
# CS 644: Introduction to Big Data

Daqing Yun (daqing.yun@njit.edu)
New Jersey Institute of Technology

# Outline

- Data Exploration
  - Objects, Types, Attributes, Values, etc.
  - Summary Statistics
  - Visualization
  - Quality Issues
- Data Preprocessing
  - Data Understanding
  - Data Preparation
  - A Roadmap of Available Techniques
- A List of Major Issues

Some of these slides are based on: Tan et al., Introduction to Data Mining; and many other online resources (links are included at the bottom of the slides)

# The Process


**Data Sourcing/Representation**
Where are the dataset from

**Data Cleaning**
To remove noise and inconsistent data

**Data Integration**
Multiple data sources may be combined

**Data Selection/Reduction**
Data relevant to the analysis task are retrieved from the database

**Data Transformation/Consolidation/Reduction**
Data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations

Data Preprocessing

**Data Mining**
The essential process where intelligent methods are applied to extract data patterns

**Pattern Evaluation**
To identify the truly interesting patterns representing knowledge

**Knowledge Presentation**
Visualization and knowledge representation techniques are used to present mined knowledge to users

# Data Exploration

# Data Objects

- Collection of data *objects* and their *attributes*
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

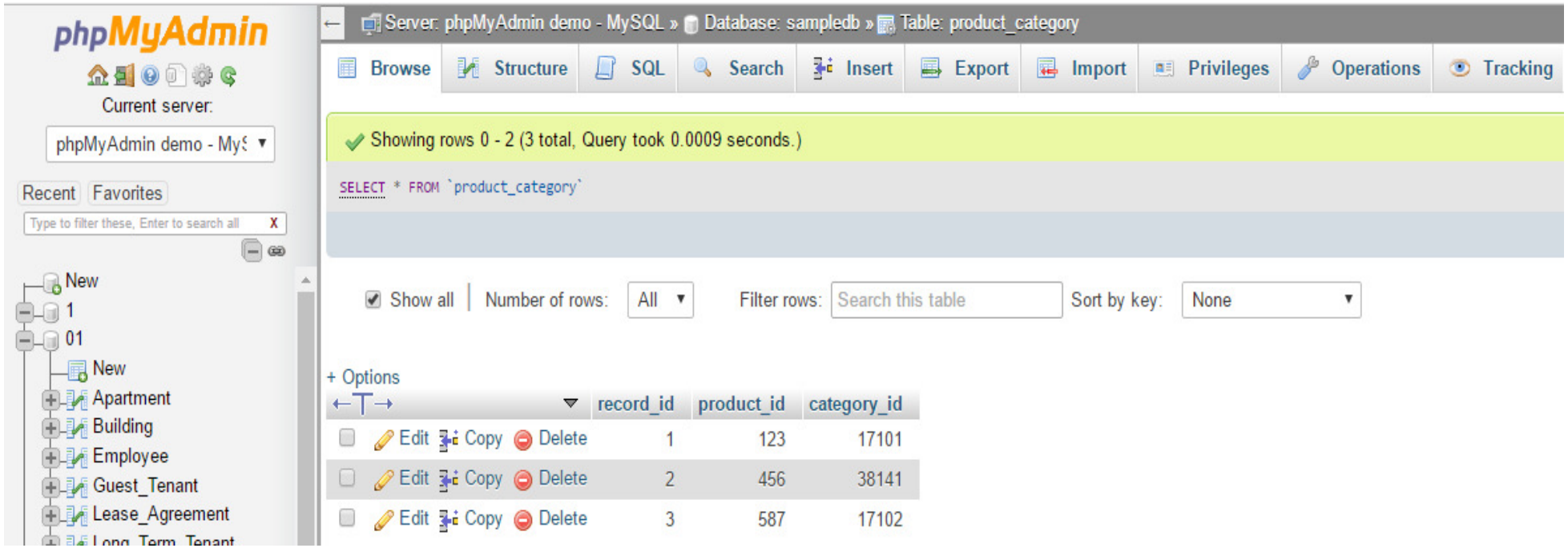
*Attributes*

*Objects*

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

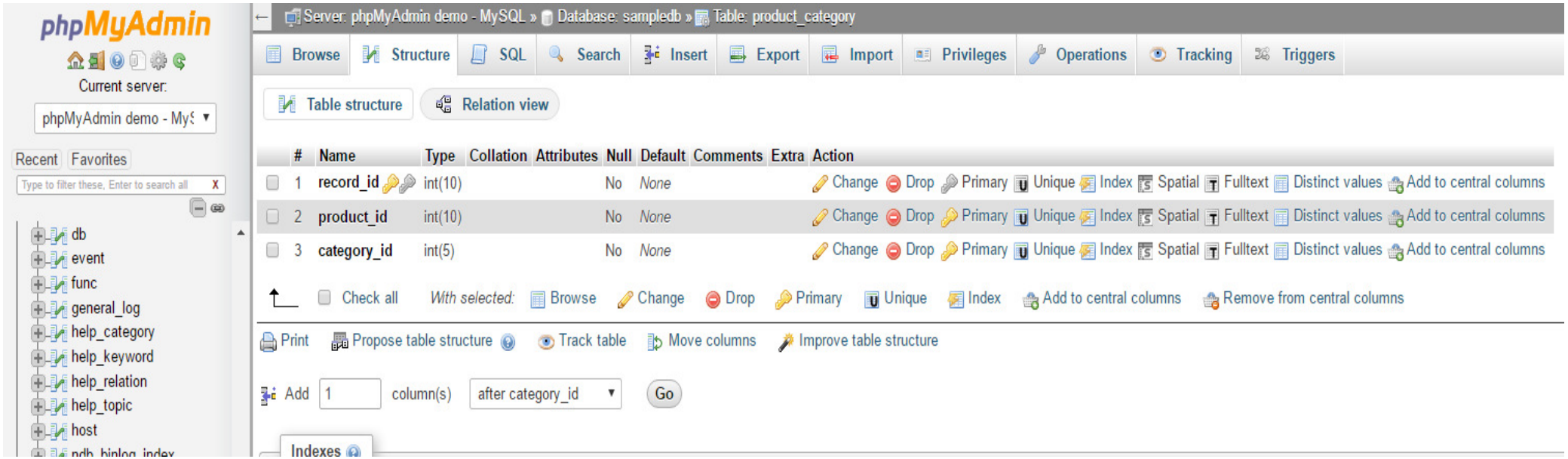
# Data Objects

- Data objects are typically described by attributes, e.g.,
  - Data tuples if stored in a database, e.g., MySQL
  - The rows of database correspond to data objects



# Attributes

- An attribute is a data field, a characteristic or feature of data object
- Attribute, dimension, feature, and variable are often used interchangeably
- A set of attributes used to describe a given object is called an attribute vector (or feature vector)



# Attribute Values

- Attribute values are numbers or symbols assigned to an attribute
- Distinction between attributes and attributes values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: attribute values for ID and age are integers
    - But properties of attribute values can be different, e.g., ID has no limit but age has a maximum and minimum value

# Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: IDs, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1 to 10), grades, height in {tall, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit
  - Ratio
    - Examples: temperatures in Kelvin, length, time, counts

# Properties of Attributes Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Addition: + -
  - Multiplication: * /

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order, & addition
  - Ratio attribute: all 4 properties

# Types of Data

- Different types of attributes

| | Attribute Type | Description | Examples | Operations |
|---|---|---|---|---|
| Categorical Qualitative | Nominal | Nominal attribute values only distinguish. (=, ≠) | zip codes, employee ID numbers, eye color, sex: {*male, female*} | mode, entropy, contingency correlation, $\chi 2$ test |
| | Ordinal | Ordinal attribute values also order objects. (<, >) | hardness of minerals, {*good, better, best*}, grades, street numbers | median, percentiles, rank correlation, run tests, sign tests |
| Numeric Quantitative | Interval | For interval attributes, differences between values are meaningful. (+, - ) | calendar dates, temperature in Celsius or Fahrenheit | mean, standard deviation, Pearson's correlation, *t* and *F* tests |
| | Ratio | For ratio variables, both differences and ratios are meaningful. (*, /) | temperature in Kelvin, monetary quantities, counts, age, mass, length, current | geometric mean, harmonic mean, percent variation |

# Types of Data

- Permissible transformations

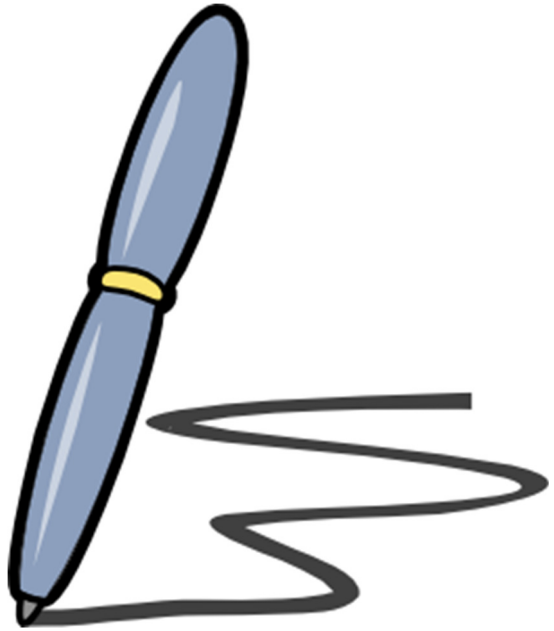
| | Attribute Type | Transformation | Comments |
|---|---|---|---|
| Categorical Qualitative | Nominal | Any permutation of values | If all employee ID numbers were reassigned, would it make any difference? |
| | Ordinal | An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function | An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}. |
| Numeric Quantitative | Interval | *new_value = a * old_value + b* where a and b are constants | Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree). |
| | Ratio | *new_value = a * old_value* | Length can be measured in meters or feet. |

# Discrete and Continuous Attributes

- Discrete attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables
  - Binary attributes are a special case of discrete attributes
- Continuous attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

# Types of Data Sets

- **Record data**
  - Data matrix
  - Document data
  - Transaction data
- **Graph-based data**
  - WWW
  - Molecular structure
- **Ordered data**
  - Spatial data
  - Temporal data
  - Sequential data
  - Genetic sequence data



# Important Characteristics of Structured Data

- Dimensionality
  - Number of attributes that the objects possess
  - Curse of dimensionality
- Sparsity
  - Only presence counts
  - Practically an advantage as usually only present data need to be stored and manipulated
- Resolution
  - Patterns depend on the scale
  - e.g., surface of the Earth

# Record Data

- Data that consist of a collection of records, each of which consist of a fixed set of attributes

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

# Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such data set can be represented by an *m* by *n* matrix, where there are *m* rows, one for each record, and *n* columns, one for each attributes

| Projection of x Load | Projection of y load | Distance | Load | Thickness |
|---|---|---|---|---|
| 10.23 | 5.27 | 15.22 | 2.7 | 1.2 |
| 12.65 | 6.25 | 16.22 | 2.2 | 1.1 |

# Document Data

- Each document becomes a “term” vector
  - Each term is a component (attribute) of the vector
  - The value of each component is the number of times the corresponding term occurs in the document

| | team | coach | play | ball | score | game | win | lost | timeout | season |
|---|---|---|---|---|---|---|---|---|---|---|
| Document 1 | 3 | 0 | 5 | 0 | 2 | 6 | 0 | 2 | 0 | 2 |
| Document 2 | 0 | 7 | 0 | 2 | 1 | 0 | 0 | 3 | 0 | 0 |
| Document 3 | 0 | 1 | 0 | 0 | 1 | 2 | 2 | 0 | 3 | 0 |

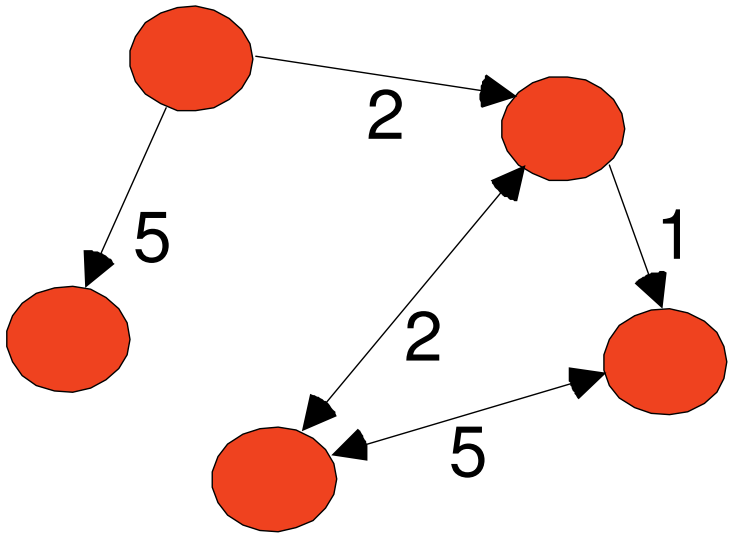
# Transaction Data

- A special type of record data, where
  - Each record (transaction) involves a set of items
  - Consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items

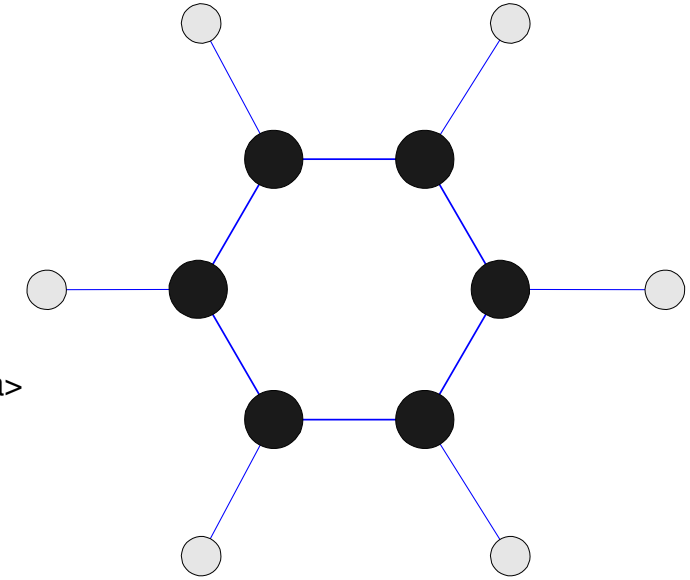
| *TID* | *Items* |
|---|---|
| 1 | Bread, Coke, Milk |
| 2 | Beer, Bread |
| 3 | Beer, Coke, Diaper, Milk |
| 4 | Beer, Bread, Diaper, Milk |
| 5 | Coke, Diaper, Milk |

