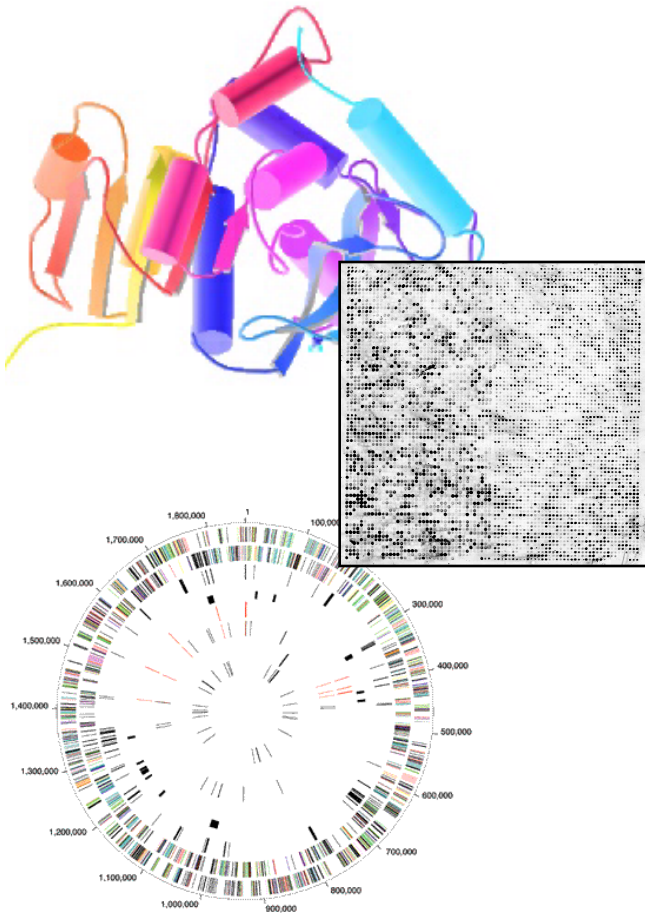


Biomed. Data Science: Unsupervised Datamining



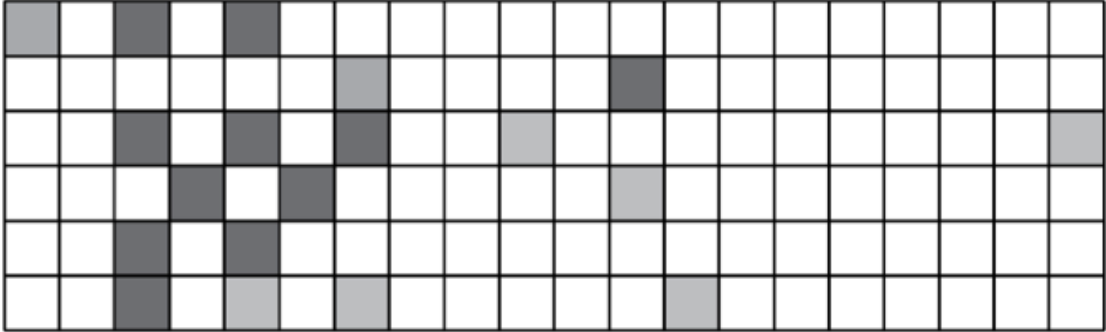
Mark Gerstein, Yale University
gersteinlab.org/courses/452
(last edit in spring '20, pack #9)

Structure of Genomic Features Matrix

1

Sites along the genome

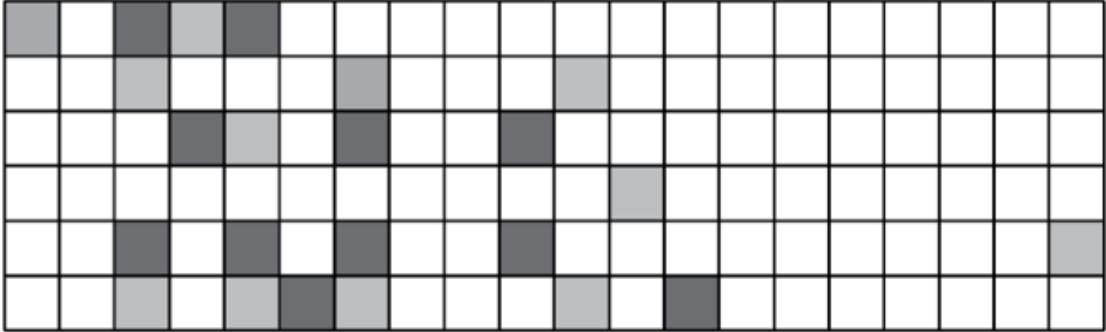
Factors
and
Chromatin
Modifications
(different
tissues)



...

⋮ ⋮

RNA
(different
tissues)



...

Unsupervised Mining

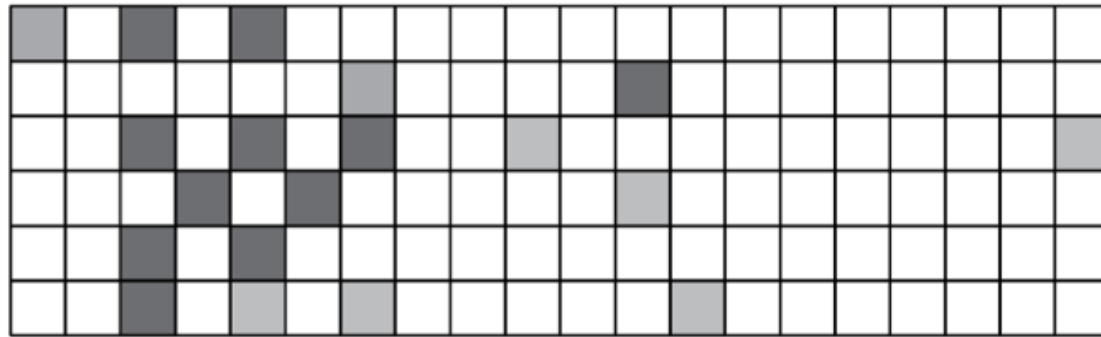
- Simple overlaps & enriched regions
- Clustering rows & columns (networks)
- PCA/SVD (theory + appl.)
- Biplot
- RCA
- CCA
- tSNE
- LDA
- (Variational Autoencoders)

Genomic Features Matrix: Deserts & Forests

1

Sites along the genome

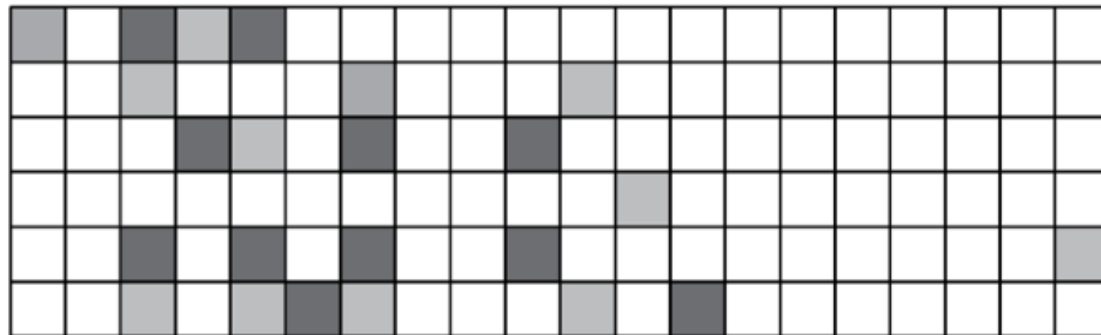
Factors
and
Chromatin
Modifications
(different
tissues)



...

⋮ ⋮

RNA
(different
tissues)



...

⋮ ⋮



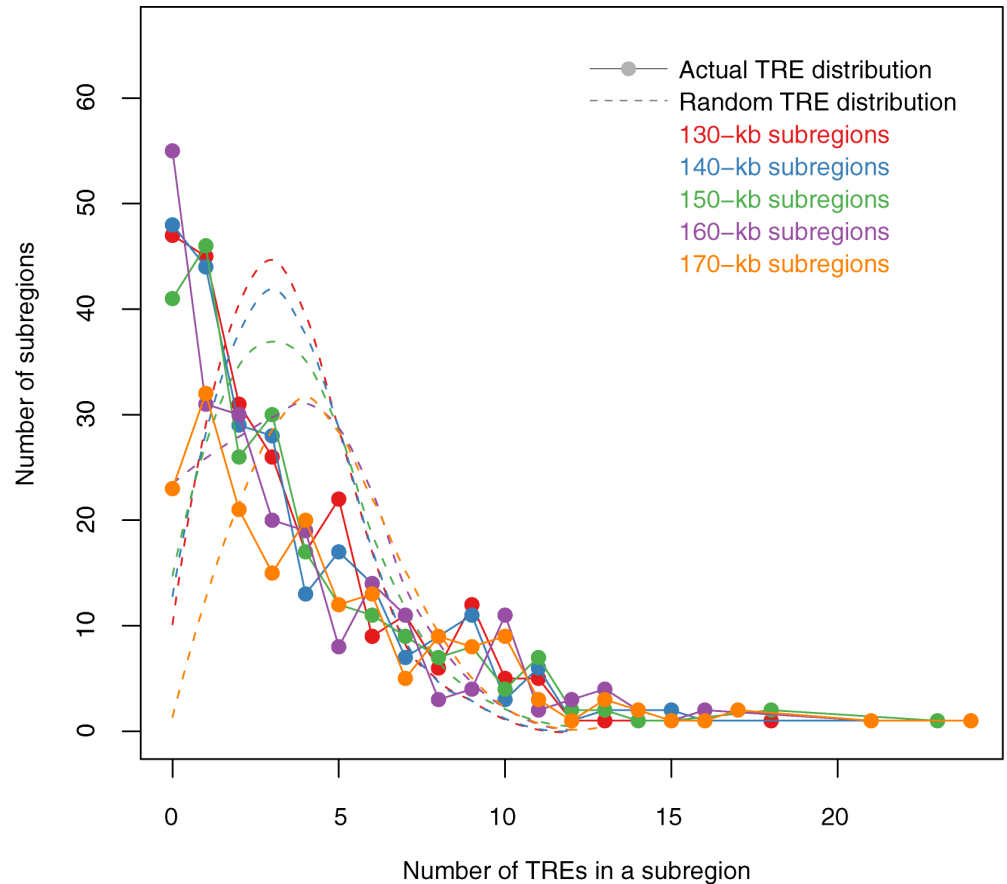
Forest



Desert

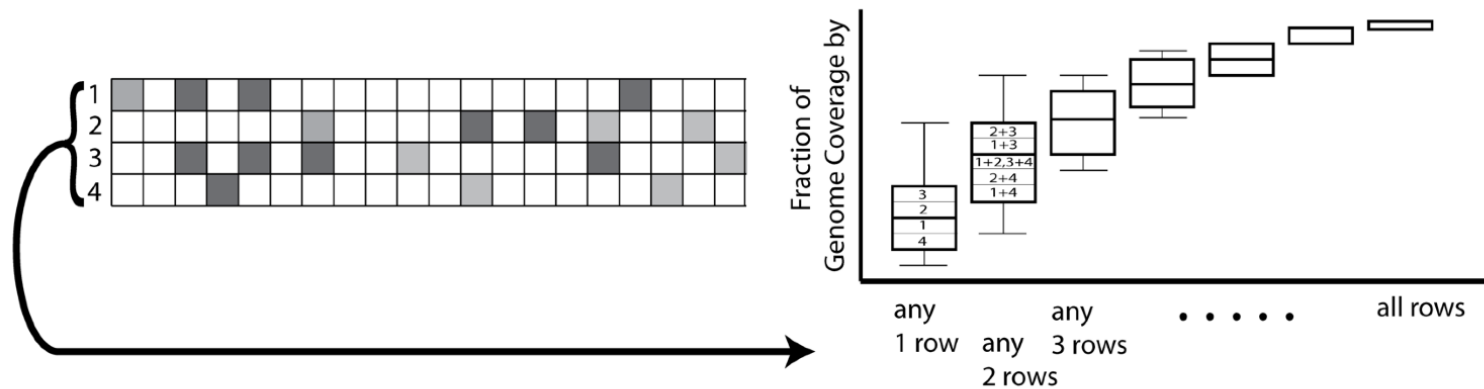
Non-random distribution of TREs

- 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.

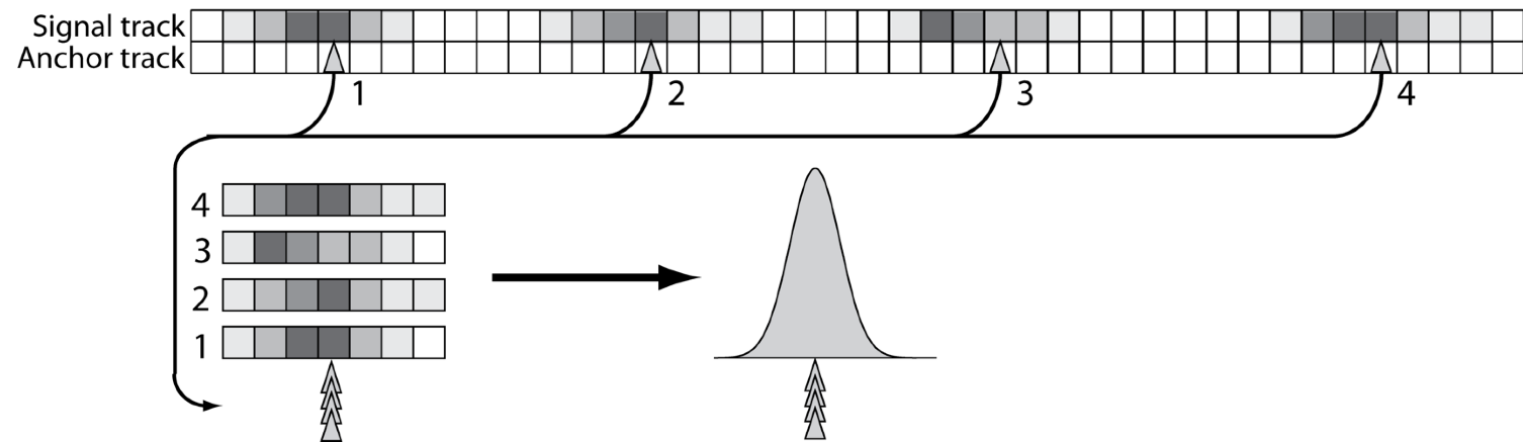


Aggregation & Saturation

B Saturation Analysis



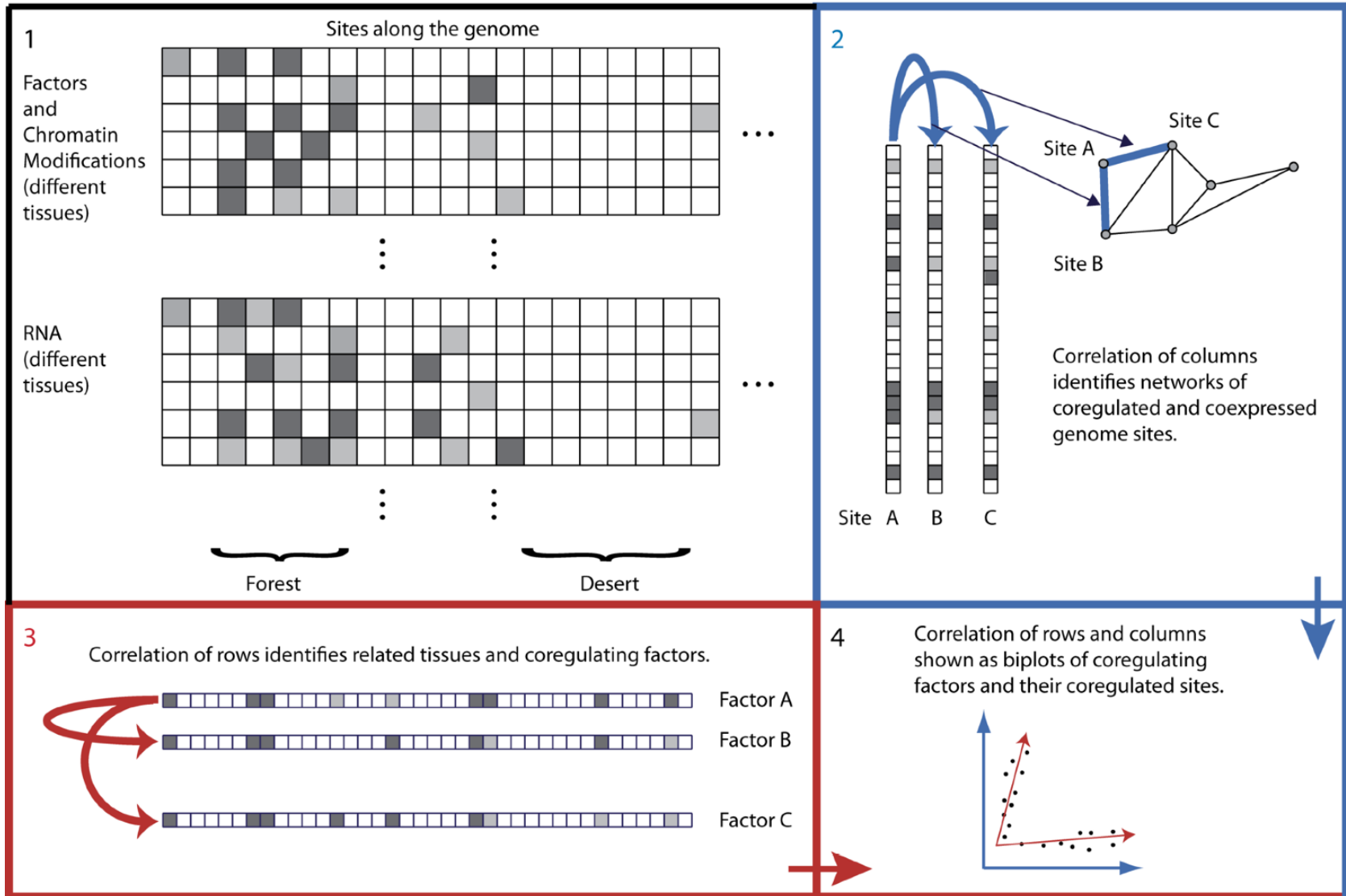
C Aggregation Analysis



Unsupervised Mining

Clustering Columns & Rows of the
Data Matrix

Correlating Rows & Columns



Spectral Methods

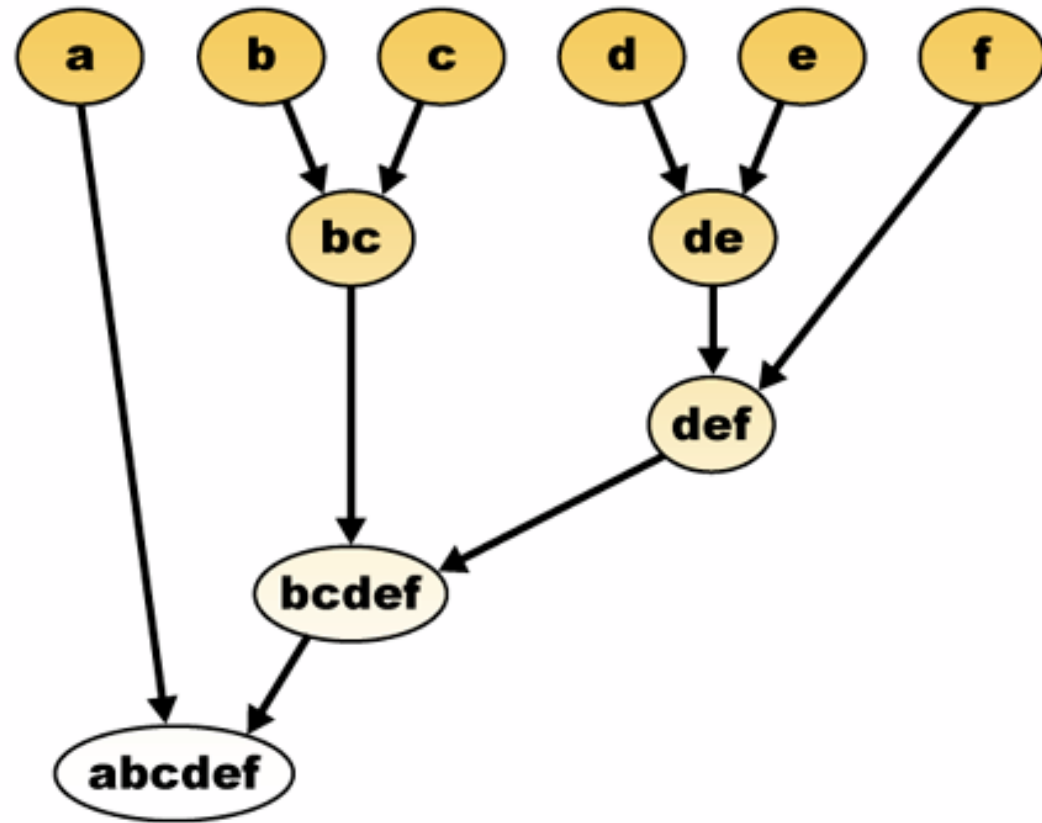
Outline & Papers

- 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.

