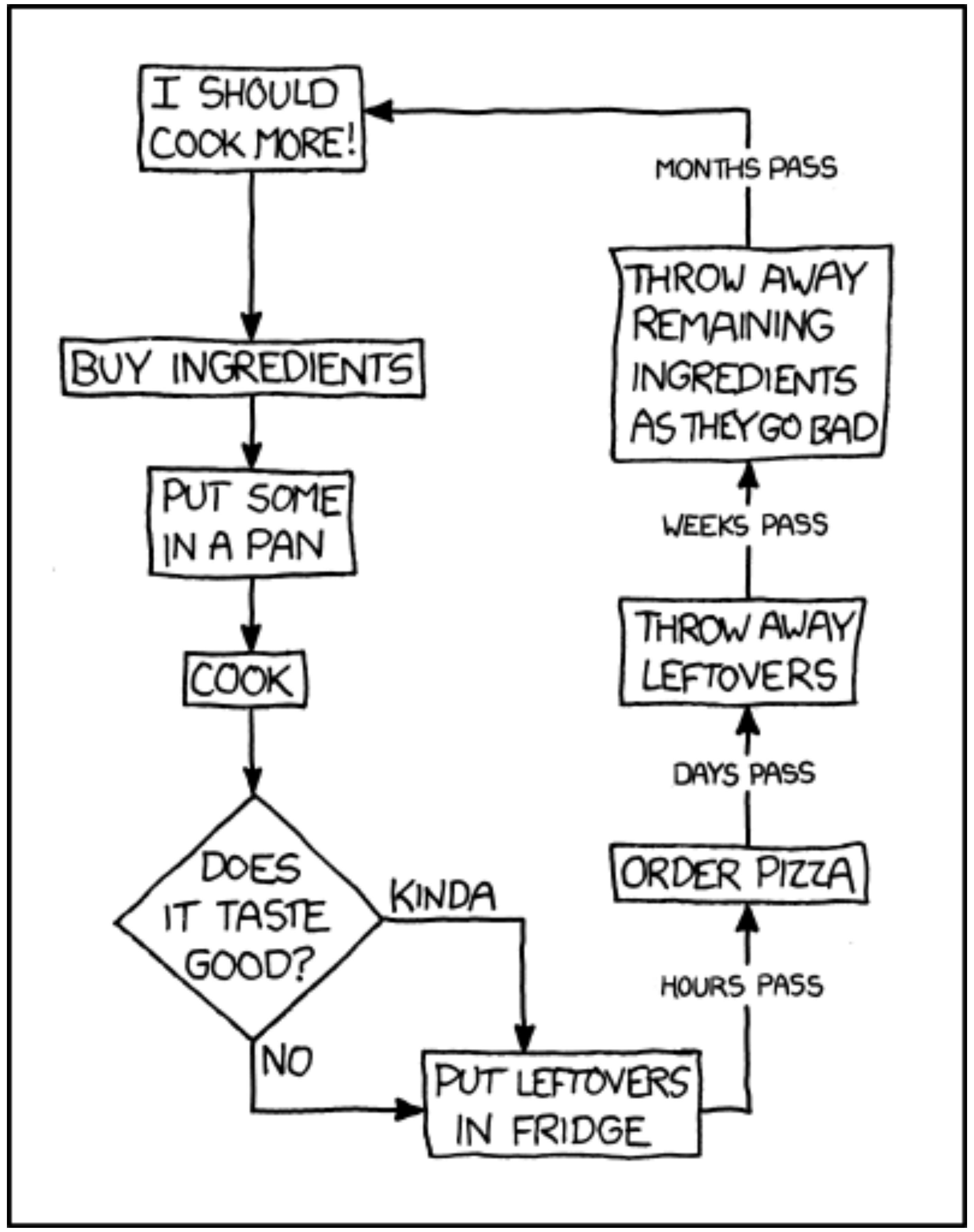
CS171 Visualization

Alexander Lex
alex@seas.harvard.edu

Tables Part II



[xkcd]

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

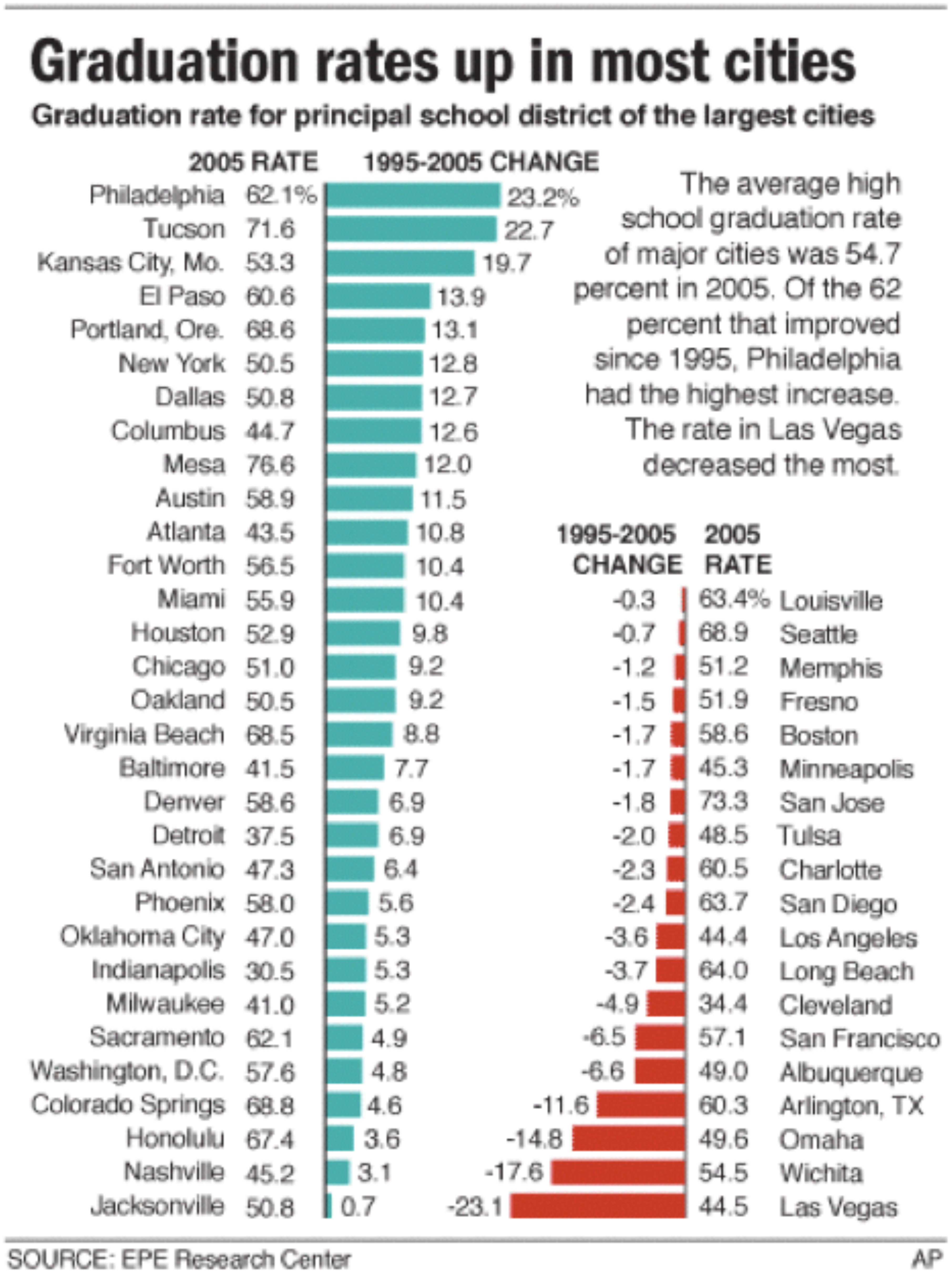
Updates

- Design Studio moved to Thursday
- Project Proposal moved to HW 4

Tables & Multi-Dimensional Data

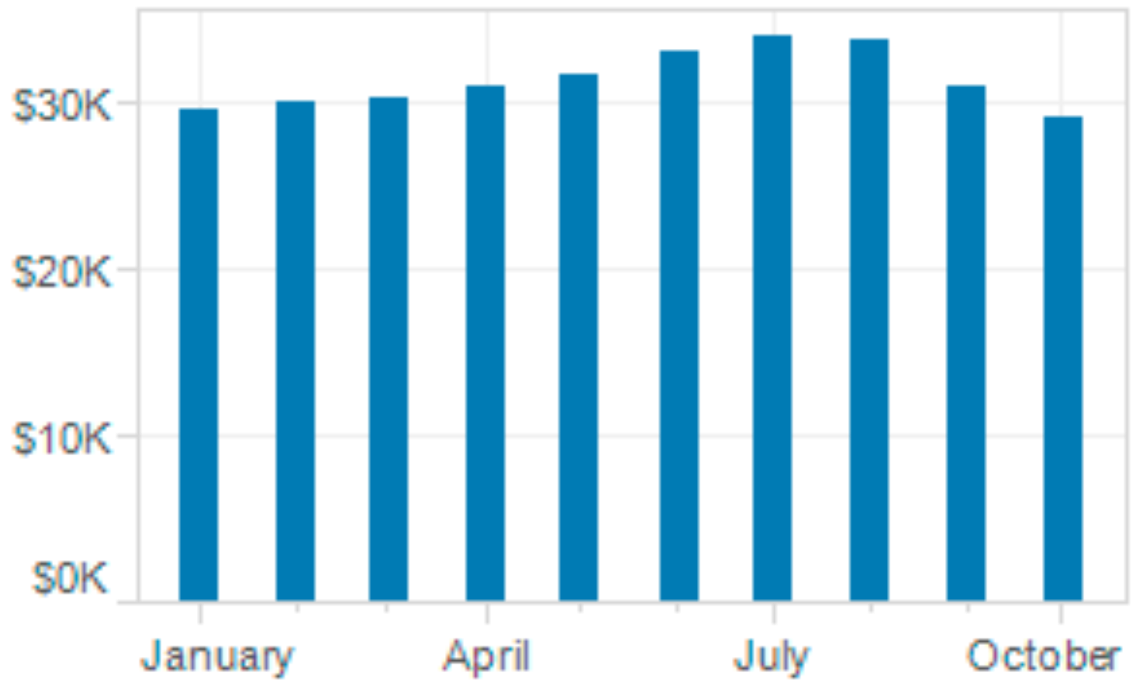
Comparisons

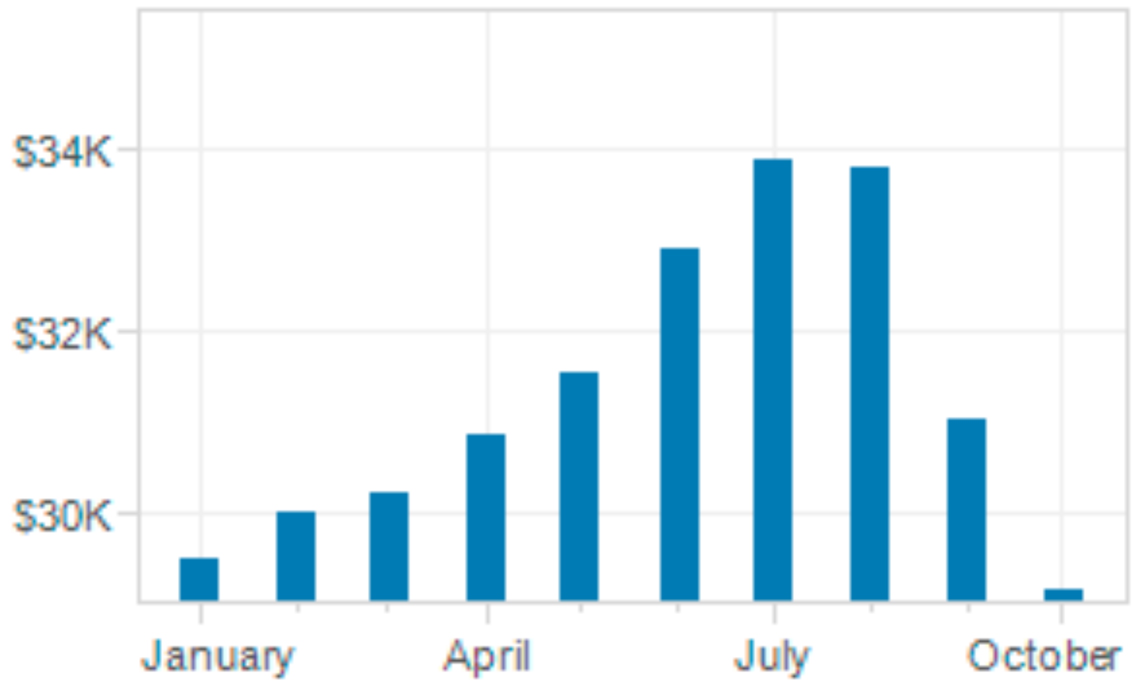
Direction

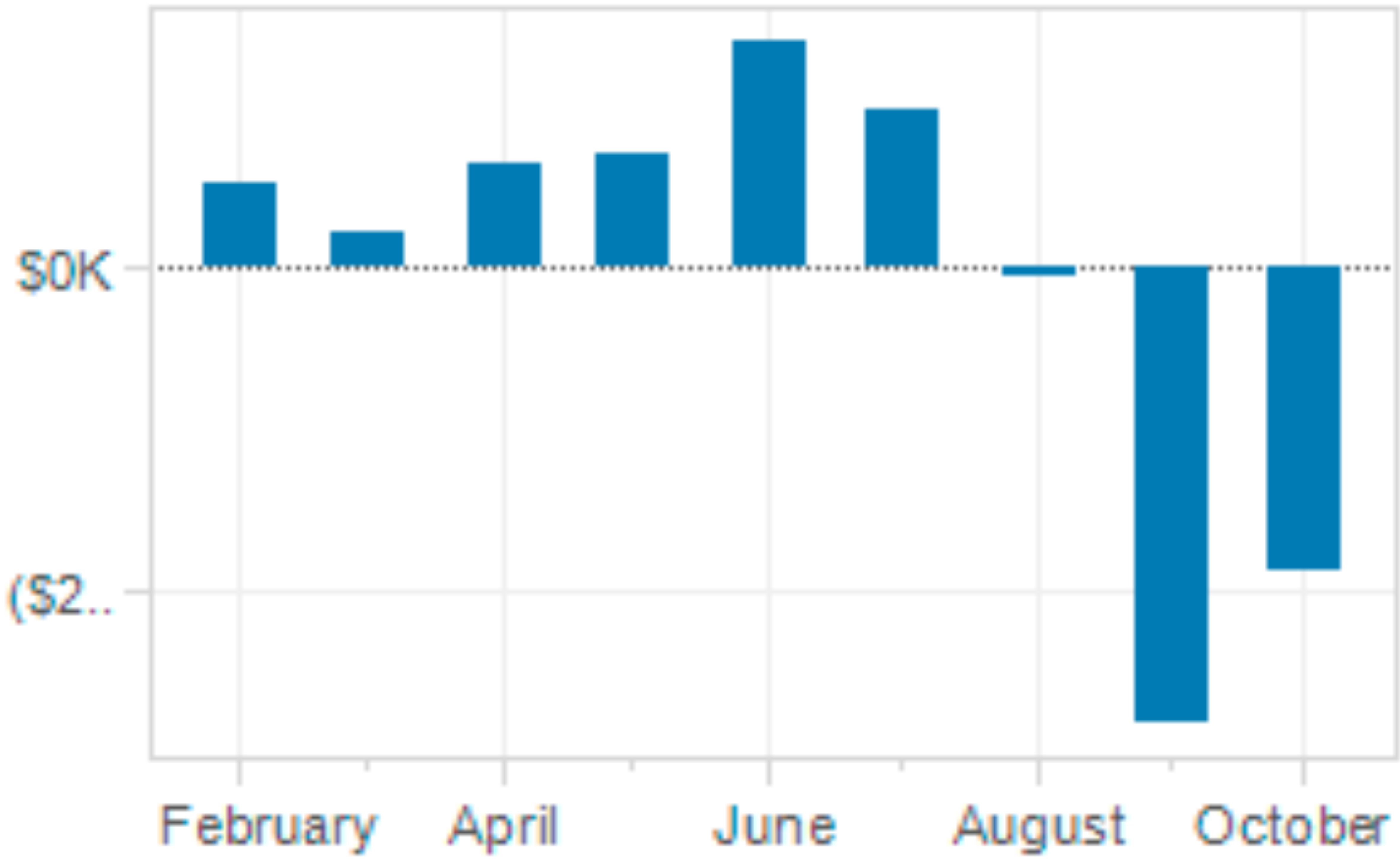


Nicolas Rapp

Plot Change Instead

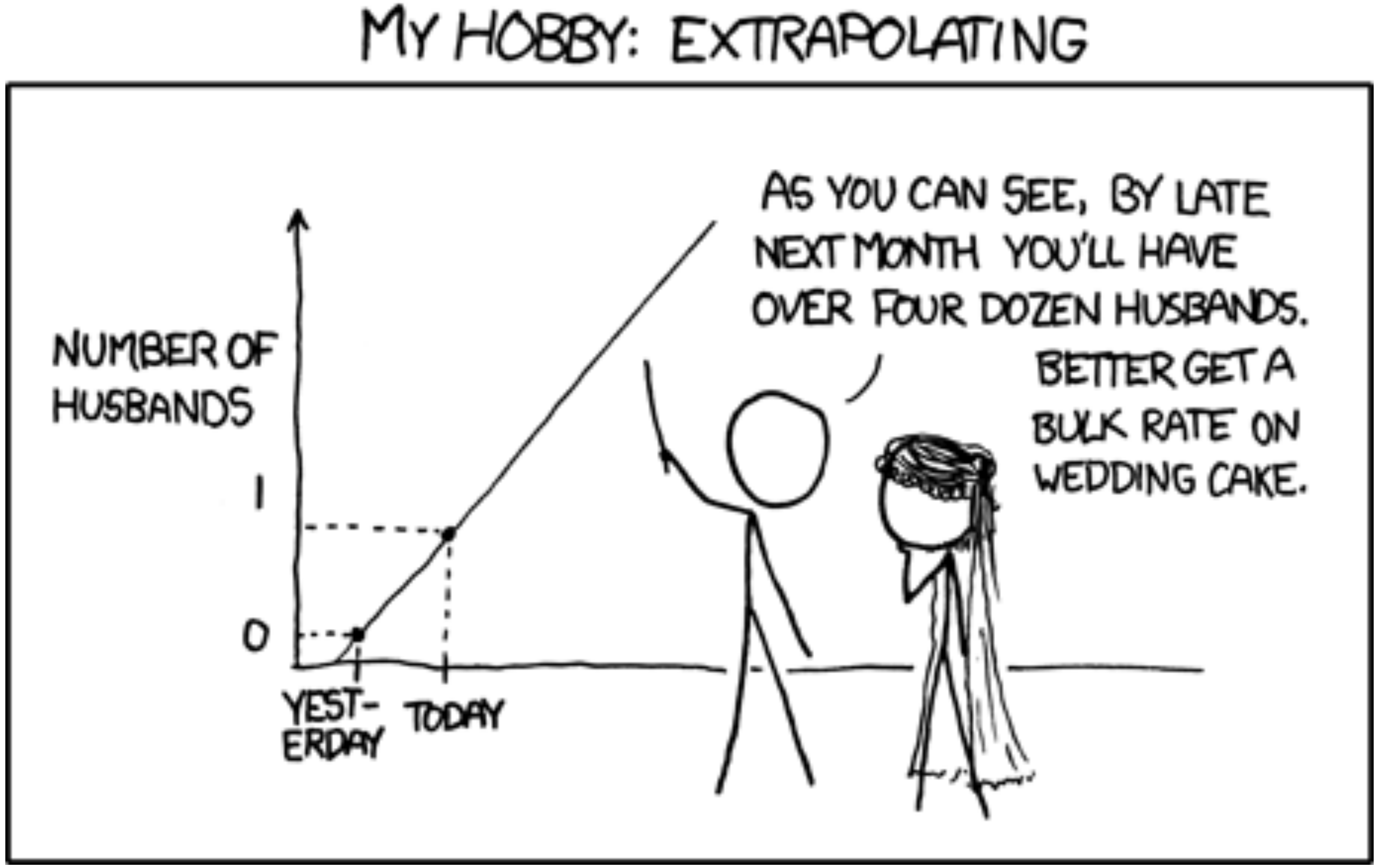






https://eagereyes.org/basics/baselines

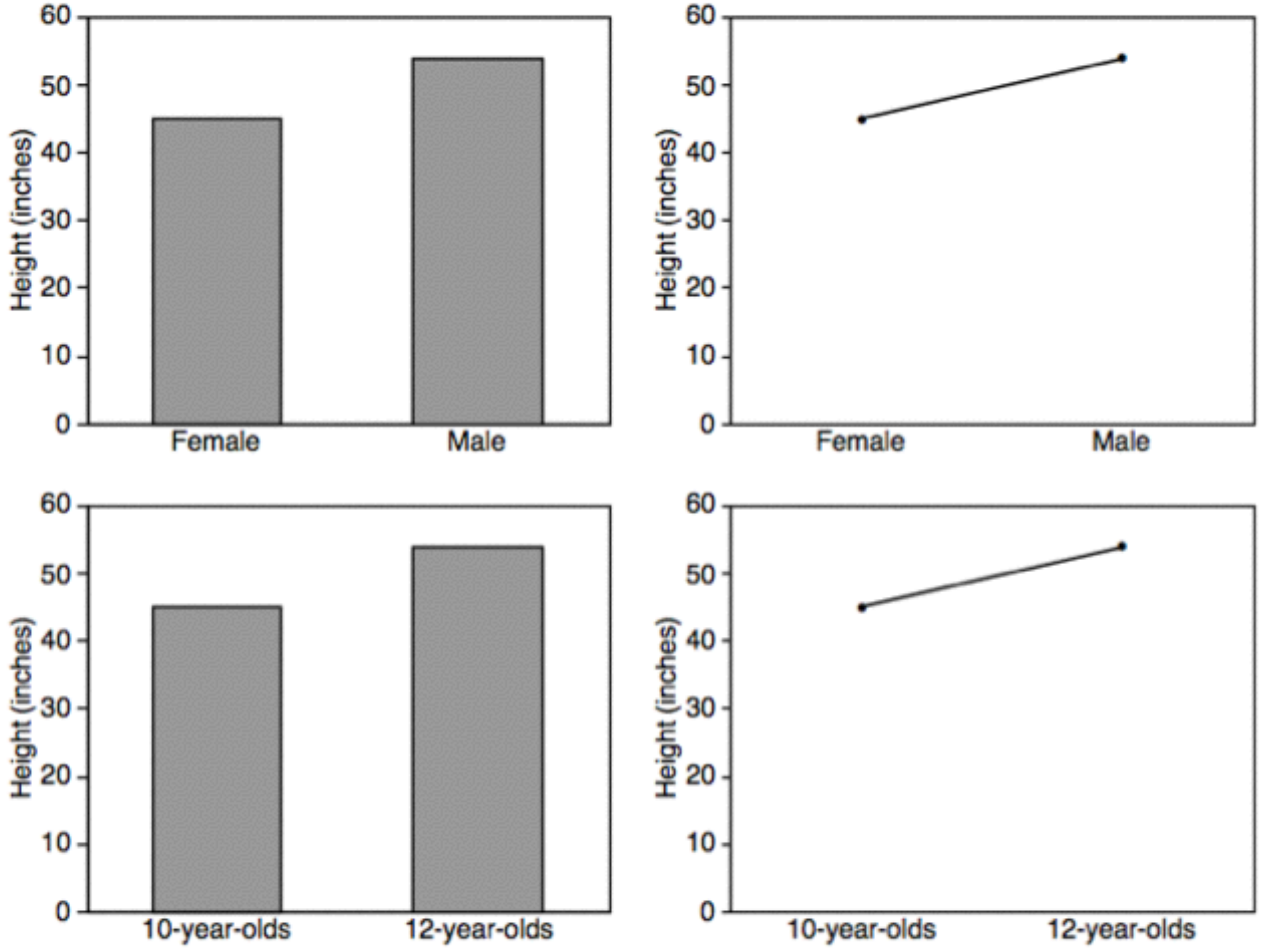
Trends Over Time



http://xkcd.com/605/

Bars vs. Lines

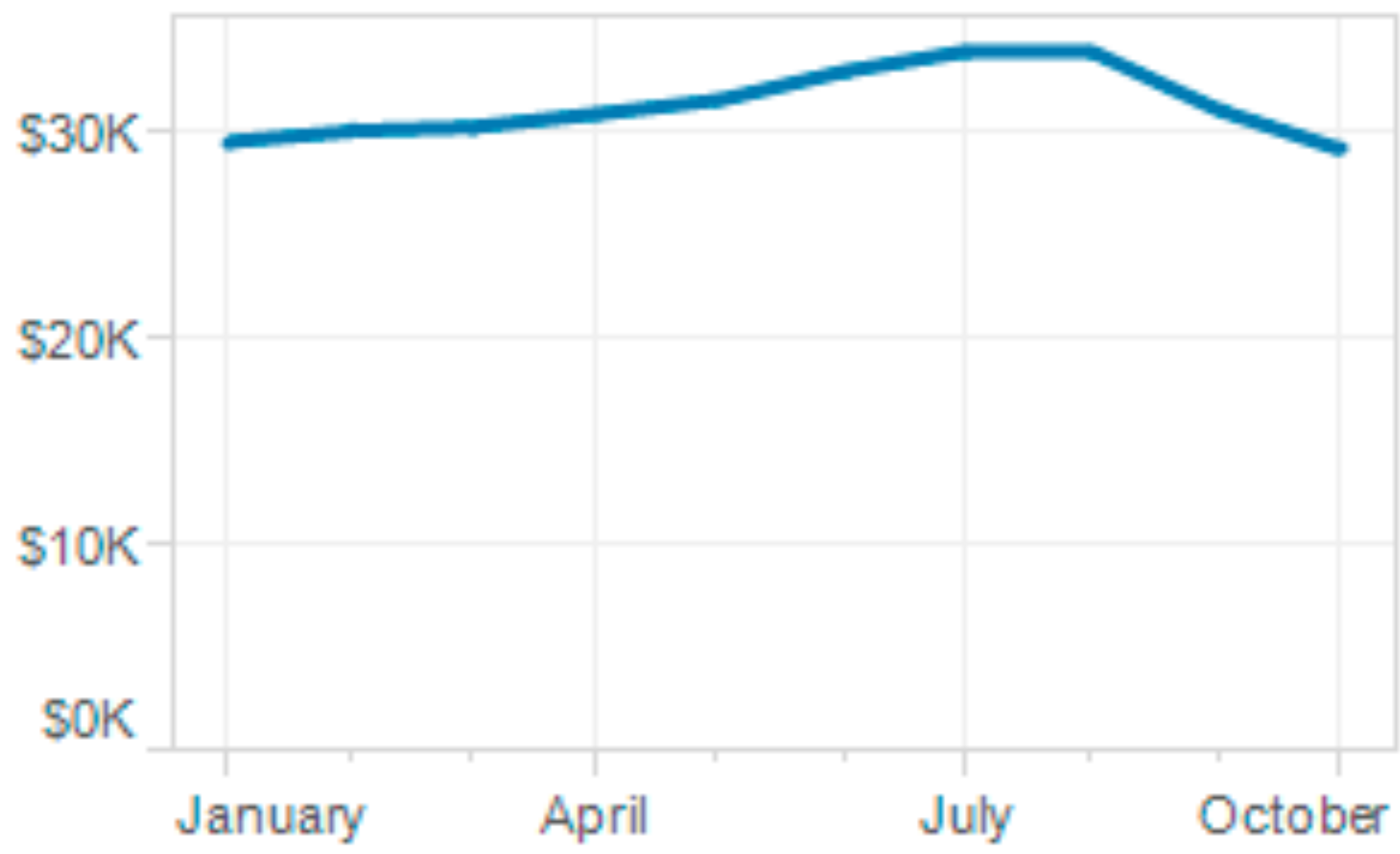
Lines imply connections & sampling from continuous data.
Do not use for categorical data.



