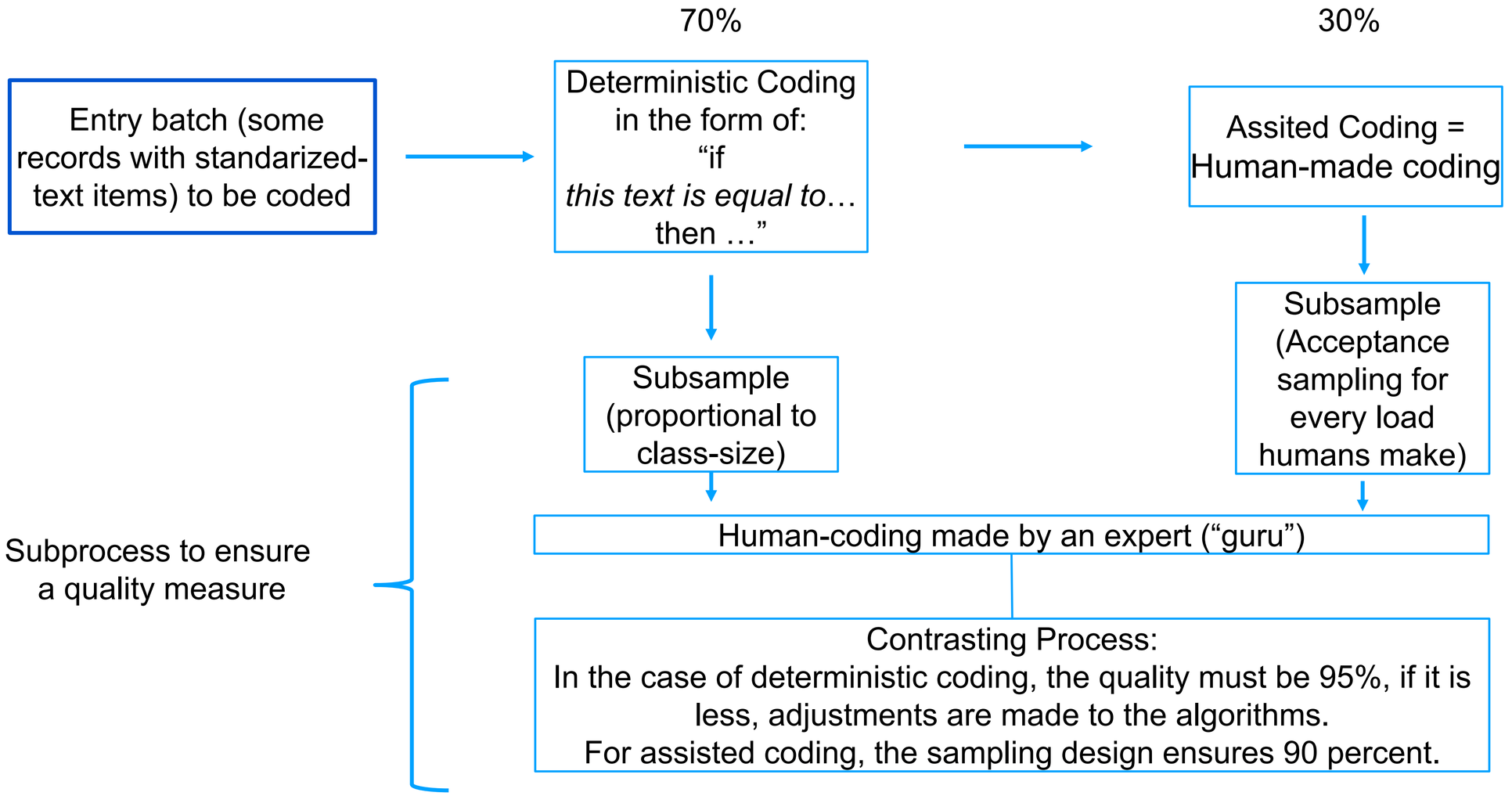
Natural Language Processing for Economic Activity and Occupation

