INTRODUCTION TO MACHINE LEARNING

- The field of study known as machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.
- Machine learning is a sub-field of artificial intelligence that involves the development of algorithms and statistical models that enable computers to improve their performance in tasks through experience. These algorithms and models are designed to learn from data and make predictions or decisions without explicit instructions.

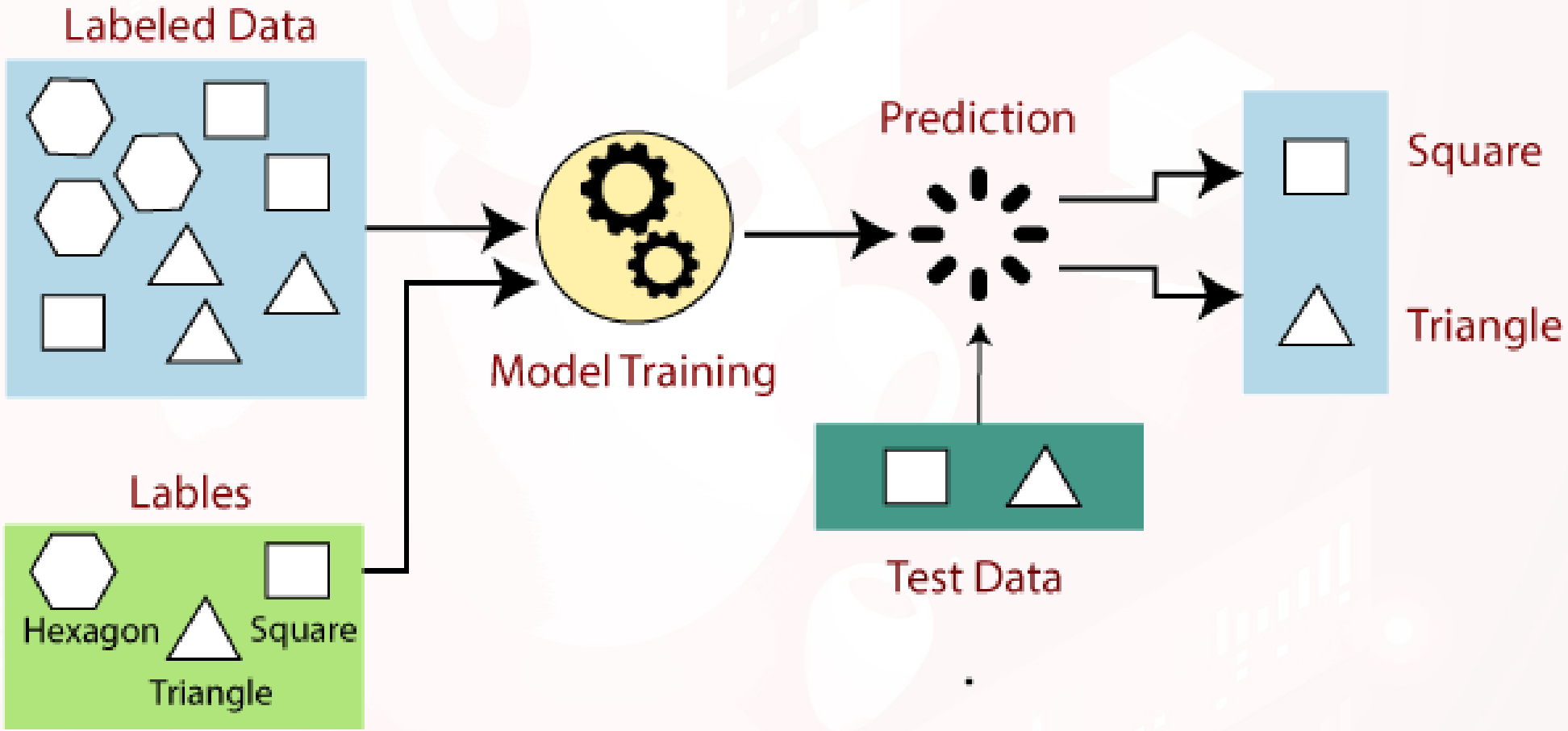
TECHNOLOGY STACK

- Anaconda
- Integrated Development Environment (IDE): Pycharm
- Python
- Jupyter Notebook

SUPERVISED MACHINE LEARNING

- The main goal in supervised learning is to learn a model from labeled training data that allows us to make predictions about unseen or future data. Here, the term supervised refers to a set of samples where the desired output signals (labels) are already known.

SUPERVISED MACHINE LEARNING



SUPERVISED ML STEPS

- Supervised machine learning model involves two steps namely:
- **Learning (Training):** Learn a model using the training data.
- **Testing:** Test the model using unseen test data to assess model accuracy.
  - Accuracy = No of correct classifications/Total no of test cases.

