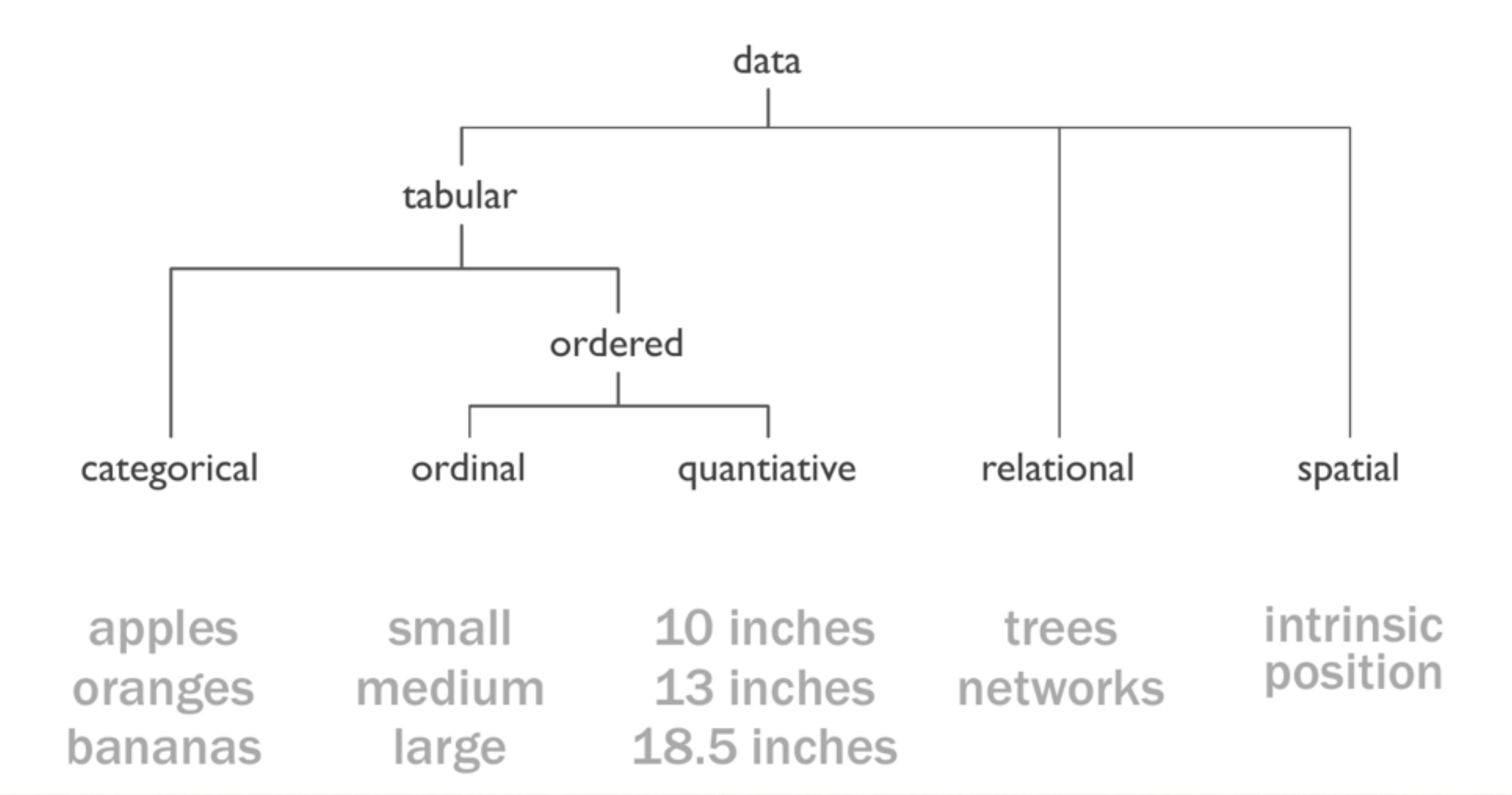
Michael Grossberg

Data Visualization

Multi and High
Dimensional Data

Review: Data Types

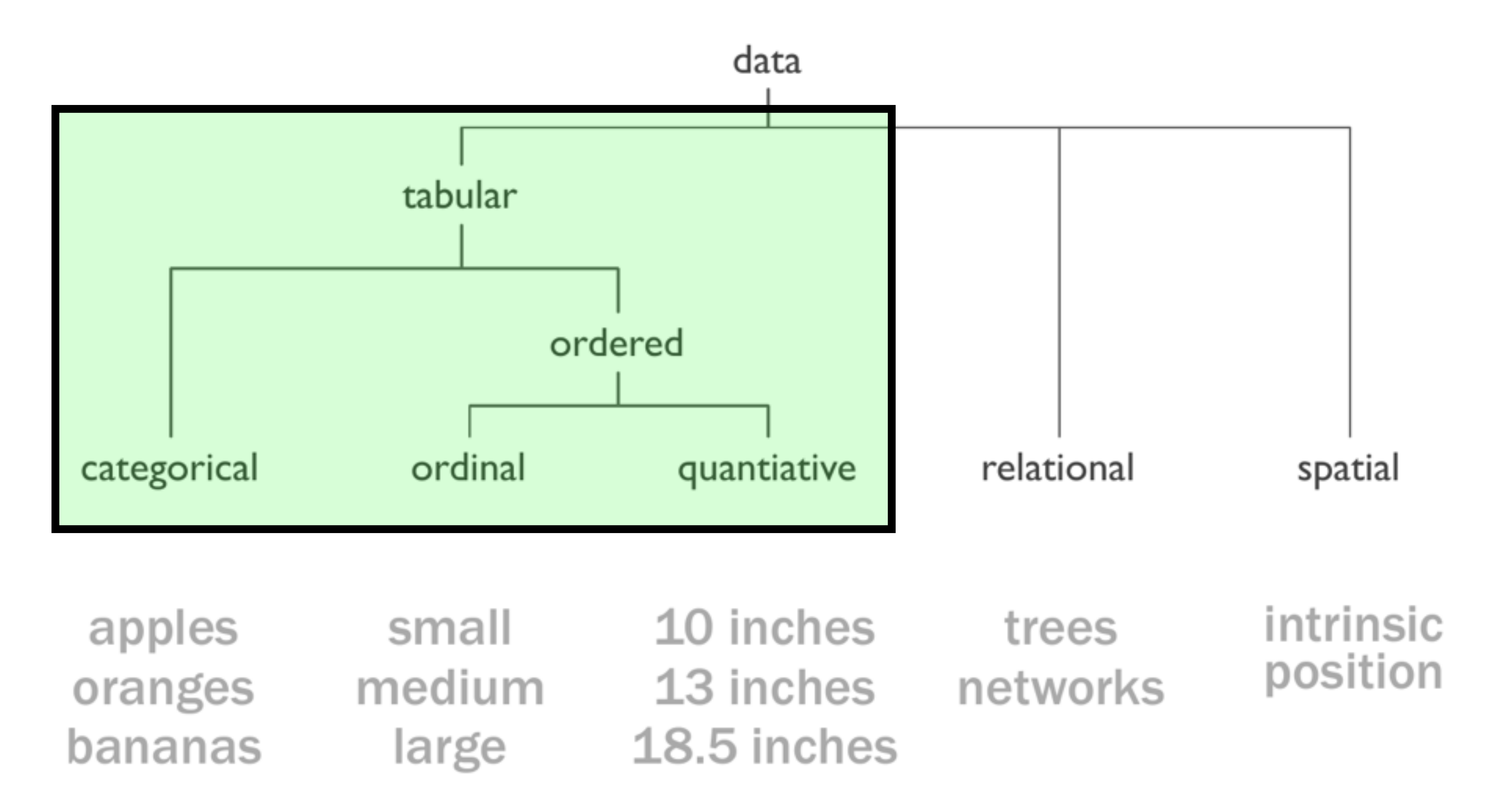


Multi-Dimensional Data

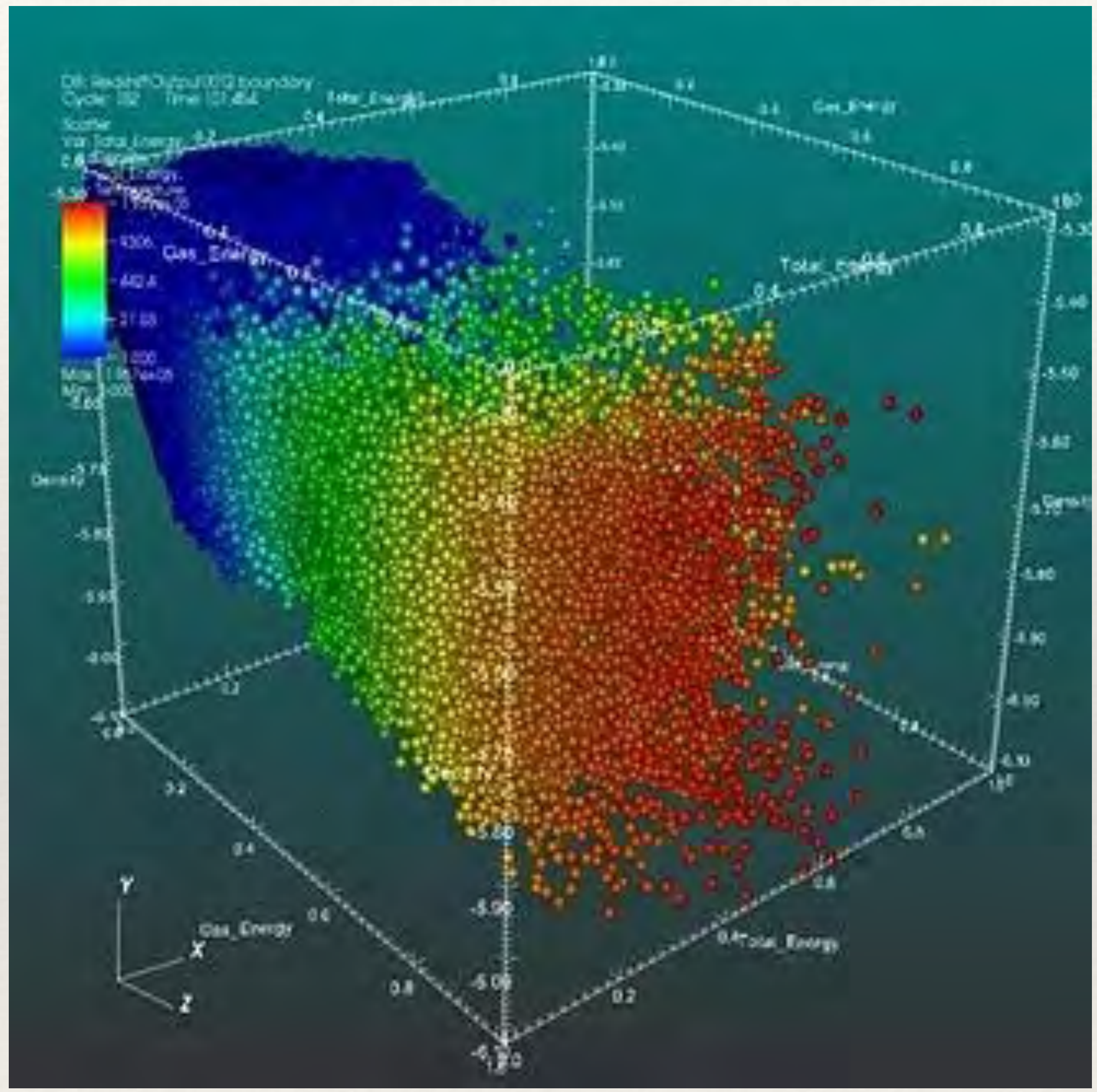
- Tabular data, containing
  - rows (records)
  - columns (dimensions)
  - rows >> columns
  - identifiers introduce semantics
- Independent & dependent variables
  - dependent are f(independent)

	Age	Gender	Height
Bob	25	M	181
Alice	22	F	185
Chris	19	M	175

Muli-Dimensional Type



Limits to Displaying Dimensions



Limit? 4D? 5D? … 10D!?!

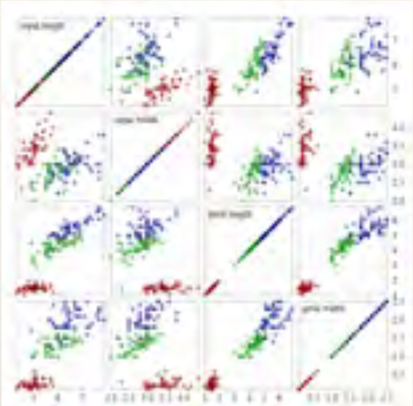
High-Dimensional Data Visualization

- How many dimensions?
  - ~50 –tractable with “just” vis
  - ~1000 –need analytical methods
- How many records?
  - ~ 1000 –“just” visis fine
  - >> 10,000 –need analytical methods
- Homogeneity
  - Same data type?
  - Same scales?

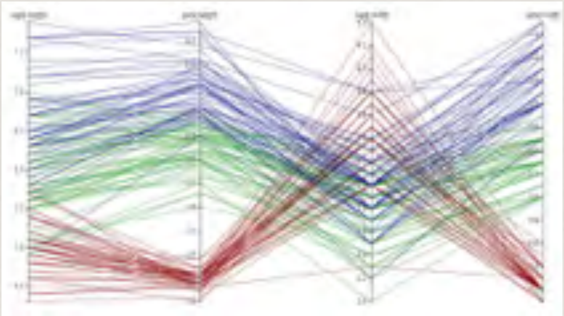
	Age	Gender	Height
Bob	25	M	181
Alice	22	F	185
Chris	19	M	175

	BPM 1	BPM 2	BPM 3
Bob	65	120	145
Alice	80	135	185
Chris	45	115	135

Analytic Component



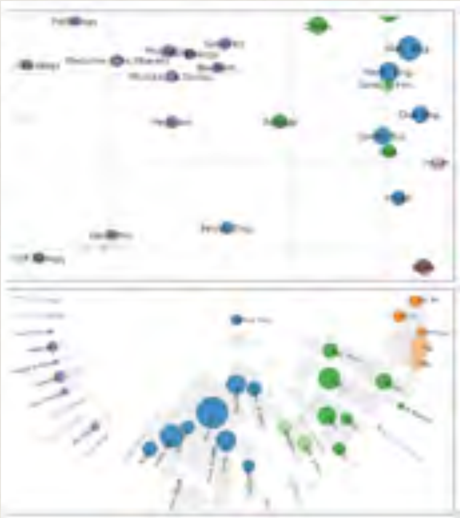
Scatterplot Matrices



Parallel Coordinates

Pixel-based visualizations / heat maps

Multi-dimensional Scaling



LDA

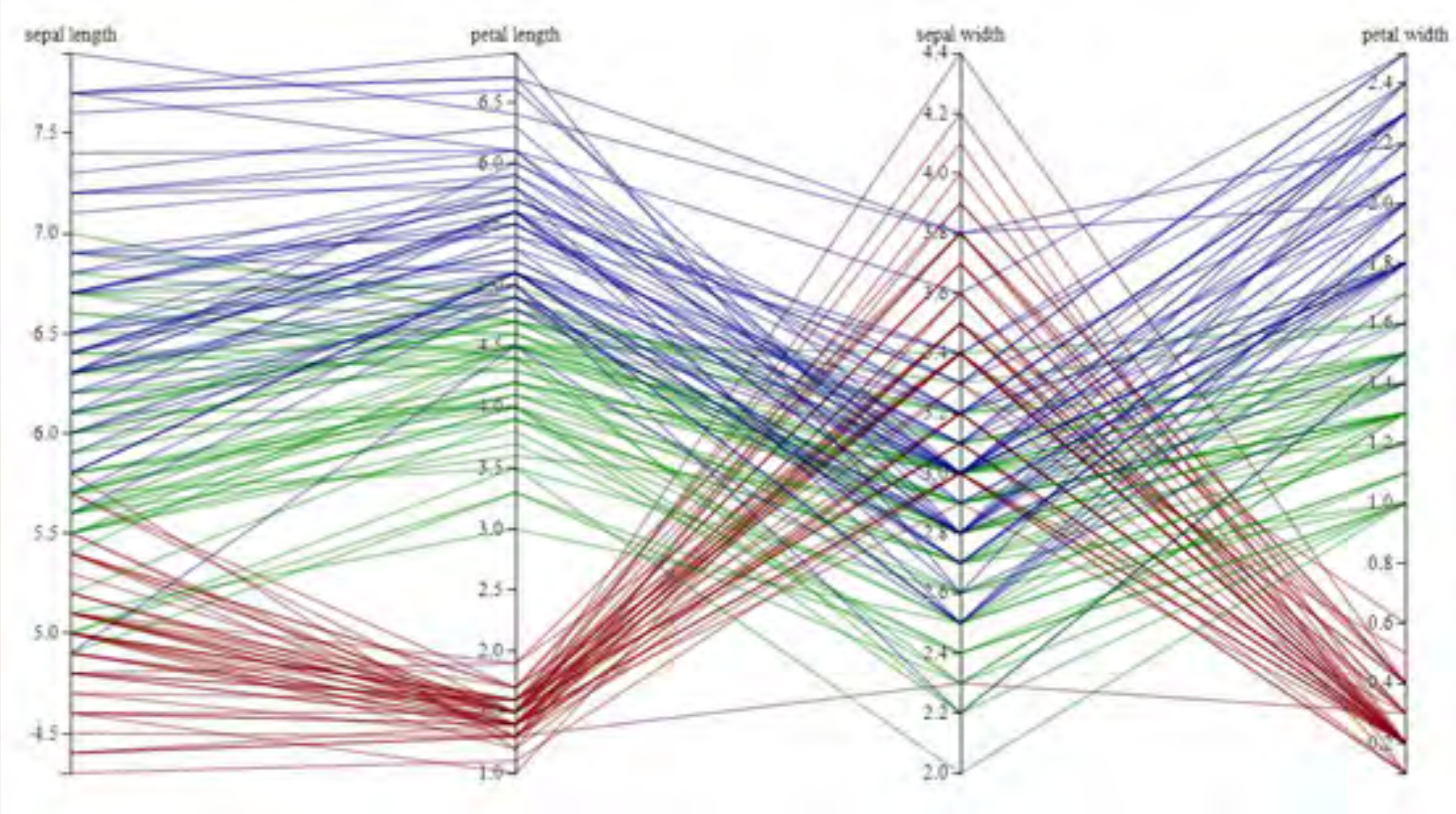
no / little analytics

strong analytics component

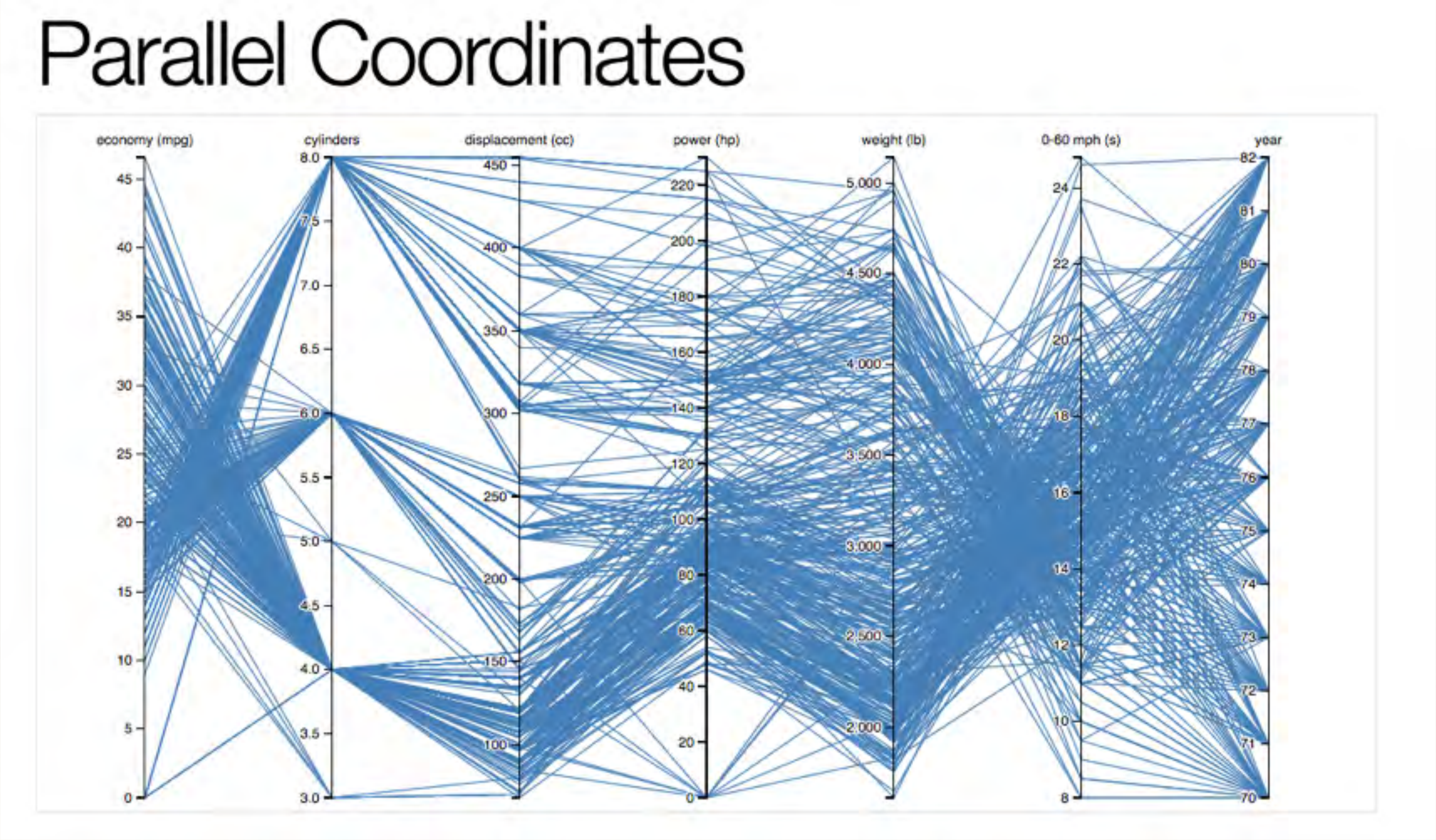
Geometric Methods

Parallel Coordinates

- Each axis represents dimension
- Lines connecting axis represent records
- Suitable for
  - all tabular data types
  - heterogeneous data



D3 Parallel Coordinates



http://bl.ocks.org/jasondavies/1341281

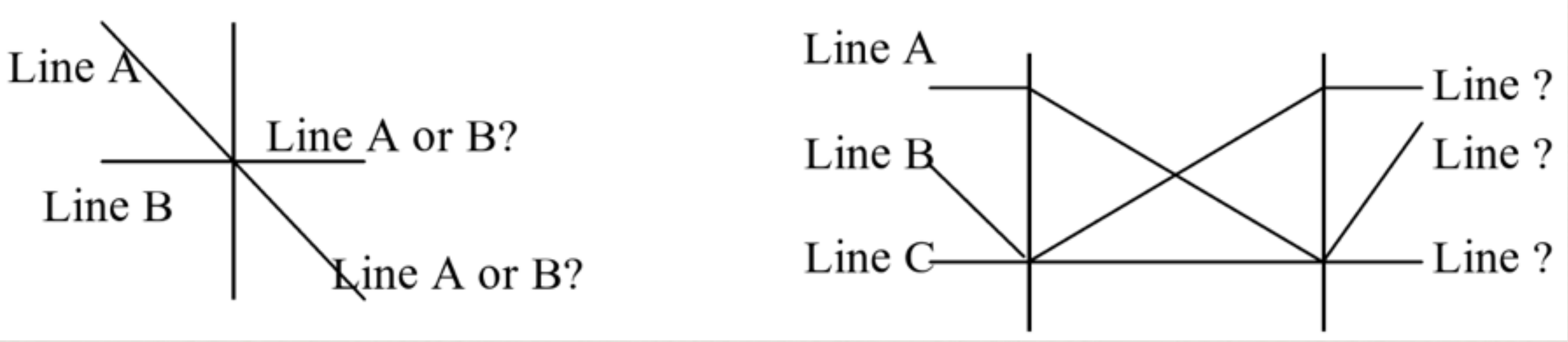
Parallel Coordinates

- Shows primarily relationships between adjacent axis
- Limited scalability (~50 dimensions, ~1-5k records)
- Transparency of lines
- Interaction is crucial
- Axis reordering
- Brushing
- Filtering

- Algorithmic approaches:
- Choosing dimensions
- Choosing order
- Clustering & aggregating records

