

Research on AI application in coding and processing data in Labour Force Survey.



Bui Thi Thu Phuong
National Statistics Office of Vietnam

Contents

-  1 Background
-  2 Labour Force Survey in Vietnam
-  3 Implementation process
-  4 Results and solutions
-  5 Implementation plan

Background

```
graph TD; Background([Background]) --> International[International]; Background --> National[National context];
```

The diagram illustrates the background factors influencing data science. At the top is a blue oval labeled 'Background'. Two arrows originate from this oval: an orange arrow pointing to a box on the left titled 'International', and a blue arrow pointing to a box on the right titled 'National context'. The 'International' box contains three bullet points about data complexity, algorithm optimization, and AI application. The 'National context' box contains two bullet points about demand for fast data and repeated surveys.

International

- Data is getting bigger, more diverse and more complex
- Algorithms are getting more and more optimized
- Application of AI in many fields

National context

- Demand for fast and accurate data provision
- Many repeated surveys with unchanged questionnaires

Labour Force Survey in Vietnam

Monthly data
collecting and
processing

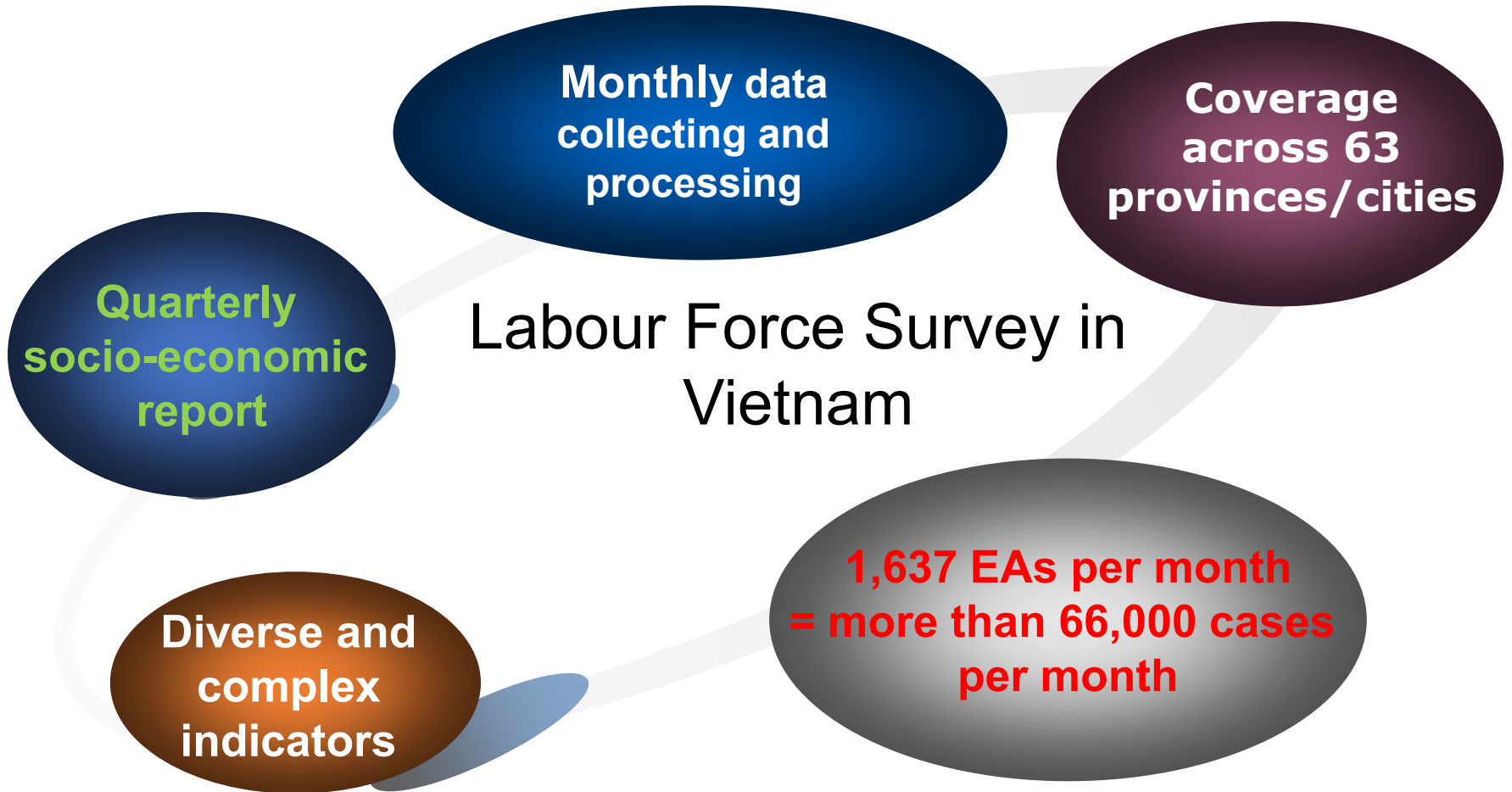
Coverage
across 63
provinces/cities

Quarterly
socio-economic
report

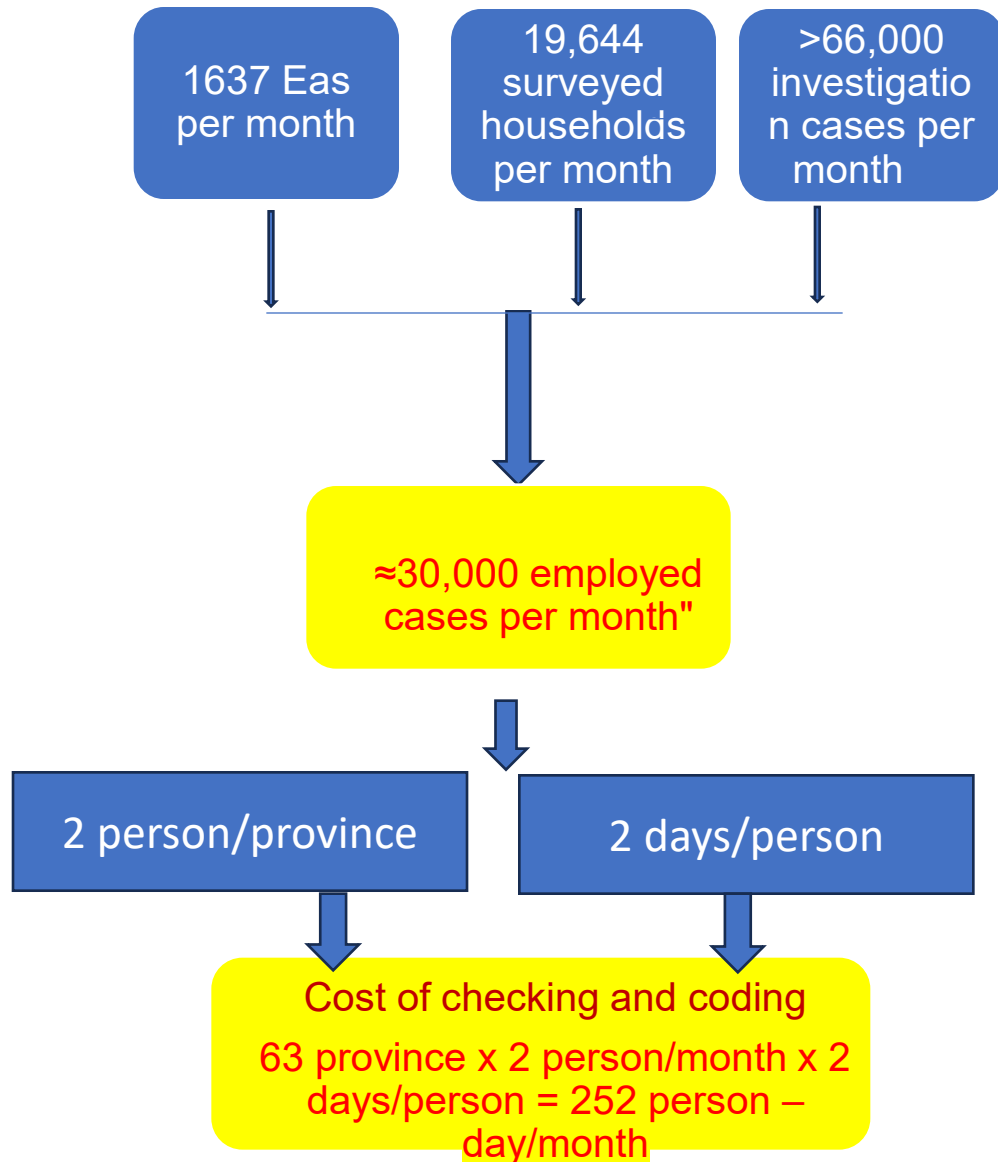
Labour Force Survey in
Vietnam

Diverse and
complex
indicators

1,637 EAs per month
= more than 66,000 cases
per month



Occupational coding

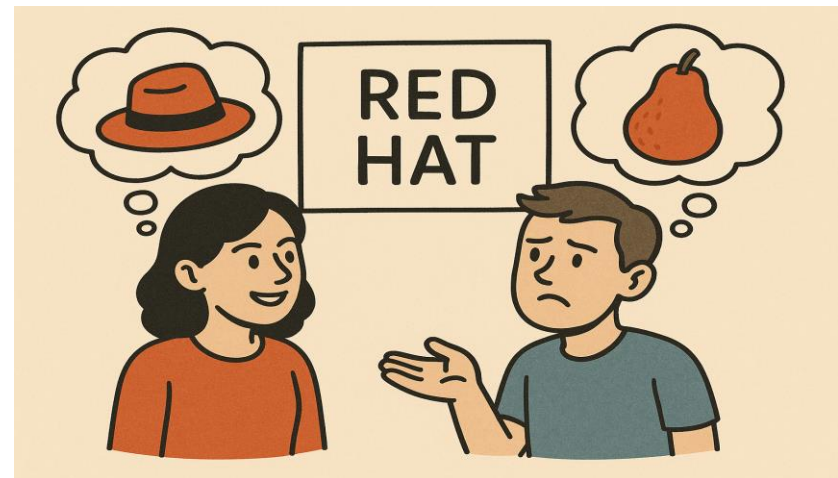


Occupational coding

The respondent's written answers were incomplete or insufficient, leading to incorrect coding by the coder



