TALL
SOM FORTELLER

Machine learning in text classification at Statistics Norway

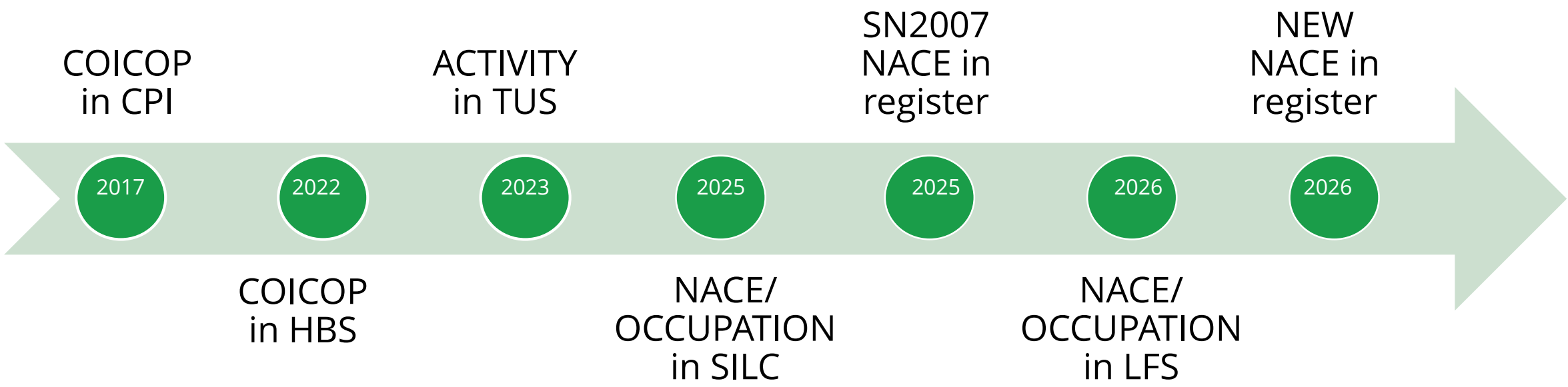
BORISKA TOTH, SUSIE JENTOFT, RUBEN MUSTAD

AIML4OS – WP10

22-01-2025

Statistisk sentralbyrå
Statistics Norway

ML classification projects



Statistisk sentralbyrå
Statistics Norway

Masters and research projects

Statistisk sentralbyrå
Statistics Norway

Norway's HBS and TUS

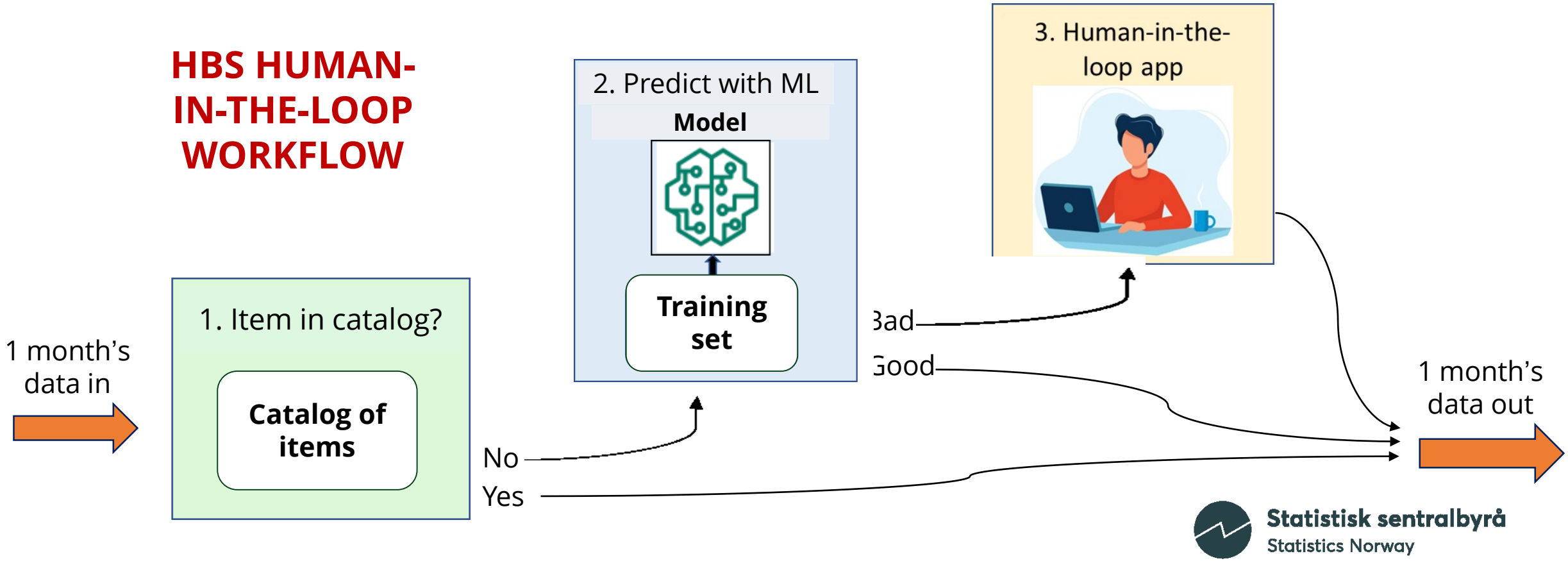
	HBS	TUS
Time period	2022	Aug 2022-Aug 2023
Size of survey	12000 people (1 week)	8800 people (2 days)
Data collection	Scanned receipts + OCR, or manually in web app	Manually in web app
Coding scheme	COICOP 2018	ACL 2018
Num. of classes	338	145
Example of code	"apple juice 1L" → 01.2.1.0	"surfing the net"→ 551
Num. items to code	~60,000 unique	~70,000

