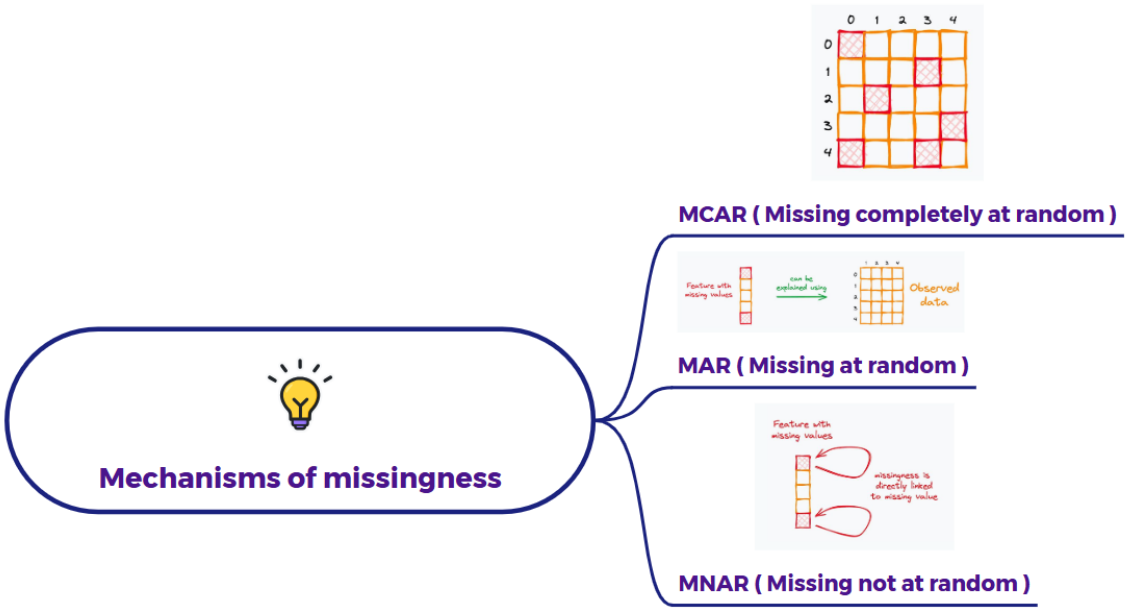
Different Mechanisms of Missingness



Let's break down the three main mechanisms of missingness in data science: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR), with clear examples for each.

1. Missing Completely At Random (MCAR)

- **Definition:** Missing Completely At Random means that the probability of a data point being missing is **independent** of both the observed values and the missing values themselves. There's no systematic reason for the data to be missing; it's purely due to chance.
- **Key Characteristic:** The missingness is unrelated to any other variable in the dataset.
- **Example:**

Imagine you're conducting a survey, and some of the paper questionnaires were accidentally dropped in a puddle, making a few random responses illegible. The missing answers on those specific questionnaires are MCAR because the fact that the data is missing has nothing to do with what the respondents actually answered (the missing value itself) or any other information you collected (like their age, gender, or income). The missingness is solely due to the random event of the questionnaires getting wet.

- **Dataset:** Survey responses with columns like Age, Income, Favorite Color, and Response to Question X.
- **Missingness:** Some values in Response to Question X are missing because those specific survey sheets were damaged.
- **Why MCAR:** Whether a particular Response to Question X is missing is not related to the respondent's age, income, favorite color, or what their actual response to Question X was. It's a random occurrence affecting a subset of the surveys.

2. Missing At Random (MAR)

- **Definition:** Missing At Random means that the probability of a data point being missing **depends on the observed values** but is **not related to the missing value itself**. In other words, we can predict the pattern of missingness based on other variables in the dataset.
- **Key Characteristic:** The missingness is systematic but can be explained by other observed variables.
- **Example:**

Consider a health survey where people are asked about their weight and income. Suppose men are less likely to report their weight than women. The missingness in the Weight variable is MAR because the probability of Weight being missing is related to the observed Gender of the respondent. However, within the male group, the missingness of weight is not related to their actual weight.

- **Dataset:** Health survey data with columns like Age, Gender, Income, and Weight.
- **Missingness:** Some values in the Weight column are missing, and the proportion of missing weights is higher for males than for females.
- **Why MAR:** The missingness of Weight is related to the observed Gender. We can use the information in the Gender column to understand the pattern of missingness. However, the fact that a specific male's weight is missing doesn't depend on what his actual weight is.

3. Missing Not At Random (MNAR)

- **Definition:** Missing Not at Random means that the probability of a data point being missing **is related to the missing value itself**. There's a systematic reason for the missingness that is inherent to the unobserved data.
- **Key Characteristic:** The missingness cannot be fully explained by other observed variables in the dataset.
- **Example:**

Think about a survey asking about drug use. Individuals who have used illegal drugs might be less likely to answer the question truthfully, leading to a higher rate of missing values for the Drug Use variable among those who actually engaged in drug use. The missingness here is directly related to the unobserved information (whether they used drugs). We cannot fully predict who didn't answer based on other observed variables like age or income.

  - **Dataset:** Survey data with columns like Age, Income, Education Level, and Drug Use.
  - **Missingness:** Some values in the Drug Use column are missing, and it's suspected that those who did use drugs are more likely to leave this question unanswered.
  - **Why MNAR:** The missingness of Drug Use is likely related to the actual (missing) value of Drug Use. People who used drugs are more inclined to not report it. We cannot fully explain this missingness using other observed variables like age or income.

Why is understanding the mechanism important?

The mechanism of missingness has significant implications for how you should handle missing data:

- **MCAR:** Statistical analyses performed on MCAR data are generally unbiased, but you still lose statistical power due to the reduced sample size. Simple methods like complete case analysis (removing rows with any missing values) might be acceptable, though not always efficient.

- **MAR:** More sophisticated imputation techniques that leverage the observed variables to predict the missing values are often used for MAR data. Complete case analysis can introduce bias under MAR.
- **MNAR:** Handling MNAR data is the most challenging. Simple imputation methods can lead to biased results. Addressing MNAR often requires more advanced statistical modeling or incorporating external information about the missingness process. If the MNAR mechanism is not properly addressed, any subsequent analysis or modeling can be severely biased.