**Content**

[1. Machine Learning concepts / terminology 3](#_Toc20008657)

[1. Glossary of Machine Learning, Statistics and Data Science terms 3](#_Toc20008658)

[1.1.1. to infer 3](#_Toc20008659)

[1.1.2. Bias-Variance Trade-off 3](#_Toc20008660)

[1.1.3. accuracy 5](#_Toc20008661)

[1.1.4. Cross Entropy 5](#_Toc20008662)

[1.1.5. Median 5](#_Toc20008663)

[1.1.6. PCA – Principal Component Analysis 5](#_Toc20008664)

[1.1.7. Label and One Hot Encoding 5](#_Toc20008665)

[1.1.8. Validation data 6](#_Toc20008666)

[1.1.9. RecordIO Format 6](#_Toc20008667)

[2. AWS – Machine Learning Algorythms 7](#_Toc20008668)

[1. Logistical Regression 7](#_Toc20008669)

[2. Support Vector Machines 8](#_Toc20008670)

[3. Decision Trees 9](#_Toc20008671)

[4. Random Forest 9](#_Toc20008672)

[5. K-Means 10](#_Toc20008673)

[6. K-Nearest Neighbour 12](#_Toc20008674)

[3. Non-AWS technologies around machine learning 14](#_Toc20008675)

[1. Apache Spark 14](#_Toc20008676)

[2. Hadoop 14](#_Toc20008677)

[3. Keras 14](#_Toc20008678)

[4. CNTK – Microsoft Cognitive Toolkit 15](#_Toc20008679)

[4. AWS Services 16](#_Toc20008680)

[1. SageMaker 16](#_Toc20008681)

[2. Transcribe 16](#_Toc20008682)

[3. Gluon 16](#_Toc20008683)

[4. Comprehend 16](#_Toc20008684)

[5. Lex 16](#_Toc20008685)

[6. Rekognition 16](#_Toc20008686)

[7. Polly 17](#_Toc20008687)

[8. Alexa 17](#_Toc20008688)

[9. Kinesis 17](#_Toc20008689)

[10. EMR 17](#_Toc20008690)

[5. Videos 18](#_Toc20008691)

[1. 2017:Machine Learning Sate of the Union 18](#_Toc20008692)

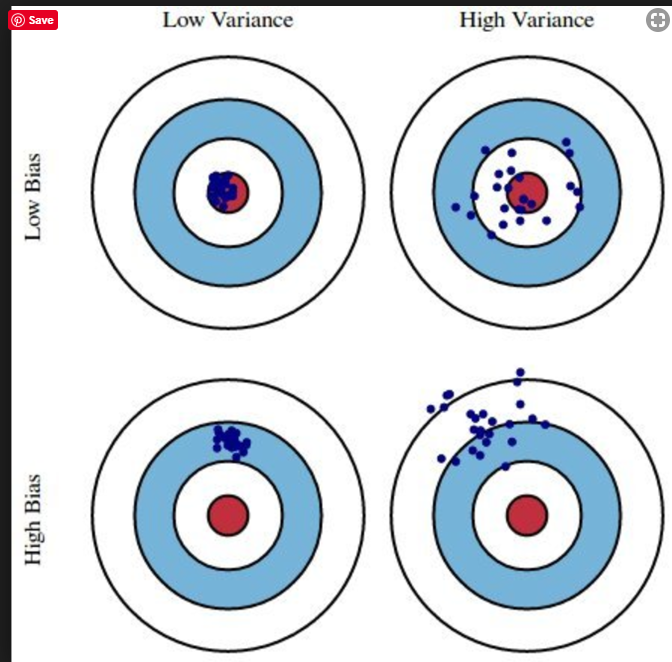
1. Machine Learning concepts / terminology
   1. Glossary of Machine Learning, Statistics and Data Science terms
      1. to infer

Use a given model in production to make inference/predictions. Get a real-world unlabeled data and ask the model to label for us.

