Data Visualization

Lecture 6

Introduction to Data Preprocessing

&

Types of Data

Types of Data

- Understanding the different data types can help us identify correct preprocessing techniques & convert the data appropriately.

- Furthermore, it will also enable us to perform the best visualizations and uncover hidden knowledge.

Types of Data

- **Types of data based on structure:**

✔ **Structured data**

This type of data is usually composed of numbers or words. They are usually stored in Relational databases

✔ **Unstructured data**

Including texts, images, videos, speech/audio, and so on.

✔ **Semi-structured**

Use of metadata or tags that provide additional information about the data elements. For example (XML)

Types of Data

✔**Numeric/Quantitative data**

Can be represented through numbers. For example (sales price)

✔**Discrete**

Data takes on discrete values or whole numbers i.e. numbers without decimal points. For example: number of houses in a city, the number of consumers in a grocery store over the last month, and so on.

✔**Continuous**

Data takes on integer values i.e. numbers with decimal values. For example (house prices)

Types of Data

✔Categorical/Qualitative data

This encompasses data that can be represented through words. It usually defines groups or categories. For example (movie ratings(good, average, bad), country of birth of individuals & so on.)

✔Ordinal

This type of data has an inherent ordering present within the categories. For instance, if you consider movie ratings with good, average & bad as the different categories, good has a higher ranking than average which is higher than bad.

✔Nominal

This type of data has categories that don’t have any particular order or ranking associated with them. The total number of categories is usually finite in this type of data as well. Examples will be the country of birth of individuals

Data Formats

- **Tabular data**

  This usually refers to the collection of data from multiple different data types as shown in Figure 1. The tabular data consists of multiple features/columns with each of them having a particular data type.

Item ID	Item name	Sale price	Quantity	Item category	Avg user review	Average sales per day
IT20TN22	Eggs	5.75	24	Dairy	Good	125
IT20TN23	Bagels	4.25	12	Bakery	Poor	150
IT20TN24	Waffles	1.99	8	Frozen Foods	Average	45
IT20TN25	Milk	5.8	1	Dairy	Good	100

