통계청
Statistics Korea

# Use of Artificial Intelligence for Census Data Processing

## - A study on automatic coding of industry & occupation classification -

The 31st PCC, Nov. 30. 2022

Minju Kim, Eunsook Jung
(KOSTAT Population Census Division)

# 1 Introduction

## ➢ Korea's census reports on industry & occupation

- KOSTAT has been collecting industry & occupation data through census since 1960 (20% of population)
- Takes a long time to classify the data due to a big amount of self responses* which aren't refined well

  * Self response rate in Korea: (2015) 48.6%, (2020) 43.9%

### Data collection

Entering text at field enumeration or internet/telephone survey

**Survey items**

- Industry: business name, business area
- Occupation: position, department, individual task

### Data processing

Code classification at the editing step (4 digits)

**Coding procedures**

① Industry coding

② Occupation coding

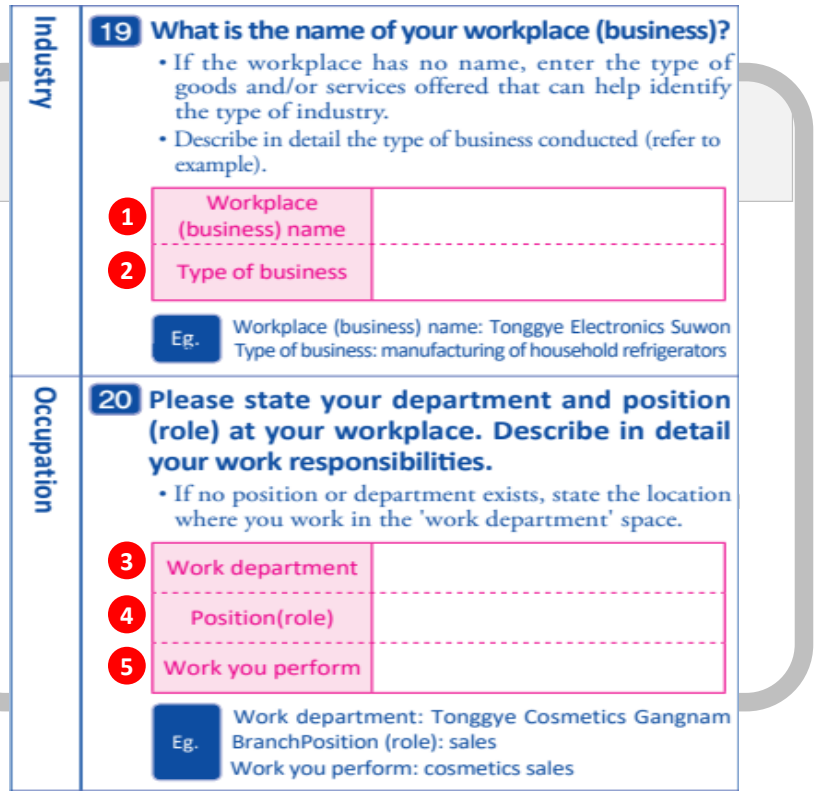
③ Cross coding of industry and occupation

# 1 Introduction

## ➢ Korea's census reports on industry & occupation

- KOSTAT has been collecting industry & occupation data through census since 1960 (20% of population)
- Takes a long time to classify the data due to a big amount of self responses* which aren't refined well

  * Self response rate in Korea: (2015) 48.6%, (2020) 43.9%



**Industry**

**19 What is the name of your workplace (business)?**

- If the workplace has no name, enter the type of goods and/or services offered that can help identify the type of industry.
- Describe in detail the type of business conducted (refer to example).

| | |
|---|---|
| 1 Workplace (business) name | |
| 2 Type of business | |

Eg. Workplace (business) name: Tonggye Electronics Suwon
Type of business: manufacturing of household refrigerators

**Occupation**

**20 Please state your department and position (role) at your workplace. Describe in detail your work responsibilities.**

- If no position or department exists, state the location where you work in the 'work department' space.

| | |
|---|---|
| 3 Work department | |
| 4 Position(role) | |
| 5 Work you perform | |

