Automated Survey Coding at the Bureau of Labor Statistics

David Oh & Brandon Kopp
Data Scientist
Bureau of Labor Statistics

JPSM Data Collection Class Presentation
April 7th, 2025

About Us

David Oh

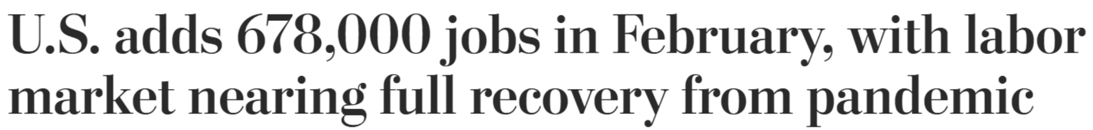
- Supervisory Data Scientist in the Office of Compensation and Working Conditions
- M.S. in Economics
- Started at BLS in 2017 as an Economist
- Transitioned to data science in January 2020

Brandon Kopp

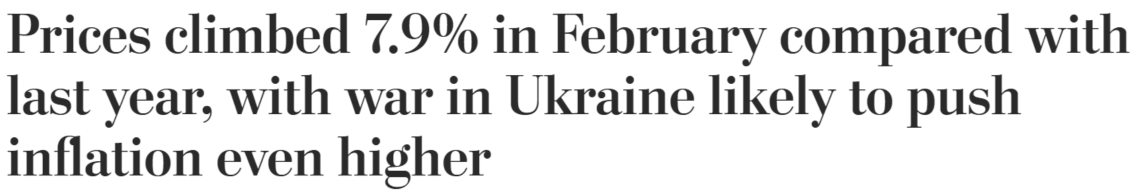
- Director of the Data Science Research Center in the Office of Survey Methods Research
- Ph.D. in Social Psychology
- Started at BLS in 2010 as a Research Psychologist
- Transitioned to data science in February 2022

The U.S. Bureau of Labor Statistics

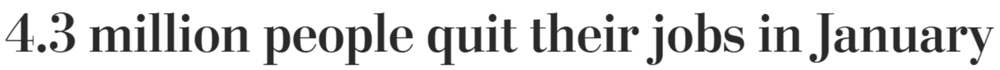
- The Bureau of Labor Statistics is the principal fact-finding agency for the federal government in the broad area of labor economics and statistics.



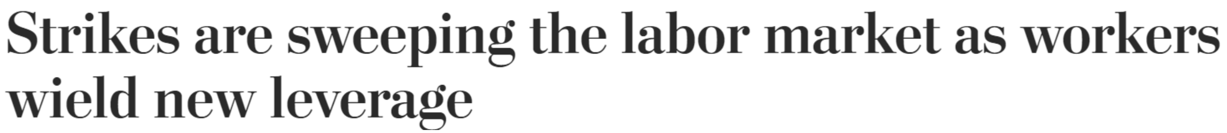
U.S. adds 678,000 jobs in February, with labor market nearing full recovery from pandemic



Prices climbed 7.9% in February compared with last year, with war in Ukraine likely to push inflation even higher



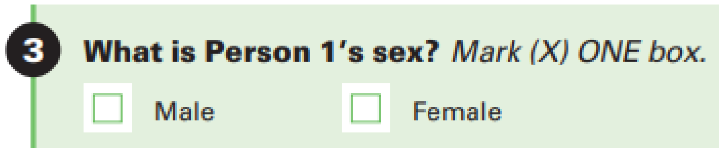
4.3 million people quit their jobs in January



Strikes are sweeping the labor market as workers wield new leverage

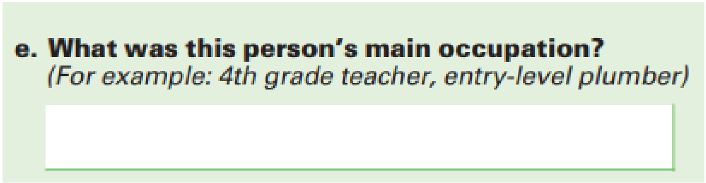
Classification → Aggregation

- What makes statistics meaningful is the ability to make inferences and draw conclusions about circumstances affecting groups or categories
- In many cases, these categorizations are relatively straightforward to apply (demographics, state of residence, etc.) and we have respondents select them themselves.
- In other cases, the range of categorization possibilities is so large, or distinctions between categories so detailed, that making these classifications is burdensome and/or error-prone (occupations, industries, etc.)
- The process of applying these categorizations after the data is collected is called "coding"



3 **What is Person 1's sex?** *Mark (X) ONE box.*

□ Male □ Female



e. What was this person's main occupation?
(For example: 4th grade teacher, entry-level plumber)

Source: 2022 American Community Survey (ACS)

Aggregate Statistics Requiring Classification

Survey of Occupational Injuries and Illnesses

+290.8%

% increase in cases with Days Away from Work between 2019 and 2020 for **Registered Nurses**

Source

Occupational Employment and Wage Statistics

$149,530

Mean annual wage for **Data Scientists** in the **Software Publishing** industry as of May 2023

Source

Occupational Requirements Survey

99.3%

% of **Software Engineers** with a sedentary strength level required for job in 2024

Source

Consumer Price Index

-3.1%

% increase in **gasoline (all types)** index over the 12-month period ending in February 2025

Source

Quarterly Census of Employment and Wages

1,487

Total employment in the **Marketing Research and Public Opinion Polling** industry in DC in Sept 2024

Source

American Time Use Survey

5.15

Average number of hours per day spent on **Leisure and Sports Activities (Including Travel)** in 2023

Source

Classification Systems at BLS

System	Abbv	Focus/Subject	Surveys Using System	Link
Standard Occupational Classification	SOC	• Occupations	• Occupational Requirements Survey • Survey of Occupational Injuries and Illnesses • Census of Fatal Occupational Injuries • Occupational Employment and Wage Statistics • Employment Projections	Link
North American Industry Classification System	NAICS	• Industries	• Quarterly Census of Employment and Wages • Current Employment Statistics • Job Openings and Labor Turnover Survey	Link
Occupational Injuries and Illness Classification System	OIICS	• Event Types • Body Parts • Injury Causes • Injury Type	• Survey of Occupational Injuries and Illnesses • Census of Fatal Occupational Injuries	Link
Universal Classification Codes	UCC	• Products • Services	• Consumer Expenditure Survey	Link
Entry Level Item	ELI	• Products • Services	• Consumer Price Index	Link
Activity Coding Lexicon	--	• Activities	• American Time Use Survey	Link

Exploring Survey Coding Automation

- Across all BLS Surveys, thousands of hours are devoted to manual, human coding of responses every year
- There would be substantial benefits if some or all of this survey coding could be automated

Rules-Based

- If job_title contains “janitor” then “372011”

Dictionary Lookup

match_title	soc_code
custodian	372011
school custodian	372011
office custodian	372011
…	…
bulldozer operator	537199

Machine Learning

- Provide an algorithm with labeled data and let it learn the rules

A Brief History of Automated Coding at BLS

- In 2014, Alex Measure, an economist responsible for reviewing human coding of injury data, began to explore partial automation of coding using machine learning
- That work has advanced to the point that, now, over 90% of codes in the SOII were applied by a neural network “autocoder”
- This work also inspired many other BLS programs to pursue automation for their own coding purposes

Ongoing Automated Coding Projects at BLS

In “Production”

- Survey of Occupational Injuries and Illnesses
  - Occupation & Worker Injuries
- Occupational Employment and Wage Statistics
  - Occupation
- Consumer Price Index
  - Products and Services
- Consumer Expenditure Survey
  - Products and Services

In Development

- Producer Price Index
  - Products and Services
- Quarterly Census of Employment and Wages
  - Industry
- Occupational Requirements Survey
  - Occupation
  - Job Tasks

Business Case for Automation of Survey Coding

- Decrease time spent on manual coding activities
  - Free up human coders to work on higher value work (e.g., gaining cooperation from respondents)
  - Reduce cost associated with manual coding
  - Could allow for more timely release of data products
- Could improve (or at least maintain) accuracy

