High Dimensional Data

Alark Joshi

High dimensional data

- Data with multiple dimensions, multiple variables or multiple attributes
- Cars dataset
  - Economy
  - Cylinders
  - Displacement
  - Power
  - Weight
  - Mph
  - Year

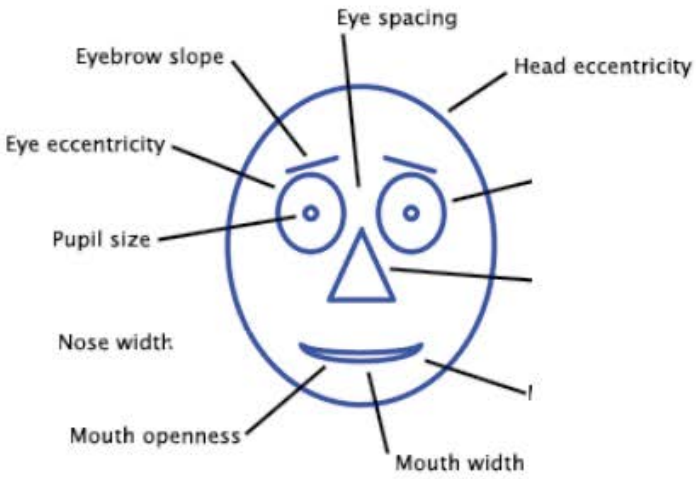
Scatterplots

- Great for visualizing 2D data
- Plot data attributes on x- and y-axis
- Scatterplot Matrix can be used to visualize multiple attributes

Scatterplot Matrix

- http://mbostock.github.com/d3/ex/splom.html

Chernoff Faces



Randomly selected parameters.
Change face.

Animation with random parameters.

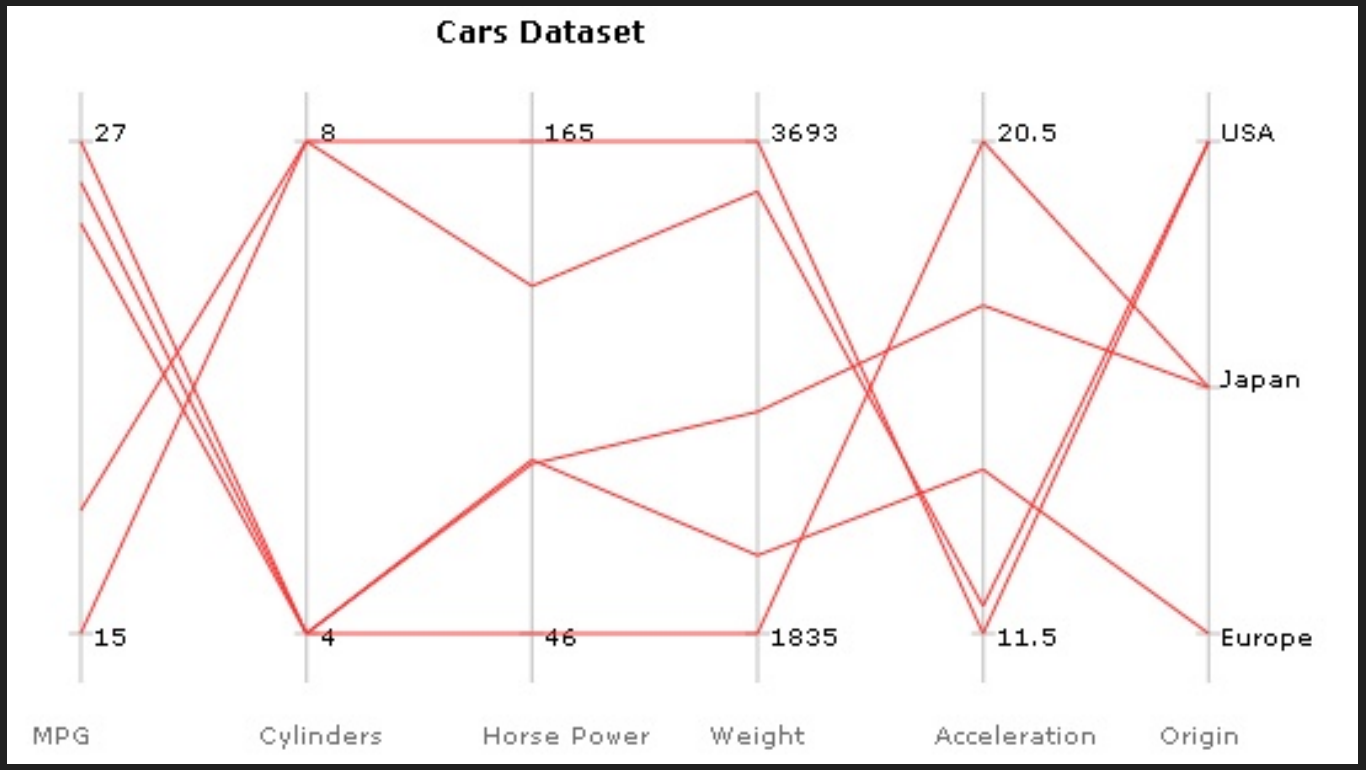
All 0.

All 0.5.

All 1.

Parallel Coordinates

- Instead of having only 2 orthogonal axes (scatter plots), have parallel axes



Parallel Coordinates

- Connect variables for each data entity with a line

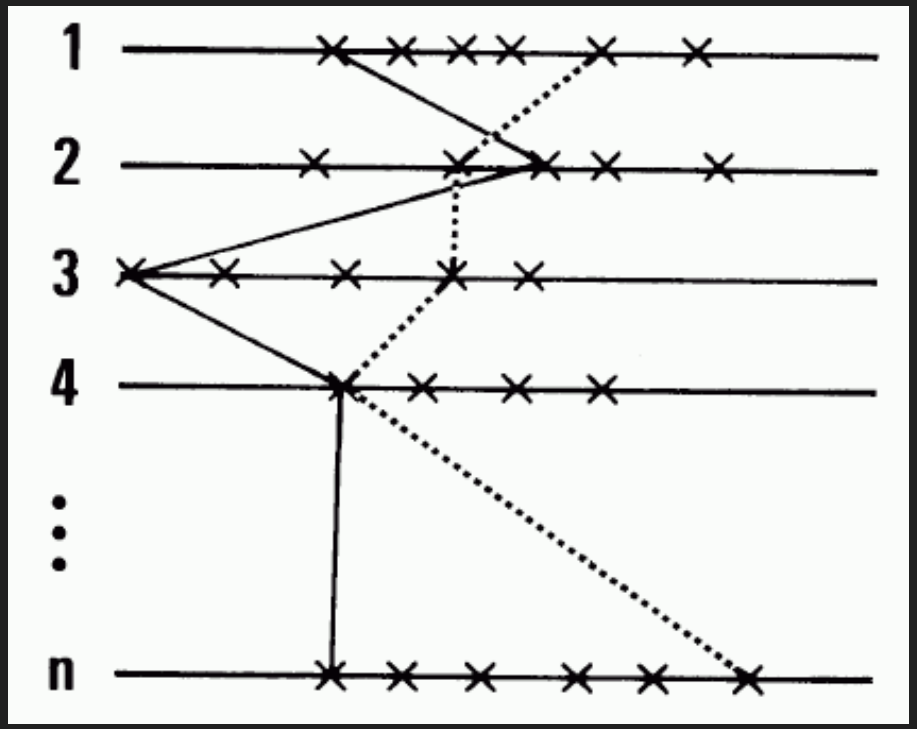


Image credits: Hyperdimensional Data Analysis Using Parallel Coordinates, Edward J. Wegman Journal of the American Statistical Association , Vol. 85, No. 411 (Sep., 1990), pp. 664-675

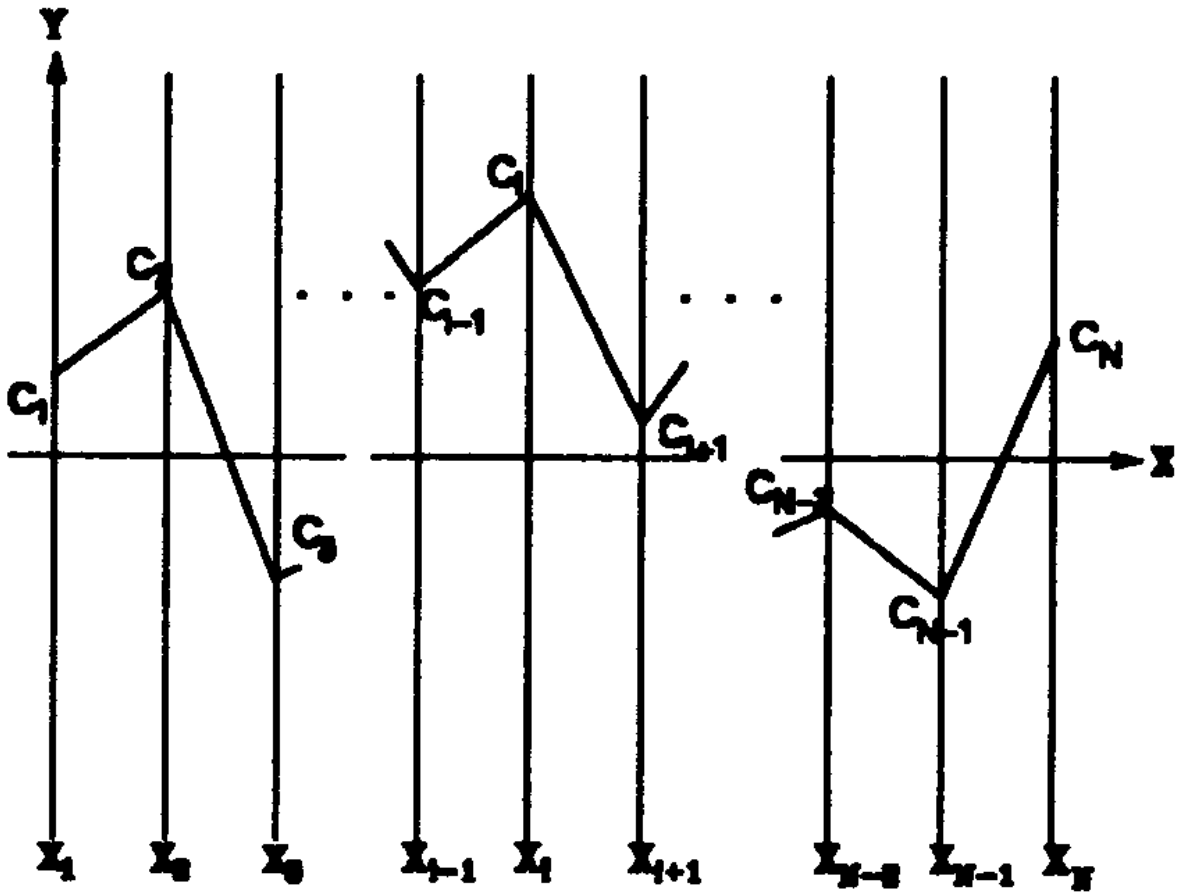


Figure 1: – Parallel axes for R^N.

The polygonal line shown represents the point $C = (c_1 , \ldots , c_{i-1} , c_i , c_{i+1} , \ldots , c_N)$.

Five-dimensional hypersphere

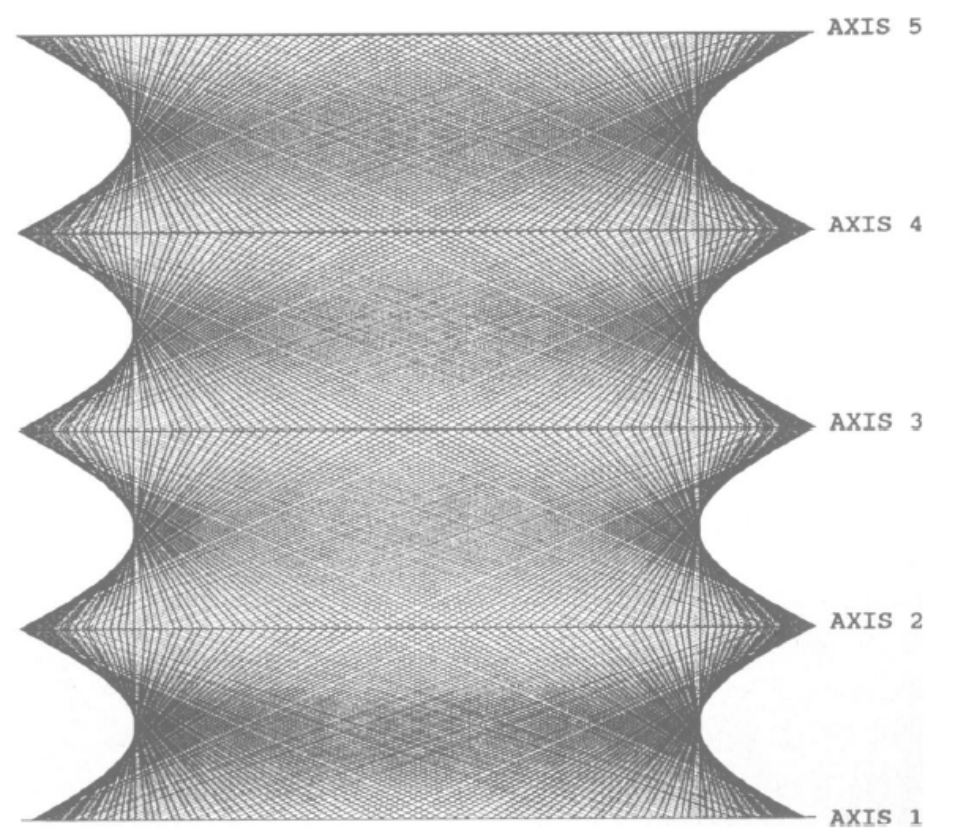
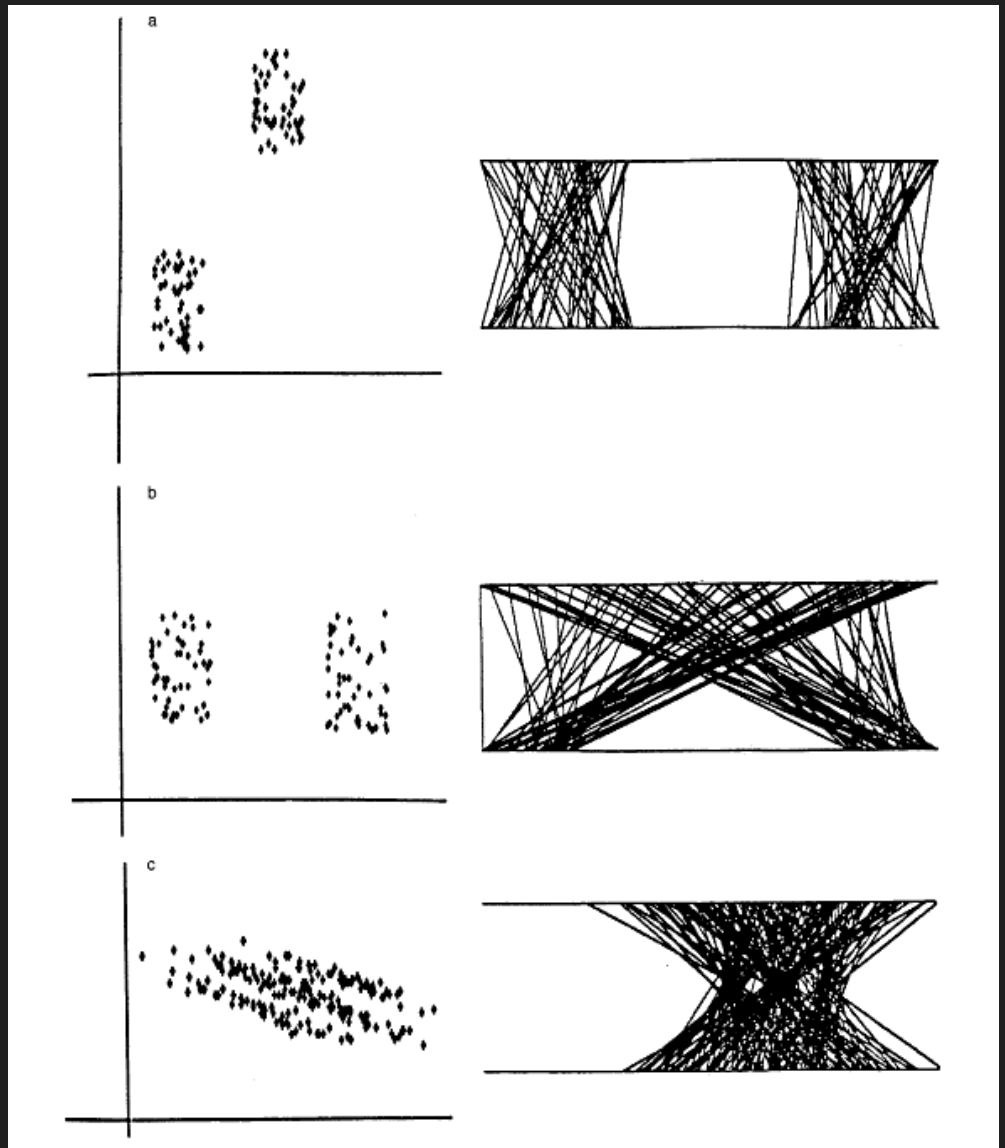


Figure 4. Parallel Coordinate Plot of a Five-Dimensional Hypersphere. Every axis pair is correspondingly the parallel coordinate plot of a circle. Notice the hyperbolic envelopes.

