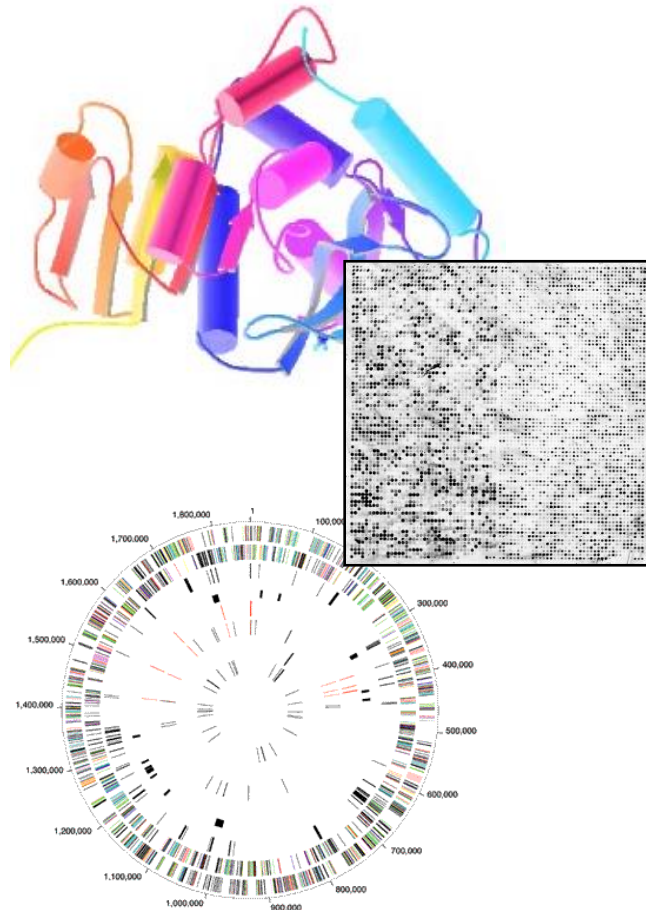


Unsupervised Datamining D: SVD Extensions



Mark Gerstein, Yale University
gersteinlab.org/courses/452
(last edit in spring '21, pack #9d, final)

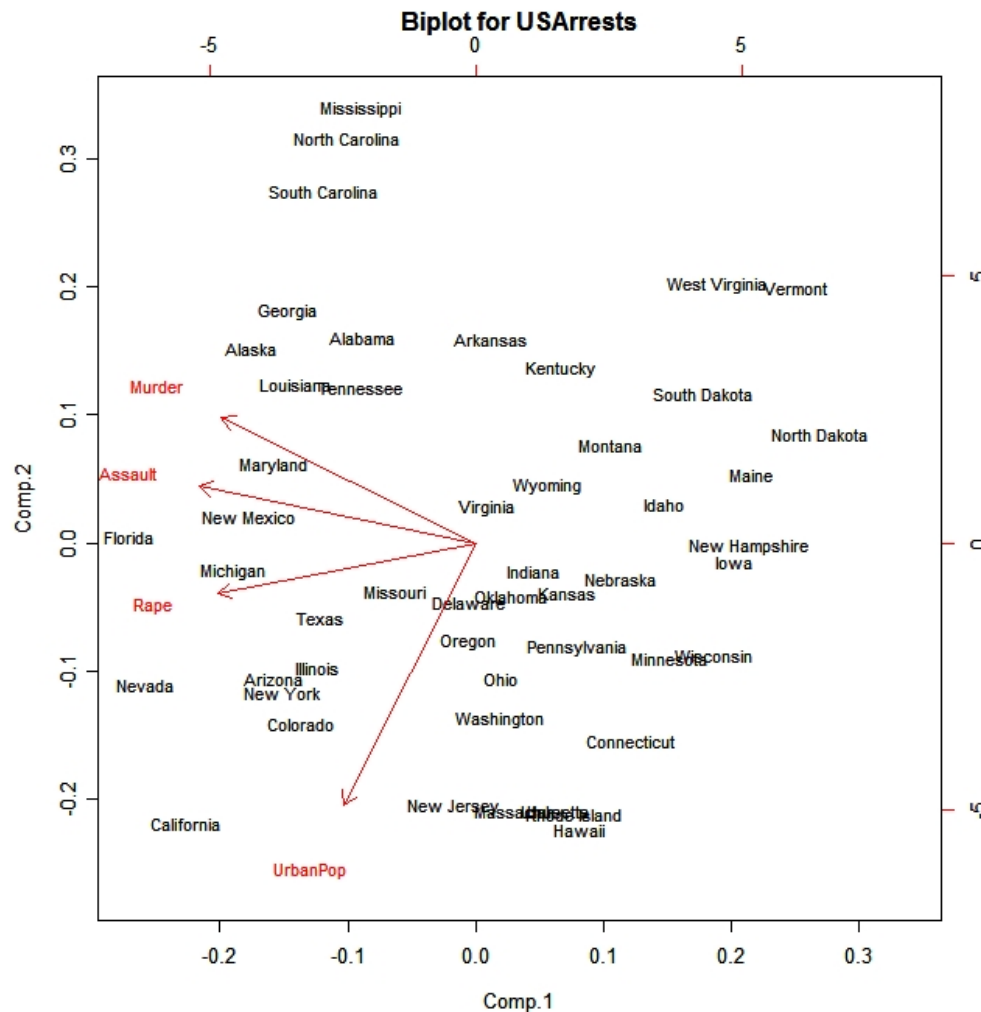
Unsupervised Mining

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

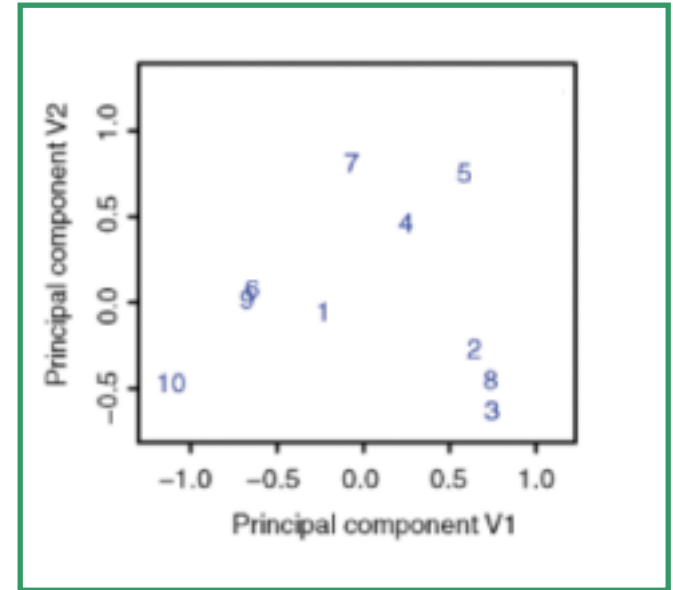
Sites: 1,2,3...