* + 1. Bias-Variance Trade-off

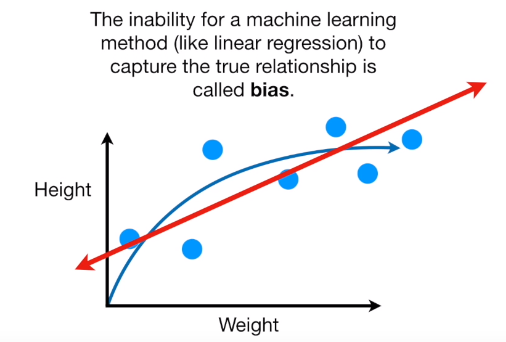
1. **Bias** error is useful to quantify how much on an average are the predicted values different from the actual value. A high bias error means we have a under-performing model which keeps on missing important trends.
2. **Variance** on the other side quantifies how are the prediction made on same observation different from each other. A high variance model will over-fit on your training population and perform badly on any observation beyond training. If an algorithm has low variability – it produces consistent predictions across different datasets.

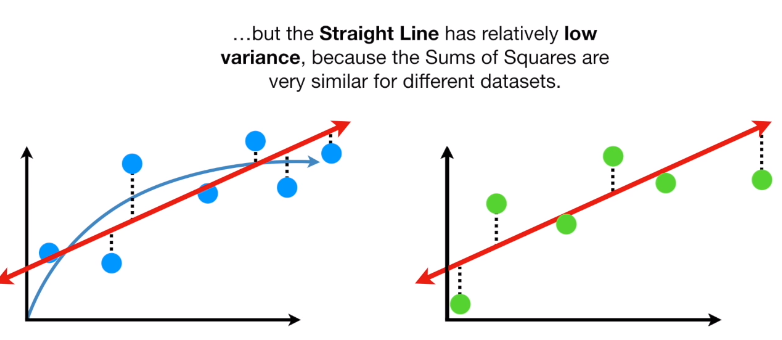
In order to have a perfect fit in the model, the bias and variance should be balanced which is bias variance trade off.

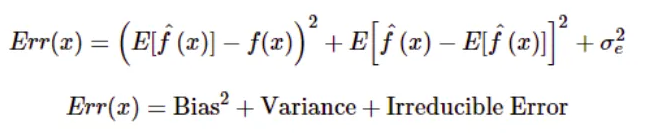
A high bias means the prediction will be inaccurate. Intuitively, bias can be thought as having a ‘bias’ towards people. If you are highly biased, you are more likely to make wrong assumptions about them. An oversimplified mindset creates an unjust dynamic: you label them accordingly to a ‘bias.’

“Bias is the algorithm’s tendency to consistently learn the wrong thing by not taking into account all the information in the data (underfitting).”

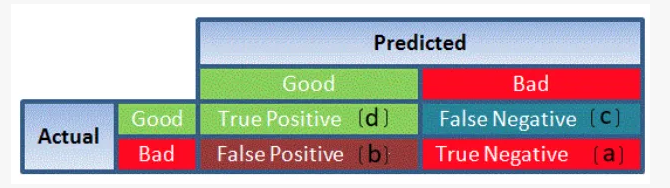
Variance is the algorithm’s tendency to learn random things irrespective of the real signal by fitting highly flexible models that follow the error/noise in the data too closely (overfitting).”







* regularization
* boosting
* bagging
  + 1. accuracy



* + 1. Cross Entropy

In information theory, the cross entropy between two probability distributions and over the same underlying set of events measures the average number of bits needed to identify an event drawn from the set, if a coding scheme is used that is optimized for an “unnatural” probability distribution , rather than the “true”. Cross entropy can be used to define the loss function in machine learning and optimization.

