Categories of Missing Data

The categories of "missingness" can have important implications for statistical bias and power. The three categories of missing data are:

Missing completely at random or MCAR:

- When data are MCAR, missing cases are, on average, identical to non-missing cases, with respect to the feature matrix. Complete case analysis will reduce the power of the analysis, but will not affect bias.

Missing at random or MAR:

- When data are MAR the missing data often have some dependence on measured values, and models can be used to help impute what the likely data would be. For example, in a Major League Baseball (MLB) survey, there might be a disproportionate number of male respondents completing all of the questions as compared to females, since males comprise most of the MLB's viewership.

Missing not at random or MNAR:

- In this case the missing data depend on unmeasured or unknown variables. There is no information available to account for the missingness.

Of these three categories, the best case scenario is that the data are MCAR. It should be noted that imputing values under the other two types of missingness can result in an increase in bias. Many strategies have been proposed for the management and replacement of missing values—a process known as **imputation**. In this unit we will illustrate the method of simple imputation, as well as two more sophisticated strategies: Multiple imputation and Bayesian imputation. In the case study that follows we will also exemplify the iterative process for deciding which strategy is best.