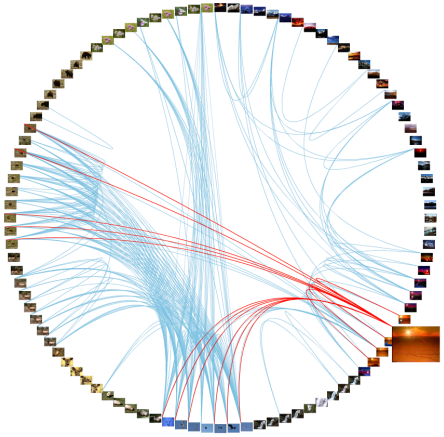
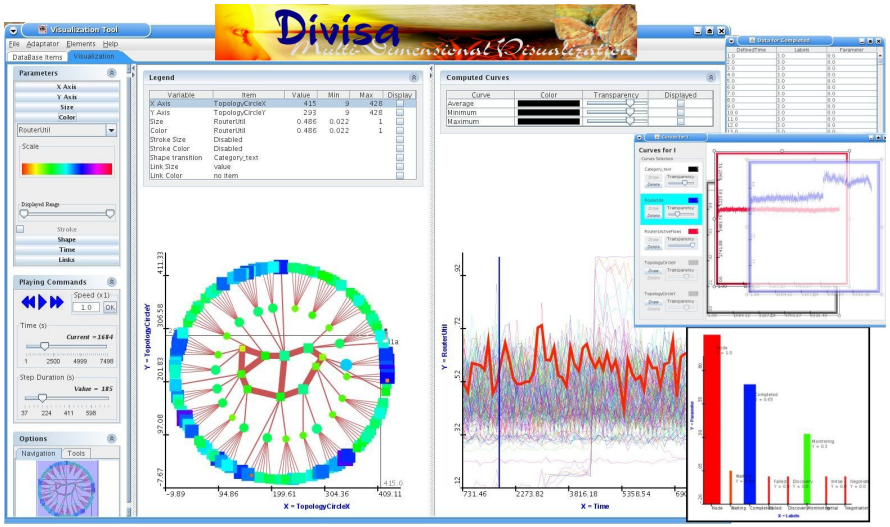
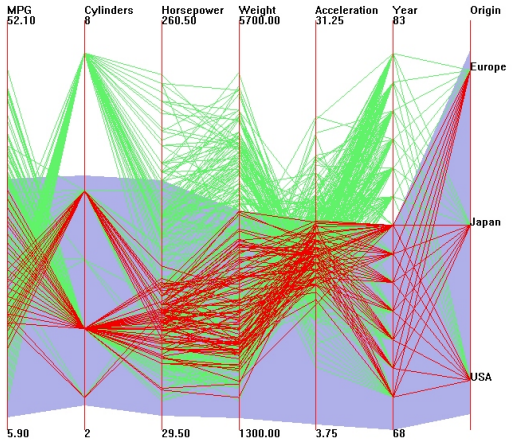




IAT 814 Visualization Multidimensional Methods

Lyn Bartram



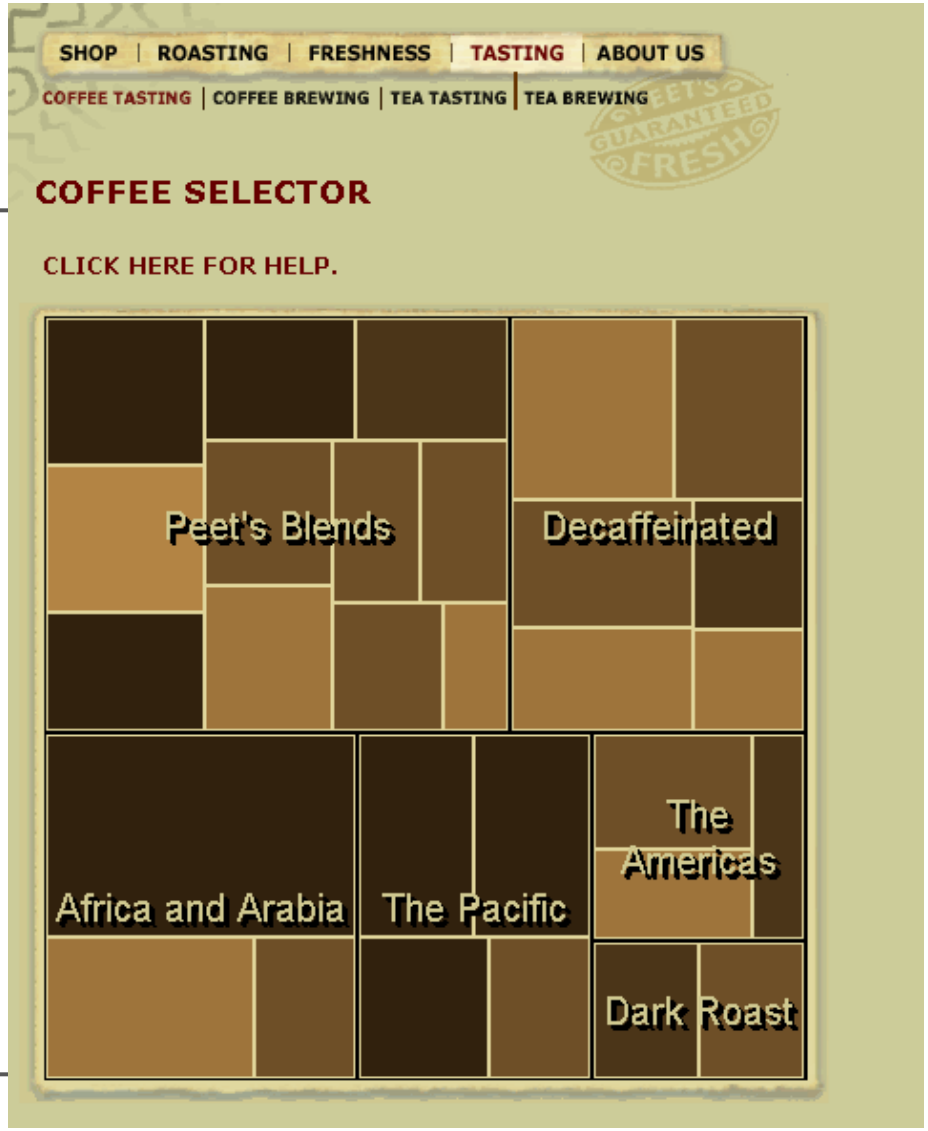
SFU

SCHOOL OF INTERACTIVE ARTS + TECHNOLOGY

Spatial layouts cont.

Space-filling display

- total area used by the layout is equal to the total area available in the view
- Affords powerful use of colour
- No white space

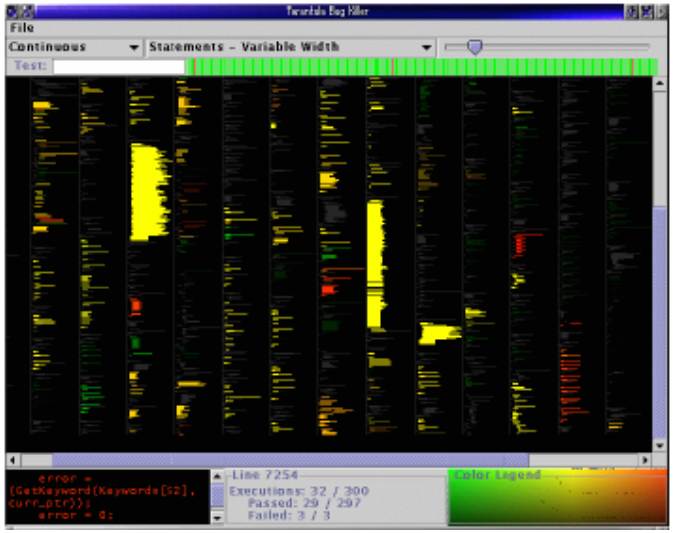


Spatial Layouts cont from last week

Pixel –oriented displays

- Also called “dense” displays
- Basic idea: use tightly packed tiny marks, often a single pixel in size
- Only 2d position and colour useful

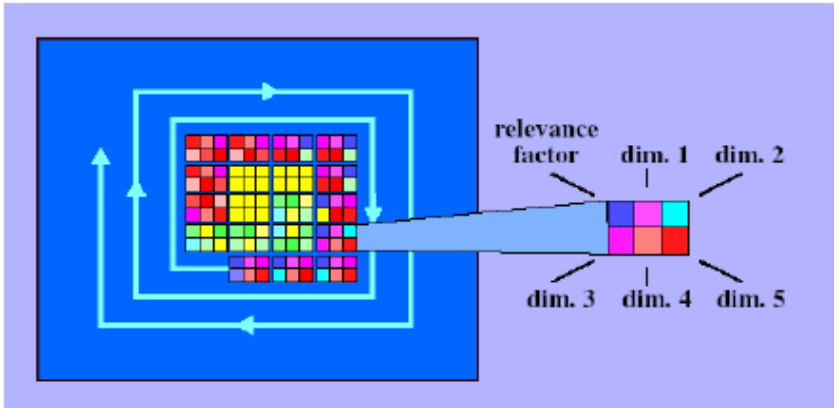
Idiom	Dense Layouts
What: Data	Text with numbered lines (source code, test results log).
What: Derived	Two quantitative attributes (test execution results).
How: Encode	Dense layout. Spatial position and line length from text ordering. Color channels of hue and brightness.
Why: Task	Locate faults, summarize results and coverage.
Scale	Lines of text: ten thousand.



J.A.. Jones, M.J. Harrold, and J. Stasko. “Visualization of Test Information to Assist Fault Localization.” Proceedings of the International Conference on Software Engineering (ICSE), pp. 467–477. ACM, 2002.

VisDB

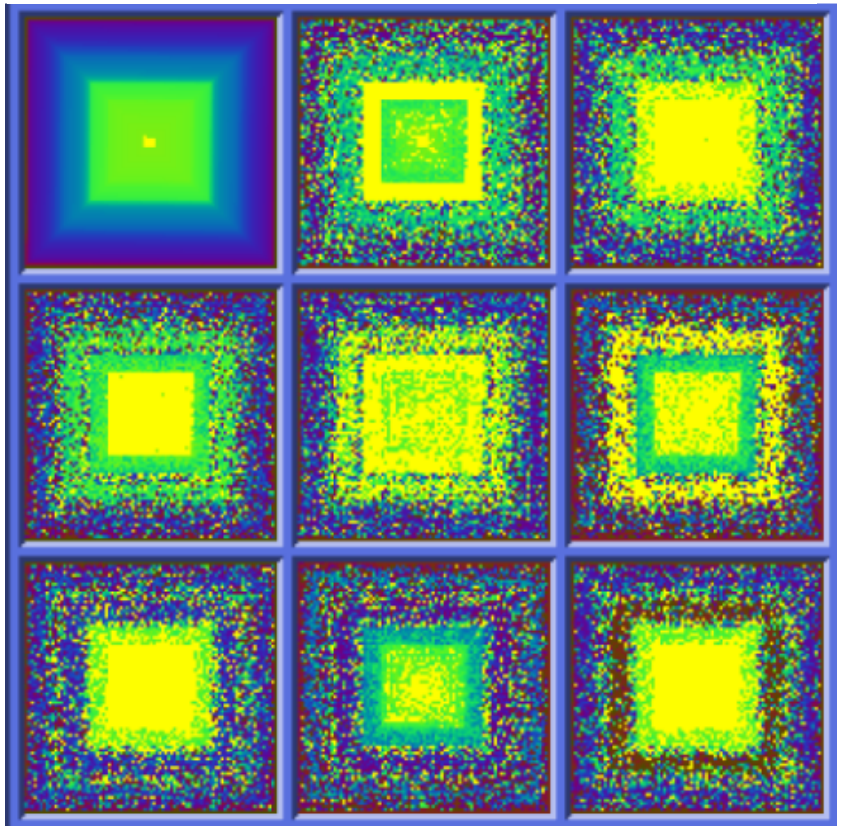
- Pixels are colored according to 5 attributes and relevance
  - 6 attribute glyph
- Local ordering
  - Spiral path

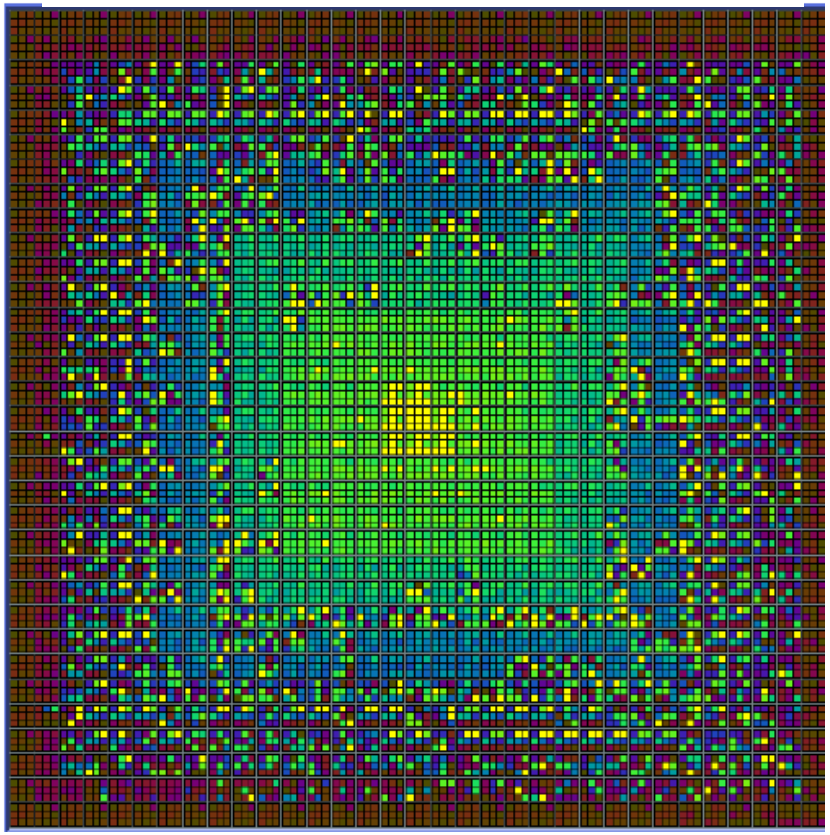


VisDB: Database Exploration using Multidimensional Visualization, Keim and Kriegel, IEEE CG&A, 1994

VisDB:

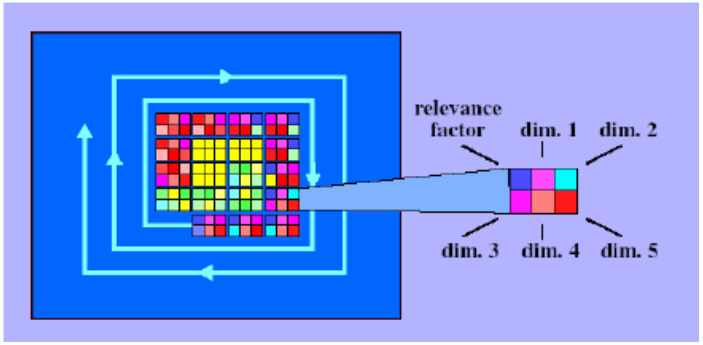
- One view per attribute, coloured by relevance —— Aggregate view using multi-attribute **glyph** –



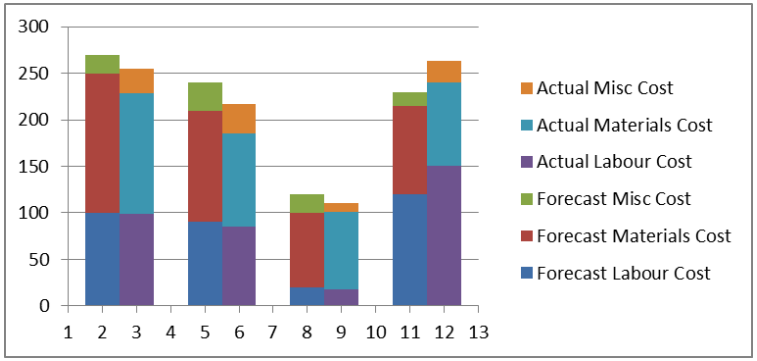


Glyphs

- VisDB uses multi-dimensional **glyphs**



- Stacked bars are simple glyphs



Glyphs

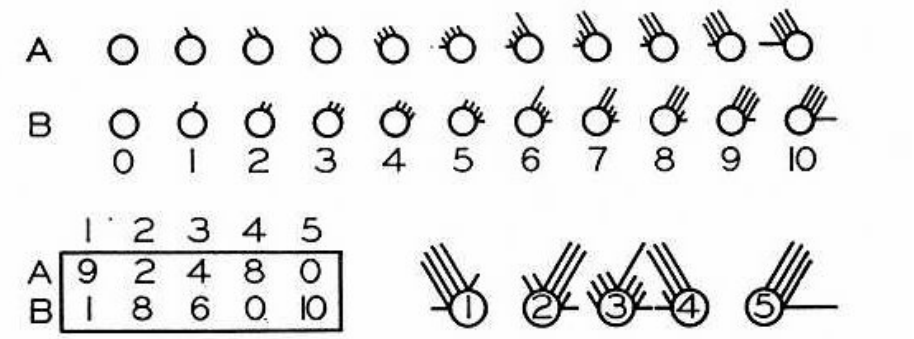
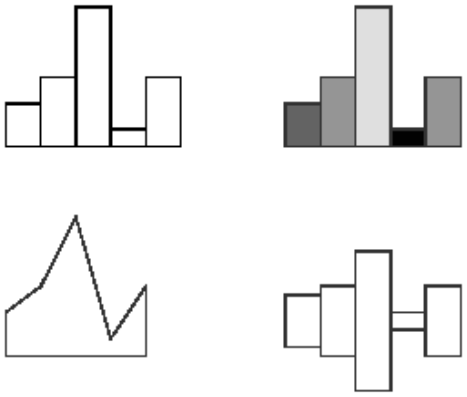


FIGURE II. Top: variables A and B are coded in values from 1 to 10 each. Lower left: the records of five individuals with respect to A and B. Lower right: the corresponding metroglyphs for each of the five individuals.

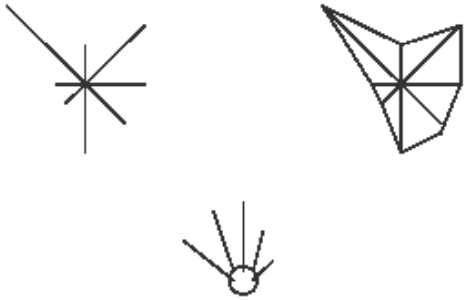
- Glyphs are graphical entities which convey one or more data values via attributes such as shape, size, color, and position

- geometric attributes: shape, size, orientation, position, direction/magnitude of motion
- appearance attributes: color, texture, and transparency

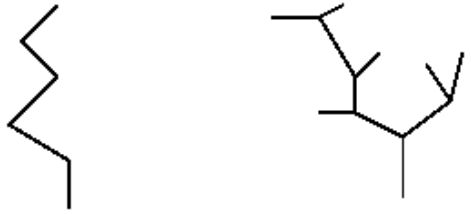
Examples



Variations on Profile glyphs

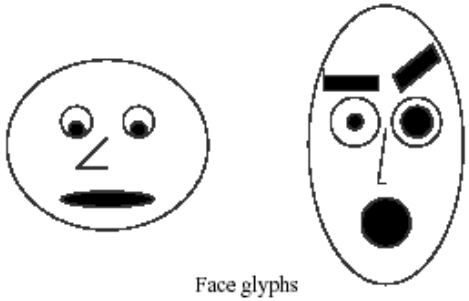


Stars and Anderson/metroglyphs



Sticks and Trees

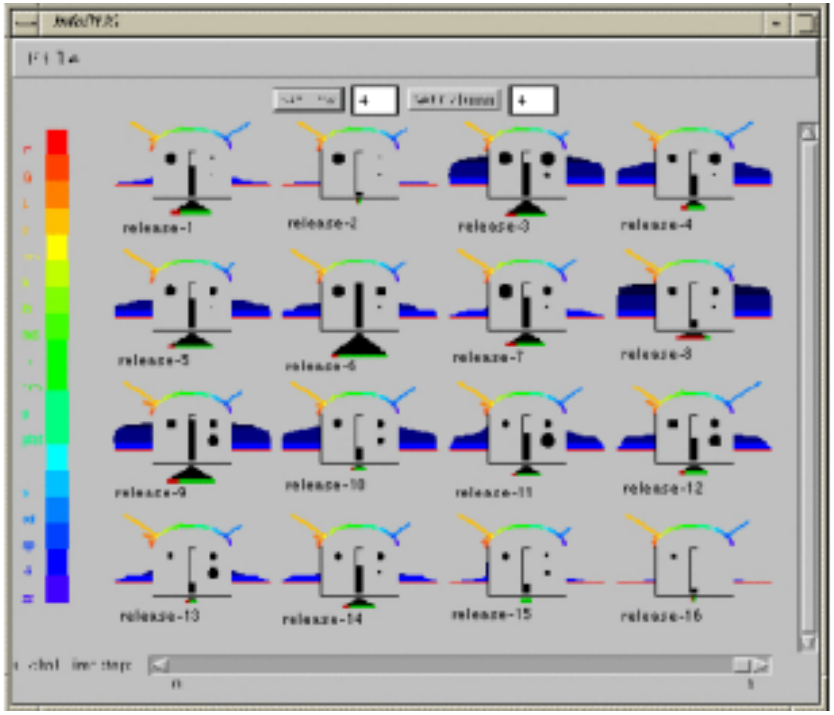
Autoglyph and box glyph

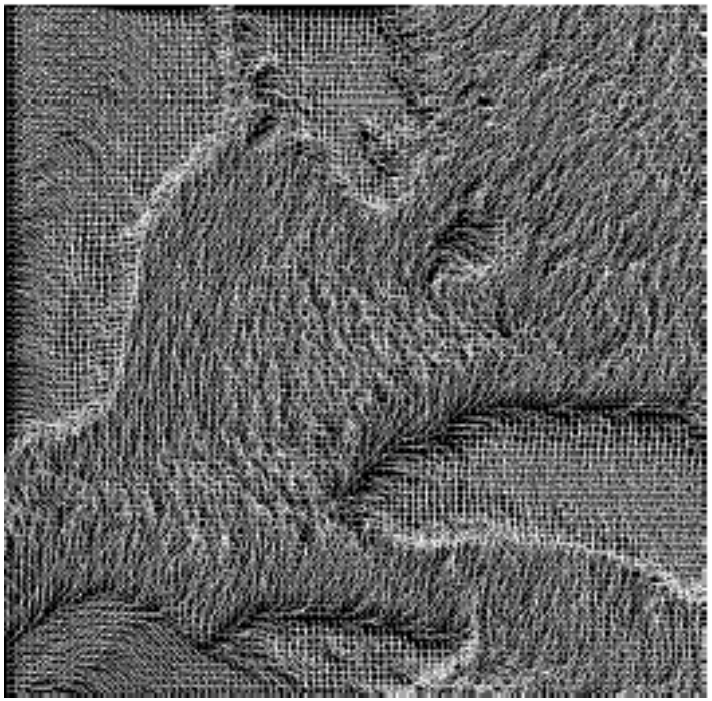


Face glyphs

Arrows and Weathervanes

Macro (icons) vs micro (texture)

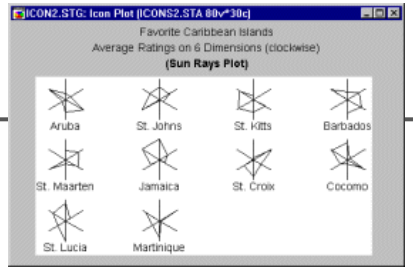




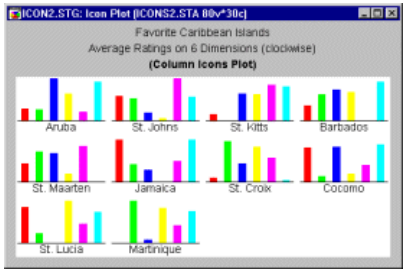
Icon plots

- Each icon represents a singular multivariate case :emergent visual appearance as a unique “visual identity”
- Useful for small data sets with up to 10 or so variables
- Limitations?
  - Small data sets, small dimensions
  - Ordering of variables may affect perception

