

Espaliers

A Visualization Method for Big Data

Max Robinson* Institute for Systems Biology
Gustavo Glusman

Greg Eley Scimentis, LLC

Joseph G. Vockley Inova Translational Medicine Institute
John E. Niederhuber



Big Data: The Object/Attribute Matrix

$$A = \begin{matrix} & \langle a_{1j}, \dots, a_{mj} \rangle & n \text{ Attributes} \\ \begin{matrix} \langle a_{1i}, \dots, a_{mi} \rangle \\ m \text{ Objects} \end{matrix} & \begin{matrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{matrix} \end{matrix}$$

$a_{ij} \geq 0$ Non-negative!

Example: text mining
 a_{ij} = Occurrences of word j in text i

1. Population genetics (genotypes)
2. Gene expression matrix
3. Social network graph



1. Population genetics: genotypes

1000 Genomes, phase 3 genotypes

$$A = \begin{matrix} & n = 212,961 \text{ variant positions (1\%, AC} \geq 5) \\ \begin{matrix} m = 2,504 \\ \text{DNA samples} \end{matrix} & \begin{matrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{matrix} \end{matrix}$$

a_{ij} = sample i 's genotype at variant j

81,029,508 of the $a_{ij} > 0$ (15%)

Variant position j : 1 2 3 4
Variant allele: C A G C
Reference Sequence: ACGTATACCTCAAGAGATTTCGAGTGCCTCGA

$i = 3$ Maternal allele: C A T G
Paternal allele: T A G G
Genotype: 1 2 1 0
 $a_{31} \ a_{32} \ a_{33} \ a_{34}$



1. Population genetics: genotypes

1000 Genomes, phase 3 genotypes

$$A = \begin{matrix} & n = 212,961 \text{ variant positions (1\%, AC} \geq 5) \\ \begin{matrix} m = 2,504 \\ \text{DNA samples} \end{matrix} & \begin{matrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{matrix} \end{matrix}$$

a_{ij} = sample i 's genotype at variant j

81,029,508 of the $a_{ij} > 0$ (15%)

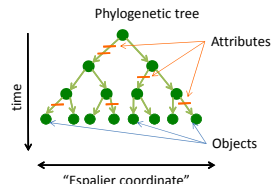
Σ allele count (AC)

Variant position j : 1 2 3 4
Variant allele: C A G C
Reference Sequence: ACGTATACCTCAAGAGATTTCGAGTGCCTCGA

$i = 3$ Maternal allele: C A T G
Paternal allele: T A G G
Genotype: 1 2 1 0
 $a_{31} \ a_{32} \ a_{33} \ a_{34}$



