INSTITUTE FOR SOCIAL RESEARCH • SURVEY RESEARCH CENTER
SURVEY RESEARCH OPERATIONS
UNIVERSITY OF MICHIGAN

A coding application

Gina-Qian Cheung and Cheng Zhou

TSG

Background 1-Every job has a code

US Census Bureau	
2010 Occupation Code List	
last updated: August 12, 2011	
Occupation 2010 Description	**2010 Census Code**
The 2010 census occupation classification list has 539 codes including 4 military codes.	
Computer and information research scientists	**1005**
Computer systems analysts	**1006**
Information security analysts	**1007**
Computer programmers	**1010**
Software developers, applications and systems software	**1020**
Web developers	**1030**
Computer support specialists	**1050**
Database administrators	**1060**
Network and computer systems administrators	**1105**
Computer network architects	**1106**
Computer occupations, all other	**1107**
Actuaries	1200
Mathematicians	1210
Operations research analysts	1220
Statisticians	1230
Miscellaneous mathematical science occupations	1240

American Community Survey	
Industry Code List	
Industry 2010 Description	**2010 Census Code**
271 codes	
Legal services	7270
Accounting, tax preparation, bookkeeping, and payroll services	7280
Architectural, engineering, and related services	7290
Specialized design services	7370
Computer systems design and related services	7380
Management, scientific, and technical consulting services	7390
Scientific research and development services	7460
Advertising and related services	7470
Veterinary services	7480
Other professional, scientific, and technical services	7490
Management of companies and enterprises	7570
Employment services	7580
Business support services	7590
Travel arrangements and reservation services	7670
Investigation and security services	7680
Services to buildings and dwellings (except cleaning during construction and immediately after construction)	7690

Background 2- A real coding case

A respondent gave following information:

Job tile: Manager, Sales person.

Job duty: sell to public . supervise 2 people . use computer . licesnse from state of TN to sell manufactured homes. sell and supervise and delivery of manufactured homes . calling to arrange deliveries .

Industry: Housing industry . private . national company . manufacturing and sales . build houses and commerical residences . #employees location 8 . all locations employees probably 30000 .

Our senior coder coded:
Occupation: 4700 First-line supervisors of retail sales workers

Background 3 – Design requirement

- Need to develop a new coding application to accommodate data collected from different sources.
- Being able to code different types of coding, like Occupation/Industry, cognition, opinion, etc.
- Use the advanced technology, like machine learning, to speed up the coding process.
- **Focus of this presentation:** machine learning methods to predict occupation and industry codes.

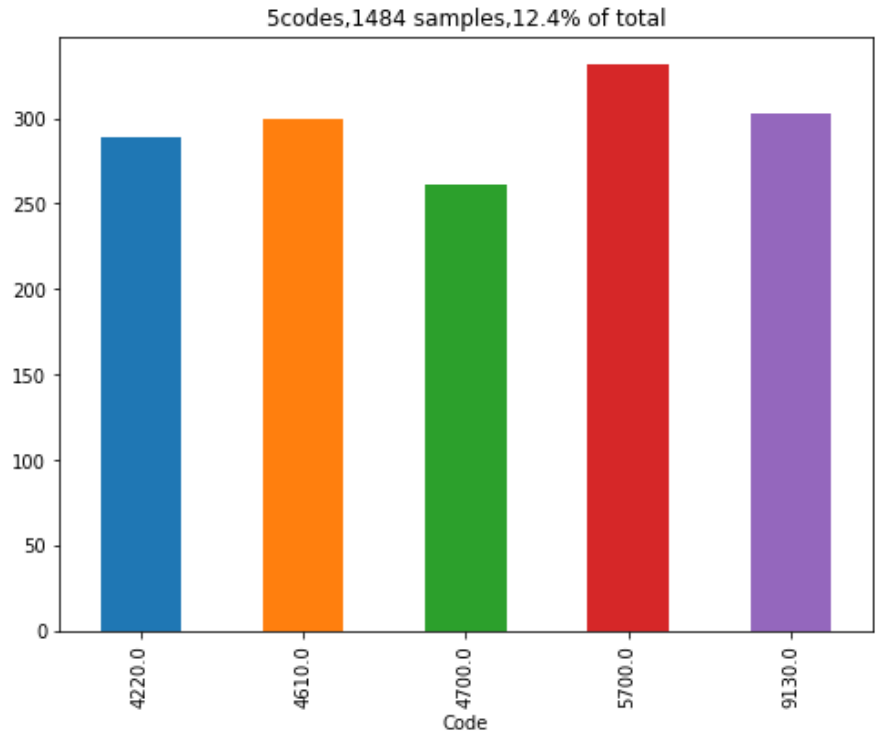
Training and Testing data description -1

- HRS2012/2014/2016 current job information.
- Industry description(Ind1), Job title (Occ1), Job Duty (Occ2)
- ~13K records in total
- ~12K records after getting rid of records in Spanish by checking against the stop words (roughly 7-8% are in Spanish).
- Each record has already been coded with **Industry code** and **Occupation code.**
- 80% of total selected data will be used as training data to train the model and the rest 20% will be used to test the model's prediction ability.

