Data Preparation

(Data pre-processing)

INTRODUCTION TO DATA PREPARATION

Why Prepare Data?

- Some data preparation is needed for all mining tools
- The purpose of preparation is to transform data sets so that their information content is best exposed to the mining tool
- Error prediction rate should be lower (or the same) after the preparation as before it

Why Prepare Data?

- Preparing data also prepares the miner so that when using prepared data the miner produces better models, faster

- GIGO - good data is a prerequisite for producing effective models of any type

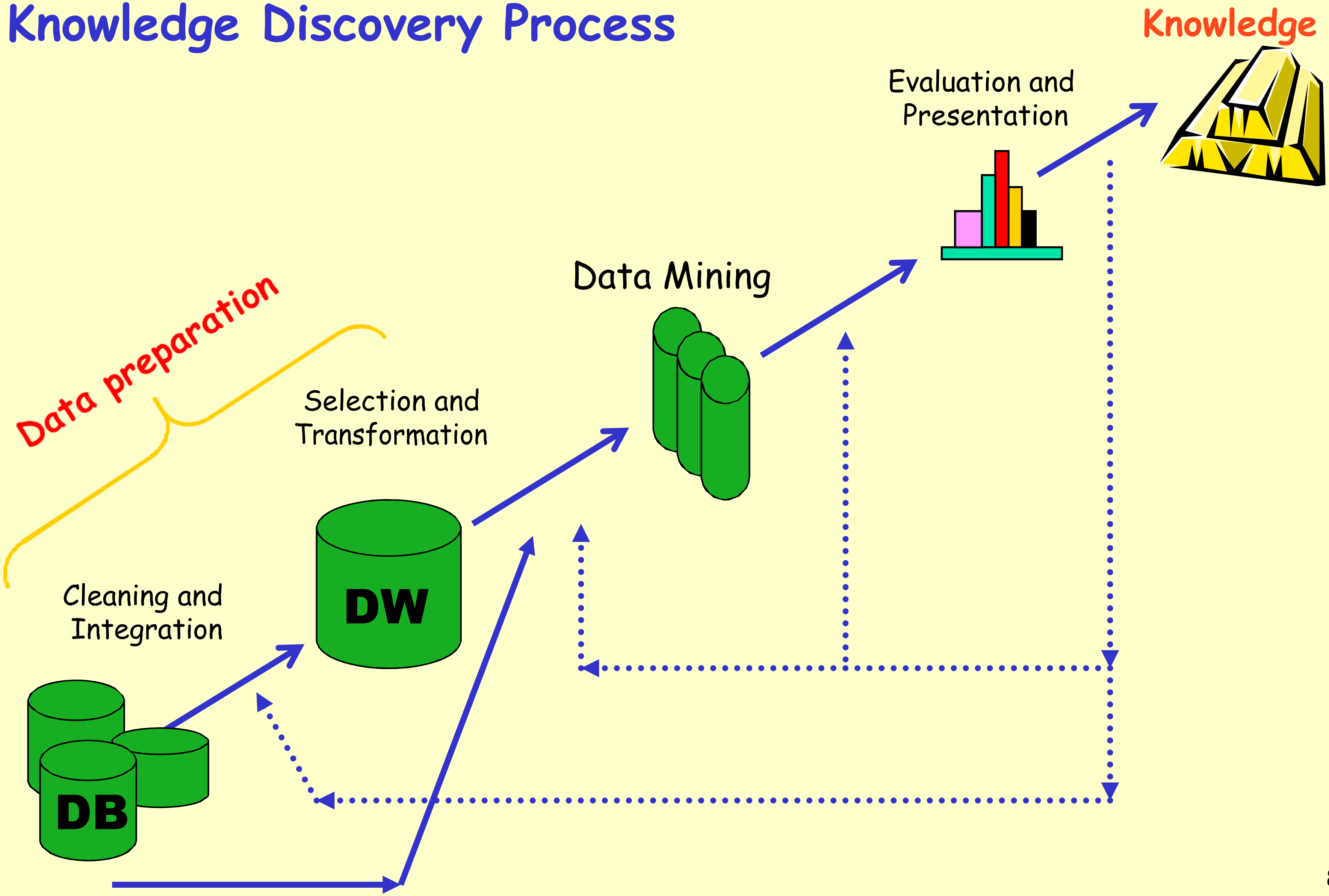
Why Prepare Data?

