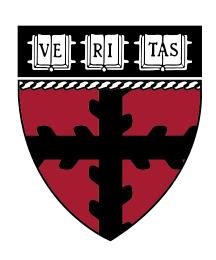
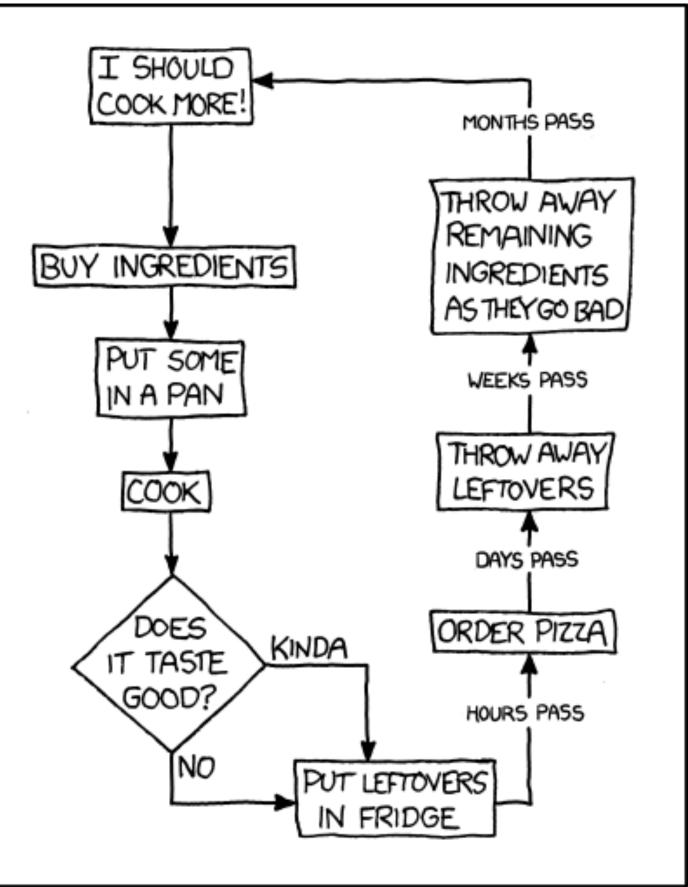
CS171 Uisualization

Alexander Lex alex@seas.harvard.edu

Tables Part II



HARVARD School of Engineering and Applied Sciences



Next Week

Reading: VAD, Chapters 9 Lecture 11: Text & Documents Lecture 12: Homework 3 Design Studio Sections: view coordination, linking & brushing Updates Design Studio moved to Thursday Project Proposal moved to HW 4

Tables & Multi-Dimensional Data

Comparisons

Direction

Graduation rates up in most cities

2005 RA

- Philadelphia 62.
 - Tucson 71.
- Kansas City, Mo. 53.
 - El Paso 60.
 - Portland, Ore. 68.
 - New York 50.
 - Dallas 50.
 - Columbus 44.
 - Mesa 76.
 - Austin 58.
 - Atlanta 43.
 - Fort Worth 56.
 - Miami 55.
 - Houston 52.
 - Chicago 51.
 - Oakland 50.
 - Virginia Beach 68.
 - Baltimore 41.
 - Denver 58.
 - Detroit 37.
 - San Antonio 47.
 - Phoenix 58.
- Oklahoma City 47.
 - Indianapolis 30.
 - Milwaukee 41.
 - Sacramento 62.
- Washington, D.C. 57.
- Colorado Springs 68.
 - Honolulu 67.
 - Nashville 45.
 - Jacksonville 50.

SOURCE: EPE Research Center

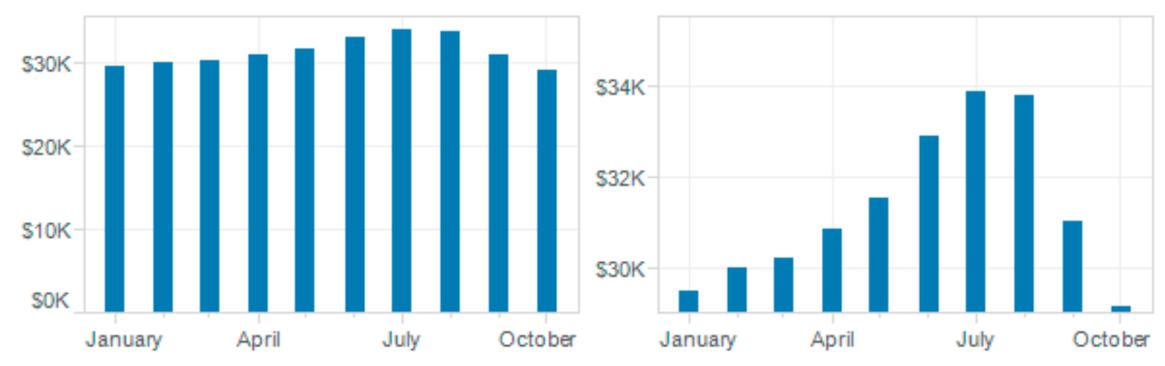
Graduation rate for principal school district of the largest cities

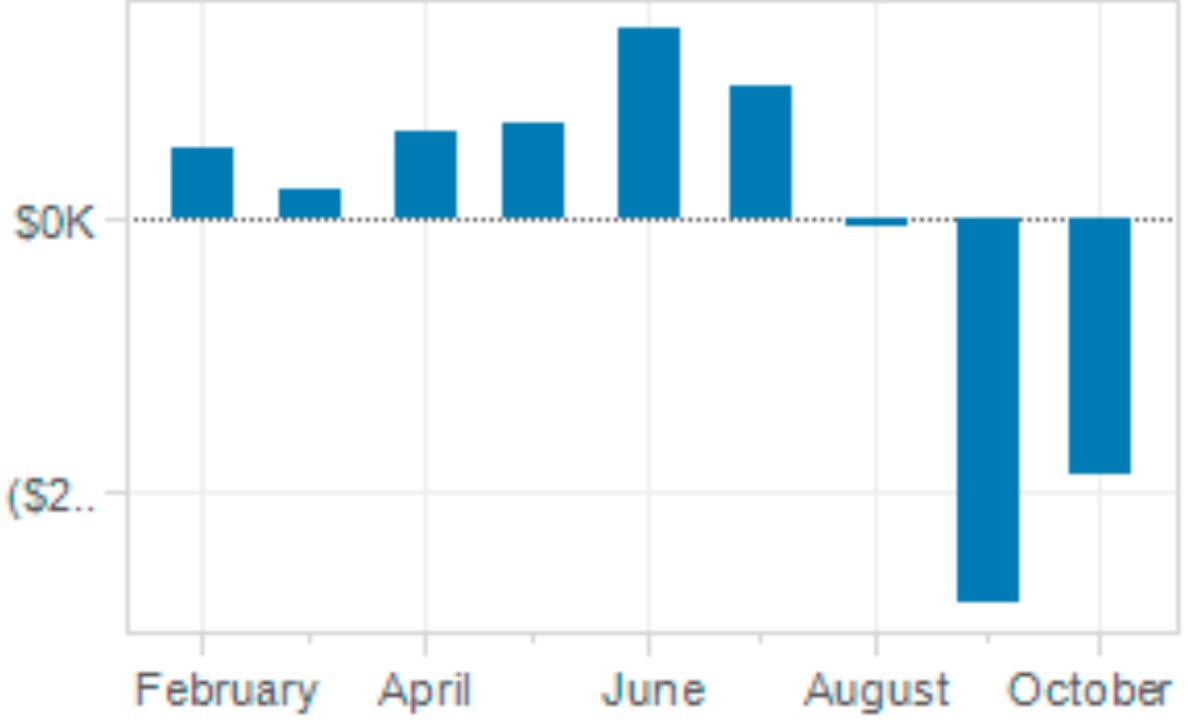
-	-			-	
ATE	1995-200	05 CHANGE	-	he eur	reas bish
.1%		23.2%			erage high
.6		22.1		-	lation rate
.3		1.0.1			was 54.7
.6	13				Of the 62
.6	13.				improved
.5	12.	.8 sind	ce 19	995, Ph	iladelphia
.8	12.	7 had	the I	highest	increase.
.7	12.	6 7	The r	ate in l	.as Vegas
.6	12.0	0	dec	reased	the most.
.9	11.5	j.			
.5	10.8	1995-2	005	2005	
.5	10.4	CHAI	NGE	RATE	
.9	10.4	-0	.3	63.4%	Louisville
.9	9.8	-0	7 🛔	68.9	Seattle
.0	9.2	-1	.2 🚺	51.2	Memphis
.5	9.2	-1	.5 📕	51.9	Fresno
.5	8.8	-1	.7 📕	58.6	Boston
.5	7.7	-1	.7 📕	45.3	Minneapolis
.6	6.9	-1	.8 📕	73.3	San Jose
.5	6.9	-2	.0 📕	48.5	Tulsa
.3	6.4	-2	.3 📕	60.5	Charlotte
0.0	5.6	-2		63.7	San Diego
.0	5.3	-3.	6	44.4	Los Angeles
.5	5.3	-3.		64.0	Long Beach
.0	5.2	-4.9			Cleveland
1	4.9	-6.5		57.1	San Francisco
.6	4.8	-6.6		49.0	Albuquerque
.8	4.6	-11.6		60.3	Arlington, TX
.4	3.6	-14.8			Omaha
.2	3.1	-17.6		54.5	Wichita
.8	0.7 -2	23.1	10000	44.5	Las Vegas

AP

Nicolas Rapp

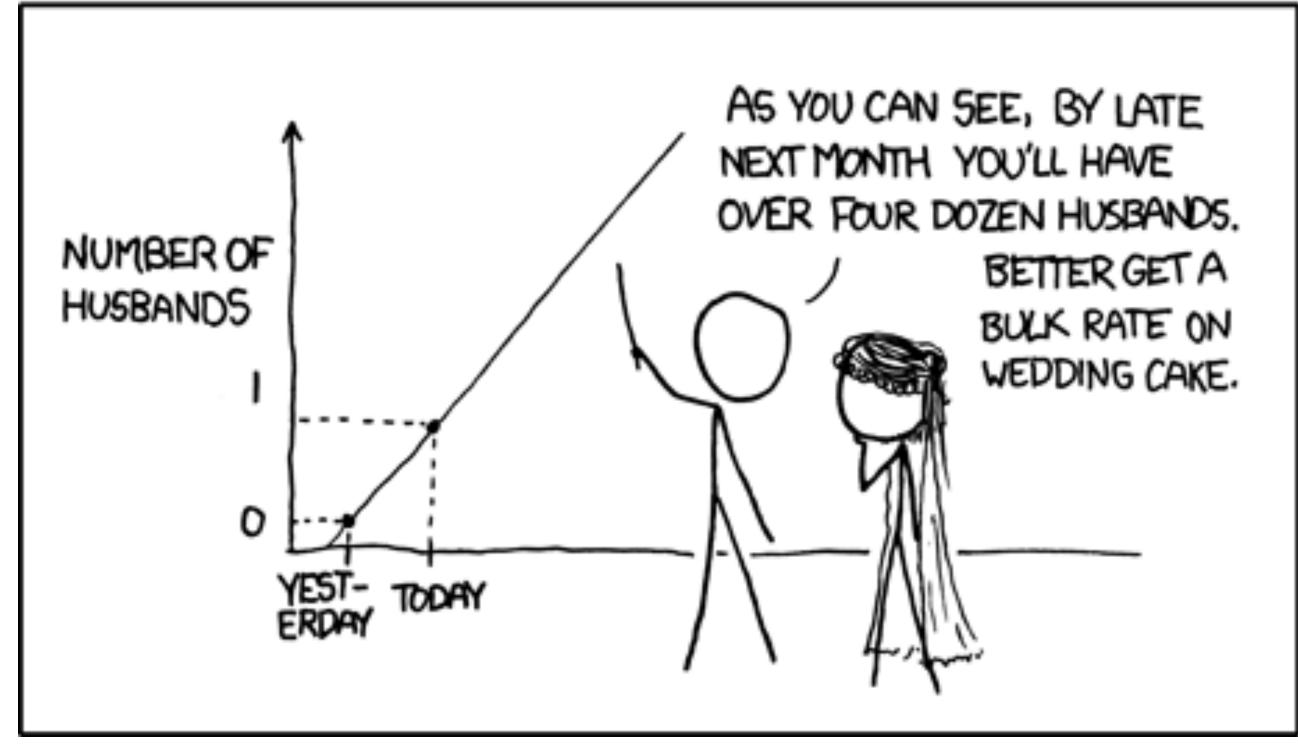
Plot Change Instead





https://eagereyes.org/basics/baselines

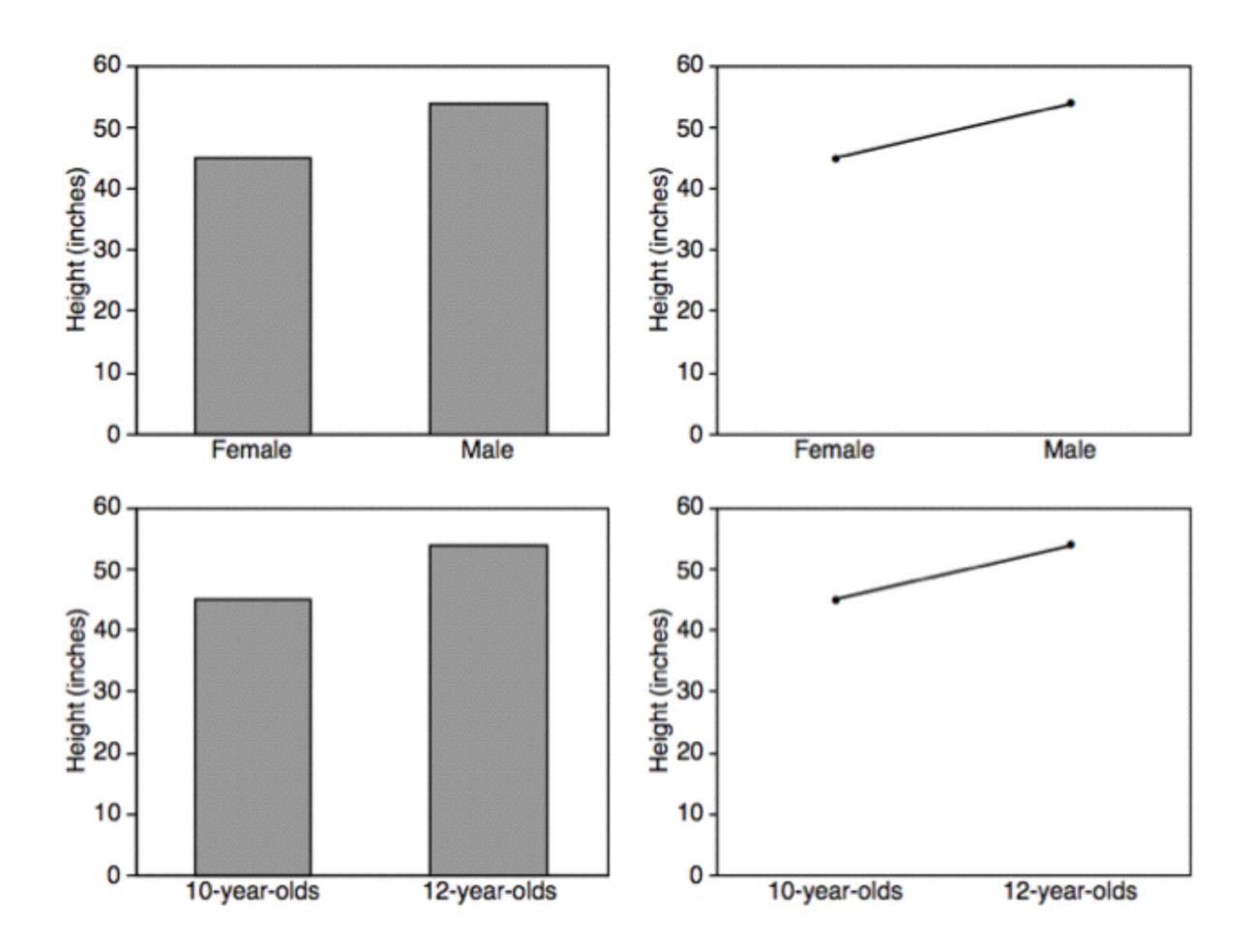
Trends Over Time My HOBBY: EXTRAPOLATING



http://xkcd.com/605/

Bars vs. Lines

Lines imply connections & sampling from continuous data. Do not use for categorical data.



Zacks 1999

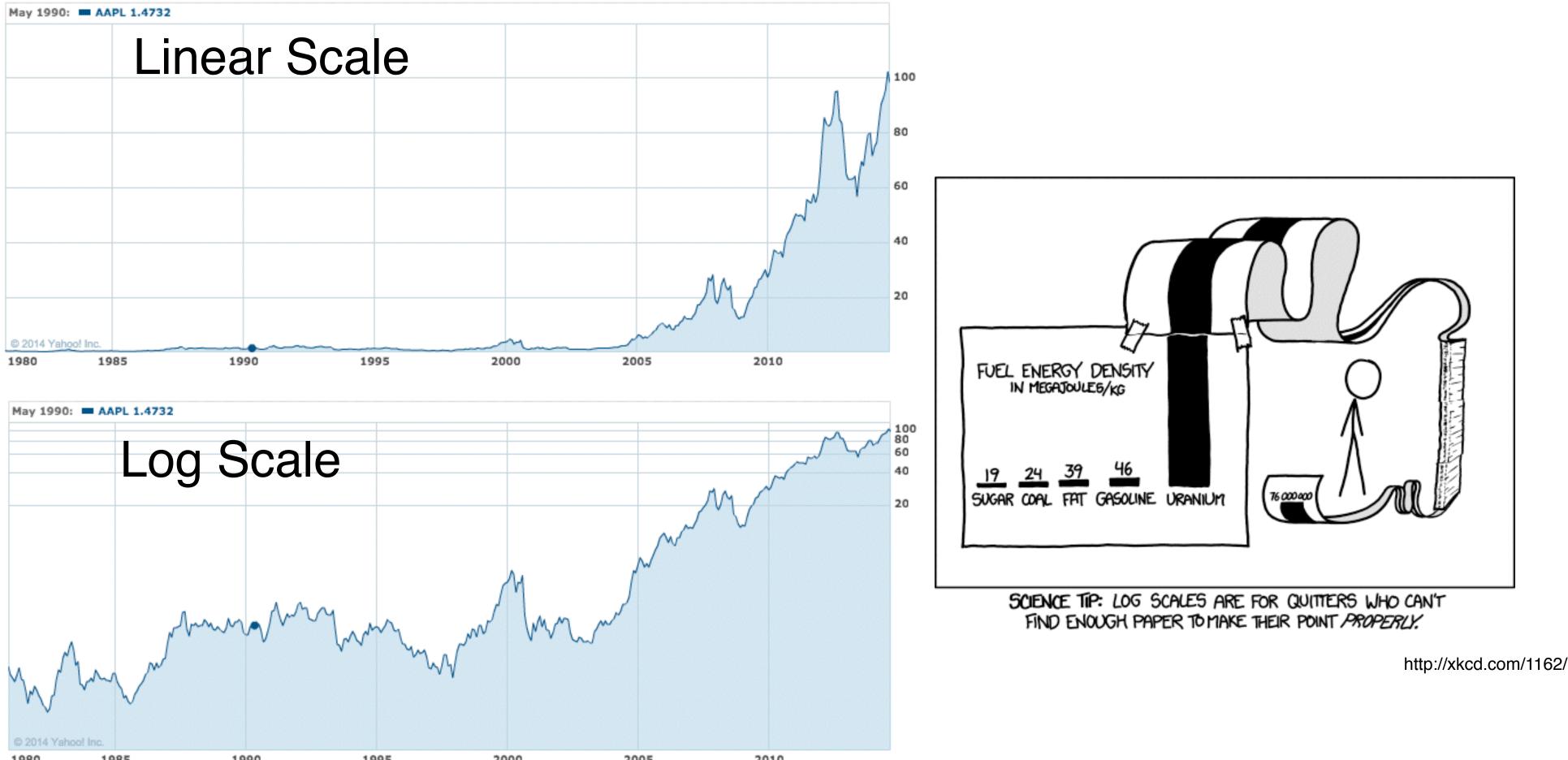
Baseline Problem (again)

