Data Visualization

Vidya.N

Introduction

- Researchers from the University of Berkeley estimate that every year about 1.5 Exabytes (= 1.5 Milhon Terabytes) of data are generated, most of which are available in digital form.
- Even simple transactions of everyday life, are typically recorded by computers.

Introduction(Contd.)

- The data are collected because people believe that it is a potential source of valuable information.
- Extracting information from these data is difficult .
- So far we have learnt a lot of tools and methods in our course to extract information from data.

Why Data Visualization?

- With today's data management systems, it is only possible to examine quite small portions of the data.
- This is like a drop in the ocean when dealing with millions of data items.
- Having no possibility to explore the large amounts of data that have been collected, the data becomes useless and the databases become data 'dumps'

Data Visualization

- **Data visualization** is a general term that describes any effort to help people understand the significance of **data** by placing it in a visual context.
- The basic idea of visual data mining is to present the data in some visual form, allowing the user to gain insight into the data, draw conclusions, and directly interact with the data.
- Patterns, trends and correlations that might go undetected in text-based **data** can be exposed and recognized easier with **data visualization** software**.**

Benefits of Data Visualization

- Can easily deal with highly non-homogeneous and noisy data.
- It is intuitive and requires no understanding of complex mathematical or statistical algorithms or parameters.
- Visualization can provide a qualitative overview of the data, allowing data phenomena to be isolated for further quantitative analysis.

Benefits of Data Visualization (Contd)

- Allows a faster data exploration and often provides more interesting results, especially in cases where automatic algorithms fail.
- Provides a much higher degree of confidence in the findings of the exploration.

Visual Exploration Paradigm

- It follows a three step process:
- **Overview** :User identifies interesting patterns or groups in the data and focuses on one or more of them.
- **Zoom and filter :**User drills-down and access details of the data.
- **Details-on-demand** : Provide further details when requested for.

Classification of Visual Data Analysis Techniques

- The techniques can be classified based on three criteria :
  - The data type to be visualized,
  - The visualization technique, and
  - The interaction technique used
- The three dimensions of our classification can be assumed to be orthogonal .
- A specific system may support different data types and it may use a combination of visualization and interaction techniques.

Data type to be visualized

- **One-dimensional data**
  - *E.g : time series of stock prices(temporal data)*
  - *Method : The Circle Segments Technique*
- **Two-dimensional data**
  - *E.g : geographical data*
  - *Method : x-y Plot*
- **Multi-dimensional data**
  - *E.g: tables from relational databases*
  - *Method : Parallel Coordinates Technique*

Parallel coordinate Technique

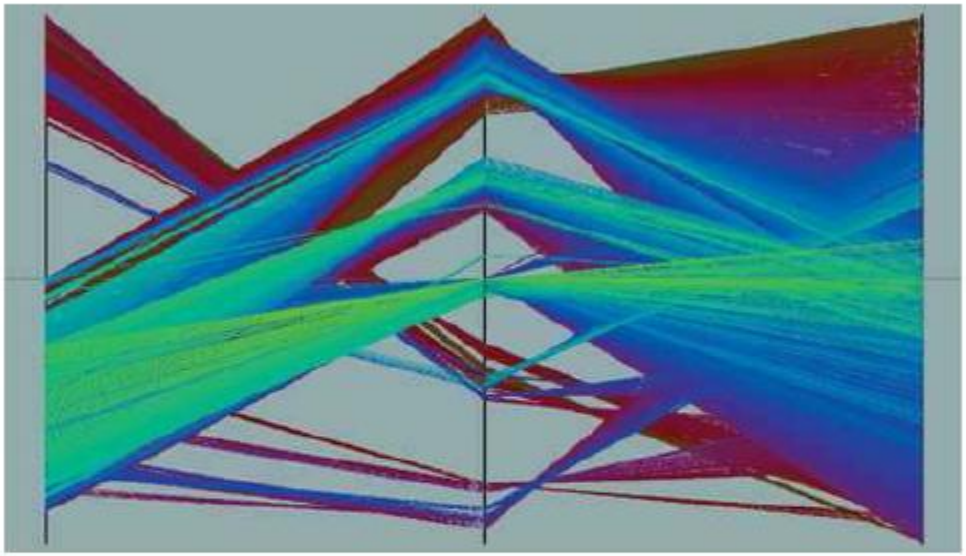


Fig. 11.4. The Parallel Coordinates Technique [280] belongs to the class of geometrically-transformed displays. In conjunction with similarity-based coloring, it displays each multi-dimensional data item as a polygonal line intersecting the dimension axes at the position corresponding to the data value for the dimension. *(from [301] ©IEEE)*

