Multivariate Analysis

Introduction to Multivariate Analysis

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Describe attributes of multivariate data visualization

Terms

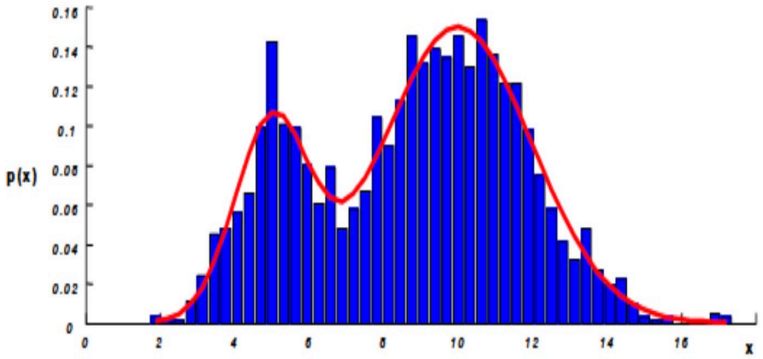
Multivariate Data

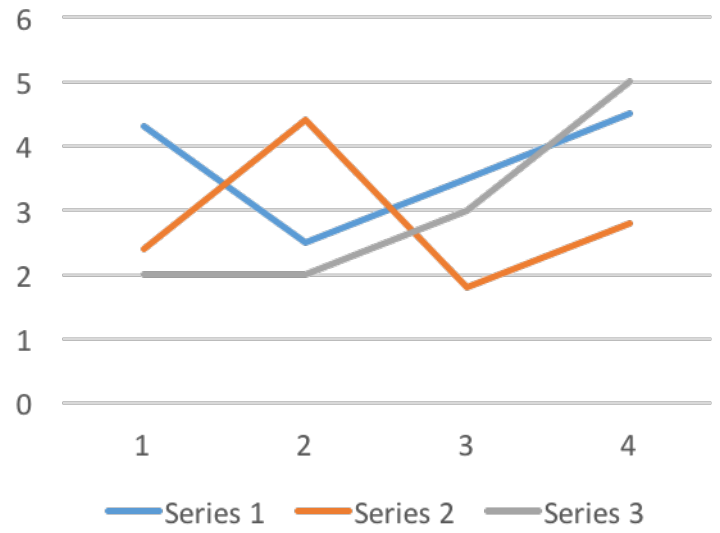
- Any statistical technique used to analyze data from more than one variable
- Used to process information in a meaningful way

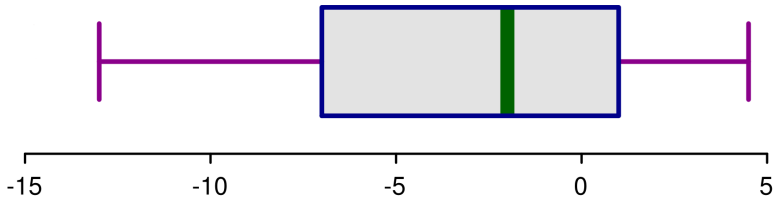
Curse of Dimensionality

- The more dimensions in a visualization, the less effective standard computational and statistical techniques become.

Univariate Visualization







Representation

What are the two main ways of presenting multivariate data sets?

- Directly (textually) – Tables
- Symbolically (pictures) – Graphs

How do we decide which to use, and when?

Tables

Use tables when:

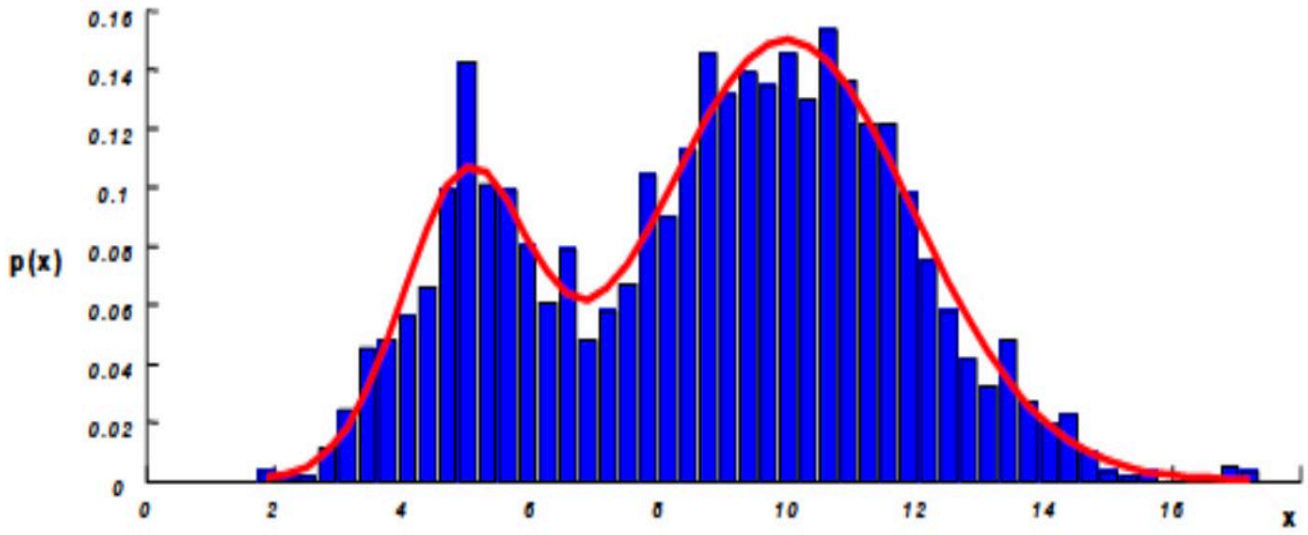
- The document will be used to look individual value
- The document will be used to compare individual values
- Precise values are required
- The quantitative info to be communicated involves more than one unit of measure

Category 1	Category 2	Category 3	Category 4
value	value	value	value
value	value	vale	valye

Graphs

Use graphs when

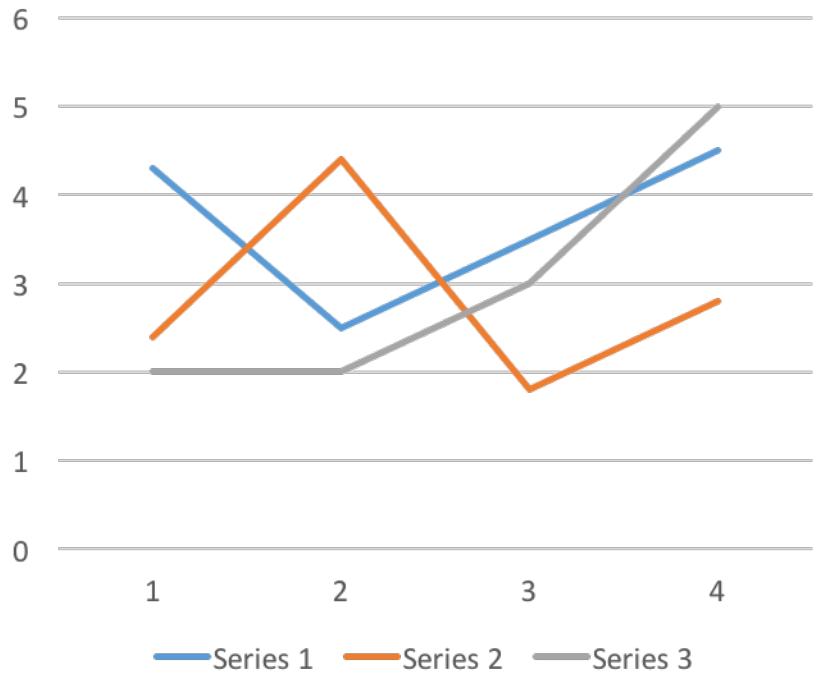
- The message is contained in the shape of the values
- The document will be used to reveal relationships among values



Graphs

Graph

- Visual display that illustrates one or more relationships among entities
- Shorthand way to present information
- Allows a trend, pattern or comparison to be easily comprehended



Task-Centric Graphing

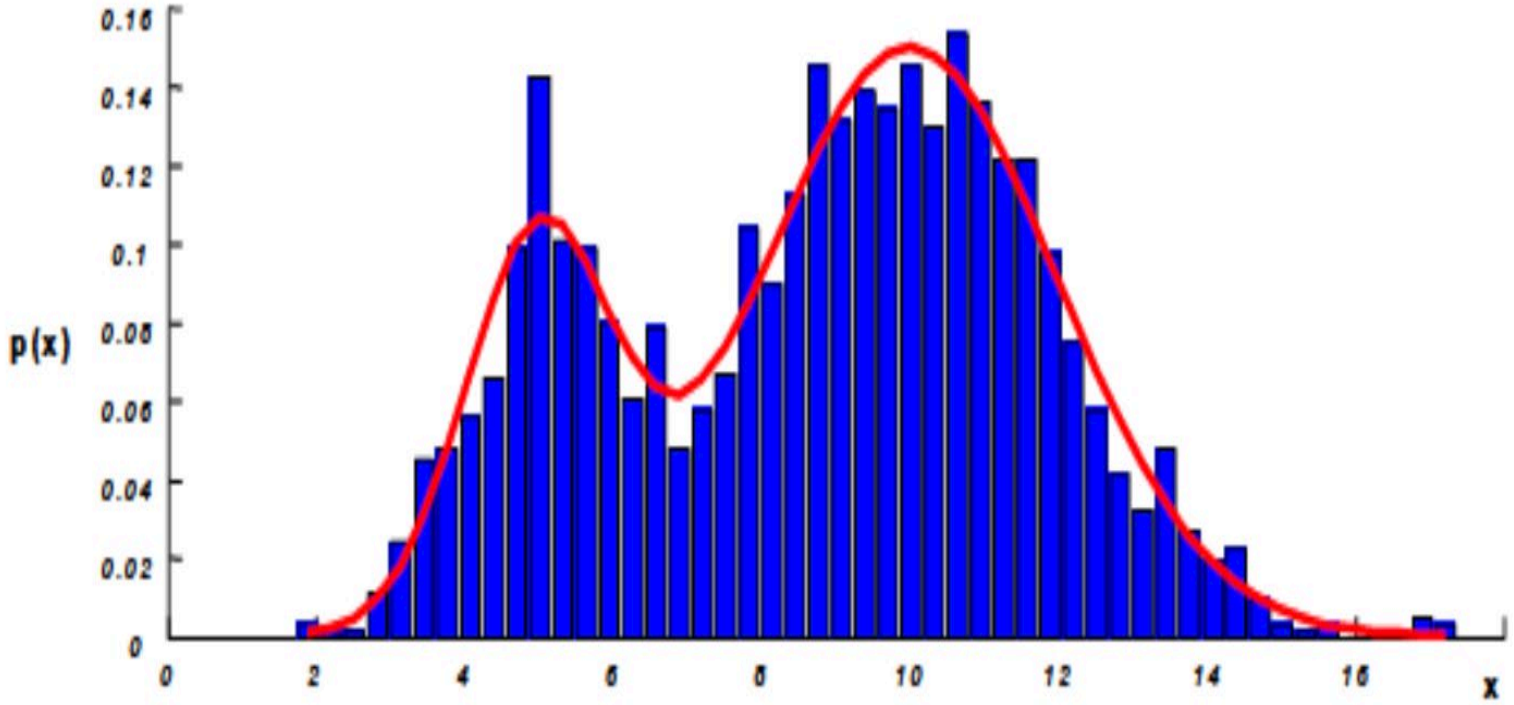
Why do you need a graph?

What questions are being answered?

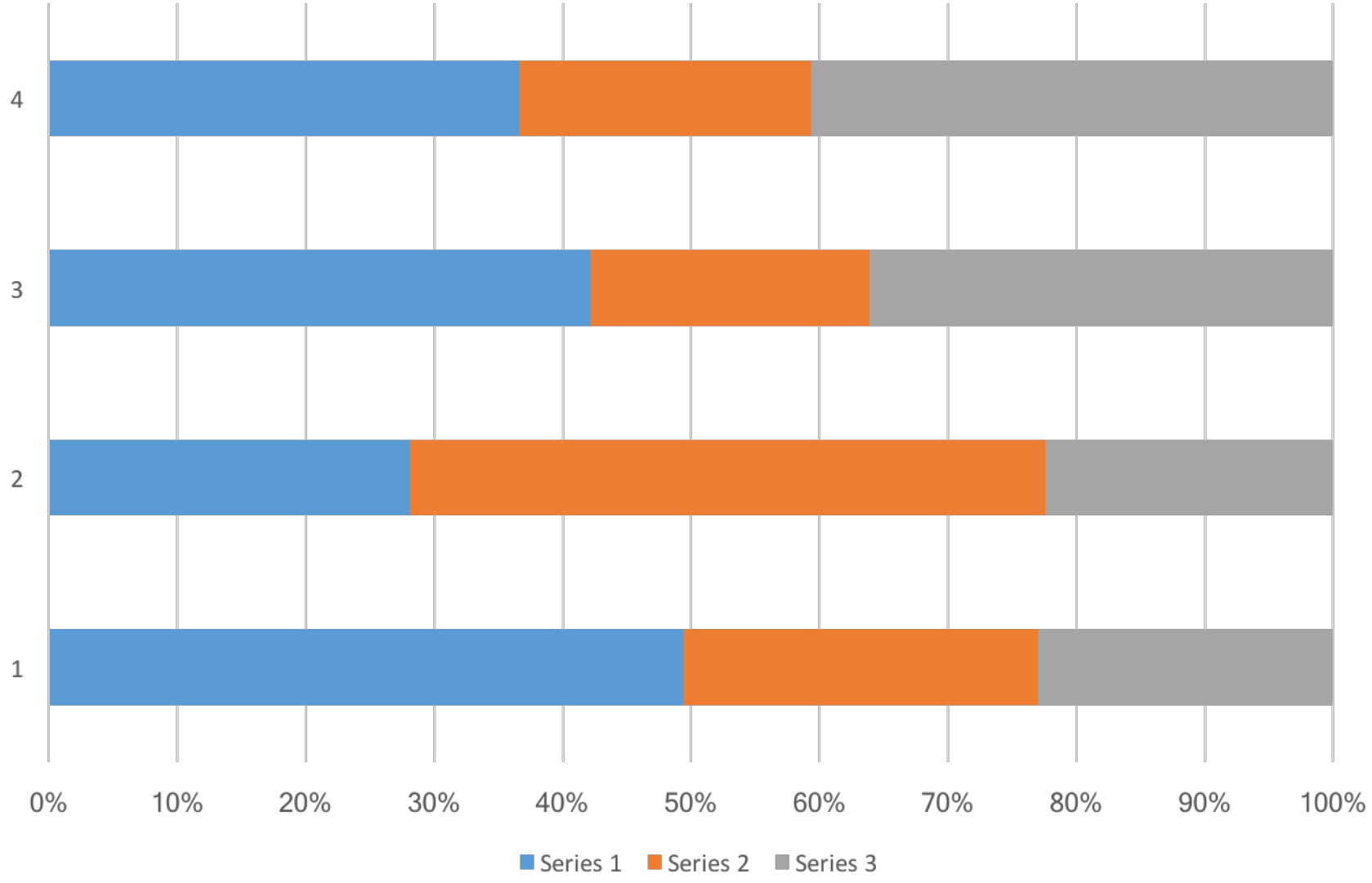
What data is needed to answer those questions?

Who is the audience?

