

# Machine Learning for In-Instrument Product Code Search

Emily Wiley (she/her)

Chief, Statistics Modernization Branch, U.S. Census Bureau

October 22, 2024

# Outline

- Background and motivation
- Machine learning classification model
- 2022 Economic Census Results
- Further research

# Background: Economic Census

- The Economic Census (EC) is the official measure of the United States' businesses and overall economy, conducted in years ending in 2 and 7
- Around 4 million business locations, covering most industries and all geographic areas of the United States, receive surveys tailored to their primary business activity
- The EC collects data on key economic data items, such as revenue, payroll, employment, and **products and services**

# Background: NAPCS

- The **North American Product Classification System (NAPCS)** is a hierarchical classification system for goods and services that is independent of NAICS industry
- NAPCS was introduced in the 2017 Economic Census
- There are 7,234 unique NAPCS codes

# NAPCS Reporting: 2017 Economic Census

- Businesses classify their revenue into one or multiple NAPCS codes
- Respondents choose from a prelist or write in a free-text description

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

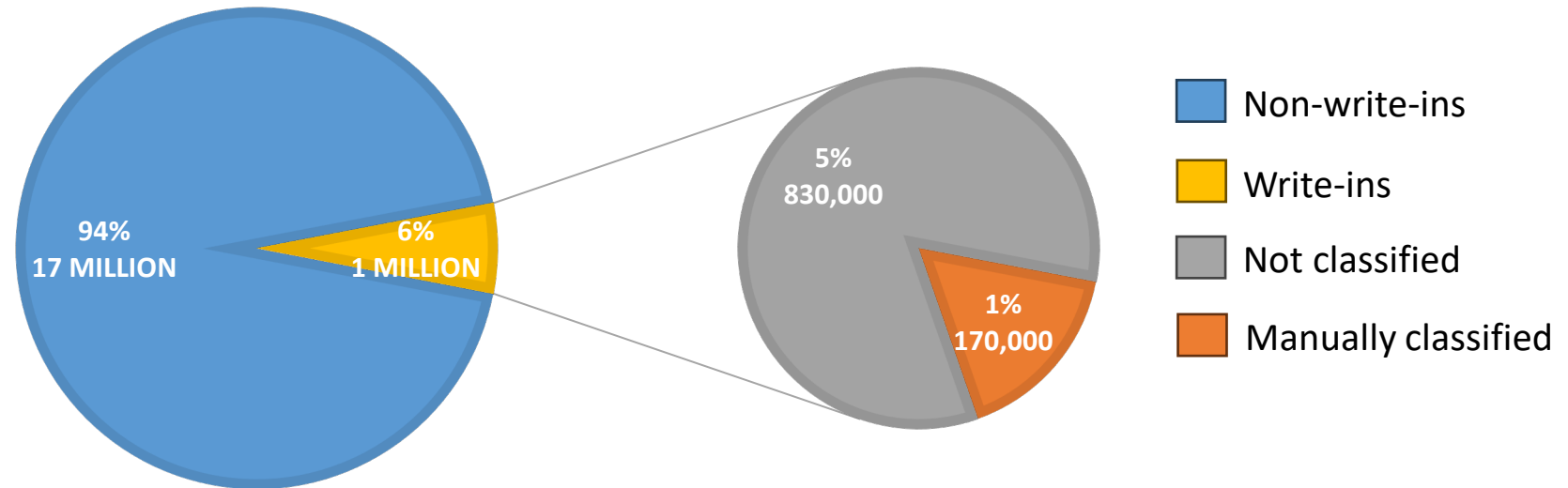
Of the \$,000.00 of Sales, Shipments, Receipts, or Revenue reported in **Item 5**, what was the value for each product or service?

