**One Sentence Problem Statement**

**Short: There is not a common definition of a Data Scientist**

**Long: Who and/or what is a Data Scientist is undecided - leading to unmet expectations, misfit hires, and lost time/resources for employers, employees, and applicants.**

**Jodi’s Script**

Slide 3: Good morning everyone, my name is Jodi Pafford from Aurora, Colorado. I am an innovative analyzer who loves Natural Language Processing, jigsaw puzzles, and watching it snow outside. Am I a Data Scientist?

Slide 13: To build a dataset, we used the python library, Beautiful Soup, to scrape indeed.com for job postings. Beautiful Soup allowed us to go from the HTML code within the website to a nice clean csv export.

Slide 14: Our dataset was created by searching 6 different job titles: Data Scientist, Data Analyst, Data Engineer, Database Administrator, Software Engineer, and Statistician.

Slide 15: We specifically searched those 6 titles in the top 16 locations for data scientists across the United States. We collected 8,738 unique job postings in our dataset. The most were collected from California.

Slide 16: Data Scientist job postings contained 20% of our dataset. Software Engineer was approximately One-third of our dataset and and using our 16 cities, Statistician job posting make-up only 2% of our dataset.

Slide 17: Indeed.com allows companies to post jobs in any format. This means that a job posting usually include “Equal Opportunity” Language, company information, and varying use of styles making the data cleaning process difficult.

Slide 18: So, we used BeatifulSoup to extract bulleted text beneath bold headings. We used the headings of: Education, Qualifications, Responsibilities, Requirements, and Skills. We then combined all of these into one corpus.

Slide 19: Next we were able to start our NLP pipeline. We used the Python libraries, Natural Language TookKit (NLTK) and spaCy to remove stop words and eliminate extra symbols, and line breaks.

Slide 20: In order to transform our words into something usable for analysis, we used Tensor Flow hub’s Universal Sentence Encoder to turn our corpus into 512 dimensional feature vectors that were ready for analysis.

Slide 28-29: Now that we had our analysis complete, we went back to our job postings to pull out the most common words. Here you can see that Data Scientist, Data Analyst, and Statistician all share common words like “Analysis”, “Statistics”, “Communication”, and “Team”. A Data Scientist adds tools and skills such as “Programming”, “Engineering”, and “Machine Learning”.

Slide 30: Last May, I was an Excel geek working in education who wanted to know more. Through this program I have learned these skills and more. I am proud to say that I AM a Data Scientist.