BEACON: An Industry Classification Tool

Eurostat Webinar

April 30, 2024

Daniel Whitehead

Sarah Pfeiff

Economic Statistical Methods Division

U.S. Census Bureau

United States® Census Bureau

Disclaimer

Any opinions and conclusions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau. The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data [Project No. P-7504847, Disclosure Review Board (DRB) approval number: CBDRB-FY23-EMSD001-011.]

United States® Census Bureau

Outline

- Background
- Motivating Problem
- BEACON Methodology
- Future Work
- Demo

United States® Census Bureau

North American Industry Classification System (NAICS)

- Establishments are physical locations where business is conducted
- U.S. Census Bureau classifies establishments by NAICS industry based on primary business activity
- Used throughout survey life cycle
  - Sample selection
  - Data collection
  - Publication
- Hierarchical 6-digit structure
  - First two digits represent sector
  - Additional non-zero digits add detail

Example NAICS Structure

NAICS	Description
51	Information
515	Broadcasting (except Internet)
5151	Radio and Television Broadcasting
51511	Radio Broadcasting
515112	Radio Stations

United States® Census Bureau

Economic Census (EC)

- Conducted by the Census Bureau for years ending in “2” or “7”
- Covers approximately eight million establishments, most industries, and all geographic areas of the U.S.
- Key statistics
  - Total number of establishments
  - Total number of employees
  - Value of sales, shipments, receipts, and revenue
  - Total annual payroll
- Data products are presented by NAICS and geography

United States® Census Bureau

Primary Business or Activity Question

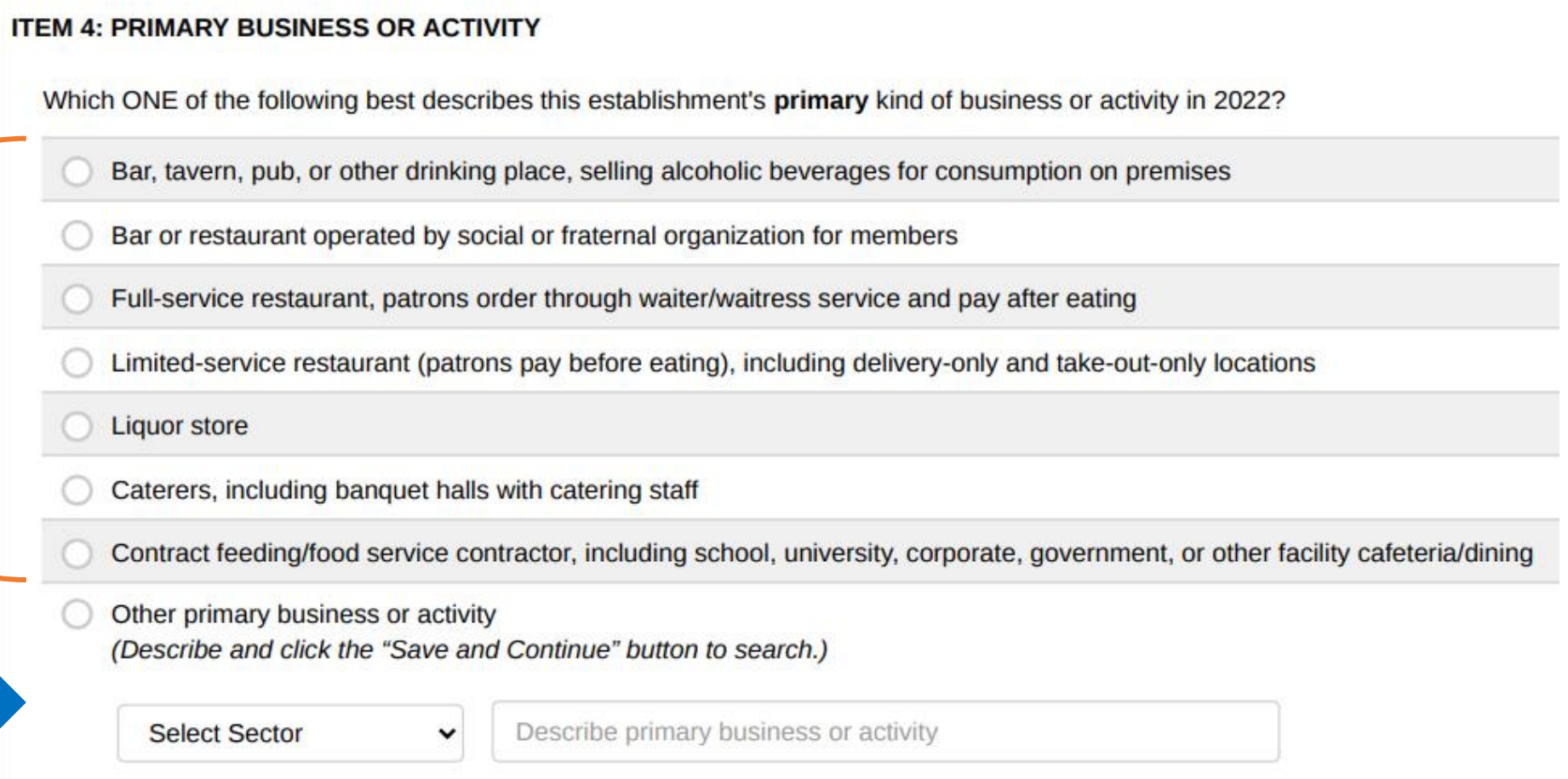
- Question asks respondents to describe their business
- Respondent is presented with prelisted descriptions based on an estimated NAICS code at the time of mailout
- Respondent can also provide a short, open-ended response
- There are hundreds of thousands of these so-called “write-in” responses every EC
- Clerical analysis of write-in text is mostly manual
- Using more automated methods can improve efficiency

United States® Census Bureau

Primary Business or Activity Question (cont.)

Prelisted descriptions based on estimated NAICS code

Write-in



ITEM 4: PRIMARY BUSINESS OR ACTIVITY

Which ONE of the following best describes this establishment's **primary** kind of business or activity in 2022?

