Data Analytics Lifecycle

- Data science projects differ from BI projects
  - More exploratory in nature
  - Critical to have a project process
  - Participants should be thorough and rigorous
- Break large projects into smaller pieces
- Data Analytics Lifecycle **defines the analytics process and best practices from discovery to project completion**.

Data Analytics Lifecycle

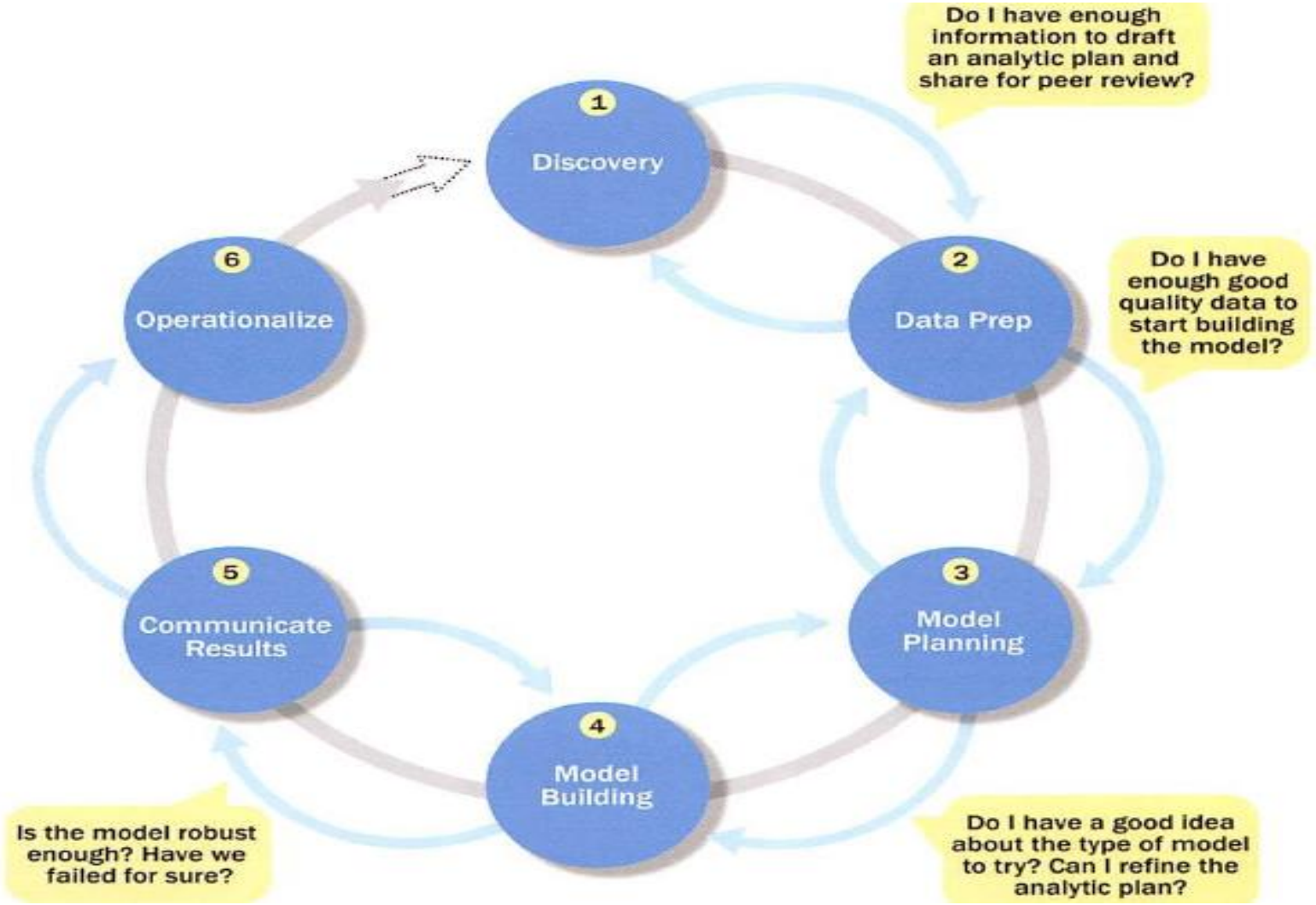
- **Data Analytics lifecycle Phases:**

1. Discovery Phase
2. Data Preparation Phase
3. Model Planning Phase
4. Model Building Phase
5. Communicate Result
6. Operationalize

