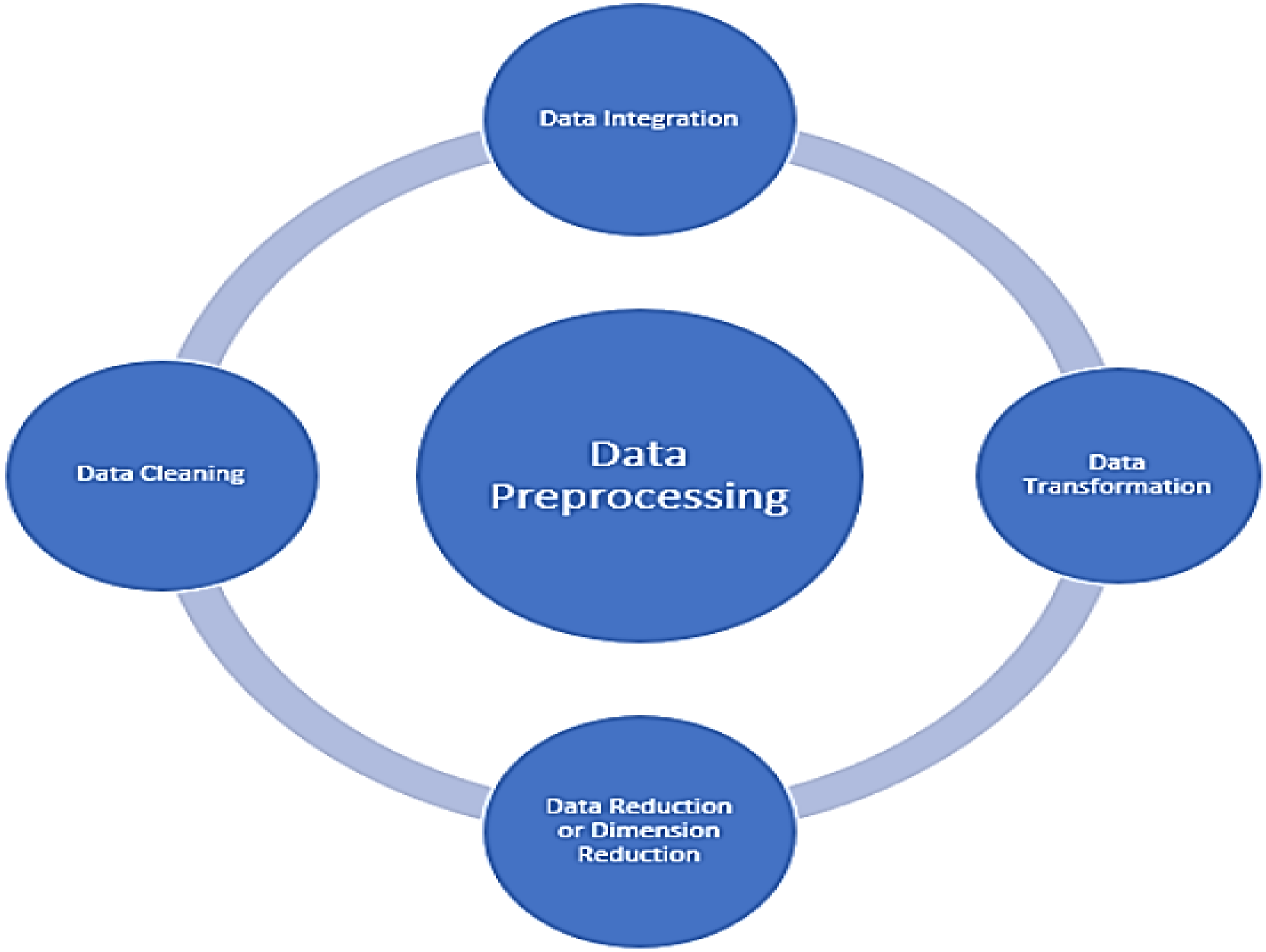
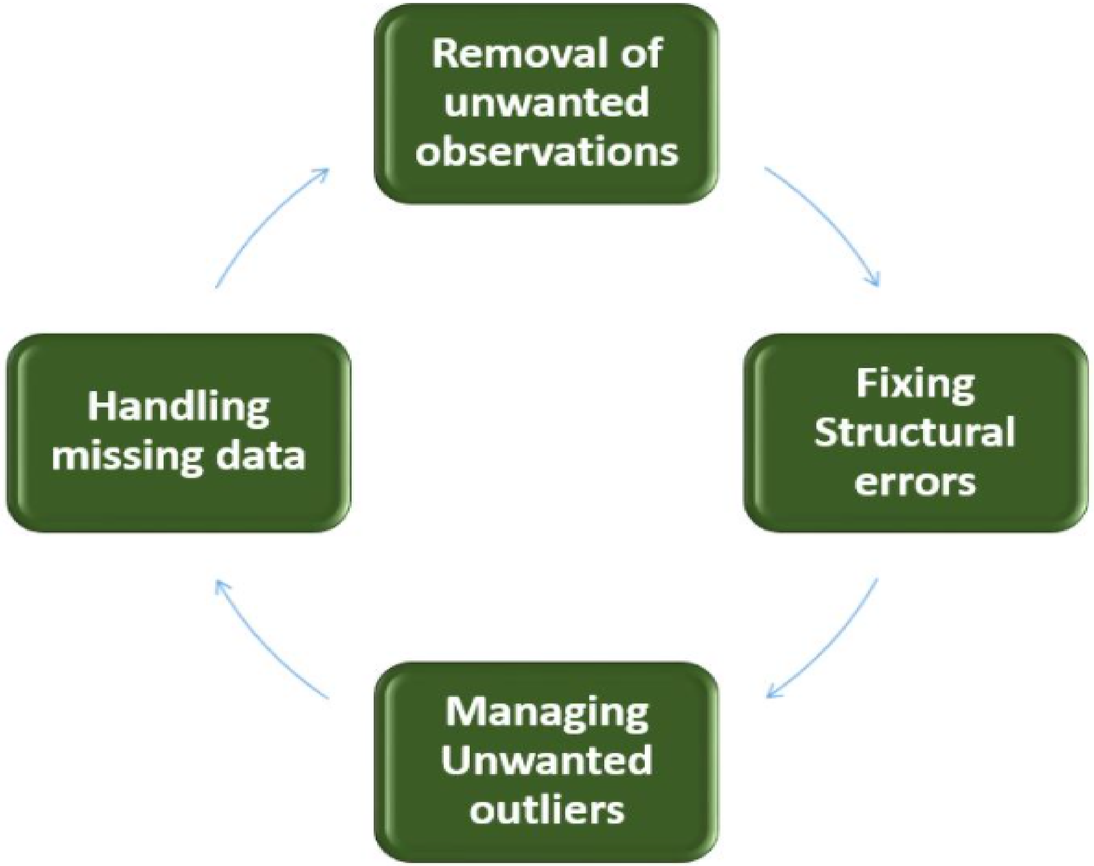
Data Preprocessing