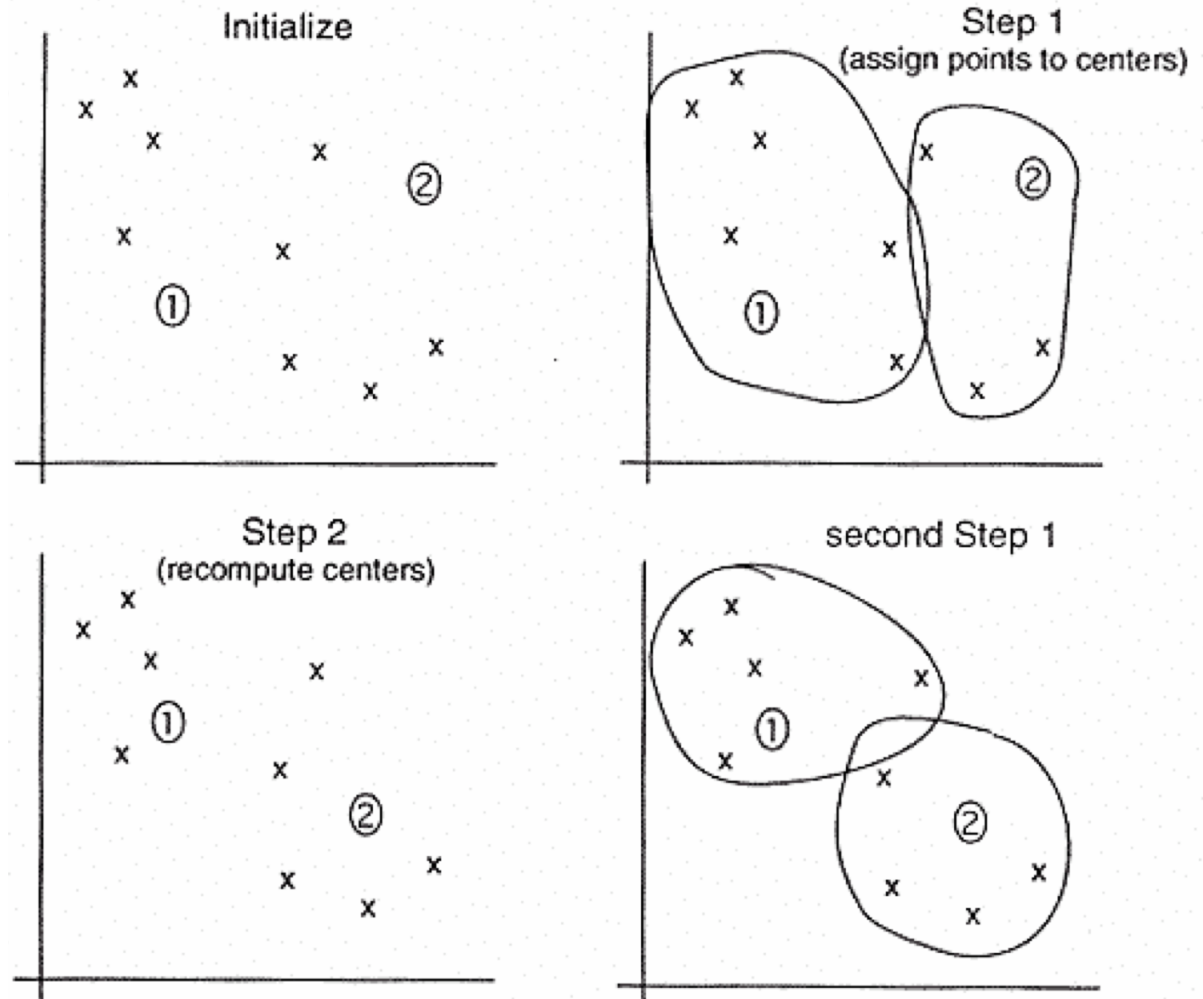
Expression Clustering

Agglomerative Clustering

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

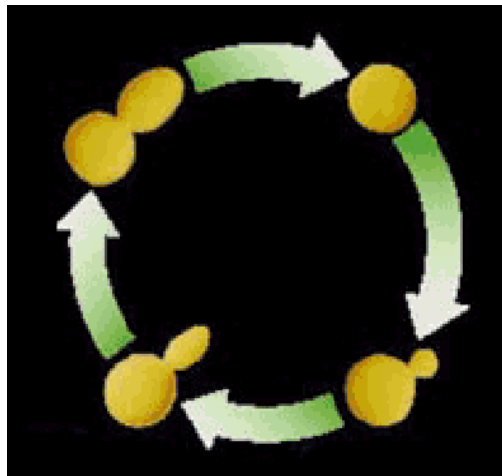


K-means

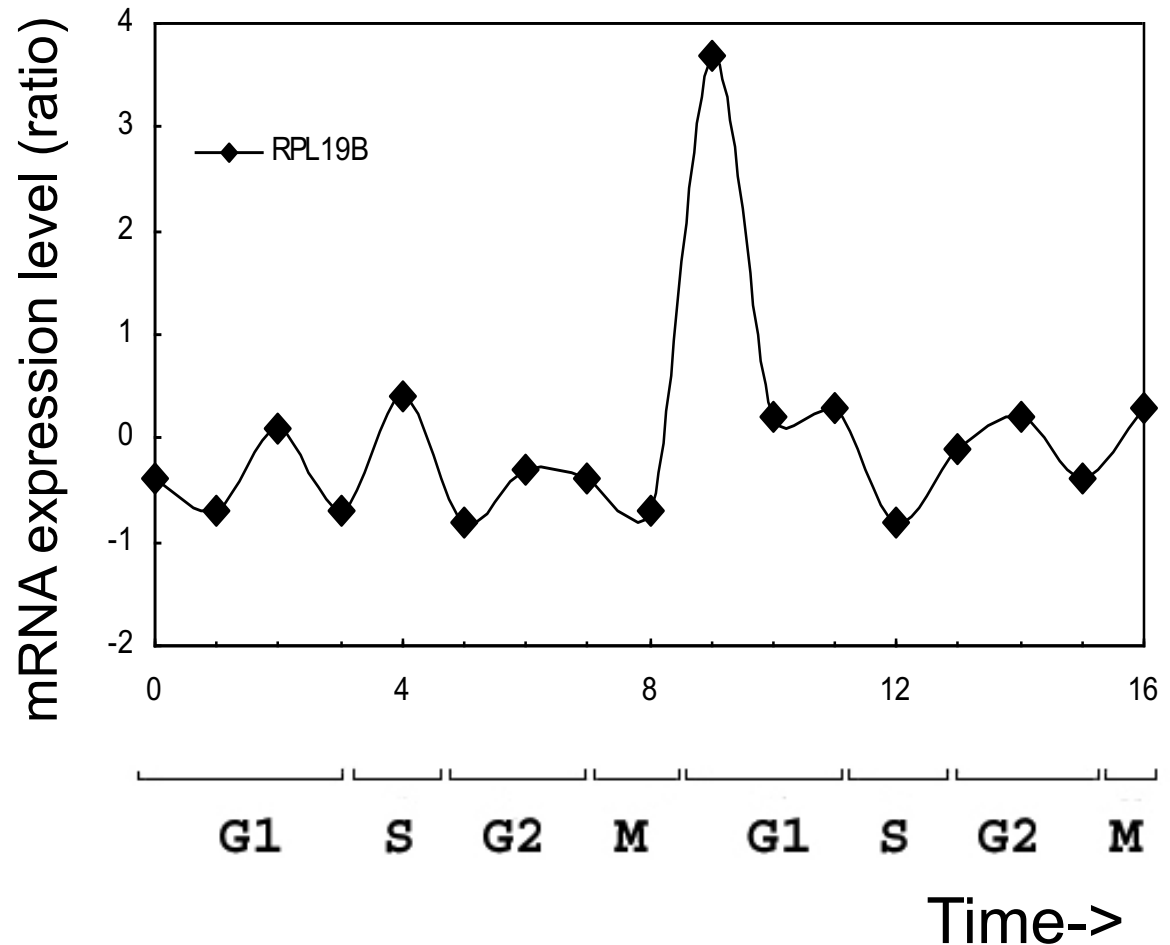


- 1) Pick ten (i.e. k ?) random points as putative cluster centers.
- 2) Group the points to be clustered by the center to which they are closest.
- 3) Then take the mean of each group and repeat, with the means now at the cluster center.
- 4) Stop when the centers stop moving.

Clustering the yeast cell cycle to uncover interacting proteins

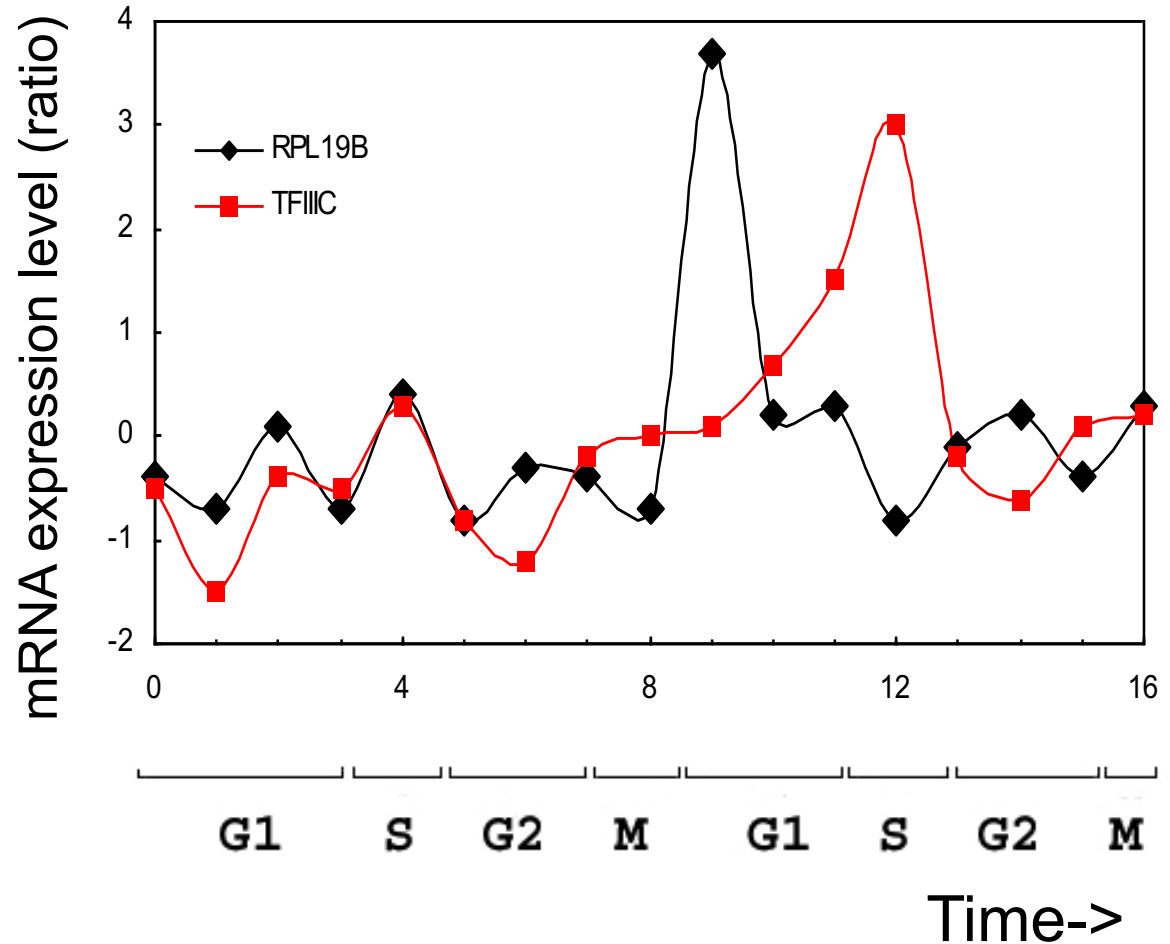
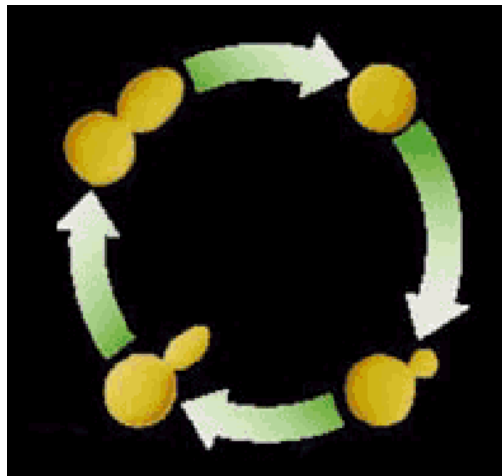


[Brown, Davis]



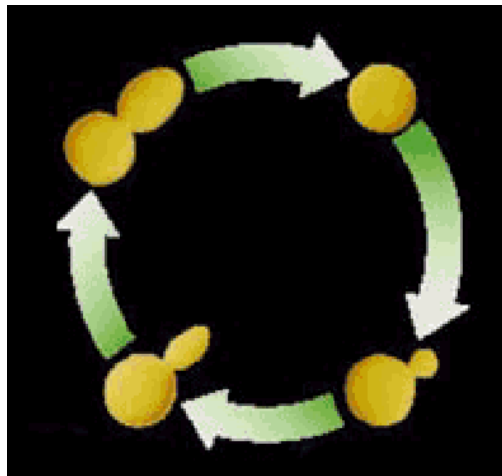
Microarray timecourse of
1 ribosomal protein