Univariate Graph



Bivariate Case – Stacked Bars



Multivariate Analysis

Introduction to Scatterplots

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

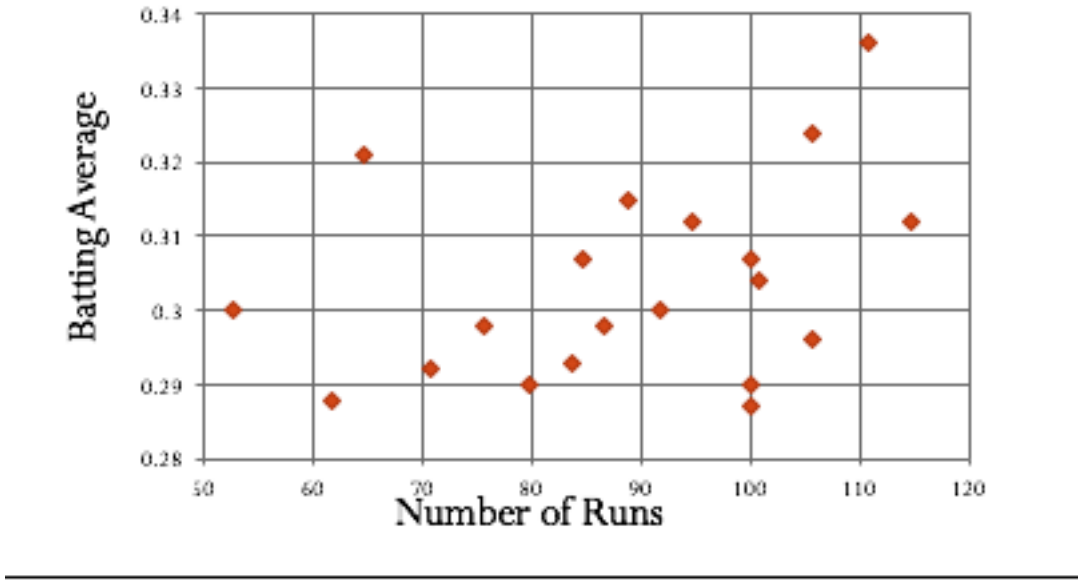
Objective

Apply methods of visualizing discrete data values along two axes

Bivariate Case - Scatterplots

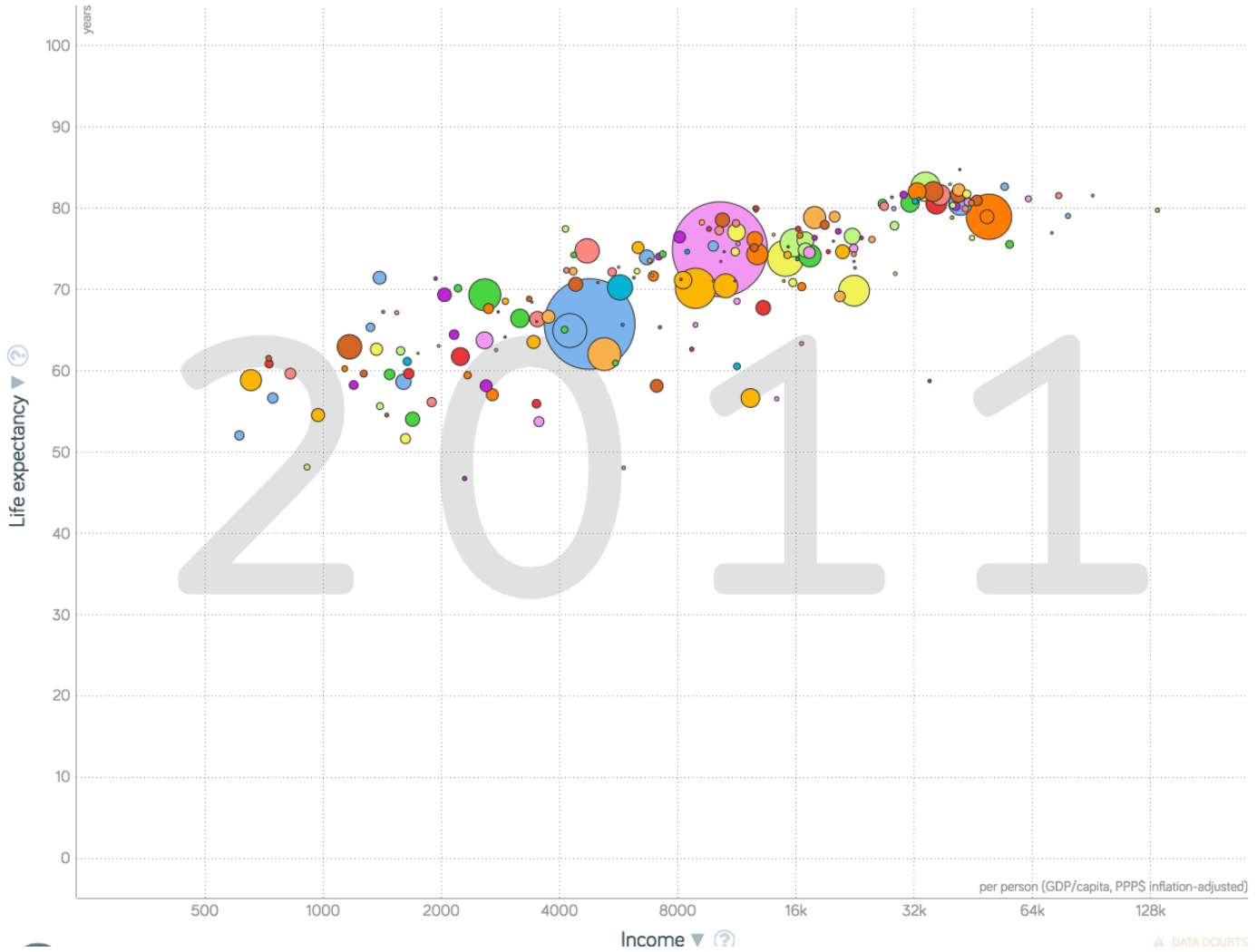
- Visualizes discrete data values along two axes
- Used as a means of analyzing bivariate relationships
- Quick means of assessing outliers, clusters and distributions
- Putting a line through the data can help assess trends, but can also mislead viewer

Scatterplots (placeholder)



Note on ppt: NEED NEW EXAMPLE FROM PREVIOUS DATA

Multivariate Case - Scatterplot



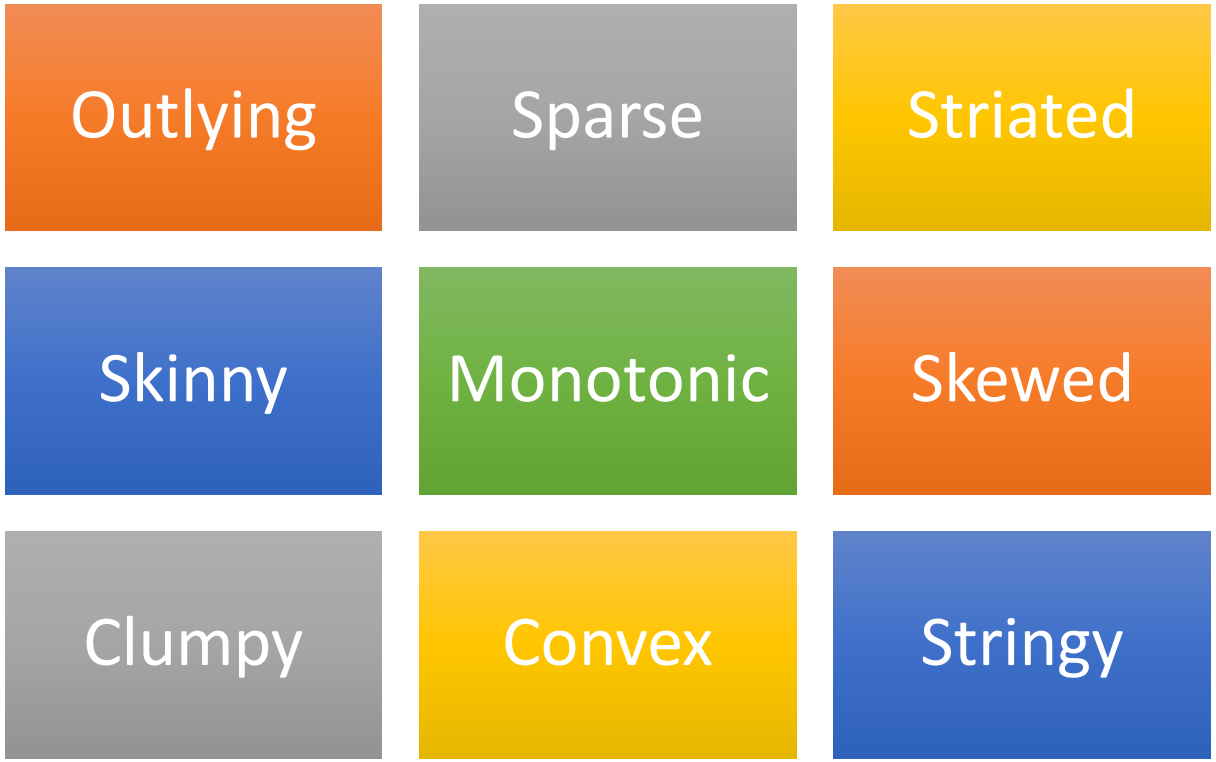
https://www.gapminder.org

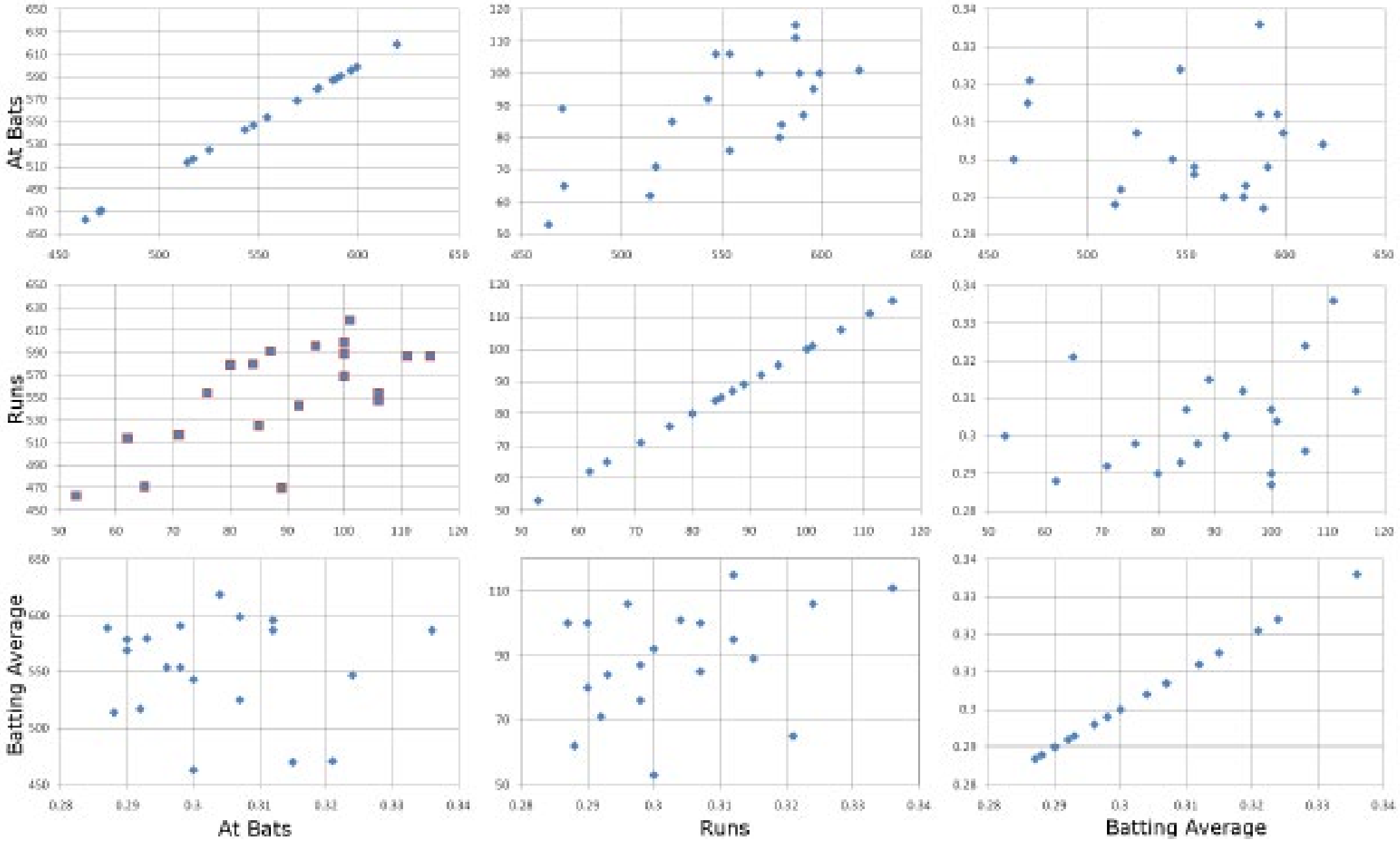
Scagnostics: Scatterplot Diagnostics

- Graph-theoretic measures for detecting a variety of structural anomalies in a geometric graph representation of scatterplot data
- Ratings can be used to pick views that show particular structures that are of interest to the user
- Coined by Tukey, it is an exploratory graphical technique to help determine notable relationships between two variables

Scagnostics

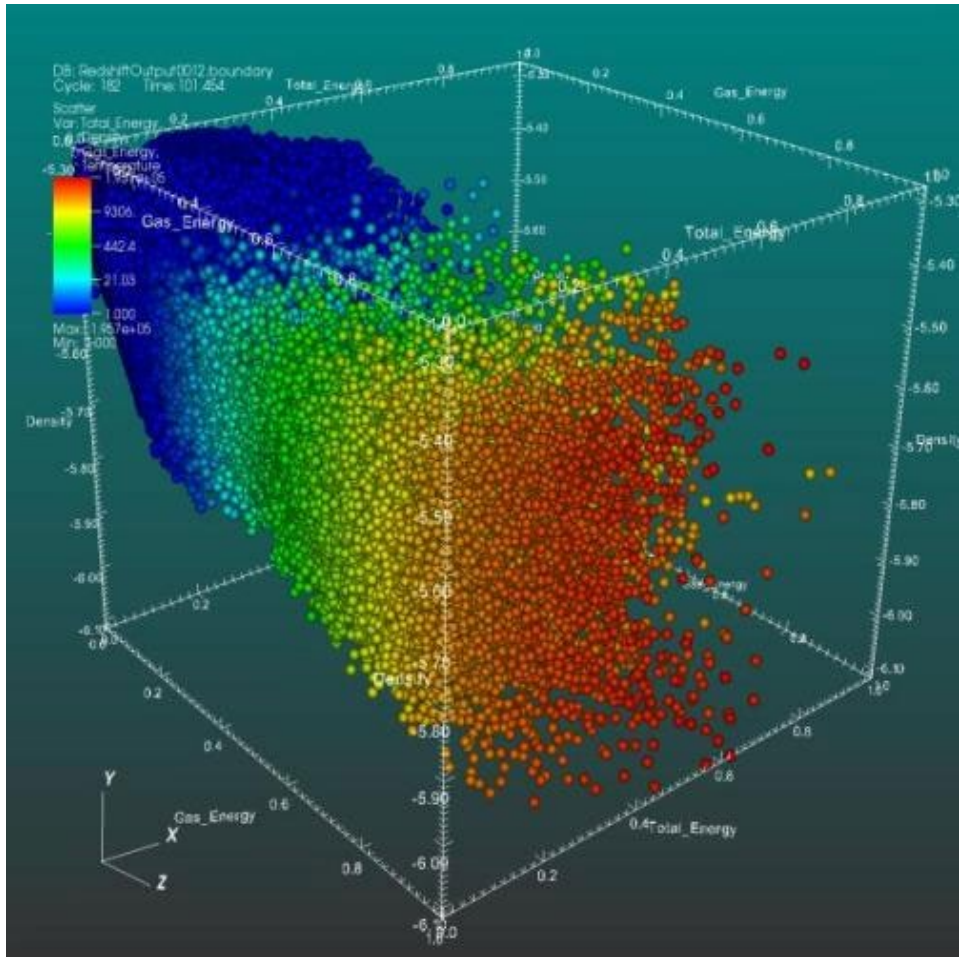
Wilkinson et al. propose nine scagnostic measures to characterize the scatterplots.