- Data need to be formatted for a given software tool
- Data need to be made adequate for a given method
- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=""
  - noisy: containing errors or outliers
    - e.g., Salary="-10", Age="222"
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age="42" Birthday="03/07/1997"
    - e.g., Was rating "1,2,3", now rating "A, B, C"
    - e.g., discrepancy between duplicate records
    - e.g., ***Endereço:*** travessa da Igreja de Nevogilde ***Freguesia:*** Paranhos

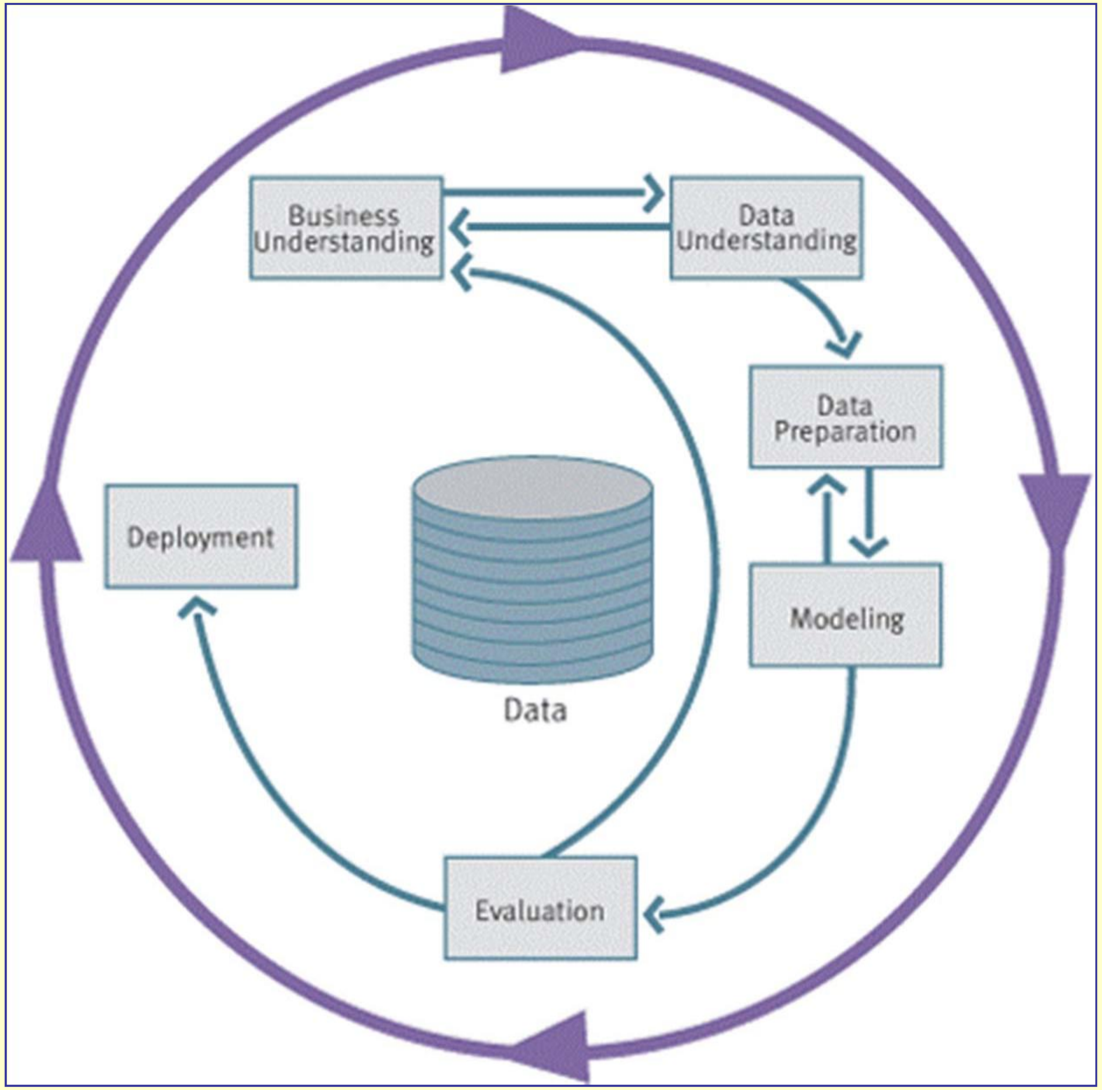
Major Tasks in Data Preparation

- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data
- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results

