**Title: A Data Science Approach to Defining a Data Scientist**

**Abstract**

In this paper, we present a novel approach for the problem of not having a common definition and list of skillsets for a ‘Data Scientist.’ As it relates to the field of data science, adoption has spanned nearly all industries and disciplines. This has increased the necessity for a wide range of technical skills, soft skills, experiences, and education to work on applications within data science (i.e. Artificial Intelligence, Machine Learning, Big Data Analytics, Data Visualization, etc.). The result is an overlap and ambiguity of various roles such as data scientist, data engineer, data architect, data analyst, research scientist, and research analyst. To solve the problem, we collected numerous job postings from multiple sources that share similar attributes and job descriptions to form our data set. Natural Language Processing (NLP) and advanced Machine Learning (ML) was performed to identify and visualize common patterns of job descriptions/summaries, experiences, skills, educational preferences to arrive at our key findings on the definition and list of skillsets of a data scientist. A secondary finding verified various industries (and even within the same industry) experience this same problem. Our conclusion is a Data Scientist is an individual who discovers valuable insights from large amounts of data – possessing programming skills, well versed in statistics and mathematics, and can adequately visualize and communicate the findings.

**Introduction**