CSE 564
Visualization & Visual Analytics

High-Dimensional Data

Klaus Mueller

Computer Science Department
Stony Brook University and SUNY Korea

Lecture	Topic	Projects
1	Intro, schedule, and logistics	
2	Applications of visual analytics, basic tasks, data types	
3	Introduction to D3, basic vis techniques for non-spatial data	Project #1 out
4	Data assimilation and preparation	
5	Data reduction and notion of similarity and distance	
6	Visual perception and cognition	
7	Visual design and aesthetics	Project #1 due
8	Dimension reduction	Project #2 out
9	Data mining techniques: clusters, text, patterns	
10	Cluster analysis: numerical data	
11	Cluster analysis: categorical data	
12	Spatial data origins: medical imaging, scientific simulation	
13	Techniques to visualize spatial data: volume visualization	
14	Intro to GPU programming	
15	Techniques to visualize spatial data: flow visualization	Project #3 out
16	Midterm #1	Project #2 due
17	Illustrative rendering	
18	High-dimensional data	Project #3 due
19	Correlation and causal modeling	
20	Principles of interaction	Final project proposal due
21	Visual analytics and the visual sense making process	
22	Evaluation and user studies	
23	Visualization of time-varying, time-series, streaming data	
24	Visualization of graph data	Final Project preliminary report due
25	Visualization of text data	
26	Midterm #2	
27	Data journalism	
	Final project presentations	Final Project slides and final report due

Understanding High-D Objects

Feature vectors are typically high dimensional

- this means, they have many elements
- high dimensional space is tricky
- most people do not understand it
- why is that?

- well, because you don't learn to see high-D when your vision system develops

Object permanence (Jean Piaget)

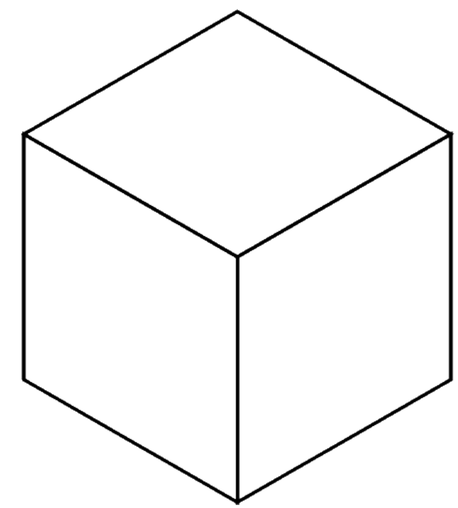
- the ability to create mental pictures or remember objects and people you have previously seen
- thought to be a vital precursor to creativity and abstract thinking

High-D Space is Tricky

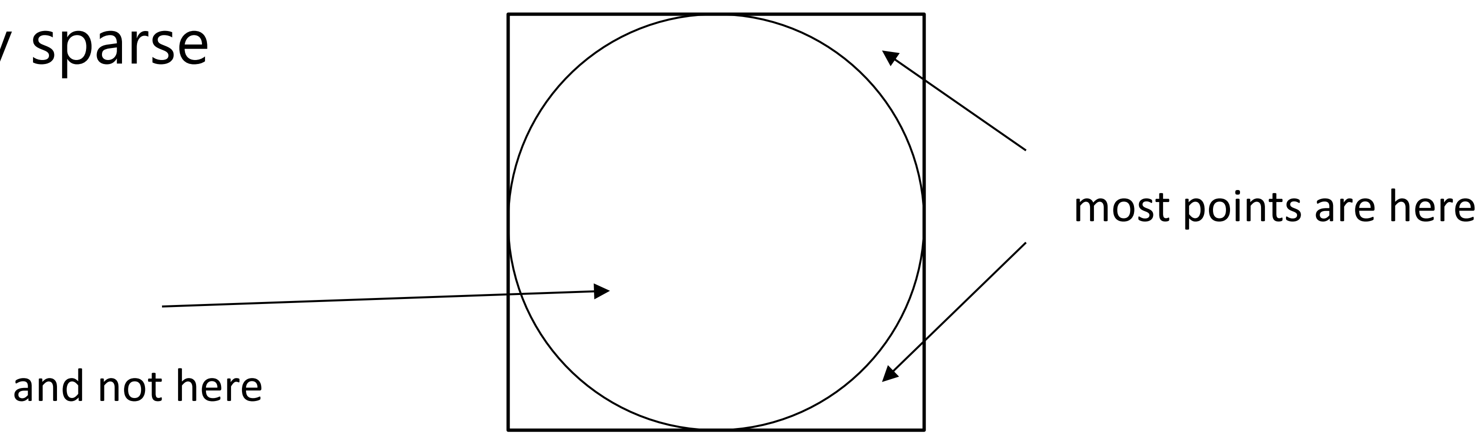
The curse of dimensionality

As $n \to \infty$

- Cube: side length l, diagonal d, volume V
- $V \to \infty$ for $l > 1$
- $V \to 0$ for $l < 1$
- $V = 1$ for $l = 1$
- $d \to \infty$

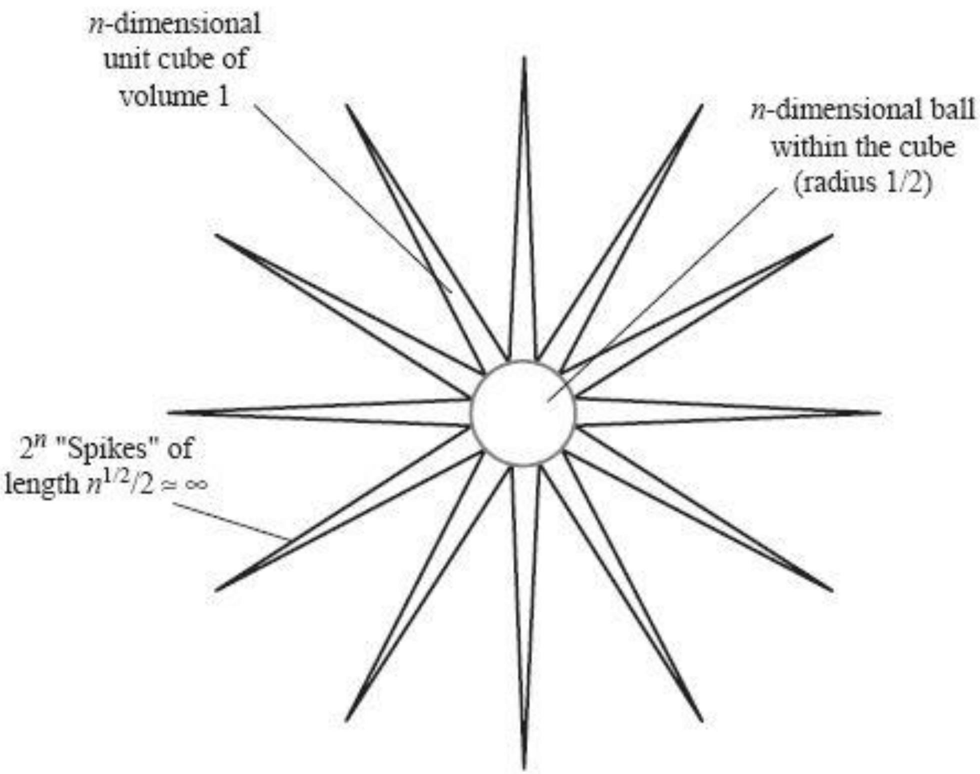


and very sparse



High-D Space is Tricky

Essentially hypercube is like a "hedgehog"



Curse of Dimensionality

Points are all at about the same distance from one another

- concentration of distances
- fundamental equation (Bellman, '61)

$$\lim_{n\to\infty} \frac{Dist_{\max} - Dist_{\min}}{Dist_{\min}} \to 0$$

- so as *n* increases, it is impossible to distinguish two points by (Euclidian) distance
  - unless these points are in the same cluster of points

Sparseness Demonstration

Space gets extremely sparse

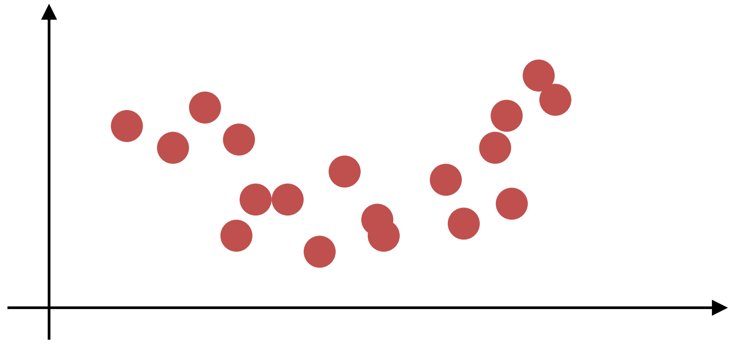
- with every extra dimension points get pulled apart further
- distances become meaningless

Sparseness Demonstration

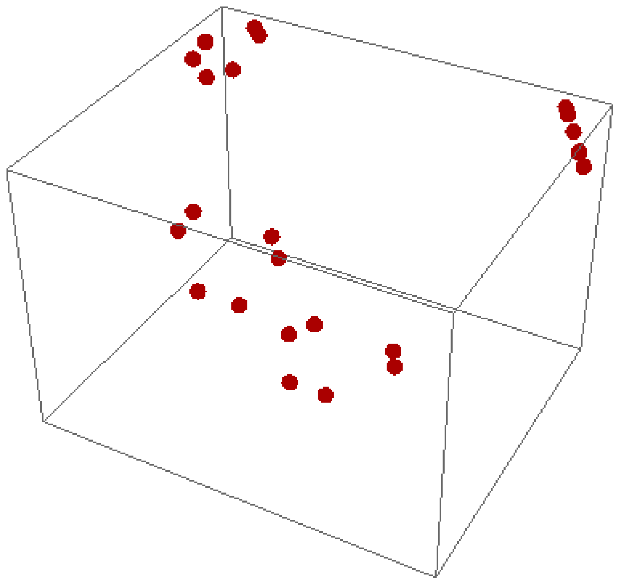
Space gets extremely sparse

- with every extra dimension points get pulled apart further
- distances become meaningless

1D – points are very close



2D – points spread apart



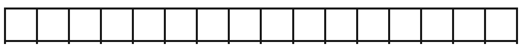
3D – getting even sparser

4D, 5D, … – sparseness grows further

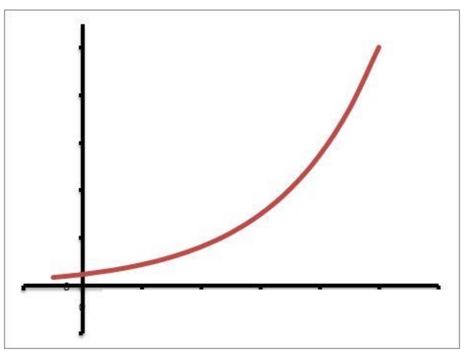
Space and Memory Management

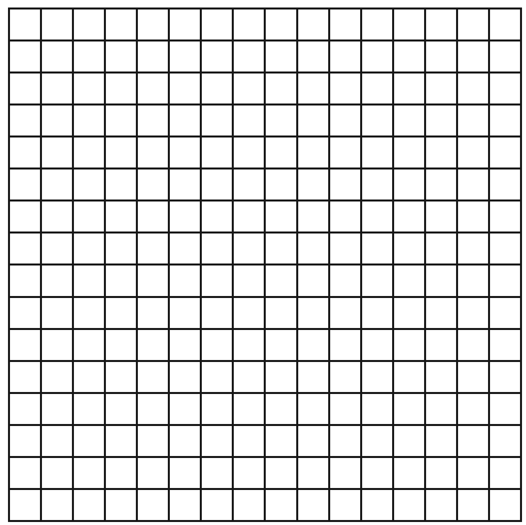
Indexing (and storage) also gets very expensive

- exponential growth in the number of dimensions

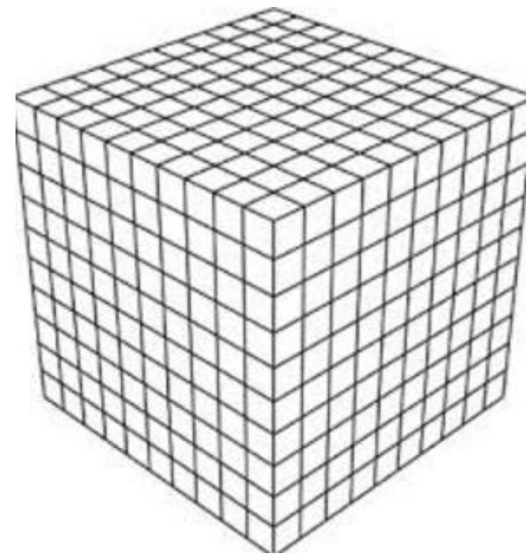


16 cells





16^2 = 256 cells

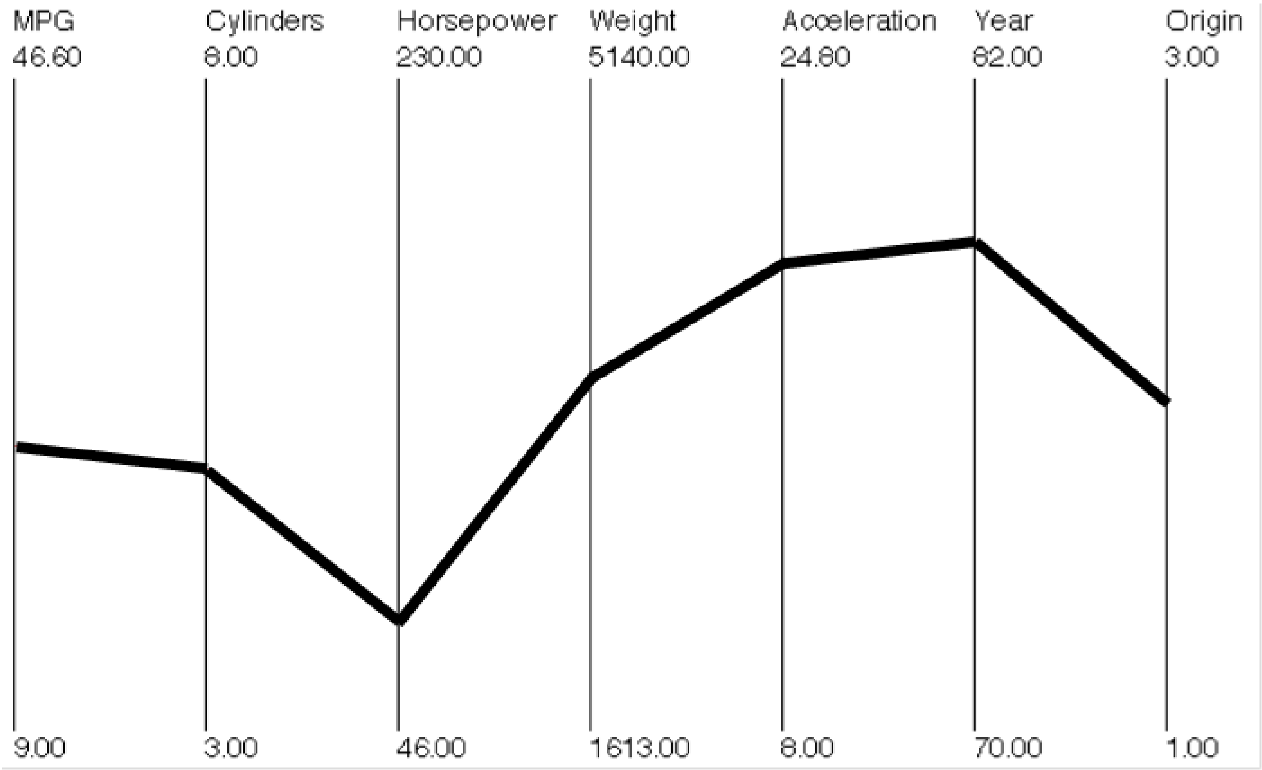


16^3 = 4,096 cells

- 4D: 65k cells   5D: 1M cells   6D: 16M cells   7D: 268M cells
- keep a keen eye on storage complexity