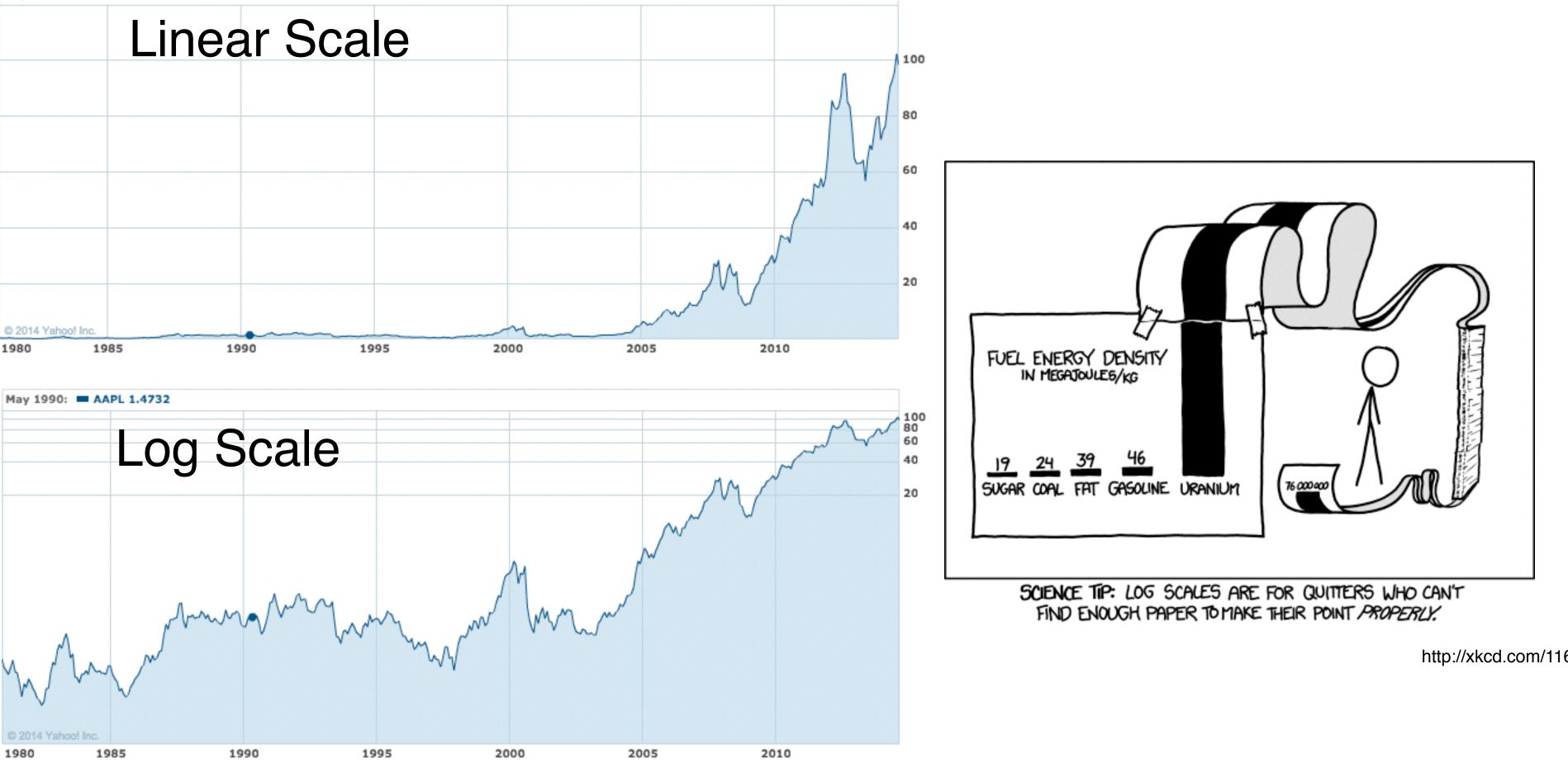




https://eagereyes.org/basics/baselines

Linear vs. Logarithmic Scale





Apple Stock Price

Aspect Ratios

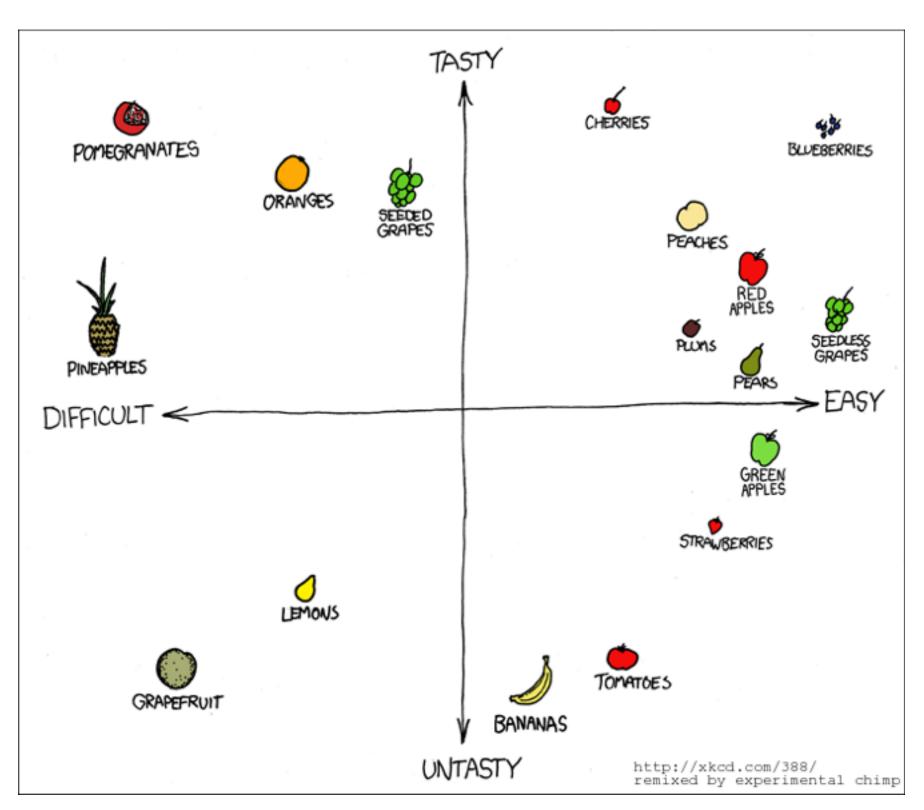
Rule of Thumb: Banking to 45° (average line slope: 45°)



