Bias in Commodity Flow Survey ML: A Case Study

Christian Moscardi

United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Disclaimer: Any opinions and conclusions expressed herein are those of the author(s) and do not represent the views of the U.S. Census Bureau or the Bureau of Transportation Statistics. All results have been reviewed to ensure that no confidential information is disclosed.

The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied.

(Approval ID: CBDRB-FY20-ESMD002-010)

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

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

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?

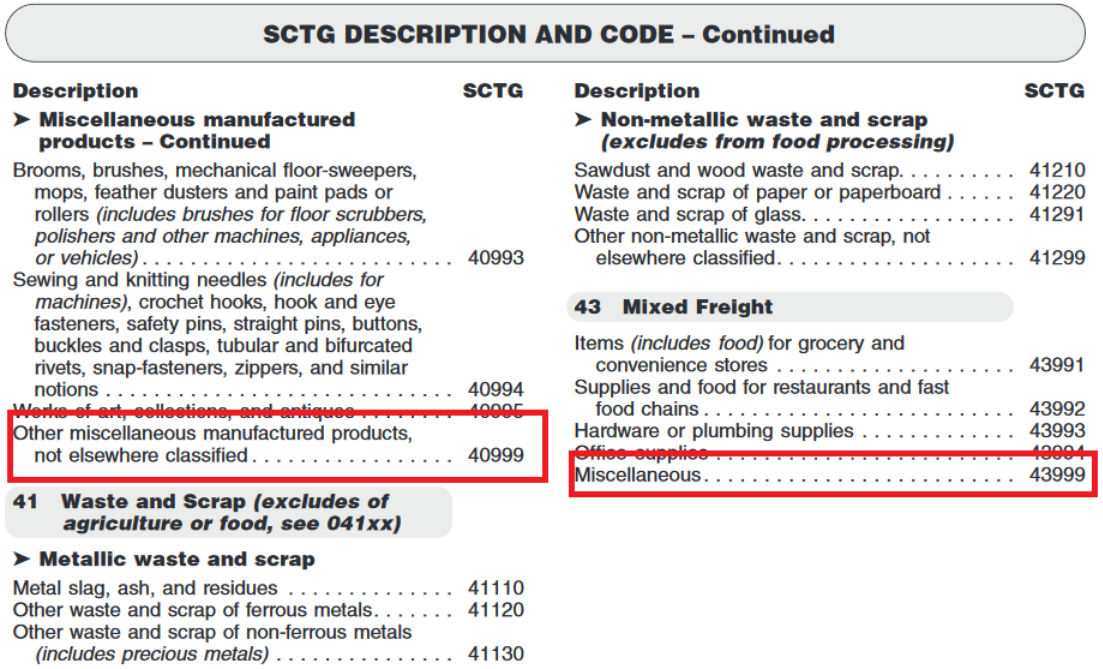
(Yes.)

What are the potential impacts of bias?

- **Impact on product**: increased or decreased estimates of freight activity by product type
  - Survey informs transportation planning decisions & resource allocation
  - Could lead to increased/decreased spending on various types of infrastructure
  - Bias potential: State/local DOTs could misunderstand the types of products being brought into / exported from their communities, misallocate resources
- **Impact on process:** reduced time spent assigning product codes & correcting data
  - We see limited risk given this activity strictly reduces respondent + analyst workload

Initial Investigation

- Data
  1. Labelled Records (6.4M) from 2017 CFS
  2. For training, de-duplicating leads to ~500,000 unique descriptions
- Miscellaneous Codes



SCTG DESCRIPTION AND CODE – Continued

Description	SCTG
➤ **Miscellaneous manufactured products – Continued**	
Brooms, brushes, mechanical floor-sweepers, mops, feather dusters and paint pads or rollers *(includes brushes for floor scrubbers, polishers and other machines, appliances, or vehicles)*	40993
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	40994
Works of art, collections, and antiques	40995
Other miscellaneous manufactured products, not elsewhere classified	40999
41 Waste and Scrap *(excludes of agriculture or food, see 041xx)*	
➤ **Metallic waste and scrap**	
Metal slag, ash, and residues	41110
Other waste and scrap of ferrous metals	41120
Other waste and scrap of non-ferrous metals *(includes precious metals)*	41130

