Welcome to Data Science Online Bootcamp

Building Your First ML Model

dφ

Democratizing Data Science Learning

Learning Objectives

Machine Learning & its use-cases

ML Keywords

Building our First ML Model

The ABC of Machine Learning

What is Machine Learning?

- Machine learning provides systems the ability to **automatically learn and improve from experience without being explicitly programmed**.

- It allow computers to discover hidden and useful insights

- **In nutshell, Machine Learning is a new way of communicating your wishes to a computer.**

How does Machine Learning (ML) model?

- For now, let's consider it is a Magical box that help us to predict what we want. In the below case we want to predict whether an incoming email should land in our inbox or spam box. We will discuss more about ML models soon.

In other terms this is nothing but **data**. This data will have variables such as: sender email id, subject of email, email body etc

Once the incoming emails go through the Machine Learning Model it categorizes and predicts whether a mail should go in your inbox or spam box

Machine Learning is used in..

- **Fraud detection** - **Eg:** Credit card fraud detection. It will help us to detect whether a transaction is fraud or not.

- **Email spam filtering** - **Eg:** Helps in categorising whether a particular email should go in inbox or spam box.

- **Recommendation engines** - **Eg:** E-commerce platforms like Amazon can recommend you a similar product based on your previously browsed list of products

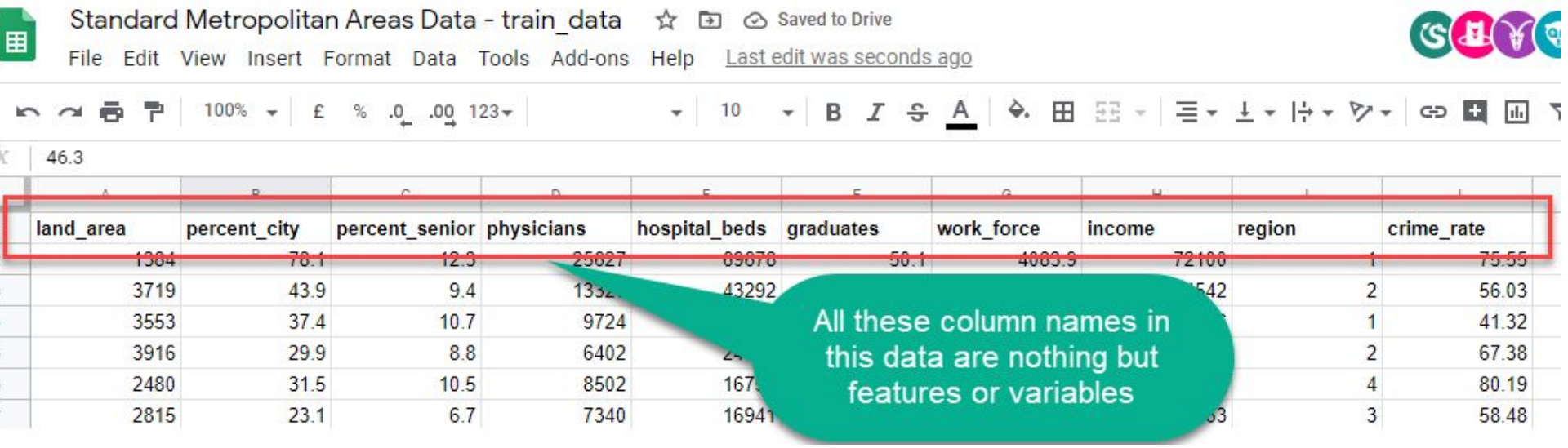
- and many more!!!

Let’s understand some keywords in ML!

Variables/features

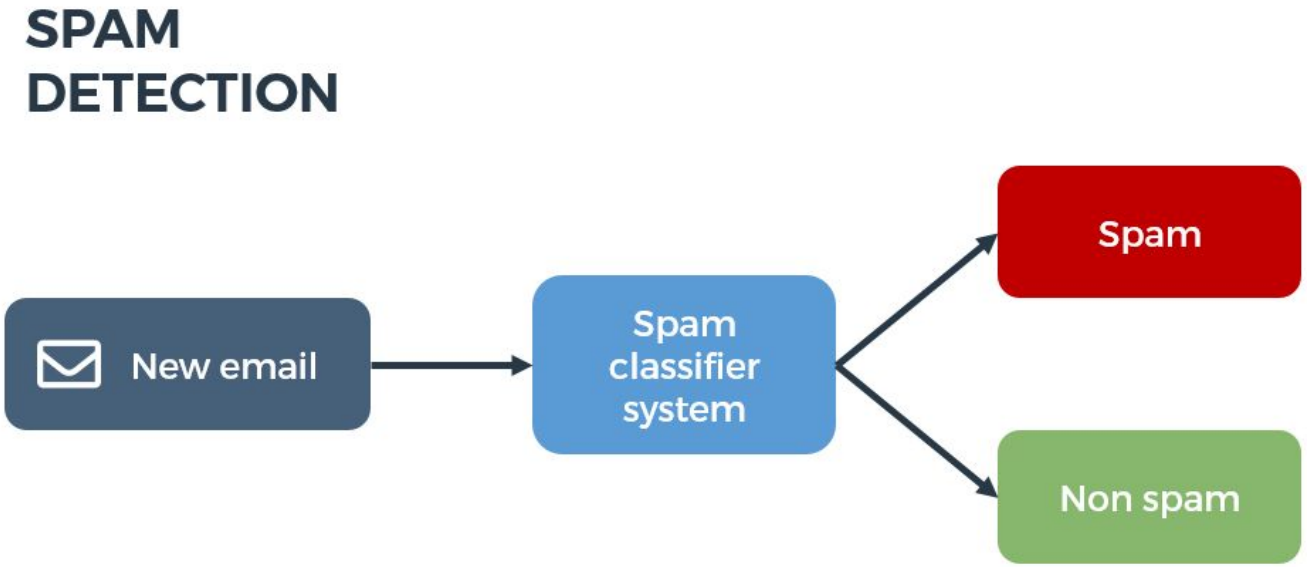
- **Features or Variables:** These are the the most common terms that we would come across from now on.

