**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 and I am NOT a Data Scientist, but I am an innovative analyzer who loves jigsaw puzzles, doing homework (NOT hand gesture), and watching it snow outside.

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.

Slide 16: The make-up of our data set included 1,742 Data Scientist job postings. We had the most Software Engineer job postings and least Statistician job posting.

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 turn 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 in the job postings. Here you can see that Data Scientist, Data Analyst, and Statistician all share common words like “Research”, “Analysis”, and “Statistics”. A Data Scientist adds tools and skills such as Programming, Modeling, Engineering, and Machine Learning.

Slide 30: When considering the insights of our research, it is important to consider how it might be used negatively. Employers could use our information to profile candidates or Candidates could use this information to try to cheat an Applicant Tracking System.