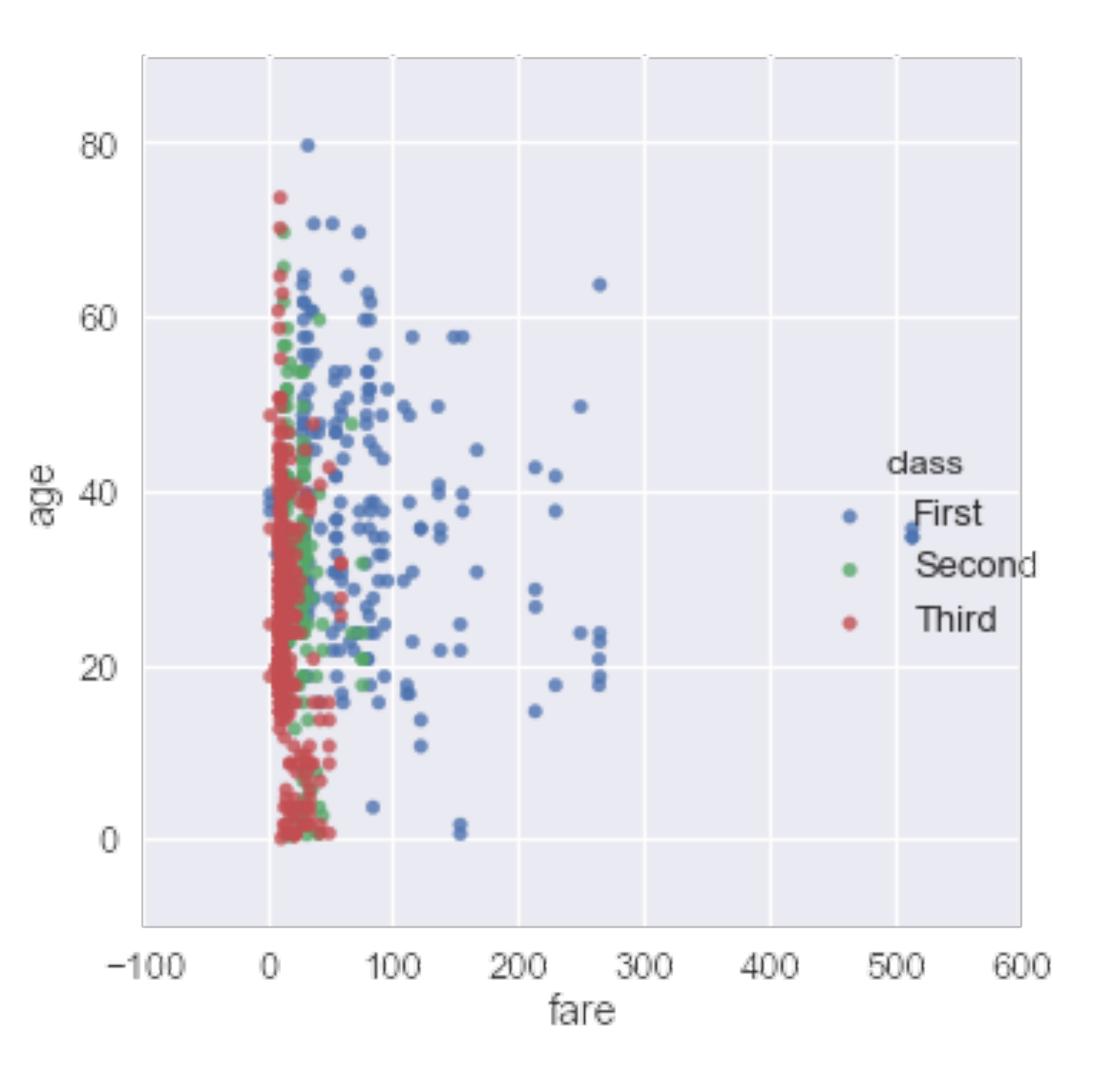


eagereyes.org

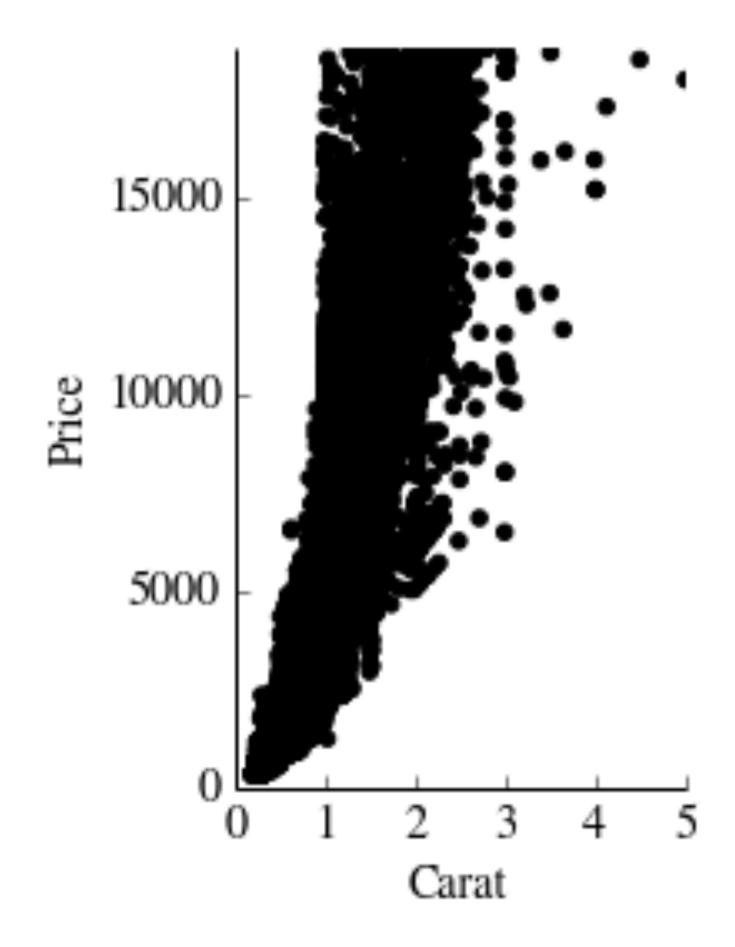
Correlations

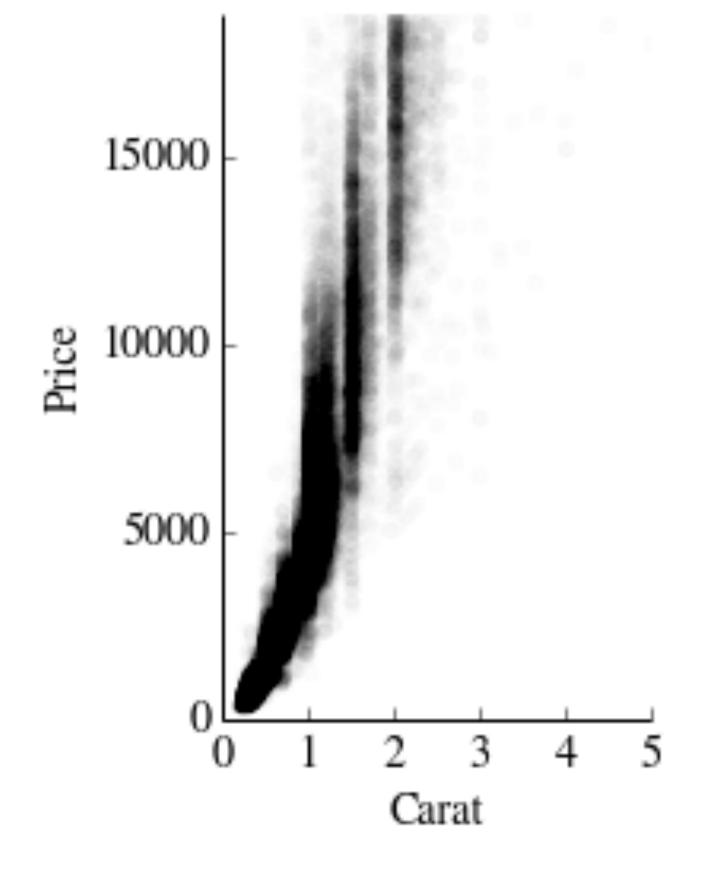


Scatterplots



Overplotting

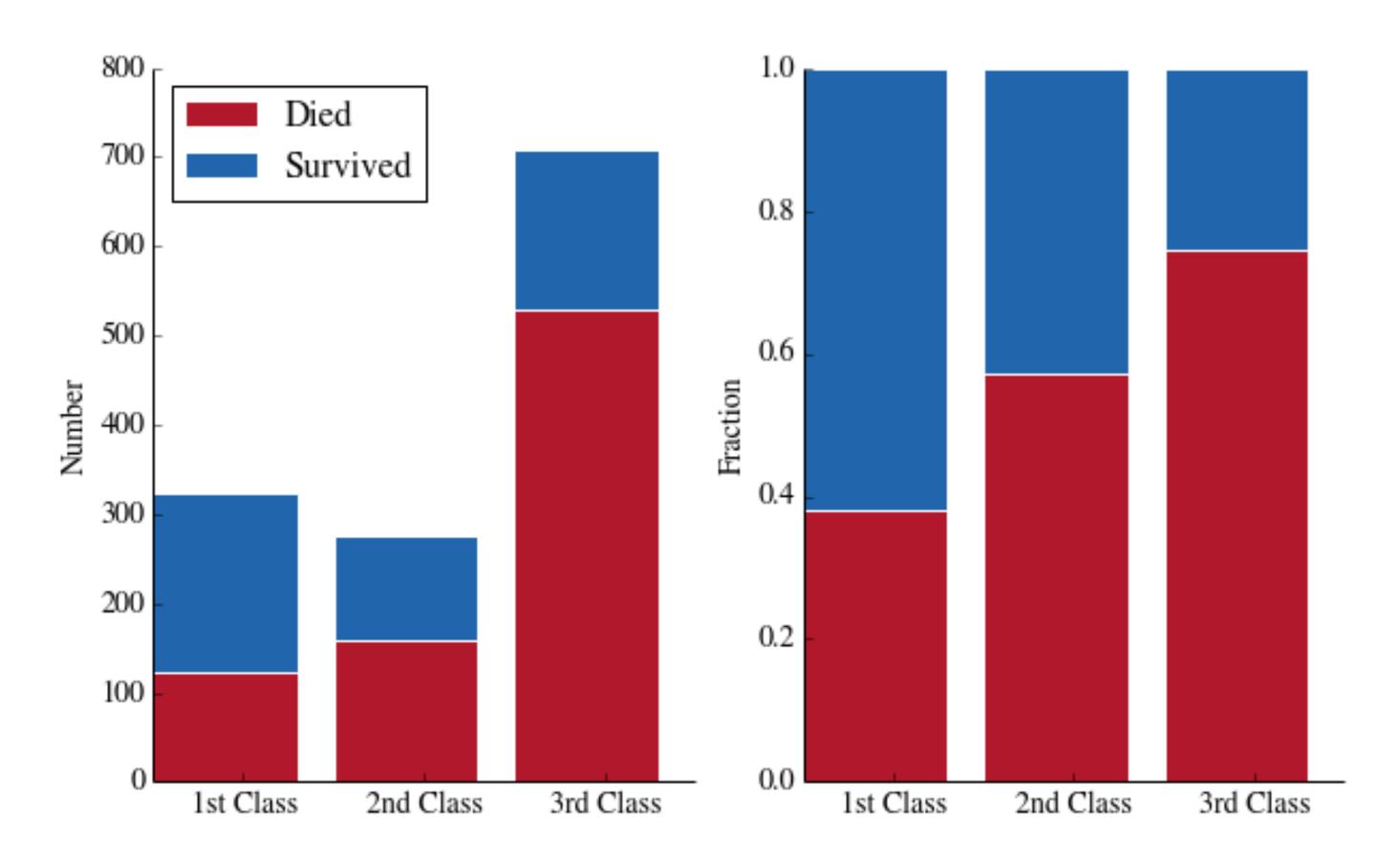




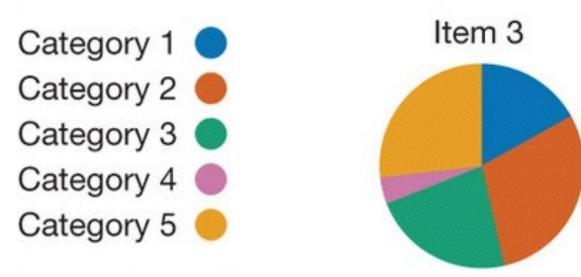
alpha = 1/100

Compositions

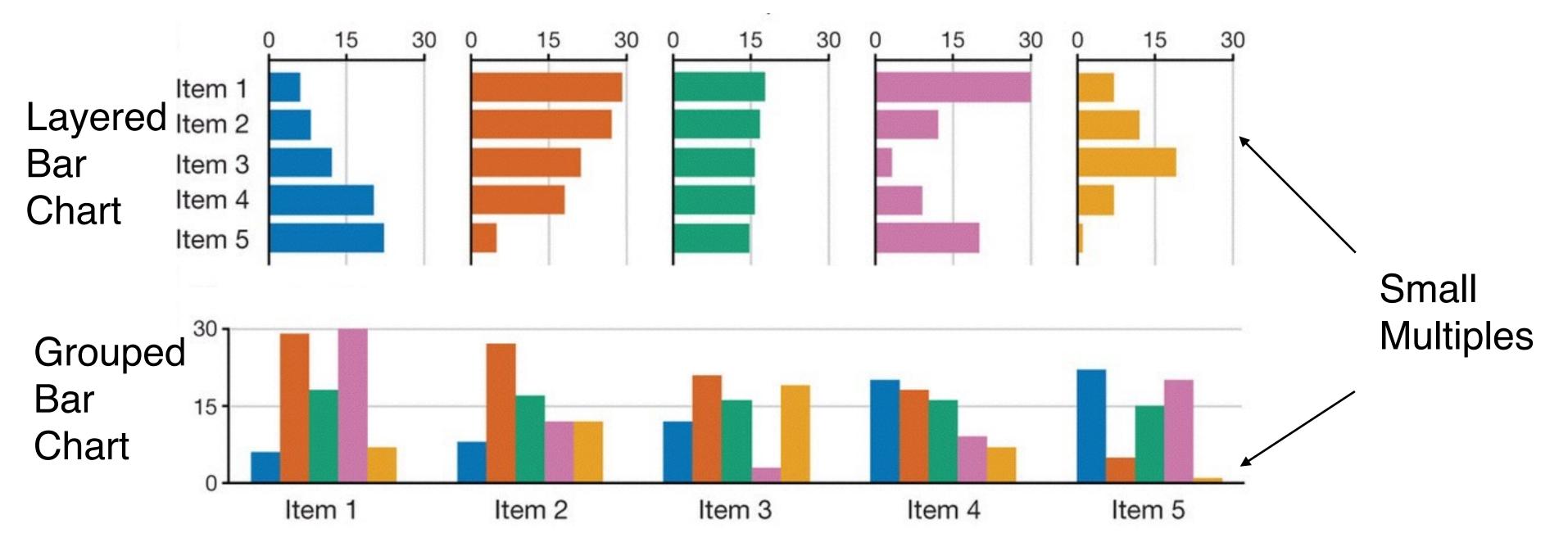
Stacked Bar Chart

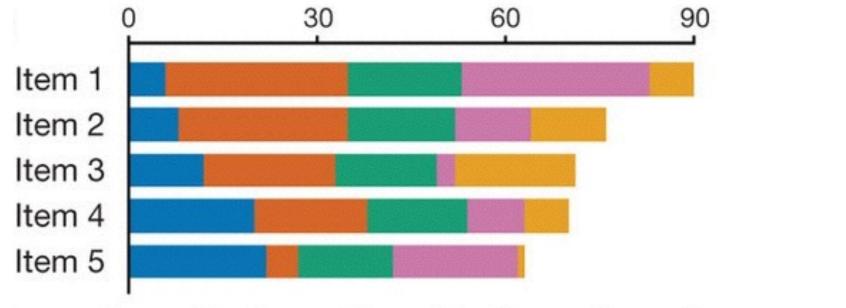


Comparison of bar chart types



Pie Chart

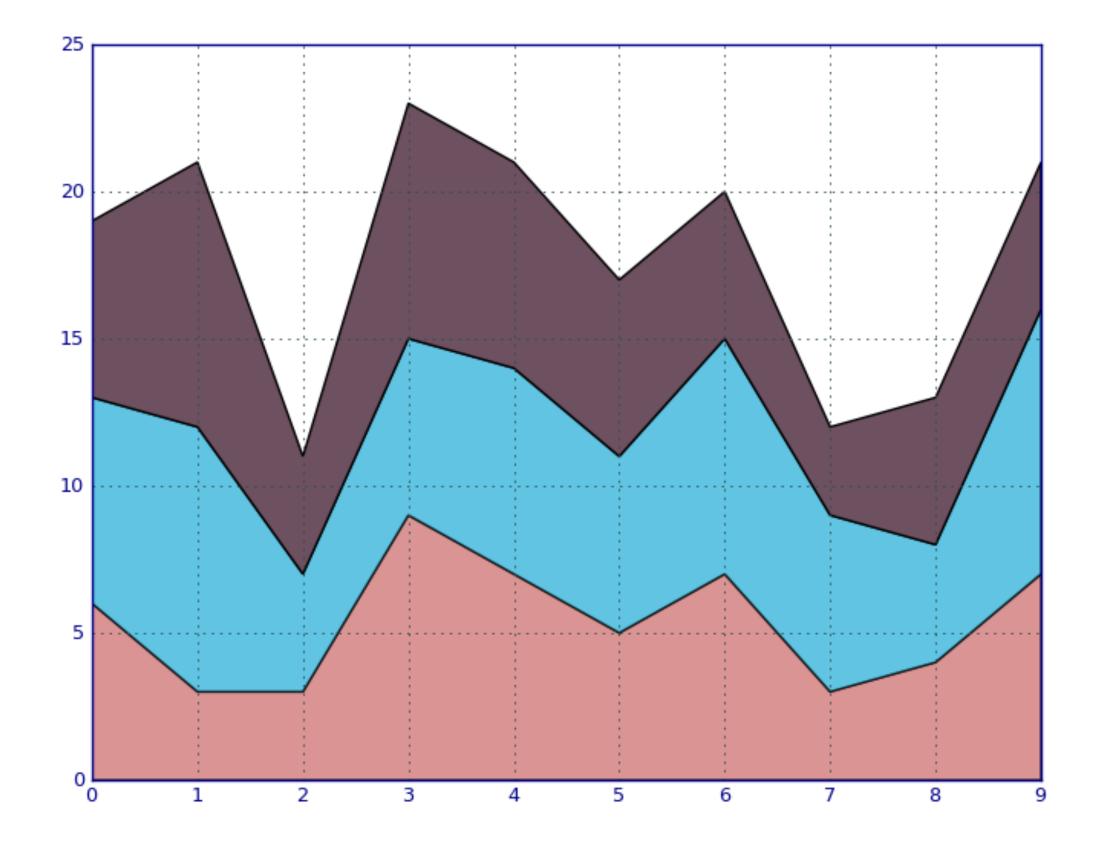




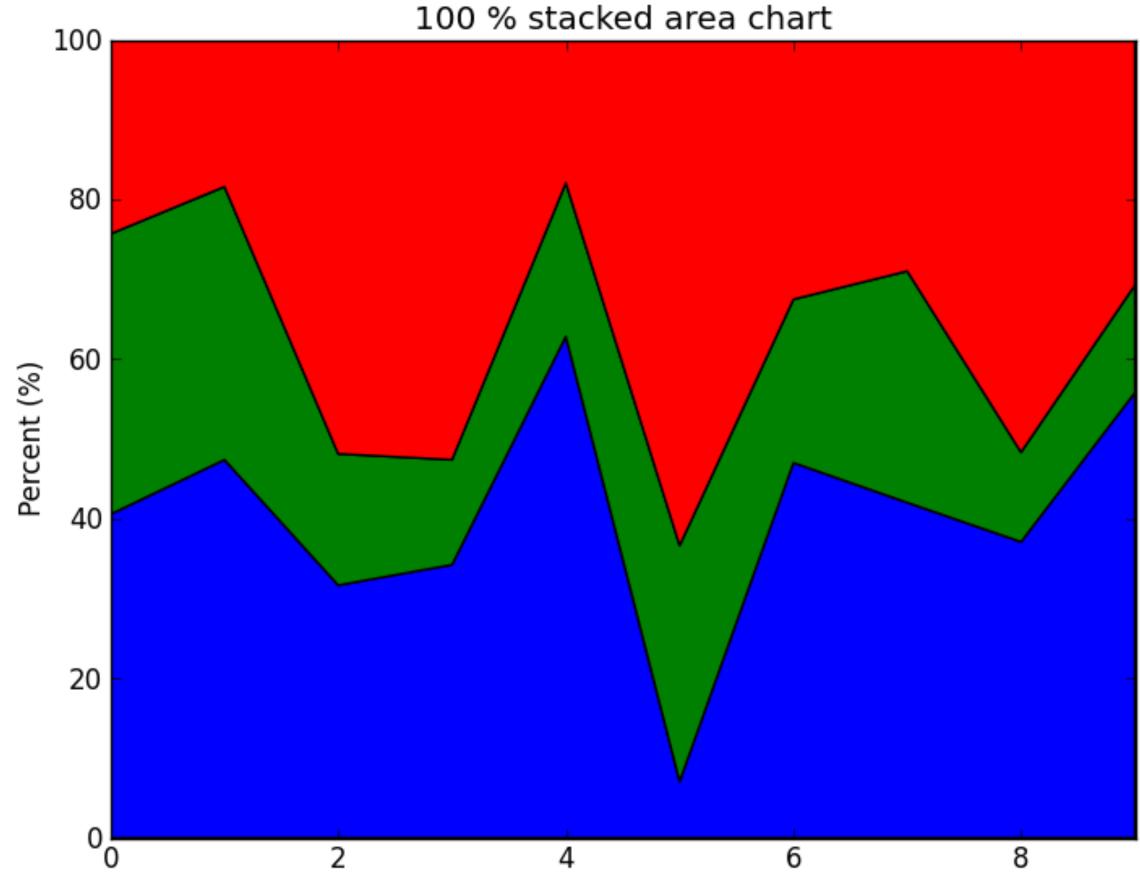
Stacked bar chart

Streit & Gehlenborg, PoV, Nature Methods, 2014

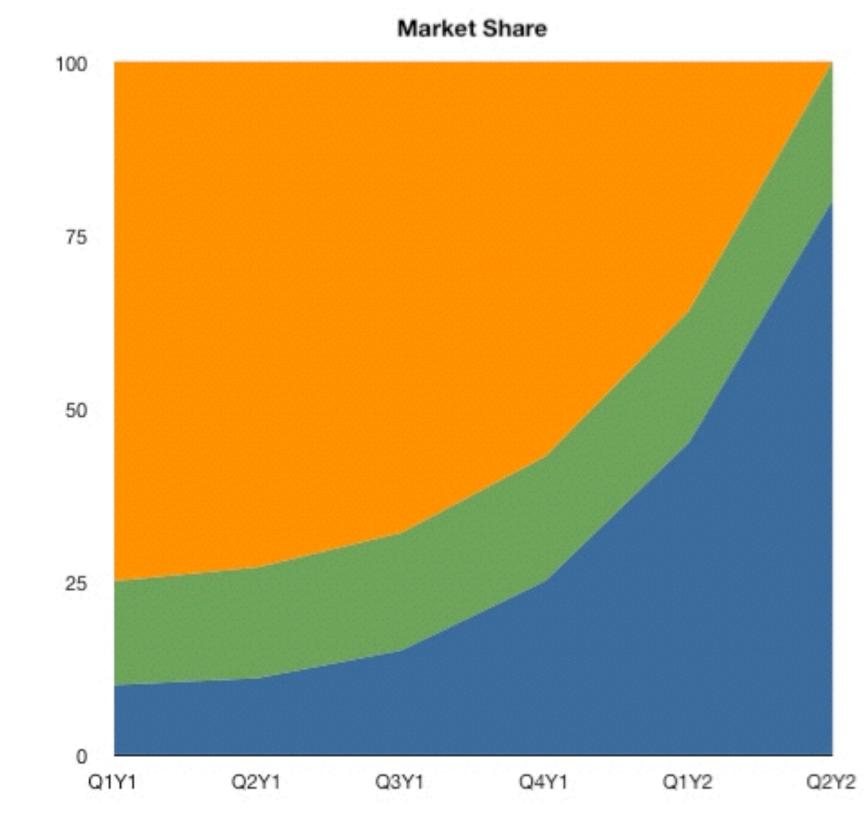
Stacked Area Chart

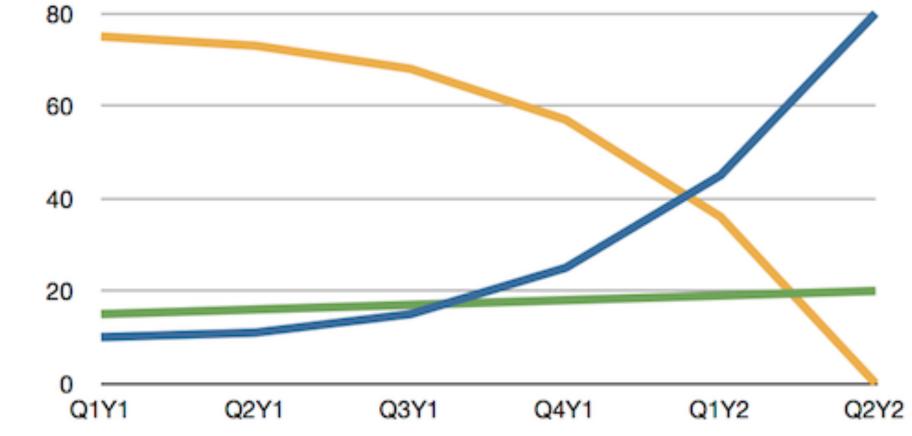


100% Stacked Area Chart



Stacked Area vs. Line Graphs



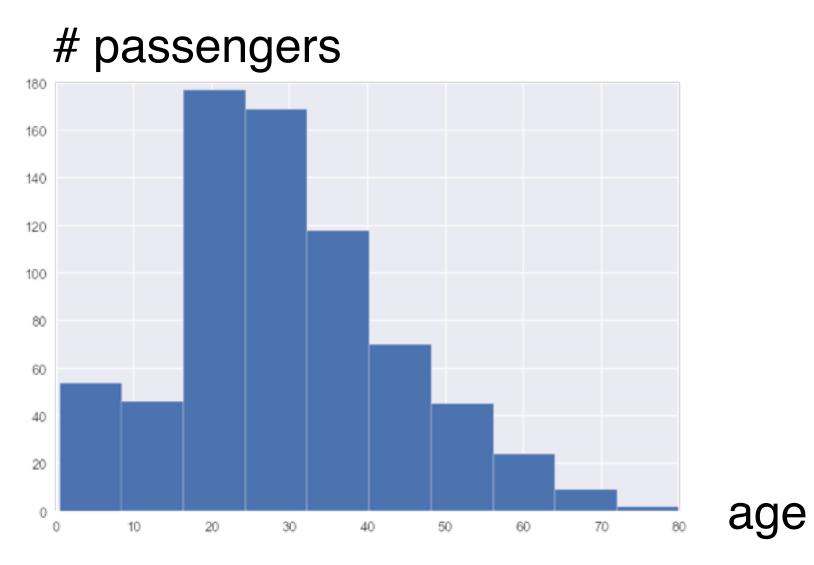


leancrew.com & Practically Efficient

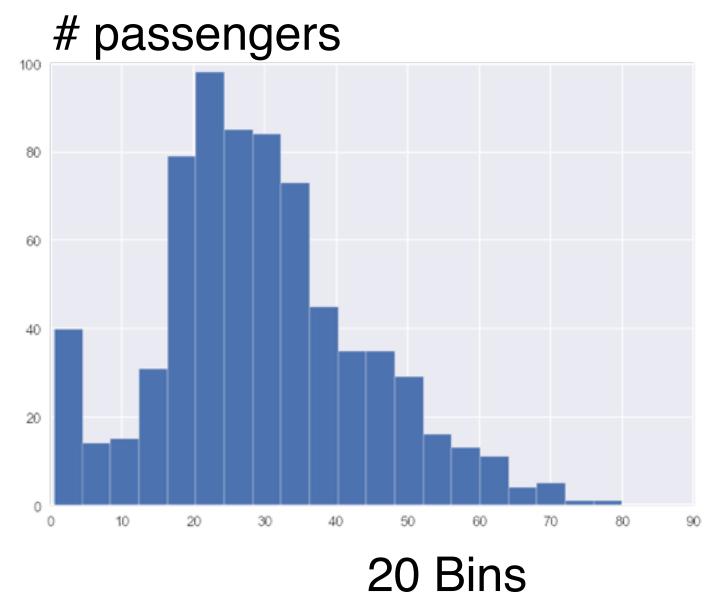
Distributions

Histogram

#bins hard to predict
make interactive!
rule of thumb: #bins = sqrt(n)

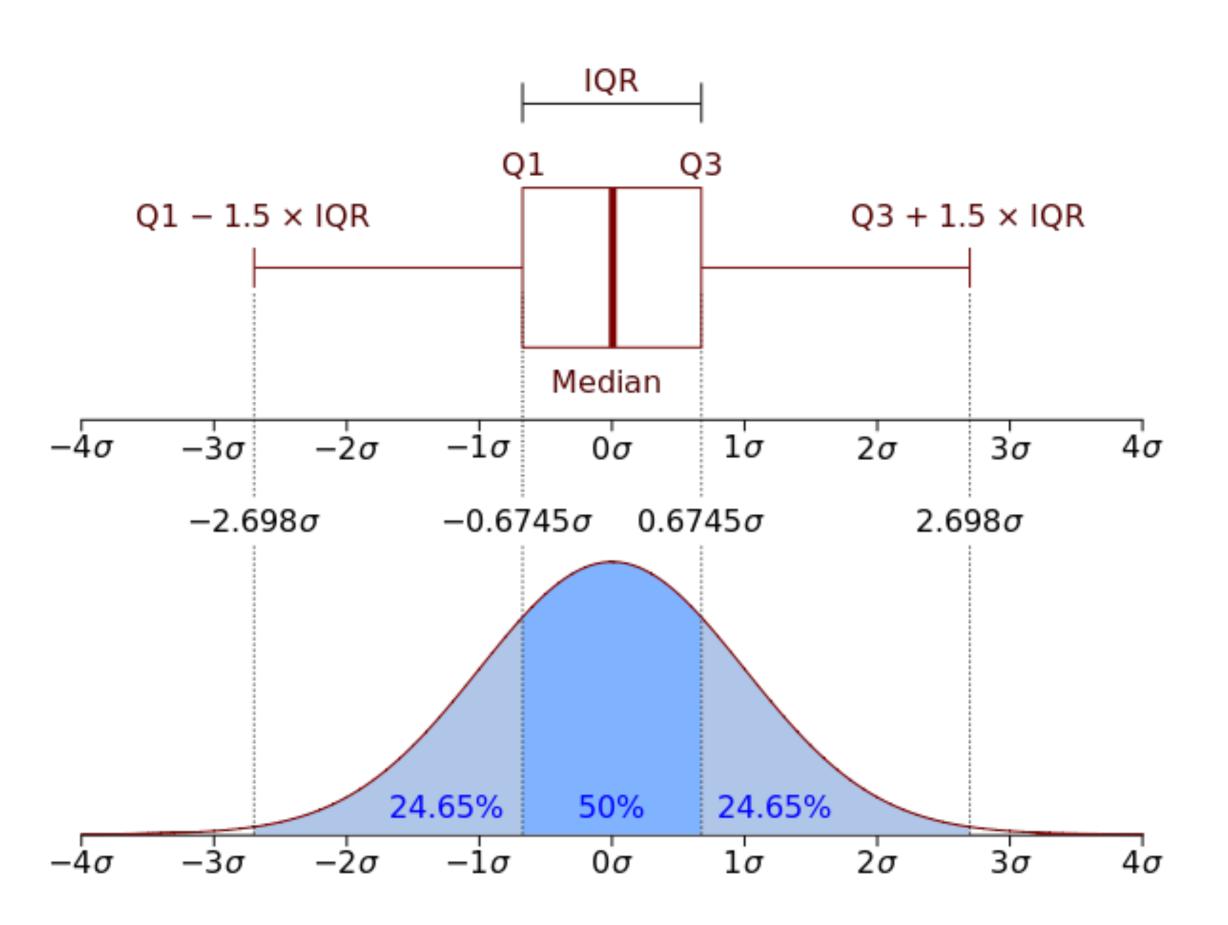






age

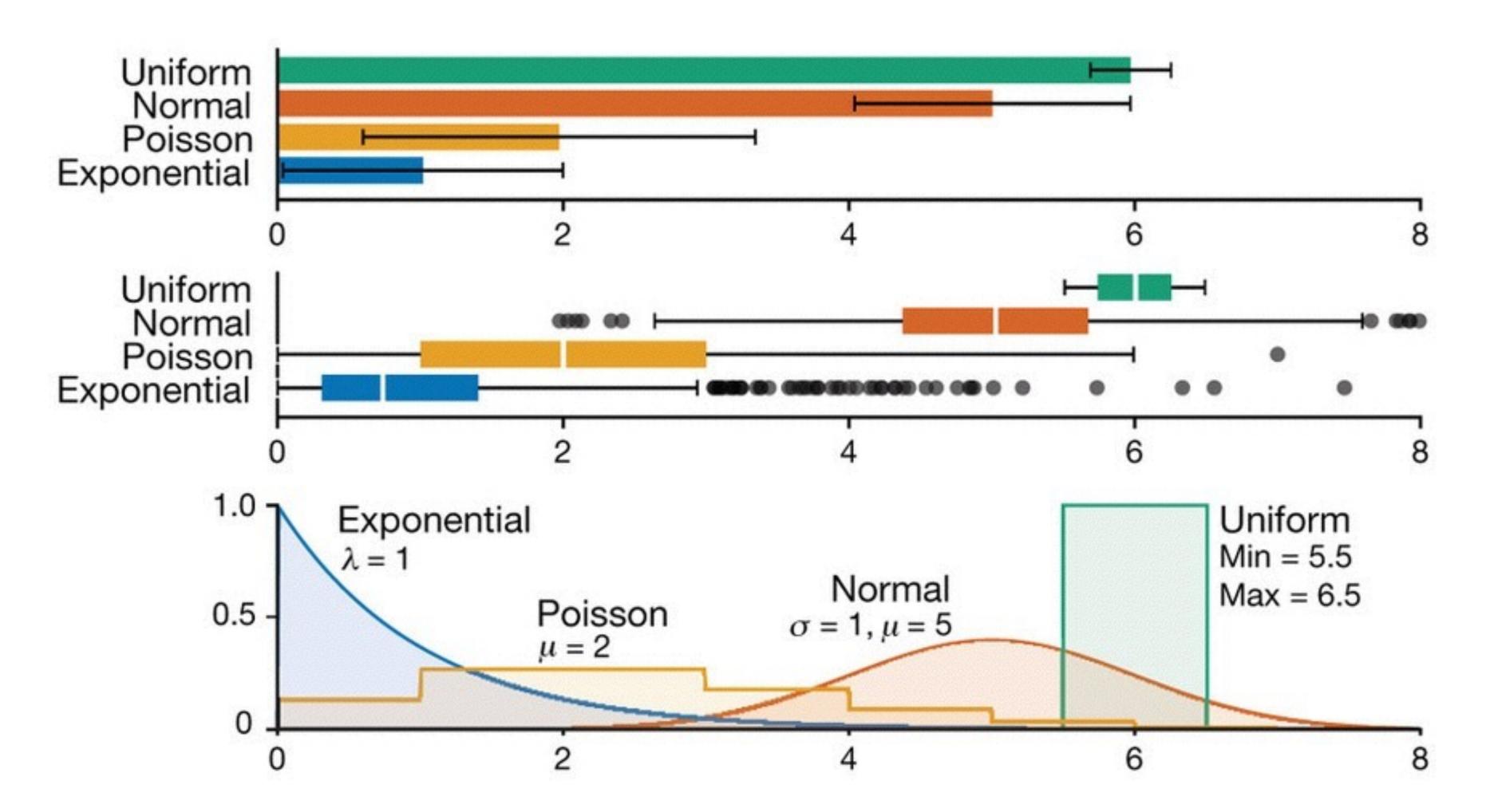
Box Plots



Wikipedia

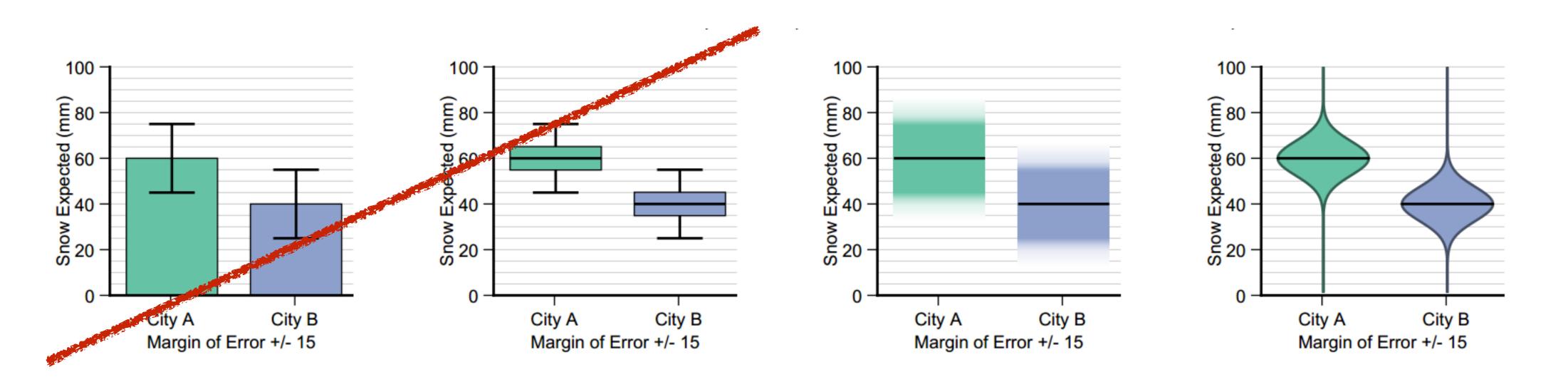
aka Box-and-Whisker Plot

Comparison



Streit & Gehlenborg, PoV, Nature Methods, 2014

Showing Expected Values & Uncertainty



Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error Michael Correll, and Michael Gleicher

Highdimensional Data

What is High-dimensional Data?

Tabular data, containing

- rows (items)
- columns (attributes or items)

rows >> columns

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

High-Dimensional Data Visualization Homogeneity

How many dimensions?

~50 - tractable with "just" vis

~1000 – need analytical methods

How many records?

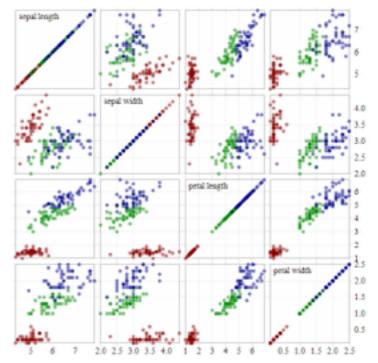
 \sim 1000 – "just" vis is fine

>> 10,000 – need analytical methods

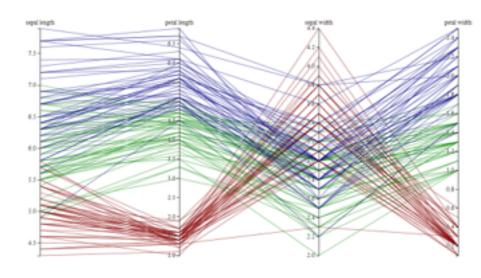
Same data type? Same scales?