Image credits: Hyperdimensional Data Analysis Using Parallel Coordinates, Edward J. Wegman Journal of the American Statistical Association , Vol. 85, No. 411 (Sep., 1990), pp. 664-675

Clustering in scatterplots vs PC



Clustering separated in x and y

Clustering separated in x but not in y

Clustering not separated in either projection

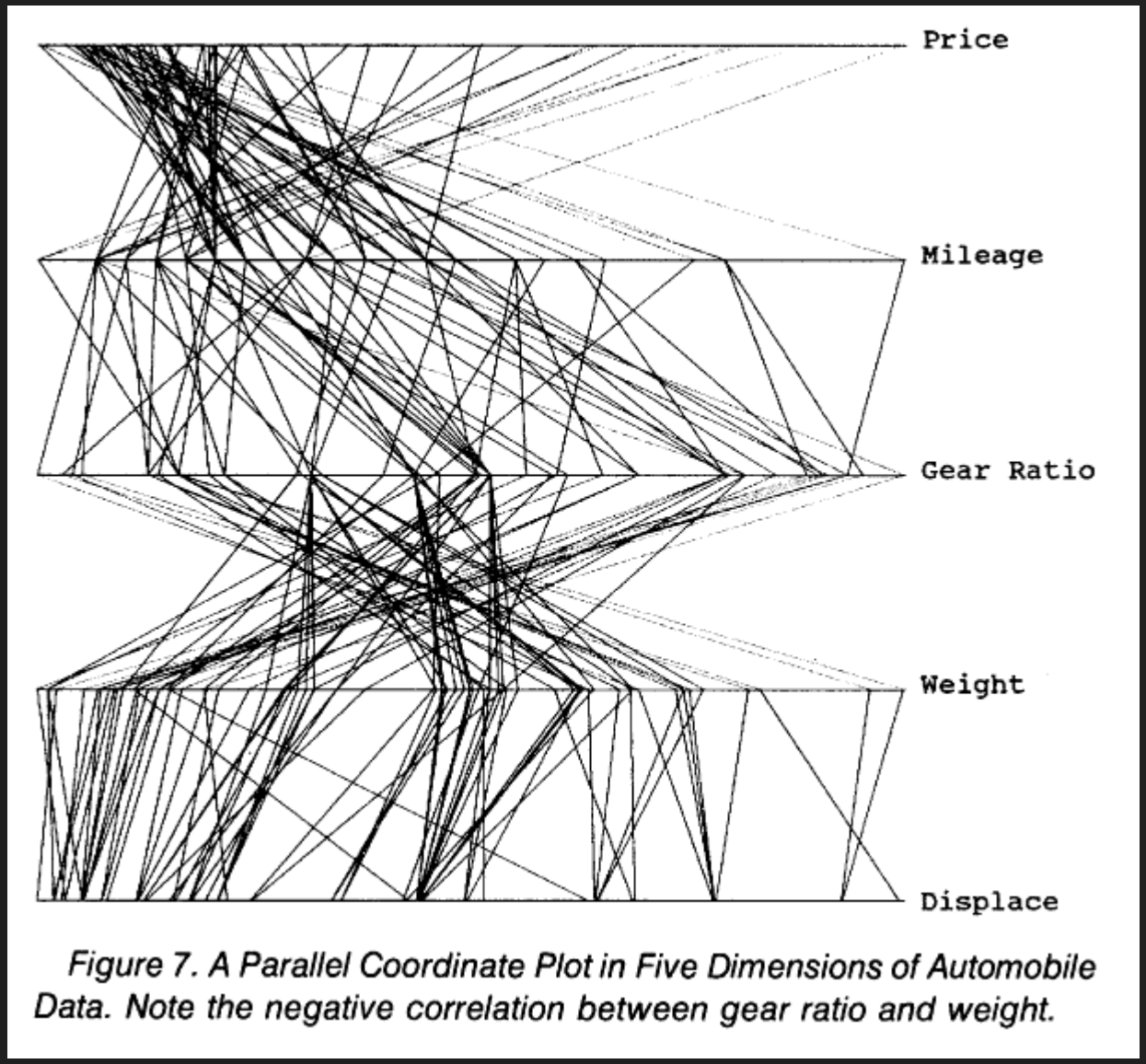


Figure 7. A Parallel Coordinate Plot in Five Dimensions of Automobile Data. Note the negative correlation between gear ratio and weight.

Image credits: Hyperdimensional Data Analysis Using Parallel Coordinates, Edward J. Wegman Journal of the American Statistical Association , Vol. 85, No. 411 (Sep., 1990), pp. 664-675

PC Plot showing American Cars

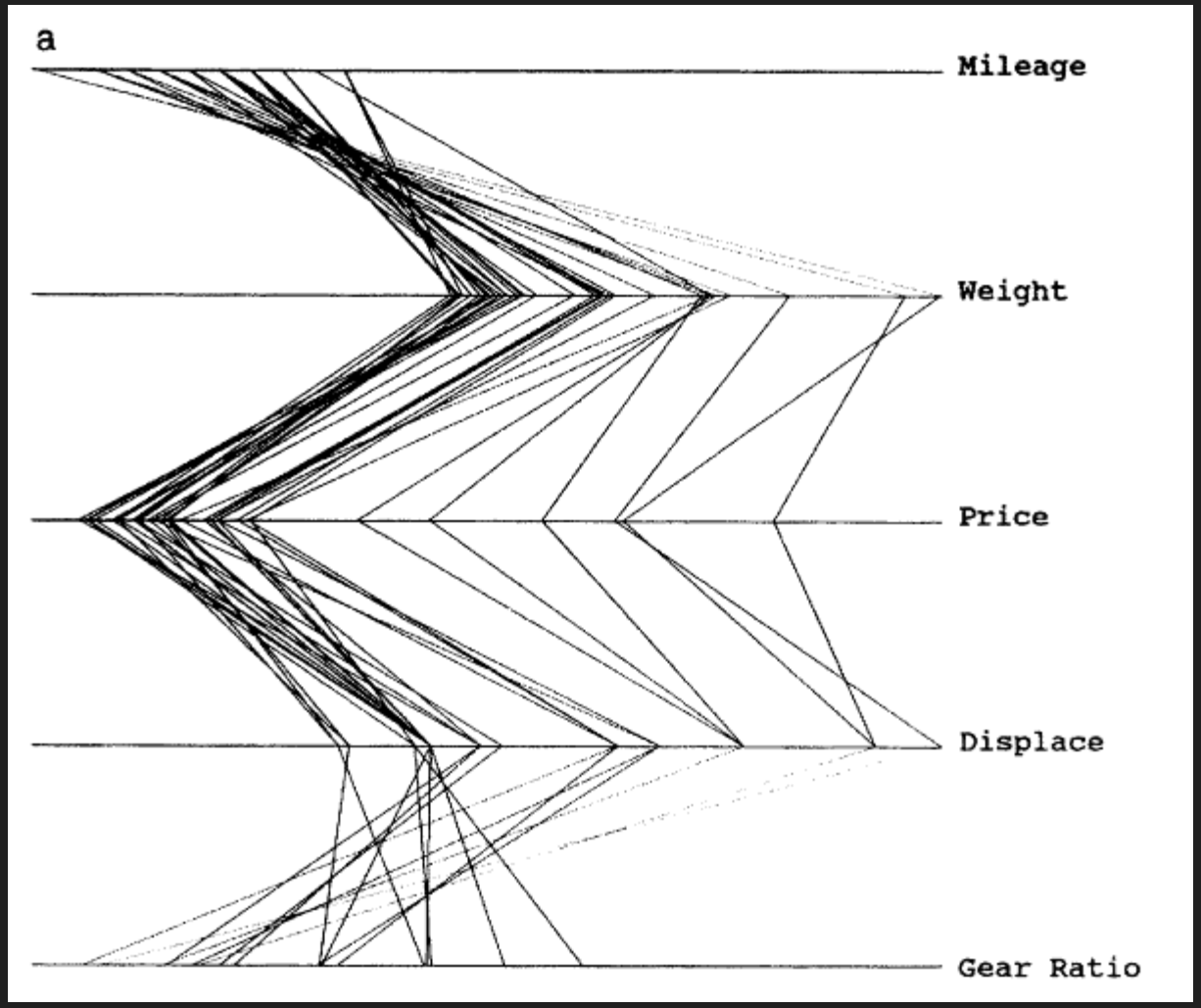


Image credits: Hyperdimensional Data Analysis Using Parallel Coordinates, Edward J. Wegman Journal of the American Statistical Association , Vol. 85, No. 411 (Sep., 1990), pp. 664-675

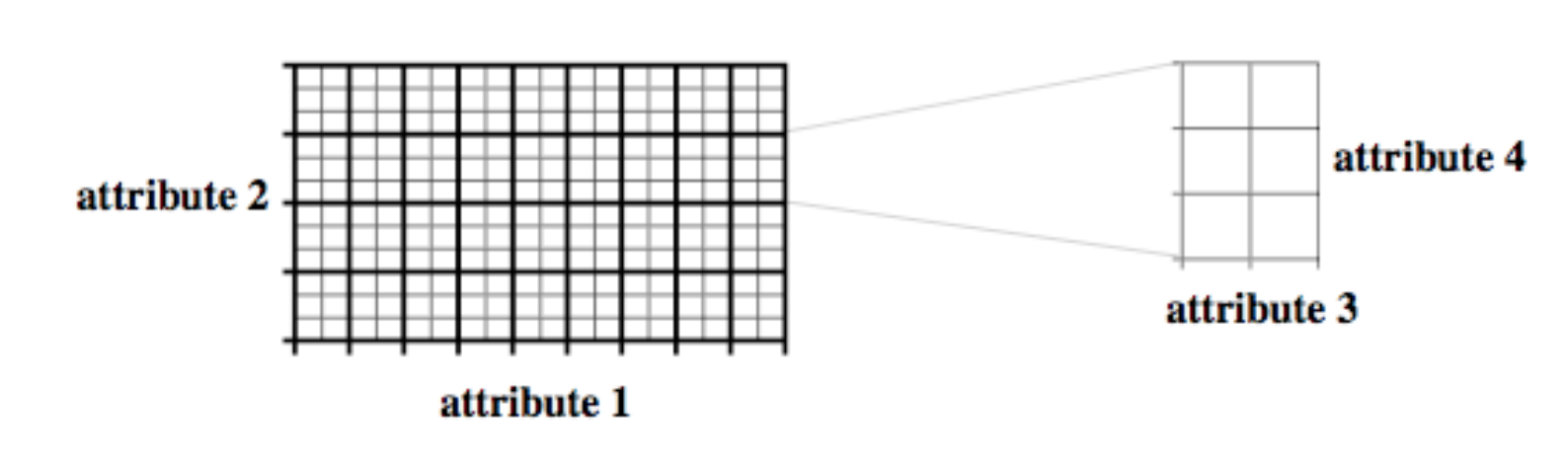
Demo

- Parallel coordinates in D3
  - http://bl.ocks.org/1341281

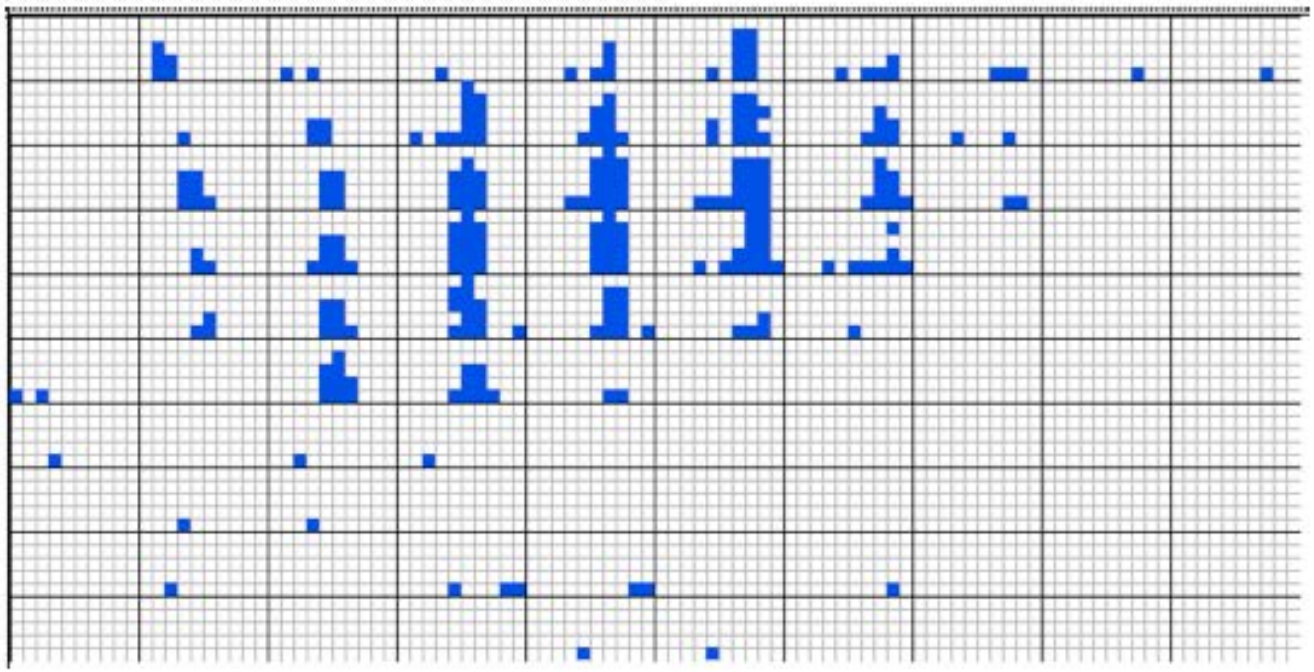
PC: Axis Ordering

- Geometric interpretations
  - Hyperplane, hypersphere
  - Points do have an intrinsic order
- Nominal data
  - No intrinsic order
  - Indeterminate/arbitrary order
    - Weakness of many techniques
    - Downside: human-powered search
    - Upside: Powerful interaction technique
- In most implementations, a user can interactively swap axes

Dimensionality Stacking

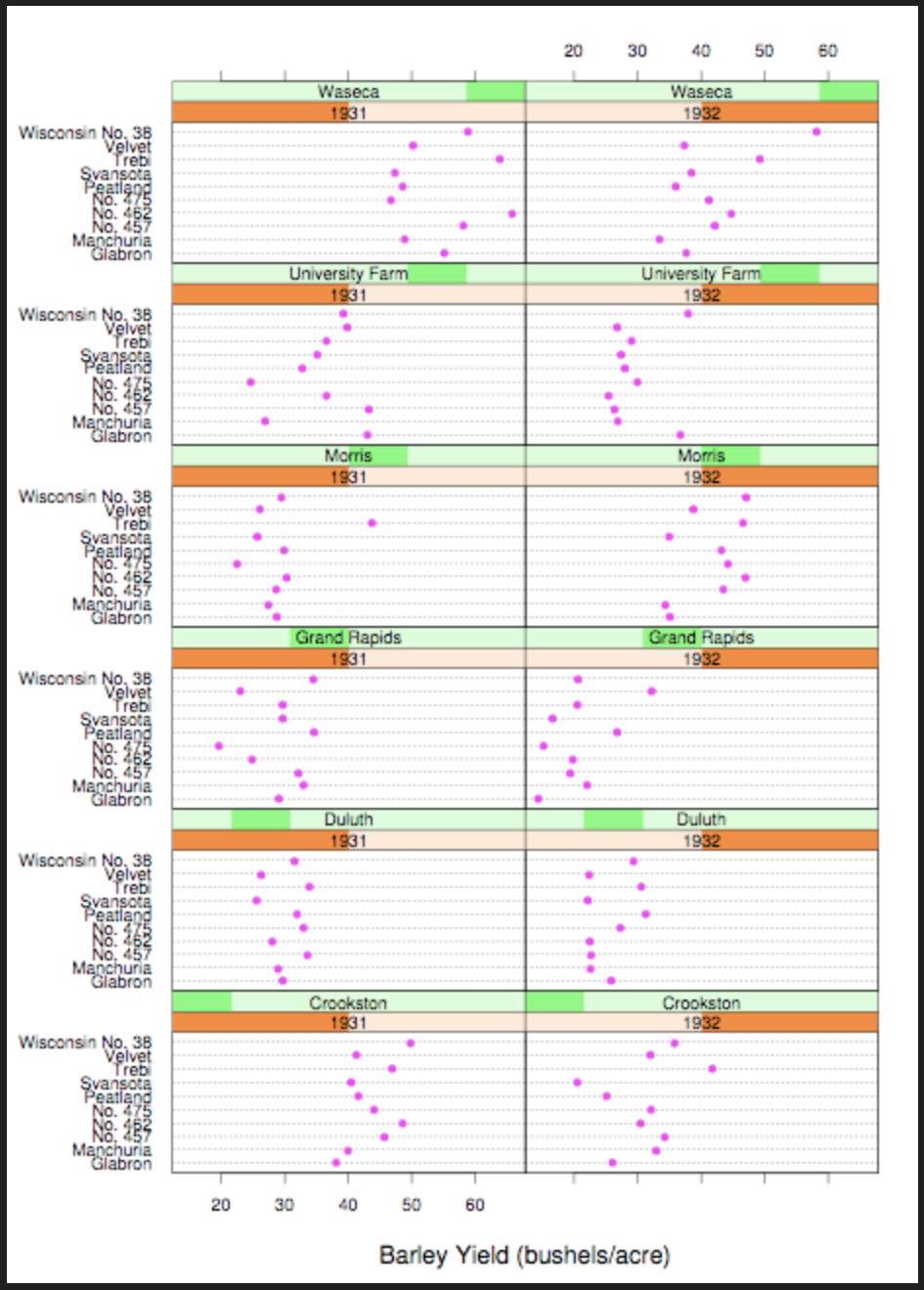


Dimensionality Stacking

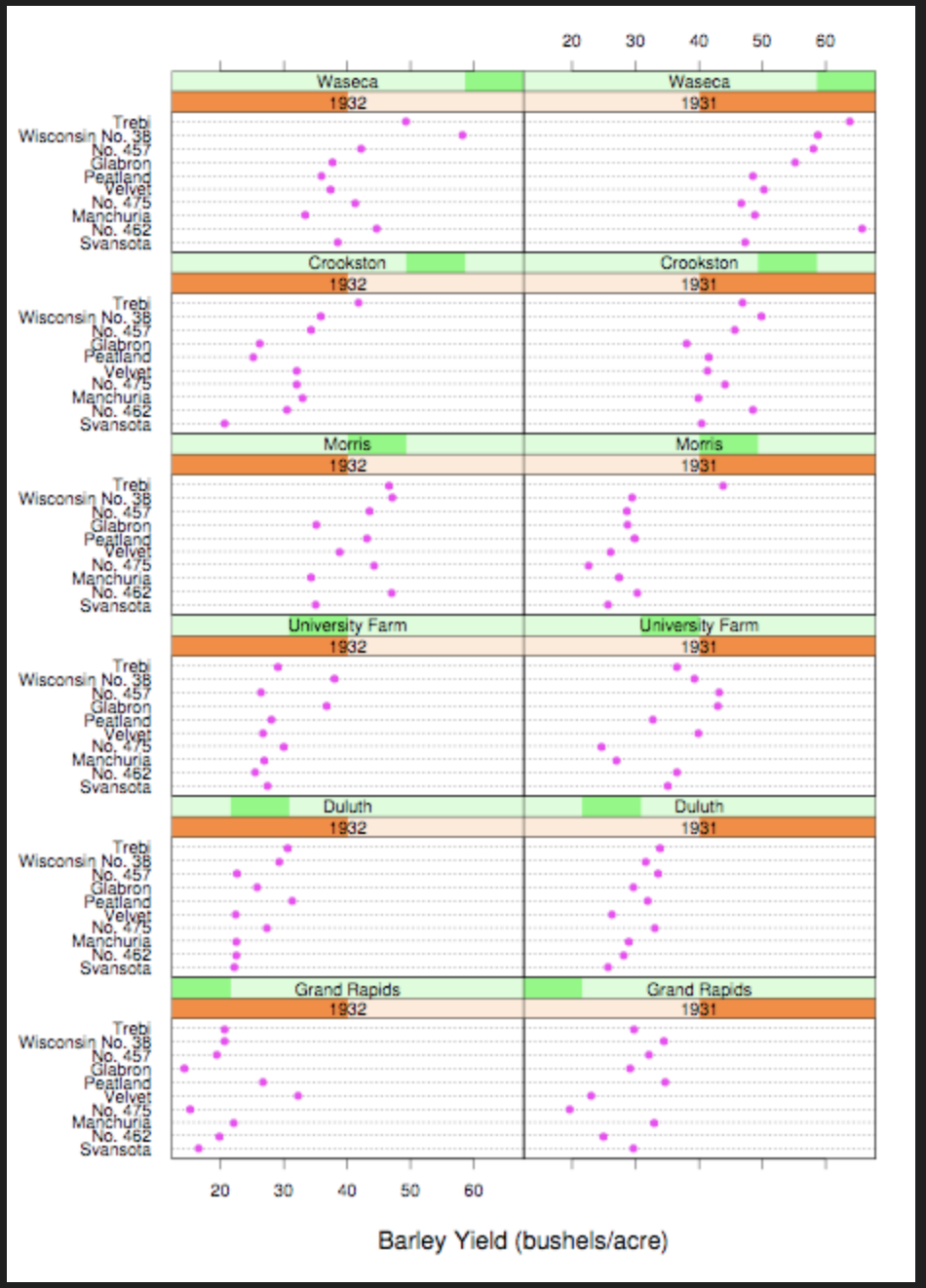


visualization of oil mining data with longitude and latitude mapped to the outer x-, y- axes and ore grade and depth mapped to the inner x-, y- axes

Image credits: Matt Ward et al.



