

Big Data Fundamental



Agenda

[Day 2]

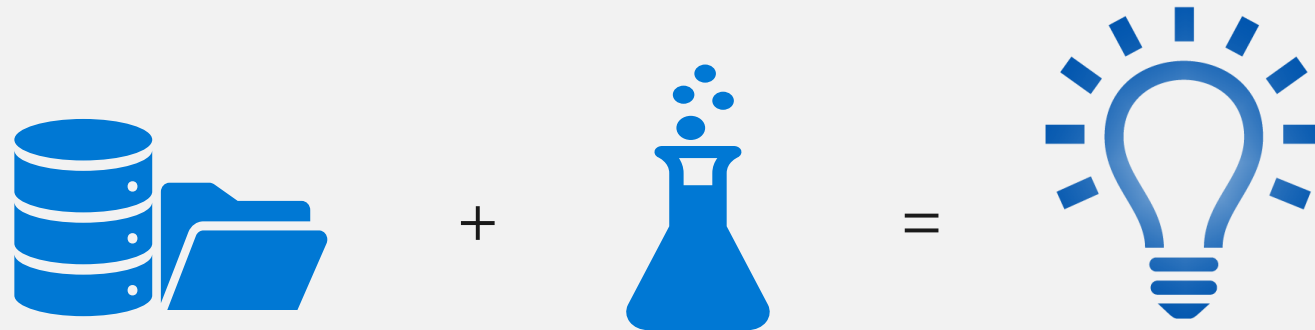
- Data Science
- Machine Learning
- Types of Machine Learning
- Steps to do Machine Learning / Predictive Analytics
- Performance Matrix
- Data Analytics (Hands-on)
- Big Data – Career Path

Resource Link

<http://arif.works/bdf/>

What is Data Science?

Apply **Scientific Methods** to extract **Knowledge** from **Data**.

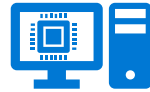


Scientific Methods



Statistics

Designed for inference about the relationships between variables



Machine Learning

Designed to make the most accurate predictions possible



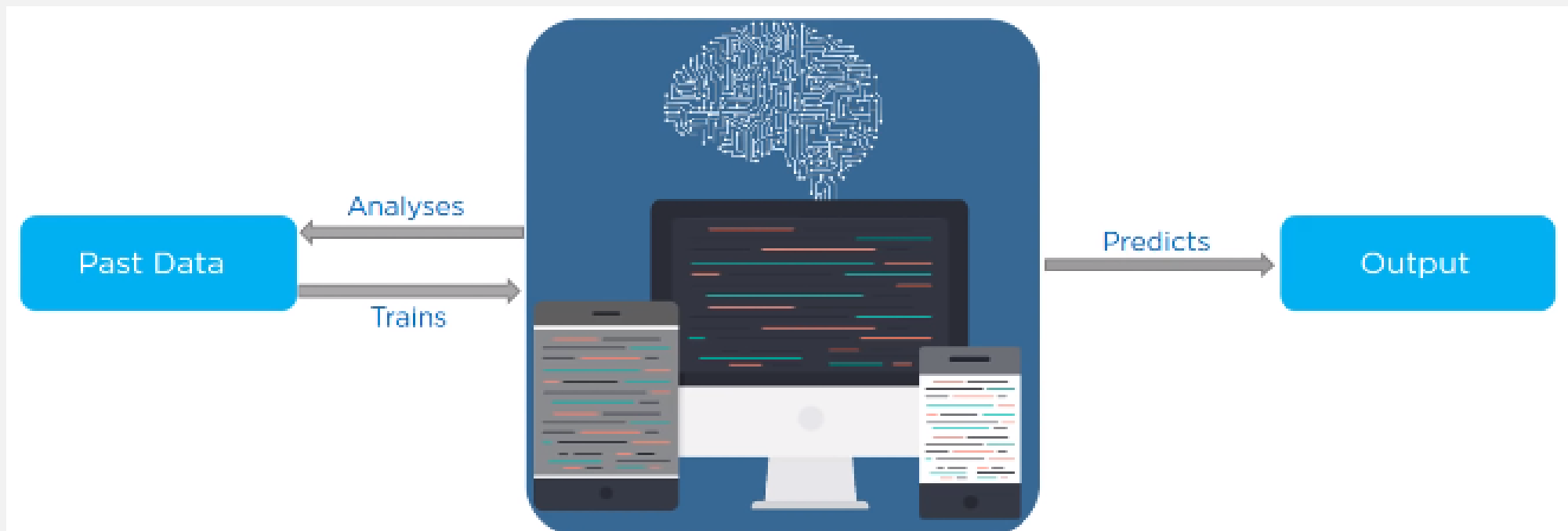
Artificial Intelligence

Designed to mimic human behavior using ML and Deep Learning



Machine Learning

Machine (computer) tries to find the pattern (self-learn) from the data without being explicitly programmed.



