But we start with a brief discussion of the Friedman article and the relationship between Data Mining and Statistics
Data Mining and Statistics: What’s the Connection?

- What did you think about this 1997 article by Jerry Friedman, a noted statistician?
  - Did it help you understand the difference between these 2 fields?
  - What comments or points resonated with you?
  - Do you think this old article is extremely dated or that the points are no longer relevant?
Key points in Friedman Article

- Data Mining is a vaguely defined field
  - Numerous definitions
  - Defined partially by methods (decision trees, neural nets, nearest neighbor)
  - Associated with commercial products
  - Largely a commercial enterprise
    - To sell hardware and software
      - Initially hardware manufacturers were in the DM business to sell hardware (e.g., Silicon Graphics, IBM)
Friedman: Contents of DM Packages

- Data Mining Packages had:
  - Easy to use graphical interface and graphical output of models and results
  - Include certain methods
    - Classification methods, clustering, association analysis

- Data Mining Packages do **not** include:
  - Hypothesis testing and experimental design
  - ANOVA, Linear/logistic regression*, factor analysis
  * Not true now
Friedman: Is DM Intellectual Discipline?

- According to Friedman only the methodology is an intellectual discipline
  - But thinks in the future DM will be as computer power increases
- Can probably conclude it is now, but may not be as coherent as other disciplines
- How would you describe it as an intellectual discipline?
  - I would say that it emphasizes computational methods to analyze data
Friedman: Should Statistics include DM?

- If DM part of Stats then:
  - DM published in Stats journals
  - DM taught in stats departments
  - Answer unclear (then)

- Many disciplines started in stats and perhaps could have stayed there:
  - Pattern recognition, machine learning, neural networks, data visualization
Friedman: What is Statistics?

- Statistics defined by a set of tools
  - Probability theory, Real analysis, Decision theory

- Probabilistic inference based on mathematics
  - Some felt that stats should remain focused on this area of success
  - Stats journals required proofs; DM rarely uses proofs

- Statistics was not defined by a set of problems, namely data analysis
  - Methods not using probability not included
  - This is an odd way to define a field, most disciplines defined by set of problems it addresses
What Happened to Statistics?

- Friedman raised issues but did not say what should happen.
  - He listed possible futures and its implications

- What actually happened?
  - Clearly DM took off on its own.
  - Largely defined by problems it can address
  - Includes statistics but there is a preference for computational/algorithmic methods
    - May diminish in time if data science takes over.

- In 2018 Stats programs still do not incorporate data mining (but require programming)
What is data exploration?

- A preliminary exploration of the data to better understand its characteristics.

- Key motivations of data exploration include
  - Helping to select the right tool for preprocessing or analysis
  - Making use of humans’ abilities to recognize patterns
    - People can recognize patterns not captured by data analysis tools

- Related to the area of Exploratory Data Analysis (EDA)
  - Created by statistician John Tukey
  - Seminal book is Exploratory Data Analysis by Tukey
  - A nice online introduction can be found in Chapter 1 of the NIST Engineering Statistics Handbook

Techniques Used in Data Exploration

- In EDA, as originally defined by Tukey
  - The focus was on visualization
  - Clustering and anomaly detection were viewed as exploratory techniques
  - In data mining, clustering and anomaly detection are major areas of interest, and not thought of as just exploratory

- We will focus on
  - Summary statistics
  - Visualization
    - Note we offer a course in data visualization
  - Online Analytical Processing (OLAP)
Iris Sample Data Set

- Many of the exploratory data techniques are illustrated with the Iris Plant data set.
  - Can be obtained from the UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html](http://www.ics.uci.edu/~mlearn/MLRepository.html)
  - From the statistician Douglas Fisher
  - Three flower types (classes):
    - Setosa
    - Virginica
    - Versicolour
  - Four (non-class) attributes
    - Sepal width and length
    - Petal width and length

Summary Statistics

- Summary statistics are numbers that summarize properties of the data
  - Summarized properties include frequency, location and spread
    - Examples: location - mean
                 spread - standard deviation
  - Most summary statistics can be calculated in a single pass through the data
Frequency and Mode

- The frequency of an attribute value is the percentage of time the value occurs in the data set
  - Given the attribute ‘gender’ and a representative population, ‘female’ occurs about 50% of the time
- The mode of an attribute is most frequent value
- The notions of frequency and mode are typically used with categorical data
For continuous data, the notion of a percentile is more useful

- For instance, the 50th percentile for attribute x is the x-value such that 50% of all values of x are less than it
Measures of Location: Mean & Median

