Implementing Interactive Classification Tools in the 2022 Economic Census

Emily Wiley

Daniel Whitehead

U.S. Census Bureau

October 27, 2022

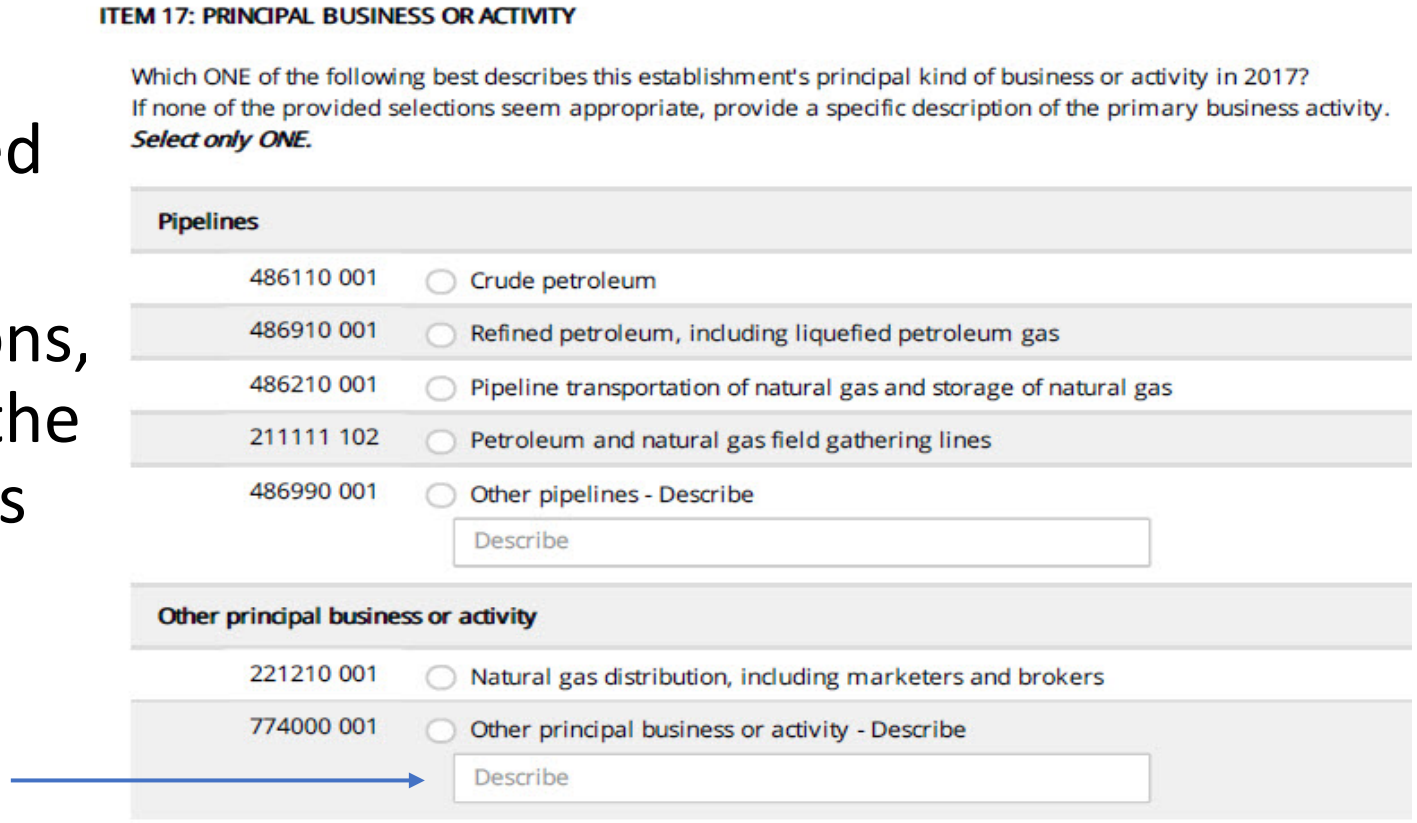
Outline

- Background and motivation
- BEACON overview
- SINCT overview
- Field Test results
- Lessons learned and conclusions

United States® Census Bureau

North American Industry Classification System (NAICS)

- Business establishments are classified by NAICS code based on primary business activity
- There are prelisted descriptions, but the respondent also has the option of writing in a business description



ITEM 17: PRINCIPAL BUSINESS OR ACTIVITY

Which ONE of the following best describes this establishment's principal kind of business or activity in 2017?
If none of the provided selections seem appropriate, provide a specific description of the primary business activity.
Select only ONE.

