CSE 564
Visualization & Visual Analytics

Visualizing High-Dimensional Data: Linear Methods

Klaus Mueller

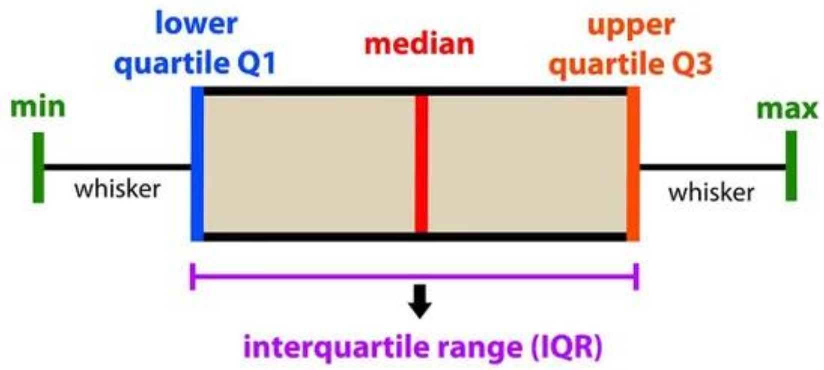
Computer Science Department
Stony Brook University

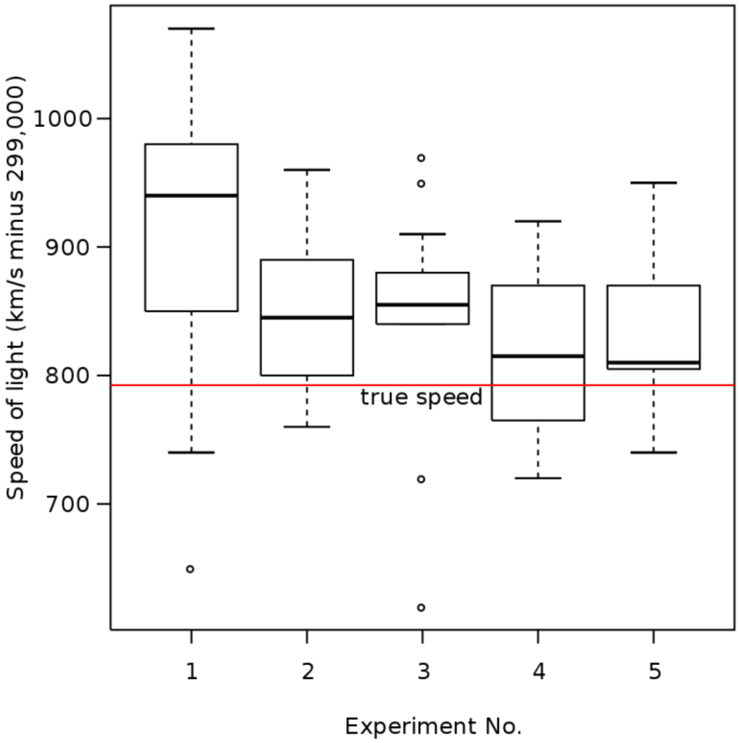
Lecture	Topic	Projects
1	Intro and logistics	
2	Basic visualizations and tasks, data types, examples, ethical considerations	
3	Data preparation (cleaning, imputation, data set integration)	
4	AI-assisted coding for VIS applications (design, debugging, refactoring)	Project #1 out
5	Big data and data reduction (distance/sim metrics, intro to clustering)	
6	High-D data: concept, subspaces, dimension reduction, PCA	
7	Cluster analysis: hierarchical, density, model, embedding, temporal	
8	Perception and cognition (human visual system, color, contrast)	Project #2(a) out
9	Visual design and aesthetics	
10	Visualization of multivariate and high-D data: linear methods, projections	
11	Vis. of multivariate and high-D data: non-linear methods, embeddings	
12	Visualization and AI: mutual support and capabilities (VIS4AI, AI4VIS)	Project #2(b) out
13	Principles of interaction: drive what is visualized, analyzed & how (HCI4VIS)	
14	Visual analytics (VA), human-centered AI, mixed-initiative system	
15	Midterm #1 (tentative date)	
16	VA system design and evaluation, collaborative VA, uncertainty, provenance	
17	Midterm #1 discussion (tentative date)	Final proj. proposal call out
18	Visualization of hierarchical data	
19	Visualization of maps and data with geo-reference	
20	Visualization of graphs, networks (incl. derivation of causal networks)	Final project proposal due
21	Vis. of time-varying, time-series, streaming data, progressive visualization	
22	Visualization of text, LLMs, and semantic data	
23	Ed Tufte revisited: principles, critiques and limits, responsible visualization	
24	Design of effective infographics	Final proj. prelim report due
25	Foundations scientific and medical visualization, intro to volume rendering	
26	Scientific visualization	Bonus project out (Vol Ren)
27	Story telling with data, data journalism	
28	Midterm #2 (tentative date)	
Final	Final project demo on zoom (public)	All final proj. materials due

Visualizing Distributions

Box Plots

You may have heard about box plots





Tend to be bewildering to many

- hard to interpret

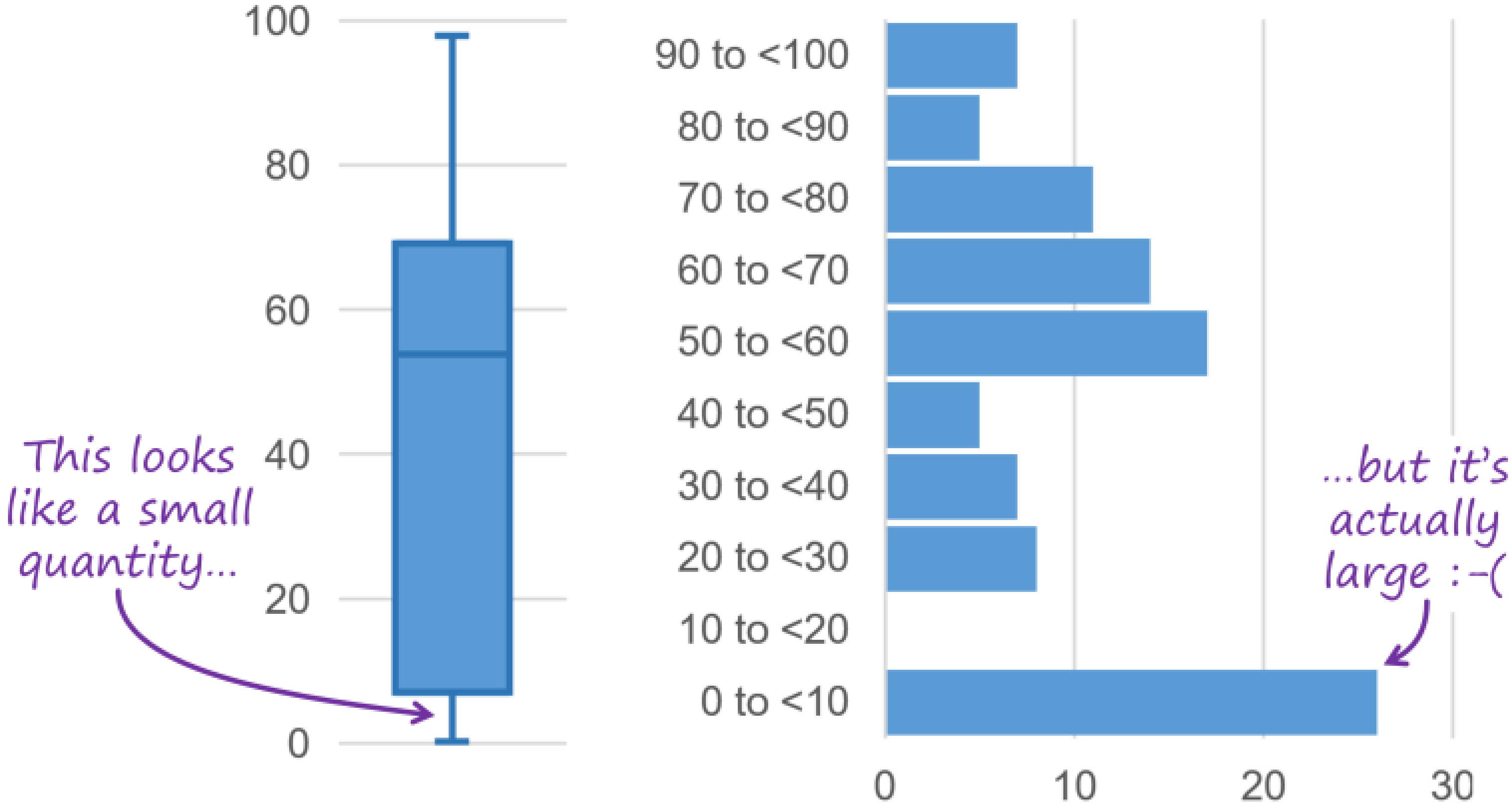
They can also give the wrong representation of data

- assume normal distributed data

Box Plots

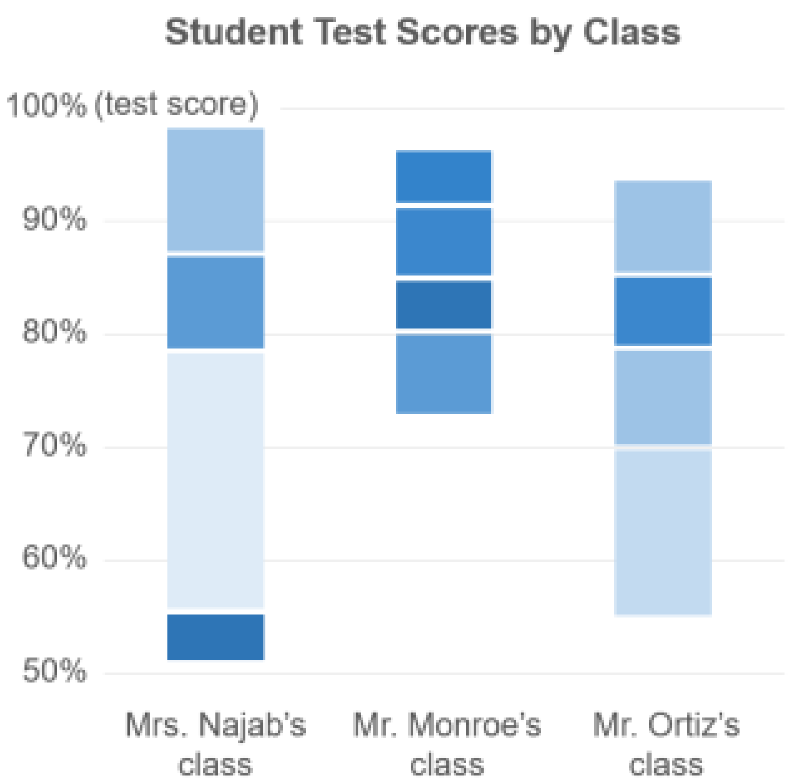
Non-normal distributed data give “wrong” box plots

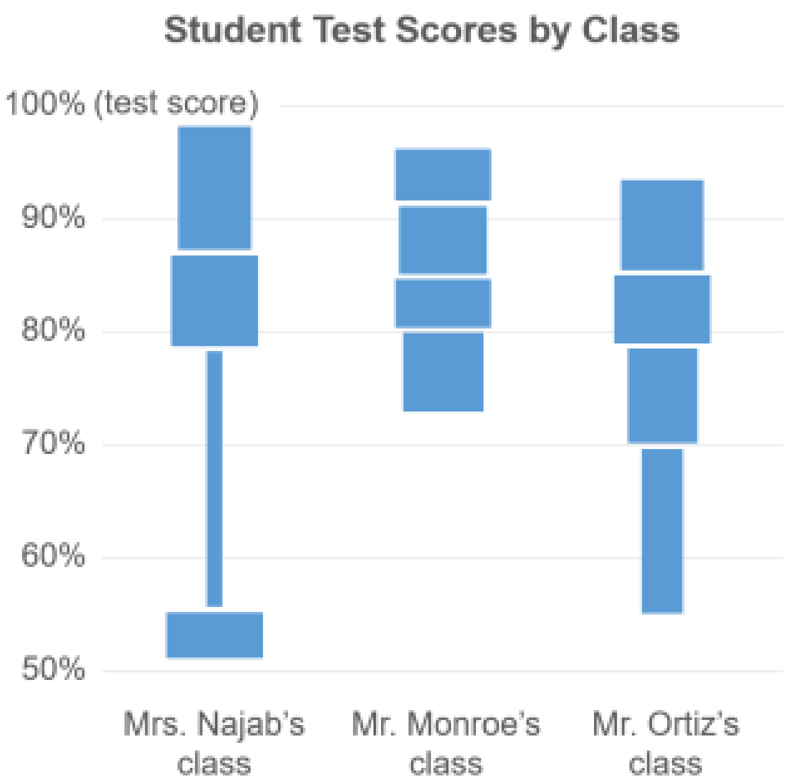
- shown here: data on student test scores



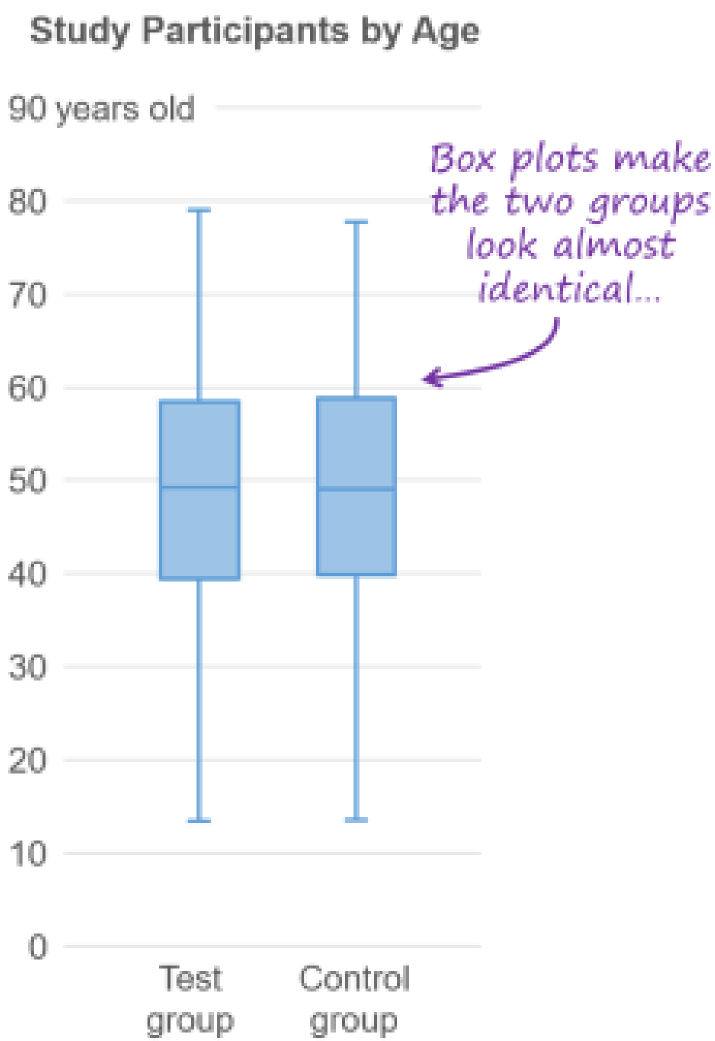
Density Plots

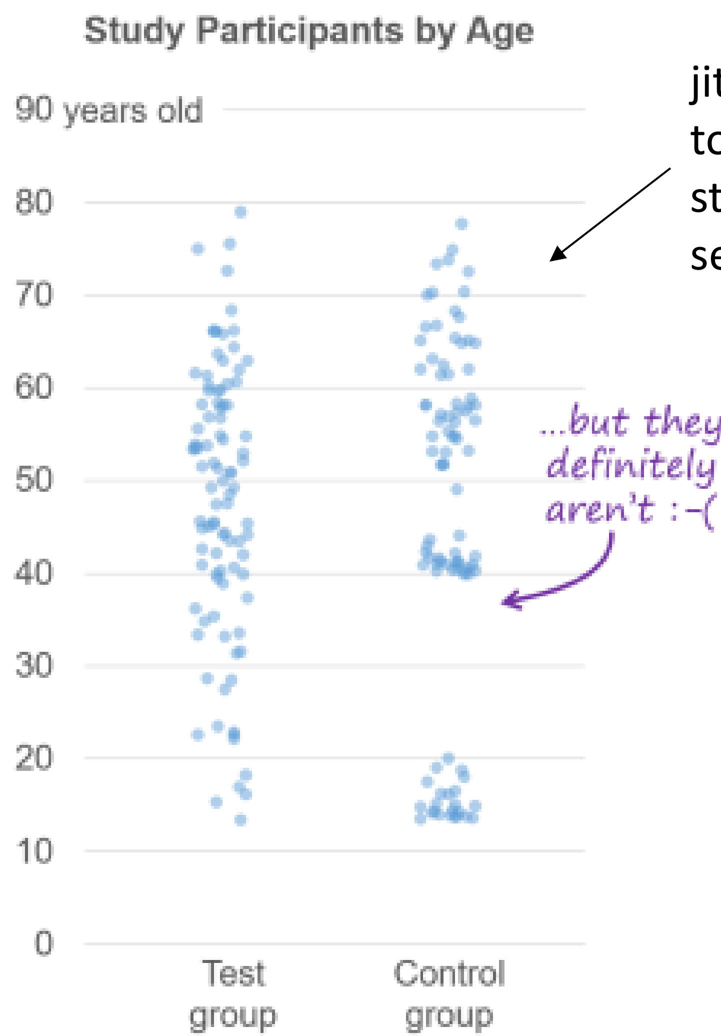
Same data than last side, multiple classes





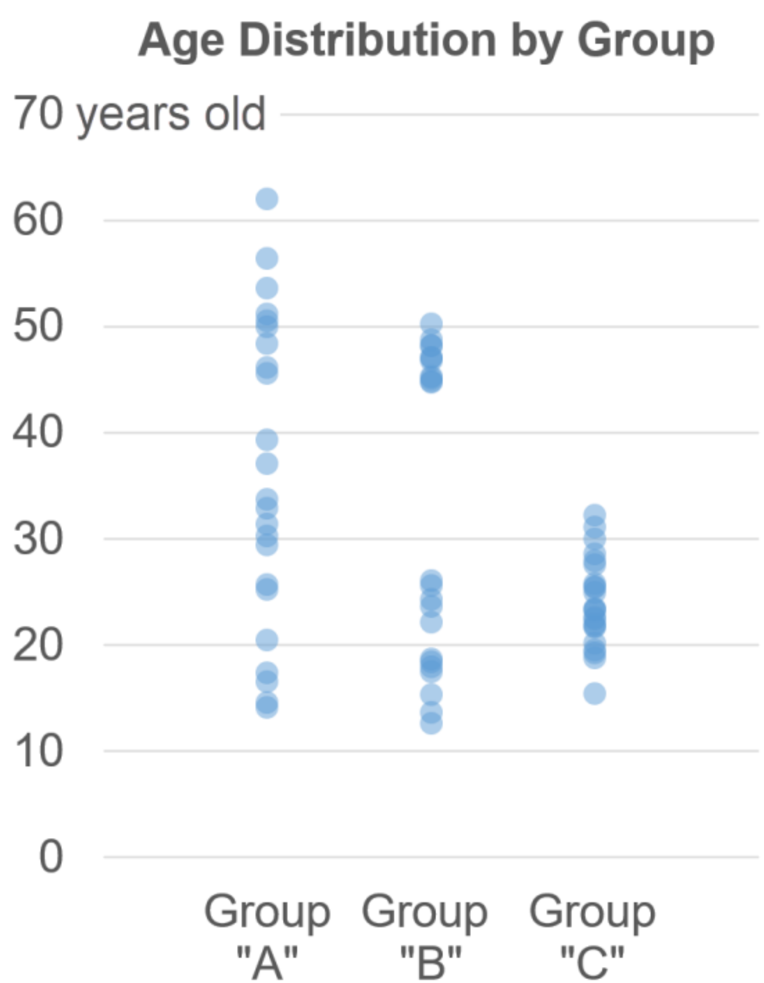
Strip Plots

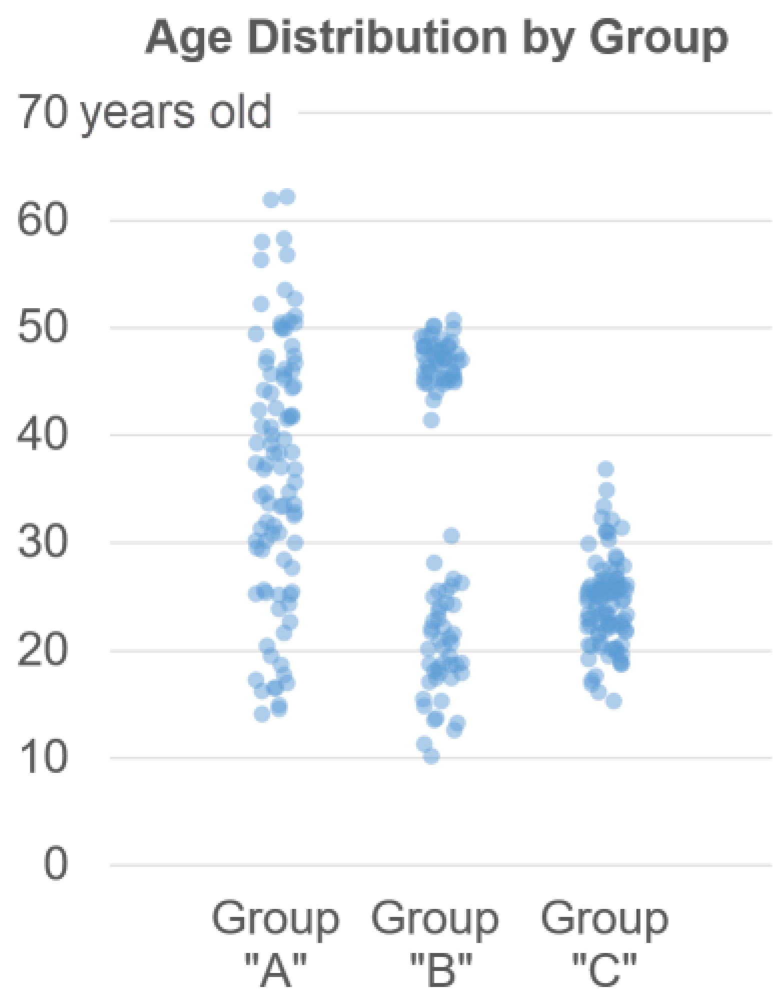




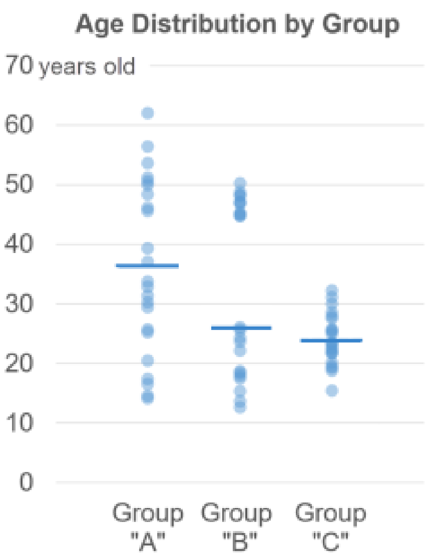
jittered points to make the stats easier to see

