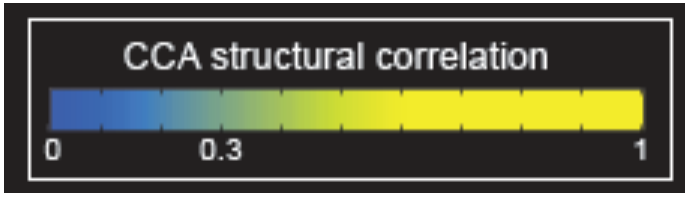


The goal of this technique is to interpret cross-variance matrices
We do this by defining a change of basis.

Gianoulis et al., PNAS 2009

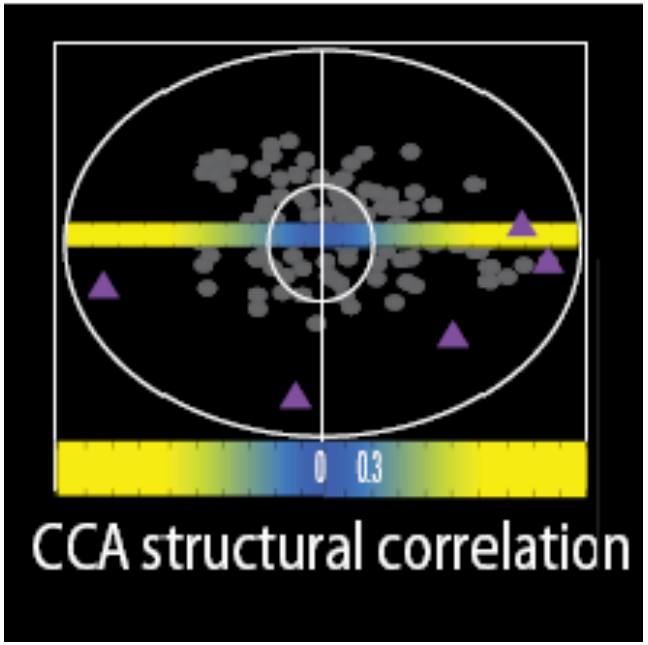


Strength of Pathway co-variation with environment



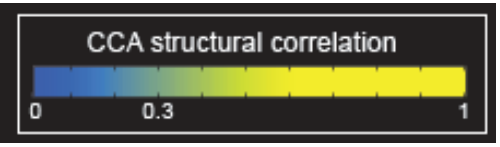
Environmentally invariant

Environmentally variant



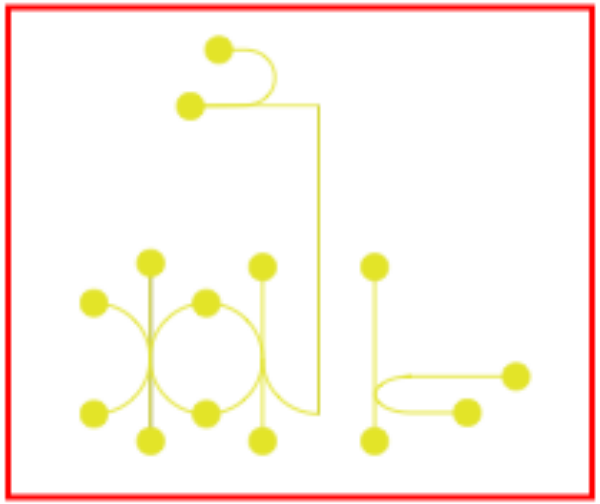
Gianoulis et al., *PNAS* 2009

Conclusion #1: energy conversion strategy, temp and depth

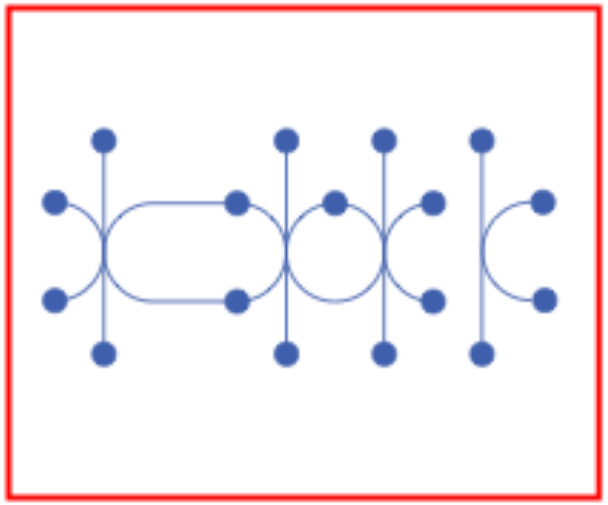


KEGG

Photosynthesis



Oxidative Phosphorylation



Gianoulis et al., *PNAS* 2009