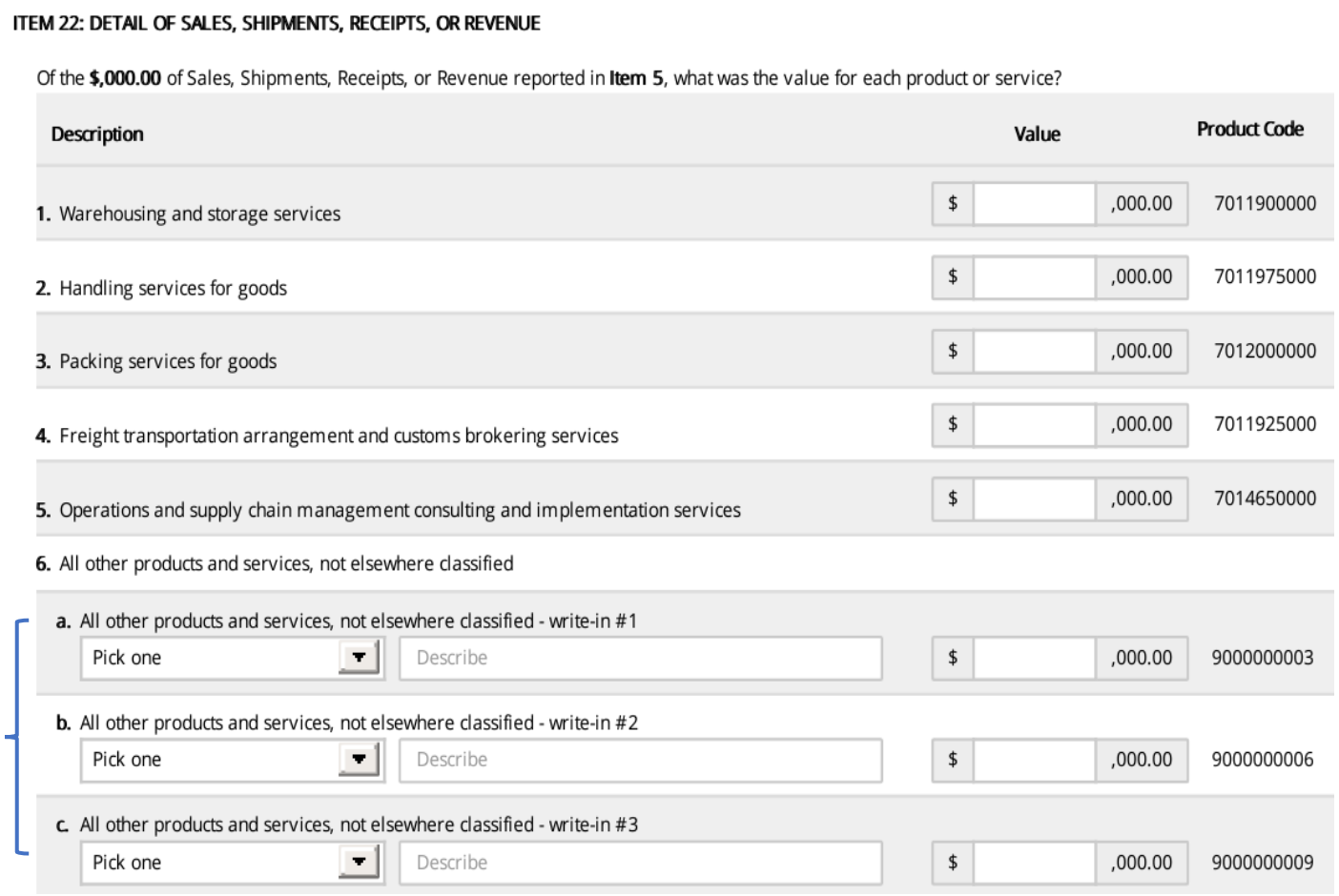
Pipelines		
486110 001	○	Crude petroleum
486910 001	○	Refined petroleum, including liquefied petroleum gas
486210 001	○	Pipeline transportation of natural gas and storage of natural gas
211111 102	○	Petroleum and natural gas field gathering lines
486990 001	○	Other pipelines - Describe
		Describe
Other principal business or activity		
221210 001	○	Natural gas distribution, including marketers and brokers
774000 001	○	Other principal business or activity - Describe
		Describe

Source: 2017 Economic Census

United States® Census Bureau

North American Product Classification System (NAPCS)