At Bats
Runs
Batting Average
At Bats
Runs
Batting Average

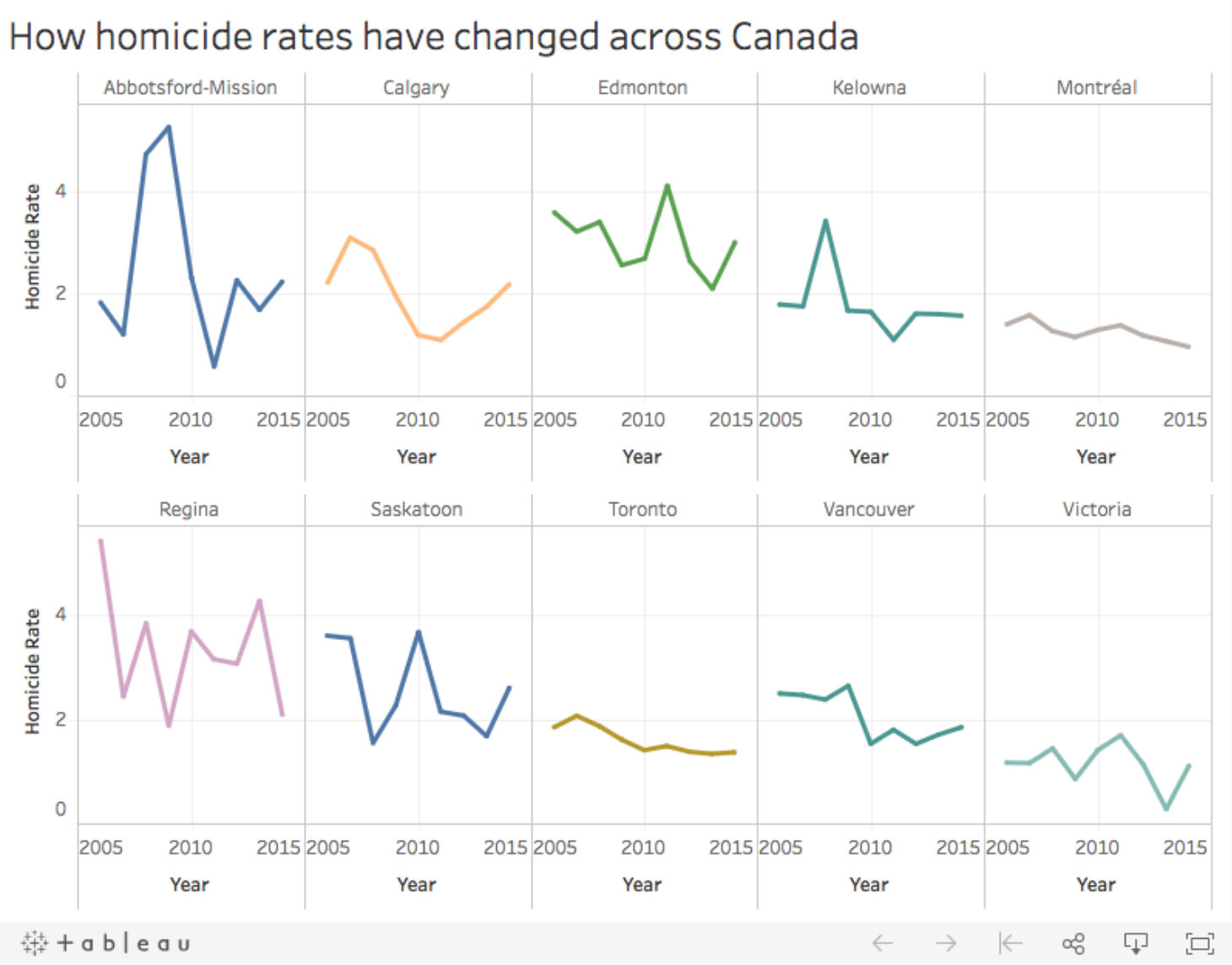
Interaction

- Adding interaction allows viewers to visualize other combinations of variables



https://commons.wikimedia.org/wiki/File:Scatter_plot.jpg

Creating Small Multiples



Multivariate Analysis

Mosaic Plots

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Apply methods of visualizing discrete data values along two axes

Introduction to Mosaic Plots

Graphical display that allows you to examine the relationship among two or more categorical variables

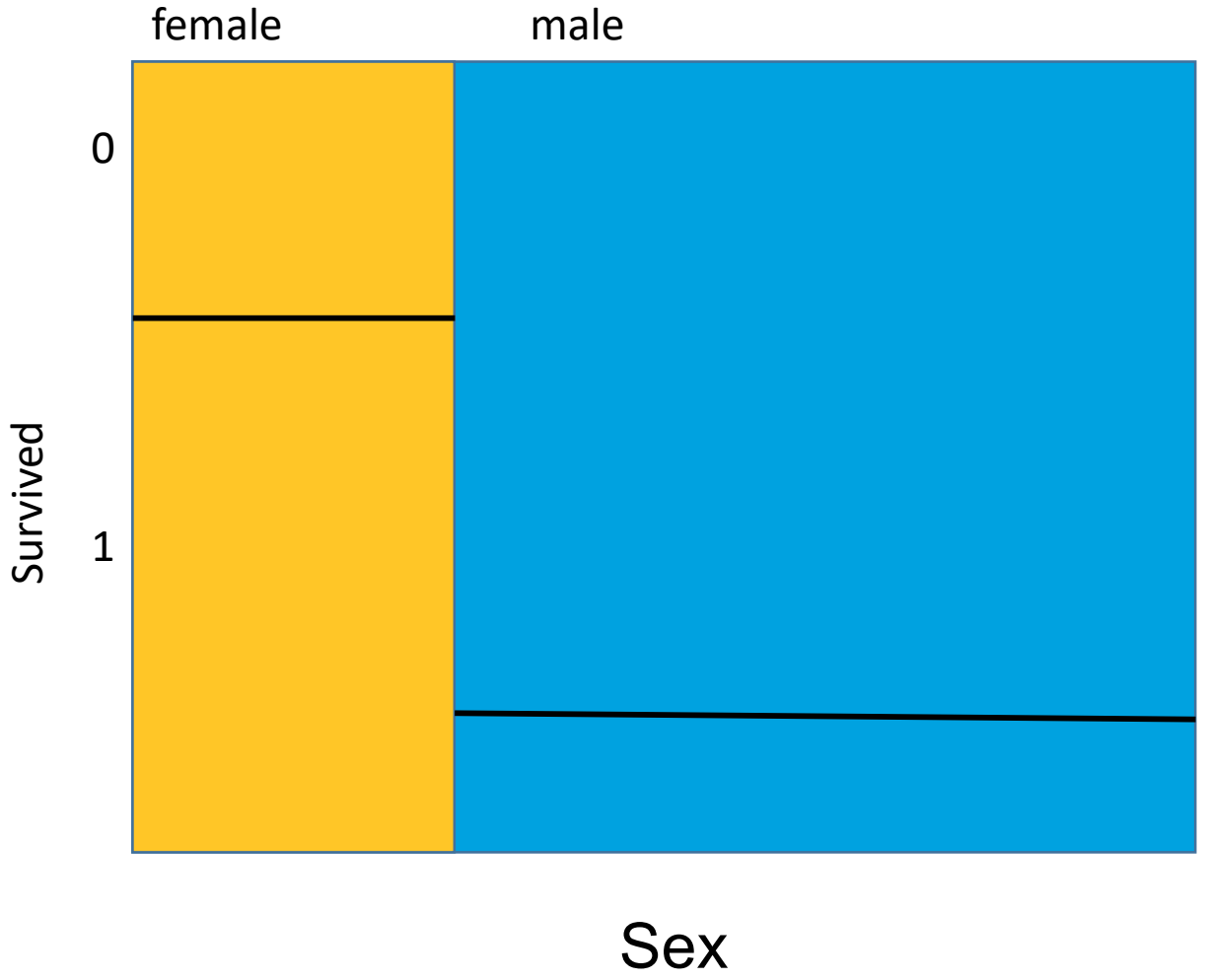
To create:

- Start as a square with length one
- Divide first into horizontal bars whose widths are proportional to the probabilities associated with the first categorical variable
- Next each bar is split vertically by the conditional probability of the second categorical variable

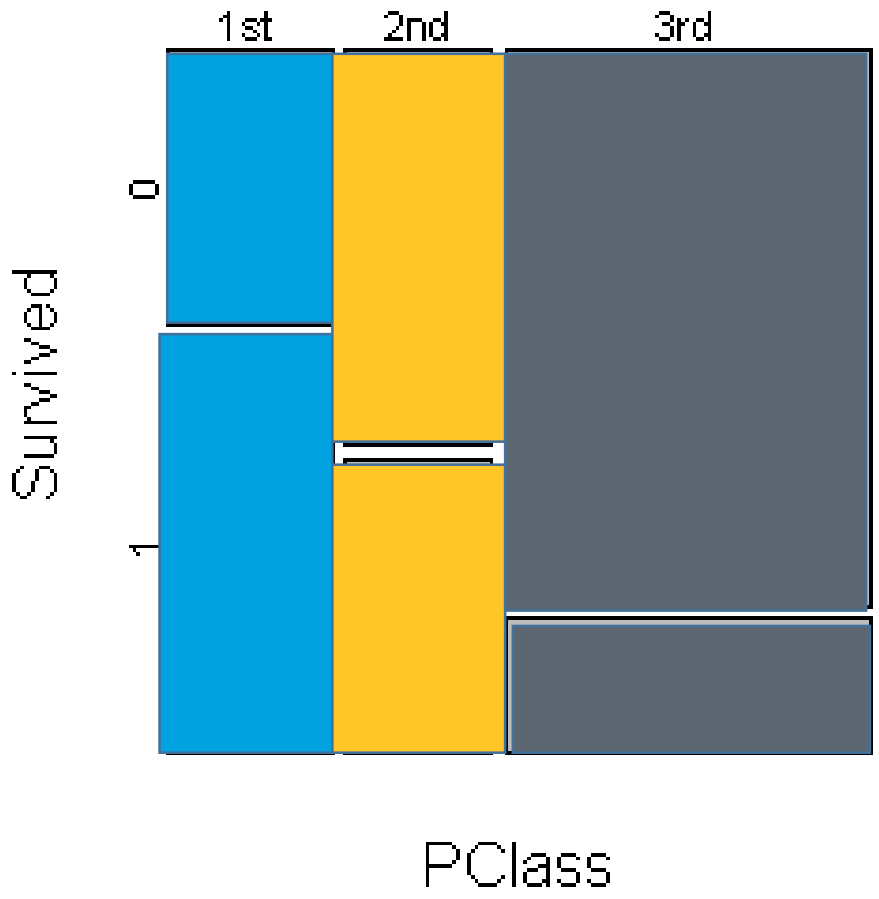
Example: Mortality rates

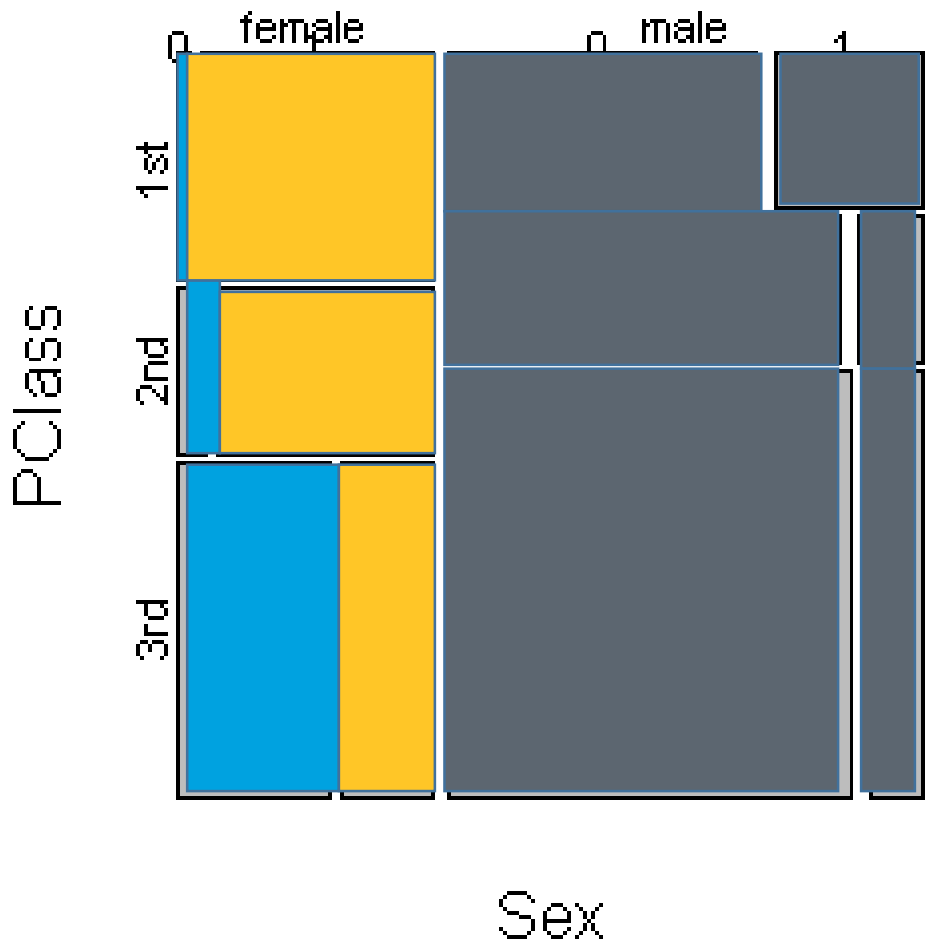
Adults	*Survivors*		*Non-Survivors*	
	Male	*Female*	*Male*	*Female*
1st Class	57	140	118	4
2nd Class	14	80	154	13
3rd Class	75	76	387	89
Crew	192	20	670	3

Children	*Survivors*		*Non-Survivors*	
	Male	*Female*	*Male*	*Female*
1st Class	5	1	0	0
2nd Class	11	13	0	0
3rd Class	13	14	35	17
Crew	0	0	0	0



Examples


1st
2nd
3rd
0
1
Survived
PClass


female
male
0
1
1st
2nd
3rd
PClass
Sex

Mosaic Plots

- It is tempting to dismiss mosaic plots because they represent counts as rectangular areas and so provide a distorted perceptual encoding
- In fact, the important encoding is the length
- At each stage, the comparison of interest is of the length of the sides

Multivariate Analysis

Pixel Based Displays

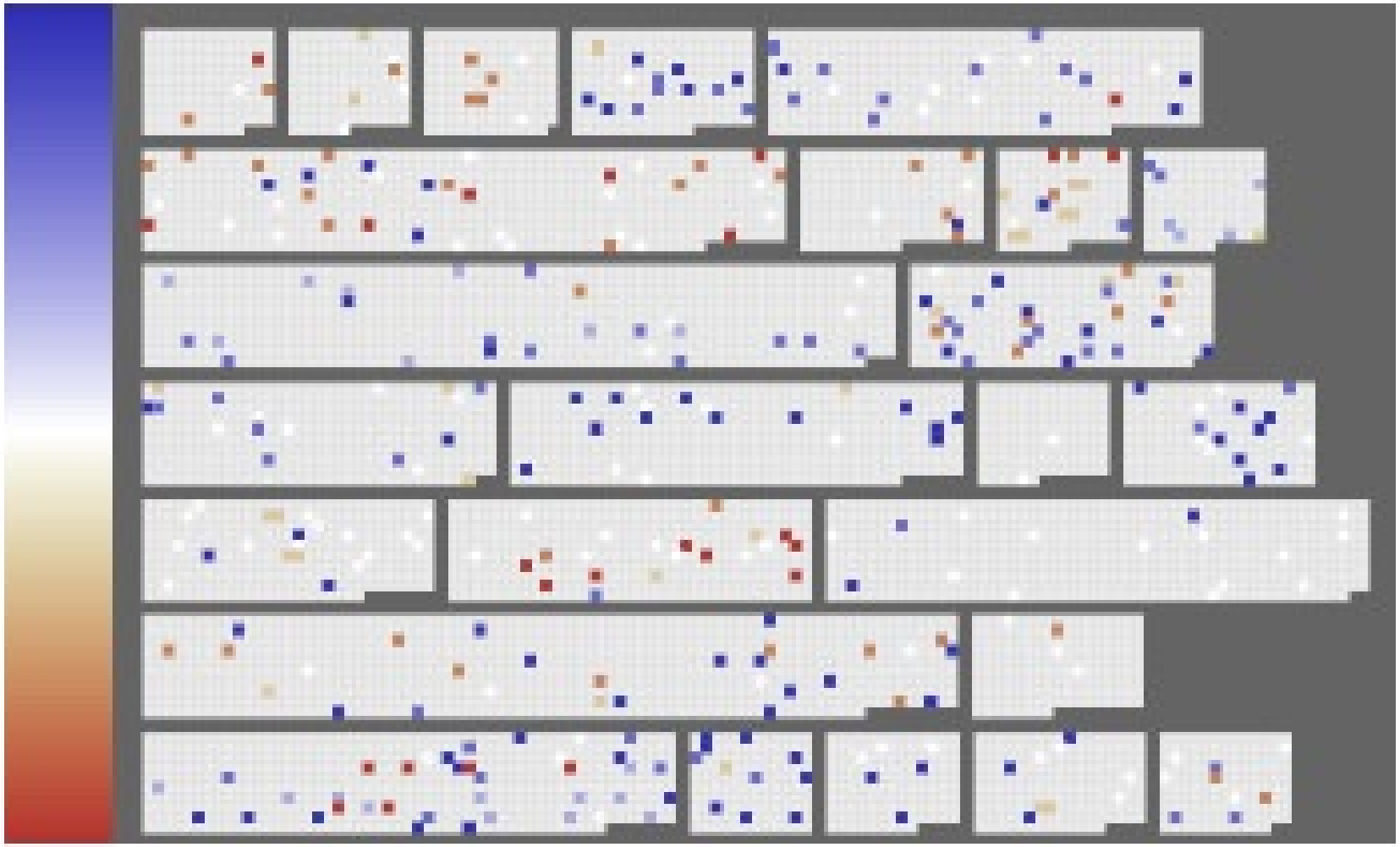
Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