Clustering the yeast cell cycle to uncover interacting proteins

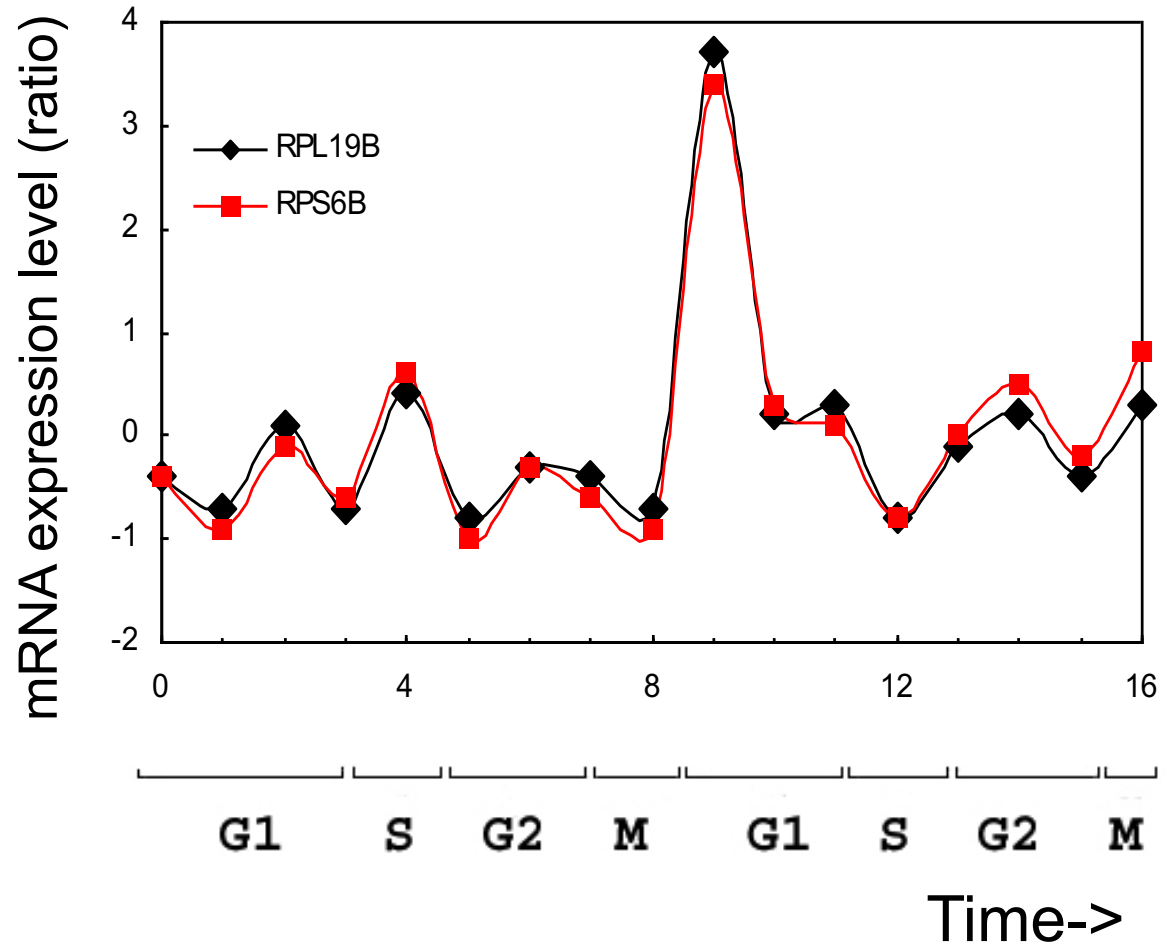


Random relationship from ~18M

Clustering the yeast cell cycle to uncover interacting proteins

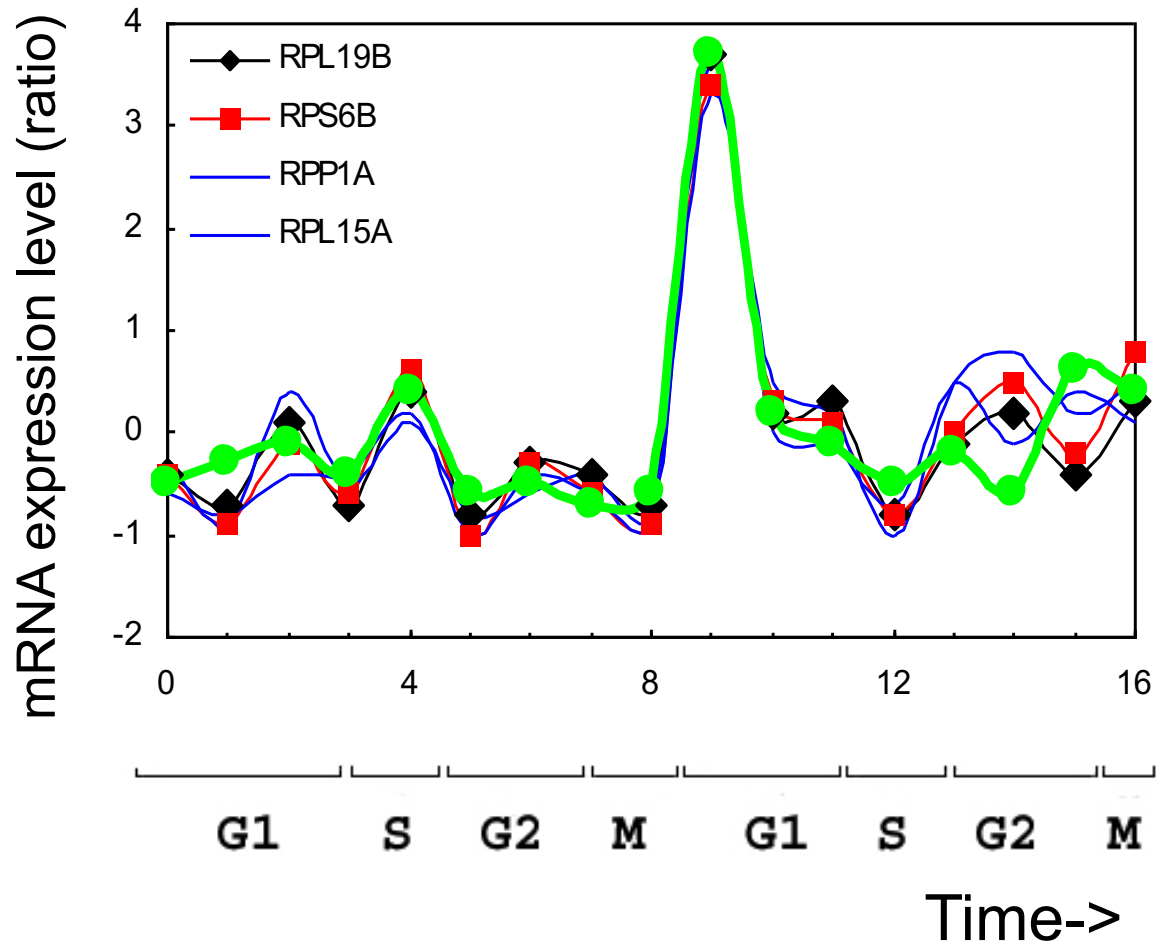
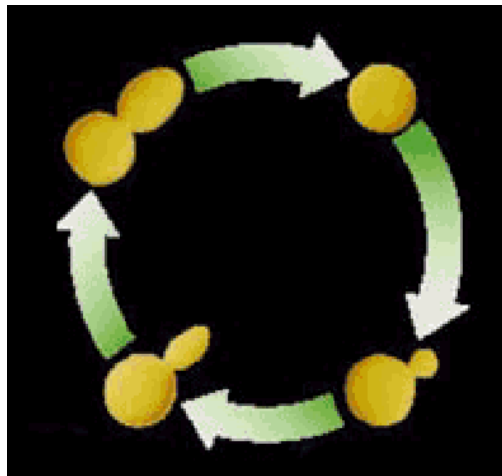


[Botstein; Church, Vidal]



Close relationship from 18M
(2 Interacting Ribosomal Proteins)

Clustering
the
yeast cell
cycle to
uncover
interacting
proteins

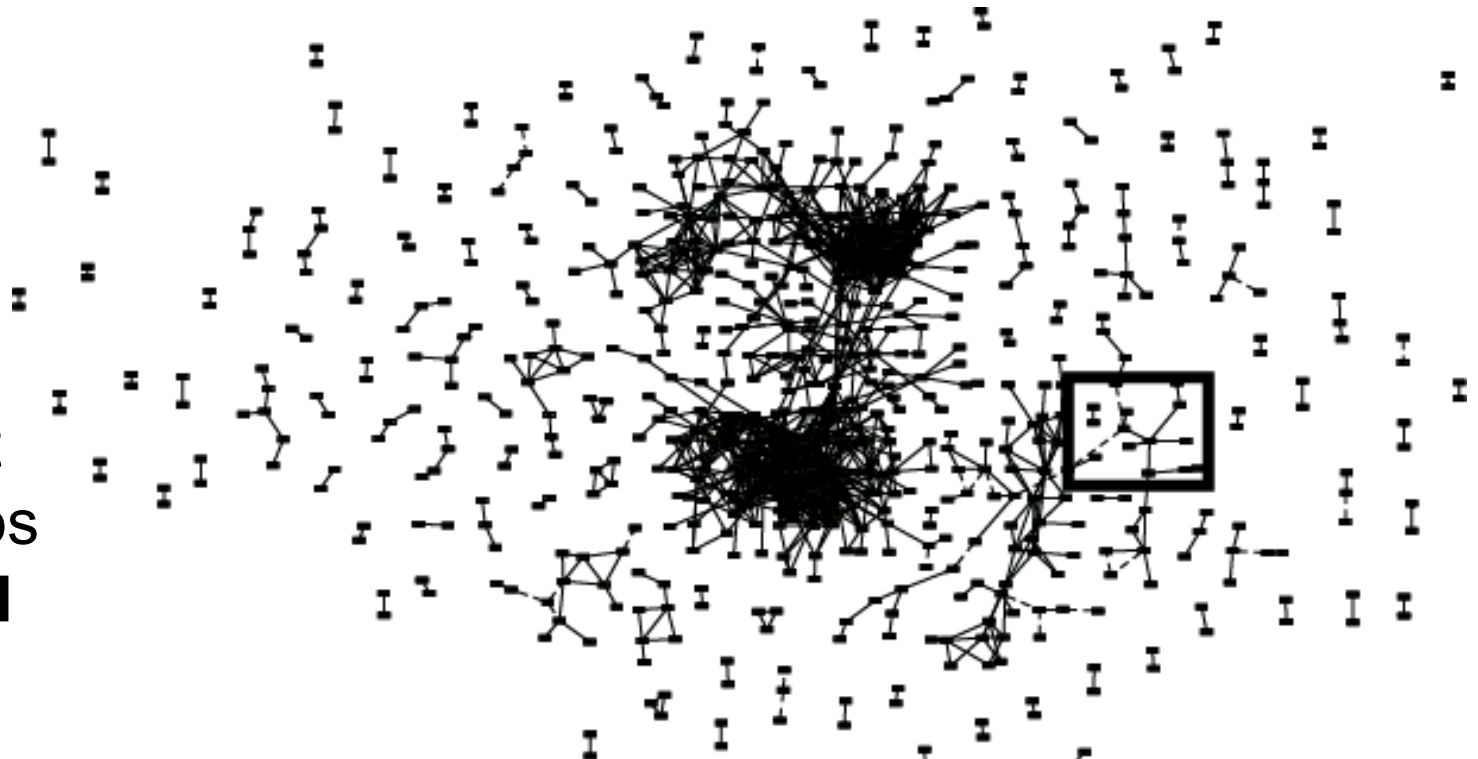


Predict Functional Interaction of
Unknown Member of Cluster



Global Network of Relationships

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

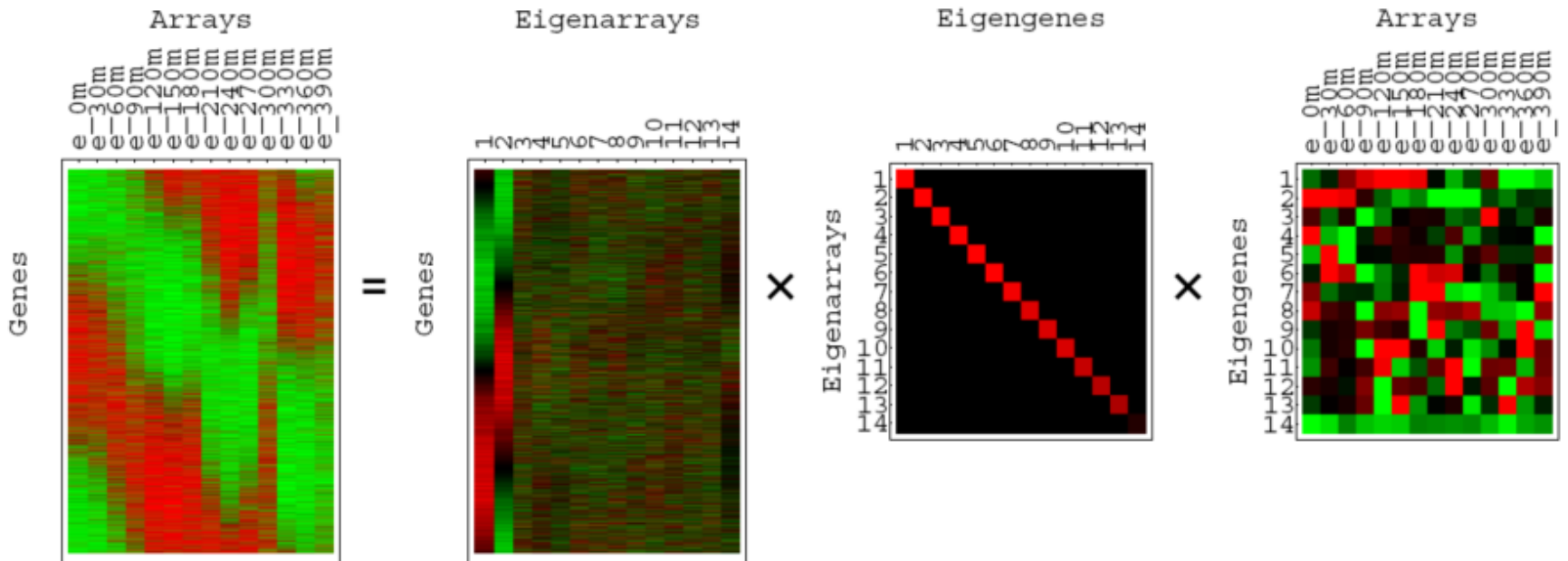


Unsupervised Mining

SVD

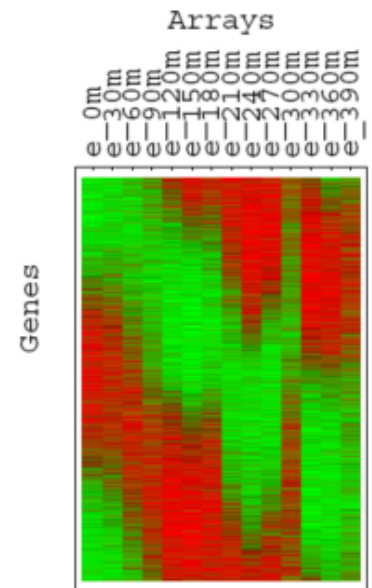
Puts together slides prepared by
Brandon Xia with images from
Alter et al. papers

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



$$A = USV^T$$

- A is any rectangular matrix ($m \geq n$)
- 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: $r (\leq n)$
- A is a linear transformation that maps vector x in row space into vector Ax in column space

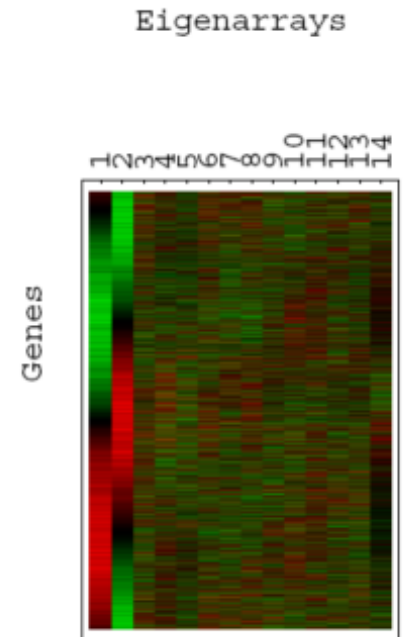


$$A = USV^T$$

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

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

- $\mathbf{u}_1, \dots, \mathbf{u}_n$ in U are eigenvectors of AA^T
 - $AA^T = USV^T V S U^T = US^2 U^T$
 - “Left singular vectors”

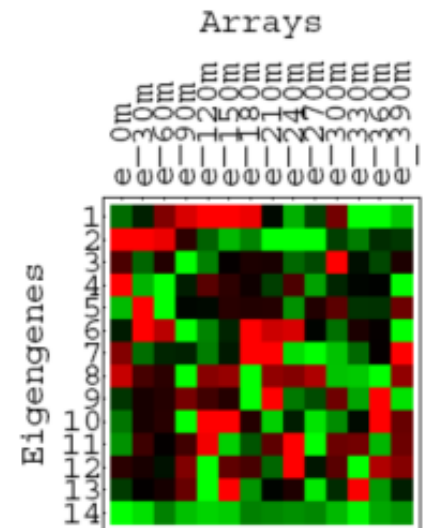


$$A = USV^T$$

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

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

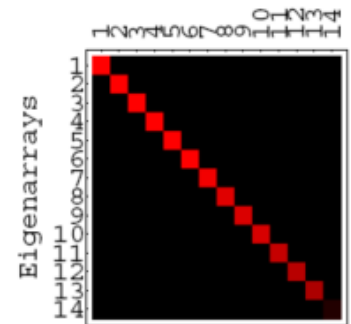
- $\mathbf{v}_1, \dots, \mathbf{v}_n$ in V are eigenvectors of $A^T A$
 - $A^T A = V S U^T U S V^T = V S^2 V^T$
 - “Right singular vectors”



$$A = USV^T$$

- 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 $A^T A$ and AA^T

Eigenvalues



$$AV = US$$

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

SVD as sum of rank-1 matrices

- $A = USV^T$
- $A = s_1 \mathbf{u}_1 \mathbf{v}_1^T + s_2 \mathbf{u}_2 \mathbf{v}_2^T + \dots + s_n \mathbf{u}_n \mathbf{v}_n^T$
- $s_1 \geq s_2 \geq \dots \geq s_n \geq 0$

an outer product ($\mathbf{u}\mathbf{v}^T$) giving a matrix rather than the scalar of the inner product

- What is the rank- r matrix \hat{A} that best approximates A ?

– Minimize
$$\sum_{i=1}^m \sum_{j=1}^n (\hat{A}_{ij} - A_{ij})^2$$

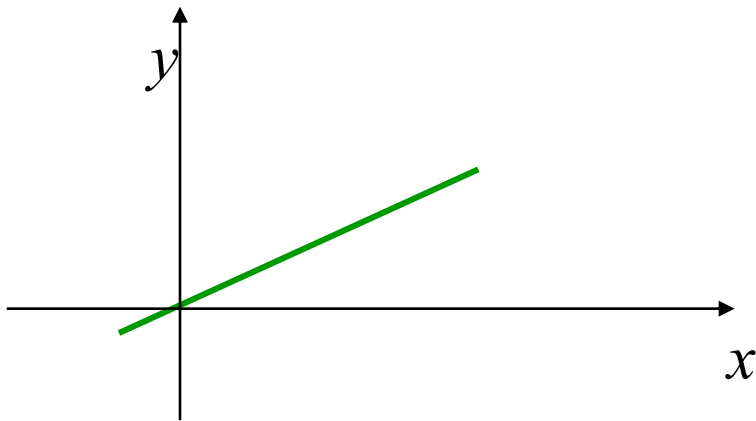
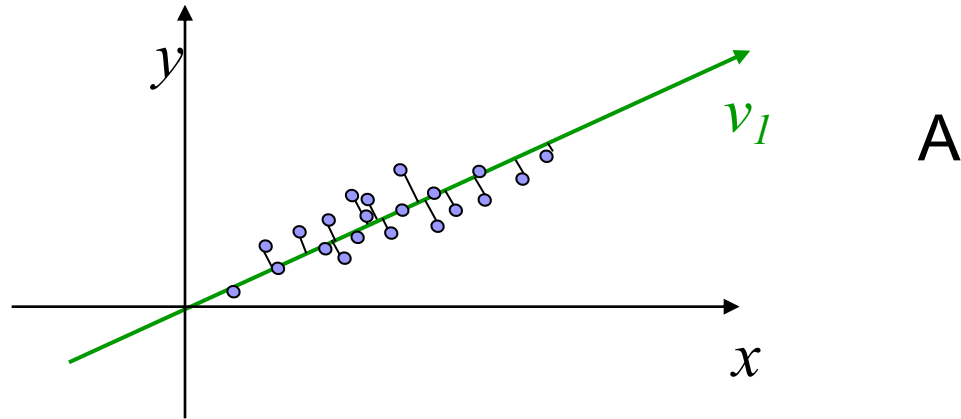
LSQ approx. If $r=1$, this amounts to a line fit.