# Graph Data

- Examples: generic graph, HTML links, and a molecule




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```



Benzene molecule: $C_6H_6$

# Ordered Data

- Attributes have relationships that involve order in time or space
- Sequences of transactions

| Time | Customer | Items Purchased |
|---|---|---|
| t1 | C1 | A, B |
| t2 | C3 | A, C |
| t2 | C1 | C, D |
| t3 | C2 | A, D |
| t4 | C2 | E |
| t5 | C1 | A, E |

| Customer | Time and Items Purchased |
|---|---|
| C1 | (t1: A,B) (t2:C,D) (t5:A,E) |
| C2 | (t3: A, D) (t4: E) |
| C3 | (t2: A, C) |

# Ordered Data

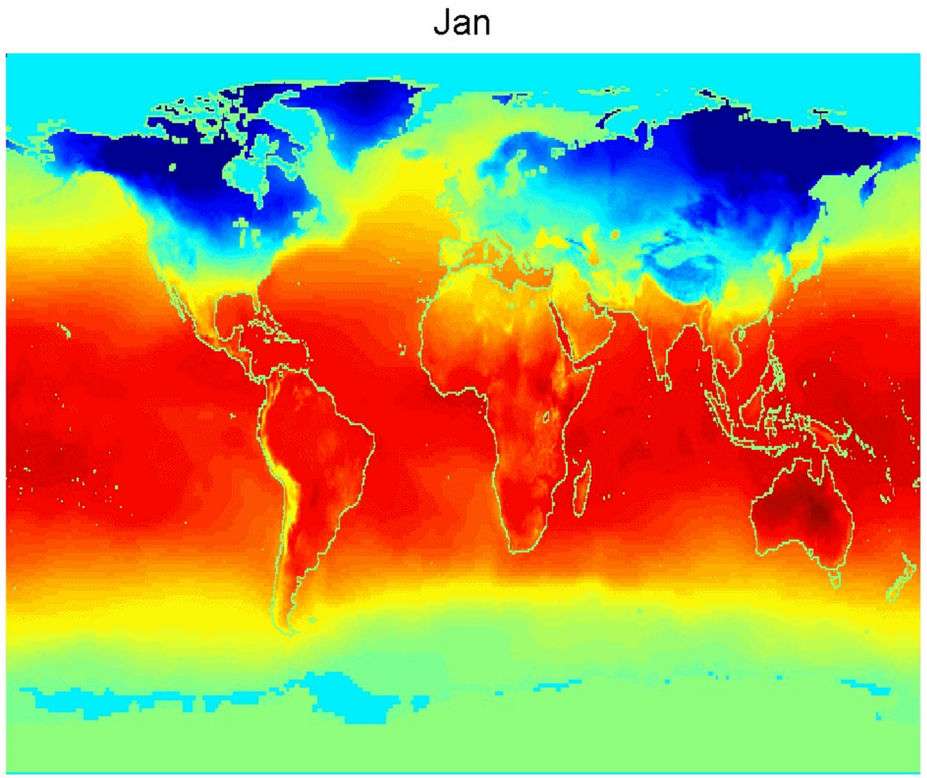
- Attributes have relationships that involve order in time or space
- Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

# Ordered Data

- Attributes have relationships that involve order in time or space
- Spatial-temporal data

Average Monthly Temperature of land and ocean



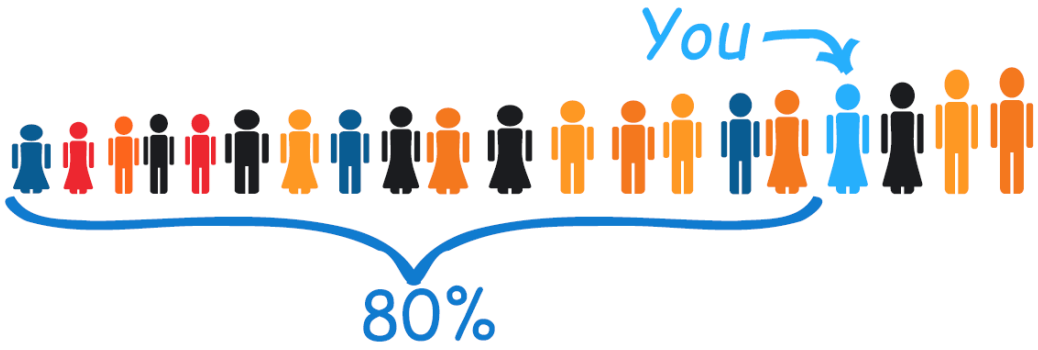
# Summary Statistics – Frequencies

- Given a set of unordered categorical values, there is not much that can be done to further characterize the values except to compute the frequency with which each value occurs for a particular set of data
- Given a categorical attribute $x$, which can take values $\{v_1, v_2, \ldots, v_k\}$ and a set of $m$ objects, the frequency of a value $v_i$ is defined as

$$\text{frequency}(v_i) = \frac{\text{number of objects with attribute value } v_i}{m}$$

# Summary Statistics – Percentiles

- For ordered data, it is more useful to consider the percentiles of a set of values
- Given an ordinal or continuous attribute $x$ and a number of $p$ between 0 and 100, the $p^{th}$ percentile $x_p$ is a value of $x$ such that $p\%$ of the observed values of $x$ are less than $x_p$



# Summary Statistics – Central Tendency

- Mean – the most common and effective numeric measure of the *center* of a set of data is the (arithmetic) *mean*

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

- Median – the data element in the *middlemost*, might not be unique if number of elements is even, average of the two middlemost values

- Mode – the value that occurs most frequently in the data set; data sets with one, two, and three modes are respectively called *unimodal*, *bimodal*, and *trimodal*

# Summary Statistics – Dispersion

- Indication of how spread out a data distribution is
- Variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = (\frac{1}{N}\sum_{i=1}^{N}x_i^2) - \bar{x}^2$$

where $\bar{x}$ is the mean value of the observations

- Standard deviation
  - the square root of variance

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} = \sqrt{(\frac{1}{N}\sum_{i=1}^{N}x_i^2) - \bar{x}^2}$$

# Multivariate Summary Statistics – Covariance Matrix

- For multivariate data with continuous values, the spread of the data is commonly captured by the covariance matrix *S*,

$$s_{ij} = \text{covariance}(x_i, x_j) = \frac{1}{m-1}\sum_{k=1}^{m}(x_{ki} - \overline{x_i})(x_{kj} - \overline{x_j})$$

*ij*th entry of *S*

number of objects

values of the *i*th attribute for the *k*th object

- Note that covariance($x_i$, $x_i$) = variance($x_i$), the covariance matrix has the variances of the attributes along the diagonal

# **Multivariate Summary Statistics –** Correlation Matrix

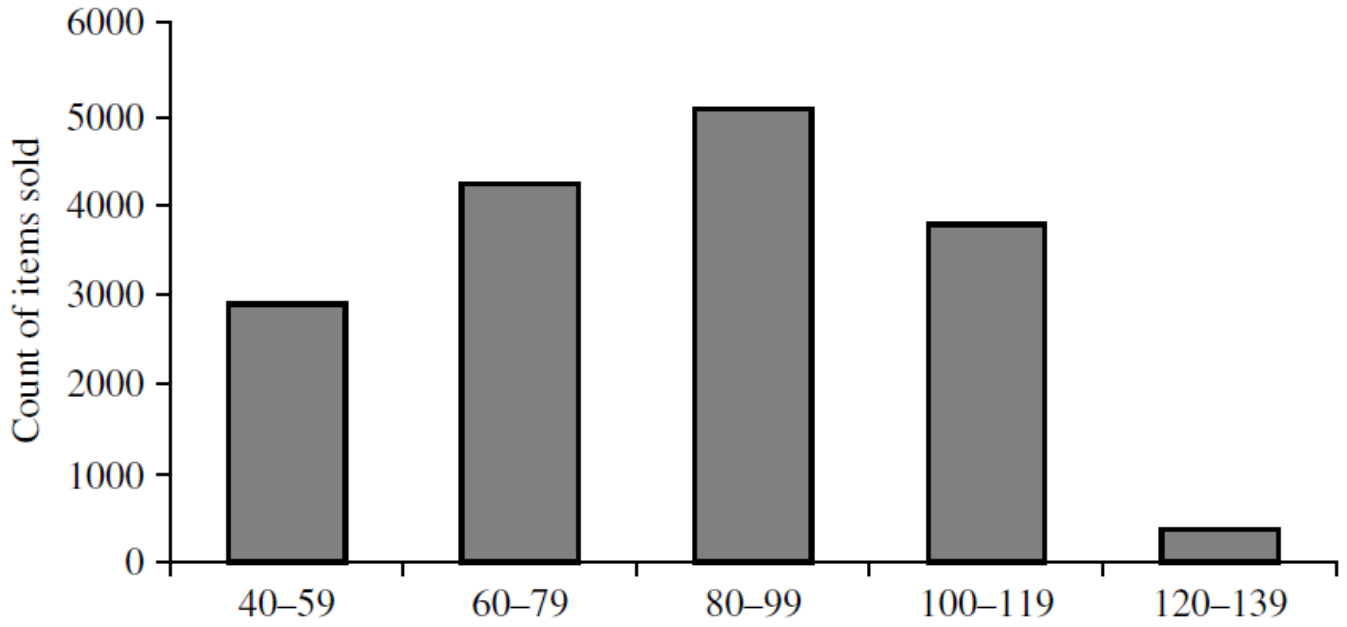
- The covariance of two attributes is a measure of the degree to which two attributes vary together and depends on the magnitudes of the variables
  - A value near 0 indicates that two attributes do not have a (linear) relationship
  - But it is not possible to judge the degree of relationship between two variables by looking only at the value of the covariance
- The correlation of two attributes immediately gives an indication of how strongly two attributes are (linearly) related, correlation is preferred to covariance for data exploration
  - The $ij^{th}$ entry of the correlation matrix $R$, is the correlation between the $i^{th}$ and $j^{th}$ attributes of the data

$$r_{ij} = \text{correlation}(x_i, x_j) = \frac{\text{covariance}(x_i, x_j)}{s_i s_j}$$

variances of $x_i$ and $x_j$

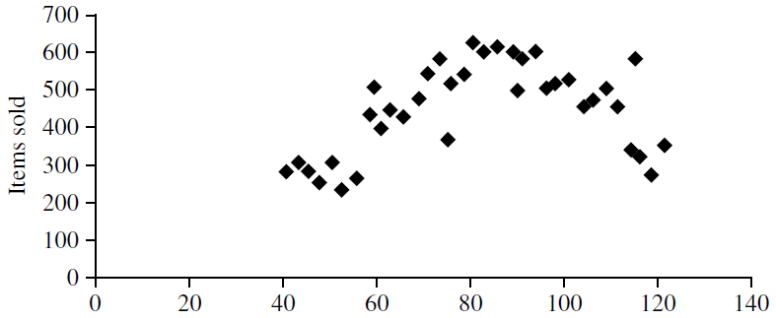
# Visualization – Graphic Displays

- Frequency histograms – the resulting graph is more commonly known as *bar chart*

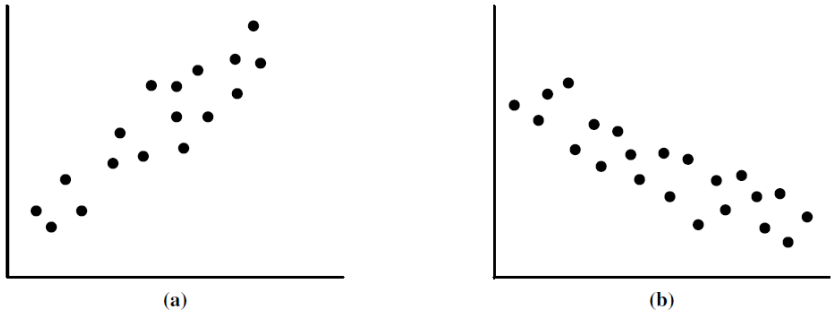


# Visualization – Graphic Displays

- Scatter plots – each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane



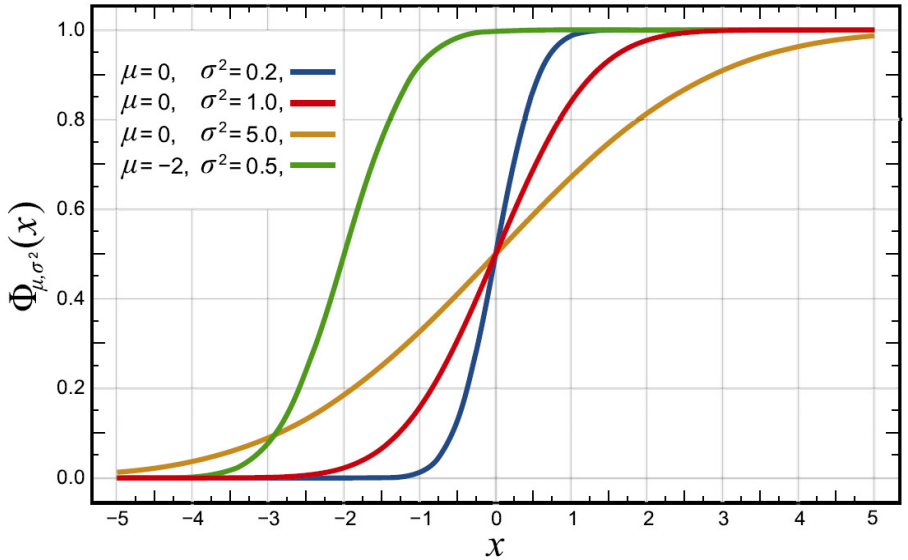

- Scatter plots can be used to find (a) positive or (b) negative correlations



# Visualization – Graphic Displays

## Plot of Empirical Cumulative Distribution Function (ECDF)

- A type of diagram that shows the distribution of the data quantitatively
- A CDF shows the probability that a point is less than a value
- An ECDF shows the fraction of points that are less than a value
- The ECDF is a step function