SUPERVISED ML PROBLEMS

- Supervised machine learning problems can be grouped into two major categories:
- **Classification Problems:** This is when the output variable is a category such as male/female, red/blue or rose/lilly.
- **Regression Problems:** A regression problem is when the output variable is a real value such as weight or house price.

COMMON SUPERVISED ML ALGORITHMS

- Some of the common supervised machine learning algorithms include:
- Decision tree.
- K-Nearest Neighbors.
- Support Vector classifier (SVC).
- Logistic Regression.
- Linear Regression.

ADVANTAGES of SUPERVISED ML

- Some of the advantages of supervised ML include:
- It allows you to be very specific with the definition of lables.
- You are able to determine the number of classes you want to have.
- The input data is very well known and is labeled.

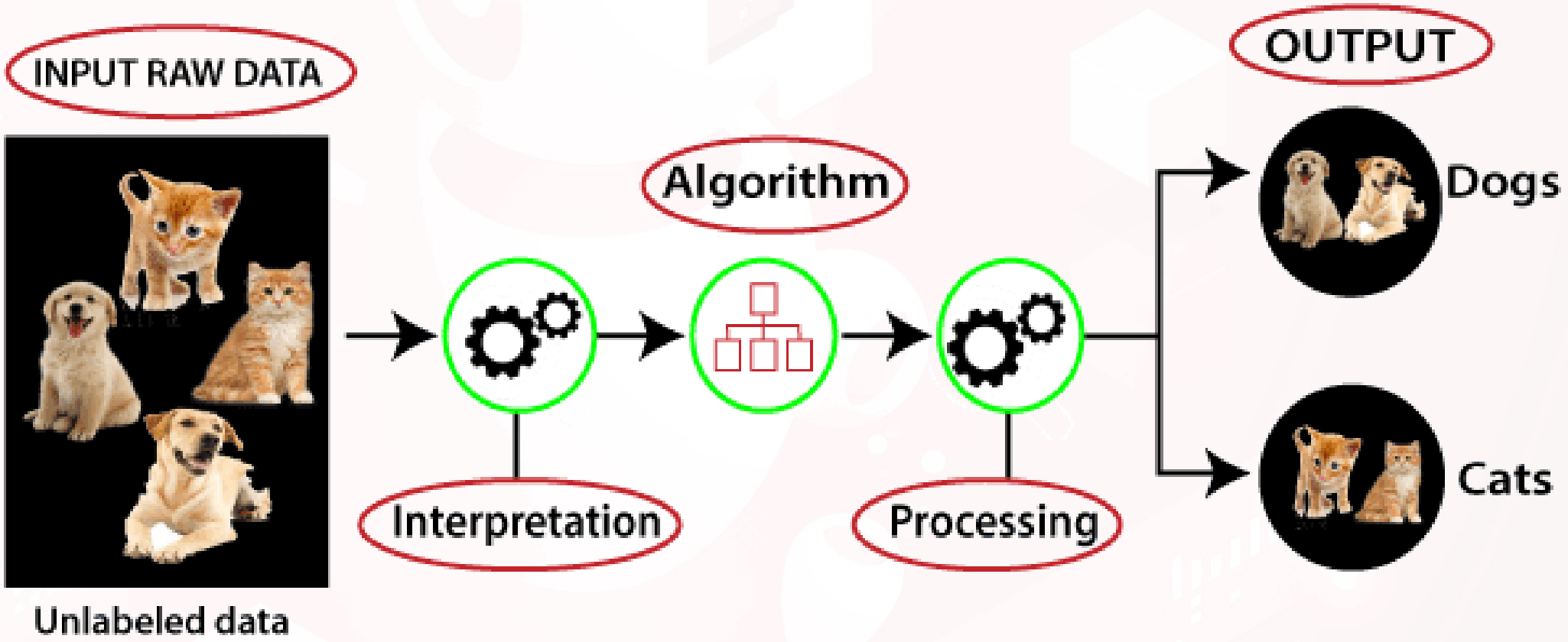
DISADVANTAGES of SUPERVISED ML

- Some of the disadvantages of supervised ML include::
- Growing data requires continuous training.
- Cannot discover new classes apart from the already labled ones during training.

UNSUPERVISED MACHINE LEARNING

- Here we only have input data X with no corresponding output variables. The goal is to learn more about the data as a result there is no correct answer and there is no teacher. The algorithm is left on its own to discover and present interesting structure in the data.

UNSUPERVISED MACHINE LEARNING



UNSUPERVISED ML GROUPS

- Supervised machine learning problems can be grouped into clustering and association problems:
- **Clustering:** A clustering problem is where you want to discover the inherent groupings in the data for example grouping customers by their purchasing power.
- **Association:** This is where you want to discover the rules that describe large portions of your data, such as people that buy X and also tend to by Y.