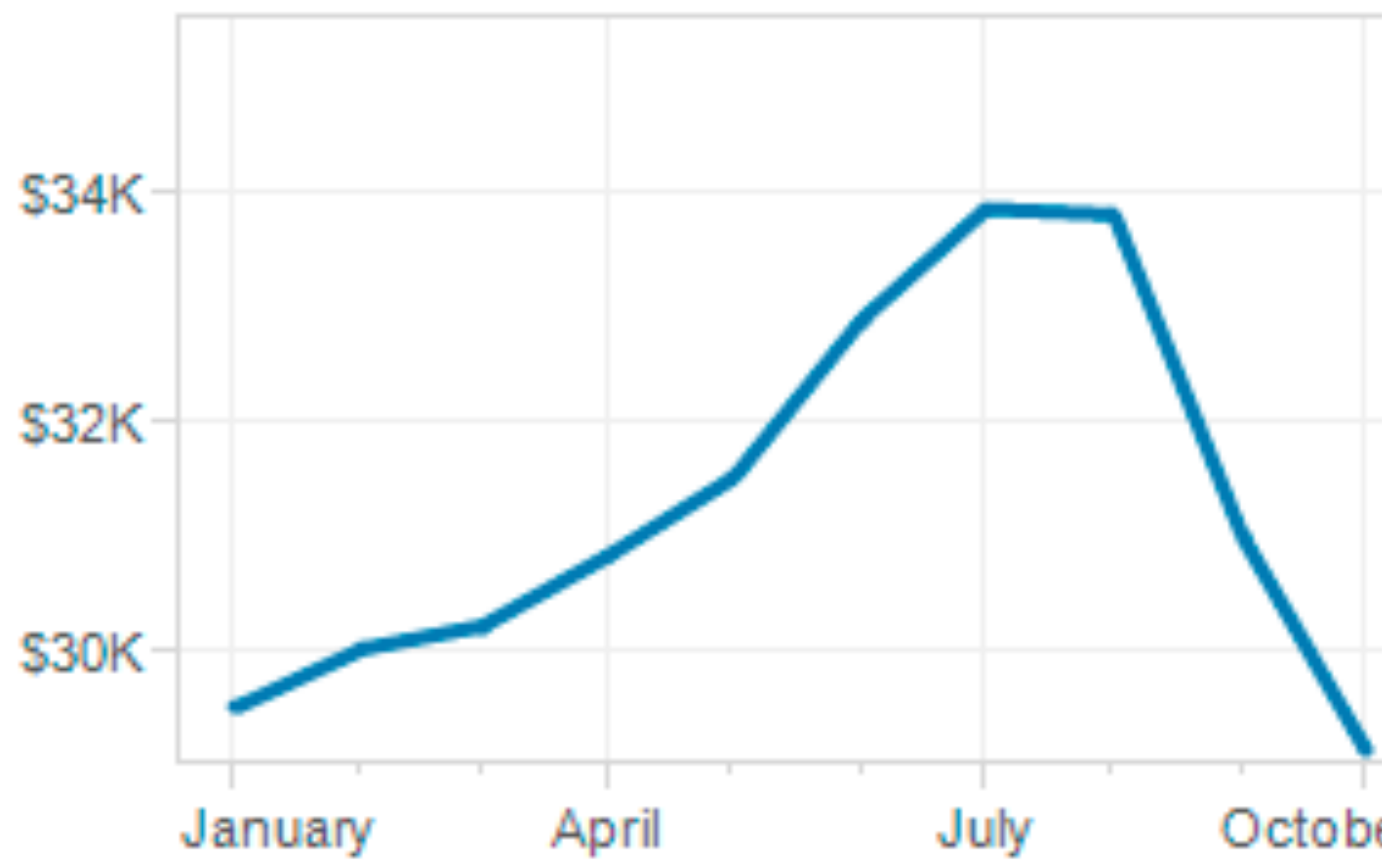
Zacks 1999

Baseline Problem (again)

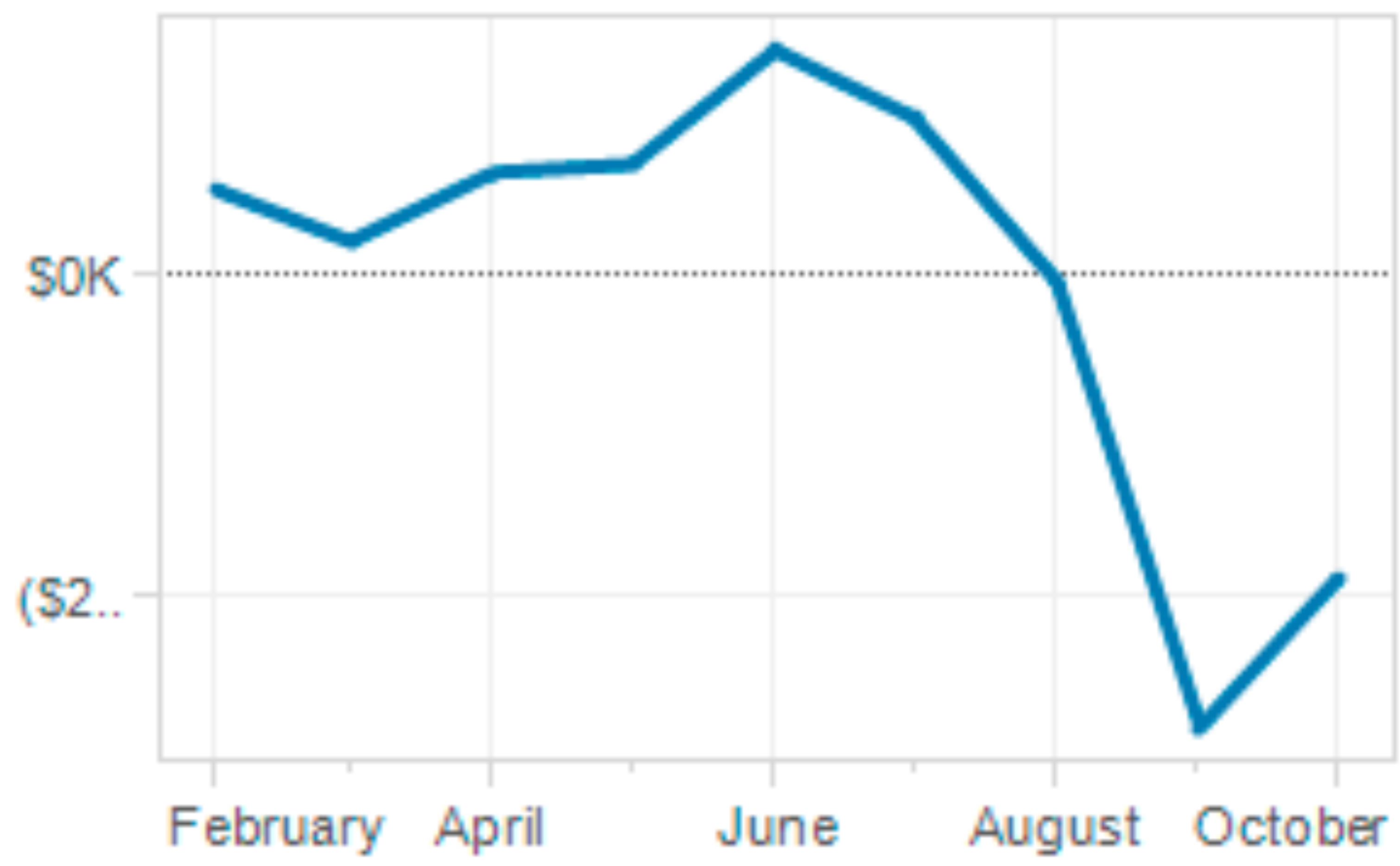
True Baseline



Clipped Baseline

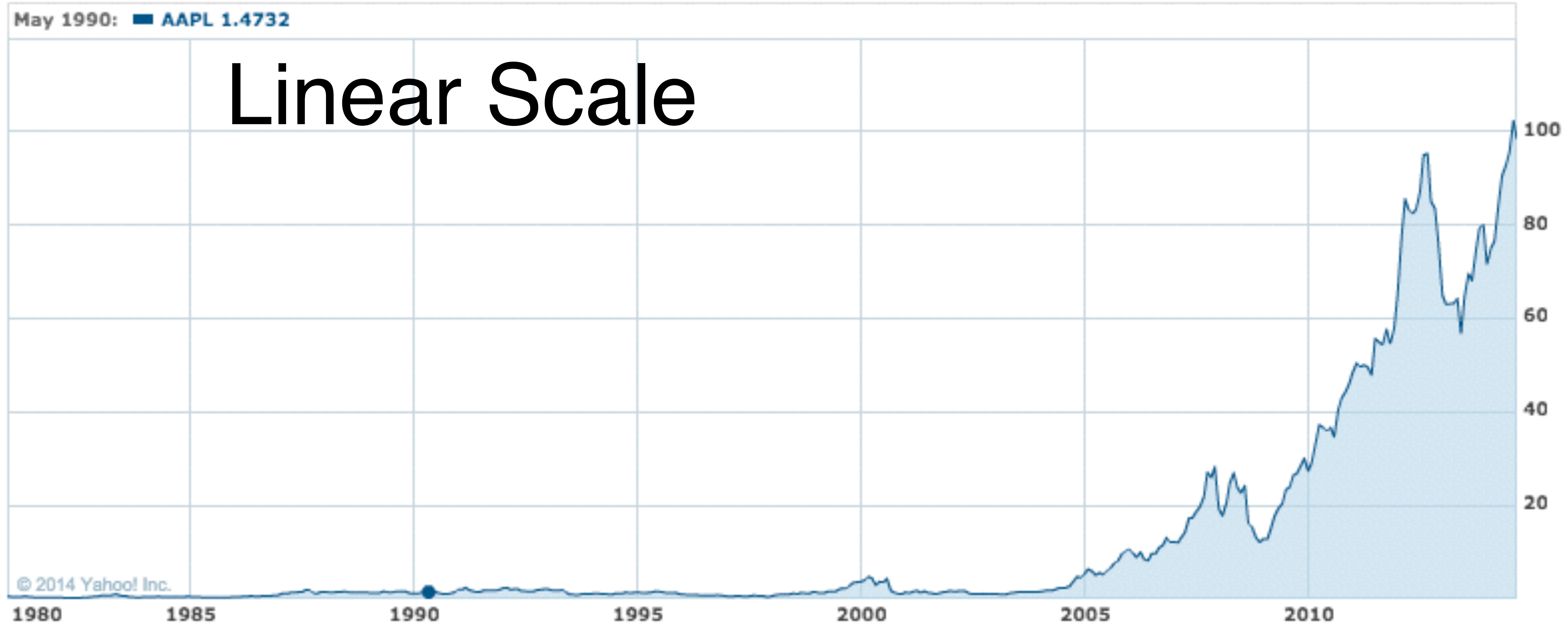


Plotting Change

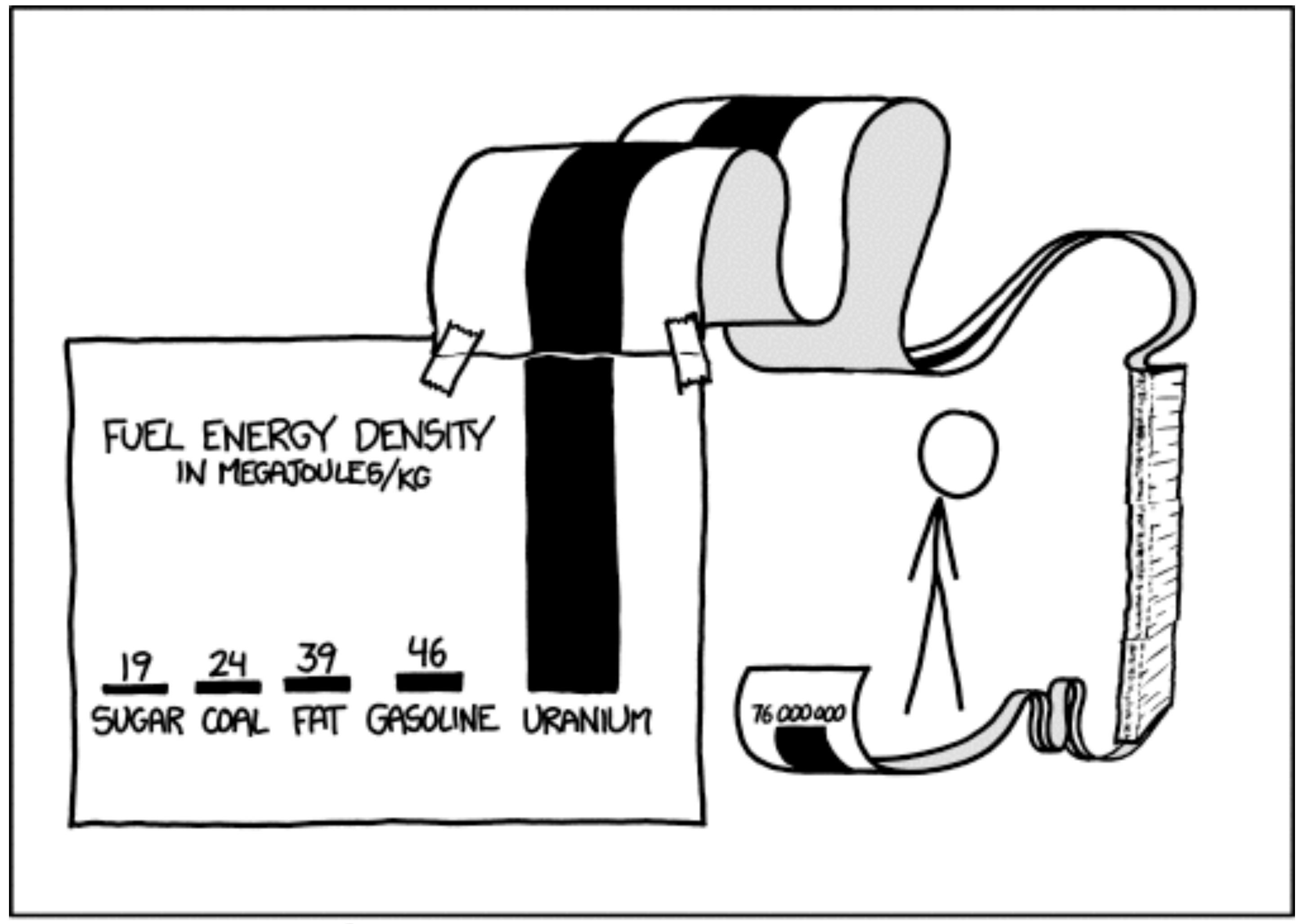


https://eagereyes.org/basics/baselines

Linear vs. Logarithmic Scale



Apple Stock Price



SCIENCE TIP: LOG SCALES ARE FOR QUITTERS WHO CAN'T FIND ENOUGH PAPER TO MAKE THEIR POINT *PROPERLY.*

http://xkcd.com/1162/

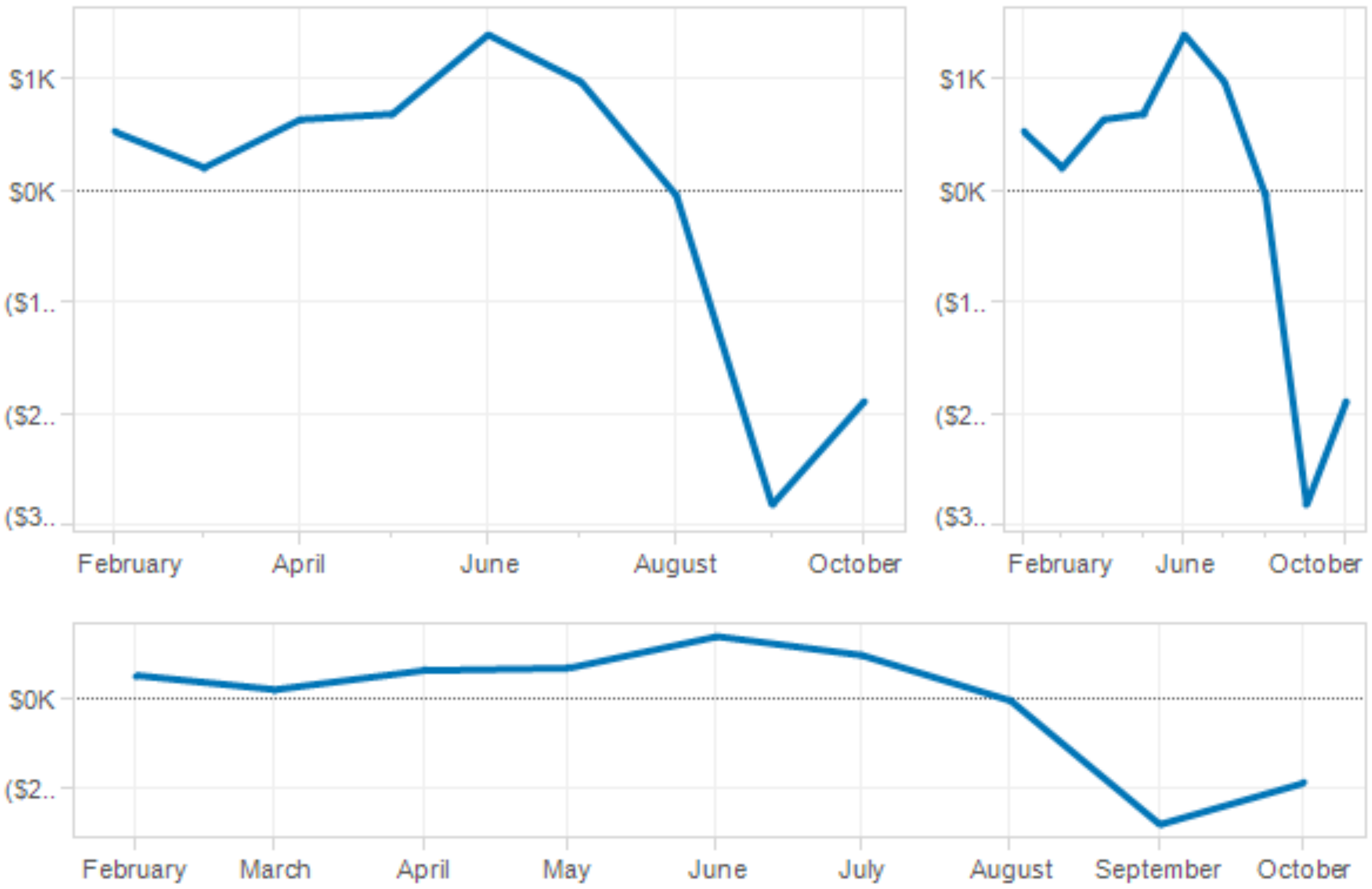
http://finance.yahoo.com/echarts?s=AAPL

Aspect Ratios

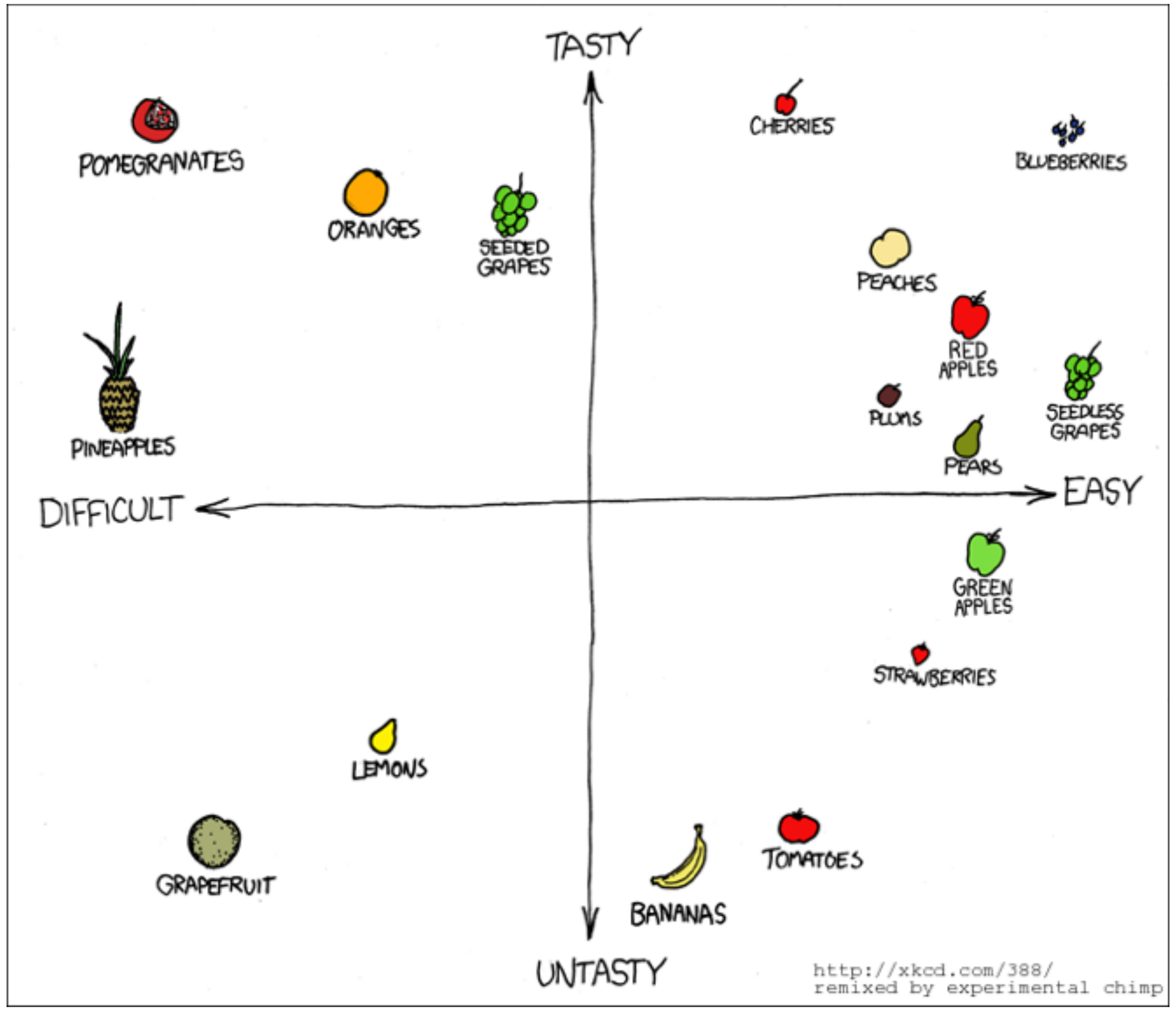
Rule of Thumb:

Banking to 45º

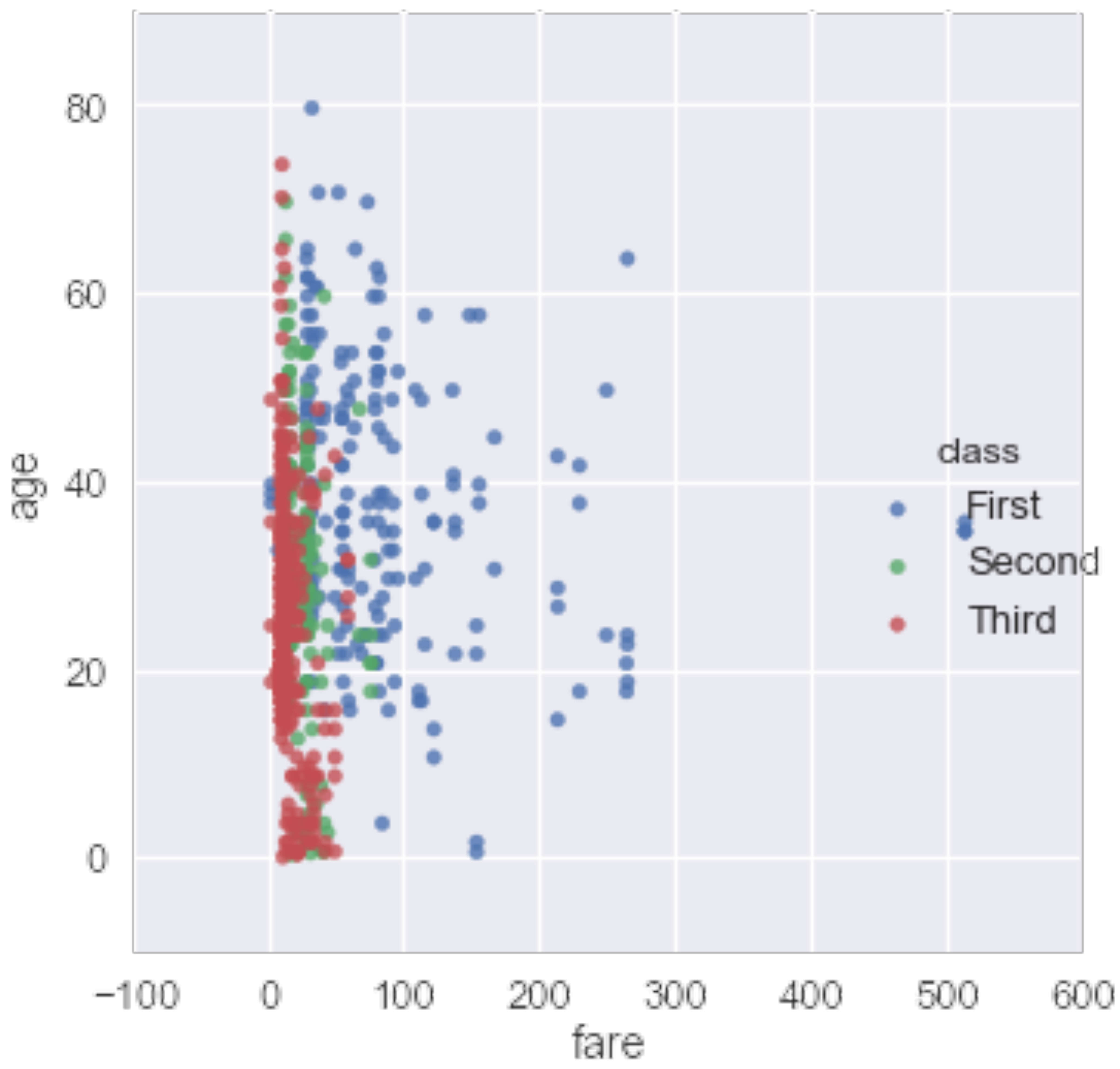
(average line slope: 45º)



