1 Can you please summarize your prior/current work related to the current position?

1.1 I have worked with Linux Foundation for log data analysis, it is an end-to-end project, including Data Generation, Data Anonymization, Data Cleaning, Parsing, Method Selection and Comparison, Model Implementation, Model Optimization. Also, I have worked with Intel Corporation but mainly on Data Engineering side, more content about script implementation and optimization, server cluster maintain and distributed system optimization and development. I also have multiple project experiences mainly about predictions, classifications, artificial intelligence and root-cause analysis.

1.2 For more details, please refer to my resume

2 What is your motivation/interest for the current position?

I shall divide this into two parts: why Data Scientist and why UPenn

2.1 Why Data Scientist? Personally speaking, I like to deal with different problems in different domains with different backgrounds, facing different customers instead of facing same problem every day. And with the development of information technology, data is used in basically every domain of our life, so analyzing data is fun and also useful, which is also the very reason I chose Data Science as my major, because this is a master key, which can be used in most cases.

2.2 Why UPenn? Firstly, I am very interested in working in universities, I like the atmosphere of researching, learning and innovation. University is a place with great vitality. And, University of Pennsylvania is one of the most famous universities in the world, and more importantly, a member of Ivy League. Though I haven't be enough lucky to learn at UPenn, but according to my friends who studies there, UPenn is one of the best among the above characteristics. So I believe that working at UPenn would be a great choice.

3 What is your experience in programming/coding/scripting? Which programming language(s), development platform(s), operating system(s) you most commonly use?

3.1 I am proficient in using Python and SQL. I can also fluently use C++, Julia and Matlab. I used C# and Java as my working language years ago so a bit rusty in those two but I believe I can quickly get proficient again if needed.

3.2 I can work in Linux OS and Windows OS

3.3 I have course-project-level understanding of AWS

4 Describe two data analysis projects you've actively participated in the past

4.1 My last data analysis project is completed weeks ago, about software log data analysis, to predict potential critical failures and return the information back to software developers for better development. We use LSTM to build time-series model to complete classification between normal messages and abnormal messages. We parse natural language-based log messages into two parts: log event type (containing static parts, like "xxx starts") and parameter

vectors (containing dynamic parts, like IP address, task names, numbers, etc.), which are also the features we use to train the model. Labels are generated with domain knowledge. We can acquire a more than 70% accuracy on a randomly generated log.

4.2 I shall mention another project, which is doing medical information classification. The data is UCI diabetes records and to predict readmission. I didn't mention it in my resume for this project was also done weeks ago, if you are interested, I can provide code. I deal with data cleaning (fillna), feature encoding and feature selection, dimension reduction using both PCA and LDA. And I use multiple classification methods, including multinomial Linear Regression, Decision Tree, and two Integrated Algorithms: Random Forest and AdaBoost. For this three-classification problem, my methods can reach at most 68% accuracy.

5 Do you have a publicly accessible repository of your projects/code (e.g. GitHub account)? If not, could you send example code?

5.1 My Github: https://github.com/Patronus-Yichen-Li

6 Briefly describe your knowledge and experience in statistical data analysis, data visualization techniques and machine learning.

6.1 Statistical Data Analysis: Linear Regression, Prediction, Correlation Analysis, Inferential Statistics, Estimation, Hypothesis Testing

6.2 Data Visualization: I have used Tableau for a while, but mostly just use plotting tools and packages in Matlab and Python, and also Excel.

6.3 Machine Learning: Classification (Logistic Regression, Clustering (KNN, K-Means, etc.), Naïve Bayes), Regression, Decision Tree, Integrated Algorithm (Bagging & Boosting), Neural Network (RNN and LSTM are the most familiar ones)

7 In addition to excellent technical skills this position also requires strong organizational skills. Could you briefly describe/comment on your organizational skills?

In my previous team projects, I always act as a leading role in topic selection, method discussion, time arrangement and division of labor. At the same time, I also actively participate in working on the project content to complete my own part. My strengths in organizational ability lie in my strong communication skills and good awareness and understanding of team members' abilities. I can coordinate and coordinate the work of each team member and assign the work that everyone is best at.

8 Do you have experience working with medical imaging? Please describe briefly, if you do.

No, I have little knowledge in imaging analysis and also have no experience with medical imaging. But I do have experience with medical information prediction, classification and signal processing.