- **Features and Variables both are the same in a dataset**, they are often interchangeably used. So there is no need to worry about it!



Target/Label Variable

- The target variable or label of a dataset is the feature of a dataset about which you want to gain a deeper understanding.
- It is the variable that is, or should be the output.
- In the example of detecting spam emails, the label will be the category the email belongs to, i.e it will be either ‘spam’ or ‘not spam’.



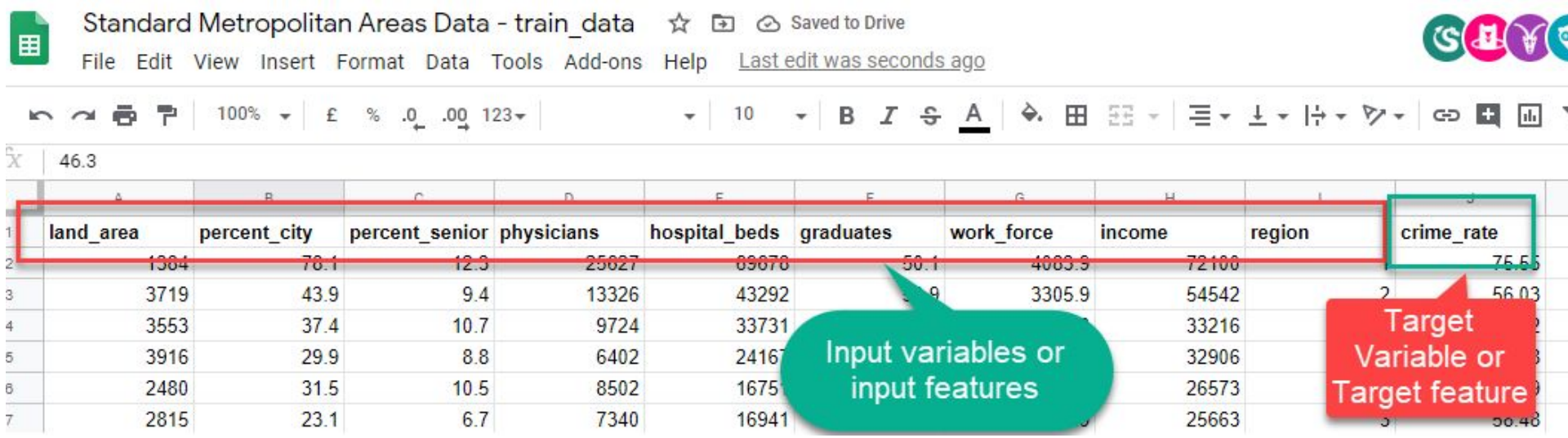
Predictor/Input Variables

- One or more variables that are used to determine (or predict) the 'Target Variable' are known as Input Variables. They are sometimes called Predictor Variable as well.

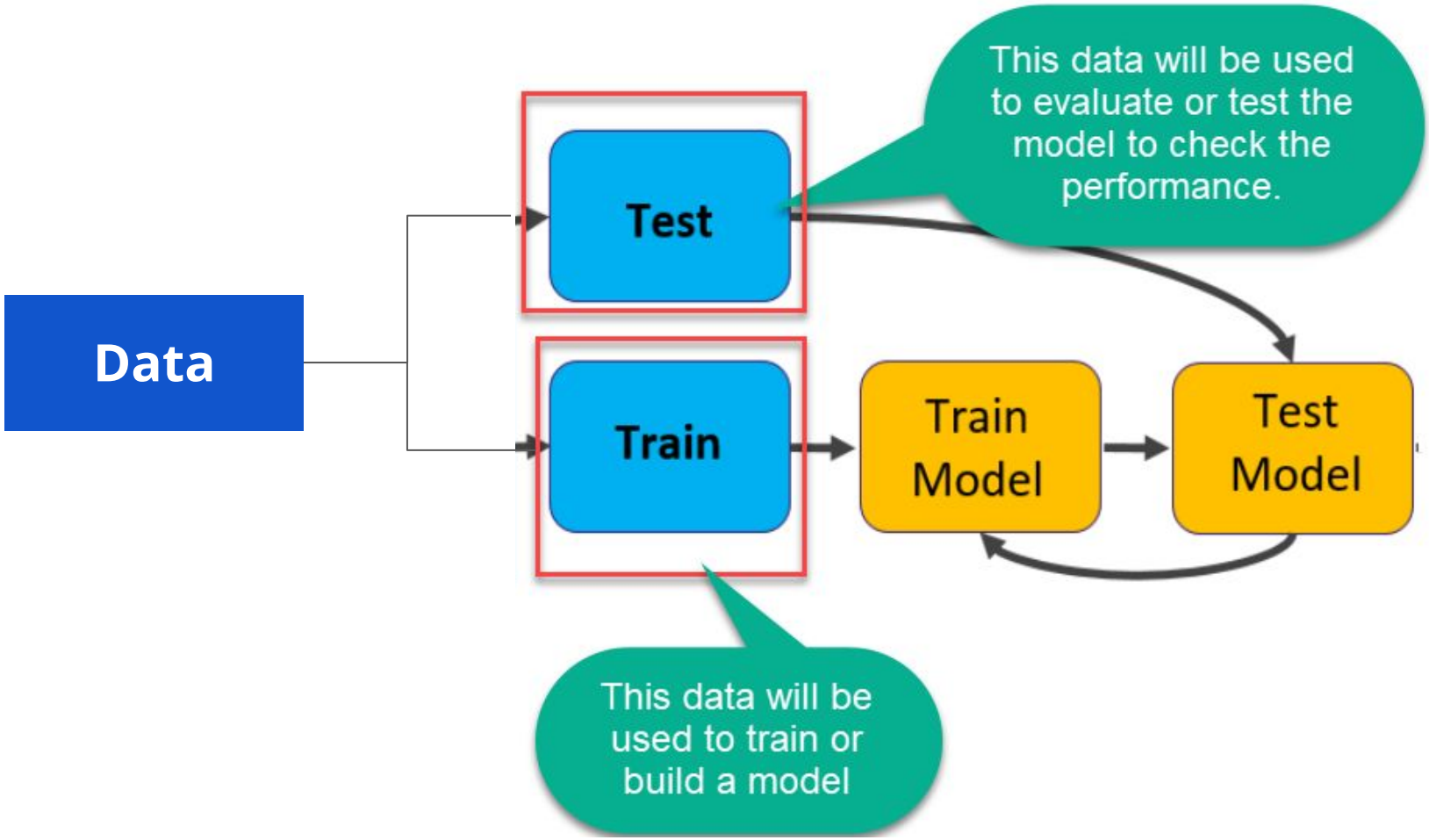
- In the spam detector example, the features could include the following:
  - words in the email text
  - sender's address
  - time of day the email was sent
  - email contains the phrase "congrats you won $1 billion - share your bank details."

