Data Preprocessing

Mrinal Das

Example: Manufacturing Chemical Products

Associated Data

Month	Temp	Time	Speed	Water	Material	Fragrance
March	40	20	300	75	15	10
April	39	20	300	70	15	15
May	40	18	200	80	15	5
June	40	20	250	80	15	5

What can we get from this ?

Associated Data

Month	Temp	Time	Speed	Water	Material	Fragrance
March	40	20	300	75	15	10
April	39	20	300	70	15	15
May	40	18	200	80	15	5
June	40	20	250	80	15	5

Month	Quality Index
March	5
April	7
May	2
June	3

There is a variation in quality.

Descriptive

But why ?

Diagnostic

Inference

Month	Temp	Time	Speed	Water	Material	Fragrance
March	40	20	300	75	15	10
April	39	20	300	70	15	15
May	40	18	200	80	15	5
June	40	20	250	80	15	5

Month	Quality Index
March	5
April	7
May	2
June	3

`If Water < 75 then QI > 4`

Diagnostic

Inference

Month	Temp	Time	Speed	Water	Material	Fragrance	Quality Index
March	40	20	300	75	15	10	5
April	39	20	300	70	15	15	7
May	40	18	200	80	15	5	2
June	40	20	250	80	15	5	4
January	40	18	200	75	15	10	5
February	40	20	200	80	15	5	3

`If Water < 75 then QI > 4`

More data increases confidence.

Prediction

Month	Temp	Time	Speed	Water	Material	Fragrance	Quality Index
March	40	20	300	75	15	10	5
April	39	20	300	70	15	15	7
May	40	18	200	80	15	5	2
June	40	20	250	80	15	5	4
January	40	18	200	75	15	10	5
February	40	20	200	80	15	5	3
July	39	20	300	80	15	5	?

```
If Water < 75 then QI > 4
```

Predictive analytics

Prescriptive

Month	Temp	Time	Speed	Water	Material	Fragrance	Quality Index
March	40	20	300	75	15	10	5
April	39	20	300	70	15	15	7
May	40	18	200	80	15	5	2
June	40	20	250	80	15	5	4
January	40	18	200	75	15	10	5
February	40	20	200	80	15	5	3
July	39	20	300	80	15	5	?
August	?	?	?	?	?	?	?

How should we set the parameters for better quality ?

Prescriptive

Month	Temp	Time	Speed	Water	Material	Fragrance	Quality Index
March	40	20	300	75	15	10	5
April	39	20	300	70	15	15	7
May	40	18	200	80	15	5	2
June	40	20	250	80	15	5	4
January	40	18	200	75	15	10	5
February	40	20	200	80	15	5	3
July	39	20	300	80	15	5	?
August	?	?	?	?	?	?	?

1. Create a virtual model of the physical system.
2. Simulate the system with various parameters.
3. Use predictive modeling to find the target values.
4. Infer the logic and set the parameters.

Prescriptive

Month	Temp	Time	Speed	Water	Material	Fragrance	Quality Index
March	40	20	300	75	15	10	5
April	39	20	300	70	15	15	7
May	40	18	200	80	15	5	2
June	40	20	250	80	15	5	4
January	40	18	200	75	15	10	5
February	40	20	200	80	15	5	3
July	39	20	300	80	15	5	?
August	?	?	?	80	?	?	?

1. Create a virtual model of the physical system.
2. Simulate the system with various parameters.
3. Use predictive modeling to find the target values.
4. Infer the logic and set the parameters.

Keep water > 75

Prescriptive

1. Create a virtual model of the physical system.
2. Simulate the system with various parameters.
3. Use predictive modeling to find the target values.
4. Infer the logic and set the parameters.

