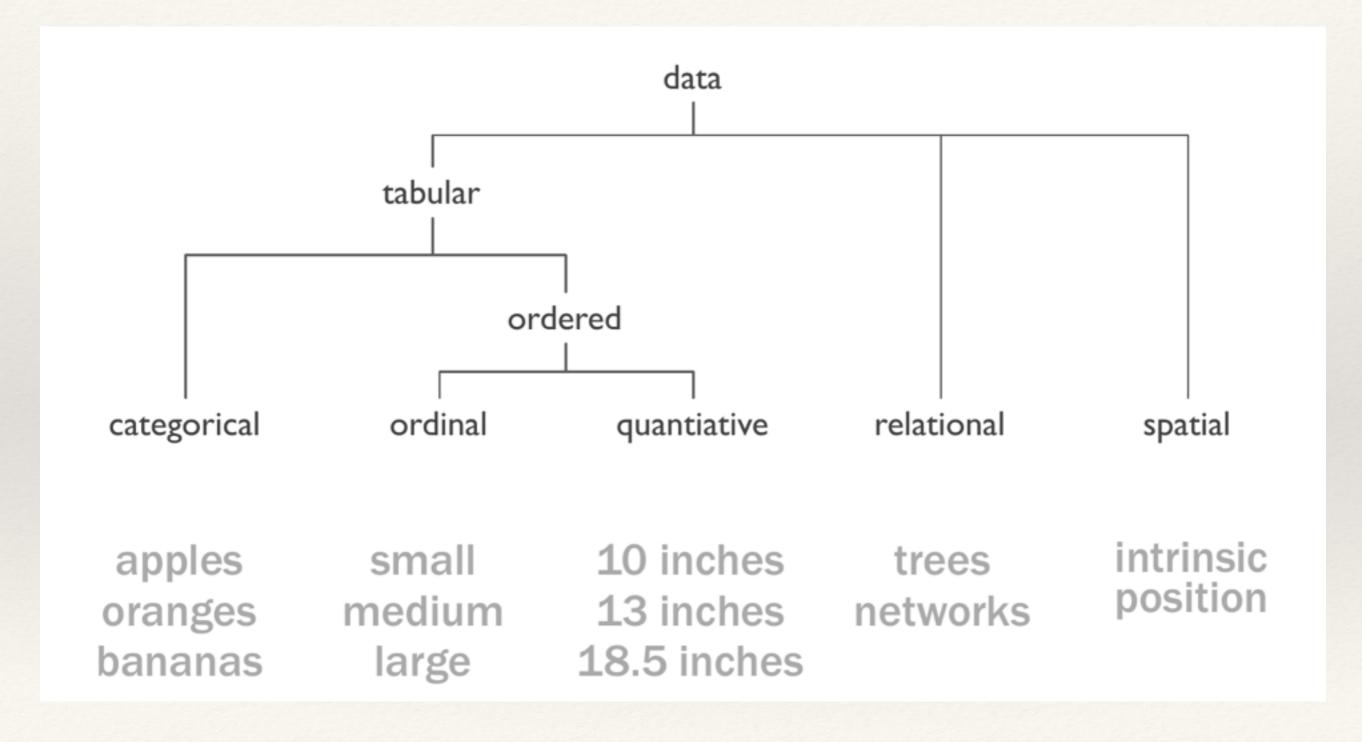
Michael Grossberg

#### Data Visualization

Multi and High Dimensional Data

### Review: Data Types



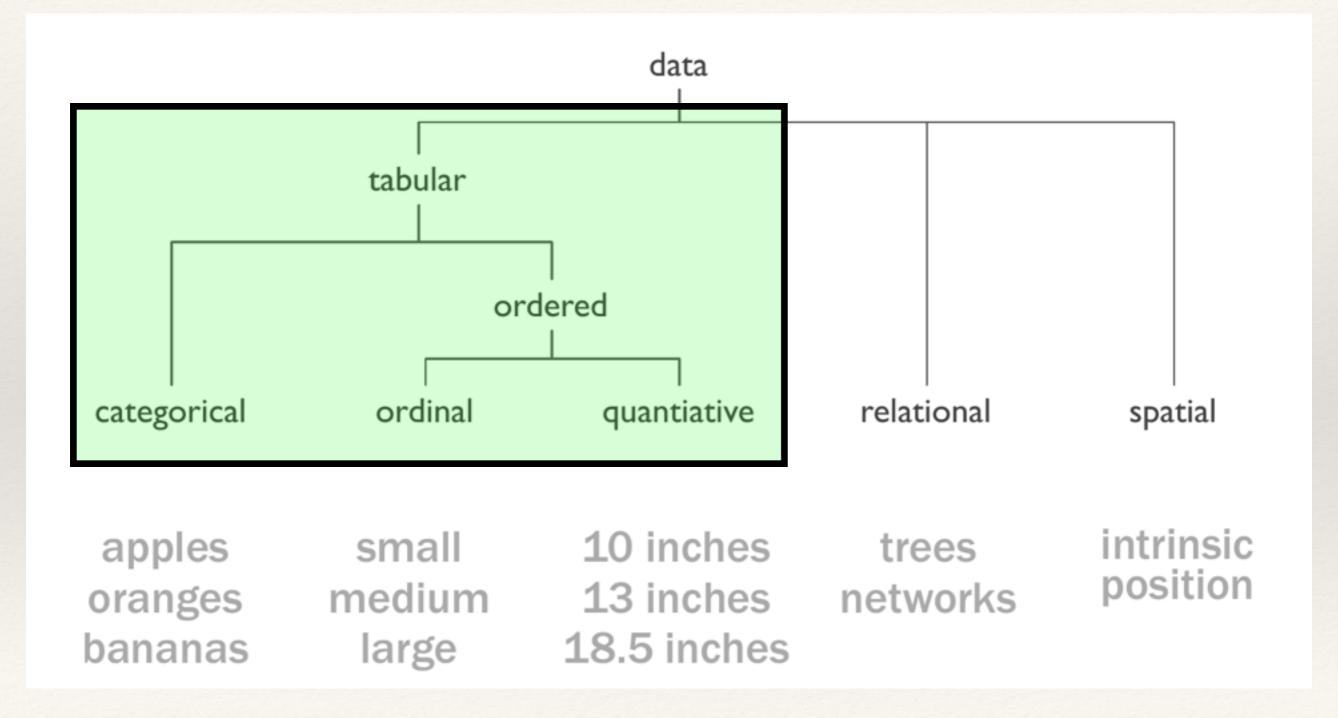
#### Multi-Dimensional Data

- Tabular data, containing
  - \* rows (records)
  - columns (dimensions)
  - \* rows >> columns

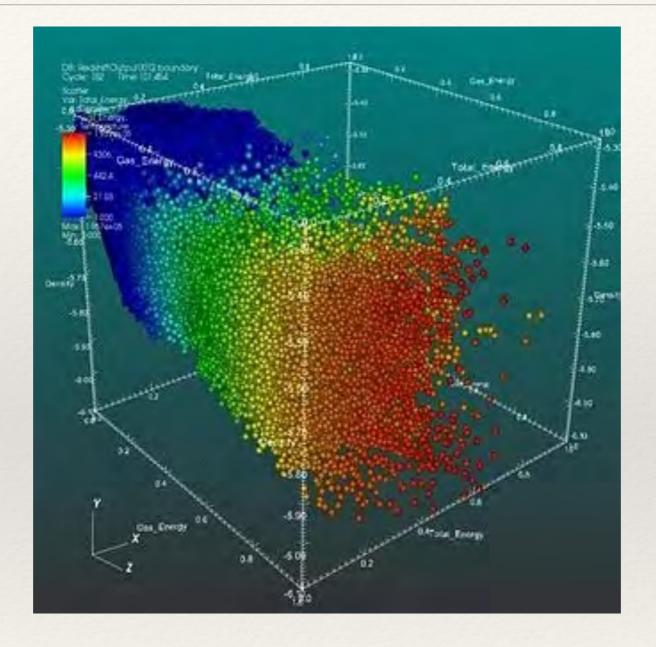
	Age	Gender	Height
Bob	25	м	181
Alice	22	F	185
Chris	19	М	175

- identifiers introduce semantics
- \* Independent & dependent variables
  - dependent are f(independent)

## Muli-Dimensional Type



### Limits to Displaying Dimensions



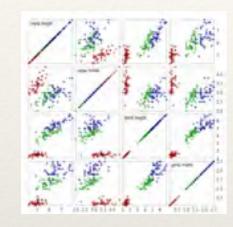
#### Limit? 4D? 5D? ... 10D!?!

#### High-Dimensional Data Visualization

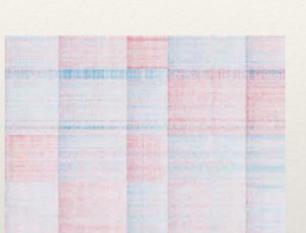
- \* How many dimensions?
  - ~50 –tractable with "just" vis
  - ~1000 –need analytical methods
- \* How many records?
  - ~ 1000 "just" visis fine
  - >> 10,000 –need analytical methods
- \* Homogeneity
  - Same data type?
  - Same scales?

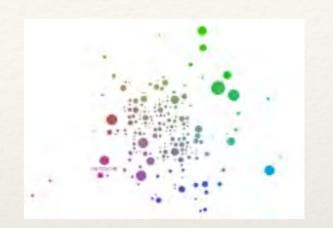
	Age	Gender	Height		
Bob	25	М	181		
Alice	22	F	185		
Chris	19	М	175		
	BPM 1	BPM 2	BPM 3		
Bob	65	120	145		
Alice	80	135	185		
	45	115	135		

## Analytic Component

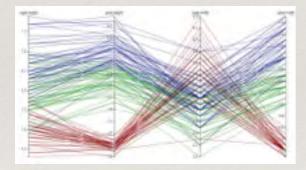


Scatterplot Matrices



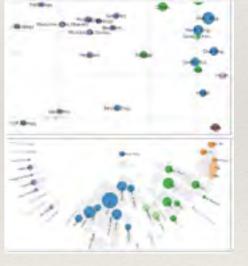


#### Multi-dimensional Scaling



Parallel Coordinates

Pixel-based visualizations/heat maps





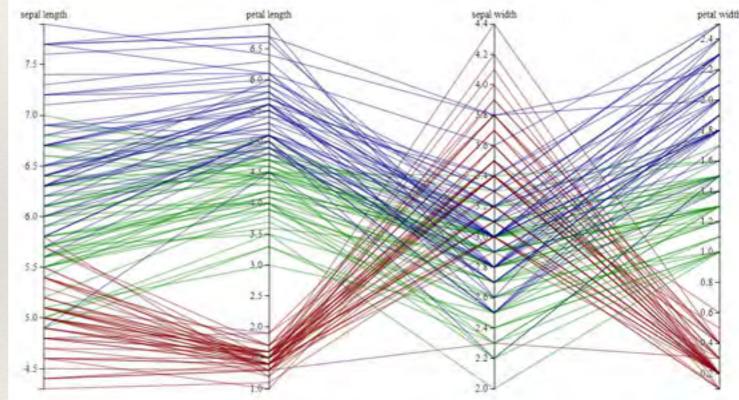
no / little analytics

strong analytics component

#### Geometric Methods

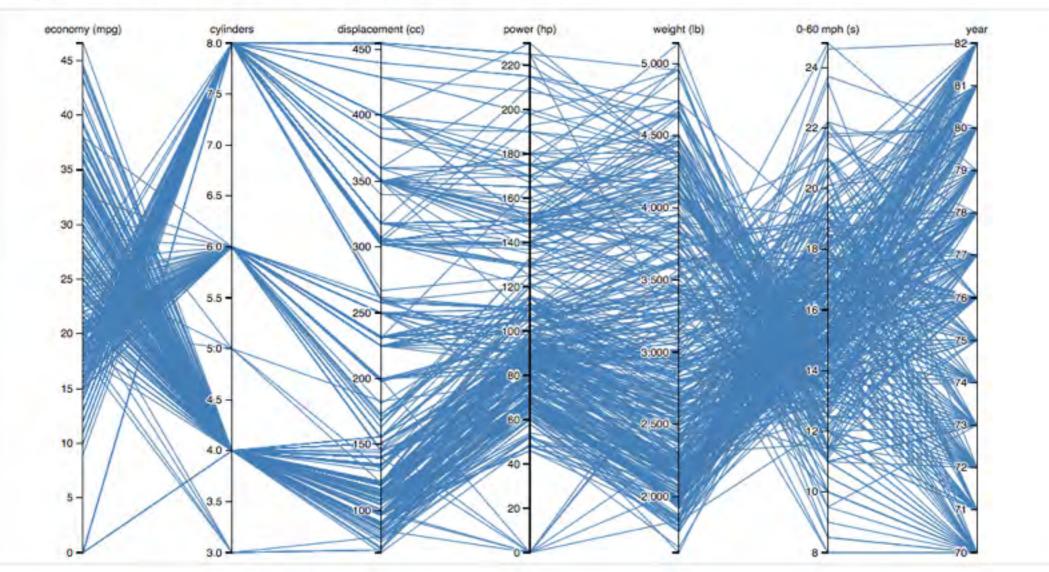
#### Parallel Coordinates

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



#### D3 Parallel Coordinates

#### Parallel Coordinates



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

#### Parallel Coordinates

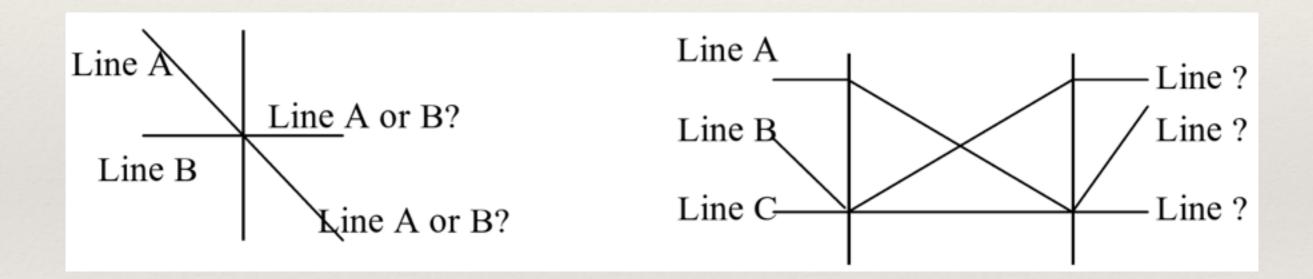
- Shows primarily relationships between adjacent axis
- Limited scalability (~50 dimensions, ~1-5k records)
- Transparency of lines
- Interaction is crucial
- \* Axis reordering
- Brushing
- \* Filtering