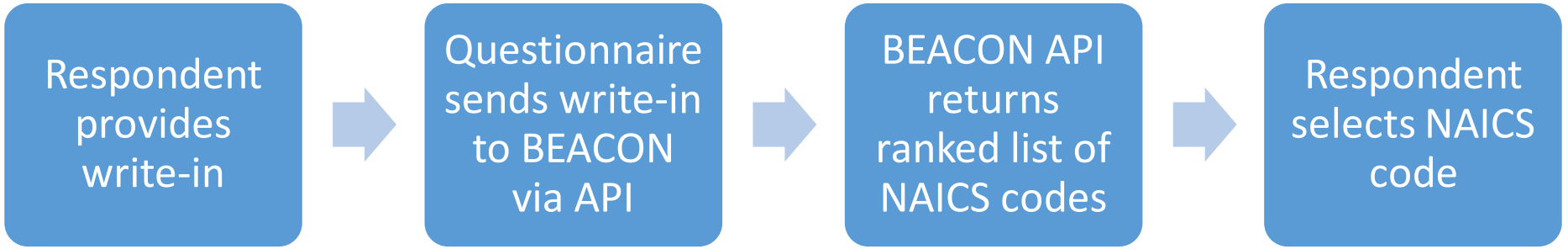
- ○ Bar, tavern, pub, or other drinking place, selling alcoholic beverages for consumption on premises
- ○ Bar or restaurant operated by social or fraternal organization for members
- ○ Full-service restaurant, patrons order through waiter/waitress service and pay after eating
- ○ Limited-service restaurant (patrons pay before eating), including delivery-only and take-out-only locations
- ○ Liquor store
- ○ Caterers, including banquet halls with catering staff
- ○ Contract feeding/food service contractor, including school, university, corporate, government, or other facility cafeteria/dining
- ○ Other primary business or activity
  (Describe and click the "Save and Continue" button to search.)

Select Sector | Describe primary business or activity

Source: https://bhs.econ.census.gov/ombpdfs2022/export/2022_AF-72240_su.pdf

BEACON Overview

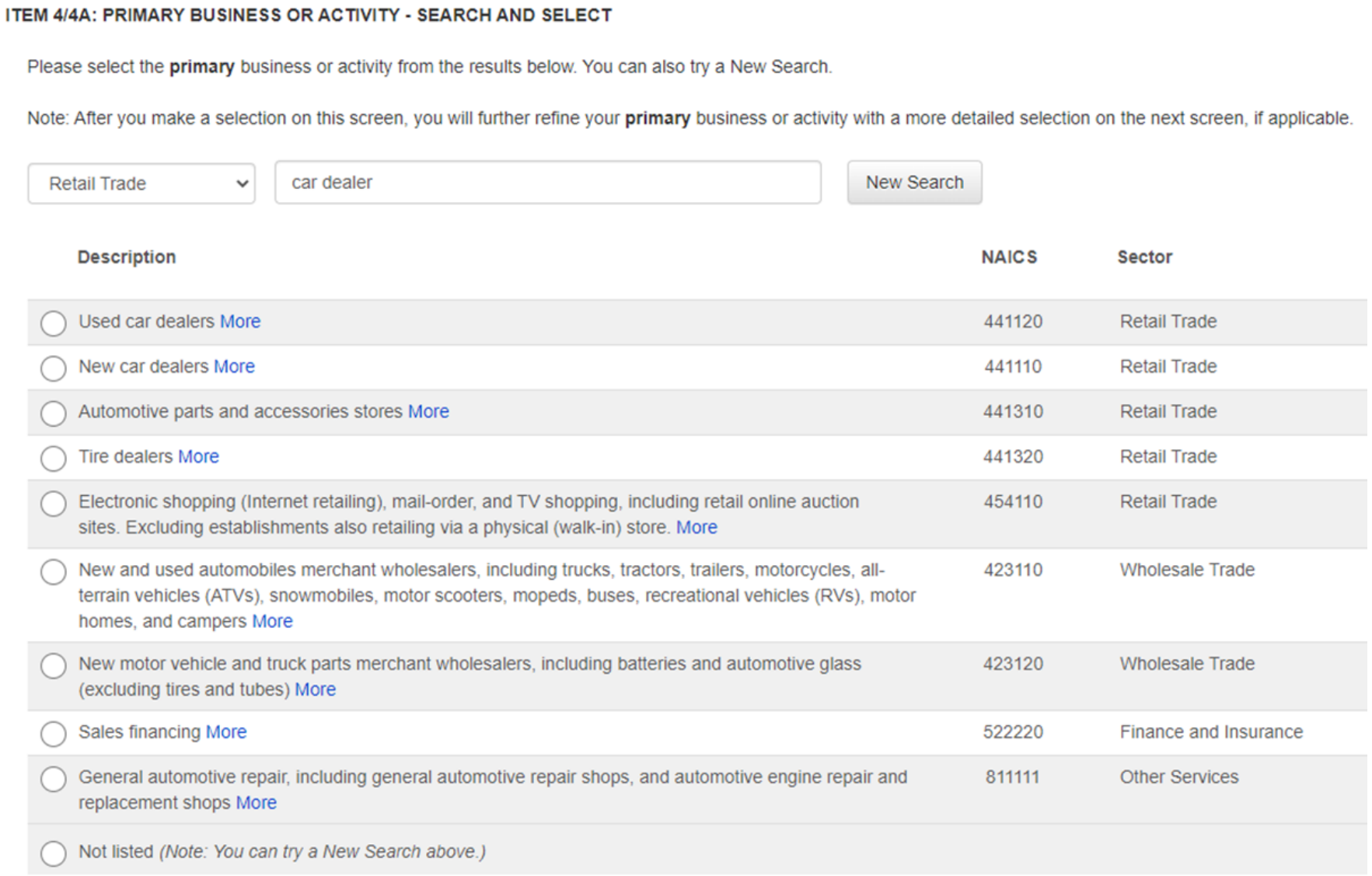
- Business Establishment Automated Classification of NAICS
- Text classification tool used in the 2022 EC that aims to
  - Help respondents self-classify their NAICS code in real time
  - Ensure respondents receive appropriate industry questions
  - Reduce clerical work associated with write-ins
- Uses an Application Programming Interface (API)



United States® Census Bureau

BEACON Results Screen

- Respondent entered the description “car dealer”
- Possible actions
  - Select a NAICS code
  - Click on “More” to view NAICS-specific page on www.census.gov/naics
  - Try a new search
  - Go back to the prelist screen
  - Select “Not listed”



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.