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

Source: 2017 Economic Census

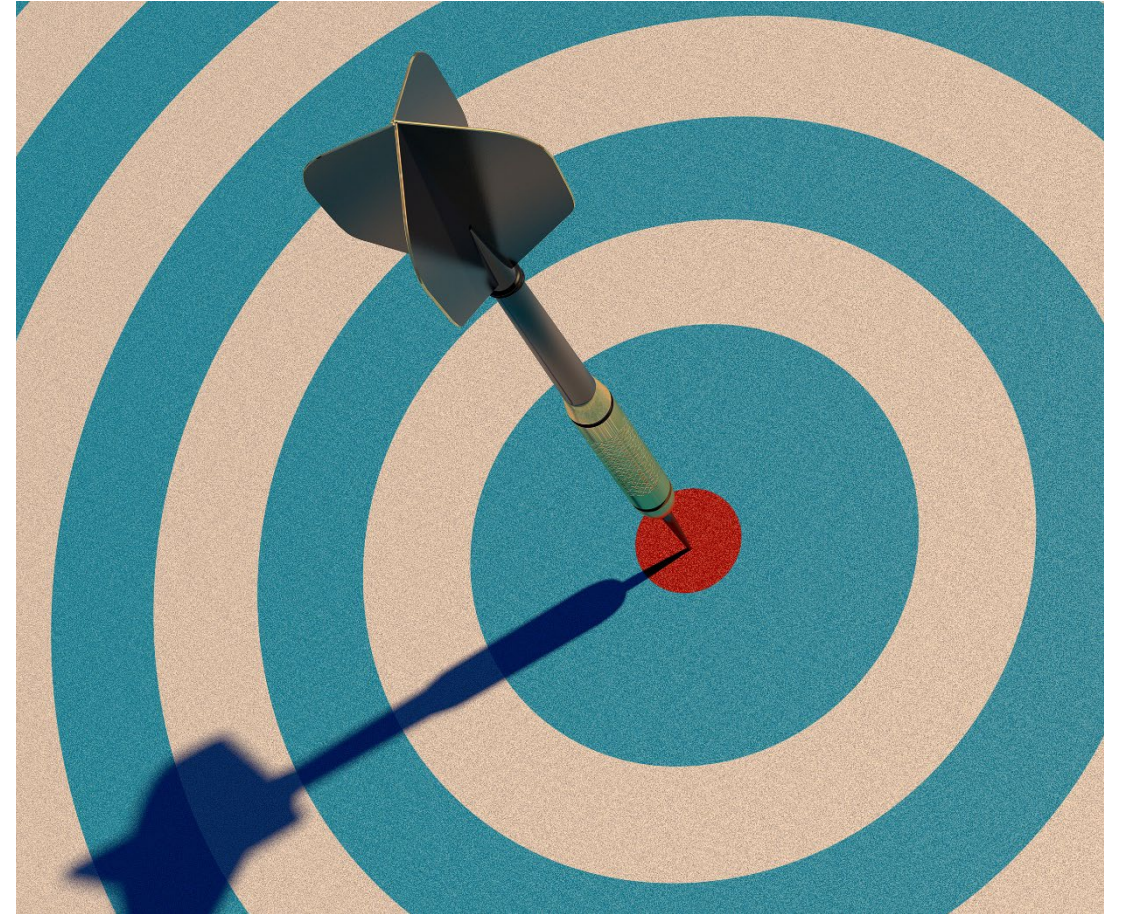
# Motivation for a Machine Learning Model

2017 NAPCS WRITE-INS  
1,000,000 WRITE-INS  
18 MILLION TOTAL NAPCS



# Goal for the 2022 Economic Census

Develop a machine learning model for NAPCS assignment and integrate it into the electronic reporting instrument for the 2022 Economic Census