- $\hat{A} = s_1 \mathbf{u}_1 \mathbf{v}_1^T + s_2 \mathbf{u}_2 \mathbf{v}_2^T + \dots + s_r \mathbf{u}_r \mathbf{v}_r^T$
- Very useful for matrix approximation

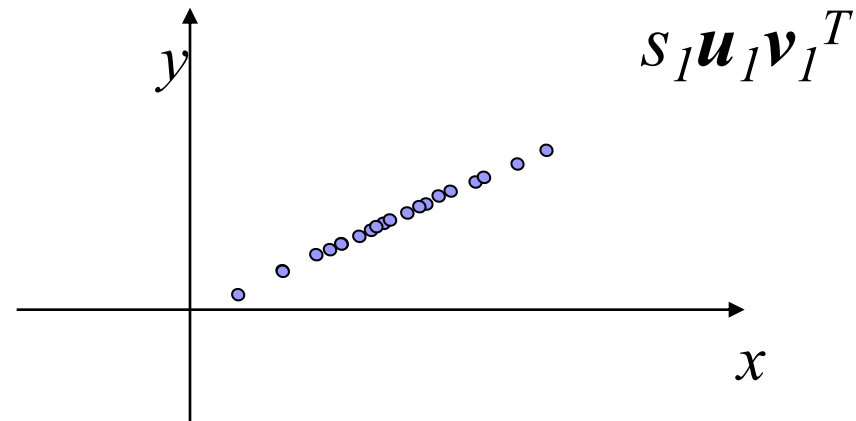
Examples of (almost) rank-1 matrices

- 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



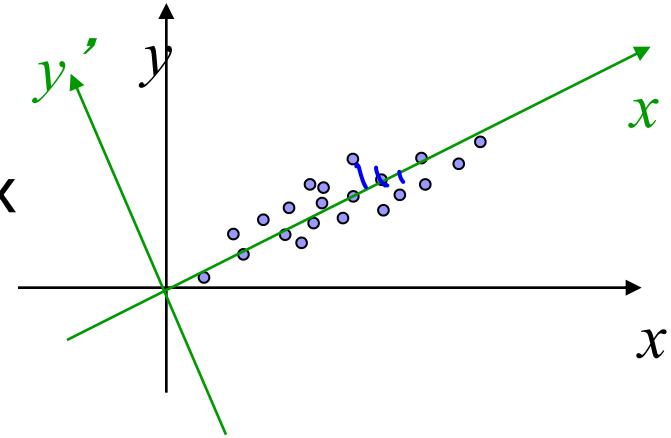
This line segment that goes through origin approximates the original data set



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



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

$$I \mathbf{v}_i = \mathbf{v}_i$$

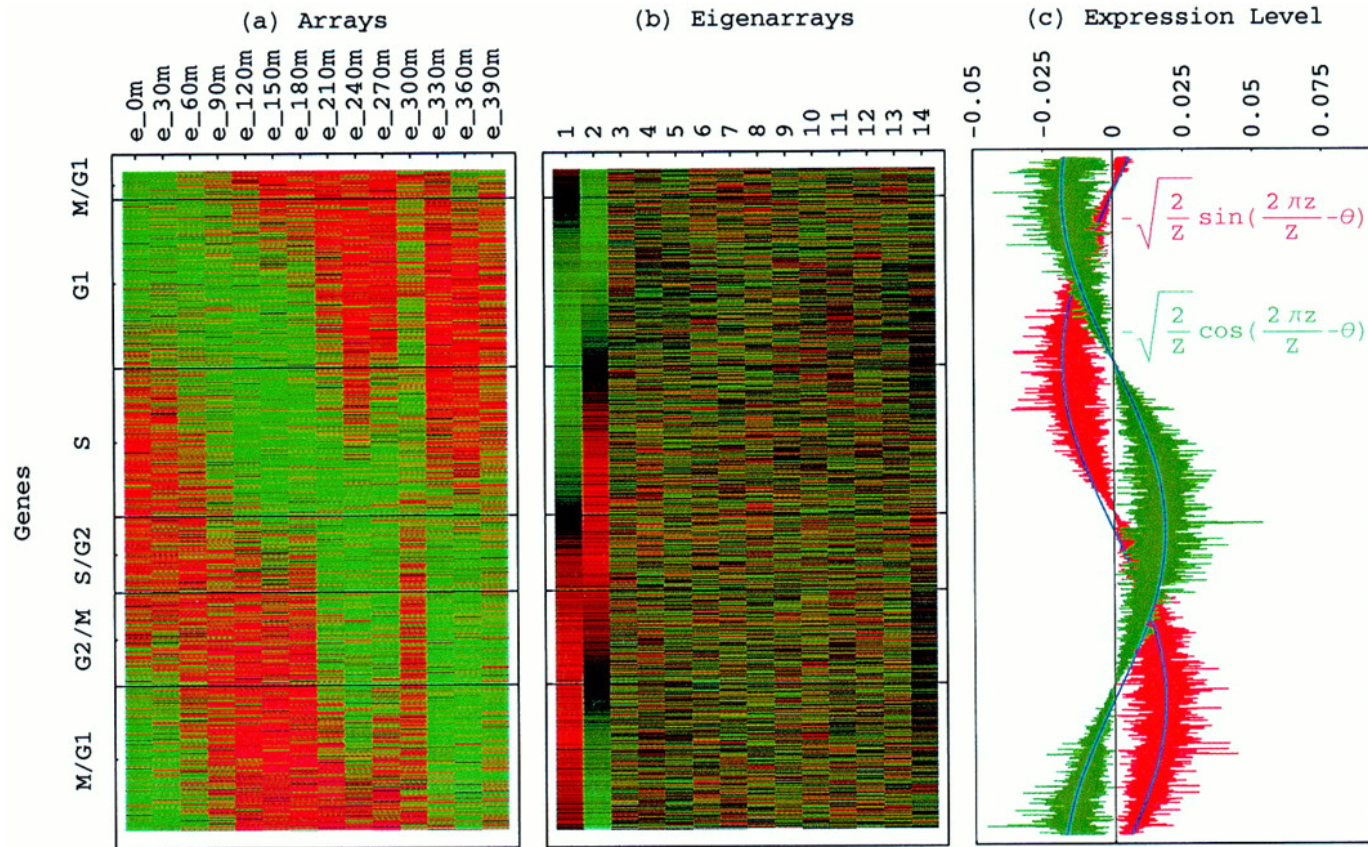
What about geometry of SVD in column space?

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

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

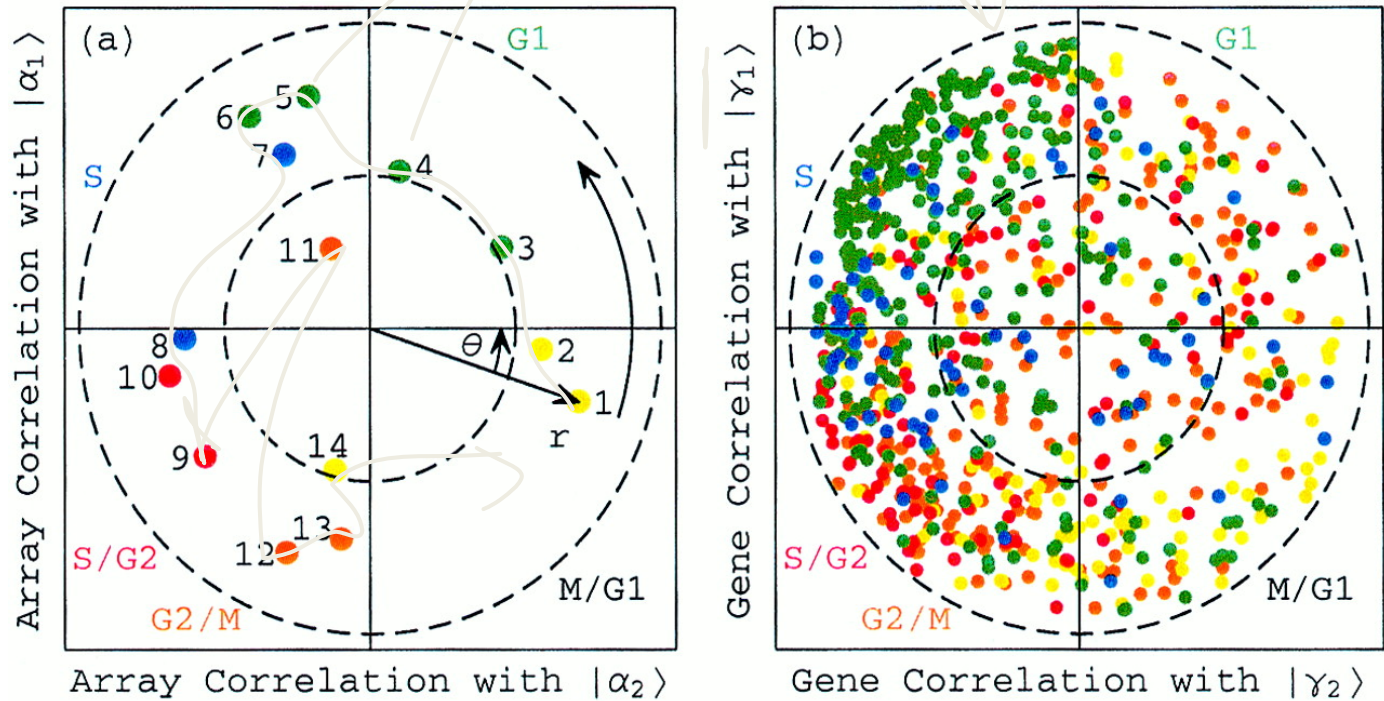


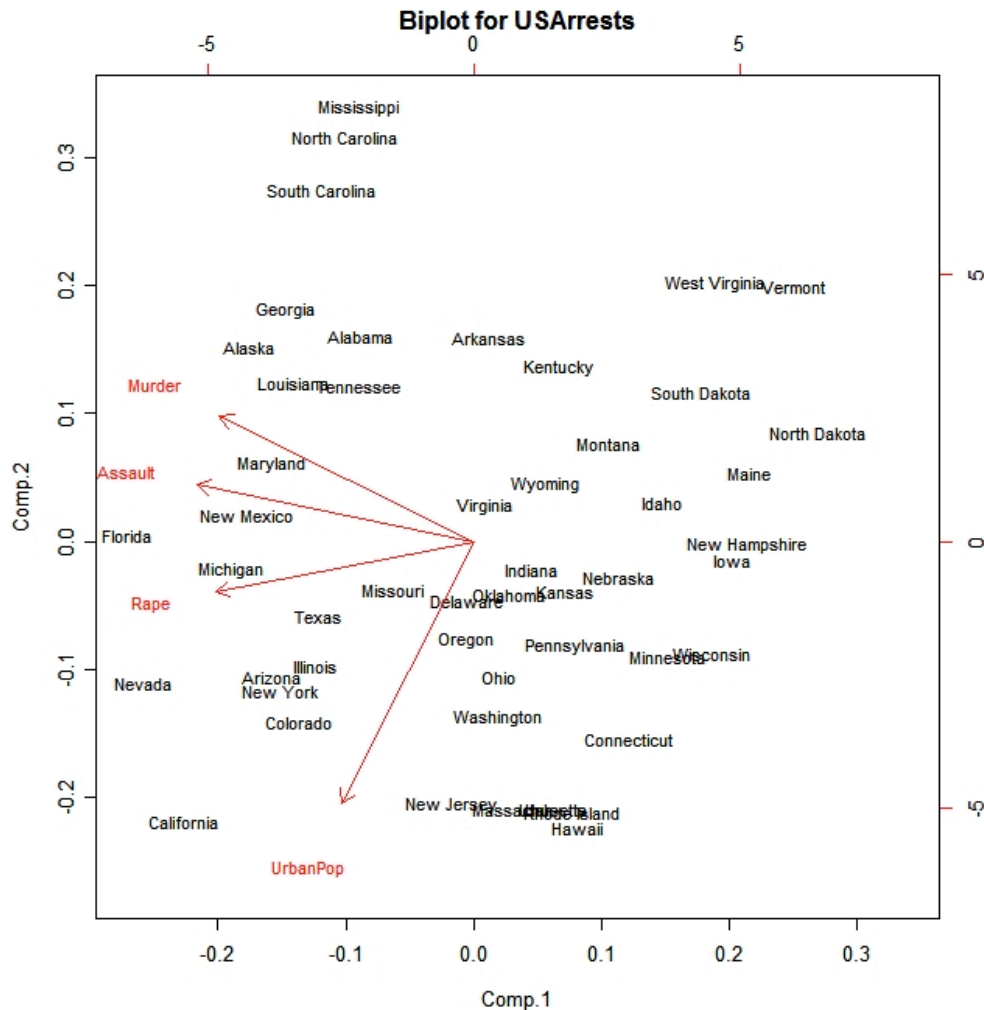
Fig. 2. Normalized elutriation expression in the subspace associated with the cell cycle. (a) Array correlation with $|\alpha_1\rangle_N$ along the y-axis vs. that with $|\alpha_2\rangle_N$ along the x-axis, color-coded according to the classification of the arrays into the five cell cycle stages, M/G₁ (yellow), G₁ (green), S (blue), S/G₂ (red), and G₂/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

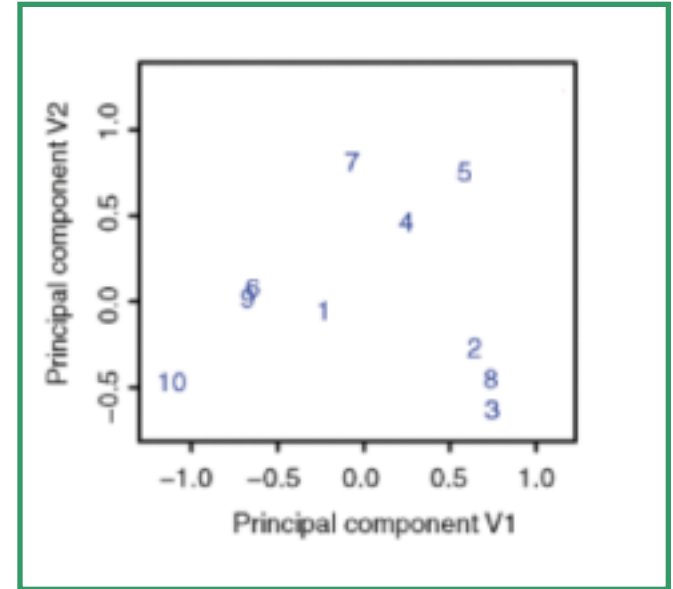
Sites: 1,2,3...

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

$A^T A$ (TF-TF corr.)

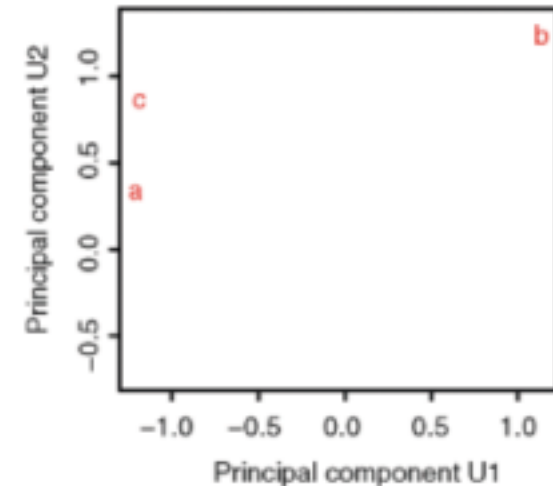


A^T

	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

$A A^T$ (site-site correlation)



Biplot to Show Overall Relationship of TFs & Sites

TFs: a, b, c...

Genomic

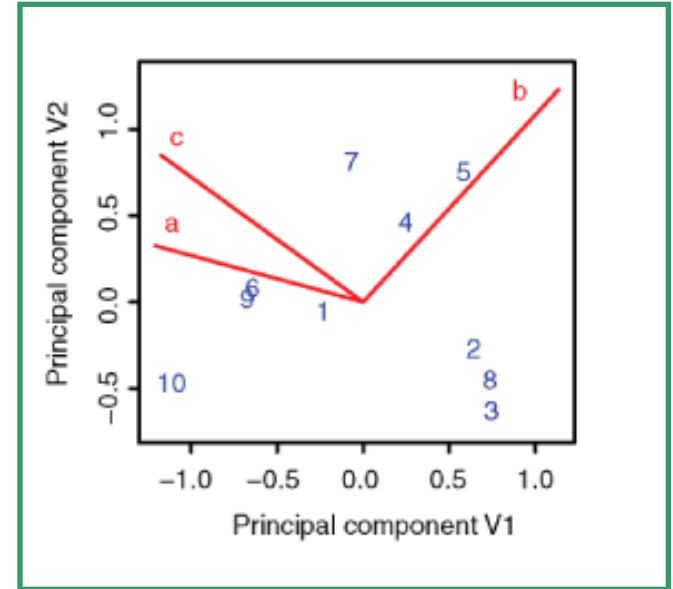
Sites: 1,2,3...

$$A = USV^T$$

	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

$A^T A$ (TF-TF corr.)

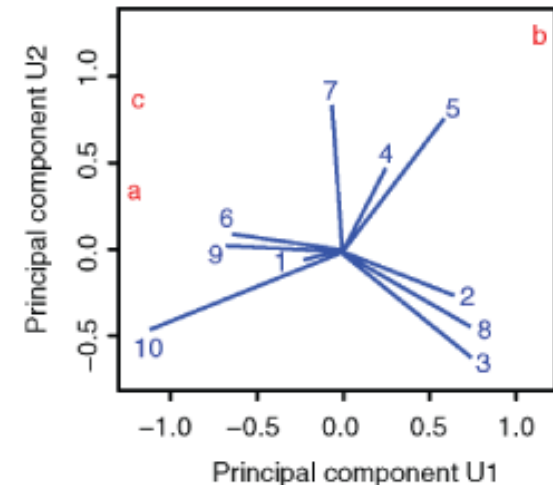


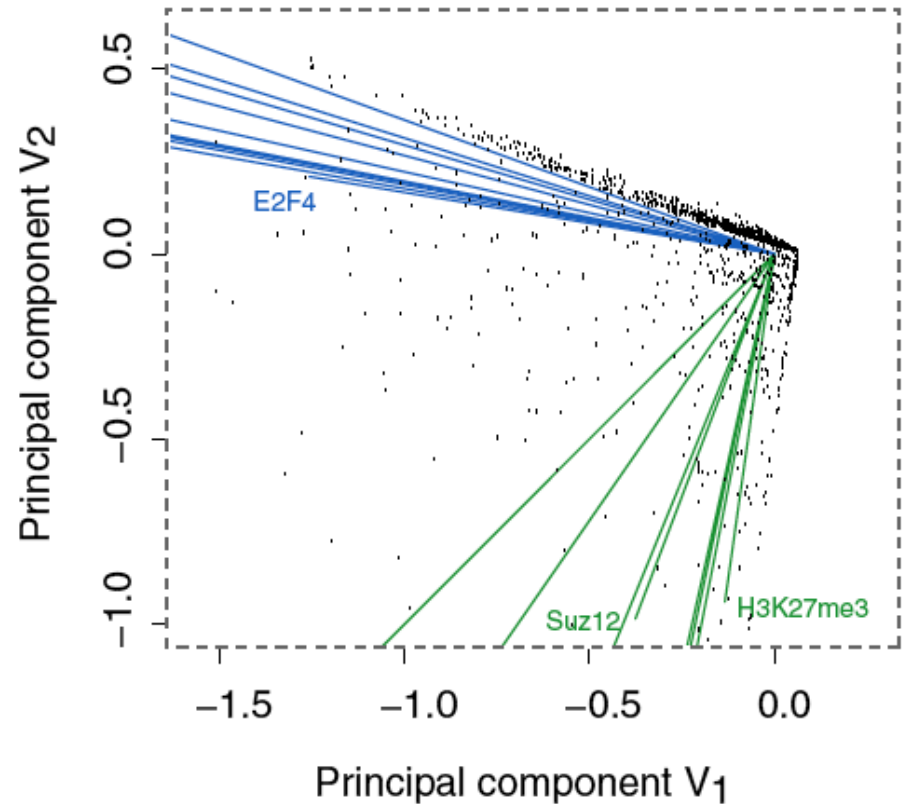
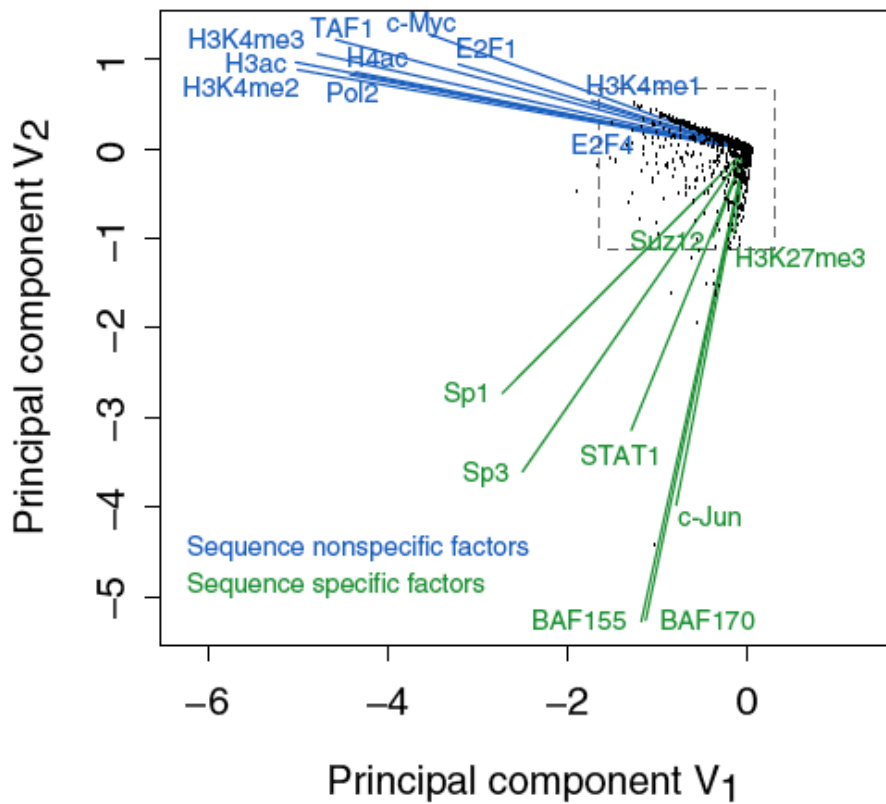
A^T

