

CSE/ISE 332

Introduction to Visualization

Visualizing High-Dimensional Data: Linear  
Methods

Klaus Mueller

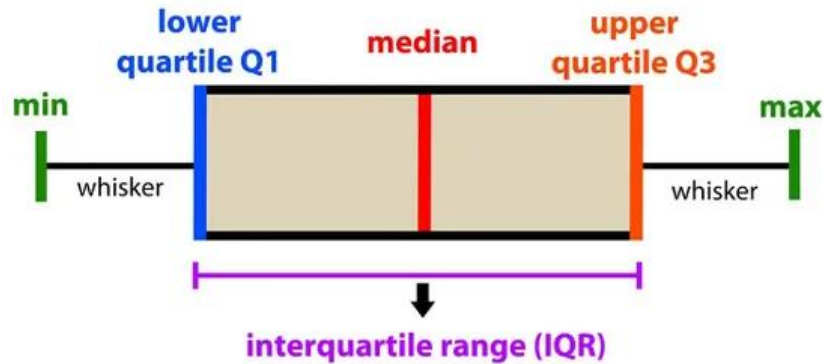
Computer Science Department  
Stony Brook University

| 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)                   | Project #1 out             |
| 4       | AI-assisted coding for VIS applications (design, debugging, refactoring)        |                            |
| 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             | Project #2(a) out          |
| 8       | Perception and cognition (human visual system, color, contrast)                 |                            |
| 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            | Project #2(b) out          |
| 12      | Visualization and AI: mutual support and capabilities (VIS4AI, AI4VIS)          |                            |
| 13      | Principles of interaction: drive what is visualized, analyzed & how (HCI4VIS)   |                            |
| 14      | Midterm recap                                                                   |                            |
| 15      | Midterm                                                                         |                            |
| 16      | Visual analytics, human-centered AI, mixed-initiative & collaborative VA        | Capstone proposal due      |
| 17      | Visualization system design & evaluation: nested model, study designs           |                            |
| 18      | Visualization of hierarchical data                                              |                            |
| 19      | Visualization of maps and data with geo-reference                               |                            |
| 20      | Vis. of time-varying, time-series, streaming data, progressive visualization    |                            |
| 21      | Critiquing visualization: Tufte's views, integrity, uncertainty, responsibility | Capstone prelim report due |
| 22      | Design of effective infographics                                                |                            |
| 23      | Foundations scientific & medical visualization, computer graphics               |                            |
| 24      | Intro to volume rendering                                                       |                            |
| 25      | Scientific visualization                                                        |                            |
| 26      | Capstone project demos (4-5 minutes + 2 minutes Q&A)                            | All capstone materials due |
| 27      | Capstone project demos (4-5 minutes + 2 minutes Q&A)                            |                            |
| 28      | Capstone project demos (4-5 minutes + 2 minutes Q&A)                            |                            |
| Final   | Final exam                                                                      |                            |

# Visualizing Distributions

# Box Plots

You may have heard about box plots

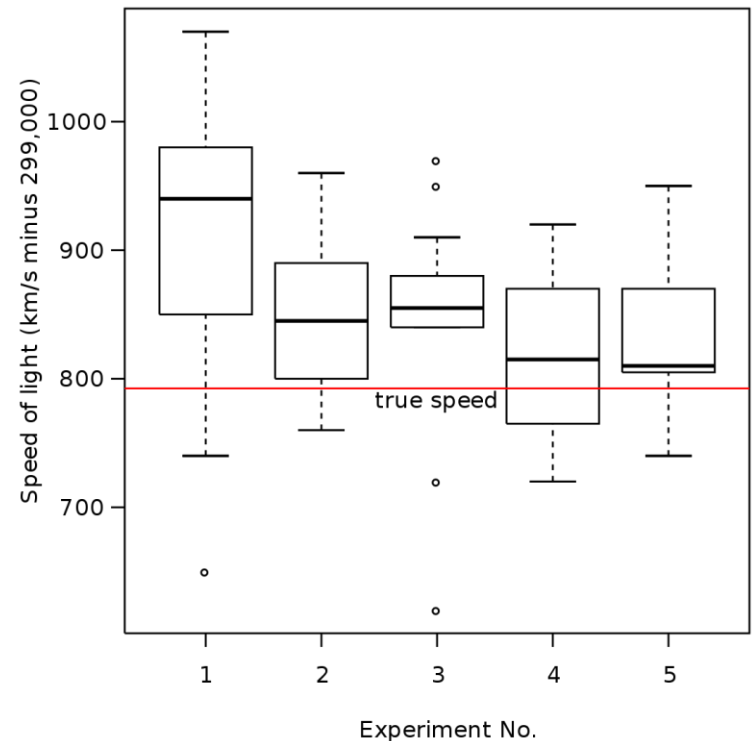


Tend to be bewildering to many

- hard to interpret

They can also give the wrong representation of data

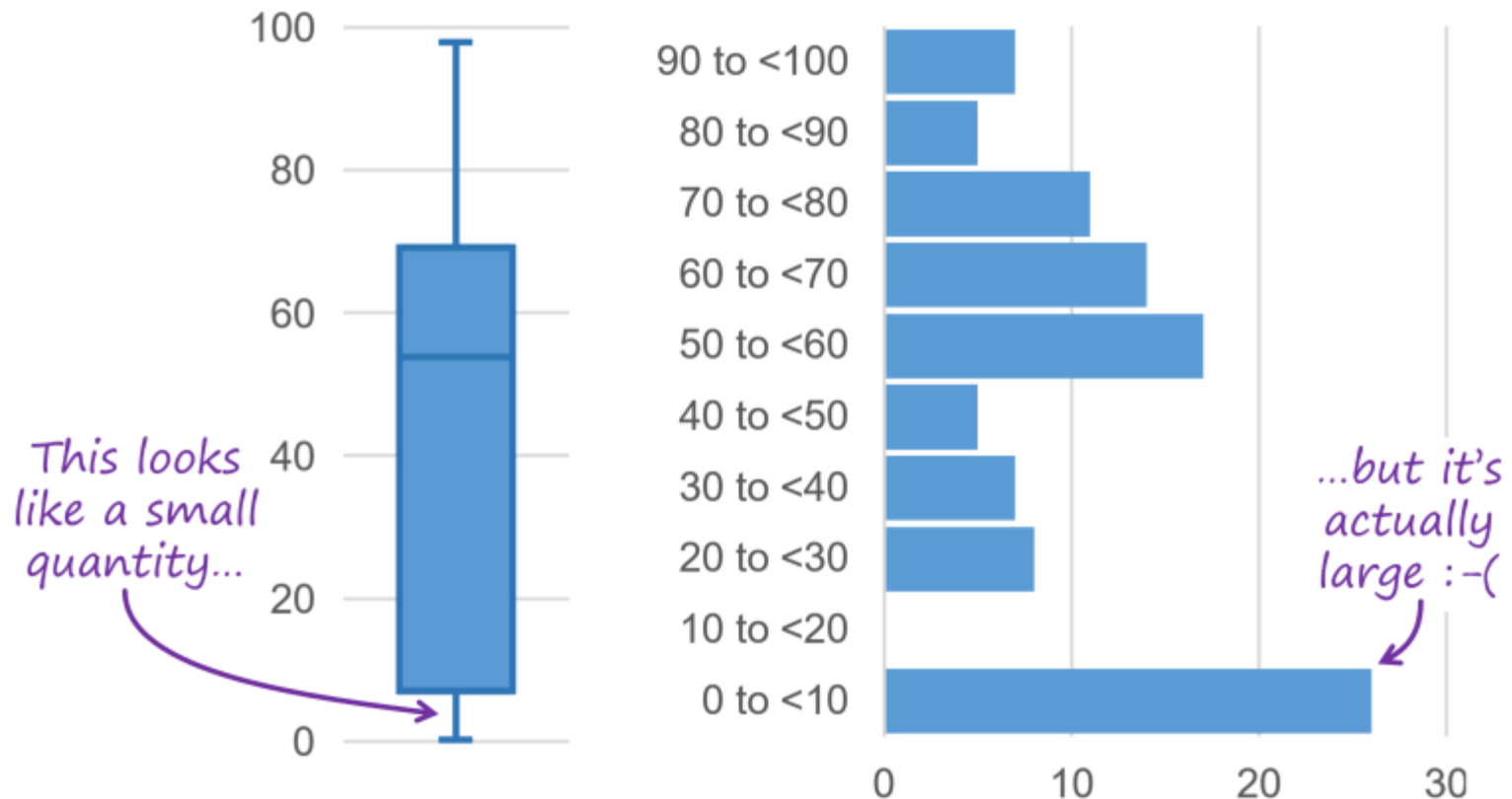
- assume normal distributed data



# Box Plots

Non-normal distributed data give “wrong” box plots

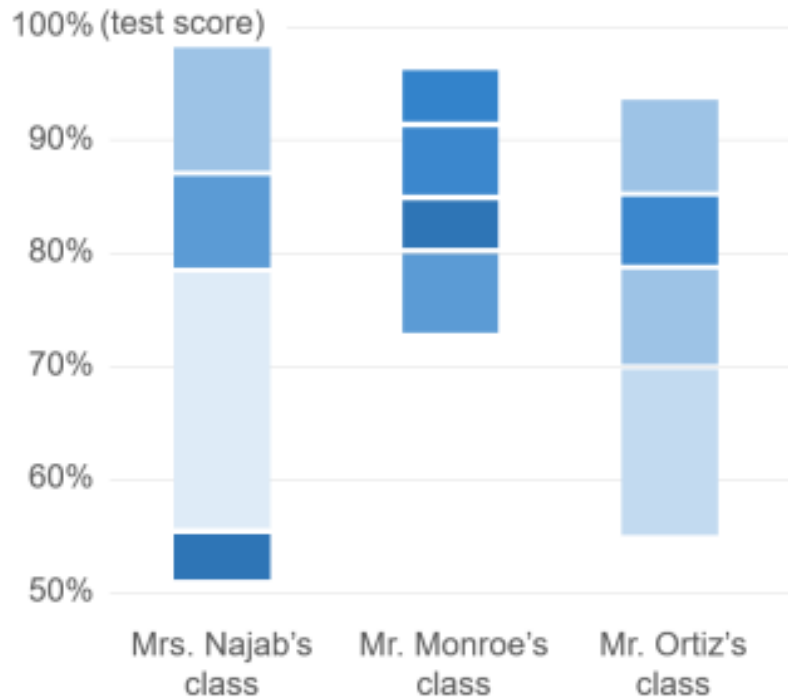
- shown here: data on student test scores



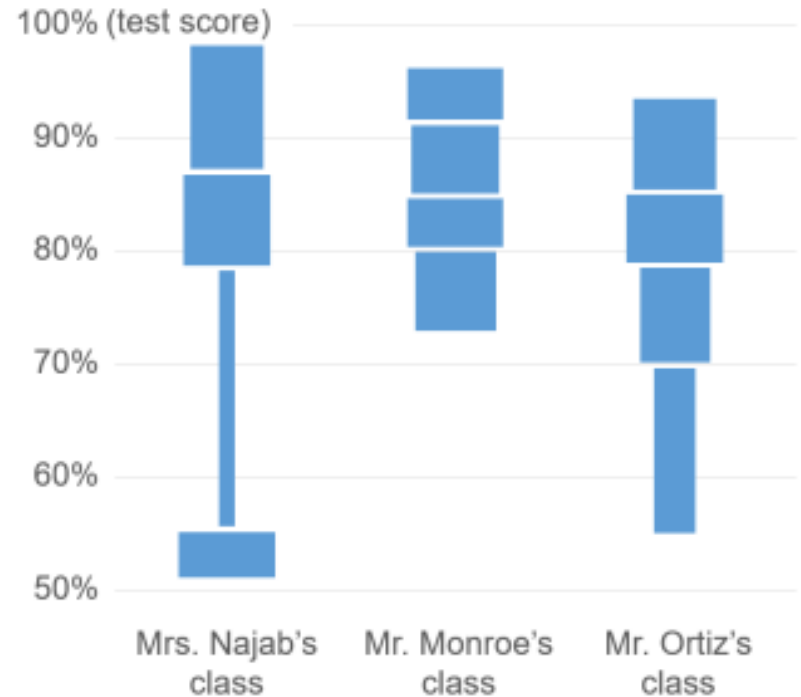
# Density Plots

Same data than last slide, multiple classes

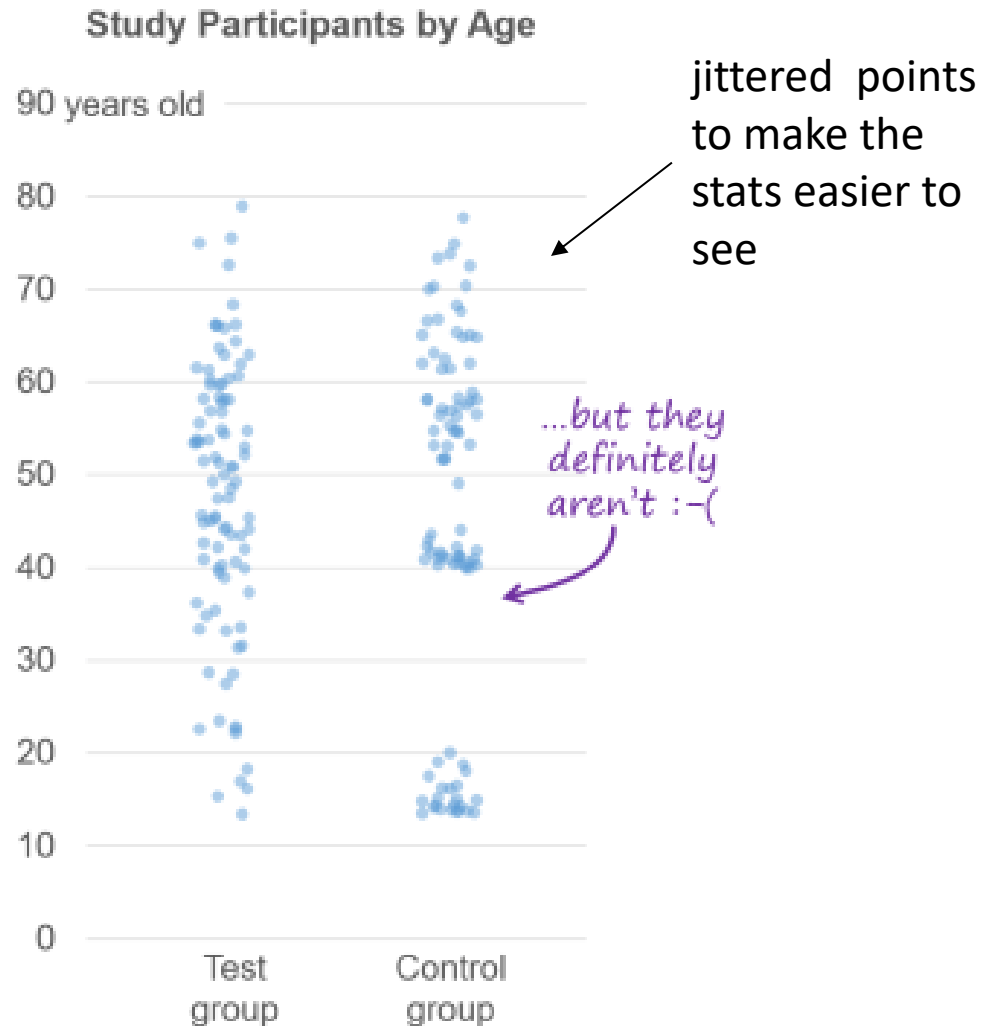
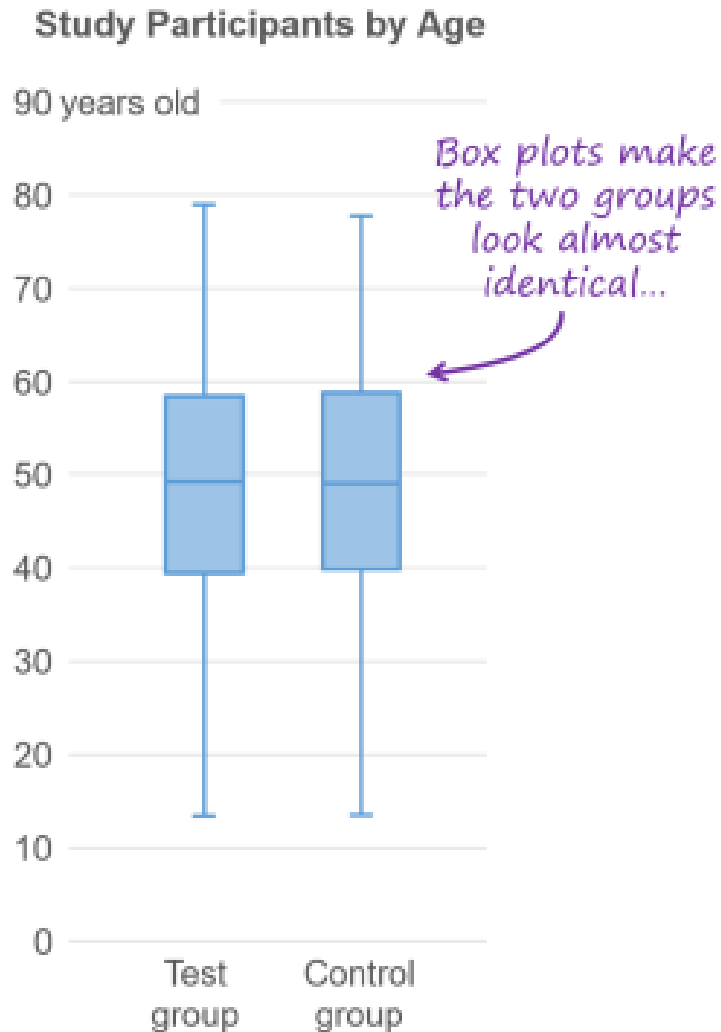
Student Test Scores by Class