Alphabetical



Median Value

Pixel-oriented techniques

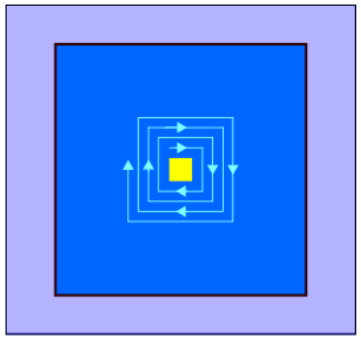


Figure 1:
Spiral Shaped Arrangement
of one Dimension

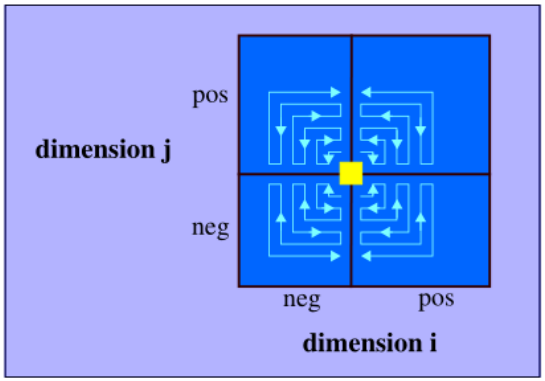


Figure 3: 2D-Arrangement of one Dimension

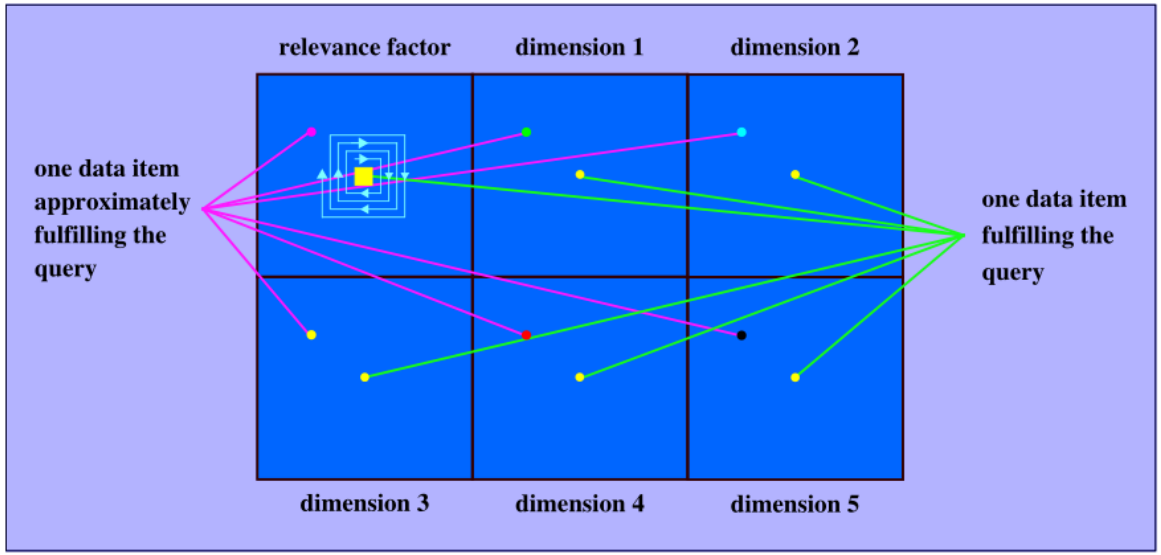


Figure 2: Arrangement of Windows for Displaying five-dimensional Data

Image credits: Daniel A. Keim and Hans-Peter Kriegel. 1995. VisDB: a system for visualizing large databases. In *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*

Pixel-oriented techniques

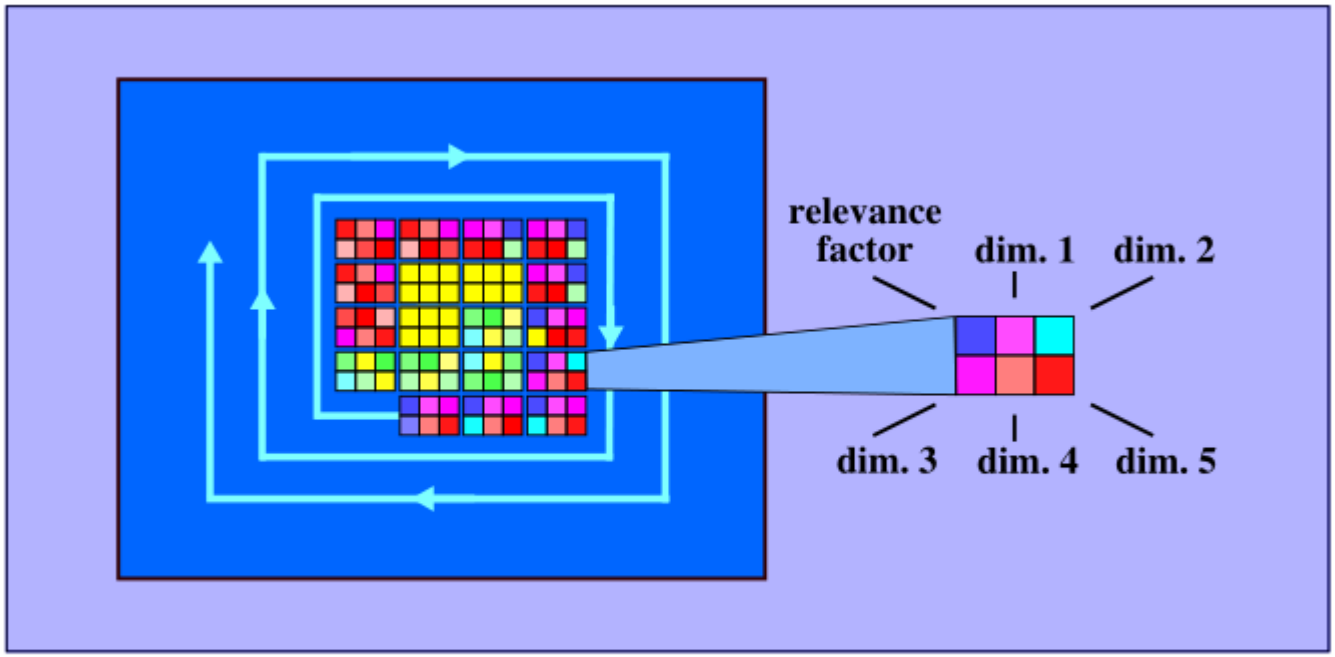
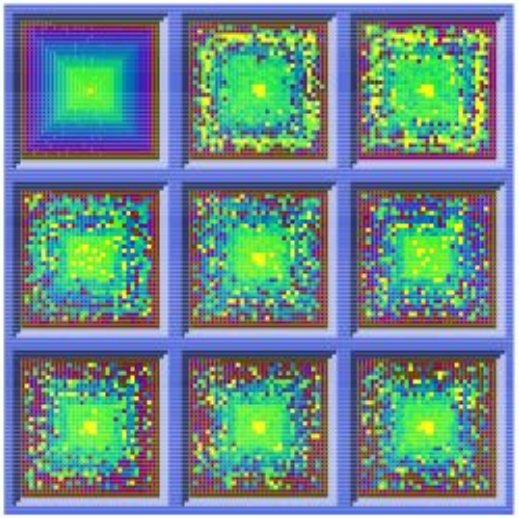


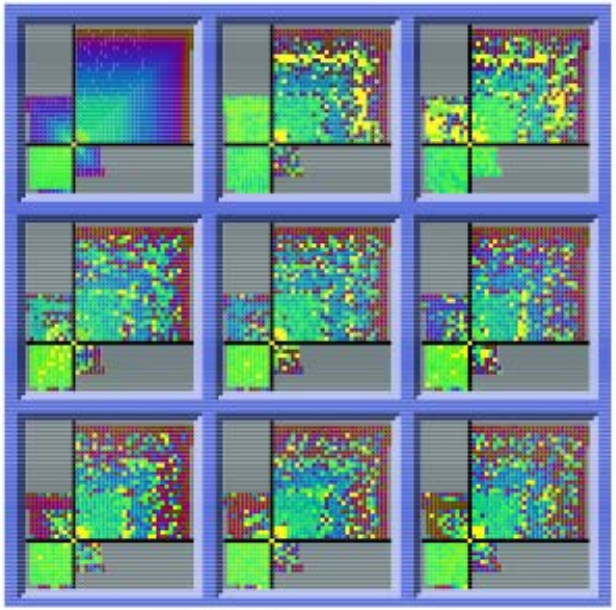
Figure 4: Grouping Arrangement for five-dimensional Data

Image credits: Daniel A. Keim and Hans-Peter Kriegel. 1995. VisDB: a system for visualizing large databases. In *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*

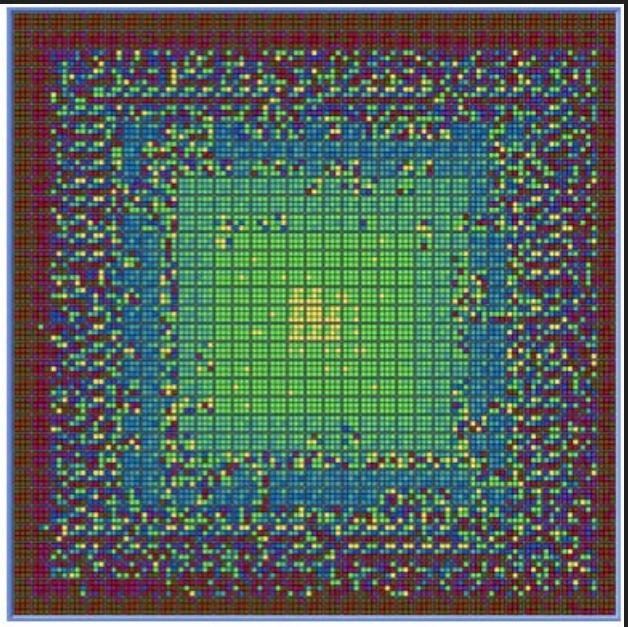
Visualizing 8-dimensional data



a. Basic Visualization Technique



b. 2D-Arrangement



c. Grouping Arrangement

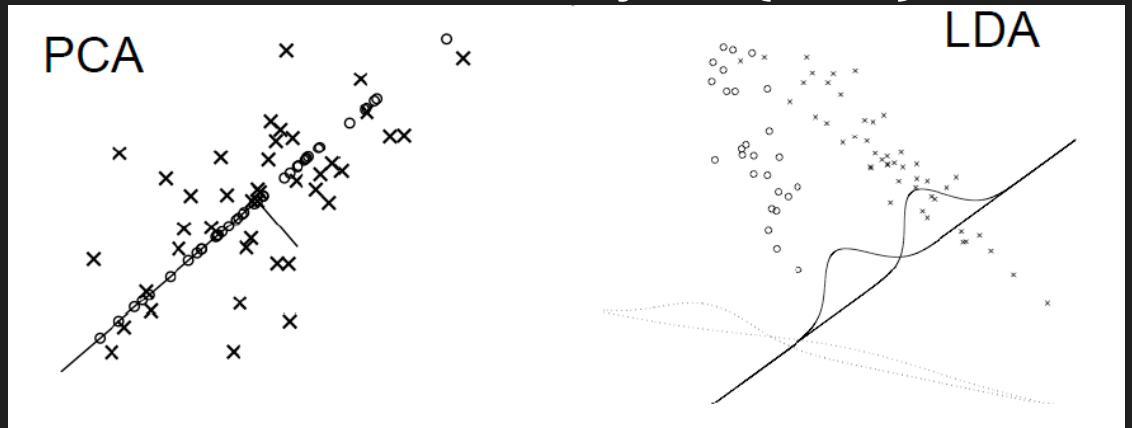
Image credits: Daniel A. Keim and Hans-Peter Kriegel. 1995. VisDB: a system for visualizing large databases. In *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*

Dimensionality Reduction

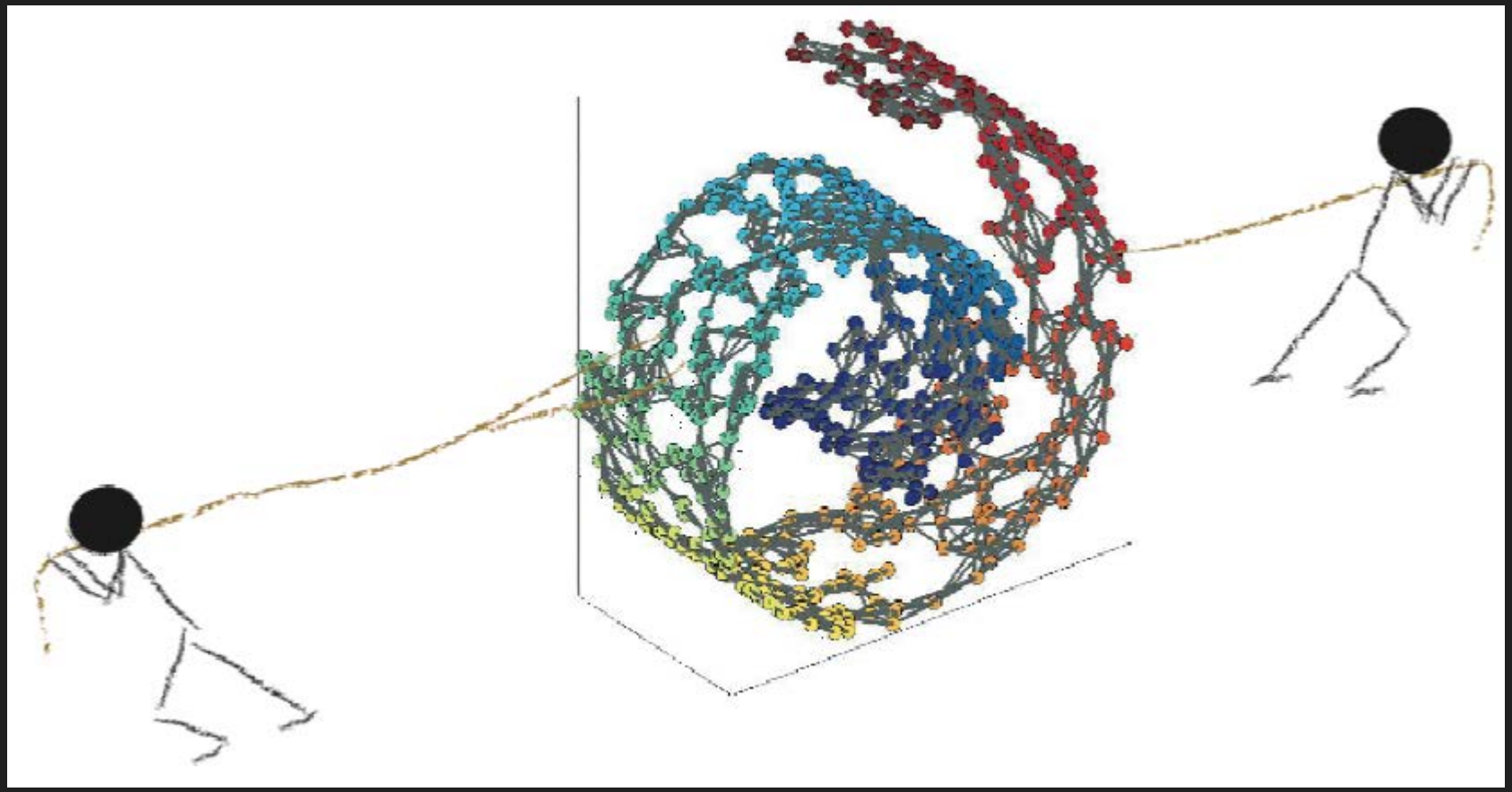
- Mapping multidimensional space into space of fewer dimensions
  - Typically 2D for clarify
  - 1D/3D possible
  - Preserve and communicate variance in data as much as possible
  - Show underlying structure of data
- Linear vs non-linear approaches

Linear Dimensionality Reduction

- Based on linear projections
- Given dimensions has a strong meaning
- Preserve the linearity in the layout
- Examples:
  - Principal Component Analysis (PCA)
  - Independent Component Analysis (ICA)
  - Linear Discriminant Analysis (LDA), …



Problems for Linear Approaches



Non-linear Dimensionality Reduction

- Does not assume any inherent meaning to given dimensions
- Minimize differences between interpoint distances in high and low dimensions
- Examples:
  - Multidimensional scaling (MDS)
  - Isomap
  - Local linear embedding (LLE)

Isomap

- 4096 D to 2D
- 2D: wrist rotation, fingers extension

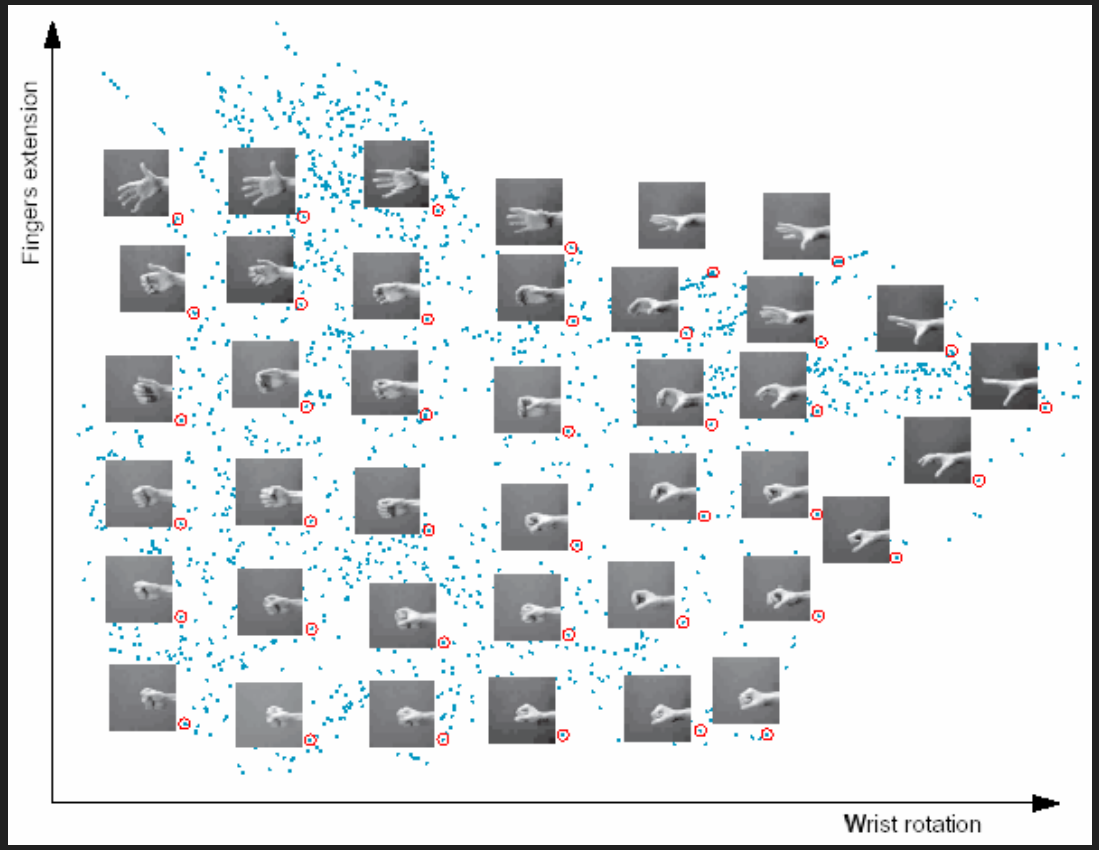


Image credits: Global Geometric Framework for Nonlinear Dimensionality Reduction. Tenenbaum, de Silva and Langford. Science 290 (5500): 2319-2323, 22 December 2000,