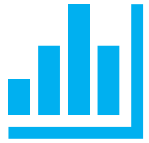
Note: After you make a selection on this screen, you will further refine your **primary** business or activity with a more detailed selection on the next screen, if applicable.

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.)*		

Methodology Overview

- Rich training data
- Thorough text cleaning algorithm
- Large dictionary of features
  - Words
  - 2- and 3-word combinations
  - Full-length descriptions
- Optimized model ensemble using machine learning



3.9+ million observations

28,000+ text cleaning rules

475,000+ model features in dictionary

3 information retrieval sub-models in ensemble

United States® Census Bureau

Training Data

- Data sources
  - Historical write-in responses to the EC
  - Frequent write-in text that was autocoded during 2017 EC
  - Business descriptions from Internal Revenue Service (IRS) SS-4 forms
  - Classification Assistance Tool (CAT) items
  - Harmonized System (HS) commodity descriptions
- Variables
  - Business description text
  - Corresponding NAICS code

Text	NAICS
This is a convenence store.	445120
automobile MFG	336111
We rapair watches & jewelry.	811490

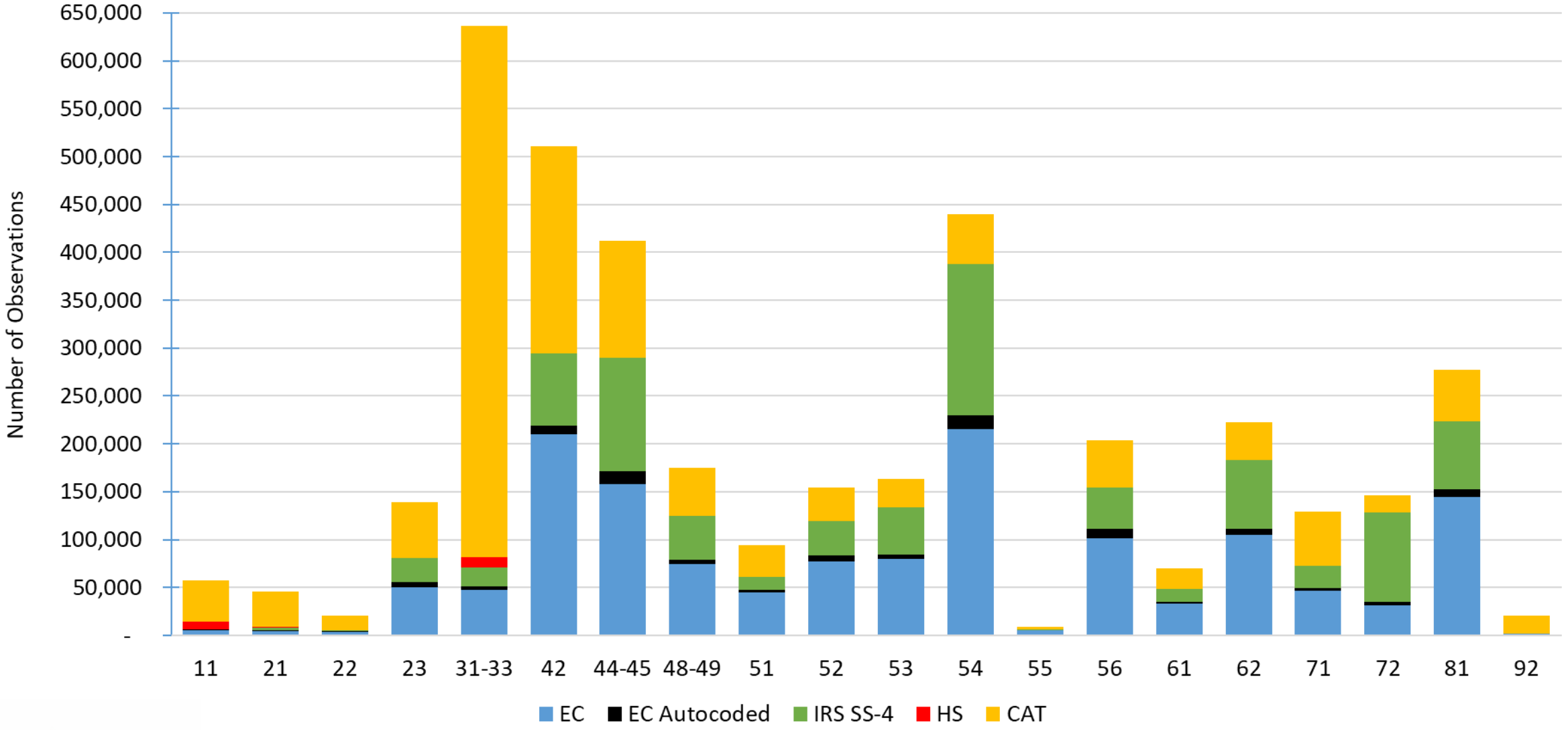
United States® Census Bureau

Data Source	Number of Observations	Advantages	Disadvantages
EC	1,439,000	• Represents target population • Reflects natural language	• Descriptions not perfectly classified • Descriptions contain misspellings
EC Autocoded	*98,000	• Improves consistency with autocoding during 2017 EC	• Relatively small data source
IRS SS-4	865,000	• Reflects natural language	• Descriptions not perfectly classified • Descriptions contain misspellings
CAT	*1,507,000	• Provides a rich vocabulary • Descriptions classified correctly	• Does not always reflect natural language
HS	21,000	• Increases sample sizes for sectors not represented well in other data sources • Descriptions classified correctly	• Relatively small data source • Covers only three sectors (11, 21, and 31-33)

