## **1) Can you enumerate the various differences between Supervised and Unsupervised Learning?**

Ans: Supervised learning is a type of machine learning where a function is inferred from labeled training data. The training data contains a set of training examples.

Unsupervised learning, on the other hand, is a type of machine learning where inferences are drawn from datasets containing input data without labeled responses. Following are the various other differences between the two types of machine learning:

Algorithms Used – Supervised learning makes use of Decision Trees, K-nearest Neighbor algorithm, Neural Networks, Regression, and Support Vector Machines. Unsupervised learning uses Anomaly Detection, Clustering, Latent Variable Models, and Neural Networks.

Enables – Supervised learning enables classification and regression, whereas unsupervised learning enables classification, dimension reduction, and density estimation.

Use – While supervised learning is used for prediction, unsupervised learning finds use in analysis

## **2) What do you understand by the Selection Bias? What are its various types?**

Answer: Selection bias is typically associated with research that doesn’t have a random selection of participants. It is a type of error that occurs when a researcher decides who is going to be studied. On some occasions, selection bias is also referred to as the selection effect.

In other words, selection bias is a distortion of statistical analysis that results from the sample collecting method. When selection bias is not taken into account, some conclusions made by a research study might not be accurate. Following are the various types of selection bias:

Sampling Bias – A systematic error resulting due to a non-random sample of a populace causing certain members of the same to be less likely included than others that results in a biased sample.

Time Interval – A trial might be ended at an extreme value, usually due to ethical reasons, but the extreme value is most likely to be reached by the variable with the most variance, even though all variables have a similar mean.

Data – Results when specific data subsets are selected for supporting a conclusion or rejection of bad data arbitrarily.

Attrition – Caused due to attrition, i.e. loss of participants, discounting trial subjects or tests that didn’t run to completion.

## **3) Please explain the goal of A/B Testing.**

Ans: A/B Testing is a statistical hypothesis testing meant for a randomized experiment with two variables, A and B. The goal of A/B Testing is to maximize the likelihood of an outcome of some interest by identifying any changes to a webpage.

A highly reliable method for finding out the best online marketing and promotional strategies for a business, A/B Testing can be employed for testing everything, ranging from sales emails to search ads and website copy.

## **4) How will you calculate the Sensitivity of machine learning models?**

Ans: In machine learning, Sensitivity is used for validating the accuracy of a classifier, such as Logistic, Random Forest, and SVM. It is also known as REC (recall) or TPR (true positive rate).

Sensitivity can be defined as the ratio of predicted true events and total events i.e.:

Sensitivity = True Positives / Positives in Actual Dependent Variable

Here, true events are those events that were true as predicted by a machine learning model. The best sensitivity is 1.0 and the worst sensitivity is 0.0.

## **5) Could you draw a comparison between overfitting and underfitting?**

Ans: In order to make reliable predictions on general untrained data in machine learning and statistics, it is required to fit a (machine learning) model to a set of training data. Overfitting and underfitting are two of the most common modeling errors that occur while doing so.

Following are the various differences between overfitting and underfitting:

Definition – A statistical model suffering from overfitting describes some random error or noise in place of the underlying relationship. When underfitting occurs, a statistical model or machine learning algorithm fails in capturing the underlying trend of the data.

Occurrence – When a statistical model or machine learning algorithm is excessively complex, it can result in overfitting. Example of a complex model is one having too many parameters when compared to the total number of observations. Underfitting occurs when trying to fit a linear model to non-linear data.

Poor Predictive Performance – Although both overfitting and underfitting yield poor predictive performance, the way in which each one of them does so is different. While the overfitted model overreacts to minor fluctuations in the training data, the underfit model under-reacts to even bigger fluctuations.

## **6) Between Python and R, which one would you pick for text analytics, and why?**

Ans: For text analytics, Python will gain an upper hand over R due to these reasons:

The Pandas library in Python offers easy-to-use data structures as well as high-performance data analysis tools.