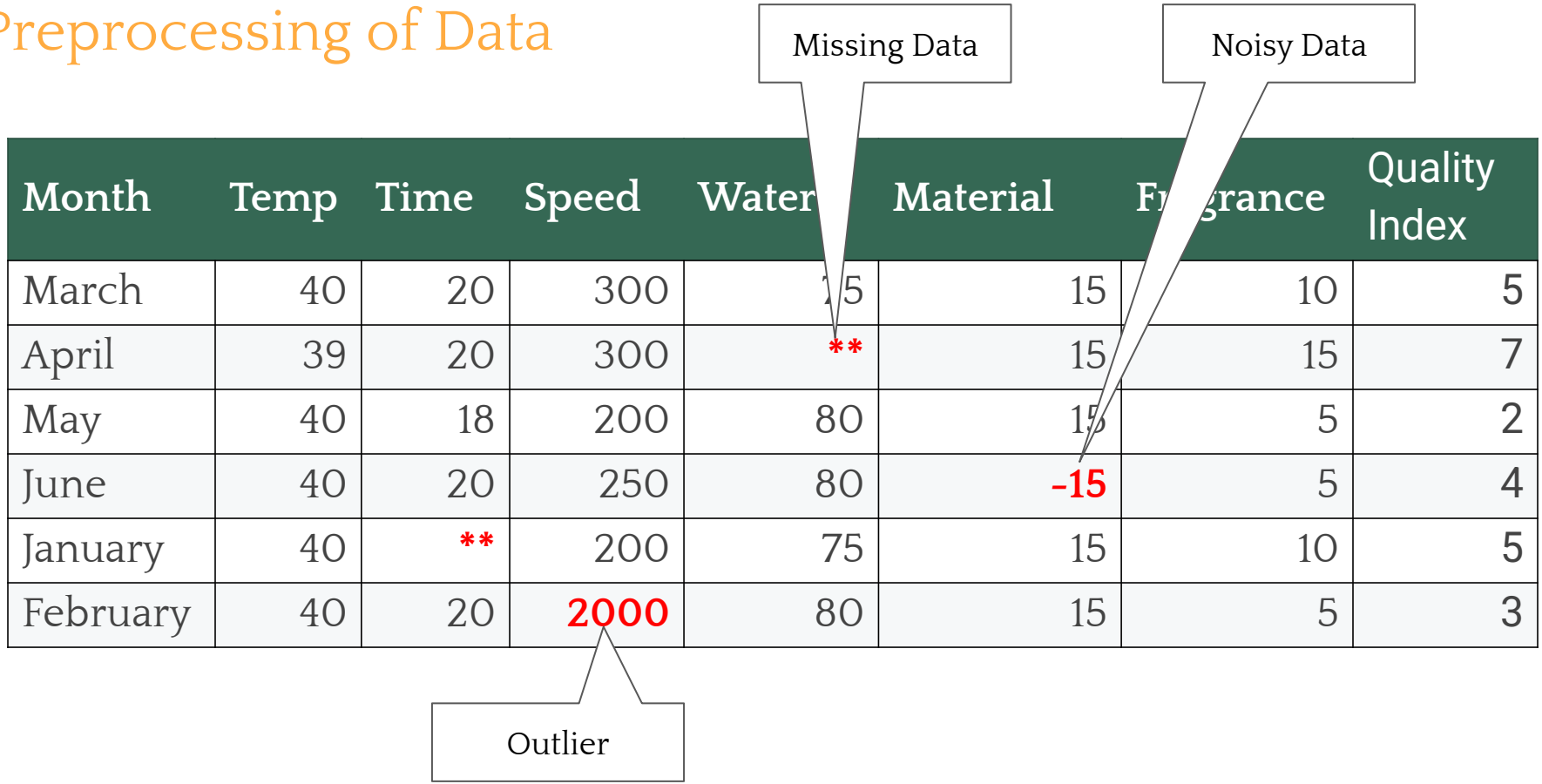
Difficult problem, need advanced techniques and large data.