Goals

- Preserve and communicate as much variance as possible
- Find and display clusters
  - Compare/evaluate with previous clustering algorithms
- Understand structure
  - Absolution position is not reliable
  - Fine grained structure not reliable

Hierarchical Parallel Coordinates

- YH Fua, MO Ward, and IA Rundensteiner (1999), Hierarchical Parallel Coordinates for Exploration of Large Datasets, Proceedings of IEEE Visualization '99, pp. 43-50.
- Interactive visualization of large multivariate data sets
- Proposed a number of novel extensions to the parallel coordinates display technique
- Presentation by Danny

Dimension Ordering

- Determining dimension ordering important
  - Heuristic
  - Divide and conquer
    - Iterative hierarchical clustering
    - Representative dimensions

Dimension Ordering

- Choices
  - Similarity metrics
  - Importance metrics (variance, etc.)
  - Ordering algorithms
    - Optimal
    - Random swap
    - Simple depth-first traversal

Dimension Filtering

- Interaction
  - Structure-based brushing
  - Focus + context
  - Manual interaction through UI components

InterRing – Hierarchical Data Navigation

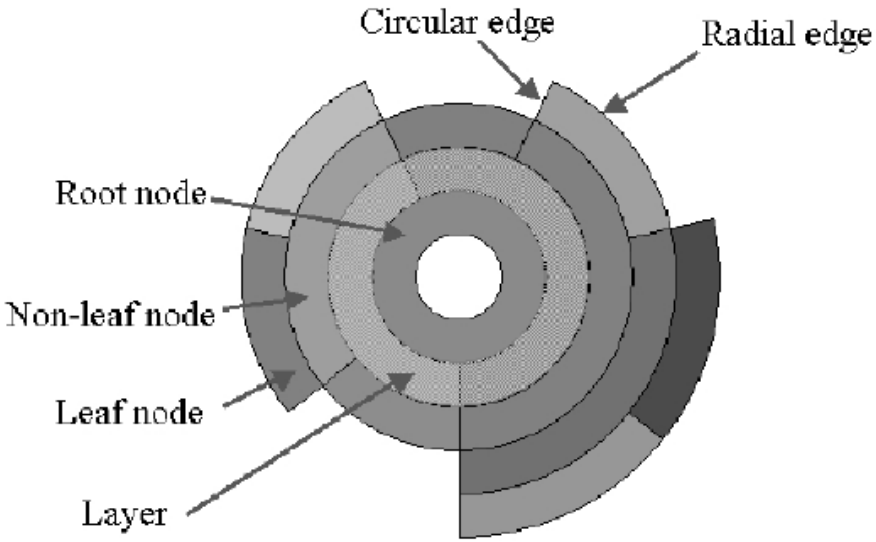
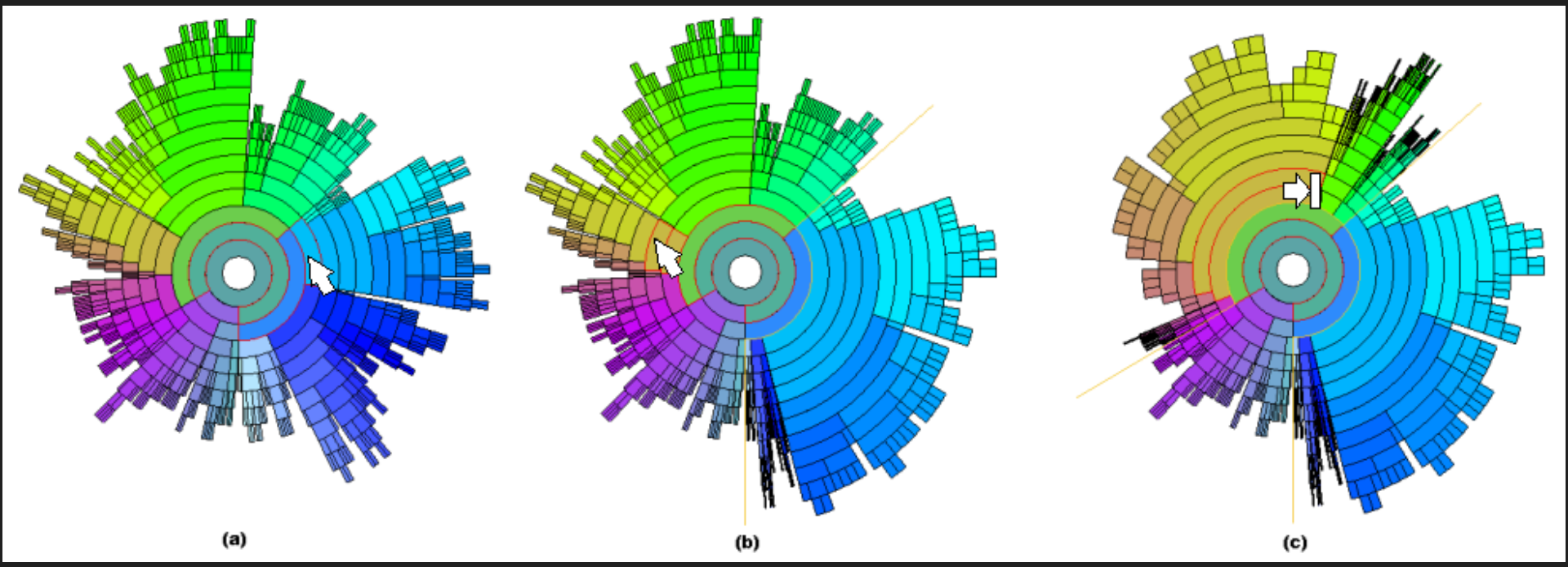


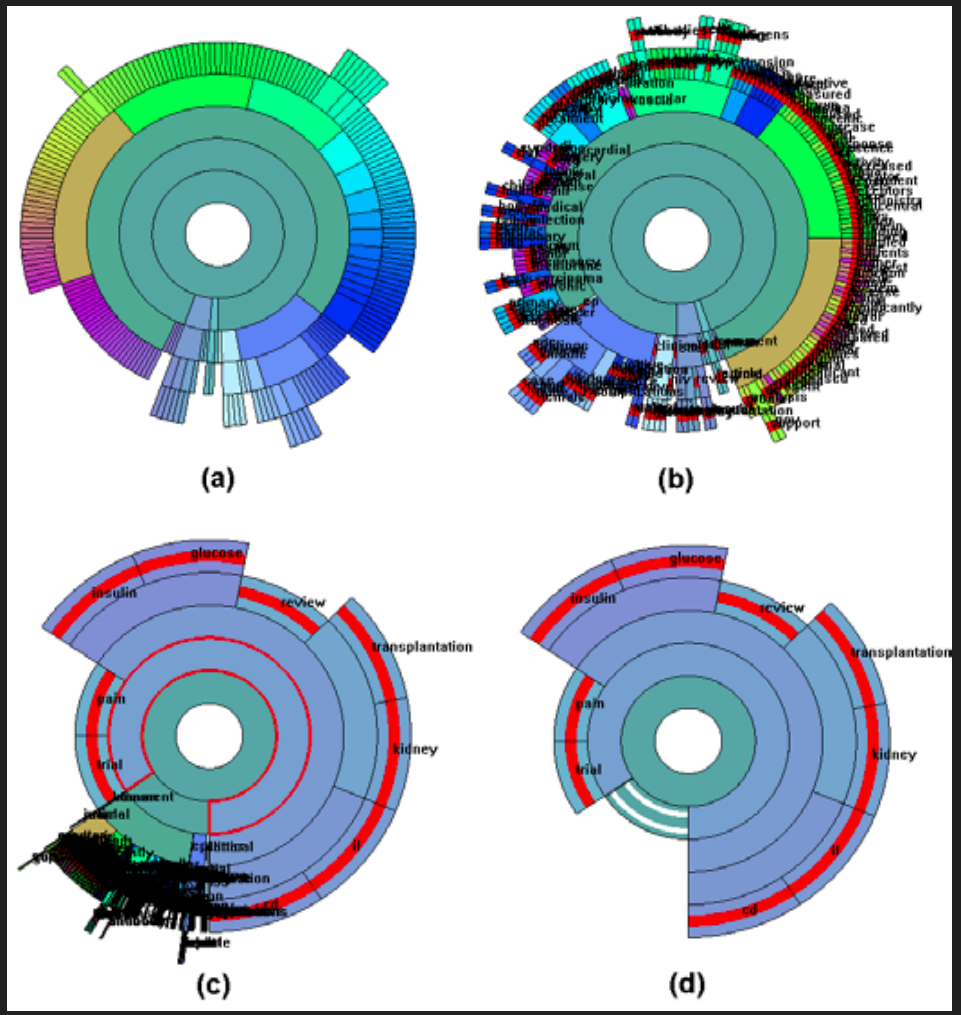
Figure 1: An RSF display with labeled components

Image credits: Jing Yang, Matthew O. Ward, Elke A. Rundensteiner, and Anilkumar Patro. 2003. InterRing: a visual interface for navigating and manipulating hierarchies. *Information Visualization* 2, 1 (March 2003), 16-30.

InterRing - MultiFocus Distortion



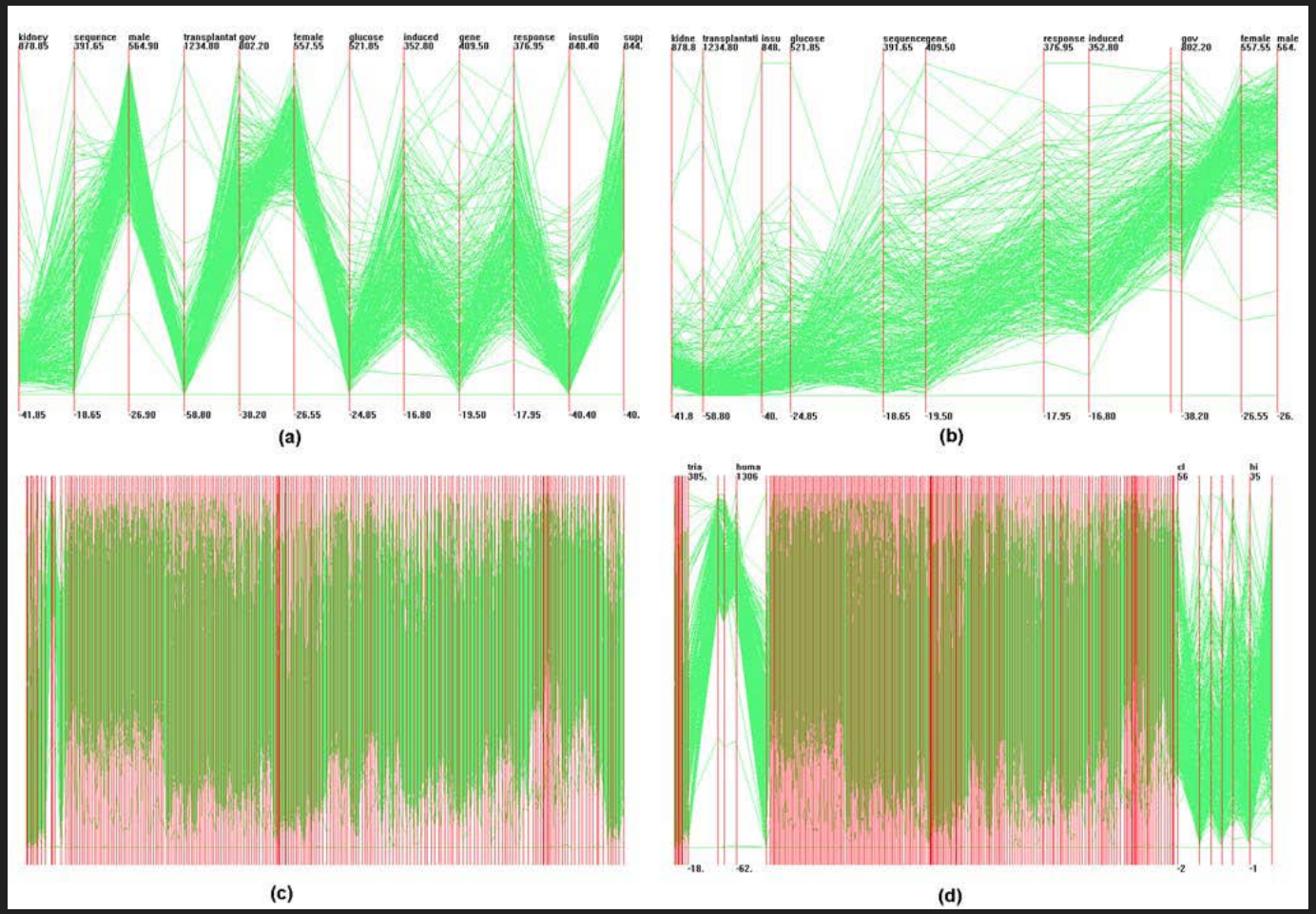
Filtering Interfaces - InterRing



Raw, order, distort and rollup (filter)

Image Credits: Jing Yang, Wei Peng, Matthew O. Ward and Elke A. Rundensteiner, "Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration of High Dimensional Datasets", IEEE Symposium on Information Visualization 2003 (InfoVis 2003), pp 105 - 112, October 2003.

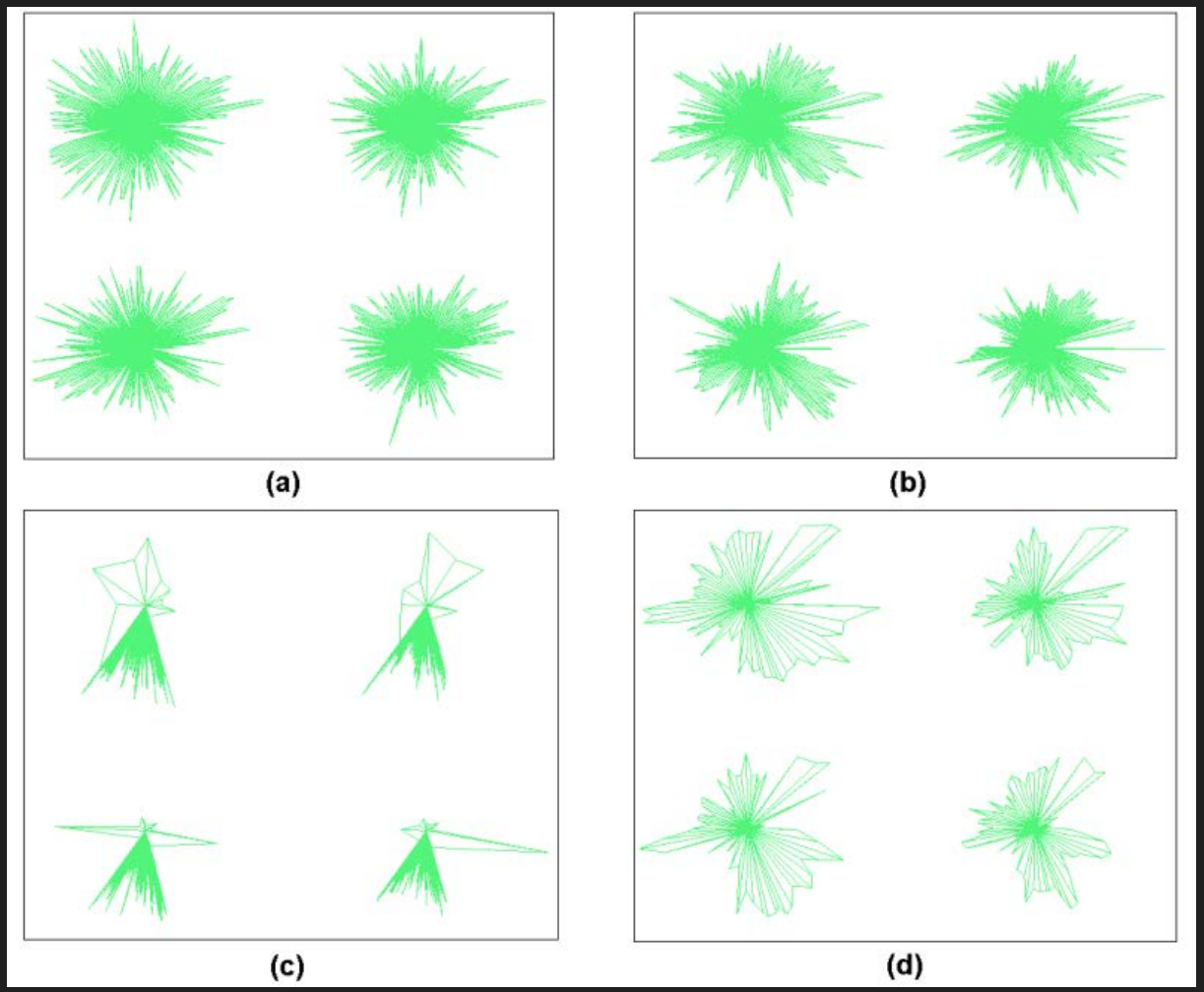
Filtering Interfaces – Parallel Coordinates



Raw, order/space, zoom and filter

Image Credits: Jing Yang, Wei Peng, Matthew O. Ward and Elke A. Rundensteiner, "Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration of High Dimensional Datasets", IEEE Symposium on Information Visualization 2003 (InfoVis 2003), pp 105 - 112, October 2003.