Recap: Parallel Coordinates

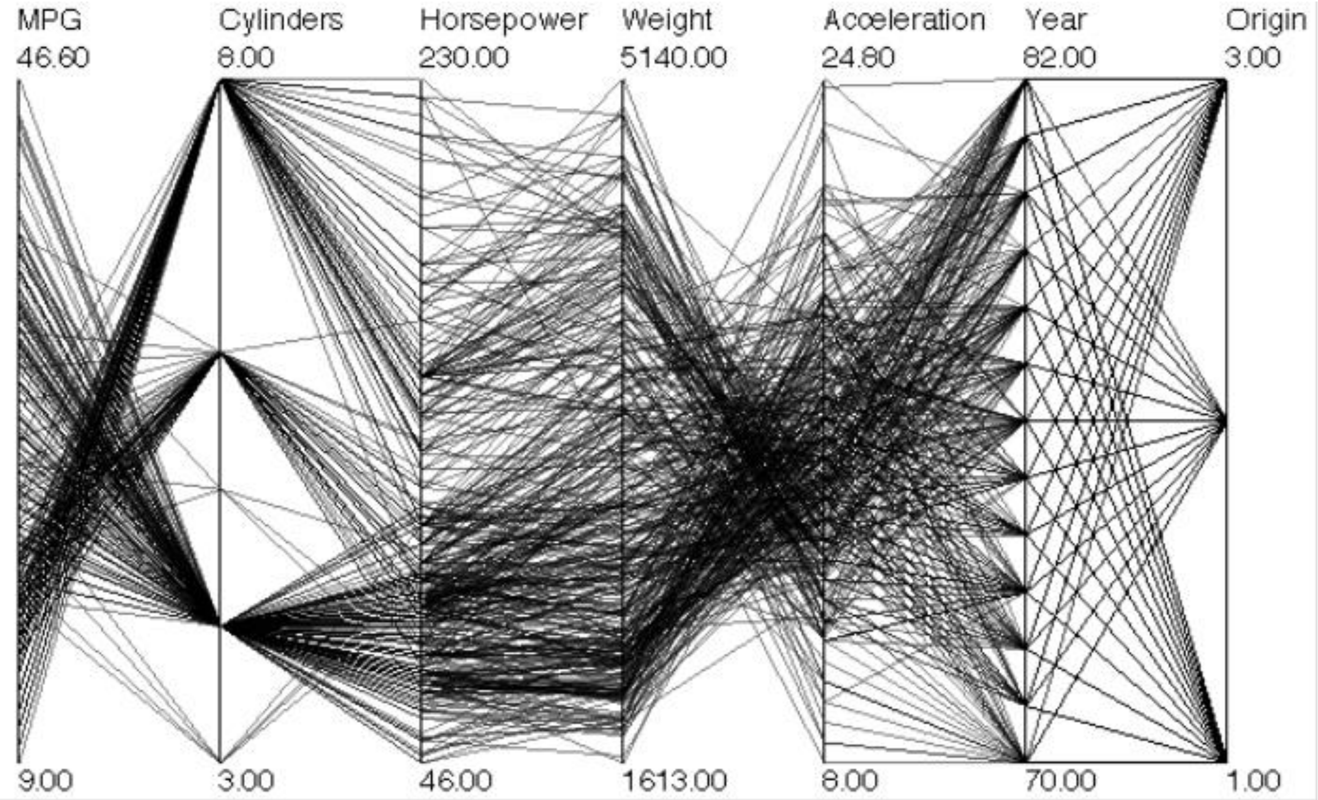
Parallel Coordinates – 1 Car



The N=7 data axes are arranged side by side

- in parallel

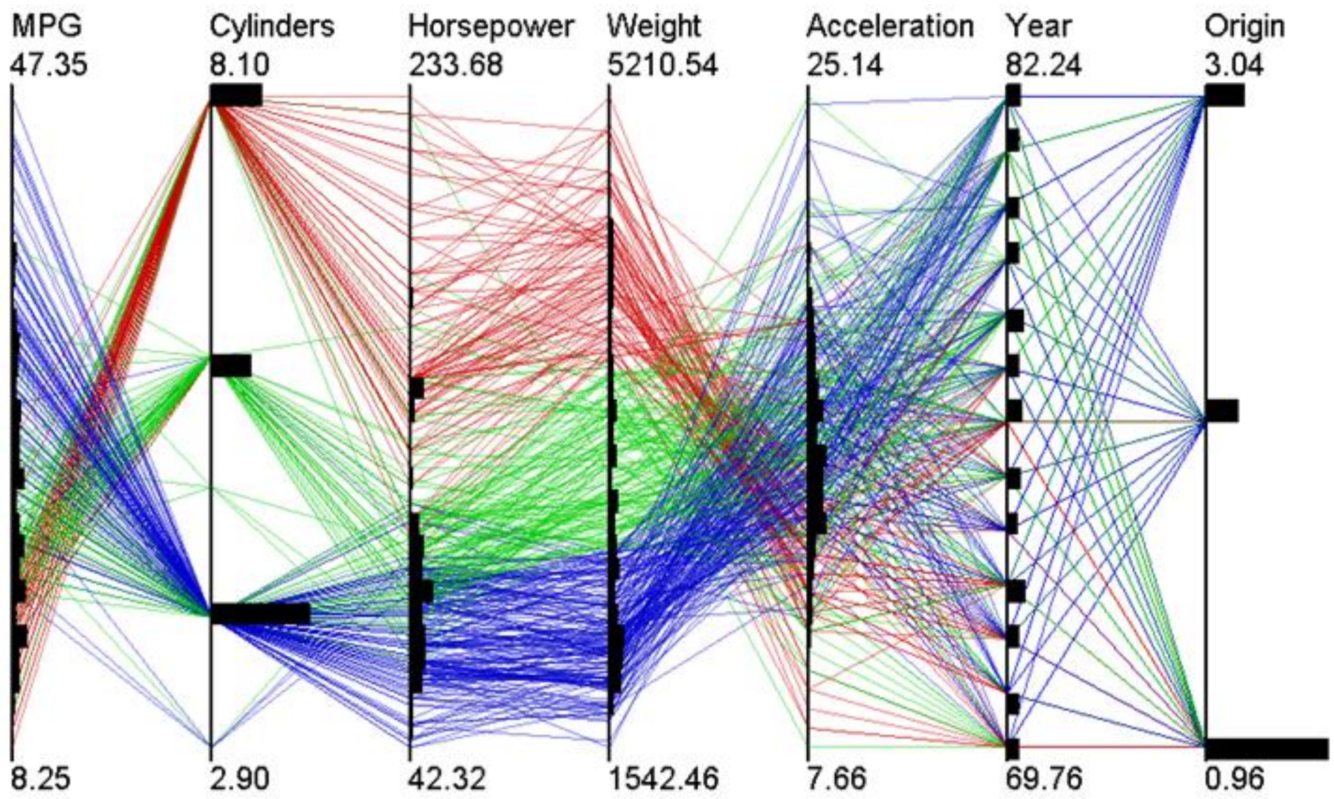
Parallel Coordinates – 100 Cars



Hard to see the individual cars?

- what can we do?

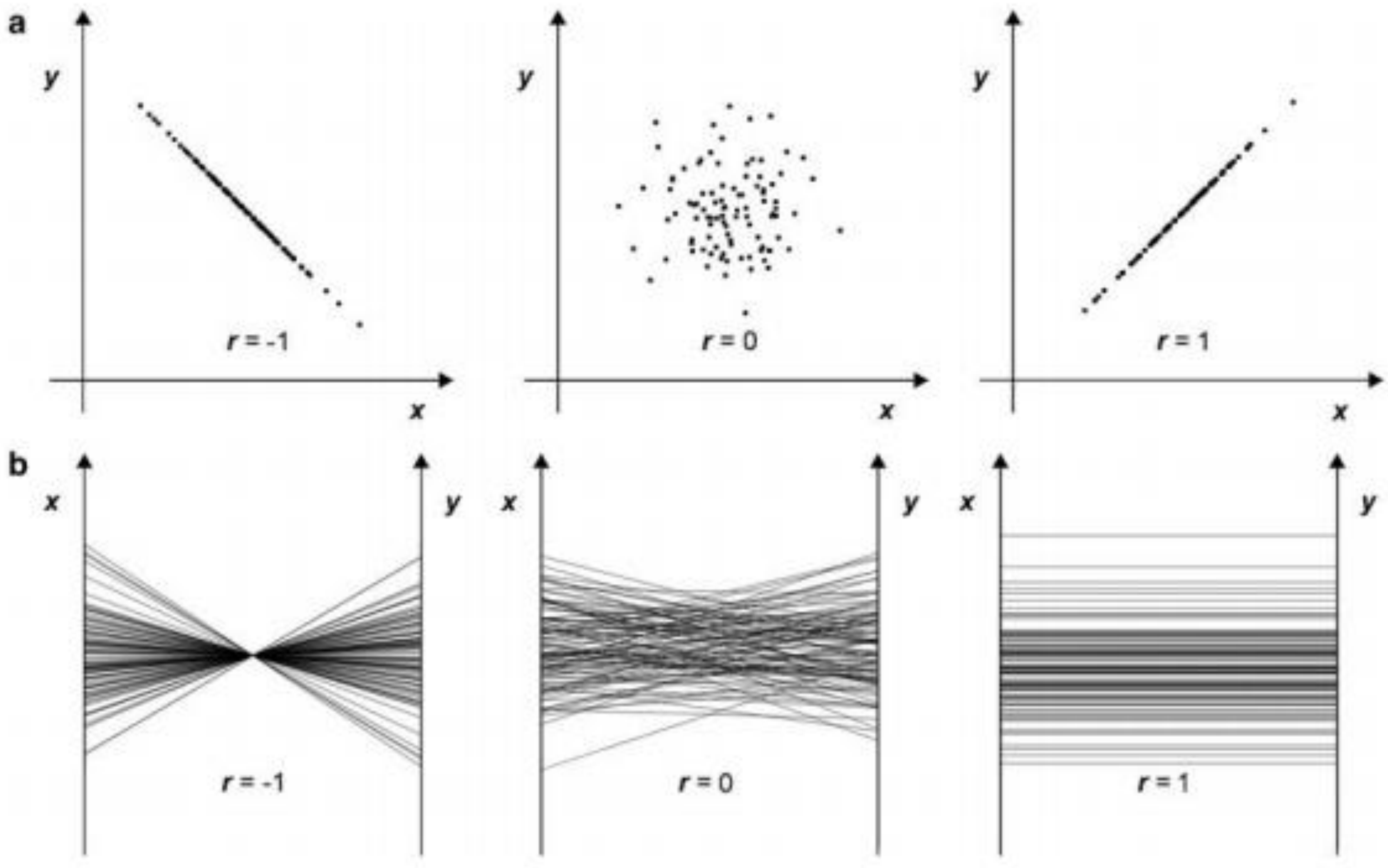
Parallel Coordinates – 100 Cars



Grouping the cars into sub-populations

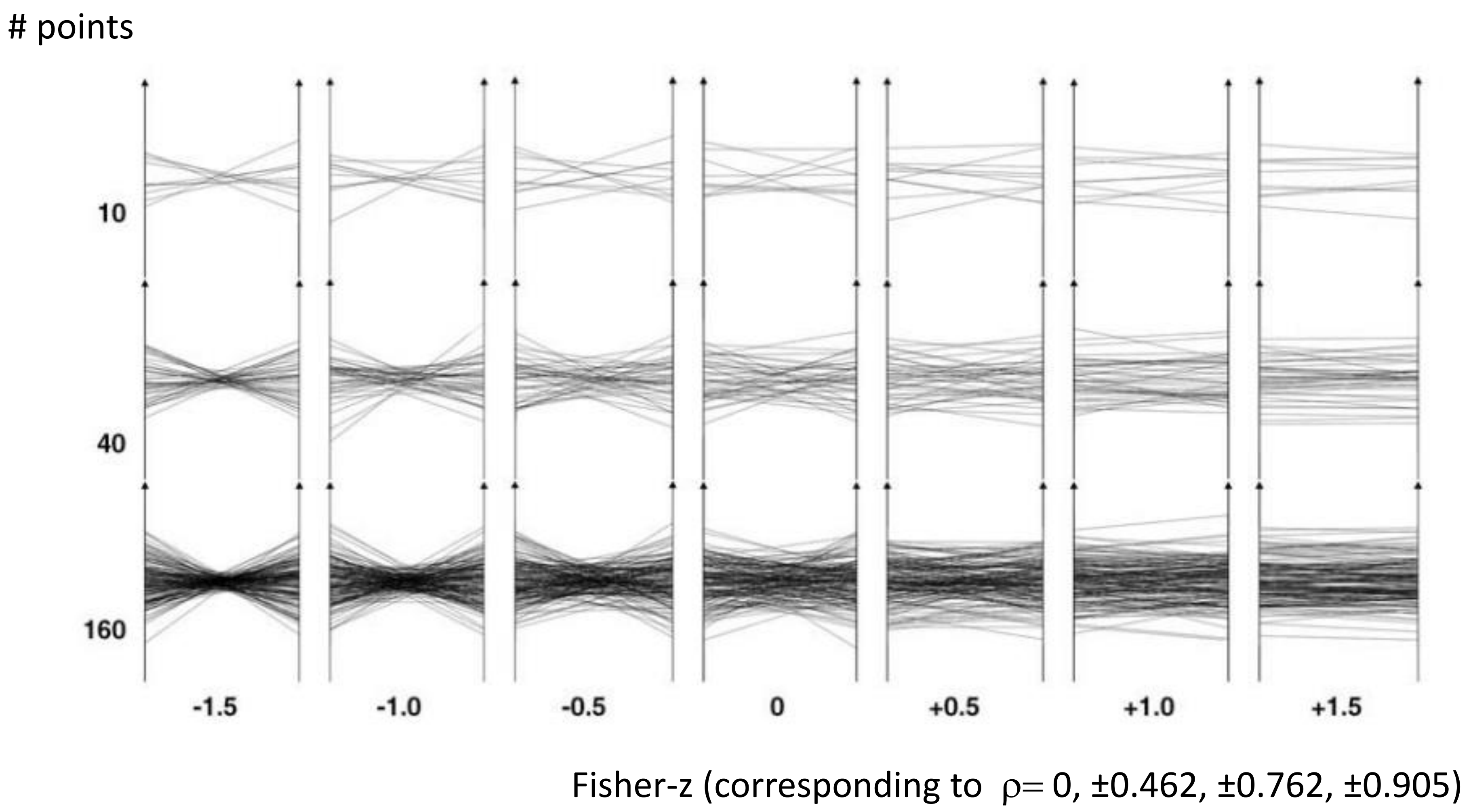
- we perform clustering
- an be automated or interactive (put the user in charge)

Patterns in Parallel Coordinates

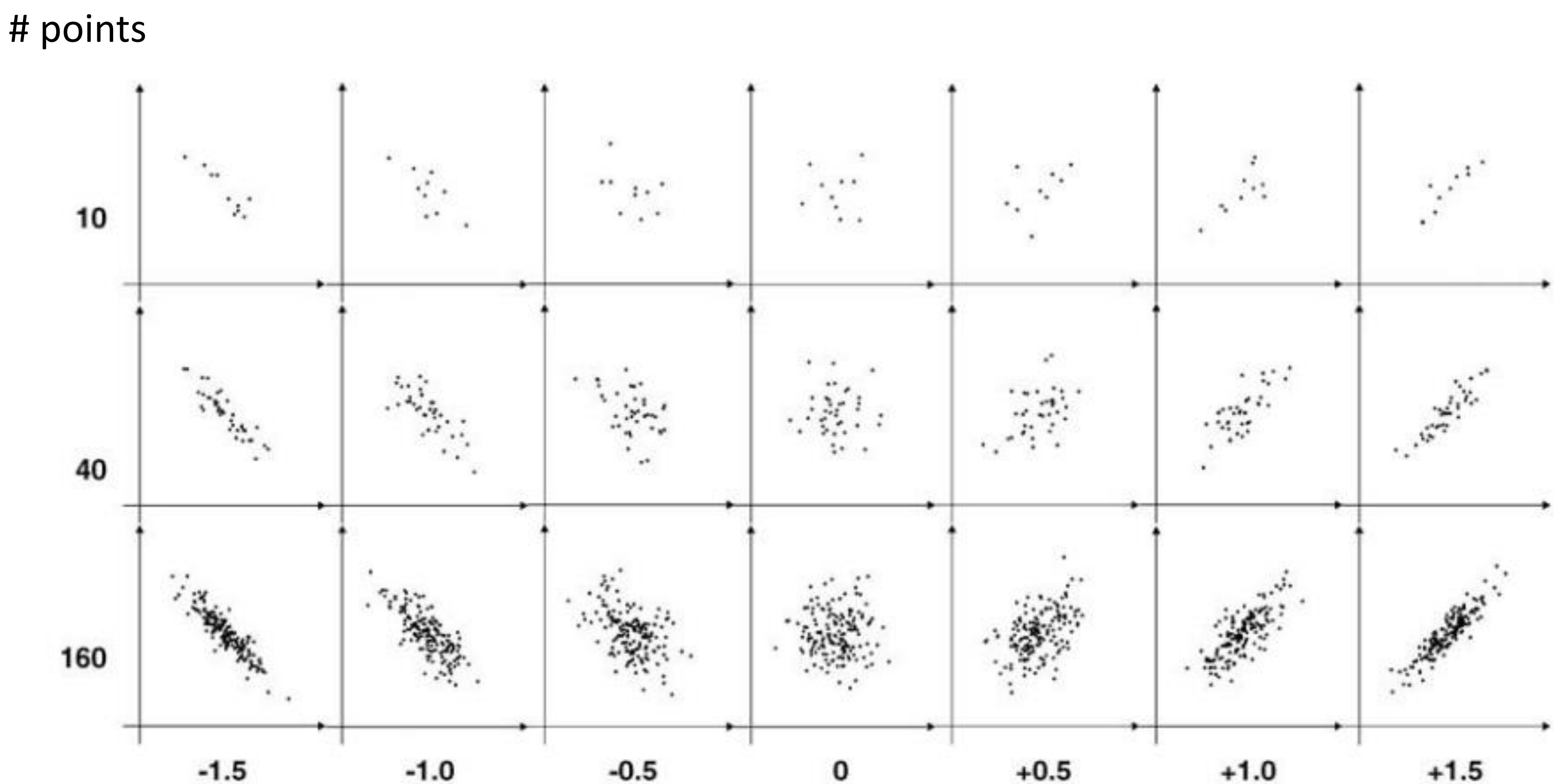


correlation r=-1.0 r=0 r=1.0

Patterns in Parallel Coordinates



PATTERNS IN SCATTERPLOTS



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

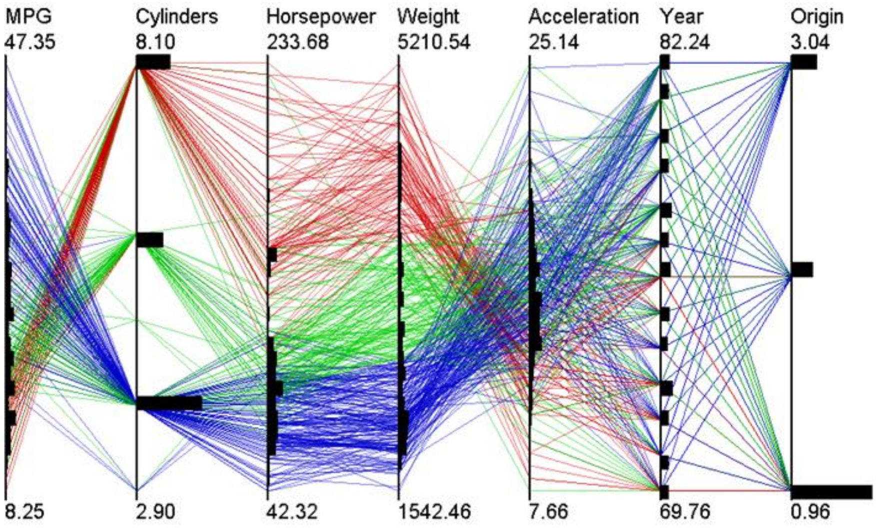
Li et al. found that <u>twice as many</u> correlation levels can be distinguished with scatterplots

Information Visualization Vol. 9, 1, 13 – 30

Axis Reordering Problem

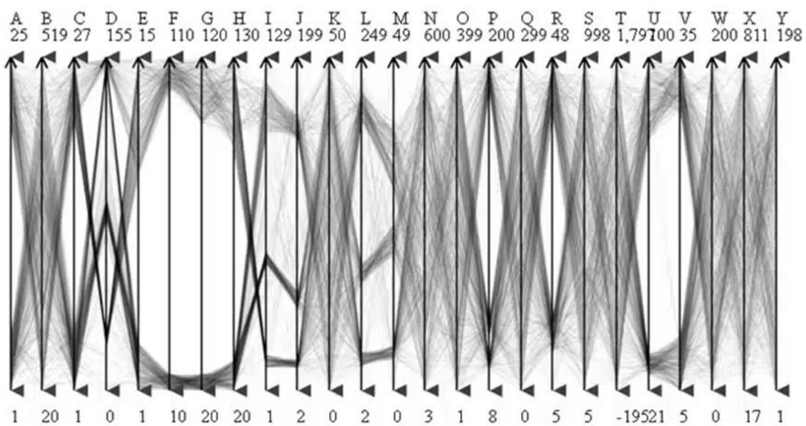
There are n! ways to order the n dimensions

- 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

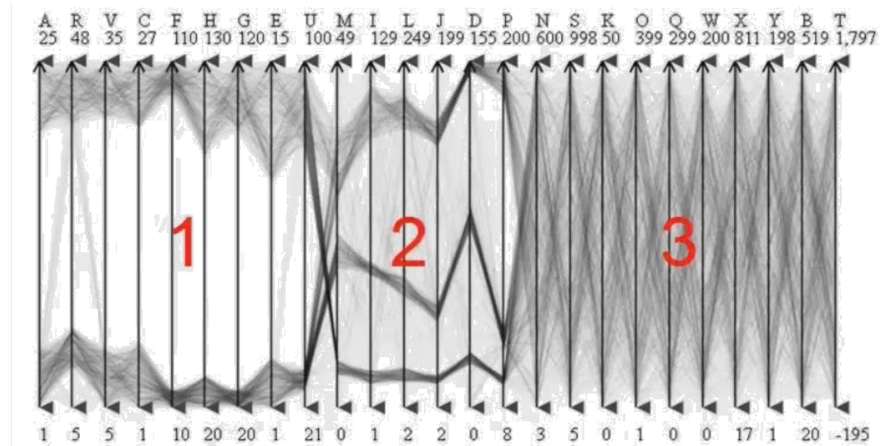


Axis Reordering Problem

The below is not an optimal ordering, why?



have



want

- what characteristics makes for an insightful pairwise ordering?
- what measure should we optimize to get this?
- yes, the correlation!