Goal

- To research the extent to which NLP can be incorporated in the current coding production-process:
  - Reduce the human-assisted workload
  - Reduce the timing
  - Matain or improve the quality
  - The hardest ones: Economic Activity, Occupation

Current Coding Process = Ground Truth



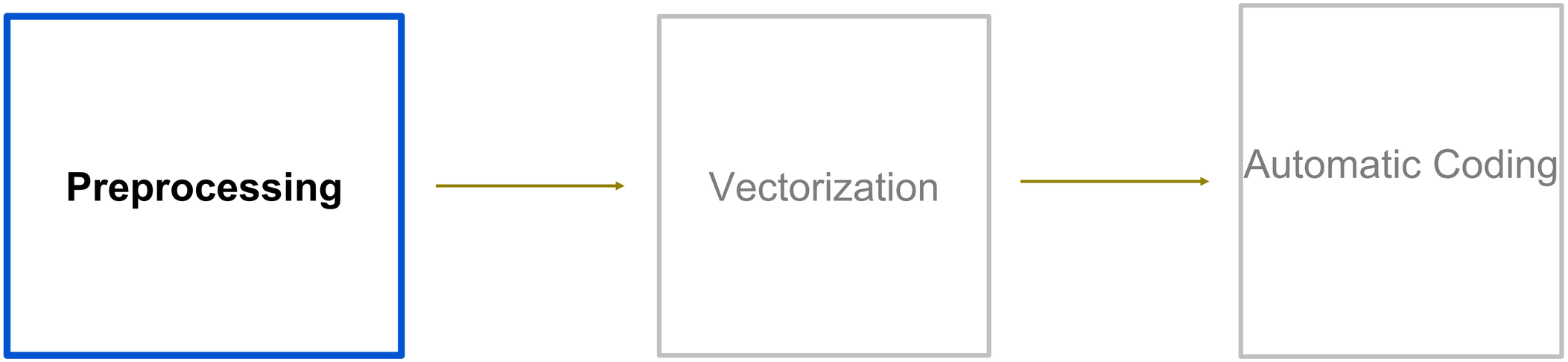
NLP Stages

- 2018 National Household Income and Expenditure Survey (ENIGH). 158k coded records

  - Features:
    - {**Ocupation:** ‘COCINERO DE ANTOJITO MEXICANO’}
    - {**Task:** ‘PREPARAR VENDER DE ANTOJITO MEXICANO EN LOCAL’}
    - {**Business activity :** ´PREPARAR VENDER DE ANTOJITO MEXICANO EN LOCAL AL PUBLICO EN GENERAL´}
    - {**Company name:** ‘HUARACHE MIMI’}
    - **13 aditional items:** academic level, company size,…

  - Variable to clasify (either Economic Activity or Occupation): {7221}
    - Hierarchical code: first two digits account for the sector

Stages



Preprocessing for ML Coding

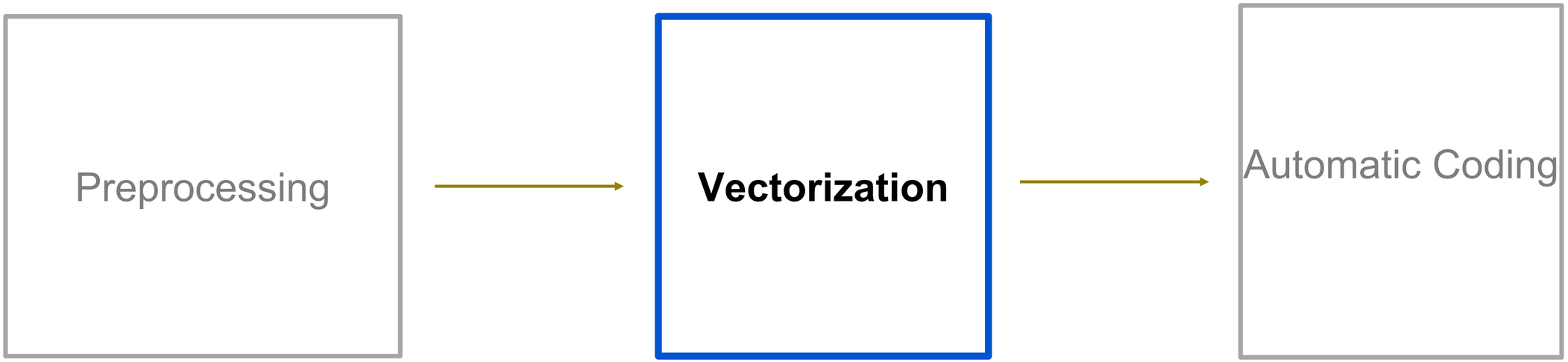
- Ortographic correction
- Lemmatization
- Stemming

