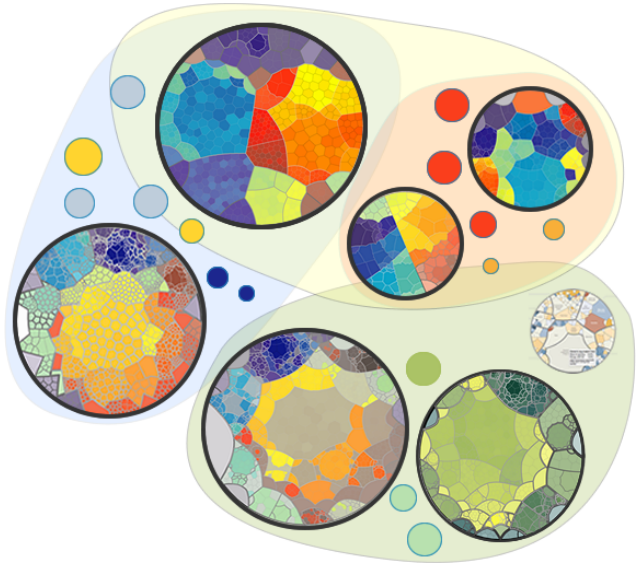


DICON: Visual Analysis on Multidimensional Clusters

design, concepts & scenarios

Nan Cao@HKUST

July 2010

Introduction

- Background
  - Clustering analysis have been proved as an useful technique and applied to many domains in data mining.
- Motivation
  - Clustering is different from classification, ***no*** existing category with clear semantic meanings are provided. The results of multidimensional clustering are usually difficult for users to understand. They are also difficult for analyzer to explain.
  - Data patterns hidden in n – dimensions (n >= 3) are usually difficult to find. Even the 2 dimensional patterns are hard to find when too many dimensions exist (Which two need to be selected ?).
  - Automatic clustering will not necessary generate 100% correct clusters. Incorrect clusters may misleading users' decision making.
- Problems to address
  - How to understand the clustering result? (Why entities are clustered and how are they related?)
  - How to find patterns hidden in the high dimensional clusters ?
  - How to estimate the qualities of the clusters in an intuitive way?
  - How to refine the cluster results which has a poor quality?

Topic: Cluster Analysis
Link: http://en.wikipedia.org/wiki/Cluster_analysis

Contents [hide]

Research Scope

- Problem
  - How to visualize multidimensional cluster results in an intuitive way for explanation, pattern detection and evaluation ?
- Goal
  - Understand the meaning of the multidimensional clustering results
  - Find high dimensional patterns
  - Evaluate and adjust the automatic analysis result to generate high quality clusters.

Is this problem a new problem?

- Compared with traditional ICON
  - No one use an ICON design to visualize cluster information
- Compared with PCP
  - Encode more statistic information provide more intuitive visual clues
- Compared with Scatter Plot & Projection
  - A novel enhancement of the traditional scatter plot visualization which provides more semantic meanings that facilitates:
    - Cluster Explanation
    - Cluster Comparison
    - Cluster Estimation

Related Work

- Multidimensional Visualization
- Visualization Multidimensional Clusters
  - Scatter Plot, Heatmap, Parallel Coordinates, Treemap
- ICON based Multivariate Visualizations
  - Glyph based ICON
  - Chernoff faces
  - Stick figure
  - Pixel based ICONs
- Interactive Cluster Exploration and Analysis
  - Notrix, HCE, Rolling the Dice, Scatter Plot Matrix
- ***No existing*** work fulfills all our problems

Contributions

- Seamlessly combine the scatter plot and a novel ICON design to explain the statistical information of multivariate data and cluster analysis result in a more intuitive way.

- provide an more intuitive way to understand the reason why entities are clustered together from multiple clusters compared to PCP and other existing technologies.

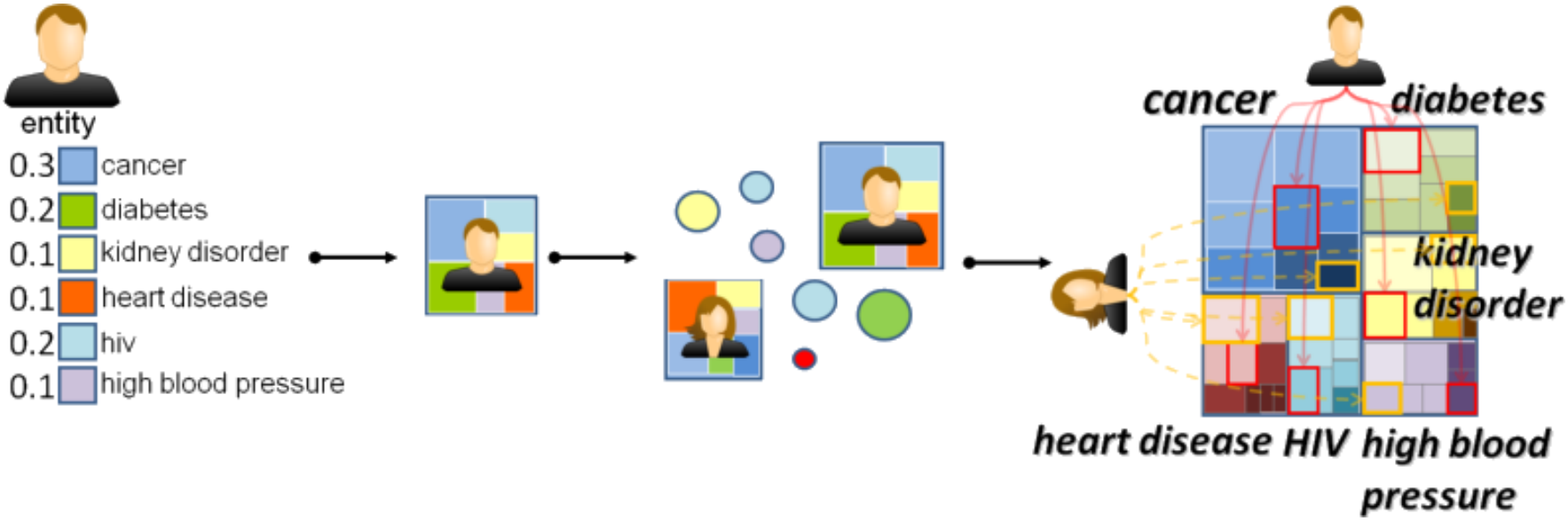
- Provide a new approach for high dimensional cluster comparison and cluster pattern detection
  - Cluster similarities
  - Feature co-occurrences within a specified cluster
  - Cluster Quality Patterns : What is a good / bad cluster ?
  - Other visual clues: when using different dimensions to encode size and color respectively

- Provided novel interactions for cluster manipulation to give valuable feedback to underlying analysis.
  - Merge & Split
  - Categorical Grouping
  - Cluster Refinement

Design Principles

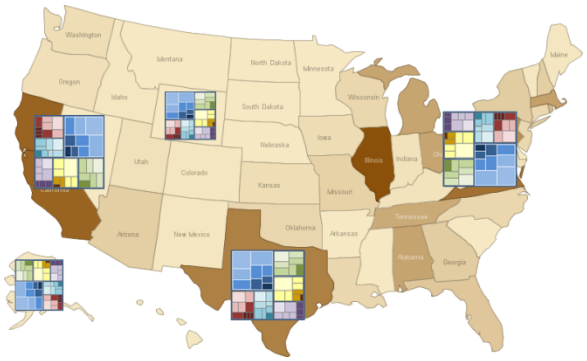
- Multi-Granularity
  - How features are distributed in an entity or a cluster. How clusters distributive in the information
  - Rationale : provide distribution details at different level facilities understanding the data from easy to hard.
- Uniform and Consistent
  - Exactly the same encoding for features, entities and clusters.
  - Rationale : minimize the learning curve
- Multivariate
  - Visualize the variances of the cluster entities and.
  - Rationale : users are more easily to find a visual pattern if we provide some visual clues and scent. Without multivariate, it is impossible to provide visual clues.
- Organized and Stabilized
  - The features, entities and clusters should be well organized together in a stabilized way.
  - Rationale : facilitates visual explanation and comparison. Users are more easily to remember an figure with some innate **disciplines.**
- Rich Interactive

