Data Analytics

By:
Divyansh Srivastava

Data Scientist
Decimal Point Analytics,
Mumbai, India.

Bachelor of Technology '22
IIT Ropar,
Punjab, India

Introduction

Data is a collection of facts, such as numbers, words, measurements, observations or just descriptions of things.

Data is getting generated at a massive rate, by the minute

Organizations are trying to explore every opportunity to make sense of this data.

What is Data Analytics?

- It is the process of exploring and analyzing large datasets to make predictions and boost data-driven decision making.
- Data analytics allows us to collect, clean, and transform data to derive meaningful insights.
- It helps to answer questions, test hypotheses, or disprove theories.

Types of Data Analytics

Descriptive Analytics - It tells you what has happened, done by exploratory data analysis.

Predictive Analytics - It tells you what will happen, achieved by building predictive models.

Prescriptive Analytics - It tells you how to make something happen, done by deriving key insights and hidden patterns from the data.

Steps involved in Data Analytics

Data Collection – through databases, web servers, social media etc

Data Preparation - remove unwanted and redundant values, converting it into the right format

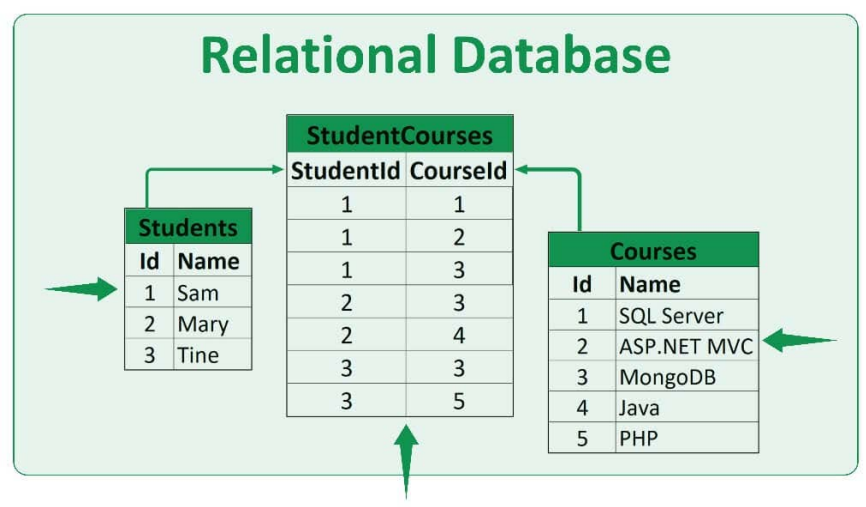
Data Exploration - various data visualization techniques to find unseen trends in the data

Data Modeling - predictive models using machine learning

Result interpretation - derive meaningful results and deployment

Data Collection - Databases

Key	Document
1001	{ "CustomerID": 99, "OrderItems": [{ "ProductID": 2010, "Quantity": 2, "Cost": 520 }, { "ProductID": 4365, "Quantity": 1, "Cost": 18 }], "OrderDate": "04/01/2017" }
1002	{ "CustomerID": 220, "OrderItems": [{ "ProductID": 1285, "Quantity": 1, "Cost": 120 }], "OrderDate": "05/08/2017" }



It is an organized collection of data stored and accessed electronically

- A database is usually controlled by a Database Management System (DBMS).

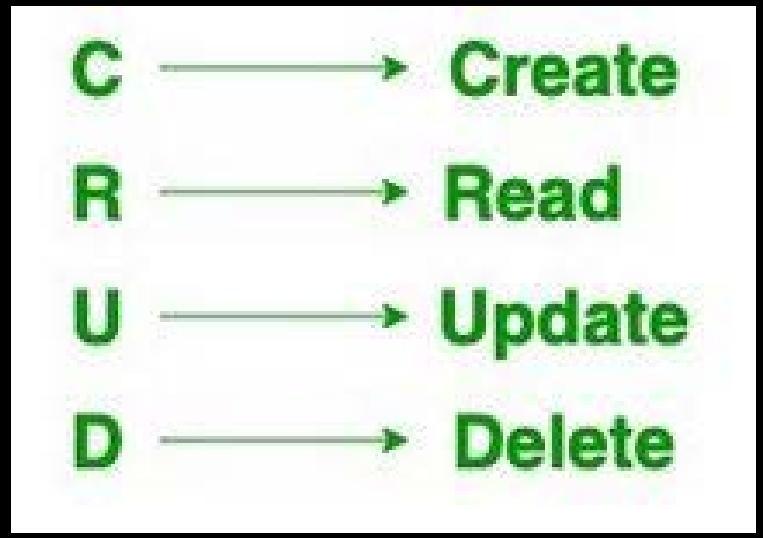
Two of the commonly used databases are:

- ***Relational Databases*** - Items in RDB are organized as a set of tables with columns and rows. It provides the most efficient and flexible way to access structured information.
- ***Non-Relational Databases*** - allows unstructured and semi structured data to be stored and manipulated. Useful for more complex data manipulation in webpages etc.

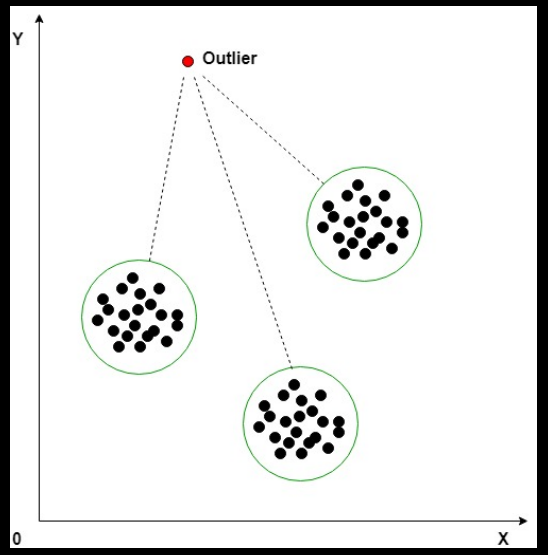
Structured Query Language (SQL)

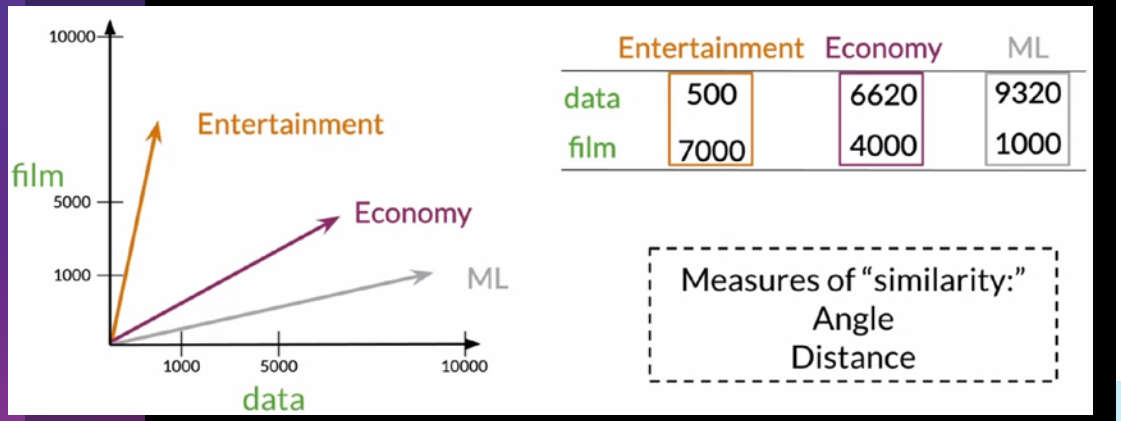
```
CREATE TABLE Employee (EMPId Int Identity NOT NULL,
 EmpNo varchar (10), SSN varchar (10), DOB DATE,
 CreatedDt Datetime, CreatedBy varchar(10));

INSERT INTO Employee (EmpNO, SSN, DOB, CreatedDt,
CreatedBy)
VALUES (1,'1234567890','2000-01-01', GETDATE(), 'system');

SELECT EmpNo, SSN, DOB
FROM Employee
WHERE EmpNo = '1';

UPDATE Employee
SET EmpNo = '3', SSN = '4984564512', DOB = '1998-01-01'
WHERE EmpId = 3;

DELETE FROM Employee
WHERE EmpId = 10;
```