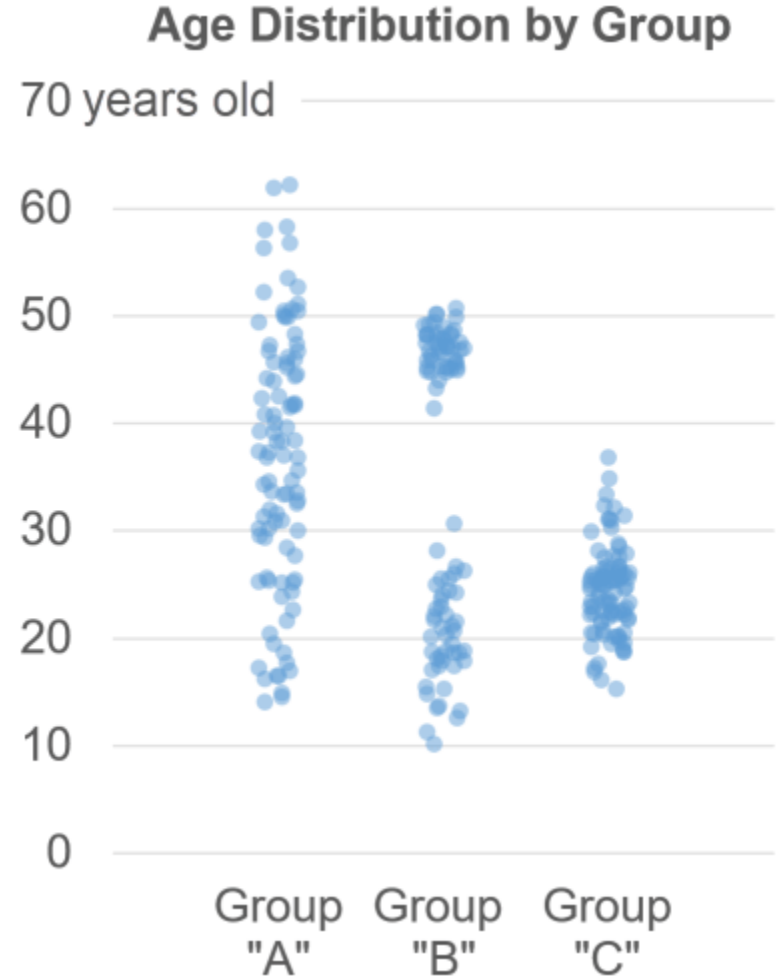
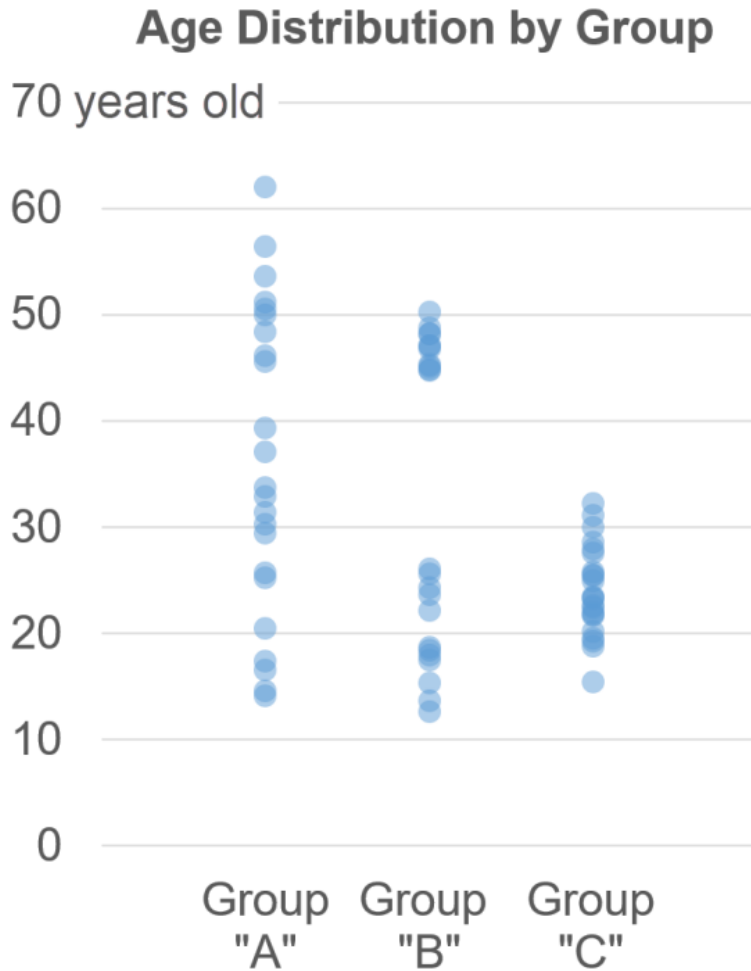
Student Test Scores by Class



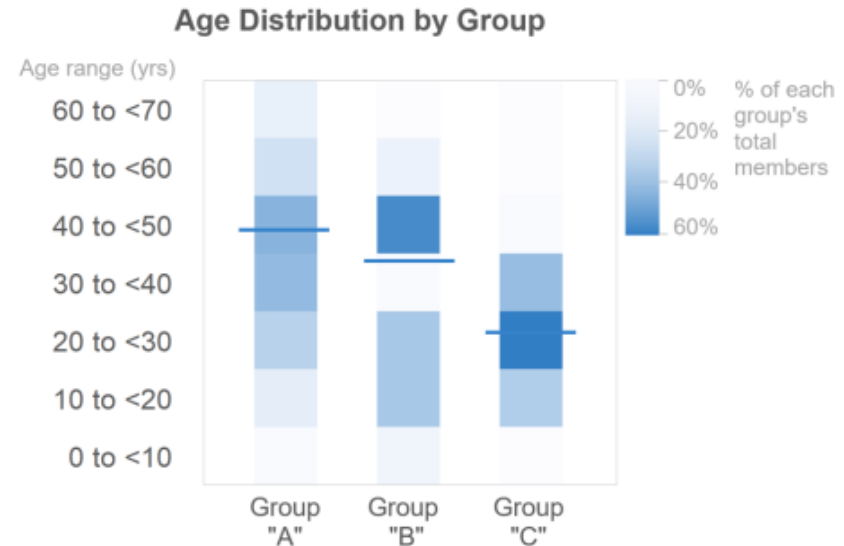
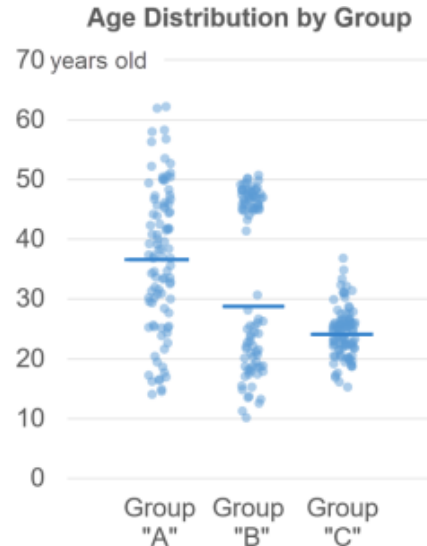
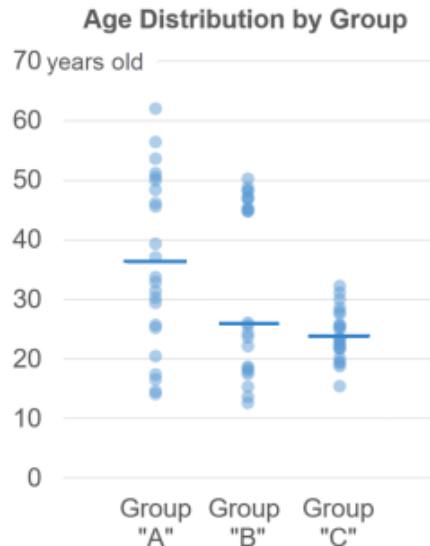
# Strip Plots



# Semitransparent vs. Jittering



# Comparison

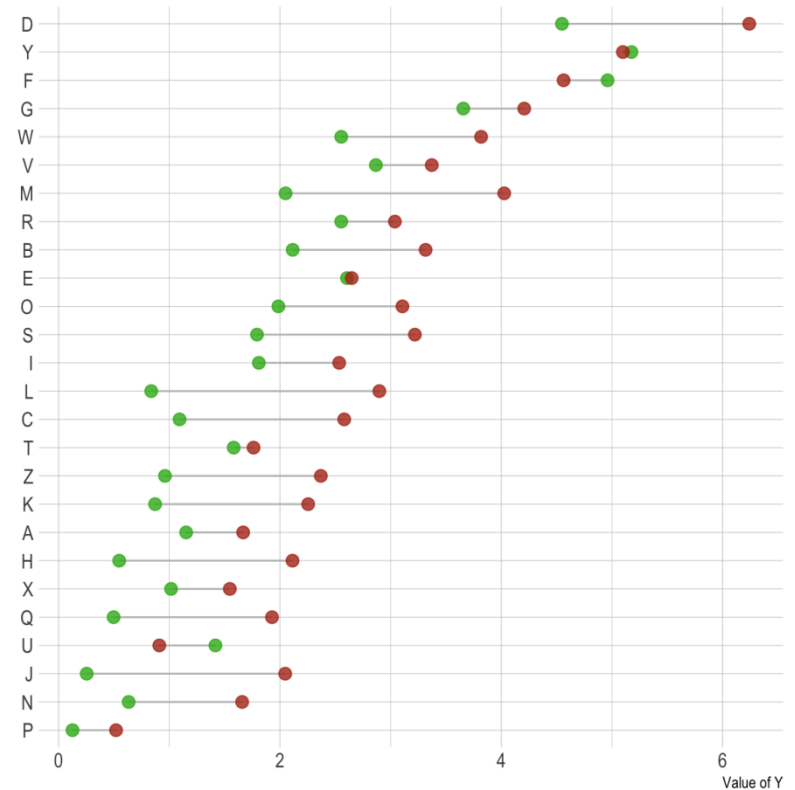
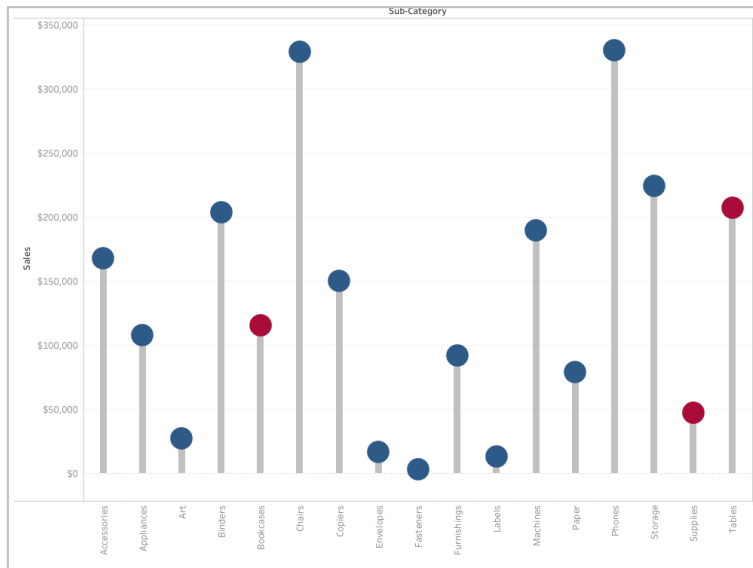


With median lines

Read more here:

<https://nightingaledvs.com/ive-stopped-using-box-plots-should-you/>

# Lollipop Charts

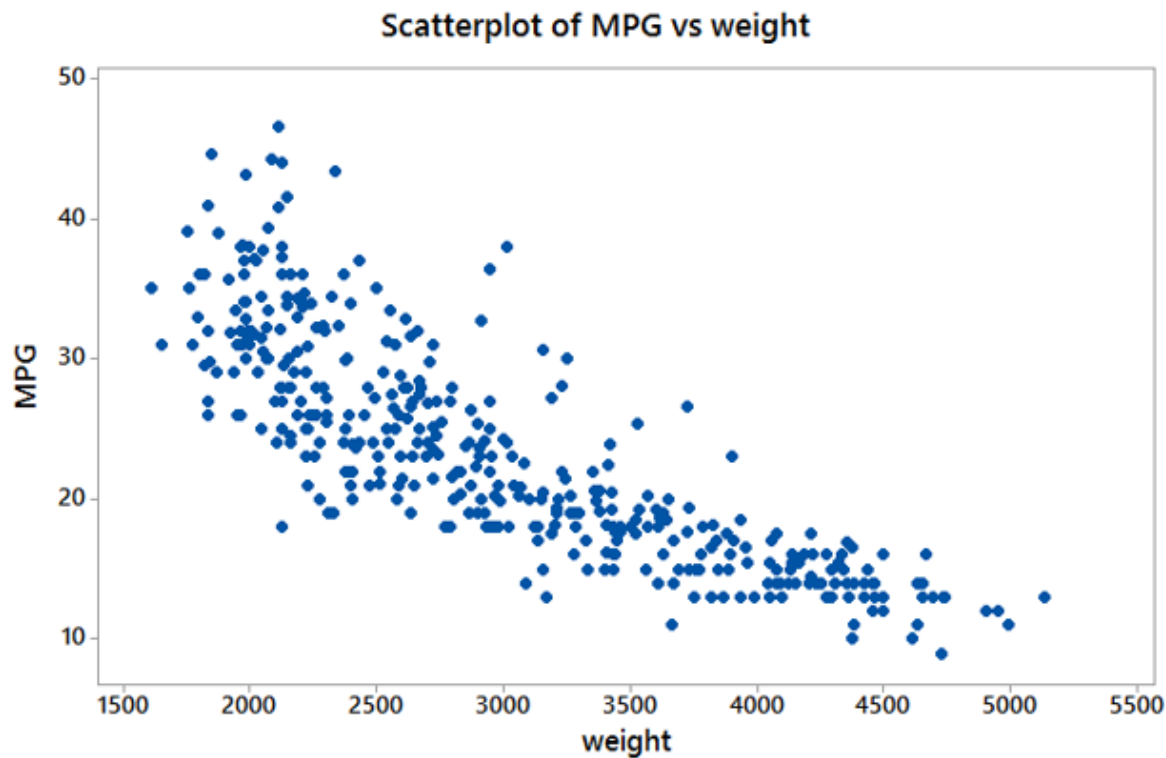


makes it easier to see and compare positions than scatter plots

# Scatterplots

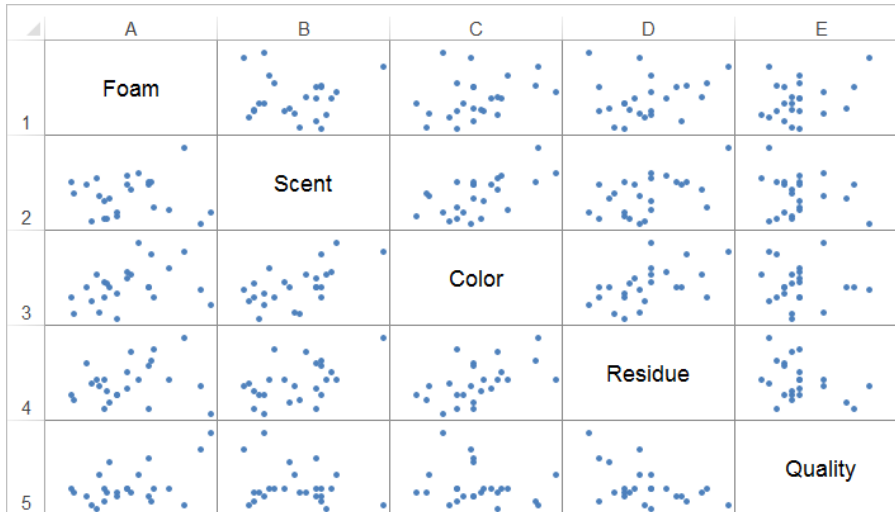
# Scatterplots

Projection of the data items into a bivariate basis of axes

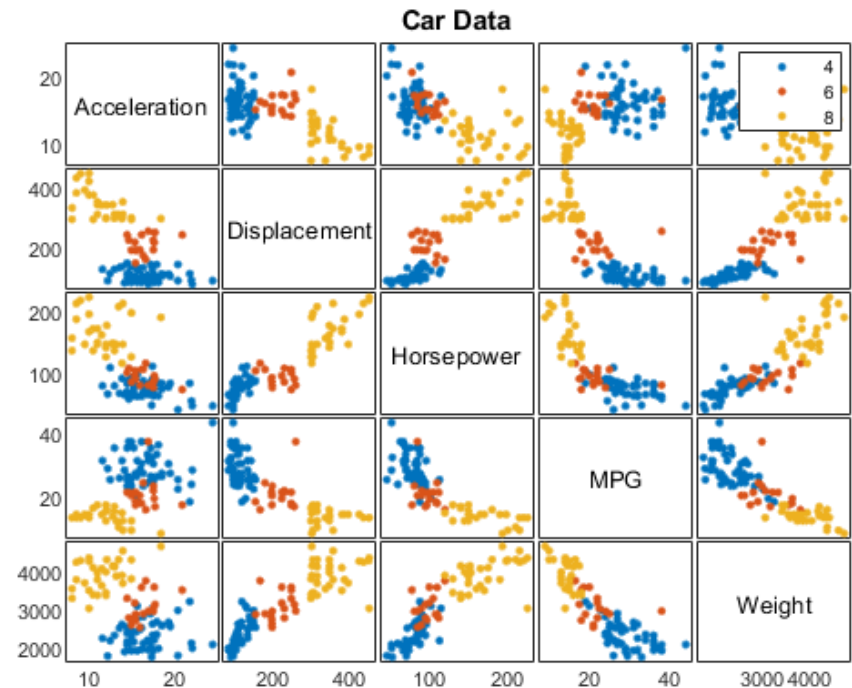


But what if you have more than two variables?

# 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?

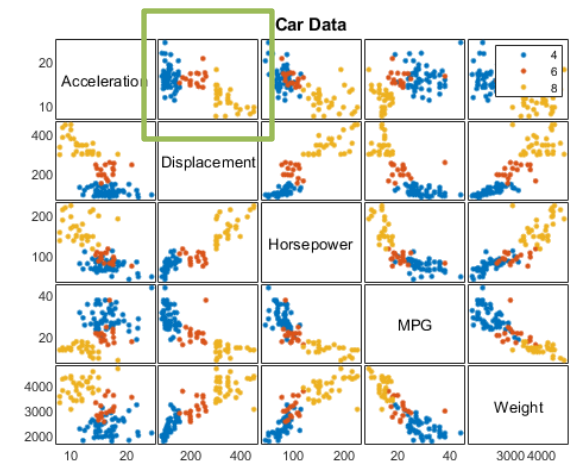
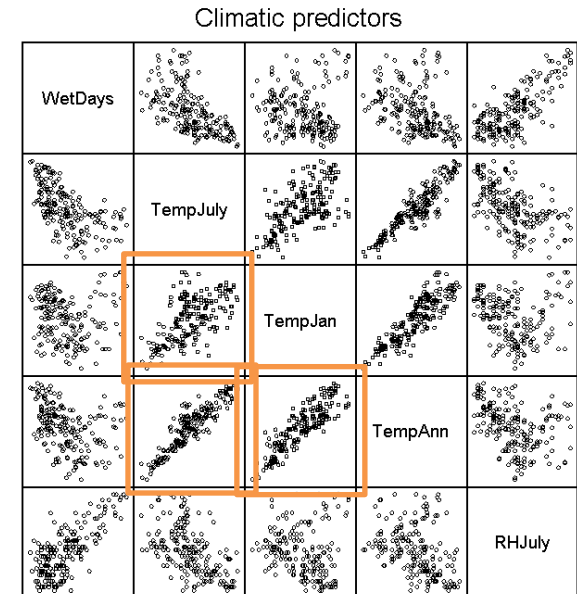
- distributes  $n(n-1)$  bivariate relationships over a set of tiles
- for  $n=4$  get 16 tiles
- can use  $n(n-1)/2$  tiles

For even moderately large  $n$ :

- there will be too many tiles

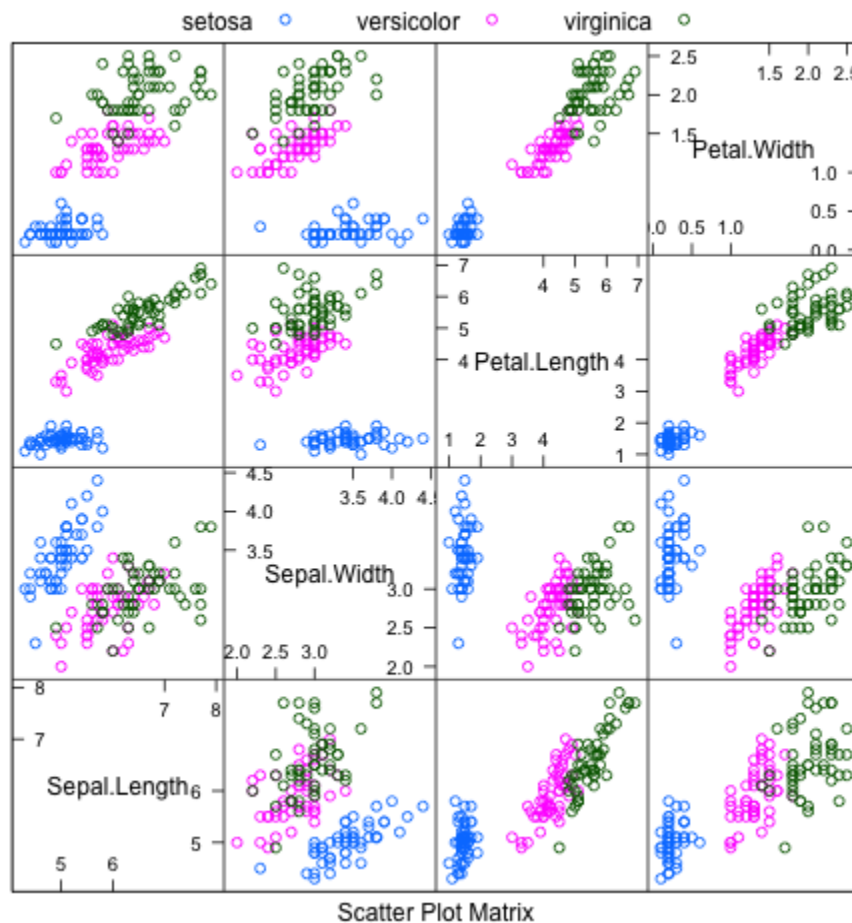
Which plots to select?

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



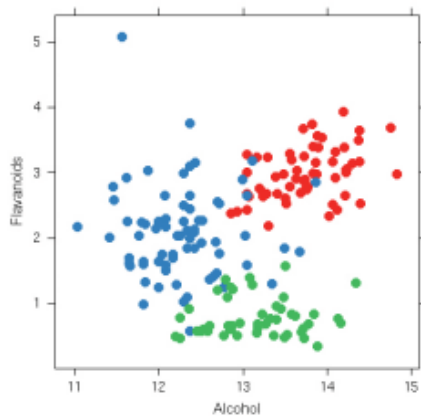
# Tile Selection

Select the most interesting tiles and show them to the user

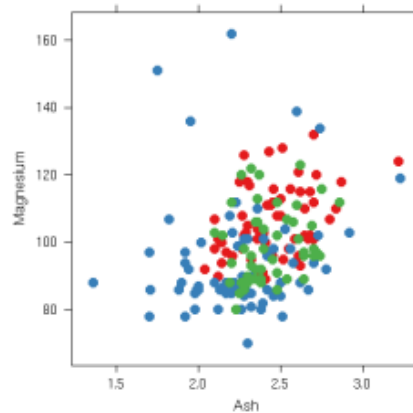


# Automated Tile Selection

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



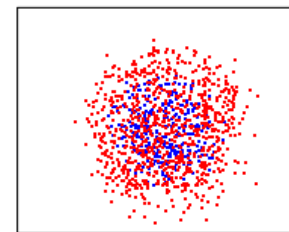
(a) DSC=90



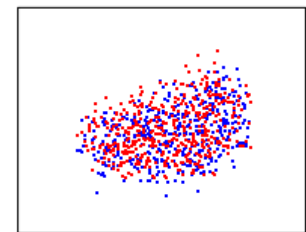
(b) DSC=49

$$\text{DSC} = \frac{|\{x' \in v(X) : \mathbf{CD}(x', \text{centr}'(c_{\text{label}(x)})) = \text{true}\}|}{k}$$

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

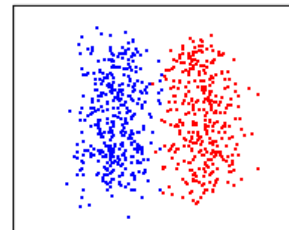


(d) 29

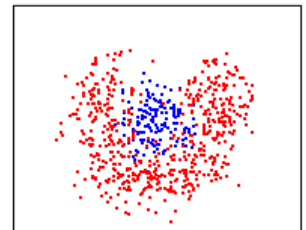


(e) 15

bad



(a) 99



(b) 74

OK

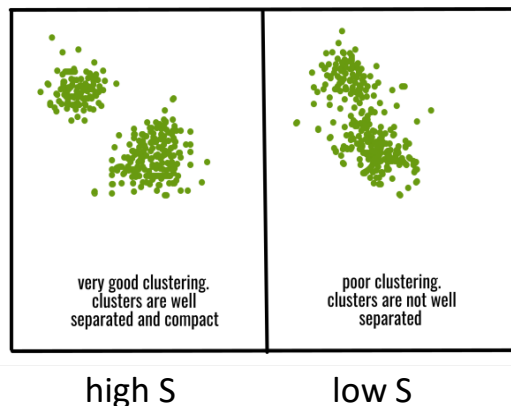
# Another Metric: Silhouette Score

## 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  $s(i)$



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

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

# Dunn Index

Favors clusters that (1) are compact and (2) are well isolated

$$DI_m = \frac{\min_{1 \leq i < j \leq m} \delta(C_i, C_j)}{\max_{1 \leq k \leq m} \Delta_k}$$

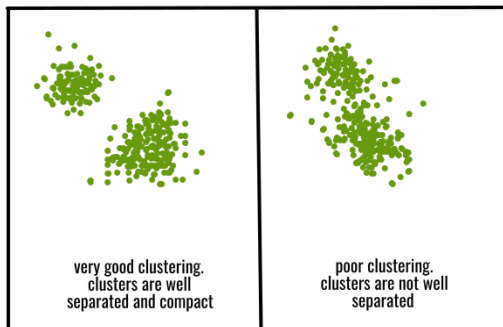
← min separation

← max spread

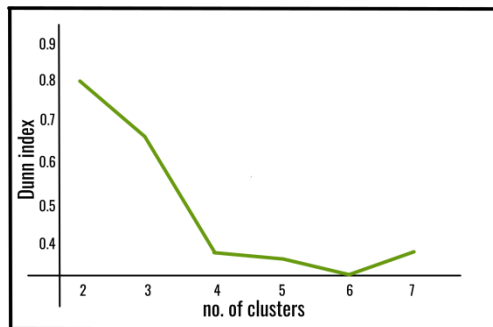
$$\Delta_i = \frac{\sum_{x \in C_i} d(x, \mu)}{|C_i|}, \mu = \frac{\sum_{x \in C_i} x}{|C_i|}$$

calculates distance of all the points from the mean.

$\delta(C_i, C_j)$  be this intercluster distance metric, between clusters  $C_i$  and  $C_j$ .



high Dunn Index      low Dunn Index



determine the quality of  
k-means clustering



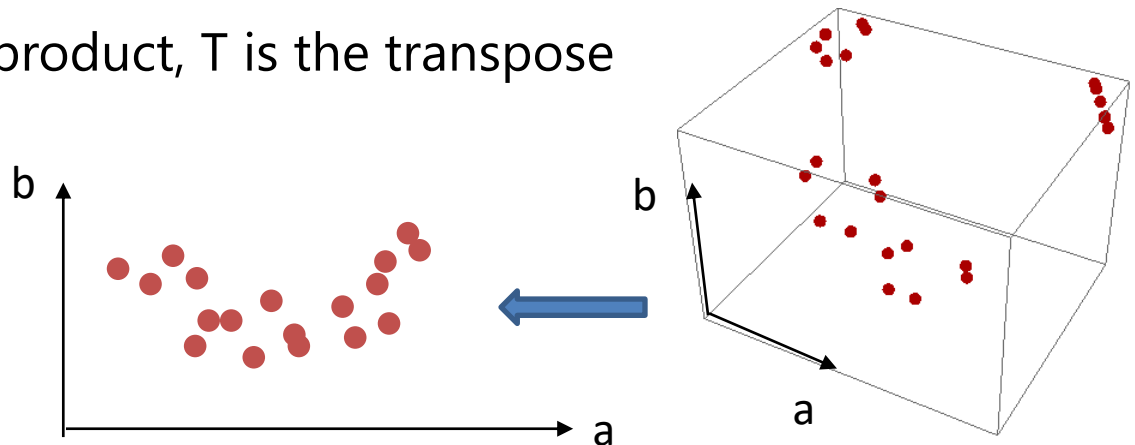
# Biplots

# Projection Operations

How does 2D projection work in practice?