Related topics: causality analysis, inverse problem

Quality of Data

Month	Temp	Time	Speed	Water	Material	Fragrance	Quality Index
March	40	20	300	75	15	10	5
April	39	20	300	**	15	15	7
May	40	18	200	80	15	5	2
June	40	20	250	80	-15	5	4
January	40	**	200	75	15	10	5
February	40	20	2000	80	15	5	3

Preprocessing of Data



Month	Temp	Time	Speed	Water	Material	Fragrance	Quality Index
March	40	20	300	75	15	10	5
April	39	20	300	**	15	15	7
May	40	18	200	80	15	5	2
June	40	20	250	80	-15	5	4
January	40	**	200	75	15	10	5
February	40	20	2000	80	15	5	3

Data Preprocessing

Real-world data are often **incomplete, inconsistent, noisy, and unstructured**. Before performing analysis or building machine learning models, the data must be transformed into a clean and usable form. This process is known as **data preprocessing**.

Definition:

Data preprocessing is the process of converting raw data into a clean, consistent, and structured format suitable for analysis and modeling.

Why preprocessing is necessary ?

- Improve data quality.
- Handle missing and erroneous values.
- Remove inconsistencies and duplicates.
- Reduce noise and outliers.

- Enhance model accuracy.
- Reduce computational complexity.
- Make data suitable for visualization and statistical analysis.

Problem with missing values

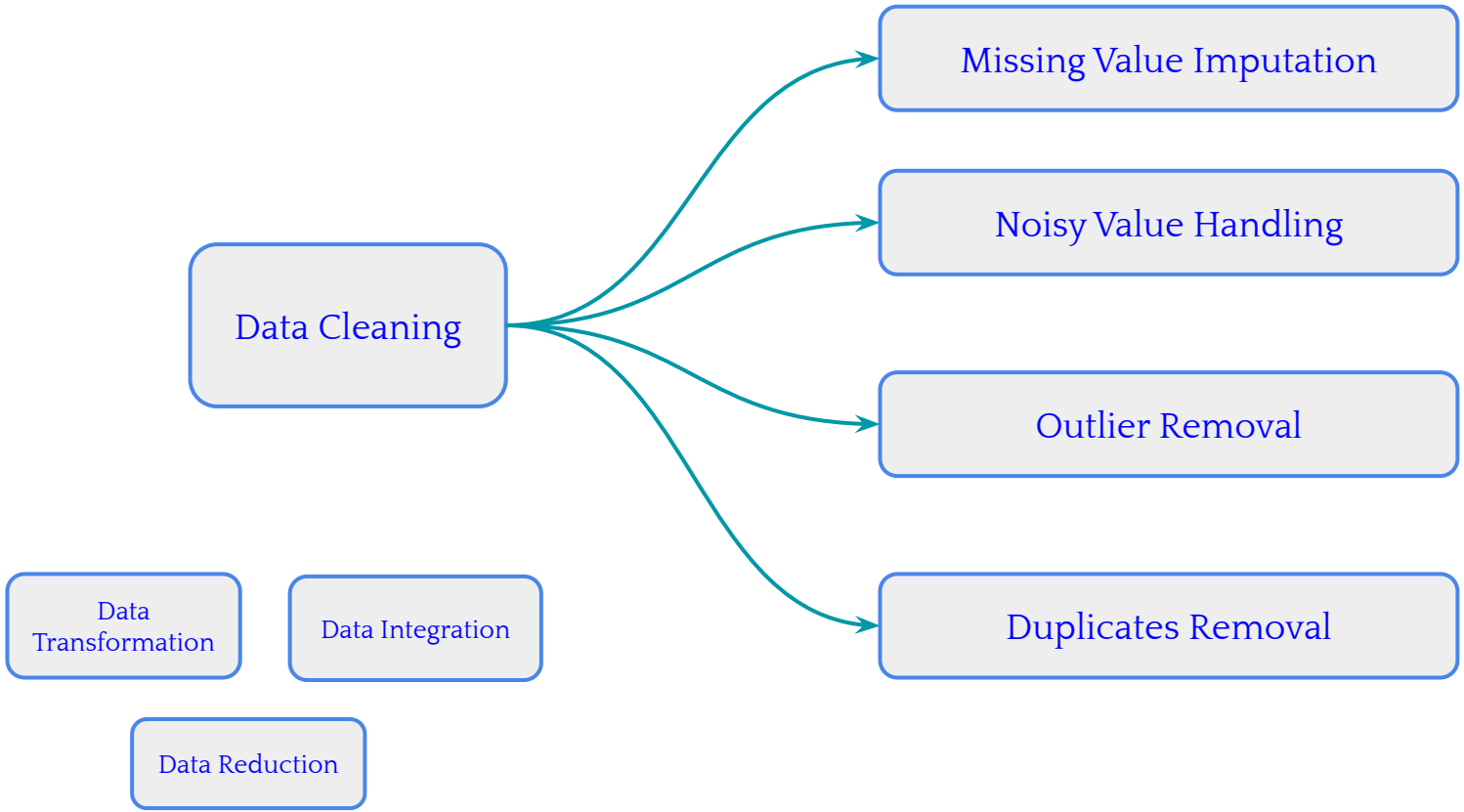
Customer ID	Age	Income	City
101	25	50000	Delhi
102	?	60000	Mumbai
103	30	NA	Chennai
103	30	NA	Chennai