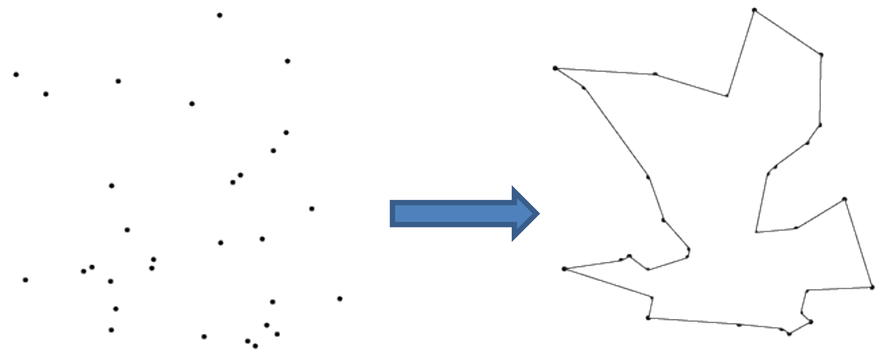
Optimizing the Axis Ordering

For each axis pair, compute correlation of attributes

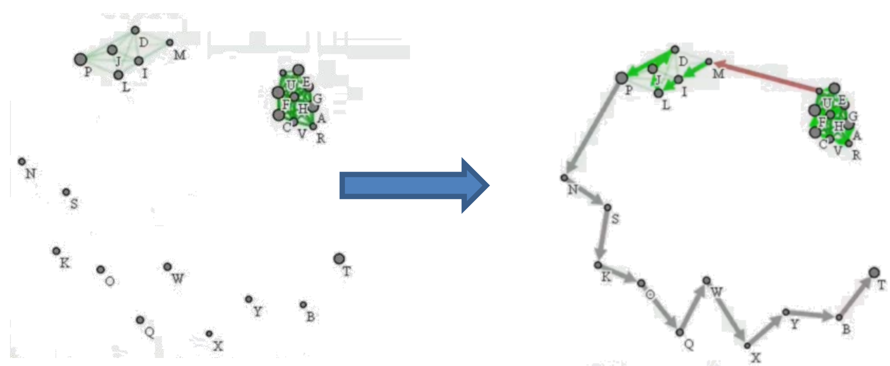
Do MDS and compute optimal-cost path across all attributes

What algorithm does this?

- Traveling Salesman Solver



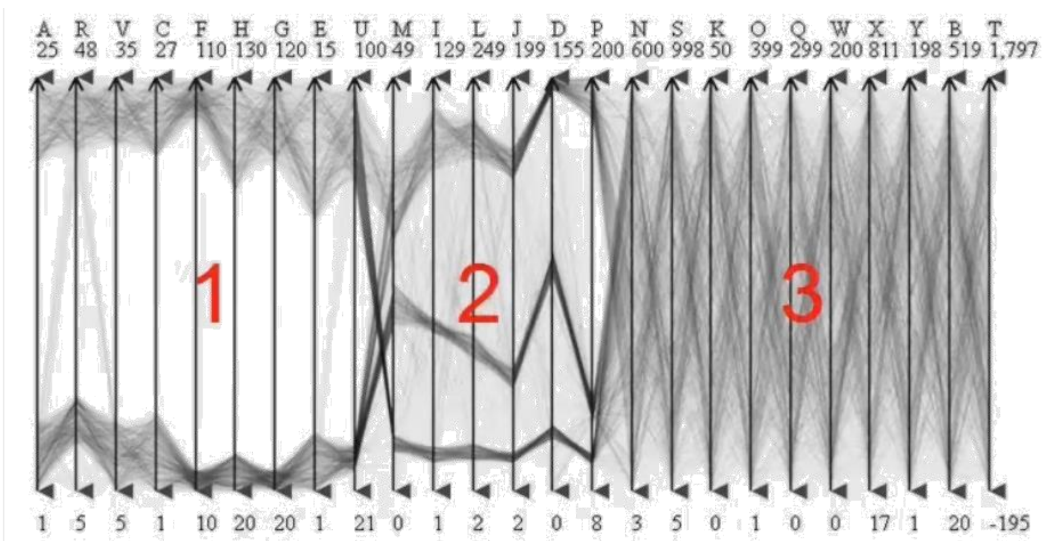
Correlation plot
for the data on last slide

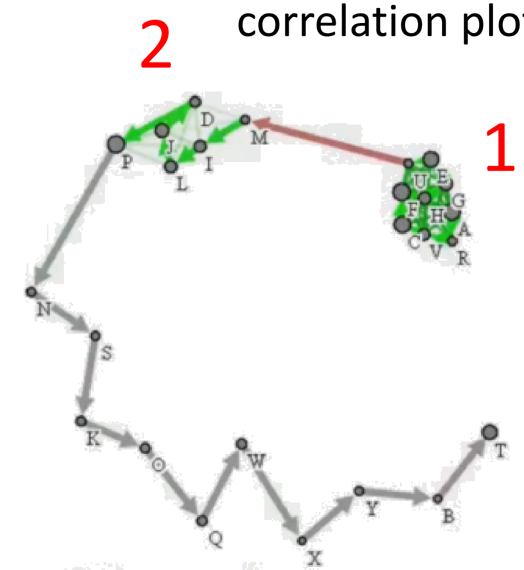


Axis Reordering Advantages

This ordering is better, why?

attribute correlation plot

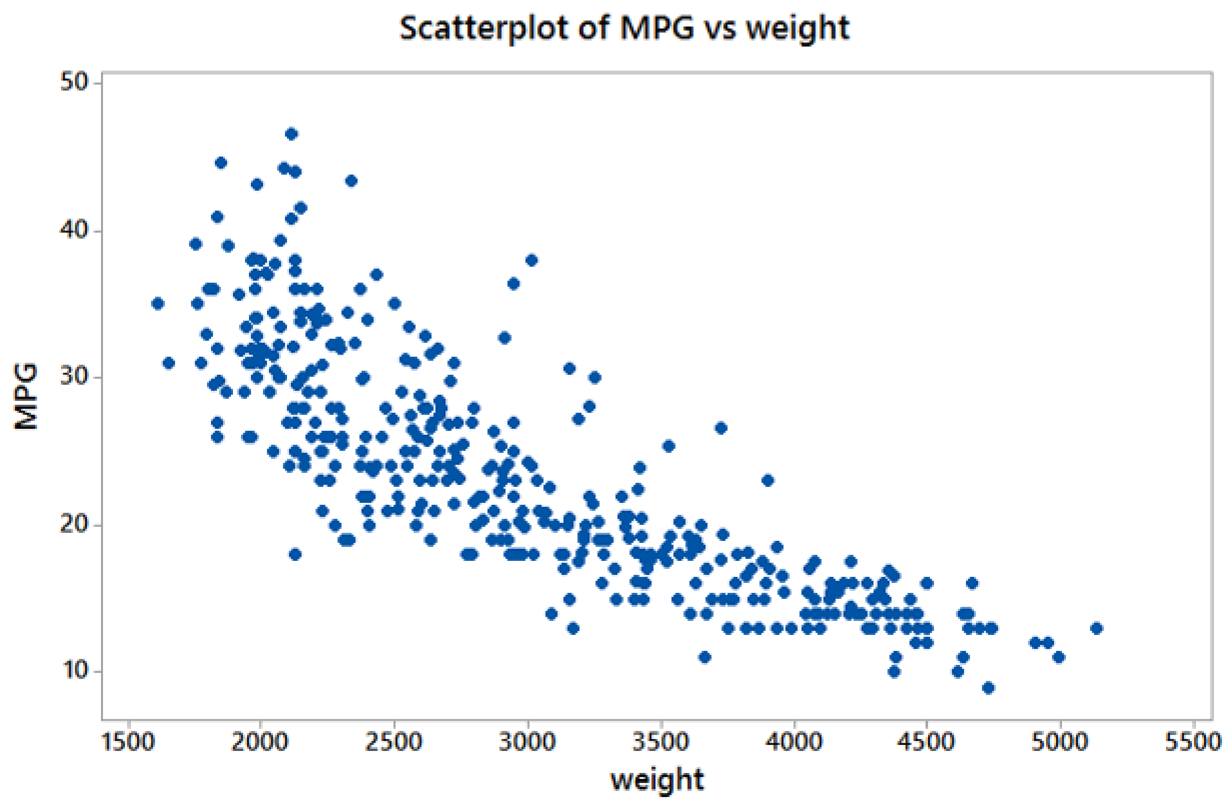




- because it doesn’t waste axis pairs on uncorrelated relationships
- only region 3 is uncorrelated
- regions 1 and 2 are subspace clusters

Scatterplots

Projection of the data items into a bivariate basis of axes



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

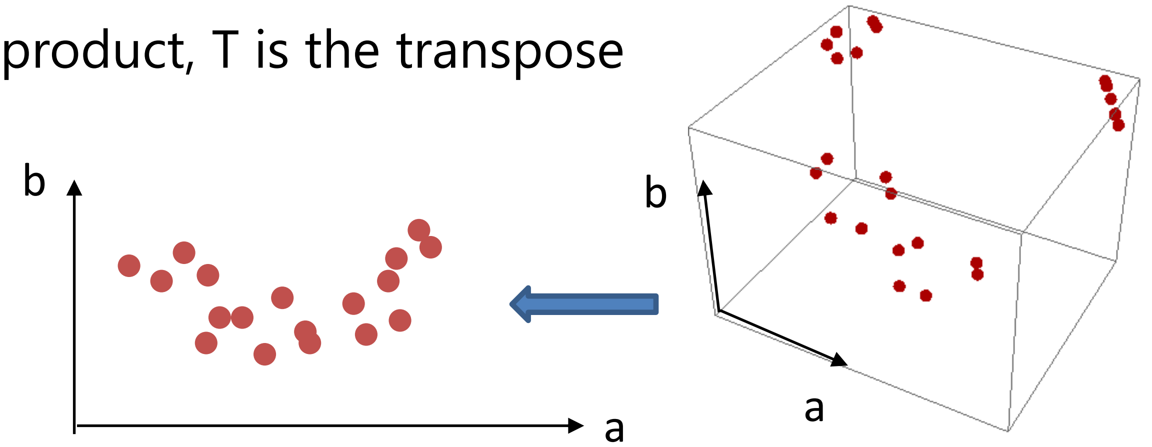
  $a = \{a_1. a_2, a_3, \dots a_N\}$

  $b = \{b_1. b_2, b_3, \dots b_N\}$
- a projection $\{x_a, x_b\}$ of x into the 2D basis spanned by {a, b} is:

  $x_a = a \cdot x^T$

  $x_b = b \cdot x^T$

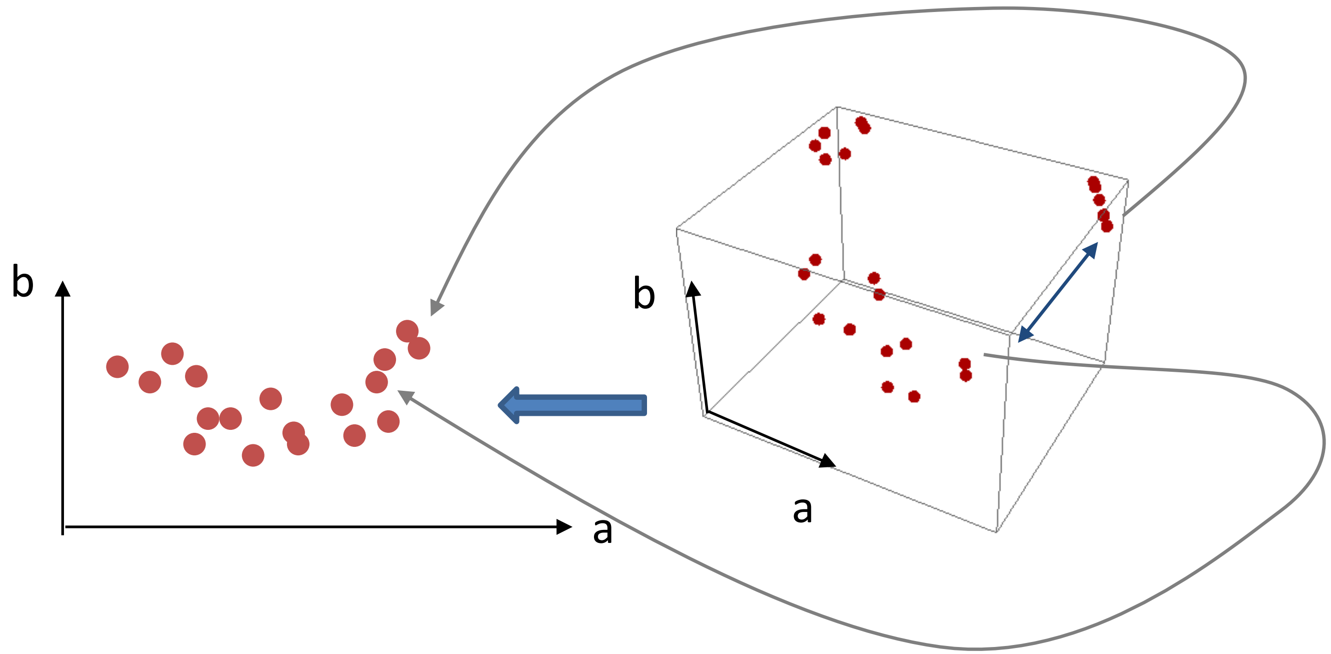
  where · is the dot product, T is the transpose



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*



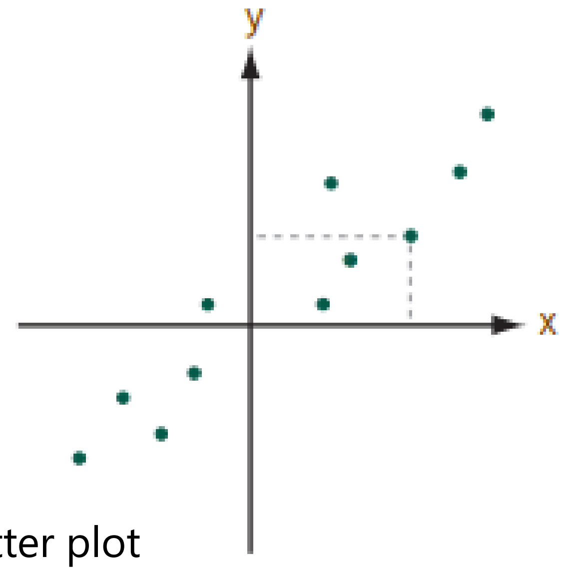
BIPLOTS

Plots data points and dimension axes into a single visualization

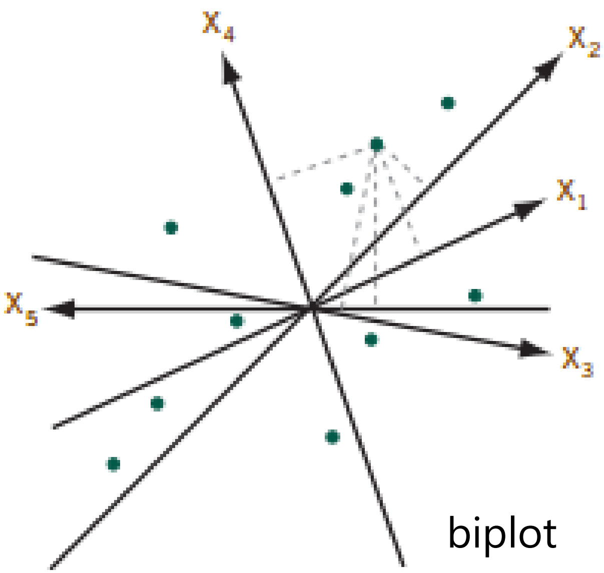
- uses first two PCA vectors as the basis to project into
- find plot coordinates [x] [y]

  for data points: [PCA_1 · data vector] [PCA_2 · data vector]

  for dimension axes: [PCA_1[dimension]] [PCA_2[dimension]]