UNSUPERVISED ML ALGORITHMS

- Some of the common ML algorithms include:
- K-Means clustering.
- K-Nearest Neighbor (KNN).
- Dimensionality Reduction.
- Hierarchical clustering.

ADVANTAGES of UNSUPERVISED ML

- It has less complexity when compared with supervised machine learning algorithm.
-

DISADVANTAGES of UNSUPERVISED ML

- You cannot be specific about the definition of the sorting classes and output.
- Its difficult to prove the accuracy of results.
- Lack of specificity with the output classes limits its industrial applicability.

Questions

Intentionally Left Blank

RECOMMENDATION ENGINES

Recommender systems are a subclass of information filtering system that seek to predict the rating or preference that a user would give to an item.

Systems for recommending items (e.g. books, movies, CD's, web pages, newsgroup messages) to users based on examples of their preferences.

In this topic our focus will be on: Content based collaborative filtering.

TECHNIQUES OF RECOMMENDATION

1. Collaborative Filtering: This is mostly used on social media engines such as Facebook, Netflix and Twitter. The main logic behind this method is to find users who have similar taste and preferences to the target user then use this subset to provide recommendations. This method works best when you have a large number of user preferences. This approach focuses mainly on user attributes such as location and gender.

TECHNIQUES OF RECOMMENDATION

2. Content based Systems: This recommends items similar to those a user has liked (browsed/purchased) in the past. The major focus is on items and not other user's opinions.

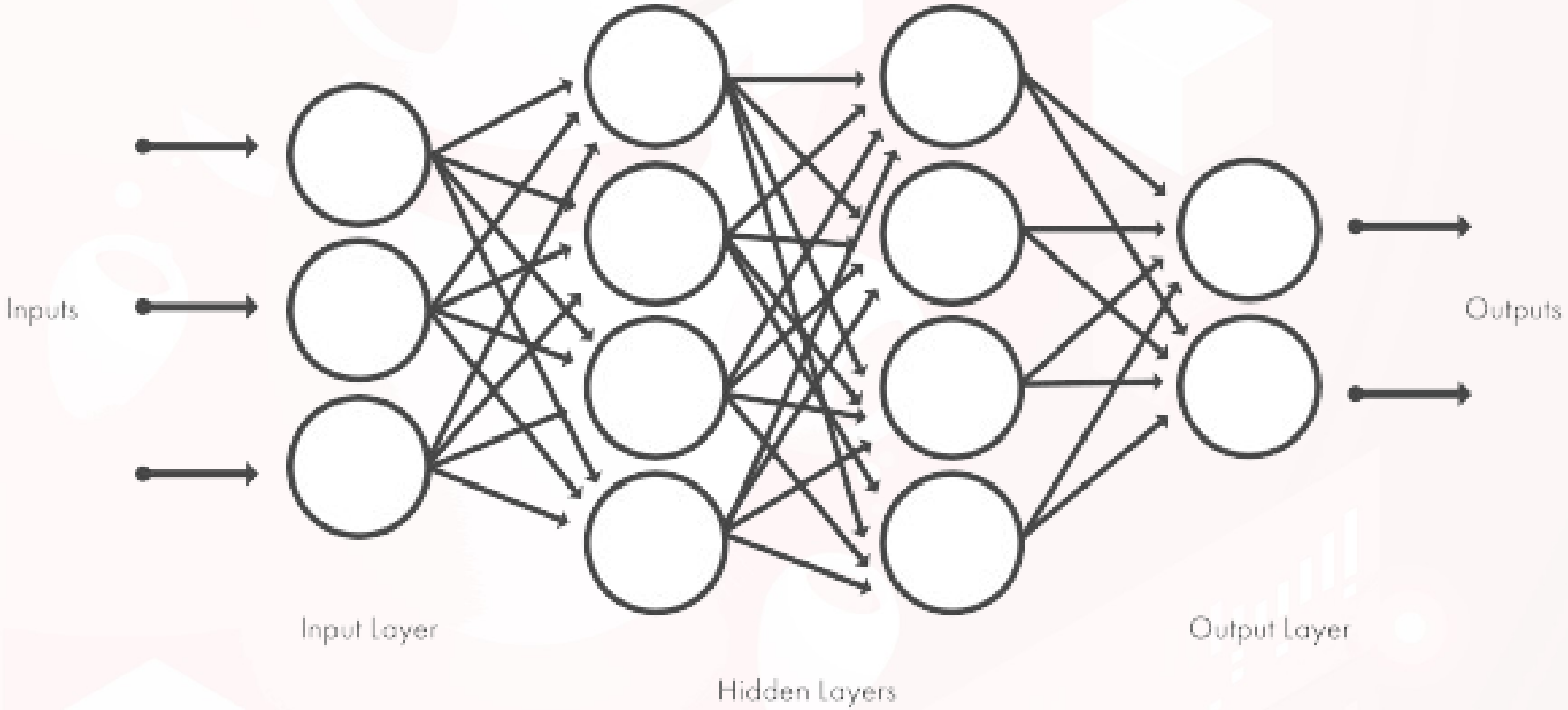
This method has the advantage of not requiring the data on other users to make a recommendation.