Data type to be visualized

- **Text and hypertext**
  - *E.g :Web documents*
- Hierarchies and graphs
  - *E.g: e-mail interrelationships among people, their shopping behavior, the file structure of the hard disk, or the hyperlinks in the World Wide Web.*
- Algorithms and software
  - *E.g: such as debugging operations*

Skitter graph Internet Map

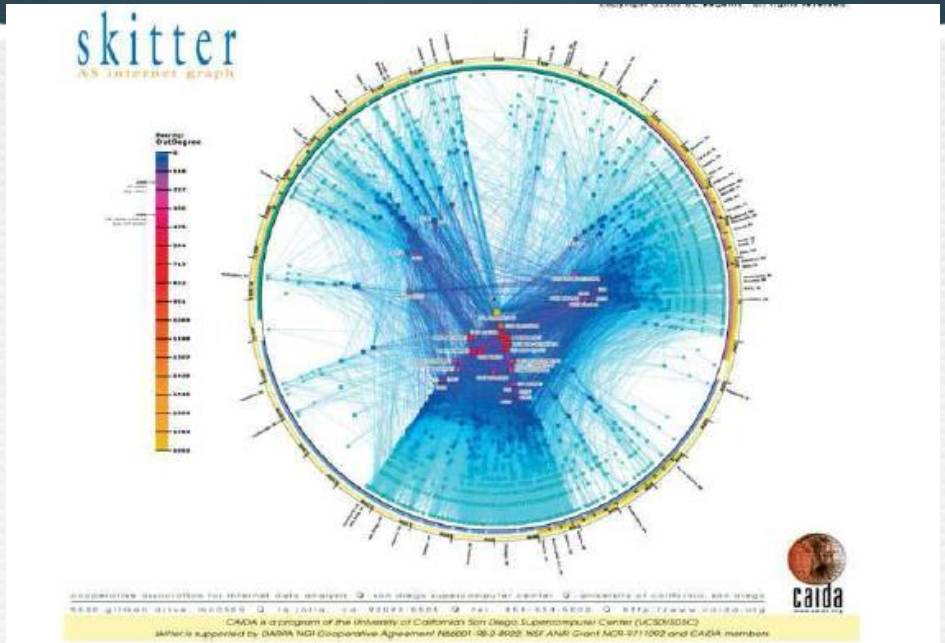


Fig. 11.5. The Skitter Graph Internet Map shows the global structure of the internet. The nodes represent autonomous systems which in general correspond to ISPs. The nodes are positioned according to their geographical longitude in polar coordinates. The more important nodes with a high number of connections are positioned towards the center and are colored accordingly. The visualization shows the high connectivity within North America and the strong connections between Europe (or Asia) and the US. In contrast, direct connections between Europe and Asia are rare. The data shown have been collected in two weeks of October 2000. *(used by permission of the Cooperative Association for Internet Data Analysis, "Skitter Graph" ©2000 UC Regents. Courtesy University of California)*

The visualization technique used

- Standard 2D/3D displays, such as bar charts and x-y plots
- Geometrically-transformed displays, such as landscapes and parallel coordinates
- Icon-based displays, such as needle icons and star icons
- Dense pixel displays, such as recursive patterns and circle segments
- Stacked displays, such as treemaps and dimensional stacking