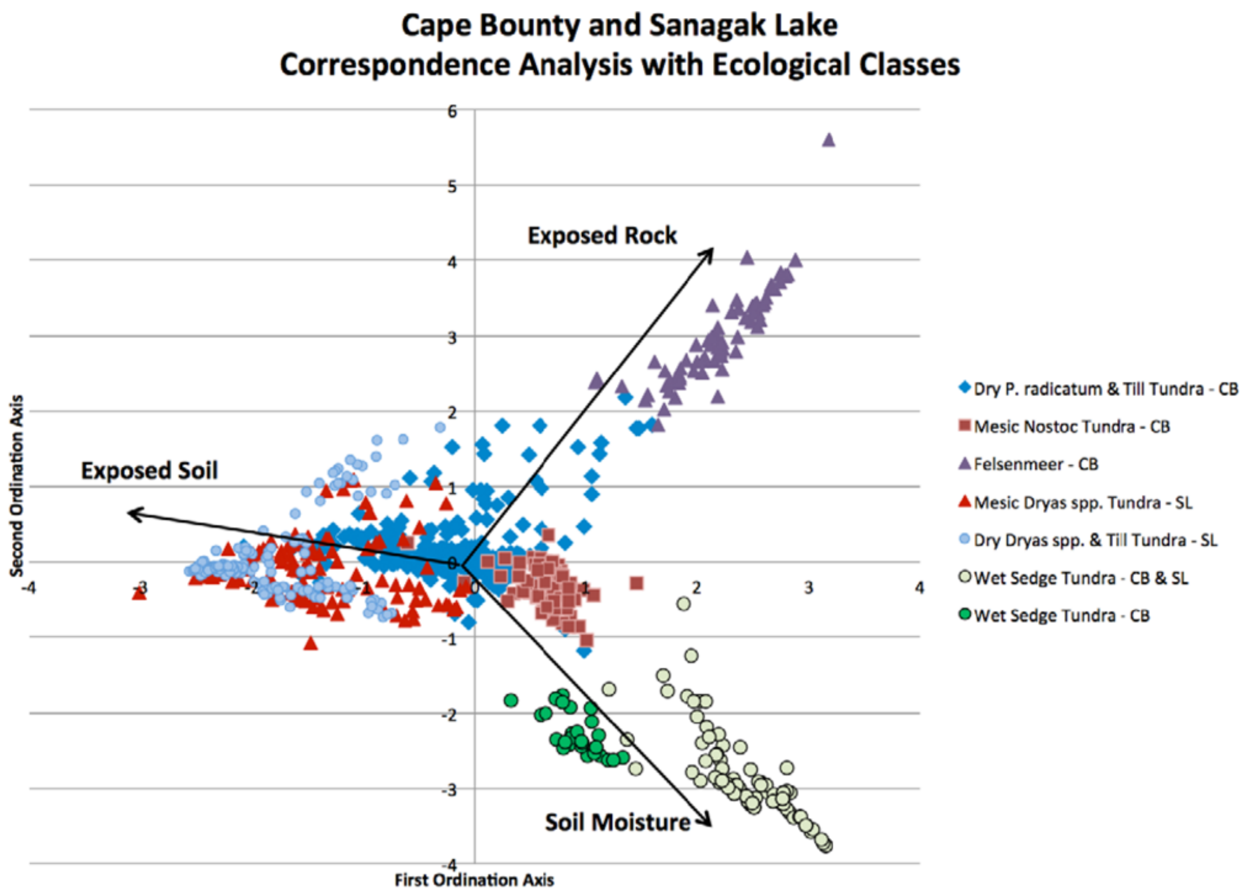
scatter plot



biplot

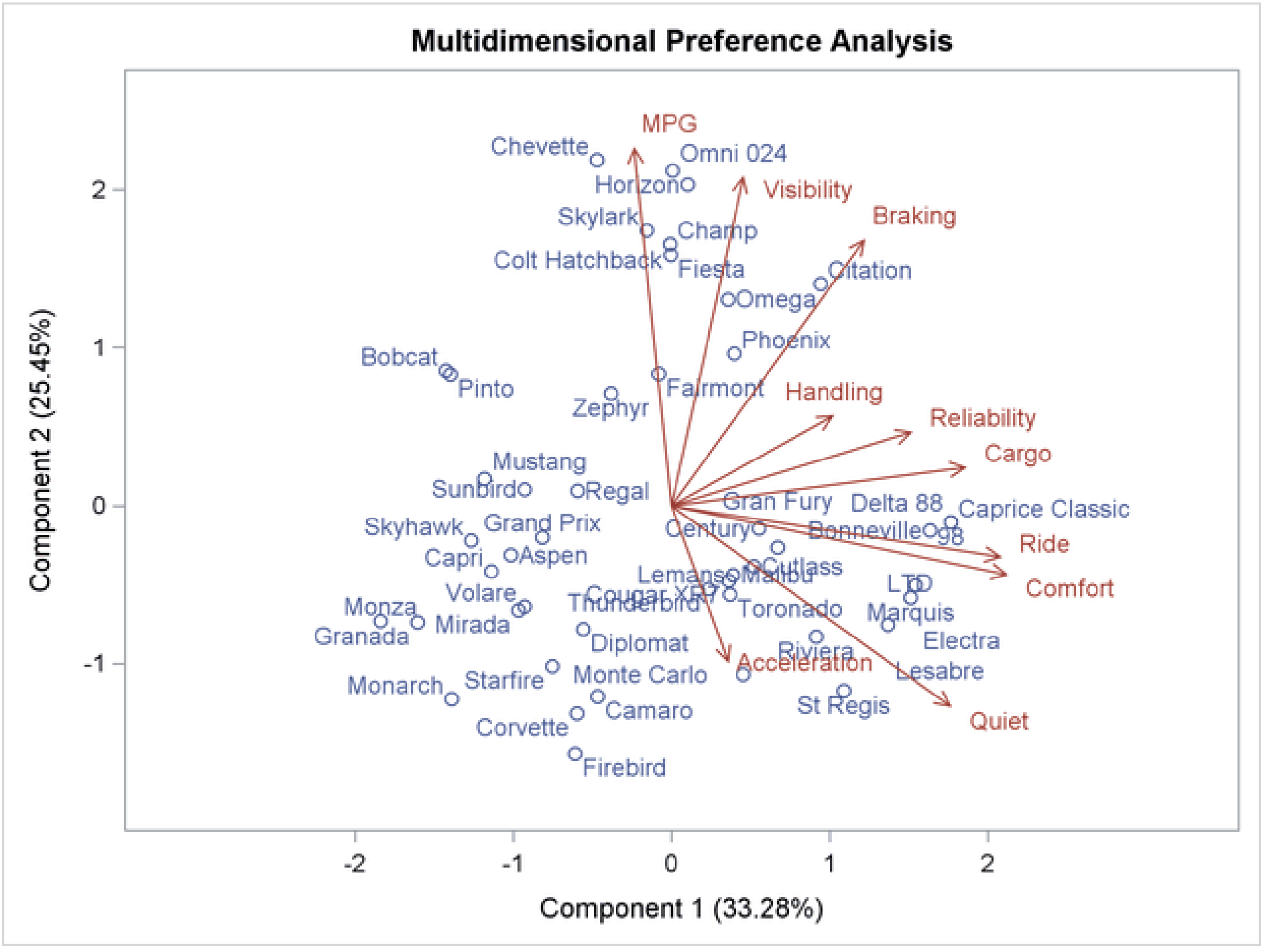
Biplots in Practice

See data distributions into the context of their attributes



Biplots in Practice

See data points into the context of their attributes



Biplots – A Word of caution

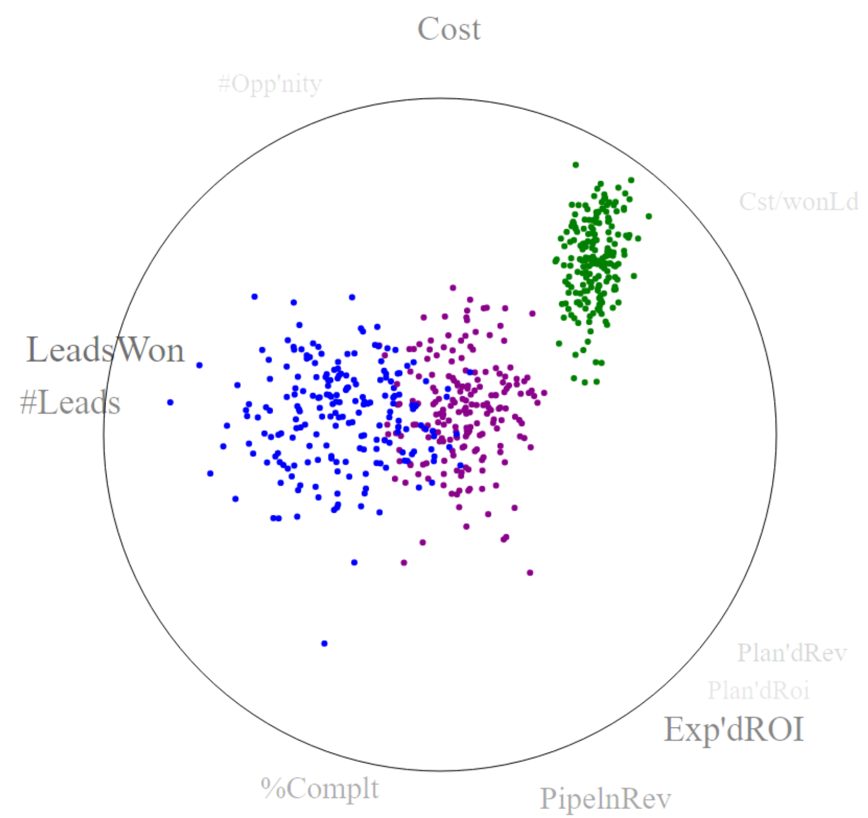
Do be aware that the projections may not be fully accurate

- 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 .....



Meet the *Subspace Voyager*

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

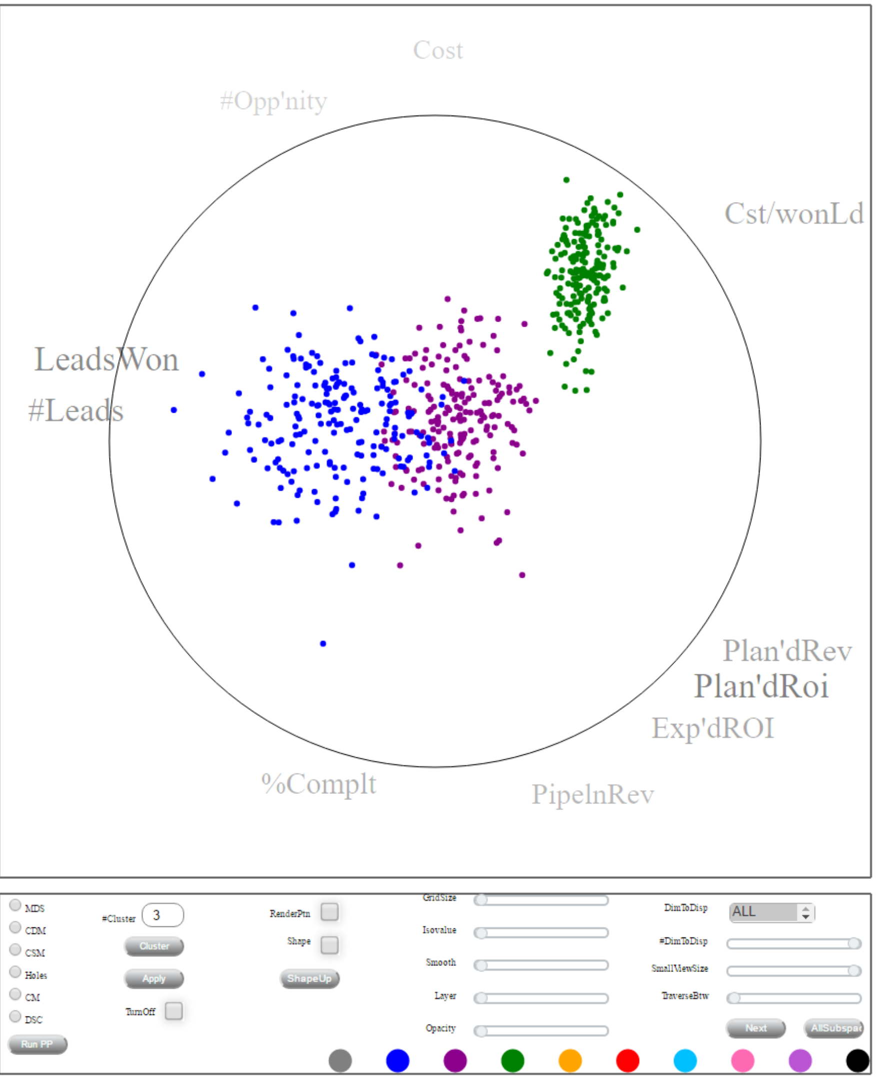
- clustering
- classification
- reducing clusters to intrinsic dimensionality via local PCA

Allows users to interactively explore these lower-D subspaces

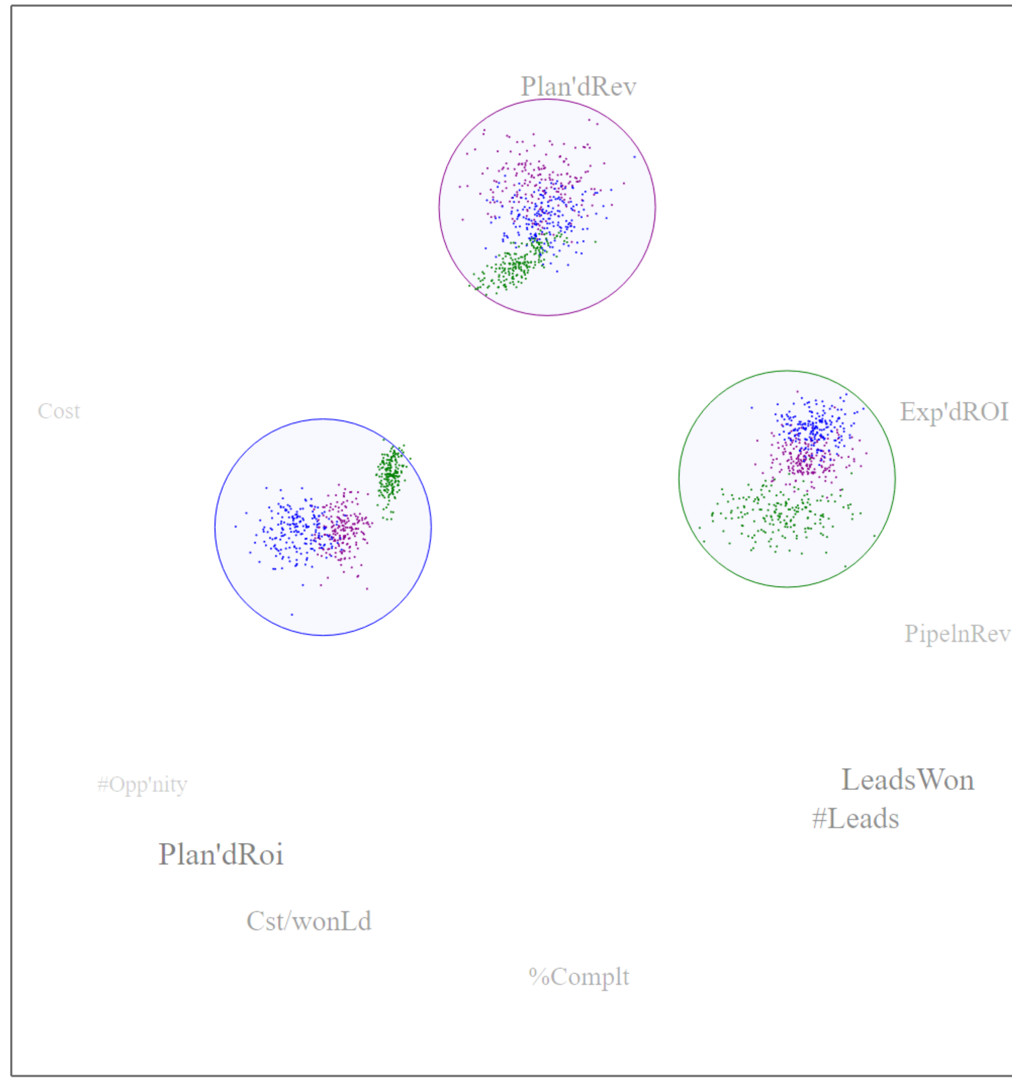
- explore them as a chain of 3D subspaces
- transition seamlessly to adjacent 3D subspaces on demand
- save observations as you go (and return to them just as well)

Visualize Raw Data w/the Subspace Voyager

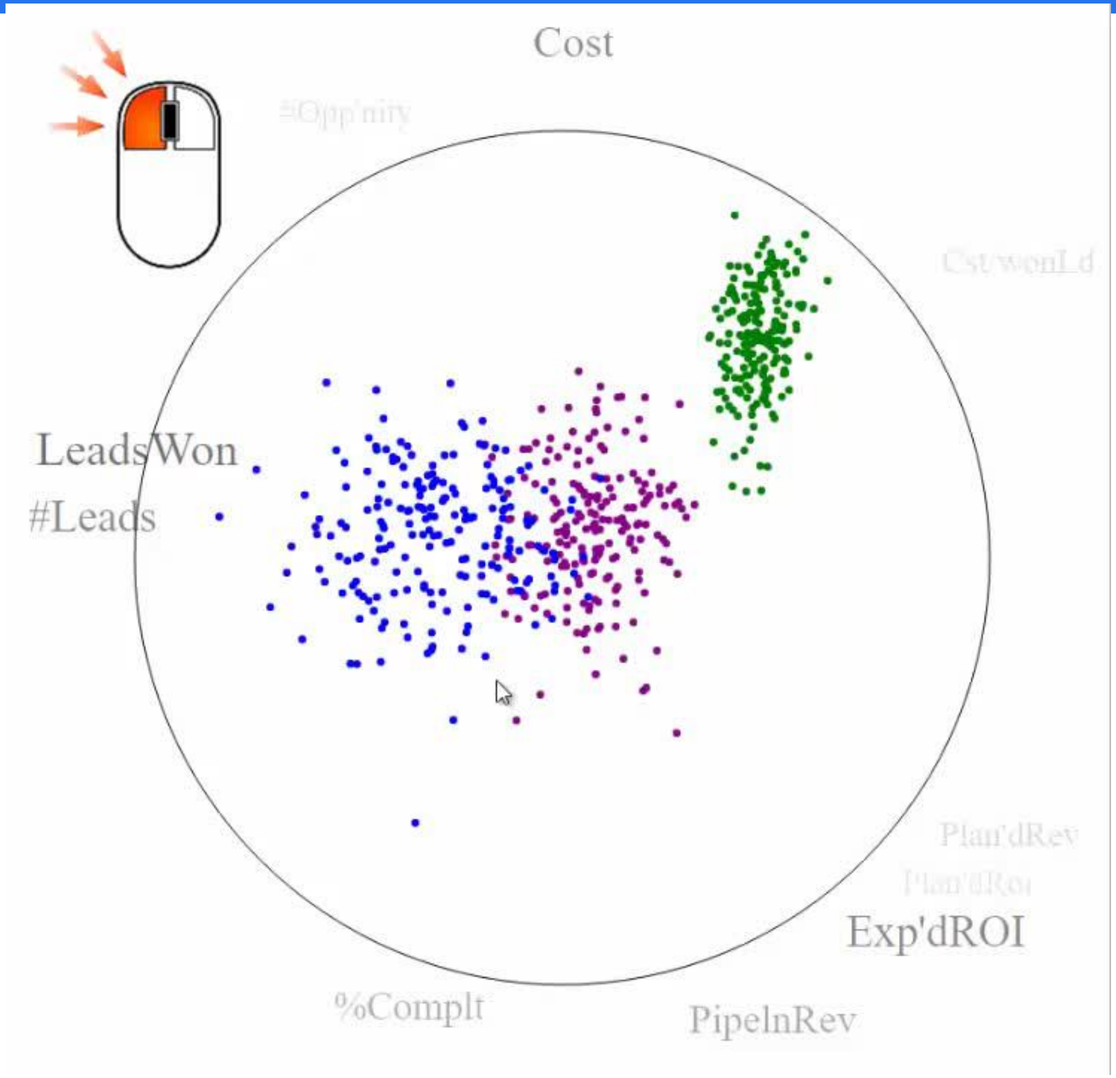
Interactive Scatterplot



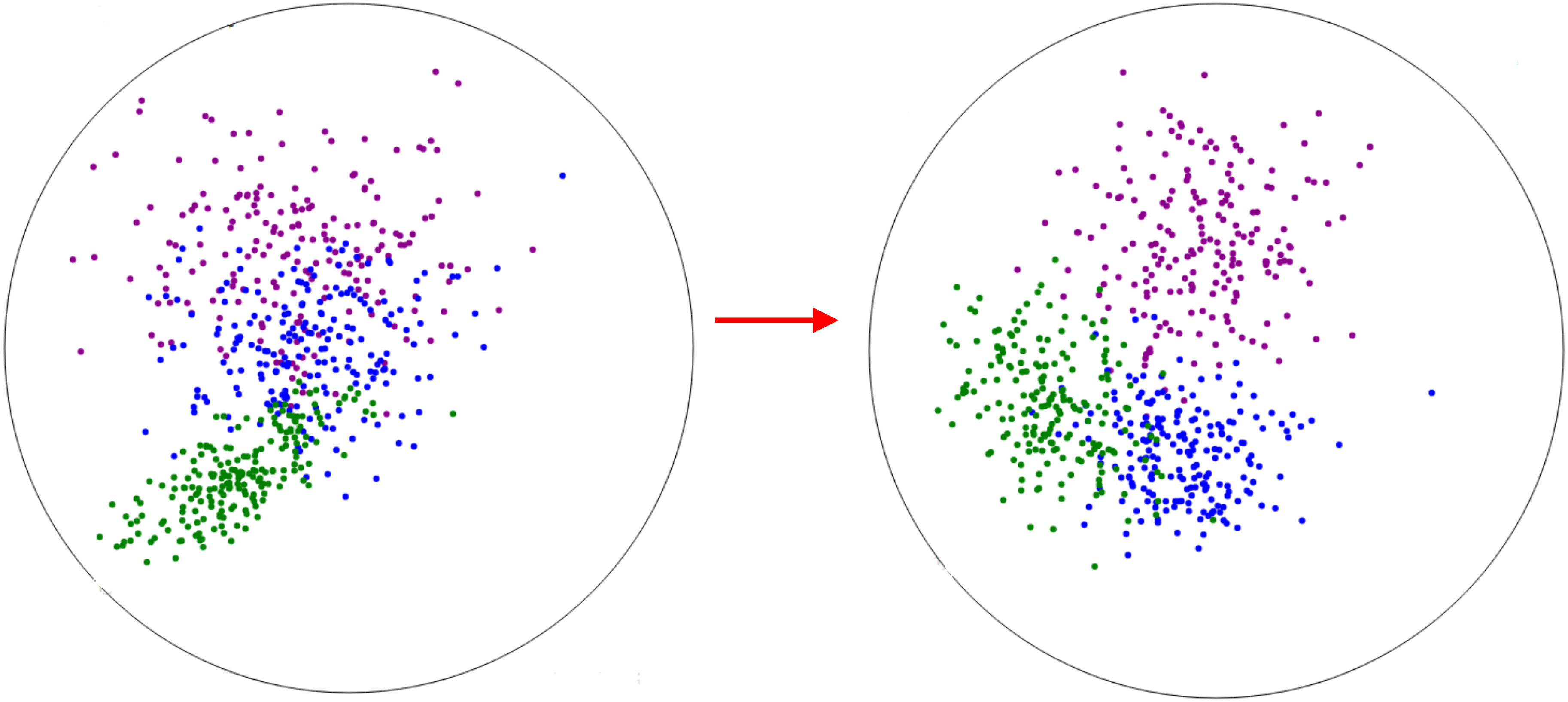
Subspace Trail Map



Trackball-Based Cluster Exploration



Interactive View Optimizer



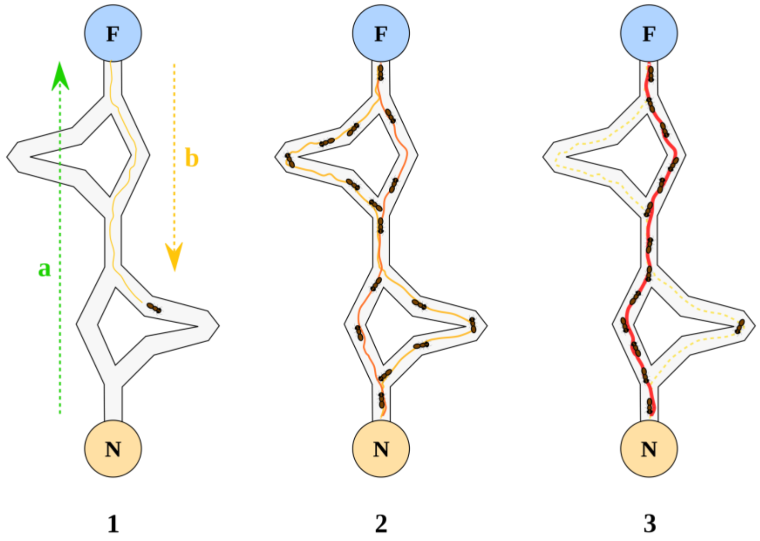
Uses genetic-algorithm driven projection pursuit