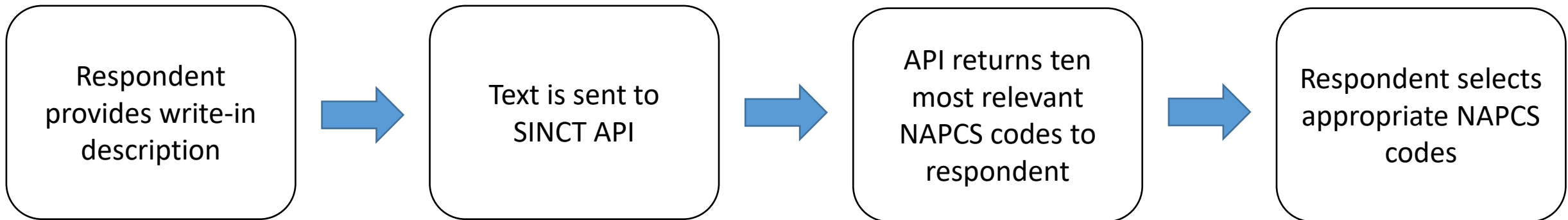


# What is SINCT?

Smart Instrument NAPCS Classification Tool



Source: istockphoto.com





# SINCT Training Data

- Initially:
  - Training data – 225,000
  - NAPCS codes – 7,234
- Subject matter experts removed 950 NAPCS codes
  - Training data – 149,500
  - NAPCS codes – 6,284

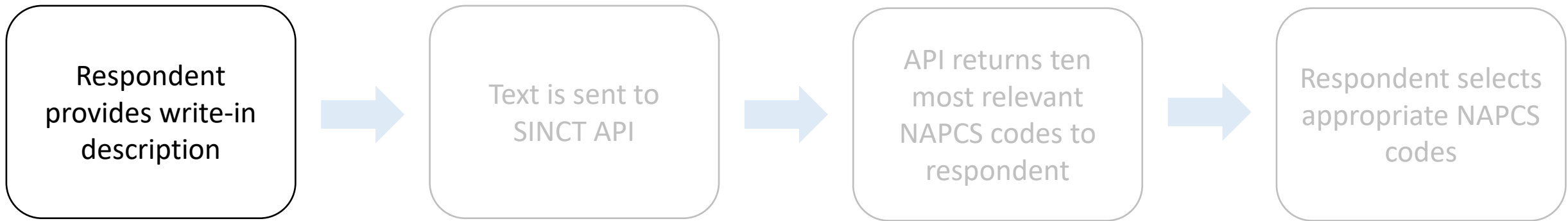
<b>Data Source</b>	<b>Number of Observations</b>	<b>Advantages</b>	<b>Disadvantages</b>
<b>Economic Census</b>	180,000	<ul style="list-style-type: none"> <li>• Represents target population</li> <li>• Reflects natural language</li> </ul>	<ul style="list-style-type: none"> <li>• Descriptions not classified perfectly</li> <li>• Descriptions contain misspellings</li> </ul>
<b>Classification Analytical Processing System</b>	24,000	<ul style="list-style-type: none"> <li>• Provides a rich vocabulary</li> <li>• Descriptions are classified correctly</li> </ul>	<ul style="list-style-type: none"> <li>• Text does not always reflect natural language</li> </ul>
<b>Subject matter expert (SME) examples</b>	11,500	<ul style="list-style-type: none"> <li>• Incorporates institutional knowledge</li> <li>• Can be continuously updated with SME feedback</li> </ul>	<ul style="list-style-type: none"> <li>• Small data source</li> </ul>
<b>NAPCS title file</b>	9,100	<ul style="list-style-type: none"> <li>• Definitive source of NAPCS descriptions</li> </ul>	<ul style="list-style-type: none"> <li>• Small data source</li> </ul>

# What is SINCT?

Smart Instrument NAPCS Classification Tool



Source: istockphoto.com



# SINCT Search

Search for product/service not listed

X

Use the search 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.