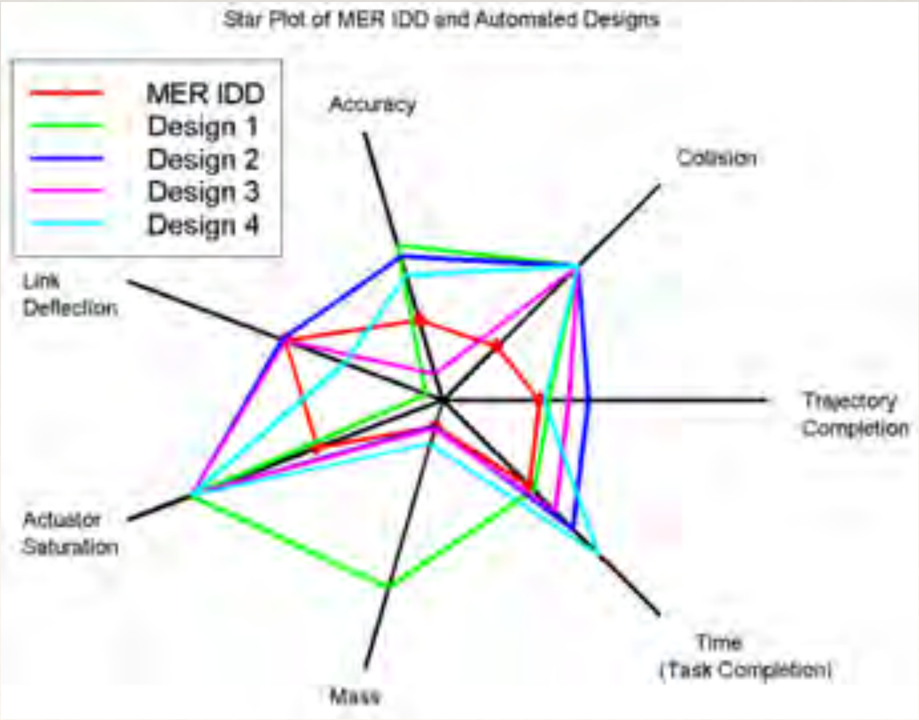
500-Axis Parallel Coordinate

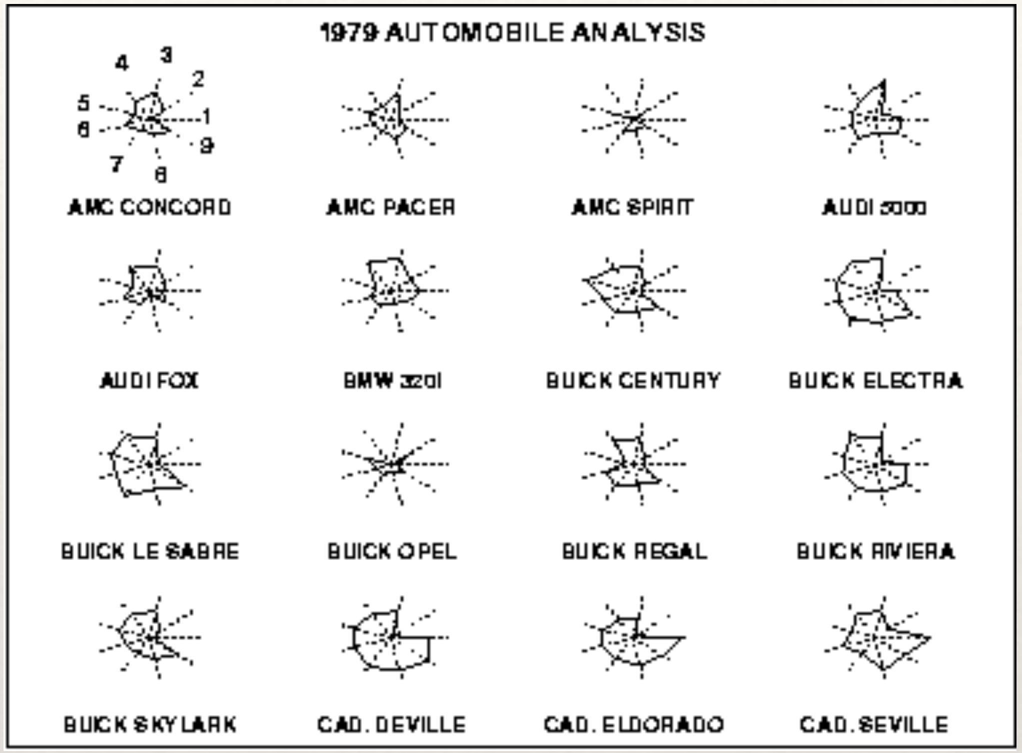
Ambiguities



Star Plot/Radar Plot

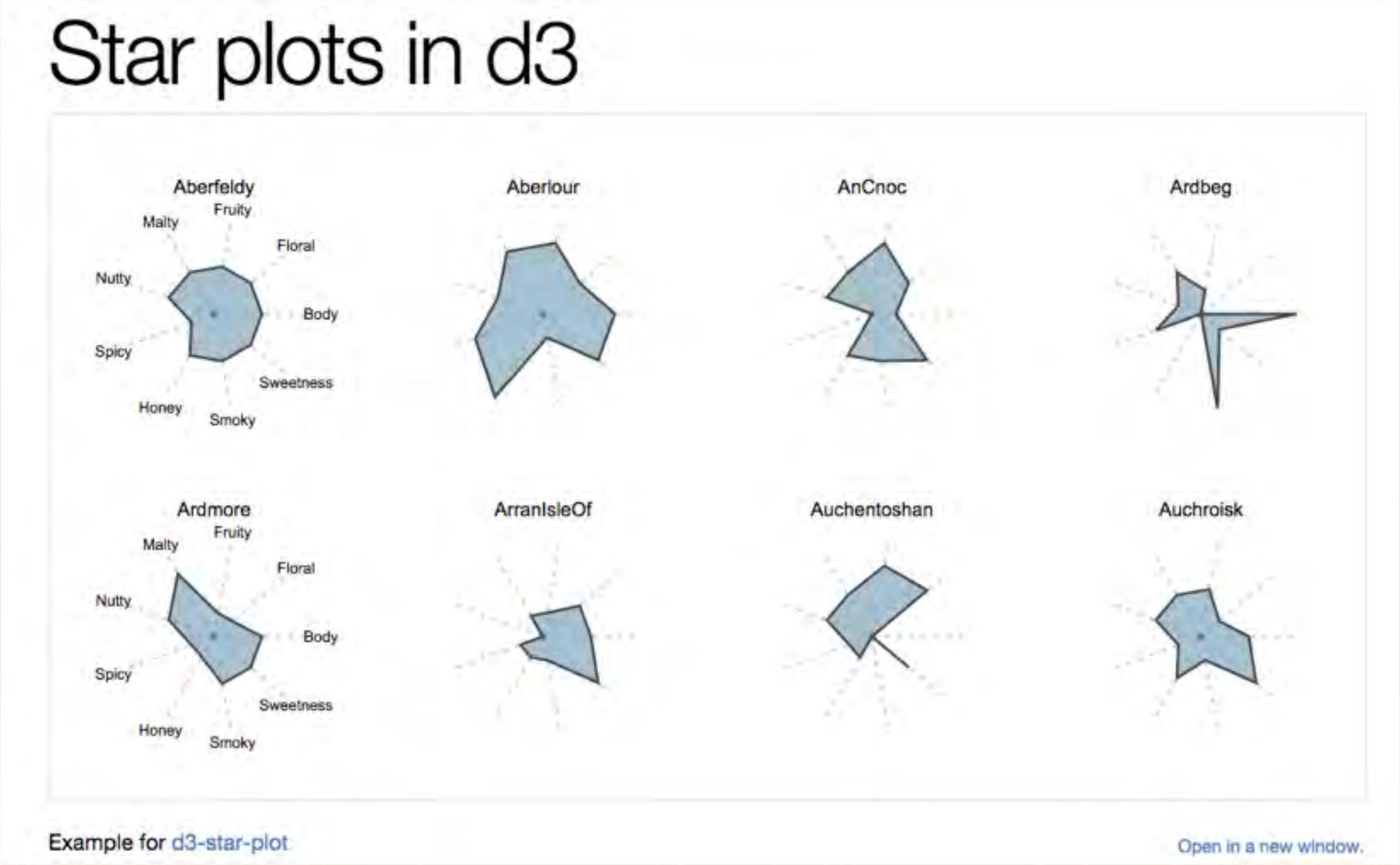
- Similar to parallel coordinates
- Radiate from a common origin





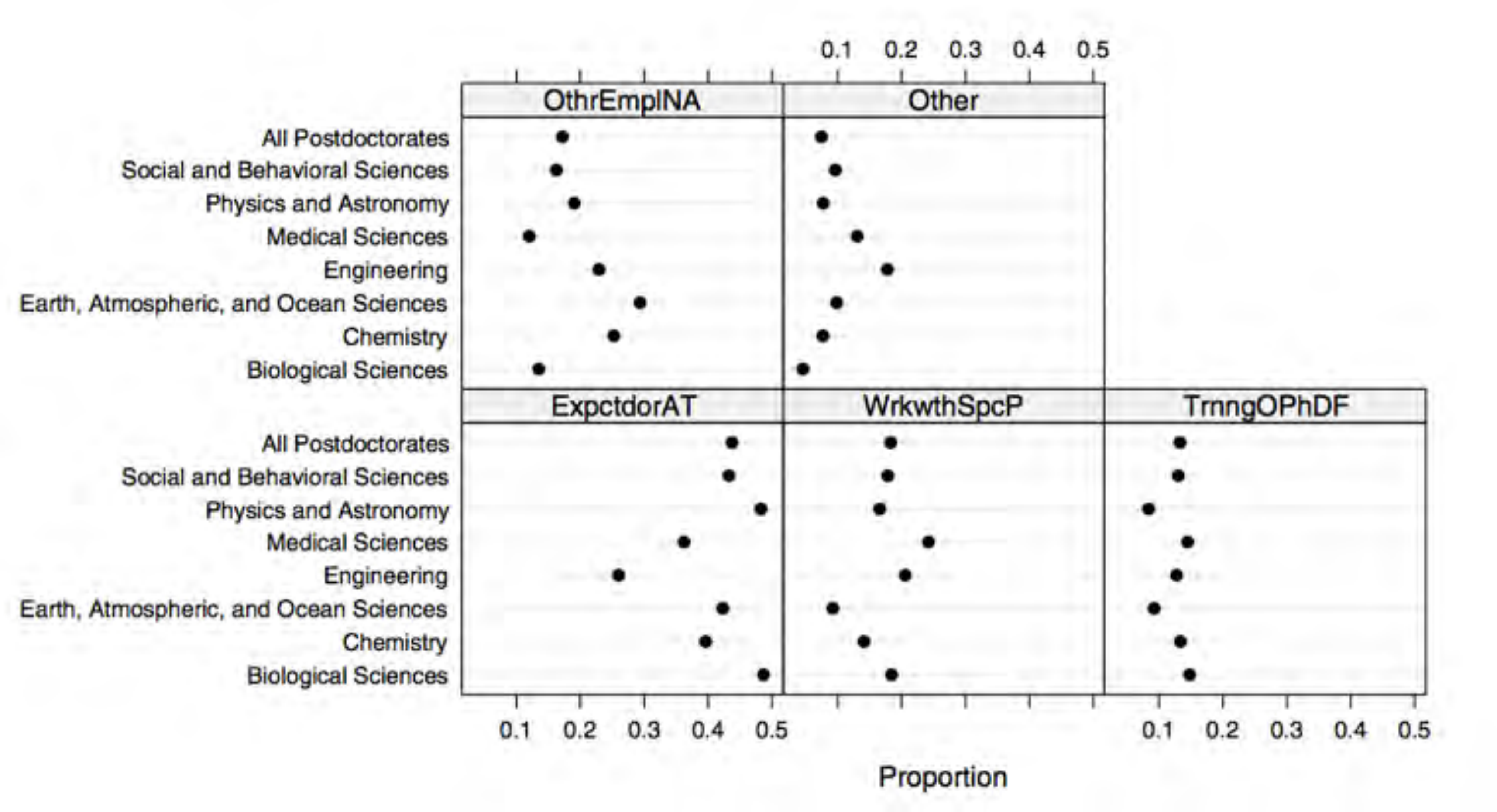
Coekin 1969

D3: Star Plot



http://bl.ocks.org/kevinschaul/8213691

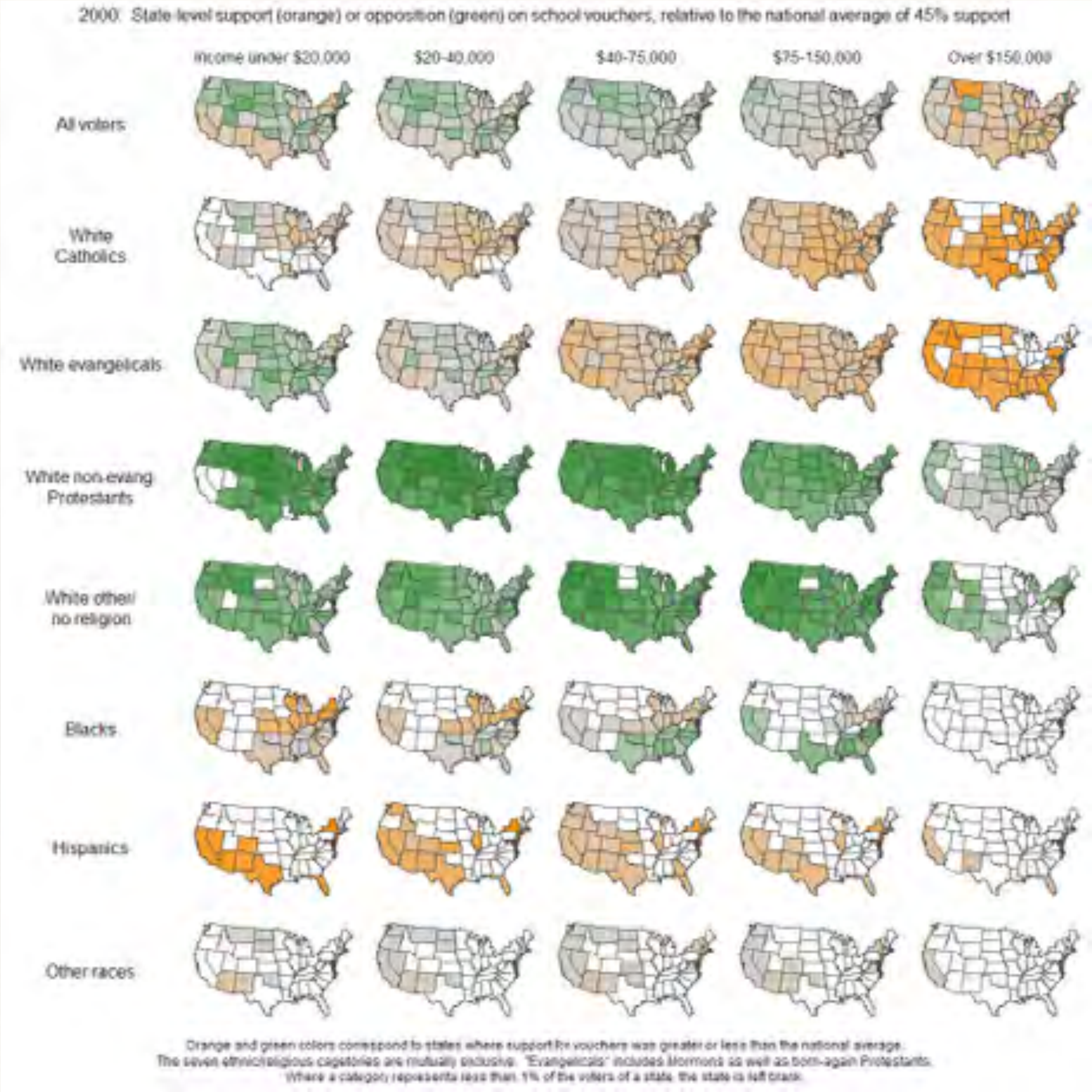
Trellis Plots



Multiple Plots (often with shared axis)

Small Multiples

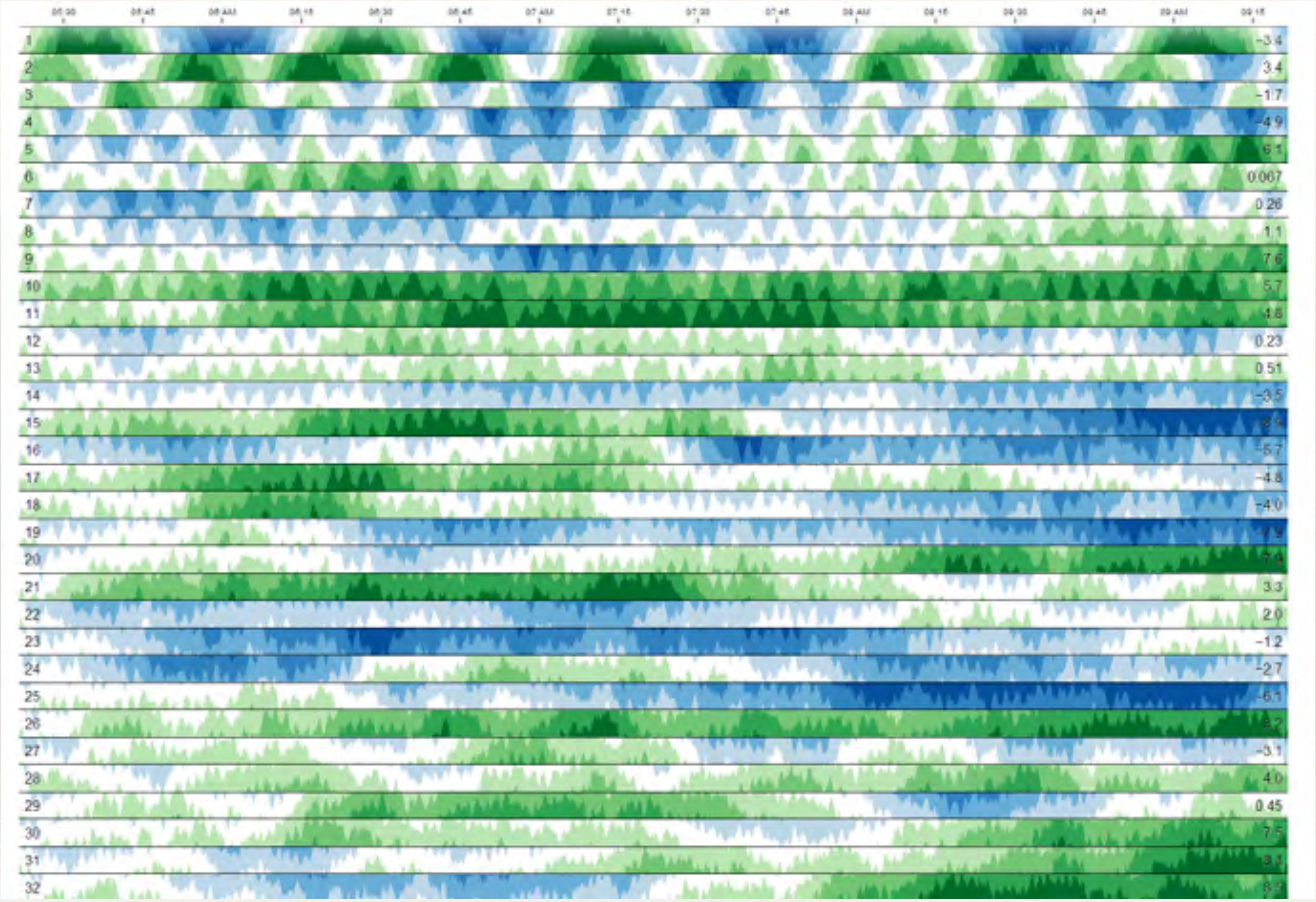
- Like Trellis
  - More plots
  - Axis may not be important
  - Each plot point on axis
- Use multiple views to show different partitions of a dataset



public support for vouchers, Andrew Gelman 2009

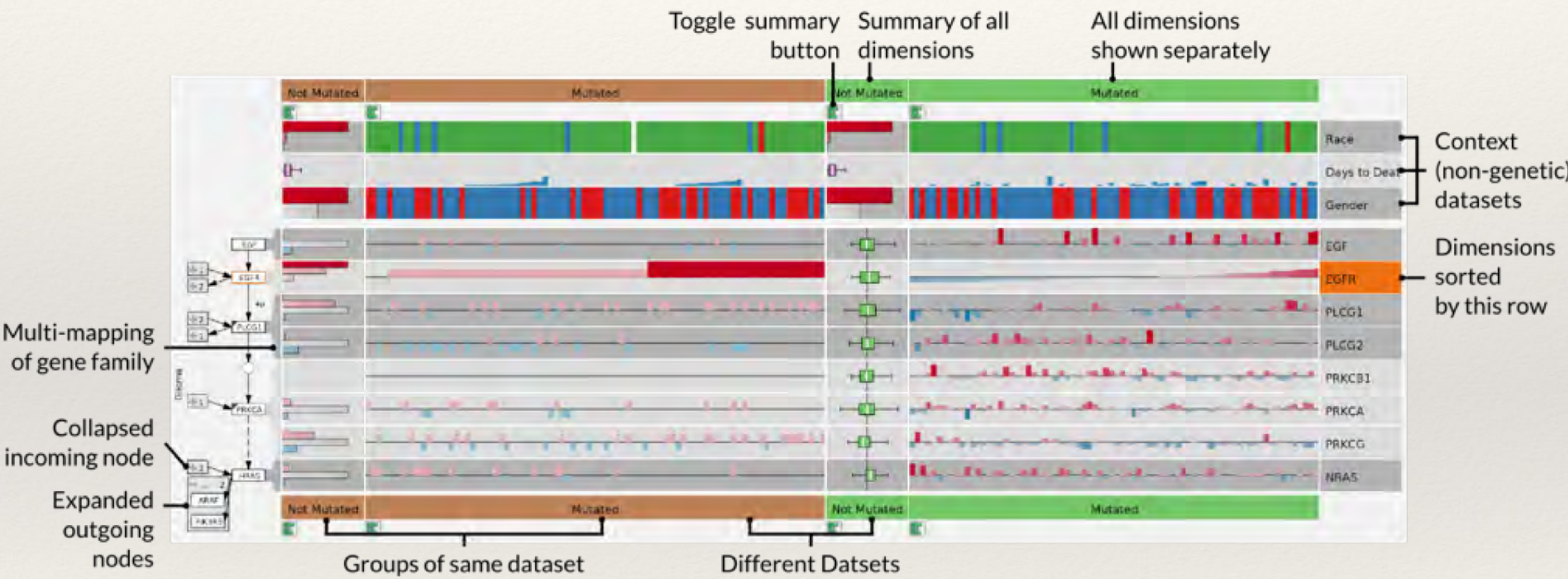
http://andrewgelman.com/2009/07/15/hard_sell_for_b/

Multiple Line Charts



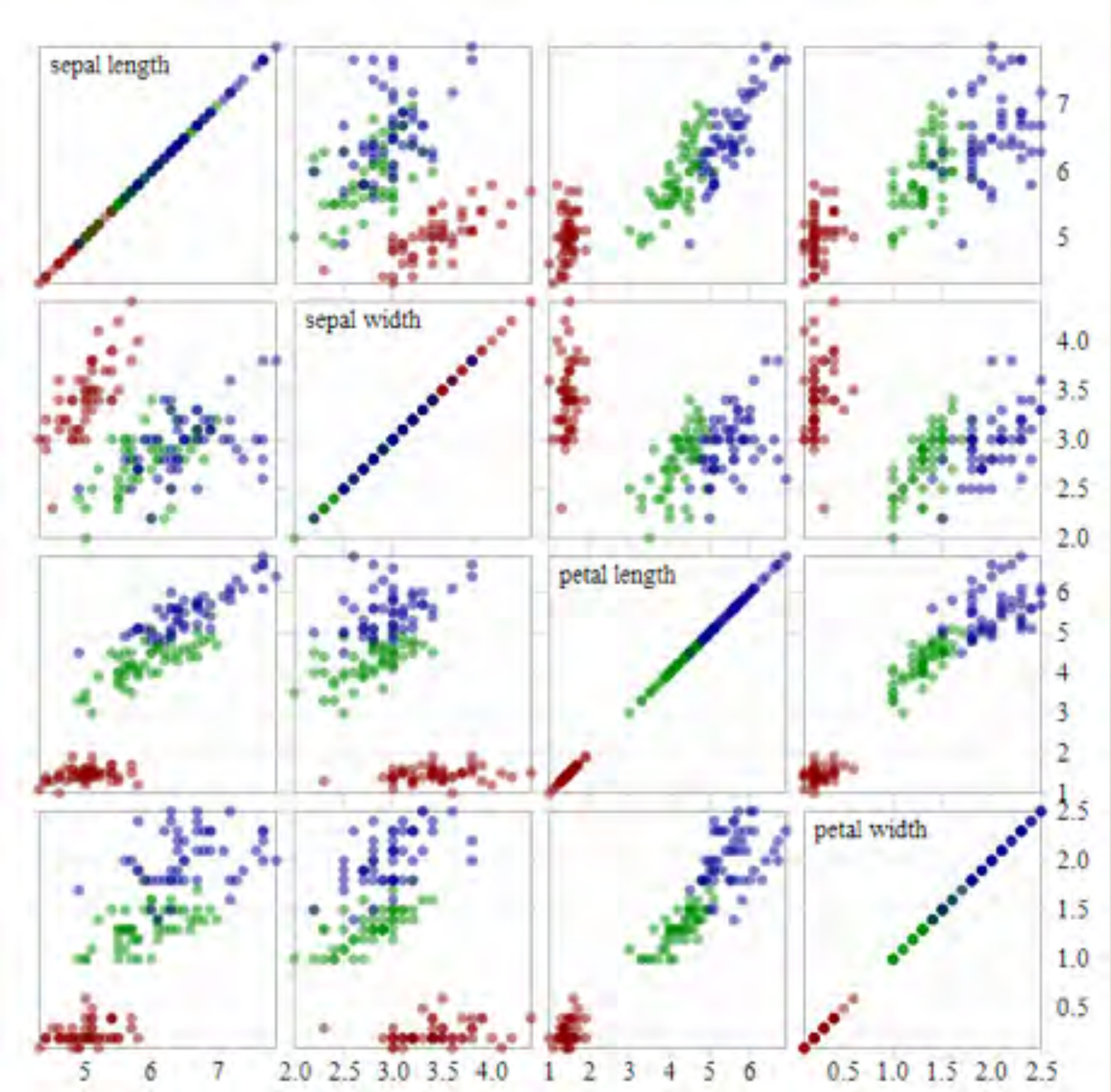
http://square.github.io/cubism/

Combining Various Charts



Scatterplot Matrices (SPLOM)

- Matrix of size d*d
- Each row/column is one dimension
- Each cell plots a scatterplot of two dimensions

