A CLOUD GURU

Data Preparation Concepts

Brock Tubre
INSTRUCTOR

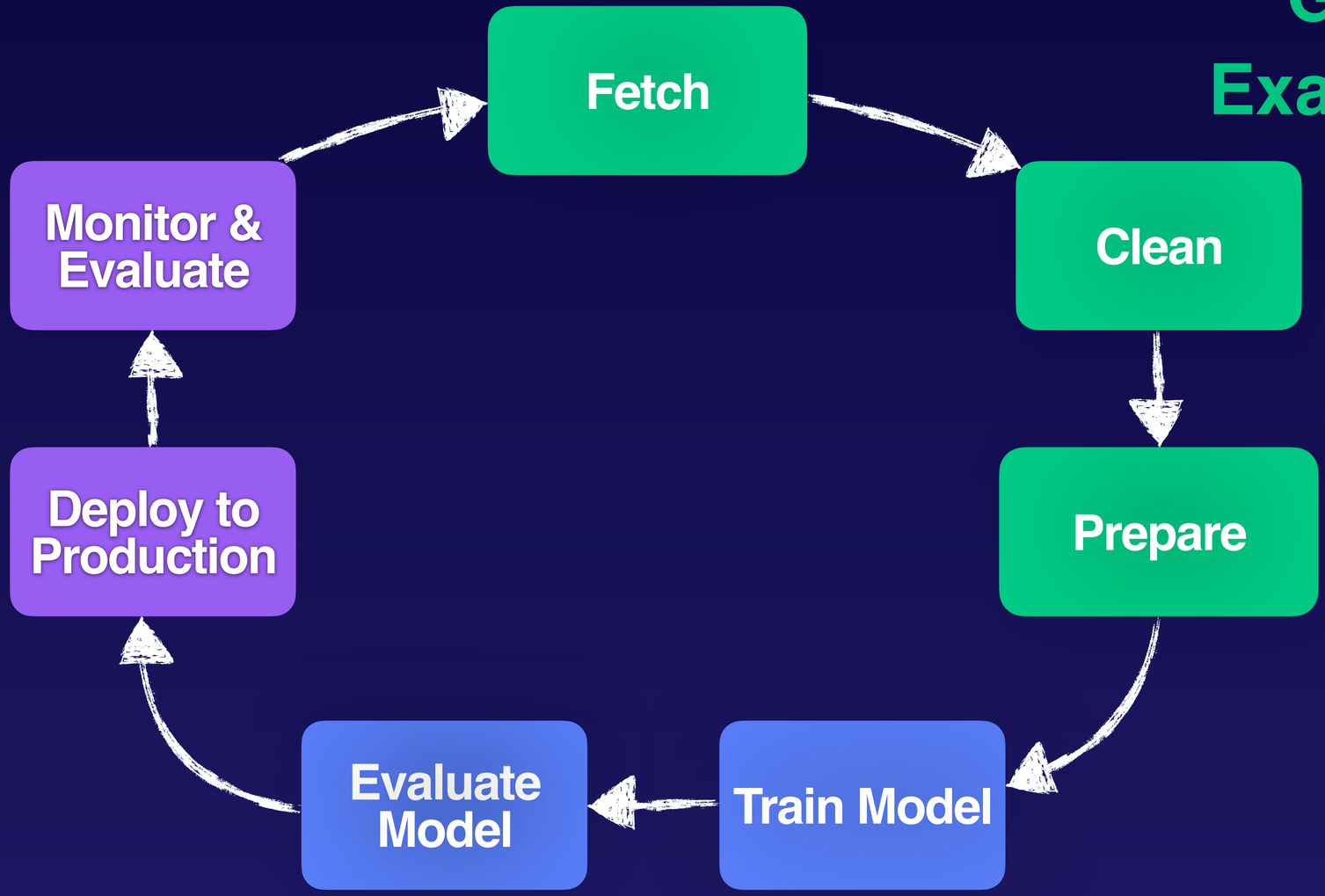
A CLOUD GURU

Data Preparation

Cleaning, Scrubbing, Sanitizing

Just like most types of vegetables there is always some type of upfront cleaning or human manipulation process that needs to be done before it is consumed.