Several view quality metrics are available

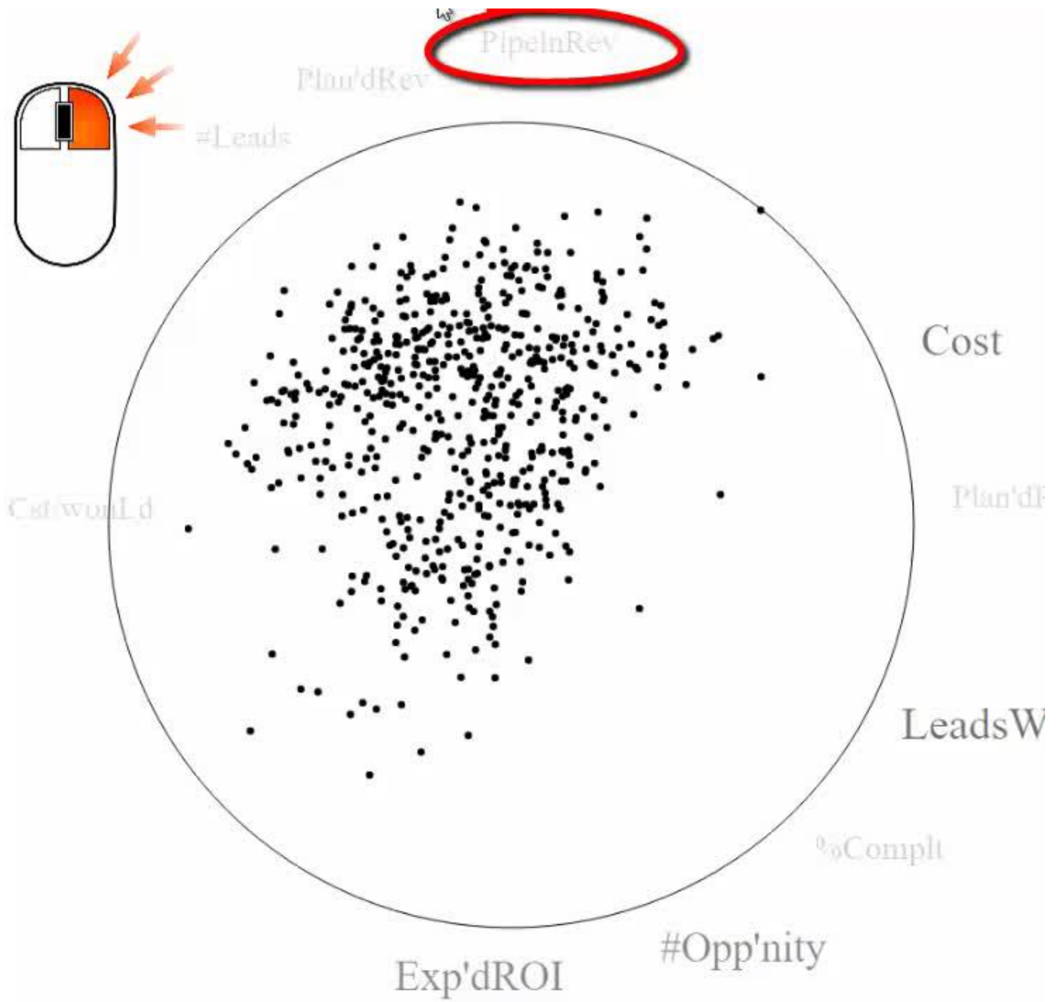
(Genetic) Ant Colony Algorithm

Generate many views and score them (one per ant)

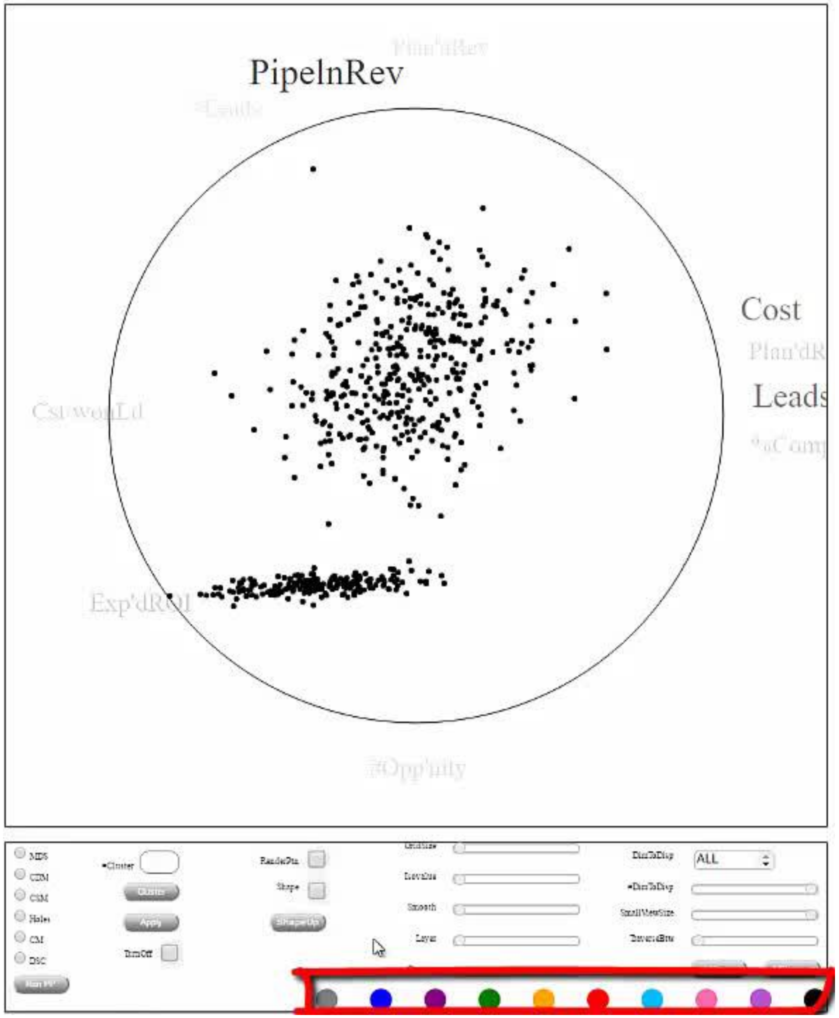
- poor scoring ants die and well-scoring ants survive
- sub paths of high scoring receive pheromone
- pheromone entices ants to take this path again
- each path variation is a parameter choice
- best view corresponds to the path that is converged on



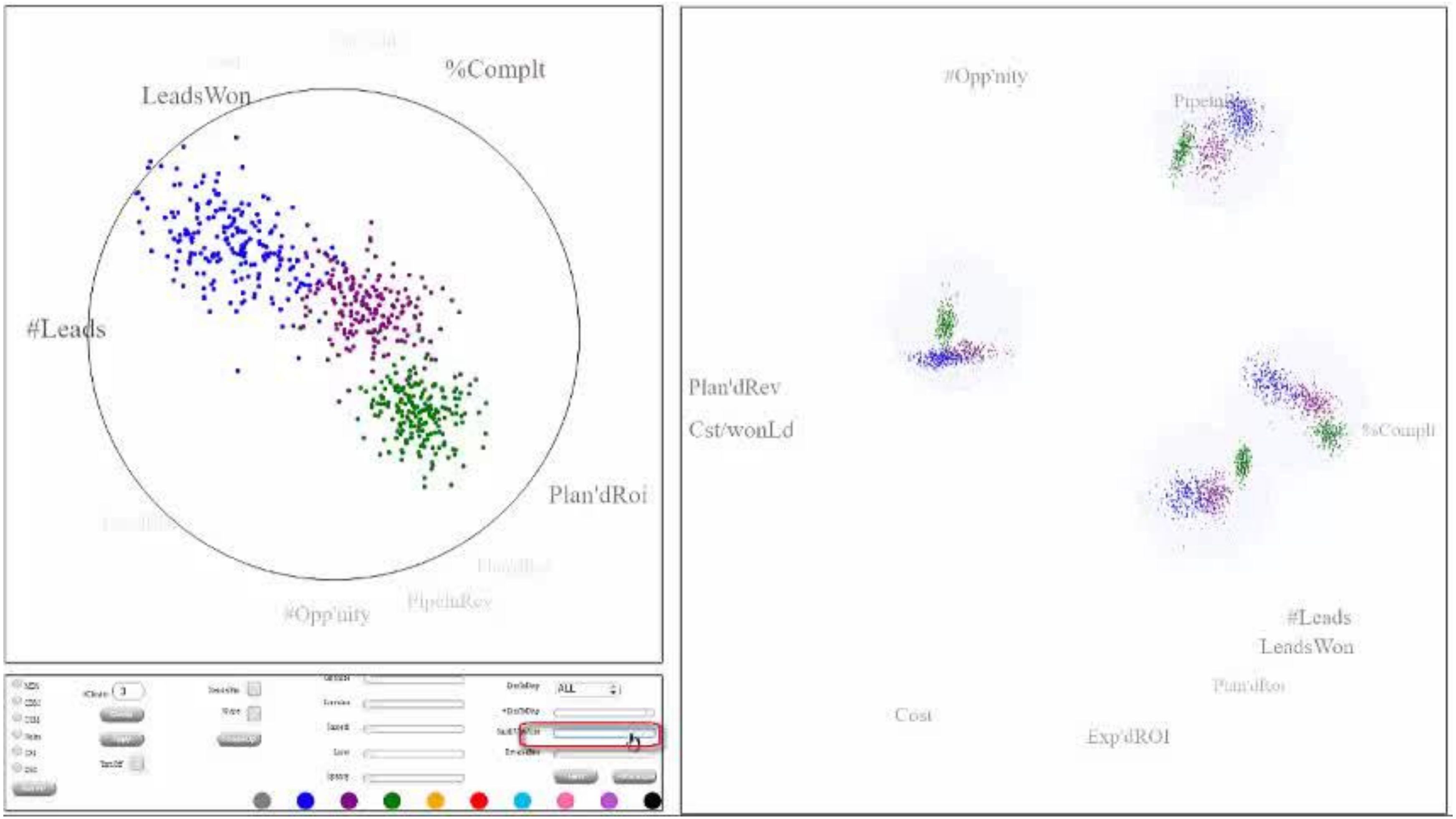
Chase Interesting Clusters – Transition to Adjacent 3D Subspaces



