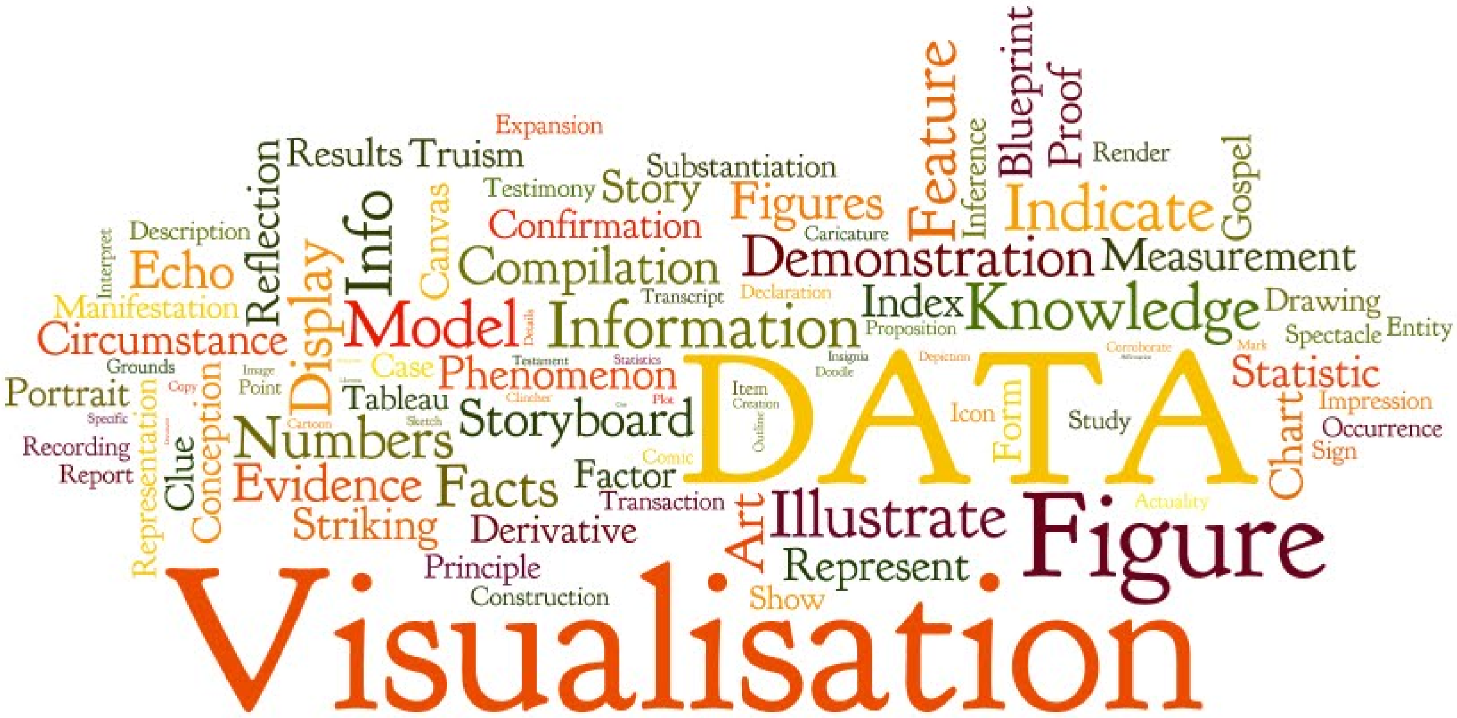


13. Visualization Techniques for Classification & Clustering

Prof. Tulasi Prasad Sariki
SCSE, VIT, Chennai

www.learnersdesk.weebly.com

KDD Process

- Selection
  - Obtain data from all of sources
- Preprocessing
  - After selecting the data, clean it to make sure it is consistent
- Transformation
  - After preprocessing the data, analyze the format/amount of data
- Data Mining
  - Once the data is in a useable format, apply various algorithms based upon the results trying to be achieved
- Interpretation/Evaluation
  - Finally, present the results of the data mining step to the user, so that the results can be used to solve the business need at hand

Importance of Data Visualization

- The final step in the KDD process :
- Highly dependent on the Data Visualization technique
- Bad/inappropriate technique may result in misunderstanding
- Misunderstanding may cause an incorrect (or no) decision

It is important to consider that the KDD process is useless if the results are not understandable

Suggested Direction

- Need to determine techniques that balance simplicity with completeness
- If this can be done for non-expert users
  - Simplicity & Completeness → Understanding
  - Understanding → Trust
  - Trust → more use of KDD/DM
  - Result will be:
    - Better business value
    - Higher ROI

Common Visualization Techniques

- Visualization techniques dependent upon
  - The type of data mining technique chosen
  - The underlying structure and attributes of the data

Classification

- Decision Trees
- Scatter Plots
- Axis-Parallel Decision Trees
- Circle Segments
- Decision Tables