What is the average age ?

Types of Preprocessing

- Data Cleaning
- Data Transformation
- Data Integration
- Data Reduction

Types of Data Cleaning



Missing Value Imputation

Example:

Student	Marks
A	85
B	?
C	92

Methods for Handling Missing Values

1. Delete records containing missing values.
2. Replace with mean value.
3. Replace with median value.
4. Replace with mode.
5. Use interpolation.
6. Predict missing values using machine learning

Noisy Data Handling

Example:

```
10, 12, 13, 11, 120, 14
```

Techniques for Noise Removal

- Binning
- Regression
- Moving averages
- Smoothing techniques

Outlier Removal

Example:

Salary (₹)
35,000
38,000
40,000
42,000
3,50,000

Detection Methods

- Box plots
- Z-score method
- Interquartile Range (IQR)

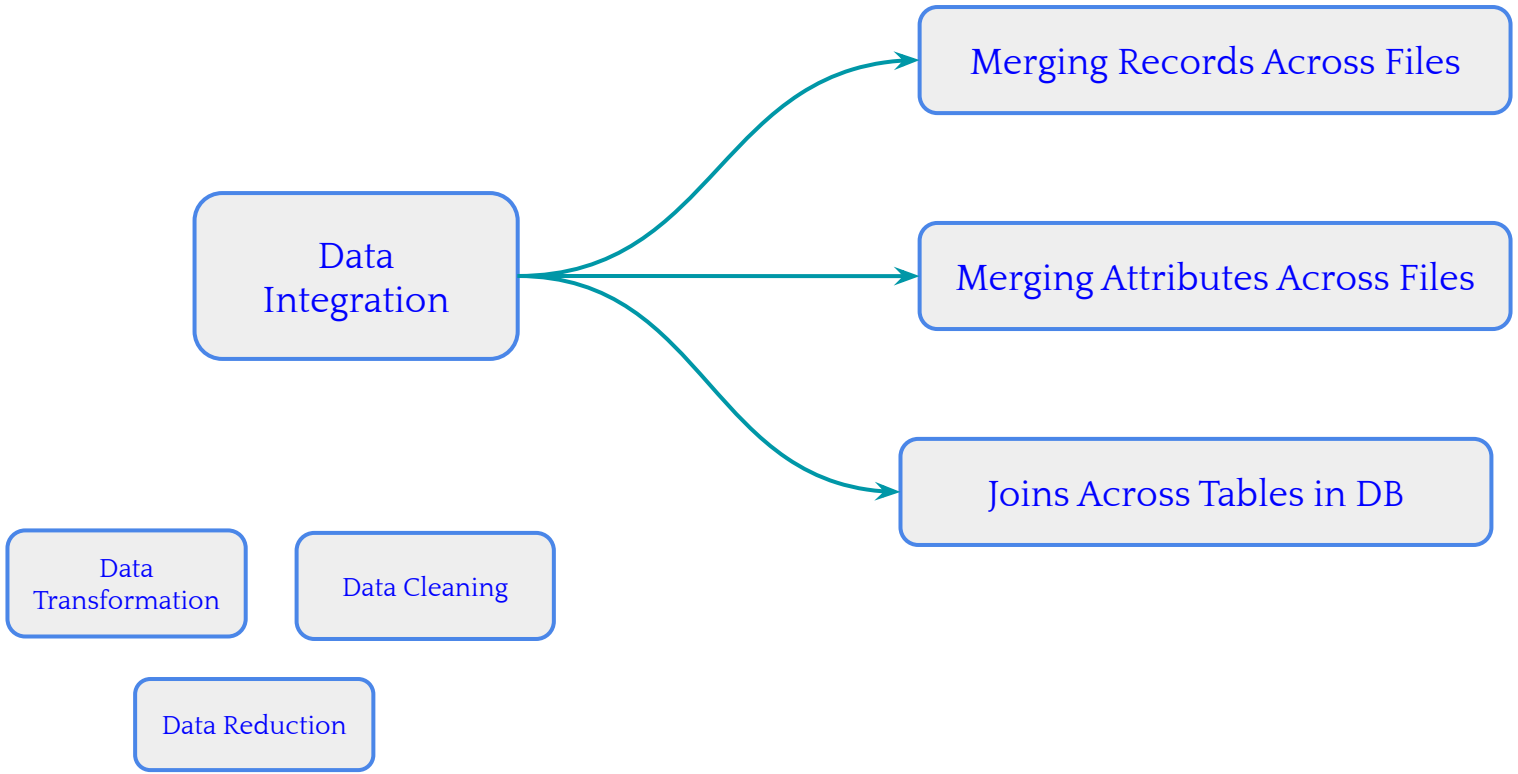
Duplicate Removal

Example:

ID
101
102
101

Identify duplicate entries and remove

Types of Data Integration



Data Integration: Join across tables in DB

Example

Customer Database

Customer ID	Name
101	Rahul

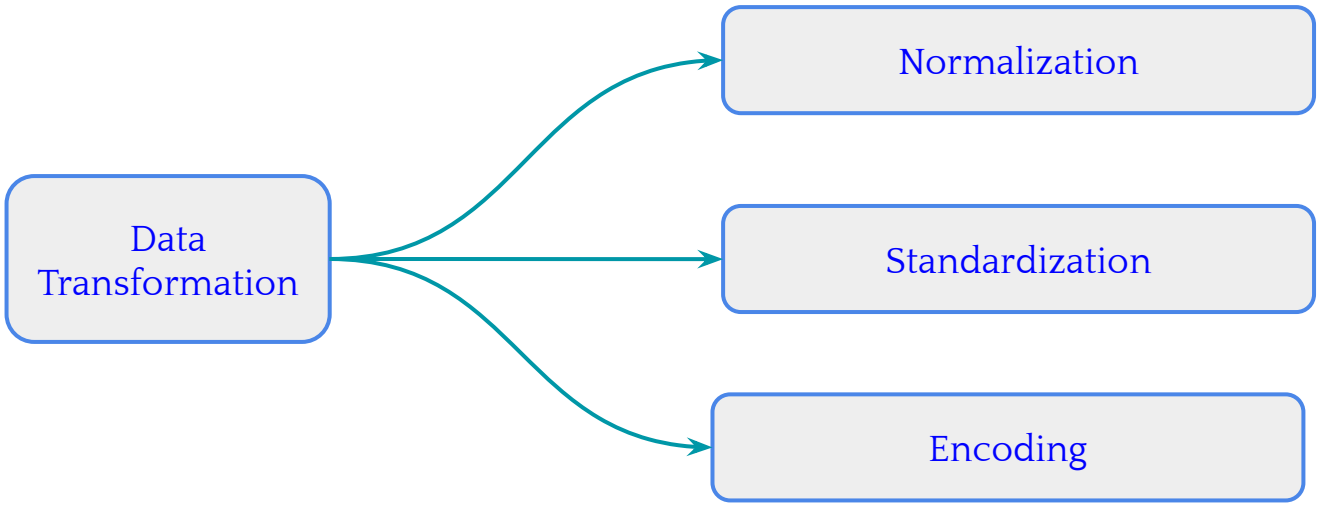
Purchase Database

Customer ID	Amount
101	5000

After integration:

Customer ID	Name	Amount
101	Rahul	5000

Types of Data Transformation



Data Integration

Data Cleaning

Data Reduction

Normalization

Car	Weight (kg)	Price (₹)
A	900	600,000
B	1200	850,000
C	1500	1,200,000

Ranges of the numerical values are very different across columns.