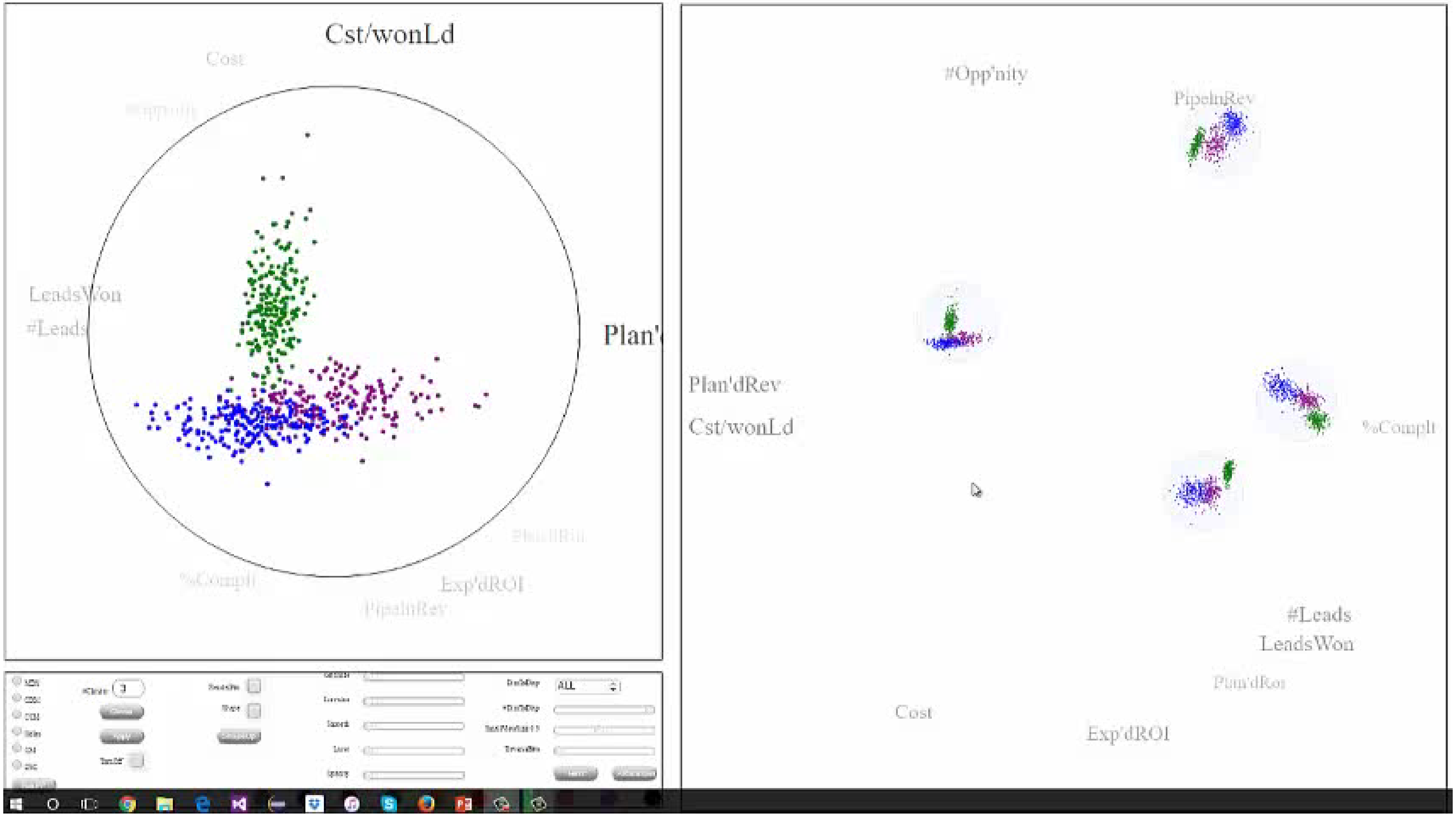
Edit and Annotate Clusters



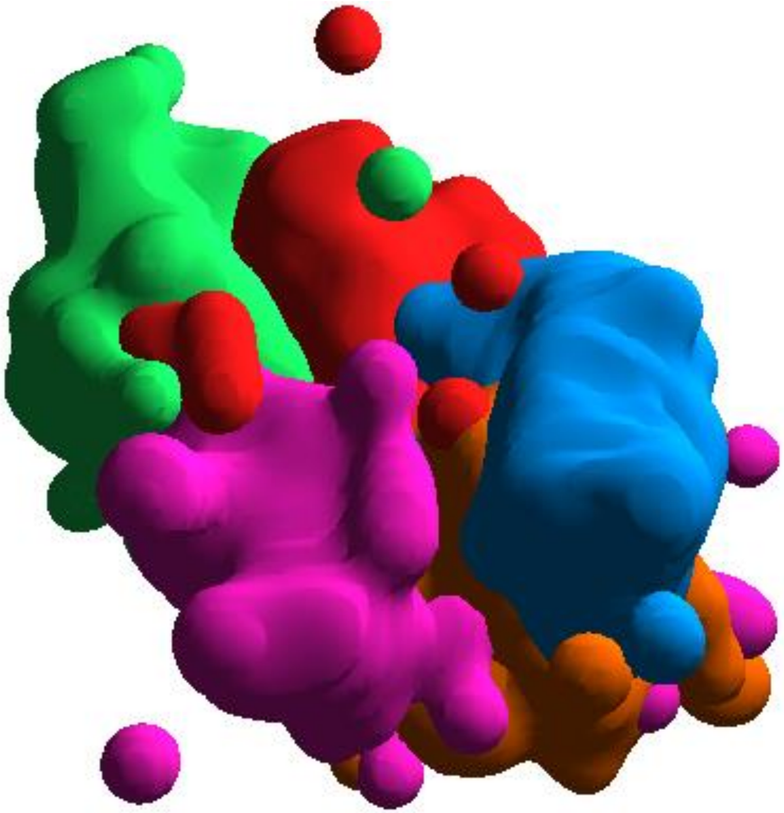
The Subspace Trail Map



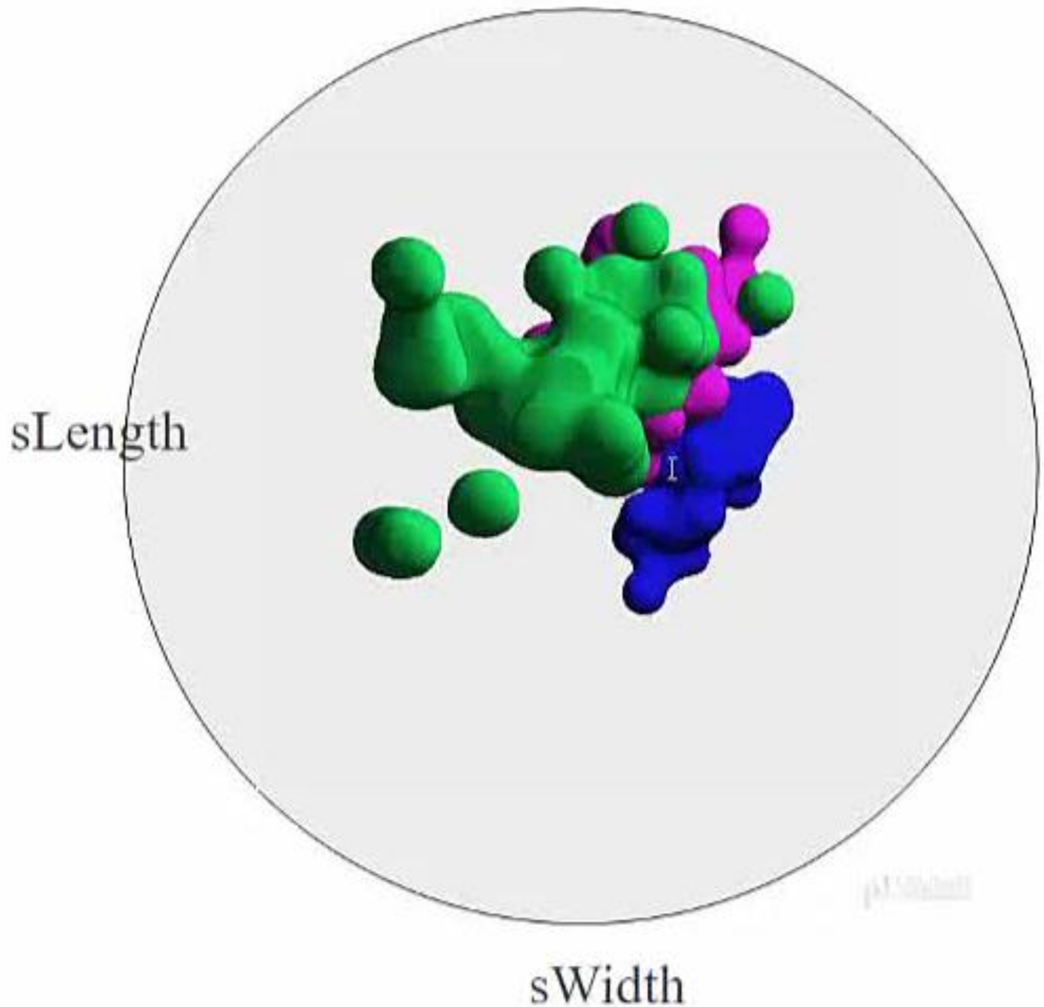
Walk The Subspace Trail Map



Clarify Spatial Relationships



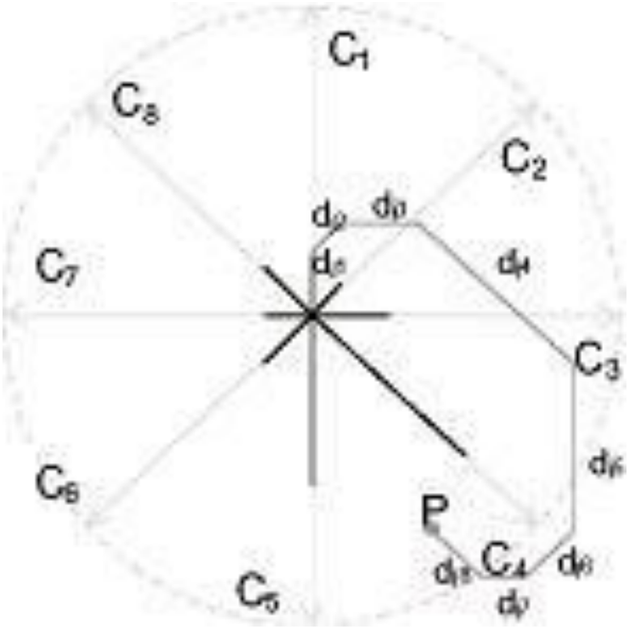
Clarify Spatial Relationships



Star Coordinates

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

- a point P is plotted as a vector sum of all axis coordinates

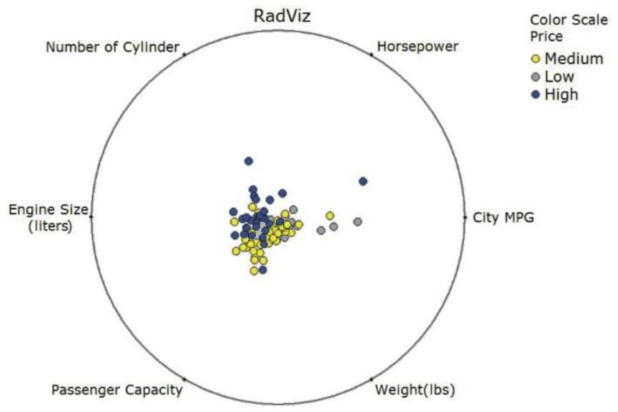


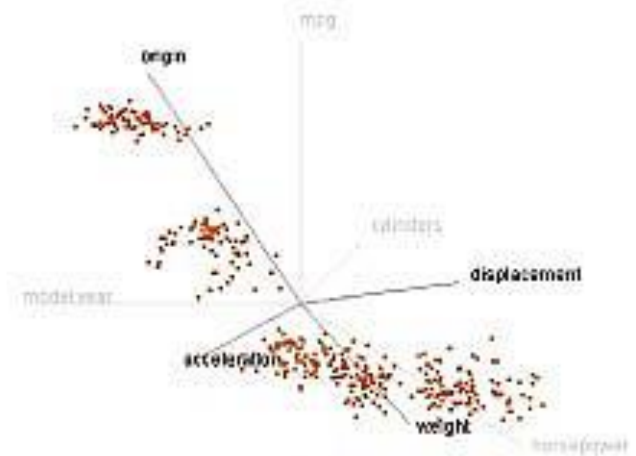
Star Coordinates

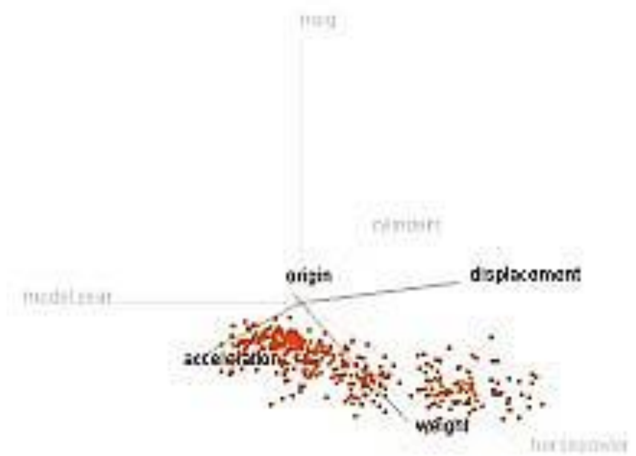
Operations defined on Star Coords

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

Similar paradigm: RadViz







Commonalities

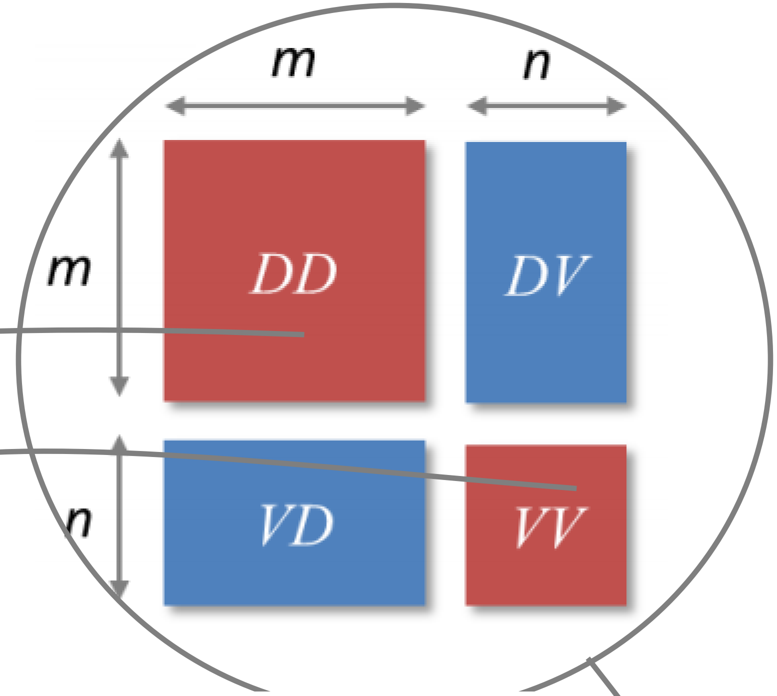
All of these scatterplot displays share the following characteristics

- allow users to see the data points in the context of the variables
- but can suffer from projection ambiguity
- some offer interaction to resolve some of these shortcomings
- but interaction can be tedious

Are there visualization paradigms that can overcome these problems?

- yes, algorithms that optimize the layout to preserve distances or similarities in high-dimensional space
- what is this algorithm?
- yes, MDS (Multi-Dimensional Scaling)
- we have discussed MDS before (so we will skip further discussion)

Uses of MDS

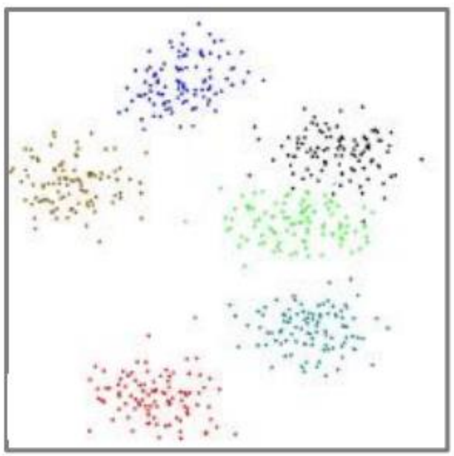


data similarity matrix (DD)

attribute similarity matrix (VV)

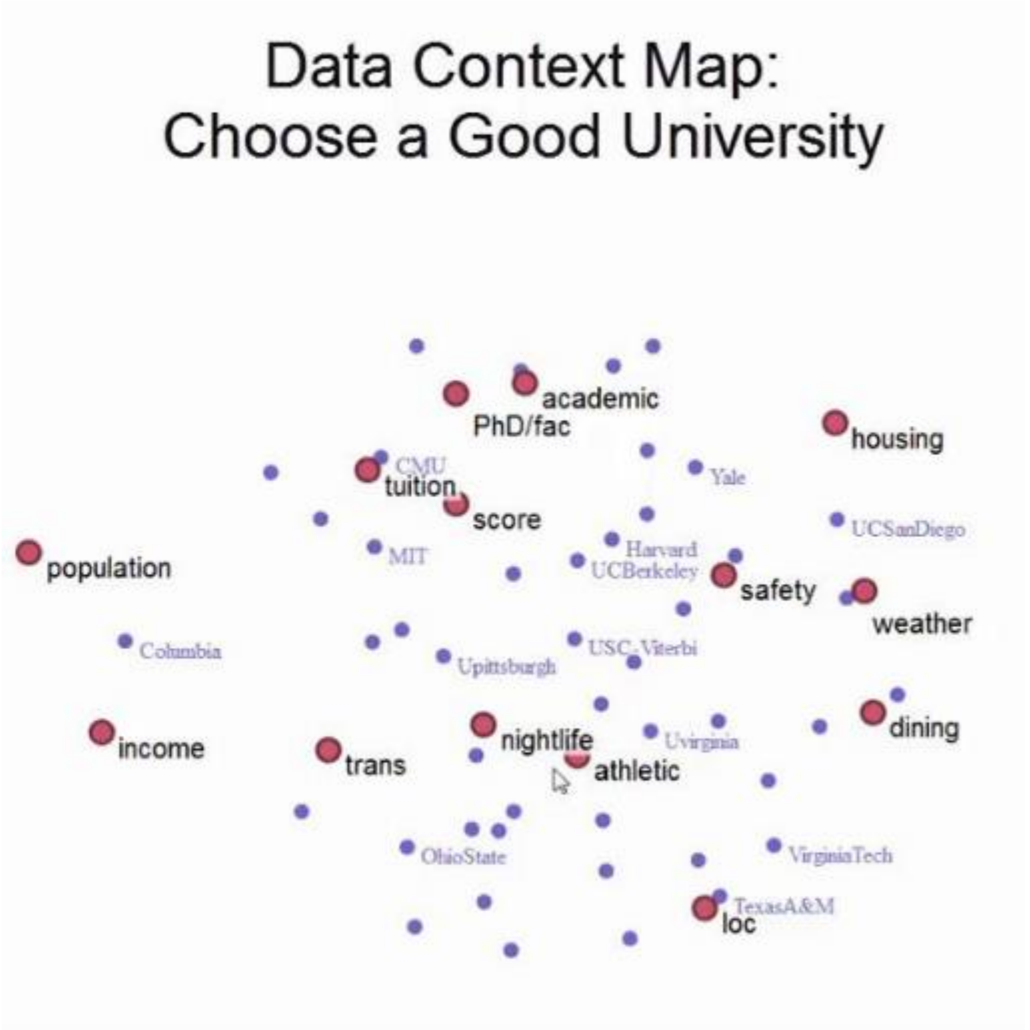
+ data-attribute similarity matrix (VD, DV)

Data layout



Yields the Data Context Map

Data visualized in the context of the attributes

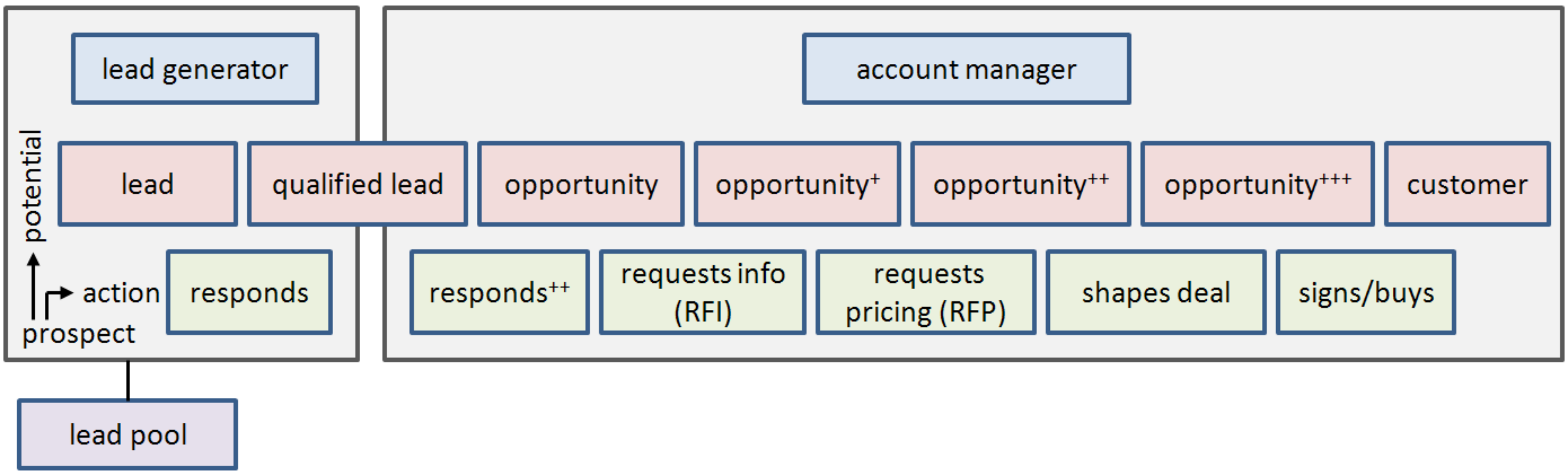


S. Cheng, K. Mueller, "The Data Context Map: Fusing Data and Attributes into a Unified Display," *IEEE Trans. on Visualization and Computer Graphics*, 22(1): 121-130, 2016.

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

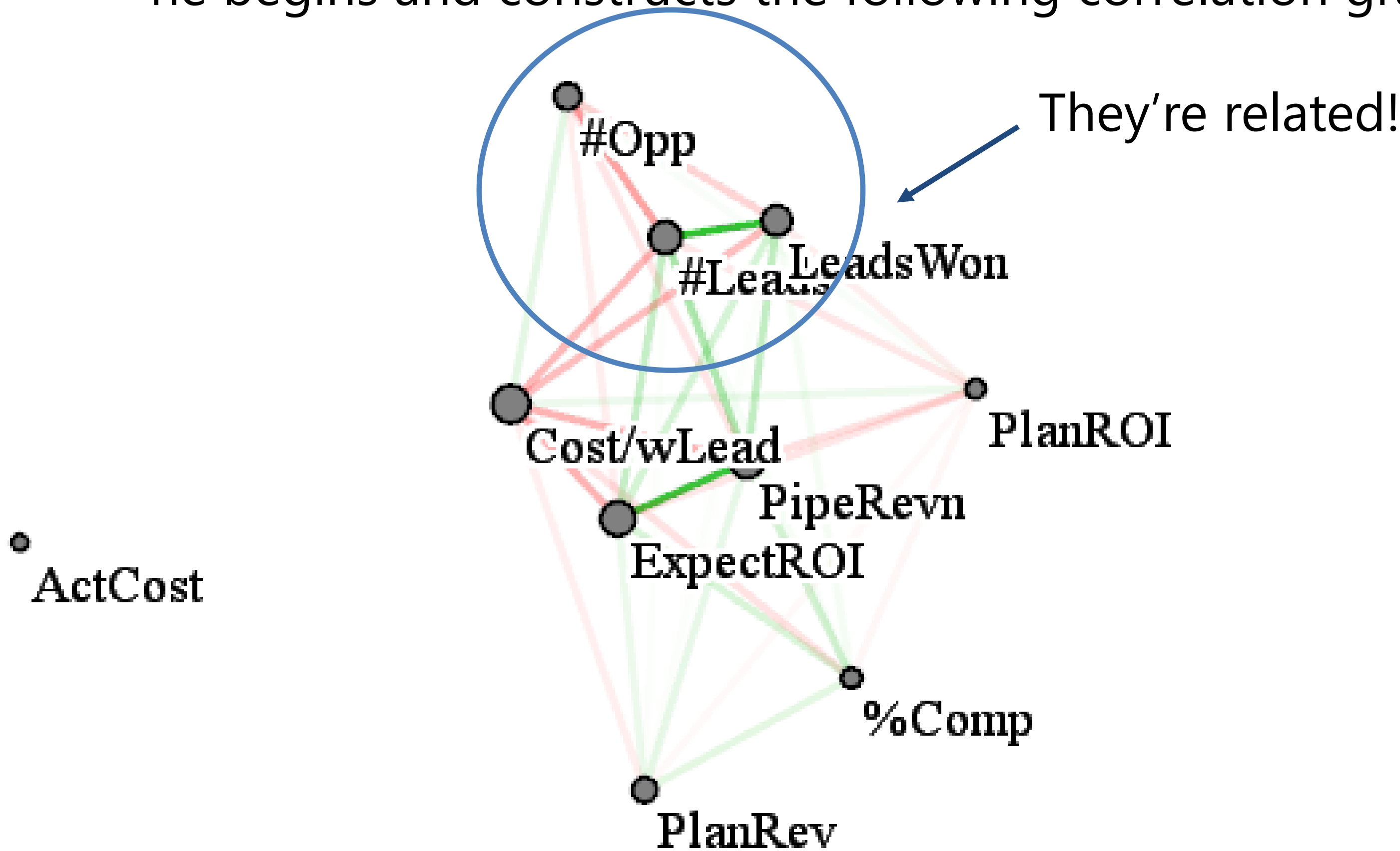
Evidence:

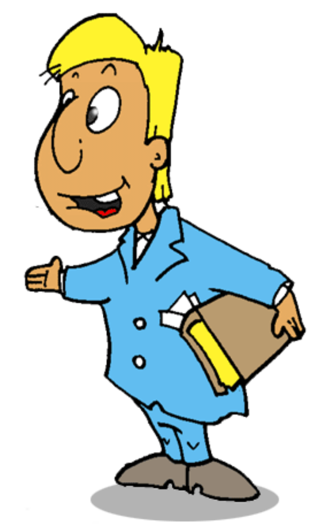
- data of three sales teams with a couple of hundred sales people in each team

Jim Begins

Meet Jim, one of the sales strategy analysts

- he begins and constructs the following correlation graph



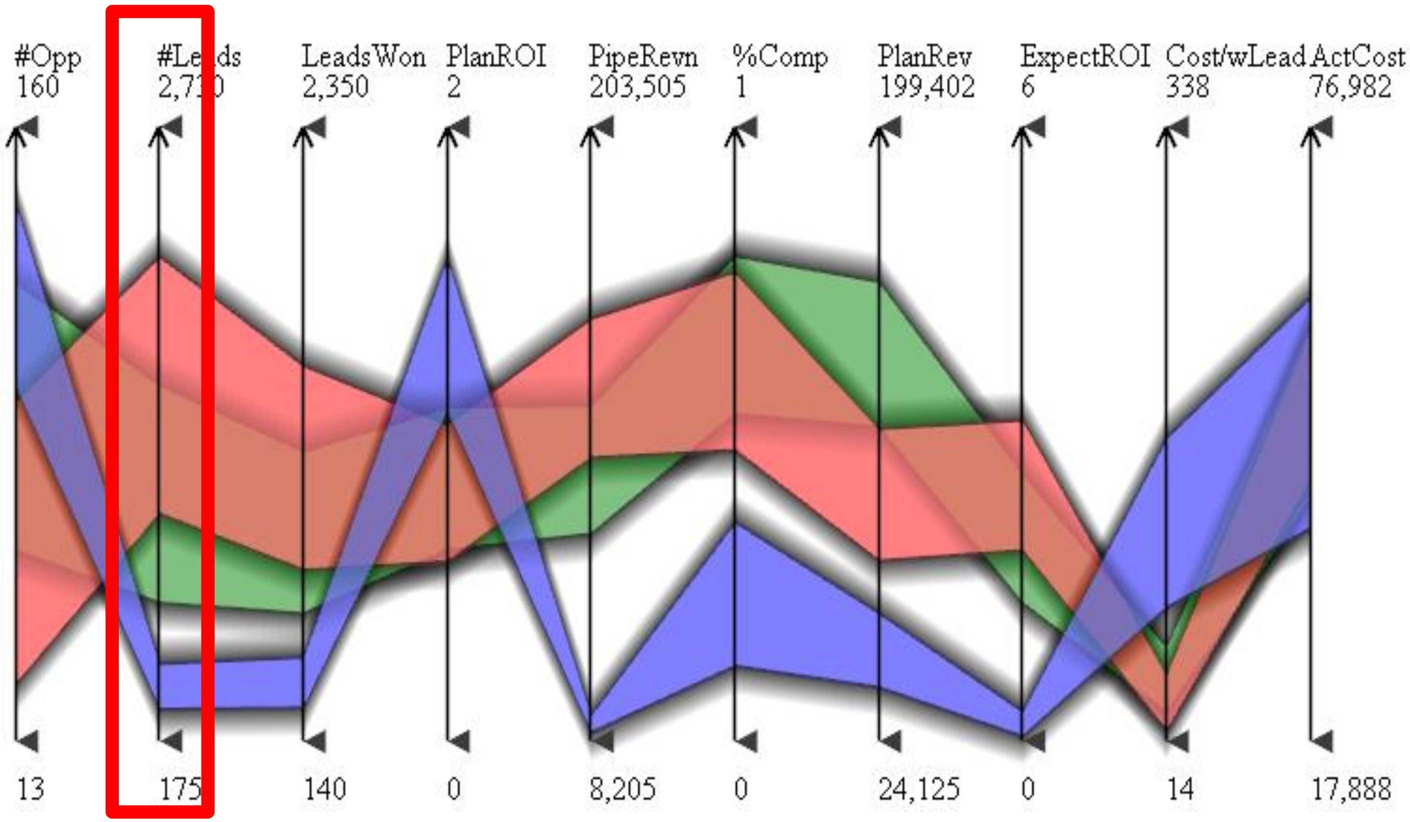


Jim

Jim's Story

He asks the TSP to compute an initial route

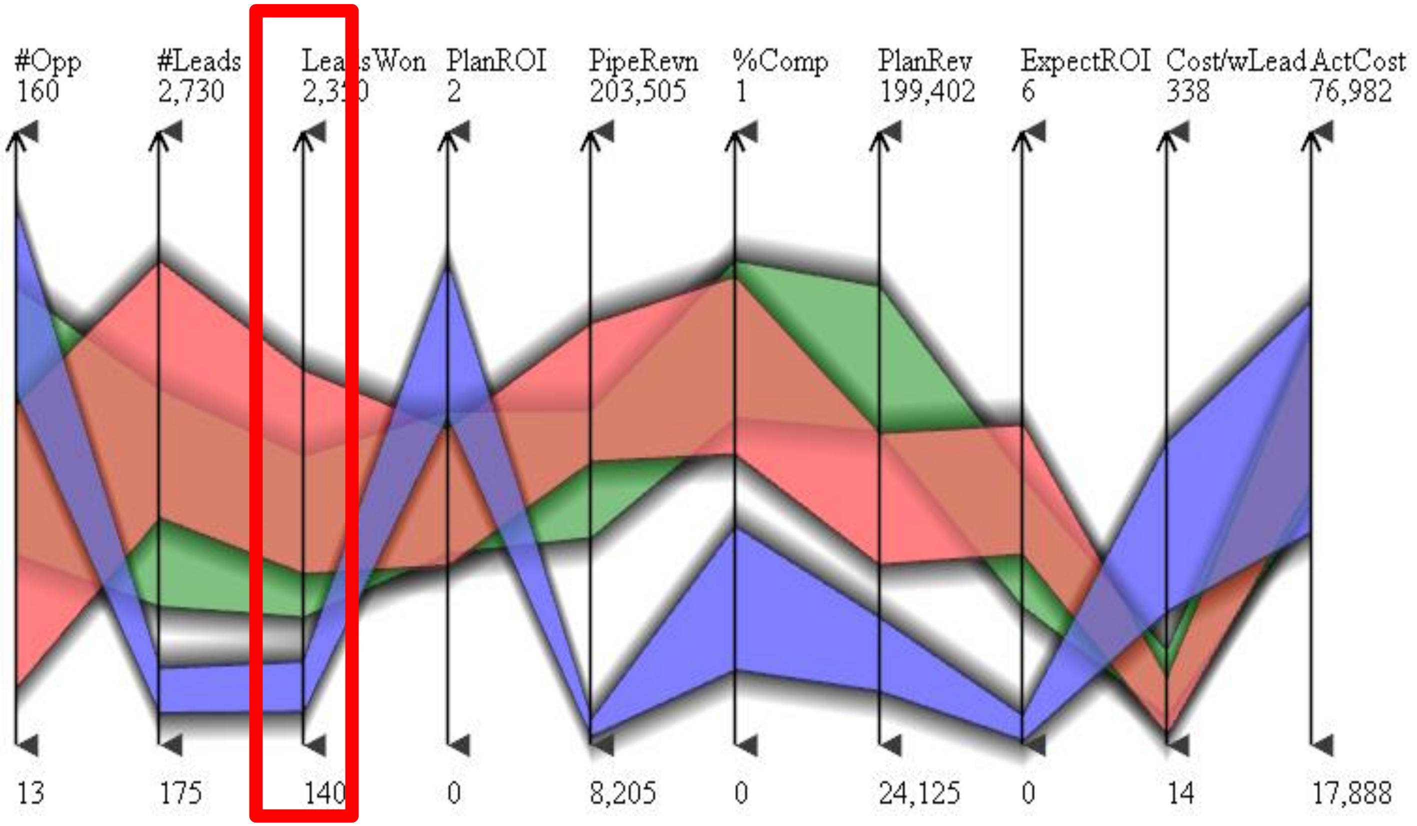
It gives rise to this parallel coordinate display



Jim's Story

He asks the TSP to compute an initial route

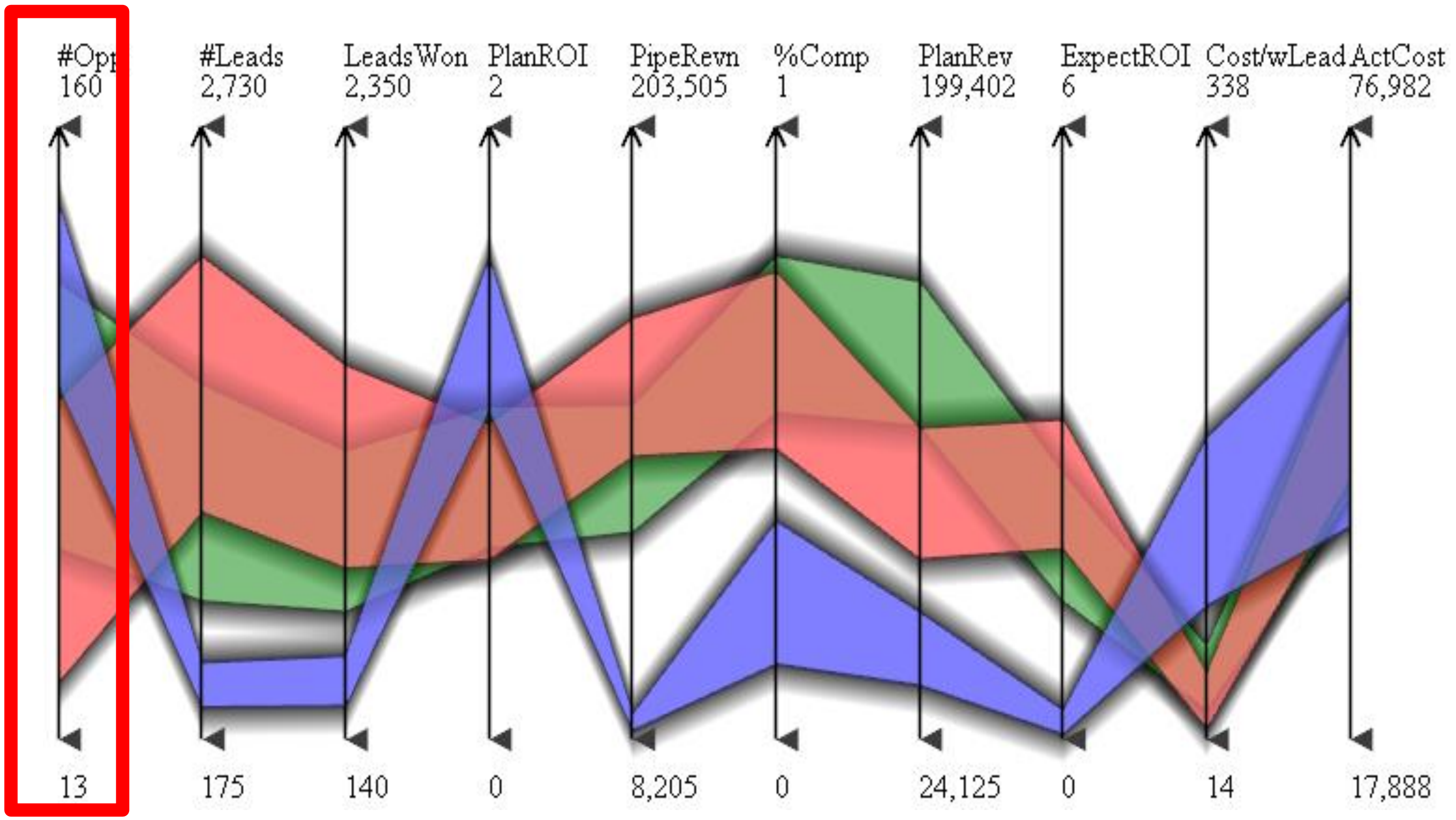
It gives rise to this parallel coordinate display



Jim's Story

He asks the TSP to compute an initial route

It gives rise to this parallel coordinate display

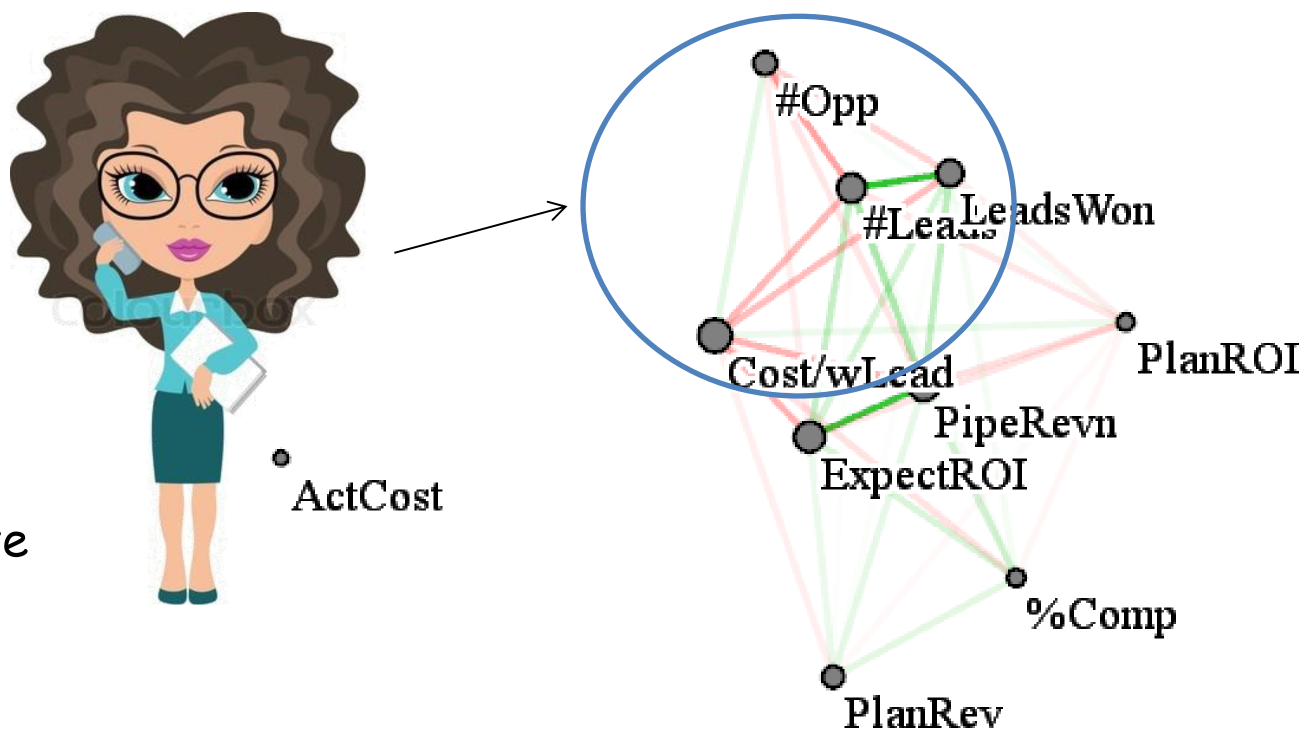


Kate Steps In

Now meet Kate, another sales analyst in the meeting room:

> “Hey, cost/won lead is nearby and it has a positive correlation with #opportunities but also a negative correlation with #won leads”

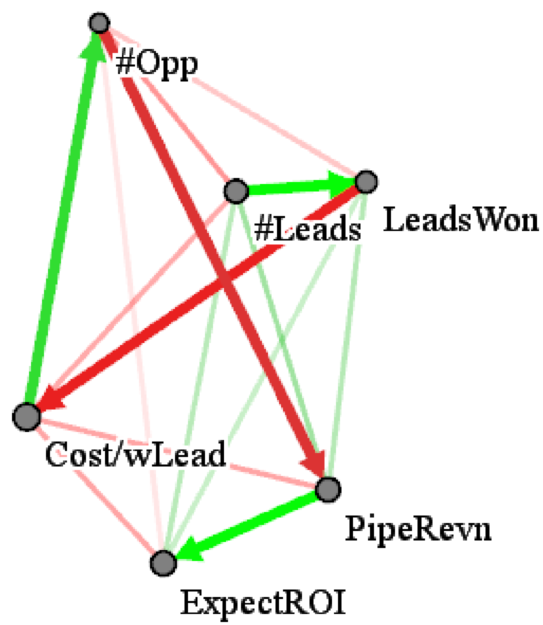
Kate



Kate's Story

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

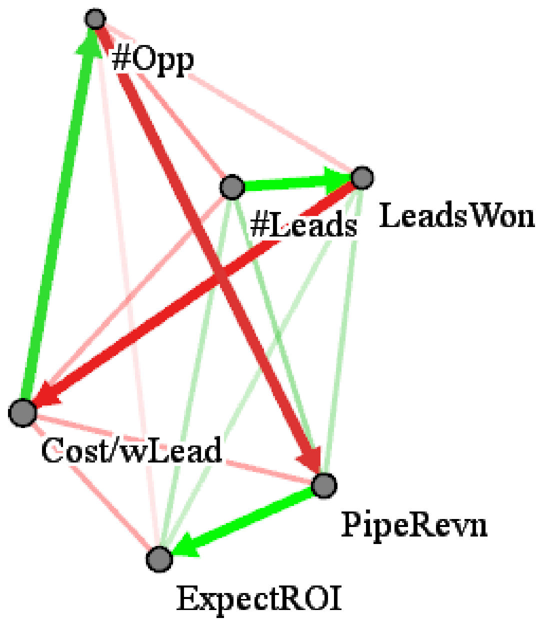
- so she uses the mouse and designs the route shown

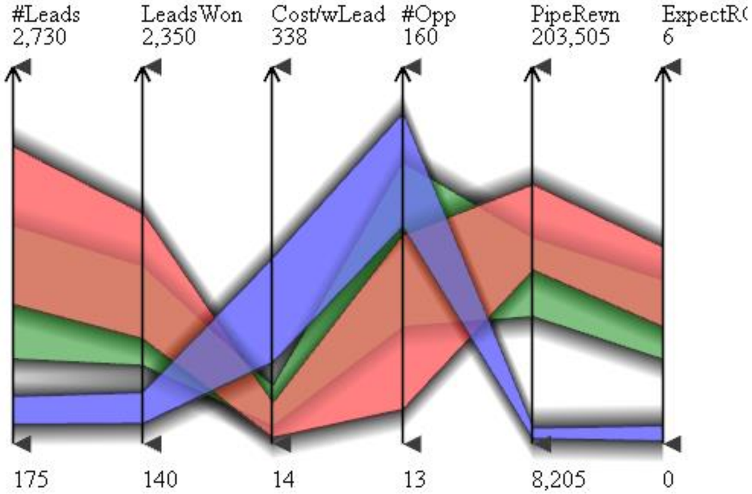


Kate's Story

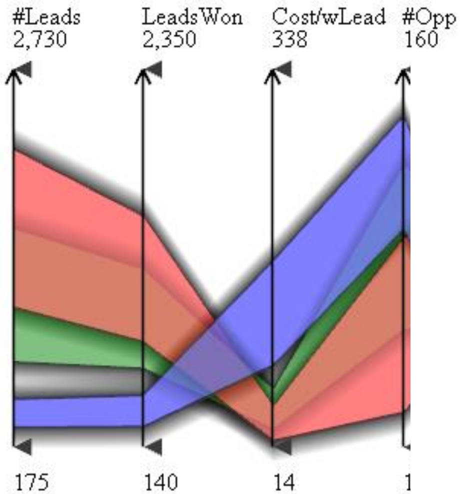
"Let's go and make a more revealing route!"