Data Preparation as a step in the Knowledge Discovery Process



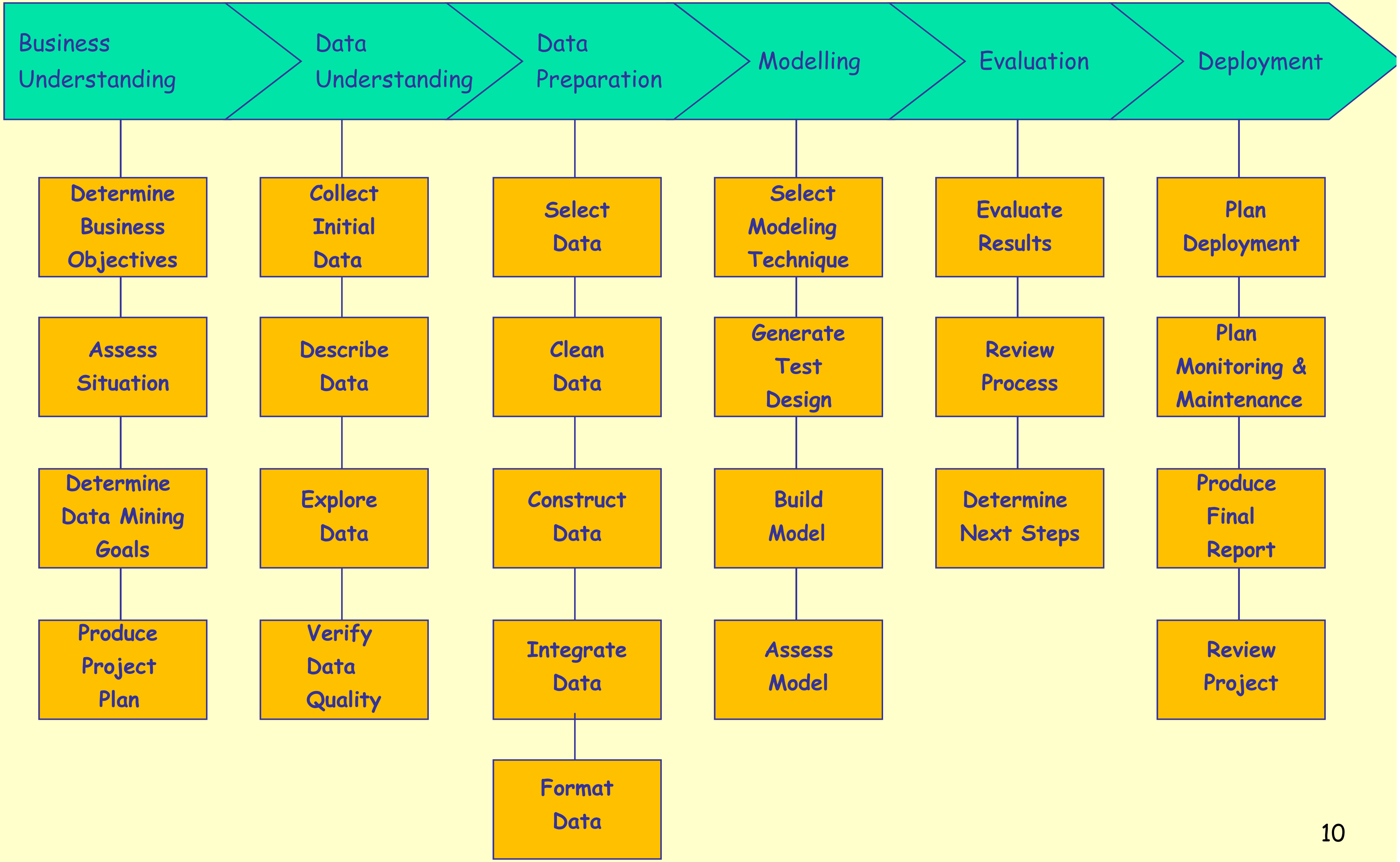
CRISP-DM



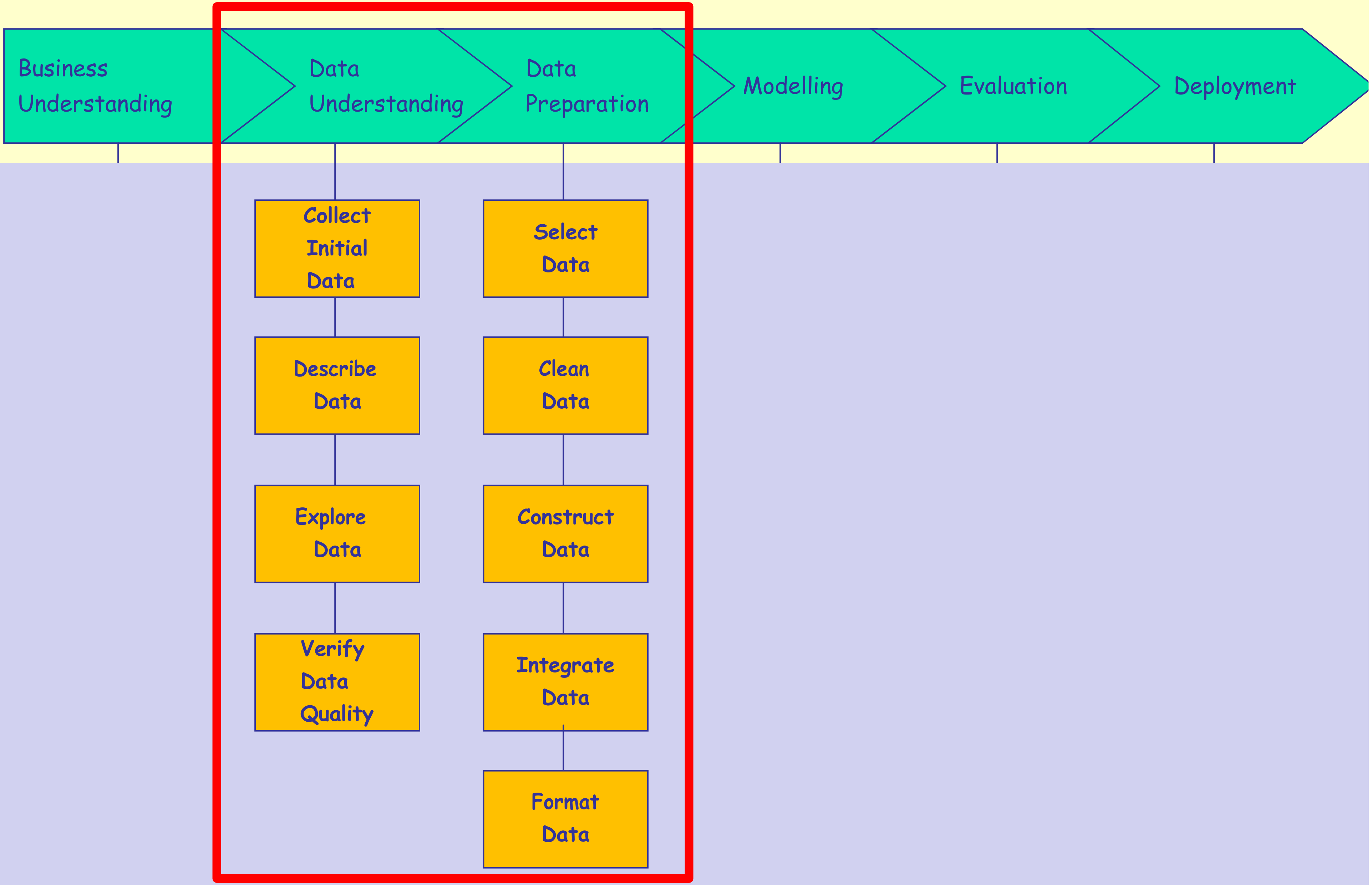
CRISP-DM is a comprehensive data **mining methodology** and process model that provides anyone—from novices to data mining experts—with a complete blueprint for conducting a data mining project.

A methodology enumerates the steps to reproduce success

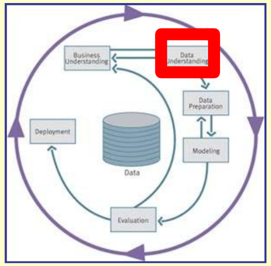
CRISP-DM Phases and Tasks



CRISP-DM Phases and Tasks

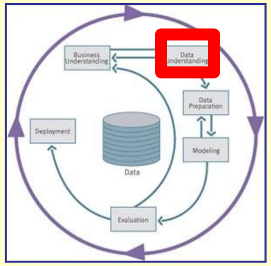


CRISP-DM: Data Understanding



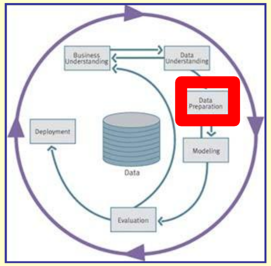
- **Collect data**
  - List the datasets acquired (locations, methods used to acquire, problems encountered and solutions achieved).
- **Describe data**
  - Check data volume and examine its gross properties.
  - Accessibility and availability of attributes. Attribute types, range, correlations, the identities.
  - Understand the meaning of each attribute and attribute value in business terms.
  - For each attribute, compute basic statistics (e.g., distribution, average, max, min, standard deviation, variance, mode, skewness).