Machine Learning Cycle

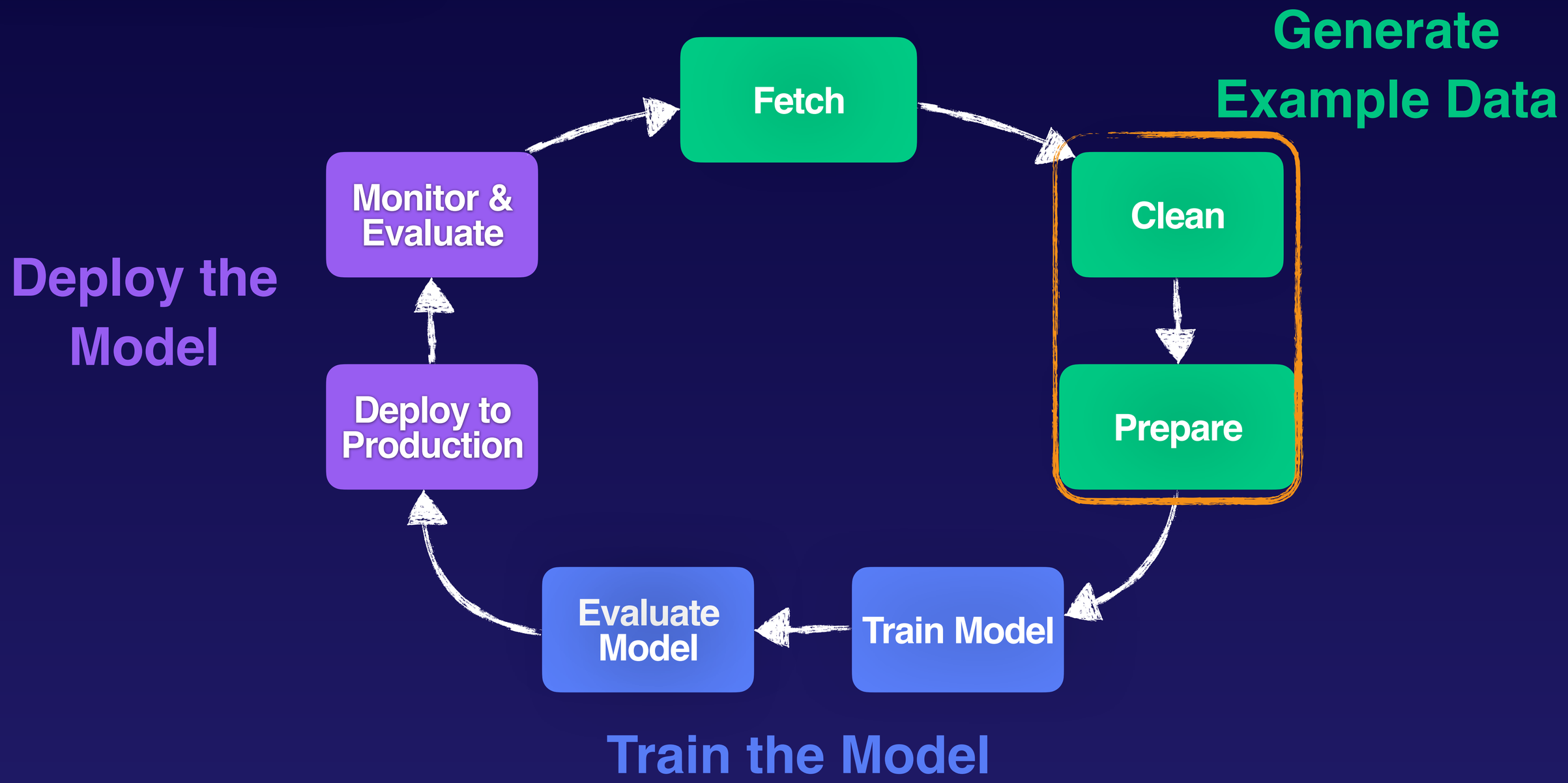


Generate Example Data

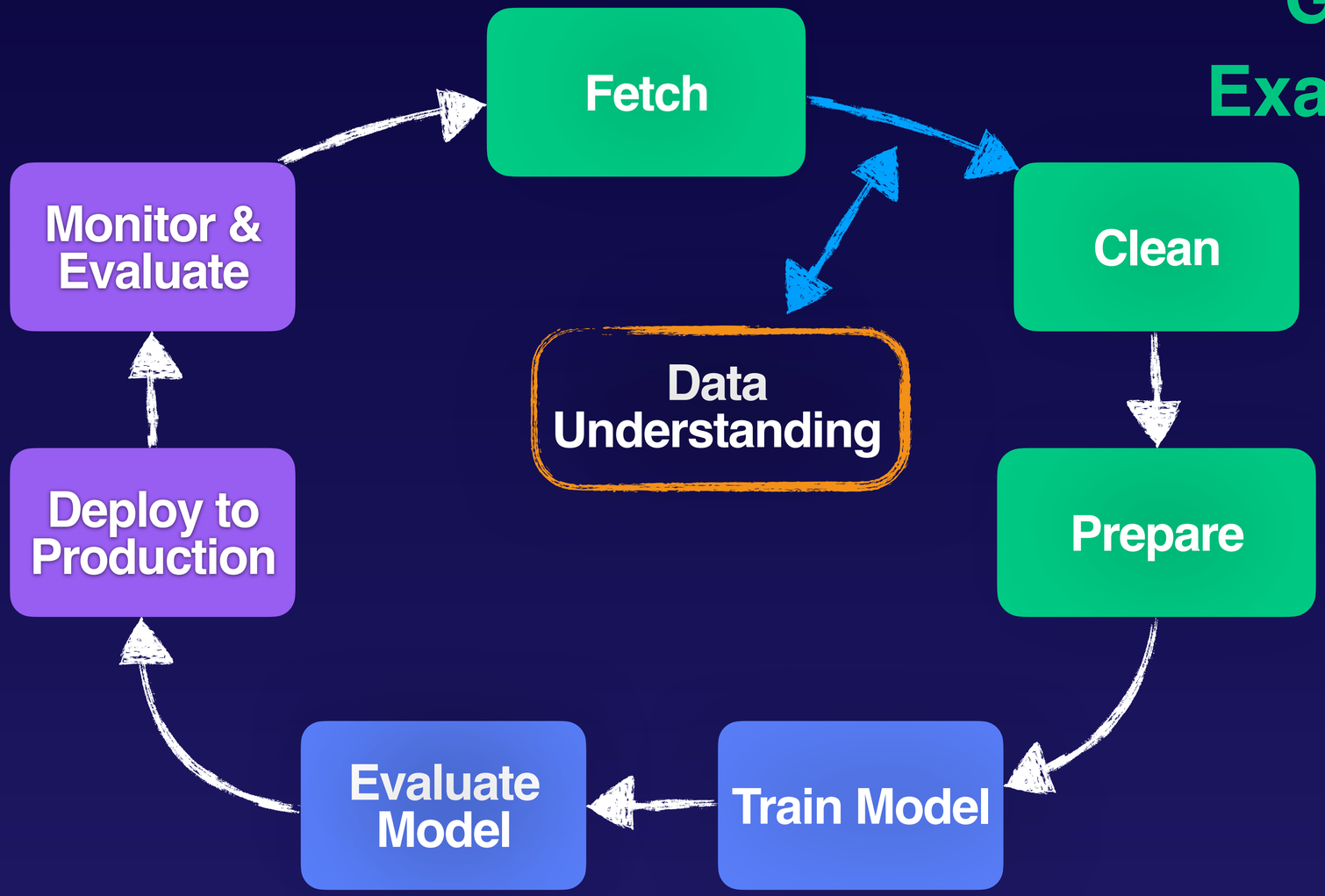
Deploy the Model

Train the Model

Machine Learning Cycle



Machine Learning Cycle



Generate Example Data

Deploy the Model

Train the Model

Data Preparation

Data Preparation is the process of transforming a dataset using different techniques to prepare it for model training and testing.

Changing our dataset so it is ready for Machine Learning.

Data Preparation

1 **Categorical Encoding**

Converting categorical values into numeric values using mappings and one-hot techniques.

2 **Feature Engineering**

Transforming features so they are ready for ML algorithms. Ensures the relevant features are used for the problem at hand.

3 **Handling Missing Values**

Removing incomplete, incorrect formatted, irrelevant or duplicated data.

Data Preparation Example

ID	Name	Evil	Affiliation
1	Luke	No	Rebels
2	Leia	NULL	REB
3	Han	0	Rebels
4	Vadar	1	Empire
5	Han	0	Rebels
6	Jabba the Hutt	1	
7	Greedo	0	Bounty Hunter

Data Preparation Example

ID	Name	Evil	Affiliation
1	Luke	No	Rebels
2	Leia	NULL	REB
3	Han	0	Rebels
4	Vadar	1	Empire
5	Han	0	Rebels