- The mean is the most common measure of the location of a set of points.
  - The mean is very sensitive to outliers.
  - Thus the median is also commonly used
    - Median is the middle value when the values are sorted

\[
\text{mean}(x) = \bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i
\]

\[
\text{median}(x) = \begin{cases} 
  x_{(r+1)} & \text{if } m \text{ is odd}, \text{i.e., } m = 2r + 1 \\
  \frac{1}{2}(x_r + x_{(r+1)}) & \text{if } m \text{ is even}, \text{i.e., } m = 2r
\end{cases}
\]
Measures of Spread: Range & Variance

- Range is the difference between the max and min
- The variance or standard deviation is the most common measure of the spread of a set of points.

\[
\text{variance}(x) = s^2_x = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})^2
\]

- However, this is also sensitive to outliers, so that other measures are often used.

\[
\text{AAD}(x) = \frac{1}{m} \sum_{i=1}^{m} |x_i - \bar{x}|
\]

\[
\text{MAD}(x) = \text{median} \left( \{|x_1 - \bar{x}|, \ldots, |x_m - \bar{x}|\} \right)
\]

\[
\text{interquartile range}(x) = x_{75\%} - x_{25\%}
\]

Bessel’s correction: \(m-1\) is used rather than \(m\) when the true population mean is not known and partially corrects for the resulting bias.

Visually shown in box plots (and in SEEQ results)
Visualization

Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.

- Visualization of data is one of the most powerful and appealing techniques for data exploration.
  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns
Example: Sea Surface Temperature

- Sea Surface Temperature (SST) for July 1982
  - Tens of thousands of data points summarized in single figure
Representation

- Is the mapping of information to a visual format
- Data objects, their attributes, and the relationships among data objects are translated into graphical elements such as points, lines, shapes, and colors.

Example:

- Objects are often represented as points
- Their attribute values can be represented as the position of the points or the characteristics of the points, e.g., color, size, and shape
- If position is used, then the relationships of points, i.e., whether they form groups or a point is an outlier, is easily perceived.
Arrangement

- Is the placement of visual elements within a display
- Can make a large difference in how easy it is to understand the data
- Example:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>6</th>
<th>1</th>
<th>3</th>
<th>2</th>
<th>5</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Selection

- Is the elimination or the de-emphasis of certain objects and attributes
- Selection may involve choosing a subset of attributes
  - Dimensionality reduction is often used to reduce the number of dimensions to two or three
  - Alternatively, pairs of attributes can be considered
- Selection may also involve choosing a subset of objects
  - A region of the screen can only show so many points
  - Can sample, but want to preserve points in sparse areas
Visualization Techniques: Histograms

- **Histogram**
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins

- **Example:** Petal Width (10 and 20 bins, respectively)
Two-Dimensional Histograms

- Show joint distribution of values of 2 attributes
- Example: petal width and petal length
  - What does this tell us?
Visualization Techniques: Box Plots

- **Box Plots**
  - Invented by J. Tukey
  - Another way of displaying the distribution of data
  - Following figure shows the basic part of a box plot
Example of Box Plots

- Box plots can be used to compare attributes
Visualization Techniques: Scatter Plots

- Scatter plots
  - Attributes values determine the position
  - Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
  - Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
  - It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes
    - See example on the next slide
Scatter Plot of Iris Attributes
Visualization Techniques: Contour Plots

- Contour plots
  - Useful when a continuous attribute is measured on a spatial grid
  - They partition the plane into regions of similar values
  - The contour lines that form the boundaries of these regions connect points with equal values
  - The most common example is contour maps of elevation
  - Can also display temperature, rainfall, air pressure, etc.
    - An example for Sea Surface Temperature (SST) is provided on the next slide
Contour Plot: SST Dec, 1998
Visualization Techniques: Parallel Coordinates

- **Parallel Coordinates**
  - Used to plot the attr. values of high-dimensional data
  - Uses parallel axes rather than perpendicular axes
  - The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line
  - Thus, each object is represented as a line
  - Often, the lines representing a distinct class of objects group together, at least for some attributes
  - Ordering of attributes is important in seeing such groupings
Parallel Coordinates Plots for Iris Data
Other Visualization Techniques

- **Star Plots**
  - Similar approach to parallel coordinates, but axes radiate from a central point
  - The line connecting the values of an object is a polygon

- **Chernoff Faces**
  - Approach created by Herman Chernoff
  - This approach associates each attribute with a characteristic of a face
  - The values of each attribute determine the appearance of the corresponding facial characteristic
  - Each object becomes a separate face
  - Relies on human’s ability to distinguish faces
Star Plots for Iris Data