Python has a faster performance for all types of text analytics R is a best-fit for machine learning than mere text analysis.

## **7) Please explain the role of data cleaning in data analysis.**

Ans: Data cleaning can be a daunting task due to the fact that with the increase in the number of data sources, the time required for cleaning the data increases at an exponential rate.

This is due to the vast volume of data generated by additional sources. Also, data cleaning can solely take up to 80% of the total time required for carrying out a data analysis task.

Nevertheless, there are several reasons for using data cleaning in data analysis. Two of the most important ones are:

Cleaning data from different sources helps in transforming the data into a format that is easy to work with Data cleaning increases the accuracy of a machine learning model

## **8) What do you mean by cluster sampling and systematic sampling?**

Ans: When studying the target population spread throughout a wide area becomes difficult and applying simple random sampling becomes ineffective, the technique of cluster sampling is used. A cluster sample is a probability sample, in which each of the sampling units is a collection or cluster of elements.

Following the technique of systematic sampling, elements are chosen from an ordered sampling frame. The list is advanced in a circular fashion. This is done in such a way so that once the end of the list is reached, the same is progressed from the start, or top, again.

## **9) Please explain Eigenvectors and Eigenvalues.**

Ans: Eigenvectors help in understanding linear transformations. They are calculated typically for a correlation or covariance matrix in data analysis.

In other words, eigenvectors are those directions along which some particular linear transformation acts by compressing, flipping, or stretching.

Eigenvalues can be understood either as the strengths of the transformation in the direction of the eigenvectors or the factors by which the compressions happens.

## **10) Can you compare the validation set with the test set?**

Ans: A validation set is part of the training set used for parameter selection as well as for avoiding overfitting of the machine learning model being developed. On the contrary, a test set is meant for evaluating or testing the performance of a trained machine learning model.

## **11) What do you understand by linear regression and logistic regression?**

Ans: Linear regression is a form of statistical technique in which the score of some variable Y is predicted on the basis of the score of a second variable X, referred to as the predictor variable. The Y variable is known as the criterion variable.

Also known as the logit model, logistic regression is a statistical technique for predicting the binary outcome from a linear combination of predictor variables.

## **12) Please explain Recommender Systems along with an application.**

Ans: Recommender Systems is a subclass of information filtering systems, meant for predicting the preferences or ratings awarded by a user to some product.

An application of a recommender system is the product recommendations section in Amazon. This section contains items based on the user’s search history and past orders.

## **13) What are outlier values and how do you treat them?**

Ans: Outlier values, or simply outliers, are data points in statistics that don’t belong to a certain population. An outlier value is an abnormal observation that is very much different from other values belonging to the set.

Identification of outlier values can be done by using univariate or some other graphical analysis method. Few outlier values can be assessed individually but assessing a large set of outlier values require the substitution of the same with either the 99th or the 1st percentile values.

There are two popular ways of treating outlier values:

To change the value so that it can be brought within a range.

To simply remove the value.

## **14) Please enumerate the various steps involved in an analytics project.**

Ans: Following are the numerous steps involved in an analytics project:

Understanding the business problem.

Exploring the data and familiarizing with the same.

Preparing the data for modeling by means of detecting outlier values, transforming variables, treating missing values, et cetera.

Running the model and analyzing the result for making appropriate changes or modifications to the model (an iterative step that repeats until the best possible outcome is gained).

Validating the model using a new dataset. Implementing the model and tracking the result for analyzing the performance of the same.

## **15) Could you explain how to define the number of clusters in a clustering algorithm?**

Ans: The primary objective of clustering is to group together similar identities in such a way that while entities within a group are similar to each other, the groups remain different from one another

Generally, the Within Sum of Squares is used for explaining the homogeneity within a cluster. For defining the number of clusters in a clustering algorithm, WSS is plotted for a range pertaining to a number of clusters. The resultant graph is known as the Elbow Curve.