Clustering

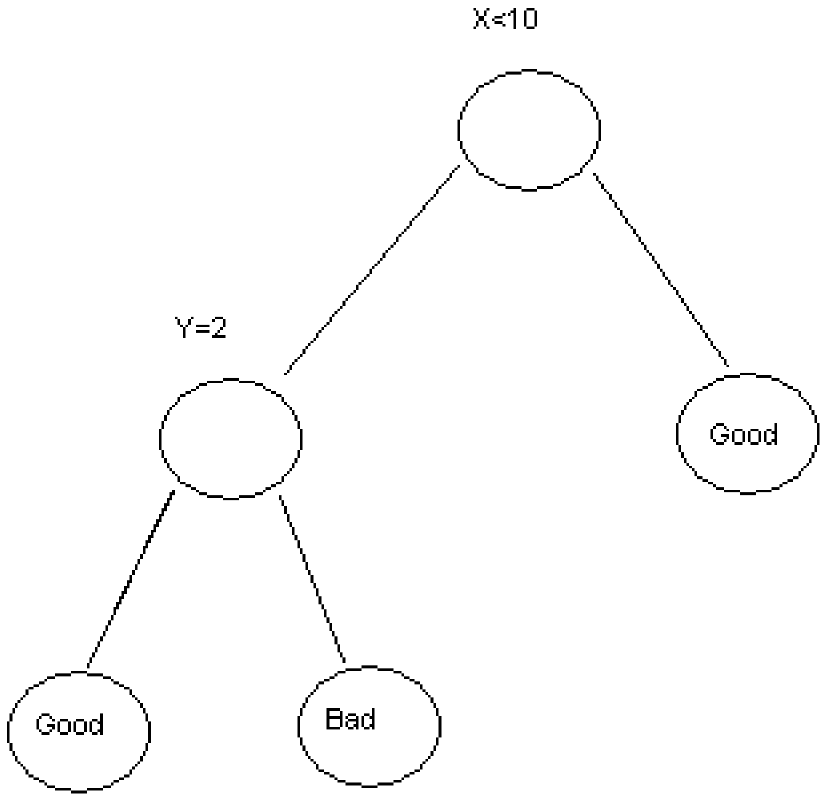
- Scatter Plots
- Dendrograms
- Smoothed Data Histograms
- Self-Organizing Maps
- Proximity Matrixes

Classification

Decision Tree

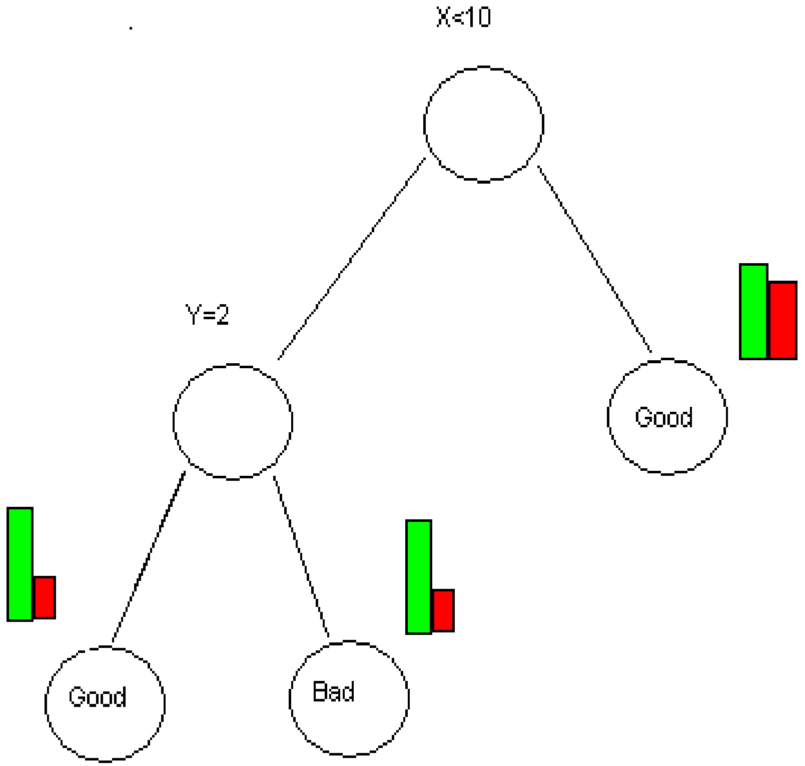
Information limited to

- Attributes
- Splitting values
- Terminal node class assignments



Decision Tree with Histograms

- Data mining rarely classify 100% of the data correctly:
  - Include the success of properly classifying the data - histogram added for each terminal node
  - Percentage of data that was classified correctly/incorrectly
  - Assists users in determining if the classification is ‘good enough’

