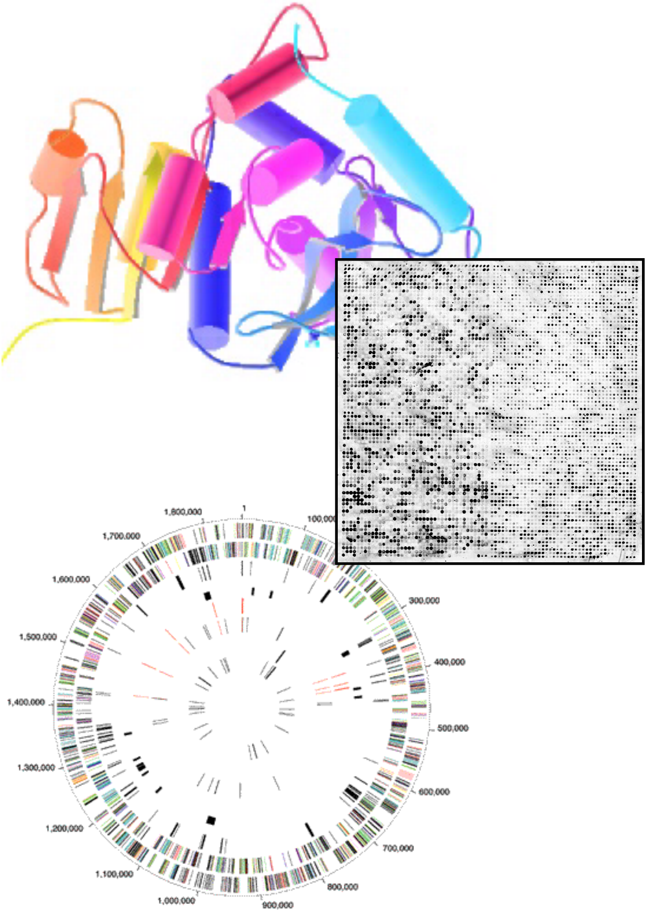
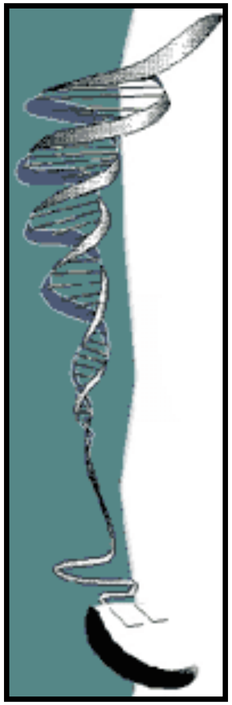
Biomed. Data Science:

Unsupervised Datamining



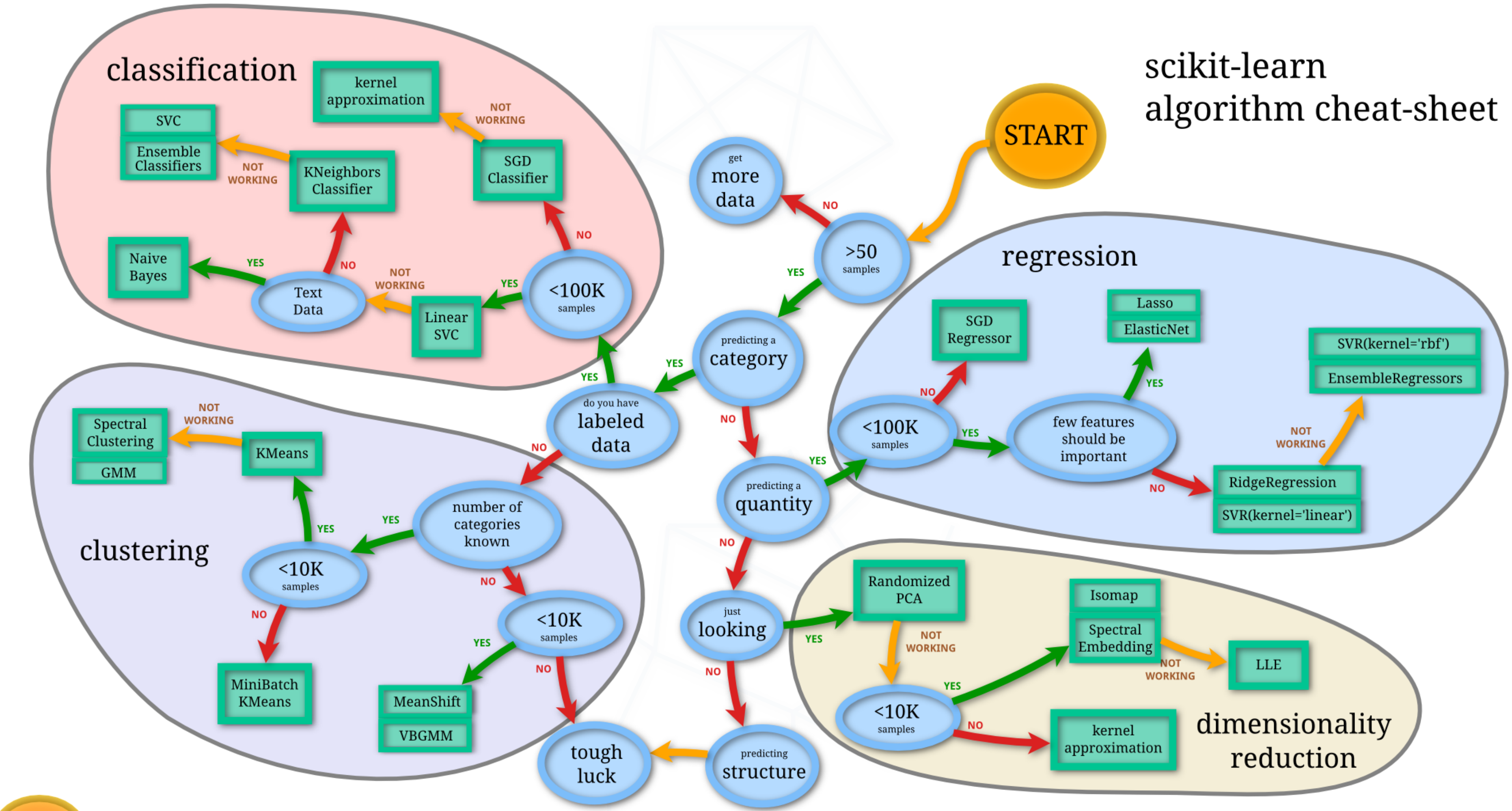


Mark Gerstein, Yale University

gersteinlab.org/courses/452

(last edit in spring ’18)

The World of Machine Learning

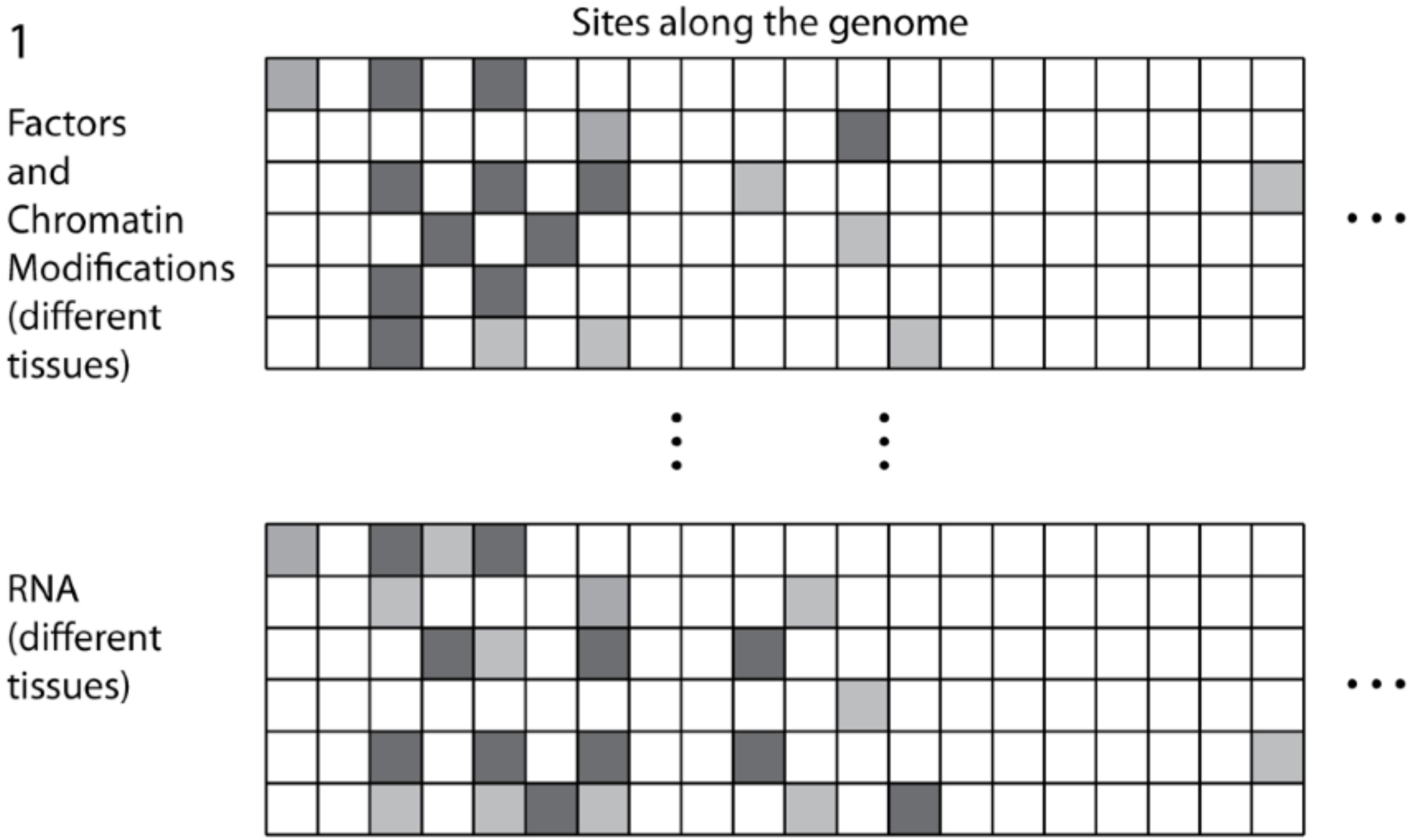


Back

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

Abstract Overview: Supervised vs Unsupervised Mining

Structure of Genomic Features Matrix



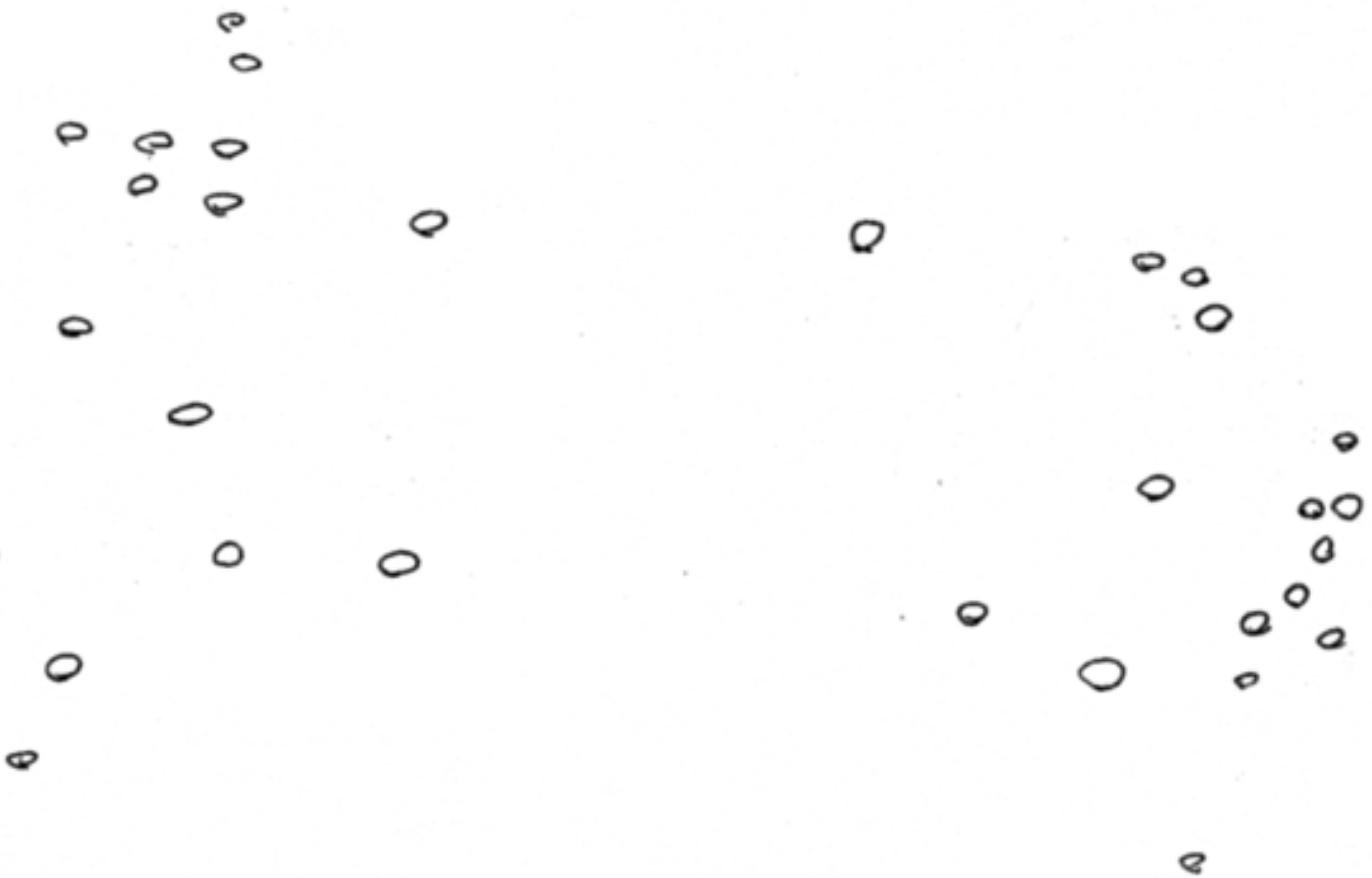
1

Sites along the genome

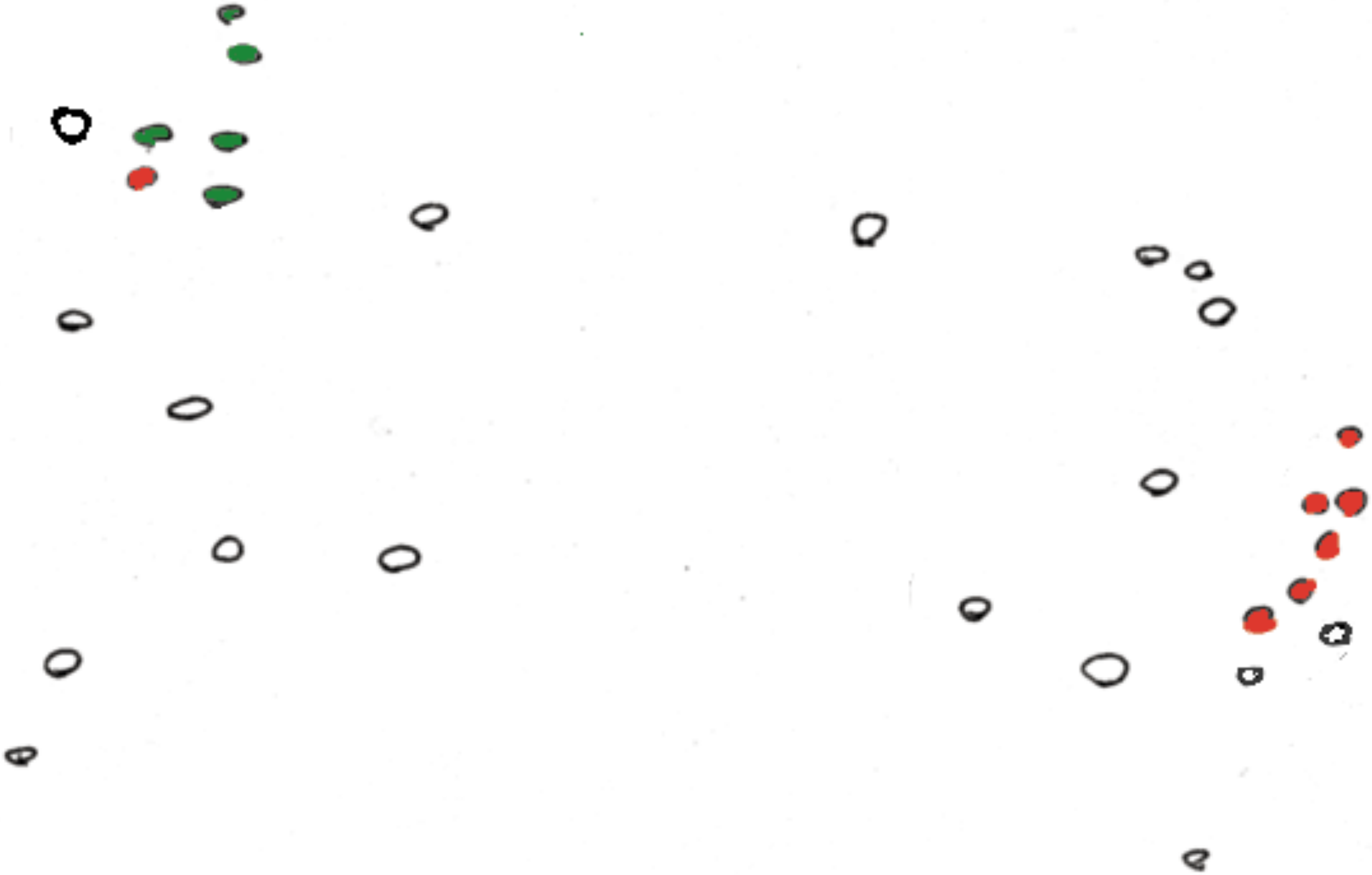
Factors and Chromatin Modifications (different tissues)

RNA (different tissues)

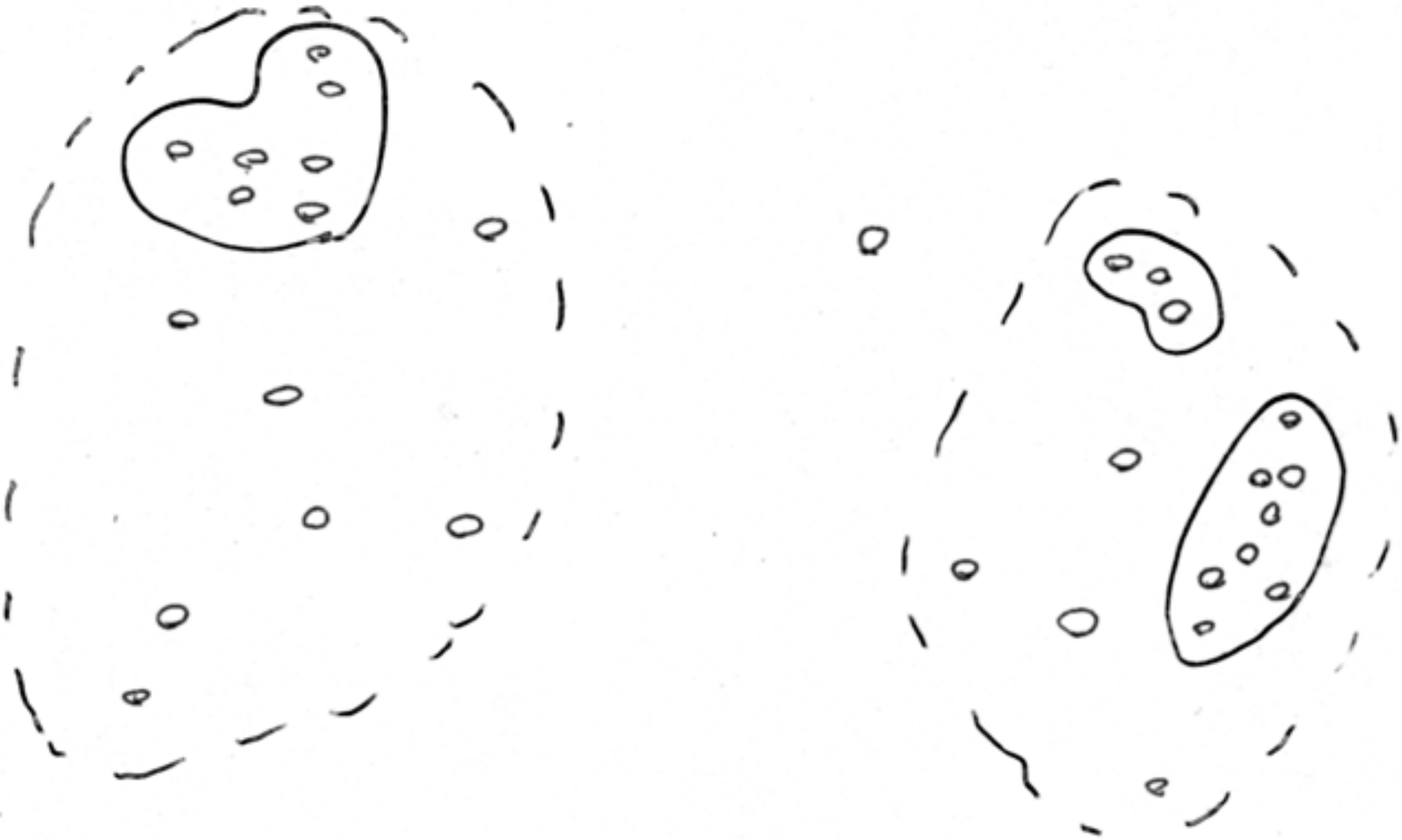
Represent predictors in abstract high dimensional space



