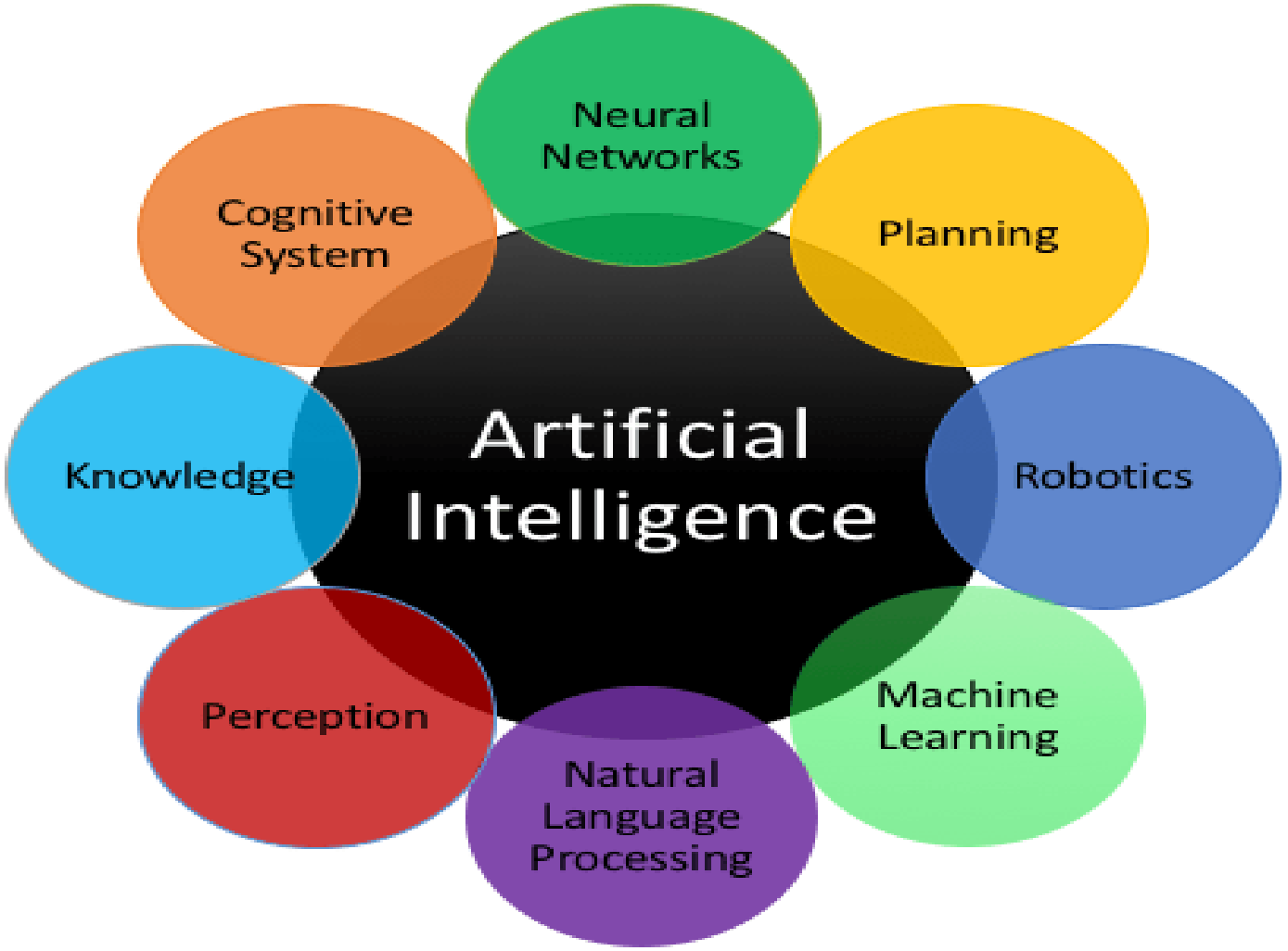
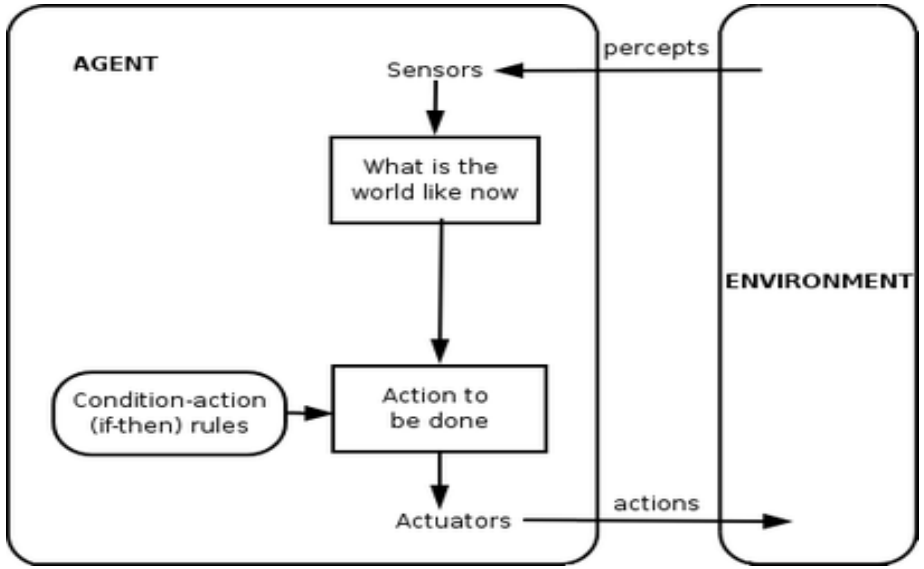
Artificial Intelligence for Effective Governance

What is Artificial Intelligence



AI (Artificial Intelligence) is the simulation of human intelligence processes by machines, especially computer systems.

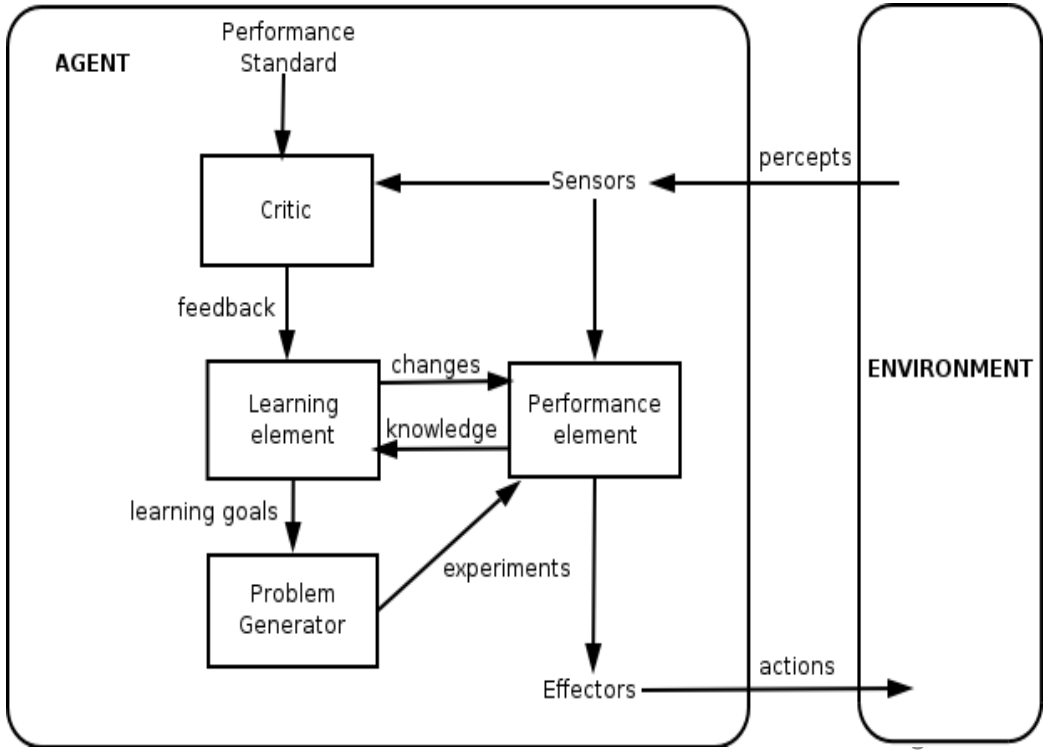
Intelligent Agents



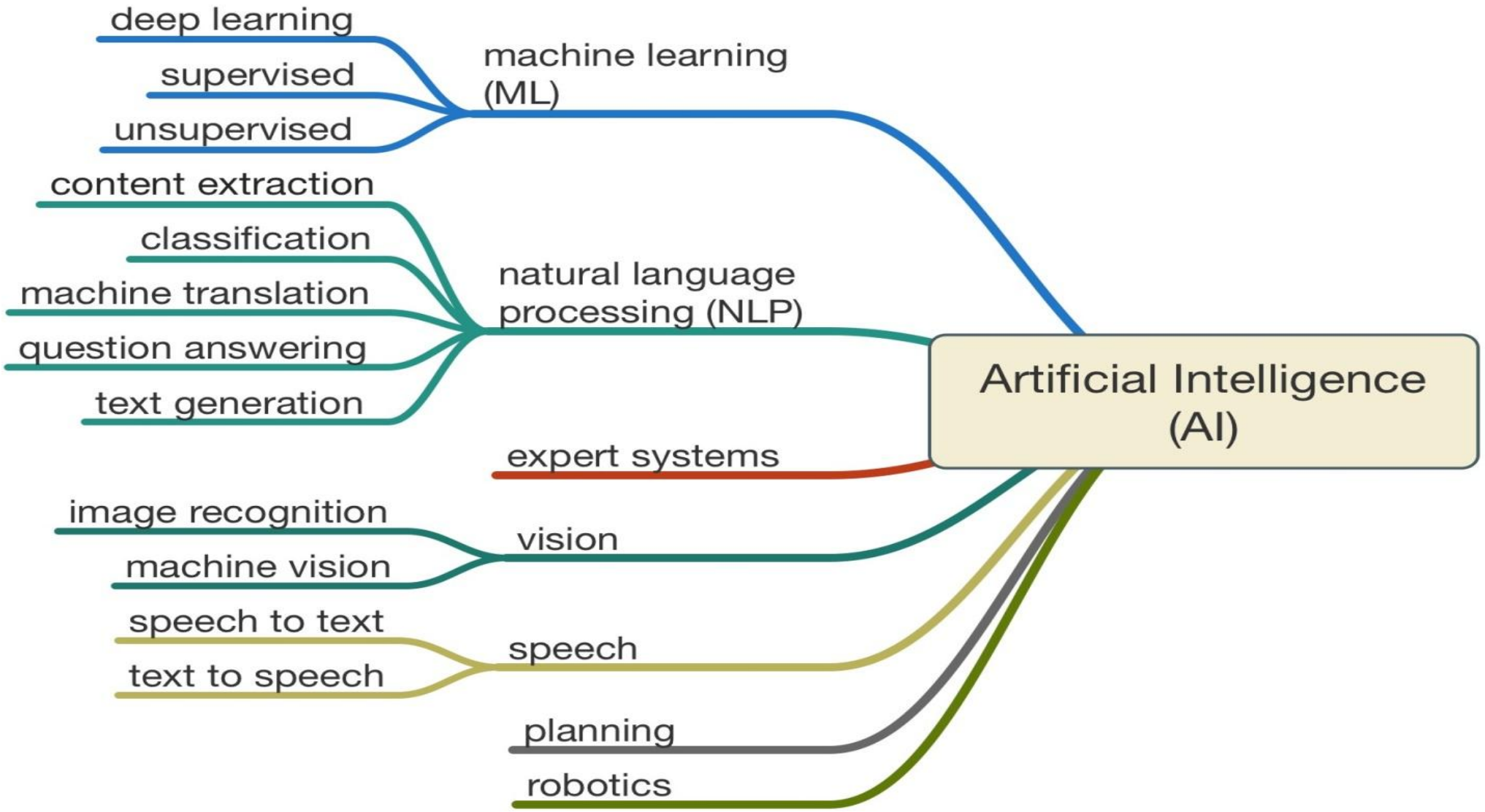
In AI, an **intelligent agent** (**IA**) refers to an autonomous entity which directing its activity towards achieving goals (i.e. it is an agent), acts upon an environment using observation through sensors and consequent actuators. A Simple reflex machine or agent, such as a thermostat, is considered an example of an intelligent agent