* Includes duplicates and variations of original observations

United States® Census Bureau

Training Data Breakdown by Sector and Source



Sources: 2002, 2007, 2012, 2017, 2022 Economic Census; Internal Revenue Service; Classification Assistance Tool; and Harmonized System
Date: 29 March 2023

Text Cleaning

- Removal of extraneous words/symbols
  - Remove extra white space and common “stop” words (“the”, “and”, “or”, etc.)
  - Account for numbers and punctuation
- Correct common misspellings
  - Map misspelled words to stem of true word
  - For example, “manifactur” → “manufactur”
- Stem
  - Apply prefix/suffix stripping rules to reduce number of word variations
  - For example, “manufacturing” → “manufactur”, “cars” → “car”
- Lemmatize
  - Map synonyms and abbreviations to a common concept
  - For example, “mfg” → “manufactur”, “auto” → “car”

United States® Census Bureau

Text Cleaning (cont.)

Text	Clean Text
new & ussed car dealer - ship	new used car dealership
automobile MFG	car manufactur
3PL	thirdparti logist
long dist trckng	long distanc trucking
Mini Golf with Juicebar!	minigolf juic bar
We rapair watches & jewelry.	repair watch jewelri
This is a convenence store.	conveni store
we do liq dist	liquor distribut
NAICS code #722511	722511

United States® Census Bureau

Dictionary

- Words and word combinations that BEACON recognizes
  - Individual words
  - 2- and 3-word combinations
  - Full-length/exact descriptions

- Data dictionary contains model features for component models

- Associations between words and NAICS codes influence predictions
  - "tutor" is highly associated with NAICS 611691 – Exam Preparation and Tutoring
  - "store" occurs in many NAICS codes and is therefore less predictive

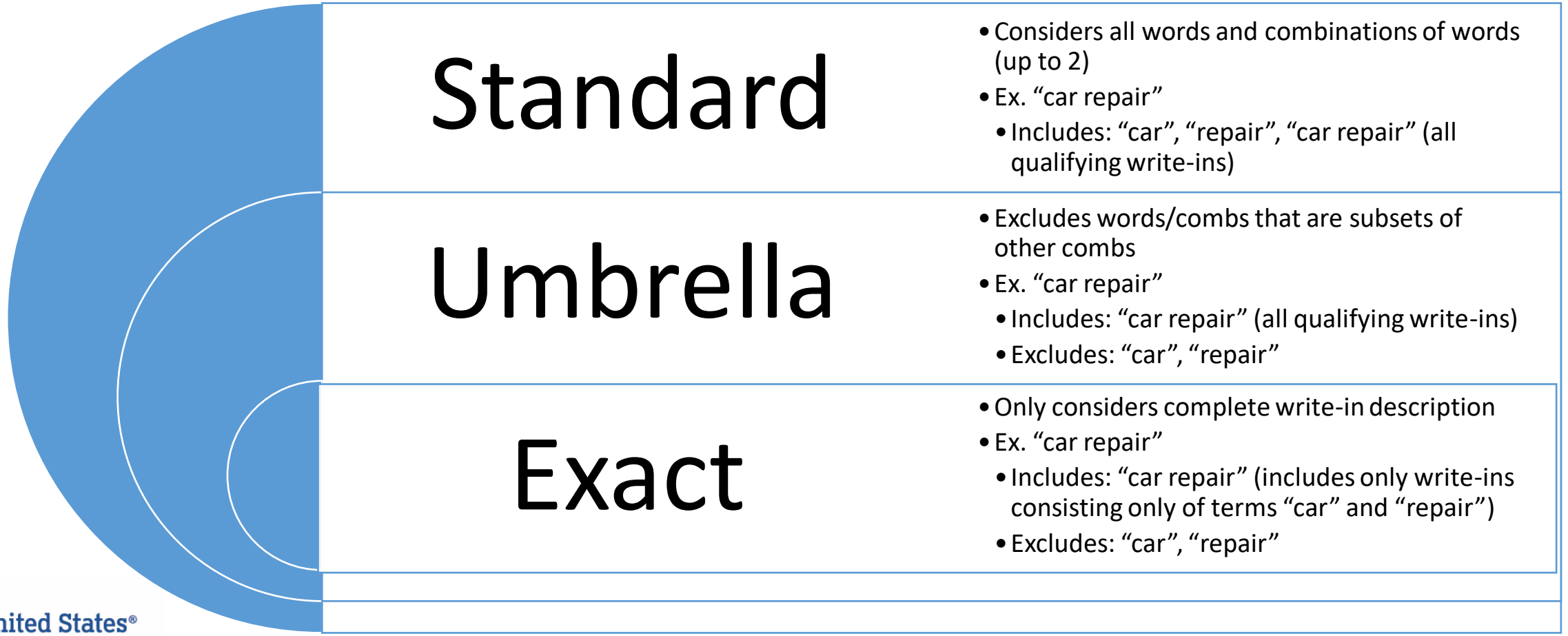
United States® Census Bureau

Model Ensemble

- Information retrieval models look at how words, combinations, and entire descriptions are distributed across NAICS codes
- Three information retrieval sub-models use different features
- Individual predictions are averaged, yielding relevance scores