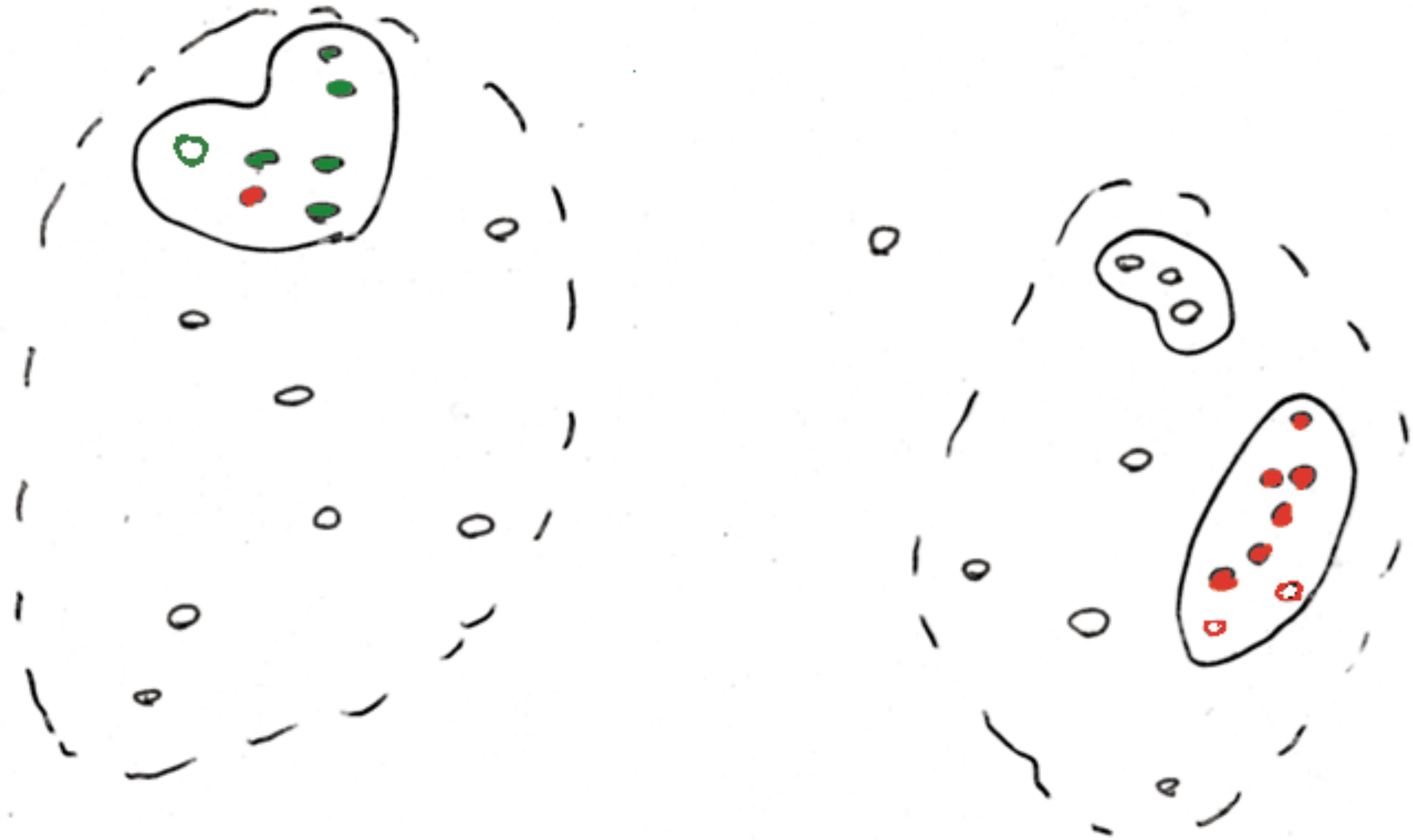
“Label” Certain Points



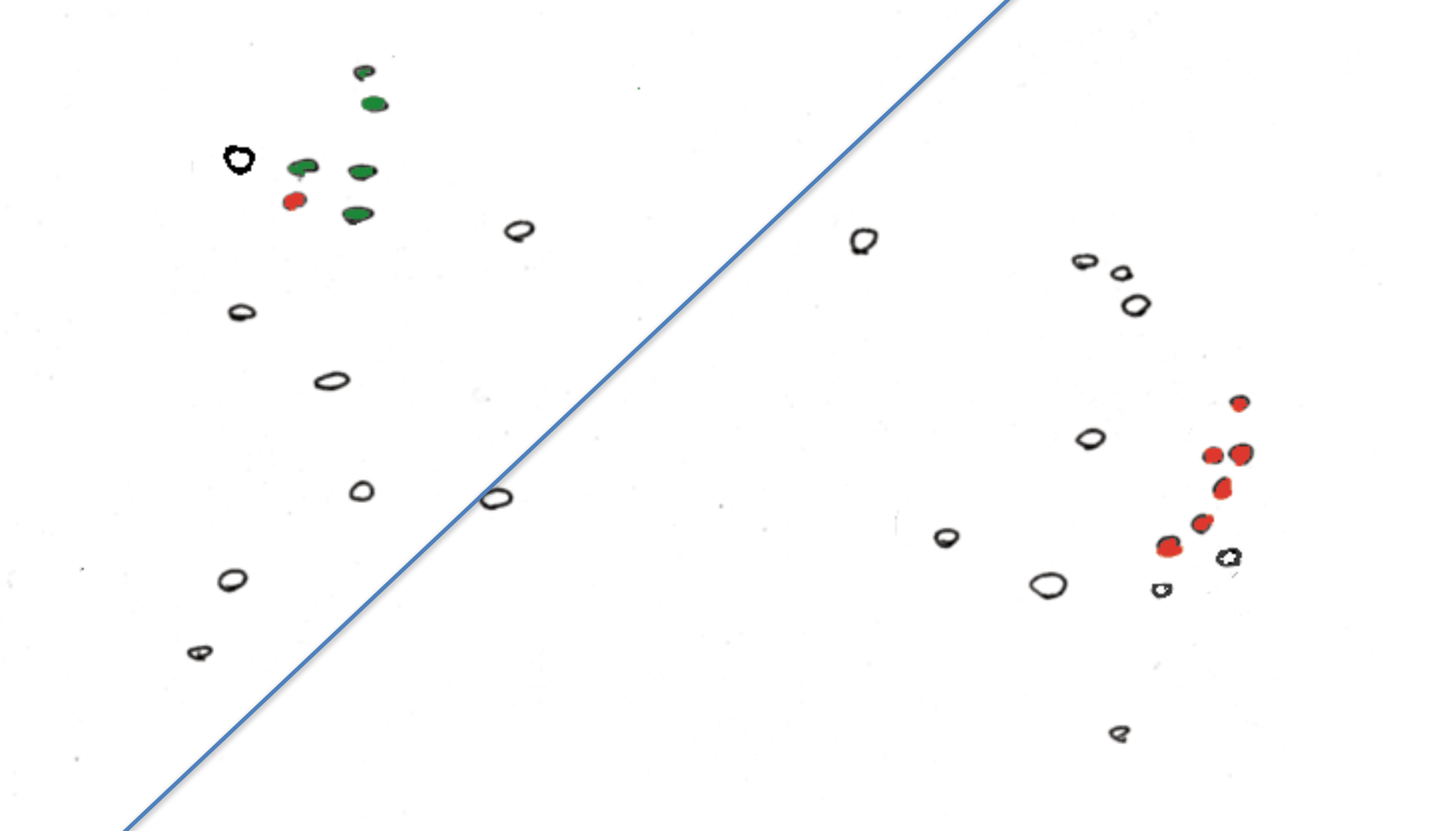
“Cluster” predictors (Unsupervised)



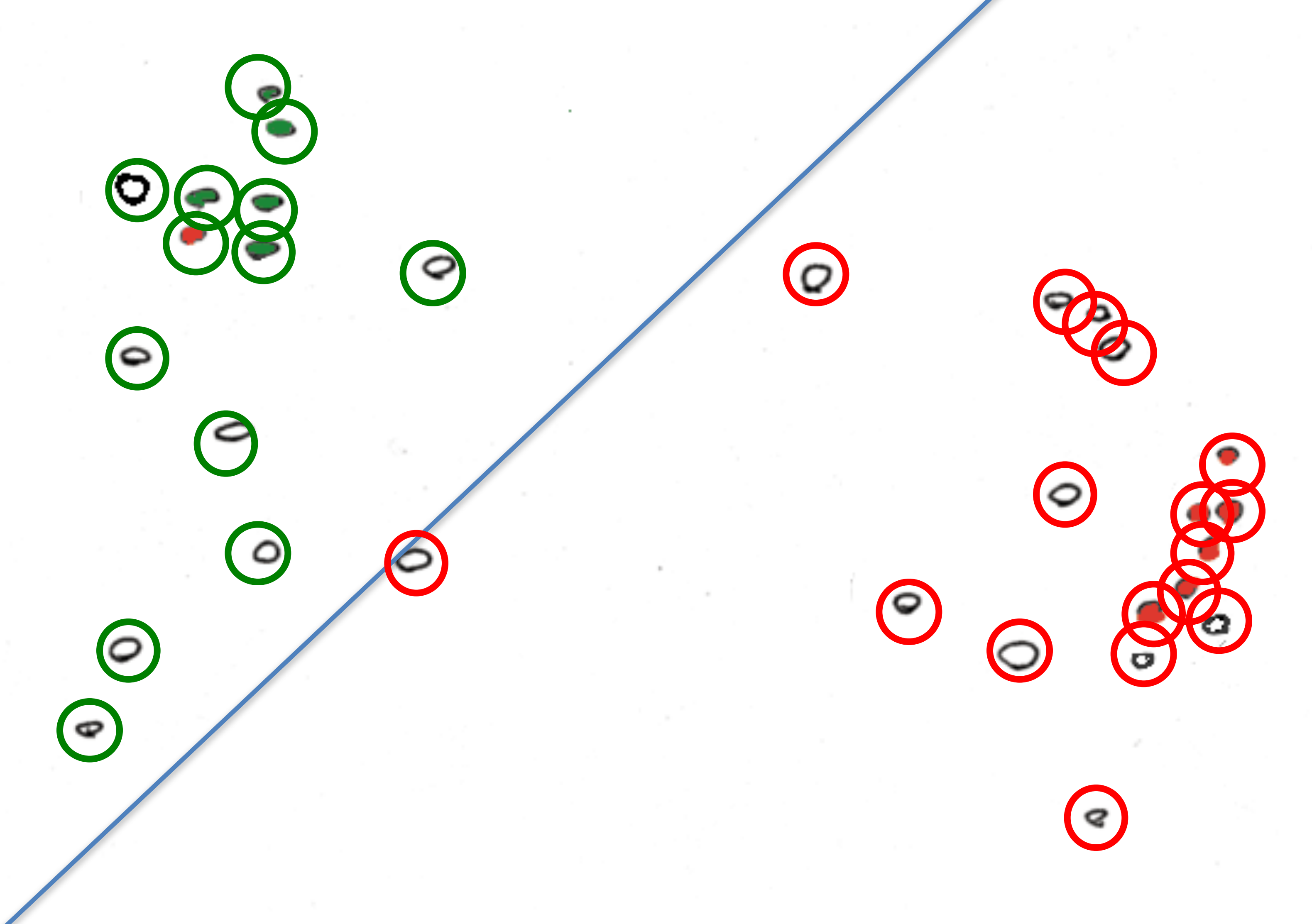
Use Clusters to predict Response (Unsupervised, guilt-by-association)



Develop Separator Based on Labeled Points (Supervised)



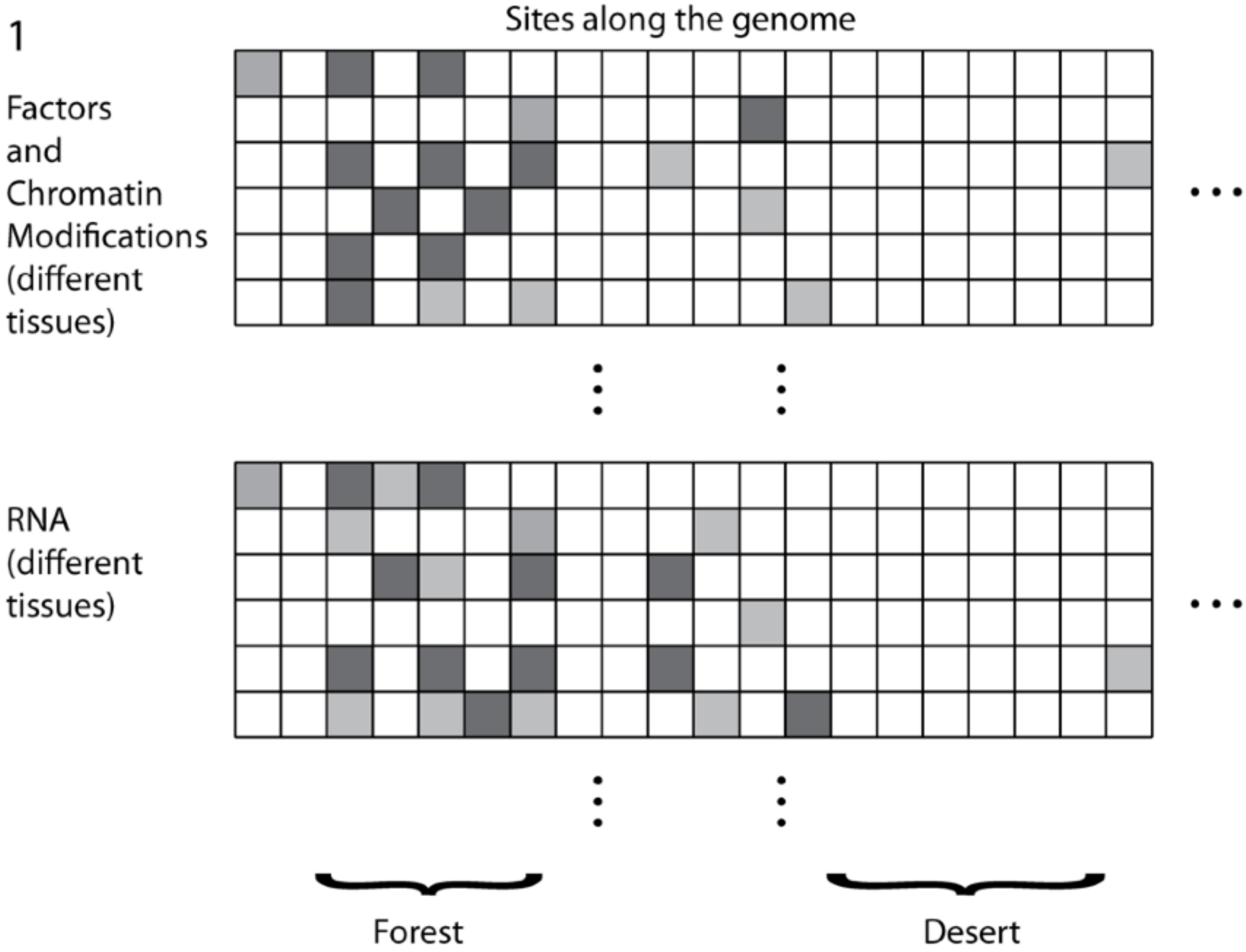
Predict based on Separator (Supervised)



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



Sites along the genome

1

Factors and Chromatin Modifications (different tissues)

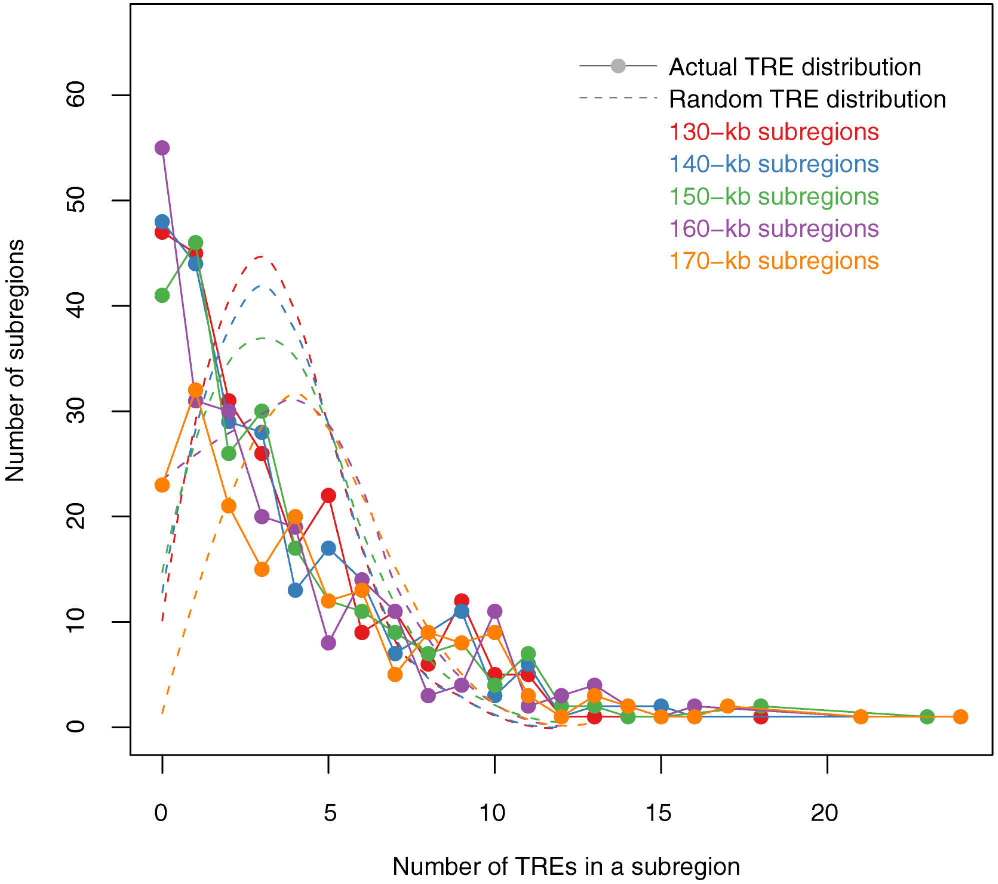
RNA (different tissues)

Forest

Desert

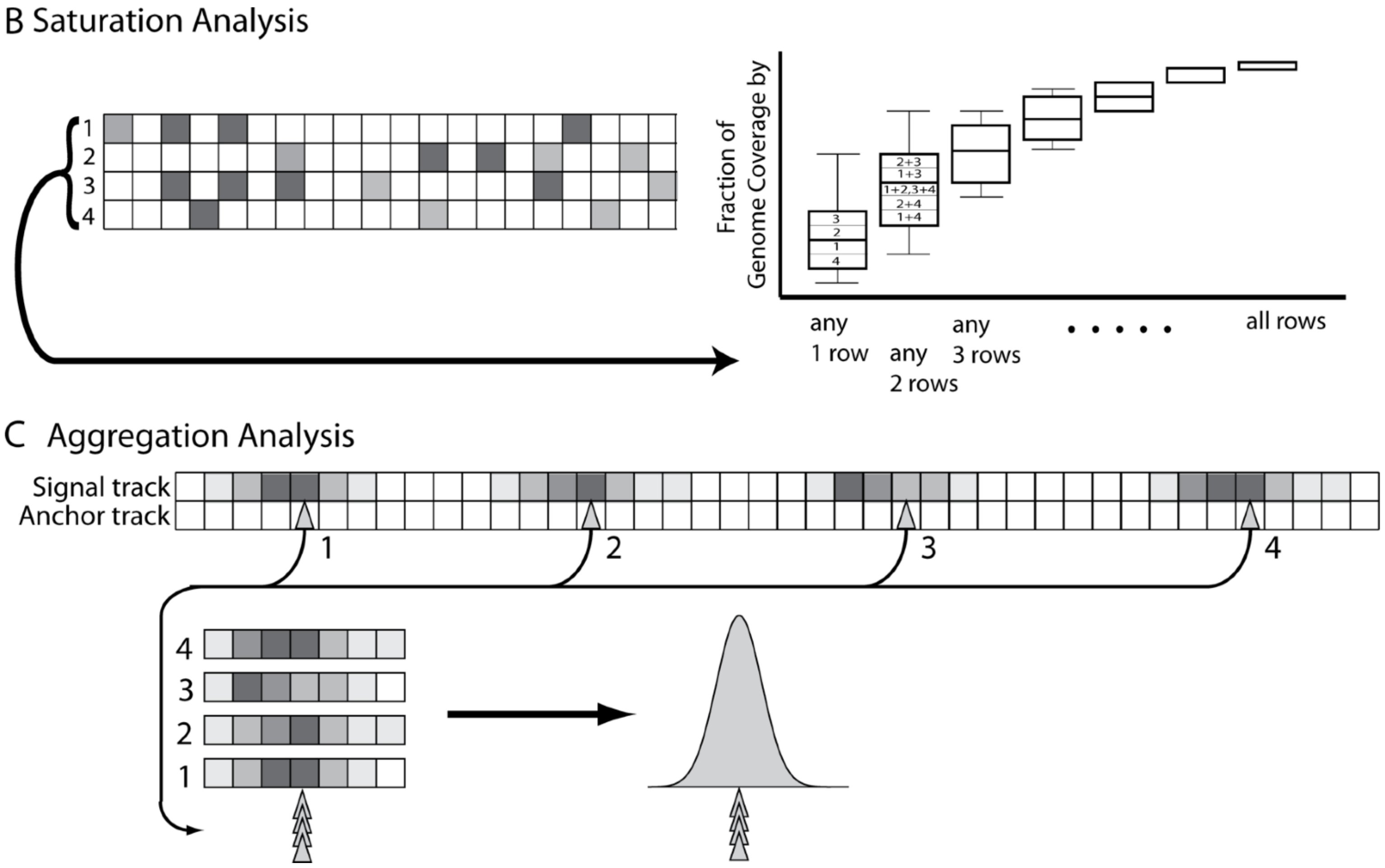
Non-random distribution of TREs

- TREs are not evenly distributed throughout the encode regions ($P < 2.2 \times 10^{-16}$).
- 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

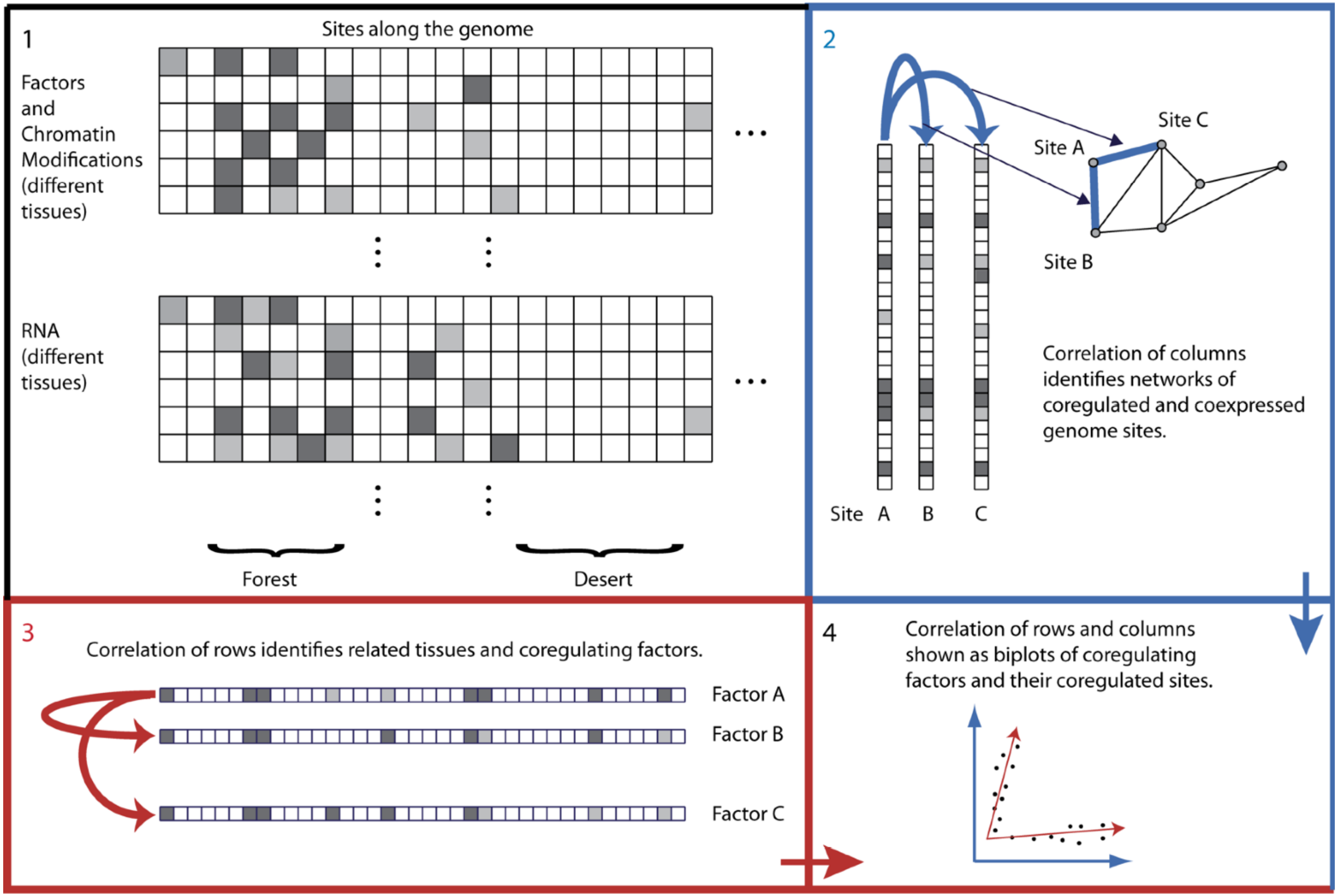


[*Nat. Rev. Genet.* (2010) 11: 559]

Unsupervised Mining

Clustering Columns & Rows of the Data Matrix

Correlating Rows & Columns



Correlation of columns identifies networks of coregulated and coexpressed genome sites.