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

#### 500-Axis Parallel Coordinate

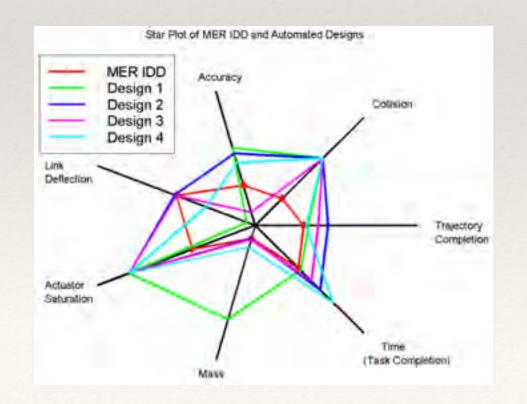


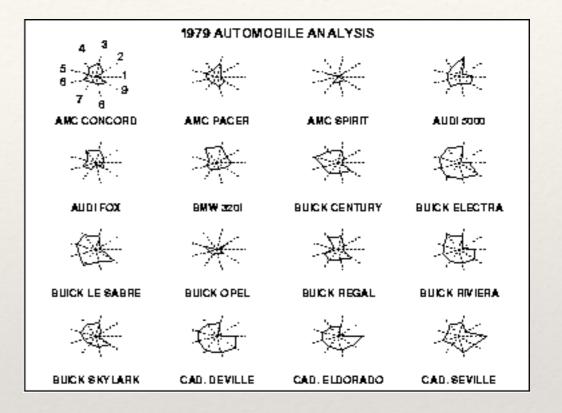
### Ambiguities



#### Star Plot/Radar Plot

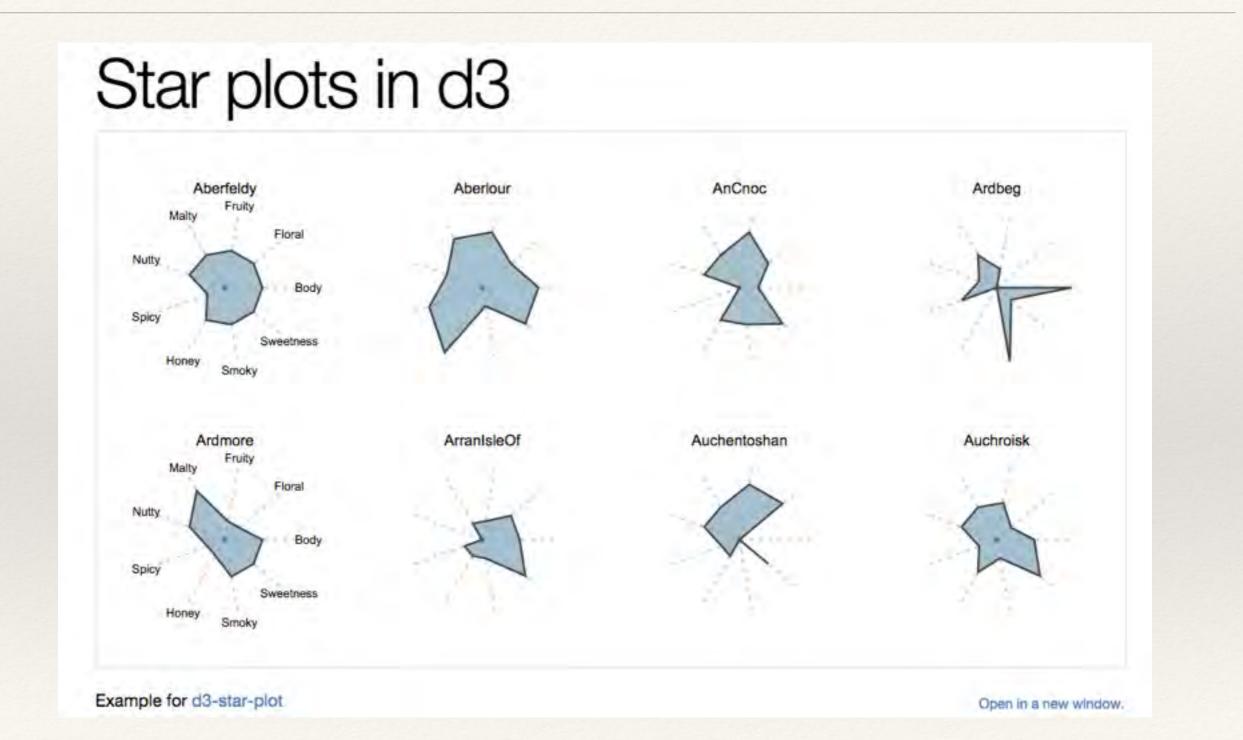
- Similar to parallel coordinates
- Radiate from a common origin





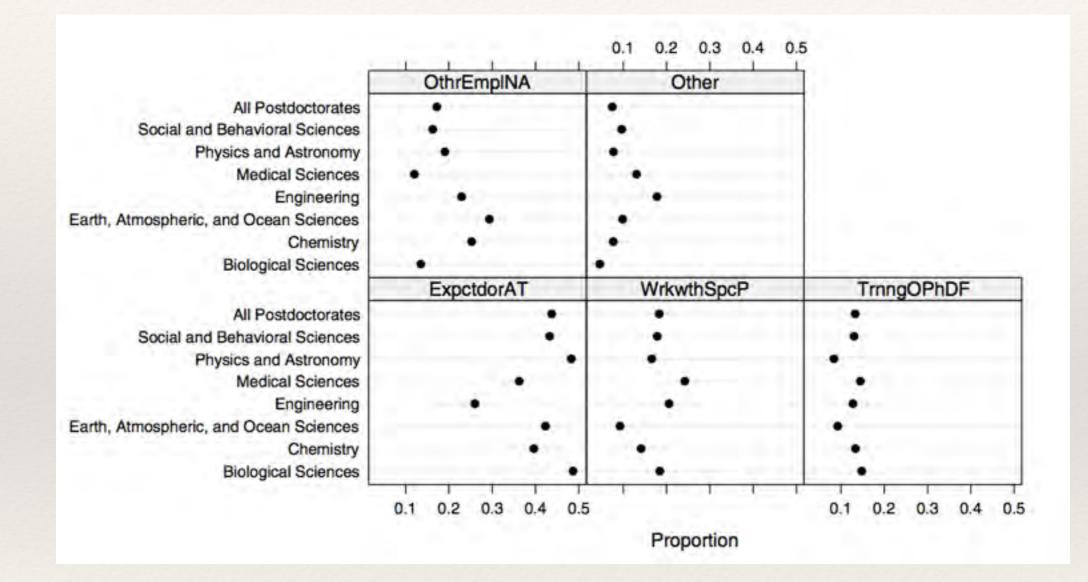
Coekin 1969

#### D3: Star Plot



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

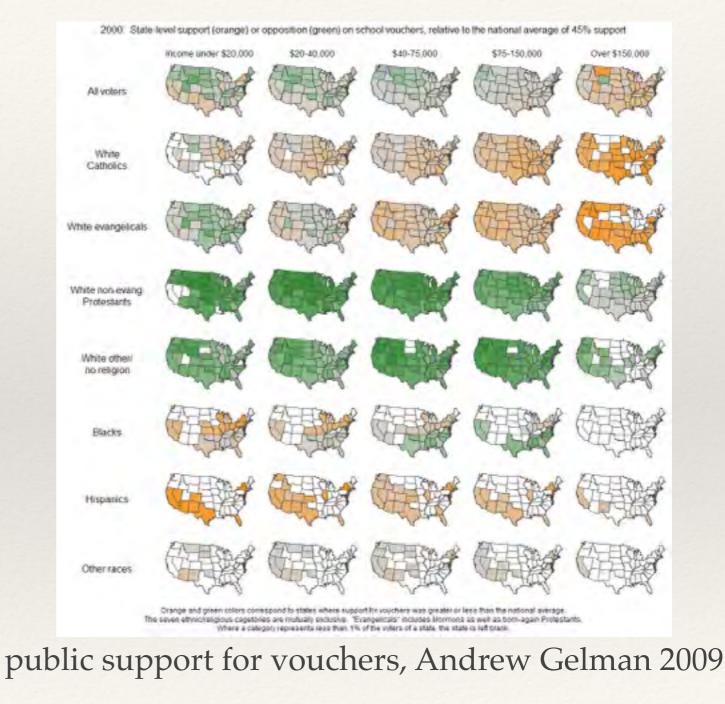
#### **Trellis Plots**



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



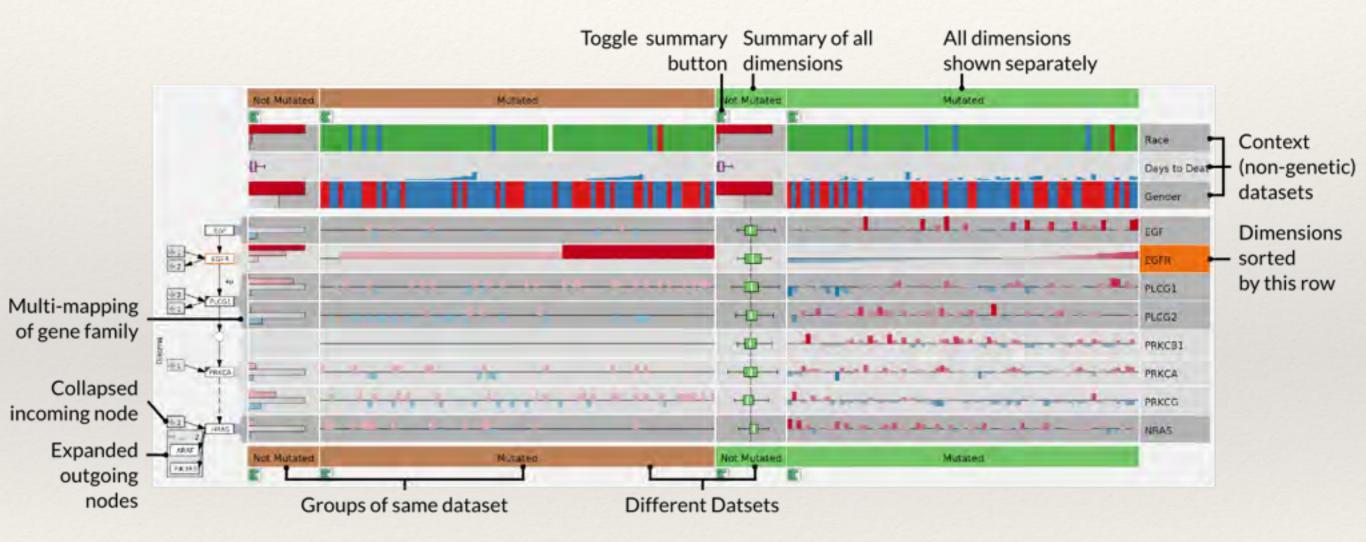
http://andrewgelman.com/2009/07/15/hard\_sell\_for\_b/

## Multiple Line Charts

00,00	20,42	10,441	an in	60,30	08,48	07 AM	17.16	07,00	17.40	06 AM	00.16	00,00	00,40	19 ANI	09.18
1 and	100	A STATE	12.00		100		-	1.1	The states		-	1.1.1	1.0	1.000	-3.4
2			1000								No. Y		-	-	3.4 -1.7
4				-			1					1		1	-49
5			1		111		111		1 10	-	1.00	1771	1 Bad		61
6		The second	NAV.		10.4	LAU	17.3	-	1.11	V.V			M Bar	1.24.1	0.067
1	A. AL		A	The second			A State	· ·		Y at 1	VAN	N. A	1.4.1	ANK ANK	0.26
8		RA I		A-D	A /12	LAN	Vand	Value,			1.1	a al	10.00		1,1
9	A A A	A	1000	100	1000		A AL	APP	Sec. A.	N. N. 1	1.9.4	1.10	A. 64		76
10	1.1			N V				<b>A RA</b>		4				114	5.7
11							- Y -					++++	9.4	-	0.23
12	-	A + 4 A		-		-				A 11			***		
14	1197		WYY Y	1000	***	TTT	TTTT	-		Part I	444		P.P.C.W.	1.11	0.51
15			Sec.	No. of Concession, Name	1.14.10		A	1	1.14	1111		100.0	T. T	And Ball	
16	day 1		1.1.1	A ALK	A	1.2.3.5	1.64	PHIL.				1000		A MAR	-57
17		<b>1</b> 1	MALA I	TAL IN		ALL MA	100	ALC: NOT	Aster	MAA		A Ar		T. Phy	-4.8
.18	4444	Contract of	1.1.1	104.0	1.1.1	11111		1111	10.10.10	ALL HAL	44.44		1000	1444 LA	-4.0
19	1	100	1	dia.	ALC: NO			MAA . 44	ALMAN,	ALCONA,	- 44-		ALC: NO	PPER DATE	A A STAL
20	100.00	Mar Land	44.14	N. Deside	1.1.1.1.1	A.M.	a balant	10000	1.1.1		- M.	Ac as A	1 and	A. 4.844	33
21	11 1TH			-		A	Asia dina		-		1.0	11	1000	ADAU	20
23	-	1911	Traine To	The state	18 J. Ball	10.00	-	-	11.11		TTO: T	1.	1.4.4	Pl al la	-12
24	A Aus		1.1.1	V. Land	A POINT	THE OWNER WAT		and the	Sec. Property	1. 11		A. A.A. A.	AL A	State 1	-27
25 14 44	a shade	and all	166	State of the second second	1.1	24	-	1-7-C		A March	ALALA	ALL SH	AL ANALY		-6.1
28	1.54, 11	Val. L.	MOR	diam'r	1	1. 1.	100		1.11	and the second			A A A A A A A A A A A A A A A A A A A	"moule	1.2
27	MALL	and the second	1.1			and the	44		110	1.11	a late		A Delitera	× (48.	-3.1
28		1	and the	Budday -		ALC: NO	a state	A Della Land	. a. Perk	4	2.111	A.A.A.A.A.A.A.A.A.A.A.A.A.A.A.A.A.A.A.	100	414 1	4.0
29		MAL IT	110	and the second	A CONTRACTOR OF		10.001	A a company		Marrie Marrie			1000	11 11	0.45
30	10.0	11.14	- the	10-11		1 12	the state	ALC: NO				-	The second		84
32	1.1.1	1.1	The second	al al al	A.A.K.			C. La Martin	and and	1.101	CAR Y W	Trees and			6.5

