# Welcome to Machine Learning 101


DPhi

# DPhi Vision:

Build data culture & Democratize Data Science Learning

# My Introduction

- Co-founder - DPhi
- Founding team - Complidata (Data Science & Marketing)
- Entrepreneur - Problem-Solver - Data Scientist - Marketer.
- Love building and growing communities/teams.
- Led lean teams across various startups.

# Learning Objectives

**ABC of Machine Learning & its use-cases**

**Types of Machine Learning**

**ML Keywords**

**Problem Solving**

DPhi

# The ABC of Machine Learning

DPhi

# What is Machine Learning?

- Machine learning is the study of computer algorithms **that improve automatically through experience.**
- It allow computers to discover **hidden and useful insights**
- **In nutshell, Machine Learning is a new way of communicating your wishes to a computer.**

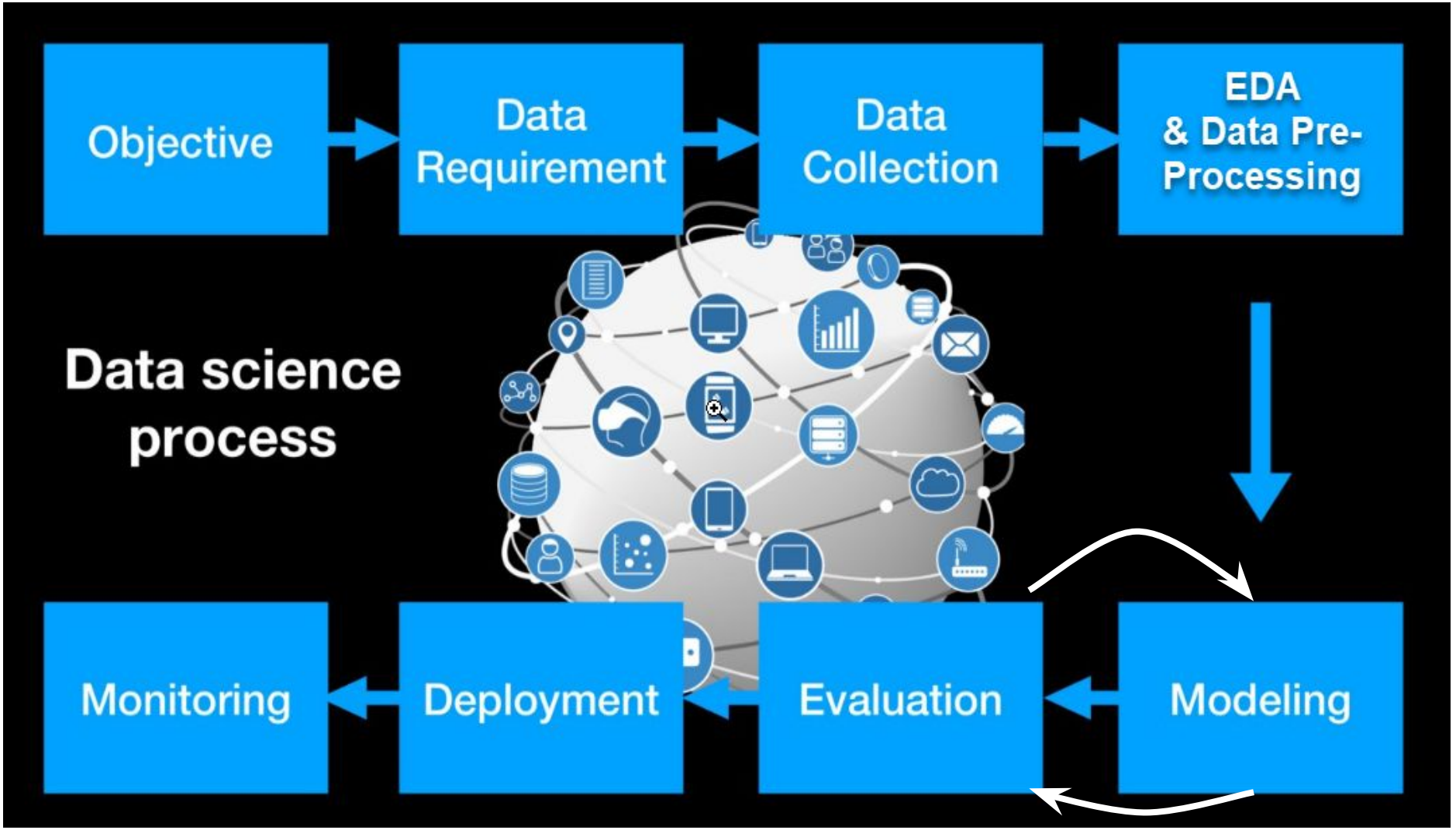
# What is a Machine Learning (ML) model?