Description	SCTG
➤ **Non-metallic waste and scrap *(excludes from food processing)***	
Sawdust and wood waste and scrap	41210
Waste and scrap of paper or paperboard	41220
Waste and scrap of glass	41291
Other non-metallic waste and scrap, not elsewhere classified	41299
43 Mixed Freight	
Items *(includes food)* for grocery and convenience stores	43991
Supplies and food for restaurants and fast food chains	43992
Hardware or plumbing supplies	43993
Office supplies	43994
Miscellaneous	43999

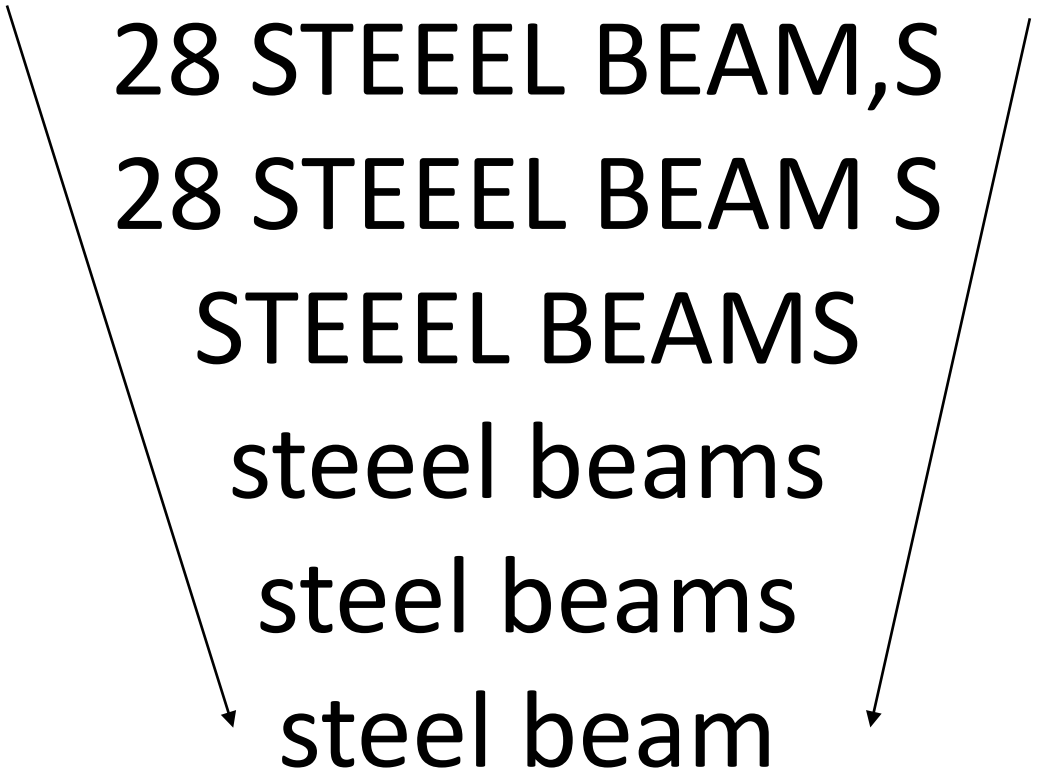
Initial Investigation

- **Solution:** remove 40999, 43999 from training data
- **Lesson learned**: know your data and work with subject experts!

Description	Product Code
ANTENNA M20 - 20L	40999
CAULKING TOOLS	40999
PLASTIC CONTINAER	43999
CLEANING FLUID 2-OZ BOTTLE	40999
26 GAGUE 5' S/L PIPE	40999
DINING TABLE	40999
MODEL SHIP	43999
PLUS DISPR LIQ-;A COMP 1GAL 48 PLT	40999
SHEAVE, IDLER, ASSEMBLY	40999
CONTROL DEVICE	40999

Initial Model

- Preprocessing
  1. Remove numeric, spell-check, stem
- Feature engineering
  1. “Bag-of-words” + TF-IDF scores
  2. NAICS codes for industry context
- Modelling
  - Logistic Regression, “elastic net” regularization
  - Cross-validate, hold out test set, etc.
- Initial Results: 50% “accuracy”



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?