	EMPId	EmpNo	SSN	DOB	CreatedDt	CreatedBy
1	1	1	1234567890	2000-01-01	2021-10-22 17:38:17.133	system
2	2	2	0123456789	1999-01-01	2021-10-22 17:39:58.517	system
3	3	3	4984564512	1998-01-01	2021-10-22 17:40:01.947	system
4	4	4	1231544984	2000-02-01	2021-10-22 19:45:10.147	system
5	5	5	5487946598	2001-01-01	2021-10-22 19:45:10.147	system
6	6	6	8789453115	2002-01-01	2021-10-22 19:45:10.147	system
7	7	7	4598135651	2003-02-01	2021-10-22 19:45:10.147	system
8	8	8	8979845654	2004-01-01	2021-10-22 19:45:10.147	system
9	9	9	7897165468	2005-01-01	2021-10-22 19:59:49.813	system
10	10	7	4598135651	2003-02-01	2021-10-22 19:59:49.813	system
11	11	8	8979845654	2004-01-01	2021-10-22 19:59:49.813	system
12	12	9	7897165468	2005-01-01	2021-10-22 19:59:49.813	system



Data Preparation / Cleaning





	Entertainment	Economy	ML
data	500	6620	9320
film	7000	4000	1000

Measures of "similarity:"
Angle
Distance

DATA CLEANING CHECKLIST

Up-to-date data

Data should be up-to-date in order to obtain maximum value from the data analysis.

Missing values

Count missing values and analyze where in the data they are missing. Missing values can disrupt some analyses and skew the results.

Duplicates

Duplicate IDs indicate multiple records for one person, e.g. someone holds multiple functions at the same time.

Numerical outliers

Numerical outliers are fairly easy to detect and remove. Define minimum and maximum to spot outliers easily.

Check IDs

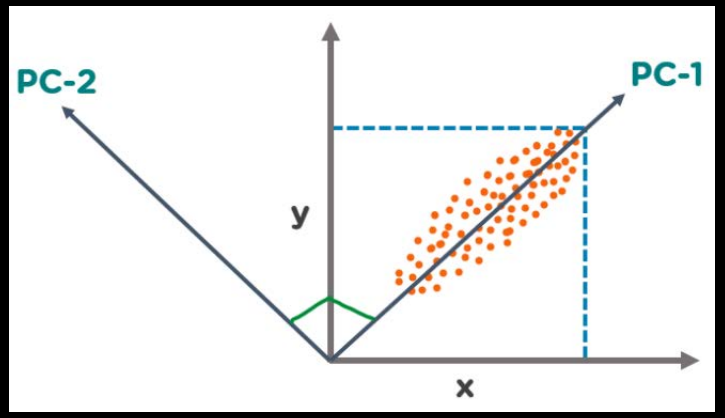
Check data labels of all the fields to see whether some categorical values are mislabeled.

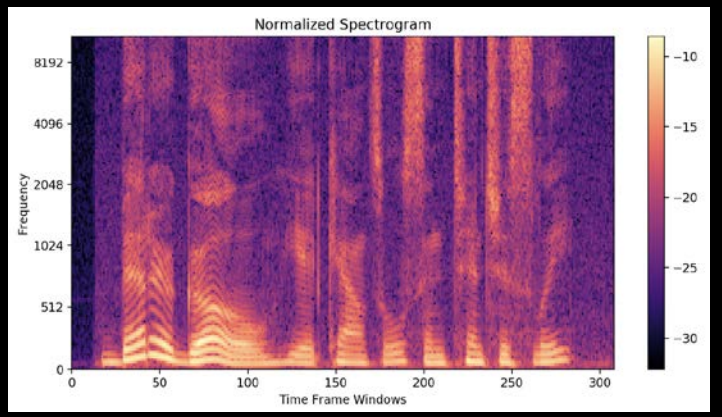
Define valid output

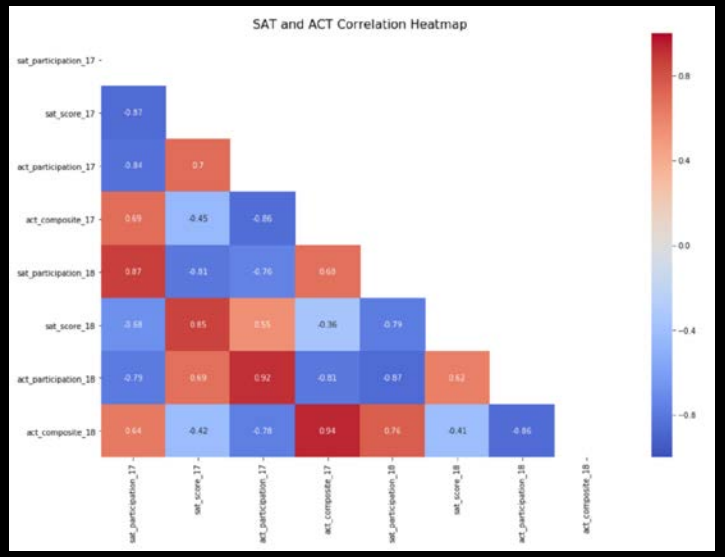
Define valid data labels for categorical data. Define data ranges for numerical variables. Non-matching data is presumably wrong.