* + 1. Median

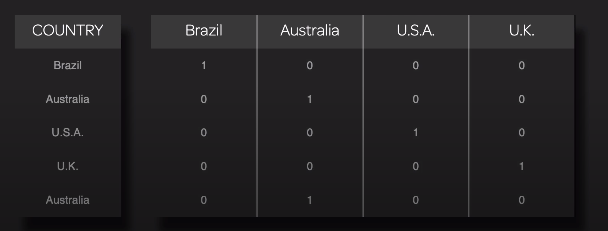
Median of a set of numbers is usually the middle value. When the total numbers in the set are even, the median will be the average of the two middle values. Median is used to measure the central tendency.

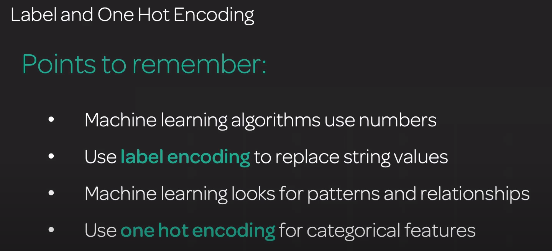
To calculate the median for a set of numbers, follow the below steps:

1. Arrange the numbers in ascending or descending order
2. Find the middle value, which will be n/2 (where n is the numbers in the set)
   * 1. PCA – Principal Component Analysis

Reduce the dimension of data in order to make it visualizabe.

* + 1. Label and One Hot Encoding





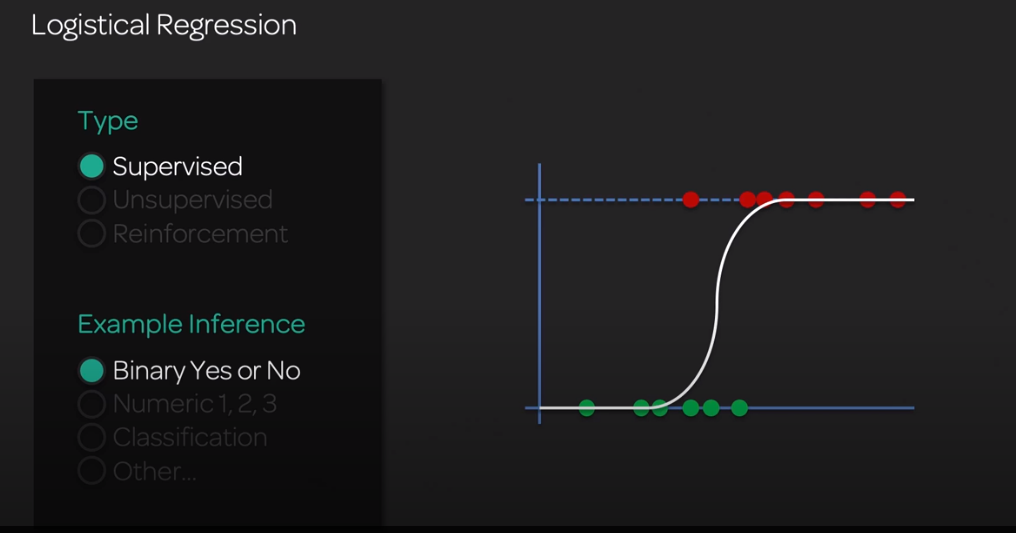
Label – replace names/countries with integers and store a table which converts back. But the ML algorithm will want to find some relationship between these numbers, maybe he will think that USA is better than UK because we replaced USA with 4 and UK with 2. This is where One Hot Encoding comes into the picture.