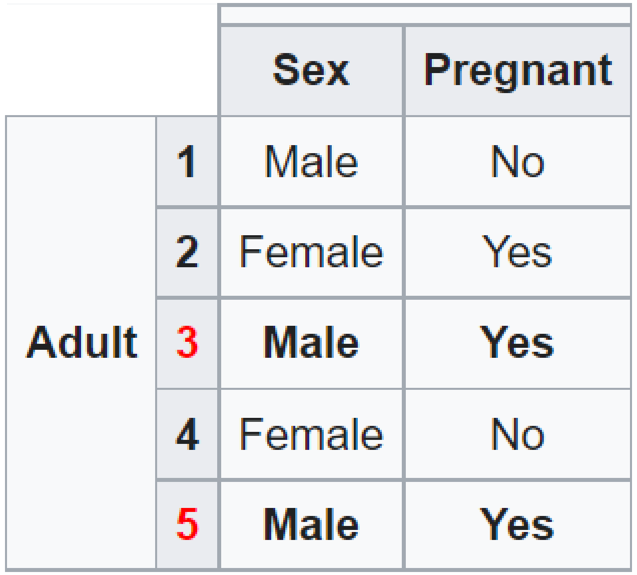


Data Cleaning

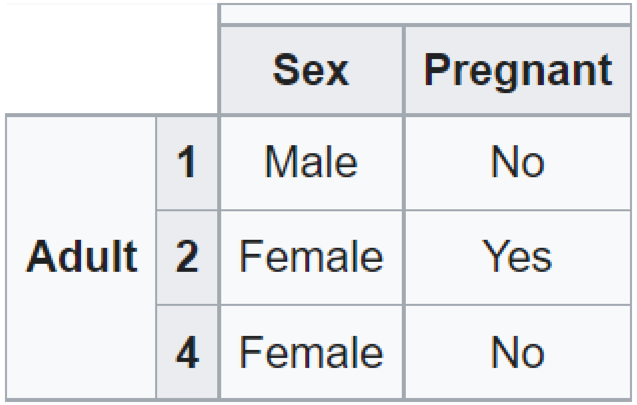


Example -Data cleaning

- We can perform a Data cleaning and choose to delete such data from our table.
- The impossible data will affect the calculation or data manipulation process .