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.)

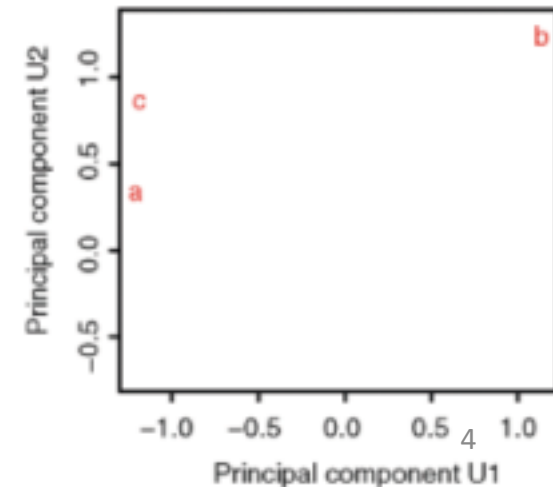


A^T

	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...

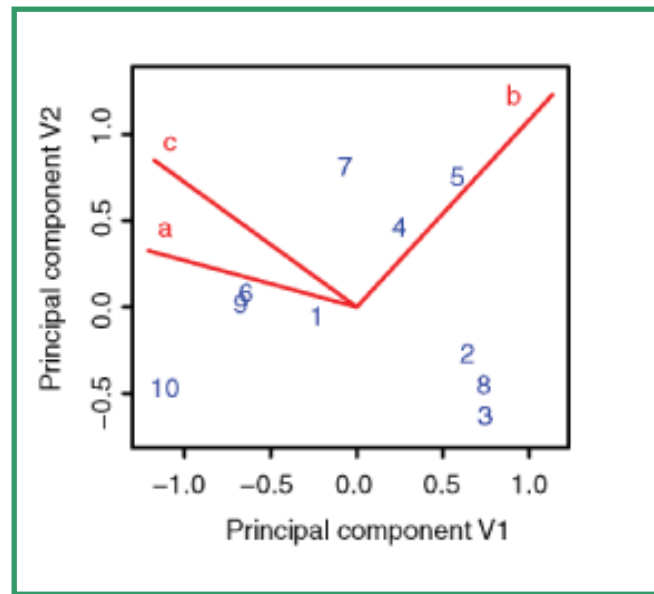
Sites: 1,2,3...

$$A = USV^T$$

	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.)

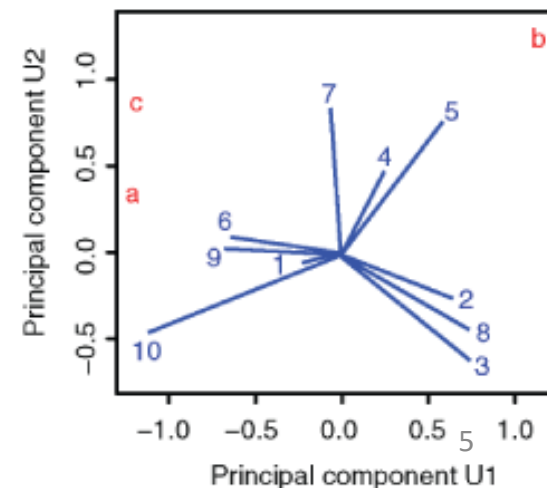


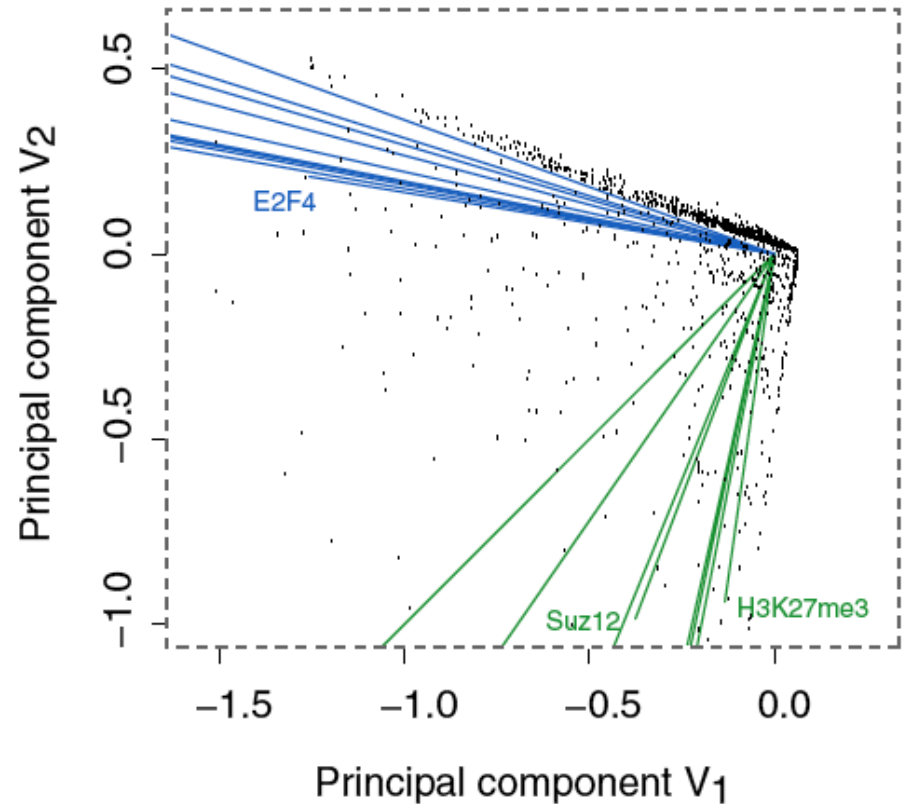
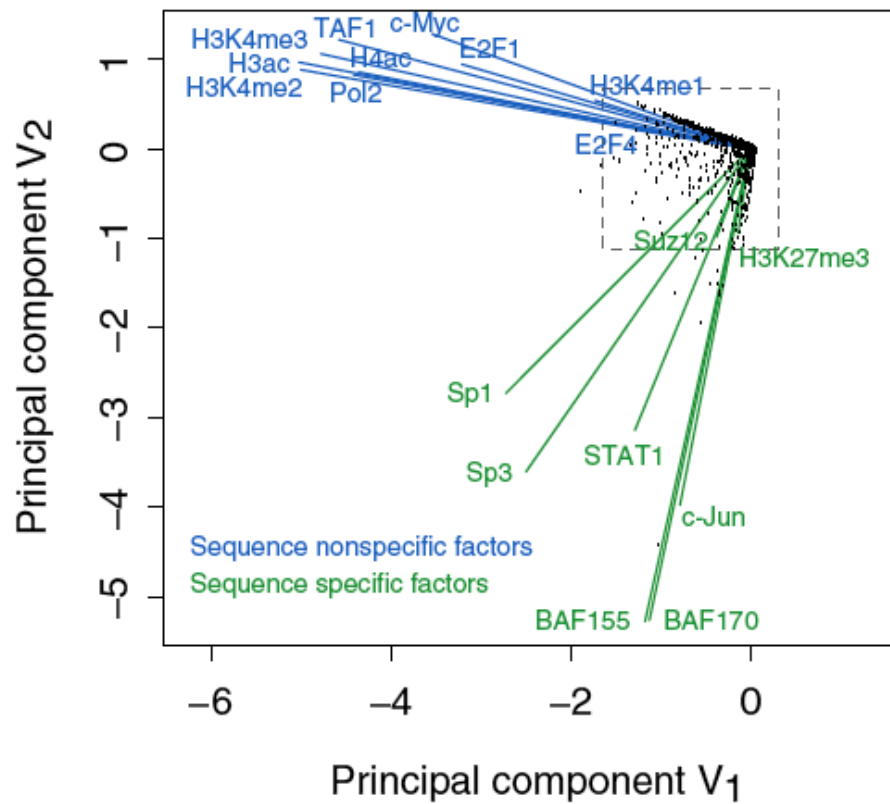
$$A A^T$$

