Extreme Visualization: Squeezing a Billion Records into a Million Pixels

Ashvani Yadav, Lt Cdr Rakesh Dash, Awanish

Department of Computer Science and Engineering
Indian Institute of Technology Kanpur (IITK)
email: {ashvaniy24, rakeshdash24, awanishk24}@cse.iitk.ac.in

SCOPE OF PRESENTATION

- Introduction to Extreme Visualization
- Challenges of Large Data Sets
- Visualization Techniques
- Various Methods – Visualizing Large Data Types
  - Timeline
  - Tree
  - Network
- Density Plot
- Conclusion

INTRODUCTION

- Traditional database searches return results in lists or tables. But with huge databases (millions or billions of records), this isn't very efficient.

- The goal is to help people explore and understand vast amounts of data visually—using graphs and images instead of tables.

- Challenges –

  To scale up visual presentation from millions to billions of records

  Develop compact data structure

  Rapid data filtering, aggression and display rendering

APPROACHES FOR VISUALIZING MASSIVE DATA

- Atomic Visualizations
- Aggregate Visualizations
- Density Plots

ATOMIC VISUALIZATIONS

- **Atomic** - Individual data record is treated as a distinct
- One marker(pixel or point) per data record.
- One-to-one mapping of records to visual markers

ATOMIC VISUALIZATIONS

- **Atomic** - Individual data record is treated as a distinct
- One marker(pixel or point) per data record.
- One-to-one mapping of records to visual markers

HOW ATOMIC VISUALIZATIONS WORK

- In atomic visualizations, basic charts like scatterplots, histograms, pixel-based representation and time series plots are used to represent large datasets.

- Each record is mapped to a marker (such as a pixel or point) on the screen. Users can then interact with the visualization by zooming in or out, filtering records, and using sliders or other controls to explore different parts of the data.

- To handle larger datasets, these visualizations often employ dynamic query filters. For example, users can select subsets of data by adjusting sliders.

- When filtering or zooming, the display updates in real-time (within milliseconds) to allow smooth exploration of the data.

Example

- Imagine a dataset that includes the hospitalization records of patients.
- This database contains six attributes (features) for each record which is visualized using a pixel-based method.
- The records are ordered by patient age in a square spiral, where younger patients in the outer parts and older patients in the centre.
- Here each pixel represents one data record.
- Each pixel's colour shows a different value for a particular attribute (like the number of hospital visits for a patient).

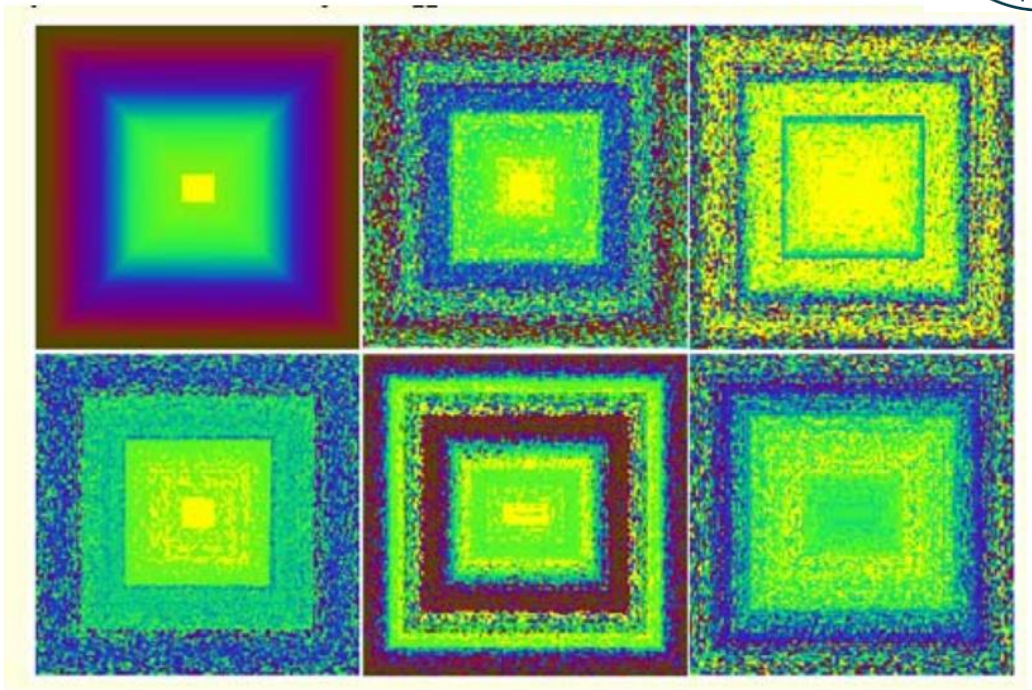


Figure 1: Pixel-based representation of database with six attributes per record. Records are arranged in a square spiral showing relationships among variables.

- The image visually tells you that younger patients had fewer hospital visits (shown by brown pixels), while older patients had more (yellow pixels).

- The other squares might show other variables ,such as office visits , medication etc.

- This method can squeeze a million records into a manageable space on the screen.

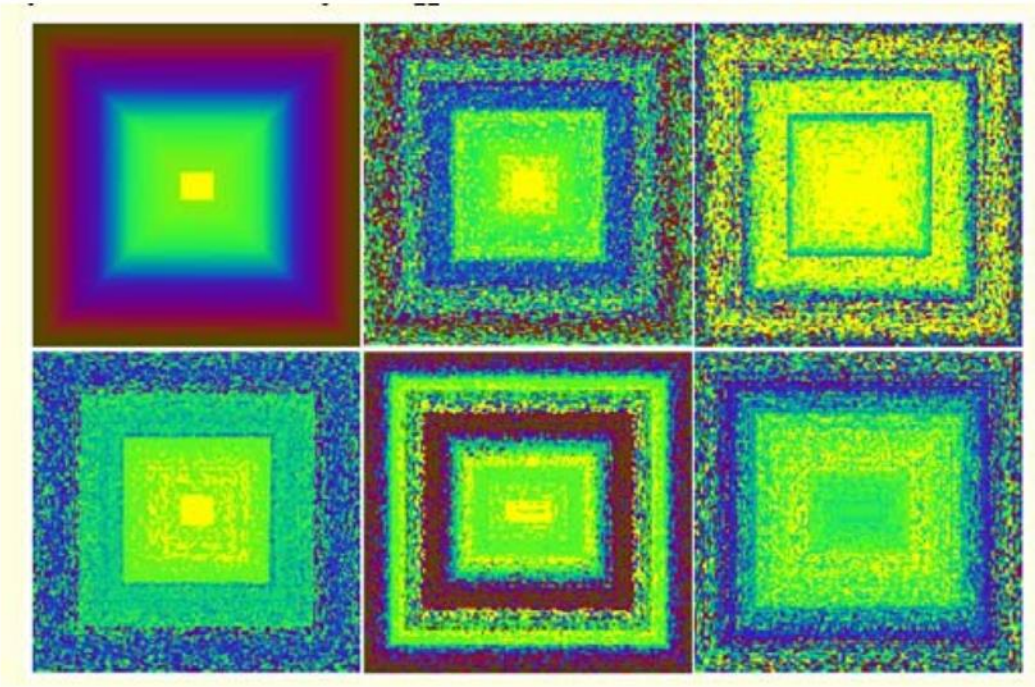


Figure 1: Pixel-based representation of database with six attributes per record. Records are arranged in a square spiral showing relationships among variables.