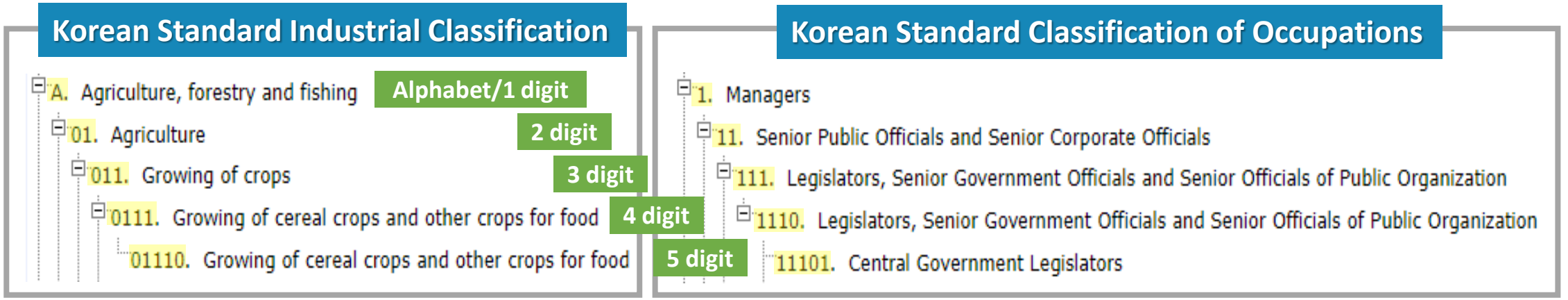
Eg. Work department: Tonggye Cosmetics Gangnam BranchPosition (role): sales
Work you perform: cosmetics sales

| 1 | 2 | | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| 직장, 사업체명 | 주된사업내용 | 산업코드(10차) | 일의 종류 | 근무부서 | 직책 | 직업코드(7차) |
| VARCHAR2 | VARCHAR2 | VARCHAR2 | VARCHAR2 | VARCHAR2 | VARCHAR2 | VARCHAR2 |
| | 아르바이트 | 8522 | | | 기타 서비스관련 단순 종사원 | 9999 |
| | 엘지텔레콤 | 6122 | | 학교 | | 9999 |
| | 삼은초등학교 | 8512 | 안전지킴이 | 안전지킴이 | 안전지킴이 | 9999 |
| | 골프공세척 | 9521 | 골프공세척 | 집 | | 9999 |
| | 정수기 | 7629 | 관리,영업 | 주)코웨이 | 팀장 | 9999 |
| | 대학생 | 8530 | | 호남대학교 | | 9999 |
| | 지킴이활동 | 9493 | 아동안전지킴이 | 부평경찰서 | 봉사활동 | 9999 |
| | 교육 | 8530 | | | 기타 서비스관련 단순 종사원 | 9999 |
| | 고등학생 | 8522 | | | 기타 서비스관련 단순 종사원 | 9999 |
| | 사진 촬영 | 7330 | | | 기타 서비스관련 단순 종사원 | 9999 |
| | 노인일자리 알선 | 8729 | 문화재 지킴 | 청간정 | 일용직 | 9999 |
| | 정수기 관련 | 7629 | 정수기코디 | | | 9999 |
| | 노인일자리알선 | 8729 | 교통안전지킴이 | | | 9999 |
| | 정수기 방문관리 | 7629 | 정수기방문관리 | 강서구 | 사원 | 9999 |
| | 혁생 | 8542 | | | 기타 서비스관련 단순 종사원 | 9999 |
| | 영어번역서비스 | 7390 | | | 기타 서비스관련 단순 종사원 | 9999 |
| | 초등학생등교 교통정리 | 8729 | 초등학생등하교 교통정리 | | | 9999 |

# 1 Introduction

➢ **Korea's Industry & Occupation Classification: KSIC** (industry) **& KSCO** (occupation)

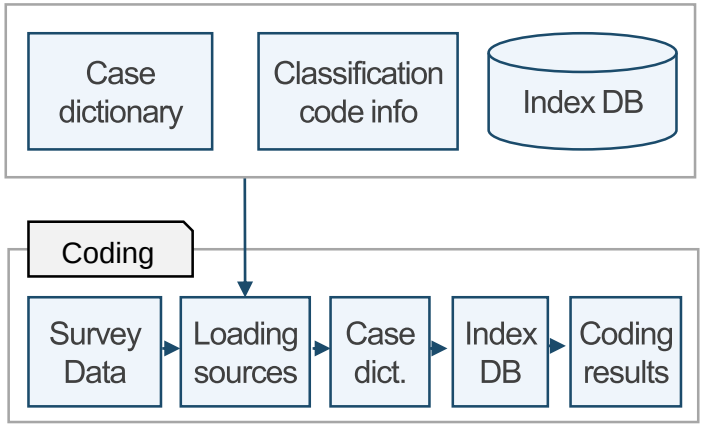
- Hierarchical 5-digit number to identify a certain industry or occupation
- Based on ISIC & ISCO (international classification)
- Collected through KOSTAT's Population Census, Local Area Labor Force Survey, etc.
- For census reports, only 4 digit codes are presented