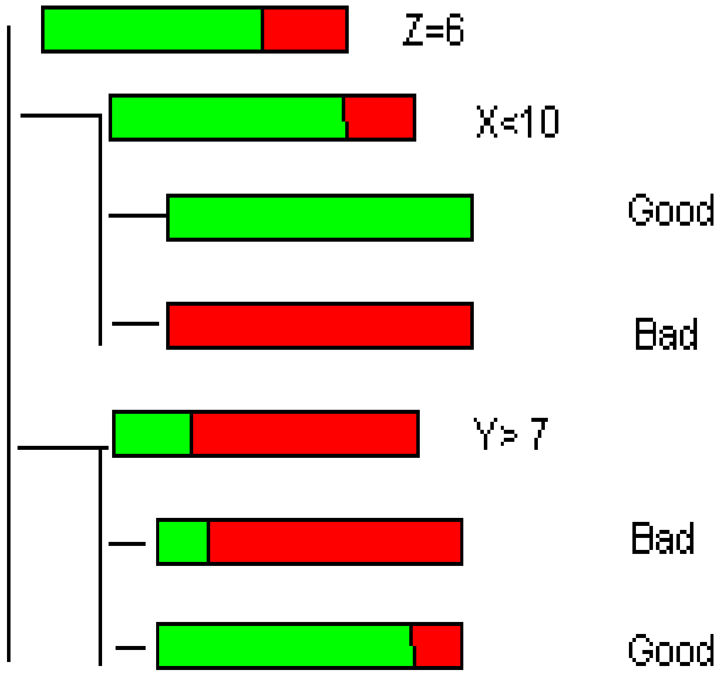


Decision Tree Different Format

Vertical representation - allows for easy user interaction

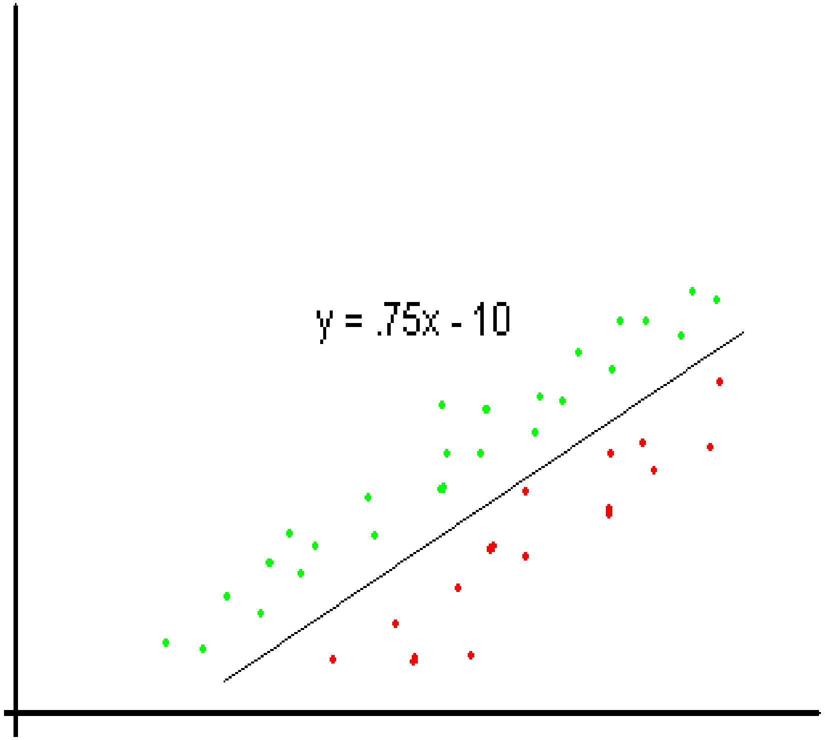
Combines the split points and classification accuracy - compactly

Key difference - colors are matched with a specific classification



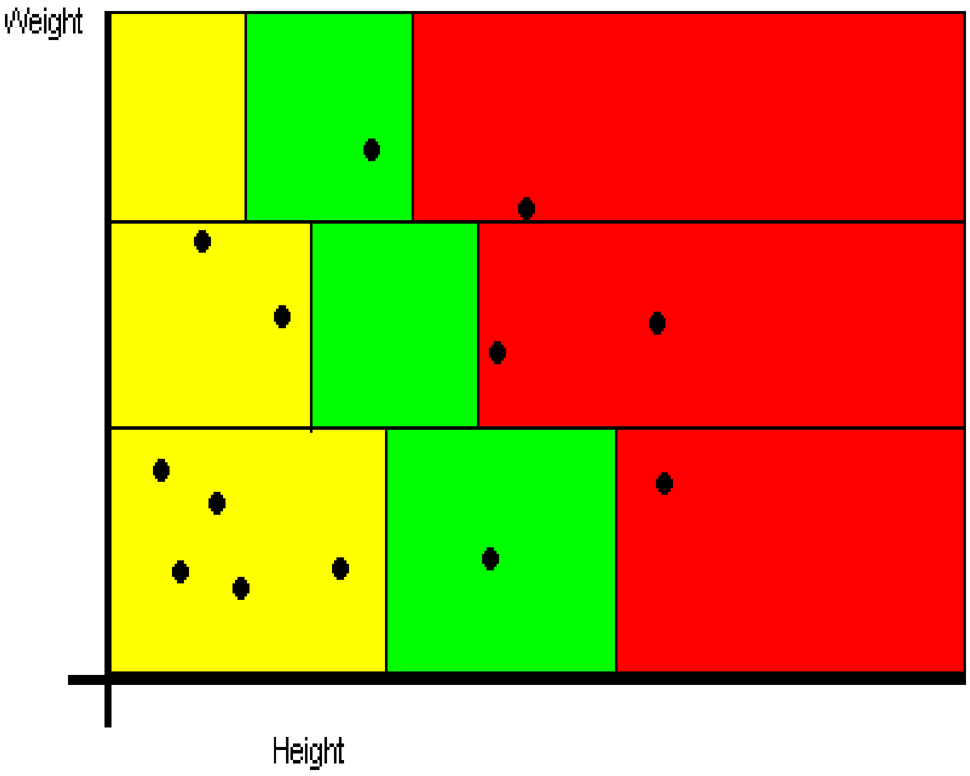
Scatter Plot with Regression Line

- Excellent way to view 2-dimensional data
- Familiar to anyone who has taken high-school algebra
- Regression lines provide descriptive techniques for classification