Statistisk sentralbyrå
Statistics Norway

1) Check in the catalog if an item is already coded

2) Machine learning: use for “good” predictions (23,000 / 60,000 unique items)

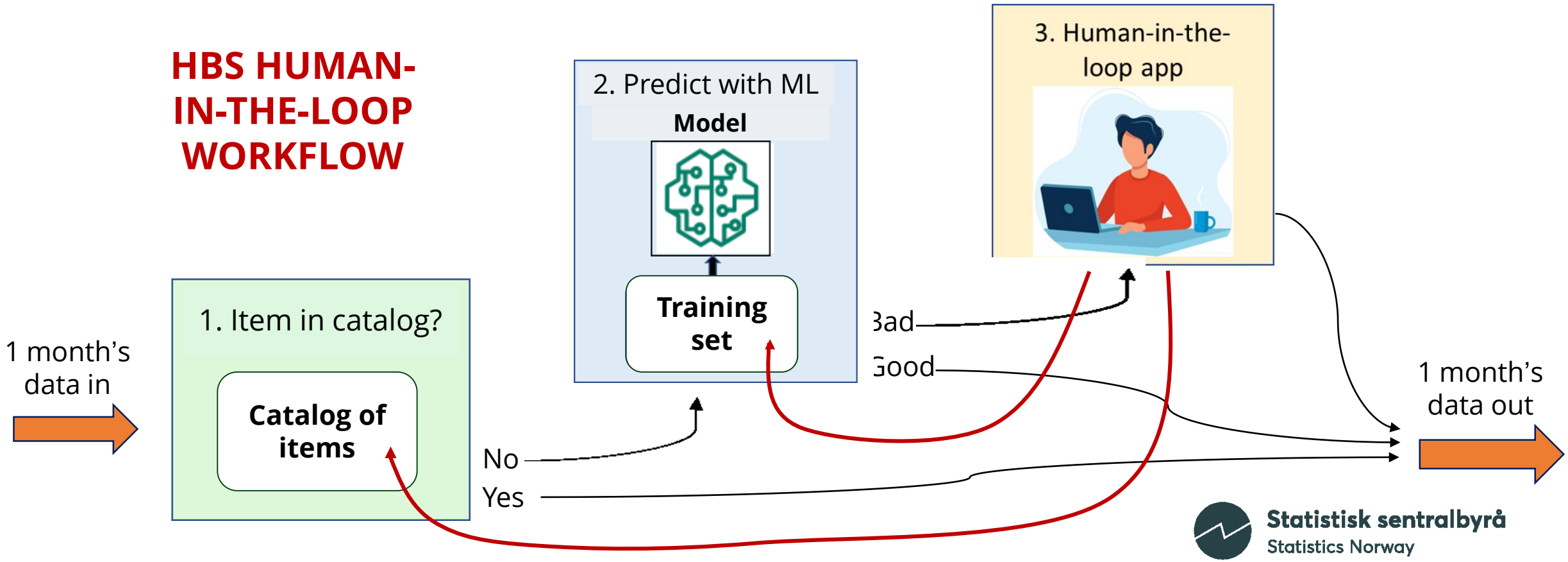
3) Human-in-the-loop app: use for “bad” predictions



1) Check in the catalog if an item is already coded

2) Machine learning: use for “good” predictions (23,000 / 60,000 unique items)

3) Human-in-the-loop app: use for “bad” predictions



Machine learning methods in HBS

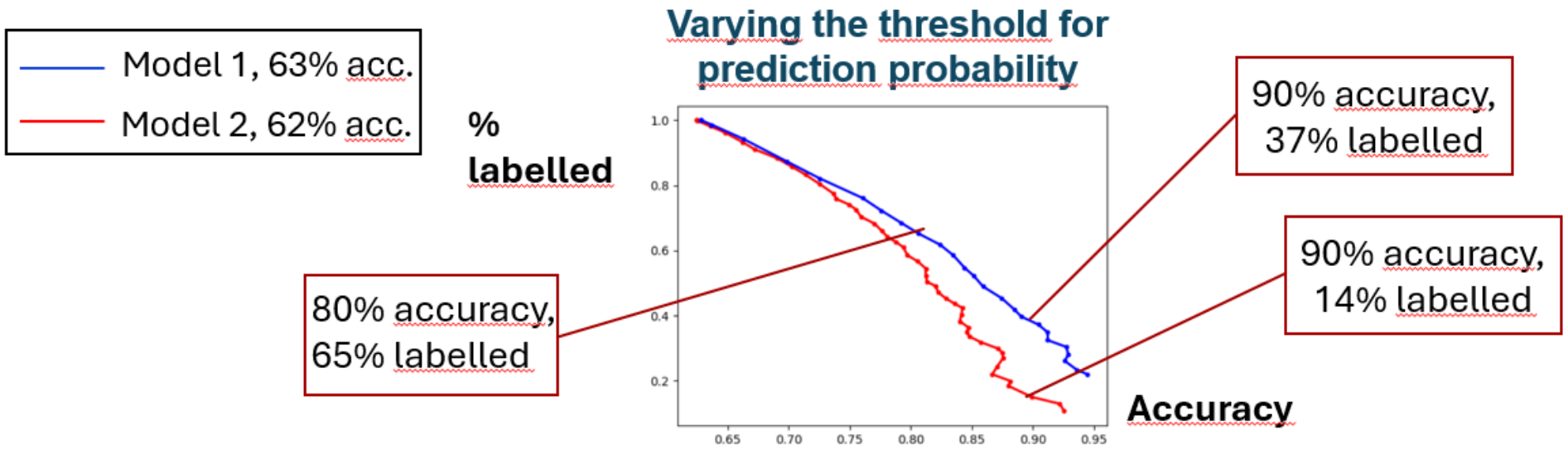
- 60,000 item training set
  - CPI catalog, manually coded, COICOP dictionaries
- Variables- 2-grams and 3-grams worked best for text features

Item name	Store name	Price	isGrocery	weight_quantity
Reebok shoes size 42	XXL Drammen	799	False	1.0
⇩	⇩			
1_ree, 1_eeb, 1_ebo, ….	1_XXL, 1_Apotek, ….			

