# Using Natural Language Processing to Improve the Commodity Flow Survey

Christian Moscardi

Commodity Flow Survey

Innovations in Freight Data Workshop – April 2019

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

United States Department of Transportation
Bureau of Transportation Statistics

# Overview

- Commodity Flow Survey
- Commissioned by BTS
- Conducted every 5 years (2012, 2017)
- Respondents provide sampling of shipments from each quarter

12027041

4

*If you prefer to complete the questionnaire online, please go to* ***https://econhelp.census.gov/cfs***

**Item F** **SHIPMENT CHARACTERISTICS**

**NOTE: Each line runs across pages 4 and 5. After entering column (I) data on page 4 for any line, continue with column (J) on page 5 for the same line.**

| Line No. (A) | Your Shipment ID Number (B) | Shipment Date (C) Month | Shipment Date (C) Day | Shipment value (excluding freight charges and excise taxes) in **whole dollars.** Estimates acceptable. (D) | Net Shipment Weight **in pounds.** Estimates acceptable. (E) | For shipments consisting of more than one commodity, report the code and description of the commodity that contributed the greatest weight of the shipment in columns (F) through (I): SCTG commodity code from accompanying booklet[1] (F) | Commodity Description[1] (G) | Is item in col. (G) Temperature controlled? (Y/N)[1,2] (H) | Is item in col (G) a hazardous material? Enter "UN" or "NA"[1] number (I) | Continue with column (J) on page 5 |
|---|---|---|---|---|---|---|---|---|---|---|
| Ex.1 | 123-5 | 4 | 26 | 224,235 | 4,840 | 34520 | Mechanical machinery | Y | | → |
| Ex.2 | 402H | 4 | 26 | 1,375 | 50,125 | 20222 | Sulfuric acid | N | 1830 | → |
| 1 | | | | | | | | | | → |
| 2 | | | | | | | | | | → |
| 3 | | | | | | | | | | → |
| 4 | | | | | | | | | | → |

**Report Online - Do Not Return**

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
*census.gov*

United States Department of Transportation
Bureau of Transportation Statistics

# Overview

- Commodity Flow Survey
- Commissioned by BTS
- Conducted every 5 years (2012, 2017)
- Respondents provide sampling of shipments from each quarter

12027041

4

*If you prefer to complete the questionnaire online, please go to* ***https://econhelp.census.gov/cfs***

**Item F** **SHIPMENT CHARACTERISTICS**

**NOTE: Each line runs across pages 4 and 5. After entering column (I) data on page 4 for any line, continue with column (J) on page 5 for the same line.**

| Line No. (A) | Your Shipment ID Number (B) | Shipment Date (C) Month | Shipment Date (C) Day | Shipment value (excluding freight charges and excise taxes) in **whole dollars.** Estimates acceptable. (D) | Net Shipment Weight **in pounds.** Estimates acceptable. (E) | For shipments consisting of more than one commodity, report the code and description of the commodity that contributed the greatest weight of the shipment in columns (F) through (I): SCTG commodity code from accompanying booklet[1] (F) | Commodity Description[1] (G) | Is item in col. (G) Temperature controlled? (Y/N)[1,2] (H) | Is item in col (G) a hazardous material? Enter "UN" or "NA",[1] number (I) | Continue with column (J) on page 5 |
|---|---|---|---|---|---|---|---|---|---|---|
| Ex.1 | 123-5 | 4 | 26 | 224,235 | 4,840 | 34520 | Mechanical machinery | Y | | → |
| Ex.2 | 402H | 4 | 26 | 1,375 | 50,125 | 20222 | Sulfuric acid | N | 1830 | → |
| 1 | | | | | | | | | | → |
| 2 | | | | | | | | | | → |
| 3 | | | | | | | | | | → |
| 4 | | | | | | | | | | → |

**Report Online - Do Not Return**

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
*census.gov*

United States Department of Transportation
Bureau of Transportation Statistics

**For shipments consisting of more than one commodity, report the code and description of the commodity that contributed the greatest weight of the shipment in columns (F) through (I)**

| SCTG commodity code from accompanying booklet[1] (F) | Commodity Description[1] (G) | Is item in col. (G) Temperature controlled? (Y/N)[1,2] (H) | Is item in col (G) a hazardous material? Enter "UN" or "NA"[1] number (I) |
|---|---|---|---|
| **34520** | **Mechanical machinery** | **Y** | |
| **20222** | **Sulfuric acid** | **N** | **1830** |