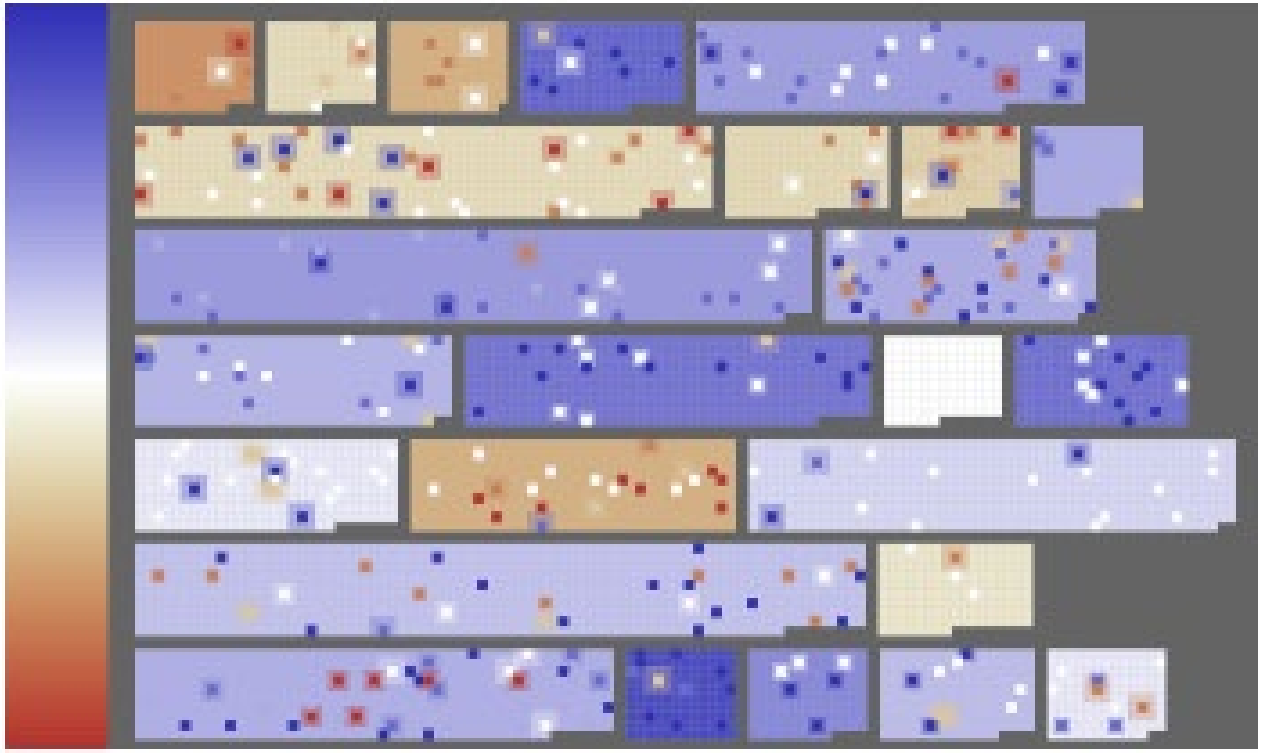
Describe attributes of multivariate data visualization

Pixel Based Displays



Designing Pixel Based Displays

- Could modify the pixel based display to incorporate components that will draw attention to the salient aspects of the data
  - Halo
  - Color
  - Distortion
  - Hatching



Multivariate Analysis

Parallel Coordinate Plots

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

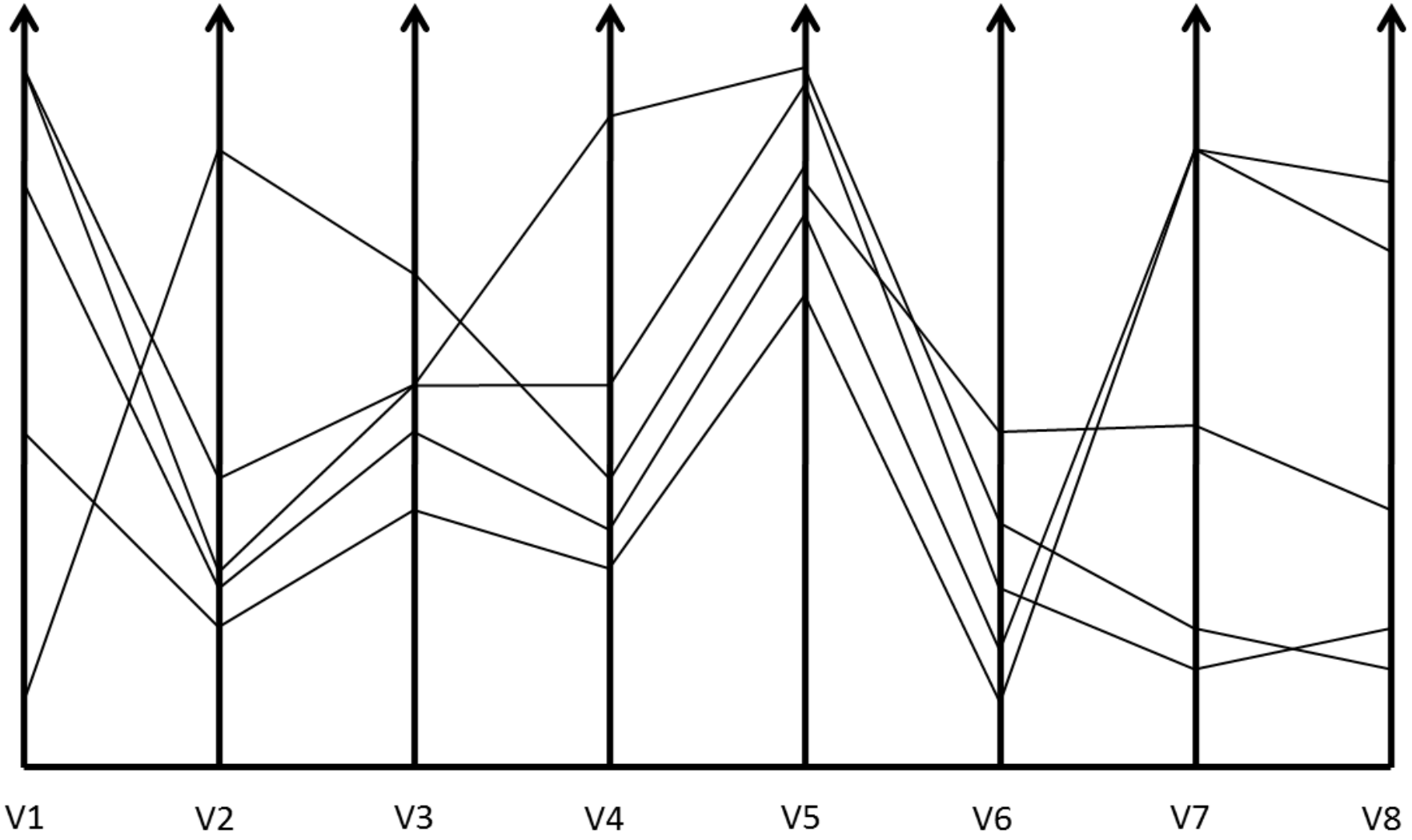
Objective

Describe attributes of multivariate data visualization

Parallel Coordinate Plots

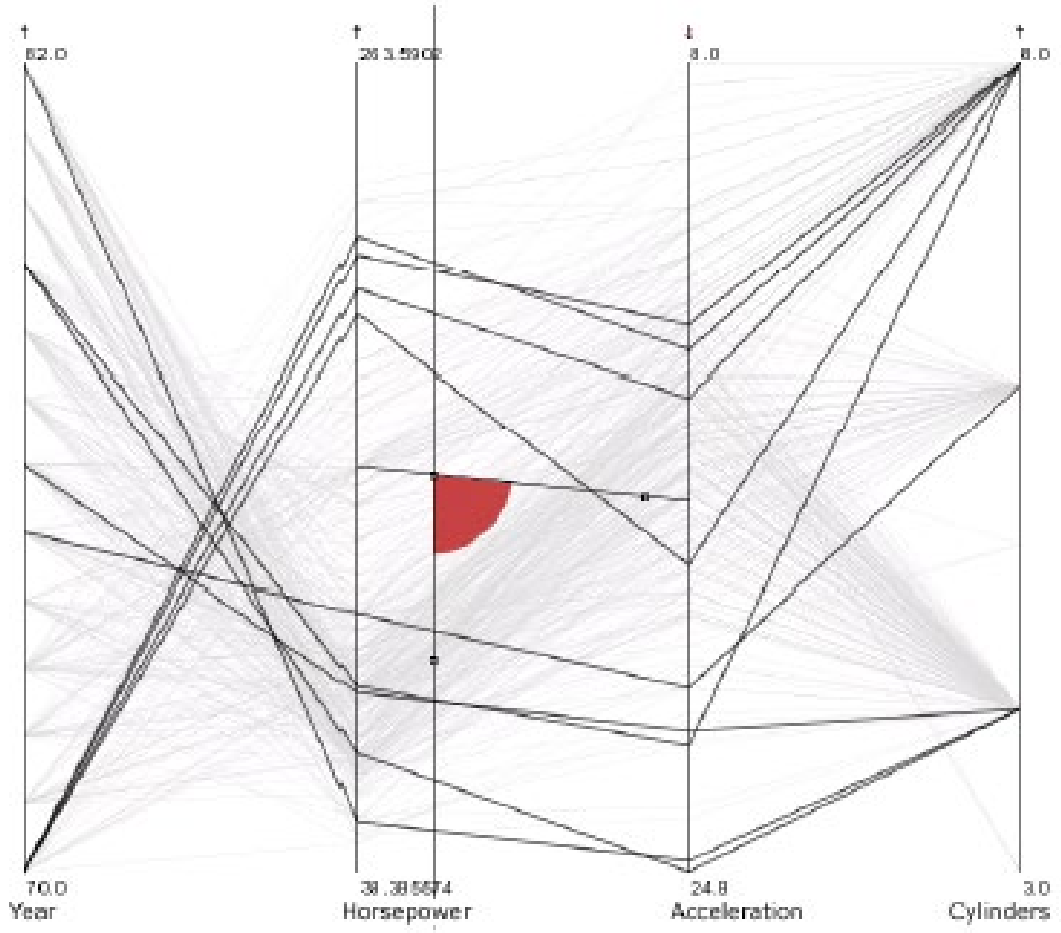
- Different variables can take different values with very different ranges
- Need to normalize data ranges
- Order of the parallel coordinate plots has a major impact on the resultant visualization
- The more variables we plot, the more lines we get and the more clutter that we get

Example



Angular Brushing

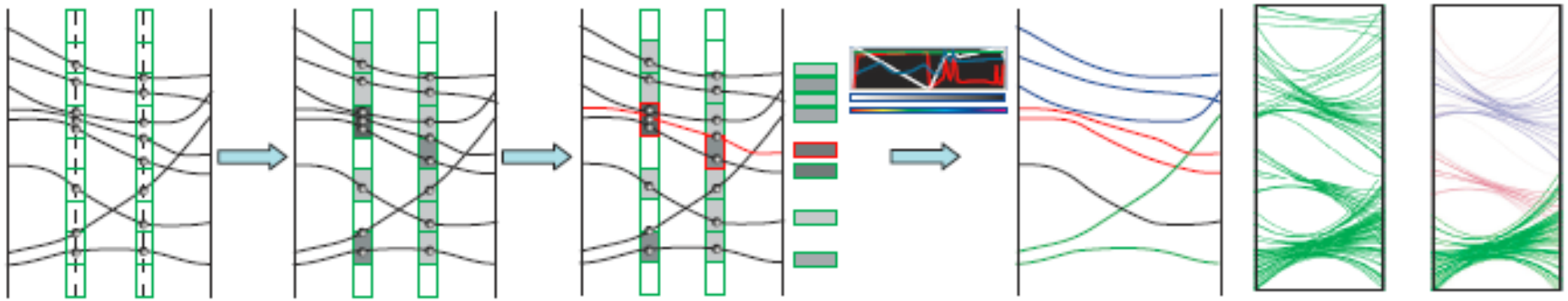
- Angle between axes indicates level of correlation
- Select subsets which exhibit a correlation along two axes by specifying angle of interest



1 –Hauser, H., Ledermann, F., Doleisch, H., "Angular Brushing of Extended Parallel Coordinates," *Proceedings of the IEEE Symposium on Information Visualization* (2002), pp. 127–130

Visual Clustering

- Apply color and opacity based on line density
- Compute local density for each line by averaging the density values of all control points
- Apply color and opacity based on user specification



Screen Space Metrics

- Creates lower-dimensional projections that provide maximum insight into the data and optimizes the parameter space for pixel-oriented visualizations
- Metrics based on a particular view of parallel coordinate plots
- Screen space metrics – depends on the size of the display
- Space between the axes is where interesting patterns occur

Screen Sized Metrics

Use a variety of metrics to try and optimize the use of the screen space

One-Dimensional Histogram Distance

- records the slope of the lines between the axes

Two-Dimensional Axis pair Histogram

- Histogram of all the lines covering both axes

Screen Sized Metrics

Use a variety of metrics to try and optimize the use of the screen space

Line Crossings

- interpret each line between a pair of axes as a directed interval

Angles of Crossing

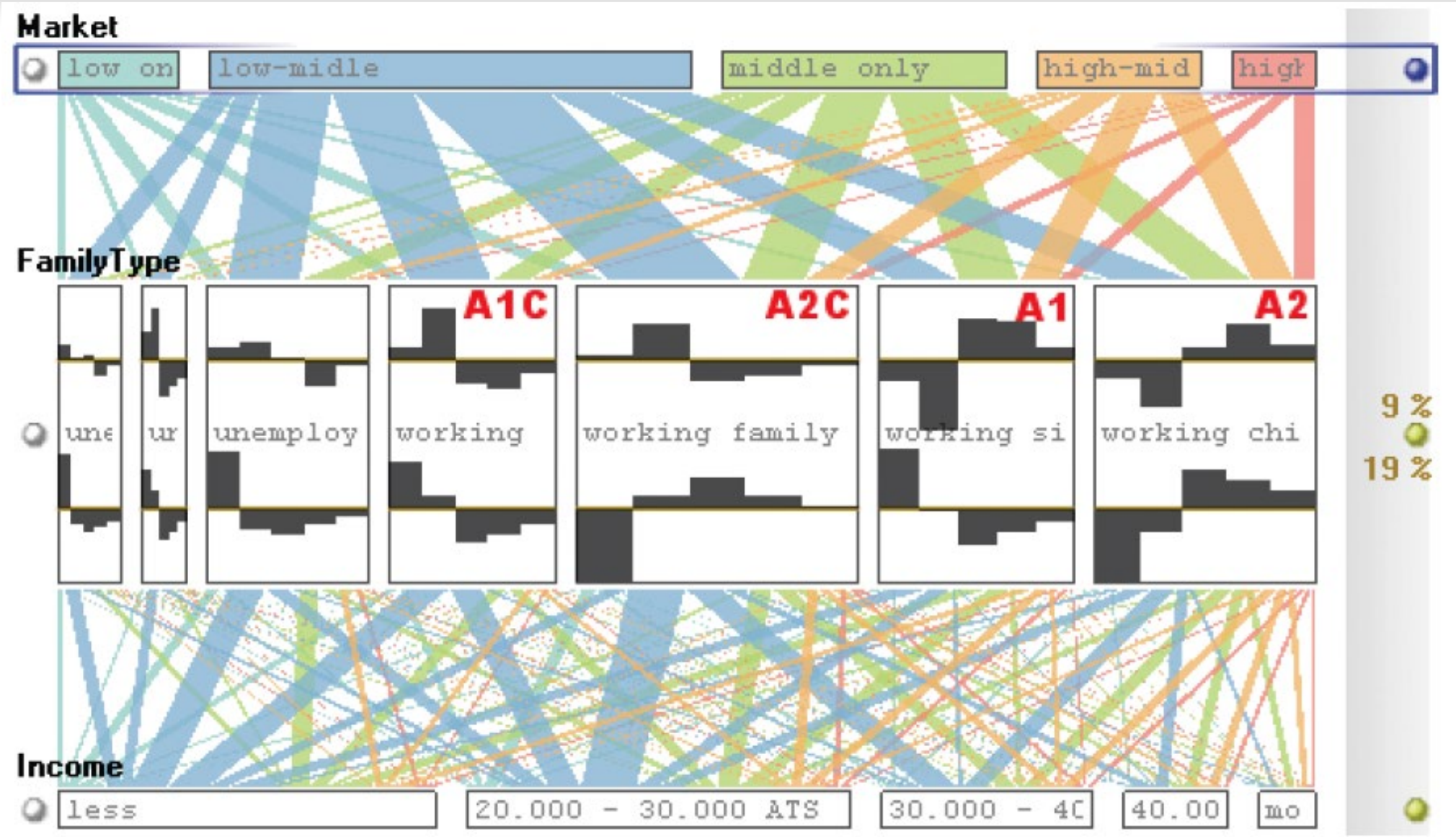
- Determine angle between line crossings