Types of Machine Learning

Supervised

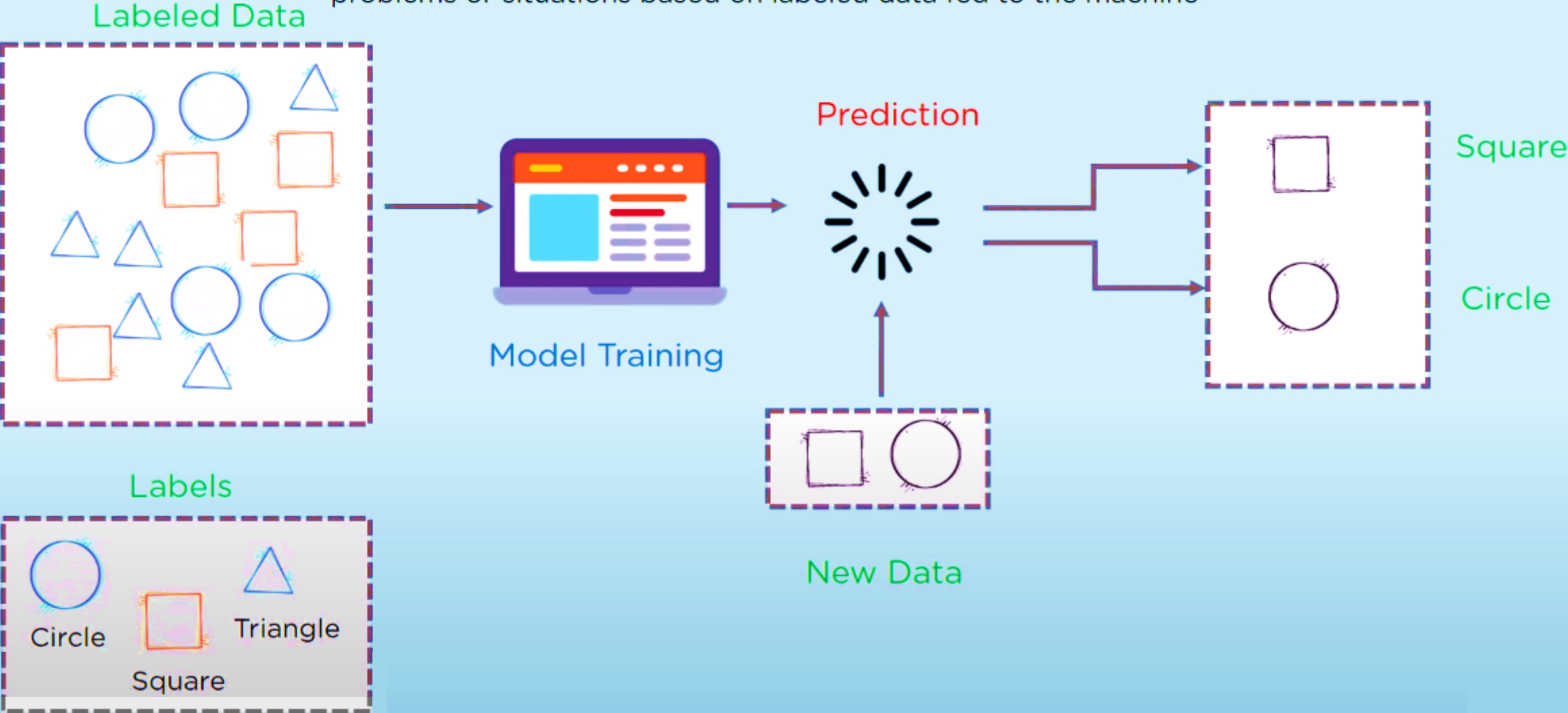
Reinforcement

Un-Supervised



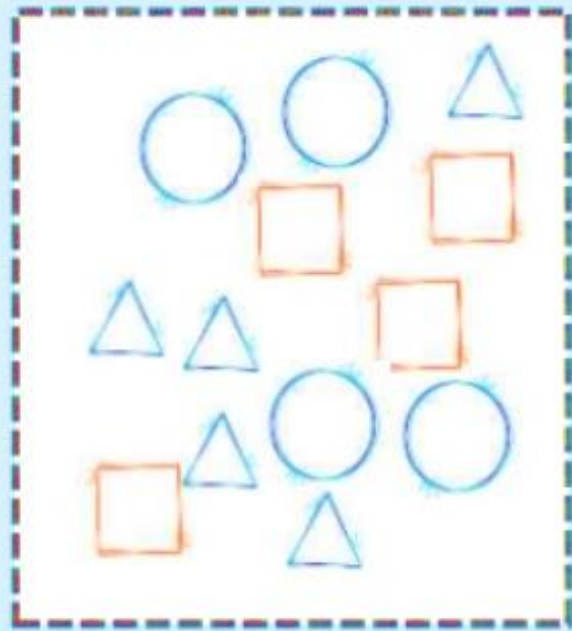
Supervised Learning

Supervised learning is a method used to enable machines to classify/ predict objects, problems