Data Analytics Lifecycle

- With six phases the project work can occur in several phases simultaneously
- The cycle is iterative to portray a real project
- Work can return to earlier phases as new information is uncovered

Data Analytics Lifecycle



Data Analytics Lifecycle

- Phase – I: Discovery
  - Learning the Business Domain
  - Resources
  - Framing the Problem
  - Developing Initial Hypotheses
  - Identifying Potential Data Sources

Data Analytics Lifecycle

- Phase – 2: Data Preparation
  - It requires analytical sandbox in which you can perform analytics for the entire duration of the project
  - Includes steps to
    - Explore
    - Preprocess
    - Conditional Data
  - Data preparation tends to be the most labor-intensive step in the analytics lifecycle
    - Often at least 50% of the data science project's time
  - The data preparation phase is a iterative process

Data Analytics Lifecycle

- Phase – 2: Data Preparation
  - In ETL users perform extract, transform, load
  - Data Analytics lifecycle → ELT or ETLT → Extract, Transform, Load and Transform.
    - early load preserves the raw data which can be useful to examine.
  - Example :
    - In credit card fraud detection, outliers can represent high-risk transactions that might be inadvertently filtered out or transformed before being loaded into the database

Data Analytics Lifecycle

- Phase – 2: Data Preparation
  - In ETL users perform extract, transform, load
  - Data Analytics lifecycle → ELT or ETLT → Extract, Transform, Load and Transform.
    - early load preserves the raw data which can be useful to examine.
  - Example :
    - In credit card fraud detection, outliers can represent high-risk transactions that might be inadvertently filtered out or transformed before being loaded into the database

Data Analytics Lifecycle

- Phase – 2: Data Preparation
  - In ETL users perform extract, transform, load
  - Data Analytics lifecycle → ELT or ETLT → Extract, Transform, Load and Transform.
    - early load preserves the raw data which can be useful to examine.
  - Example :
    - In credit card fraud detection, outliers can represent high-risk transactions that might be inadvertently filtered out or transformed before being loaded into the database

Data Analytics Lifecycle

- Phase – 3: Model Planning
  - This determines the methods and techniques to extract relationships among variables.
  - These relationship patterns will set the base for algorithms which will be used in next phase
  - It uses Exploratory Data Analysis (EDA) using various statistical formulae and visualization tools.
  - Simply, it identifies candidate models to apply to the data for clustering, classifying, or finding relationships in data.

Data Analytics Lifecycle

- Phase – 3: Model Planning
  - Activities to consider:
    - Assess the structure of the data
    - Ensure the analytic techniques enable the team to meet the business objectives and accept or reject the working hypotheses
    - Determine if the situation warrants a single model or a series of techniques as part of a larger analytic workflow
    - Research and understand how other analysts have approached this kind or similar kind of problem

Data Analytics Lifecycle

- Phase – 4: Model Building
  - Execute the models defined in Phase 3
  - Develop datasets for training, testing, and production
  - Develop analytic model on training data, test on test data.
  - It will consider whether your existing tools will suffice for running the models or it will need a more robust environment (like fast and parallel processing).
  - You will analyze various learning techniques like classification, association and clustering to build the model.