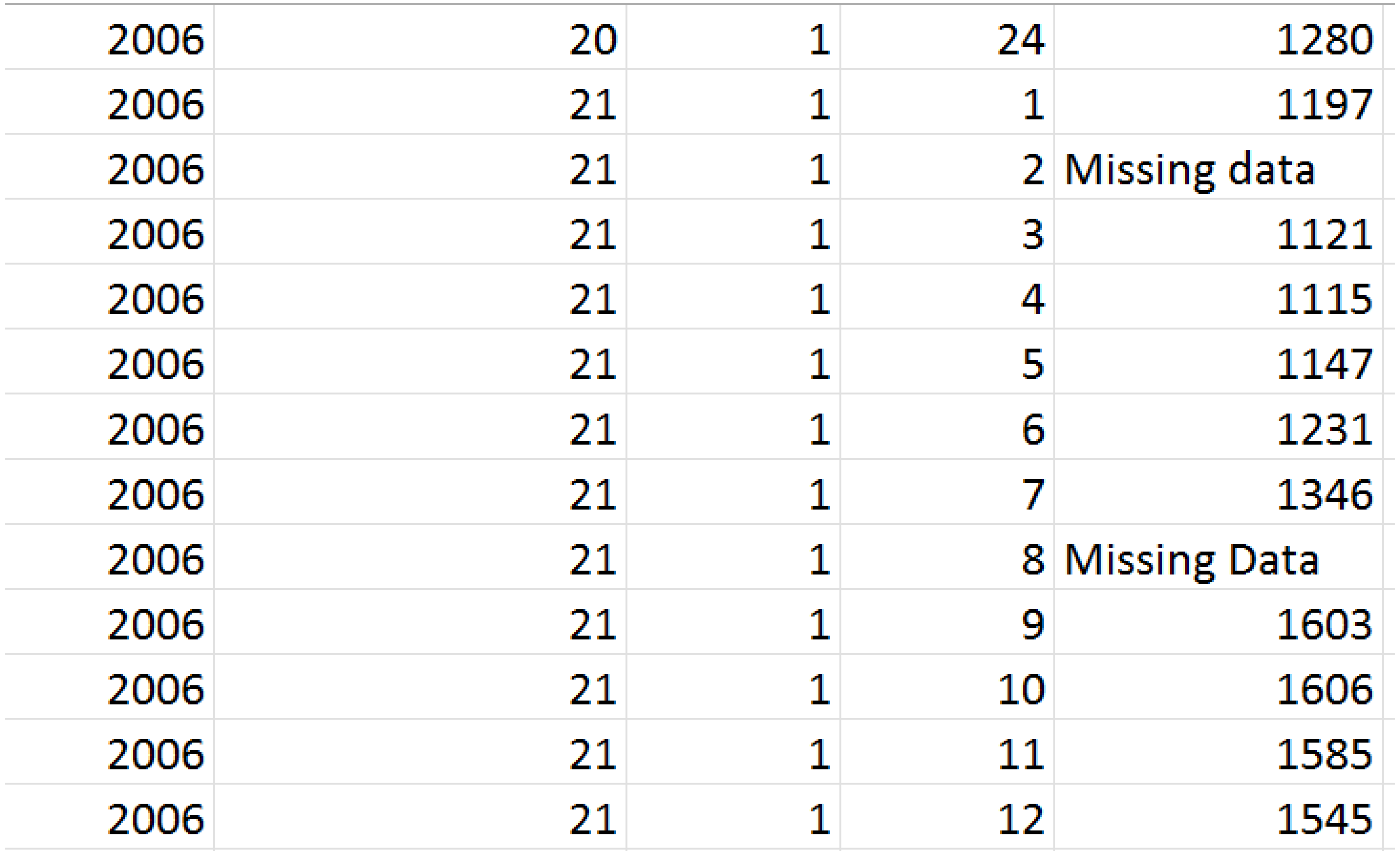


		Sex	Pregnant
Adult	1	Male	No
	2	Female	Yes
	3	**Male**	**Yes**
	4	Female	No
	5	**Male**	**Yes**



		Sex	Pregnant
Adult	1	Male	No
	2	Female	Yes
	4	Female	No

Missing Data



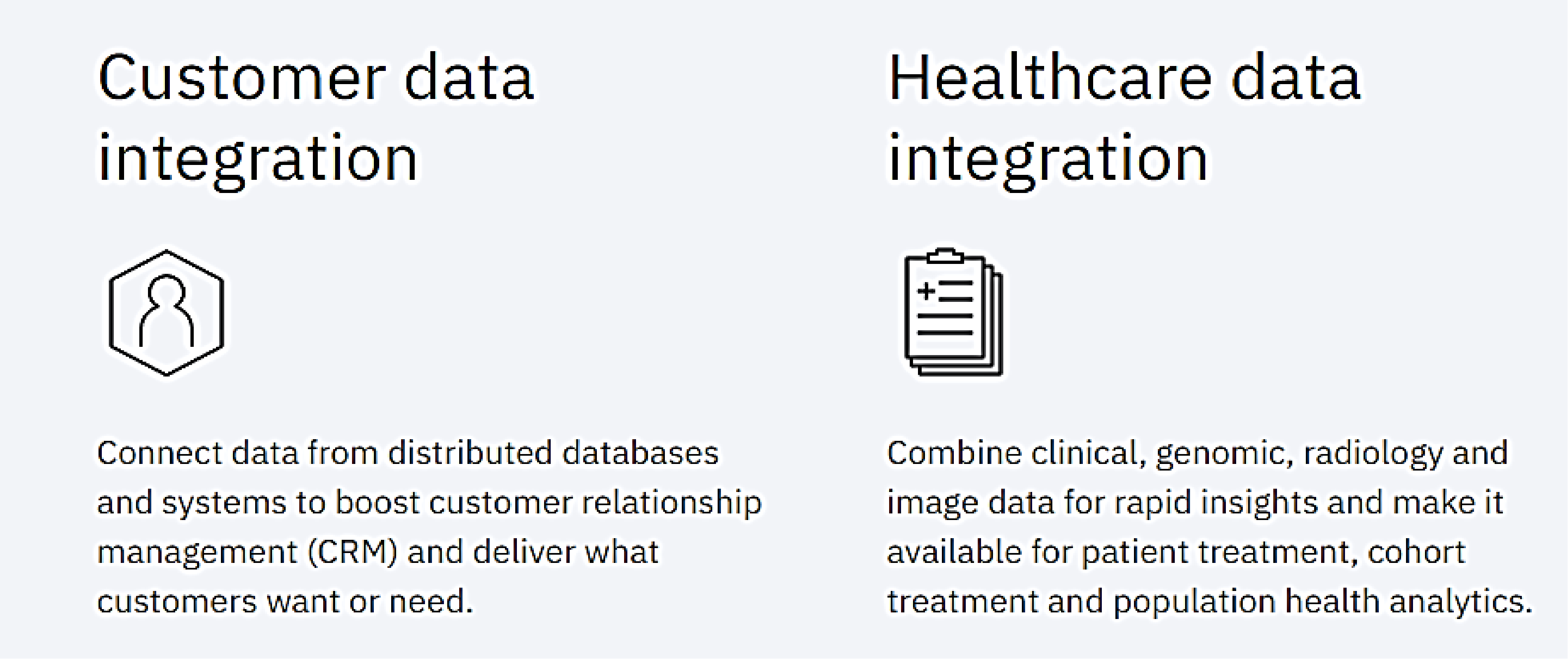
2006	20	1	24	1280
2006	21	1	1	1197
2006	21	1	2	Missing data
2006	21	1	3	1121
2006	21	1	4	1115
2006	21	1	5	1147
2006	21	1	6	1231
2006	21	1	7	1346
2006	21	1	8	Missing Data
2006	21	1	9	1603
2006	21	1	10	1606
2006	21	1	11	1585
2006	21	1	12	1545