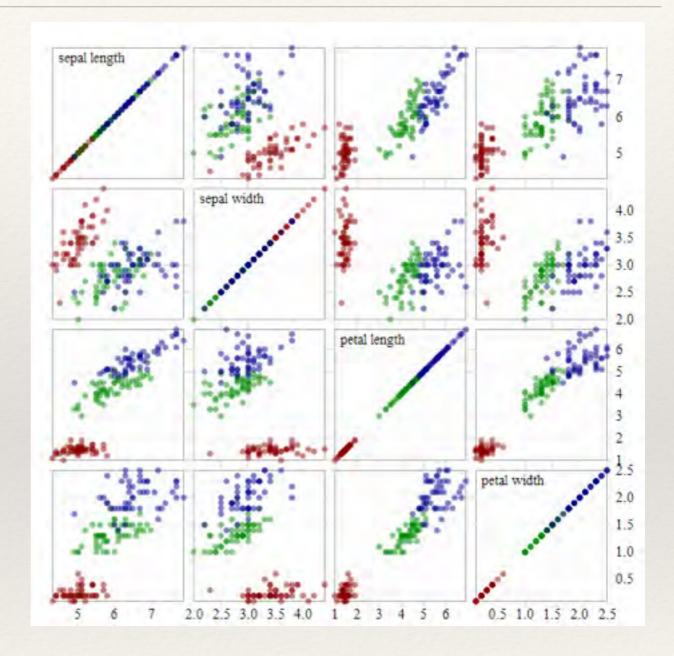
http://square.github.io/cubism/

## **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



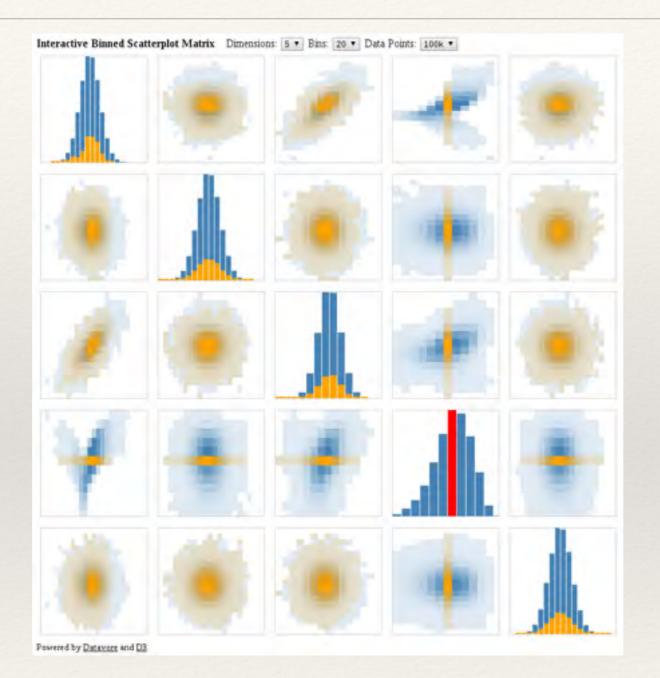
http://mbostock.github.io/d3/talk/20111116/iris-splom.html

#### 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: <u>http://vis.stanford.edu/projects/datavore/splom/</u>

## SPLOM F+C, Navigation

## Rolling the Dice

Multidimensional Visual Exploration using Scatterplot Matrix Navigation

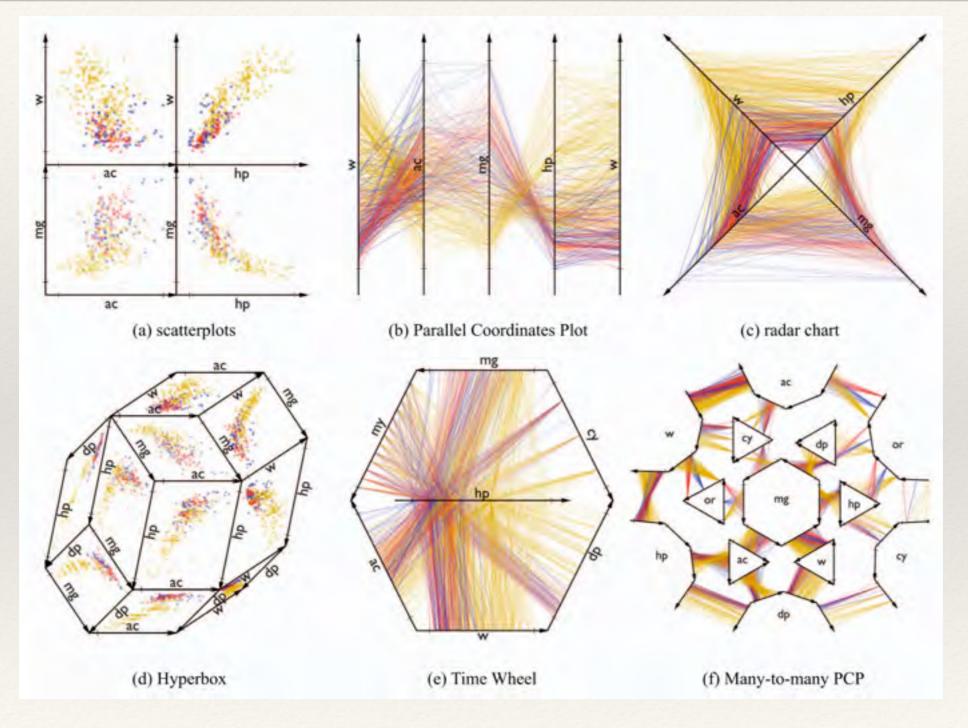
> Niklas Elmqvist Pierre Dragicevic Jean-Daniel Fekete