- N-dimensional point  $x = \{x_1, x_2, x_3, \dots, x_N\}$
- a basis of two orthogonal axis vectors defined in N-D space
$$a = \{a_1, a_2, a_3, \dots, a_N\}$$
$$b = \{b_1, b_2, b_3, \dots, b_N\}$$
- a projection  $\{x_a, x_b\}$  of  $x$  into the 2D basis spanned by  $\{a, b\}$  is:
$$x_a = a \cdot x^T$$
$$x_b = b \cdot x^T$$

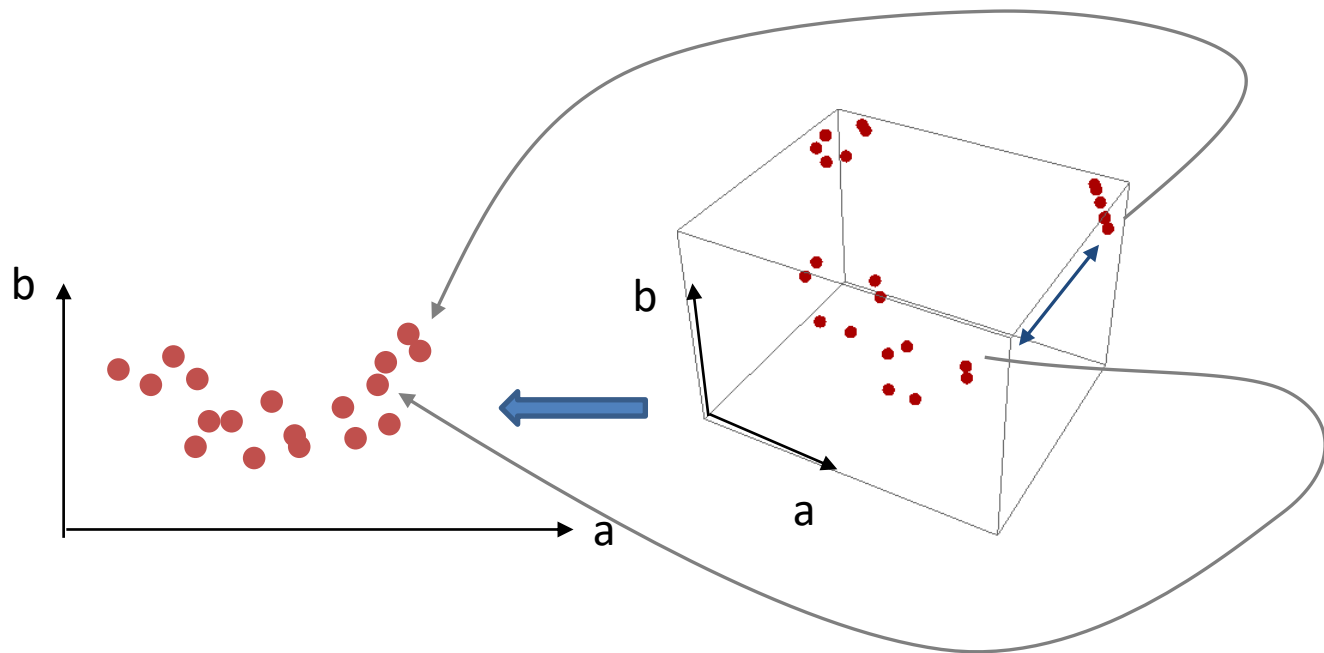
where  $\cdot$  is the dot product, T is the transpose



# Projection Ambiguity

Projection causes inaccuracies

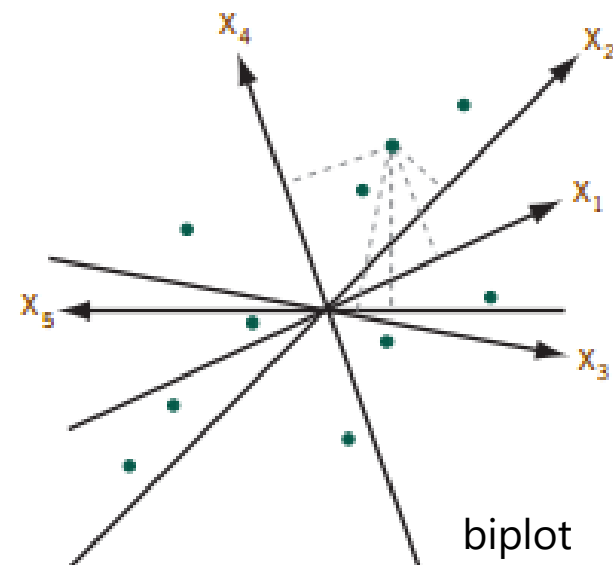
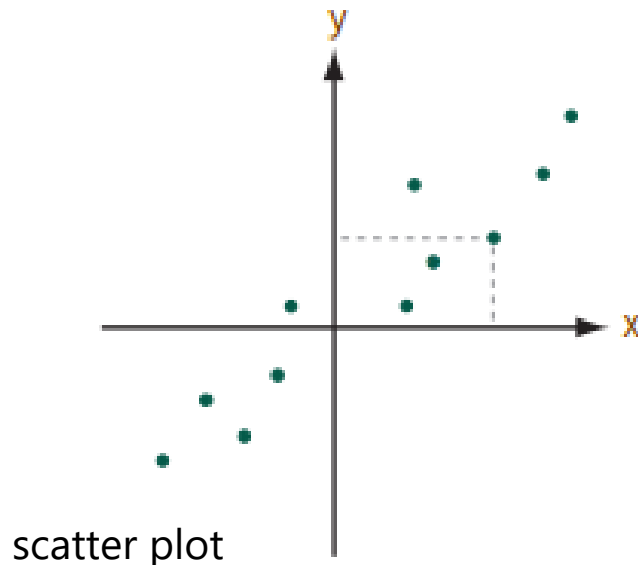
- close neighbors in the projections may not be close neighbors in the original higher-dimensional space
- this is called *projection ambiguity*



# Biplots

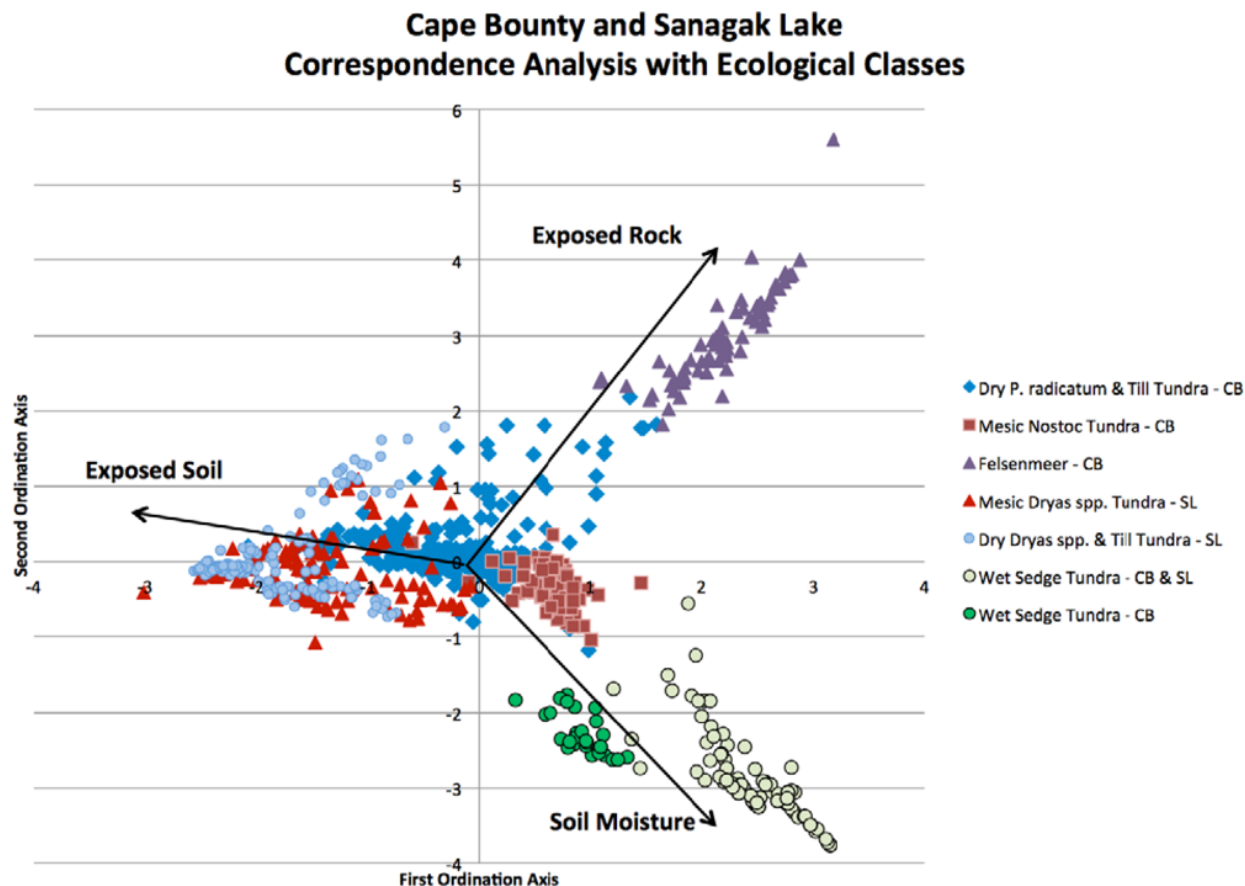
Plots data points and dimension axes into a single visualization

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



# Biplots Can Have Projection Ambiguities

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



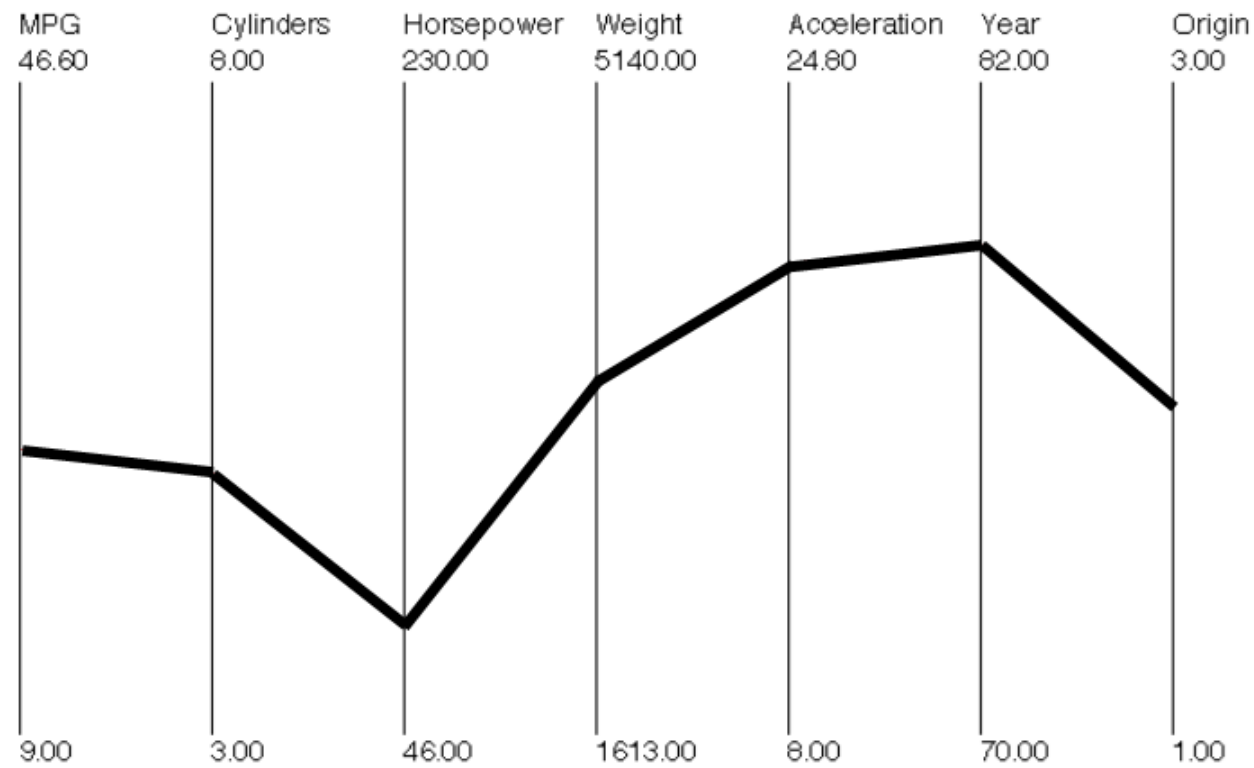
# Biplots – A Word of caution

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

# Parallel Coordinates

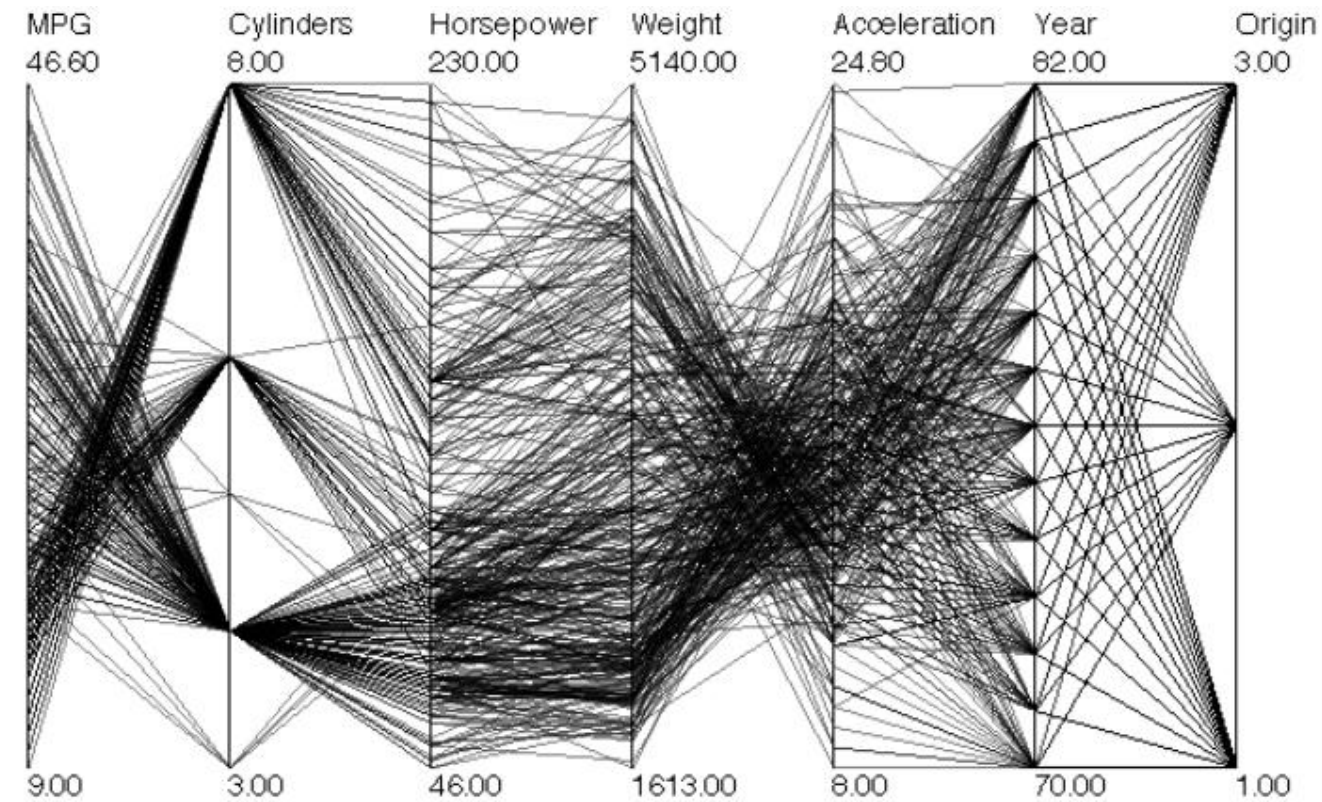
# Parallel Coordinates – 1 Car



The  $N=7$  data axes are arranged side by side

- in parallel

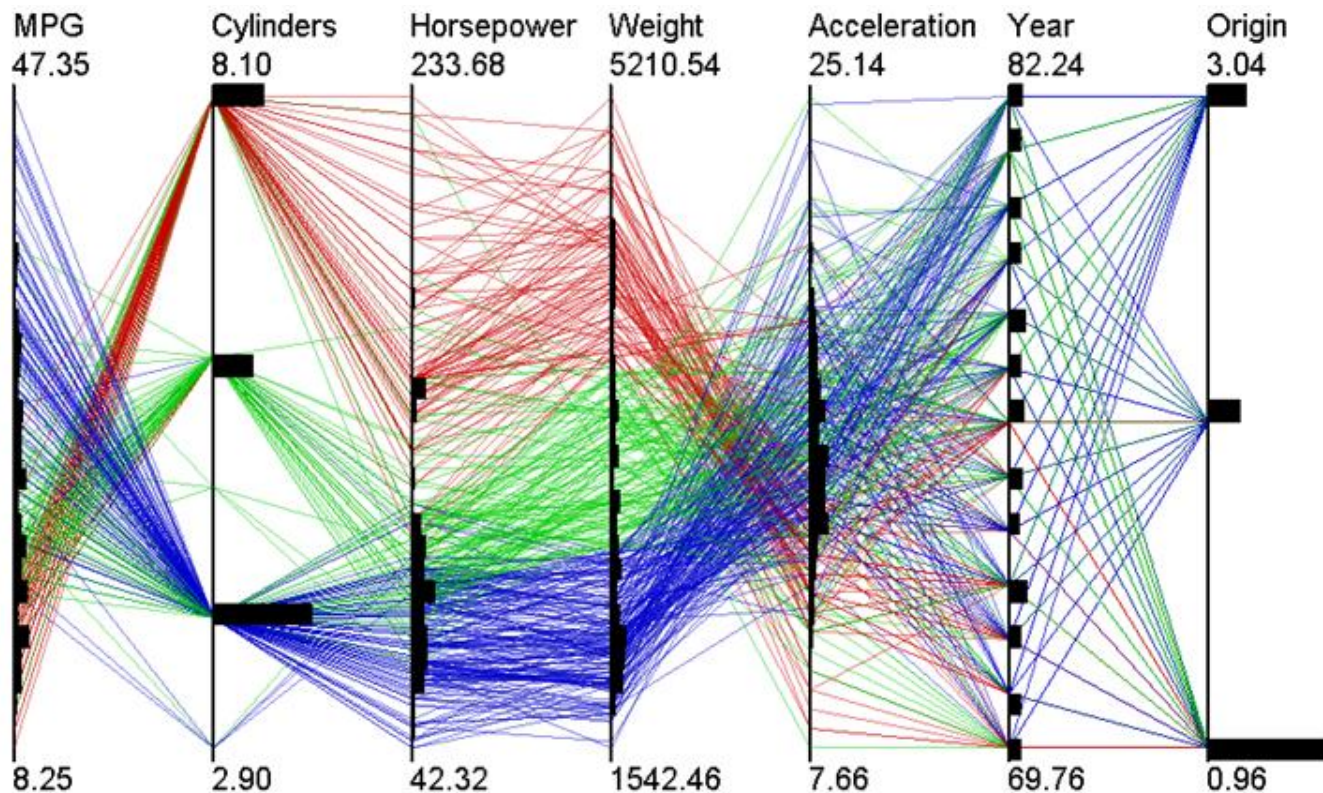
# Parallel Coordinates – 100 Cars



Hard to see the individual cars?

