A Mixed-Method Approach to Investigating Difficulty in Data Science Education
Sydney Shearer1^*, Ellie Strauss2^*, Ethan Hawk3^, Sasha Lioutikova4^, Marius Orehovschi5^, Frankie Vazquez3^*

Overview
The purpose of this study was to define a methodology to identify disconnect between students and instructors in data science classrooms through analyzing qualitative data. A combined qualitative and quantitative approach was used for analysis of survey data from three institutions. As a whole, the methods used throughout this research process provide direction for researchers in interpretation and analysis of the survey data in an efficient and time-sensitive manner. Although the research was applied to data science classrooms, this method has the potential to be applied into other fields and areas of study when performed with coordination between a field expert and a data scientist.

Dataset
Survey conducted in data science classrooms about aspects of class

- **Who was surveyed?**
  - Students (304 responses)
  - Faculty/instructors (112)
  - Teaching assistant (63)

- **Where was this conducted?**
  - Brown University
  - Smith College
  - Valparaiso University

Survey conducted in data science classrooms about aspects of class

- **What was asked?**
  - Six Likert Scale Questions
  - Used for overview of dataset
  - Four Open-Ended Questions

- **Quality analysis**
  - Proportion of relevant words in responses

- **Manual Content Analysis**
  - Categories created from inductive reasoning, responses categorized

Methods
- **TF-IDF analysis**
  - Frequency of bigrams
  - Manually labeled into categories
  - **Corpora**
    - 7 separate corpora used
    - 4 from open-ended questions
  - Responses used as documents

- **Natural Language Processing**
  - Removed stop words
  - Lemmatization
  - Tokenization
  - Bags of bigrams

- **Quality analysis**
  - Proportion of relevant words in responses

- **Manual Content Analysis**
  - Categories created from inductive reasoning, responses categorized

Conclusions/Limitations
- Manual content analysis with an automated algorithmic process gives in-depth understanding to datasets
- Data scientists should collaborate with field experts to gain insight of dataset
- Without collaboration, an in-depth analysis could not be achieved
- Data was collected during COVID-19 pandemic, resulting in a smaller dataset

Acknowledgements
- Valparaiso University
- EPIC
- TRIPODS+X-EDU NSF Grant
- DMS #1839257, 1839259, 1839270
- Ruth Wertz, Valparaiso University
- Linda Clark, Brown University
- Katherine M. Kinnaird, Smith College
- Karl R. B. Schmitt, Valparaiso University
- Bjorn Sandstede, Brown University

TF-IDF Analysis of Bigrams
- **TF-IDF: “Term Frequency, Inverse-Document Frequency”**
  - Term Frequency (TF): times a term appears in a document
  - Inverse-Document Frequency (IDF): logarithm of the inverse of the number of documents in which the term appears
  - To calculate, multiply TF by IDF to obtain a weight for each term for each document

Quality Analysis
- **Questions Asked**
  - Students (n=304)
  - Faculty (n=112)
  - Teaching assistant (n=63)

- **Surprise Questions**
  - Students (n=304)
  - Faculty (n=112)
  - Teaching assistant (n=63)

- **Topics Covered**
  - Students (n=304)
  - Faculty (n=112)
  - Teaching assistant (n=63)

- **Topics Struggled**
  - Students (n=304)
  - Faculty (n=112)
  - Teaching assistant (n=63)

Top 10 TF-IDF terms for each question/role were manually categorized into subcategories

Percentage of non-stop words by respondent role for each survey question.

TF-IDF: “Term Frequency, Inverse-Document Frequency”
- **Term Frequency (TF):** times a term appears in a document
- **Inverse-Document Frequency (IDF):** logarithm of the inverse of the number of documents in which the term appears
- To calculate, multiply TF by IDF to obtain a weight for each term for each document

- **Finds which terms are important within the documents by filtering out words that do not give insight**

- **Manual Content Analysis**
  - Categories created from inductive reasoning, responses categorized

- **who**
  - Students (n=304 responses)
  - Faculty/instructors (n=112)
  - Teaching assistant (n=63)

- **Where**
  - Brown University
  - Smith College
  - Valparaiso University

- **What was asked?**
  - Six Likert Scale Questions
  - Used for overview of dataset
  - Four Open-Ended Questions

- **Quality analysis**
  - Proportion of relevant words in responses

- **Manual Content Analysis**
  - Categories created from inductive reasoning, responses categorized

- **TF-IDF analysis**
  - Frequency of bigrams
  - Manually labeled into categories
  - **Corpora**
    - 7 separate corpora used
    - 4 from open-ended questions
  - Responses used as documents

- **Natural Language Processing**
  - Removed stop words
  - Lemmatization
  - Tokenization
  - Bags of bigrams

- **Quality analysis**
  - Proportion of relevant words in responses

- **Manual Content Analysis**
  - Categories created from inductive reasoning, responses categorized