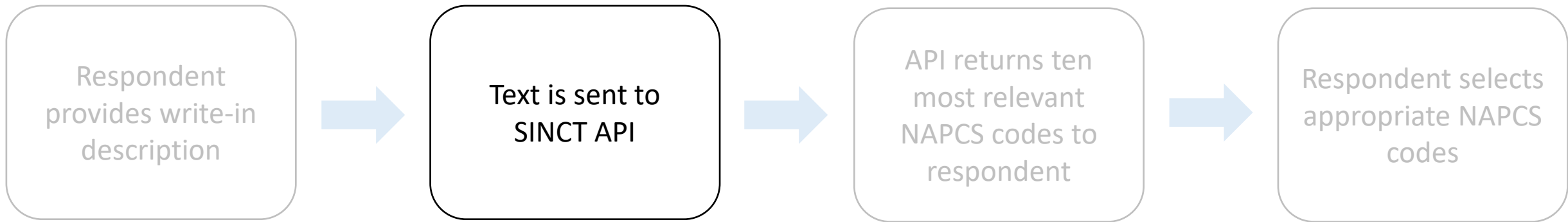
**Specify additional product or service:**

# What is SINCT?

Smart Instrument NAPCS Classification Tool

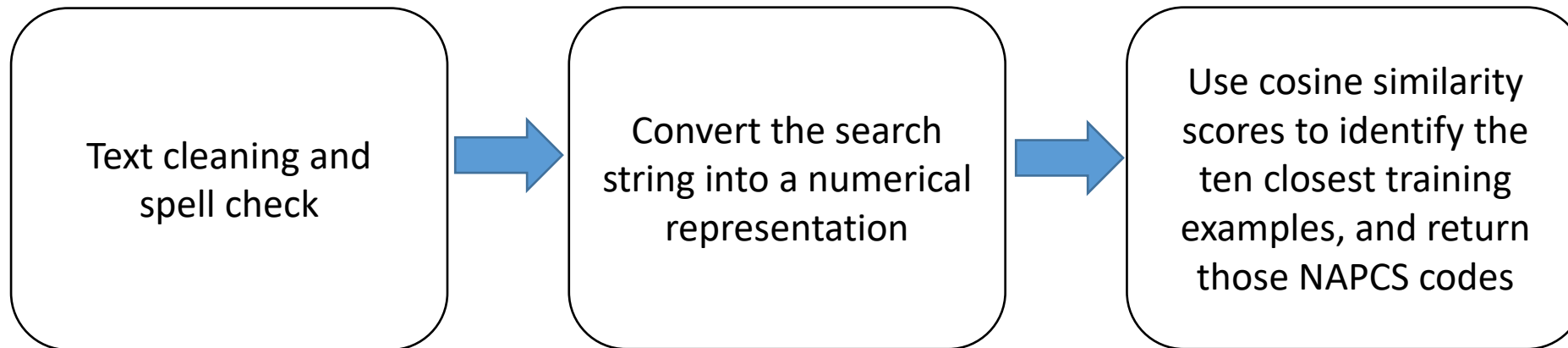


Source: istockphoto.com



# SINCT Methodology

- Doc2Vec<sup>1</sup> machine learning model
  - Unsupervised neural network for document embeddings
  - Represent words, phrases, sentences, or paragraphs as a vector