DEMO AND QA

DEEP LEARNING

- This is a machine learning technique that teaches computers to do what naturally comes to humans.
- Deep learning is highly used in driver-less cars enabling them to recognize traffic signs, lampposts and pedestrians and other image recognition applications.
- Deep Learning is based on artificial neural networks (ANNs) with multiple layers, also known as deep neural networks (DNNs). These neural networks are inspired by the structure and function of the human brain
- The most widely used architecture in deep learning are convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

DEEP LEARNING



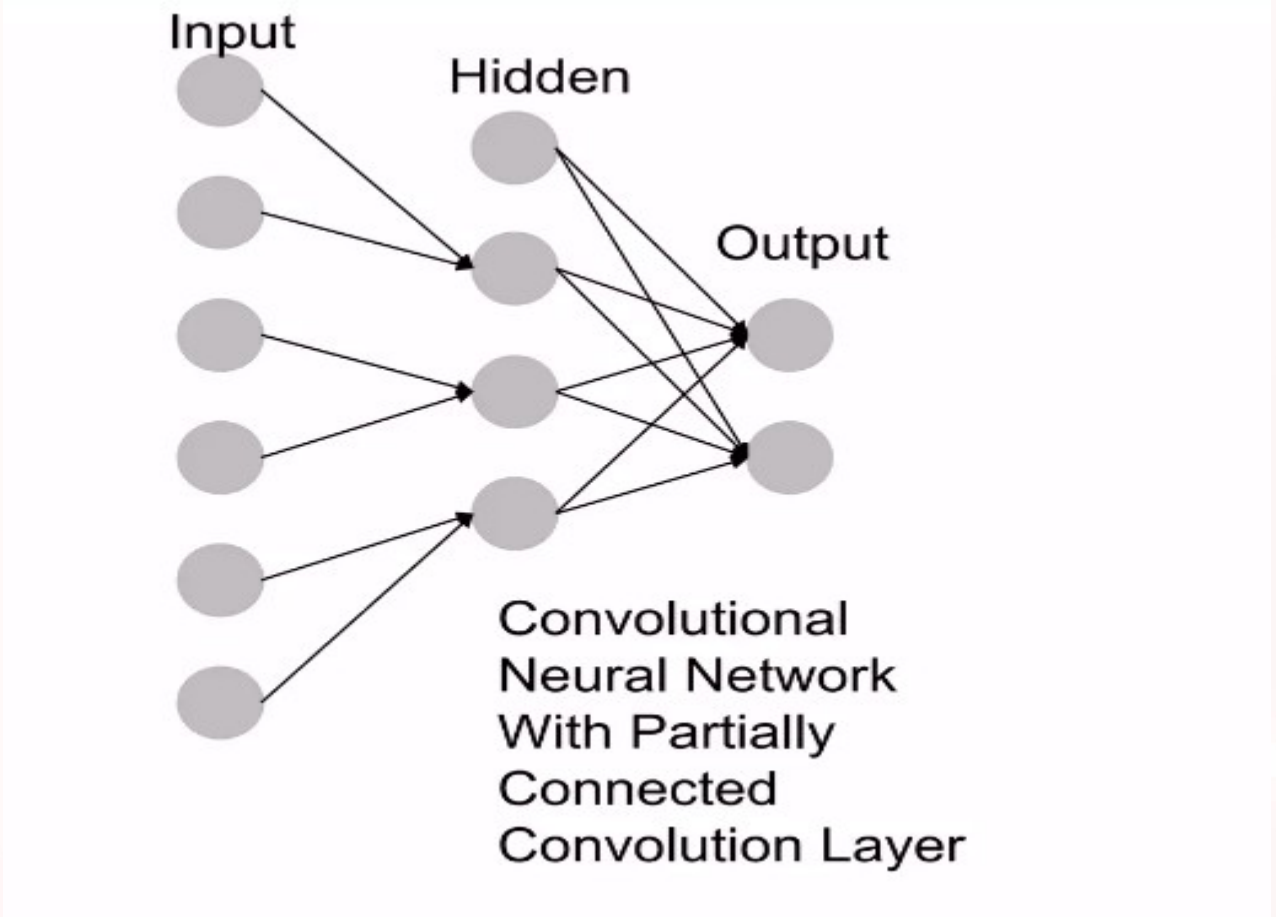
Convolutional Neural Networks

- A ConvNet is a type of feed forward artificial neural network.
- A ConveNet is made up of one or more convolutional layers and then followed by a one or more fully connected layers.
- Consider and image of size 200 * 200 * 3 (200 wide, 200 high and 3 color channels). A single fully connected neuron in the first hidden layer of the ConvNet would have
- 200 * 200 * 3 weights.
- This connectivity is wasteful and such a huge number of parameters would quickly lead to overfitting.