# Overview

- ITEM G - Other Clarifying Information

**"Pulling this information was a huge spend of time and resources.“**

**"Just glad this is over!!"**

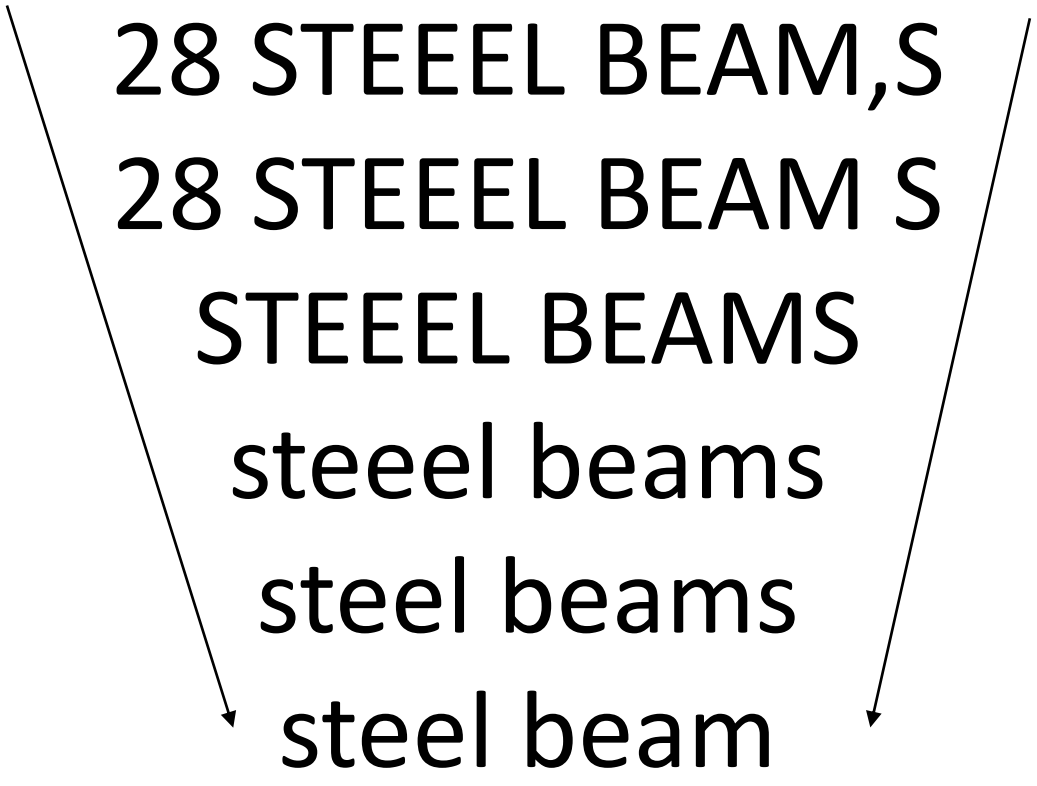
# Overview

**Using Machine Learning, can we automate the assignment of SCTG codes to shipment records?**

**(Yes.)**

# Initial Model

- Data
  1. Labelled Records (6.4M) from 2017 CFS
- Preprocessing
  1. "Throw out" SCTG 40999, 43999
     - These are miscellaneous SCTG codes
  2. Spell-check, stem, de-duplicate
  3. Left with ~400,000 unique training records
- Feature engineering
  1. "Bag-of-words" + TF-IDF scores
- Modelling
  - Logistic Regression, "elastic net" regularization
  - Cross-validate, hold out test set, etc.



# Further Investigation

- Initial results: ~50% “accuracy”
  - What does that mean?
  - Let’s call it “recovery”
- Should we use a more complex pipeline?
- Aside from 40999, 43999, ~80 more “other” codes
  - Remove these codes, recovery jumps to 64%

Other parts for motor vehicles, not elsewhere classified *(includes seat belts and seat covers, trims, plastics grilles, suspension shock-absorbers, radiators, mufflers, exhaust pipes, clutches, axles, bumpers, and steering wheels) (excludes parts for motorcycles, mopeds and armored fighting vehicles, see 36351 and 36391; engines and engine parts, see 341xx; pumps for liquids, see 34310; filters, see 34999; tires, see 24310; glass, see 313xx; lighting and signaling equipment, see 35992; ignition and starting equipment, see 35991; windshield wipers and defrosters, see 35992; seats, see 39029; and catalytic converters, see 34999)*. . . . . 36409

# Further Investigation

- E.g. **40994**
  - Sewing and knitting needles (includes for machines) crochet hooks, hook and eye **fasteners**, safety pins, straight pins, buttons, buckles and clasps, tubular and bifurcated rivets, **snap-fasteners**, zippers, and similar notions.

Image courtesy Wikimedia commons

# Further Investigation

- **Model's prediction**

- **33310**
  - Nails, screws, bolts, nuts, washers, staples except in strips, and similar **fastening** articles
- What was the NAICS Code?

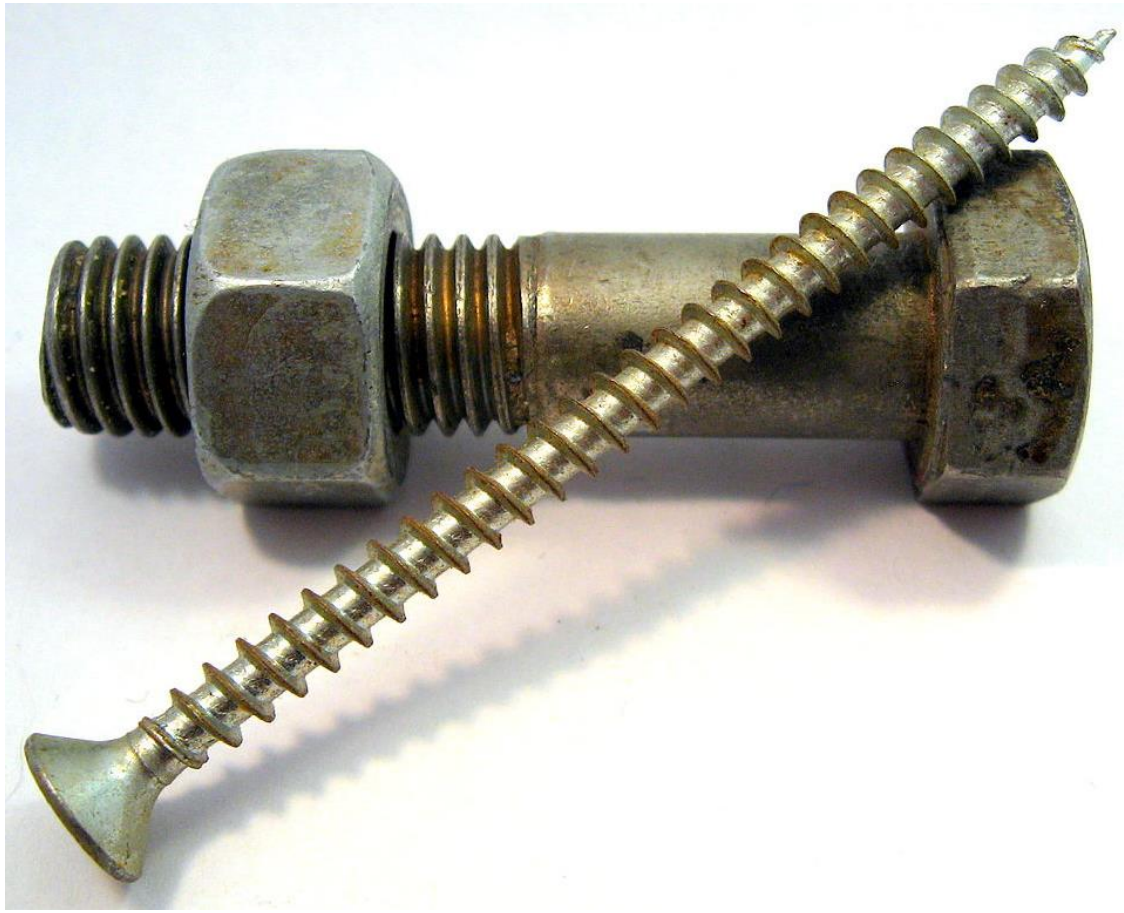
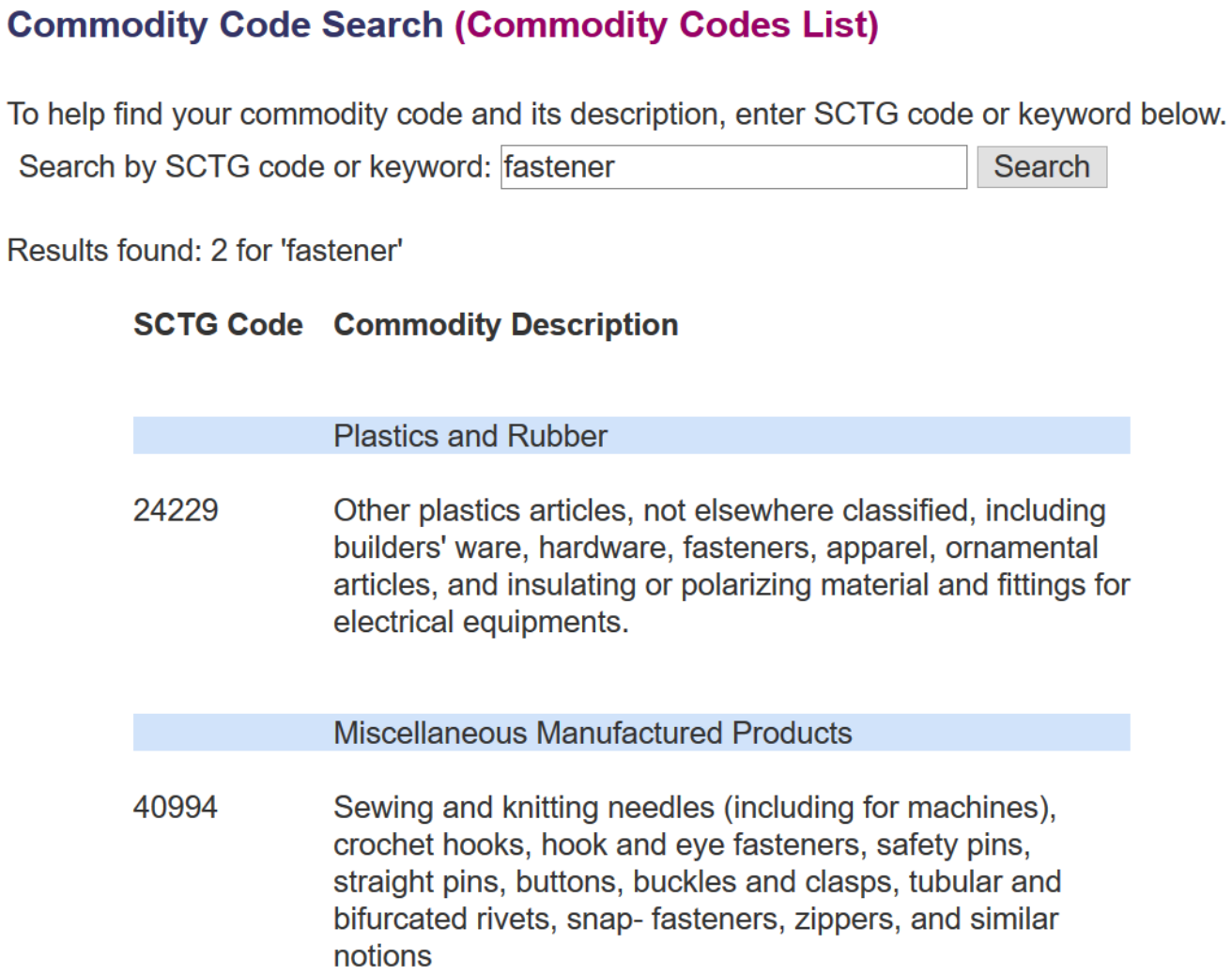