Intelligent agents may also learn or use knowledge to achieve their goals. Learning Agents have the advantage that it allows the agent to initially operate in unknown environments and to become more competent than its initial knowledge alone might allow.

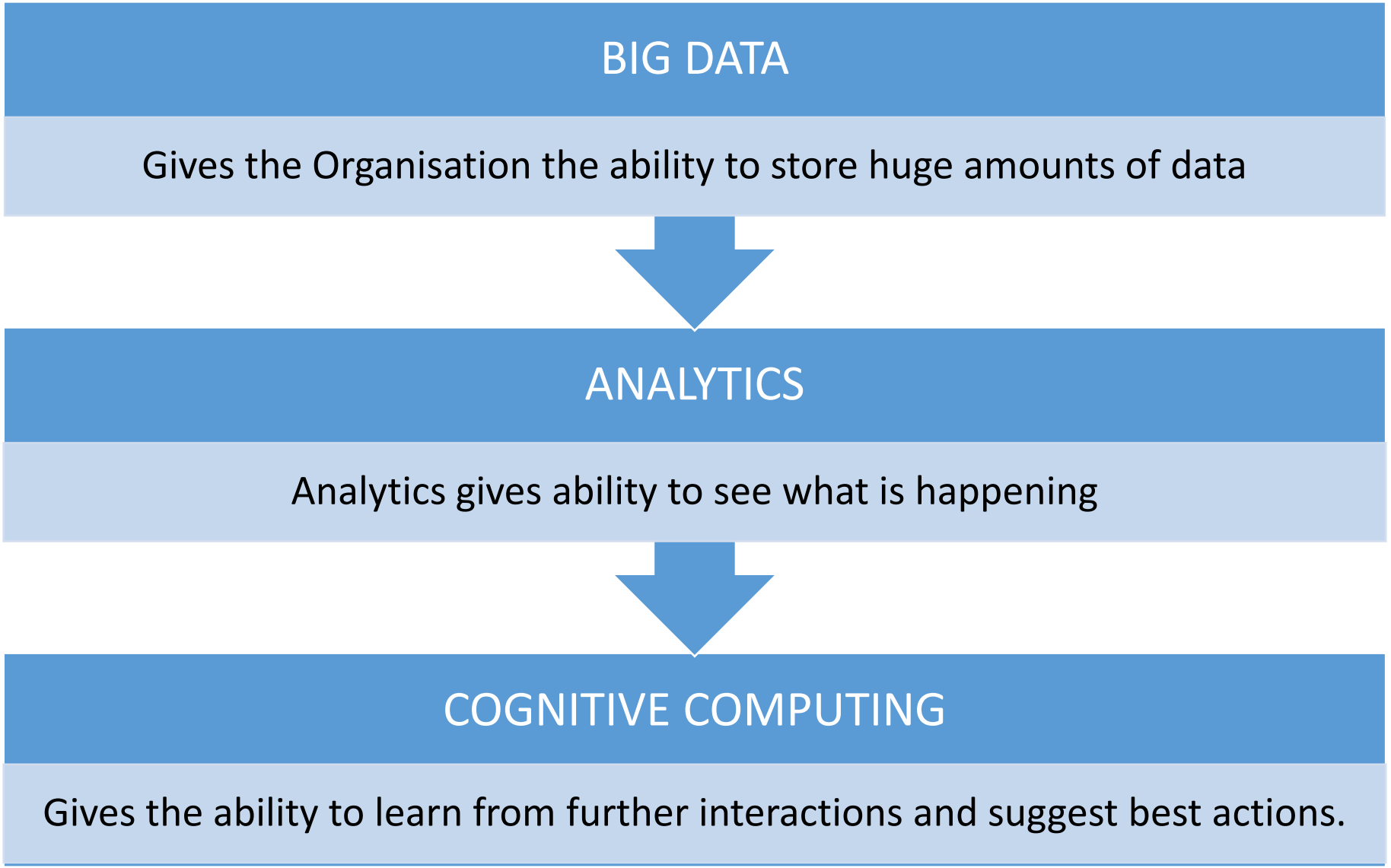
The learning element uses feedback from the "critic" on how the agent is doing and determines how the performance element should be modified to do better in the future. The performance element is what we have previously considered to be the entire agent: it takes in percepts and decides on actions. The "problem generator“ component is responsible for suggesting actions that will lead to new and informative experiences.



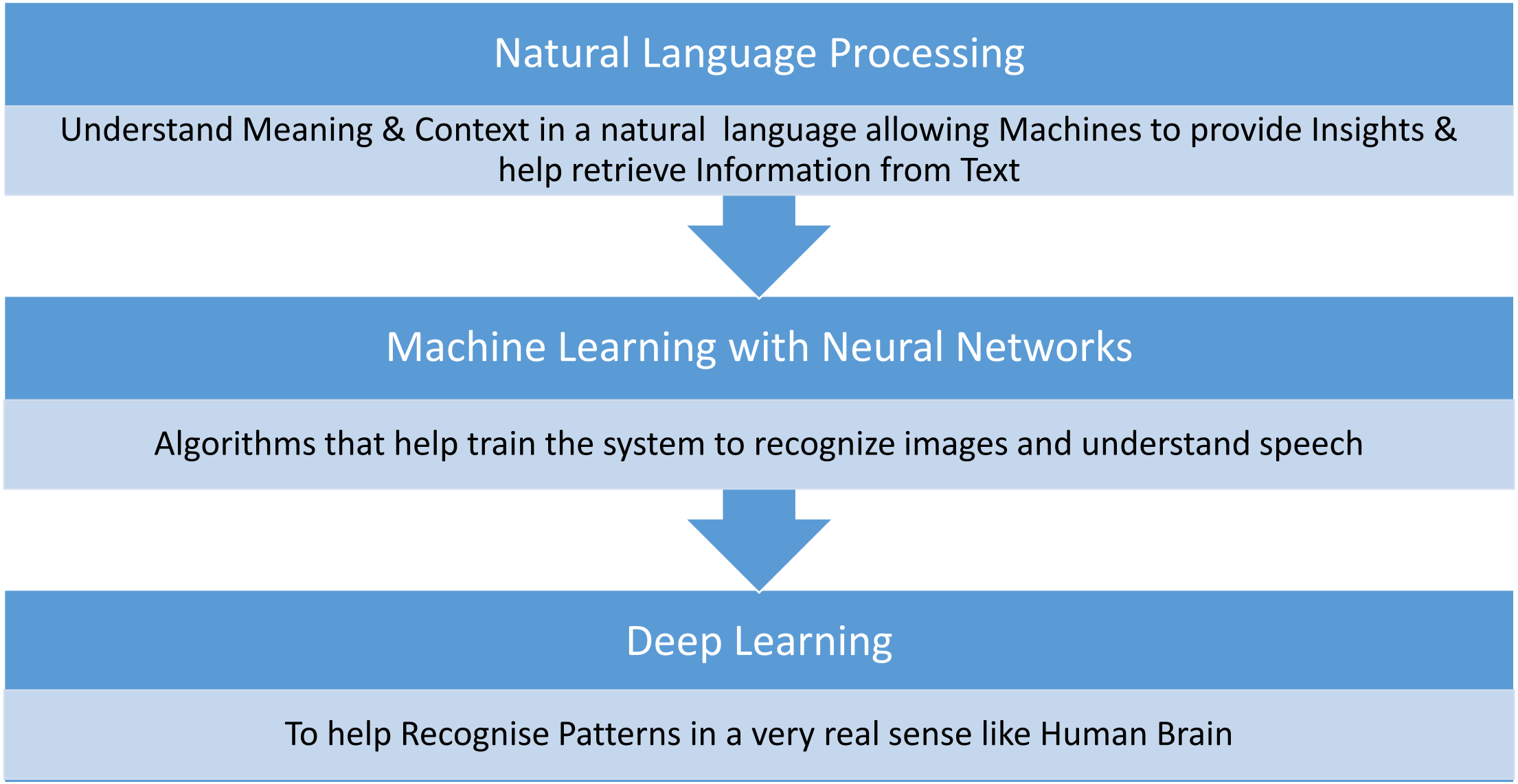
Core AI Technologies



How AI Works


BIG DATA
Gives the Organisation the ability to store huge amounts of data
ANALYTICS
Analytics gives ability to see what is happening
COGNITIVE COMPUTING
Gives the ability to learn from further interactions and suggest best actions.

How AI Works Contd..



AI & Data Quality

Machine learning aims to let algorithms learn and predict answers to problems by analysing data to make predictions on their own

The quality and depth of data will determine the level of AI applications you can achieve.