Training and Testing data description -2
Potential data issues

- Sample data are not evenly distributed. Some codes have more coverage while others may only have 1 or 2 cases.
- There are only 5 codes which have at least 250 records

Janitors and building cleaners	4220
Personal care aides	4610
First-line supervisors of retail sales workers	4700
Secretaries and administrative assistants	5700
Driver/sales workers and truck drivers	9130



Machine Learning Model Design

- Python
- Use the SVM (Support Vector Machine) /Neural network machine learning methods. Also tried other methods: logistical regression, Naive Bayes, Random Forest.
- Combined SVM and Neural network have the best success rates.

From natural language to matrix: Tf-idf, Term frequency&inverse document frequency

Step 1. Term frequency

- Consider one document/description containing 100 words wherein the word *computer* appears 10 times. The term frequency (i.e., tf) for *computer* is then (10 / 100) = 0.1.
- Another word *grocery* appears 1 times. The term frequency for *grocery* is then (1 / 100) = 0.01.
- Another word *I* appears 5 times. The term frequency for *I* is then (5 / 100) = 0.05.

Step2. Inverse documents frequency

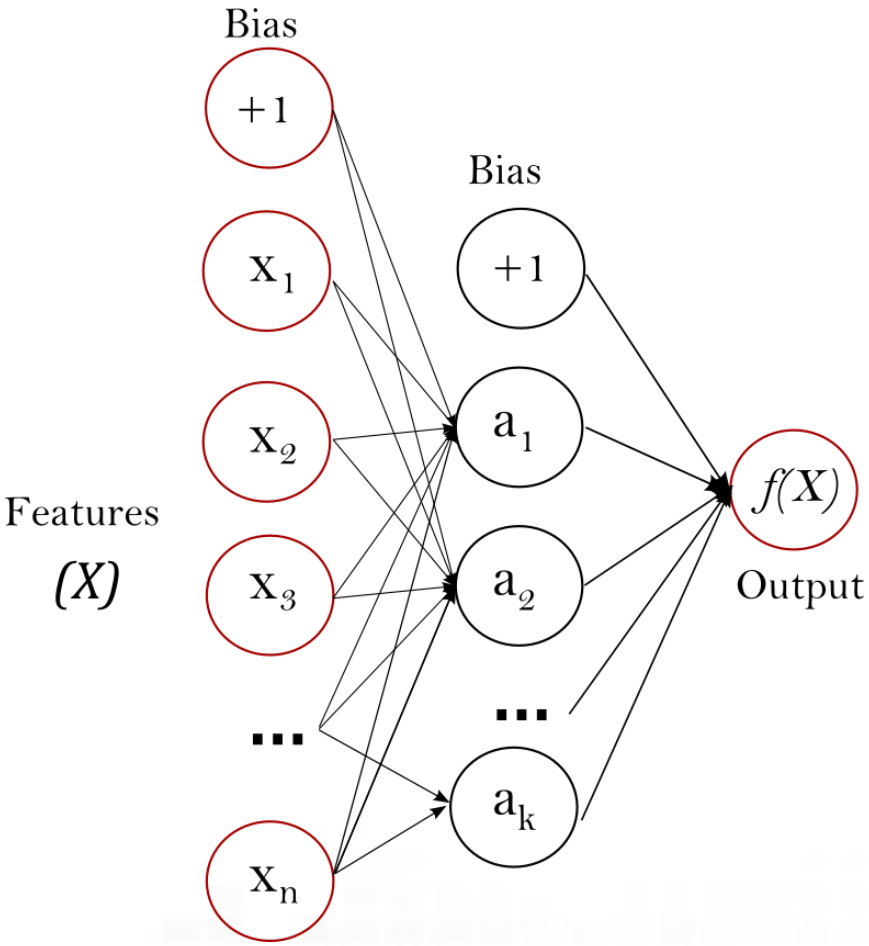
- Assume we have 1000 documents and the word *computer* appears in 10 of these. Then, the inverse document frequency (i.e., idf) is calculated as log(1000/ 10) = 2. Thus, the Tf-idf weight of *computer* is the product of these quantities: 0.1 * 2 = 0.2.

- Assume another word *I* appears in all of them. Then, the inverse document frequency (i.e., idf) is calculated as log(1000/ 1000) = 0. Thus, the Tf-idf weight of *I* is the product of these quantities: 0.05 * 0 = 0.

Final look:

0, 0, 0, 0.2, 0,….., 0.01,…,0.3, 0.1,…… 0 ->170

Machine learning method 1: Neural network models (supervised): Multi-layer Perceptron