or situations based on labeled data fed to the machine



Unsupervised Learning

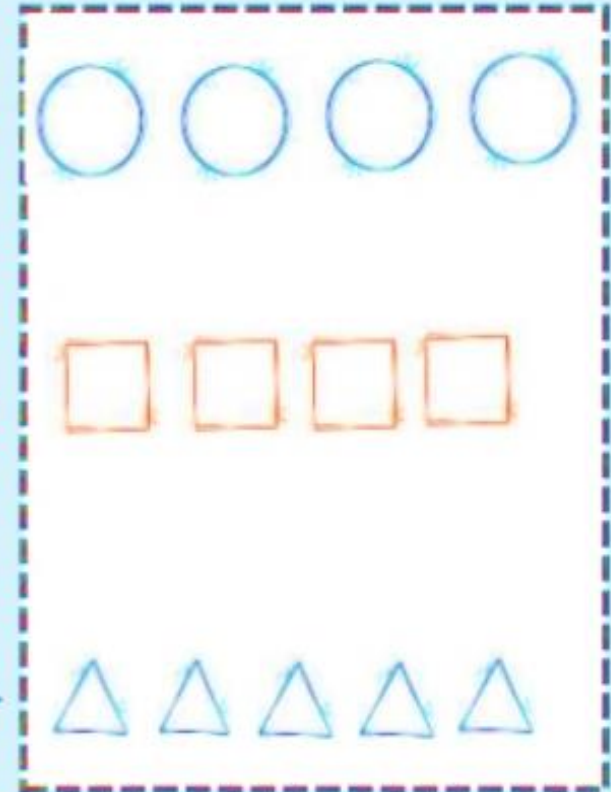
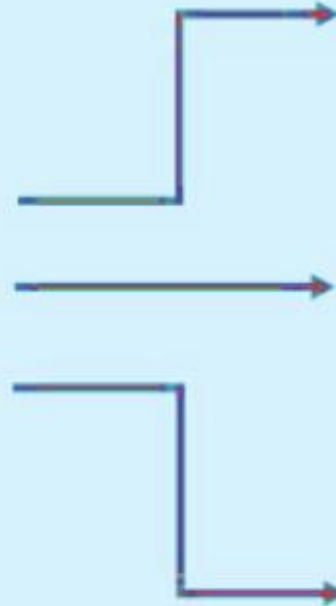
In Unsupervised learning, Machine Learning model finds the hidden pattern in an unlabeled data