Data Analytics Lifecycle

- Phase – 5: Communicate Results
  - Determine if the team succeeded or failed in its objectives
  - Assess if the results are statistically significant and valid
    - If so, identify aspects of the results that present salient findings
    - Identify surprising results and those in line with the hypotheses
  - Communicate and document the key findings and major insights derived from the analysis
  - This is the most visible portion of the process to the outside stakeholders and sponsors

Data Analytics Lifecycle

- Phase – 6: Operationalize
  - In this last phase, the team communicates the benefits of the project more broadly and sets up a pilot project to deploy the work in a controlled way
  - Risk is managed effectively by undertaking small scope, pilot deployment before a wide-scale rollout
  - During the pilot project, the team may need to execute the algorithm more efficiently in the database

Data Preprocessing Techniques

- Need of Data Preprocessing:
  - Incomplete:
    - Missing attribute values,
    - Lack of certain attributes of interest
    - Consisting of aggregated values
  - Noisy Data
    - Contains errors
    - Outliers
  - Inconsistent
    - containing discrepancies in codes or names

Data Preprocessing Techniques

- Need of Data Preprocessing:
  - Quality data is required for better identification of Model
- Major tasks in Data Processing
  - Data Cleaning
  - Data Integration
  - Data Transformation
  - Data Reduction
  - Data Discretization

Data Preprocessing Techniques

- Data Cleaning:
  - Data cleaning is important, because data is collected from different data sources.
  - Data Cleaning Tasks are:
    - Fill the missing Values
    - Identify the outliers and smooth out noisy data
    - Correct inconsistent data
    - Resolve the redundancy caused by data integration

Data Preprocessing Techniques

- How to handle missing data?
  - Ignore the tuple (loss of information)
  - Fill in missing values manually (Complex task)
  - Use a measure of central tendency for the attribute to fill the missing value
    - Middle value of the data distribution
    - Mean, median etc.
  - Use the attribute mean or median for all samples belonging to the same class as the given tuple.
  - Use the most probable value to fill in the missing value: regression , inference-based tools using Bayesian formula, decision tree, EM algorithm

Data Preprocessing Techniques

- Noisy Data: Random error or variance in a measured variable
- How to handle Noisy Data?
  - Binning Method
    - Sort the data and partition into (equal-size) bins
    - Smooth by bin means, Smooth by bin Medians, Smooth by bin Boundaries, etc
  - Clustering method, Curve-fitting, Regression
    - Detects and removes the outliers
  - Combined computer and human inspection
    - Detect the suspicious values and check by human

Data Preprocessing Techniques

- Noisy Data: Random error or variance in a measured variable
- Simple Discretization Methods: Binning
  - Equal-width(distance) partitioning:
    - It divides the range into N intervals of equal size: uniform grid
    - if A and Bare the lowest and highest values of the attribute, the width of intervals will be: W = (B-A)/N.
    - The most straightforward
    - But outliers may dominate presentation
    - Skewed data is not handled well

Data Preprocessing Techniques

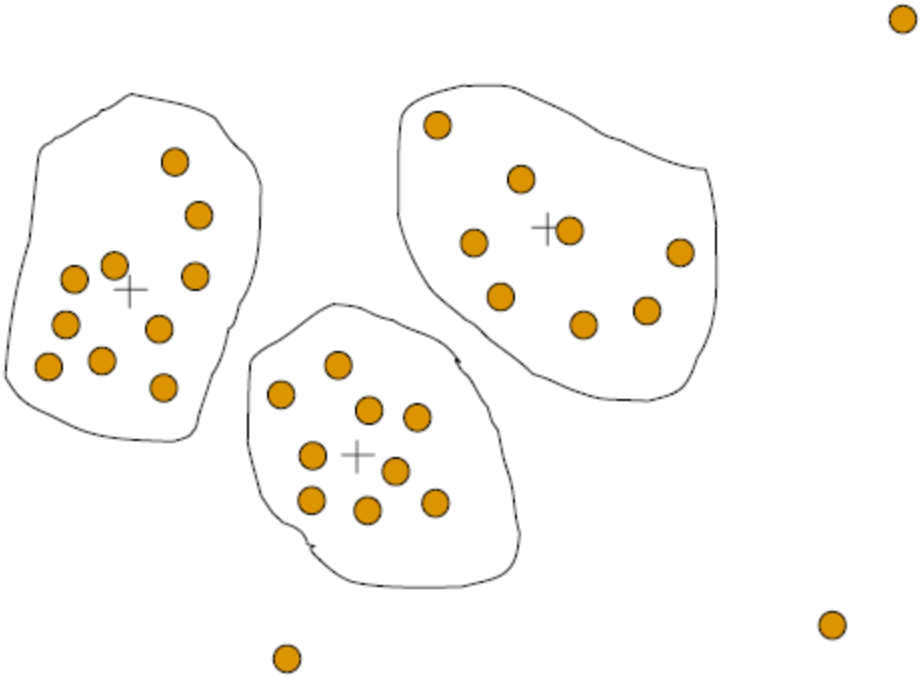
- Noisy Data: Random error or variance in a measured variable
- Simple Discretization Methods: Binning
  - Equal-depth(frequency) partitioning:
    - It divides the range into N intervals, each containing approximately same number of samples
    - Good data scaling
    - Managing categorical attributes can be tricky.
- Binning method for data mining in next slide with example