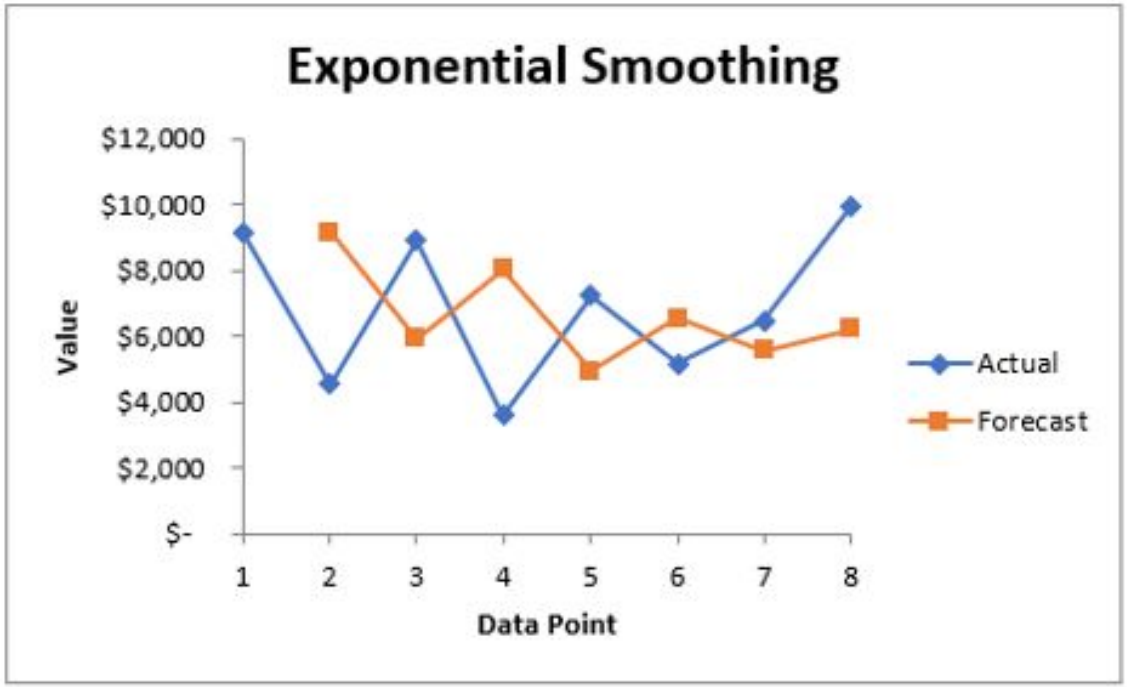
Data Formats

- **Time series data**

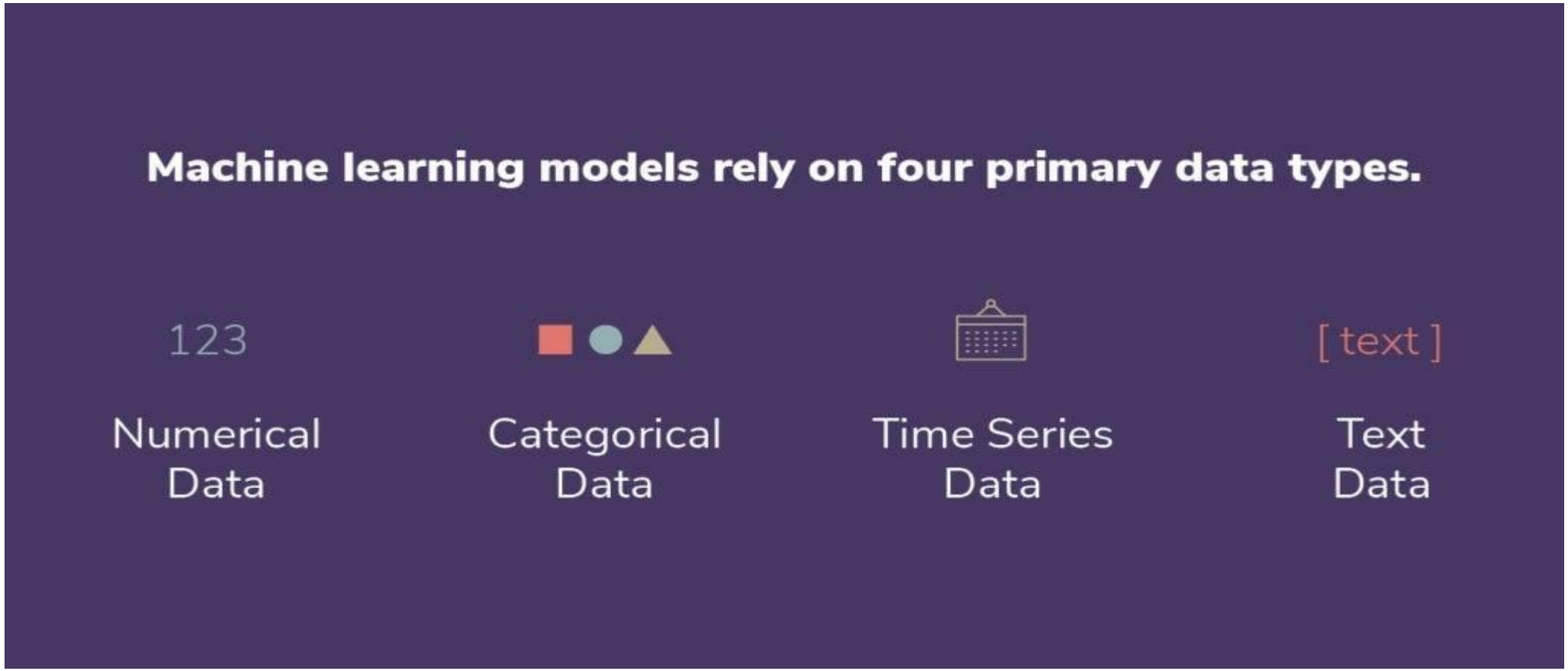
  Time series data consists of data points that are indexed at specific points in time. More often than not, this data is collected at consistent intervals. Learning and utilizing time series data makes it easy to compare data from week to week, month to month, year to year, or according to any other time-based metric you desire. The distinct difference between time series data and numerical data is that time series data has established starting and ending points, while numerical data is simply a collection of numbers that aren’t rooted in particular time periods.