## ➢ Previous classification schemes of industry & occupation code

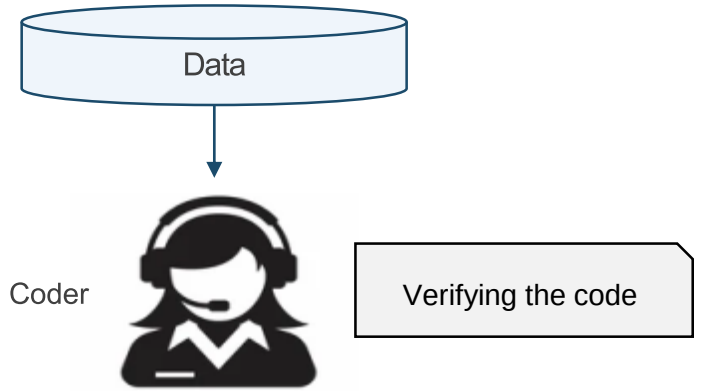
### Automatic coding system

- Keyword-based classification model
- **High accuracy, but low coverage**




### Manual coding by coders

- Determining the code by phone calls, discussions, etc.
  → **Takes a long time**
- Difference between each coder's skill and ability
  → **Inconsistency problem**



# 1 Introduction

## ➢ Considering using artificial intelligence on industry & occupation coding…

- Aritifical intelligence
  - Data-driven approach: “Let the data teach model how to predict codes”
- What’s different from the previous automatic coding system?
  - Wide coverage + ‘Confidence level’

## ➢ Previous study on industry & occupation code classification

- *Research on Automatic Census Industry/Occupation Coding and Data Analaysis*
  - Supported by KOSTAT, performed by FS (2020.6.~11.)
  - Used part of 2015 census dataset (400,000 + 600,000)
  - Comparison between coding results of 2015 census and AI model

# 2 A study on automatic classification using AI

통계청

## ➢ KOSTAT's latest research performed by FS (2022.5.~11.)

## ➢ Main goal

- To enhance **rapidity & accuracy** of industry/occupation coding procedures using AI

## ➢ Scope of the study

1. Comparison between coding results of 2020 census and new AI model
    - Consistency rate by models / digits / phases
2. Comprehensive analysis of inconsistency
    - Determining and categorizing causes of inconsistency
3. Suggestions: How to apply AI on coding procedures
    - Finding the best combination of previous auto coding, manual coding and AI