Limitations of Atomic Visualization

- **Overplotting**：When there are too many data points, they can overlap, leading to what is called "overplotting." This makes it hard to see individual data points, especially in dense areas.

  Example: In a scatterplot with millions of data points, many points might cluster together, causing them to overlap and blend, making it difficult to distinguish individual values or patterns

- **Screen Resolution Limits**: Atomic visualizations are constrained by the resolution of the display. Even with high-resolution screens, there are only limited pixels available which limiting the number of data points that can be visualized clearly.

  Example: If you have a screen resolution of 1600x1200 pixels, you can only display 1.9 million distinct data points (one per pixel). Trying to visualize more than that will result in a loss of clarity, as multiple data points will have to share the same pixel.

CONCLUSION OF ATOMIC VISUALIZATIONS

- While atomic visualizations are powerful for representing individual records and providing detailed views, they face significant limitations when it comes to large-scale datasets. Overplotting, screen resolution constraints, performance issues all become major challenges as data sizes grow.

- To handle very large datasets, other techniques like aggregation, density plots are also used.

AGGREGATE VISUALIZATION

- One Marker per Thousand Records
- Moving from Mega-pixel to Giga-pixel displays
- Brute Force Method - Directly addresses the problem of limited screen space

- **Technical Setup** -
  - Tiling multiple flat-panel monitors (50 or more) together
  - A single computer handles user interactions and input
  - Multiple computers are used to drive the displays themselves

AGGREGATE VISUALIZATION

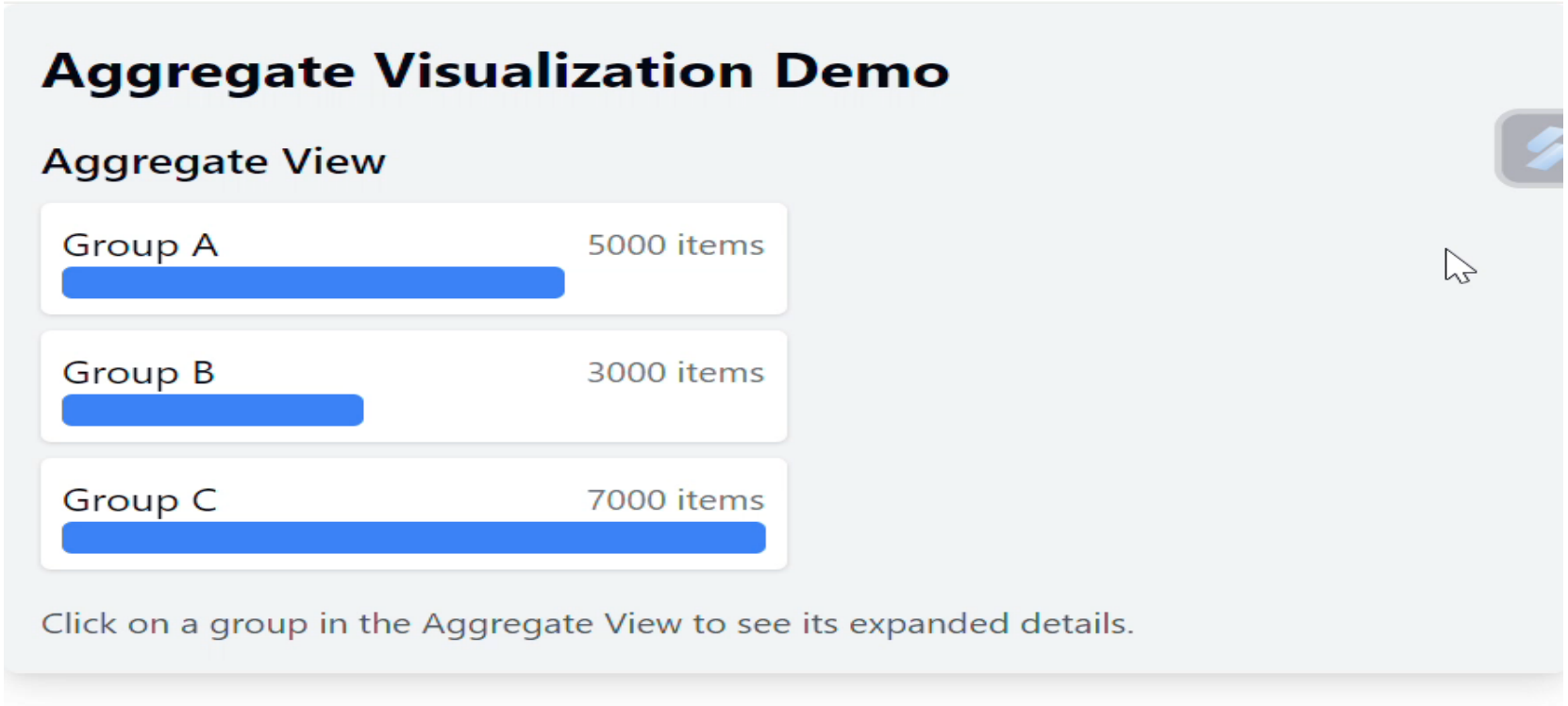
Is this method worth the cost incurred ?

AGGREGATE VISUALIZATION

- Demerit of Brute Force Method –
  - Difficulty in seeing the whole image
  - Identifying individual pixels – From large Giga-pixels
  - Physical constraints faced by user

- **A more attractive method**- Squeeze the billion record information into million pixel and view it on a common display
  - Analysing user needs
  - Suggesting novel way to aggregate data – **Aggregate Visualization**
  - Clicking on an aggregation marker will cause an expansion in place, but more effectively it will display its components in a coordinated window