❖ BLS is still working on effectively measuring these outcomes

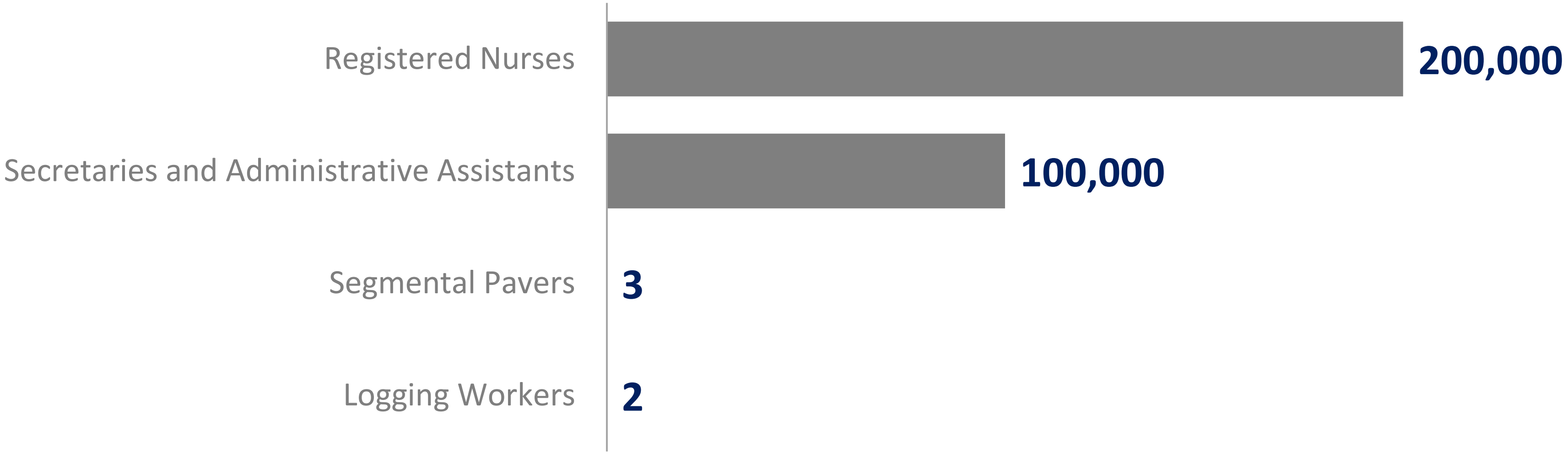
A Unique Combination of Issues

- Massively multi-class classification problem

System	Abbv	#/Type of Categories
Standard Occupational Classification	SOC	**867** Occupations
North American Industry Classification System	NAICS	**1,012** Industries
Occupational Injuries and Illness Classification System	OIICS	**537** Events **615** Natures **218** Body Parts **1,665** Sources
Universal Classification Codes	UCC	**800** Prods & Servs
Entry Level Item	ELI	**298** Prods & Servs
Activity Coding Lexicon	--	**465** Activities

A Unique Combination of Issues

- Massively multi-class classification problem
- Often huge imbalance in distribution across categories



A Unique Combination of Issues

- Massively multi-class classification problem
- Often huge imbalance in distribution across categories
- Classification is mostly based on short, sparse text data (e.g., job titles, description of business functions) that can include misspellings and odd abbreviations
- Sometimes there is not enough information to make a classification

Occupation Classification Examples

Job Title	Possible SOC Classifications	What else might help?
“Registered Nurse”	29-1141 Registered Nurses	
“Nurse”	29-1141 Registered Nurses 29-1151 Nurse Anesthetists 29-1161 Nurse Midwives 29-1171 Nurse Practitioners 29-2061 Licensed Practical and Vocational Nurses 31-1121 Home Health Aides 31-1131 Nursing Assistants	Salary Job duties Industry
“Janitor” “Custodian” “Custodial Worker” “Cleaner” “Window Washer” “Maintenance” “Building Services Worker” … 3,500+ unique job titles	37-2011 Janitors and Cleaners, Except Maids and Housekeeping Cleaners	

A Unique Combination of Issues

- Massively multi-class classification problem
- Often huge imbalance in distribution across categories
- Classification is mostly based on short, sparse text data (e.g., job titles, description of business functions) that can include misspellings and odd abbreviations
- Sometimes there is not enough information to make a classification
- Manual classification often requires in-depth skills and training
- Because classification is difficult, training data often includes incorrect codes

Standard Occupation Classification Structure

Job Title = “Info Tech Specialist 2”

```
-----------------------------------------------------------------------------
15-0000                - Computer and Mathematical Occupations
 |_15-1200             - Computer Occupations
    |_15-1250          - Software and Web Developers, Programmers, and Testers
       |_ 15-1251      - Computer Programmers
       |_ 15-1252      - Software Developers
       |_ 15-1253      - Software Quality Assurance Analysts and Testers
       |_ 15-1254      - Web Developers
       |_ 15-1255      - Web and Digital Interface Designers
-----------------------------------------------------------------------------
```

A Unique Combination of Issues

- Massively multi-class classification problem
- Often huge imbalance in distribution across categories
- Classification is mostly based on short, sparse text data (e.g., job titles, description of business functions) that can include misspellings and odd abbreviations
- Sometimes there is not enough information to make a classification
- Manual classification often requires in-depth skills and training
- Because classification is difficult, training data often includes incorrect codes
- The coding system changes periodically (every 8-10 years for SOC, every 7 years for NAICS)

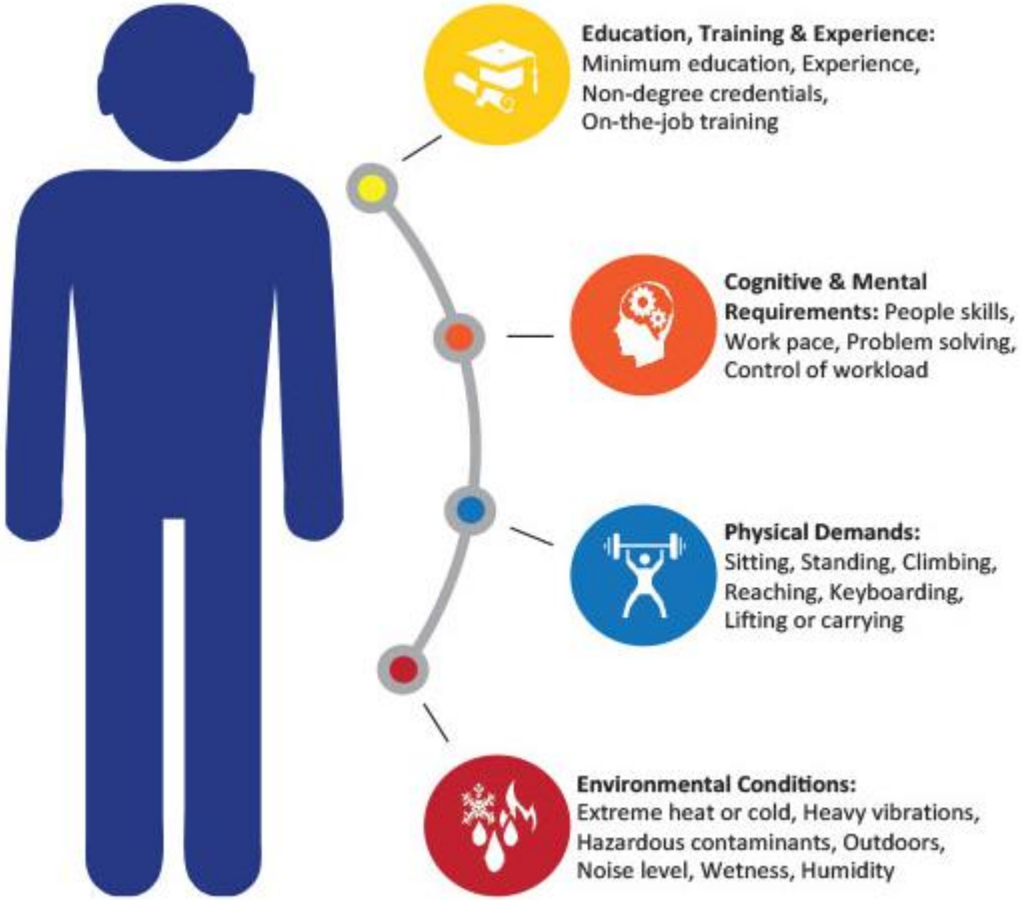
Automated Coding: Applied

Highlighting two survey programs today:

- Occupational Requirements Survey (ORS)
- Survey of Occupational Injuries and Illnesses (SOII)

Occupational Requirements Survey

- Establishment survey
- Collected on behalf of the Social Security Administration (SSA) to support adjudication of its disability programs
- Captures the requirements for a job:
  - Education, Training, and Experience
  - Mental and Cognitive Requirements
  - Physical Demands
  - Environmental Conditions
- 17,000+ descriptions of jobs each year



Education, Training & Experience: Minimum education, Experience, Non-degree credentials, On-the-job training

Cognitive & Mental Requirements: People skills, Work pace, Problem solving, Control of workload

Physical Demands: Sitting, Standing, Climbing, Reaching, Keyboarding, Lifting or carrying

Environmental Conditions: Extreme heat or cold, Heavy vibrations, Hazardous contaminants, Outdoors, Noise level, Wetness, Humidity

Occupational Requirements Survey

Example Narrative

Job title:

Hair technician

Critical tasks:

- Shampoos, cuts, colors, blow dries hair
- Recommends styling products
- Perms hair
- Waxes eyebrows and facial hair
- Creates up-dos for special occasions like weddings