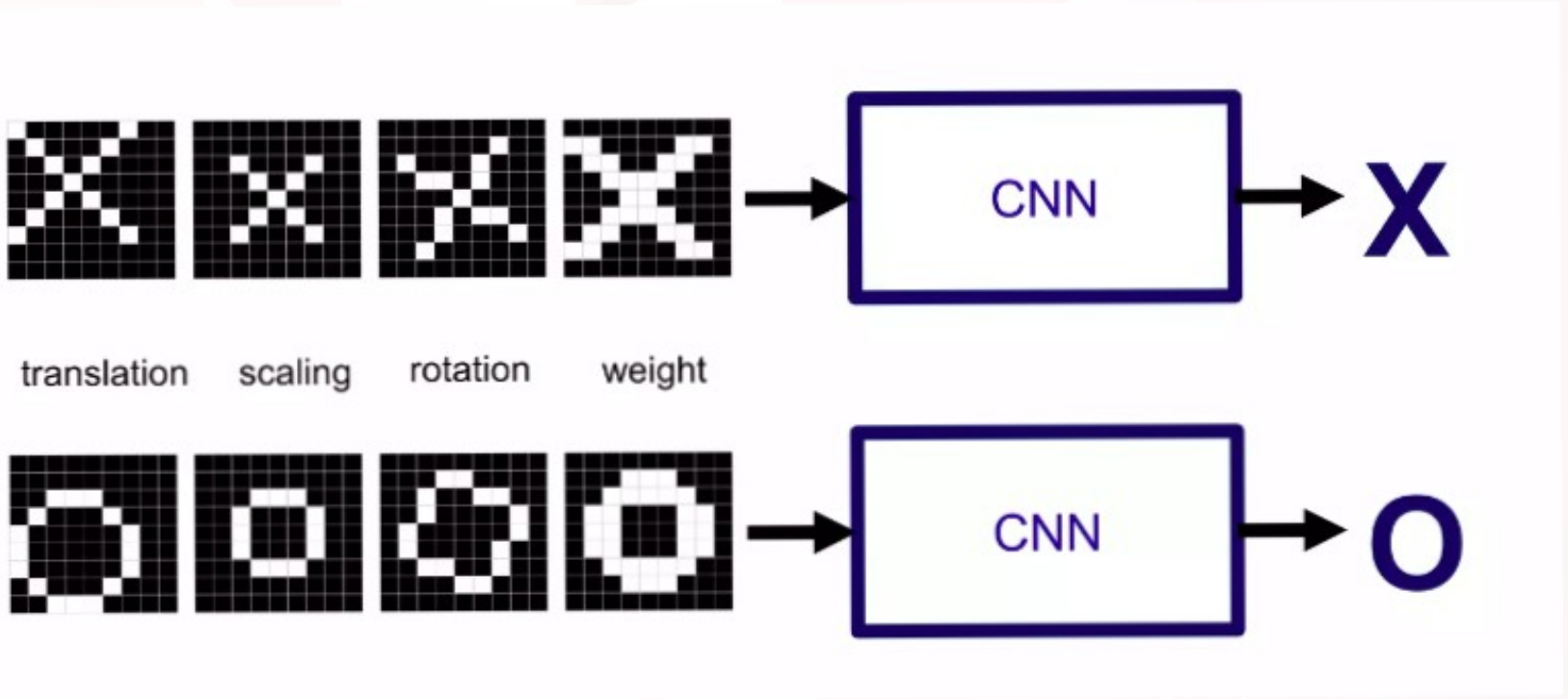
Convolutional Neural Networks

- In a ConvNet the neurons in a layer will only be connected to a small region of the layer before it instead of all the neurons in a fully-connected manner.
- This conversion happens through the convolution(featuring) layer and the pooling layers (reduction).
- The final output will be dimensions 1 *1 * N. This is because by the end of the convNet architecture we will reduce the image into a single vector of class scores (for n classes), arranged along the depth of the dimension.

