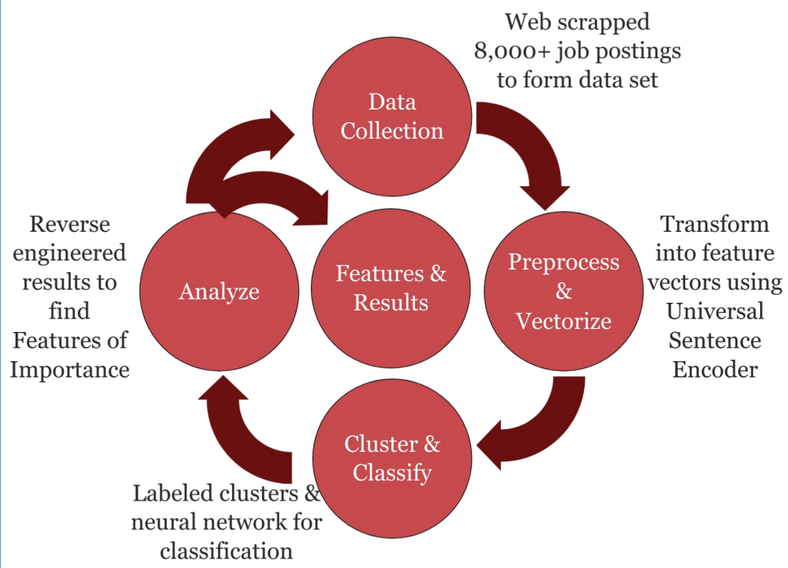
Introduction

The definition of a Data Scientist is undecided - leading to unmet expectations, misfit hires, and lost time/resources for employers, employees, and applicants. Using state-of-art NLP and Machine Learning on 8,000+ job postings, we arrive at the features that define the 'what is' and cluster other job titles to understand the 'who is' a Data Scientist.



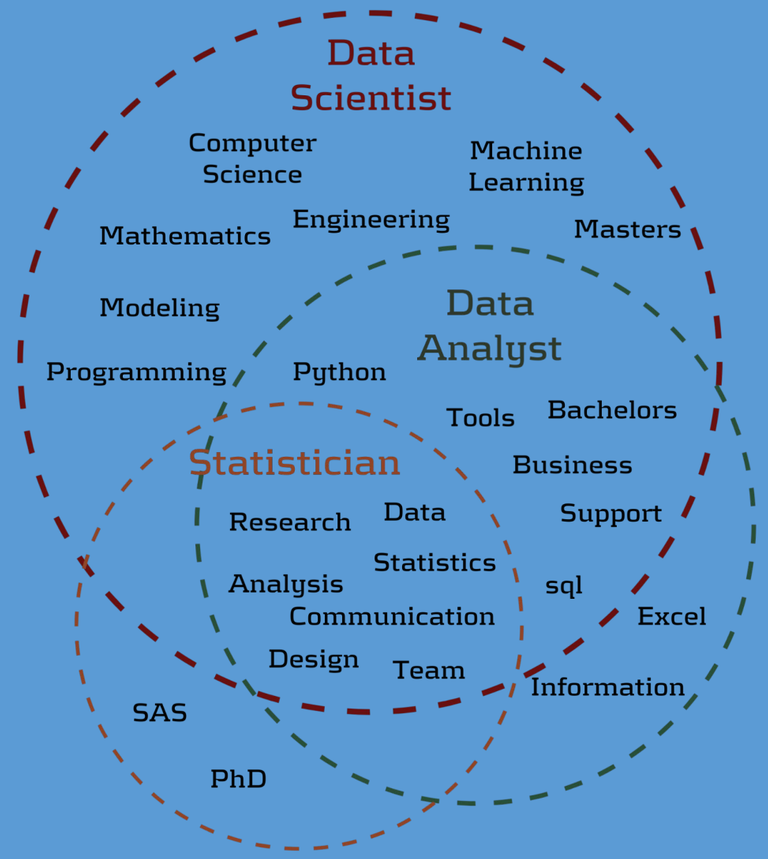
Universal Sentence Encoder for Semantic Retrieval

A pre-trained text module in Tensorflow Hub, USE is a versatile sentence embedding model that encodes a corpus into 512-dimension vector - trained with a deep averaging network (DAN) encoder and transformer encoder. Once data is converted into a vector, we can quantify the similarity/differences for all job postings. More details can be found https://tfhub.dev.

Southern Methodist University
Dallas, TX 75375, USA

A Data Science Approach to Define the Data Scientist

Most Frequent Words in Job Postings

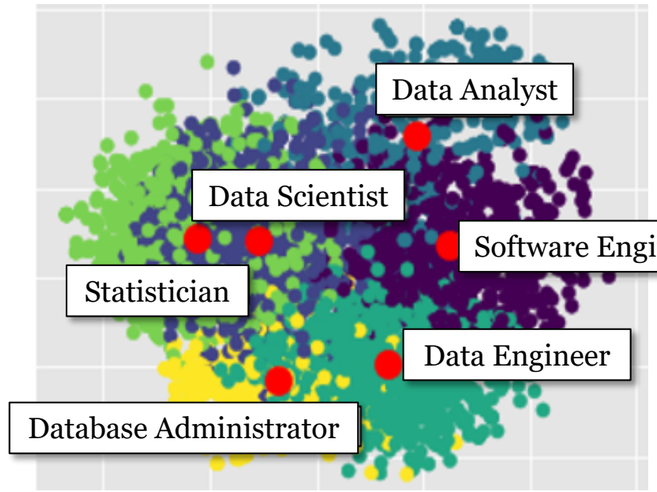


A Data Scientist codes, communicates, and collaborates – transforming data into insights using statistical, analytical, and machine learning techniques.

Andy Ho An Nguyen Jodi Pafford
Dr. Robert Slater

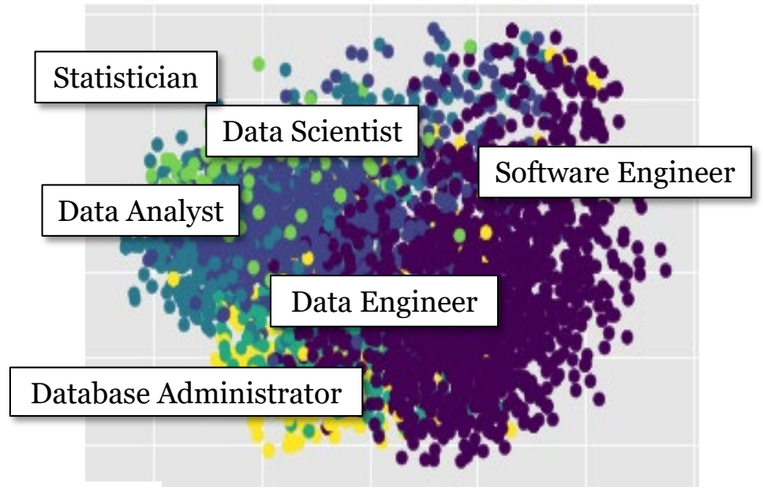
Master of Science in Data Science

k-means Clustering
54% accuracy in classification



A pair of clustering analysis was performed with PCA on the 512-dimensional vectors to visualize the similarities between the six job titles. Neural Net clustering (NNC) outperforms k-means by 31%. The NNC result provides guidance on the frequent words Venn diagram comparing Data Scientist to Data Analyst and Statistician.

Neural Net Clustering
85% accuracy in classification



Interact with our data by visiting our Tableau Website

SMU DataScience@SMU