The Elbow Curve graph contains a point that represents the point post in which there aren’t any decrements in the WSS. This is known as the bending point and represents K in K–Means.

Although the aforementioned is the widely-used approach, another important approach is the Hierarchical clustering. In this approach, dendrograms are created first and then distinct groups are identified from there.

## **16) What is Data Science?**

Ans: Data Science is a combination of algorithms, tools, and machine learning technique which helps you to find common hidden patterns from the given raw data.

## **17) What is logistic regression in Data Science?**

Ans: Logistic Regression is also called as the logit model. It is a method to forecast the binary outcome from a linear combination of predictor variables.

## **18) Name three types of biases that can occur during sampling**

Ans: In the sampling process, there are three types of biases, which are:

* Selection bias
* Under coverage bias
* Survivorship bias

## **19) Discuss Decision Tree algorithm**

Ans: A decision tree is a popular supervised machine learning algorithm. It is mainly used for Regression and Classification. It allows breaks down a dataset into smaller subsets. The decision tree can able to handle both categorical and numerical data.

## **20) What is Prior probability and likelihood?**

Ans: Prior probability is the proportion of the dependent variable in the data set while the likelihood is the probability of classifying a given observant in the presence of some other variable.

## **21) Explain Recommender Systems?**

Ans: It is a subclass of information filtering techniques. It helps you to predict the preferences or ratings which users likely to give to a product.

## **22) Name three disadvantages of using a linear model**

Ans: Three disadvantages of the linear model are:

The assumption of linearity of the errors.

You can’t use this model for binary or count outcomes.

There are plenty of overfitting problems that it can’t solve.

## **23) Why do you need to perform resampling?**

Ans: Resampling is done in below-given cases:

Estimating the accuracy of sample statistics by drawing randomly with replacement from a set of the data point or using as subsets of accessible data.

Substituting labels on data points when performing necessary tests Validating models by using random subsets.

## **24) List out the libraries in Python used for Data Analysis and Scientific Computations. SciPy**

* Pandas
* Matplotlib
* NumPy
* SciKit
* Seaborn

## **25) What is Power Analysis?**

Ans: The power analysis is an integral part of the experimental design. It helps you to determine the sample size requires to find out the effect of a given size from a cause with a specific level of assurance. It also allows you to deploy a particular probability in a sample size constraint.

## **26) Explain Collaborative filtering**

Ans: Collaborative filtering used to search for correct patterns by collaborating viewpoints, multiple data sources, and various agents.

## **27) What is bias?**

Ans: Bias is an error introduced in your model because of the oversimplification of a machine learning algorithm.” It can lead to underfitting.

## **28) Discuss ‘Naive’ in a Naive Bayes algorithm?**

Ans: The Naive Bayes Algorithm model is based on the Bayes Theorem. It describes the probability of an event. It is based on prior knowledge of conditions which might be related to that specific event.

## **29) What is a Linear Regression?**

Ans: Linear regression is a statistical programming method where the score of a variable ‘A’ is predicted from the score of a second variable ‘B’. B is referred to as the predictor variable and A as the criterion variable.

## **30) State the difference between the expected value and mean value**

Ans: They are not many differences, but both of these terms are used in different contexts. Mean value is generally referred to when you are discussing a probability distribution whereas expected value is referred to in the context of a random variable.

## **31) How do you check for data quality?**

Ans. Some of the definitions used to check for data quality are:

* Completeness
* Consistency
* Uniqueness
* Integrity
* Conformity
* Accuracy

## **32) Suppose you are given survey data, and it has some missing data, how would you deal with missing values ​​from that survey?**

Ans. This is among the important data science interview questions. There are two main techniques for dealing with missing values –

Debugging Techniques – It is a Data Cleaning process consisting of evaluating the quality of the information collected, increasing its quality, in order to avoid lax analysis. The most popular debugging techniques are –

Searching the list of values: It is about searching the data matrix for values ​​that are outside the response range. These values ​​can be considered as missing, or the correct value can be estimated from other variables

Filtering questions: It is about comparing the number of responses of a filter category and another filtered category. If any anomaly is observed that cannot be solved, it will be considered as a lost value.