United States® Census Bureau

Component Models

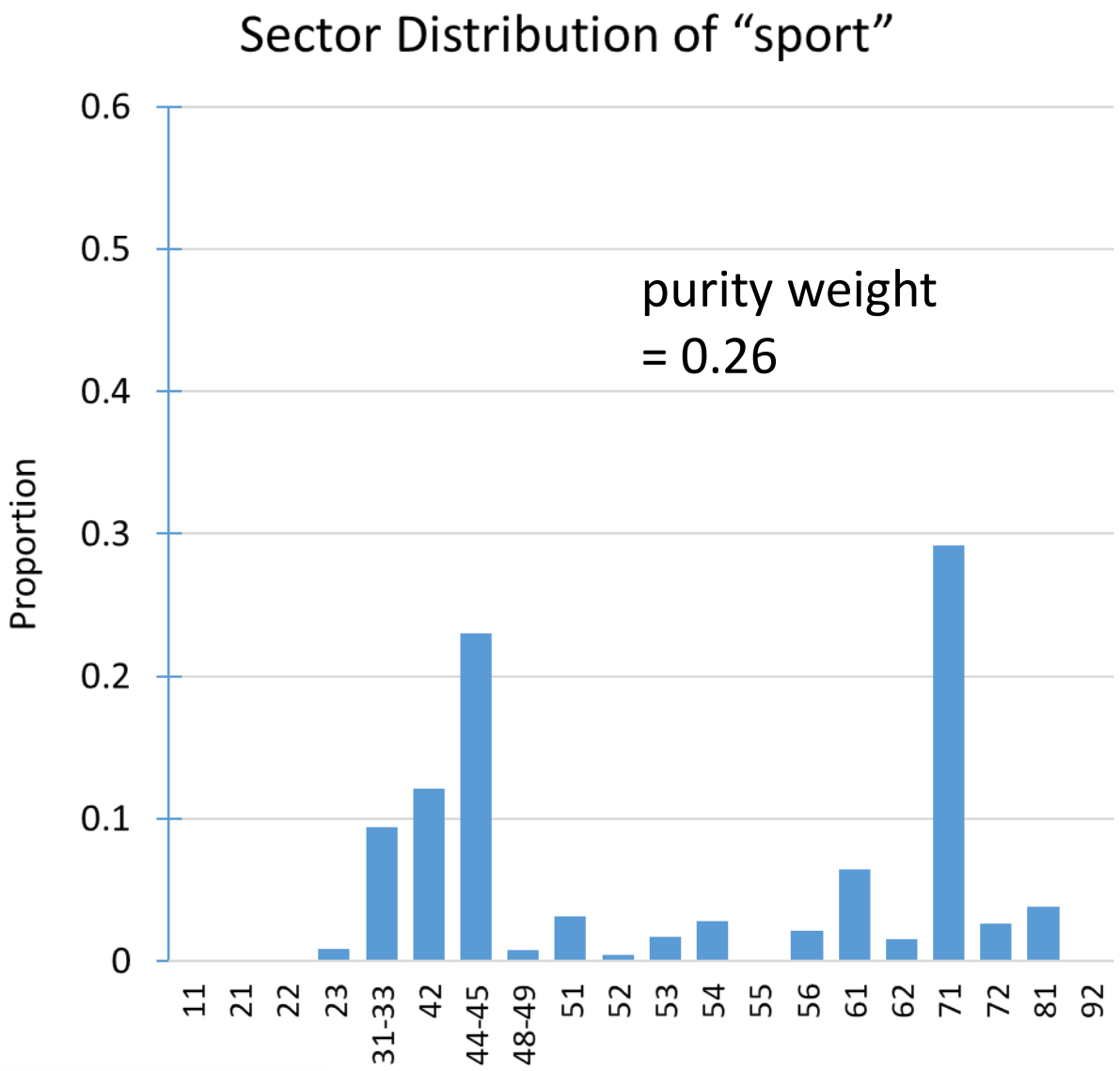


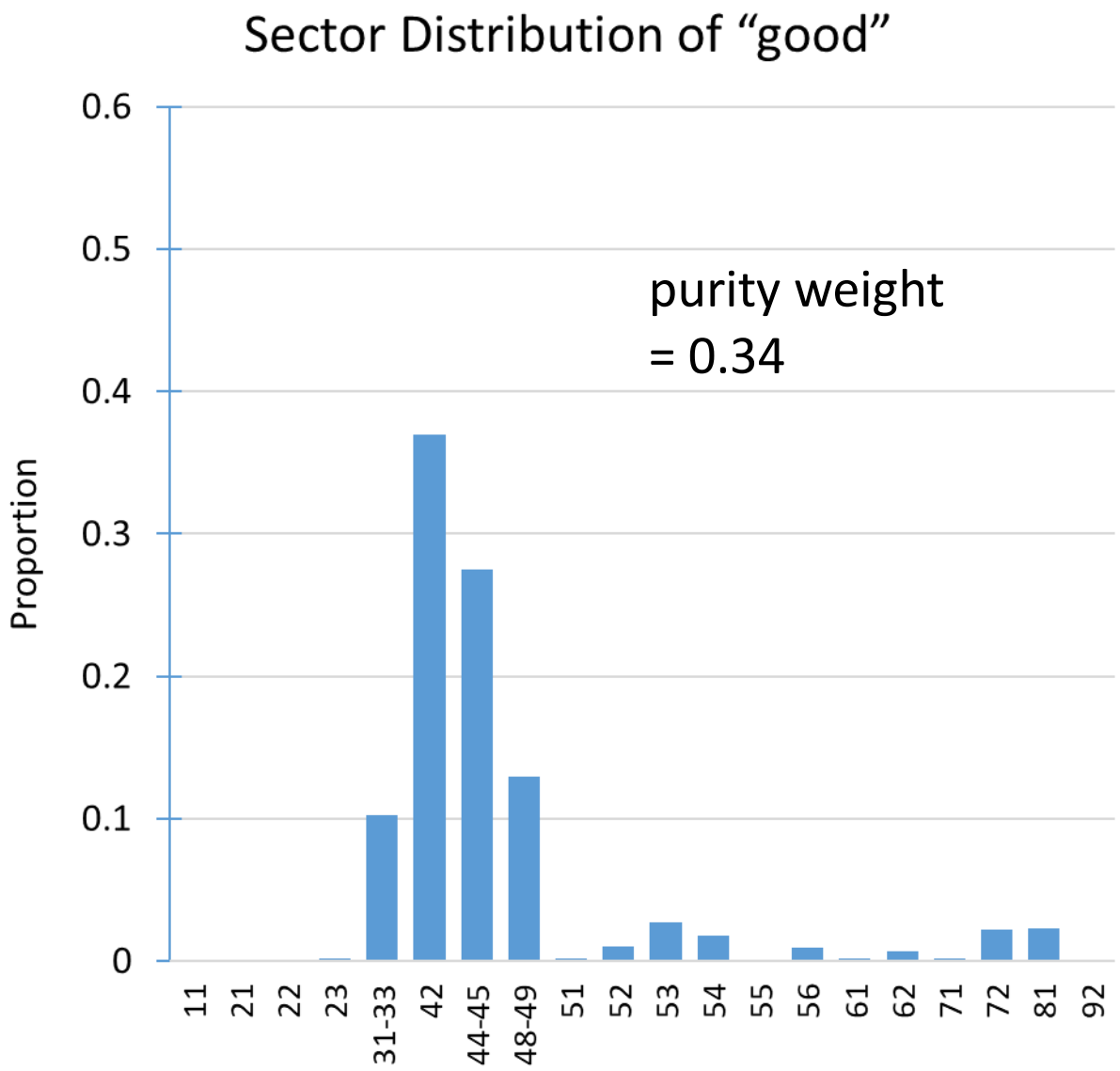
United States® Census Bureau

Purity Weights

- The NAICS distributions of the stems and stem combinations are averaged using “purity weights” that give more weight to the NAICS distributions of words that are more predictive.
- Only applied to the “standard” and “umbrella” models.
- Designed so that all terms have a weight between 0 and 1, inclusive.

