Artificial intelligence as a complementary methodology for coding tasks

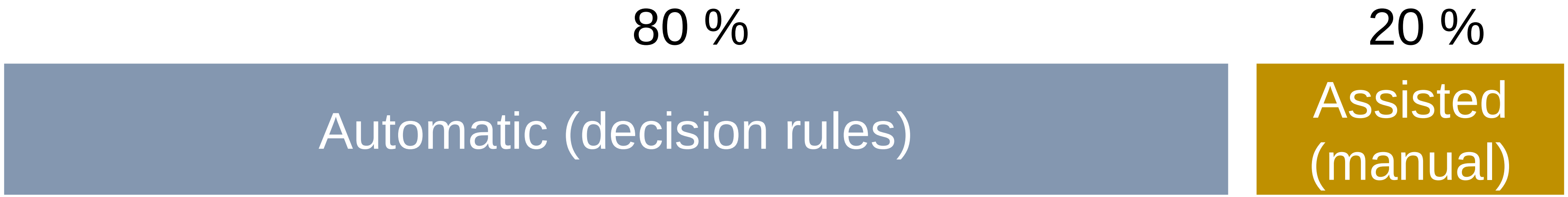
Alejandro Ruiz
Jael Pérez

INEGI

December, 2024

Occupation & Economic activity

Two coding strategies:



80 %

Automatic (decision rules)

20 %

Assisted (manual)

ENOE (quarterly):

- 40 000 records
- 260 coders
- Decentralized

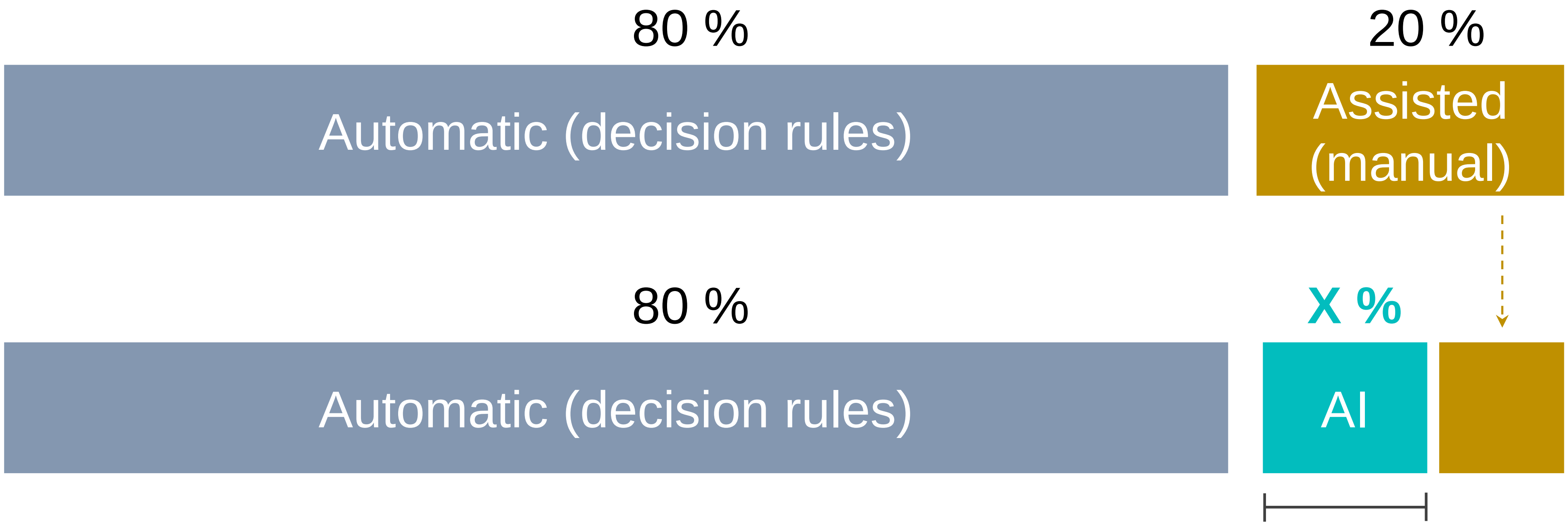
ENIGH (biennial):

- 30 000 records
- 10 coders
- Centralized

Census or Intercensal survey (quinquennial):

- + 1 million records
- 600 coders
- Centralized

Not two, but rather three strategies.