Correlation of rows identifies related tissues and coregulating factors.

Correlation of rows and columns shown as biplots of coregulating factors and their coregulated sites.

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

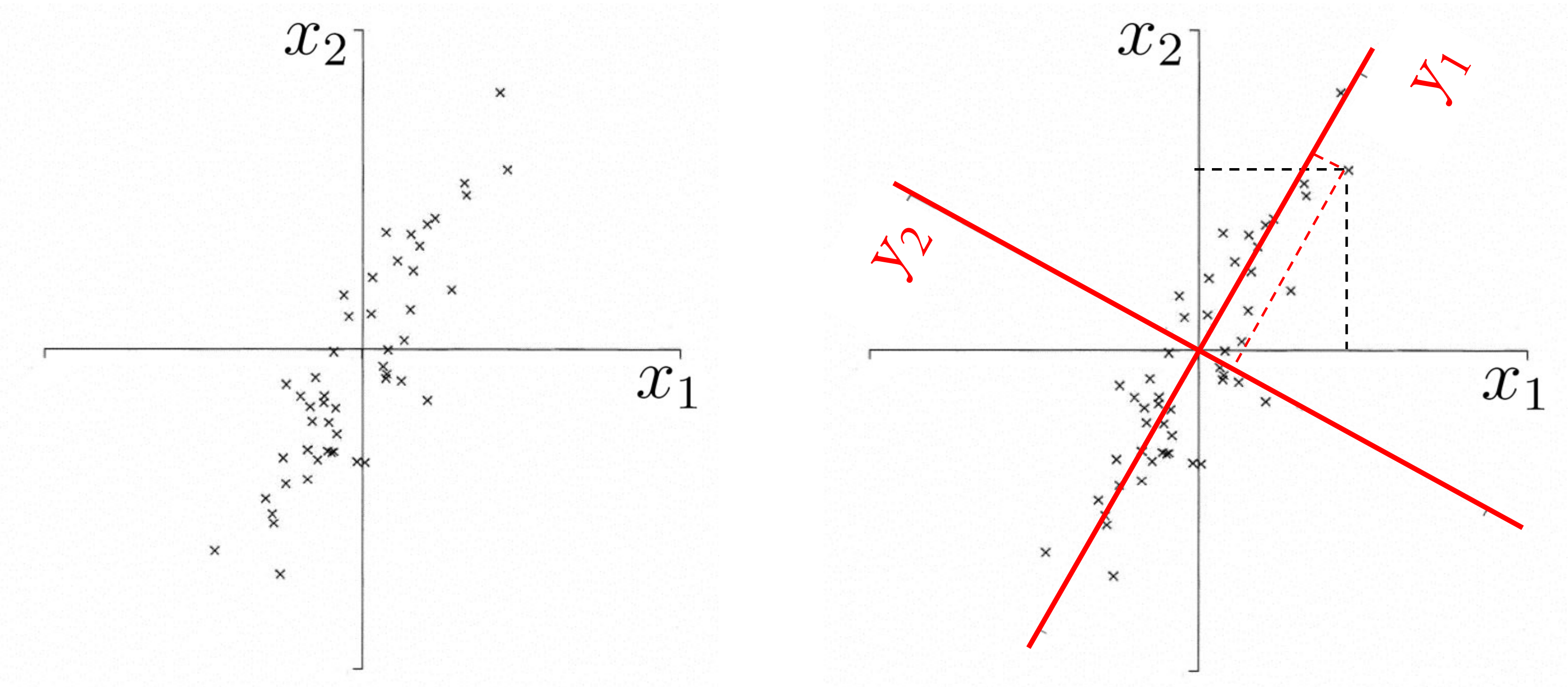
Spectral Methods Outline & Papers

- 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.

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

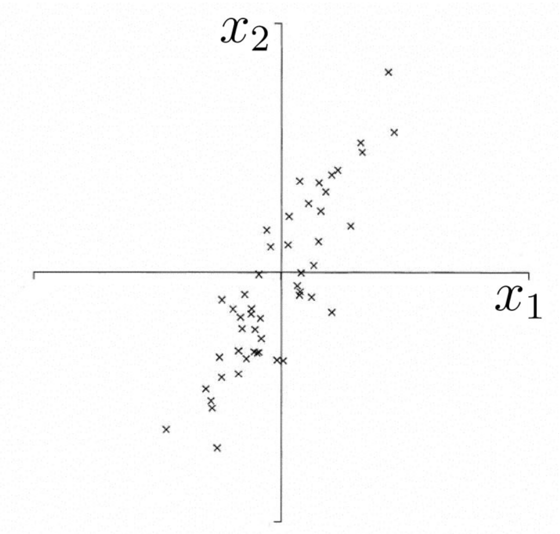


Adapted from http://www.astro.princeton.edu/~gk/A542/PCA.ppt

PCA Matrix

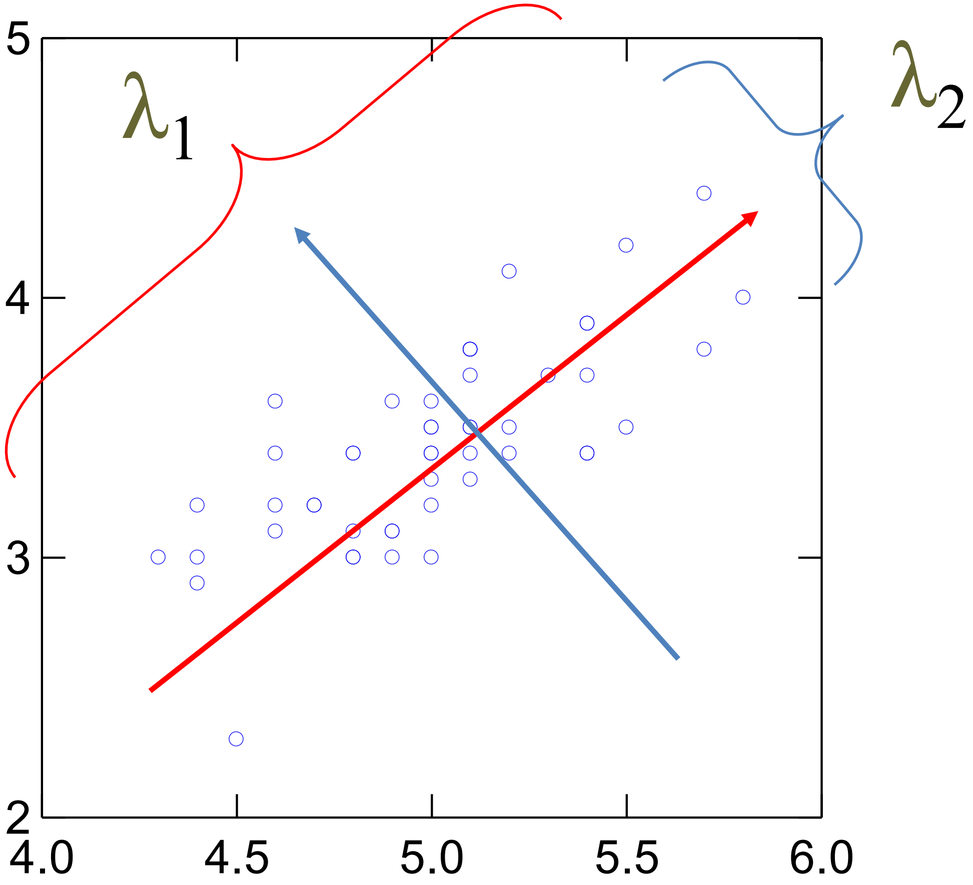
1. Start with dataset of k variables X = x_1, x_2 ... x_k and n observations.
2. Construct **covariance or correlation matrix** for variables.
3. **The Eigenvalue Problem** or Eigenanalysis: matrix diagonalization and solve for eigenvalues and eigenvectors

E.g. Start with a bunch of coordinates



Observations	X1	X2
1	2	5
2	5	6
3	4	2
4	3	7
5	9	-5
…		
n	-5	-8

Interpretation: Eigenvalues & Eigenvectors



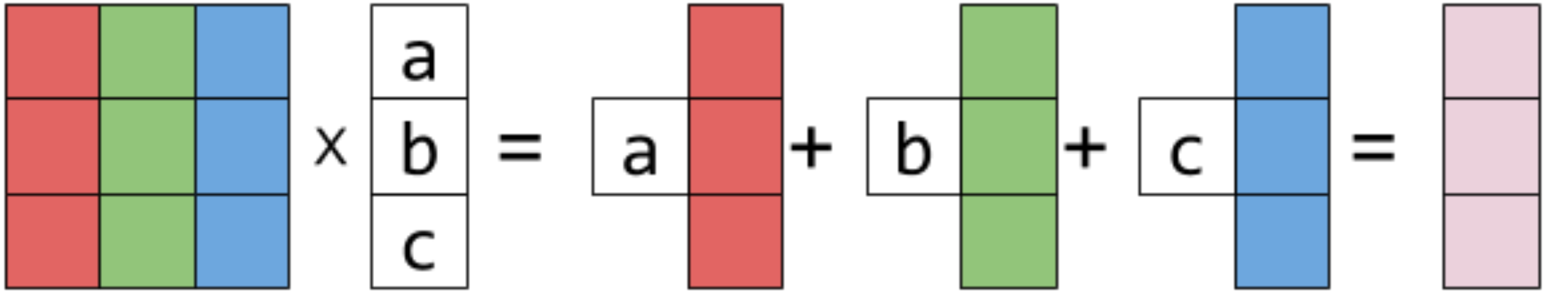
Adapted from http://myweb.dal.ca/~hwhitehe/BIOL4062/pca.ppt

Quick Refresher on Matrices

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

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

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



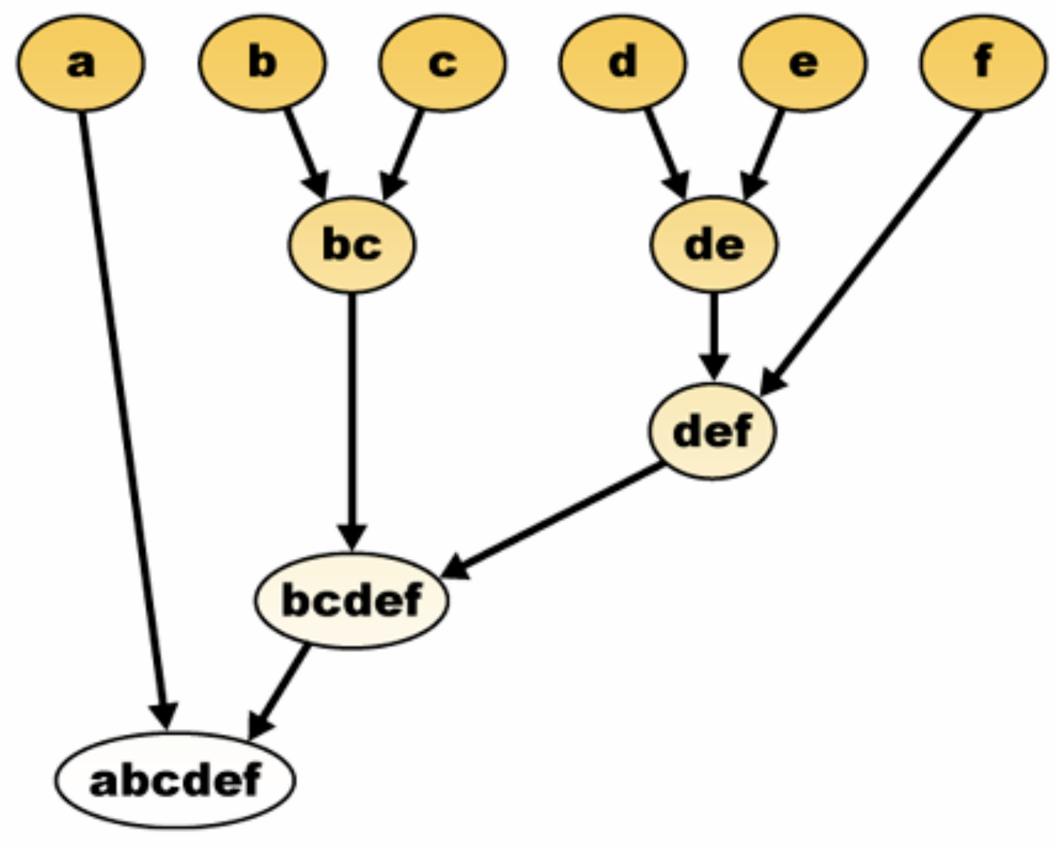
http://eli.thegreenplace.net/2015/visualizing-matrix-multiplication-as-a-linear-combination/

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

Expression Clustering

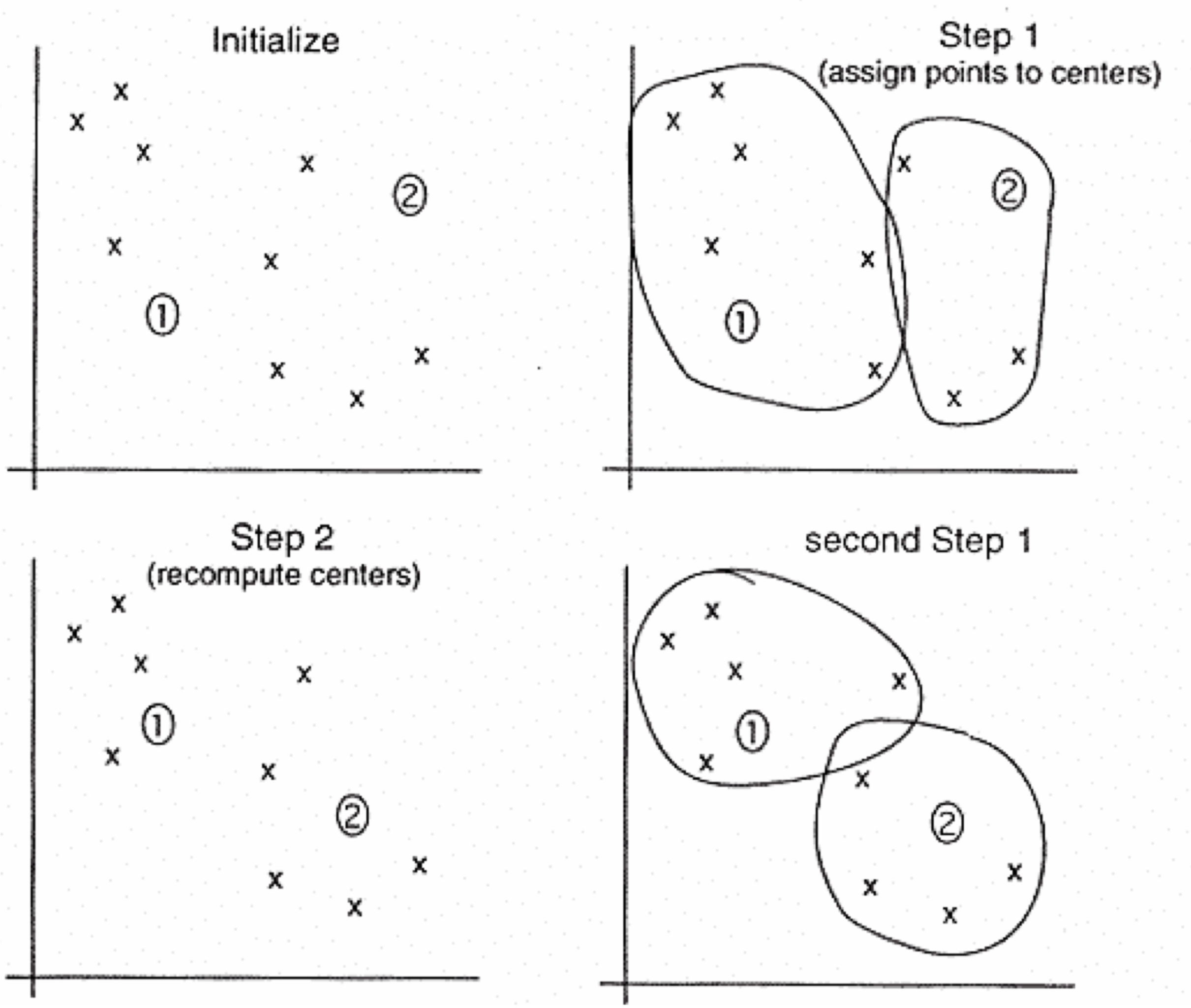
Agglomerative Clustering

- 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

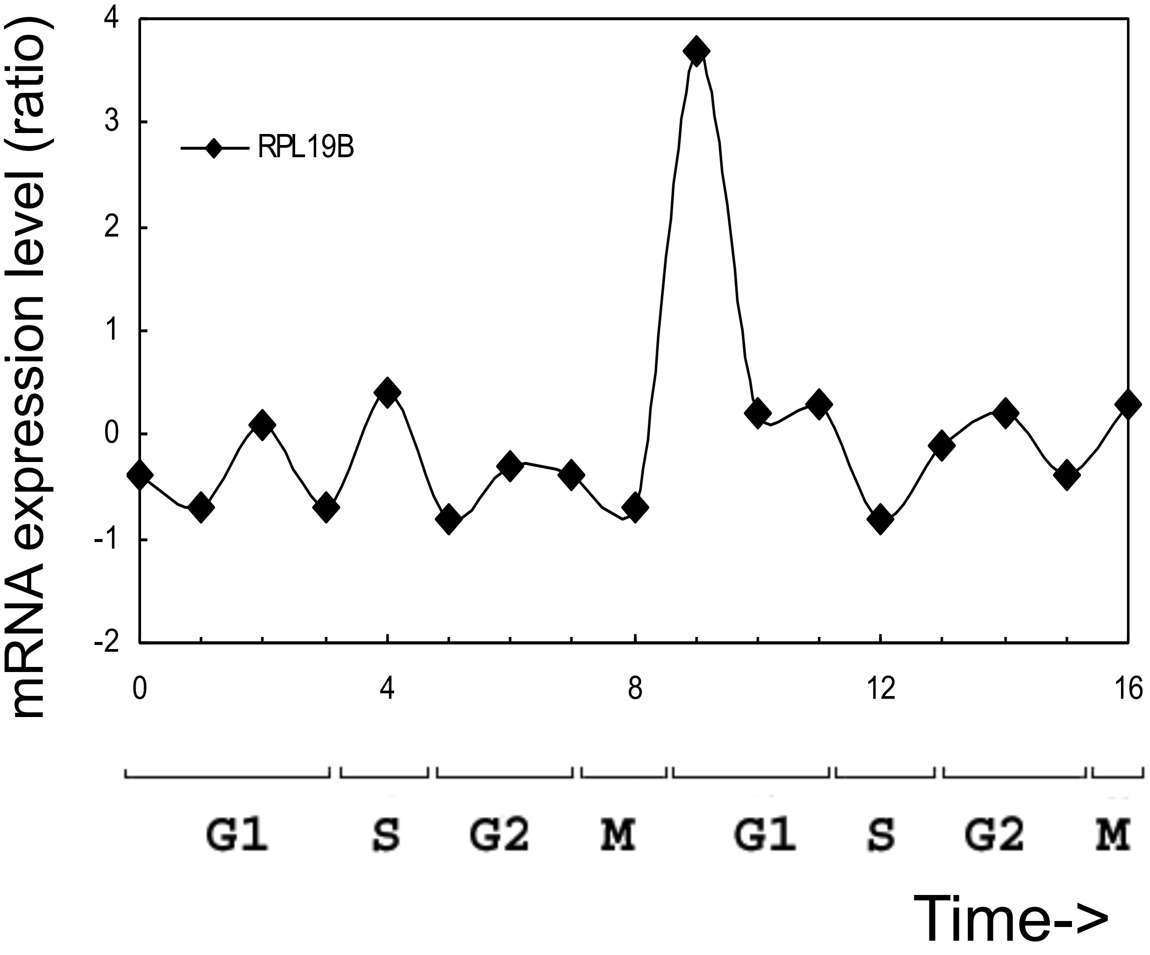
K-means



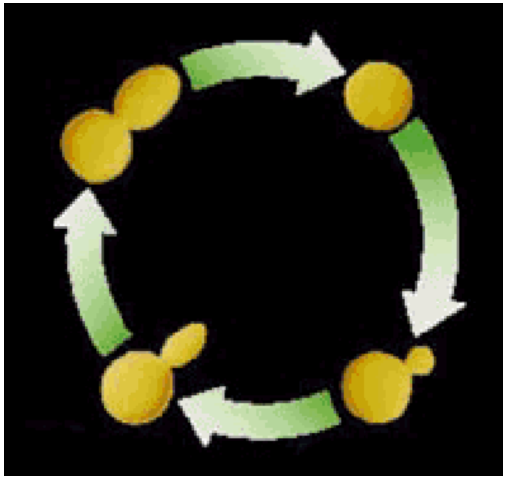
1) Pick ten (i.e. k?) random points as putative cluster centers.
2) Group the points to be clustered by the center to which they are closest.
3) Then take the mean of each group and repeat, with the means now at the cluster center.
4)Stop when the centers stop moving.