Checking for Logical Consistencies: The answers that may be considered contradictory to each other are checked.

Counting the Level of representativeness: A count is made of the number of responses obtained in each variable. If the number of unanswered questions is very high, it is possible to assume equality between the answers and the non-answers or to make an imputation of the non-answer.

Imputation Technique

This technique consists of replacing the missing values ​​with valid values ​​or answers by estimating them. There are three types of imputation:

Random imputation

Hot Deck imputation

Imputation of the mean of subclasses

## **33) How would you deal with missing random values ​​from a data set?**

Ans. There are two forms of randomly missing values:

MCAR or Missing completely at random. Such errors happen when the missing values are randomly distributed across all observations.

We can confirm this error by partitioning the data into two parts –

One set with the missing values

Another set with the non-missing values.

After we have partitioned the data, we conduct a t-test of mean difference to check if there is any difference in the sample between the two data sets.

In case the data are MCAR, we may choose a pair-wise or a list-wise deletion of missing value cases.

MAR or Missing at random. It is a common occurrence. Here, the missing values are not randomly distributed across observations but are distributed within one or more sub-samples. We cannot predict the probability from the variables in the model. Data imputation is mainly performed to replace them.

## **34) What is Hadoop, and why should I care?**

Ans. Hadoop is an open-source processing framework that manages data processing and storage for big data applications running on pooled systems.

Apache Hadoop is a collection of open-source utility software that makes it easy to use a network of multiple computers to solve problems involving large amounts of data and computation. It provides a software framework for distributed storage and big data processing using the MapReduce programming model.

Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packets of code to nodes to process the data in parallel. This allows the data set to be processed faster and more efficiently than if conventional supercomputing architecture were used.

## **35) What is ‘fsck’?**

Ans. ‘fsck ‘ abbreviation for ‘ file system check.’ It is a type of command that searches for possible errors in the file. fsck generates a summary report, which lists the file system’s overall health and sends it to the Hadoop distributed file system.

This is among the important data science interview questions and you must prepare for the related terminologies as well.

## **36) Which is better – good data or good models?**

Ans. This might be one of the frequently asked data science interview questions.

The answer to this question is very subjective and depends on the specific case. Big companies prefer good data; it is the foundation of any successful business. On the other hand, good models couldn’t be created without good data.

Based on your personal preference, you will probably choose no right or wrong answer (unless the company requires one specifically).

## **37) What are Recommender Systems?**

Ans. Recommender systems are a subclass of information filtering systems, used to predict how users would rate or score particular objects (movies, music, merchandise, etc.). Recommender systems filter large volumes of information based on the data provided by a user and other factors, and they take care of the user’s preference and interest.

Recommender systems utilize algorithms that optimize the analysis of the data to build the recommendations. They ensure a high level of efficiency as they can associate elements of our consumption profiles such as purchase history, content selection, and even our hours of activity, to make accurate recommendations.

## **38) Differentiate between wide and long data formats.**

Ans. In a wide format, categorical data are always grouped.

The long data format is in which there are a number of instances with many variables and subject variables.

## **39) What are Interpolation and Extrapolation?**

Ans. Interpolation – This is the method to guess data points between data sets. It is a prediction between the given data points.

Extrapolation – This is the method to guess data point beyond data sets. It is a prediction beyond given data points.

## **40) How much data is enough to get a valid outcome?**

Ans. All the businesses are different and measured in different ways. Thus, you never have enough data and there will be no right answer. The amount of data required depends on the methods you use to have an excellent chance of obtaining vital results.