> > INRIA

http://youtu.be/E1birsp9iYk

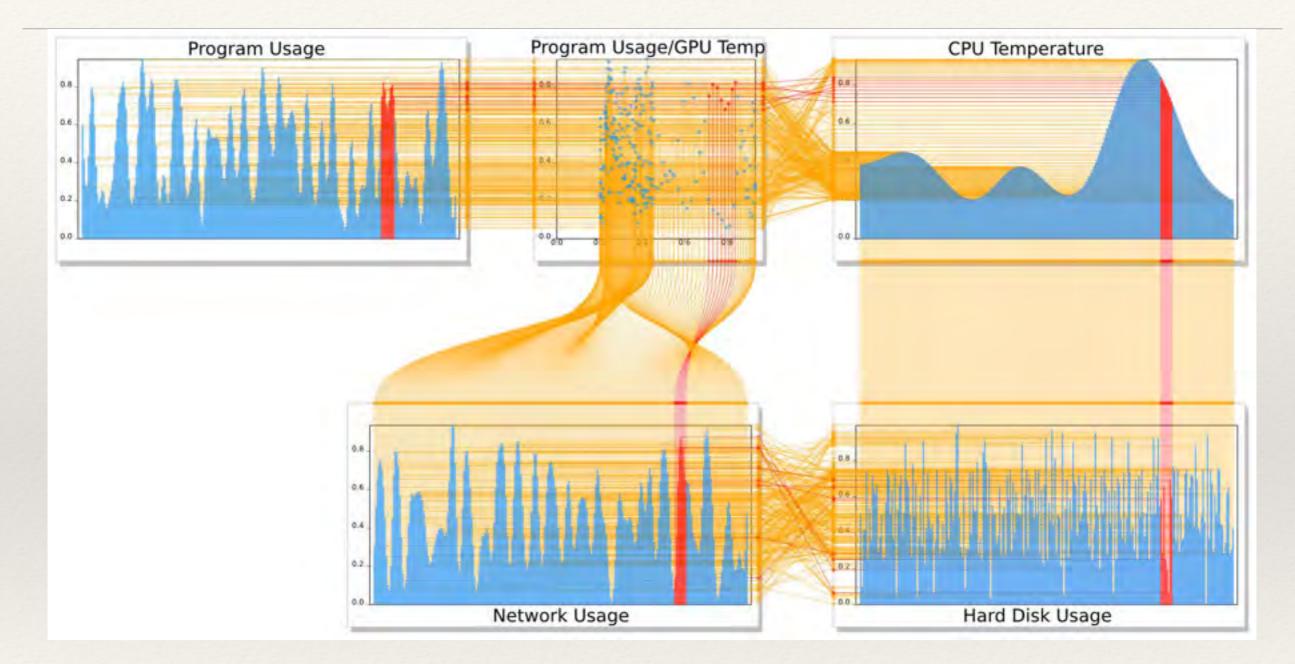
Elmqvist

#### 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-

<u>connectedCharts.pdf</u> <u>http://profs.etsmtl.ca/mmcguffin/research/</u>

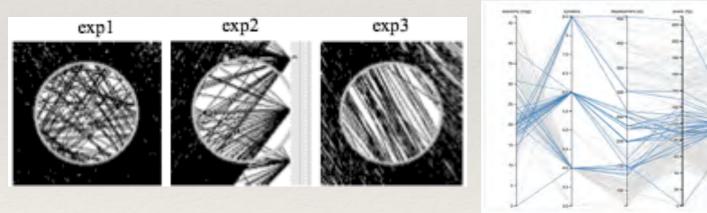
#### Data Reduction

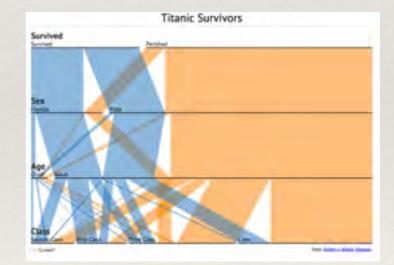
## Reducing Rows

## Sampling

## Filtering

## Clustering



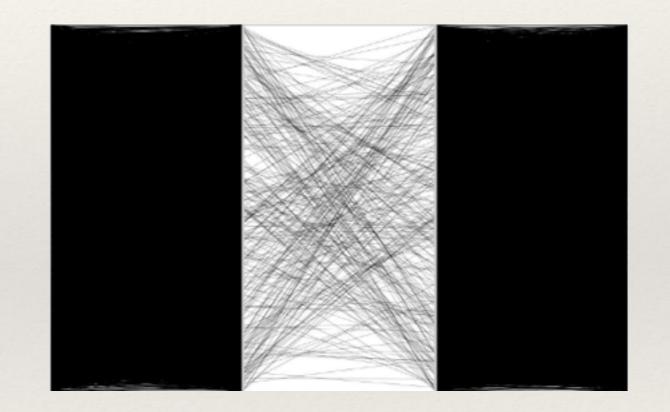


http://www.jasondavies.com/parallel-sets/

Later

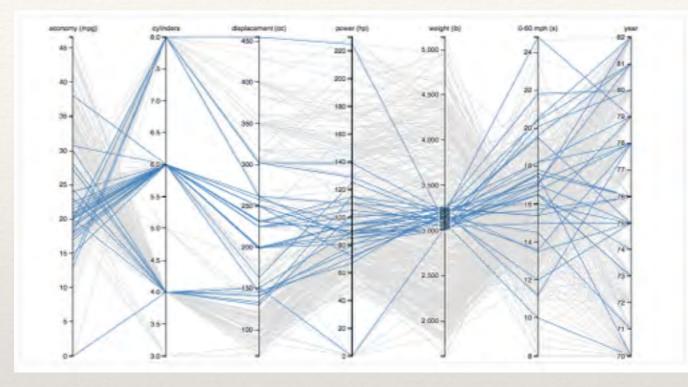
# Sampling

- \* 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

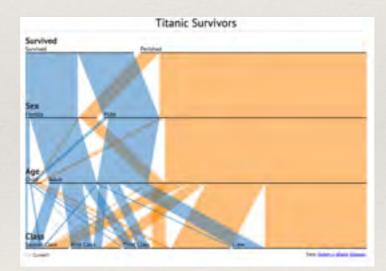
#### Filter Example



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

## Clustering



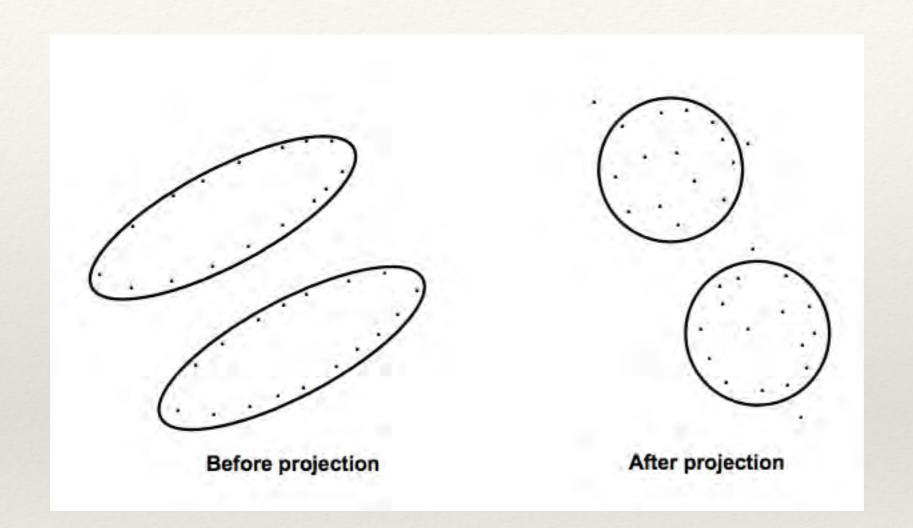


http://www.jasondavies.com/parallel-sets/

Later

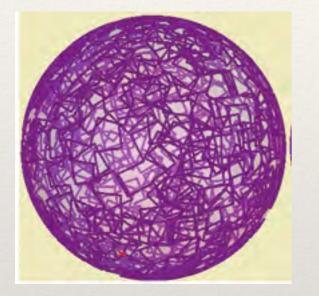
#### **Dimension Reduction**

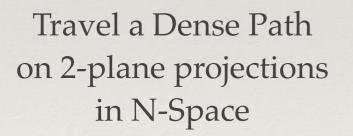
### Simple Random Projection

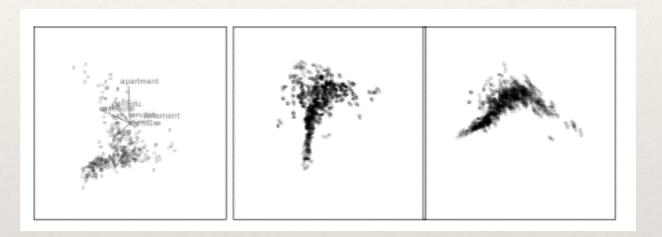


Experiments with Random Projection, Sanjoy Dasgupta, 2000

#### Grand Tour







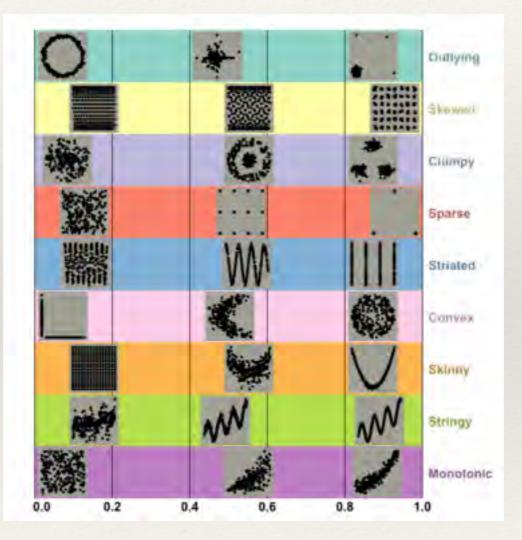
Look at all the scatterplots (Movie)

## Interesting Projections

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



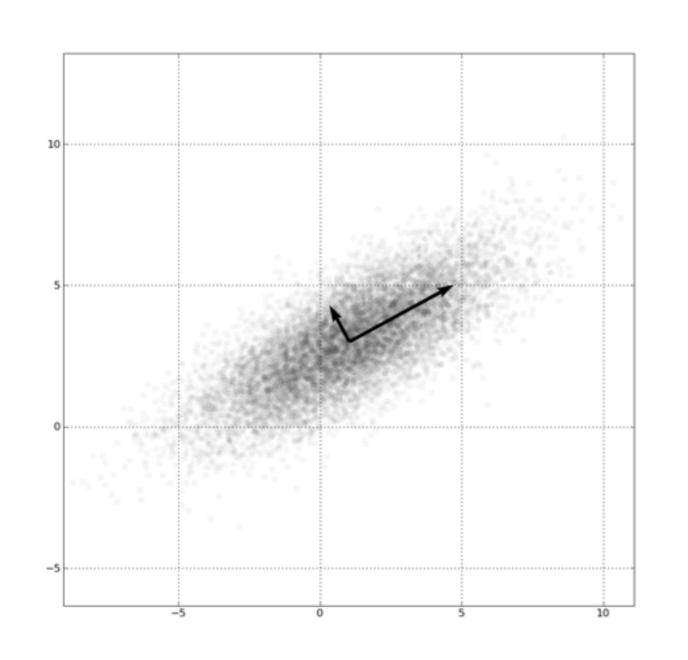
Look for projections which look interesting



TN Dang, L Wilkinson - 2014

# Searching for Projections

# Principal Component Analysis (PCA)



### 1-D mean, stdev

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

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

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

### N-D mean, (co-)variance

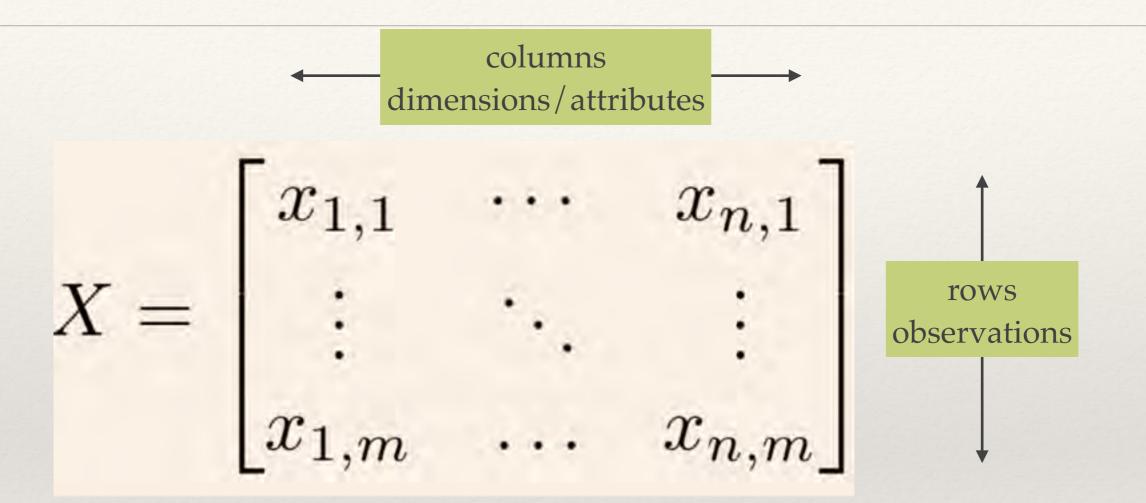
N-D Mean

$$\mu = E[\mathbf{x}] = \frac{1}{n} \sum_{i=0}^{n} \mathbf{x}_{i}$$
$$\mu = \{\mu_k\}, \ \mu_k = \frac{1}{n} \sum_{i=0}^{n} x_{i,k}$$

N-D Covariance

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

### Matrix Versions

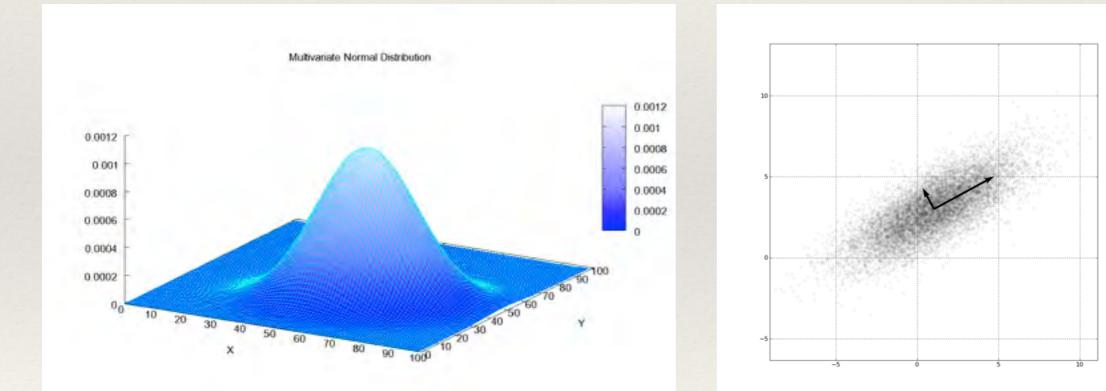


N-D Covariance

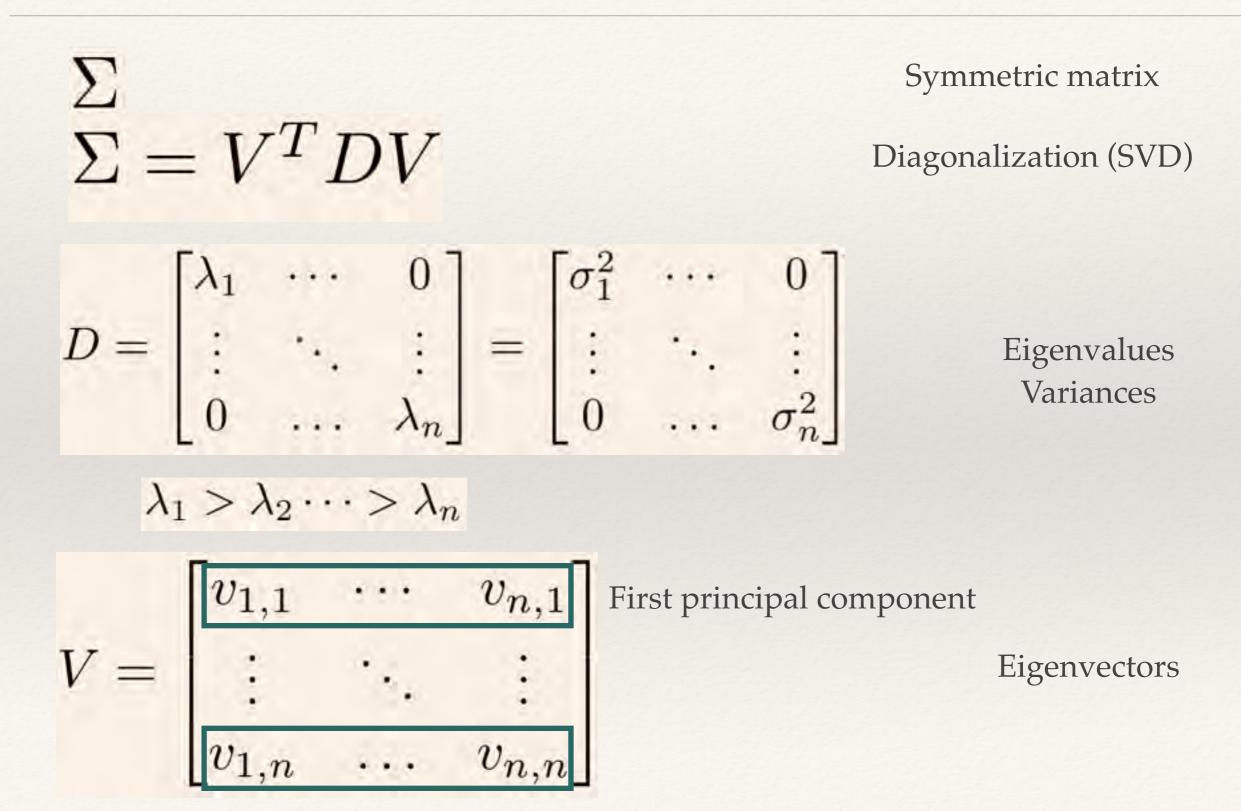
 $\Sigma = \check{X^T}\check{X}$ 

### N-D Normal Distribution

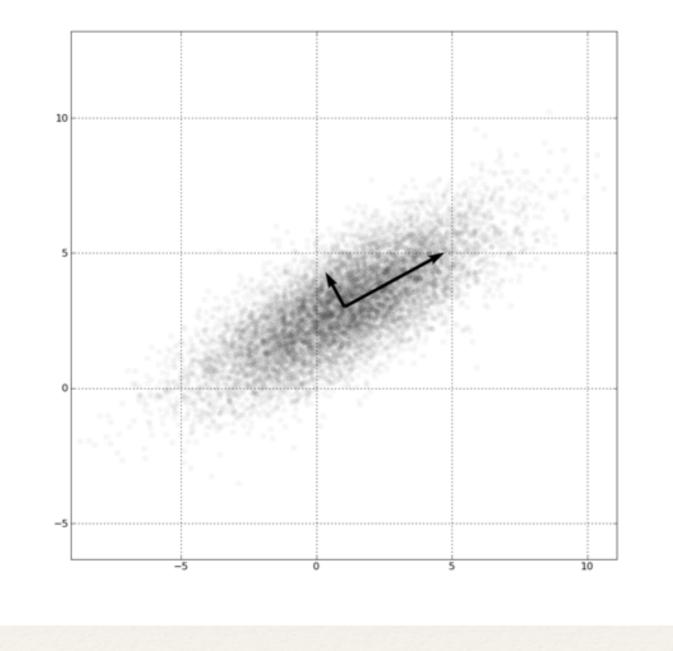
$$f_{\mathbf{x}}(x_1,\ldots,x_k) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$