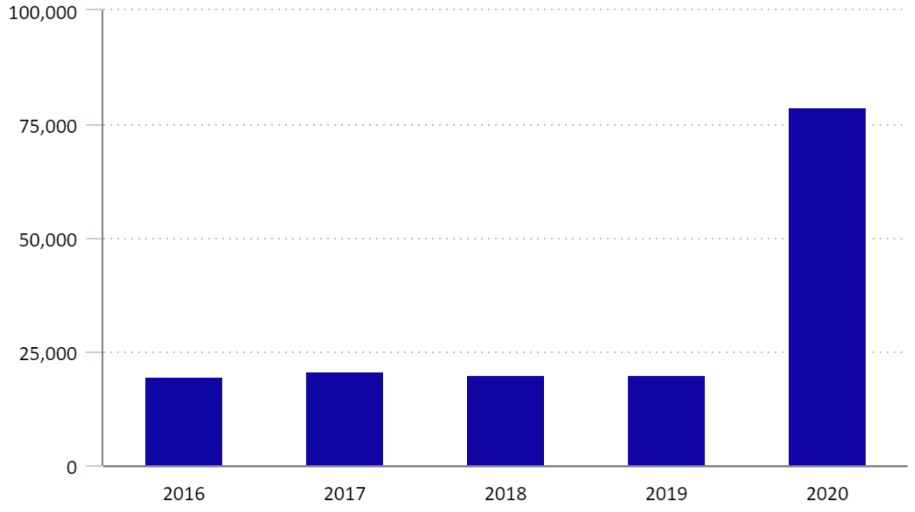
Codes Assigned

Occupation: 39-5012 (Hairdressers)

Survey of Occupational Injuries and Illnesses

- Annual establishment survey collecting injury and illness information since 1972
- Information collected:
  - Total number of cases resulting in days away from work or days of job transfer and work restrictions
  - Detailed case and demographic information about some injury or illness cases
- 200,000+ descriptions of work-related injuries and illnesses each year

Chart 1. Number of nonfatal occupational injury and illness cases involving days away from work, registered nurses, private industry, 2016–2020



Hover over chart to view data.
Source: U.S. Bureau of Labor Statistics, Survey of Occupational Injuries and Illnesses

Survey of Occupational Injuries and Illnesses

Example Narrative

Job title: sanitation worker

What was the employee doing just before the incident?
mopping floor in gym

What happened?
slipped on wet floor and fell

What part of the body was affected?
fractured right arm

What object directly harmed the employee?
wet floor

Codes Assigned

Occupation: 37-2011 (Janitor)

Nature: 111 (Fracture)

Part: 420 (Arm)

Event: 422 (Fall, slipping)

Source: 6620 (Floor)

Techniques for Coding Open-ended Responses

SOII

ORS

Manual Coding

Computer-assisted Coding

Automated Coding

Manual Coding

- Data coders assign classification codes after examining and interpreting the unstructured inputs, like open-ended responses
- Time-intensive
- Resource-intensive
- Highly variable
    - Different people interpret the same thing different ways

Techniques for Coding Open-ended Responses

SOII

ORS

Manual Coding

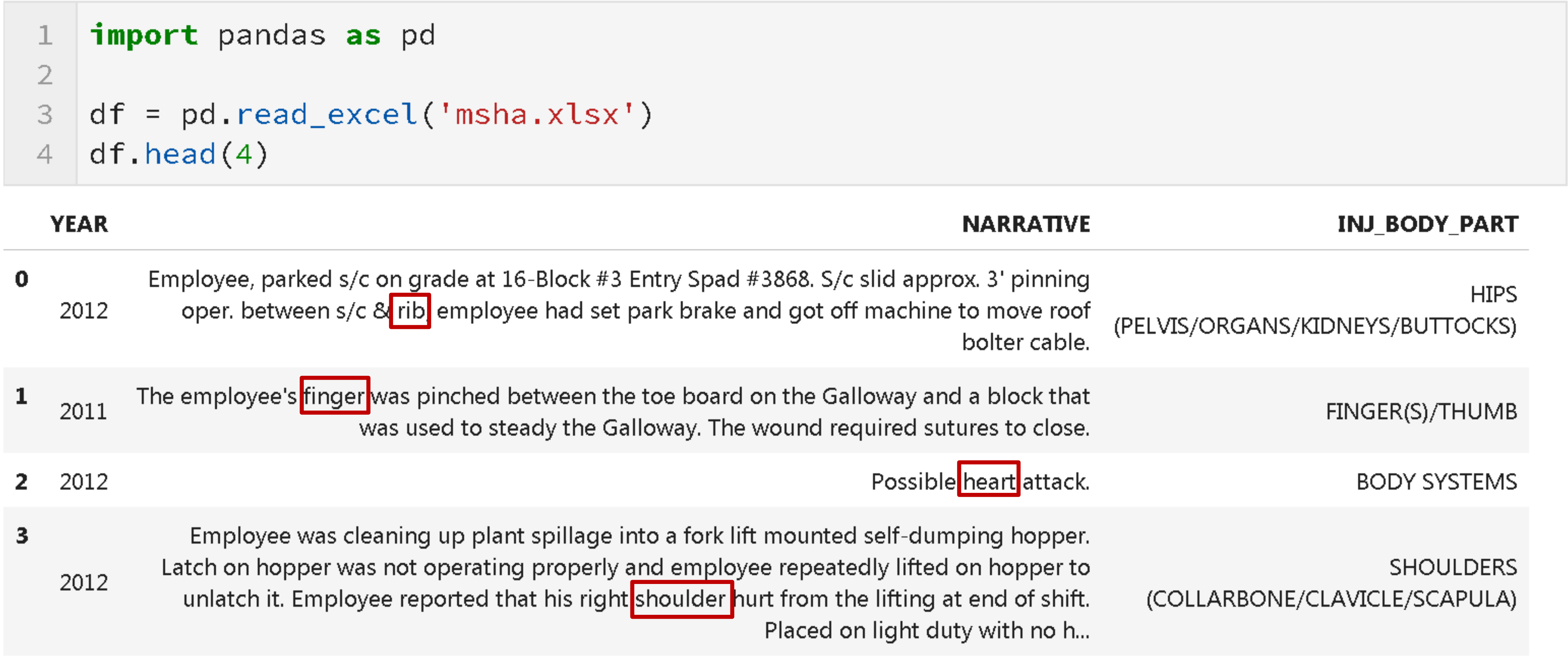
Computer-assisted Coding

Automated Coding

Computer-assisted Coding

- Data coders assign classification codes after examining and interpreting the unstructured inputs and consulting some form of feedback from a computer system
- Why machine learning?
    - Rule-based system or lookup dictionaries are impractical given the problem we are trying to solve
    - Already had large sets of manually coded historical data, making supervised machine learning easily implementable
    - Basis for most “AI” today

ML Example: Gather the Data

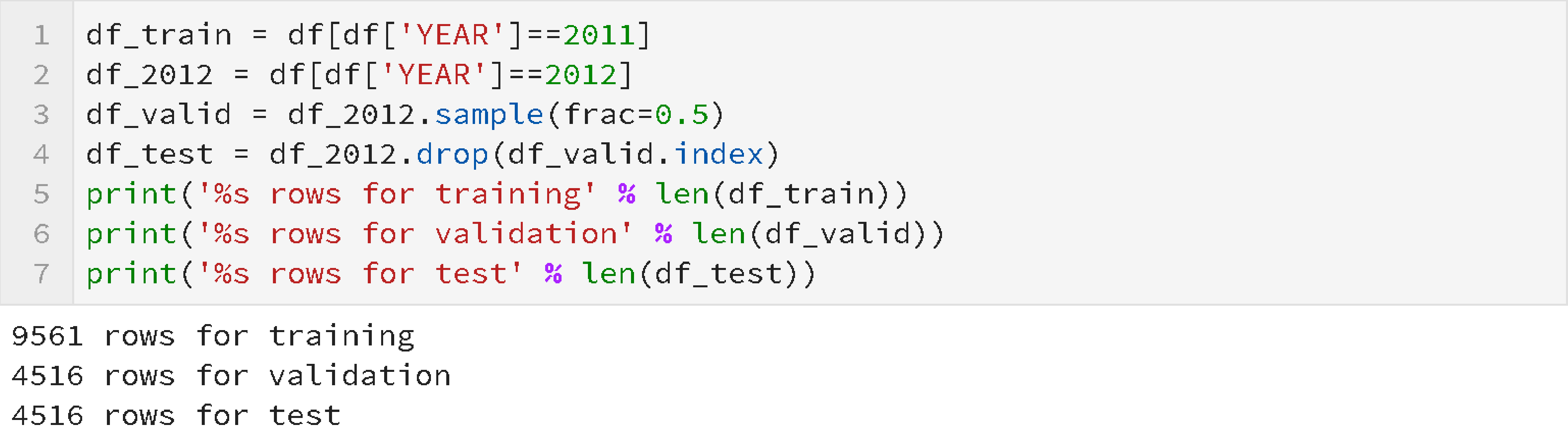


```python
import pandas as pd

df = pd.read_excel('msha.xlsx')
df.head(4)
```

