Who teaches data science concepts?

Investigations of course catalogs with mining and ML

TRIPODS+X:EDU: Investigations of Student Difficulties in Data Science Instruction

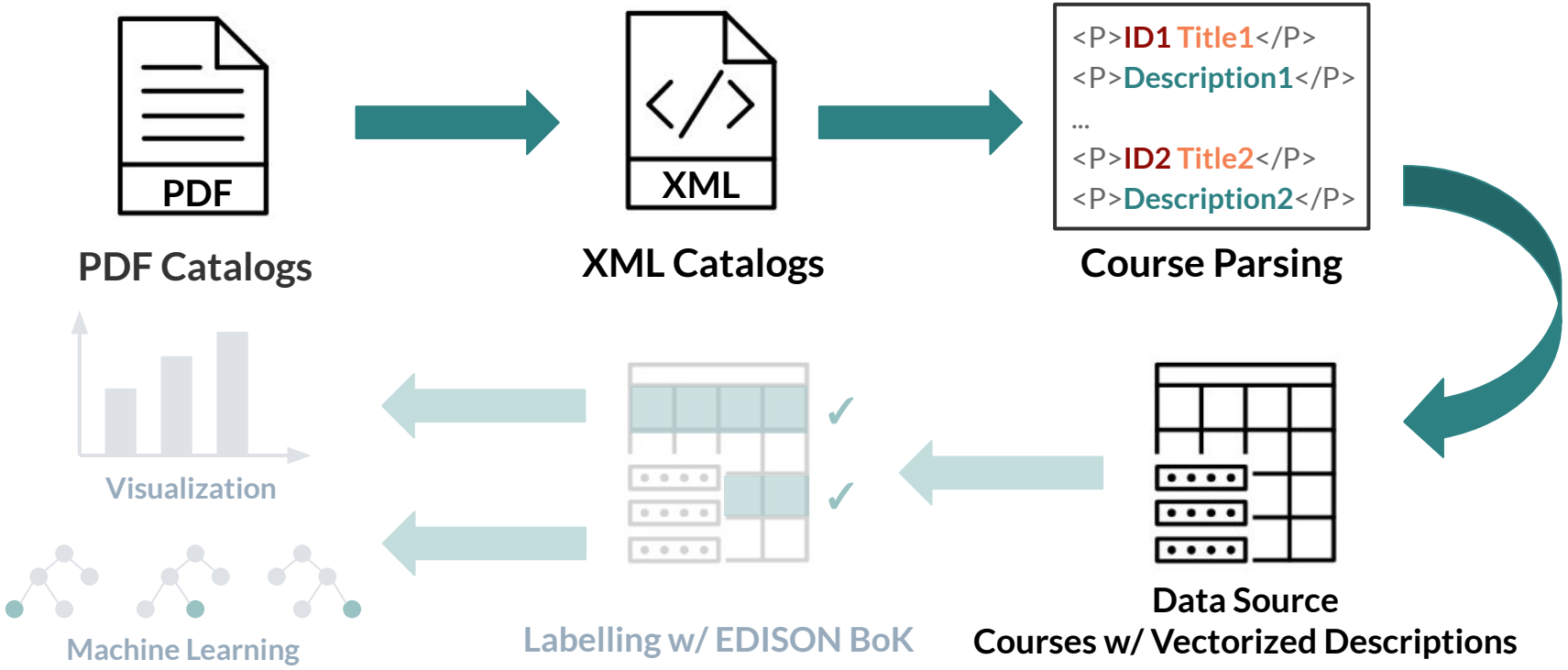
Linda E Clark, Brown University
Ethan Hawk, Valparaiso University
Katherine M Kinnaird, Smith College
Sasha Lioutikova, Yale University
Mikael Moise, Smith College
Marius Orehovschi, Colby College
Bjorn Sandstede, Brown University
*Karl R. B. Schmitt, Trinity Christian College
Sydney E Shearer, Juniata College
Ellie Strauss, Bates College
Frankie Vazquez, Valparaiso University
Ruth E.H. Wertz, Valparaiso University