It is imperative for data quality to be present for AI to be not only accurate, but impactful.

Data Readiness for AI

- Business Problems should have enough supporting data for ML Modeling Techniques to work.
- Data should be good for prediction (Standardization, Metadata Creation)
- Ensure a good understanding of input data (Annotations & Ontologies)
- Have a clear idea of trust worthiness of the data
- Plan for validation and testing as part of the data collection
- Monitor Results

Define Data Quality Requirements
Is Data Captured Online at Source
Is the Data Capture Distributed across cross cultural Domains
Is Data Granularity maintained across Distributed Data Capture
Is there possibility of same data capture through multiple entry points like CSCs, Direct Entry by Citizens, Surveys by LSG…
Is the Data across the organization captured through Predefined Templates
Has any Data Codification Standardisation practices been adopted
Does any time barred Bulk entry get posted into database without data validations
Is data captured though different systems in different locations merged
Has Deduplication process been run & are the records uniquely identifiable
Is Data Range check done on each category of data entered in the system
Is Text data entered recoded to eliminate spelling differences
Is Data Source Captured along with timestamp of entry/updation
Is Original record archived on Updation
Is Data captured at source stored on different servers in different file formats
Is same Data stored on different servers in different file formats consistently changed on Updation
Is Data backup done at scheduled intervals
Is Data restore checked for consistency at scheduled intervals

Profile, Analyse & Assess Data Quality
Are Free Text Entries like issues or remarks Grouped & analysed for frequency, for arriving at a reduced Category set for that column
Is Mean & Variance & Range calculated & distribution observed for Numerical data
Are Outliers and their frequency checked and documented
Is it documented how Mean, Variance & Range distribution change after removing outliers
Has it been brought to the notice of user department & reasons assessed
Has it been documented from user experience whether to use that variable with or without outliers
Are Zero or null values documented. Can they be replaced by mean or modal value or a randomized value within the data range
Is the Data a sparse matrix
Is it possible that the data is sparse or dense depending on target groups
Is there reasonable assurance that the data collection methods being used donot produce systematically biased data
Are Data Collection and analysis method documented in writing and being used to ensure the same procedures are followed each time?
Are mechanisms in place to prevent unauthorized changes to the data?
Is data required from any other sector to complete the use case for AI application in domain under consideration?

Define Data Quality Metrics & Thresholds
Are the entities modeled within the enterprise captured and represented uniquely within the relevant application architecture
Does the data correctly represents the “real-life” objects they are intended to model
Is data consistency maintained across datasets such that two data values drawn from separate data sets do not conflict with each other
Is data completeness ensured, it may be seen as encompassing usability and appropriateness of data values
Is Timeliness ensured, it can be measured as the time between when information is expected and when it is readily available for use
Is Data currency ensured, it may be measured as a function of the expected frequency rate at which different data elements are expected to be refreshed
Is data conforming ie. whether instances of data are stored, exchanged, or presented in a format that is consistent with the domain of values, as well as consistent with other similar attribute values
Is threshold for data conformance defined for ‘Acceptable’, ‘Questionable but usable’ and ‘Unusable’

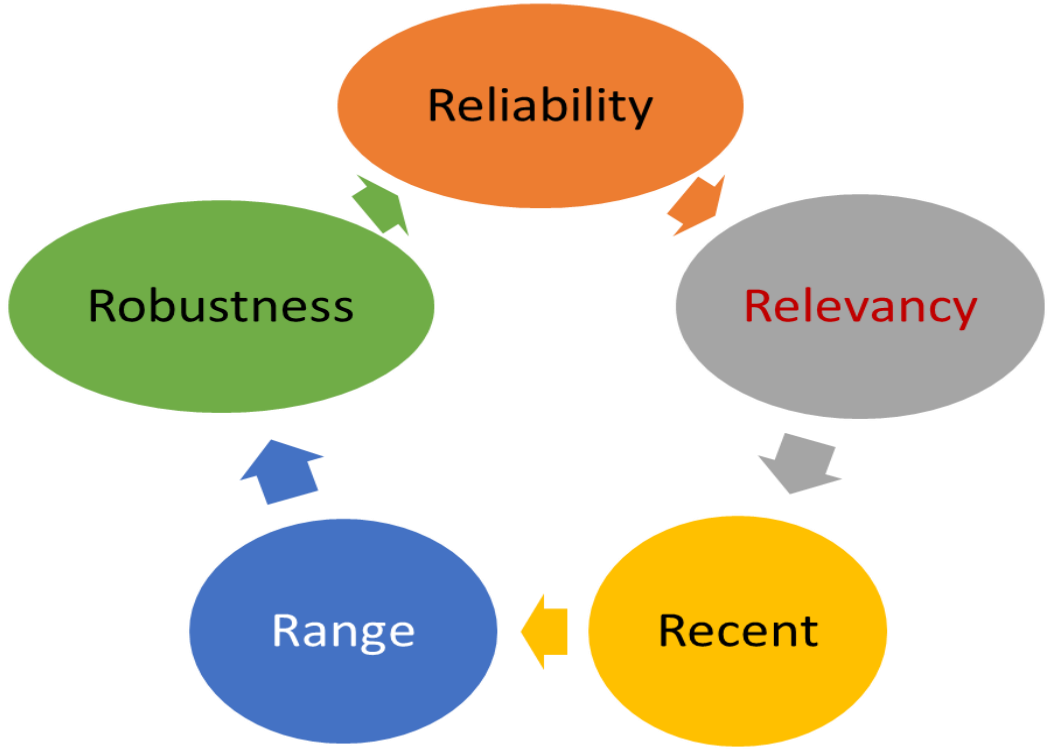
Monitor Data Quality Continuously

Are data quality thresholds set and quality alerts generated to automatically escalate data quality problems to the process owners & stakeholders
Are Data Trends tracked and changes visualised from a historical perspective
What is the Primary Data Quality Problem - Missing Data / Duplicate Data/ Addressing Standardization/ Data Enhancement/ All of Above
Is Data Quality Review Matrix prepared and in use

Nicholas Piette said the following words in this regard:

"100% of AI projects are subject to fail if there are no solid efforts beforehand to improve the quality of the data being used to fuel the applications.

Making no effort to ensure the data you are using, is absolutely accurate and trusted—in my opinion—is indicative of unclear objectives regarding what AI is expected to answer or do. I understand it can be difficult to acknowledge, but if data quality mandates aren't addressed up front, by the time the mistake is realized, a lot of damage has already been done. So make sure it's forefront."



5 R's for Data Quality – Nicholas Piette

Standard Machine Learning Algorithms

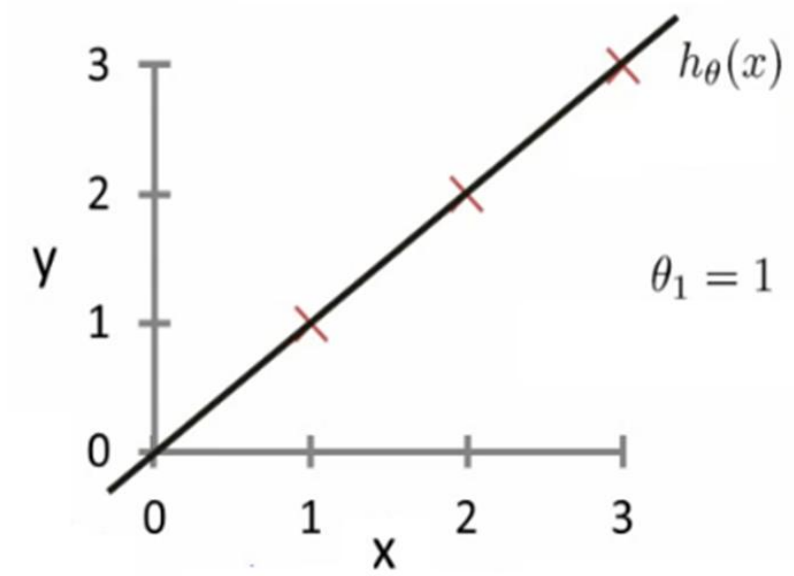
- *Machine learning algorithms are described as learning a target function (f) that best maps input variables (X) to an output variable (Y):* ***Y = f(X)***
- This is a general learning task where we would like to make predictions in the future (Y) given new examples of input variables (X)

Top Machine Learning Algorithms are

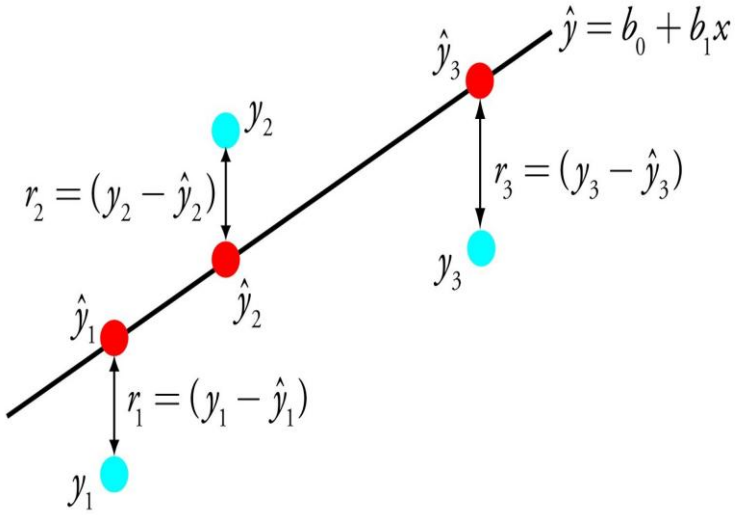
- ✓Linear Regression
- ✓Logistic Regression & Linear Discriminant Analysis
- ✓Decision & Regression Trees
- ✓Naive Bayes
- ✓k-Nearest Neighbor & Learning Vector Quantization
- ✓Support vector machine
- ✓Bagging & Random Forest
- ✓Boosting & AdaBoost
- ✓Markov

Supervised Learning — Linear Regression

Predictive modeling is primarily concerned with minimizing the error of a model or making the most accurate predictions possible. Linear Regression is for predicting values of Continuous Dependent Variables.



The representation of linear regression is an equation that describes a line that best fits the relationship between the input variables (x) and the output variables (y), by finding specific weightings for the input variables called coefficients (θ).