Based on our experience instead of a library (e.g. python)

Same process applied to both current coding and ML coding

- Q-grams (n-letter sequence) and N-grams (n-word sequence)

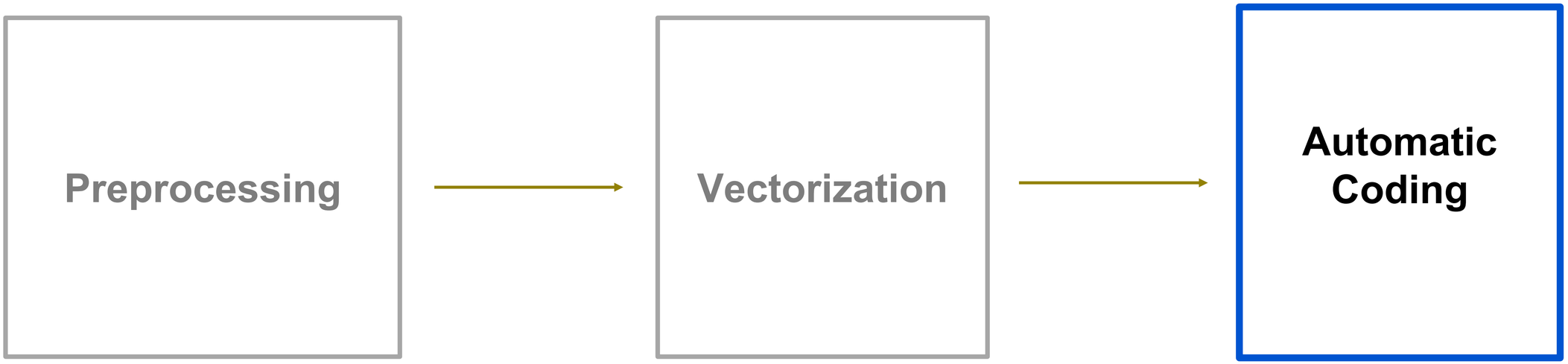
Stages



- Two TFiDF matrices (each one of 30 000 columns):
  - 2-word sequence
  - 6-letter sequence

- Concatenate the two matrices into one of 60 000 columns + 13 auxiliary variables

Stages



Ensemble ML

- ML algorithms used:
  - SVM (best accuracy among all)
  - Logistic regression
  - Random Forest
  - Neural Networks
  - XGBoost
  - K-NN
- The final assigned code is made by a voting process:
  - Equal weights for each ML-method call
  - Unequal weights (try and error)