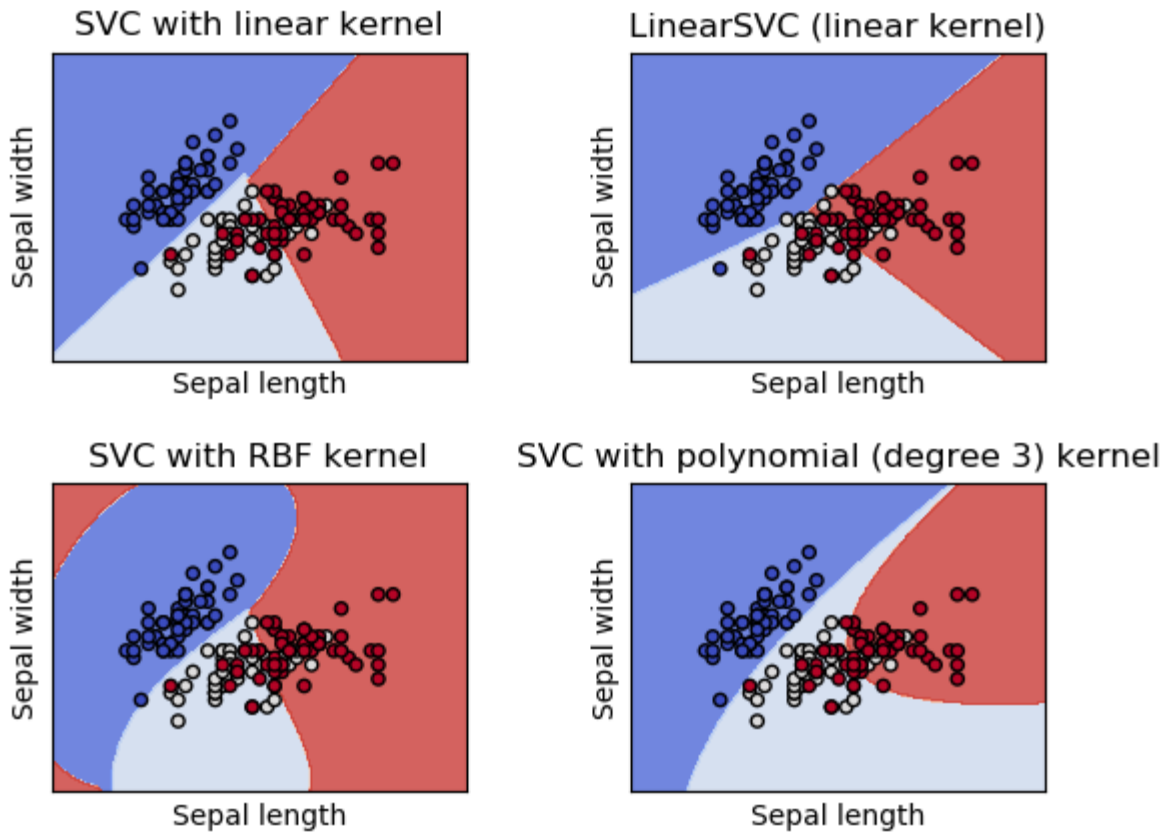


We used 100 layers.

http://scikit-learn.org/stable/modules/neural_networks_supervised.html

Machine learning method 2: SVM classifier-supporting vector machine

We used SVC with linear kernel



http://scikit-learn.org/stable/auto_examples/svm/plot_iris.html

What our model can provide? – codes and probabilities

Job tile: Manager, Sales person.

Job duty: sell to public . supervise 2 people . use computer . licesnse from state of TN to sell manufactured homes. sell and supervise and delivery of manufactured homes . calling to arrange deliveries .

Industry: Housing industry . private . national company . manufacturing and sales . build houses and commerical residences . #employees location 8 . all locations employees probably 30000 .

Prediction1	probability1	Prediction3	probability2	Prediction3	probability3
4700: First-line supervisors of retail sales workers	94%	**9130: Driver/sales workers and truck drivers**	5%	**5700: Secretaries and administrative assistants**	1%

The first prediction is correct.

What our model can provide?- 2 Another case

Job tile:

Job duty: Vacuum, dust, wash floors, Windows occassionally. I work about eight hours a week, people haven't the money to pay for it. With a crew of two people it should take about three hours, depending on the kind of work they want done.

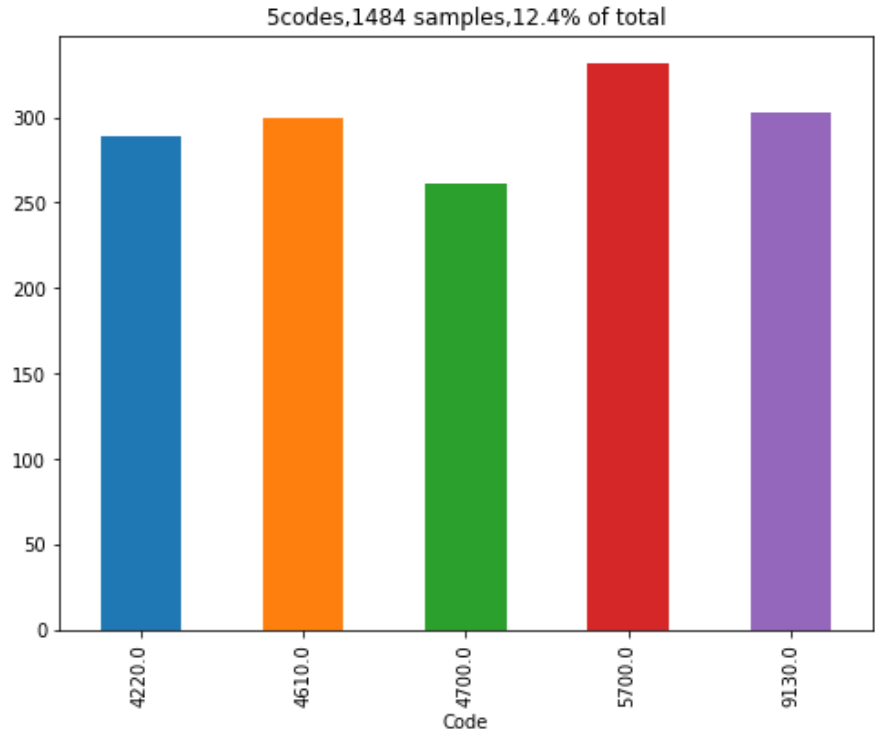
Industry: House cleaning.