* + 1. Validation data

Validation Dataset: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

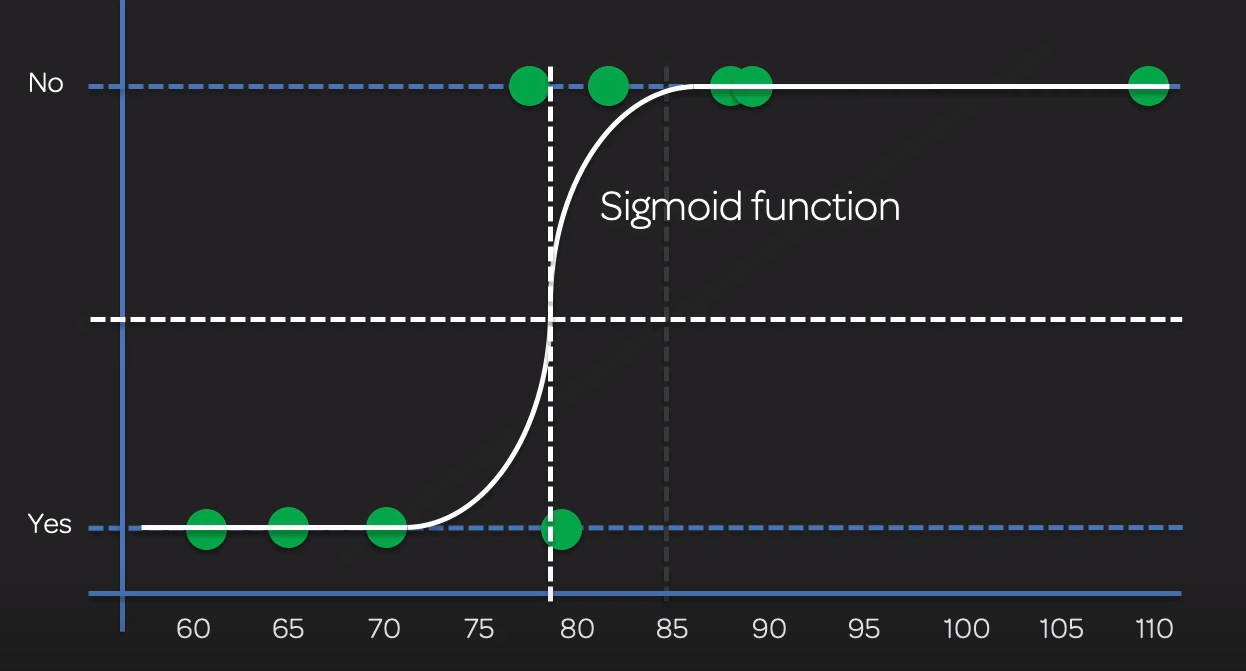
* + 1. RecordIO Format

1. AWS – Machine Learning Algorythms
   1. Logistical Regression



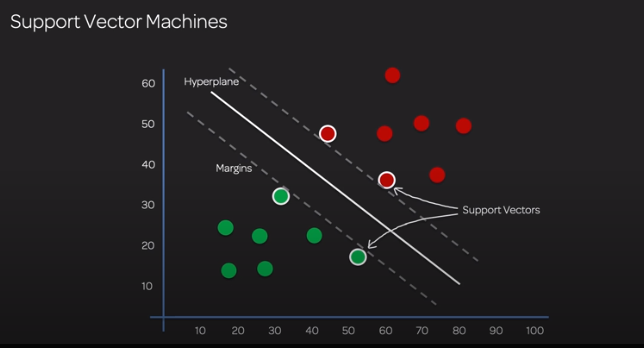


Statistical analysis tool – use when answer is ‘Yes or No’