- Businesses classify their revenue into NAPCS codes based on specific products and services
- Similar to NAICS, respondents choose from a prelist or provide a write-in



ITEM 22: DETAIL OF SALES, SHIPMENTS, RECEIPTS, OR REVENUE

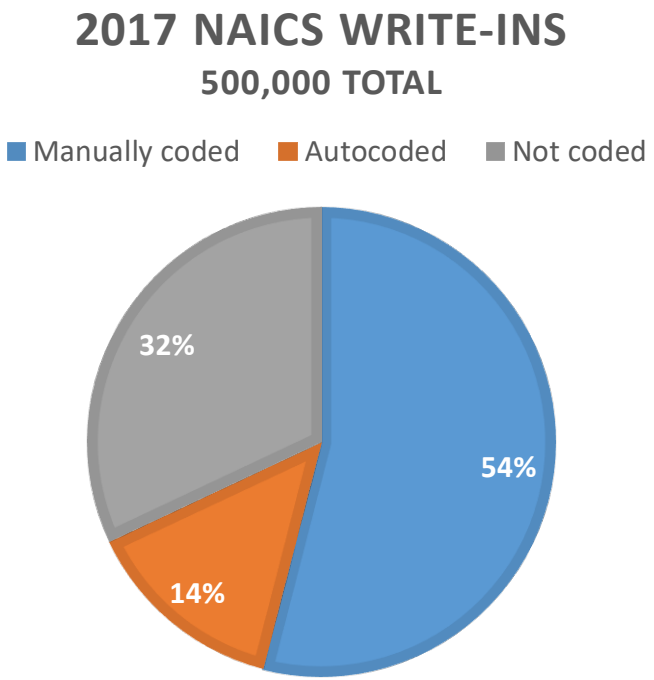
Of the **$,000.00** of Sales, Shipments, Receipts, or Revenue reported in **Item 5**, what was the value for each product or service?

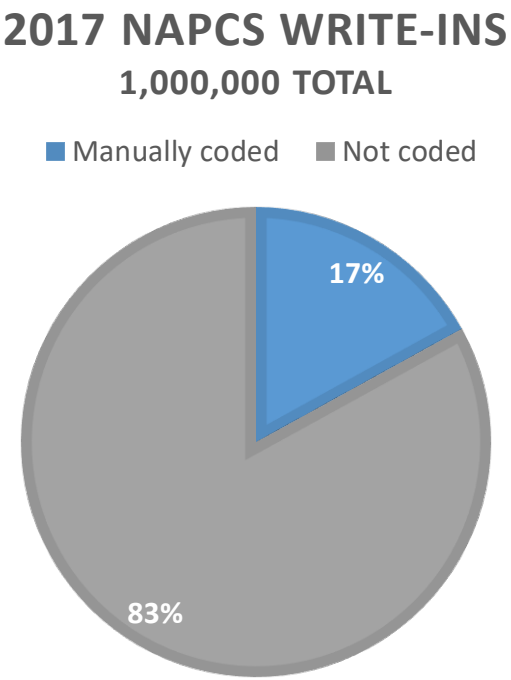
Description	Value	Product Code
1. Warehousing and storage services	$,000.00	7011900000
2. Handling services for goods	$,000.00	7011975000
3. Packing services for goods	$,000.00	7012000000
4. Freight transportation arrangement and customs brokering services	$,000.00	7011925000
5. Operations and supply chain management consulting and implementation services	$,000.00	7014650000
6. All other products and services, not elsewhere classified		
a. All other products and services, not elsewhere classified - write-in #1 Pick one Describe	$,000.00	9000000003
b. All other products and services, not elsewhere classified - write-in #2 Pick one Describe	$,000.00	9000000006
c. All other products and services, not elsewhere classified - write-in #3 Pick one Describe	$,000.00	9000000009