Funded by the National Science Foundation
DMS #1839257, 1839259, 1839270

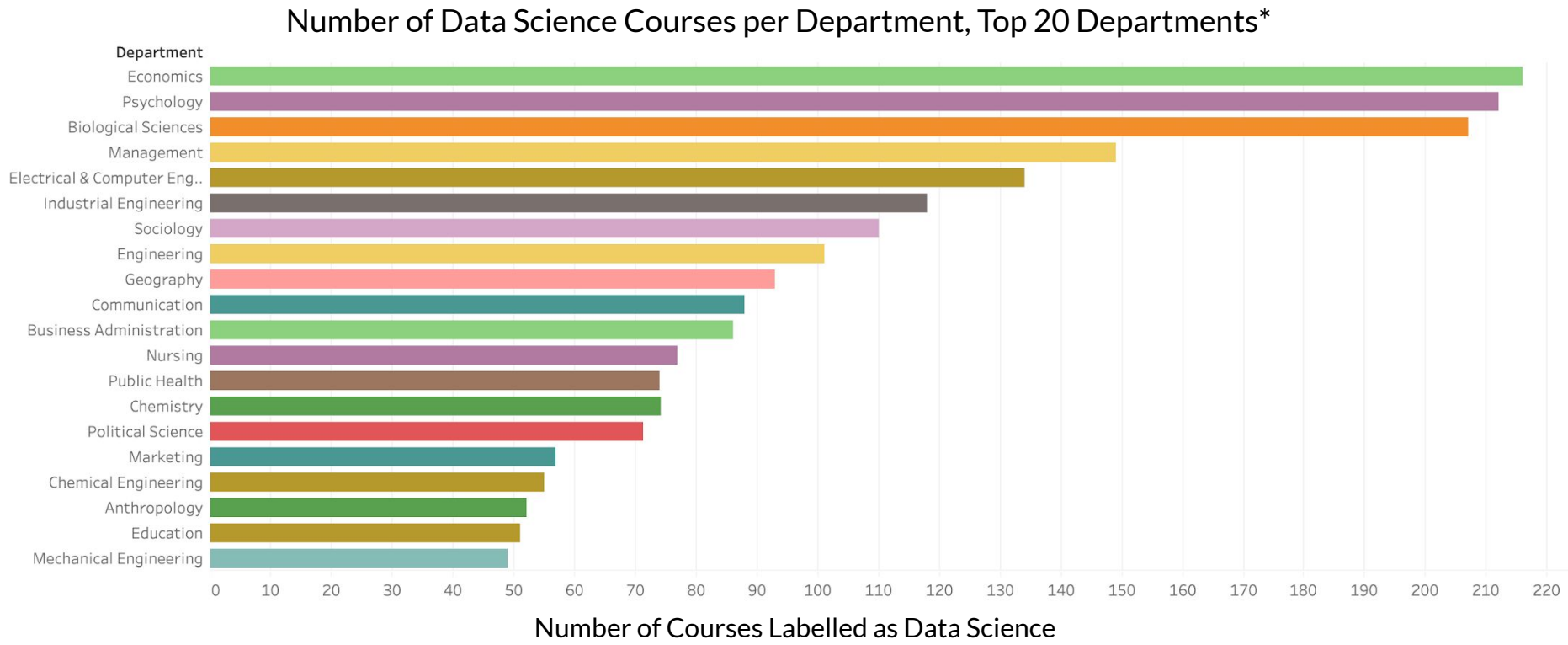
SDSS 2021

Pipeline

Long-term goal: Lay groundwork for developing a “Data Science Concept Inventory” for introductory data science courses.
Short-term goal: Identify courses (outside of data science) that develop knowledge in data science concepts



(Prototype) Visualization: Distribution of Data Science Courses



*Note: only including departments outside of the primary data science disciplines

Thank you

Questions? Email @sasha.lioutikova@yale.edu

Funded by National Science Foundation (NSF) under Grant No 1839357, 1839270, 1839259.

Who teaches data science concepts?