Clustering the yeast cell cycle to uncover interacting proteins

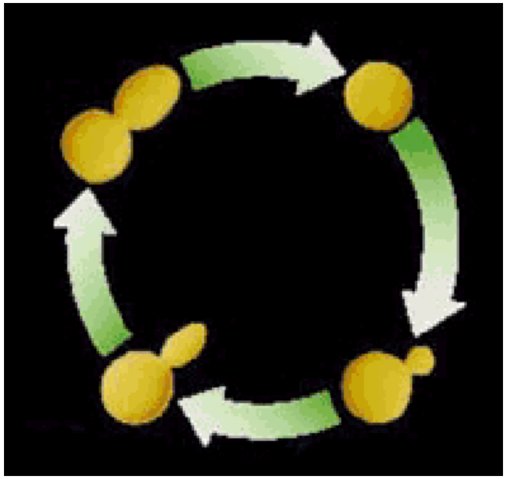
[Brown, Davis]

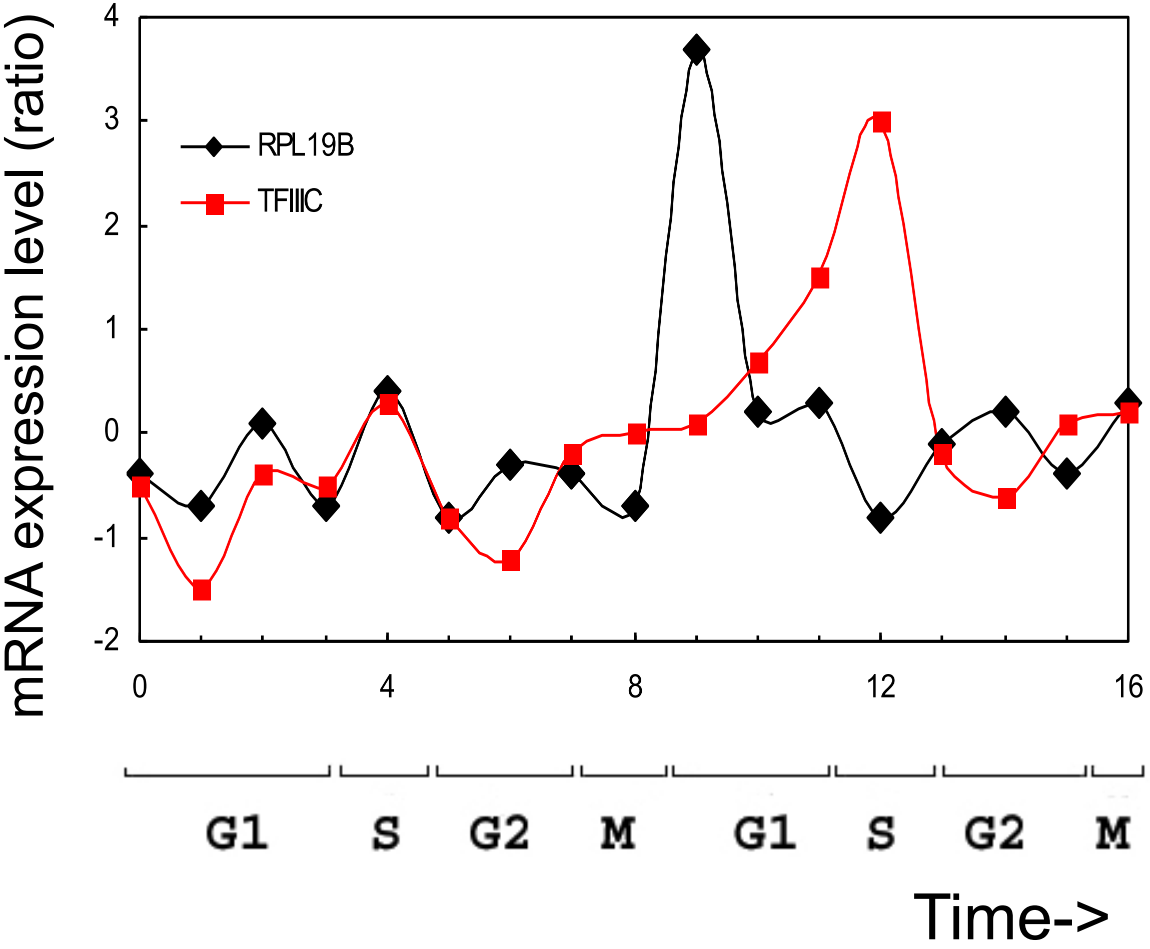


Microarray timecourse of 1 ribosomal protein



Clustering the yeast cell cycle to uncover interacting proteins

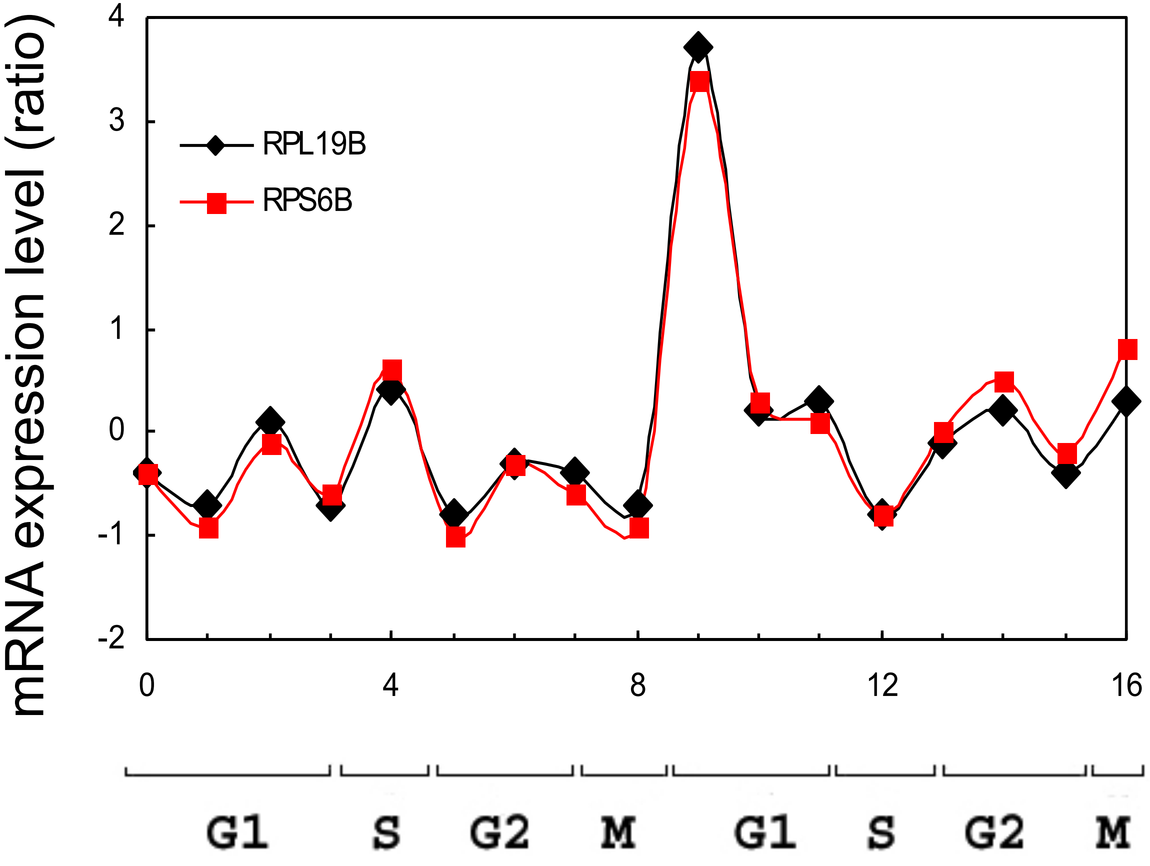




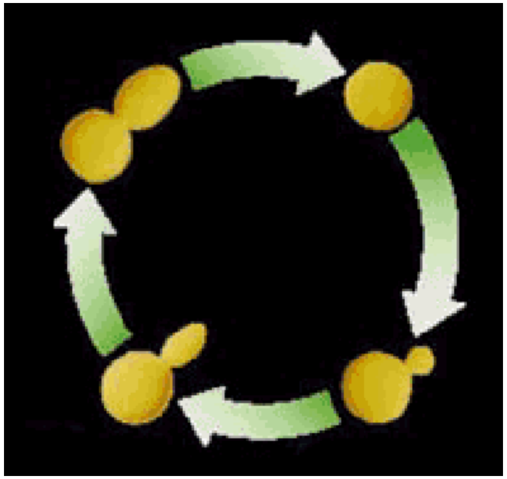
Random relationship from ~18M

Clustering the yeast cell cycle to uncover interacting proteins

[Botstein; Church, Vidal]

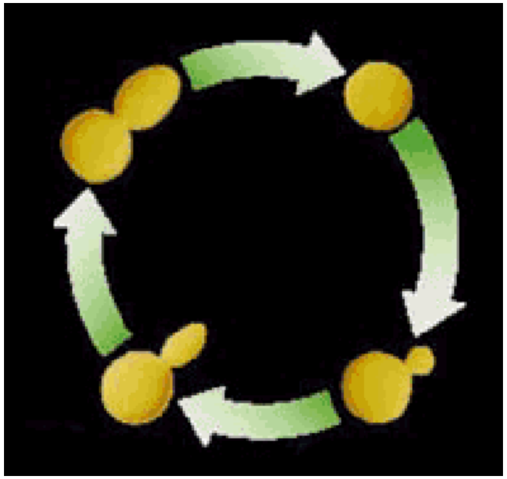


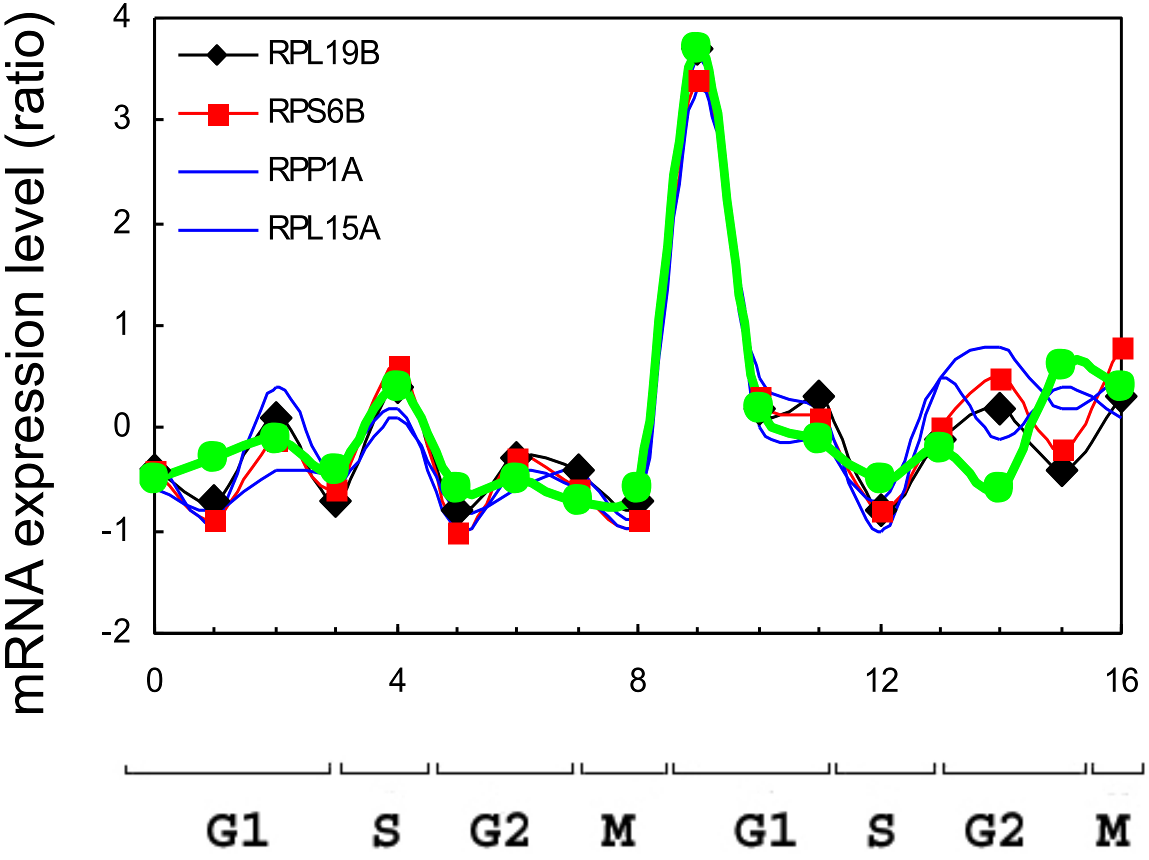
Time->



Close relationship from 18M (2 Interacting Ribosomal Proteins)

Clustering the yeast cell cycle to uncover interacting proteins



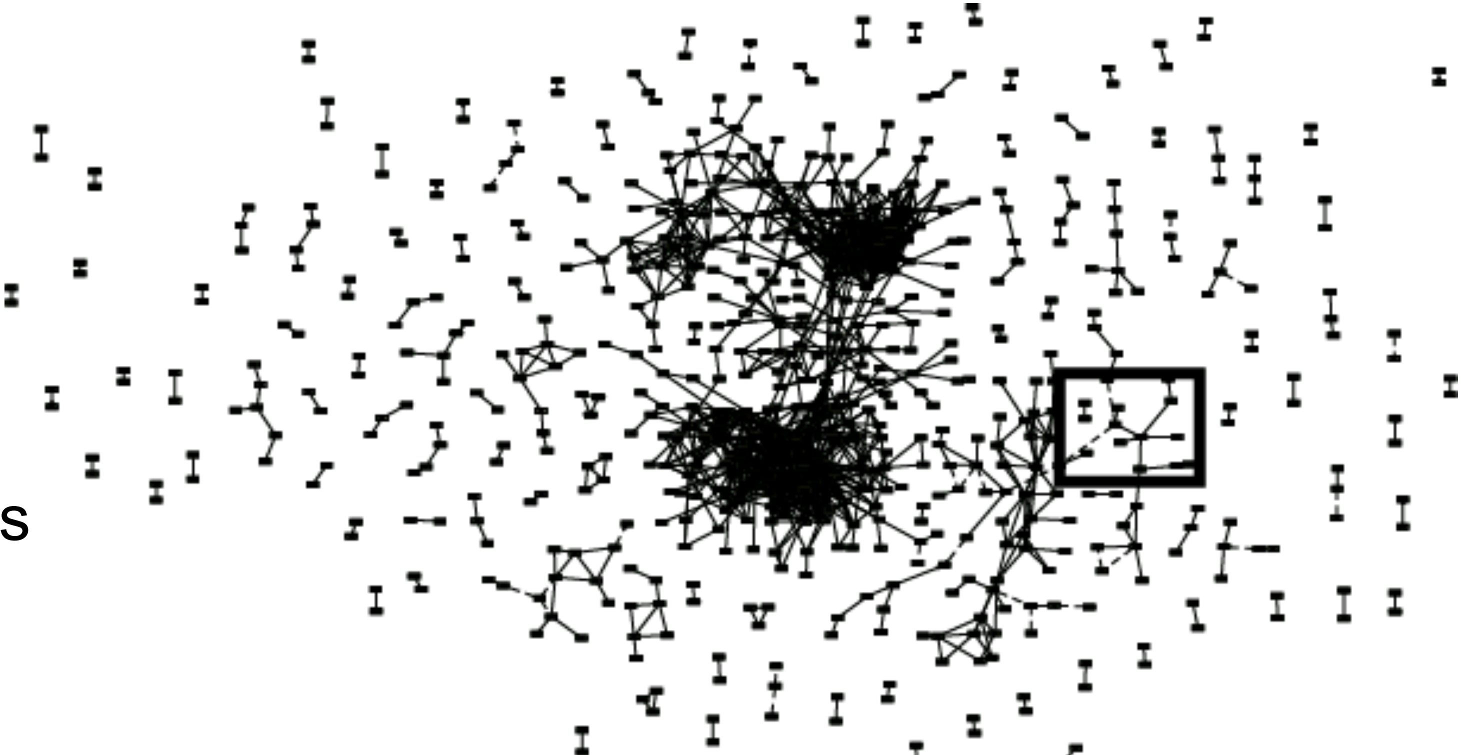


Time->

Predict Functional Interaction of Unknown Member of Cluster

●—●—●

Global Network of Relationships

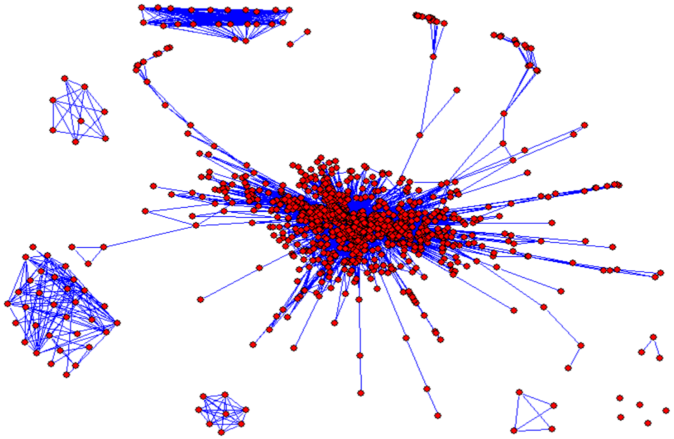


~470K significant relationships from ~**18M** possible

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

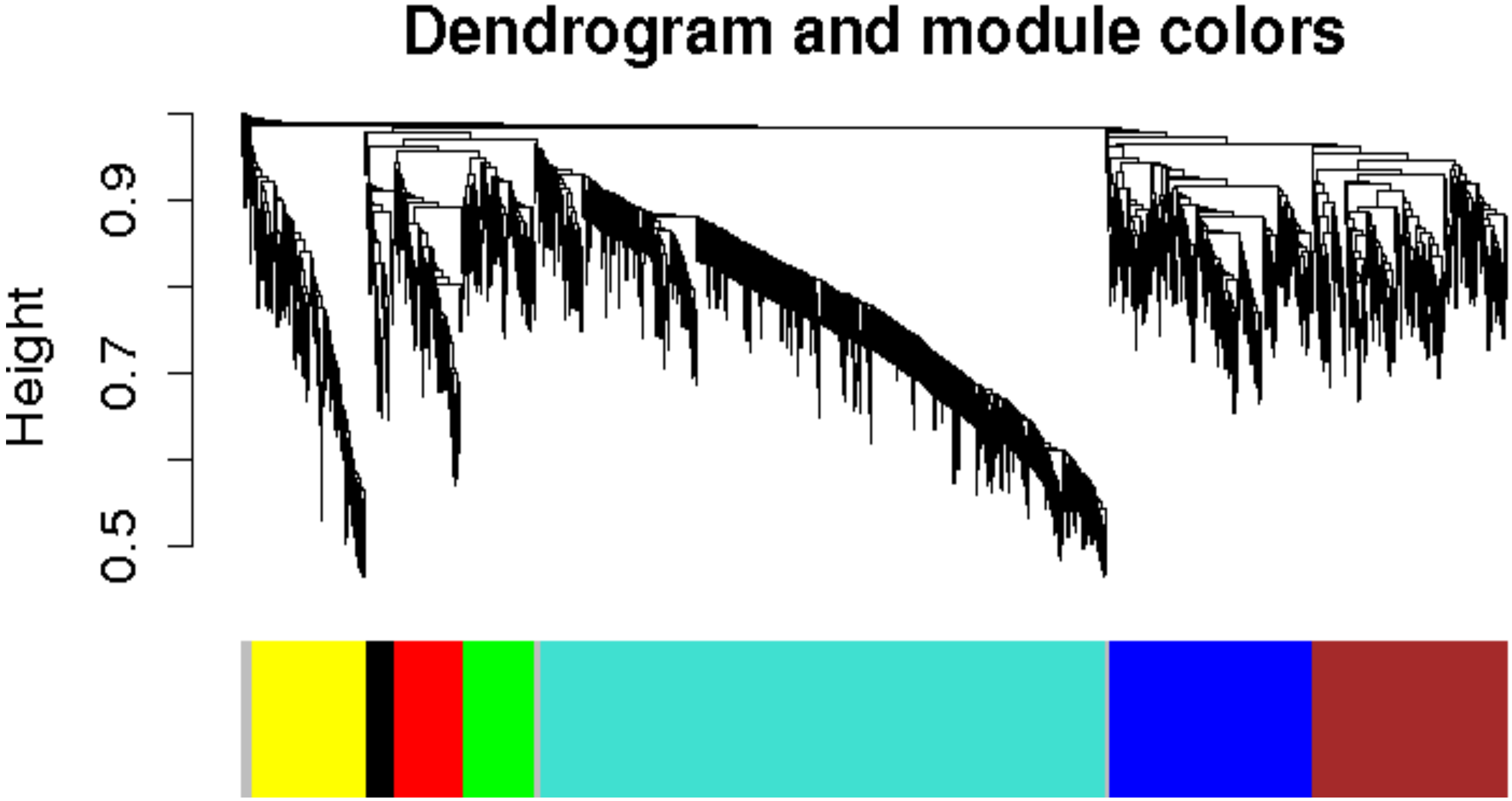
Module Detection

- 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

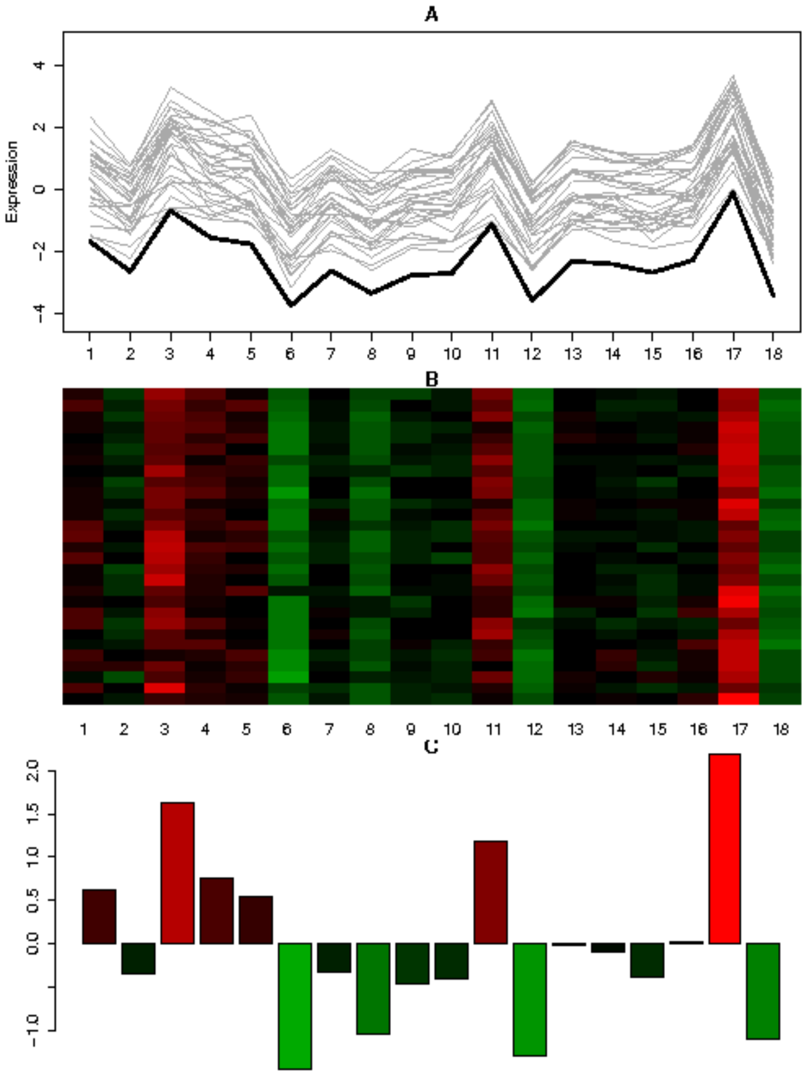
Example of module detection via hierarchical clustering

- 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

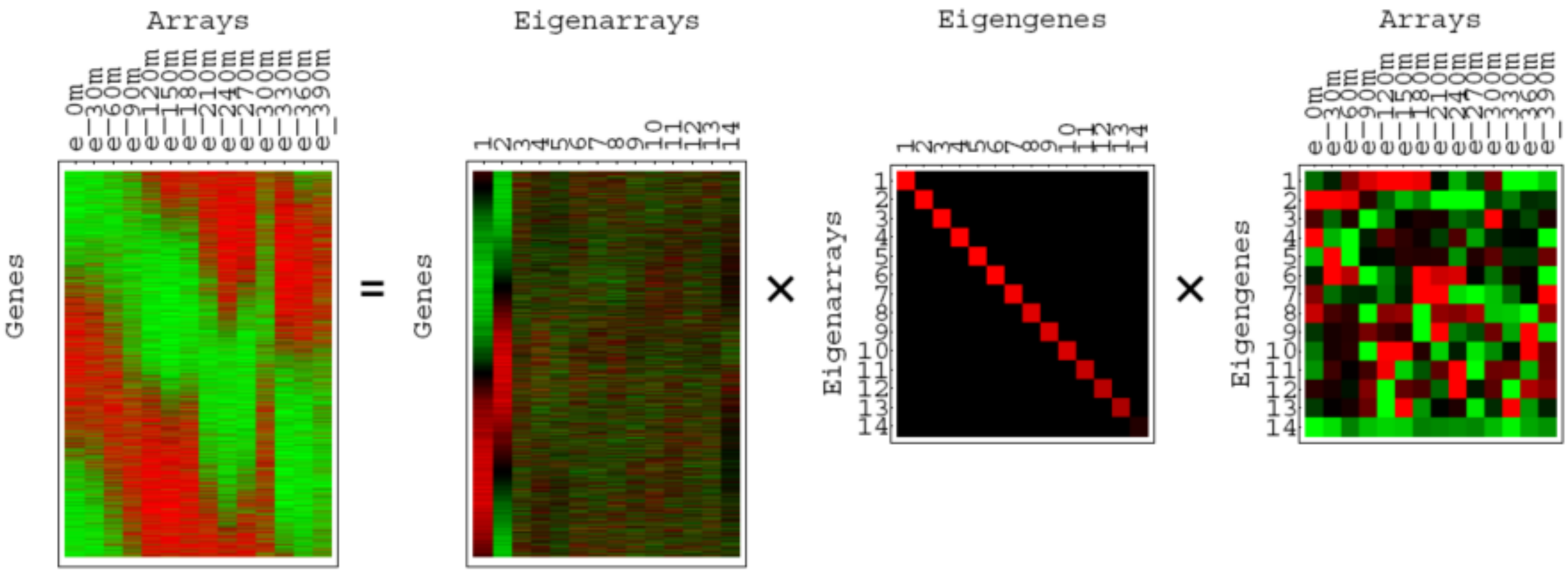
Adapted from : http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork

Unsupervised Mining

SVD

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

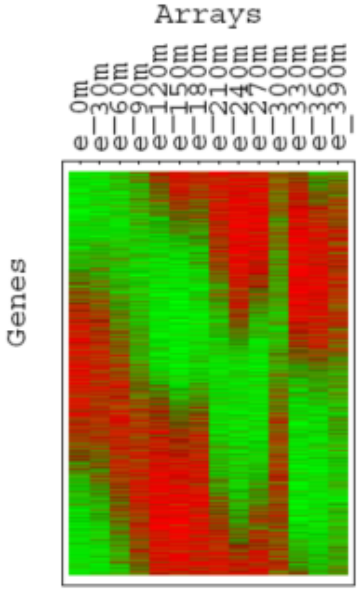
SVD for microarray data (Alter et al, PNAS 2000)



U S V^T

$$A = USV^T$$

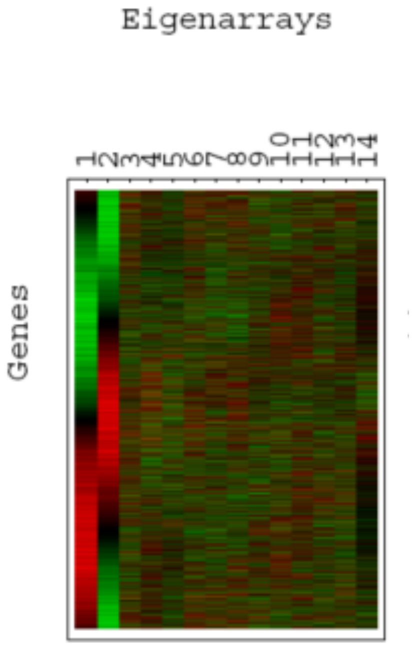
- A is any rectangular matrix (m ≥ n)
- Row space: vector subspace generated by the row vectors of A
- Column space: vector subspace generated by the column vectors of A
  - The dimension of the row & column space is the rank of the matrix A: r (≤ n)
- A is a linear transformation that maps vector x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

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

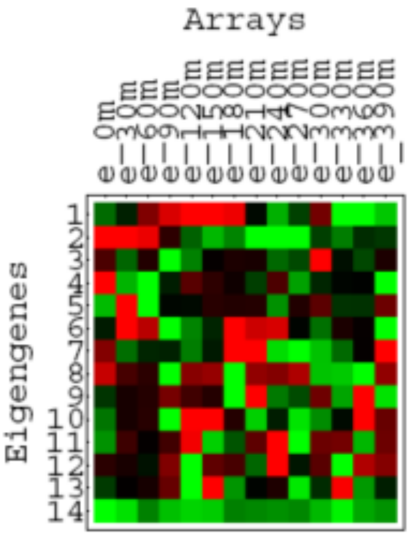


- $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ in U are eigenvectors of AA^T
  - $AA^T = USV^T VSU^T = US^2 U^T$
  - “Left singular vectors”

$$A = USV^T$$

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

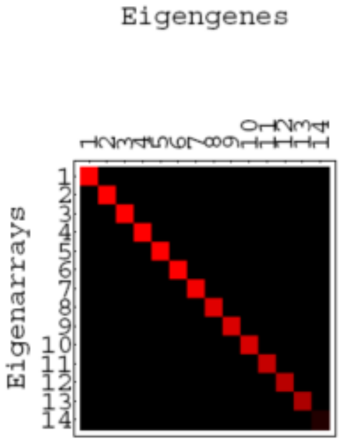
$$V = \begin{pmatrix} | & | & & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \\ | & | & & | \end{pmatrix}$$



- $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ 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 of *A* (*m* by *n*): recap

- $A = USV^T =$ (big-"orthogonal")(diagonal)(sq-orthogonal)

- $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ in U are eigenvectors of AA^T
- $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ in V are eigenvectors of A^TA
- $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ in S are nonnegative singular values of A

- $AV = US$ means each $A\boldsymbol{v}_i = s_i\boldsymbol{u}_i$
- “Every A is diagonalized by 2 orthogonal matrices”

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?

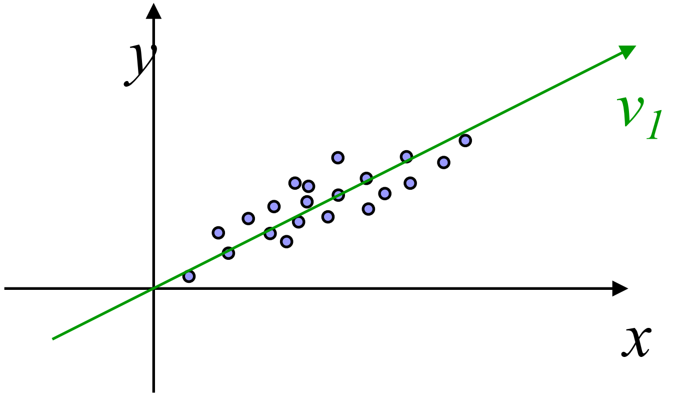
$$\begin{pmatrix} 101 & 303 & 202 \\ 102 & 300 & 201 \\ 103 & 304 & 203 \\ 101 & 302 & 204 \end{pmatrix}$$

- Signals?

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

Geometry of SVD in row space

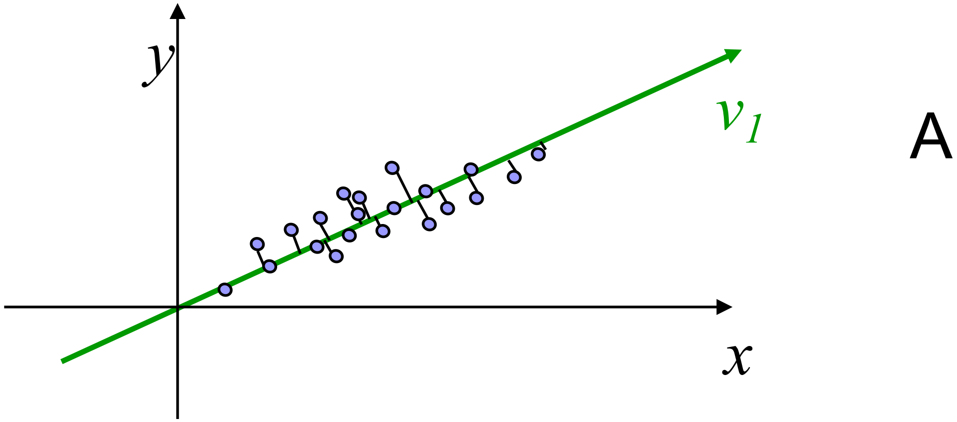
- A as a collection of m row vectors (points) in the row space of A
- $s_1 \boldsymbol{u}_1 \boldsymbol{v}_1^T$ is the best rank-1 matrix approximation for A
- Geometrically: $\boldsymbol{v}_1$ is the direction of the best approximating rank-1 subspace that goes through origin
- $s_1 \boldsymbol{u}_1$ gives coordinates for row vectors in rank-1 subspace
- $\boldsymbol{v}_1$ Gives coordinates for row space basis vectors in rank-1 subspace

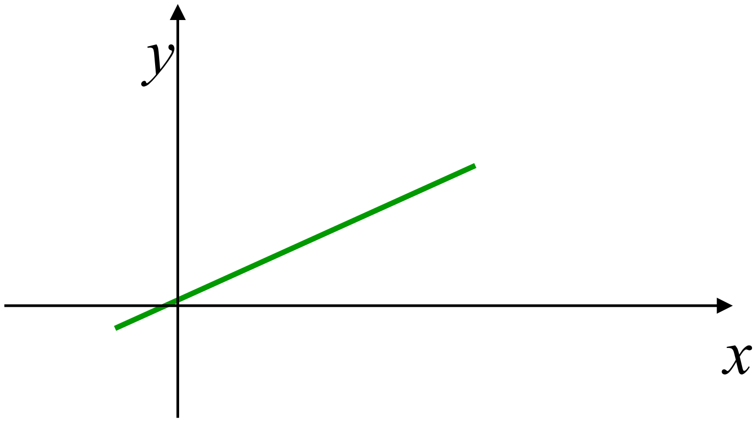


$$\boxed{A\,\mathbf{v}_\mathbf{i} = s_i \mathbf{u}_\mathbf{i}}$$

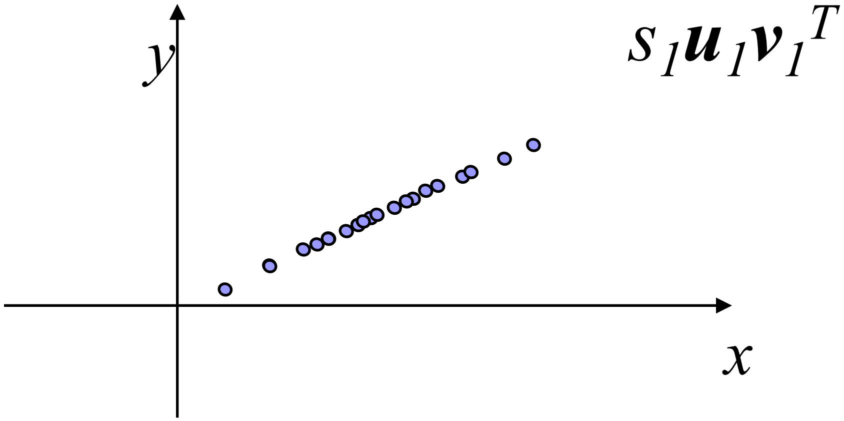
$$\boxed{I\,\mathbf{v}_\mathbf{i} = \mathbf{v}_\mathbf{i}}$$

Geometry of SVD in row space





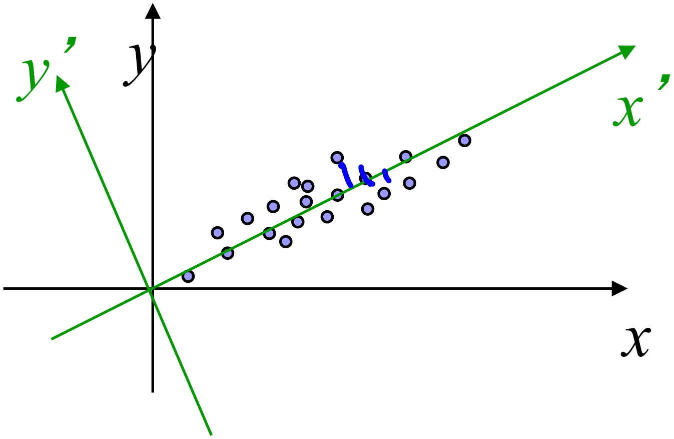
This line segment that goes through origin approximates the original data set



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



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

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

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

Geometry of SVD in row and column spaces

- Row space
  - $s_i\boldsymbol{u}_i$ gives coordinates for row vectors along unit vector $\boldsymbol{v}_i$
  - $\boldsymbol{v}_i$ gives coordinates for row space basis vectors along unit vector $\boldsymbol{v}_i$
- Column space
  - $s_i\boldsymbol{v}_i$ gives coordinates for column vectors along unit vector $\boldsymbol{u}_i$
  - $\boldsymbol{u}_i$ gives coordinates for column space basis vectors along unit vector $\boldsymbol{u}_i$
- Along the directions $\boldsymbol{v}_i$ and $\boldsymbol{u}_i$, these two spaces look pretty much the same!
  - Up to scale factors s_i
  - Switch row/column vectors and row/column space basis vectors
  - **Biplot....**

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

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

$$\boxed{A^T\,\mathbf{u_i} = s_i\mathbf{v_i}}$$

$$\boxed{I\,\mathbf{u_i} = \mathbf{u_i}}$$

Additional Points

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

$$\mathbf{A} = \begin{array}{cc} & \text{DOCS} \\ \text{TERMS} & \left(\begin{matrix} \square & & \\ & & \\ & & \end{matrix}\right) \end{array}$$

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

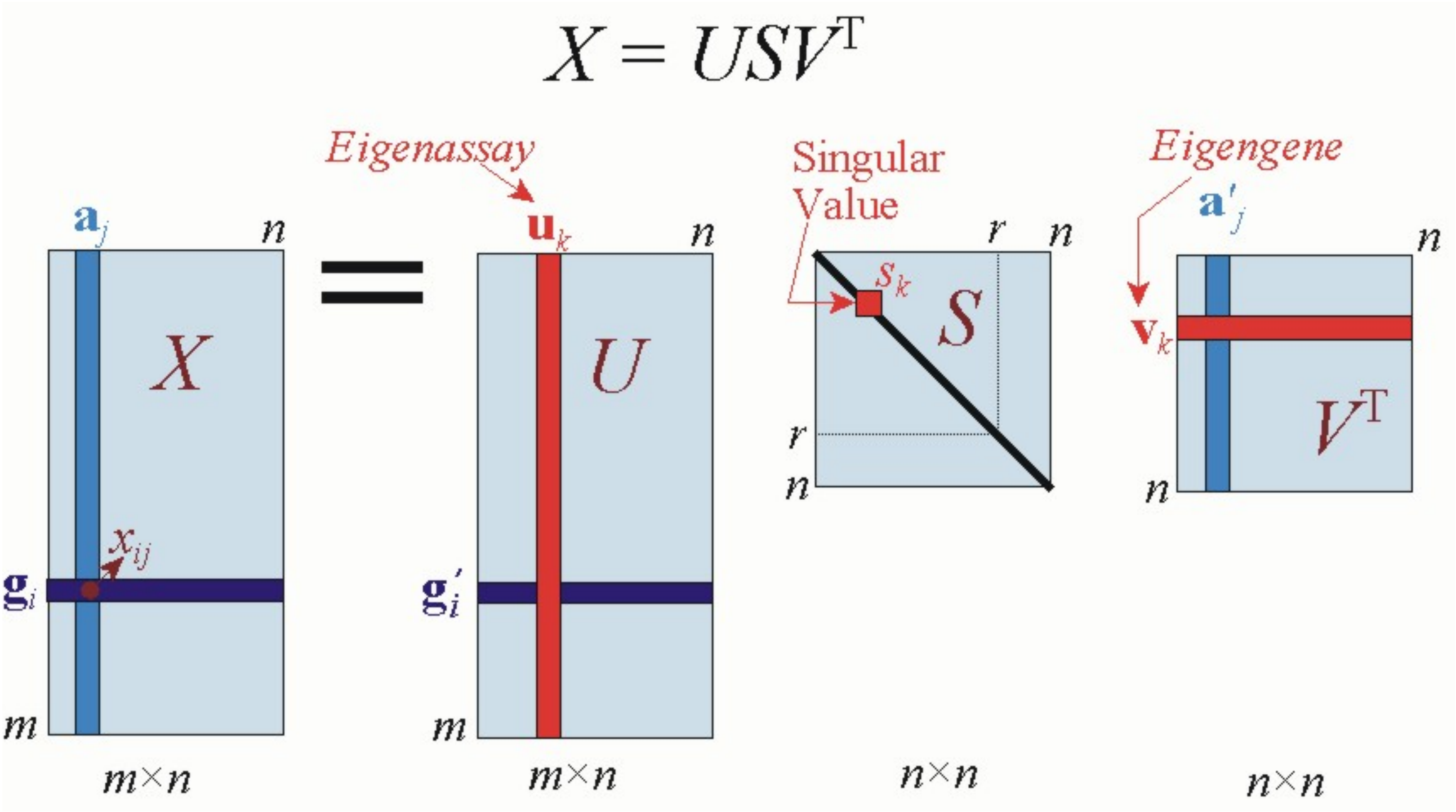
Conclusion

- SVD is the “absolute high point of linear algebra”
- SVD is difficult to compute; but once we have it, we have many things
- SVD finds the best approximating subspace, using linear transformation
- Simple SVD cannot handle translation, non-linear transformation, separation of labeled data, etc.
- Good for exploratory analysis; but once we know what we look for, use appropriate tools and model the structure of data explicitly!

Unsupervised Mining

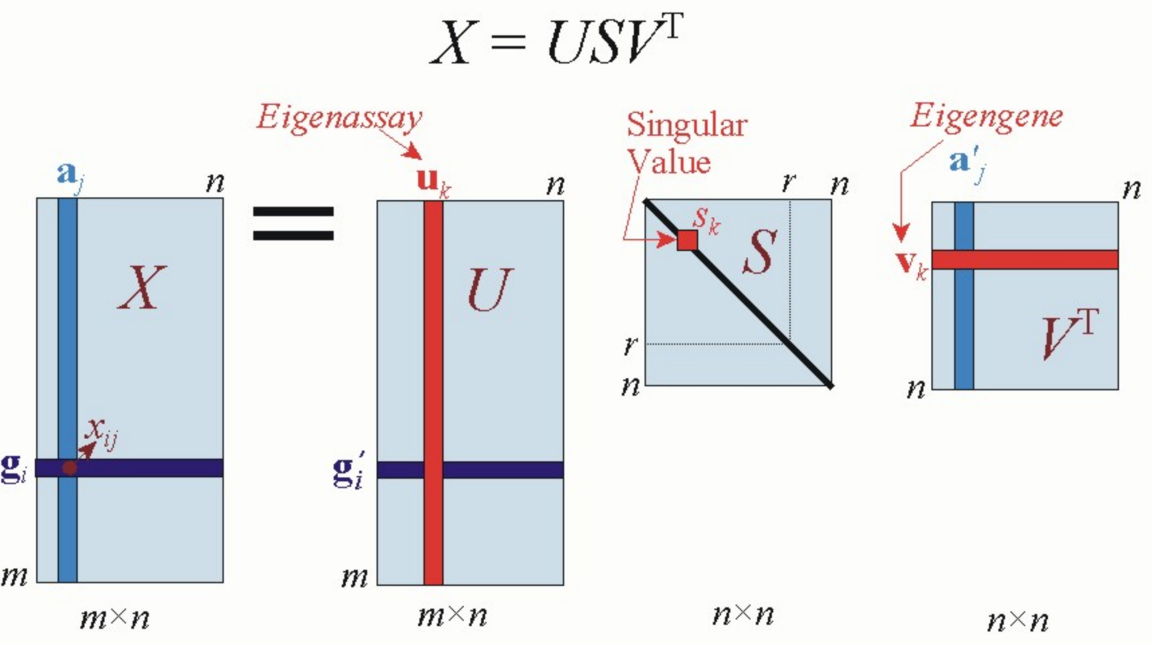
Intuition on interpretation of SVD
in terms of genes and conditions

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



Notation

- m=1000 genes
  - row-vectors
  - 10 eigengene (v_i) of dimension 10 conditions
- n=10 conditions (assays)
  - column vectors
  - 10 eigenconditions (u_i) of dimension 1000 genes



Close up on Eigengenes

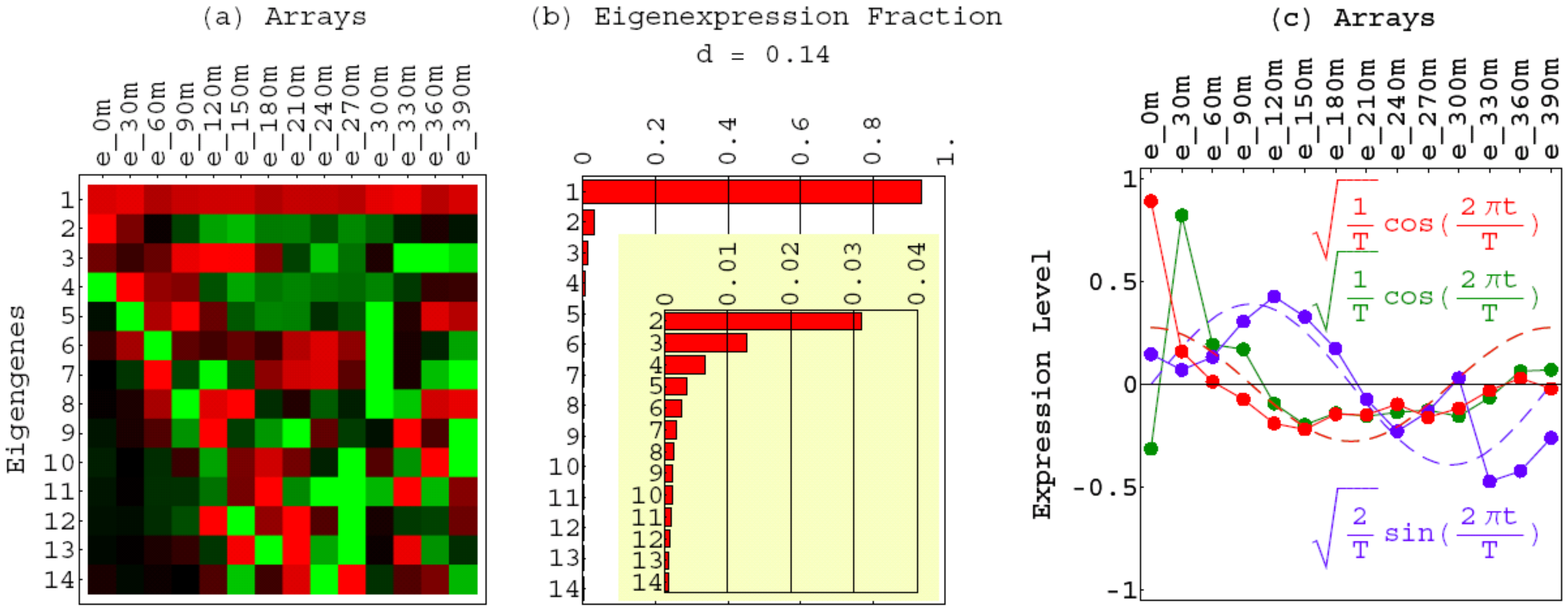
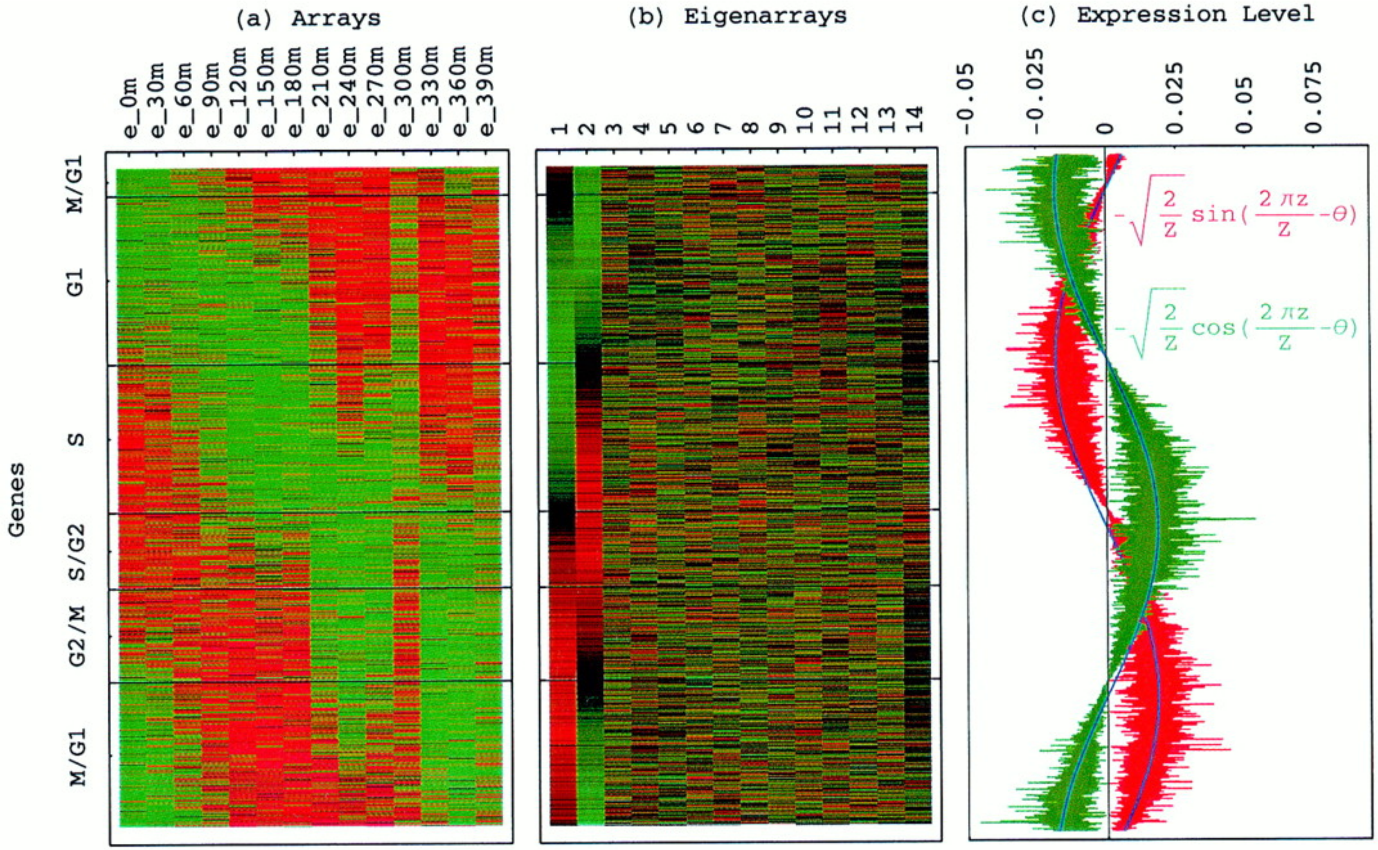