Source: 2017 Economic Census

United States® Census Bureau

Motivation for Machine Learning Applications





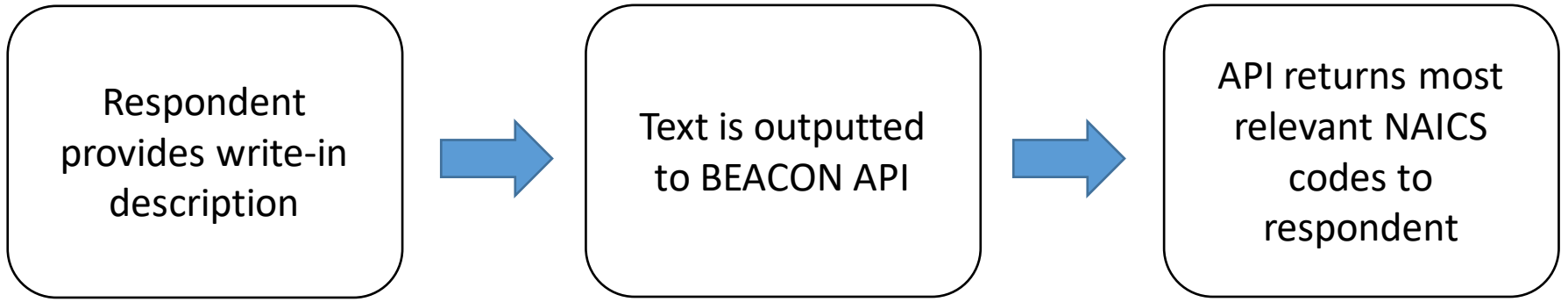
- Goal: Develop ML applications and integrate them into the electronic reporting instrument for the 2022 Economic Census

United States® Census Bureau

What is BEACON?

- Business Establishment Automated Classification of NAICS

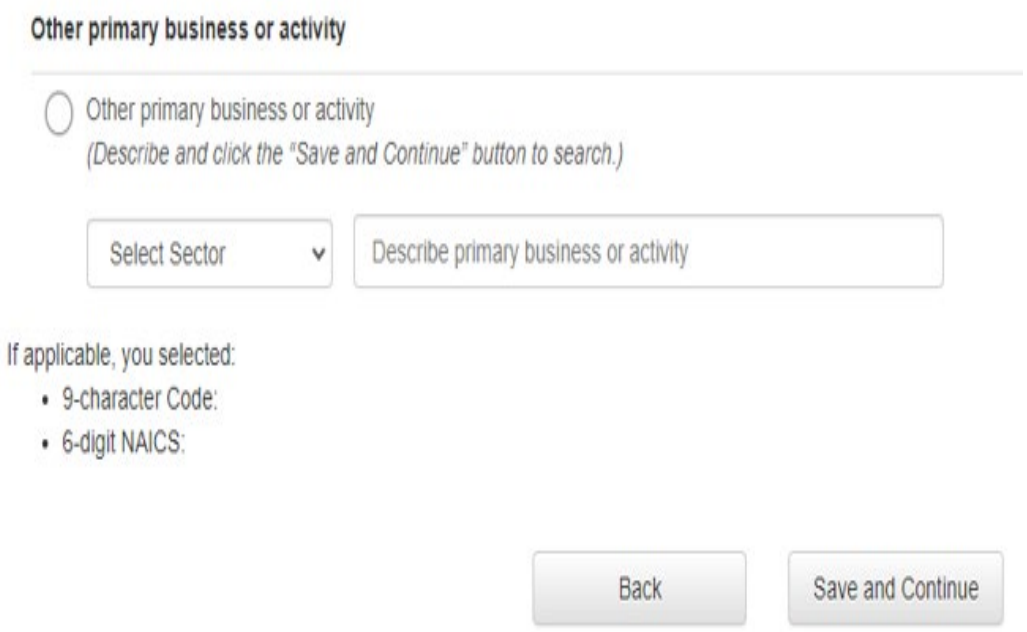
- A machine learning tool developed by the Economic Statistical Methods Division (U.S. Census Bureau) to classify NAICS for establishments based on a write-in business description



United States® Census Bureau

BEACON: Goals

- Assist respondents in self-designating their NAICS codes
- Improve accuracy of self-designated NAICS codes
- Reduce manual coding of write-ins



United States® Census Bureau

BEACON: Training Data

- Historic write-in responses to the Economic Census (EC)
- Frequent write-in text that was autocoded during 2017 EC
- Business descriptions from IRS SS-4 forms
- Classification Analytical Processing System (CAPS) items
- Harmonized System commodity description

Business Description Text	NAICS
This is a car dealership.	441110
R&D lab – medical/health	541715
we mainly repair furniture, some sales	811420

United States® Census Bureau

BEACON: Methodology

- Text cleaning
  - Remove common words and phrases (e.g., “the”, “has”, “for instance”)
  - Correct common misspellings