Image courtesy Wikimedia commons

# Further Investigation

- Manually validating, about 50% of items labelled 40994 by respondents were miscoded.
- However, the model was getting it right!
- **We can see** the workflow which led to these miscodings



**Commodity Code Search (Commodity Codes List)**

To help find your commodity code and its description, enter SCTG code or keyword below.

Search by SCTG code or keyword: fastener Search

Results found: 2 for 'fastener'

| SCTG Code | Commodity Description |
|---|---|
| | Plastics and Rubber |
| 24229 | Other plastics articles, not elsewhere classified, including builders' ware, hardware, fasteners, apparel, ornamental articles, and insulating or polarizing material and fittings for electrical equipments. |
| | Miscellaneous Manufactured Products |
| 40994 | Sewing and knitting needles (including for machines), crochet hooks, hook and eye fasteners, safety pins, straight pins, buttons, buckles and clasps, tubular and bifurcated rivets, snap- fasteners, zippers, and similar notions |

# Let’s Experiment

- Proof-of-concept: ran model on 170,000 unlabeled/invalid records
- 70,000 with probability score above predefined threshold [.5 – 1)
  - Determined by coarse inspection
- CFS Analysts validate a sample of 350 unique records

- Also wanted to determine accuracy in the [0 - .5) threshold
- Took sampling of the other 100,000 unlabeled / invalid records.
  - Model probability ranges [0 - .5)
  - 60 from each range

# Results

- Validation: **89% accurate in [.5 -1); 80% in [.4 - .5)**
  - "Accuracy" definition
- Batch-edits have saved **~1000 hours** of manual editing time