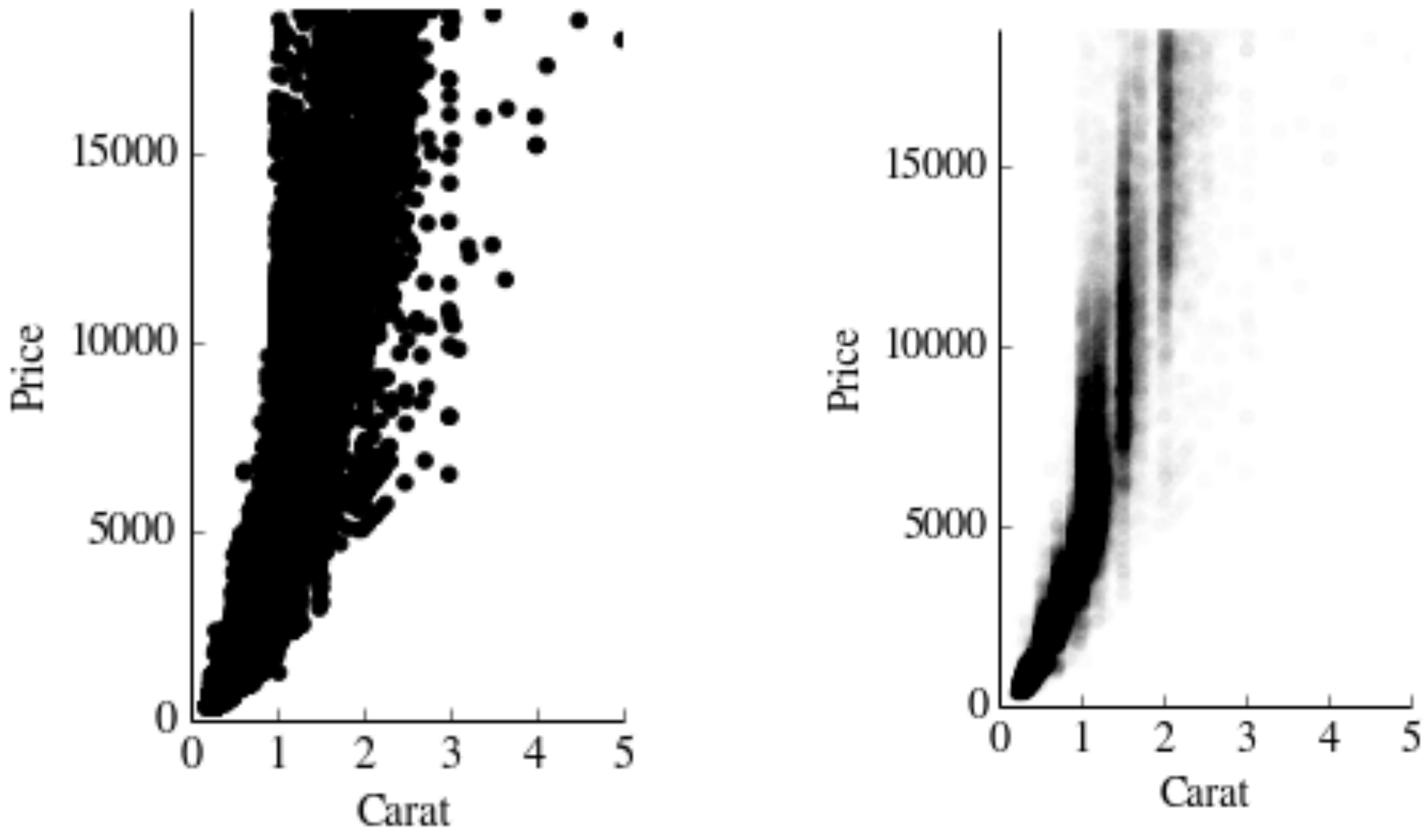
Correlations



Scatterplots



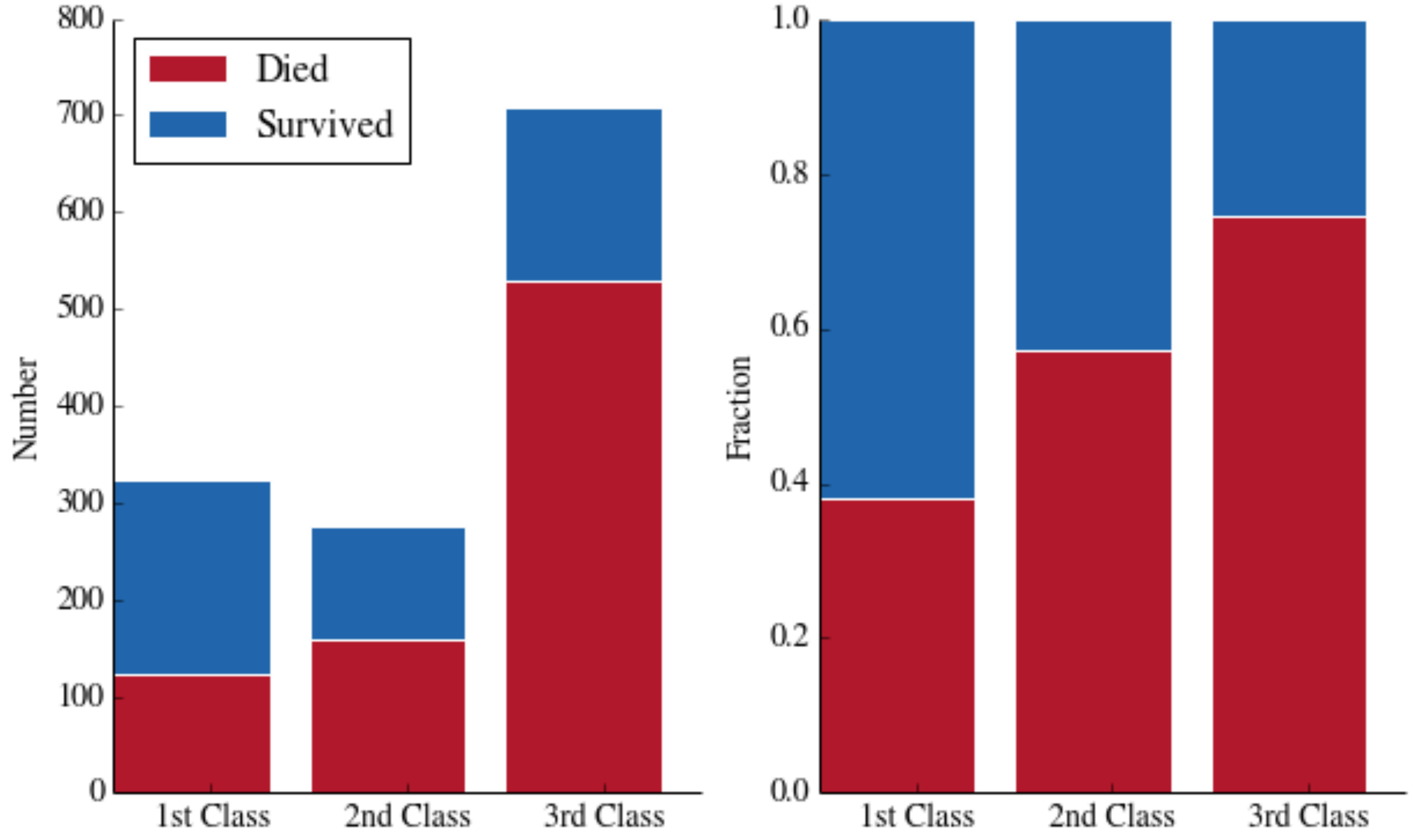
Overplotting



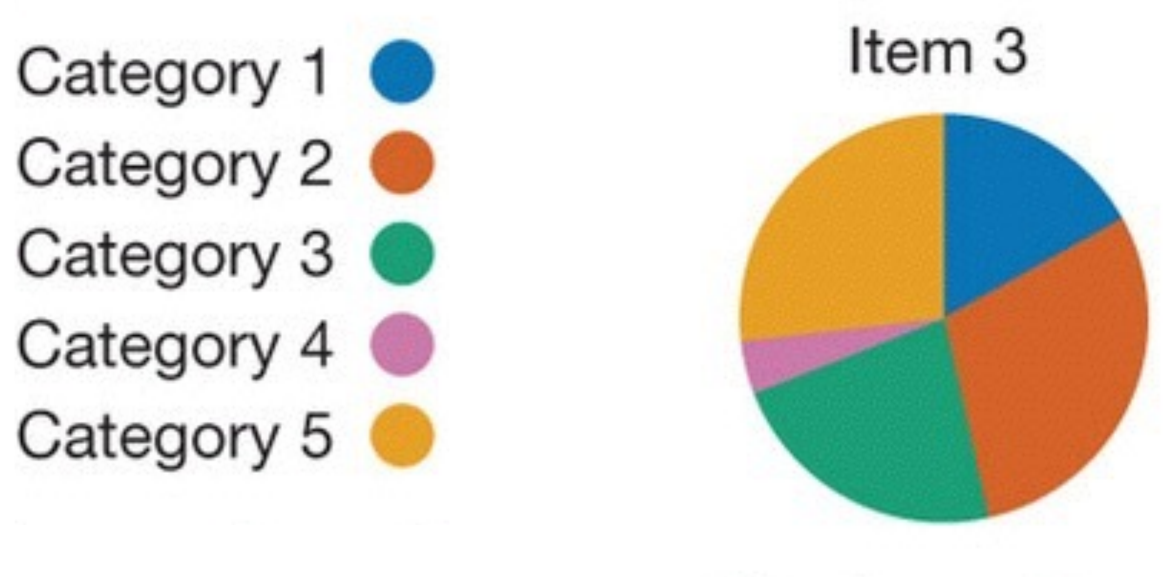
alpha = 1/100

Compositions

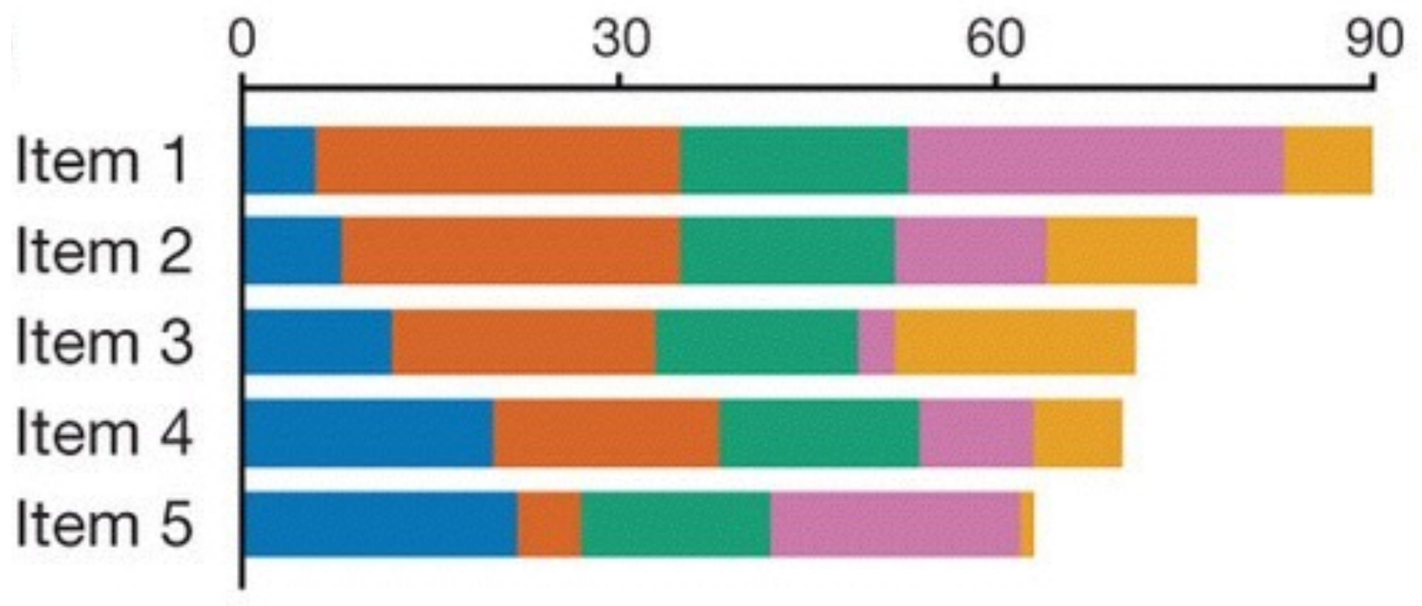
Stacked Bar Chart



Comparison of bar chart types

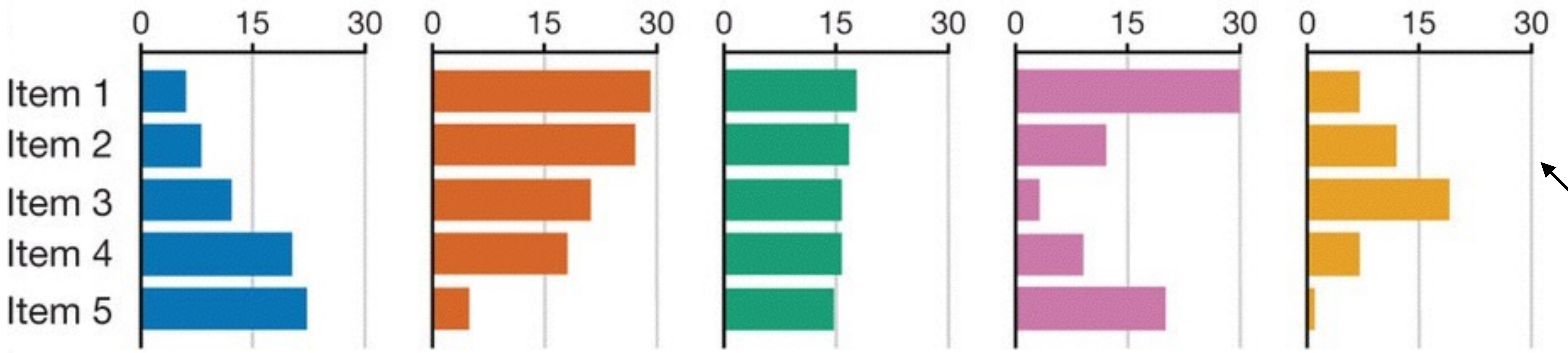


Pie Chart



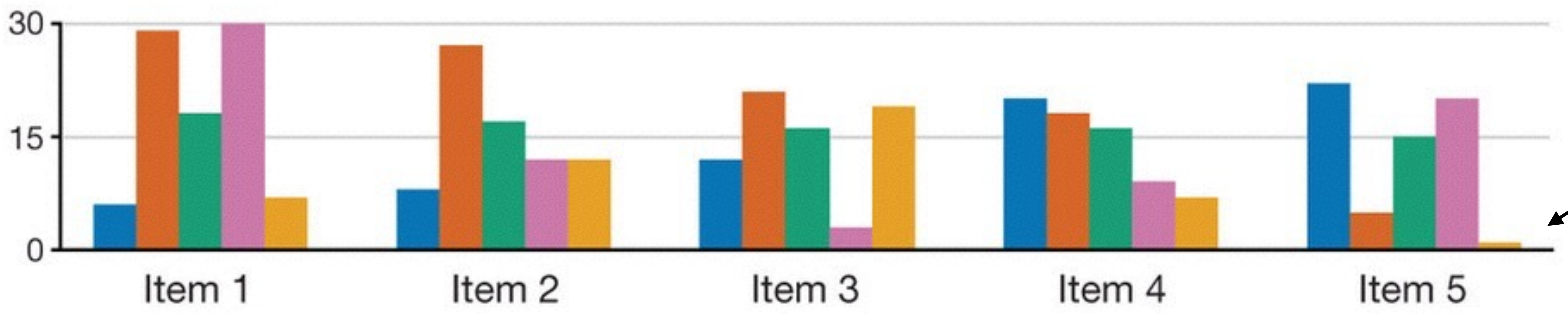
Stacked bar chart

Layered Bar Chart



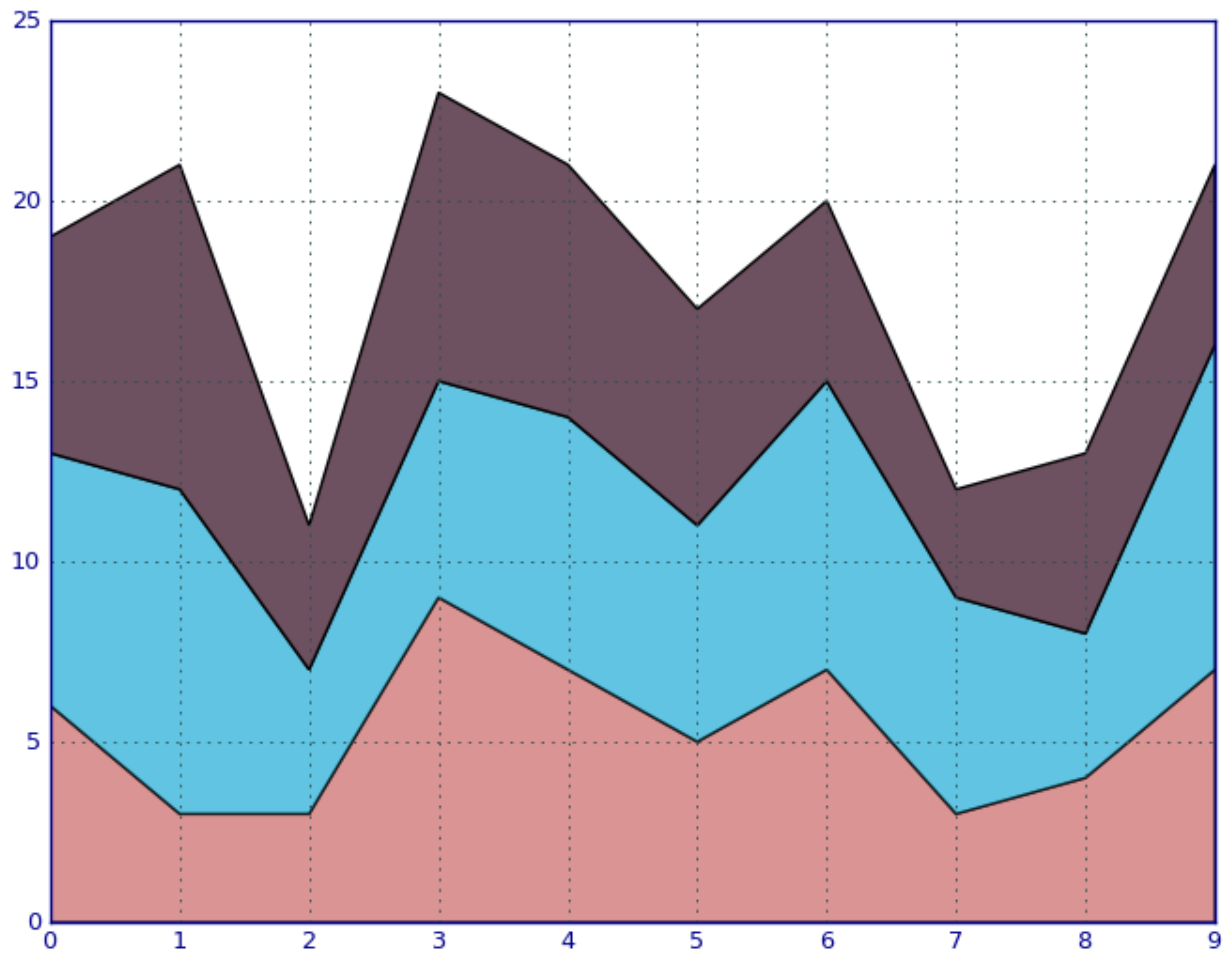
Small Multiples

Grouped Bar Chart



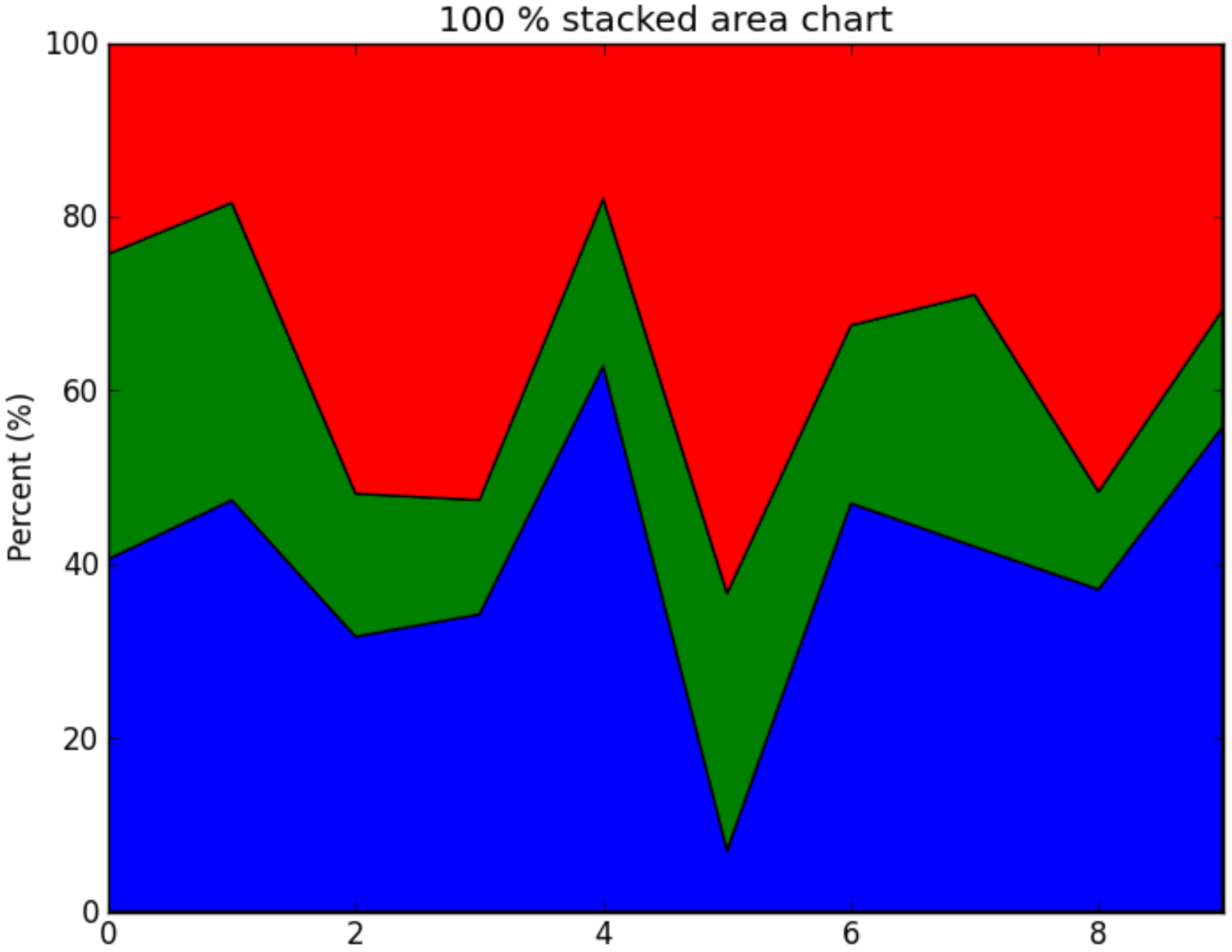
Streit & Gehlenborg, PoV, Nature Methods, 2014

Stacked Area Chart



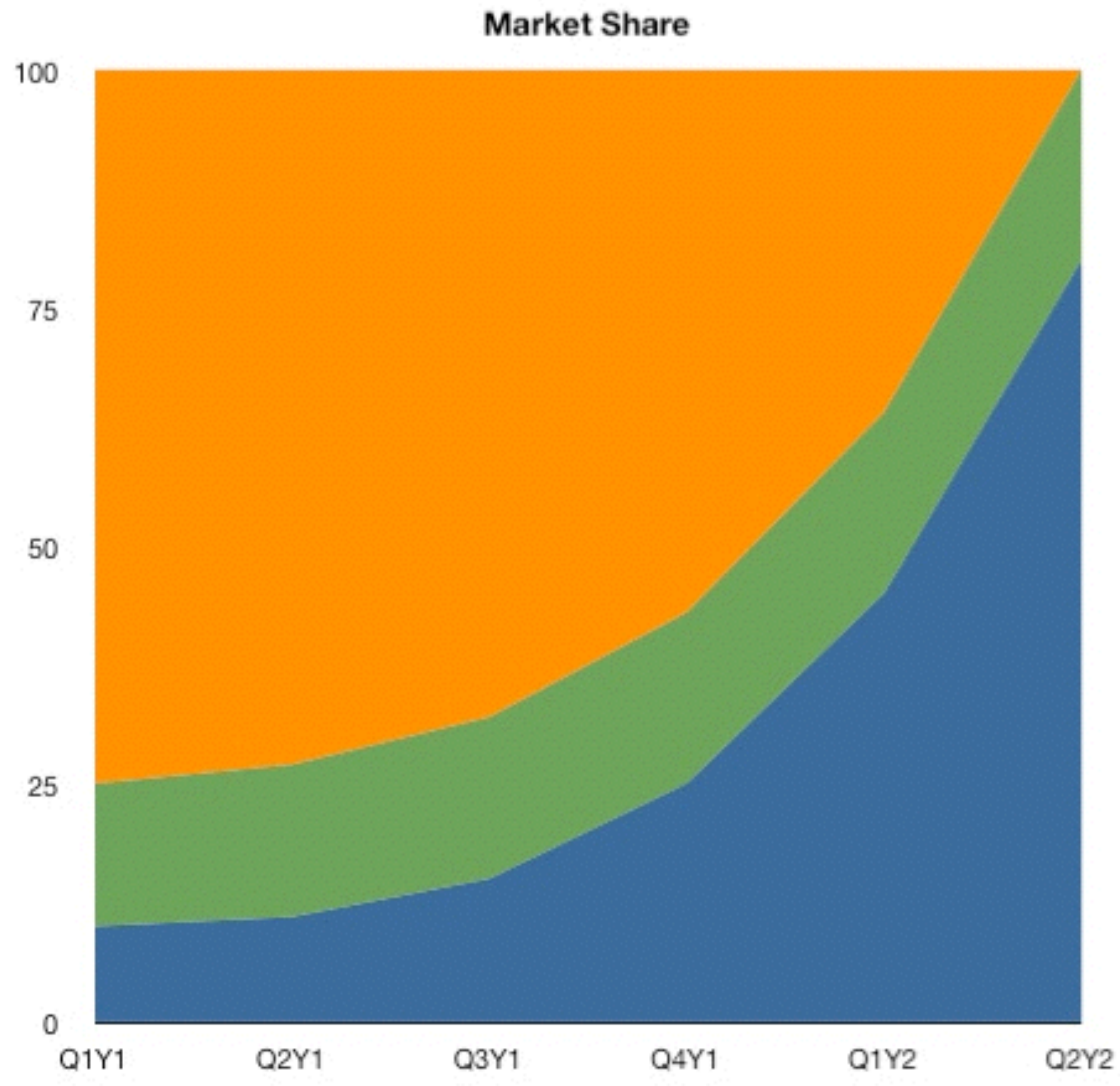
http://stackoverflow.com/questions/2225995/how-can-i-create-stacked-line-graph-with-matplotlib

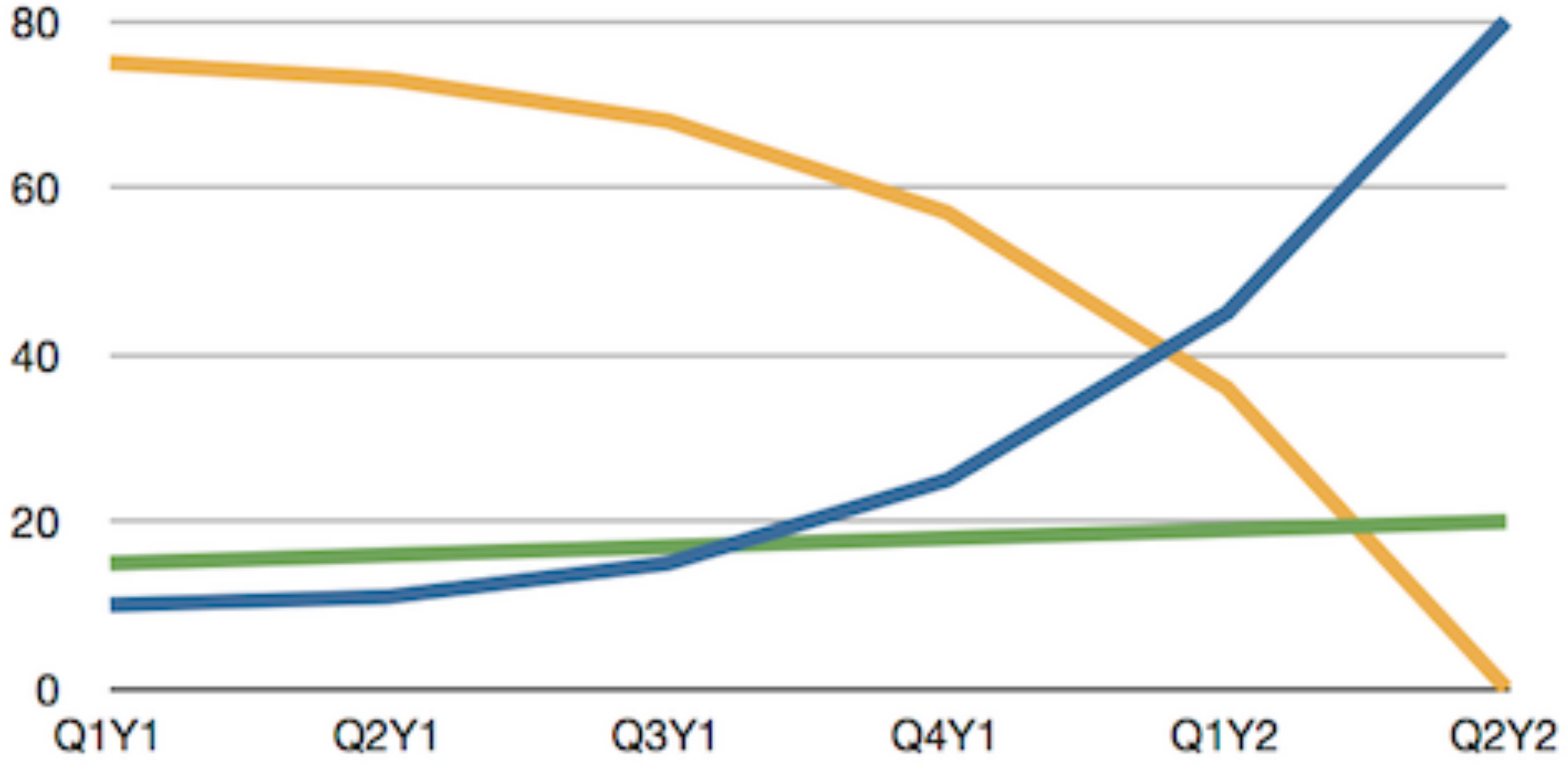
100% Stacked Area Chart



http://stackoverflow.com/questions/16875546/create-a-100-stacked-area-chart-with-matplotlib

Stacked Area vs. Line Graphs





leancrew.com &
Practically Efficient

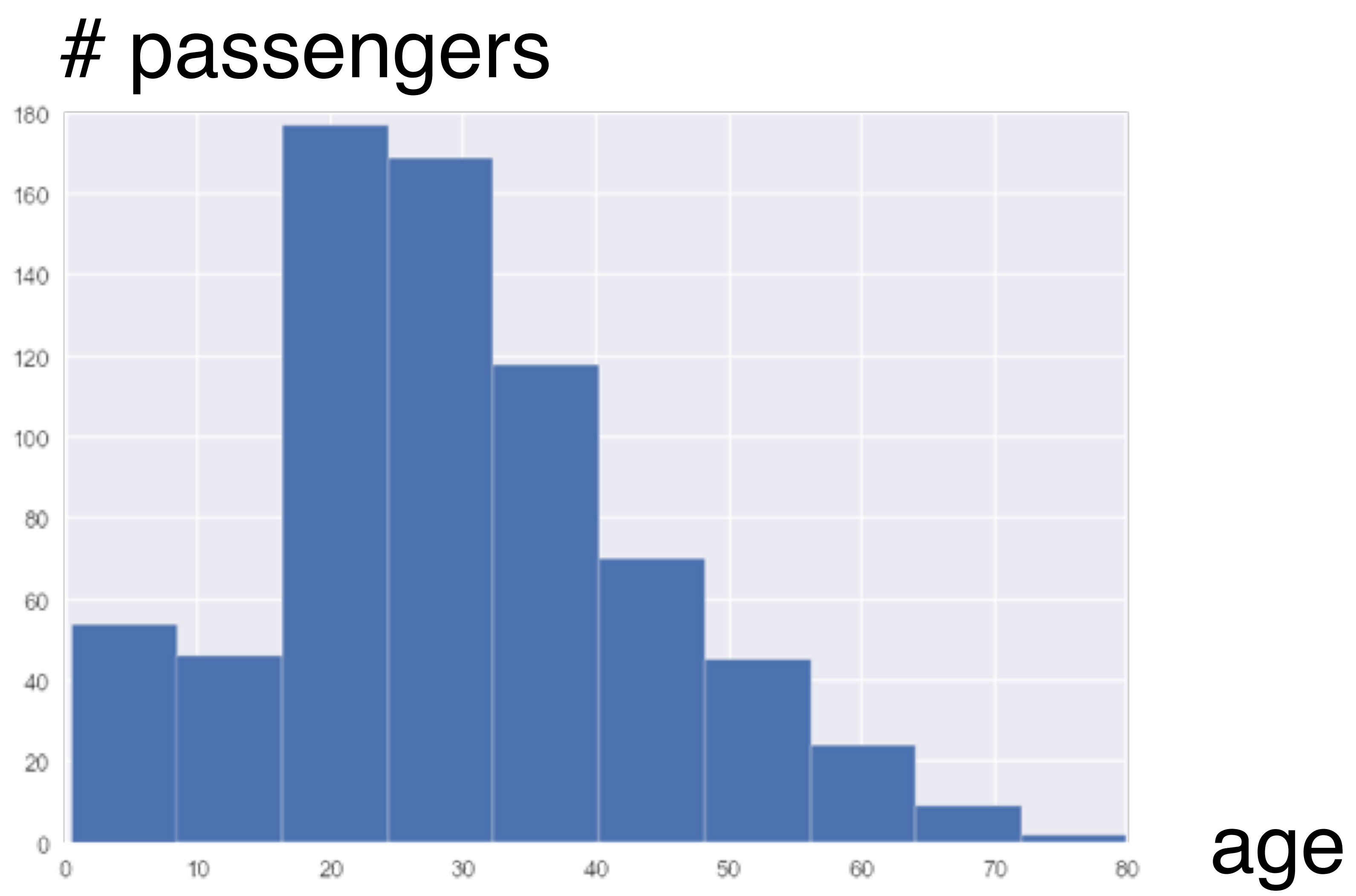
Distributions

Histogram

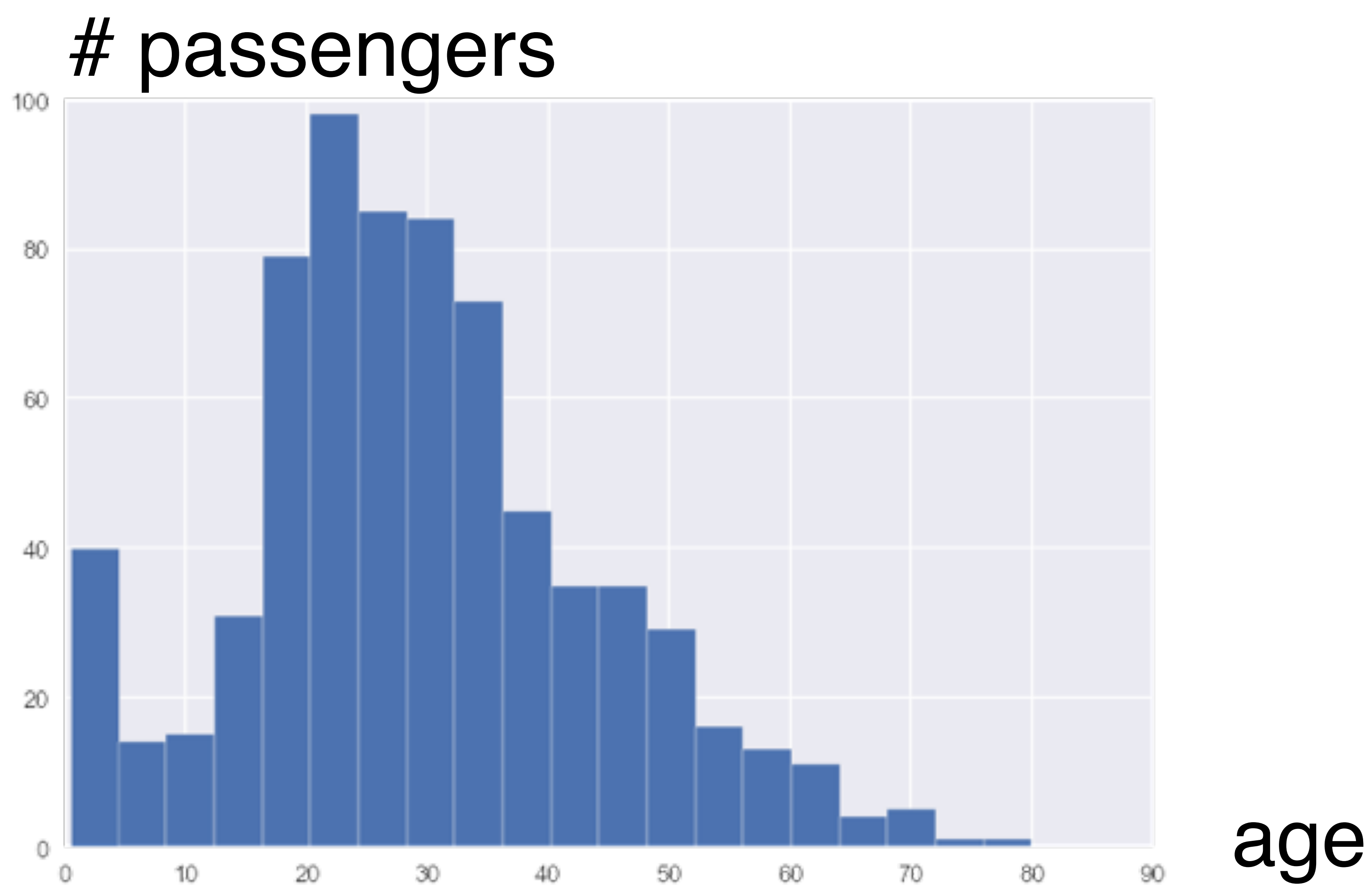
#bins hard to predict

make interactive!

rule of thumb: #bins = sqrt(n)



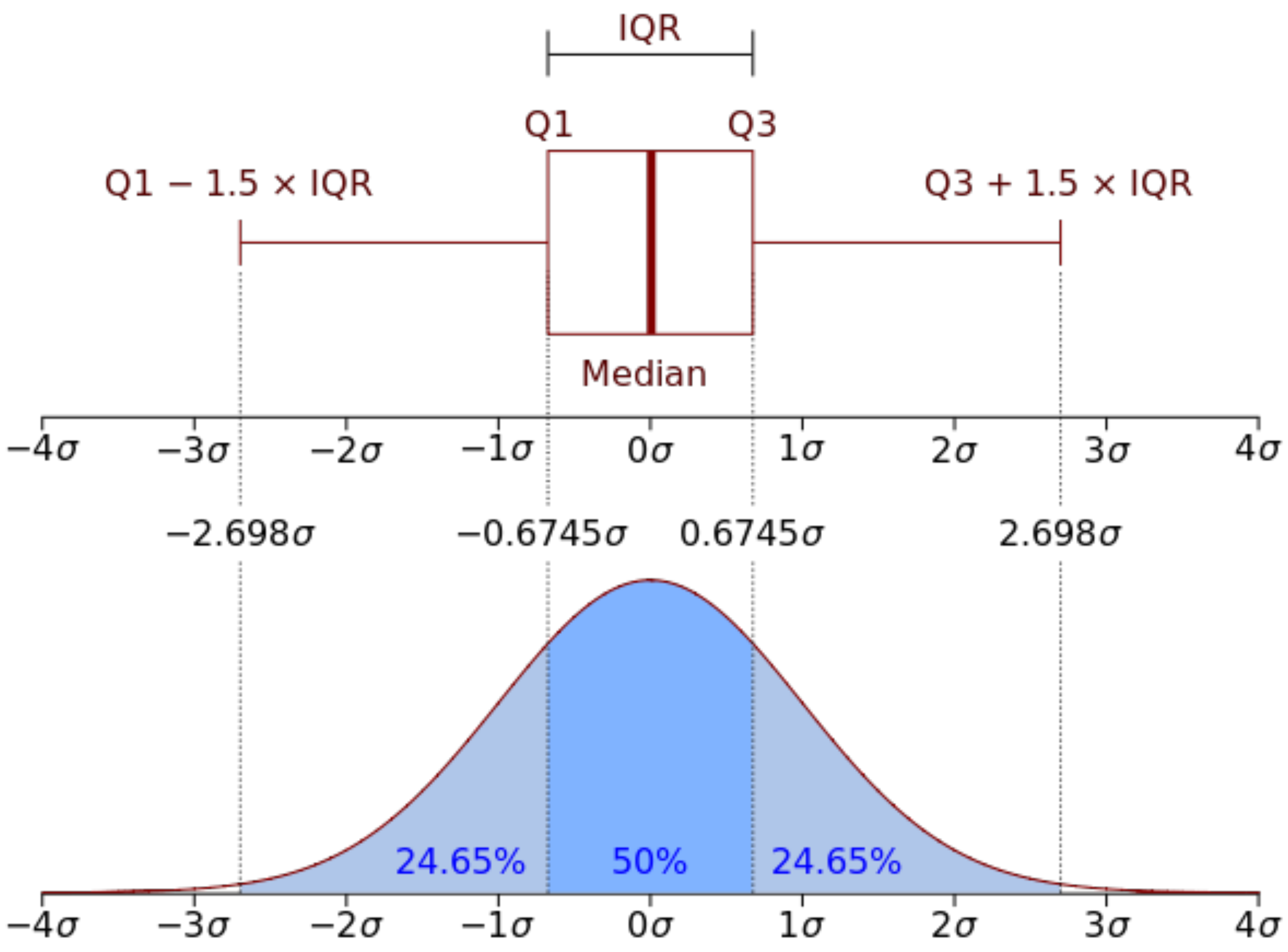
10 Bins



20 Bins

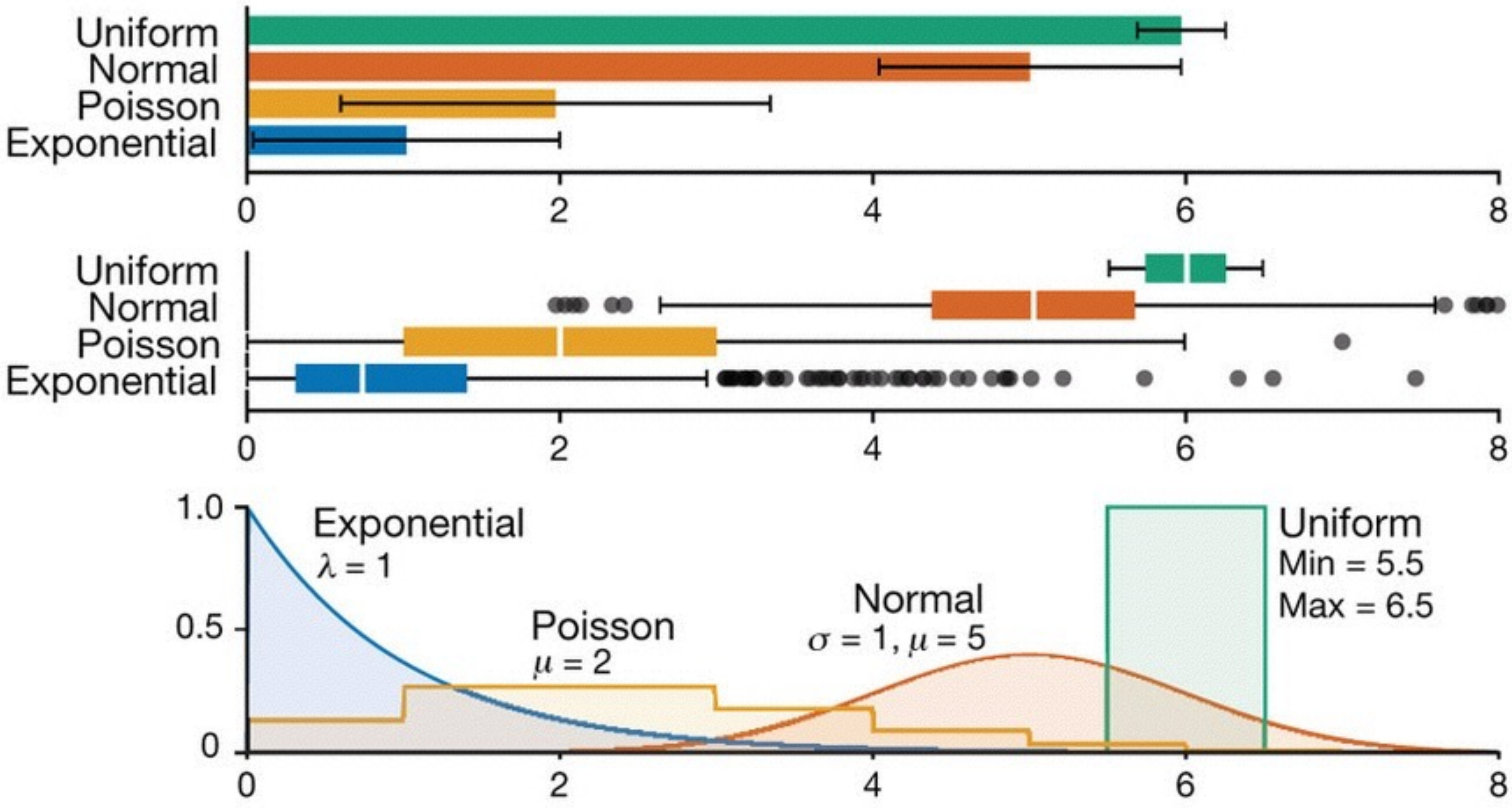
Box Plots

aka Box-and-Whisker Plot



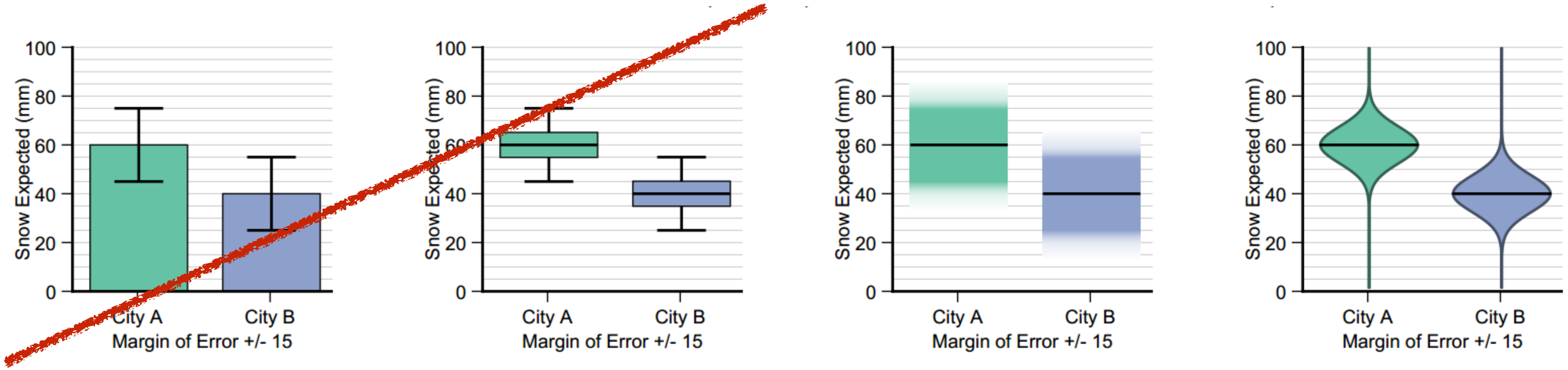
Wikipedia

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*	*M*	*181*
Alice	*22*	*F*	*185*
Chris	*19*	*M*	*175*

High-Dimensional Data Visualization

How many dimensions?