- what can we do?

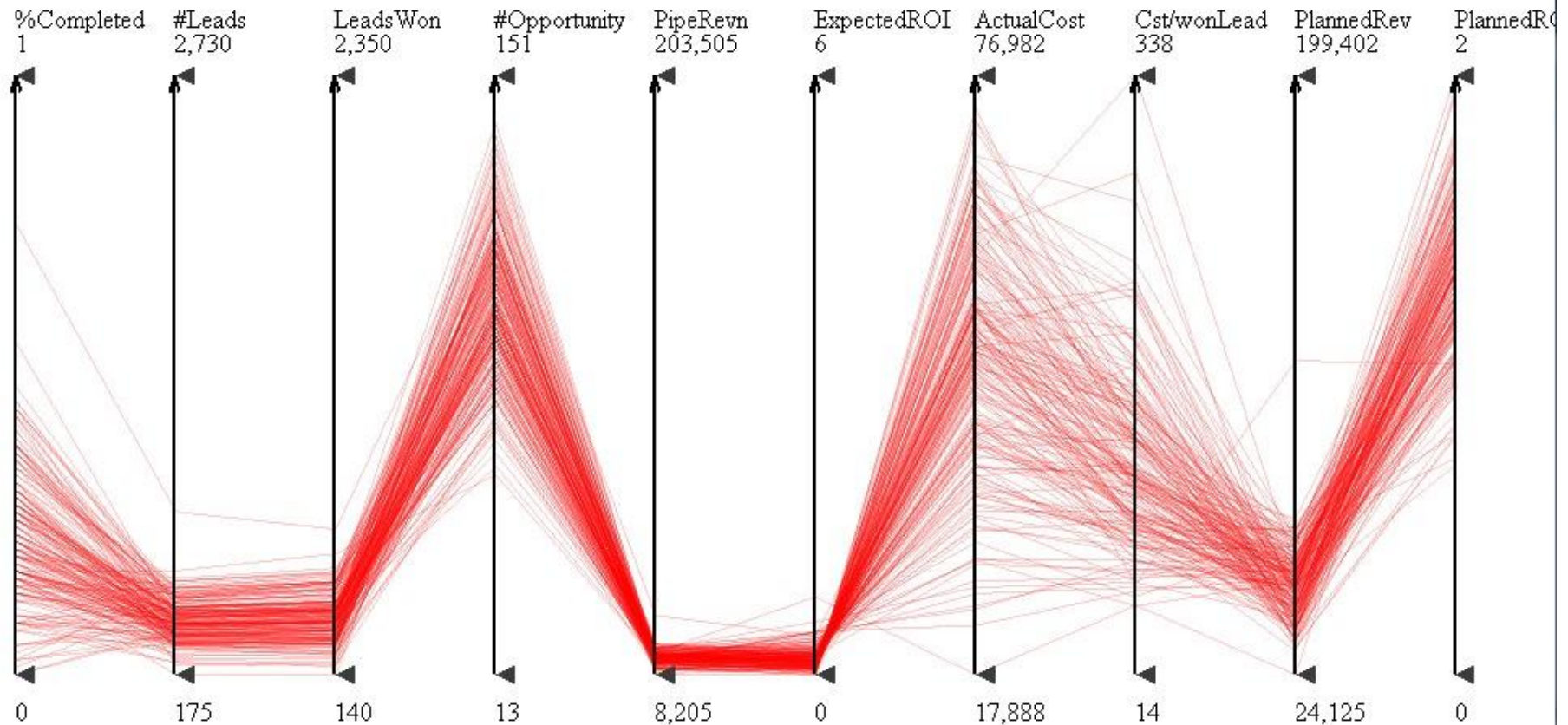
# Parallel Coordinates – 100 Cars



Grouping the cars into sub-populations

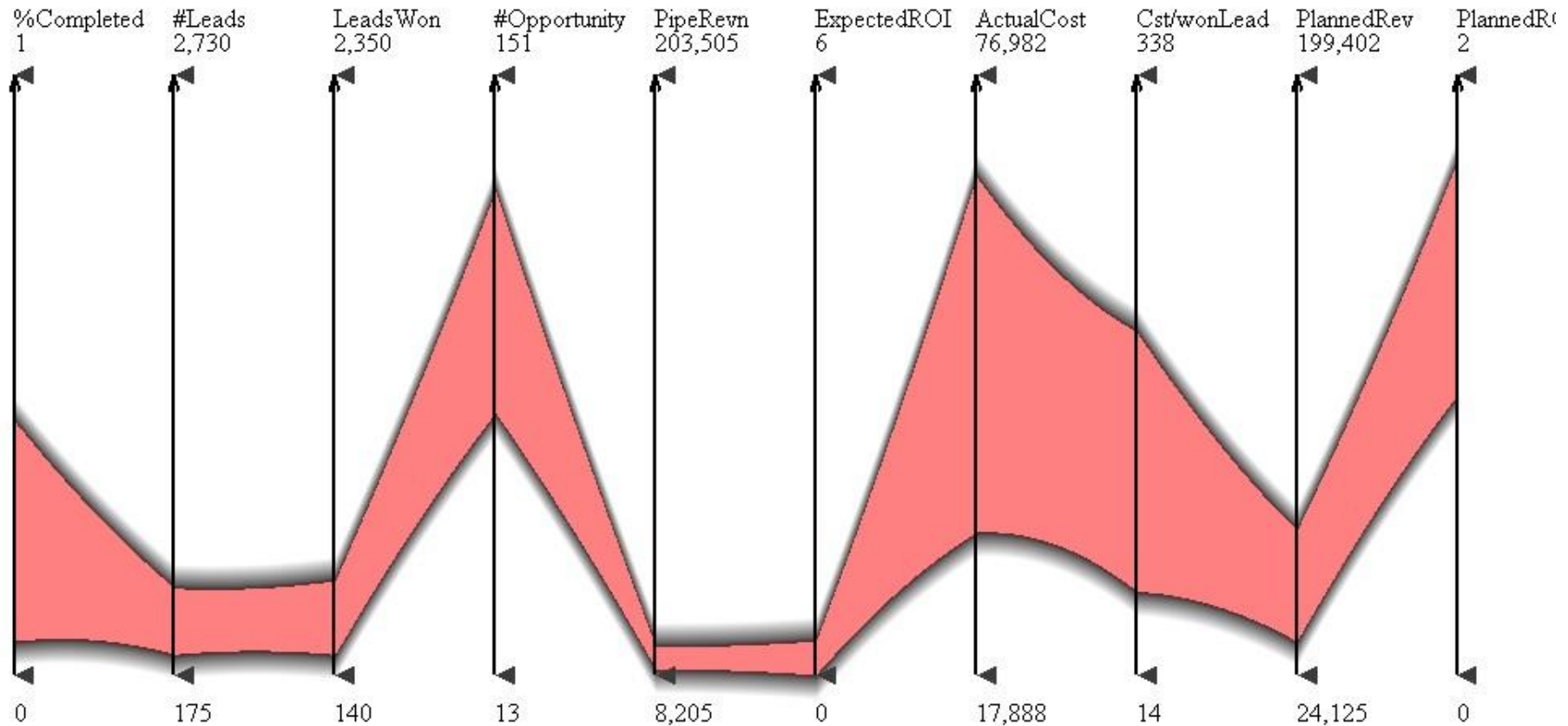
- we perform clustering
- can be automated or interactive (put the user in charge)