# A study on automatic classification using AI

## ➢ 1. Comparison between coding results of 2020 census and new AI model

### 1.1. Methodology

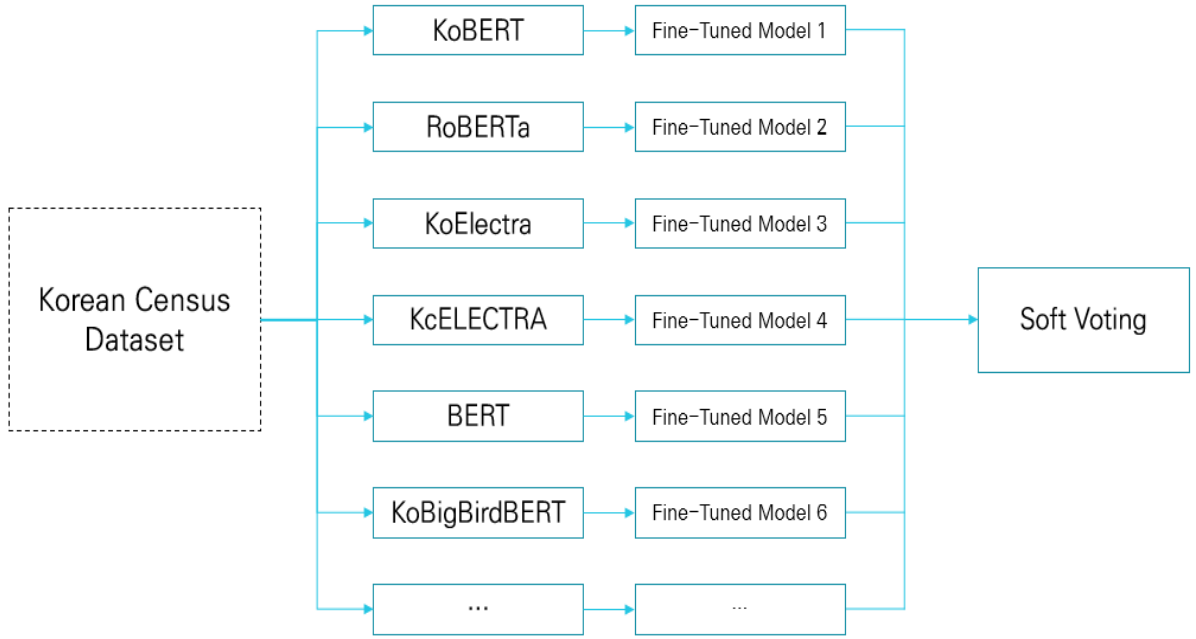
**2020 Census dataset**

| Usage | Auto coding | Manual coding | | Final version |
|---|---|---|---|---|
| | File 1 | File 2 | File 3 | File 4 |
| AI learning & prediction | | | | ○ |
| Comparing the results | ○ | ○ | ○ | ○ |

**Used items**

- Industry: business name, business area, coding results
- Occupation: sex, age, education, place of work, position, department, individual task, coding results

**Ensemble Model(Voting)**



# A study on automatic classification using AI

통계청

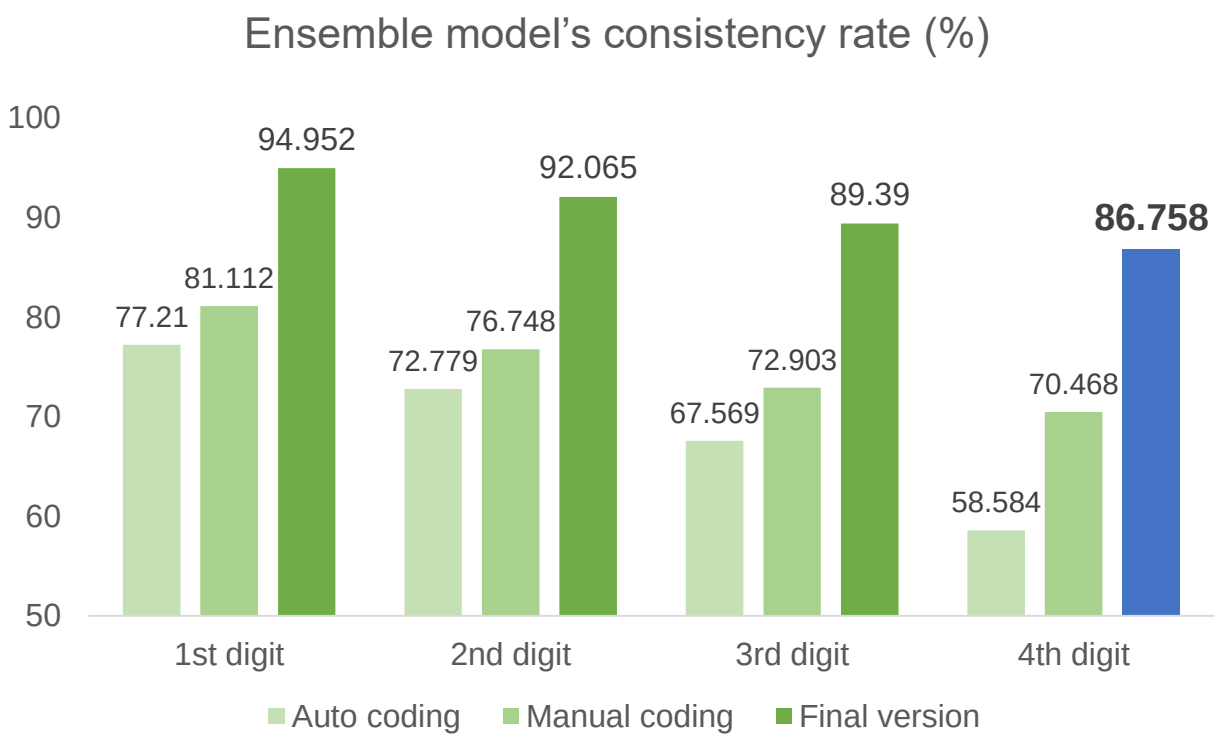
## ➢ 1. Comparison between coding results of 2020 census and new AI model