- 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

# How? The Process




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

DPhi

# 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.

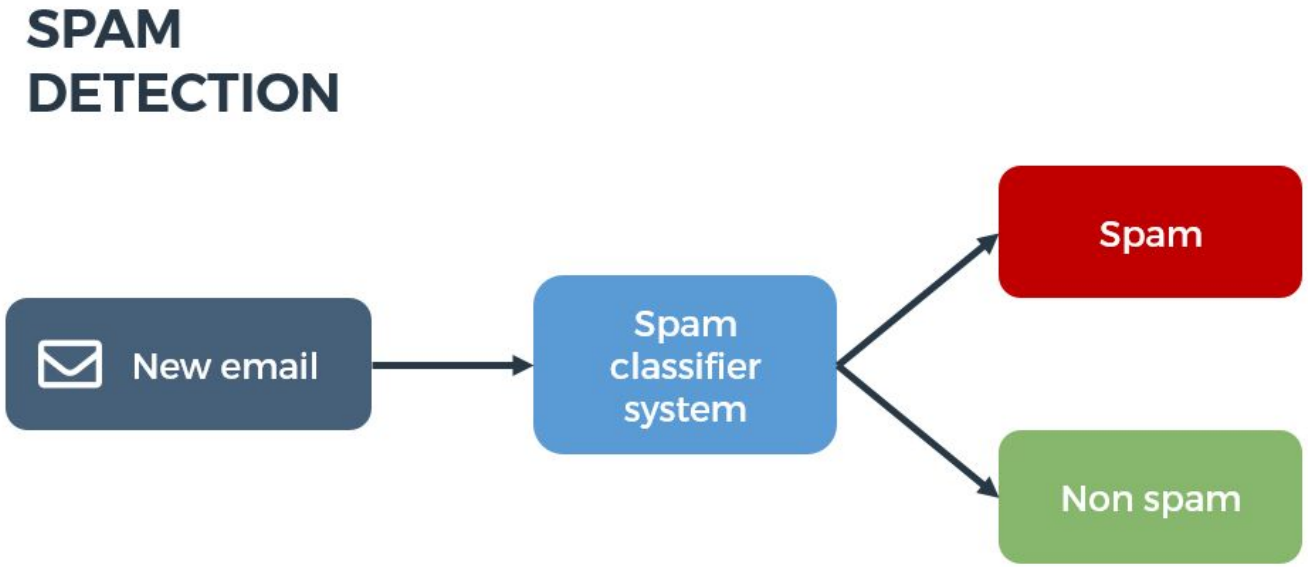
- **Recommendation engines** - **Eg:** E-commerce platforms like Amazon can recommend you a similar product based on your previously browsed list of products

- and many more!!!

**Let’s understand some keywords in ML!**

# Label/Target Variable

- The target variable or label of a dataset is the variable 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/Features

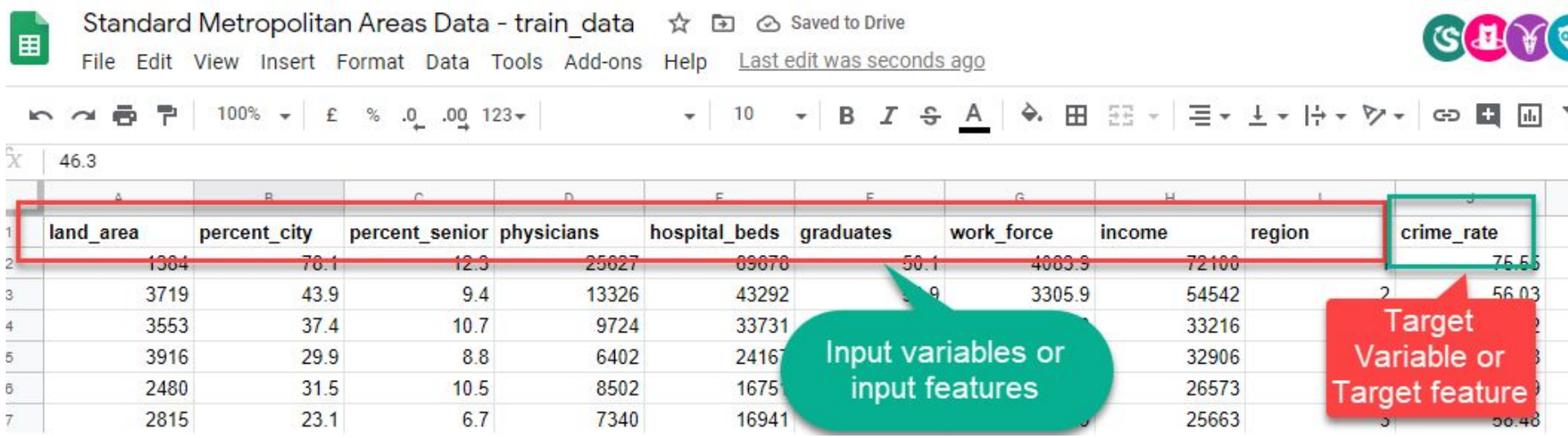
- One or more variables that are used to determine (or predict) the 'Target Variable' are known as Input Variables. They are sometimes called Predictor Variable as well.

- 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."


DPhi

# Another Example

- Standard Metropolitan Areas Data: In this dataset **we might be curious to predict "crime_rate" in future**, so that becomes our target and rest of the variables become input variables or features for building a machine learning model.