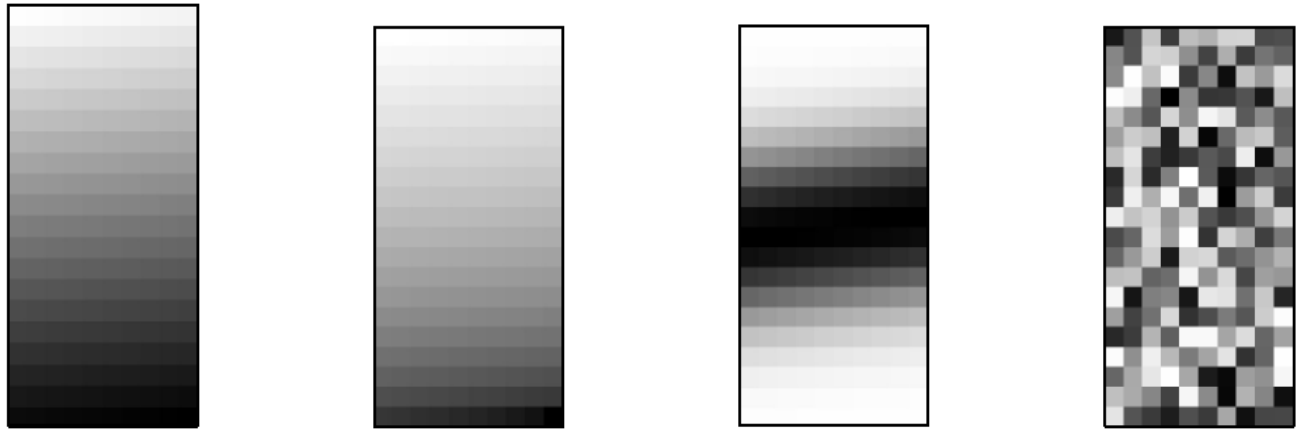

Cumulative distribution function for the normal distribution

# Visualization – Graphic Displays

- Visualization and knowledge representation techniques are used to present mined knowledge to users
- Data visualization aims to communicate data clearly and effectively through graphical representation
  - Pixel-oriented visualization techniques
  - Geometric projection visualization techniques
  - Visualizing complex data and relations
  - …

# Visualization – Graphic Displays

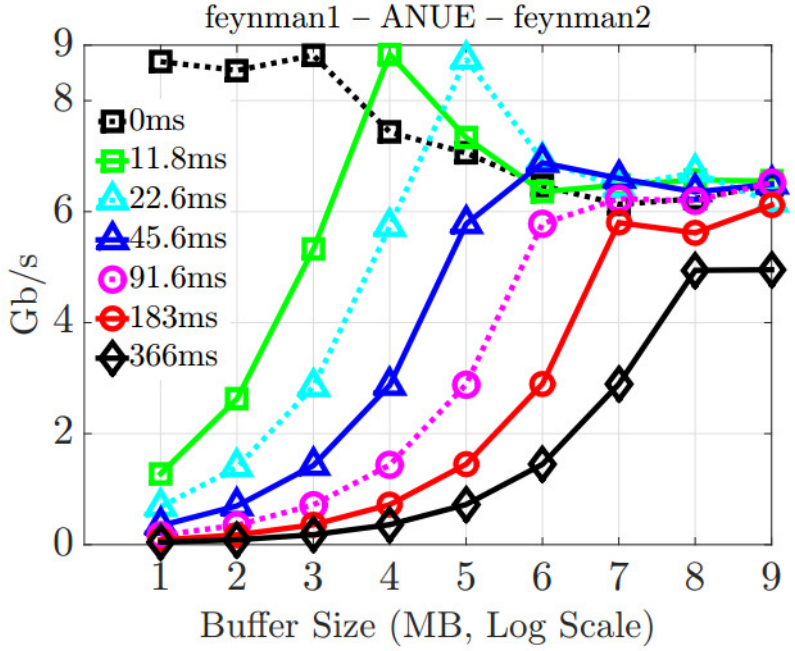
- Pixel-oriented visualization techniques
  - A simple way to visualize the value of a dimension is to use a pixel where the color of the pixel reflects the dimension's value
  - For example, a customer information table, which consists of four dimensions: income, credit limit, transaction volume, and age. Can we analyze the correlation between income and the other attributes by visualization?



(a) *income* (b) *credit_limit* (c) *transaction_volume* (d) *age*

# Visualization – Graphic Displays

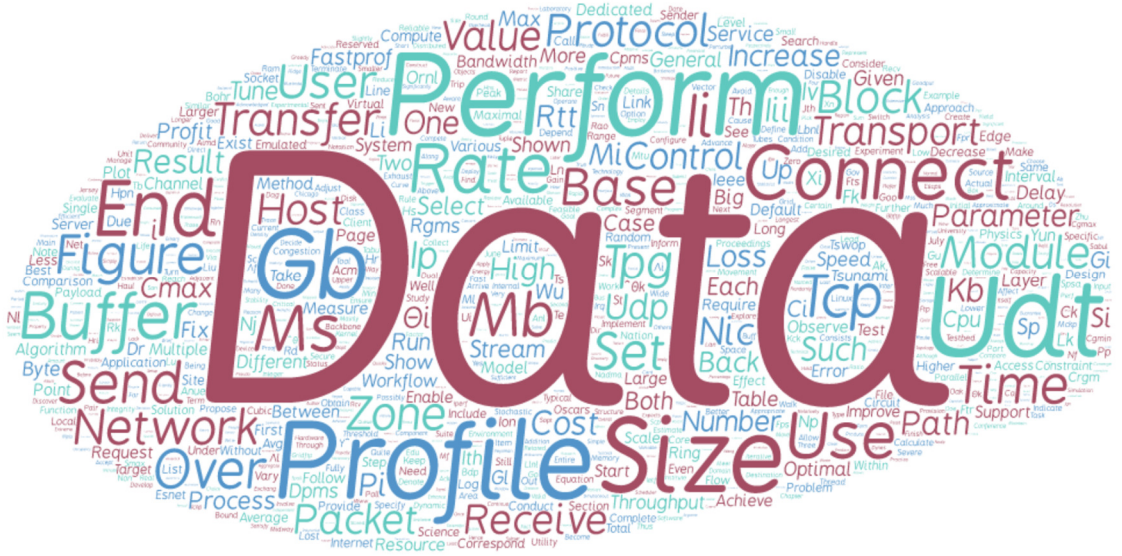
- Geometric projection visualization techniques
  - Help users find interesting projections of multidimensional datasets
  - The central challenge of the geometric projection techniques is how to visualize a high-dimensional space on a 2-D display
  - For example, scatter plot




D. Yun *etal.*, Profiling Optimization for Big Data Transfer Over Dedicated Channels. In Proc. of ICCCN 2016.

# Visualization – Graphic Displays

- Visualizing complex data and relations
  - Visualize non-numeric data, such as text and social networks
  - For example, word cloud generated based on some document




Generated based on Yun's dissertation "An Integrated Transport Solution to Big Data Movement in High-performance Networks", 2016.

# Data Quality Issues

- Reasons cause inaccurate, incomplete, inconsistent data
  - Faulty data collection instruments
  - Occurrences of human or computer errors
  - Disguised missing data, e.g., users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal information
  - Technology limitations, e.g., errors in data transmission
  - Naming convention inconsistencies
  - Data code inconsistencies
  - Format inconsistencies
  - Attributes of interest may not always be available, e.g., customer information for sales transactions
  - Malfunctions of equipment
  - …

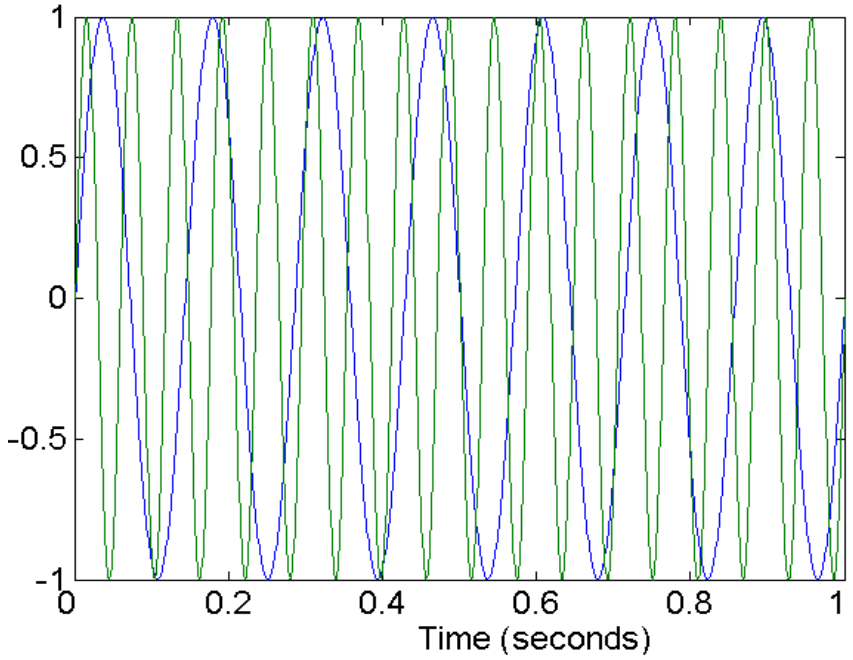
# Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

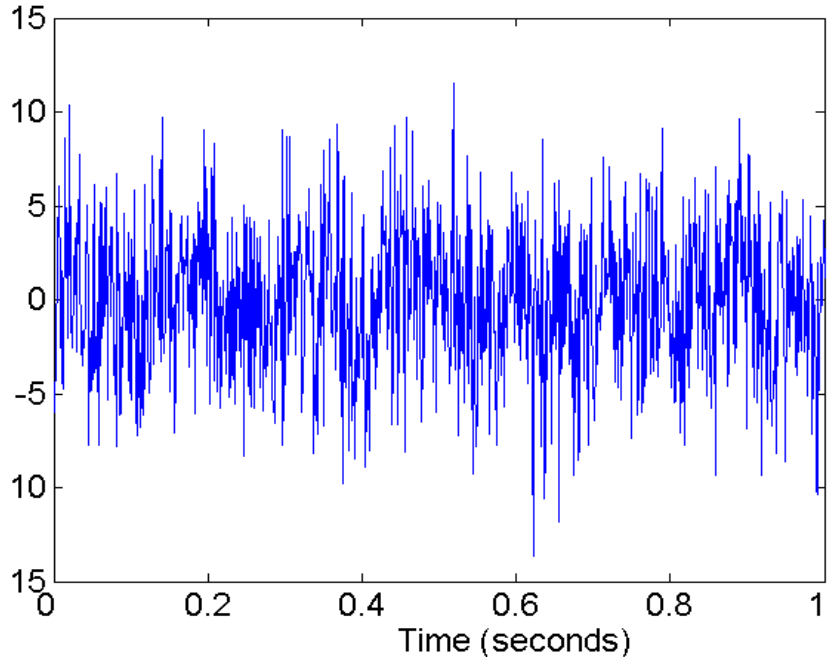
- Examples of data quality problems:
  - Noise and outliers
  - Missing values
  - Duplicate data

# Noise

- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on the television screen



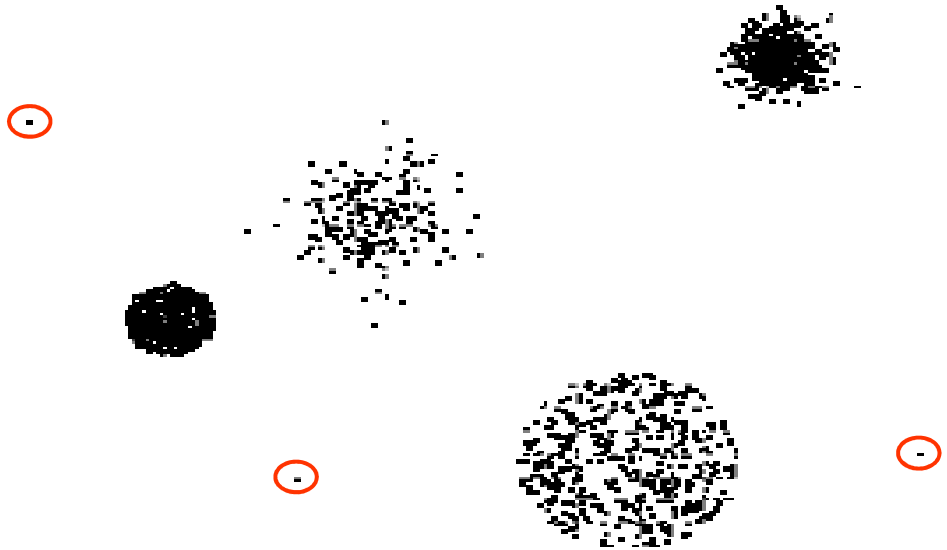

Two Sine Waves




Two Sine Waves + Noise

# Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



# Missing Values

- Reason for missing values
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., yearly income is not applicable to children)

- Handling missing values, e.g.,
  - Eliminate data objects
  - Estimate missing values
  - Ignore the missing value during analysis
  - Replace with all possible values (weighted by their probabilities)

# Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issues when merging data from heterogeneous sources ← Data Integration
- Examples
  - Same person with multiple email address
  - "`customer_id`" and "`cust_id`", the same attribute?
- Data cleaning
  - Process of dealing with duplicate data issues

# Reading

- Read Chapter 2 and Chapter 3 of the textbook *Introduction to Data Mining,* by P. Tan et al.