Prediction1	probability1	Prediction3	probability2	Prediction3	probability3
4220: Janitors and building cleaners	40%	**4230: Maids and housekeeping cleaners 33%**	33%	**4200: First-line supervisors of housekeeping and janitorial workers**	4.5%

The 2nd prediction is correct.

Model results 1: Occupation coding
Select codes with at least 250 records (5 codes only)

Janitors and building cleaners	4220
Personal care aides	4610
First-line supervisors of retail sales workers	4700
Secretaries and administrative assistants	5700
Driver/sales workers and truck drivers	9130

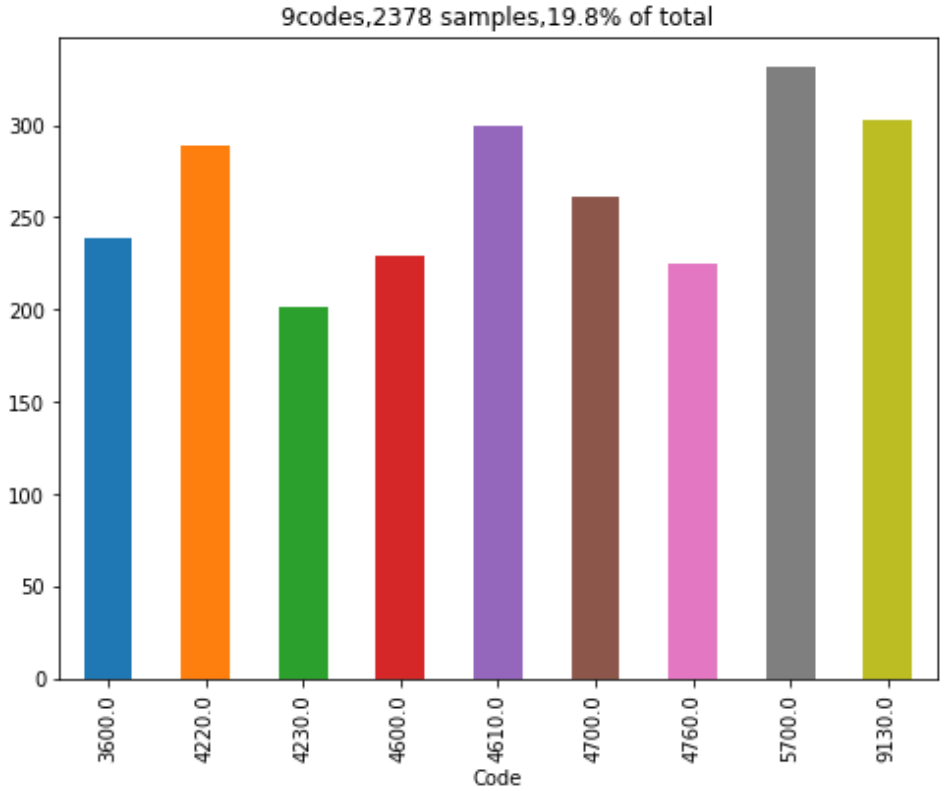


- 5 codes only, Train data size: 1484 records (80%), Test data size: 297 records (20%)
- **296 cases out of 297 are right, except**
- Prediction success rate is 99.7%!!!

Actual code: 4220 Janitors and building cleaners
Predicted code: 4700 First-line supervisors of retail sales workers
Description: Production clerk, R does night cleaning, Grocery Store