# Diagonalization

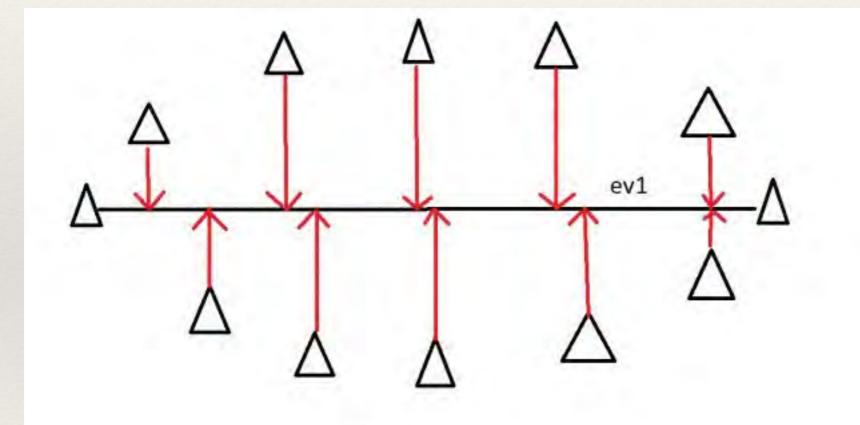


## **Eigenvectors = Principal Components**

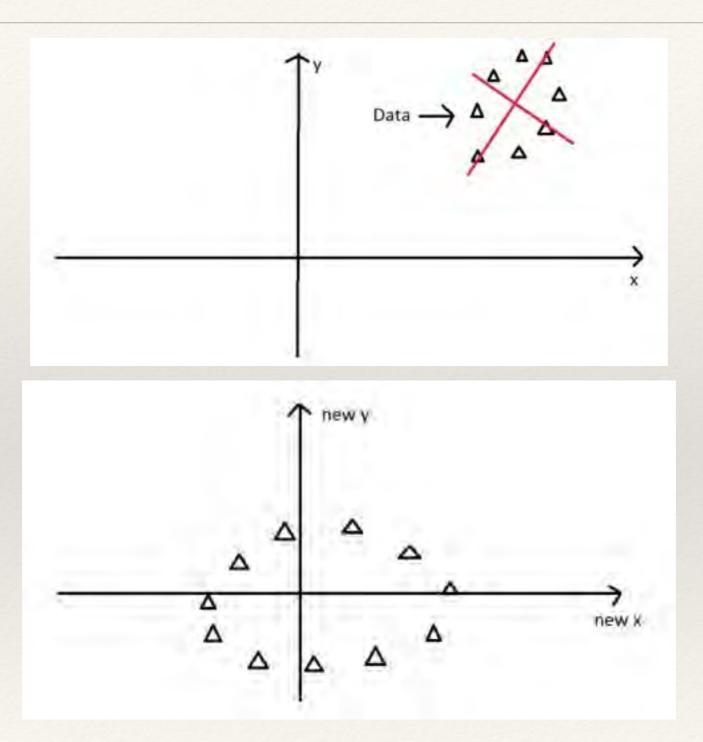


First principal component = direction of greatest variance

Least Squares



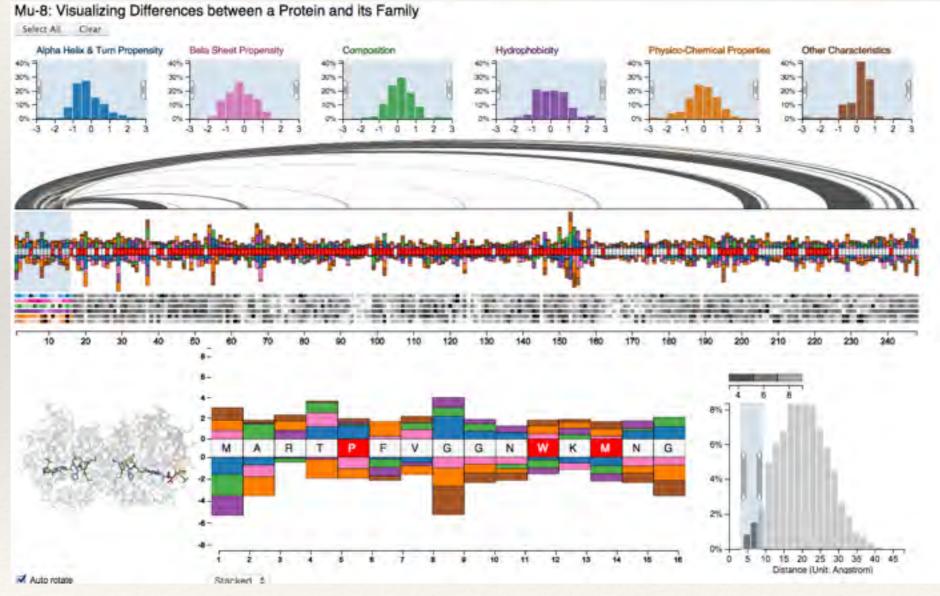
### Data Oriented Coordinates



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

### PCA in Vis Reduction

#### PCA from Harvard Cs171 student project

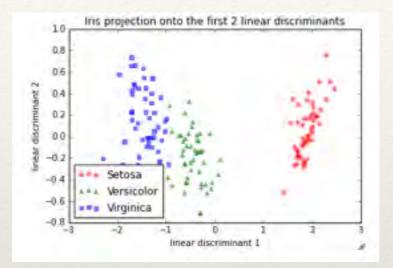


http://mu-8.com

Mercer and Pandian

### Many other Linear Component Reductions

Linear Discriminant Analysis (LDA)



#### Separate classes

Latent Semantic Analysis (LSA)



#### Also

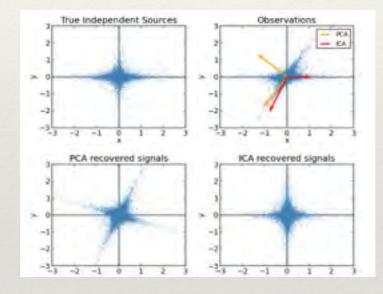
Factor Analysis

**Dictionary Learning** 

Non-negative Matrix Factorization

# **Projection Pursuit**

#### Independent Component Analysis (ICA)

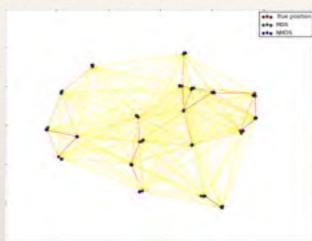


Make components independent

Maximize Entropy/Non-gausianness

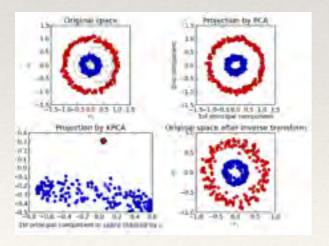
# Non-Linear Dimension Reduction

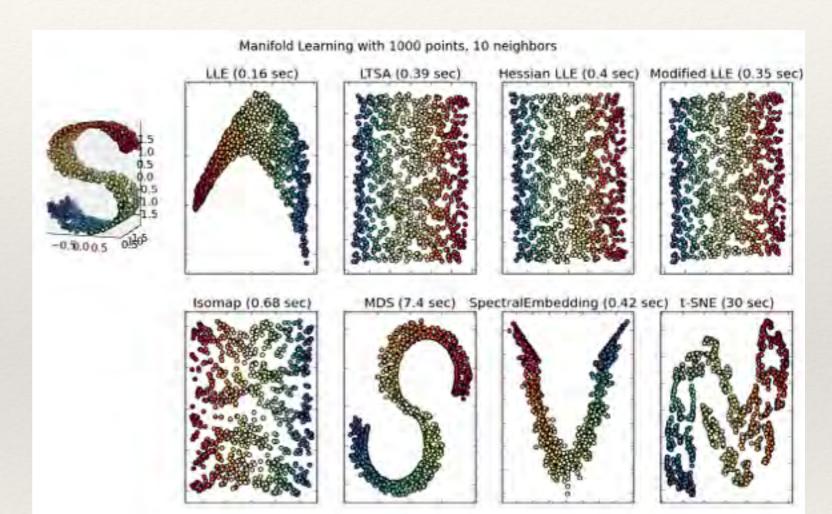
#### Multi-Dimensional Scaling (MDS)



Preserve distances between observations

Kernel PCA



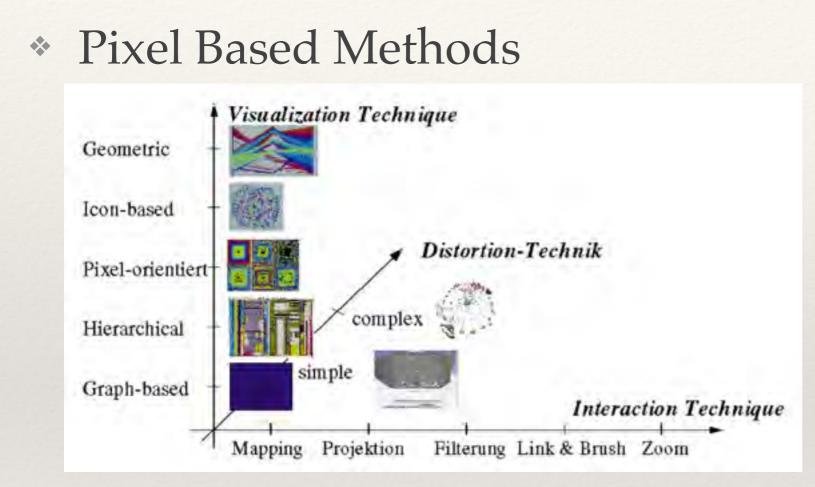


Manifold Learning

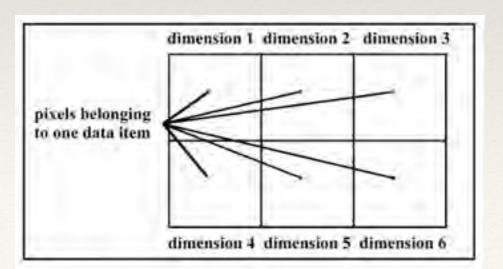
Non-Linear map to lower dimensional space

### Pixel Based Methods

### Daniel Keim

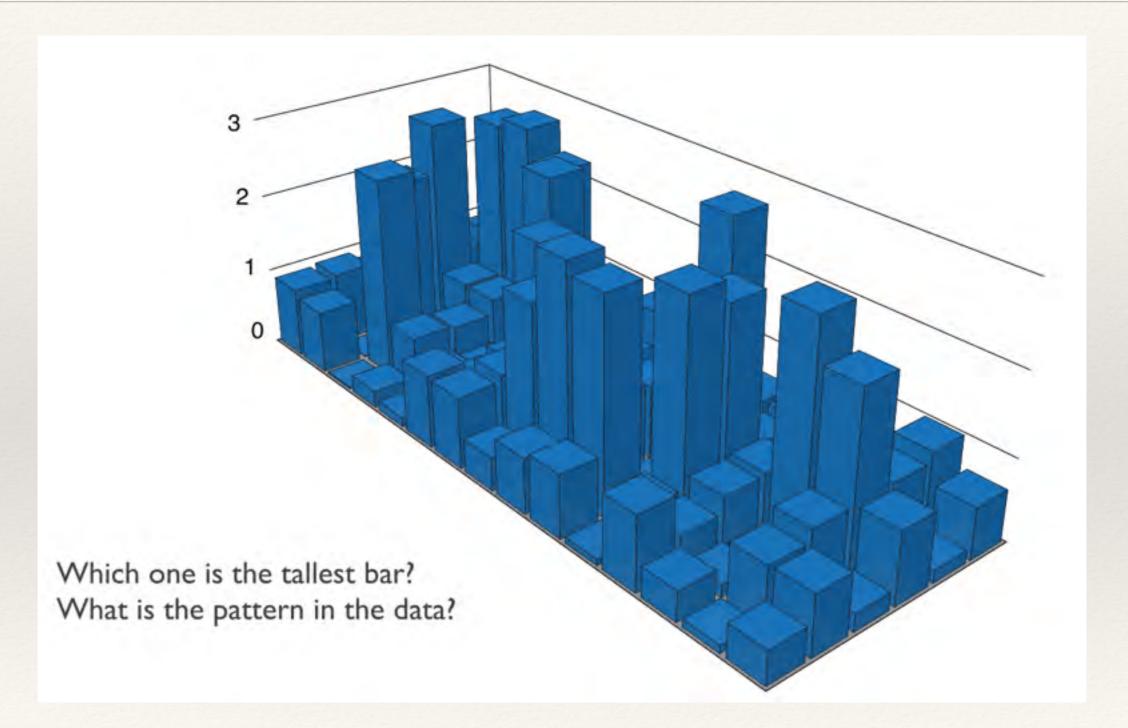






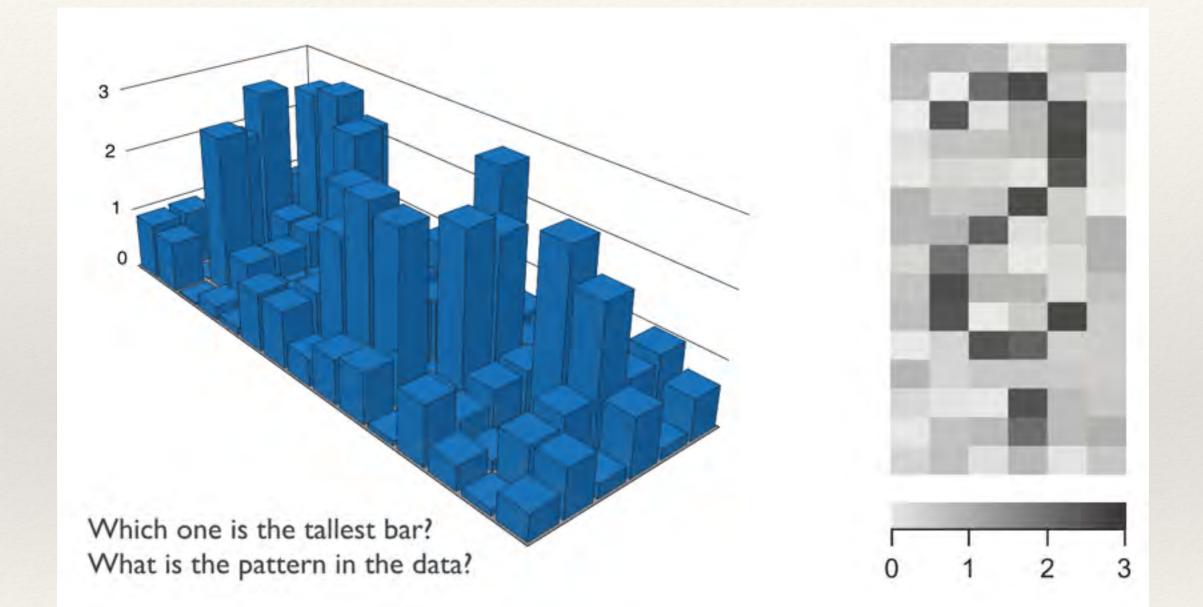
Designing Pixel-Oriented Visualization Techniques: Theory and Applications Daniel A. Keim, 2000