# Data Preprocessing

# Why?

- If you are a database manager
  - A body of data should be regarded as very clean
- If you are a data miner
  - You may have to fix many problems

- If, for example, having missing values in datasets
  - Not a problem for a database manager since what does not exist does not have to be stored
  - For a data miner, what does not exist in one field of a record might cause the entire record to be omitted from the analysis

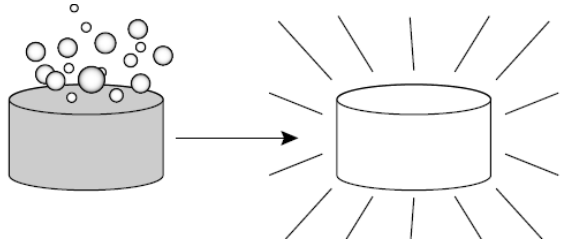
# Data Preprocessing

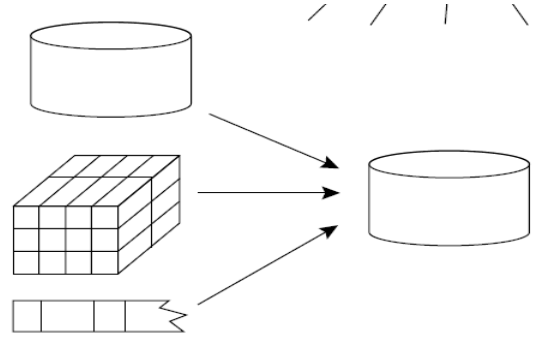
- Goal
  - To make the data more suitable for data mining
  - To improve the data mining analysis with respect to time, cost, and quality
  - To avoid “garbage in, garbage out”

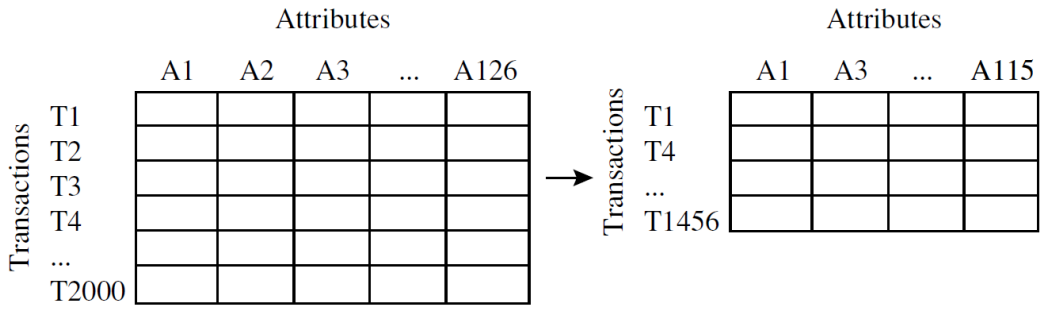
- Categories
  - Selecting objects and attributes for the analysis
  - Creating/changing the attributes

# Data Preprocessing – Major Tasks

- Data Cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data Integration
  - Include multiple sources (e.g., multiple databases, data cubes, and files) in data analysis
- Data Reduction
  - Obtain a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results, including dimensionality reduction and numerosity reduction

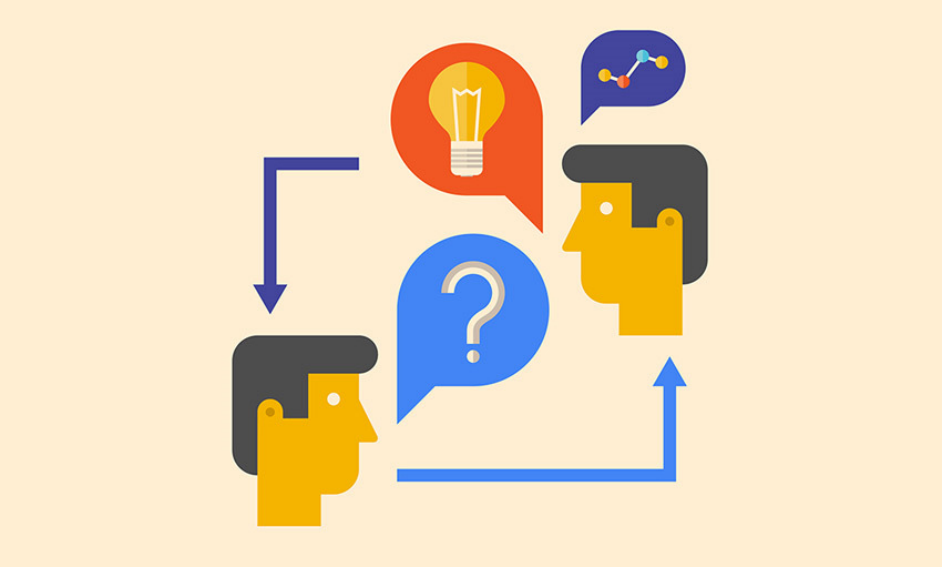






# Data Preprocessing – Techniques

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature Subset Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformation
- …



# Data Understanding

- Collect Initial Data
  - Access and load the data
  - Joining data from multiple sources and rationalizing it into on dataset
- Describe Data
  - Descriptive statistics
  - The structure of the data
- Explore Data
  - Data exploration offers an early view into the data
  - Visualizing the data to look for patterns in the data
  - A number of data issues can be uncovered during this step
  - Possibly formulate hypotheses that could lead to new data collection and experiments
- Verify Data Quality
  - Errors, outliers, and missing values/observations

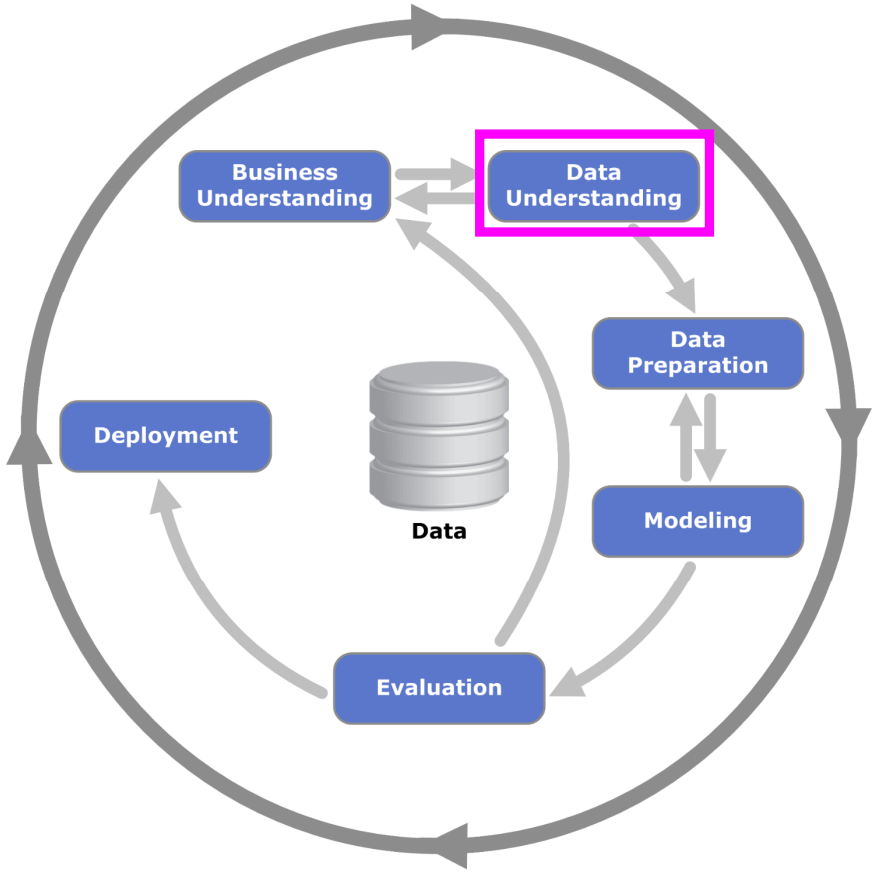



Image source: https://www.the-modeling-agency.com/crisp-dm.pdf

# Data Preparation

- The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data
- Data preparation tasks are likely to be performed multiple times, and not in any prescribed order
- Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools

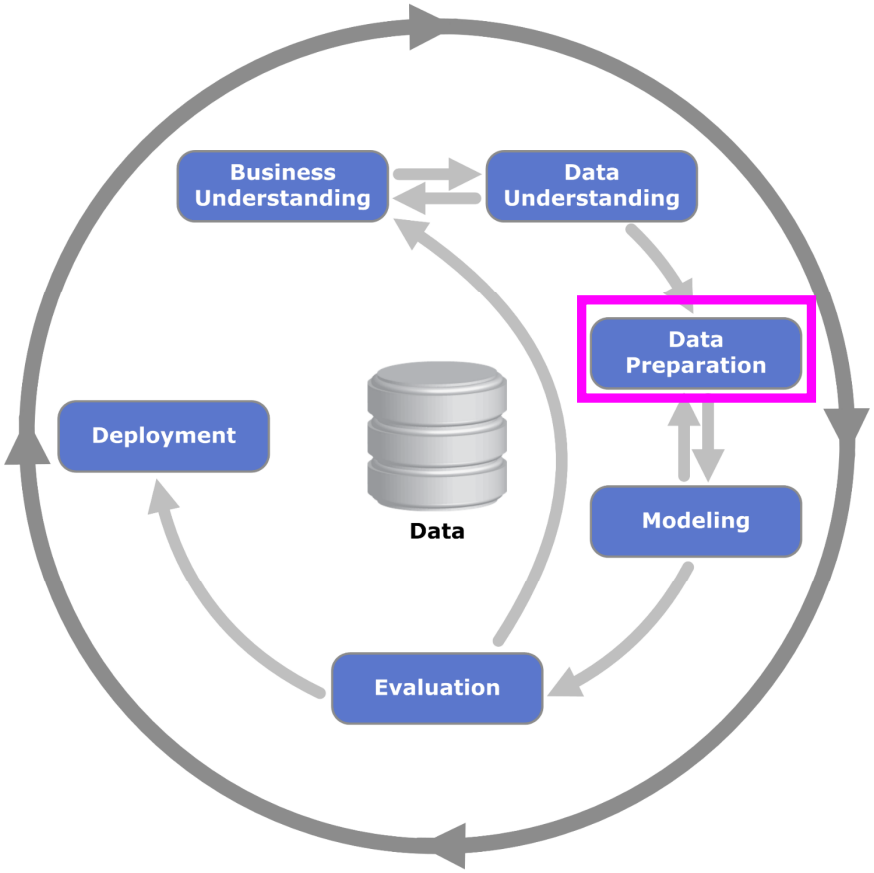
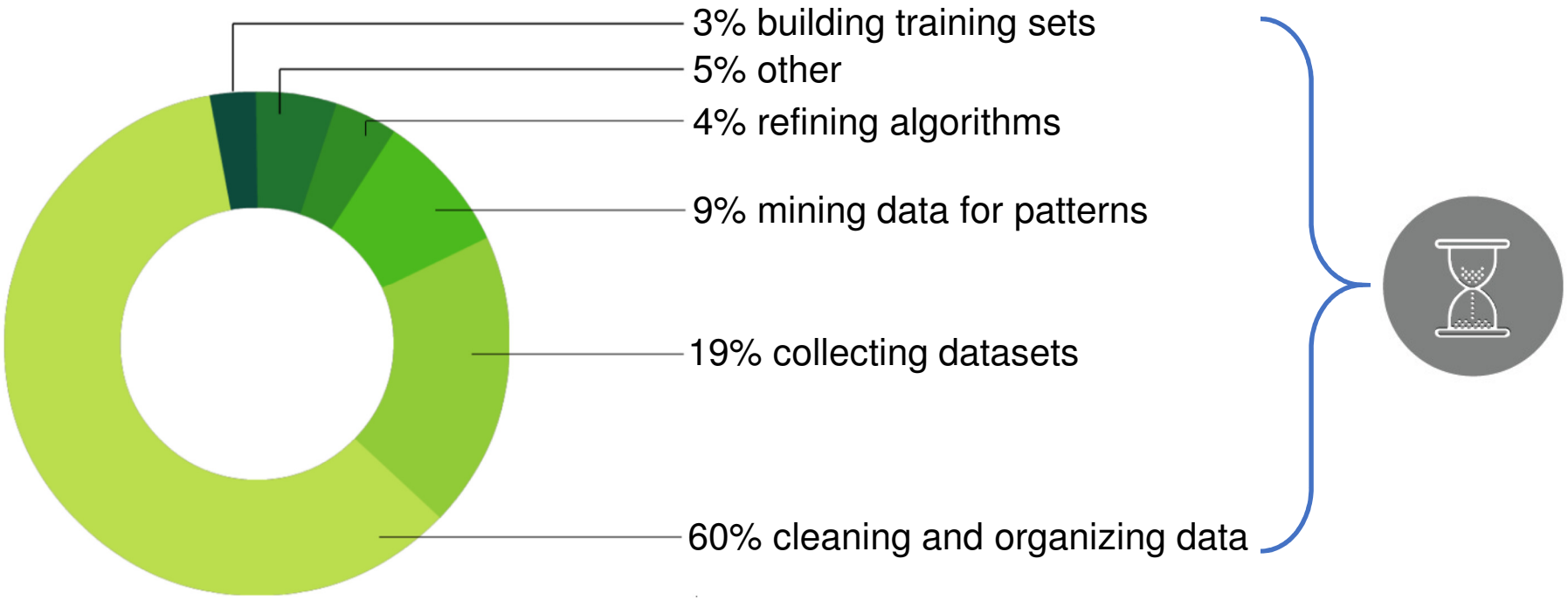



Image source: https://www.the-modeling-agency.com/crisp-dm.pdf

# Preparing data could be time-consuming

- What data scientists spend the most time doing




New York Times article reported that data scientists spend from ***50%*** to ***80%*** of their time mired in the more mundane task of collecting and preparing unruly data before it can be explored for useful nuggets.

**Messy data is by far the most time-consuming aspect of the typical data scientist's workflow**

# A Roadmap of Available Techniques

- Data Understanding
  - How do I find the data I need for modeling? ← Data Acquisition
  - How do I put all pieces of data together? ← Data Integration
  - What do the data look like? ← Data Description
  - How clean is the data set? ← Data Assessment
- Data Preparation
  - How do I clean up data? ← Data Cleaning
    - How do I handle missing values?
    - How do I handle noises?
    - How do I handle outliers/anomalies?
  - How do I express data variables? ← Data Transformation
  - Are all cases treated the same? ← Data Weighting and Balancing
  - Can I reduce the amount of data to use? ← Data Reduction
  - Can I create some new variables? ← Data Derivation

# Data Acquisition – how do I find the data I need for modeling?

- Goal: identify where the datasets are and how you can get them
- If datasets exist in different places
  - Departments, spreadsheets, databases, printed documents, handwritten notes, etc.
- If all in one place, e.g., in a data warehouse
  - Determine the best way to access the data
- If in multiple database structures
  - Query-based data extracts from database to flat files
  - High-level query language for direct access to the database
  - Low-level connections for direct access to the database

# Data Acquisition – how do I find the data I need for modeling?

- Goal: identify where the datasets are and how you can get them
- Query-based data extract
  - SQL: Structured Query Language
  - Capability: filtering input and output fields (columns) and records (rows) based on specified levels in a number of variables, aggregations (group by), sorting (order by), and subselects (select within other select, nested)
  - Statements: insert, create, update, etc.
  - Advantage: access the data the fastest (in RAM)
  - Disadvantage: duplicate and save to a file, might be impossible to fill with very large data

# Data Acquisition – how do I find the data I need for modeling?

- Goal: identify where the datasets are and how you can get them
- High-level query language
  - Optimized SQL for data mining, model-oriented
  - MQL: Modeling Query Language (Imielinski and Virmani, 1999)
  - DMQL: Data Mining Query Language (Han et al., 1996)
  - Attractive but not in standard use
  - Some data mining tools may support this approach, e.g., XML

# Data Acquisition – how do I find the data I need for modeling?

- Goal: identify where the datasets are and how you can get them
- Some data mining tools provide in-database access to data
  - Via ODBC or other proprietary low-level interfaces
  - SAS-Enterprise Miner, SPSS Clementine, and STATISTICA
- Benefits:
  - No need to move large volumes of data
  - Centralized data management and provisioning
  - Reduced unnecessary proliferation
  - Better data governance to satisfy compliance concerns

# Data Integration – how do I put all pieces of data together?

- Create a combined data structure suitable for input
  - Combine data in different fields in several different data extract files to form one field in the output
  - Combine data from several different fields into one field
- Data mining tools: most data mining tools provide some integration capabilities, e.g., merging, lookups, etc.

