Slide 4 – My name is An Nguyen, I am currently living in Grand Prairie Texas. I have a been a biologist over the last 6 years and am currently working for a Medical Device company. Am I a Data Scientist?

Slide 21 – Thanks Jodi. We visualized each of the job posting by transforming the 512 vectors into 2 features using Principal Component Analysis.

Slide 22 - The untransformed vectors are combined into one dataset without labels and passed into an unsupervised clustering algorithm (K-Means). Using the squared distance between each point and the cluster center we decided that 4 clusters is the optimal choice. The accuracy of the resulting analysis is ~50%.

Slide 23 – We also did supervised clustering with Data Scientists and Non-Data Scientist job labels. ~50% accuracy.

Slide 24 – And with each job descriptions labeled. ~50% accuracy. Note that SE is clustered in two different clusters and DE could not be found.

Slide 25 – To improve our accuracy percentage we chose to employ Neural Networks to classify each job posting. And was able to achieve a 92% accuracy in classifying a job posting as a Data Scientist or not.

Slide 26 – 80% accuracy if classifying all 6 job descriptions.

Slide 27 - We observed that there is a large overlap between Data Analyst and Statistician with Data Scientist job postings.