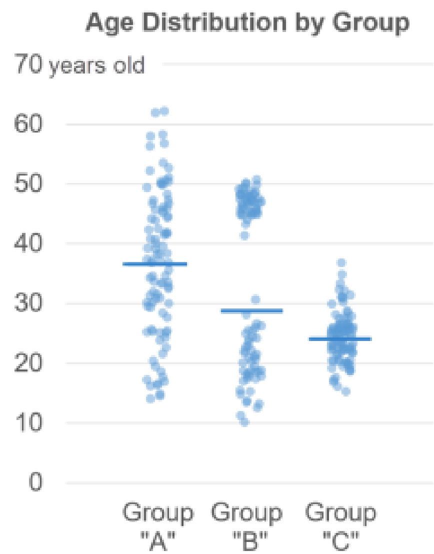
Semitransparent vs. Jittering

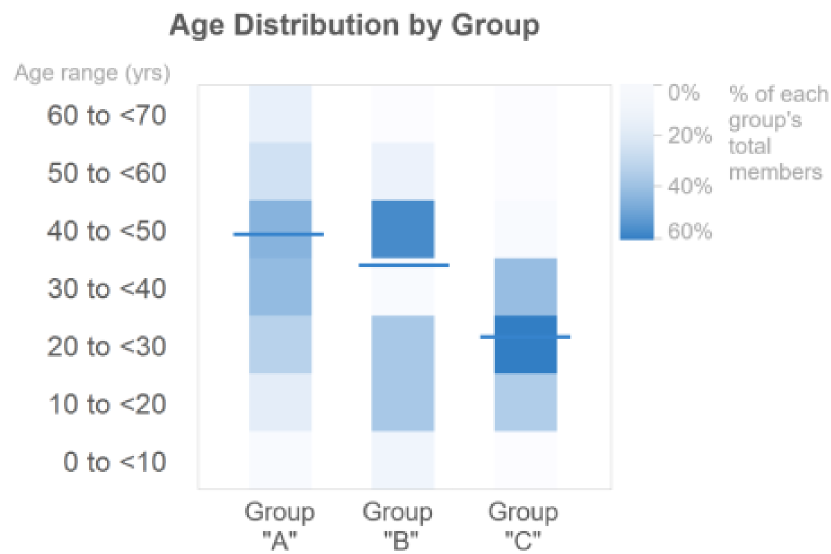




Comparison



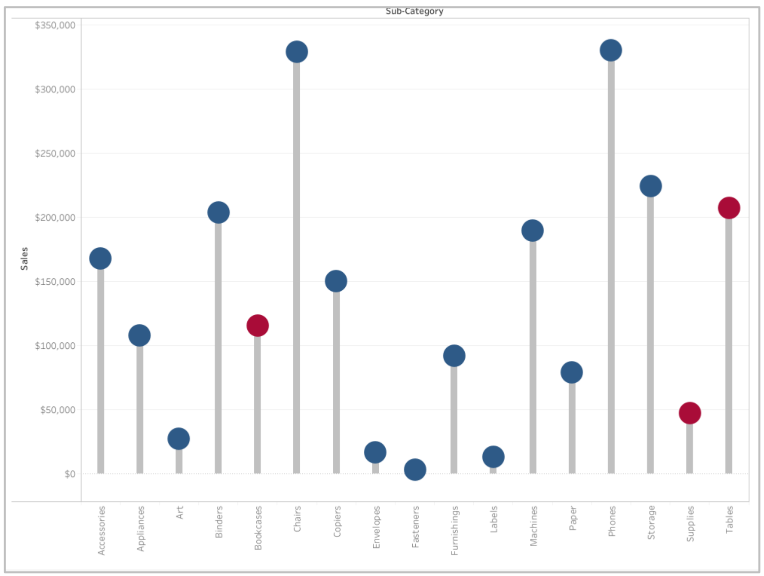


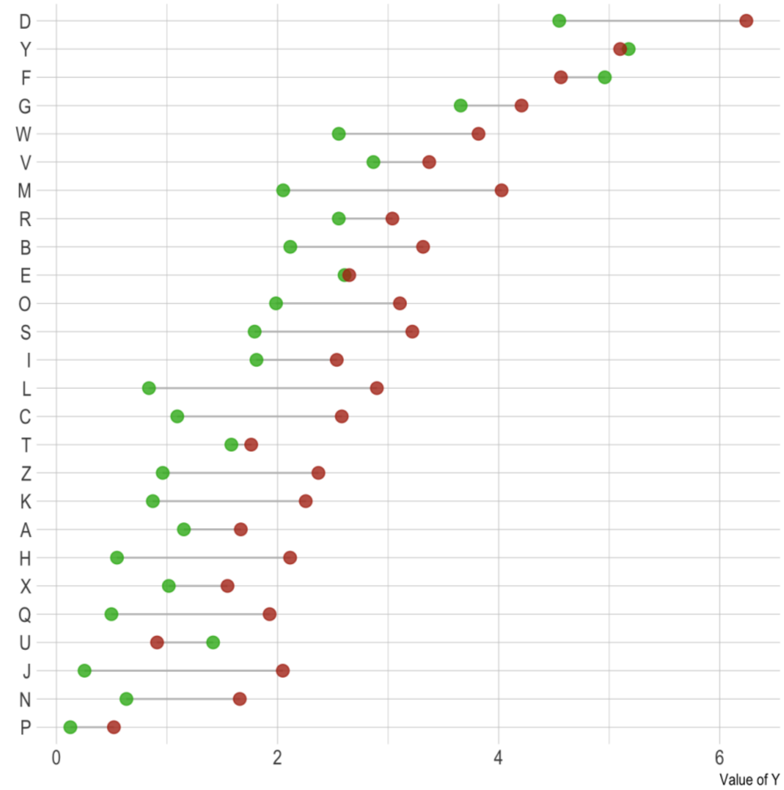


With median lines

Read more here:
https://nightingaledvs.com/ive-stopped-using-box-plots-should-you/

Lollipop Charts



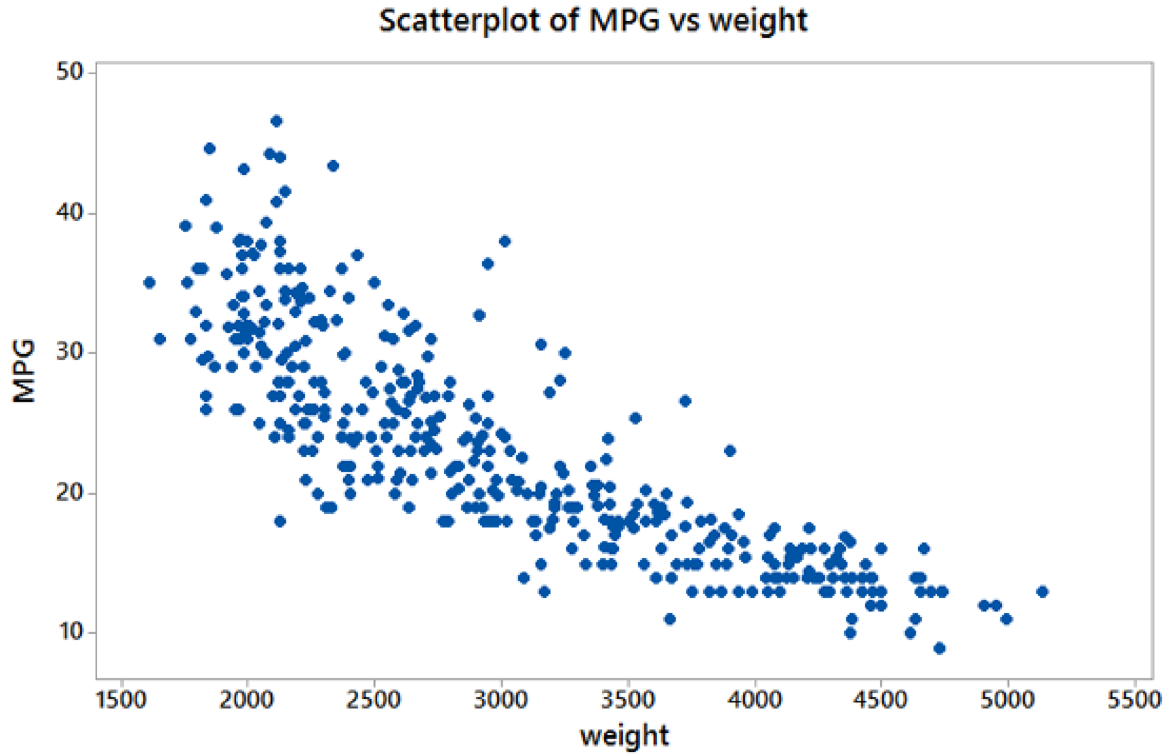


makes it easier to see and compare positions than scatter plots

Scatterplots

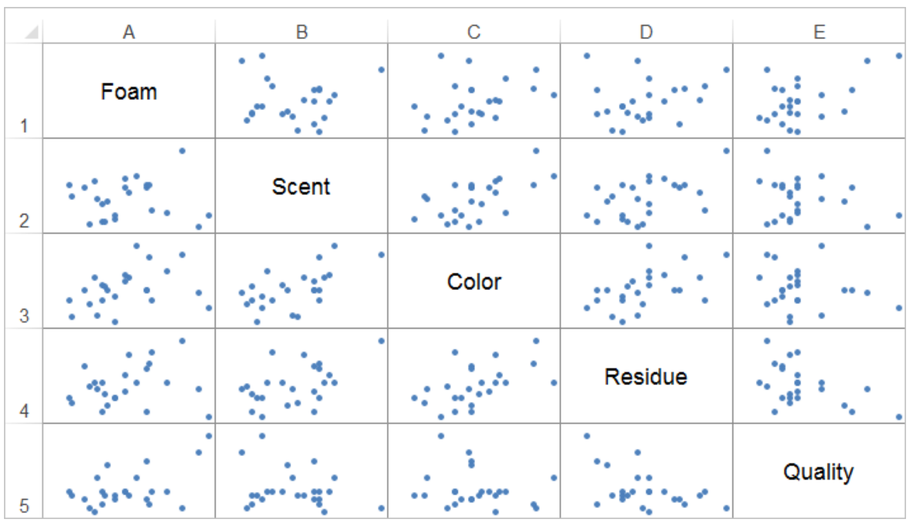
Scatterplots

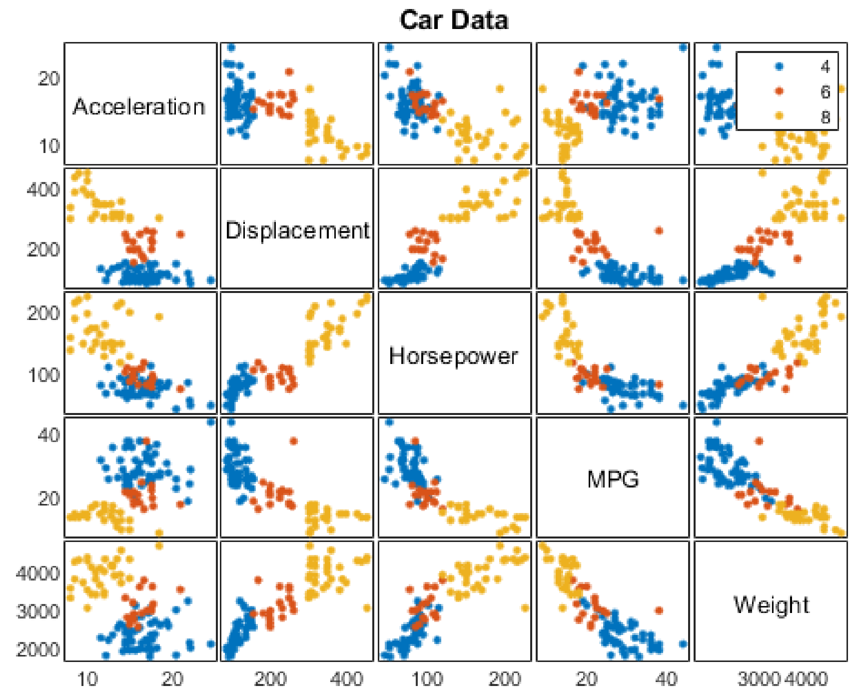
Projection of the data items into a bivariate basis of axes



But what if you have more than two variables?

Scatterplot Matrix





raw data

colored by cluster or class

Problem:

- multivariate relationships are scattered across the tiles
- difficult to see multivariate relationships
- biplots are one way to visualize these – there are others

Bivariate Tile Selection

Scatterplot Matrix: Which Bivariate Tiles to Show?

How many tiles are there?

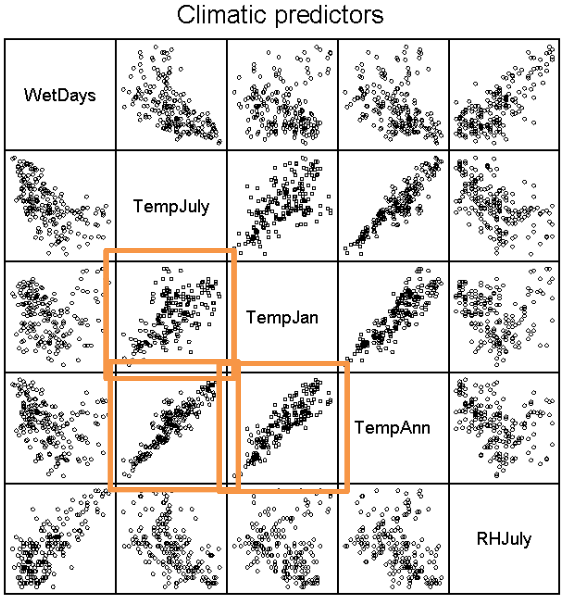
- distributes n(n-1) bivariate relationships over a set of tiles
- for n=4 get 16 tiles
- can use n(n-1)/2 tiles

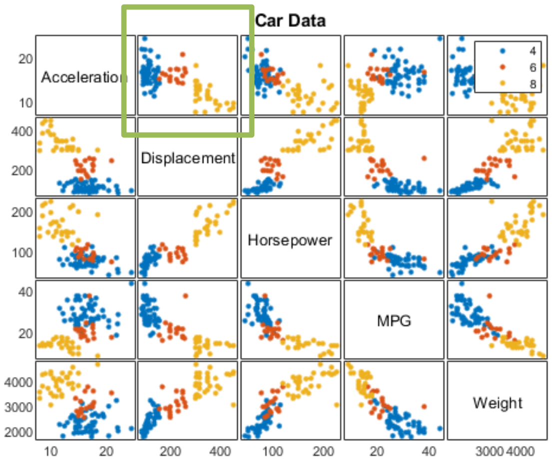
For even moderately large n:

- there will be too many tiles

Which plots to select?

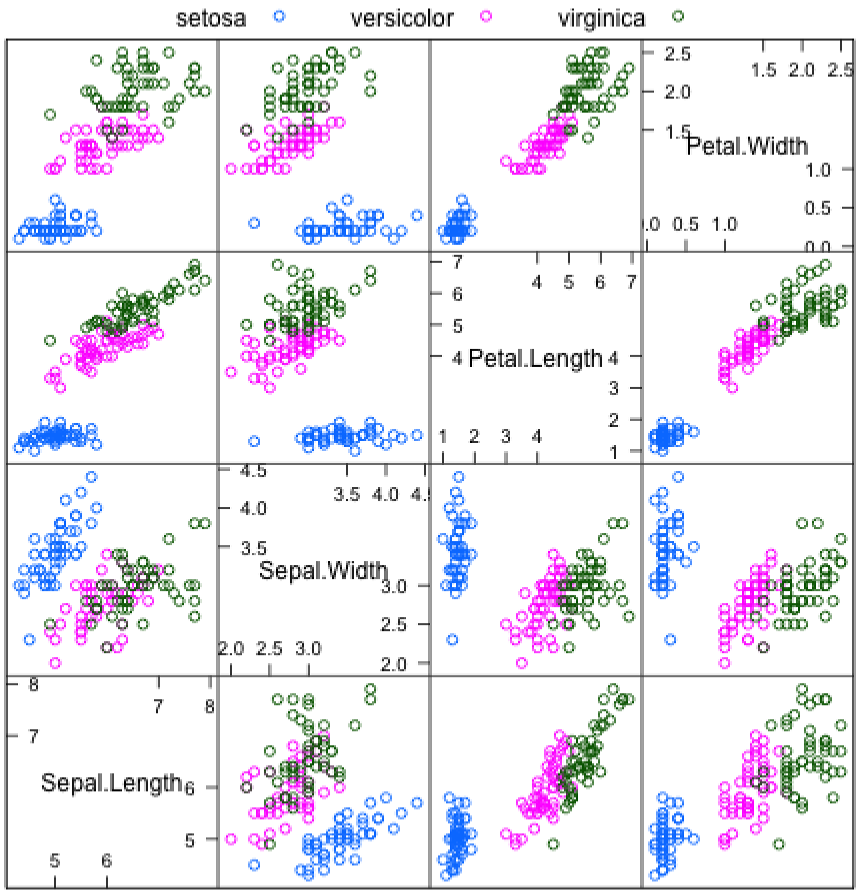
- plots that show correlations well
- plots that separate clusters well





Tile Selection

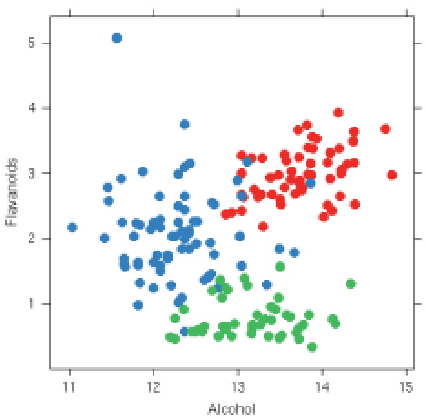
Select the most interesting tiles and show them to the user



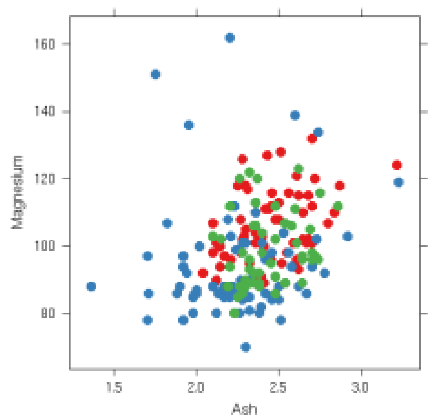
Scatter Plot Matrix

Automated Tile Selection

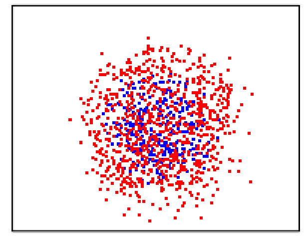
Several metrics, a good one is Distance Consistency (DSC)