AGGREGATE VISUALIZATION

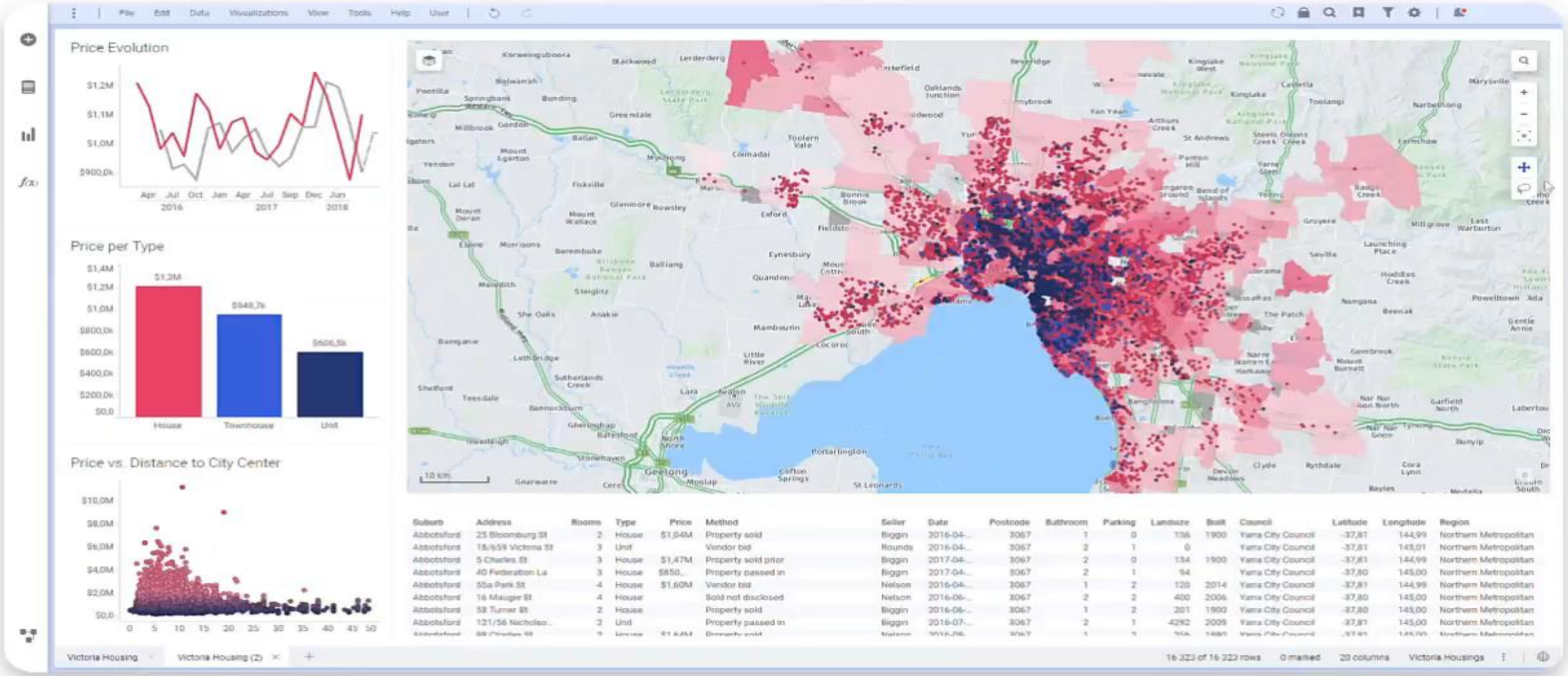


Main Slide: Aggregate View and user when clicks on it – Independent Coordinated Windows

AGGREGATE VISUALIZATION

- Coordinated Window Approach :
  - Well established in Geographical information visualization
  - A document database can be explored by seeing an overview by year and topic
  - Commercial tool Spotfire – Uses the similar strategy that supports visualization with million markers
  - New Feature enables User to select markers to initiate “on-demand” database retrieval – displayed in coordinated windows

GEOGRAPHICAL INFORMATION VISUALIZATION

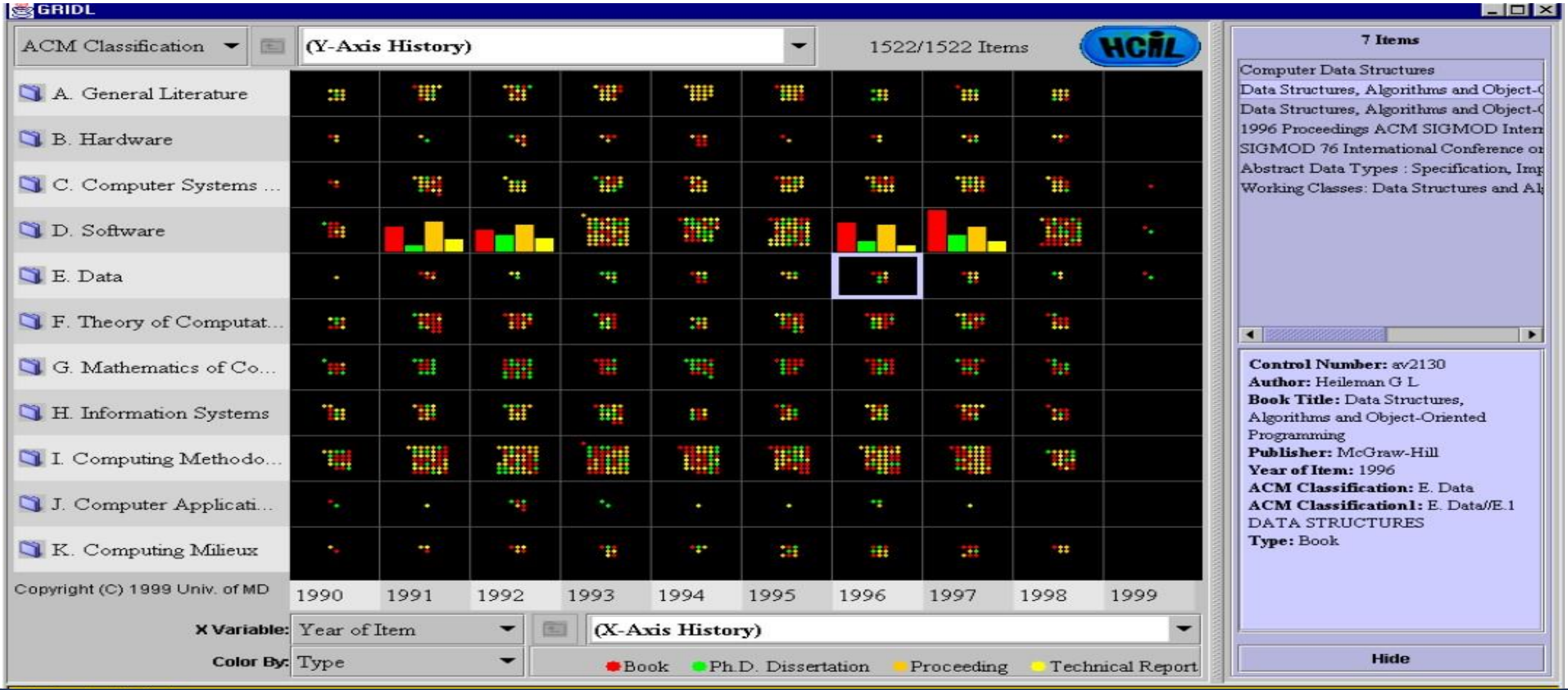


GRIDL – Graphical Interface for Digital Libraries

- Digital library search – usually textual list 10-20 items per page
- Can we view several thousand search result in 1 page?

