Slide 3

Data visualization is the process of taking data and creating images the represent the data visually. So what makes a good data visualization? The three key aspects are data, function, and design.

- Data: understanding your data so that you are using the best type of visualization to represent it
- Function: data is accurately represented by visual
- Design: visual is appealing and impactful

For example, if I have a dataset with a hundred unique data points, I am not going to use a barplot to represent my data.

Slide 4

Graphic depicting balance between the three aspects

Slide 5

The definition I have highlighted is the one that I will be using when I mention high dimensional data.

- Microarray analysis: a DNA microarray is a tool to detect the expression of certain genes
- The Curse of Dimensionality: various phenomena that occur as the number of dimensions grows. For example, definitions of density and distance become less meaningful and organizing data becomes more difficult since the data can seem dissimilar.

Slide 6

- The goal of visualization is to create a visual that makes data easy to interpret and understand for readers. Since it is complex, high dimensional data can be difficult to work with.
- Preprocessing: an essential component to data mining. Optimizes data so that it is easier to analyze.
- Dimension reordering: rearranging existing dimensions to find structure
- Dimension reduction: collapsing or removing dimensions to establish structure

Slide 8

Clutter reduction is a method of preprocessing used to help find structure within the data. This makes it possible to create visualizations of the data that are easy to interpret.

- Clutter: crowded and disordered visual entities that obscure the structure in visual displays that can affect a reader's understanding of information
- General procedure:
    - Create an initial visualization of the data. This can differ depending on what type of clutter is in the data. There are many different algorithms to address various types of clutter, but the two I will discuss are parallel coordinates and scatterplot matrices.
    - Rearrange the order of the attributes to reduce clutter
    - Create a revised visualization

Slide 9

The clutter we are reducing with parallel coordinates is the number of outliers between neighboring dimensions. This is done by reducing the proportion of outliers to total data points by reordering the dimensions.

- Introduced by Alfred Inselberg, a mathematician and computer scientist, in 1985

Slide 10

Example of parallel coordinates using the Cars dataset.

Slide 11

In scatterplot matrices, clutter is a lack of structure within the matrices. This is reduced by reorganizing the dimensions based on measures of cardinality.

- Cardinality: the number of elements in a set
- Pearson Correlation Coefficient: measure of strength of linear relationship between two variables. The scale is [-1,1].

Slide 12

Example of Scatterplot Matrix using the Cars dataset.

Slide 13

- Nearest Neighbors: simple algorithm used to solve the Traveling Salesman problem
- The Traveling Salesman: problem in which a salesman starts in one city and must travel to others without visiting the same city twice

Slide 14

Example of SBAA using the Quadriped dataset

Slide 16

A reduction algorithm that builds off of SBAA is SBAR

- Eliminates attributes that are most similar to their neighbors in A
- Used when there are too many attributes cluttering a visual or not contributing to analysis
- Removes attributes from A based on similarity

Slide 17

Comparison of default order, SBAA, and SBAR on Quadriped dataset

Slide 18

- Used in image processing, natural language processing, genomic data, speech processing
- Looks similar to a clustering algorithm but because it maps to a lower dimensional space, it is a dimension reduction algorithm
- Plots are mostly used for exploratory analysis
- Takes longer to execute than linear dimensional reduction methods

Slide 19

Comparison of linear method (PCA) vs. non linear (t-SNE)

Slide 21

With high dimensional data, it can be difficult to do classical data exploration which starts from an initial visual representation. The following algorithms are designed to find effective and expressive visualizations.

- Classified data is data that has been grouped into classes

Slide 22

- Analyzes each pixel and computes a sum of the absolute difference between pairs of pixels and sums the result

M = number of classes

P = number of pixels

Slide 24

- S_k is a quality measure for each class based on cardinality
- Sum of all class quality measures

M = number of classes

P = number of pixels

The higher the value, the smaller the overlap between classes