Convolutional Neural Networks



Convolutional Neural Networks



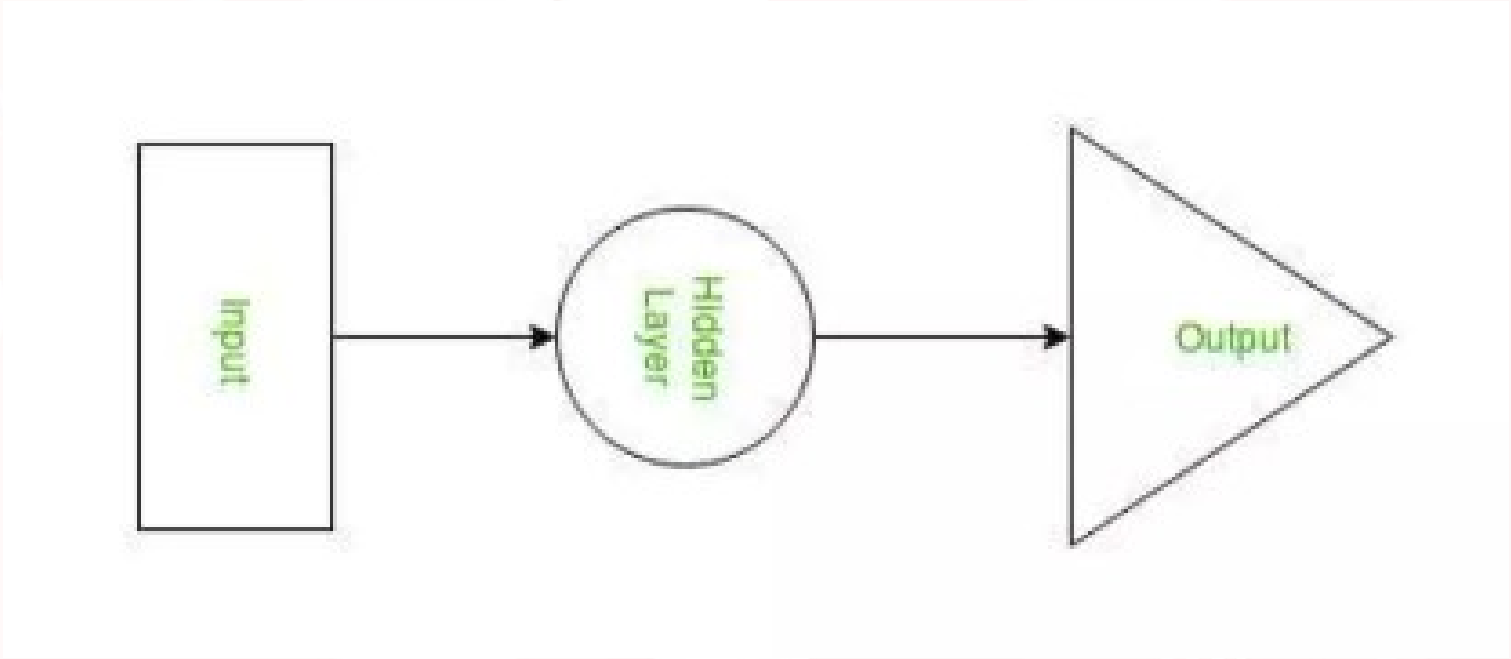
Convolutional Neural Networks Layers

- Convolution Layer: This layer is majorly responsible for featuring of the image.
- Pool layer: This layer is responsible for reduction of the features identified by convolution layer.

Recurrent Neural Networks

- Recurrent neural networks are a type of neural networks that are able to process sequential data such as time series and natural language.
- This makes them well suited for tasks such as speech recognition, natural language processing and language translation.
- All the new outputs of the network are dependent on the previously generated output for example predicting the next word in a sentence, the previous words are required and there is need to remember the previously generated words.

Recurrent Neural Networks



Recurrent Neural Networks

- The **Importance** of RNN is that for every instance you do not have to think from scratch again and again. We use the prior knowledge to get understanding of the new instance.
- Hence with RNN the machine is able to think like a human.

Training Through RNN

- A single input is provided to the network.
- Get the current state using a set of current input and previous states.
- The current state then becomes the previous state of the next time step.
- One can go on as many steps forward by joining all the information from all the previous states.
- Once all the steps are completed the final current state is used to calculate the output.
- Compare the output to the target output and get the error generated.

Intentionally left Blank

Transfer Learning

This is a research problem in ML that focuses on storing knowledge gained in while solving one problem and applying it to a different but related problem. This is mostly used with deep learning models.

Why Transfer Learning

In practice very few people train neural networks from scratch because its rare to have enough dataset.

Using pre trained network weights as initializations help solve this problem.

More so deep learning networks are expensive to train . Most models take weeks to train with very expensive GPUs.

Determining the hyper-parameters for deep learning is still mostly a black art with no official or conventionally agreed standards.

Major Scenarios with Transfer Learning

1. New data-set is a small and similar to original data-set.
2. New data-set is large and similar to the original data-set.
3. New data-set is small but very different from the original data-set.
4. New data-set is large and very different from the original data-set.