Our Methods

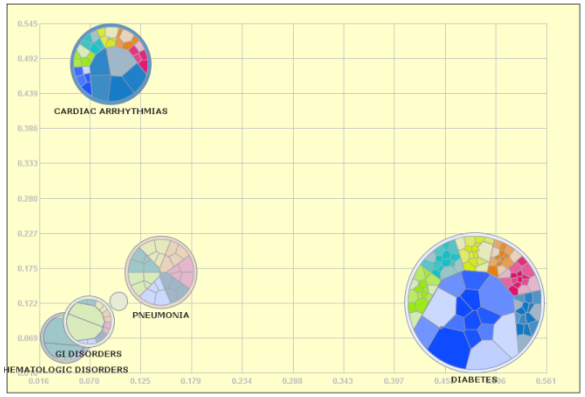


- Color Hue : Feature category
- Color Brightness : Feature SD / Variance
- Size : the feature value / number for entities in the cluster
- Distance: Similarities

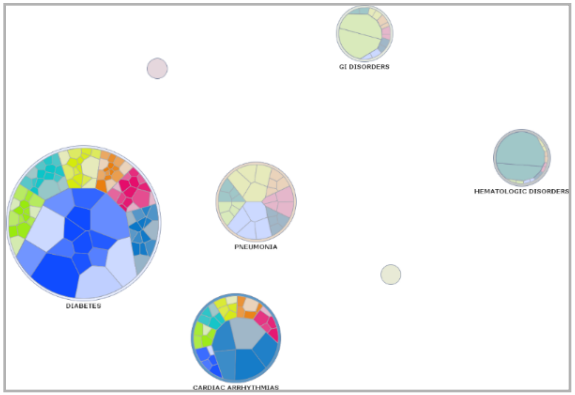
Global Layouts



Geometry Based



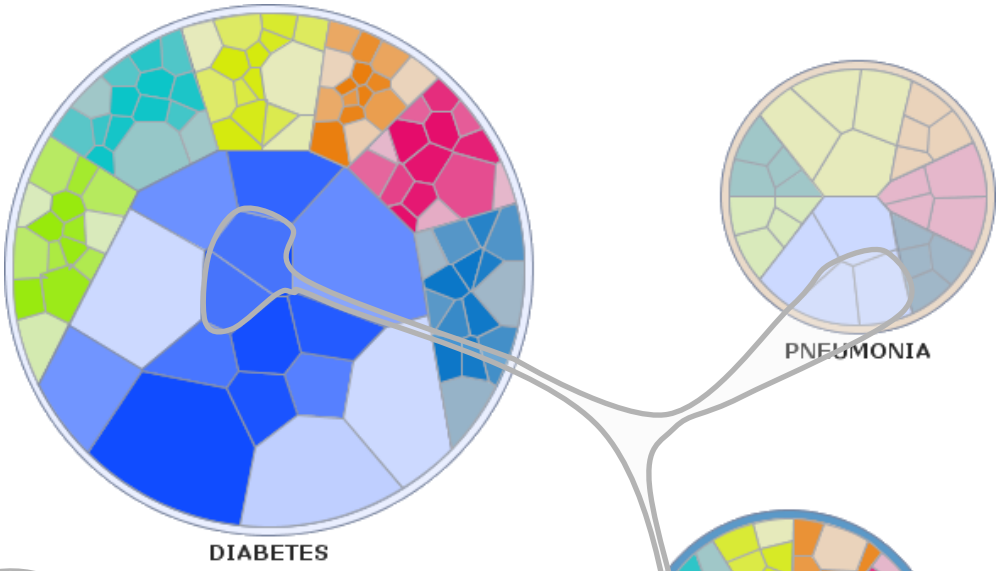
Scatter Plot



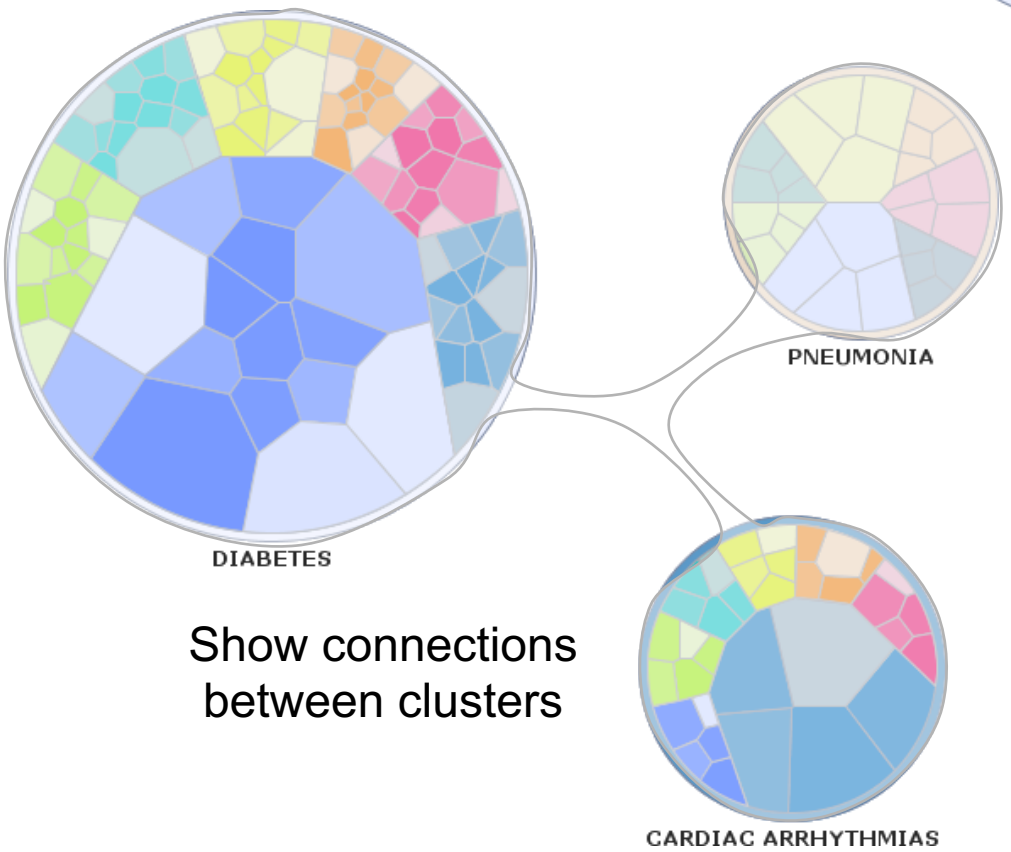
MDS

- Multi-layout Transformation

Visual Clues



Show Connection
between feature sets

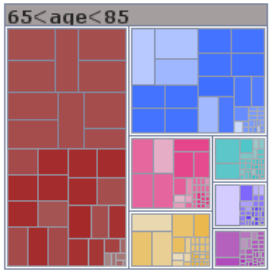


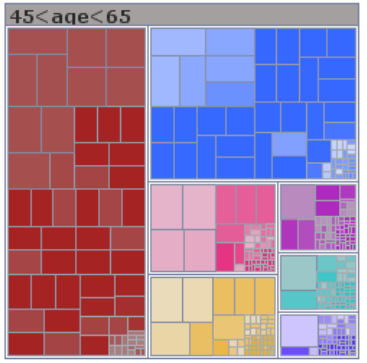
Show connections
between clusters

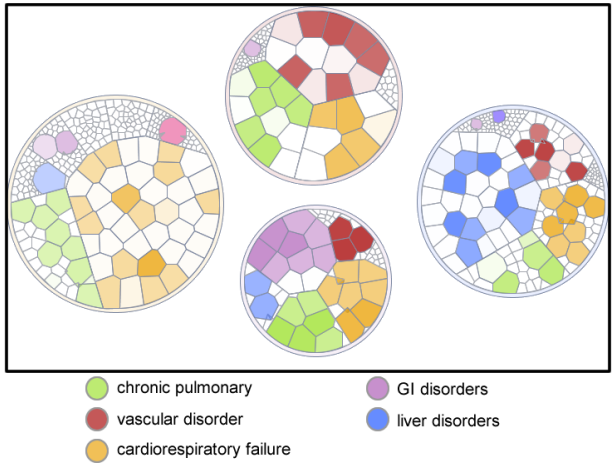
Relational Clues

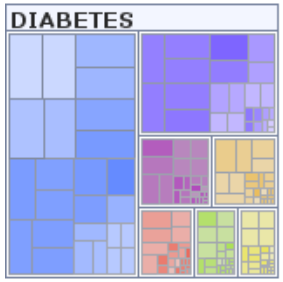
Visual Clues

- Similarity
- Co-occurrence
- Cluster Quality

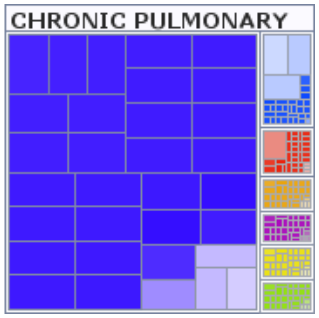








Bad cluster



Good cluster

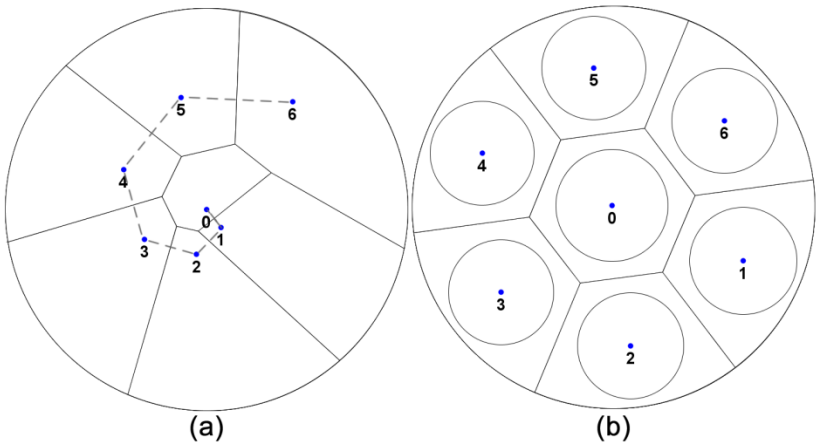
Visual Clues

- Others (Various Small Visual Clues)