# 3D Pitfall: Occlusion & Perspective



Gehlenborg and Wong, Nature Methods, 2012

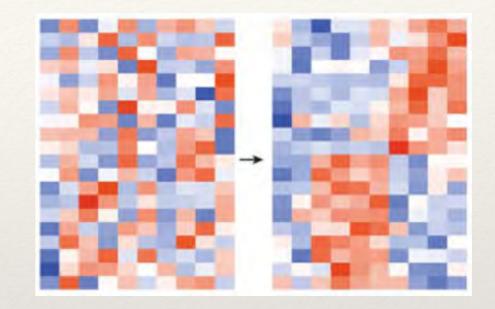
# 3D Pitfall: Occlusion & Perspective

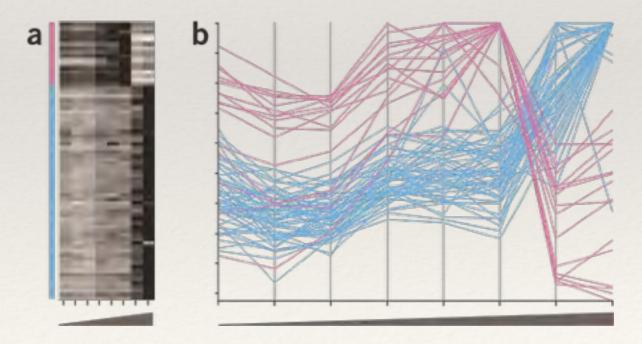


Gehlenborg and Wong, Nature Methods, 2012

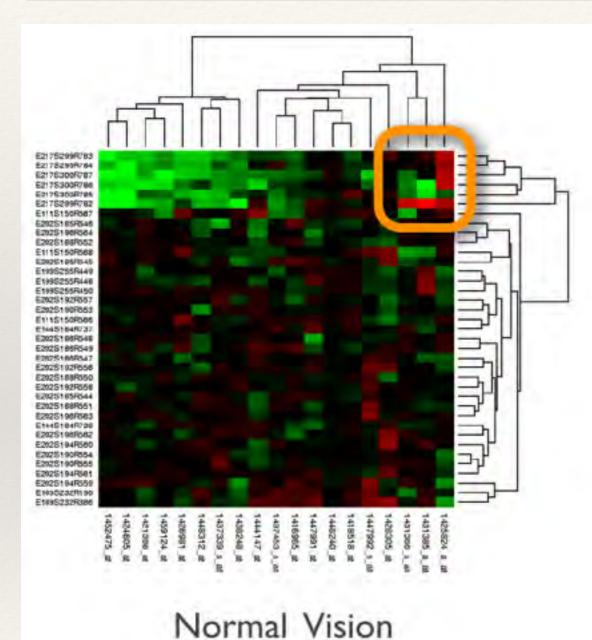
# Pixel Based Displays

- 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**



E2175299R783 E21782998784 EZ:753008787 E2175300R788 E21753568785 E2178299R782 E1118150R567 E202S185R548 E2025196F564 E2025108R052 E111S150R688 EDIDE1057645 E190825588449 F1995255R449 E199S256R450 E202S192R557 E2025190R553 E1115150R500 £1445154703/ E202S1#8R548 E202S166R549 E20251#6R547 E2%2S1v2R556 E2025188R650 F2025192R558 E2025165R044 E2025188R551 E2025196R563 E1+4E184R728 E2025198R682 E2025194R560 E20251907854 E202S190R655 E202S104R501 E2025194R559 E1490232/0190 E1698232R388 1448312 # 1438248 1444147 # 1431385 1437339 1418965 147992 5 8 1401060 1425824 4269351 2 447991 418518 428305 459124 #

> **Deuteranope Vision** ("Red-Green Blindness")

おうちに

2.04

8 'n. 8 15

.

Sa.

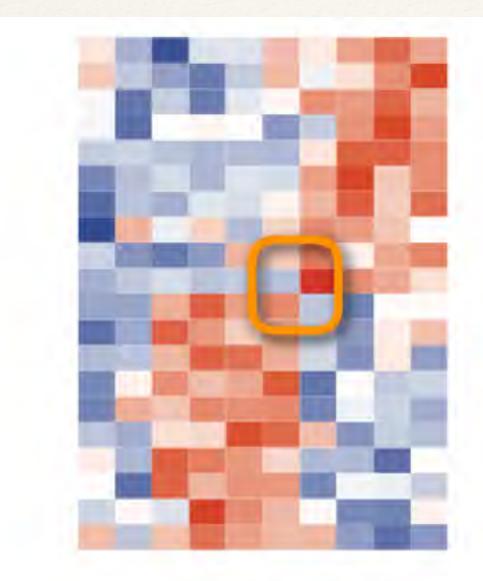
418240

8 1 a a -

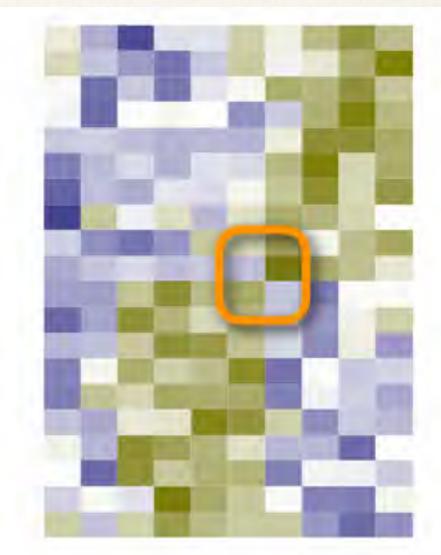
452475 424605 4138

84 6 5

# Good Color Mapping

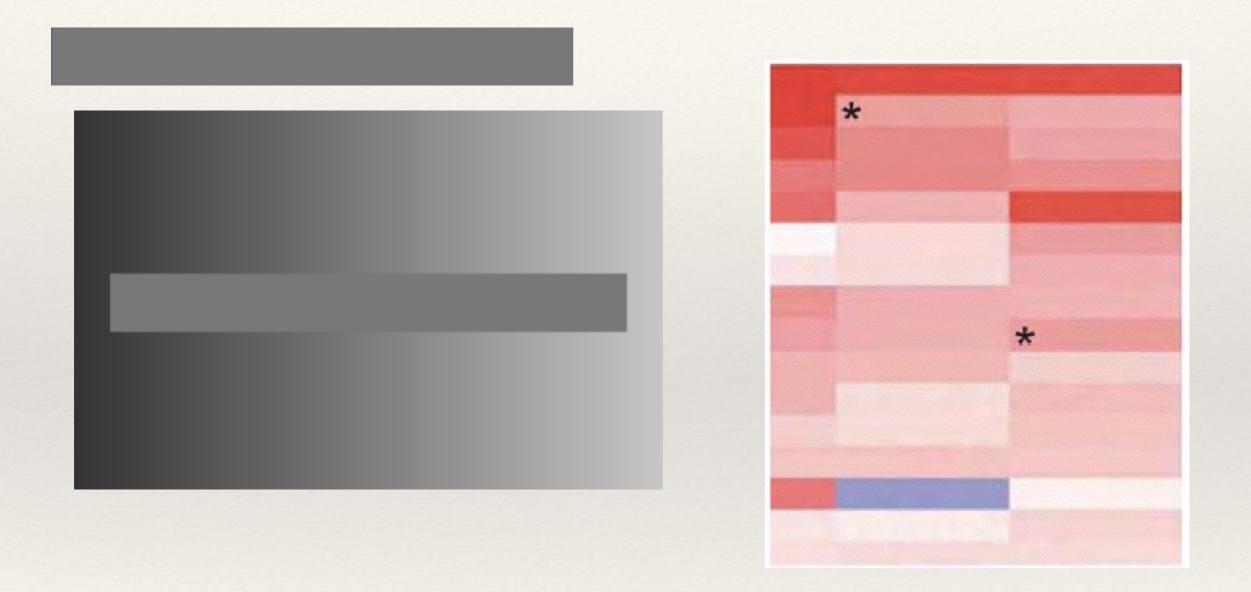


Normal Vision



Deuteranope Vision ("Red-Green Blindness")

### 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

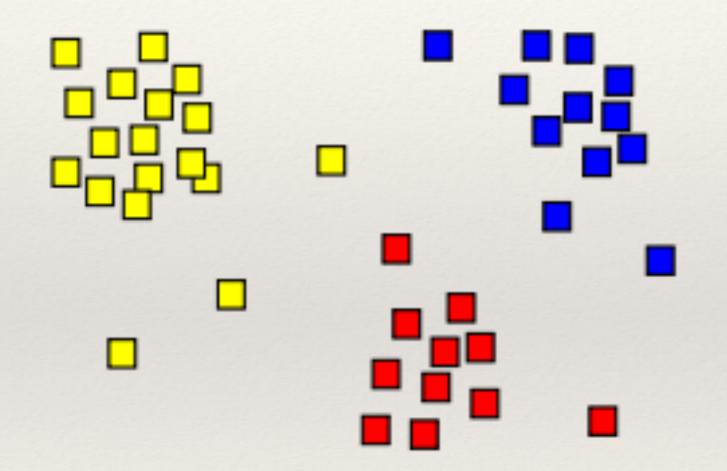
#### **Bi-Partite**



# 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

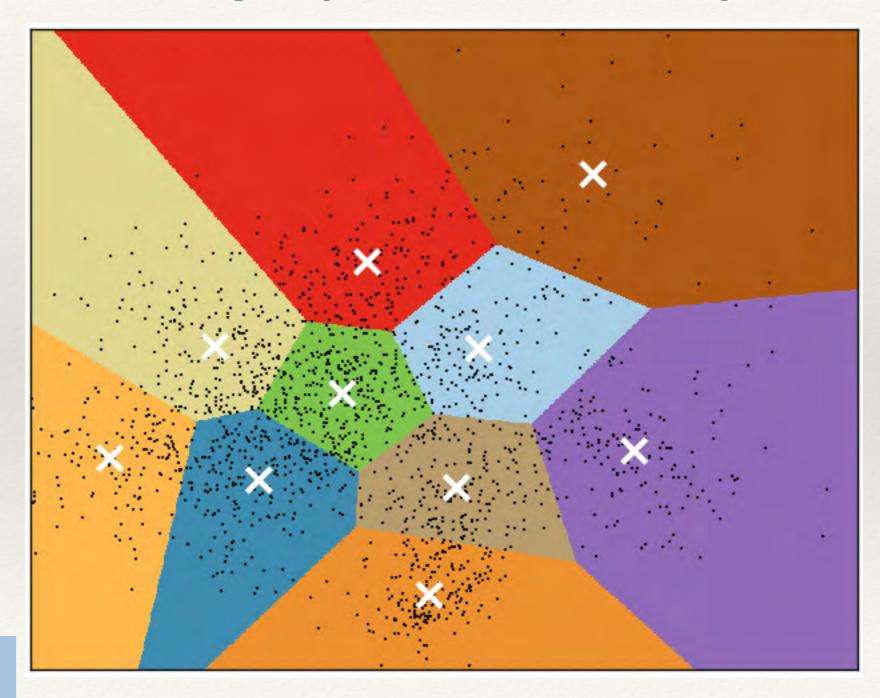
# Partition Clustering



Each Point in a Unique Class

# Centroid Based Clustering

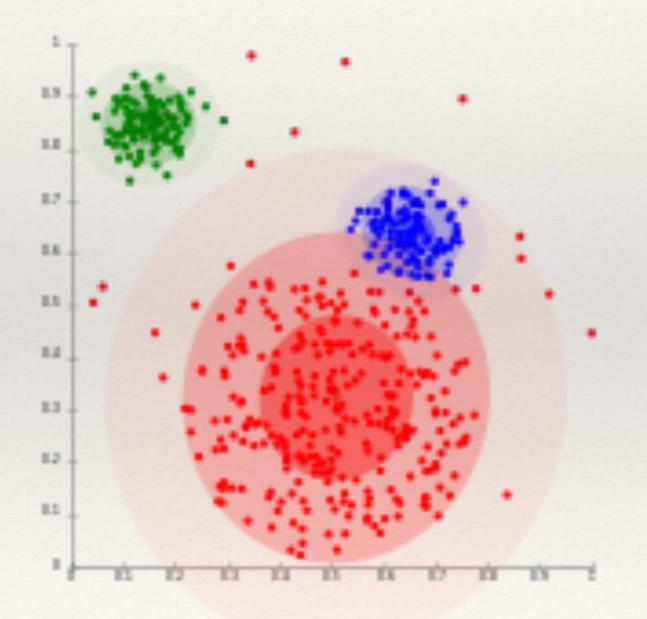
Example Algorithm: K-Means Clustering





# Distribution Based Clustering

Expectation Maximization (EM)

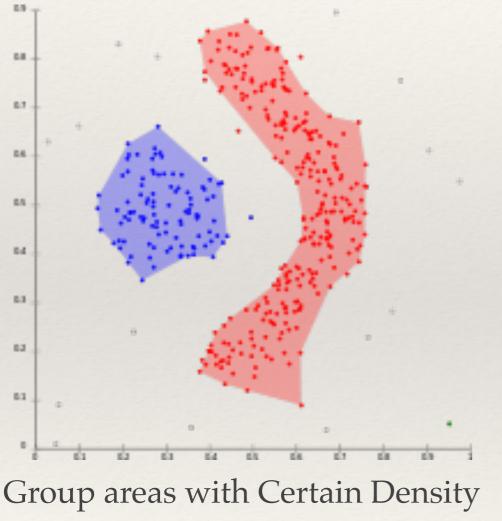


### Partition

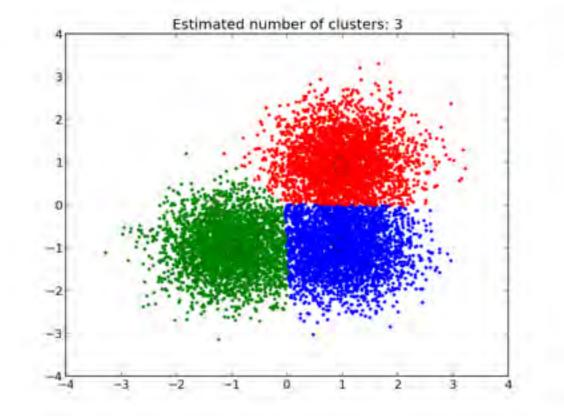
# Density Based Clustering

#### DBSCAN

#### Mean Shift

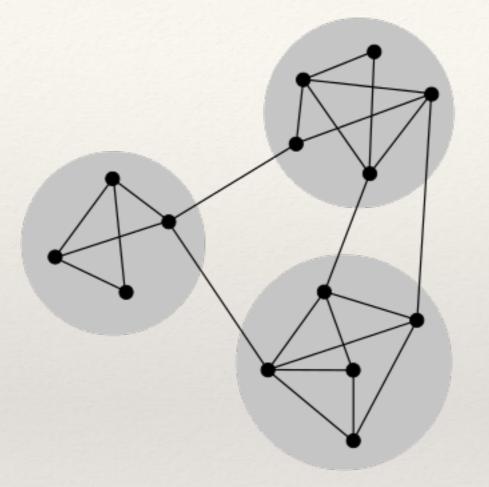


OPTICS (more general)