AIHR
BLOG & ACADEMY

Data Exploration

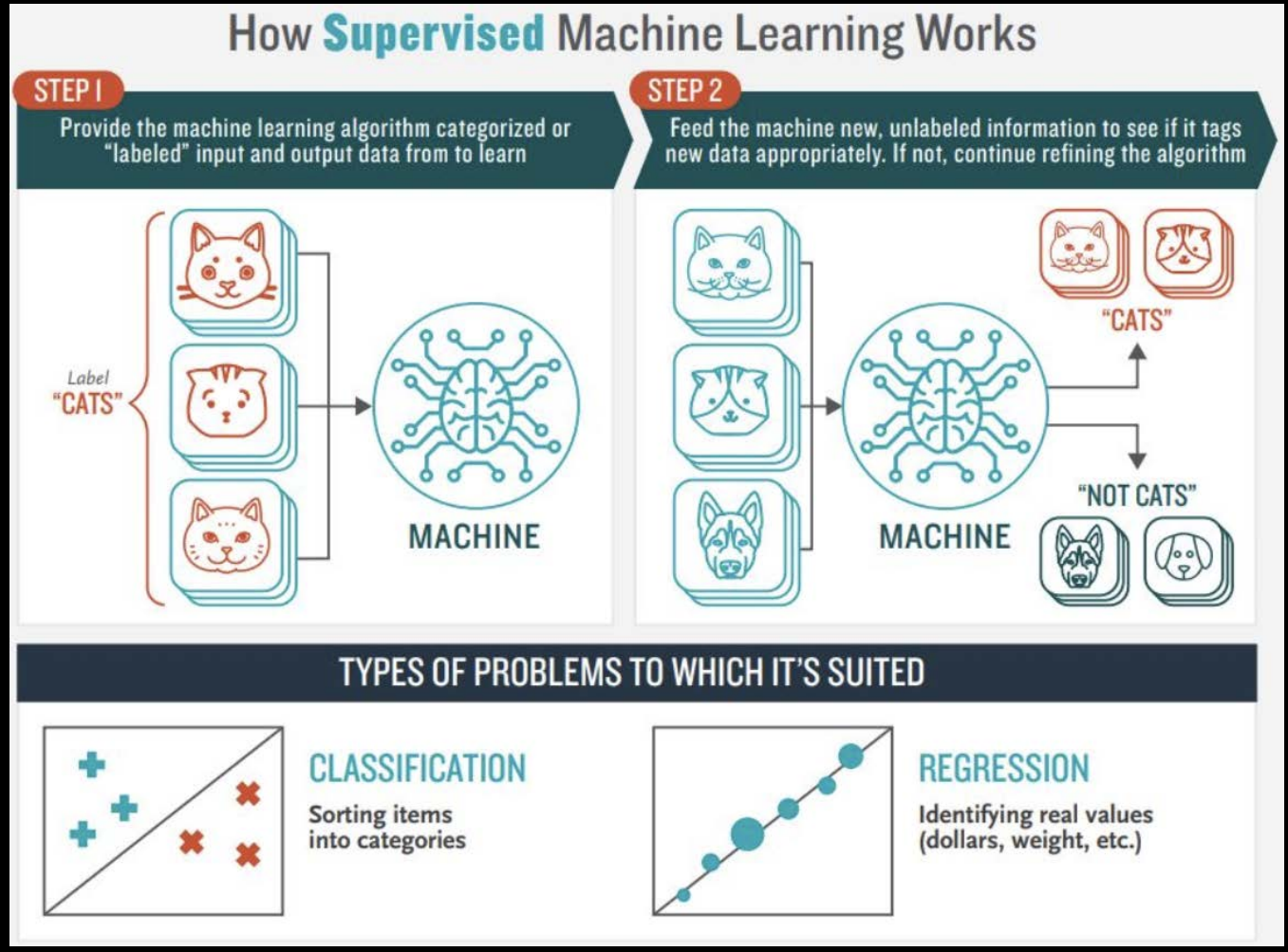




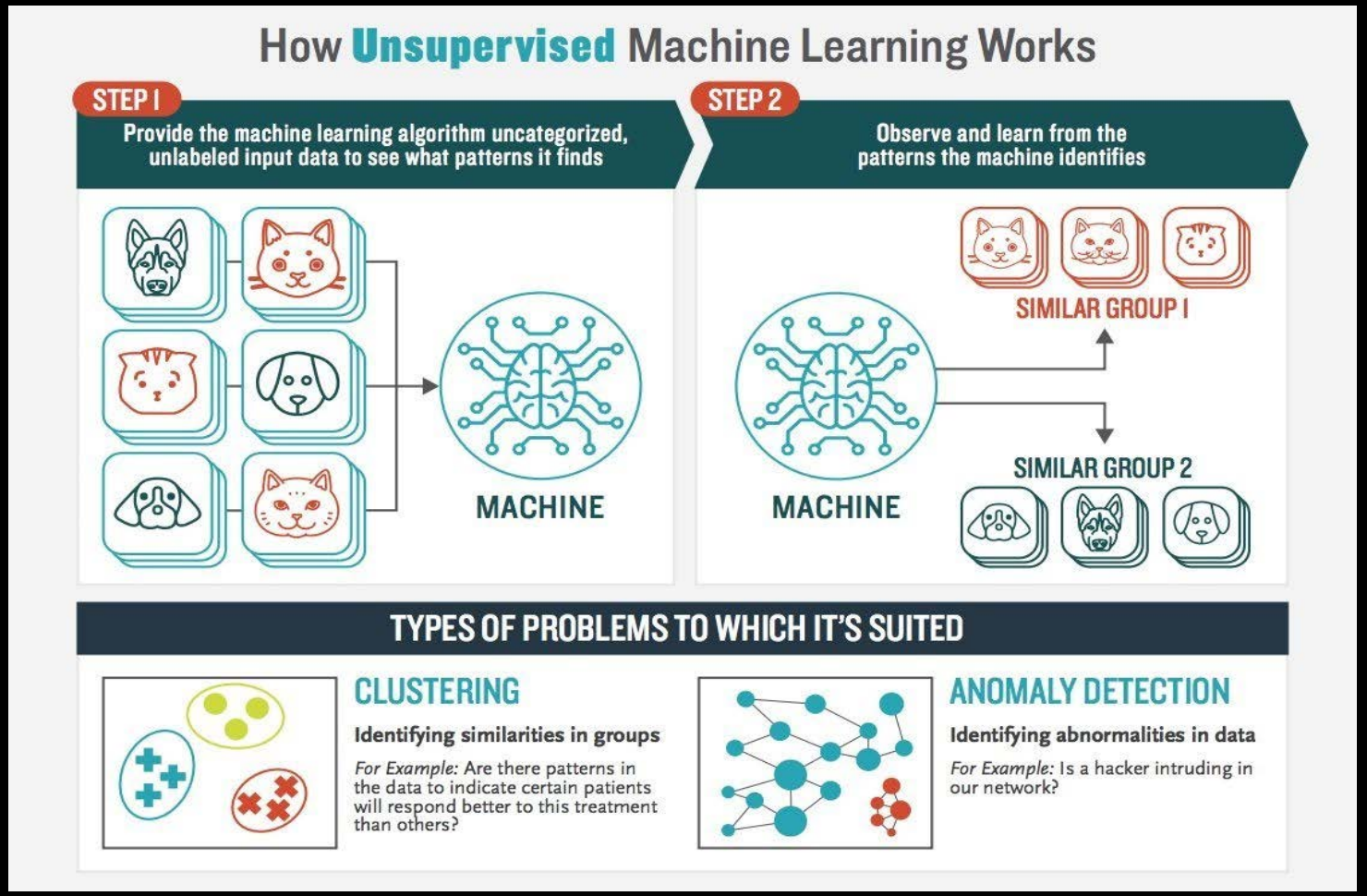


Predictive models Supervised Learning