- Tested SVM, logistic regression, and random forest on 1500 manually labelled items (pilot + survey)
- Chose random forest (top performance and computationally easy)
- 63% accuracy, 63% F1-score overall

Statistisk sentralbyrå
Statistics Norway

Machine learning methods in HBS



- Prediction probabilities- fitted model for probability of a prediction being correct
- Good prediction probabilities are more important than the overall accuracy of a model for replacing manual coding!

Statistisk sentralbyrå
Statistics Norway

Lessons learned from HBS 2022

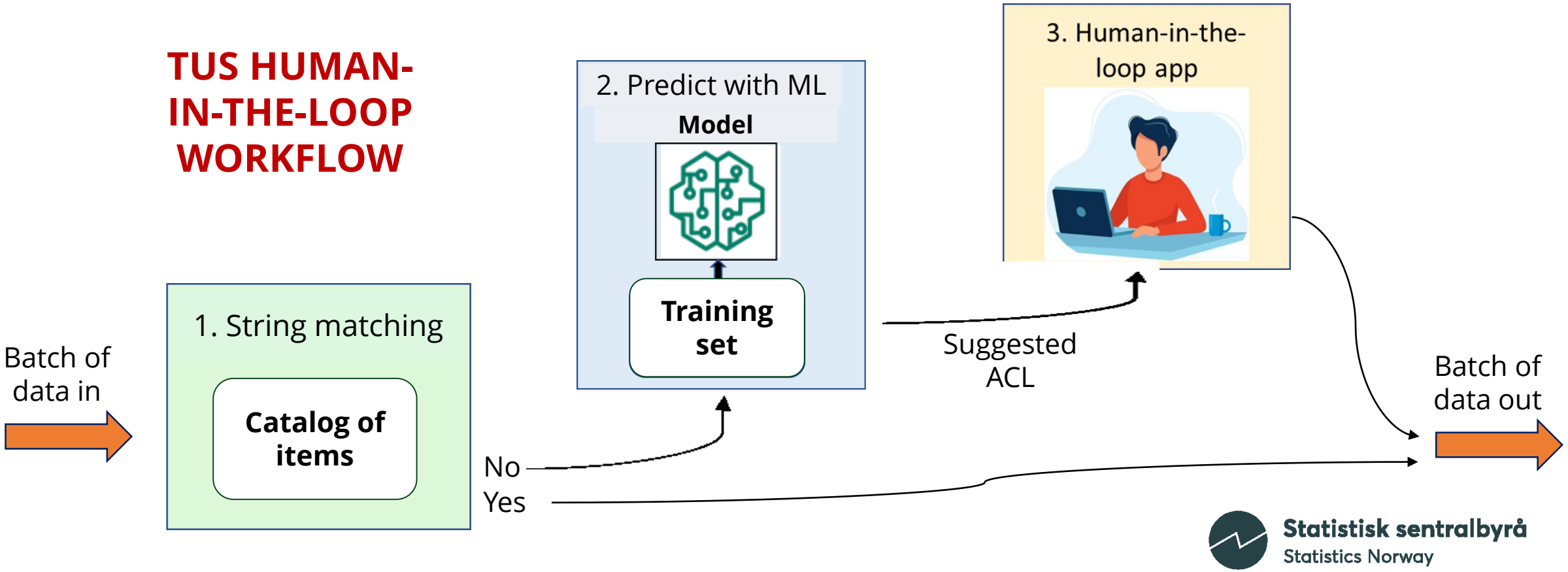
- Machine learning: traditional methods were decent
  - But huge potential to do better with LLM's
  - A measure of the system's confidence in its predictions is crucial (WP10 literature + cluster reports)
- Design of the human-in-the-loop system: good and less good decisions
  - Interface was very effective for human coding
  - Better design decision: what to store as separate items («orange juice 1l» vs «orange juice 2l»)
- Other needs for machine learning beyond coding
  - Learning the layout of receipts for better data quality
  - Imputation
  - Editing?