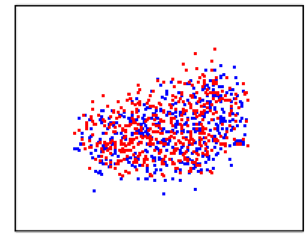


(a) **DSC=90**

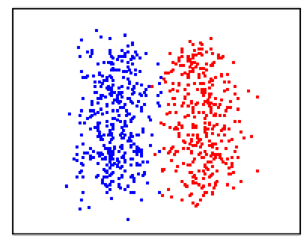


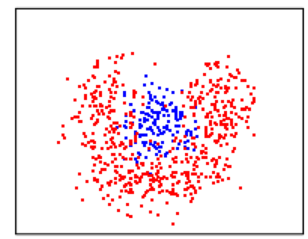
(b) **DSC=49**


(d) 29


(e) 15

bad


(a) 99


(b) 74

OK

$$\mathbf{DSC} = \frac{\left| x' \in v(X) : \mathbf{CD}(x', centr'(c_{clabel(x)}) \; = \; true \right|}{k}$$

- measures how “pure” a cluster is
- rank and pick the views with highest normalized DSC

M. Sips et al., Computer Graphics Forum, 28(3): 831–838, 2009

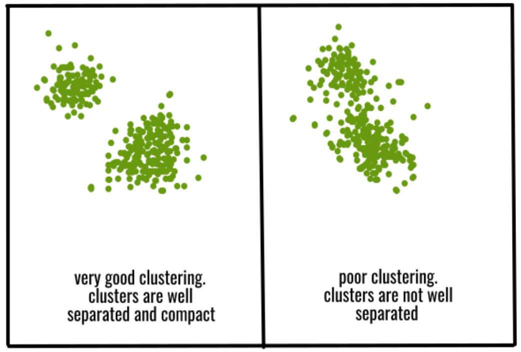
Another Metric: Silhouette Score

Comparison

- DSC → % of correctly assigned points under nearest-centroid rule
- Silhouette score → looks at good margin separability

Compute for each point *i*

- a(i): average distance of *i* to all points in its own cluster
- b(i): lowest average distance of *i* to points in any other cluster
- Overall score S is the average of all *s(i)*

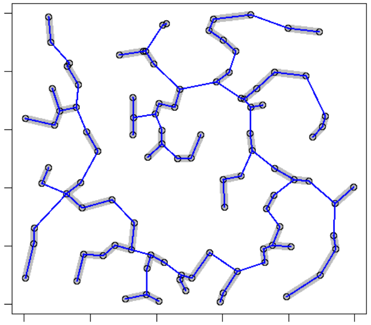


high S   low S

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

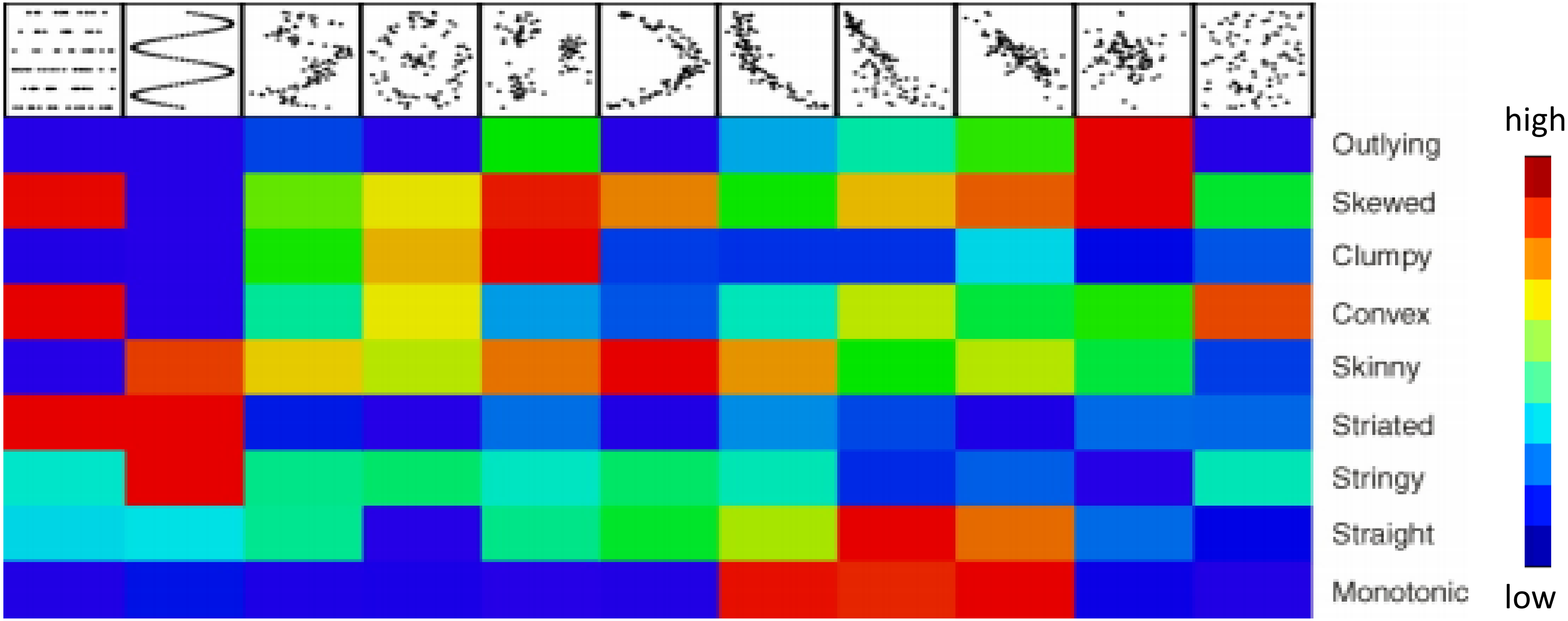
$$S = \frac{1}{N}\sum_{i=1}^{N} s(i)$$

Scagnostics



Describe scatterplot features by graph theoretic measures

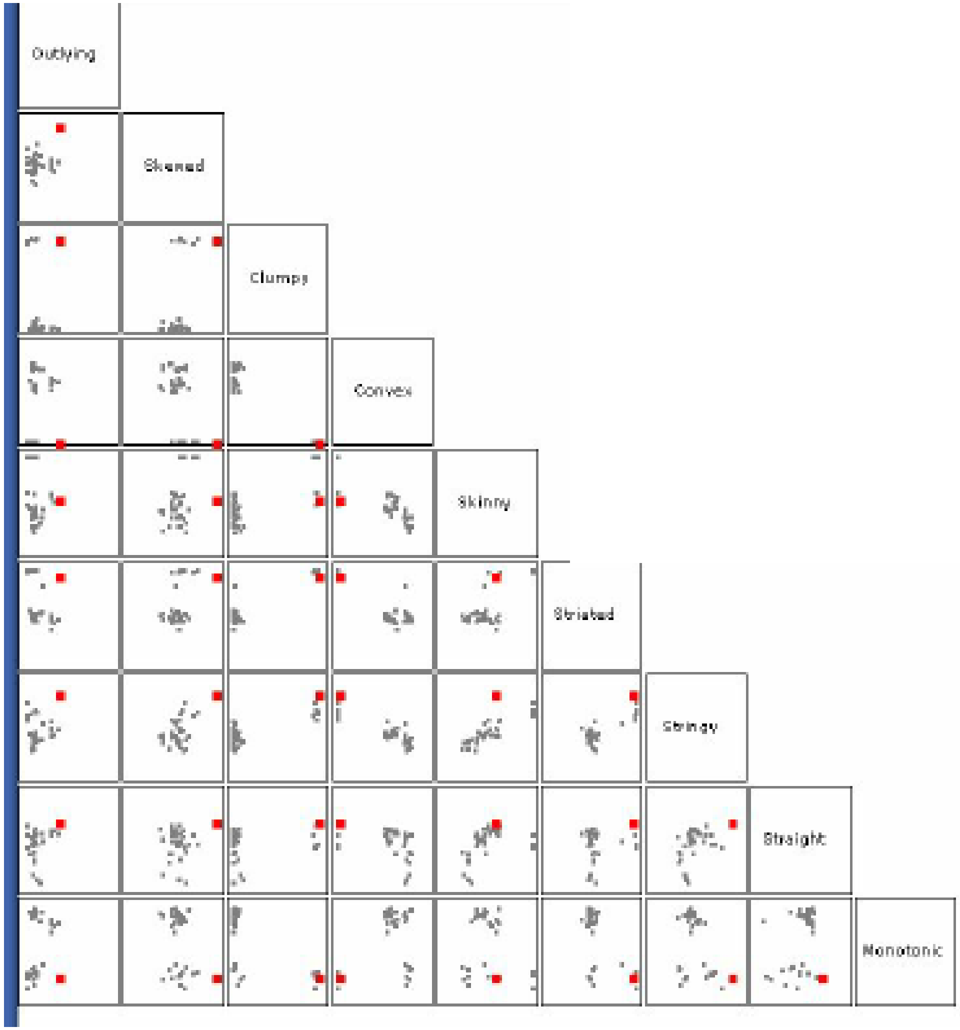
- mostly built on minimum spanning tree
- can be used to summarize large sets of scatterplots



Scatterplot of Scatterplots

Use scagnostics to quickly survey 1,000s of scatterplots

- compute scagnostics measures
- create scatterplot matrix of these measures
- each scatterplot is a point



Biplots

Projection Operations

How does 2D projection work in practice?

- N-dimensional point $x = \{x_1. x_2, x_3, \ldots x_N\}$
- a basis of two orthogonal axis vectors defined in N-D space

$$a = \{a_1. a_2, a_3, \ldots a_N\}$$

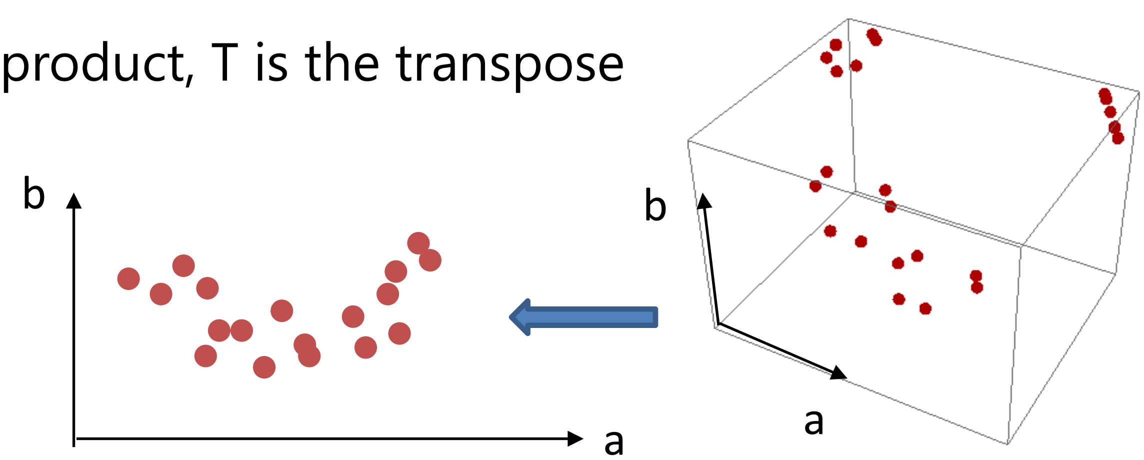
$$b = \{b_1. b_2, b_3, \ldots b_N\}$$

- a projection $\{x_a, x_b\}$ of x into the 2D basis spanned by {a, b} is:

$$x_a = a \cdot x^T$$

$$x_b = b \cdot x^T$$

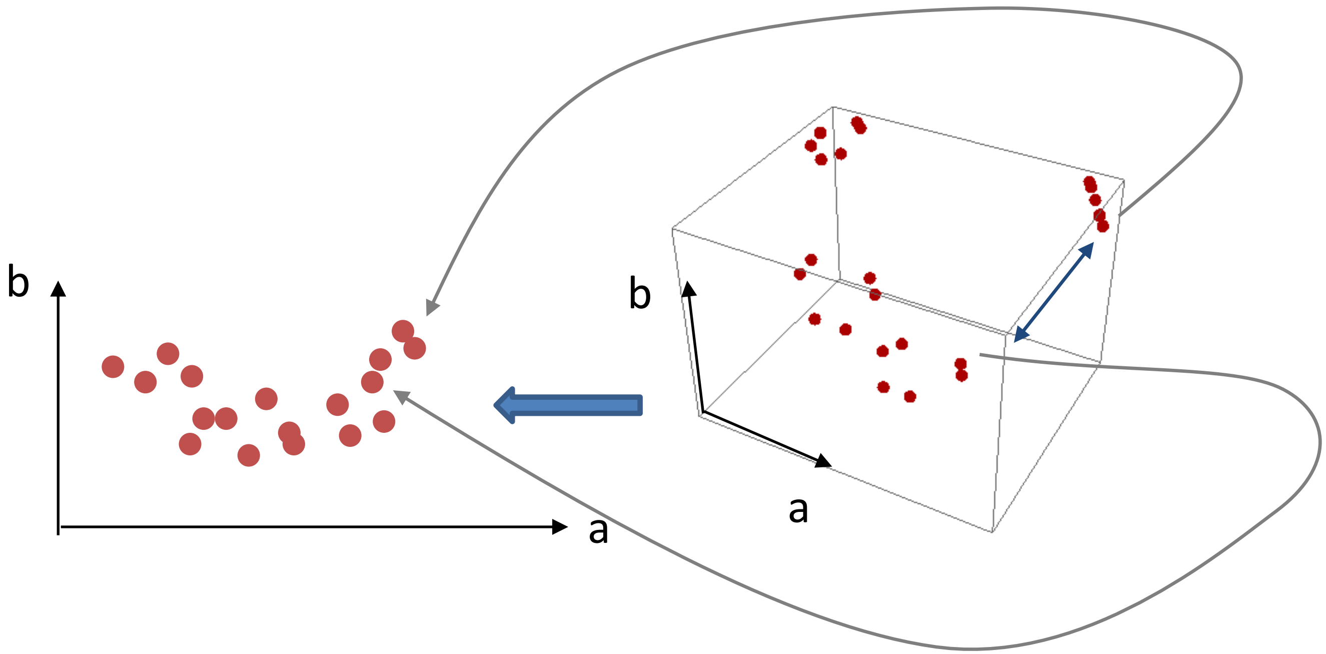
where · is the dot product, T is the transpose



Projection Ambiguity

Projection causes inaccuracies

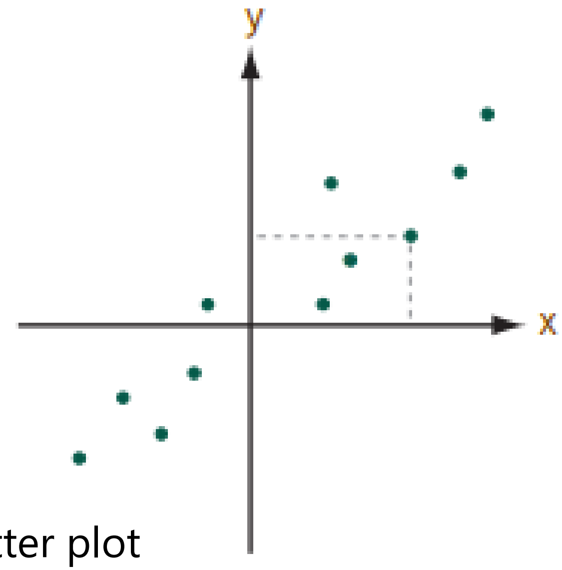
- close neighbors in the projections may not be close neighbors in the original higher-dimensional space
- this is called *projection ambiguity*



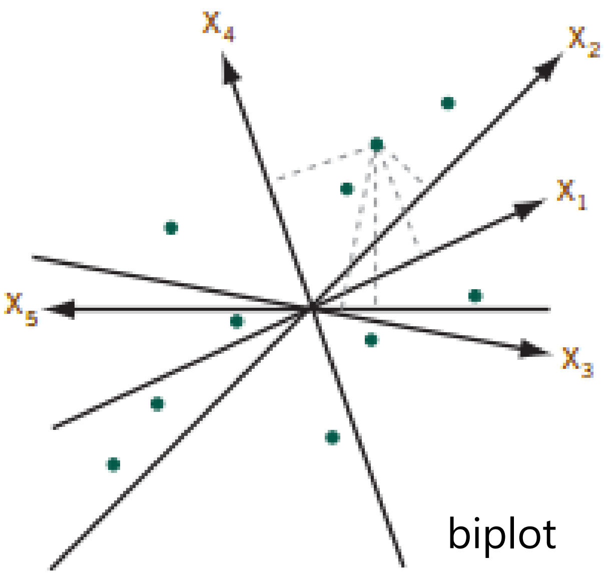
Biplots

Plots data points and dimension axes into a single visualization

- uses first two PCA vectors as the basis to project into
- find plot coordinates [x] [y]
  - for data points: [$PCA_1 \cdot$ data vector] [$PCA_2 \cdot$ data vector]
  - for dimension axes: [PCA_1[dimension]] [PCA_2[dimension]]



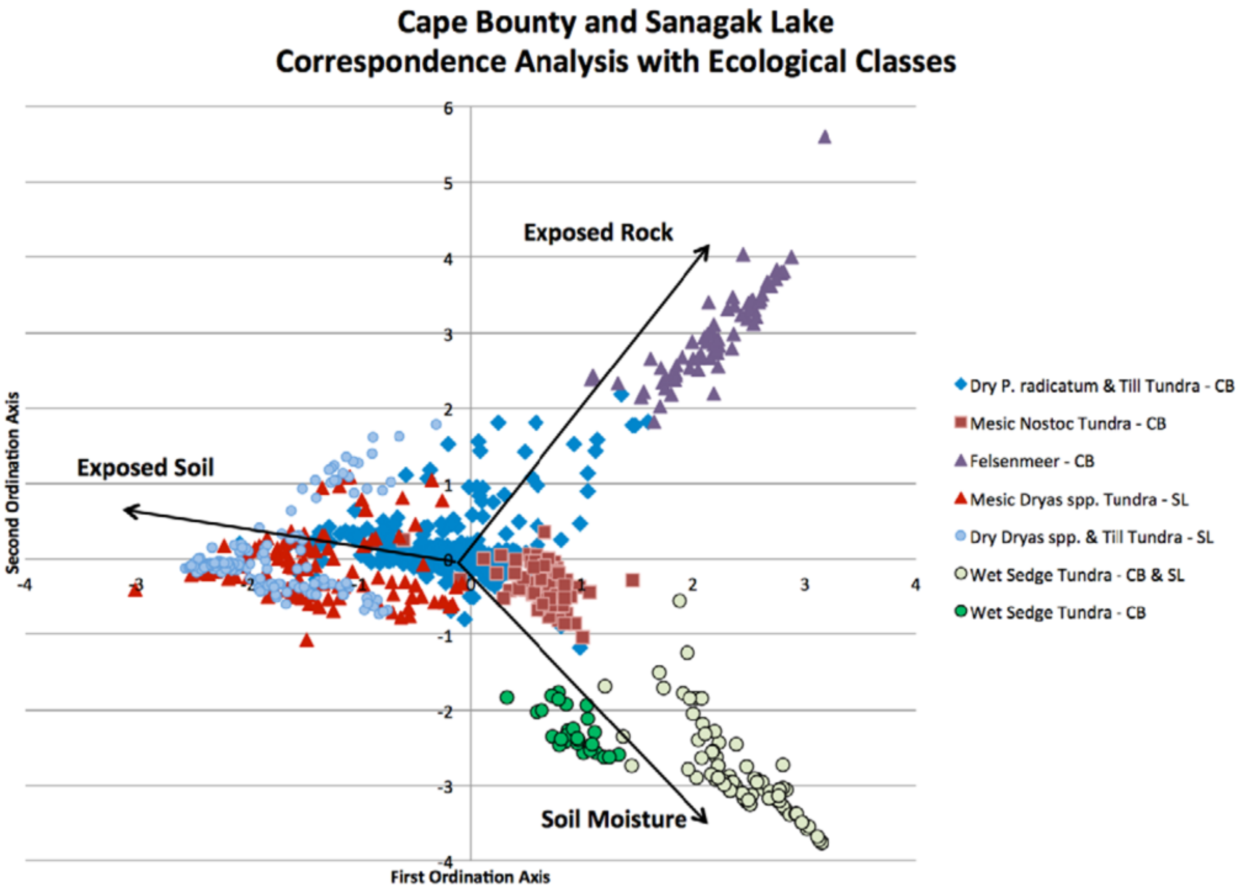
scatter plot



biplot

Biplots Can Have Projection Ambiguities

Are just a linear projection into the 2D basis generated by PCA



Biplots – A Word of caution

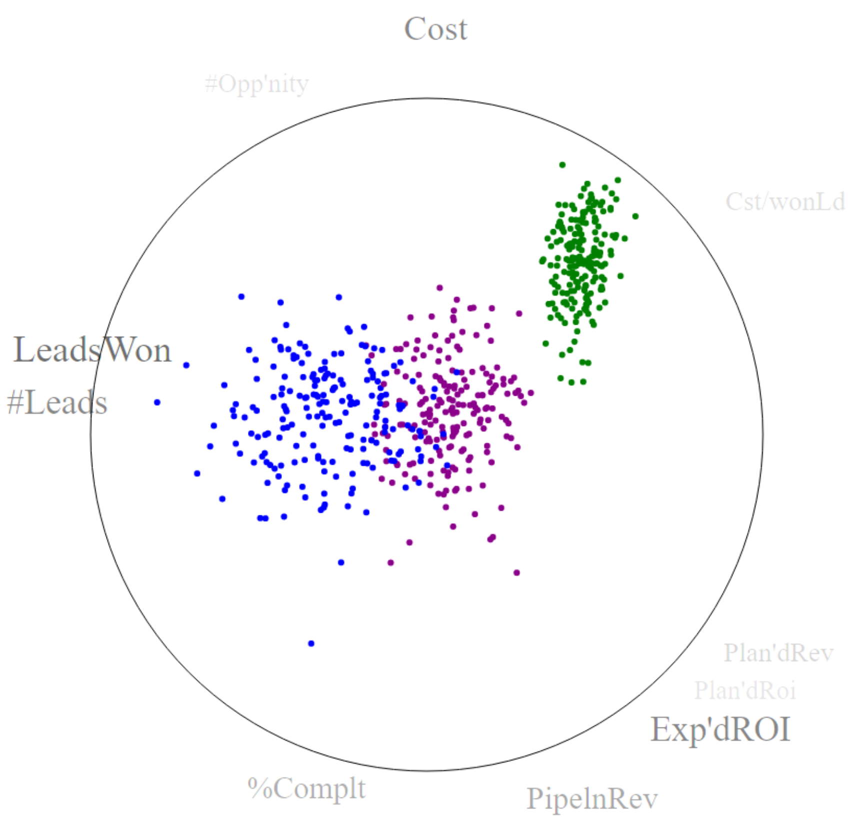
Do be aware that the projections may not be fully accurate

- you are projecting N-D into 2D by a linear transformation
- if there are more than 2 significant PCA vectors then some variability will be lost and won't be visualized
- remote data points might project into nearby plot locations suggesting false relationships → projection ambiguity
- always check out the PCA scree plot to gauge accuracy

Interactive Biplots

Also called multivariate scatterplot

- biplot-axes length vis replaced by graphical design
- less cluttered view
- but there's more to this .....