Target and Input variables

- Remember the Standard Metropolitan Areas Data used in previous slides? In that dataset **we might be curious to predict "crime_rate" in future**, so that becomes our target variable and rest of the variables become input variables for building a machine learning model.



Train and Test Set



Exams

Case 1:

- You end up reading everything in the textbook
- You performed well in the test

Case 2:

- You got the test questions through some means and you ended focusing more on it
- And obviously, you performed well

Is there any learning in case 2? - we don’t want will happen if you are exposed to new questions right?

Exams

Is there any learning in case 2? - there might be some learning (we end up learning those few questions) and we might end up getting good grades too as we know the test questions. But we don't know want will happen if we are exposed to new questions that are outside the leaked test questions right?

Let's say the school authorities learnt about the test paper leak and changed the questions (which you would know only after starting your exam and you are under prepared). Then we might end up getting not so great score. Because we have trained ourselves on those specific test questions that are leaked. Implying we just trained ourselves on a specific test paper and our learning was bad when move out of our comfort zone and expose ourself to new set of questions.

Exams

Crux of the story: We should never expose test data while training a model as it might lead to overfitting that might give you good results for that particular data but when exposed to new data we might get bad results.

Sports

- **Train:** We get trained at our local sports centers
- **Internal matches (Test):** For preparation
- **Future:** Get ourselves exposed to the real matches - national and international championship

Train and Test Dataset

- This is exactly why we need to test machine learning models on **unseen data or test data.** Otherwise, we have no way of knowing whether the algorithm has learned a generalizable pattern or has simply memorized the training data.

- **TRAINING DATA:** The observations in the training set form the experience that the algorithm uses to learn.

- **TEST DATA:** The test set is a set of observations used to evaluate the performance of the model using some performance metric. **It is important that no observations from the training set are included in the test set.** If the test set does contain examples from the training set, it will be difficult to assess whether the algorithm has learned to generalize from the training set or has simply memorized it.

Train Test Split

- Consider an Example where our Original Dataset has 1000 rows.
- When we start building our ML model we will split our datatset into two parts (70% train data and 30% test data).
- We will train our model on 70% of data i.e 700 rows and then test our model performance on 30% of data i.e 300 rows. As discussed above while testing our model we will not provide the outcome to our model for the test data although we know the outcome and instead let our model give us the outcome for those 300 rows.
- Later we will compare the outcome of our model to the original outcome of our test data to get the accuracy of our model predictions.
- For splitting our data to training and testing set we use **train_test_split method of scikit-learn library.**

Understanding Target Variable

To identify whether it is a classification or regression problem

Target Variable Classification vs Regression

Classification:

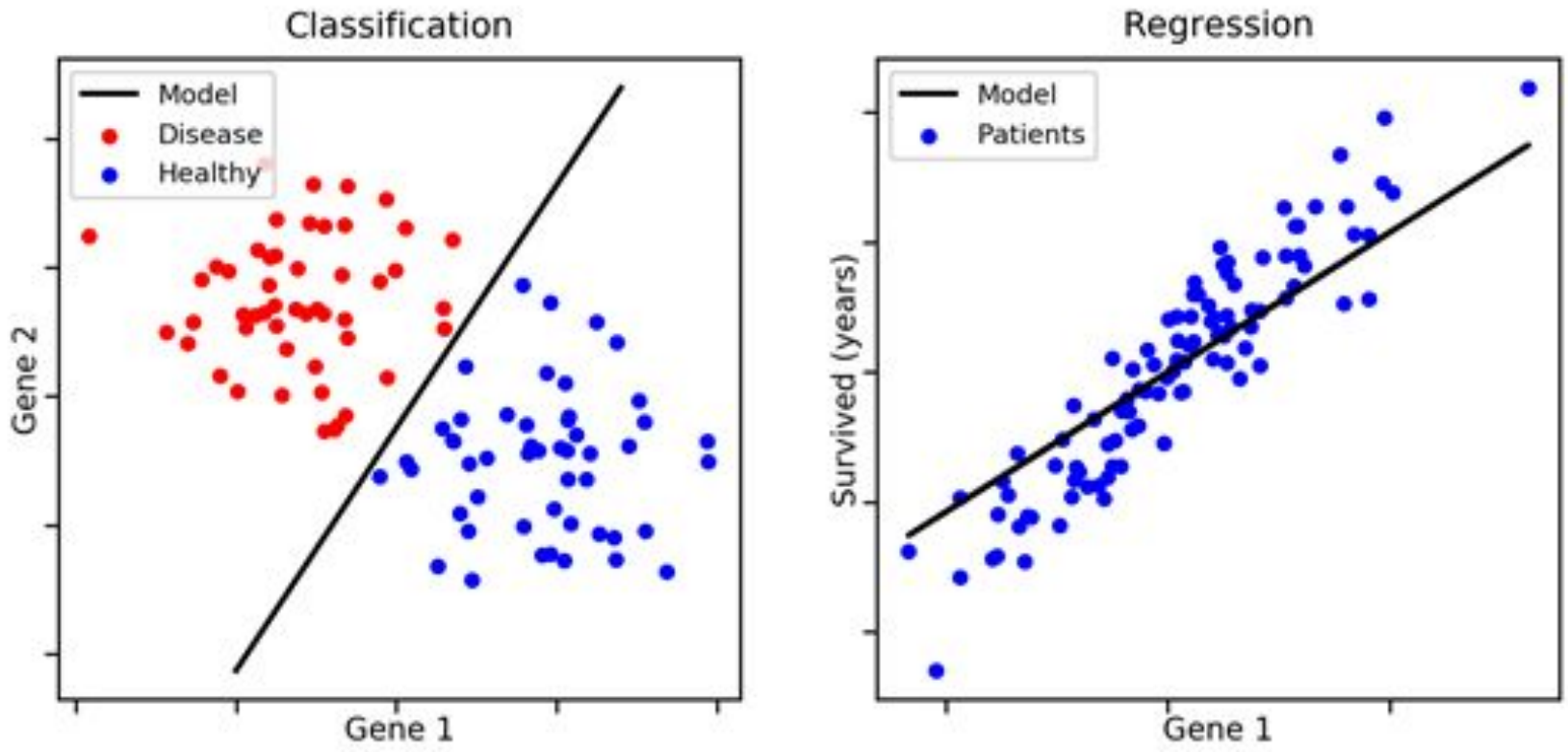
- Classify the outcome
- **Examples:**
  - Predict whether a transaction is fraud or not fraud
  - Predict whether to give loan or not
  - Predict whether to give college admission or not
  - Predict the grade (Grade A, B, C, D)
  - Note: Classification can be more than two