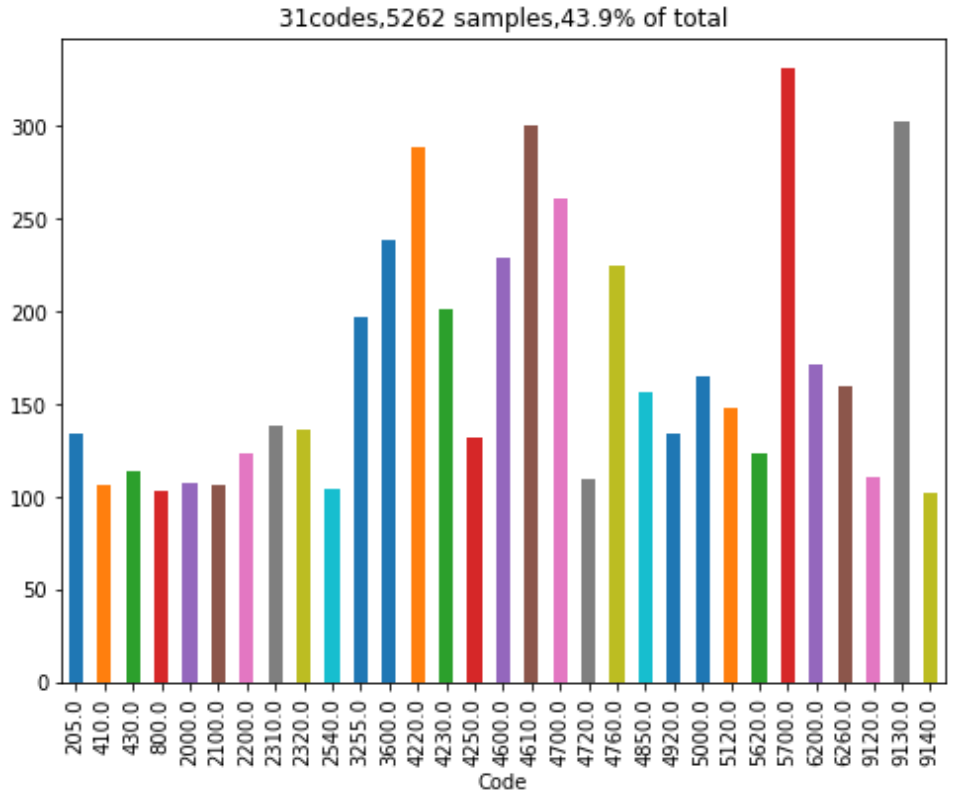
Model results 2: Occupation coding

Select codes with at least 200 records (9 codes only)



- 9 codes only, Train data size: 1902 records, Test data size: 476 records
- 432 cases out of 476 are right
- Prediction success rate is 91%

Model results 3: Occupation coding Select codes with at least 100 records



Question: We have 3 pieces of information (job title, duty an industry description) and 2 codes to predict (industry code and occupation code), what information to use to predict the two codes?

Conclusion:
Job title*3 + duty + Industry description is the best. Order matters too!

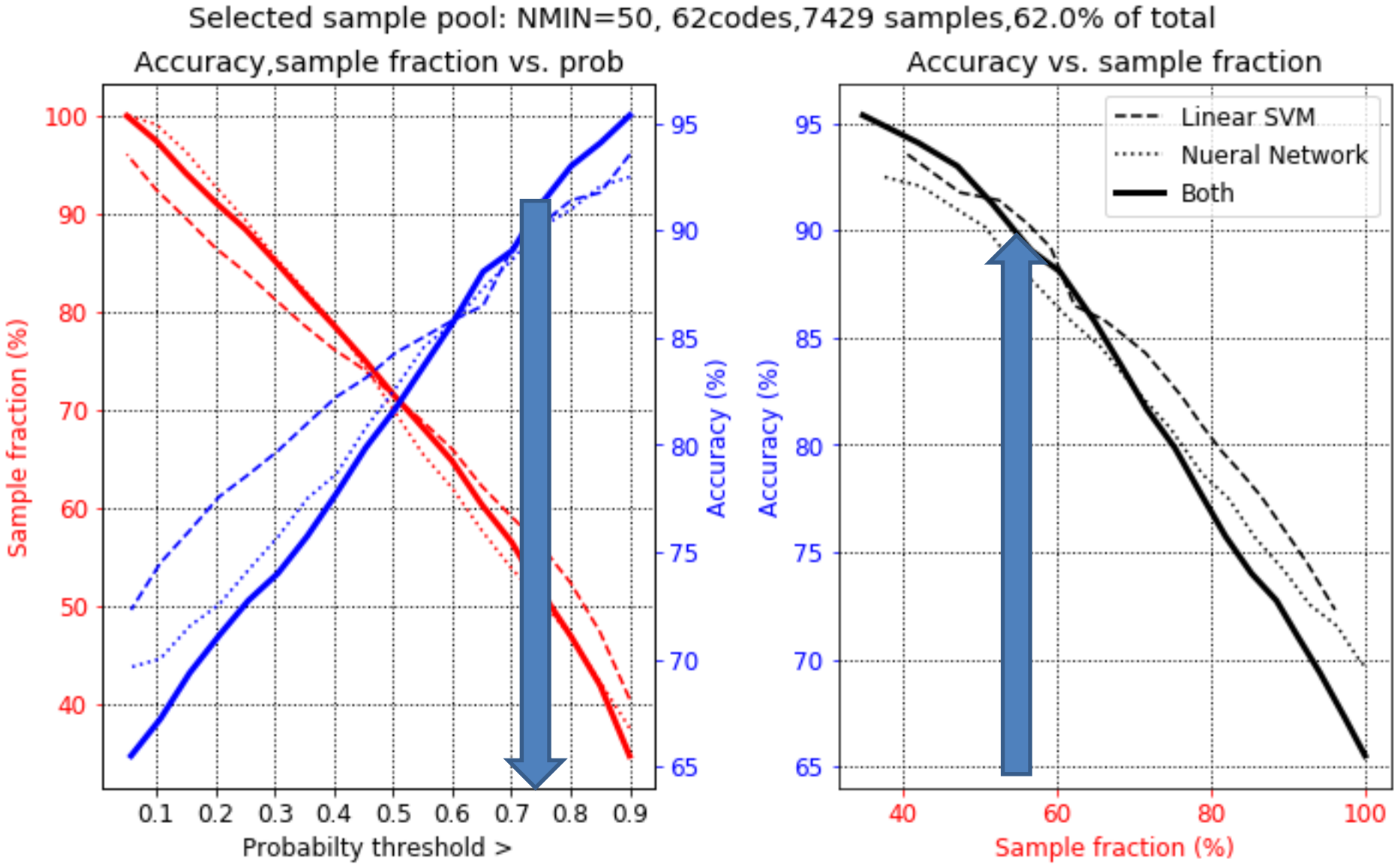
Choice of description	Job title	Job duty	Job title+duty	Job title*3+duty+industry
Success rate	45%	65%	71%	**82%**