	YEAR	NARRATIVE	INJ_BODY_PART
0	2012	Employee, parked s/c on grade at 16-Block #3 Entry Spad #3868. S/c slid approx. 3' pinning oper. between s/c & rib, employee had set park brake and got off machine to move roof bolter cable.	HIPS (PELVIS/ORGANS/KIDNEYS/BUTTOCKS)
1	2011	The employee's finger was pinched between the toe board on the Galloway and a block that was used to steady the Galloway. The wound required sutures to close.	FINGER(S)/THUMB
2	2012	Possible heart attack.	BODY SYSTEMS
3	2012	Employee was cleaning up plant spillage into a fork lift mounted self-dumping hopper. Latch on hopper was not operating properly and employee repeatedly lifted on hopper to unlatch it. Employee reported that his right shoulder hurt from the lifting at end of shift. Placed on light duty with no h...	SHOULDERS (COLLARBONE/CLAVICLE/SCAPULA)

ML Example: Separate Training, Validation and Test



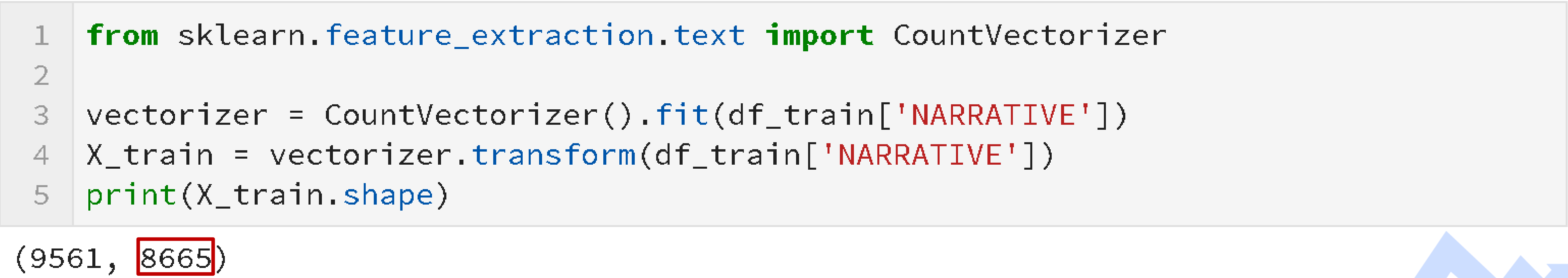
```python
df_train = df[df['YEAR']==2011]
df_2012 = df[df['YEAR']==2012]
df_valid = df_2012.sample(frac=0.5)
df_test = df_2012.drop(df_valid.index)
print('%s rows for training' % len(df_train))
print('%s rows for validation' % len(df_valid))
print('%s rows for test' % len(df_test))
```

```
9561 rows for training
4516 rows for validation
4516 rows for test
```

ML Example: Convert the Inputs

Bag of Words Representation

		x_1	x_2	x_3	x_4	x_5
Narrative		employee	finger	leg	cuts	head
employee injured finger		1	1	0	0	0
employee broke leg	→	1	0	1	0	0
cuts to finger leg and head		0	1	1	1	1



```python
from sklearn.feature_extraction.text import CountVectorizer

vectorizer = CountVectorizer().fit(df_train['NARRATIVE'])
X_train = vectorizer.transform(df_train['NARRATIVE'])
print(X_train.shape)
```

```
(9561, 8665)
```

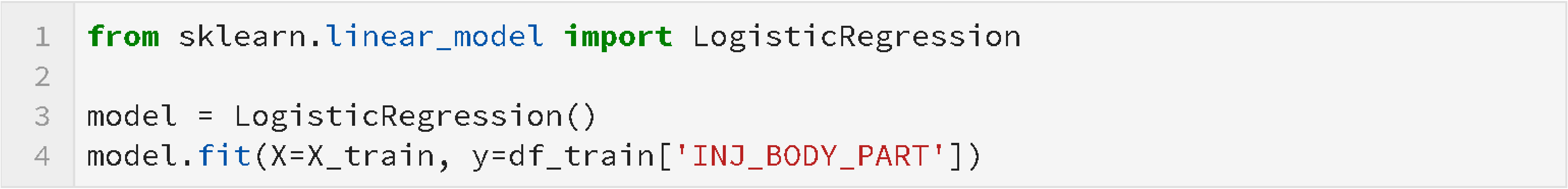
ML Example: Learn from Data

Multinomial Logistic Regression

- $P(code = "hand") = f(w_{h1}x_1 + w_{h2}x_2 + \cdots + b_h)$
- $P(code = "arm") = f(w_{a1}x_1 + w_{a2}x_2 + \cdots + b_a)$
- $P(code = "eye") = f(w_{e1}x_1 + w_{e2}x_2 + \cdots + b_e)$

Where:

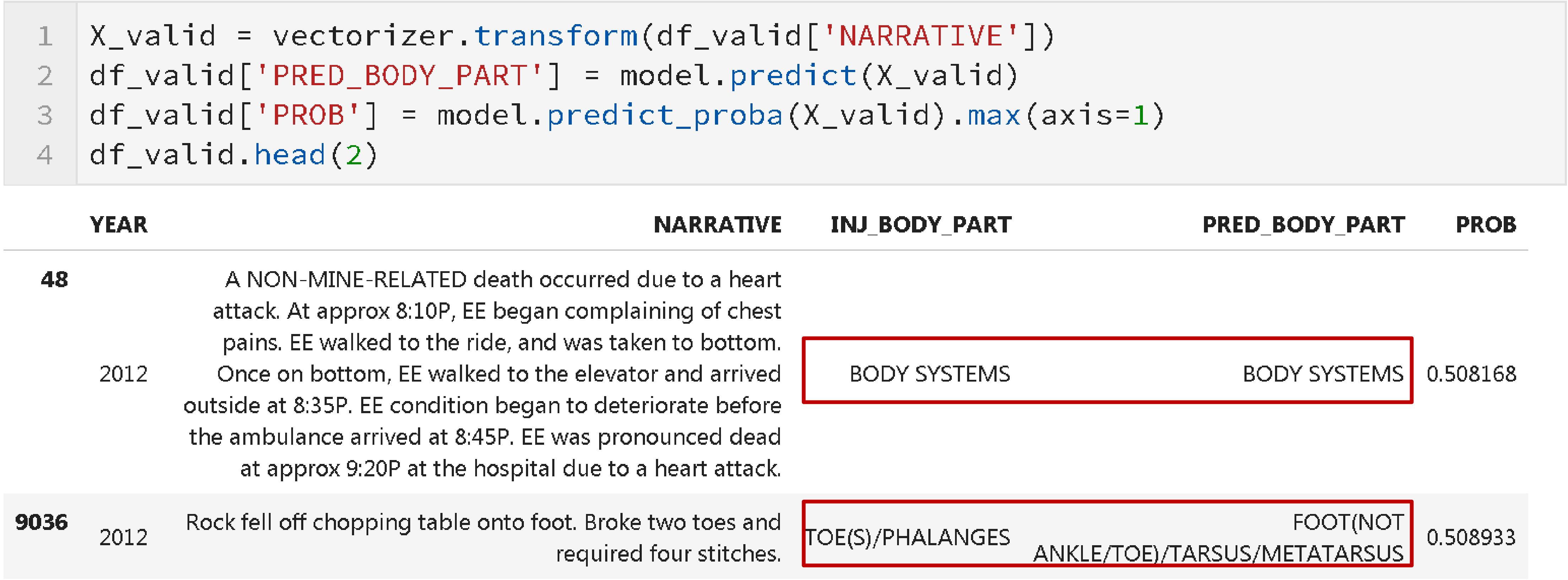
$$f(z) = \text{multinomial logistic function}$$



```python
from sklearn.linear_model import LogisticRegression

model = LogisticRegression()
model.fit(X=X_train, y=df_train['INJ_BODY_PART'])
```

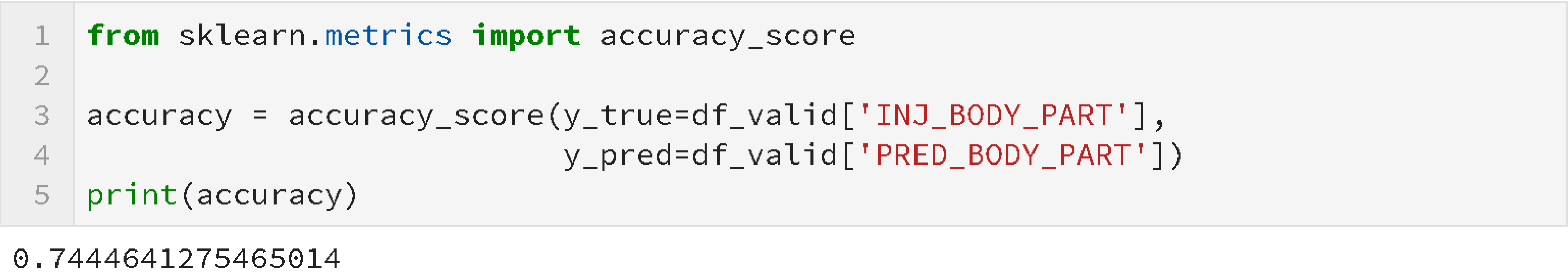
ML Example: Use the Model



```python
X_valid = vectorizer.transform(df_valid['NARRATIVE'])
df_valid['PRED_BODY_PART'] = model.predict(X_valid)
df_valid['PROB'] = model.predict_proba(X_valid).max(axis=1)
df_valid.head(2)
```