# PC With Illustrative Abstraction



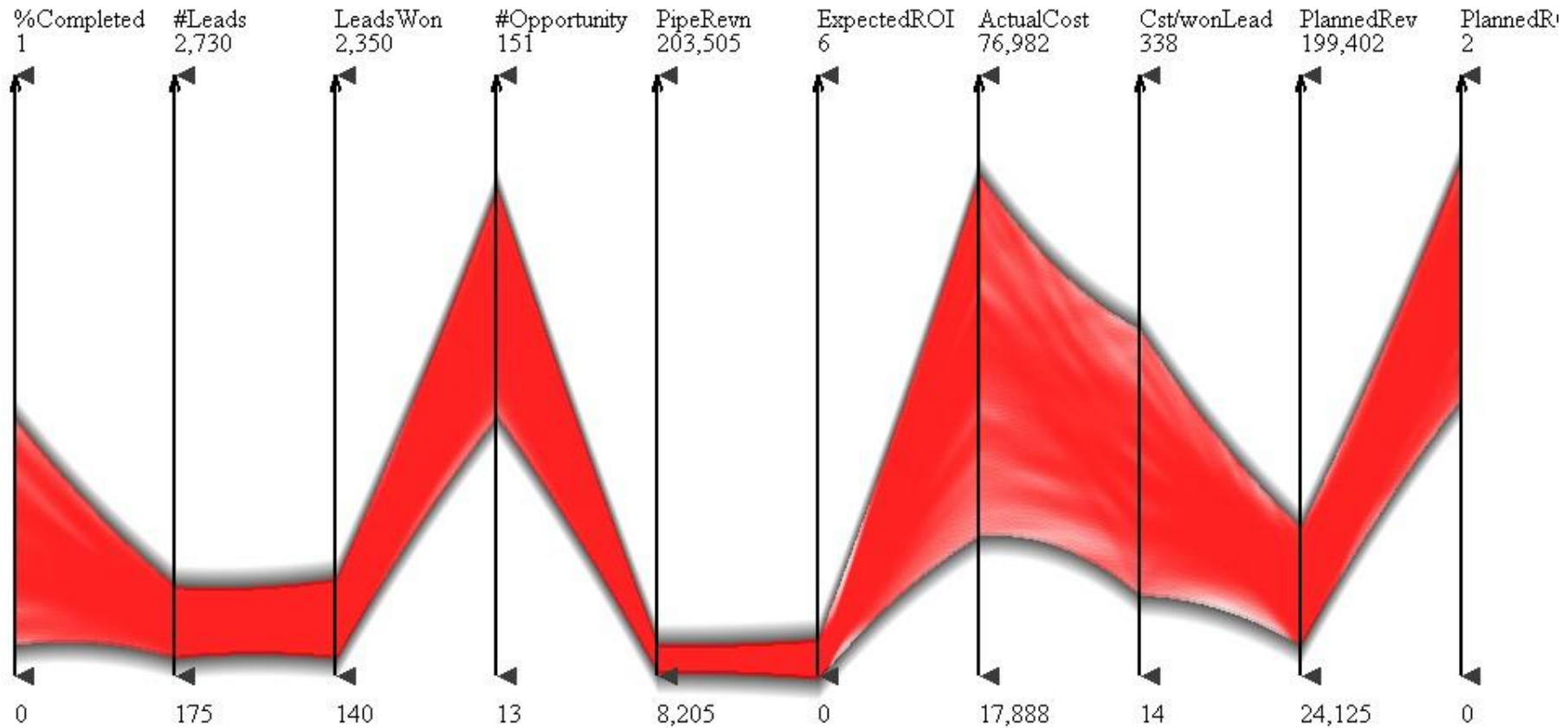
individual polylines

# PC With Illustrative Abstraction



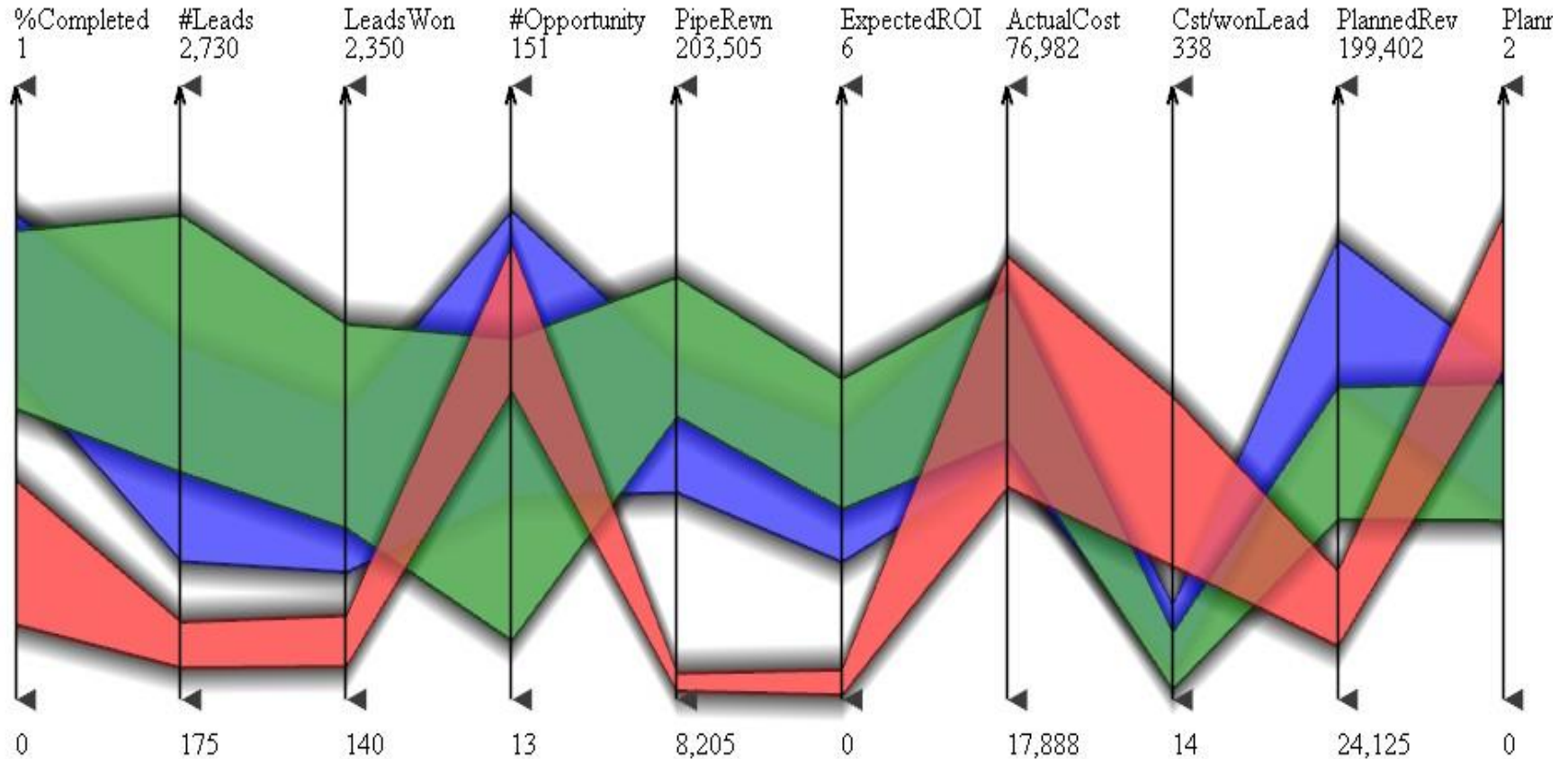
completely abstracted away

# PC With Illustrative Abstraction



blended partially

# PC With Illustrative Abstraction



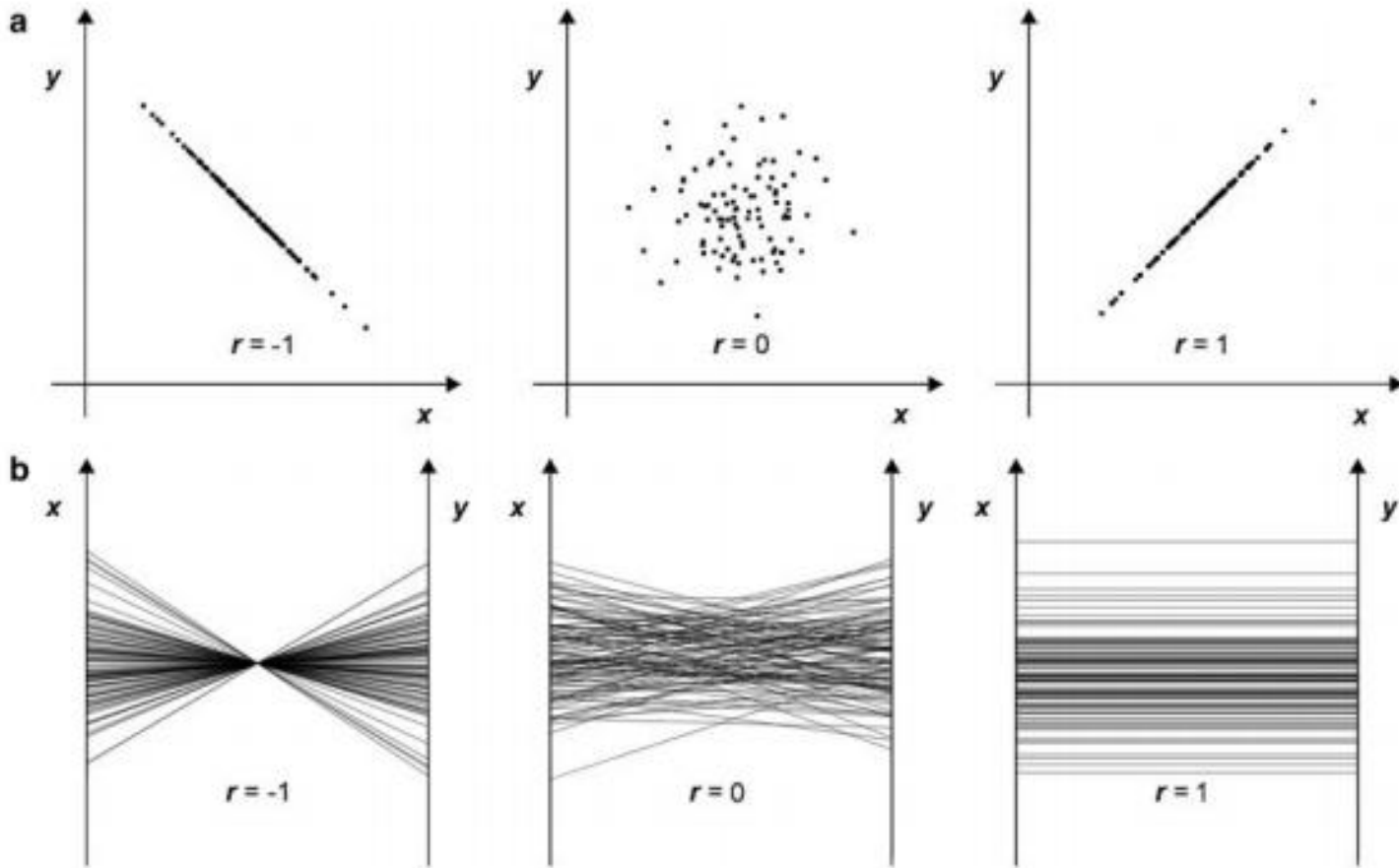
all put together – three clusters

[McDonnell and Mueller, 2008]

# Interaction is Key

Interaction in Parallel Coordinate

# Patterns in Parallel Coordinates



correlation

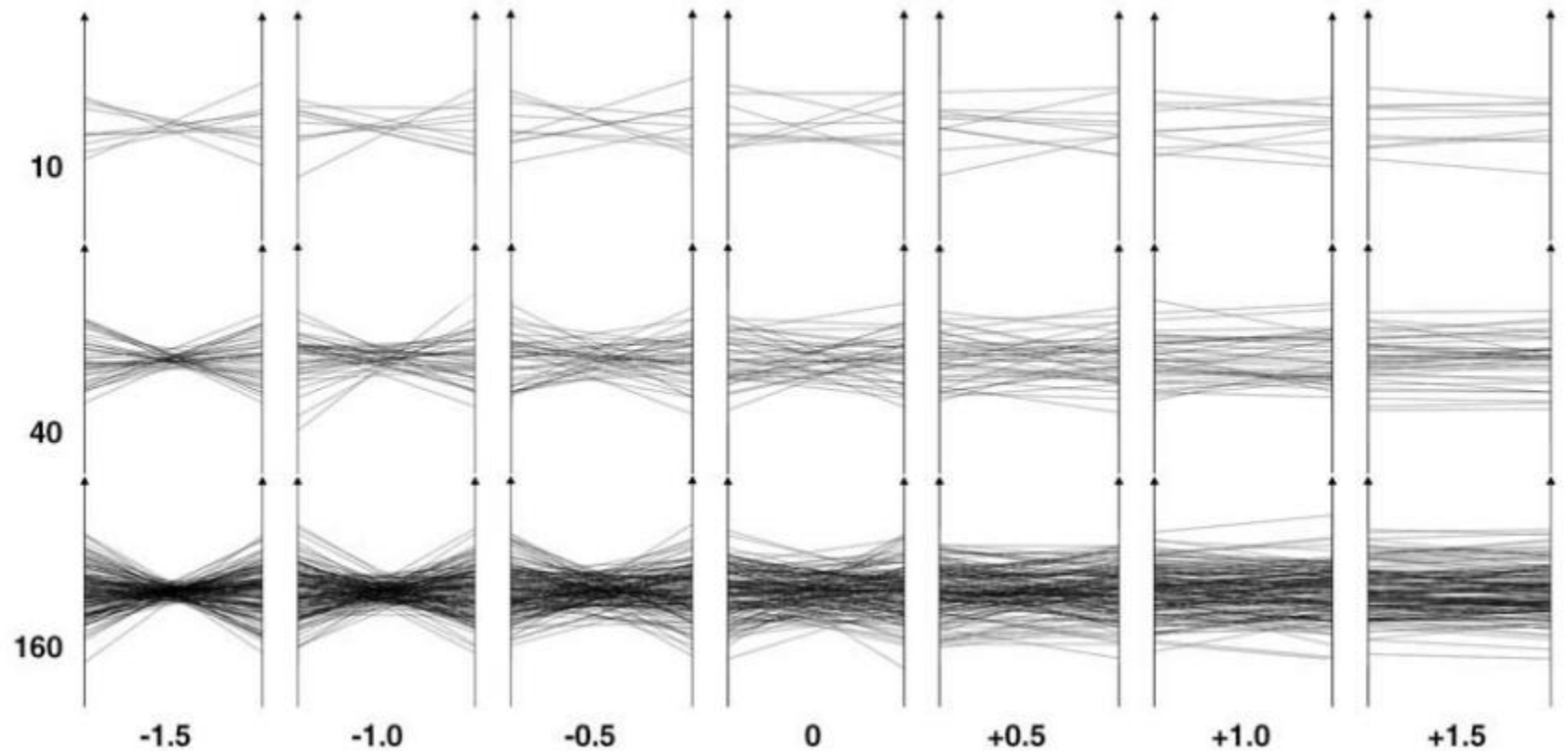
$r = -1.0$

$r = 0$

$r = 1.0$

# Patterns in Parallel Coordinates

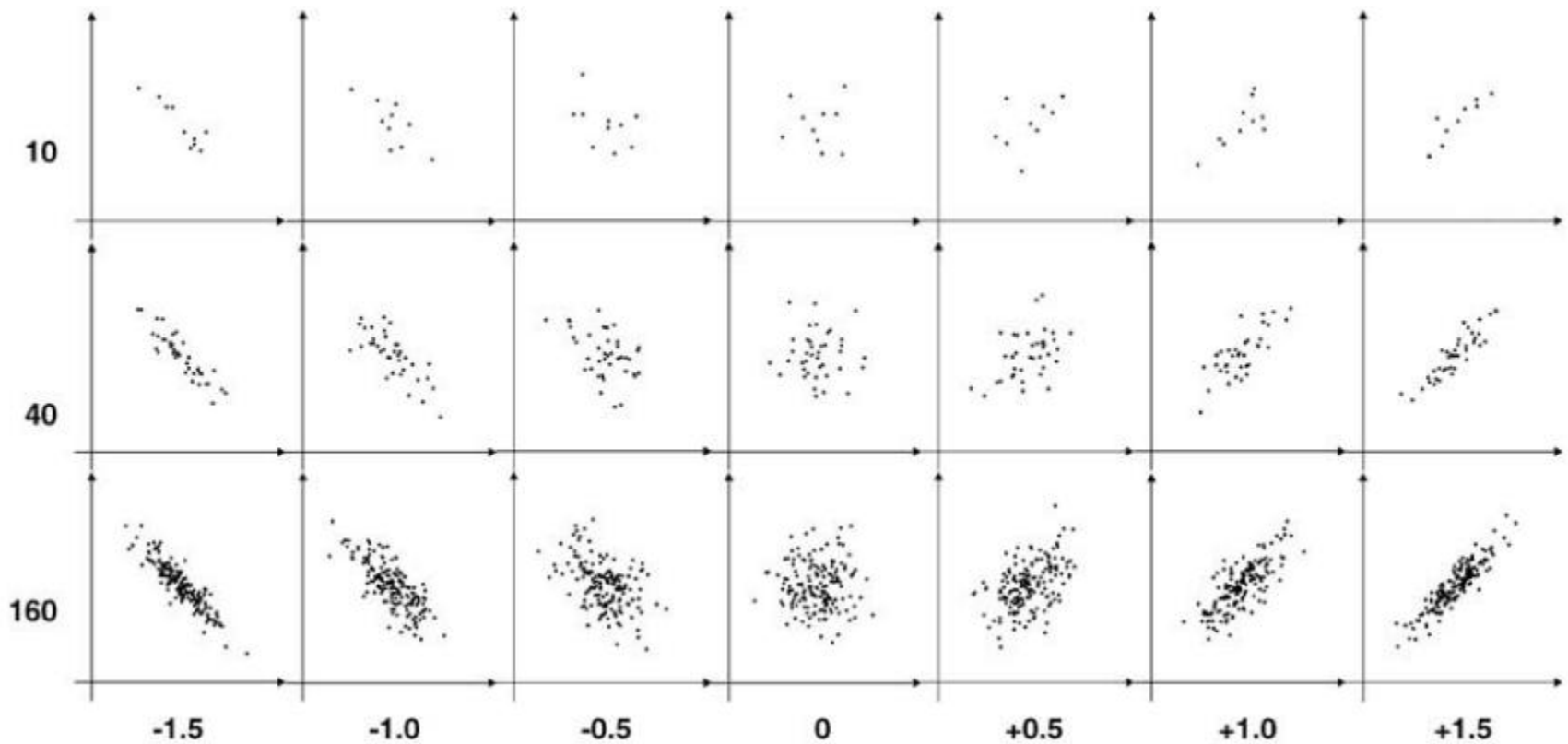
# points



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

# Patterns in Scatterplots

# points



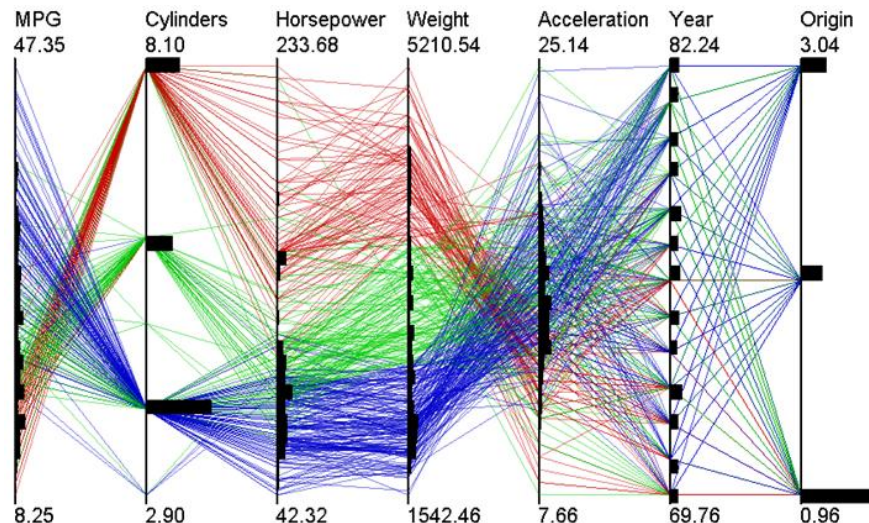
Fisher-z (corresponding to  $\rho = 0, \pm 0.462, \pm 0.762, \pm 0.905$ )

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

# 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

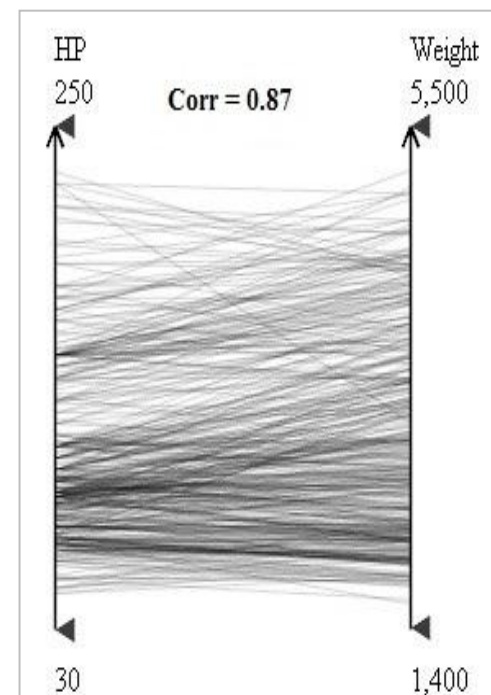
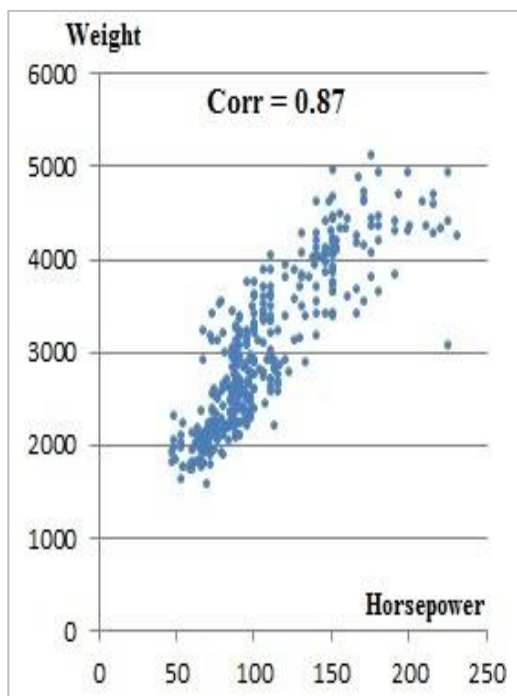


# We Need a Measure for Relationships

## 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

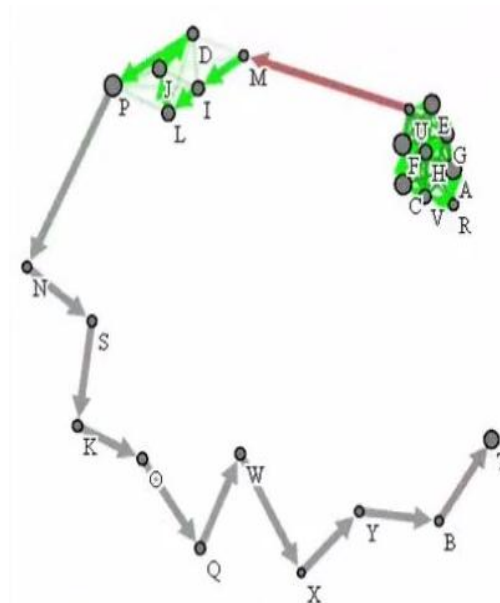
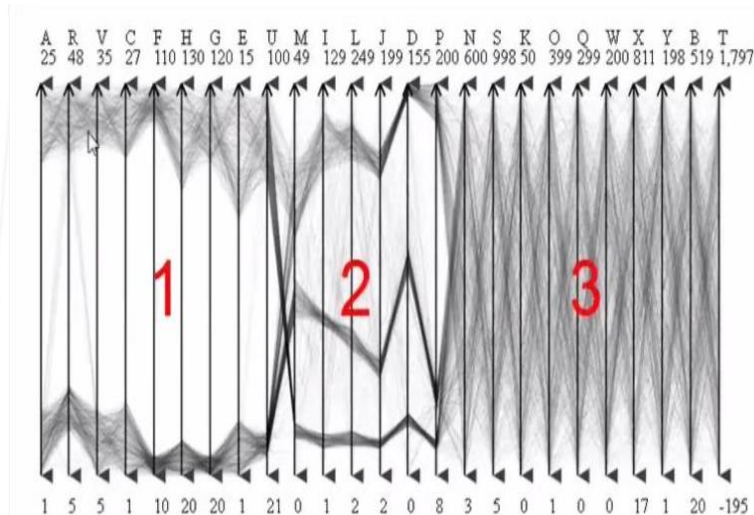
Create a correlation matrix

