Understanding the AI and DATA Landscape

Dr. Shailesh Sivan

Cochin University of Science and Technology

Hi

I'm Dr. Shailesh Sivan

B.Tech CSE, M.Tech CIS, M.Sc Mathematics, Ph.D. AI/ML

My Philosophy

Inspire Yourself, Uplift Others

Assistant Professor | Research Guide

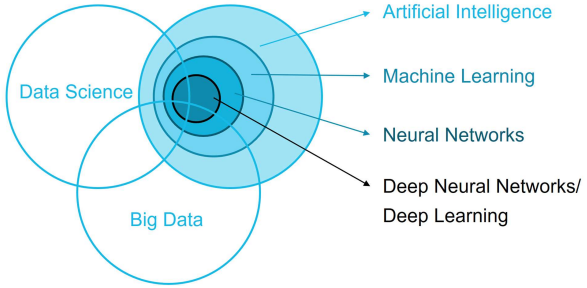
Former Sofware Engineer| Lifetime Techie

AI/ML Enthusiast | C/C++ Advocate | Algorithms & Maths

OVERVIEW

- Confused?
- Data Science and Big Data
- What is machine learning?
- Programming vs Learning
- Machine Learning Pipeline
- When to use Machine Learning?
- ML Application
- Types of Learning

CONFUSED ?



COMPUTER SYSTEMS

- An electronics device
- Capable of storing and processing data
- Can be used for controlling other devices

One which do computing

ROBOTS AND COMPUTER

Are Computers Intelligent or Dumb ?

COMPUTERS ARE DUMB ⊗ !

- Computer is not a magical device.
- It performs only those works which man can do!
- But with very high speed and reliable accuracy.
- It has no intelligence quality or thinking power

DATA, INFORMATION & KNOWLEDGE

Elephant

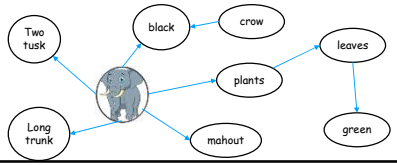
vegetarian
big
trunk
tusk
black
mahout

Data

- Biggest animal in the land
- Has two tusk
- Black in Colour
- Has a long trunk
- Eats plants
- Mahout is the keeper

Information

Knowledge



ARTIFICIAL INTELLIGENCE

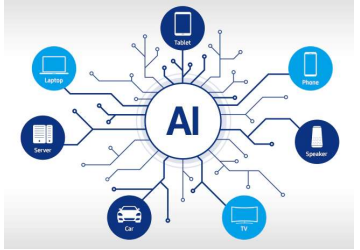
Can We make Computers Intelligent ?

Yeah Of course !

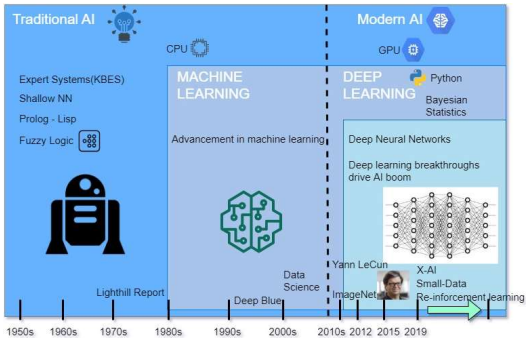
\+ Knowledge = **Artificial Intelligence**

Artificial Intelligence is a scientific domain that deals with **making intelligence** for systems with the help of **data** and **algorithms** to enable decision making, optimization, and the generation of content

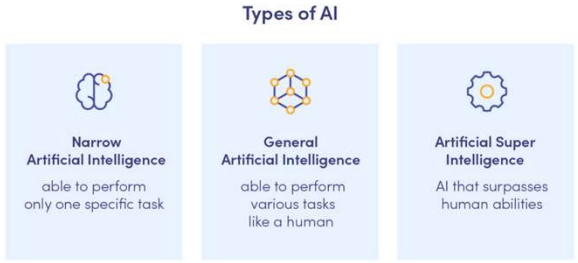
To make a system "**think like a human**"