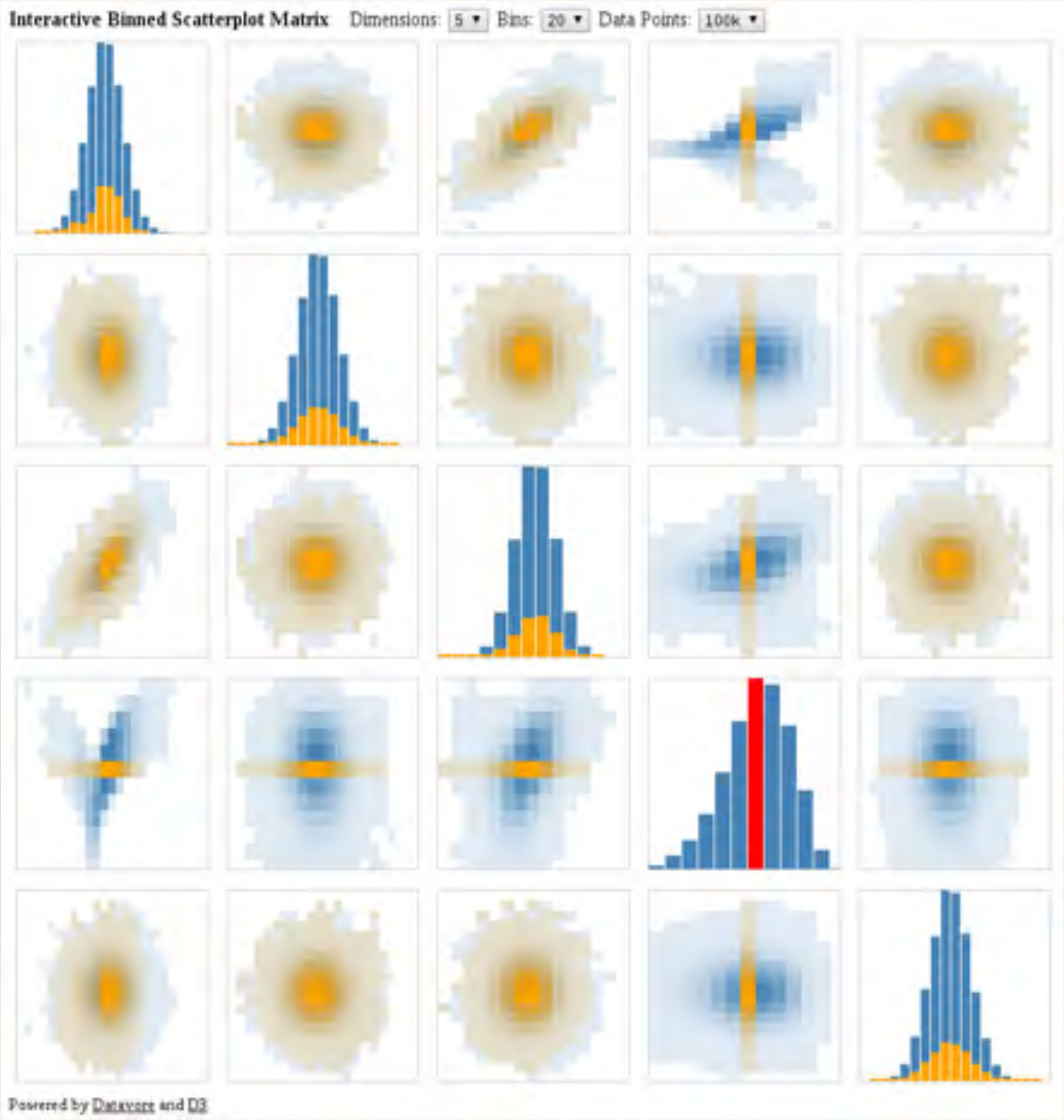


http://mbostock.github.io/d3/talk/20111116/iris-splom.html

Scatterplot Matrices

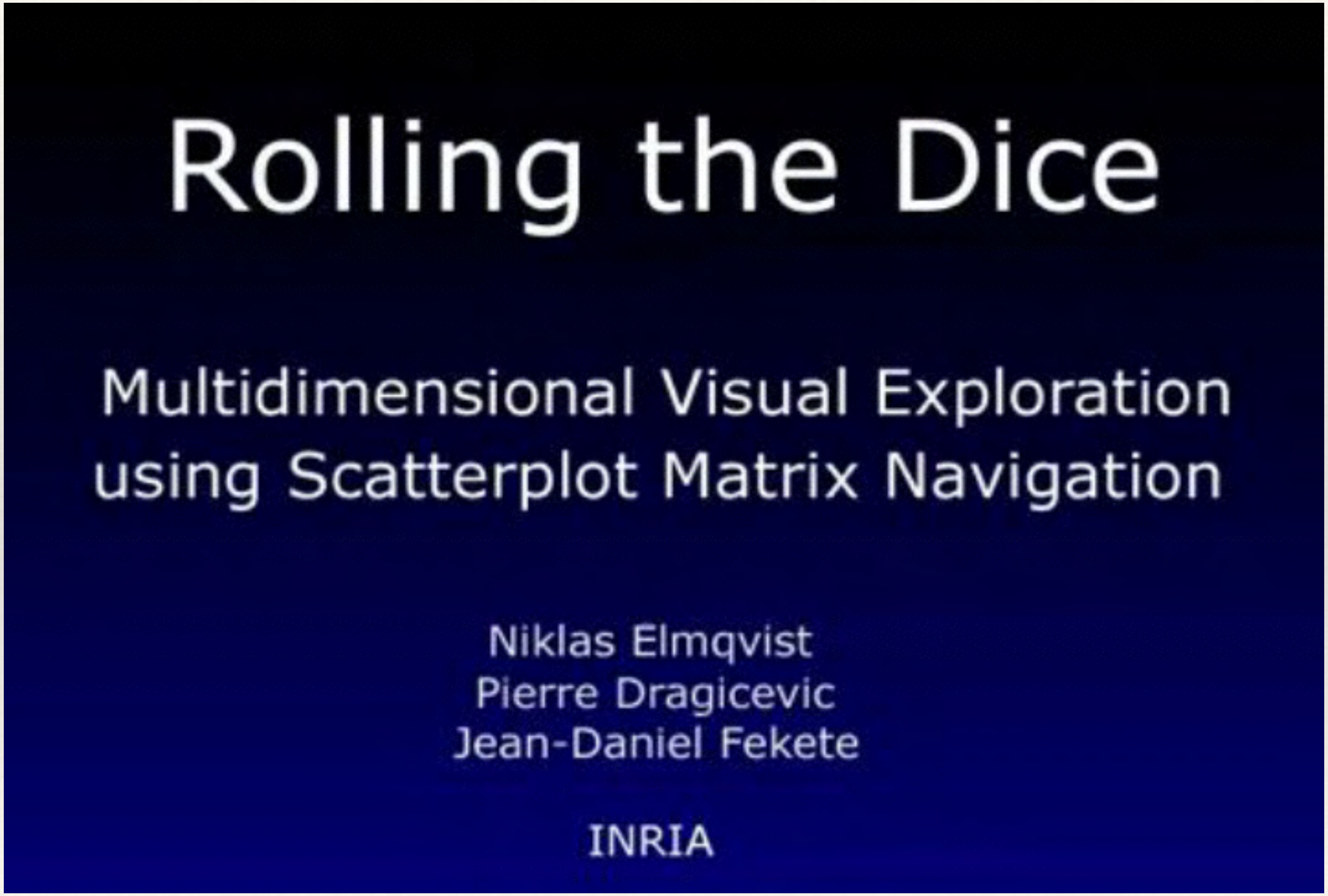
- Limited scalability (~20 dimensions, ~500-1k records)
- Brushing is important
- Often combined with "Focus Scatterplot" as F +C technique
- Algorithmic approaches:
- Clustering & aggregating records
- Choosing dimensions
- Choosing order

SPLOM Aggregation



Datavore: http://vis.stanford.edu/projects/datavore/splom/

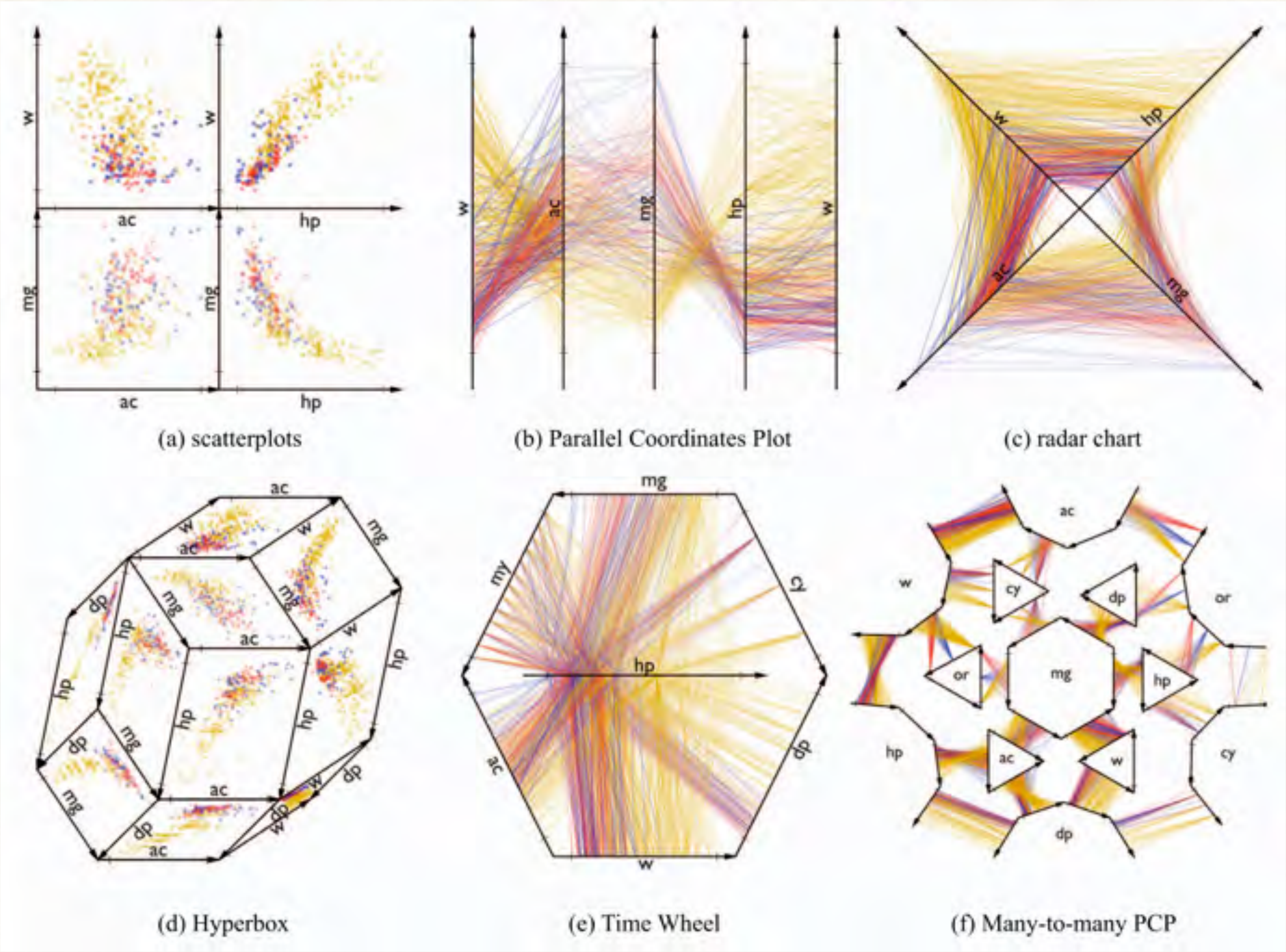
SPLOM F+C, Navigation



http://youtu.be/E1birsp9iYk

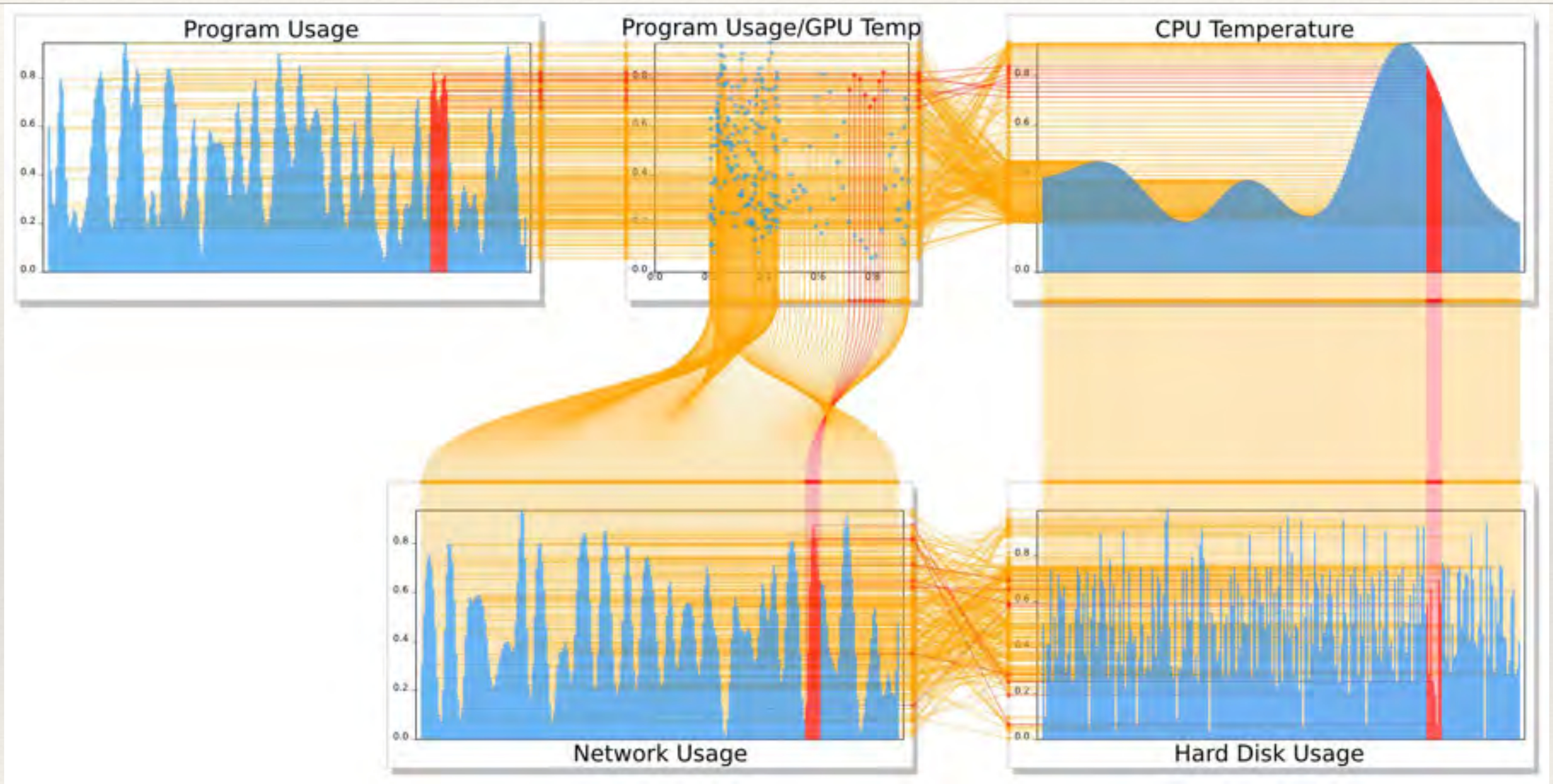
Elmqvist

Combining PCs & Sploms



Claessen & van Wijk 2011

Connected Charts



C. Viau, M. J. McGuffin 2012
http://profs.etsmtl.ca/mmcguffin/research/2012-viau-connectedCharts/viau-eurovis2012-connectedCharts.pdf
http://profs.etsmtl.ca/mmcguffin/research/

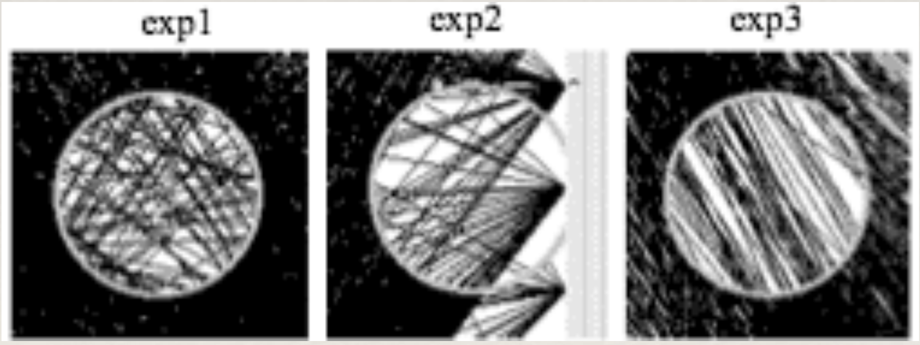
Data Reduction

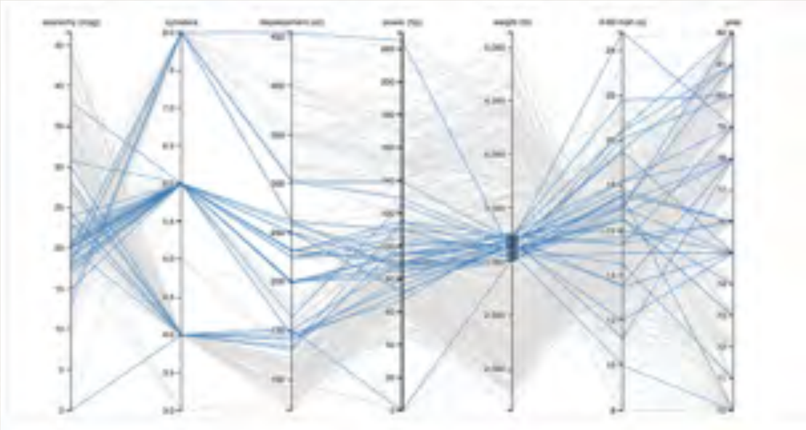
Reducing Rows

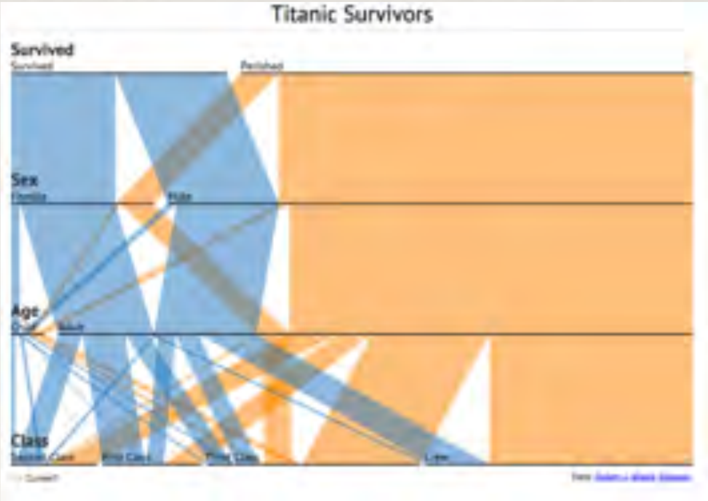
Sampling

Filtering

Clustering





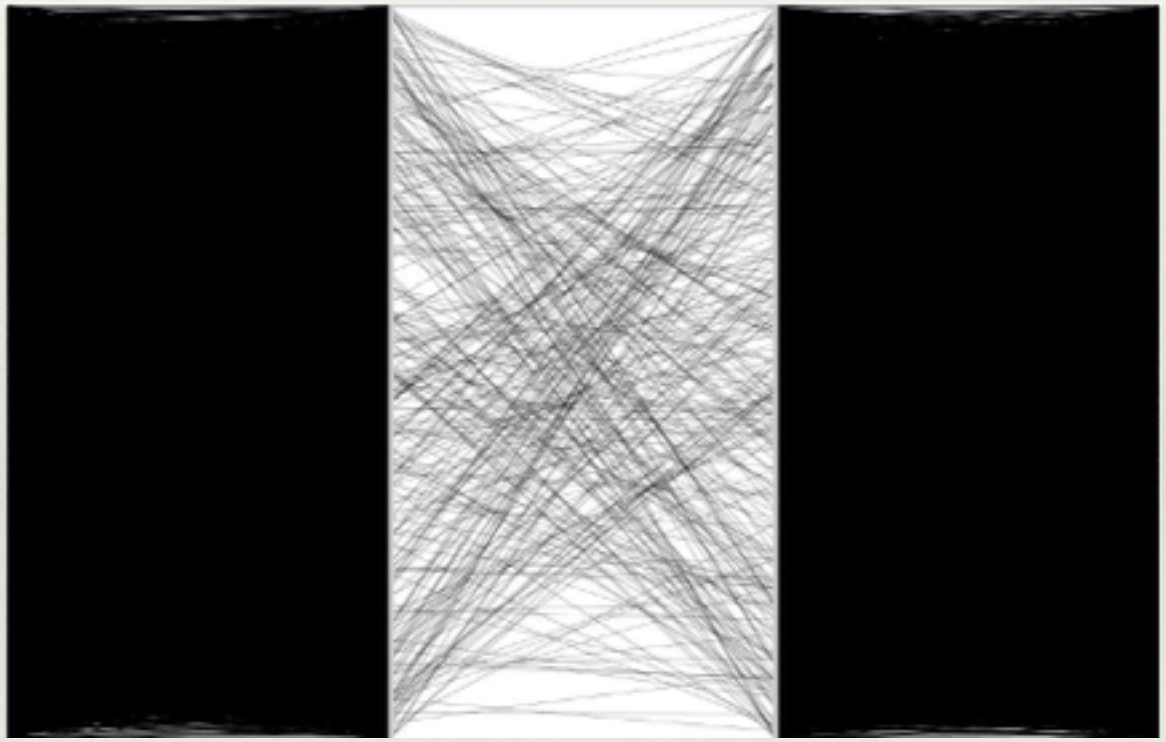


http://www.jasondavies.com/parallel-sets/

Later

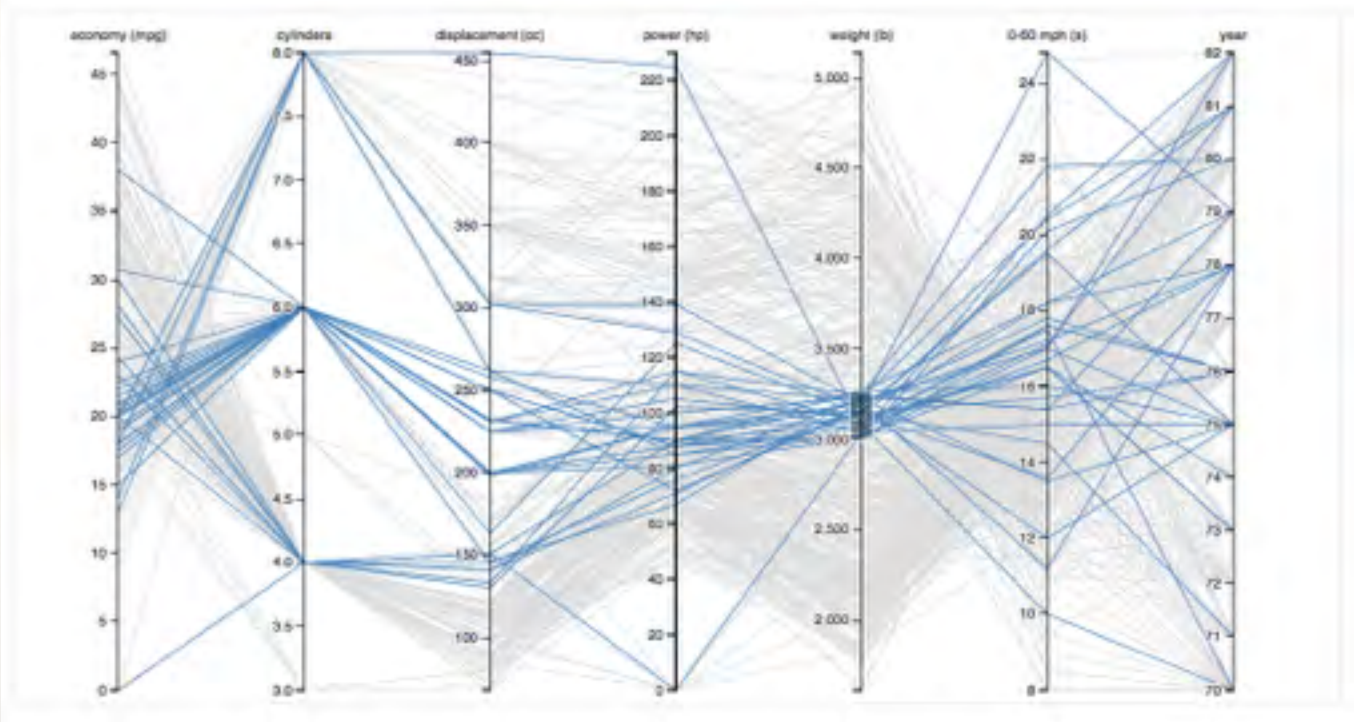
Sampling

- show (random) subset
- Efficient for large dataset
- For display purposes
- Outlier-preserving approaches



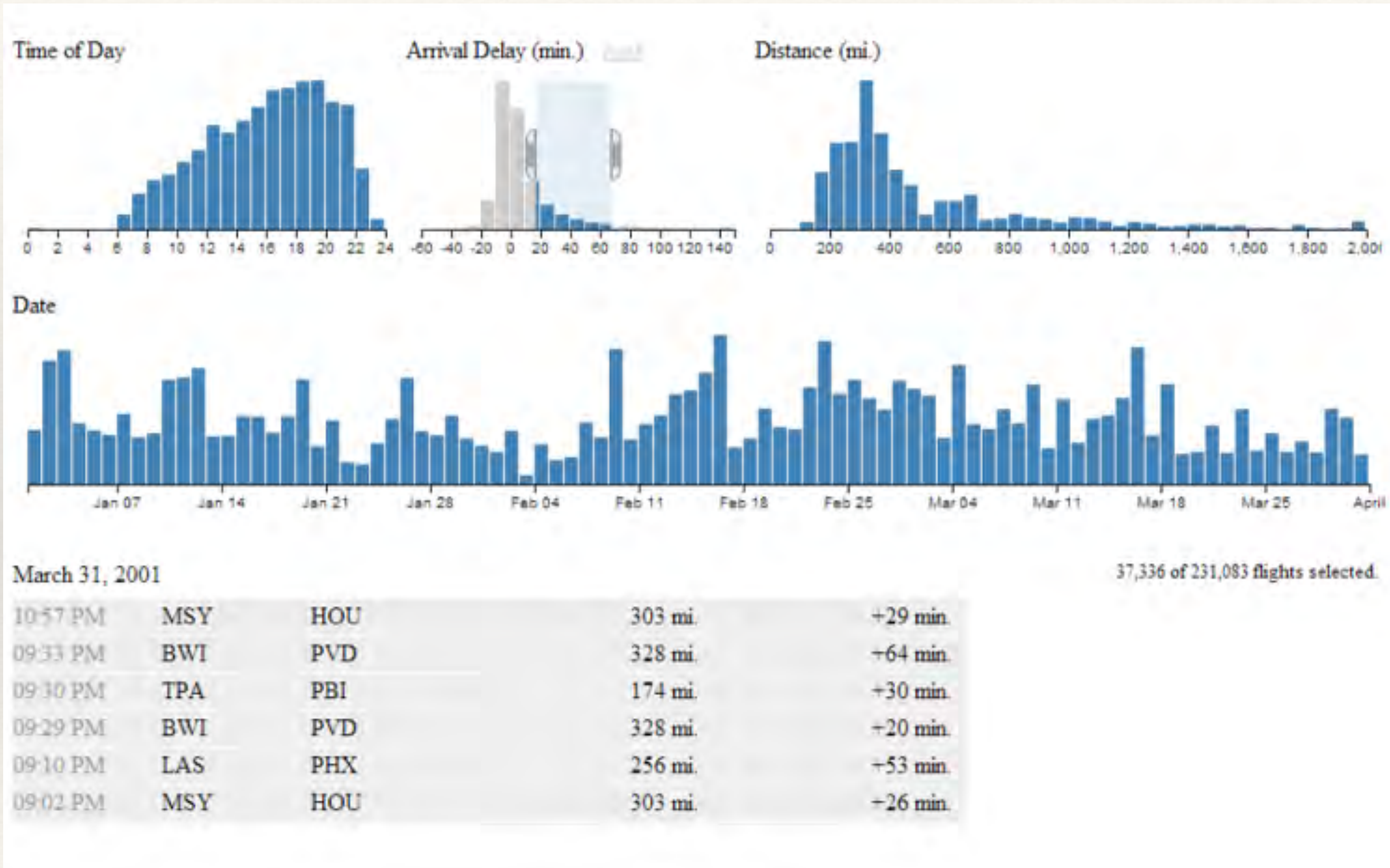
Ellis & Dix, 2006

Filtering



- Criteria to remove data
  - minimum variability
  - Range for dimension
  - consistency in replicates