80 %

Automatic (decision rules)

20 %

Assisted (manual)

80 %

Automatic (decision rules)

X %

AI

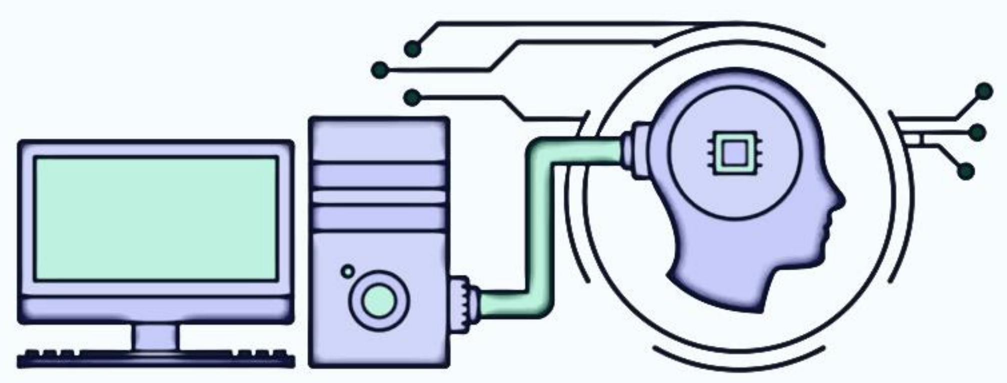
Those records for which the AI algorithm is "relatively confident" about what it is doing.

Supervised AI algorithms

Training set

Text 1 → code "A"
Text 2 → code "A"
Text 3 →code "C"
Text 4 → code "F"
…
Text n → code "D"