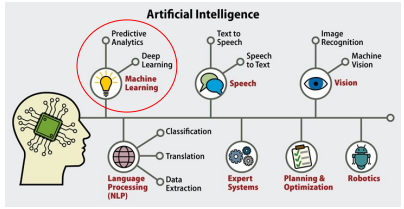


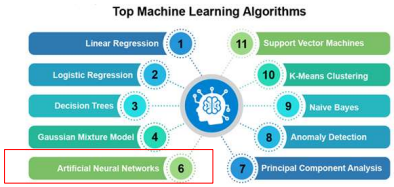
ARTIFICIAL INTELLIGENCE

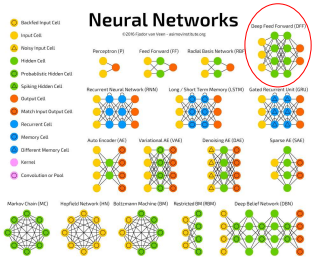


ARTIFICIAL INTELLIGENCE

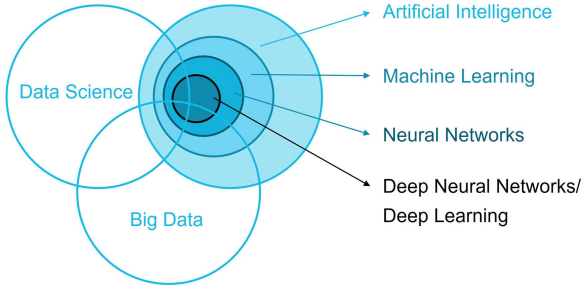




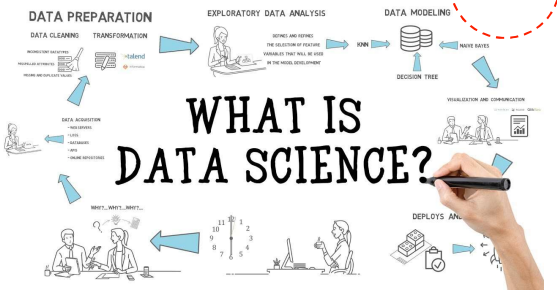




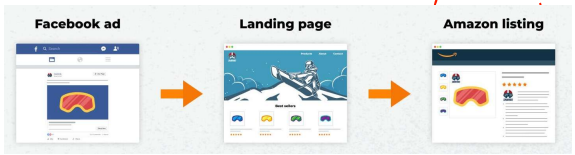
CONFUSED ?

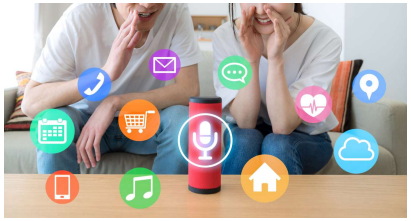


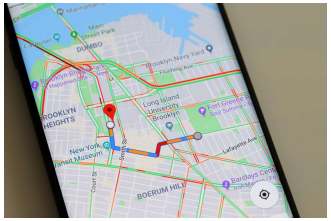
DATA SCIENCE



IS DATA IMPORTANT?







WHY DATA IS IMPORTANT ?

- Make Informed Decisions
- Get The Results You Want
- Find Solutions To Problems
- Back Up Your Arguments
- Be Strategic In Your Approaches
- Keep Track Of It All
- Know What You Are Doing Well

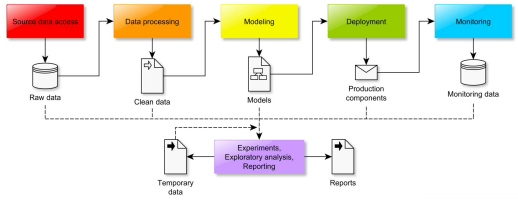
DATA IS NEW OIL

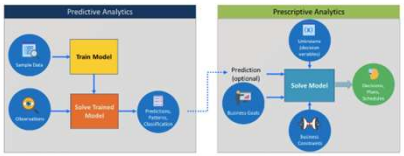
REALITY ABOUT DATA

We are now **drowing** in data ! but **starving** for insights ☹

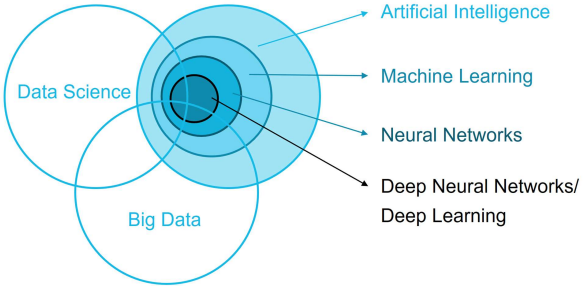
SOLUTION ?