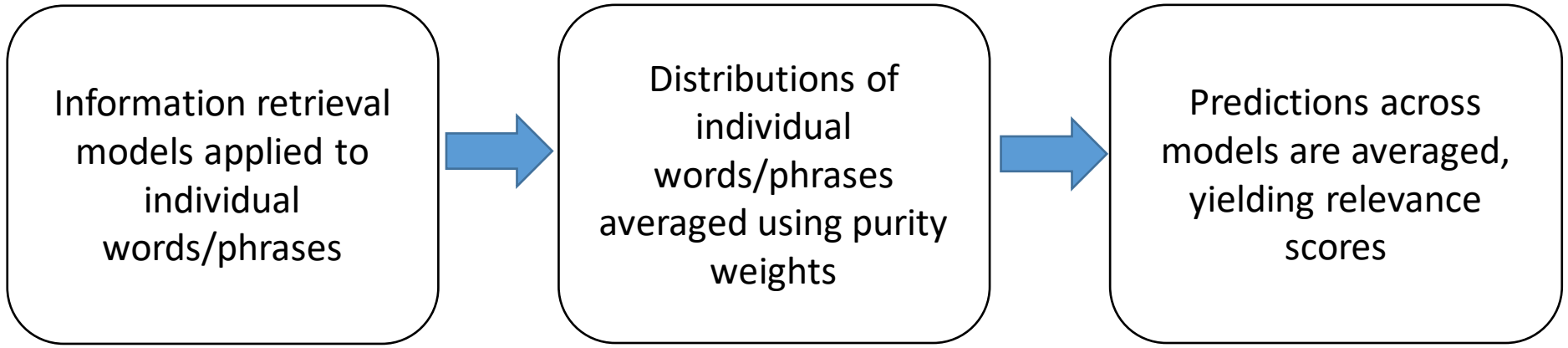
- Dictionary
  - Words and word combinations that BEACON recognizes
    - Words are cleaned, stemmed, and meet minimum frequency requirements
  - Associations between words and NAICS codes in the training data
    - “tutor” is highly associated with NAICS 611691 – Exam Preparation and Tutoring
    - “retail” occurs in many NAICS codes and is therefore less predictive

United States® Census Bureau

BEACON: Methodology

- Three separate information retrieval models

- Models are applied hierarchically
  - First: 2-digit (sector level)
  - Then: 6-digit (industry level)



United States® Census Bureau

BEACON: Methodology

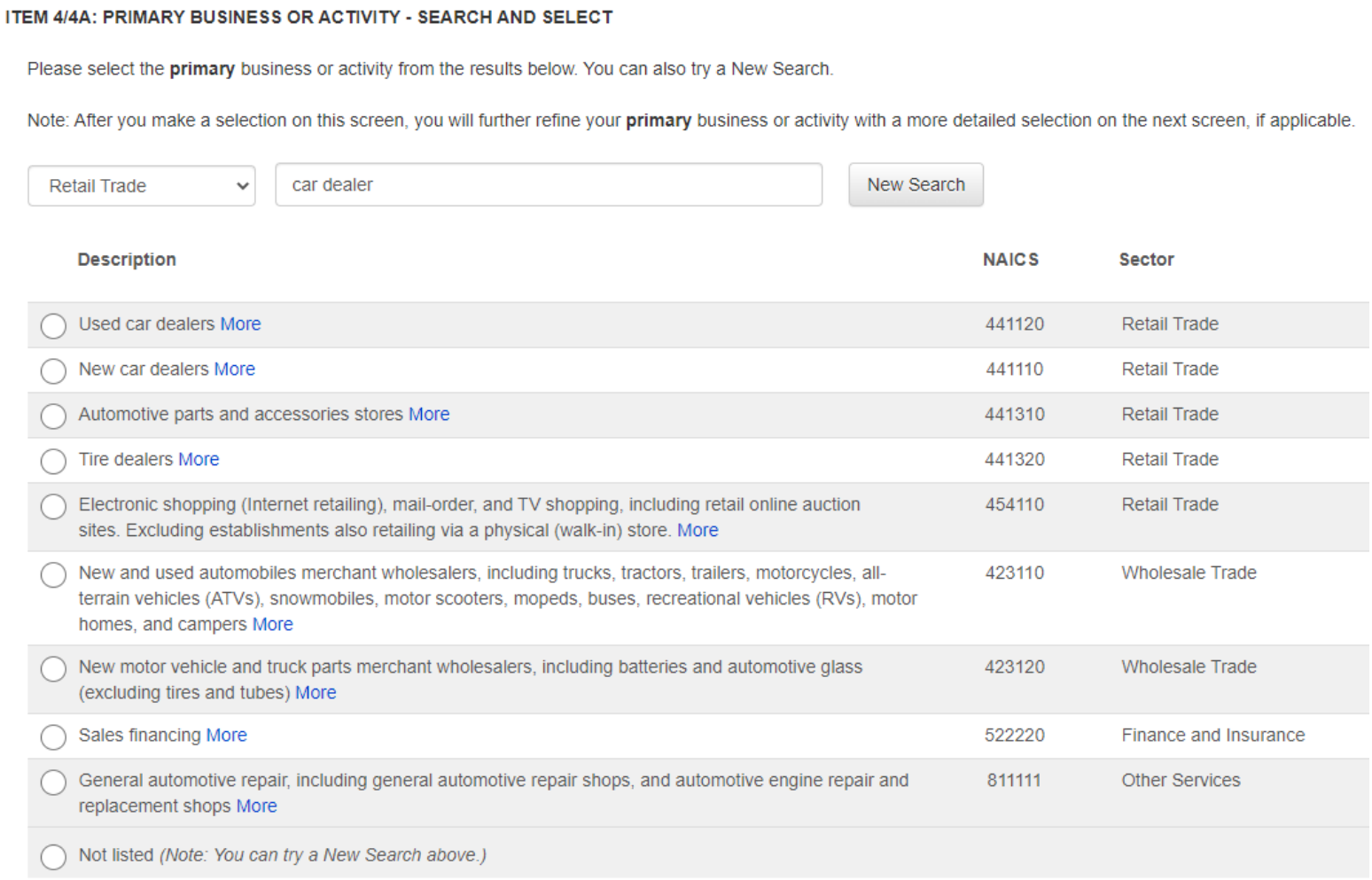
- Purity Weights
  - The NAICS distributions of the various words/combs are averaged using “purity weights” that give more weight to the NAICS distributions of words/combs that are more pure/predictive
  - The purity weight is a function of the maximum proportion