For example using the expression For example: $Y = \theta_0 + \theta_1 X$ which is called the hypothesis we will predict y given the input x and the goal of the linear regression learning algorithm is to find the values for the coefficients θ_0 and θ_1 such that we minimize the Residual errors of the dataset points and line of best fit or that we minimize the difference between predicted y ^ & the ground truth y.

Linear Regression Continued....

A pair (x(i) , y(i)) is called a training example, and the dataset that we'll be using to learn — a list of m training examples {(x(i), y(i)); i = 1, . . . , m }—is called a training set. We will also use X to denote the space of input values, and Y the space of output values. In this example, X = Y = R.

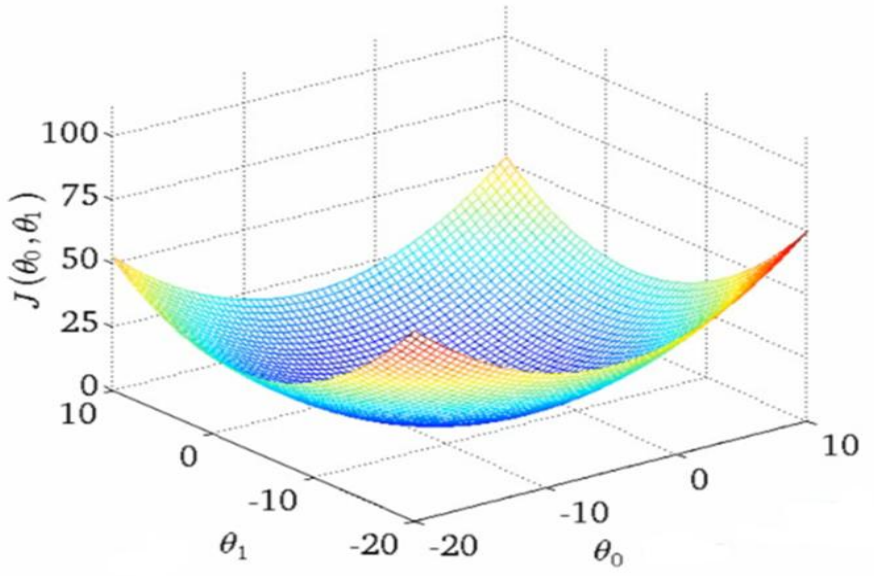
To describe the supervised learning problem slightly more formally, our goal is, given a training set, to learn a function h : X → Y so that h(x) is a "good" predictor for the corresponding value of y. For historical reasons, **this function h is called a hypothesis**.

Let's consider a Housing Saleprices dataset as given: Here, x's are a four-dimensional vector in R4 . For instance, $x(i)_1$ is the living area in square feet of the i-th house in the training set, and $x(i)_2$ is its number of bedrooms and so on…

e.g. $h_\theta(x) = 80 + 0.1x_1 + 0.01x_2 + 3x_3 - 2x_4$
Here, the θ_i's are the parameters (also called weights) parameterizing the space of linear functions mapping from X to Y. We define and try to minimize the cost function J(θ) as

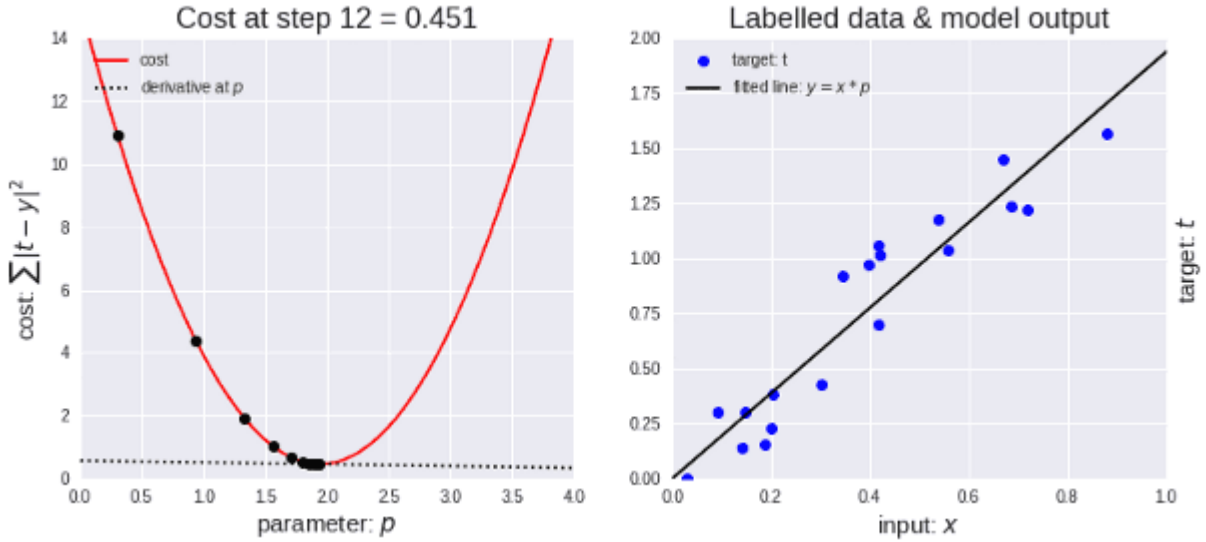
$$J(\theta_0, \theta_1, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Size (feet²)	Number of bedrooms	Number of floors	Age of home (years)	Price ($1000)
x_1	x_2	x_3	x_4	y
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178



Gradient Descent

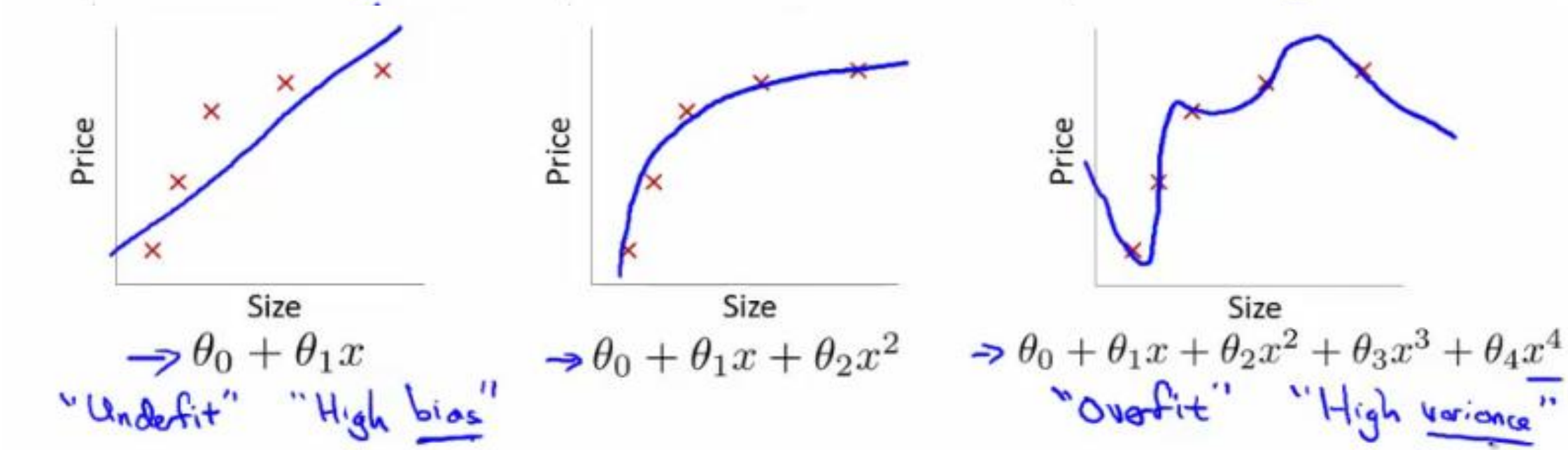
It is an ***iterative*** optimization algorithm used in machine learning to find the best results (minima of a curve). *Gradient* means the *rate* of inclination or declination of a slope. *Descent* means the instance of *descending*. The algorithm is **iterative** means that we need to get the results multiple times to get the most optimal result. The iterative quality of the gradient descent helps a under-fitted graph to make the graph fit optimally to the data.



The Gradient descent has a parameter called **learning rate.** As you can see above (left), initially the steps are bigger that means the learning rate is higher and as the point goes down the learning rate becomes more smaller by the shorter size of steps. Also, the **Cost** Function is decreasing or the cost is decreasing . When the data is too big ,we can't pass all the data to the computer at once. So, to overcome this problem we divide the data into smaller sizes and give it to our computer one by one and update the weights of the neural networks at the end of every step to fit it to the data given. Now **epoch** can be defined as one forward pass and one backward pass of all training data while **iteration** is one forward pass and one backward pass of each batch size. If all training data is divided into four batch sizes then **epoch** =1 and **iteration** =4. ... Typically an **epoch** is achieved after several **iterations**.

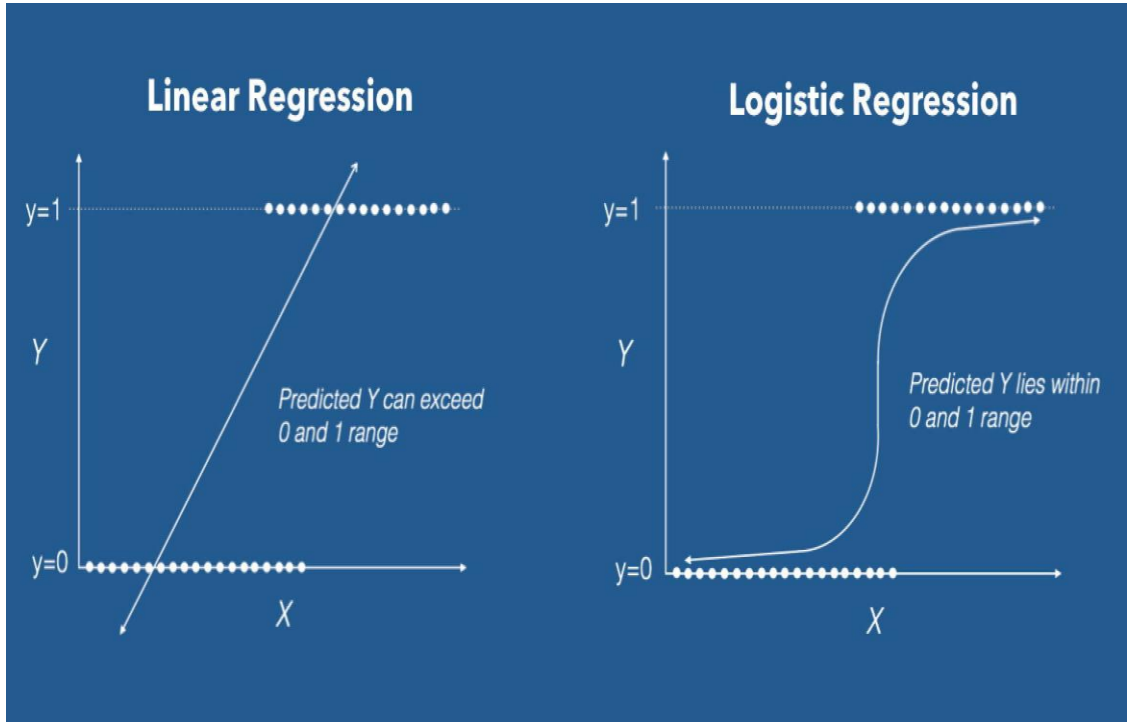
Linear Regression Contd..