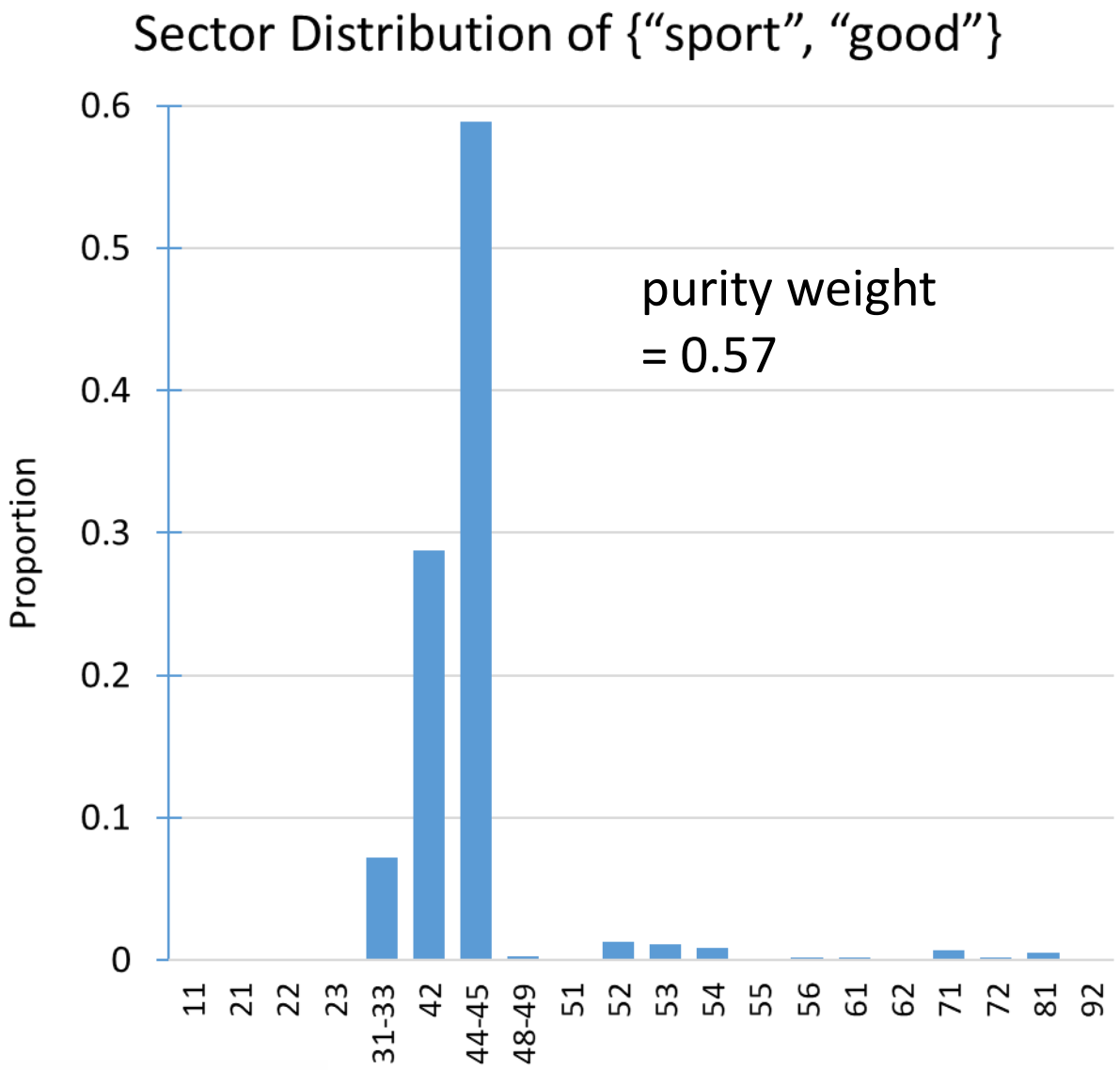
United States® Census Bureau

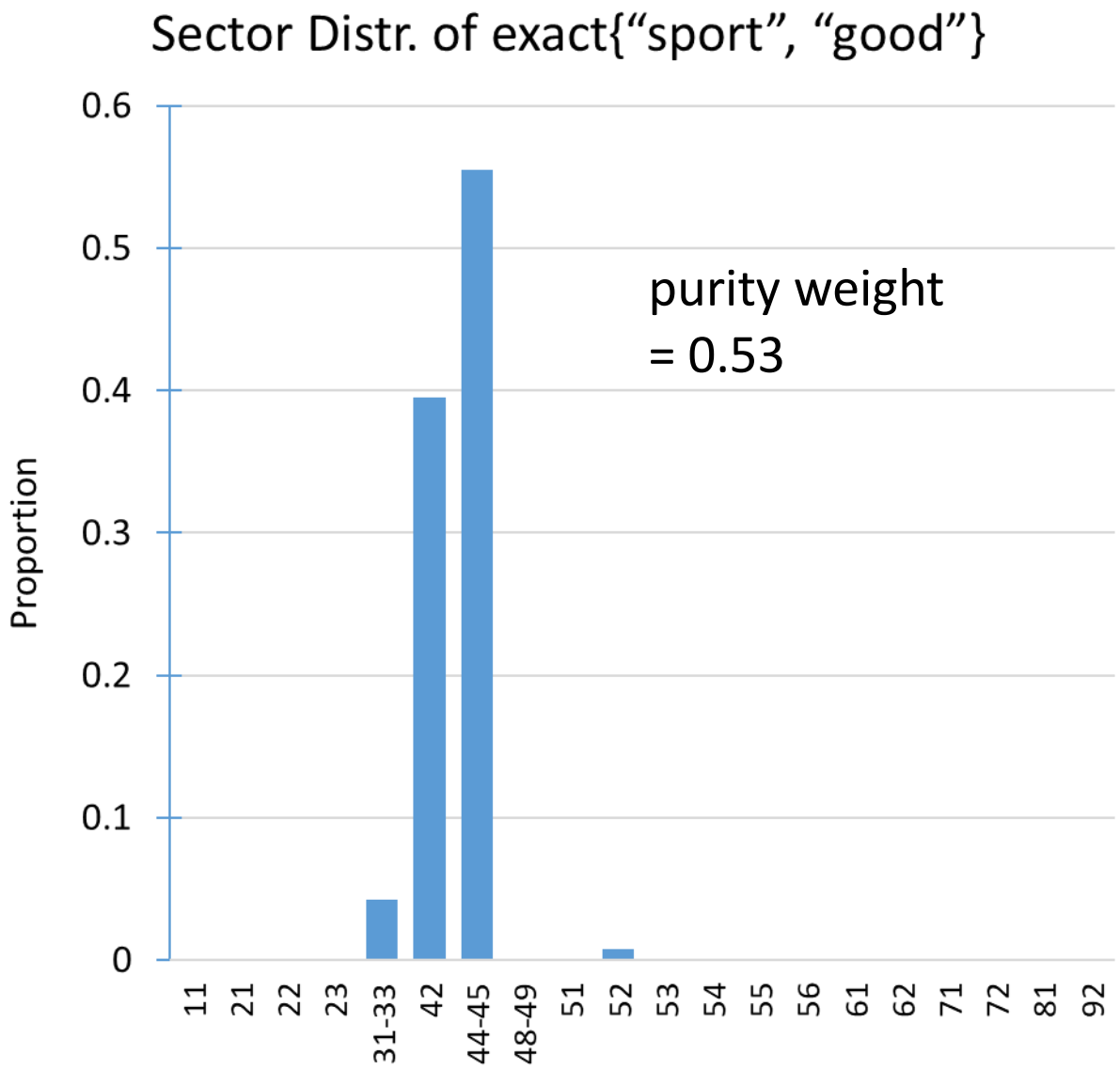




United States® Census Bureau

Sources: 2002, 2007, 2012, 2017, 2022 Economic Census; Internal Revenue Service; Classification Assistance Tool; and Harmonized System





Sources: 2002, 2007, 2012, 2017, 2022 Economic Census; Internal Revenue Service; Classification Assistance Tool; and Harmonized System

Relevance Scores

- BEACON assigns relevance scores to NAICS codes for ranking
- For each sub-model, the NAICS distributions of the relevant features are averaged using purity weights to calculate relevance scores
- The model ensemble calculates a weighted average of scores from the three sub-models, where the weights are optimized using machine learning

0.1 (Standard) + 0.6 (Umbrella) + 0.3 (Exact)

United States® Census Bureau

Hierarchical Model Structure

- Assigning relevance scores directly at the 6-digit level is challenging
- Model ensemble is applied 21 times
  - 1x to assign scores at the 2-digit level
  - 20x to assign sector-conditional scores at the 6-digit level
- Scores are calculated using the conditional probability formula

$$\text{score}(515112) = \text{score}(51) \times \text{score}(515112|51)$$

United States® Census Bureau

Future Work

- Continue to refine text cleaning algorithm
- Research more advanced models
  - Model stacking
  - Word embeddings
- Incorporate more Spanish language words
- Develop external interface for BEACON

United States® Census Bureau

Demo of Internal Interface

United States® Census Bureau

BEACON

Business Establishment Automated Classification of NAICS

Select the sector from the drop-down menu, enter a business description in the text field, and then click "Classify" to get a list of candidate 2017 NAICS codes. (The sector specifies what economic sector to give preference to in the search results. Leave unselected if no preference is desired.)