Fig. 8. Elutriation eigengenes. (*a*) Raster display of $\hat{v}^T$, the expression of 14 eigengenes in 14 arrays, with overexpression (red), no change in expression (black), and underexpression (green) around the steady state, which can be associated with the first eigengene, $|\gamma_1\rangle$. (*b*) Bar chart of the fraction of eigenexpression p_l of each eigengene $|\gamma_l\rangle$, showing more than 90% of the overall relative expression in $|\gamma_1\rangle$, about 3%, 1.5%, and 0.5% in $|\gamma_2\rangle$, $|\gamma_3\rangle$, and $|\gamma_4\rangle$, respectively, and a low entropy $d = 0.14 \ll 1$. (*c*) Line-joined graphs of the expression levels of $|\gamma_2\rangle$ (red), $|\gamma_3\rangle$ (blue), and $|\gamma_4\rangle$ (green) in the 14 arrays, and dashed graphs of normalized cosine (blue) and sine (red) of period T.

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

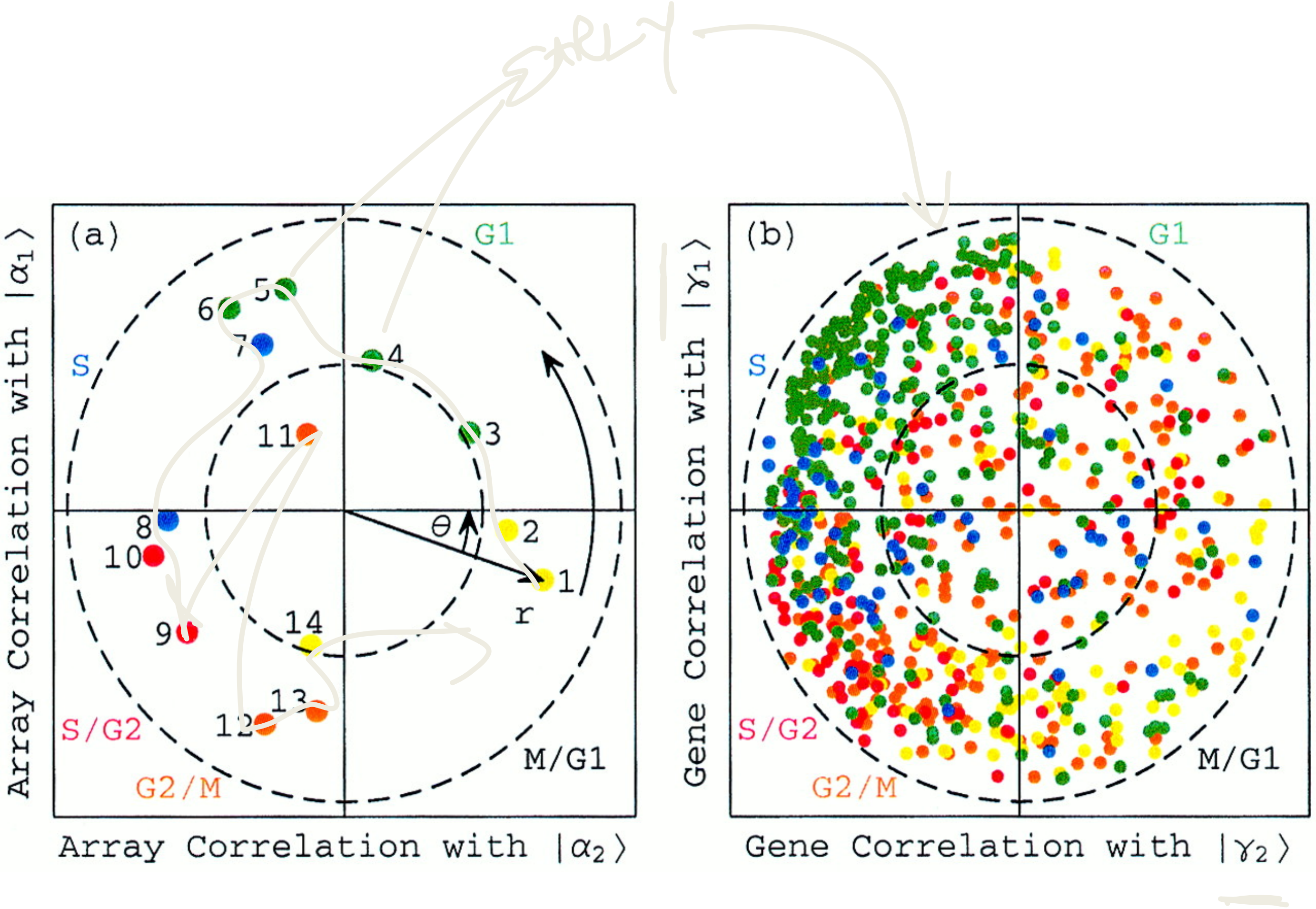
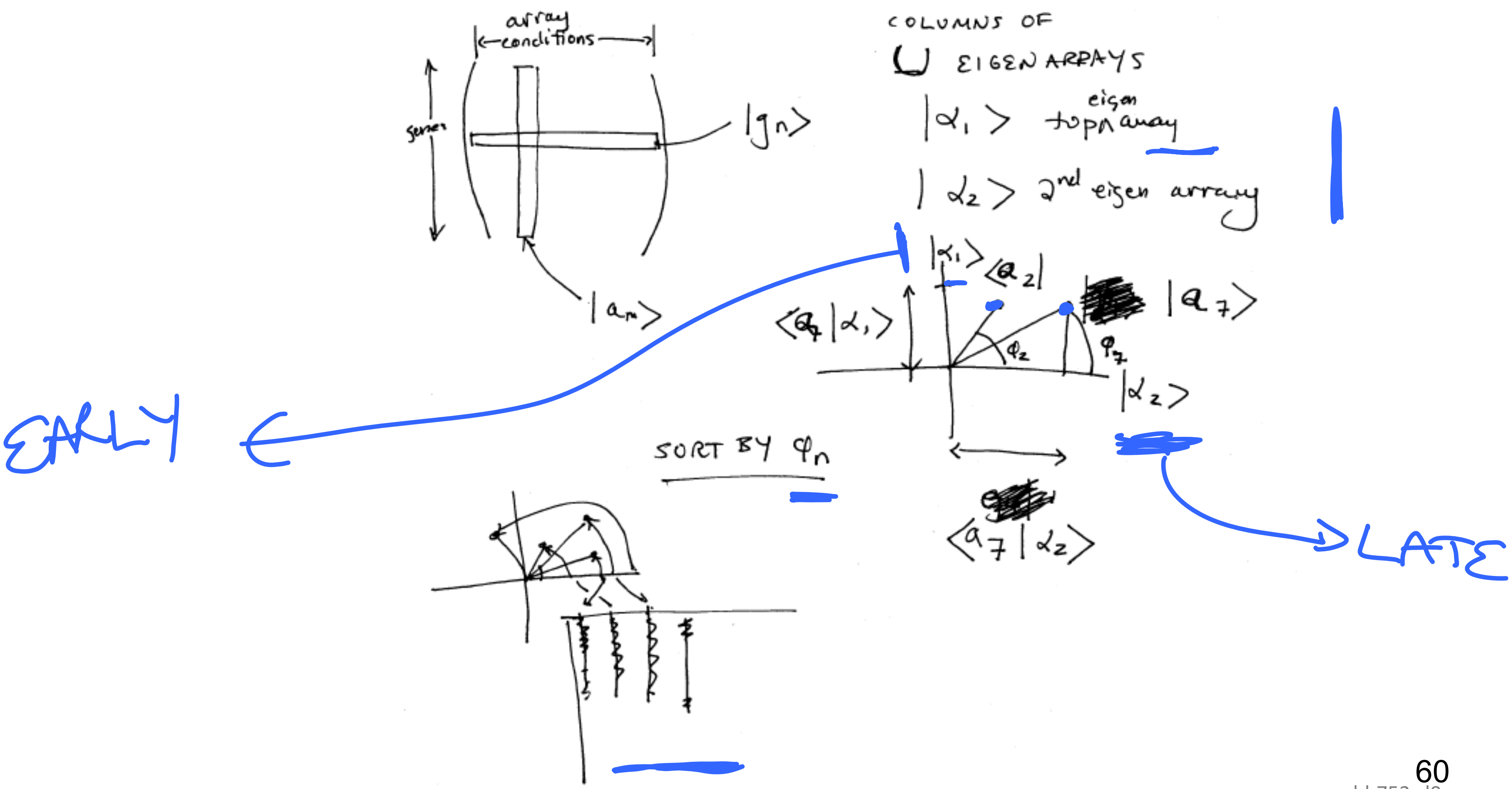


Fig. 2. Normalized elutriation expression in the subspace associated with the cell cycle. (a) Array correlation with $|\alpha_1\rangle$ along the y-axis vs. that with $|\alpha_2\rangle$ along the x-axis, color-coded according to the classification of the arrays into the five cell cycle stages, M/G_1 (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 $|\alpha_1\rangle$ and $|\alpha_2\rangle$ subspace. (b) Correlation of each gene with $|\gamma_1\rangle$ vs. that with $|\gamma_2\rangle$, 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

PNAS

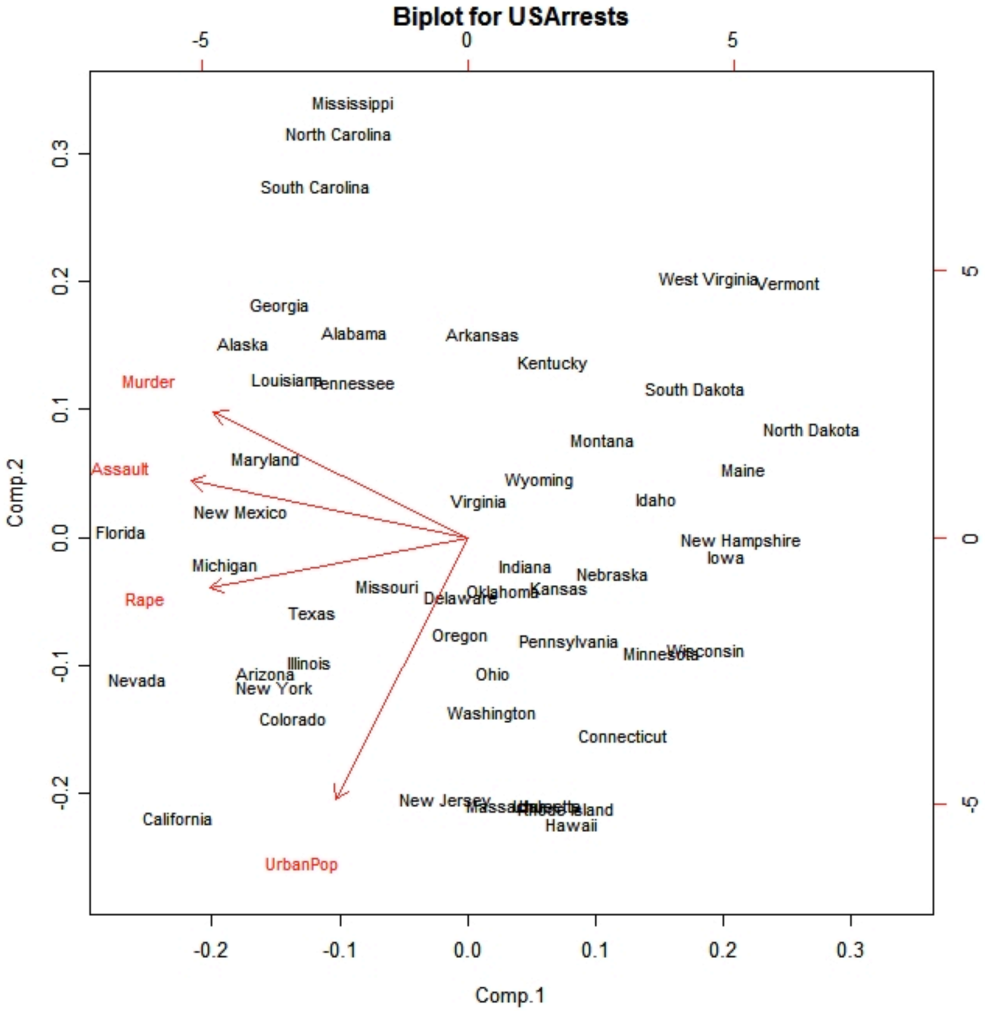
Plotting Experiments in Low Dimension Subspace



Unsupervised Mining

Biplot

Introduction



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

PCA

TFs: a, b, c...

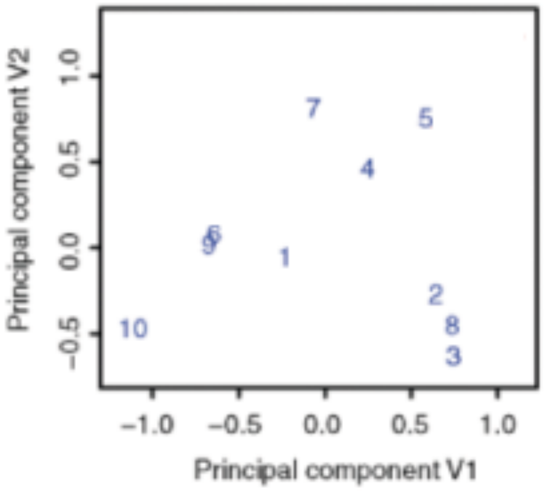
Genomic Sites: 1,2,3...

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.)

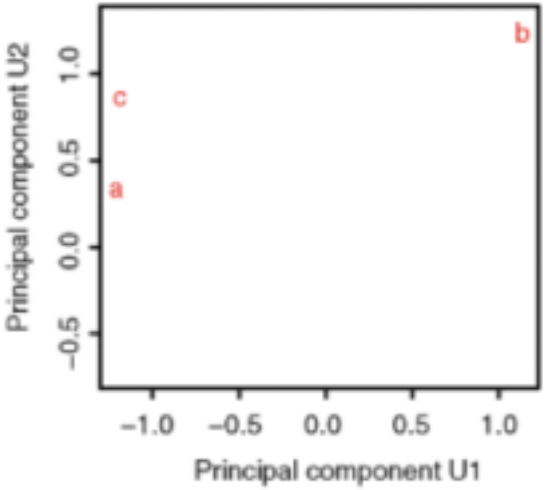


A^T

	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...

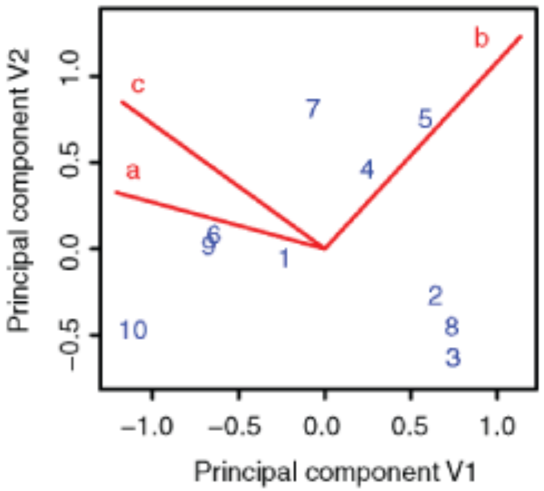
Genomic Sites: 1,2,3...

$$A=USV^T$$

	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.)

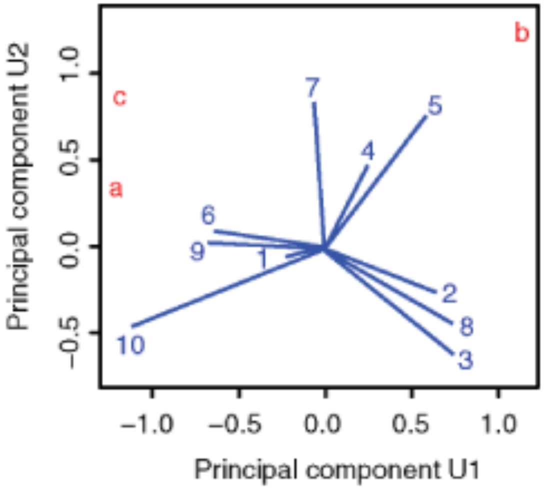


$$A^T$$

	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)



Biplot Ex

A' A

A

	TF		
Genomic bin	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

Data matrix

Variable (column) standardization

	TF		
Genomic bin	a	b	c
1	0.84	-0.23	-0.20
2	-1.06	0.29	-0.82
3	-1.06	0.03	-1.43
4	-1.06	0.55	0.82
5	-0.24	1.59	-0.20
6	0.57	-0.75	1.02
7	1.11	1.07	0.20
8	-0.78	0.29	-1.43
9	0.03	-1.01	1.43
10	1.65	-1.80	0.61

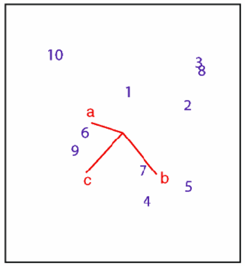
Standardized data matrix

Correlating factors

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

Correlation matrix R_1

3D scatterplot



PCA

Transpose

	Genomic bin									
Transcription factor	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

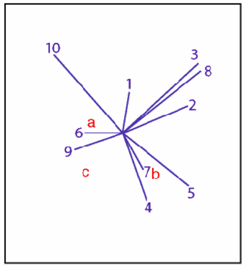
Data matrix (transposed)

Variable (column) standardization

	Genomic bin									
TF	1	2	3	4	5	6	7	8	9	10
a	-0.11	-0.90	-0.91	-0.81	-1.03	-0.26	-0.47	-0.95	-0.36	0.18
b	-0.94	-0.18	-0.16	-0.30	0.06	-0.84	-0.68	-0.09	-0.77	-1.08
c	1.05	1.08	1.07	1.12	0.97	1.10	1.15	1.04	1.13	0.90

Standardized data matrix (transposed)

10D scatterplot *



PCA

Correlating bins

	Genomic bin									
Genomic bin	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

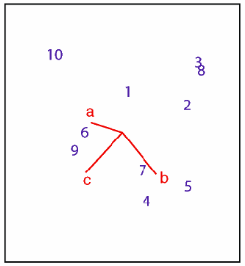
Correlation matrix R_2

Biplot Ex #2

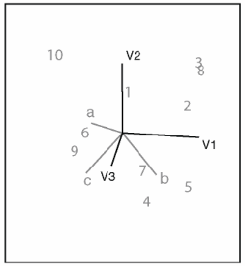
TF

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

Correlation matrix R_1

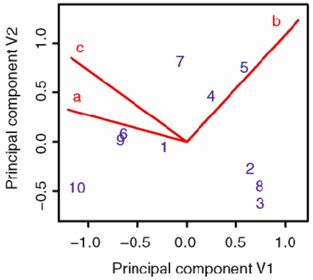


PCA *



$$A^T A = V S^2 V^T$$

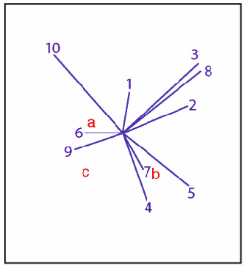
Projection *



The same rank-2 approximation of the original data matrix

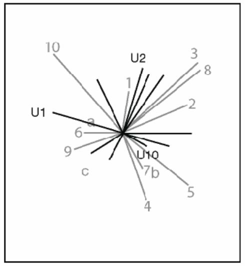
$$\mathbf{A}\,\mathbf{v}_j = \boldsymbol{u}_j\, s_j \;\&\; \mathbf{A}^T\,\mathbf{u}_j = \mathbf{v}_j\, s_j$$

$$\boxed{\mathbf{A} = (\mathbf{U}\,\mathbf{S}^r)\,(\mathbf{V}\,\mathbf{S}^{1-r})^{T}}$$

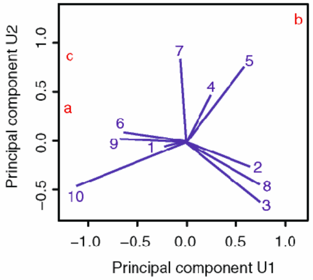


$$A A^T = U S^2 U^T$$

PCA *



Projection *



Genomic bin

2	3	4	5	6	7	8	9	10
.70	0.69	0.77	0.54	0.99	0.95	0.65	0.98	0.97
.00	1.00	0.99	0.98	0.79	0.89	1.00	0.84	0.50
.00	1.00	0.99	0.98	0.78	0.89	1.00	0.83	0.49
.99	0.99	1.00	0.95	0.85	0.94	0.98	0.89	0.59
.98	0.98	0.95	1.00	0.64	0.78	0.99	0.71	0.31
.79	0.78	0.85	0.64	1.00	0.98	0.74	1.00	0.93
.89	0.89	0.94	0.78	0.98	1.00	0.86	0.99	0.84
.00	1.00	0.98	0.99	0.74	0.86	1.00	0.80	0.43
.84	0.83	0.89	0.71	1.00	0.99	0.80	1.00	0.89
.50	0.49	0.59	0.31	0.93	0.84	0.43	0.89	1.00

Correlation matrix R_2

*

10D scatterplots are used here for illustrative purpose only.