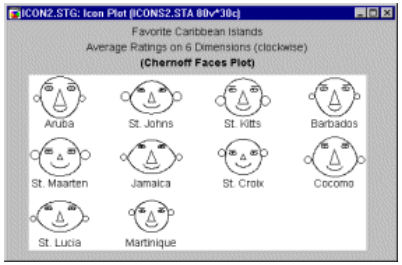
- Circular



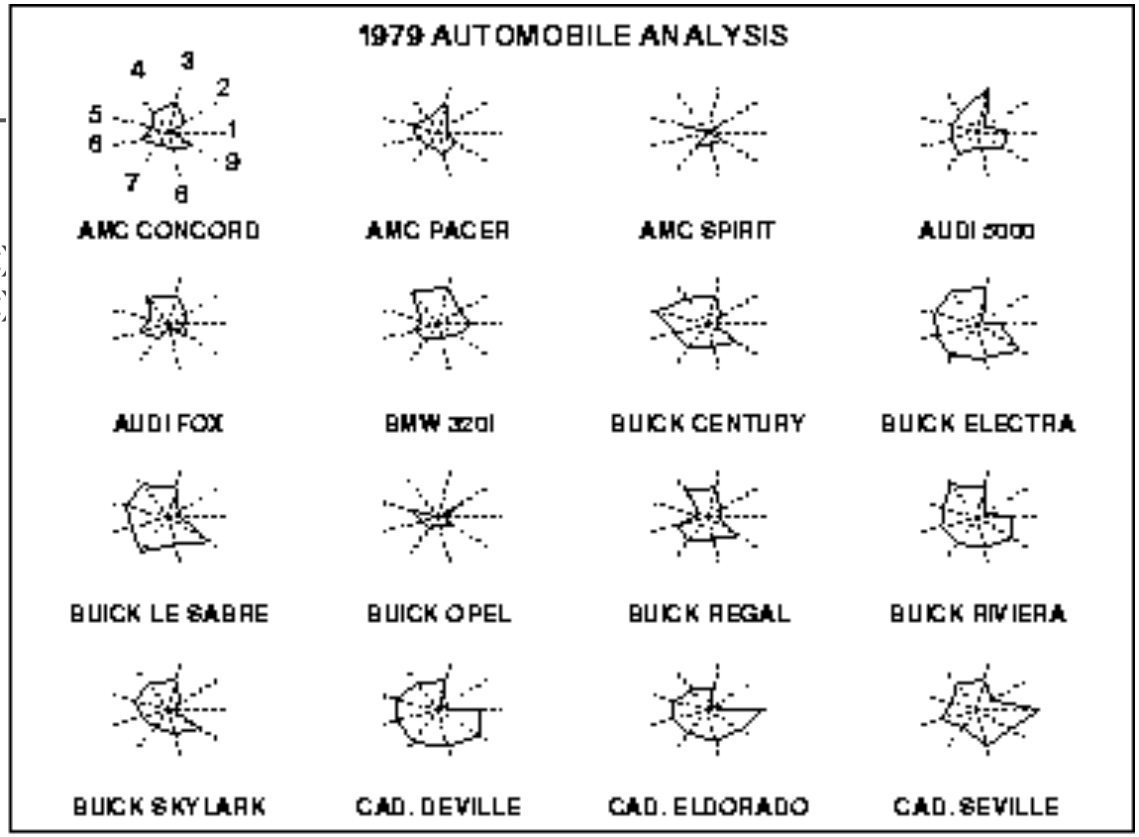
- Sequential



- Faces



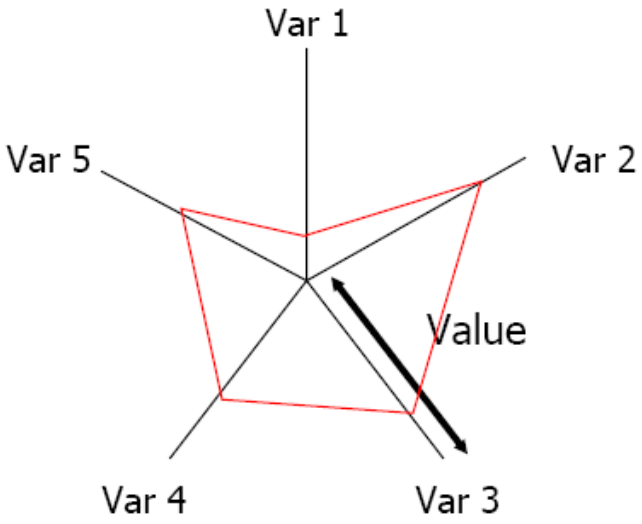
1 Price
2 Mileage (MPG)
3 1978 Repair Record (1 = Worst, 5 = Best)
4 1977 Repair Record (1 = Worst, 5 = Best)
5 Headroom
6 Rear Seat Room
7 Trunk Space
8 Weight
9 Length



Star Plots

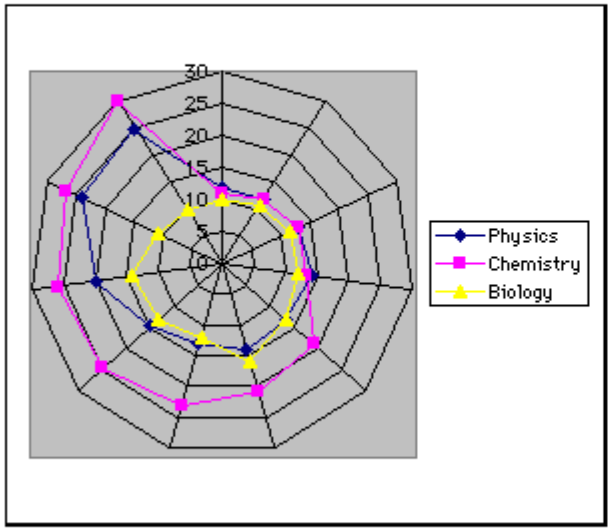
- Space out the n variables at equal angles around a circle
- Each “spoke” encodes a variable’s value
- Data point is now a “shape”

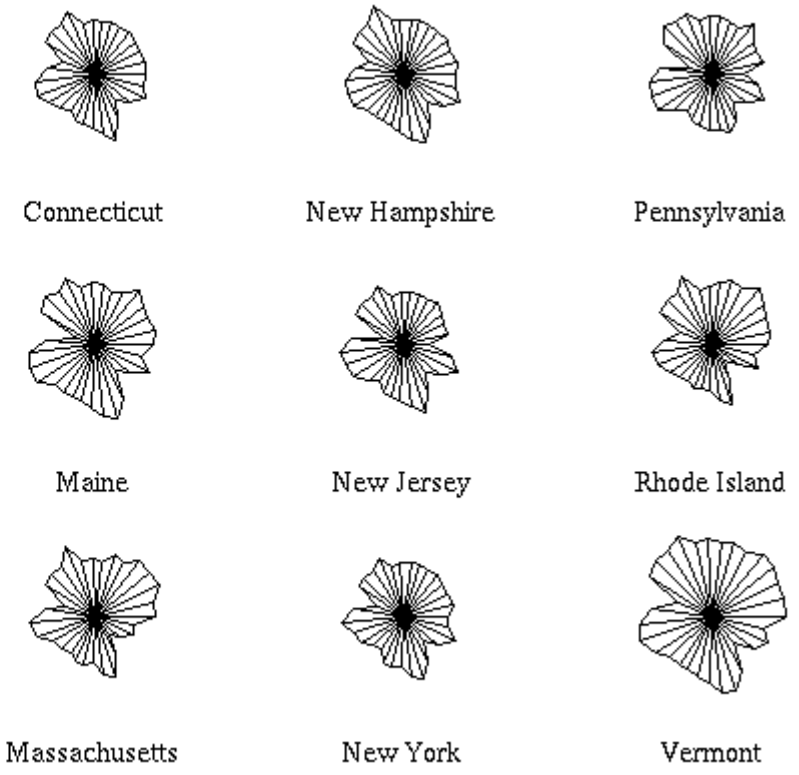
- Limitations
  - Small data sets, small dimensions
  - Ordering of variables may affect interpretation



Star Plot Examples

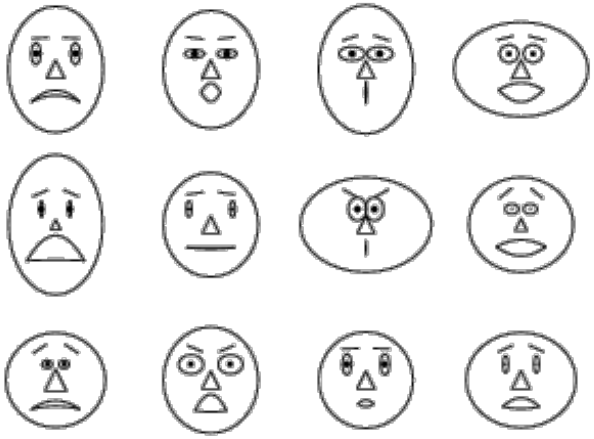
Dimensional Reorganization





Chernoff faces

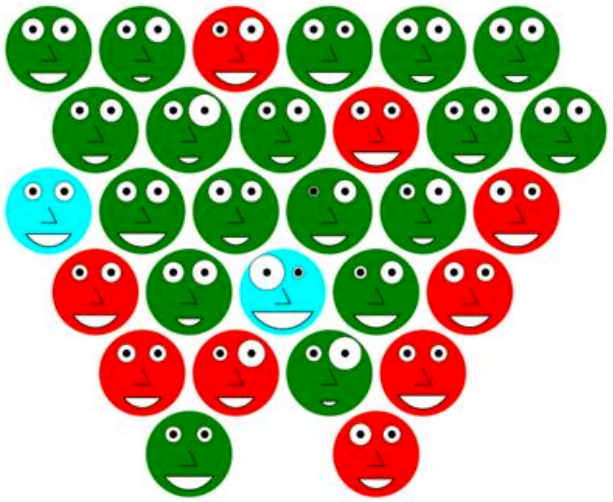
Ten parameters: head eccentricity, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, eye spacing, eye size, mouth length and degree of mouth opening



Gonick, L. and Smith, W. The Cartoon Guide to Statistics. New York: Harper Perennial, p. 212, 1993

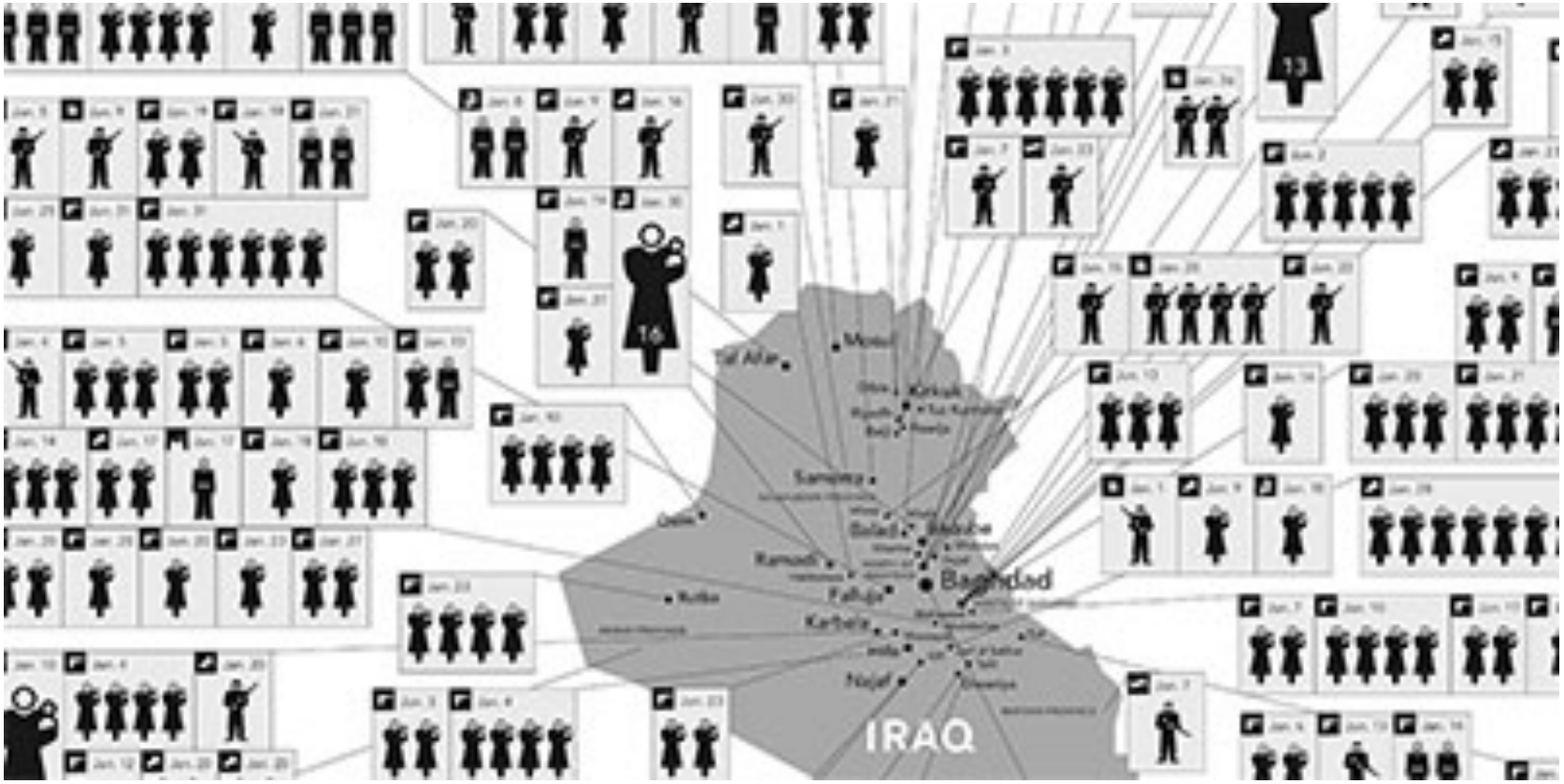
Chernoff faces applet

- In use



G. Wills. Funny Faces: Visualizing Many Variables. IBM AnalyticsZone blog, 2013.

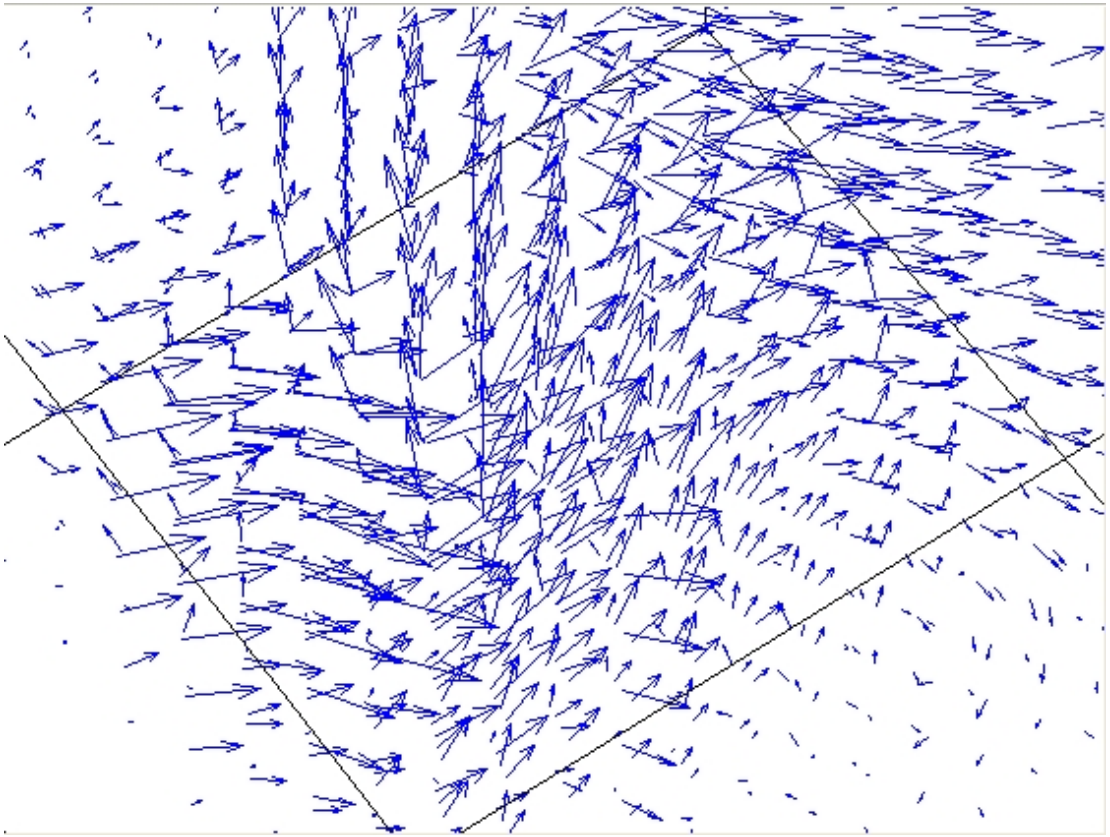
Infographics and glyphs: are they the same?

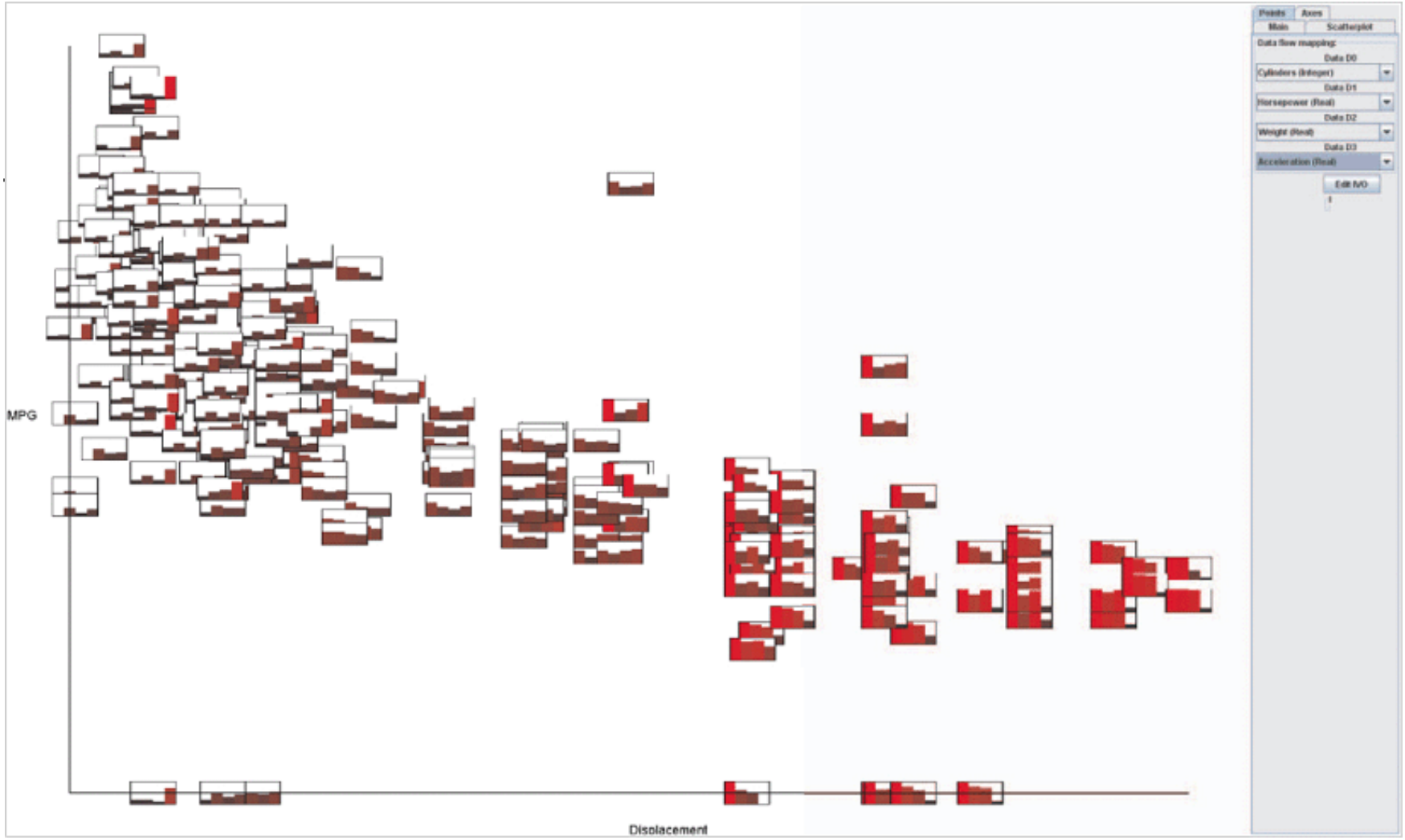


Glyph Placement Issues

1) *data-driven* (e.g., based on two data dimensions) vs. structure-driven (e.g., based on an order (explicit or implicit) or other relationship between data points)
2) Overlaps vs. non-overlaps
3) optimized screen utilization (e.g., space-filling algorithms) vs. use of white space to reinforce distances
4) Distortion vs. precision

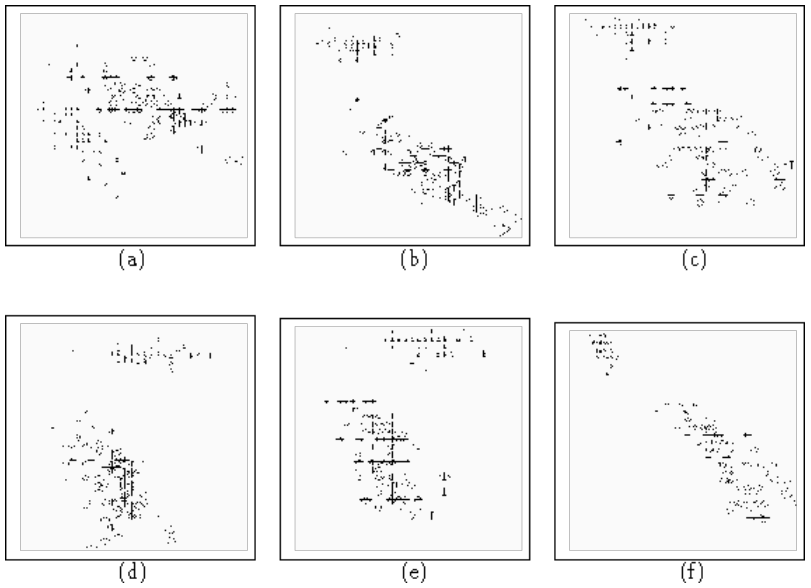
Glyph placement is an issue




MPG
Displacement
Points
Axes
Main
Scatterplot
Data flow mapping:
Data D0
Cylinders (Integer)
Data D1
Horsepower (Real)
Data D2
Weight (Real)
Data D3
Acceleration (Real)
Edit IVO

Data-driven placement

- One, two or three of the data dimensions are used as positional components
- (Subsetting!)
- DDGP – Data driven glyph placement



http://davis.wpi.edu/~matt/courses/glyphs/node6.html

SFU

Data-driven placement.

- Conveys detailed relationships between dimensions selected
- Ineffective mapping => substantial cluttering and poor screen utilization.
- Some mappings may be more meaningful than others (But, which one?).
- Bias given to dimensions involved in mapping. Thus, conveys only pairwise (or three-way, for 3-D) relations between the selected dimensions.
- Most useful when two or more of the data dimensions are spatial in nature.

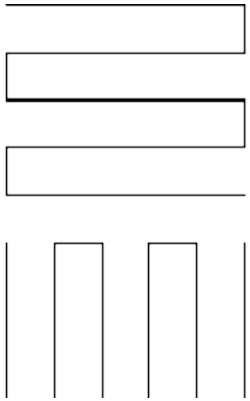
Data-Driven Placement Cont.

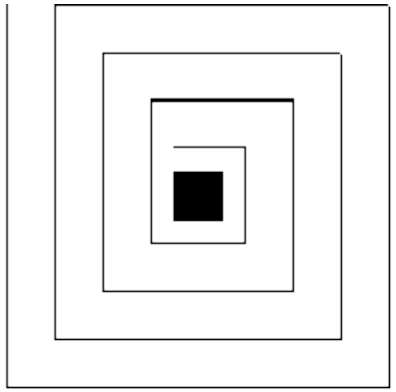
- Issues: reduce clutter and overlap
- Solution: Distortion
  - Random Jitter
  - Shift positions to minimize or avoid overlaps.
- But, how much distortion allowed?
- Selectively vary the level of detail shown in the visualization

Structure-Driven Glyph Placement

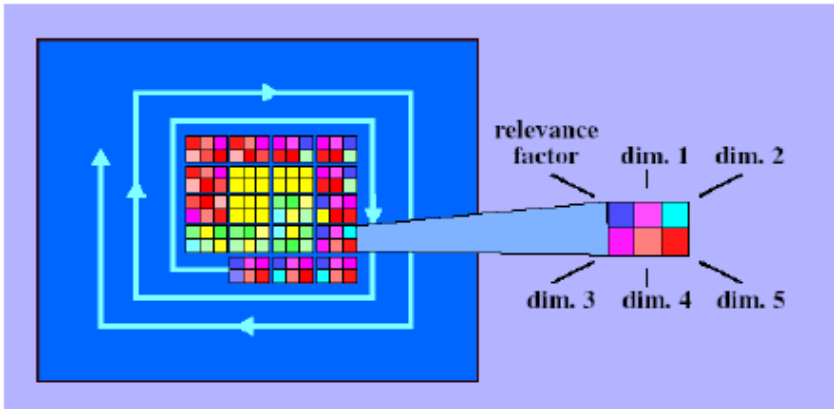
- Structure implies relationships or connectivity
- Explicit structure (one or more data dimensions drive structure) v.s.
- Implicit structure (structure derived from analyzing data)
- Common structures: ordered, hierarchical, network/ graph

Ordered Structure





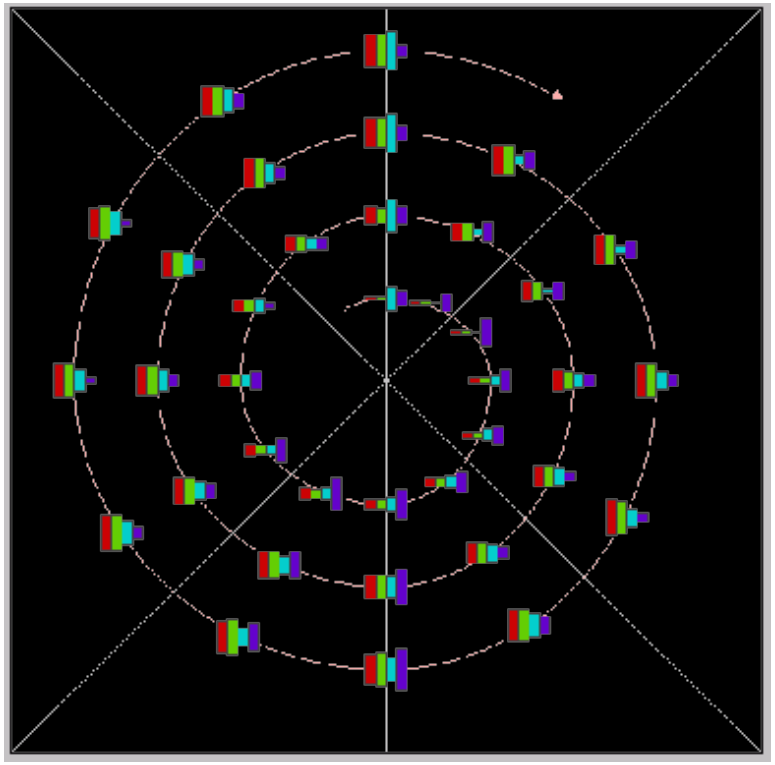
- May be linear (1-D) or grid-based (***N***-D)
- Good for detection of changes in the dimensions used in the sorting



SFU

SDGP – Ordered Structure Cont.