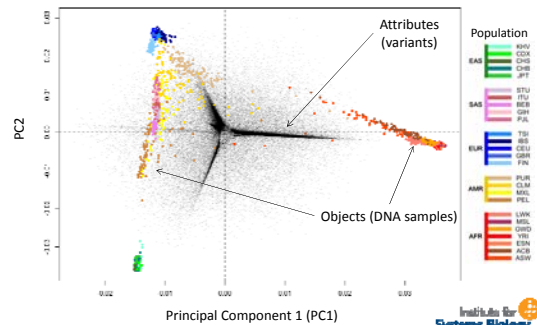
The Espalier Concept

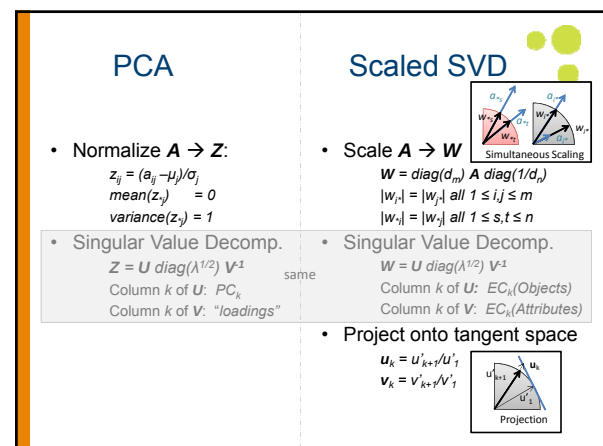
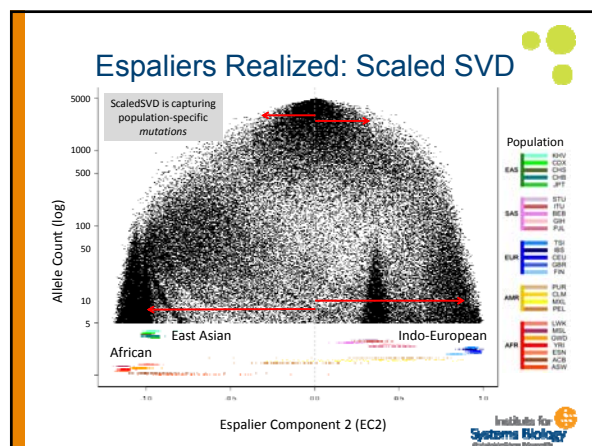
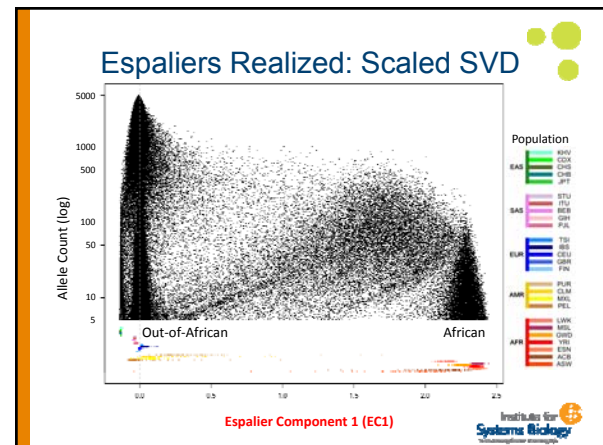
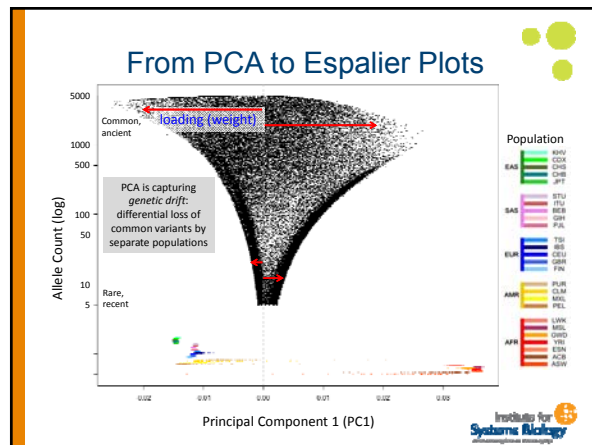
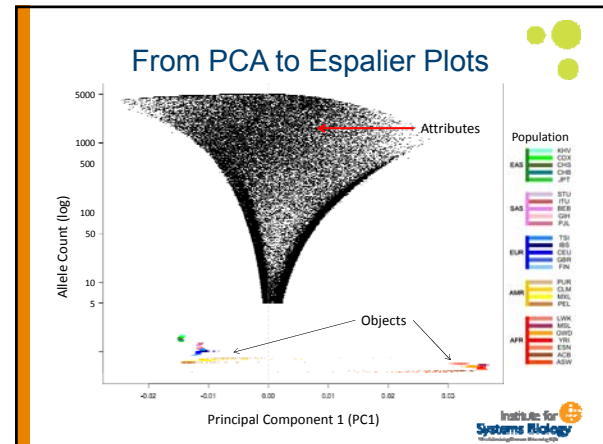
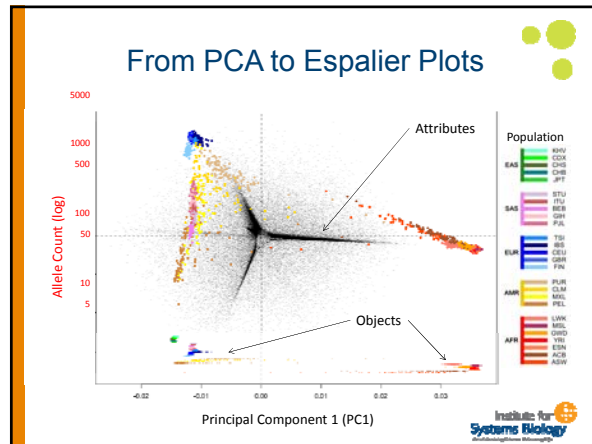