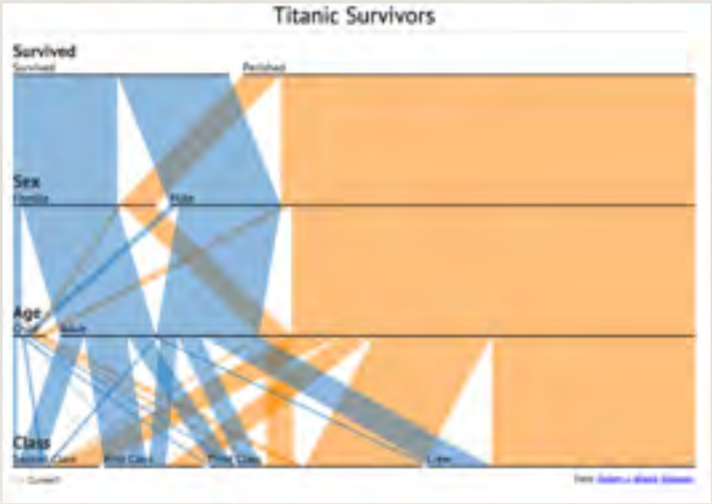
Filter Example



http://square.github.io/crossfilter/

Clustering

Clustering

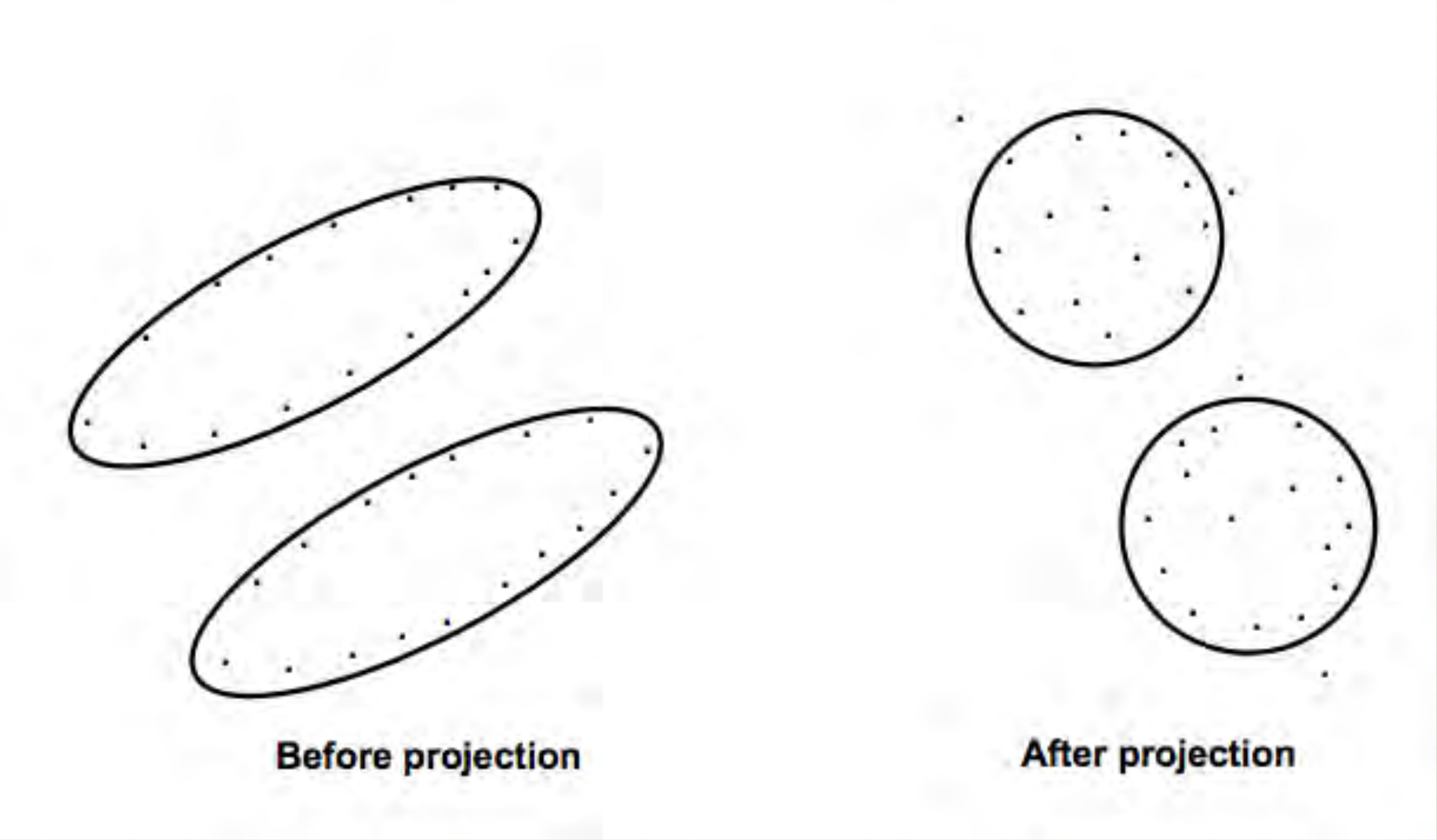


http://www.jasondavies.com/parallel-sets/

Later

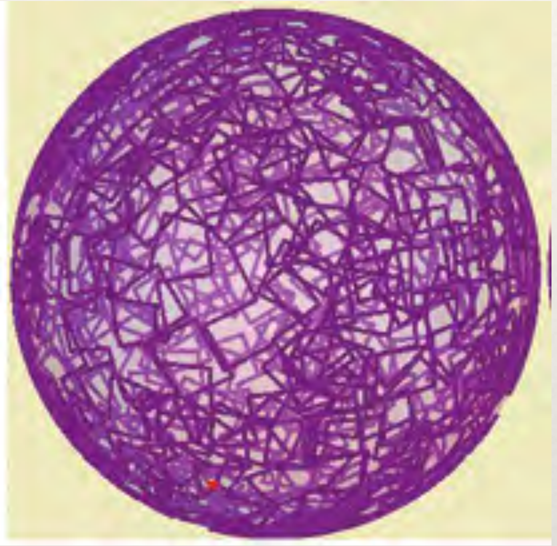
Dimension Reduction

Simple Random Projection

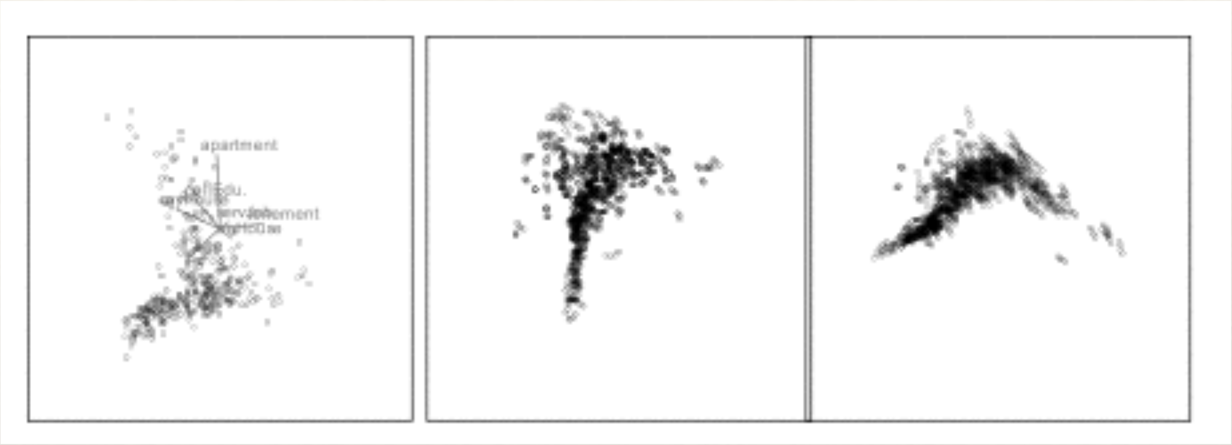


Experiments with Random Projection,
Sanjoy Dasgupta, 2000

Grand Tour



Travel a Dense Path
on 2-plane projections
in N-Space



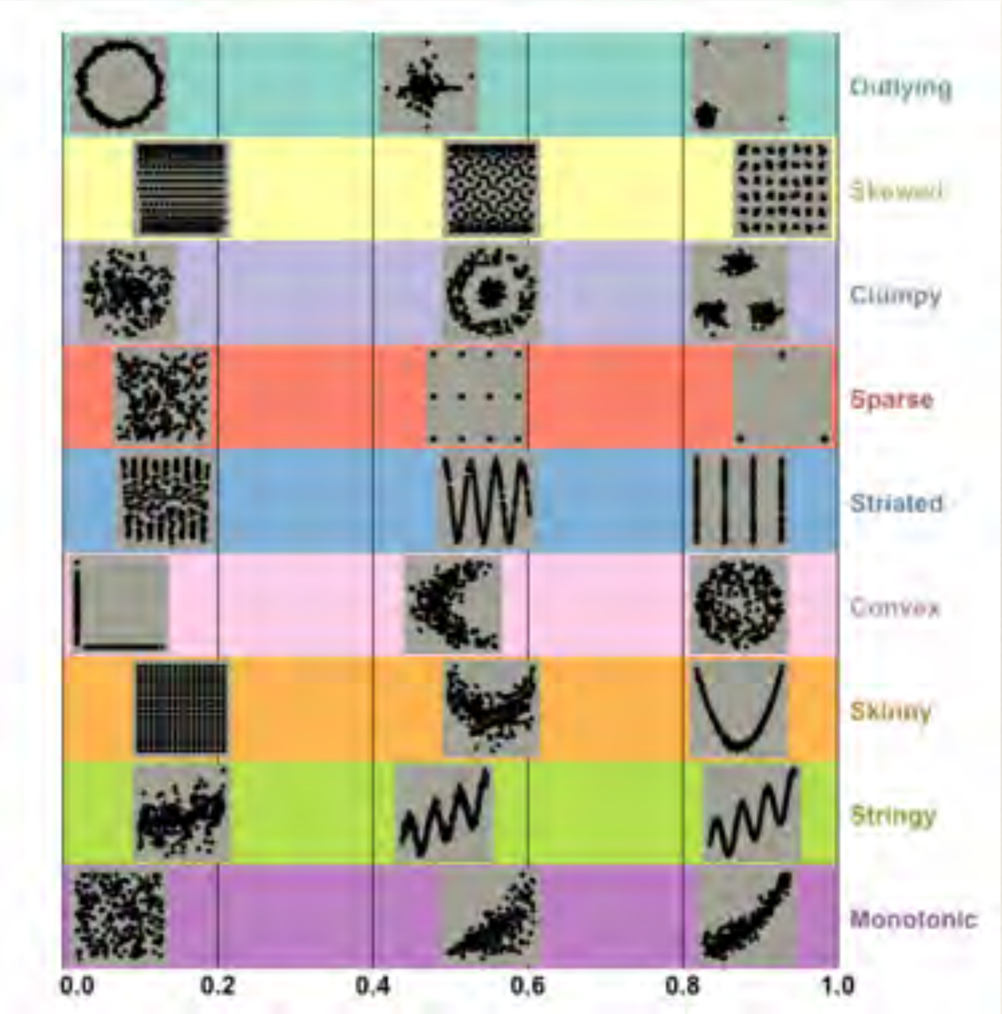
Look at all the scatterplots (Movie)

Interesting Projections

- Search for 2D scatter plots that maximize / minimize some attribute
- Clumpiness
- Variance
- Non-gausianness (entropy)

Scagnostics

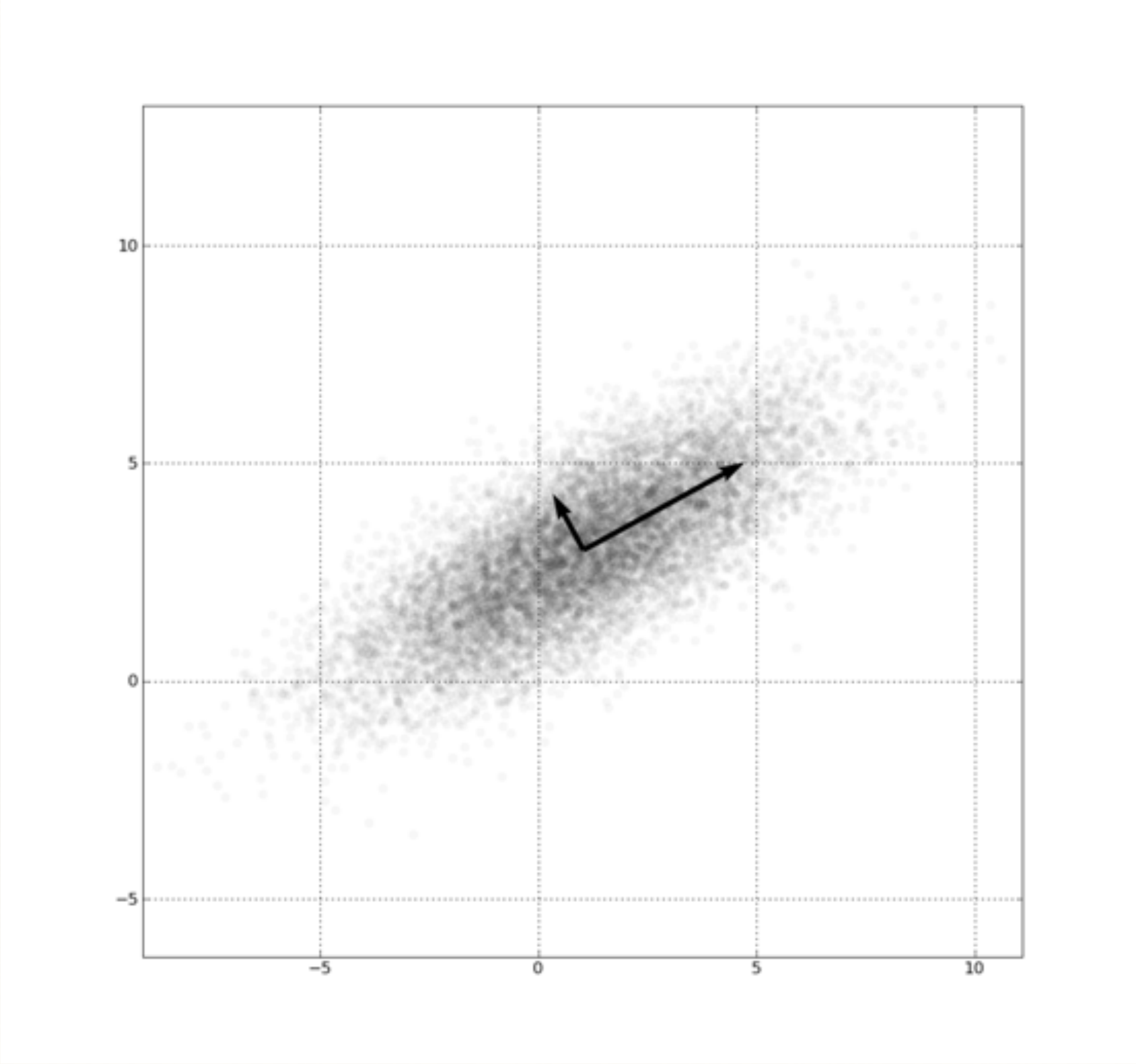
Look for projections which
look interesting



TN Dang, L Wilkinson - 2014

Searching for Projections

Principal Component Analysis (PCA)



1-D mean, stdev

Mean

$$\mu = E[x] = \frac{1}{n}\sum_{i=0}^{n} x_i$$

Variance = (Standard Deviation)^2

$$\sigma^2 = E[(x-\mu)^2] = \frac{1}{n}\sum_{i=0}^{n}(x_i - \mu)^2$$

Normal (1-D) distribution

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

N-D mean, (co-)variance

N-D Mean

$$\mu = E[\mathbf{x}] = \frac{1}{n}\sum_{i=0}^{n} \mathbf{x}_i$$

$$\mu = \{\mu_k\},\ \mu_k = \frac{1}{n}\sum_{i=0}^{n} x_{i,k}$$

N-D Covariance

$$\Sigma_{j,k} = \frac{1}{n}\sum_{i=0}^{n}(x_{i,j} - \mu_j)(x_{i,k} - \mu_k)$$

Matrix Versions

columns
dimensions / attributes

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{n,1} \\ \vdots & \ddots & \vdots \\ x_{1,m} & \cdots & x_{n,m} \end{bmatrix}$$

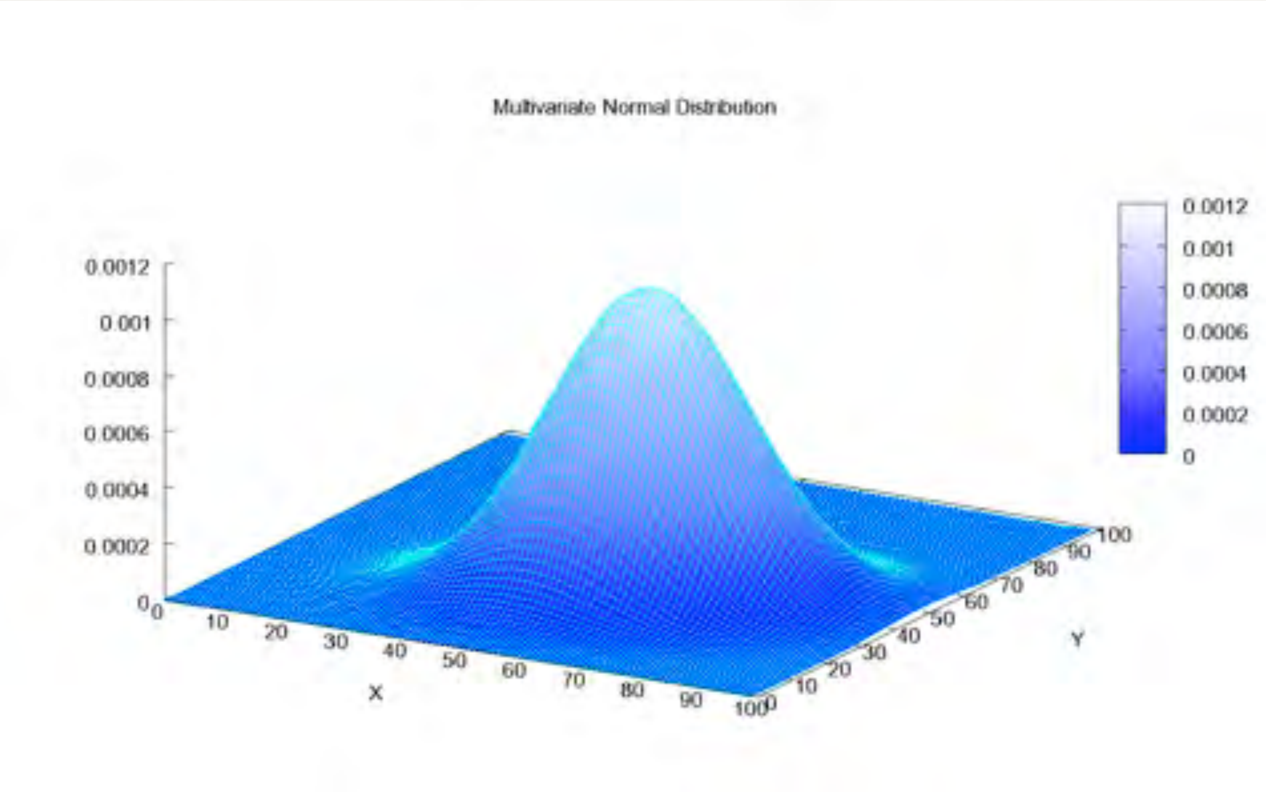
rows
observations

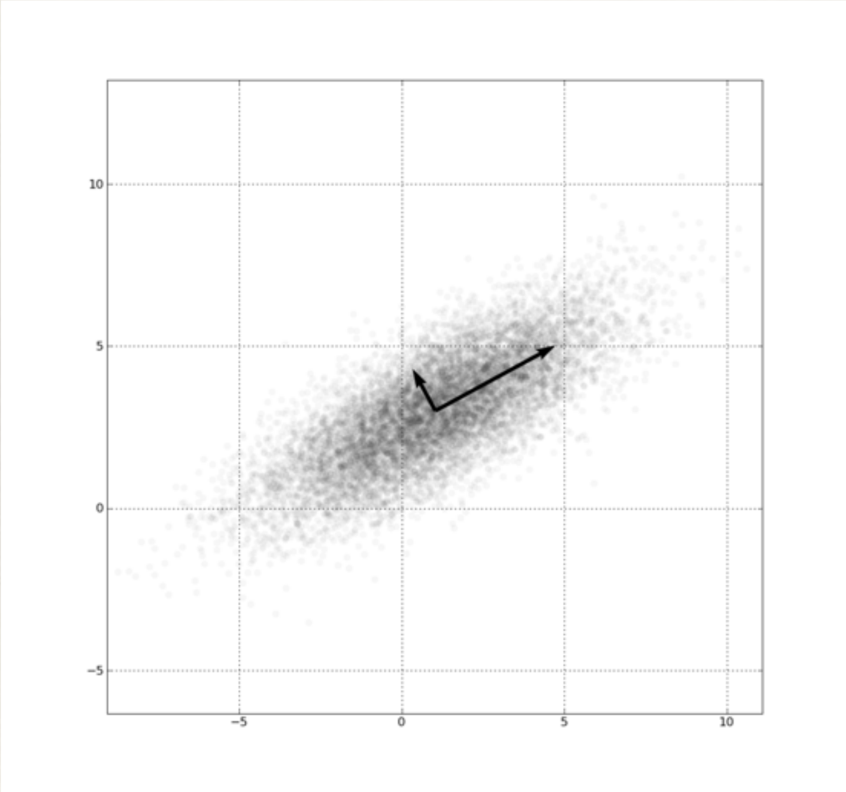
N-D Covariance

$$\Sigma = \check{X}^T \check{X}$$

N-D Normal Distribution

$$f_{\mathbf{x}}(x_1, \ldots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right),$$





Diagonalization

$$\Sigma$$

Symmetric matrix

$$\Sigma = V^T D V$$

Diagonalization (SVD)

$$D = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_n^2 \end{bmatrix}$$

Eigenvalues
Variances

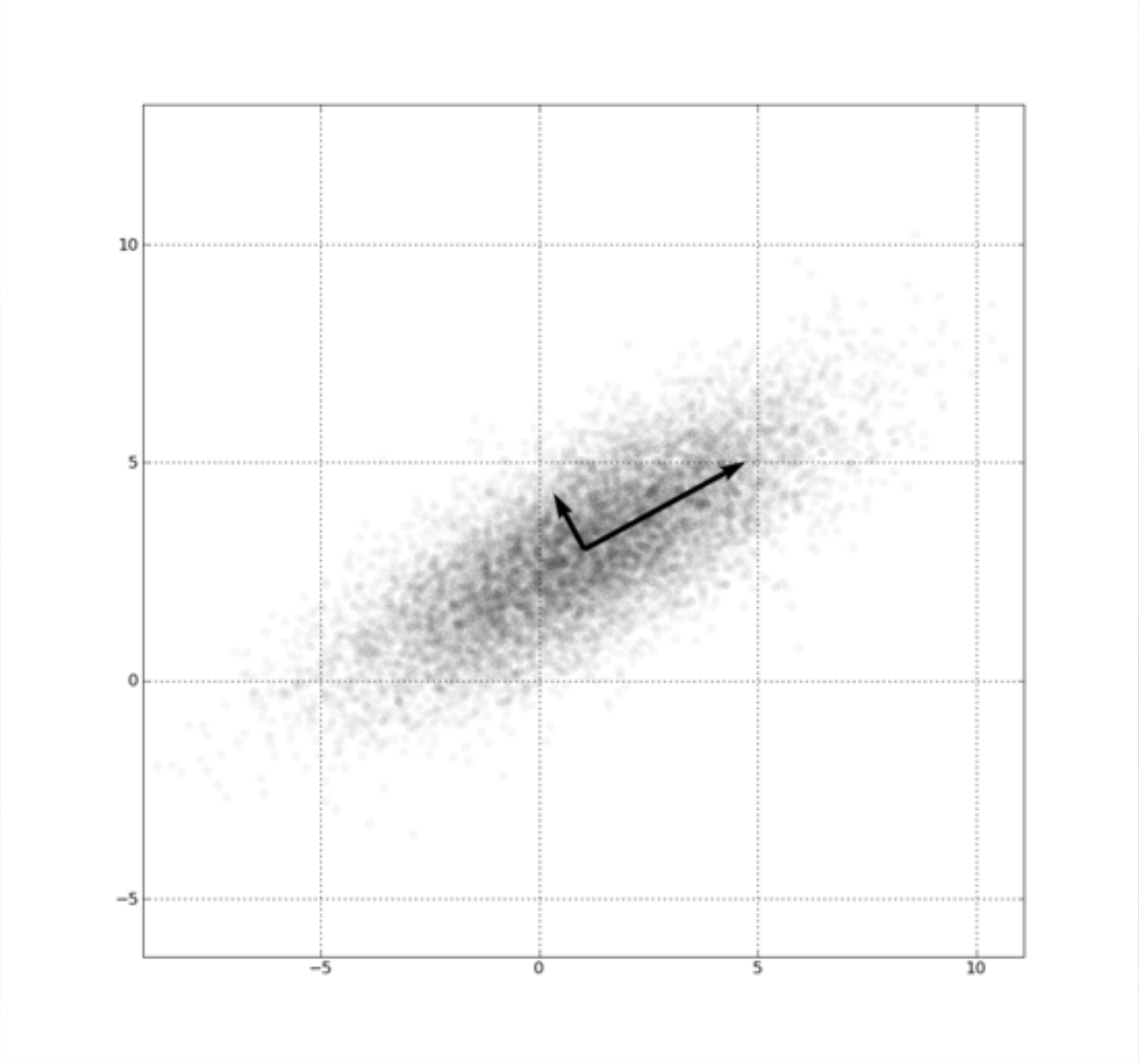
$$\lambda_1 > \lambda_2 \cdots > \lambda_n$$

$$V = \begin{bmatrix} v_{1,1} & \cdots & v_{n,1} \\ \vdots & \ddots & \vdots \\ v_{1,n} & \ldots & v_{n,n} \end{bmatrix}$$