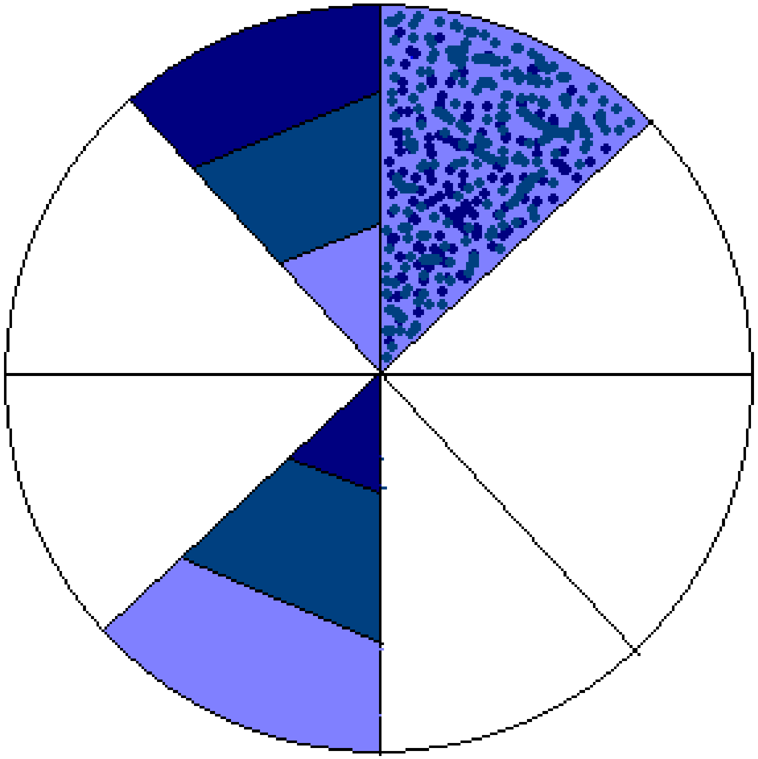
Axis-Parallel Decision Tree

- Combination Scatter Plot and Decision Tree
- Areas divided in parallel regions on the axis
- Well suited for classification problems with two attribute values
- High visibility into the impact of outliers



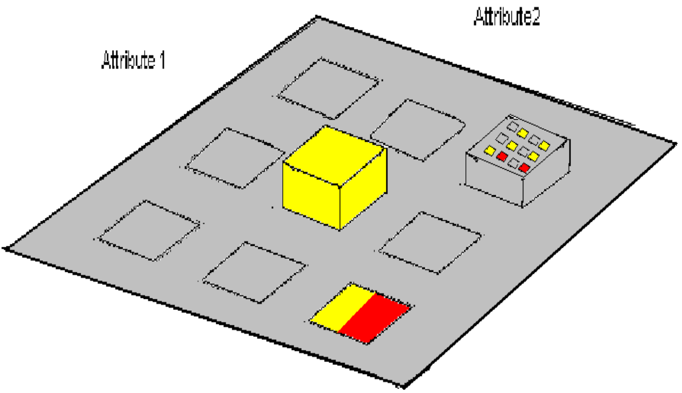
Circle Segments

- Multi-dimension data
- Maps dataset with *n* dimensions onto a circle divided by *n* segments
    - Each segment is a different attribute
    - Each pixel inside a segment is a single value of the attribute
    - Values of each attribute are then sorted (independently) and assigned a different colors based upon its class



Decision Table

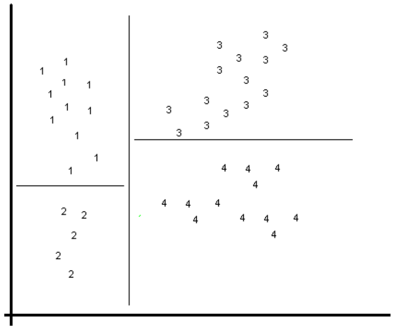
- Interactive technique
- Maps attribute data to a 2D hierarchical matrix
- Levels can be drilled down - another set of attributes
- Height of a cell conveys the number of data entities
- Cells color coded
  - Neutral color → no data in that intersection point
  - Color coded by class (percentage)