	YEAR	NARRATIVE	INJ_BODY_PART	PRED_BODY_PART	PROB
48	2012	A NON-MINE-RELATED death occurred due to a heart attack. At approx 8:10P, EE began complaining of chest pains. EE walked to the ride, and was taken to bottom. Once on bottom, EE walked to the elevator and arrived outside at 8:35P. EE condition began to deteriorate before the ambulance arrived at 8:45P. EE was pronounced dead at approx 9:20P at the hospital due to a heart attack.	BODY SYSTEMS	BODY SYSTEMS	0.508168
9036	2012	Rock fell off chopping table onto foot. Broke two toes and required four stitches.	TOE(S)/PHALANGES	FOOT(NOT ANKLE/TOE)/TARSUS/METATARSUS	0.508933

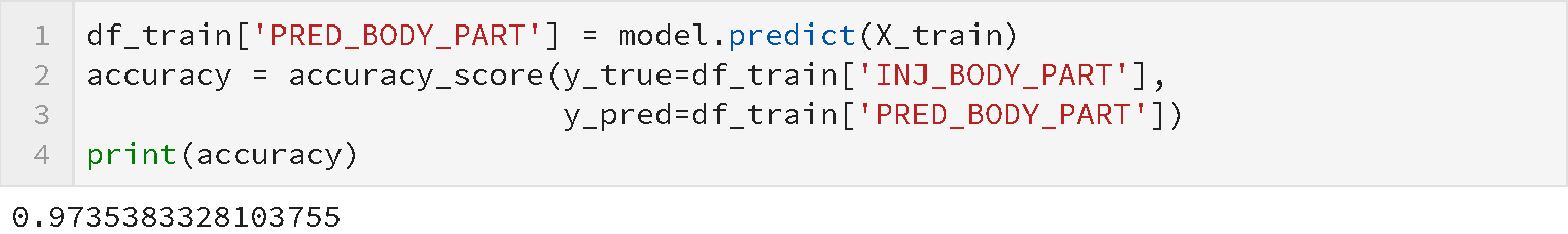
ML Example: Evaluate



```python
from sklearn.metrics import accuracy_score

accuracy = accuracy_score(y_true=df_valid['INJ_BODY_PART'],
                          y_pred=df_valid['PRED_BODY_PART'])
print(accuracy)
```

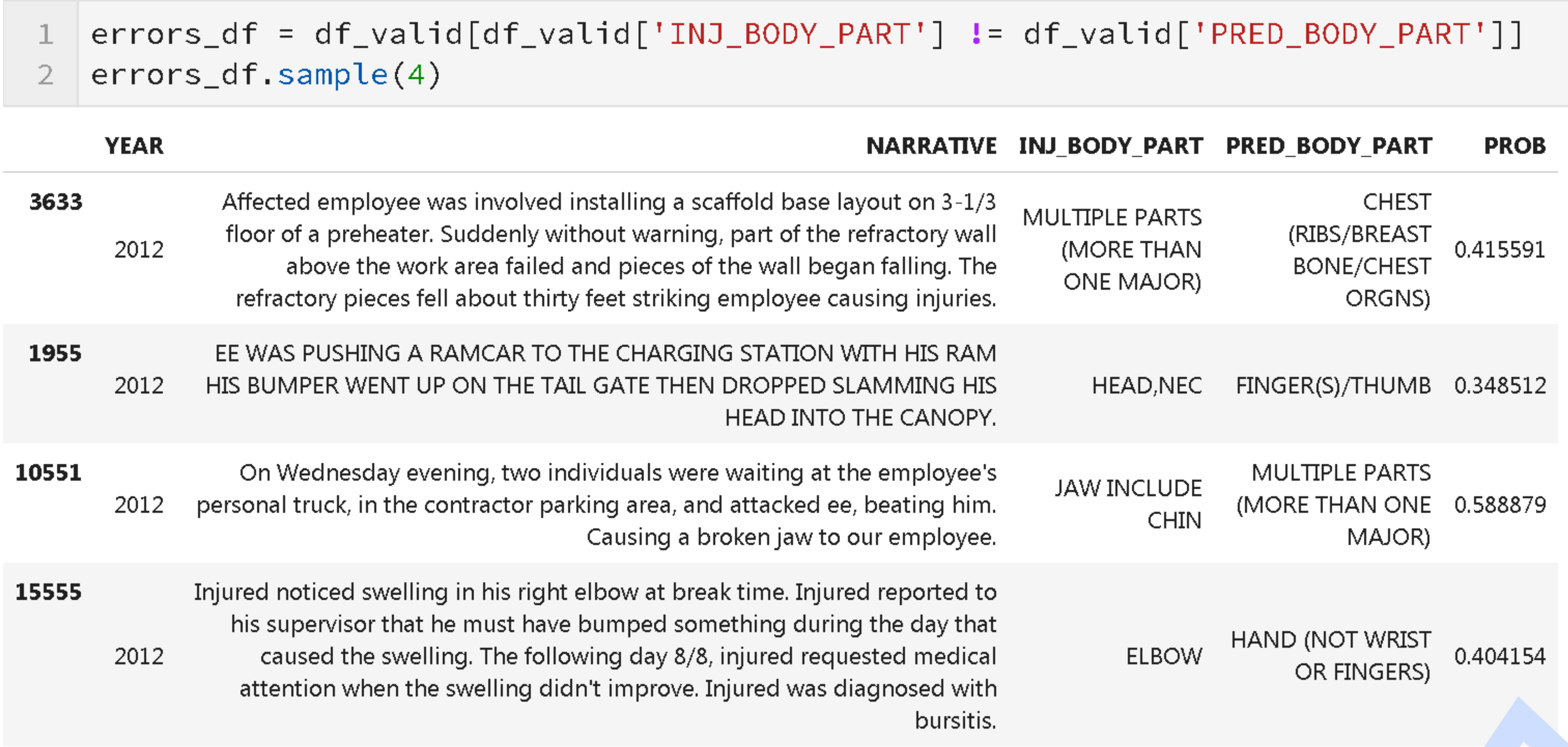
```
0.744641275465014
```



```python
df_train['PRED_BODY_PART'] = model.predict(X_train)
accuracy = accuracy_score(y_true=df_train['INJ_BODY_PART'],
                          y_pred=df_train['PRED_BODY_PART'])
print(accuracy)
```

```
0.9735383328103755
```

ML Example: Examining a Sample

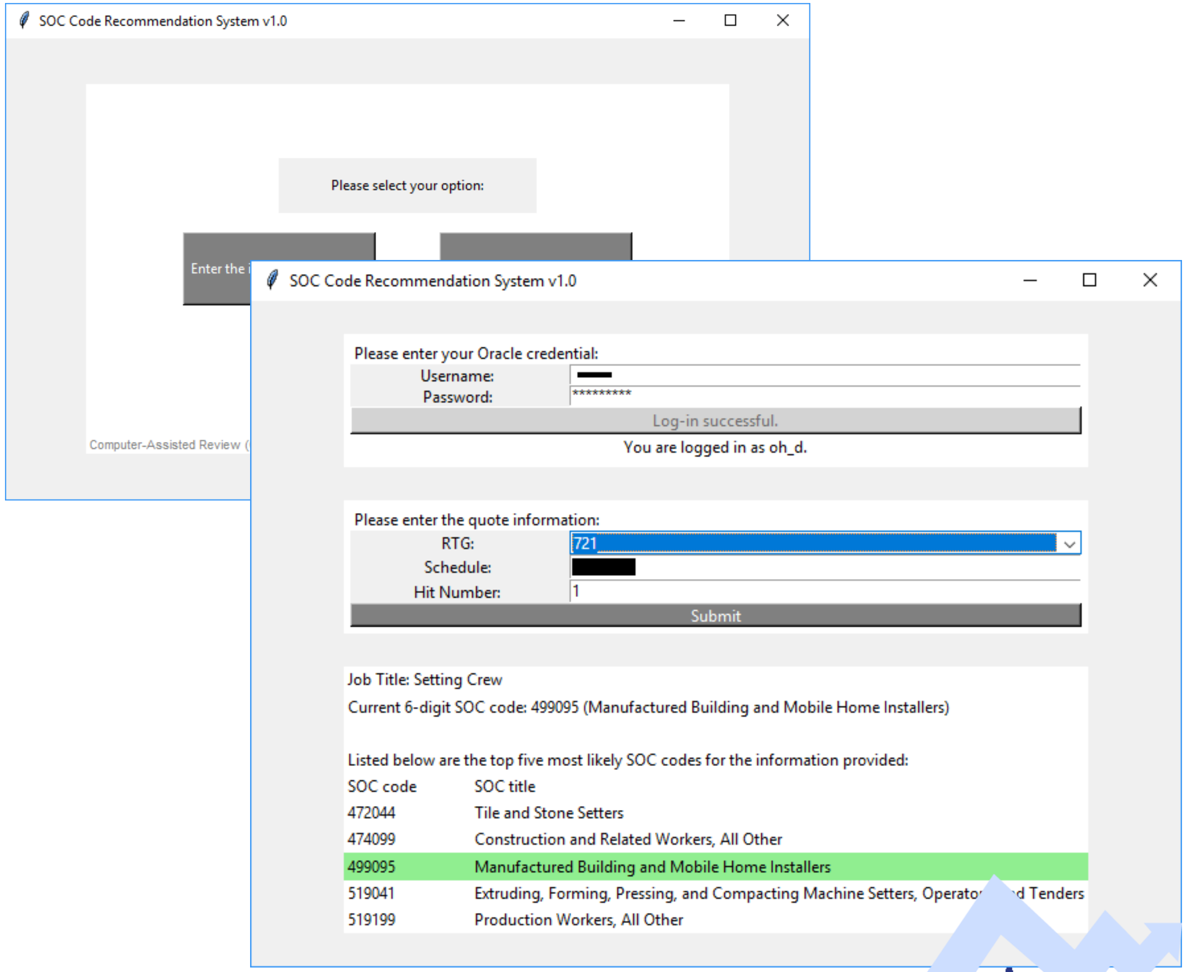


```
errors_df = df_valid[df_valid['INJ_BODY_PART'] != df_valid['PRED_BODY_PART']]
errors_df.sample(4)
```