First principal component

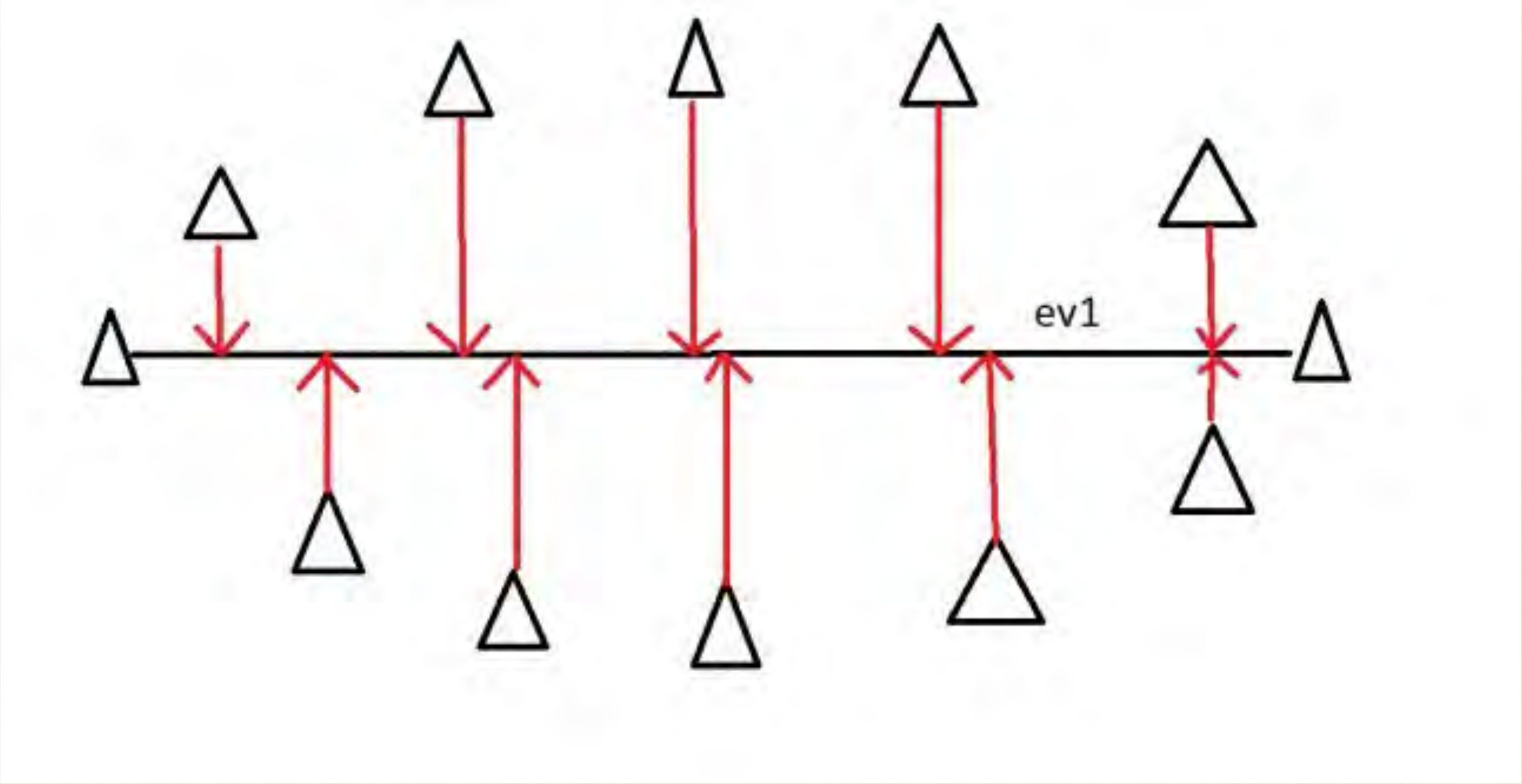
Eigenvectors

Eigenvectors = Principal Components

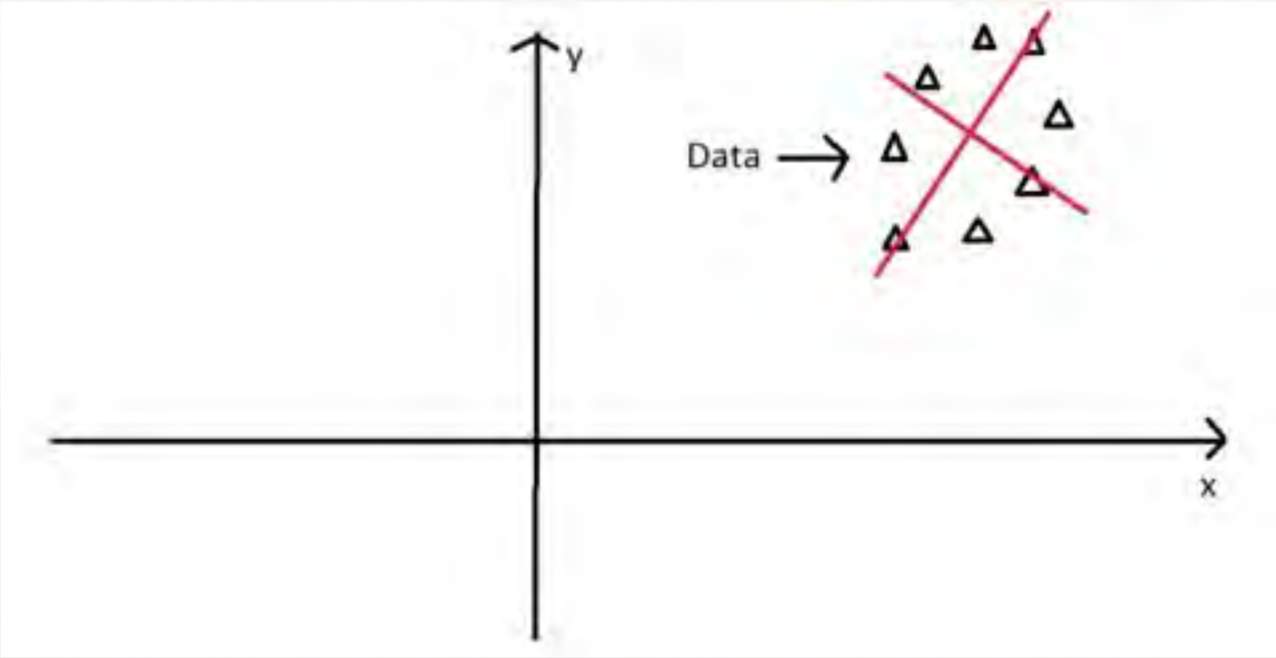


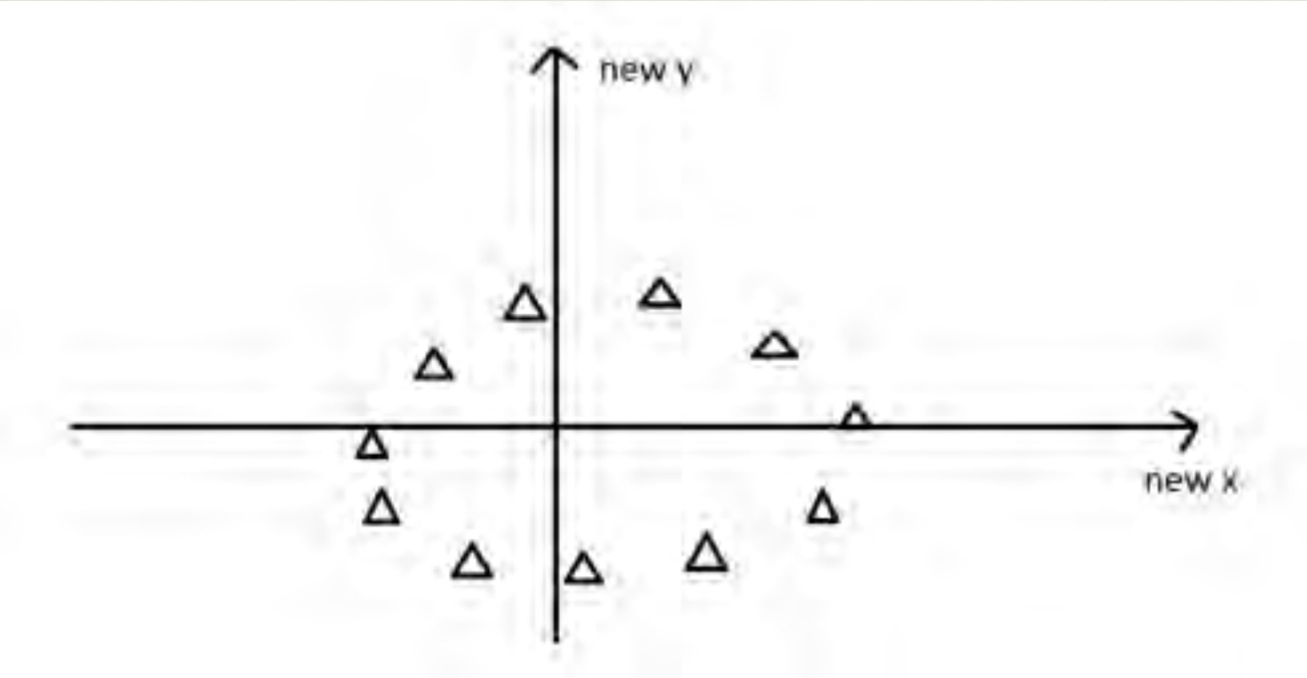
First principal component = direction of greatest variance

Least Squares



Data Oriented Coordinates

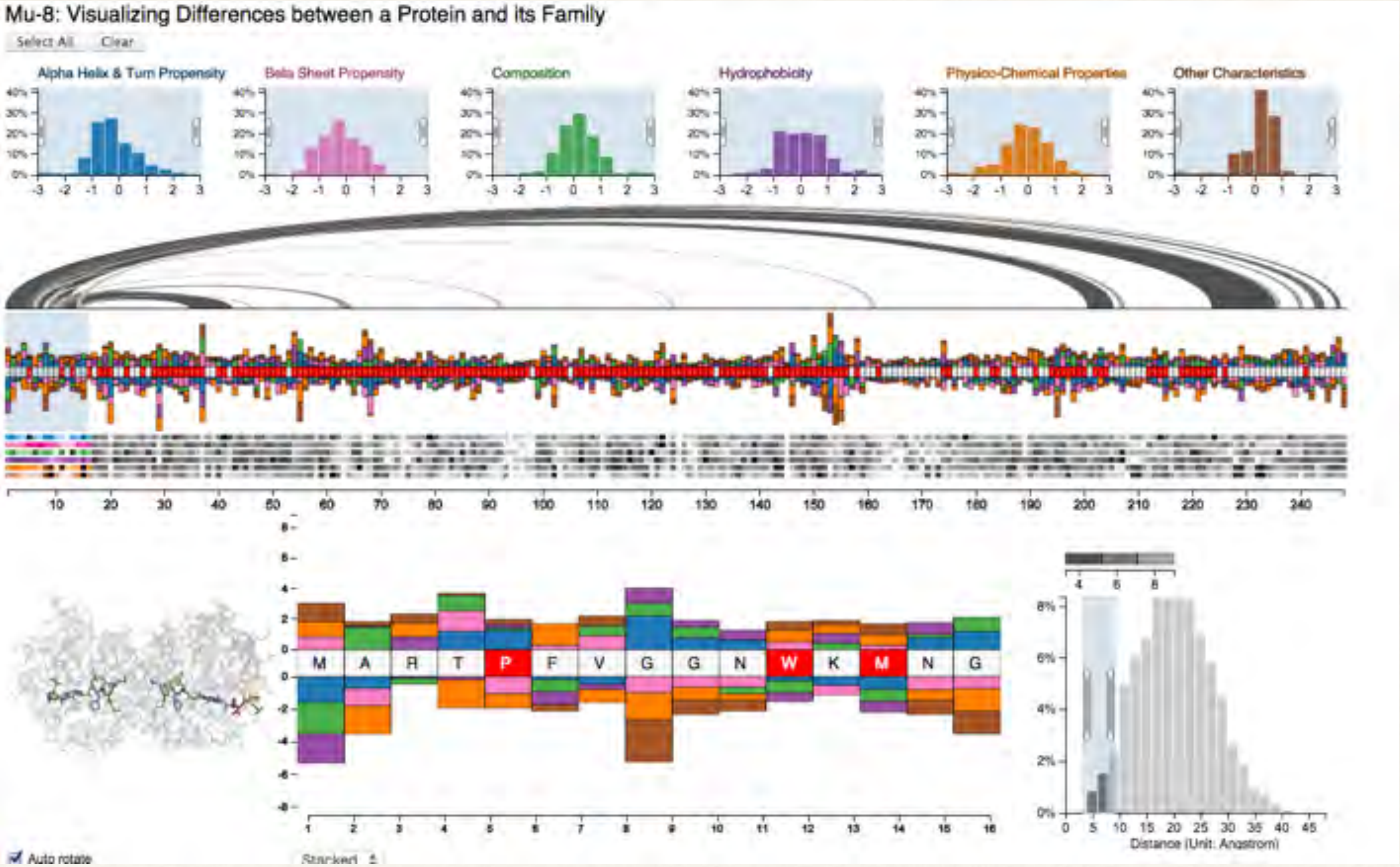




http://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-eigenvalues-and-dimension-reduction/

PCA in Vis Reduction

PCA from Harvard Cs171 student project

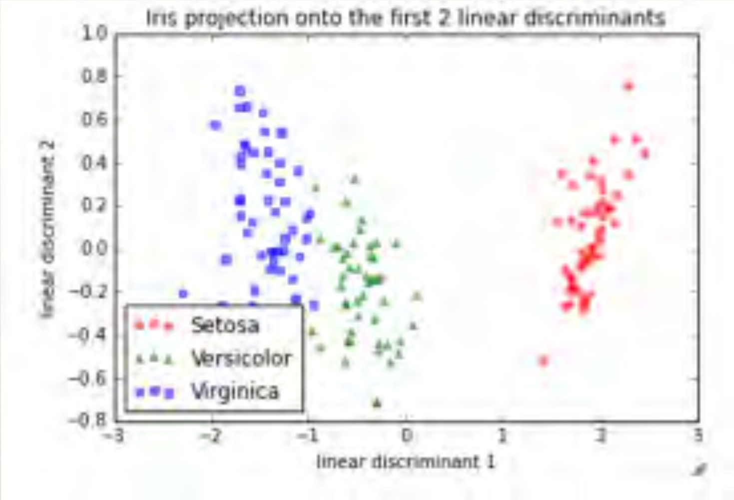


http://mu-8.com

Mercer and Pandian

Many other Linear Component Reductions

Linear Discriminant Analysis (LDA)



Separate classes

Latent Semantic Analysis (LSA)

Also

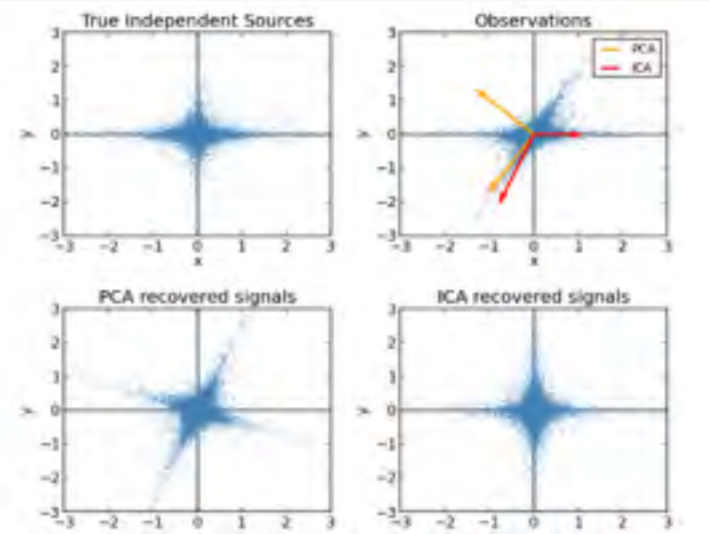
Factor Analysis

Dictionary Learning

Non-negative Matrix Factorization

Projection Pursuit

Independent Component Analysis (ICA)

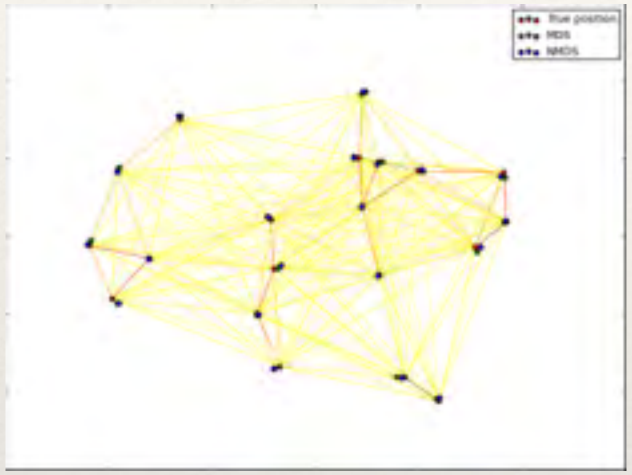


Make components independent

Maximize Entropy / Non-gausianness

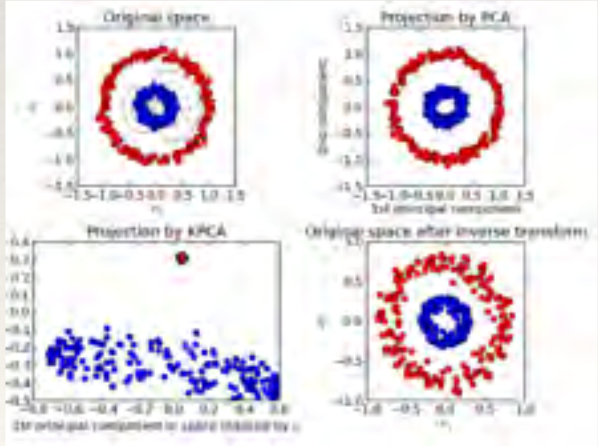
Non-Linear Dimension Reduction

Multi-Dimensional Scaling (MDS)

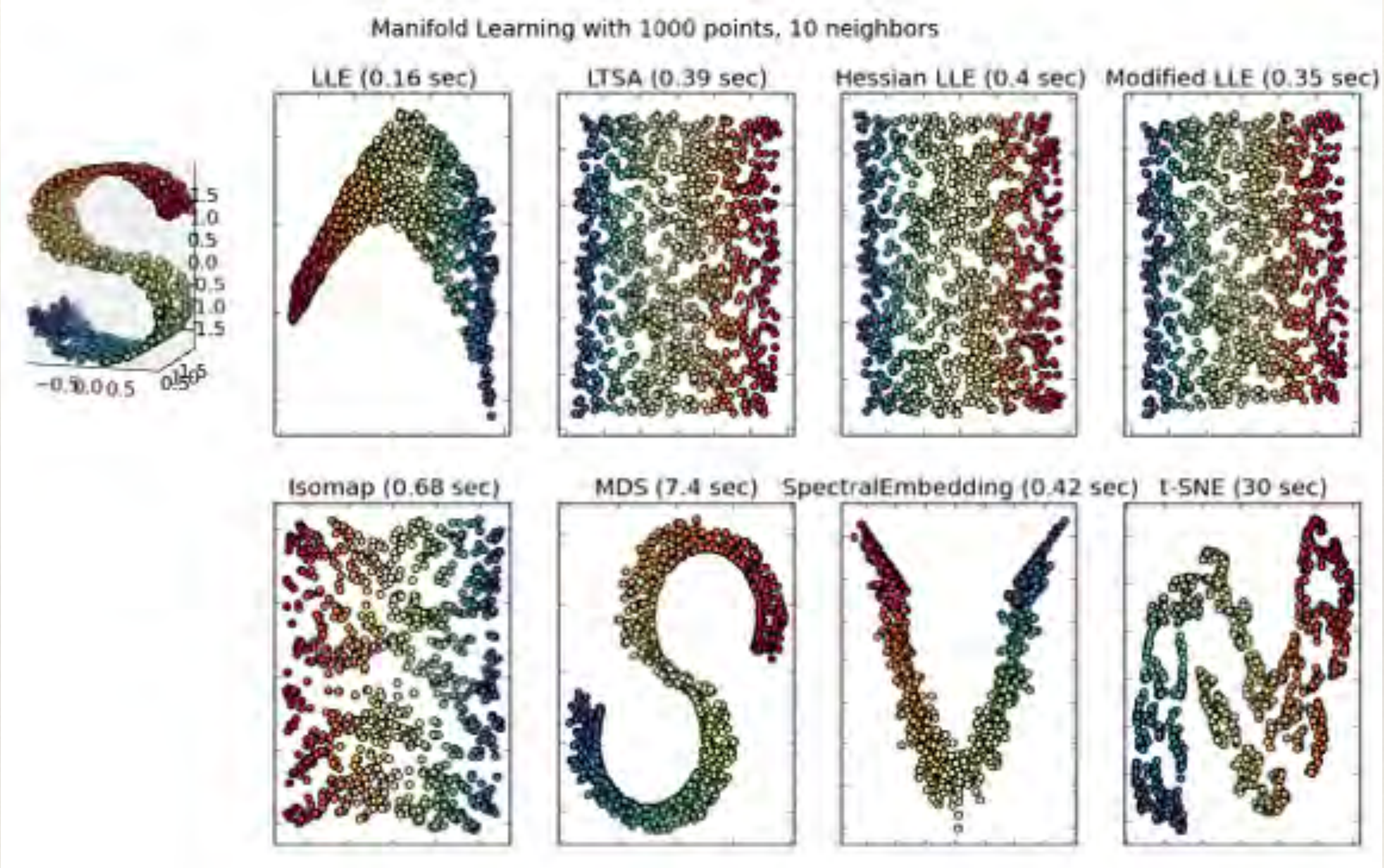


Preserve distances between observations

Kernel PCA



Manifold Learning

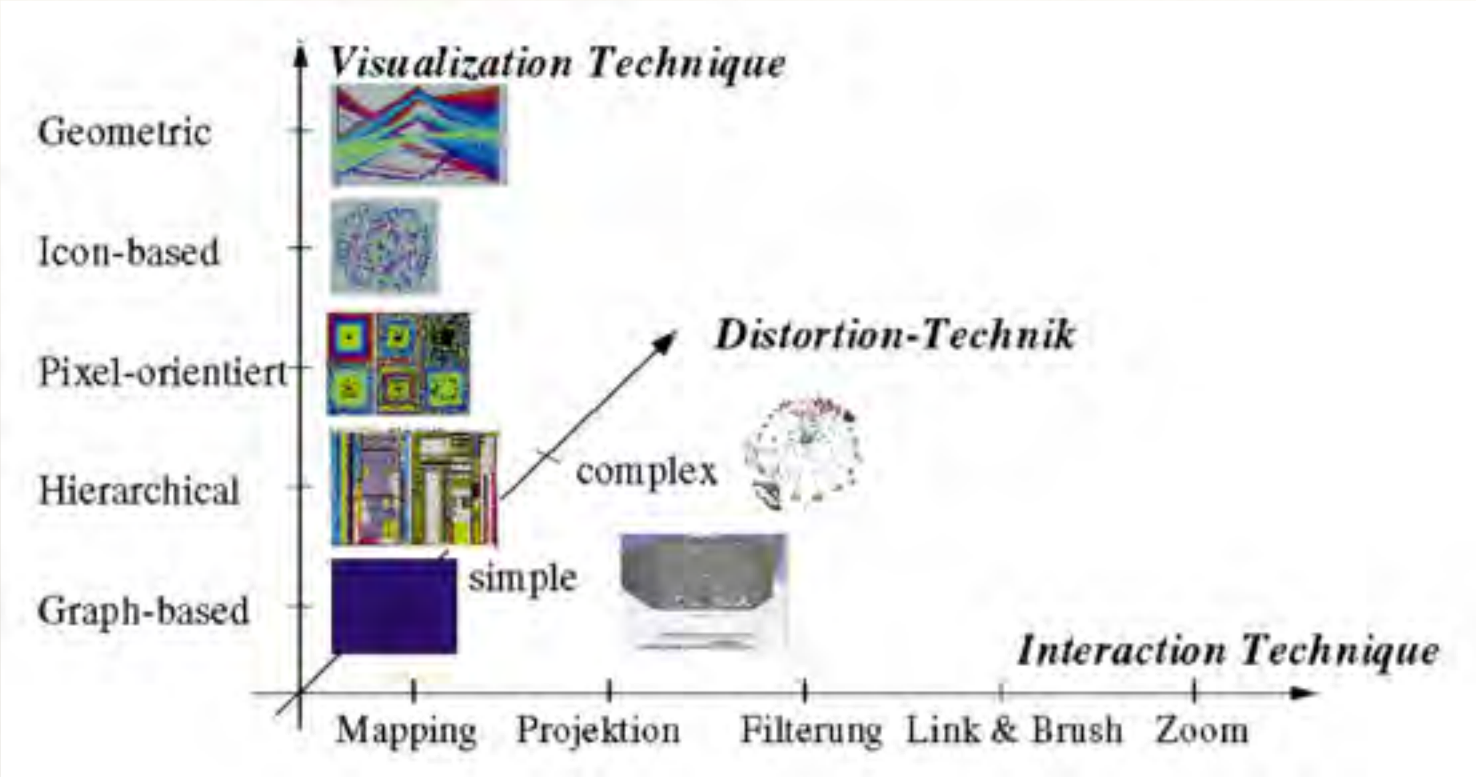


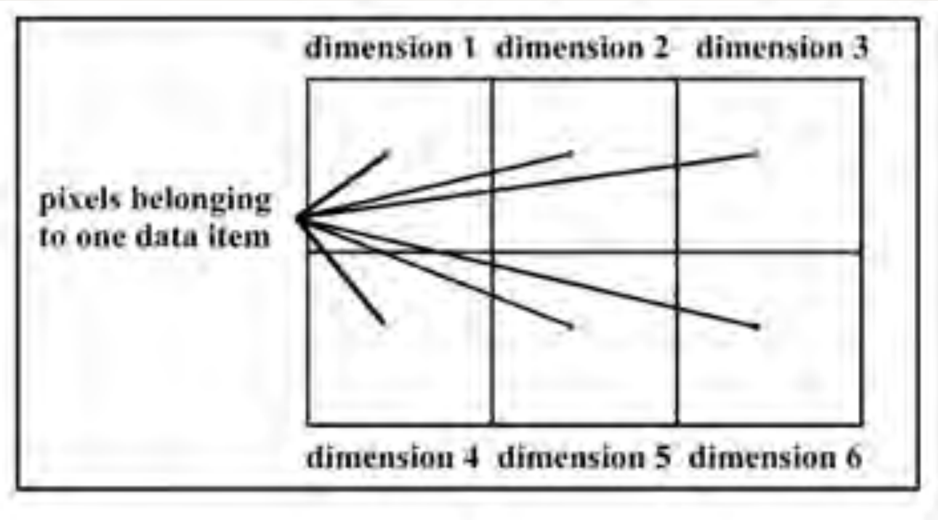
Non-Linear map to lower dimensional space

