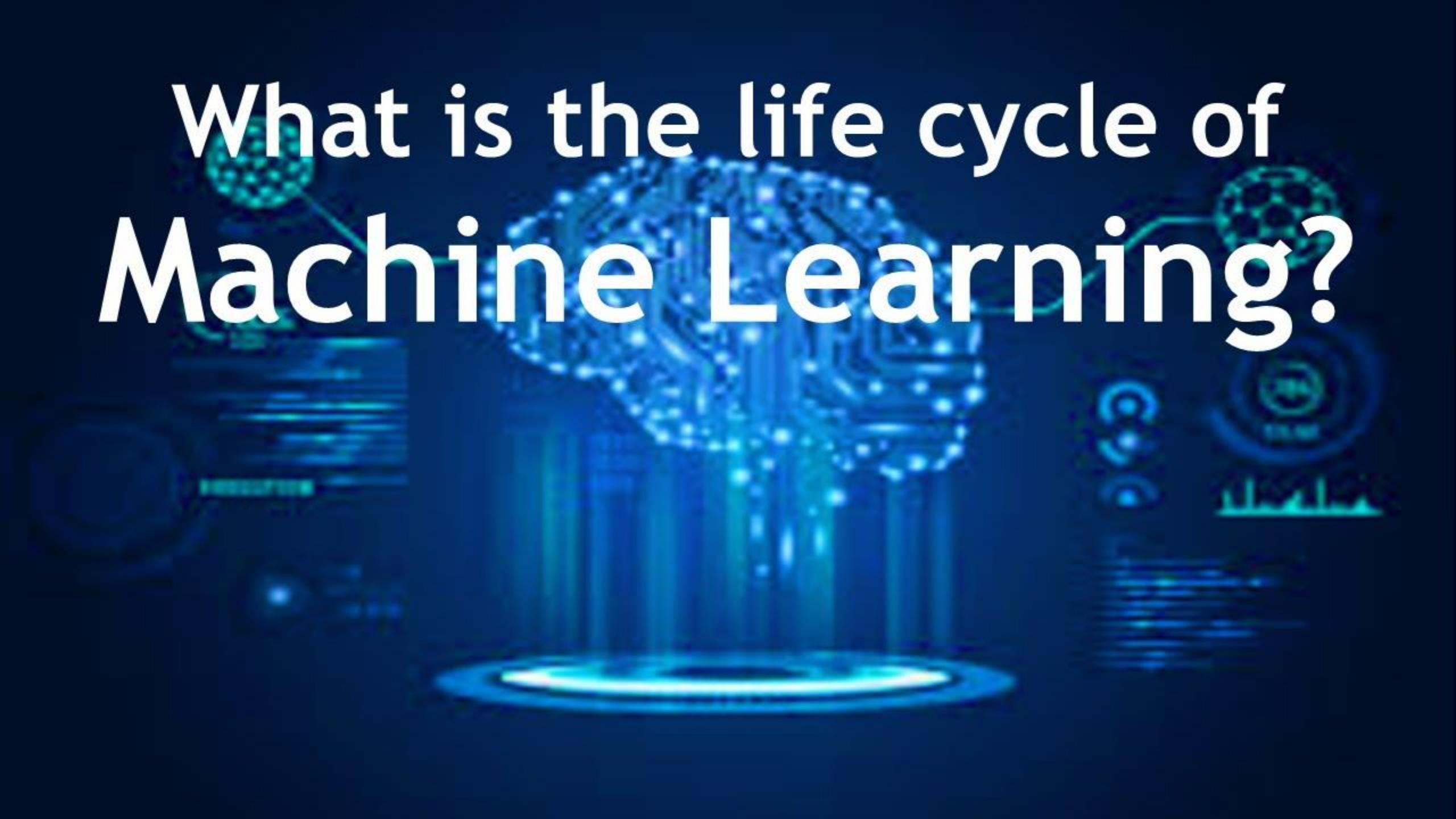
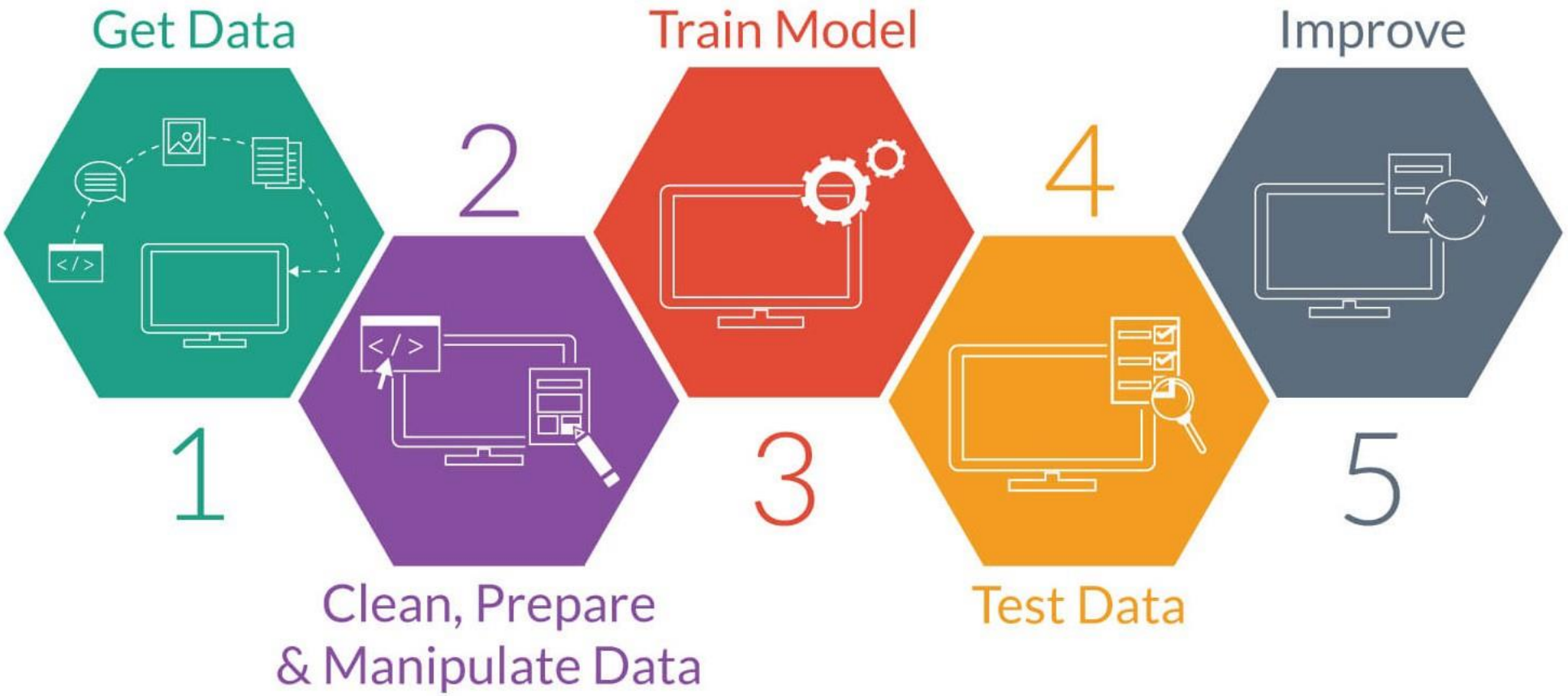
What is the life cycle of Machine Learning?