Setosa

Versicolour

Virginica
Chernoff Faces for Iris Data

Setosa

1, 51, 101

Versicolour

2, 52, 102

Virginica

3, 53, 103

4, 54, 104

5, 55, 105
OLAP

- This is not data mining but often mentioned in a data mining course
  - For now just become familiar with the terminology and basic idea
- On-Line Analytical Processing (OLAP) was proposed by E. F. Codd, the father of the relational database.
- Relational databases put data into tables, while OLAP uses a multidimensional array.
- There are a number of data analysis and data exploration operations that are easier with such a data representation.
Creating a Multidimensional Array

- Two key steps in converting tabular data into a multidimensional array.
  - First, identify which attributes are to be the dimensions and which attribute is to be the target attribute whose values appear as entries in the multidimensional array.
    - Attributes used as dimensions must have discrete values
    - The target value is typically a count or continuous value, e.g., the cost of an item
    - Can have no target variable at all except the count of objects that have the same set of attribute values
  - Second, find the value of each entry in the multidimensional array by summing the values (of the target attribute) or count of all objects that have the attribute values corresponding to that entry.
Example: Iris data

- We show how the attributes, petal length, petal width, and species type can be converted to a multidimensional array
  - First, we discretized the petal width and length to have categorical values: *low*, *medium*, and *high*
  - We get the following table - note the count attribute

<table>
<thead>
<tr>
<th>Petal Length</th>
<th>Petal Width</th>
<th>Species Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>low</td>
<td>Setosa</td>
<td>46</td>
</tr>
<tr>
<td>low</td>
<td>medium</td>
<td>Setosa</td>
<td>2</td>
</tr>
<tr>
<td>medium</td>
<td>low</td>
<td>Setosa</td>
<td>2</td>
</tr>
<tr>
<td>medium</td>
<td>medium</td>
<td>Versicolour</td>
<td>43</td>
</tr>
<tr>
<td>medium</td>
<td>high</td>
<td>Versicolour</td>
<td>3</td>
</tr>
<tr>
<td>medium</td>
<td>high</td>
<td>Virginica</td>
<td>3</td>
</tr>
<tr>
<td>high</td>
<td>medium</td>
<td>Versicolour</td>
<td>2</td>
</tr>
<tr>
<td>high</td>
<td>medium</td>
<td>Virginica</td>
<td>3</td>
</tr>
<tr>
<td>high</td>
<td>high</td>
<td>Versicolour</td>
<td>2</td>
</tr>
<tr>
<td>high</td>
<td>high</td>
<td>Virginica</td>
<td>44</td>
</tr>
</tbody>
</table>
Example: Iris data (continued)

- Each unique tuple of petal width, petal length, and species type identifies one element of the array.
- This element is assigned the corresponding count.
- All non-specified tuples are 0.
OLAP Operations: Data Cube

- The key operation of a OLAP is the formation of a data cube.
- A data cube is a multidimensional representation of data, together with all possible aggregates.
- For example, if we choose the species type dimension of the Iris data and sum over all other dimensions, the result will be a one-dimensional entry with three entries, each of which gives the number of flowers of each type.
Data Cube Example

- Consider a data set that records the sales of products at a number of company stores at various dates.
- This data can be represented as a 3-dimensional array.
- There are 3 two-dimensional aggregates (3 choose 2), 3 one-dimensional aggregates, and 1 zero-dimensional aggregate (the overall total).
OLAP Operations: Slicing & Dicing

- Slicing is selecting a group of cells from the entire multidimensional array by specifying a specific value for one or more dimensions.
- Dicing involves selecting a subset of cells by specifying a range of attribute values.
  - This is equivalent to defining a subarray from the complete array.
- In practice, both operations can also be accompanied by aggregation over some dimensions.
OLAP Operations: Roll-up & Drill-down

- Attribute values often have a hierarchical structure.
  - Each date is associated with a year, month, and week.
  - A location is associated with a continent, country, state (province, etc.), and city.
  - Products can be divided into various categories, such as clothing, electronics, and furniture.

- Note that these categories often nest and form a tree or lattice
  - A year contains months which contains day
  - A country contains a state which contains a city
OLAP Operations: Roll-up & Drill-down

- This hierarchical structure gives rise to the roll-up and drill-down operations.
  - For sales data, we can aggregate (roll up) the sales across all the dates in a month.
  - Conversely, given a view of the data where the time dimension is broken into months, we could split the monthly sales totals (drill down) into daily sales totals.
  - Likewise, we can drill down or roll up on the location or product ID attributes.