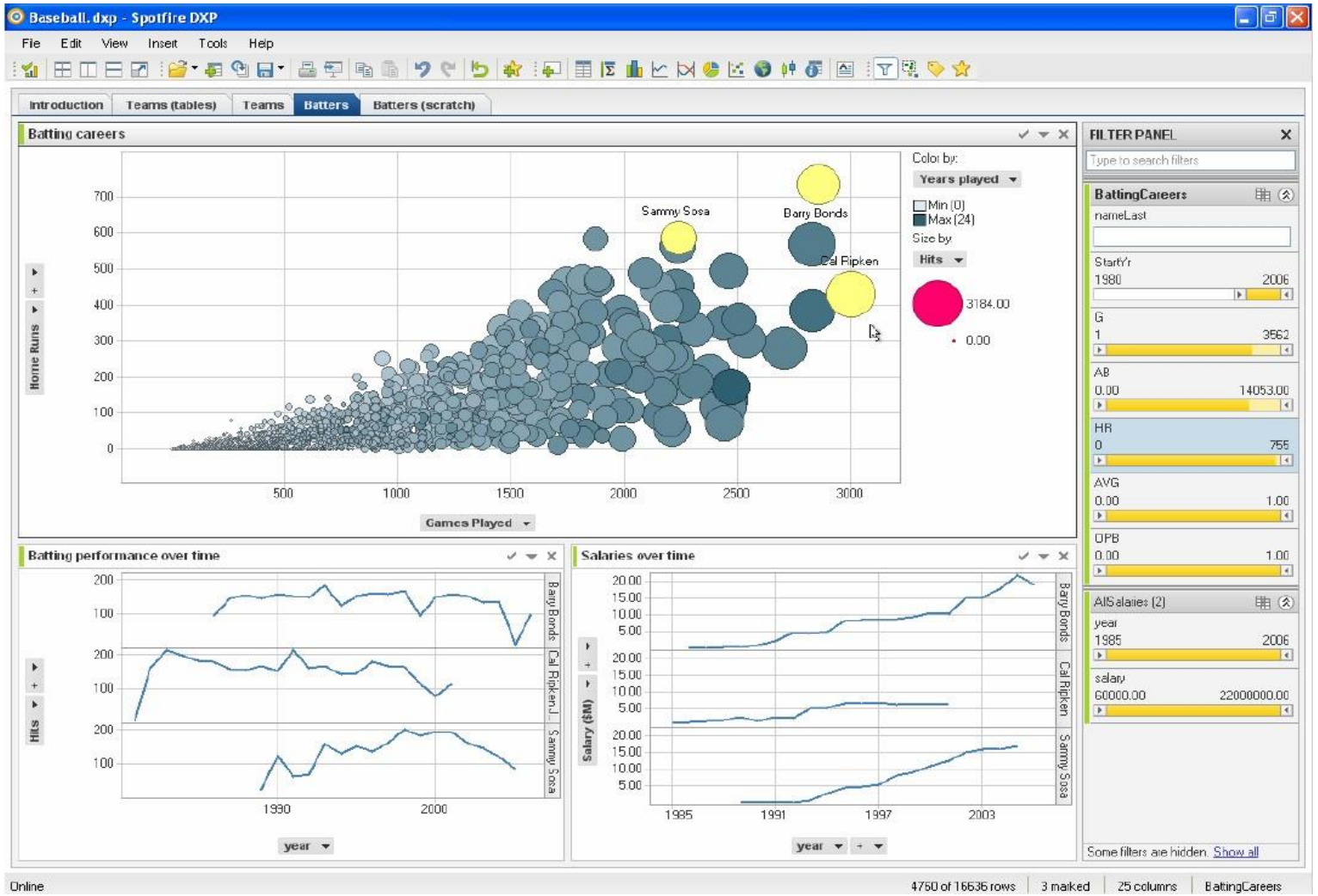
- YES – By GRIDL
  - Simplified 2D display that uses categorical (Y) and hierarchical (X) axes
  - Users appreciate the meaningful and limited number of terms on each axis
  - At each grid point of the display we show a cluster of color-coded dots or a bar chart
  - Users see the entire result set and can then click on labels to move down a level in the hierarchy

EXPLORING DOCUMENT DATABASE



ANALYZE BASEBALL STATISTICS - SPOTFIRE

- Multiple Coordinated Views: Batting Careers, Batting performance over time and Salaries Over time
- Each dot represents a player, with size and color encoding additional variables
- Notable players like Sammy Sosa and Barry Bonds are labeled
- Interactive Filtering: A filter panel on the right allows users to dynamically adjust the data shown
- Filters include date ranges, player statistics (G, AB, HR, AVG, OPB), and salaries



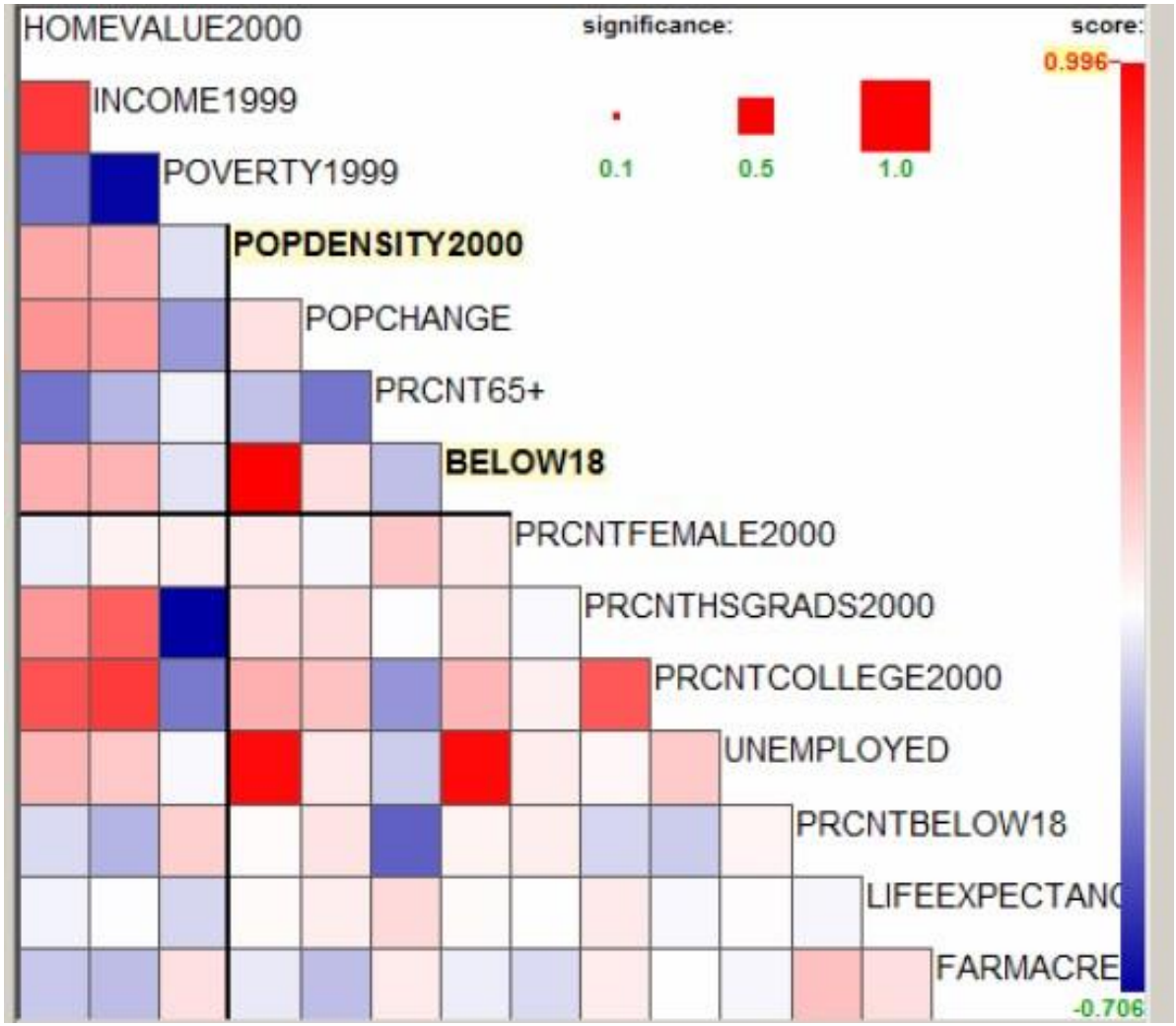
MULTI-VARIATE DATABASES

- Imagine the large group of people at ANTARAGINI 24
  - If we fly a drone over the entire area of Pro-nite ground and capture the total number of people attending - YES doable
  - Can we find how many males and females?
  - Can we find out how many are 60+ age?
- Traditional method – Histogram with age in years on the X -axis with a vertical bar indicating percent of people in each age group