What Do we need to build a machine learning model?

Building a Machine Learning model



Data Collection

- For an innovative project:
- you need to be innovative ☺ --

- Collect data yourself, simulations, use existing data in a different way, .. find a way

- In many cases, you may deal with legacy systems and extract data from them

Data Collection Methods

Primary

- Survey/ Questionnaire
- Interview
- Observation
- Experiment

Secondary

- Literature Review
- Government Database
- Commercial Database
- Web Scraping

- Data needs to be in the form of:
  - Features
    - Numerical (Quantitative)
    - Categorical (Qualitative)
      - Ordinal : categories has relative order like grades (A , B+ , C, etc.)
      - Nominal : categories has no relative order like gender (Male, Female)

  - Labels
    - Continues : in regression
    - Discrete : in classification

- **Now I have the data, can I build the model?**

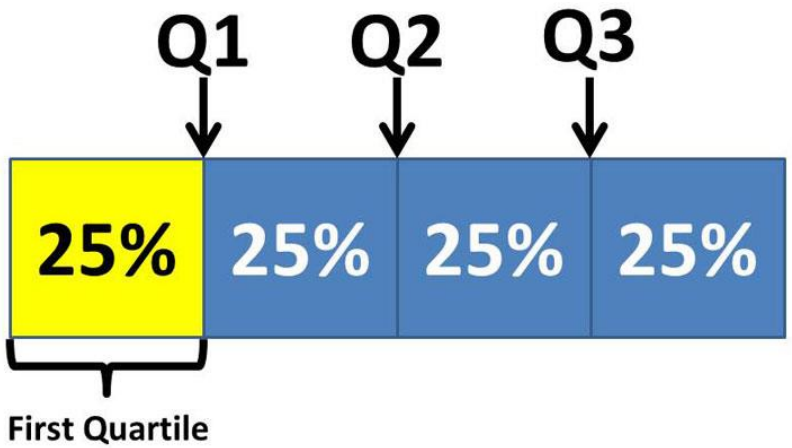
- **Not yet, you must prepare the data**

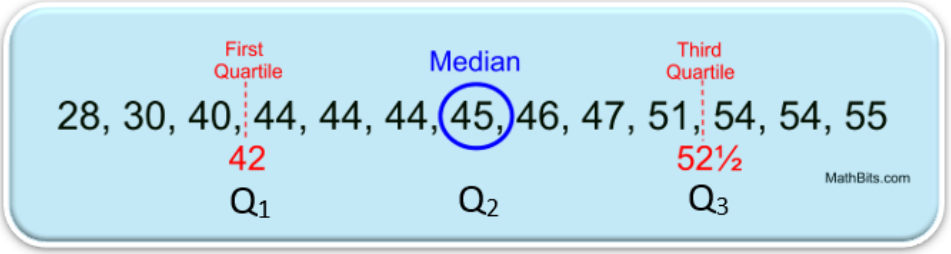
Data Preparation

- Combine data from various sources
- Check data schema:
  - Usually, the source of the data has schema describing the data
- Clean your Data: identify and handle errors in your data.
  - Handle Missing data
  - Handle outliers
- Data Transformation: change the scale of some/all variables. Why?? We will discuss it later.
- Feature Selection: select these features that are most relevant to your task.
- Feature Engineering: combine features, derive new variables, dimensionality reduction, etc.

Exploratory Data Analysis

- We can’t prepare the data without first exploring the data.
- Quantify missing data
- Identify numerical and categorical variables
- Determine unique values (cardinality) in categorical features
- Check rare/ dominant categories in categorical features
- Highlight outliers
- Identify linear relationships
- Identify a normal distribution
- Check histograms


Q1
Q2
Q3
25%
25%
25%
25%
First Quartile


First Quartile
Median
Third Quartile
28, 30, 40, 44, 44, 44, 45, 46, 47, 51, 54, 54, 55
42
52½
Q1
Q2
Q3
MathBits.com

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

Label Encoded Data

Team	Points
0	25
0	12
1	15
1	14
1	19
1	23
2	25
2	29

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

One-Hot Encoded Data

Team_A	Team_B	Team_C	Points
1	0	0	25
1	0	0	12
0	1	0	15
0	1	0	14
0	1	0	19
0	1	0	23
0	0	1	25
0	0	1	29