## **41) What is a Microsoft Azure Solution Architect?**

Ans. The Azure Solution Architect is a leadership position, he/she drives revenue and market share providing customers with insights and solutions leveraging the Microsoft Azure services to meet their application, infrastructure, and data modernization and cloud needs, to uncover and support the business and IT goals of our customers.

This role will demonstrate the business value of the Microsoft Platform and drive technical decisions at the customer, thus securing long-term sustainable growth for Microsoft.

## **42) What are the different cloud deployment models?**

Ans: Following are the three cloud deployment models:

Public Cloud: The infrastructure is owned by your cloud provider and the server that you are using could be a multi-tenant system.

Private Cloud: The infrastructure is owned by you or your cloud provider gives you that service exclusively. For eg: Hosting your website on your servers, or hosting your website with the cloud provider on a dedicated server.

Hybrid Cloud: When you use both Public Cloud, Private Cloud together, it is called Hybrid Cloud. For Example: Using your in-house servers for confidential data, and the public cloud for hosting your company’s public-facing website. This type of setup would be a hybrid cloud.

Go through this Microsoft Azure Blog to get a clear understanding of Cloud Deployment Models!

## **43) I have some private servers on my premises, also I have distributed some of my workloads on the public cloud, what is this architecture called?**

Ans: This type of architecture would be a hybrid cloud. Why? Because we are using both, the public cloud, and on-premises servers i.e the private cloud.

## **44) What are the three main components of the Windows Azure platform?**

Ans: Three most important components of the Windows Azure platform are:

* Compute
* Storage
* AppFabric

You can find these components in the form of Azure Compute, Azure AppFabric, and Azure Storage.

## **45) Explain the advantage of the Azure CDN?**

Ans: Azure CDN stands for Content Delivery Network. It has three advantages: quick responsiveness, help in saving the bandwidth and reduce the load time.

## **46) Explain the importance of the Azure HDInsight?**

Ans: HDInsight is part of Hadoop components. It helps in processing a huge amount of data in an effective, smooth and quick manner. It even provides full control to manage the configuration of the clusters and software installed.

## **47) Define the Role in Azure?**

Ans: In simple language, it can be understood as the set of permission that helps in performing read and write operation. Azure RBAC contain around 120 roles.

## **48) Explain the deployments slot in Azure**

Ans: Deployment slots are present under the Azure Web App Service. There are two types of slot present in Azure Web App: Production slot and Staging slot. The production slot is the default one in which the app runs, but staging slots are the ones that help in testing the application usability before promoting to the production slot.

## **49) How two Virtual Network can communicate with each other?**

Ans: To establish communication between two Virtual Network we need to create a Gateway subnet. The gateway subnet is configured while specifying the range of the Virtual network. It takes the use of IP addresses to specify the quantity of subnet to be contained.

## **50) What are the different types of Storage areas in Azure?**

Ans: BLOB: BLOBs offer a component for storing a lot of content or binary data, for example, pictures, audio, and visual documents. They can scale up to 200 terabytes and can be acquired by utilizing REST APIs.

Table: Tables represent storage areas across machines for information that is in the form of properties on the cloud.

Queue: The sole target of a queue is to empower communication among Web and Worker Role instances. They help in storing messages that may be accessed by a customer.

## **51) Give the various advantages of using Azure ML Studio**

Ans: Azure ML Studio is the most popular features as it has a complete package that helps in Classification, Ranking, Clustering, Anomaly Detection, Recommendation, and Ranking. Due to the presence of drag and drop utility, processes become easy to perform. The various framework supported by the Azure ML Studio includes TensorFlow, SparkML, Microsoft Cognitive Toolkit and so on.