Visualization Techniques

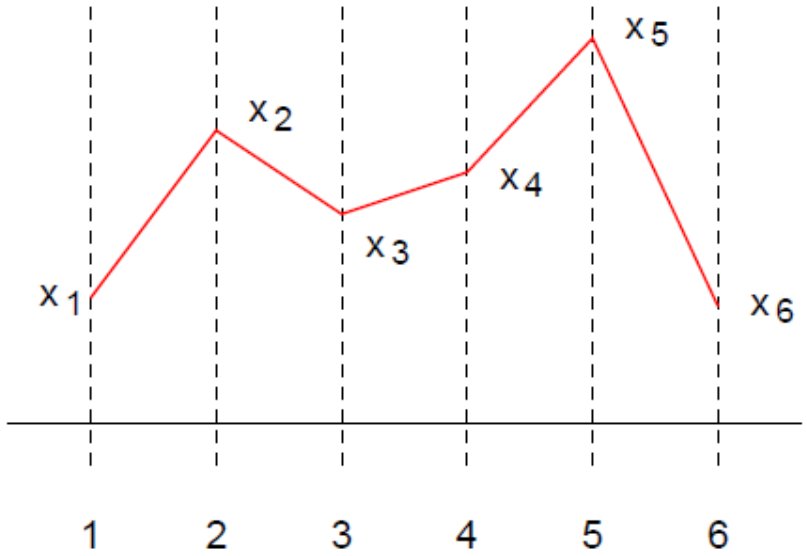
- Geometrically-Transformed Displays:
  - finding "interesting" transformations of multi-dimensional data sets.
  - E.g: Scatter Plot and Parallel visualization Techniques

Parallel Visualization Technique

- The Parallel Coordinate Technique maps the k-dimensional space onto the two display dimensions by using k axes that are parallel to each other (either horizontally or vertically oriented) and are evenly spaced across the display.
- The axes correspond to the dimensions and are linearly scaled from the minimum to the maximum value of the corresponding dimension.
- Each data item is presented as a chain of connected line segments, intersecting each of the axes at the location corresponding to the value of the dimension considered.

Parallel Visualization Technique

- A variable = a vertical axis
- An object = a piecewise linear curve
- $(x_1, \ldots, x_p)$ is mapped to the polyline that joins $(1, x_1)$, $(2, x_2)$, ..., (p, x_p)



Parallel Visualization Technique

- Pros:
  - reasonable variable scalability
  - powerful after proper training
- Cons:
  - learning curve
  - major overlapping problem
  - variable order

Visualization Techniques

- **Iconic Displays**
  - map the attribute values of a multi-dimensional data item to the features of an icon.
  - Icons may be defined arbitrarily
    - little faces ,needle icons , star icons , stick figure icons , color icons or Tile Bars
  - **Stick figure technique :**Two dimensions are mapped to the display dimensions and the remaining dimensions are mapped to the angles and/or limb length of the stick figure icon