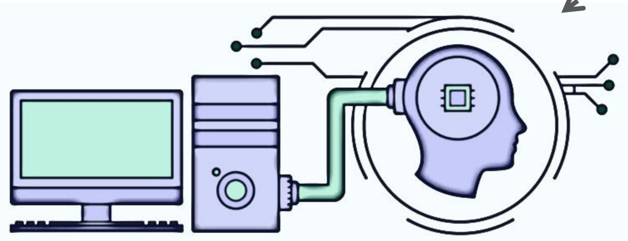
Transformers



https://datascientest.com/es/deep-learning-definicion

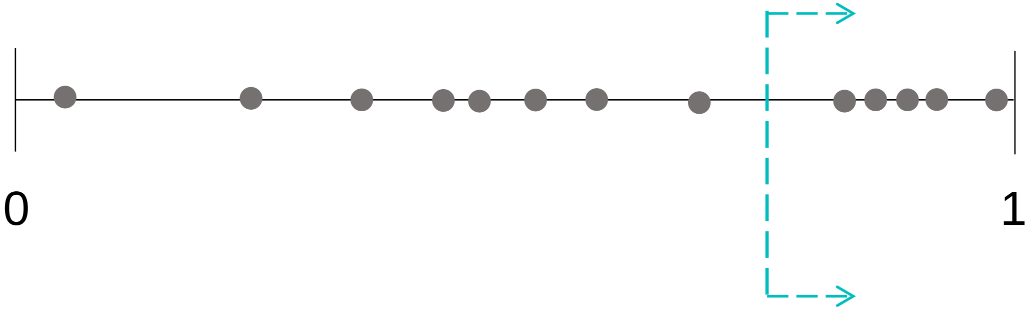
Supervised AI algorithms

New text



1) Code
2) Score

Score



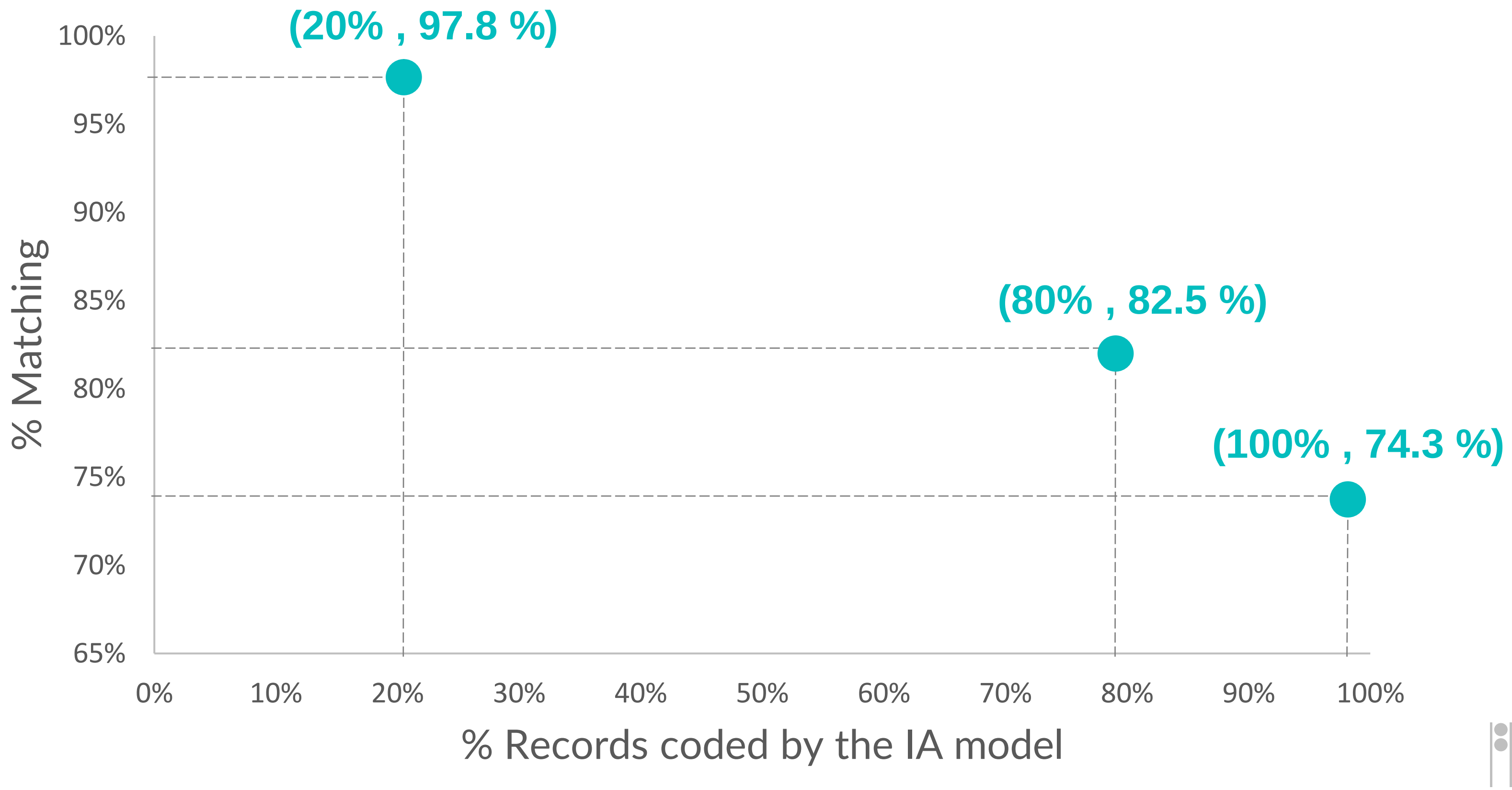
Threshold

Trade-off:

% records
vs
% matching

Trade-off in the test set

Def. **Matching percentage**: % records for which the IA code is the same as the Manual code.



ENIGH Results

ENIGH

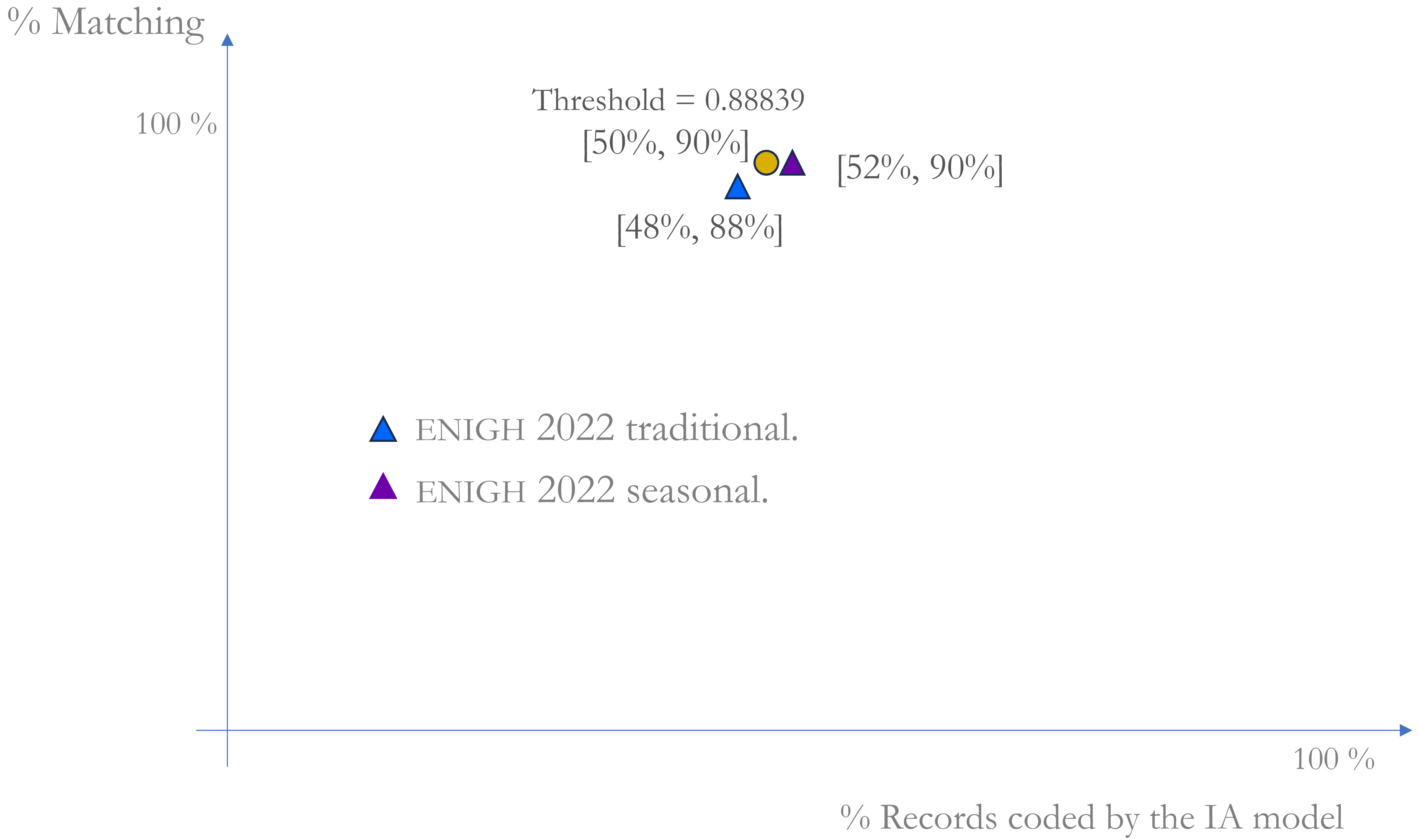
ENIGH 2016, ENIGH 2018 & ENIGH 2020:

- Model training: 100% of records automatically coded + 80% of records manually coded.
- First test (test set): 20 % of the remaining manually coded.

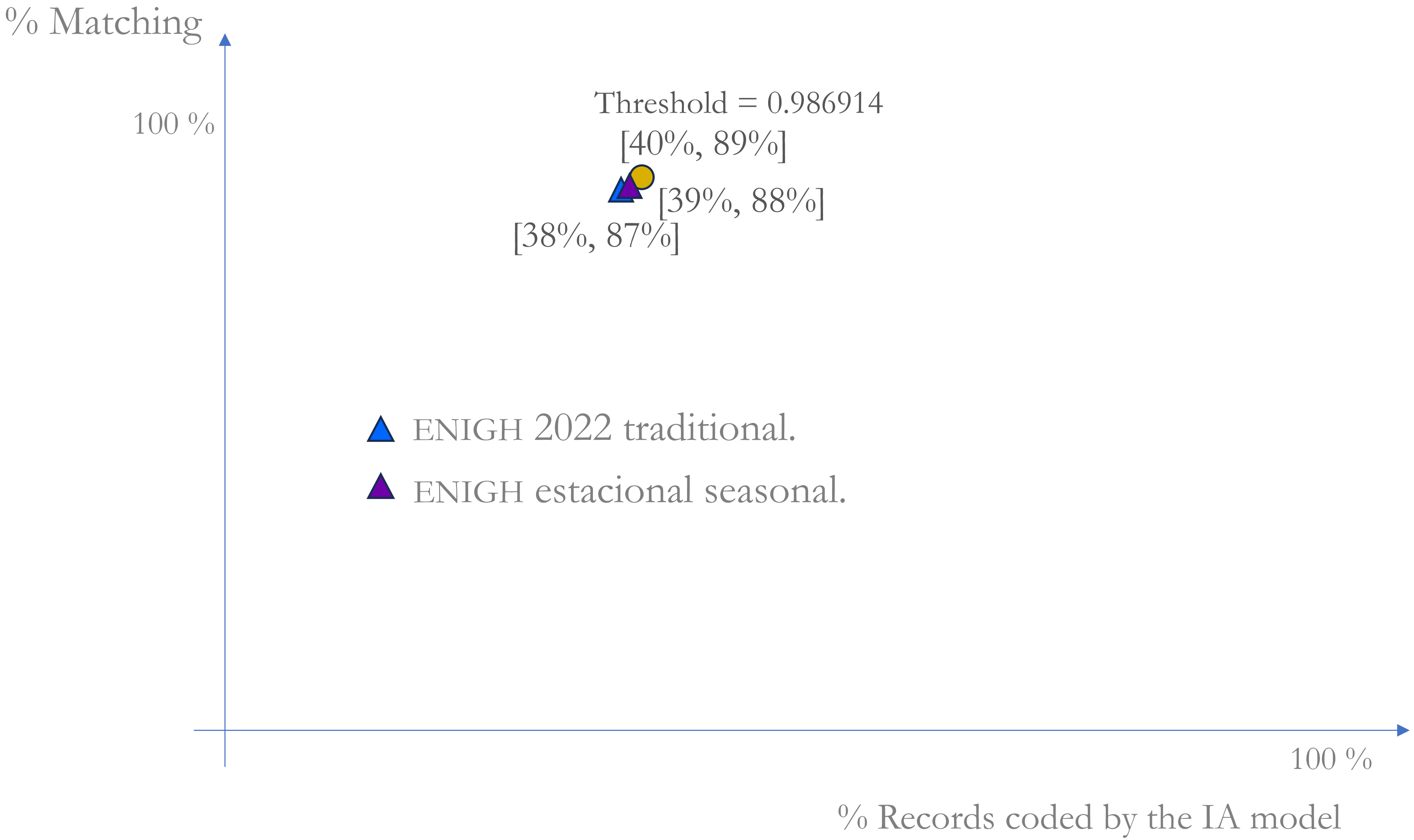
Model evaluation:

- ENIGH 2022 Traditional
- ENIGH 2022 Seasonal

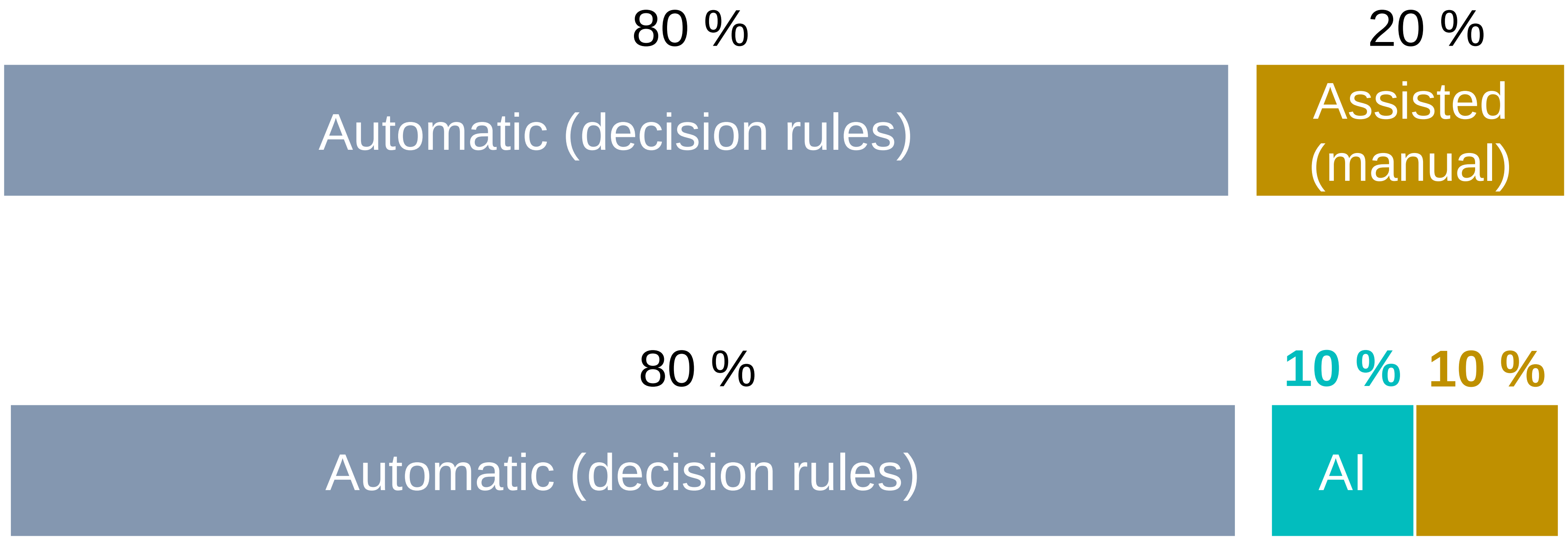
Prediction vs Evaluation, Economic activity



Prediction vs Evaluation, Occupation



ENIGH



80 %

20 %

Automatic (decision rules)

Assisted (manual)

80 %

10 %

10 %

Automatic (decision rules)