### *Extracting value from large amounts of data {and making human sense of it is one of the primary challenge of data science*

### *Introduction to Data Science*

1.Find the data

2.Extract and acquire the data

3.Clean and transform the data

4.Understand the relationships in the data  
and build a model

5.Mine for additional data

6.Evaluate and refine the model

7.Communicate the results

### *Applied Data Science at Microsoft*



Cortana {contextual understanding}

Bing {ranking, recommendations}

Windows 10 {Insider telemetry}

Windows Hello {next gen security}

Skype Translator {voice input}

HoloLens {object recognition}

Xbox {gamer matchmaking}

Microsoft Research {innovation}

### *Microsoft Cognitive Toolkit*

Speed & scalability {while maintaining accuracy}

Commercial-grade AI {used in Cortana & Bing}

Familiar {use Python on Linux or Windows}

C#, R, Spark, Docker, and other open source coming soon

### *So What is CNTK*

CNTK expresses (nearly) arbitrary neural networks by composing simple building blocks into complex computational networks, supporting relevant network types and applications.

CNTK is production-ready: State-of-the-art accuracy, efficient, and scales to multi-GPU/multi-server.

Lego-like composability allows CNTK to support a wide range of networks, e.g.

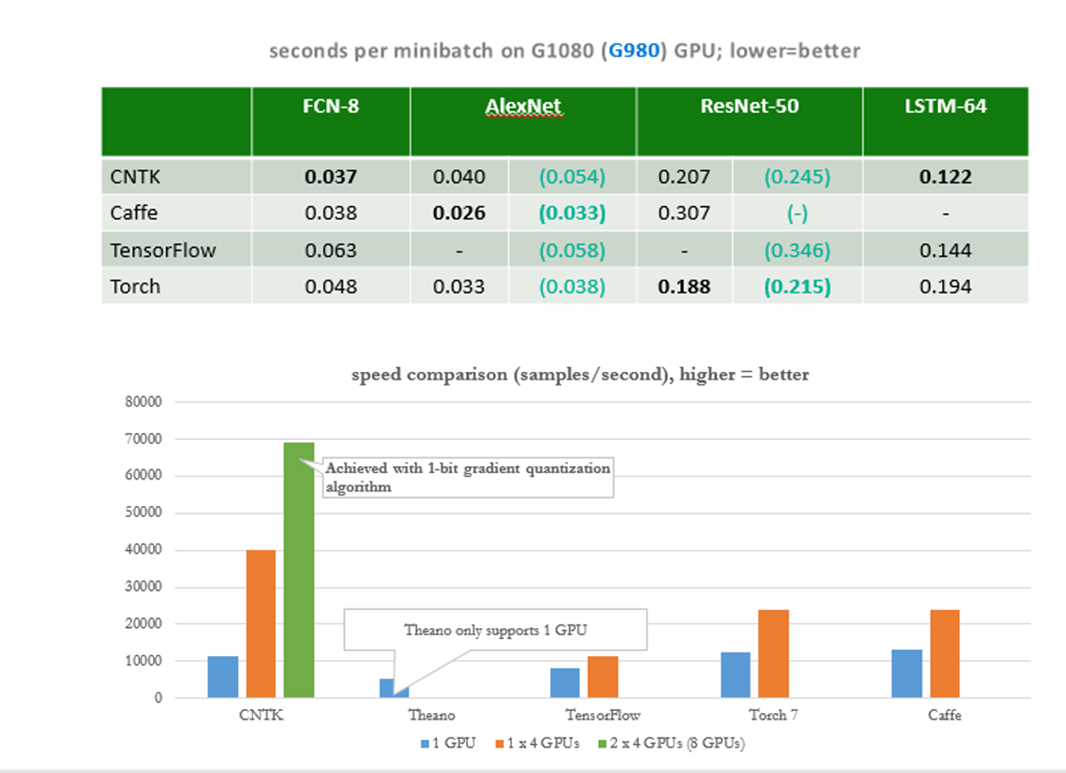
* feed-forward DNN
* RNN, LSTM, GRU
* convolution
* DSSM

sequence-to-sequence

CNTK is ideal for a range of applications including

1. speech
2. vision
3. text
4. and combinations
5. Large data sets is the key here; performant

