Use of Public Data to Generate NAICS Codes for Business Establishments.

John Cuffe (Corresponding), Sudip Bhattacharjee: U.S. Census Bureau/UConn

Ugochukwu Etudo: UConn

Dmitry Zhdanov: Georgia State

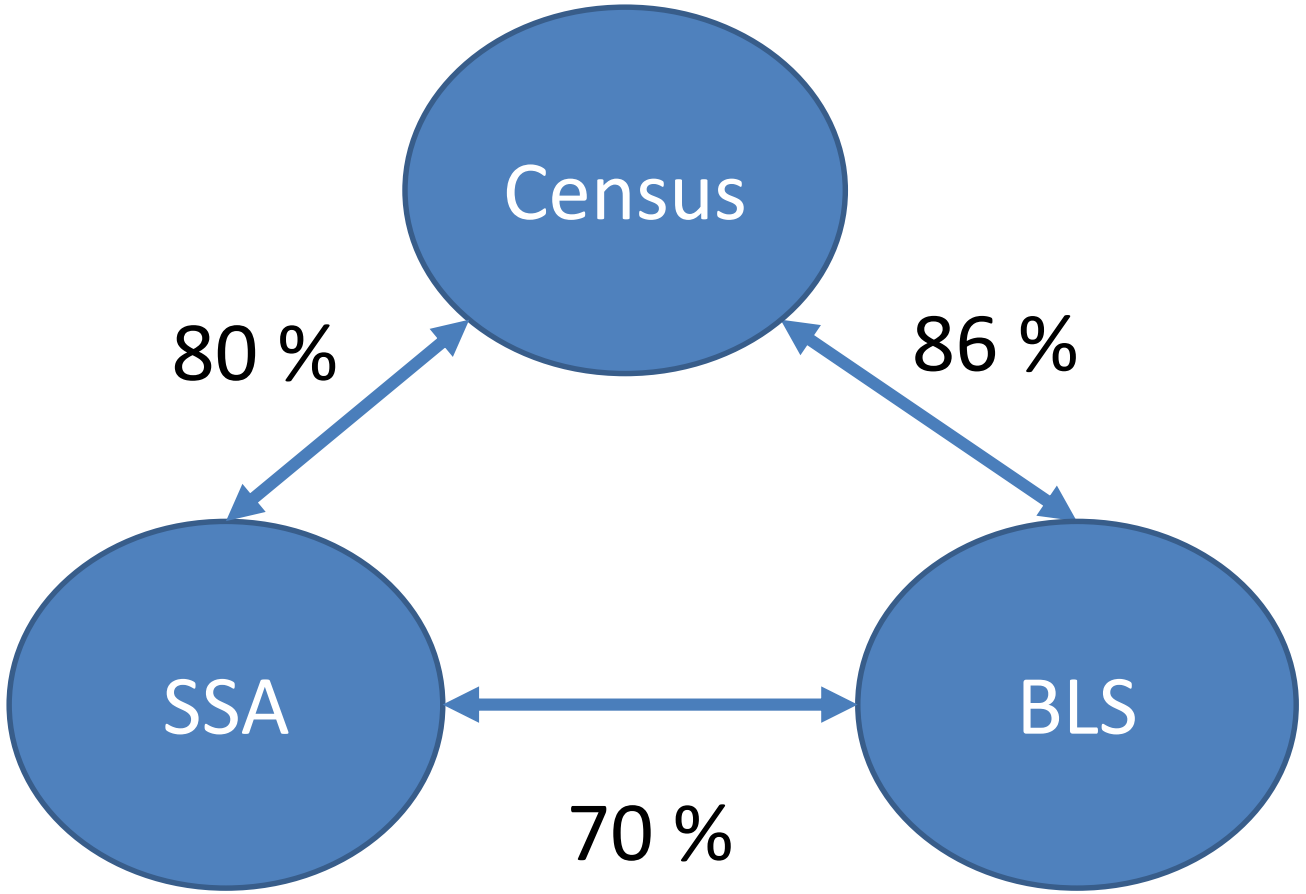
All views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. All results have been reviewed to ensure no confidential data have been disclosed.