Data Preprocessing Techniques

* Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

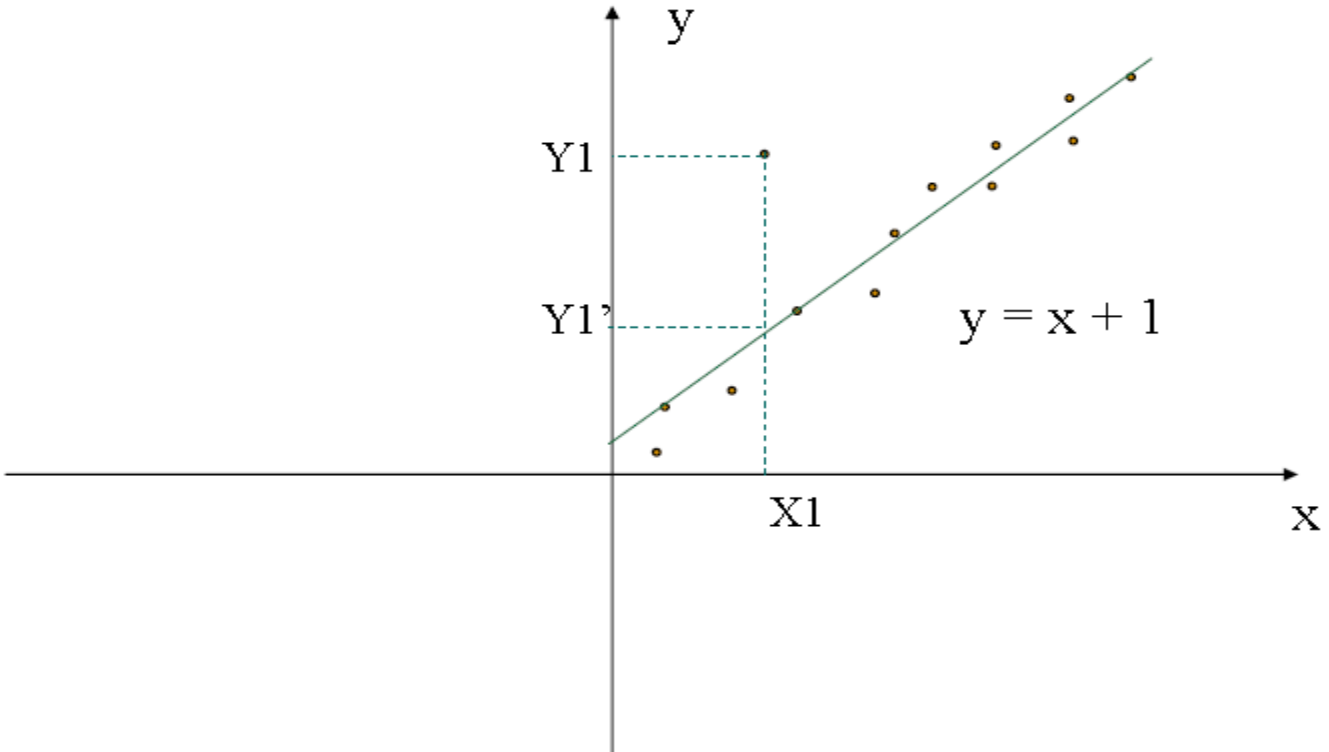
Data Preprocessing Techniques

- Noisy Data: Random error or variance in a measured variable
- Cluster Analysis:



Data Preprocessing Techniques

- Noisy Data: Random error or variance in a measured variable
- Regression:



Data Preprocessing Techniques

- Data Integration:
  - combines data from multiple sources into a coherent store → Data Warehouse
- Schema integration:
  - integrate metadata from different sources.
  - Entity identification problem example: A.custID = B.custNumber
  - Solution: entity identification → problems → meta data
- Detecting and resolving data value conflicts:
  - for the same real world entity, attribute values from different sources are different
  - possible reasons: different representations, different scales, e.g., metric vs. British units

Data Preprocessing Techniques

- Handling Redundant data in Data Integration:
  - The situations which leads to redundancy are
    - Inconsistency: inconsistencies in attributes or dimension naming, example: Derivable data
    - Object Identification: The same attribute may have different names in different databases

- Detected by correlational analysis:
  - helps to determine how strongly one attribute implies the other attribute.
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Preprocessing Te

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

- Handling Redundant data in Data Integration:
  - Correlation Analysis: (**Pearson's product moment coefficient**)
    - It lies between -1 and +1
    - The higher the coefficient value the stronger the correlation
    - if the correlation coefficient is greater than 0 that means A and B are positively correlated.
    - If correlation coefficient is equal to 0 then we call that the attribute A and B are independent that means they are not correlated.
    - if the correlation coefficient is negative then we say that A and b are negatively correlated.

Data Preprocessing Te

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- Handling Redundant data in Data Integration:
  - Correlation Analysis: (**Pearson's product moment coefficient**)
    - chi-square test to determine the correlation.
    - Suppose A has c-distinct values. $a_1, a_2, a_3 \ldots a_c$
    - Suppose B has r-distinct values. $b_1, b_2, b_3 \ldots b_r$
    - Use the **contingency table** to describe the tuples of A and B, with c-values of A making up columns and the r-values of B making up the rows.
    - A table summarization of two categorical variables in this form is called a contingency table. → contingency table
    - Joint event has its own cell or slot in the table. → Joint event (A_i, B_j).

Data Preprocessing Techniques

- Handling Redundant data in Data Integration:
  - Correlation Analysis: (**Pearson's product moment coefficient**)
    - Joint event has its own cell or slot in the table. → Joint event (A_i,B_j).

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

    - Where o_{ij} is the observed frequency of the joint event (A_i,B_j) and e_{ij} is the expected frequency of (A_i,B_j).

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N},$$

    - where N is the number of data tuples, count(A = a_i) is the number of tuples having value a_i for A like wise count(B = b_j)

Data Preprocessing Techniques

- Handling Redundant data in Data Integration:
  - Correlation analysis of categorical attributes using person chi-square statistic:
- Suppose that a group of 1,500 people was surveyed. The **gender** of each person was noted. Each person was polled as to whether their preferred type of reading material was **fiction** or **nonfiction. Thus, we have two attributes, gender and preferred reading**. The observed frequency (or count) of each possible joint event is summarized in the contingency table. Where the numbers in parentheses are the expected frequencies

Data Preprocessing Techniques

- Handling Redundant data in Data Integration:
  - Correlation analysis of categorical attributes using person chi-square statistic:

A 2×2 contingency table for the data of Example 2.1. Are *gender* and *preferred_Reading* correlated?

	male	*female*	Total
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Data Preprocessing Techniques