Microsoft’s open-source deep-learning toolkit



* Ease of use: what, not how
* Fast
* Flexible
* First class on Linux and Windows
* OpenSource

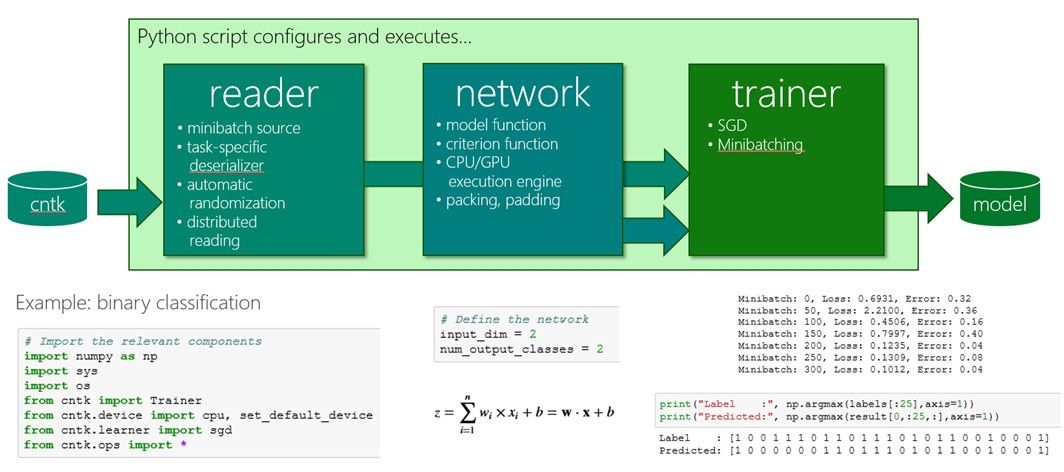
### *Getting Started with CNTK*

[https://notebooks.azure.com/n/1zbIwzaANic/notebooks/CNTK\_101\_LogisticRegression.ipynb](https://notebooks.azure.com/n/1zbIwzaANic/notebooks/CNTK_101_LogisticRegression.ipynb" \t "https://techcommunity.microsoft.com/t5/educator-developer-blog/deep-learning-with-microsoft-cognitive-toolkit-cntk/ba-p/_blank)

1.Configure reader, network, learner

2.Train & evaluate (multi-phase for layer building)

3.Deploy offline from Python



Binaries & Tutorials

[https://github.com/Microsoft/CNTK/wiki](https://github.com/Microsoft/CNTK/wiki" \t "https://techcommunity.microsoft.com/t5/educator-developer-blog/deep-learning-with-microsoft-cognitive-toolkit-cntk/ba-p/_blank)

Deep-learning VM Toolkit [https://azure.microsoft.com/en-us/marketplace/partners/microsoft-ads/dsvm-deep-learningtoolkit/...](https://azure.microsoft.com/en-us/marketplace/partners/microsoft-ads/dsvm-deep-learningtoolkit/" \t "https://techcommunity.microsoft.com/t5/educator-developer-blog/deep-learning-with-microsoft-cognitive-toolkit-cntk/ba-p/_blank)