Select Sector

sporting goods

Classify

Advanced Parameters

Score Threshold 1

Max Candidates 10

Unit Type SU

Results

Standardized Text: **sport good**

Rank	NAICS	Score	Sector Description	NAICS Description
1	451110	41.58	Retail Trade	Sporting goods stores More
2	425120	15.39	Wholesale Trade	Agents and brokers, including manufacturer's representatives and auctions More
3	423910	13.11	Wholesale Trade	Sporting and recreational goods and supplies merchant wholesalers, including swimming pool equipment and supplies More
4	339920	6.01	Manufacturing	Sporting and athletic goods manufacturing, excluding apparel, footwear, and small arms and ammunition More
5	454110	3.72	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
6	448210	1.83	Retail Trade	Shoe stores More
7	453310	1.23	Retail Trade	Used merchandise and antiques stores, excluding general merchandise auction houses More
8	454390	1.16	Retail Trade	Other direct selling establishments, including specialized and general merchandise not sold from permanent locations, retail trade More
9	522298	1.08	Finance and Insurance	All other nondepository credit intermediation including pawn shops, title loans, title pawn, car title lending, factoring, agency of foreign bank - primarily commercial finance, and federally-sponsored credit agencies not elsewhere classified, etc. More

United States® Census Bureau

Sources: 2002, 2007, 2012, 2017, 2022 Economic Census; Internal Revenue Service; Classification Assistance Tool; and Harmonized System

BEACON

Business Establishment Automated Classification of NAICS

Select the sector from the drop-down menu, enter a business description in the text field, and then click "Classify" to get a list of candidate 2017 NAICS codes. (The sector specifies what economic sector to give preference to in the search results. Leave unselected if no preference is desired.)