1) Use string matching to check if an activity is in the catalog of coded items

2) If not, predict ACL with machine learning

3) Predicted ACL assists with human coding in the human-in-the-loop app

---



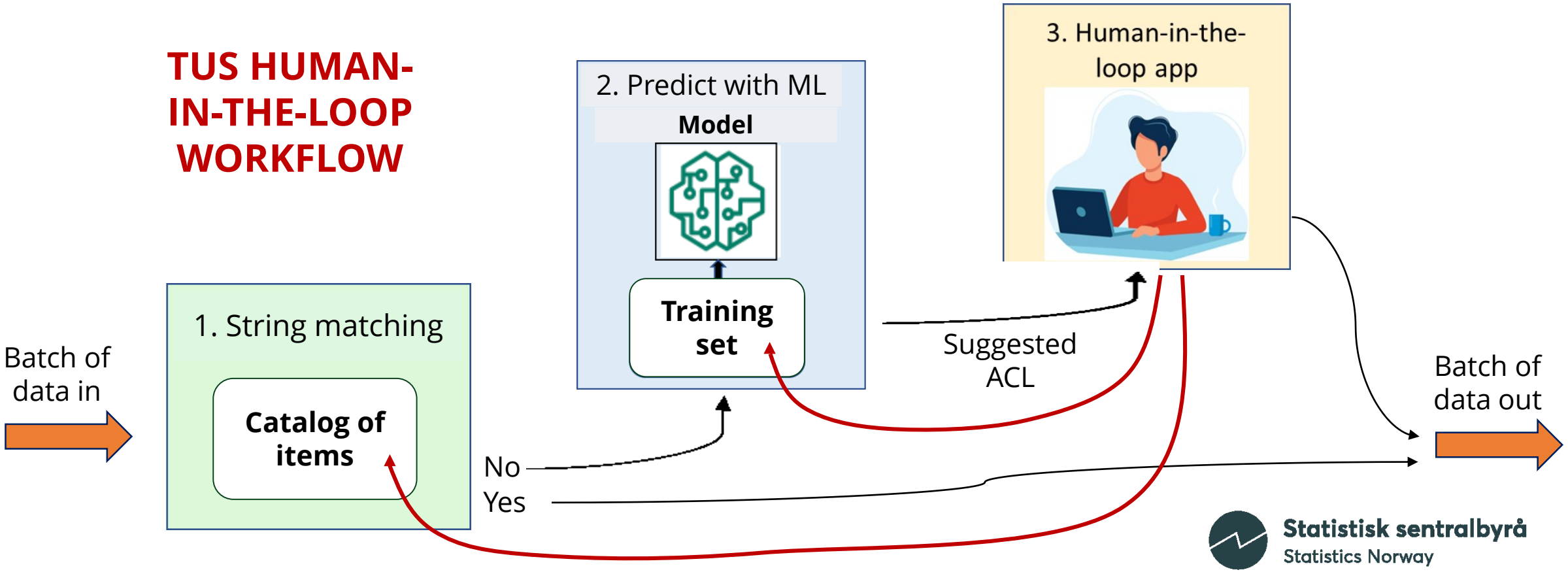
Statistisk sentralbyrå
Statistics Norway

1) Use string matching to check if an activity is in the catalog of coded items

2) If not, predict ACL with machine learning

3) Predicted ACL assists with human coding in the human-in-the-loop app

---



Statistisk sentralbyrå
Statistics Norway

Machine learning methods in TUS

- Training set (~17,000 items) and test set (~5,000 items)
- String matching procedure to check if an activity has already been coded
  - Find percent similarity between two strings based on edit distance
  - Rule: at least 3 matches in the catalog (and TypeOfTravelTo matches)

- If no match is found, use machine learning:

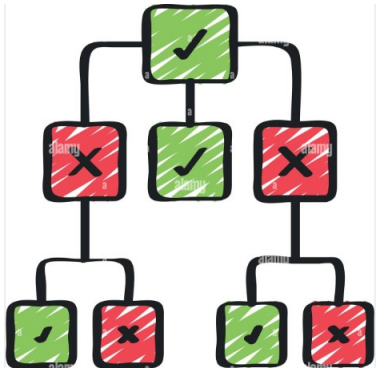
1) Create word embedding from activity text with GloVe
   - Map each word to a 100-dimensional vector capturing semantics
   - Take the sum of these vectors over all words in the activity text

Machine learning methods in TUS

- If no match is found, use machine learning:

2) Classification using gradient boosting

3) a



Use as features:

  - Text of activity as word embedding, AND:
  - Other variables: time of day, day of week, duration of activity, type of travel to activity, etc

- Classifier used was gradient boosting

  XGBoost + LightGBM, with soft voting

- Results: 77% F1-score on the test set

Statistisk sentralbyrå
Statistics Norway

Lessons learned from TUS 2022/23

- Overall a very helpful system
- Sharing of resources with HBS

- Context (activities before and after) would help, computational challenges
- Many special cases handled poorly
  - Seasonal activities, multiple activities, named entities, English text
- Limitations of GloVe representations of words
  - Summing vectors across words in the text, out-of-vocabulary words

⟹ Use LLM's!

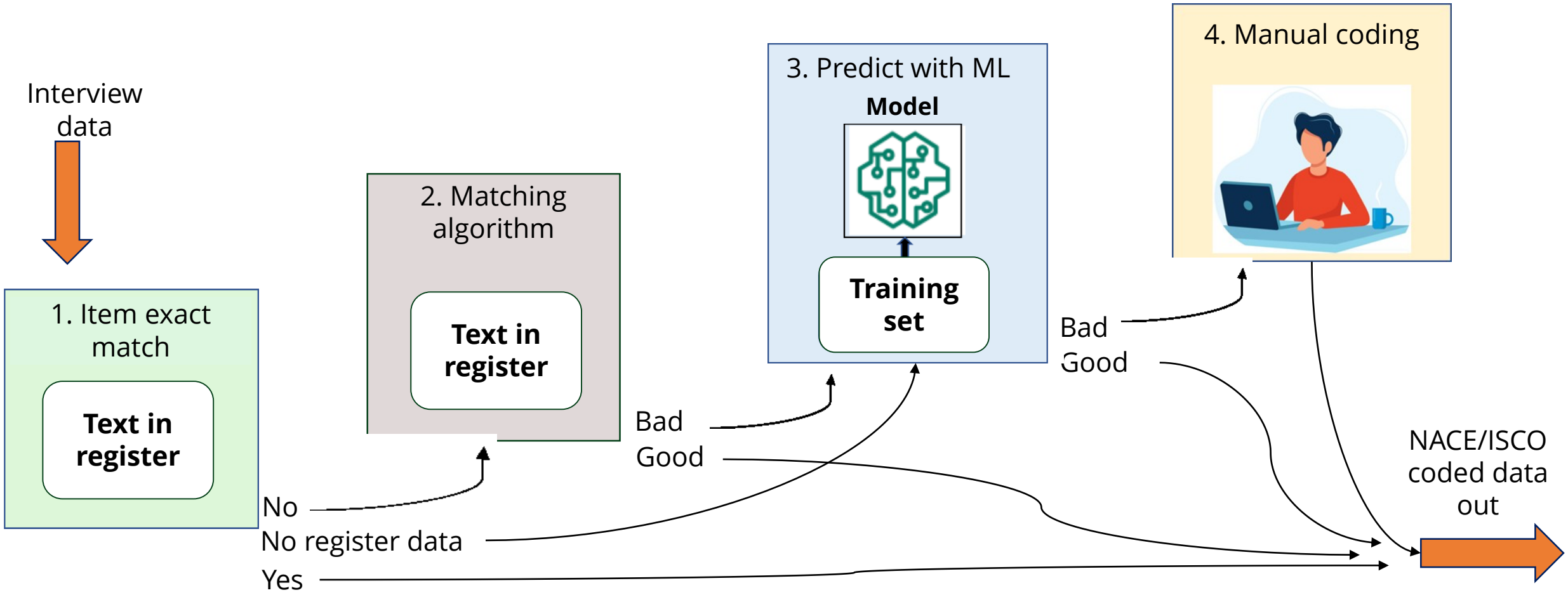
Statistisk sentralbyrå
Statistics Norway