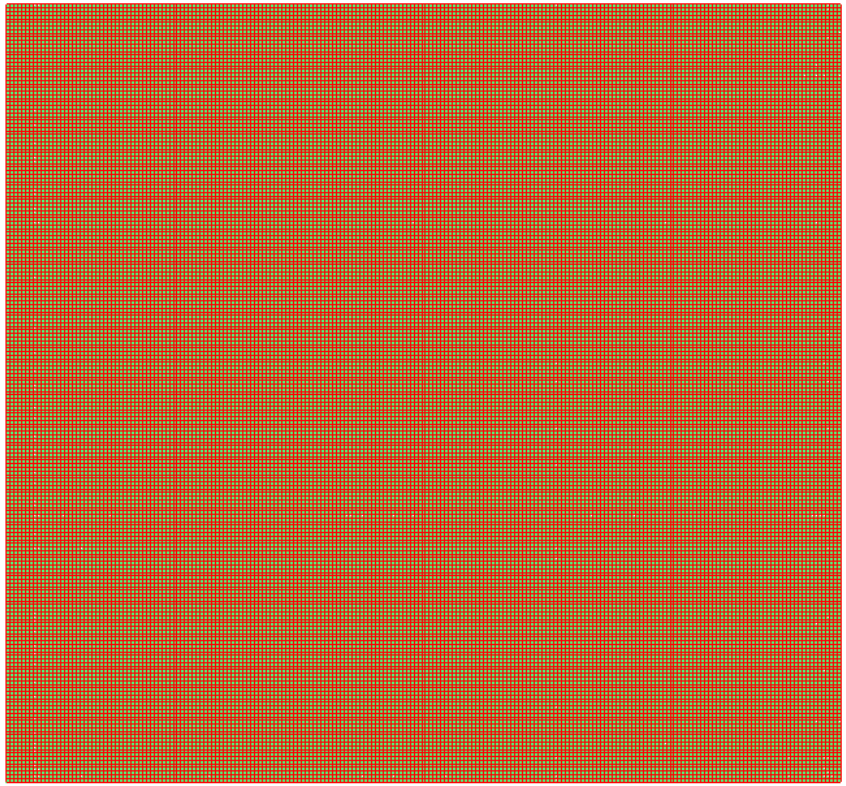
Filtering Interfaces - InterRing



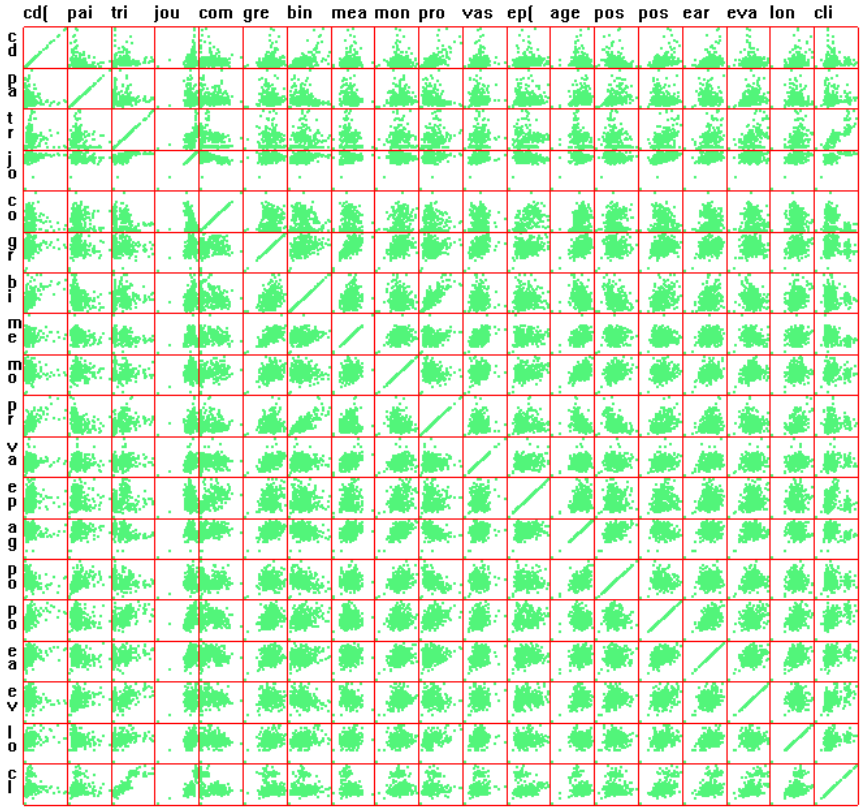
Raw, order/space, distort and filter

Image Credits: Jing Yang, Wei Peng, Matthew O. Ward and Elke A. Rundensteiner, "Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration of High Dimensional Datasets", IEEE Symposium on Information Visualization 2003 (InfoVis 2003), pp 105 - 112, October 2003.

Filtering Interfaces - InterRing



(a)



(b)

Raw and filter

Image Credits: Jing Yang, Wei Peng, Matthew O. Ward and Elke A. Rundensteiner, "Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration of High Dimensional Datasets", IEEE Symposium on Information Visualization 2003 (InfoVis 2003), pp 105 - 112, October 2003.

Polaris

- Multiscale Visualization Using Data Cubes, Chris Stolte, Diane Tang and Pat Hanrahan, Proc. InfoVis 2002.
- Stolte, C., Tang, D., and Hanrahan, P., Polaris: a system for query, analysis, and visualization of multidimensional databases, Commun. ACM 51, 11 (Nov. 2008), 75-84.

Large, Multi-Dimensional Databases

- Data acquisition not a problem anymore
- Extracting useful meaning from the data is a challenge
- “Path of exploration is unpredictable”
- Analysts want to be able to change the type of data and the visualization technique to examine the data
- Need to be able to visualize large subsets of data

Polaris

- An interactive exploration system that facilitates exploration of large, multi-dimensional relational databases
- Treat each attribute as a data cube (n-dimensional databases = n data-cubes)
- Polaris can facilitate multi-dimensional data exploration through a table-based display

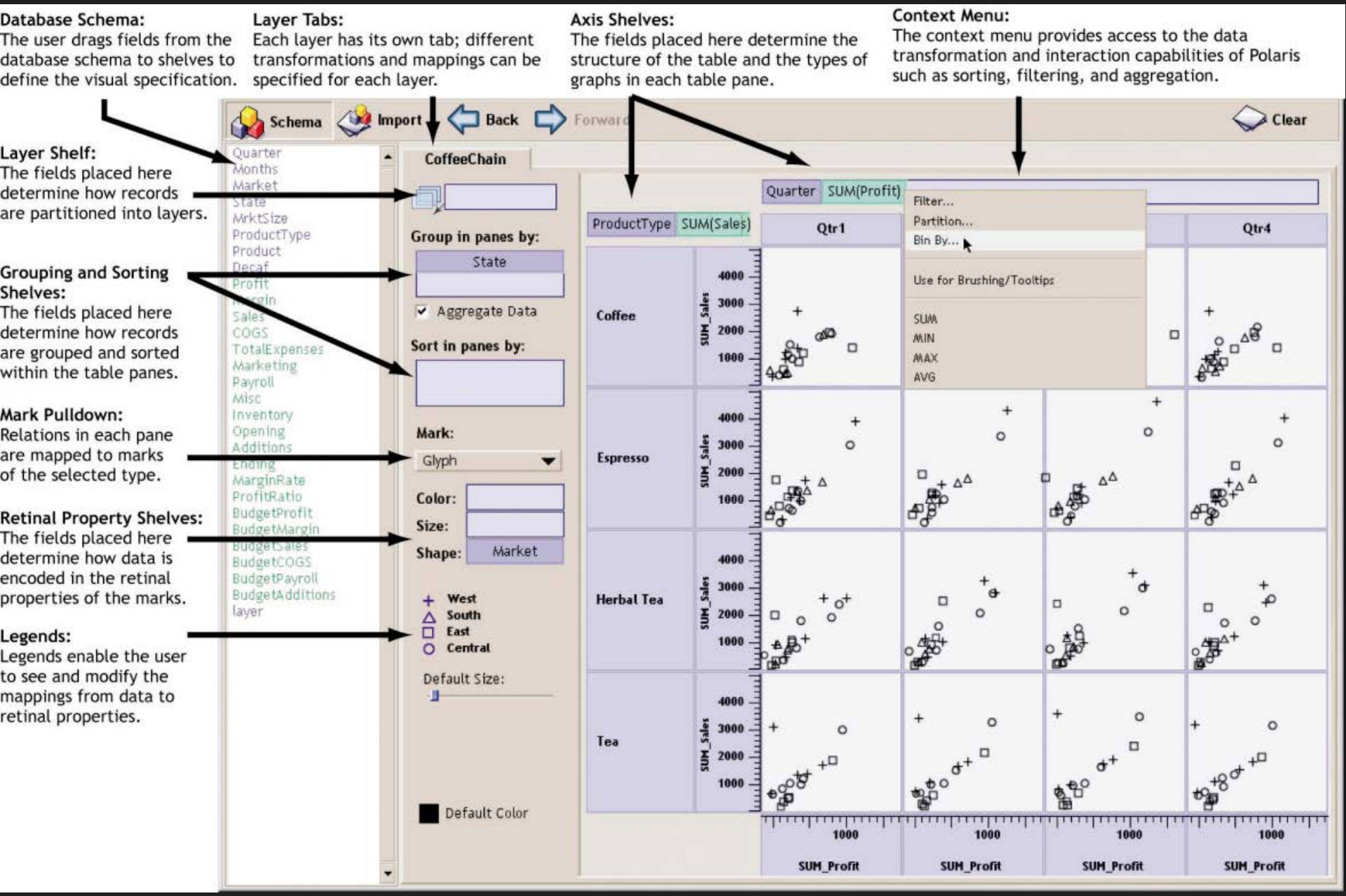
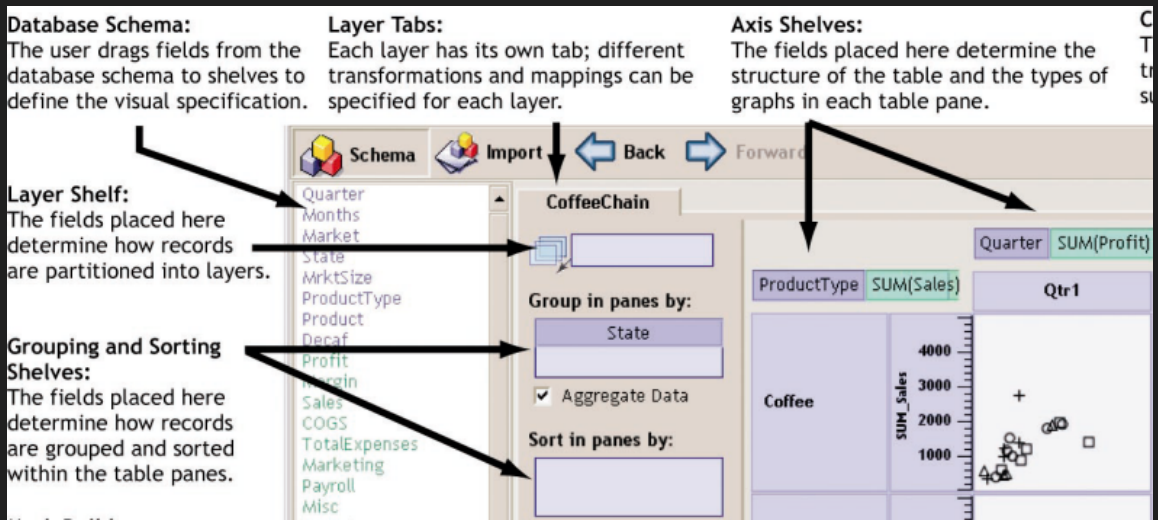


Image credits: Chris Stolte, Diane Tang, and Pat Hanrahan. 2008. Polaris: a system for query, analysis, and visualization of multidimensional databases. *Commun. ACM* 51, 11 (November 2008), 75-84.

Table Algebra

- Define a formal mechanism to specify table configurations
- Consists of three separate expressions
  - Two expressions define the x and y axes of the table
  - Third expression defines the z-axis (partitions the display into layers)



Operators

$$A \times B = \{a_1, \ldots, a_n\} \times \{b_1, \ldots, b_m\} = \{a_1 b_1, \ldots, a_1, b_m, a_2 b_1, \ldots, a_2 b_n, \ldots, a_n b_1, \ldots, a_n b_m\}$$

- Cross (x) operator: Cartesian product

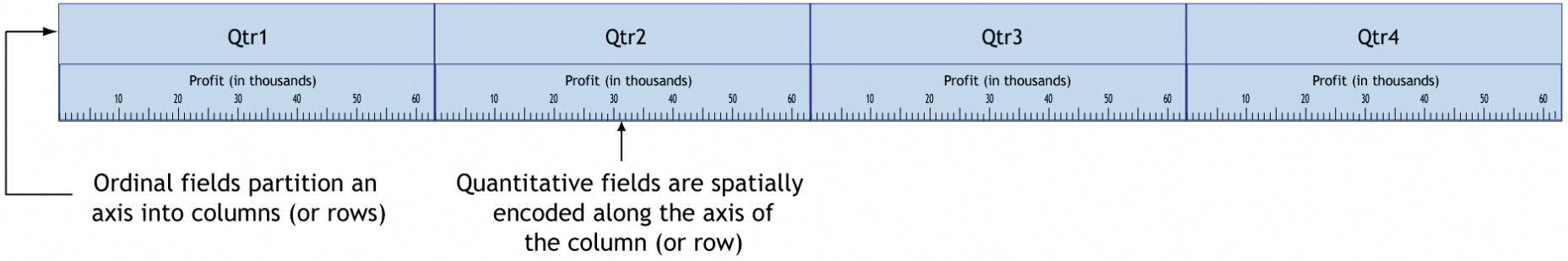
O = Quarter = {Qtr1, Qtr2, Qtr3, Qtr4} = Qtr1 + Qtr2 + Qtr3 + Qtr4:

Qtr1	Qtr2	Qtr3	Qtr4

O x O = Quarter x Product = {(Qtr1,Coffee), (Qtr1,Espresso), (Qtr1,Herbal Tea), (Qtr1, Tea), (Qtr2, Coffee) ... (Qtr4, Tea)}:

Qtr1				Qtr2				Qtr3				Qtr4			
Coffee	Espresso	Herbal Tea	Tea	Coffee	Espresso	Herbal Tea	Tea	Coffee	Espresso	Herbal Tea	Tea	Coffee	Espresso	Herbal Tea	Tea

O x Q = Quarter x Profit = {(Qtr1,Profit), (Qtr2, Profit), (Qtr3, Profit), (Qtr4, Profit)}:



$$A/B = \{a_i b_j \mid \exists r \in R \text{ st } A(r) = a_i \ \& \ B(r) = b_i\}.$$

Operators

- Nest (/) operator: A/B = B within A

O = Quarter = {Qtr1, Qtr2, Qtr3, Qtr4} = Qtr1 + Qtr2 + Qtr3 + Qtr4:

O/O = Quarter / Month = {(Qtr1,Jan), (Qtr1,Feb), (Qtr1,Mar), (Qtr2, Apr), (Qtr2, May) ... (Qtr4, Dec)}:

The set entry (Qtr4,Nov) corresponds to this column

$$A + B = \{a_1, \ldots, a_n\} + \{b_1, \ldots, b_m\}$$
$$= \{a_1, \ldots, a_n, b_1, \ldots, b_m\}$$

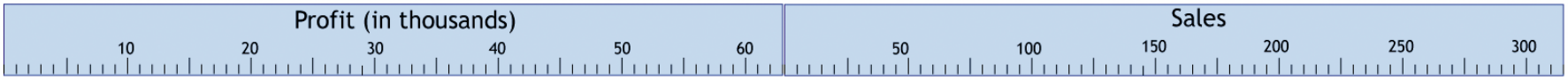
Operators

- Concatenation (+) operator

O = Quarter = {Qtr1, Qtr2, Qtr3, Qtr4} = Qtr1 + Qtr2 + Qtr3 + Qtr4:

O + O = Quarter + Product = {Qtr1, Qtr2, Qtr3, Qtr4, Coffee, Espresso, Herbal Tea, Tea}:

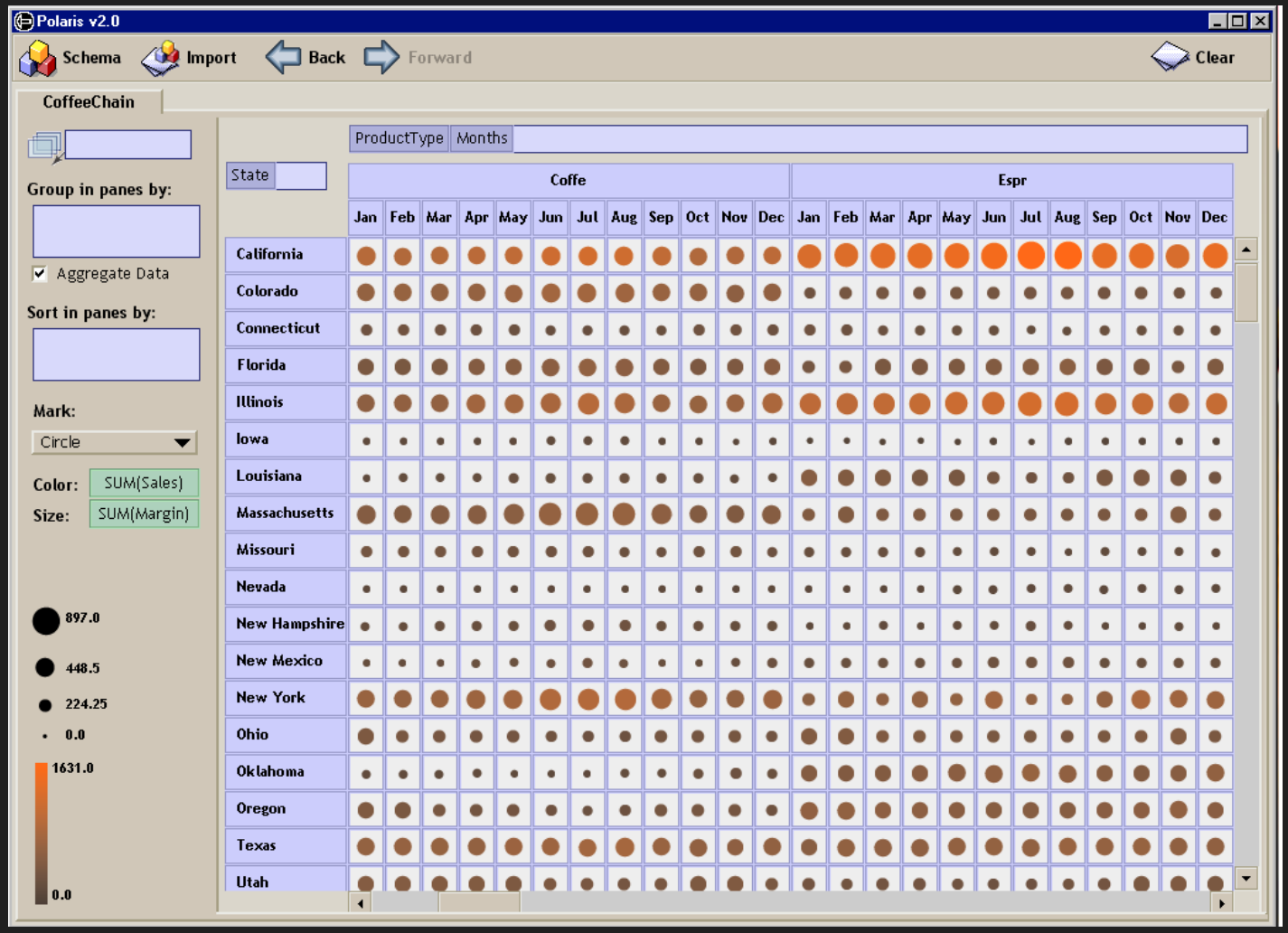
Q + Q = Profit + Sales = {Profit, Sales}:



Space of Graphics

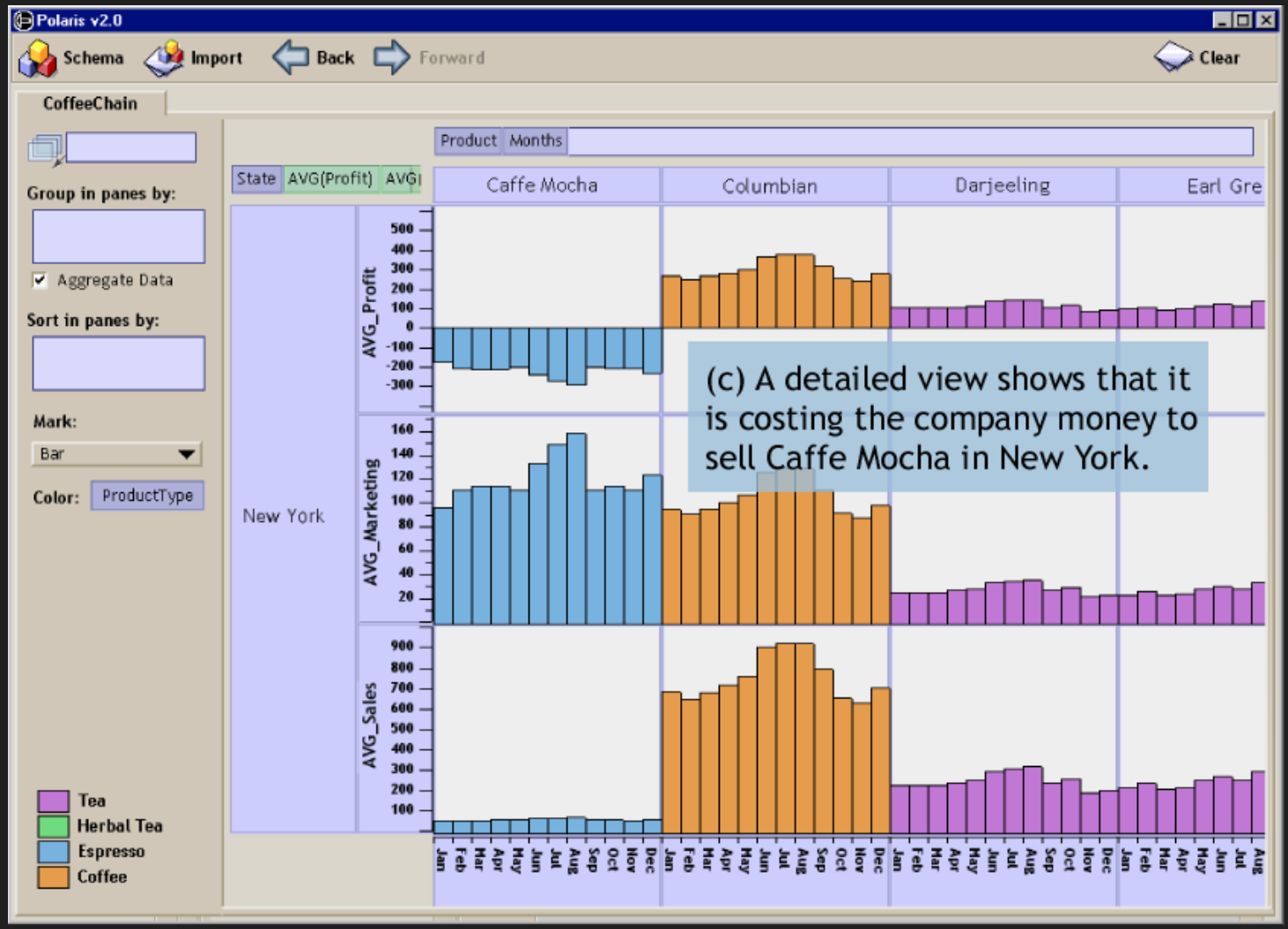
- Structured into three families
  - Ordinal-Ordinal
  - Ordinal-Quantitative
  - Quantitative - Quantitative

Ordinal-Ordinal



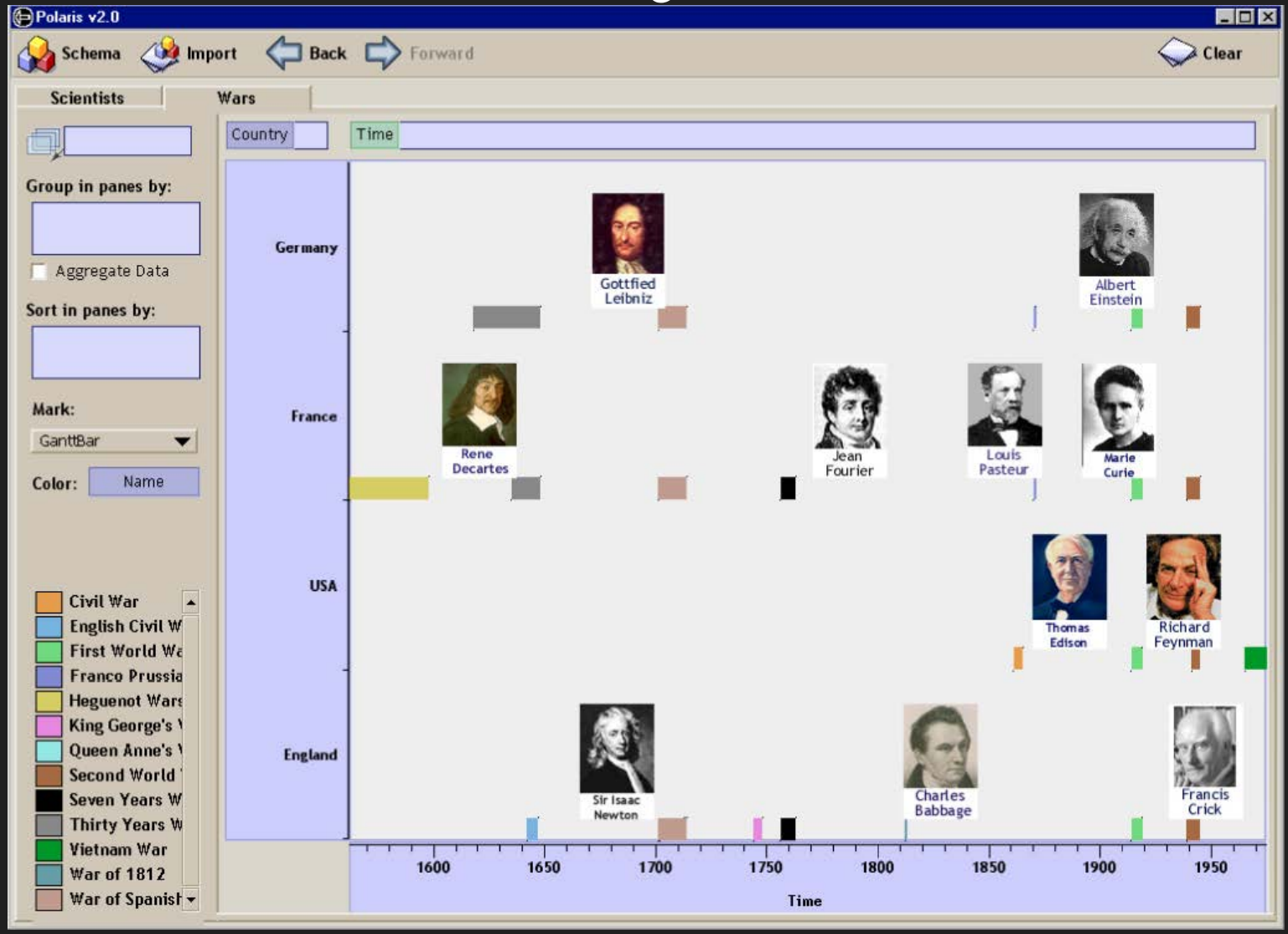
Sales and margins vs product type, month and state for the items sold

Ordinal - Quantitative



Matrix of bar charts is used to study independent variables – product and month

Ordinal - Quantitative

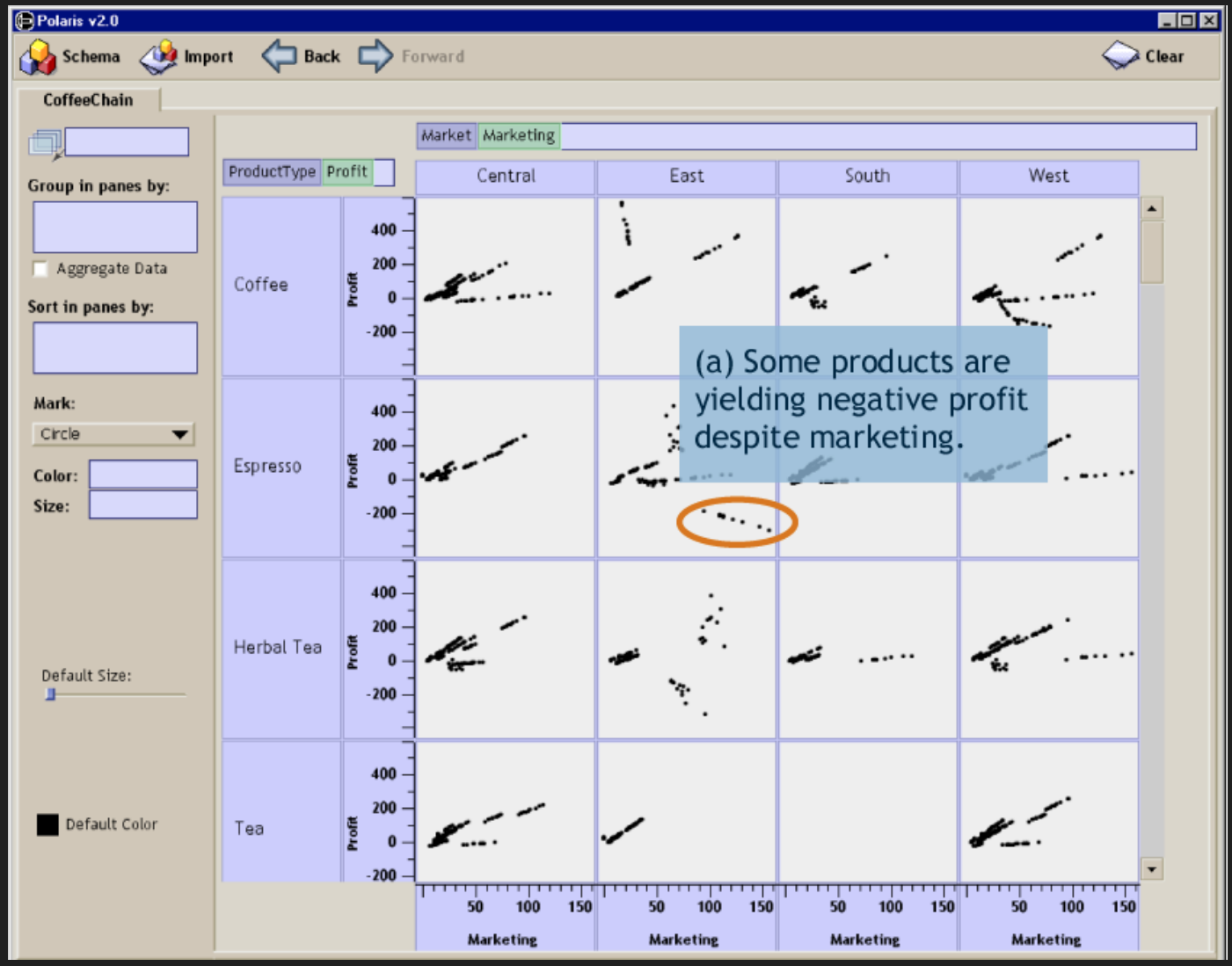


Major wars over the last five hundred years and additional layer of major scientists (country and date of birth)

Ordinal - Quantitative

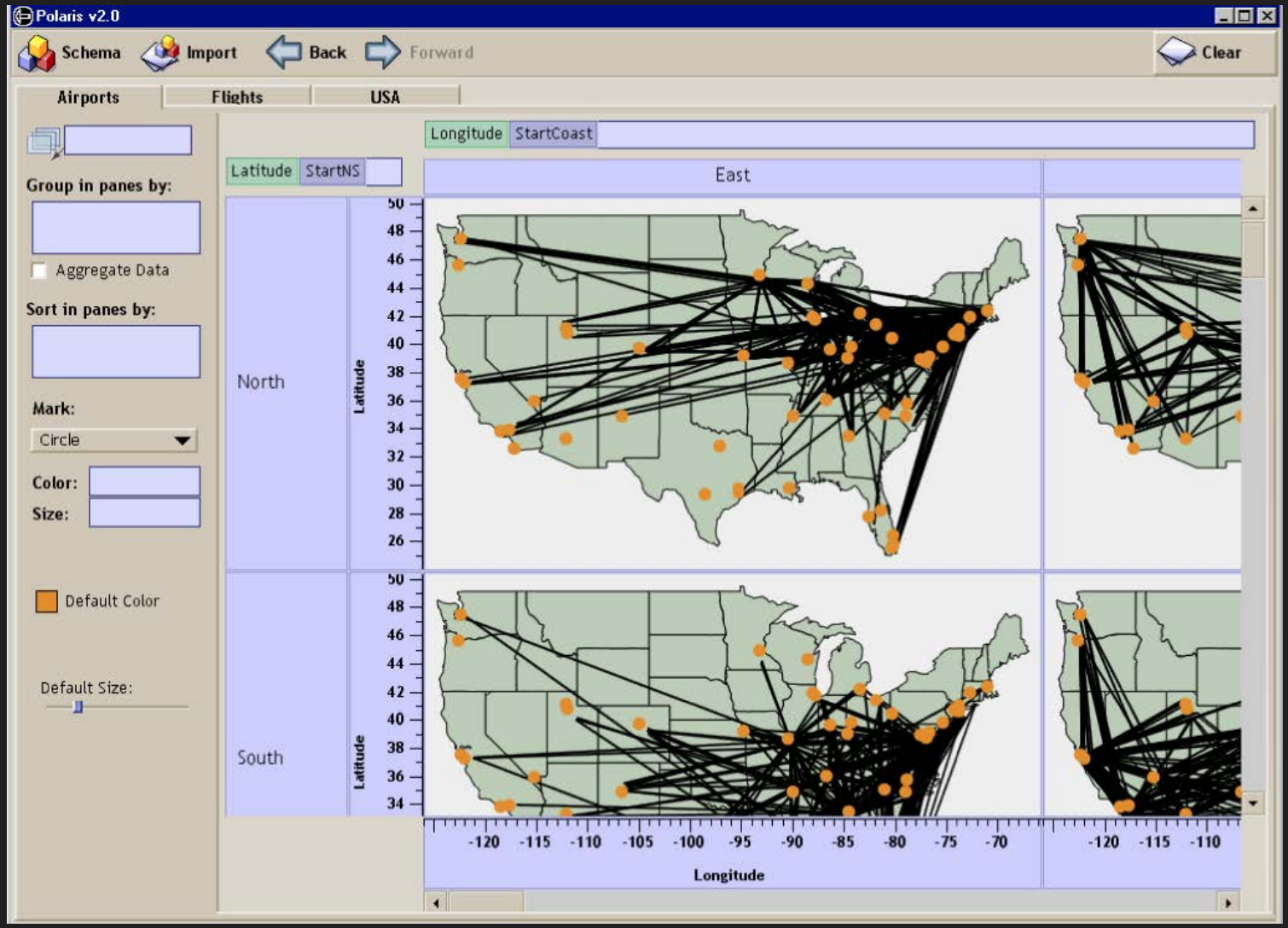
Thread scheduling and locking activity on a CPU within a multiprocessor computer