A summary of success rates from SVM

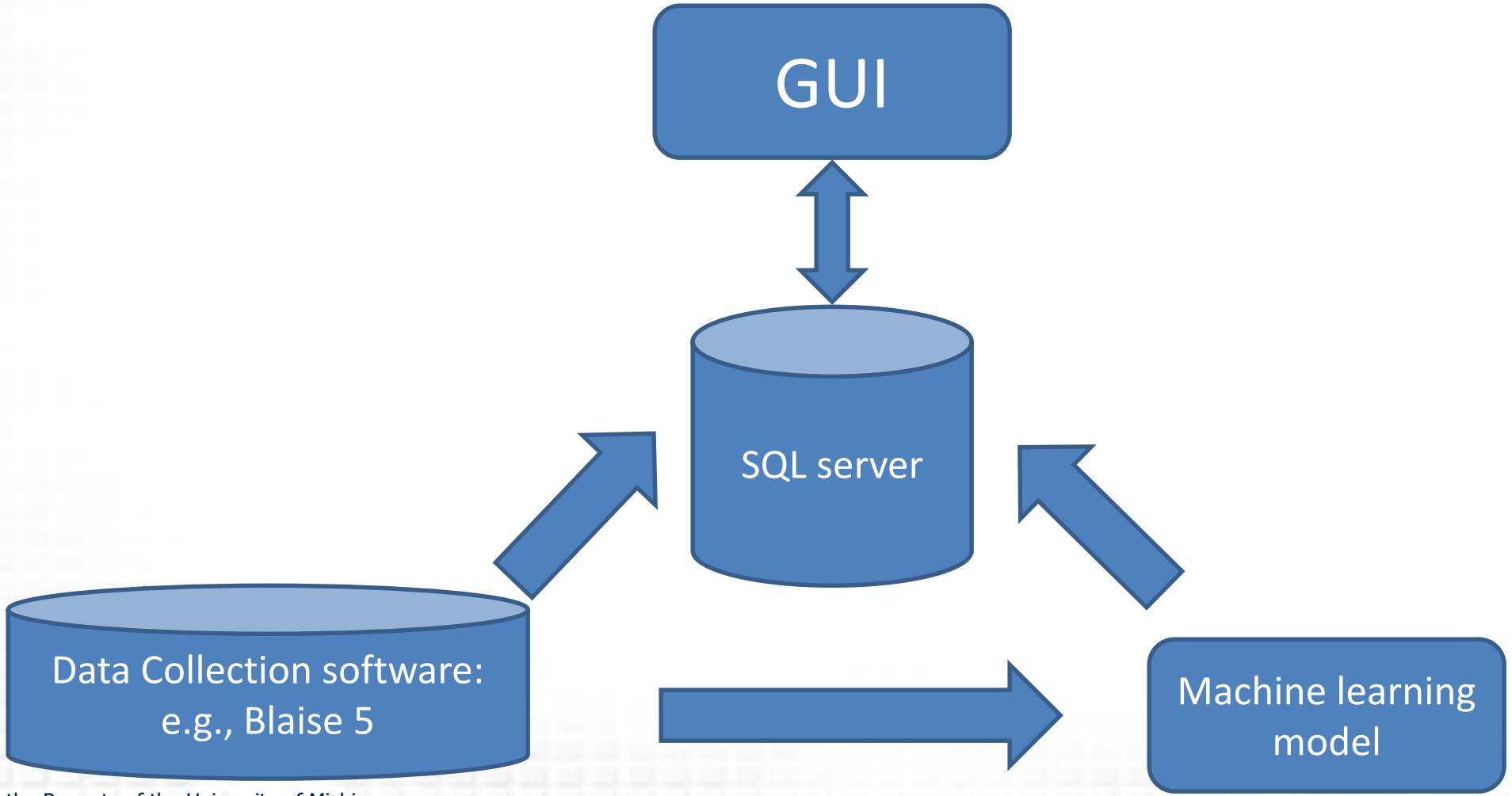
Selecting codes w/ at least N records	Occupation coding Success Rate (% of total samples selected, # of codes)	Industry coding Success Rate (% of total samples, # of codes)
N>=250 for all codes	99.4% (12.4% of total sample selected, 5 codes)	93% (27%, 7 codes)
N>=200	91% (19.8%, 9 codes)	89% (35%, 11 codes)
N>=100	82% (44%,31 codes)	84% (57%, 29 codes)
N>=50	71% (62%,62 codes)	79% (73%,55 codes)
N>=10	61%(91%,216 codes)	66% (96%, 162 codes)
N>=5	58%(96%,283 codes)	
All	57%(100%,494 codes)	

Best results for Occupation coding is achieved when repeating the job title 3 times in the string and setting hf=3 in the tf-idf function. This has little impact on the industry coding.