Goals

- Our project aims to investigate new methods of data collection and analysis can improve the process for generating NAICS codes.
  - Lower costs.
  - Reduce respondent burden.
- How
  - Combine publicly-available data from websites and APIs with natural language processing, machine learning methods.

Problem

- No known ground truth.
- Current auto-coding methods…
  - Relies on almost 2-decade old EIN applications.
  - Strain on SMEs.
  - Rely on logistic regression.
- Burdensome to respondents.
- Not all sources agree.



Current Method For New Businesses: Autocoder

- Developed in early 2000s (Kornbau, 2005)
  - Uses textual primary business and merchandise sold from SS4 form.
  - Generates likelihood score for NAICS classification based on Logistic Regression.
  - Uses 2002-2004 EIN applications as training data.
  - Hand coding and QA (Census)
    - 10,000 records per quarter

Proposal

- We can combine publicly available data, Census data, and machine learning techniques to generate NAICS codes.
  - Reduce costs:
    - More time for SMEs to focus on difficult-to-code cases.
    - Enhance survey efficiency.
  - Reduce Respondent burden:
    - Respondents to not have to identify precise NAICS codes.
  - Enhance data quality:
    - Rich, textual data provides more information than only a few words.
    - Generate data faster.

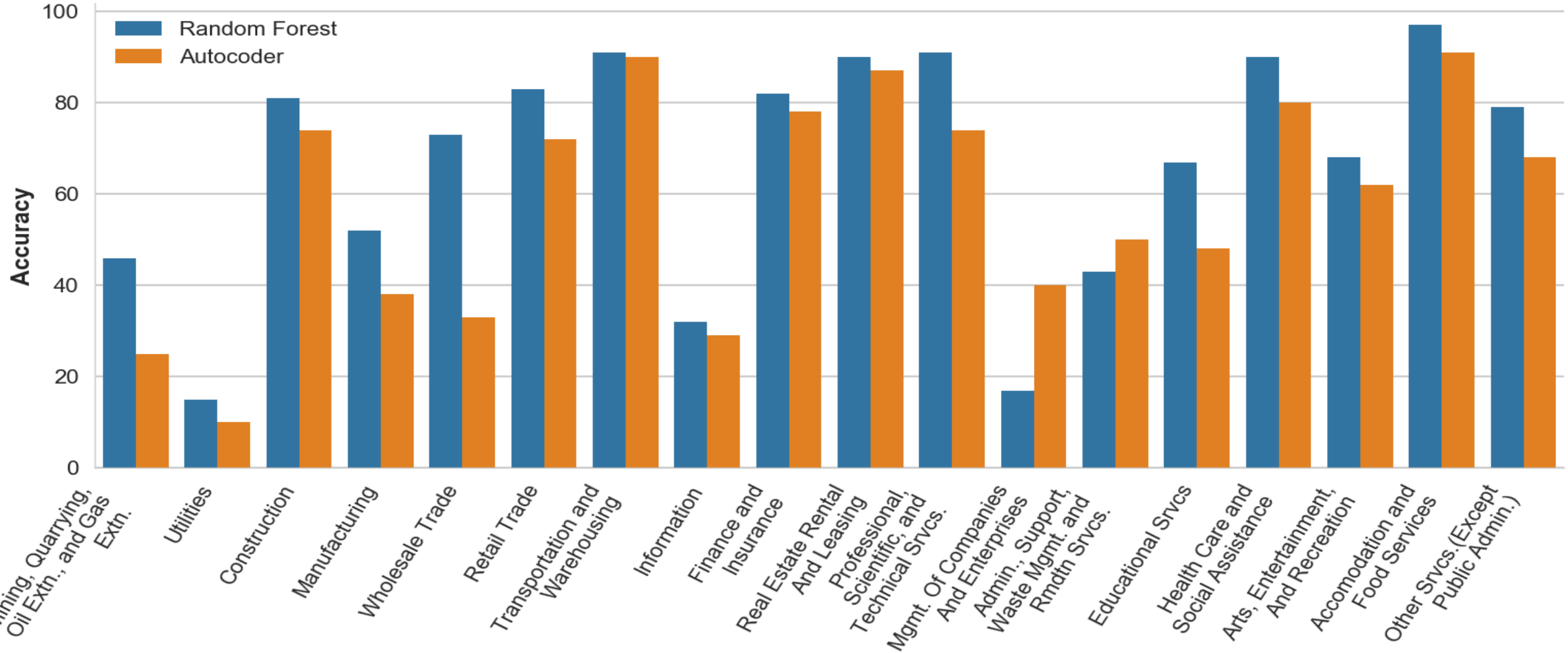
Current Data Collection and Methods

- Took 130,000 single-unit establishments from Business Register
- Found 48,000 in Google Places API
  - Searched for Types tags, website text, and user reviews.
- Fit into simple TF-IDF vectorizor, take results into Random Forest, Naive Bayes classifiers.

Common Words in Some Industries

Industry	Most Common Words in Google	Most Common Words in SS4
Information	new, 2018, business, news, contact	media, internet, production, publishing, software
Health Care And Social Assistance	doctor, care, staff, office, time	care, therapy, medical, physical, doctor
Retail Trade	store, staff, friendly, time, best	store, sales, convenience, retail, internet
Accomodation And Food Services	food, staff, friendly, time, best	restaurant, food, bar, shop, retail
Mining, Quarrying, Oil Extn., And Gas Extn.	safety, contact, oil, energy, construction	oil, gas, oilfield, drilling, energy
Finance And Insurance	insurance, financial, business, investment, help	investment, insurance, management, financial, advice

Results: Comparing to Autocoder



Research Plan

- Current results show we are at least competitive with current methods.
- We will expand upon our findings:
  - Larger sample.