	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)

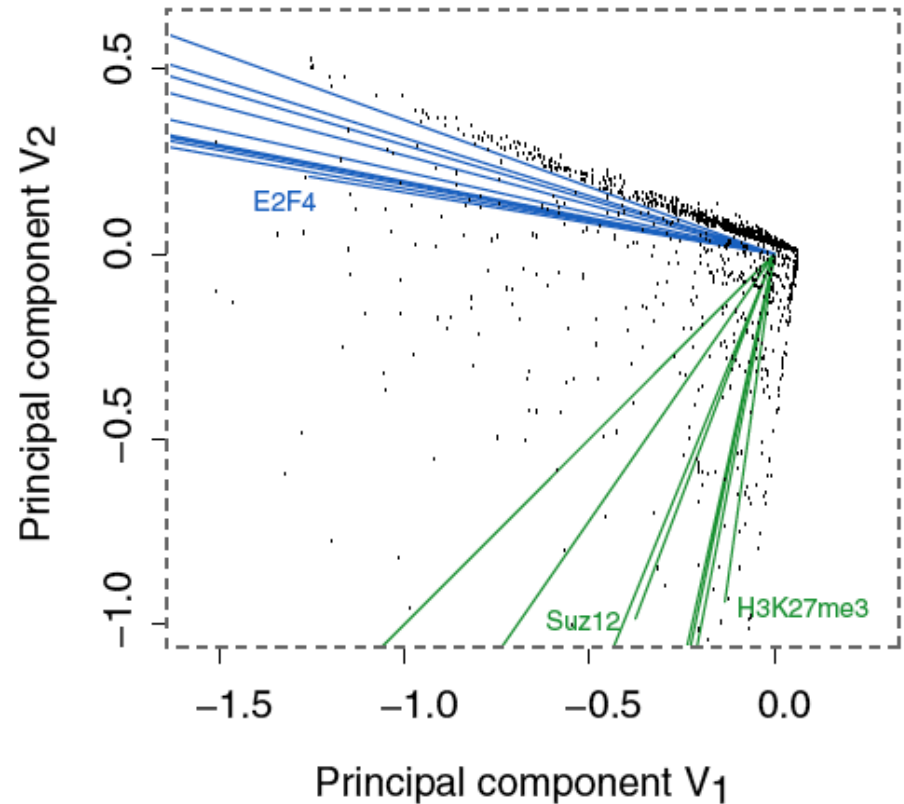
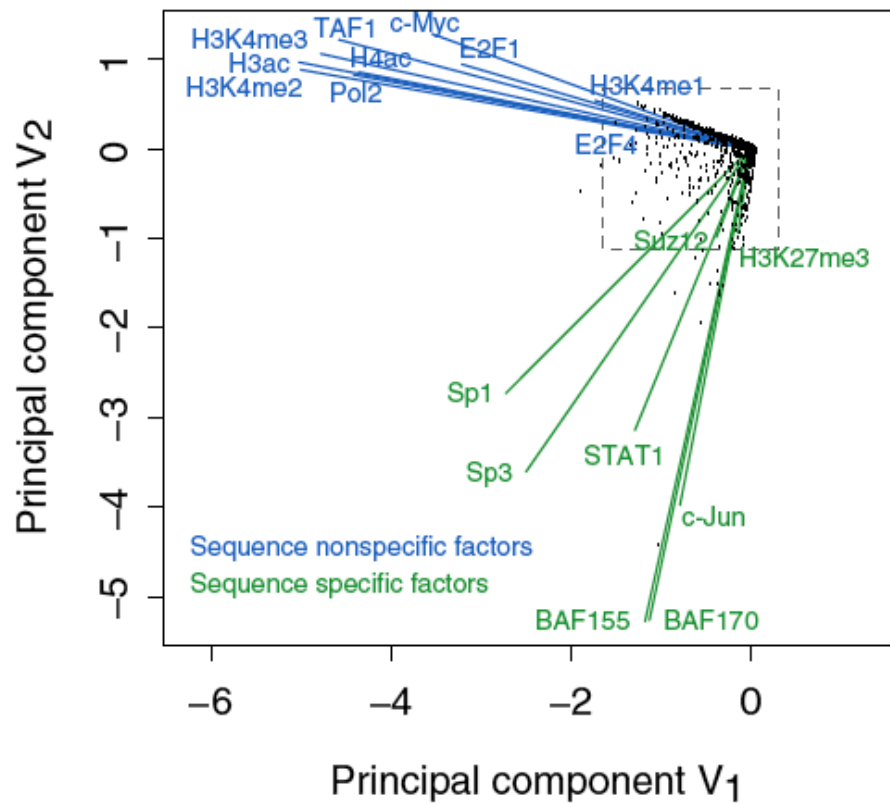




Results of Biplot

- 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

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.