Regression

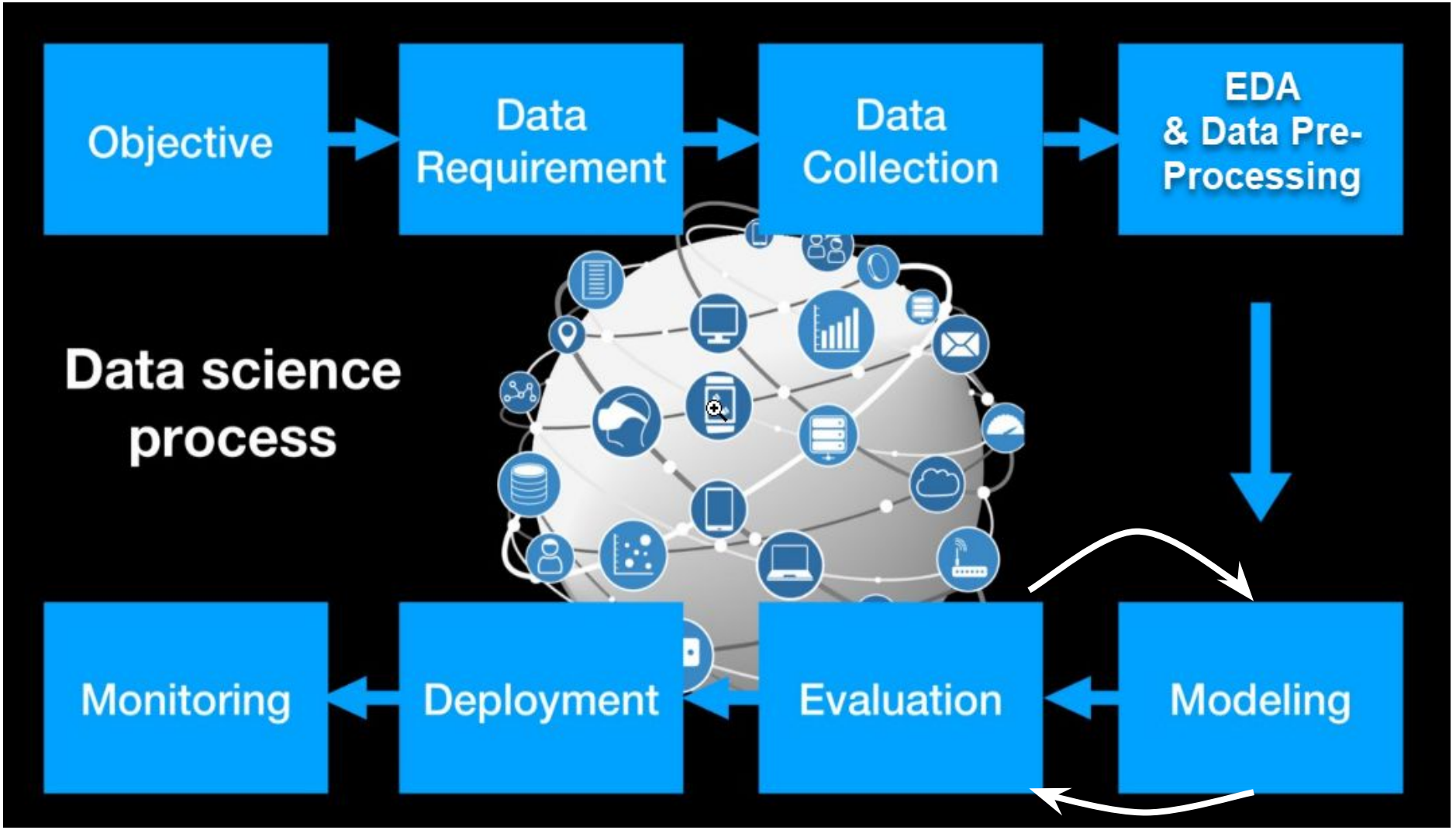
- Regression is the problem of predicting a continuous outcome (a numeric outcome)
- **Examples:**
  - Predict house price
  - Predict crime rate

Machine Learning

Classification Vs Regression



Data Science Modeling Process



Credits: https://towardsdatascience.com/data-science-modeling-process-fa6e8e45bf02