| Name | Address | City | State | Zipcode |
|---|---|---|---|---|
| John Brown | 1234 E St. | Chicago | IL | 60610 |
| Jean Blois | 300 Day St. | Houston | TX | 77091 |
| Neal Smith | 325 Clay St. | Portland | OR | 97201 |

+

| Name | Address | Product | Sales Date |
|---|---|---|---|
| John Brown | 1234 E. St. | Mower | 1/3/2007 |
| John Brown | 1234 E. St. | Rake | 4/16/2006 |
| Neal Smith | 325 Clay St. | Shovel | 8/23/2005 |
| Jean Blois | 300 Day St. | Hoe | 9/28/2007 |

=

| Name | Address | City | State | Zipcode | Product1 | Product2 |
|---|---|---|---|---|---|---|
| John Brown | 1234 E. St. | Chicago | IL | 60610 | Mower | Rake |
| Neal Smith | 325 Clay St. | Portland | OR | 97201 | Shovel | |
| Jean Blois | 300 Day St. | Houston | TX | 77091 | Hoe | |

# Data Integration – how do I put all pieces of data together?

- Combine data from multiple sources into a coherent data set
- Entity identification problem
  - How can equivalent real-world entities from multiple sources be matched up?
  - Are `customer_id` and `cust_number` the same attribute?
- Redundancy and correlation analysis
  - An attribute (e.g., yearly income) may be redundant if it can be derived from another attribute of set of attributes
  - Some redundancies can be detected by correlation analysis
- Tuple duplication
  - Duplication should be detected at the tuple level, e.g., there might be two or more identical tuples/records for a given unique data entry case
  - Inconsistencies often arise between various duplicates, due to inaccurate data entry or updating some but not all data occurrences, e.g., if a purchase order database contains attributes for the purchaser's name and address instead of a key, discrepancies can occur such as purchaser's name appearing with different addresses
- Data value conflict detection and resolution
  - For example, a weight attribute may be stored in metric units in one system and British units in another, e.g., kilograms vs. pounds
  - For example, course grades, A to F vs. 1 to 10 vs. 0 to 100
  - Resolving these issues is case-specific

# Data Description – what do the data look like?
## Initial Data Analysis (IDA)

- IDA is an essential part of nearly every analysis
  - The structure of the data
  - The quality of the data: errors, outliers, and missing observations
  - Descriptive statistics
    - Mean, median, mode, variance, standard deviation, minimum, maximum, frequency table, histograms, etc.
  - Graph displays
- The data are modified according to the analysis
  - Adjust extreme observations, estimate missing values, transform variables, bin data, form new variables, etc.
- For simple descriptive metrics
  - Microsoft Excel Analysis Tool Pack add-on might be sufficient
  - STATISTICA Data Miner and SPSS provide more robust descriptive tools

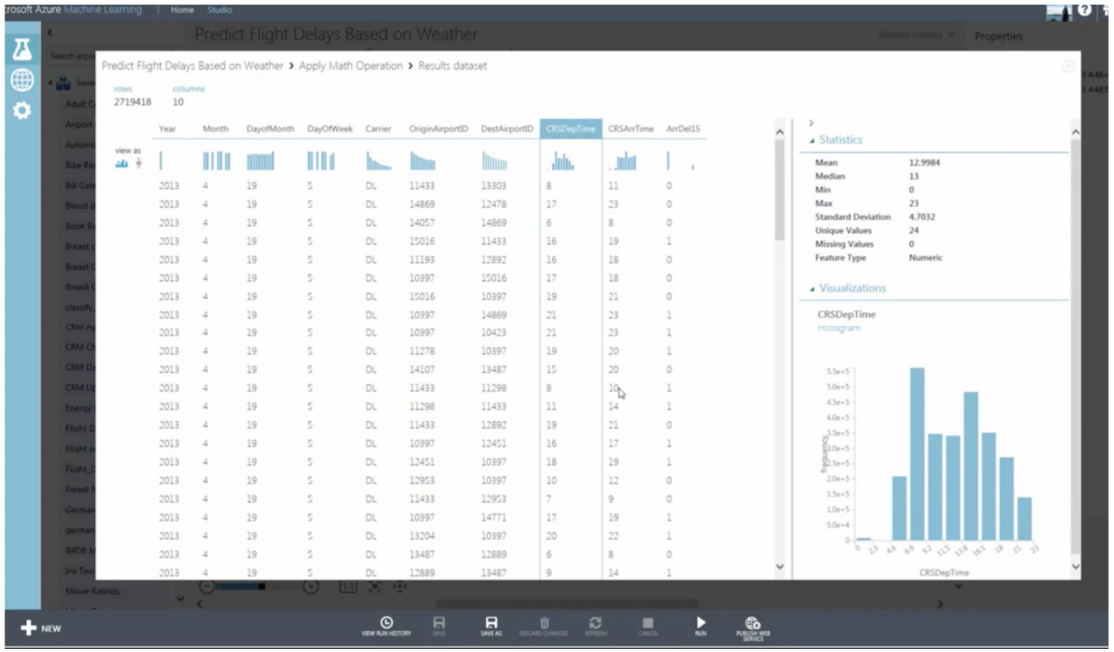
# **Data Description – what do the data look like?**
## Exploration Data Analysis (EDA)

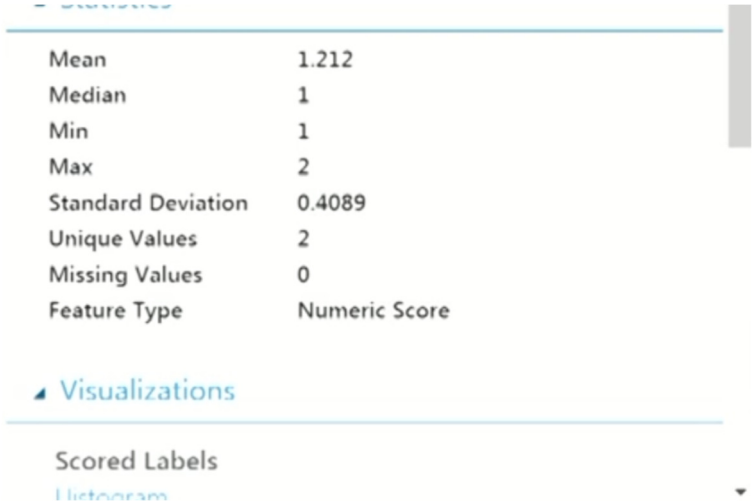
- EDA is an approach to analyze data for purpose of formulating hypotheses that are worth testing
  - Visualization techniques are often used
  - Encourages explorations and possible hypotheses could lead to new data collection and experiments

- It is important to understand what you ***CAN*** do before you can learn to measure how ***WELL*** you seem to have done it
  - " To learn about data analysis, it is right that each of us try many things that do not work – that we tackle more problems than we make expert analyses of. We often learn less from an expertly done analysis than from one where, by not trying something, we missed an opportunity to learn more."

# Data Description – what do the data look like?

- Examine the “gross” or “surface” properties of the acquired data and report on the results
- Statistical Summary Chart shows the distribution of each variable and provides descriptive statistics



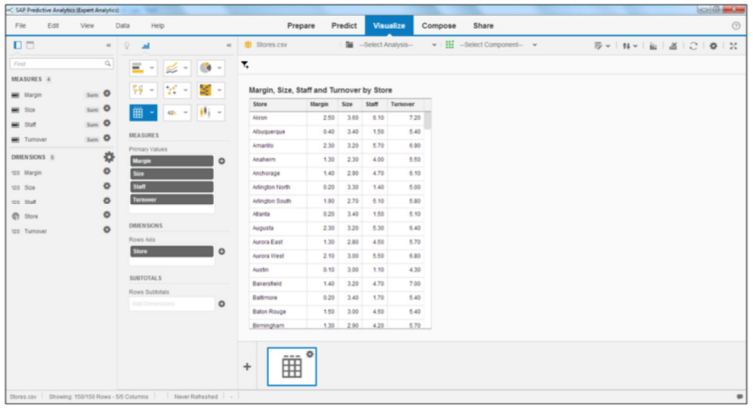



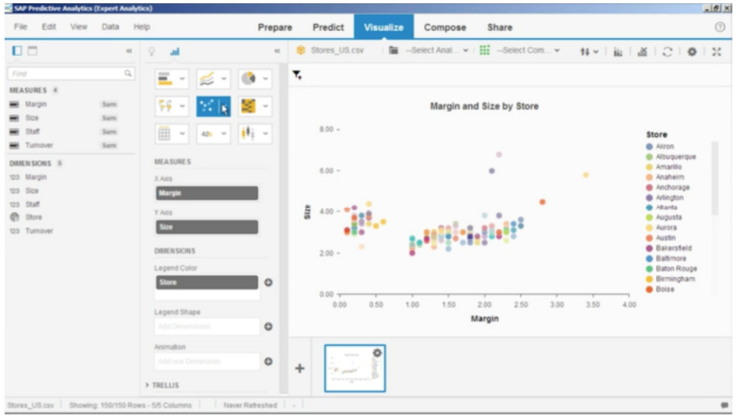

**Data description report**: describe the data that has been acquired including its format, its quality (e.g., the number of records and fields in each table), the identities of the fields and any other surface features which have been discovered. Evaluate whether the data acquired satisfies your requirements

Image source: Microsoft Azure Machine Learning

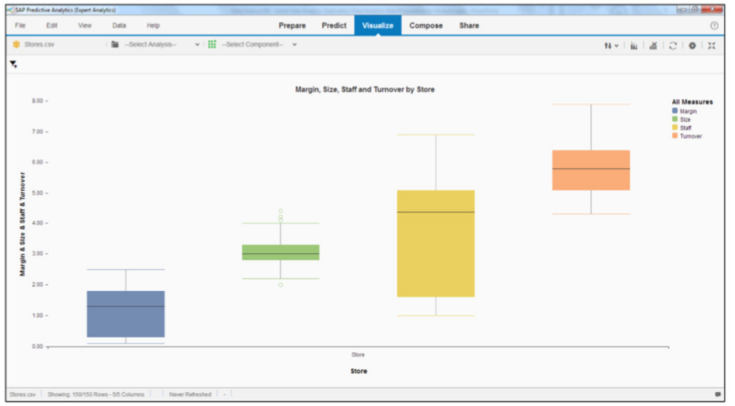
# Data Description – what do the data look like?

- Data exploration and visualization








- Box-plots can help identify outliers
- Density plots and histogram show the spread of the data
- Scatter plots can describe bivariate relationships

# Data Description – what do the data look like?

## Let us take a closer look at the data

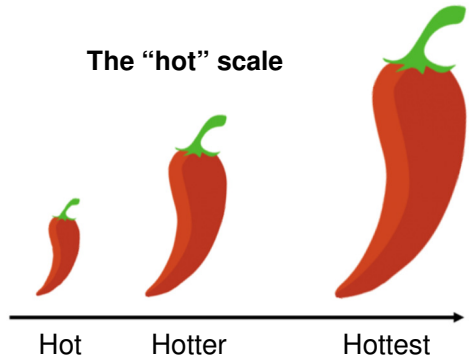
**Categorical variable / Nominal variable**

- It is a discrete (categorical), qualitative variable that characterizes, describes, or names an element of a population
- The order of the categories does not matter

| ID | Account Name | Age | Gender | Annual Income | Membership | Satisfaction Level |
|---|---|---|---|---|---|---|
| 1 | Jack Lennon | 43 | Male | 125,000 | Gold | Very dissatisfied |
| 2 | Steve Ive | 35 | Male | 100,000 | Gold | Dissatisfied |
| 3 | Sherry Jones | 38 | Female | 145,000 | Silver | Neutral |
| 4 | Peter Lorenz | 31 | Male | 96,000 | Gold | Satisfied |
| 5 | Bill McCartney | 25 | Male | 85,000 | Bronze | Very satisfied |
| 6 | John Carter | 18 | Male | 234,000 | Silver | Very dissatisfied |
| 7 | Kelly Mills | 19 | Female | 97,000 | Bronze | Very satisfied |
| 8 | Bono Sinead | 57 | Male | 135,000 | Gold | Neutral |
| 9 | James Scott | 44 | Male | 460,000 | Silver | Very satisfied |
| 10 | Jens Schneider | 29 | Male | 150,000 | Gold | Satisfied |

**Ordinal variables**

- An ordinal variable is discrete (categorical), qualitative variable that has order, e.g., gold, silver, bronze
- The order of the categories does matter

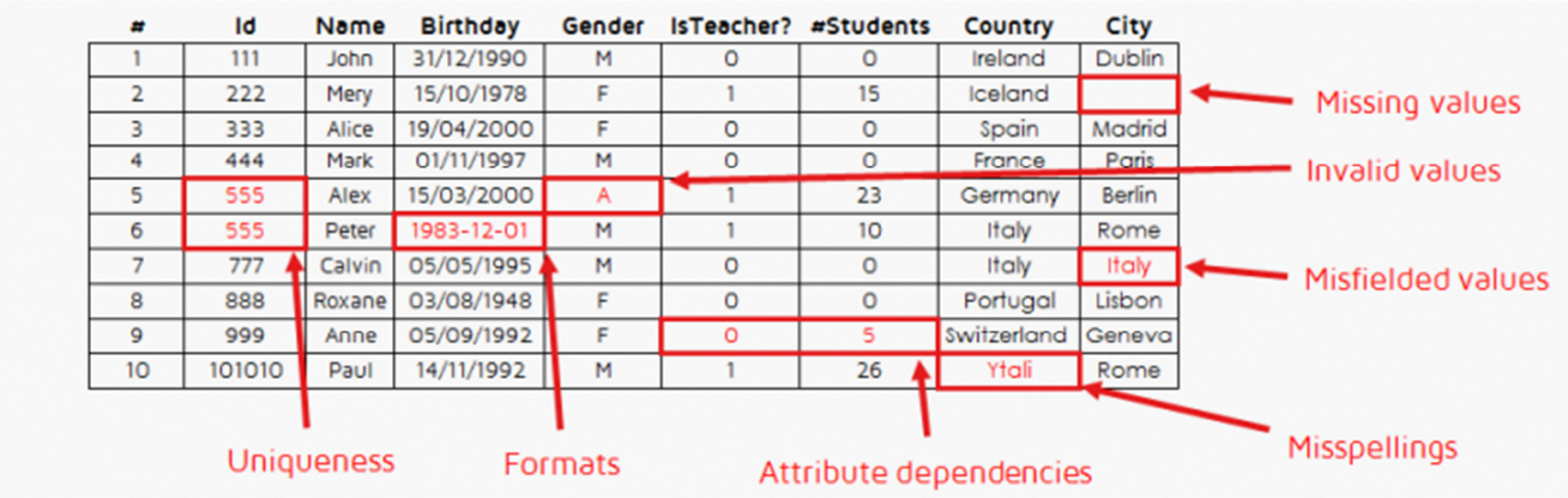



**Continuous variables**

- It is a quantitative variable
- It is a real number that can take any value (with fractions/decimal places) between two specific numbers
- It accommodates all basic arithmetic operations (addition, subtraction, multiplication, and division)