- What are Attributes ?
  - Like Age, Gender……are all individual attributes
  - **Histograms** are effective for most data types, from binary (male/female, YES/NO) to categorical variables (drama/action/mystery/etc.)
  - What if we are given a data with n – records and k attributes ?

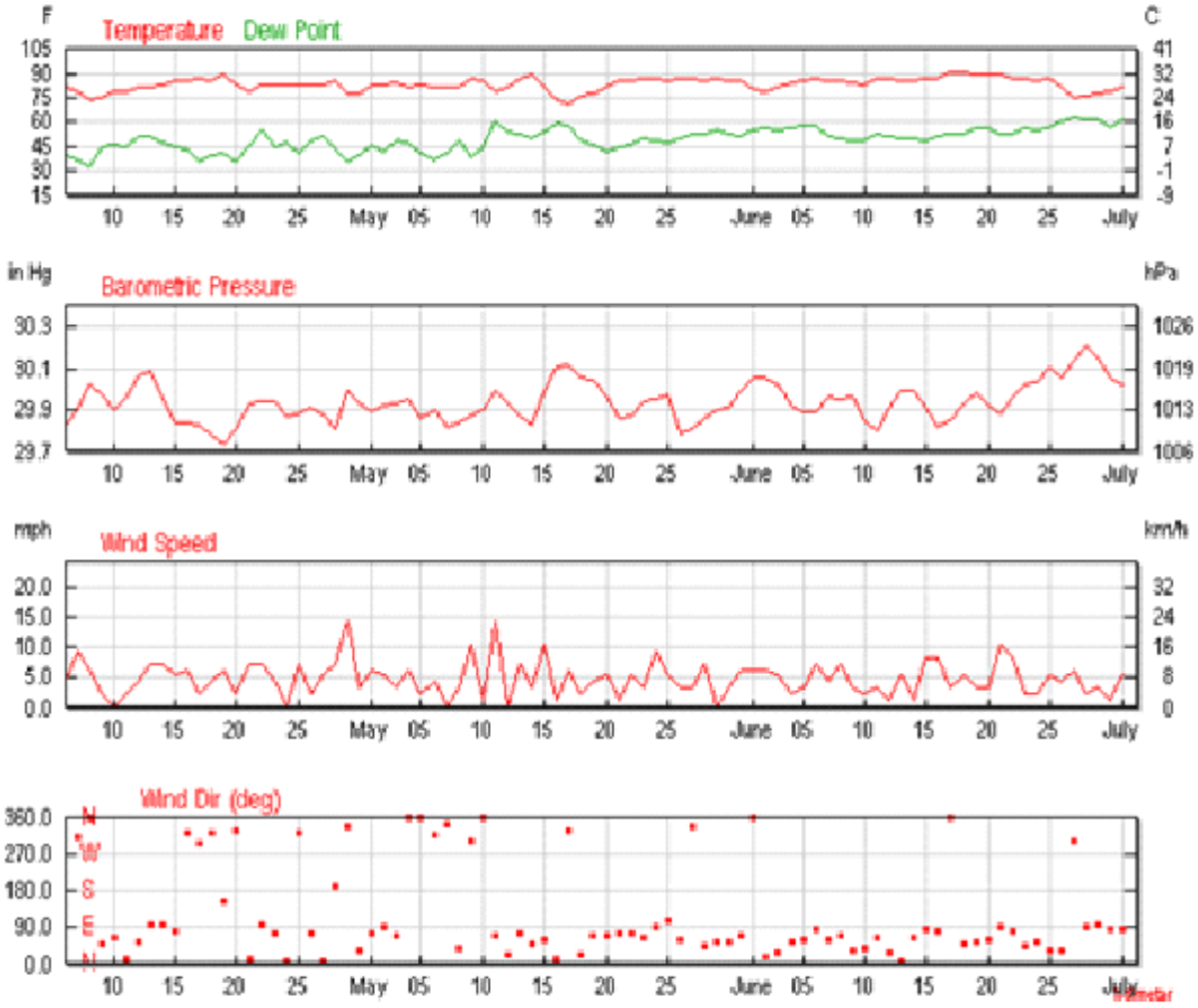
HIERARCHICAL CUSTERING EXPLORATION

- Database with n- records and k – attributes
- For analysis - pair-wise relation between attributes
- User would want to understand k (k-1)/2 pairwise relation
- Even for n = 1 billion – there may be 10 – 100 attributes
- HCE – Strategy of ranking strength of feature
- Rank by Feature Framework-
  - Each cell represents the correlation between two variables
  - Color Coding:-
    - Red indicates strong positive correlations
    - Blue indicates strong negative correlations
    - Lighter shades indicate weak correlations
    - White – No correlation
  - **14** Attributes of **3128** US states [14x13/2 =91]



TIME SERIES - CLUSTERING

- Time series database – grow large in number of time points (daily, weekly, monthly etc.)
- Environment sensor – capture Temp, Humidity, Pressure every minute for a year = 5,25,600 data points for each variable , 1000 sensor = well over billion data points
- Solution by Aggregation – Aggregation by sensor or month is useful by clustering into smaller number of **meta time series**