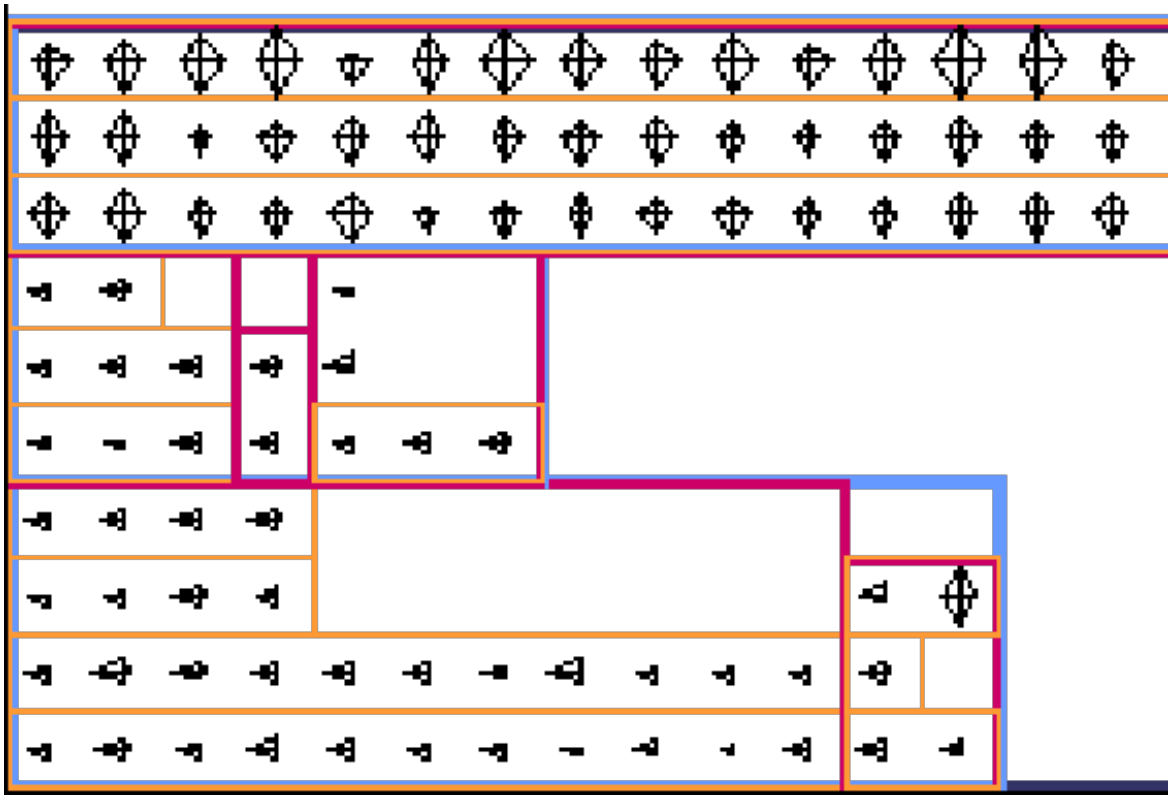
- Common linear ordering include raster scan, circular, and recursive space-filling patterns
- Dimensions (from left to right): Dow Jones average, Standard and Poors 500 index, retail sales, and unemployment.
- Data for December radiate straight up (the 12 o'clock orientation). Low unemployment, High Sales
- Note distance and size are not constant between years, just months



SFU

SDGP – Hierarchical Structure

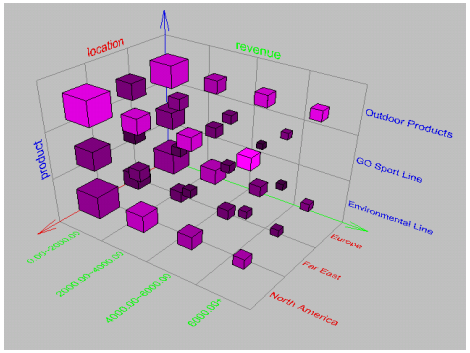
- Treemaps



SFU

The Core Problem Statement

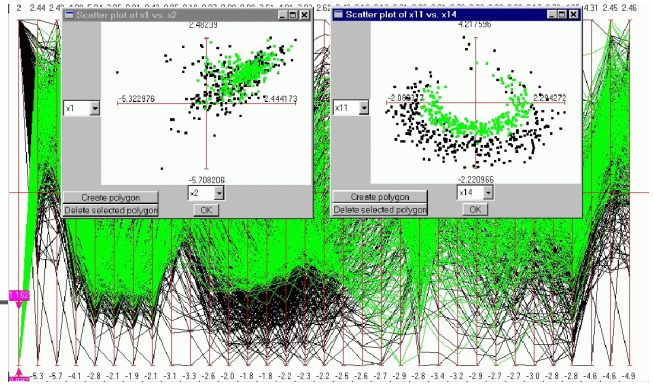
- How to effectively present more than 3 dimensions of information in a visual display with 2 (to 3) dimensions?
  - very large, multivariate data sets?

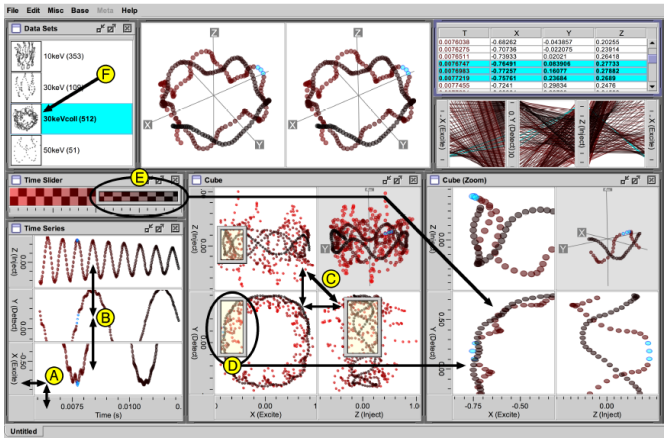


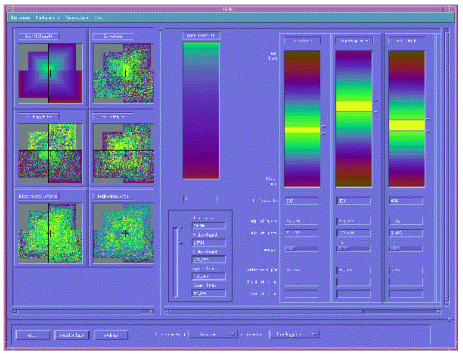
- No loss of information [Inselberg]
  - Minimal complexity
  - Any number of dimensions
  - Variables treated uniformly
  - Objects remain recognizable across transformations
  - Easy / intuitive conveyance of information
  - Mathematically / algorithmically rigorous

Characteristics

- “Data-dense displays” (large number of dimensions and/ or values)
  - Often combine color with position / proximity representing relevance “distance”
- Often provide multiple views
  - Retinal properties of marks
  - Gestalt concepts, e.g., grouping

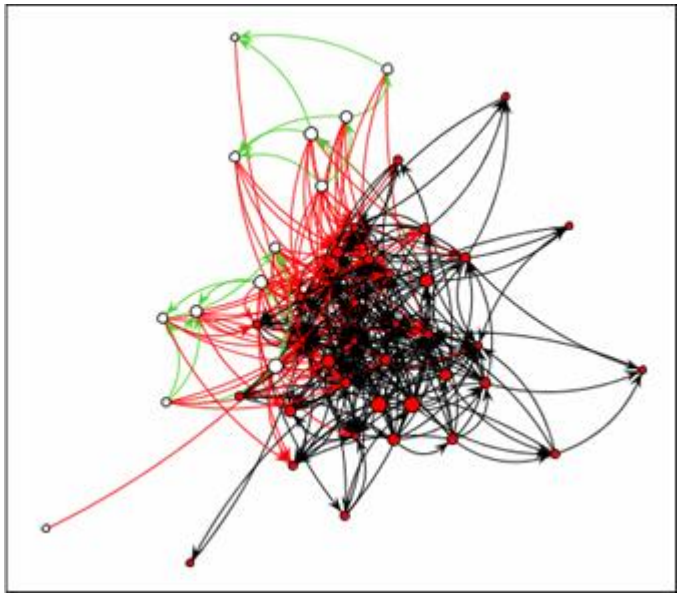


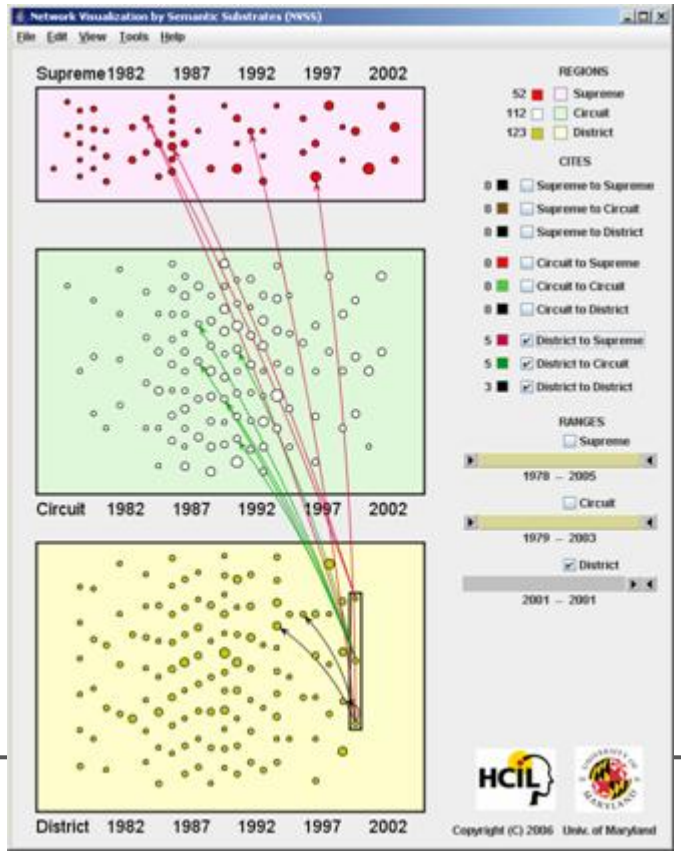




Network Visualization by Semantic Substrates
[Shneiderman & Aris, IEEE TVCG 2006]

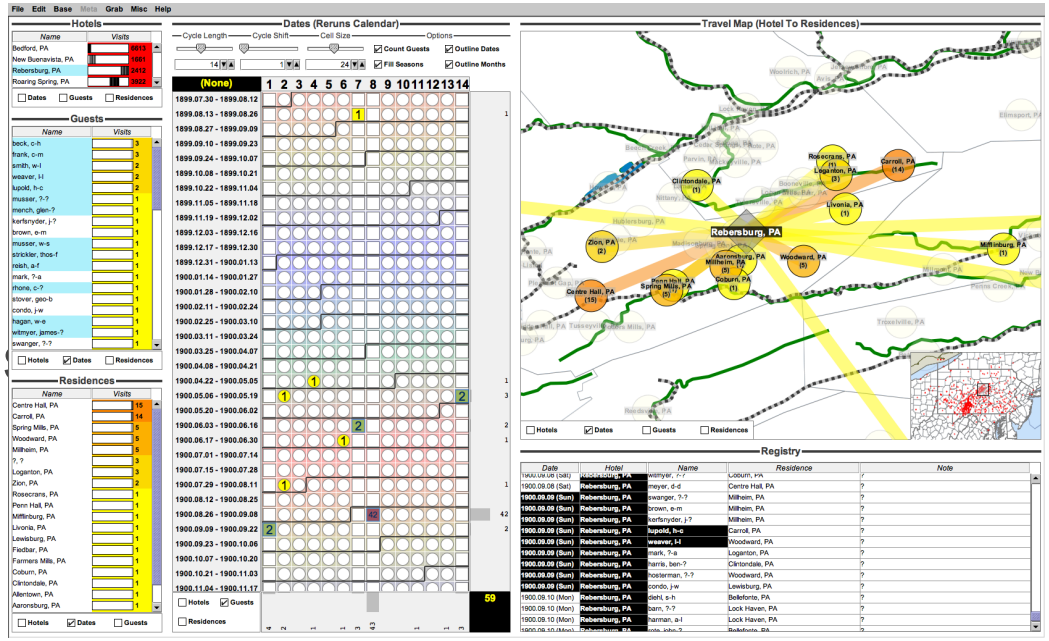
- http://hcil.cs.umd.edu/video/2006/substrates.mpg





Improvise [Weaver]

- System for building coordinated multiple views
- E.g. Hotel registrations in Pennsylvania



http://www.cs.ou.edu/~weaver/improvise/index.html

PaperLens

- Understanding research trends in conferences using PaperLens Lee et al., CHI'05 extended abstracts
- http://www.cs.umd.edu/hcil/paperlens/PaperLens-Video.mov

MiDAVisT

- 5 categorical and 6 quantitative variables

Multivariate Visualization

- 2 variables: scatter plots, etc
- 3 variables:
  - 3-dimensional plots (Look impressive, but often not used well)
  - Can be cognitively challenging to interpret
  - Alternatives: overlay color-coding (e.g., categorical data) on 2d scatter plot
- 4 variables:
  - 3d with color or time
  - Can be effective in certain situations, but tricky

- Higher dimensions
  - Generally difficult
  - Scatter plots, icon plots, parallel coordinates: all have weaknesses
  - Complexity, coordination and visual fragmentation all problematic

Much More than 3D

- Fundamentally, we have 2 display dimensions
- For data sets with >2 variables, we must project data down to 2D
- Come up with visual mapping that locates each dimension into 2D plane
- Computer graphics 3D->2D projections

Methods for Visualizing Multivariate Data

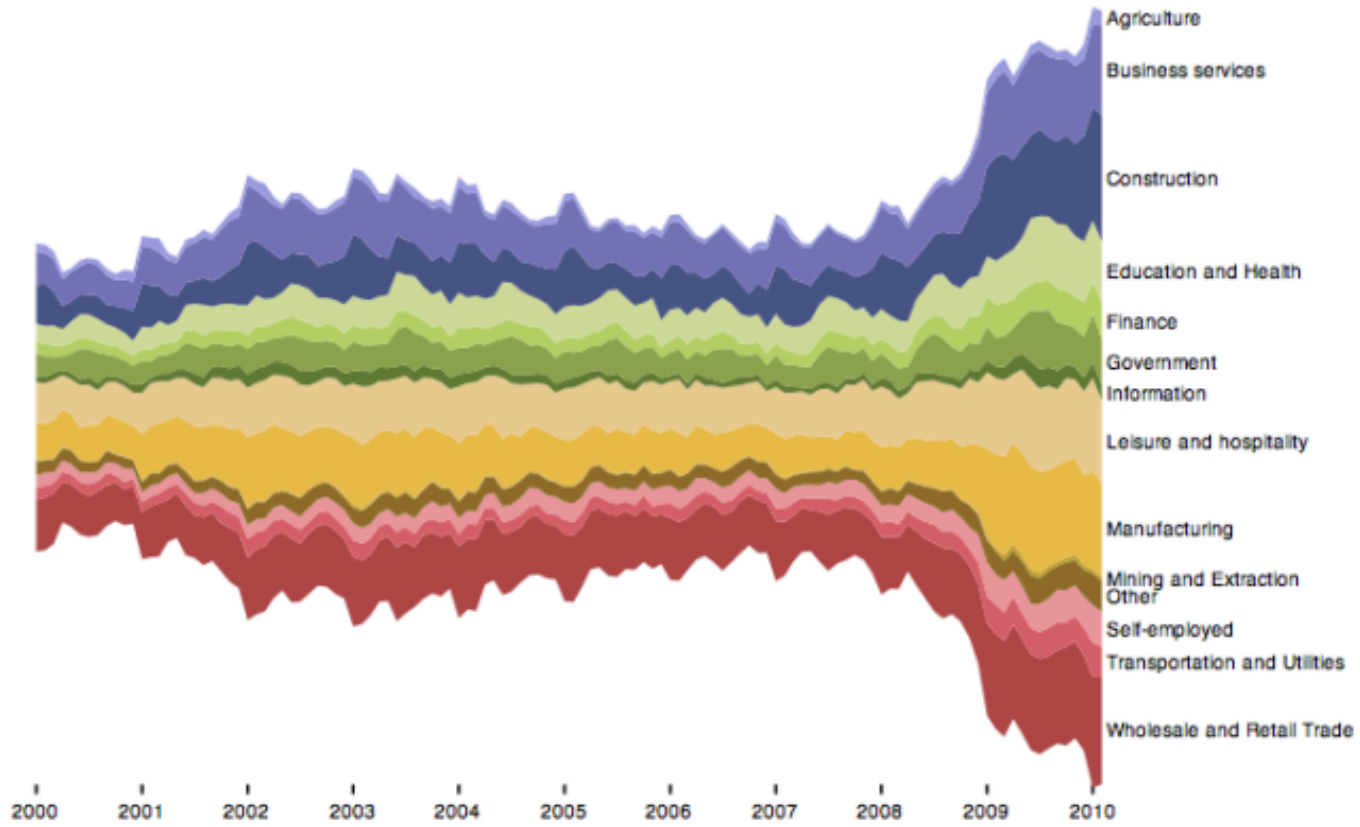
- Univariate splitting (small multiples)
- Dimensional Subsetting (bivariate)
- Dimensional Embedding
- Dimensional Reduction (later)
- Dimensional Reorganization

- Coordinated multiple views

Multiple view methods

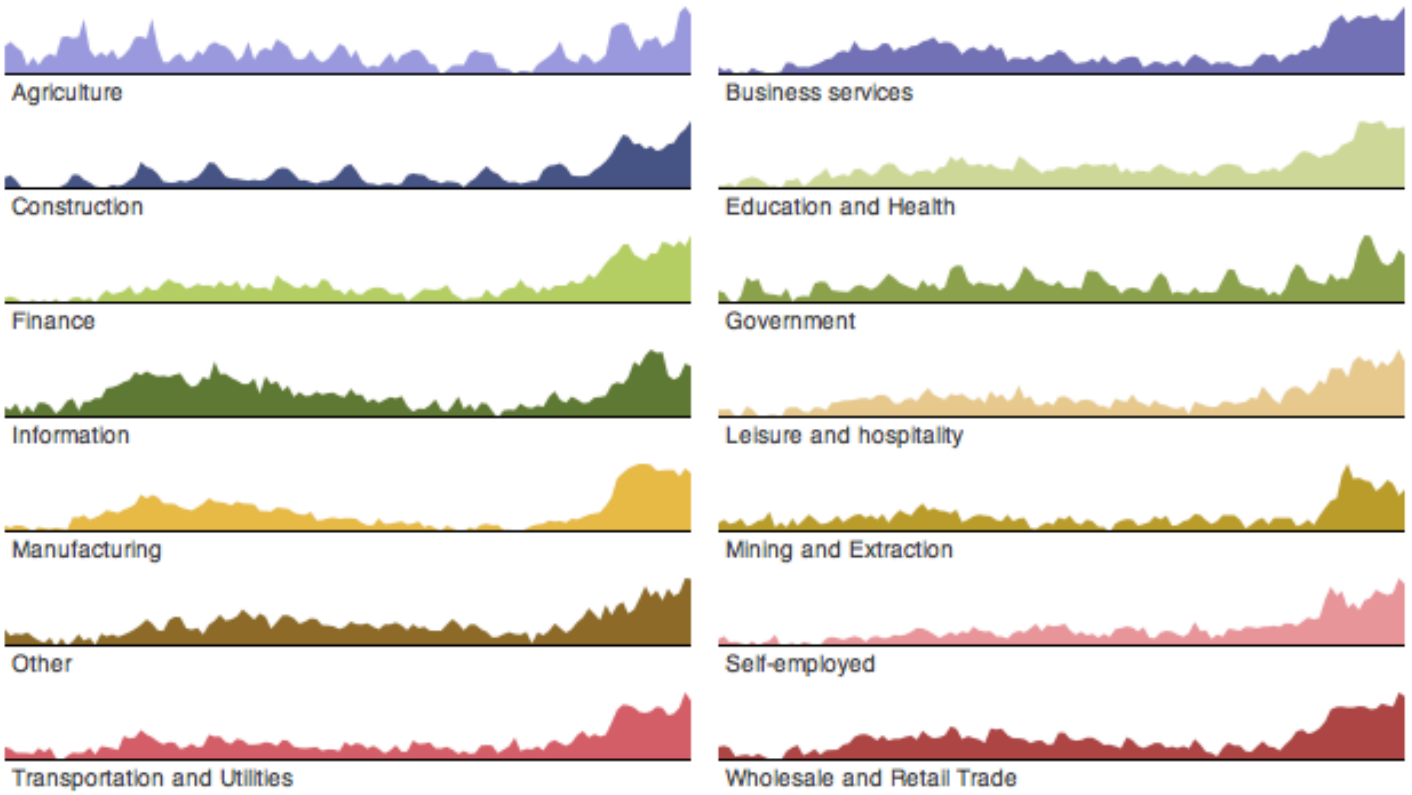
- view choices
  - encoding: same or multiform
  - dataset: same or small multiple
  - data: all or subset (overview/detail)
  - spatial ordering of views
- Animation?
  - Data change over time
- many combinations possible

Why doesn't this work?



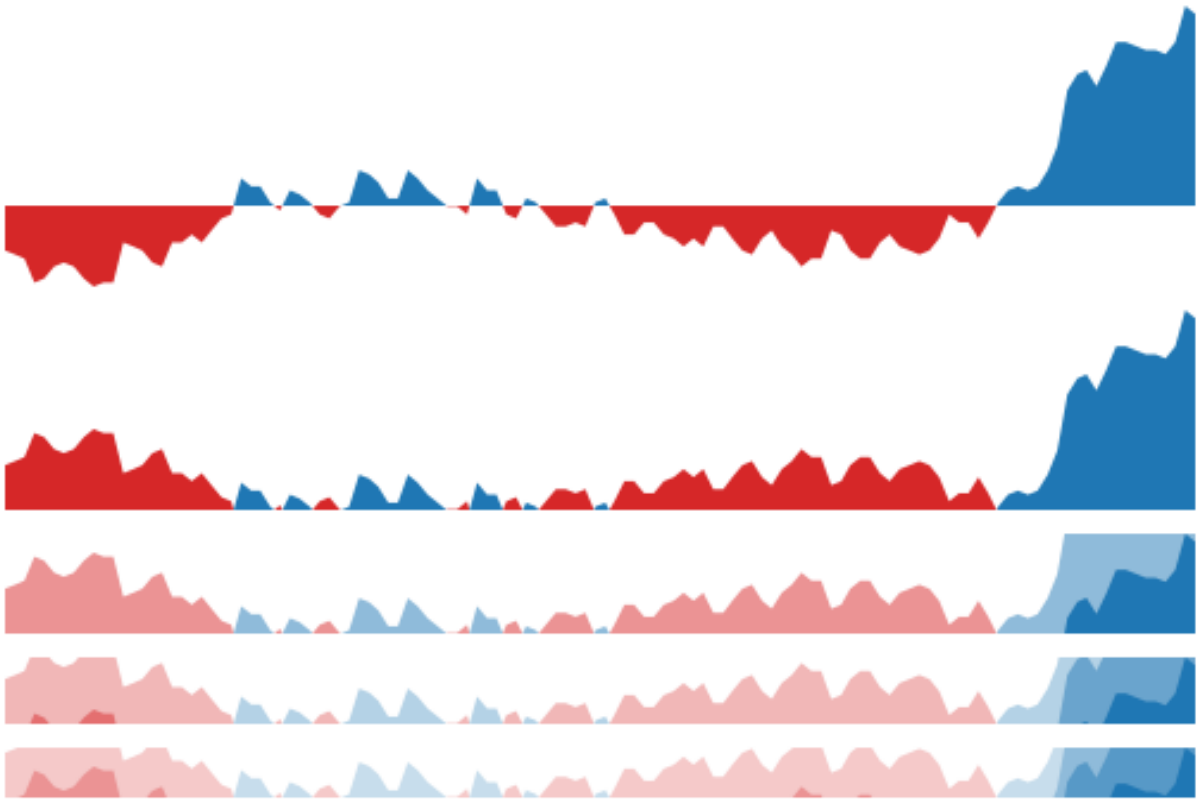
J. Heer, M. Bostock and V. Ogievetsky, A Tour Through the Visualization Zoo.

Small multiples



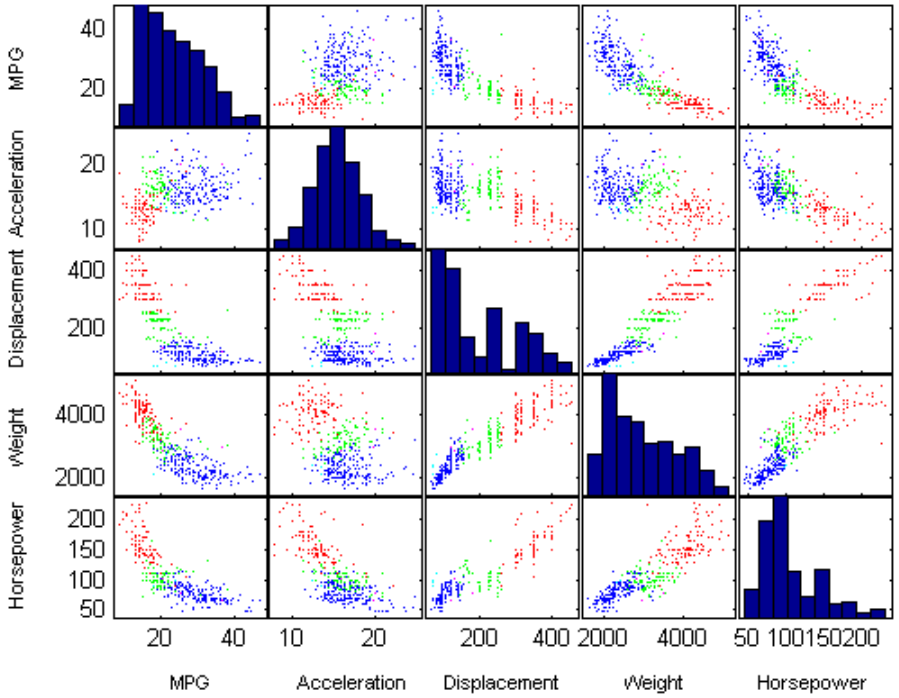
J. Heer, M. Bostock and V. Ogievetsky, A Tour Through the Visualization Zoo.