    - Use salting records from previous scrape.
    - Additional scraping activities underway.
  - Match to BR using MAMBA data linkage tool (Cuffe and Goldschlag, 2018).
  - Improve predictive model.
    - Full use of doc2vec (Le and Mikolov, 2014).
    - Hierarchical modelling.

Thank You!

John Cuffe
John.cuffe@census.gov

Sudip Bhattacharjee
Sudip.bhattacharjee@census.gov

BACKUP

Methods: Overview

- Gather population of relevant businesses.
- Use Google Places API to gather textual information such as user reviews and website text on businesses.
- Natural Language Processing to create numeric representation of text.
- Machine Learning models to generate predictions at 2 digit NAICS level.

Methods: Web Scraping

- Use the Business Register records to search Google Places API.
- Gather:
  - Type tags (e.g. restaurant, auto repair).
  - User reviews.
  - Visible HTML text from website.

Example return:

{
‘website’ : ’www.KrustyBurger.com’
‘types’ : [‘restaurant’, ‘meal_delivery’, ‘meal_takeaway’],
‘vicinity’ : ‘Springfield’
‘reviews’ : [‘mmm…I don’t mind the taste’, ‘why are these called steamed hams when they are clearly fried?’, ‘Try the Clogger!’]
}

Methods: Natural Language Processing

- Natural Language Processing is a type of method designed to give mathematical meaning to human language.
- Extremely powerful as a vector reduction tool.
- Here, our goals are to find factor loadings to generate likelihood of presence in a given NAICS code.

Methods: Natural Language Processing

- NLP can be used to identify
  - Parts of speech (Noun, Verb)
  - Tokens of related words (San and Francisco probably combined to San Francisco)
  - Find the probability of words appearing in a corpus of text.
  - Identify sentiment.
- We took two approaches:
  - Word Vectorization
  - Doc2vec modelling

Approach 1: Word Vector Creation

- The matrix right shows a vector representation of a phrase.
  - Manageable with small vocabularies
  - Strips context from words
- For our data, take the 5,000 most common words in textual data, via term frequency inverse document frequency.

$$\begin{pmatrix} Do \\ Or \\ Do \\ Not \\ There \\ Is \\ No \\ Try \end{pmatrix} = \begin{pmatrix} Do & Or & Not & There & Is & No & Try \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Approach 2: doc2vec models

- Doc2vec (Le and Mikolov, 2014) are a class of models that seek to identify latent classes from sentances.
  - E.g. Related words (e.g. ‘England’ and ‘Queen’ should be more common than pairing ‘England’ and ‘Sunshine’)
- As with the word vector methods, goal is to generate a vector.
  - Here, the vectors are latent, rather than explicit.