- so she uses the mouse and designs the route shown





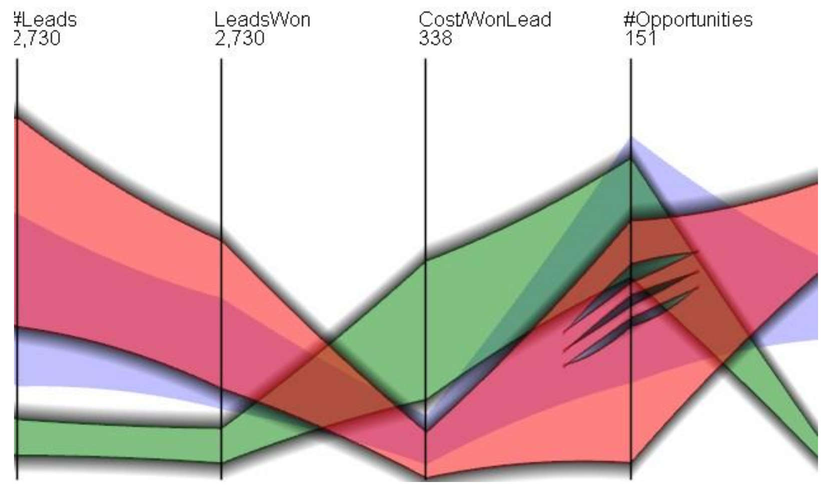
The Big Insight



It is now immediately obvious:

- the blue team employs a very different strategy than the green and the red teams.
- it generates far fewer leads but spends much more resources on each → this gives it an advantage in the final outcome.
- the blue team is also much more consistent than the other teams, as indicated by the much narrower band
- what else can we see?

Further Insight



Kate notices something else:

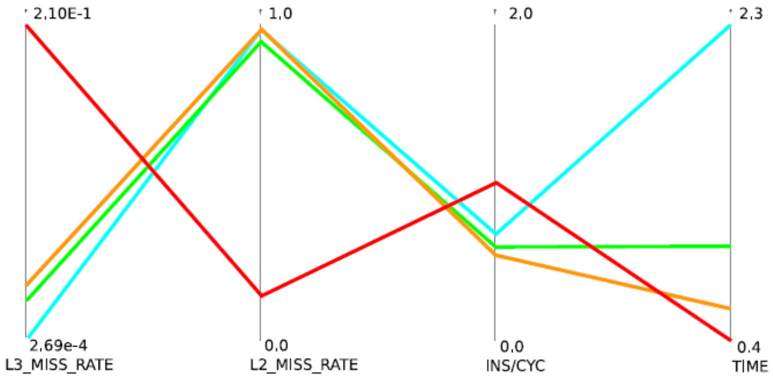
- now looking at the red team
- there seems to be a spread in effectiveness among the team
- the team splits into three distinct groups

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

Recent Reviewer Comment

From a paper sent to a software visualization conference:

Figure 8

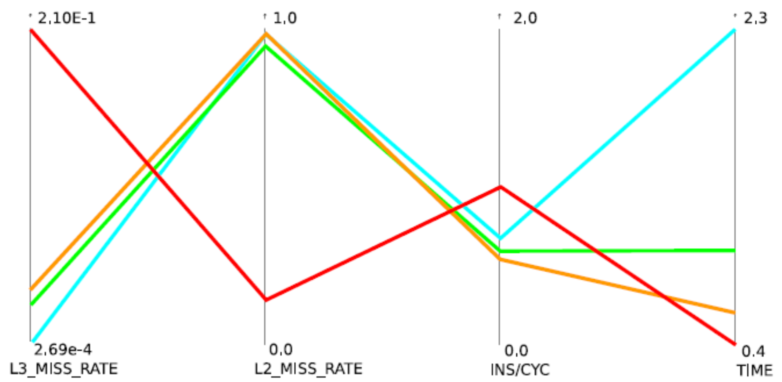


- Multiple visualizations appear to present categorical data as line graphs, which seems a strange choice.

Recent Reviewer Comment

From a paper sent to a software visualization conference:

Figure 8

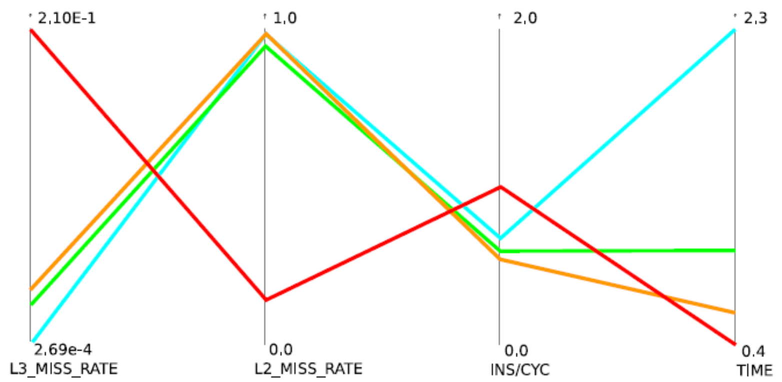


- Multiple visualizations appear to present categorical data as line graphs, which seems a strange choice. Figure 8, for example, at first sight appeared to be showing a change over time, but in fact further inspection shows that the different x-coordinates are almost entirely unrelated to one another 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.

How to Teach Mainstream Users

Learning Visualizations by Analogy

Puripant Ruchikachorn and Klaus Mueller

Stony Brook University

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.

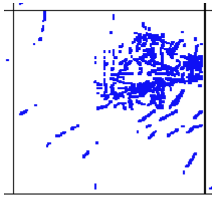
Results

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

Scatterplot For Two Attributes

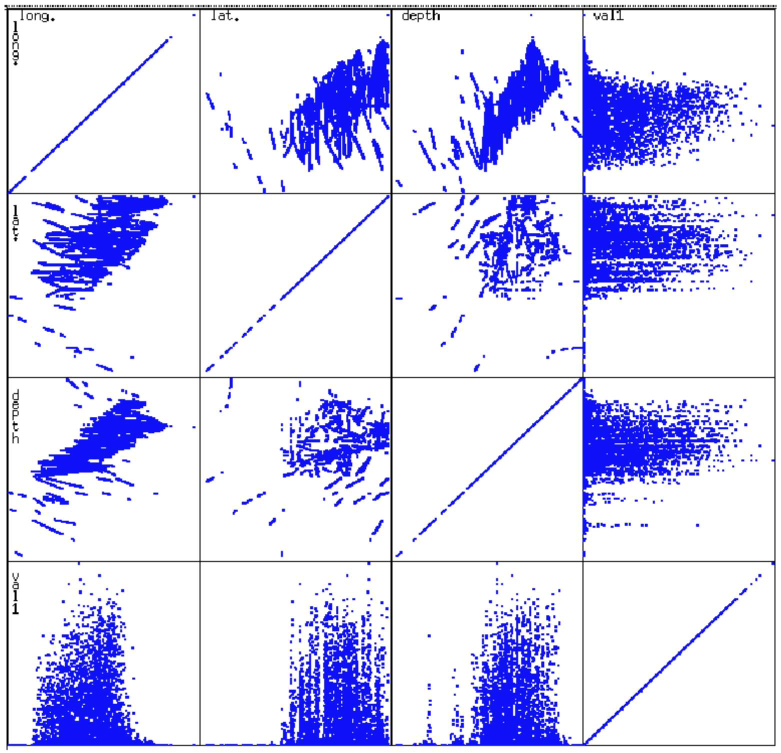
Appropriate for the display of bivariate relationships



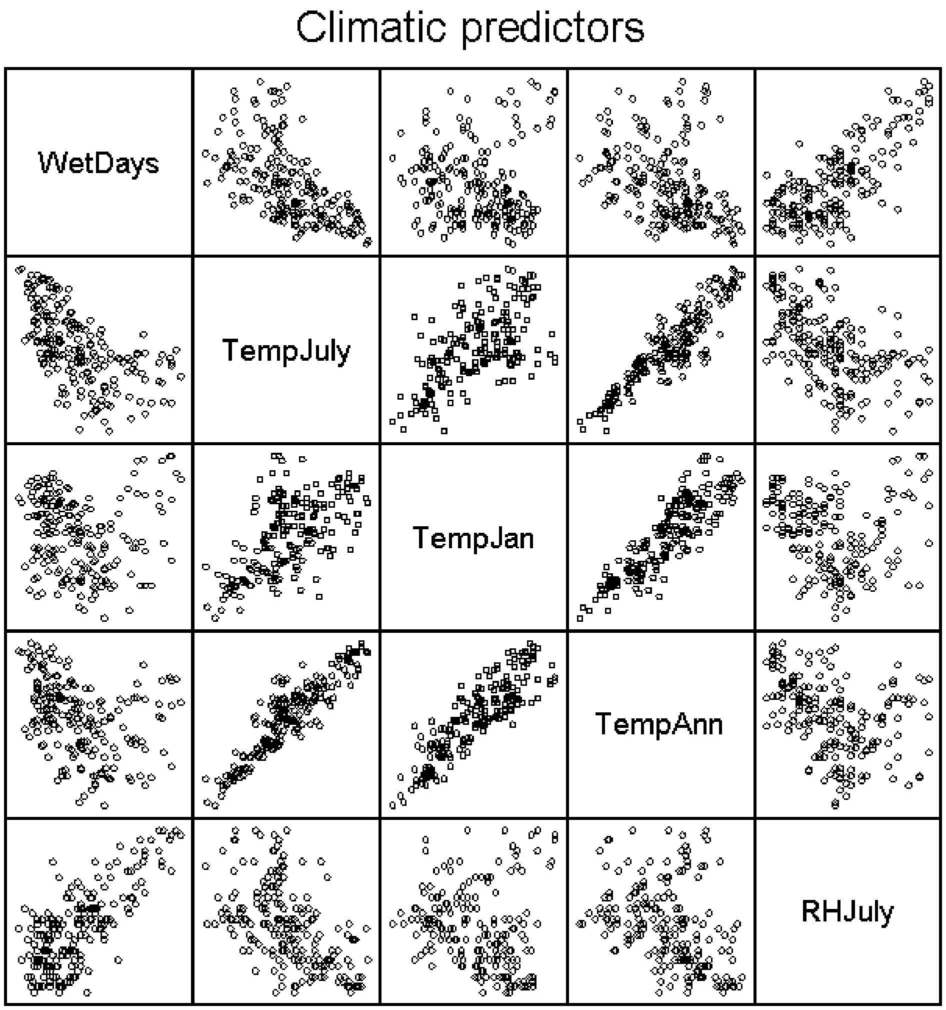
Scatterplot For many Attributes

What to do when there are more than two variables?

- can arrange multivariate relationships into scatterplot matrices
- not overly intuitive to perceive multivariate relationships



Scatterplot Matrix (SPLOM)



Scatterplot Matrix

Scatterplot version of parallel coordinates

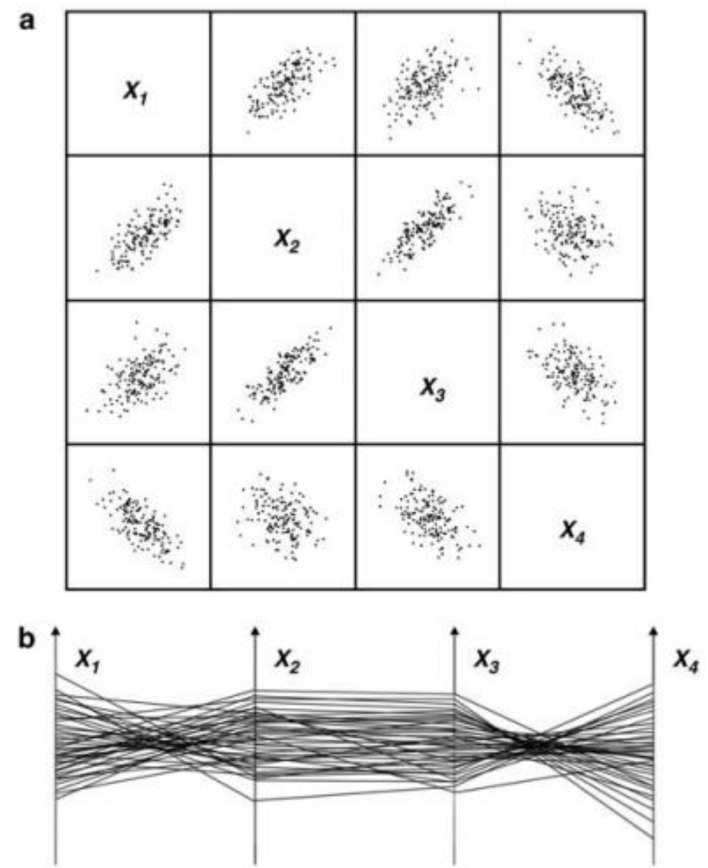
- distributes n(n-1) bivariate relationships over a set of tiles
- for n=4 get 16 tiles
- can use n(n-1)/2 tiles

For even moderately large n:

- there will be too many tiles

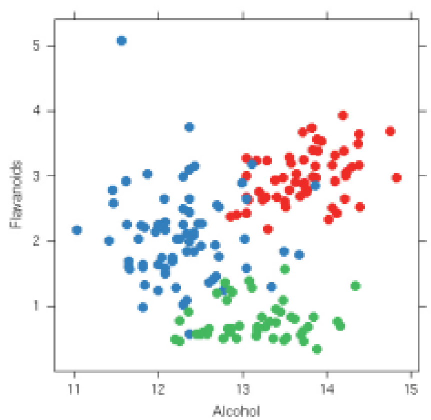
Which plots to select?

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

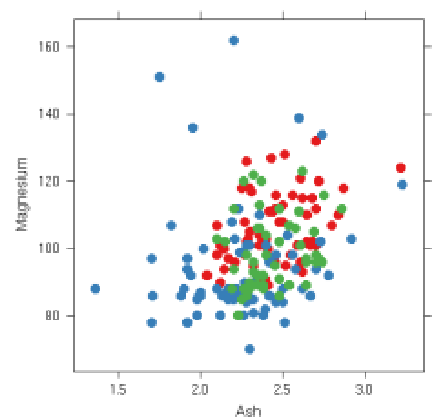


Automated Scatterplot Selection

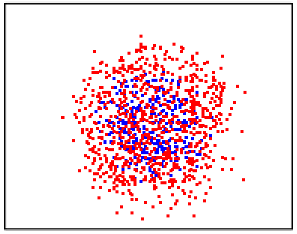
Several metrics, a good one is Distance Consistency (DSC)



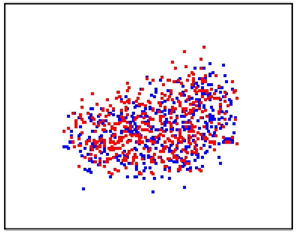
(a) **DSC=90**



(b) **DSC=49**

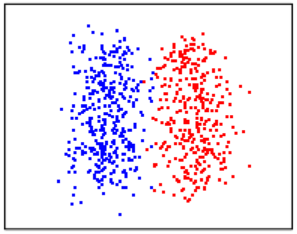


(d) 29

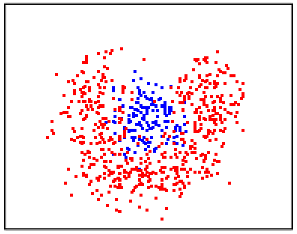


(e) 15

bad



(a) 99



(b) 74

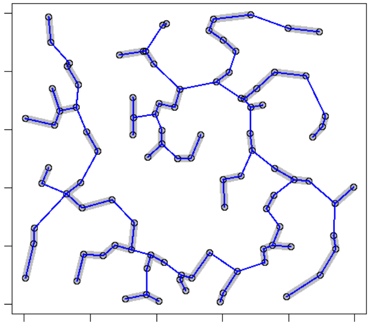
OK

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

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

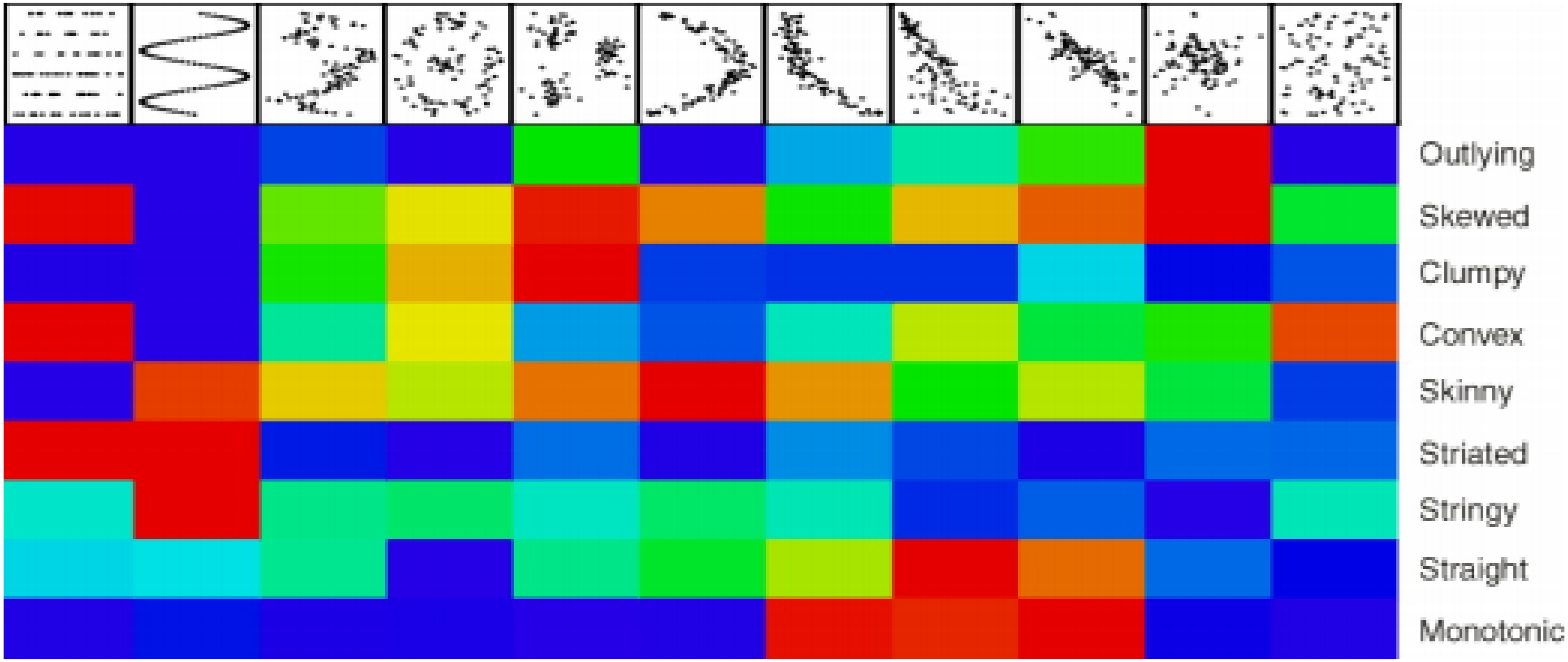
M. Sips et al., Computer Graphics Forum, 28(3): 831–838, 2009

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