~50 – tractable with “just” vis

~1000 – need analytical methods

How many records?

~ 1000 – “just” vis is fine

>> 10,000 – need analytical methods

Homogeneity

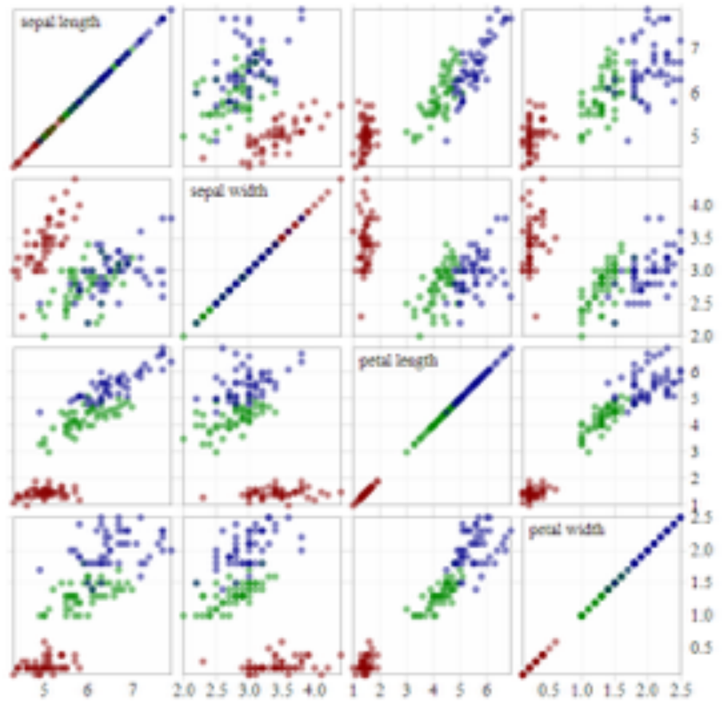
Same data type?

Same scales?

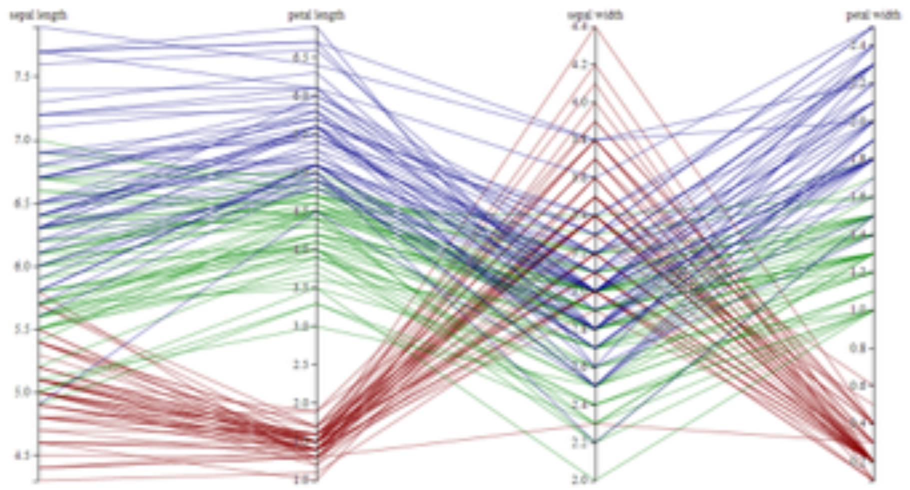
	Age	Gender	Height
Bob	*25*	*M*	*181*
Alice	*22*	*F*	*185*
Chris	*19*	*M*	*175*

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

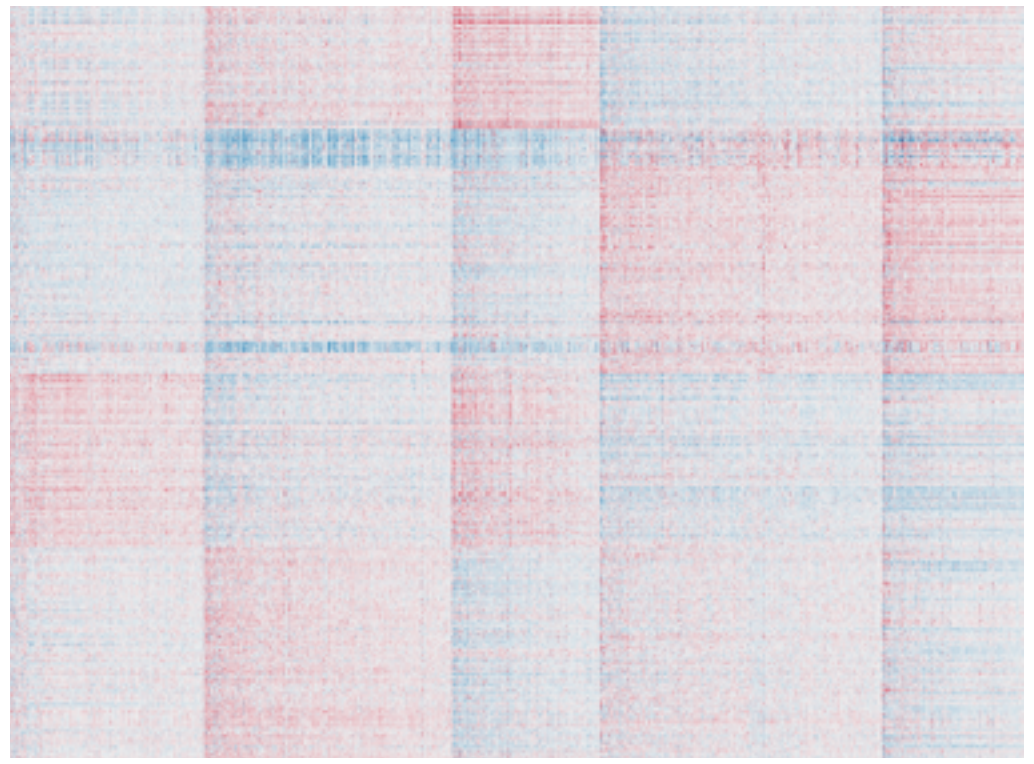
Analytic Component



Scatterplot Matrices
[Bostock]

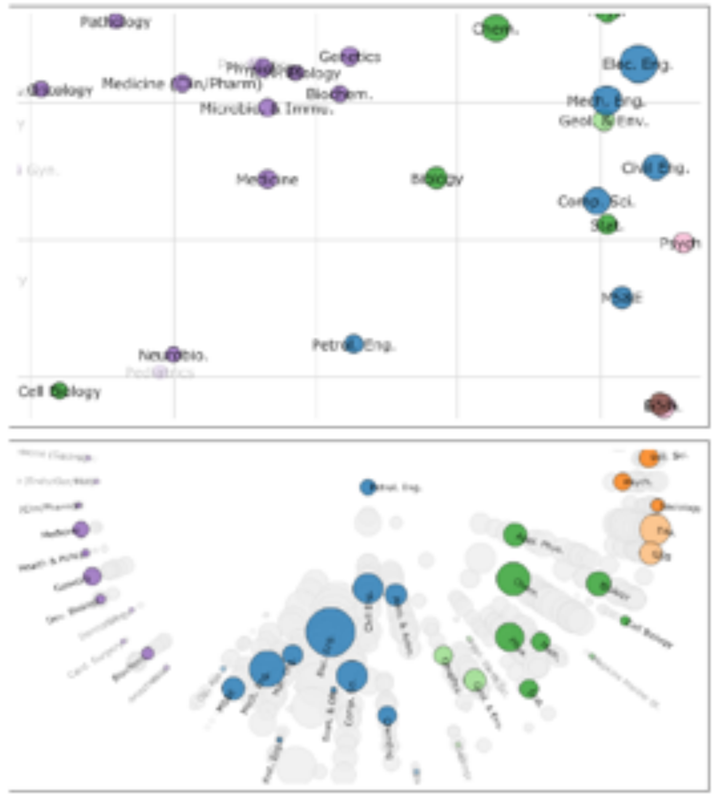


Parallel Coordinates
[Bostock]



Pixel-based visualizations / heat maps

Multidimensional Scaling
[Doerk 2011]



[Chuang 2012]

no / little analytics

strong analytics component

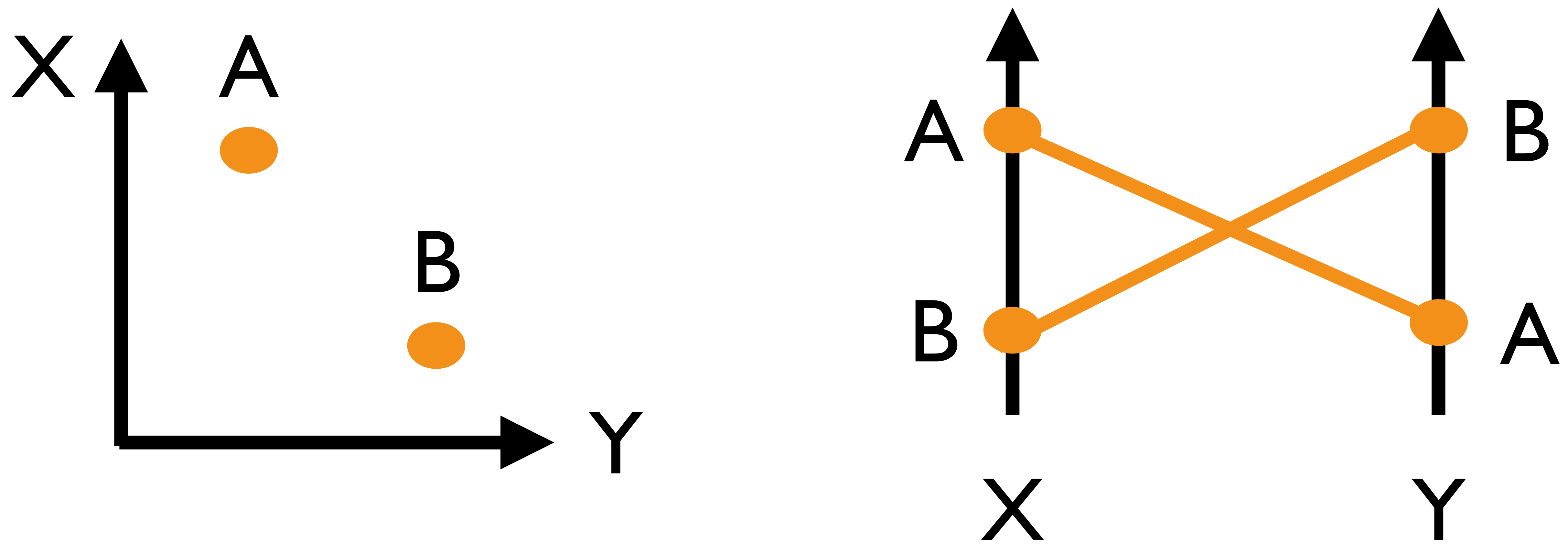
Geometric Methods

Parallel Coordinates (PC)

Inselberg 1985

Axes represent attributes

Lines connecting axes represent items



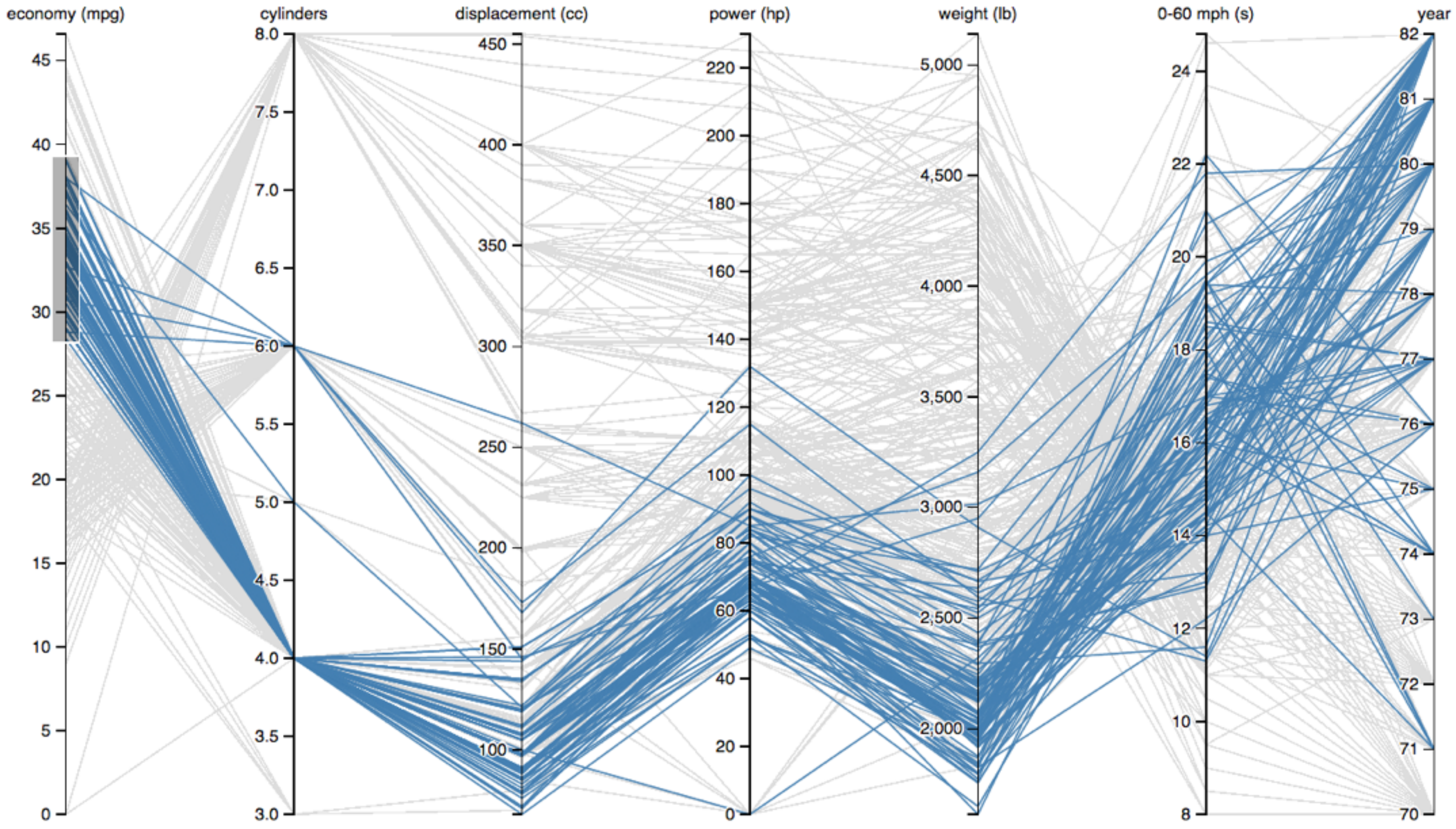
Parallel Coordinates

Each axis represents dimension

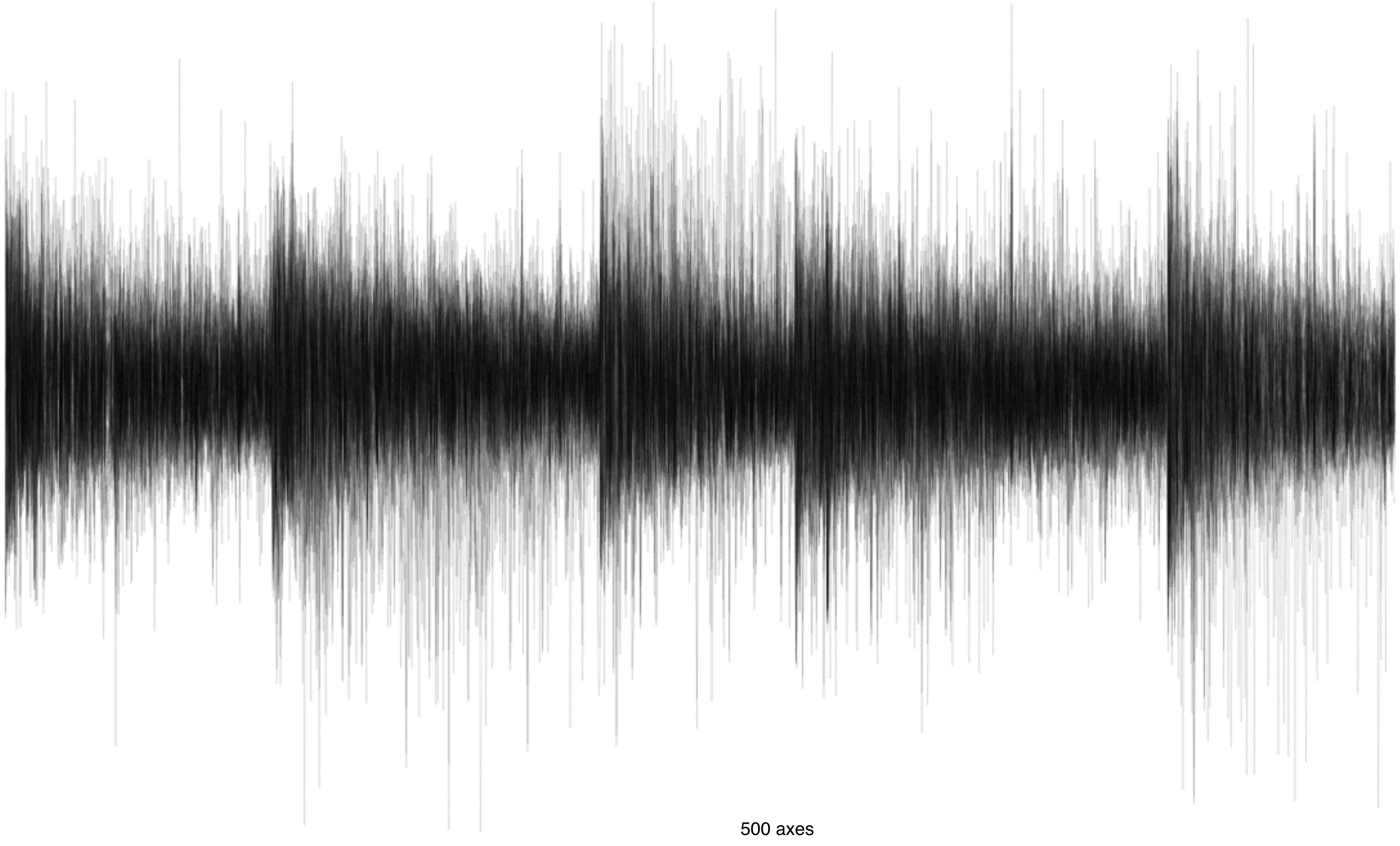
Lines connecting axis represent records

Suitable for

- all tabular data types
- heterogeneous data



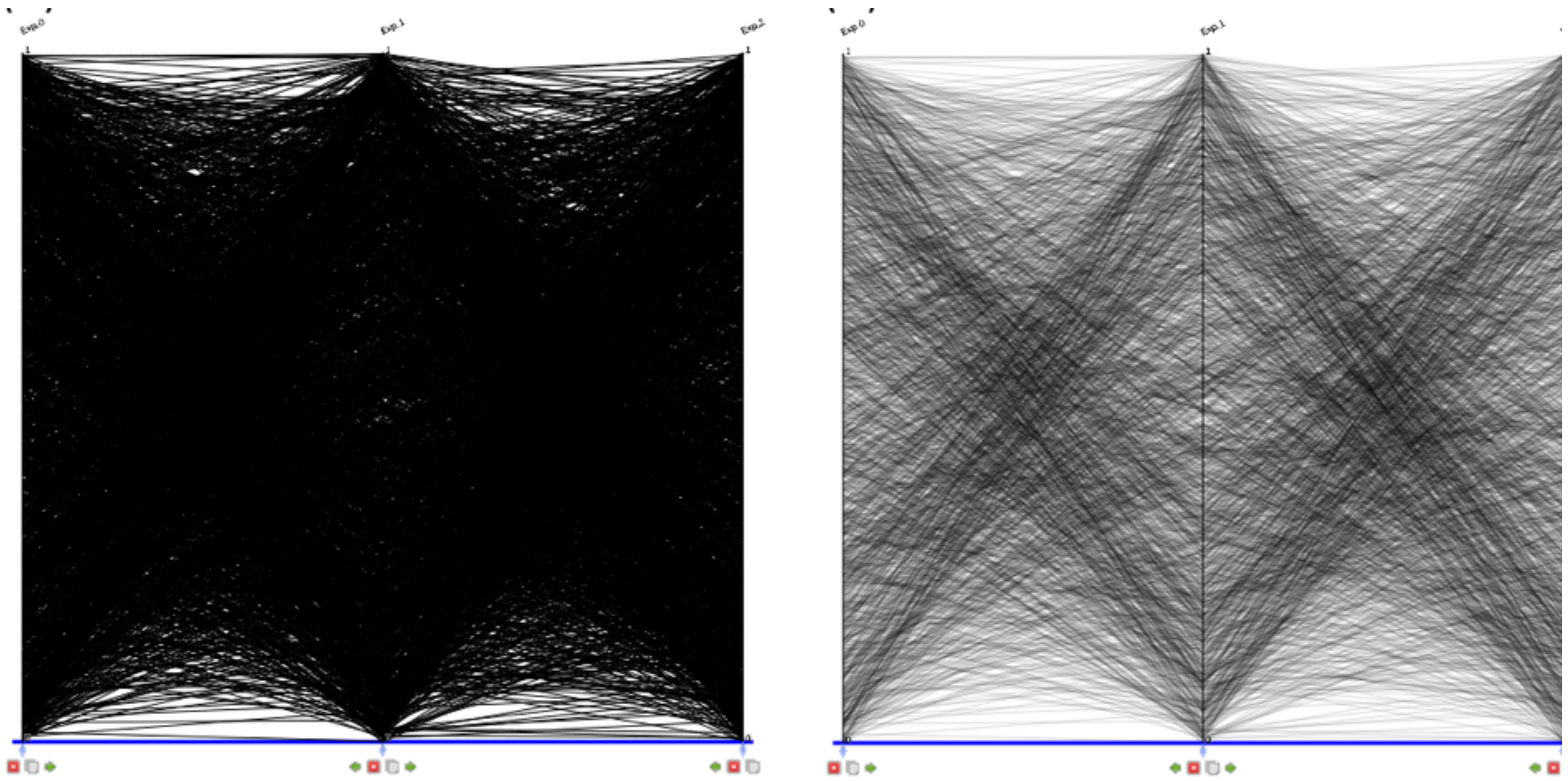
PC Limitation: Scalability to Many Dimensions



PC Limitation: Scalability to Many Items

Solutions:

- Transparency
- Bundling, Clustering
- Sampling

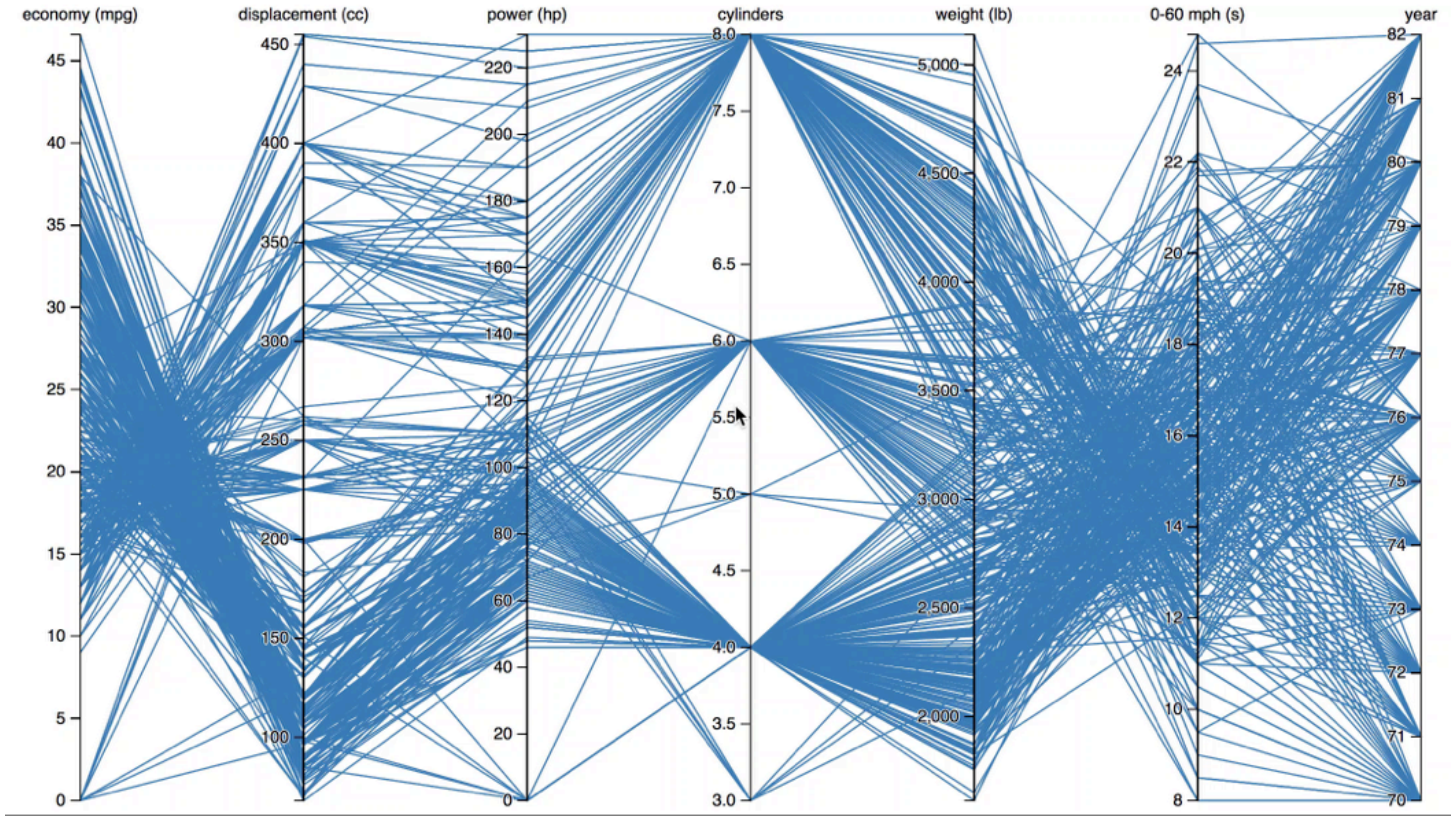


PC Limitations

Correlations only between adjacent axes

Solution: Interaction

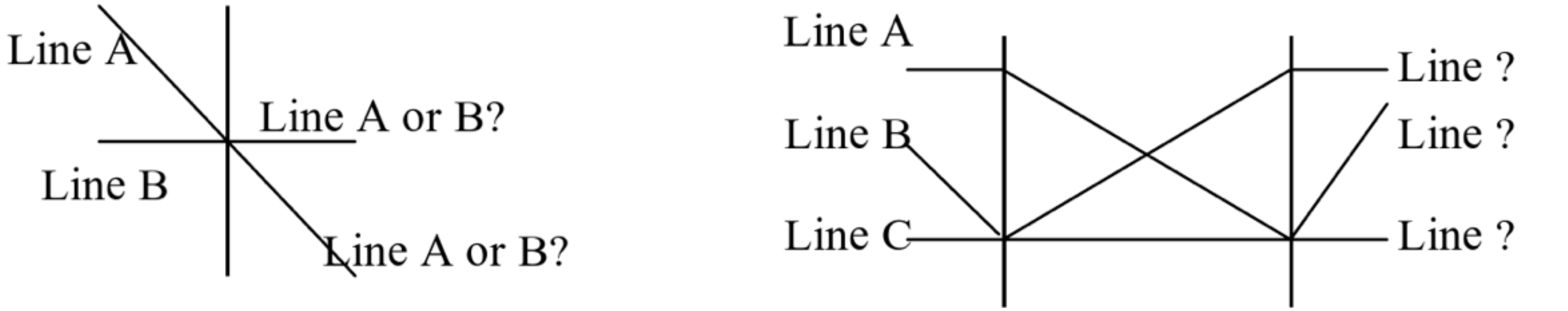
- Brushing
- Let user change order

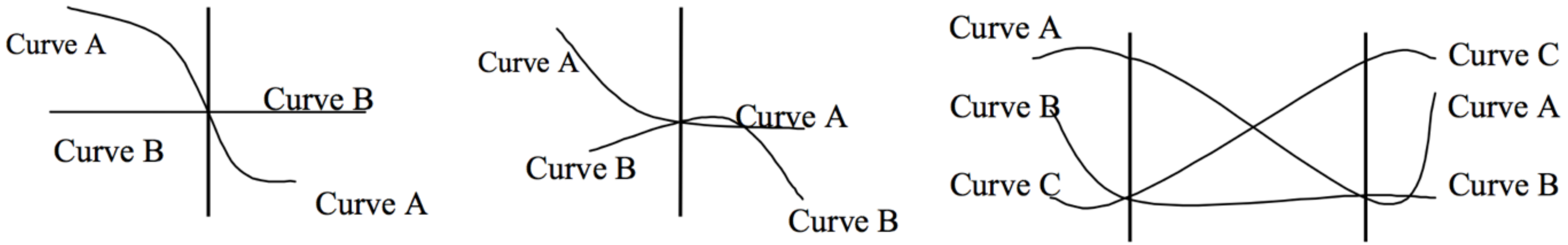


PC Limitation: Ambiguity

Solutions:

- Brushing
- Curves





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

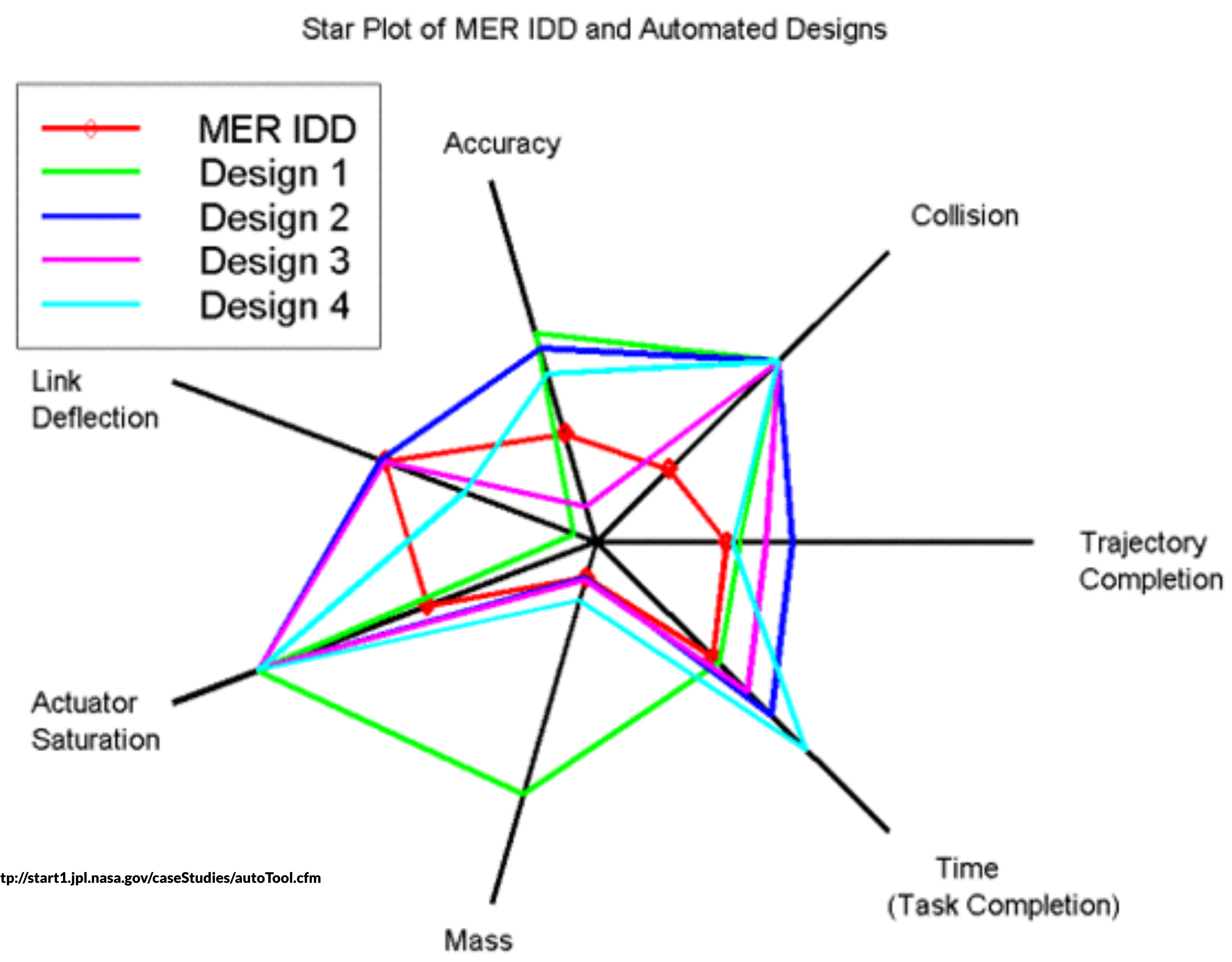
http://bl.ocks.org/jasondavies/1341281

Star Plot

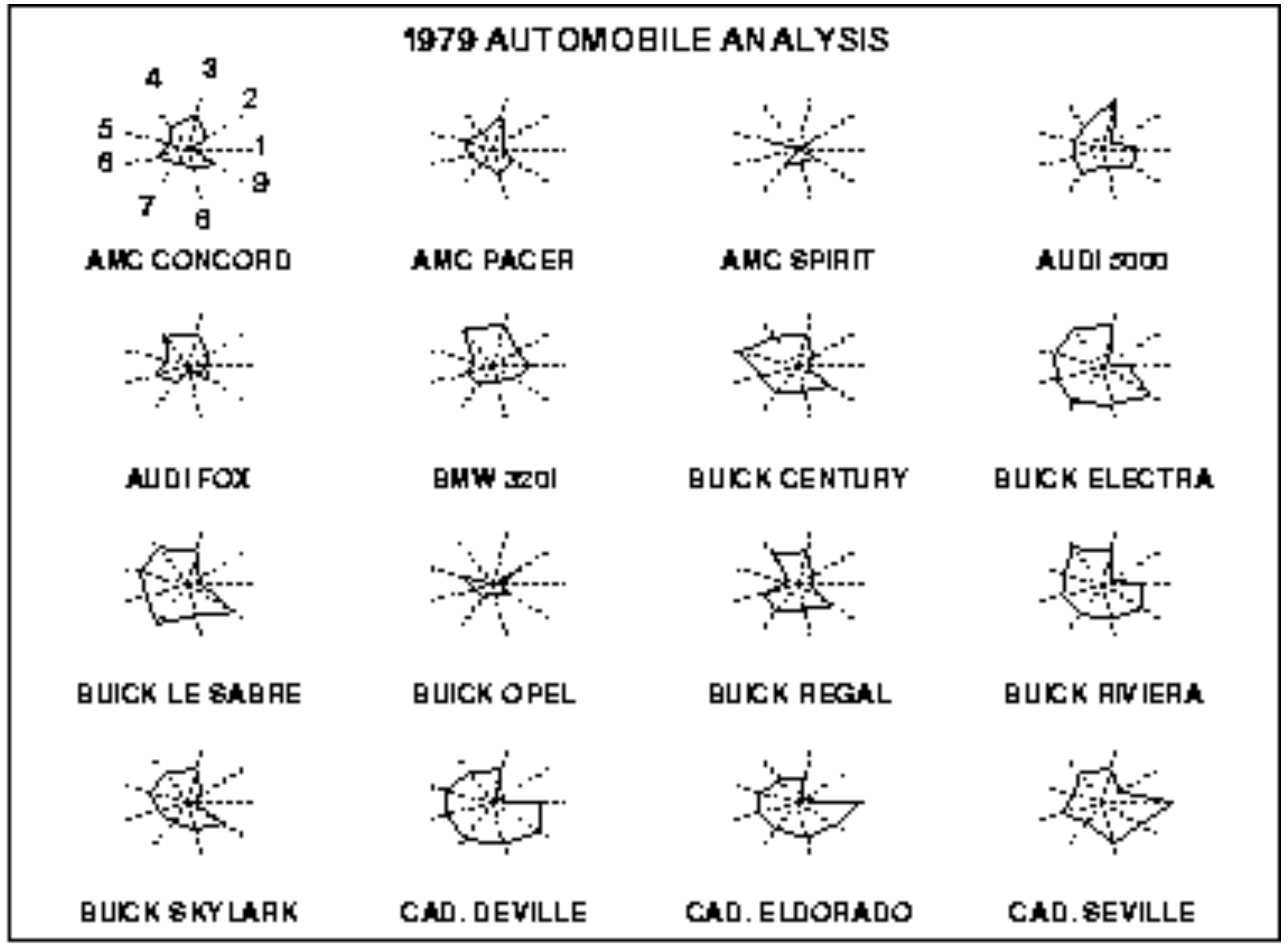
[Coekin1969]

Similar to parallel coordinates

Radiate from a common origin



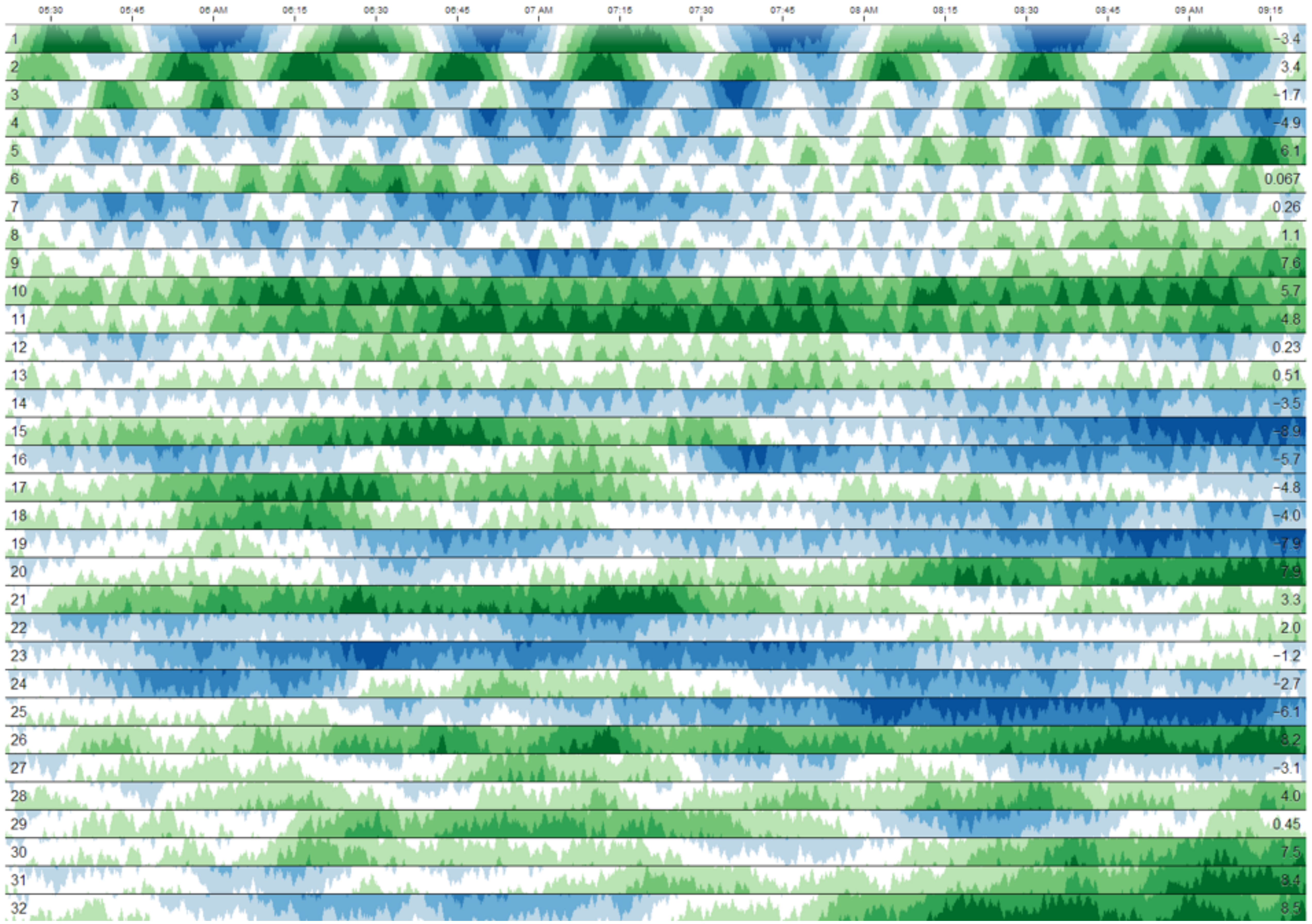
http://start1.jpl.nasa.gov/caseStudies/autoTool.cfm



http://www.itl.nist.gov/div898/handbook/eda/section3/starplot.htm

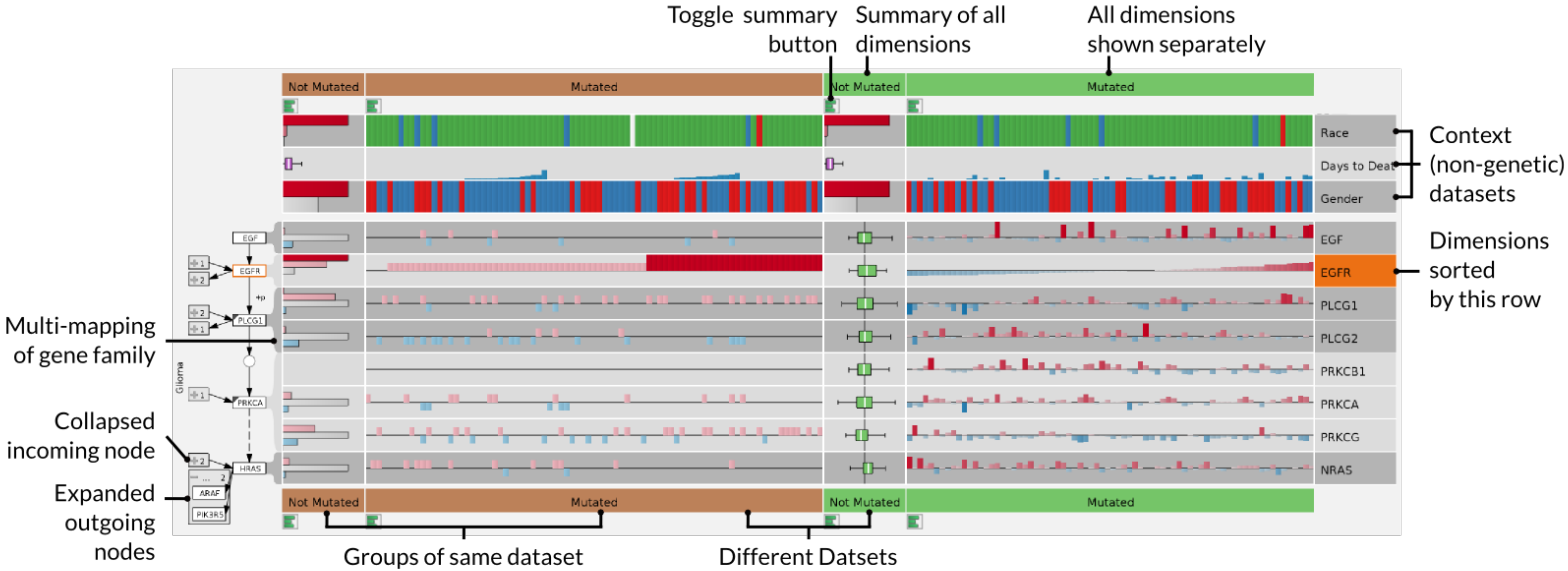
http://bl.ocks.org/kevinschaul/raw/8833989/

Multiple Line Charts



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

Combining Various Charts

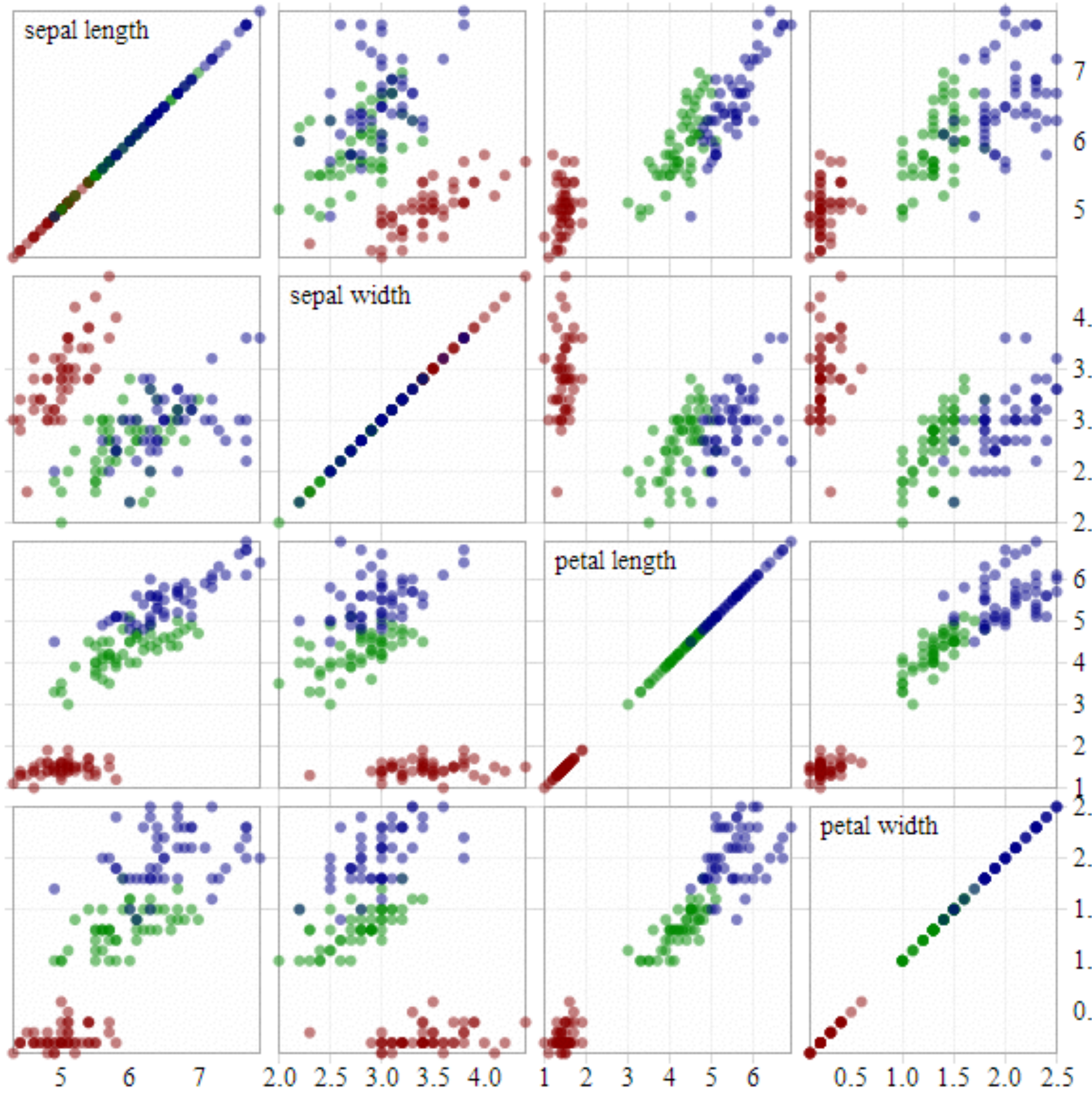


Scatterplot Matrices (SPLOM)

Matrix of size d*d

Each row/column is one dimension

Each cell plots a scatterplot of two dimensions



Scatterplot Matrices

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

Brushing is important

Often combined with “Focus Scatterplot” as F+C technique

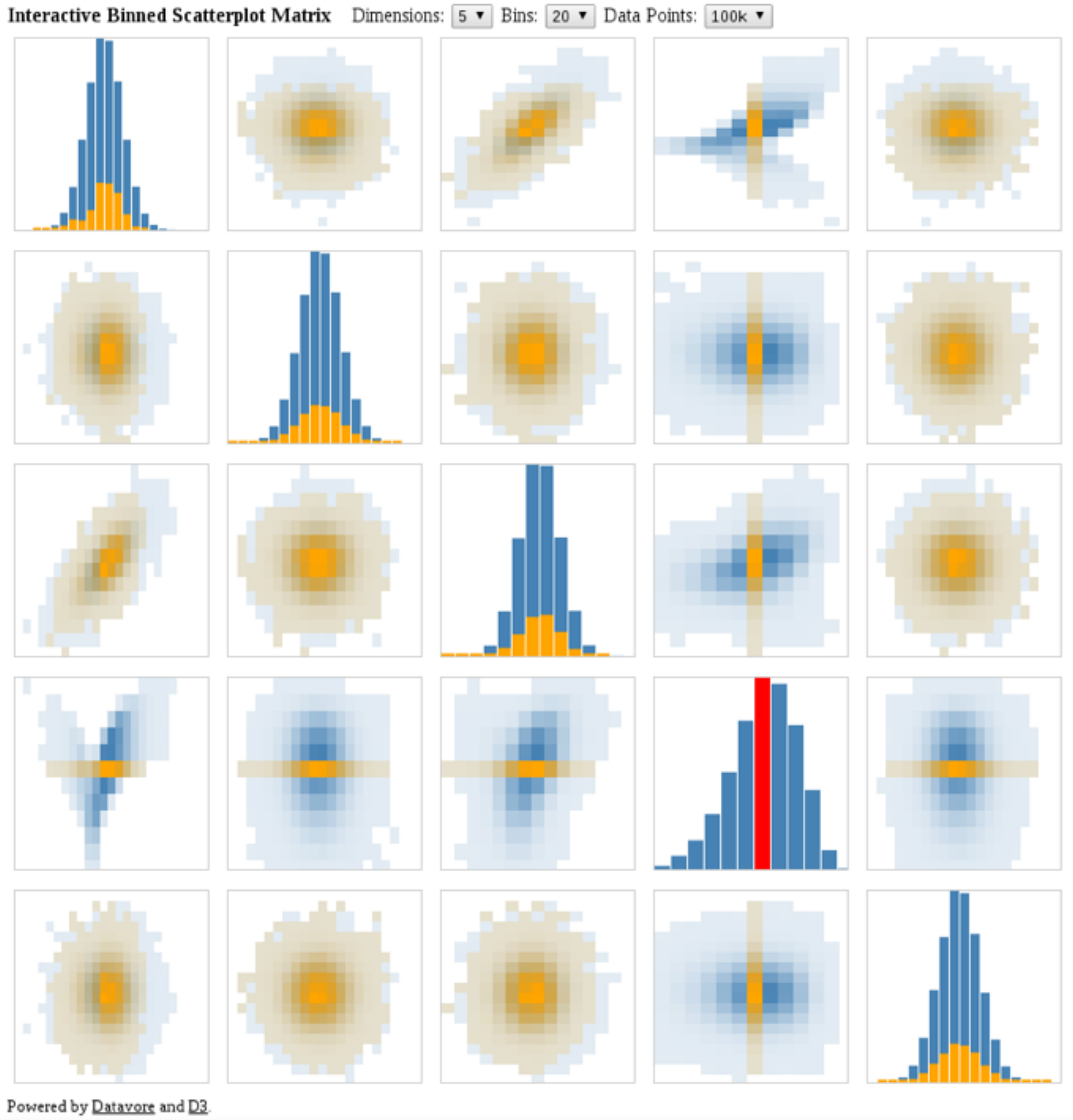
Algorithmic approaches:

Clustering & aggregating records

Choosing dimensions

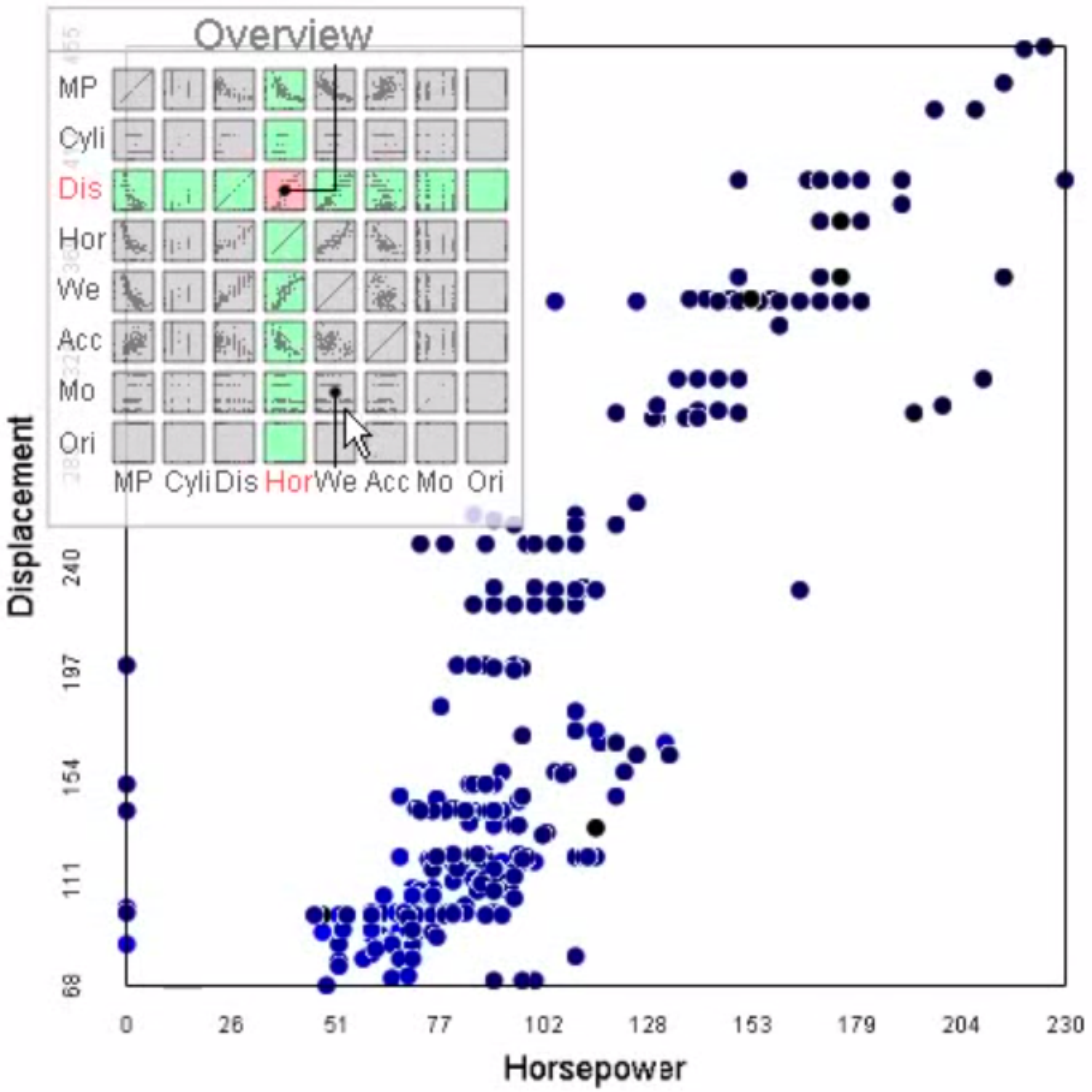
Choosing order

SPLOM Aggregation - Heat Map



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

SPLOM F+C, Navigation

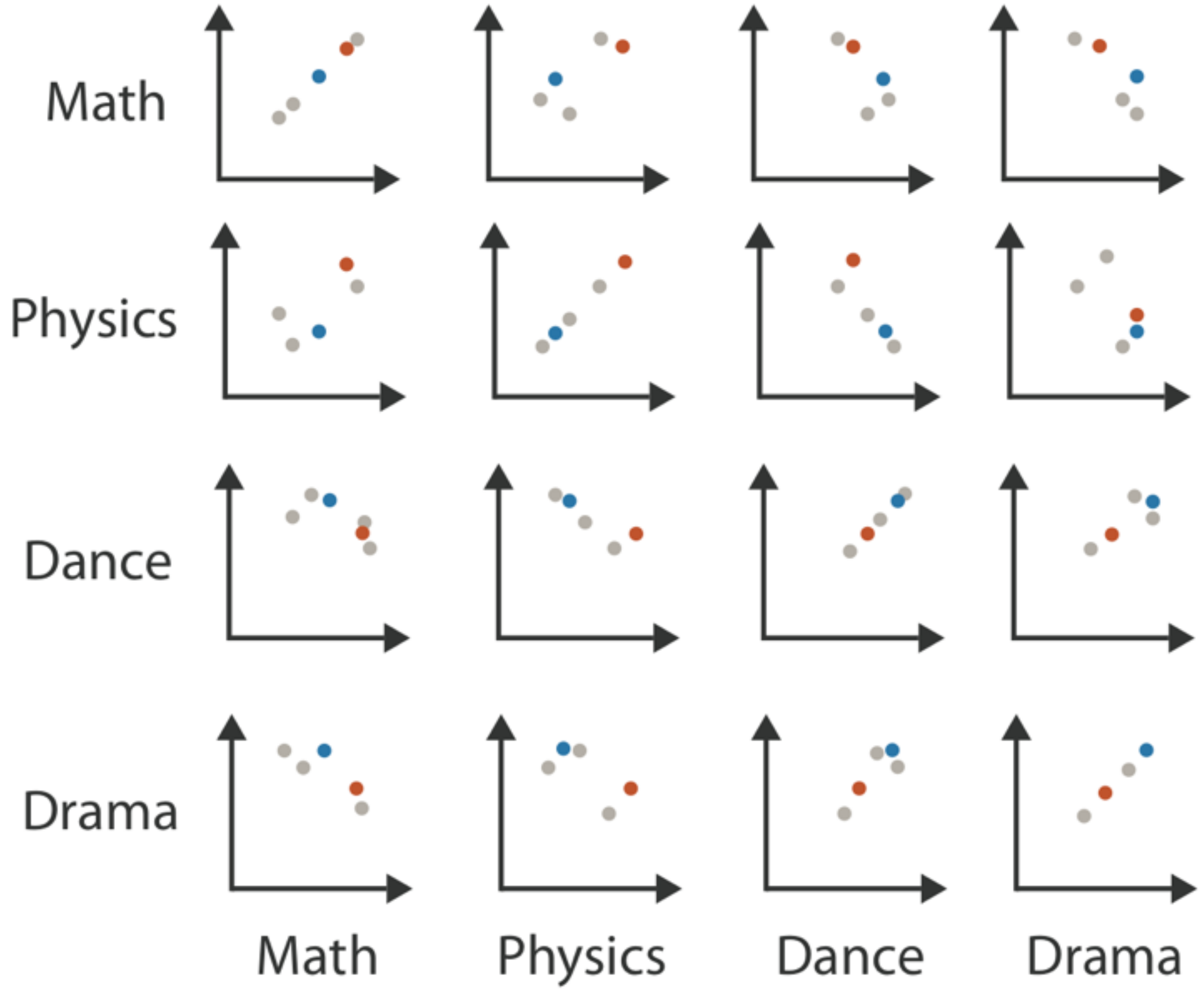


[Elmqvist]

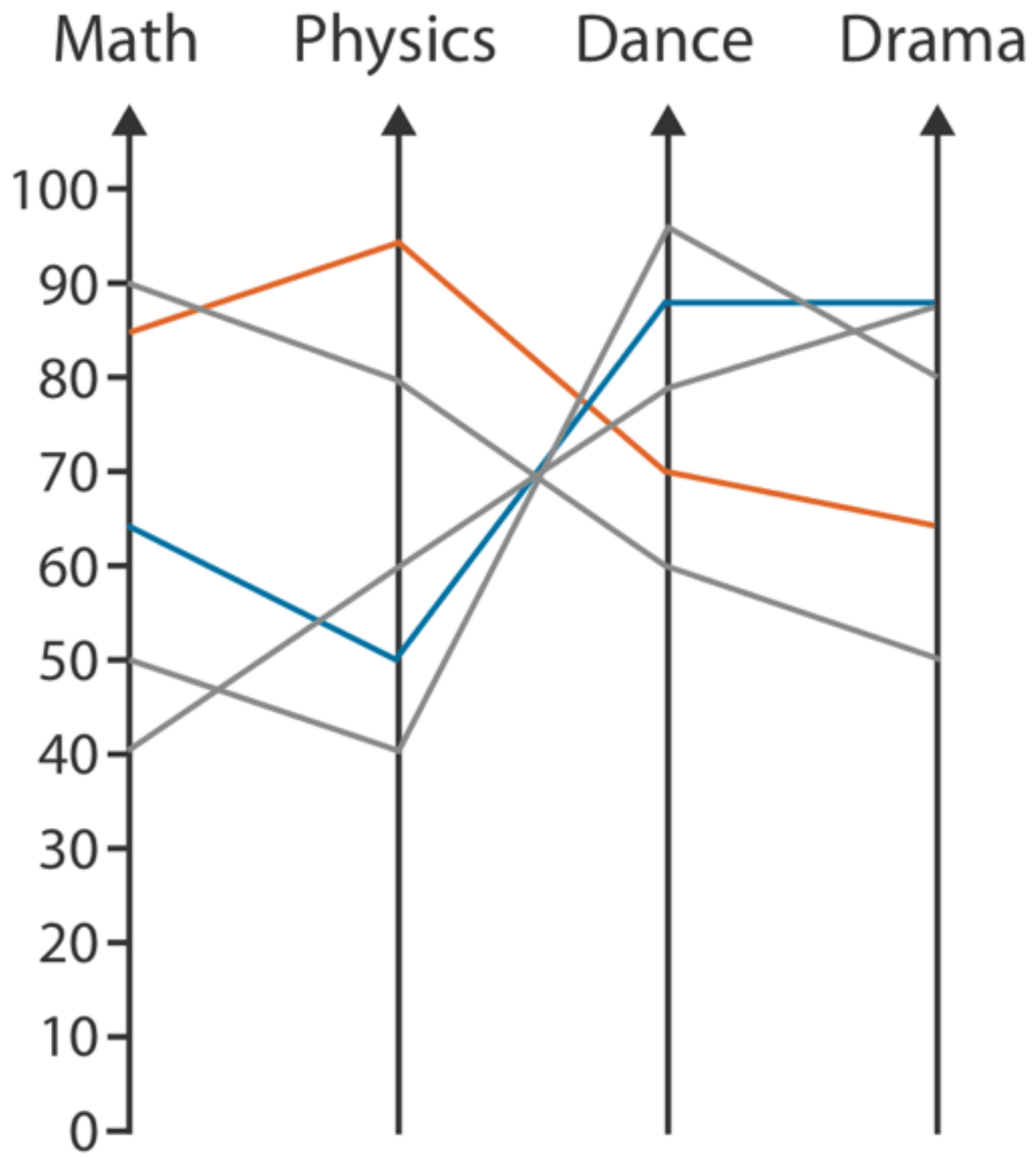
Table

Math	Physics	Dance	Drama
85	95	70	65
90	80	60	50
65	50	90	90
50	40	95	80
40	60	80	90