Clustering

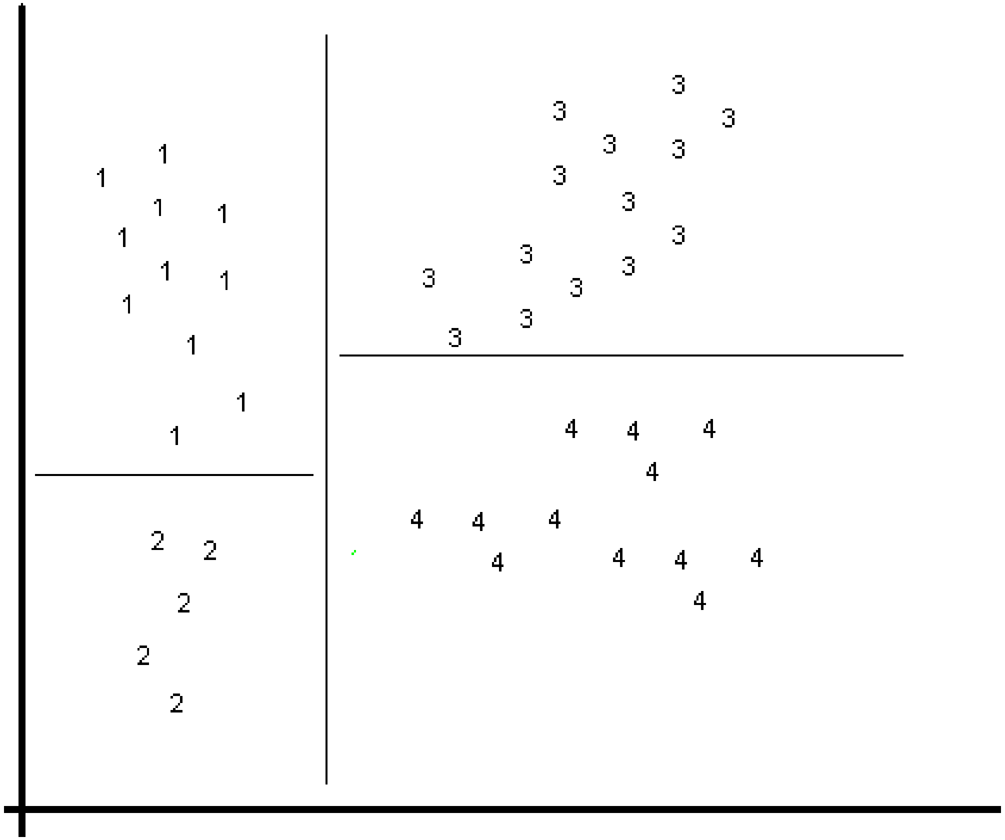
Scatter Plot

- Extensions include, displaying points in:
  - Various sizes and colors to indicate additional attributes
  - Shading of points to introduce a third dimension
  - Using different brightness levels of the same color to represent continuous values for the same attribute
  - Using various points or classification identifiers (i.e., numbers, symbols)
  - Using various glyphs to display additional attributes