Methods: The Population

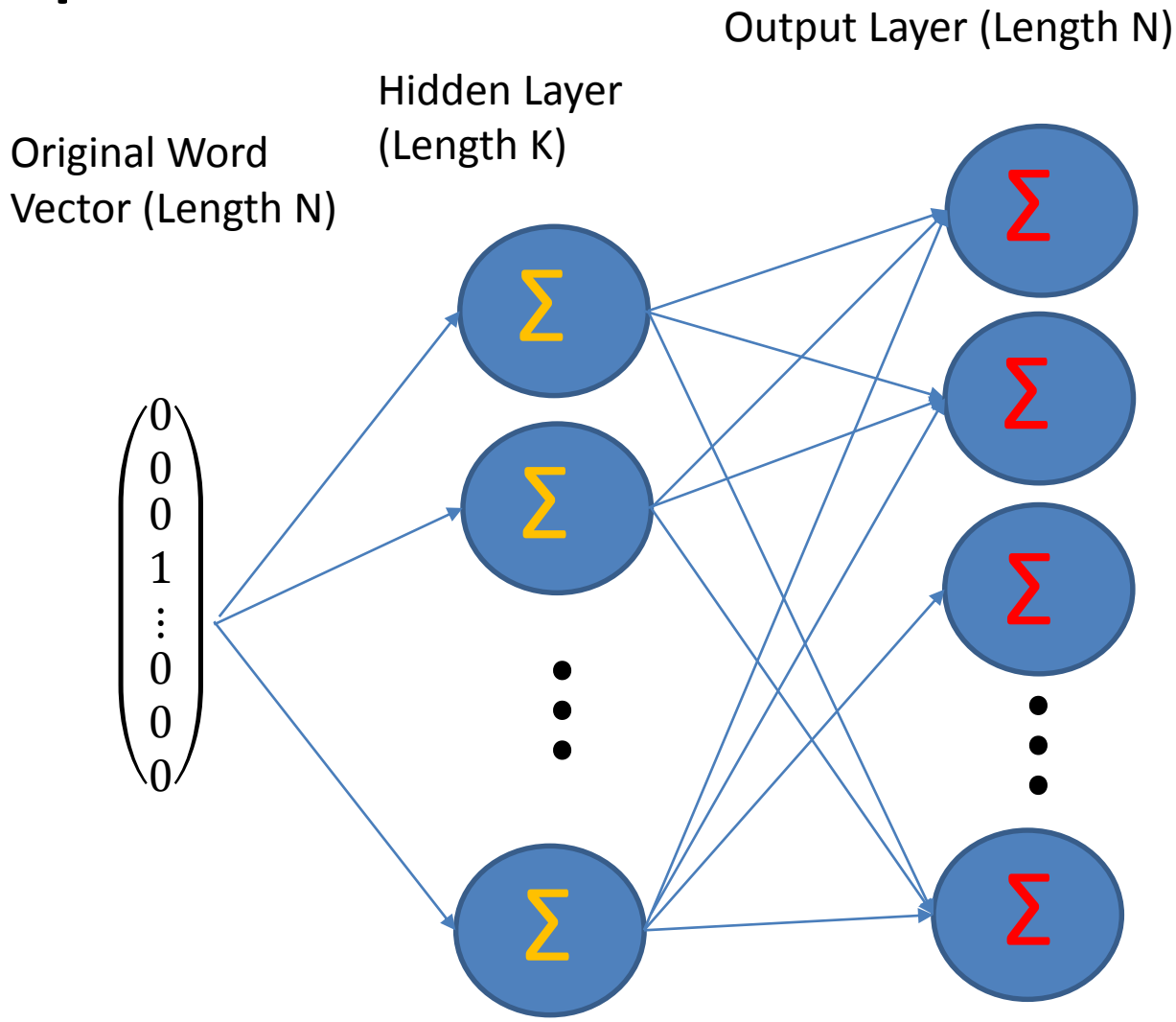
- Use 2012 Economic Census as ground truth.
- Find businesses
  - With SS4 between 2010 and 2012
  - Appear as employer business between 2012 and 2016.
  - Appear as a single establishment.