Parallel Sets

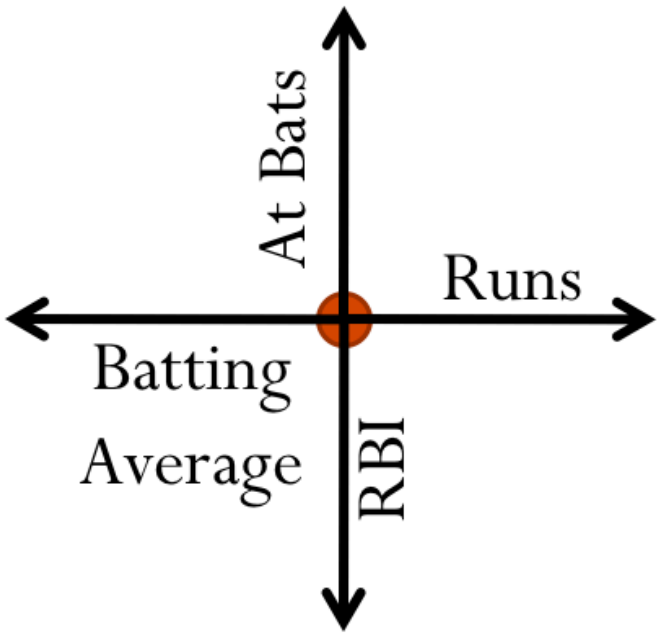
- Visualization method adopting parallel coordinate layout but uses frequency based representation
- Layout similar to parallel coordinate plots
- Continuous axes replaced with boxes
- Used for categorical data

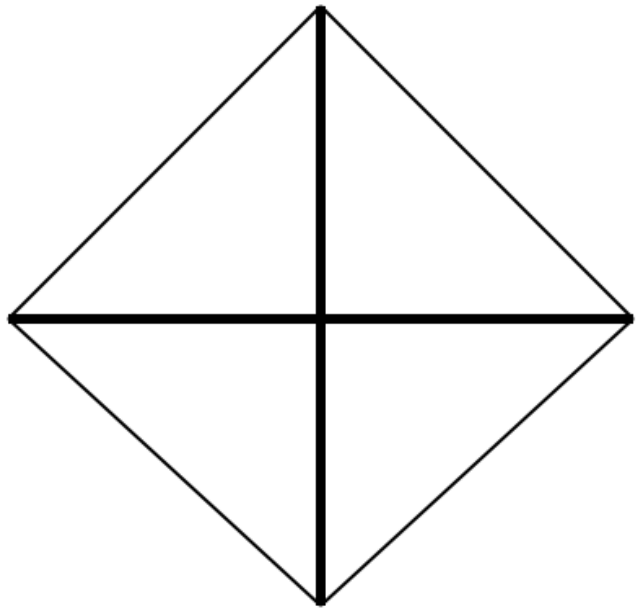
Robert Kosara, Fabian Bendix, and Helwig Hauser. 2006. Parallel Sets: Interactive Exploration and Visual Analysis of Categorical Data. *IEEE Transactions on Visualization and Computer Graphics* 12, 4 (July 2006), 558-568

Parallel Set

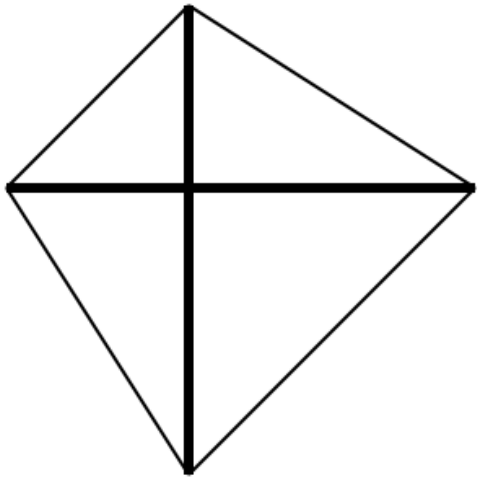


Star Plot

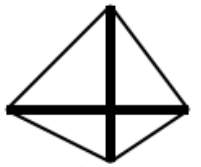




Gonzalez

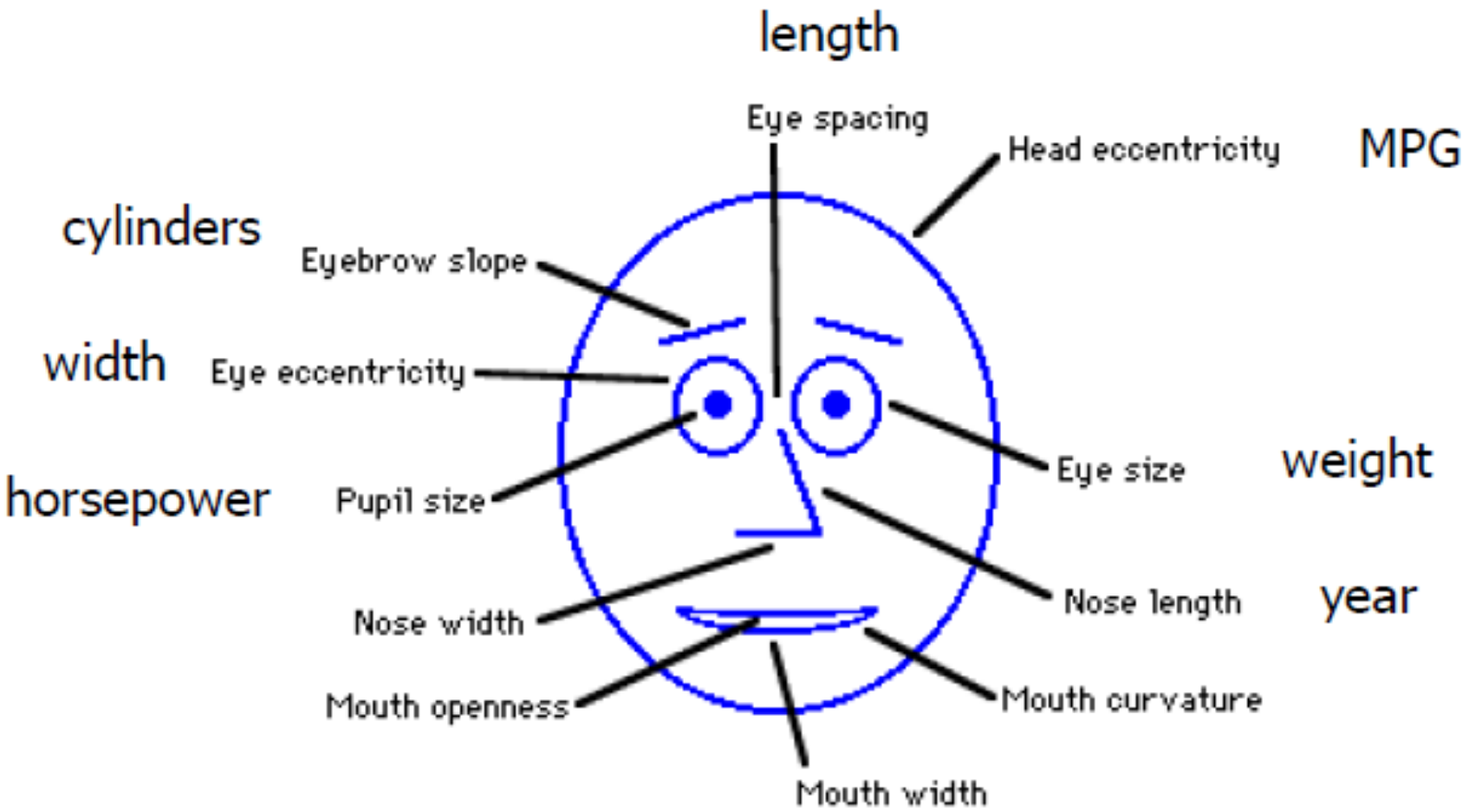


Votto



Infante

Multivariate Case: Chernoff Faces



Multivariate Analysis

Text Visualization

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

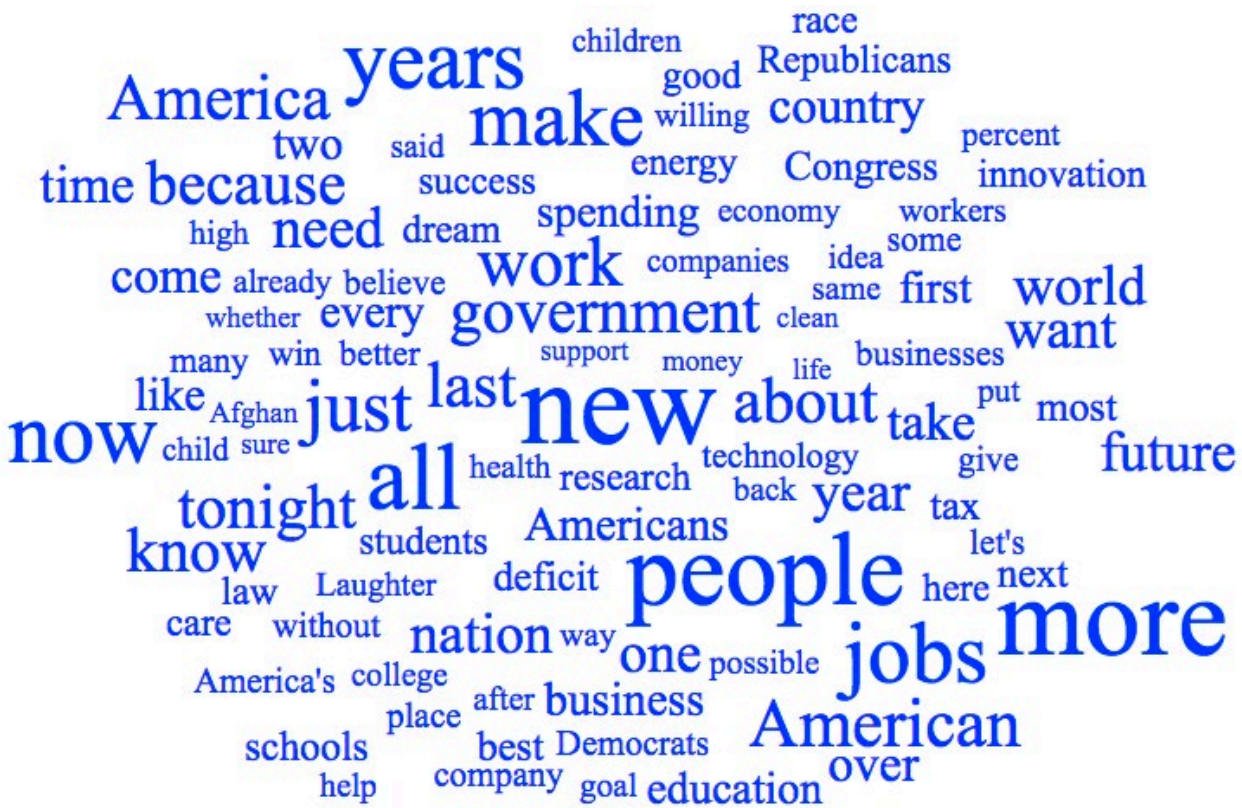
Describe attributes of multivariate data visualization

Text Visualization

- Visual representation for text data where words are placed and scaled based on some statistical measures
- Font size is typically determined by the number of instances a word is used

Example: Word Clouds

2002 State of the Union Address by U.S. President Bush



2011 State of the Union Address by President Obama

Multivariate Analysis

Supervised Learning

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

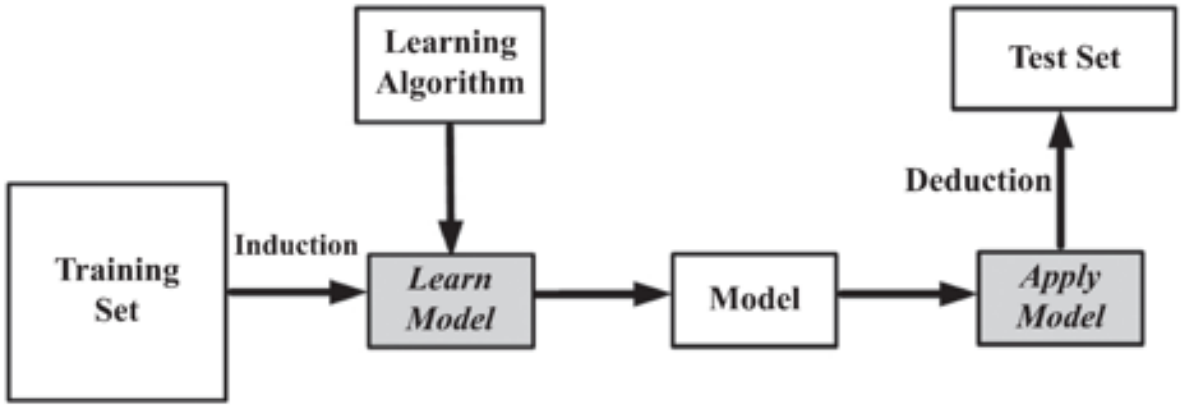
Define supervised learning and describe supervised learning methods

A Twitter Example

ID	Celebrity	Verified Account	# Followers	Influential?
1	Yes	No	1.25M	No
2	No	Yes	1M	No
3	No	Yes	600K	No
4	Yes	Unknown	2.2M	No
5	No	No	850K	Yes
6	No	Yes	750K	No
7	No	No	900K	Yes
8	No	No	700K	No
9	Yes	Yes	1.2M	No
10	No	Unknown	950K	Yes

Supervised Learning: The Process

- We are given a set of labeled records/instances
  - In the format (***X***,*y*)
  - ***X*** is a vector of features
  - *y* is the class attribute (commonly a scalar)
- [Training] supervised learning task is to build a model that maps **X** to y
  - Find a mapping, *m*, such that *m(**X**) = y*
- [Testing] Given an unlabeled instance (***X'***,?), we compute *m(**X'**)*
  - E.g, spam/non-spam prediction



Supervised Learning Algorithms

Classification

- Decision Tree Learning
- *k*-Nearest Neighbor Classifier

Regression

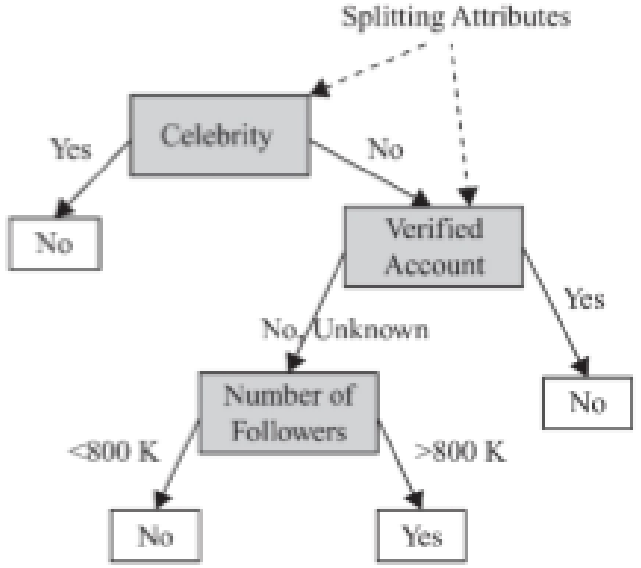
- Linear Regression

Decision Tree Learning

Decision Trees

- A decision tree is learned from the dataset
  - Training data with known classes
- Learned tree is later applied to predict class attribute value of new data
  - Test data with unknown classes
  - Only feature values are known