### 1.2. Consistency rate (Industry)

**By models**

| AI Model | Consistency rate (%) |
|---|---|
| **Ensemble model** | **86.758** |
| **KcBERT-large** | **85.889** |
| **KoBigBirdBERT-base** | **85.830** |
| **KcBERT-base** | **85.828** |
| **BERT-base** | **85.707** |
| **KcELECTRA v3 (Base Discriminator)** | **85.195** |
| RoBERTa-large | 85.161 |
| RoBERTa-base | 85.055 |
| RoBERTa-small | 84.997 |
| KcELECTRA-base | 84.711 |
| KoBERT | 28.944 |

**By digits & phases**



# A study on automatic classification using AI

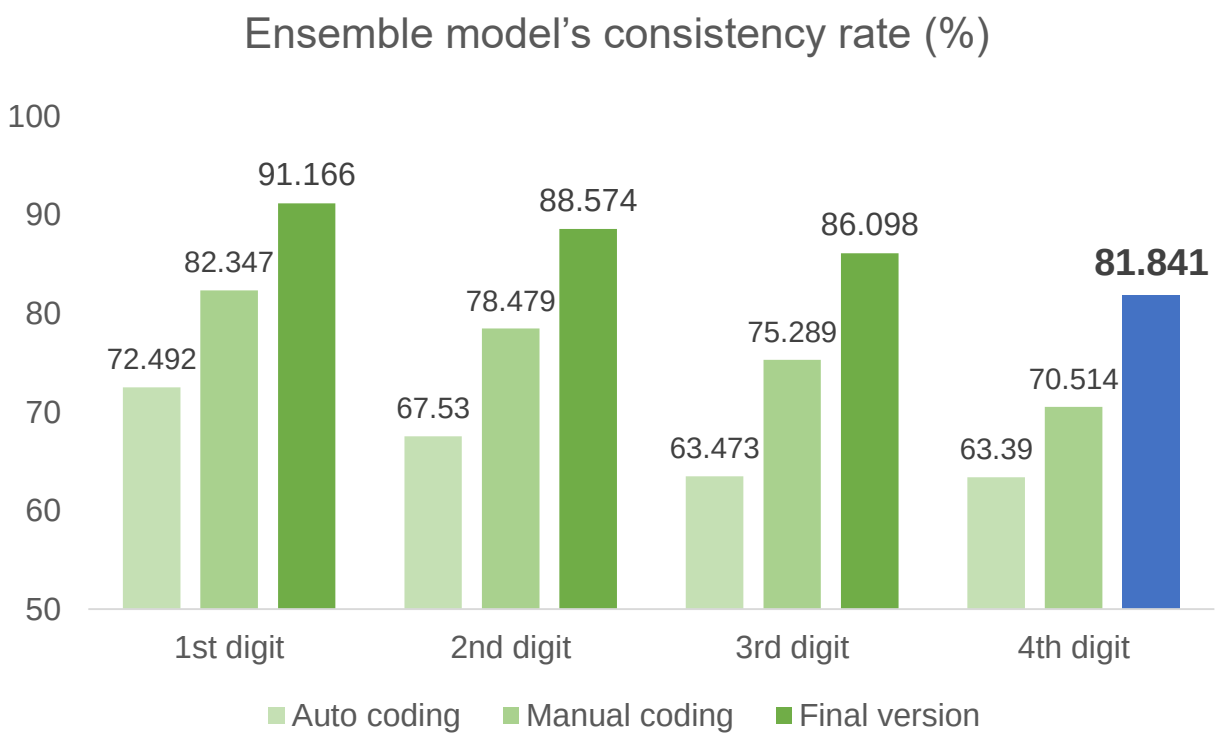
## ➢ 1. Comparison between coding results of 2020 census and new AI model

### 1.2. Consistency rate (Occupation)

**By models**

| AI Model | Consistency rate (%) |
|---|---|
| **Ensemble model** | **81.841** |
| **KoBigBirdBERT-base** | **81.045** |
| **BERT-base** | **80.956** |
| **KcBERT-base** | **80.808** |
| **KcBERT-large** | **80.511** |
| **KcELECTRA v3 (Base Discriminator)** | **80.227** |
| RoBERTa-base | 80.130 |
| RoBERTa-small | 80.117 |
| RoBERTa-large | 80.023 |
| KcELECTRA-base | 79.860 |
| KoBERT | 29.677 |