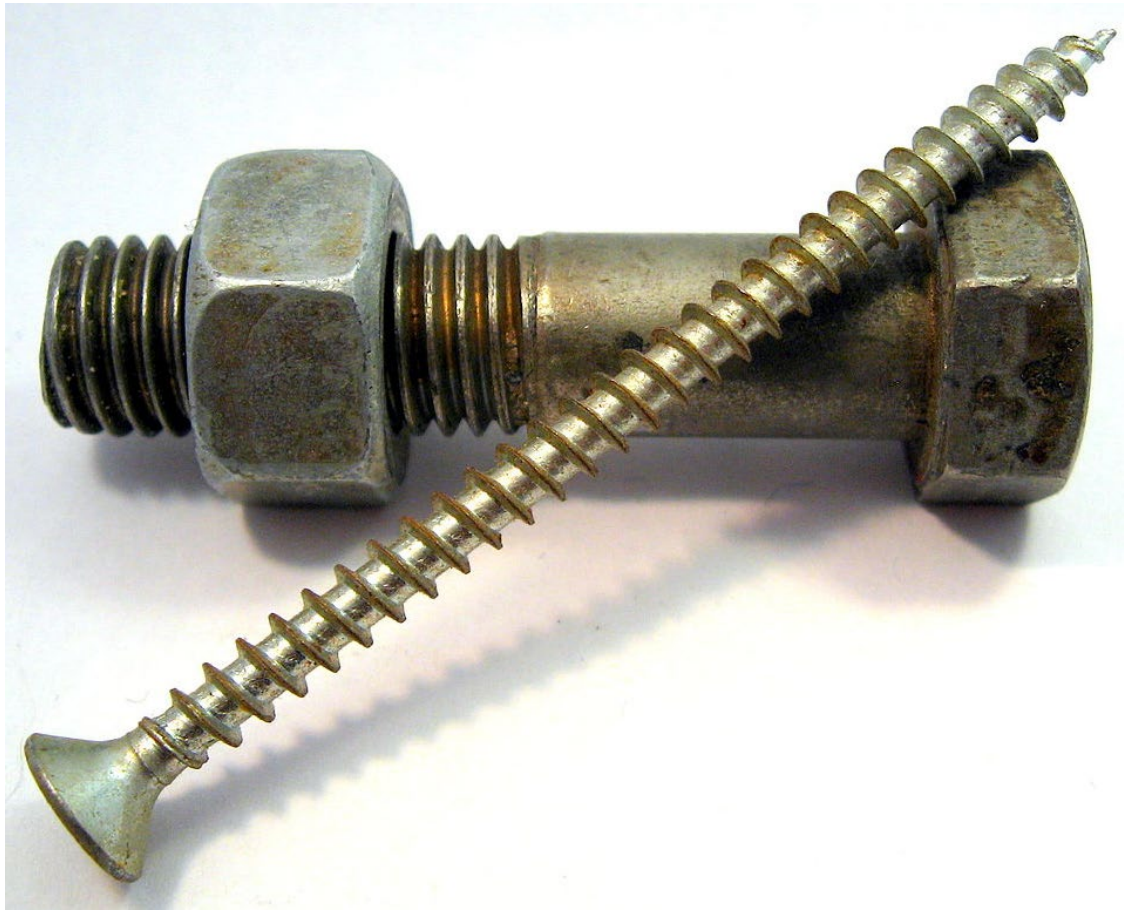


Image courtesy Wikimedia commons

Further Investigation

- Manually validating, about 50% of items labelled 40994 by respondents were miscoded**.**
- **We can see** the workflow which led to these miscodings
- We use the model to help target classification errors in data.