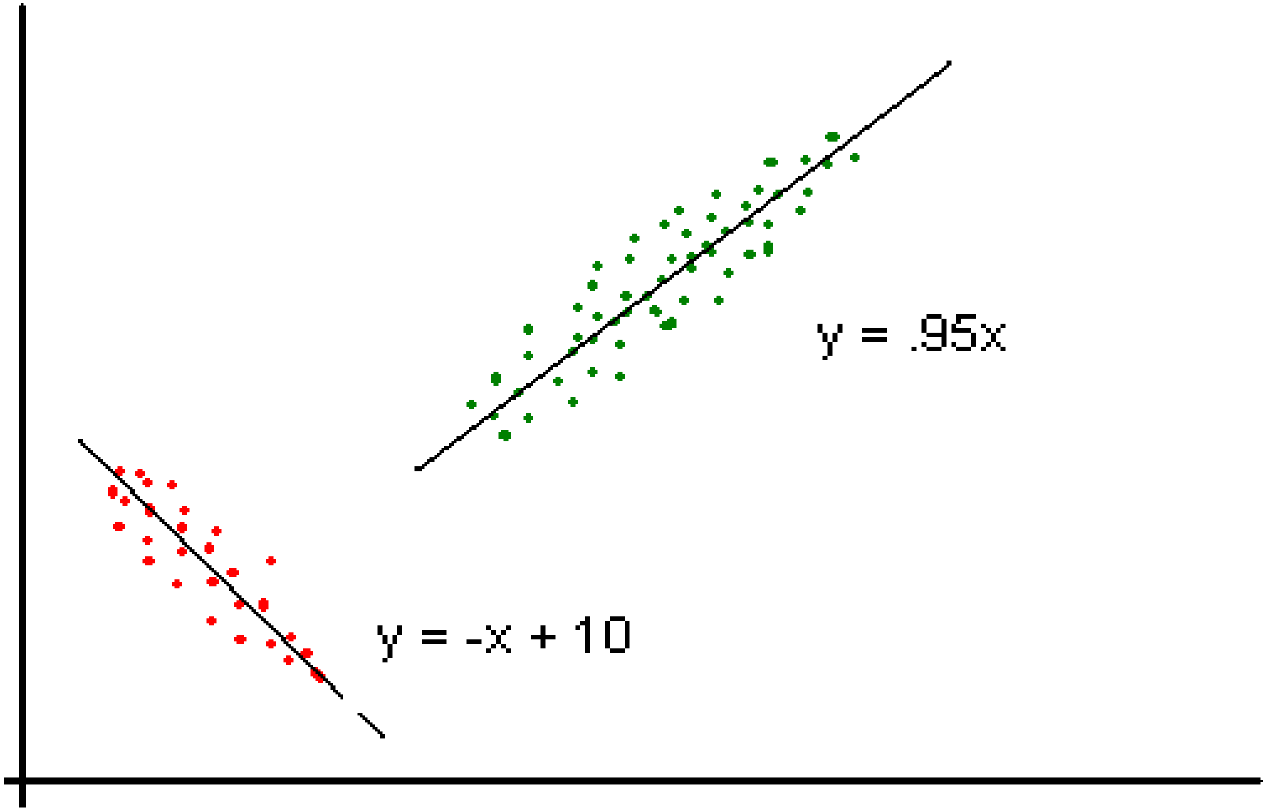


Scatter Plot

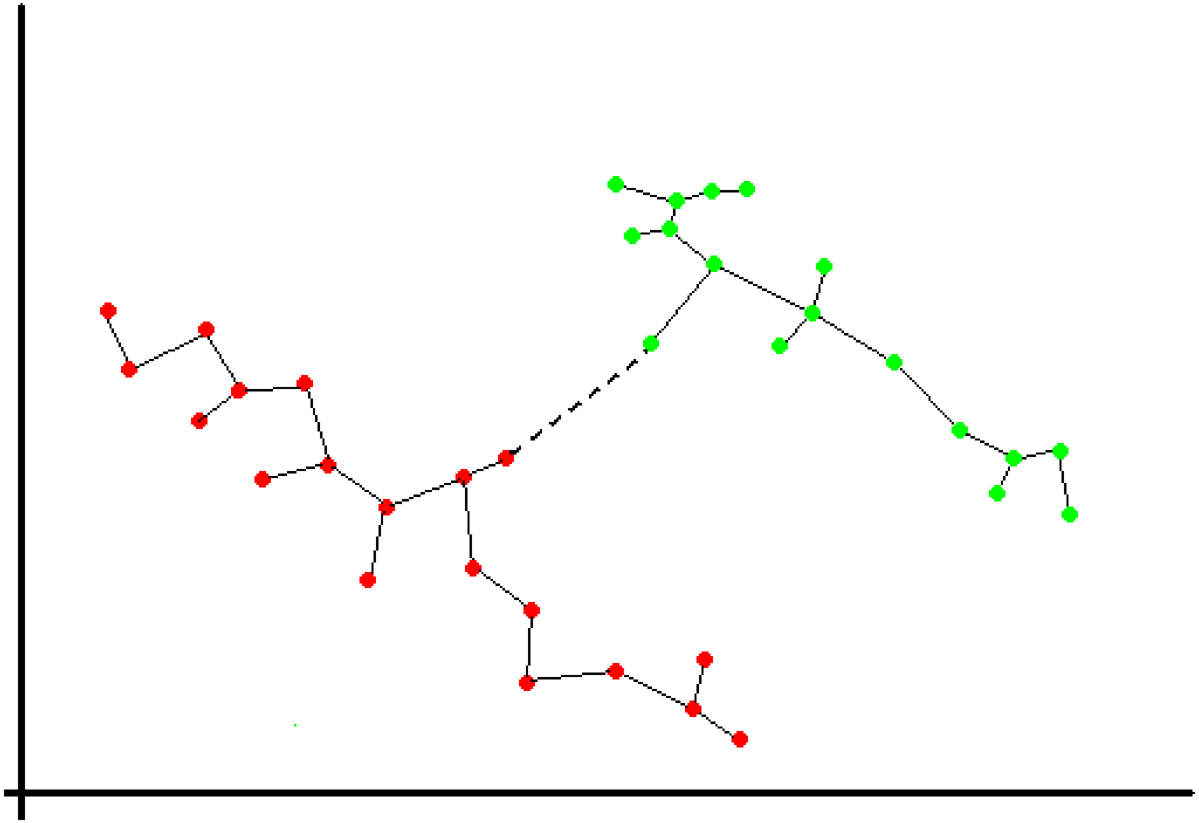
- Map decision trees on top of scatter plots to describe clusters



