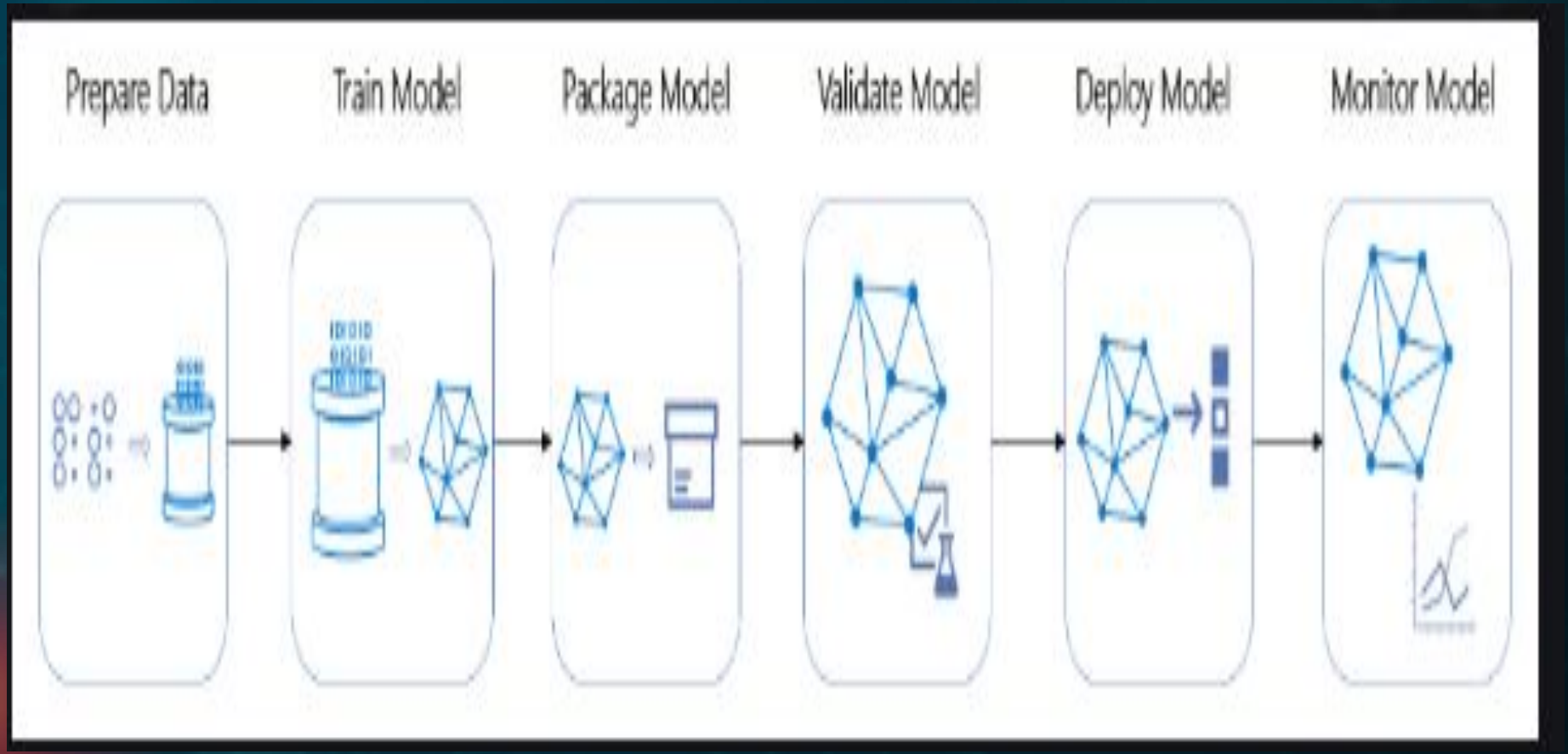
# Day - 3
# ML Pipeline & Data preprocessing

# CONTENT

# ML Pipeline

- O — Obtaining our data
- S — Scrubbing / Cleaning our data
- E — Exploring / Visualizing our data will allow us to find patterns and trends
- M — Modeling our data will give us our predictive power as a wizard
- N — Interpreting our data


Prepare Data
Train Model
Package Model
Validate Model
Deploy Model
Monitor Model

# Obtain Your Data

- Database Management: MySQL, PostgresSQL,MongoDB
- Querying Relational Databases
- Retrieving Unstructured Data: text, videos, audio files, documents
- Distributed Storage: Hadoops, Apache Spark/Flink

# Scrubbing / Cleaning Your Data

- Examine the data: understand every feature you're working with, identify errors, missing values, and corrupt records
- Clean the data: throw away, replace, and/or fill missing values/errors

Find patterns in your data through visualizations and charts.
Extract features by using statistics to identify and test significant variables

# Exploring (Exploratory Data Analysis)

# Modeling (Machine Learning)

After cleaning your data and finding what features are most important, using your model as a predictive tool will only enhance your business decision making

- Machine Learning: Supervised/Unsupervised algorithms
- Evaluation methods
- Machine Learning Libraries: Python (Sci-kit Learn) / R (CARET)
- Linear algebra & Multivariate Calculus

# Updating Your Model

As your model is in production, its important to update your model periodically, depending on how often you receive new data. The more data you receive the more frequent the update.he introduction to new features will alter the model performance either through different variations or possibly correlations to other features.

# DATA PREPROCESSING

- Preparing the proper input dataset, compatible with the machine learning algorithm requirements.
- Improving the performance of machine learning models.

```
# importing required values
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline

# read the train data
train_data = pd.read_csv('dataset/train_kOBLwZA.csv')

# check for the null values
train_data.isna().sum()
```

# Feature Engineering

Feature engineering is the process of using domain knowledge of the data
to create features (feature vectors) that make machine learning algorithms work.
feature vector is an n-dimensional vector of numerical features that represent some object.
Many algorithms in machine learning require a numerical representation of objects,
since such representations facilitate processing and statistical analysis.