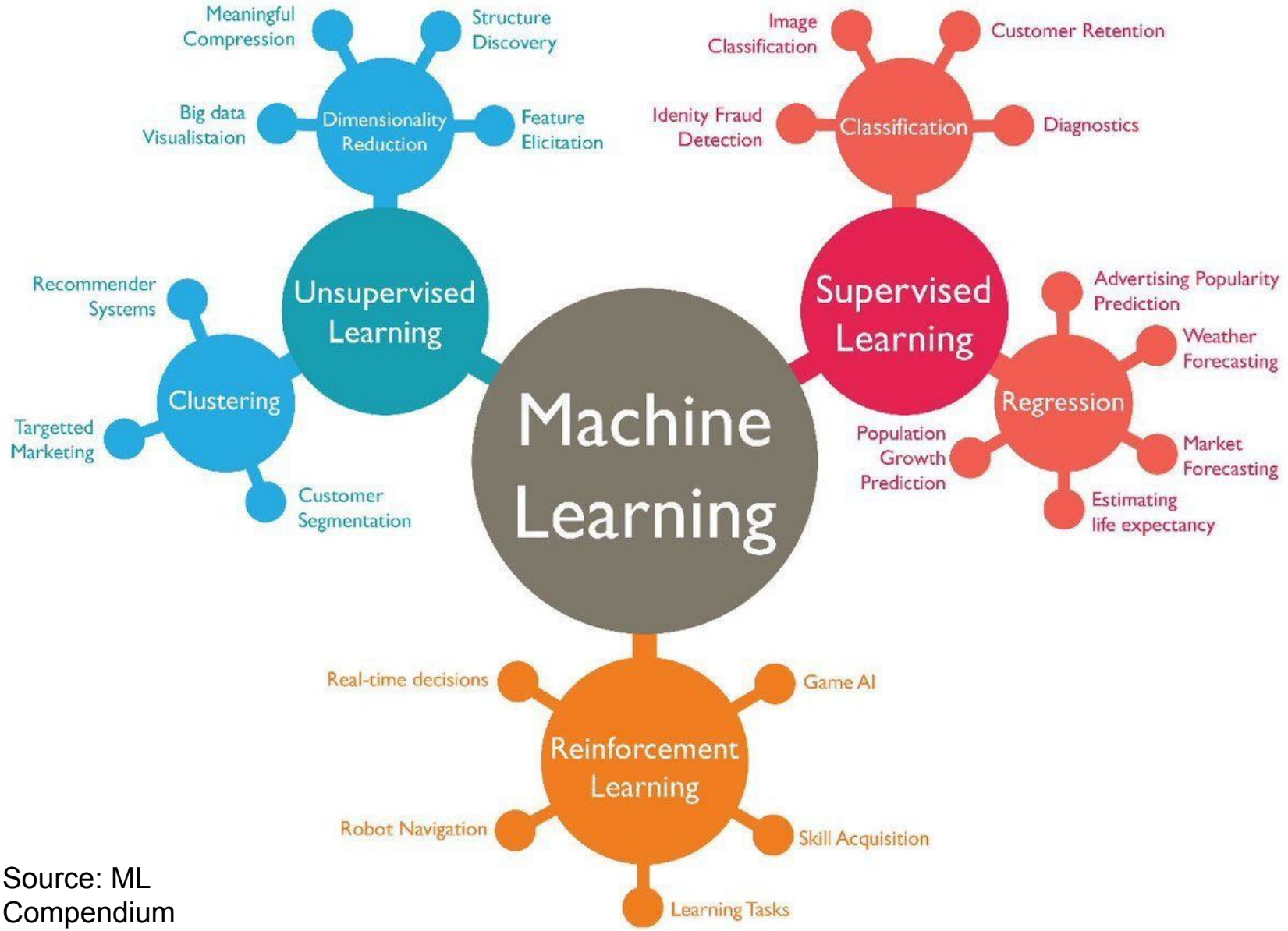



DPhi

# Types of Machine Learning

- **Supervised Learning**
  - **Classification**
  - **Regression**
- **Unsupervised Learning**
- **Reinforcement Learning**

# Types of Machine Learning




Source: ML Compendium

DPhi

# Supervised Learning

- Makes machine learn explicitly
- Data with clear defined output is given
- Direct feedback is given
- Predicts outcome/ future
- Resolve classification and regression problem


DPhi

# Unsupervised Learning

- Machine understand the data (identifies patterns/ structures)
- Evaluation is qualitative or indirect
- Does not predict or find anything specific

# Reinforcement Learning

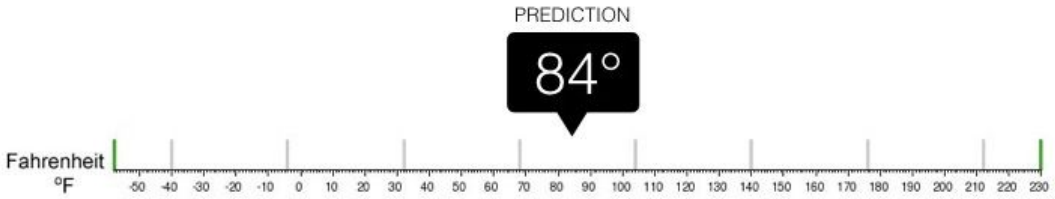
- An approach to AI
- Reward based learning
- Learn from positive and negative reinforcement
- Machine learns how to act in certain environment
- To maximize reward or minimize punishment

# Supervised Learning

## Types of Problems

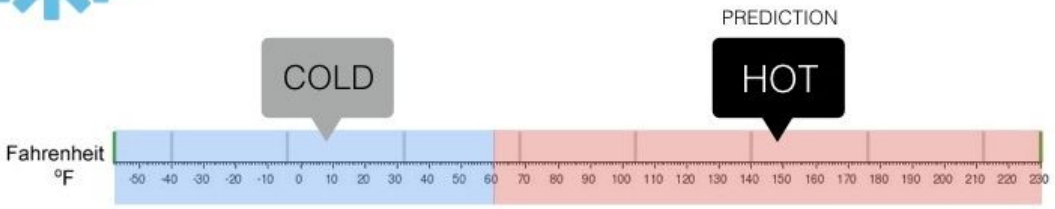
**Regression**

What is the temperature going to be tomorrow?




**Classification**

Will it be Cold or Hot tomorrow?