Quantitative - Quantitative



Number of attributes of different products sold by a coffee chain

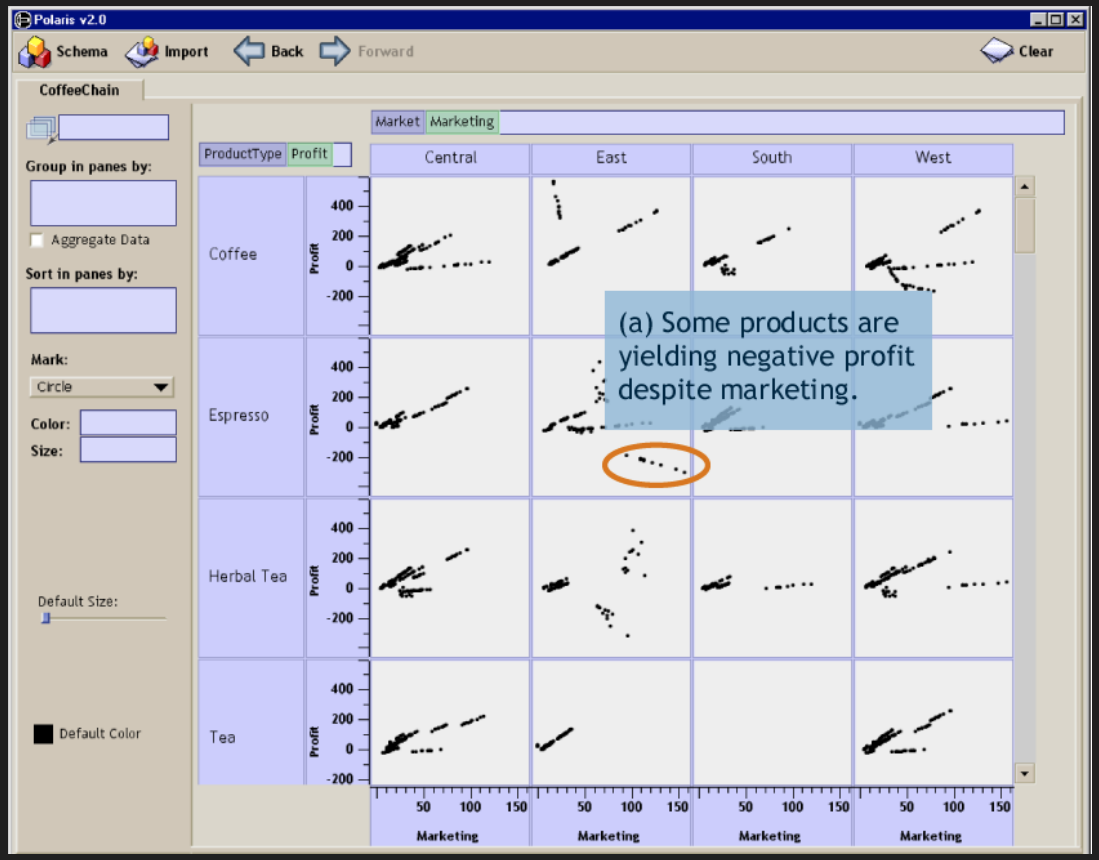
Quantitative - Quantitative



Flight scheduling varies with the region of the country the flight originated in

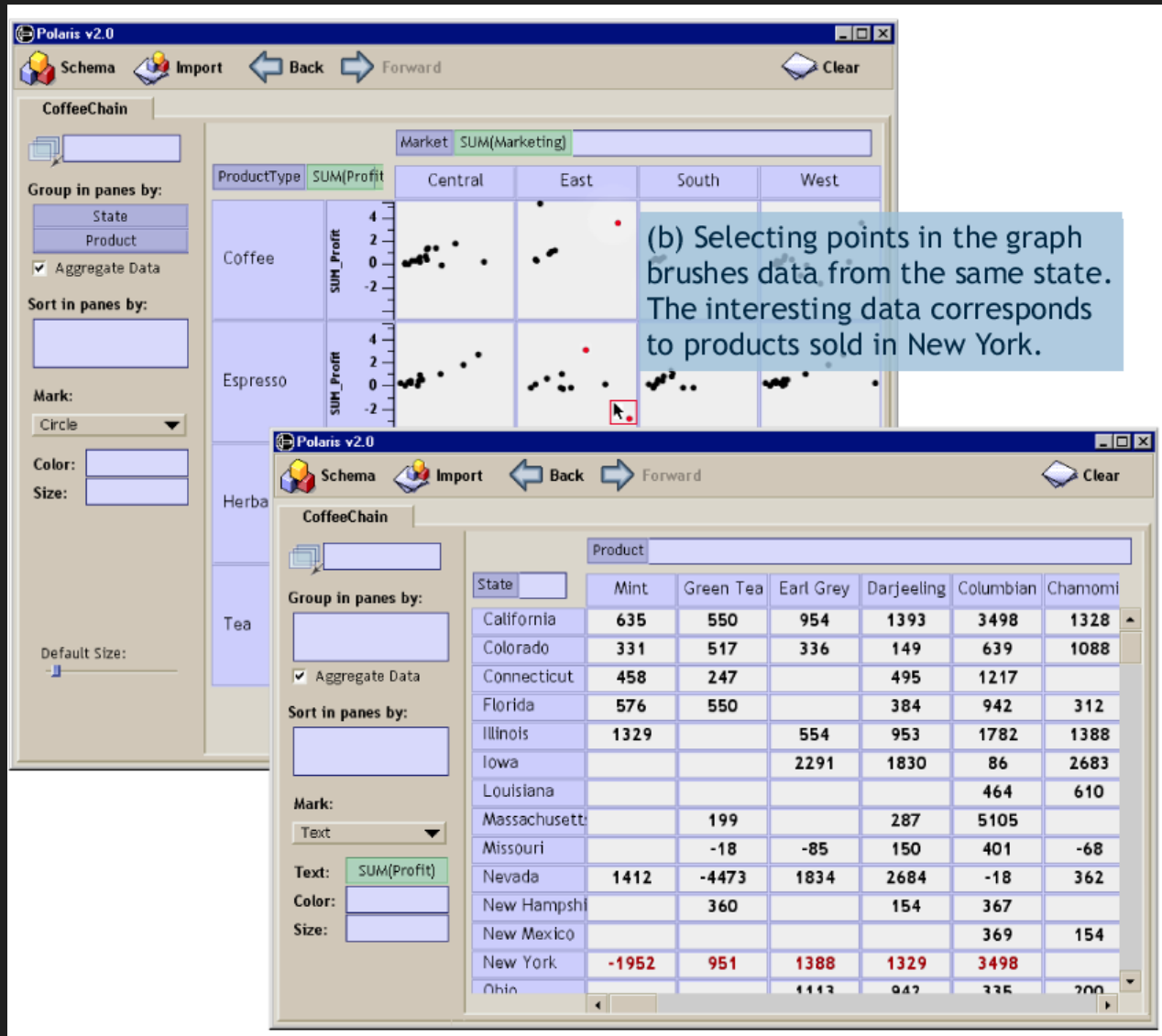
Scenarios

- CFO told to cut expenses – Scatterplots of marketing costs and profit categorized by product type and market



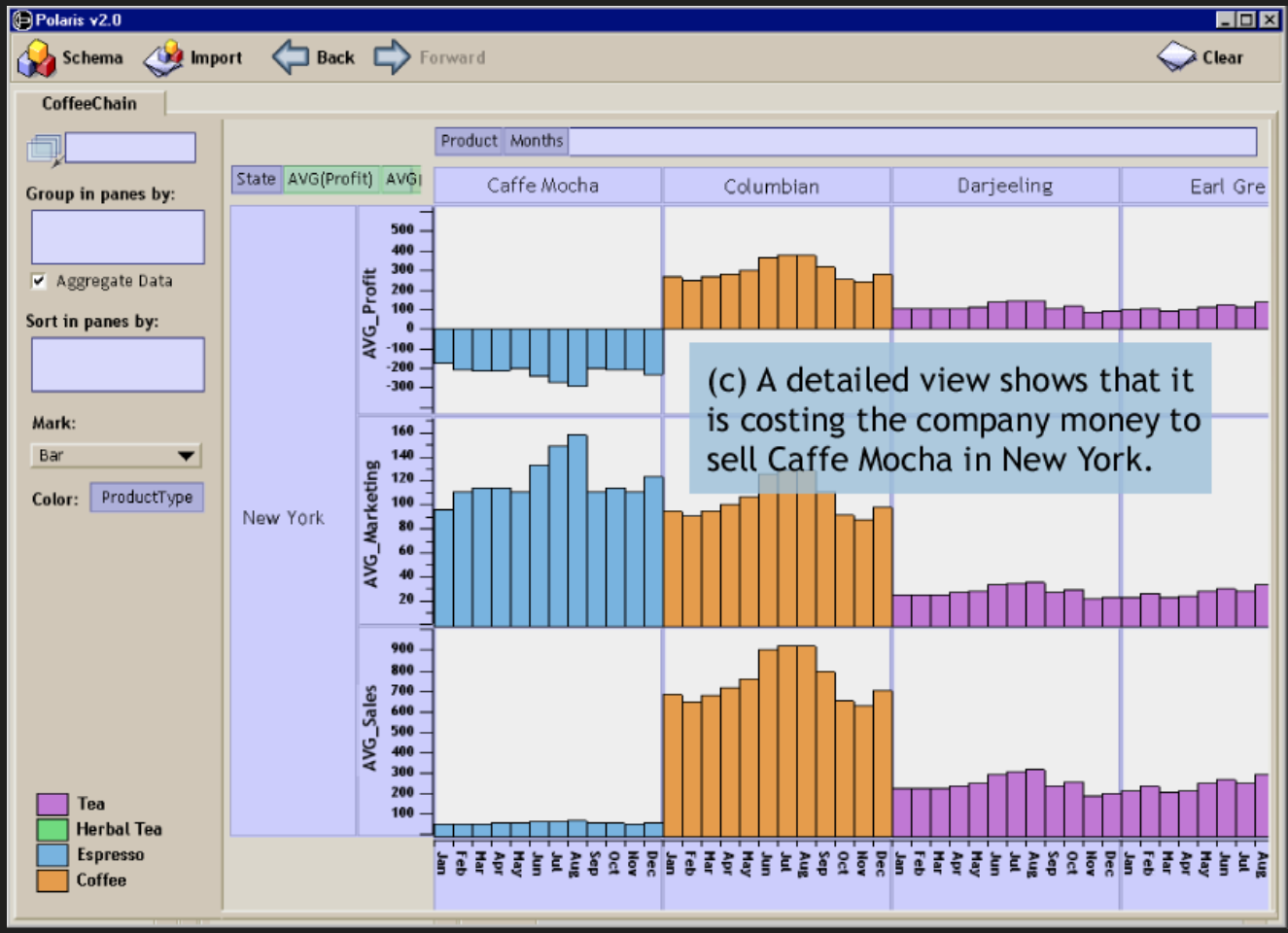
Scenarios

- Further investigate by creating two linked displays
- Scatter plot and text table



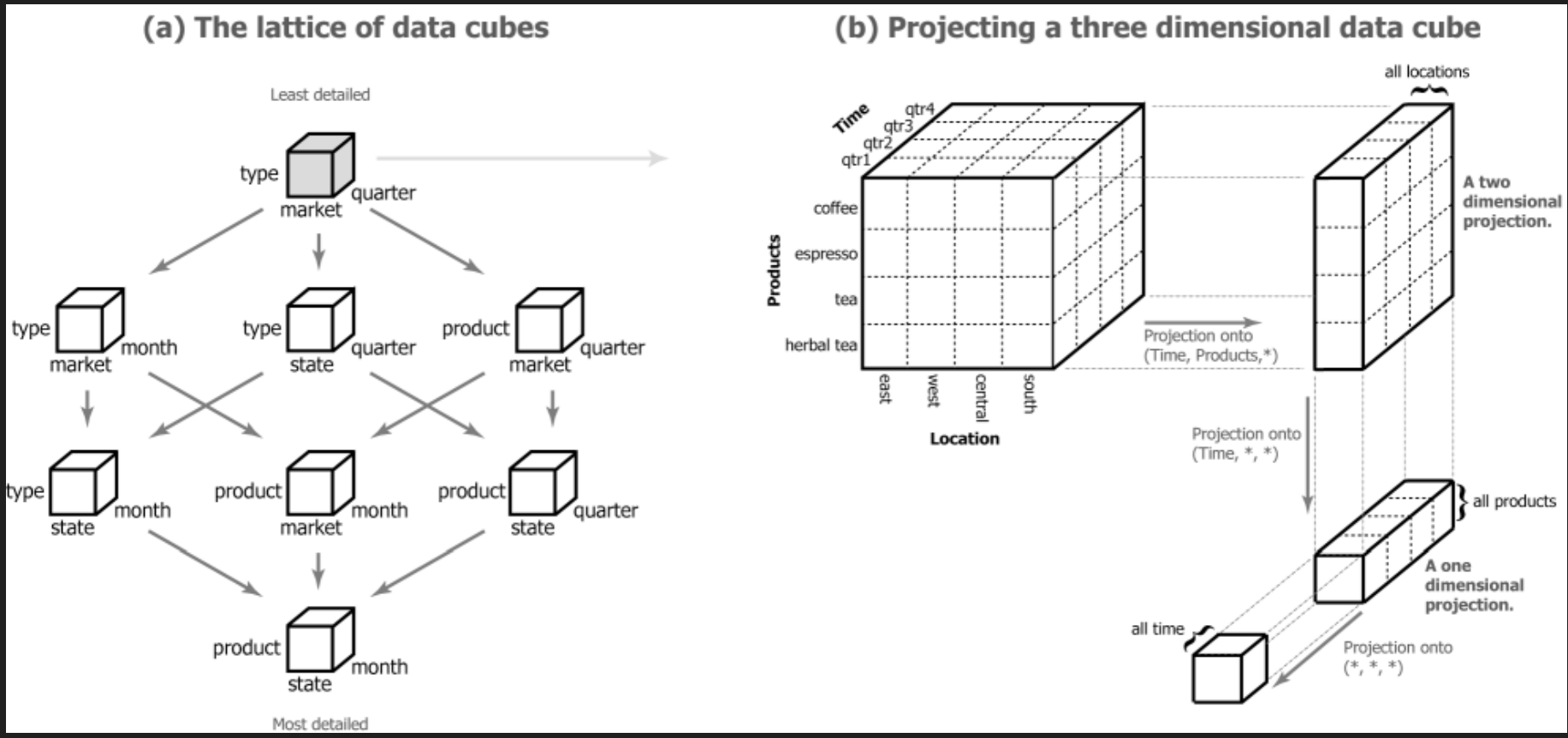
Scenarios

- Third display to visualize data for each product sold in New York
- Café Mocha



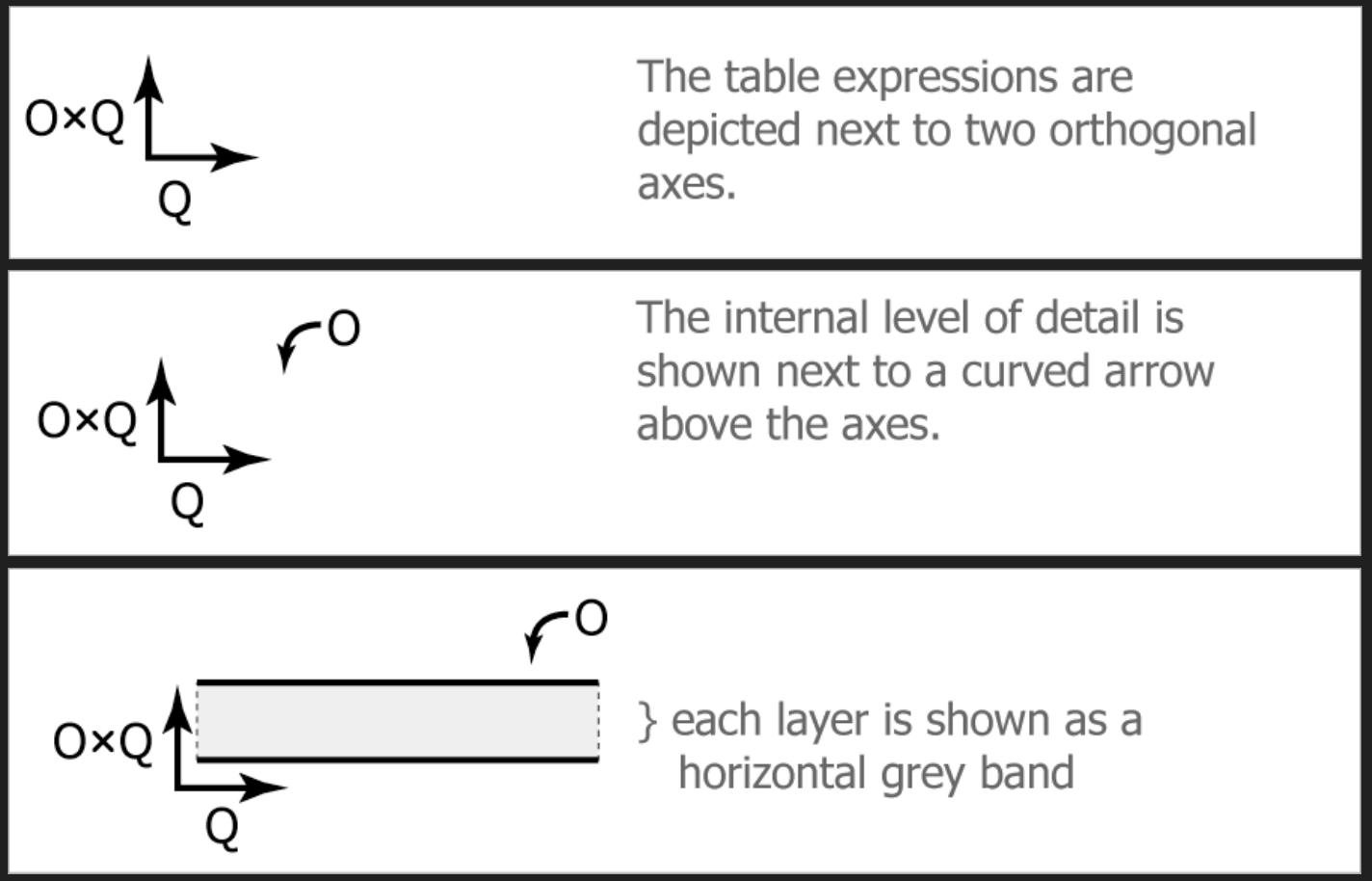
Multiscale Visualizations

- Data Abstraction through Data Cubes



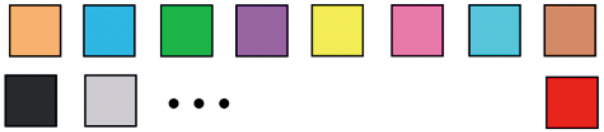
Multiscale Visualizations

- Visual Abstraction: Polaris



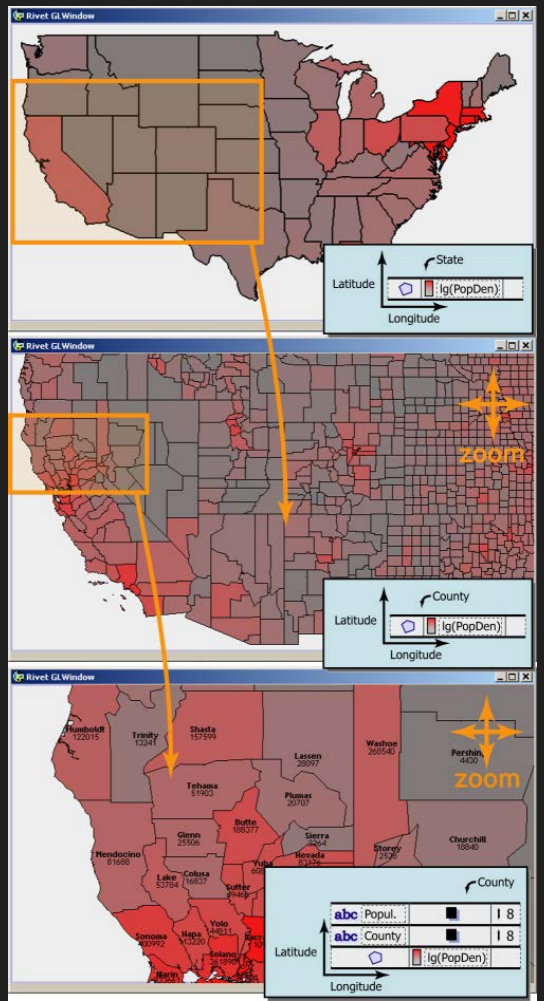
Multiscale Visualizations

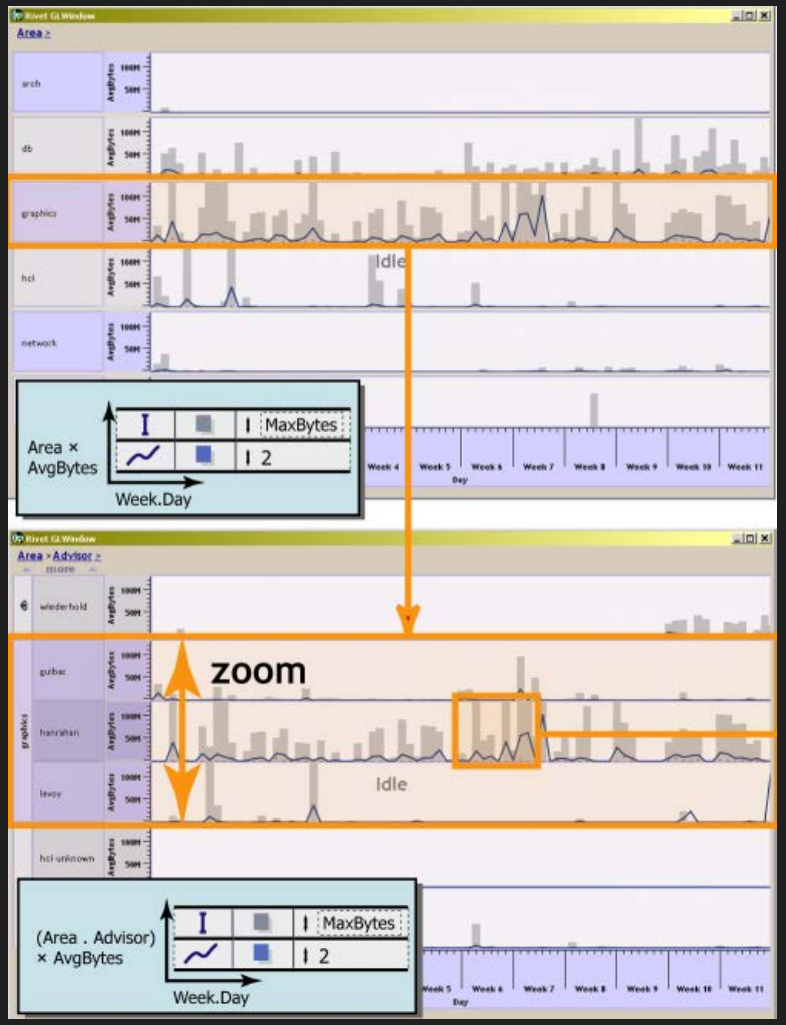
- Visual Abstraction: Polaris

property	marks	ordinal/nominal mapping	quantitative mapping
shape	glyph		
size	rectangle, circle, glyph, text		
orientation	rectangle, line, text		
color	rectangle, circle, line, glyph, y-bar, x-bar, text, gantt bar		