6	Jabba the Hutt	1	
7	Greedo	0	Bounty Hunter

Formatting

Data Preparation Example

ID	Name	Evil	Affiliation
1	Luke	No	Rebels
2	Leia	NULL	REB
3	Han	0	Rebels
4	Vadar	1	Empire
5	Han	0	Rebels
6	Jabba the Hutt	1	
7	Greedo	0	Bounty Hunter

Formatting

Missing Values

Data Preparation Example

ID	Name	Evil	Affiliation
1	Luke	No	Rebels
2	Leia	NULL	REB
3	Han	0	Rebels
4	Vadar	1	Empire
5	Han	0	Rebels
6	Jabba the Hutt	1	
7	Greedo	0	Bounty Hunter

Formatting

Duplicates

Missing Values

Data Preparation Example

ID	Name	Evil	Affiliation
1	Luke	No	Rebels
2	Leia	NULL	REB
3	Han	0	Rebels
4	Vadar	1	Empire
5	Han	0	Rebels
6	Jabba the Hutt	1	
7	Greedo	0	Bounty Hunter

Formatting

Duplicates

Missing Values

Invalid Data

Data Preparation Example

ID	Name	Evil	Affiliation
1	Luke	No	Rebels
2	Leia	NULL	REB
3	Han	0	Rebels
4	Vadar	1	Empire
5	Han	0	Rebels
6	Jabba the Hutt	1	
7	Greedo	0	Bounty Hunter

Formatting

Duplicates

Encoding

Missing Values

Invalid Data

Data Preparation

Data Preparation

?

Options for Data Preparation

SageMaker & Juypter Notebooks

ETL jobs in AWS Glue

Options for Data Preparation

SageMaker & Juypter Notebooks

ETL jobs in AWS Glue

Adhoc **Reusable**