Source: Pinterest

DPhi

# Supervised Learning

**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

# CLASSIFICATION VS REGRESSION

Student Profile

CLASSIFICATION

*Predicting Student*

Pass Or Fail

Student Profile

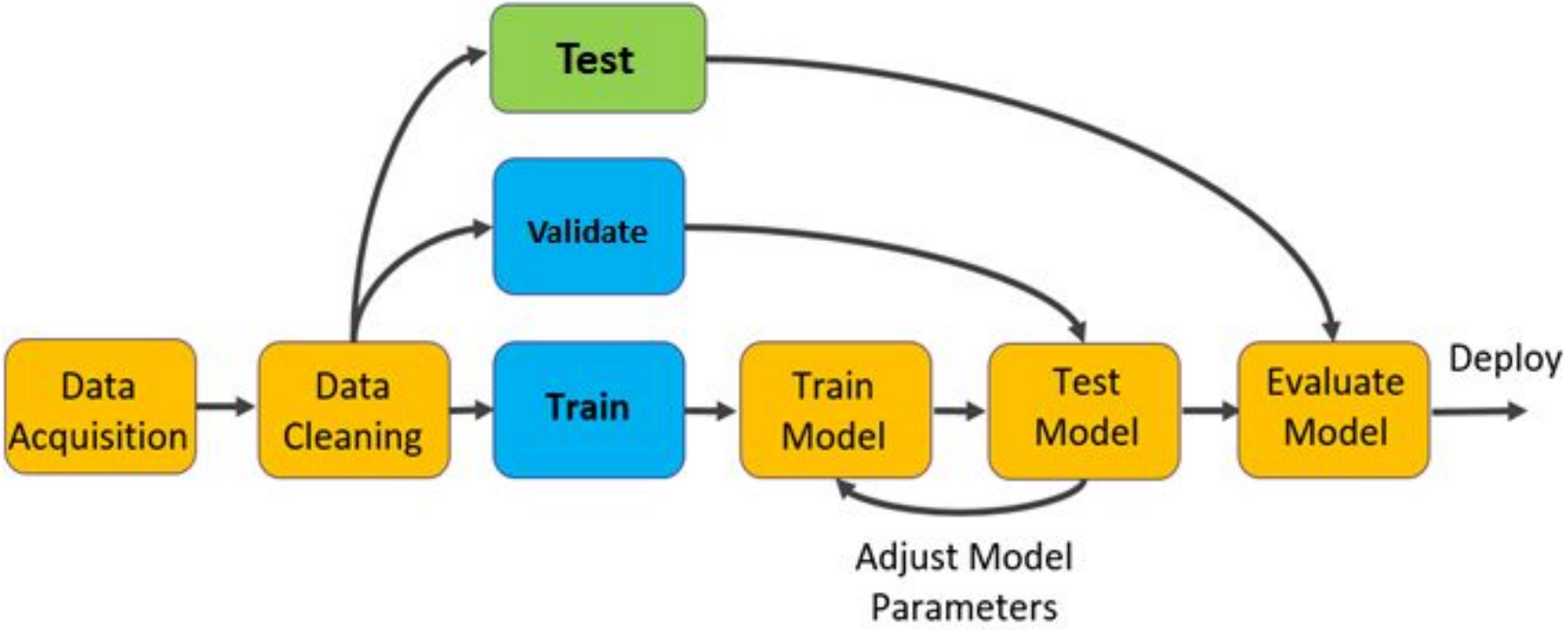
REGRESSION

*Predicting Student Marks*

Percentage

@ dataaspirant.com

# Supervised Machine Learning Workflow



# Train, Validation and Test Dataset

**Learn**
**(Train)**

Practice by doing Exercise
(Validate)

Take Exam
(Test)

Source: DPhi DS Bootcamp Tan

DPhi

# Train, Validation and Test Dataset

Learn
(Train)

**Practice by doing Exercise**
**(Validate)**

Take Exam
(Test)

# Train, Validation and Test Dataset

Learn
(Train)

Practice by doing Exercise
(Validate)

**Take Exam**
**(Test)**

DPhi

# Exams

- What if the paper gets leaked?
- And, what if the university get to know that and the test paper changes in the last minute?

**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.

# Train, Validation and Test Dataset


| | Purpose | Yield | Used for Model training | Used for Parameter tuning |
|---|---|---|---|---|
| **Train Data** | To learn patterns from the data. | A model that makes near-expected predictions | Yes | Yes |
| **Validation Data** | To understand model behaviour and generalizibility on unseen data. | Insights on how to tune your model. | No | Yes |
| **Test Data** | To understand how the model would perform in real world scenario. | A completely unbiased estimate of model performance. | No | No |

# In today's hands on session

For simplicity and to make it beginner friendly, **we will be using train and test split only**. We will get to this later

# Problem Solving

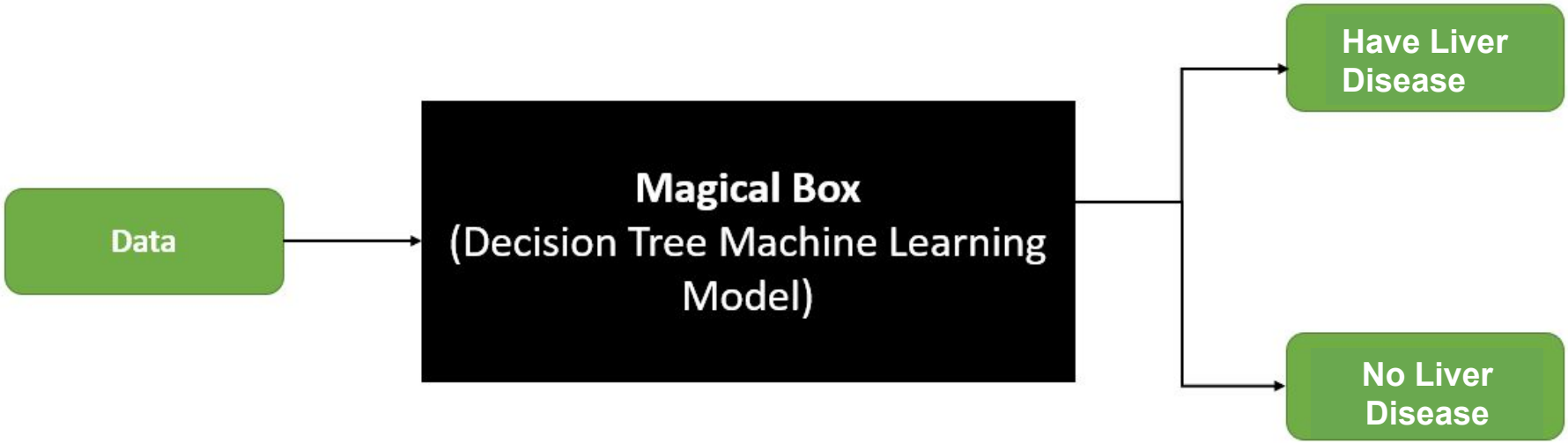
- **Define Objective or understand the problem statement**
- Data Requirements
- Data Collection
- **Exploratory Data Analysis**
- **Data Pre-processing**
- **Build a model**
    - Understand whether it is a regression or classification problem
- **Evaluate**
- **Optimise**
- **Production**
- **Monitor**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

# Objective/Problem Statement

The goal of the model is to **predict whether a patient has liver disease or not**, given a set of data points.

DPhi

# Objective/Problem Statement




DPhi

# Data Requirements & Collection

We have the data!