Two sets: Training (75%) and Test (25%). The following are based on the test set

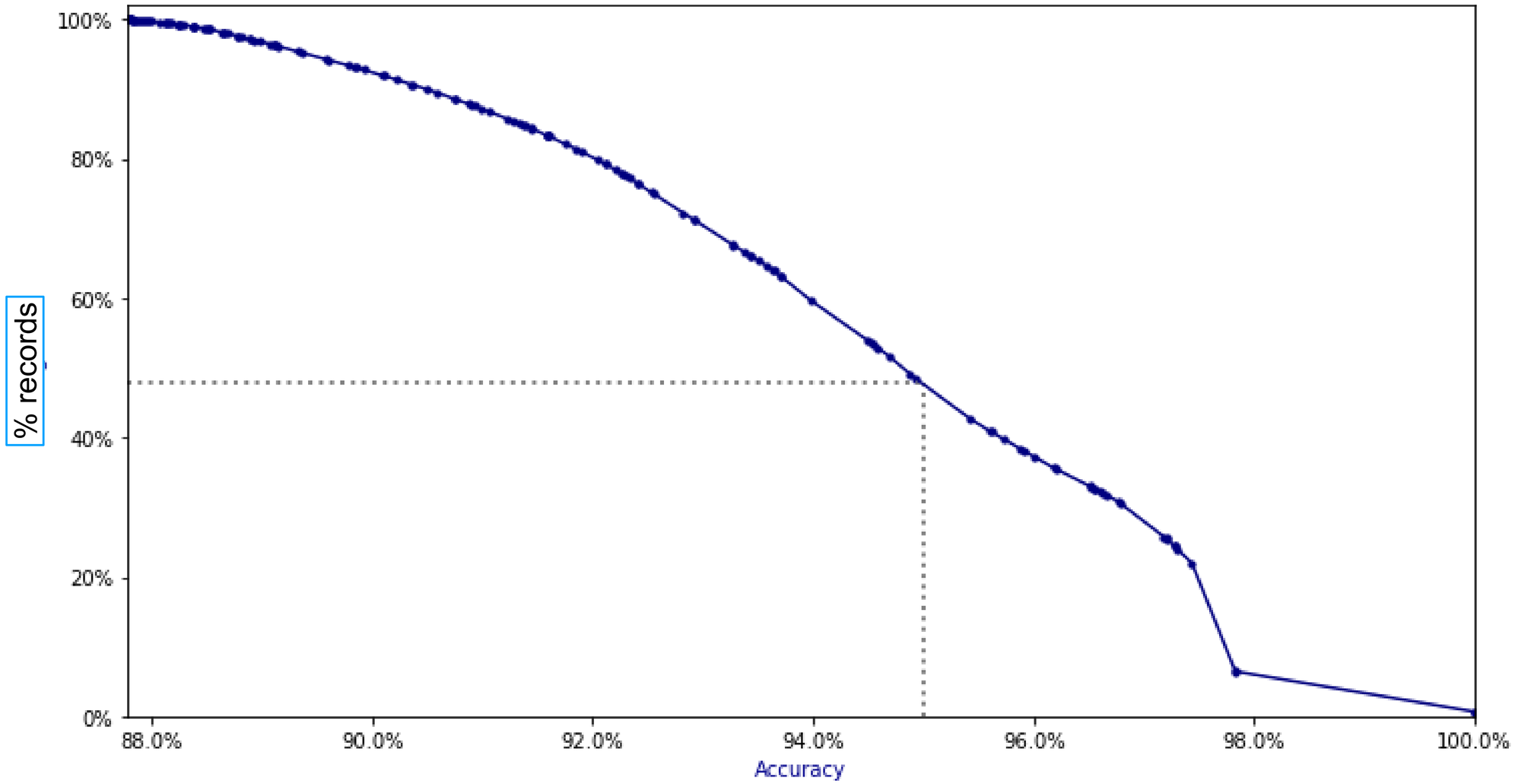
Economic Activity (157 classes)