	Age	Gender	Height
Bob	25	Μ	181
Alice	22	F	185
Chris	19	Μ	175
	BPM 1	BPM 2	BPM 3
Bob	BPM 1 65	BPM 2 120	BPM 3 145
Bob Alice			

Analytic Component

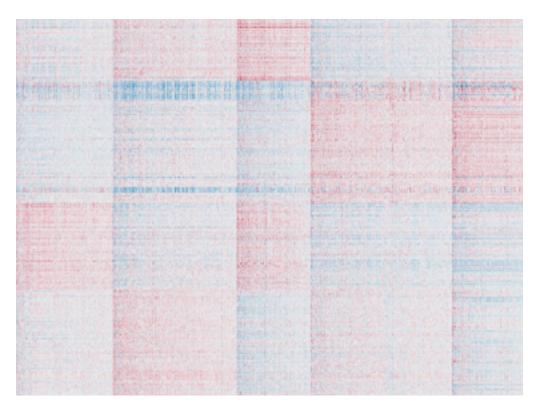


Scatterplot Matrices [Bostock]



Parallel Coordinates [Bostock]

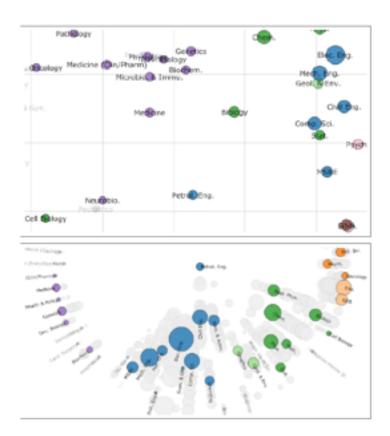
no / little analytics



Pixel-based visualizations / heat maps



Multidimensional Scaling [Doerk 2011]



[Chuang 2012]

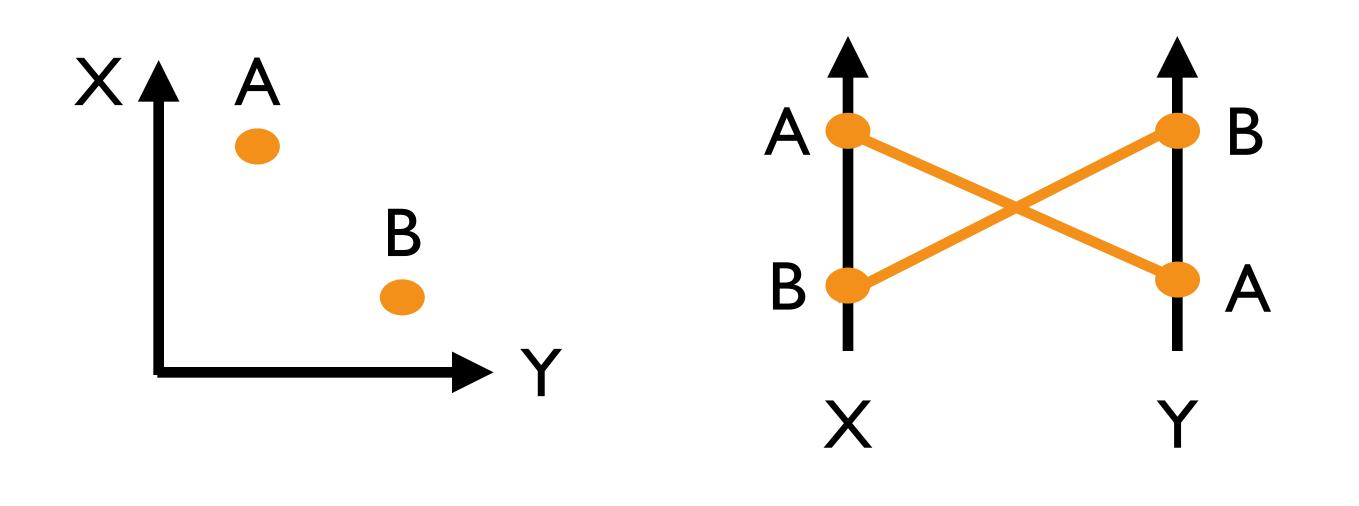
strong analytics component



Geometric Methods

Parallel Coordinates (PC) Inselberg 1985

Axes represent attributes Lines connecting axes represent items



Parallel Coordinates

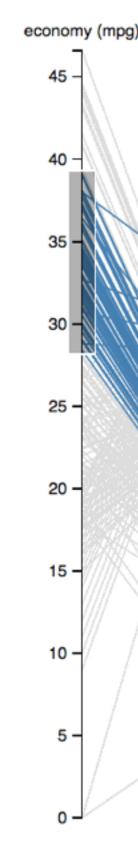
Each axis represents dimension

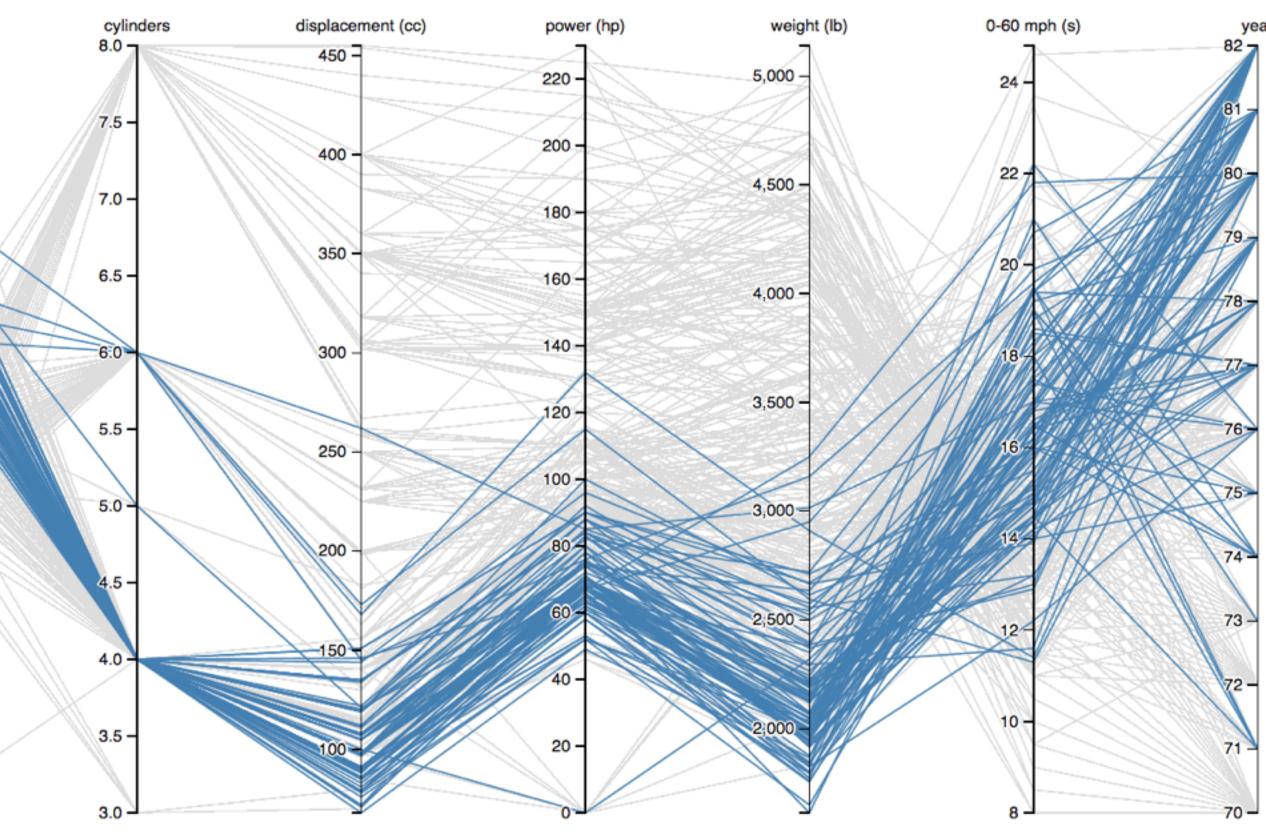
Lines connecting axis represent records

Suitable for

all tabular data types

heterogeneous data







PC Limitation:

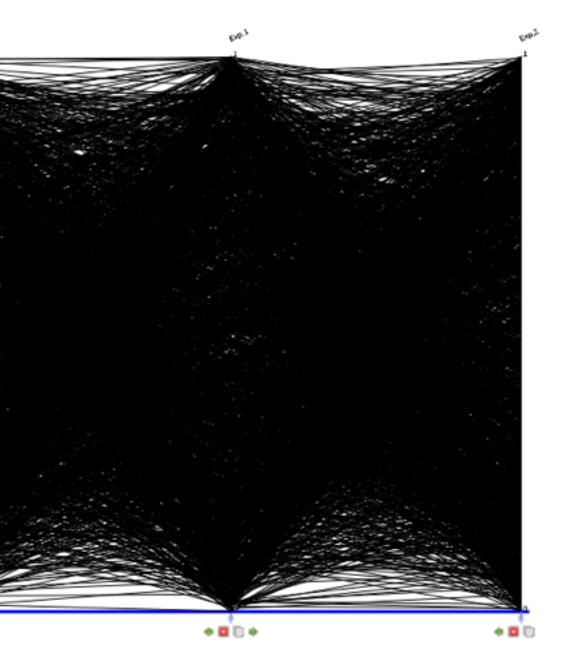


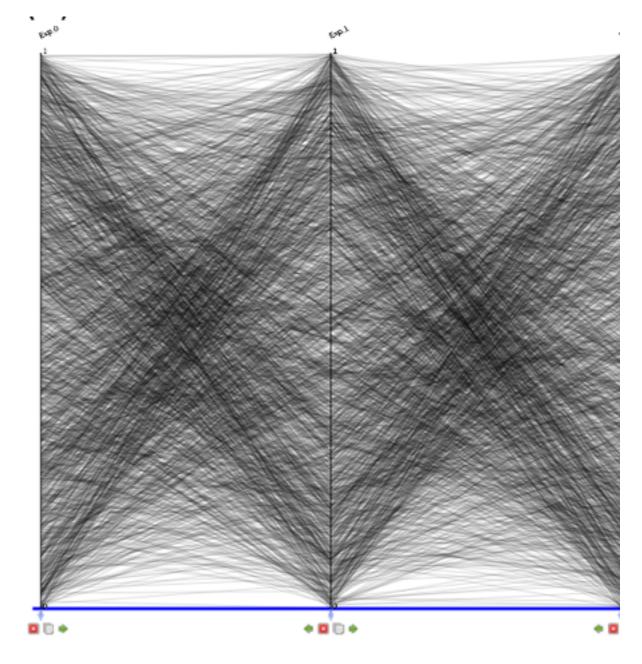
500 axes

PC Limitation: Scalability to Many Items Solutions:

Transparency Bundling, Clustering Sampling



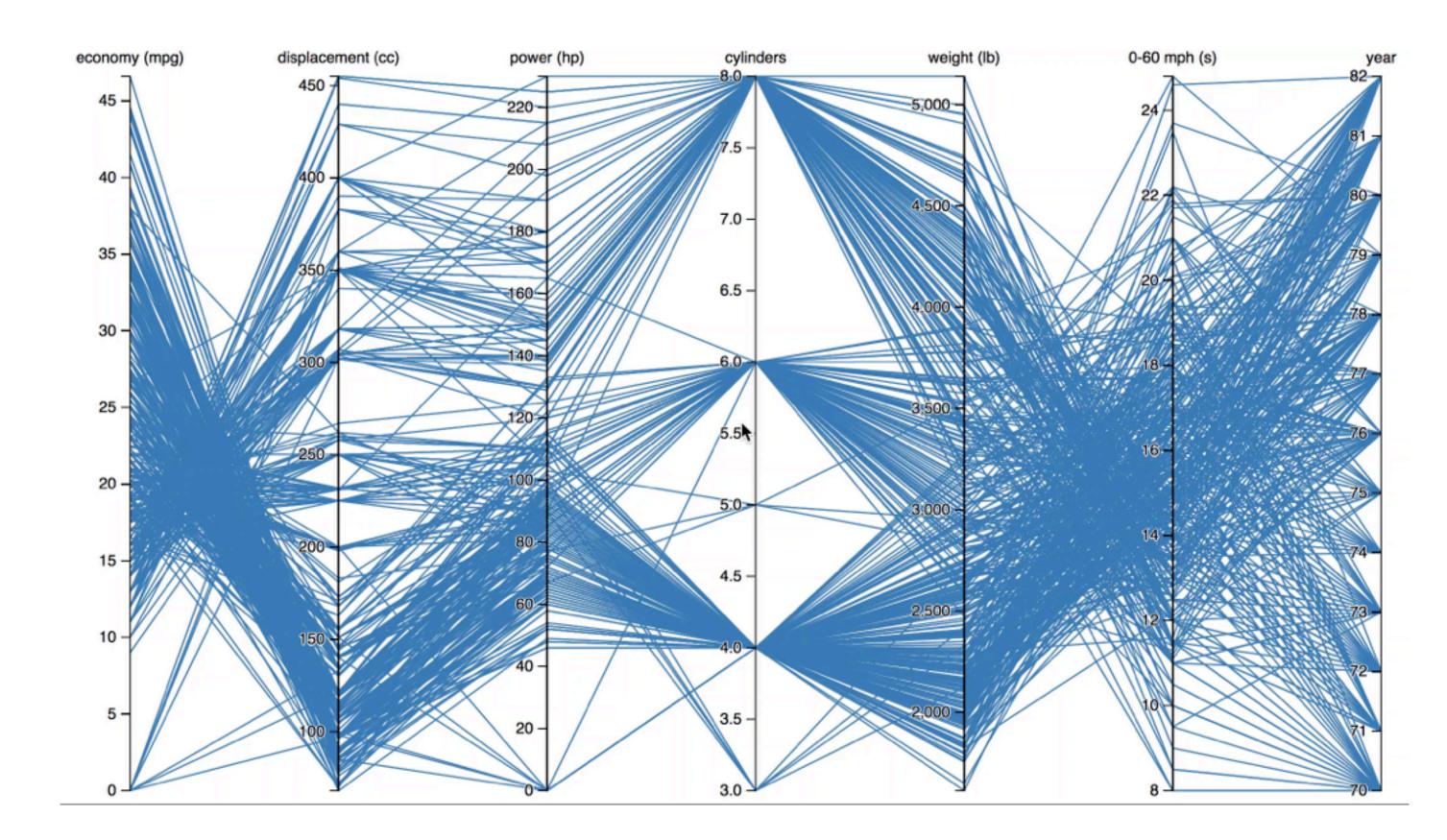




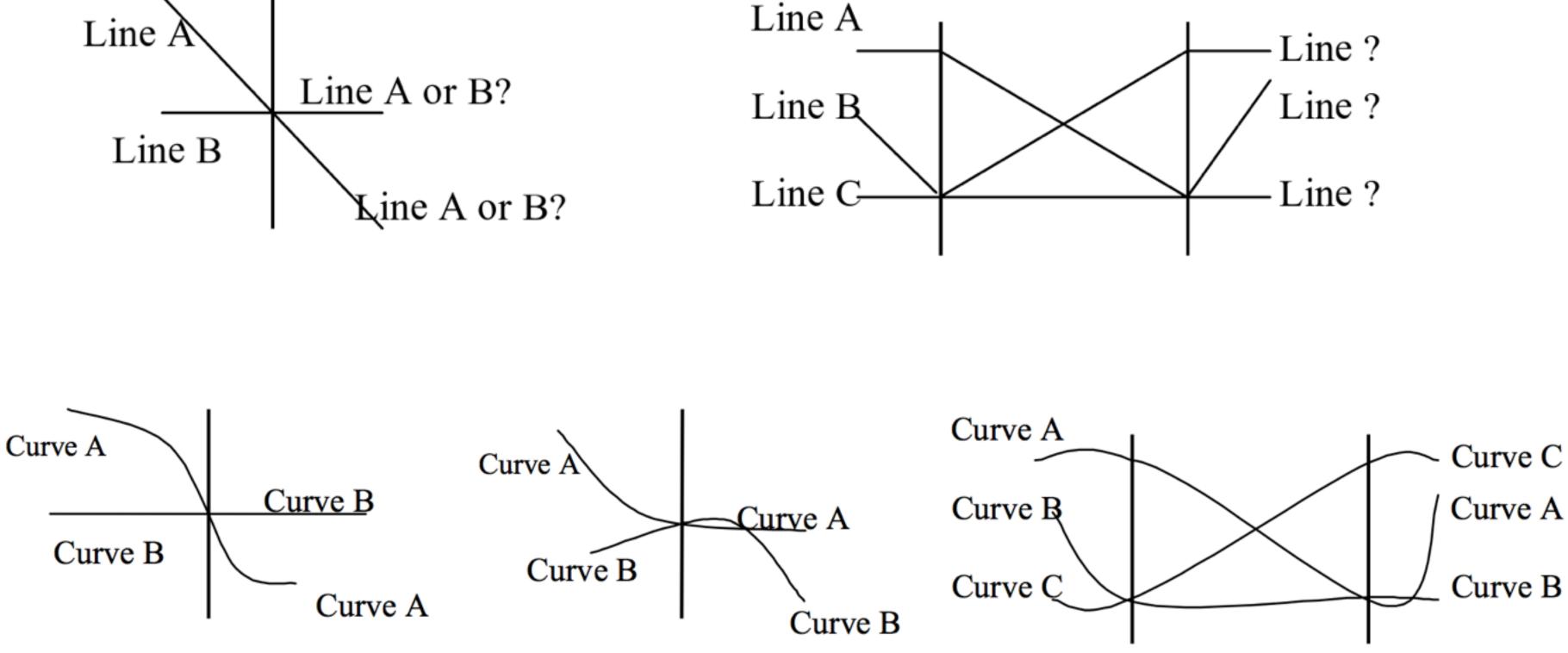
PC Limitations Correlations only between adjacent axes

Solution: Interaction Brushing

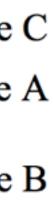
Let user change order



PC Limitation: Ambiguity Solutions: Line À Brushing Curves Line B



Graham and Kennedy 2003



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

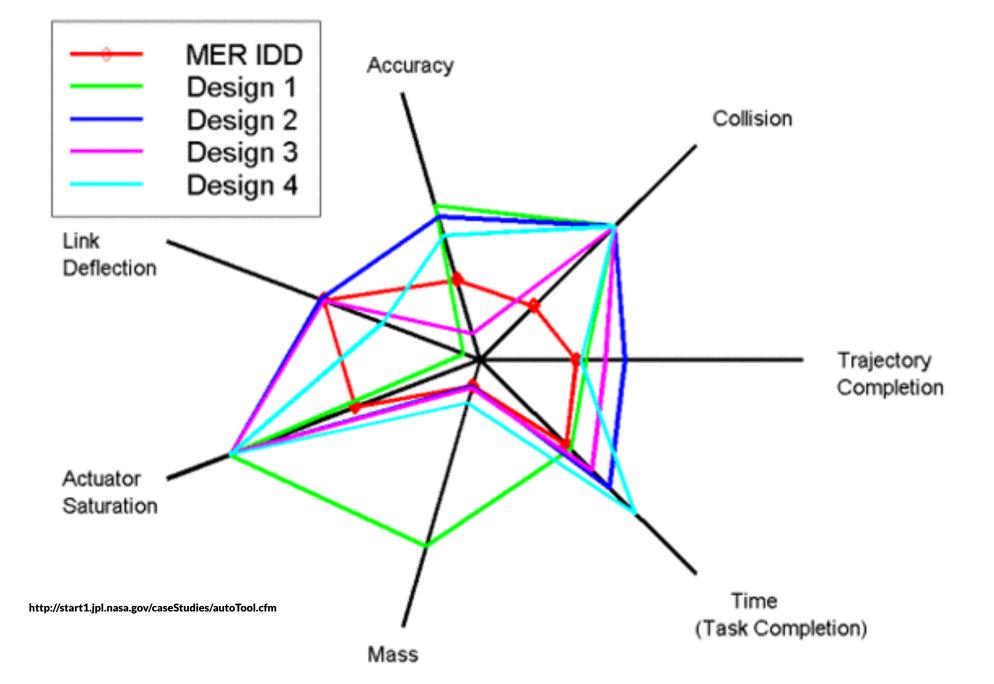
Algorithmic support: Choosing dimensions Choosing order Clustering & aggregating records