Time series data

Year	Quarter	Revenue	Smoothed Levels	Standard Errors
2020	1	$ 9,150	#N/A	#N/A
	2	$ 4,560	$ 9,150	#N/A
	3	$ 8,920	5937	#N/A
	4	$ 3,615	8025.1	#N/A
2022	1	$ 7,245	4938.03	4058.545347
	2	$ 5,150	6552.909	3350.09361
	3	$ 6,480	5570.8727	2985.478535
	4	$ 9,950	6207.26181	1644.868454



Types of Data

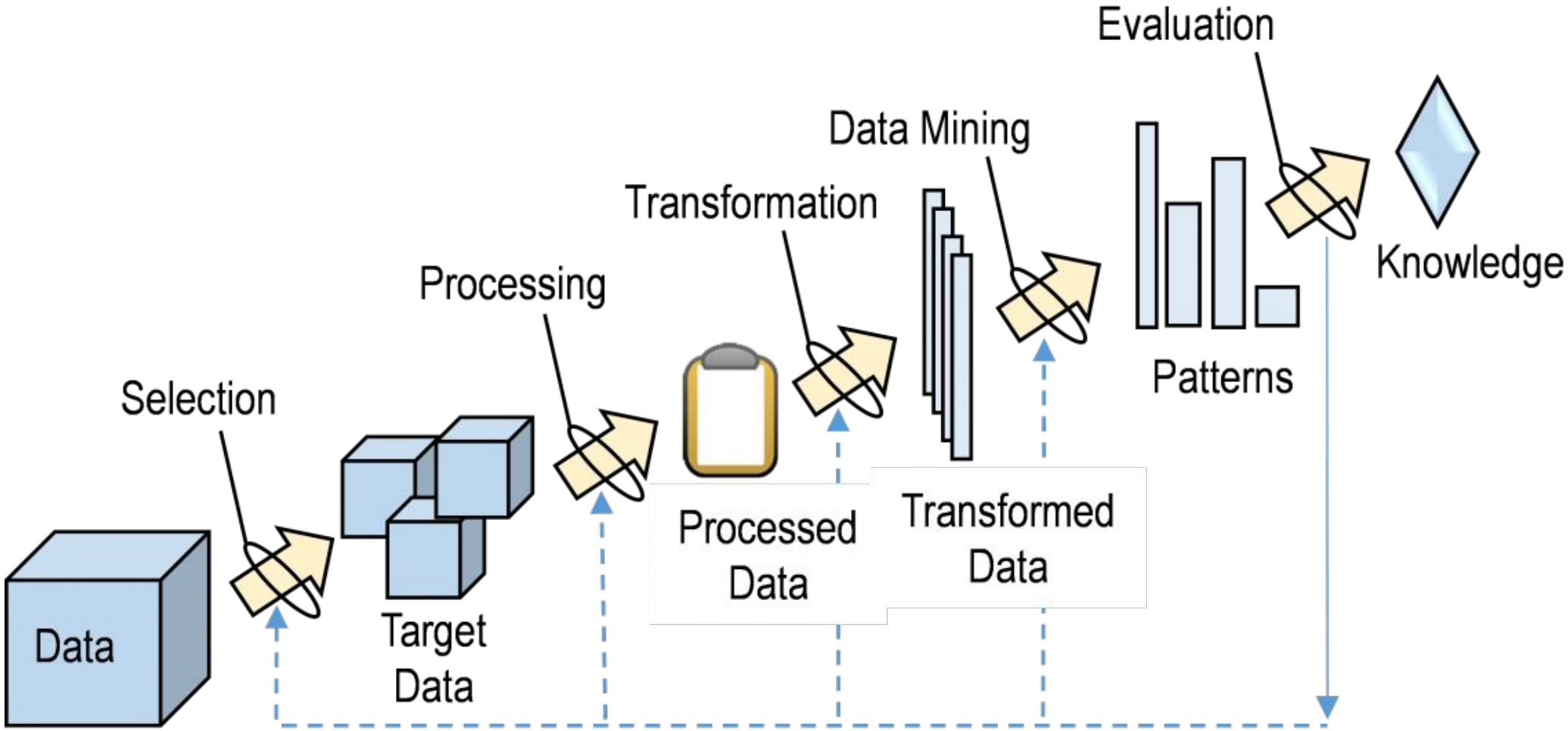


Data Preprocessing

What is data preprocessing?

- The fundamental purpose of data preparation is to manipulate and transform raw data so that the information content enfolded in the data set can be exposed or made more easily accessible.

Data preprocessing in Data Mining and ML process