Investigations of course catalogs with mining and ML

TRIPODS+X:EDU: Investigations of Student Difficulties in Data Science Instruction

Linda E Clark, Brown University
Ethan Hawk, Valparaiso University
Katherine M Kinnaird, Smith College
Sasha Lioutikova, Yale University
Mikael Moise, Smith College
Marius Orehovschi, Colby College
Bjorn Sandstede, Brown University
*Karl R. B. Schmitt, Trinity Christian College
Sydney E Shearer, Juniata College
Ellie Strauss, Bates College
Frankie Vazquez, Valparaiso University
Ruth E.H. Wertz, Valparaiso University

Funded by the National Science Foundation
DMS #1839257, 1839259, 1839270

Catalog Mining Overview

MAIN GOAL: Pull course descriptions from college course catalogs to **identify classes in a variety of disciplines teaching the core topics of data science**

- Maximized the amount of courses pulled from each catalog
  - Convert PDF to XML
  - Trim XMLs
  - Clean XMLs
  - Fix spacing issues
  - Parser handles multiple cases
- Vectorization & Classification
  - EDISON Body of Knowledge
- Begin work on Machine Learning
  - Random Forest
  - Near Miss Under Sampling

XML Trimming & Cleaning

MAIN GOAL: Reduce the amount of data sent through the parser.

- Manually found start and end lines for course descriptions for each catalog
- Subsetted the catalogs using their start and end lines

MAIN GOAL: Fix broken XML structure by reintroducing tree structure to the XML, fixing broken tags, and removing invalid characters.

- Each trimmed catalog was given the same root tag
- Tags that had broken pairs were removed
- Invalid UTF-8 characters were:
  - Converted to their UTF-8 version
  - Deleted if they did not have a UTF-8 version

Reintroducing Spaces

MAIN GOAL: Fix spacing issues present in some XML files.

- Pre-semantic processing
  - Add spaces based on brackets and punctuation first
  - Use regex to identify course codes and pad them with spaces
- Semantic processing
  - Split a “word,” keep the split if it results in at least one meaningful word
- Only applied on words that are 8 characters or longer
- e.g. “Anoptionwillbeavailableforstudentsinterestedinreadingthepoemsin Arabic.” → ['An', 'option', 'will', 'be', 'available', 'for', 'students', 'interested', 'in', 'reading', 'the', 'poems', 'in', 'Arabic']

Parsing Courses

MAIN GOAL: Pull out course IDs and descriptions from the course catalog XMLs

Formats handled by our parser:

Normal	Multiple Tag Description	One-Tag Course*
<P>**ID Title**</P> <P>**Description**</P>	<P>**ID Title**</P> <P>**Desc**</P> <P>**rip**</P> <P>**tion**</P>	<P>**ID Title Description**</P>

- Formats recognized based on element text and type
- IDs recognized based on regular expression (e.g. CS 100, AMATH-5000A)

*This format is considered only for catalogs known to have this format, determined manually

Vectorization & Classification

MAIN GOAL: Assign numeric values to the descriptions and classify courses as data science courses to allow for machine learning.

- Vectorize descriptions in a matrix
- **EDISON BoK** used to **classify courses**
  - Labeled as a data science course if its description contains any topic in the BoK
- Vectorized descriptions and classes (0: non-dsci, 1: dsci) used for machine learning

data	mining	dog	data mining	metadata	is dsci
1	1	0	1	0	1
1	1	1	0	0	0
0	1	0	0	0	0
1	0	0	0	1	1