  - Use Size to encode feature value, color to encode the SD to the mean vector
    - Small Size + Dark Color : A none important cluster member
    - Small Size + Light Color : A none important member that don’t belongs to the cluster
    - Large Size + Dark Color : A key member of the cluster
    - Large Size + Light Color : A key outlier of the cluster

ICON Layout Algorithms

- Multiple hierarchical ICON designs are made:
  - Pie-Chart TreeMap ICON
    - Breaks the design principle 2 or 4:
      - “uniform and consistent”
      - “organized and stabilize”
  - Rectangular TreeMap ICON
    - Breaks the design principle 4
  - Voronoi ICON
    - We proposed an new optimal constraint voronoi icon layout to satisfy all the design principles

Voronoi ICON Layout



- Centroidal Voronoi
- Weighted Voronoi
- Layout Model

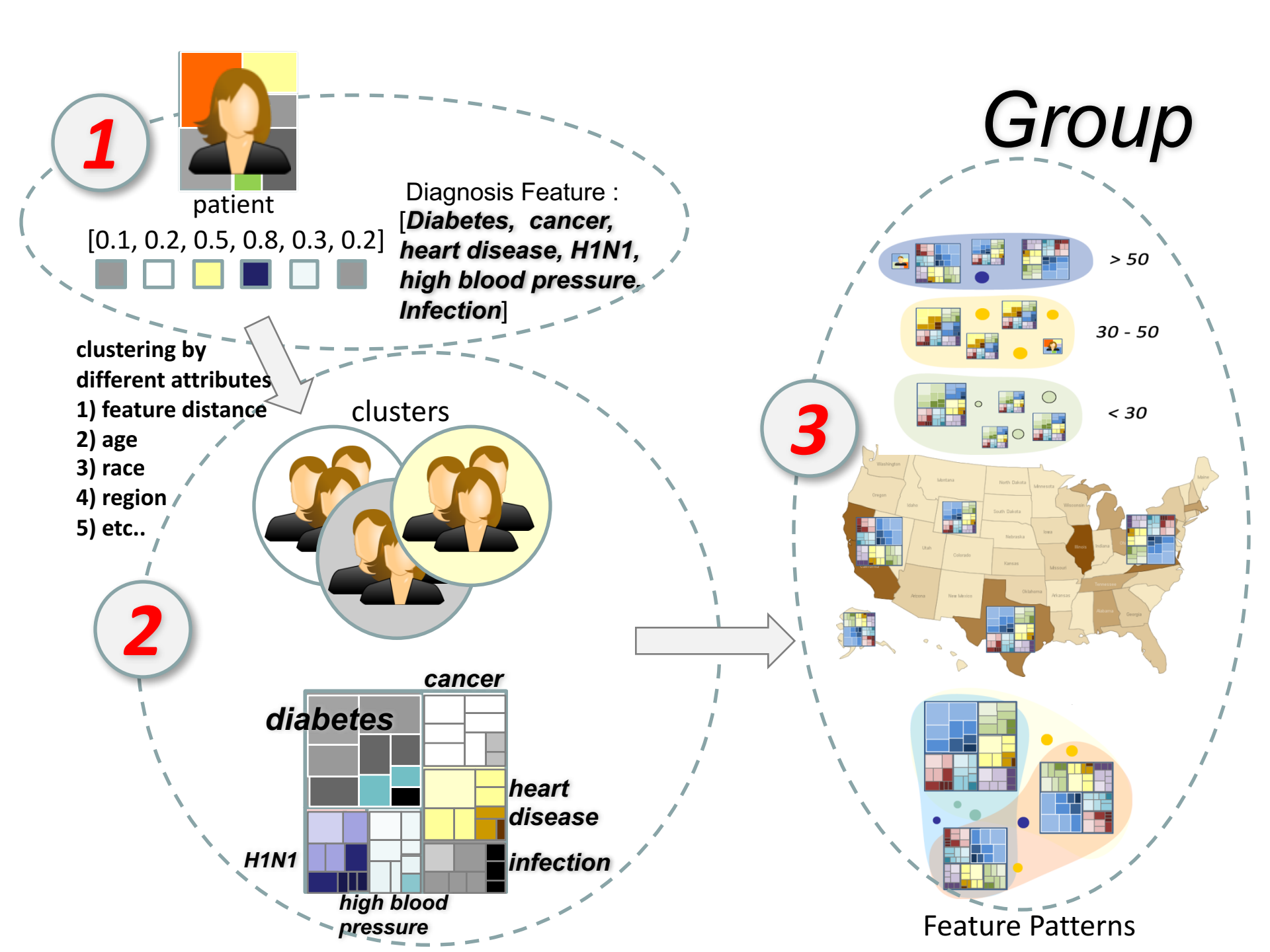
$$\mu_1 \sum_i |X_i - c_i|^2 + \mu_2 \sum_{i<j} (\omega_{ij}|X_i - X_j| - d_{ij})^2 + \mu_3 \sum_i |X_i - pre(X_i)|^2$$