### Partition

# Simple Graph Cutting Methods

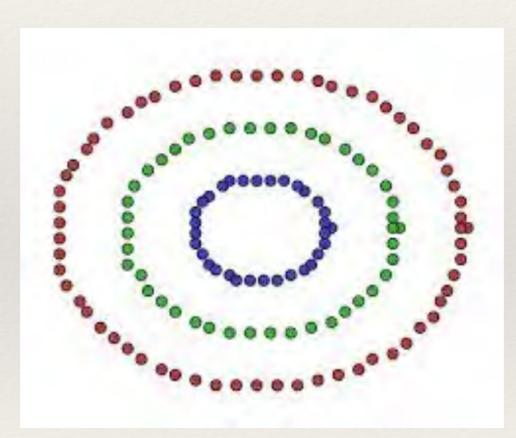


- (1) Similarity Score
- (2) Pick Edge Threshold
- (3) Cut (only connect stronger edges)
- (4) Compute Connected components



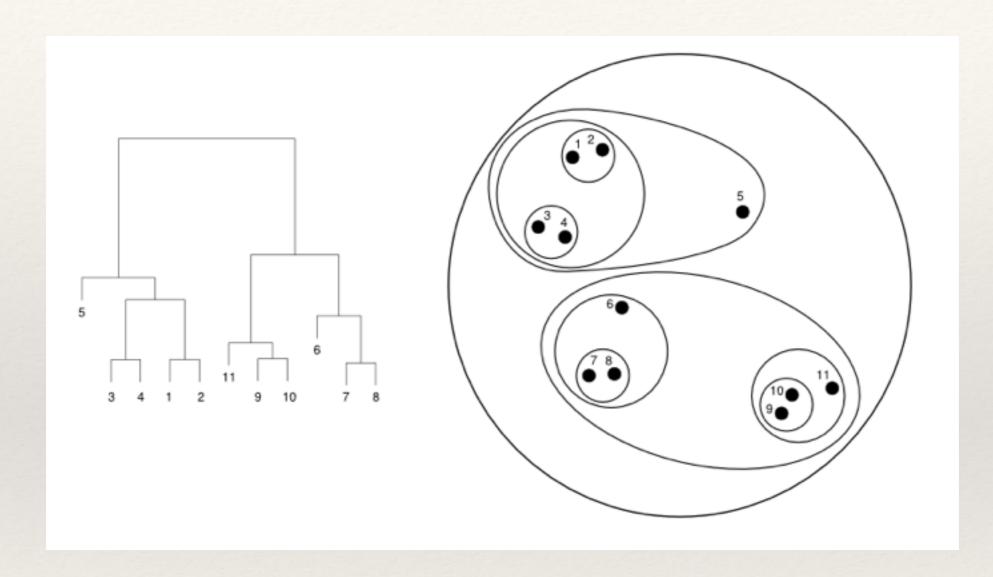
# Graph Cut Based

### Spectral Clustering





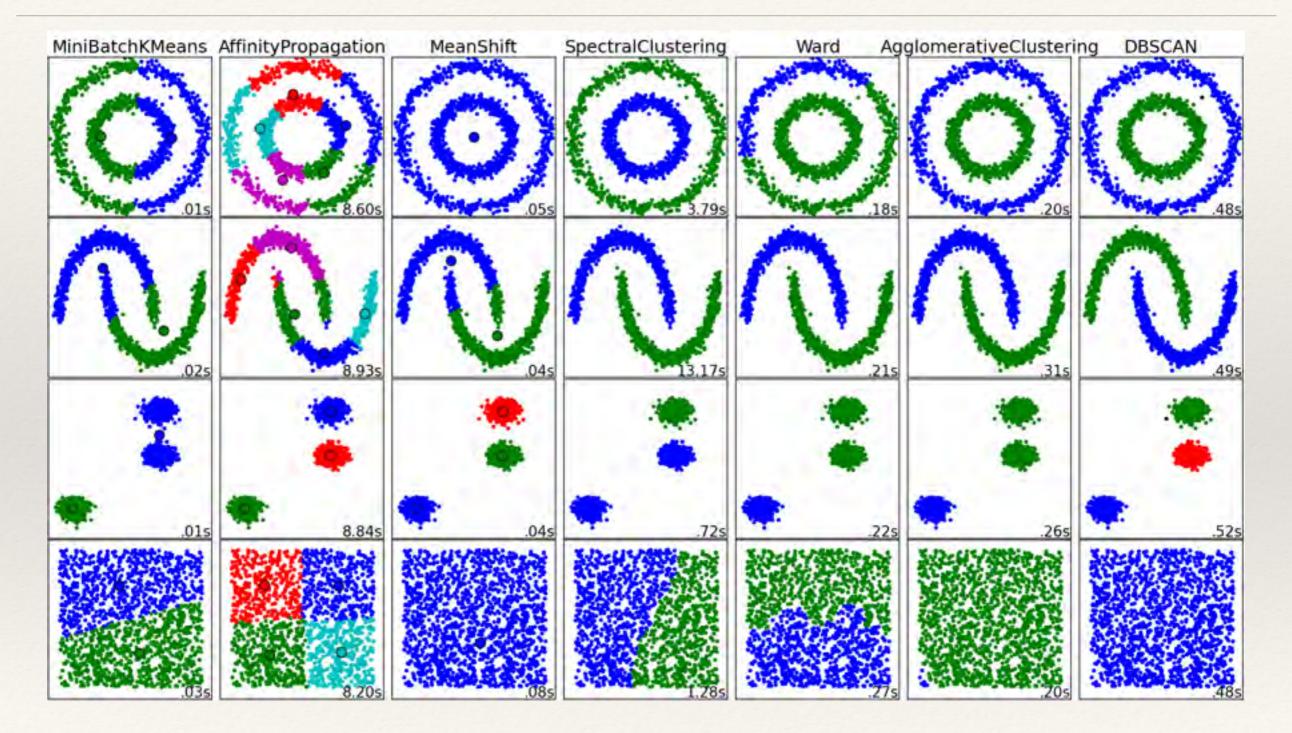
# Hierarchical Clustering



Example: Ward Clustering

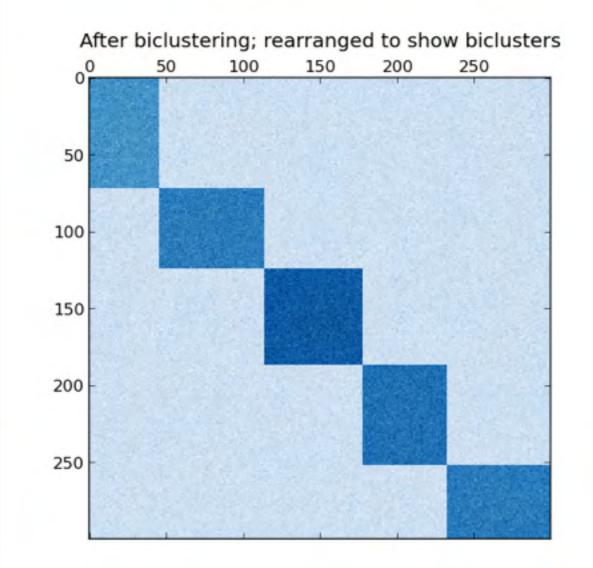


# Comparison



Different Clustering Based on Different Assumptions

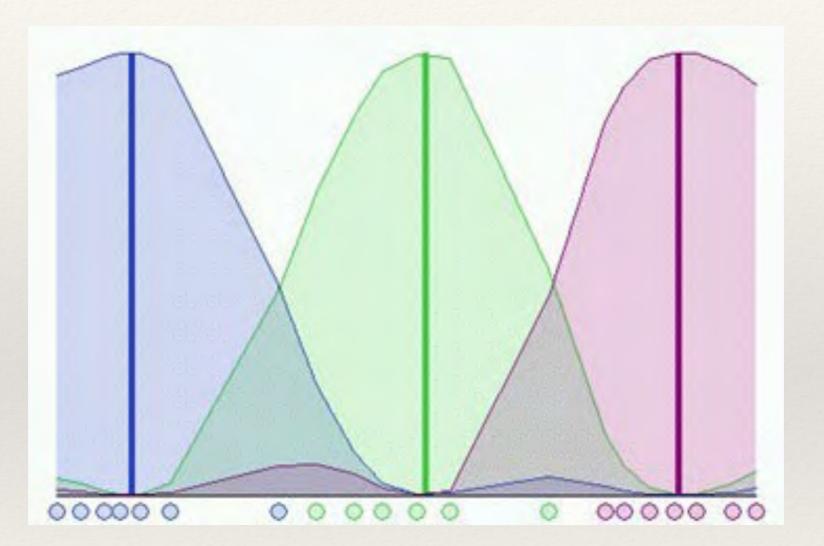
# Bipartite Clustering (Biclustering)



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
- Aggregation

Titanic Survivors

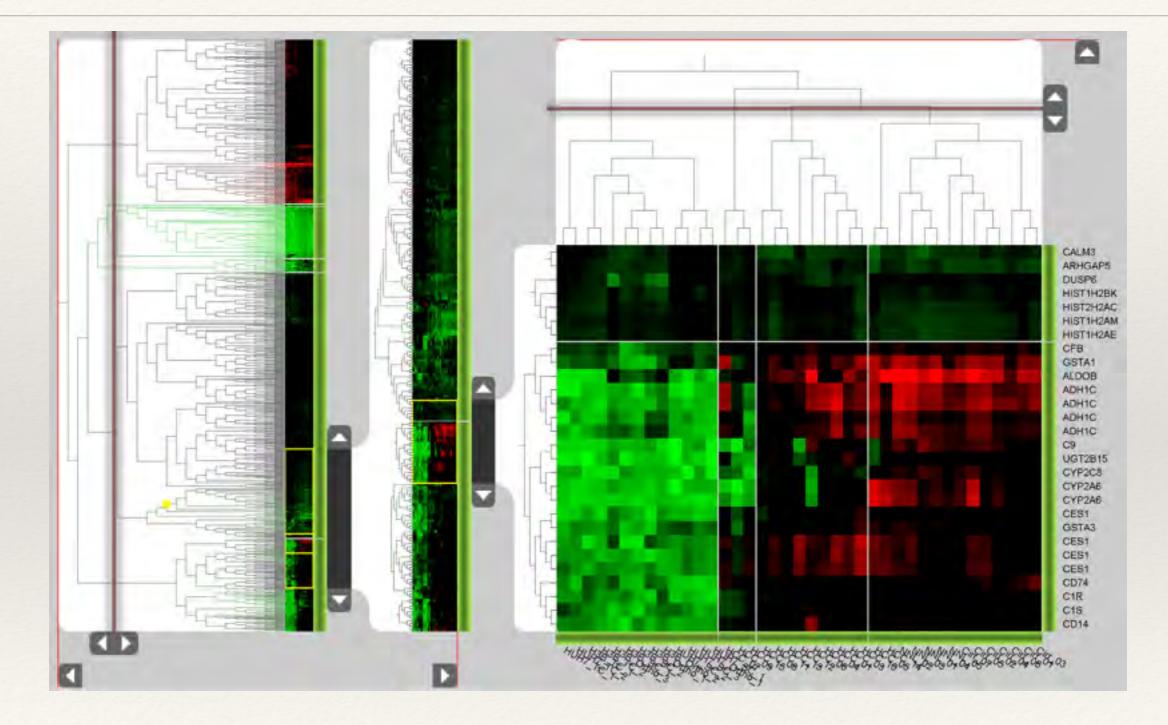
http://www.jasondavies.com/parallel-sets/

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

# **Clustered Heat Map**

			Martin State	Carlos and a second second	
	A statement of the stat	ing and spectrum dates and	The later of the second	the strange of the st	And the property of the second s
	Mainter	and the state of the second		Berther Mark	State
Neural	The second second	and the first of the second		Republic and a start of the second se	
	a share we as he had a share a sha		State of the second second second second	and the second s	Look I want to be a series
		the second se			
and the second	Statement and the second		DECEMPTO	A CONTRACTOR OF A DESCRIPTION	une l'an anna baine anna
	2012	and the second	and succession and the		Standard March
Cassical		An application of the second second second	Sector States	PERSONAL PROPERTY AND A	The state of the state
REAL STR	were strain a local	search and service and the	TE PROPERTY IN	States and the second second second	and the second in the second second
1.1.1.1.1.1.1.1	Secondary and the second	\$2.444 (A) # 17.1 (A) # 1	E.P.	Contraction of the State of the	E E HULL AND
E 1 2 2 2	State of the state of the state of	Sugar States and states a	Million and	Printerio di Anto del Constanti da	The second second second
	In In Adversion even with a role	CONTRACTOR DESCRIPTION PERMIT	Barren anne anan a - a	Full Adding Schule and Schule and A	THE PACE METAL
LIS LA VEL	al measure destruction of the			and the second of the state of the	Distanting the state manter
CALL C				Contrast of the second states and the	and with them in
			and the state of		E State of the second
Mesenchymai 160 mile de Rodenseel		Super Constant of the State	The state of the	and the second s	and the real of the second
<b>小村</b> 村村村村	and the second of the second second	建理 和公司的 新市	ALL ADD ALL	Frankis - Briefferen - To y	and the transfer of the second
					the state of the s
ALL BRIDE	「「「「「「「「」」」	國南部的有利的	部省组织		and the second second
1 H. H.	SALS, THE STREET			and the second second	Provide and the second
1 Sterrow	the second second				
or fair has anno	LAN D COMPANY AND A STORE OF		Strange al	Market Strate .	Provide Victoria
S. S. S. S. COL	Station in the sec	至1451560.000°C。100	Station of the	and the state of the local of the	1 Sector
Proneural		A Contraction of the	water and the support of the sup-	See 10 - See Anno 1 - State See Anno 1	A STATE OF A
The Part of the The	10-1-1-10 C		and the second		Entry This
- 医二乙二至	FR. F. T. March	And the second sec	man faith 1 1	Distance in the	
State Carlow Mar	and the second second			Cherry Mr. Physics Press	Service and a