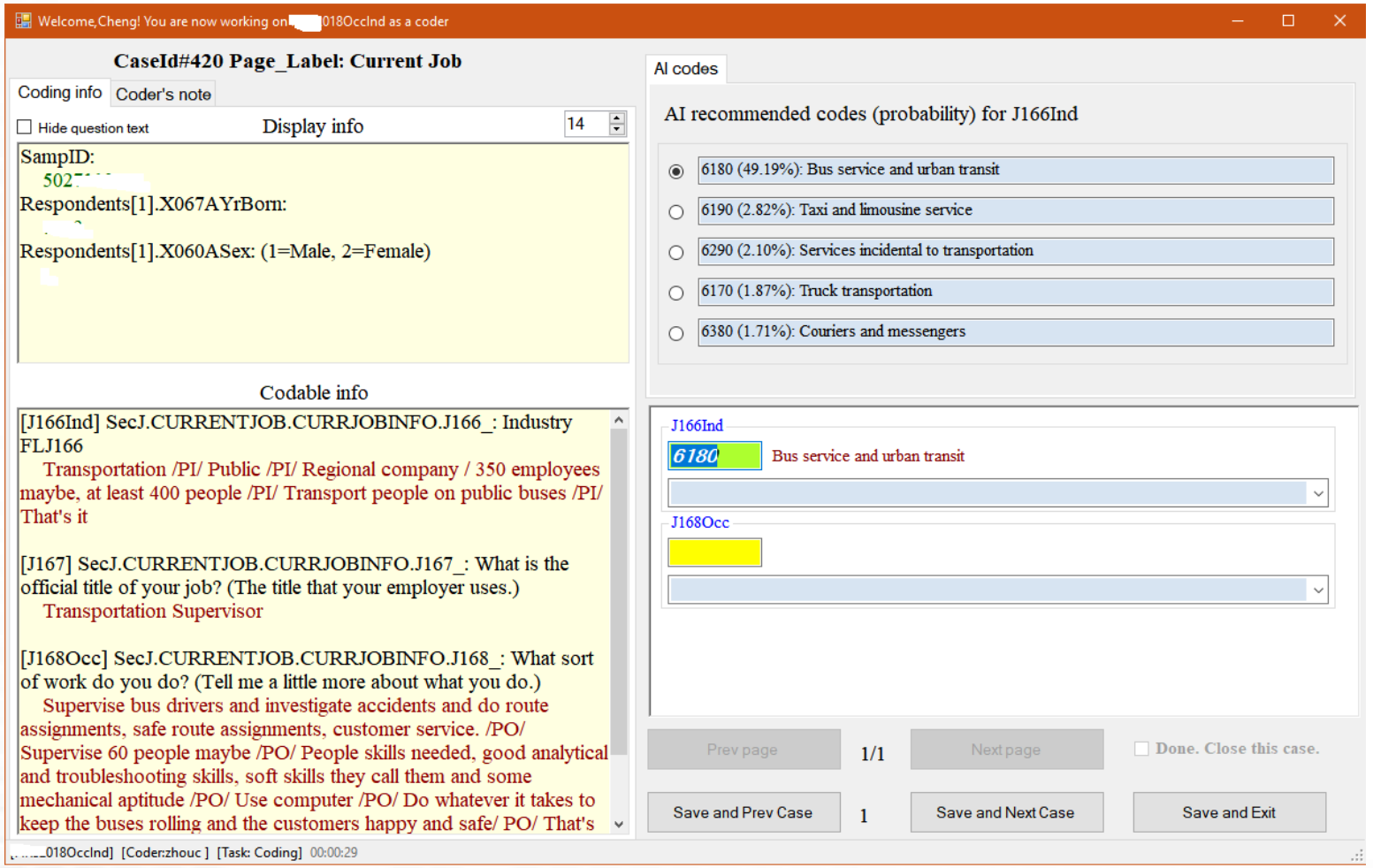
Automated coding using two supervised machine learning methods