Classification in SILC

Current - 2024

- Register data on employee and occupation linked before intervju
- NACE & ISCO coded after intervju (2-digit only):
  - Some direct matches coded directly
  - Rest manual (800 NACE and 4500 ISCO)
- Double-coding study: 7%-18% difference in 2-digit occupation

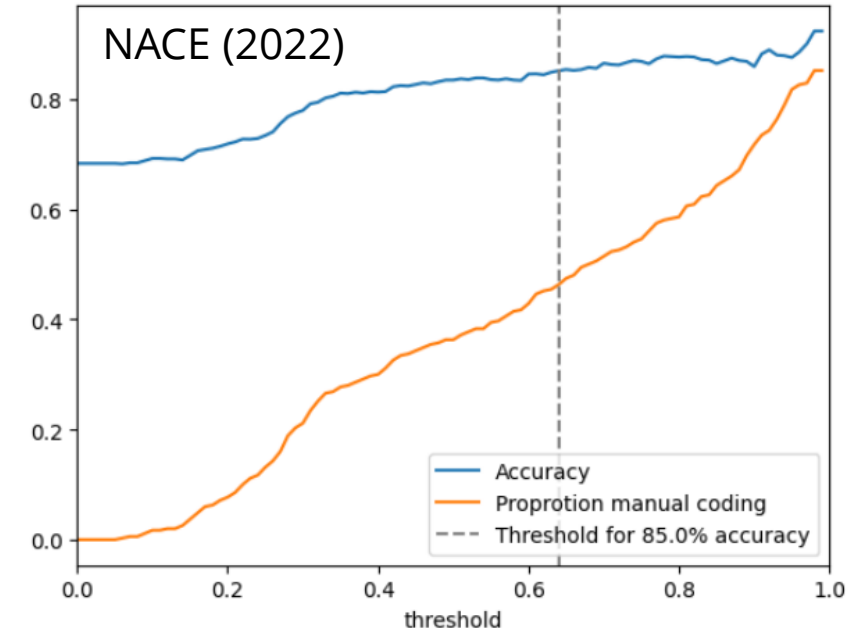
Classification in SILC from 2025

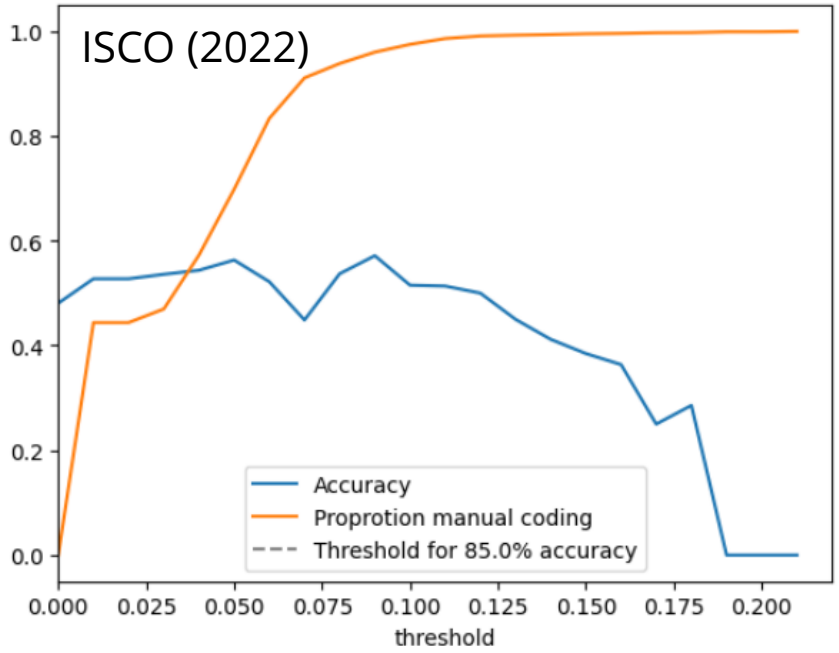


Statistisk sentralbyrå
Statistics Norway

Matching

- Good, monthly register data on employees, companies - delayed
- Relevant register info linked after
- Ratcliff-Obershelp algorithm from *difflib* (recursive longest common match)

