Questions from Exams

- **Your deep neural network seems to converge on different solutions with different accuracy each time you train it. What's a likely explanation?**
  Large batch sizes tend to get stuck, at random, inside "local minima" instead of the correct solution.
- **Your neural network's accuracy on its training data is increasing beyond the accuracy on test or validation data. What might be a valid thing to try to prevent this overfitting?**
  Dropout layers force the network to spread out its learning throughout the network, and can prevent overfitting resulting from learning concentrating in one spot. Early stopping would be another valid answer.
- **You're implementing a machine learning model for fraud detection, where most of your training data does not indicate fraud. The cost of a incorrectly identifying an actual fraudulent transaction is much higher than the cost of incorrectly identifying a non-fraudulent transaction. Which metric should you focus on for your model?**
  Recall is appropriate when you care most about false negatives, which in this case is incorrectly identifying fraudulent transactions as non-fraudulent.
- **What is the simplest way to manage automating the archiving or deletion of old data in your S3 data lake?**
  - Use S3 Lifecycle Rules
- **A Kinesis Data Stream's capacity is provisioned by *shards*. What is the maximum throughput of a single shard?** 1MB/s or 1000 messages/s
- **Which Amazon service is appropriate for connecting video data from cameras to backend systems to analyze that data in real time?**
  Kinesis video streams connect video producers, including a variety of cameras, to the AWS cloud for further processing with services such as Rekognition, SageMaker, or EC2.
- **What is the underlying platform for Glue ETL?**
  A serverless Apache Spark platform
- **Which AWS data store provides a highly scalable data warehouse (for OLAP) that can query your S3 data lake directly?**
  When using Redshift Spectrum, Redshift can query S3 data directly - in addition to many other data sources.
- **Which kind of graph is best suited for visualizing outliers in your training data?**
  - Box & whisker, or just "box plots", organize your data into quartiles, and display outliers in the outer quartiles.
- **What sort of data distribution would be relevant to flipping heads or tails on a coin?**
  - Binomial distributionare used for binary classifications of discrete events, such as flipping a coin.
- **What is a serverless, fully-managed solution for querying unstructured data in S3?**
  - Athena, when paired with a Glue crawler, can issue SQL queries directly on S3 data in a fully-managed setting.
- **Which imputation technique for missing data would produce the best results?**
  - **MICE** Multiple Imputation by Chained Equations finds relationships between features and is one of the most advanced imputation methods available. Using machine learning techniques such as KNN and deep learning are also good approaches.
- **You have a dump of social media posts related to your company, and which to classify them based on sentiment. Which service could perform this task?** Sentiment analysis is just one of the many NLP results AWS Comprehend can give you on text input.
- **Even though you are constantly feeding it new data, you're finding that your recommendations from Amazon Personalize are becoming less relevant over time. How might you address the issue?**A full retrain (passing trainingMode=full) is recommended at least once a week.

  This was discussed in Lecture 116: **Amazon Personalize**
- **You are developing a computer vision system that can classify every pixel in an image based on its image type, such as people, buildings, roadways, signs, and vehicles. Which SageMaker algorithm would provide you with the best starting point for this problem?** The **SageMaker Semantic Segmentation** algorithm is specifically designed for pixel-level classification tasks and is well-suited for identifying and categorizing different objects in an image, such as people, buildings, roadways, signs, and vehicles.

- **You wish to use a SageMaker notebook within a VPC. SageMaker notebook instances are Internet-enabled, creating a potential security hole in your VPC. How would you use SageMaker within a VPC without opening up Internet access?**
  1.Create a VPC with No Internet Gateway
  2.Use a VPC Endpoint for SageMaker
  3.Set Up a VPC Endpoint for S3
- **A large news website needs to produce personalized recommendations for articles to its readers, by training a machine learning model on a daily basis using historical click data. The influx of this data is fairly constant, except during major elections when traffic to the site spikes considerably. Which system would provide the most cost-effective and simplest solution?Configure the Notebook Instance**

  Publish click data into Amazon S3 using **Kinesis Firehose,** and process the data nightly using **Apache Spark and MLLib** using **spot instances** in an **EMR** cluster. Publish the model's results to for producing recommendations in real-time. The use of spot instances in response to anticipated surges in usage is the most cost-effective approach for scaling up an EMR cluster. Kinesis streams is over-engineering because we do not have a real-time streaming requirement.
- **You are developing an autonomous vehicle that must classify images of street signs with extremely low latency, processing thousands of images per second. What AWS-based architecture would best meet this need? SageMaker Neo** is designed for compiling models using TensorFlow and other frameworks to edge devices such as Nvidia Jetson. The **low latency** requirement requires an edge solution, where the classification is being done within the vehicle itself and not over the air. Rekognition (which doesn't have an "edge mode," but does integrate with DeepLens) can't handle the very specific classification task of identifying different street signs and what they mean.
- **Your company wishes to monitor social media, and perform sentiment analysis on Tweets to classify them as positive or negative sentiment. You are able to obtain a data set of past Tweets about your company to use as training data for a machine learning system, but they are not classified as positive or negative. How would you build such a system?** A machine learning system needs labeled data to train itself with; there's no getting around that. Only the Ground Truth answer produces the positive or negative labels we need, by using humans to create that training data initially. Another solution would be to use natural language processing through a service such as Amazon Comprehend.
- **Your automatic hyperparameter tuning job in SageMaker is consuming more resources than you would like, and coming at a high cost. What are TWO techniques that might reduce this cost?**
  - Use logarithmic scales on your parameter ranges
  - Use less concurrency while tuning
- **After training a deep neural network over 100 epochs, it achieved high accuracy on your training data, but lower accuracy on your test data, suggesting the resulting model is overfitting. What are TWO techniques that may help resolve this problem?**

  Early stopping is a simple technique for preventing neural networks from training too far, and learning patterns in the training data that can't be generalized. Dropout regularization forces the learning to be spread out amongst the artificial neurons, further preventing overfitting. Removing layers, rather than adding them, might also help prevent an overly complex model from being created - as would using fewer features, not more.
- **You are training an XGBoost model on SageMaker with millions of rows of training data, and you wish to use Apache Spark to pre-process this data at scale. What is the simplest architecture that achieves this?** The SageMakerEstimator classes allow tight integration between Spark and SageMaker for several models including XGBoost, and offers the simplest solution. You can't deploy SageMaker to an EMR cluster, and XGBoost actually requires LibSVM or CSV input, not RecordIO.**Which is an application that emits data records as they are generated?**

  Data Producer
- **Which is an AWS service or distributed Kinesis application that retrieves data from Kinesis Data Streams?**

  Data Consumer
- **Which is a logical grouping of shards?**

  Data Stream
- **Which is a highly configurable library that puts data into an Amazon Kinesis data stream?** Amazon Kinesis Producer Library
- **Which is a pre-built Java application that collects and sends data to your Amazon Kinesis stream?**

Amazon Kinesis Agent

- **Which is a fully managed service that automatically scales to match data throughput?**

Kinesis Data Firehose

- **Which automatically provisions and scales infrastructure to read streaming media?**

Kinesis Video Streams

- **In Kinesis Data Streams, users need to manually add or remove shards to scale.**
- Kinesis data Streams can not write directly to S3. It needs a Kinesis Consumer Library app to receive the data and then write it to S3
- Kinesis Data Firehose can stream directly to S3
- Which AWS migration service allows you to copy to and from both heterogeneous or homogeneous databases?

DMS