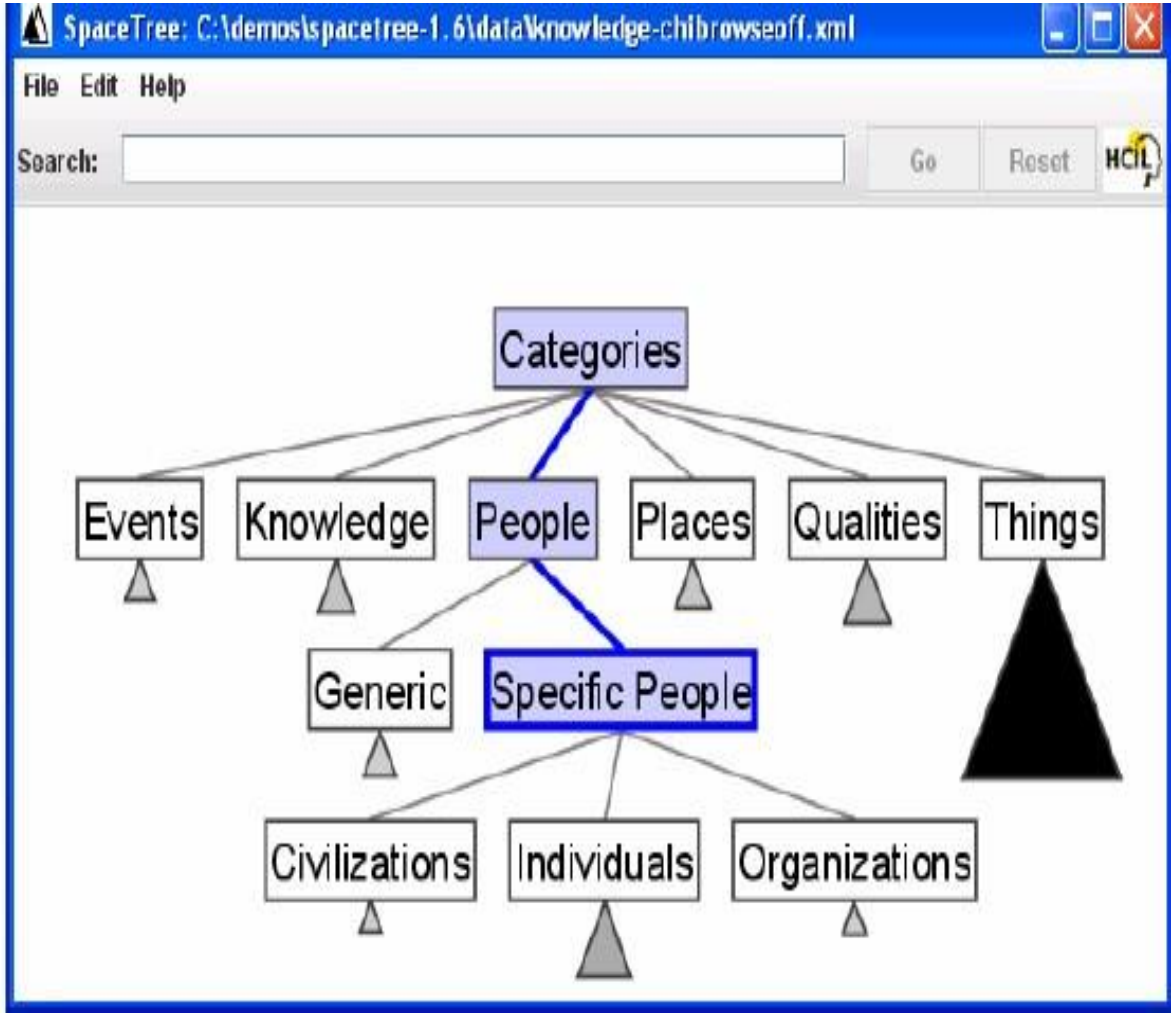


TREE STRUCTURE

- Tree structures are widely used due to their ability to represent hierarchical relationships

- Aggregation and representation:
  - Higher levels can effectively summarize lower levels
  - First few levels of a tree can often provide a good overview of the entire structure.

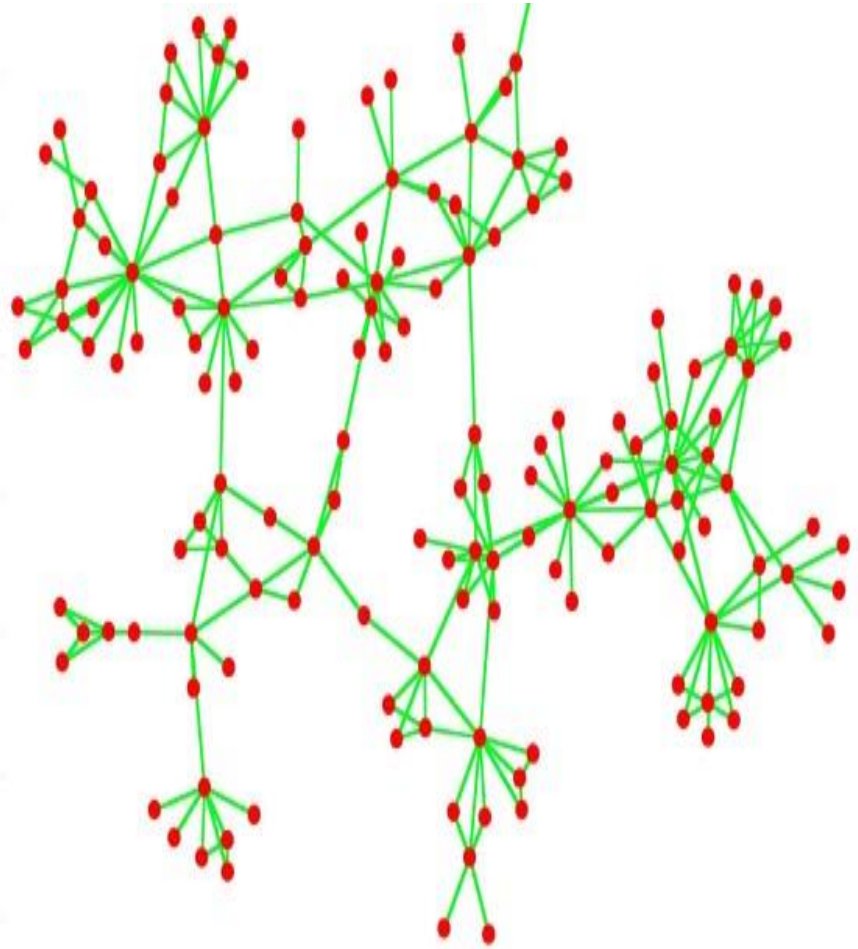
- **Space Tree** –Initially root and 1st level node shown, user can open lower levels “On Demand”
- **Darkness-** No. of Nodes & **Height-**No. of levels