NLTK DEMO

TENSORFLOW

1. This is an open source deep learning library developed by Google.

2. Tensorflow accepts data in in the form of higher multi-dimensional arrays of higher dimensions called Tensors.

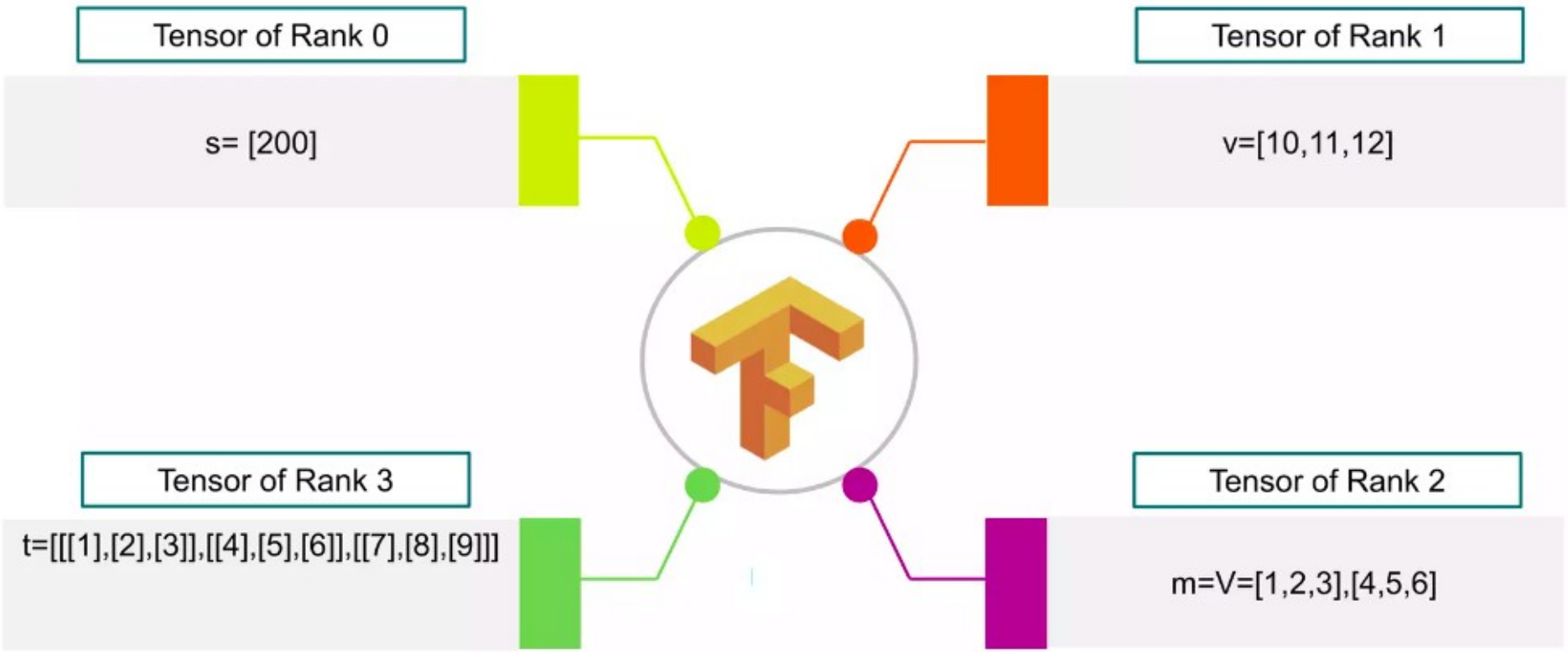
3. A **Tensor** can be defined as a geometrical objects over vector spaces, whose coordinates obey certain laws of transformation under change basis.

4. Tensors are merely a generalization of of scalars and vectors. A scalar is a zero rank tensor and a vector is a first rank tensor.

TENSORFLOW

Tensor of Dimensions[3,3,3]

TENSOR RANKS



MACHINE LEARNING APIs.

Apache Spark MLlib module is specifically designed for machine learning use cases.

It supports both Java, Scala, Python and Scala thus enhancing its usability.

It contains many machine learning algorithms some of which include:

MACHINE LEARNING APIs.

1. Classification: Logistic regression, naive Bayes
2. Regression: Linear regression.
3. Decision Trees and Random forests.
4. Recommendation: Alternating Least Square (ALS).
5. Clustering: K-Means.

MACHINE LEARNING Workflow Utilities.

1. Feature transformation: Standardization and Normalization.
2. Model evaluation and hyper-parameter tuning.
3. ML persistence: Saving and loading of models.

Intentionally left Blank

INTRODUCTION TO BIG DATA AND HADOOP.