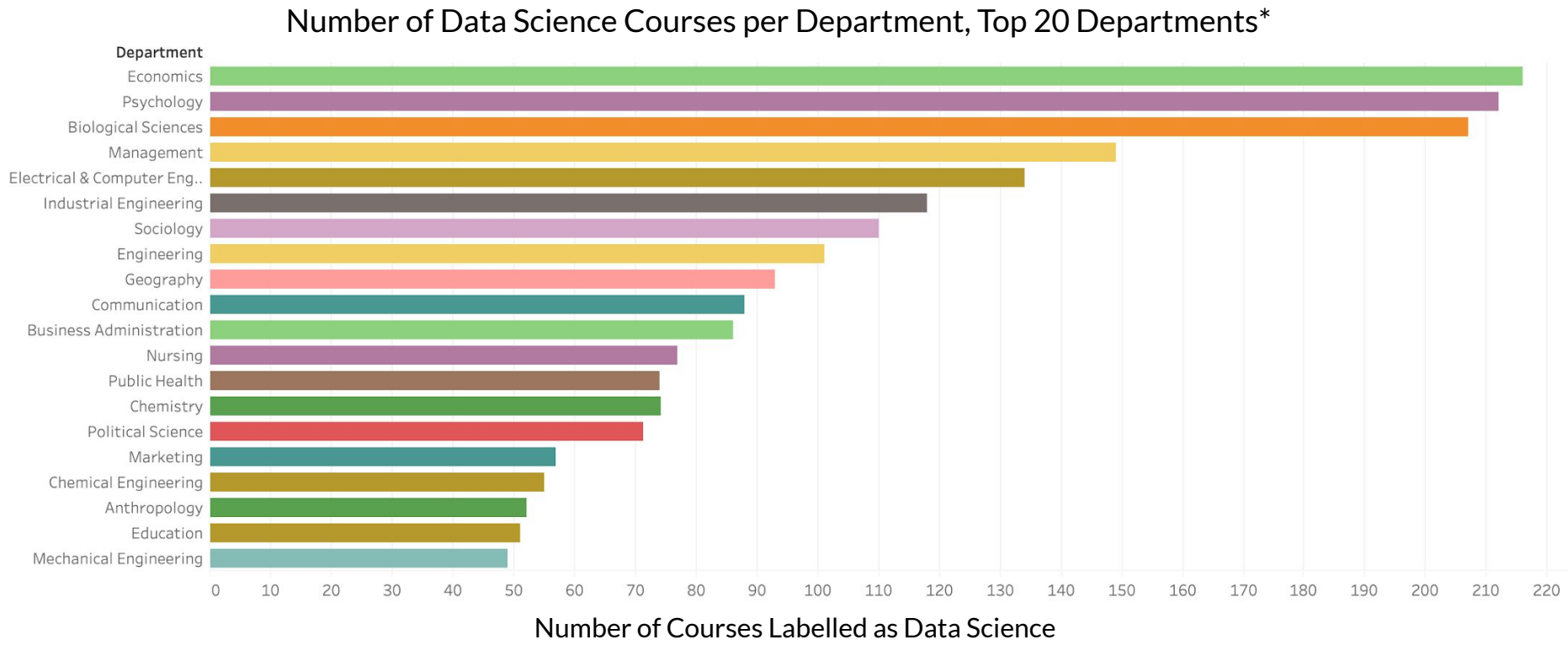
Machine Learning Pipeline

MAIN GOAL: Classify courses as data science using ML on course descriptions

Input Data: vectorized (tokenized) course descriptions

- Setting up Data:
  - Labeled as “data science” or “non-data science” based on EDISON BoK terms
  - Apply stratified k-fold (k=10) with undersampling by a Near Miss algorithm
    - Set aside one fold as Validation set
- Random Forest machine learned
  - Resulting model is trained using tuned random forest parameters
  - Model is validated on Validation set
- Predict on additional catalogs

(Prototype) Visualization: Distribution of Data Science Courses



*Note: only including departments outside of the primary data science disciplines

Machine Learning Pipeline, Revisited

MAIN GOAL: Classify courses as data science using ML on course descriptions

Input Data: vectorized (tokenized) course descriptions

- Setting up Data:
  - ***Labeled as "data science" or "non-data science" based on EDISON BoK terms***
  - Apply stratified k-fold (k=10) with undersampling by a Near Miss algorithm
    - Set aside one fold as Validation set
- Random Forest machine learned
  - Resulting model is trained using tuned random forest parameters
  - ***Model is validated on Validation set***
- Predict on additional catalogs

Future Work

- Incorporate modular testing of pipeline
- Vectorization and Classification
- Apply trained algorithm to larger, tokenized data-set

Thank you

Funded by National Science Foundation (NSF) under Grant No 1839357, 1839270, 1839259.