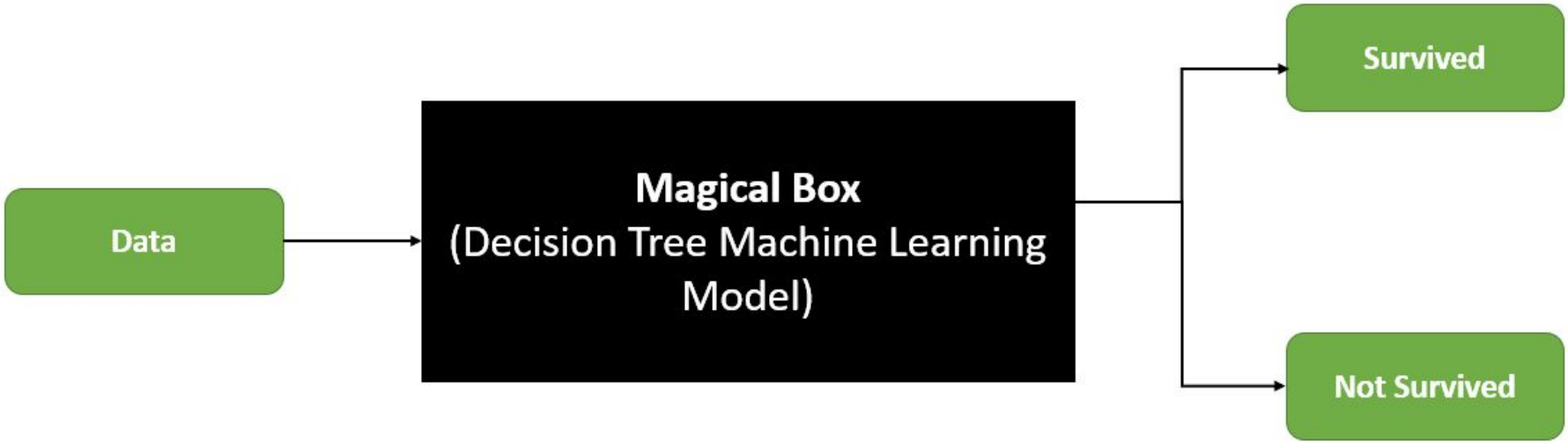
Problem Solving

- **Define Objective or understand the problem statement**
- Data Requirements
- Data Collection
- **Exploratory Data Analysis**
- **Data Pre-processing**
- **Build a model**
- **Evaluate**
- **Optimse**
- **Production**
- **Monitor**
- **You keep Optimising it every now and then**

Objective/Problem Statement

The goal of the model is to **predict whether a passenger survived or not in the Titanic disaster**, given their age, class and a few other features.

Objective/Problem Statement



Data

We have the data!

Understanding the Data

- PassengerId - this is a just a generated Id
- Pclass - which class did the passenger ride - first, second or third
- Name - self explanatory
- Sex - male or female
- Age
- SibSp - were the passenger's spouse or siblings with them on the ship
- Parch - were the passenger's parents or children with them on the ship
- Ticket - ticket number
- Fare - ticket price
- Cabin
- Embarked - port of embarkation
- Survived - did the passenger survive the sinking of the Titanic?

Explore the data

Let’s get to the notebook

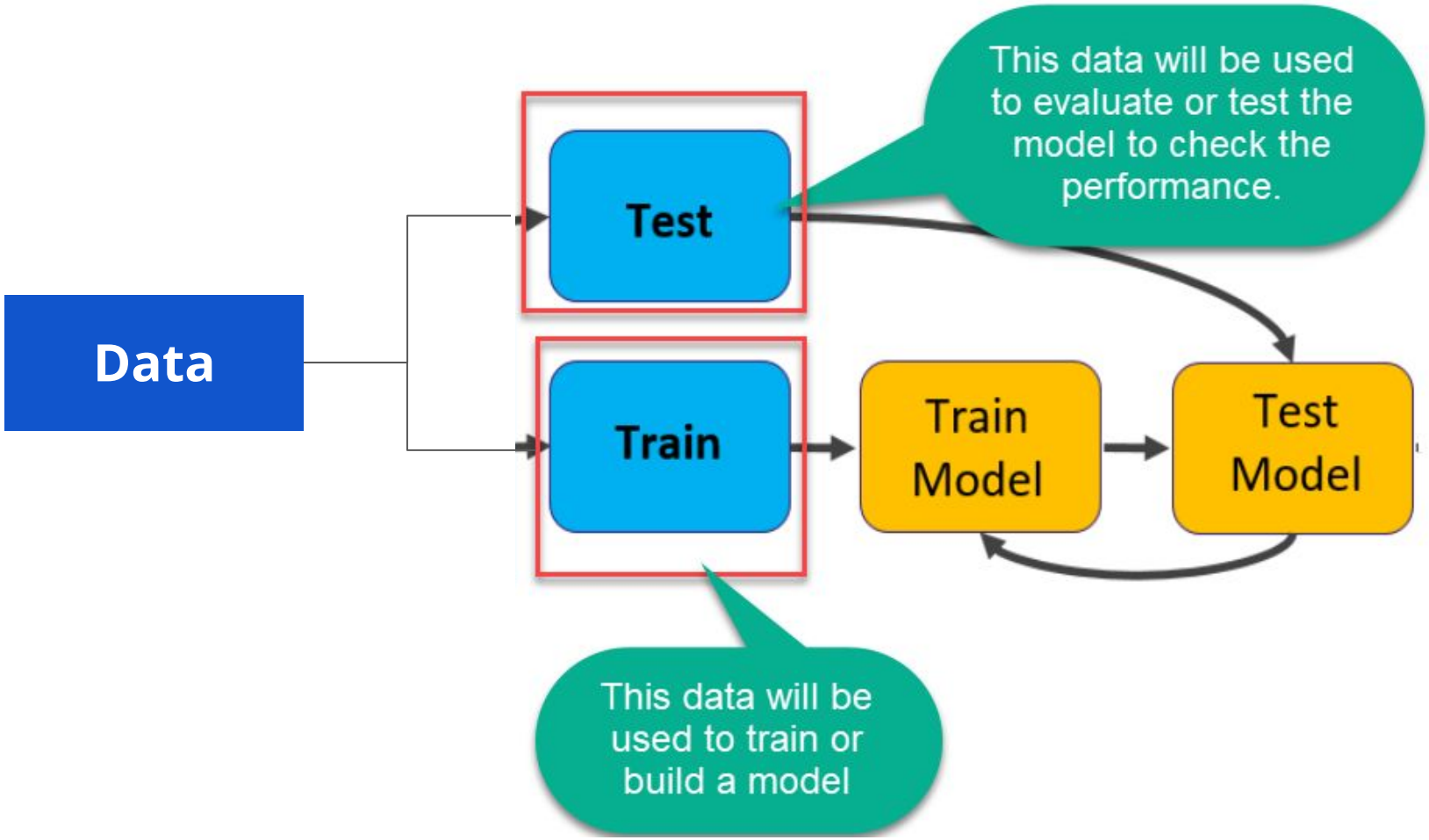
Missing Values

Omitting Irrelevant Variables/Columns

You shouldn’t drop columns or variables just like that! Unless there is a strong premise.