Run a mass-spring model

Run Traveling Salesman on the correlation nodes

Use it to order your parallel coordinate axes via TSP

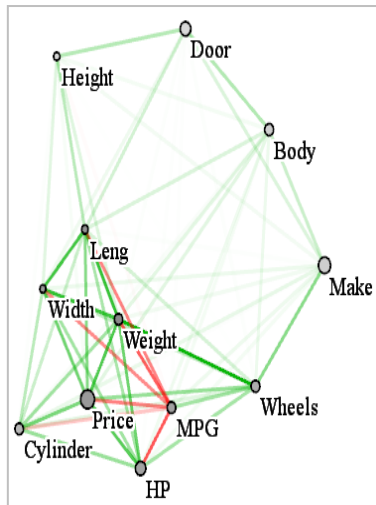
Z. Zhang, K. McDonnell, K. Mueller, "A Network-Based Interface for the Exploration of High-Dimensional Data Spaces, " *IEEE Pacific Vis*, 2012



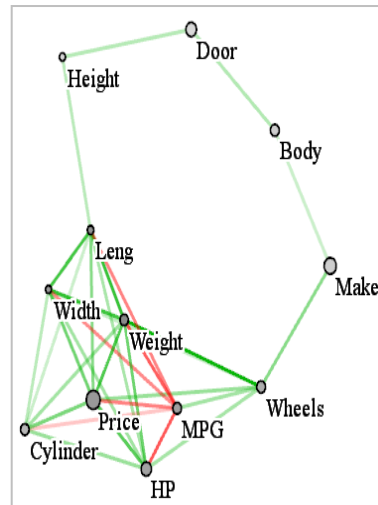
|      | MRK   | MSFT   | PFE    | PG    | T      | TRV    | UTX    | VZ    | WMT    | XOM   |
|------|-------|--------|--------|-------|--------|--------|--------|-------|--------|-------|
| MRK  | 1     | 0.39   | 0.72   | -0.43 | 0.57   | 0.031  | -0.26  | 0.61  | -0.11  | -0.25 |
| MSFT | 0.39  | 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  | -0.067 | 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 | -0.067 | -0.033 | 0.026 | 1      | 0.832 |
| XOM  | -0.25 | 0.24   | -0.72  | 0.9   | 0.54   | 0.81   | 0.93   | 0.36  | 0.832  | 1     |

# Interaction with the Correlation Network

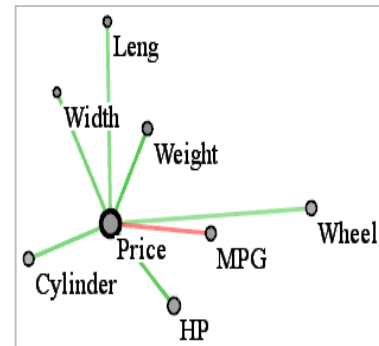
- Vertices are attributes, edges are correlations
  - vertex: size determined by  $\sum_{j=0}^D \frac{|correlation(i,j)|}{D-1} \quad j \neq i$
  - edge length is a measure of  $(1-|correlation|)$
  - edge: color/intensity  $\rightarrow$  sign/strength of correlation



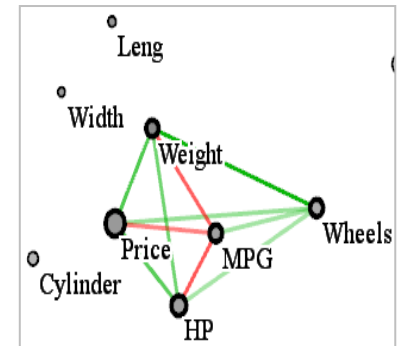
all edges



filtered by strength

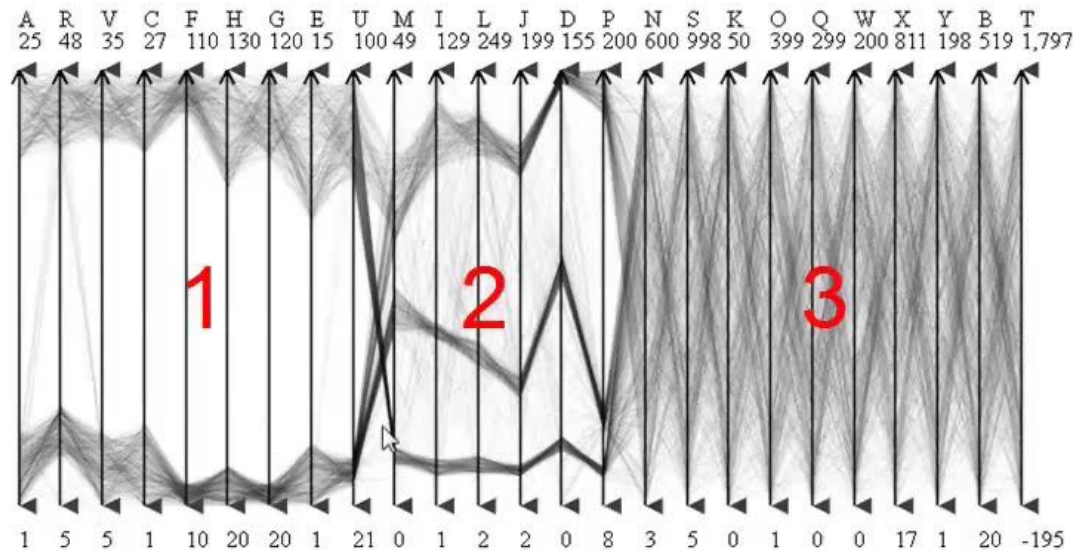


attribute centric

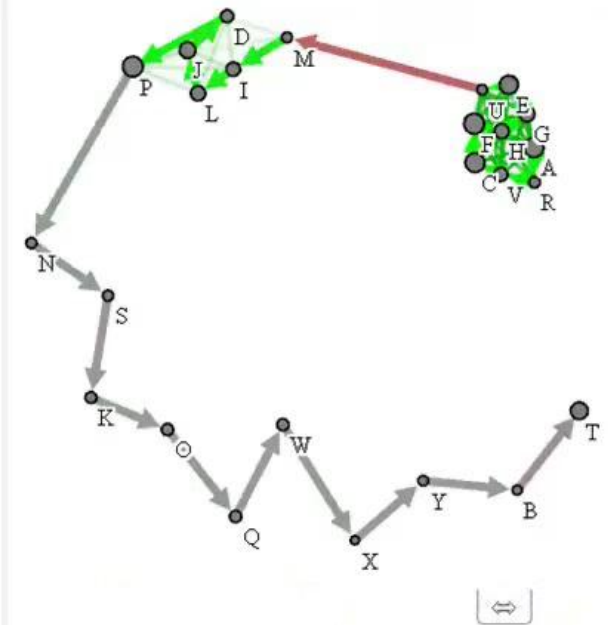


subset of attributes

# Multiscale Zooming



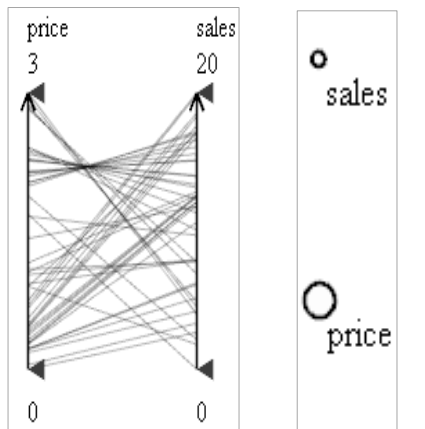
3 subspaces are well separated.



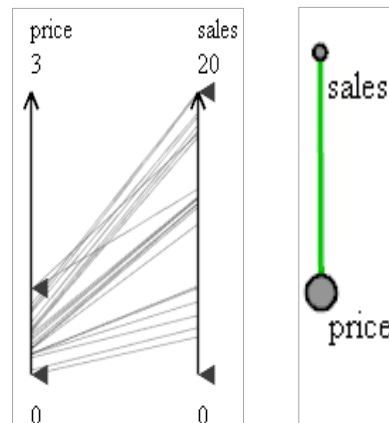
# Bracketing and Conditioning

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

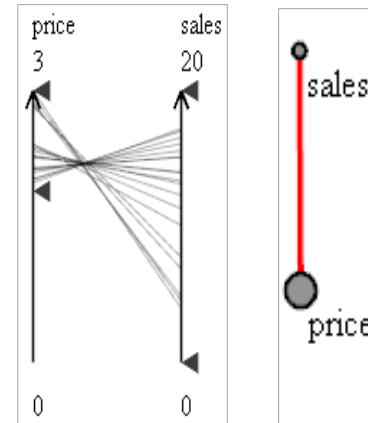
- this limits the derived relationships to this value range
- such limits are commonplace in targeted marketing, etc.



no bracketing



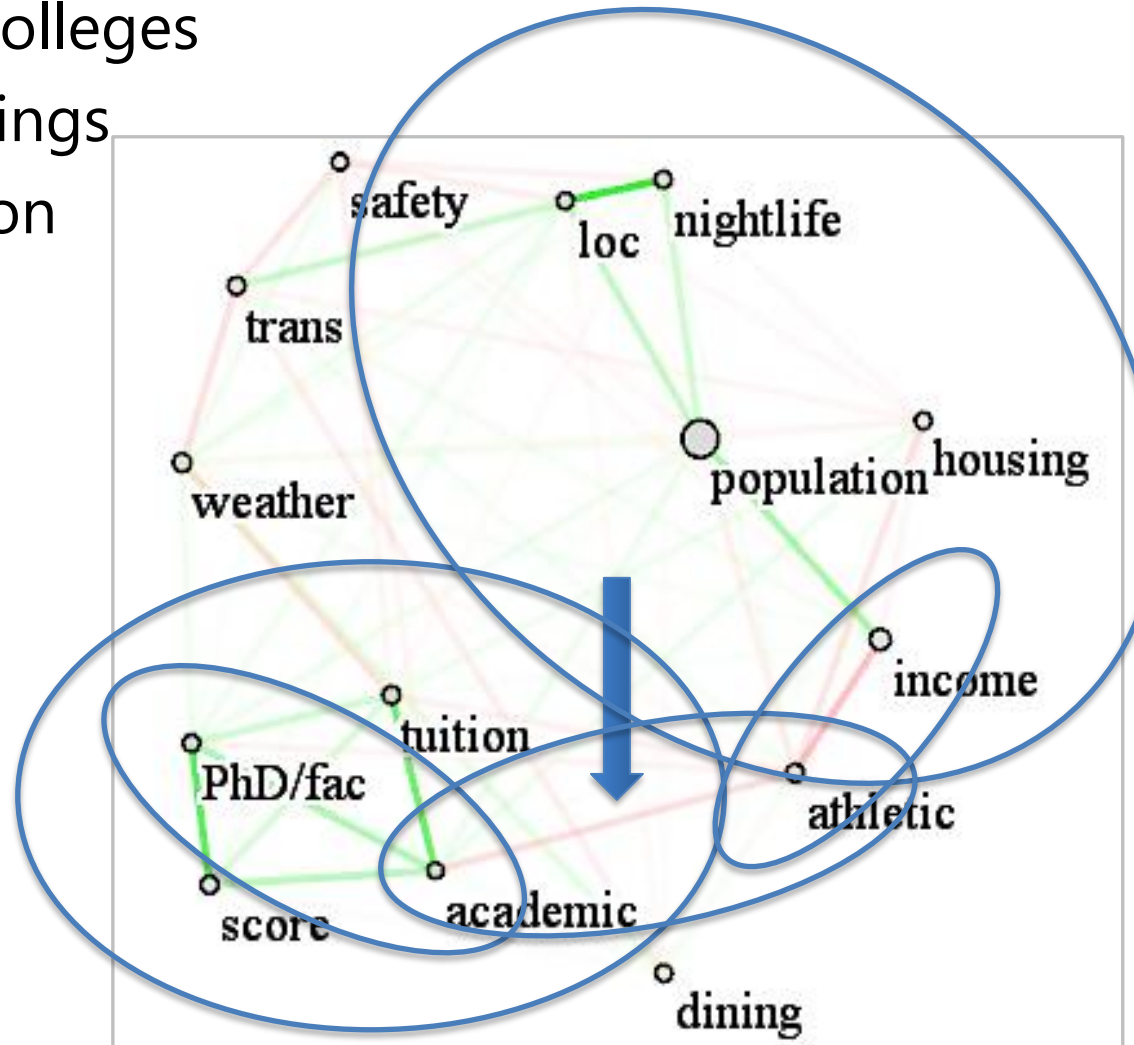
lower price range



higher price range

# Correlation Plots Are Powerful

Fused dataset of 50 US colleges  
US News: academic rankings  
College Prowler: survey on  
campus life attributes

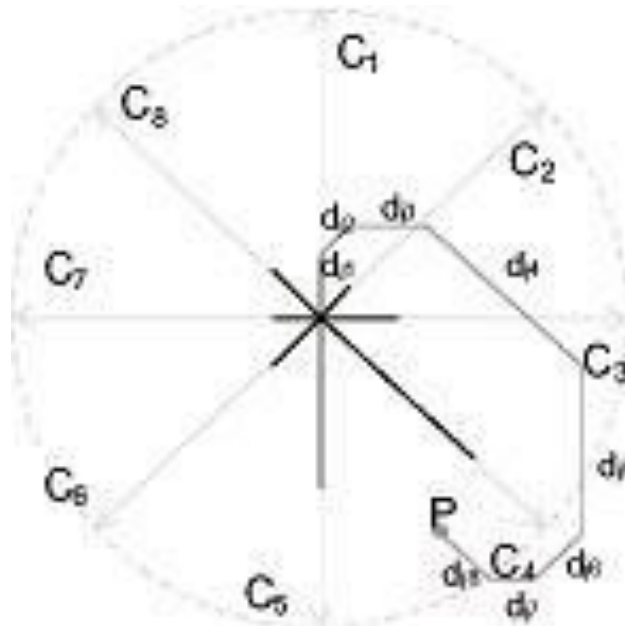


# Radial Layouts

# Star Coordinates

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

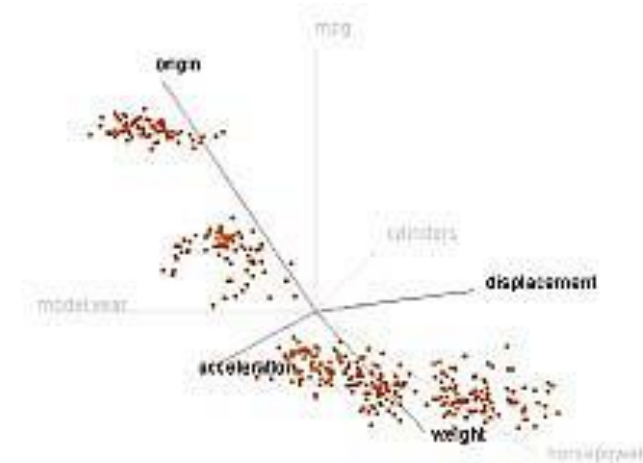
- no prior PCA and subsequent projection
- instead, a point  $P$  is plotted as a vector sum of all axis coordinates