- Relevance scores
  - Range in value between 0 and 100
  - Reflect how confident BEACON is that the NAICS code is correct

United States® Census Bureau

BEACON: Results

- From these scores, BEACON returns a ranked list of NAICS codes at the 6-digit level



ITEM 4/4A: PRIMARY BUSINESS OR ACTIVITY - SEARCH AND SELECT

Please select the **primary** business or activity from the results below. You can also try a New Search.

Note: After you make a selection on this screen, you will further refine your **primary** business or activity with a more detailed selection on the next screen, if applicable.

Retail Trade | car dealer | New Search

Description	NAICS	Sector
Used car dealers More	441120	Retail Trade
New car dealers More	441110	Retail Trade
Automotive parts and accessories stores More	441310	Retail Trade
Tire dealers More	441320	Retail Trade
Electronic shopping (Internet retailing), mail-order, and TV shopping, including retail online auction sites. Excluding establishments also retailing via a physical (walk-in) store. More	454110	Retail Trade
New and used automobiles merchant wholesalers, including trucks, tractors, trailers, motorcycles, all-terrain vehicles (ATVs), snowmobiles, motor scooters, mopeds, buses, recreational vehicles (RVs), motor homes, and campers More	423110	Wholesale Trade
New motor vehicle and truck parts merchant wholesalers, including batteries and automotive glass (excluding tires and tubes) More	423120	Wholesale Trade
Sales financing More	522220	Finance and Insurance
General automotive repair, including general automotive repair shops, and automotive engine repair and replacement shops More	811111	Other Services
Not listed *(Note: You can try a New Search above.)*		

United States® Census Bureau

What is SINCT?

- Smart Instrument NAPCS Classification Tool
- Developed by Economy-Wide Statistics Division
- Two distinct versions
  - SINCT 1.0: TF-IDF
  - SINCT 2.0: Doc2Vec