Lack of consistency: Coding based on the subjective judgment of the coder



Occupational coding

Prolonged data processing time

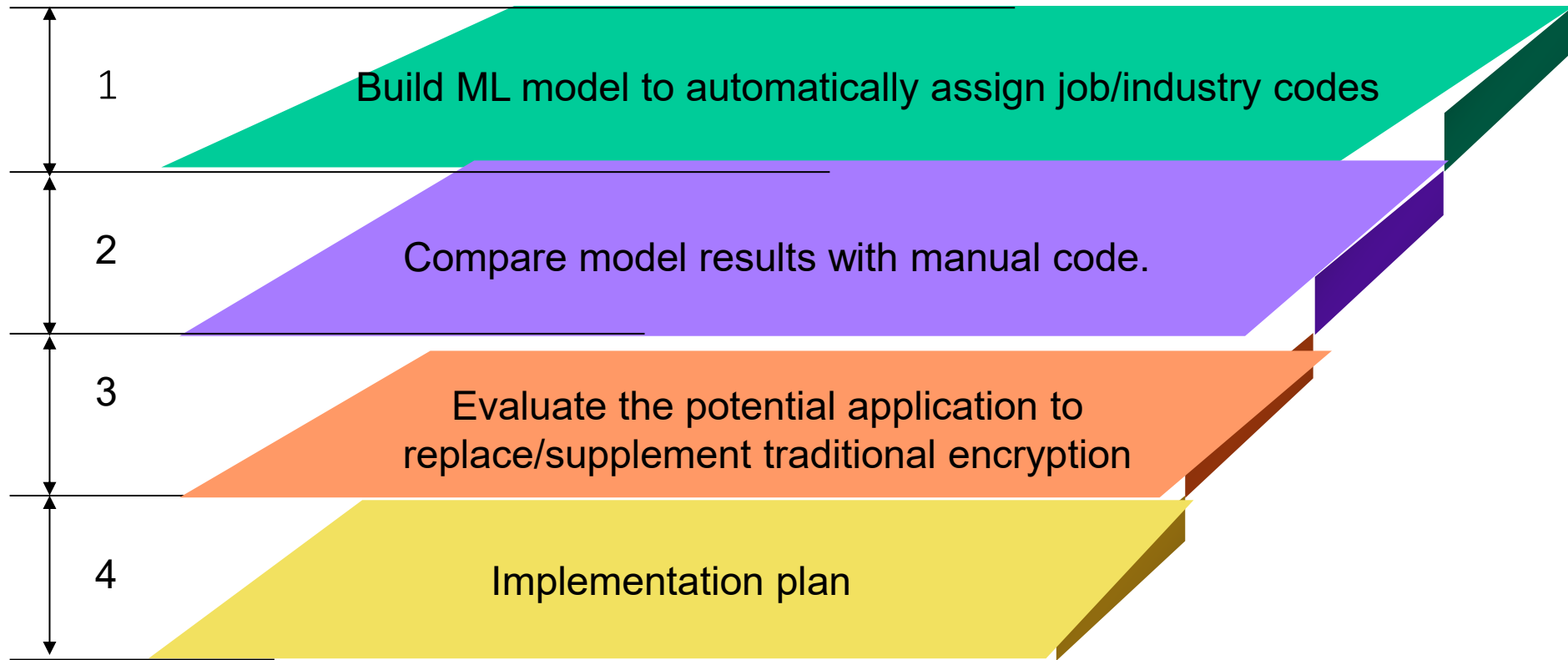
Additional data processing costs

Low accuracy

Lack of consistency

Research to
Develop
Machine
Learning–Based
Automatic
Coding
Guidelines

Objective



IMPLEMENTATION PROCESS

Data sources

- LFS survey data from 2021 to 2024
- Data size: over 400,000 cases
- Variables in the data: "main_task", "job_title", "education_level", "economic_sector", "job code"



**Merge data, define description columns,
code columns**



TOOL: PYTHON

IMPLEMENTATION PROCESS

Text preprocessing	Text Normalization, Word Separation, Stop Word Removal, Word Normalization, Acronym Handling, Combine Text Fields
Train-test split	Train-test ratio: 80-20
TF-IDF Vectorization	Convert data (especially text, images, or discrete features) into a numerical form (number vectors) that the model can understand and process
Test models and evaluate to choose the best model	LogisticRegression, LinearSVC, Naives Bayer, CosineCentroid
Test models and evaluate to choose the best model	Predicting job codes on new data

MODEL EVALUATION RESULTS

Model	Accuracy	Train_time(s)	Predict_time(s)
LogisticRegression	0.700753	342.5639112	0.666987419
LinearSVC	0.707015	1414.737299	1.167625427
MultinomialNB	0.614936	35.45636511	1.075924158
CosineCentroid	0.542409	130.090924	18.46308208

Best Model: Linear SVC

Strength: Best performance, Accuracy (0.70715)

PREDICTIONS FOR NEW DATA

Data used for prediction: LFS data Q2 2025 with manual coding