CRISP-DM: Data Understanding



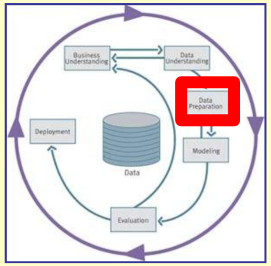
- **Explore data**
  - Analyze properties of interesting attributes in detail.
    - *Distribution, relations between pairs or small numbers of attributes, properties of significant sub-populations, simple statistical analyses.*
- **Verify data quality**
  - Identify special values and catalogue their meaning.
  - Does it cover all the cases required? Does it contain errors and how common are they?
  - Identify missing attributes and blank fields. Meaning of missing data.
  - Do the meanings of attributes and contained values fit together?
  - Check spelling of values *(e.g., same value but sometime beginning with a lower case letter, sometimes with an upper case letter).*
  - Check for plausibility of values, e.g. all fields have the same or nearly the same values.

CRISP-DM: Data Preparation



- **Select data**
  - Reconsider data selection criteria.
  - Decide which dataset will be used.
  - Collect appropriate additional data (internal or external).
  - Consider use of sampling techniques.
  - Explain why certain data was included or excluded.
- **Clean data**
  - Correct, remove or ignore noise.
  - Decide how to deal with special values and their meaning *(99 for marital status).*
  - Aggregation level, missing values, etc.
  - Outliers?

CRISP-DM: Data Preparation



- **Construct data**
  - Derived attributes.
  - Background knowledge.
  - How can missing attributes be constructed or imputed?
- **Integrate data**
  - Integrate sources and store result (new tables and records).
- **Format Data**
  - Rearranging attributes *(Some tools have requirements on the order of the attributes, e.g. first field being a unique identifier for each record or last field being the outcome field the model is to predict).*
  - Reordering records *(Perhaps the modelling tool requires that the records be sorted according to the value of the outcome attribute).*
  - Reformatted within-value *(These are purely syntactic changes made to satisfy the requirements of the specific modelling tool, remove illegal characters, uppercase lowercase).*

TYPES OF DATA

Types of Measurements

- Nominal scale
- Categorical scale
- Ordinal scale
- Interval scale
- Ratio scale

Qualitative

Quantitative

Discrete or Continuous

More information content

Types of Measurements: Examples

- Nominal:
  - ID numbers, Names of people
- Categorical:
  - eye color, zip codes
- Ordinal:
  - rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
- Interval:
  - calendar dates, temperatures in Celsius or Fahrenheit, GRE (Graduate Record Examination) and IQ scores
- Ratio:
  - temperature in Kelvin, length, time, counts

Types of Measurements: Examples

Day	Outlook	Temperature	Humidity	Wind	*PlayTennis?*
1	Sunny	85	85	Light	No
2	Sunny	80	90	Strong	No
3	Overcast	83	86	Light	Yes
4	Rain	70	96	Light	Yes
5	Rain	68	80	Light	Yes
6	Rain	65	70	Strong	No
7	Overcast	64	65	Strong	Yes
8	Sunny				
9	Sunny				
10	Rain				
11	Sunny				
12	Overcast				
13	Overcast				
14	Rain				

Day	Outlook	Temperature	Humidity	Wind	*PlayTennis?*
1	Sunny	Hot	High	Light	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Light	Yes
4	Rain	Mild	High	Light	Yes
5	Rain	Cool	Normal	Light	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Light	No
9	Sunny	Cool	Normal	Light	Yes
10	Rain	Mild	Normal	Light	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Light	Yes
14	Rain	Mild	High	Strong	No

Data Conversion

- Some tools can deal with nominal values but other need fields to be numeric

- Convert ordinal fields to numeric to be able to use ">" and "<" comparisons on such fields.
  - A → 4.0
  - A- → 3.7
  - B+ → 3.3
  - B → 3.0