Horizon graphs



Dimensional Subsetting

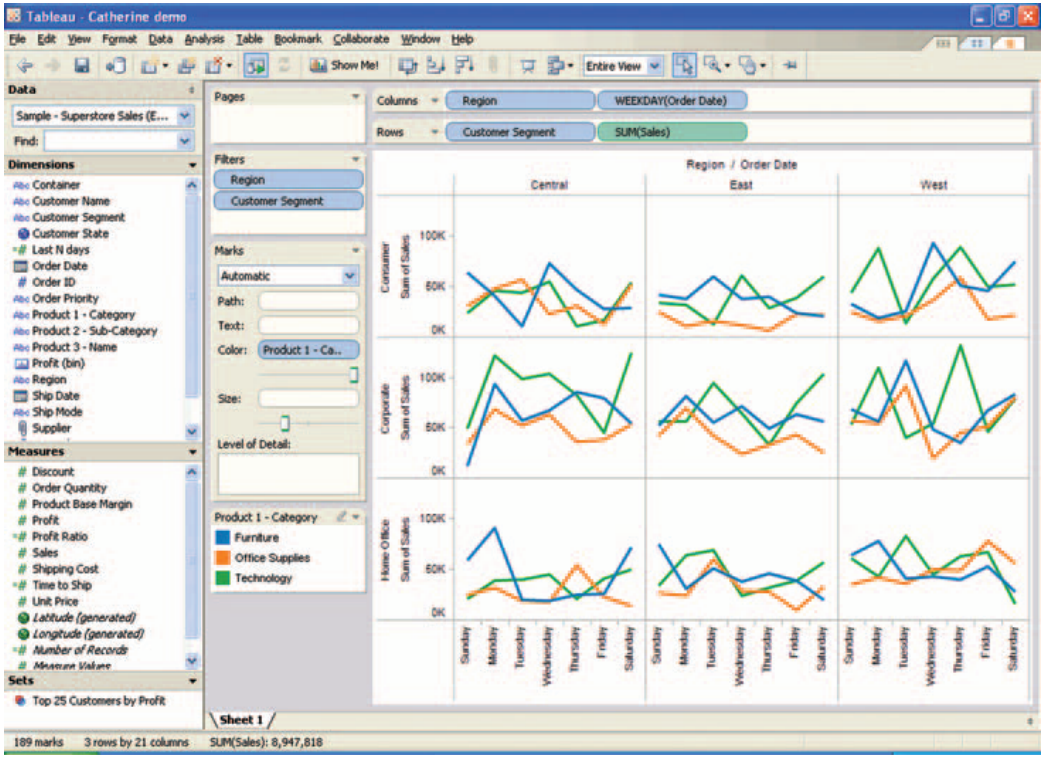
- In a n-dimensional data set subset the data into categories in order to compare the patterns of data between the resulting subsets
- Lay out the resulting small multiples in sequential or 2d ordered grid



Adapted from M. Ward

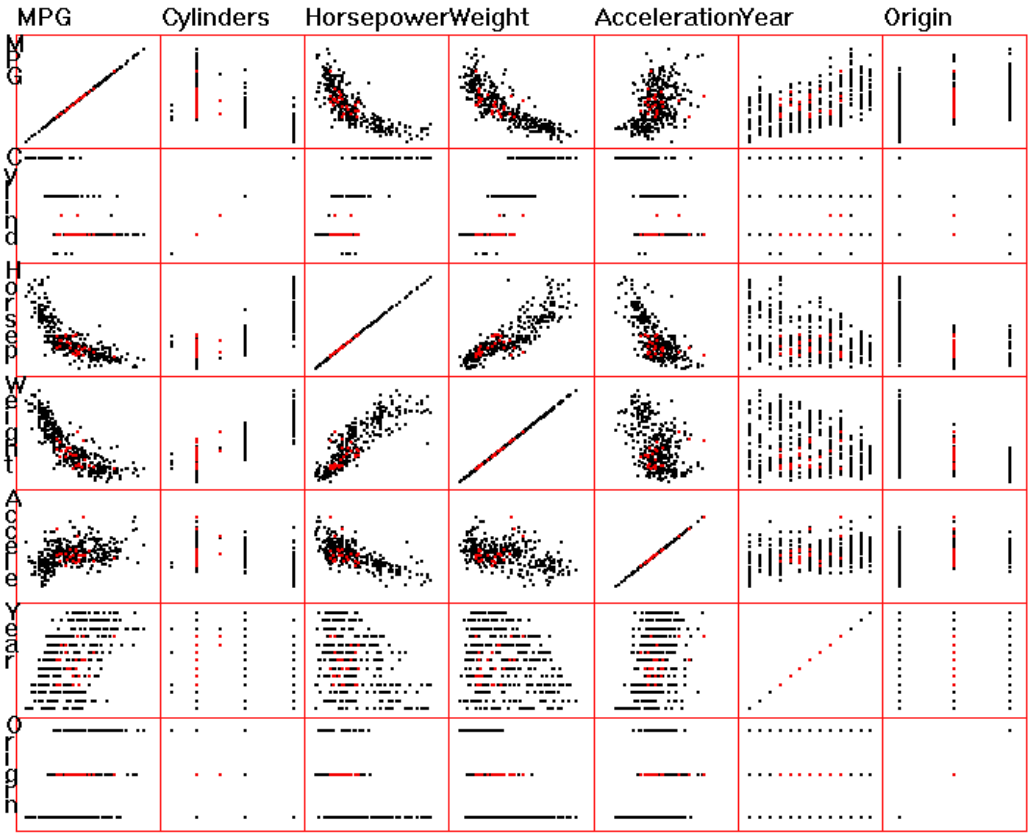
Simple subsetting

- Breaks large number of attributes into smaller groups
- a series or grid of small similar graphics or charts allows easy comparison.
- can be ordered along one or two categorical axes



Scatterplot Matrix (SPLOM)

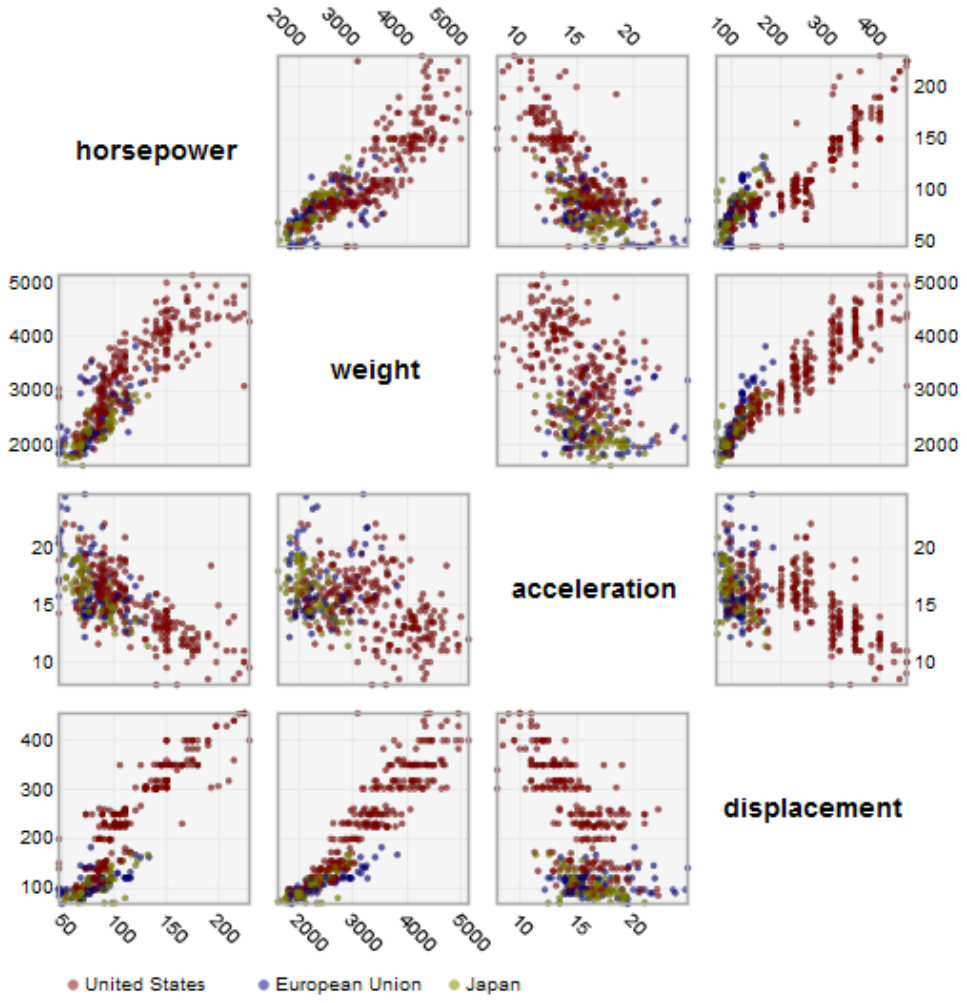
- Scatterplot matrix displays all pairwise plots
- Selection allows linkage between views
- Clusters, trends, and correlations readily discerned between pairs of dimensions
- Important for statistical distributions



Adapted from M. Ward

SPLOM

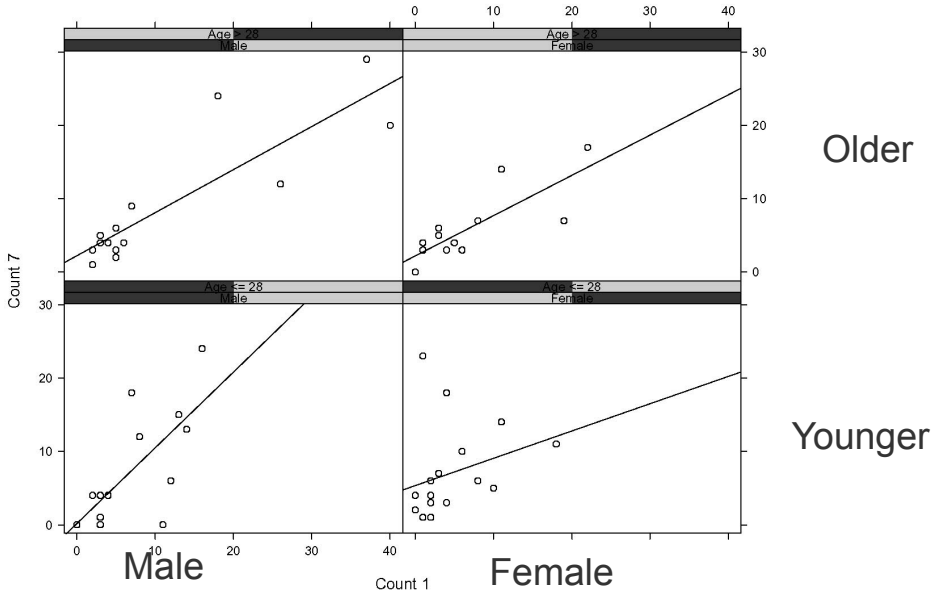
- More elegant layout sacrifices alignment but improves indexing



J. Heer, M. Bostock and V. Ogievetsky,
A Tour Through the Visualization Zoo.

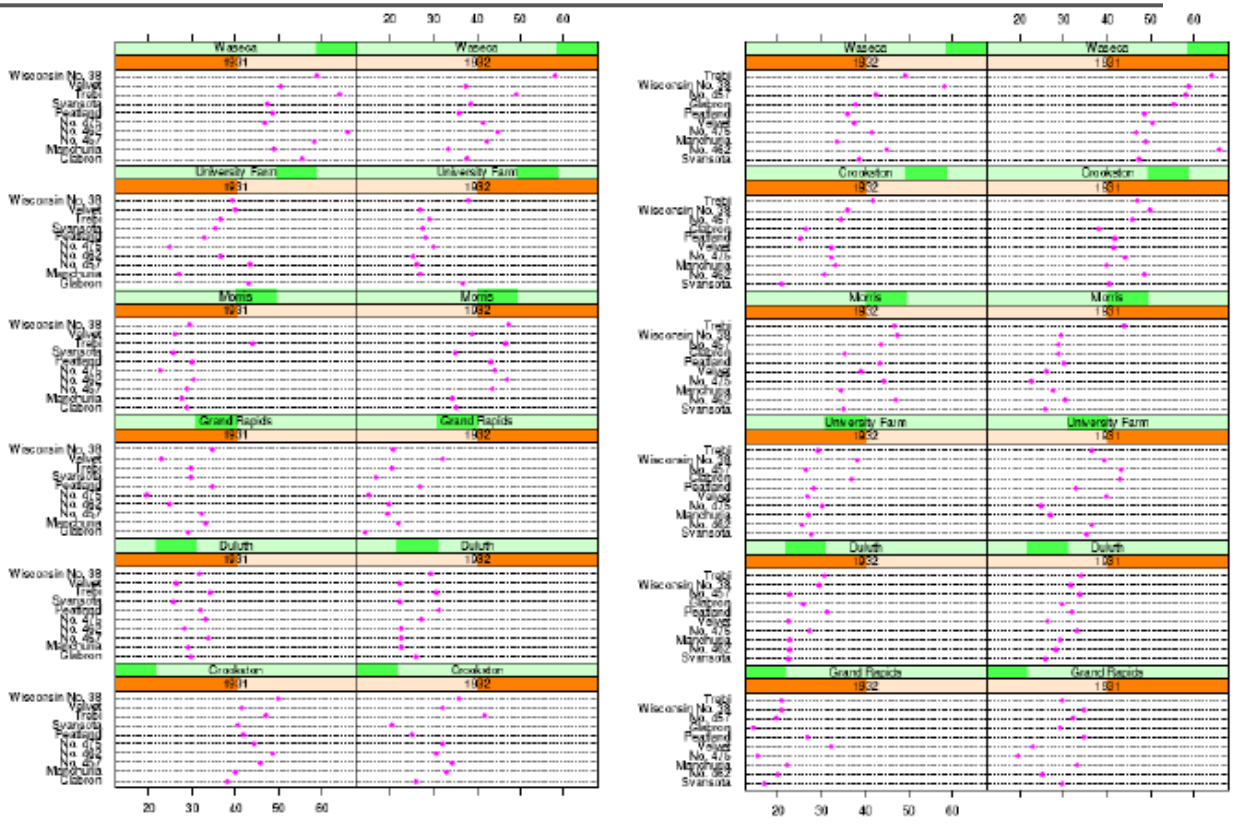
Trellis Plot

- Trellised visualizations enable you to quickly recognize similarities or differences between different categories in the data



Trellis plots

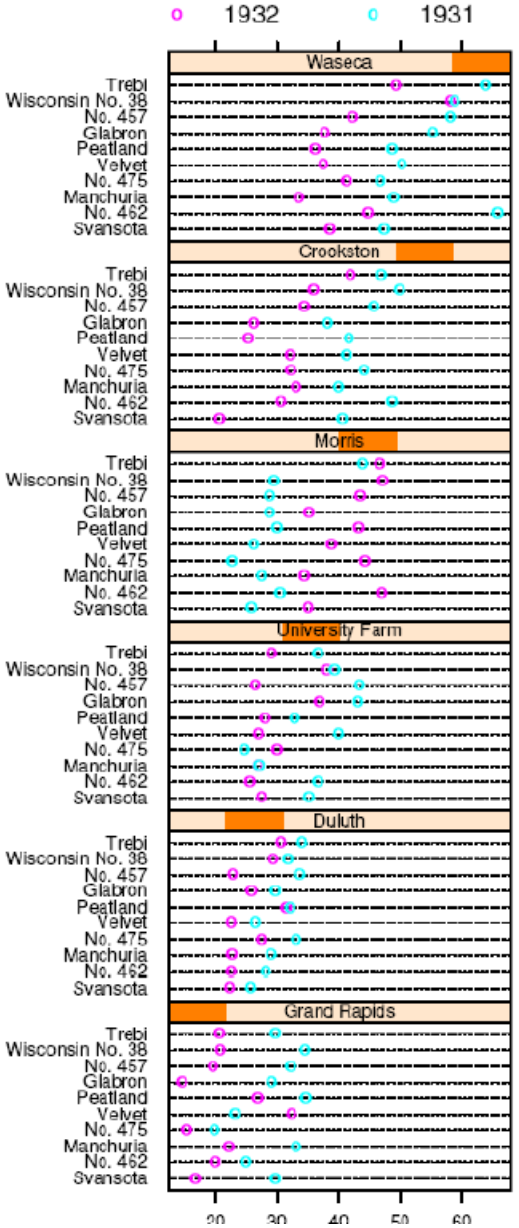
- Can be sorted on direct or derived attributes



[The Visual Design and Control of Trellis Display. Becker, Cleveland, and Shyu. JCSG 5:123-155 1996]

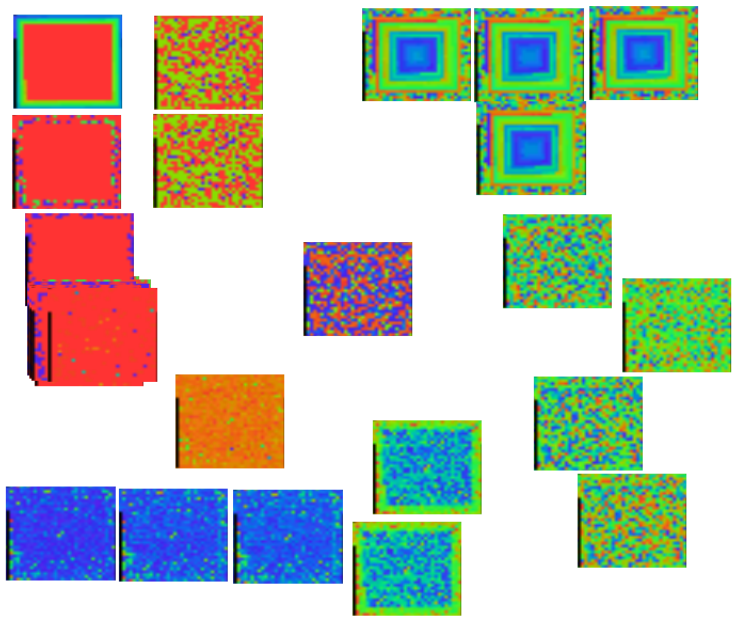
Trellis plots

- conditioning/trellising: choose structure
  - pick how to subdivide into panels
  - pick x/y axes for panels
  - explore space with different choices
    - multiple conditioning
- ordering
  - large-scale: between panels
  - small-scale: within panels
  - main effects: sort by group median
  - derived space, from categorical to ordered



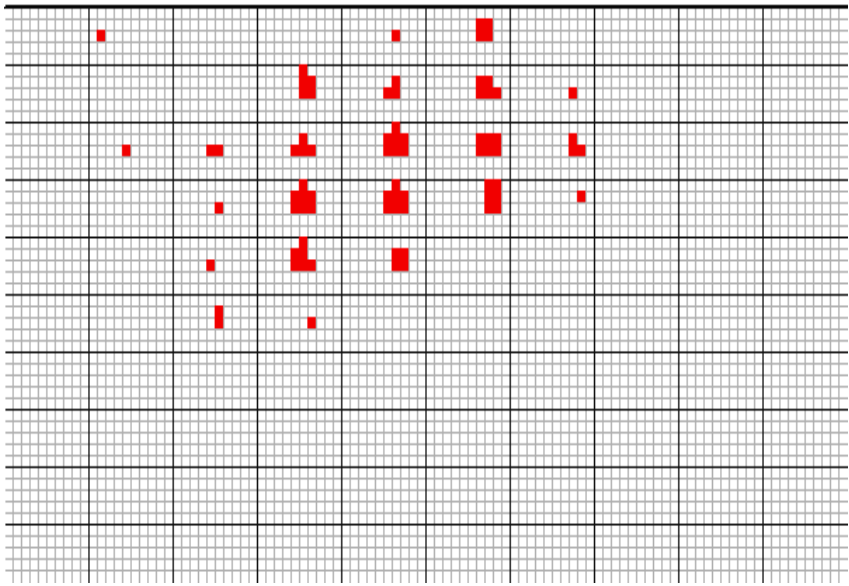
Dimensional Subsetting (2)

- Pixel-oriented techniques lay out a series of univariate displays
- Values are conveyed via color
- Records are ordered temporally, by value, or by a user query



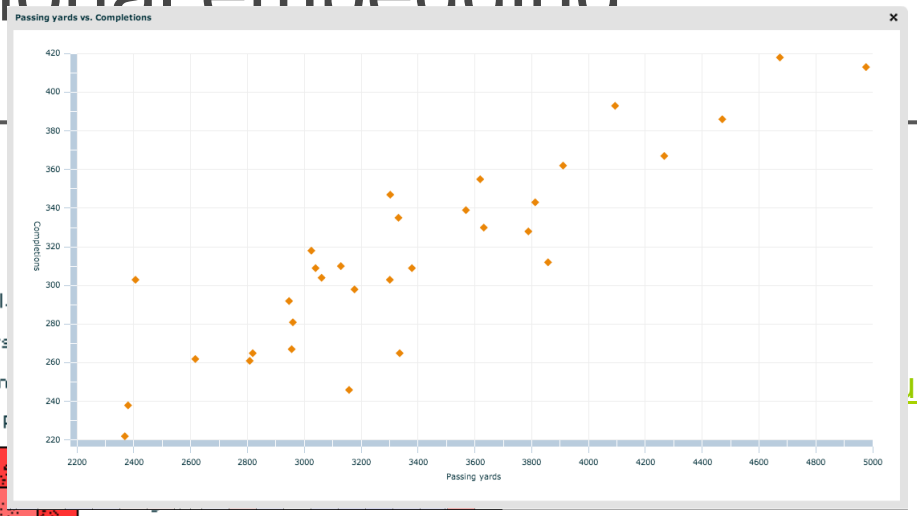
Dimensional Embedding

- Dimensional stacking divides data space into bins
- Each N-D bin has a unique 2-D screen bin
- Screen space recursively divided based on bin count for each dimension
- Clusters and trends manifested as repeated patterns



SFU

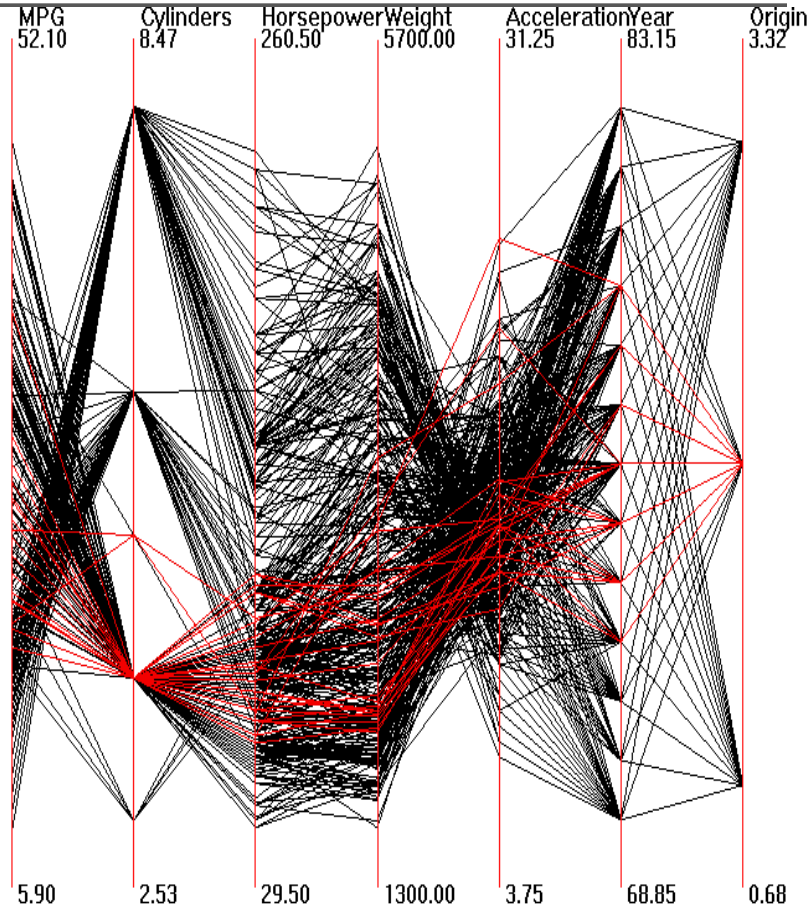
Dimensional embedding



ualizing NFL 2008

Dimensional Reorganization

- Parallel Coordinates creates parallel, rather than orthogonal, dimensions.
- Data point corresponds to polyline across axes
- Clusters, trends, and anomalies discernable as groupings or outliers, based on intercepts and slopes



SFU

Data Table Idiom

	Dimensions	Case1	Case2	Case3
Variable1		Value11	Value21	Value31
Variable2		Value12	Value22	Value32
Variable3		Value13	Value23	Value33

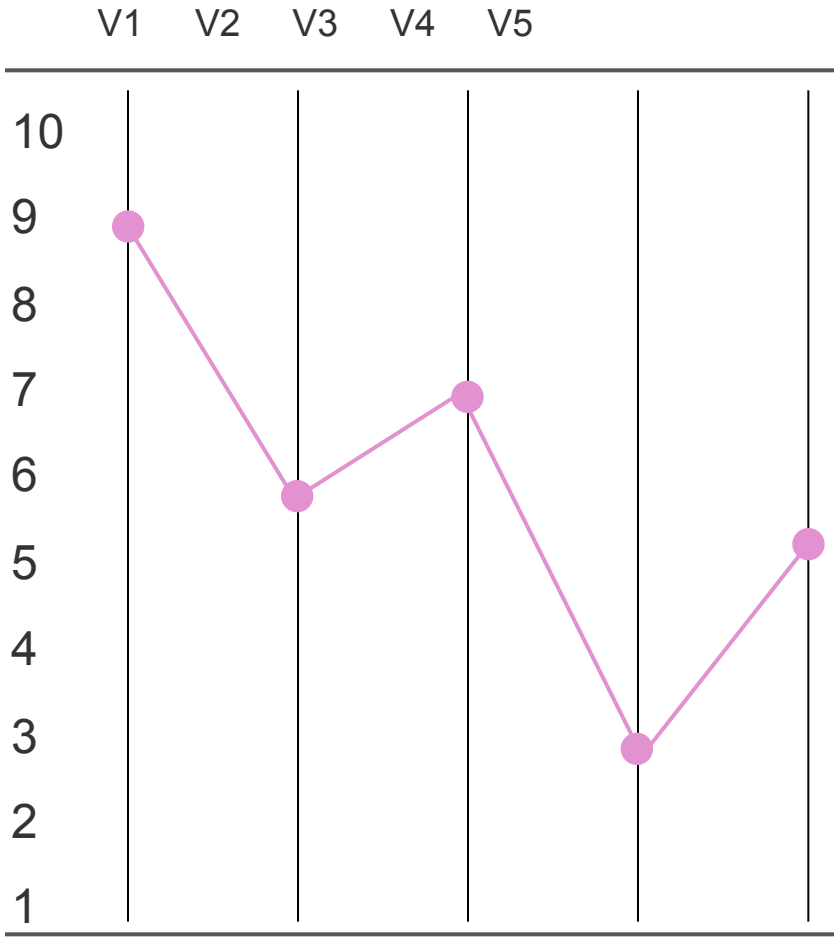
- Think of as a function
  - f(case1) = <Val11, Val12,…>