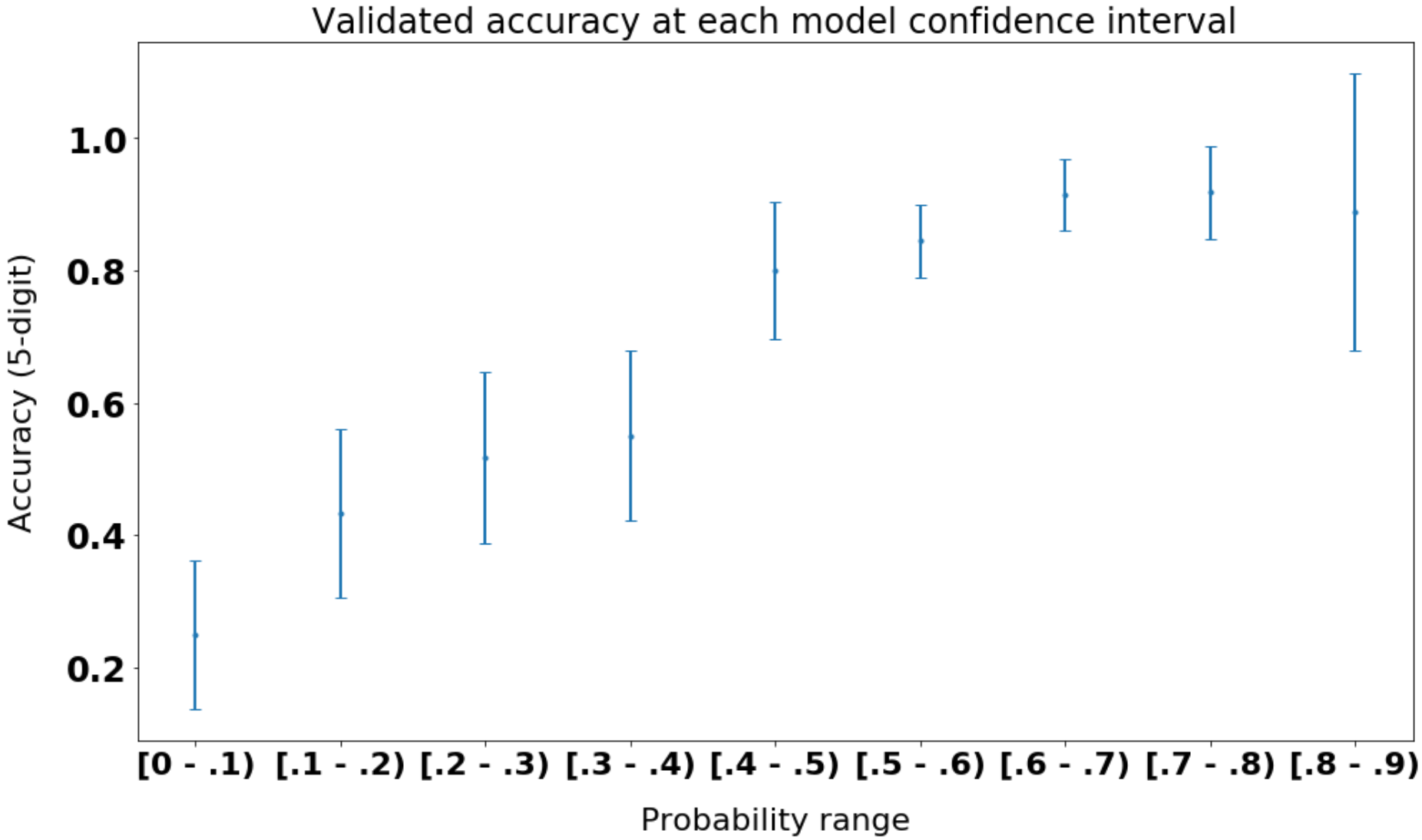



*Figure: validation accuracy for each model probability / confidence range. Bars are 95% Bernoulli CI*

# Applications – Batch Classifier

**Classify Commodities**

Choose File No file chosen  Submit

Download

| Description | NAICS | pred_0 | prob_0 | pred_0_desc | pred_1 | prob_1 | pred_2 | prob_2 |
|---|---|---|---|---|---|---|---|---|
| Hand tools, small mechanical appliances for food preparation, and blades for saws | 4223 | 40999 | 0.320 | Other Miscellaneous manufactured products, not elsewhere classified | 33321 | 0.056 | 24229 | 0.014 |
| Cutlery, including cutlery plated with precious metal, razors, scissors, shears, swords, daggers, and similar arms (excludes cutlery of precious metal, and cutlery clad with precious metal, see 40942) | 4223 | 33999 | 0.224 | Other Articles of non-precious metal, not elsewhere classified (except backed or printed foil, see 324xx, and musical instruments, see 40992) | 40941 | 0.076 | 40999 | 0.048 |
| Interchangeable tools for hand-or machine-tools, including for construction and mining tools | 4223 | 40999 | 0.372 | Other Miscellaneous manufactured products, not elsewhere classified | 33321 | 0.023 | 35999 | 0.016 |
| Locks, mountings and fittings, racks and similar fixtures, and automatic door closers, of base metal | 4219 | 40999 | 0.439 | Other Miscellaneous manufactured products, not elsewhere classified | 33340 | 0.023 | 32300 | 0.013 |
| Other Metal containers with a capacity not exceeding | 4212 | 40999 | 0.165 | Other Miscellaneous manufactured | 33999 | 0.057 | 32499 | 0.027 |

# Applications – Top Words

Browse...

Upload File

Search Text:

Top Numbers Shown: **15**

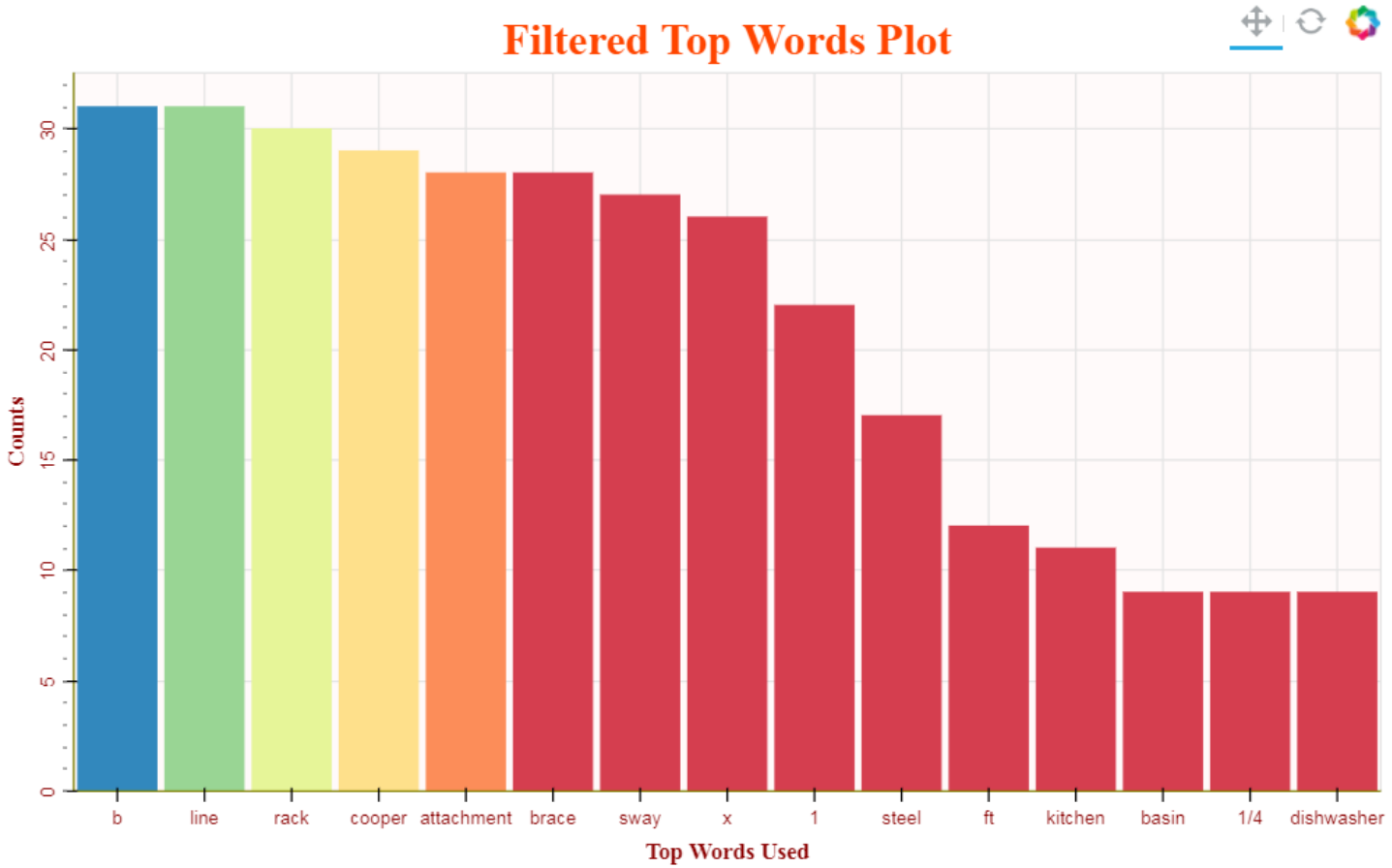
Words-Count | NAICS | Words-TFIDF

SHIP_SCTG

33999

Plot

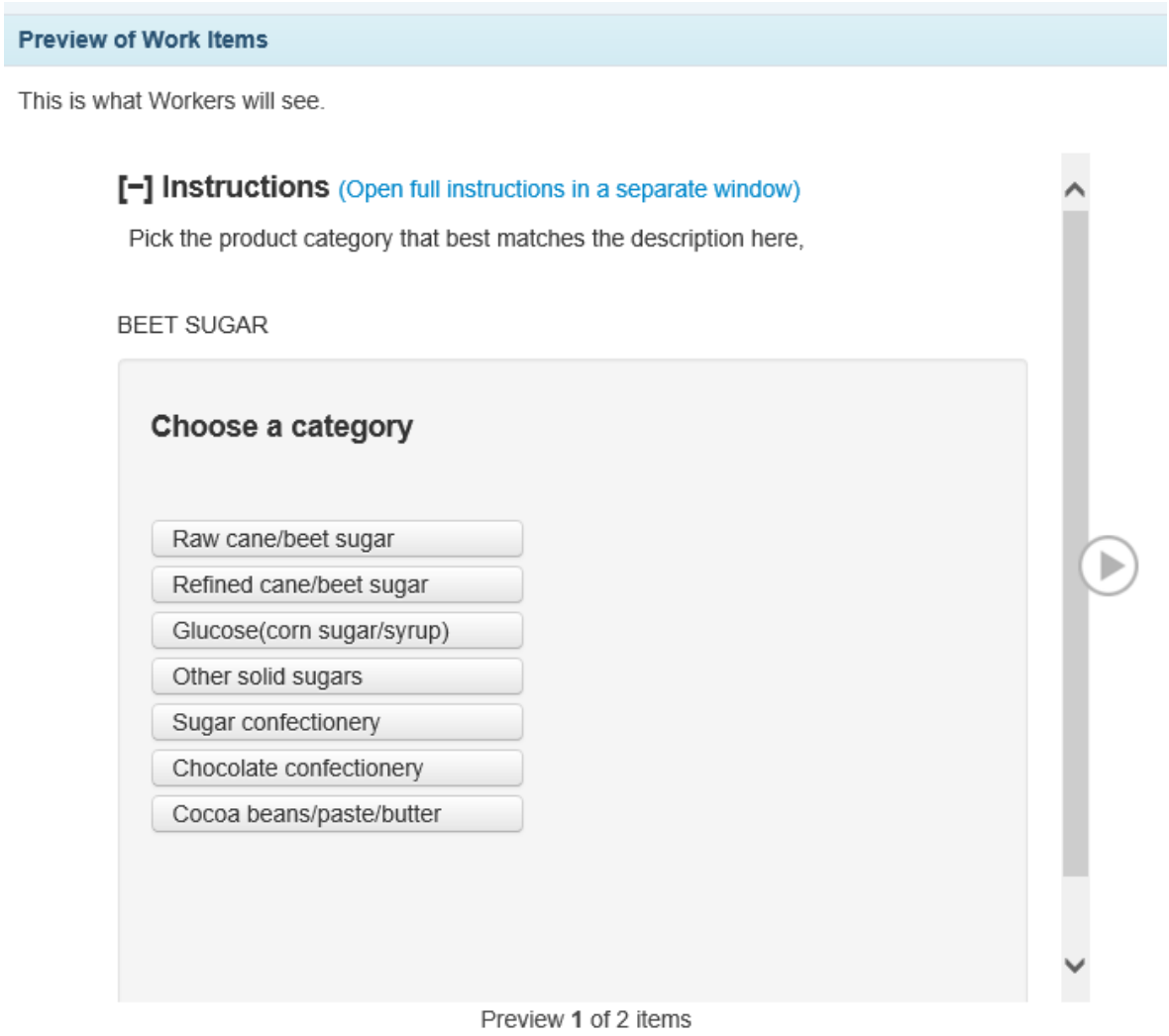
Download




Status: Success

# Future Work

- To move forward, we need cleaner training data
- **Mechanical Turk**
  - Start w/ predictions from current best model
  - Relabel ~15,000 records from NAICS index
  - Turkers choose among top 7-10 predictions from model



# MTurk - Implementation

- **How do we ensure quality?**
- Gateway task
  - Label 50 “gold standard” records
  - Must be at least 60% accurate on min. 5 records
- “Quadruple-key entry”
  - 4 workers label each record
  - Take a vote
  - Total disagreement? This record needs manual investigation.
- Continuous Validation
  - Inter-rater agreement
  - Include more gold standard during actual task