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

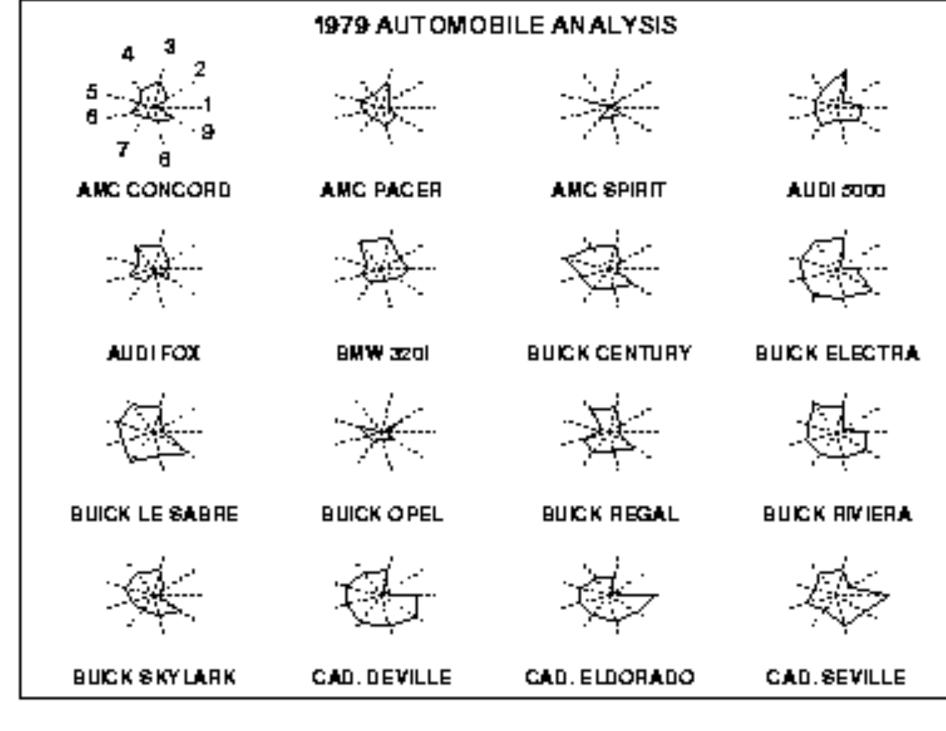
Star Plot

Similar to parallel coordinates Radiate from a common origin

Star Plot of MER IDD and Automated Designs



[**Coekin1969**]



http://bl.ocks.org/kevinschaul/raw/8833989/





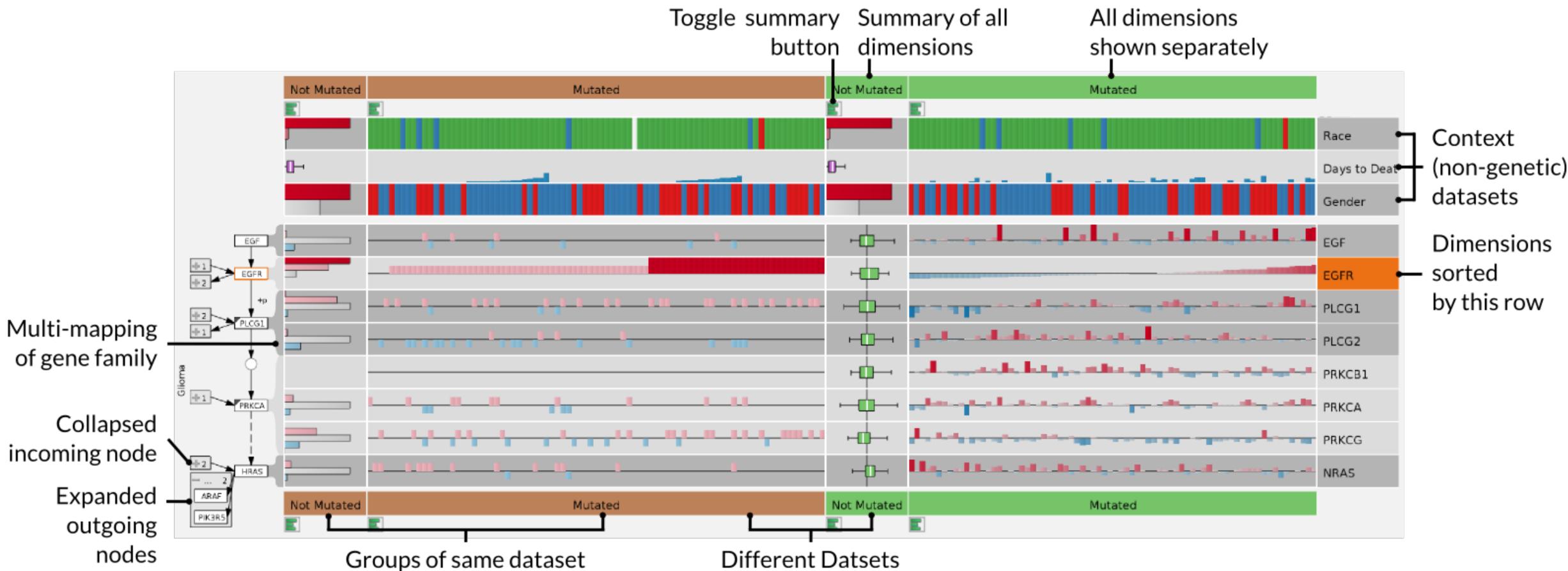
Multiple Line Charts

05:30	05:45	08 AM	06:15	08:30	06:45	07 AM	07:15	07:30	07:45 I	08 AM	08:15	08:30	08:45	MA CO	09:15
1			1810						1 A A		and the second	a la cart	1111	A Contraction	-3.4
2		Although the second	1.1					Sec. 1				1.4			3.4
3			11 A V												-1.7
4				And Shake			84 - C.	A. J.			<u> </u>		1.1.1	1	-4.9
5								1.1.1		1.	AA		A 1		6.1
6	. A. (A)			A Second	A. A.			A4	1	\mathbf{V}	A			N. A.M.	0.067
7			A 1 4 1	<u> </u>		N.Y		Sec.			V. ***	A State of the second		1AVC	0.26
8			A 1944				1.1.1	1 A 1993				A 4	<u>a.</u>		1.1
9	949 A.M							AJA		$M_{\rm ext}$			1		7.6
10		A A A	<u> </u>	AAA						<u> 17. YA</u>		A STA	A A A		5.7
11			ARA	A. A.		1 1 1		<u>yy</u>							4.8
12					A	4 . Y									0.23
13									1 141	A a a a				A	0.51
14						TTNI	TIT.		LAIMAN						-3.5
15	A Andrea	****	. <u></u>					1.1			<u>. A A A .</u>		A. 34.4	A. S. A. S. A	-8.9
16			a data a			<u>Likini v</u>	1 Alexandre					A second		<u>A.A.A.</u>	-5.7
17						<u>A.A. A.M</u>			14 A A A A						-4.8
18	***	TAAM													-4.0
19 20											. u A A				
20			A A A A A	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	-	- ANI	ABAAM	A & 4 44				4 4 4	A AA		79
21				ARE LAND	<u>k et a se dette</u>	1 . A	A has diverse			<u>k</u>					3.3
22 23		A	. A		A & A &	Mar in								TY I I I	2.0
23		Ay				<u> </u>		1	A K			<u>a.k.</u>			-1.2
25	A LAN	LLARA M.					- 444				A.A A				-2.7
100 2 2 2 2 2	and the state of a		A4.4 . A4											يطفع إربعاقه	-0.1
26		A A Board	All talls		-	1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1						A			-3.1
28			and the second se				Adden and a	a bei be week	an adread					A A A A A A A A A A A A A A A A A A A	4.0
29	W WAY	· · · · ·		A A A A	A and a	An. Auto	A	ALC: NO.			Martin Martin				0.45
30				1.000		and a second		A. I. Sauce	A			- 7 -	111		7.5
31	11 T		FFTTT	1.1.1	- thread	1 17.01	S. P. An	1.44	A. A.			A CONTRACTOR			8.4
32			-				ALLA ALLA	A. A.A.		and the second		ALL ALLAN		A MARKED	8.5
YE MULA		11 A. A. M. 11		THE REPORT OF	1. A.A.	AAAAAAAA	- T 1			1000		1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	100	0.0



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

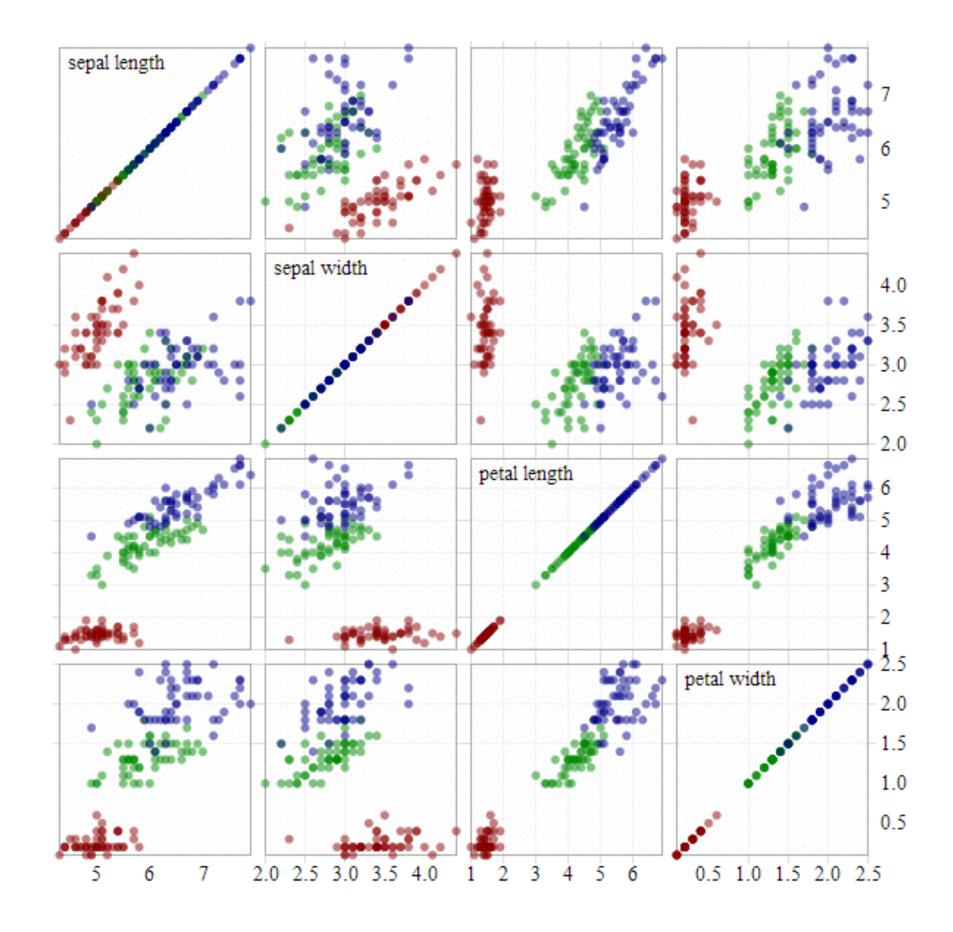
Combining Various Charts





Scatterplot Matrices (SPLOM)

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



Scatterplot Matrices

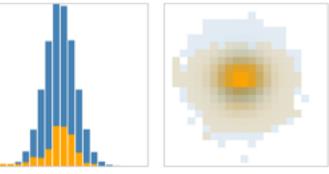
Limited scalability (~20 dimensions, ~500-1k records)

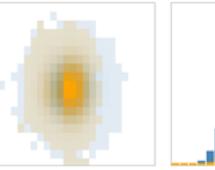
Brushing is important

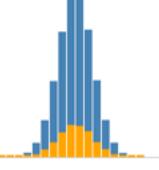
Often combined with "Focus Scatterplot" as F+C technique Algorithmic approaches: Clustering & aggregating records Choosing dimensions Choosing order

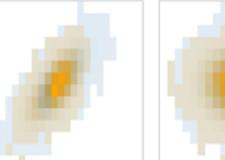
SPLOM Aggregation - Heat Map

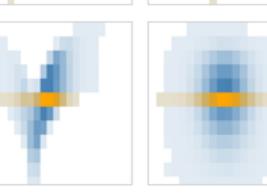
Interactive Binned Scatterplot Matrix Dimensions: 5 V Bins: 20 V Data Points: 100k V

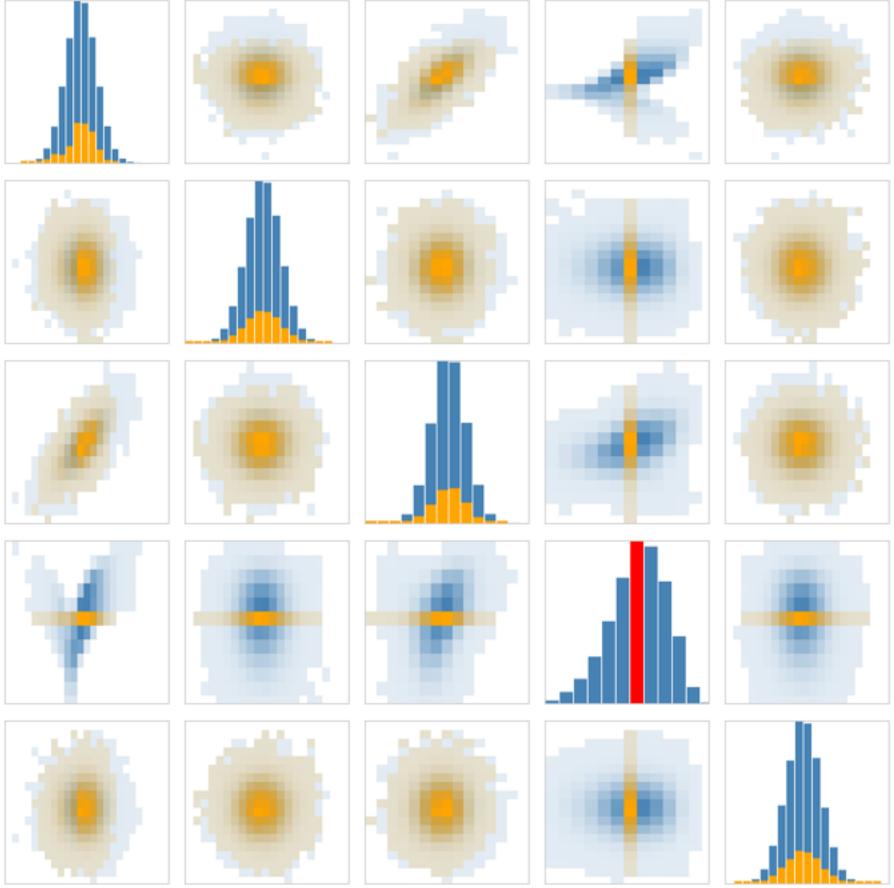








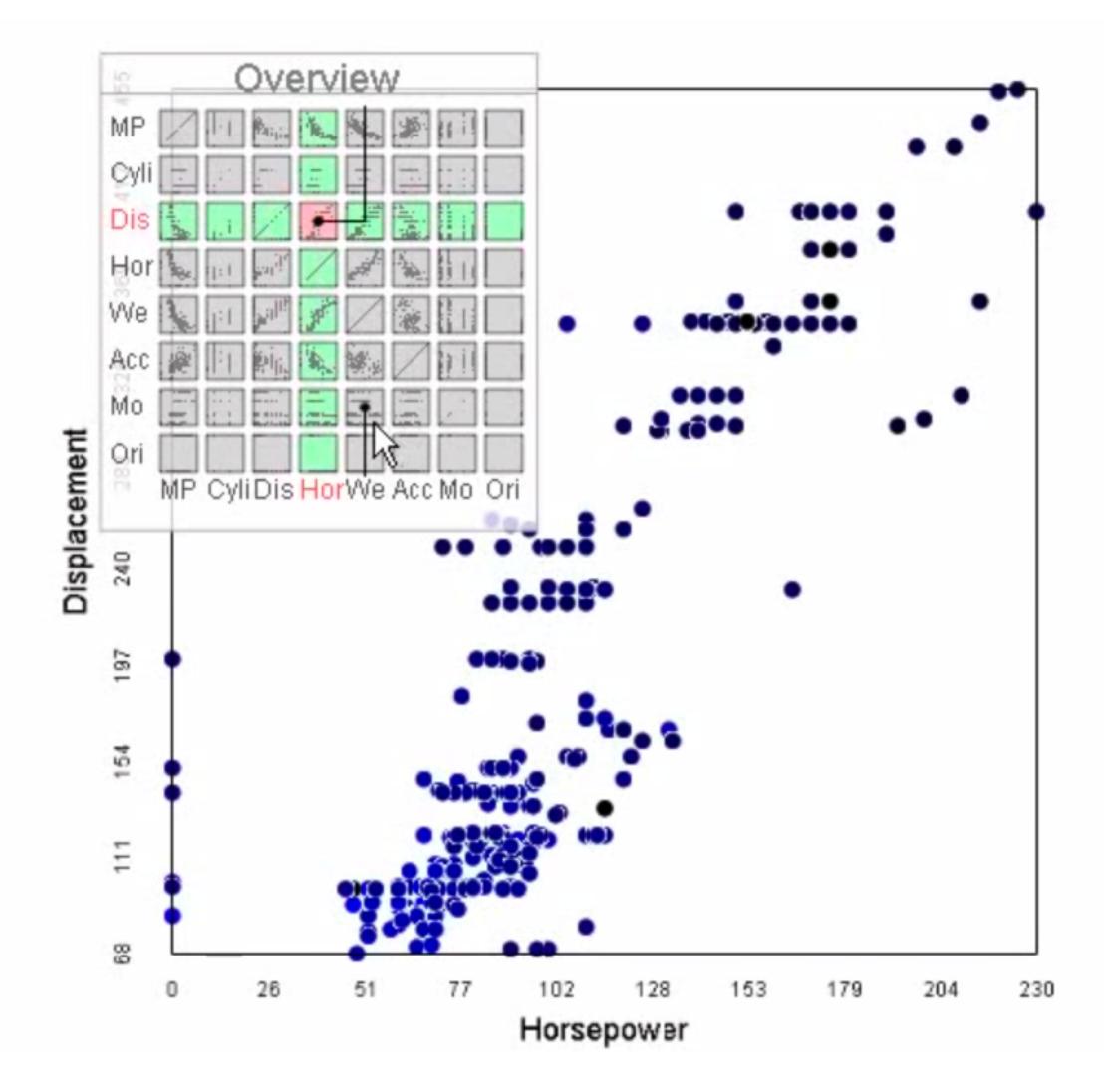




Powered by Datavore and D3.

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

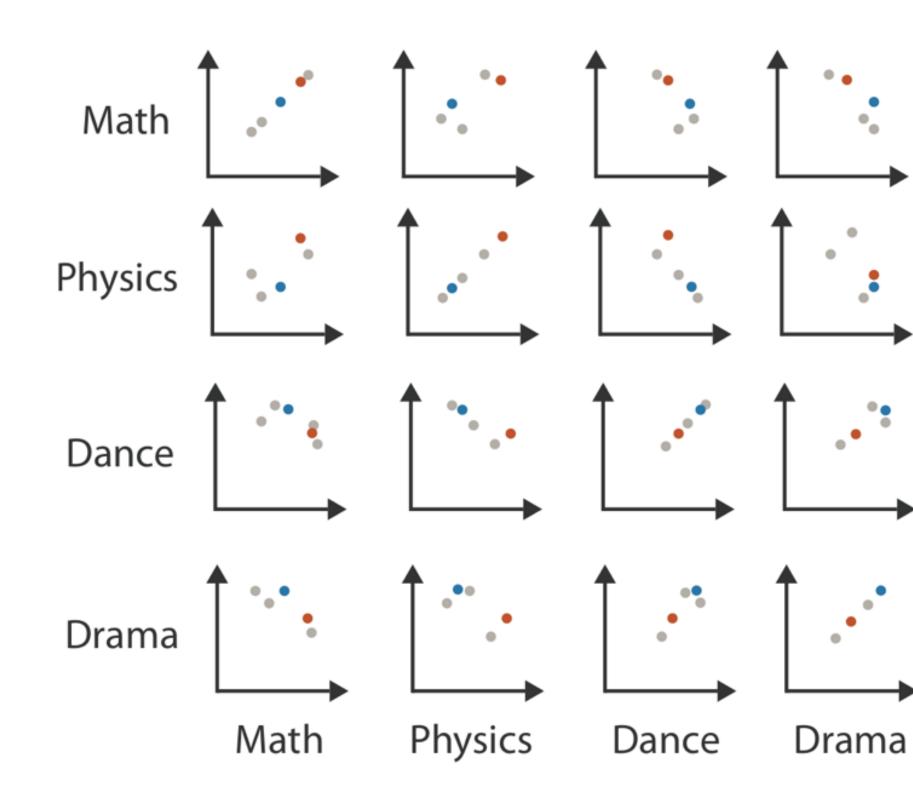
SPLOM F+C, Navigation



[Elmqvist]

Math	Phys
85	95
90	80
65	50
50	40
40	60

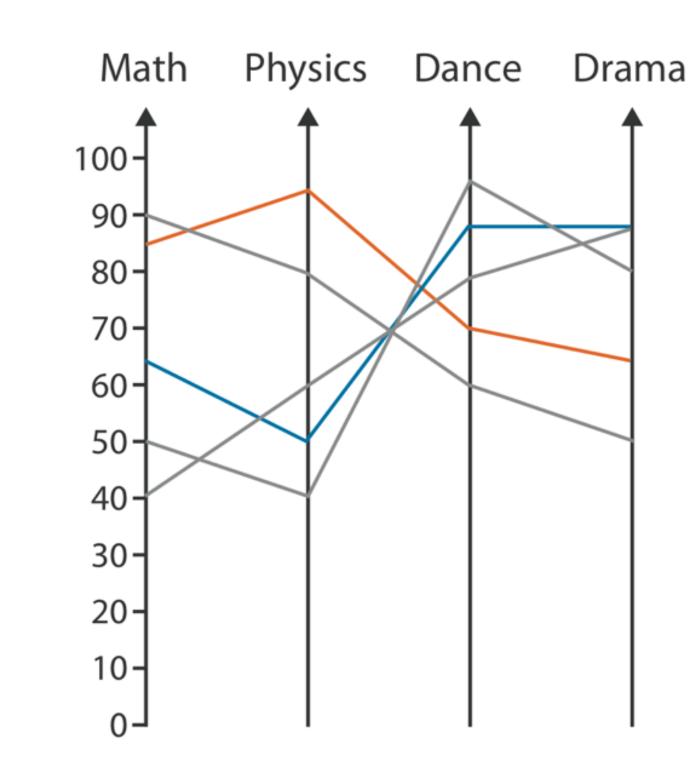




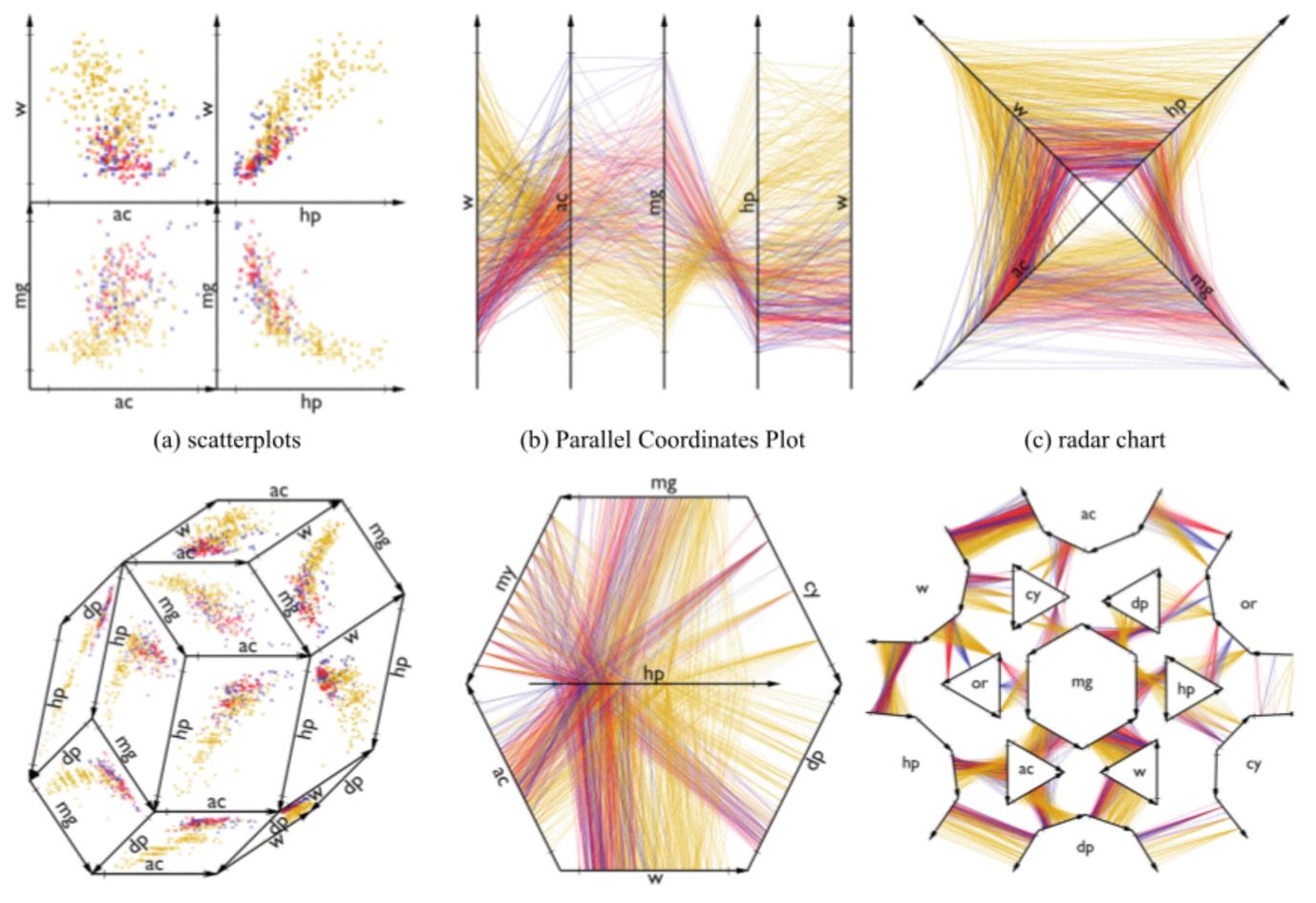
Table

sicsDanceDrama70656050909095808090

Parallel Coordinates



Flexible Linked Axes (FLINA)