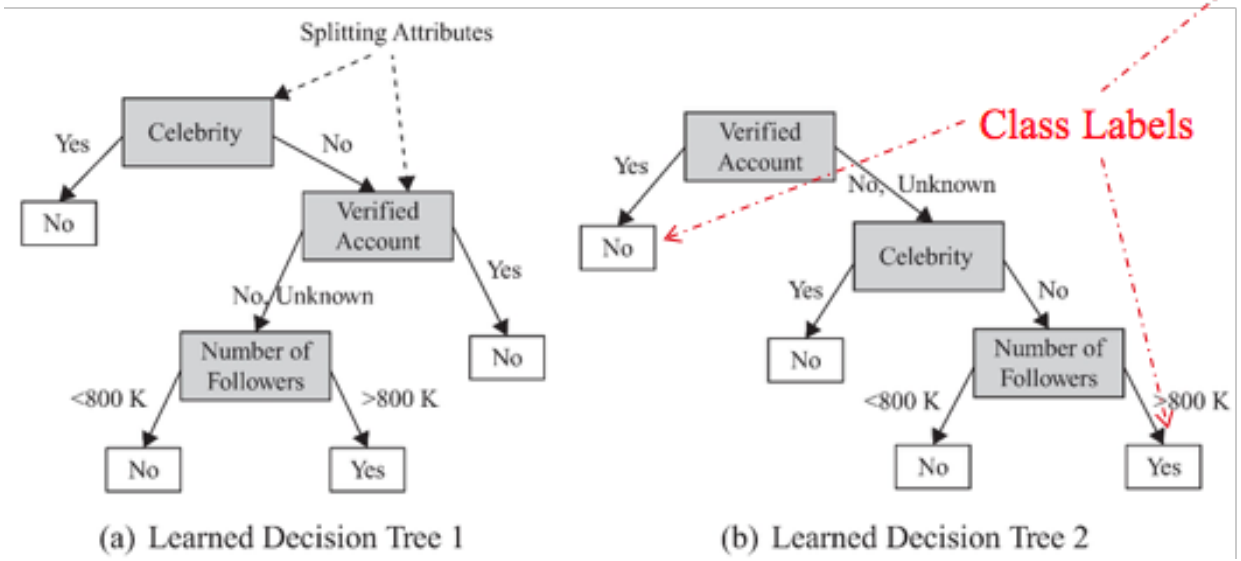
ID	Celebrity	Verified Account	# Followers	Influential?
1	Yes	No	1.25M	No
2	No	Yes	1M	No
3	No	Yes	600K	No
4	Yes	Unknown	2.2M	No
5	No	No	850K	Yes
6	No	Yes	750K	No
7	No	No	900K	Yes
8	No	No	700K	No
9	Yes	Yes	1.2M	No
10	No	Unknown	950K	Yes



Decision Trees Example

Multiple decision trees can be learned from the same dataset.

ID	Celebrity	Verified Account	# Followers	Influential?
1	Yes	No	1.25M	No
2	No	Yes	1M	No
3	No	Yes	600K	No
4	Yes	Unknown	2.2M	No
5	No	No	850K	Yes
6	No	Yes	750K	No
7	No	No	900K	Yes
8	No	No	700K	No
9	Yes	Yes	1.2M	No
10	No	Unknown	950K	Yes



(a) Learned Decision Tree 1

(b) Learned Decision Tree 2

Decision Tree Construction

- Decision Trees are constructed recursively
- After selecting a feature for each node, different branches are created
- Training set is then partitioned into subsets based on feature values
- When selecting features, we prefer features that partition the set of instances into subsets that are more *pure*
- A *pure* subset has instances that all have the same class attribute value

Stopping Criteria for Decision Tree Induction

When reaching pure (or highly pure) subsets under a branch

- Decision tree construction process no longer partitions the subset
- Creates a leaf node under the branch
- Assigns class attribute value (or the majority class attribute value) for subset instances as the leaf's predicted class attribute value

Measuring Purity

- To measure purity we can use/minimize entropy
- Over a subset of training instances, *T*, with a binary class attribute (values in {+,-}), the entropy of *T* is defined as:

$$entropy(T) = -p_+ \log p_+ - p_- \log p_-$$

Entropy Example

Assume there is a subset T that has 10 instances:

- Seven instances have a positive class attribute value
- Three have a negative class attribute value

- Denote *T* as [7+,3-]
- The entropy for subset *T* is:

$$entropy(T) = -\frac{7}{10}\log\frac{7}{10} - \frac{3}{10}\log\frac{3}{10} = 0.881$$

Entropy Values vs. Purity

$$entropy(T) = -p_+ \log p_+ - p_- \log p_-$$

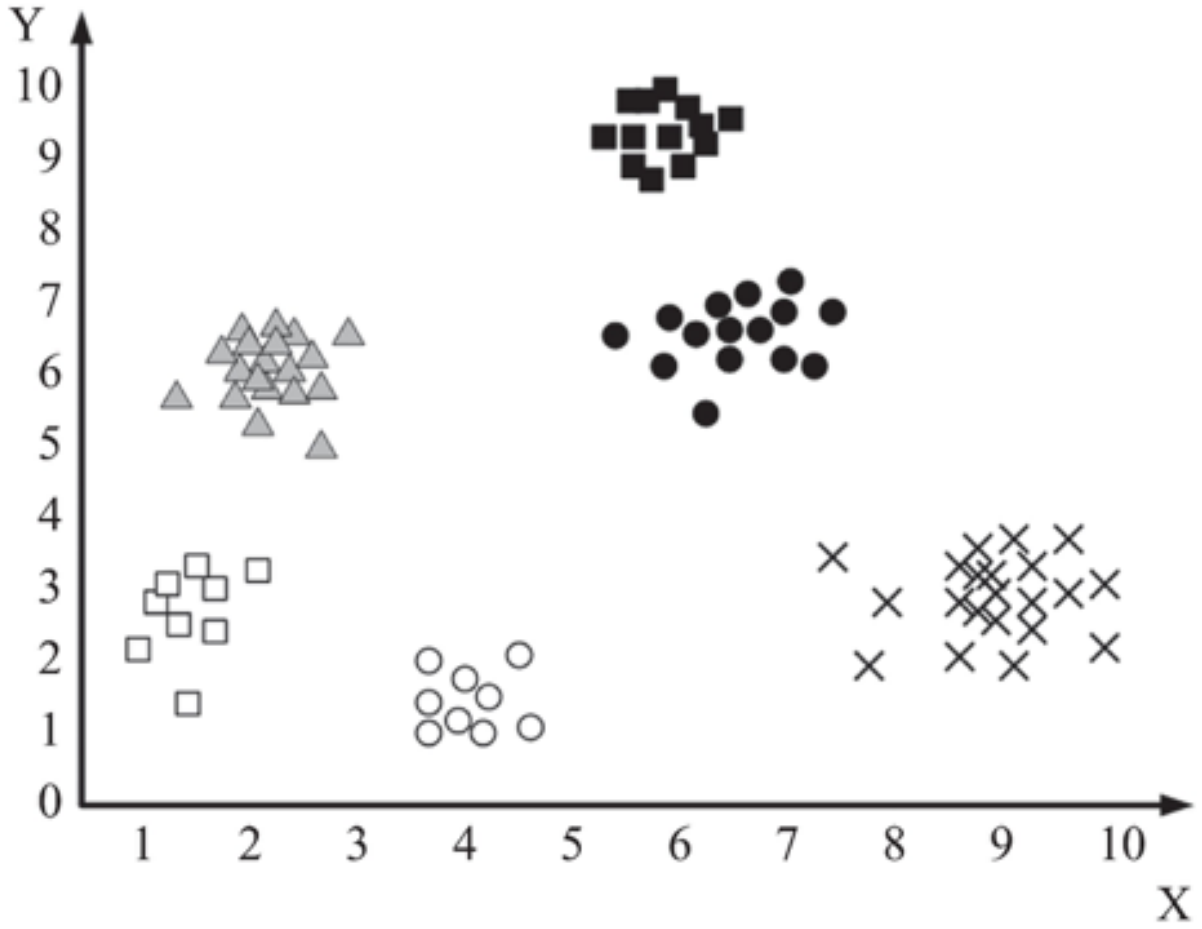
- In a pure subset, all instances have the same class attribute value (**entropy is 0**)
- When is it **1**?
- If the subset contains an unequal number of positive and negative instances
  - The entropy is between 0 and 1

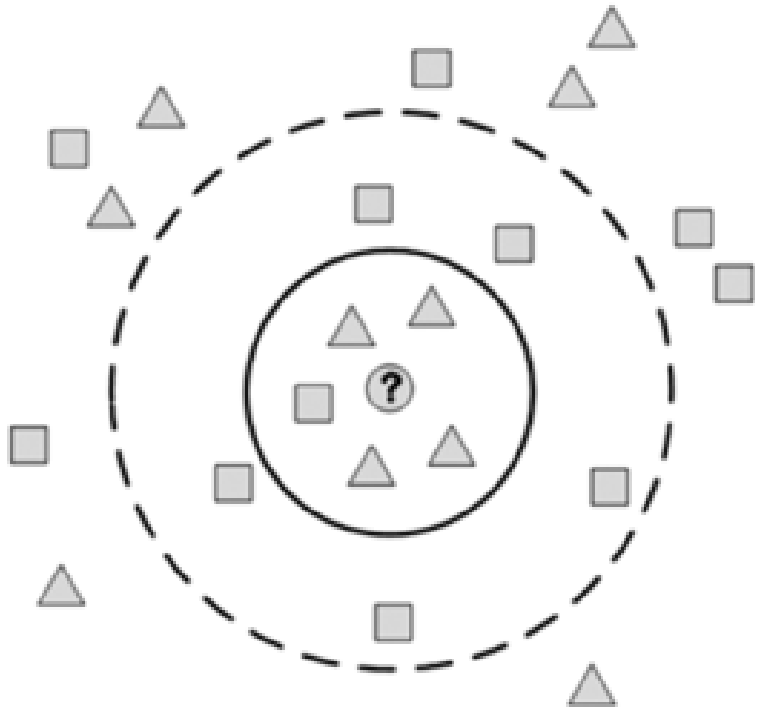
Nearest Neighbor Classifier

An Intuitive Illustration of *k*-NN

No.	Outlook (O)	Temperature (T)	Humidity (H)	Play Golf (PG)
1	sunny	hot	high	N
2	sunny	mild	high	N
3	overcast	hot	high	Y
4	rain	mild	high	Y
5	sunny	cool	normal	Y
6	rain	cool	normal	N
7	overcast	cool	normal	Y
8	sunny	mild	high	?

k-NN: Example





k-Nearest Neighbors

- *k*-nearest neighbors or *k*-NN
- Uses *k* nearest instances, called **neighbors**, to perform classification
- Instance being classified is assigned label (class attribute value) that majority of its *k* neighbors are assigned
- When *k* = 1, the closest neighbor's label is used as predicted label for instance being classified
- To determine the neighbors of an instance, we need to measure its distance to all other instances based on some distance metric

k-NN: Example

No.	Outlook (O)	Temperature (T)	Humidity (H)	Play Golf (PG)
1	sunny	hot	high	N
2	sunny	mild	high	N
3	overcast	hot	high	Y
4	rain	mild	high	Y
5	sunny	cool	normal	Y
6	rain	cool	normal	N
7	overcast	cool	normal	Y
8	sunny	mild	high	?

Similarity between row 8 and other data instances;
(Similarity = 1 if attributes have the same value, otherwise similarity = 0)

Data instance	Outlook	Temperature	Humidity	Similarity	Label	K	Prediction
2	1	1	1	3	N	1	N
1	1	0	1	2	N	2	N
4	0	1	1	2	Y	3	N
3	0	0	1	1	Y	4	?
5	1	0	0	1	Y	5	Y
6	0	0	0	0	N	6	?
7	0	0	0	0	Y	7	Y

k-NN: Algorithm

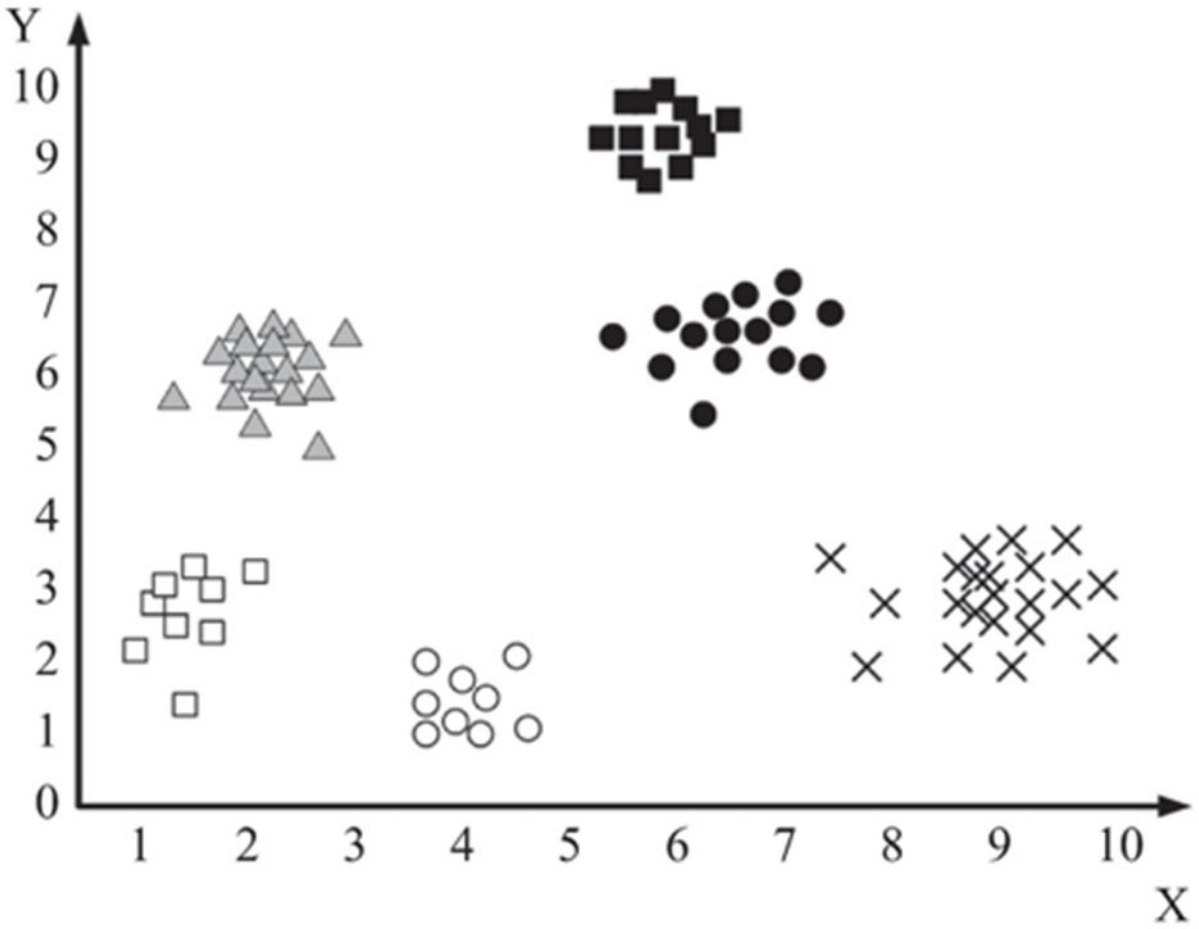
Algorithm 5.1 *k*-Nearest Neighbor Classifier

Require: Instance i, A Dataset of Real-Value Attributes, k (number of neighbors), distance measure d

1: **return** Class label for instance i
2: Compute k nearest neighbors of instance i based on distance mea- sure d.
3: l = the majority class label among neighbors of instance i. If more than one majority label, select one randomly.
4: Classify instance i as class l

k-NN: A Lazy Learning Algorithm

- Does *k*-NN learn?
- How fast is *k*-NN?