Iconic Display

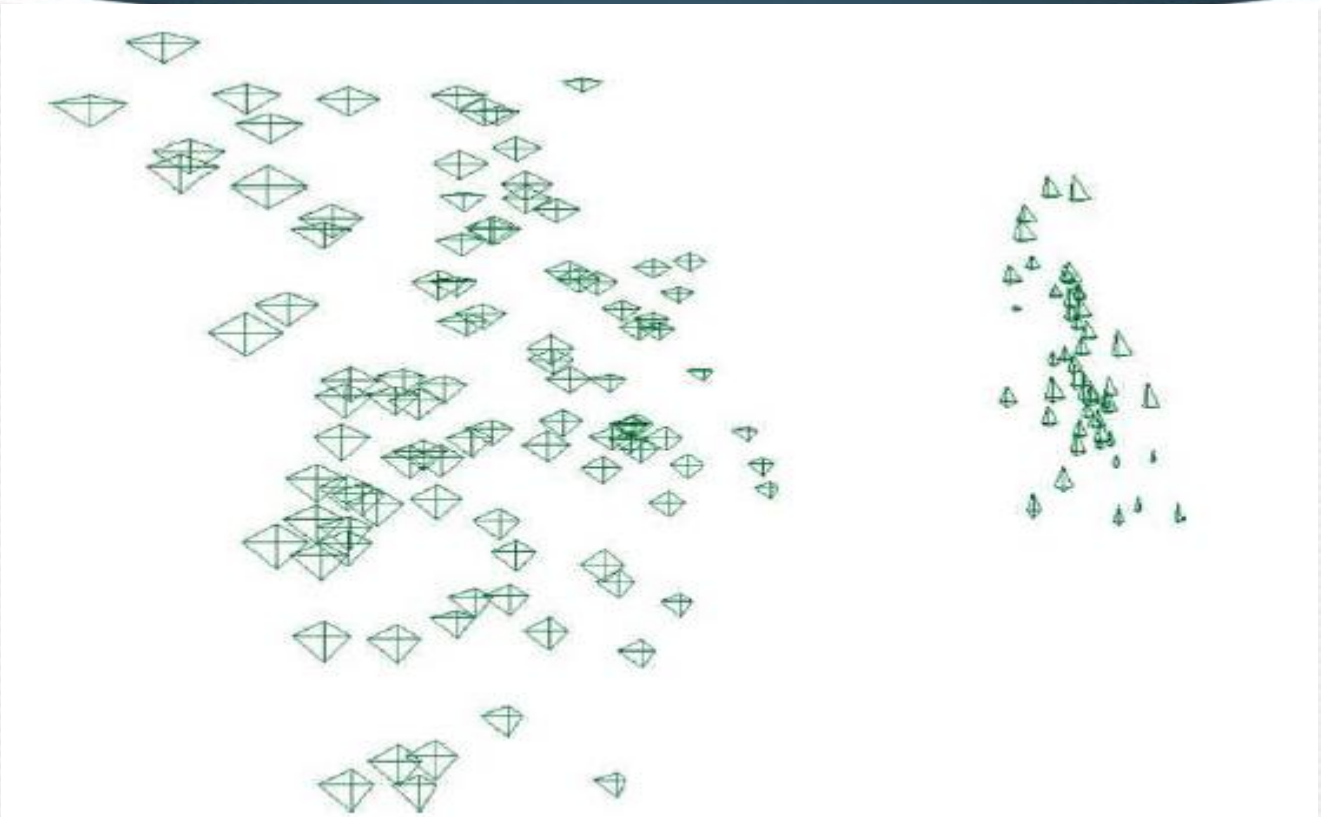


Fig. 11.6. The iris data set – displayed using star glyphs positioned based on the first two principal components. *(generated using XmdvTool [533])*

Visualization Techniques

- Dense Pixel Displays
  - map each dimension value to a colored pixel and group the pixels belonging to each dimension into adjacent areas
  - use one pixel per data value, the techniques allow the visualization of the largest amount of data possible on current displays.
  - E.g: Recursive pattern technique and the circle segments technique .