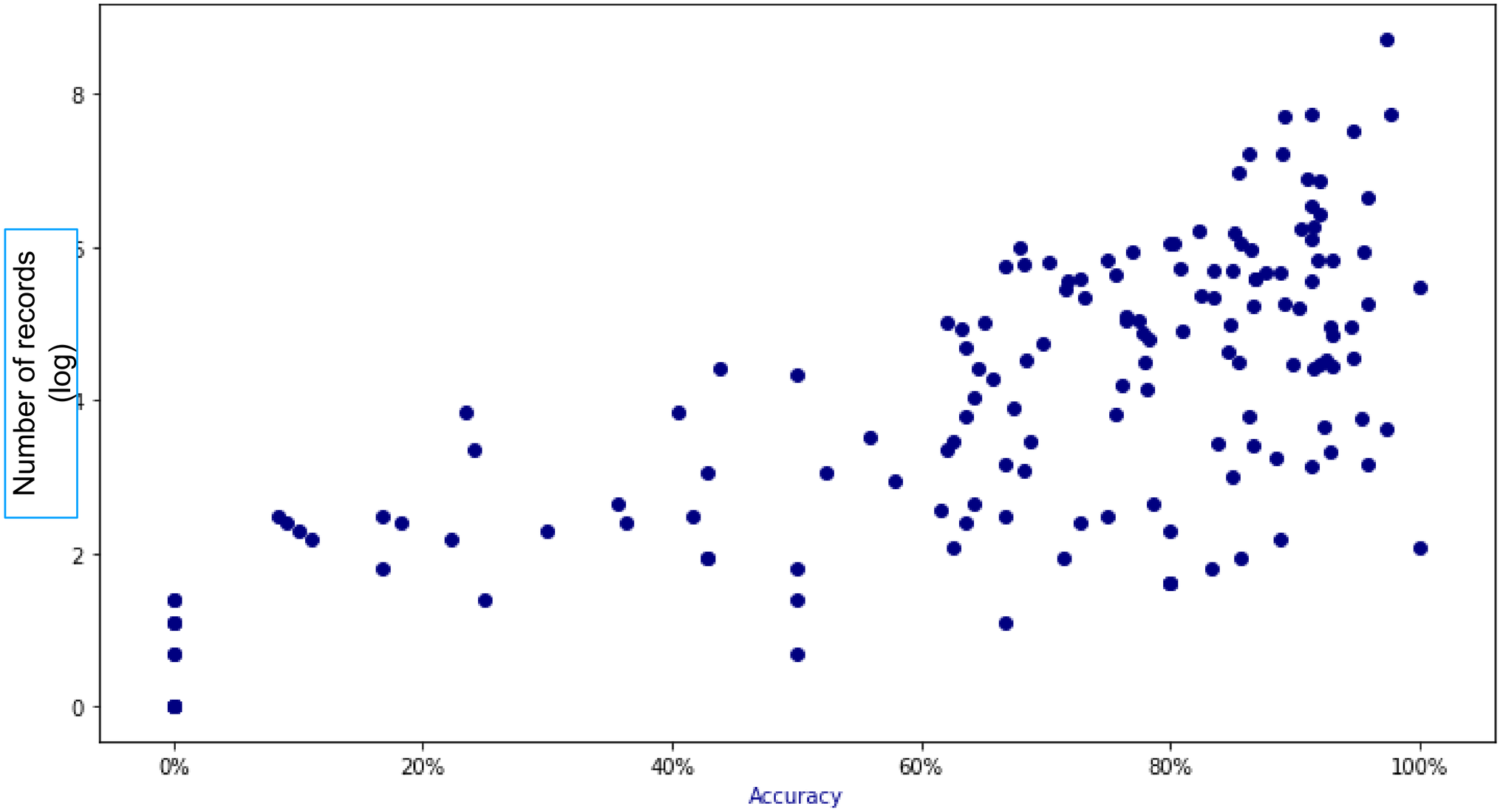
	Accuracy	Precision	Recall	F1
Ensemble with equal weights	**0.8905**	**0.6925**	**0.6149**	**0.6365**
Ensemble with different weights	**0.8921**	**0.6767**	**0.6420**	**0.6512**

Occupation (461 classes)

	Accuracy	Precision	Recall	F1
Ensemble with equal weights	**0.8447**	**0.6441**	**0.5384**	**0.5639**
Ensemble with different weights	**0.8505**	**0.6437**	**0.5637**	**0.5831**

Accuracy vs records coded automatically (percentage)