The cause of poor performance in machine learning is either **overfitting or underfitting** the data. In the first diagram the Linear Regression function in Blue line is underfit to the samples given in red crosses — in which the data clearly shows structure not captured by the model. In the Middle Diagram a polynomial of degree 2 models the sample truly, say we had added few extra features and fit $y = \sum_{j=0}^{2} \theta_j x_j$, then we obtain a slightly better fit to the data, and in the last figure the Blue Line going through all the points in training set is an overfit.



Supervised Learning — Logistic Regression

Logistic Regression is used when the dependent or target variable is categorical. Usually used where there are two possible outcome values often labeled as "0" and "1", which represent outcomes such as pass/fail, win/lose , whether the email is spam or ham, predicting if a given mass of tissue is benign or malignant etc.. Linear regression does not have this capability of predicting a value between 0 and 1.



Unlike linear regression, the prediction for the output is transformed using a non-linear function called the logistic function hΘ(x) = sigmoid (Z) If ‘Z’ goes to infinity, Y(predicted) will become 1 and if ‘Z’ goes to negative infinity, Y(predicted) will become 0.

$h_\Theta(x) = P(Y=1|X; theta)$

Probability that Y=1 given X which is parameterized by ‘theta’.

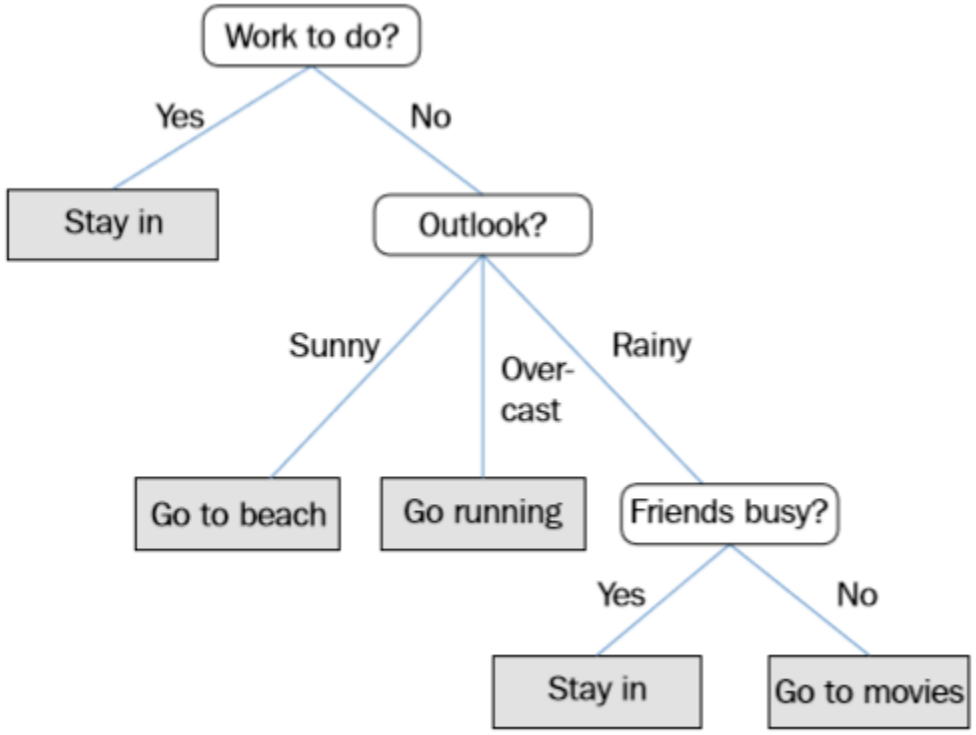
$P(Y=1|X; theta) + P(Y=0|X; theta) = 1$

$P(Y=0|X; theta) = 1 - P(Y=1|X; theta)$

Supervised Learning —Decision Trees

The representation of the decision tree model is a binary tree. Each node represents a single input variable (x) and a split point on that variable (assuming the variable is numeric). The leaf nodes of the tree contain an output variable (y) which is used to make a prediction.

Predictions are made by walking the splits of the tree until arriving at a leaf node and output the class value at that leaf node. Trees are fast to learn and are also accurate for a broad range of problems and do not require any special preparation for your data.

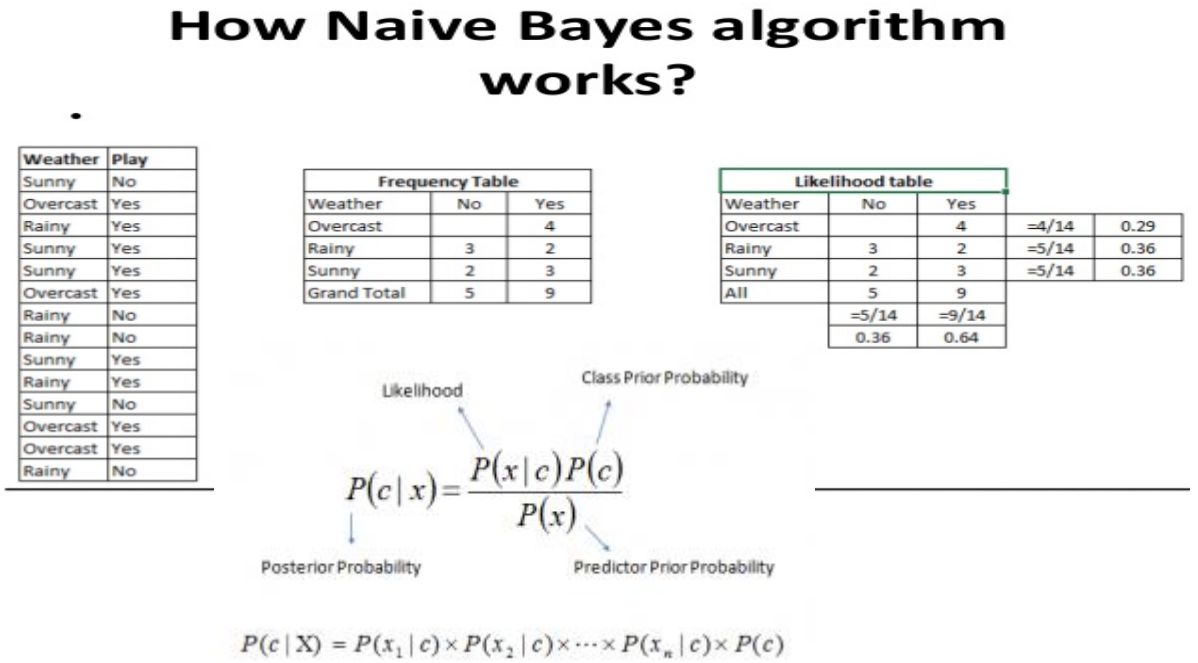


The performance of a tree can be further increased by ***pruning***. *It involves* ***removing the branches that make use of features having low importance***. This way, we reduce the complexity of tree, and thus increasing its predictive power by reducing overfitting.

Supervised Learning —Naive Bayes

Naive Bayes is a simple but surprisingly powerful algorithm for predictive modelling. The model is comprised of two types of probabilities : 1) The probability of each class; and 2) The conditional probability for each class given each x value. Once calculated, the probability model can be used to make predictions for new data using Bayes Theorem.

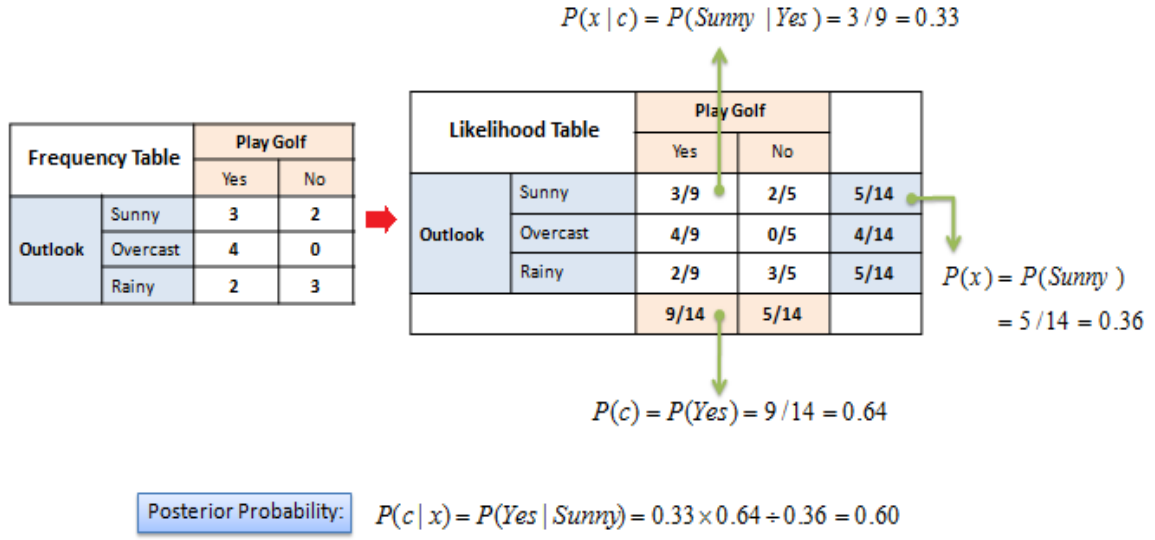
Naive Bayes is called naive because it assumes that each input variable is independent. It remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other with word frequencies as the features.



Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

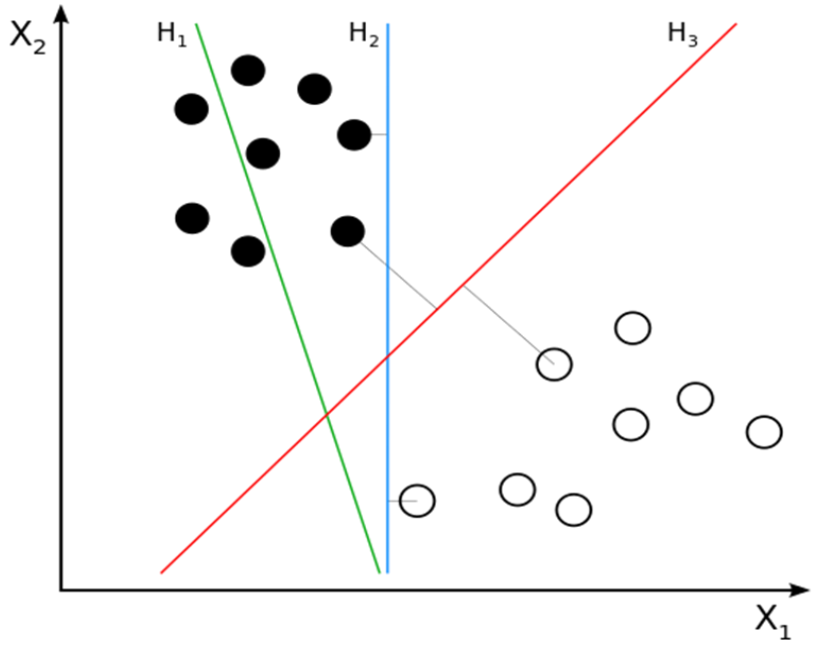
Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		



Supervised Learning - Support Vector Machines

Support Vector Machines are perhaps one of the most popular and talked about machine learning algorithms. A hyperplane is a line that splits the input variable space. In SVM, a hyperplane is selected to best separate the points in the input variable space by their class, either class 0 or class 1.

In two-dimensions, you can visualize this as a line and let's assume that all of our input points can be completely separated by this line. **The SVM learning algorithm finds the coefficients that results in the best separation of the classes by the hyperplane. The distance between the hyperplane and the closest data points is referred to as the margin.**

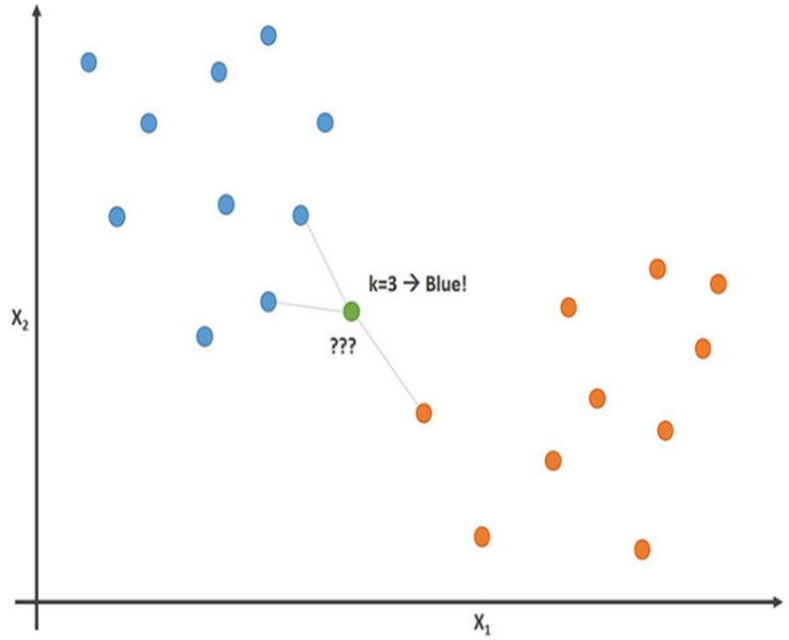


The best or optimal hyperplane that can separate the two classes is the line that has the largest margin. Only these points are relevant in defining the hyperplane and in the construction of the classifier. These points are called the support vectors. They support or define the hyperplane. In practice, an optimization algorithm is used to find the values for the coefficients that maximizes the margin.

Supervised Learning - K Nearest Neighbour

The KNN algorithm is very simple and very effective. The model representation for KNN is the entire training dataset. Predictions are made for a new data point by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those K instances.

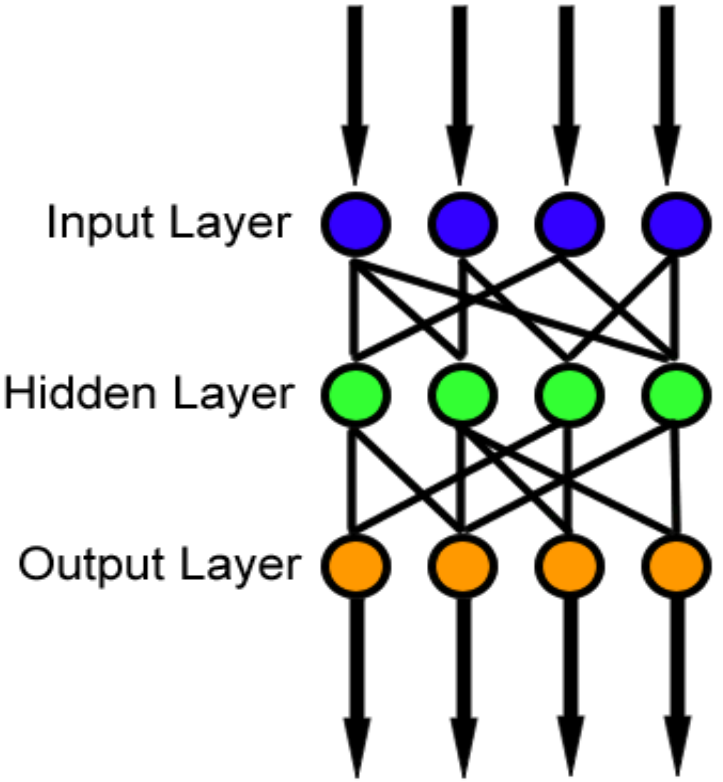
For regression problems, this might be the mean output variable, for classification problems this might be the mode (or most common) class value. The trick is in how to determine the similarity between the data instances. The simplest technique if your attributes are all of the same scale (all in inches for example) is to use the Euclidean distance, a number you can calculate directly based on the differences between each input variable.



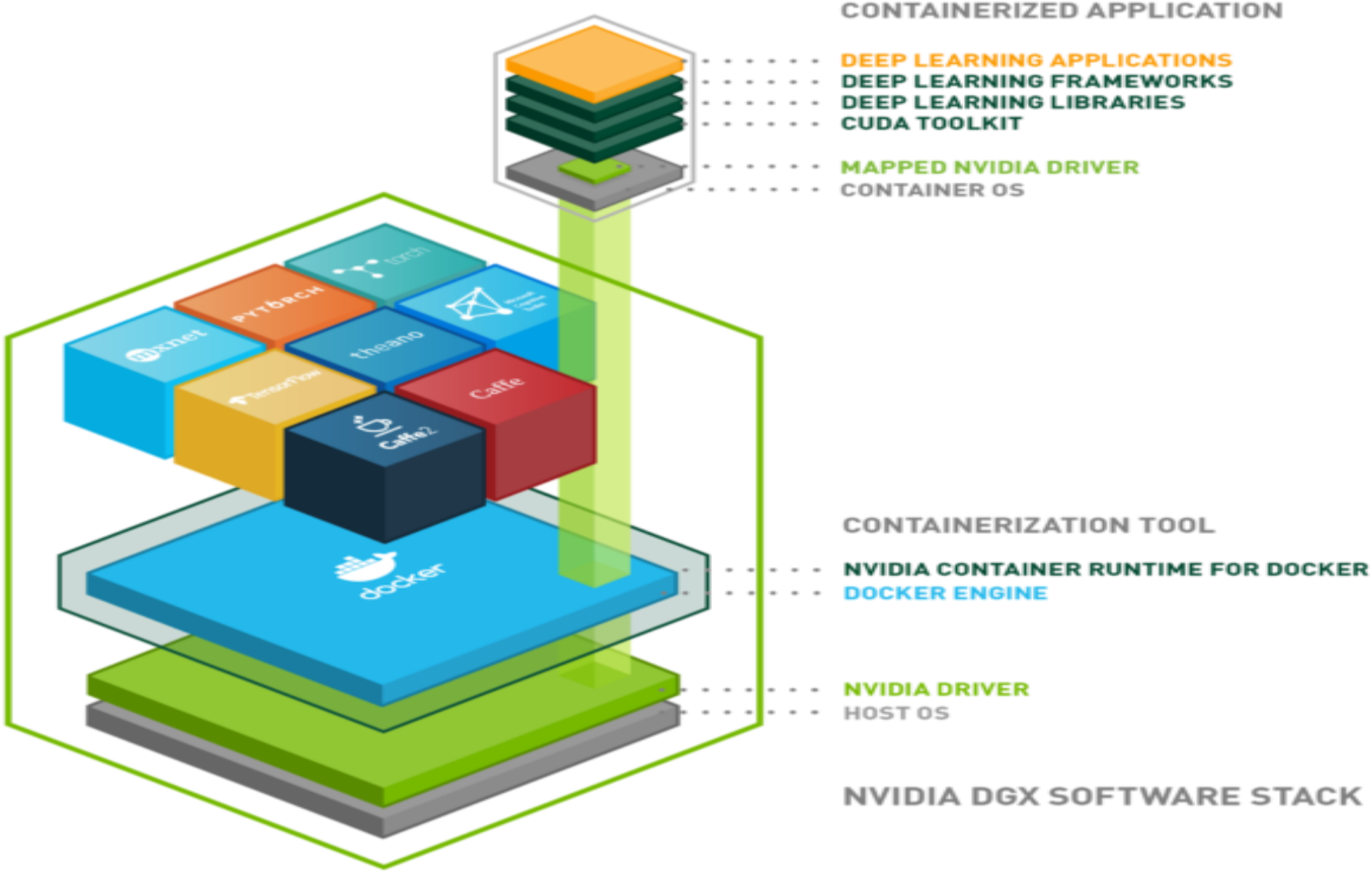
KNN can require a lot of memory or space to store all of the data, but only performs a calculation (or learn) when a prediction is needed, just in time. The idea of distance or closeness can break down in very high dimensions (lots of input variables) which can negatively affect the performance of the algorithm on your problem.