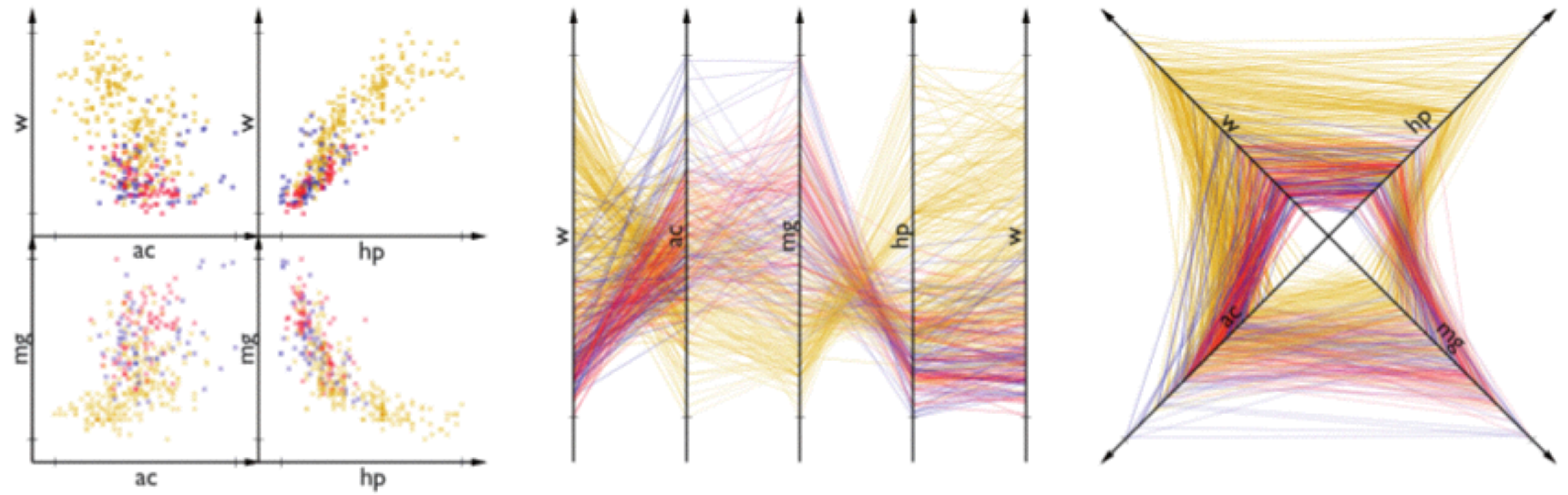
Scatterplot Matrix



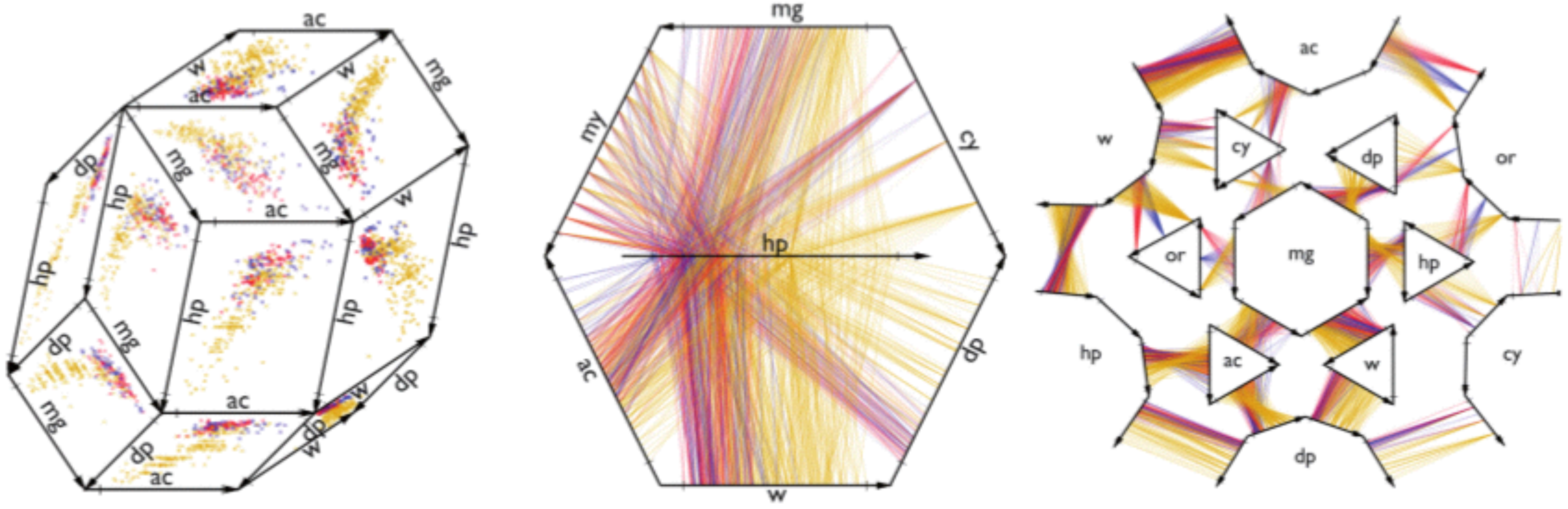
Parallel Coordinates



Flexible Linked Axes (FLINA)



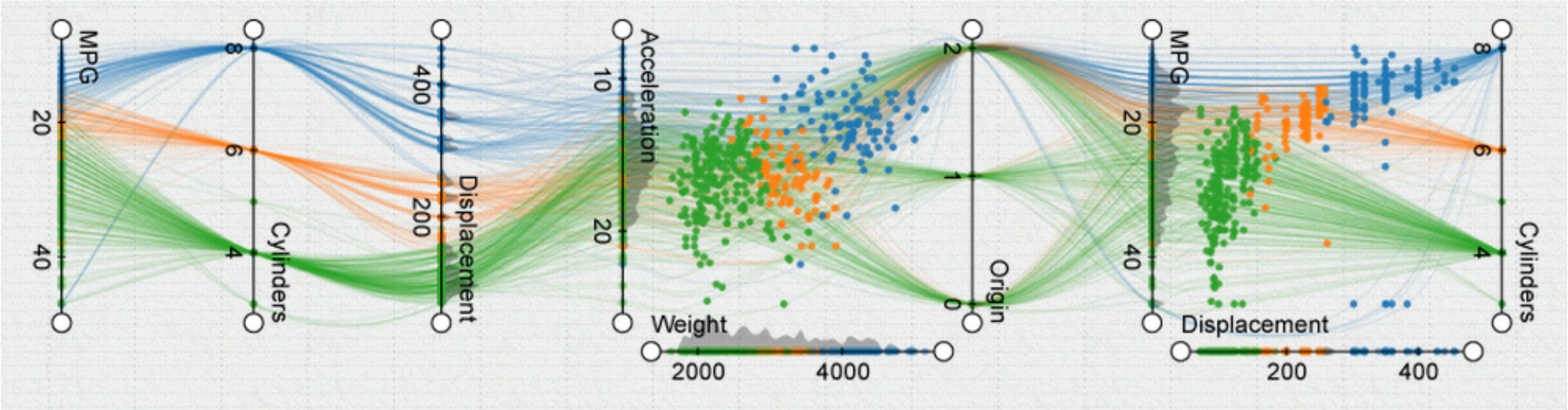
(a) scatterplots (b) Parallel Coordinates Plot (c) radar chart



(d) Hyperbox (e) Time Wheel (f) Many-to-many PCP

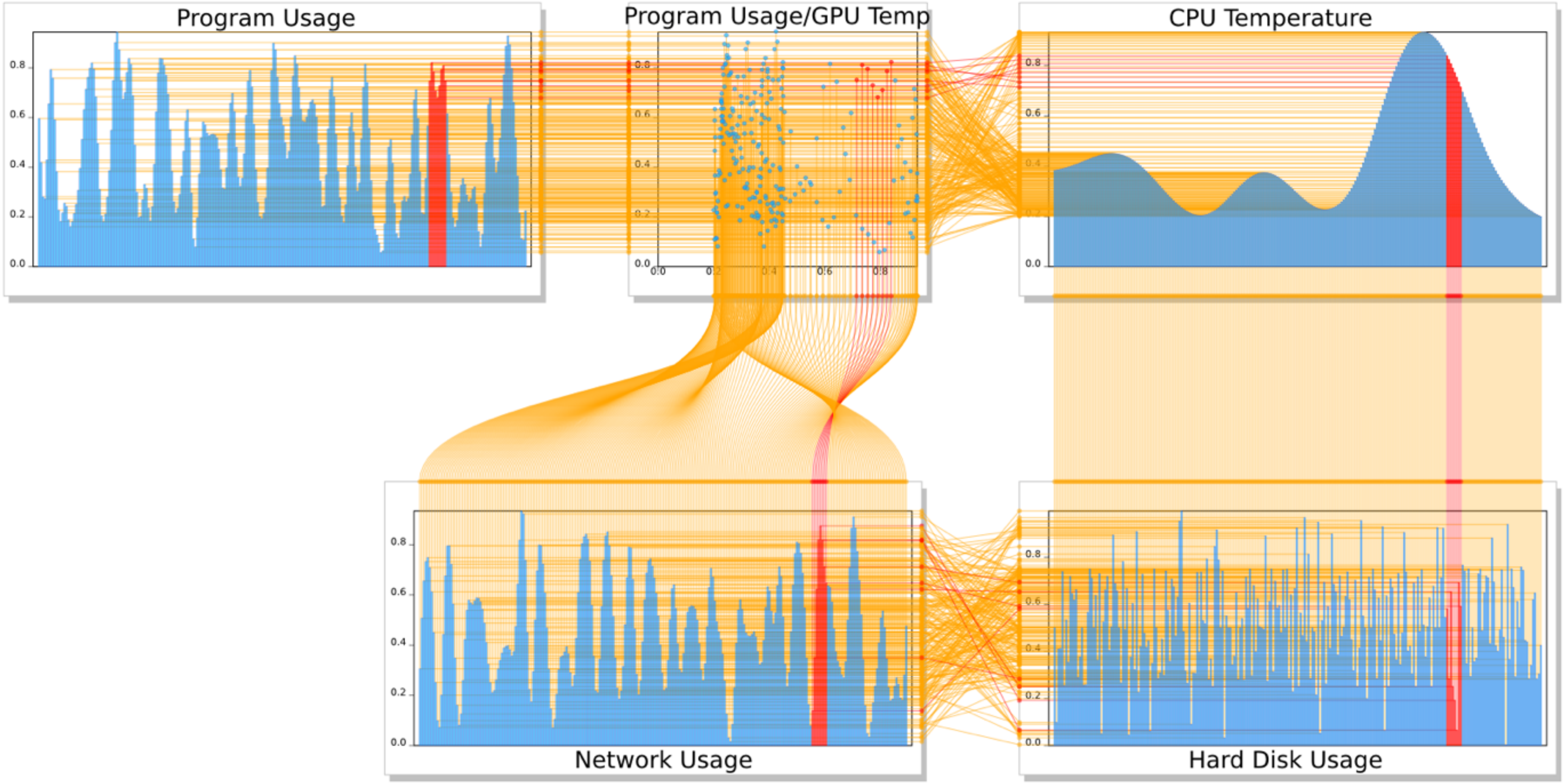
Claessen & van Wijk 2011

Web-based implementation of FLINA concept



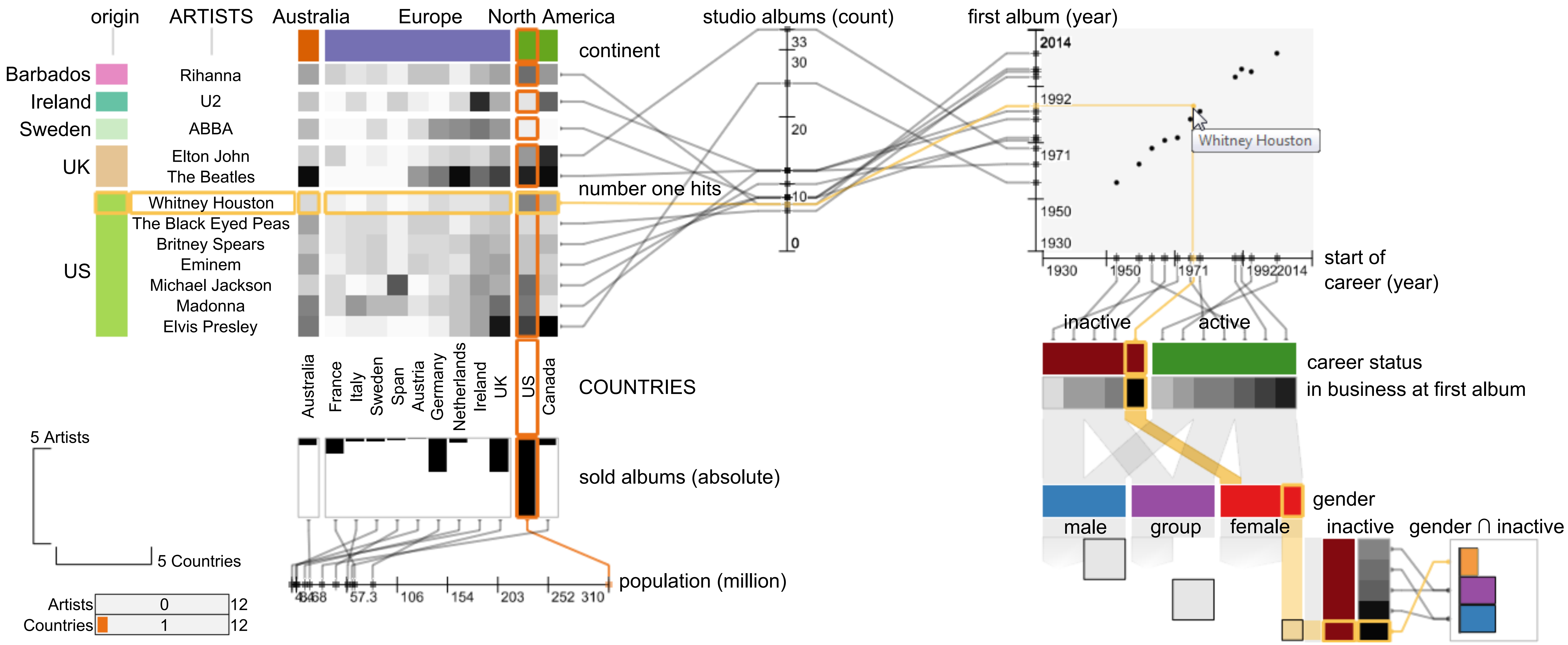
http://vis.pku.edu.cn/mddv/val/

Connected Charts



Viau & McGuffin 2012

Domino



Gratzl et al. 2014

Data Reduction

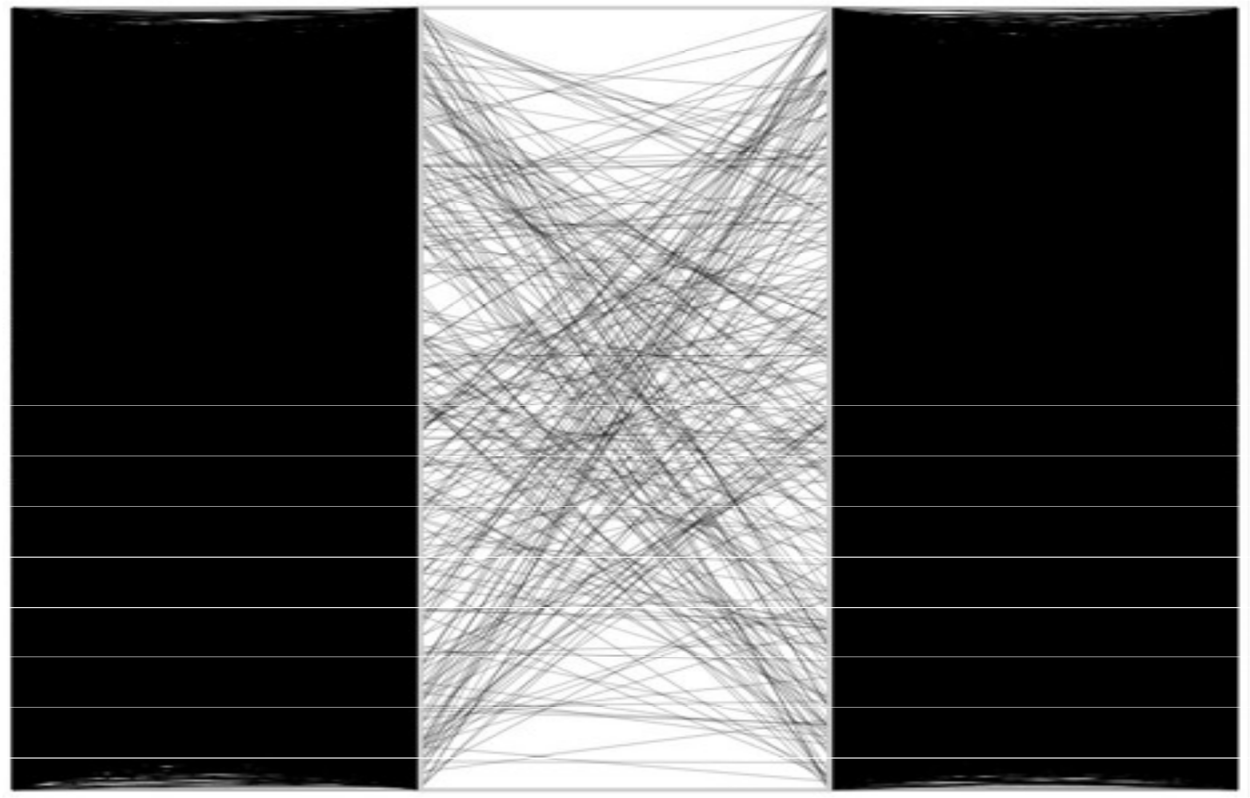
Sampling

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

Efficient for large dataset

Apply only for display purposes

Outlier-preserving approaches



[Ellis & Dix, 2006]

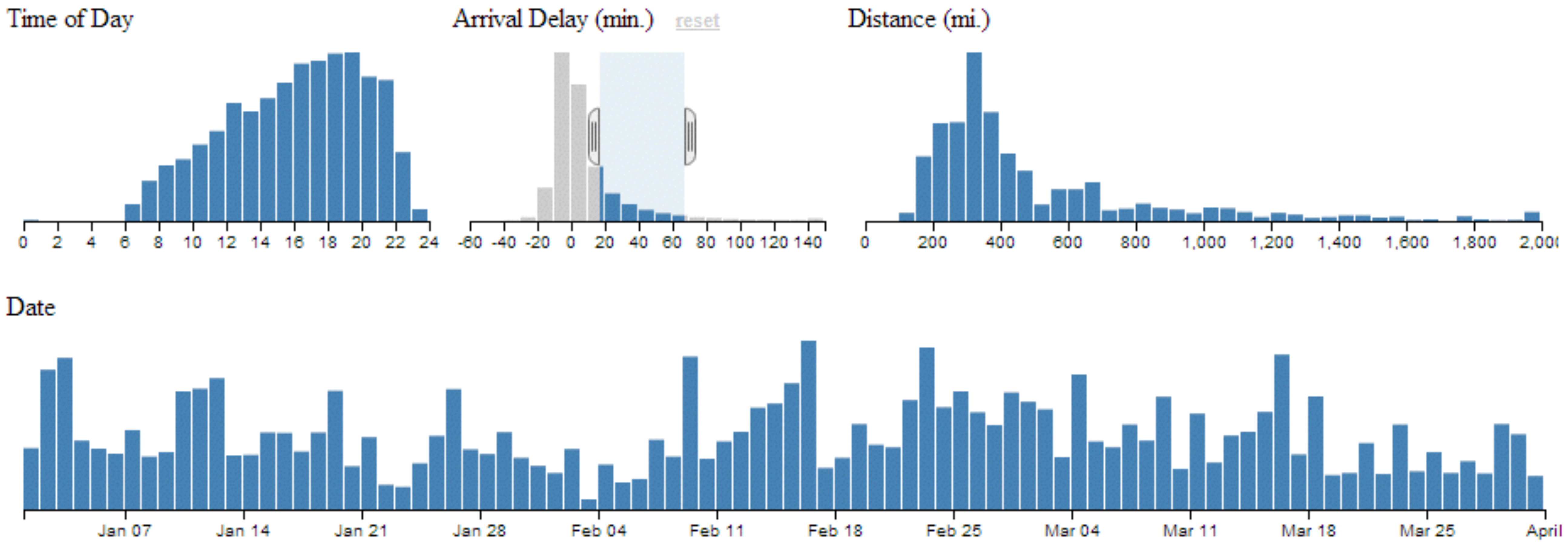
Filtering

Define criteria to remove data, e.g.,

- minimum variability
- > / < / = specific value for one dimension
- consistency in replicates, …

Can be interactive, combined with sampling

Filter Example



March 31, 2001

37,336 of 231,083 flights selected.

10:57 PM	MSY	HOU	303 mi.	+29 min.
09:33 PM	BWI	PVD	328 mi.	+64 min.
09:30 PM	TPA	PBI	174 mi.	+30 min.
09:29 PM	BWI	PVD	328 mi.	+20 min.
09:10 PM	LAS	PHX	256 mi.	+53 min.
09:02 PM	MSY	HOU	303 mi.	+26 min.

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

Pixel Based Methods

Pixel Based Displays

Each cell is a “pixel”, value encoded in color / value

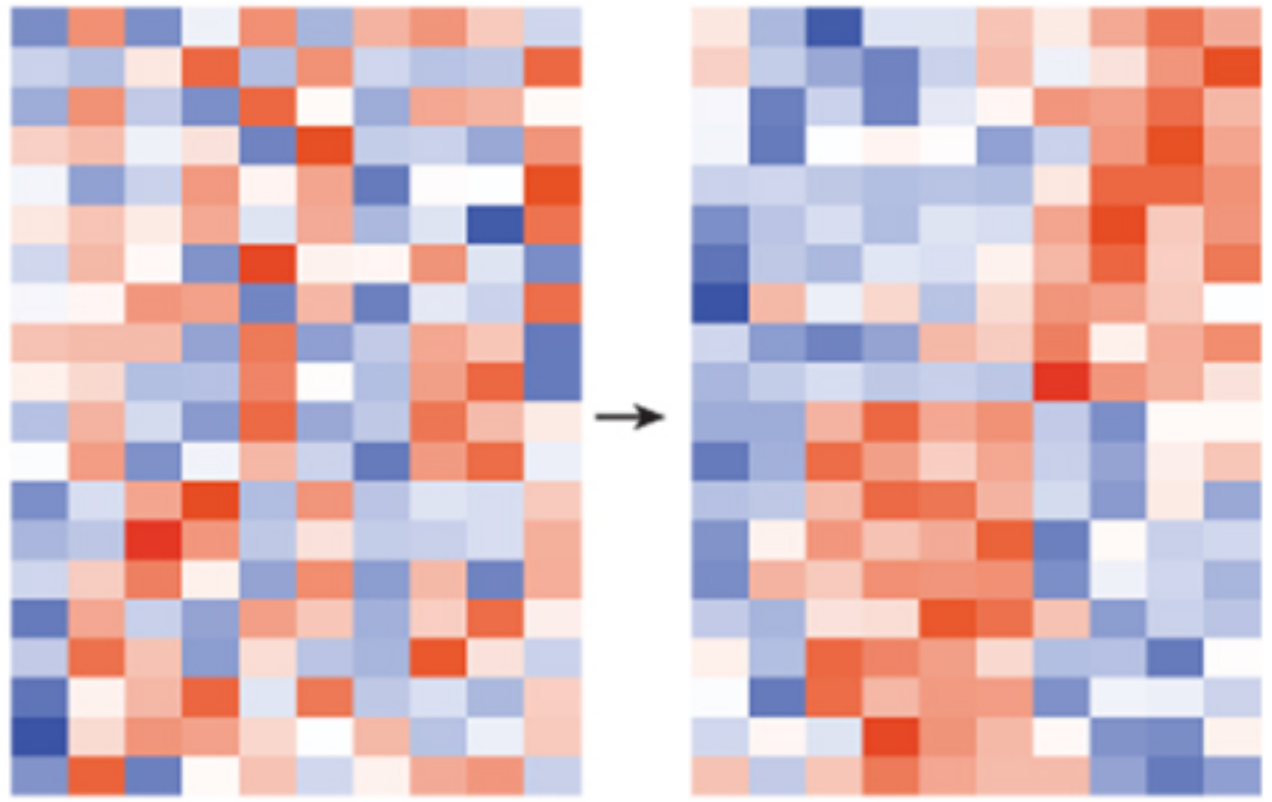
Meaning derived from ordering

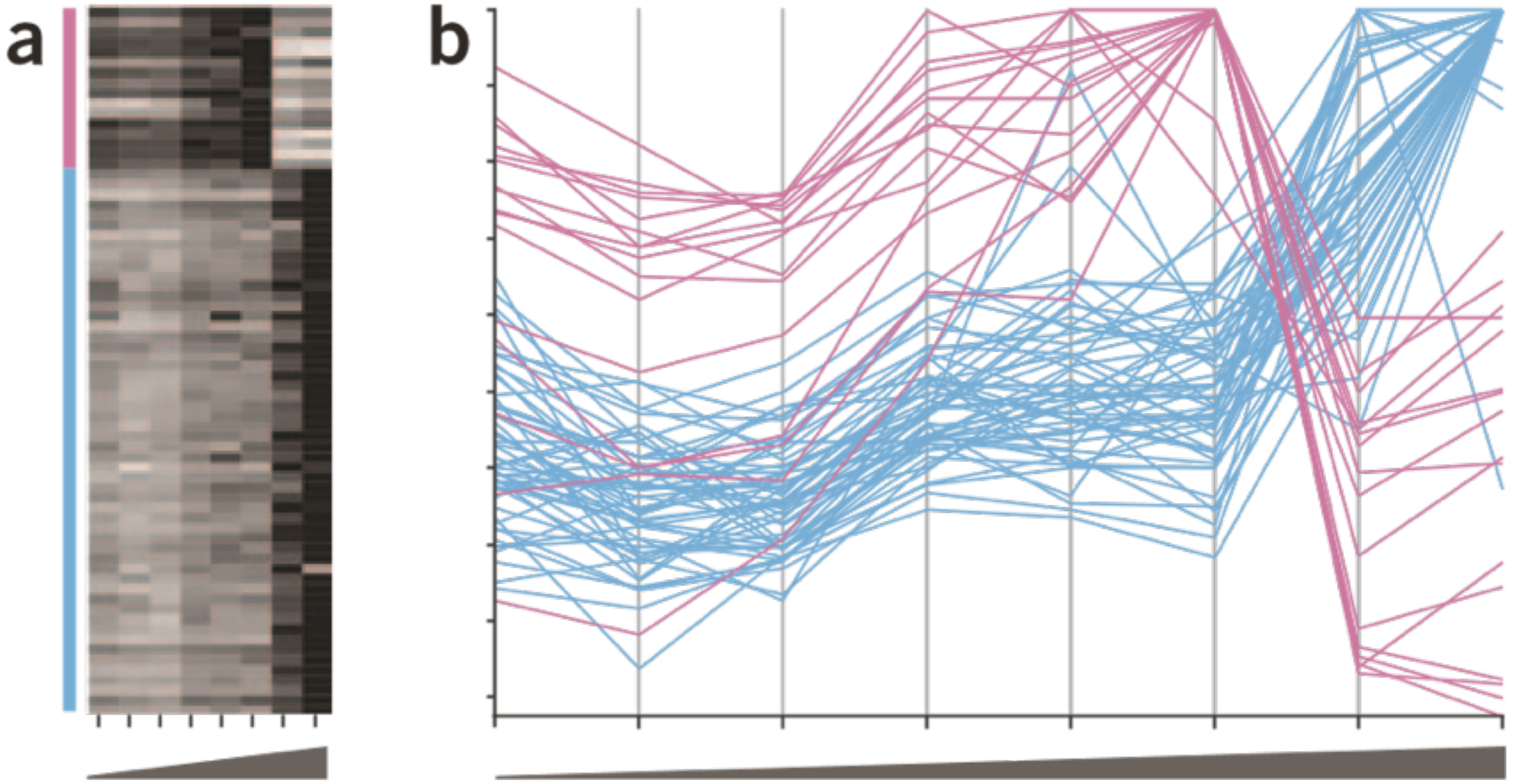
If no ordering inherent, clustering is used

Scalable – 1 px per item

Good for homogeneous data

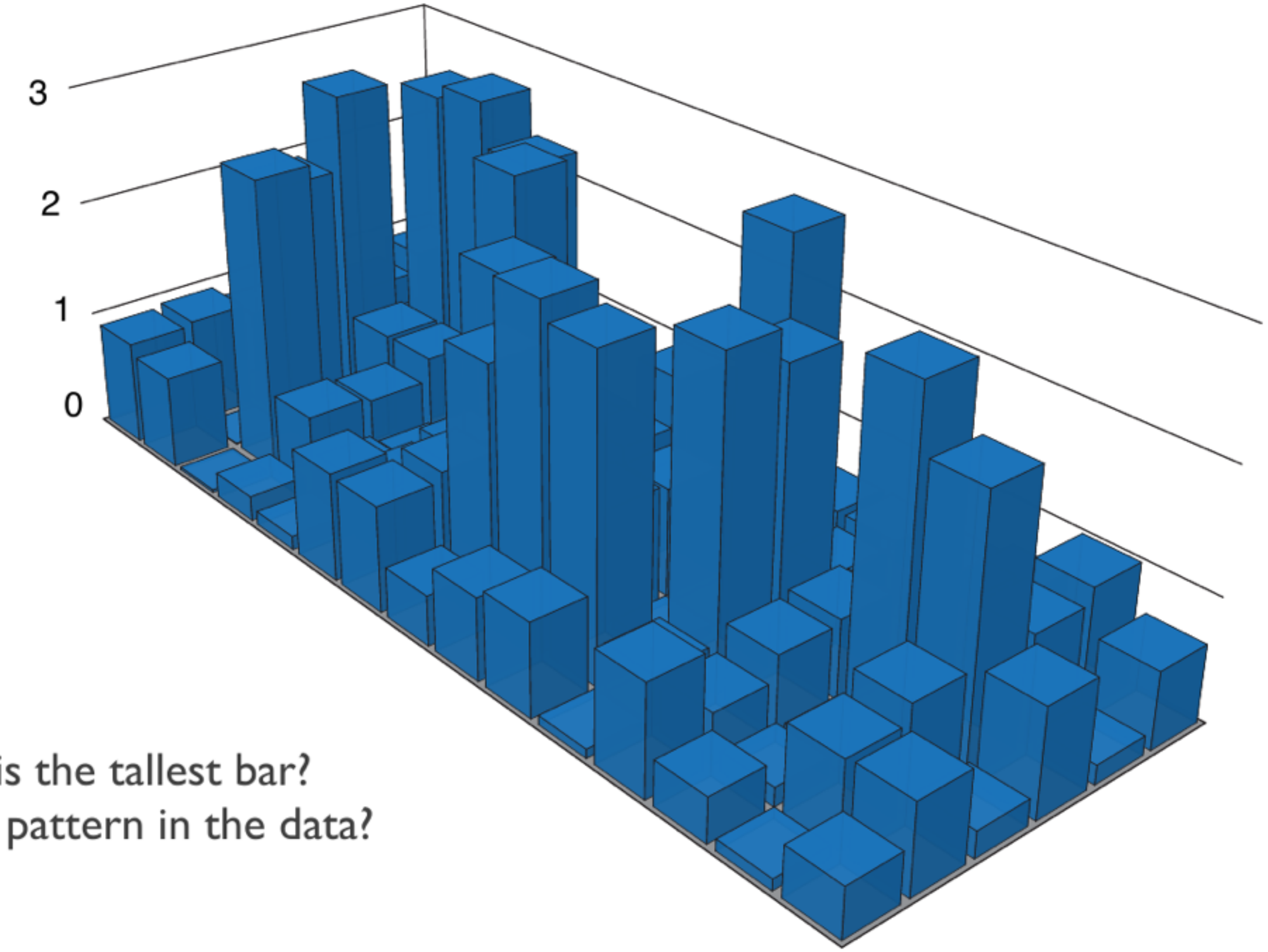
same scale & type





[Gehlenborg & Wong 2012]