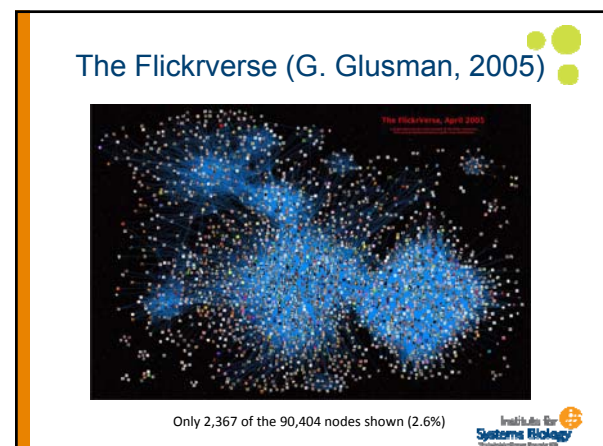
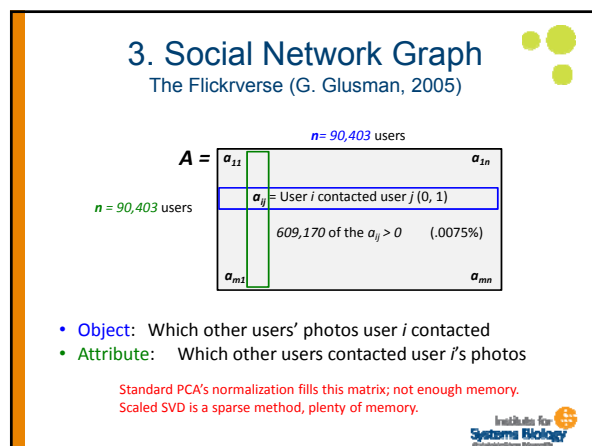
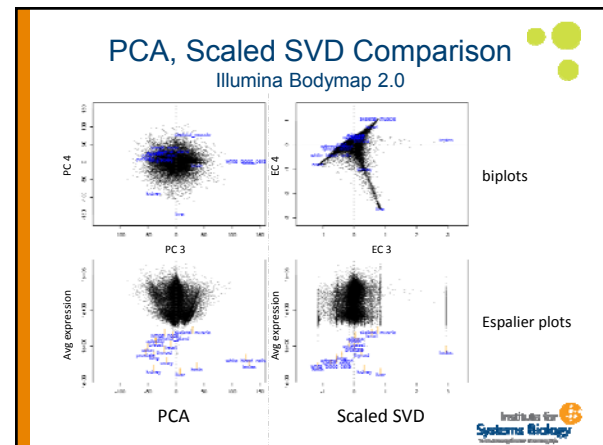
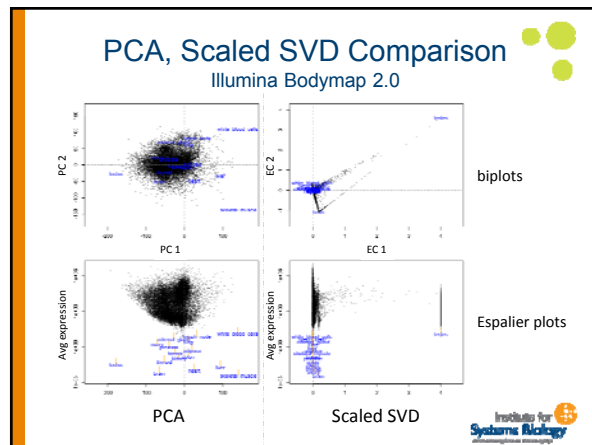
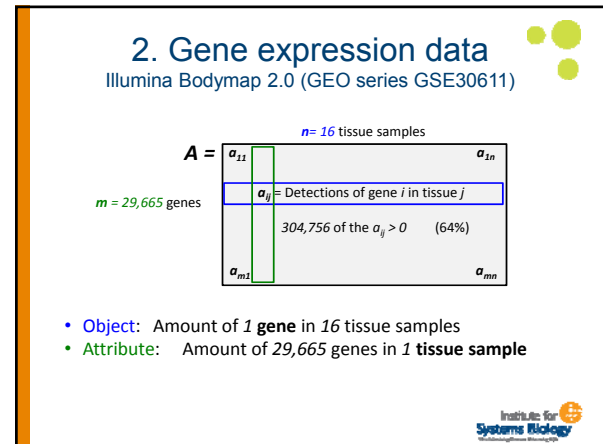
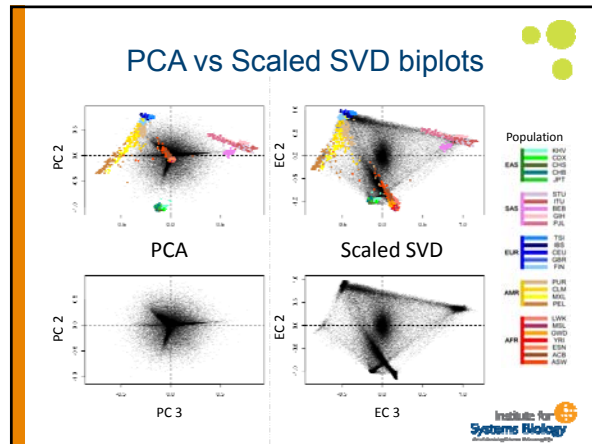
The correlation structure of a data matrix places objects and attributes in related positions along a *shared coordinate*.