Deep-dive lecture & labs

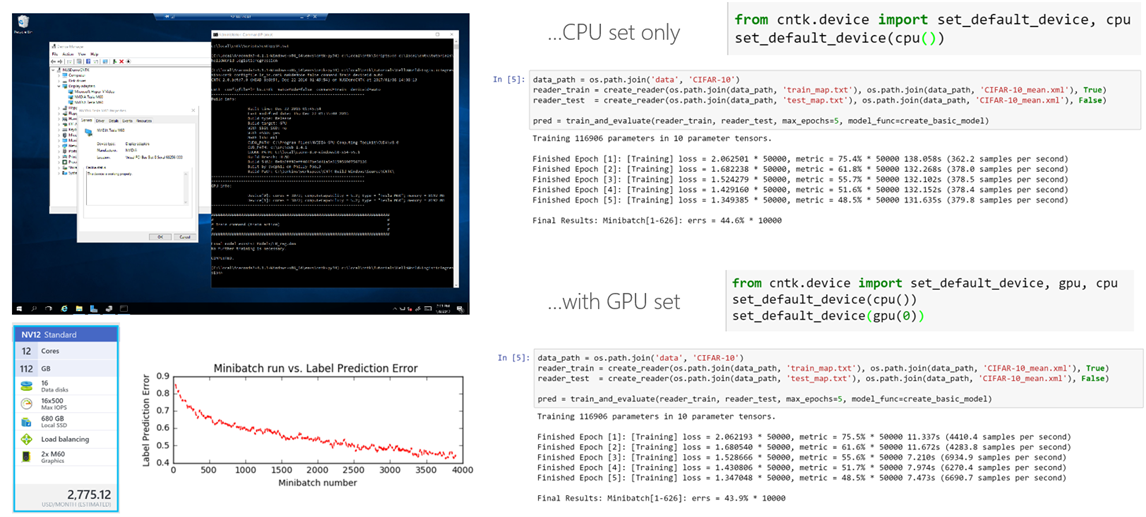
[https://www.youtube.com/watch?v=pl-kbFJ1KzU](https://www.youtube.com/watch?v=pl-kbFJ1KzU" \t "https://techcommunity.microsoft.com/t5/educator-developer-blog/deep-learning-with-microsoft-cognitive-toolkit-cntk/ba-p/_blank)

Team Q&A

[http://stackoverflow.com/questions/tagged/cntk](http://stackoverflow.com/questions/tagged/cntk" \t "https://techcommunity.microsoft.com/t5/educator-developer-blog/deep-learning-with-microsoft-cognitive-toolkit-cntk/ba-p/_blank)

Escience Whats New in CNTK 2.0 [https://esciencegroup.com/2016/11/10/cntk-revisited-a-new-deep-learning-toolkit-release-from-mi...](https://esciencegroup.com/2016/11/10/cntk-revisited-a-new-deep-learning-toolkit-release-from-microsoft/" \o "https://esciencegroup.com/2016/11/10/cntk-revisited-a-new-deep-learning-toolkit-release-from-microsoft/" \t "https://techcommunity.microsoft.com/t5/educator-developer-blog/deep-learning-with-microsoft-cognitive-toolkit-cntk/ba-p/_blank)

### *CNTK with GPU Example Using Azure NV12 Series Servers with 12 Cores 128GB Ram and 2 x M60 NVIDIA GPU*



### *Getting Started with Microsoft Cognitive Toolkit  Webinar*

0 Likes

**[Like](https://techcommunity.microsoft.com/t5/blogs/v2/blogarticlepage.kudosbuttonv2.kudoentity:kudoentity/kudosable-gid/378280?t:ac=blog-id/EducatorDeveloperBlog/article-id/14&t:cp=kudos/contributions/tapletcontributionspage" \o "Click here to give likes to this post.)**

1 Comment

[nicazure](https://techcommunity.microsoft.com/t5/user/viewprofilepage/user-id/1410897)

[nicazure](https://techcommunity.microsoft.com/t5/user/viewprofilepage/user-id/1410897" \t "https://techcommunity.microsoft.com/t5/educator-developer-blog/deep-learning-with-microsoft-cognitive-toolkit-cntk/ba-p/_self)

Occasional Visitor

‎Nov 28 2022 10:51 AM

command = SpeechTrain  
stderr = "$Root$\$RunName$.log"  
speechTrain = [  
modelPath = "$Root$\$RunName$.cn"  
SGD = [  
reader = [  
features = [  
type = "real"  
dim = "$DataSet1\_Dim$"  
file = "$DataSet1\_Features$"  
]  
]  
]  
]

https://www.tutorialspoint.com/microsoft\_cognitive\_toolkit/microsoft\_cognitive\_toolkit\_introduction.htm