Algorithm 1: VoronoiIconLayout()

Data: $S(s_i,...s_n), V, \varepsilon, pre(s_1),...,pre(s_n)$
Result: coordinates of each site $X(X_1, X_2, ..., X_n)$
begin
  if *pre(S) is not empty* **then**
    $X_i' = pre(s_i)$;
  else
    $X_i' =$ random locations within V;
  $stress' = 10000$ *//give a very large value*;
  while $ratio > \varepsilon$ **do**
    //the coordinate update based on stress majorization;
    $X_i = \frac{\sum_{i<j} \omega_{ij}(x_j + d_{ij}(x_i' - x_j')inv(||X_i' - X_j'||))}{\sum_{i<j} w_{ij}}$;
    compute voronoi tessellation v_i according to X_i;
    compute c_i according to v_i;
    $X_i = \mu_1 \cdot (X_i - c_i) + \mu_2 \cdot X_i + \mu_3 \cdot (X_i - pre(s_i))$;
    AdjustWeight($w_i, a_i, a_{idesired}, X$);
    $str_1 = \sum_i |X_i - c_i|^2$;
    $str_2 = \sum_i |X_i - pre(X_i)|^2$;
    $str_3 = \sum_{i<j} (|X_i - X_j| - d_{ij})^2$;
    $stress = \sum_k (\mu_k \cdot str_k)$;
    $ratio = (stress' - stress)/stress'$;
    $\mu_k = \mu_k \cdot (1 + (stress - str_k)/stress)$;
    normalize μ_k;
    $X_i' = X_i$;
    $stress' = stress$;

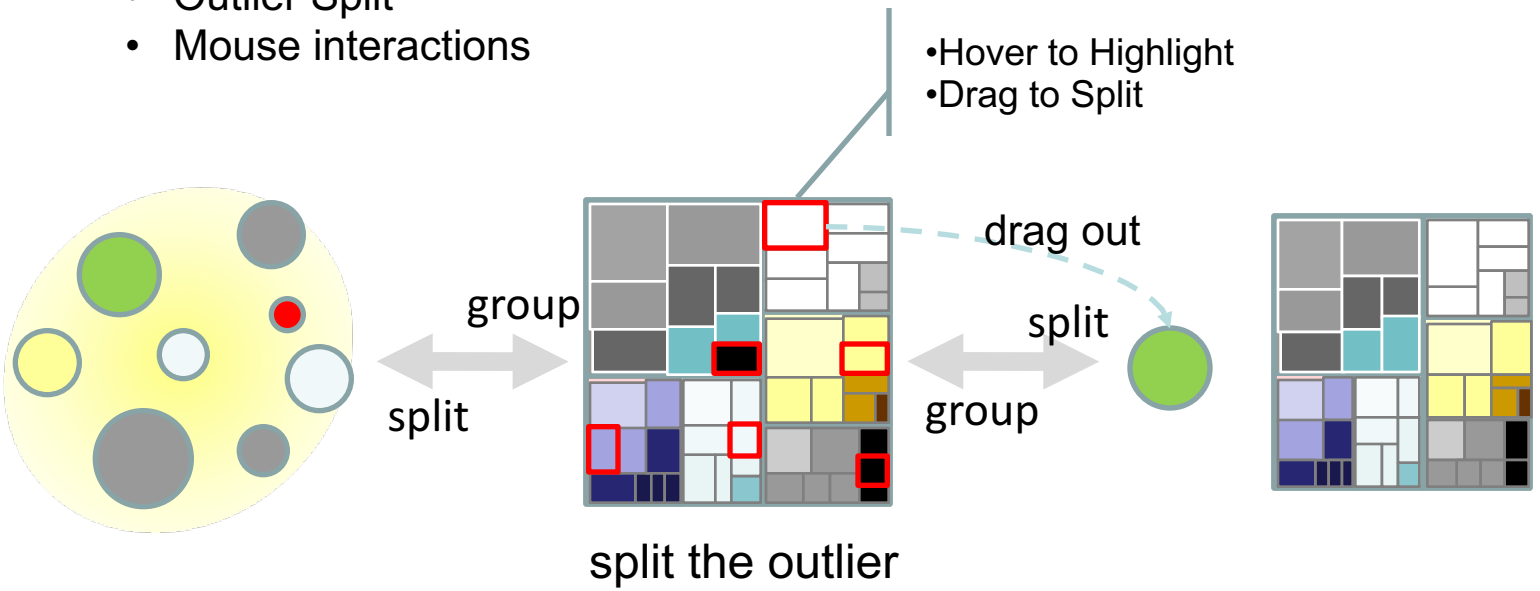
Algorithm 2: AdjustWeight()

Data: the weight w_i of s_i, area a_i of region v_i defined by s_i,
  desired area $a_{idesired}$ of region v_i, $X(X_i,...,X_n)$
Result: new weight of s_i
begin
  $w_i = w_i \cdot (1 + (a_{idesired} - a_i)/a_{idesired})$
  if $w_i < 1$ **then**
    $w_i = 1$
  $ratio = min\{|X_i - X_j|^2/(w_i + w_j)\}$;
  if $ratio < 1$ **then**
    $w_i = w_i \cdot ratio$