Source: istockphoto.com

United States® Census Bureau

SINCT: Training Data

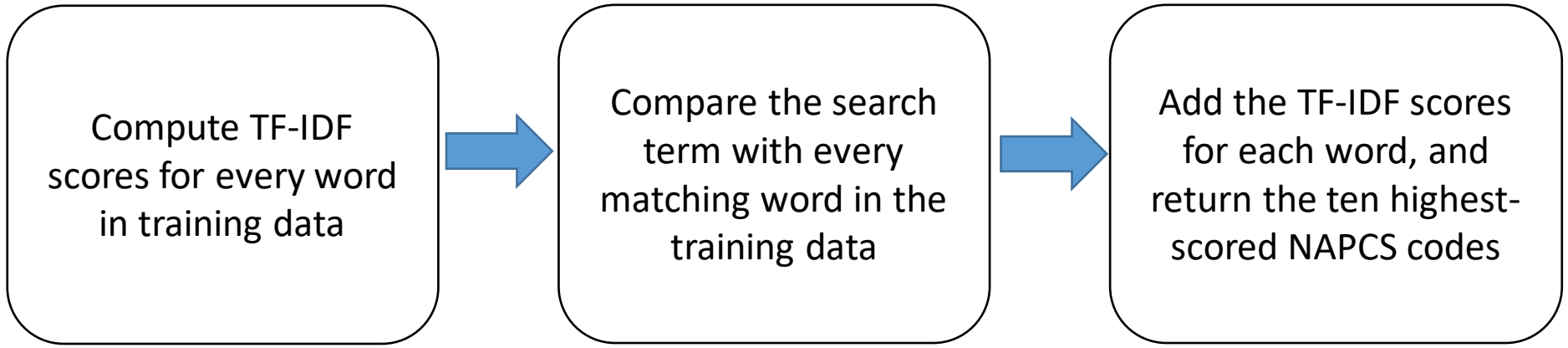
- Reclassified write-ins from the 2012 and 2017 Economic Census
- Classification Analytical Processing System (CAPS) items
- Subject matter expert examples
- NAPCS title file

NAICS	Search Term	NAPCS
423860	Airplane	4002000003
481111	Airplane	7003075000

United States® Census Bureau

SINCT 1.0: Methodology

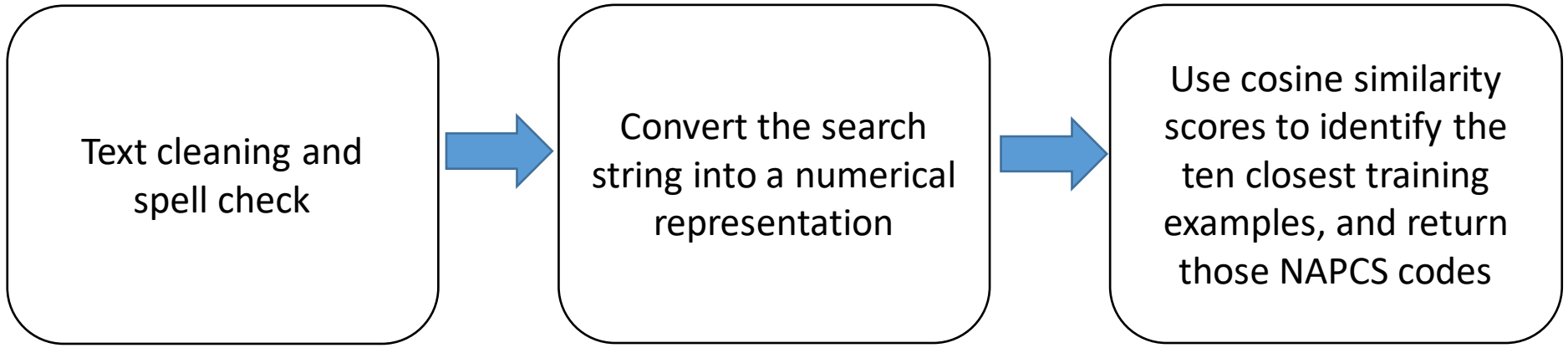
- Term frequency-inverse document frequency (TF-IDF)
  - Compute the frequency of a word or phrase within a document, and then adjust that calculation to account for common words



