510 DATA SCIENCE

Lecture 05

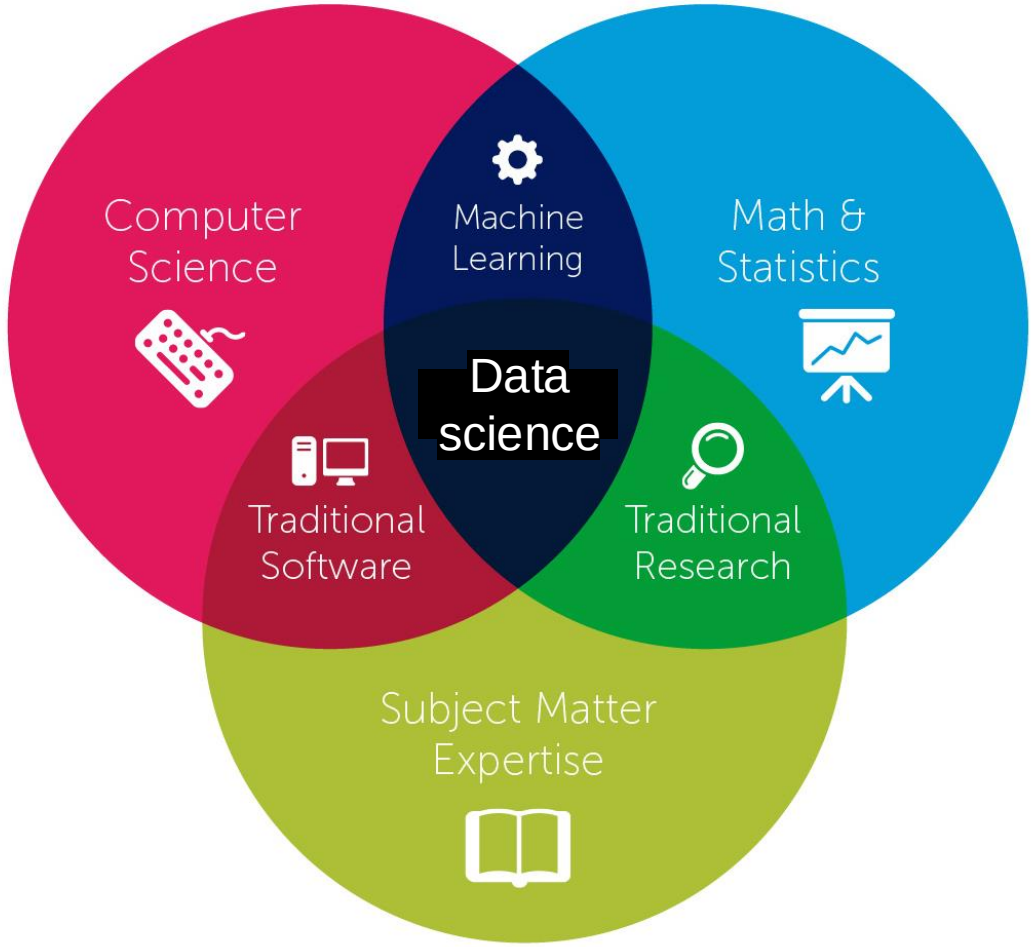
Fall 2024

Instructor: Assoc. Prof. Şener Özönder

Email: sener.ozonder@bogazici.edu.tr

Institute for Data Science & Artificial Intelligence

Boğaziçi University



Data Imputation

- Data imputation refers to the process of replacing missing or invalid data with substituted values. Proper data imputation can improve the quality of the data and lead to better insights and models.

1. Mean Imputation: Replace the missing value with the mean of the available data.

- Usage: Suitable when the data is normally distributed. It's simple and retains the dataset's overall mean.

- Drawback: It doesn't preserve the variability in the data, leading to an underestimation of variance.

2. Median Imputation: Replace the missing value with the median (middle value) of the available data.

- Usage: Useful for skewed data or data with outliers because the median is robust to such anomalies.

Data Imputation

3. Mode Imputation: Replace the missing value with the mode (most frequent value) of the available data.

- Usage: Recommended for categorical data.

4. Random Imputation from a distribution: Fill missing values randomly if the underlying distribution is known or can be guessed.

- Usage: When you want to preserve the data distribution but don't want to introduce potential biases with mean or median imputation.

5. K-Nearest Neighbors (KNN) Imputation: For a missing value, find the 'k' training samples closest in distance and estimate the missing value based on their values.

- Usage: When the data has underlying patterns that can be captured with KNN. It's more computationally intensive than mean or median imputation.

6. Interpolation: For ordered data (like time series), fill missing values by interpolating between valid data points.

- Usage: Time-series data or data with a clear order.

Data Imputation

7. Model-Based Imputation: Model the records with complete cases to predict the missing values in incomplete records.

- Usage: When there's a clear relationship between the missing variable and other variables.

8. Multiple Imputation: Instead of filling each missing value once, multiple imputations generate multiple datasets with different imputed values, reflecting the uncertainty of imputation. Analysis is then conducted on each dataset separately, and results are aggregated.

9. Forward Fill & Backward Fill: Used mainly in time series data. Forward fill takes the previous value and fills the next missing value. Backward fill takes the next valid value and fills the previous missing value.

10. Constant Value Imputation: All missing values are replaced by a constant. This is more meaningful when the missingness has a specific significance.