NETWORKS – COARSENING

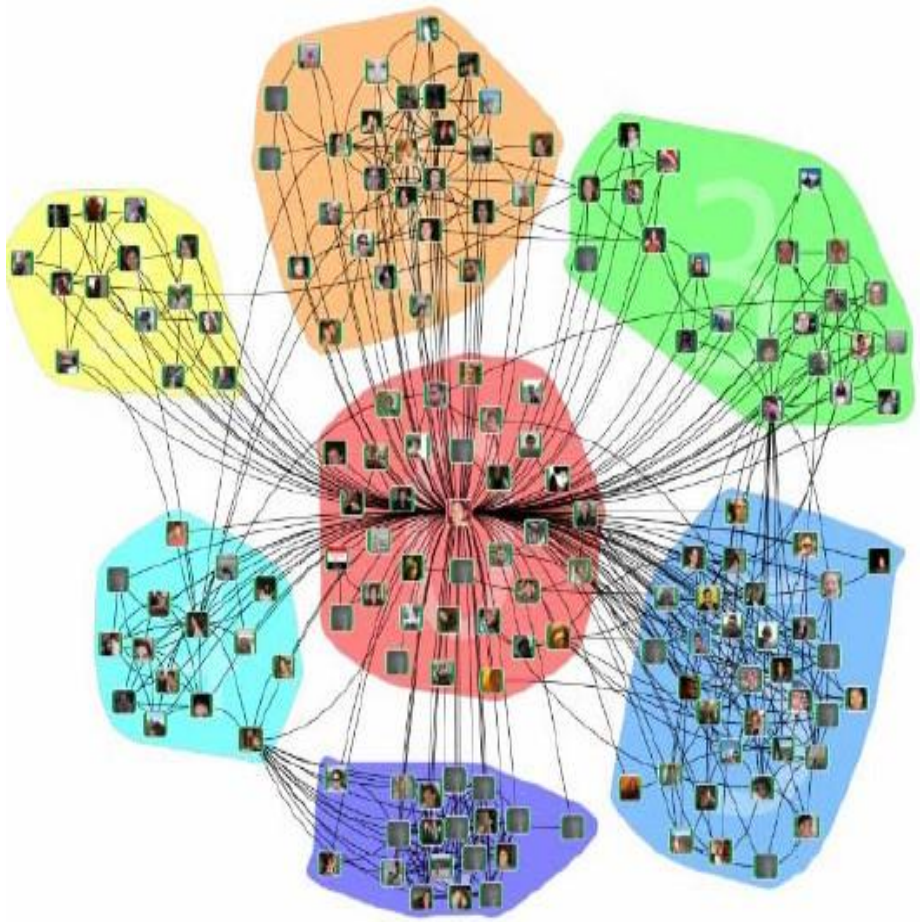
- Coarsening of network –
  - 152-node network is a simplified version of a much larger network of 46,480 node
- Node representation:
  - Red Dots – Represent nodes in coarsened network
  - Green edge – connection between coarsened nodes

- **GreenMax** –claimed to be able to scale to multi-million network.
- **Information Preservation- V**isualization still conveys important structural information about the original network, such as major clusters, hubs, and the overall connectivity pattern.



NETWORKS – COMMUNITY STRUCTURE

- Community Structure of Network – Network divided into 7 different communities represented by 7 colours
- Node representation: Each black square – individual nodes i.e facebook user here in this picture.
- The nodes within each community are clustered together, indicating stronger connections within the community.
- The red community at the center appears to be the most connected to other communities – nodes having max. connection – influential people in facebook
- Each community can be replaced by a single aggregate node

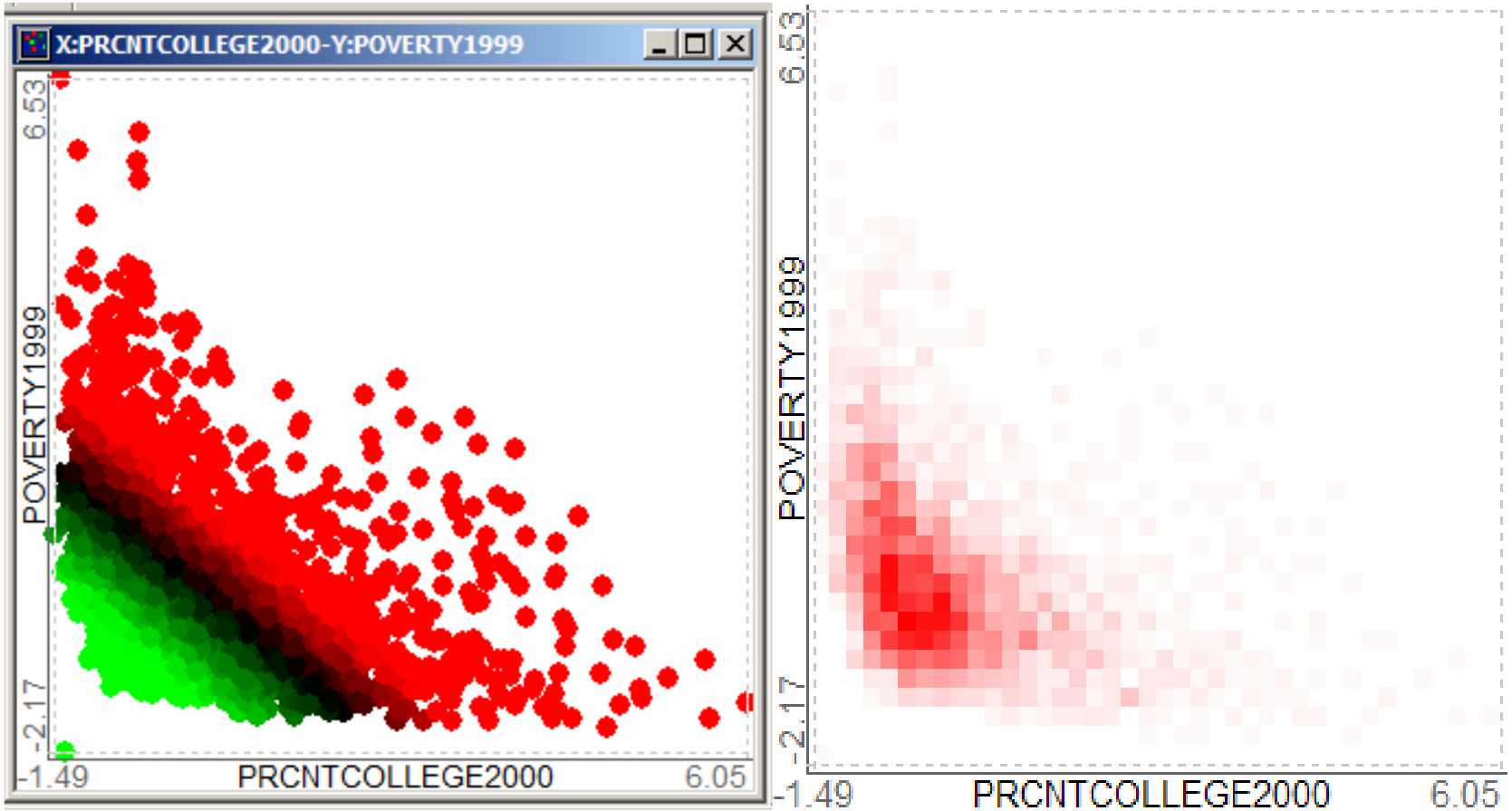


Why Density Plot ?