	YEAR	NARRATIVE	INJ_BODY_PART	PRED_BODY_PART	PROB
3633	2012	Affected employee was involved installing a scaffold base layout on 3-1/3 floor of a preheater. Suddenly without warning, part of the refractory wall above the work area failed and pieces of the wall began falling. The refractory pieces fell about thirty feet striking employee causing injuries.	MULTIPLE PARTS (MORE THAN ONE MAJOR)	CHEST (RIBS/BREAST BONE/CHEST ORGNS)	0.415591
1955	2012	EE WAS PUSHING A RAMCAR TO THE CHARGING STATION WITH HIS RAM HIS BUMPER WENT UP ON THE TAIL GATE THEN DROPPED SLAMMING HIS HEAD INTO THE CANOPY.	HEAD,NEC	FINGER(S)/THUMB	0.348512
10551	2012	On Wednesday evening, two individuals were waiting at the employee's personal truck, in the contractor parking area, and attacked ee, beating him. Causing a broken jaw to our employee.	JAW INCLUDE CHIN	MULTIPLE PARTS (MORE THAN ONE MAJOR)	0.588879
15555	2012	Injured noticed swelling in his right elbow at break time. Injured reported to his supervisor that he must have bumped something during the day that caused the swelling. The following day 8/8, injured requested medical attention when the swelling didn't improve. Injured was diagnosed with bursitis.	ELBOW	HAND (NOT WRIST OR FINGERS)	0.404154

Computer-assisted Coding: SOII and ORS

- SOII autocoder was initially used to flag cases coded by humans that the autocoder predicted differently and had high confidence in
- ORS autocoder is used in an application where data reviewers can query to see a recommendation of codes



SOC Code Recommendation System v1.0

Please select your option:

Enter the i

Computer-Assisted Review (

SOC Code Recommendation System v1.0

Please enter your Oracle credential:

Username: [redacted]

Password: *********

Log-in successful.

You are logged in as oh_d.

Please enter the quote information:

RTG: 721

Schedule: [redacted]

Hit Number: 1

Submit

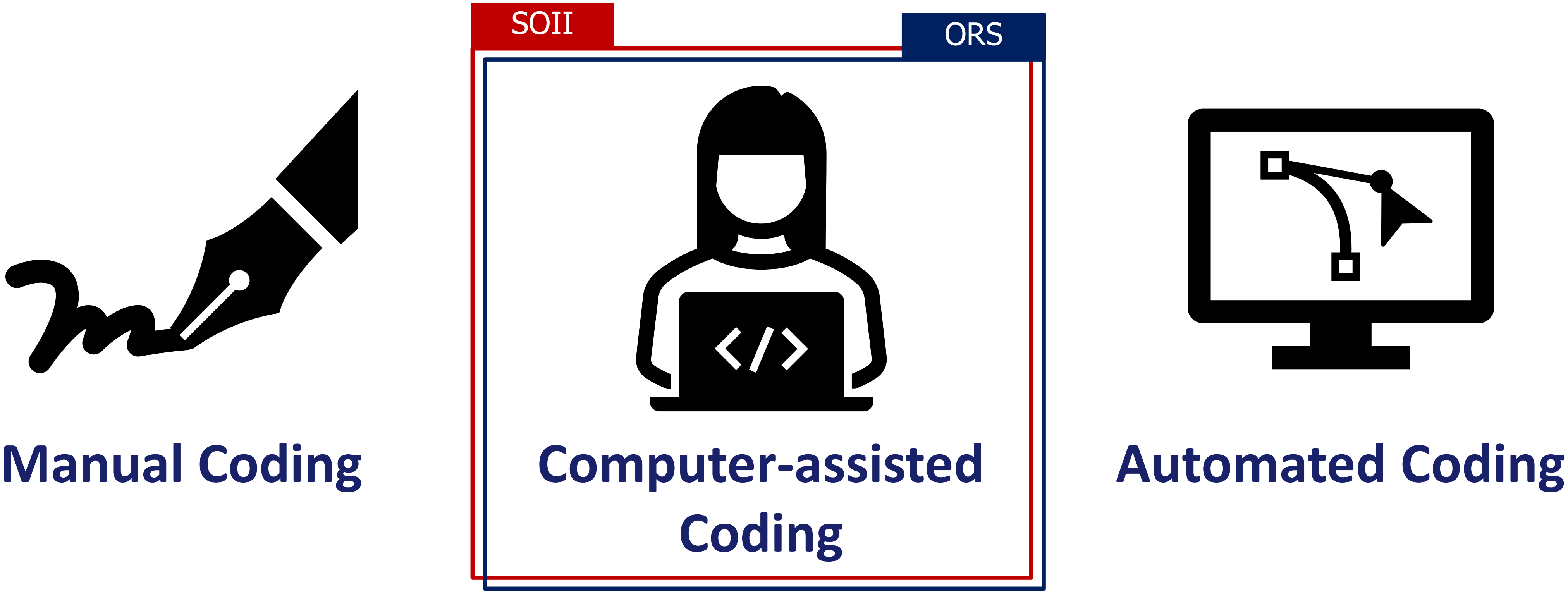
```
Job Title: Setting Crew
Current 6-digit SOC code: 499095 (Manufactured Building and Mobile Home Installers)

Listed below are the top five most likely SOC codes for the information provided:
SOC code    SOC title
472044      Tile and Stone Setters
474099      Construction and Related Workers, All Other
499095      Manufactured Building and Mobile Home Installers
519041      Extruding, Forming, Pressing, and Compacting Machine Setters, Operato   d Tenders
519199      Production Workers, All Other
```

Autocoder as a Review Tool: a Pilot Study

- A pilot study using a randomized, controlled trial showed that computer assistance in occupation coding review leads to…
  - An improvement in occupation coding
  - No apparent biases on the reviewers
  - An increase in time spent reviewing
- We are continuing to experiment with…
  - Providing no codes when probability < threshold
  - Providing Top N codes (usually 5 or 2)
  - Ordering Top N choices by probability (or not)
  - Including probability (or not)
  - Providing alternative indicators of confidence (e.g., 🔴 🟡 🟢)