Why Data Preprocessing?

1. Real data could be dirty and could drive to the extraction of useless patterns/rules.

- **This is mainly due to**:
  - Incomplete data: lacking attribute values, …
  - Data with noise: containing errors or outliers
  - Inconsistent data (including discrepancies)

Why Data Preprocessing?

2. Data preprocessing can generate a smaller dataset than the original, which allows us to improve the efficiency in the Data Mining process.

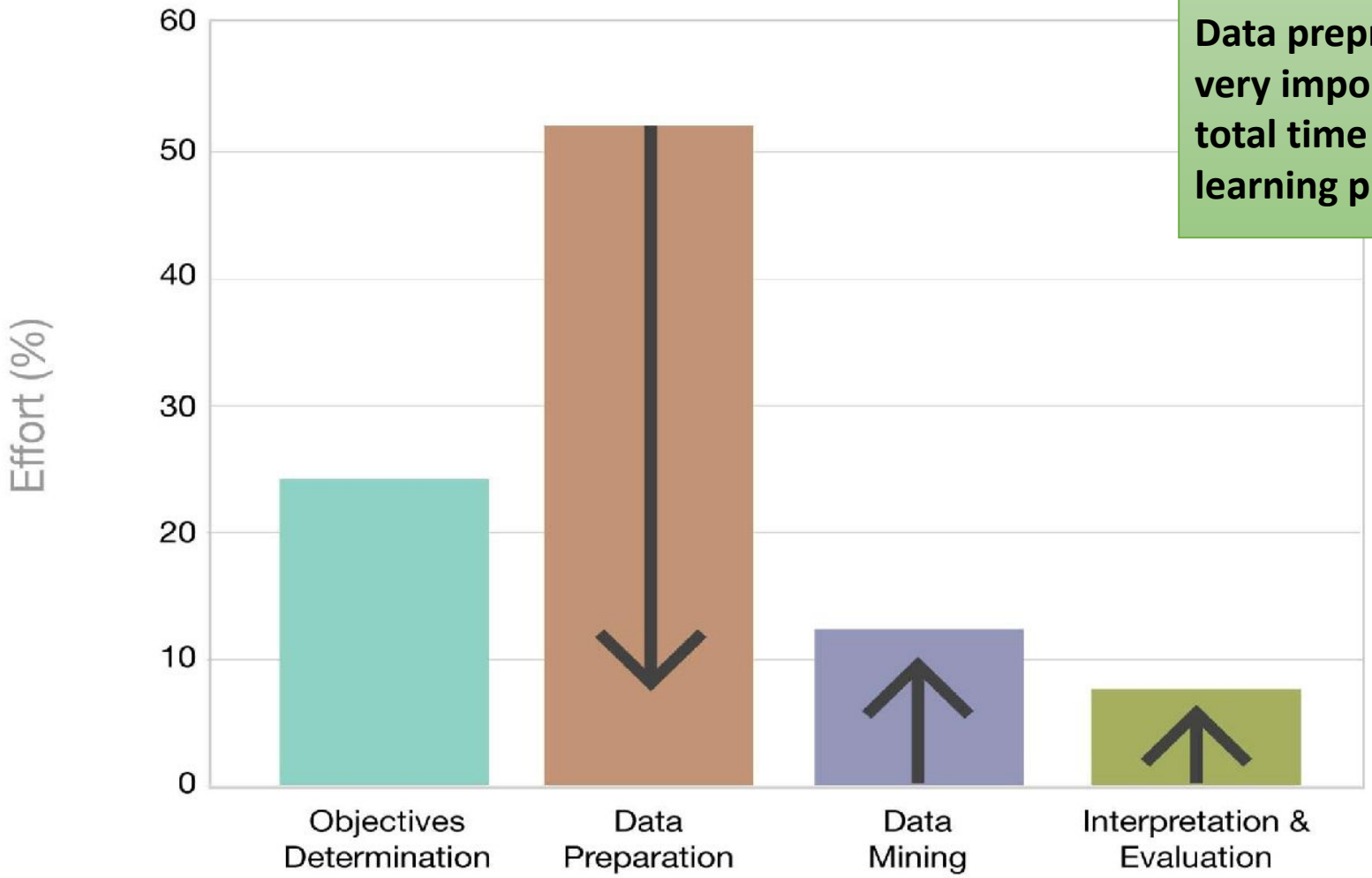
- **This performing includes Data Reduction techniques**:
  - Feature selection
  - sampling or instance selection
  - discretization