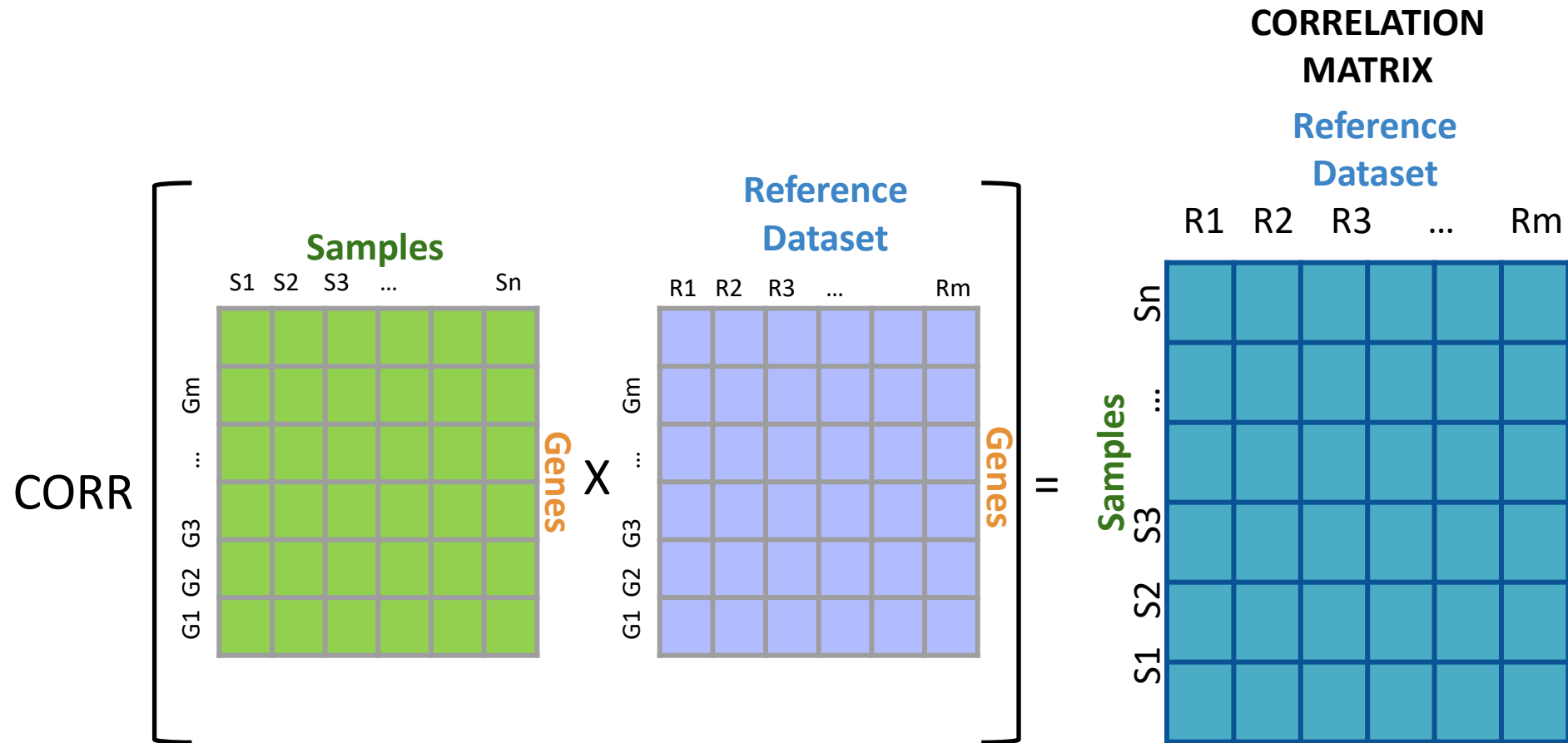
Unsupervised Mining

RCA

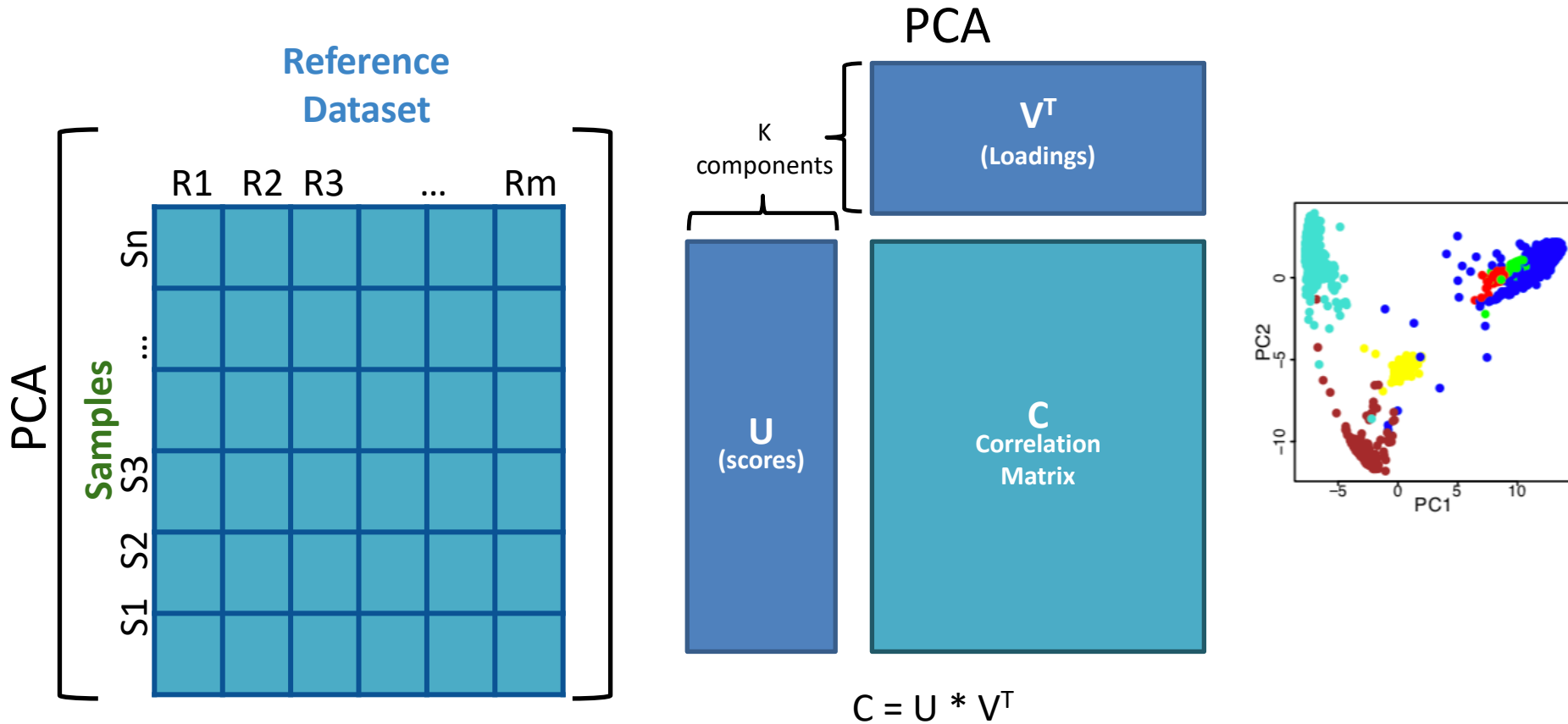
What is RCA?