Techniques for Coding Open-ended Responses



Manual Coding

SOII

ORS

Computer-assisted Coding

Automated Coding

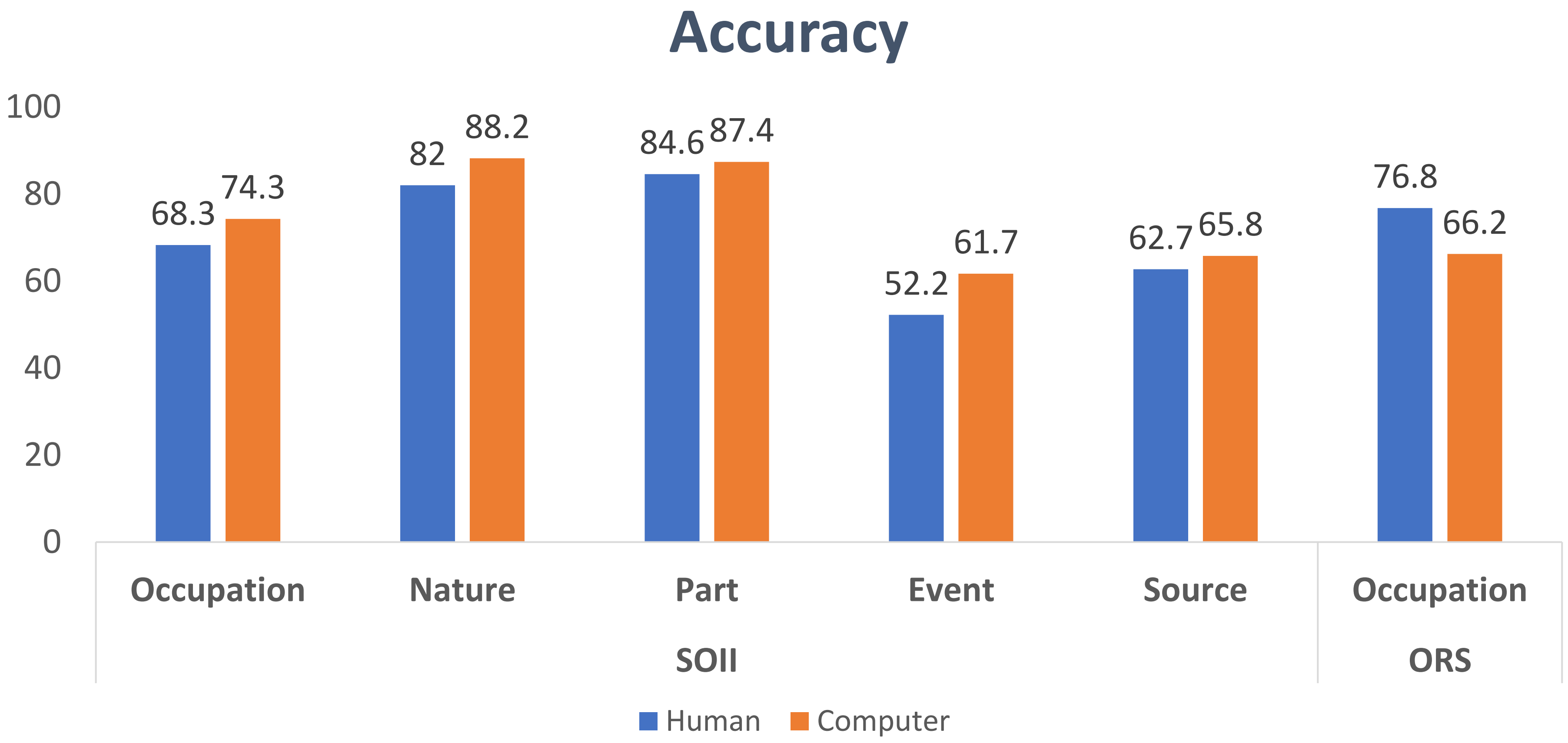
Automated Coding

- The computer system assigns classification codes for some or all cases without human intervention
- Should we…
  - Not autocode anything?
  - Autocode some things?
  - Autocode everything?
- Need a gold standard dataset

Gold Standard Dataset

- A dataset used as the strongest test case to measure the accuracy of human coder and autocoder
- Constructed by taking a representative sample of the data then having multiple experts blindly assign classification codes resulting in a final set of codes that are considered "gold standard"

Human vs. Computer Coding

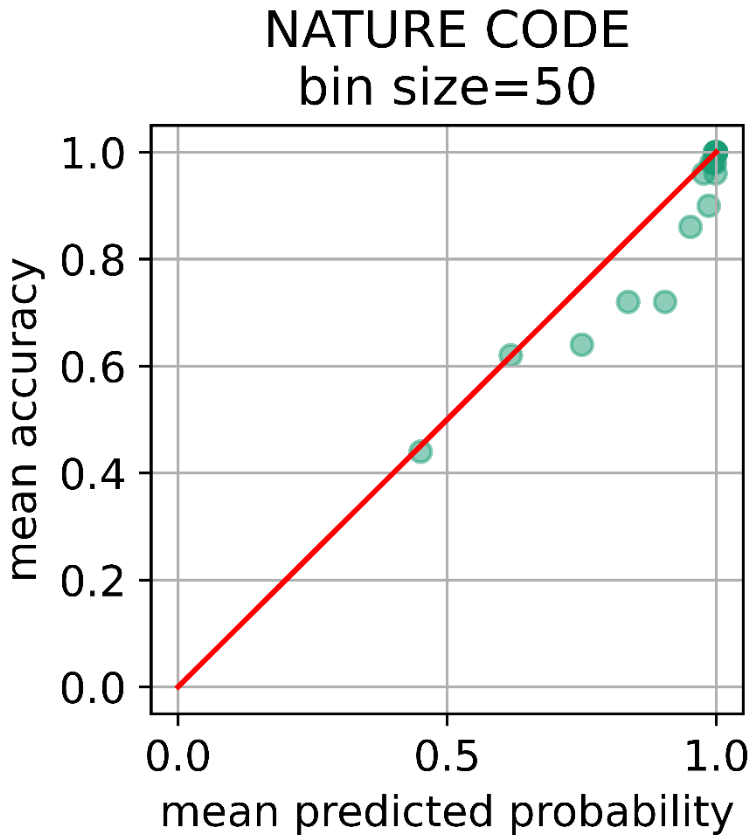


Why Not Autocode Everything?

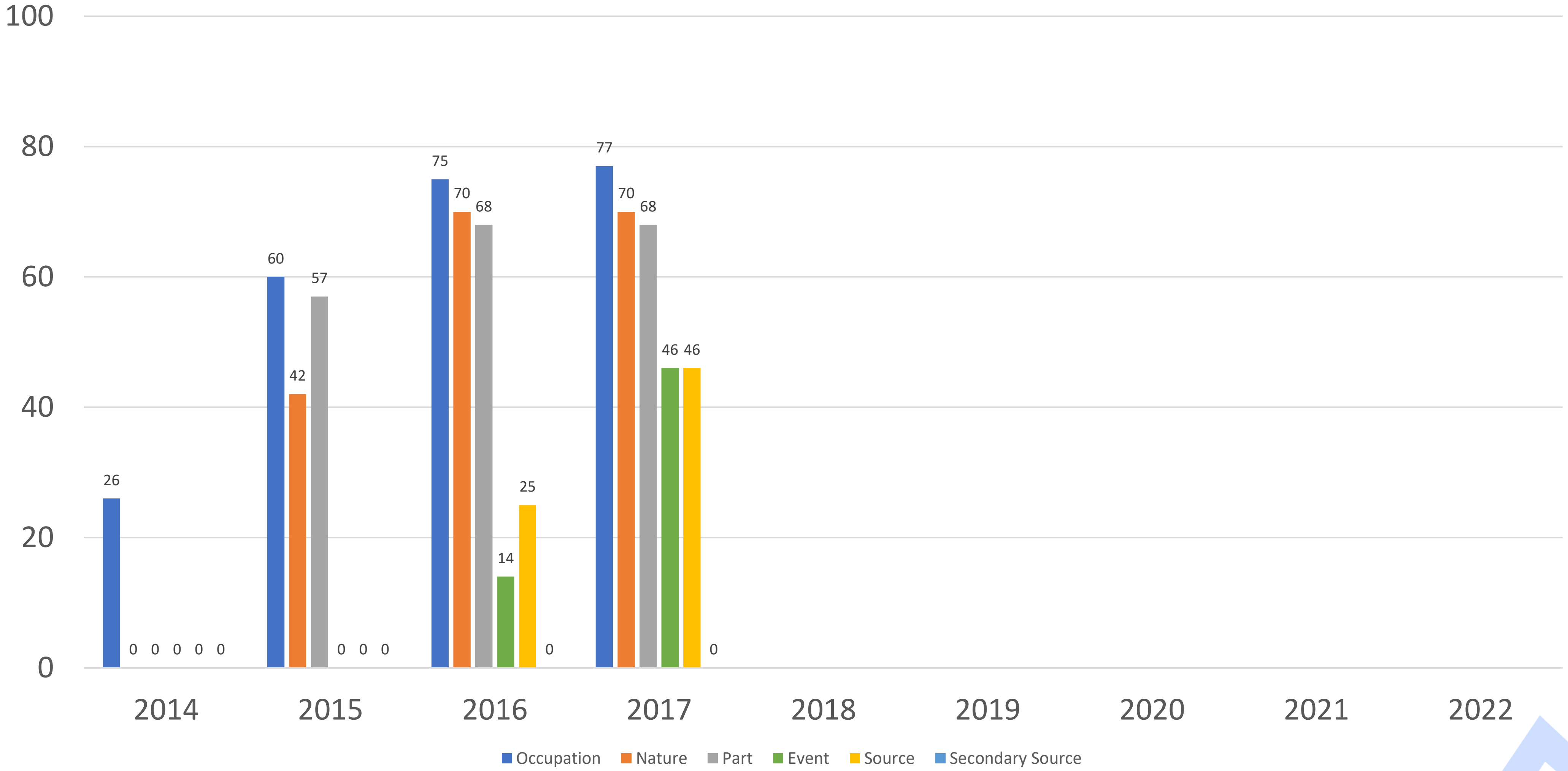
- Accuracy is an aggregate measure of performance
  - There might be certain areas where manual coding (human) outperforms automated coding and vice versa
- Situations where humans are better
  - We can see it in the samples of errors
  - We can see it in the algorithm
    - Algorithm can’t pick up new things
    - Sometimes additional information has to be gathered