Why Data Preprocessing?

3. No quality data, no quality mining results!

- **Data preprocessing techniques generate**:
  - “quality data”, driving us to obtain “quality patterns/rules”.

Why Data Preprocessing?



Data preprocessing spends a very important part of the total time in a Machine learning process.

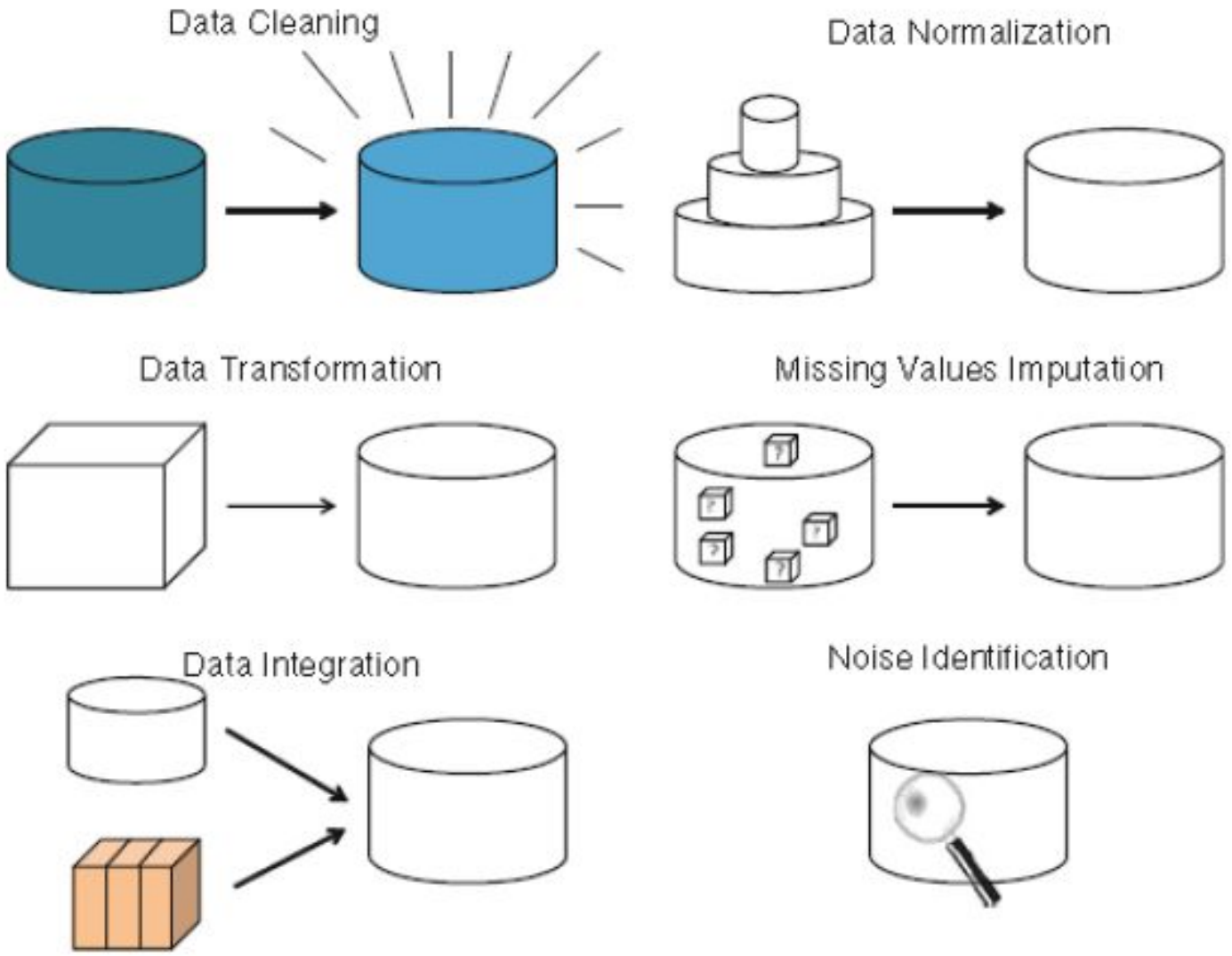
What is included in data preprocessing?

- Real databases usually contain noisy data, missing data, and inconsistent data, …

Major Tasks in Data Preprocessing

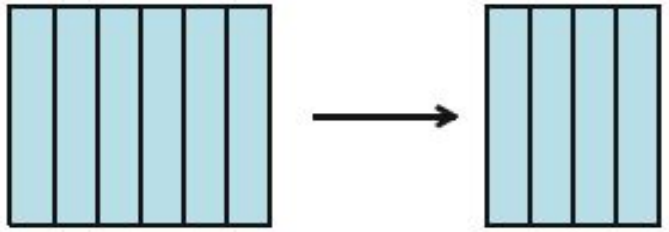
- Data integration. Fusion of multiple sources in a Data Warehousing.
- Data cleaning. Removal of noise and inconsistencies.
- Missing values imputation.
- Data Transformation.
- Data reduction.

What is included in data preprocessing?



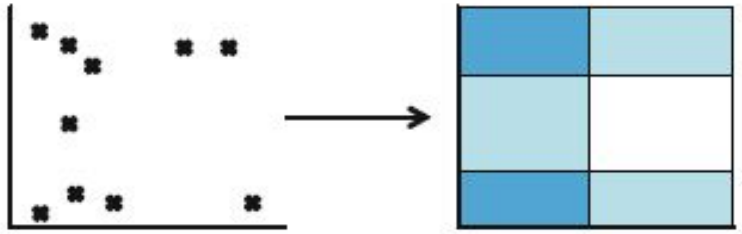
What is included in data preprocessing?