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 w/ analysts
- 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% correct in [.5 -1); 80% in [.4 - .5)
- Can apply this as an imputation strategy for those 100k
- Also ran similar procedure to relabel some of those pesky 40999, 43999 codes

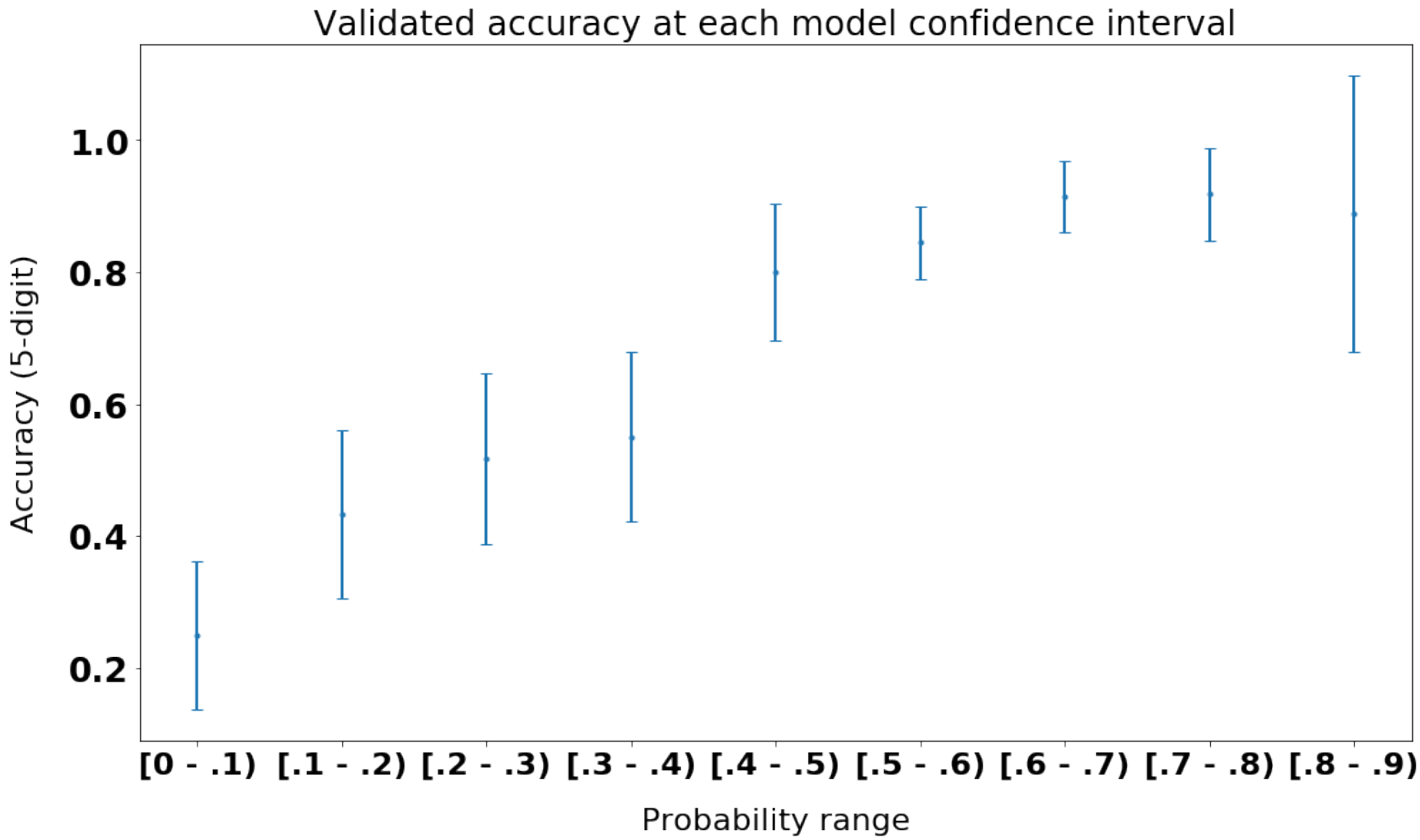
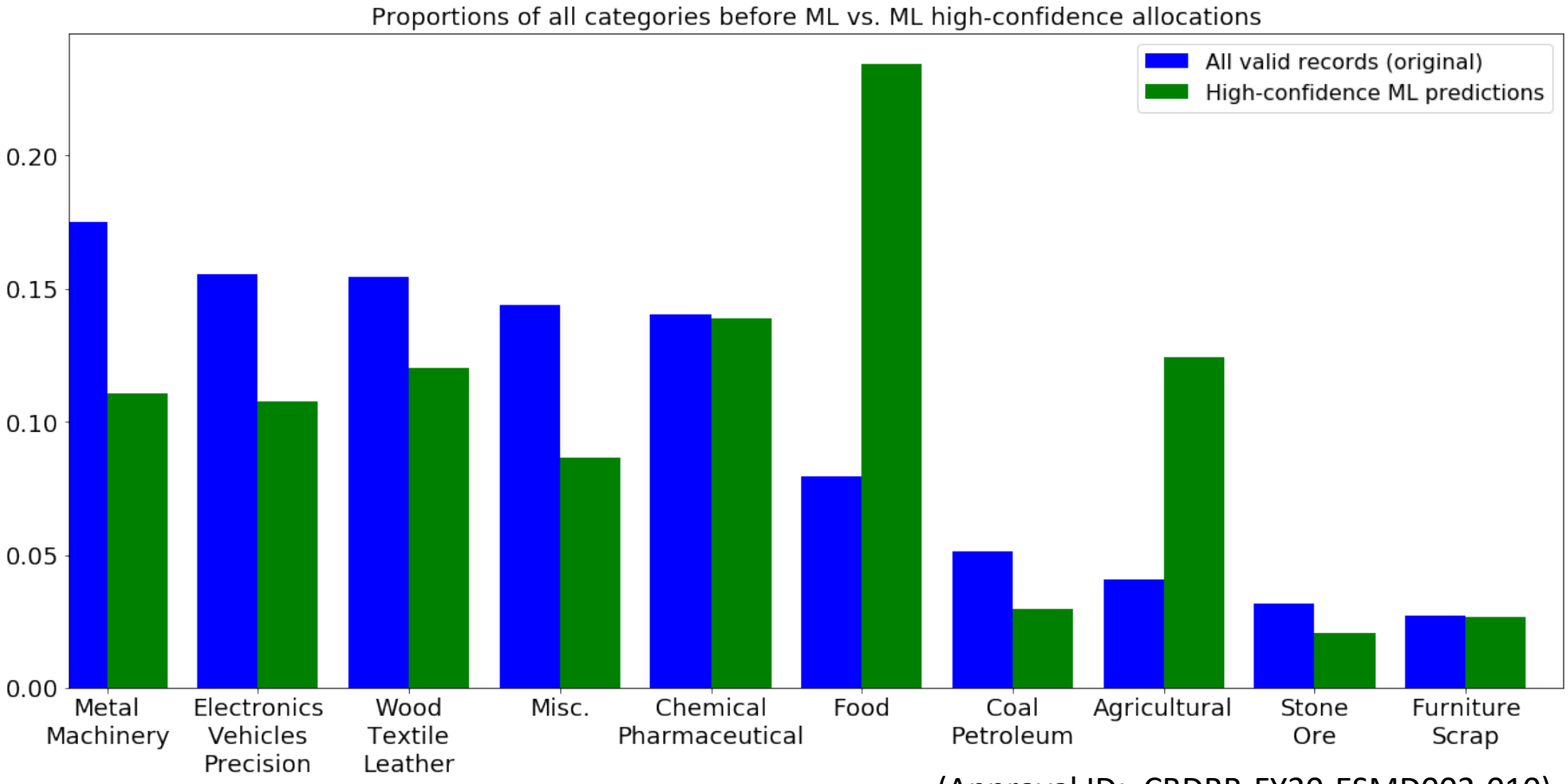