Data Science Pipeline

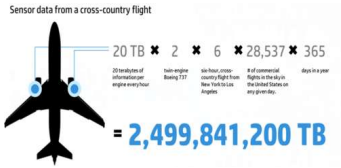




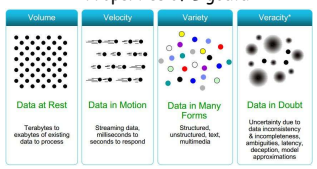
CONFUSED ?



Big Data

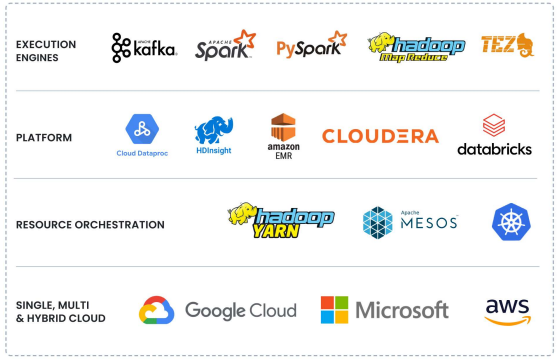


Properties of Bigdata



Is Big Data = Large Data ?

Big Data



WHAT IS MACHINE LEARNING?

"Learning is any process by which a system improves performance from experience."
- Herbert Simon

Definition by Tom Mitchell (1998):

Machine Learning is the study of algorithms that
- improve their performance **P**
- at some task **T**
- with experience **E**.

A well-defined learning task is given by **<P,T,E>**

This is a shirt we used to wear.



Is this a shirt ?



Ok


This is also a shirt

Is this a shirt ?



Ok


This is also a shirt

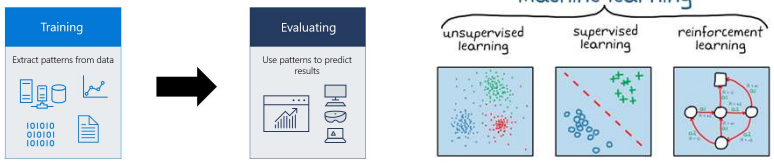
Yes, these are all shirts

PROGRAMMING VS LEARNING

Traditional Programming

Machine Learning

MACHINE LEARNING



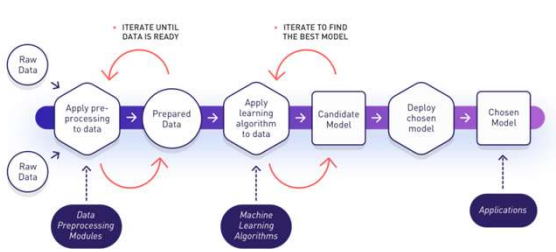
Problems with Uncertainity

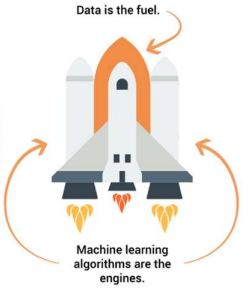
Approximating the patterns to Generalization