Feature Selection



Instance Selection

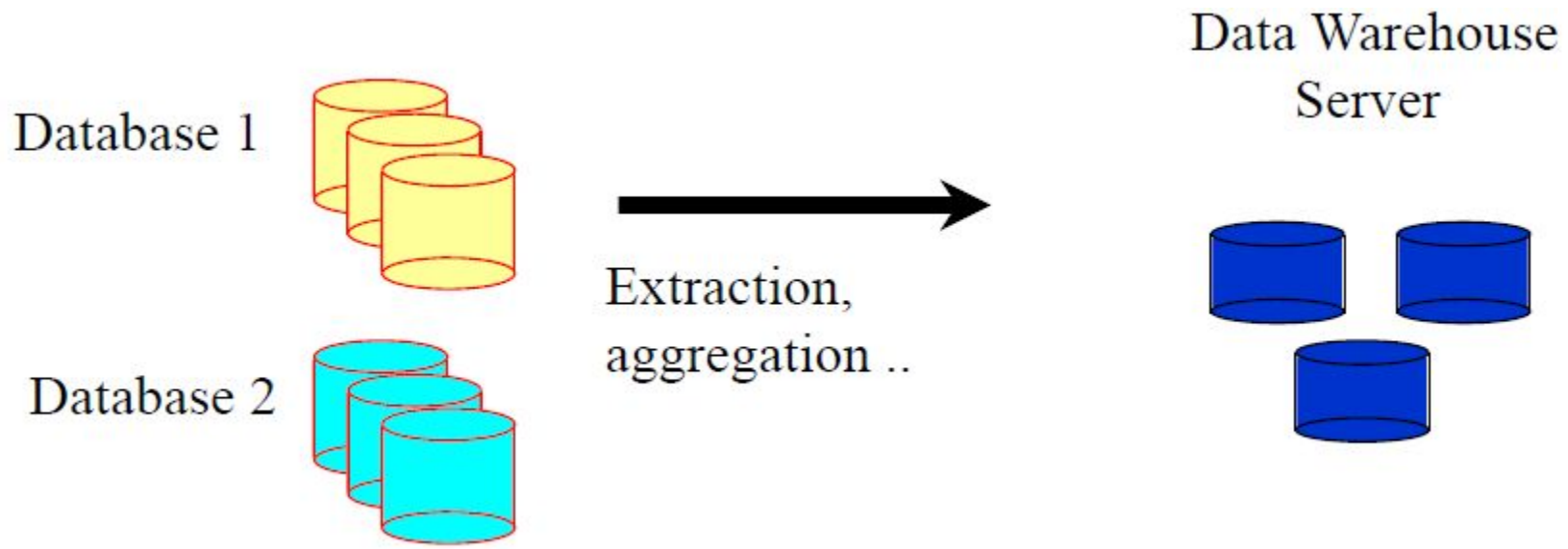
Discretization



Data Integration

- Obtain data from different information sources.
- Address problems of codification and representation.
- Integrate data from different tables to produce homogeneous information,

...



Data Cleaning

- Fix inconsistencies
- Fill/impute missing values,
- Smooth noisy data,
- Identify or remove outliers …

Data Cleaning: Inconsistent data

Age="42"
Birth Date="03/07/1997"

Data Transformation

- To transform data in the best way possible to the application of Data Mining algorithms.

Transformation typical operations

- Aggregation. i.e. Sum of the totality of month sales in an unique attribute called annual sales,…
- Data generalization. It is to obtain higher degrees of data from the currently available, by using concept hierarchies.
  - ✔ Streets cities
  - ✔ Numerical age {young, adult, half-age, old}
- Normalization: Change the range [-1,1] or [0,1].
- Lineal transformations, quadratic, polynomial, …

Data Normalization and Standardization

- To convert the values of an attribute to a better range.