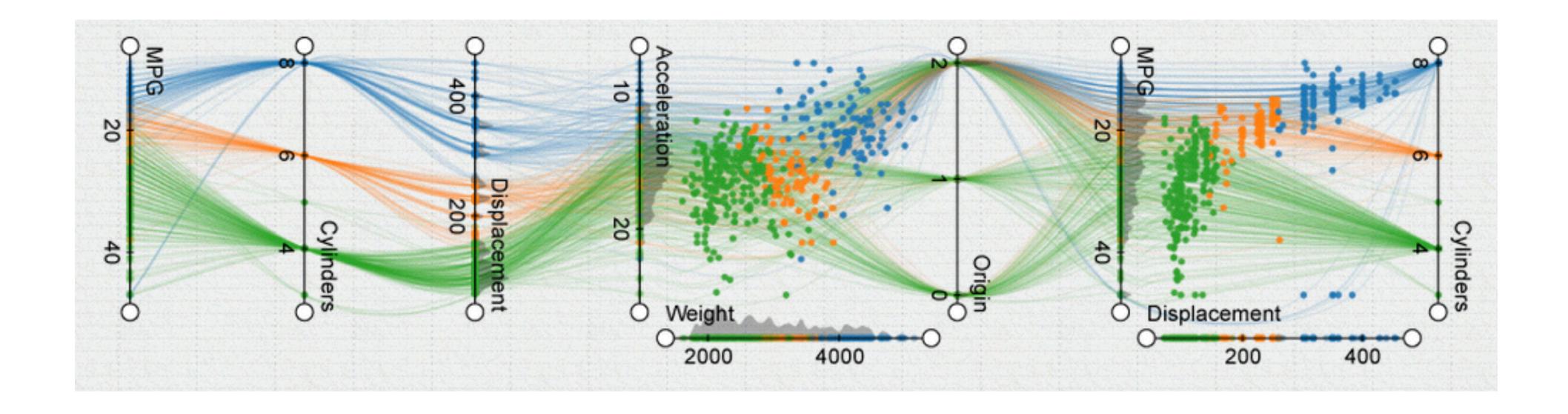


(d) Hyperbox

(e) Time Wheel

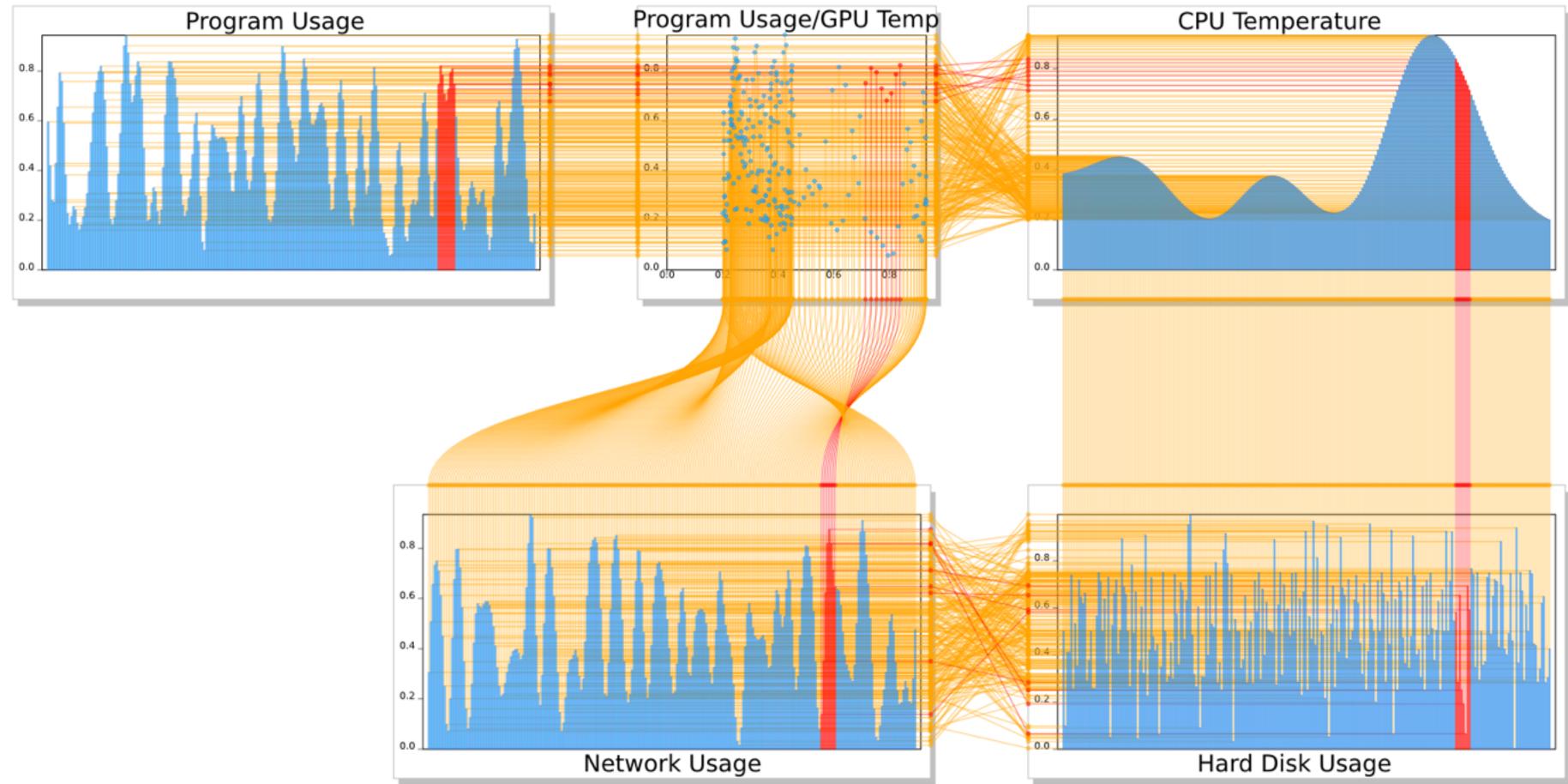
(f) Many-to-many PCP

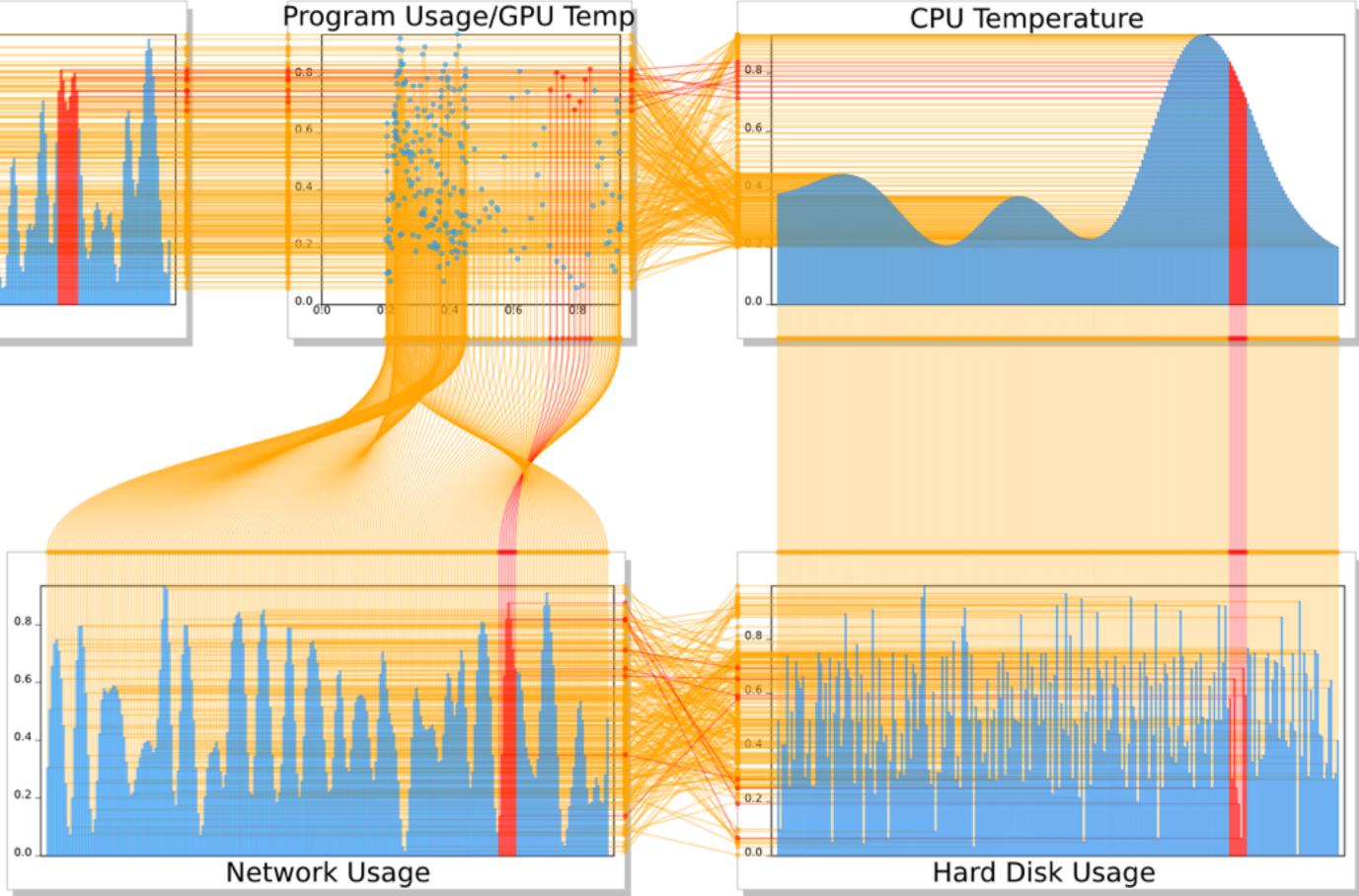
Web-based implementation of FLINA concept



http://vis.pku.edu.cn/mddv/val/

Connected Charts

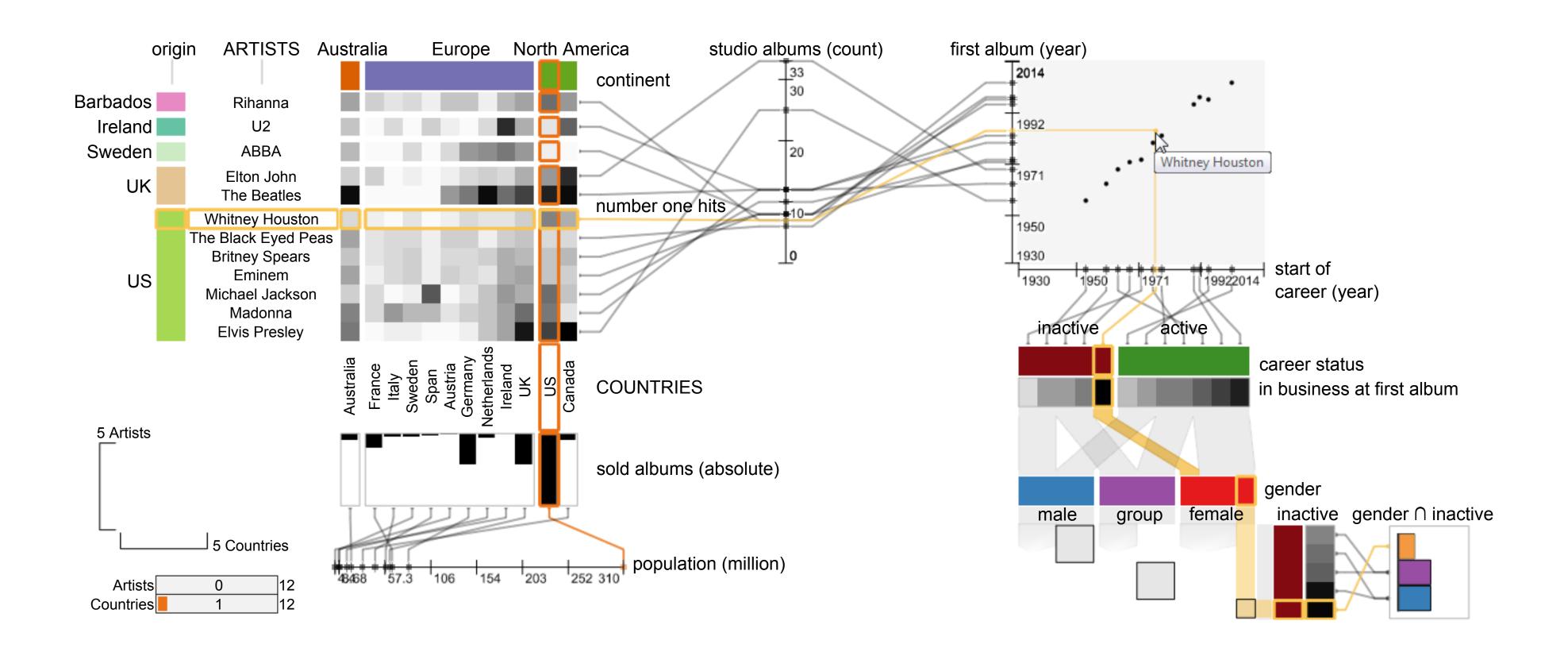






Viau & McGuffin 2012

Domino



Gratzl et al. 2014

Data Reduction

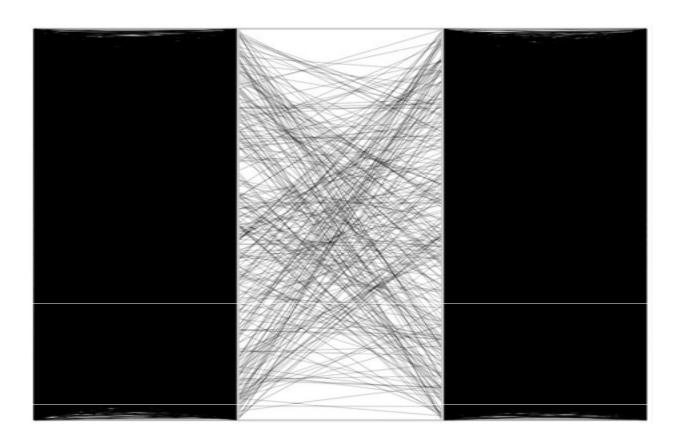
Sampling

Don't show every element, show a (random) subset

Efficient for large dataset

Apply only for display purposes

Outlier-preserving approaches



[Ellis & Dix, 2006]

Filtering Define criteria to remove data, e.g.,

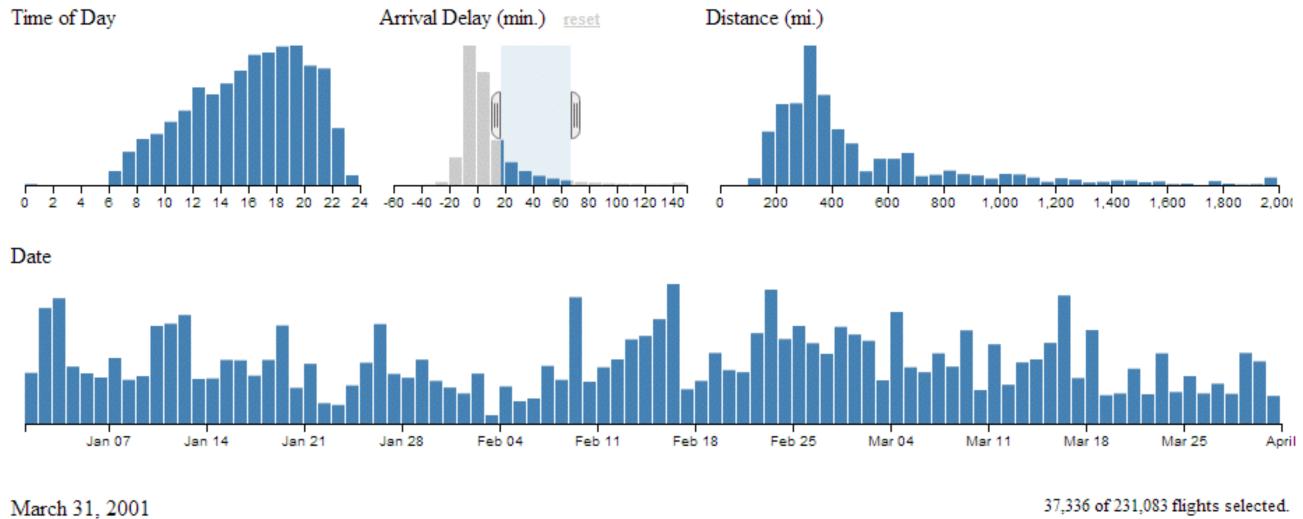
minimum variability

> / < / = specific value for one dimension</p>

consistency in replicates, ...

Can be interactive, combined with sampling

Filter Example



10:57 PM	MSY	HOU
09:33 PM	BWI	PVD
09:30 PM	TPA	PBI
09:29 PM	BWI	PVD
09:10 PM	LAS	PHX
09:02 PM	MSY	HOU

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

37,336 of 231,083 flights selected.

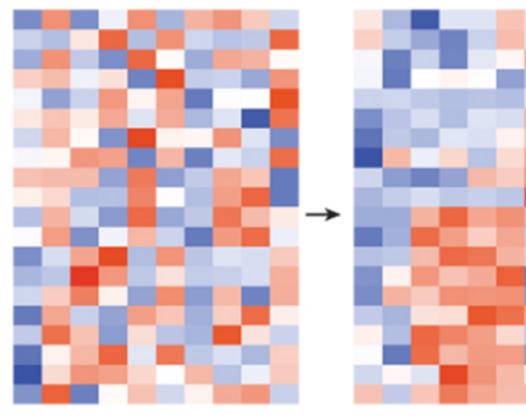
303 mi.	+29 min.
328 mi.	+64 min.
174 mi.	+30 min.
328 mi.	+20 min.
256 mi.	+53 min.
303 mi.	+26 min.

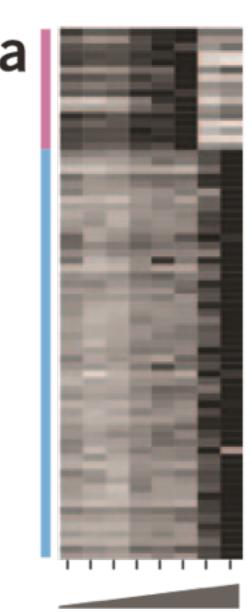
Pixel Based Methods

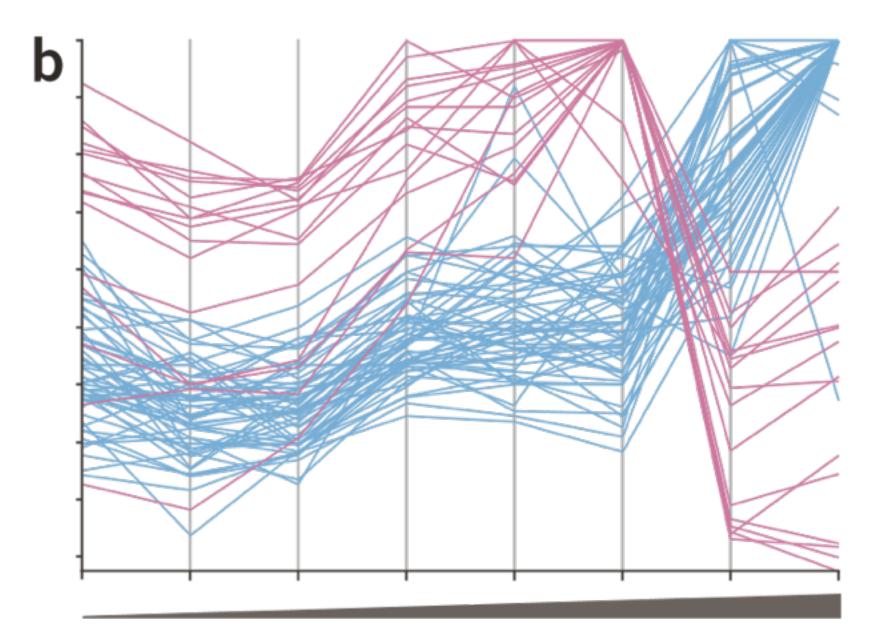
Pixel Based Displays

- 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







[Gehlenborg & Wong 2012]

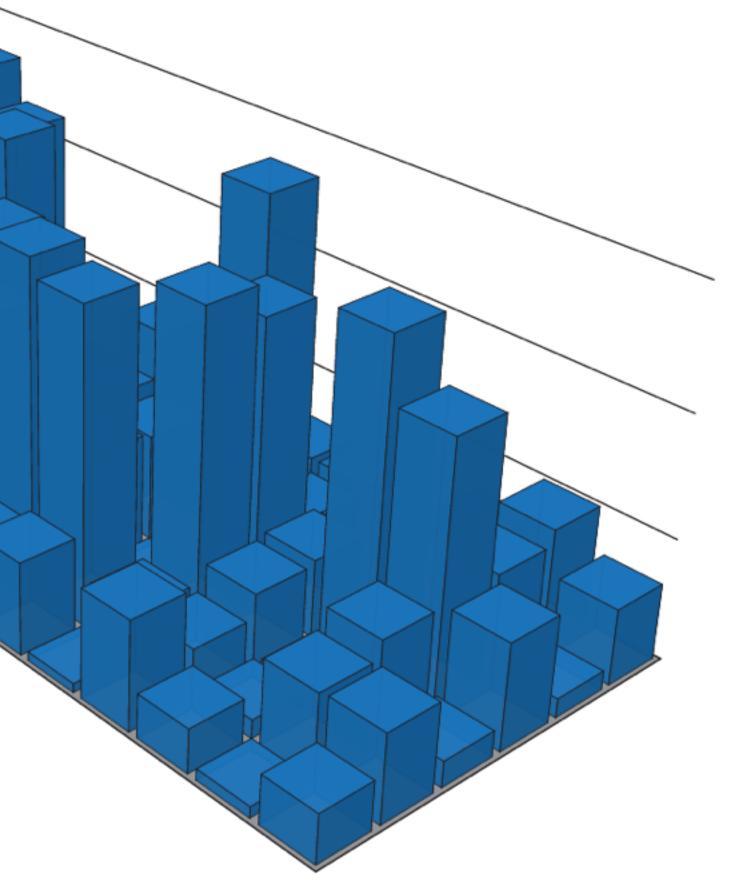


3D Pitfall: Occlusion & Perspective

Which one is the tallest bar? What is the pattern in the data?