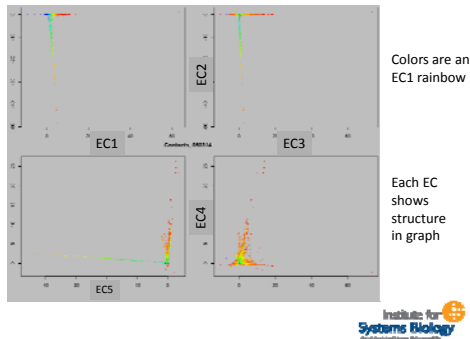
Population Structure via Principal Component Analysis



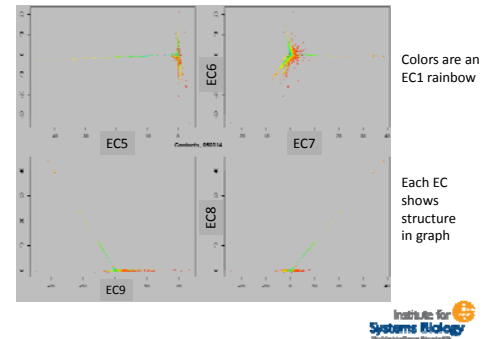




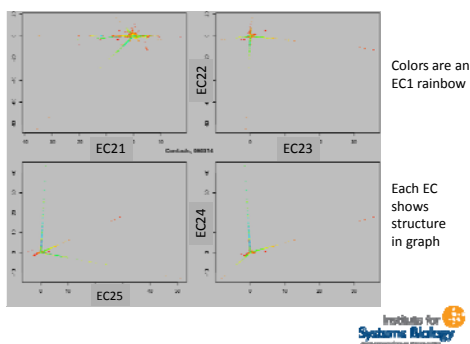
The Flickrverse, analyzed by Scaled SVD



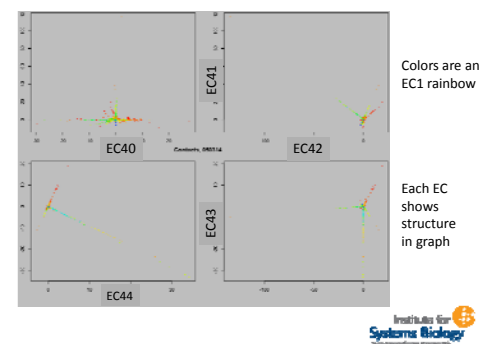
The Flickrverse, analyzed by Scaled SVD



The Flickrverse, analyzed by Scaled SVD



The Flickrverse, analyzed by Scaled SVD



Summary

- Scaled SVD--linear data reduction method
 - Designed for non-negative data
 - Geometric rationale (Espalier)
 - Big Data: preserves sparseness of matrix
 - Appropriate first step for co-clustering
 - Transpose of matrix yields identical result
 - Axis orientation, ordering properties of SVD, like PCA
 - *Qualitatively different results from PCA*
- Espalier Plots
 - One espalier coordinate vs. row, column sums
 - Grouped Objects, Attribute span data quantity range
 - Espalier coordinate biplots

Acknowledgements

Thanks to my co-authors:

Gustavo Glusman, Greg Eley, Joseph G. Vockley,
and John E. Niederhuber

and to these members of the Hood and Shmulevich groups at ISB for
helpful discussions of this work:

Nigel Clegg, Anna Stittrich, Theo Knijnenberg,
Bill Longabaugh, Alissa Severson, Chris Lausted,
and Ilya Shmulevich

Thanks especially to our collaborators at

The Inova Translational Medicine Institute
for the opportunity to pursue this research!