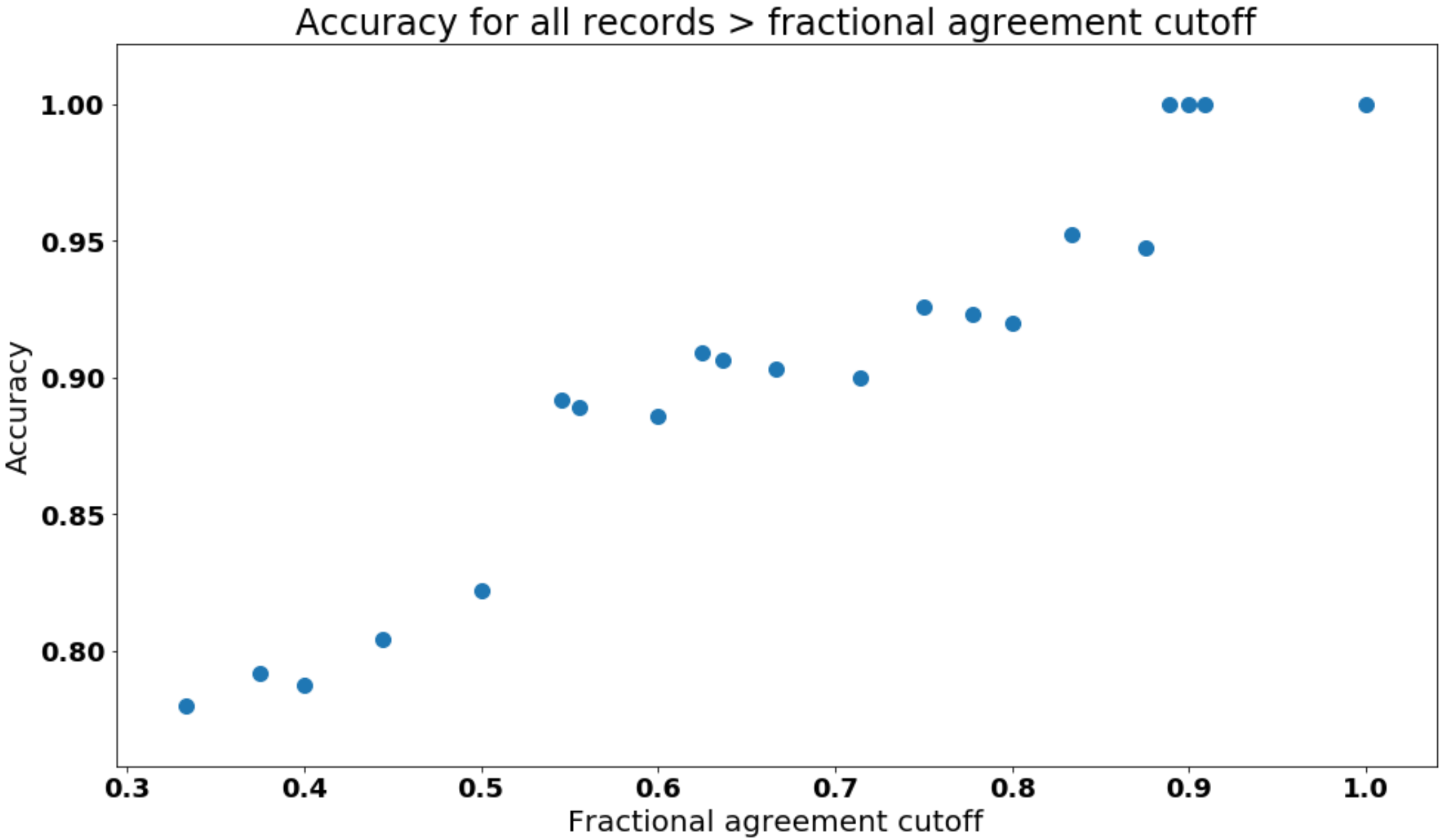



*Figure caption in slide notes.*

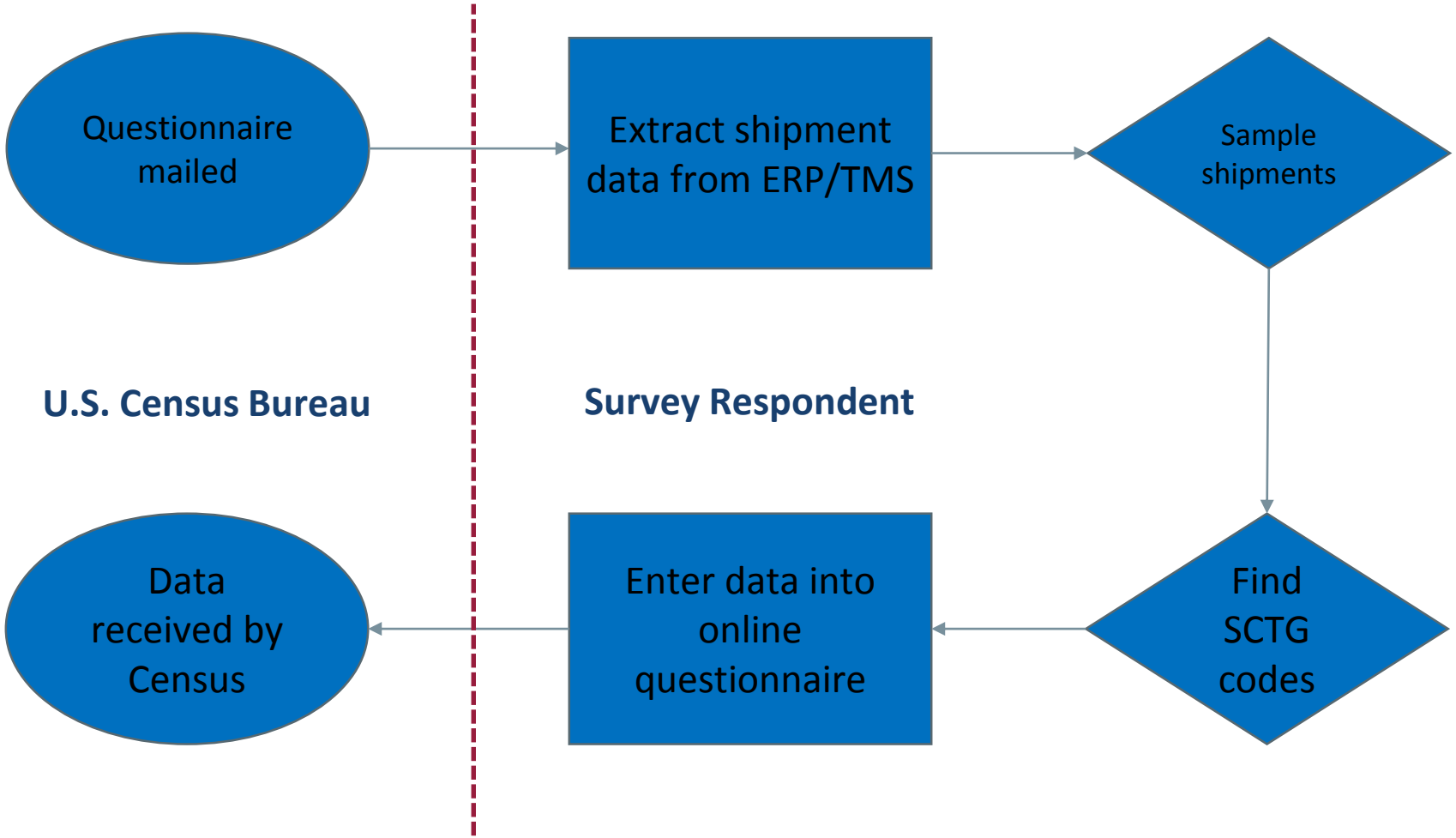
# Future Work – Bigger Picture

Q: How do we reduce suppressions in our current estimates or provide more detail for data users?
A: More data

