**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. 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 searched indeed.com using the <some of> the top 16 locations for data scientists across the United States. We collected 8,738 unique job postings in our dataset. Of those, we had a total of 1,742 Data Scientist job Postings pulled.

Slide 16 (Edit image on this slide): California had the most Data Scientist job postings with 517.

Although Texas had the second largest number of total postings collected, New York, Washington state, and Massachusetts each had more Data Scientist job postings compared to Texas at 170.

Slide 17: Once we had all the data, we had to ensure it was all usable. Indeed.com allows companies to post job in any format. This means that a job posting often includes “Equal Opportunity” Language, company information, and varying use of styles making the use of Beautiful Soup Difficult.

Slide 18: So, we used BeatifulSoup to extract bulleted text beneath bold headings. We used the bold words: 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 use the Python libraries, NLTK and spaCy to remove stop words and eliminate extra symbols, and line breaks. We were finally ready to analyze the data.

Slide 20: NOT SURE HOW TO TIE THIS SLIDE IN WITH MINE. IT REALLY SHOULD GO AFTER THE KNN, RIGHT?? I can talk about it here, but there’d be no explanation as to why I did a Venn Diagram with just these three job titles.

Slide 30: When considering the applying the insights of our findings, it is important to consider how it might be used negatively. One of the ethical considerations of this type of work is to consider that employers may use it to profile candidates. (Need one more very short sentence here)

Slide 31: Additionally, this type of analysis could lead to a person cheating the automated Applicant Tracking System by falsifying information. We know both these things happen in real life, but we feel that this project may contribute to that line of thinking.