Unlabeled Data



Model Training



Output

Types of Machine Learning

Classification

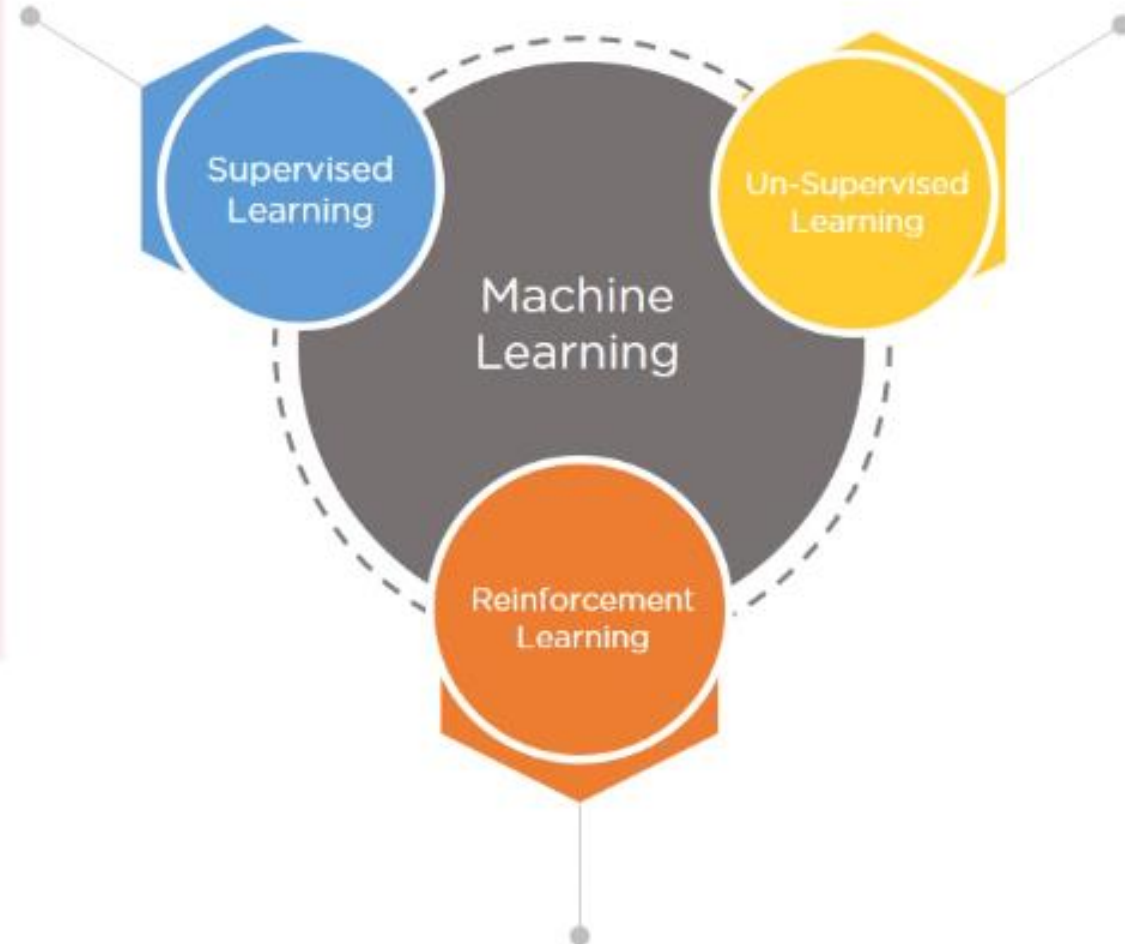
- Fraud Detection
- Email Spam Detection
- Image Classification

Categorical

Regression

- Weather Forecasting
- Risk Assessment
- Score Prediction

Numerical



Types of Machine Learning

Supervised learning, algorithms are trained using marked data, where the input and the output are known.

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

labels

⚙ Set of inputs ~ [Features] / [Independent Variables] / [X]

⚙ Outputs ~ [Labels] / [Dependent Variables] / [Y]

Types of Machine Learning

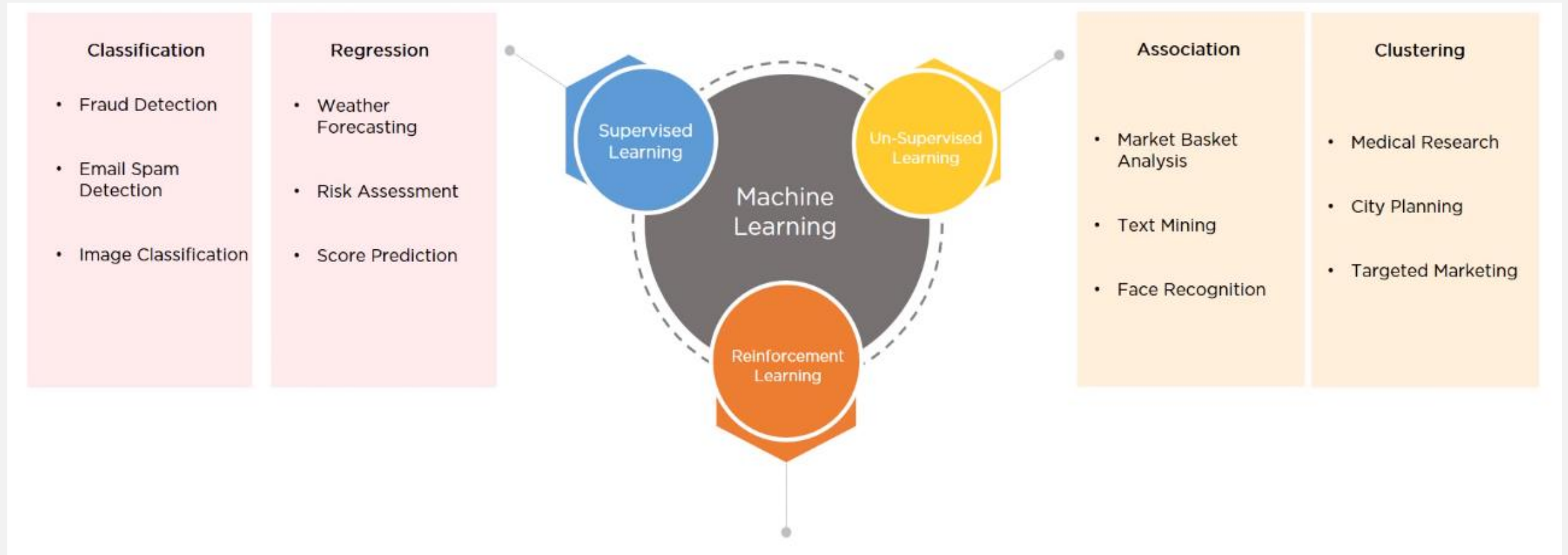
User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