0

З

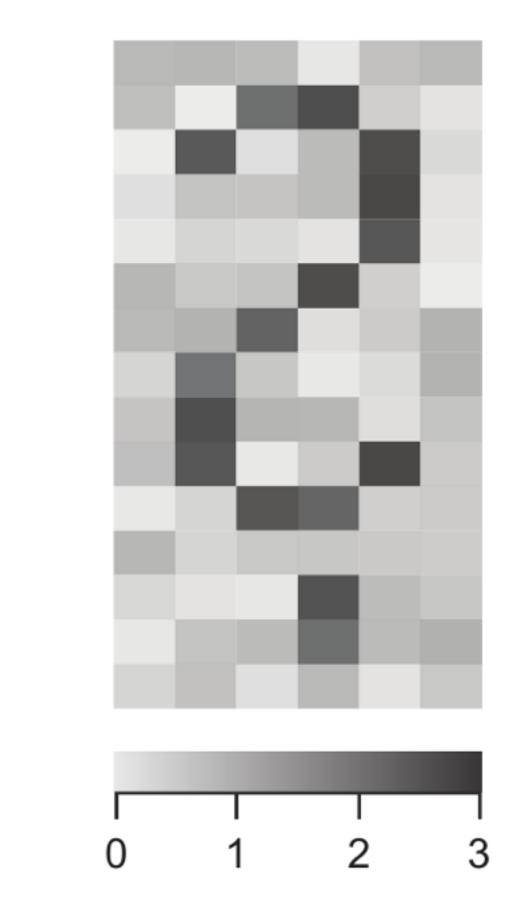


[Gehlenborg and Wong, Nature Methods, 2012]

3D Pitfall: Occlusion & Perspective

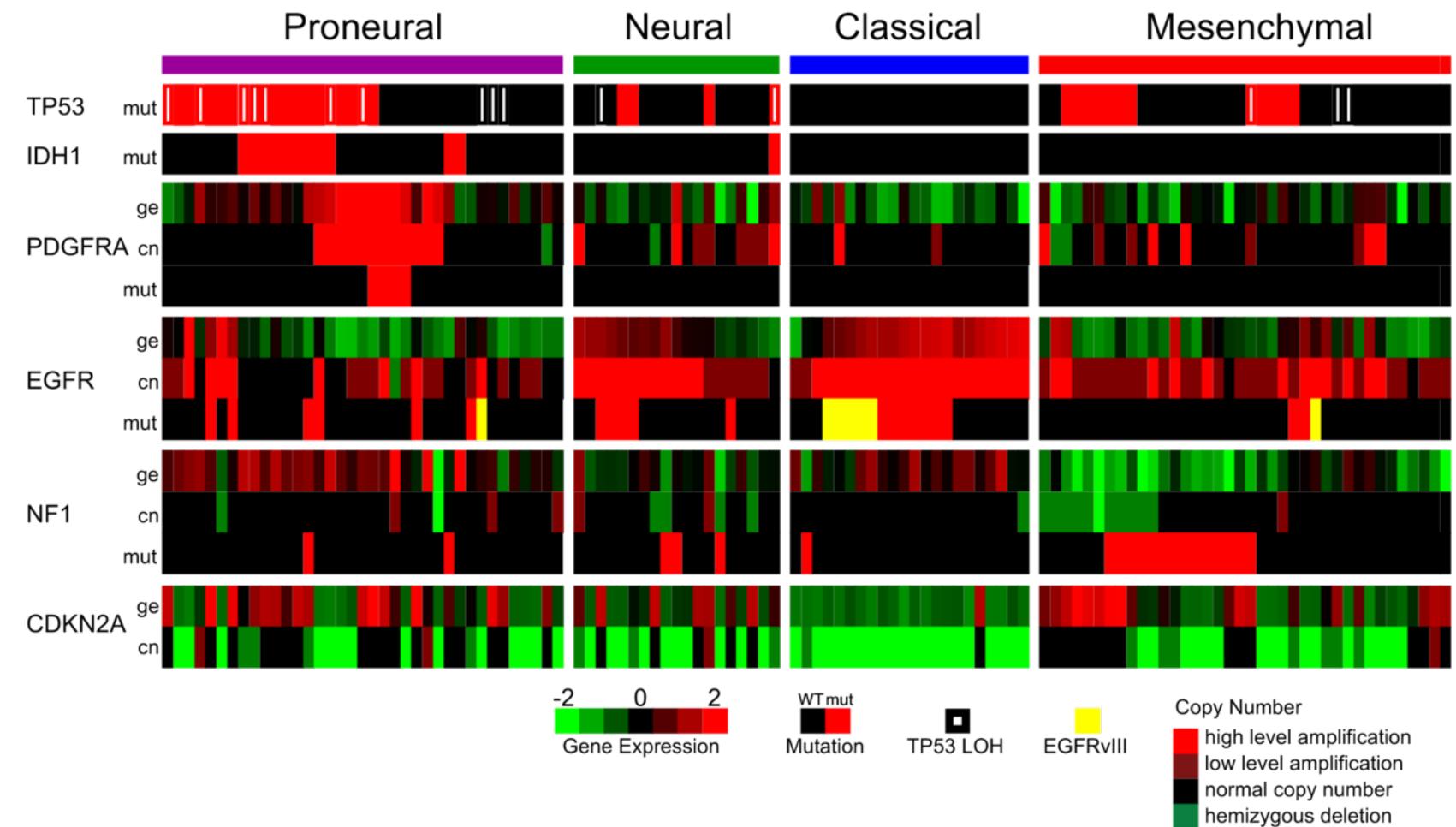
Which one is the tallest bar? What is the pattern in the data?

з



[Gehlenborg and Wong, Nature Methods, 2012]

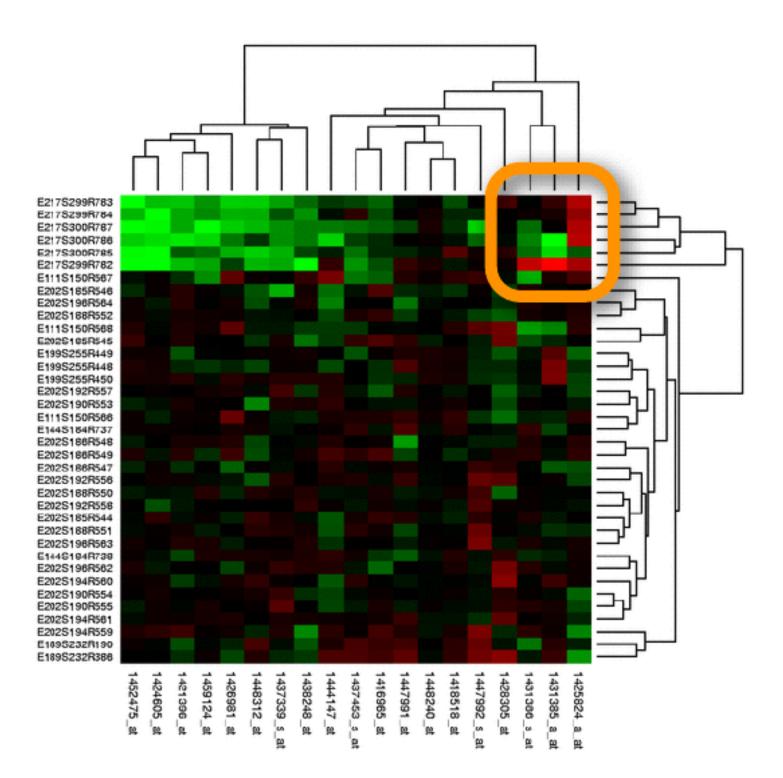
Heterogeneous Data?



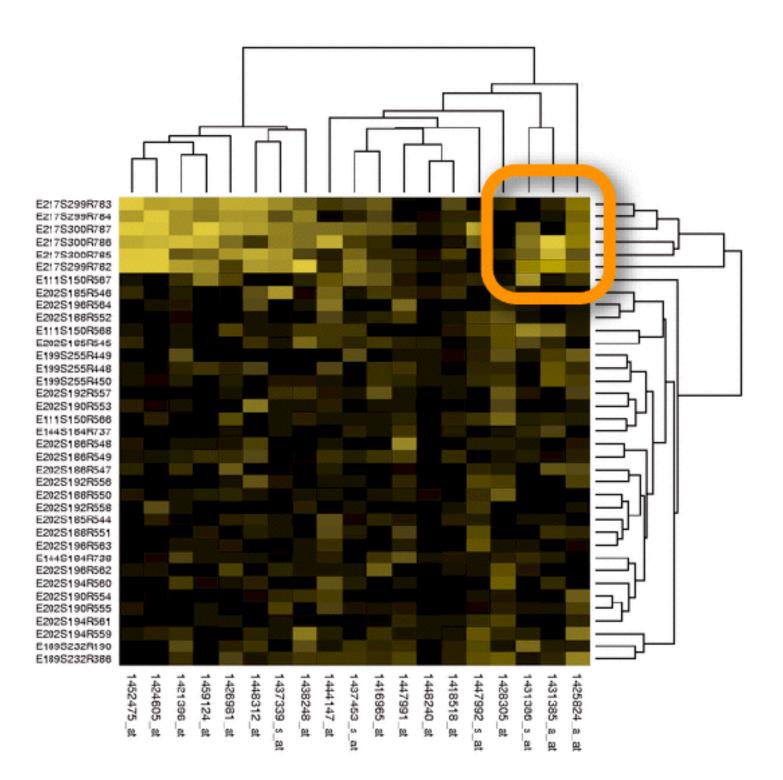
[Verhaak 2012]

homozygous deletion

Bad Color Mapping

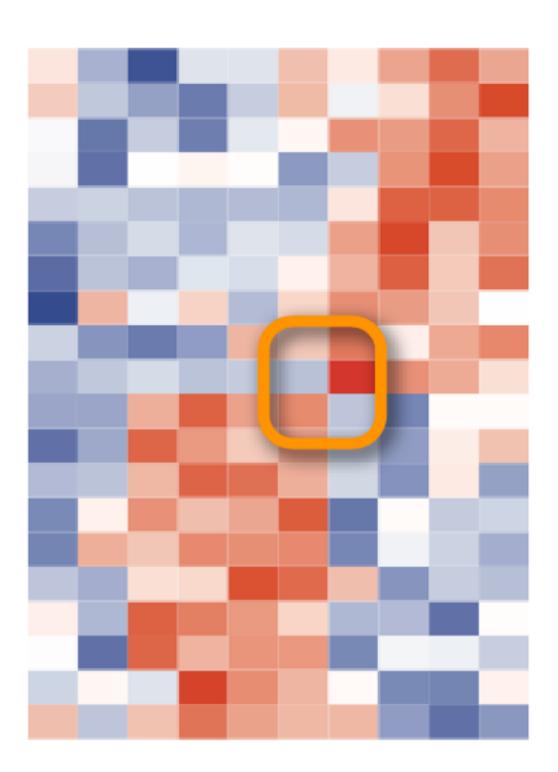


Normal Vision

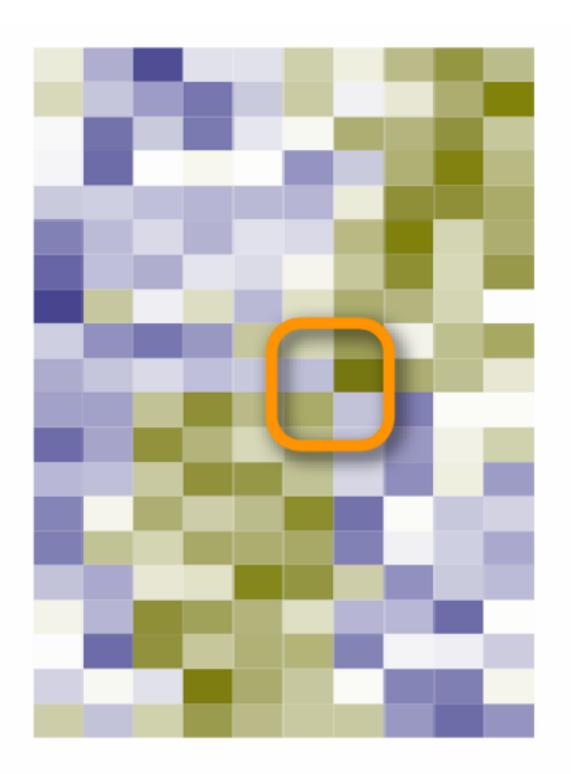


Deuteranope Vision ("Red-Green Blindness")

Good Color Mapping

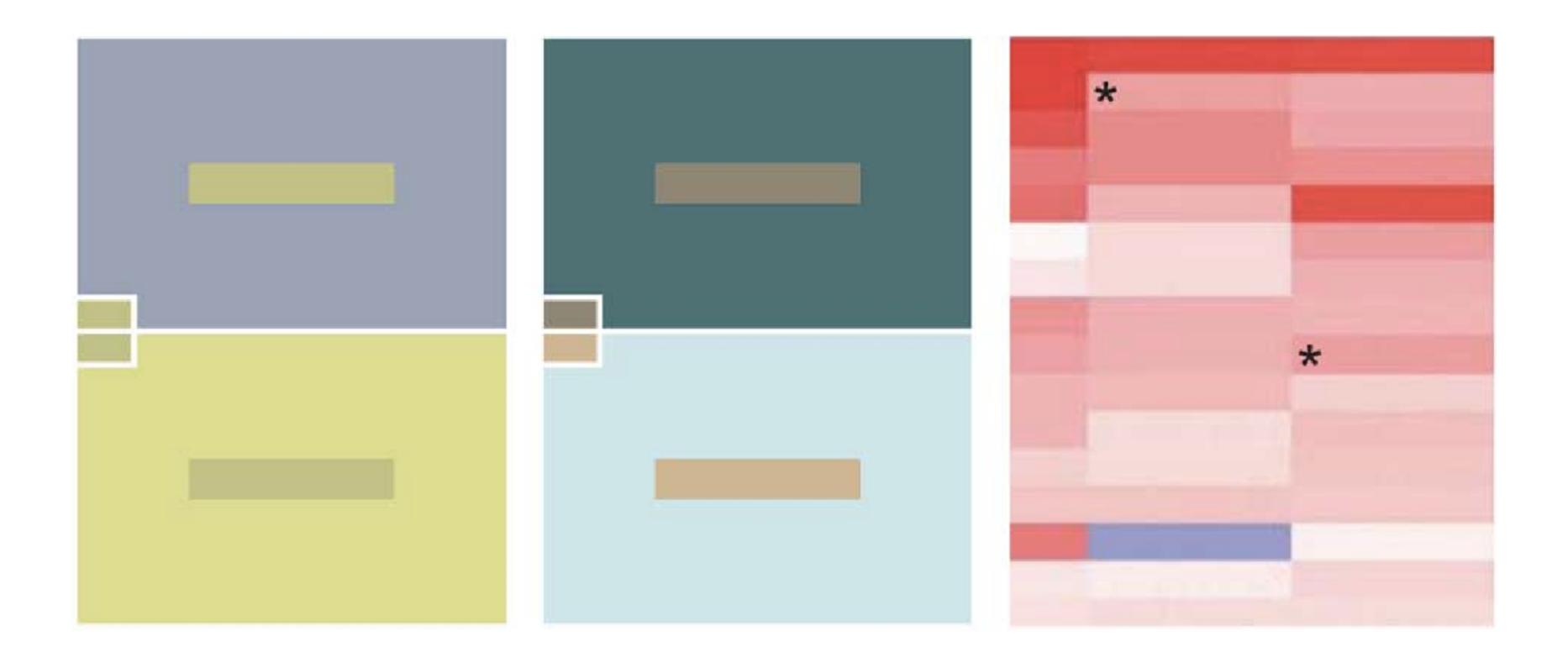


Normal Vision



Deuteranope Vision ("Red-Green Blindness")

Color is relative!





Clustering

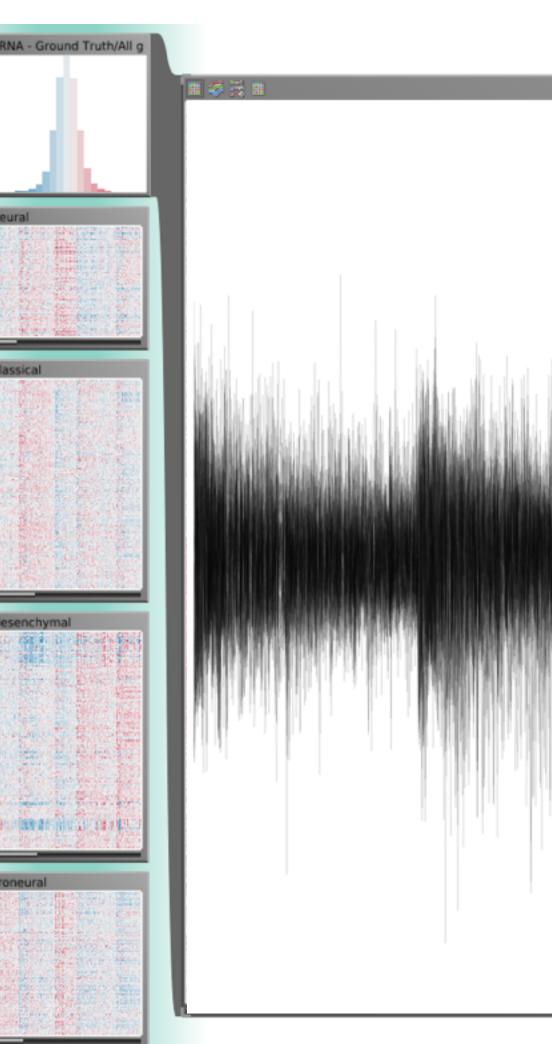
- Classification of items into "similar" bins
- Based on similarity measures
 - Euclidean distance, Pearson correlation, ...
- Partitional Algorithms
 - divide data into set of bins
 - # bins either manually set (e.g., kmeans) or automatically determined (e.g., affinity propagation)

- Hierarchical Algorithms Produce "similarity tree" – dendrogram
- **Bi-Clustering**
- Clusters dimensions & records
- Fuzzy clustering
- allows occurrence of elements in multiples clusters

Clustering Applications

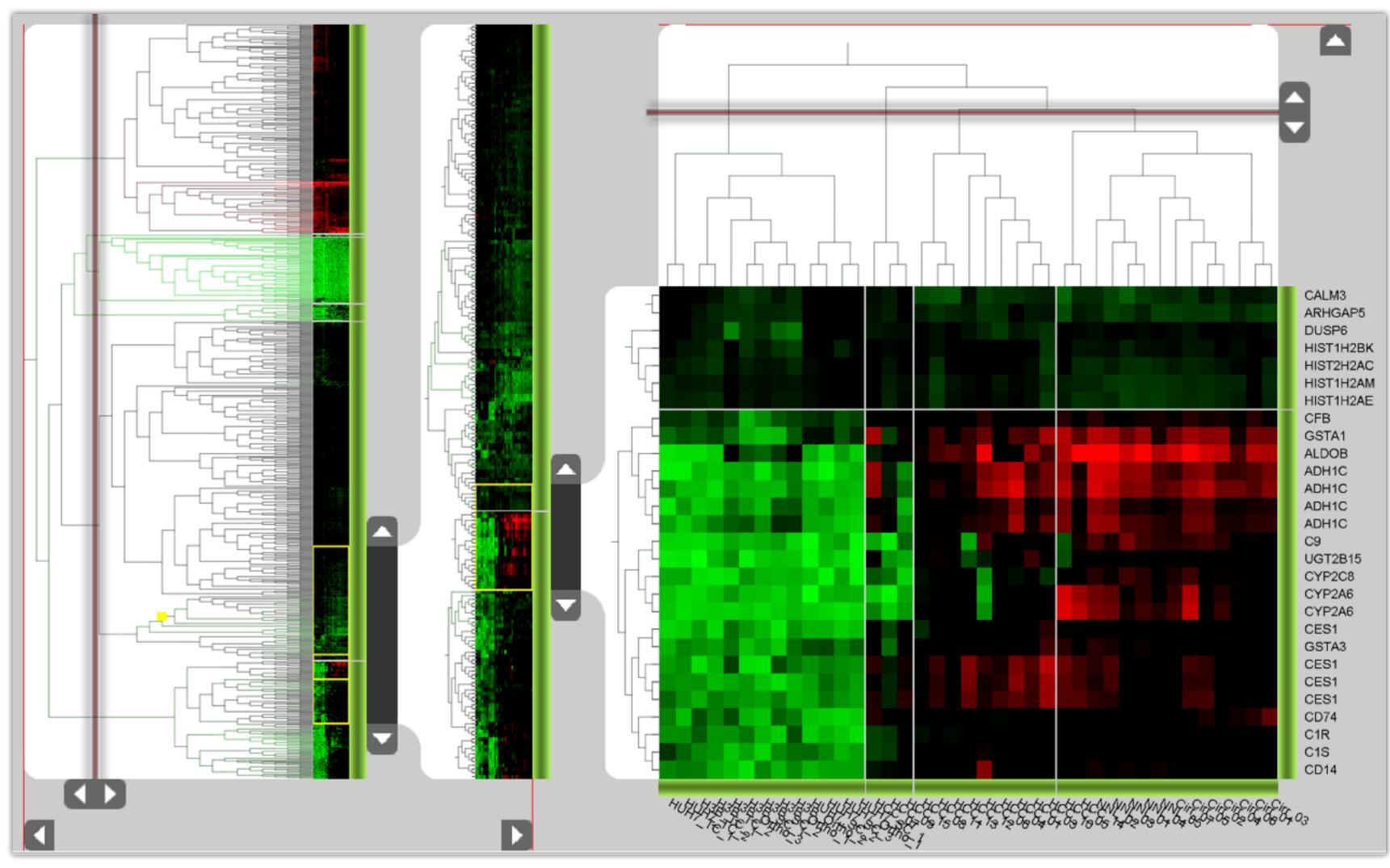
Clusters can be used to order (pixel based techniques) brush (geometric techniques) aggregate Aggregation cluster more homogeneous than whole dataset statistical measures, distributions, etc. more meaningful