This creates problem in many methods.

Normalization will rectify this.

Normalization

Car	Weight (kg)	Price (₹)
A	900	600,000
B	1200	850,000
C	1500	1,200,000
D	1800	1,500,000

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

- x = original value
- $x_{\min}$ = minimum value
- $x_{\max}$ = maximum value

$x_{\min} = 900$ $x_{\max} = 1800$

$$\frac{900 - 900}{1800 - 900} = 0$$

$$\frac{1200 - 900}{1800 - 900} = \frac{300}{900} = 0.33$$

$$\frac{1500 - 900}{900} = \frac{600}{900} = 0.67$$

$$\frac{1800 - 900}{900} = 1$$

Normalization

Car	Weight (kg)	Price (₹)
A	900	600,000
B	1200	850,000
C	1500	1,200,000
D	1800	1,500,000

Car	Normalized Weight	Normalized Price
A	0	0
B	0.33	0.28
C	0.67	0.67
D	1	1

$$x_{\min} = 900 \qquad x_{\max} = 1800$$

$$\frac{900 - 900}{1800 - 900} = 0 \qquad \frac{1200 - 900}{1800 - 900} = \frac{300}{900} = 0.33$$

$$\frac{1500 - 900}{900} = \frac{600}{900} = 0.67 \qquad \frac{1800 - 900}{900} = 1$$

Standardization

$$z = \frac{x - \mu}{\sigma}$$

- x = original value
- μ = mean
- σ = standard deviation

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

- x = original value
- $x_{\min}$ = minimum value
- $x_{\max}$ = maximum value

Standardization

$$z = \frac{x - \mu}{\sigma}$$

- x = original value
- μ = mean
- σ = standard deviation

Car	Weight (kg)	Price (₹)
A	900	600,000
B	1200	850,000
C	1500	1,200,000
D	1800	1,500,000

$$\mu = \frac{900 + 1200 + 1500 + 1800}{4} = \frac{5400}{4} = 1350 \text{ kg}$$

Weight (x)	$x - \mu$	$(x - \mu)^2$
900	-450	202,500
1200	-150	22,500
1500	150	22,500
1800	450	202,500

$$202500 + 22500 + 22500 + 202500 = 450000$$

$$\sigma = \sqrt{\frac{450000}{4}} = \sqrt{112500} \approx 335.41$$

Standardization

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{900 - 1350}{335.41} = \frac{-450}{335.41} \approx -1.34$$

$$z = \frac{1200 - 1350}{335.41} = \frac{-150}{335.41} \approx -0.45$$

$$z = \frac{1500 - 1350}{335.41} = \frac{150}{335.41} \approx 0.45$$

$$z = \frac{1800 - 1350}{335.41} = \frac{450}{335.41} \approx 1.34$$

Car	Weight (kg)	Price (₹)
A	900	600,000
B	1200	850,000
C	1500	1,200,000
D	1800	1,500,000

Car	Weight (kg)	Normalized Weight	Standardized Weight
A	900	0	-1.34
B	1200	0.33	-0.45
C	1500	0.67	0.45
D	1800	1	1.34