PCA: the correlation matrix is eigen-decomposed; then the principal components are added to the original space.

Projection: the points and axes in the original space are projected onto the plane defined by the top two principal components.

Biplot Ex #3

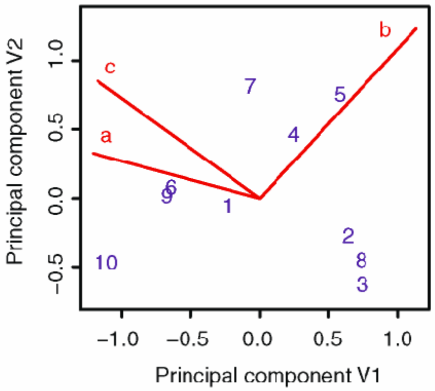
$$\boxed{A\,\mathbf{v}_i = s_i\mathbf{u}_i}$$

$$\boxed{A^T\,\mathbf{u}_i = s_i\mathbf{v}_i}$$

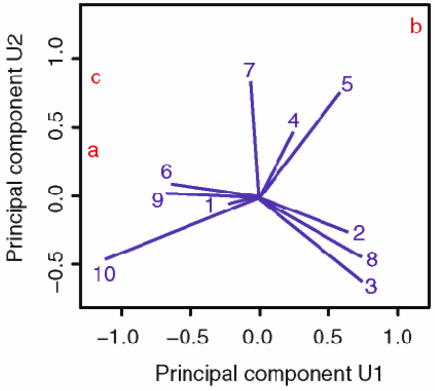
Assuming s=1,

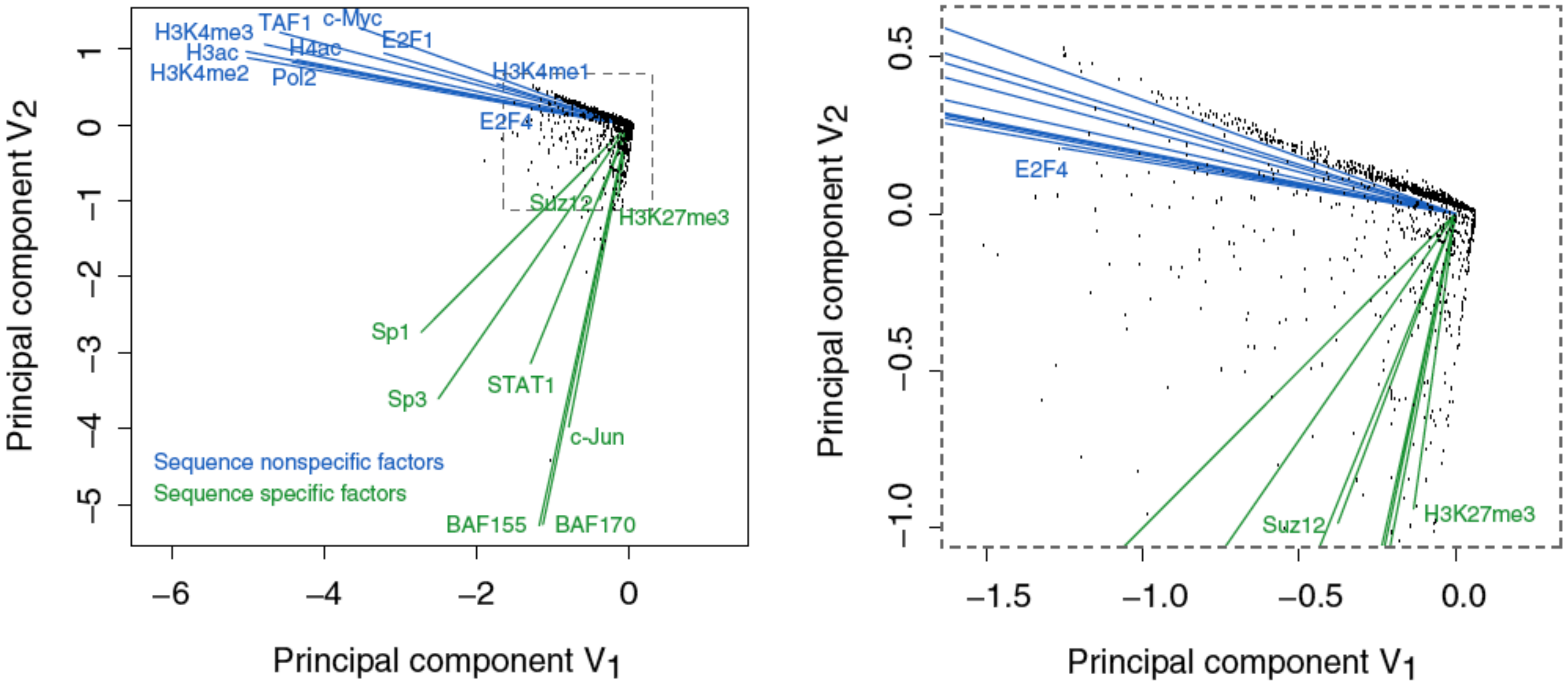
$A\mathbf{v}_i = \mathbf{u}_i$

$A^T\mathbf{u}_i = \mathbf{v}_i$



The same rank-2 approximation of the original data matrix

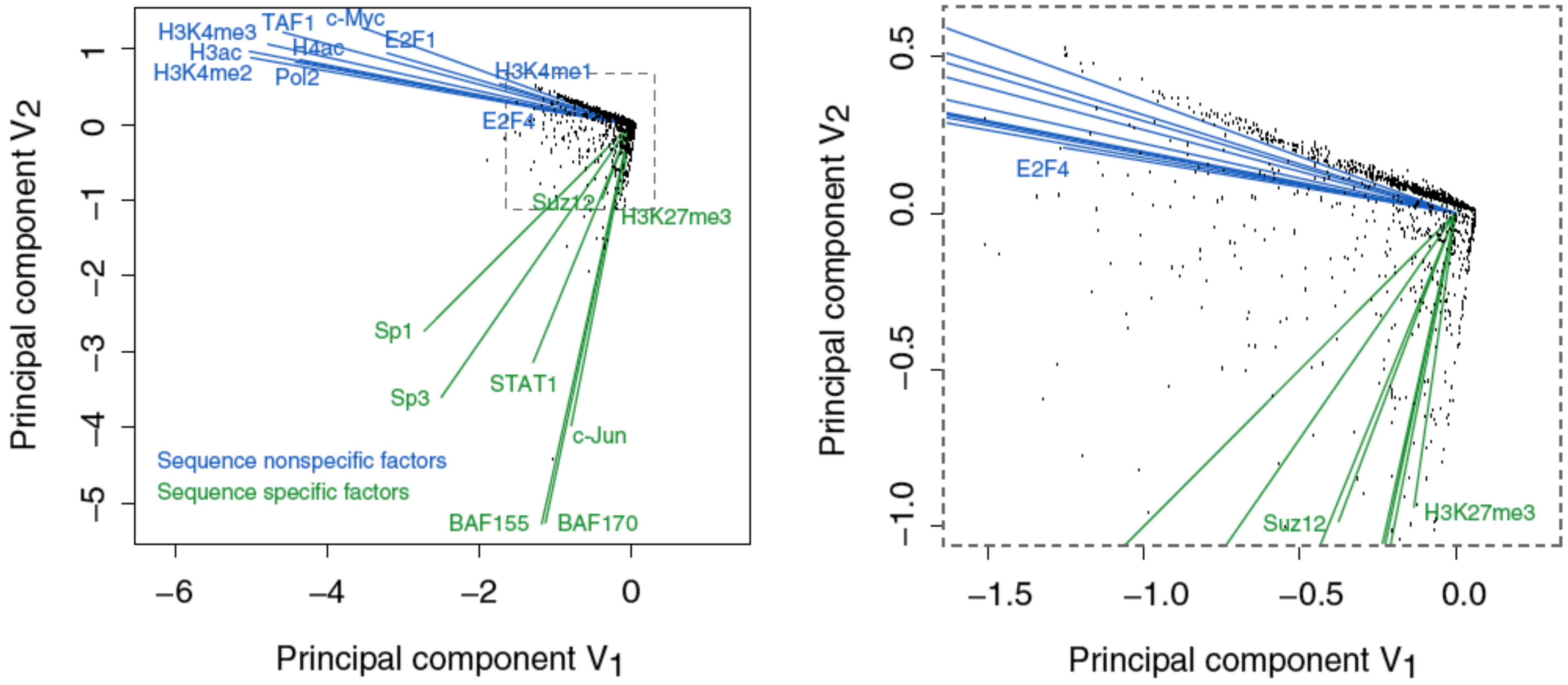




Results of Biplot

- 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

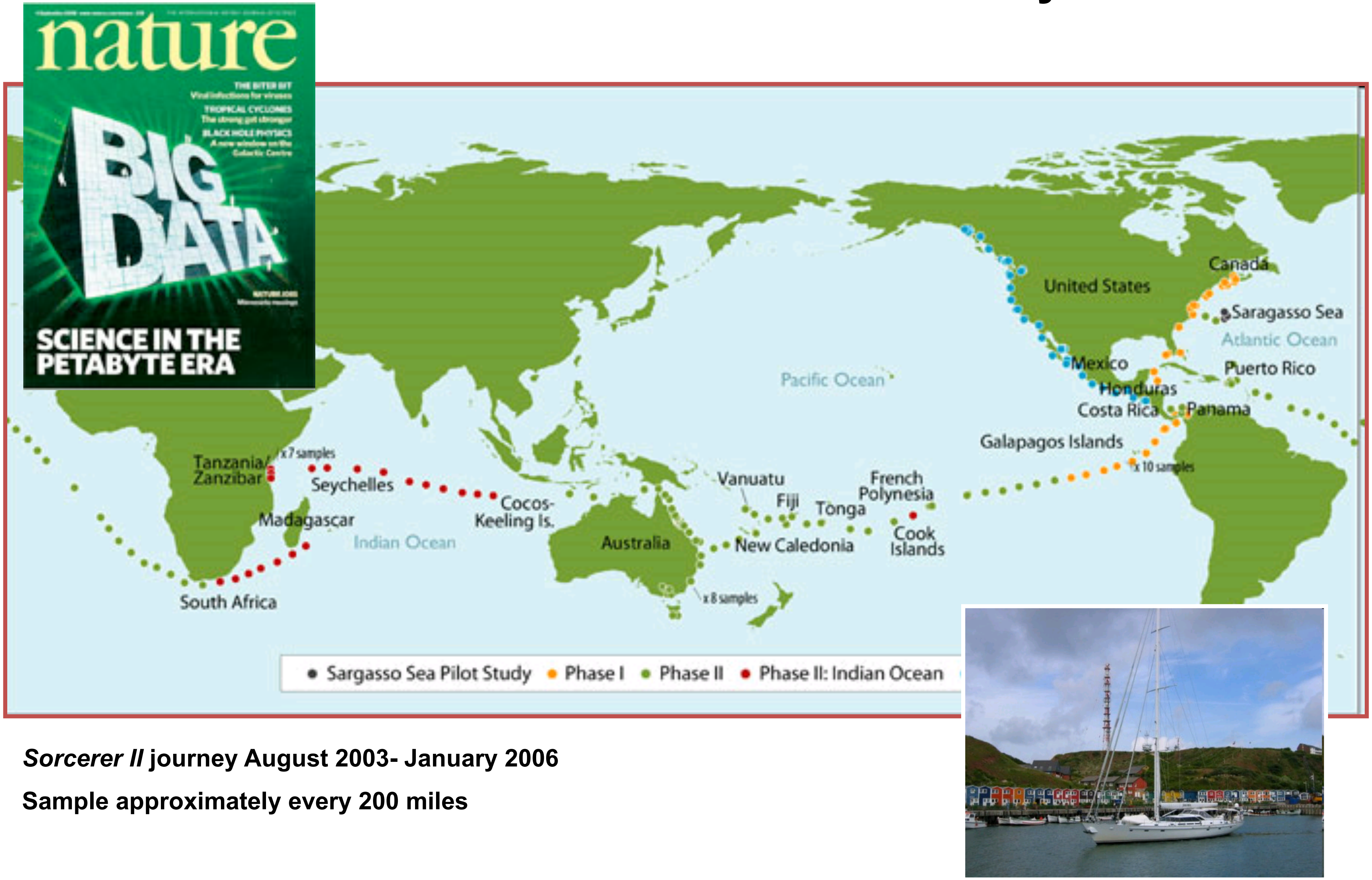
- Biplot groups TFs into sequence-specific and sequence-nonspecific clusters.
  - c-Myc may behave more like a sequence-nonspecific TF.
  - H3K27me3 functions in a transcriptional regulatory process in a rather sequence-specific manner.
- Genomic Bins are associated with different TFs and in this fashion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

Zhang et al. (2007)
Gen. Res.

Unsupervised Mining

CCA

Sorcerer II Global Ocean Survey

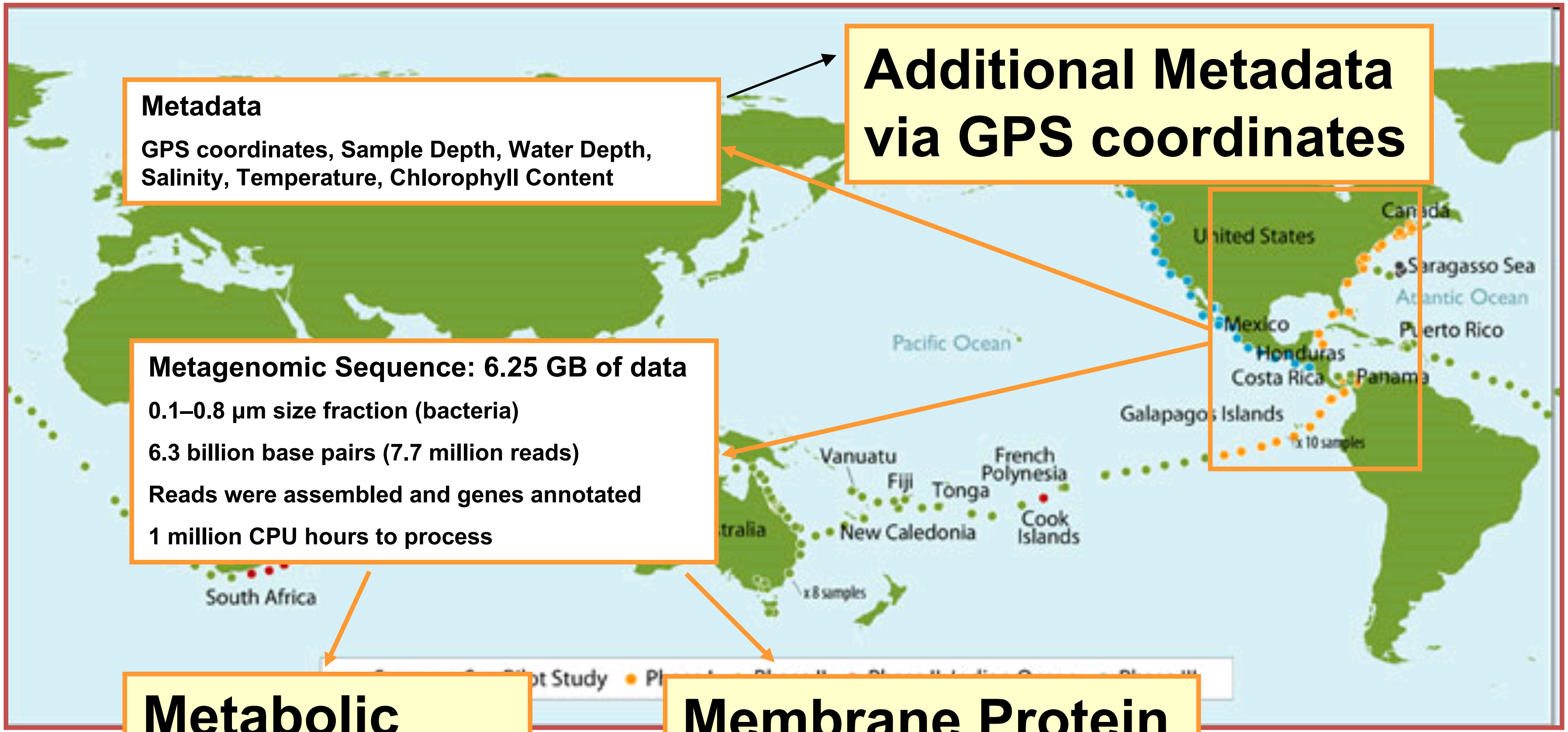


Sorcerer II journey August 2003- January 2006

Sample approximately every 200 miles

Rusch, et al., _PLOS Biology_ 2007

Sorcerer II Global Ocean Survey



Metadata

GPS coordinates, Sample Depth, Water Depth, Salinity, Temperature, Chlorophyll Content

Additional Metadata via GPS coordinates

Metagenomic Sequence: 6.25 GB of data

0.1–0.8 µm size fraction (bacteria)

6.3 billion base pairs (7.7 million reads)

Reads were assembled and genes annotated

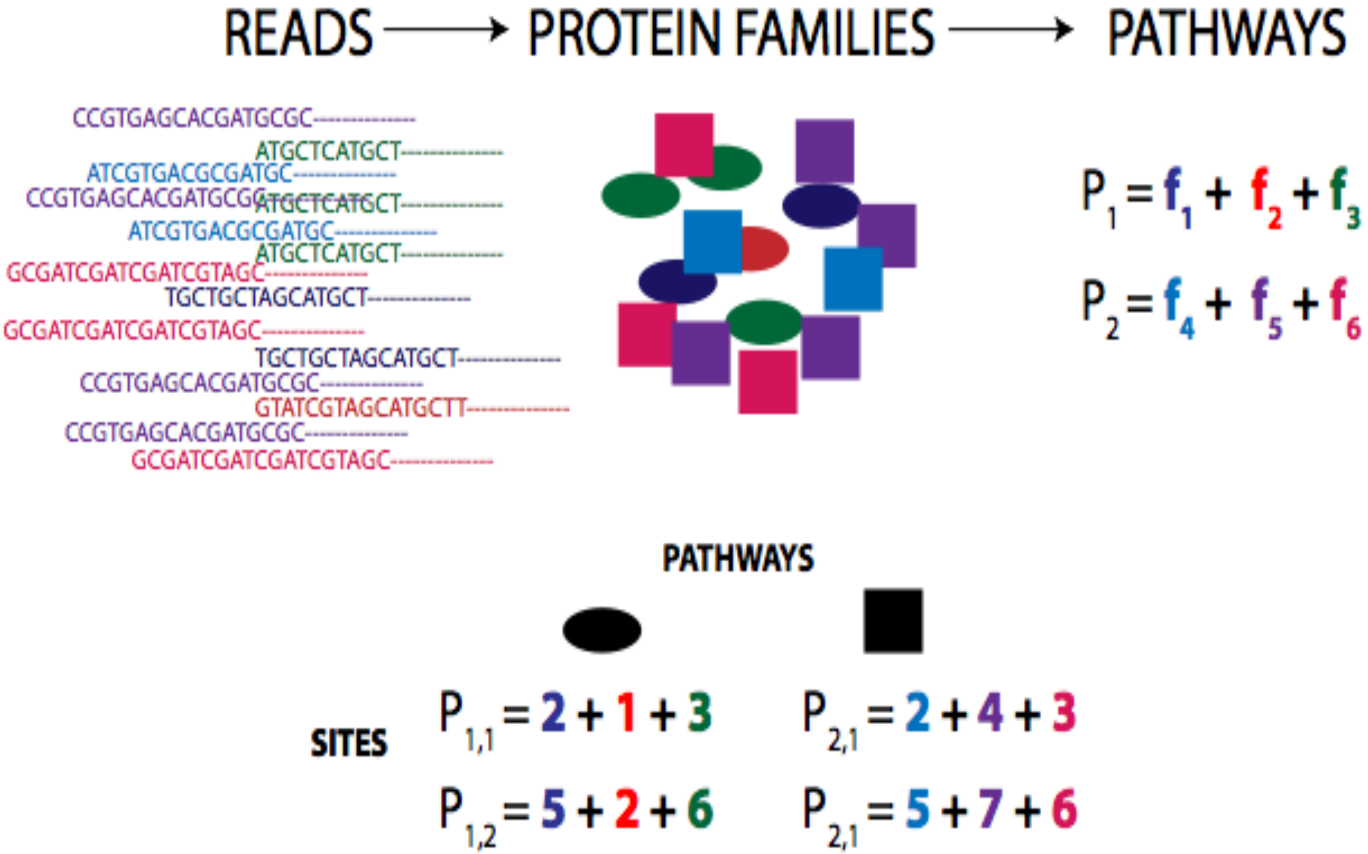
1 million CPU hours to process

Metabolic Pathways

Membrane Protein Families

Rusch, et al., *PLOS Biology* 2007

Mapping Raw Metagenomic Reads to a Matrix of Familes or Pathways for each Site



$$P_1 = f_1 + f_2 + f_3$$

$$P_2 = f_4 + f_5 + f_6$$

$$P_{1,1} = 2 + 1 + 3 \qquad P_{2,1} = 2 + 4 + 3$$

$$P_{1,2} = 5 + 2 + 6 \qquad P_{2,1} = 5 + 7 + 6$$

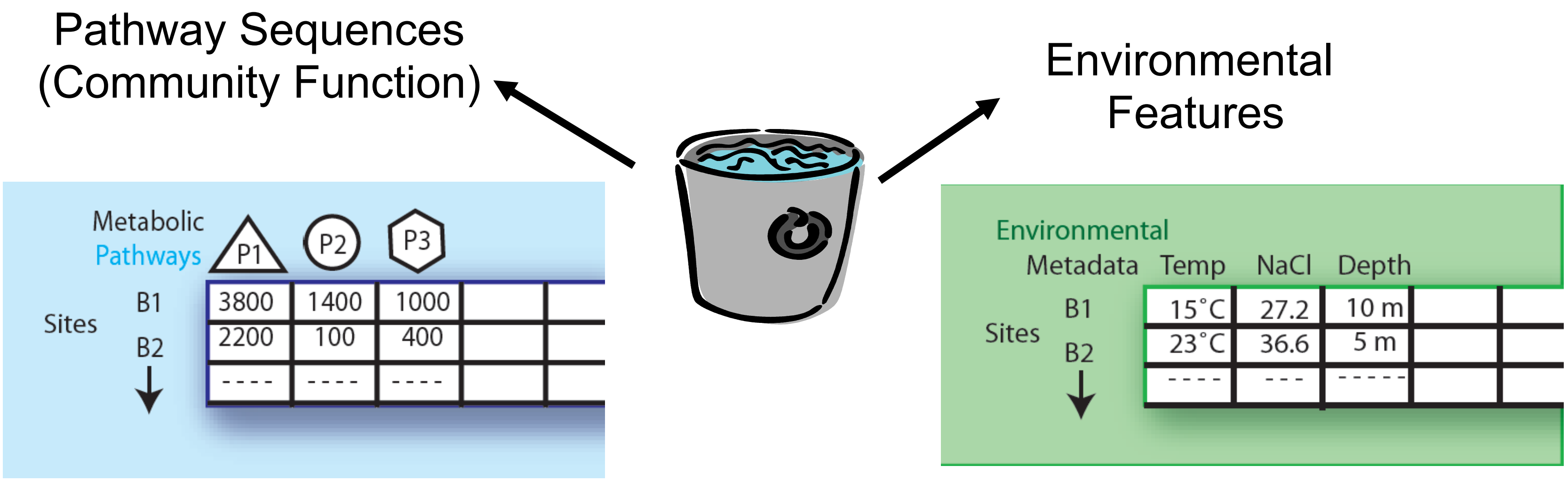
Canada
United States
Saragasso Sea
Atlantic Ocean
Mexico
Puerto Rico
Honduras
Costa Rica
Panama
Galapagos Islands
x 10 samples