Clustered Heat Map

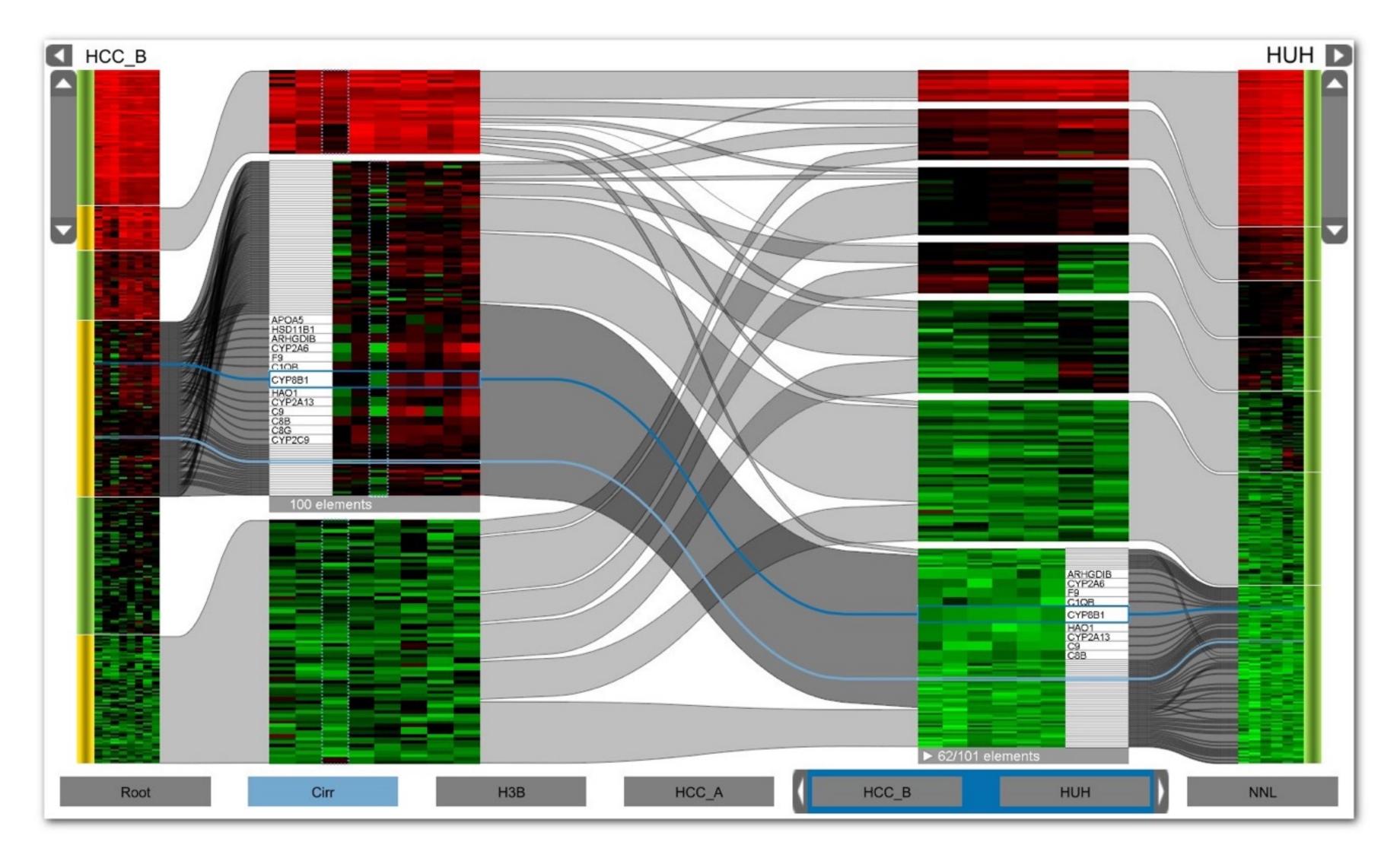




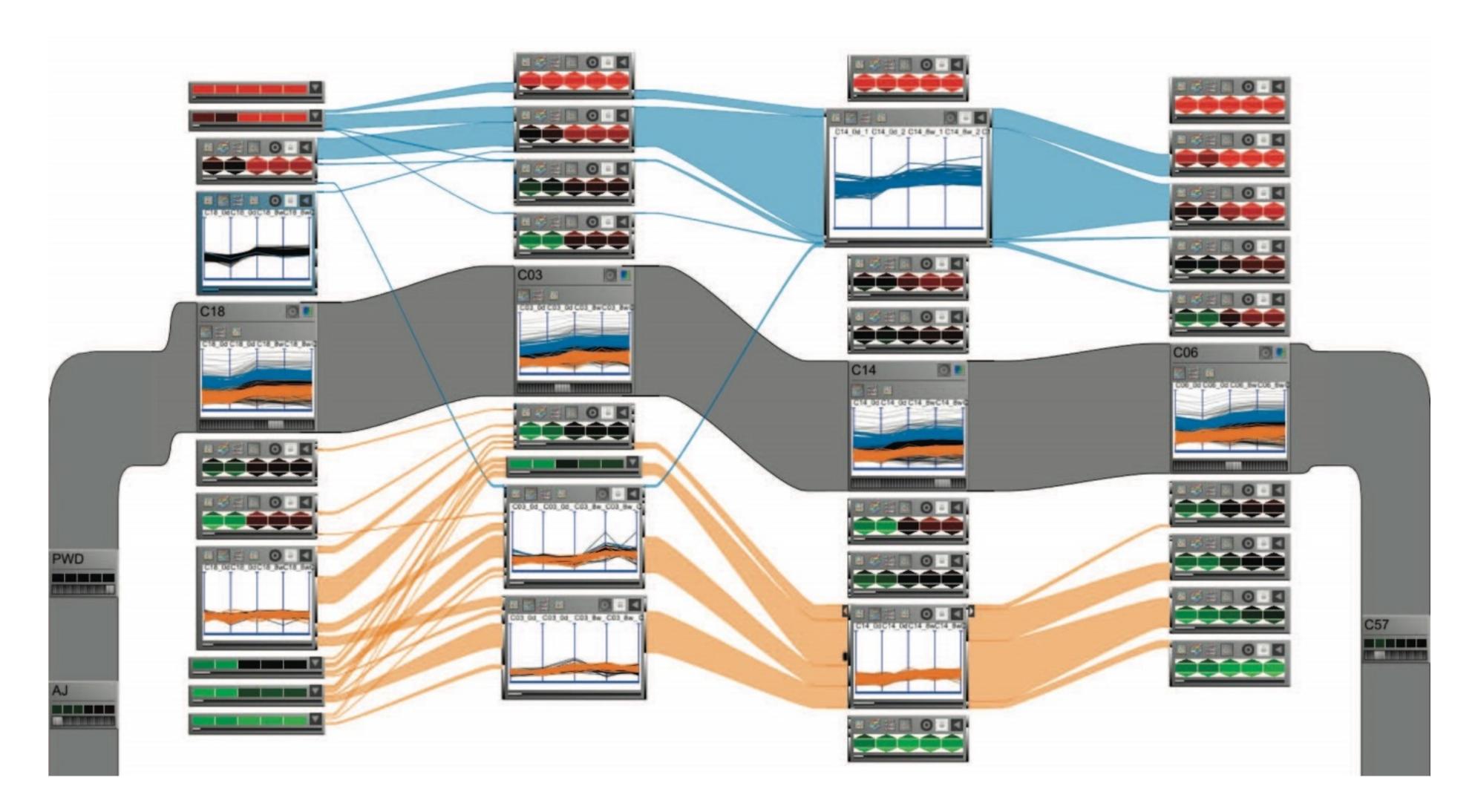
F+C Approach, with Dendrograms



Cluster Comparison

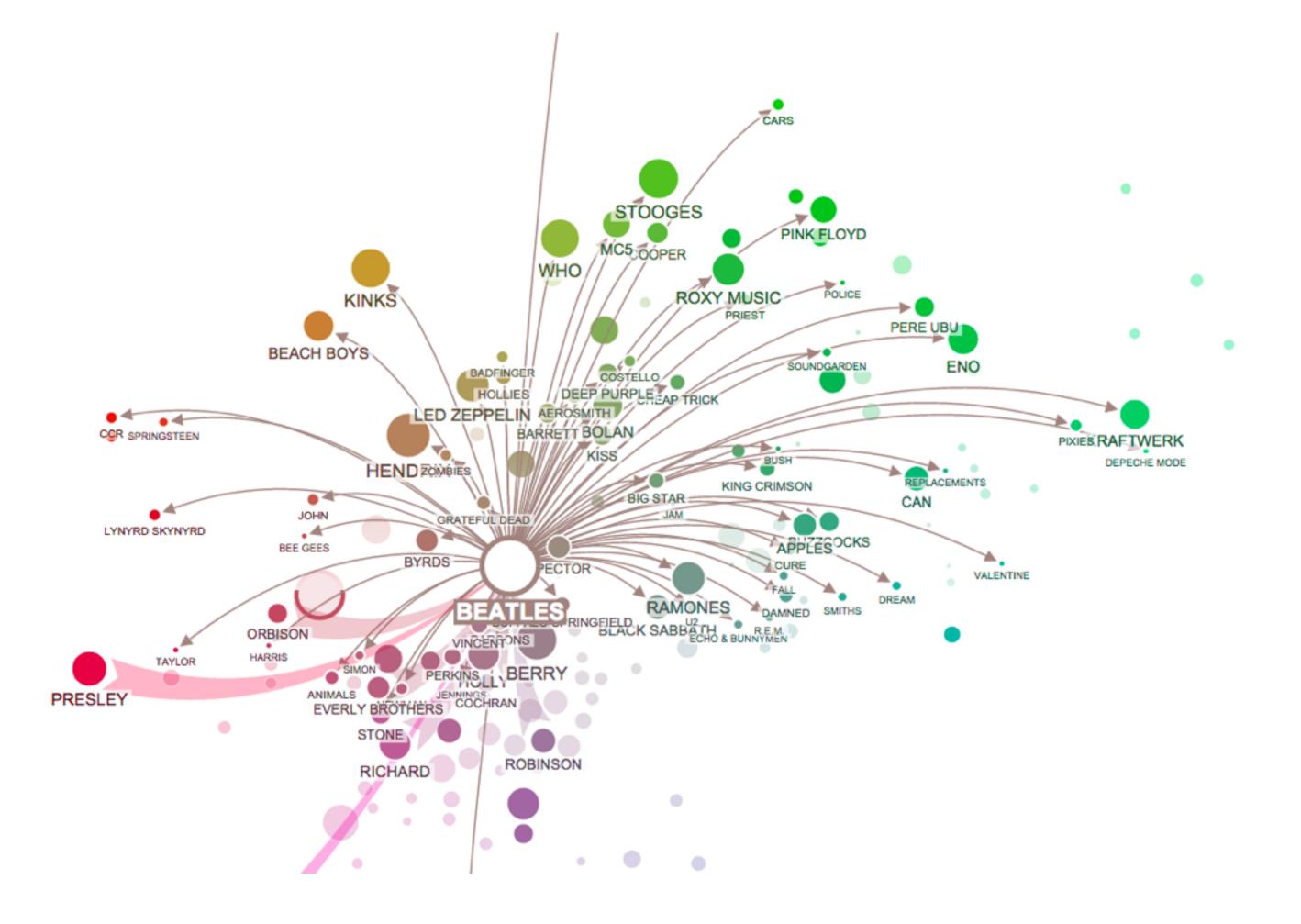


Aggregation



Design Critique

EdgeMaps: <u>http://goo.gl/q8Cv7t</u>

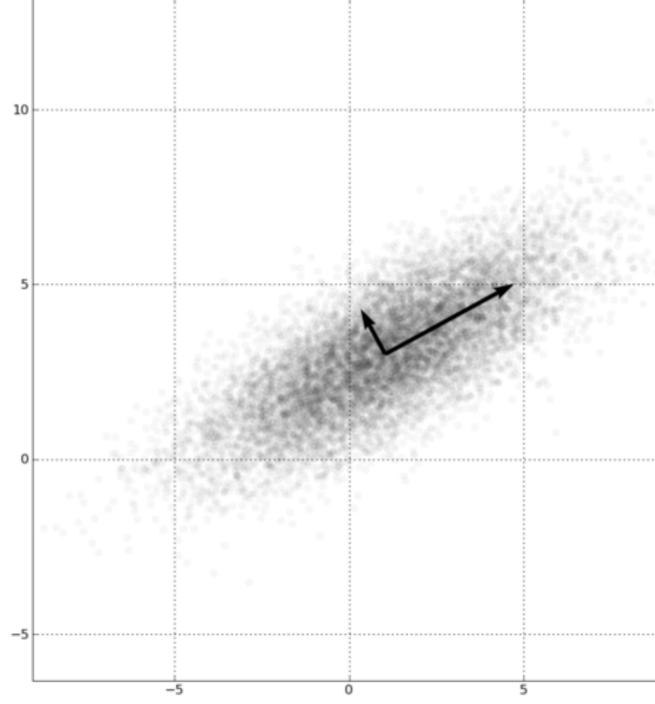


http://mariandoerk.de/edgemaps/demo/#music

Dimensionality Reduction

Dimensionality Reduction

- Reduce high dimensional to lower dimensional space
- Preserve as much of variation as possible
- Plot lower dimensional space Principal Component Analysis
 - linear mapping, by order of variance



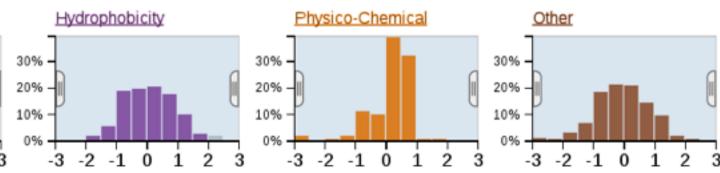


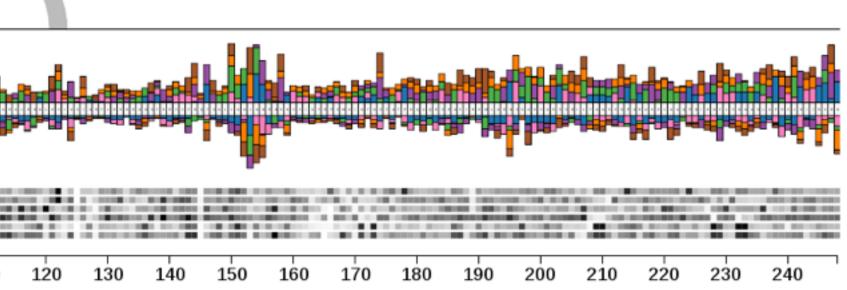
PCA Example – CS 171 Project

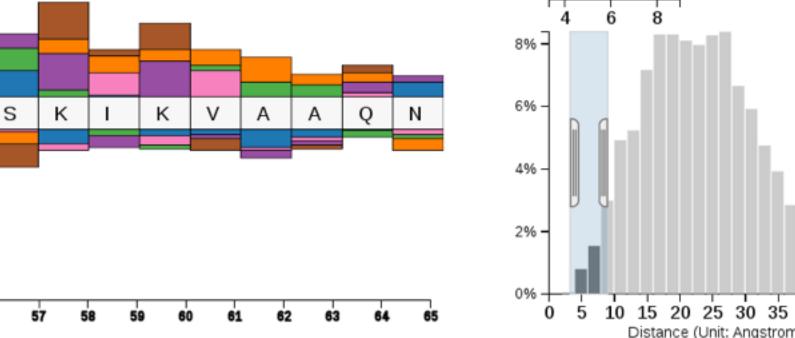
2013

Alpha 30% 20% 10% -3 -2		1 2	30% - 20% - 10% - 0% - 3 -	Beta Sh	l	1 2	309 209 109 09 3		2 -1		2 3
10	20	30	40		60 4 -	70	80	9	0	100	110
	A CONTRACT				2 - R 0 - R 0 - R 2 - 4 4 - 6 6 - 8 - 8		A		K	G	S
				-	50 Stack	51 ked	52	53	54	55 T	56 5

http://mu-8.com/





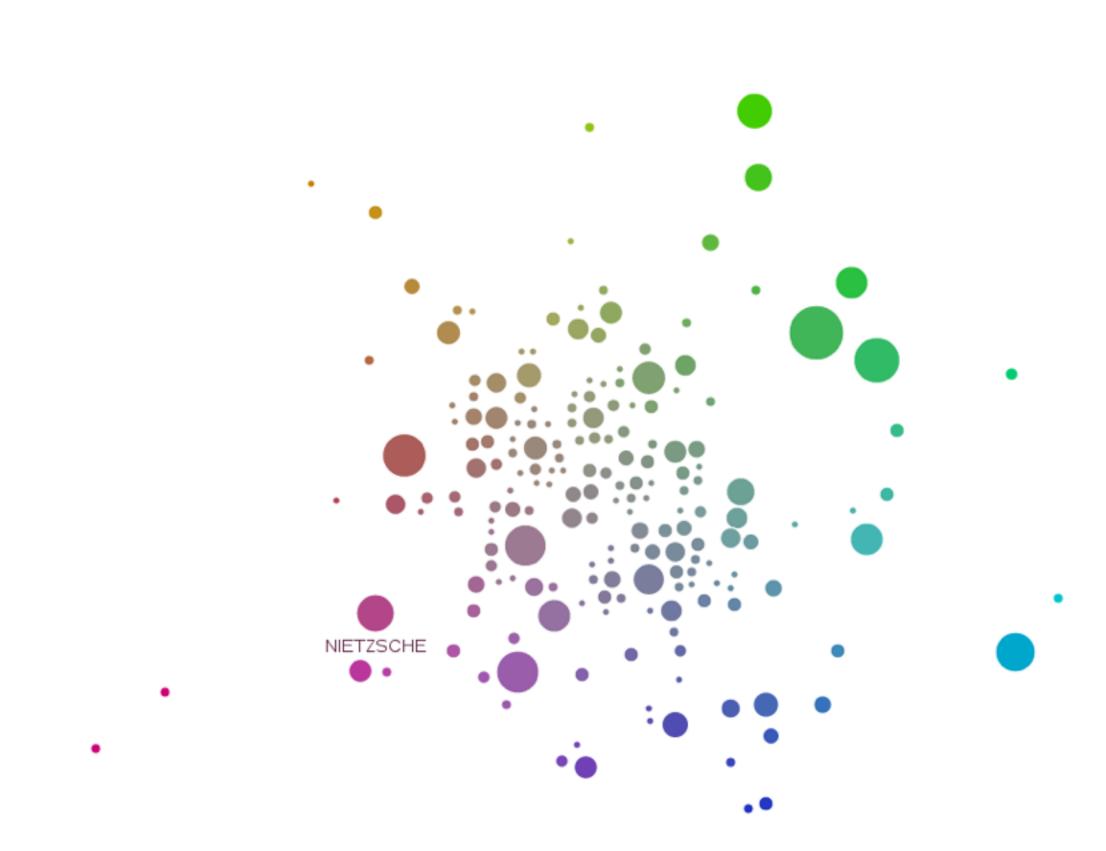


[Mercer & Pandian]

Multidimensional Scaling

Nonlinear, better suited for some DS

Popular for text analysis



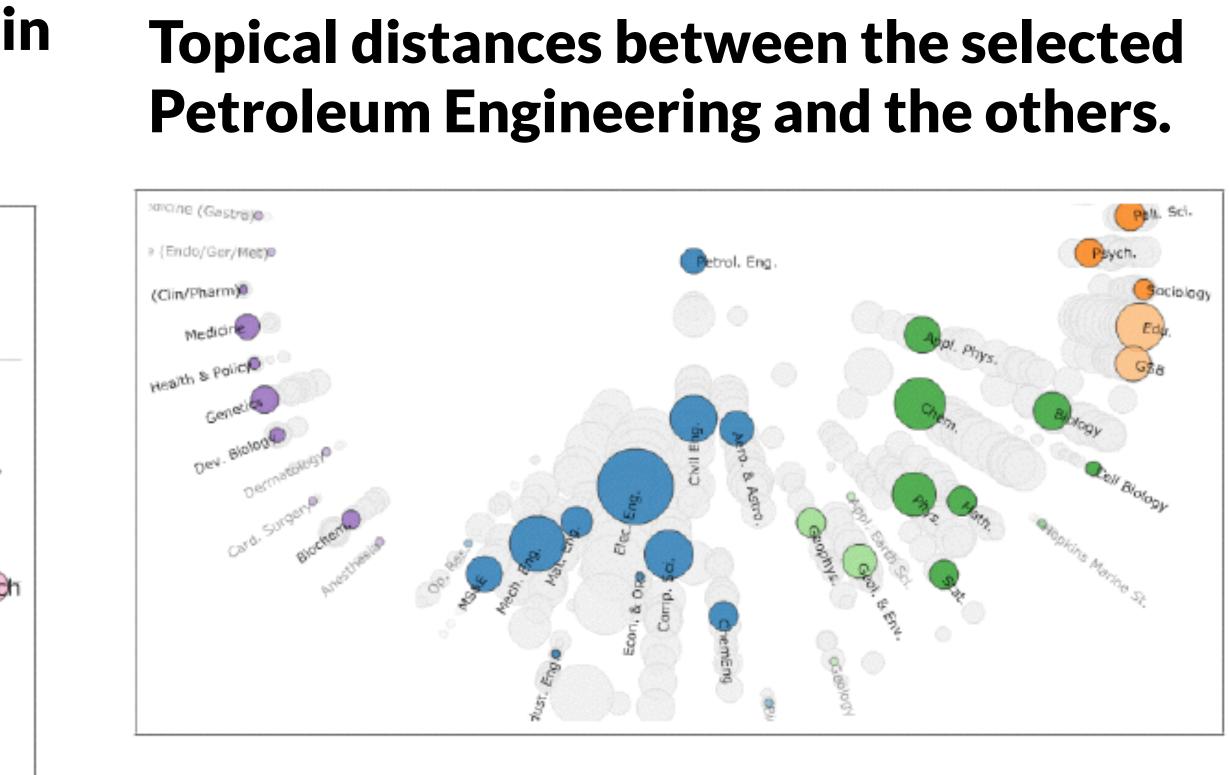
[Doerk 2011]

Can we Trust Dimensionality Reduction?

Topical distances between departments in a 2D projection



http://www-nlp.stanford.edu/projects/dissertations/browser.html



[Chuang et al., 2012]