- Multi-valued, unordered attributes with small no. of values
  - e.g. Color=Red, Orange, Yellow, …, Violet
  - for each value *v* create a binary "flag" variable *C_v* , which is 1 if Color=*v*, 0 otherwise

Conversion: Nominal, Many Values

- Examples:
  - US State Code (50 values)
  - Profession Code (7,000 values, but only few frequent)

- Ignore ID-like fields whose values are unique for each record
- For other fields, group values "naturally":
  - e.g. 50 US States → 3 or 5 regions
  - Profession – select most frequent ones, group the rest
- Create binary flag-fields for selected values

Parsing and Transformation

Smoothing: Smoothing is a process of removing noise from the data.

Aggregation: Aggregation is a process where summary or aggregation operations are applied to the data.

Generalization: In generalization low-level data are replaced with high-level data by using concept hierarchies climbing.

Normalization: Normalization scaled attribute data so as to fall within a small specified range, such as 0.0 to 1.0.

Attribute Construction: In Attribute construction, new attributes are constructed from the given set of attributes.

Scalability

Scalability is the capability of a system, network, or process to handle a growing amount of work, or its potential to be enlarged to accommodate that growth

- Administrative scalability: The ability for an increasing number of organizations or users to easily share a single distributed system.
- Functional scalability: The ability to enhance the system by adding new functionality at minimal effort.
- Geographic scalability: The ability to maintain performance, usefulness, or usability regardless of expansion from concentration in a local area to a more distributed geographic pattern.
- Load scalability: The ability for a distributed system to easily expand and contract its resource pool to accommodate heavier or lighter loads or number of inputs. Alternatively, the ease with which a system or component can be modified, added, or removed, to accommodate changing load.
- Generation scalability: The ability of a system to scale up by using new generations of components. Thereby, heterogeneous scalability is the ability to use the components from different vendors.

Data Cleaning

Data cleaning

Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - <u>incomplete</u>: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation*=" " (missing data)
  - <u>noisy</u>: containing noise, errors, or outliers
    - e.g., *Salary*="−10" (an error)
  - <u>inconsistent</u>: containing discrepancies in codes or names, e.g.,
    - *Age*="42", *Birthday*="03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"
    - discrepancy between duplicate records
  - <u>Intentional</u> (e.g., *disguised missing* data)
    - Jan. 1 as everyone's birthday?

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably

- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems

- technology limitation
- inconsistency in naming convention

- Other data problems which require data cleaning
  - duplicate records
  - incomplete data

inconsistent data

How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering

- detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Data Cleaning as a Process

- Data discrepancy detection
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and

clustering to find outliers)

- Data migration and integration
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
- Integration of the two processes
  - Iterative and interactive (e.g., Potter's Wheels)

Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values
- Methods
  - Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization, data cube construction

- Normalization: Scaled to fall within a smaller, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Discretization: Concept hierarchy climbing

Normalization

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]
  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to
- **Z-score normalization** (μ: mean, σ: standard deviation):
  - Ex. Let μ = 54,000, σ = 16,000. Then
- **Normalization by decimal scaling**

Where j is the smallest integer such that Max(|v'|) < 1

$$v' = \frac{v}{10^j}$$

Discretization

- Three types of attributes
  - Nominal—values from an unordered set, e.g., color, profession
  - Ordinal—values from an ordered set, e.g., military or academic rank
  - Numeric—real numbers, e.g., integer or real numbers
- Discretization: Divide the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Reduce data size by discretization
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
  - Prepare for further analysis, e.g., classification

Data Discretization Methods

- Typical methods: All the methods can be applied recursively
  - Binning
    - Top-down split, unsupervised
  - Histogram analysis
    - Top-down split, unsupervised
  - Clustering analysis (unsupervised, top-down split or bottom-up merge)

- Decision-tree analysis (supervised, top-down split)
- Correlation (e.g., χ^2) analysis (unsupervised, bottom-up merge)

Simple Discretization: Binning

- Equal-width (distance) partitioning
  - Divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are 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
- Equal-depth (frequency) partitioning
  - Divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (**equi-depth**) bins:

- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

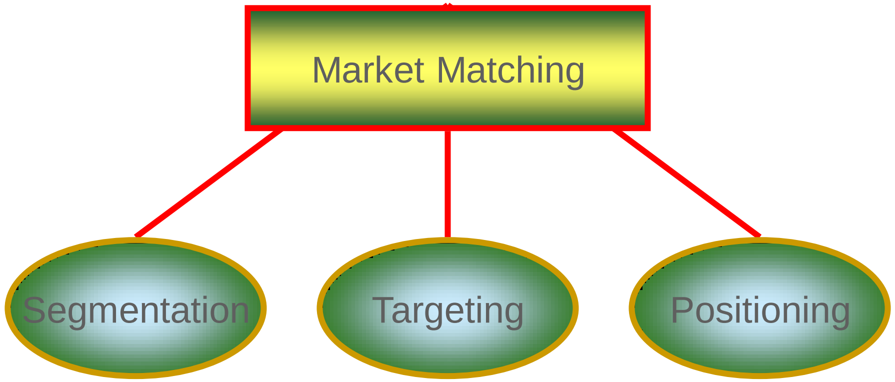
* Smoothing by **bin means**:

- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

* Smoothing by **bin boundaries**:

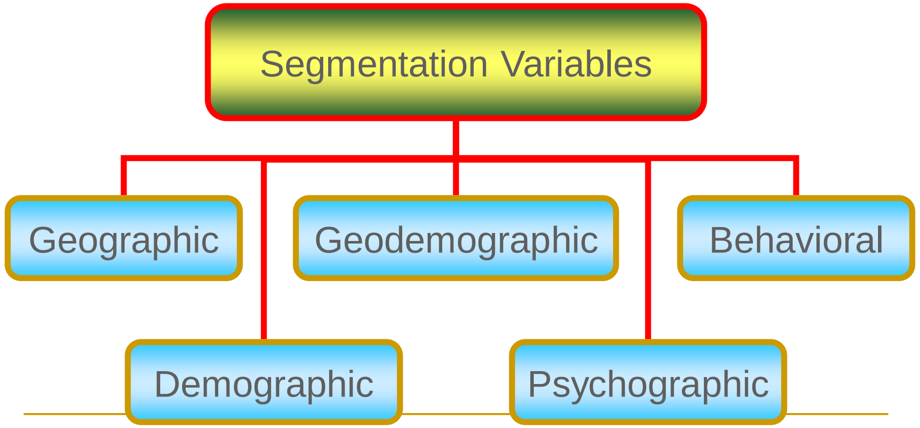
- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

Market Matching Strategy



- **_Segmentation_**
- Act of dissecting the marketplace into submarkets that require different *marketing mixes*

Segmentation Variables



- Marketers may use a single variable
- Marketers may use two or more variables

Geographic Segmentation

- Division of the market based on the *location* of the target market
- People living in the same area have *similar needs and wants* that differ from those living in other areas
- Climate
- Population density
- Taste
- Micromarketing

Demographic Segmentation

- Partitioning of the market based on factors such as
  - age
  - gender
  - marital status
  - income
  - occupation
  - education

ethnicity

Geodemographic Segmentation

- A *hybrid* segmentation scheme
- Based on notion that people who live close to one another are likely to have similar financial means, tastes, preferences, lifestyles and consumption habits

Psychographic Segmentation

- Partitioning of the market based on *lifestyle* and *personality* characteristics
- Marketers use it to further refine a target market
- Its appeal lies in the vivid and practical profiles of consumer segments that it can produce
- Accomplished by using AIO inventories

Behavioral Segmentation

- Partitioning of the market based on attitudes toward or reactions to a product and to its promotional appeals
- Behavioral segmentation can be done on the basis of:

1. Usage rate
2. Benefits sought from a product
3. Loyalty to a brand or a store

Exploratory analysis

Statistics

- Powerful tools… we must use them for good.
  - Be sure our data is valid and reliable
  - Be sure we have the right type of data
  - Be sure statistical tests are applied appropriately
  - Be sure the results are interpreted correctly
  - Remember… numbers may not lie, but people can
- Statistics
- Numerical representations of our data

- Can be:
  - Descriptive statistics summarize data.
  - Inferential statistics are tools that indicate how much confidence we can have when we generalize from a sample to a population.
-

Sampling & Statistics

- Statistics depend on our sampling methods:
- Probability or Non-probability? (i.e. Random or not?)

- Statistics: What's What?

- Descriptive objectives/ research questions:
- Comparative objectives/ hypotheses

- Descriptive statistics
- Inferential Statistics

Descriptive Statistics

- Number
- Frequency Count
- Percentage
- Deciles and quartiles
- Measures of Central Tendency (Mean, Midpoint, Mode)
- Variability
- Variance and standard deviation
- Graphs
- Normal Curve

Descriptive Statistics

- Can be applied to any measurements (quantitative or qualitative)
- Offers a summary/ overview/ description of data. Does <u>not</u> explain or interpret.