	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

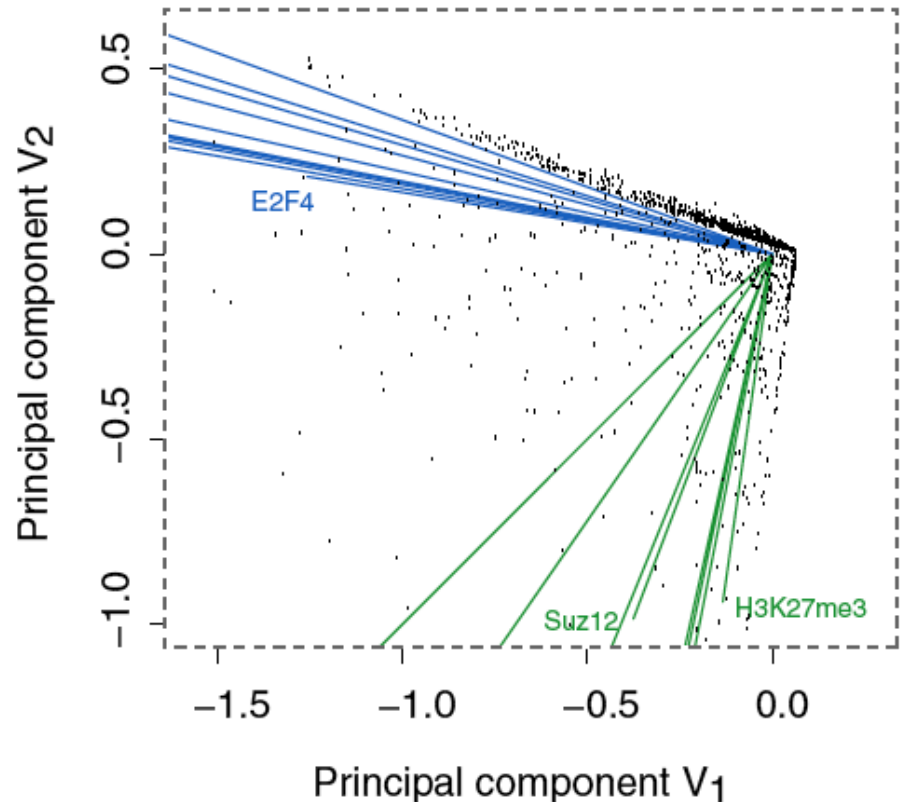
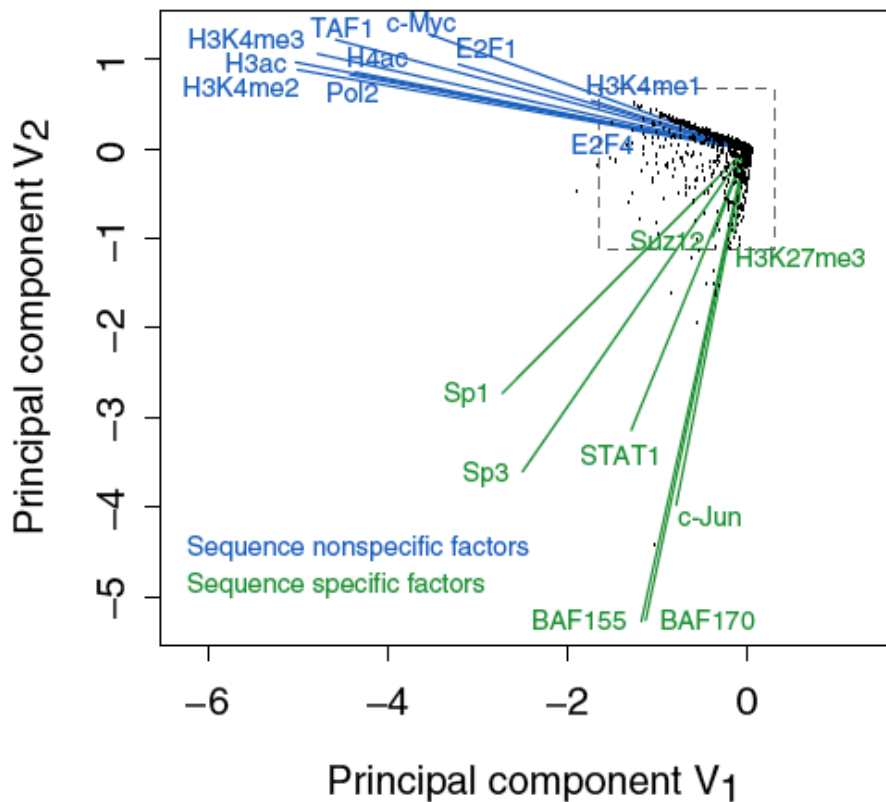
$A A^T$ (site-site correlation)





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



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

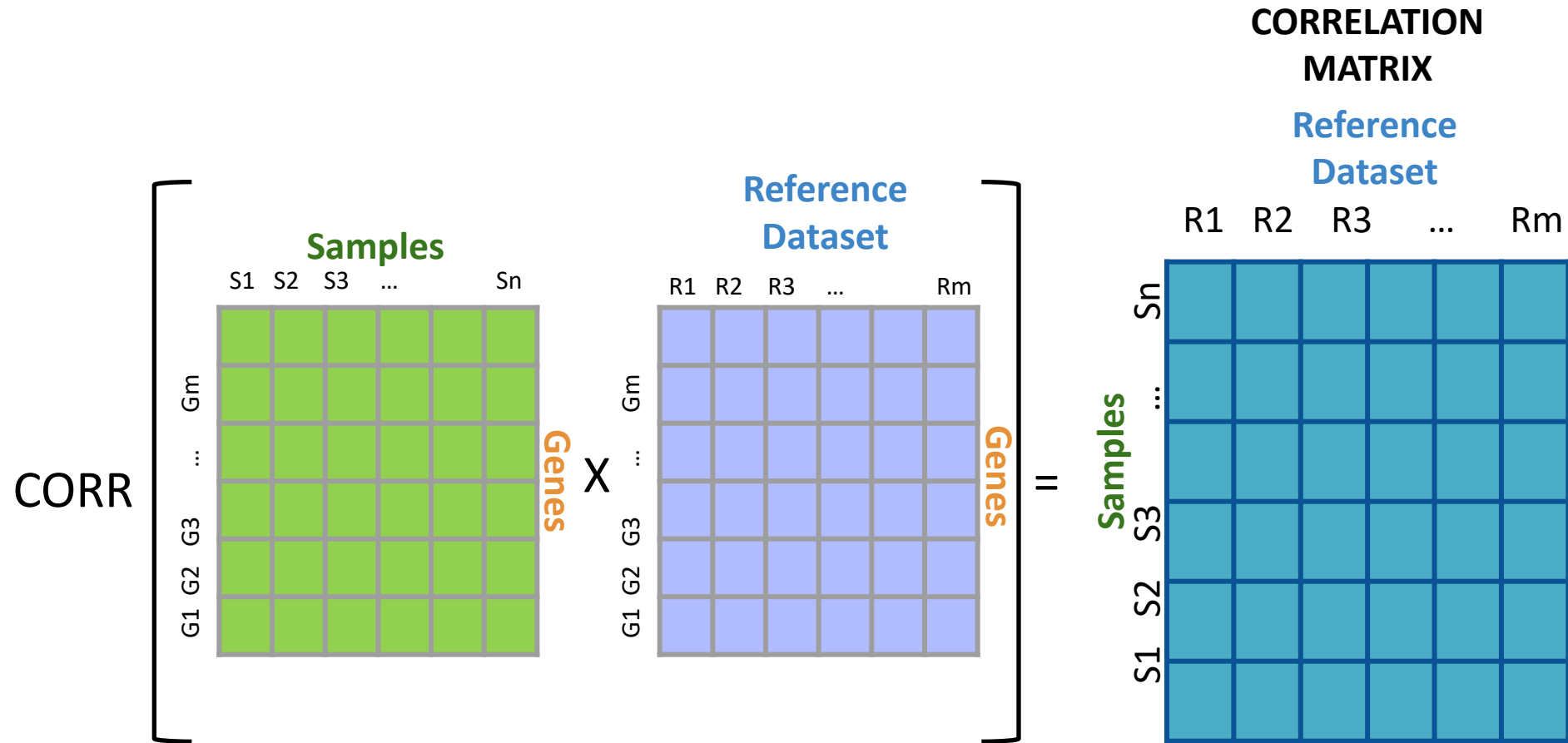
Unsupervised Mining

RCA

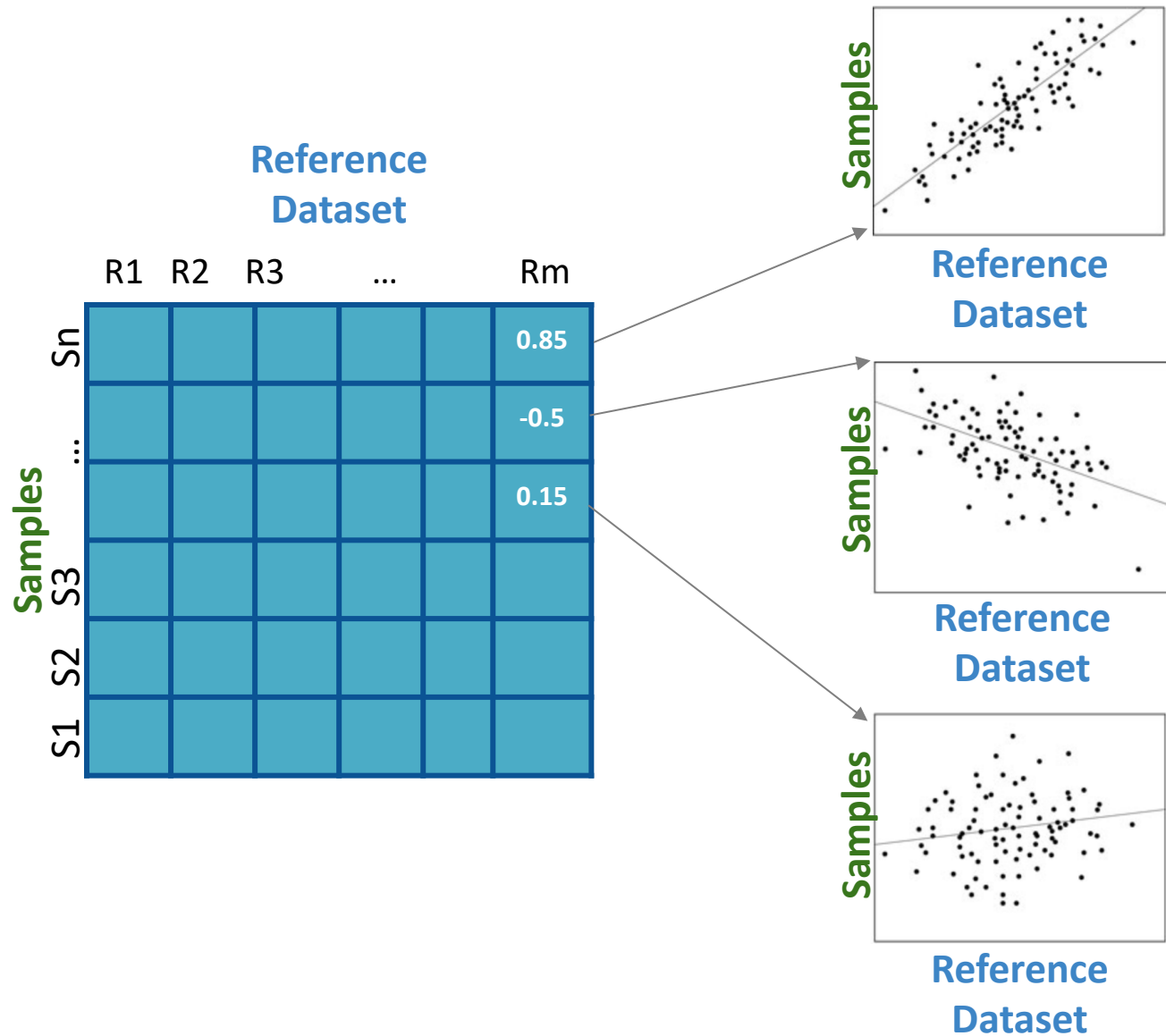
What is RCA?

- RCA stands for **Reference** Component Analysis
- RCA is an algorithm that expands the standard PCA to address noisy data:
 - Batch effect
 - Low signal to noise datasets
- It is still an unsupervised clustering method but, RCA adds external information to address noisy data:
 - Instead of projecting the original data into new axis
 - It first correlates the original data to a reference panel
 - And then, performs PCA on the correlations
- In single-cell or bulk RNA-seq