Pixel Based Methods

Daniel Keim

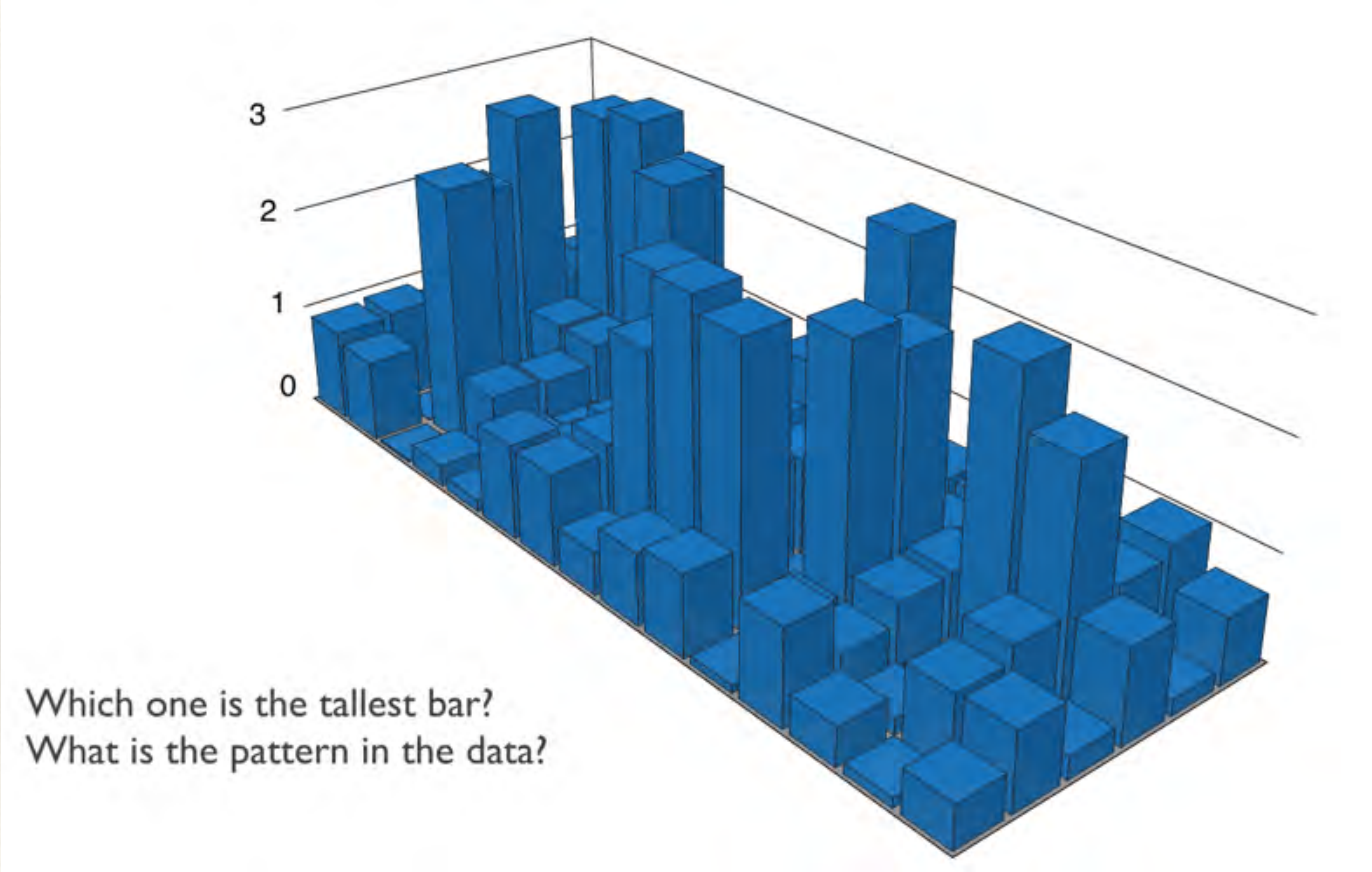
- Pixel Based Methods





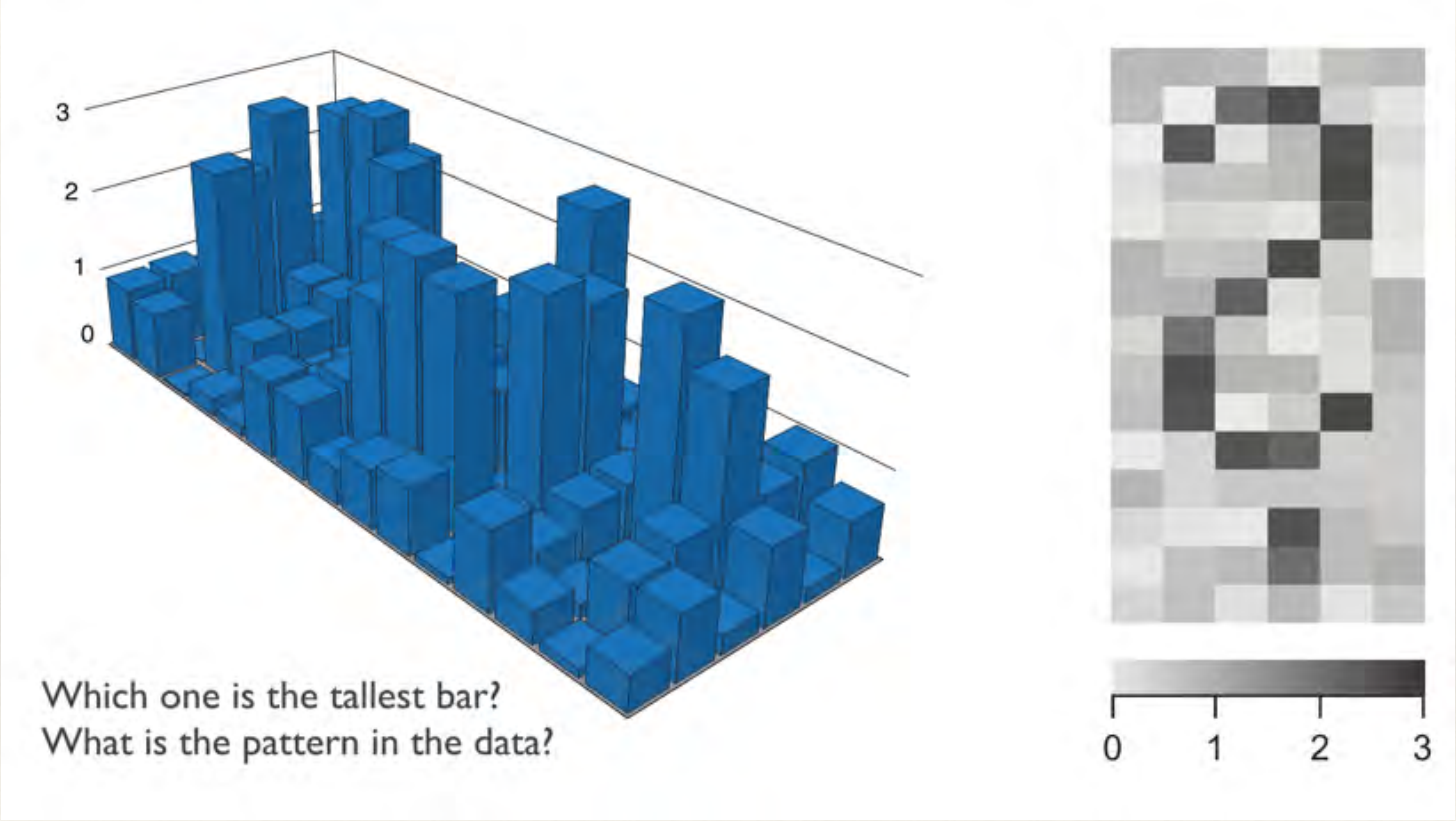
Designing Pixel-Oriented Visualization Techniques: Theory and Applications
Daniel A. Keim, 2000

3D Pitfall: Occlusion & Perspective



Gehlenborg and Wong, Nature Methods, 2012

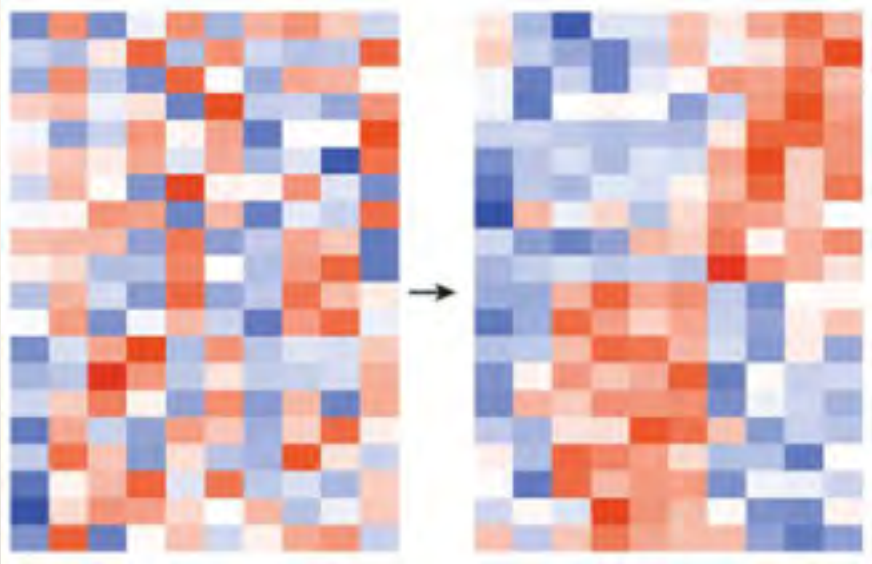
3D Pitfall: Occlusion & Perspective

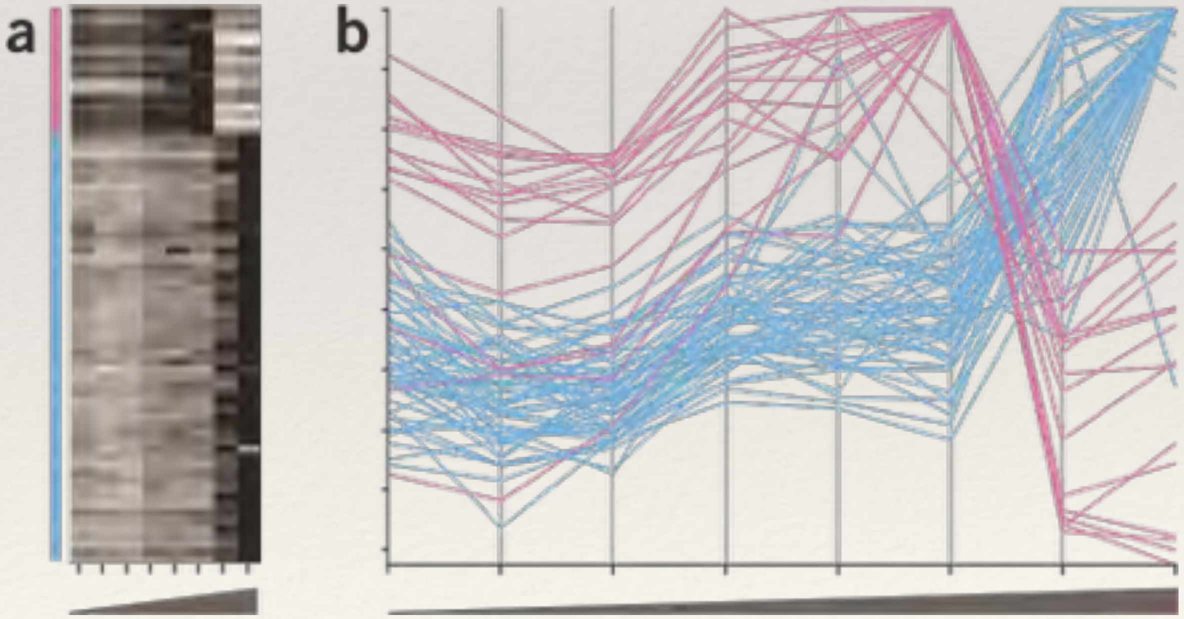


Gehlenborg and Wong, Nature Methods, 2012

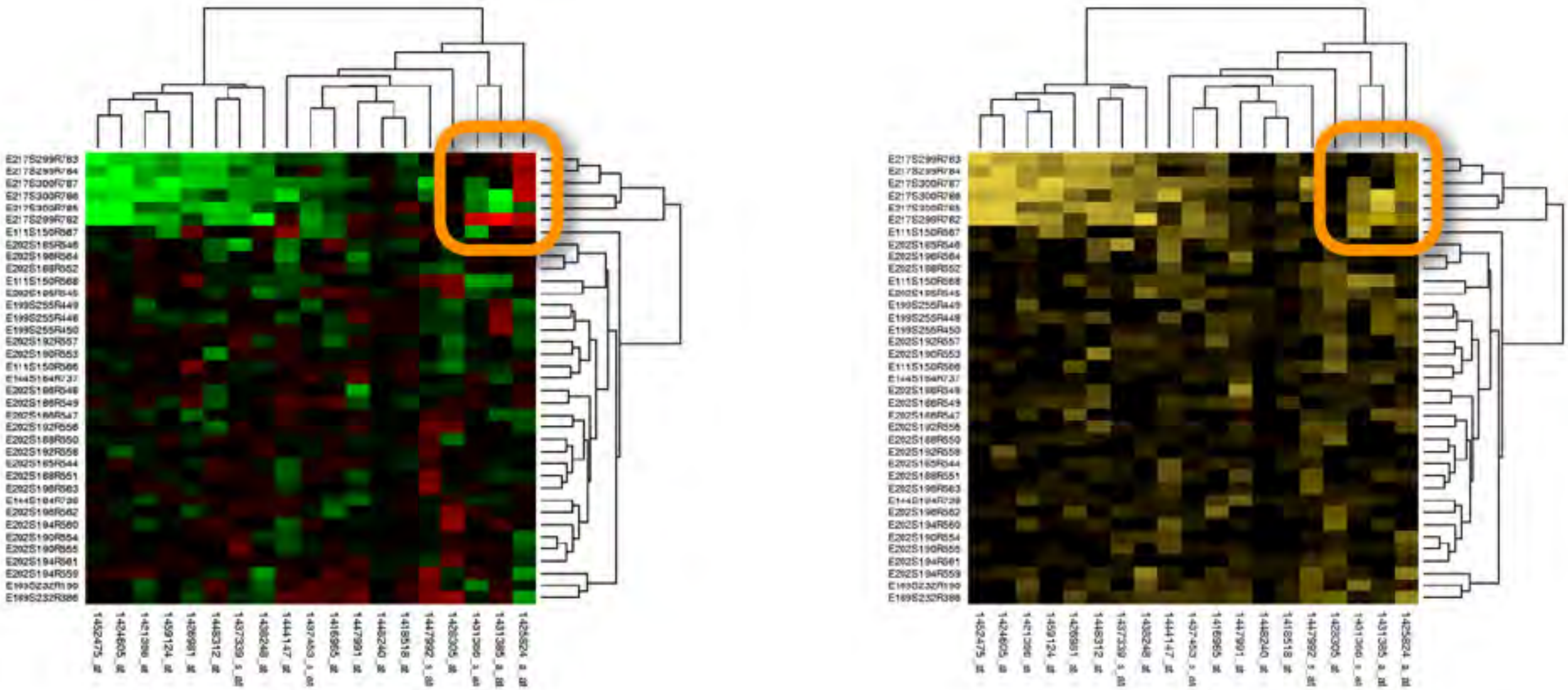
Pixel Based Displays

- Each cell is a "pixel", value encoded in color / value
- Meaning derived from ordering
- If no ordering inherent, clustering is used
- Scalable –1 px per item
- Good for homogeneous data
  - same scale & type





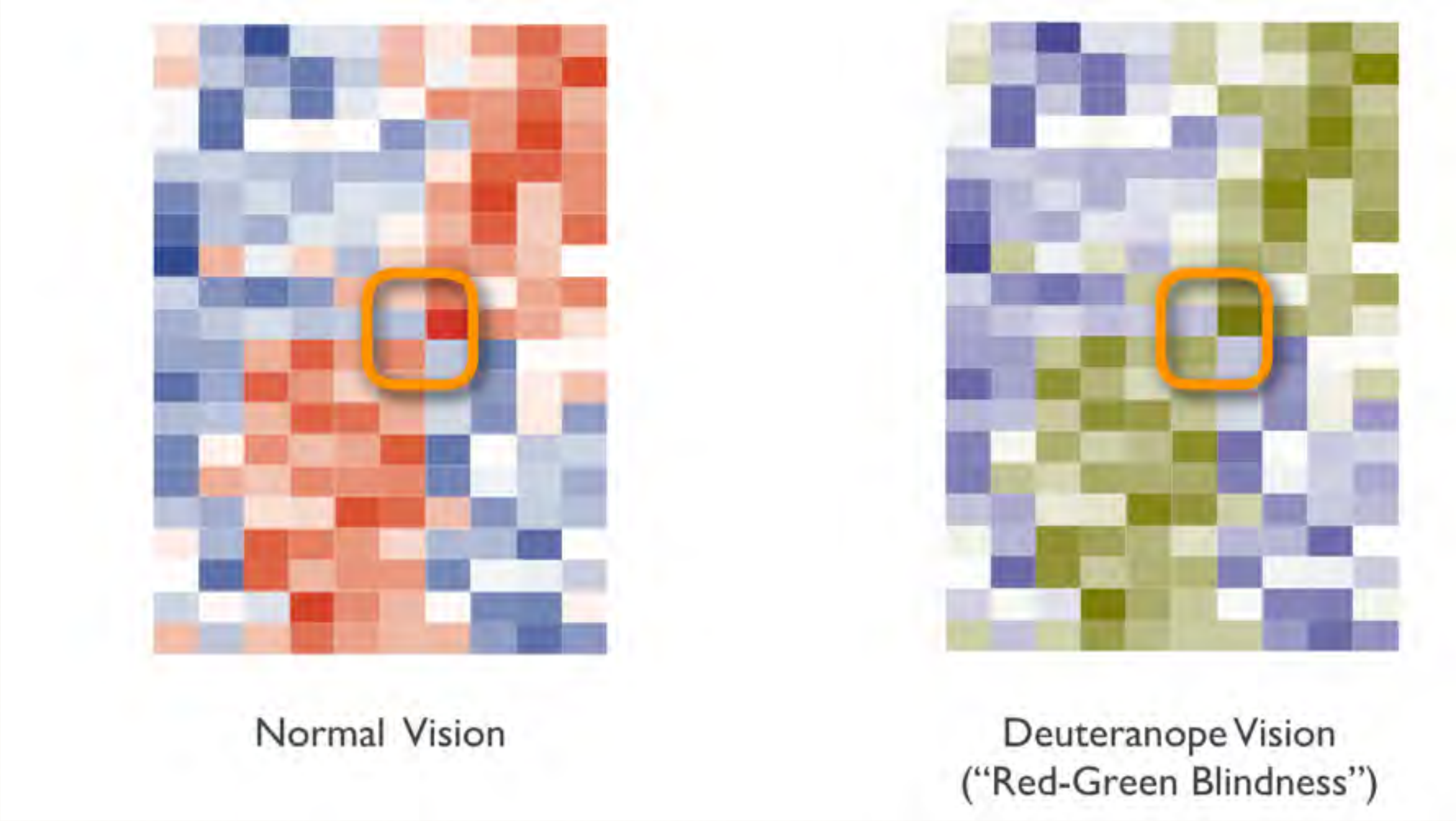
Bad Color Mapping



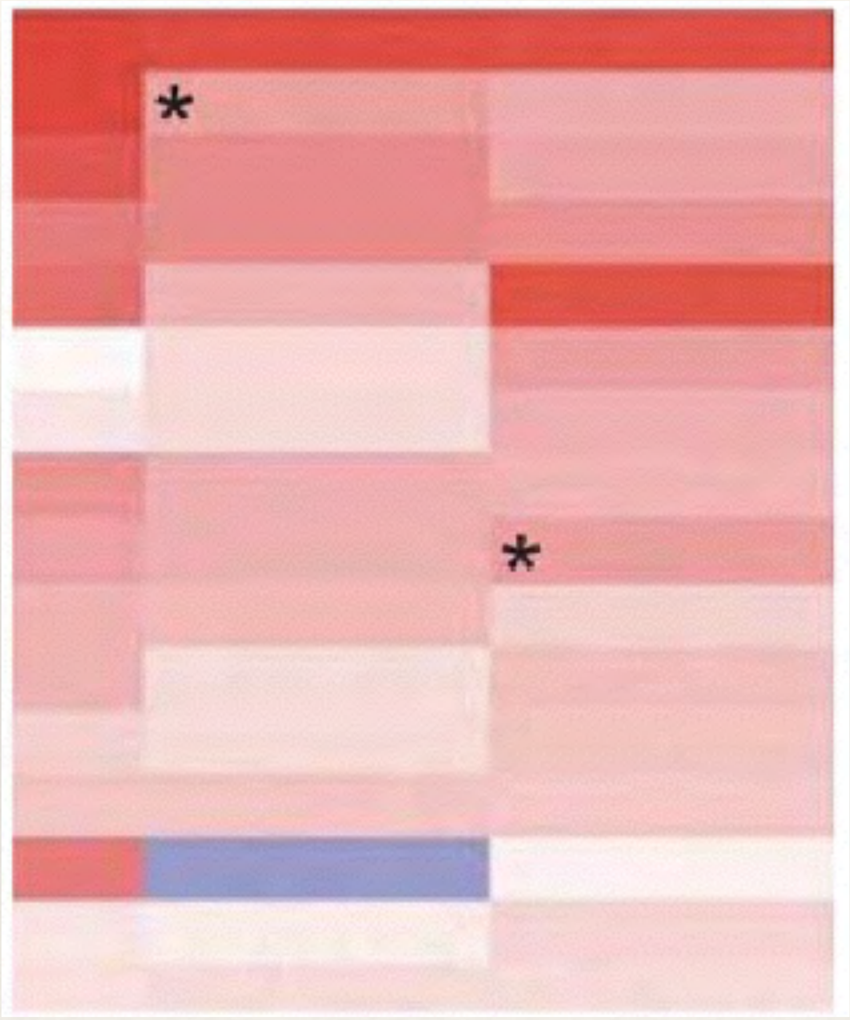
Normal Vision

Deuteranope Vision
("Red-Green Blindness")

Good Color Mapping



Color is relative!



Machine Learning

- Supervised Learning
  - A.K.A. Classification
  - Known labels (subset of rows)
  - Algorithms: label unlabeled rows
- Unsupervised Learning
  - A.K.A Clustering
  - Algorithm: label based on similarity
- Semi-Supervised Learning
  - Do both

Clustering

Partition

Hierarchical

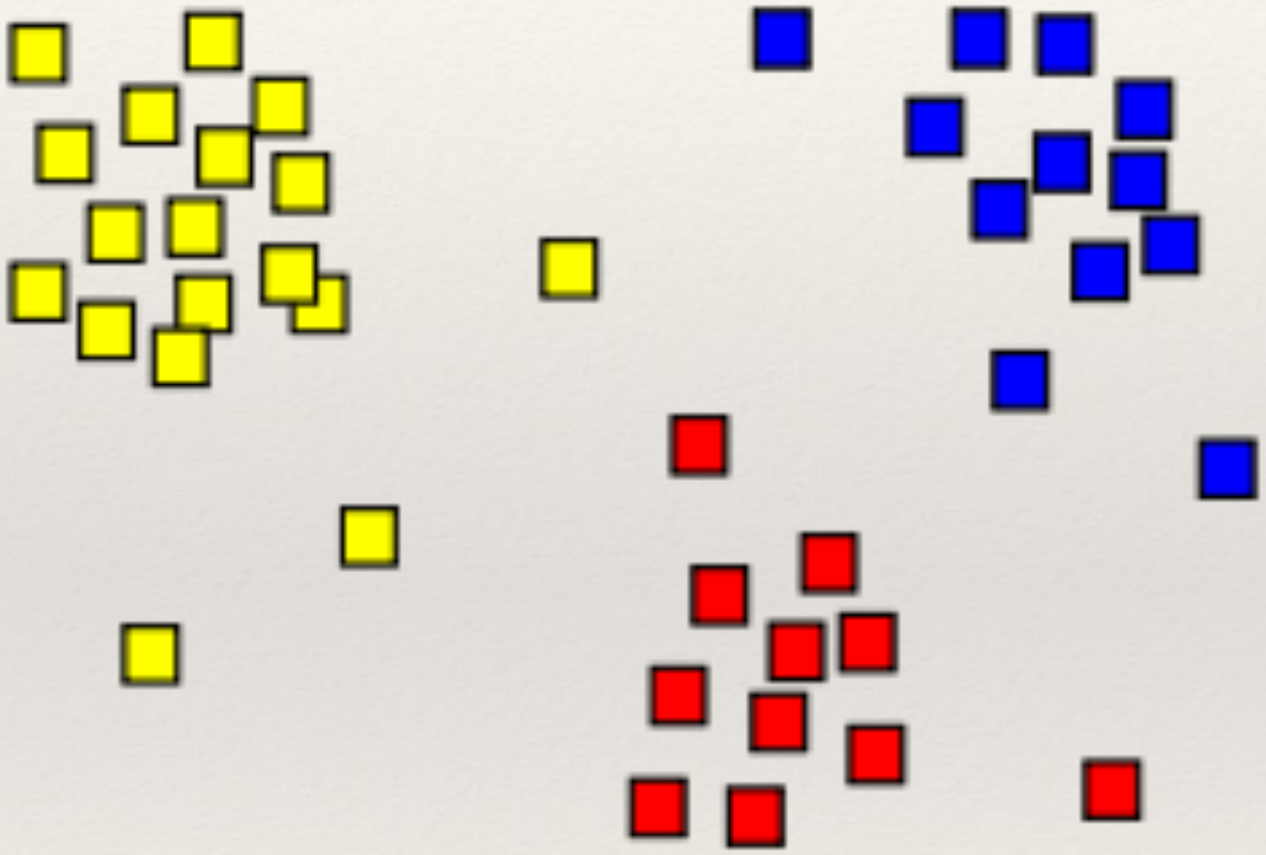
Bi-Partite

Fuzzy

Machine Learning

- Supervised Learning
  - A.K.A. Classification
  - Known labels (subset of rows)
  - Algorithms: label unlabeled rows
- Unsupervised Learning
  - A.K.A Clustering
  - Algorithm: label based on similarity
- Semi-Supervised Learning
  - Do both