Figure: validation accuracy for each model probability / confidence range. Bars are 95% Bernoulli CI

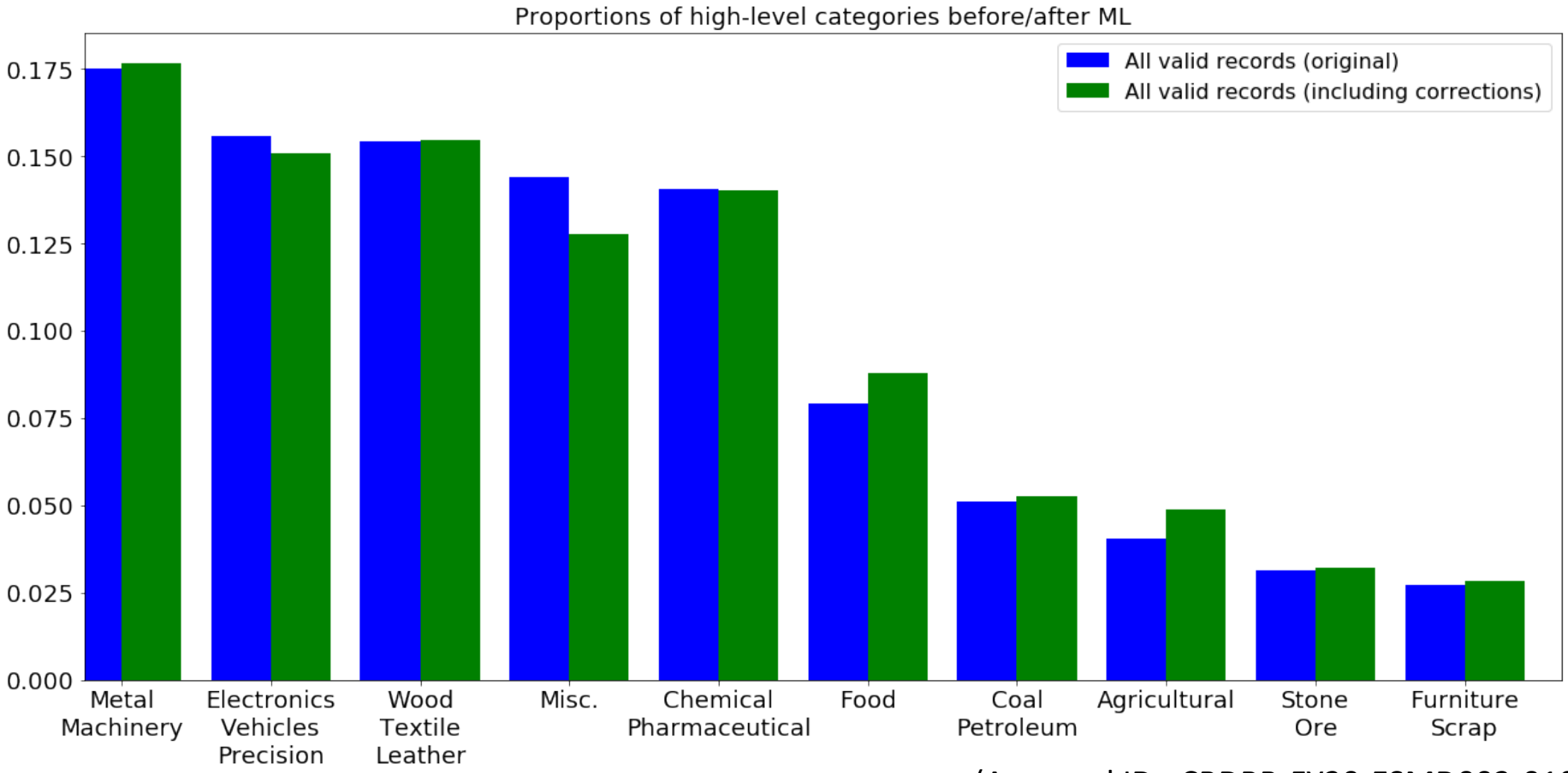
(Approval ID: CBDRB-FY20-ESMD002-010)

What about Bias?



(Approval ID: CBDRB-FY20-ESMD002-010)

What about Bias?



(Approval ID: CBDRB-FY20-ESMD002-010)

What biases affect CFS ML?

- **Response bias (data)**
  - Codebook is burdensome; search tool itself introduced bias by making only certain codes visible; misc. codes
- **Omitted variable bias (model + data)**
  - E.g. in database, product descriptions are limited to 150 chars
  - Writing limits # of words respondents can provide
- **Automation + Confirmation bias (model + data)**
  - Analysts may be predisposed to confirm the model's classification
- **System drift (data + classification scheme)**
  - Products themselves change over time; new products introduced. Does coding scheme capture this?

Zoom out: what are the potential impacts of autocoding bias?