B. Wang, K. Mueller, "The Subspace Voyager: Exploring High-Dimensional Data along a Continuum of Salient 3D Subspaces," *IEEE TVCG, 2018*

Meet the *Subspace Voyager*

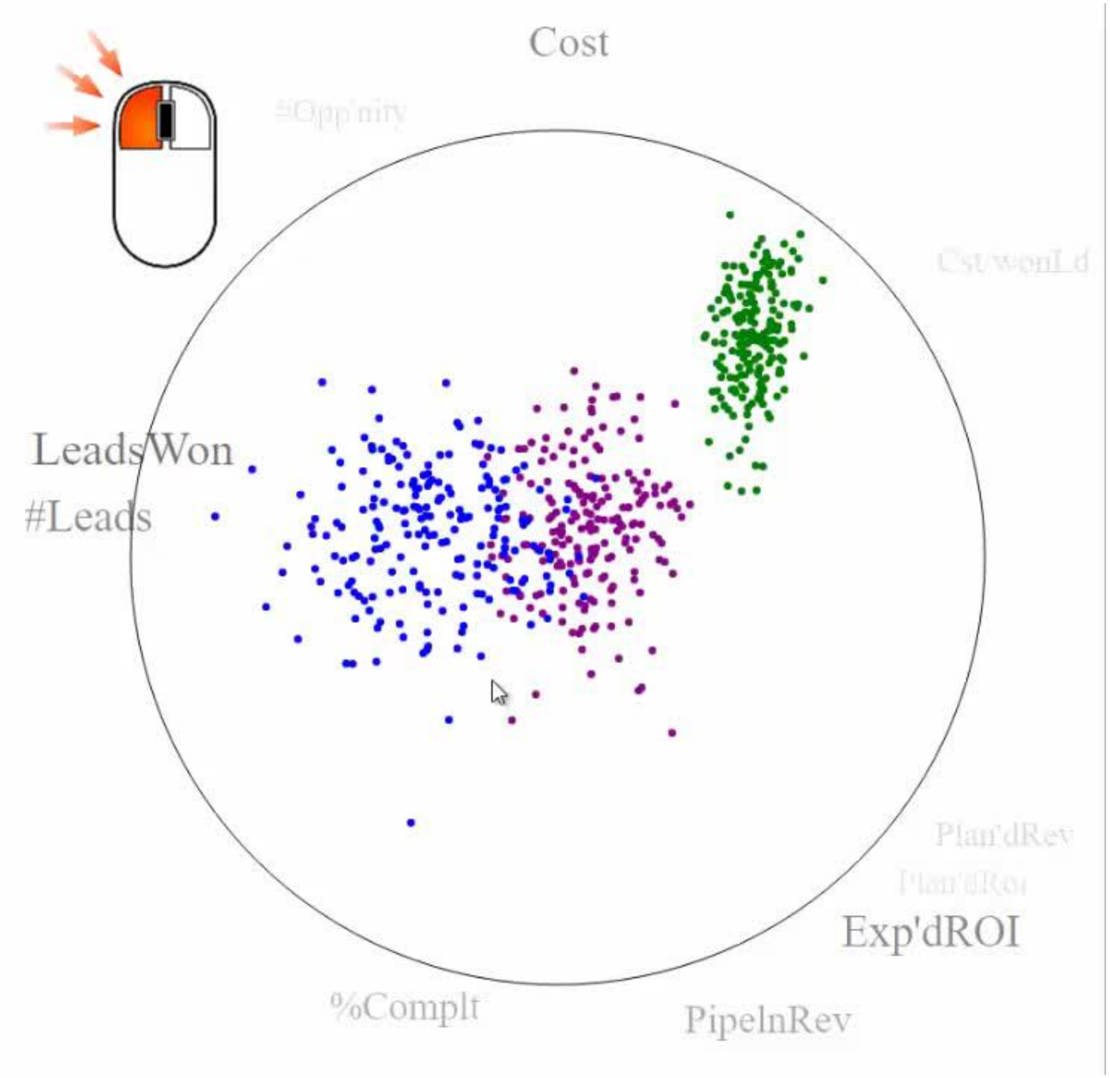
Decomposes high-D data spaces into lower-D subspaces by

- clustering
- classification
- reducing clusters to intrinsic dimensionality via local PCA

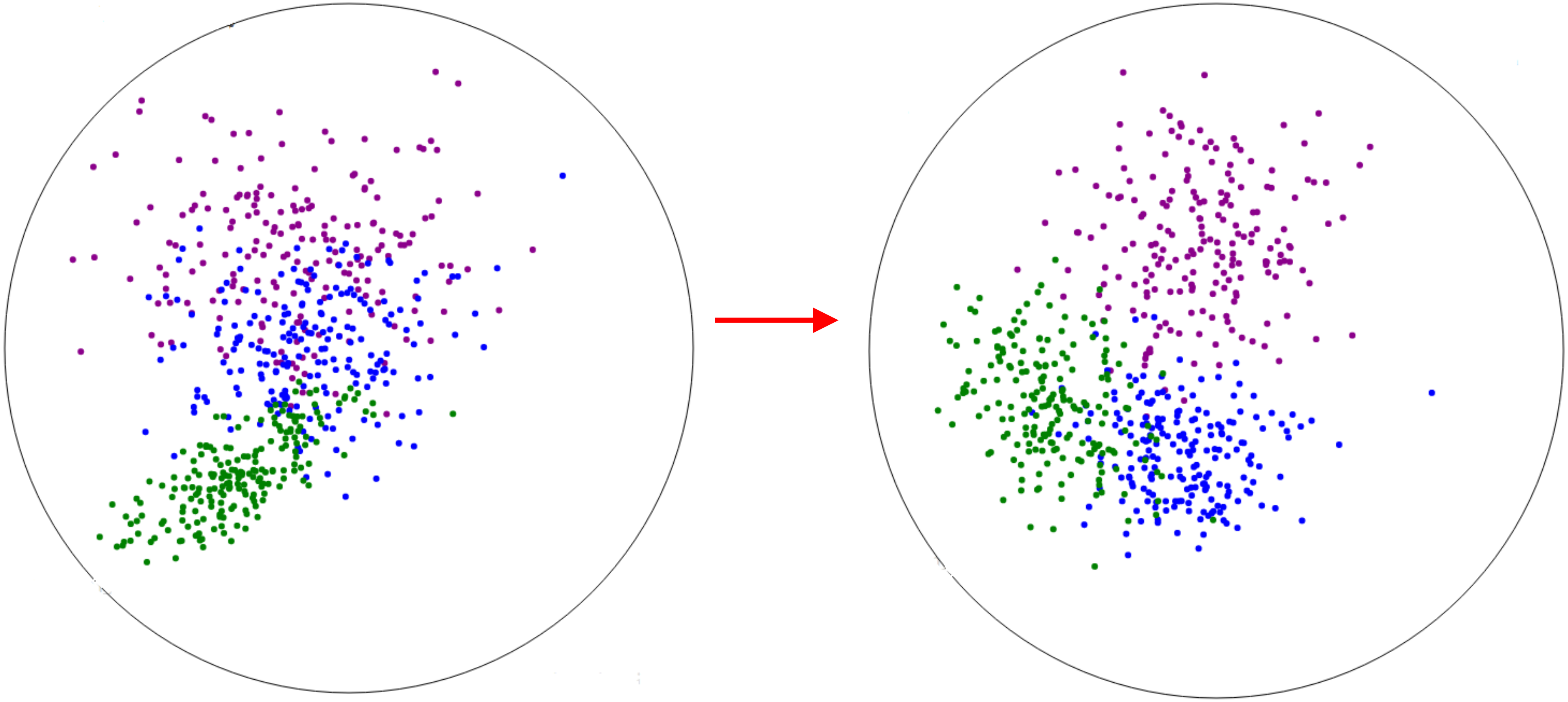
Allows users to interactively explore these lower-D subspaces

- explore them as a chain of 3D subspaces
- transition seamlessly to adjacent 3D subspaces on demand
- save observations as you go (and return to them just as well)

Trackball-Based Cluster Exploration



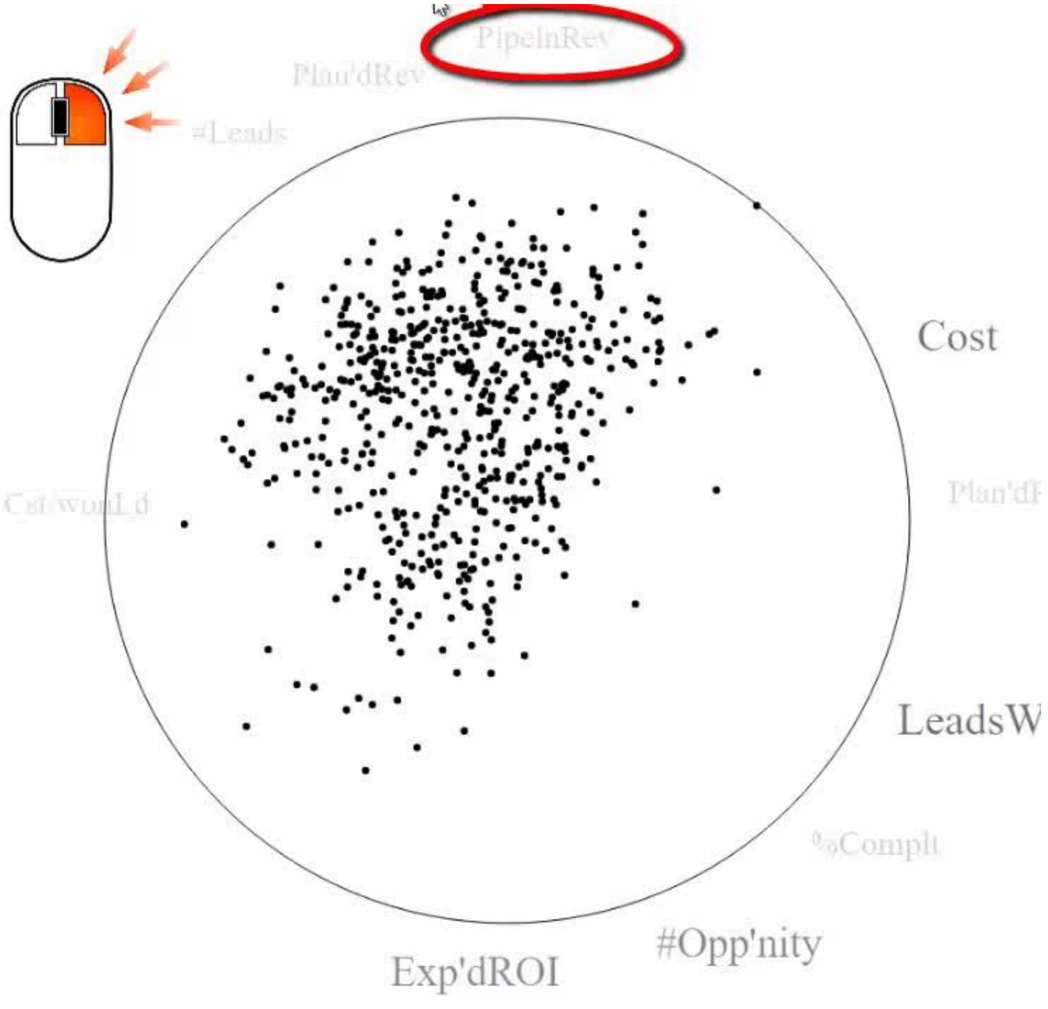
Interactive View Optimizer



Uses genetic-algorithm driven projection pursuit

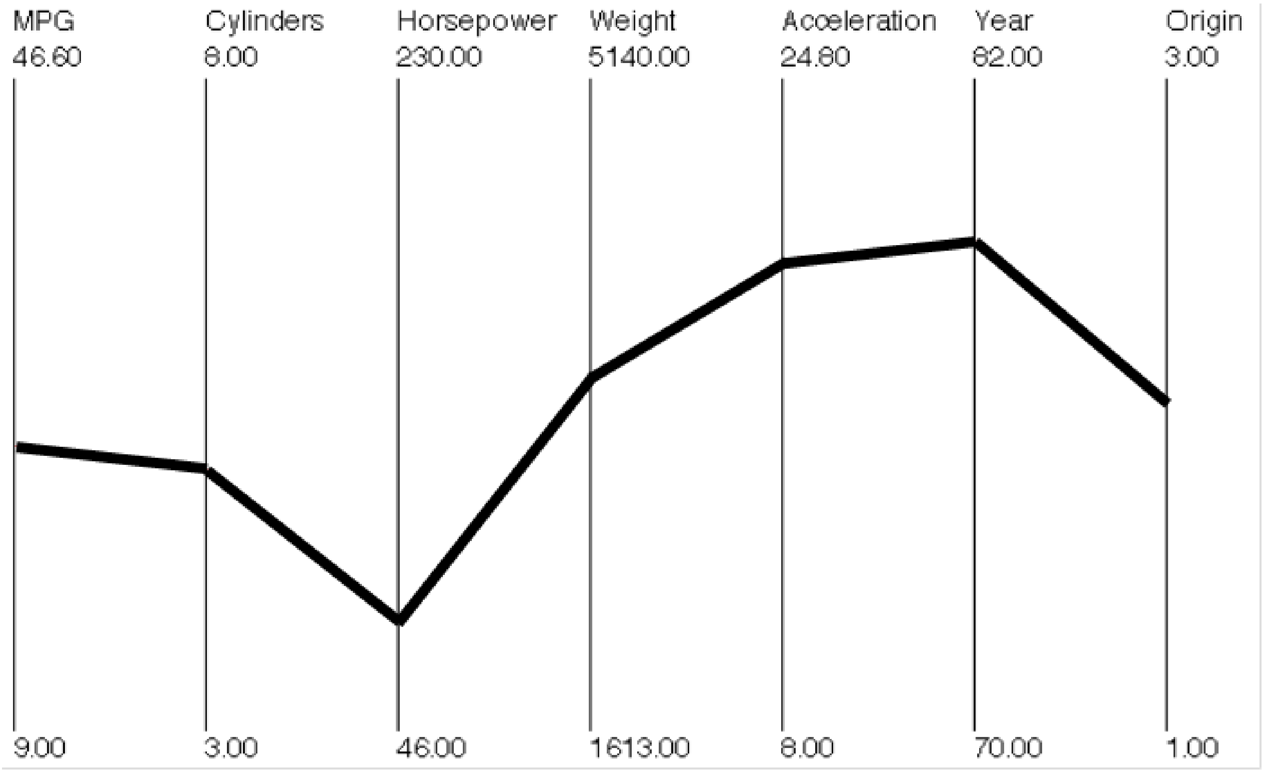
Several view quality metrics are available