Id, port, cabin and name

Split the data into train and test

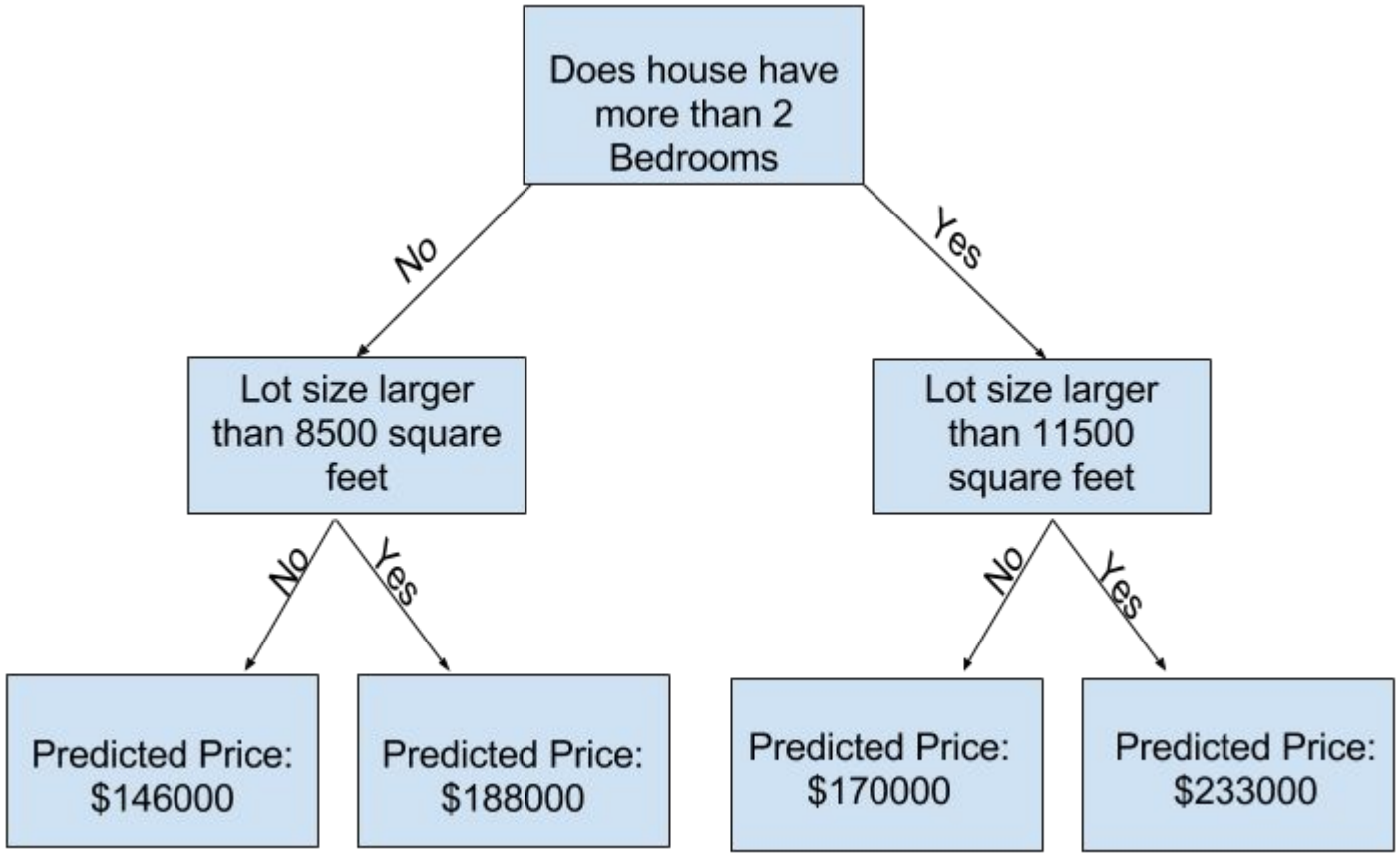


Model Building - Decision Tree

Now what is this decision tree?