**By digits & phases**



# A study on automatic classification using AI

## ➢ 2. Comprehensive analysis of inconsistency

### 2.1. Coders' work on inconsistency

- Skilled coders worked for months, analyzing 30,177 (ind.) & 32,024 (occ.) data
- Choosing an appropriate code among 2020 census data and AI model's code
- Determining the cause of inconsistency

**Coders' work form on inconsistency**

**Industry**

| Text | | Results | | Coders' choice | | |
|---|---|---|---|---|---|---|
| Business name | Type of business | 2020 census No.1 | AI No.2 | No.1 or No.2 | Cause(s) of inconsistency | |
| | | Code/Name | Code/Name | | | |
| | | .. | .. | | | |
| | | .. | .. | | | |
| | | .. | .. | | | |

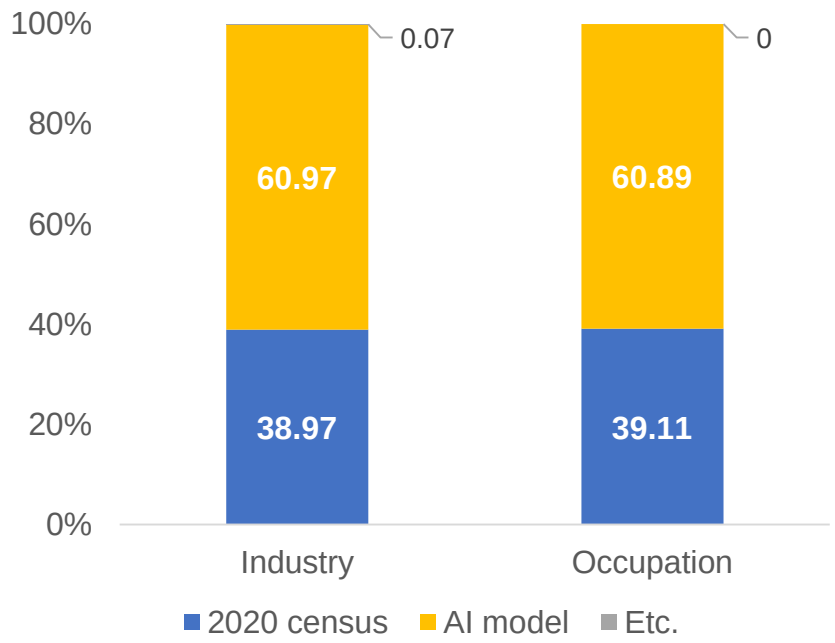
**Occupation**

| Text | | | Results | | Coders' choice | | |
|---|---|---|---|---|---|---|---|
| Position | Depart-ment | Individual task | 2020 census No.1 | AI No.2 | No.1 or No.2 | Cause(s) of inconsistency | |
| | | | Code/Name | Code/Name | | | |
| | | | .. | .. | | | |
| | | | .. | .. | | | |
| | | | .. | .. | | | |

## ➢ 2. Comprehensive analysis of inconsistency

### 2.2. Causes of inconsistency

Share of chosen code (%)




Major cause of inconsistency

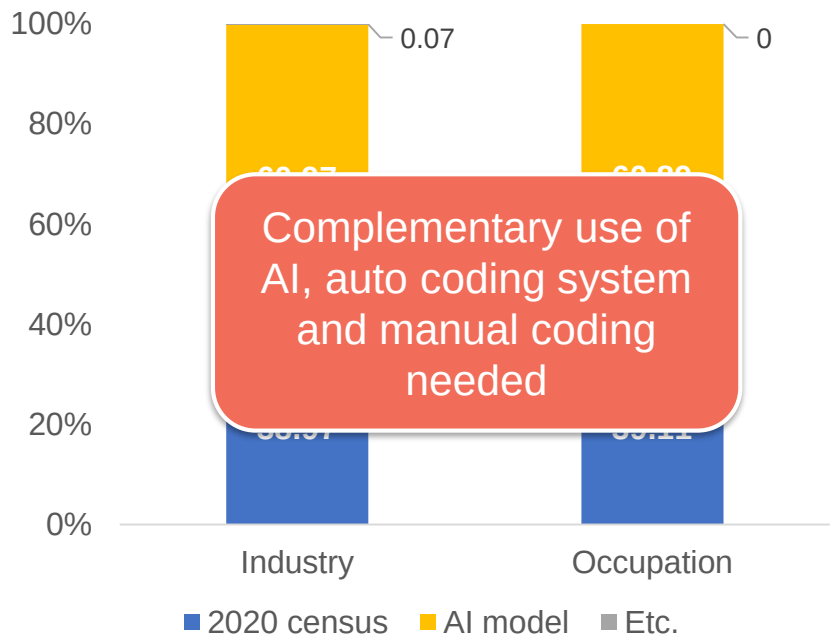
| Cause of inconsistency | Industry | | Occupation | |
|---|---|---|---|---|
| | Number | Share(%) | Number | Share(%) |
| **Uncertain meaning of the data** | 11,721 | **38.8** | 11,177 | **34.9** |
| **Short length of the data** | 14,553 | **48.2** | 15,421 | **48.2** |
| Missing ‘Business name’ or ‘Type of business’ | 1,298 | 4.3 | 4,026 | 12.6 |
| Related to ‘Other’ on taxonomy | 53 | 0.2 | 0 | 0.0 |
| Additional info needed (e.g. sex, age, education, …) | 0 | 0.0 | 0 | 0.0 |
| Etc. | 2,552 | 8.5 | 1,400 | 4.4 |
| **Total** | 30,177 | 100.0 | 32,024 | 100.0 |