- Handling Redundant data in Data Integration:
  - Correlation analysis of categorical attributes using person chi-square statistic:

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.$$

  - Assume the degree of freedom is (r-1) x (c-1) → (2-1) x (2-1) = 1
  - For 1 degree of freedom, the chi-square value needed to reject the hypothesis at the **0.001** significance level is **10.828.**
  - **Since our computed value is above this, we can reject the hypothesis that gender and preferred reading are independent and conclude that the two attributes are (strongly) correlated for the given group of people.**

Data Preprocessing Techniques

- Data Transformation:
  - The data are transformed or consolidate into forms appropriate for mining
  - Data transformation can involve following:
    - Smoothing: removes noise from data
    - Aggregation: summarization, etc
    - Generalization: Concept of hierarchy climbing
    - Normalization: scale to fall within a small, specified range
    - Attribute or feature construction: new attributes are constructed from given set of attributes.

Data Preprocessing Techniques

- Data Transformation:
  - Normalization:
    - Min-max normalization
    - Z-score normalization
    - Normalization by decimal scaling

Data Preprocessing Techniques

- Data Transformation:
  - Min-max normalization:
    - It performs the linear transformation on original data
    - It preserves the relationships among the original data values.
    - Suppose min_A and max_A are the minimum and maximum values of an attribute, A.
    - It maps the value v of A to **v'** in the range [new_min_A, new_max_A] using formula

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Data Preprocessing Techniques

- Data Transformation:
  - Min-max normalization: Example
    - Suppose that the minimum and maximum values for the attribute income are 12,000 and 98,000 respectively. We would like to map income to the range [0.0,1.0]. By min-max normalization, a value of 73,000 for income is transformed to
    - (73,600 – 12,000)/(98,000-12,000) * (1.0-0.0) + 0.0 = 0.716

Data Preprocessing Techniques

- Data Transformation:
  - Z-score normalization: (**Zero – mean normalization**)
    - The values of an attribute A, are normalized based on the mean and standard deviation of A.
    - One of the value – v of A. is normalized to **v'** using

$$v' = \frac{v - \bar{A}}{\sigma_A};$$

    - Where $\bar{A}$ and σ_A are the mean and standard deviation of attribute A, respectively.
    - It is useful when the actual min and max of attribute A are unknown.

Data Preprocessing Techniques

- Data Transformation:
  - Z-score normalization: Example
    - Suppose that the mean and standard deviation of the values for the attribute income are 54000 and 16000 respectively.
    - The z-score normalization value of 73000 for income attribute A is transformed to
    - (73,600 – 54,000) /16,000 = 1.225

Data Preprocessing Techniques

- Data Transformation:
  - Normalization by decimal scaling:
    - It normalizes by moving the decimal point of values of attribute A.
    - The number of decimal points moved depends on the maximum absolute values of A.
    - The value – v of A is normalized to **v'** by using following formula $$v' = \frac{v}{10^j}$$
    - Where j is the smallest integer such that max(|v'|) < 1

Data Preprocessing Techniques

- Data Transformation:
  - Normalization by decimal scaling: ex
    - Suppose that the recorded values of A range from −986 to 917.
    - The maximum absolute value of A is 986.
    - To normalize by decimal scaling, we therefore divide each value by 1,000 (i.e, j=3) that −986 normalizes to −0.986 and 917 normalizes to 0.917.

Data Preprocessing Techniques

- Data Reduction:
  - Warehouse may store terabytes of data.
  - Complex data analysis/mining may take a very long time to run on the complete data set
  - So, Data Reduction is required
    - Obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

Data Preprocessing Techniques

- Data Reduction strategies:
  - Data cube aggregation
    - Aggregation operations are applied on the data in the construction of data cube
  - Attribute subset selection
    - Irrelevant, weakly relevant, redundant attributes may be removed
  - Dimensionality reduction
    - Use mechanism to reduce the data set size
  - Numerosity reduction
    - Data are replaced by alternatives