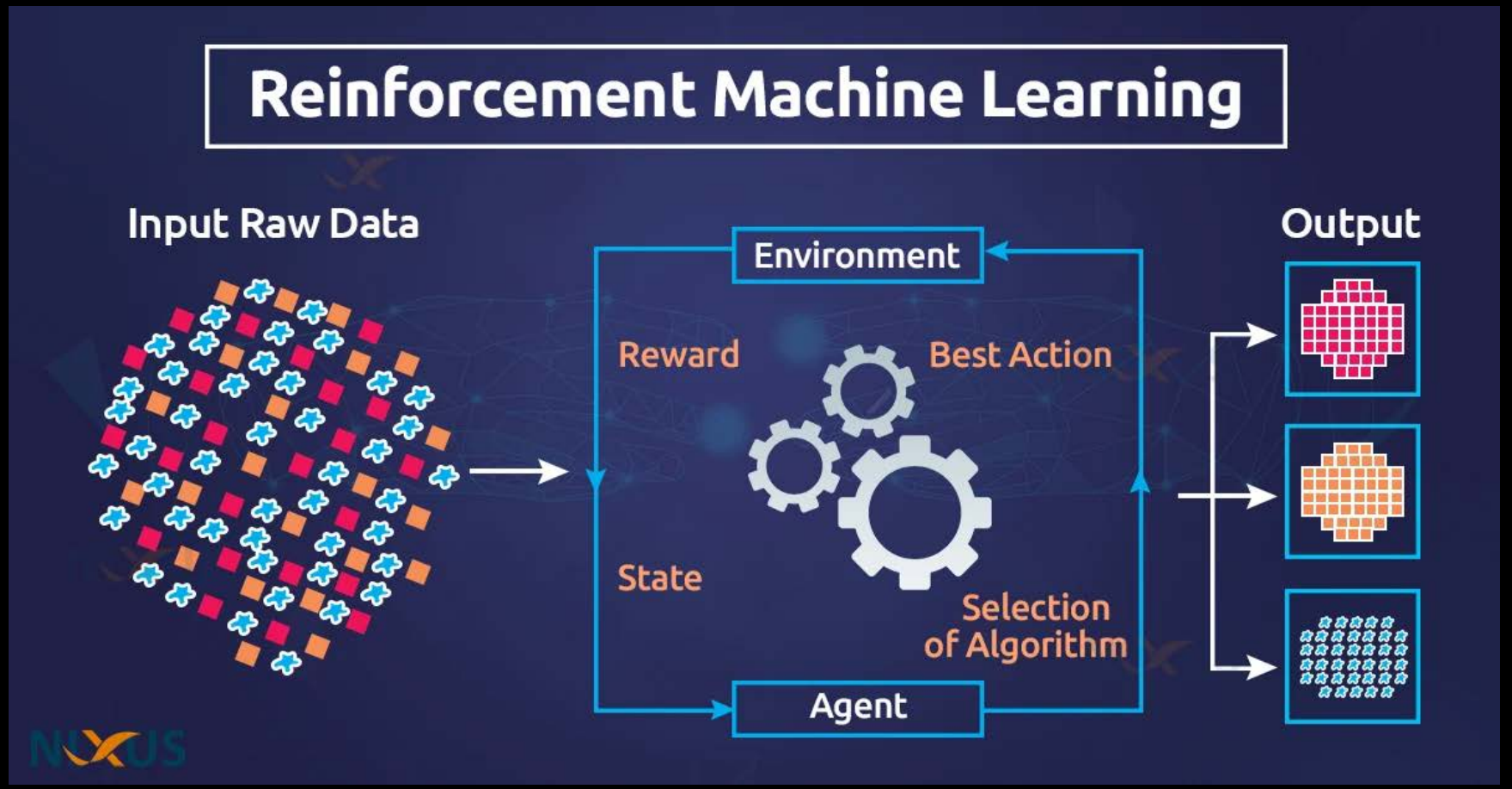
- Initially model knows nothing
- The objective is to minimize the loss function which represents the difference between the actual label and the label predicted by the model.
- Multiple iterations are run on the model in order to improve accuracy.