Scatter Plot with Regression Lines

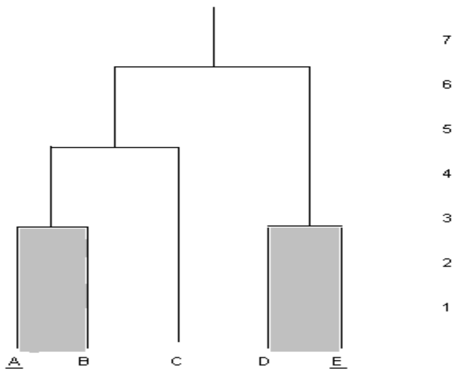


Scatter Plot with Min Spanning Tree



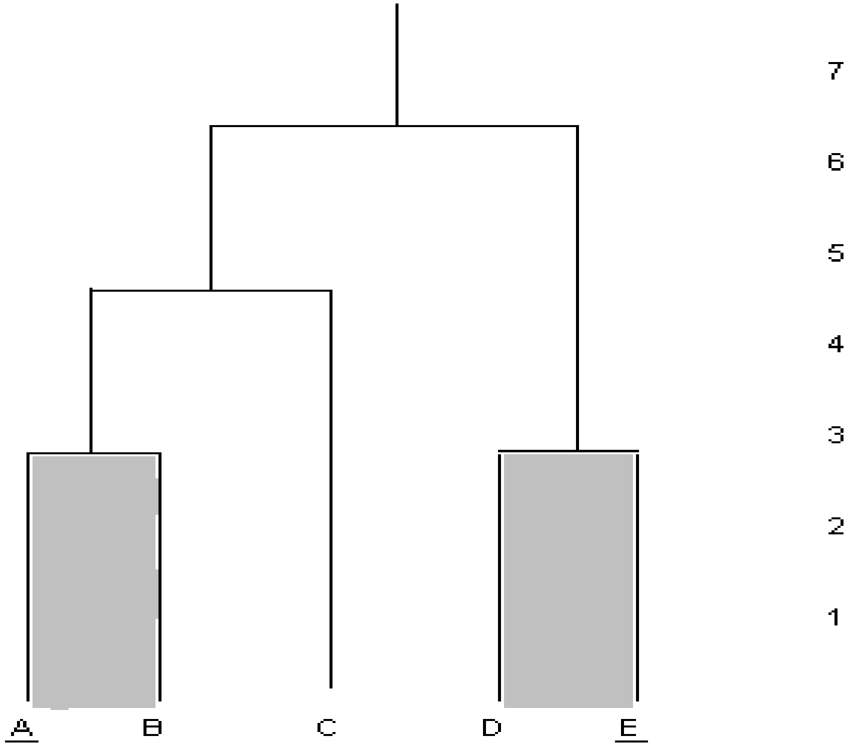
Dendrogram

- Intuitive representation - hierarchical decomposition of data into sets of nested clusters.
- From an agglomerative perspective:
  - Each leaf - a single data entity
  - Each internal node - the union of all data entities in its sub-tree
  - The root - the entire dataset
  - The height of any internal node - the similarity between its ‘children’.



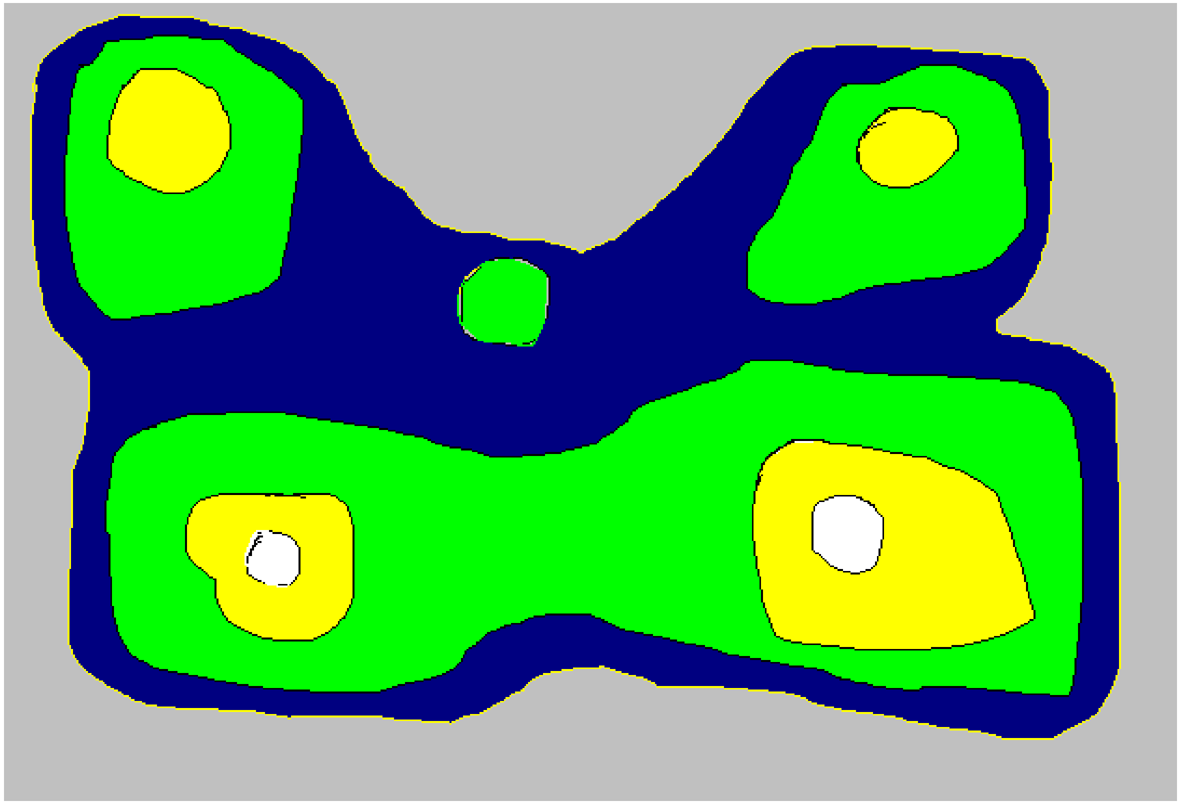
Dendrogram with Exemplars

- The “most typical member of each cluster” [Wishart99]
  - Underlined labels of the leafs
  - Done in combination with shading to identify the clustering level



Smoothed Data Histogram

- Represents data on a ‘display map’
- Similar data items are located close to each other
- More defined the clusters – lighter colors



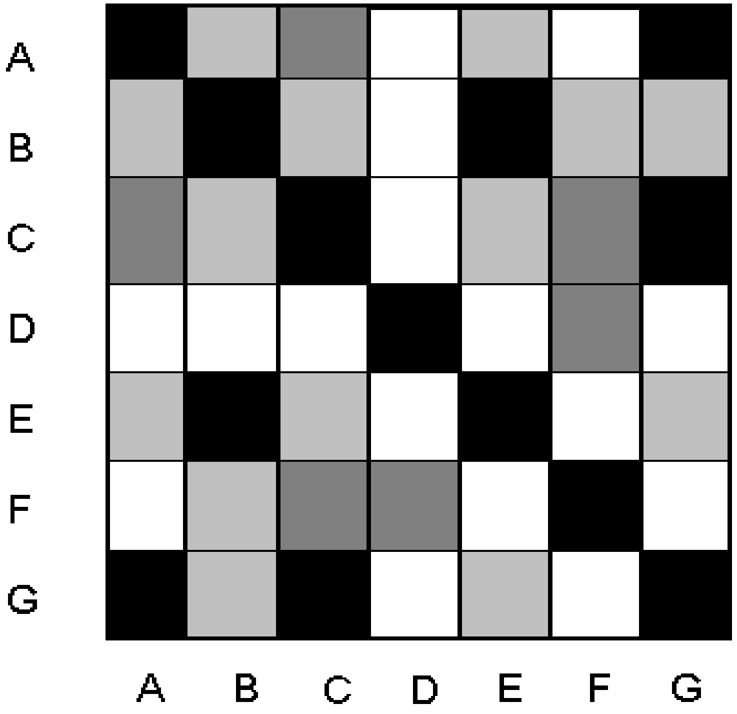
Self-Organizing Map ‘Grid’