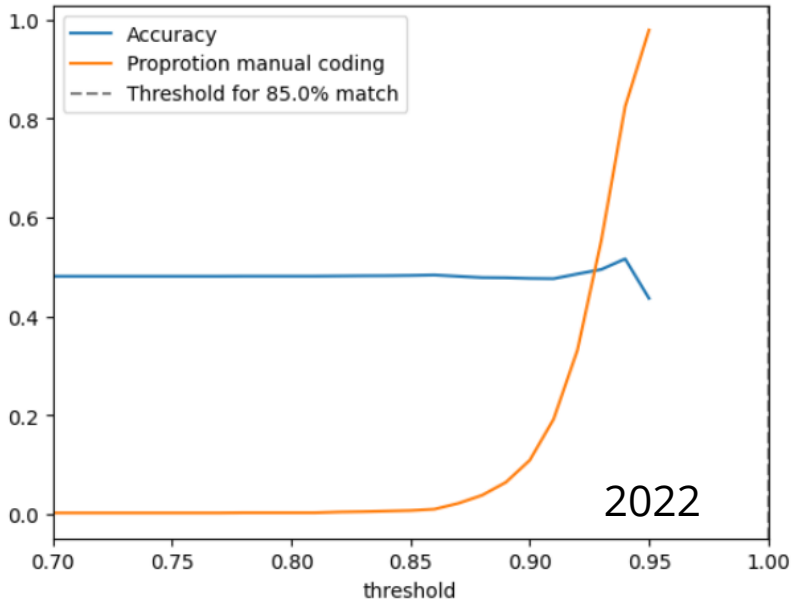




Matching - occupation

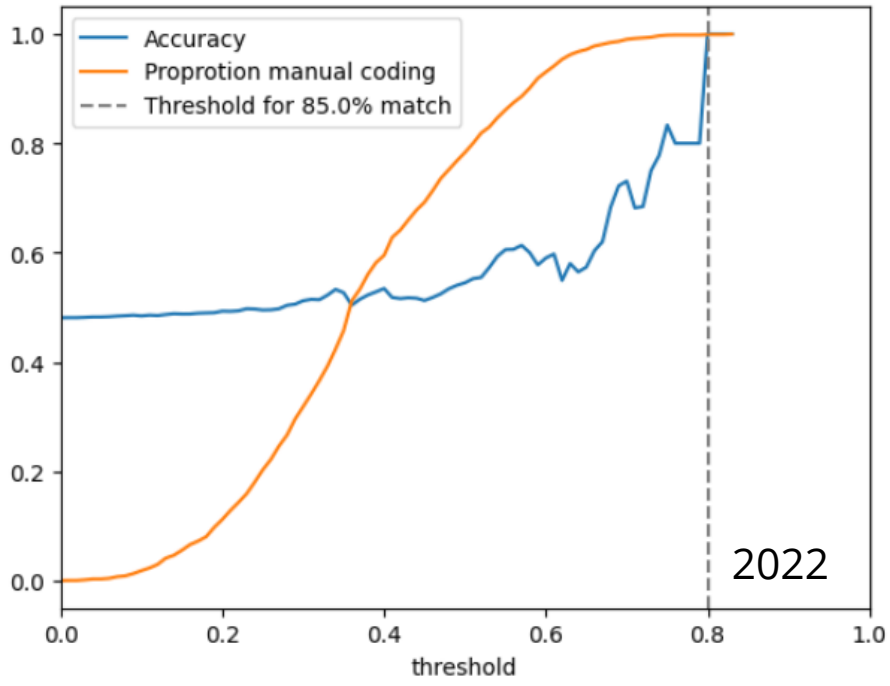
- Word embeddings + cosine similarity
- Averaged for text string