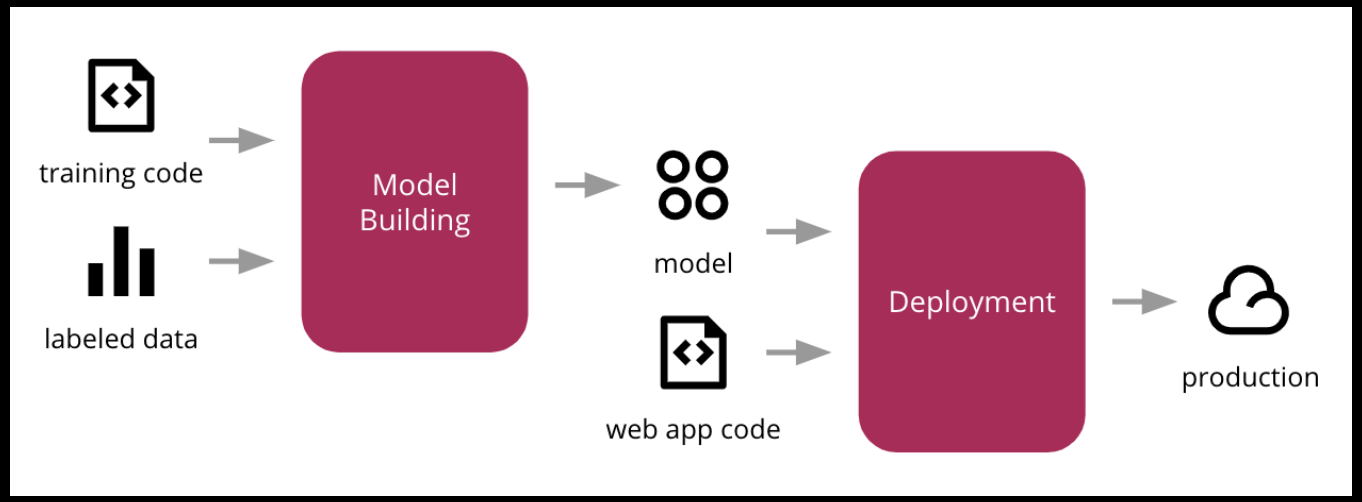
Predictive models
Unsupervised Learning

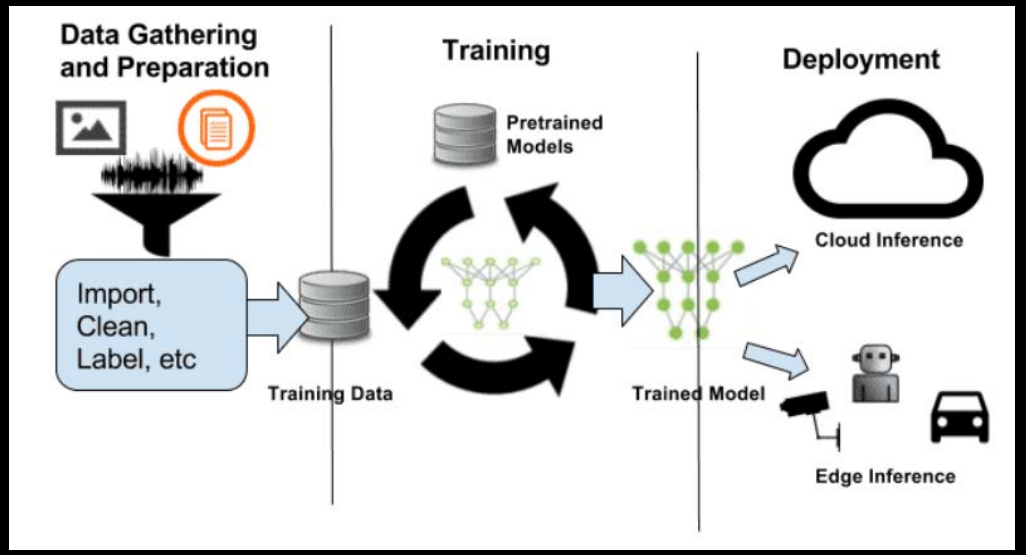


Predictive models
Reinforcement Learning



Deployments





Applications

Banking and e-commerce industries to detect fraudulent transactions.

Healthcare sector uses data analytics to improve patient health by detecting diseases before they happen

Data analytics finds its usage in inventory management to keep track of different items.

Logistics companies use data analytics to ensure faster delivery of products by optimizing vehicle routes.

Marketing professionals use analytics to reach out to the right customers and perform targeted marketing.

Thank You

In case of any queries please connect with me through the following email:

divyanshsrivastava1009@gmail.com