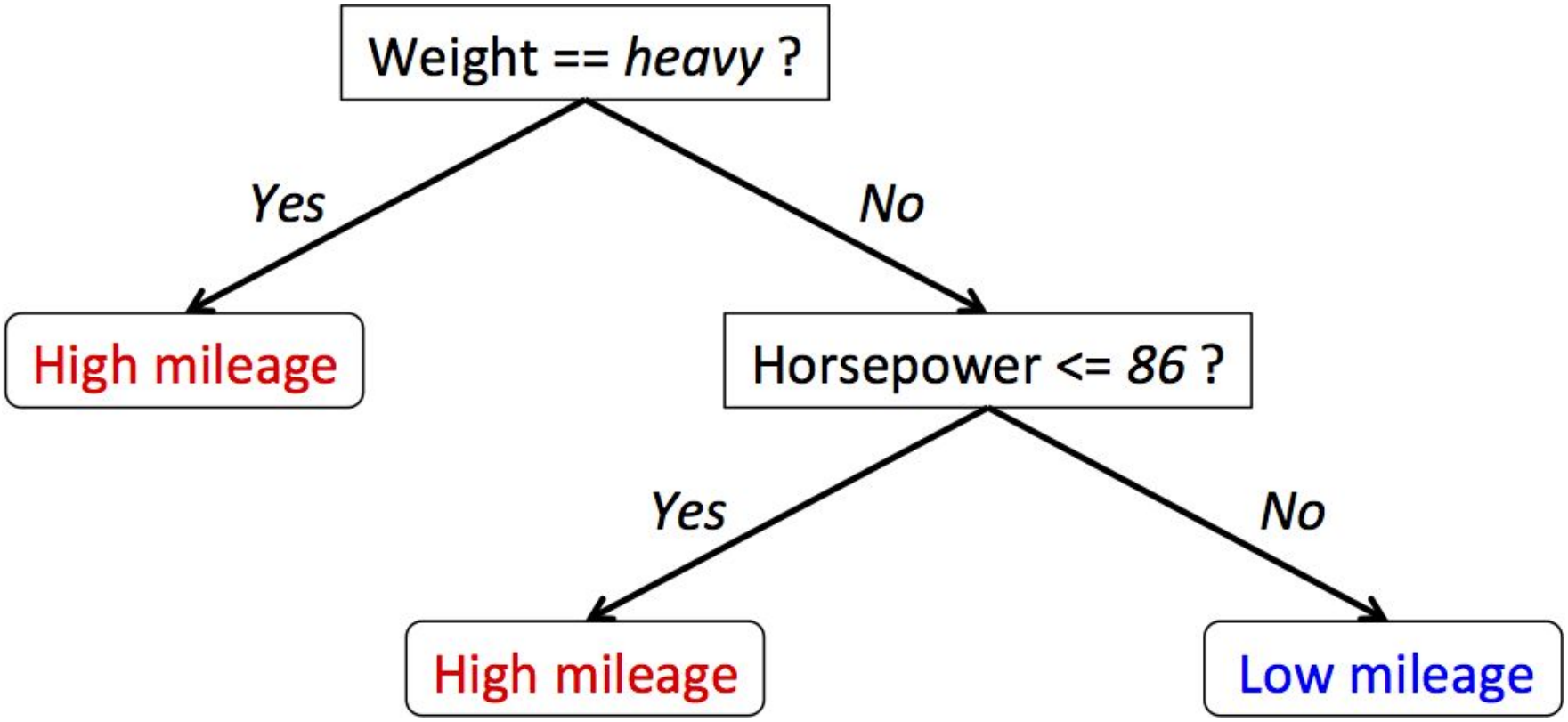
Well, we all might have seen it by now!

Decision Tree Examples

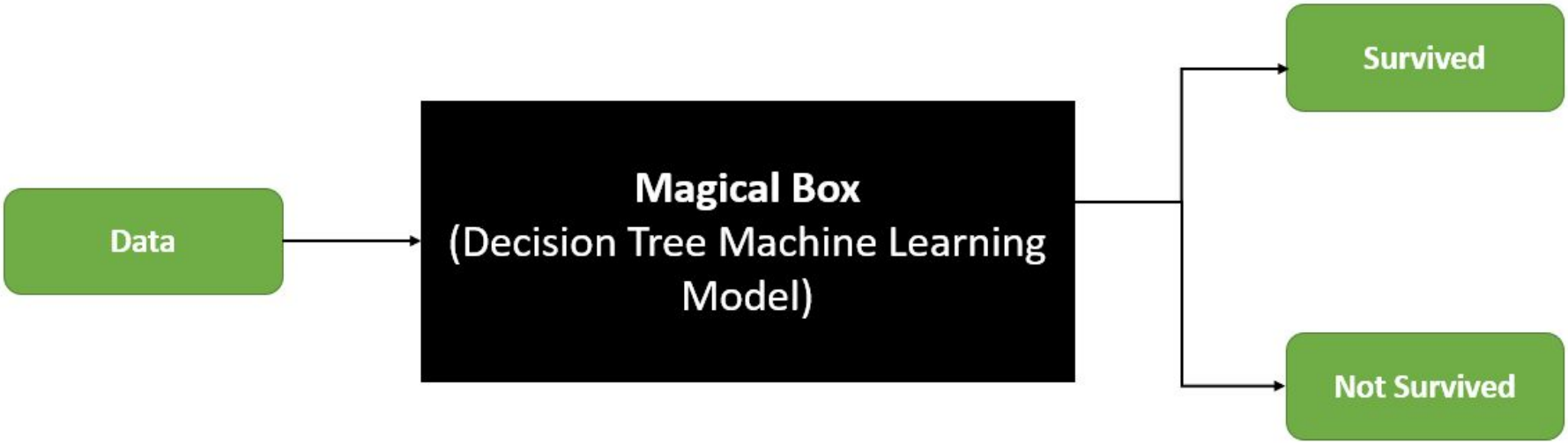


Decision Tree Examples

for Car Mileage Prediction



Now what next?