job_code	main_task	job_title	education_level	economic_sector
5211	Nhân viên bán xăng dầu	Nhân viên	3,00	Kinh doanh xăng dầu
3434	Nấu ăn	Nhân viên	3,00	Dịch vụ ăn uống
5246	Bán hàng nước	Tự làm	1,00	Bán hàng nước



Best model

- Prediction result: “predict_job_code” is generated on the prediction data file

job_title	education_level	economic_sector	text	predicted_job_code
Nhân viên	3,00	Kinh doanh xăng dầu	Nhân viên bán xăng ...	5223
Nhân viên	3,00	Dịch vụ ăn uống	Nấu ăn Nhân viên 3....	5120
Tự làm	1,00	Bán hàng nước	Bán hàng nước Tự l...	5211

PREDICTIONS FOR NEW DATA

Find top 1 and top 3

AccuracyTop 1: Predicts correctly the first time,
Top 3: Safe “suggestions”, reduces errors, especially
when job_code has many classes and data is highly
diverse.



Python code

Top 1- Accuracy: 0,7015
Top 3 - Accuracy: 0,8594

PREDICTIONS FOR NEW DATA

TOP - 3



 main_task	 job_title	 education_level	 economic_sector	 text	 top1_pred	 top2_pred	 top3_pred
bộ đội	không	5,00	bộ đội	bộ đội khô...	120.0	130.0	320.0
kỹ thuật máy	nhân viên	5,00	an ninh qu...	kỹ thuật m...	2152.0	220.0	130.0
bộ đội	bộ đội	4,00	quốc phòng	bộ đội bộ đ...	120.0	130.0	110.0
bộ đội	không	3,00	quốc phòng	bộ đội khô...	130.0	120.0	110.0
bộ đội chu...	bộ đội	3,00	an ninh qu...	bộ đội chu...	130.0	110.0	120.0
bộ đội chu...	bộ đội	2,00	an ninh qu...	bộ đội chu...	130.0	110.0	120.0
bộ đội chu...	bộ đội	2,00	an ninh qu...	bộ đội chu...	130.0	110.0	120.0
quản lý đư...	quản lý đư...	5,00	quản lý đư...	quản lý đư...	1739.0	3112.0	3139.0
phó ban thi...	phó ban thi...	5,00	quản lý nh...	phó ban thi...	2422.0	1232.0	2212.0
...	...	...	...	...	...	...	...