Dense Pixel Displays

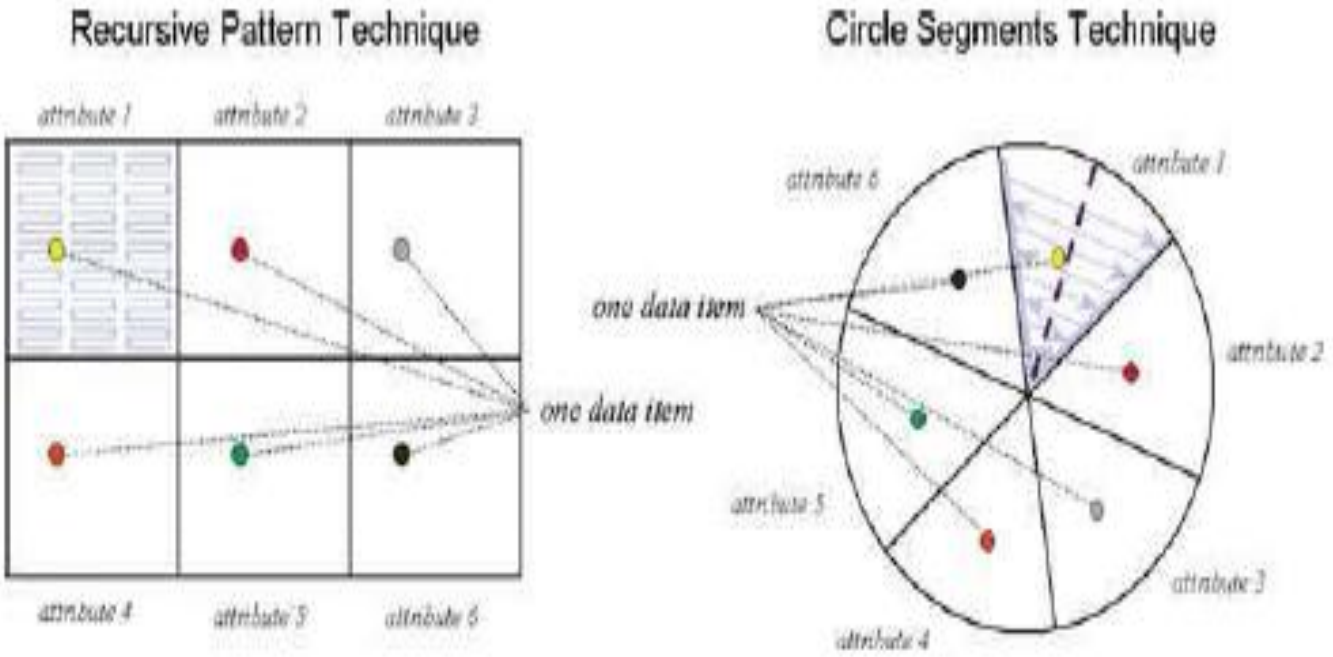


Fig. 11.7. The basic idea of Recursive Pattern and Circle Segments techniques.

Dense Pixel Displays

Advantages of DPDs:

Excellent information density that takes advantage of limited display space

They make it easy to identify patterns, especially correlations between different variables.

Disadvantages:

For some data with non-natural ordering, complicated sorting/clustering algorithms are needed to make the pixel display useful for comparing trends.

It is not suited for use in cases where you want to see clear individual values (it is more useful for pattern finding / as an overview)

Visualization Techniques

- Stacked Displays
  - To present data partitioned in a hierarchical fashion.
  - In the case of multi-dimensional data, the data dimensions to be used for partitioning the data and building the hierarchy have to be selected appropriately.
  - E.g: Dimensional Stacking
  - Embed one coordinate system inside another coordinate system, i.e. two attributes form the outer coordinate system, two other attributes are embedded into the outer coordinate system.
  - Divide the outermost level coordinate system into rectangular cells. Within the cells, the next two attributes are used to span the second level coordinate system.
  - This process may be repeated multiple times.

Stacked Displays

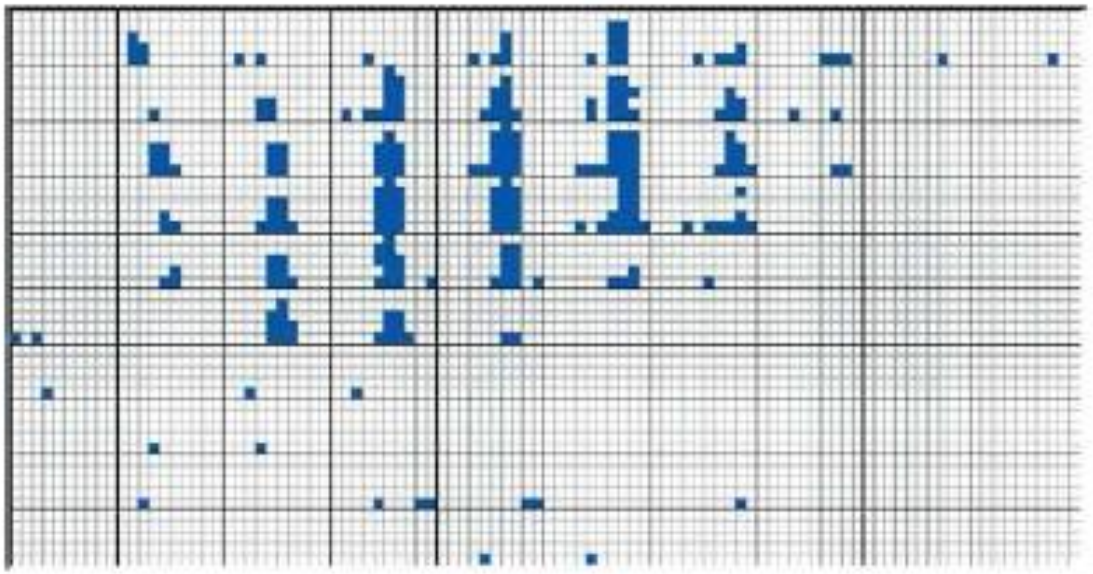
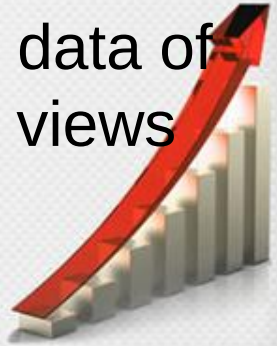