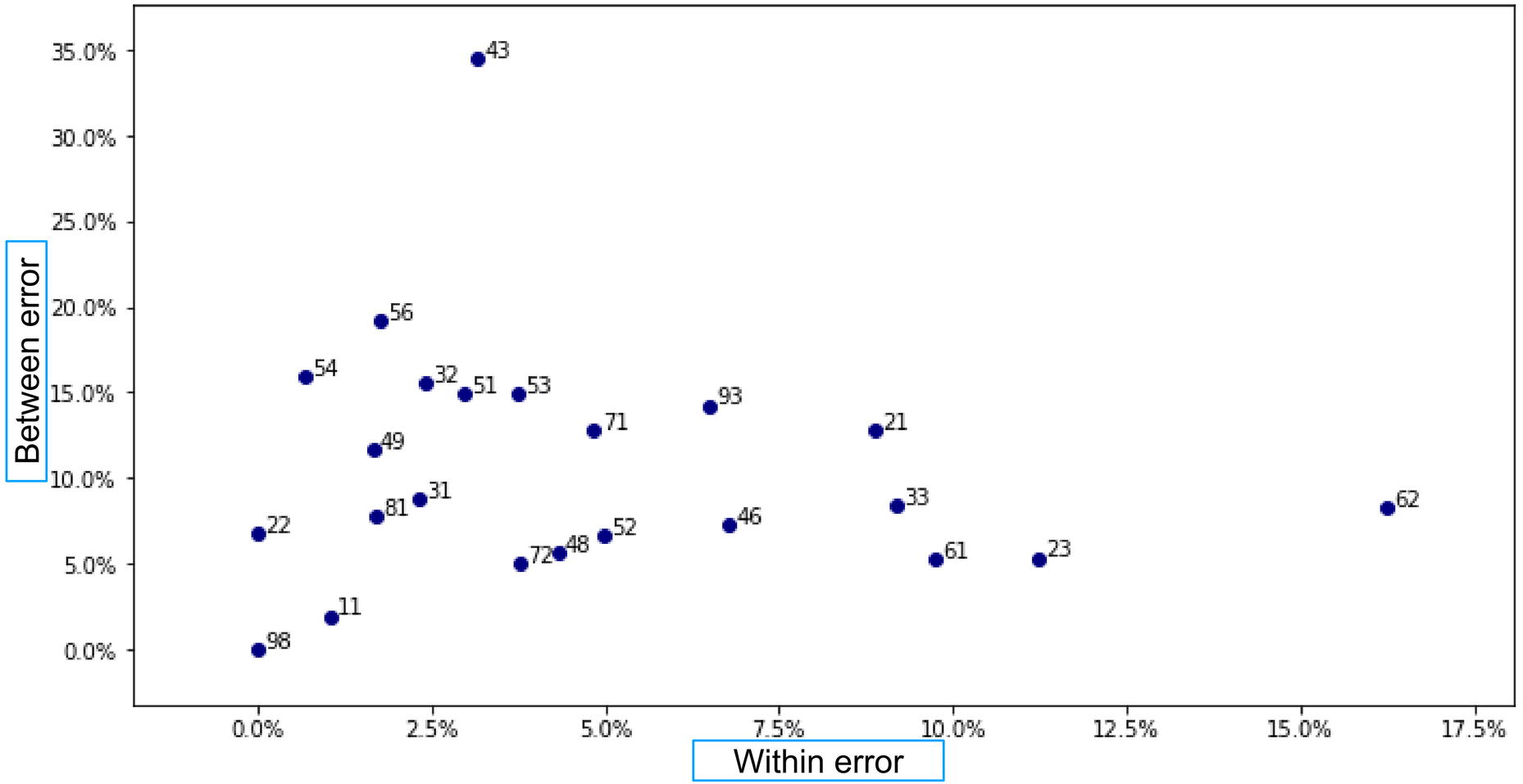


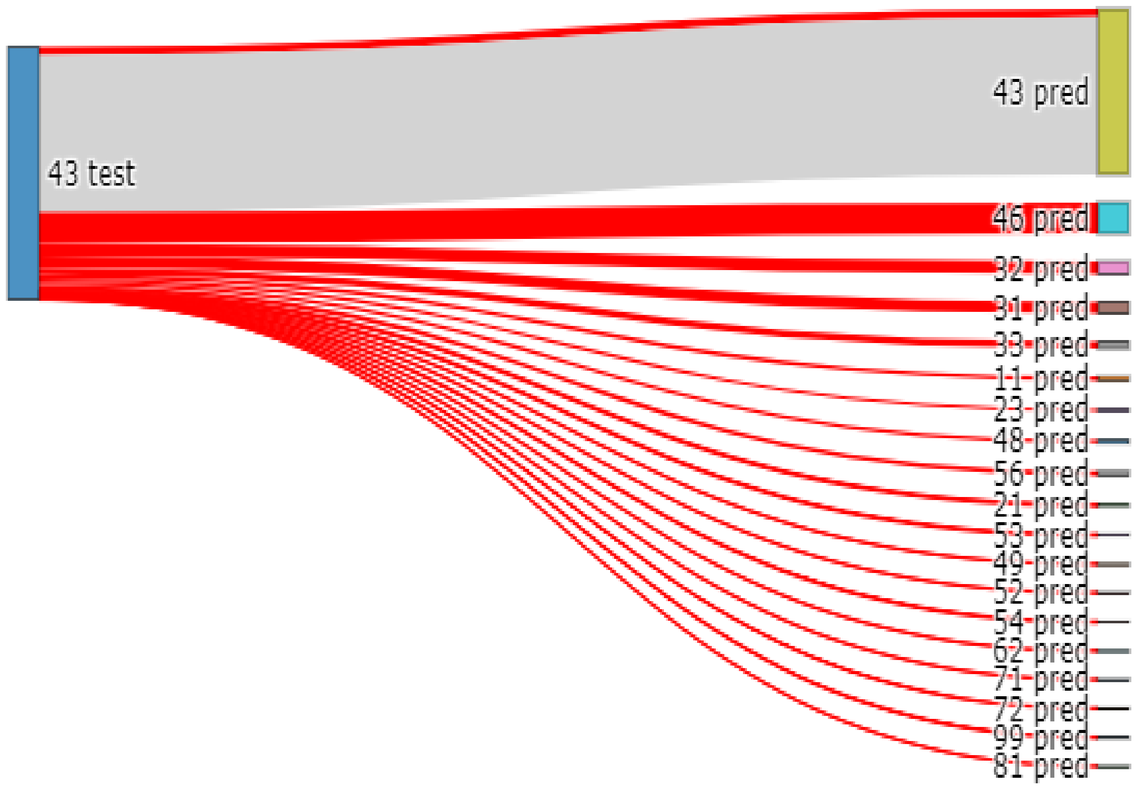

Accuracy vs number of records (log), by code
Number of records (log)
8
2
0
0%
20%
40%
60%
80%
100%
Accuracy