Partition Clustering



Each Point in a Unique Class

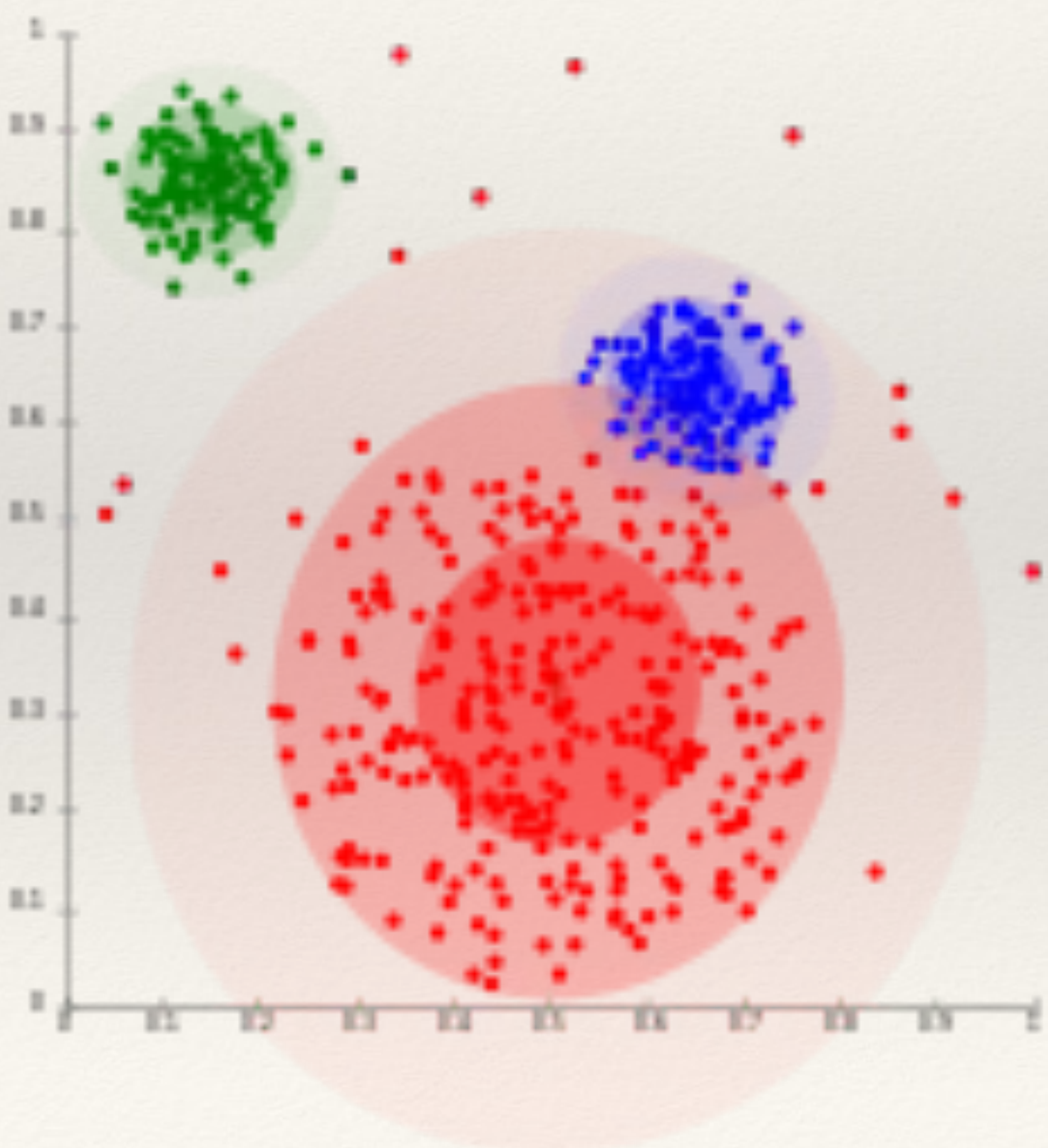
Centroid Based Clustering

Example Algorithm: K-Means Clustering

Partition

Distribution Based Clustering

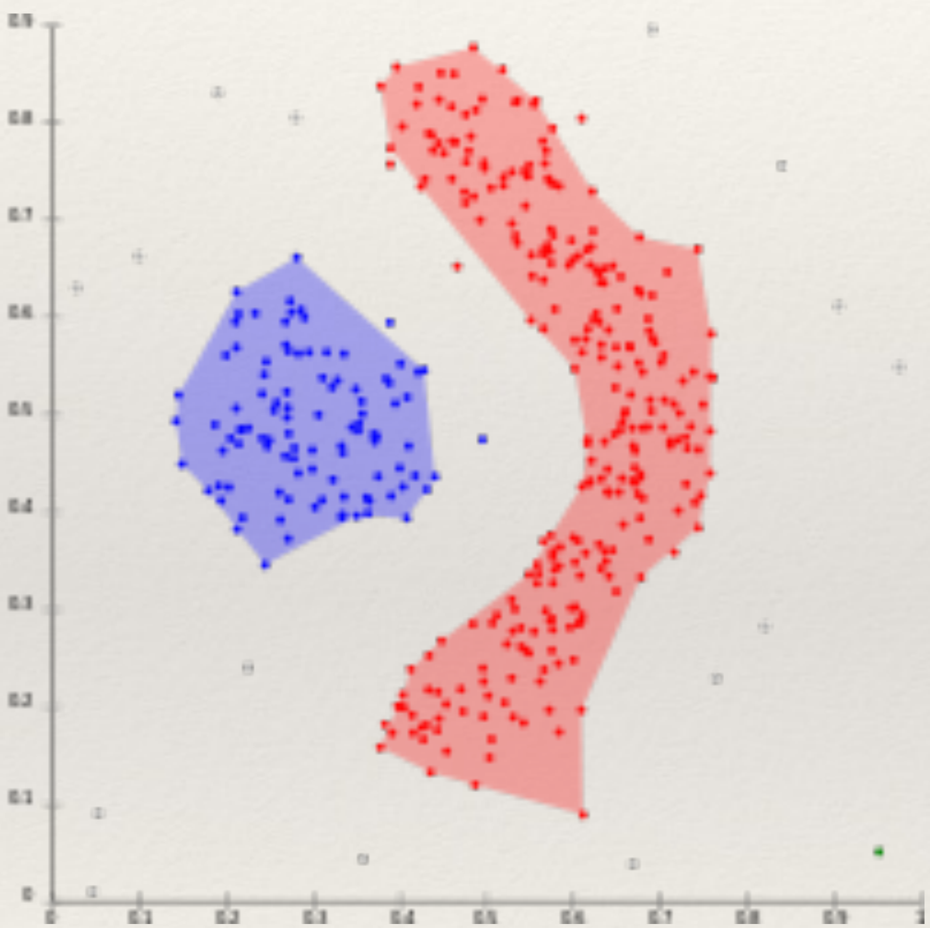
Expectation Maximization (EM)



Partition

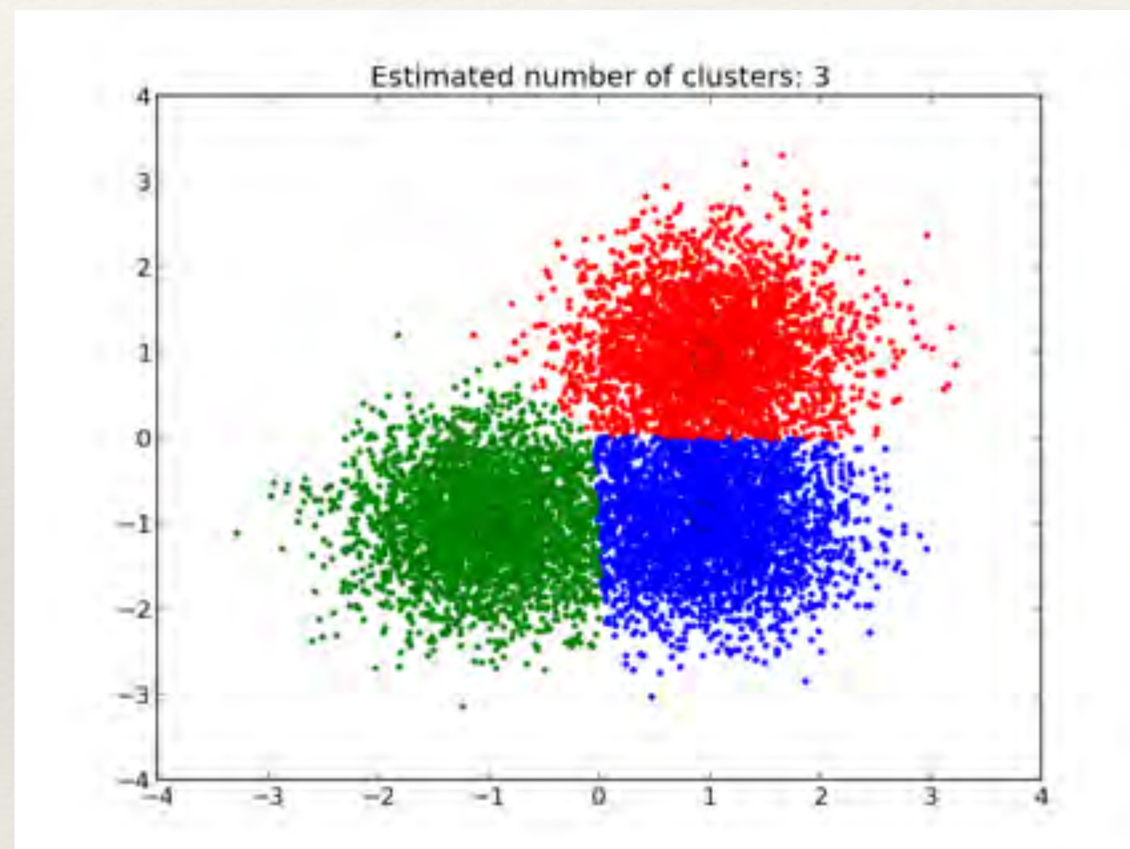
Density Based Clustering

DBSCAN



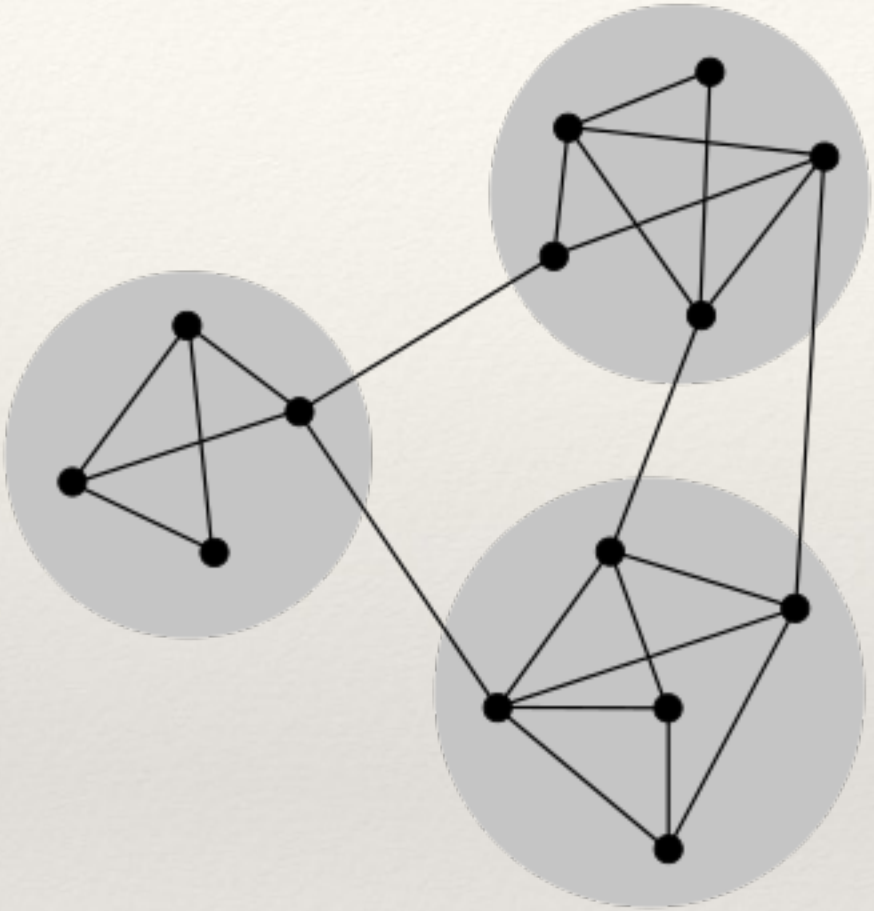
Group areas with Certain Density
OPTICS (more general)

Mean Shift



Partition

Simple Graph Cutting Methods

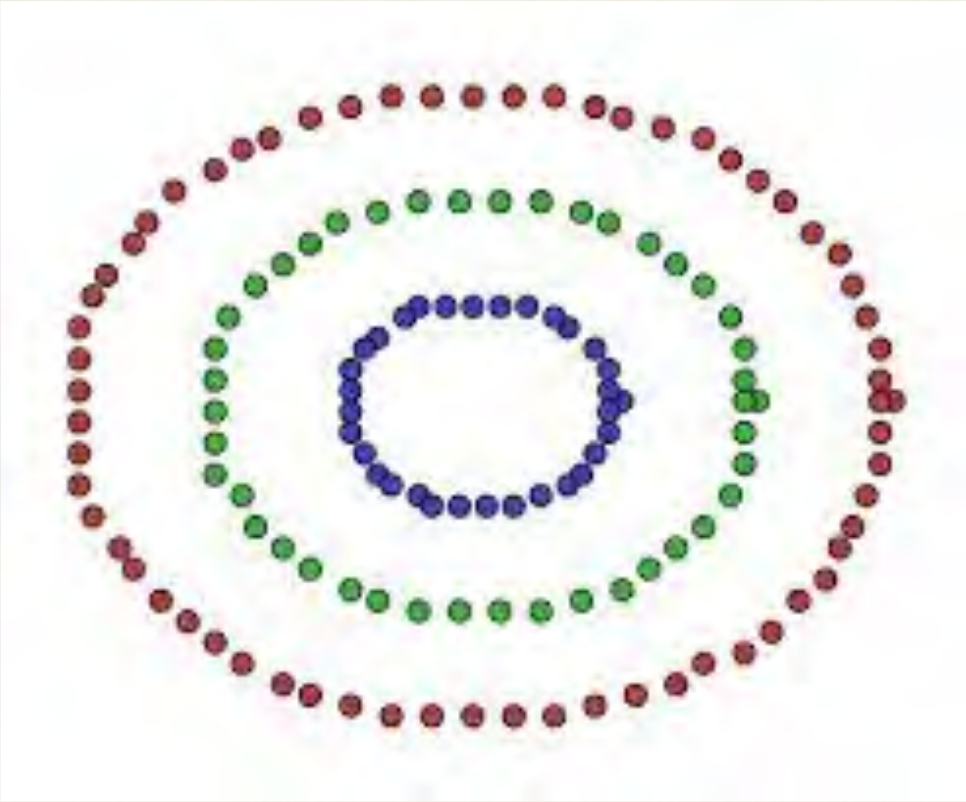


(1) Similarity Score
(2) Pick Edge Threshold
(3) Cut (only connect stronger edges)
(4) Compute Connected components

Partition

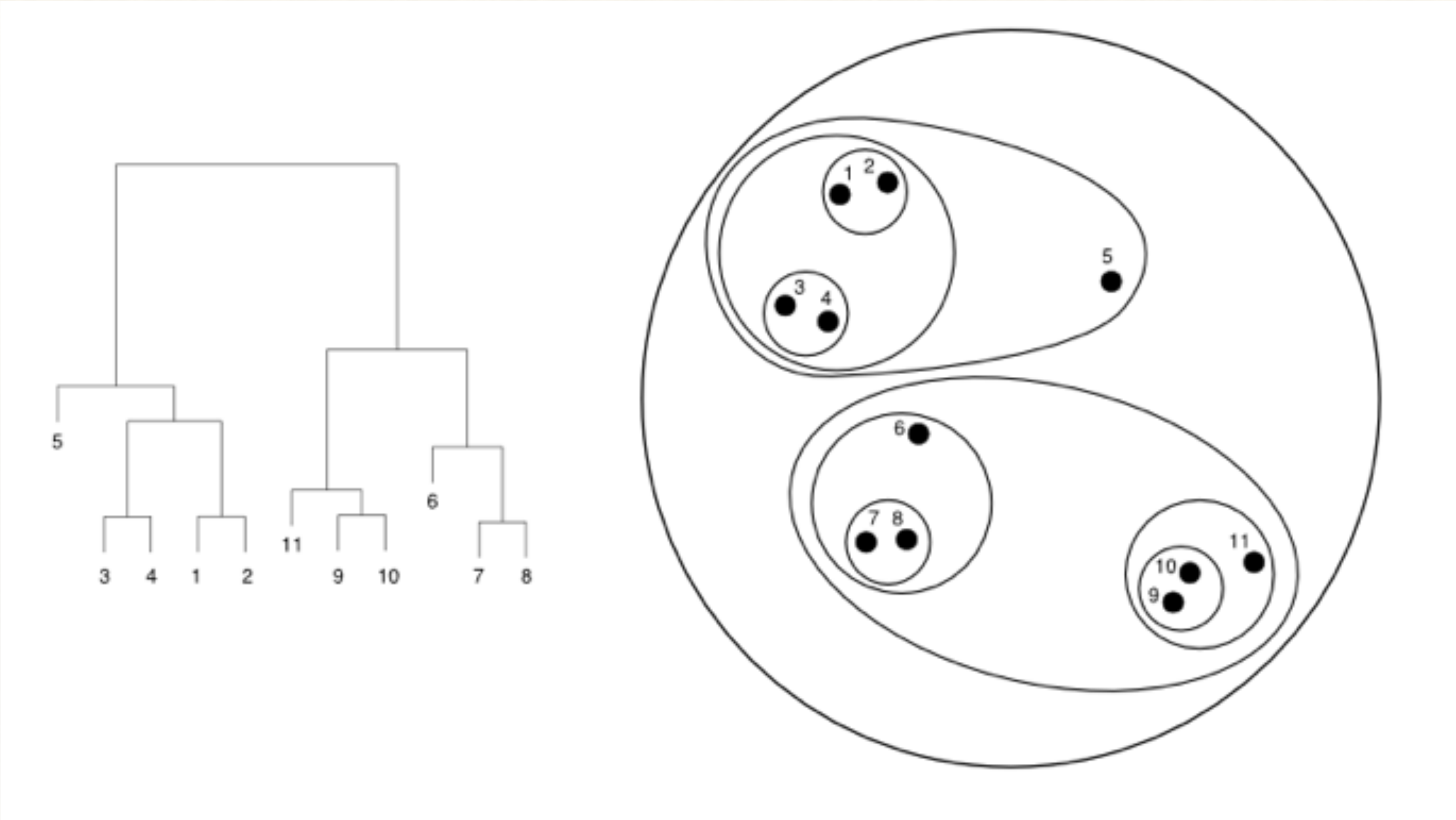
Graph Cut Based

Spectral Clustering



Partition

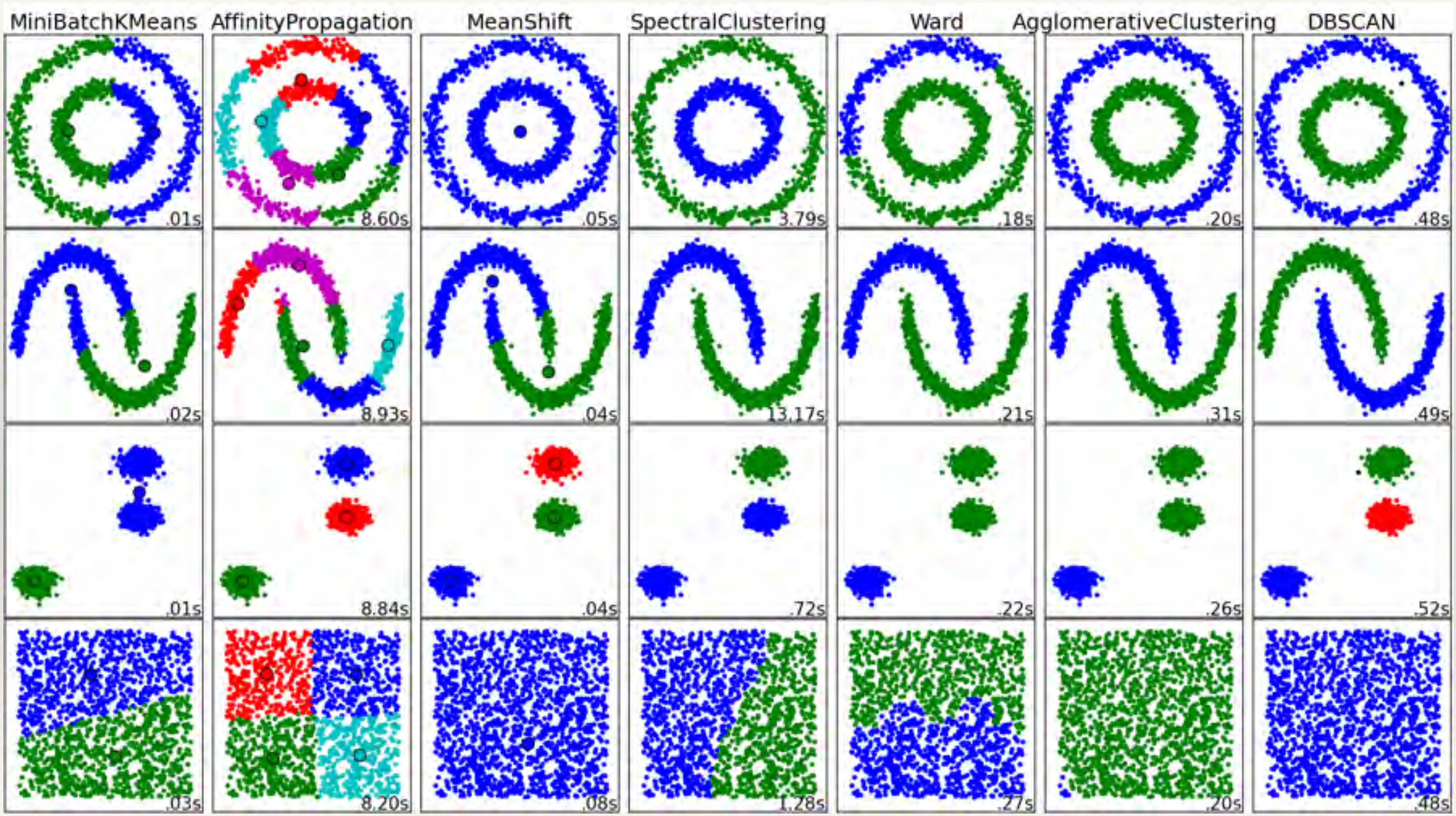
Hierarchical Clustering



Example: Ward Clustering

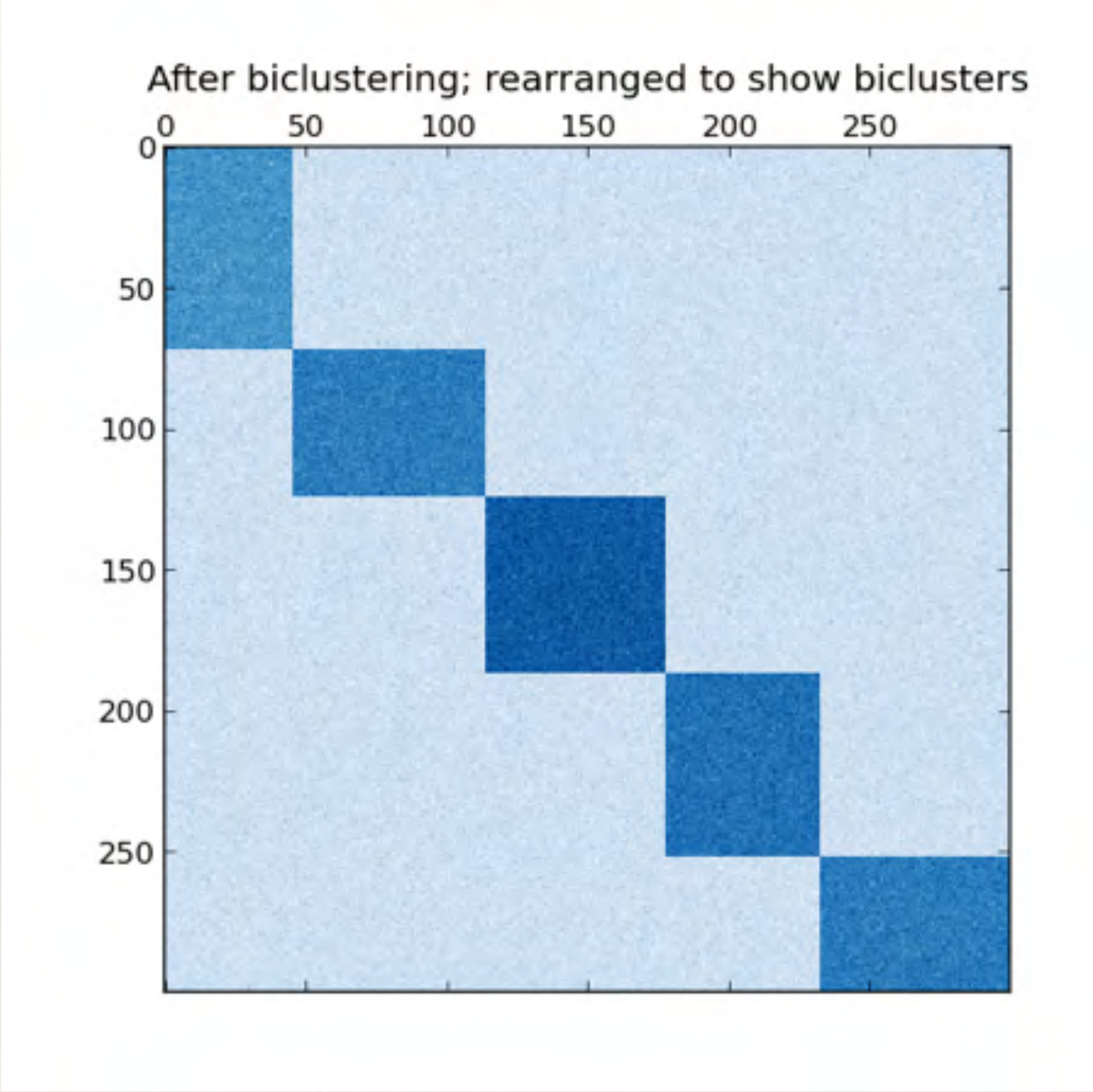
Hierarchical

Comparison



Different Clustering Based on Different Assumptions

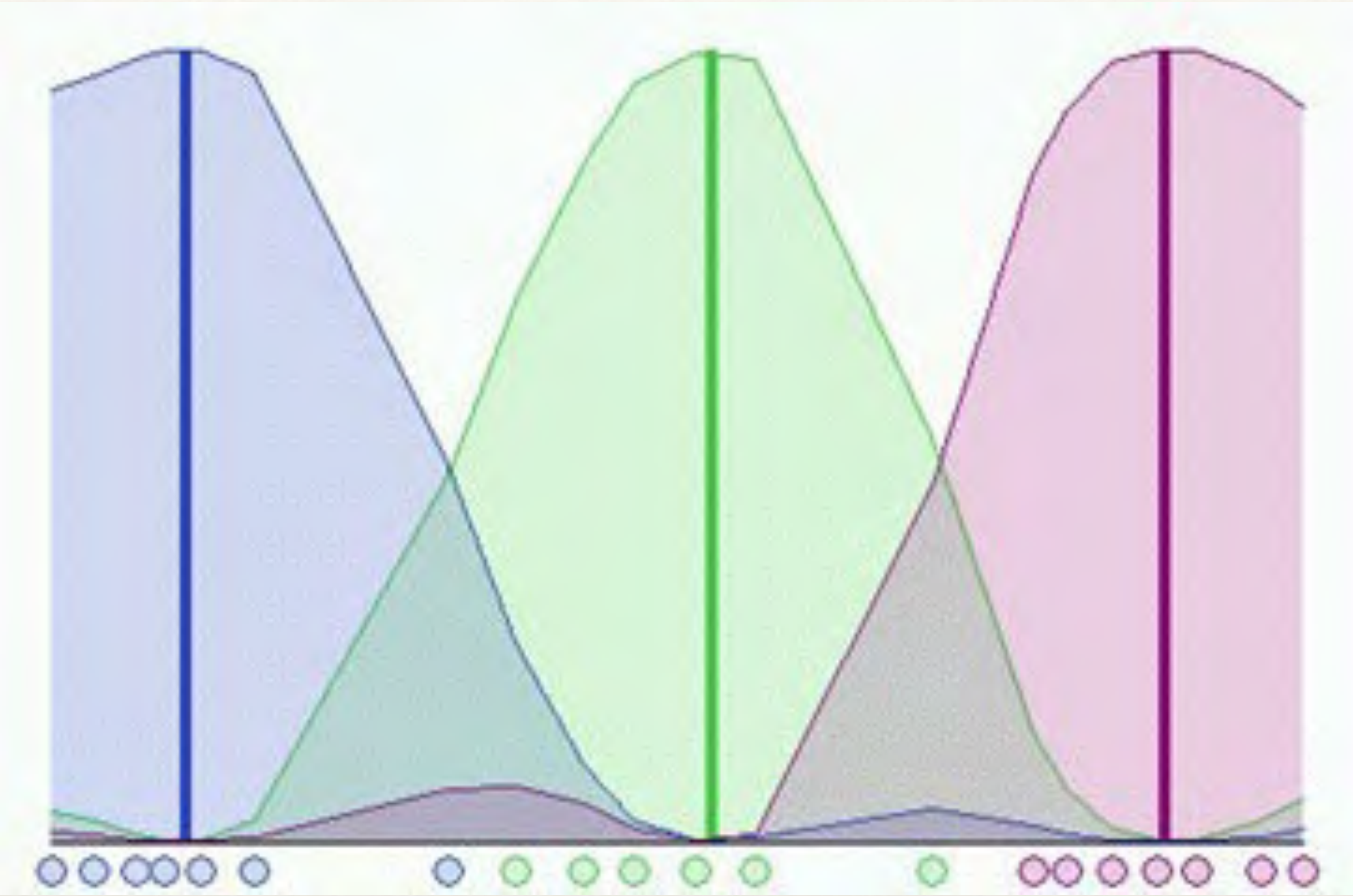
Bipartite Clustering (Biclustering)