United States® Census Bureau

SINCT 2.0: Methodology

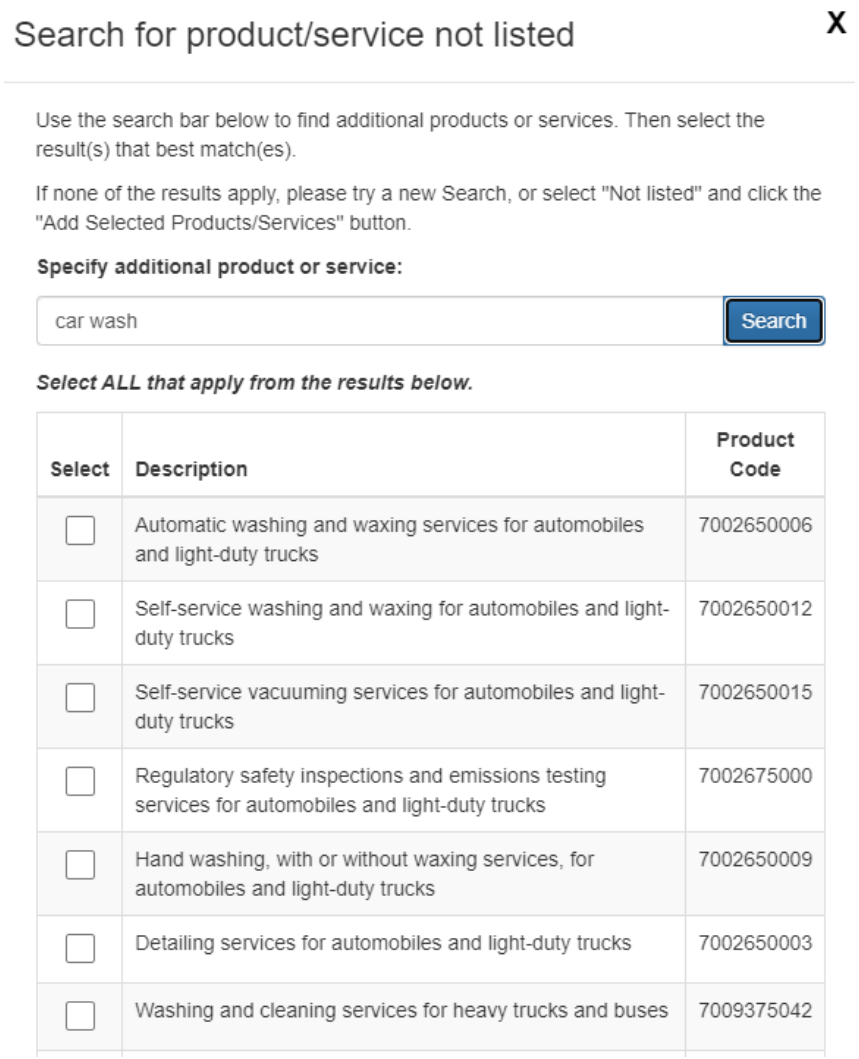
- Doc2Vec machine learning model
  - Neural network model
  - Represent words, phrases, sentences, or paragraphs as a vector



United States® Census Bureau

SINCT: Results

- SINCT returns the ten highest-scored NAPCS codes to respondents
- Respondents can select from those results, perform a new search, or leave their search term as a write-in



Search for product/service not listed

X

Use the search bar below to find additional products or services. Then select the result(s) that best match(es).

If none of the results apply, please try a new Search, or select "Not listed" and click the "Add Selected Products/Services" button.

Specify additional product or service:

car wash | Search

Select ALL that apply from the results below.

Select	Description	Product Code
☐	Automatic washing and waxing services for automobiles and light-duty trucks	7002650006
☐	Self-service washing and waxing for automobiles and light-duty trucks	7002650012
☐	Self-service vacuuming services for automobiles and light-duty trucks	7002650015
☐	Regulatory safety inspections and emissions testing services for automobiles and light-duty trucks	7002675000
☐	Hand washing, with or without waxing services, for automobiles and light-duty trucks	7002650009
☐	Detailing services for automobiles and light-duty trucks	7002650003
☐	Washing and cleaning services for heavy trucks and buses	7009375042

United States® Census Bureau

2021 Industry Classification Report Field Test

- The 2021 Economic Census Industry Classification Report (Refile) was repurposed as a field test for BEACON and SINCT

- Approximately 37,000 establishments
  - 12,000 truth deck
  - 25,000 non-truth deck

United States® Census Bureau

Field Test: Results

- BEACON
  - Returned correct NAICS code 90% of the time
  - Respondents selected it 83% of the time

- SINCT
  - Returned correct NAPCS code 74% of the time
  - Respondents only selected it 50% of the time

United States® Census Bureau

Lessons Learned

- BEACON performed as expected
  - Speed and concurrent request requirements were met
  - 90% accuracy rate

- SINCT performance was lower than expected
  - 6% timeout rate
  - 74% accuracy rate
  - Solution: overhaul the model, improving speed and accuracy

United States® Census Bureau

Lessons Learned

- Respondents often select “none of these”, even if the correct result is presented
  - Solution: update wording and instructions for main mailing

- Additional training data
  - Added 7,000 examples each to BEACON and SINCT’s training data

United States® Census Bureau

Conclusions

BEACON and SINCT should significantly reduce the number of unclassified write-ins in the 2022 Economic Census

Analysts will spend less time and will code a higher percentage of write-ins

Data quality will be improved

United States® Census Bureau

Thank You!

Emily Wiley

Emily.L.Wiley@census.gov

Daniel Whitehead

Daniel.Whitehead@census.gov

United States® Census Bureau