Data with hidden patterns

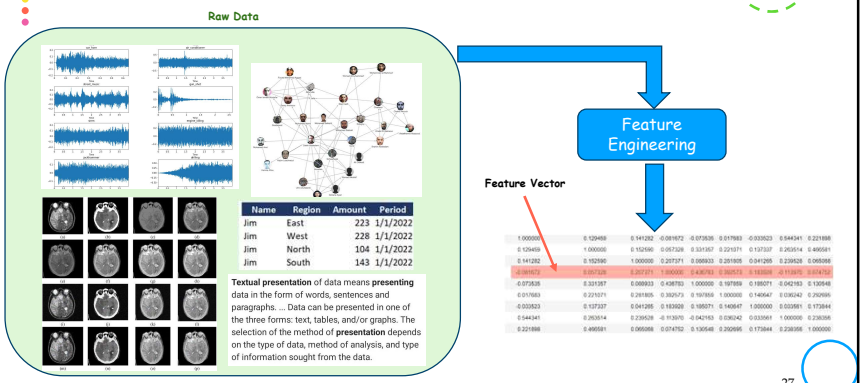
Minimum error to predict unseen data

MACHINE LEARNING PIPELINE





DATA AS FEATURE VECTORS



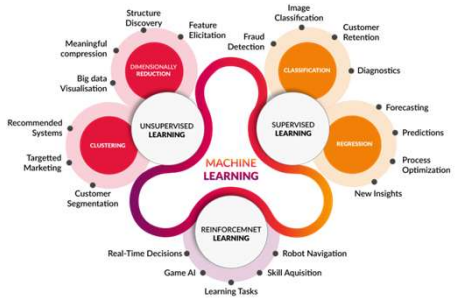
WHEN TO MACHINE LEARNING?

ML is used when:

- Human expertise does not exist (navigating on Mars)
- Humans can't explain their expertise (speech recognition)
- Models must be customized (personalized medicine)
- Models are based on huge amounts of data (genomics)

Learning isn't always useful - There is no need to "learn" to calculate payroll

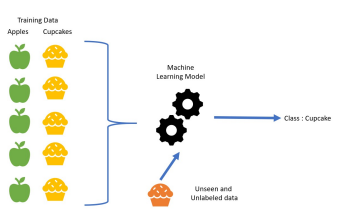
TYPES OF LEARNING



SUPERVISED LEARNING

Input : Labeled Data

X $(features)$	Y $(labels)$
$x_{11}, x_{12}, x_{13}, \ldots\ldots\ldots\ldots x_{1n}$	y_1
.	.
.	.
.	.
$x_{k1}, x_{k2}, x_{k3} \ldots\ldots\ldots\ldots x_{kn}$	y_k



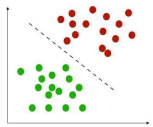
Goal : Construct a predictor $f : X \rightarrow Y$

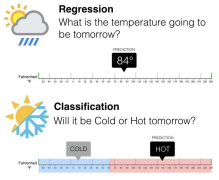
to minimize the error between $\hat{Y}, Y$
where, $\hat{Y} = f(X)$

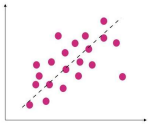
Use : using predictor to predict $\hat{y} = f(\hat{x})$

For the unknown input $\hat{x}$

SUPERVISED LEARNING







Classification

- features and discrete labels
- maps an input to discrete label(class)
- Eg: spam or not, type of cancer

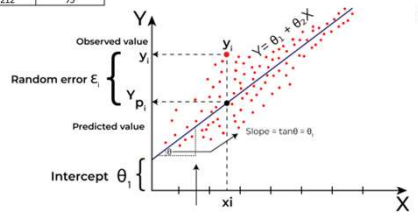
Regression

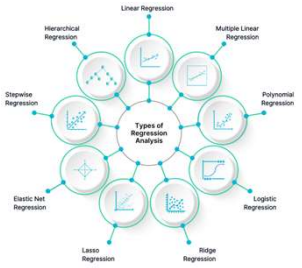
- features and continues real values
- Predict a real value for an input
- Eg: gold price, temperature

REGRESSION ALGORITHMS

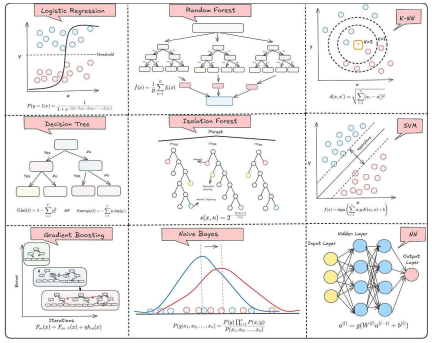
Weight (lbs)	Height (inches)
140	60
155	62
159	67
179	70
192	71
200	72
212	75