# Data Assessment – how clean is the data set?

- Data is rarely clean and oftentimes you have data quality issues




| # | Id | Name | Birthday | Gender | IsTeacher? | #Students | Country | City |
|---|---|---|---|---|---|---|---|---|
| 1 | 111 | John | 31/12/1990 | M | 0 | 0 | Ireland | Dublin |
| 2 | 222 | Mery | 15/10/1978 | F | 1 | 15 | Iceland | |
| 3 | 333 | Alice | 19/04/2000 | F | 0 | 0 | Spain | Madrid |
| 4 | 444 | Mark | 01/11/1997 | M | 0 | 0 | France | Paris |
| 5 | 555 | Alex | 15/03/2000 | A | 1 | 23 | Germany | Berlin |
| 6 | 555 | Peter | 1983-12-01 | M | 1 | 10 | Italy | Rome |
| 7 | 777 | Calvin | 05/05/1995 | M | 0 | 0 | Italy | Italy |
| 8 | 888 | Roxane | 03/08/1948 | F | 0 | 0 | Portugal | Lisbon |
| 9 | 999 | Anne | 05/09/1992 | F | 0 | 5 | Switzerland | Geneva |
| 10 | 101010 | Paul | 14/11/1992 | M | 1 | 26 | Ytali | Rome |

The typical data quality issues that arise are:

- **Incomplete**: data lacks attributes or containing missing values
- **Noisy**: data contains erroneous records or outliers
- **Inconsistent**: data contain conflicting records or discrepancies

https://quantdare.com/data-cleansing-and-transformation/

# Data Assessment – how clean is the data set?

## What kind of issues affect the quality of data?

- Invalid values
  - Some datasets have well-known values, e.g., gender
- Formats
  - The most common issue. It is easy to get values in different formats like a name written as “Name, Surname” or “Surname, Name”
- Attribute dependencies
  - When the value of a feature depends on the value of another feature, e.g., if we have some school data, the “number of students” is related to whether the person “is teacher?” If someone is not a teacher he/she cannot have any students
- Uniqueness
  - It is possible to find repeated data in features that only allow unique values, e.g., we cannot have two products with the same identifier
- Missing values
  - Some features in the dataset may have blank or null values
- Misspellings
  - Incorrectly written values
- Misfielded values
  - When a feature contains the values of another


https://quantdare.com/data-cleansing-and-transformation/

# Data Assessment – how clean is the data set?

- Goal: locate the problems in the data and decide how to handle them
- Some problems will become evident during data description operations
- Data profiling and analysis of the impact of poor-quality data
- Data distribution of each variable
  - The central tendency of data in the variable
  - Any potential outlier
  - The number of and distribution of blanks across all the cases
  - Any suspicious data, like miscodes, training data, system test data, or just plain garbage
- *Your findings could and should be presented in the form of a report and listed as a milestone in the project plan*
- Data cleaning…

# Data Cleaning – how do I clean up data?

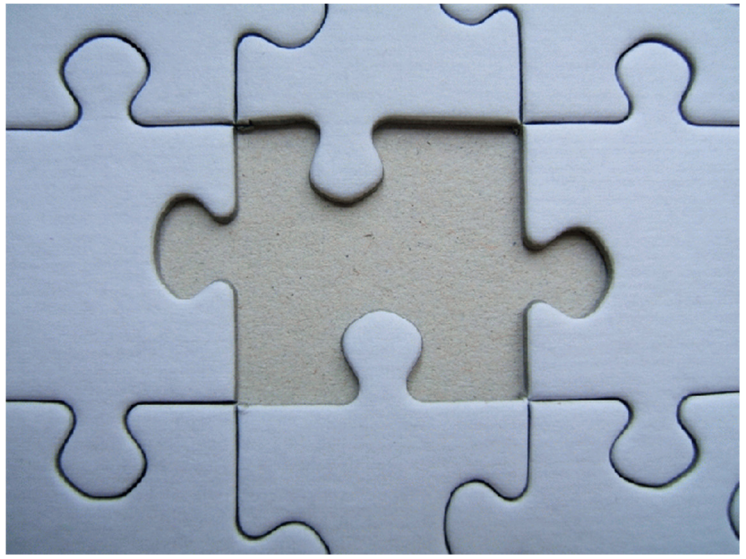
- Correct “bad” data, filter some bad data out of the data set, and filter out data that are too detailed for use in your model
- Validating codes against lists of acceptable values
  - Check the contents of each variable in all records to make sure that all of the contents are valid entries all over the data set
- Deleting particularly “dirty” records
  - Values like blanks, or noise are inappropriate for the data set, and should be removed from the records otherwise they will only confuse the model “signal” and decrease the predictive power of the model. Automatic techniques are very common way to implement this task

# Data Cleaning – how do I clean up data?

## How do I handle missing values?

### Why to deal with missing values?

- Missing values in a dataset can be caused by many reasons, e.g., errors or overlooked observations, device limitations, etc.
- When missing value are presented, certain algorithms may not work or you may not have the desired result
- Missing data affect some models more than others
- Even for models that handle missing data, they can be sensitive to it (missing data for certain variables can result in poor predictions)

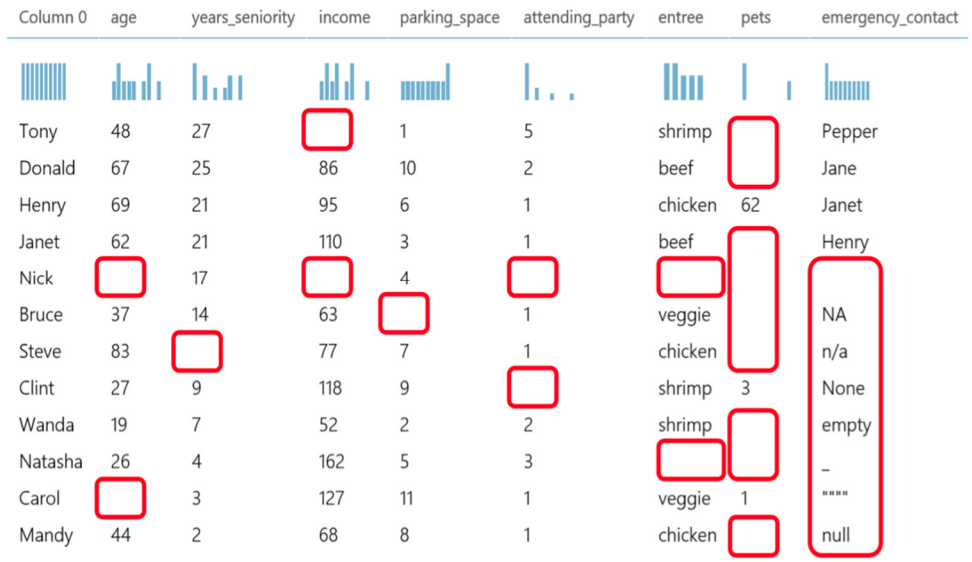


# Data Cleaning – how do I clean up data?

## How do I handle missing values?

## Data Imputations

- The operation of deciding what data to use to fill these blanks in records
- “Do the least harm”
  - By making the right assumption
- Missing Completely at Random (MCAR)
  - The probability of missing values in one variable is unrelated to the value of the variable itself, or to values of any other variable
- Missing at Random (MAR)
  - The probability of a value’s being missing in one variable is unrelated to the probability of missing data in another variable, but may be related to the value of the variable itself



| Column 0 | age | years_seniority | income | parking_space | attending_party | entree | pets | emergency_contact |
|---|---|---|---|---|---|---|---|---|
| Tony | 48 | 27 | | 1 | 5 | shrimp | | Pepper |
| Donald | 67 | 25 | 86 | 10 | 2 | beef | | Jane |
| Henry | 69 | 21 | 95 | 6 | 1 | chicken | 62 | Janet |
| Janet | 62 | 21 | 110 | 3 | 1 | beef | | Henry |
| Nick | | 17 | | 4 | | | | |
| Bruce | 37 | 14 | 63 | | 1 | veggie | | NA |
| Steve | 83 | | 77 | 7 | 1 | chicken | | n/a |
| Clint | 27 | 9 | 118 | 9 | | shrimp | 3 | None |
| Wanda | 19 | 7 | 52 | 2 | 2 | shrimp | | empty |
| Natasha | 26 | 4 | 162 | 5 | 3 | | | _ |
| Carol | | 3 | 127 | 11 | 1 | veggie | 1 | """" |
| Mandy | 44 | 2 | 68 | 8 | 1 | chicken | | null |

# Data Cleaning – how do I clean up data?

## How do I handle missing values?

**List-wise (or case-wise) deletion: the entire record (with missing values) is deleted (ignored)**

- Can be used for any kind of data
- No special statistical methods required, safest when MCAR
- Good for data with variables that are completely independent
  - The effect of each variable on the dependent variable is not affected by any other variables
- Applicable and appropriate to data sets suitable for regression
- Non-missing information in the record is lost, and the total information content of data will be reduced
- Can produce biased estimates if data are MAR
  - For example, if salary level depends positively on education level

# Data Cleaning – how do I clean up data?

## How do I handle missing values?

- **Pair-wise deletion**: all the cases with values for a variable will be used to calculate the covariance of that variable
  - Justified only if the data are MCAR; can lead to significant bias in the estimators if data are MAR
- **Mean substitution**: replace the missing values with mean (if numerical)
  - Or other central tendency, e.g., median, mode, which might be calculated based on all samples belonging to the same class as the given record
- **Reasonable value imputation**
  - Non-mean substitution, e.g.,, it is more reasonable to replace a missing value for number of children with zero rather than replace it with the mean or the median number of children based on all the other records (many couples are childless)
- **Dummy substitution**: replace missing values with a dummy value
  - Use a global constant to fill in missing value, unknown for categorical or 0 or infinity for numerical values; simple but sometimes misleading
- **Most "probable" value substitution**
  - **Frequent substitution**: replace missing values with the most frequent item (if categorical)
  - **Regression (interpolation) substitution**: use a regression method to replace missing values with regressed (interpolated) values

*Fill in the missing value manually could be time consuming and not always feasible, some data mining and statistical packages provide a facility for imputation with mean values*

# Data Cleaning – how do I clean up data?

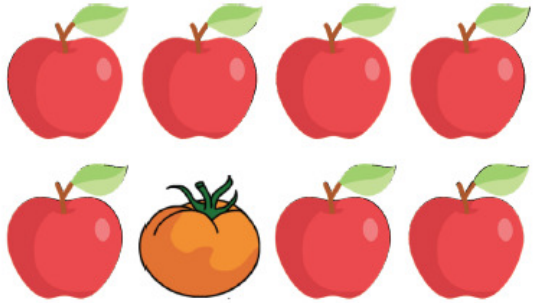
## How do I handle noises?

- Noisy data: anything that is not “true” data, could be caused by random error or variance in a measurement
- Goal: smooth the data to remove noise
- **Binning method**: consulting its “neighborhood”, i.e., the values around it
  - In smooth by bin means, each value in a bin is replaced by the mean value of the bin
  - In smooth by bin medians, each value in a bin is replaced by the bin median
  - In smooth by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries, and each bin value is then replaced by the closest boundary value

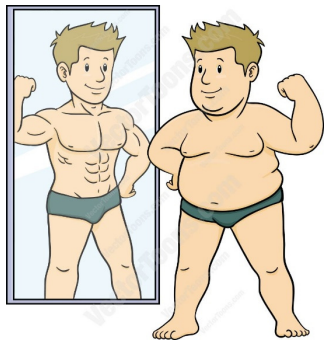
# Data Cleaning – how do I clean up data?

## How do I handle outliers/anomalies?

- An outlier is a data point that distinctly different from the rest
  - It may bring about problems by distorting the predictive model
  - What is an outlier is somewhat subjective
  - Can be very common in multidimensional data
  - Some models are less sensitive (more robust) to outliers than others
  - Can be result of bad data collection, which can legitimate extreme (or unusual) values
  - Sometimes outliers could be of our interests, and other times they just get in the way
- Causes of outliers



**Data from different classes**



**Data measurement and collection Errors**

**Natural variation**

# Data Cleaning – how do I clean up data?

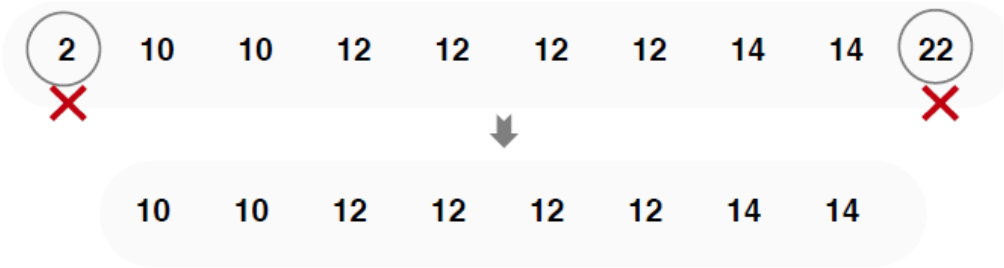
## How do I handle outliers/anomalies?

- How to deal with and outlier should depend on the cause
- Keep outliers
  - In many applications, outliers provide crucial information, e.g., in credit card fraud detection, outliers may indicate purchases that fall outside of a customer's usual buying patterns
- Exclude outliers
  - Trimming/Truncation: trimming discards the outliers
  - Winsorizing: replaces the outliers with the nearest "non-suspect" data
  - Outlier analysis: group similar values together in a cluster, values fall outside of the cluster are considered outliers

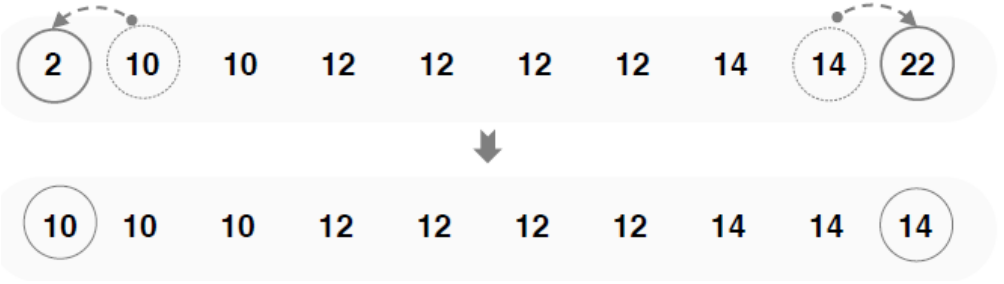
# Data Cleaning – how do I clean up data?