DPhi

# Understanding the Data

- Age of the patient
- Gender of the patient
- Total Bilirubin
- Direct Bilirubin
- Alkaline Phosphotase
- Alamine Aminotransferase
- Aspartate Aminotransferase
- Total Protiens
- Albumin
- Albumin and Globulin Ratio
- liver_disease: 1 --> Have liver disease, 0 --> No liver disease

# Explore the data

Let’s get to the notebook

DPhi

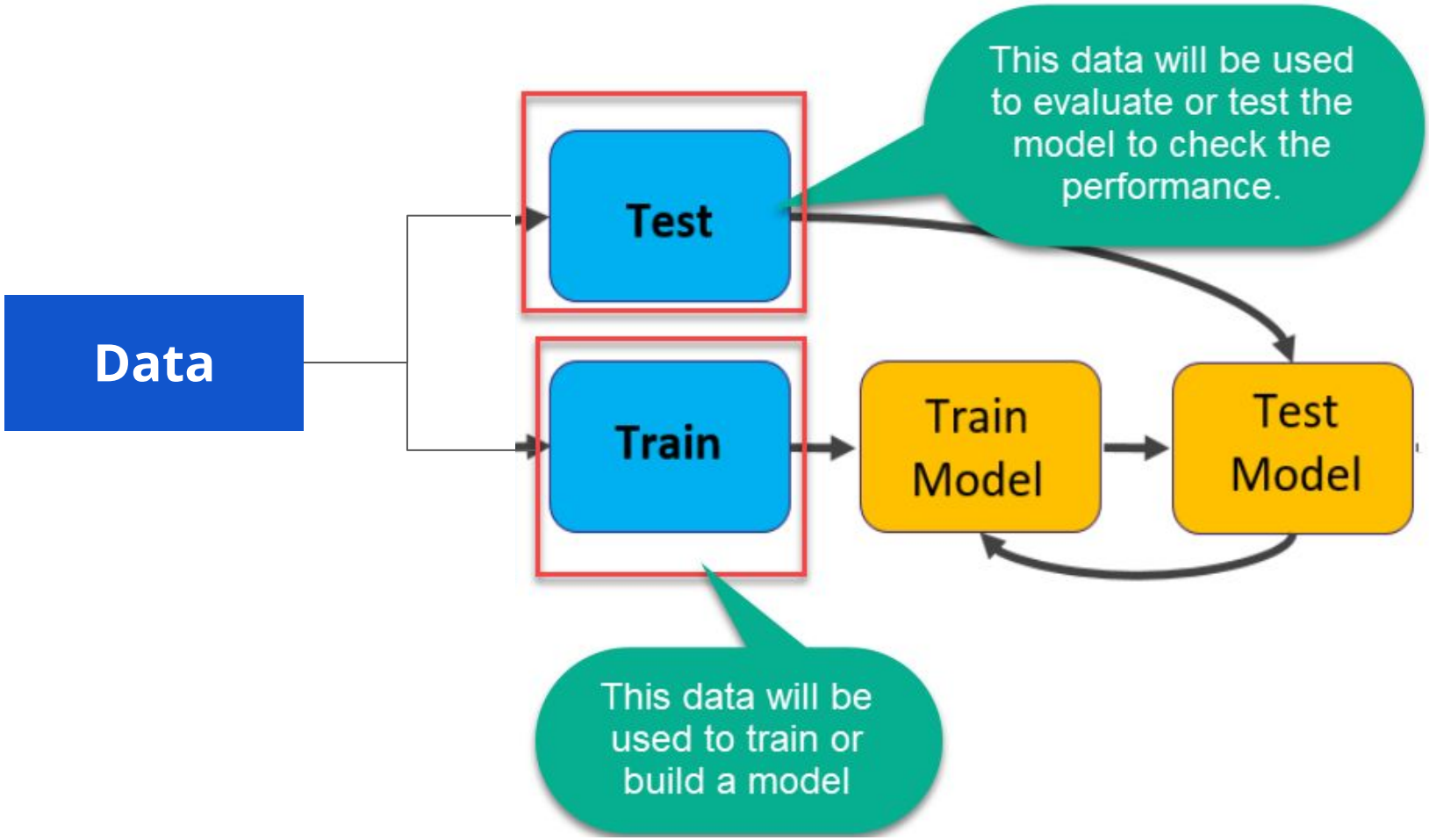
# Missing Values

# Are all data points relevant?

Subject Matter Expertise & Feature Selection

DPhi

# Split the data into train and test

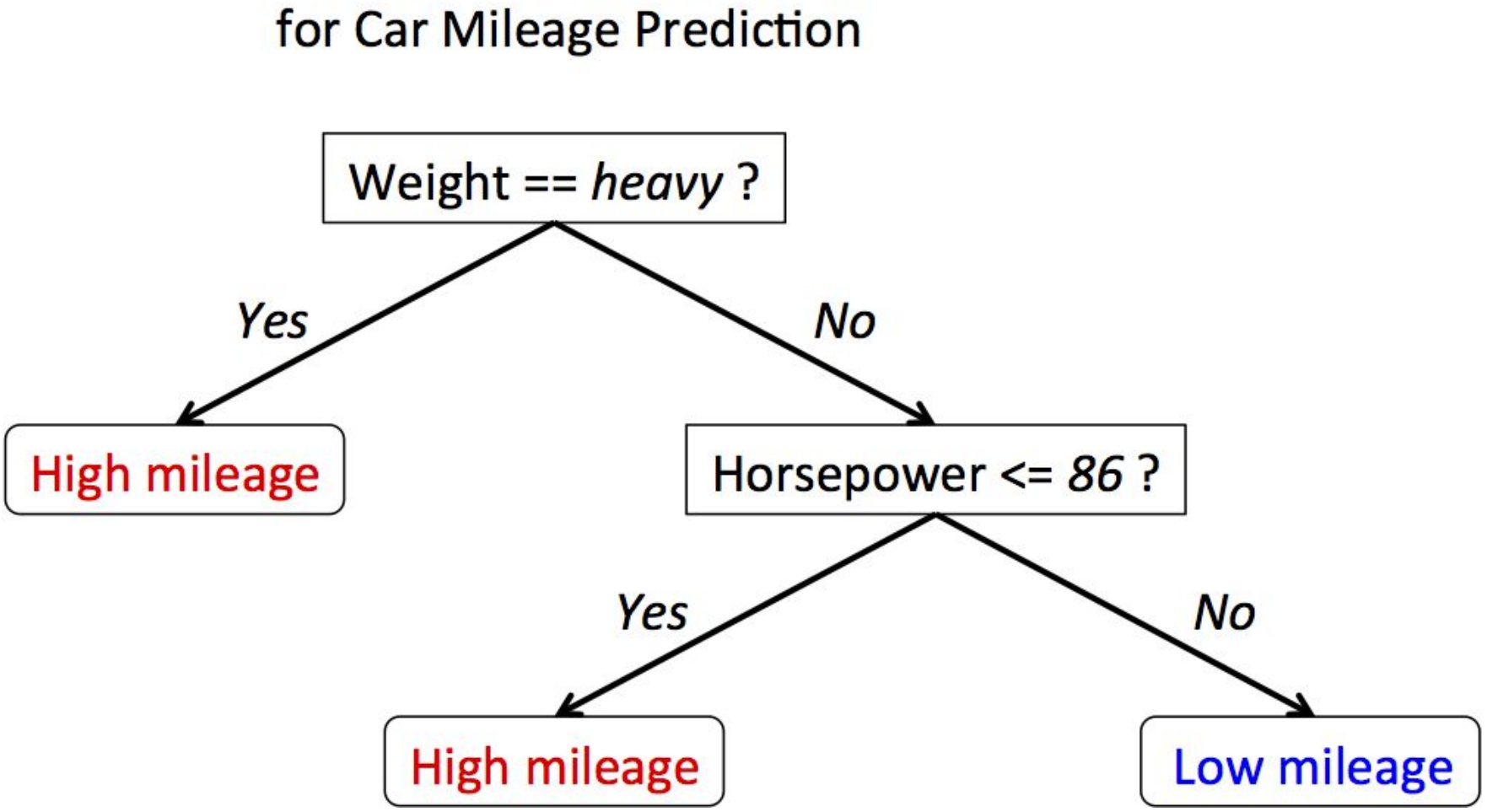


# Model Building - Decision Tree

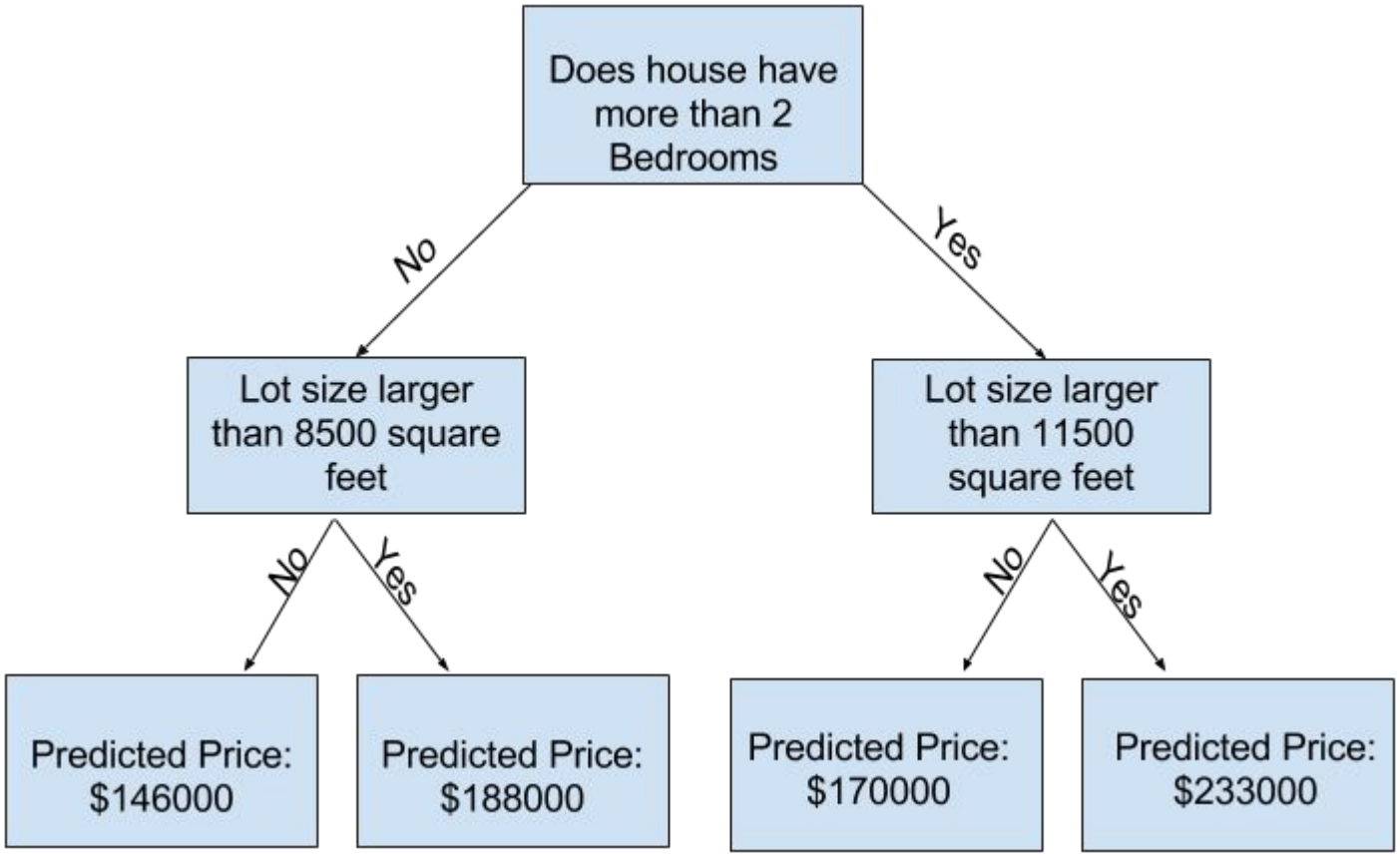
Now what is this decision tree?

Well, we all might have seen it by now!