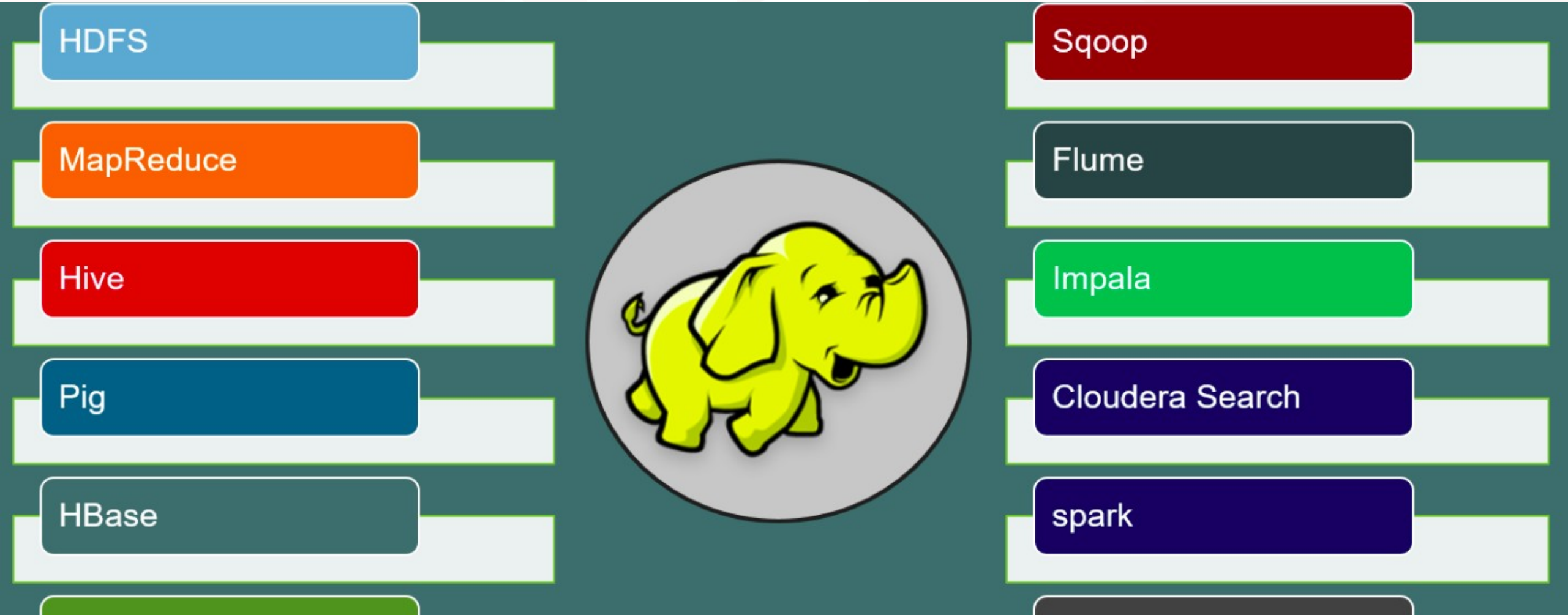
Hadoop ecosystem is neither a programming language nor a service, its a platform or framework for solving big data problems.

It can be considered as a suite which encompasses a number of services (ingesting, storing, analyzing and maintaining) inside it.

The diagram below shows the most used Hadoop ecosystem technologies:

Predictive Analytics
Drilling sense out of data

INTRODUCTION TO BIG DATA AND HADOOP.



HDFS.

Hadoop distributed file system is used scale data from a single node to upto all the nodes in the cluster.

Its one of the initial core components of Hadoop.

MapReduce.

MapR is a programing framework within the hadoop ecosystem that is used for batch processing of huge datasets.

It uses the concept of splitting the data into multiple small use data sets, works on the data sets and combines the results back together into one final out put. A process known as reducing.

Its one of the native components of hadoop.

HBase.

This is a NoSQL database that is responsible for storing the data of HDFS.

SQOOP.

This is designed to transfer data between HDFS and traditional RDMS databases such as Oracle,MSSQL, MysQL.

FLUME.

Hadoop streaming service. Flume is ideal for ingesting real time data from event streams from multiple systems.

SPARK.

This is an open source cluster computing framework that supports machine learning, business intelligence, streaming,graph analytics and batch processing.

Due to its in memory technology, its 100 times faster that MapR.

Apache PIG.

Pig compliments MapR in analyzing huge datasets using SQL like language (very loose SQL).

It easily allows the user to create UDFs for future reuse.

During compilations it converts into map reduce functions before executions.

Apache IMPALA.

This is a high performance SQL engine which runs on top of hadoop.

It supports a dilect of sql known as Impala SQL.

Apache HIVE.

This is similar to Impala. Hence its suited for data processing and ETL.

Apache Oozie.

Oozie is used for management of jobs.

Ideally Oozie is a workflow system much more like Airflow.

Apache Hue.

This is basically an SQL editor like Dbeaver. It provides an interface for running SQL queries to Hive,Impala,Mysql,SparkSQL etc.