PREDICTED RESULTS

70%
match

- The match between manual and ML code is: 70 (consistent with existing studies)
- If referring to Top - 3, the match probability is up to 85.9%

30%
difference

Data

- Inconsistent, missing information
- Data imbalance between groups

Model

- Algorithm limitations
- How to predict
- Training data: insufficient samples, imbalance

Human

- Subjective, inconsistent judgment
- Humans can understand deep semantics

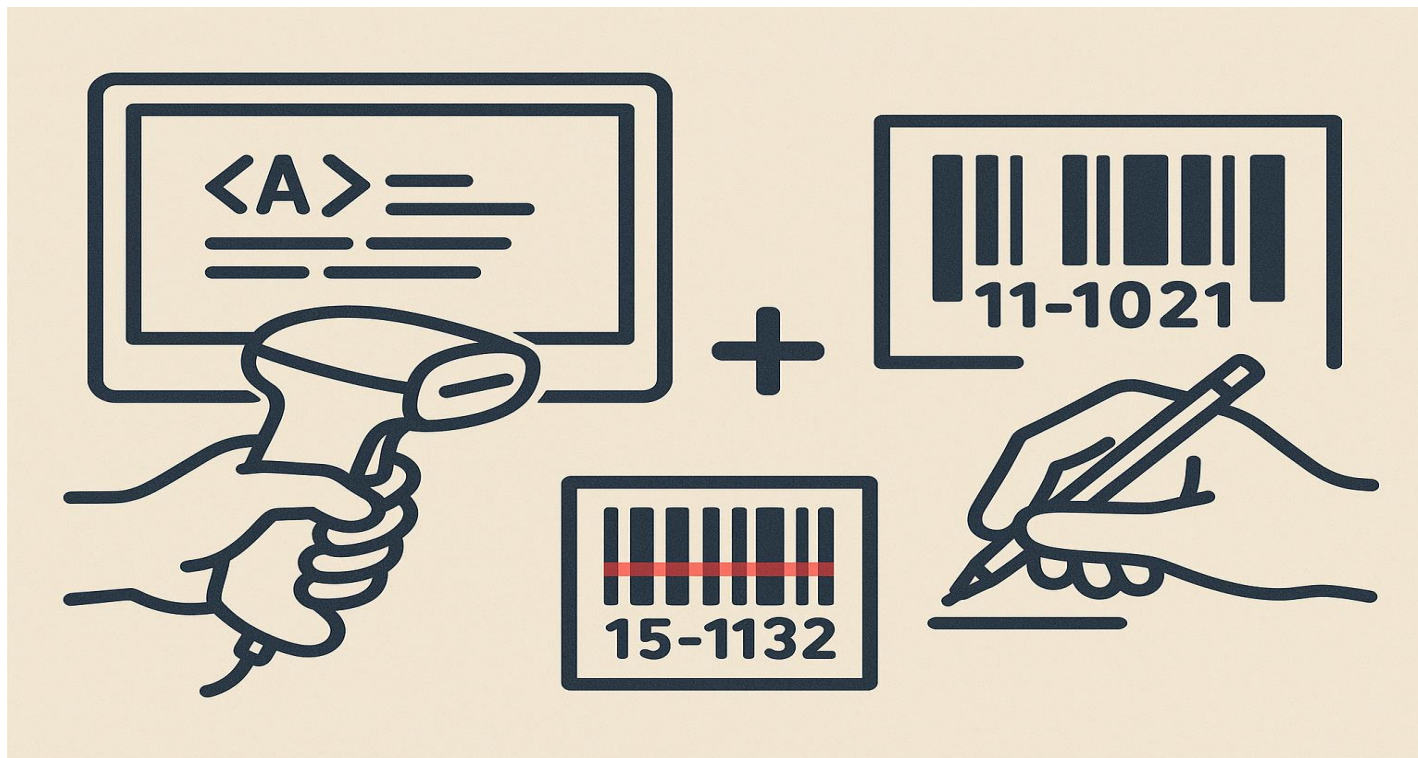
PREDICTED RESULTS

	Automatic coding	Manual coding	Automatic and manual combination
Processing speed	20 minutes	4 days	1 days
Accuracy	70%	95-97%	96-98%
Human resource costs		2 people working for 2 days	1 people working for 1 days