- 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

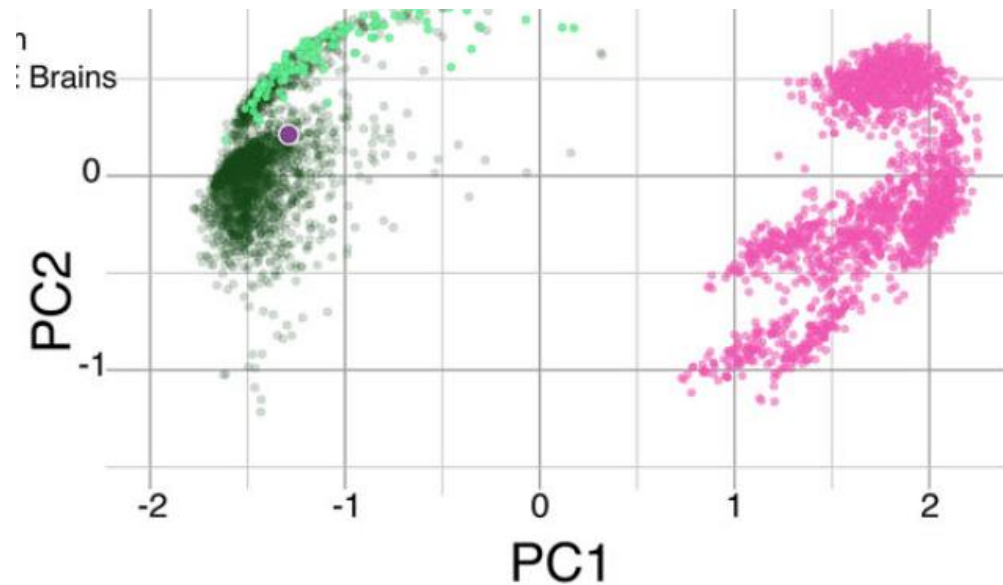
Projection to external dataset



PCA on correlation matrix



Placing
Brain expression data
in context of all other Body Tissues
(expression from GTEx)