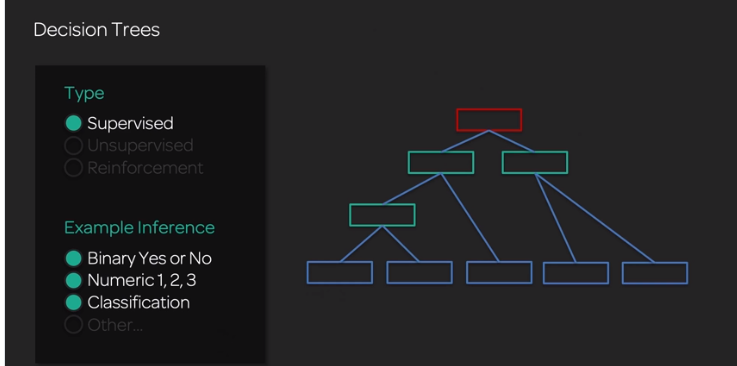


* 1. Support Vector Machines

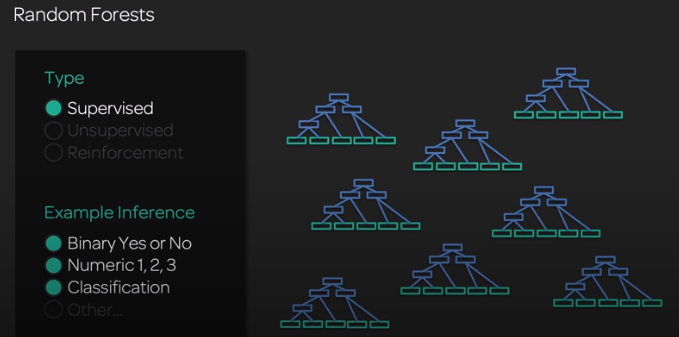
trying to determine the best separation of 2 clusters



* 1. Decision Trees



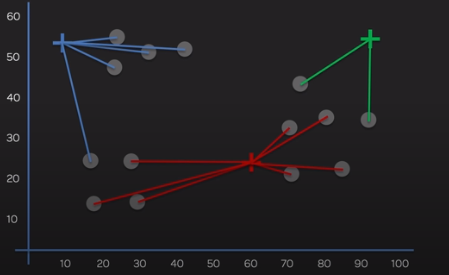
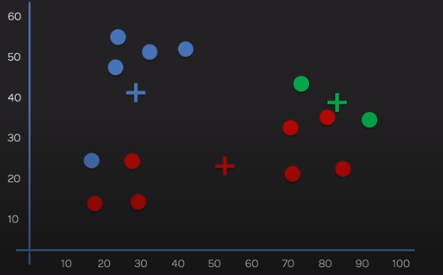
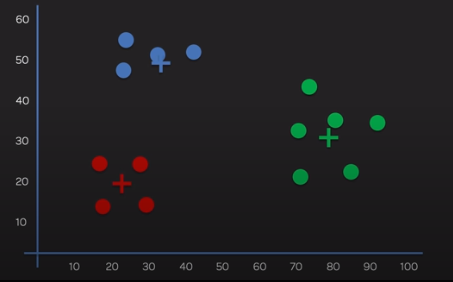
* 1. Random Forest

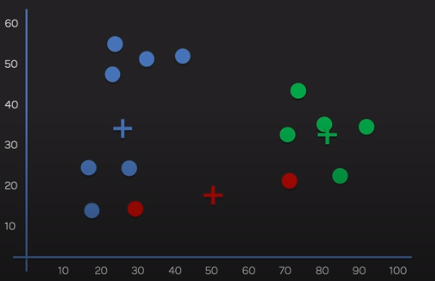


* 1. K-Means



* Data exploration
* Customer classification

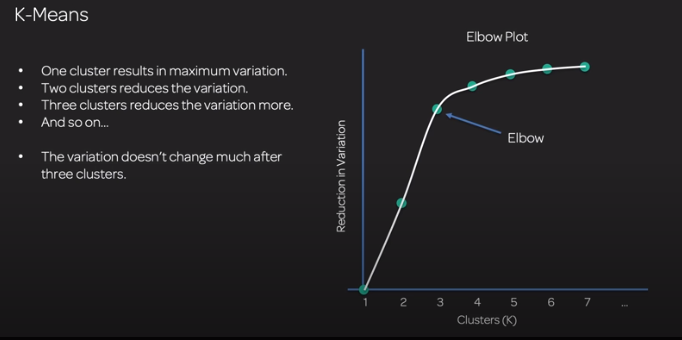
K represents a number. How much category it should find?



Total variation: determines how good our solution is.