Table 1: Counts of BR Establishments, 2012-2015

Type	Count
Appear in the Business Register	21,390,000
Appear as Single Unit	15,620,000
Coded by 2012 ECON Census	4,532,000
SS4 Between 2010 and 2012	130,000

The Skip-Gram Model

- Word- and Doc2Vec models build Neural Networks that seek to identify the likelihood of an output word given the words surrounding it.



Adapted from: http://mccormickml.com/assets/word2vec

Methods: Machine Learning

- Current Autocoder uses logit
  - Not ideal—heavy assumptions, which may not hold with new data.
- Our proposal:
  - Random Forest Classifier, using vectors of weights/word counts for each of the following:
    - SS4 Business Name + Primary Merchandise Sold
    - When available:
      - Google Type Tags
      - Google User Reviews
      - Website Text

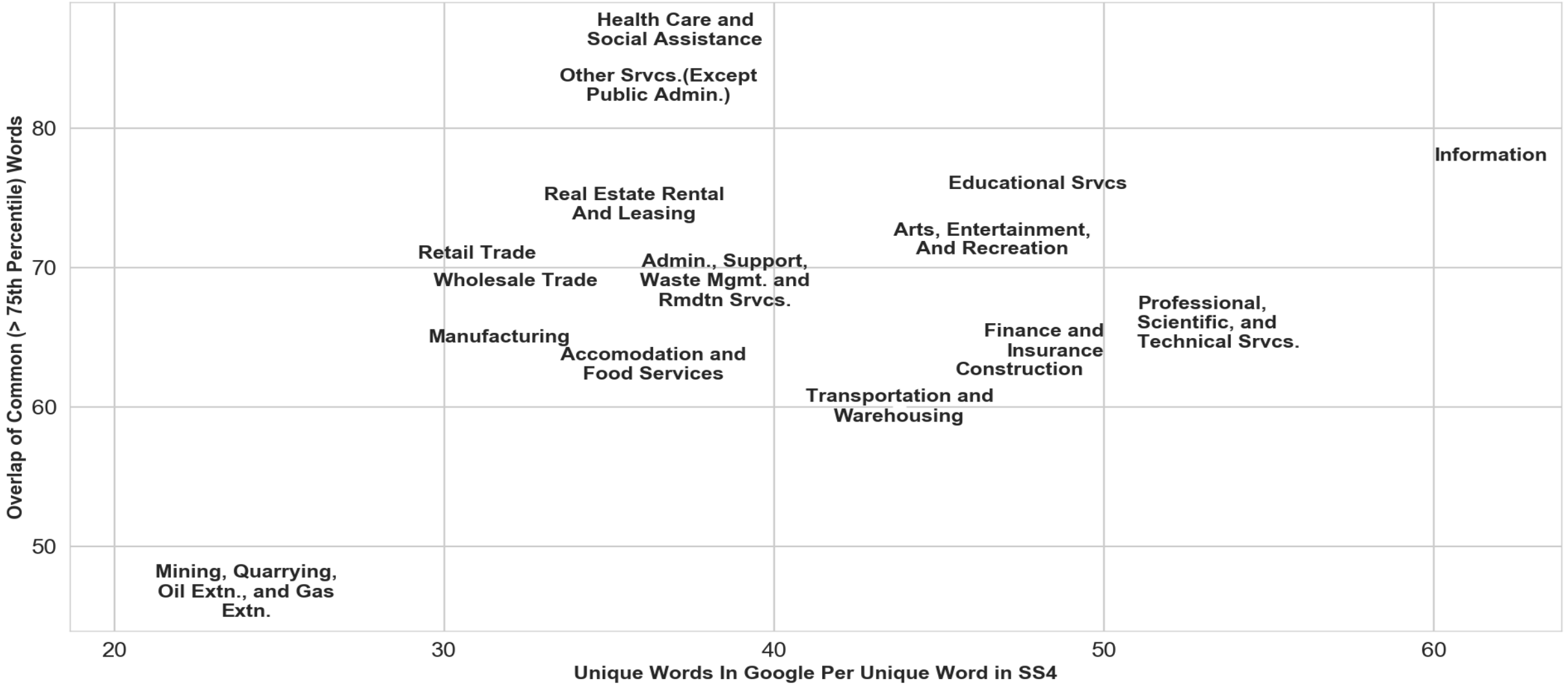
Previous Work: Strategy

- We ran different models, attempting to correctly identify 2-digit NAICS codes from the 2012 Economic Census, with the following features:
  - SS4 Business Name and Primary Merchandise Sold (SS4) (80,000)
  - SS4 + Google Types (Types) (28,000)
  - SS4 + Types + Google Reviews + Website Text (22,000)
- Take ensemble of predictions, compare to Autocoder.

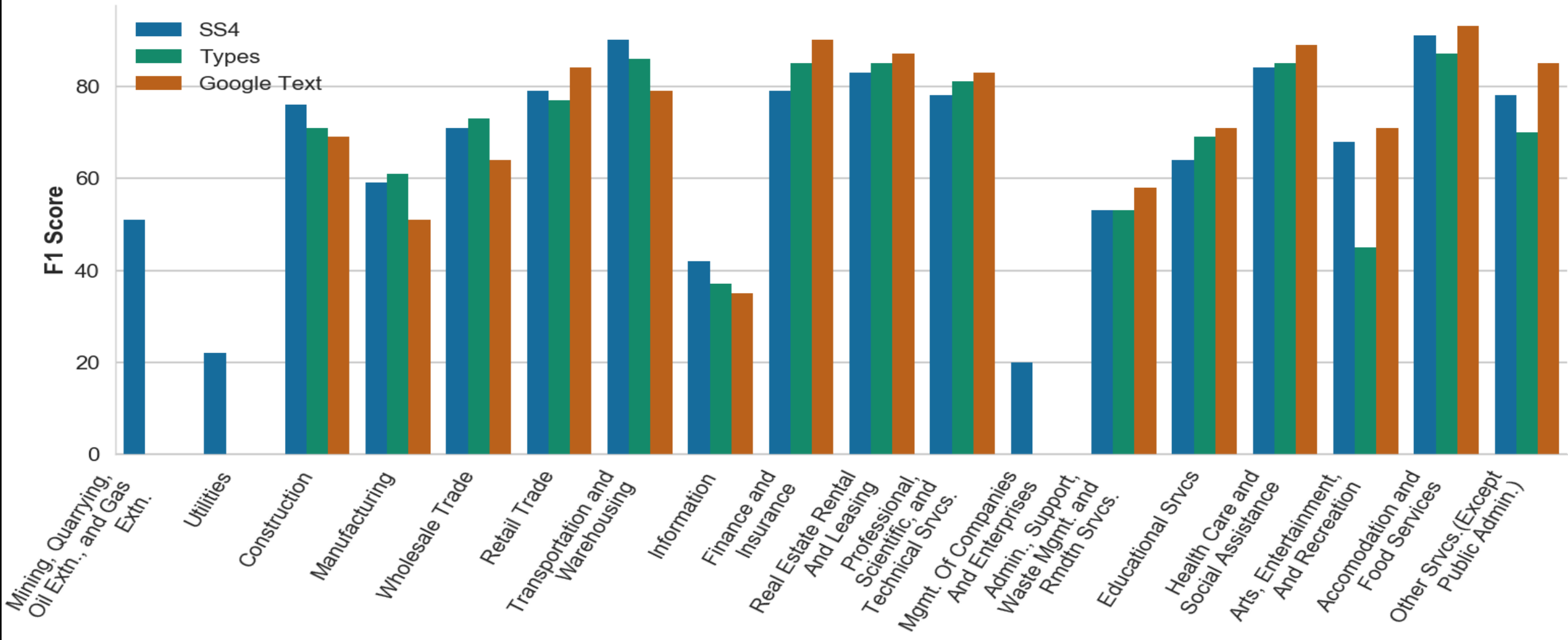
Previous Results: Google Scrape

- Most likely to appear in Google API:
  - Accommodation and Food Services (~35%), Other Services, Retail Trade, Health Care (all ~30%)
- Least likely to appear in Google API:
  - Construction
  - Mining
  - Utilities
  - Management of Companies