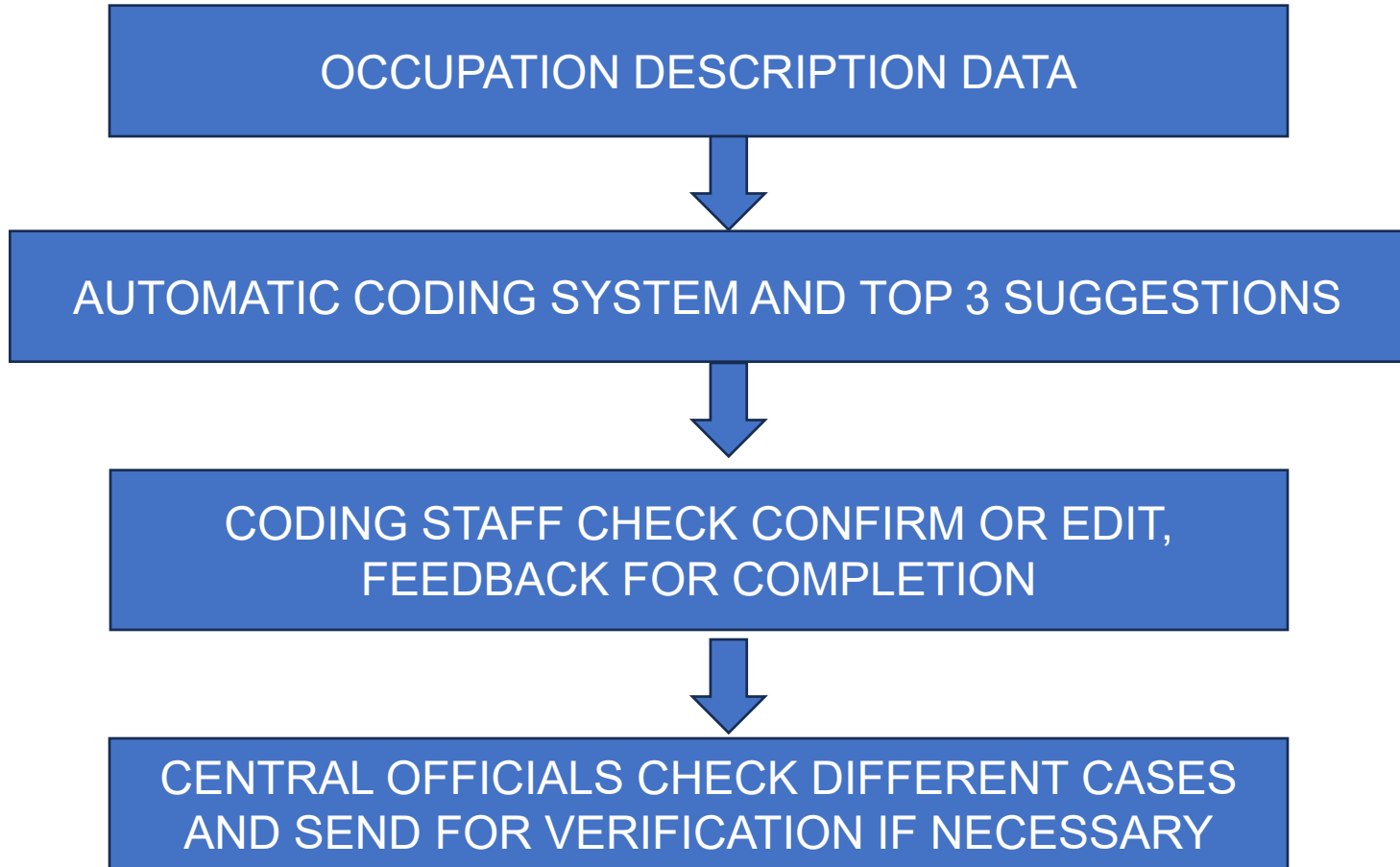
Combination of manual and automatic coding

SOLUTION



Combination of manual and automatic coding

SOLUTION



Lessons and Experiences

- Input data is very important: inaccurate description, data imbalance...
- It is necessary to pilot and evaluate the accuracy with many aspects: increase the volume of training data set, handle data imbalance, add independent variables...
- It is not advisable to rely entirely on machines but still need human control.

THE WAY FORWARD

1

Pilot

- Evaluate the feasibility
- Collect feedback to improve the model

2

Enhance data quality

- Expand the training dataset
- Update corrected cases for retraining
- Improve data cleaning and reduce imbalance

3

Test new algorithms

- Deep Learning
- Gradient Boosting



**Thank
You!**