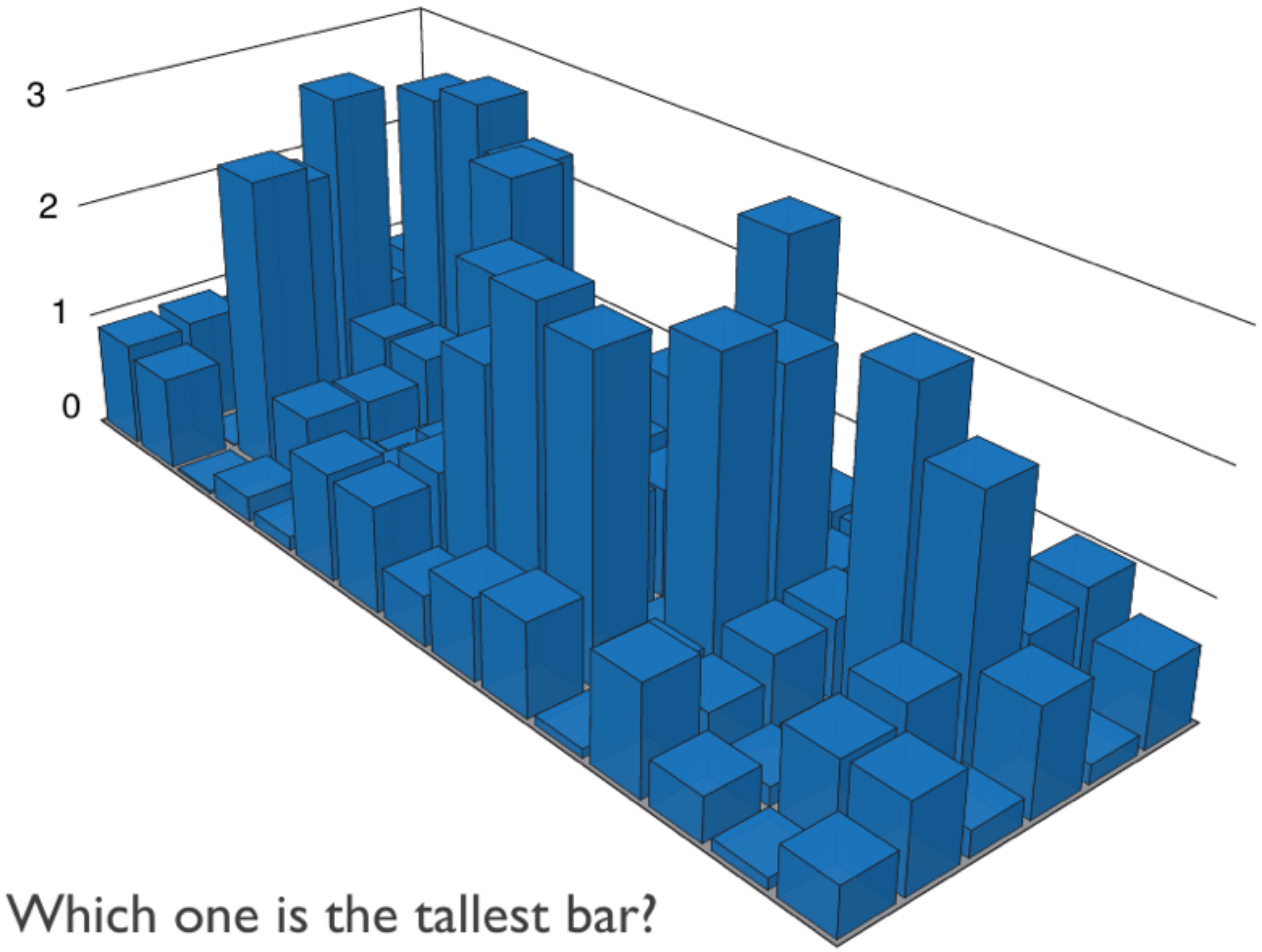
3D Pitfall: Occlusion & Perspective



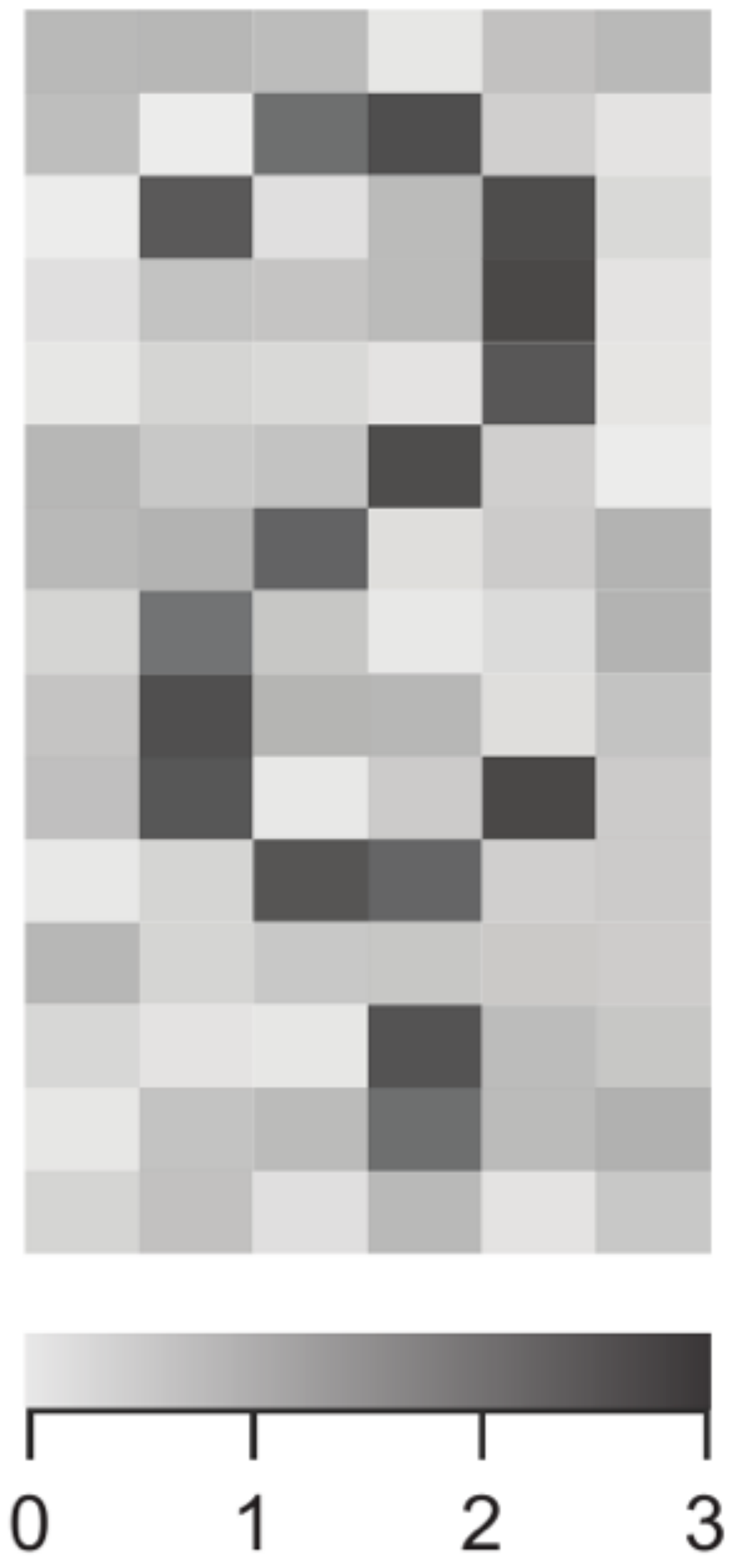
Which one is the tallest bar?
What is the pattern in the data?

[Gehlenborg and Wong, Nature Methods, 2012]

3D Pitfall: Occlusion & Perspective

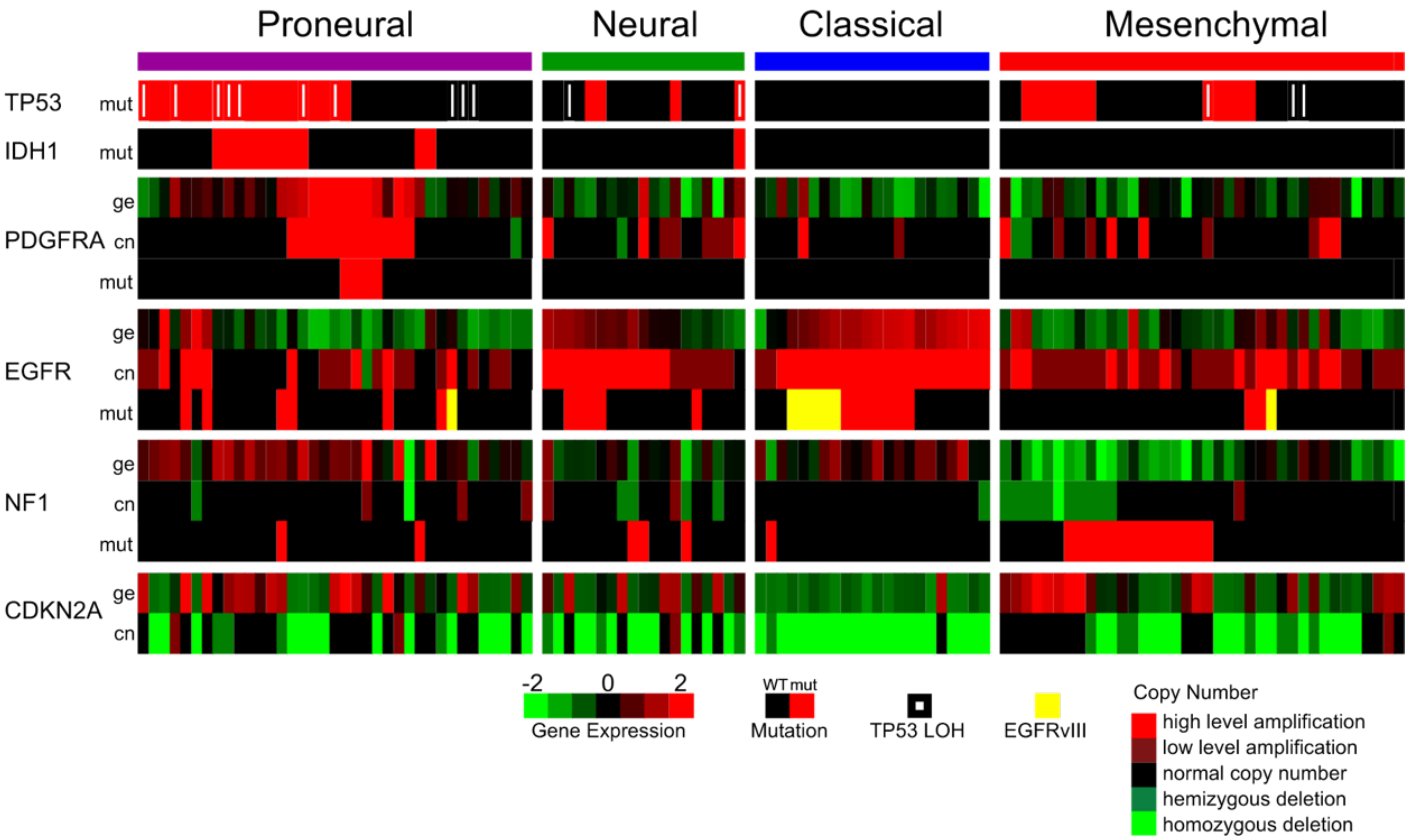


Which one is the tallest bar?
What is the pattern in the data?



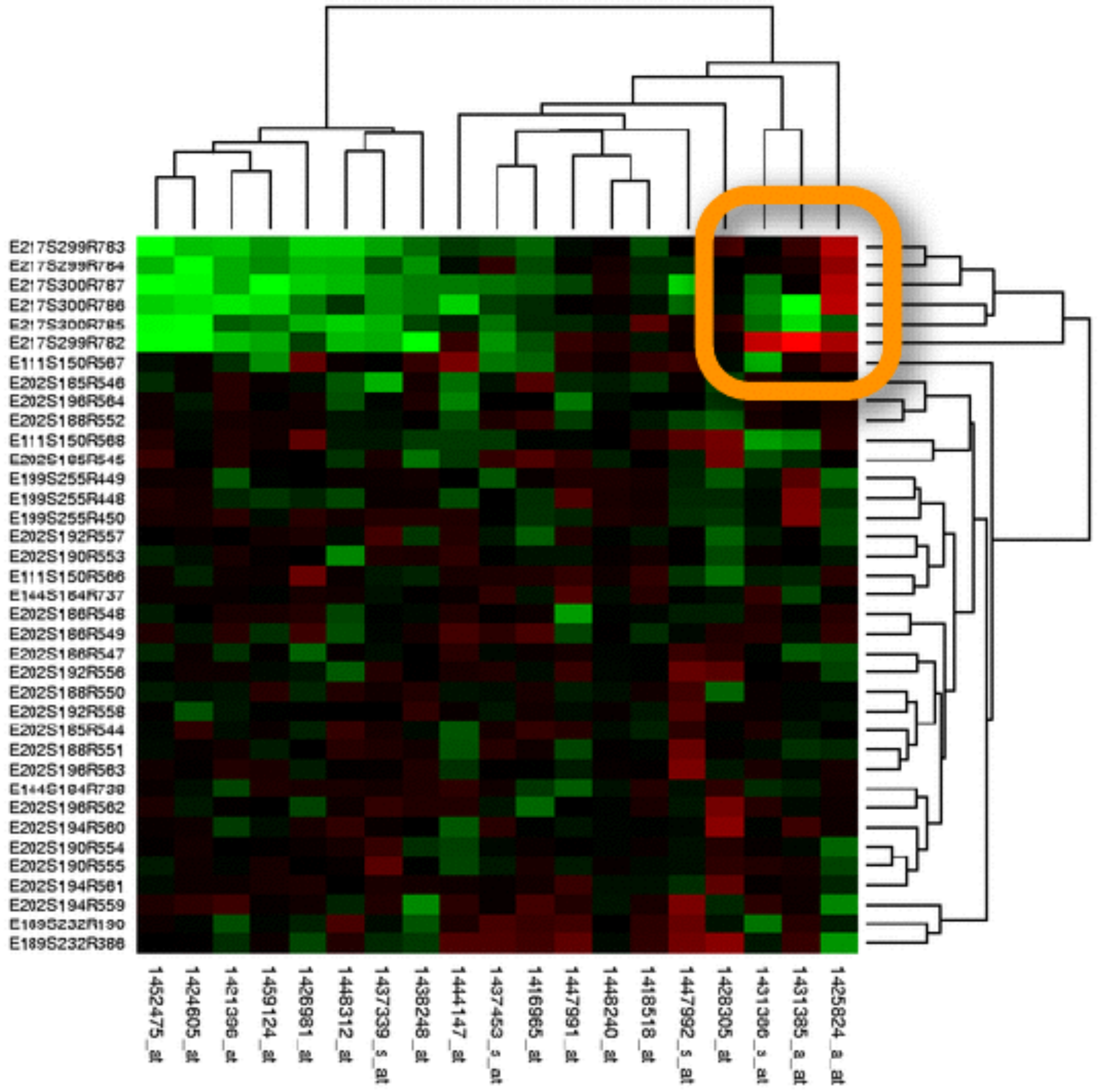
[Gehlenborg and Wong, Nature Methods, 2012]

Heterogeneous Data?

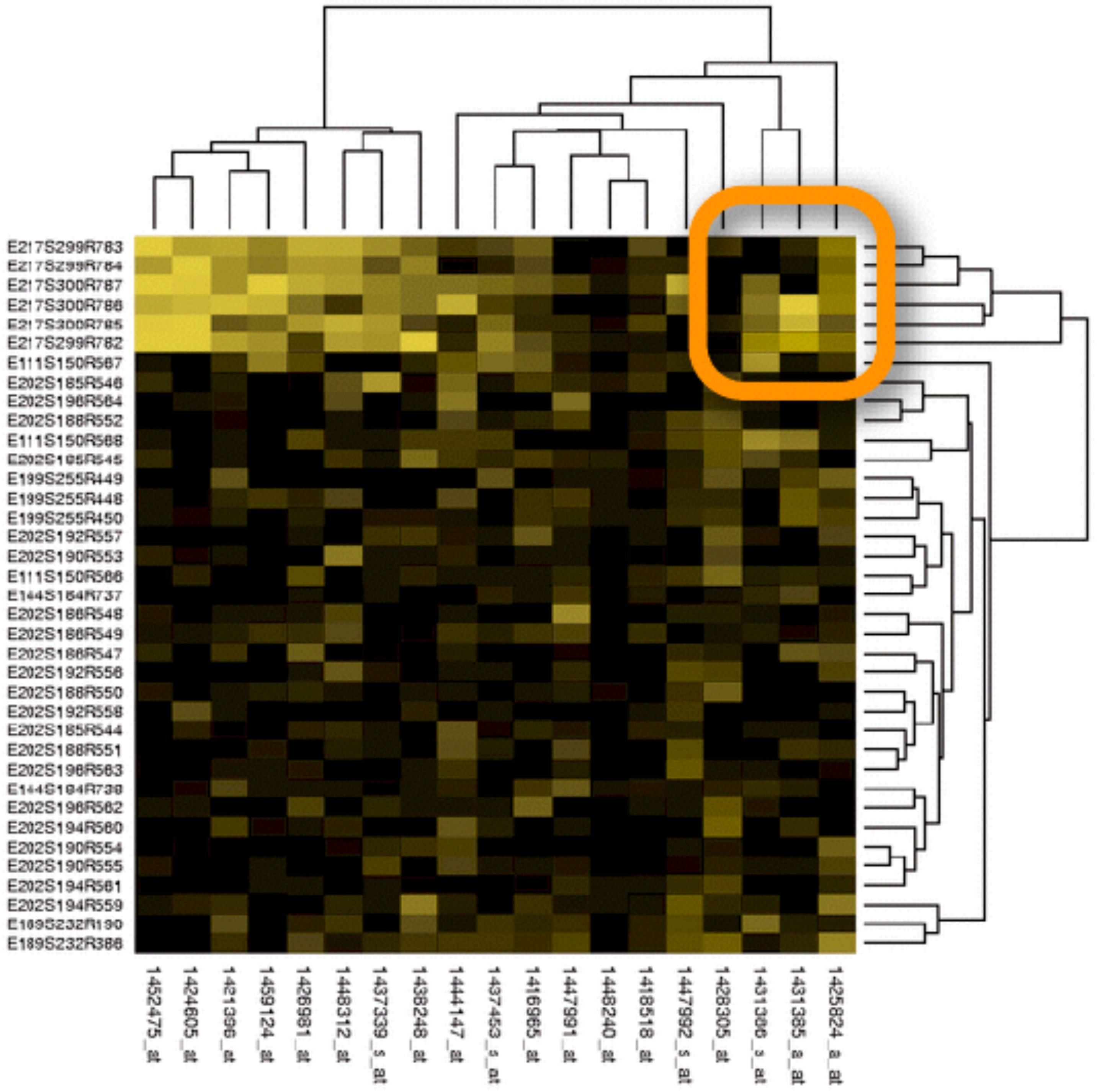


[Verhaak 2012]

Bad Color Mapping

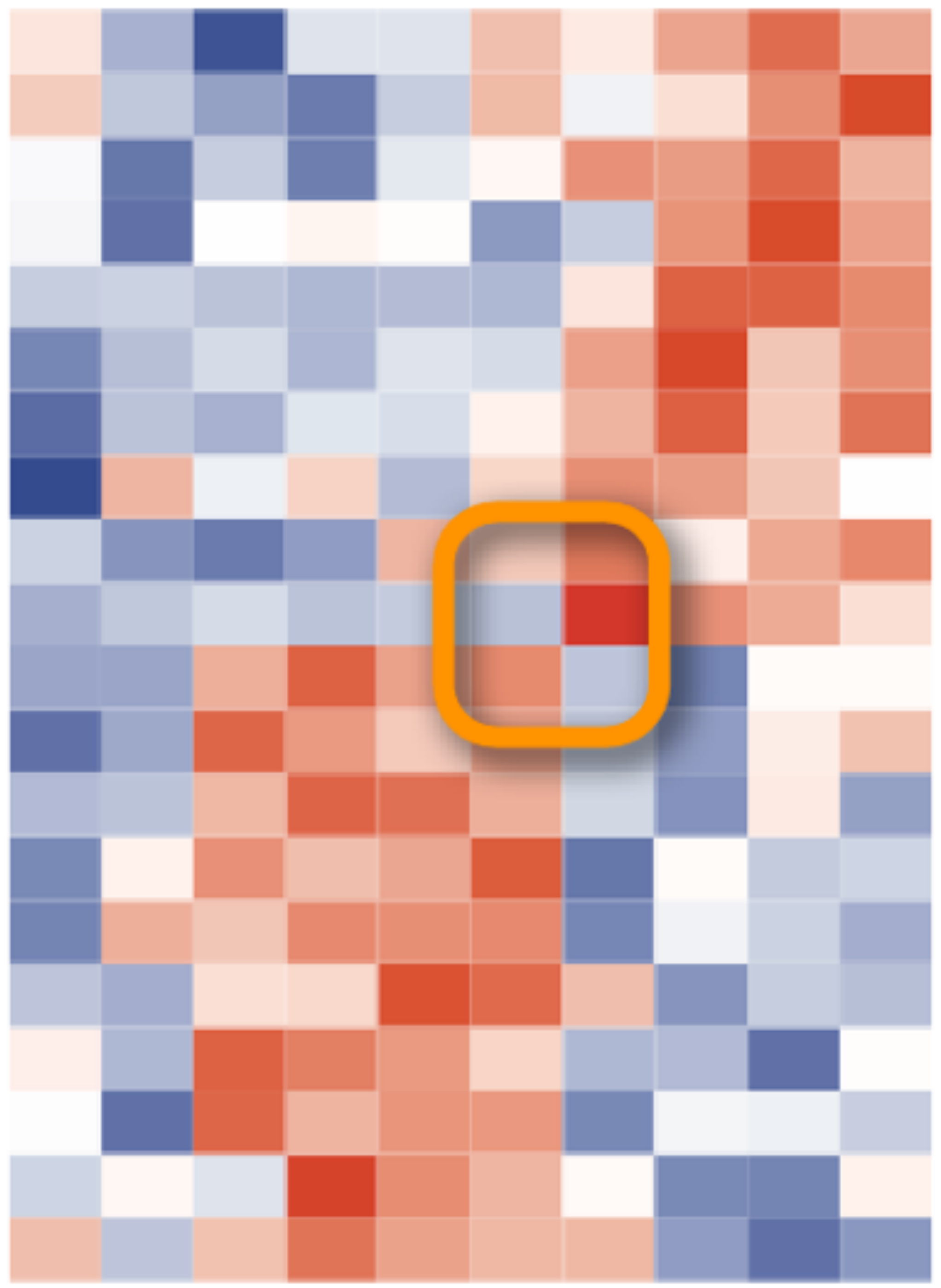


Normal Vision

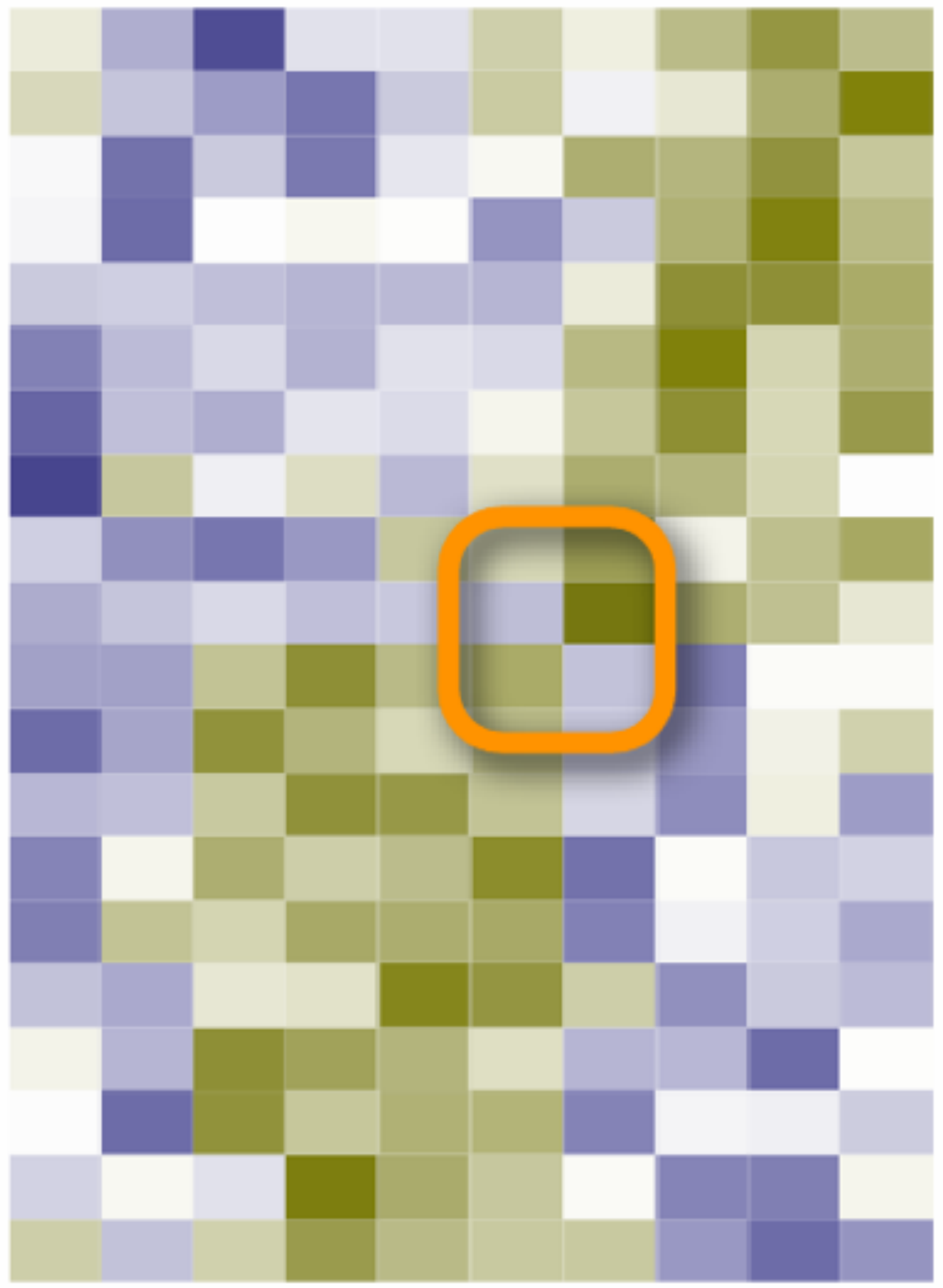


Deuteranope Vision
(“Red-Green Blindness”)

Good Color Mapping

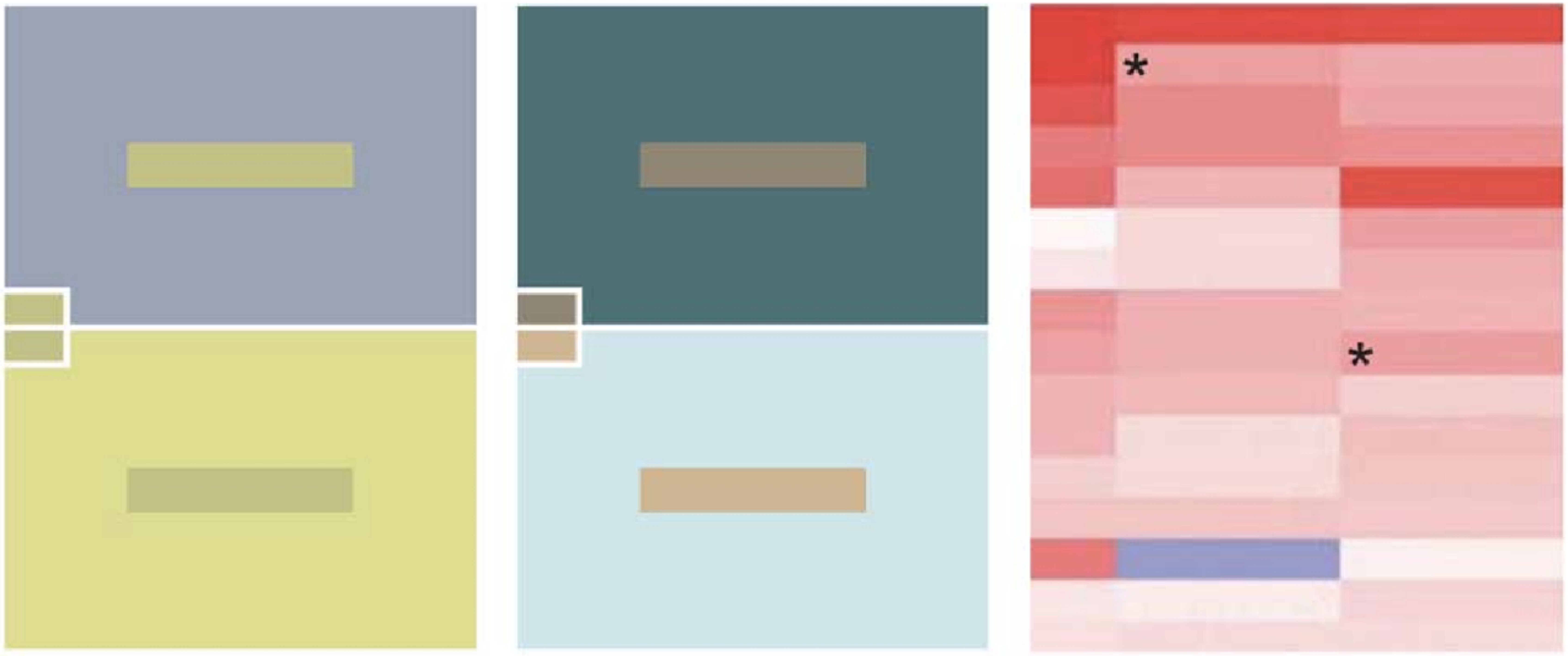


Normal Vision



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., k-means) 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