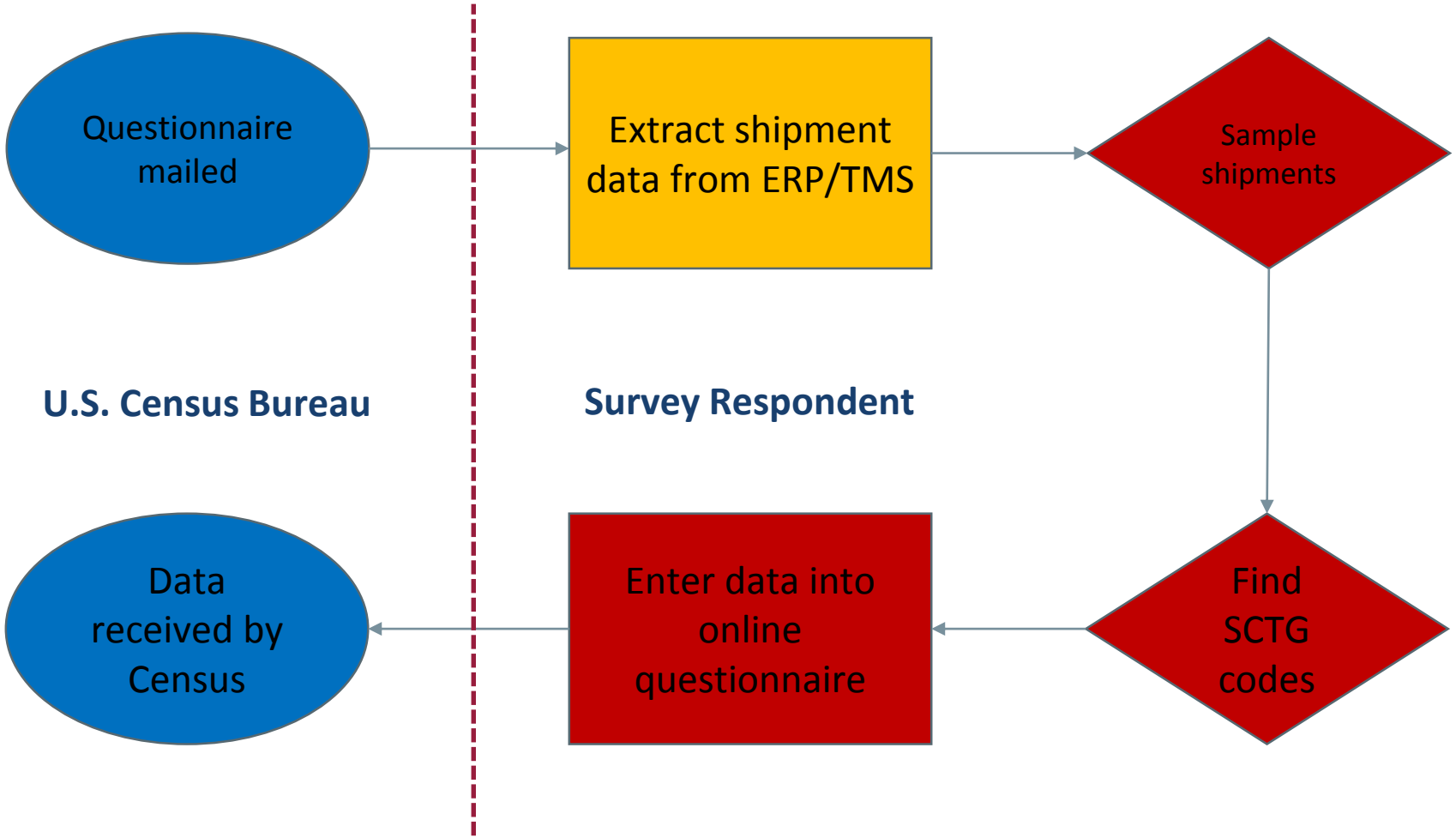
Q: How do we collect data more often than every 5 years?
A: Reduce respondent burden

Q: How do we collect more data AND reduce respondent burden?