Interpolation/Optimization

- The popular INEDI (Improved New Edge Directional Interpolation) method is used .
- The edge directed interpolation algorithm estimates the local **covariance coefficients** from low-resolution images and then these are used to adapt the interpolation at a higher resolution based on geometric duality between LR covariance and HR covariance.

Data Integration

- **Data integration** involves combining data residing in different sources .



Customer data integration

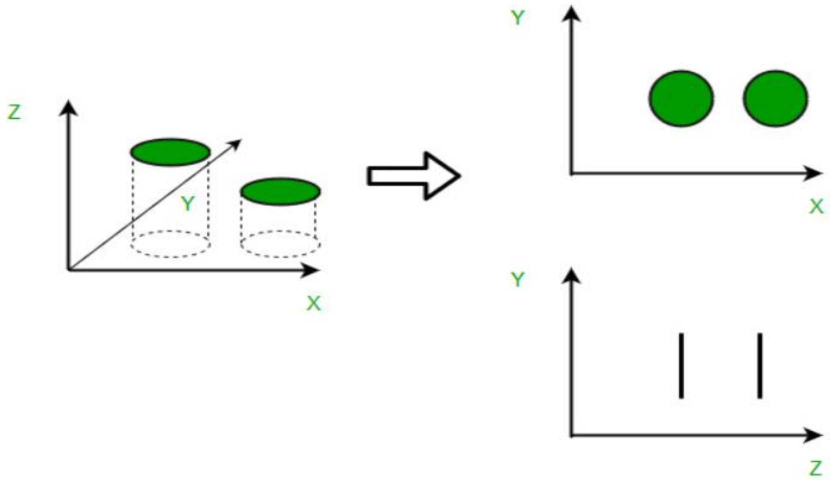
Connect data from distributed databases and systems to boost customer relationship management (CRM) and deliver what customers want or need.

