Lecture 2

There are 4 types of Data sets

Record:

- Relational Records
- Data Matrix (Numerical Matrix and Crosstabs)
- Document data (Term Frequency Vector)
- Transaction Data

Graph and Network:

- The Web
- Social or Info Nets
- Molecular Structures

Ordered Data Sets:

- Video data
- Temporal data (time-series)
- Sequential data
- Genetic sequence data

Spatial, IMG and MM:

- Spatial data (Maps)
- IMG data
- Video Data

Attribute: a data field representing a feature of a data object.

- **Nominal**
  - Categories, states, names
- **Binary**
  - Nominal but with only two states
    - Symmetric Binary
      - Both outcomes equal (Gender)
    - Asymmetric Binary
      - Not equally important (Medical Test)
- **Ordinal**
  - Values have meaningful order but magnitude is unknown (sizes, grades, ranks)
- **Numeric**
    - Quantity (integer or real)
    - Interval Scaled
      - Measured on a scale of equal sized units with no true zero (Temperature, dates)
    - Ratio Scaled
      - Has a zero point (Temperature in Kelvin, length, counts, monetary values)

Discrete Attribute: Has a finite or countably infinite set of values, sometimes represented as integer variables (Special Case: Binary Attributes)

Continuous Attribute: Has real numbers as values represented using Floating point variables (Weight, Height, Temperature)

Motivation: To understand data

- Central Tendency (Mean, Median, Mode)
- Variation
- Spread

Data Dispersion:

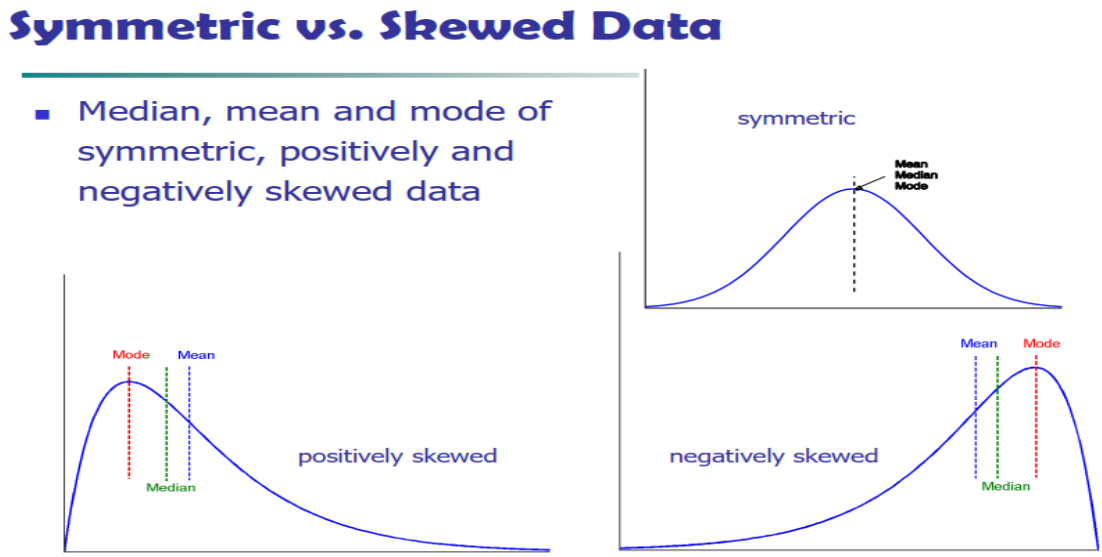
- Median, Max, Min
- Quantiles
- Outliers
- Variance

Numerical Dimensions (Sorted Intervals):

- Data Dispersion (Analyzed with multiple levels of precision)
- Boxplot or Q-Analysis on sorted intervals

Dispersion Analysis on Computed Measures:

- Folding measures into numerical dimensions
- Boxplot or Q-Analysis on transformed cube.



Quartiles: Q1 (25^{th} %), Q3 (75^{th} %)

Inter quartile range: IQR= Q3-Q1

Five NO. Summary: min, Q1, median, Q3, max

Outlier: a value higher or lower than IQR x 1.5

Variance

Standard Deviation: Square root of Variance

Box Plot: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually (Graphic display of 5 NO. summary)

Histogram: x-axis are values, y-axis are frequencies (Better than Boxplot)

Quantile Plot: each value Xi is paired with Fi

Quantile-Quantile Plot: graphs the quantiles of one univariant distribution against the corresponding quantiles of another

Scatter Plot: each pair of values is a pair of coordinates (Positive Correlation, Negative Correlation, No Correlation)

Data Visualization

- **Gain insight** by mapping data onto graphical primitives
- **Provide a qualitative view** of large data sets
- **Search for patterns, relations, irregularities**
- **Find interesting region and parameters** for more quantitative analysis
- **Provide Visual Proof** of derived computer representation

Visualization Methods:

- Pixel Oriented
    - For a data set of **X** dimensions create **X** windows on the screen
    - Values of records are mapped to **X** pixels
    - Colors of Pixels reflect Values
    - Can also be done in a Circle Segment
- Geometric Projection
    - Direct Visualization
    - Scatterplot and Scatterplot Matrices
    - Landscapes
        - Data must be transformed into a 2D spatial representation
    - Parallel Coordinates
        - N equally spaced axes parallel to the screen axis and correspond to attributes
        - Scaled to [Min : Max] range of attribute
        - Data items correspond to the polygonal line intersecting the axes
    - Projection Pursuit
    - Prosection Views
    - Hyperslice

- Icon Based
  - Chernoff faces
    - Display Values on a 2D surface (10 x 10)
  - Stick Figures
    - Two attributes mapped to axes, remaining attributes mapped to angle or length of limbs
  - Shape Coding
    - Use shape to encode info
  - Color Icons
    - Use Color to encode info
  - Tile bars
    - Use small icons to represent relevant feature vectors
- Hierarchical (Partitioning into Subspaces)
  - Dimensional Stacking
    - Partitioning of N-Dimensional attribute space in 2D subspaces, stacking.
    - Attribute Value ranges into classes, using the important ones on the outside
    - Ordinal Data w/ Low cardinality with no more than 9D
    - Must map Dimensions properly
  - Worlds within Worlds
    - Assign the function and two most important parameters to innermost world
    - All other parameters constant
    - N-Vision, Auto Visual

    - Tree Map
        - Screen-filling partitioning of the screen into regions depending on the attribute values
    - Cone trees
        - build a 2D circle tree that arranges its nodes in concentric circles centered on the root node
        - Up to 1000 nodes
        - Cannot avoid overlaps
    - InfoCube
        - 3D Technique where info is displayed as nested semitransparent cubes
- Visualizing Complex Data
    - Non-Numerical Data
    - Social Networks
    - Tag Cloud
        - Importance is by font size and color'

Similarity

- Numerical measure of how alike
- Range [0,1]

Dissimilarity

- Numerical measure of how different
- Min is often 0, upper limit differs

Proximity

- Refers to Similarity or Dissimilarity

Z-score = $\frac{x-\mu}{\sigma}$

X is the raw score, μ is the mean and σ is the standard deviation

- Minkowski Distance
  - Manhattan (power of P)
  - Supremum (Max of two points)
  - Euclidean (Power of 2)
- Cosine Similarity