Means of Central Tendency

- <u>Averages</u>
  - Mode: most frequently occurring value in a distribution (any scale, most unstable)
  - Median: midpoint in the distribution below which half of the cases reside (ordinal and above)
  - Mean: arithmetic average- the sum of all values in a distribution divided by the number of cases (interval or ratio)

Median (Mid-point)

- Example (11 test scores)

61, 61, 72, 77, 80, 81, 82, 85, 89, 90, 92

The median is 81 (half of the scores fall above 81, and half below)

Median (Mid-point)

- Example (6 scores)

3, 3, 7, 10, 12, 15

Even number of scores= Median is half-way between these scores

Sum the middle scores (7+10=17) and divide by 2

17/2= **8.5**

Median

- Insensitive to extremes

3, 3, 7, 10, 12, 15, 200

Mean: Arithmetic Average

- Mean is half the sum of a set of values:
- Scores: 5, 6, 7, 10, 12, 15
- Sum: 55
- Number of scores: 6
- Computation of Mean: 55/6= **9.17**

Mean

- Influenced by extremes
- Only appropriate with interval or ration data

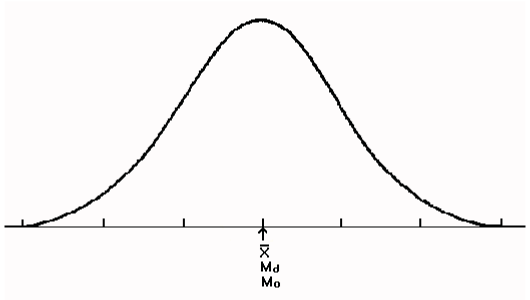
- Is this four-point scale ordinal or interval?

1= Strongly Agree	3=Disagree
2=Agree	4=Strongly Disagree

Mode: Frequency

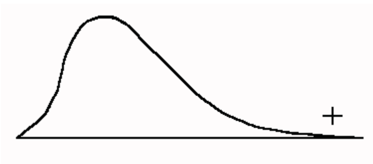
- Mode is the most frequently occurring value in a set.
- Best used for nominal data.

Normal Curve



Distribution: Skewness

- Skewed to the right (positive) or left (negative)
- An extremely hard test that results in a lot of low grades will be skewed to the right:



- Allows for comparisons across variables
  - i.e. is there a relation between one's occupation and their reason for using the public library?
- Hypothesis Testing

Error Type

- Type I error
  - Reject the null hypothesis when it is really true
- Type II error
  - Fail to reject the null hypothesis when it is really false

Probability

- By using inferential statistics to make decisions, we can report the probability that we have made a Type I error (indicated by the *p* value we report)
- By reporting the p value, we alert readers to the odds that we were incorrect when we decided to reject the null hypothesis

Levels of significance

- The level of significance is the predetermined level at which a null hypothesis is not supported. The most common level is $p < .05$
  - P =probability
  - < = less than (> = more than)

Particular Tests

- Chi-square test of independence: two variables (nominal and nominal, nominal and ordinal, or ordinal and ordinal)
  - Affected by number of cells, number of cases
  - 2-tailed distribution= null hypothesis
  - 1-tailed distribution= directional hypothesis
  - Cramer's V, Phi
      - **example**

Inferential Statistics (2)

- Correlation—the extent to which two variables are related across a group of subjects
  - Pearson r
    - It can range from -1.00 to 1.00
    - -1.00 is a perfect inverse relationship—the strongest possible inverse relationship
    - 0.00 indicates the complete absence of a relationship
    - 1.00 is a perfect positive relationship—the strongest possible direct relationship
    - The closer a value is to 0.00, the weaker the relationship
    - The closer a value is to -1.00 or +1.00, the stronger it is
  - Spearman rho

More tests

- t-test
  - Test the difference between two sample means for significance
  - pretest to posttest
  - Relates to research design
  - Perhaps used for information literacy instruction

Analysis of variance

- Regression analysis (including step-wise regression)

More tests

Analysis of variance (ANOVA) tests the difference(s) among two or more means

- It can be used to test the difference between two means
- So use t-test or ANOVA?
- KEY: ANOVA also can be used to test the difference among ***more than two means in a single test—which cannot be done with a t test***