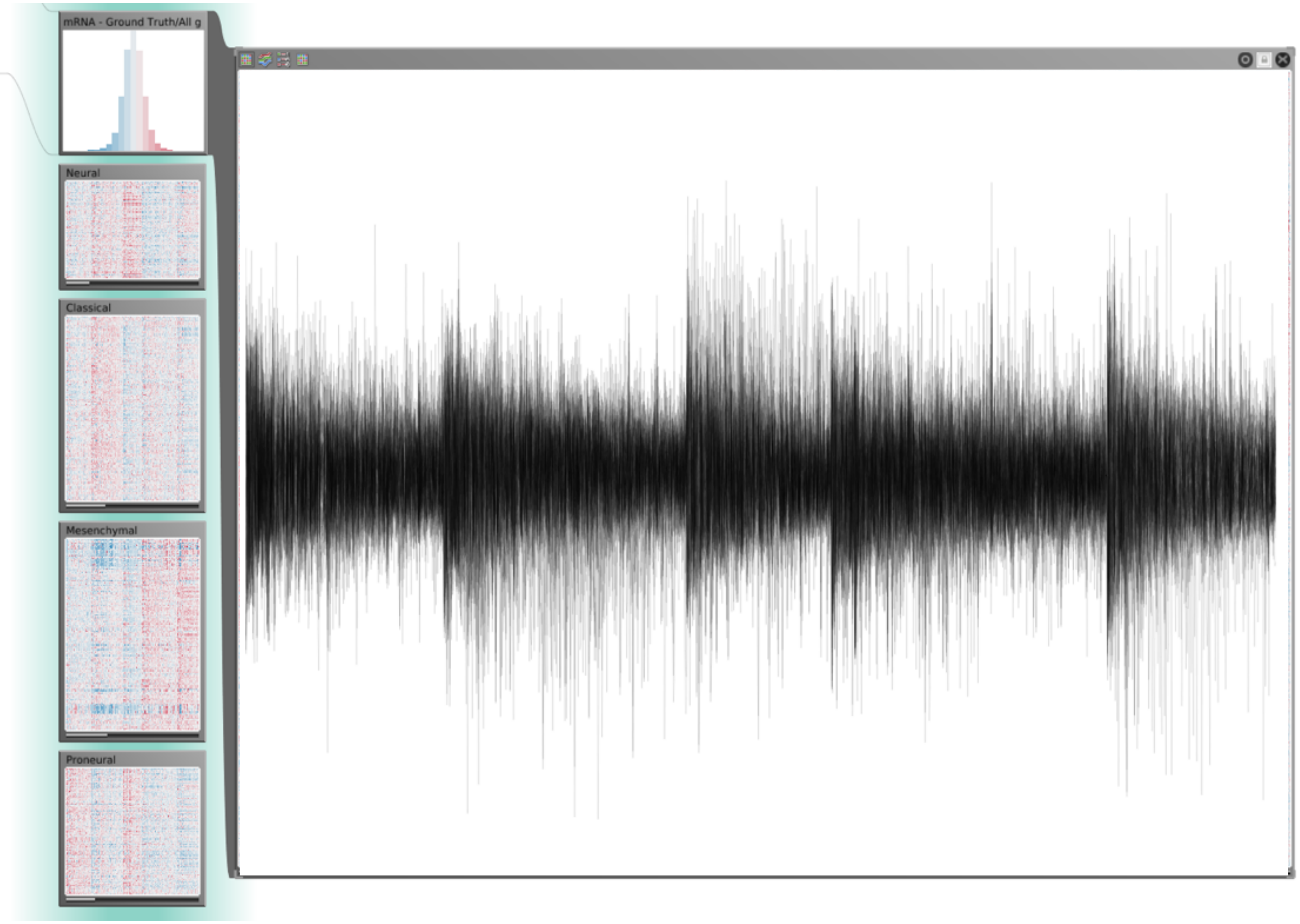
Clusters can be used to

- order (pixel based techniques)
- brush (geometric techniques)
- aggregate

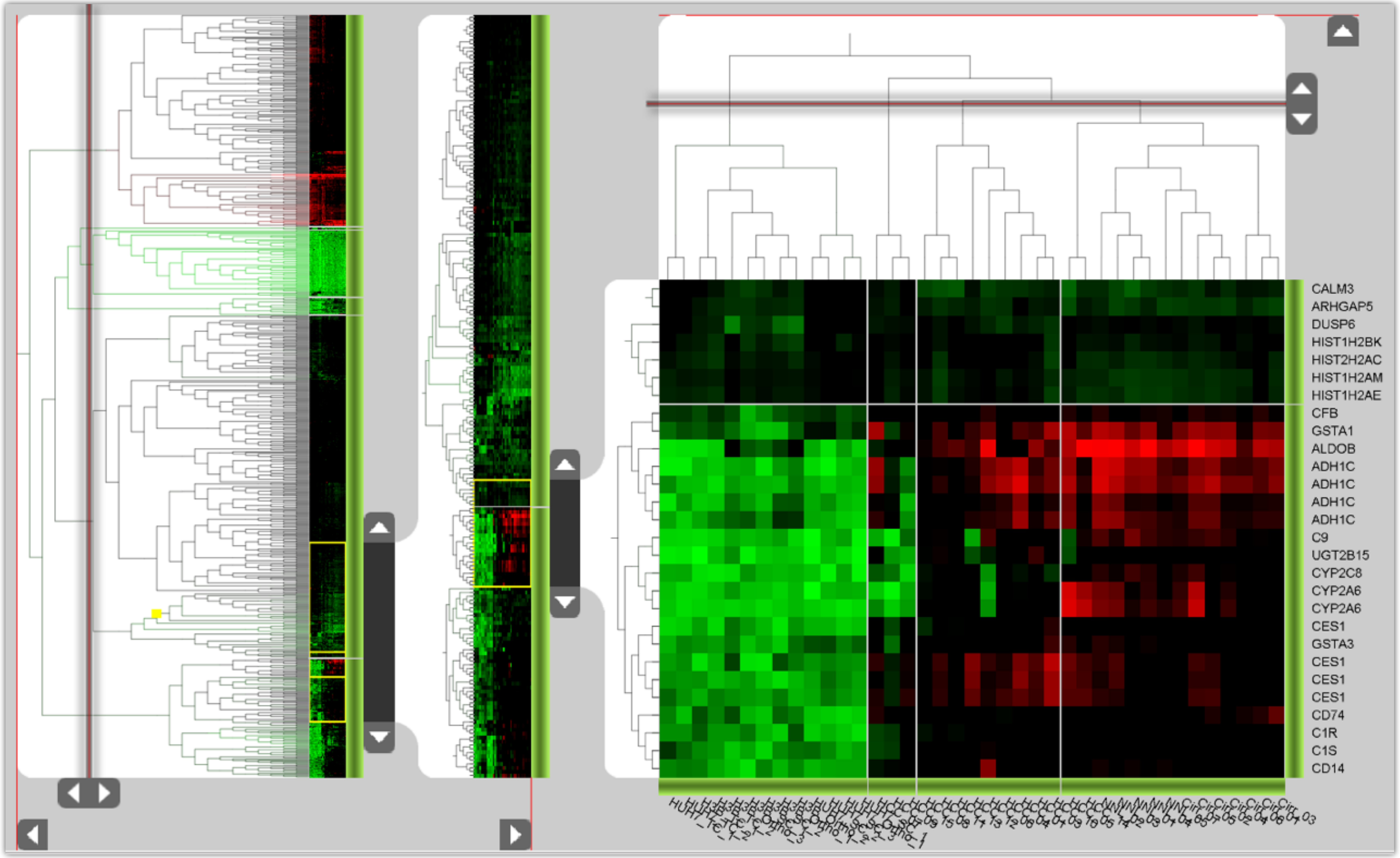
Aggregation

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

Clustered Heat Map

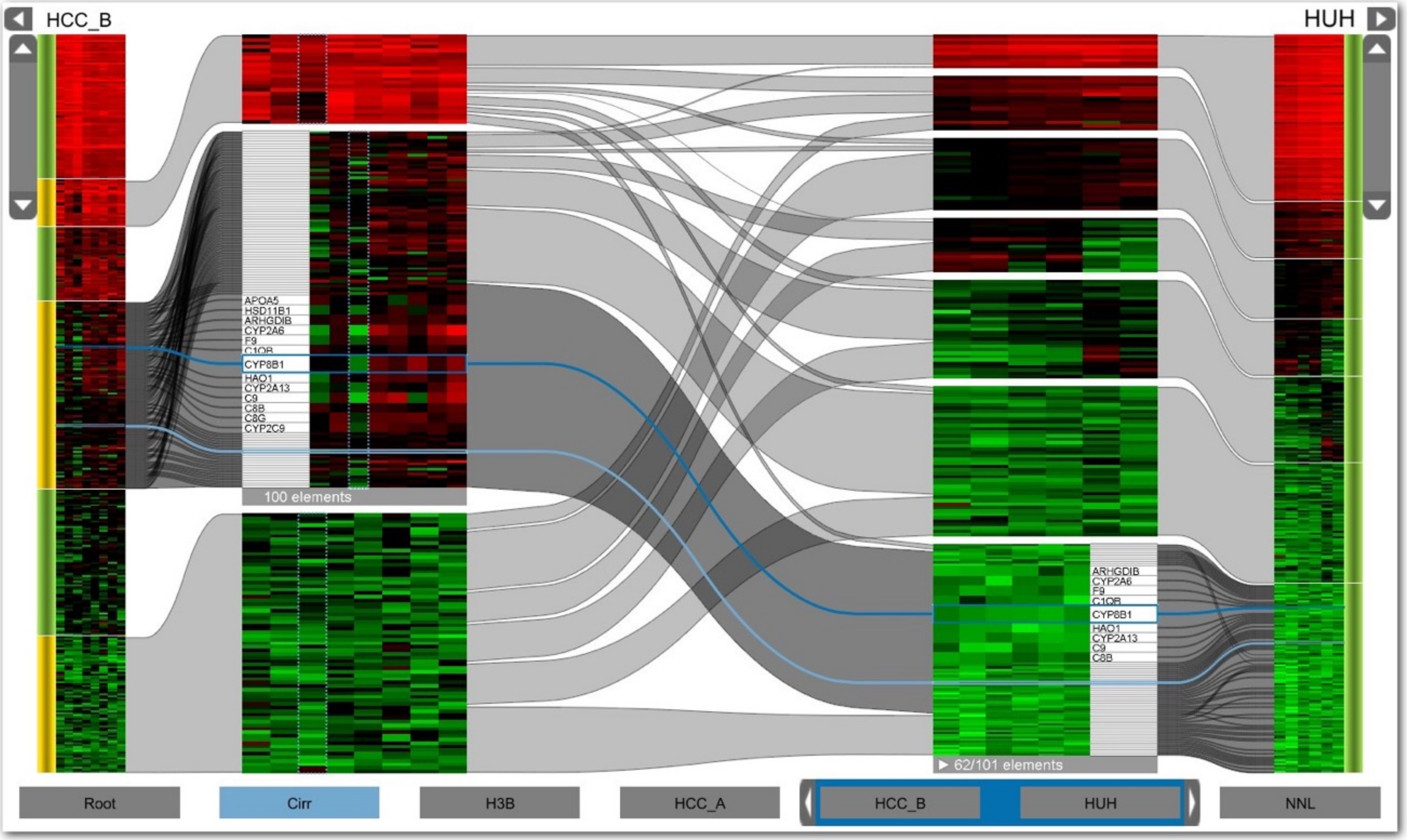


F+C Approach, with Dendrograms

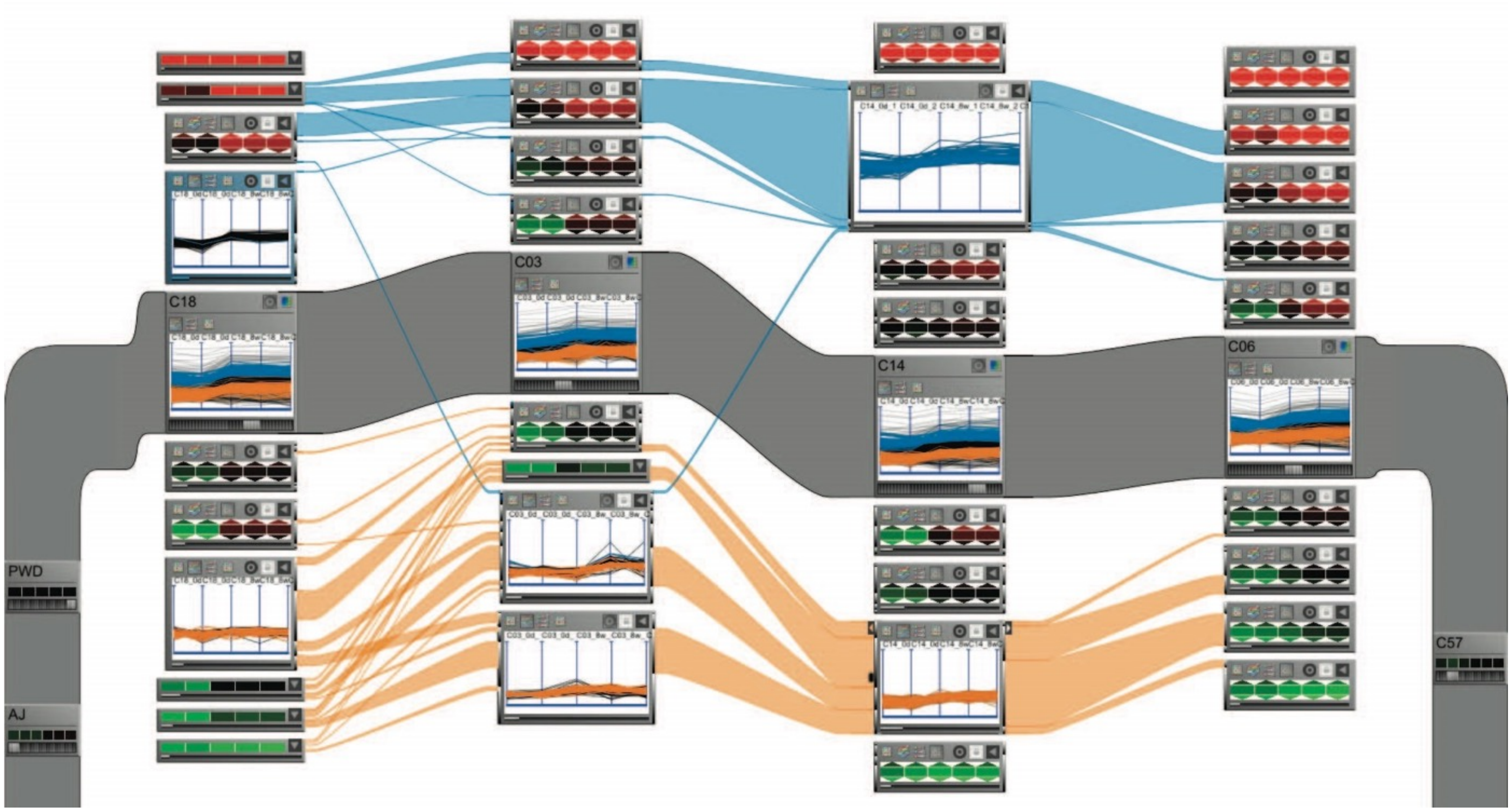


[Lex, PacificVis 2010]

Cluster Comparison

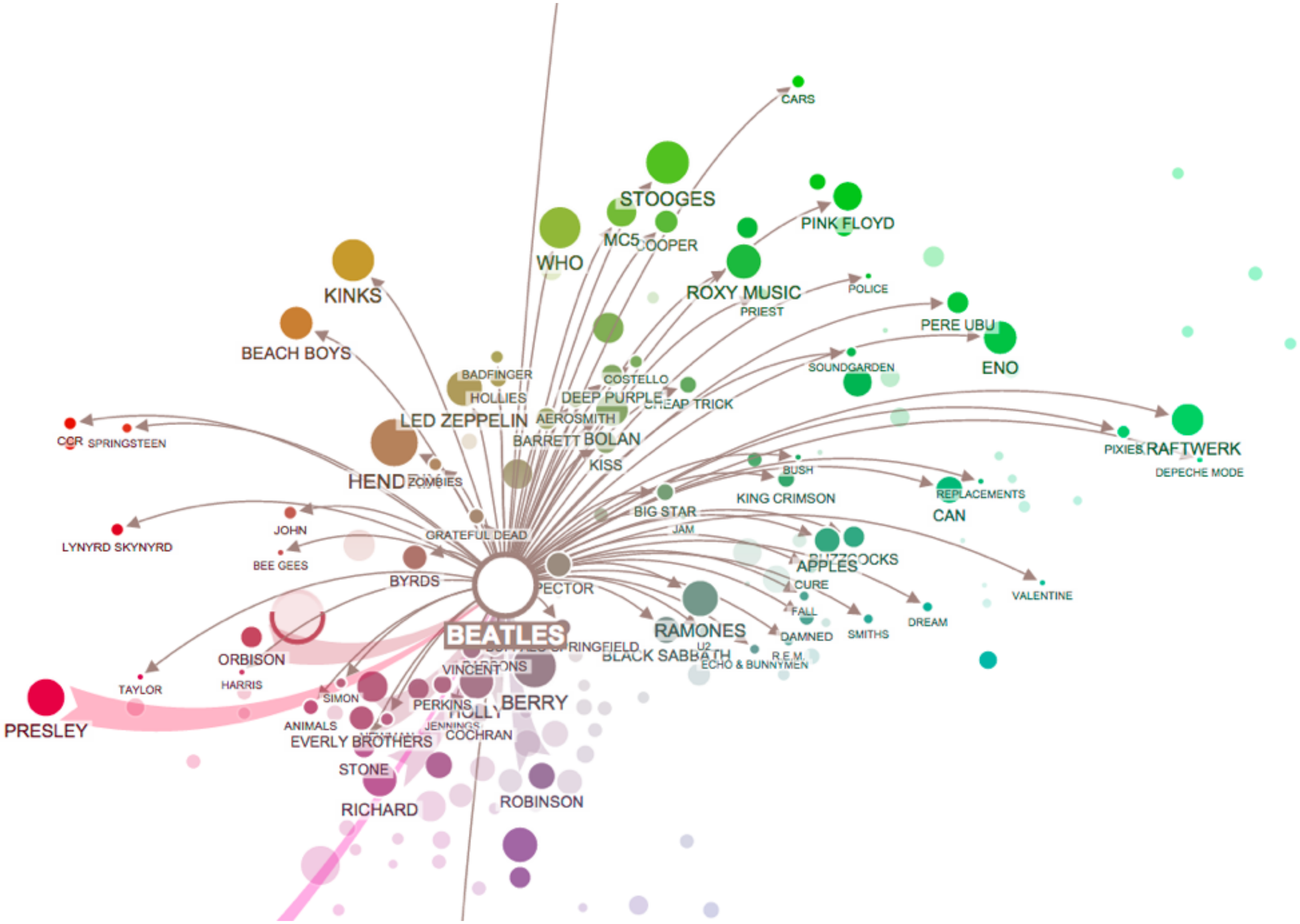


Aggregation



Design Critique

EdgeMaps: http://goo.gl/q8Cv7t



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

Dimensionality Reduction

Dimensionality Reduction

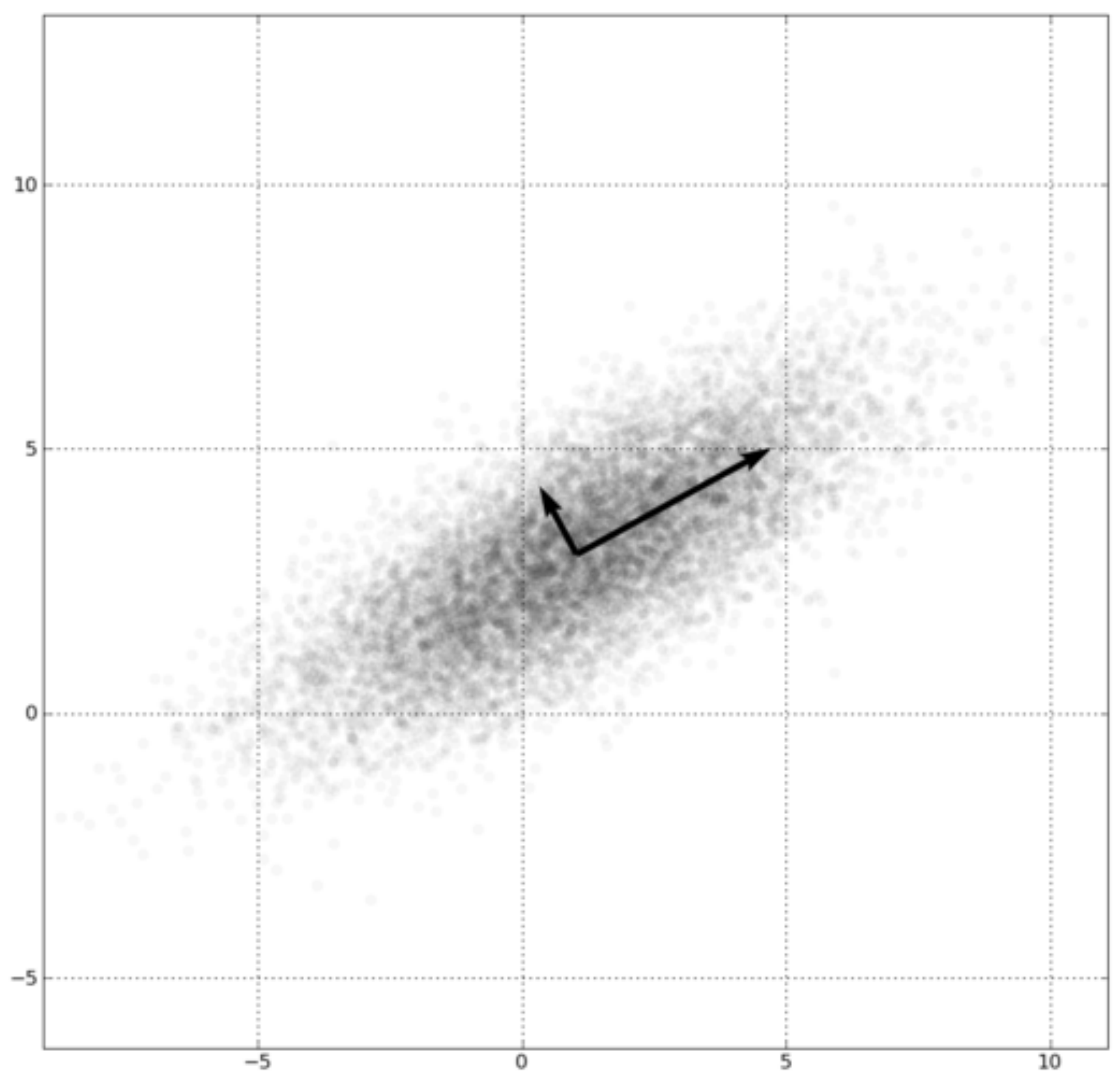
Reduce high dimensional to lower dimensional space

Preserve as much of variation as possible

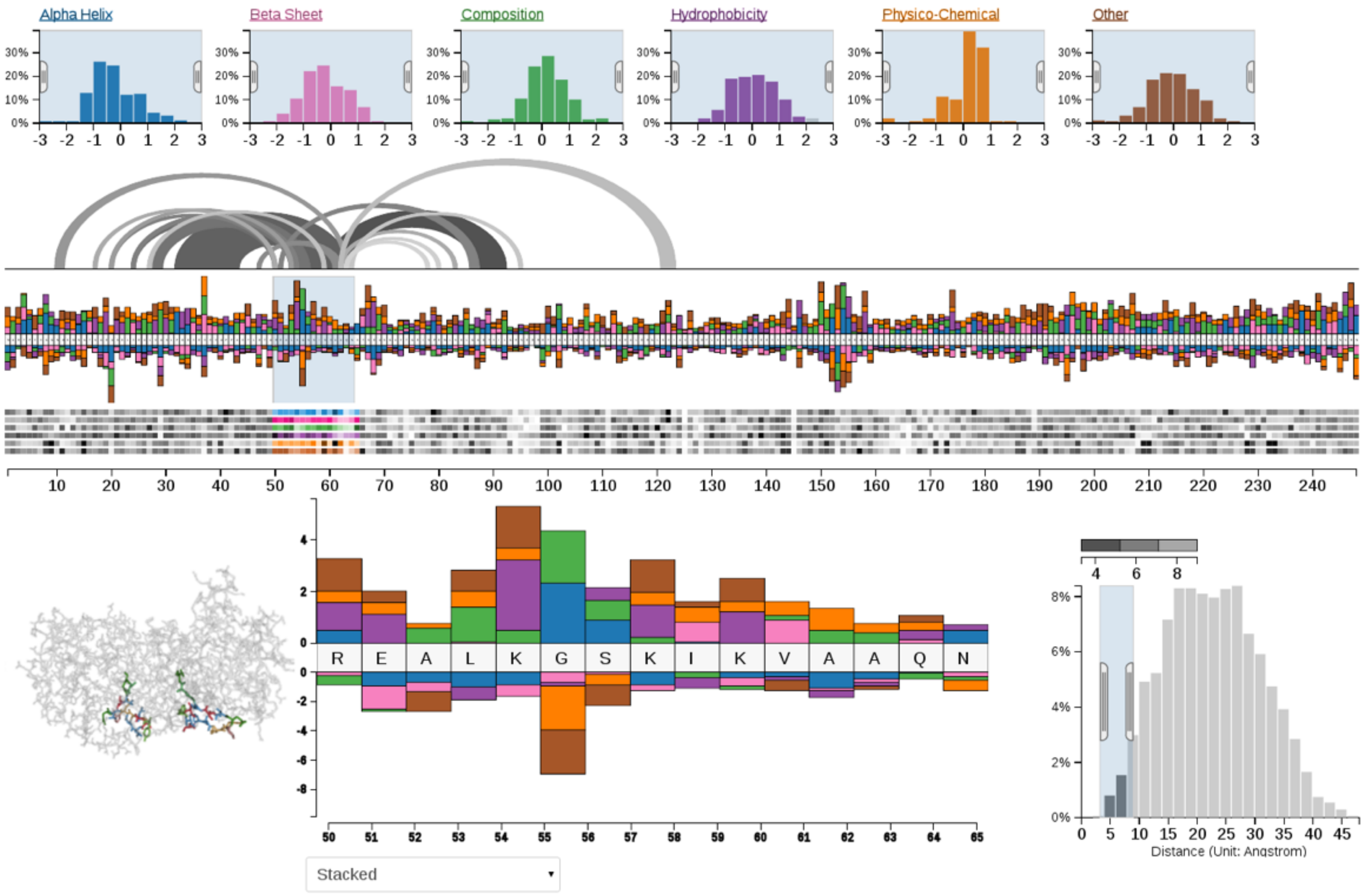
Plot lower dimensional space

Principal Component Analysis (PCA)

linear mapping, by order of variance



PCA Example – CS 171 Project 2013



http://mu-8.com/

[Mercer & Pandian]

Multidimensional Scaling

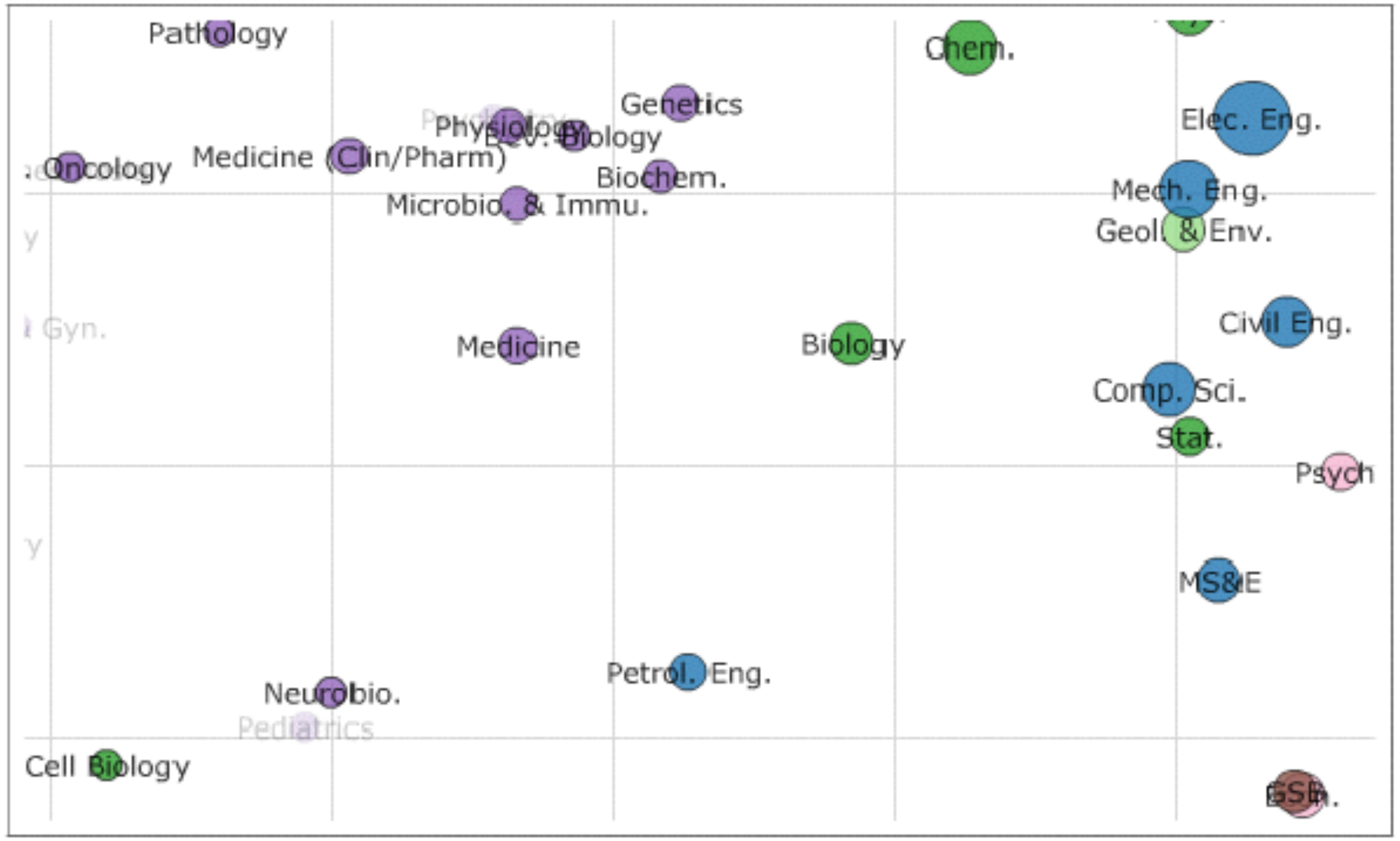
Nonlinear, better suited for some DS

Popular for text analysis

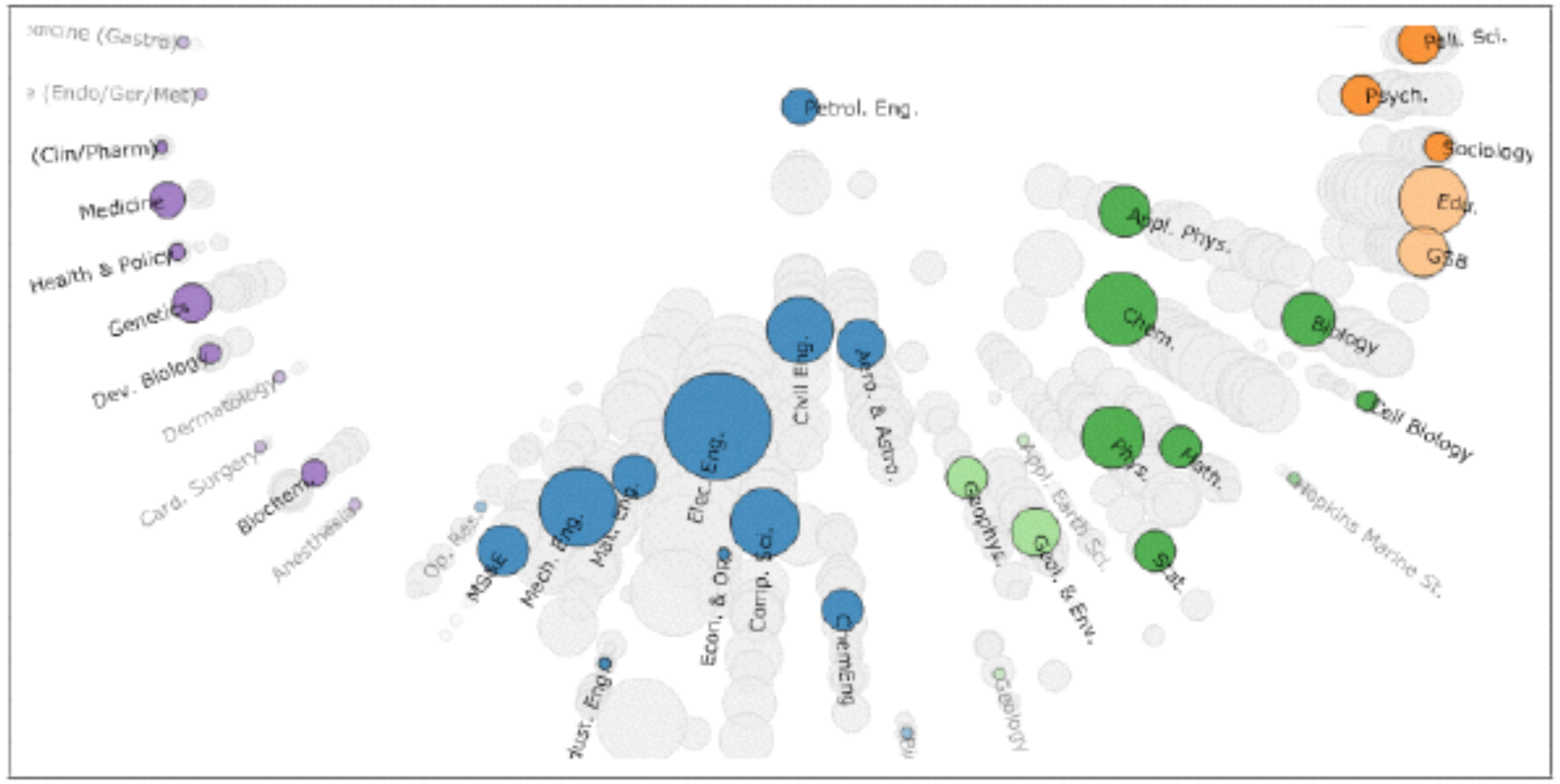
[Doerk 2011]

Can we Trust Dimensionality Reduction?

Topical distances between departments in a 2D projection



Topical distances between the selected Petroleum Engineering and the others.



[Chuang et al., 2012]

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