AI

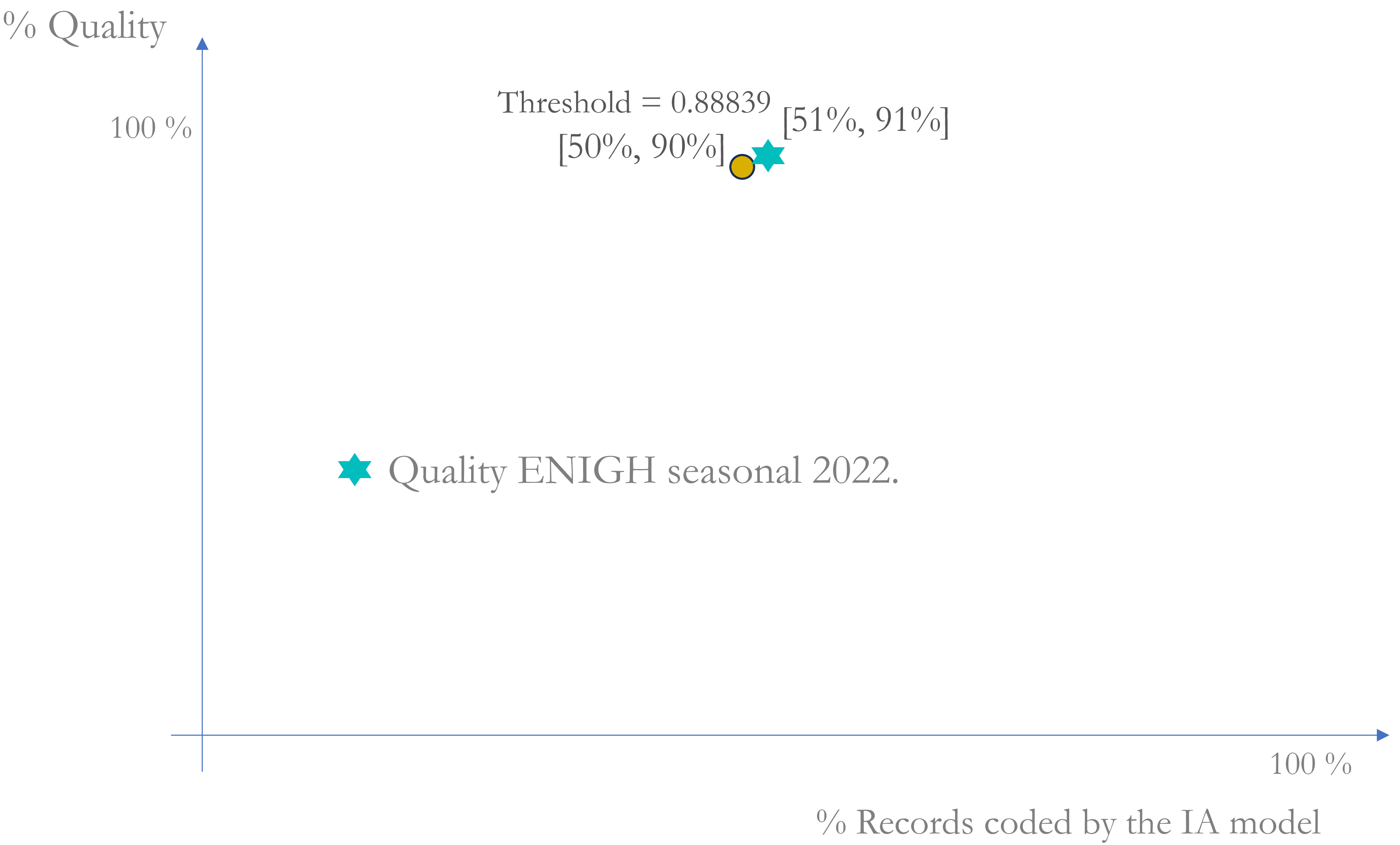
ENIGH coding quality

Matching is not the same as quality.

To **measure quality**:

- Coding experts validated the results of the AI model (ENIGH seasonal).
- Def Quality percentage: % of records in which the AI code matches the experts' code.

AI coding quality, Economic activity



ENOE Results

ENOE goal

Goal: to increase the coding quality for Occupation and Economic activity variables.

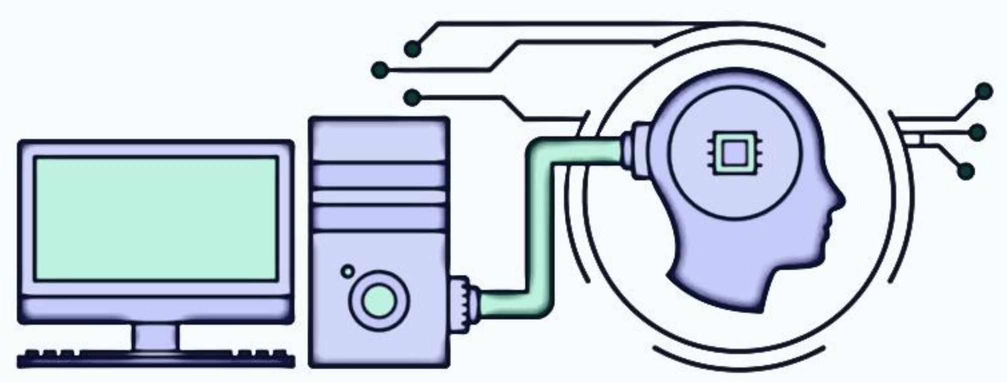
- ENOE is the most widely used for labor-related topics.
- It is the largest continuous household survey.
- Manual coding is carried out by each of the 32 states in the country:
  1. *There is significant quality variance across states.*
  2. *The mean quality is inferior to other similar surveys.*
- INEGI invested in a ground-truth database to train the model.