Multiscale Visualizations

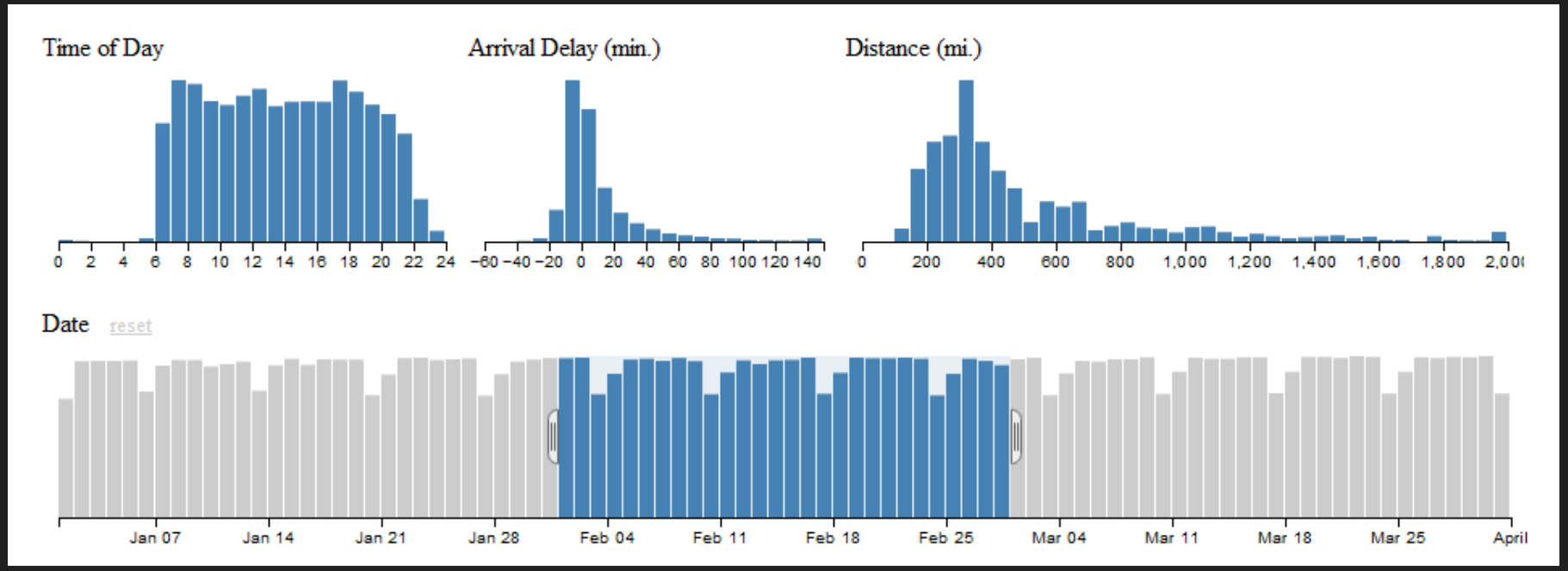
- Zoom Graphs:



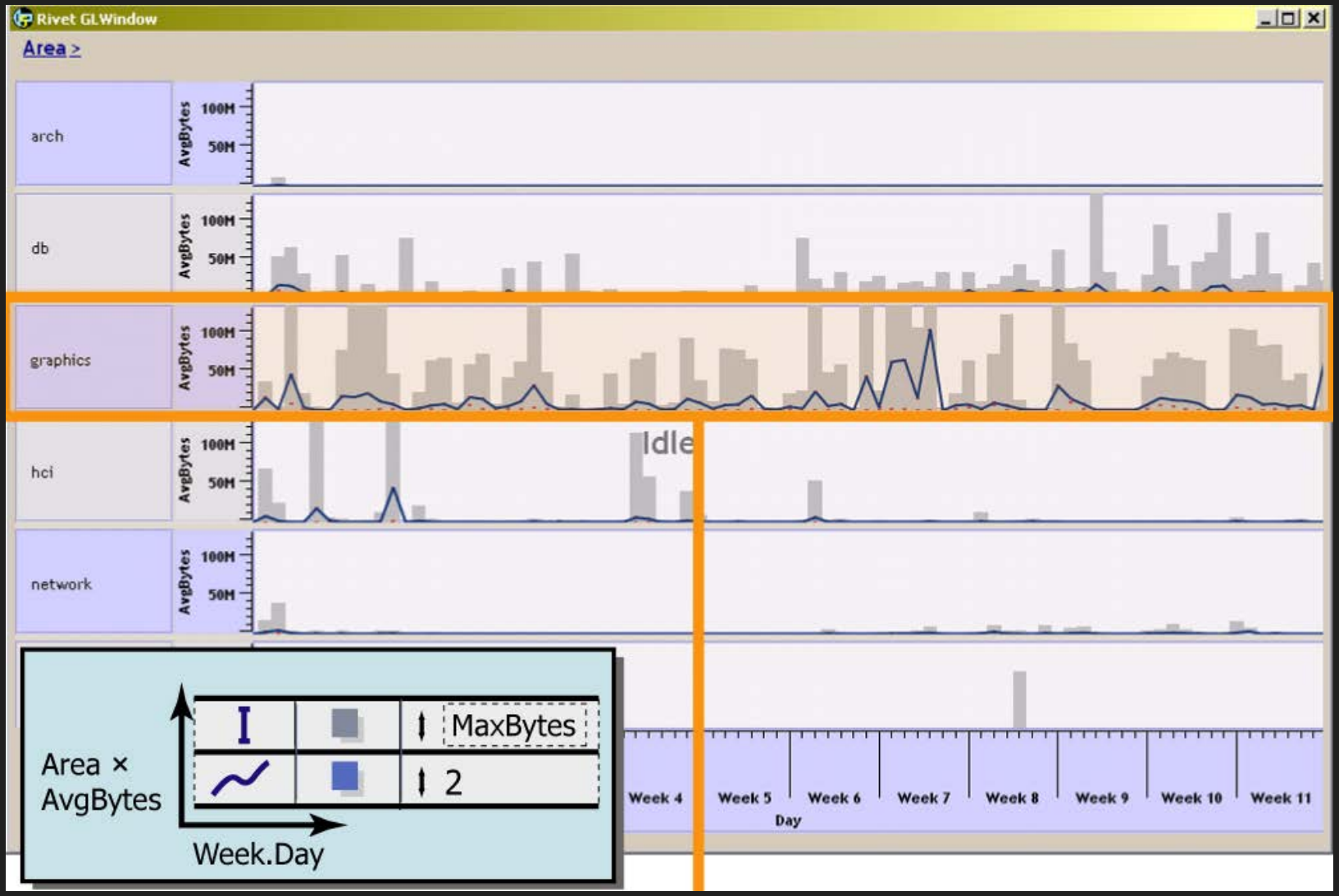


Fast Multidimensional Filtering for Coordinated Views

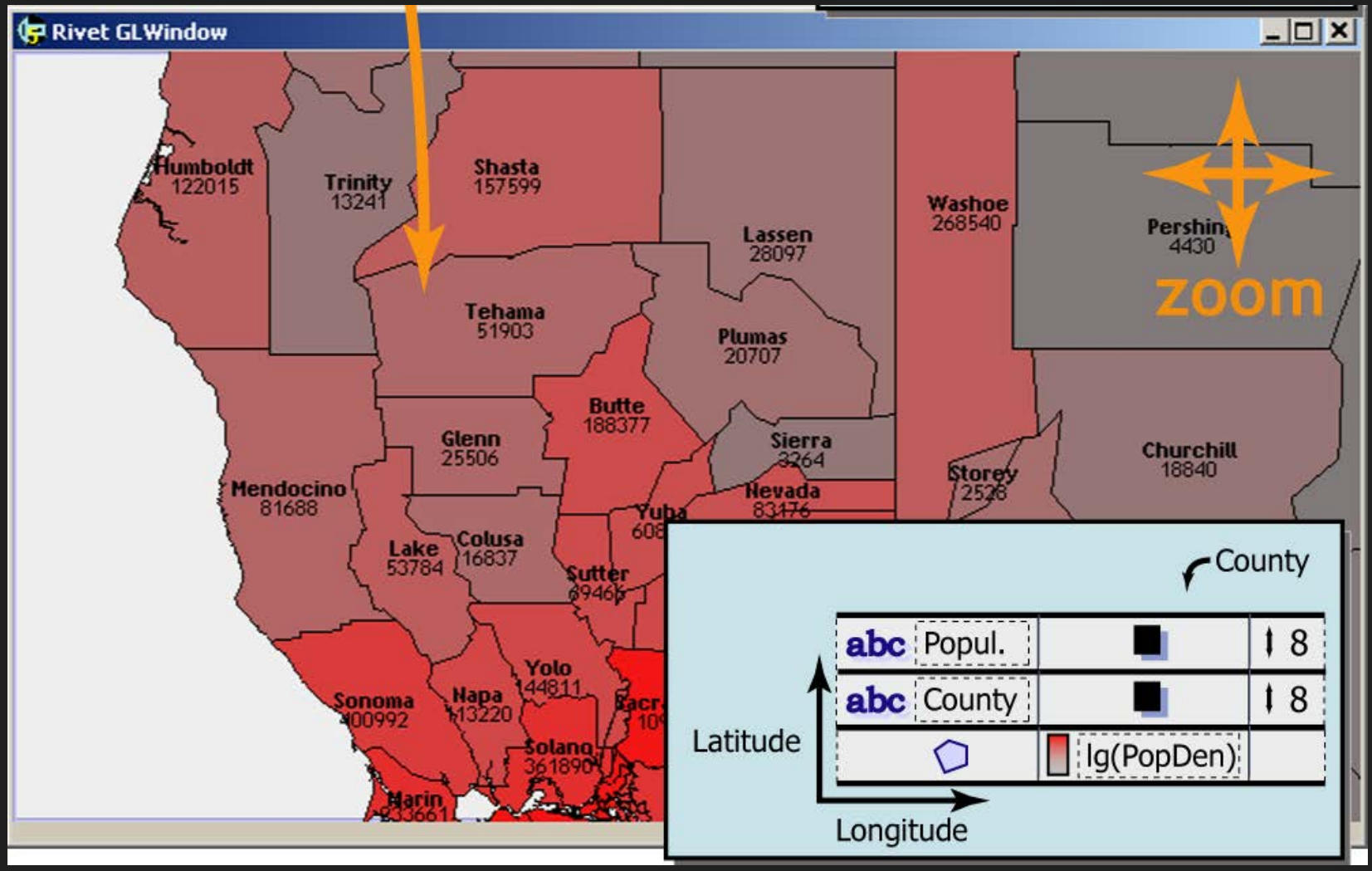
- http://square.github.io/crossfilter/



Multiscale Design Patterns: Chart Stacks

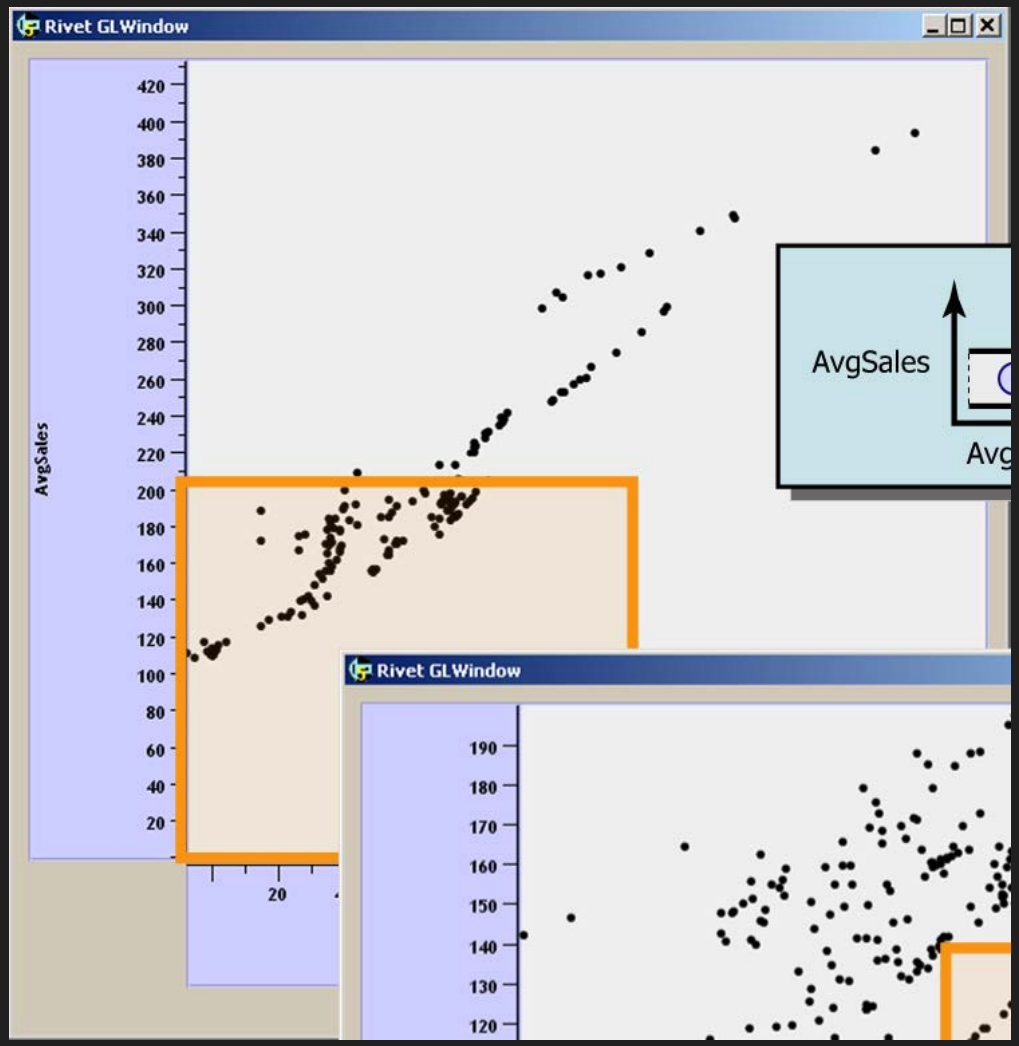


Multiscale Design Patterns: Thematic Maps

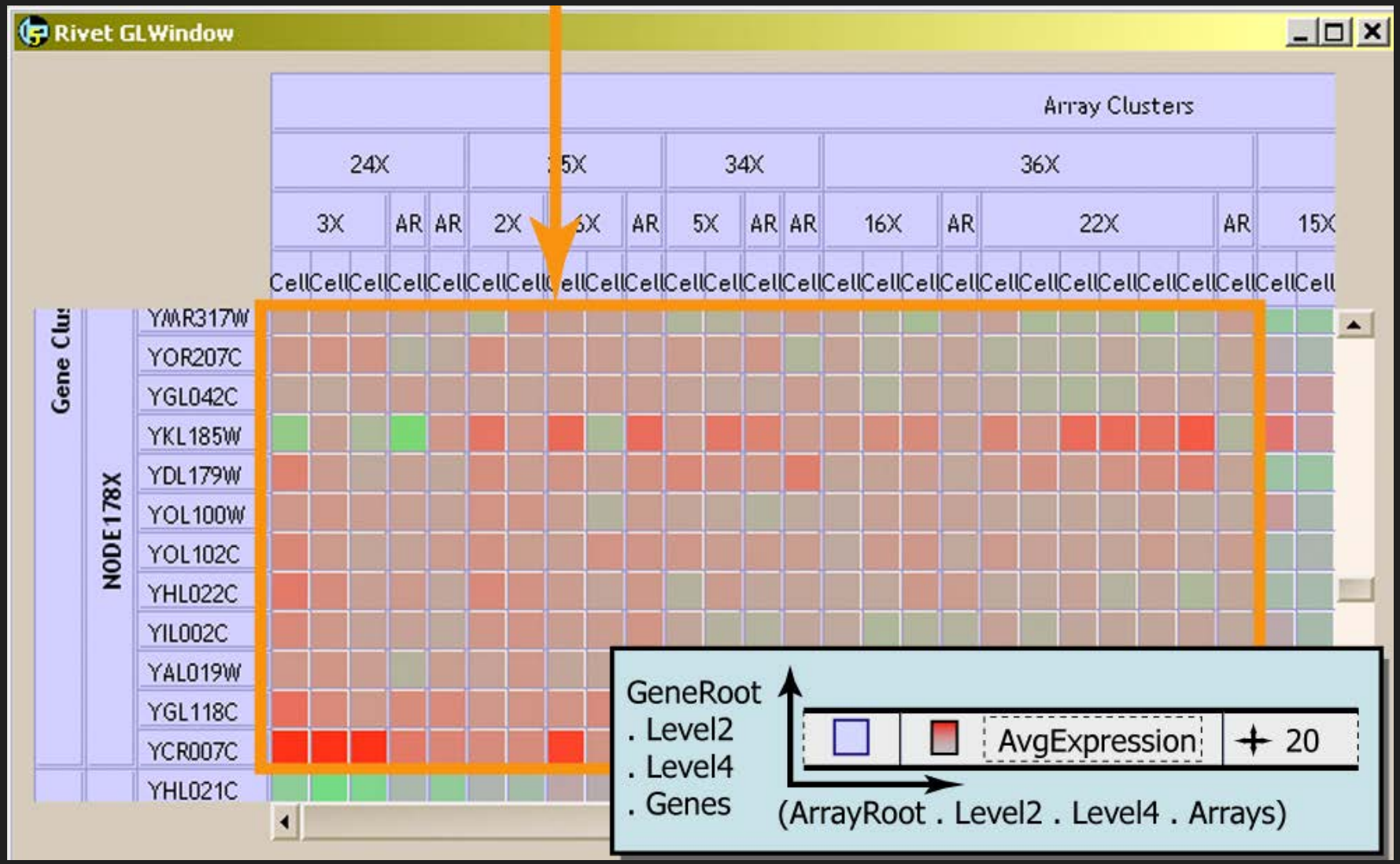


Population density

MDP: Dependent Quantitative-Dependent Quantitative Scatterplots



Multiscale Design Patterns: Matrices



Evaluation of Multivariate Trend Visualization

- Mark A. Livingston, Jonathan W. Decker: Evaluation of Trend Localization with Multi-Variate Visualizations. IEEE Trans. Vis. Comput. Graph. 17(12): 2053-2062 (2011)
- Evaluates the multivariate visualization techniques