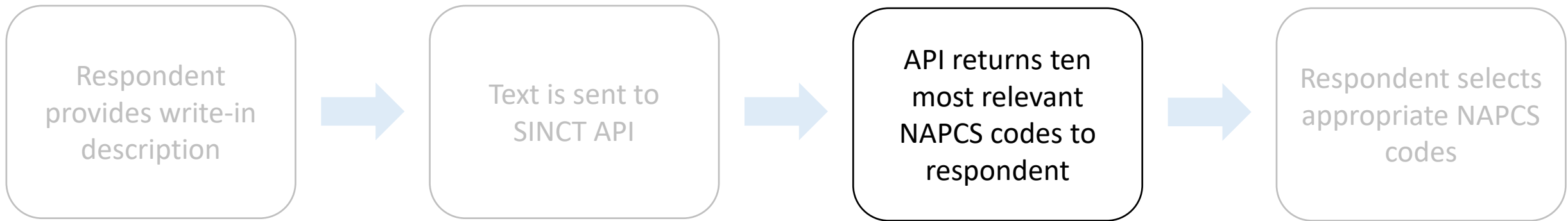


# What is SINCT?

Smart Instrument NAPCS Classification Tool



Source: istockphoto.com



# SINCT Results

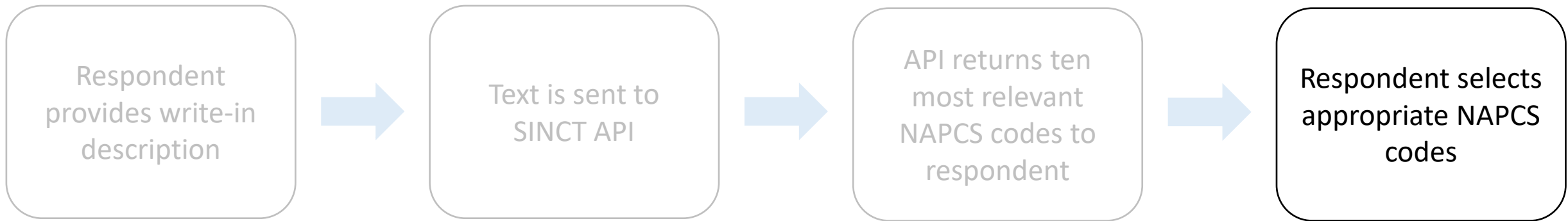
Rank	NAPCS Code	NAPCS Description	Score
1	7002650006	Automatic washing and waxing services for automobiles and light-duty trucks	0.970
2	7002650012	Self-service washing and waxing for automobiles and light-duty trucks	0.9696
3	7002650009	Hand washing, with or without waxing services, for automobiles and light-duty trucks	0.9649
4	7002650015	Self-service vacuuming services for automobiles and light-duty trucks	0.9584
5	7009375042	Washing and cleaning services for heavy trucks and buses	0.9582
6	7002650003	Detailing services for automobiles and light-duty trucks	0.9574
7	7002675000	Regulatory safety inspections and emissions testing services for automobiles and light-duty trucks	0.9463
8	5001950031	Retail sales of automotive supplies, including appearance and maintenance chemicals; automotive paint; antifreeze; functional fluids; etc.	0.8868
9	5001725015	Retail sales of other automotive fuels	0.8844
10	7003135000	Sightseeing services by rail	0.8830

# What is SINCT?

Smart Instrument NAPCS Classification Tool



Source: istockphoto.com





# SINCT Results

- 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 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.*