## How do I handle outliers/anomalies?

- **Example for Trimming**
  - Eliminate the outliers “2” and “22”




- **Example for Winsorizing**
  - Assign outlier the next highest or lowest value found in the sample that is not an outlier. In this example, “10” and “14” are not outliers and used to replace the outliers “2” and “22”




**Trimming or Winsorizing less than 5% of data points**

- It will not likely affect the hypothesis testing outcome

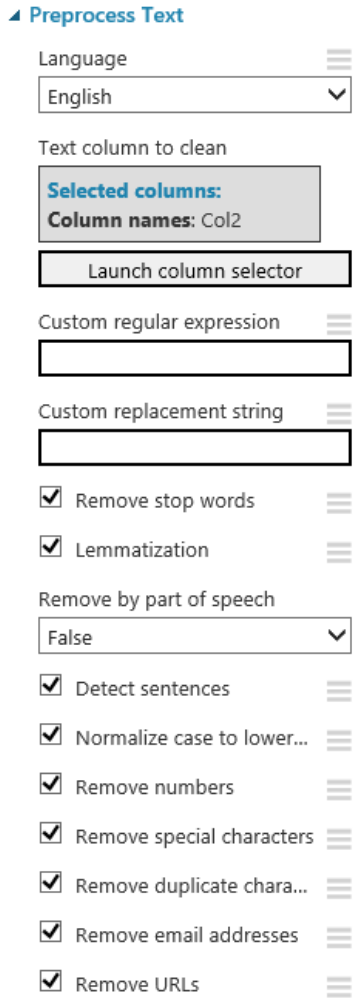
**Trimming or Winsorizing greater 5% of data points**

- It may affect the outcome results
  - Reduce the power of analysis
  - Makes sample less representative
  - May affect normalcy of data
  - Consider transforming data, choosing an alternate outcome variable or data analysis technique

https://www.youtube.com/watch?v=ls-JHKUDL2k

# Data Cleaning – how do I clean up data?
## Text Cleaning

- Improper text encoding handling while writing/reading text leads to information loss
- Inadvertent introduction of unreadable characters, e.g., null, may also affect text parsing
- Unstructured text such as tweets, product reviews, or search queries usually requires some preprocessing before it can be analyzed
- For example
  - Replacing special characters and punctuation marks with spaces
  - Normalizing case
  - Removing duplicate characters
  - Removing user-defined or built-in stop-words
  - Word stemming




**Example for Cleaning and preprocessing text dataset in Azure Machine Learning Studio**

- Remove stop words - common words such as "the" or "a" - and numbers, special characters, duplicated characters, email addresses, and URLs.
- Convert the text to lowercase, lemmatize the words, and detect sentence boundaries that are then indicated by "|||" symbol in pre-processed text

https://docs.microsoft.com/en-us/azure/machine-learning/studio/text-analytics-module-tutorial

# Data Transformation – how do I express data variables?

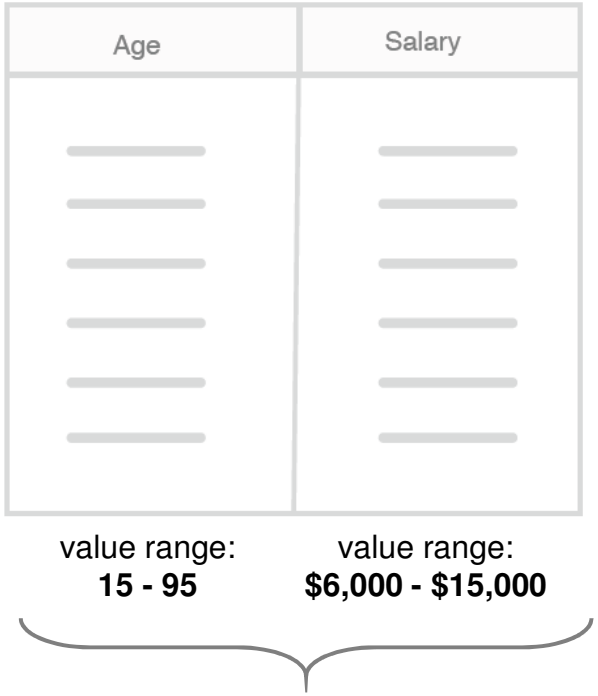
- **Data normalization** rescales numerical values to a specified range
- Numerical variable
  - Min-Max normalization: linearly transform the data to a range, say between 0 and 1, where the min value is scaled to 0 and max value to 1

$$x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$$

  - Z-score normalization (or standardization): scale data based on mean and standard deviation: divide the difference between the data and the mean by the standard deviation.

$$z = \frac{x - \mu}{\sigma}, \text{ i.e., z-score} = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

  - Decimal scaling: scale the data by moving the decimal point of the attribute value



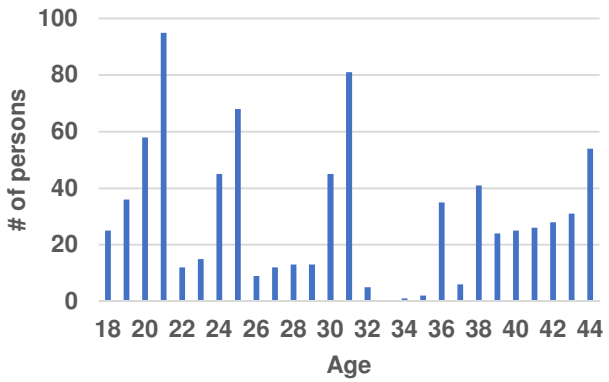

Different scales in a dataset may be problematic in some cases where certain machine algorithms require data to be in the same scale

https://en.wikipedia.org/wiki/Feature_scaling

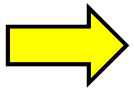
https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/prepare-data

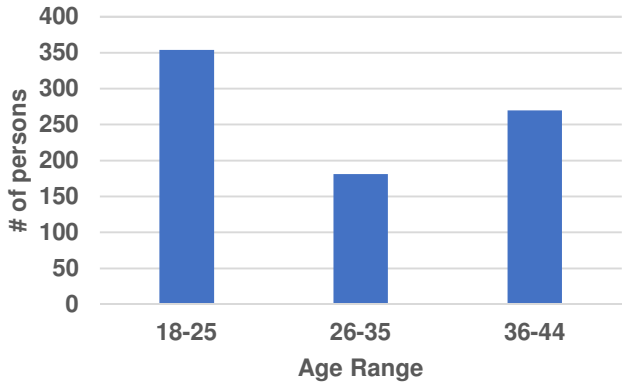
# Data Transformation – how do I express data variables?

- **Data discretization** converts continuous attributes by “binning” to categorical attributes for ease of use with certain learning methods
  - A numeric variable may have many different values and for some algorithms this may lead to very complex models
  - For example, the raw values of a numeric attribute (e.g., age) are replaced by interval labels (e.g., 0 to 10, 11 to 20, etc.) or conceptual labels (e.g., youth, adult, senior)




**Raw data – No binning**






**Manual binning**

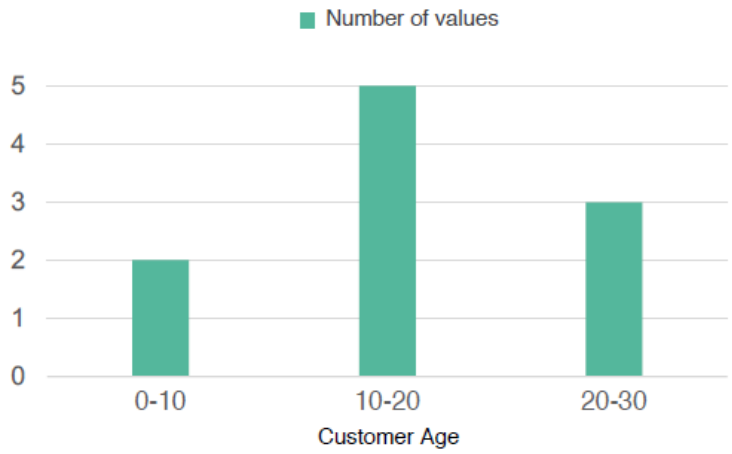
- Discretization is the process of putting values into buckets so that there are a limited number of possible states. The buckets are treated as ordered and discrete values. You can discretize both numeric and string columns
  - Think about your letter-based final grades of CS644

- Binning helps to improve model performance. It captures non-linear behavior of continuous variables
- It minimizes the impact of outliers. It removes “noise” from large numbers of distinct values
- it makes the models more explainable – grouped values are easier to display and understand. It improves model build speed – predictive algorithms build much faster as the number of distinct values decreases

# Data Transformation – how do I express data variables?

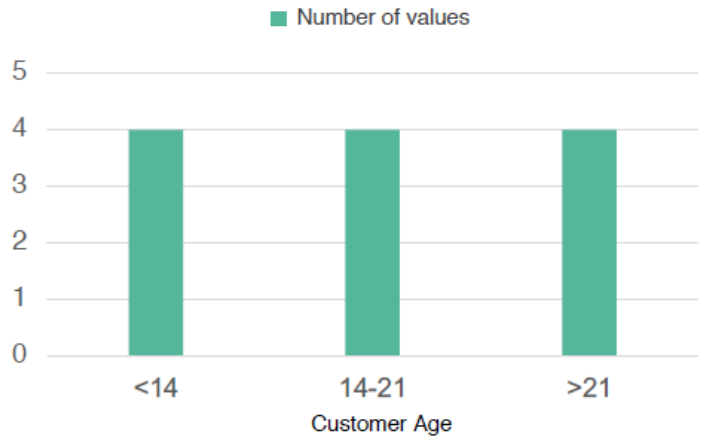
Data can be **discretized** by converting continuous values to categorical attributes or intervals

- Equal-width binning (by distance)




  - Divide the range of all possible values of an attribute into N groups of the same size, and assign the values that fall in a bin with the bin number

- Equal-height binning (by frequency)




  - Divide the range of all possible values of an attribute into N groups, each containing the same number of instances, then assign the values that fall in a bin with the bin number

# Data Transformation – how do I express data variables?

- Transform data so that the resulting mining process may be more efficient, and the patterns found may be easier to understand
- Smoothing: removing noise from data, binning, regression, and clustering
- Attribute construction (feature construction): new attributes are constructed and added from the given set of attributes to help the mining process
- Aggregation: summary or aggregation operations are applied to the data, e.g., the daily sales data may be aggregated so as to computer monthly and yearly total amounts
- Concept hierarchy generation from nominal data: the attributes such as street can be generalized to high-level concepts, like city or country

# Data Transformation – how do I express data variables?

- Numerical variables: discretization, normalization, standardization
- Textual values may represent an underlying numerical progression
  - For example, Monday, Tuesday, Wednesday, Thursday, or Friday. But when you relate them to stress building up during the work week, you can treat them as a numeric variable
- Categorical variables, e.g., integers values vs. textual values

| Case | Color | Color-Red | Color-Blue | Color-Yellow | Color-Green |
|---|---|---|---|---|---|
| 1 | Red | 1 | 0 | 0 | 0 |
| 2 | Blue | 0 | 1 | 0 | 0 |
| 3 | Yellow | 0 | 0 | 1 | 0 |
| 4 | Green | 0 | 0 | 0 | 1 |
| 5 | Blue | 0 | 1 | 0 | 0 |

# Data Weighting and Balancing – are all cases treated the same?

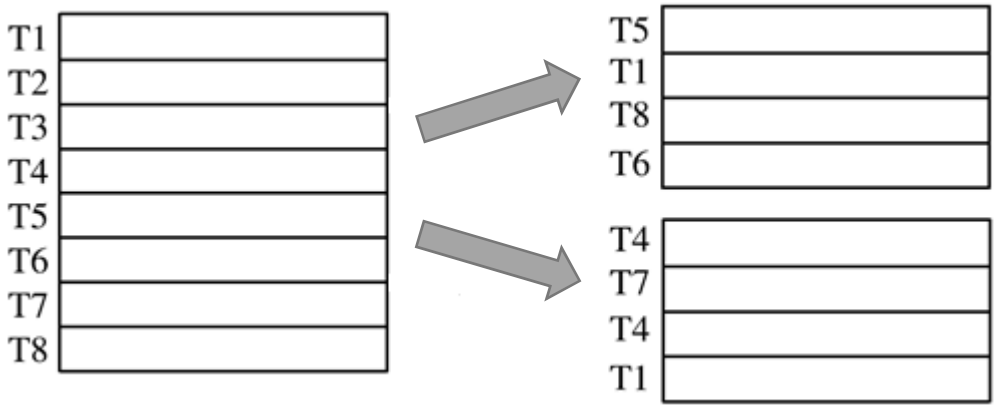
- Sometimes, you may want to weight each data point with data in another variable to calculate relationships consistent with reality
  - For example, a data value input from sensor-A may be twice as **accurate** as data from sensor-B. In this case, it would be wise to apply a weight of 2 for all values input by sensor-A and a weight of 1 for values input from sensor-B

- Some machine learning algorithms can learn case by case, and no metrics need to be calculated
  - For example, back-propagation ANNs assign random weights to each variable on the first pass through the data; in subsequent passes through the data (usually 100s or 1000s enough to reach the final stabled set of weights), the **weights are adjusted** according to the effects of variables in each case

# Data Reduction – can I reduce the amount of data to use?

**Obtain a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results**

- **Record sampling**
  Sample the data records and only choose the representative subset from the data
- **Attribute sampling**
  Select only a subset of the most important attributes from the data



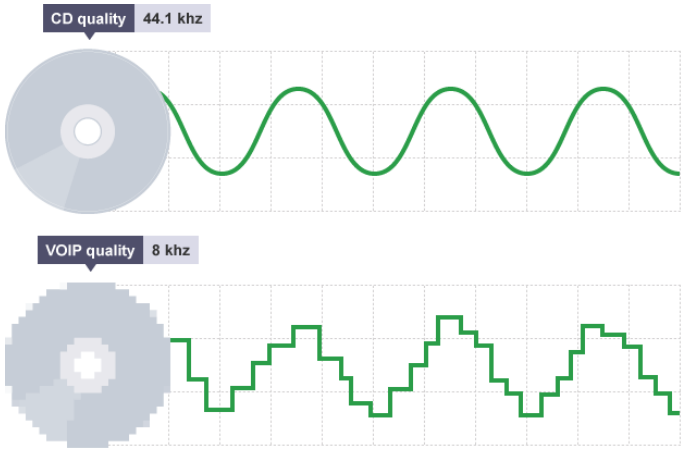

- If the dataset you plan to analyze is large, it's usually a good idea to down-sample the data to reduce it to a smaller but representative and more manageable size. This facilitates data understanding, exploration, and feature engineering