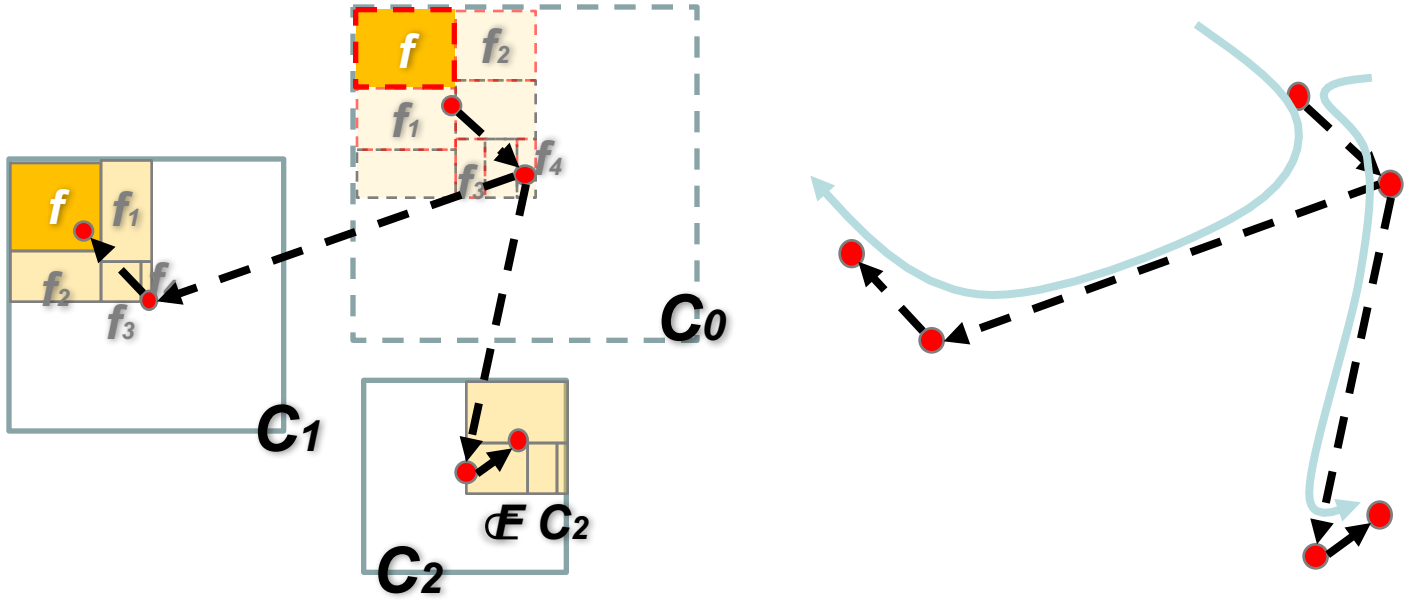

Group
1
patient
[0.1, 0.2, 0.5, 0.8, 0.3, 0.2]
Diagnosis Feature :
[Diabetes, cancer, heart disease, H1N1, high blood pressure, Infection]
clustering by different attributes
1) feature distance
2) age
3) race
4) region
5) etc..
clusters
2
cancer
diabetes
heart disease
H1N1
infection
high blood pressure
3
> 50
30 - 50
< 30
Feature Patterns

Interactions

- Cluster Manipulation
  - Merge : mouse interactions
  - Split
    - Binary Split
    - Outlier Split
    - Mouse interactions



Animation Path Bundle



- Bundle the path to make the transition trend more clear (Demo)

Evaluation

- **T1: People at the elder ages have a high probability to get heart diseases. (Cluster Explanation)**
- **T2: The people in the group 49 – 60 and 61 – 80 pretend to get similar diseases (Cluster Comparison)**
- **T3: The diseases GI disorder, Chronic Pulmonary and GI Disorder are usually accompanied with each other (Cluster Pattern Detection)**
- **T4: When grouping by ages the cluster quality is better than cluster grouping by patients' primary disease. (Visual Cluster Evaluation)**

Initial Results (Demo)

Remaining Tasks

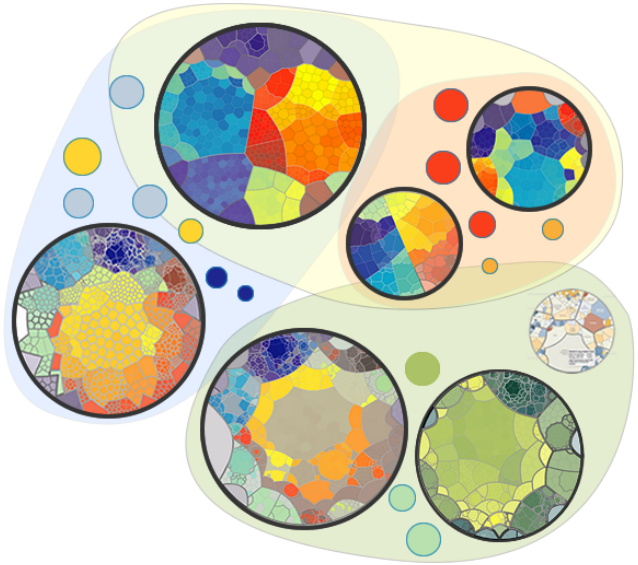
- Major Problem
  - Evaluation
  - Some of the coding
- Workload
  - When task ready, participants ready, should be finished within one week
  - Coding need an additional week
- Mile Stone

  >>> finish the coding before March

  >>> finish the user study before 10th March

  >>> finish the writing of case study before 25th Feb

  >>> finish the first draft before 15th March



SmartIcon: Multivariate Feature Visualization

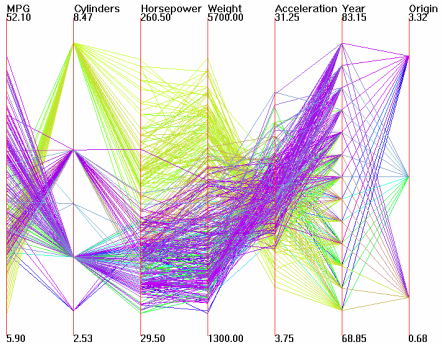
design, concepts & scenarios

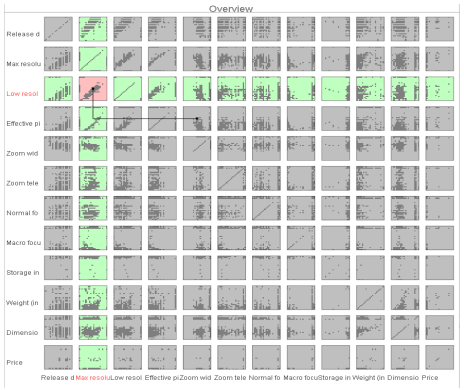
Nan Cao @ HKUST

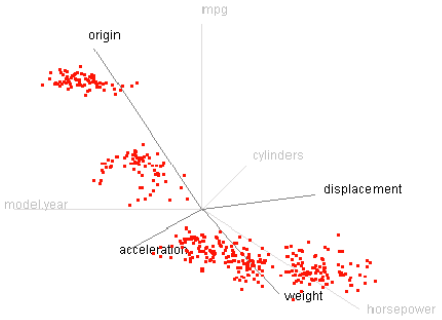
July 2010

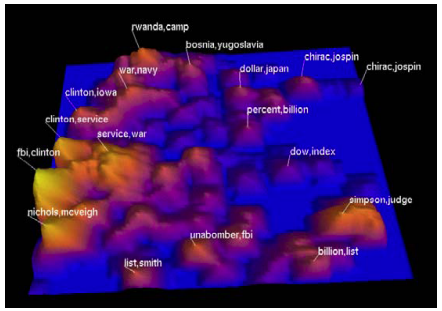
Prior Art: Geometric Techniques

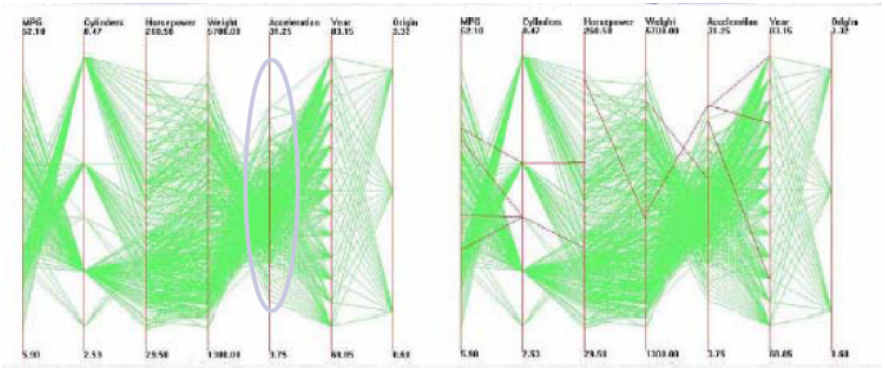
- Basic idea: Visualization of geometric
  - transformations and projections of data
  - Scatterplot-matrices [And 72, Cle 93]
  - Parallel coordinates [Ins 85, ID 90]
  - Star coordinates [Kan 2000]
  - Landscapes [Wis 95]
  - Projection Pursuit Techniques [Hub 85]
  - Prosection Views [FB 94, STDS 95]
  - Hyperslice [WL 93]