Fig. 11.8. Dimensional Stacking visualization of drill hole mining data.
(used by permission of M. Ward, Worcester Polytechnic Institute ©IEEE)

Interaction technique

- **Dynamic Projection**, that allows smooth navigations through the data space.
- **Interactive Filtering**, to enable users to isolate subsets of data for focused analysis .
- **Zooming**, to enlarge data for detailed analysis
- **Distortion**, to increase the screen space allocated to areas of interest while preserving the context of the entire data set
- **Linking and Brushing**, to enable users to select data of interest in one view and see it highlighted in other views



Interaction technique

- **Dynamic Projection :**
  - It is an automated navigation operation.
  - The basic idea is to dynamically change the projections in order to explore a multi-dimensional data set.
  - E.g : Grand Tour system
    - Shows all interesting - i.e., those exhibiting desirable properties such as well-separated clusters - two-dimensional projections of a multi-dimensional data set as a series of scatterplots.
    - Note that the number of possible projections is exponential in the number of dimensions, i.e., it is intractable for large dimensionality
  - **Link to watch:** https://www.youtube.com/watch?v=x0tqO3jrNWw

Interaction technique

- Interactive Filtering :
  - Is a combination of selection and view enhancement.
  - In exploring large data sets, it is important to interactively partition the data set into segments and focus on interesting subsets.
  - This can be done by a direct selection of the desired subset (browsing) or by a specification of properties of the desired subset (querying).
  - Eg : the Magic Lens
    - filter the data directly in the visualization.
    - The data under the magnifying glass is processed by the filter and displayed in a different way than the remaining data set
    - . Magic Lenses show a modified view of the selected region, while the rest of the visualization remains unaffected.
  - Link to watch : https://www.youtube.com/watch?v=uzRt_vLOLdM

Interaction technique

- Zooming:
  - Data is in a highly compressed form to provide an overview of the data but at the same time allow a variable display of the data at different resolutions.
  - Zooming does not only mean displaying the data objects larger, but also that the data representation may automatically change to present more details on higher zoom levels.
  - The objects may, for example, be represented as single pixels at a low zoom level, as icons at an intermediate zoom level, and as labeled objects at a high resolution.

Interaction technique

- Zooming:
  - Eg: TableLens approach
    - Represent each numerical value by a small bar.
    - All bars have one-pixel height and the lengths are determined by the attribute values.
    - This means that the number of rows on the display can be nearly as large as the vertical resolution and the number of columns depends on the maximum width of the bars for each attribute.
    - The initial view allows the user to detect patterns, correlations, and outliers in the data set.
    - In order to explore a region of interest the user can zoom in, with the result that the affected rows (or columns) are displayed in more detail, possibly even in textual form.