# A study on automatic classification using AI

## ➢ 2. Comprehensive analysis of inconsistency

### 2.2. Causes of inconsistency

**Share of chosen code (%)**




Complementary use of AI, auto coding system and manual coding needed

**Major cause of inconsistency**

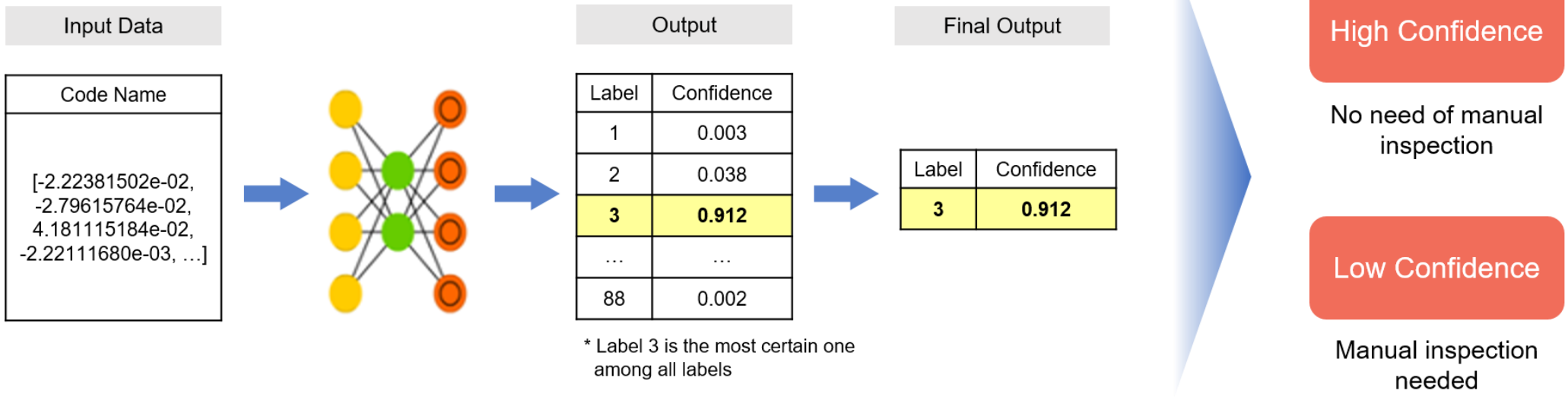
| Cause of inconsistency | Industry | | Occupation | |
|---|---|---|---|---|
| | Number | Share(%) | Number | Share(%) |
| **Uncertain meaning of the data** | 11,721 | **38.8** | 11,177 | **34.9** |
| **Short length of the dat** | | **48.2** | 15,421 | **48.2** |
| Missing ‘Business name’ or ‘Type | | 4.3 | 4,026 | 12.6 |
| Related to ‘Other’ on taxon | | 0.2 | 0 | 0.0 |
| Additional info needed (e.g. sex, age, education, …) | 0 | 0.0 | 0 | 0.0 |
| Etc. | 2,552 | 8.5 | 1,400 | 4.4 |
| **Total** | 30,177 | 100.0 | 32,024 | 100.0 |