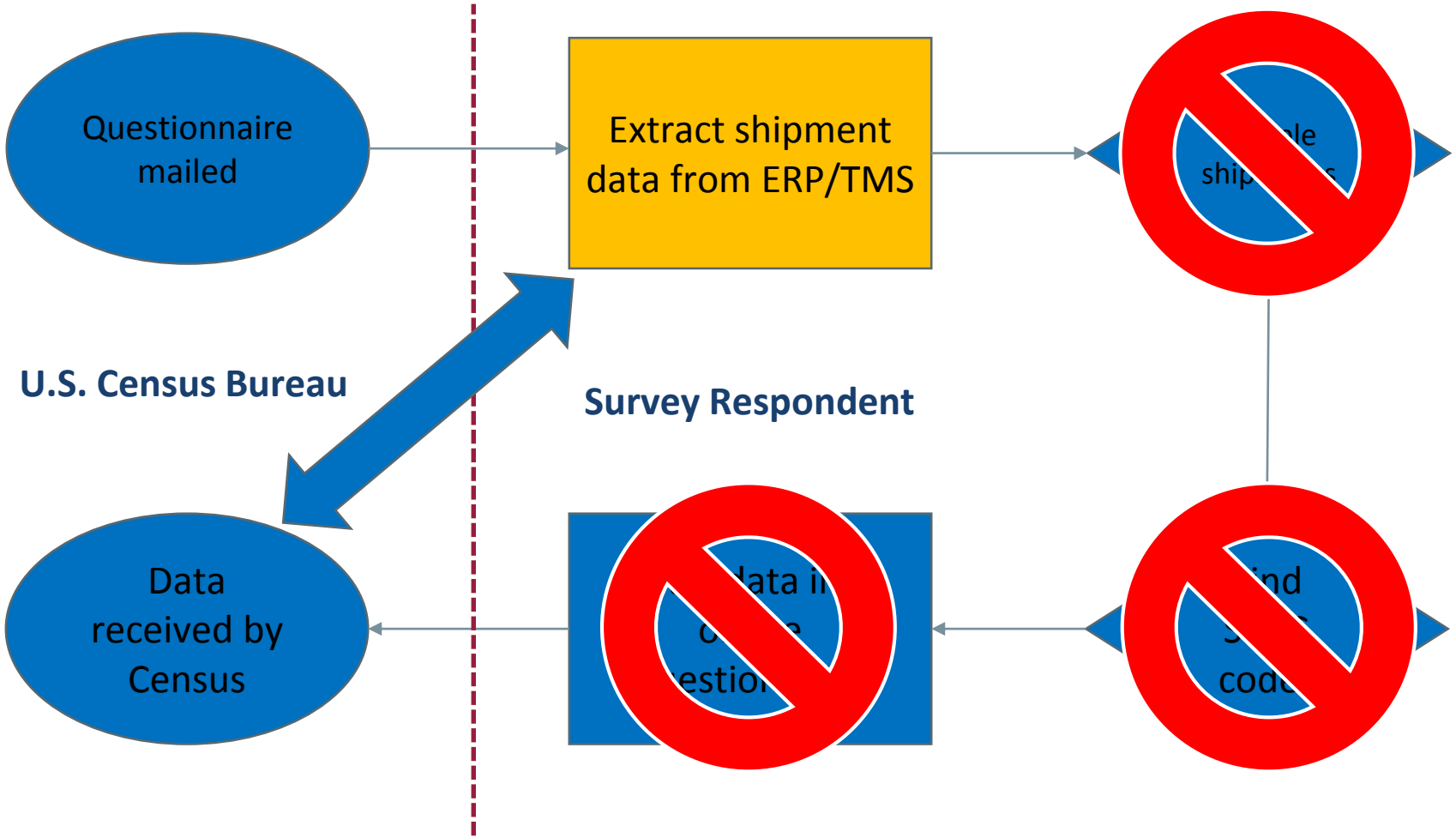
# How the survey process works



# Respondent Burden



# How do we collect more data AND reduce respondent burden?

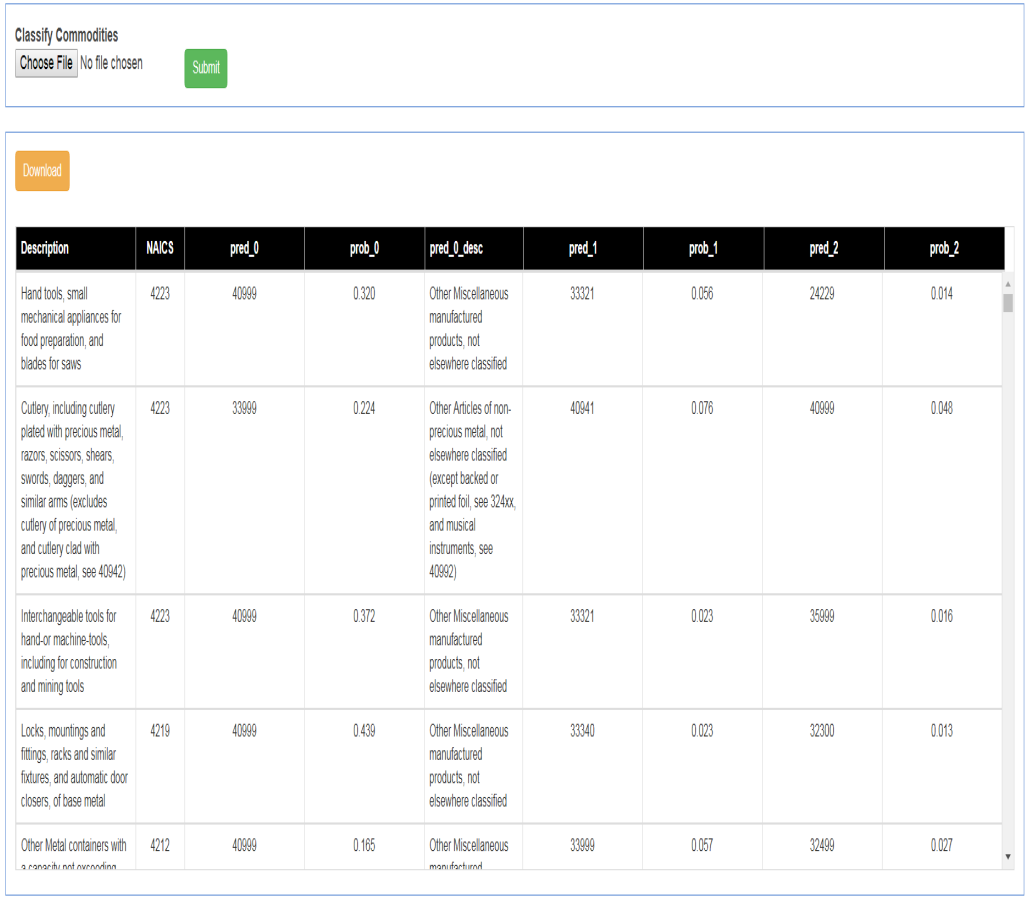


# Automated transfers of shipment records from shippers to Census can provide

- More data
- More often
- Higher quality data
- Less respondent burden
- For less money?

# We want your input!

- SCTG search/classify tool
- What would you care about the most?
  - E.g. More Timely?
  - E.g. More granular geography?
- Spring 2020 CFS Workshop




| Description | NAICS | pred_0 | prob_0 | pred_0_desc | pred_1 | prob_1 | pred_2 | prob_2 |
|---|---|---|---|---|---|---|---|---|
| Hand tools, small mechanical appliances for food preparation, and blades for saws | 4223 | 40999 | 0.320 | Other Miscellaneous manufactured products, not elsewhere classified | 33321 | 0.056 | 24229 | 0.014 |
| Cutlery, including cutlery plated with precious metal, razors, scissors, shears, swords, daggers, and similar arms (excludes cutlery of precious metal, and cutlery clad with precious metal, see 40942) | 4223 | 33999 | 0.224 | Other Articles of non-precious metal, not elsewhere classified (except backed or printed foil, see 324xx, and musical instruments, see 40992) | 40941 | 0.076 | 40999 | 0.048 |
| Interchangeable tools for hand-or machine-tools, including for construction and mining tools | 4223 | 40999 | 0.372 | Other Miscellaneous manufactured products, not elsewhere classified | 33321 | 0.023 | 35999 | 0.016 |
| Locks, mountings and fittings, racks and similar fixtures, and automatic door closers, of base metal | 4219 | 40999 | 0.439 | Other Miscellaneous manufactured products, not elsewhere classified | 33340 | 0.023 | 32300 | 0.013 |
| Other Metal containers with a capacity not exceeding | 4212 | 40999 | 0.165 | Other Miscellaneous manufactured | 33999 | 0.057 | 32499 | 0.027 |

# Thank you!

- **Christian**: Christian.L.Moscardi@census.gov