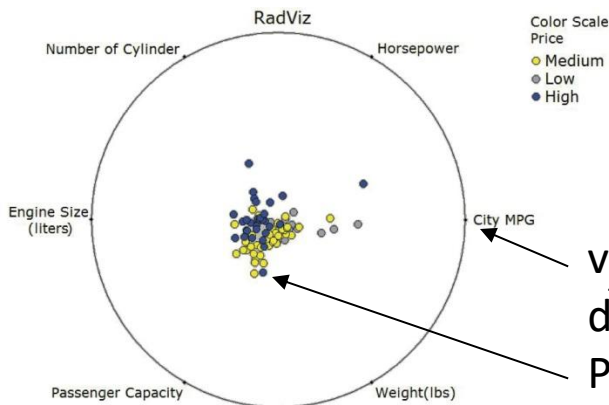
# Star Coordinates

## 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



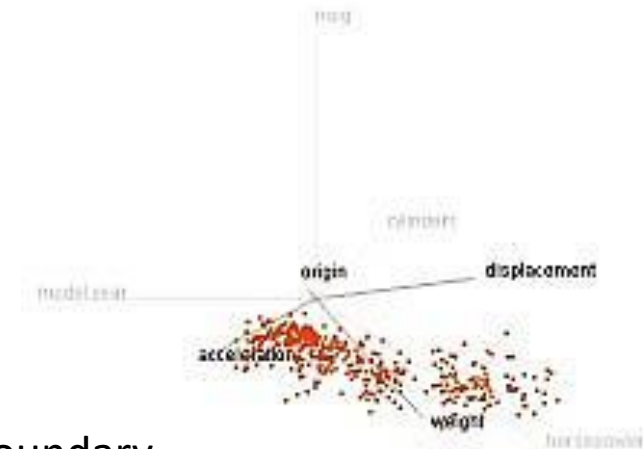
$$P = \sum_{i=1}^n w_j v_j$$

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

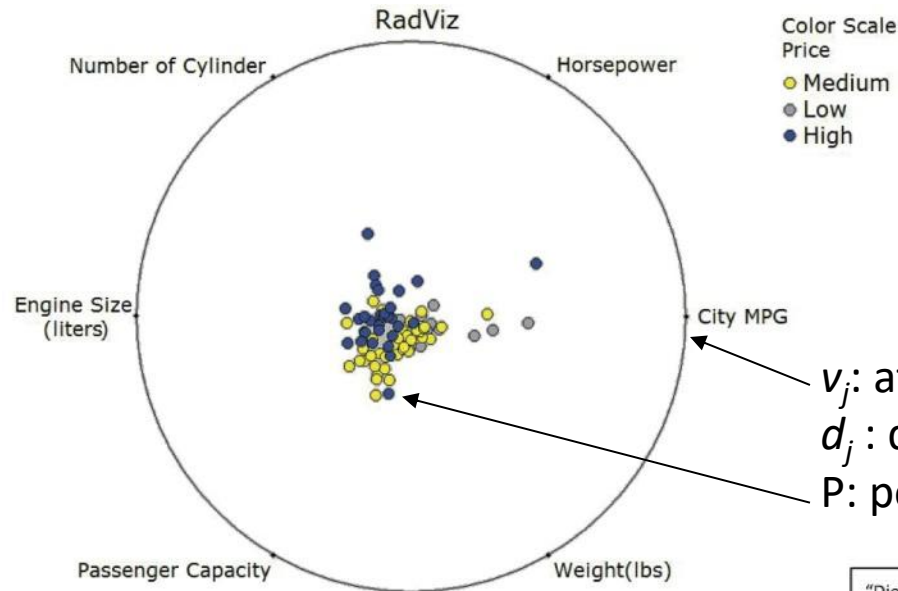
$v_j$ : attribute coordinates on disk boundary

$d_j$ : data vector values

$P$ : point location in RadViz disk



# RadViz

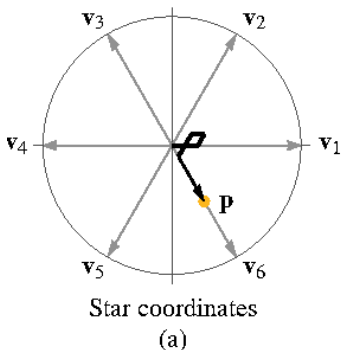


$$P = \sum_{i=1}^n w_i v_i$$

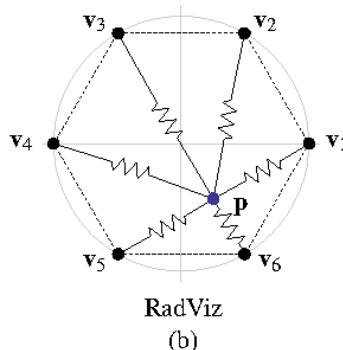
$$w_i = d_i / \sum_{k=1}^n d_k$$

$v_j$ : attribute coordinates on disk boundary  
 $d_j$ : data vector values  
 $P$ : point location in RadViz disk

$$\frac{x}{1^T x} = (0.2, 0.1, 0, 0.1, 0.2, 0.4)$$

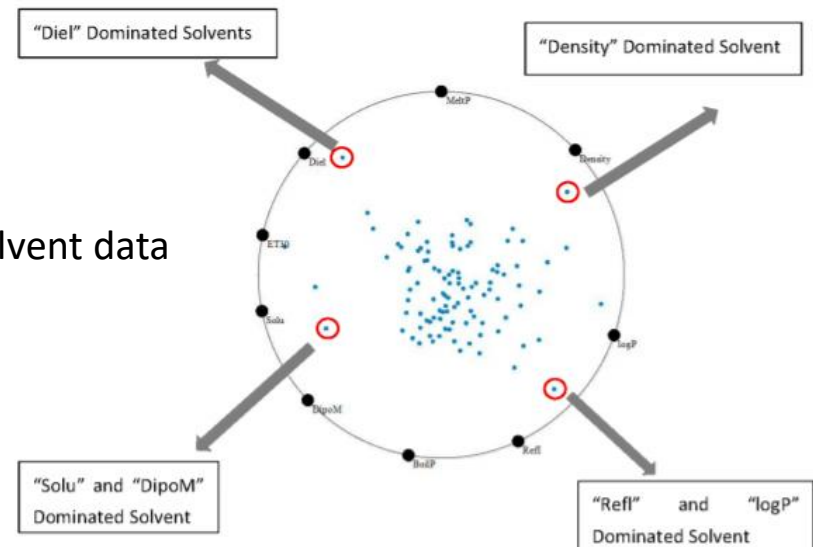


$$x = (0.5, 0.25, 0, 0.25, 0.5, 1)$$



Comparison with Star-coordinates

Solvent data



# Radar Chart

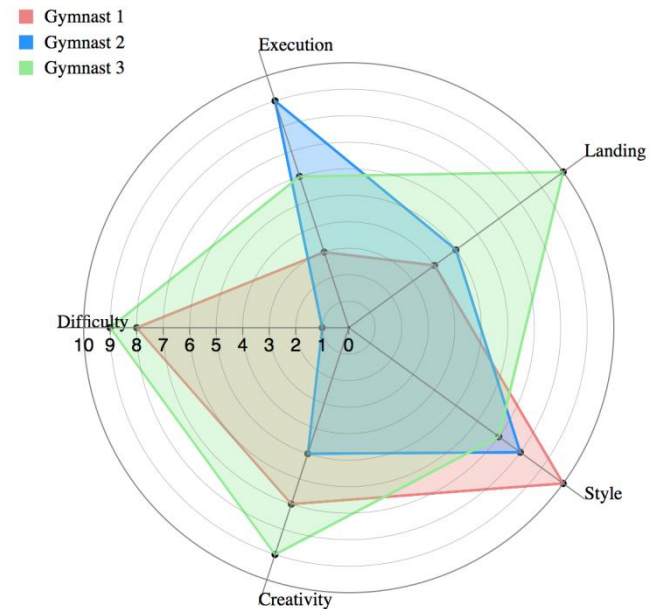
Equivalent to a parallel coordinates plot, with the axes arranged radially

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

## Disadvantages

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

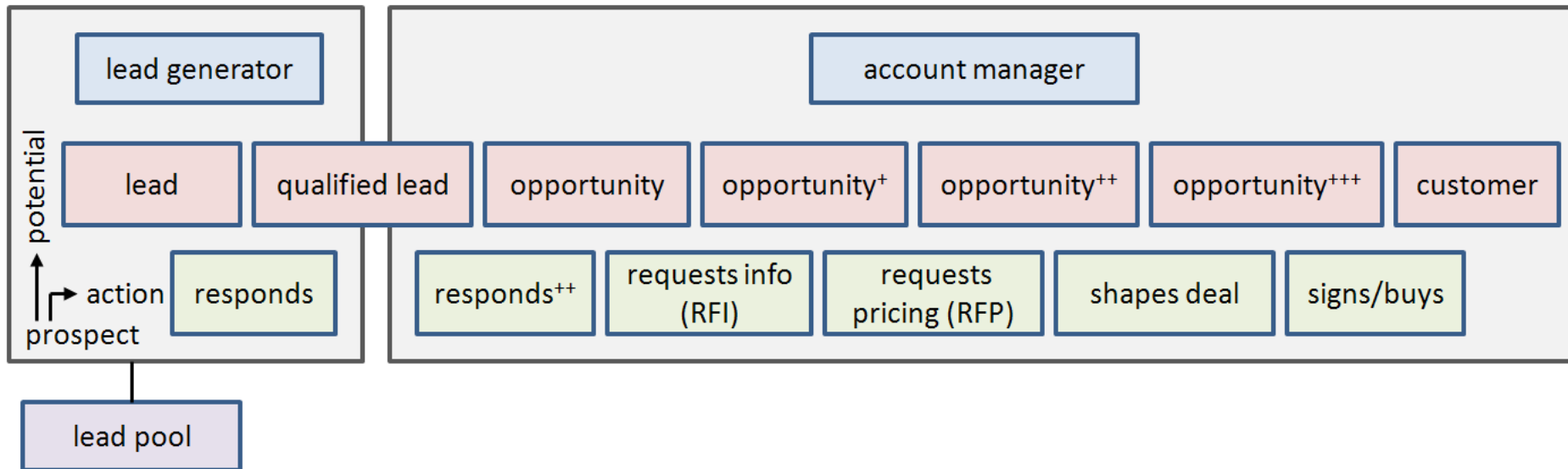
**Gymnast Scoring Radar Chart**



# 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

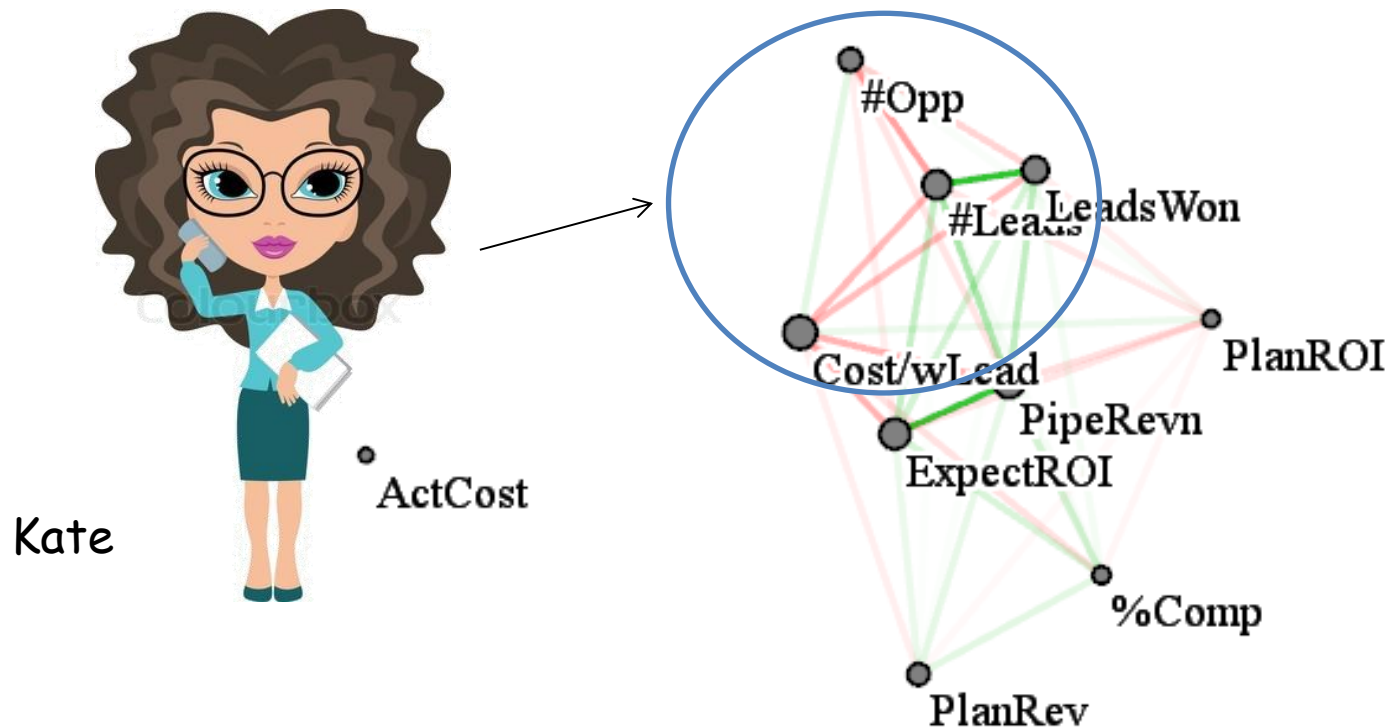
## Evidence:

- 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:

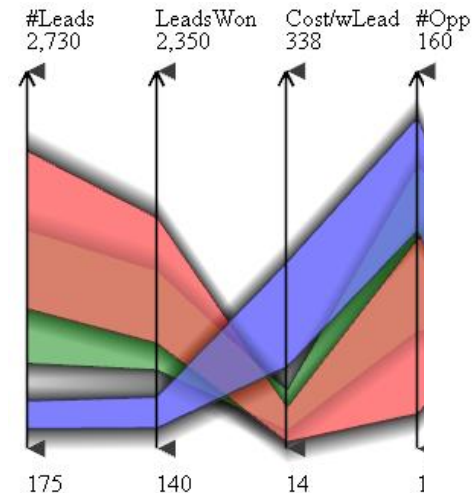
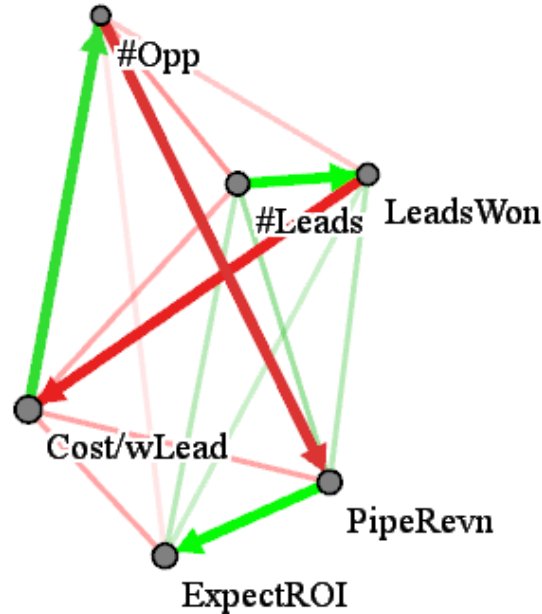
"OK...let's see, cost/won lead is nearby and it has a positive correlation with #opportunities but also a negative correlation with #won leads"



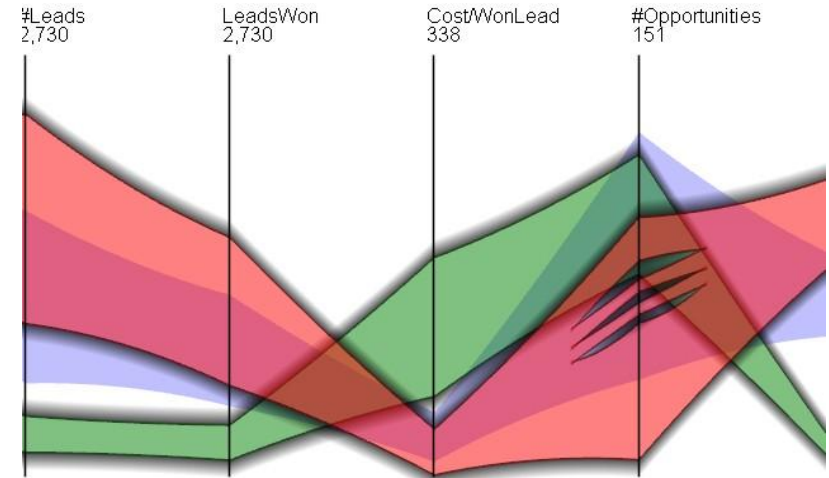
# Kate Designs the Narration

“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"

Next Lecture:  
Non-Linear Projection Techniques  
(Embeddings)