Manufacturing

sporting goods

Classify

Advanced Parameters

Score Threshold 1

Max Candidates 10

Unit Type SU

Results

Standardized Text: **sport good**

Rank	NAICS	Score	Sector Description	NAICS Description
1	339920	6.01	Manufacturing	Sporting and athletic goods manufacturing, excluding apparel, footwear, and small arms and ammunition More
2	451110	41.58	Retail Trade	Sporting goods stores More
3	425120	15.39	Wholesale Trade	Agents and brokers, including manufacturer's representatives and auctions More
4	423910	13.11	Wholesale Trade	Sporting and recreational goods and supplies merchant wholesalers, including swimming pool equipment and supplies More
5	454110	3.72	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
6	448210	1.83	Retail Trade	Shoe stores More
7	453310	1.23	Retail Trade	Used merchandise and antiques stores, excluding general merchandise auction houses More
8	454390	1.16	Retail Trade	Other direct selling establishments, including specialized and general merchandise not sold from permanent locations, retail trade More
9	522298	1.08	Finance and Insurance	All other nondepository credit intermediation including pawn shops, title loans, title pawn, car title lending, factoring, agency of foreign bank - primarily commercial finance, and federally-sponsored credit agencies not elsewhere classified, etc. More

United States® Census Bureau

Sources: 2002, 2007, 2012, 2017, 2022 Economic Census; Internal Revenue Service; Classification Assistance Tool; and Harmonized System

BEACON

Business Establishment Automated Classification of NAICS

Select the sector from the drop-down menu, enter a business description in the text field, and then click "Classify" to get a list of candidate 2017 NAICS codes. (The sector specifies what economic sector to give preference to in the search results. Leave unselected if no preference is desired.)

Select Sector

sporting goods distributor

Classify

Advanced Parameters

Score Threshold 1

Max Candidates 10

Unit Type SU

Results

Standardized Text: **sport good distribut**

Rank	NAICS	Score	Sector Description	NAICS Description
1	423910	45.41	Wholesale Trade	Sporting and recreational goods and supplies merchant wholesalers, including swimming pool equipment and supplies More
2	425120	21.17	Wholesale Trade	Agents and brokers, including manufacturer's representatives and auctions More
3	451110	7.94	Retail Trade	Sporting goods stores More
4	454110	7.44	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
5	713990	2.65	Other Services	All other amusement and recreation industries, miniature golf courses, archery or shooting ranges, day camps (except instructional), billiard parlors, recreational or youth sports teams, boating clubs (without marinas), dance halls, or riding stables More
6	713940	1.2	Other Services	Fitness centers, gyms, and other recreational sports centers, such as yoga studios, swimming pools, skating rinks, and sports ball facilities, fields, and courts More

United States® Census Bureau

Sources: 2002, 2007, 2012, 2017, 2022 Economic Census; Internal Revenue Service; Classification Assistance Tool; and Harmonized System

BEACON

Business Establishment Automated Classification of NAICS

Select the sector from the drop-down menu, enter a business description in the text field, and then click "Classify" to get a list of candidate 2017 NAICS codes. (The sector specifies what economic sector to give preference to in the search results. Leave unselected if no preference is desired.)

Select Sector

We rapair watches & jewelry.

Classify

Advanced Parameters

Score Threshold: 1

Max Candidates: 10

Unit Type: SU

Results

Standardized Text: **repair watch jewelri**

Rank	NAICS	Score	Sector Description	NAICS Description
1	811490	64.06	Other Services	Other personal and household goods repair and maintenance, including clocks, jewelry, musical instruments, watches, garments, bicycles, motorcycles, boats, and motorboats More
2	448310	27.38	Retail Trade	Jewelry stores, including watches and clocks More
3	423940	6.23	Wholesale Trade	Jewelry (costume and fashion) merchant wholesalers, including watches, diamonds, gemstones, silverware, trophies, precious metals, and coins (excluding precious metal ores) More

United States® Census Bureau

Sources: 2002, 2007, 2012, 2017, 2022 Economic Census; Internal Revenue Service; Classification Assistance Tool; and Harmonized System

References

- Aggarwal, C.C. (2018). *Machine learning for text*. Cham: Springer International Publishing.
- Dumbacher, B. and Whitehead, D. (2022). Industry self-classification in the Economic Census. *2022 Proceedings of the American Statistical Association*, *Section on Statistical Learning and Data Science*, 1049–1064.
- Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M.A., Meira Jr., W. (2011). Word co-occurrence features for text classification. *Information Systems*, *36*(5), 843–858.
- U.S. Census Bureau. (2017). 2017 North American Industry Classification System Manual. https://www.census.gov/naics/reference_files_tools/2017_NAICS_Manual.pdf
- Whitehead D., Dumbacher B. (2023). Ensemble modeling techniques for NAICS classification in the Economic Census. *Proceedings of the 2023 Federal Committee on Statistical Methodology Research and Policy Conference*.
- Wiley, E. and Whitehead, D. (2022). Implementing interactive classification tools in the 2022 Economic Census. *Proceedings of the 2022 Federal Committee on Statistical Methodology Research and Policy Conference*.

United States® Census Bureau

Contact Information

- Sarah.Pfeiff@census.gov
- Daniel.Whitehead@census.gov

United States® Census Bureau