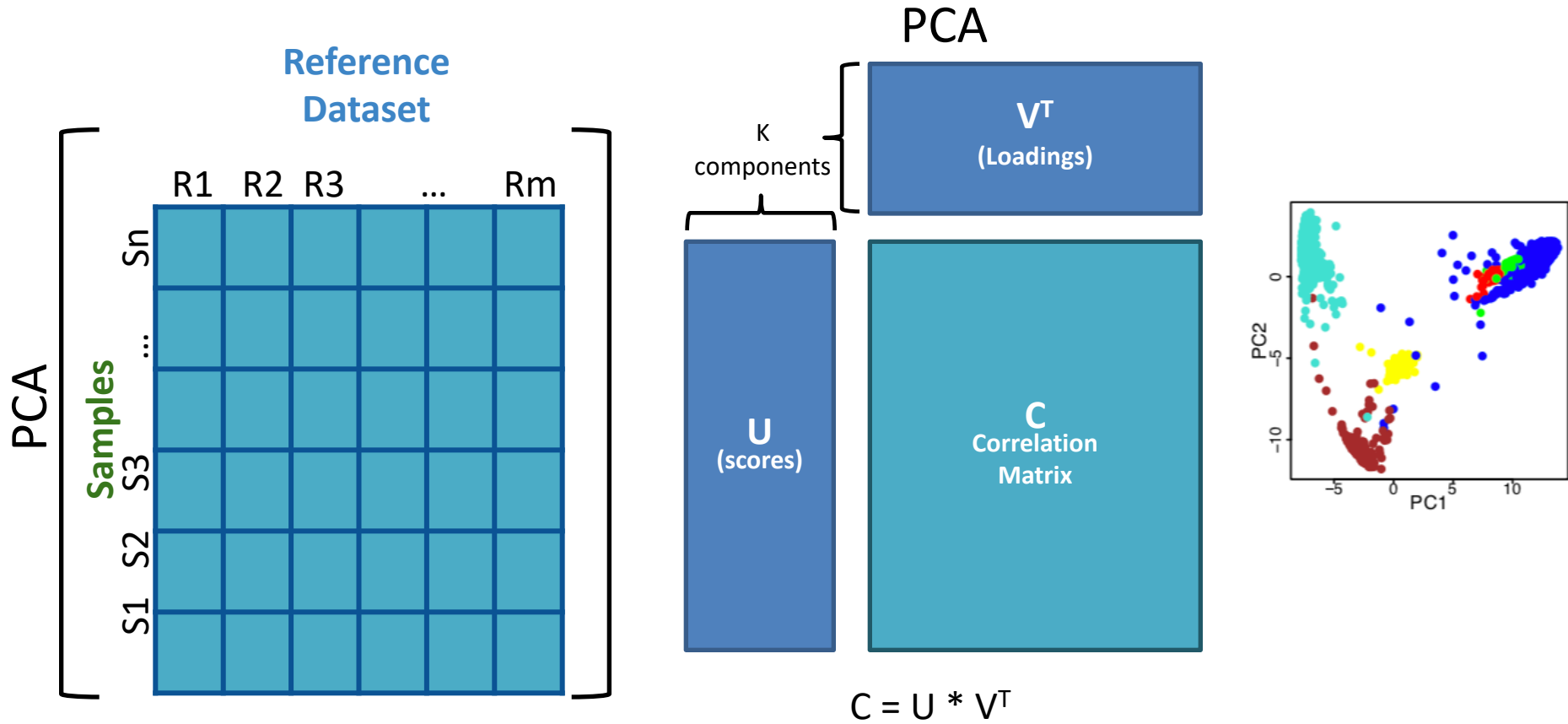
Projection to external dataset



Correlation matrix



PCA on correlation matrix



Unsupervised Mining

CCA

Sorcerer II Global Ocean Survey

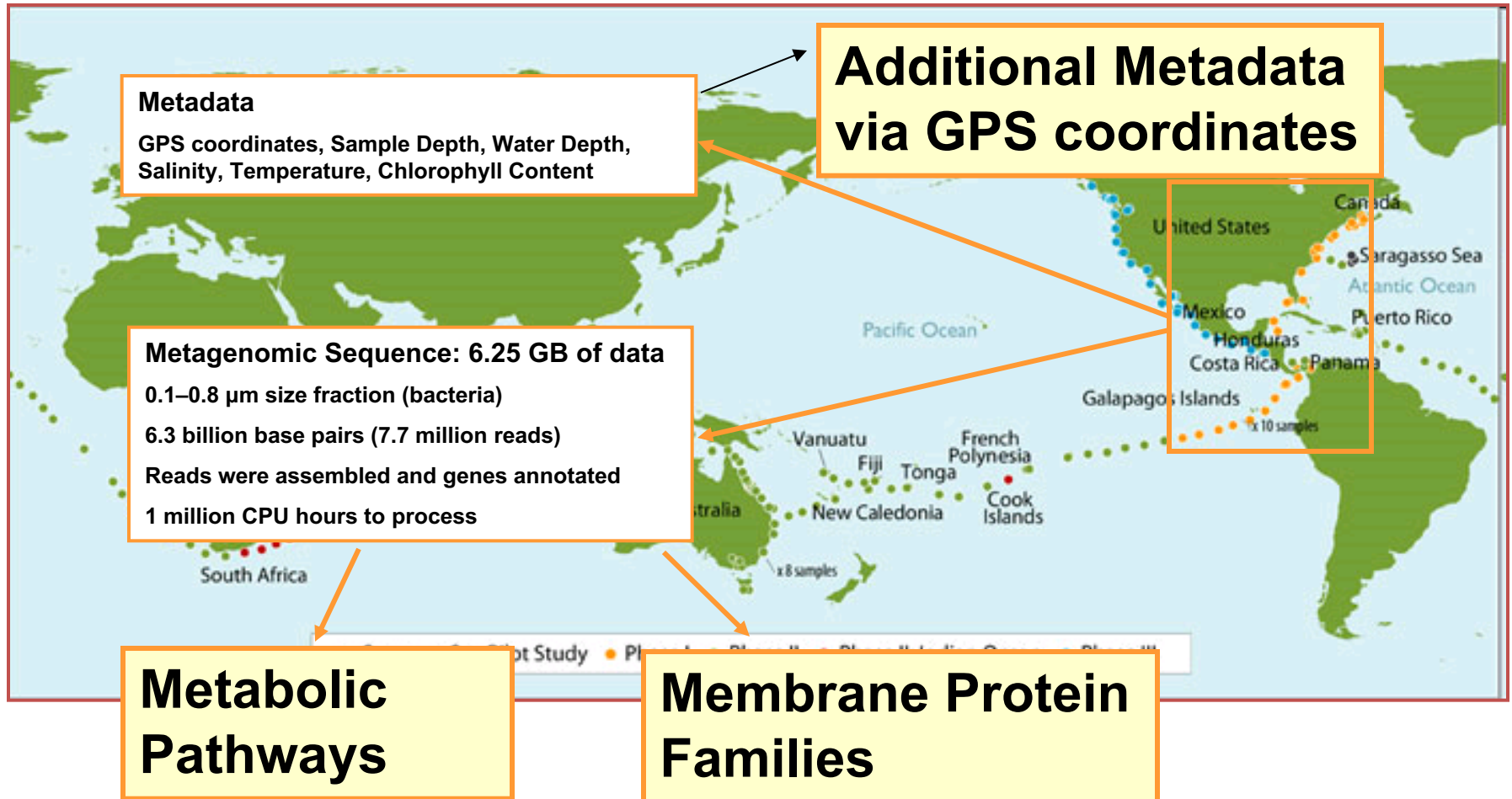


Sorcerer II journey August 2003- January 2006

Sample approximately every 200 miles



Sorcerer II Global Ocean Survey

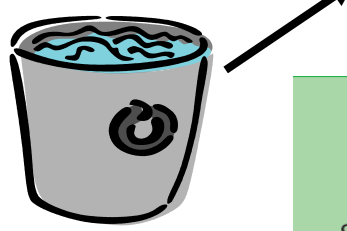


Pathway Sequences (Community Function)

Metabolic Pathways

Sites

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



Environmental Features

Environmental

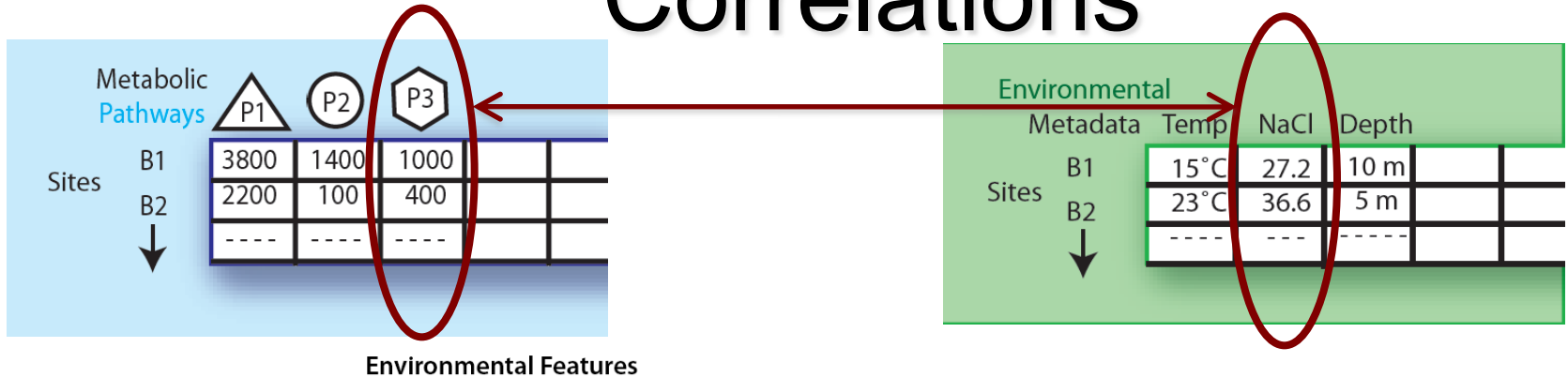
Metadata

Sites

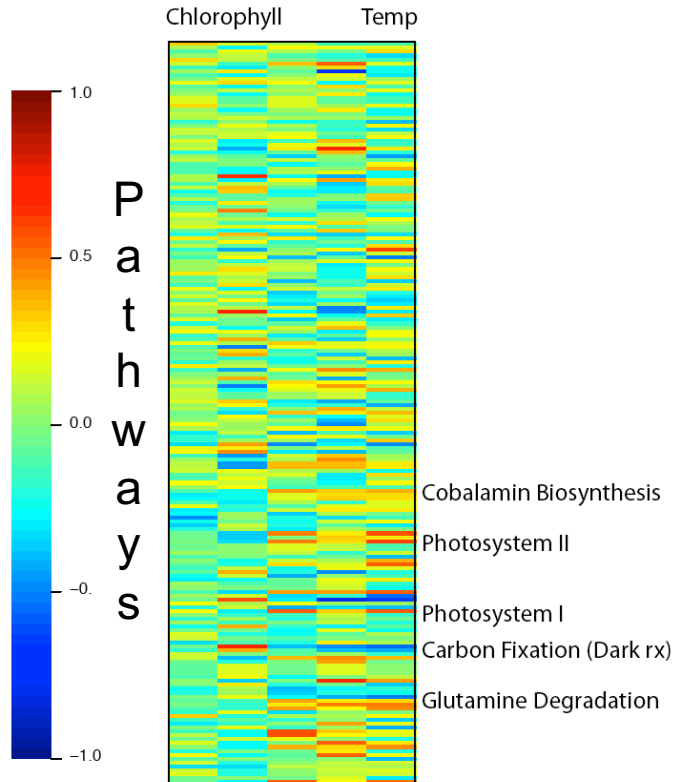
	Temp	NaCl	Depth		
B1	15°C	27.2	10 m		
B2	23°C	36.6	5 m		
↓	----	----	----		

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

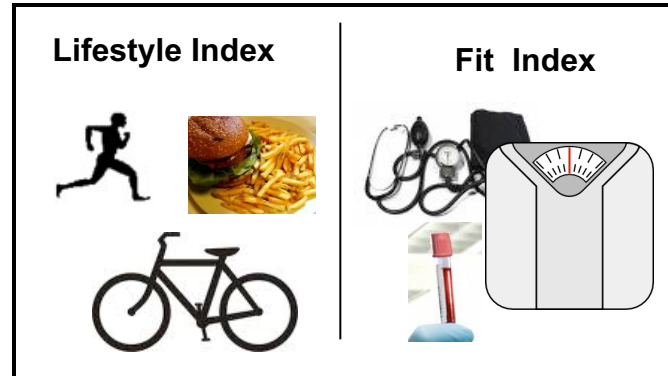
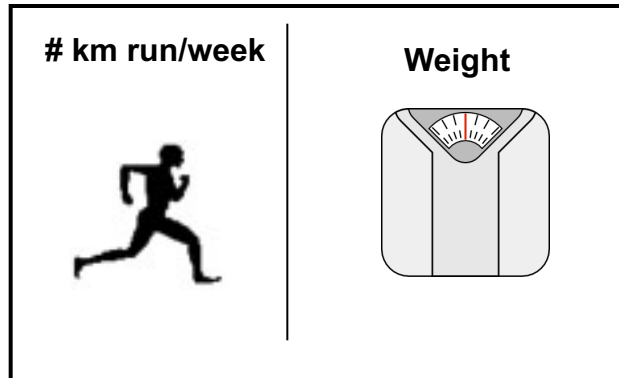
Simple Relationships: Pairwise Correlations



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



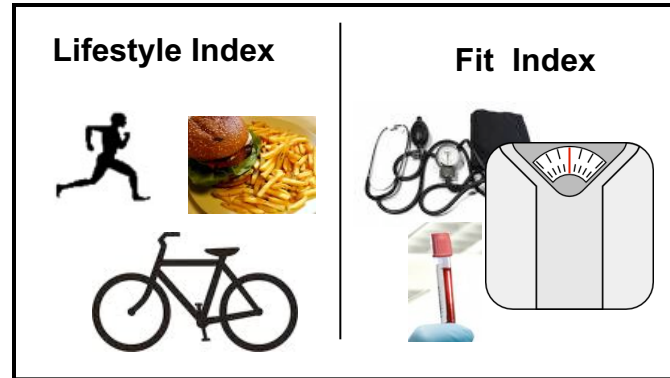
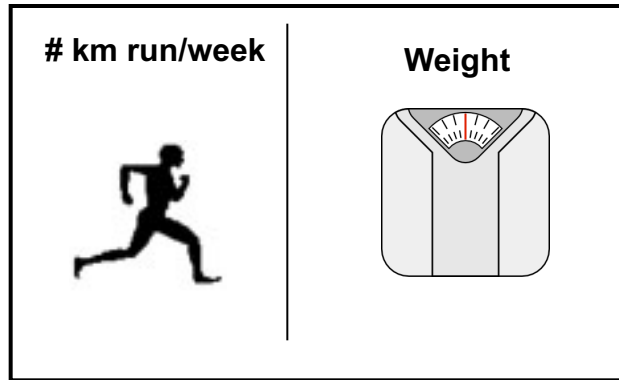
Canonical Correlation Analysis: Simultaneous weighting



$$\text{Lifestyle Index} = a \text{ } \img alt="Silhouette of a person running" data-bbox="388 643 438 708" \text{ } + b \text{ } \img alt="A burger and fries" data-bbox="536 631 628 726" \text{ } + c \text{ } \img alt="A bicycle" data-bbox="701 636 808 721" \text{ }$$

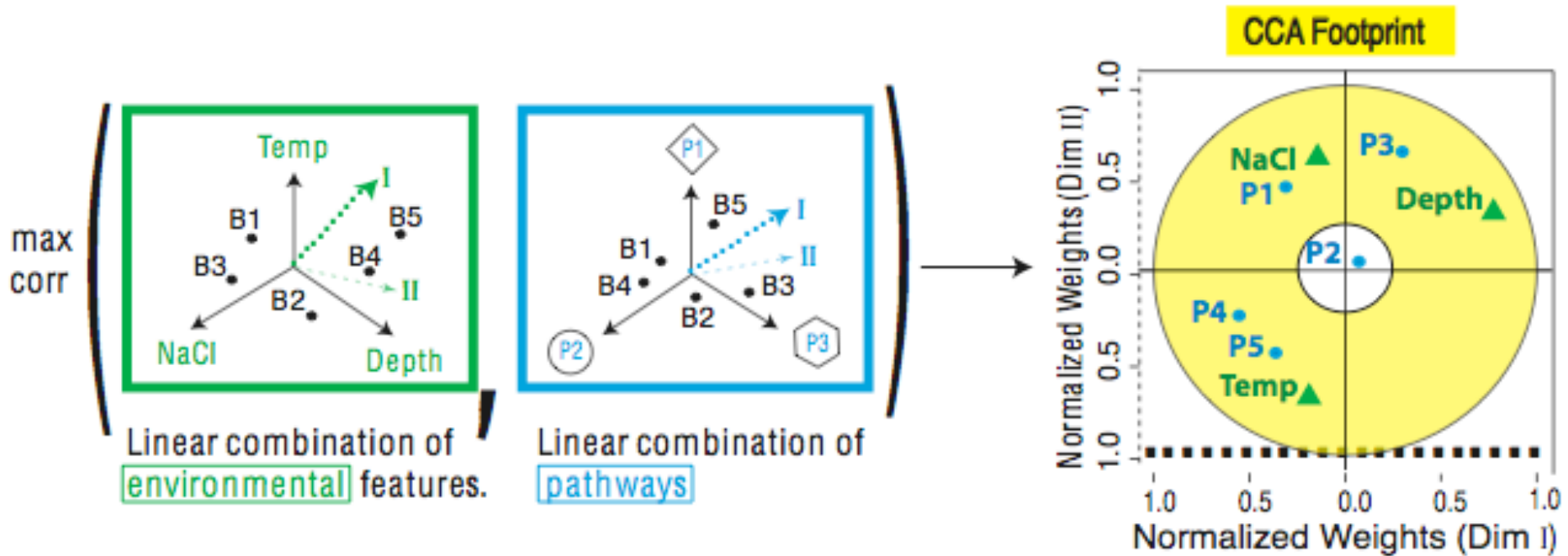
$$\text{Fit Index} = a \text{ } \img alt="A stethoscope and a blood pressure cuff" data-bbox="301 801 418 901" \text{ } + b \text{ } \img alt="A blood pressure monitor" data-bbox="503 776 558 891" \text{ } + c \text{ } \img alt="A scale with a dial" data-bbox="653 801 733 908" \text{ }$$

Canonical Correlation Analysis: Simultaneous weighting

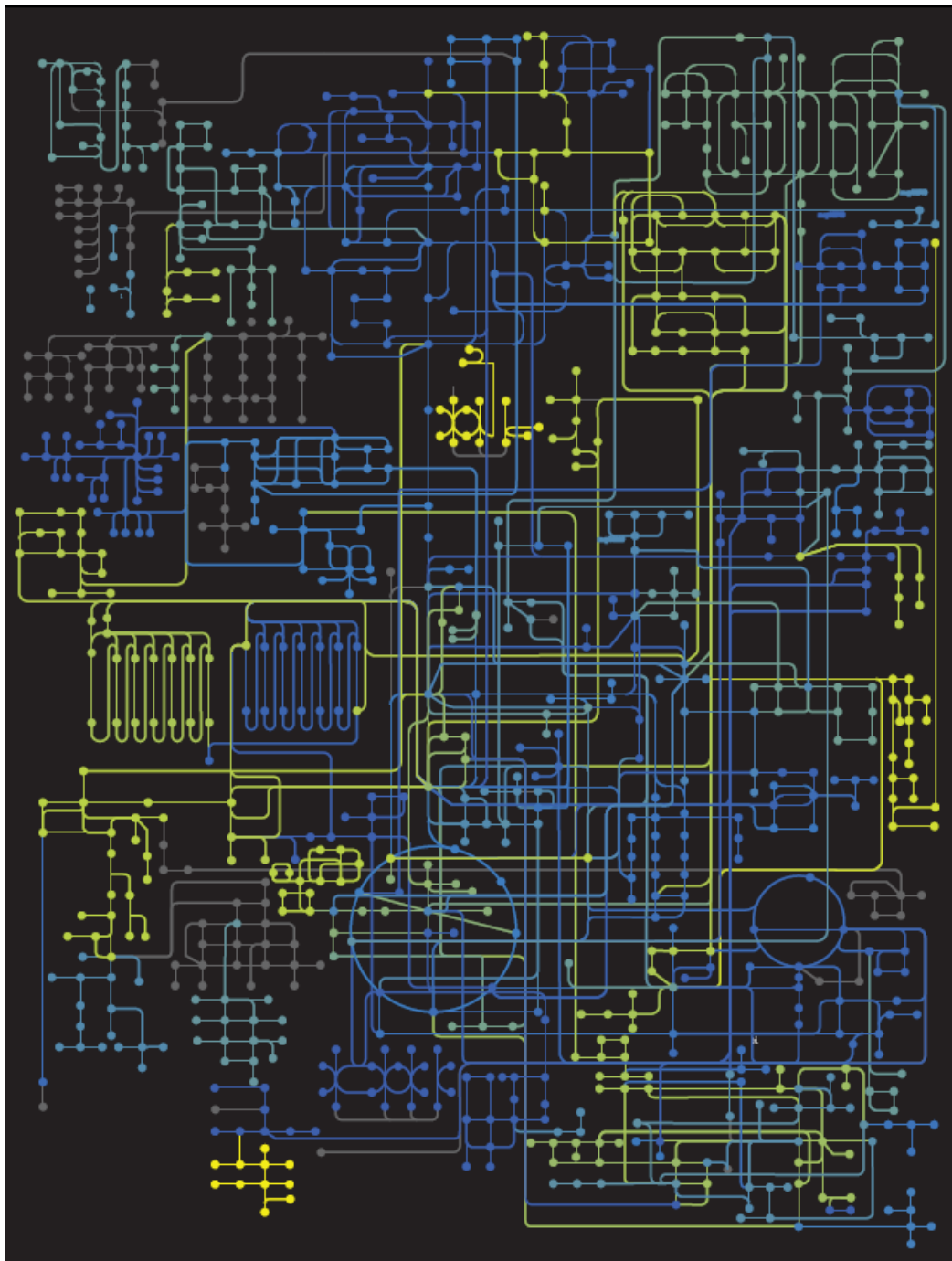


Life	<p>Environmental Features</p> <p>Temp etc</p>	<p>Metabolic Pathways/ Protein Families</p> <p>Photosynthesis etc</p>
Fit	<p>Chlorophyll</p>	<p>Lipid Metabolism</p>

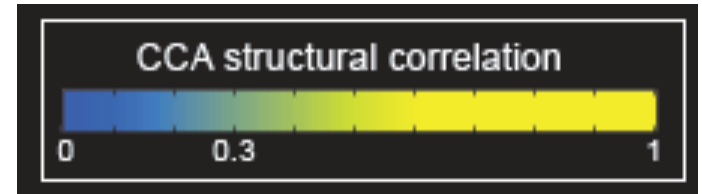
CCA: Finding Variables with Large Projections in "Correlation Circle"