Encoding

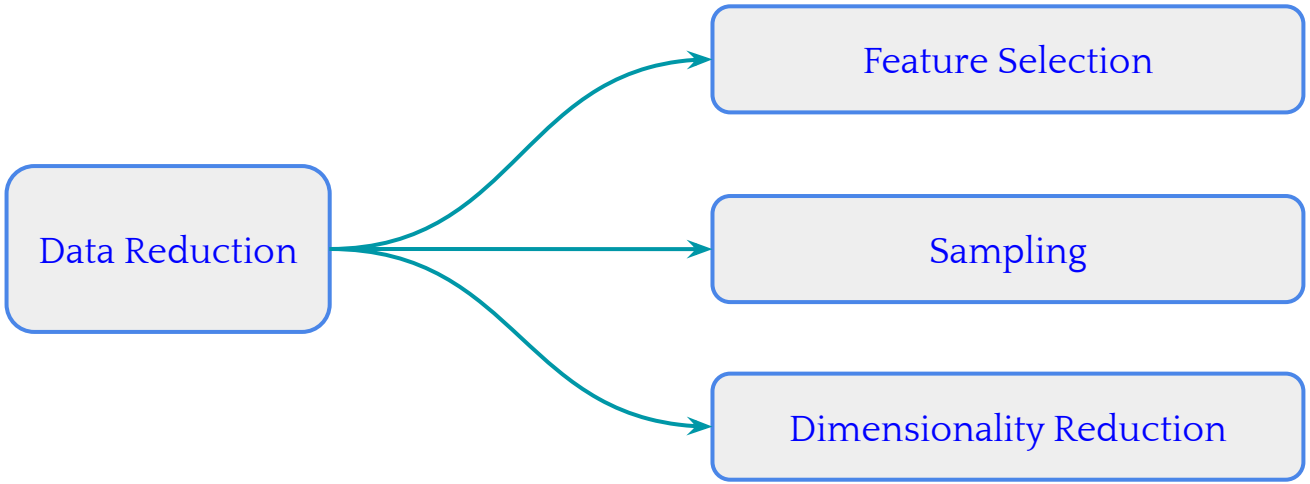
Label Encoding

Color	Color
Red	0
Blue	1
Green	2

One-Hot Encoding

Red	Blue	Green
1	0	0
0	1	0
0	0	1

Types of Data Reduction



Data Integration

Data Cleaning

Data Transformation

Feature Selection

Feature selection is the process of selecting the most relevant features (variables or attributes) from a dataset while removing irrelevant, redundant, or less useful features.

The objective is to build a model using only the features that contribute significantly to the prediction.

Example

Car	Engine (cc)	Mileage (km)	Age (Years)	Color	Price (₹)
A	1200	20,000	2	Red	8,50,000
B	1500	40,000	5	Blue	6,50,000
C	1800	10,000	1	White	12,00,000

Columns are features: Engine, Mileage, Age, Color

Price is the target

Question is are all features required to predict the price ?

Example

Car	Engine (cc)	Mileage (km)	Age (Years)	Color	Price (₹)
A	1200	20,000	2	Red	8,50,000
B	1500	40,000	5	Blue	6,50,000
C	1800	10,000	1	White	12,00,000

Columns are features: Engine, Mileage, Age, Color

Price is the target

Question is are all features required to predict the price ?

Feature selection may remove Color

Benefits of Feature Selection

- Reduces computation time
- Simplifies the model
- Makes results easier to interpret
- Reduces storage requirements

Sampling for Data Reduction

Sampling is a data reduction technique in which **a small subset of data is selected from a large dataset** to represent the characteristics of the entire dataset.

Instead of analyzing millions of records, analysts study a representative sample, which reduces computation time and storage requirements while still providing meaningful insights.

Can avoid:

- Large storage
- High computational power
- Long processing time

Sampling Applications

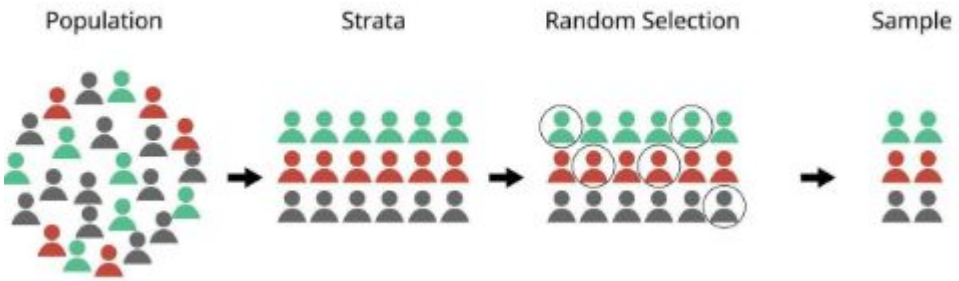
- Opinion polls
- Medical research
- Manufacturing quality inspection
- Customer satisfaction surveys
- Machine learning model training
- Big data analytics