Chase Interesting Clusters – Transition to Adjacent 3D Subspaces



Parallel Coordinates

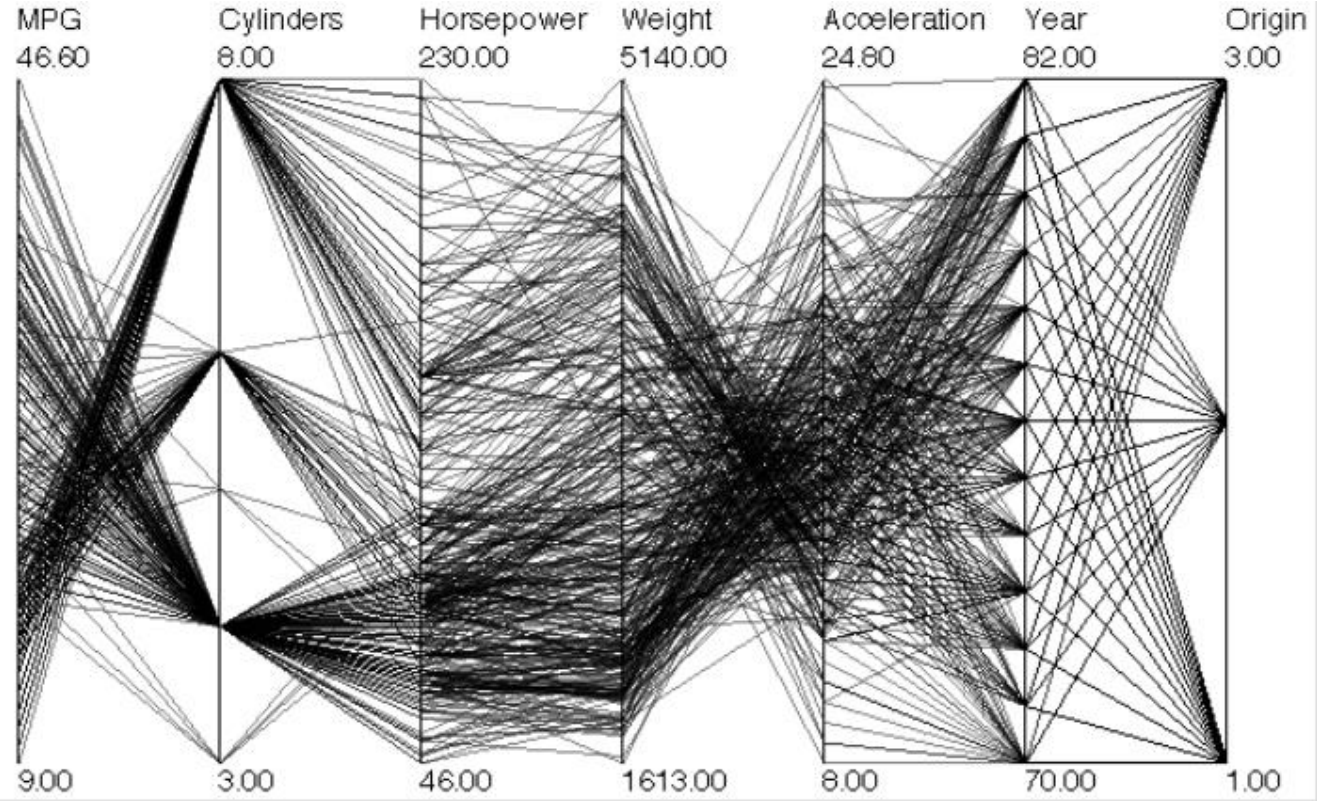
Parallel Coordinates – 1 Car



The N=7 data axes are arranged side by side

- in parallel

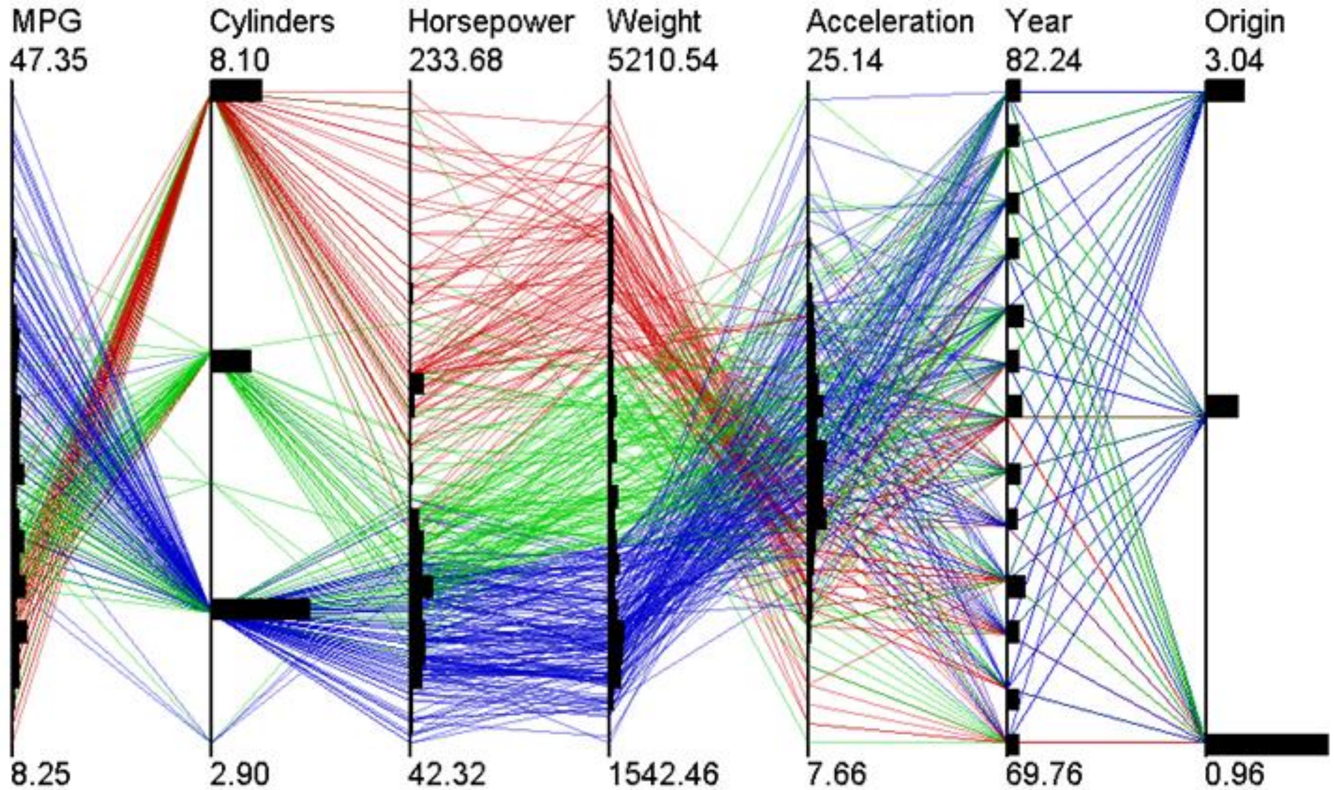
Parallel Coordinates – 100 Cars



Hard to see the individual cars?

- what can we do?

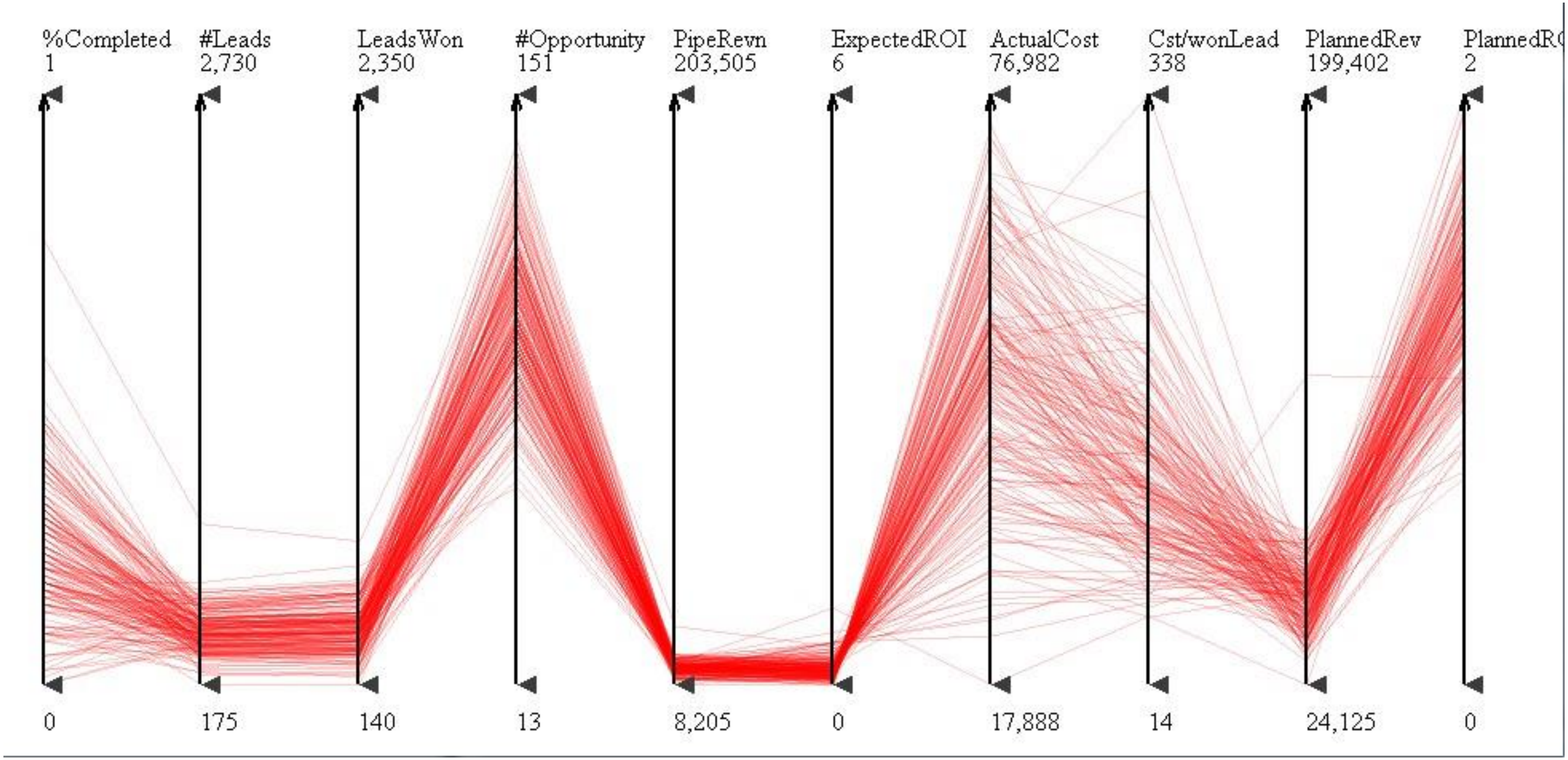
Parallel Coordinates – 100 Cars



Grouping the cars into sub-populations

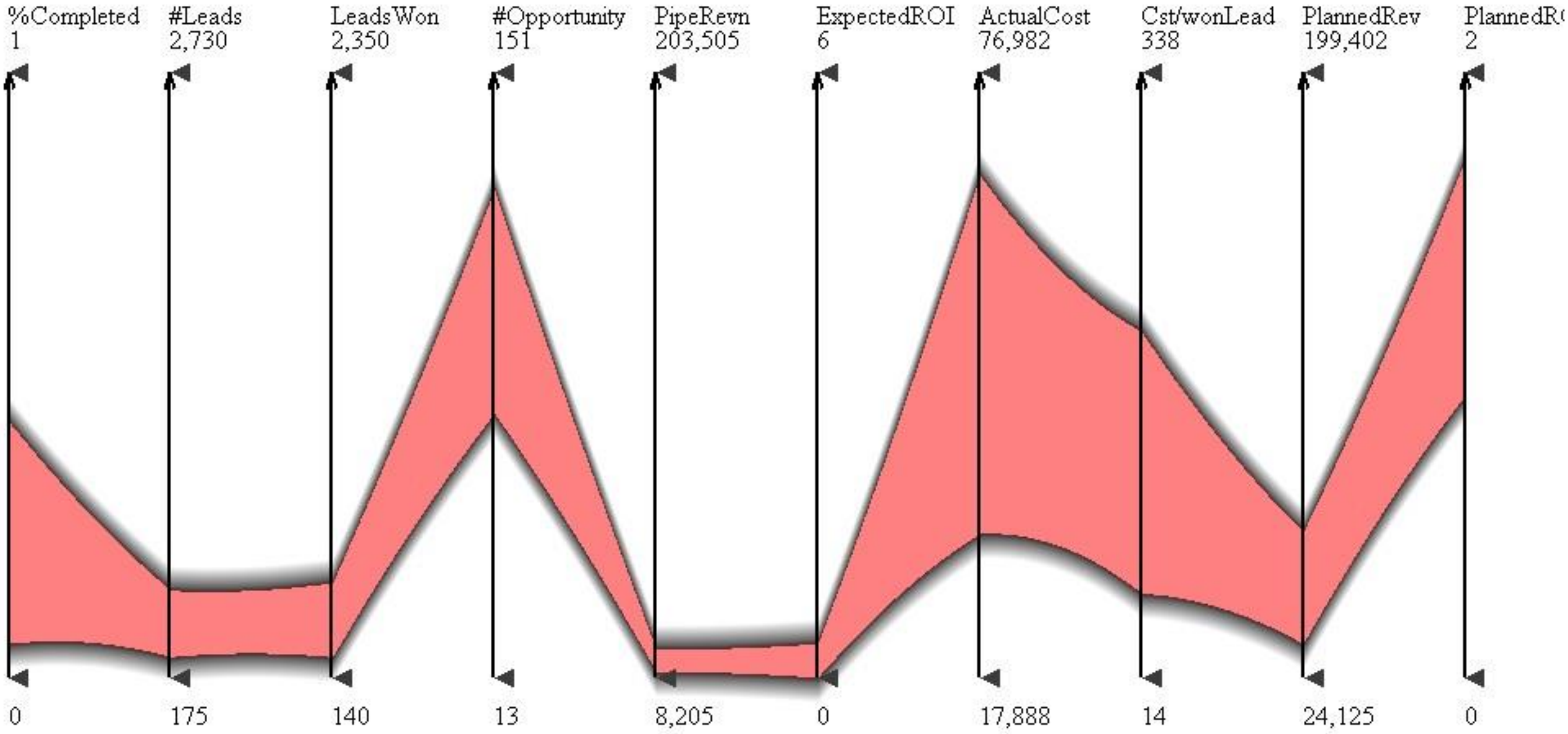
- we perform clustering
- an be automated or interactive (put the user in charge)