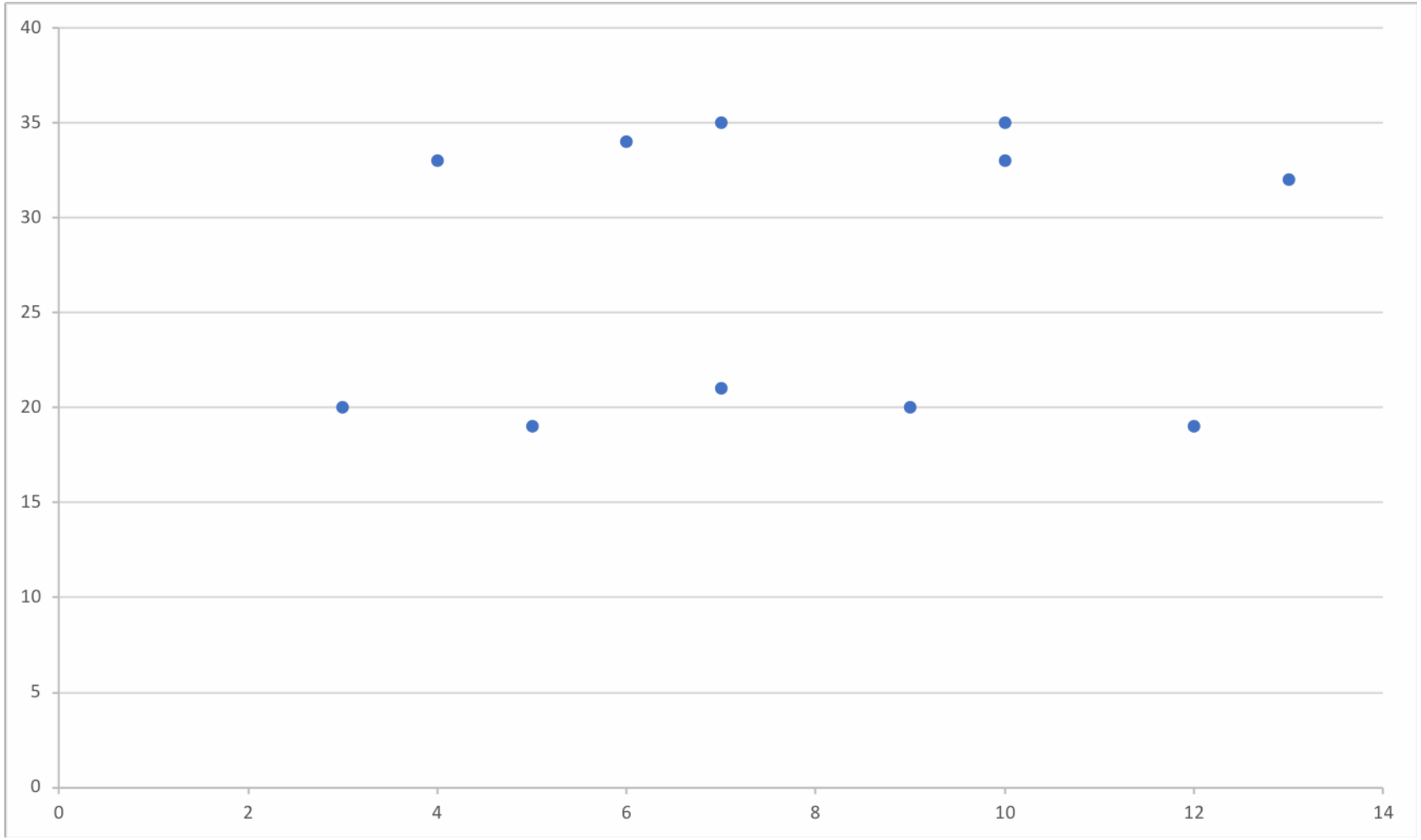
Regression

$$y = c$$

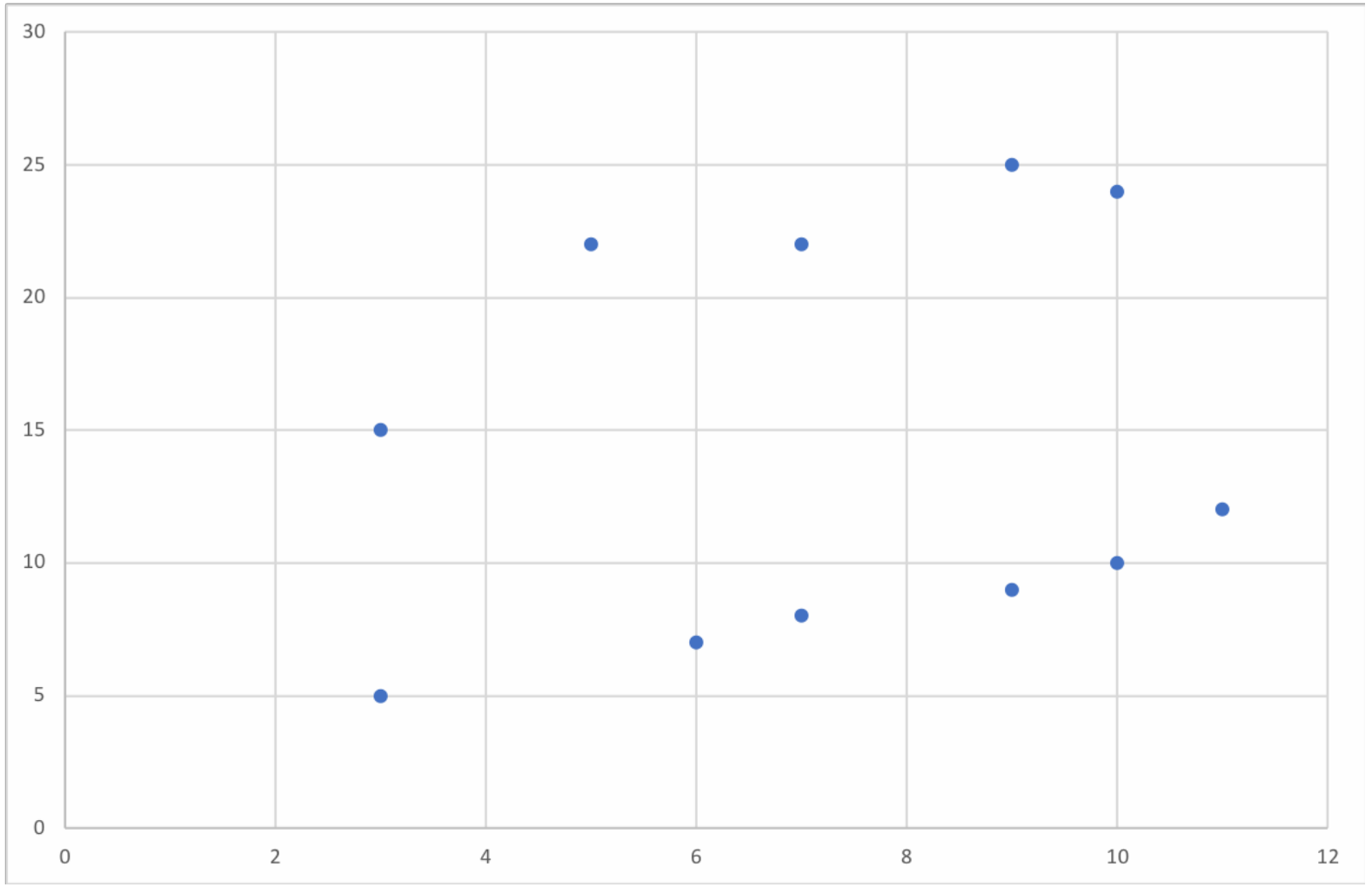
$$y = \beta_0 + \beta_1 x$$

$$\varepsilon = y - (\beta_0 + \beta_1 x)$$

Approximation



Linear Approximation



Regression

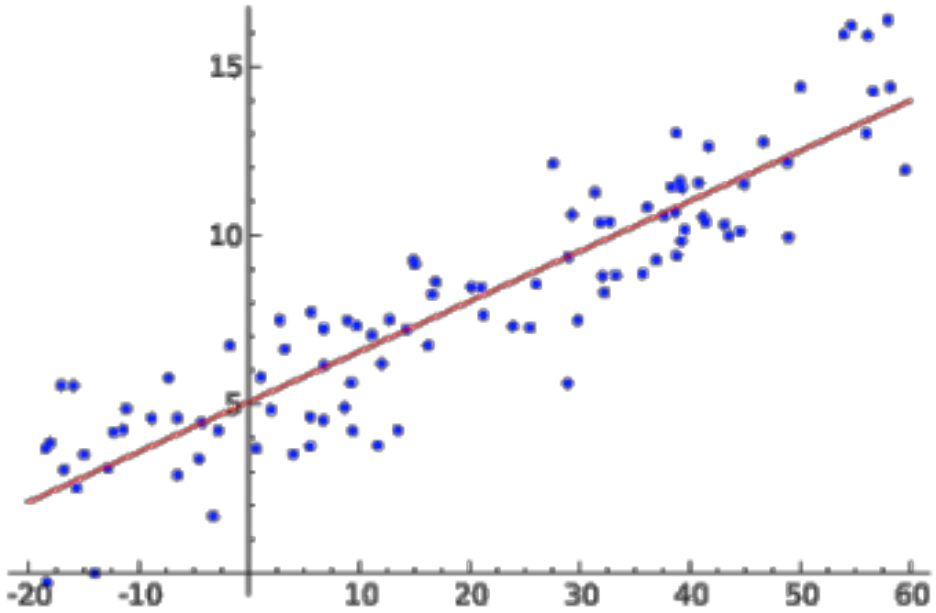
In regression,

- Class values are real numbers as class values
  - In classification, class values are categorical

$$y \approx f(X)$$

Class attribute
(dependent variable)
$y \in R$

Features
(regressors)
$X = (x_1, x_2, \ldots, x_m)$



GOAL:
Find the relation between
y and vector $X = (x_1, x_2, \ldots, x_m)$

Linear Regression

$$Y = XW + \epsilon$$

$$\epsilon^2 = ||\epsilon^2|| = ||Y - XW||^2$$

Linear Regression

- we assume the relation between the class attribute *Y* and feature set ***X*** is linear
- *W* represents the vector of regression coefficients

Regression

- can be solved by estimating *W* and epsilon using the provided dataset and the labels *Y*
- “Least squares” is a popular method to solve regression

Multivariate Analysis

Supervised Learning Evaluation

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Describe Methods for Evaluating Supervised Learning

Evaluating Supervised Learning: Why and How

Training/Testing framework: A training dataset is used to train a model
the model is evaluated on a test dataset.

The correct labels of a test dataset are unknown

When testing, the labels from this test set are removed

Some Basic Methods of Evaluation

Dividing the training set into train/test sets:

Leave-one-out training

- Divide training set into *k* equally sized partitions
- Use all folds but one to train and one left out for testing

k-fold cross validation training

- Divide training set into *k* equally sized sets
- Run algorithm *k* times
- In round *i*, use all folds but fold *i* for training and fold *i* for testing
- average performance of algorithm over *k* rounds measures performance of algorithm

Measures used in Evaluation

- Class labels are discrete, measure accuracy by dividing number of correctly predicted labels (*C*) by total number of instances (*N*)

$$\text{accuracy} = \frac{C}{N}$$

$$\text{error rate} = 1 \text{ - accuracy}$$

- More sophisticated approaches of evaluation:
  - AUC
  - F-measure

Beyond Accuracy Measure

- Labels cannot be predicted precisely
- Set a margin to accept or reject the predictions
  - Example:
    When the observed temperature is 71, any prediction range of 71±0.5 can be considered as a correct prediction
- Use correlation between predicted labels and ground truth

Multivariate Analysis

Unsupervised Learning

Ira A. Fulton Schools of Engineering
Arizona State University

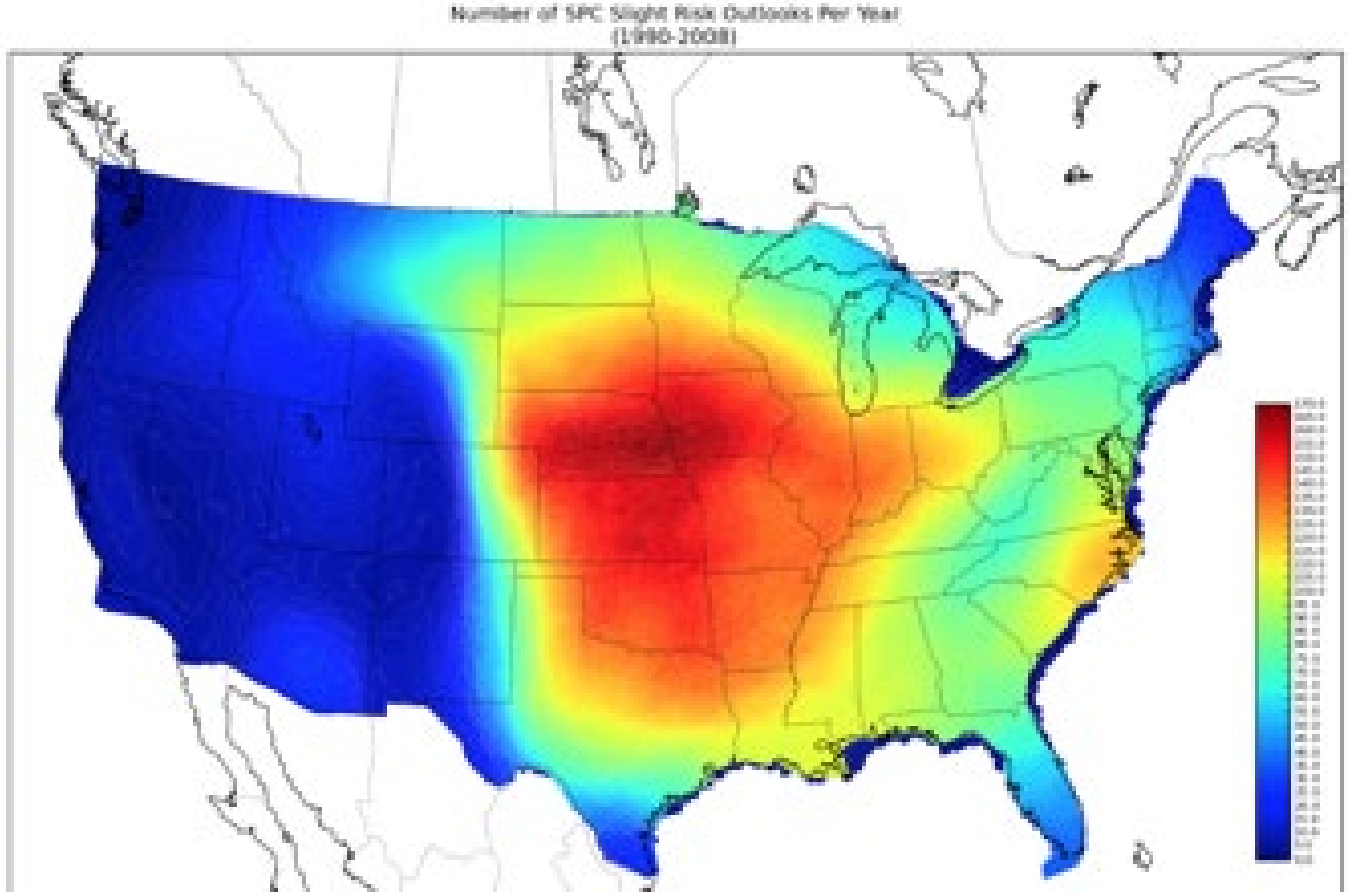
Objective

Objective

Define unsupervised learning and describe unsupervised learning methods

Unsupervised Learning

- Clustering is a form of unsupervised learning
- Clustering algorithms group together similar items

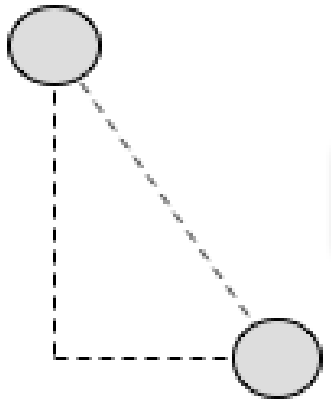


Kernel Density Estimation

Measuring Distance/Similarity in Clustering

- Clustering Goal: Group together similar items
- Instances are put into different clusters based on distance to other instances
- Any clustering algorithm requires a distance measure

The most popular (dis)similarity measure for continuous features are ***Euclidean Distance*** and ***Pearson Linear Correlation***