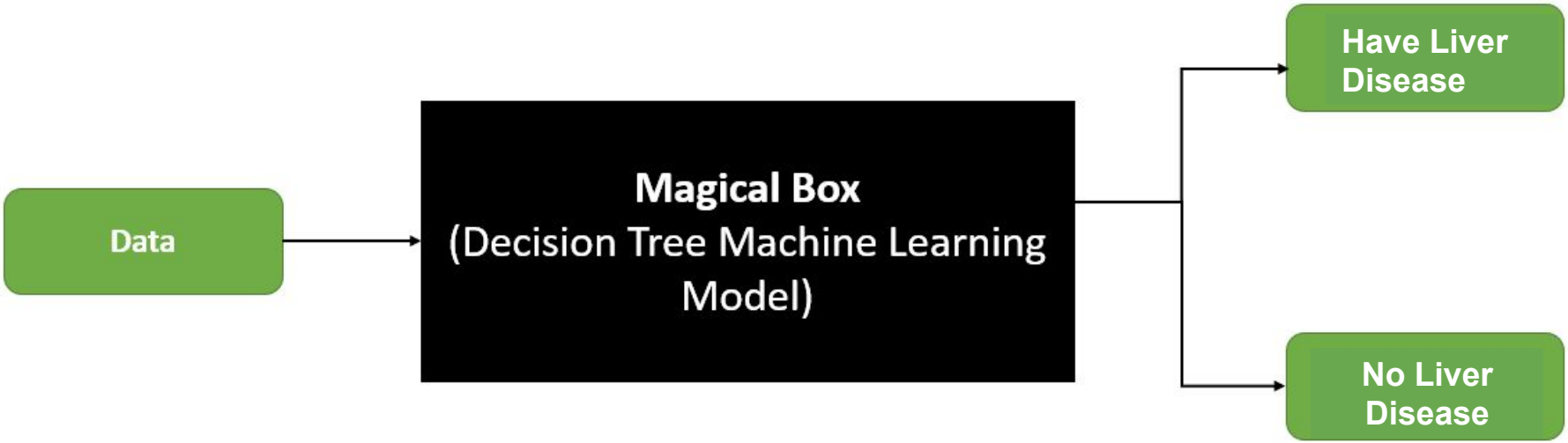
# Decision Tree Examples



# Decision Tree Examples



# Now what next?




Let's do it!

DPhi

# Model Evaluation

**Evaluate on test dataset to check the performance!**

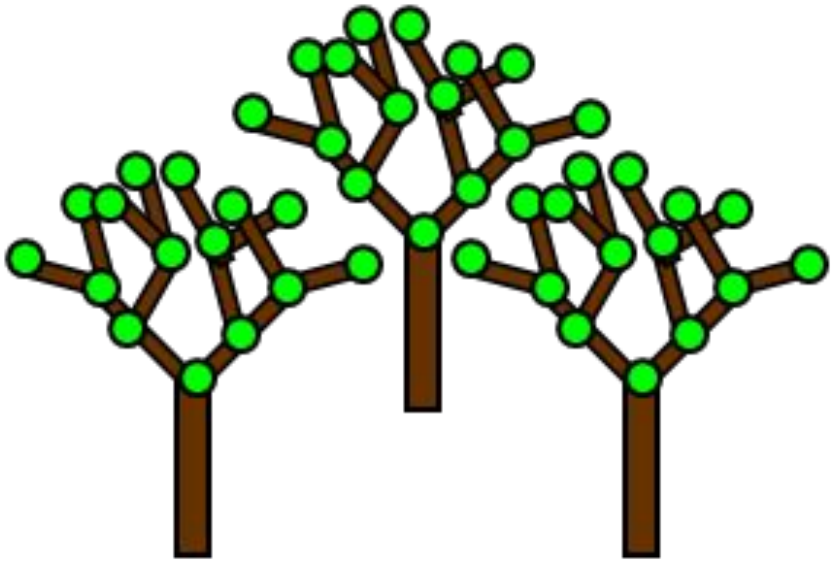
Well, we build a model on historical data and expose them to new data that we would see in future. Technically the model will be exposed to unseen data

# Random Forest

Random forest is a flexible, easy to use machine learning algorithm that produces, a great result most of the times even without hyper-parameter tuning.

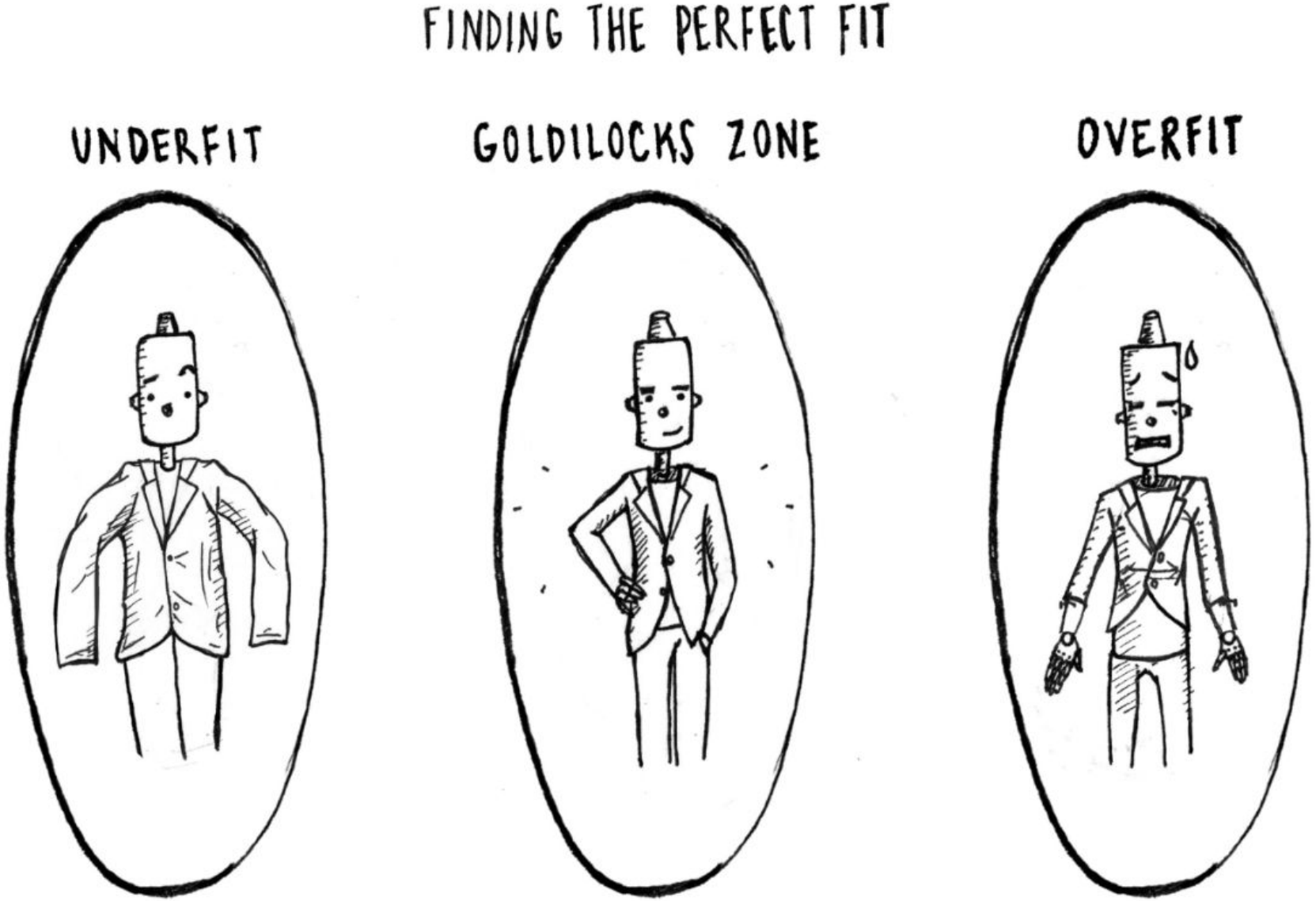
It is also one of the most used algorithms, because of its simplicity and diversity (**can be used for both classification and regression tasks**).