Spreadsheets

- Tables allocate a unique space per value
  - Case + Variable
  - 1 case per row
  - 1 variable per column

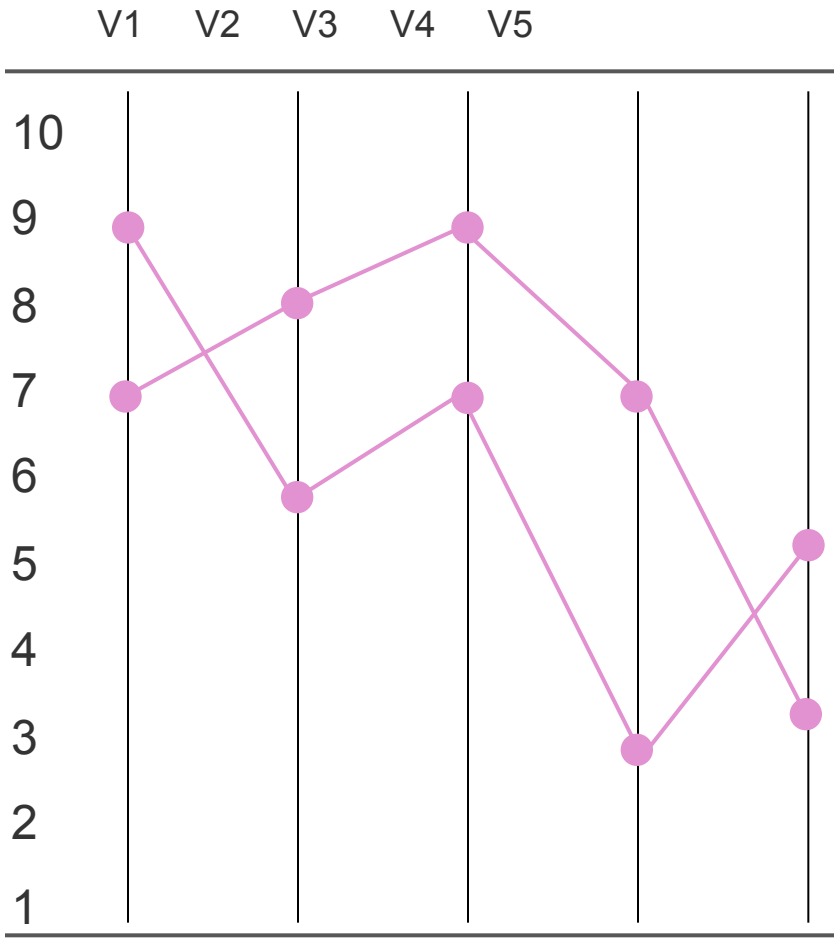
Case	V1	V2	V3	V4	V5
1	9	6	7	3	5
2	7	8	9	7	3

Parallel Coordinates (Inselberg 1997)



Case	V1	V2	V3	V4	V5
1	9	6	7	3	5
2	7	8	9	7	3

Parallel Coordinates



Case	V1	V2	V3	V4	V5
1	9	6	7	3	5
2	7	8	9	7	3

Parallel Coordinates

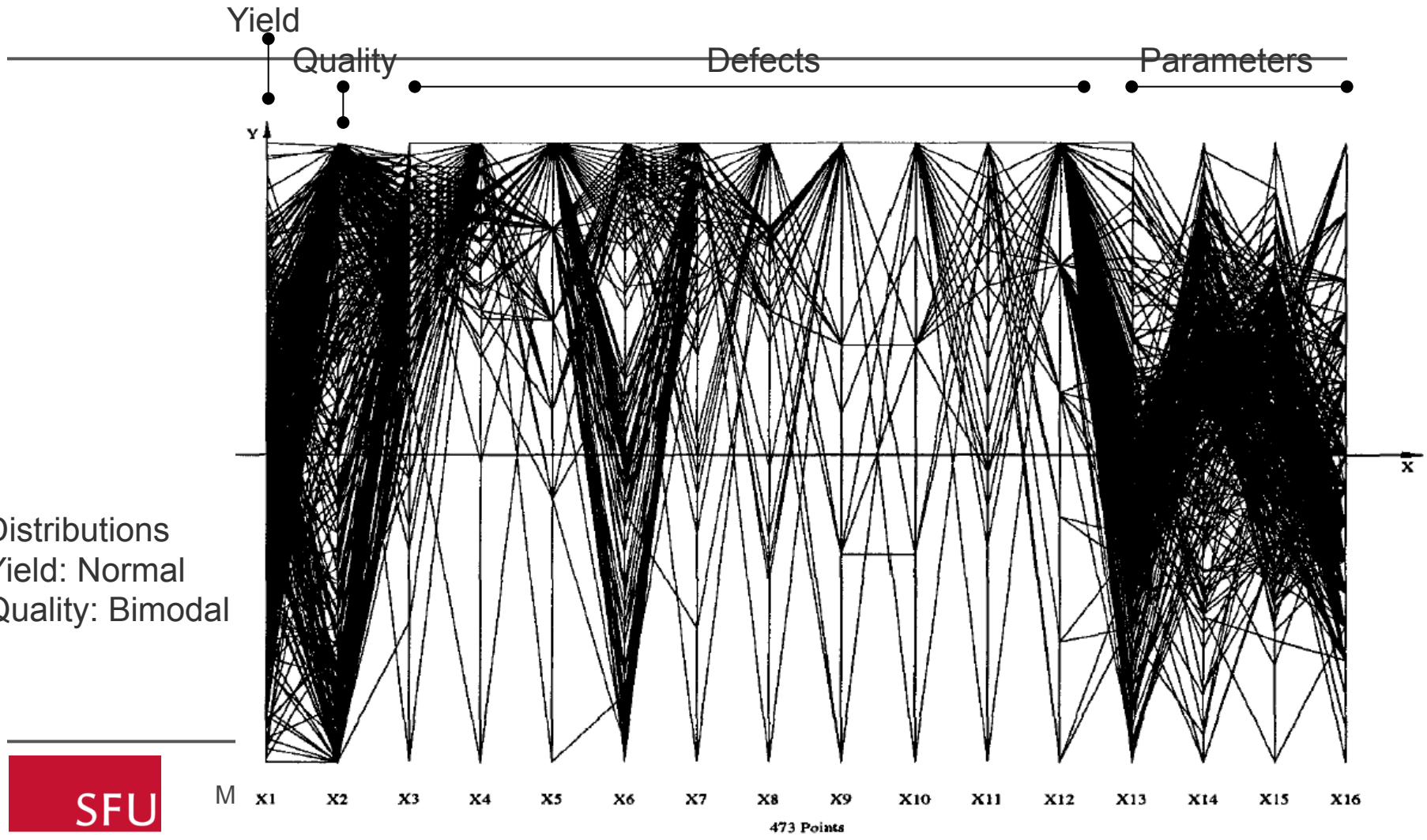
- Each column of space is assigned a variable
- Vertical Scale to left
- Each data case is a polyline that puts a vertex on each column at its corresponding data value

Example Problem

- VLSI chip manufacture
- Want high quality chips (high speed) and a high yield batch (% of useful chips)
- Able to track defects
- Hypothesis: No defects gives desired chip types
- 473 batches of data
  - A. Inselberg, “Multidimensional Detective” InfoVis 1997.

The Data

- 16 variables
  - X1 - yield
  - X2 - quality
  - X3-X12 - # defects (inverted)
  - X13-X16 - physical parameters



Distributions
Yield: Normal
Quality: Bimodal

SFU

Top Yield & Quality

- Split in parameters

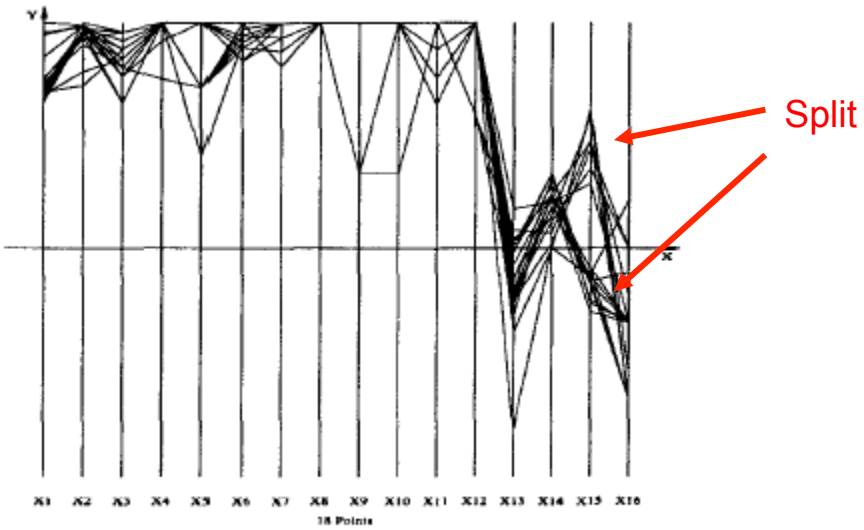


Figure 2: The batches high in Yield, $X1$, and Quality, $X2$.

Minimal Defects

- Not best yield & quality

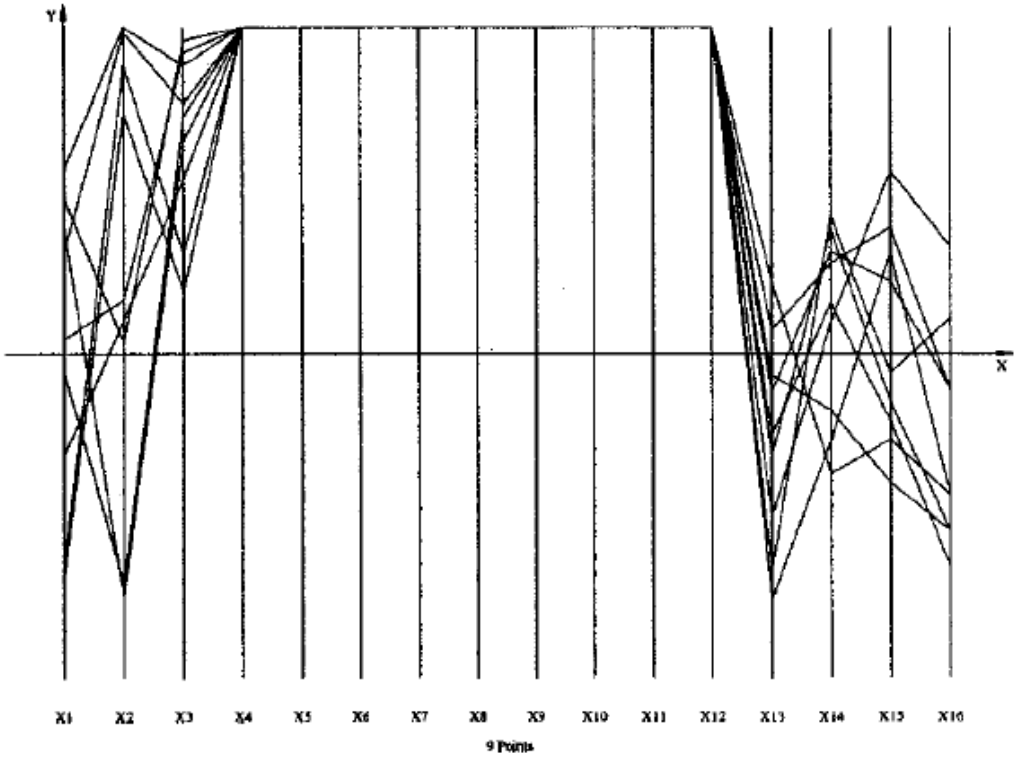


Figure 3: The batches with zero in 9 out of the ten defect types.

Best Yields

Parameters that give best yields cause 2 types of defects

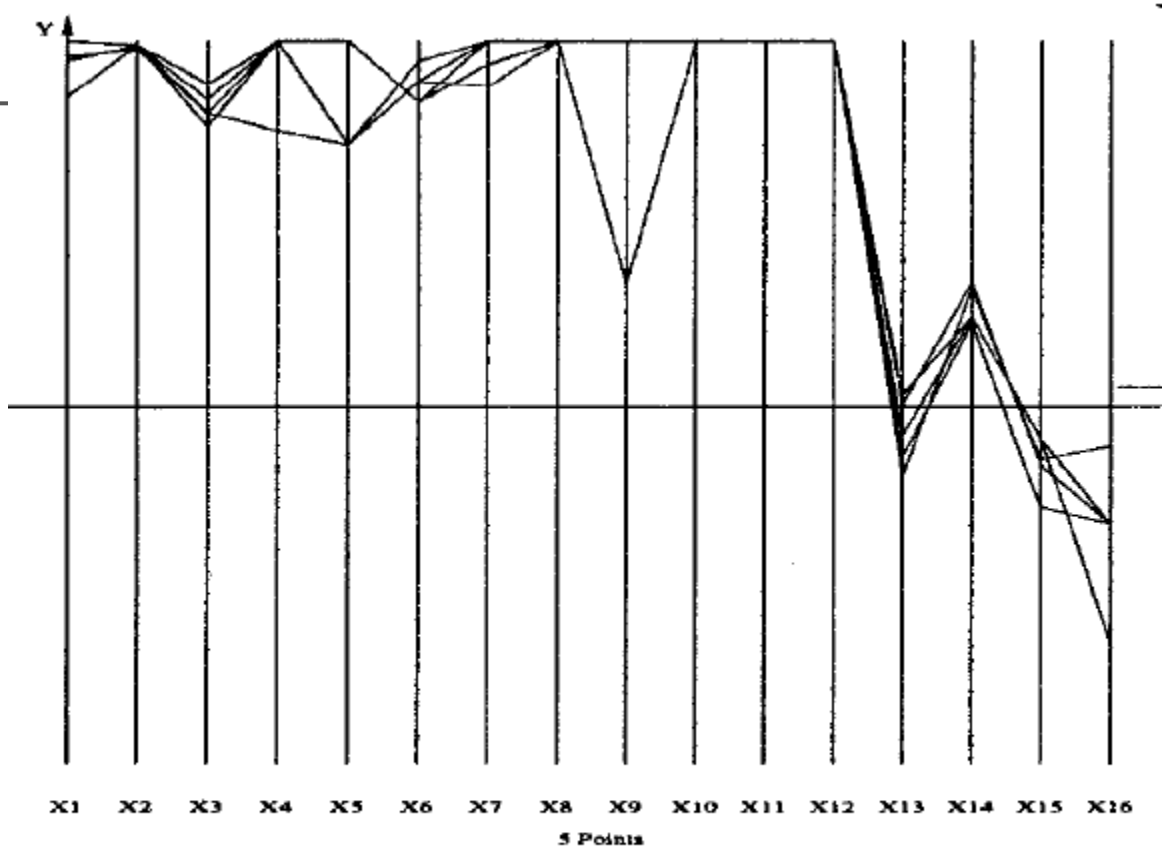


Figure 6: Batches with the highest Yields do not have the lowest defects in $X3$ and $X6$.

Go back to first diagram, looking at defect categories.

Notice that X6 behaves differently than the rest.

Allow two defects, where one defect in X6.

This results in the very best batch appearing.

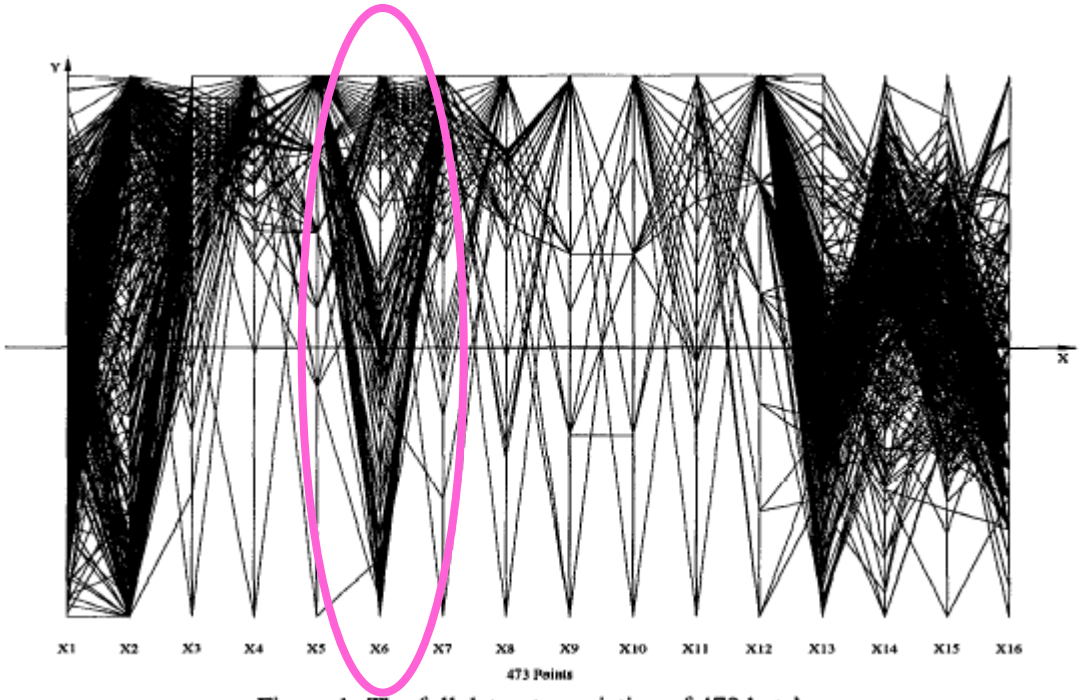


Figure 1: The full dataset consisting of 473 batches

SFU

Multidimensional Detective

- Fig 5 and 6 show that high yield batches don't have non-zero values for defects of type X3 and X
  - Don't believe your assumptions …
- Looking now at X15 we see the separation is important
  - Lower values of this property end up in the better yield batches

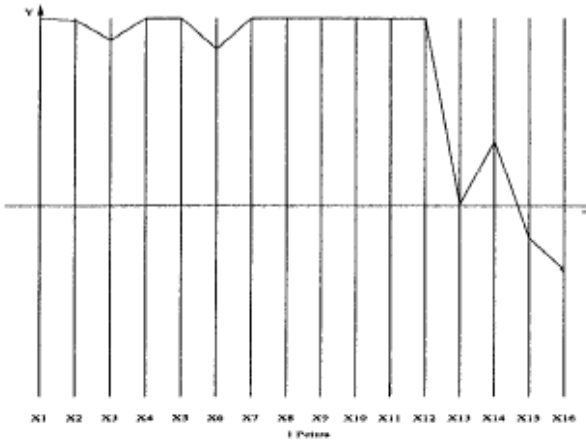


Figure 5: The best batch. Highest in Yield, $X1$, and very high in Quality, $X2$.

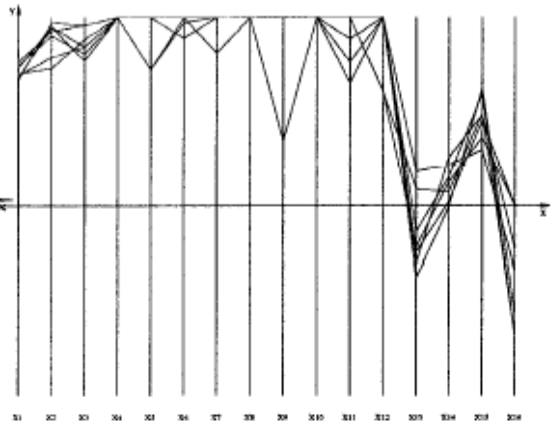
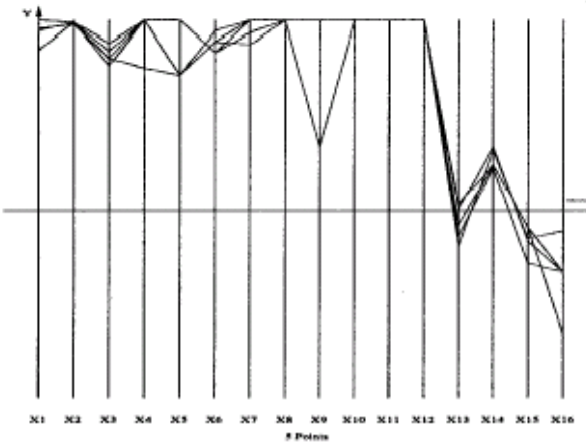
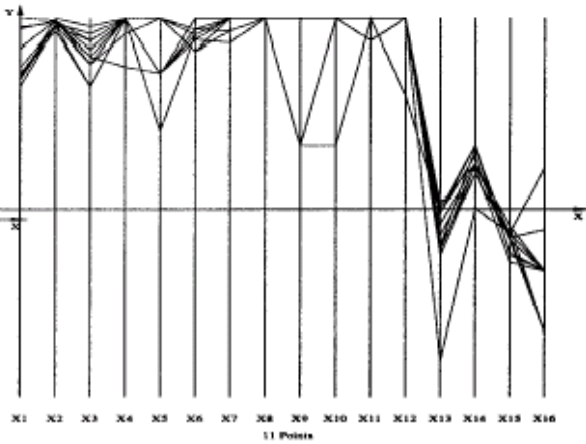


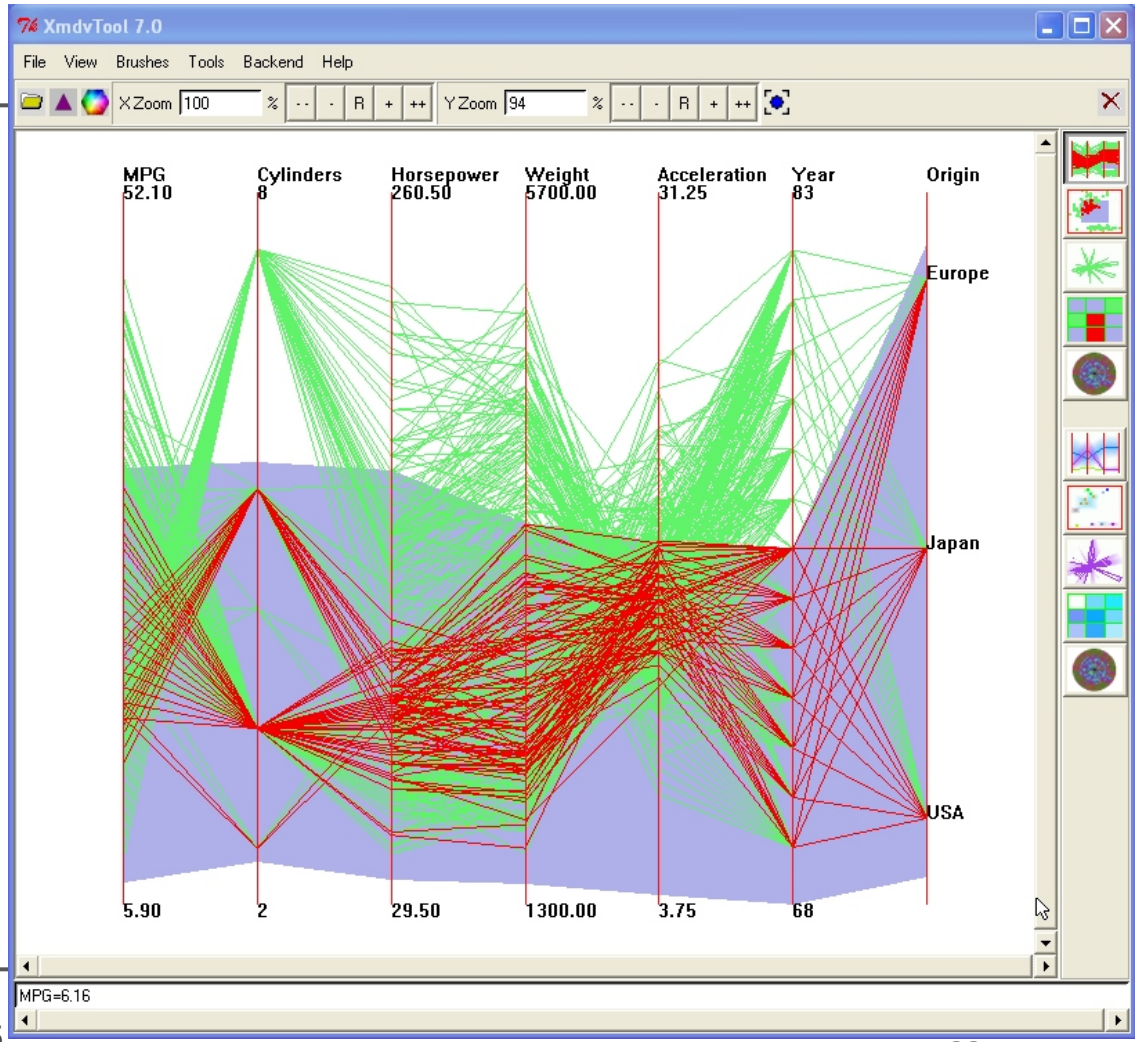
Figure 7: Upper range of split in $X15$





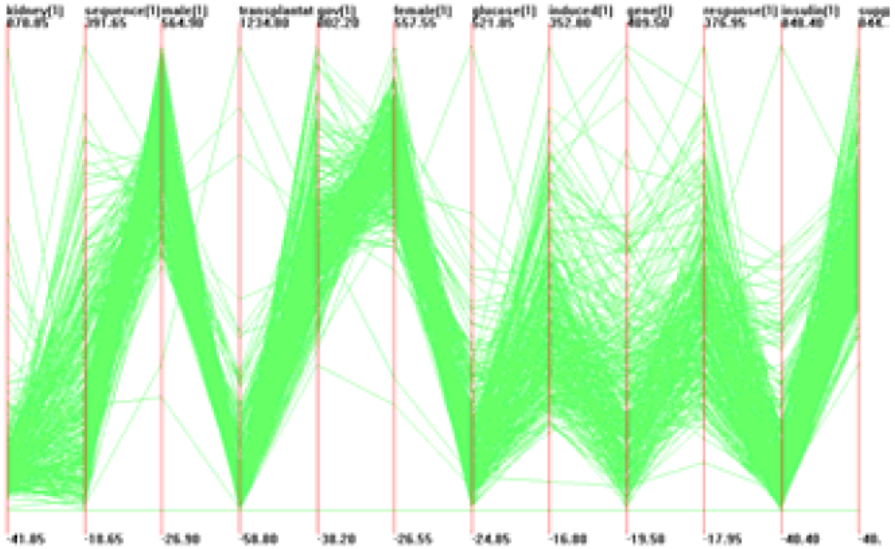
XmdvTool

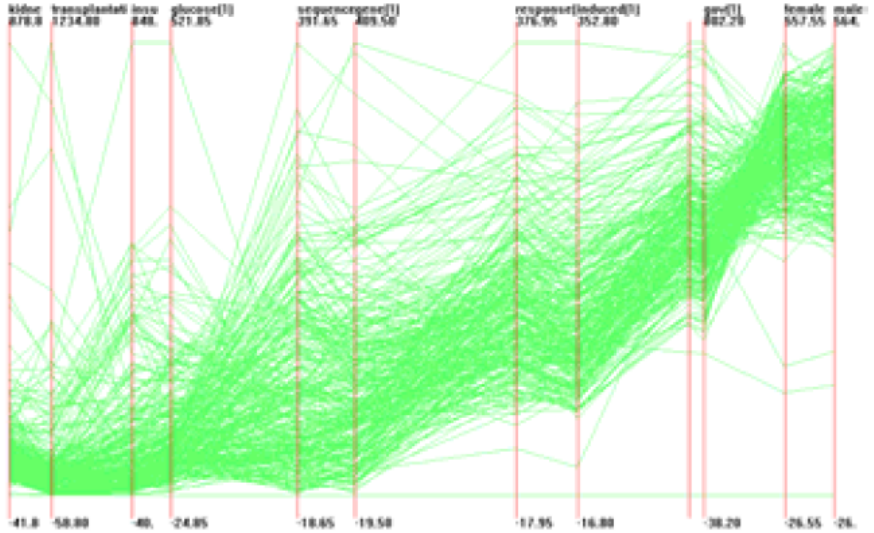
- Matt Ward, WPI
- Does Parallel Coords



Dimensional Reordering

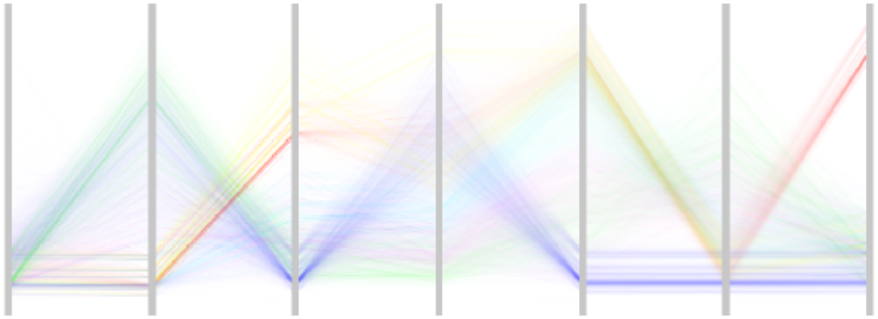
- Which Dimensions are most alike?
- Sort dimensions according to similarity