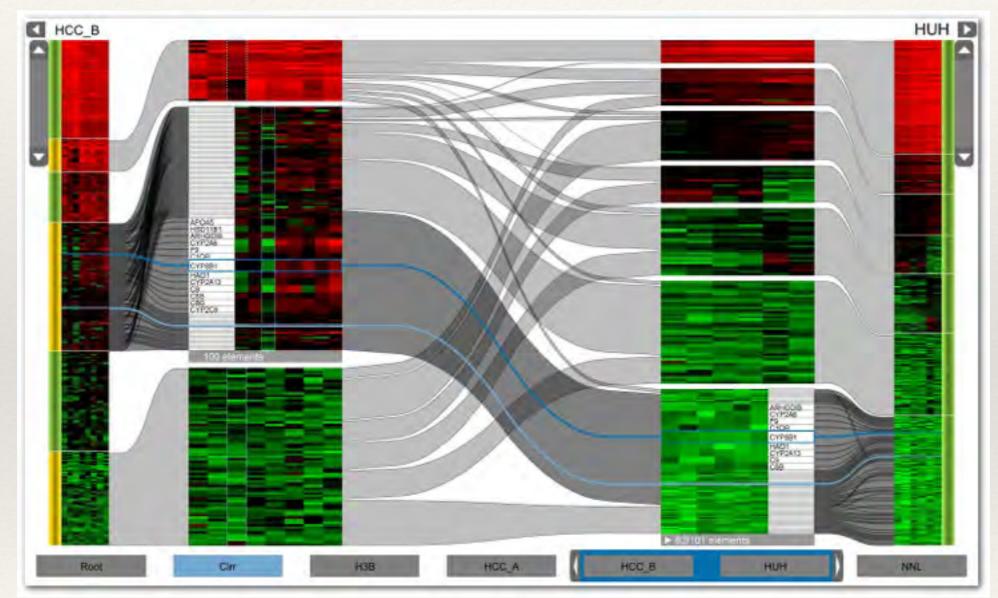
# F+C Approach, with Dendrograms



Lex, PacificVis, 2010

# Cluster Comparison

#### Caleydo Matchmaker

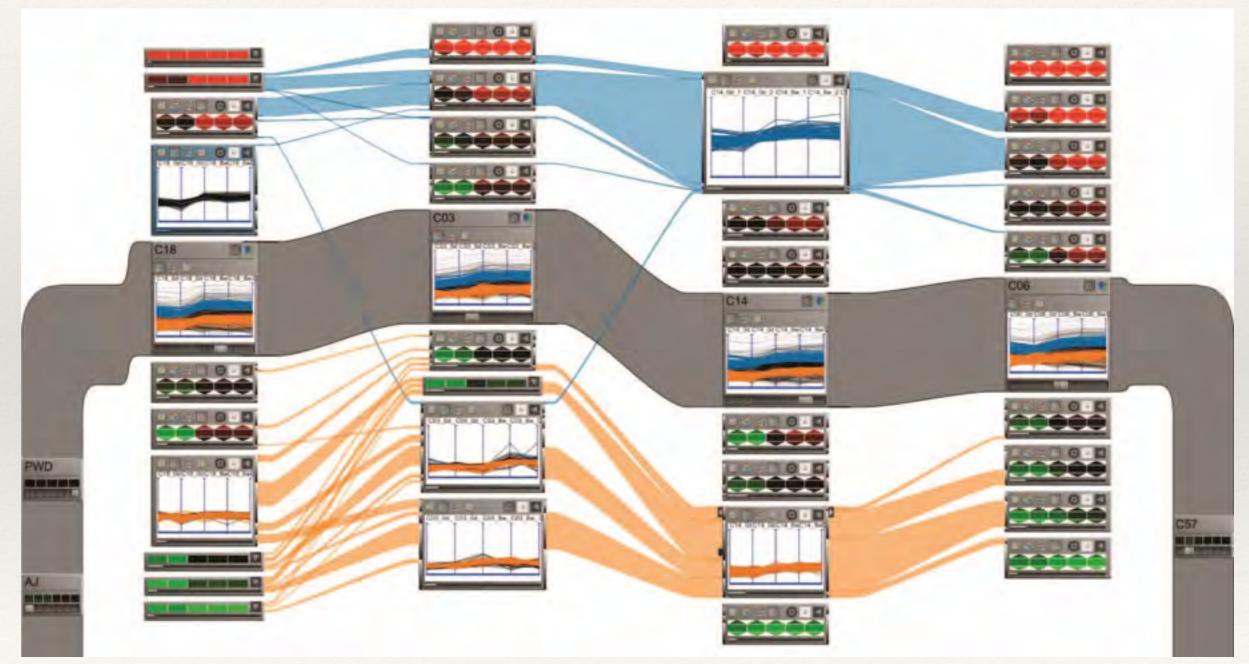


Lex, Streit, Partl, Kashofer, Schmalstieg 2010

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

# Aggregation

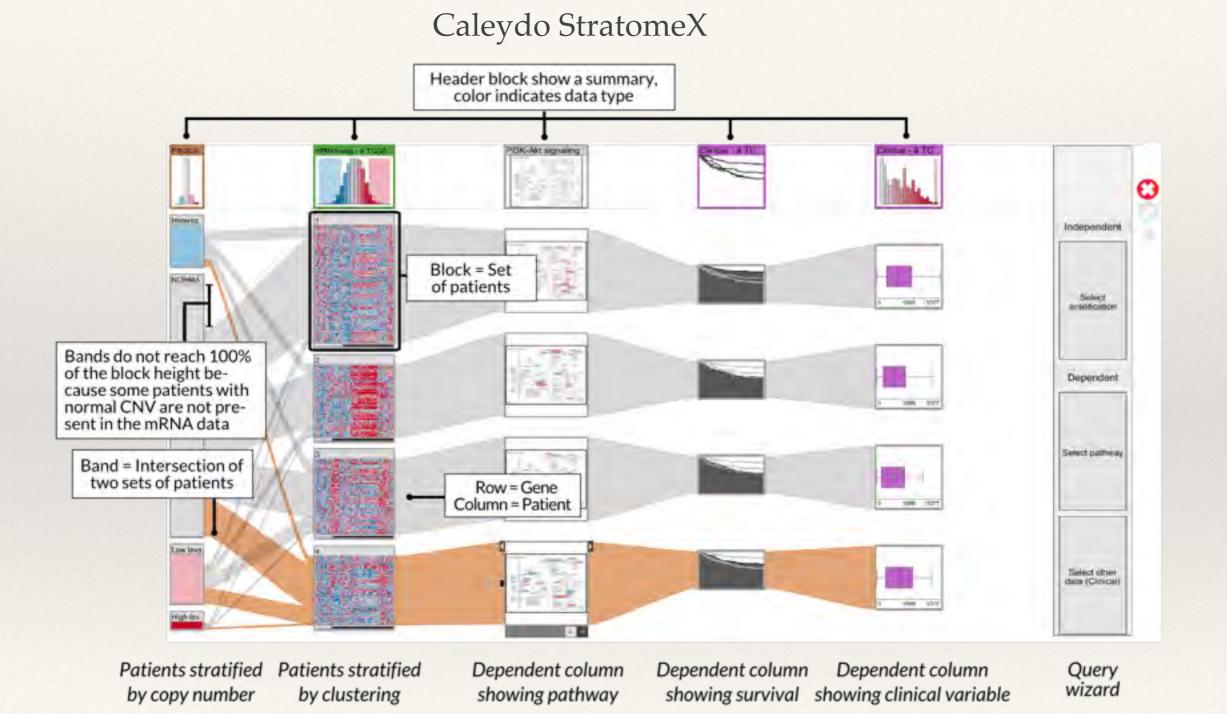
#### VisBricks



http://people.seas.harvard.edu/~alex/papers/2011\_infovis\_visbricks.pdf

# Heterogeneous Data

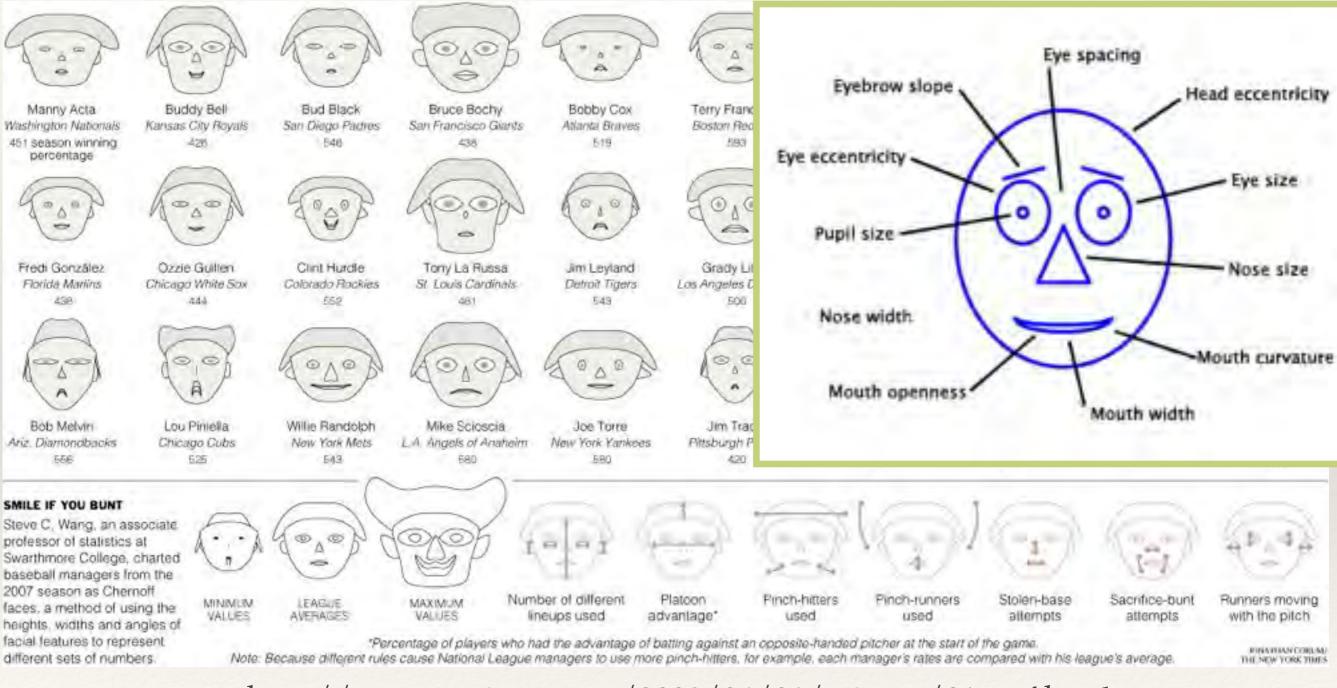
# Heterogeneous Variables



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



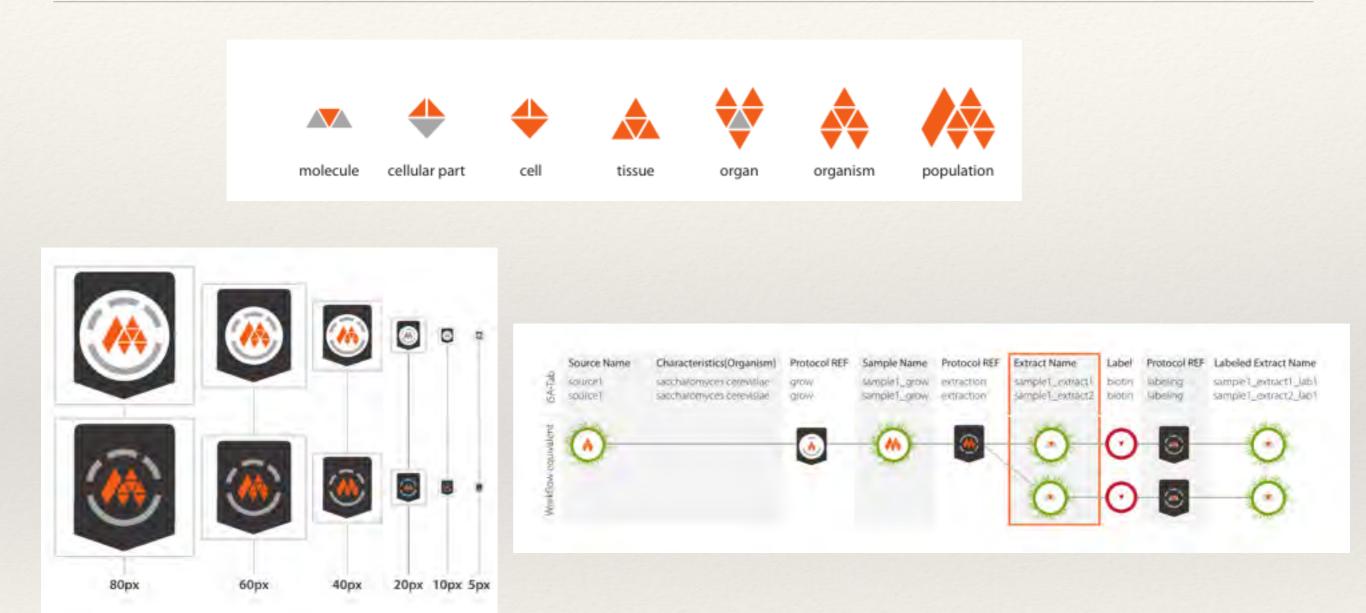
### **Chernoff Faces**



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

Fail?

# Complex Glyphs for Bio Workflows



Maguire 2012