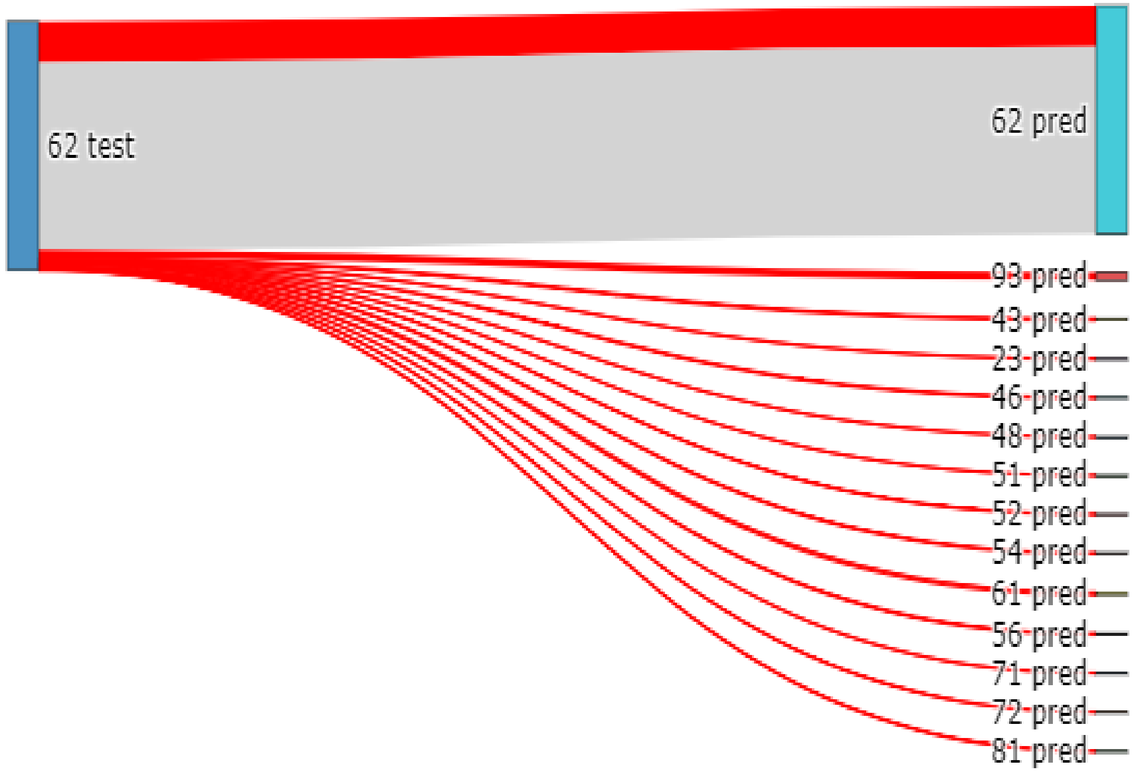
- The first two digits for each code (class) represent the sector which the code belongs to:
  - 7751, 7755: same sector

- Definitions:
  - Within-sector error: when the ML-algorithm codes a record it does it wrong, but the code it assigns does belong to the true sector: 4431 (real) vs 4434 (ML coding)
  - Between-sector error: when the ML-algorithm codes a record it does it wrong, and also the code it assigns does not belong to the true sector: 4431 (real) vs 3178 (ML coding)

Percentage of misclassified records
Between vs Within Errors by Sector




43 test
43 pred
46 pred
32 pred
31 pred
33 pred
11 pred
23 pred
48 pred
56 pred
21 pred
53 pred
49 pred
52 pred
54 pred
62 pred
71 pred
72 pred
99 pred
81 pred


62 test
62 pred
93 pred
43 pred
23 pred
46 pred
48 pred
51 pred
52 pred
54 pred
61 pred
56 pred
71 pred
72 pred
81 pred

Two-stage Automatic Coding

- First Stage:
  - Develop a SVM model on the training set in which we try to predict the Sector based on the textual features
  - Apply the above model to predict the Sector for the test set: **predicted_sector**
- Second Stage:
  - Develop a SVM model on the training set in which we try to predict the Class (4 digits) based on the following features: text + **real_sector**
  - Apply the above model to predict the Class for the test set, based on the following features: text + **predicted_sector** (first stage)

- We have not developed an ensemble algorithm for this two-stage approach

Economic Activity

	Accuracy	Precision	Recall	F1
One stage (6-grams, 2-words)	0.8793	0.6372	0.6760	0.6511
Two stages (6-grams, 2-words)	0.8774	0.6600	0.6452	0.6443

Occupation

	Accuracy	Precision	Recall	F1
One stage (6-grams, 2-words)	0.8188	0.5353	0.5918	0.5531
Two stages (6-grams, 2-words)	0.8312	0.5786	0.5730	0.5648

What is next

- Try state-of-the-art NLP methods for vectorizing
- Measure the performance on a new wave of records (ENIGH 2020)
- Develop a framework in which ML can be a part of the whole coding process and be subjected to quality measure

Conociendo México

01 800 111 46 34

www.inegi.org.mx

atencion.usuarios@inegi.org.mx

INEGI Informa