Families Matrix

	Fam 1	Fam 2	...	...	Fam 151
Site 1	.01	.02			
Site 2	0	.01			
...					
Site 29					

$$\frac{\text{\# counts Fam 2}}{\text{\#total protein counts at site 2}}$$

Patel et. al., *Genome Research* 2010



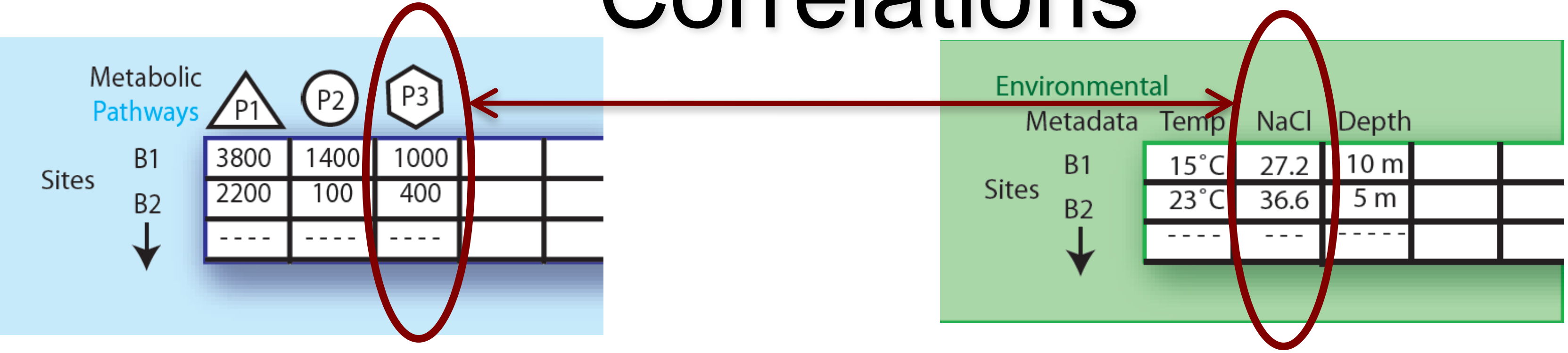
Sites	Metabolic Pathways	P1	P2	P3
	B1	3800	1400	1000
	B2	2200	100	400
	↓	----	----	----

Sites	Environmental Metadata	Temp	NaCl	Depth
	B1	15°C	27.2	10 m
	B2	23°C	36.6	5 m
	↓	----	---	-----

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

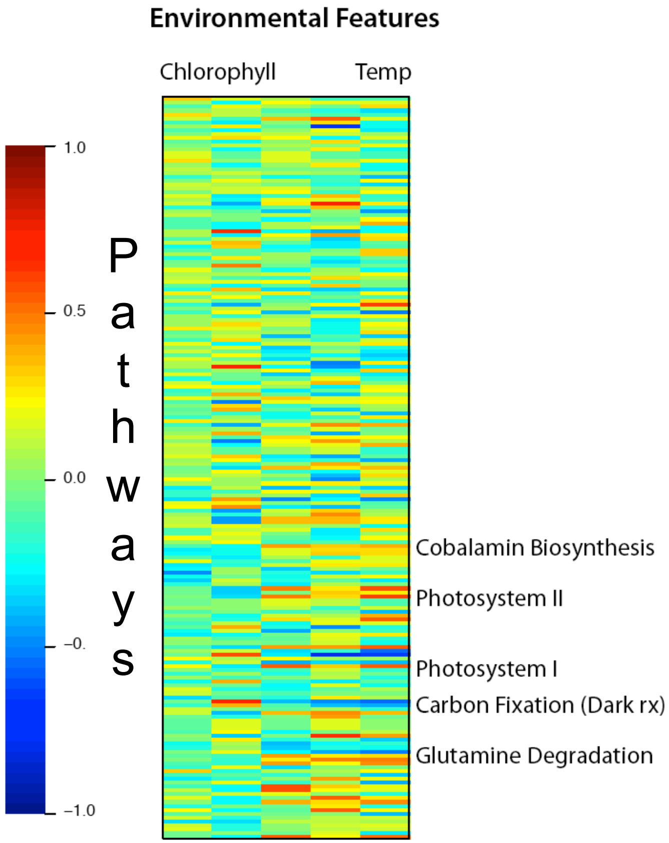
[Rusch et. al., (2007) PLOS Biology;
Gianoulis et al., PNAS (in press, 2009]

Simple Relationships: Pairwise Correlations



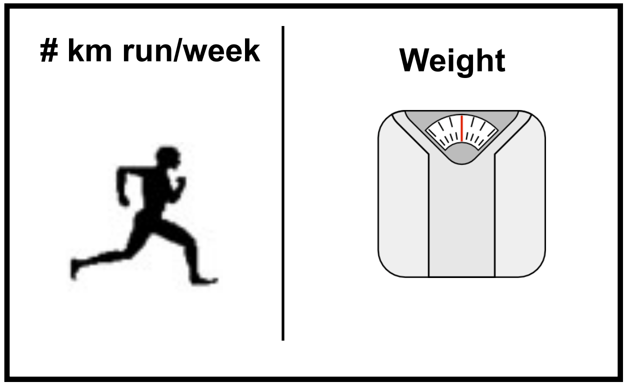
Metabolic Pathways	P1	P2	P3
Sites B1	3800	1400	1000
B2	2200	100	400
	----	----	----

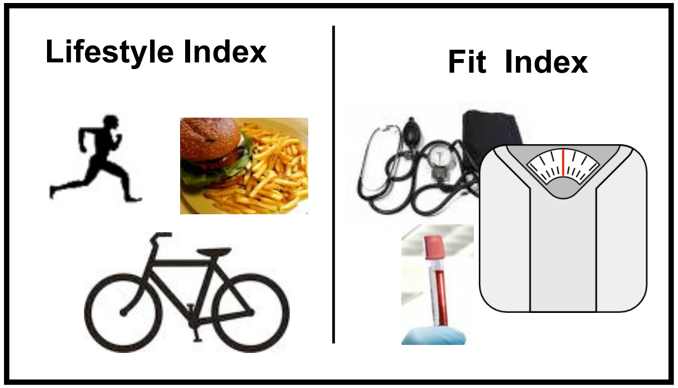
Environmental Metadata	Temp	NaCl	Depth
Sites B1	15˚C	27.2	10 m
B2	23˚C	36.6	5 m
	----	---	-----



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

Canonical Correlation Analysis: Simultaneous weighting

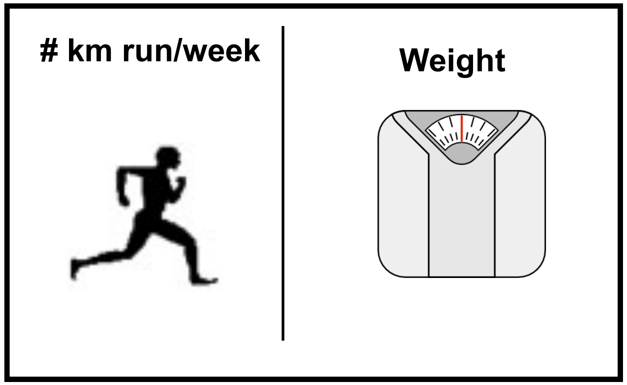


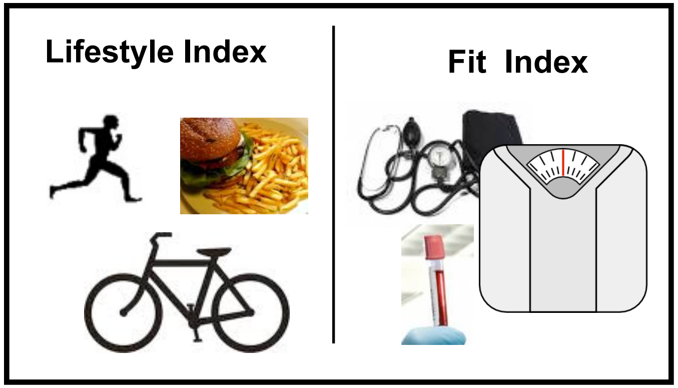


Lifestyle Index = a +b + c

Fit Index = a + b +c

Canonical Correlation Analysis: Simultaneous weighting





Life

Fit

Environmental Features

Temp

etc

Chlorophyll

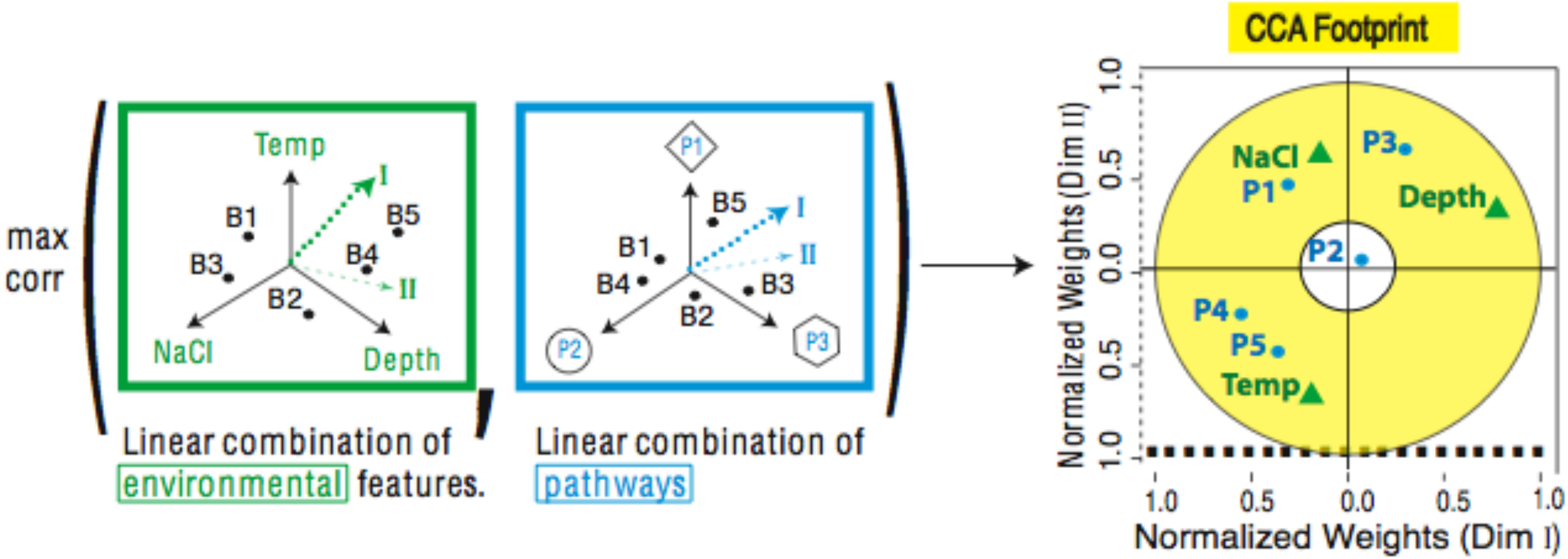
Metabolic Pathways/ Protein Families

Photosynthesis

etc

Lipid Metabolism

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


max corr
Temp
I
B1
B5
B3
B4
II
B2
NaCl
Depth
Linear combination of environmental features.
P1
B5
I
B1
II
B4
B3
B2
P2
P3
Linear combination of pathways
CCA Footprint
NaCl
P3
P1
Depth
P2
P4
P5
Temp
1.0
0.5
0.0
0.5
1.0
Normalized Weights (Dim II)
1.0
0.5
0.0
0.5
1.0
Normalized Weights (Dim I)

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

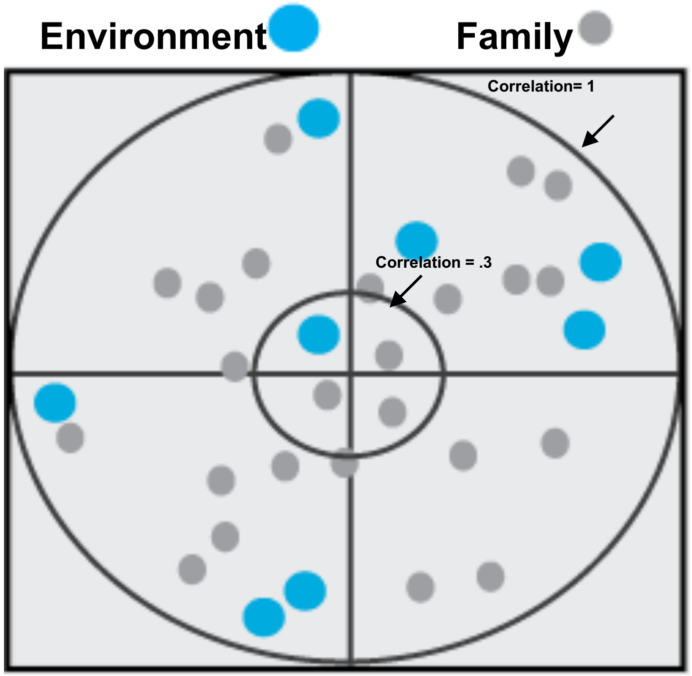
CCA results

We are defining a change of basis of the cross co-variance matrix

We want the correlations between the projections of the variables, X and Y, onto the basis vectors to be mutually maximized.

Eigenvalues→ squared canonical correlations

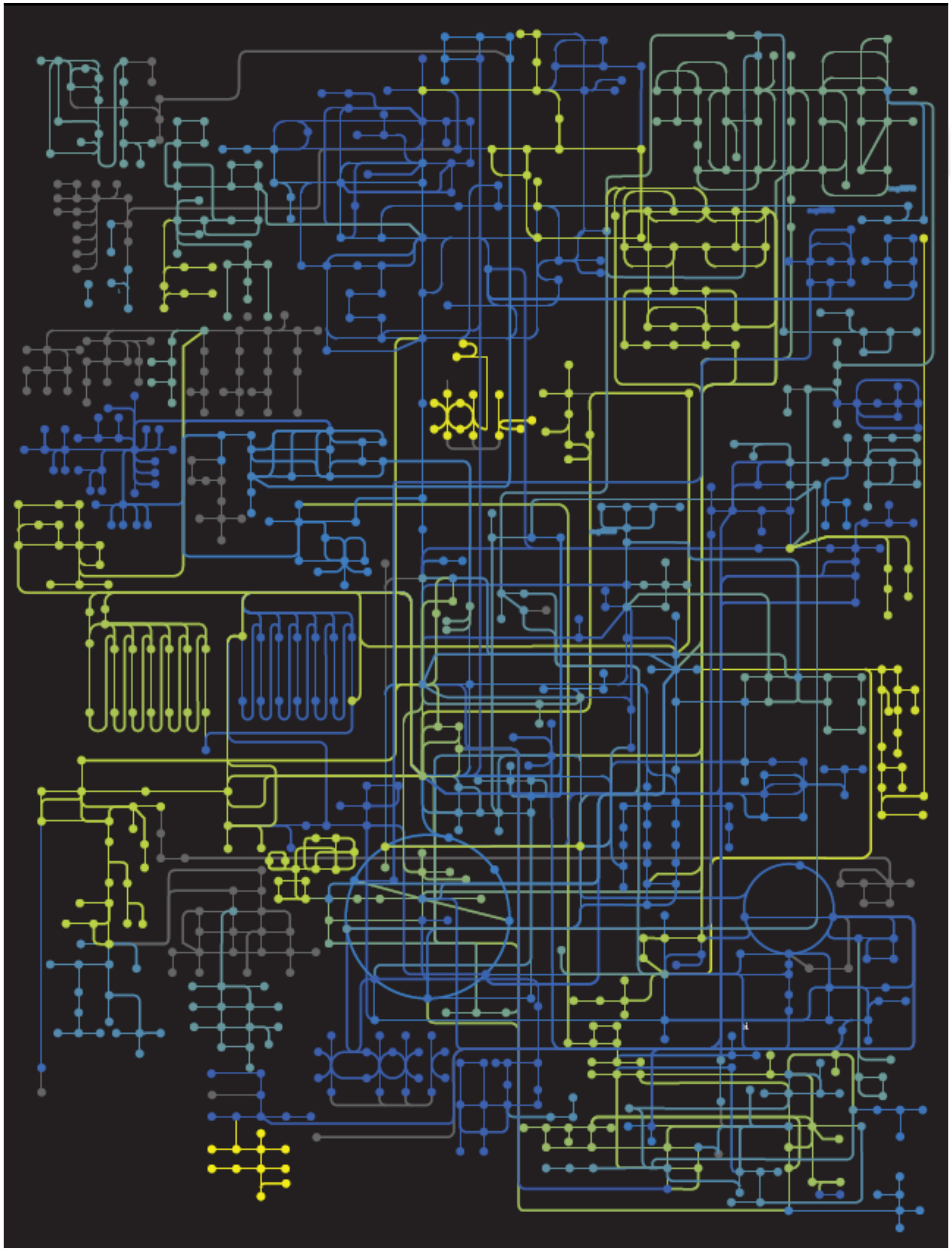
Eigenvectors→ normalized canonical correlation *basis vectors*



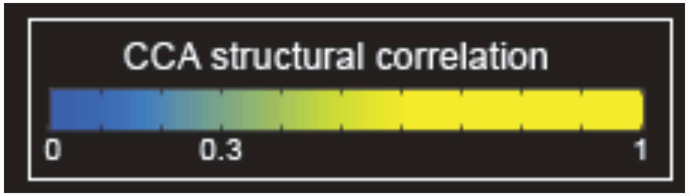
This plot shows the correlations in the first and second dimensions

Correlation Circle: The closer the point is to the outer circle, the higher the correlation

Variables projected in the same direction are correlated

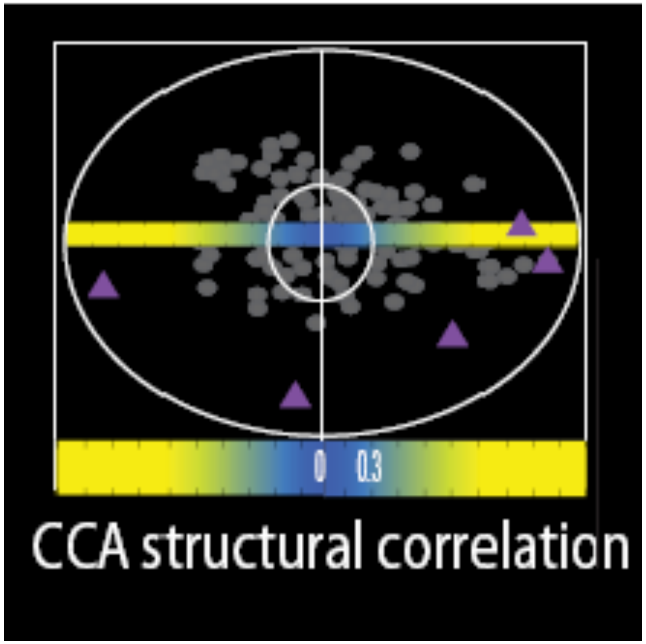


Strength of Pathway co-variation with environment



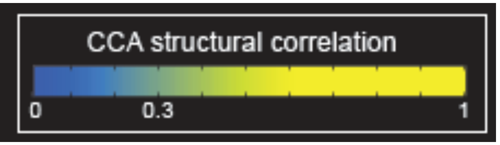
Environmentally invariant

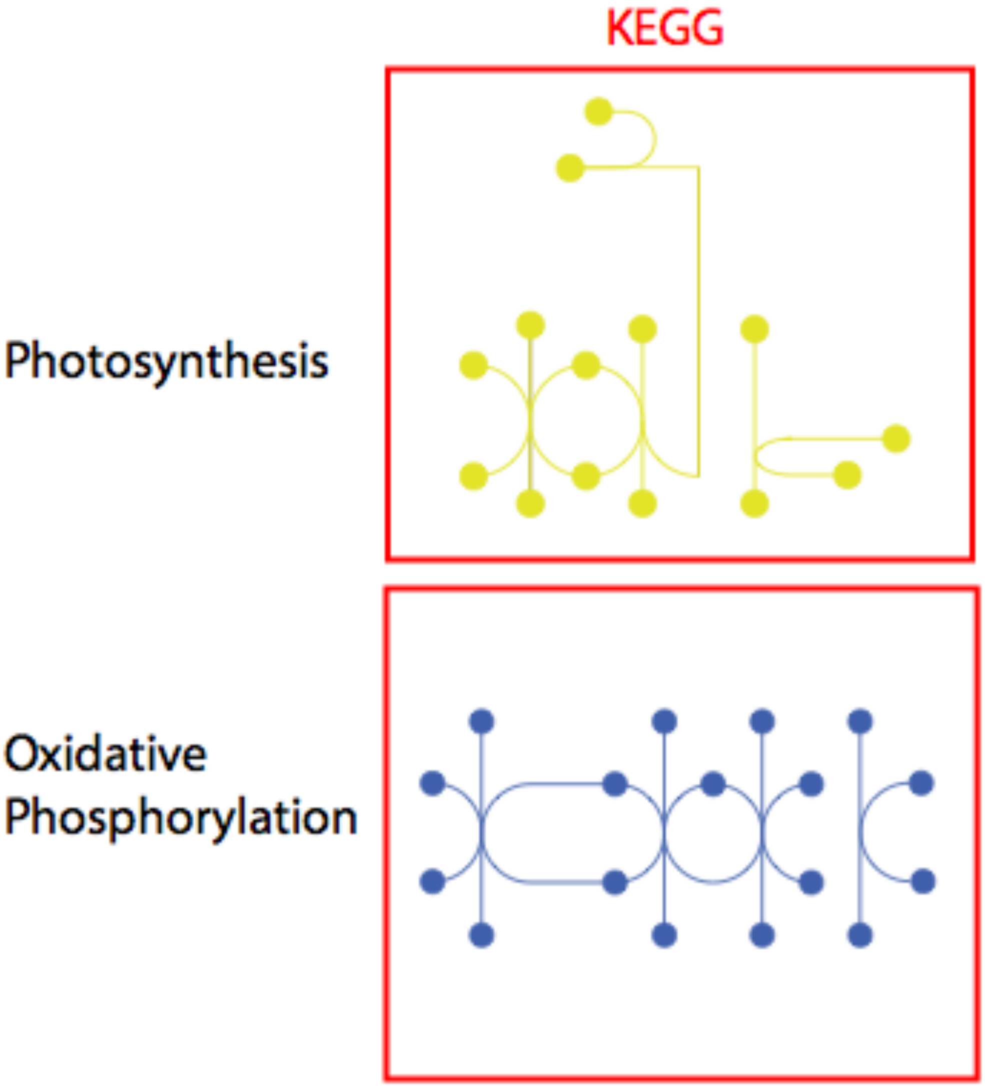
Environmentally variant



Gianoulis et al., *PNAS* 2009

Conclusion #1: energy conversion strategy, temp and depth




KEGG

Photosynthesis

Oxidative Phosphorylation

Gianoulis et al., *PNAS* 2009