NorBERT



- Sentence embeddings + cosine similarity

distiluse-base-multilingual-cased-v2



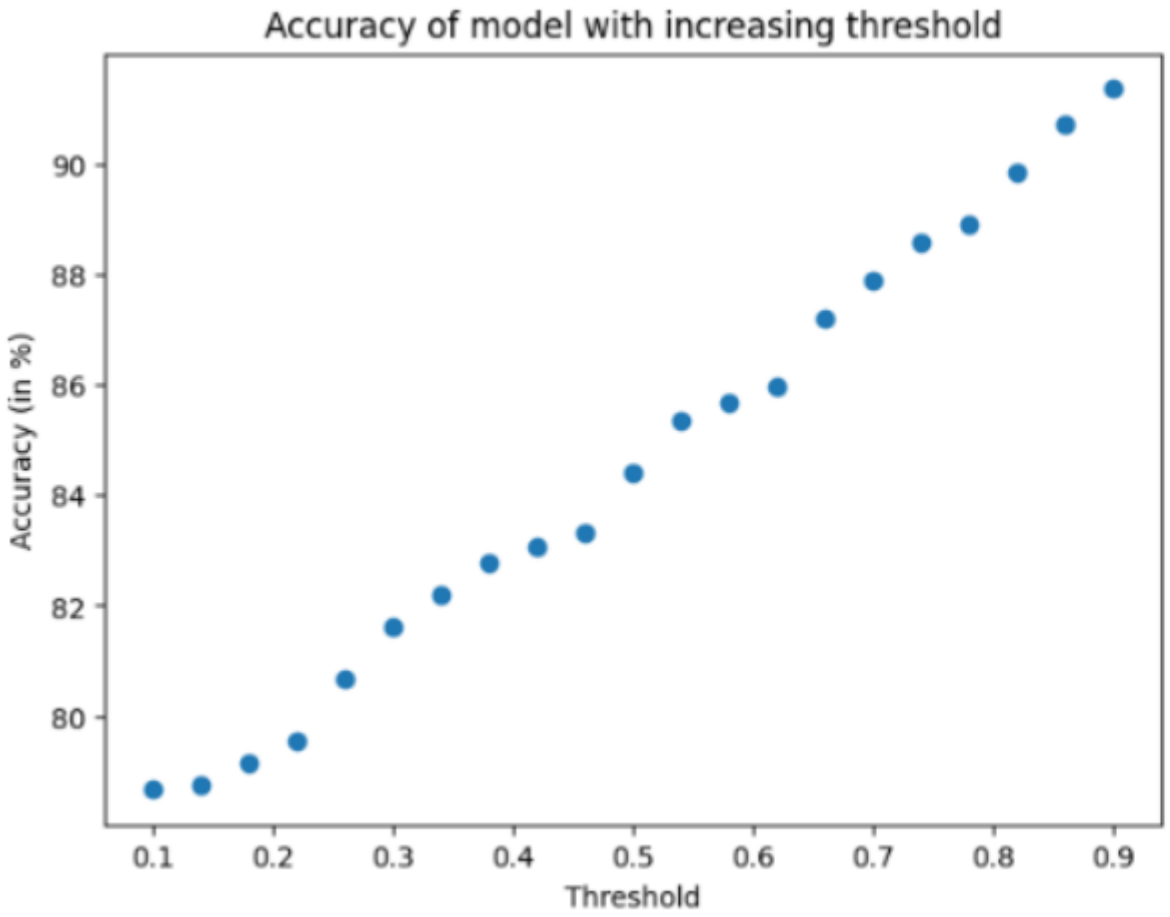
Prediction algorithms - NACE

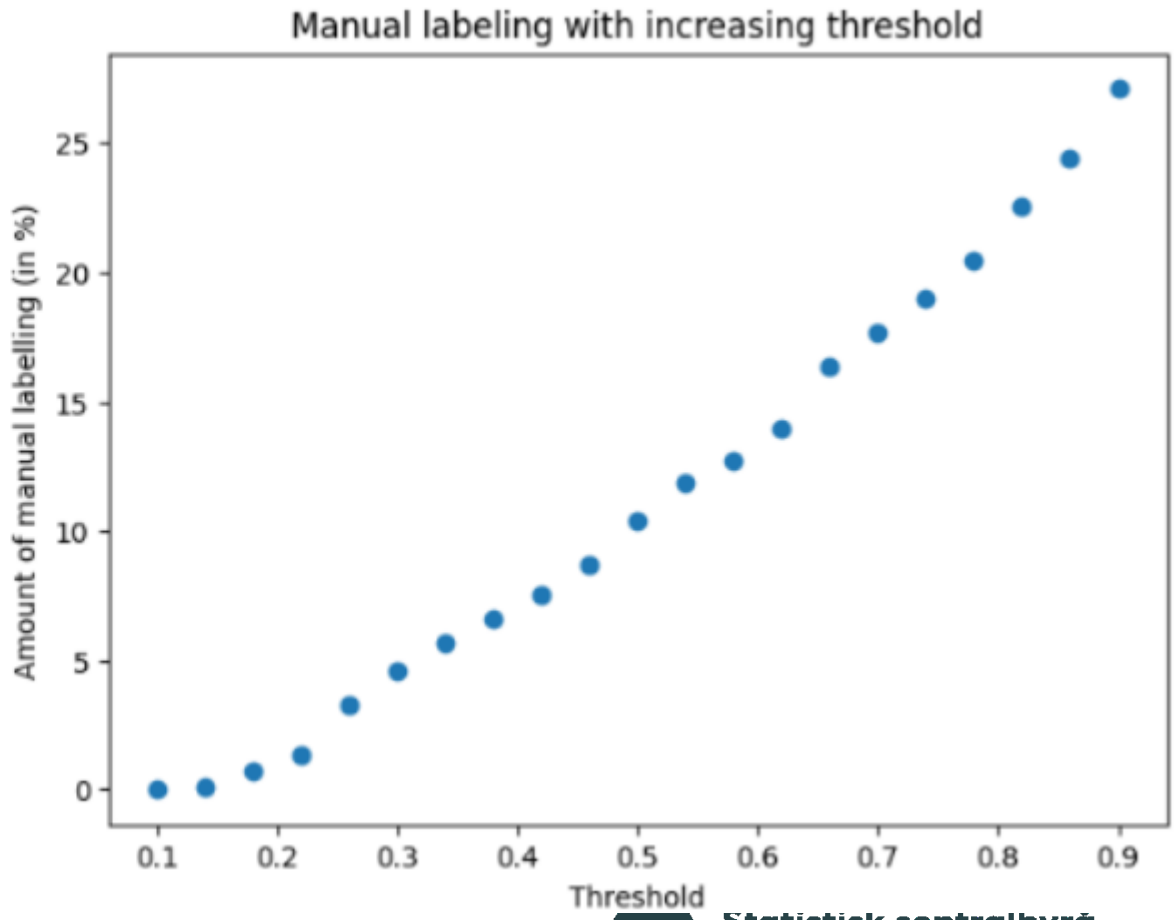
Embeddings + Model	Accuracy	Weighted F1-score	Macro F1-score
Word2Vec + LightGBM	56.7%	57.5%	48.6%
Bag of words + LightGBM	62.8%	63.8%	53.6%
BERT + SVC	63.7%	64.2%	55.2%
Tf-idf + LightGBM	67.3%	67.9%	56.6%
BERT + LightGBM	72.2%	72.8%	62.8%
Tf-idf + SVC	75.1%	75.7%	66.8%
Fine-tuned BERT + XGBoost	77.3%	77.4%	70.2%
Fine-tuned BERT + LightGBM	**78.7%**	**79.1%**	**71.7%**

NACE - modelling

- Pre-trained BERT on Norwegian text
- Combining it with XGBoost/LightGBM
- Survey + company activity (Brønnøysundregisteret)
  - Varying quality, from one word to a few sentences
- "other"-class for classes with few samples
- XGBoost/LightGBM were tuned, where best parameters on validation set