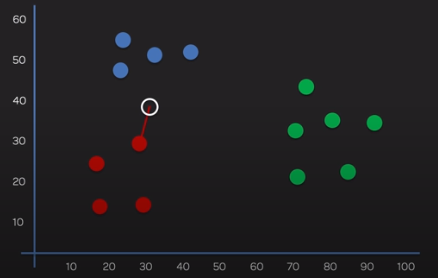
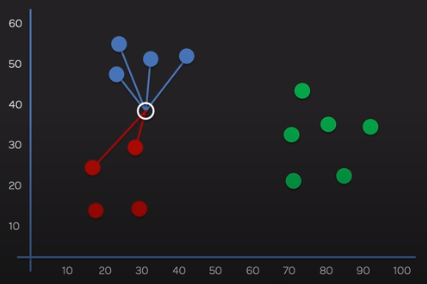


Helps us to determine the number of K (f you would have as many clusters as data points, you’d have 0 variation)



* 1. K-Nearest Neighbour



* Recommendation engine
* Similar articles and objects
* Make K large enough to reduce the influence of outliers
* Make K small enough that classes with a small sample size don’t lose influence

e.g. there are 2 clusters, one w 4 data points one w 200. After K>8 you will be closer and closer to the second one, so you don’t want to stop at e.g. K=6

1. Non-AWS technologies around machine learning
   1. Apache Spark

Apache Spark is an open-source cluster computing framework. Spark can be deployed in a variety of ways, provides native bindings for the Java, Scala, Python, and R programming languages, and supports SQL, streaming data, and machine learning. Some of the key features of Apache Spark are listed below:

* Speed − Spark helps to run an application in Hadoop cluster, up to 100 times faster in memory, and 10 times faster when running on disk
* Spark supports popular data science programming languages such as R, Python, and Scala
* Spark also has a library called MLlIB which includes basic machine learning including classification, regression, and clustering
  1. Hadoop

Hadoop is an open source distributed processing framework used when we have to deal with enormous data. It allows us to use parallel processing capability to handle big data. Here are some significant benefits of Hadoop:

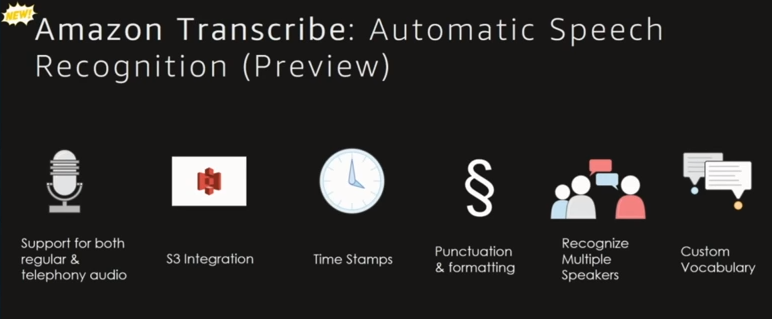
* Hadoop clusters work and keeps multiple copies to ensure reliability of data. A maximum of 4500 machines can be connected together using Hadoop
* The whole process is broken down into pieces and executed in parallel, hence saving time. A maximum of 25 Petabyte (1 PB = 1000 TB) data can be processed using Hadoop
* In case of a long query, Hadoop builds back up data-sets at every level. It also executes query on duplicate datasets to avoid process loss in case of individual failure. These steps makes Hadoop processing more precise and accurate
* Queries in Hadoop are as simple as coding in any language. You just need to change the way of thinking around building a query to enable parallel processing
  1. Keras

Keras is a simple, high-level neural network library, written in Python. It is capable of running on top of Tensorflow and Theano. This is done to make design and experiments with Neural Networks easier.