Random forest **builds multiple decision trees and merges them together** to get a more accurate and stable prediction.

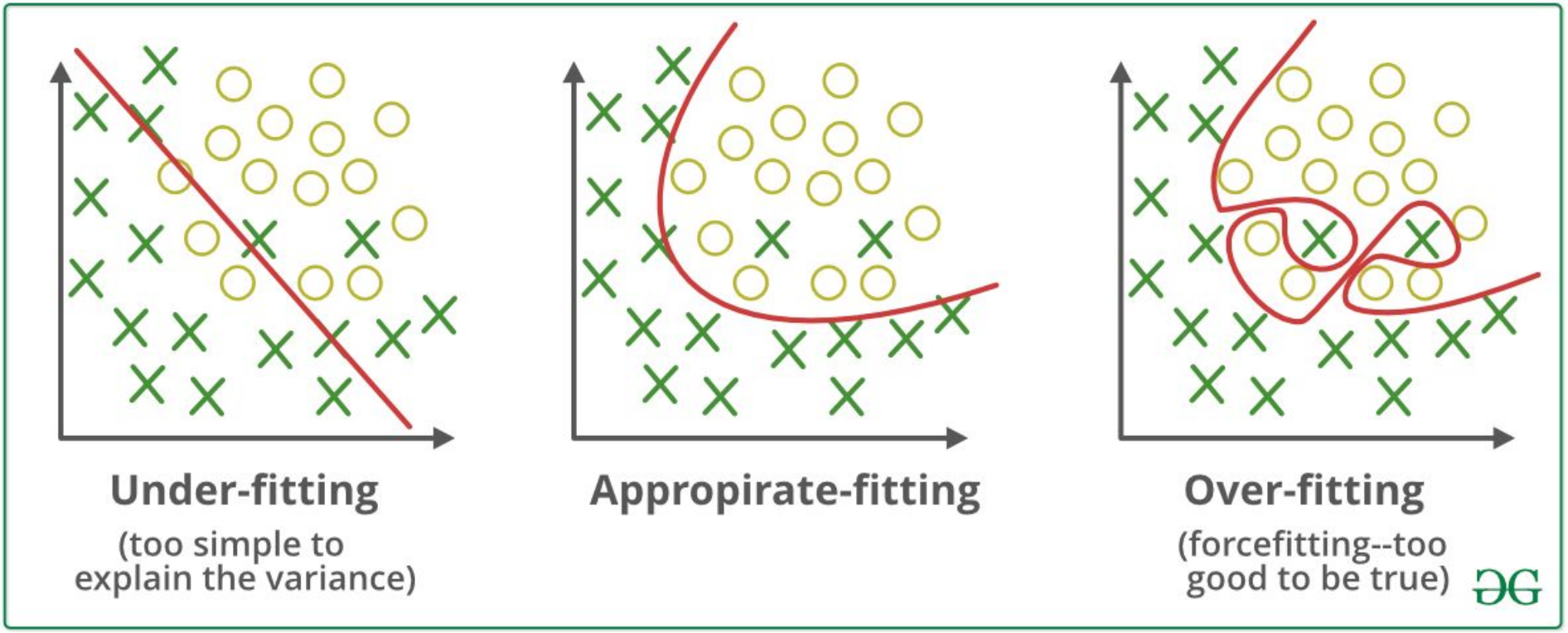


DPhi

# Overfitting - Underfitting



# Overfitting - Underfitting



# Model Evaluation

!! We are not done yet,

We can improvise it significantly.

# 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.

# What are we achieving out of it?

- Technically an efficient model can **empower doctors who are treating Liver Patients**
- **Avoid Errors:** Feeding in the patient records could help assess doctors to understand the significance of the disease.
- Practical Advanced Use-Case: Parkinson Disease using Kinect
- What is Parkinson Disease? Disorder of the central nervous system that affects movement. Patients often visit doctors in a few months and their concerns are usually influenced by recency factors.

ML/Data Science can empower people to arrive at a meaningful decision using Data.

That’s it for the day. Thank you!

**We can stay connected:**

Linkedin: https://www.linkedin.com/in/chanukyapatnaik/

Twitter: @chanukya_p

Medium: @chanukya

DPhi