Table Lens

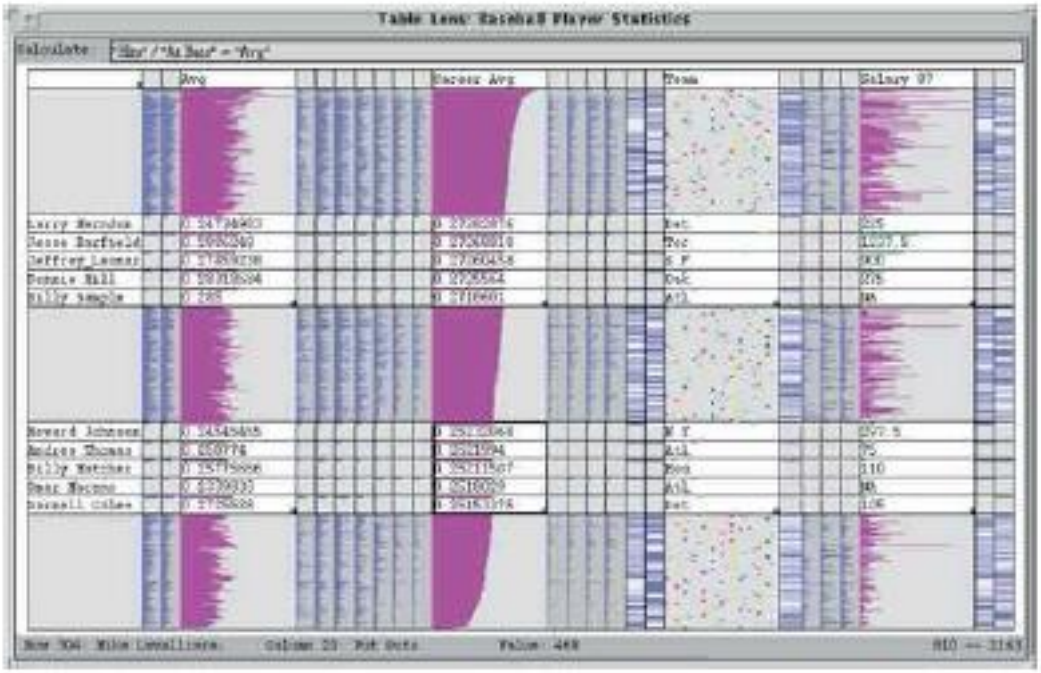


Fig. 11.9. The Table Lens approach. *(used by permission of R. Rao, Xerox PARC ©ACM)*

Interaction technique

- Distortion:
  - Supports the data exploration process by preserving an overview of the data during drill-down operations.
  - The basic idea is to show portions of the data with a high level of detail while others are shown with a lower level of detail.
  - Popular distortion techniques are hyperbolic and spherical distortions .
  - Often used on hierarchies or graphs but may also be applied to any other visualization technique.

Fish-eye Lens

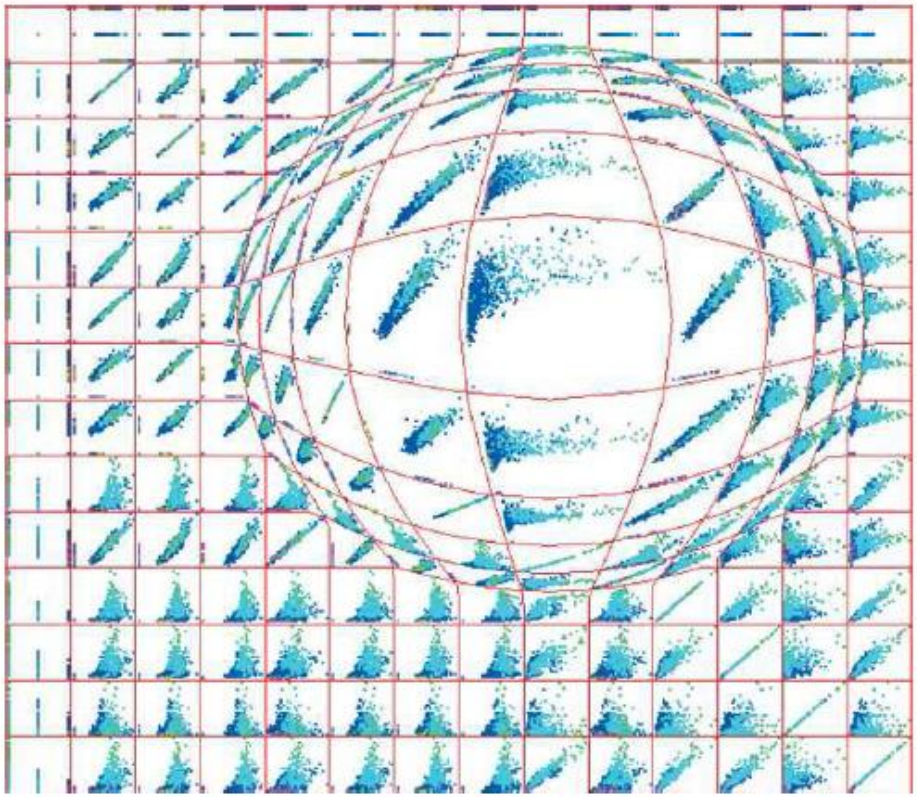


Fig. 11.10. A scatterplot matrix with part of the display distorted using a fisheye lens. *(used by permission of M. Ward, Worcester Polytechnic Institute)*

Interaction technique

- **Brushing and Linking :**
  - Brushing is an interactive selection process for communicating the selected data to other views of the data set.
  - The idea of linking and brushing is to combine different visualization methods to overcome the shortcomings of individual techniques.
  - linking and brushing can be applied to visualizations generated by all visualization techniques described above.
  - As a result, the brushed points are highlighted in all visualizations, making it possible to detect dependencies and correlations.
  - Interactive changes made in one visualization are automatically reflected in the other visualizations.
  - Connecting multiple visualizations through interactive linking and brushing provides more information than considering the component visualizations independently.