Deep Learning

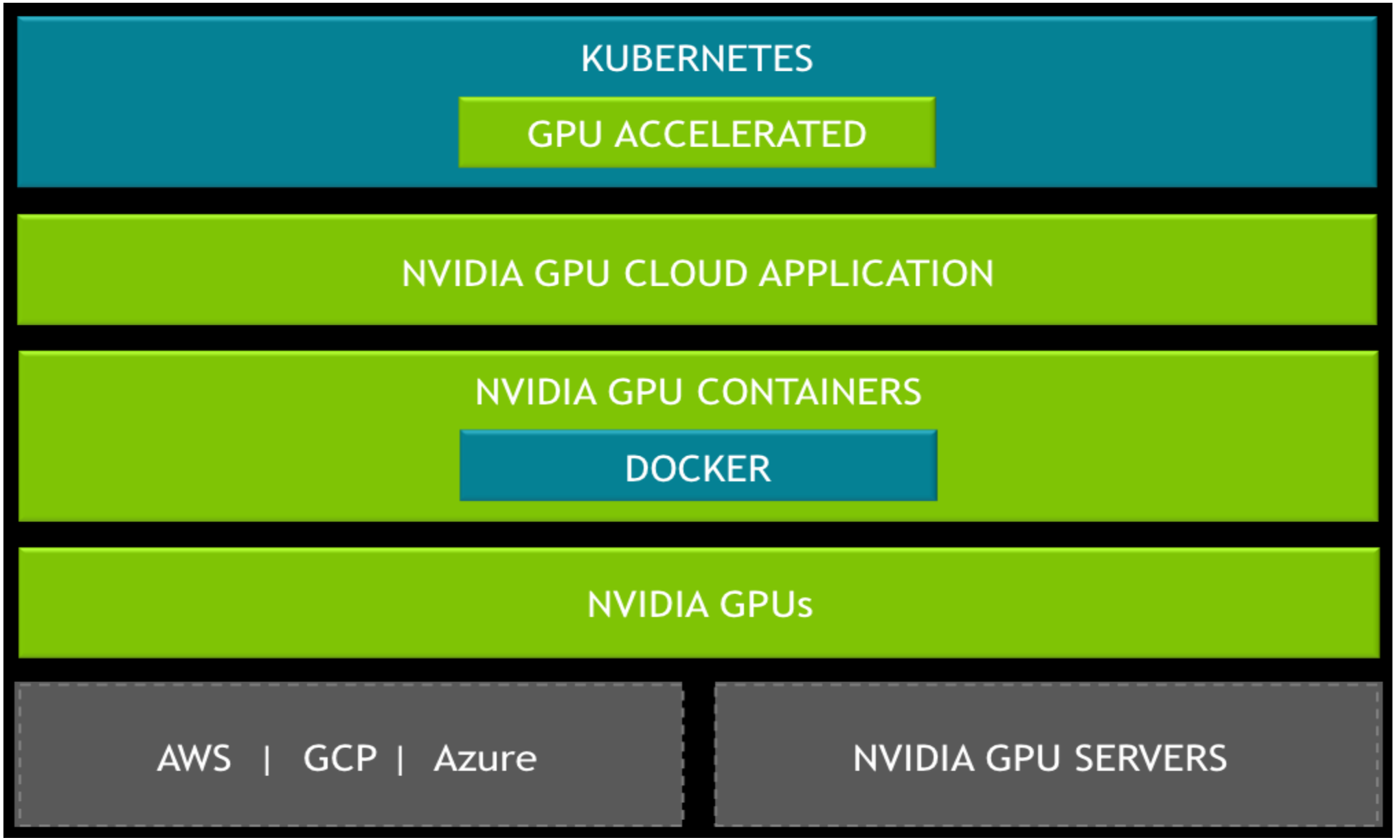
- Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.
- **Deep Learning** *achieves great power and flexibility by learning to represent the world as nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones.*
- In the example of image recognition it means identifying light/dark areas before categorizing lines and then shapes to allow face recognition.Each neuron or node in the network represents one aspect of the whole and together they provide a full representation of the image.
- Each node or hidden layer is given a weight that represents the strength of its relationship with the output and as the model develops the weights are adjusted.

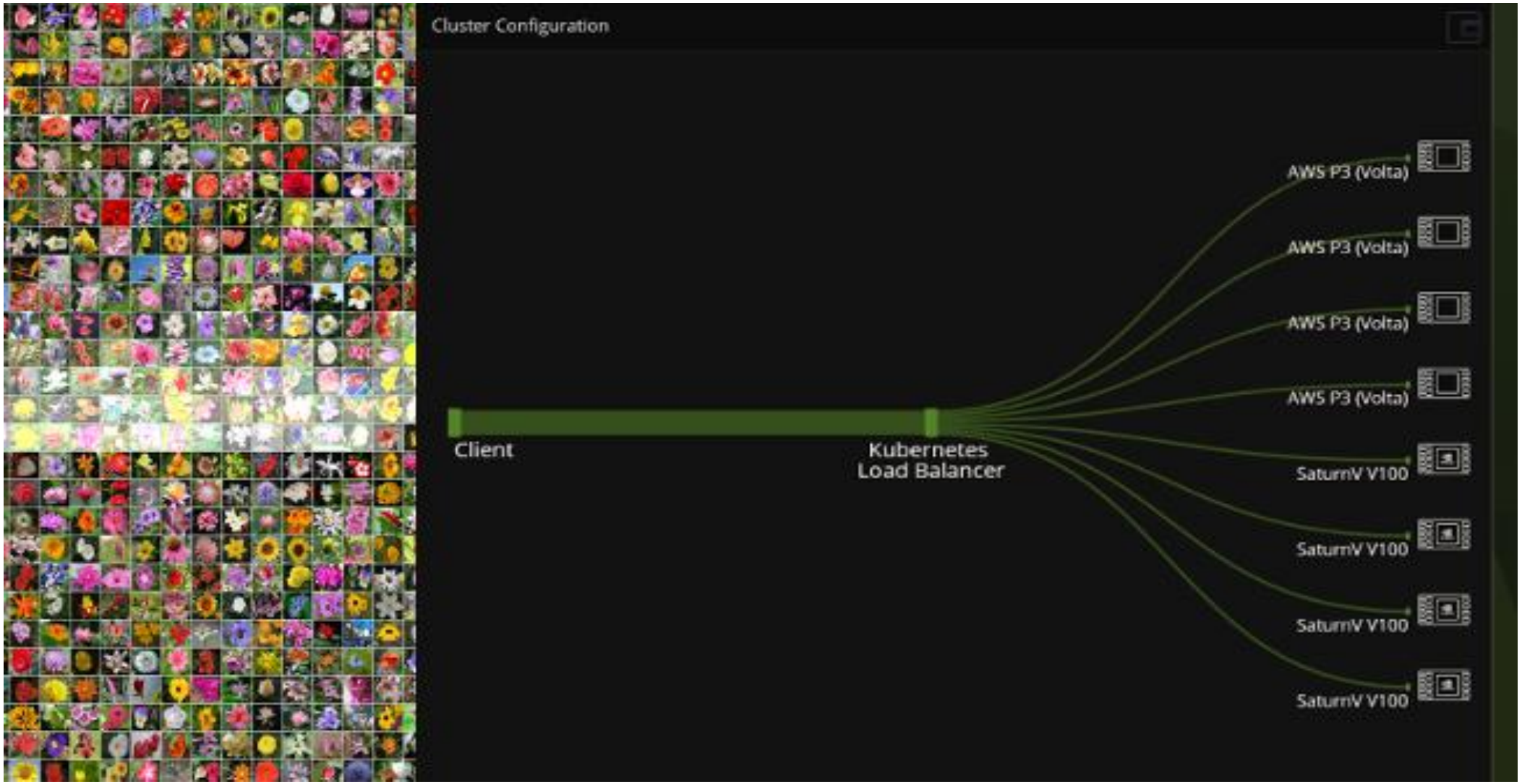


DGX 1 Architecture



AI Platform as a Service – Resource management with Kubernetes




Cluster Configuration
Client
Kubernetes
Load Balancer
AWS P3 (Volta)
AWS P3 (Volta)
AWS P3 (Volta)
AWS P3 (Volta)
SaturnV V100
SaturnV V100
SaturnV V100
SaturnV V100

AI in Citizen Services

- Citizen Engagement (Queries, Grievances, Assistance)
- Search Services based on Data/ Text Mining
- Language Translation
- Voice Based Interfaces/ Chatbots

Navigating AI in Government

- Make AI a part of Goal Based, Citizen Centric Program
- Get Citizen Input
- Build upon Existing Resources
- Be data Prepared and Tread carefully with Privacy
- Mitigate ethical risks and avoid AI decision making
- Augment employees, do not replace them

Cyber security, Safety, Legal and Ethical issues

Build Trust by making encryption and tokenization, identity and access management, security intelligence logs part of work culture to minimize risk.

With AI, it becomes easier to correlate data even with anonimisation... and remove Privacy.

Data protection and the ownership of insights from the data needs to be detailed out under Data Protection Act.

Government of India Initiatives on AI

4 Committees Constituted

- Committee on platforms and data for AI
- Committee on cyber security, safety, legal & ethical issues
- Leveraging AI for identifying national missions in key sectors
- Mapping technological capabilities, key policy enablers required across sectors, skilling , re-skilling and R&D

Niti Aayog

- National policy on AI, outlining the scope of research for the adoption and commercialisation of the technology

Commerce Ministry

- India's Industrial Revolution 4.0 - Commerce Ministry Sets Up Task Force Dedicated To AI

Centre of Excellence in AI @ NIC

- Honorable Minister of Electronics & IT, Law & Justice, Sh. Ravi Shankar Prasad inaugurated the COE-AI on 10th January 2019.
- Vision is to ensure that India is not a passive participant and user of AI based technologies developed elsewhere, but can actively influence AI development path.