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

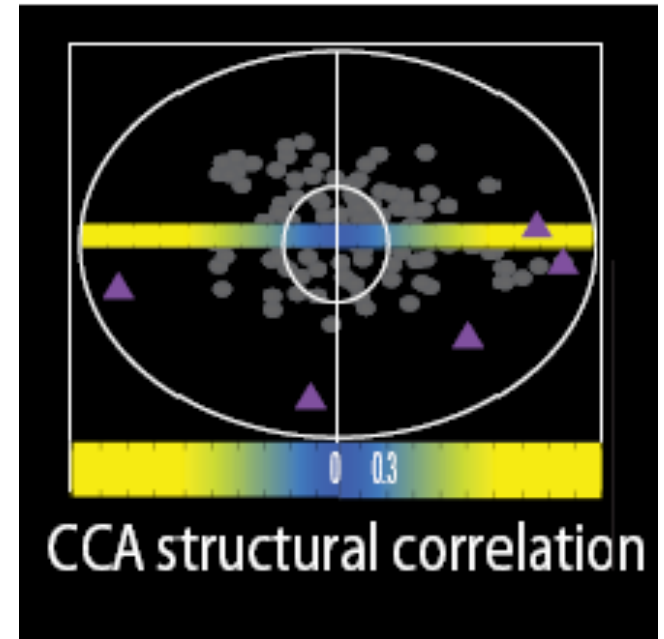


Strength of Pathway co-variation with environment

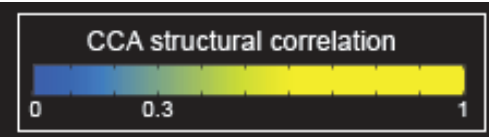


Environmentally
invariant

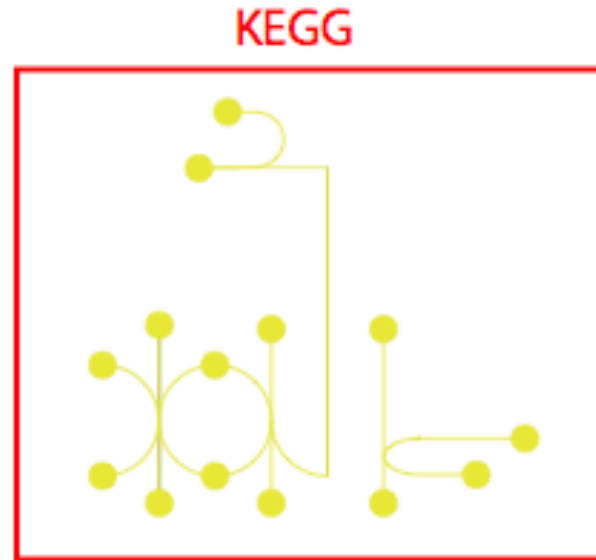
Environmentally
variant



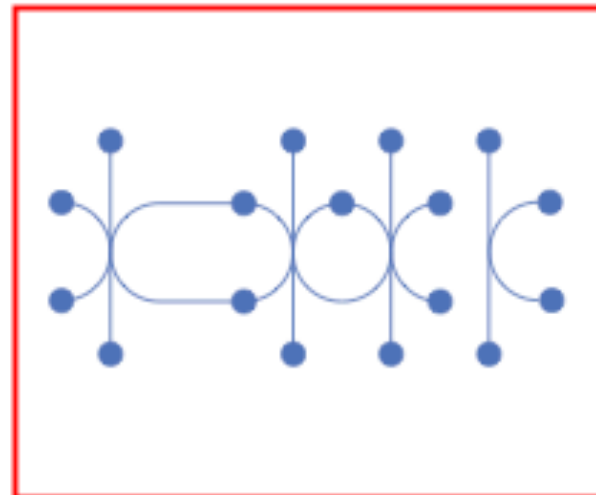
Conclusion #1: energy conversion strategy, temp and depth



Photosynthesis



Oxidative
Phosphorylation

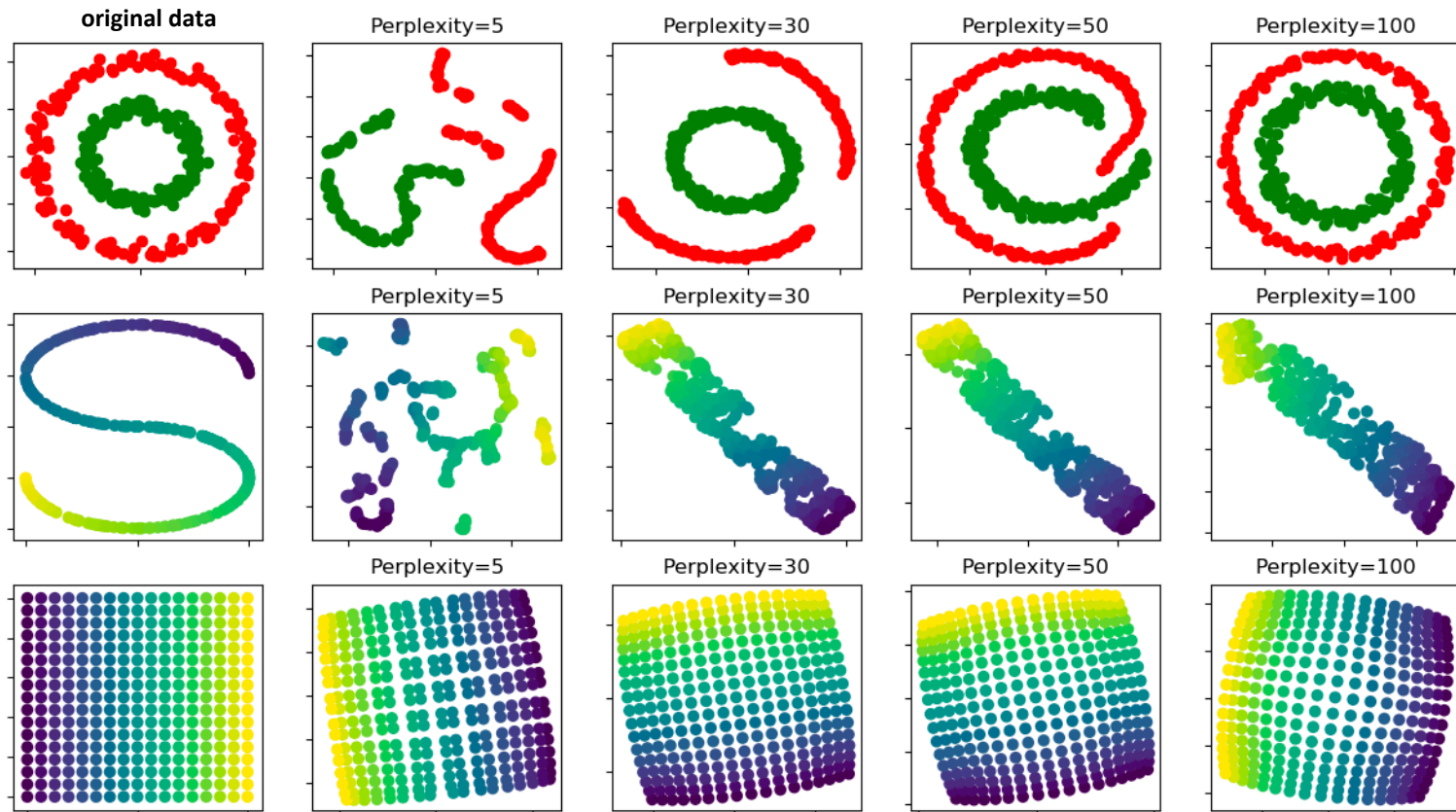


Unsupervised Mining

tSNE

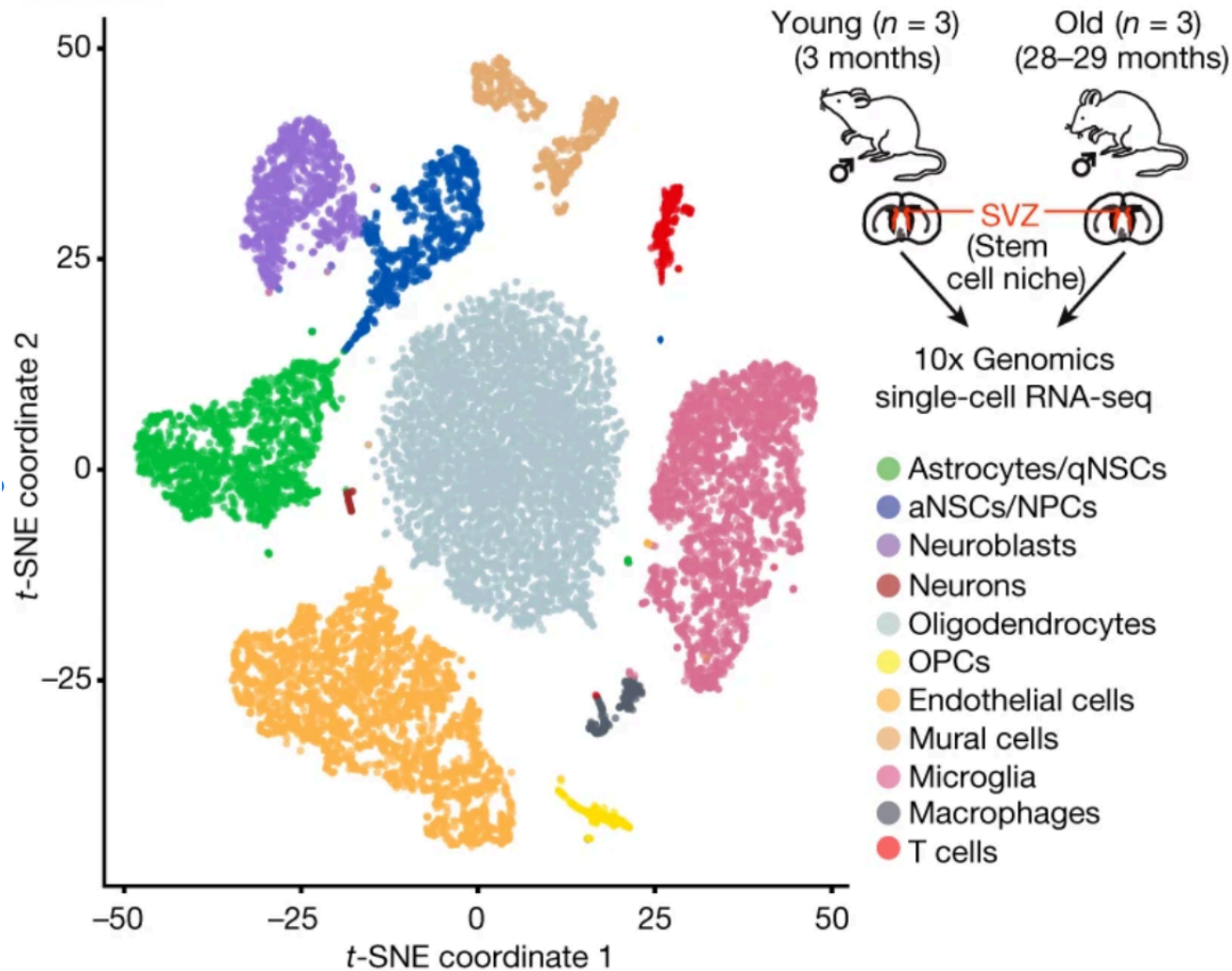
tSNE

a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets



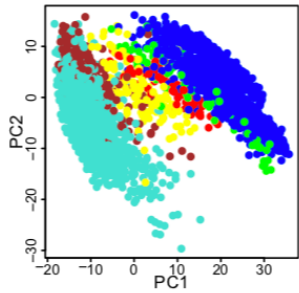
hyperparameters ‘perplexity’ really matter; Cluster sizes in a t-SNE plot mean nothing; Distances between clusters might not mean anything

Example: t-SNE clustering of **14,685** single-cell transcriptomes

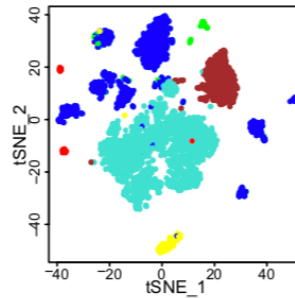


Comparison on real datasets (melanoma scRNA-seq dataset)

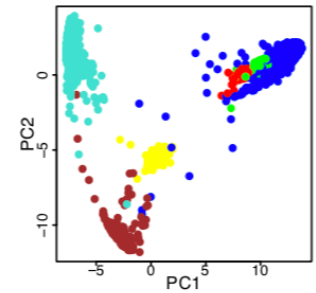
PCA



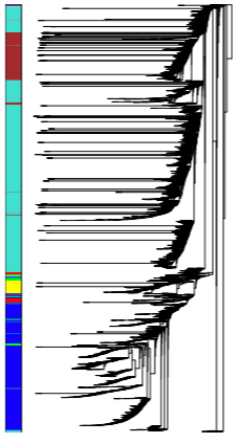
Seurat



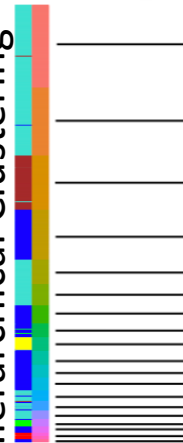
RCA



Hierarchical Clustering



Hierarchical Clustering

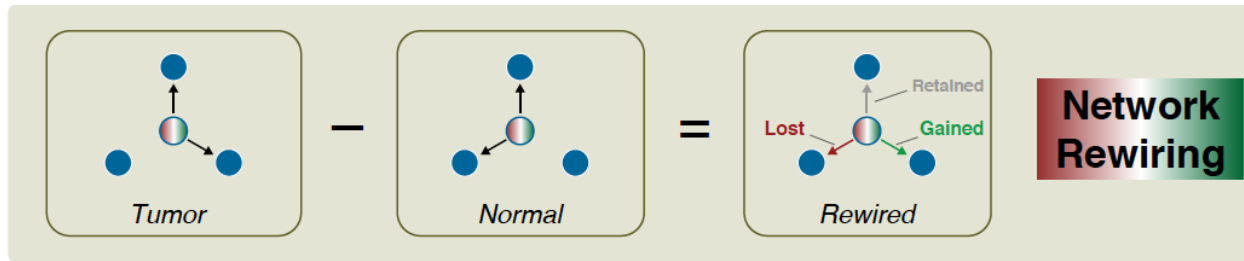


Hierarchical Clustering

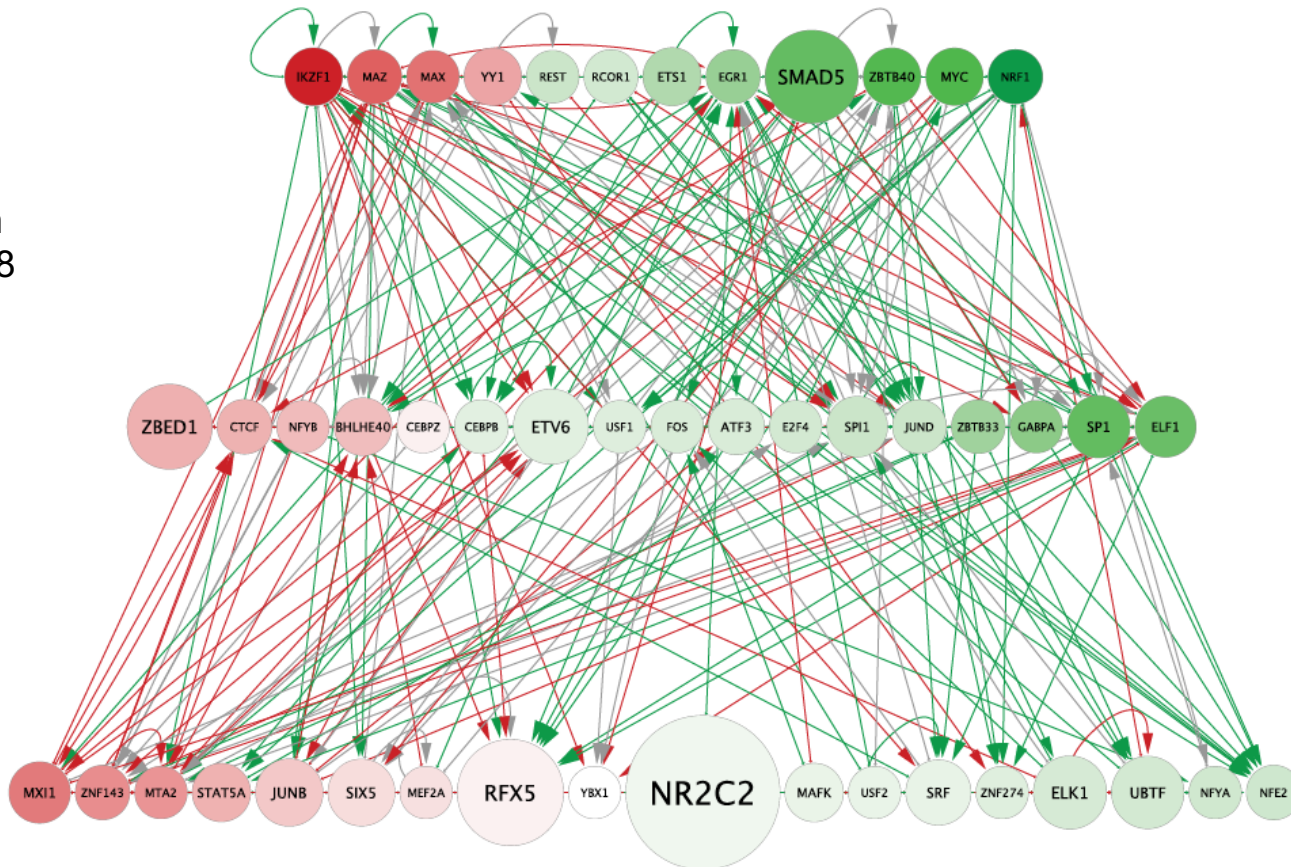


Unsupervised Mining

LDA



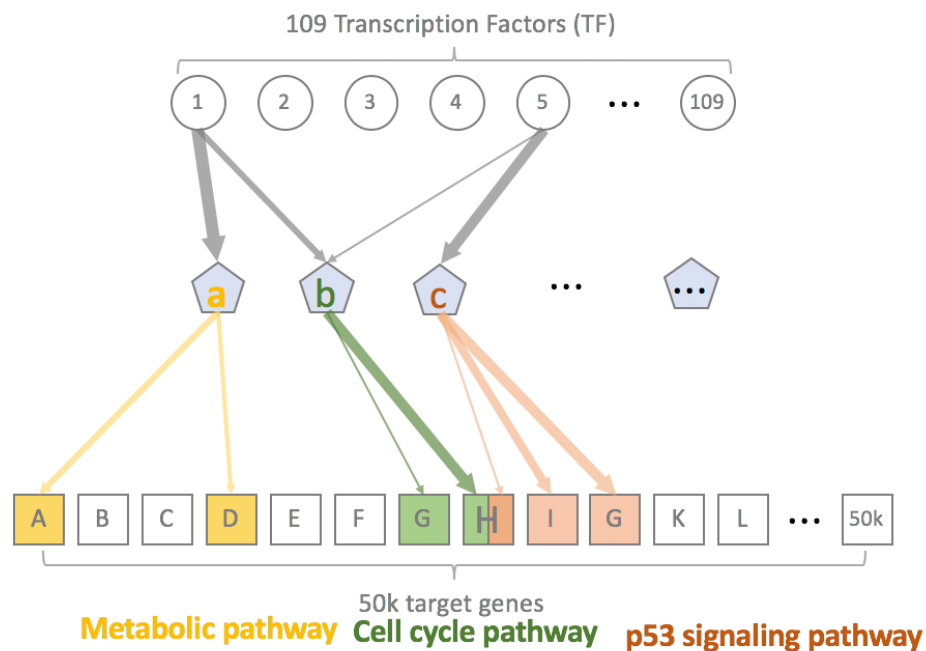
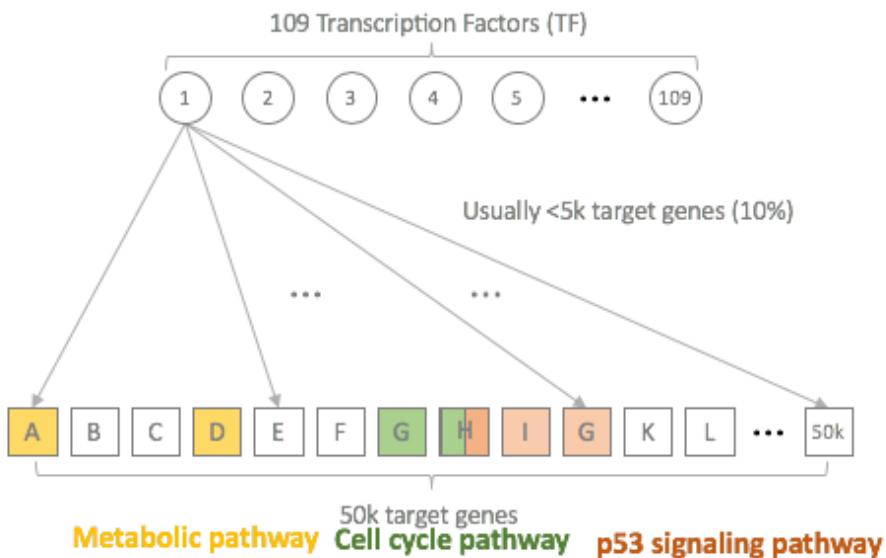
Rewired edges in comparison of GM12878 to K562 109 node TF-TF network (approx. CML)