Linear Regression

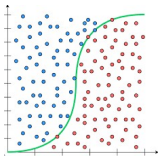


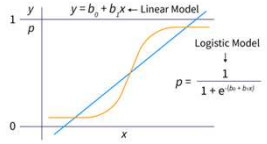


CLASSIFICATION ALGORITHMS



Logistic Regression

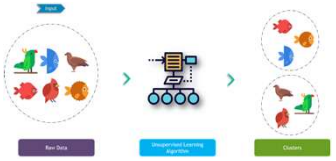




UNSUPERVISED LEARNING

Input : Unlabeled Data

X (features)
$x_{11}, x_{12}, x_{13}, \dots, x_{1n}$
.
.
.
$x_{k1}, x_{k2}, x_{k3}, \dots, x_{kn}$



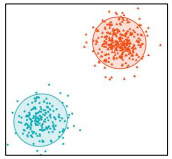
Goal : Construct a analyzer to find the relationship between inputs

hidden

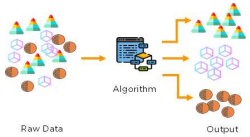
$x_1, x_2, x_3, \dots, x_k$

Use : Group or associate inputs according to their similarity

UNSUPERVISED LEARNING



Clustering

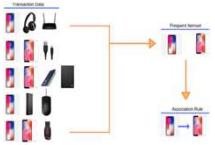


ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
...	...

market basket transactions

{Diapers, Beer} Example of a frequent itemset

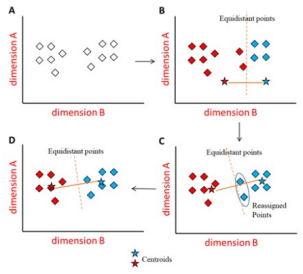
{Diapers} → {Beer} Example of an association rule



Itemset and rule mining

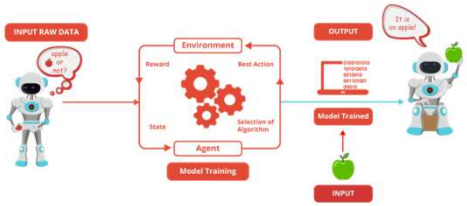
CLUSTERING ALGORITHM

K-Means Clustering



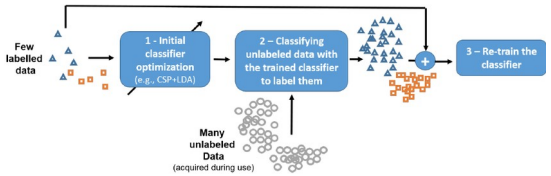
Type of Clustering Algorithm	Visual Overview	Description	Algorithm(s)
Centroid-based		Cluster points based on proximity to centroid	KMeans KMeans++ KMedoids
Connectivity-based		Cluster points based on proximity between clusters	Hierarchical Clustering (Agglomerative and Divisive)
Density-based		Cluster points based on their density instead of proximity	DBSCAN OPTICS HDBSCAN
Graph-based		Cluster points based on graph distance	Affinity Propagation Spectral Clustering
Distribution-based		Cluster points based on their likelihood of belonging to the same distribution.	Gaussian Mixture Models
Compression-based		Transform data to a lower dimensional space and then perform clustering	BIRCH

REINFORCEMENT LEARNING



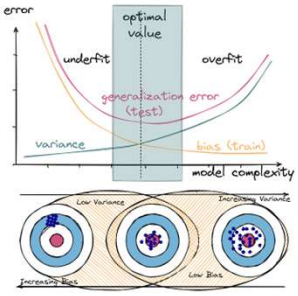
- Use software agents
- Based on rewards
- Objective is to maximize rewards for better learning

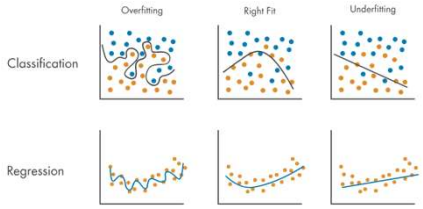
SEMI SUPERVISED LEARNING



In a nutshell, semi-supervised learning (SSL) is a machine learning technique that uses a small portion of labeled data and lots of unlabeled data to train a predictive model

BIAS VARIANCE AND ISSUES IN ML





Dr. Shailesh Sivan
+91 8907230664
shaileshsivan@cusat.ac.in
https://shaileshsivan.info


QUESTIONS