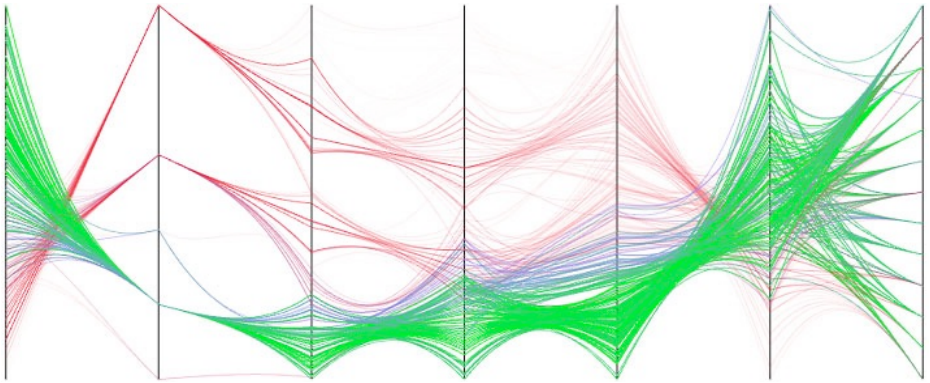


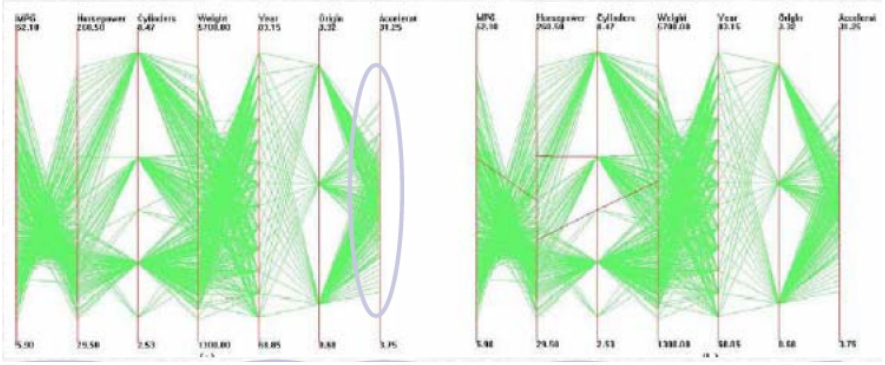


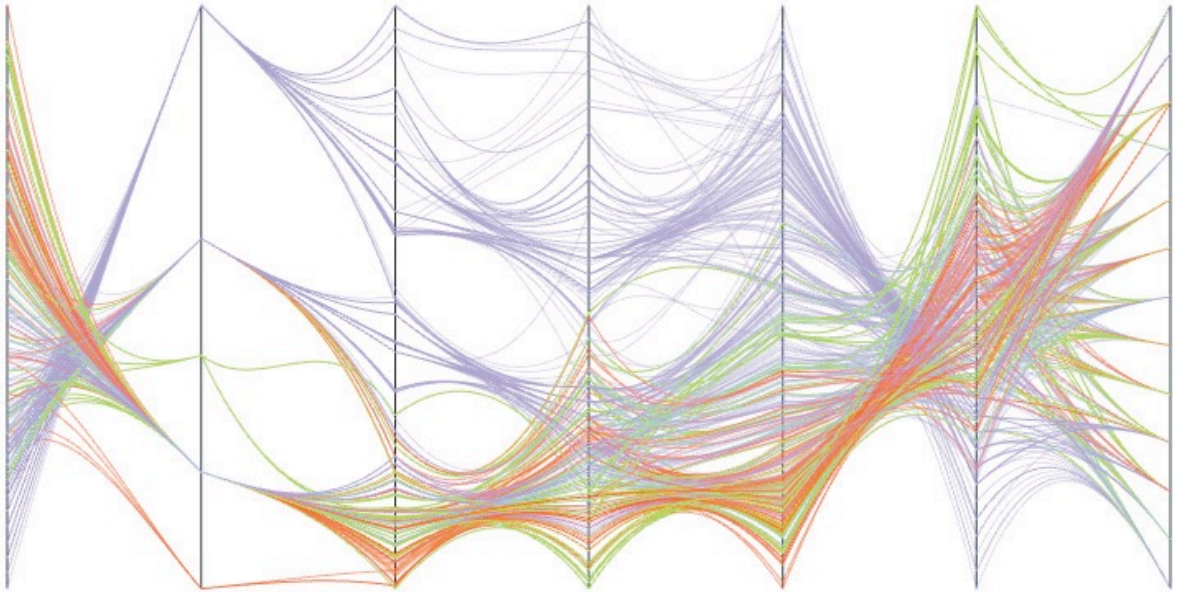


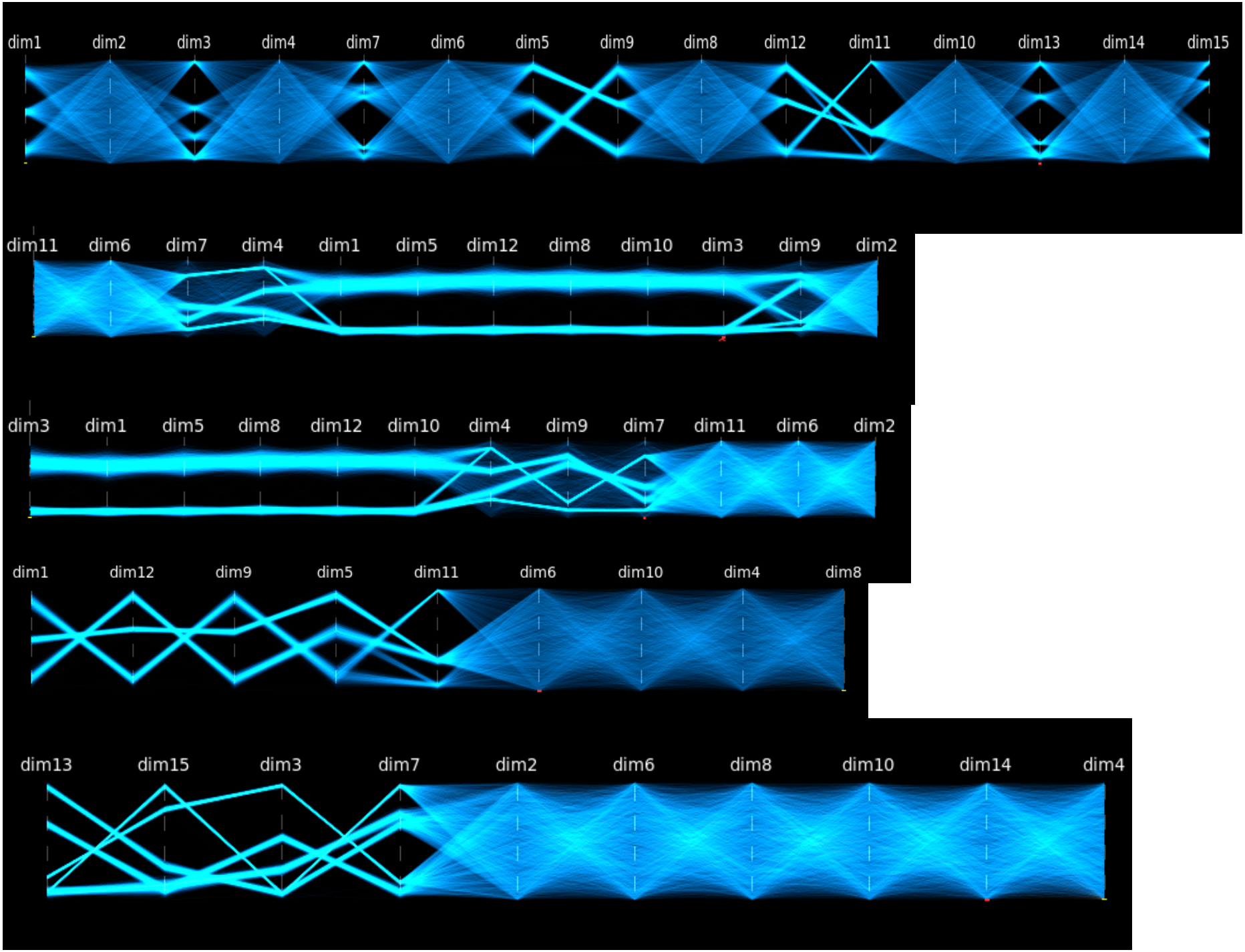










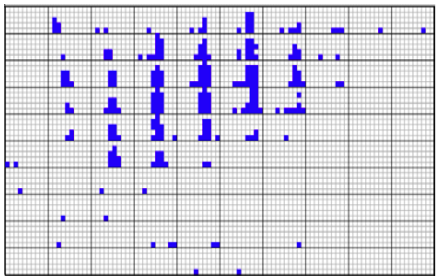

dim1 dim2 dim3 dim4 dim7 dim6 dim5 dim9 dim8 dim12 dim11 dim10 dim13 dim14 dim15
dim11 dim6 dim7 dim4 dim1 dim5 dim12 dim8 dim10 dim3 dim9 dim2
dim3 dim1 dim5 dim8 dim12 dim10 dim4 dim9 dim7 dim11 dim6 dim2
dim1 dim12 dim9 dim5 dim11 dim6 dim10 dim4 dim8
dim13 dim15 dim3 dim7 dim2 dim6 dim8 dim10 dim14 dim4

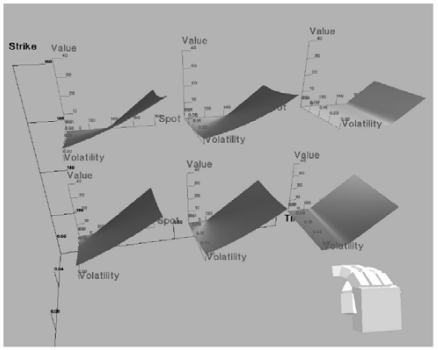
Prior Art

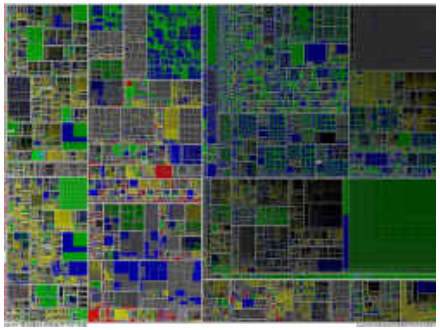
- Visualization for multi-dimensional dataset (see backup for details)
  - Geometric techniques
    - *Scatterplot matrices*, *parallel coordinates*, landscapes, …
  - Hierarchical techniques
    - *Dimensional stacking*, worlds-within-worlds, Treemap, ConeTree
  - Icon-based techniques
    - *Star glyphs*, stick figures, shape-coding, color icons, chernoff faces
  - Pixel-oriented techniques
    - Recursive pattern, circle segments, spiral, axes, techniques