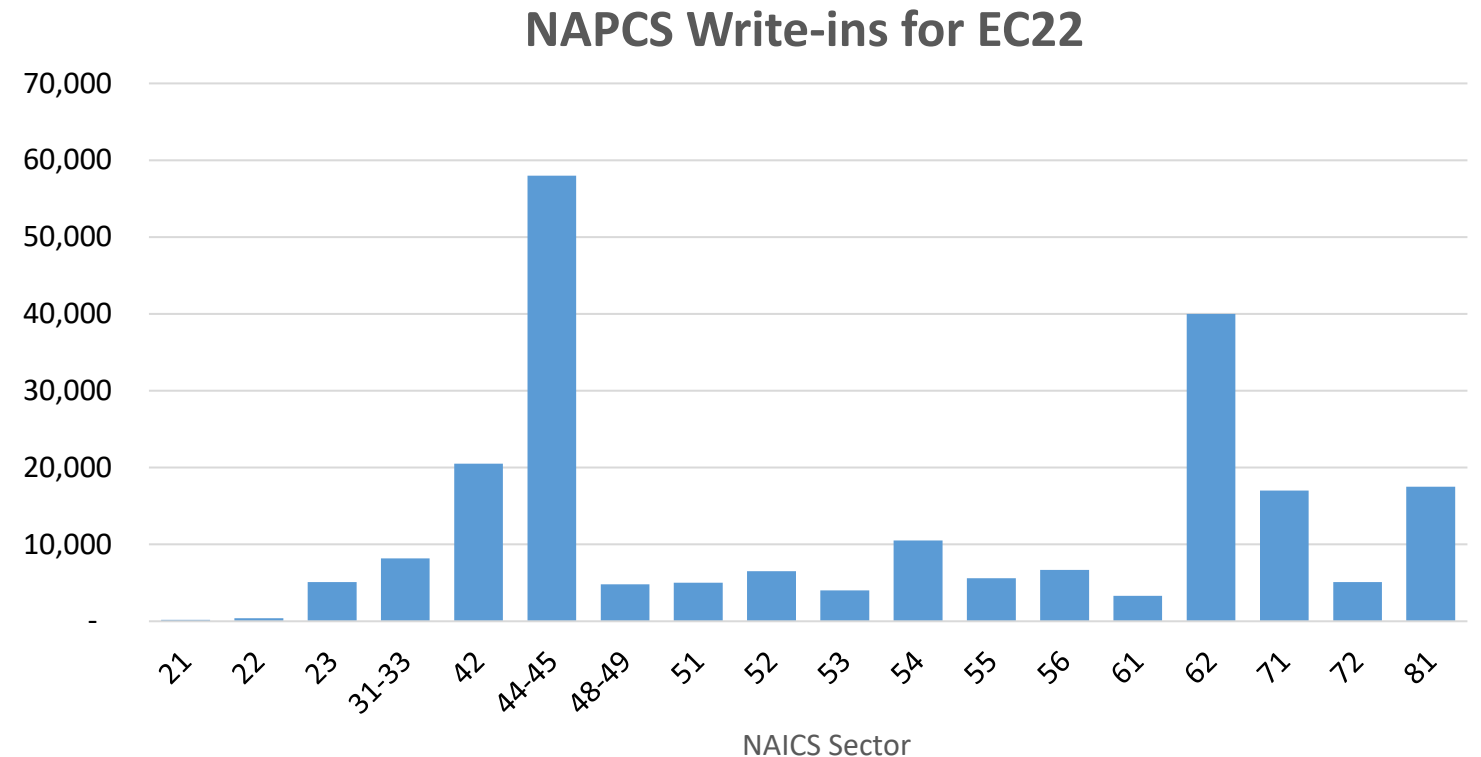
Select	Description	Product Code
<input type="checkbox"/>	Automatic washing and waxing services for automobiles and light-duty trucks	7002650006
<input type="checkbox"/>	Self-service washing and waxing for automobiles and light-duty trucks	7002650012

Not listed (Note: You can try a new Search above.)

Add Selected Products/Services

# Results: 2022 Economic Census

- EC17 to EC22:
  - **78% reduction!**
  - **220,000 write-ins, down from 1,000,000**
- Remaining Write-ins
  - Retail (44-45) - 26.4%
  - Healthcare (62) - 18.1%
  - Wholesale (42) - 9.2%
  - Other Services (81) - 8.0%



# Remaining Write-Ins

<b>Write-in term</b>	<b>Count</b>
Other	10,000
Miscellaneous	9,000
General retail	6,500
Financial	6,000
Automotive tools	6,000
Automotive electrical	6,000
Interest	3,000

+ 70,000 additional terms!

# Further Research

- What will SINCT look like in the 2027 Economic Census?
  - Capability for spreadsheet reporters
  - Spanish-language version
- Can we extend SINCT to other surveys?
  - Annual Integrated Economic Survey
- Can we extend SINCT to other product/service codes?
  - Harmonized System codes
  - UPC/GTINs – barcodes
- Can we use transformer-based embeddings?

# Thank You!

[Emily.L.Wiley@census.gov](mailto:Emily.L.Wiley@census.gov)

[Clayton.G.Knappenberger@census.gov](mailto:Clayton.G.Knappenberger@census.gov)