- Depends on the application
  - Veterans' disability codes
    - **Product:** disability codes assigned to claims
    - **Process:** can the claims of certain demographics be processed faster?
  - ACS occupation codes
    - **Product:** ACS estimates of occupations by demographics
    - **Process:** 70% of cases are currently human coded

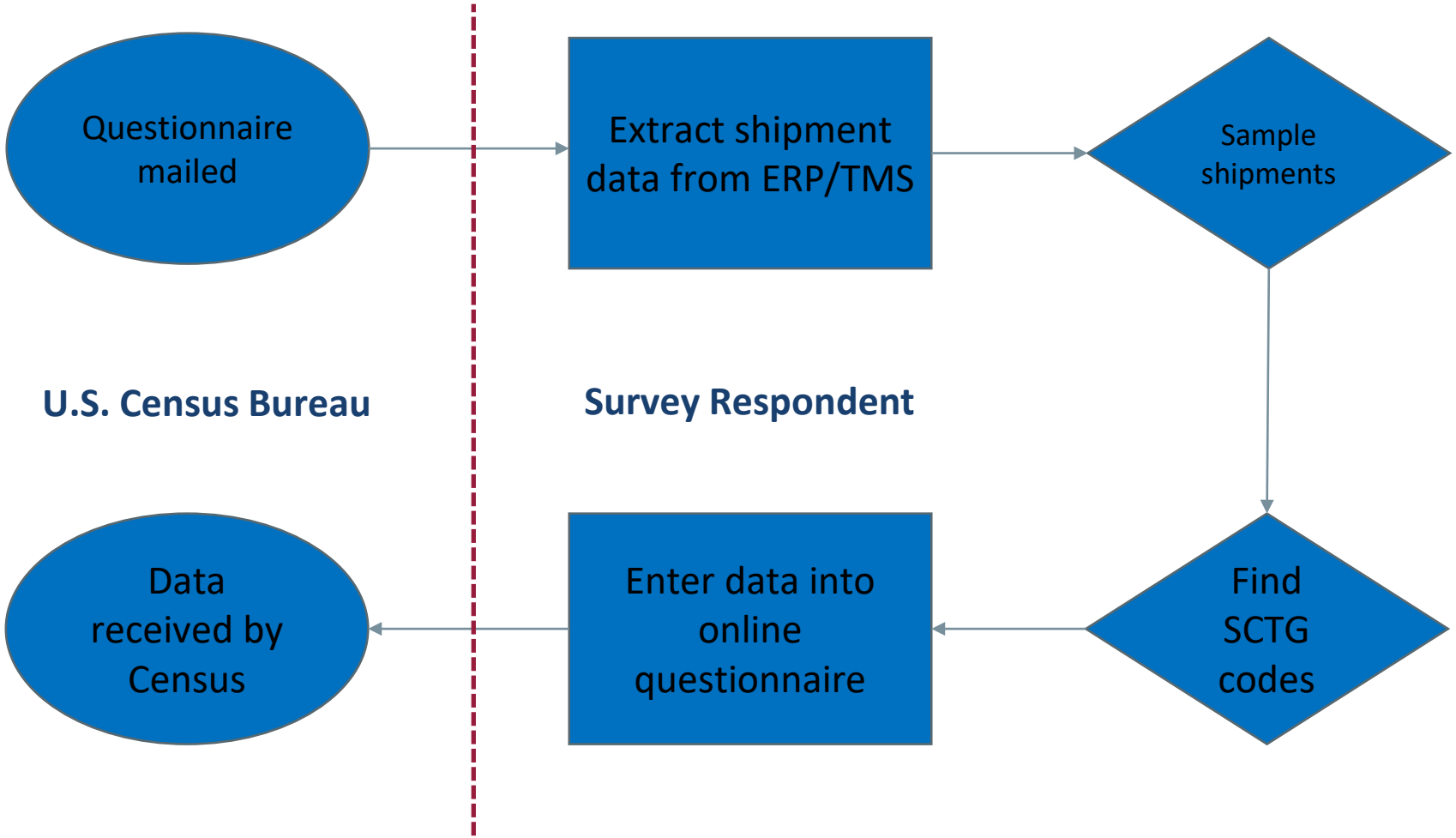
Thank you!

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

References

1. https://towardsdatascience.com/understanding-and-reducing-bias-in-machine-learning-6565e23900ac

Biggest issue: Response Bias



Biggest issue: Response Bias