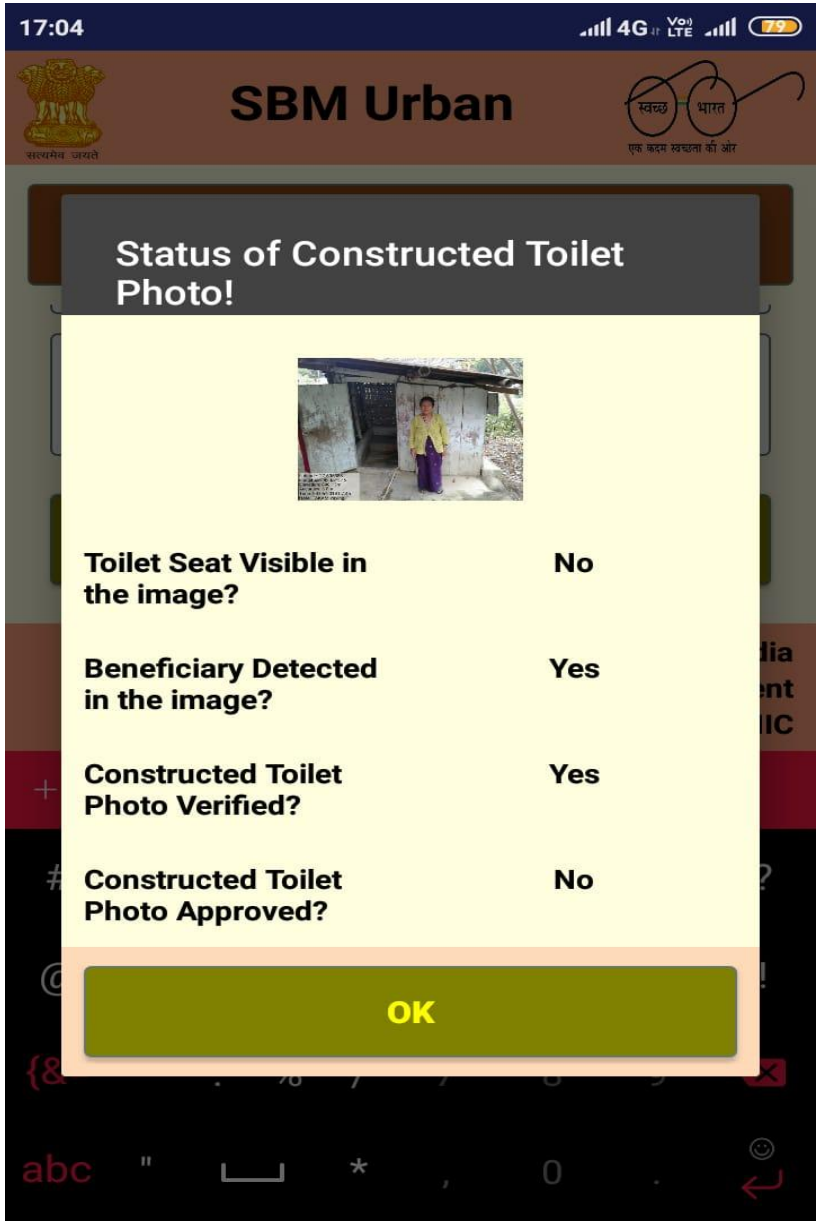
Few Proof of Concepts demonstrated by COE-AI@NIC

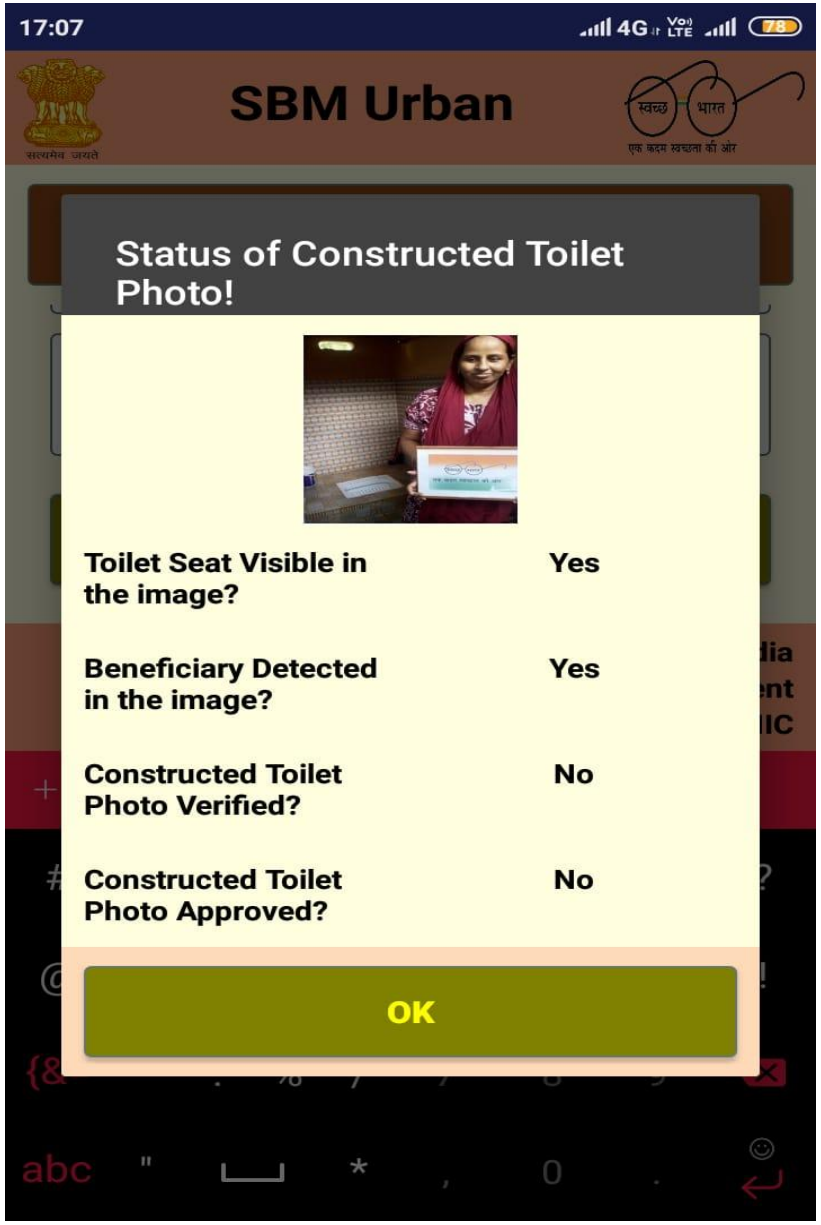
- AI has the ability to automate repetitive tasks, make connections, see relationships and make predictions with reasonable levels of accuracy.
- Object Detection in uploaded images of constructed Individual House hold toilets from Swachh Bharat Mission project in both Urban & Rural areas.
- There were eight and a half crore images uploaded in the two ministries combined, which were responsible for getting these govt. schemes implemented at grass root level. Management wanted to know efficacy.
- This POC has been matured into a model with high average precision and recall and low misclassification scores.
- Text Classification & Text Summarisation of Case Order Outcomes was demonstrated. There are 3 crore case outcome pdfs available due to eCourts computerization project.
- This POC has been asked by eCourts Committee of SCI to be matured into full blown model to help assist judges in lower judiciary. It will involve Cognitive Search, Deep Learning & Machine Translation.
- Also Demonstrated was POC in Numeric Modeling for analysis of AT&C Losses in Power Sector across feeders over a period of 22 months since National Power Portal is in existence.
- Power sector has also demonstrated interest in AI intervention in power management.

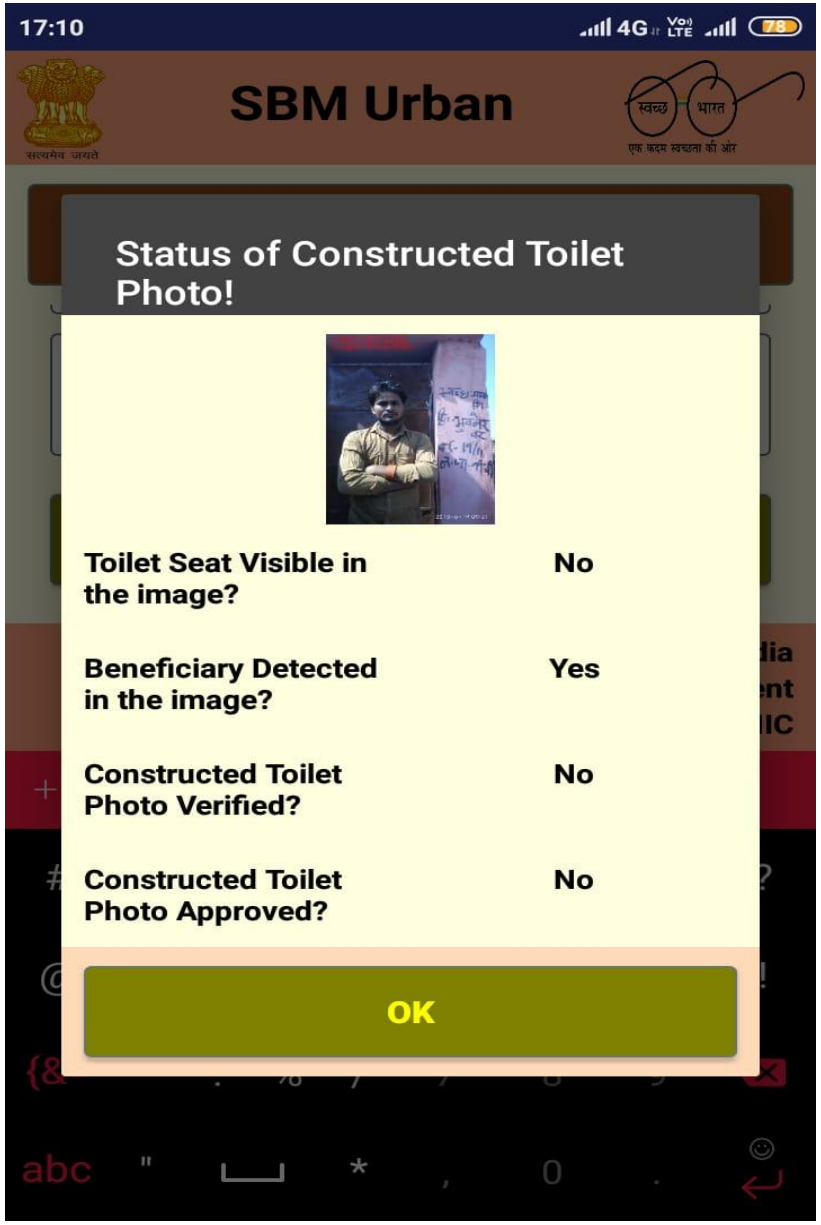
Image Analytics – Object Detection

Beneficiary & Toilet Seat Detection in IHHL Constructed Toilet Photos uploaded in Swachh Bharat Urban with accuracy 99%

Image Analytics in eGovernance – A practical example







Contact us :
011- 24305211 / 24305747
Website :
http://ai.nic.in