Sampling Methods

Random Sampling

Stratified Sampling

Reservoir Sampling

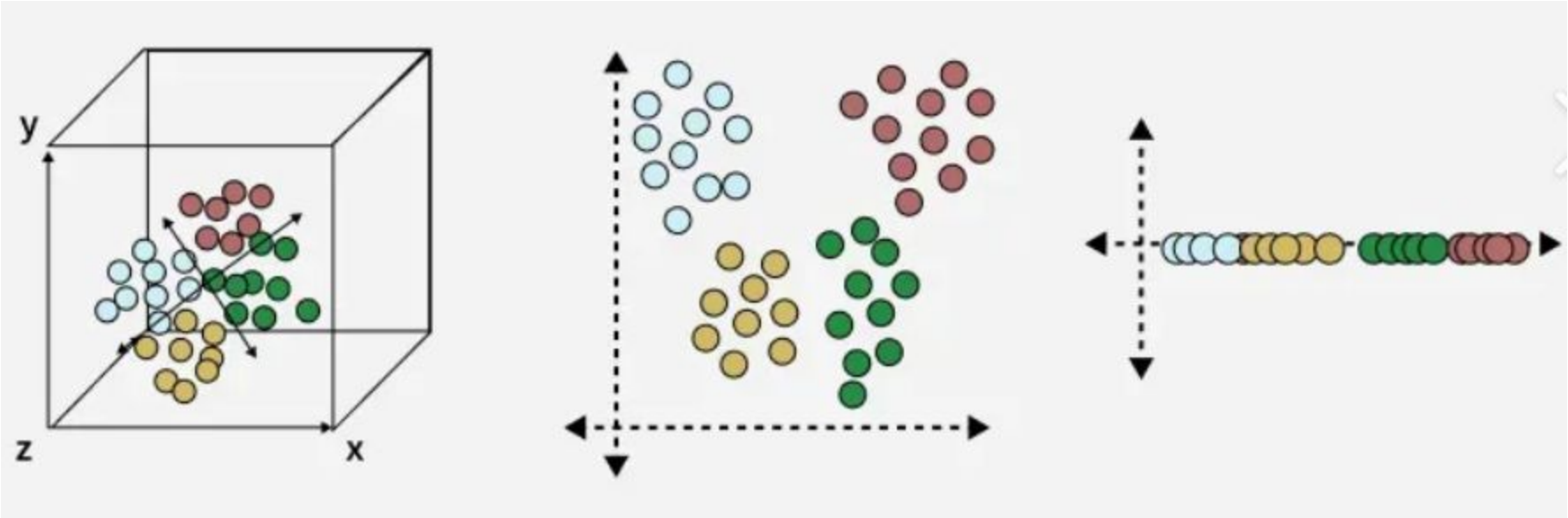


Dimensionality Reduction

Dimensionality reduction is a data reduction technique that reduces the number of features (variables or dimensions) in a dataset while preserving as much useful information as possible.

Instead of working with hundreds or thousands of features, we transform the data into a smaller set of informative features.

Reduce Dimension



Quiz-3: Part A

Raw Data

Temperature	Speed	Defect
80	120	No
?	115	No
95	140	Yes
95	140	Yes

How will you preprocess this dataset ?

Quiz-3: Part B

1. What are preprocessing steps ?
2. Which is the most important step ?

3. Write 2 things that you liked/understood today.
4. Write 2 things that you did not like/understand today.

5. Consider the case study you picked in the last session.
    a. Revise the case study based if required
    b. Identify the preprocessing steps
    c. Describe the descriptive, diagnostic, predictive, and prescriptive analytics