A CLOUD GURU

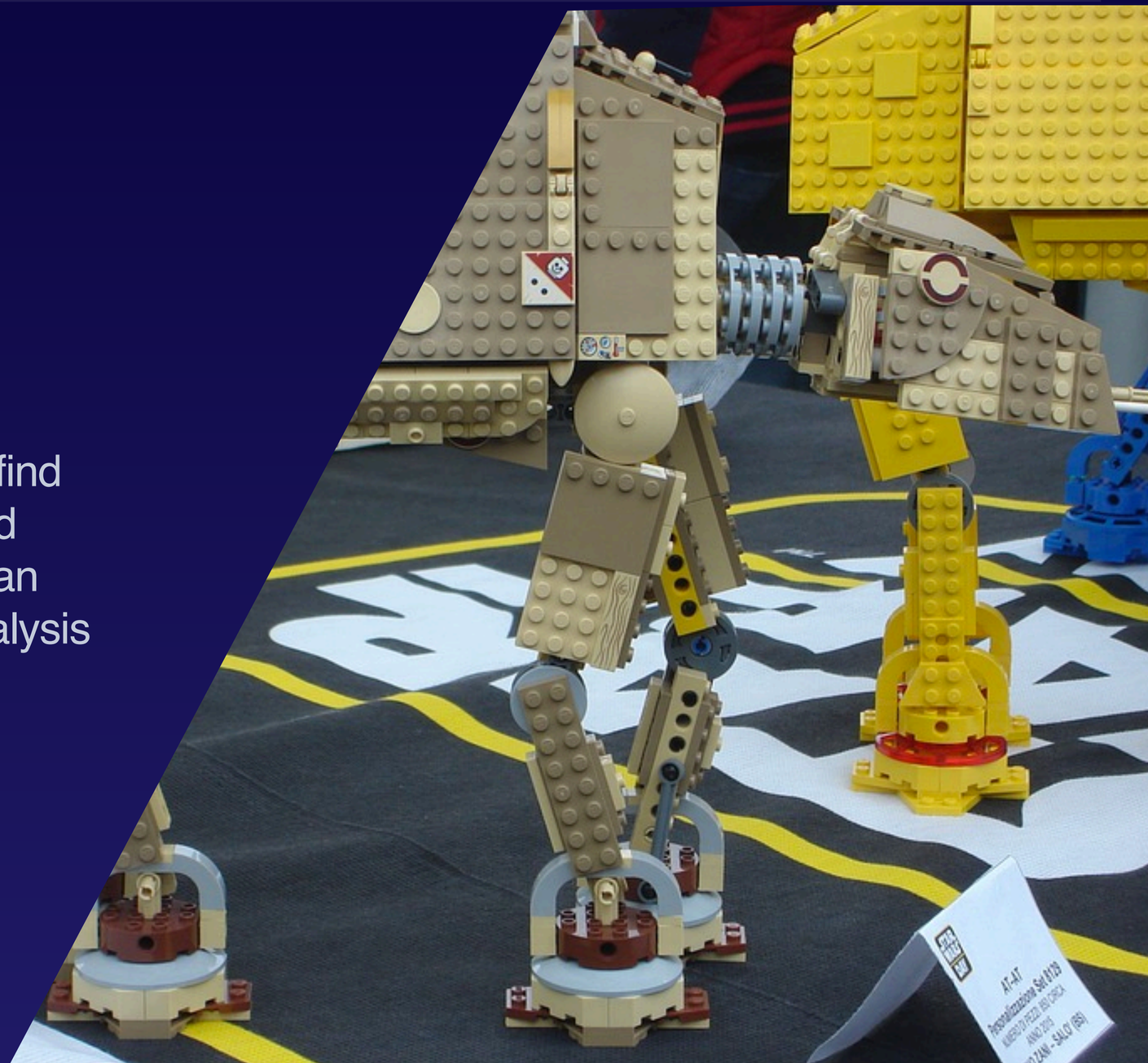
Handling Missing Values

Brock Tubre
INSTRUCTOR

Missing Data

Missing Lego Pieces

Imagine opening up a set of legos only to find several pieces missing. It's never a desired situation. Missing values in your dataset can be just as frustrating and interfere with analysis and your models prediction.



Handling Missing Values

Missing Data

Missing data can be represented in many different ways (*null, NaN, NA, None, etc.)* and handling missing values is an important data preparation step.

Why are values missing?

Why are values missing in the first place?…

Missingness Mechanisms

1

Missing at Random (MAR)

Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data.

2

Missing Completely at Random (MCAR)

The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.

3

Missing not at Random (MNAR)

Two possible reasons are that the missing value depends on the hypothetical value or missing value is dependent on some other variable's value.

How to handle missing values

Technique	Why this works	Ease of Use
Supervised learning	Predicts missing values based on the values of other features	Most difficult, can yield best results
Mean	The average value	Quick and easy, results can vary
Median	Orders values then choses value in the middle	Quick and easy, results can vary
Mode	Most common value	Quick and easy, results can vary
Dropping rows	Removes missing values	Easiest but can dramatically change datasets

How to handle missing values

Technique	Why this works	Ease of Use
Supervised learning	Predicts missing values based on the values of other features	Most difficult, can yield best results
Mean	The average value	Quick and easy, results can vary
Median	Orders values then choses value in the middle	Quick and easy, results can vary
Mode	Most common value	Quick and easy, results can vary
Dropping rows	Removes missing values	Easiest but can dramatically change datasets

Replacing data is known as imputation