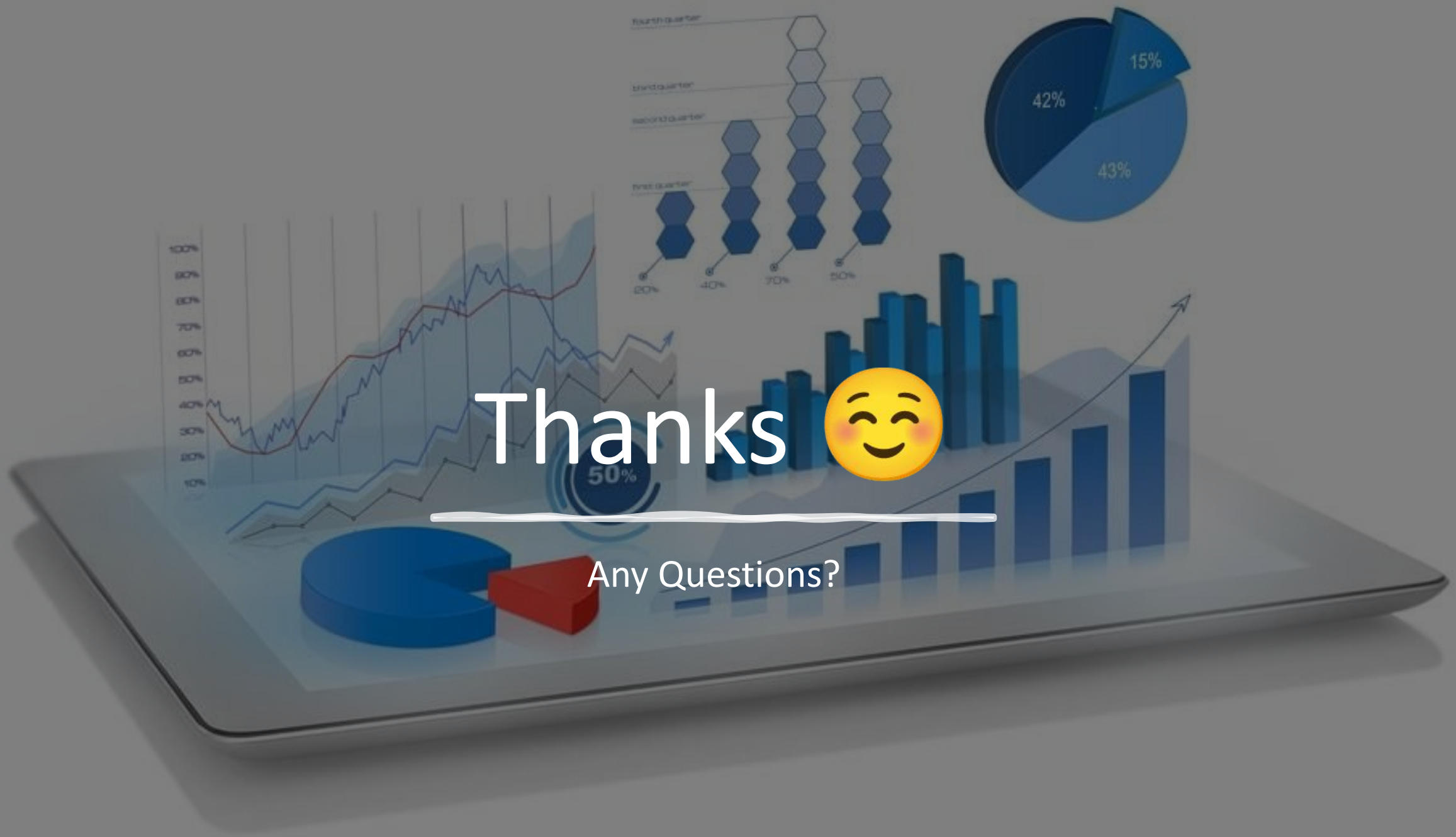
Some techniques

- Z-score standardization

$$v' = \frac{v - \overline{A}}{\sigma_A}$$

- min-max normalization

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new_{\max_A} - new_{\min_A}) + new_{\min_A}$$



Thanks 😊

Any Questions?