Tell me about yourself

* Currently work as an Electrical Engineer for IBM. Background in Electrical Engineering.
* We work with a lot of data so two years ago I asked my manager if I can pursue a degree in data science and machine learning.
* Immediately after starting at my courses at Berkeley, I started applying what I’ve been learning and made up my own data science projects at work.
* My very first project was creating an algorithm to detect a defect called Fin Residue based on the data we have.
* I impressed my boss so much that he made me present what I did for the Vice President and then later the Senior Vice president.
* I think I would call myself a trail blazer in my team at IBM because all of a sudden people were migrating from SAS to Python and SQL and words like statistically significant were being thrown around, and a lot of my coworkers were asking me to help them automate some of their routine work.
* I also started holding weekly data science discussions to go over some of the things I’ve learned in class to some of my other coworkers.
* Even though I have the freedom to choose my data science projects at IBM I still have a lot of standard responsibilities. Also I felt really limited in what I can learn because at IBM I’m the one teaching other people and there’s no one to teach me on how to become a better data scientist.
* That’s why I’m now trying to pursue an official data science/ machine learning role and grow my career in this realm.

Tell me about a project you’ve done recently

* Our team tends to look at data purely as averages, yields/mean of wafer or lot.
* We want to take a more granular approach at looking into the fail locations on the wafer as well but this takes a lot more manual work.
* Once you realize the regional pattern on the wafer, you can then set up code to look at those specific regions.
* However, it’s a very manual approach to identify the special patterns of our wafers.
* I implemented a K Means Clustering method to cluster similar wafers
* Then using that as labeled data, fed it into a CNN classifier
* This project so far is more theoretical than useful.
* An idea I have is to use features extracted from the final layers of a CNN to do clustering. It might not be necessary, but it would be an interesting experiment.

What was your approach to the project?

Why did you take this approach?

Why Microsoft? What’s your interest in the role?

* I really like my team and my job at IBM but I feel like it’s limiting my growth as a Data Scientist and Machine Learning Engineer.
* I wanted to find a role that can really help me grow as a data scientist and machine learning engineer.
* I can apply to various hedge funds and banks but my passion is in Tech so I would really like to stay in tech.
* My thought process for choosing a company to be at long term is that the company has to be dominant in a field and growing.
* Microsoft is a historic company and Windows is still the dominant OS world wide so there isn’t any fear of the company failing.
* Also Microsoft Azure is one of the top cloud platforms available and it’s still growing which is why Microsoft’s stocks are steadily going up without lots of large fluctuations.

K Means Clustering

* Pick your k
* Pick k random data points as our starting clusters
* Assign all other points to the k clusters
* Calculate the mean of each cluster
* Re-assign all points
* Keep going until the clusters no longer change
* Calculate variation, then repeat process
* Choose the clustering with least variation
* Plot reduction in variation vs. K to find ideal K (elbow method)

CNN

* Pass a filter over the image
* The filter performs a convolution operation
* Then passed through a nonlinear activation function

KNN

* Uses entire data set as training set (no split for testing set)
* When an outcome is required for a new data instance, the KNN algorithm goes through the entire data set to find the k-nearest instances to the new instance, or the k number of instances most similar to the new record, and then outputs the mean of the outcomes (for a regression problem) or the mode (most frequent class) for a classification problem. The value of k is user-specified.
* The similarity between instances is calculated using measures such as Euclidean distance and Hamming distance.

Out of bag error / BAGGING?

* Bootstrap Aggregating
* Random Forest
* Bootstrap the dataset, train multiple weaker learners, aggregate the results to make prediction
* Out of bag error is to look use trees that don’t use a specific row to train, make those trees predict the result.

How to deal with imbalanced classes

* Imbalanced data set - often you model under performs, or you over guess the larger set, look at confusion matrix to see what’s happening. You can under sample the large set but …, you can over sample the small set but … Use informed oversampling and informed under sampling. Use Knn to generate samples that are similar to the your small set, then apply informed under sampling which is….

Confusion Matrix, TP v FP, TN v FN <- when is either one important

-Random Forest - what makes it different from Decision Trees?

-Gradient Boosting - why vs Random Forest?

-Logistic Regression

-L1/L2 Regularization

-Experimental Design (you want to test if a new configuration for an Azure VM improves boot up time) < - how do you create random samples for both classes?

-Power/P-Value/How many samples do you need for an experiment

-Applied data - lots of NLP / Word vectorization questions.

-Let's say you want to create an alerting system for whether or not a team should receive alert emails

-How do you initialize original labels?

-How do you create features from text in emails? other attributes? (e.g. when the alert was sent)

-Preprocessing techniques for text

-Lots of questions on DS projects I've worked on in the past

Coding problem:

Given a fasta file: e.g.

>Fragement 1

AATGCCAGTCAGTAGGCCTAAGCTAGATCCC

>Fragement 2

ATTTGGCCAGTACT[ATGGGCCAGG] <=> [ATGGGCCAGG]GGGATCAAATTGCATGC

>Fragement 3

Output how many characters each fragement overlaps the previous fragement:

E.g.

>Fragement 1 - Fragement 2

0

>Fragement 2 - Fragement 3

10

Case- max 50 chars

result - maximum substring overlap between the two strings. Begin

ATGGGCCAGGGGGATCAAATTGCATGC

ATTTGGCCAGTACTATGGGCCAGG

1. **How would you scale this problem, what if there are thousands of genes in each file and several files**
2. **What questions will you ask to tackle that problem.**
3. **I talked about distributing on several machines using like map reduce concepts, having distributed memory to store shared information. The conditions that need to be considered. Use of hashset for comparing the gene**

1. Check for ladder string and given a string return the one with highest score substring

Ladder string is a string where the letter before and after the current letter are the either letter before or after in lexicographical(alphabetical) order.

b can have a, c next to it

c can have b or d next to it

a       b       c       d      c

   +1      +1     +1    -1

A b g is not ladder string

-1 when the letter is the previous letter to the current letter in the lexicographical order

+1 when the letter is the next letter to the current letter in the lexicographical order

1. <https://leetcode.com/problems/intersection-of-two-linked-lists/>

Intersection of linked lists

1. There is a feature called Azure advisor which prompts recommendations to improve many different aspects of a Virtual Machine, optimizing cost, security performance, reliability.The advisor only pops up for users that don’t have the optimal configurations but not for the ones that already have the optimal ones.

    We want to test a new recommendation to improve the VM performance which is the startup latency how would you go about testing that. What other metric will help in accessing

1. Say two teams have built a email classifier that predicts actionable/no actionable

 then which teams classifier would you pick

If you had to develop the classifier what would you do?

1. Given a time series signal, where for each week we calculate the performance metric of a VM

    How do you know if one week's value is an anomaly.

1. Many questions about the project I worked on
2. What kind of impact
3. What kind of  technical challenges did you face in the project?
4. Why did you select one model vs other
5. Why did you use rest vs json rpc
6. How many requests was your system managing
7. In your model why did you pick a specific sample size
8. On average how many historical EF values were there for a patient. In the EF model I explained.
9. Any kind of design decisions or choosing, what was the reason.
10. In case of microservices at safeguard how many servers were approximately running each service.
11. Why did you use microservices for the safeguard, what was the reason to pick microservices
12. Explain NER model how it works?

7. Why did you choose this role/company

8. How will you deal with situation where you are inundated with information and have steep learning curve

9. Situations of Collaboration and adaptability

10. How to transfer large data from database and process, the data can’t fit in memory.