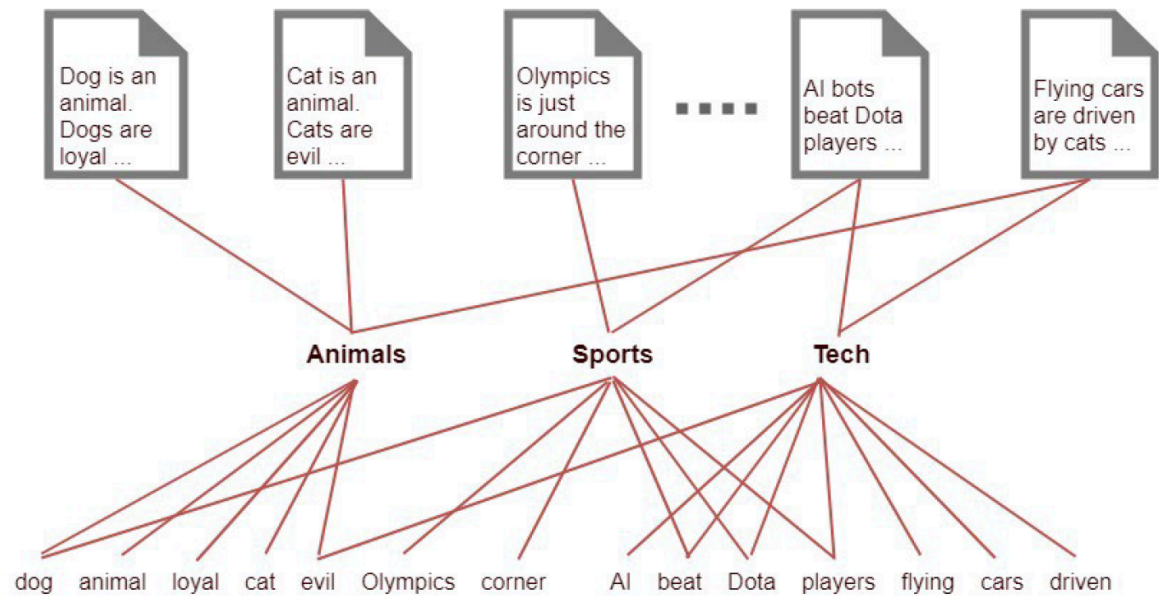
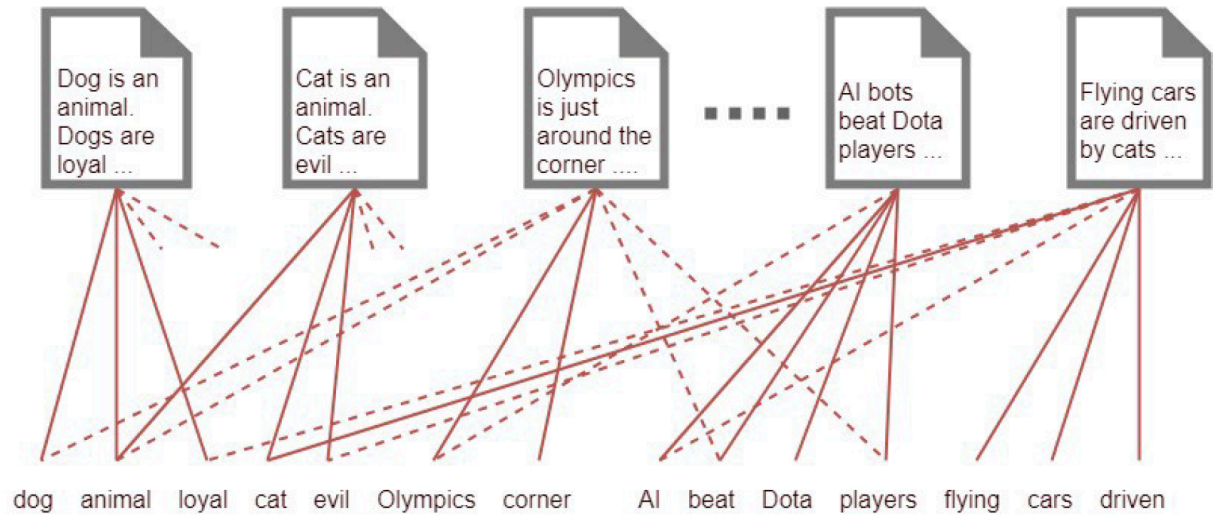
From *TF* →
gene (109×50,000)
to *TF* →
pathway (109×50)

Hidden Layer
(50 biological
pathways?)

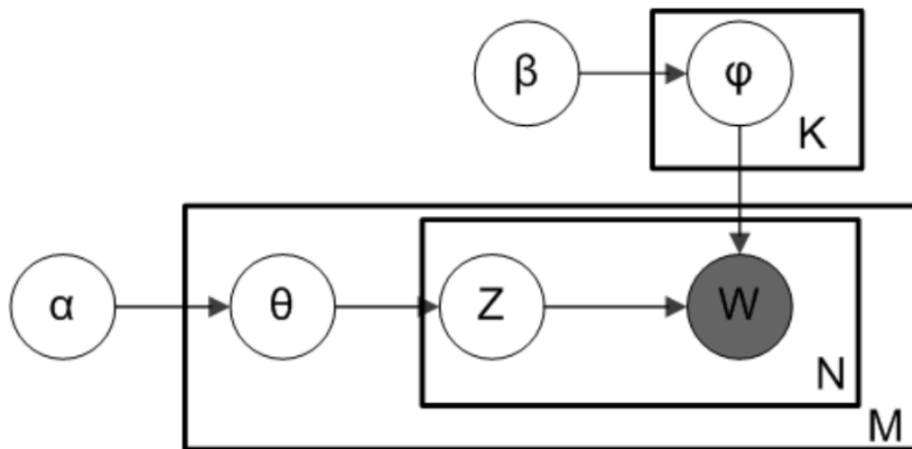
(in press)]



From
dimension
reduction
view



Diagram



$$\begin{aligned}\varphi_{k=1\dots K} &\sim \text{Dirichlet}_V(\beta) \\ \theta_{d=1\dots M} &\sim \text{Dirichlet}_K(\alpha) \\ z_{d=1\dots M, w=1\dots N_d} &\sim \text{Categorical}_K(\theta_d) \\ w_{d=1\dots M, w=1\dots N_d} &\sim \text{Categorical}_V(\varphi_{z_{dw}})\end{aligned}$$

α is the parameter of the Dirichlet prior on the per-document topic distributions

β is the parameter of the Dirichlet prior on the per-topic word distribution

θ_i is the topic distribution for document i

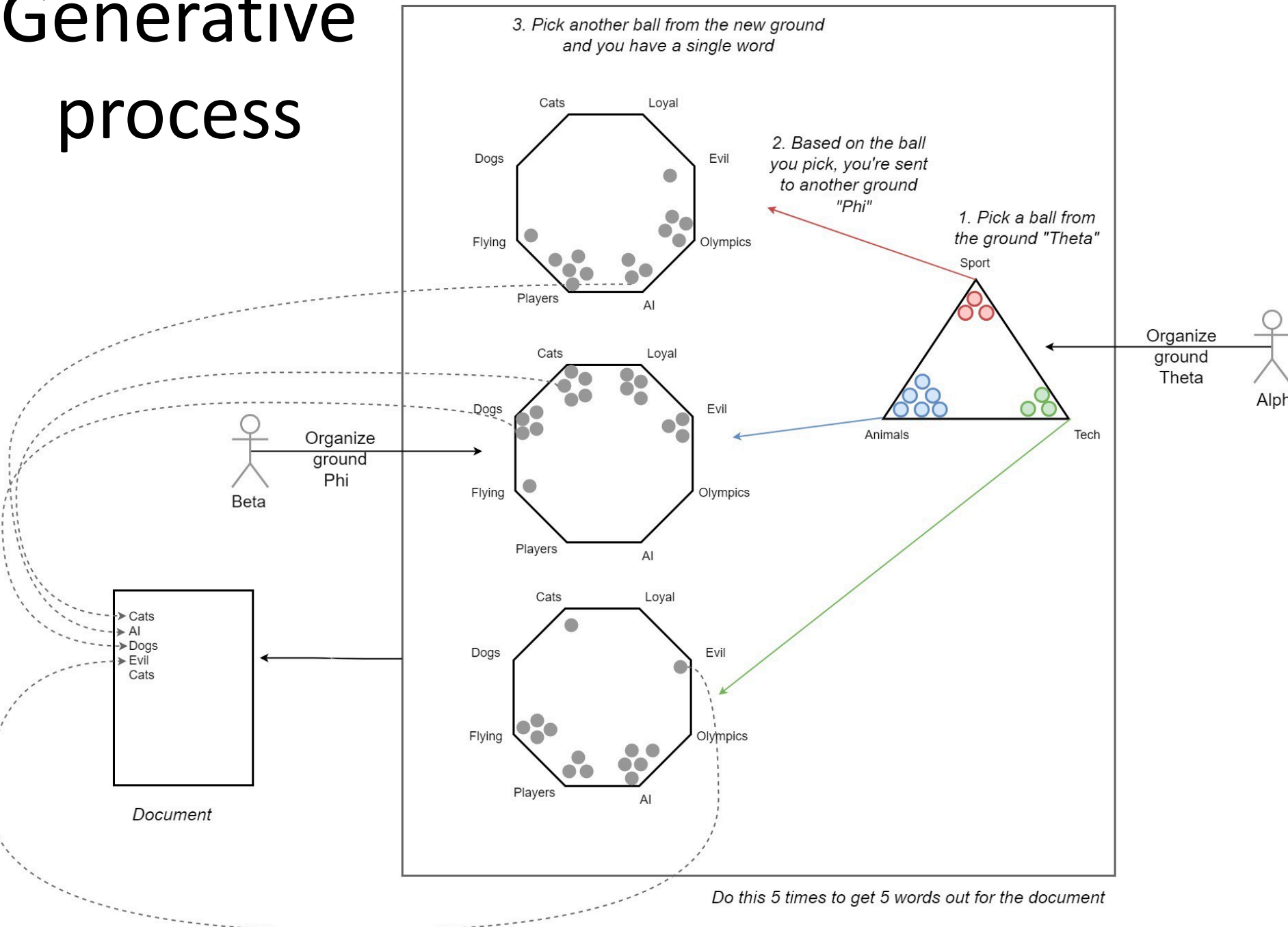
φ_k is the word distribution for topic k

z_{ij} is the topic for the j th word in document i

w_{ij} is the specific word, and

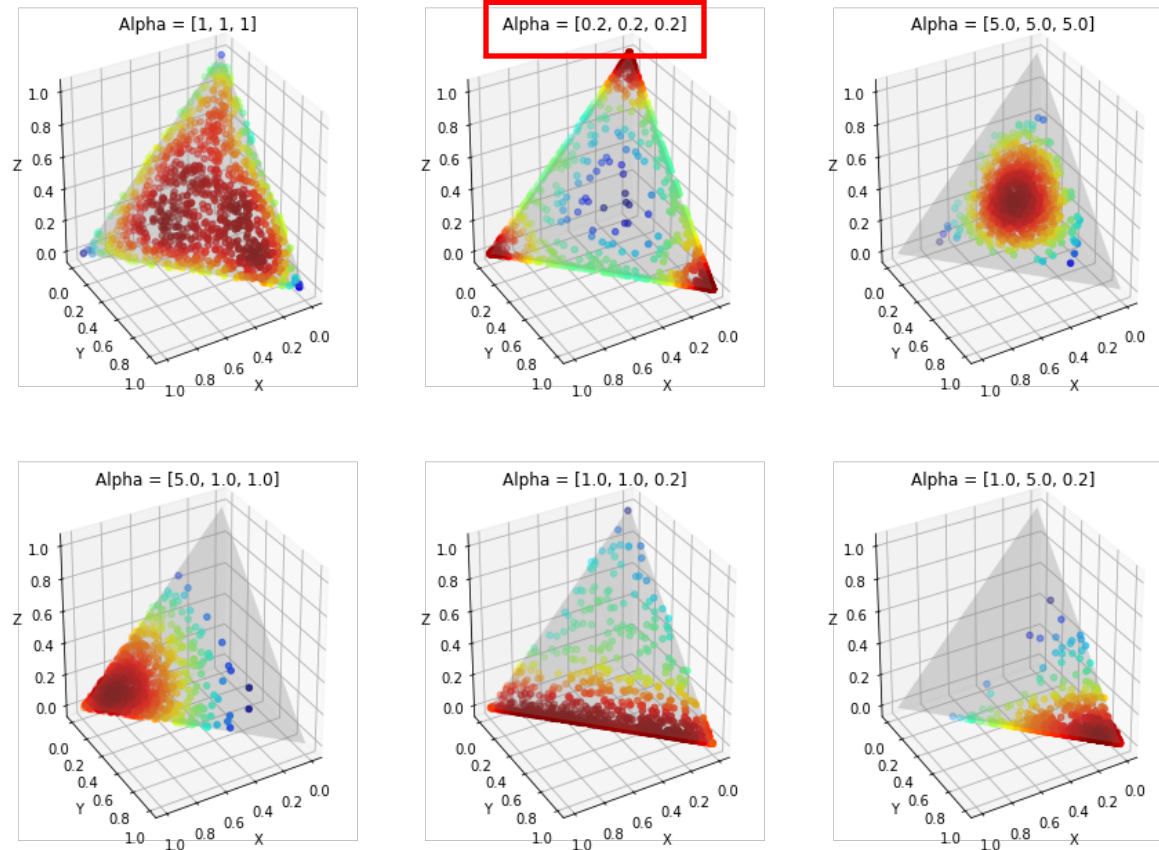
K is the number of topics, N is the number of word in a document, M is the number of Documents.

Generative process



The sparsity is important

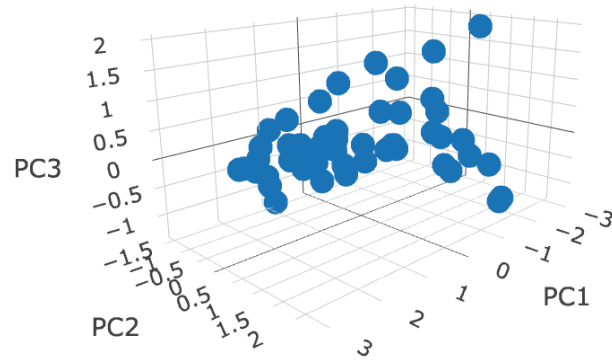
hyperparameter of Dirichlet distribution enable the sparsity of document to topic (θ) and word to topic (ϕ) distribution, make LDA works better than others similar methods most of time.



In LDA analysis, alpha should be tuned for topic distribution(θ)

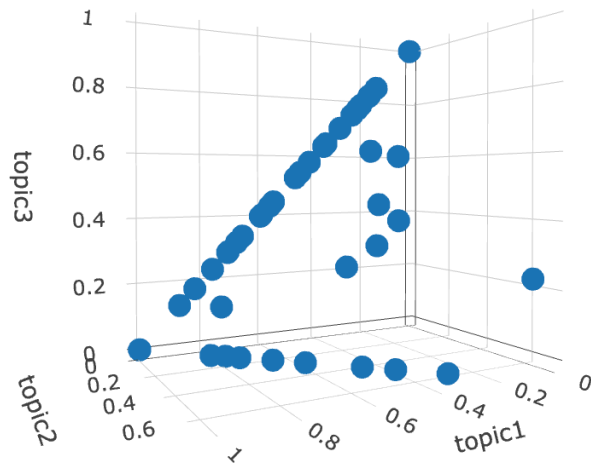
How the distribution of θ changes with different α values

Comparison of sparsity using **USArrests** dataset in a three-dimensional space



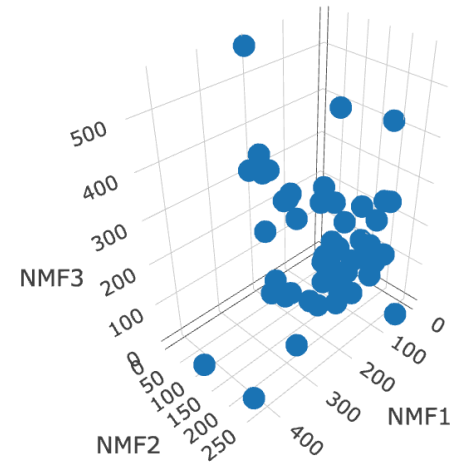
PCA

Rotation= XW (loading)



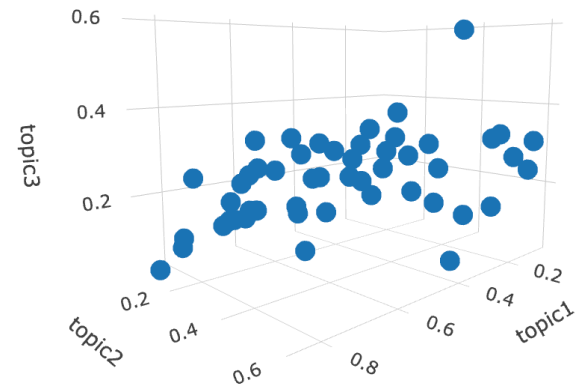
LDA K=3, alpha=0.2

θ



NMF (rank=3) $X=W*H$

W



LDA K=3, alpha=5

θ