Threshold for NACE



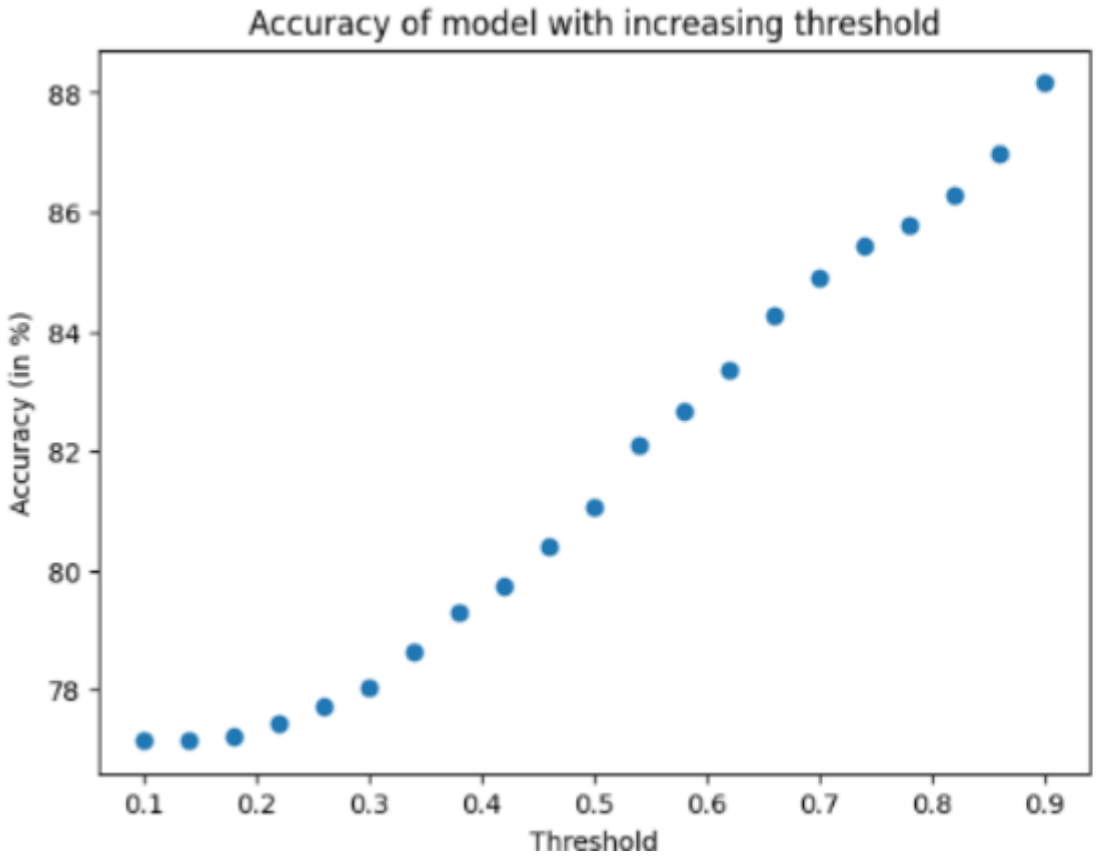


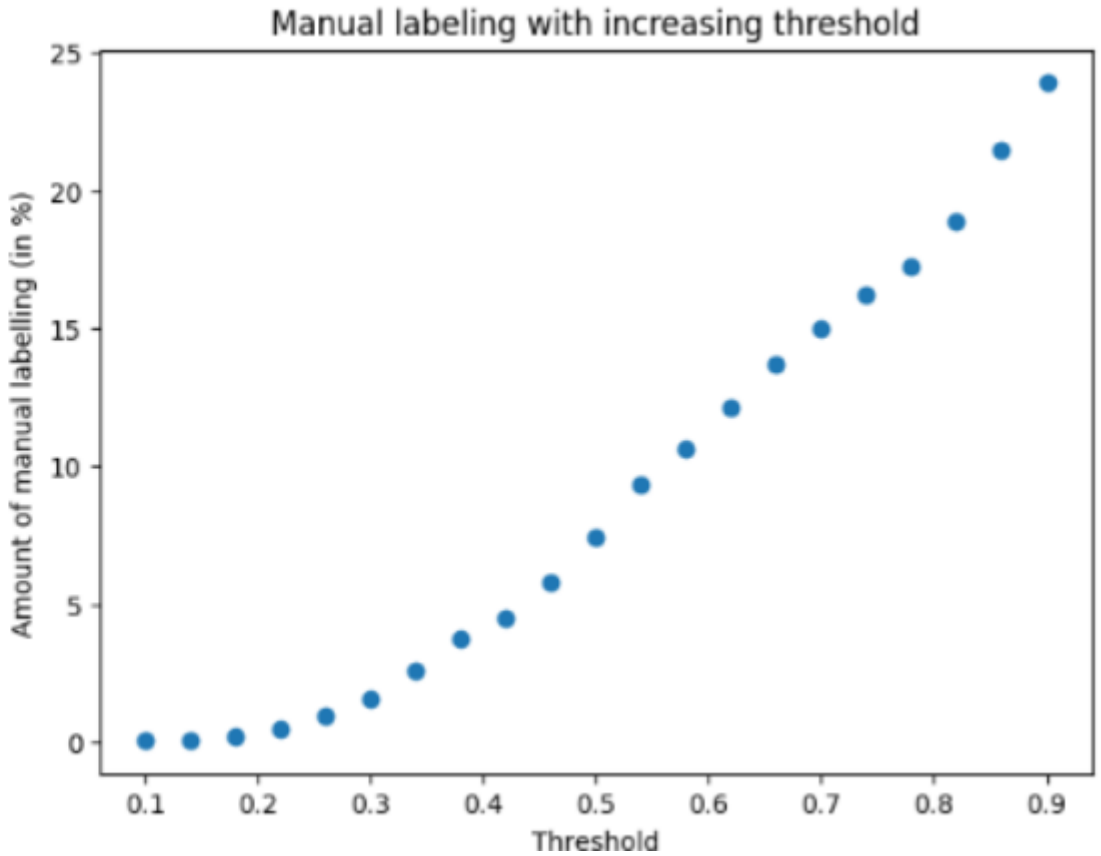
Statistisk sentralbyrå
Statistics Norway

Prediction algorithms - ISCO

- Results:
  - Accuracy: 77.4%
  - Micro f1: 77.3%
  - Macro f1: 69.2%
- Dropped company description

Threshold for ISCO





Statistisk sentralbyrå
Statistics Norway

Future plans

- Similar NACE/ISCO coding for LFS
- Back conversion of NACE – implementation sommer 2025
- New NACE: Masters student, AIML4OS
- Coicop classification for HBS 2026
- General framework for automatic classification 2025