PC With Illustrative Abstraction



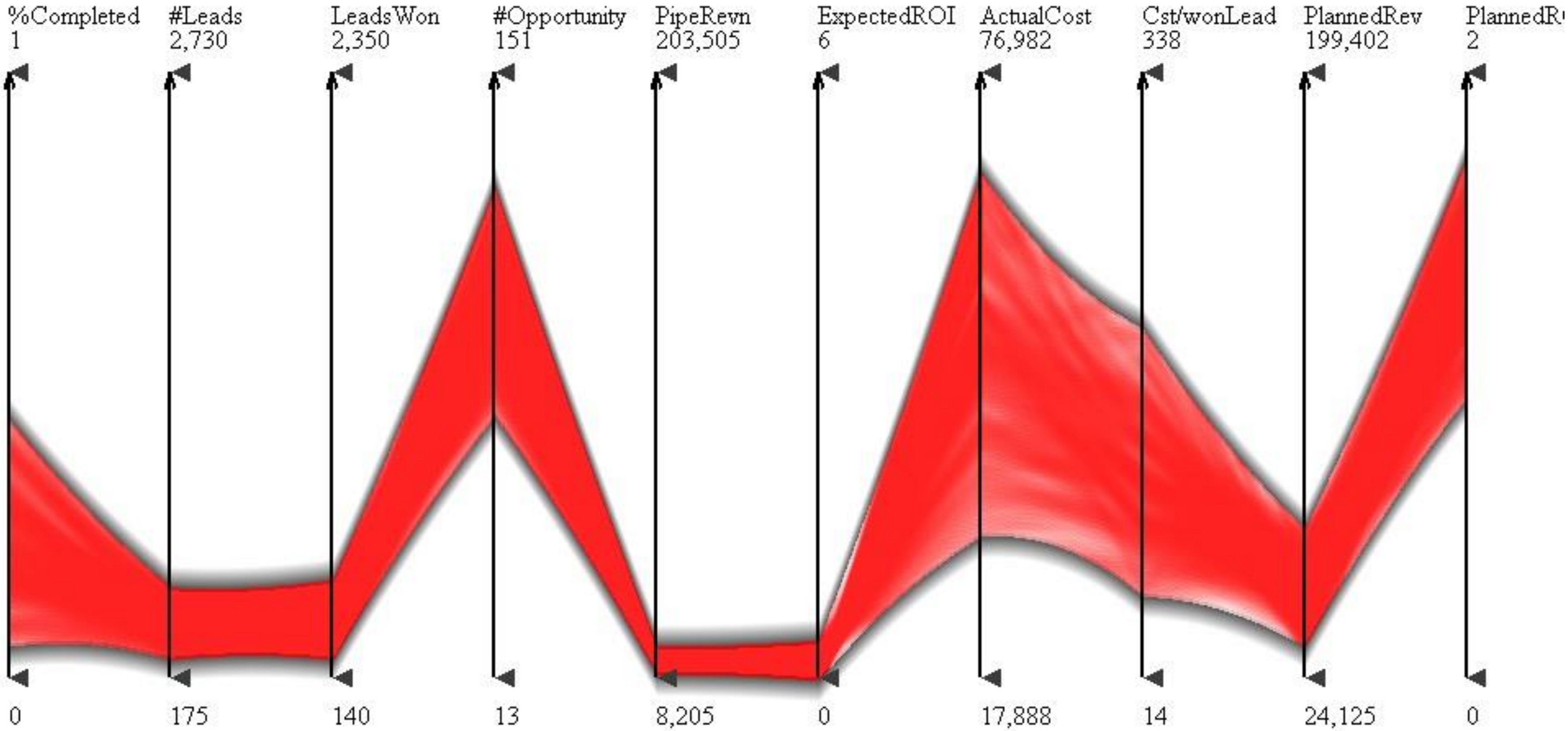
individual polylines

PC With Illustrative Abstraction



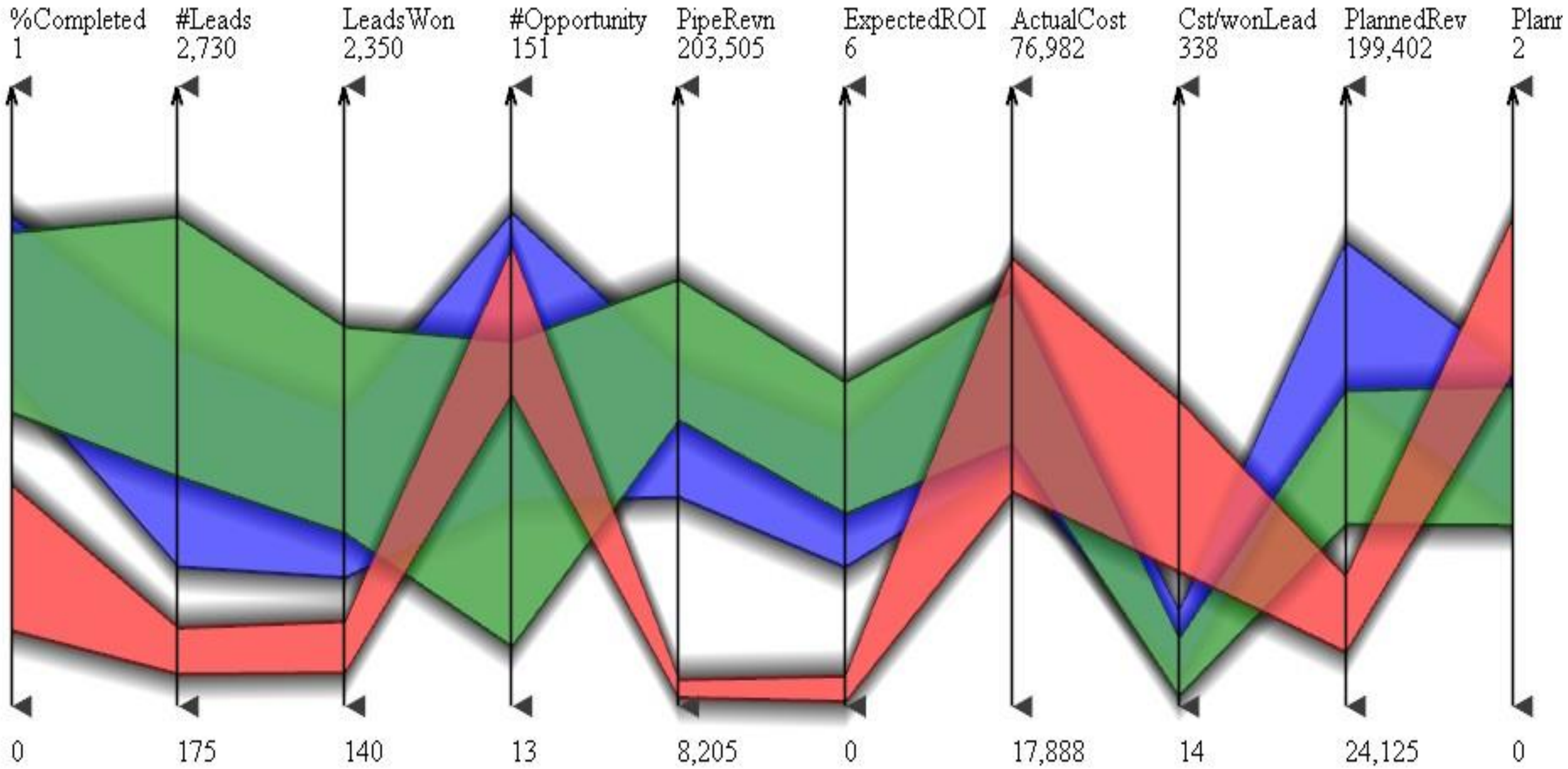
completely abstracted away

PC With Illustrative Abstraction



blended partially

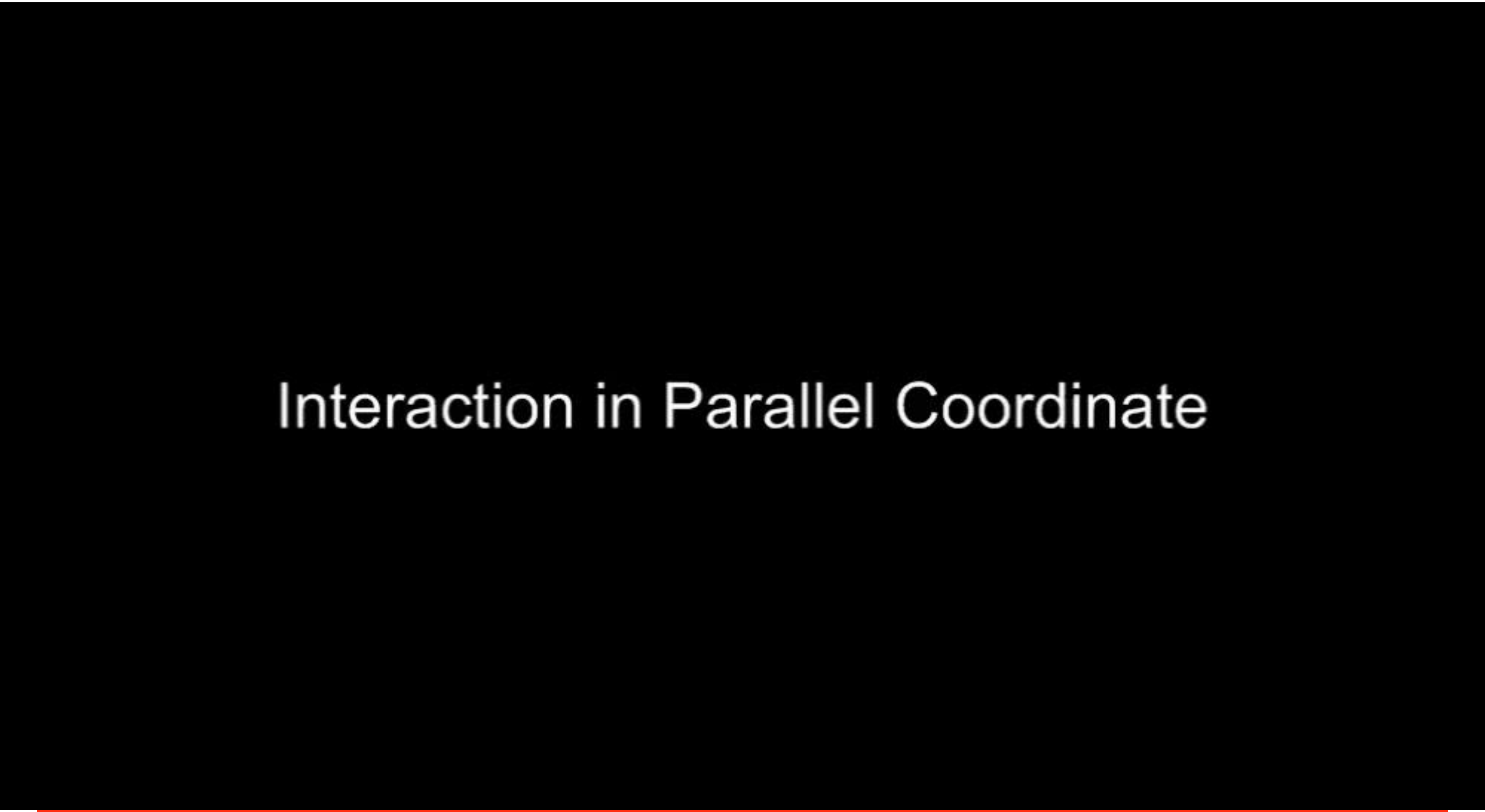
PC With Illustrative Abstraction



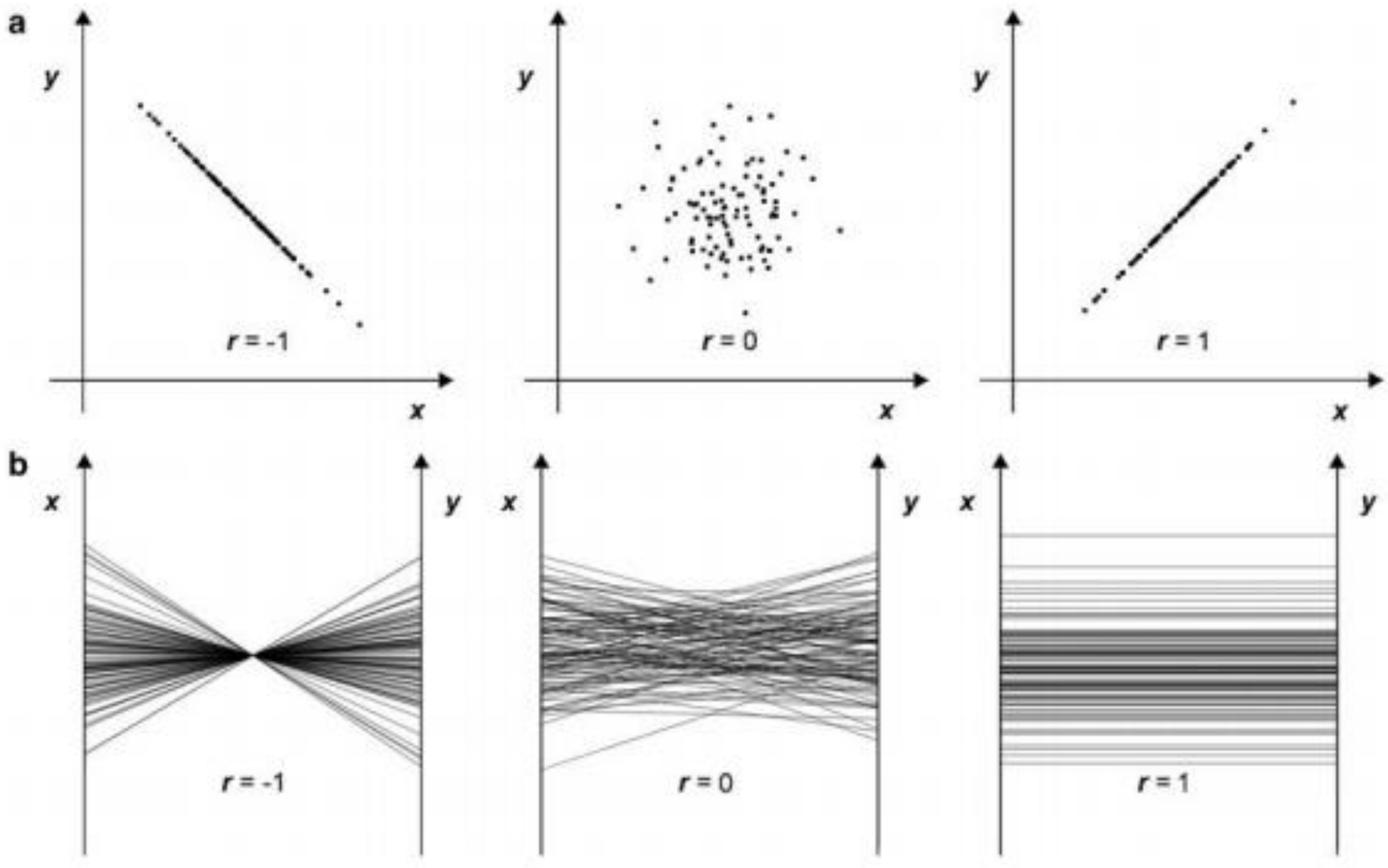
all put together – three clusters

[McDonnell and Mueller, 2008]

Interaction is Key



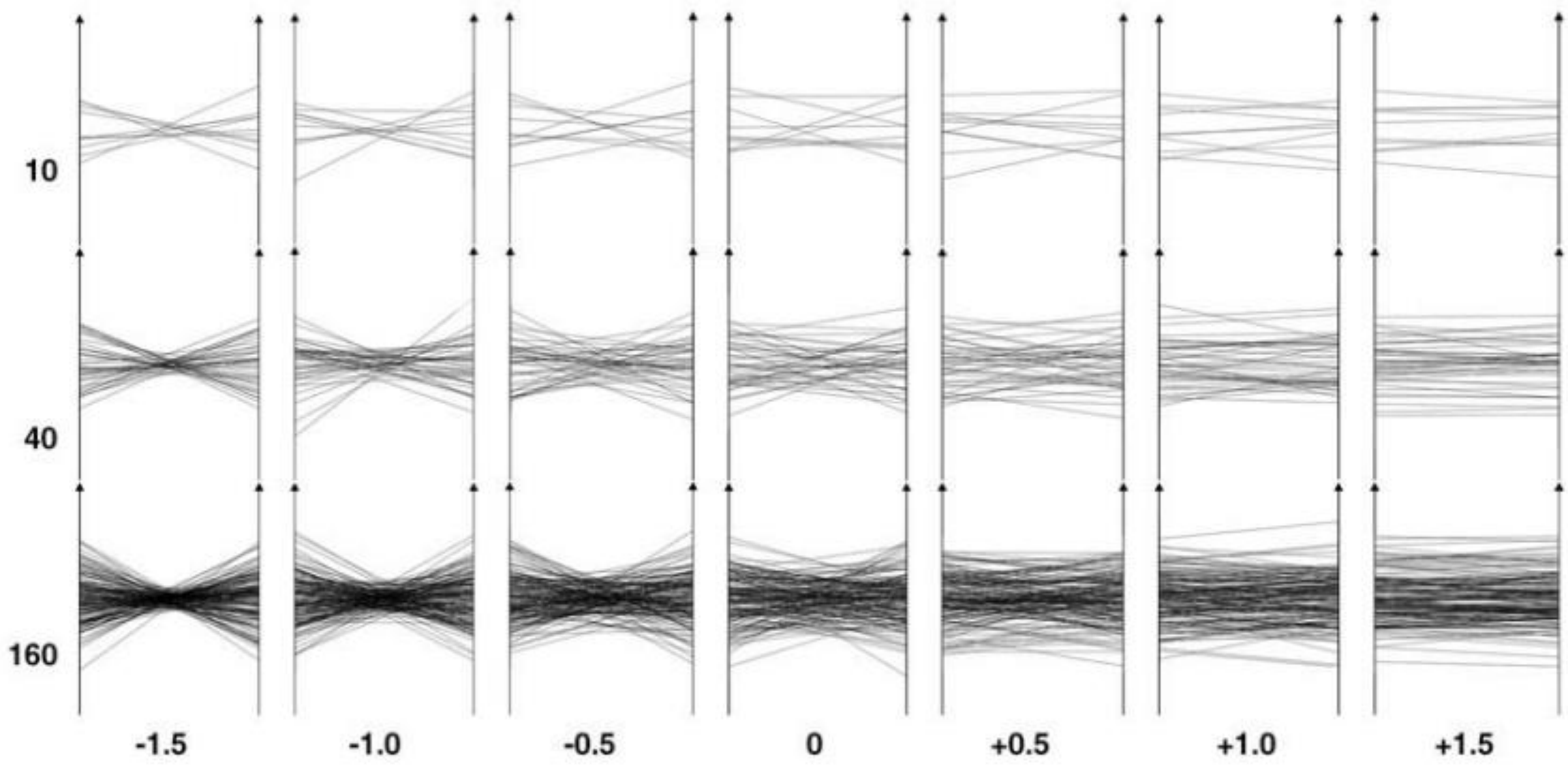
Patterns in Parallel Coordinates



correlation r=-1.0 r=0 r=1.0

Patterns in Parallel Coordinates

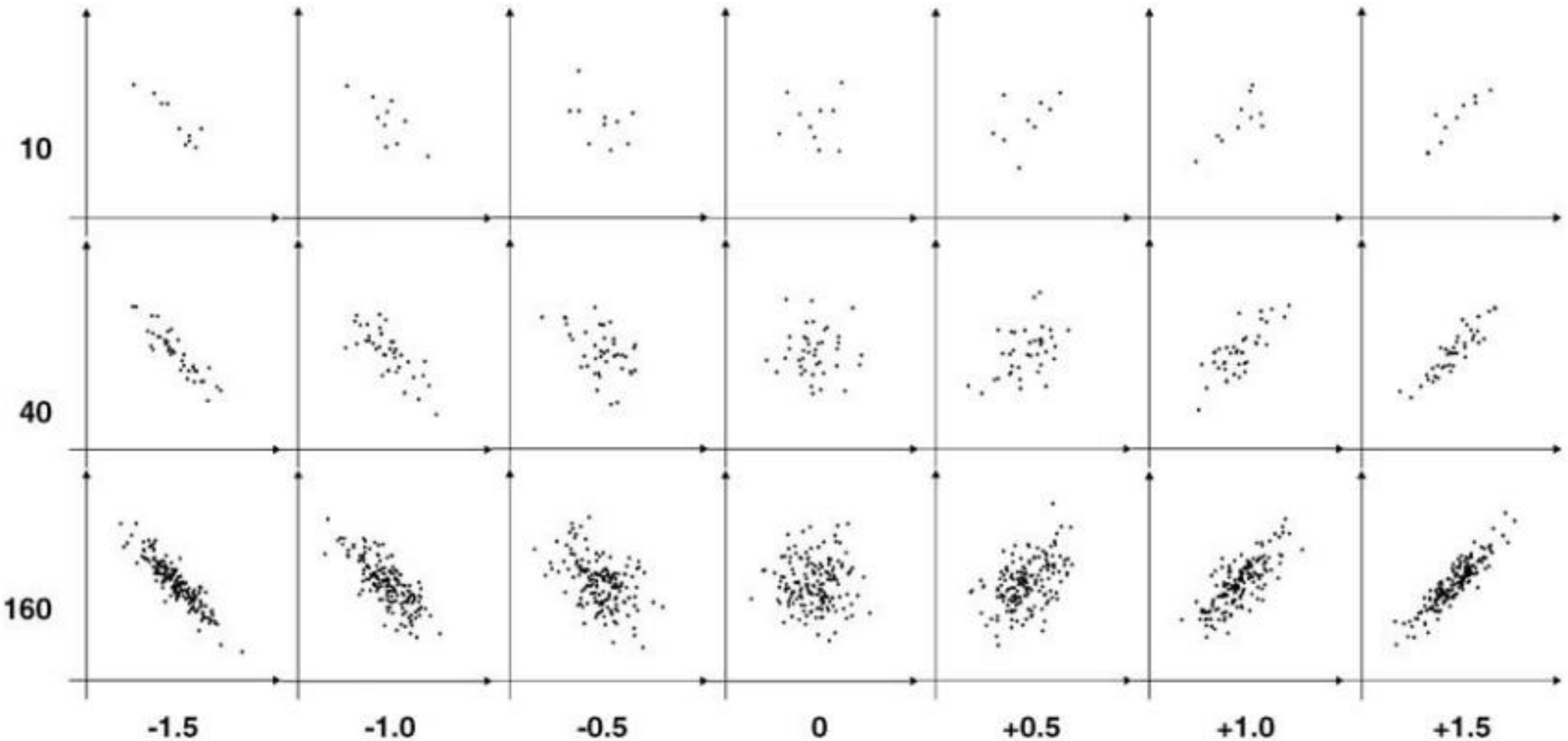
points



Fisher-z (corresponding to $\rho = 0, \pm 0.462, \pm 0.762, \pm 0.905$)

Patterns in Scatterplots

points



Fisher-z (corresponding to $\rho = 0, \pm 0.462, \pm 0.762, \pm 0.905$)

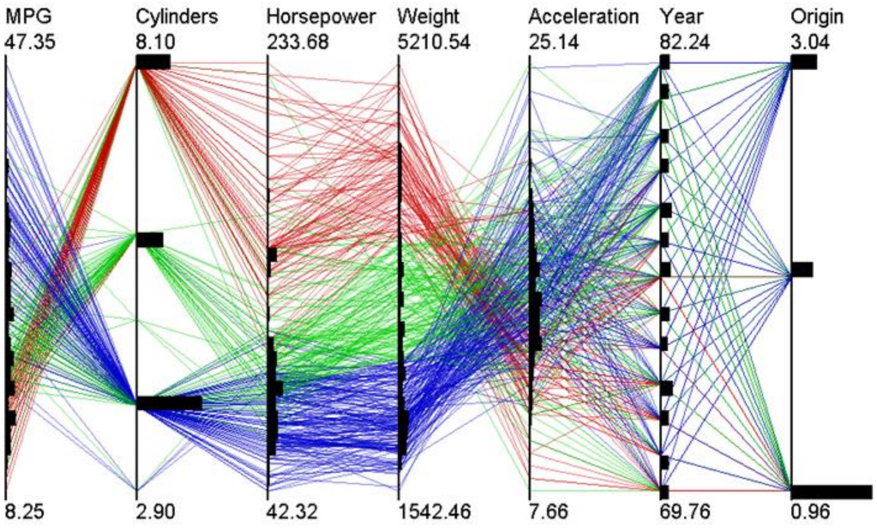
Li et al. found that <u>twice as many</u> correlation levels can be distinguished with scatterplots

Information Visualization Vol. 9, 1, 13 – 30

Axis Reordering Problem

There are n! ways to order the n dimensions

- how many orderings for 7 dimensions?
- 5,040
- but since can see relationships across 3 axes a better estimate is $n!/((n-3)!\,3!) = 35$
- still a lot of axes orderings to try out → we need help

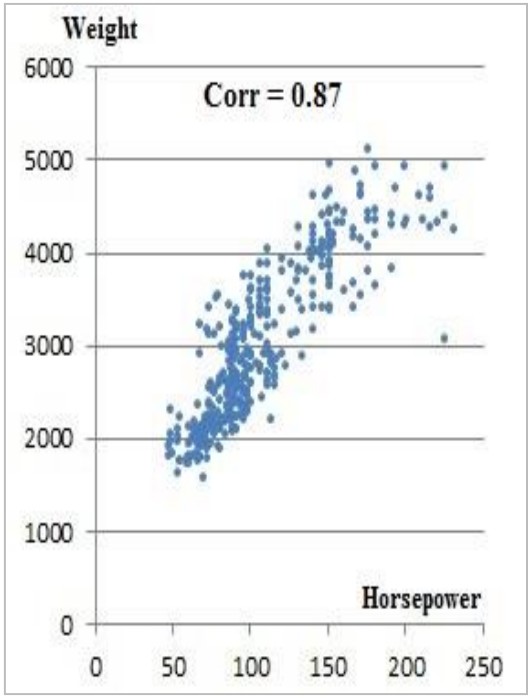


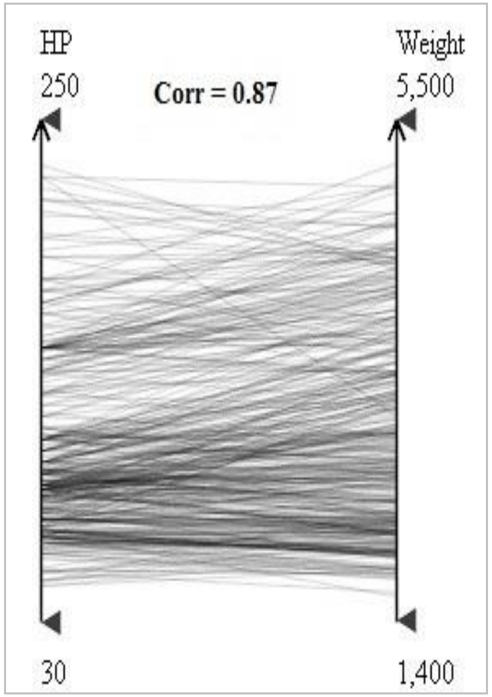
We Need a Measure for Relationships

Correlation

- a statistical measure that indicates the extent to which two or more variables fluctuate together

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$





Building the Correlation Matrix

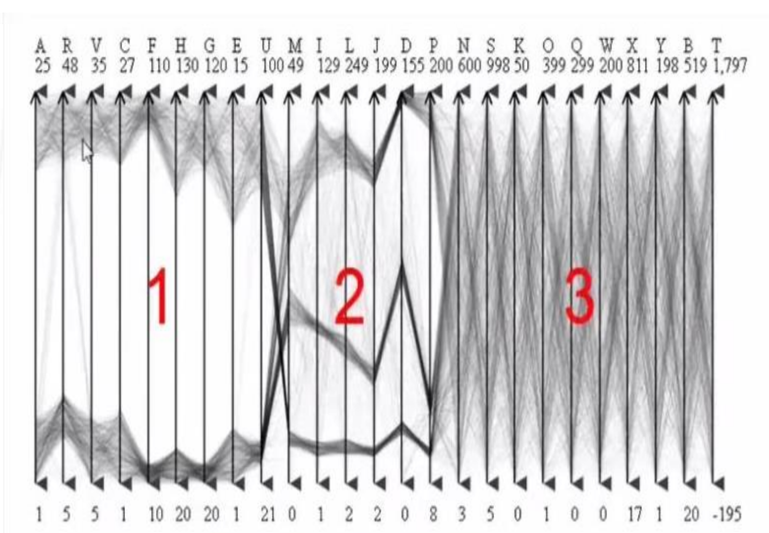
Z. Zhang, K. McDonnell, K. Mueller, "A Network-Based Interface for the Exploration of High-Dimensional Data Spaces, " *IEEE Pacific Vis*, 2012

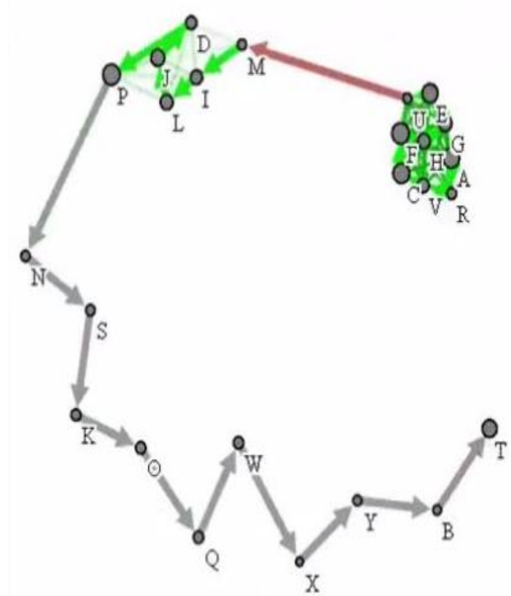
Create a correlation matrix

Run a mass-spring model

Run Traveling Salesman on the correlation nodes

Use it to order your parallel coordinate axes via TSP

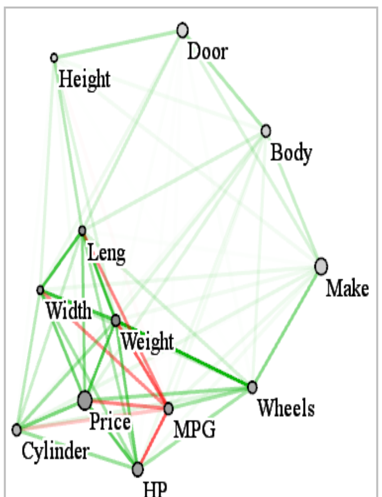




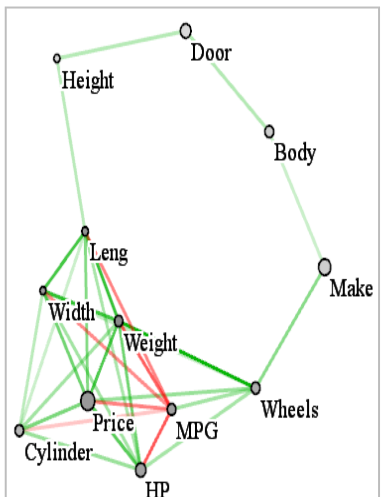
	MRK	MSFT	PFE	PG	*T*	TRV	UTX	VZ	WMT	XOM
MRK	1.	0.39	0.72	−0.43	0.57	0.031	−0.26	0.61	−0.11	−0.25
MSFT	0.39	1.	0.14	0.11	0.56	0.25	0.25	0.67	−0.074	0.24
PFE	0.72	0.14	1.	−0.77	0.08	−0.37	−0.65	0.19	−0.077	−0.72
PG	−0.43	0.11	−0.77	1.	0.25	0.68	0.92	0.086	0.072	0.9
T	0.57	0.56	0.08	0.25	1.	0.65	0.46	0.87	−0.059	0.54
TRV	0.031	0.25	−0.37	0.68	0.65	1.	0.83	0.43	−0.0067	0.81
UTX	−0.26	0.25	−0.65	0.92	0.46	0.83	1.	0.27	−0.033	0.93
VZ	0.61	0.67	0.19	0.086	0.87	0.43	0.27	1.	0.026	0.36
WMT	−0.11	−0.074	−0.077	0.072	−0.059	−0.0067	−0.033	0.026	1.	0.032
XOM	−0.25	0.24	−0.72	0.9	0.54	0.81	0.93	0.36	0.032	1.