- More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset

# Data Reduction – can I reduce the amount of data to use?

## Data sampling

- It can reduce the number of data cases submitted to the modeling algorithm
- It can help you select only those cases in which the response patterns are relatively homogeneous
- It can help you balance the occurrence of rare events for analysis by machine learning tools
- Finally, simple random sampling can be used to divide the data set into **three data sets** for analysis

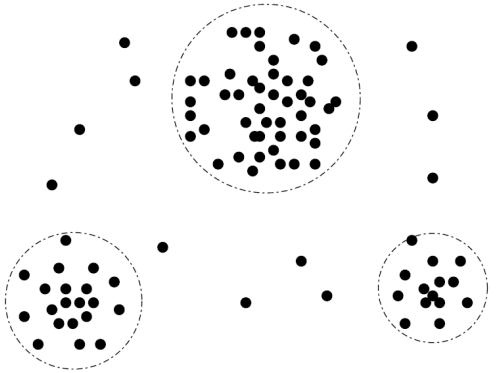



- **Training set**: these cases are randomly selected for use in training the model
- **Validation set**: these cases are used to assess the predictability of the model, before refining the model or adding model enhancements
- **Test set**: these cases are used to test the final performance of the model after all modeling is done

# Data Reduction – can I reduce the amount of data to use?

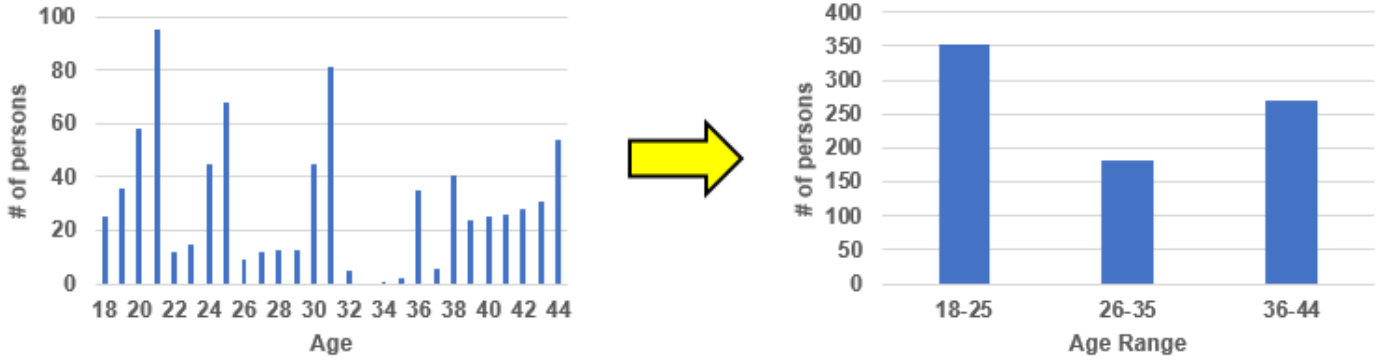
- Aggregation
  - Data records are gathered and expressed in a summary form
  - e.g., the daily revenue numbers of a restaurant chain over the past 20 years can be aggregated to monthly revenue to reduce the size of the data

- Clustering
  - Non-parametric data reduction
  - Group similar objects into the same group and dis-similar objects into different groups
  - Represented by its diameter, the maximum distance between any two objects in the same cluster; or centroid distance, which is the average distance of each cluster object from the cluster centroid

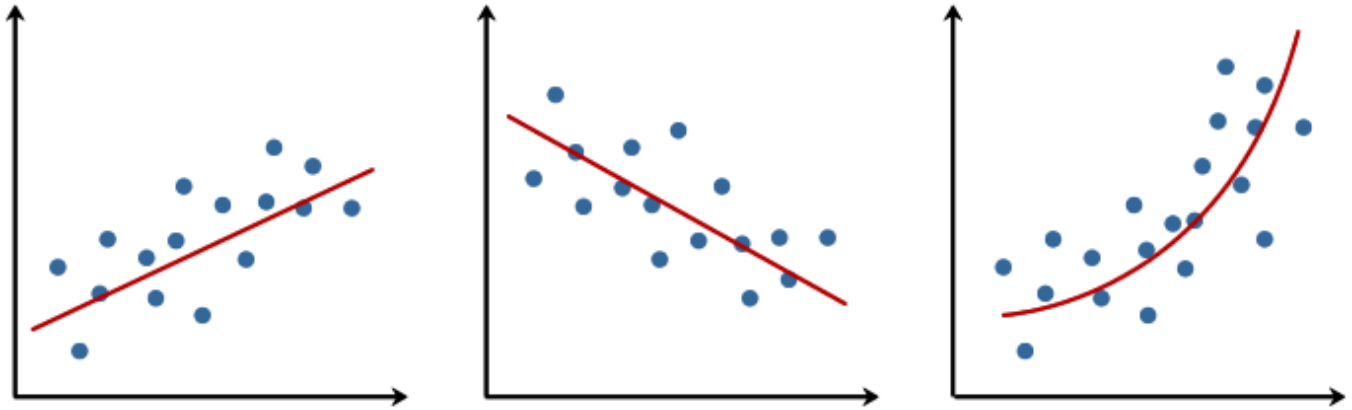


# Data Reduction – can I reduce the amount of data to use?

- Data discretization
  - A technique used to reduce the effects of minor observation errors




- Regression
  - Approximation of the given data
  - Given two series of data vectors X and Y, if we could find the approximate relationship between X and Y, say Y = wX + b, then we only need to handle the relationship between X and Y rather than every pair of the two data series ($x_i$, $y_i$)
  - Solvers: SAS, SPSS, S-Plus
  - Book, Numerical Recipes in C



# Data Reduction – can I reduce the amount of data to use?

- Dimensionality reduction: reduce the number of random variables
- Correlation coefficients
- Observations:
  - Most of the variables have relatively high and significant **correlations** with crime rate
  - For Charles River proximity is relatively low and insignificant

Correlations of Some Variables in the Boston Housing Data Set. Correlations in Red Are Significant at the 95% Confidence Level

| | Crime Rate | Nonretail Bus Acres | Charles River | Dist to Empl Centers | Property Tax Rate |
|---|---|---|---|---|---|
| Crime Rate | 1.000000 | 0.406583 | −0.055892 | −0.379670 | 0.582764 |
| Nonretail Bus Acres | 0.406583 | 1.000000 | 0.062938 | −0.708027 | 0.720760 |
| Charles River | −0.055892 | 0.062938 | 1.000000 | −0.099176 | −0.035587 |
| Dist to Empl Centers | −0.379670 | −0.708027 | −0.099176 | 1.000000 | −0.534432 |
| Property Tax Rate | 0.582764 | 0.720760 | −0.035587 | −0.534432 | 1.000000 |

Charles River may not be a good variable to include in the model

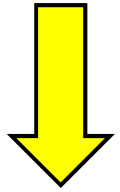
# Data Reduction – can I reduce the amount of data to use?

- Dimensionality reduction: reduce the number of random variables
- Correlation coefficients
- Observations:
  - None of the correlations of the predictor variables is greater than 0.90
  - If a correlation between two variables exceeded 0.90 (a common rule-of-thumb threshold)

Correlations of Some Variables in the Boston Housing Data Set. Correlations in Red Are Significant at the 95% Confidence Level

| | Crime Rate | Nonretail Bus Acres | Charles River | Dist to Empl Centers | Property Tax Rate |
|---|---|---|---|---|---|
| Crime Rate | 1.000000 | 0.406583 | −0.055892 | −0.379670 | 0.582764 |
| Nonretail Bus Acres | 0.406583 | 1.000000 | 0.062938 | −0.708027 | 0.720760 |
| Charles River | −0.055892 | 0.062938 | 1.000000 | −0.099176 | −0.035587 |
| Dist to Empl Centers | −0.379670 | −0.708027 | −0.099176 | 1.000000 | −0.534432 |
| Property Tax Rate | 0.582764 | 0.720760 | −0.035587 | −0.534432 | 1.000000 |

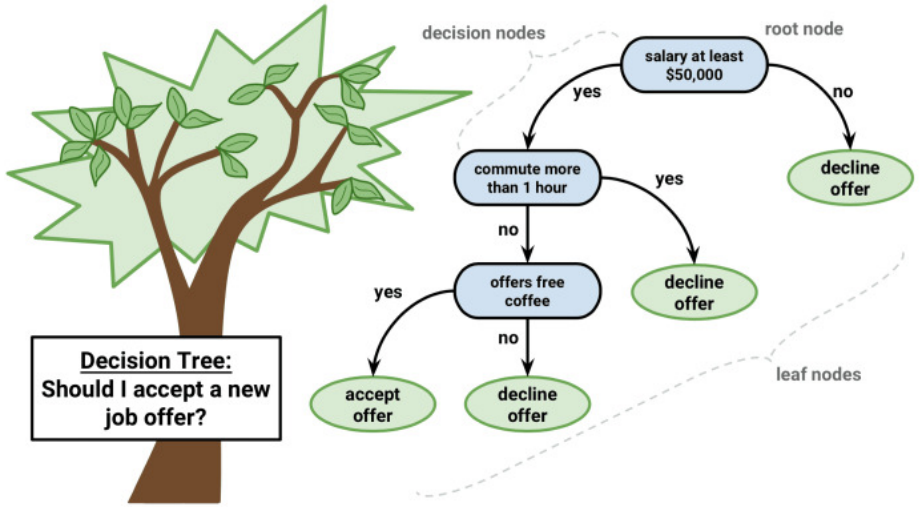
Too collinear to include in the model



Too much overlap in the effects of the collinear variables, making interpretation of the results problematic

# Data Reduction – can I reduce the amount of data to use?

- Dimensionality reduction: reduce the number of random variables
- Wavelet Transformation
  - The wavelet transformed data can be truncated
  - A compressed approximation of the data can be retained by storing only a small fraction of the strongest of the wavelet coefficients
- Principal Component Analysis
  - Convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables
- Decision Tree Induction
  - Gini index, Information Gain, Gain Ratio
- Data Compression
  - Transformation are applied so as to obtain a reduced or compressed representation of the original data, e.g., Huffman coding, LZW, etc.



# Data Deviation – can I create some new variables?

## Assignment or derivation of the target variable

- The operation involving defines the modeling goal in terms of available input variables
- The goal is to “hit” the target variable with the prediction of the model
  - The target variable can be selected from among the existing variables in the data set
  - e.g., the target variable for a model of equipment failure could be the presence or absence of a failure date in the data record

## Derivation of new predictor variables

- New variables can be created from a combination of existing variables, e.g., create distance attribute based on latitude and longitude

$$= \text{acos}(\sin(\text{Lat}\,1) * \sin(\text{Lat}\,2) + \cos(\text{Lat}\,1) * \cos(\text{Lat}\,2) * \cos(\text{Lon}\,2 - \text{Lon}\,1)) * 3934.31$$

## Attribute-oriented induction of generalization variables

- A generalization technique to create a higher-level (more general) expression variable from a list of detailed categories in a variable
  - e.g., form a concept generalization, “white collar worker”, based on specific levels in a number of other variables, e.g., “yearly salary”, “homeowner”, and “number of cars”

# Reading

- Read Chapter 2 and Chapter 3 of *Data Mining Concepts and Techniques* by J. Han et al.
- Read Chapter 2 and Chapter 3 of *Introduction to Data Mining* by P. Tan et al.
- Preprocessing Data in Azure Machine Learning Studio
  - https://azure.microsoft.com/en-us/resources/videos/preprocessing-data-in-azure-ml-studio/

# Processing Issues

- Where is processing hosted?
  - Distributed servers / cloud (e.g., Amazon EC2)
- Where is data stored?
  - Distributed storage (e.g., DFS, Amazon S3)
- What is the programming model?
  - Distributed processing (e.g., MapReduce)
- How data is stored and indexed?
  - High-performance schema-free databases (e.g., MongoDB)
- What operations are performed on data?
  - Analytics / sematic processing (e.g., R)

# A List of Major Issues in Data Mining

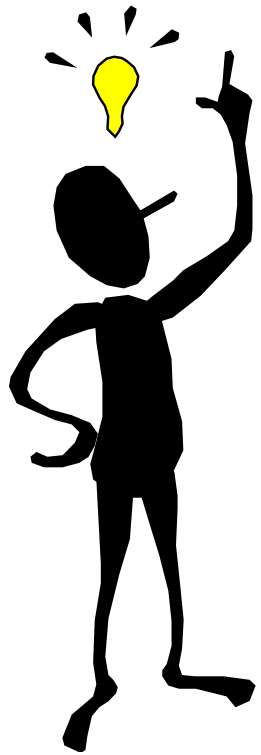
- Mining of different kinds of information in databases
  - It is necessary to integrate data from diverse sources, including data warehouses, data marts, excel spreadsheets, text documents, and images
- Interactive mining of knowledge at multiple levels of abstraction
  - Account-level data must be combined with individual-level data and coordinated with data with different time-grains (daily, monthly, etc.)
- Incorporation of background information
  - Some of the most powerful predictor variables are those gathered from outside the corporate database
- Data mining query languages and ad hoc data mining
  - Data miners must interface closely with database management systems to access data. Structured Query Language (SQL) is the most common query tool used to extract data from large databases
- Presentation and visualization of data mining results
  - Presenting highly technical results to nontechnical managers can be very challenging
- Handling “noisy” or incomplete data
  - Many items of data (fields) for a given customer or account (a record) are often blank. One of the most challenging tasks in data mining is filling those blanks with intuitive values

# A List of Major Issues in Data Mining

- Pattern evaluation - the “interestingness” problem
  - Many patterns may exist in a data set. The challenge for data mining is to distinguish those patterns that are “interesting” and useful to solve the data mining problem at hand
- Efficiency and scalability of data mining algorithms
  - Efficiency of a data mining algorithm can be measured in terms of its predictive power and the time it takes to produce a model
- Parallel, distributed, and incremental mining algorithms
  - Large data mining problems can be processed much more efficiently using a divide and conquer approach with multiple processors in parallel computers
- Handling of relational and complex types of data
  - Much input data might come from relational databases (a system of normalized tables linked together by common keys). Other input data might come from complex multidimensional databases
- Mining info from heterogeneous/global information systems
  - Data mining tools must have the ability to process data input from very different database structures

# Next

- Big Data Analytics Platforms



# Thanks ! ☺

# Questions ?