Unsupervised Mining

CCA

Sorcerer II Global Ocean Survey

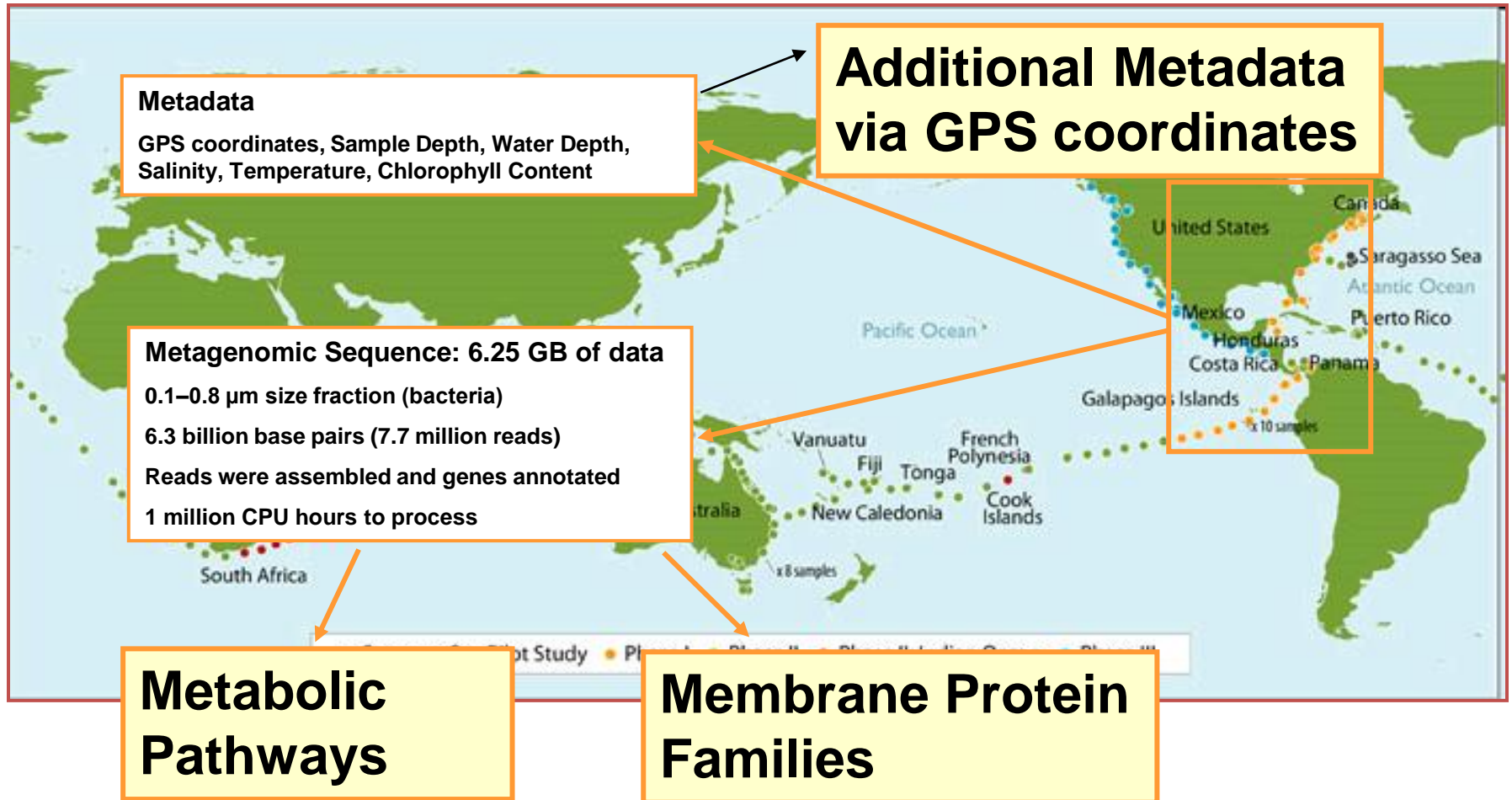


Sorcerer II journey August 2003- January 2006

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

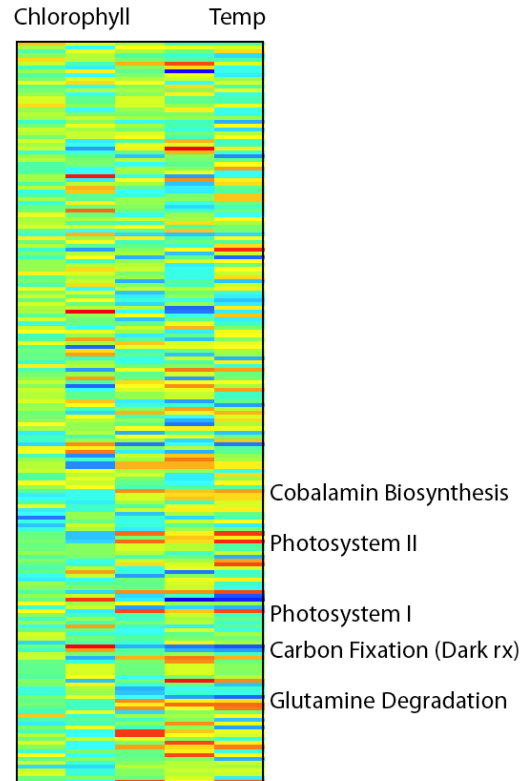
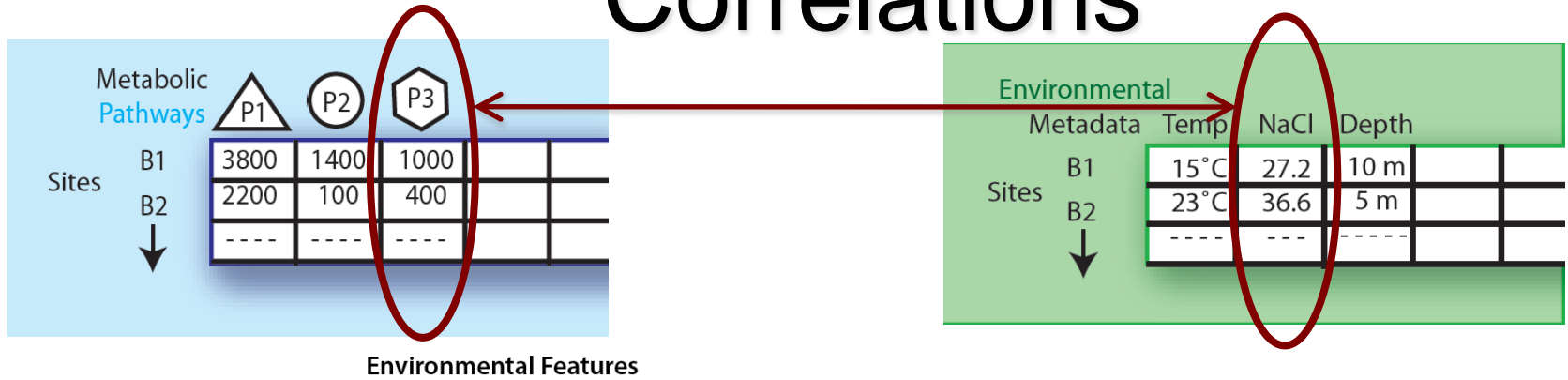
Metadata

Sites

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

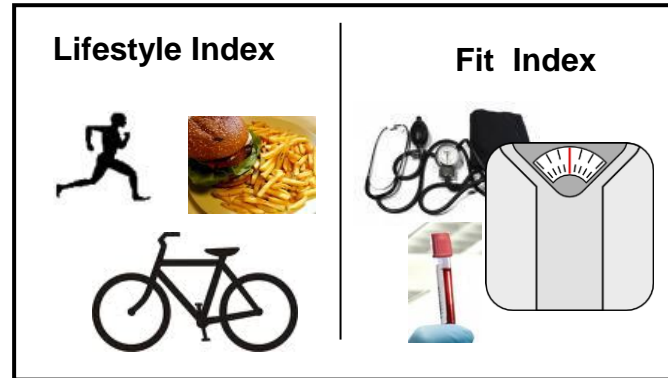
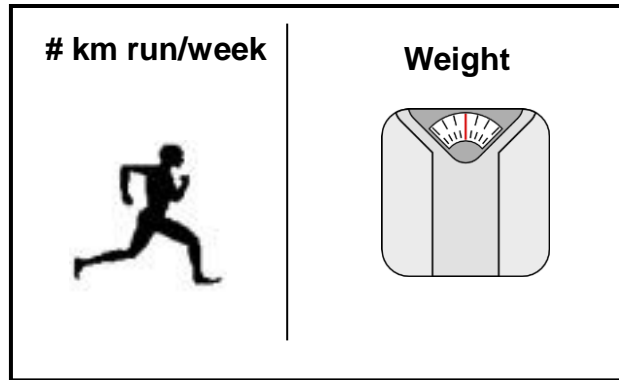
Expressing data as matrices indexed by site, env. var., and pathway usage