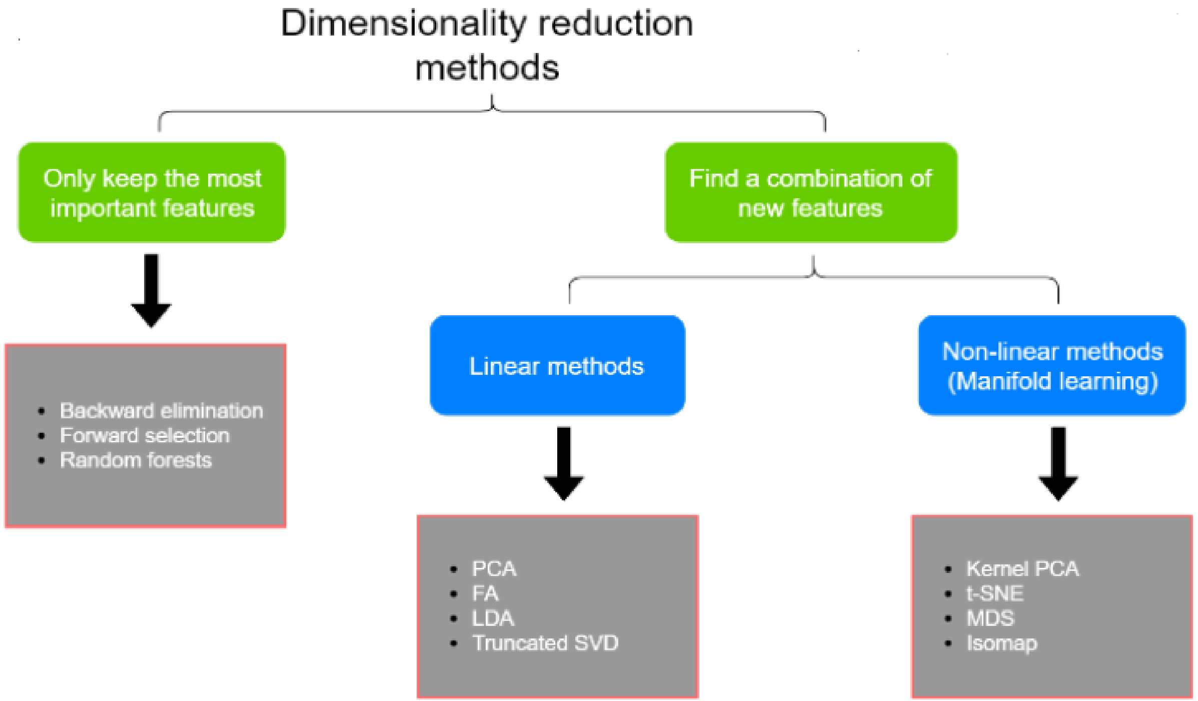
Healthcare data integration

Combine clinical, genomic, radiology and image data for rapid insights and make it available for patient treatment, cohort treatment and population health analytics.

Data transformation and **Dimension Reduction**

- **Data transformation** is the process of converting data from one format or structure into another format or structure.(Finding Max and Min,Rounding…)
- **Dimension Reduction:**





Dimensionality reduction methods

Only keep the most important features

- Backward elimination
- Forward selection
- Random forests

Find a combination of new features

Linear methods

- PCA
- FA
- LDA
- Truncated SVD

Non-linear methods (Manifold learning)

- Kernel PCA
- t-SNE
- MDS
- Isomap

Data Preprocessing

Reading the Dataset

```
# Import the libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
# Import the dataset
 dataset = pd.read_csv('/content/Salary_Data.csv’)
 X = dataset.iloc[:, :-1].values
 y = dataset.iloc[:, 1].values
```

Handling Missing Data

- To handle missing data,typical methods include imputation(Replace with one and deletion(ignores missing values)
- By integer-location based indexing(iloc),we split data into inputs and outputs, where the former takes the first two columns while the latter only keeps the last column.

Data Preprocessing….

- For numerical values in inputs that are missing,we replace the“None”entries with the mean value of the same column.
- inputs, outputs = data.iloc[:, 0:2], data.iloc[:, 2]
- inputs = inputs.fillna(inputs.mean())
- print(inputs)
- A=[1 2 3]
- [4   none 6]
- *Please refer text book-1 page 51 Example*

Practice questions

- Describe preprocessing and explain how do we handle Missing data
- W.A.P to handle missing data in an array.