ENOE

Training set =
Ground-truth dataset

Text 1 → code “A”
Text 2 → code “A”
Text 3 →code ”C”
Text 4 → code “F”
…
Text n → code “D”

Transformers



https://datascientest.com/es/deep-learning-definicion

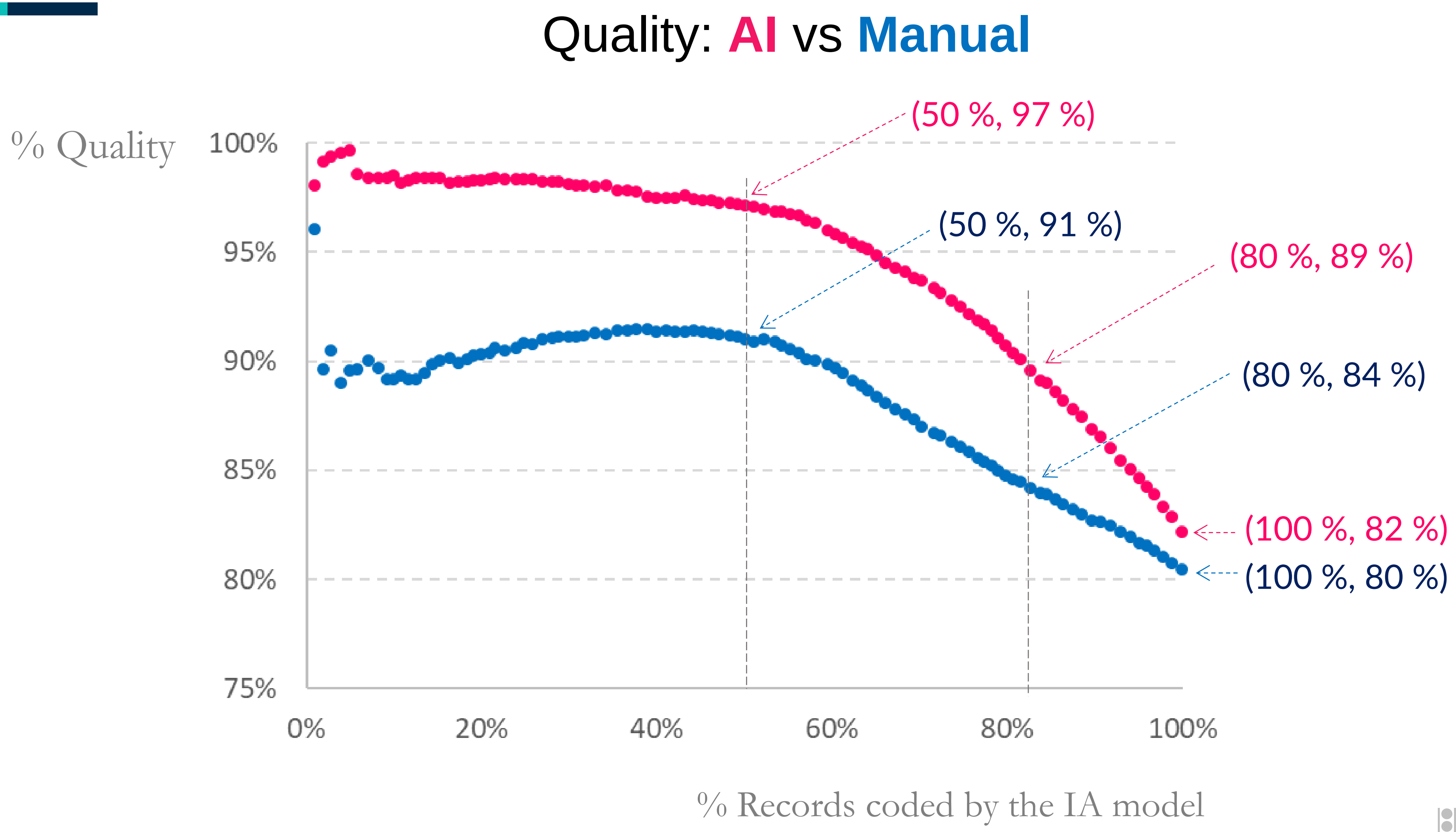
ENOE

This ground-truth dataset allows two measurements:

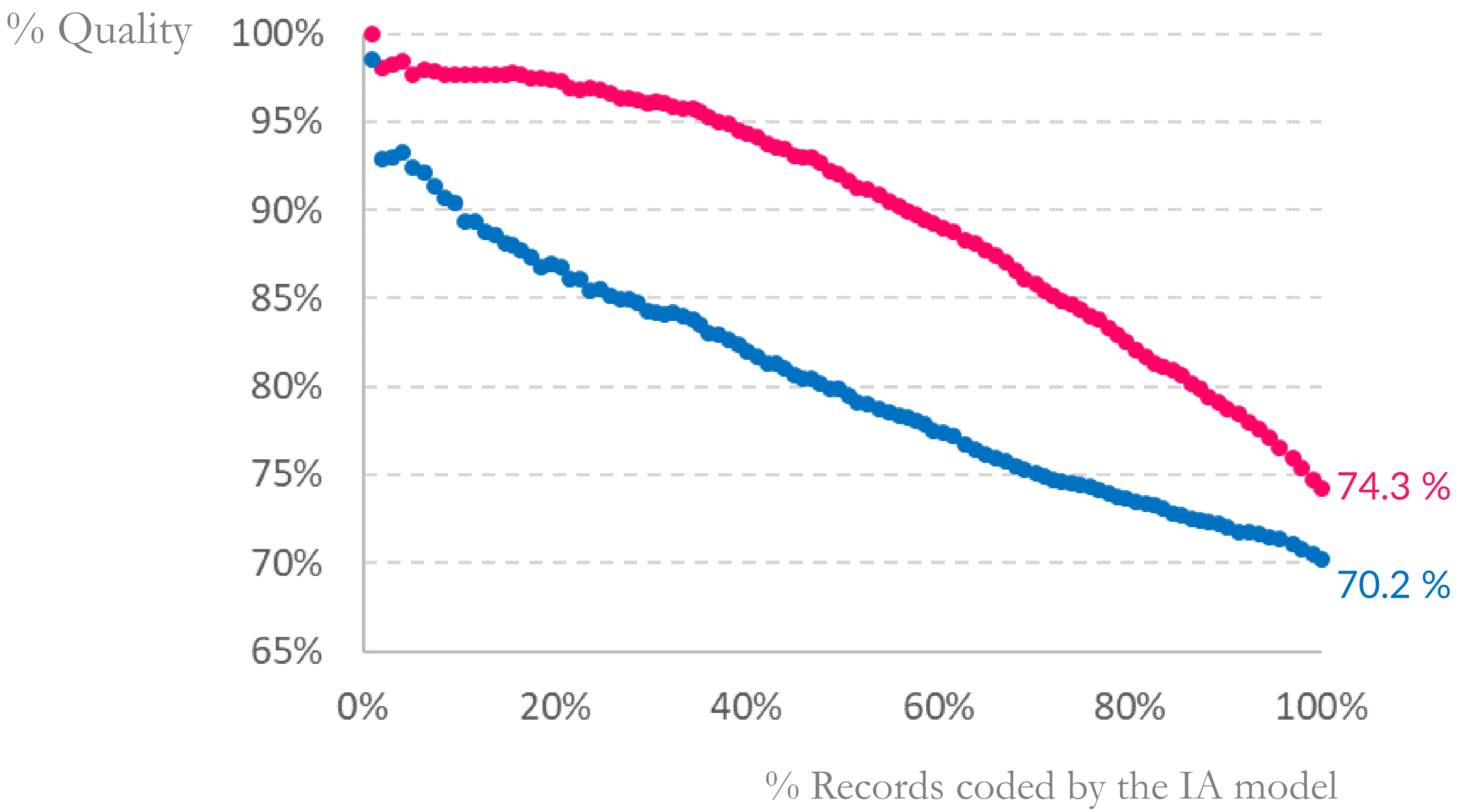
1) AI quality: AI code vs Experts' code

2) Manual quality: Manual code vs Experts'code

ENOE, Economic activity



ENOE, Occupation



Conclusions

- The single most important element in those AI models is the **input database**.
- **For the ENIGH**, we aimed to replicate the **original manual patterns**; thus, we used the original databases to train the AI model.
- **For the ENOE,** we aimed to assess whether we could achieve **better quality** than the manually coded version.

  i. Yes, but we need a high-quality input database.

  ii. The quality gap between the manual coding and the AI coding could be significant, leading to potential changes in the distribution of Occupation and Economic activity.

  iii. We are continuing to evaluate in order to draw more robust conclusions.

Thanks!

INEGI