Simple Relationships: Pairwise Correlations



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

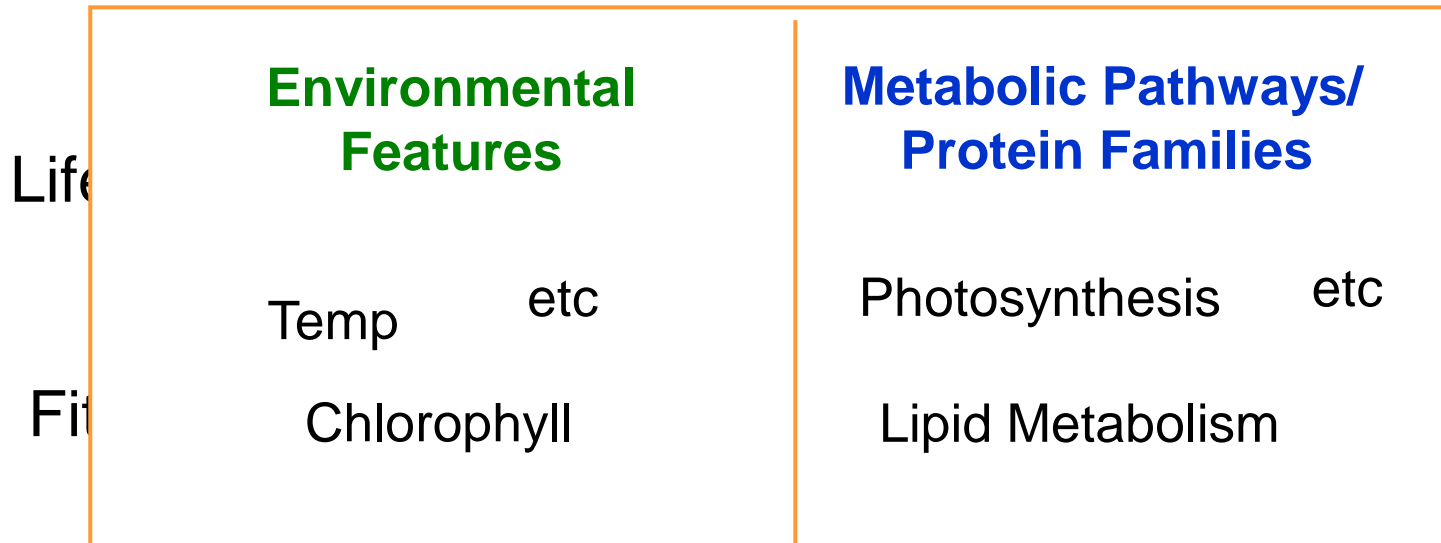
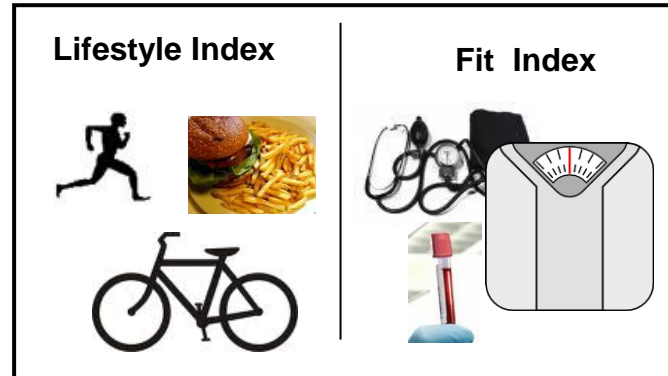
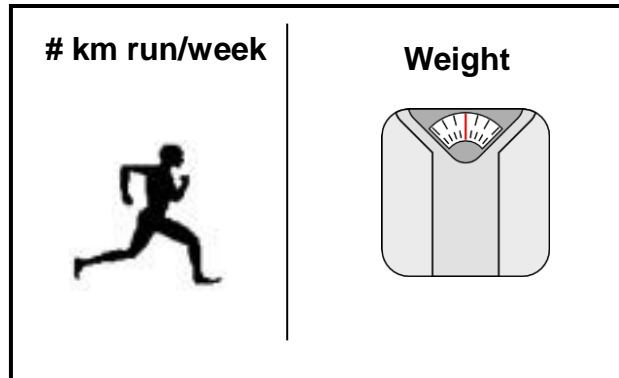
Canonical Correlation Analysis: Simultaneous weighting



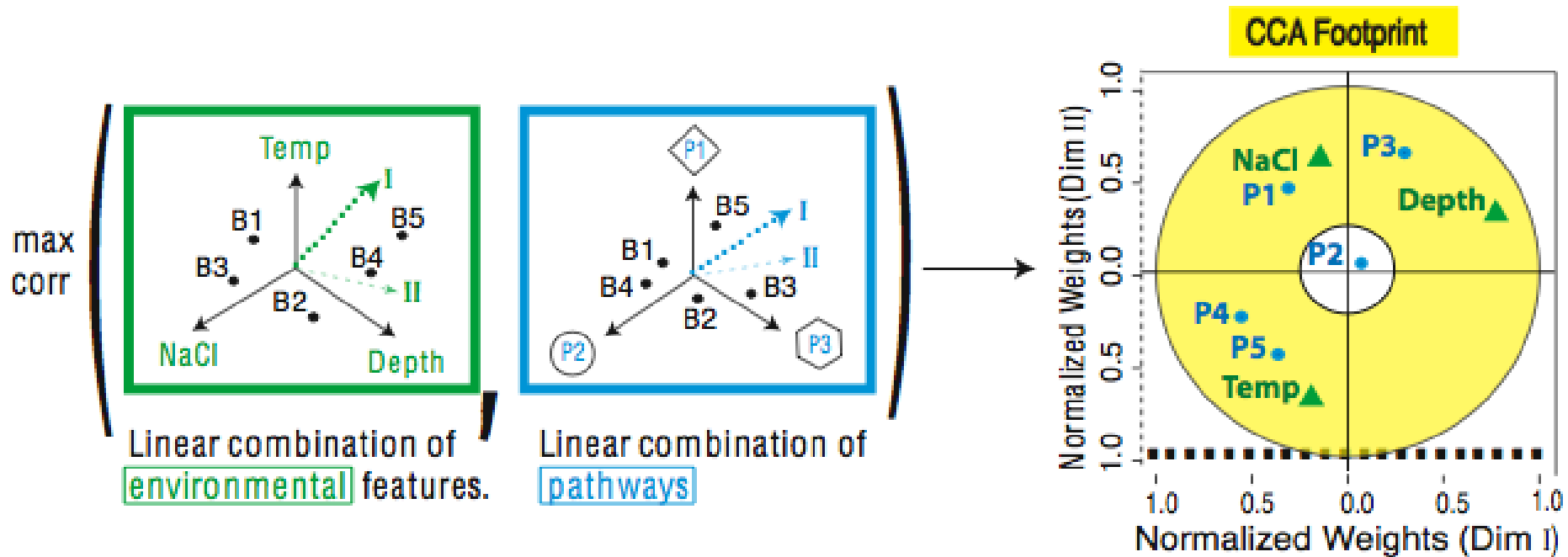
$$\text{Lifestyle Index} = a \text{  + b \text{  + c \text{ $$