Figure A: CLASSIFICATION

Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.7150879	3.278597116
13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
24.23155922	988.796875	19.74790765	318.3214111	0.329656571

Figure B: REGRESSION

Types of Machine Learning



Types of Machine Learning

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

unlabeled



Types of Machine Learning



Labeled Data



Direct feedback



Predict output

Supervised

vs

Unsupervised



Non-labeled data

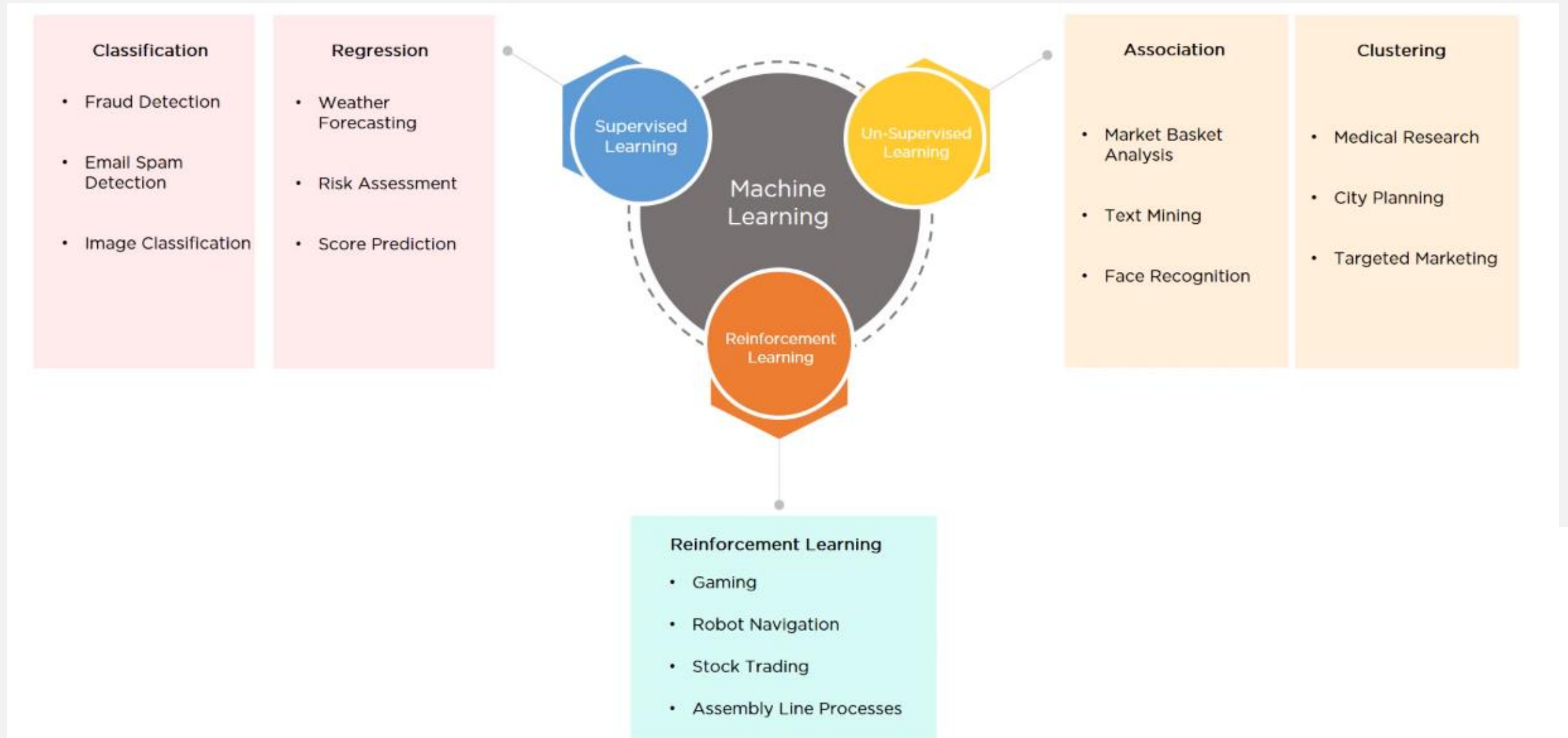


No feedback



Find hidden structure in data

Types of Machine Learning



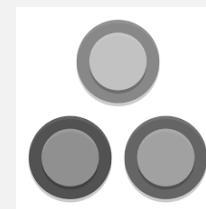
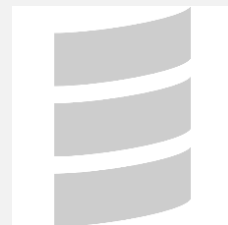
Processing Steps for Machine Learning



Major ML supported Languages

Python / R / Java / Scala / Spark / Julia / **No Code**

These language provide all necessary ML packages



Hands-on

Supervised Machine Learning

IDE : No Code Machine Learning with [Azure Machine Learning Studio \(Classic\)](#)

Hands-on

Step 1 : Please go to this site <https://studio.azureml.net/>

Step 2 : Use any Microsoft Account to Register and Login

Step 3 : Let's do some Prediction

Big Data – Career Path

DEMAND

Big Data is going to have an impact on global GDP of **\$15 Trillion** by 2030.

Big Data market is predicted to be worth **\$47 Billion** by 2018.

By 2018, the US alone is going to face a shortage of **140K to 190K** people with deep analytical skills.

1.5 Million data managers will be needed by 2018.

AREAS WHERE BIG DATA IS USED	REQUIRED TECHNICAL SKILLS
<p> Farmers around the world are using sensor data to reinvent their farms.</p>	<div style="display: flex; justify-content: space-between;"> <div style="width: 45%;"> <p>1 Apache Hadoop</p> <p>3 NoSQL</p> <p>5 Machine learning and data mining</p> <p>7 Statistical and quantitative analysis</p> <p>9 Data visualization</p> </div> <div style="width: 45%;"> <p>2 Apache Spark</p> <p>4 SQL</p> <p>6 Creativity and problem solving</p> <p>8 General purpose programming languages</p> </div> </div>