Interaction with the Correlation Network

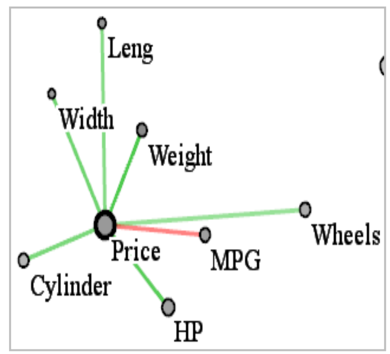
- Vertices are attributes, edges are correlations
  - vertex: size determined by $\sum_{j=0}^{D} \frac{|correlation(i,j)|}{D-1} \; j \neq i$
  - edge length is a measure of (1-|correlation|)
  - edge: color/intensity → sign/strength of correlation



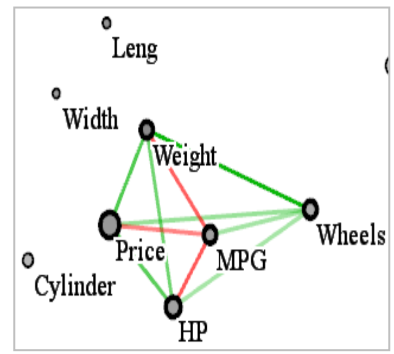
all edges



filtered by strength

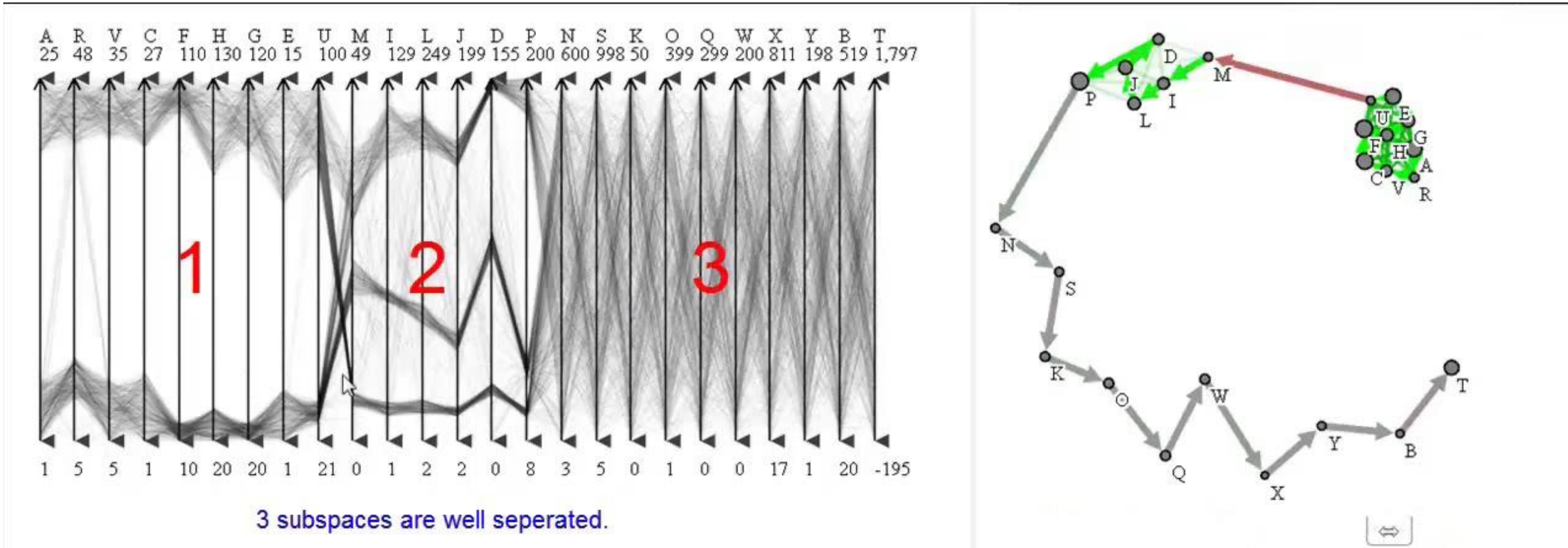


attribute centric



subset of attributes

Multiscale Zooming

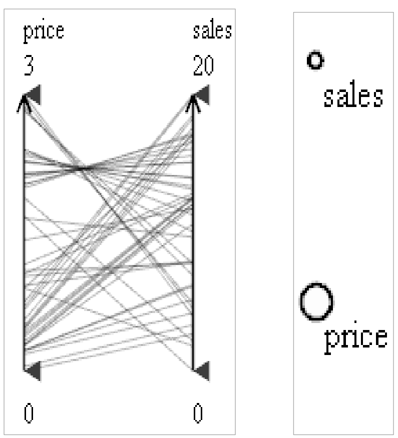


Z. Zhang, K. McDonnell, K. Mueller, "A Network-Based Interface for the Exploration of High-Dimensional Data Spaces, " *IEEE Pacific Vis*, 2012

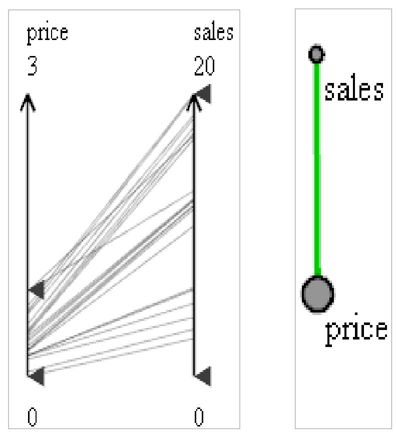
Bracketing and Conditioning

Correlation strength can often be improved by constraining a variable's value range

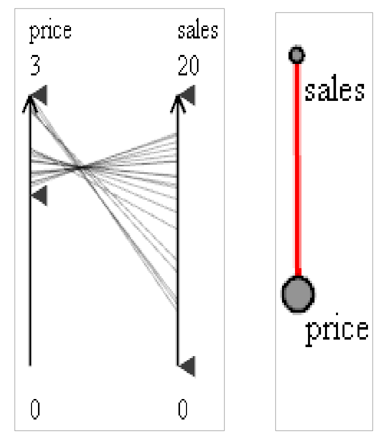
- this limits the derived relationships to this value range
- such limits are commonplace in targeted marketing, etc.



no bracketing



lower price range



higher price range

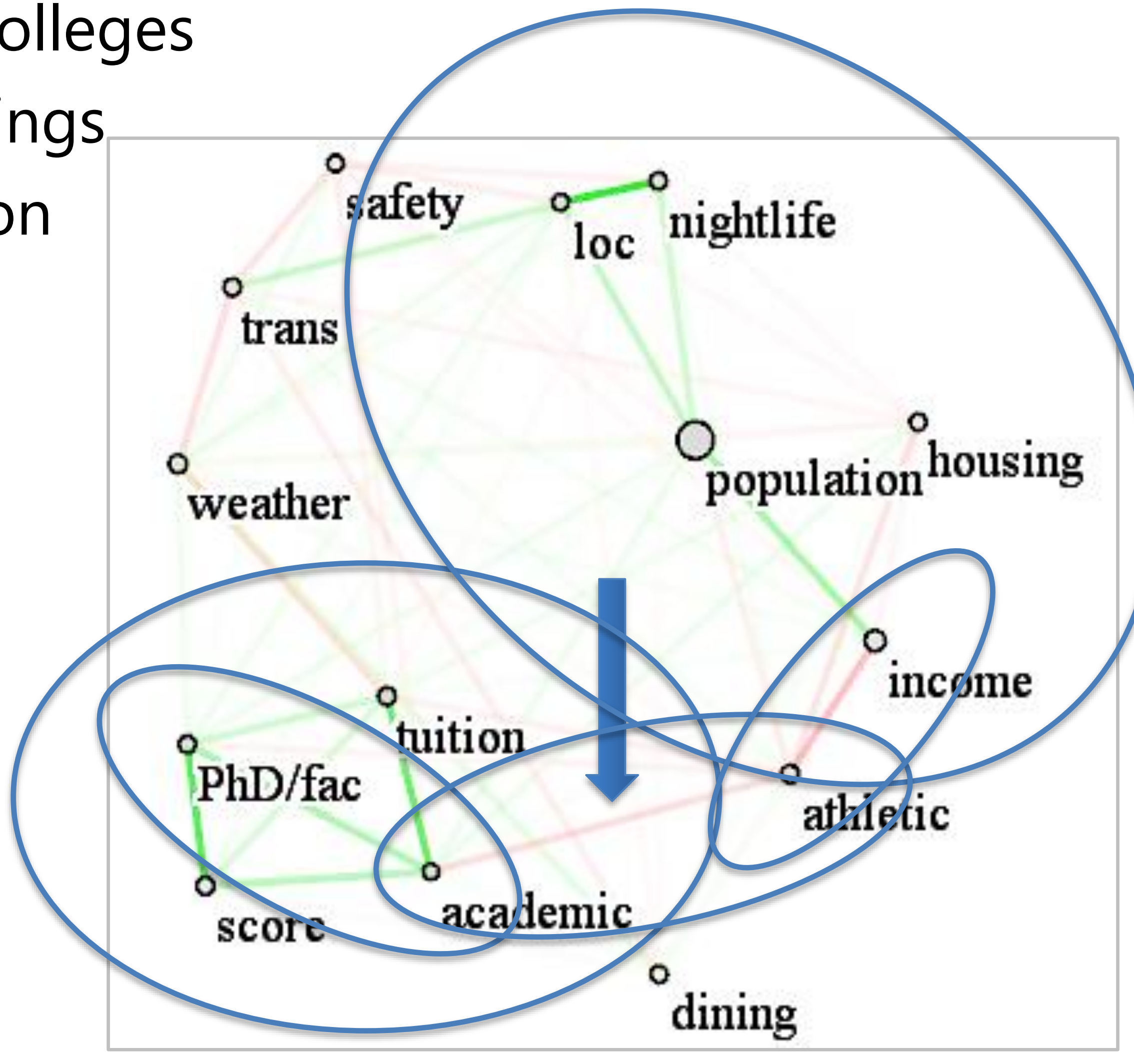
Z. Zhang, K. McDonnell, E. Zadok, K. Mueller, "Visual Correlation Analysis of Numerical and Categorical Data on the Correlation Map," *IEEE TVCG,* 2015.

Correlation Plots Are Powerful

Fused dataset of 50 US colleges

US News: academic rankings

College Prowler: survey on campus life attributes

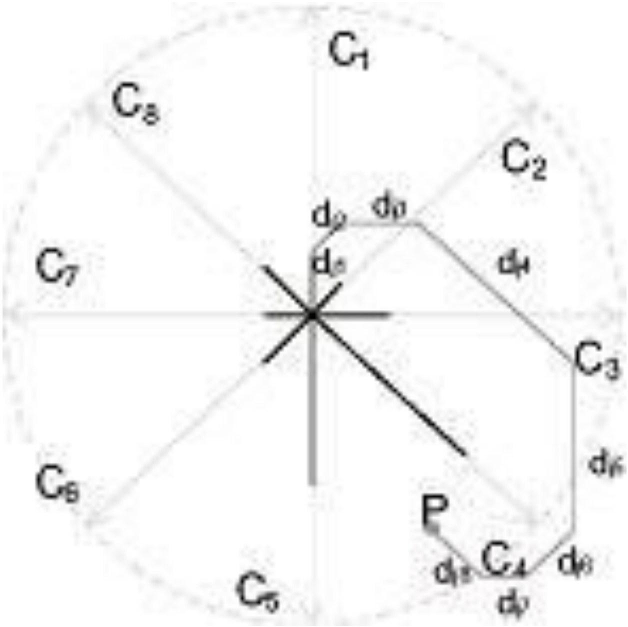


Radial Layouts

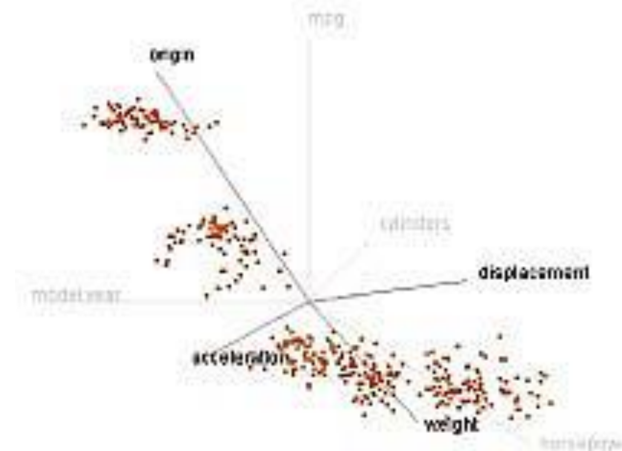
Star Coordinates

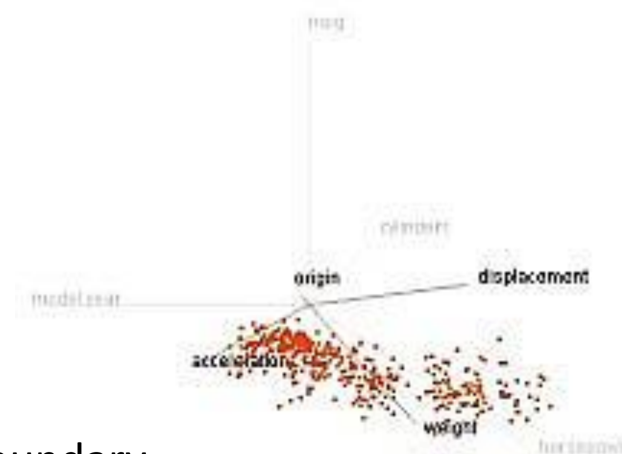
Coordinate system based on axes positioned in a “star”, or circular pattern

- no prior PCA and subsequent projection
- instead, a point P is plotted as a vector sum of all axis coordinates



Star Coordinates

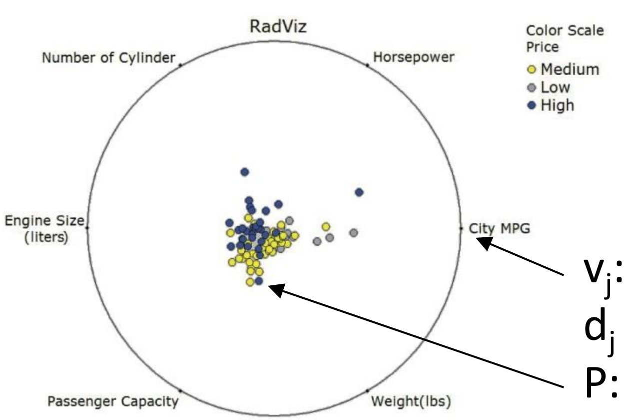




Operations defined on Star Coords

- scaling changes contribution to resulting visualization
- axis rotation can visualize correlations
- also used to reduce projection ambiguities

Similar paradigm: RadViz



$$P = \sum_{i=1}^{n} w_j v_j$$

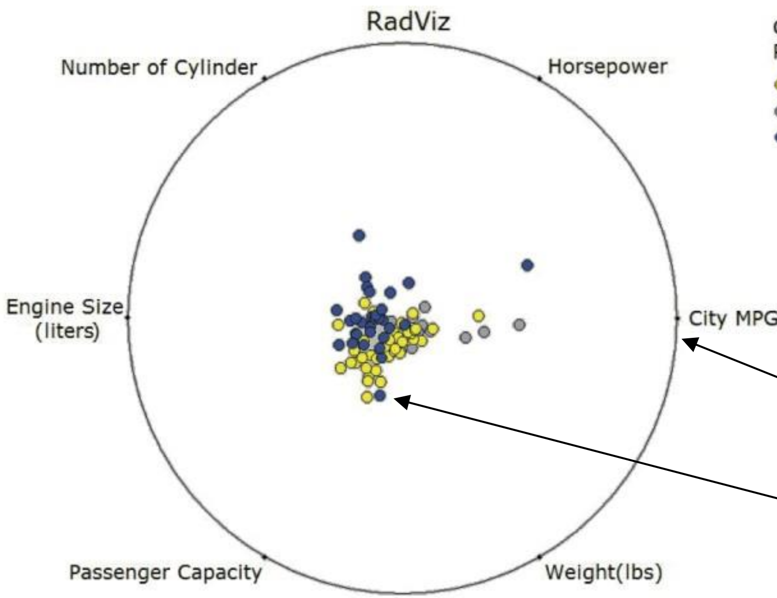
$$w_i = d_j / \sum_{k=1}^{n} d_k$$

v_j: attribute coordinates on disk boundary

d_j : data vector values

P: point location in RadViz disk

RadViz



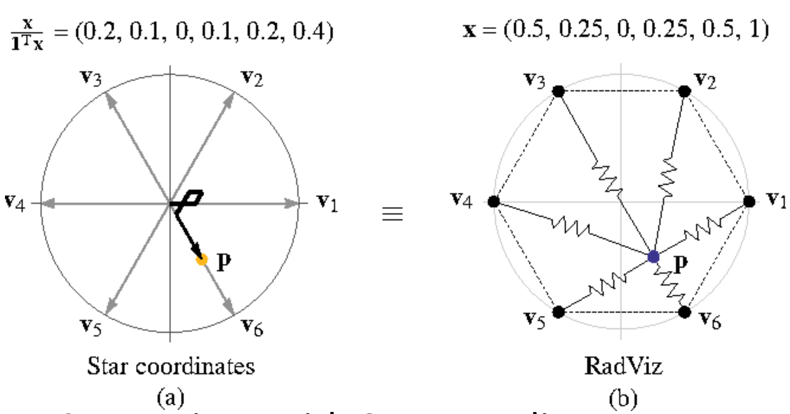
$$P = \sum_{i=1}^{n} w_j v_j$$

$$w_i = d_j / \sum_{k=1}^{n} d_k$$

v_j: attribute coordinates on disk boundary

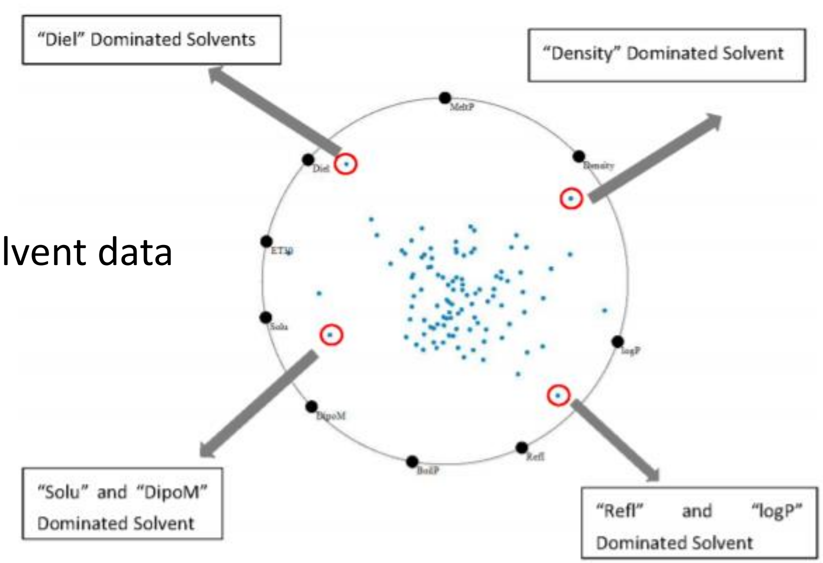
d_j : data vector values

P: point location in RadViz disk



Comparison with Star-coordinates

Solvent data



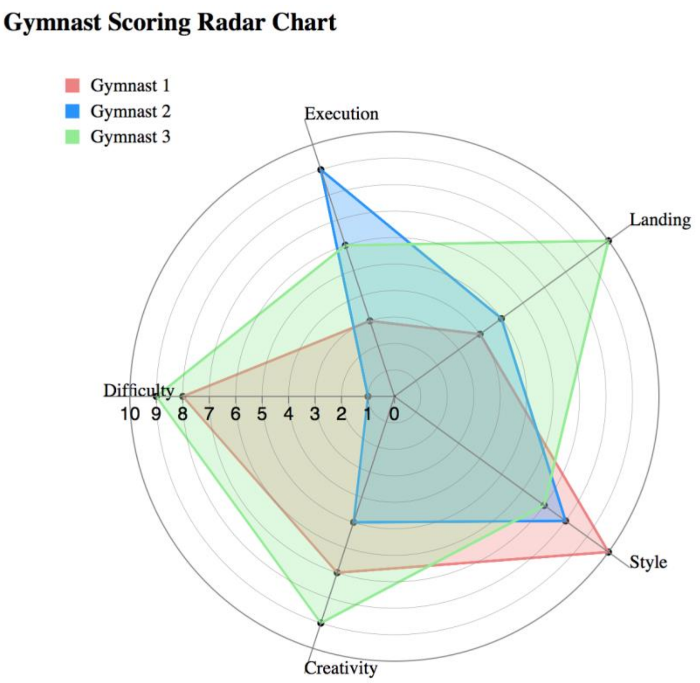
Radar Chart

Equivalent to a parallel coordinates plot, with the axes arranged radially

- each star represents a single observation
- can show outliers an commonalities nicely