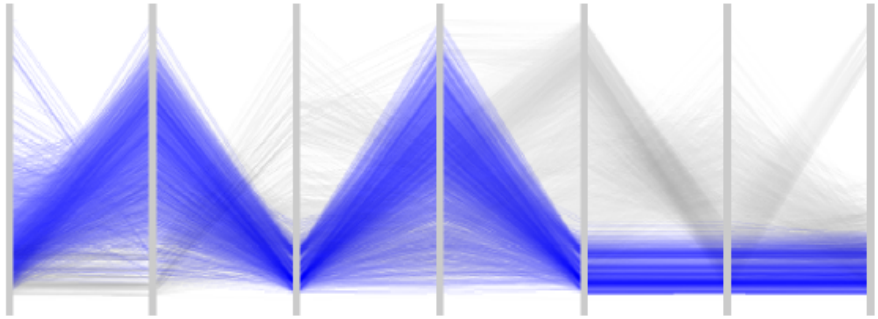


Advanced Graphics

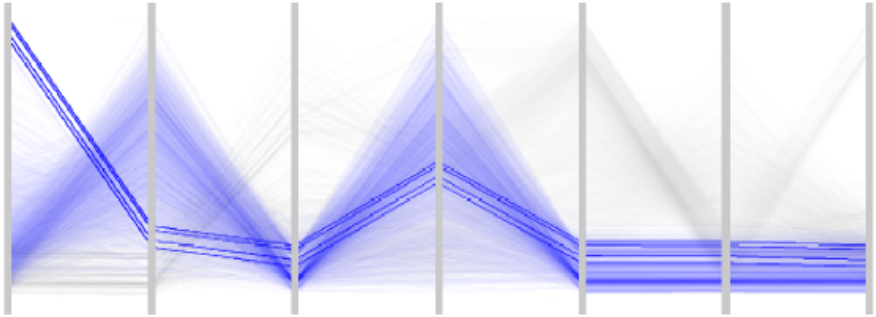
- Johanson et al InfoVis 2005



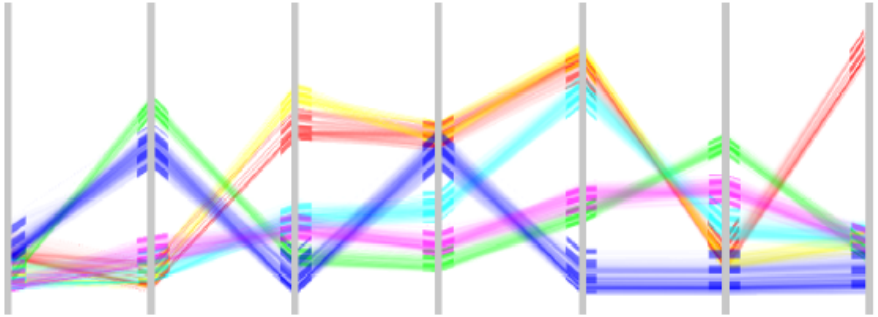
(a) A linear transfer function has been applied to the high-precision texture in order to prevent cluttering and to provide overview of the data.



(b) A logarithmic transfer function is applied to a selected cluster. The structure is preserved and emphasis is put on the low density regions.



(c) Local cluster outliers are enhanced. A square root transfer function is used and the outliers are visible even through high-density regions.

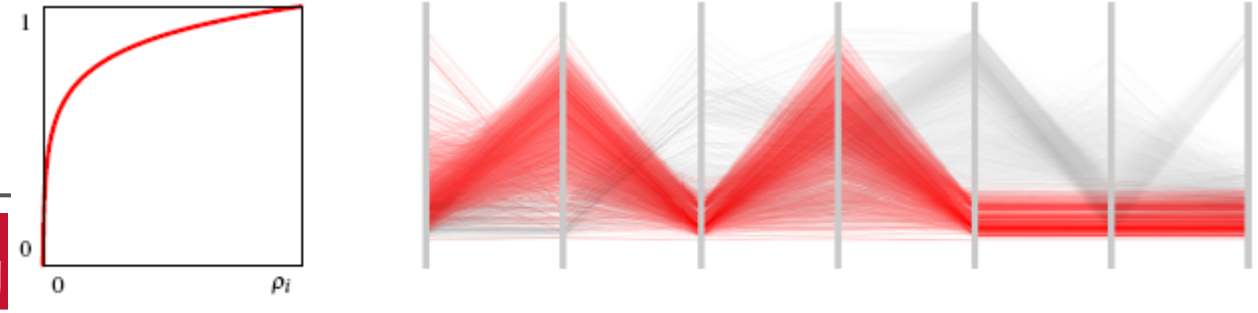


(d) A complementary view of the clusters with uniform bands. 'Feature animation' presents statistics about the clusters and acts as a guidance.

Use texture mapping

- Pre-process data into categories
- Render each category to texture
- Blend textures

(b) A square TF is used and only dense regions become visible. This facilitates analysis of many clusters simultaneously.



(d) Using a logarithmic TF puts even more emphasize on the lower density regions.

Further elaboration

- Brush individual ranges
- Display histogram per dimension

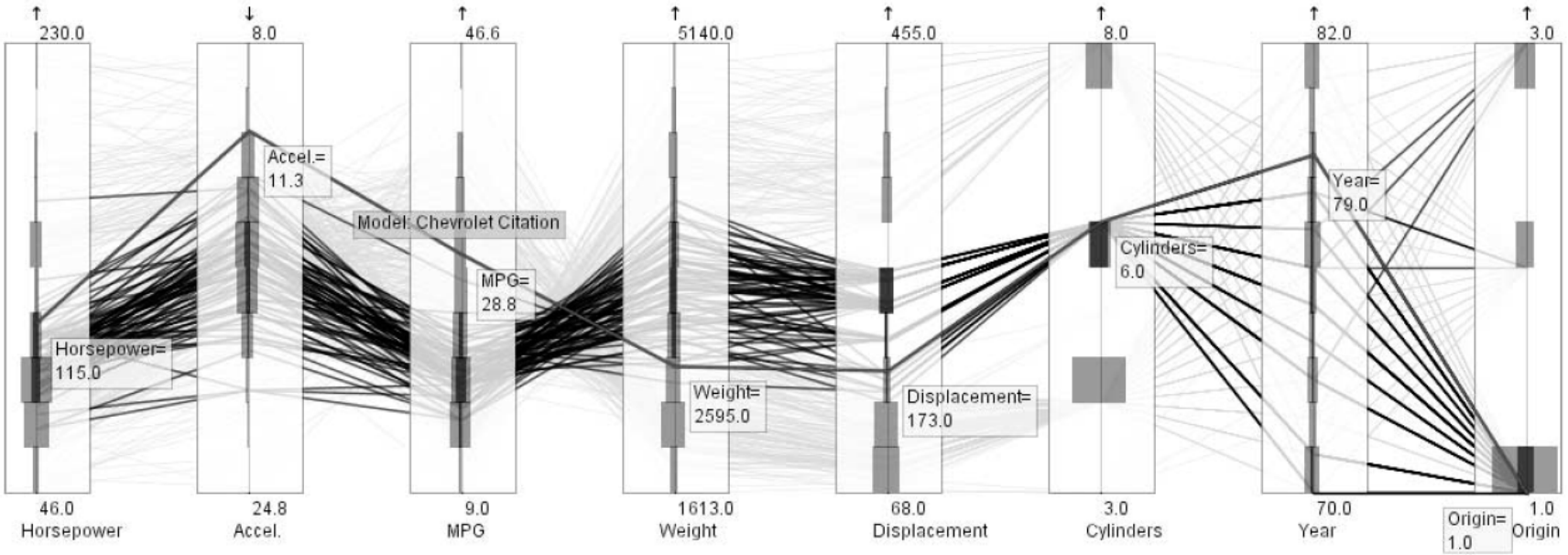


Figure 2. Extended parallel coordinates, a sample view of the *cars* data-set: cars with six cylinders were emphasized through brushing, histograms are laid over axes, and one data-point is shown with all details.

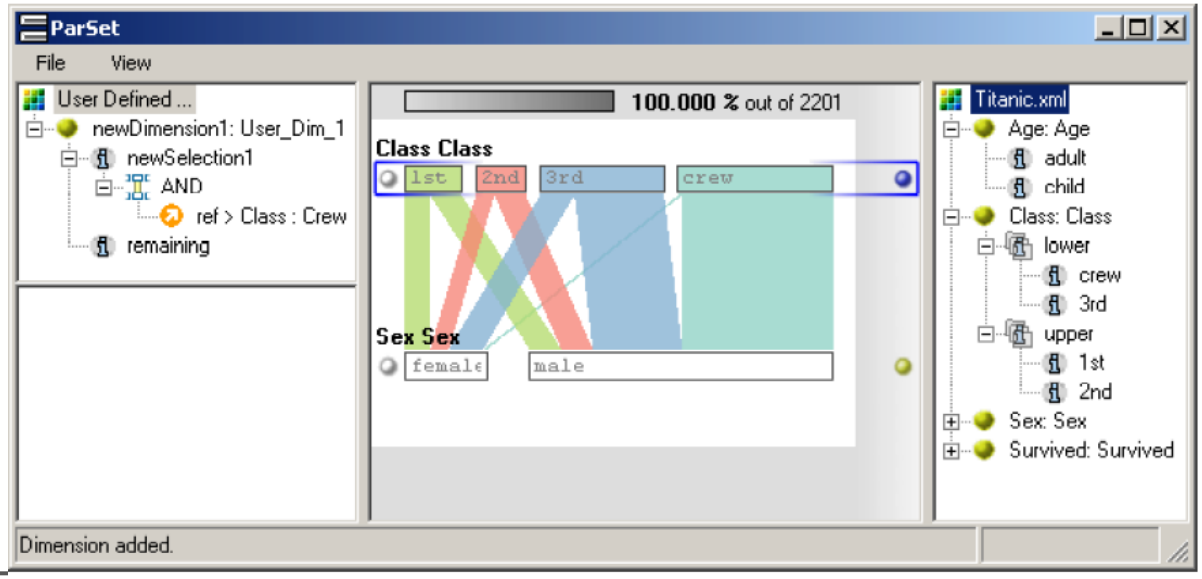
Visualizing Categories

- Titanic Disaster
- Bendix et al InfoVis 2005

Class	Sex		
	female	male	
first	145 44.6% 30.8% 6.6%	180 55.4% 10.4% 8.2%	325 14.8%
second	106 37.2% 22.6% 4.8%	179 62.8% 10.4% 8.1%	285 12.9%
third	196 27.8% 41.7% 8.9%	510 72.2% 29.5% 23.2%	706 32.1%
crew	23 2.6% 4.9% 1.1%	862 97.4% 49.8% 39.1%	885 40.2%
	470 21.4%	1731 78.6%	2201 100%

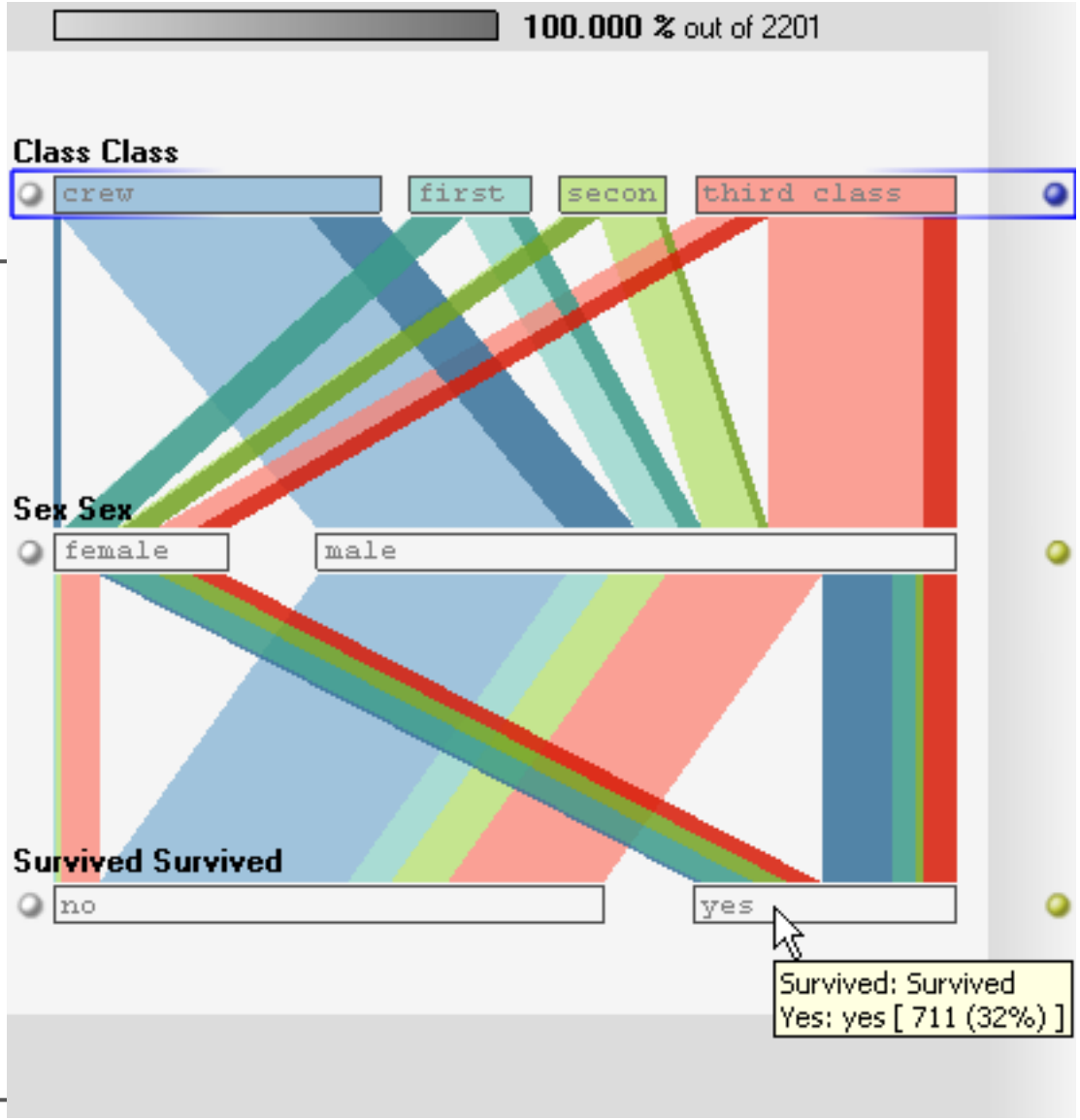
Parallel sets

- Parallel coordinates layout
  - Continuous axes replaced with boxes
- Uses frequency based representation



Parallel sets

- Continuous axes replaced with boxes
- Makes use of layering and transparency
- Good for categorical visualization?



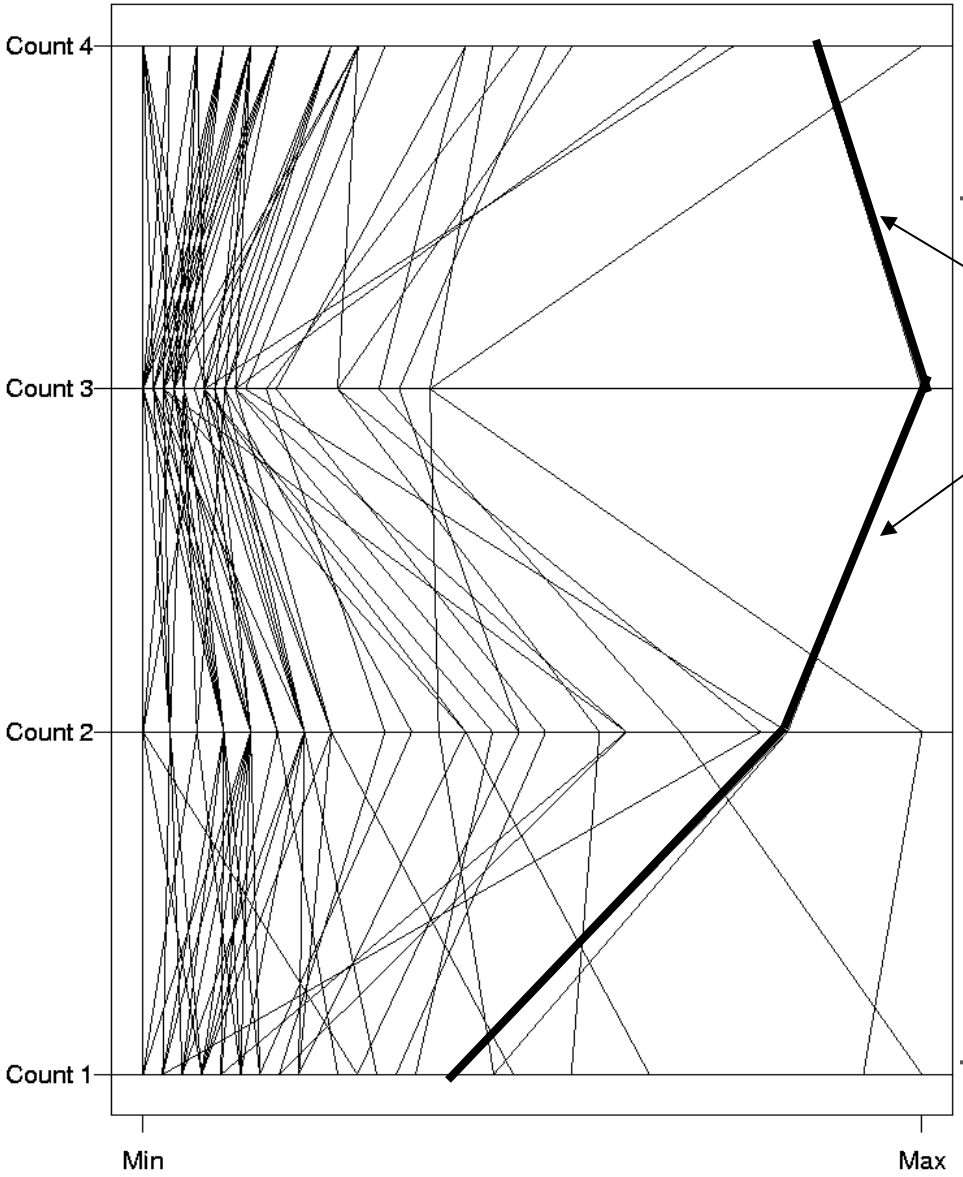
Parallel Coordinate

(epileptic seizure data from text)

dimensions
(possibly all p of them!)

often (re)ordered
to better distinguish
among interesting
subsets of n total cases

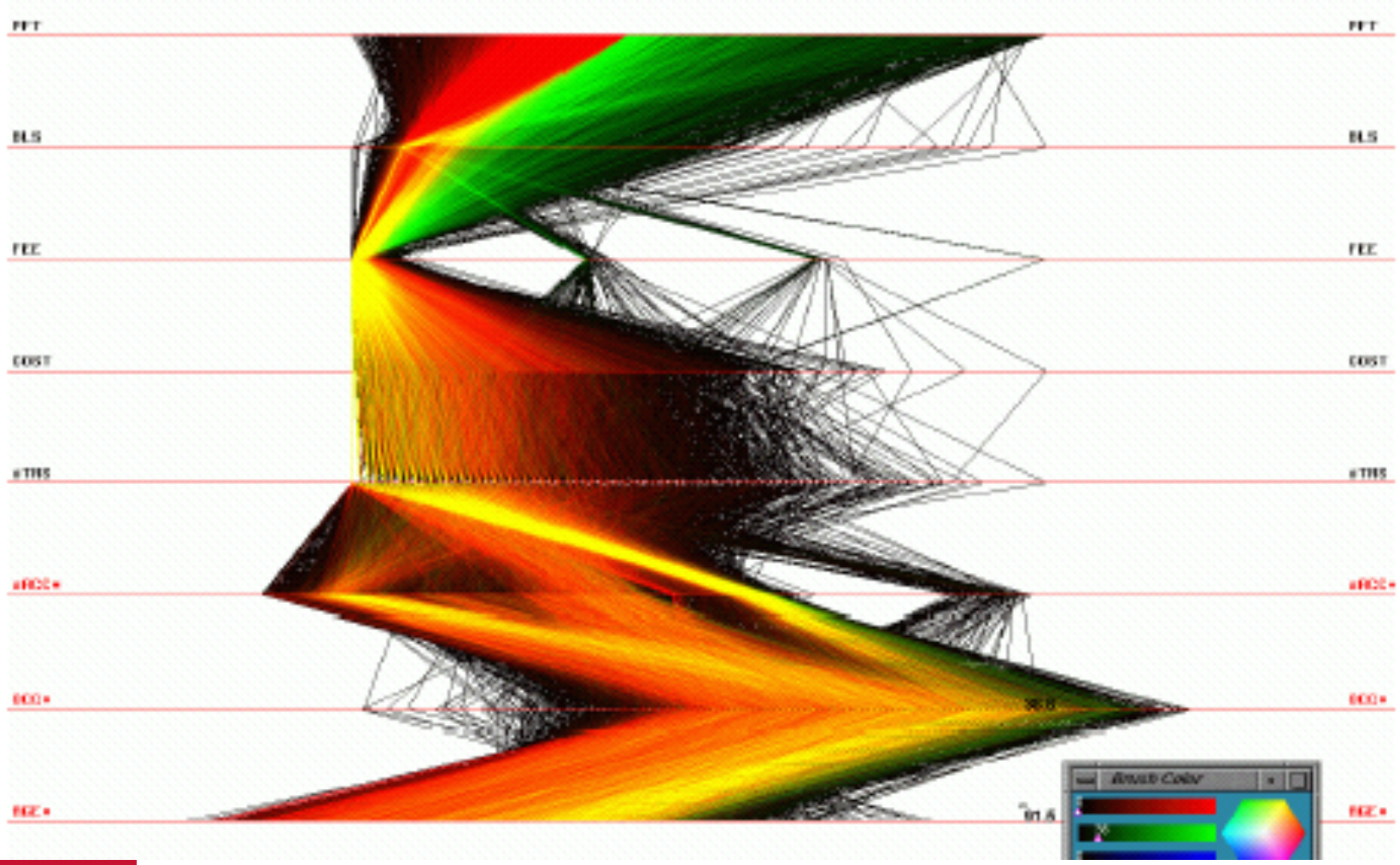
interactive
"brushing" is useful
for seeing such
distinctions



1 (of n)
cases

(this case is
a "brushed"
one, with a
darker line,
to standout
from the n-1
other cases)