Find Blocks in Data Matrices

Bi-Partite

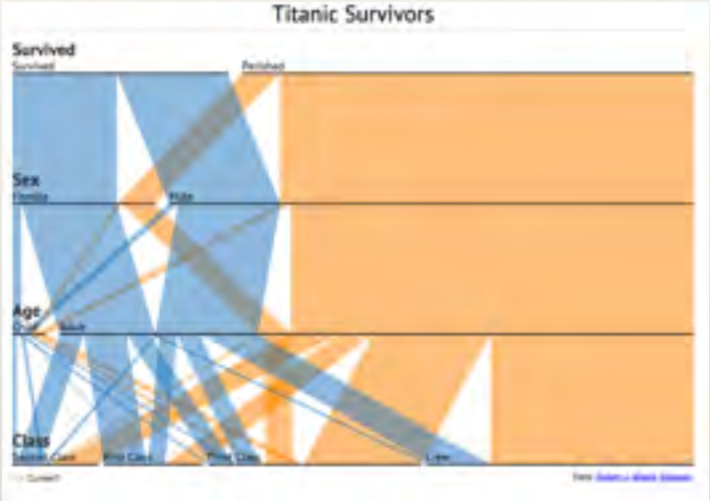
Fuzzy Clustering



Membership in Cluster is Floating Point

Clustering Applications

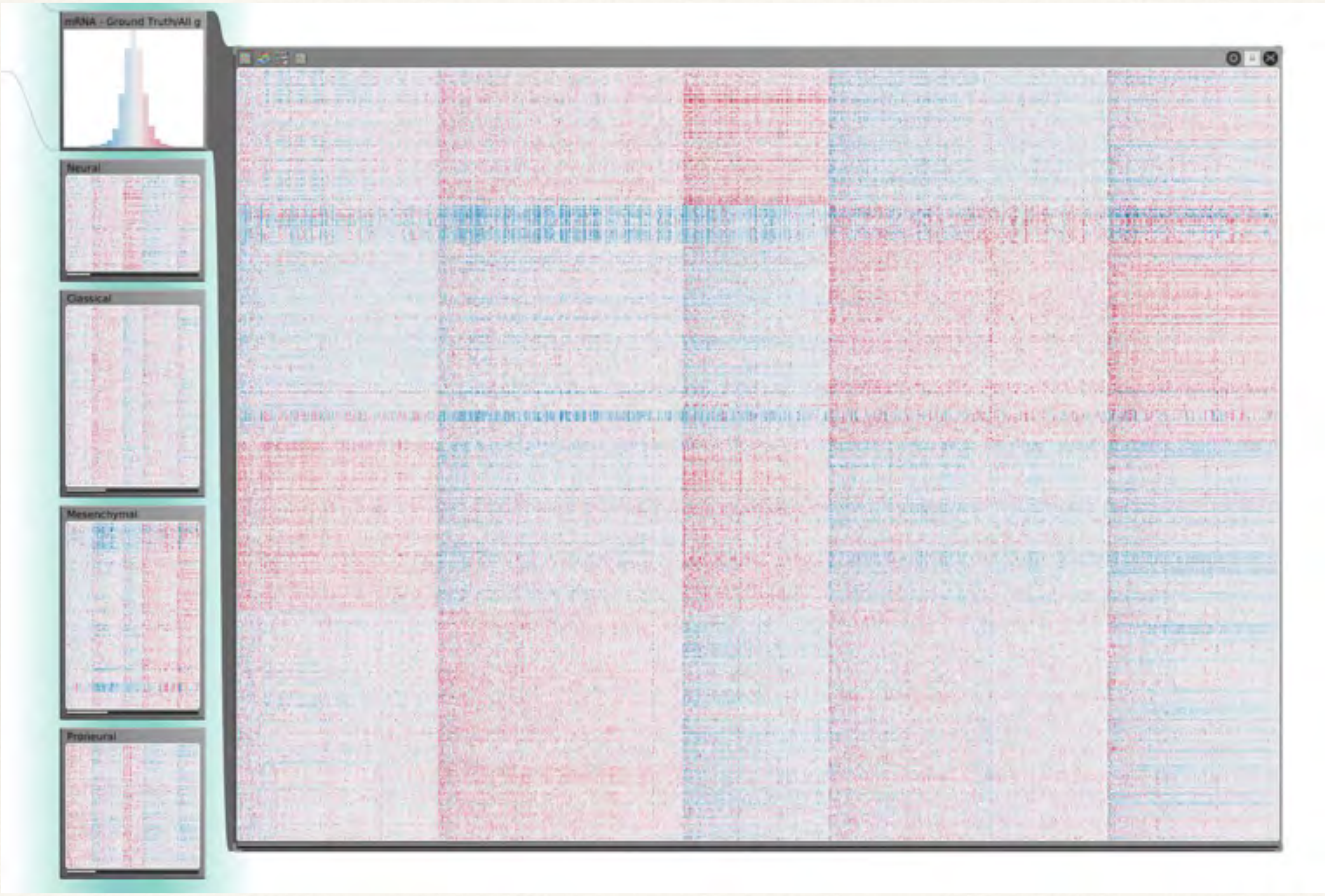
- Clusters can be used to
  - order (pixel based techniques)
  - brush (geometric techniques)
  - aggregate



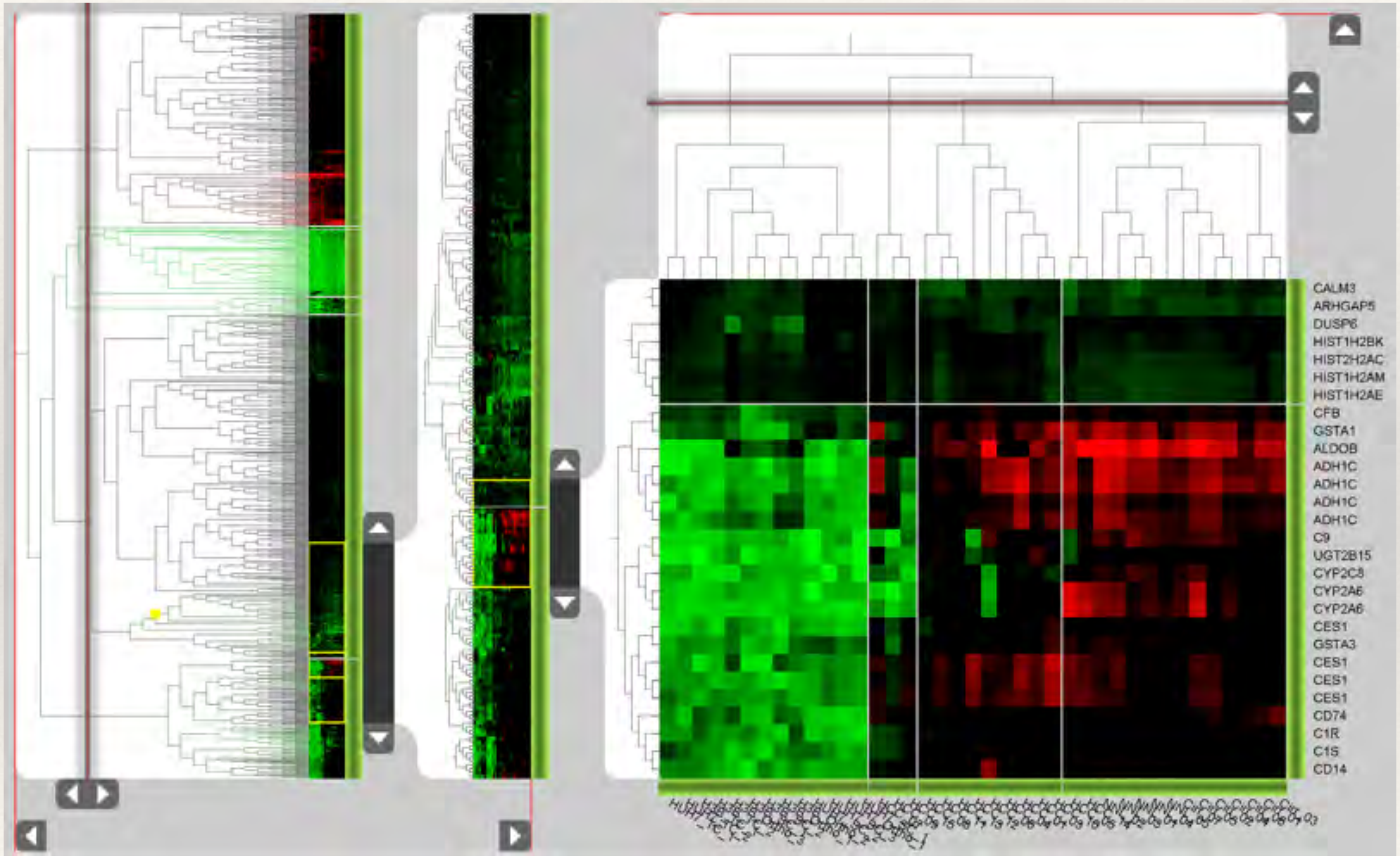
http://www.jasondavies.com/parallel-sets/

- Aggregation
  - cluster more homogeneous than whole dataset
  - statistical measures, distributions, etc. more meaningful

Clustered Heat Map



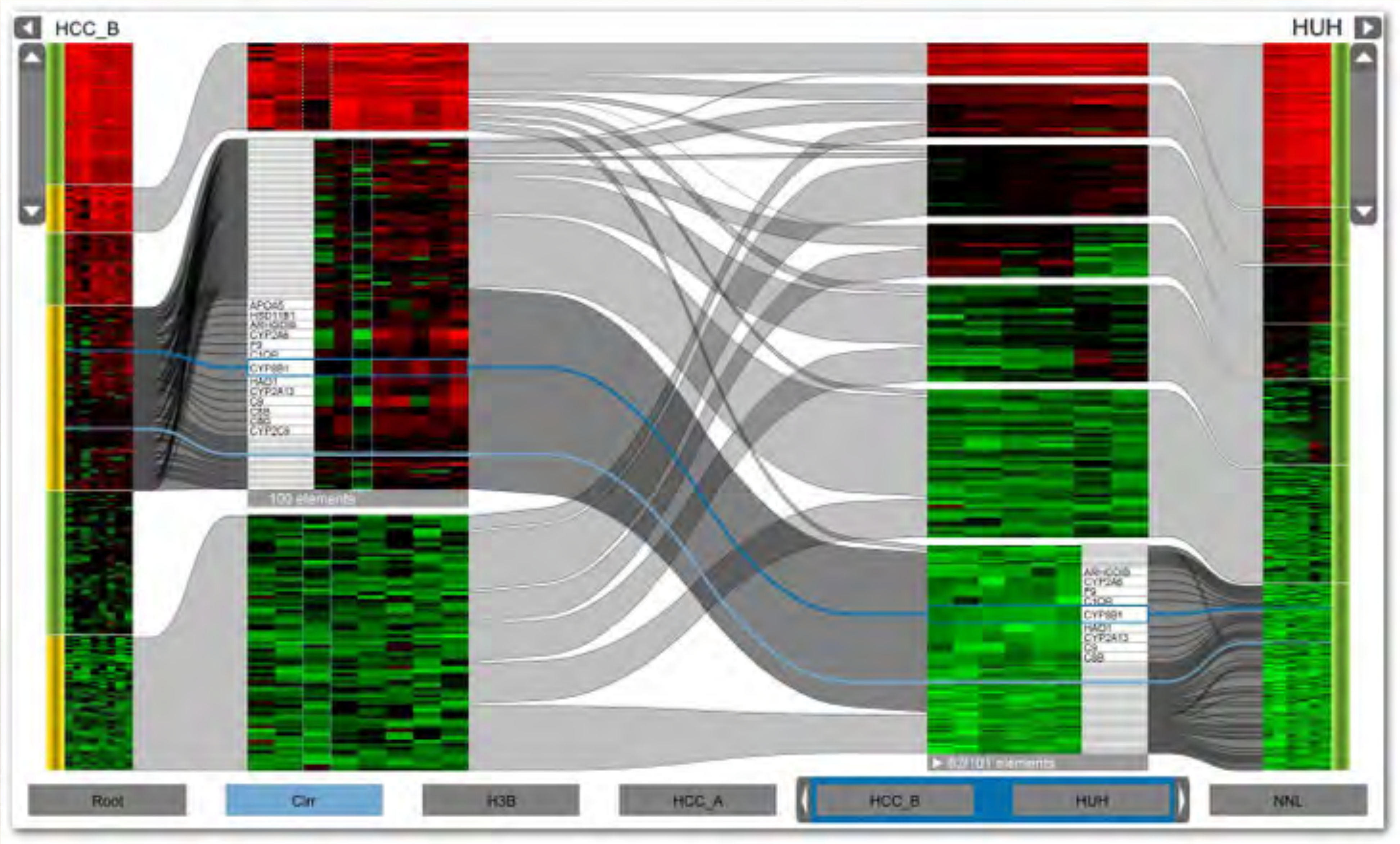
F+C Approach, with Dendrograms



Lex, PacificVis, 2010

Cluster Comparison

Caleydo Matchmaker

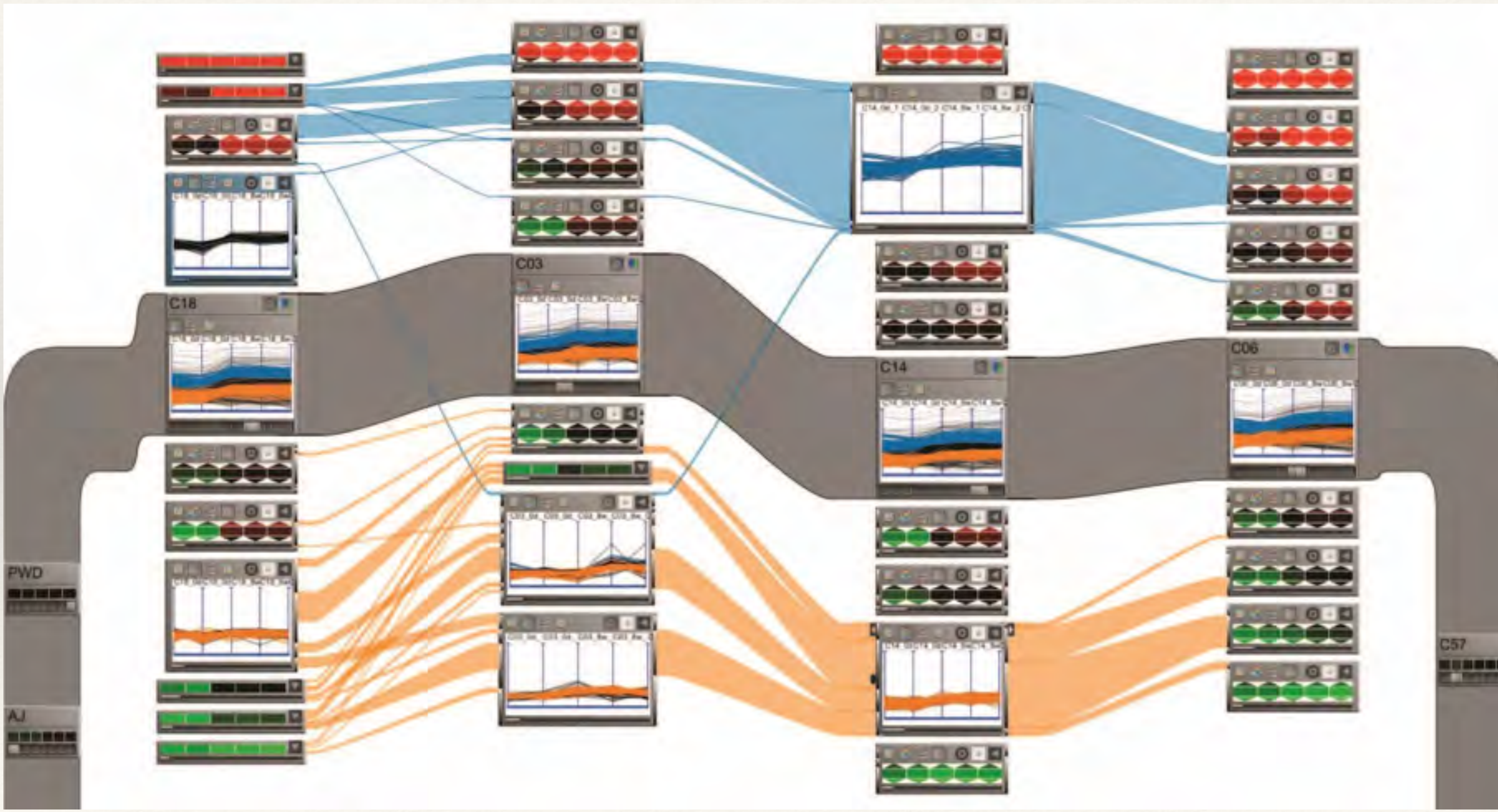


Lex, Streit, Partl, Kashofer, Schmalstieg 2010

https://www.youtube.com/watch?v=vi-G3LqHFZA

Aggregation

VisBricks

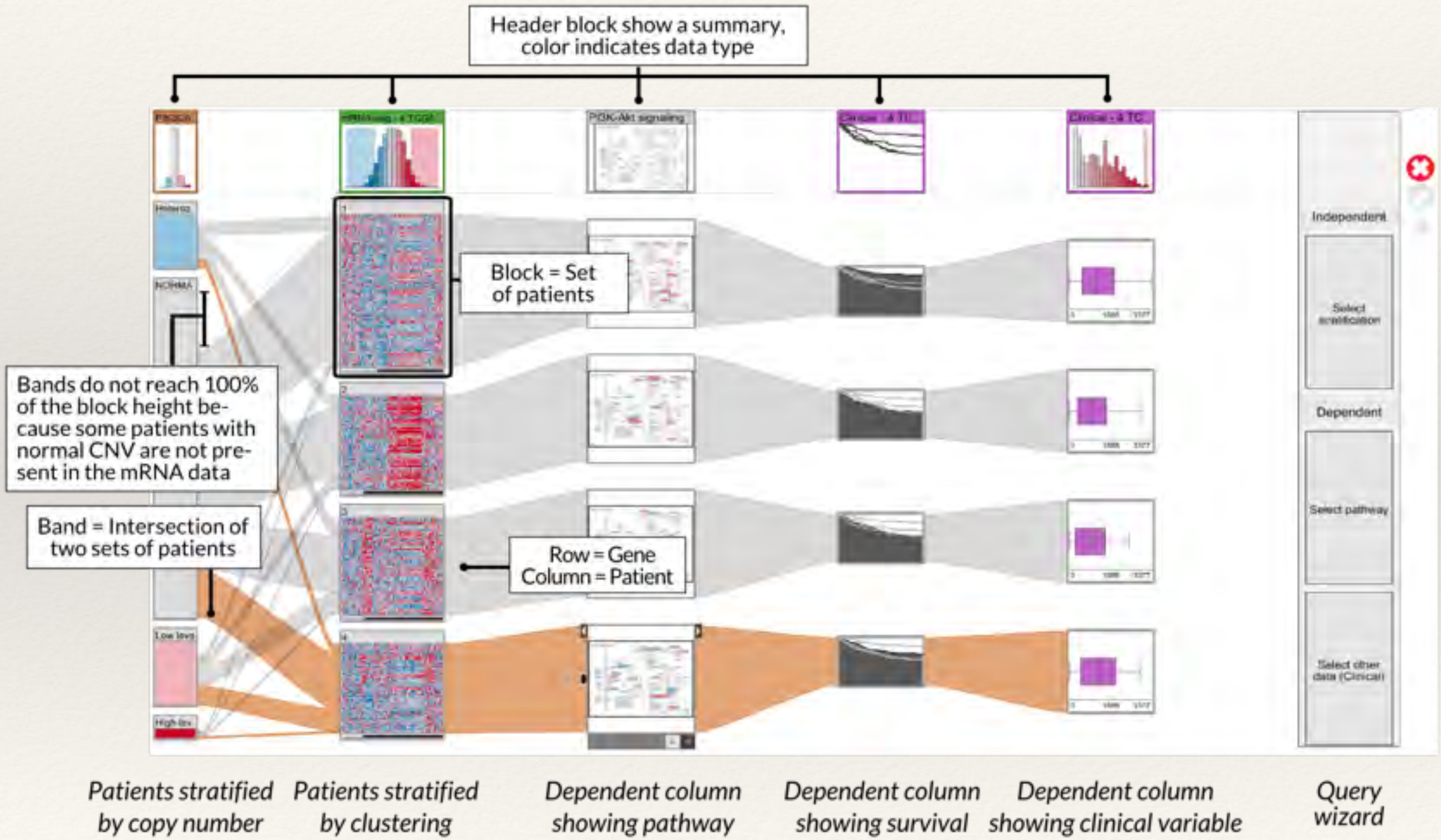


http://people.seas.harvard.edu/~alex/papers/2011_infovis_visbricks.pdf

Heterogeneous Data

Heterogeneous Variables

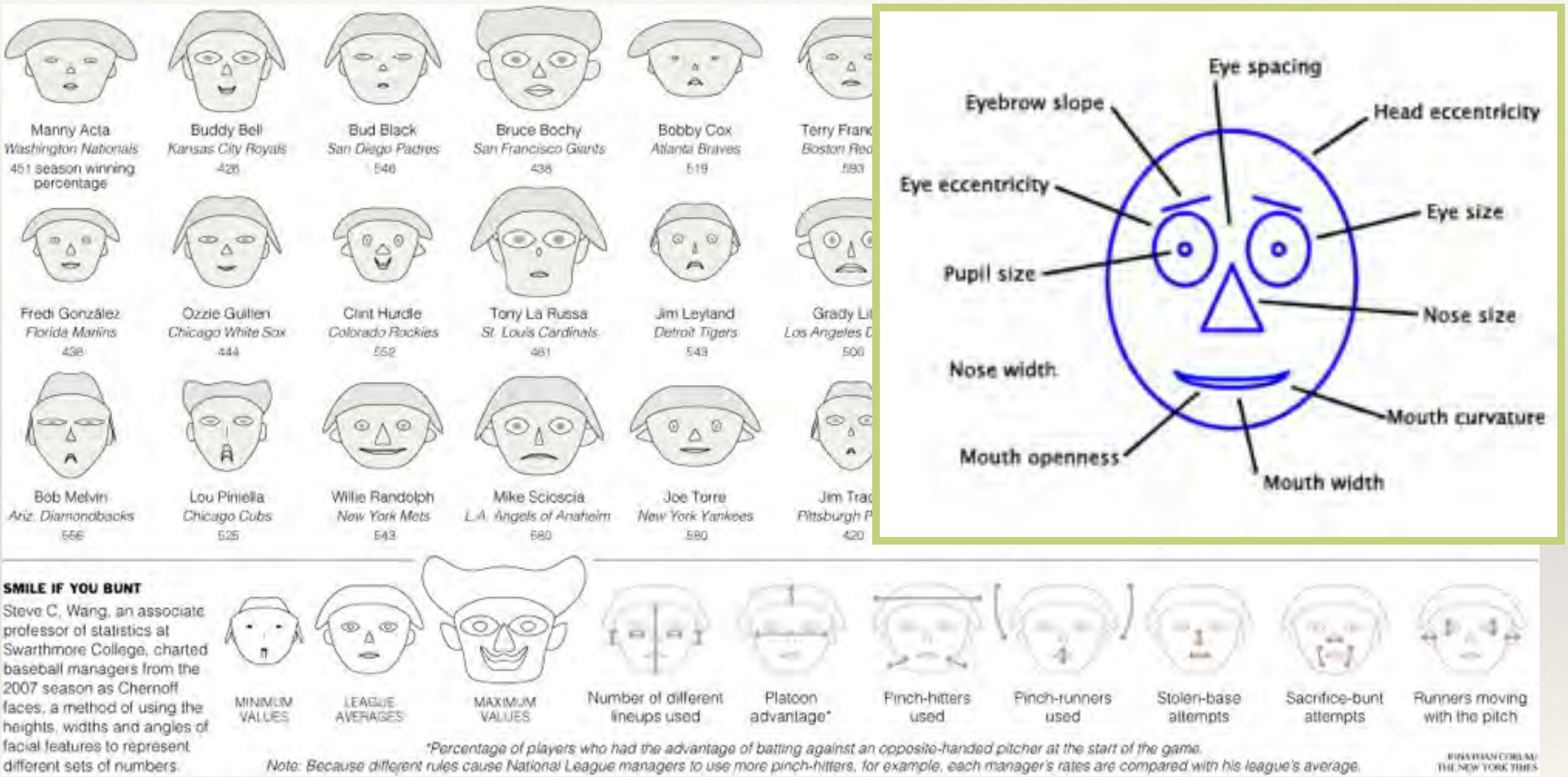
Caleydo StratomeX



https://www.youtube.com/watch?v=s2ZofJ2GVHU

Glyphs

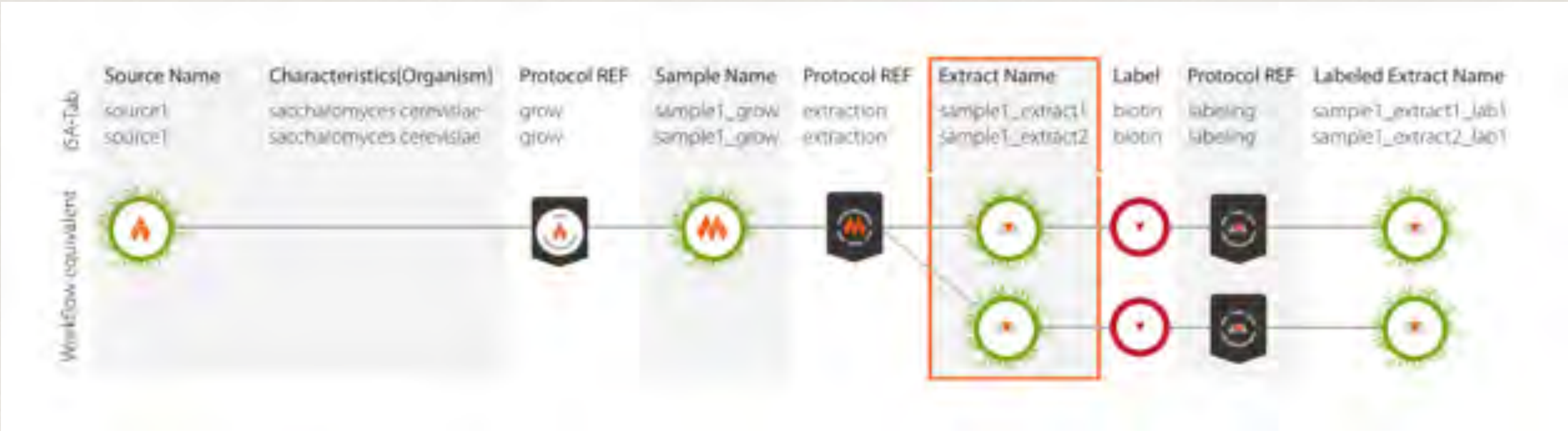
Chernoff Faces



http://www.nytimes.com/2008/04/01/science/01prof.html

Fail?

Complex Glyphs for Bio Workflows



Maguire 2012