Disadvantages

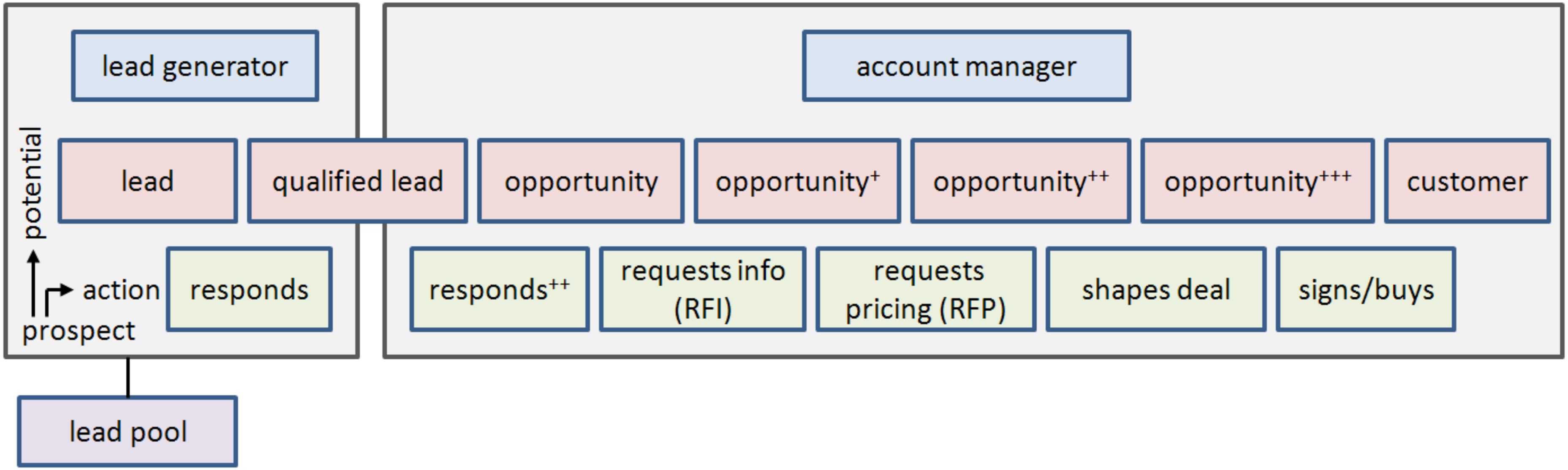
- hard to make trade-off decisions
- distorts data to some extents when lines are filled in



Telling Stories with Parallel Coordinates

Example: Sales Strategy Analysis

Anatomy of a Sales Pipeline



The Setup

Scene:

- a meeting of sales executives of a large corporation, Vandelay Industries

Mission:

- review the strategies of their various sales teams

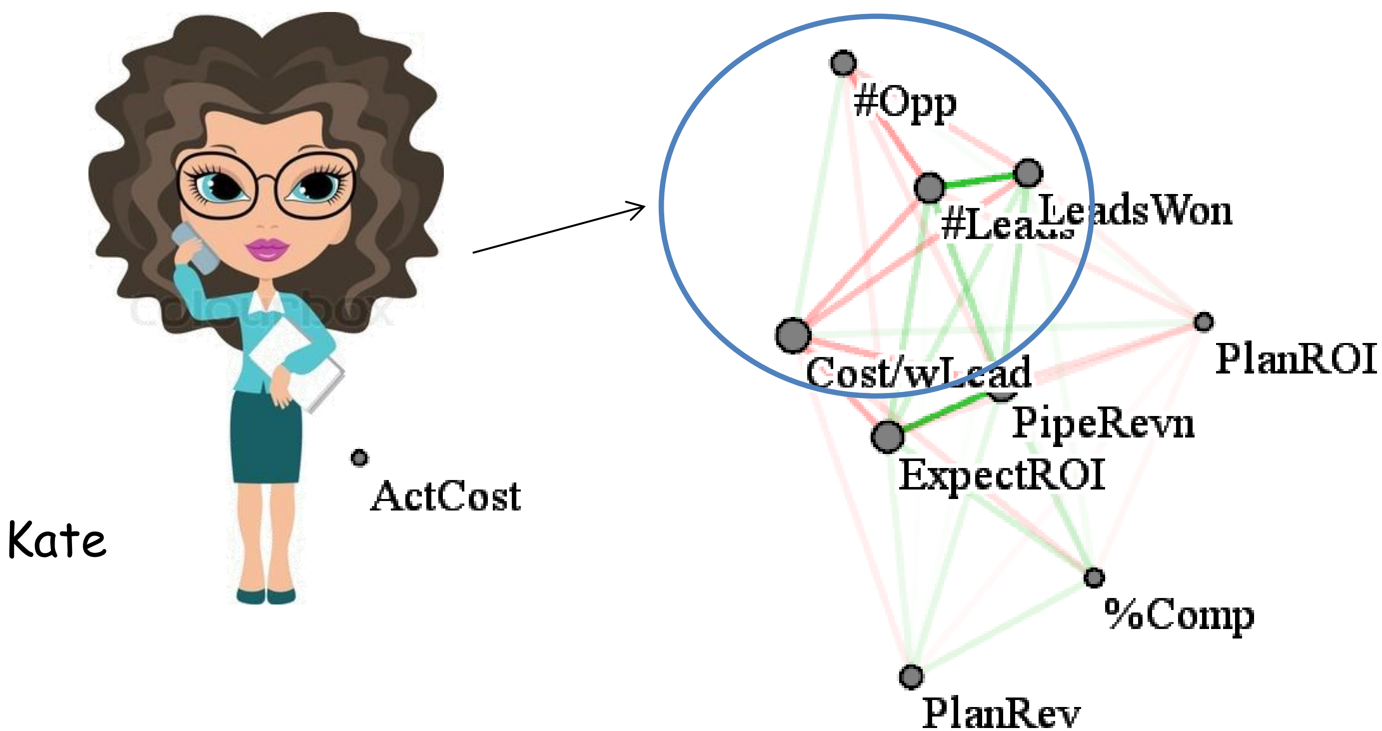
Evidence:

- data of three sales teams with a couple of hundred sales people in each team

Kate Explains It All

Meet Kate, a sales analyst in the meeting room:

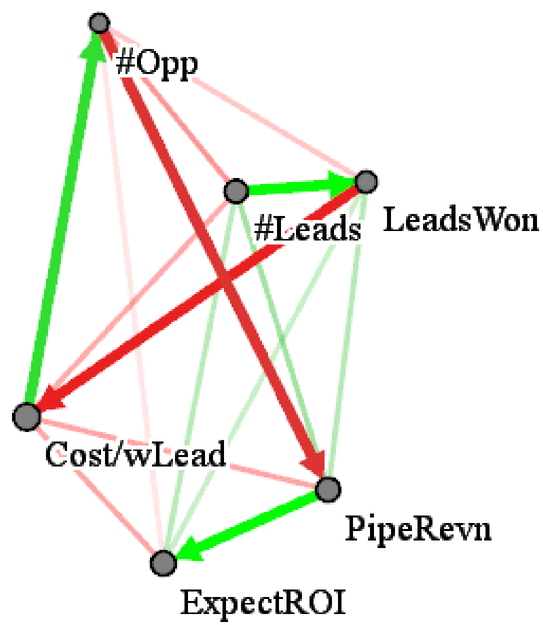
> "OK…let's see, cost/won lead is nearby and it has a positive correlation with #opportunities but also a negative correlation with #won leads"

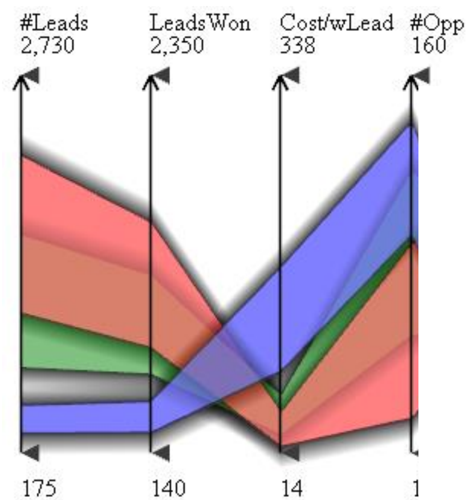


Kate Designs the Narration

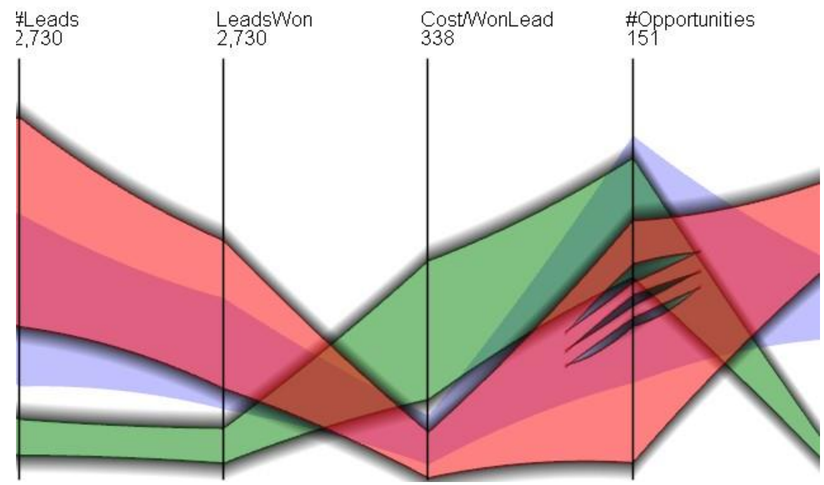
"Let's go and make a revealing route!"

- she uses the mouse and designs the route shown
- she starts explaining the data like a story …





Further Insight



Kate notices something else:

- now looking at the red team
- there seems to be a spread in effectiveness among the team
- the team splits into three distinct groups

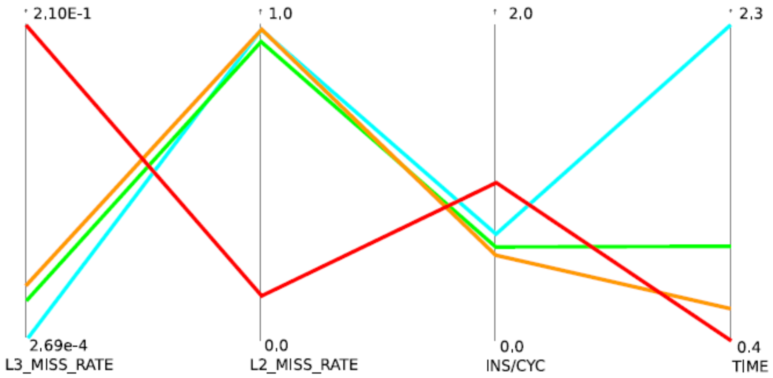
She recommends: “Maybe fire the least effective group or at least retrain them”

How to Teach Mainstream Users

Recent Reviewer Comment

From a paper sent to a software visualization conference:

Figure 8

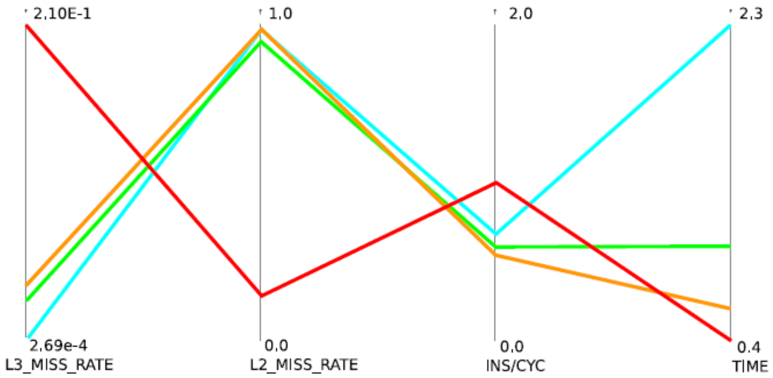


- Multiple visualizations appear to present categorical data as line graphs, which seems a strange choice.

Recent Reviewer Comment

From a paper sent to a software visualization conference:

Figure 8

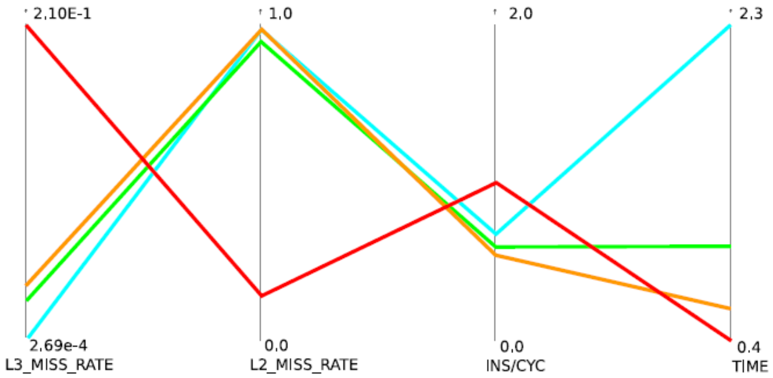


- Multiple visualizations appear to present categorical data as line graphs, which seems a strange choice. Figure 8, for example, at first sight appeared to be showing a change over time, but in fact further inspection shows that the different x-coordinates are almost entirely unrelated to one another and in no particular order.

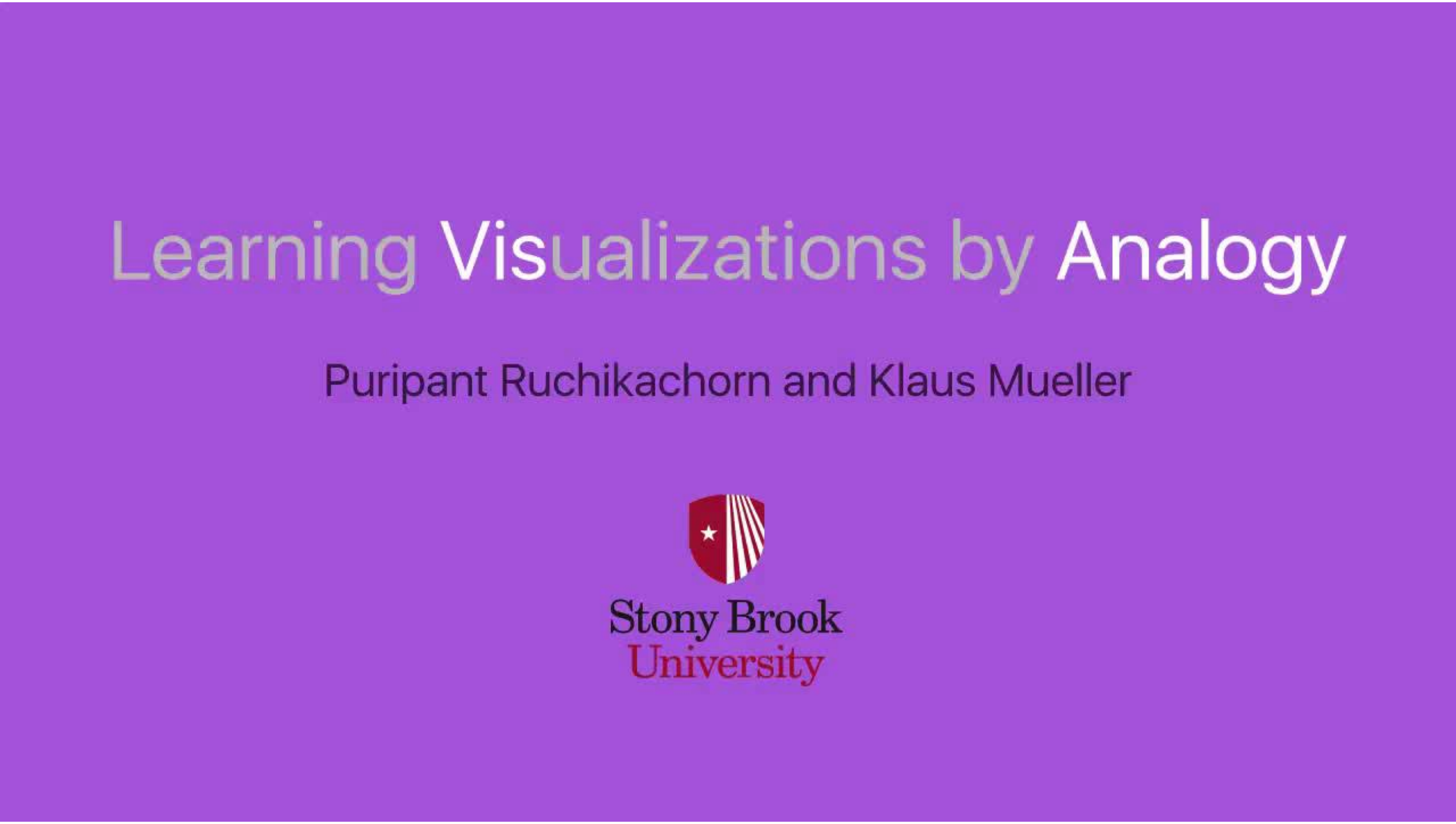
Recent Reviewer Comment

From a paper sent to a software visualization conference:

Figure 8



- Multiple visualizations appear to present categorical data as line graphs, which seems a strange choice. Figure 8, for example, at first sight appeared to be showing a change over time, but in fact further inspection shows that the different x-coordinates are almost entirely unrelated to one another and in no particular order. This is such an unusual choice that I'm not sure that I am understanding the role of the graphs correctly.



https://www.youtube.com/watch?v=mdolkHA-RpA

User Studies

Encode user responses based on task complexities

- none (0): cannot report any findings
- low (1): understand representation visual encoding
- medium (2): identify groups and outliers
- high (3): recognize correlations and trends

User Studies – Car Dataset

Visual understanding:

(1) The MPG of the orange-highlighted car is ~40% of its range
(2) There is just one line at the top of the acceleration scale
(3) Heavier cars are faster

Data Understanding:

(1) The number of cylinders of the orange-highlighted car is 4, one fifth between 3 and 8.
(2) Many cars have the same numbers of cylinders, mostly even numbers particularly 4 and 8.
(3) Heavier cars have more cylinders and hence more horsepower and speed.

Results

Participants		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
Parallel Coordinates Plot	Before	3	0	0	0	1	0	2	1	0	3	3
	After	3	2	2	1	2	2	3	2	1	3	3
	Diff.	0	2	2	1	1	2	1	1	1	0	0

D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11
0	2	3	1	1	3	1	1	2	0	3
2	3	3	3	1	3	2	2	3	2	3
2	1	0	2	0	0	1	1	1	2	0

Next Lecture:
Non-Linear Projection Techniques (Embeddings)