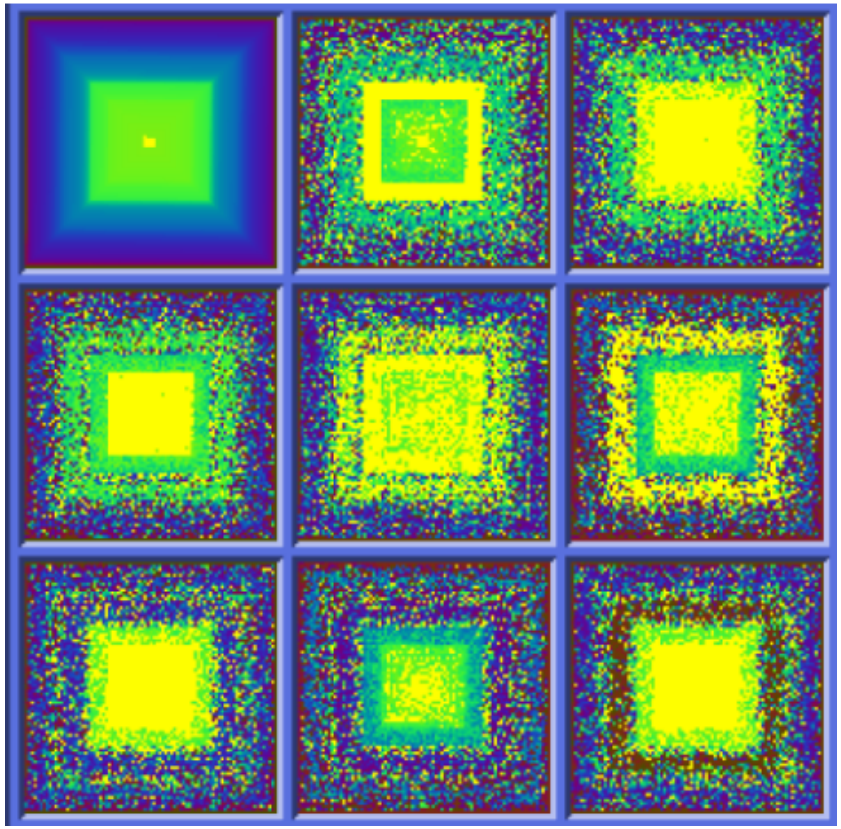
More elaborate parallel coordinates example (from E. Wegman, 1999).
12,000 bank customers with 8 variables
Additional “dependent” variable is profit (green for positive, red for negative)

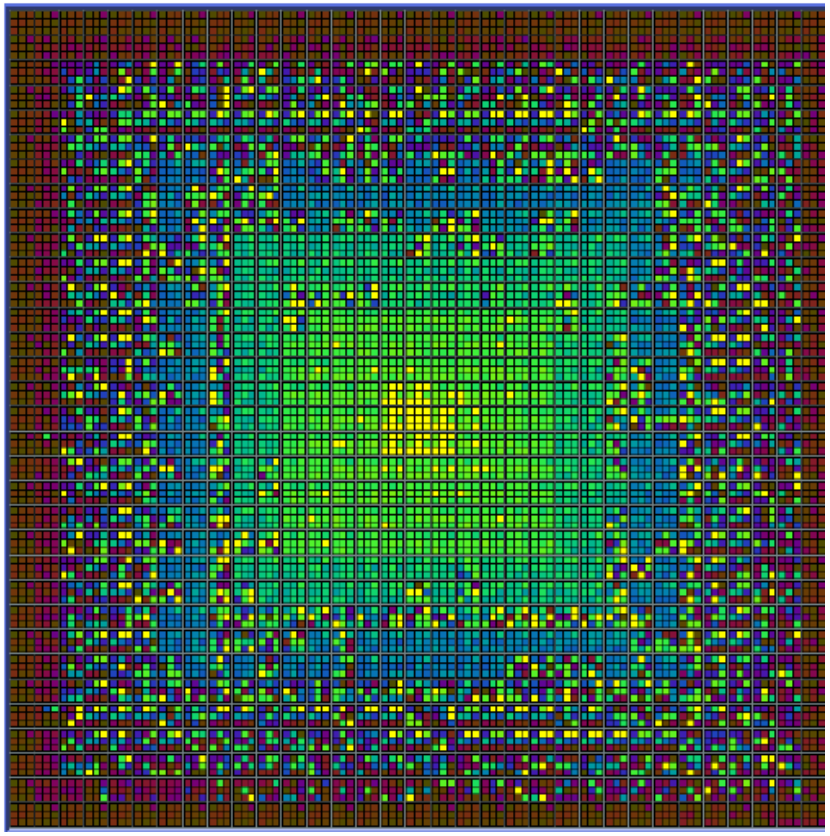


Issues

- The range of each variable can be different:
  - Must **rescale** to the vertical space available
  - Each variable rescales independently
- Hard to read parallel coord plot as a static picture
  - **Interaction required**

What else is going on here?

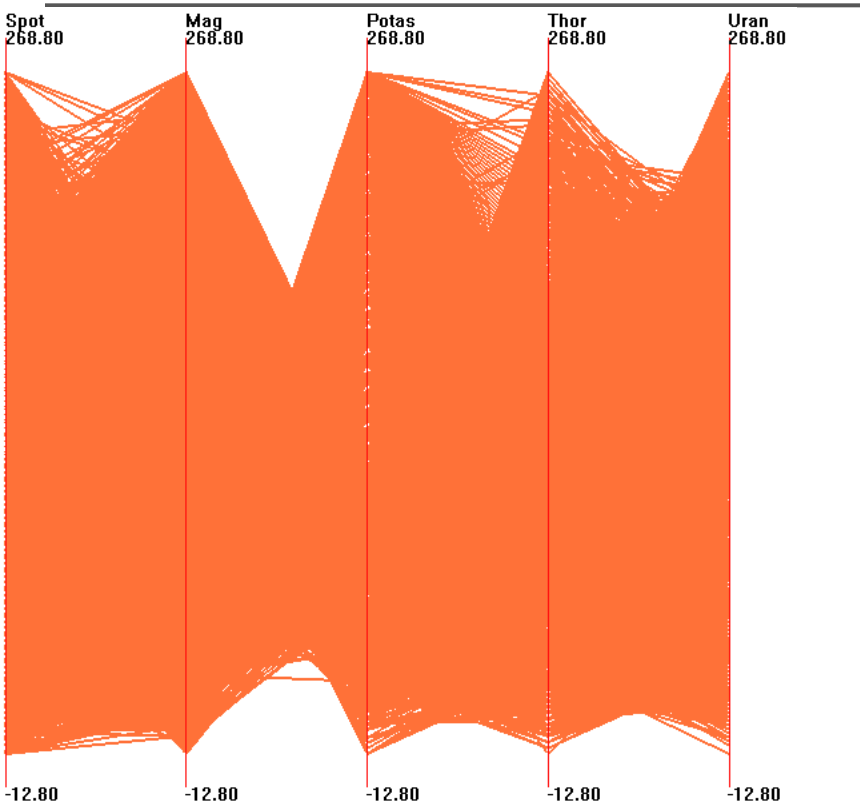


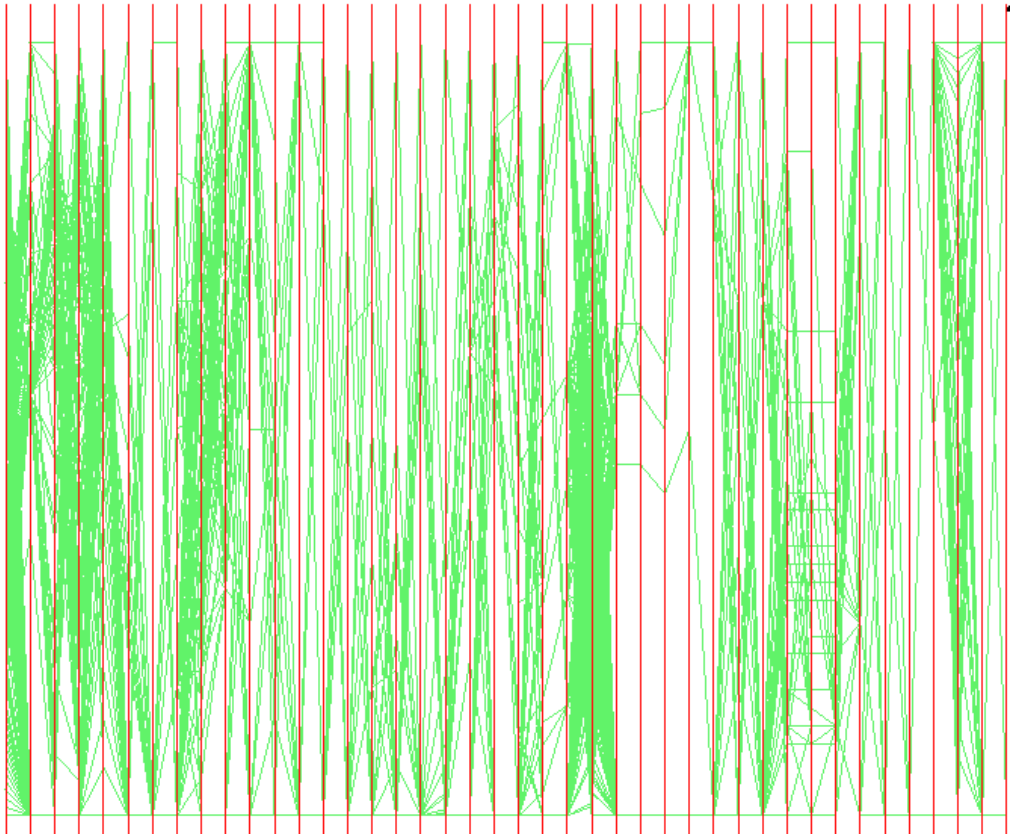


Problems with Large Data Sets

- Most techniques are effective with small to moderate sized data sets
- Large sets (> 50K records) are increasingly common
- When traditional visualizations used, occlusion and clutter make interpretation difficult

Examples of Scale Problem

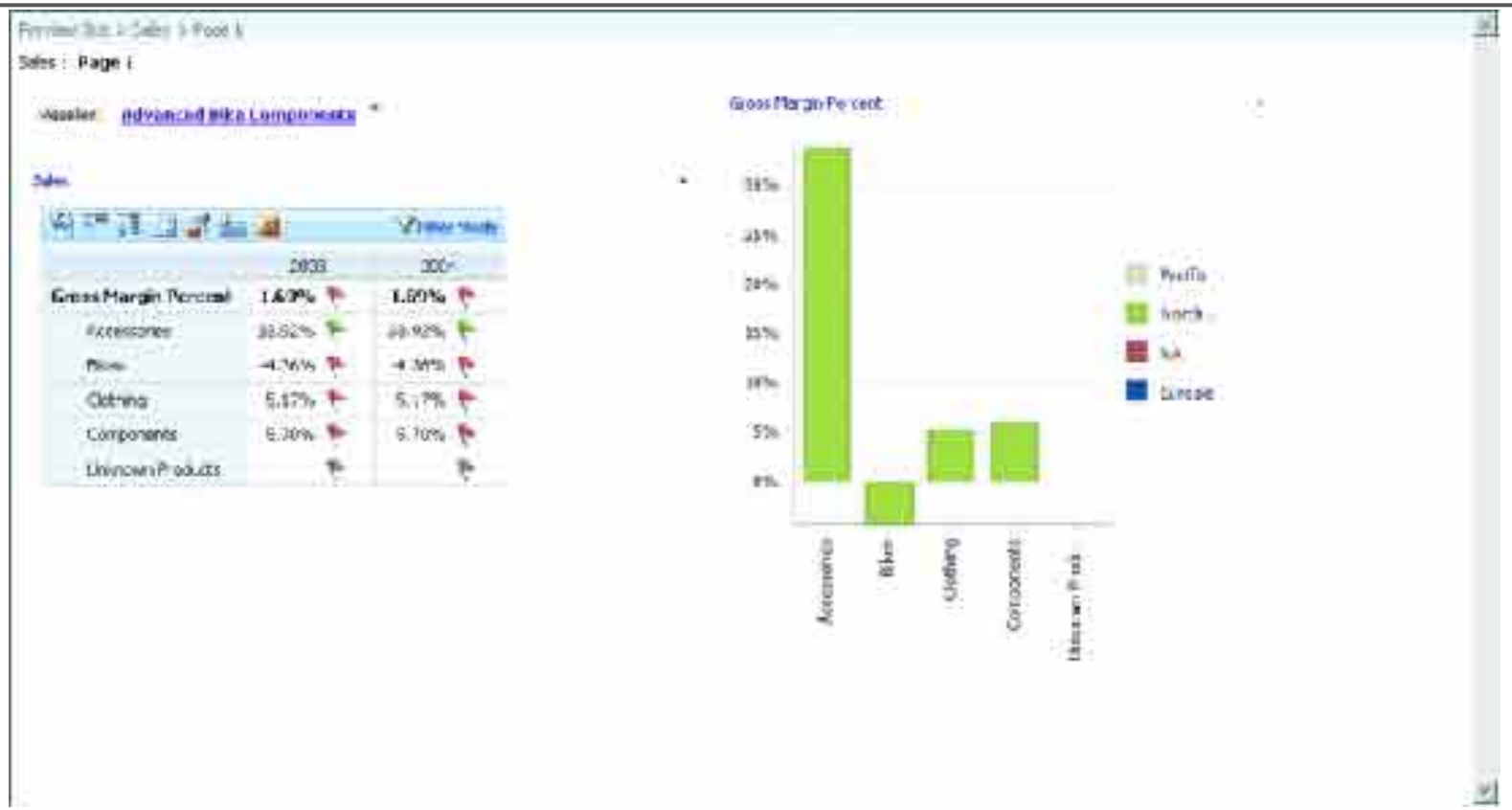




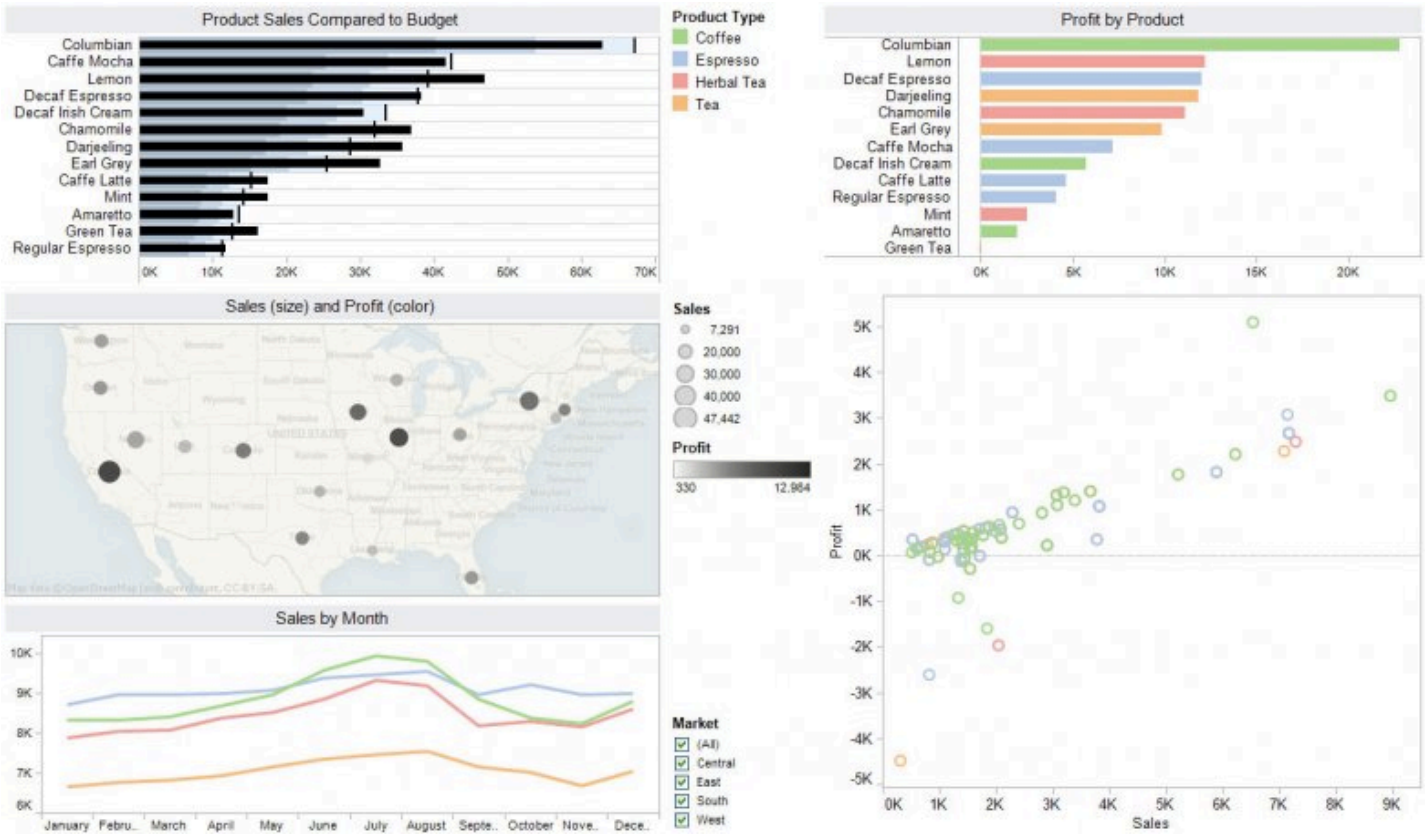
Multiple view methods

- view choices
  - encoding: same or multiform
  - dataset: same or small multiple
  - data: all or subset (overview/detail)
  - spatial ordering of views
- Animation?
  - Data change over time
- many combinations possible

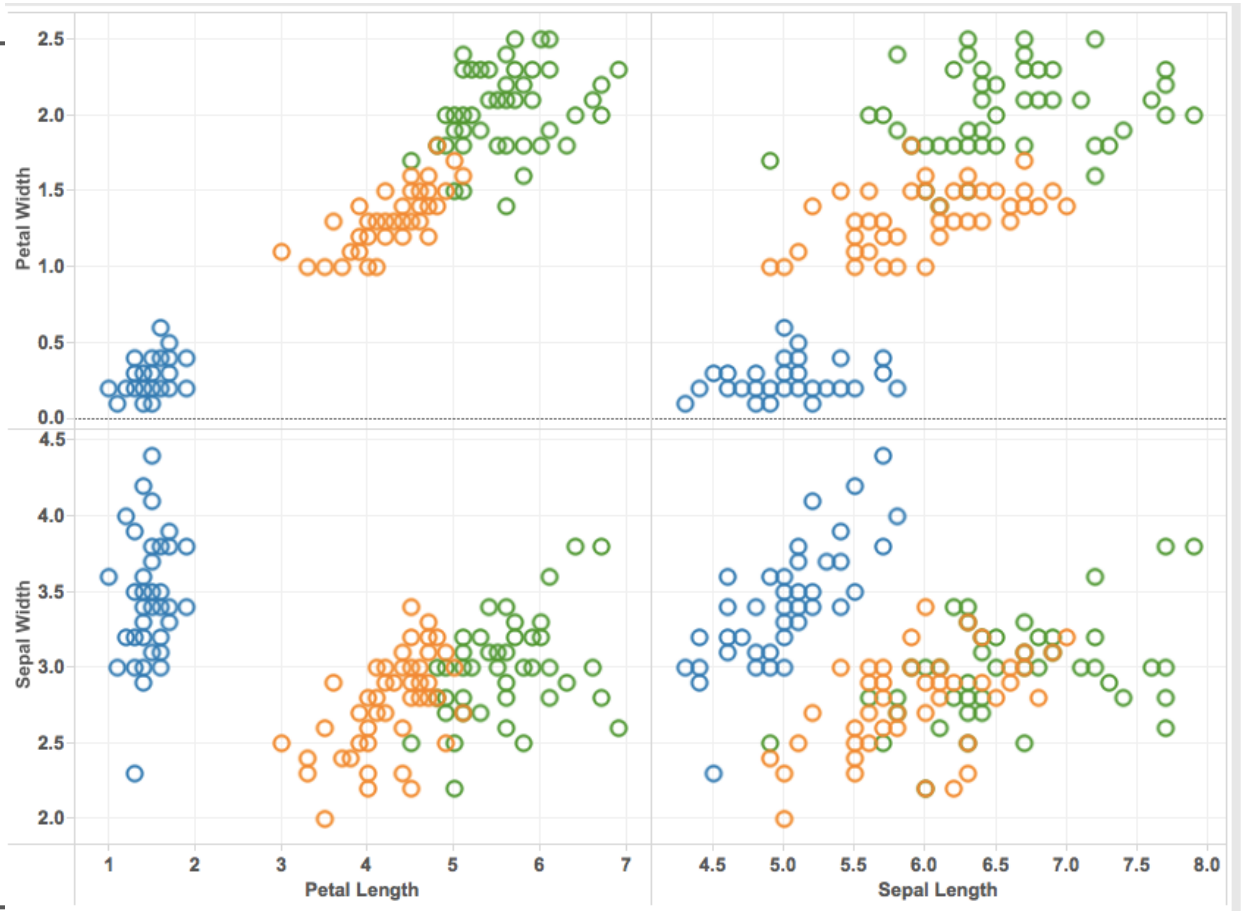
Same dataset, different encoding



Different subsets, different encoding



same encoding, different subsets

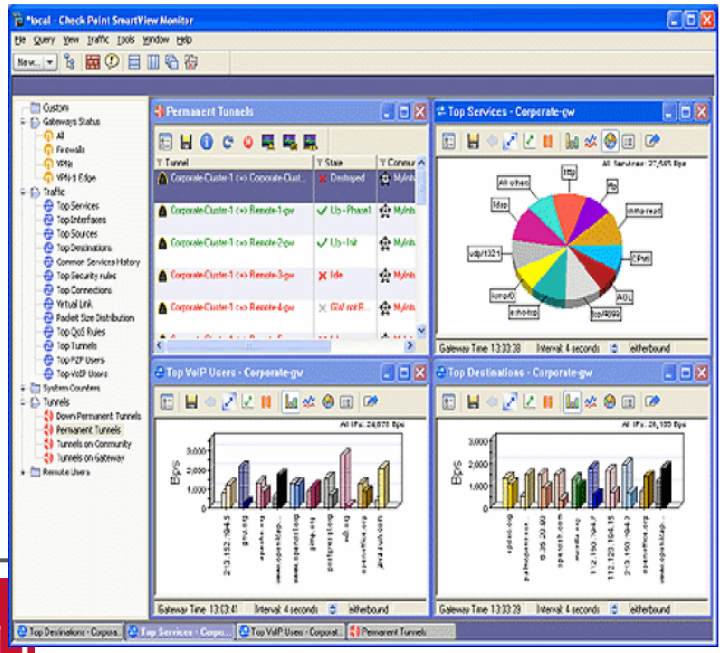


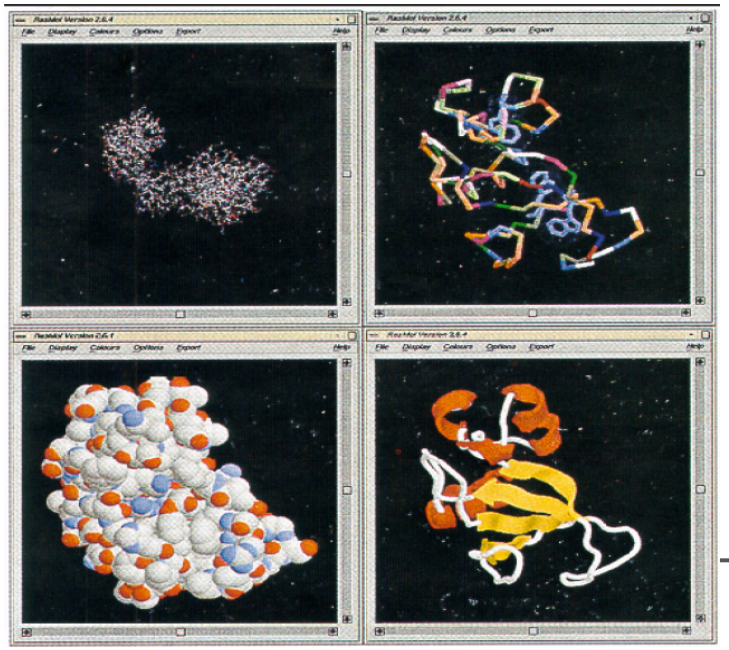
Replace, Replicate, Overlay? [Munzner]

- when to do which
- design tradeoffs
  - always replace: too much reliance on memory
  - always replicate: too many windows
  - always overlay: too much clutter in single window

Multiple Views

- "Guidelines for Using Multiple Views in Information Visualization"
  - Baldonado, Woodruff and Kichinsky AVI 00





Multiple Views: 8 Guidelines

- Rule of Diversity:
  - Use multiple views when there is a diversity of attributes
- Rule of Complementarity:
  - Multiple views should bring out correlations and/or disparities
- Rule of Decomposition: “Divide and conquer”.
  - Help users visualize relevant chunks of complex data
- Rule of Parsimony:
  - Use multiple views minimally