  - Table-based techniques
    - HeatMap, tableLens, Tabular
  - Others (Hybrid techniques)
    - Scatter plot in parallel coordinates, NodeTrix, FacetAtlas, Bubble Set
- ***New problem, no existing*** work fulfills all our problems

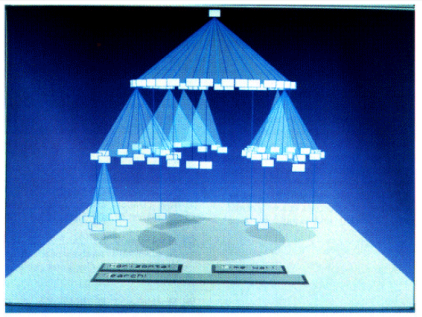
Prior Art: Hierarchical Techniques

- Basic ideas: Visualization of the data using a hierarchical partitioning into subspaces
  - Dimensional Stacking [LWW90]
  - Worlds-within-Worlds [FB 90a/b]
  - Treemap [Shn 92, Joh 93]
  - Cone Trees [RMC 91]
  - InfoCube [RG93] (OLAP)





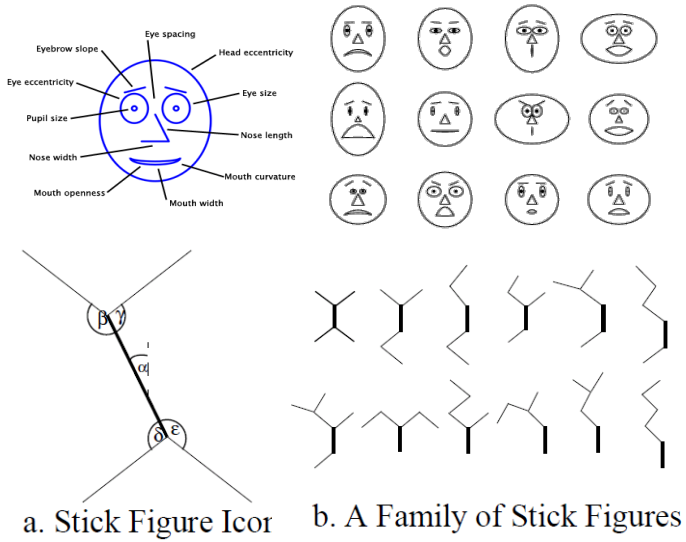


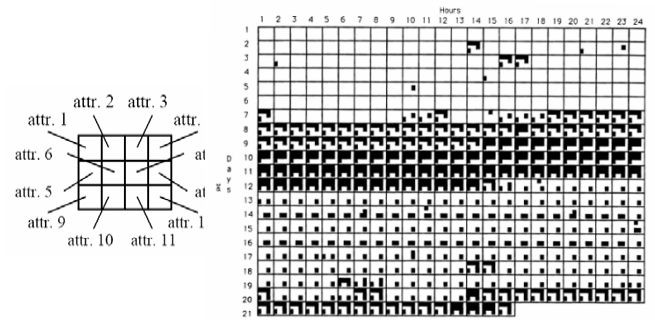


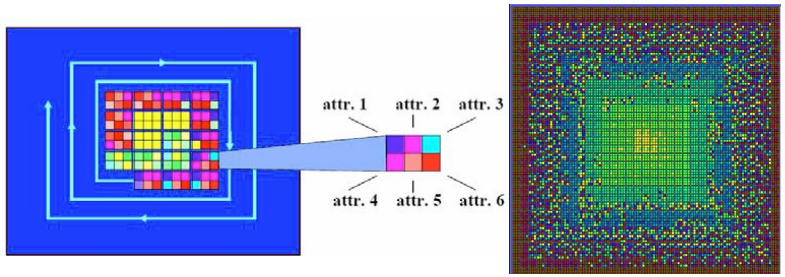
Robertson Plate 1

Prior Art: Icon-based techniques

- Chernoff face visualization
- Stick figure technique
  - two dimensions are mapped to the display dimensions and the remaining dimensions are mapped to the angles and/or limb lengths of the stick figure icon
  - the number of dimensions that can be visualized is limited
- Shape encoding
- Color Icons