$$\text{Fit Index} = a \text{  + b \text{  + c \text{ $$

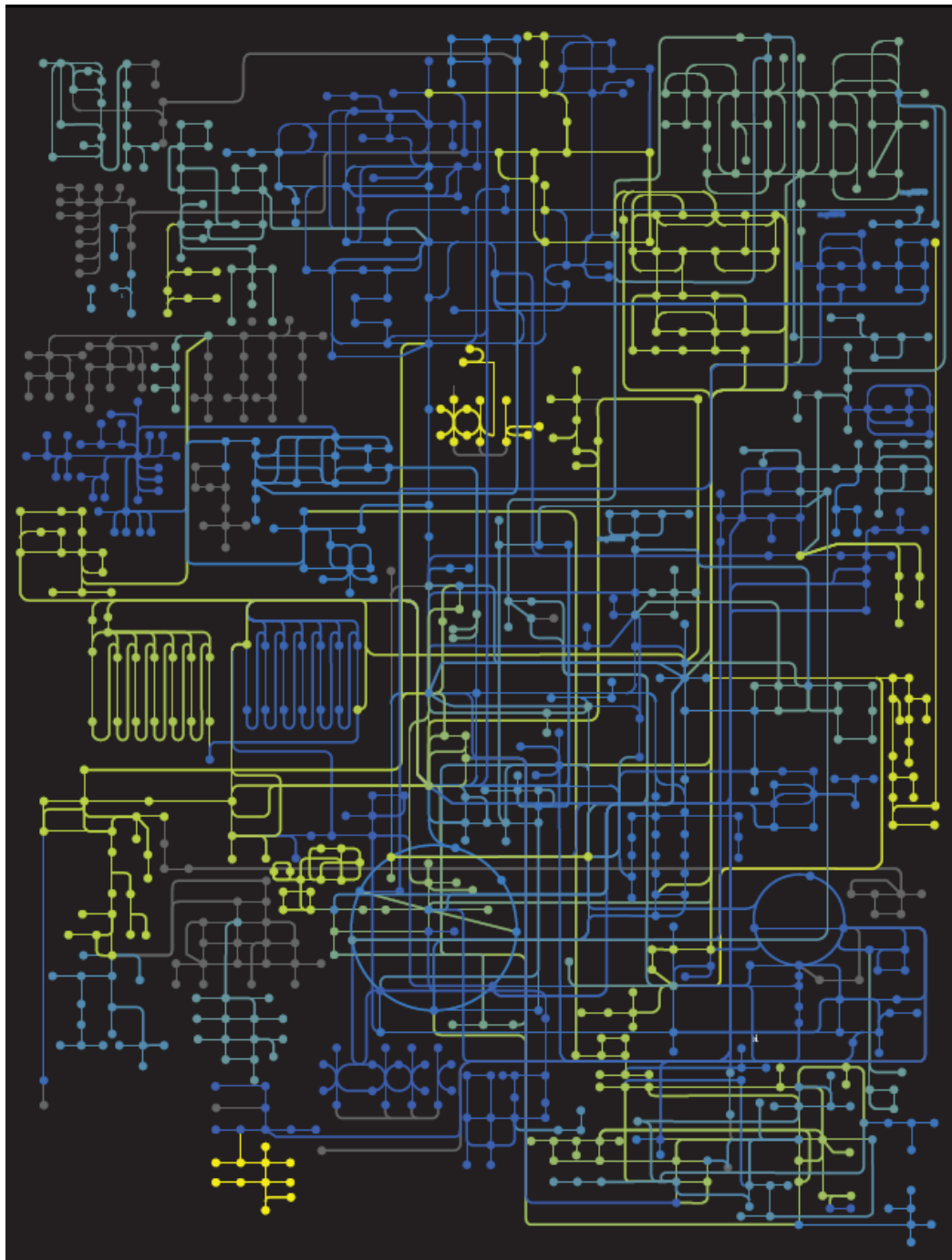
Canonical Correlation Analysis: Simultaneous weighting



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.

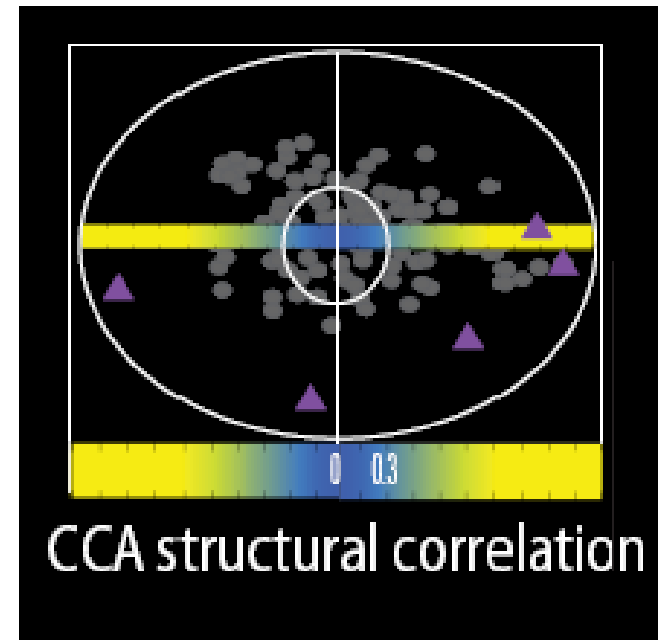


Strength of Pathway co-variation with environment

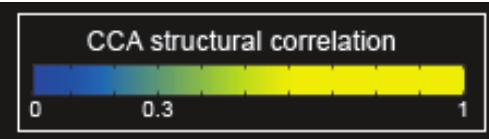


Environmentally invariant

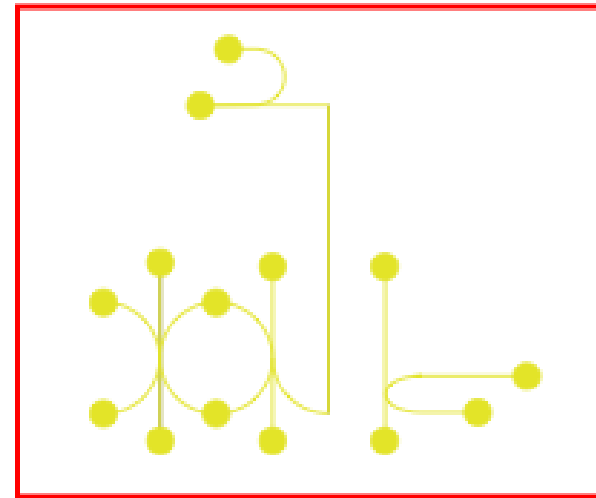
Environmentally variant



Conclusion #1: energy conversion strategy, temp and depth



Photosynthesis



Oxidative
Phosphorylation