<p> A building in the United Arab Emirates uses data to produce more energy than it consumes.</p>	
<p> Taxis in Sweden use data to cut traffic and auto emissions.</p>	
<p> Barcelona is harnessing data to build a smarter city.</p>	

- 1

Data Scientist

Data Scientists make value out of data, which has been obtained from various sources by getting meaningful insights from it. They need to have proper analytics and technical capabilities to perform such tasks. A Data Scientist can gain an average salary in the range of US\$90,000-1,100,000 per year.
- 2

Big Data Engineer

Big Data Engineers mainly focus on developing, maintaining, testing, and implementing network Big Data projects. Also, they are useful for building designs that have been offered by Solutions Architects. A Big Data Engineer can gain up to US\$95,000-150,000 annually.
- 3

Big Data Architect

Big Data Architects address any significant data problems, along with the elements. They explain the structure and behavior of a Big Data solution using a technology they are specialized in. Big Data Architects are included as the link between an organization and its Data Engineers. A Big Data Architect has an average annual salary of US\$144315.
- 4

Business Analytics Specialist

Business Analytics Specialists help in different testing activities, as well as in some development initiatives. They usually come up with unique, cost-effective solutions for solving business issues. A Business Analytics Specialist is capable of getting somewhere around US\$78,819 a year.
- 5

Data Visualization Developer

The role of a Data Visualization Developer and Analyst is to detail, design, and deliver a production guide for data visualizations to be used across an enterprise. Data Visualization Developers attach to any visual representation of data for explaining the importance of it. On average, a Data Visualization Developer earns US\$123,039 per year.