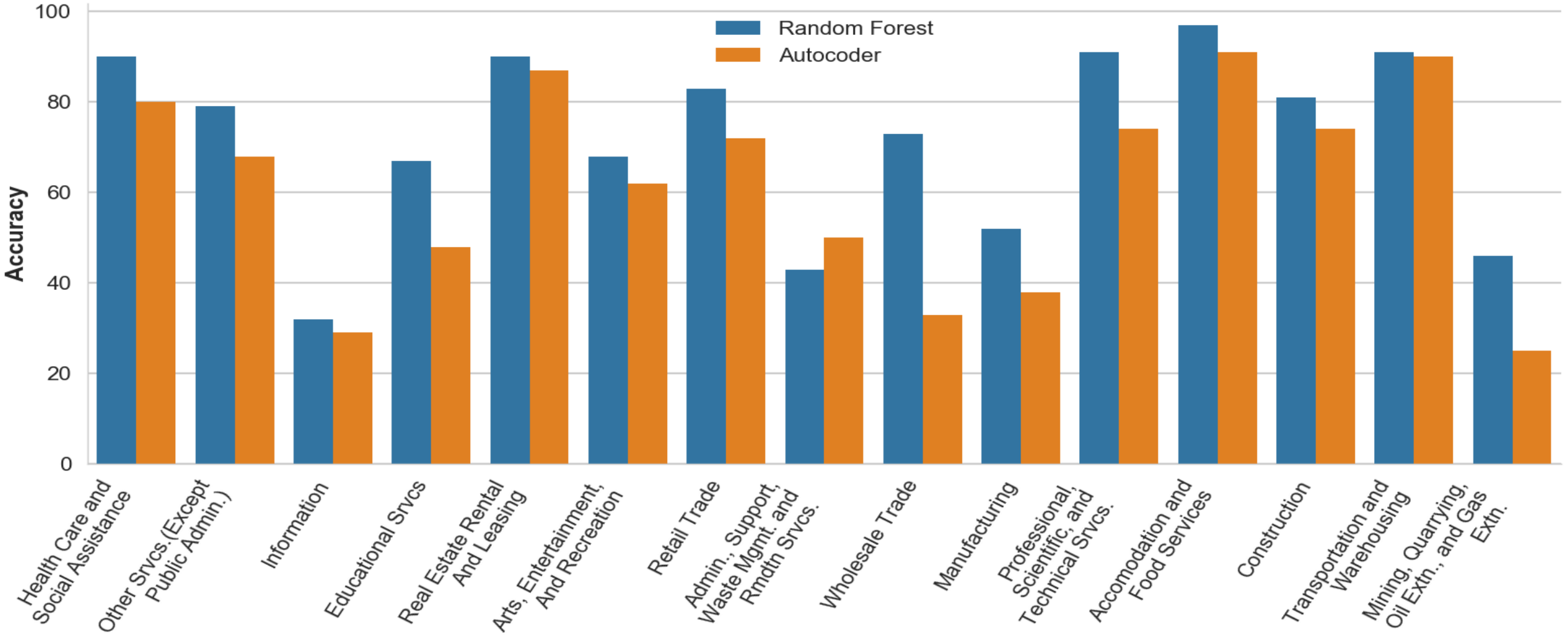
Previous Results: Data Richness



Results: Model Performance



Results: Impact of Textual Information



Weaknesses

- Survivorship bias:
  - Lower than expected match rate.
  - About 30% of all restaurants fail within a year.
- Changes over time
  - Business can change NAICS codes over time.
  - Websites more immediate source of information.
- Cleaning of Website text.
  - Better parsing and understanding of errors on NLP models.

The Future…

- Begin modelling more detailed NAICS codes.
  - Involves nested models.
- Broader target population.
  - Not just limit to 2010-12 SS4.
- Compare Google vocabulary to Autocoder.
- Test models on 2017 Economic Census.