Euclidean Distance

$$d(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Similarity Measures

X and *Y* are *n*-dimensional vectors

$$X = (x_1, x_2, \ldots, x_n)$$

$$Y = (y_1, y_2, \ldots, y_n)$$

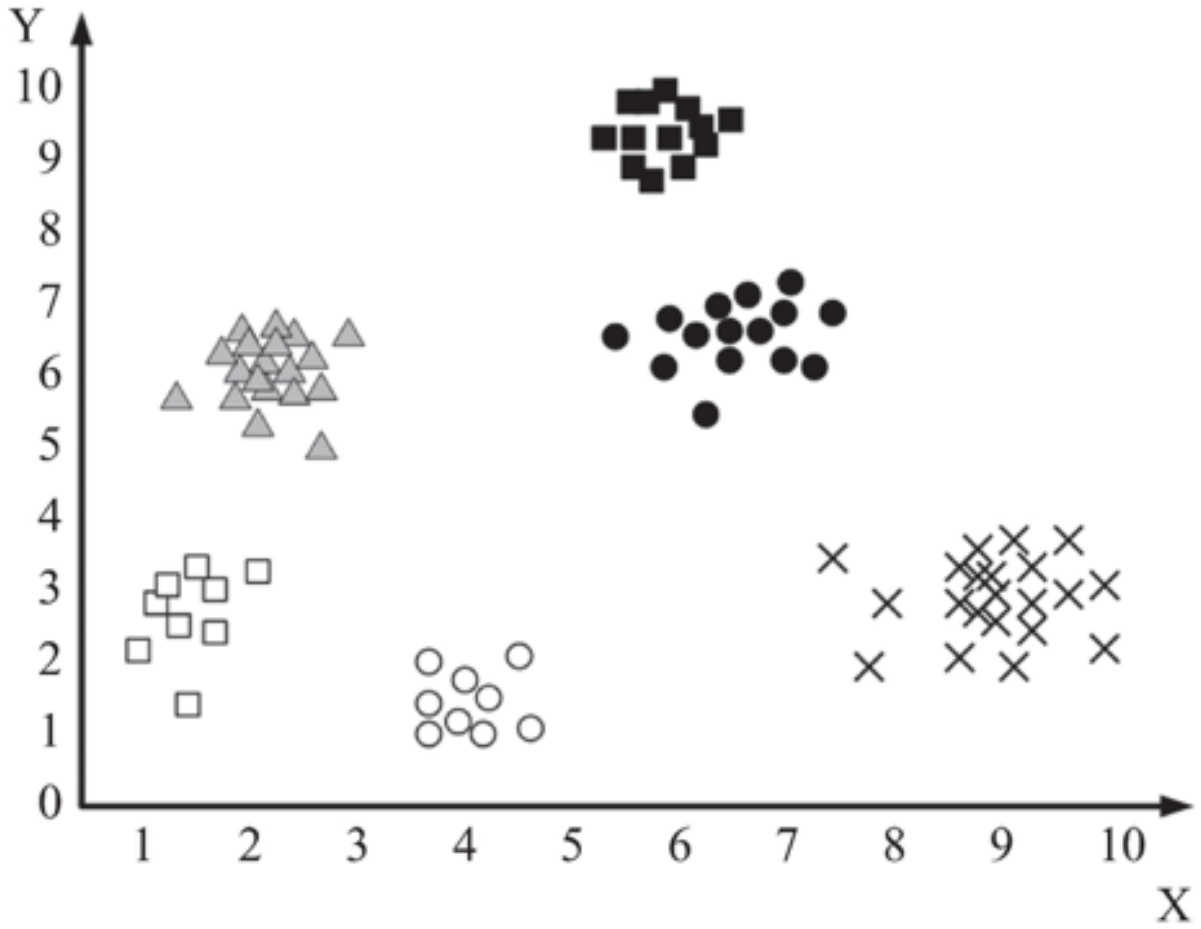
Measure Name	Formula	Description
Mahalanobis	$d(X,Y) = \sqrt{(X-Y)^T \Sigma^{-1} (X-Y)}$	X, Y are features vectors and Σ is the covariance matrix of the dataset
Manhattan (L_1 norm)	$d(X,Y) = \sum_i \lvert x_i - y_i \rvert$	X, Y are features vectors
L_p-norm	$d(X,Y) = (\sum_i \lvert x_i - y_i \rvert^n)^{\frac{1}{n}}$	X, Y are features vectors

Once a distance measure is selected, instances are grouped using it.

Clustering

- Clusters are usually represented by compact and abstract notations
- “Cluster centroids” are one common example of this abstract notation
- Partitional Algorithms (most common type):
  - Partition dataset into a set of clusters
  - Each instance is assigned to a cluster exactly once
  - No instance remains unassigned to clusters

A 2-d Data with 6 Clusters



k-Means: An intuitive and common algorithm

Given data points x_i and an initial set of k centroids $m^1_1, m^1_2, \ldots, m^1_k$ the algorithm proceeds as follow:

Assignment step:

- Assign each data point to cluster S^t_i with closest centroid
  - Each data point goes into exactly one cluster

Update step:

- Calculate new means to be centroid of data points in cluster

$$S_i^t = \{x_p : ||x_p - m_i^t|| \leq ||x_p - m_j^t|| \forall 1 \leq j \leq k\}$$

k-Means – the most commonly used

Algorithm 5.2 *k*-Means Algorithm

Require: A Dataset of Real-Value Attributes, *k* (number of Clusters)

1: **return** A Clustering of Data into *k* Clusters
2: Consider *k* random instances in the data space as the initial cluster centroids.
3: **while** centroids have not converged **do**
4: Assign each instance to the cluster that has the closest cluster centroid.
5: If all instances have been assigned then recalculate the cluster centroids by averaging instances inside each cluster
6: **end while**

Also often used as a baseline algorithm for empirical comparison

When do we stop?

The procedure is repeated until convergence:

Convergence:

- Whether centroids are no longer changing
- Equivalent to clustering assignments not changing

Convergence:

- Algorithm can be stopped when Euclidean distance between centroids in two consecutive steps is less than some small positive value

k-Means (alternative!)

As an alternative, k-means can be implemented to minimize an objective function.

Example: squared distance error

- x_j is the *j*th instance of the cluster *i*.
- $n(i)$ is the number of instances in the cluster *i*.
- c_i is the centroid of cluster *i*

$$\sum_{i=1}^{k} \sum_{j=1}^{n(i)} ||x_j^i - c_i||^2$$

Stopping Criterion

- When the difference between objective function values of two consecutive iterations of k-means algorithm is less than some small value

k-Means Discussion

- Finding global optimum of *k* partitions is computationally expensive (**NP**-hard)
- This is equivalent to finding optimal centroids that minimize objective function

- **Solution**: efficient heuristics
- **Outcome**: converge quickly to a local optimum that might not be global.
  - **Example**: running *k*-means multiple times

Means vs. Medians

- Mean is the average
- Median is the middle value
- Example: A start-up with 8 employees, 1 CTO, 1 CEO
  - 50K, 50K, 50K, 50K, 50K, 80K, 80K, 90K, 150K, 150K
- Should we use means or medians?
- Which one is easier to update?

Multivariate Analysis

Unsupervised Learning Evaluation

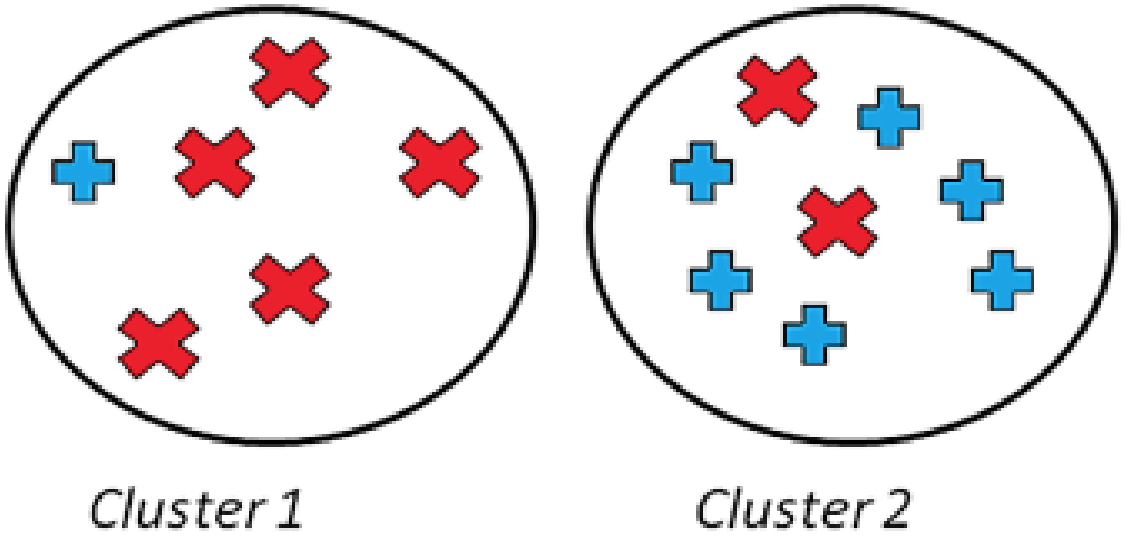
Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Describe methods for evaluating unsupervised learning

Evaluating the Clusterings

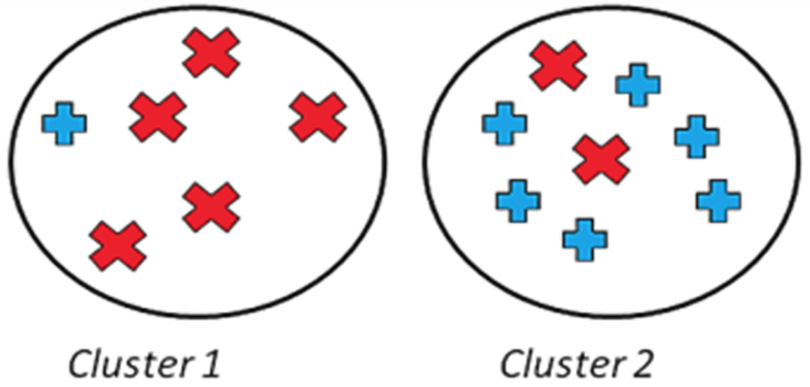


- We are given two types of objects
  - In perfect clustering, objects of the same type are clustered together

Evaluation with Ground Truth

When ground truth is available,

- We have prior knowledge on what the clustering should be, or the correct clustering
  - We can use Accuracy to measure
- But, what is the use of clustering?



Evaluation without Ground Truth

Cohesiveness

- In clustering, we are interested in clusters that exhibit cohesiveness
- In cohesive clusters, instances inside the clusters are close to each other

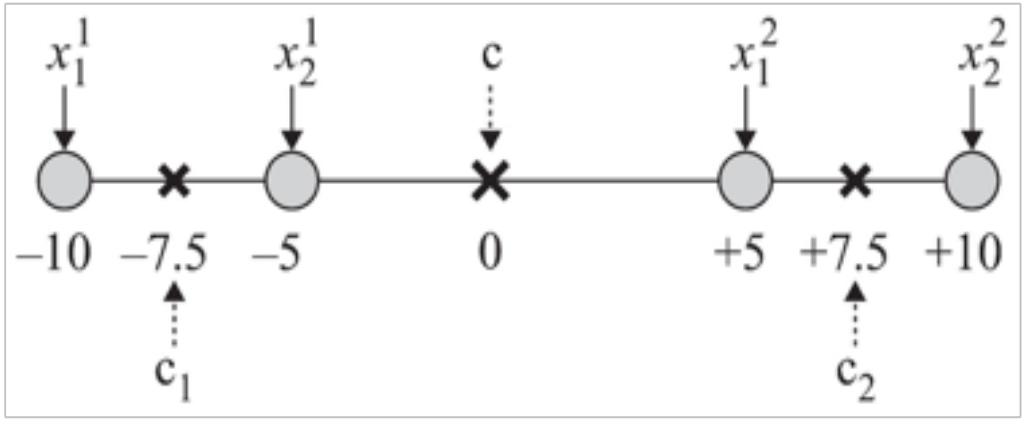
Separateness

- We are interested in clusters that are well separated from one another

Cohesiveness

- Being close to the centroid of the cluster

$$cohesiveness = \sum_{i=1}^{k} \sum_{j=1}^{n(i)} ||x_j^i - c_i||^2$$

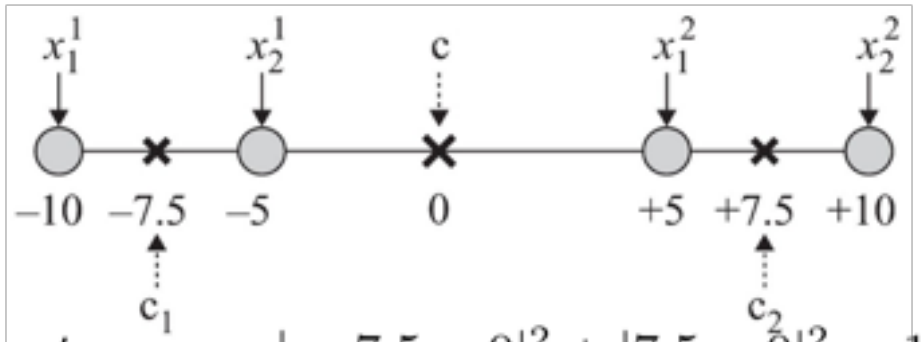


$$cohesiveness = |-10-(-7.5)|^2+|-5-(-7.5)|^2+|5-7.5)|^2+|10-7.5|^2 = 25$$

Separateness

- Cluster centroids being far from the mean of entire dataset

$$separateness = \sum_{i=1}^{k} ||c - c_i||^2$$



$$separateness = |-7.5-0|^2 + |7.5-0|^2 = 112.5$$

Silhouette Index

We are interested in clusters that are both cohesive and separate

Silhouette index

It compares:

- Average distance value between instances in the **same** cluster

To:

- Average distance value between instances in the **same** cluster

In a well-clustered dataset

- Average distance between instances in the same cluster is small (**cohesiveness**)

And

- Average distance between instances in different clusters is large (**separateness**)

Silhouette Index

For any instance *x* that is a member of cluster *C*

$$a(x) = \frac{1}{|C|-1} \sum_{y \in C, y \neq x} ||x - y||^2$$

- Compute the within-cluster average distance

Compute the average distance between *x in C* and instances in cluster *G*

$$b(x) = \min_{G \neq C} \frac{1}{|G|} \sum_{y \in G} ||x - y||^2$$

- *G* is closest to *x* in terms of the average distance between *x* in *C* and members of *G*

Silhouette Index

Our interest:
clusterings where $a(x) < b(x)$

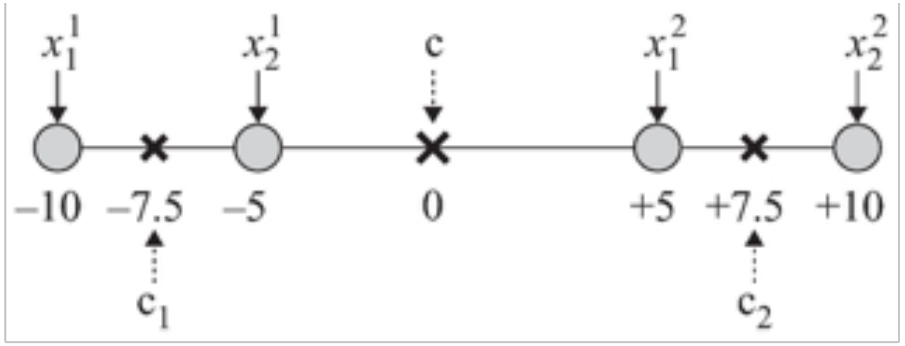
- Silhouette can take values between [-1,1]

The best case happens when for all x,
$-a(x) = 0$, $b(x) > a(x)$

$$s(x) = \frac{b(x) - a(x)}{\max(b(x), a(x))}$$

$$silhouette = \frac{1}{n} \sum_{x} s(x)$$

Silhouette Index: Example



$$a(x_1^1) = |-10-(-5)|^2 = 25 \qquad a(x_1^2) = |5-10|^2 = 25$$

$$b(x_1^1) = \frac{1}{2}(|-10-5|^2 + |-10-10|^2) = 312.5 \qquad b(x_1^2) = \frac{1}{2}(|5-(-10)|^2 + |5-(-5)|^2) = 162.5$$

$$s(x_1^1) = \frac{312.5-25}{312.5} = 0.92 \qquad s(x_1^2) = \frac{162.5-25}{162.5} = 0.84$$

$$a(x_2^1) = |-5-(-10)|^2 = 25 \qquad a(x_2^2) = |10-5|^2 = 25$$

$$b(x_2^1) = \frac{1}{2}(|-5-5|^2 + |-5-10|^2) = 162.5 \qquad b(x_2^2) = \frac{1}{2}(|10-(-5)|^2 + |10-(-10)|^2) = 312.5$$

$$s(x_2^1) = \frac{162.5-25}{162.5} = 0.84 \qquad s(x_2^2) = \frac{312.5-25}{312.5} = 0.92.$$