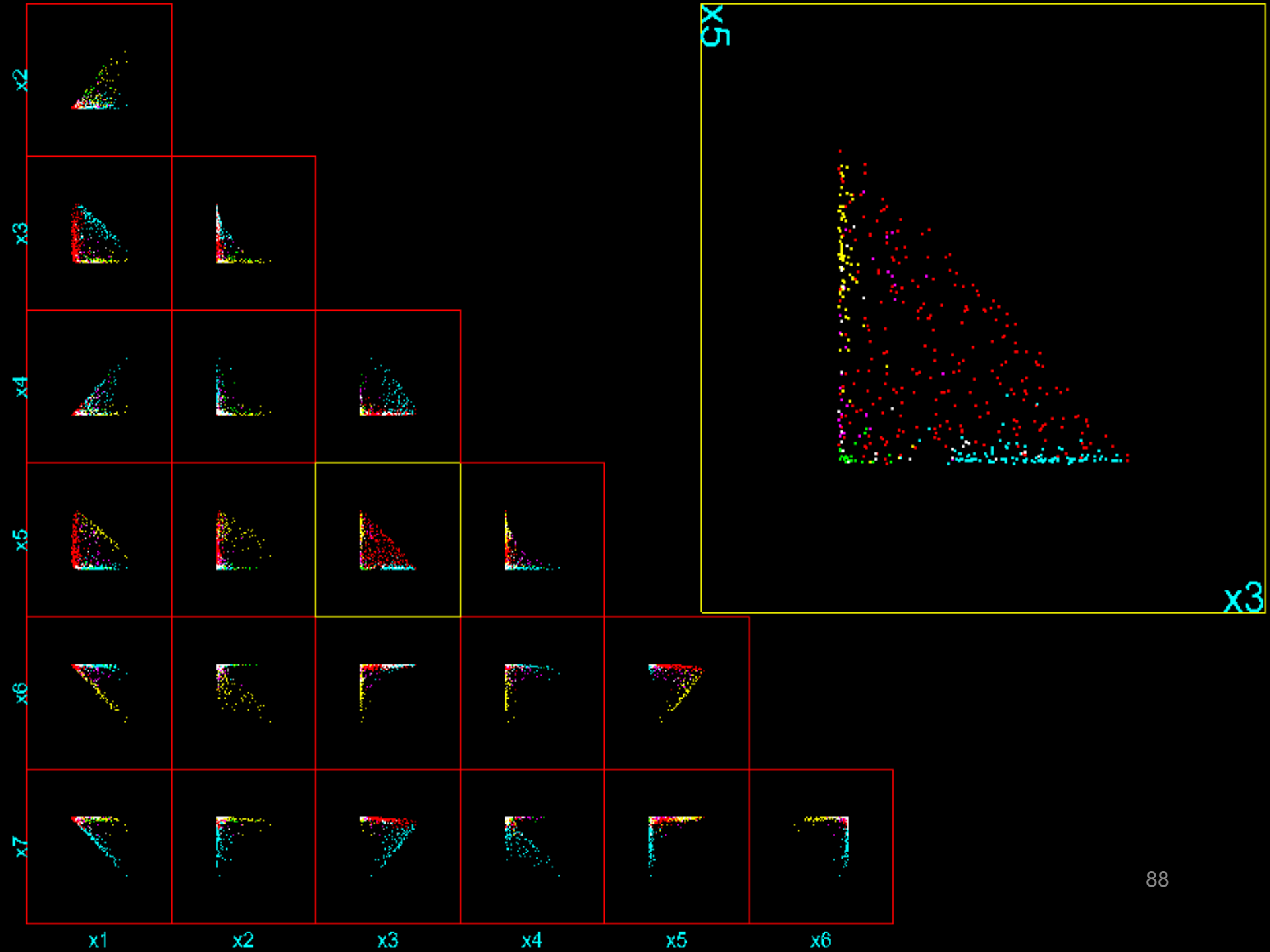
8 Guidelines Cont'd

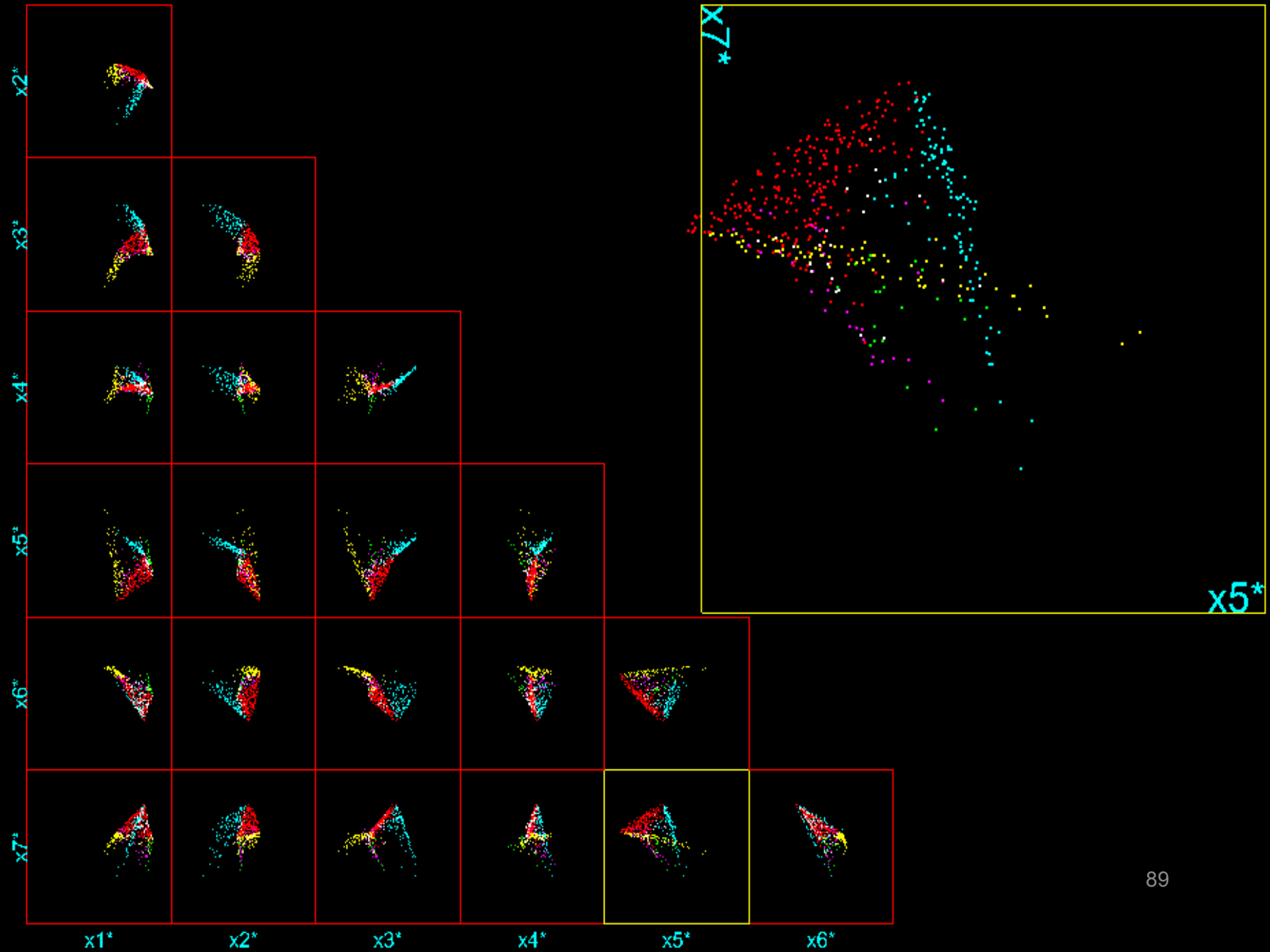
- Rule of Space/Time Resource
  - Optimization: Balance spatial and temporal benefits of presenting and using the views
- Rule of Self Evidence:
  - Use cues to make relationships apparent.
- Rule of Consistency:
  - Keep views and state of multiple views consistent
- Rule of attention management:
  - Use perceptual techniques to focus user attention

Interactive “Grand Tour” Techniques

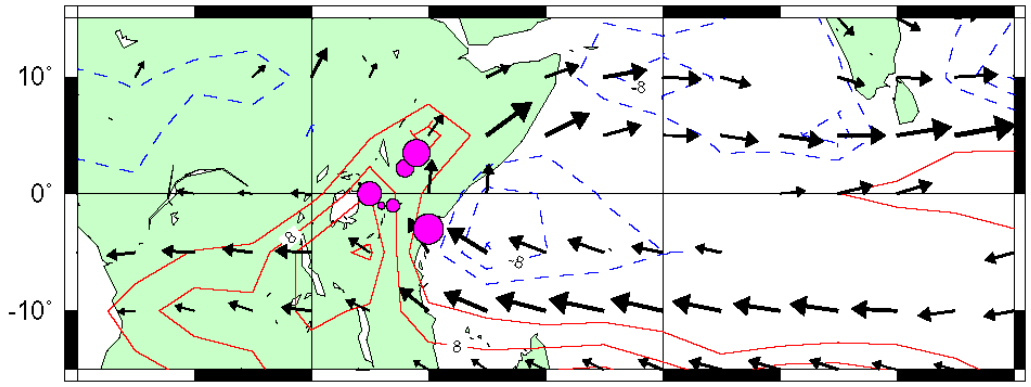
“Grand Tour” idea
- Cycle continuously through multiple projections of the data
- Cycles through all possible projections (depending on time constraints)
- Projects can be 1, 2, or 3d typically (often 2d)
- Can link with scatter plot matrices (see following example)

- e.g. XGOBI visualization package (available on the Web)
  - http://public.research.att.com/~stat/xgobi/
- Example on following 2 slides
  - 7dimensional physics data, color-coded by group, shown with
    (a) Standard scatter matrix
    (b) static snapshot of grand tour

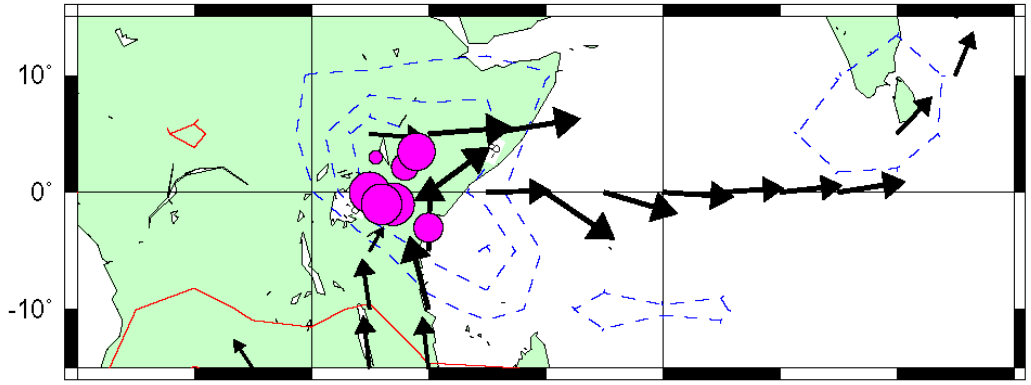

x2
x3
x4
x5
x6
x7
x1
x2
x3
x4
x5
x6
x5
x3


x2*
x3*
x4*
x5*
x6*
x7*
x1*
x2*
x3*
x4*
x5*
x6*
x7*
x5*

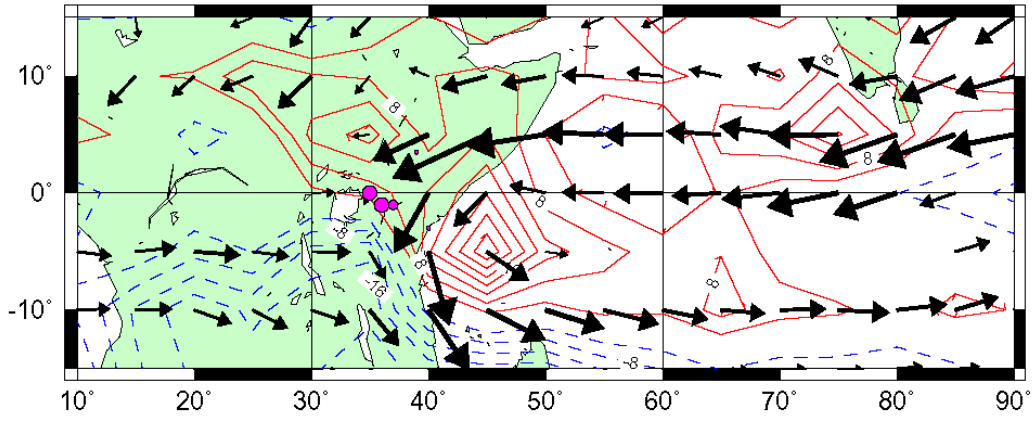
a) State 1 (830 d)



b) State 2 (1083 d) (winds x 3)



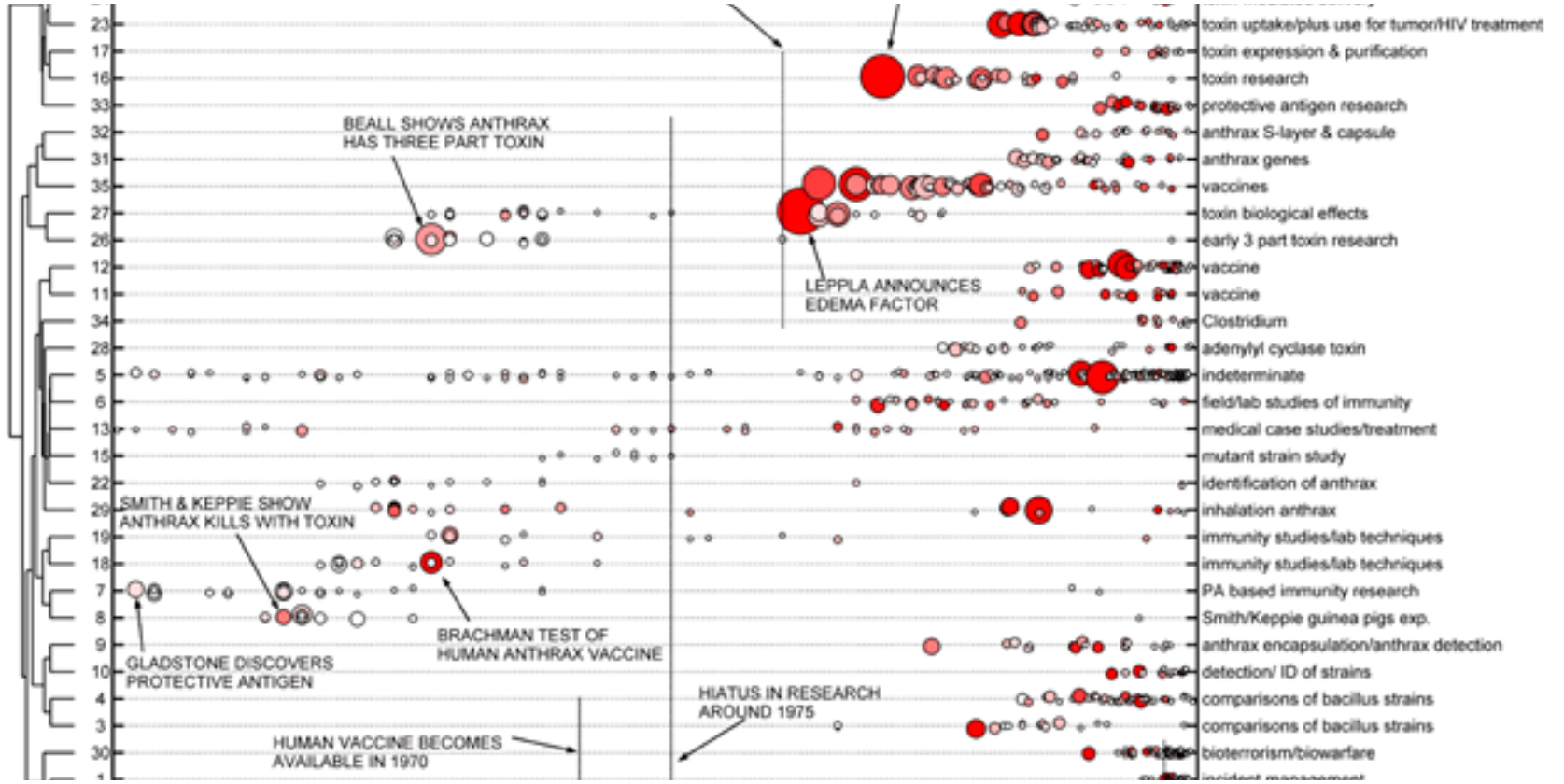
c) State 3 (755 d)



Summary of 189k newsgroups and 257M postings
Green = increase in postings in 2004 over 2003, red = decrease
Uses “treemap” technique. Details at http://jcmc.indiana.edu/vol10/issue4/turner.html

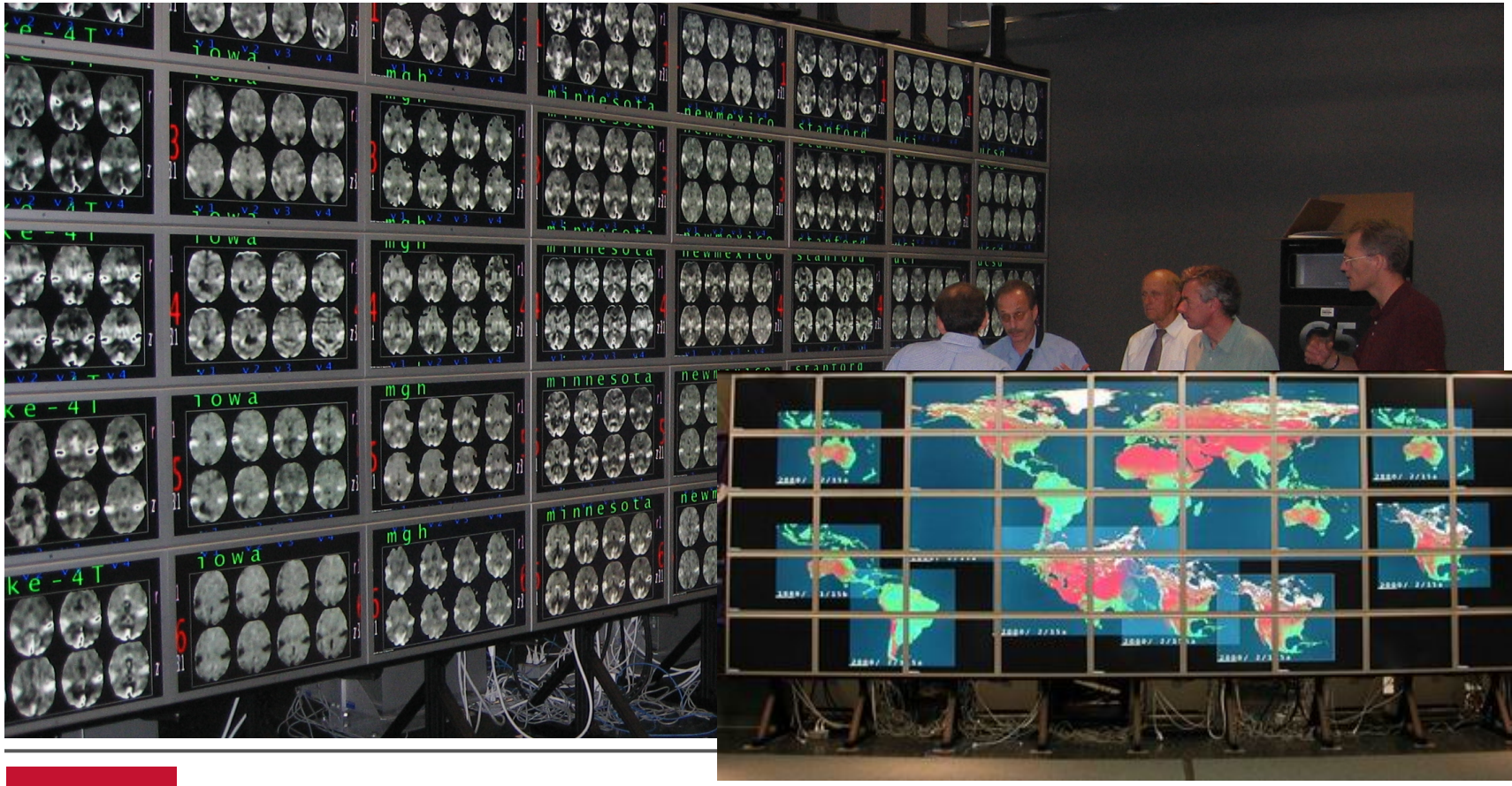
Time-line Visualization of Research Fronts

(Morris et al, JASIST, 2003)



Interactive MultiTile Visualization

(Falko Kuester's HIPerWall system, Calit2, UCI)



Common Approaches to the Problem of Scale

- Sampling
- Filtering
- Aggregation and Summarization
- Dimensionality Reduction (e.g., PCA, MDS)
- Binning
- Multiresolution Methods***

Multiple Resolutions in Visual analytics

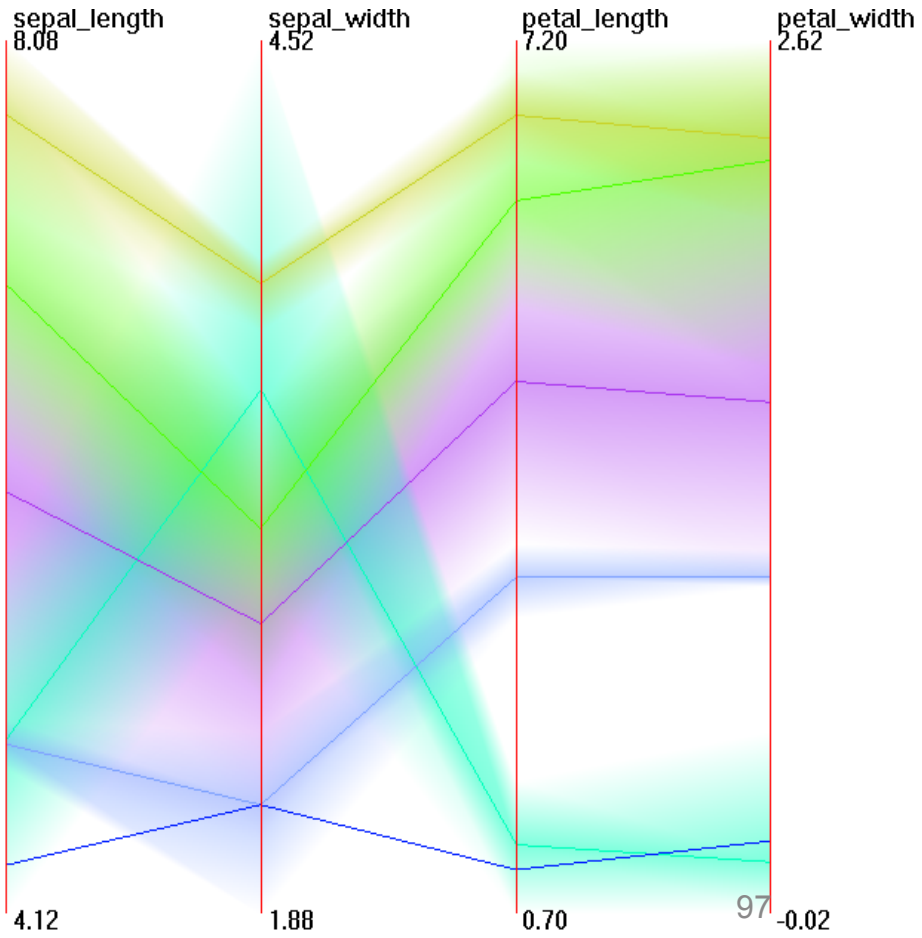
- For each target (number of records, dimensions, distinct nominal values)
  - Apply hierarchical clustering algorithm
  - Identify representative value for each non-terminal cluster
  - Compute cluster descriptors to convey contents
  - Visualize representative values using traditional tools, augmented with descriptors
  - Provide interactive tools to navigate, modify, and filter the hierarchical structure

Visualizing Large Numbers of Records: Mean-Band Method

- User specifies focus region in data space and level of detail for focused/unfocused areas
- Mean (or other derived) value for each cluster displayed in color based on its location in hierarchy
- Opacity bands around data points show population and extent of clusters

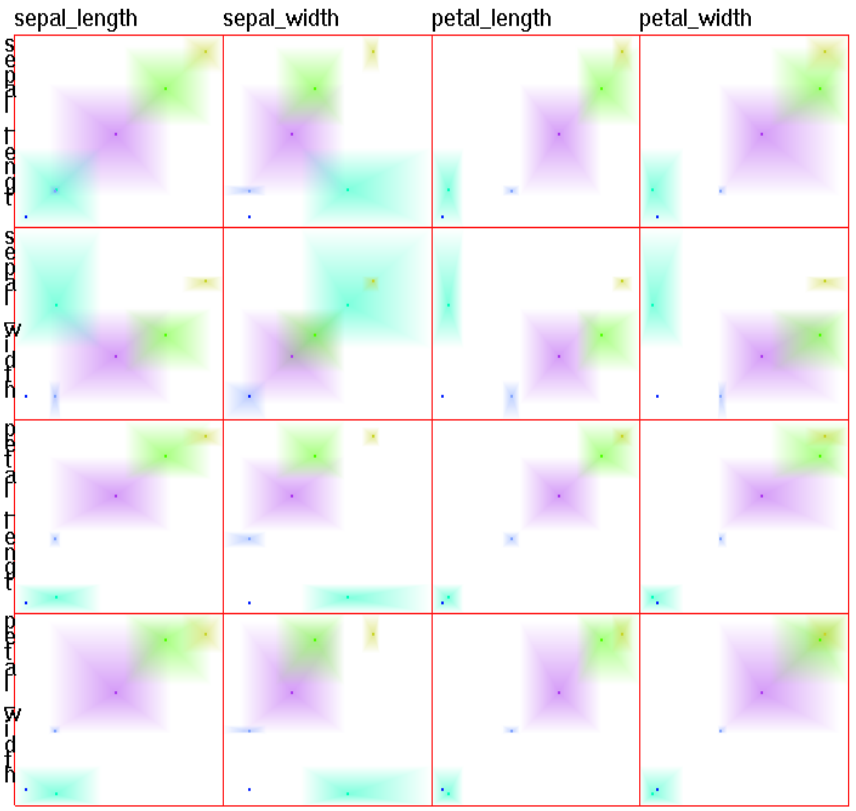
Hierarchical Parallel Coordinates

- Bands show cluster extents in each dimension
- Opacity conveys cluster population
- Color similarity indicates proximity in hierarchy



Hierarchical Scatterplots

- Clusters displayed as rectangles, showing extents in 2 dimensions
- Color/opacity consistently used for relational and population info



The Role of Interaction

- User needs to interact with display, examine interesting patterns or anomalies, validate hypotheses
- Selection allows isolation of subset of data for highlighting, deleting, focussed analysis
- Navigation allows alternate views, drill-down for details
- Direct (clicking on displayed items) vs. indirect (range sliders, text queries)
- Screen space (2-D) , data space (N-D), structure space (spatio-temporal, grids, hierarchies)

Navigating Hierarchies

- Drill-down, roll-up operations for more or less detail
- Need selection operation to identify subtrees for:
  - Exploration
  - Manipulation
  - Pruning
- Can be user-driven, data-driven, structure-driven