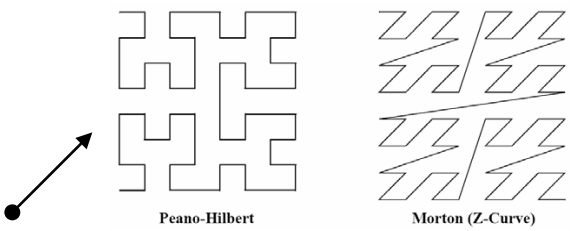


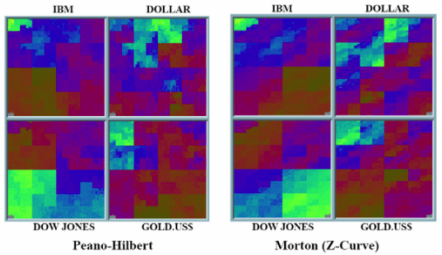


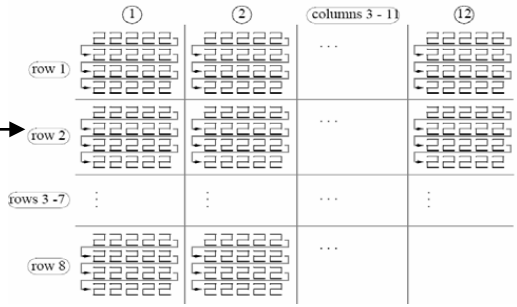


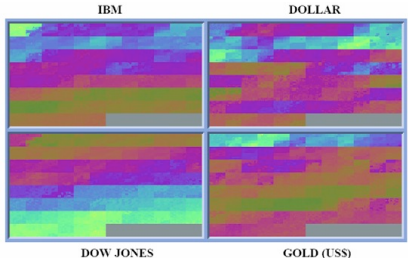
Prior Art:Pixel-Oriented Techniques

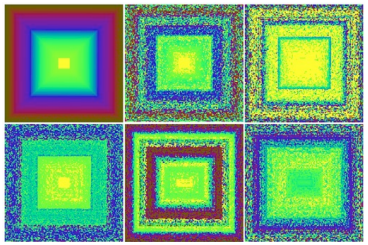
- Query Independent
  - Space-Filling Curve Arrangements
  - Recursive Pattern Technique
- Query Dependent
  - Spiral Technique
  - Axes Technique
  - Circle Segments

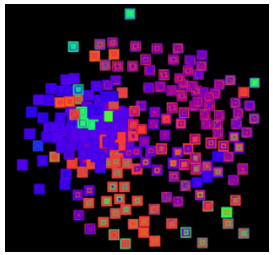


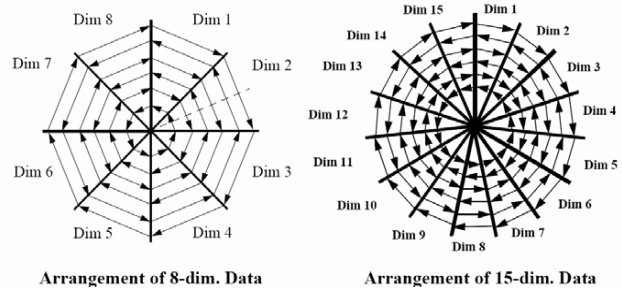


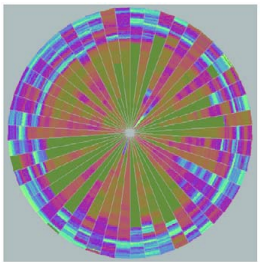






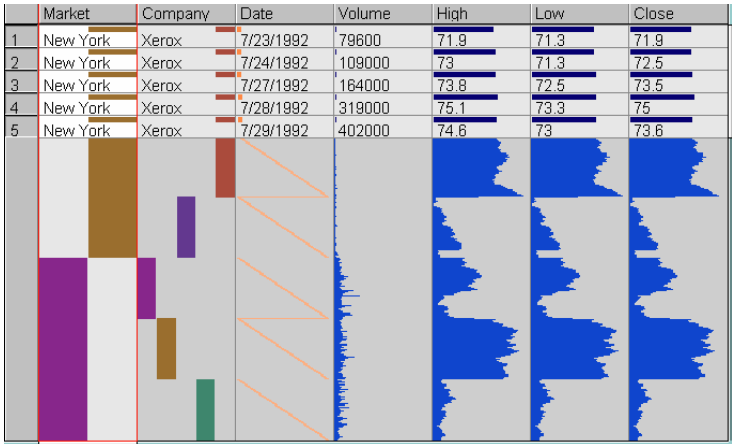


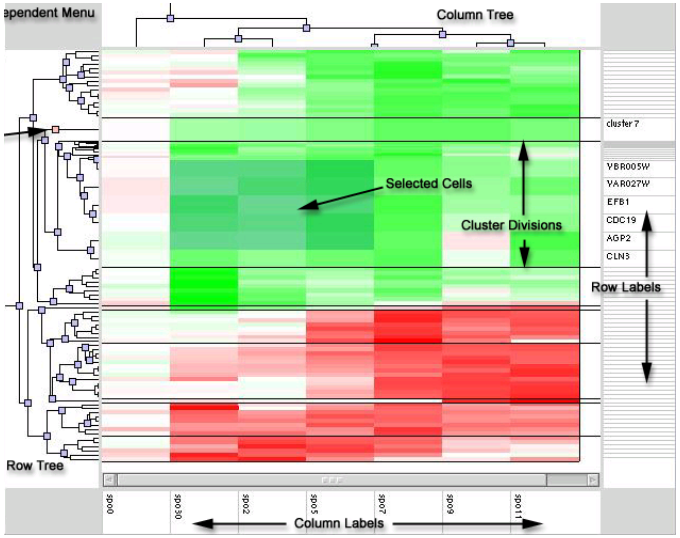


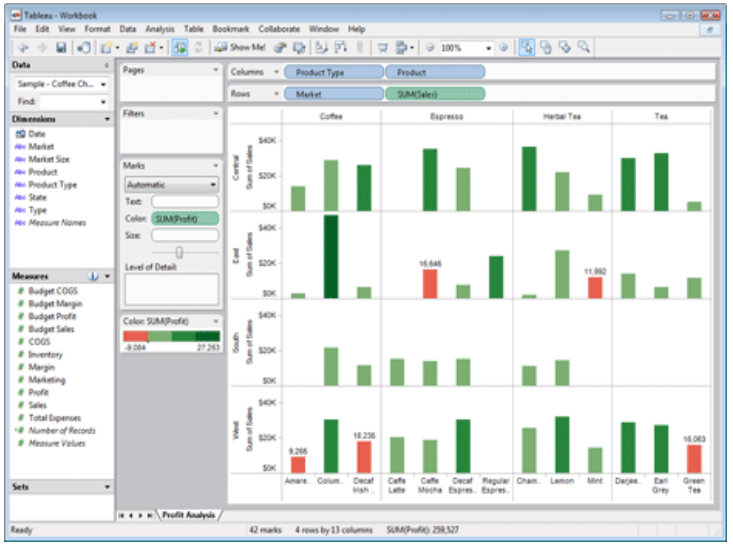


Prior Art: Table-based techniques

- Table Lens
- Tableau
- Heat Map







Prior Art: Others (Hybrid Techniques)

- NodeTrix: a Hybrid Visualization of Social Networks. Nathalie Henry, Jean-Daniel Fekete, Michael J. McGuffin, InfoVis 2007
- Scattering Points in Parallel Coordinates. Xiaoru Yuan, Peihong Guo, He Xiao, Hong Zhou, Huamin Qu, InfoVis 2009
- Bubble Sets: Revealing Set Relations with Isocontours over Existing Visualizations, Christopher Collins, Gerald Penn, Sheelagh Carpendale, InfoVis 2009
- Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation. Niklas Elmqvist, Pierre Dragicevic, Jean-Daniel Fekete, InfoVis 2008
- Interactive Dimensionality Reduction Through User-defined Combinations of Quality Metrics, Sara Johansson, Jimmy Johansson, InfoVis 2009
- FacetAtlas: Multifaceted Visualization for Rich Text Corpora, InfoVis 2010