- Especially only dealing with complex datasets
- Handles Over plotting
- Shows Data Distribution
- Effective for Large Datasets
- Useful in Multivariate Data Analysis
- Highlights Clustering and Relationships

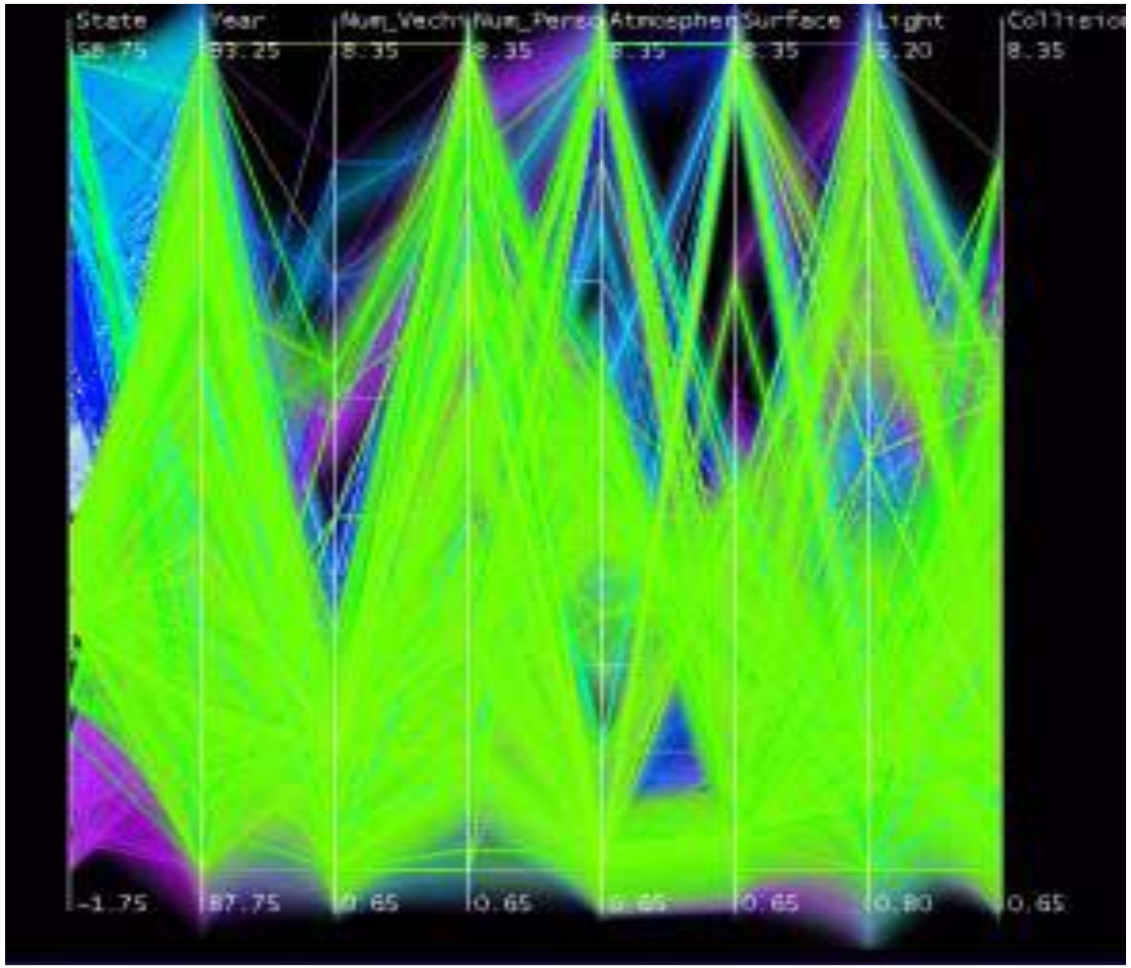
Density Plot Data Visualizations

- Graphical representation
- Distribution of data points across 2 D Scatterplot
- The scatterplot uses **color coding** to highlight data density
- This visualization helps users quickly spot patterns and anomalies within the dataset, particularly focusing on areas of high density, which are of interest for further exploration.



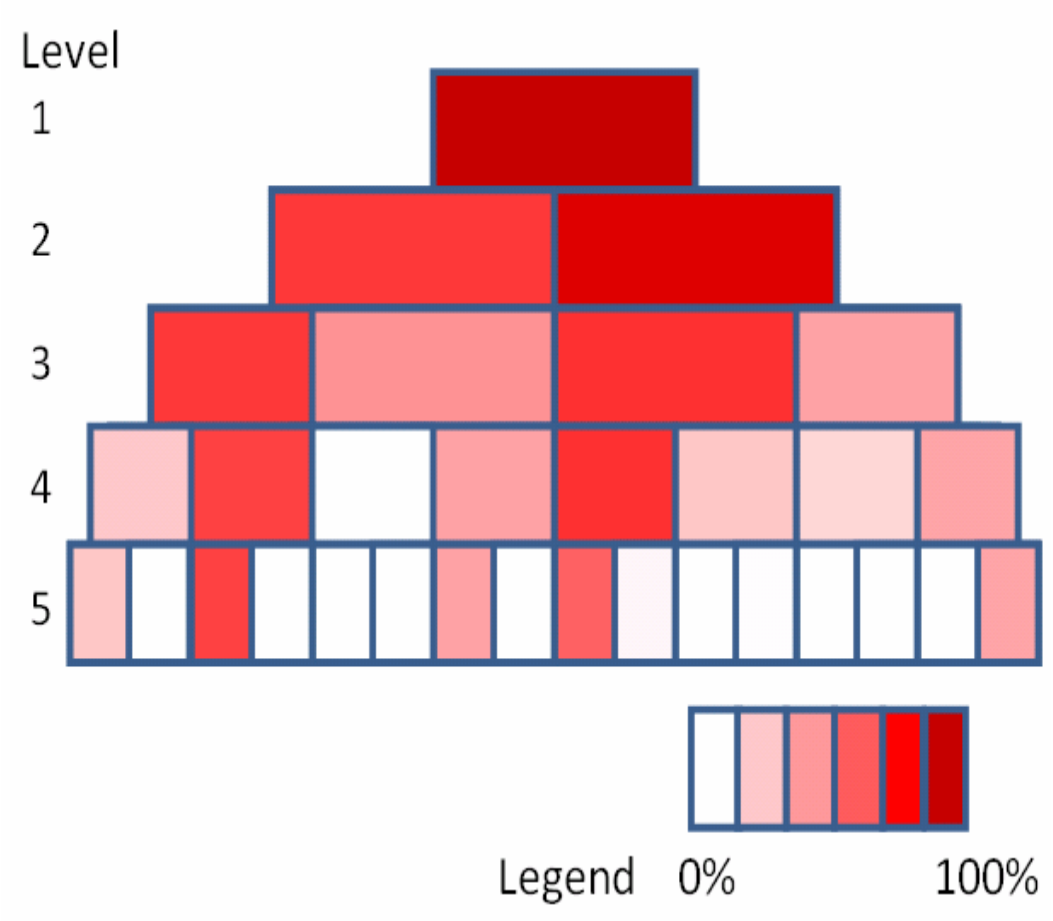
Density Plot : Multivariate Data

- 2-D projections use scattergram
- 3-D developed for time series data
- This type of visualization useful when we try to find correlations and patterns in large and complex datasets across multiple dimension
- It is commonly used in statistical analysis, machine learning, and exploratory data analysis



Density Plot : Tree maps

- Tree maps can display density by aggregating subtree counts/attributes
- This visualization technique useful for identifying, hierarchical dataset are most densely populated
- By focusing darker region user can spot areas of interest



Conclusion

- Visual exploration benefits are increasingly recognized, raising user expectations for exploring large databases
- Innovative interface design is likely more beneficial than gigapixel displays
- Meaningful aggregate visualizations promote sense-making with low display complexity
- Aggregation markers representing thousands of records can be organized to guide user clicks
- Density plots provide novel possibilities, particularly for user with a statistical inclination

THANK YOU