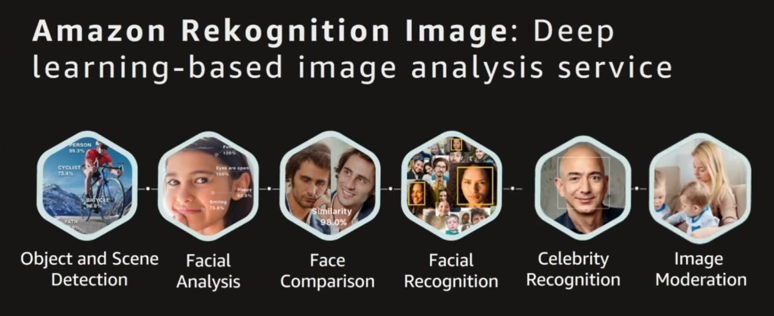
Following are some important features of Keras:

* User friendliness
* Modularity
* Easy extensibility
* Work with Python
  1. CNTK – Microsoft Cognitive Toolkit

1. AWS Services
   1. SageMaker
   2. Transcribe

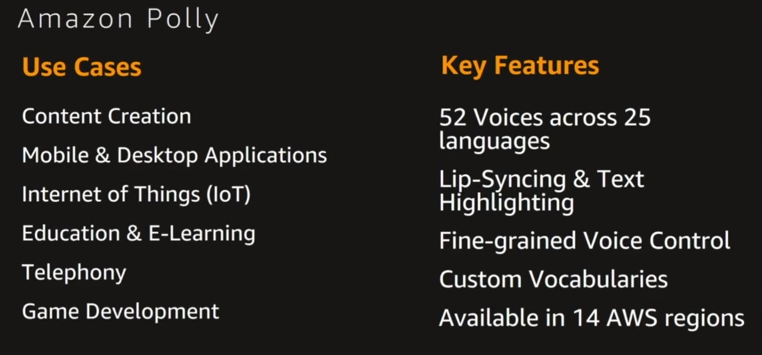


* 1. Gluon
  2. Comprehend
  3. Lex
  4. Rekognition



+ Amazon Rekognition Video. Although the video is series of images, the classical solution is not that good.

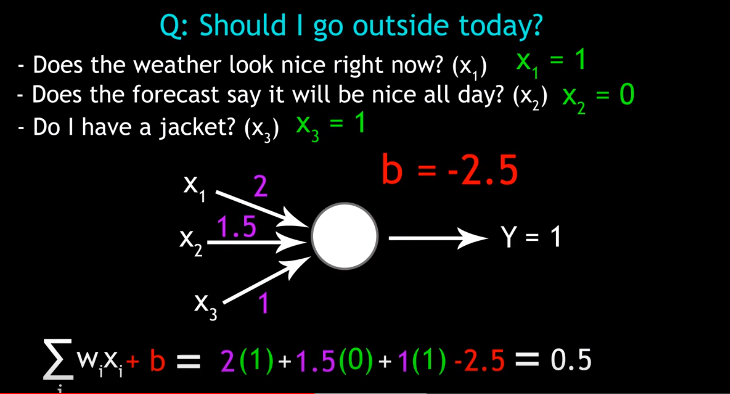
* 1. Polly



* 1. Alexa
  2. Kinesis
  3. EMR

1. Videos

Brilliant slide: Perceptron explained.



* 1. 2017:Machine Learning Sate of the Union

Framework innovation – compose the flexibility of PyTorch and scalability of Tensorfow.



**CNTK – Microsoft Cognitive Toolkit**

**Caffe2 -** CAFFE (Convolutional Architecture for Fast Feature Embedding) is a deep learning framework, originally developed at University of California, Berkeley

**Amazon Mechanical Turk (MTurk)** is a crowdsourcing website for businesses (known as Requesters) to hire remotely located "crowdworkers" to perform discrete on-demand tasks that computers are currently unable to do. It is operated under Amazon Web Services, and is owned by Amazon.[2] Employers post jobs known as Human Intelligence Tasks (HITs), such as identifying specific content in an image or video, writing product descriptions, or answering questions, among others. Workers, colloquially known as Turkers or crowdworkers, browse among existing jobs and complete them in exchange for a rate set by the employer. To place jobs, the requesting programs use an open application programming interface (API), or the more limited MTurk Requester site.[3] As of April 2019, Requesters could register from only 49 approved countries