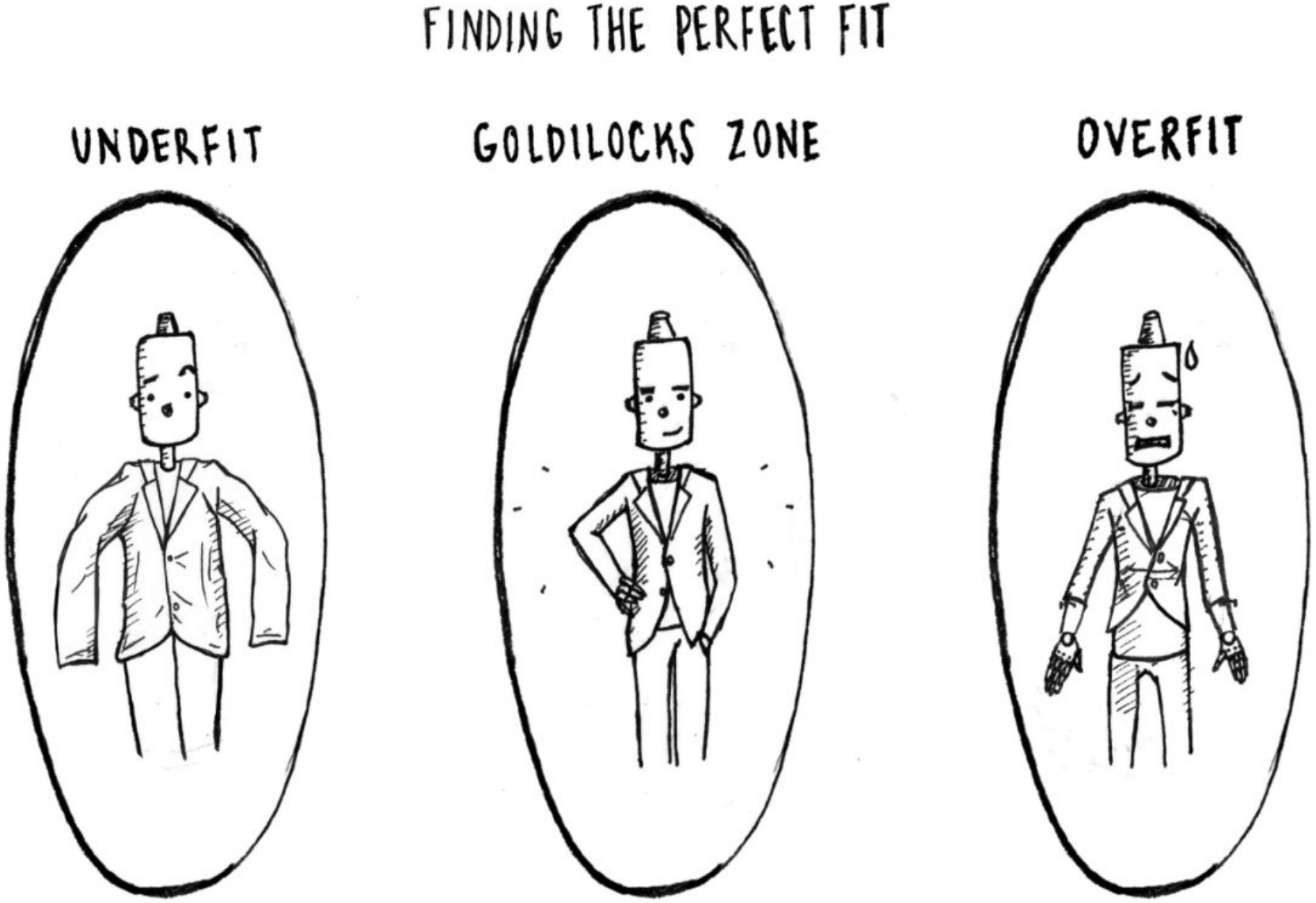
Let's do it!

Model Evaluation

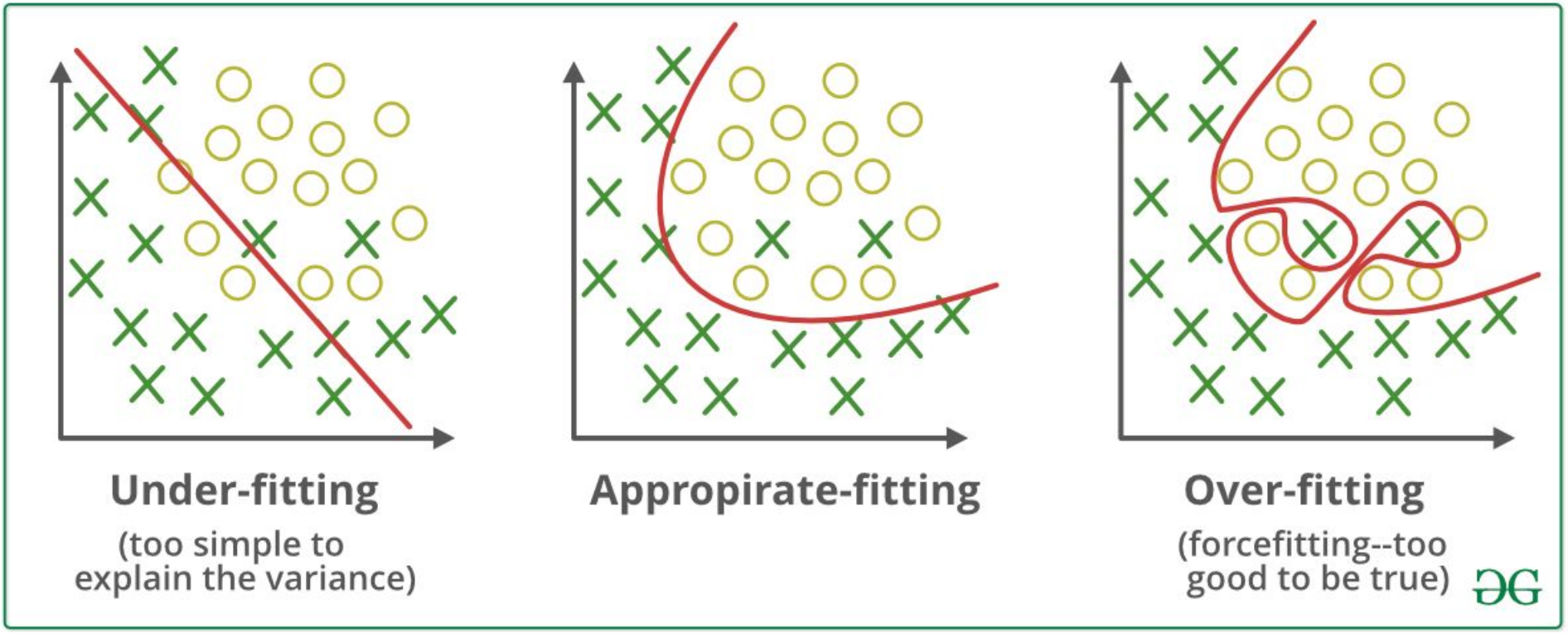
Evaluate on test dataset to check the performance!

Well, we build a model on historical data and expose them to new data that we would see in future. Technically they will be exposed to unseen data

Overfitting - Underfitting



Overfitting - Underfitting



Model Evaluation

!! We are not done yet,

We can improvise it significantly.

How? It will be followed in due course!

What else can be done in general?

- Feature Selection
- Cross validation
- Applying different ML Models
- Hyper parameter tuning etc

And as data scientist we must keep optimising and building better models that derives meaningful results.

That’s it for the day. Thank you!

Feel free to post any queries in the #help channel on Slack