Our new coding application

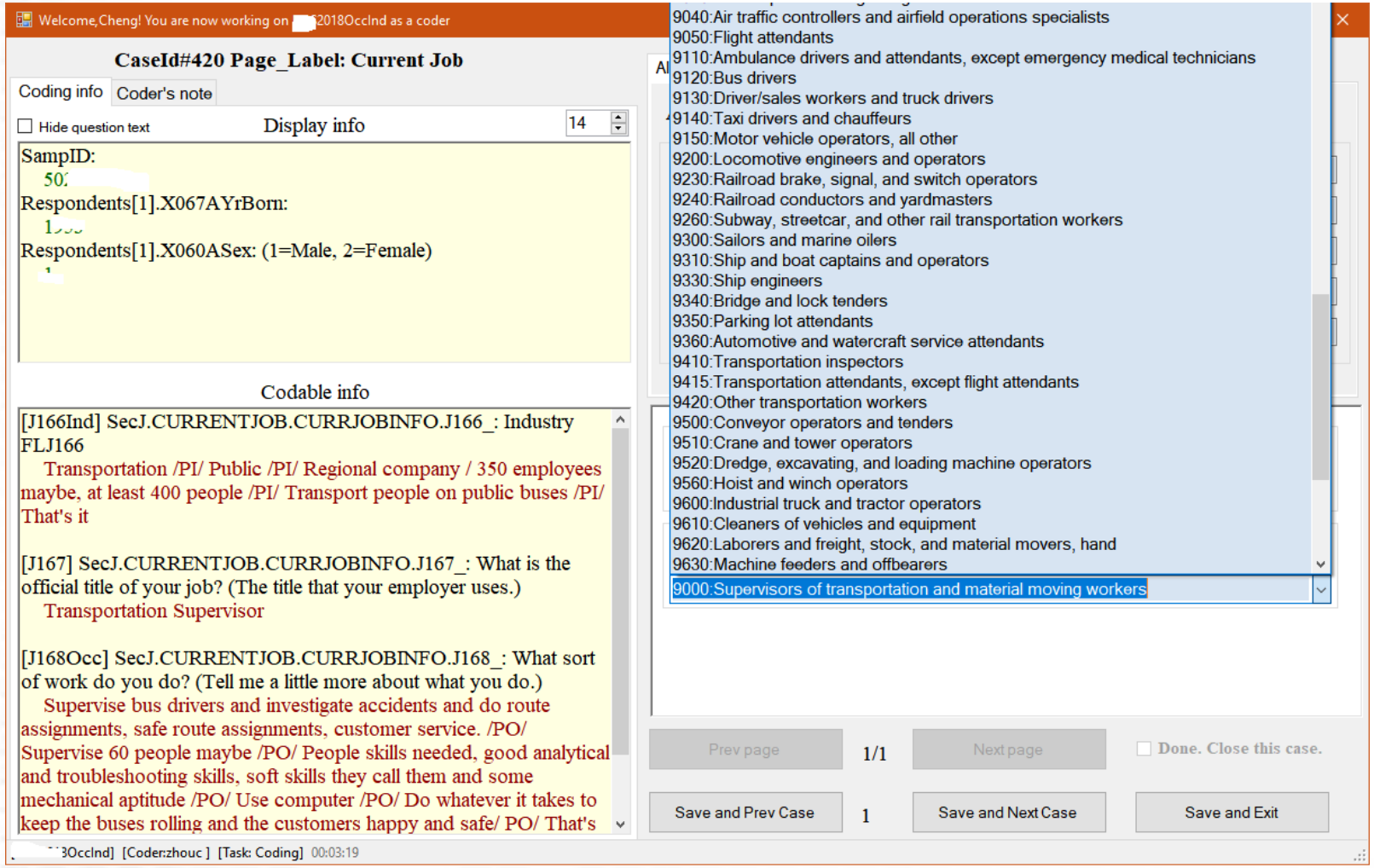


An example from the new coding application



Feature 1: Top 5 most likely codes from the machine-learning are provided to the coder

An example from the new coding application



Feature 2: Coder can search based on partial code or description

Performance of the machine-learning model

	Percentage of matching
Matching 1st code	51.4%
Matching 2nd code	11.7%
Matching 3rd code	5.0%
Matching 4th code	3.2%
Matching 5th code	1.9%
Overall	**73.3%**
Not match	**26.7%**

- 5107 coded codes from a panel study
- Spanish records are included

Summary

- We built a new coding application to accommodate different coding needs and data sources
- A machine-learning model is built for occupation/industry coding
- A over all 51% matching percentage for the 1st predicted code and 73% matching percentage for the top 5 most likely codes
- Additional training data coverage will improve the prediction success rate
- A separate Spanish model is needed.

INSTITUTE FOR SOCIAL RESEARCH • SURVEY RESEARCH CENTER
SURVEY RESEARCH OPERATIONS
UNIVERSITY OF MICHIGAN

Thank you!

LUIS (Microsoft Azure Language Understanding)

- Utterance -> Intent, e.g., 'I want to go to Seattle'->'Book a flight'
- Some limitations:
  - 500 characters for each utterance, pay attention to ". Avoid this.
  - 15000 utterance for each project
  - 500 intents
  - Import batch file: 100 utterance per file,
  - Intents need to be created manually
  - Test batch files: 1000 utterance per file.
  - Demonstrate

https://docs.microsoft.com/en-us/azure/cognitive-services/luis/home
https://www.luis.ai/

A summary of success rate of occupation coding: LUIS vs. In-house model

Selecting codes w/ at least N records	LUIS	In-house model
N>=250 for all codes (12.4% of total sample selected, 5 codes)	98.3%	99.4%
N>=200 (19.8%, 9 codes)	86%	91%
N>=100 (44%,31 codes)	74%	82%
N>=50 (62%,62 codes)	?	71%