- Source of Smoothed Data Histogram
- Numbers indicate most ‘common’ cluster

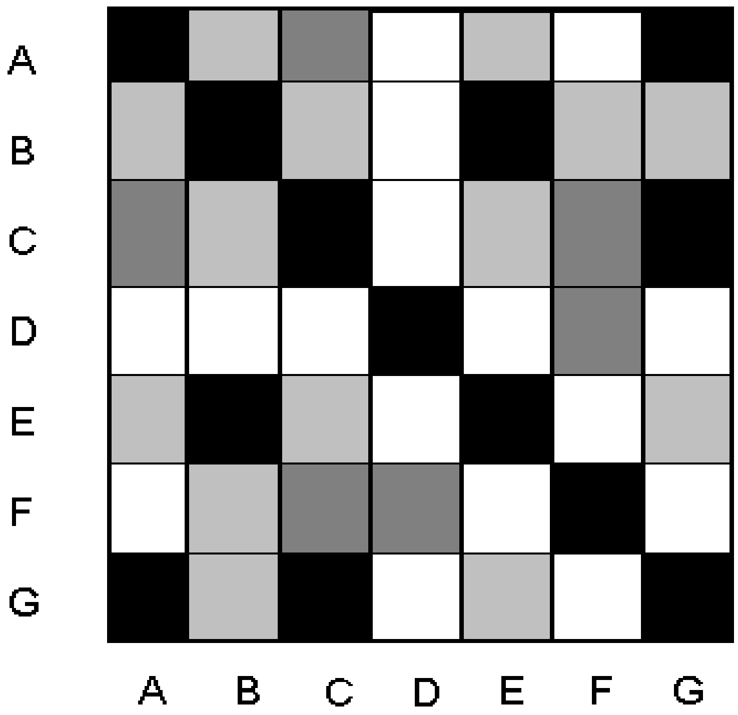
1					*5*	
2	*3*	*2*		*5*	*6*	*5*
2	*2*	*2*	*4*	*5*	*5*	*5*
7	*1*	*1*	*1*	*5*	*7*	
7	*8*	*7*	*7*	*7*	*10*	*7*
7	*9*	*7*	*7*		*11*	*7*
	8			*7*	*10*	*7*

Proximity Matrix

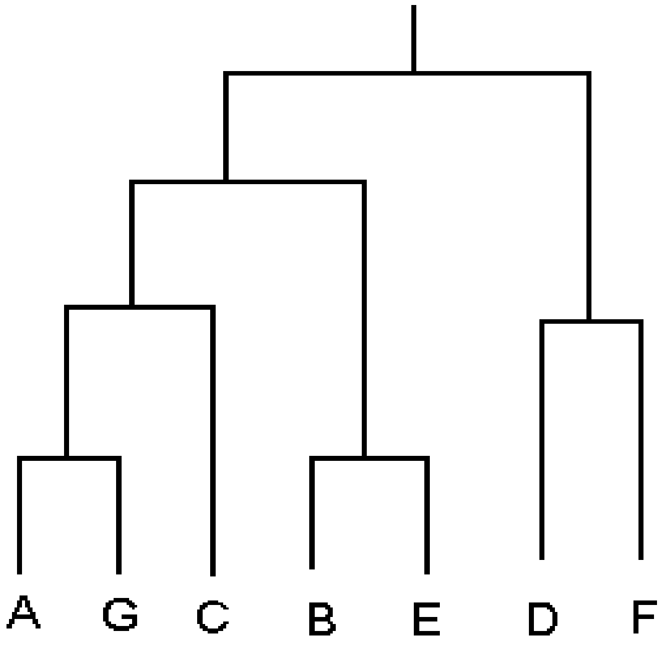
- Graphically display the relationship between data elements
- Usually symmetric, but can be sorted by the strength of relationships



Proximity Matrix and Dendrogram



	A	B	C	D	E	F	G
A	■	░	▓	□	░	□	■
B	░	■	░	□	■	░	░
C	▓	░	■	□	░	▓	■
D	□	□	□	■	□	▓	□
E	░	■	░	□	■	□	░
F	□	░	▓	▓	□	■	□
G	■	░	■	□	░	□	■



Summary

- Data visualization techniques are extremely important for understanding the KDD process
- A balance of simplicity and completeness is important
- The techniques discussed allow average users to understand the results of the KDD process
- Understanding → KDD results to be interpreted/trusted by non-expert users → extending the business value
- If data visualization techniques do not establish a high level of trust in the KDD process, the process will fail

Thank You