Predicted Probabilities and Thresholds

- Autocoder mostly knows what it doesn’t know
  - Predicted prob ≈ True prob
- At what point should human code?
  - Gold standard code + Human code + Computer code allows simulation of hybrid coding approach



% of Codes Automatically Assigned to SOII

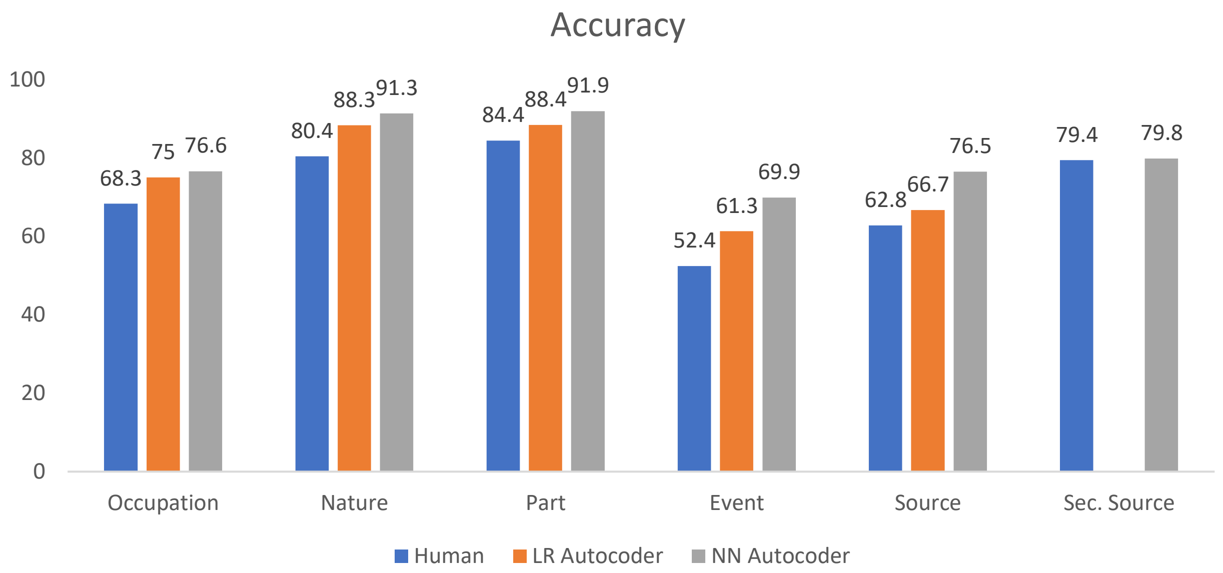


Three Problems with Our Model

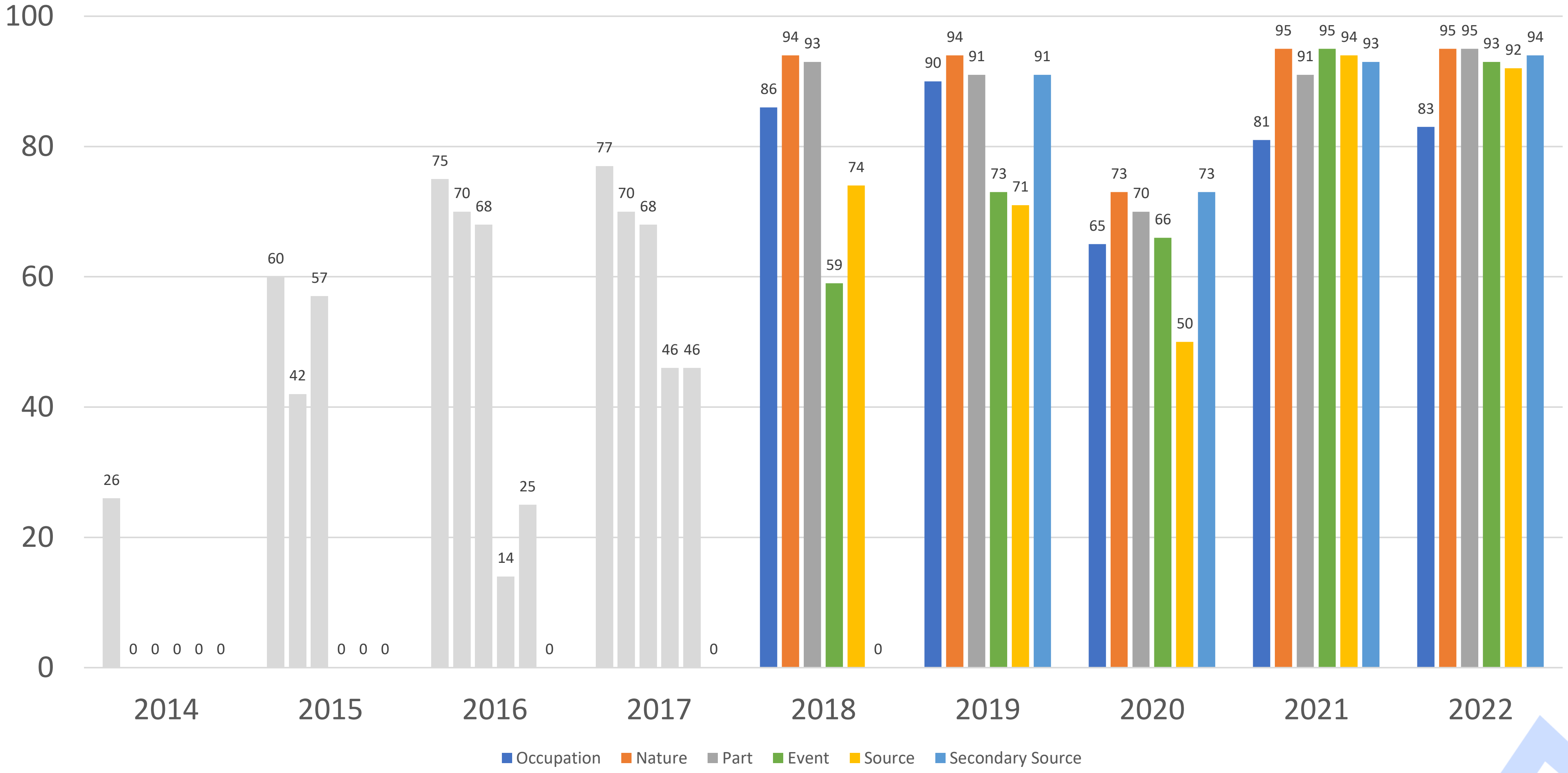
- The linear model assumption
  - “man fell on car” ≠ “car fell on man”
- The bag-of-words assumption
  - “Rock” ≠ “Rocks” ≠ “Rocky” ≠ “Stone” ≠ “Boulder” ≠ “Pebble”
- Huge training data
  - Maybe hundreds or thousands of examples per code
  - Classification systems change

Neural Network Autocoder

- Starting in 2018, we began using deep neural networks
- Timeline:
  - 2018-2020: LSTM model
  - 2021: Transformer model



% of Codes Automatically Assigned to SOII



Final Thoughts on Autocoding at BLS

- BLS has several automated coding projects in production and more that are at various stages of research and development
- This work has mostly been accomplished by staff that transitioned from other roles (economists, statisticians, etc.). BLS is highly supportive of professional development training.
- We have hired more than a dozen data scientists in the last 2 years
- We also advance this work by bringing in interns, sponsoring projects with universities, and participating in coding challenges
- We are exploring the feasibility of incorporating ML models into web-based surveys to assist respondents with live coding
- There has been increased scrutiny on how AI/ML applications may impact the public and staff at BLS

Other Classification Systems in Government

Census Industry Codes	--	• American Community Survey (Census)	Link
Census Occupation Codes	--	• American Community Survey (Census)	Link
North American Product Classification System	NAPCS	• Economic Census (Census)	Link
International Classification of Diseases	ICD-10	• National Vital Statistics System (NCHS)	Link
Injury and Product Classifications	--	• National Electronic Injury Surveillance System (CPSC)	Link
Standardized Classification of Transported Goods	SCTG	• Commodity Flow Survey (BTS)	Link
College Course Map	CCM	• Postsecondary Education Transcript Studies (NCES)	Link

Resources

- Autocoding at BLS on bls.gov
- Automated Coding of Worker Injury Narratives (PDF)
- Deep Neural Networks for Worker Injury Autocoding (PDF)

Contact Information

David Oh
Supervisory Data Scientist
BLS/OCWC
oh.david@bls.gov

Brandon Kopp
Supervisory Data Scientist
BLS/OSMR
kopp.brandon@bls.gov