Systems and Application

Introduction

- Aim of Data Analytics:
  - To gain competitive advantage using the knowledge from corporate data leading to "business intelligence" - the gathering and management of data, and the analysis of that data to turn it into useful information to improve decision making.

Areas Predominantly Used in:

1. Business and Finance
2. Science and Engineering
3. Bioinformatics
4. Medicine and Health Care

Business and Finance

- Applications include fraud detection, customer retention, cross selling, stock prediction, marketing and insurance.
- Financial Crimes Enforcement Network (FINCEN) of the US Treasury Department, developed system called "FAIS": the FINCEN AI Systems [471]. FAIS detects potential money-laundering activities from a large number of big cash transactions.
- IBM's Intelligent Miner, developed a credit card attrition model to predict which customers will stop using Mellon's credit card in the next few months. Based on the prediction results, the bank can take marketing actions to retain these customers loyalty.
- TV program schedulers would like to know the likely audience for a proposed program and the optimum time to show it. Using Clementine, the Integral Solutions Limited developed a system to predict television audiences for the BBC
- "Market basket analysis", examining associations between different products bought by customer, has been popular with organizations such as supermarkets and catalog selhng companies. The underlying technology is the so-called association rule algorithm

Science and Engineering

- Pavilion Technologies' Process Insights, an apphcation development tool that combines neural networks, fuzzy logic and statistical methods, has been successfully used by Eastman Kodak and other companies to develop chemical manufacturing and control applications to reduce waste, improve product quality and increase plant throughput .
- Data Engine is applied for data analysis in process industry. In particular, the tool has been applied to process analysis in chemical, steel and rubber industries, with savings in input materials and improvement in quality and productivity as a result .
- The Sky Image Cataloguing and Analysis Tool (SKICAT) system integrates methods from machine learning, image processing, classification and database, and is reported to be able to classify objects too faint for visual classification with high accuracy .

Bioinformatics

- Bioinformatics is concerned with the development and application of computational and mathematical methods for organising, analysing and interpreting biological data, and requires an interdisciplinary research effort from biologists, computer scientists, statisticians and in general.
- DNA microarray [464] allow the study of thousands of genes in a single experiment and provide a global view of the underlying biological process, for example by revealing which genes are responsible for a disease process, how they interact and are regulated, and which genes are being co-expressed under different experimental conditions
- It has been found that patients receiving the same diagnosis can have markedly different clinical courses and treatment responses For example, the diffuse large B-cell lymphoma has been found clinically heterogeneous: 40% of patients respond well to current therapy and have prolonged survival, while the remainder succumb to the disease .

Medicine and Health Care

- Regularities, trends and surprising events extracted from these data by IDA methods are important in assisting clinicians to make informed decisions, thereby improving health services.
- Clinicians evaluate patient's condition over time. The analysis of large quantities of time-stamped data will provide doctors with important information regarding the progression of the disease
- Guardian is an intelligent autonomous agent for monitoring and diagnosing intensive care patients recovering from cardiac surgery
- Visual field testing provides the eye care practitioner with essential information regarding the early detection of major blindness-causing diseases such as glaucoma.
- a software-based test system has been developed using machine learning techniques (e.g. neural networks and decision tree induction), an intelligent user interface and a pattern discovery model, and this system has been successfully used in several primary care settings

Other Applications

- The Associate was designed to provide the pilot with enhanced situational awareness by sorting and prioritizing data, analyzing sensor and aircraft system data, and turning the data into useful information. Based on this information, plans for achieving mission goals can be developed and presented to the pilot for approval and execution.
- Web Watcher, an operational tour guide for the WWW [292]. It learns to predict what links users will follow on a particular page, highlight the links along the way, and learn from experience to improve its advice-giving skills. The prediction is based on many previous access patterns and the current user's stated interests.
- Advanced Scout was developed for the American National Basketball Association (NBA) coaching staffs to discover interesting patterns in basketball game data. These patterns are then used to assess the effectiveness of certain coaching decisions and to formulate game strategies for subsequent games.

THANK YOU