More accurate and ample data needed

## ➢ 3. Suggestions: How to apply AI on coding procedures

### 3.1. Areas on which AI is applicable

**AI's work process**



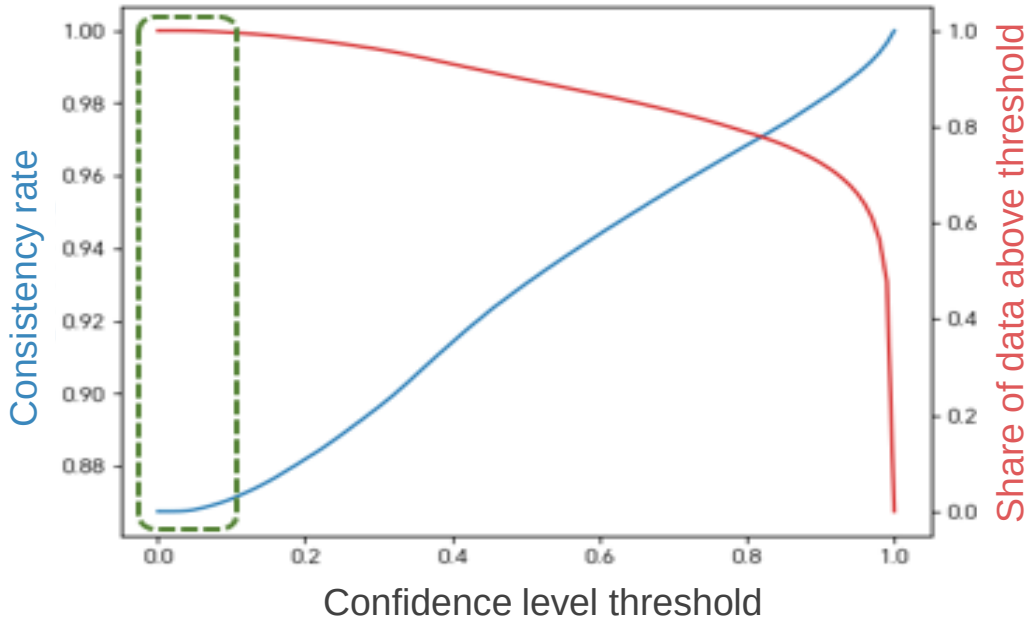
# A study on automatic classification using AI

## ➢ 3. Suggestions: How to apply AI on coding procedures

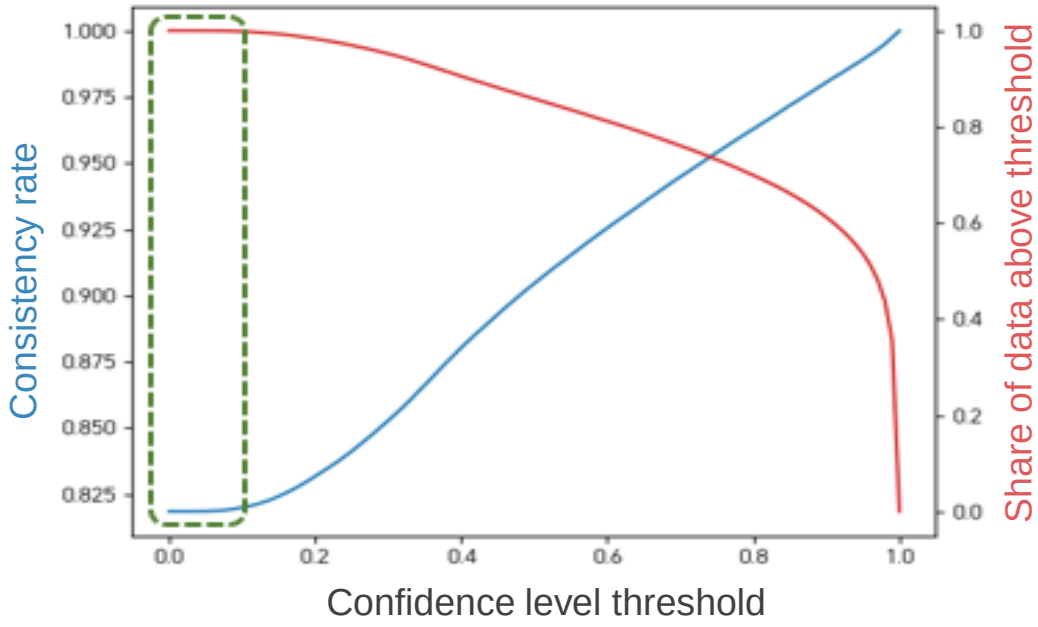
### 3.1. Areas on which AI is applicable

**Tradeoff between confidence & coverage**

**Industry**



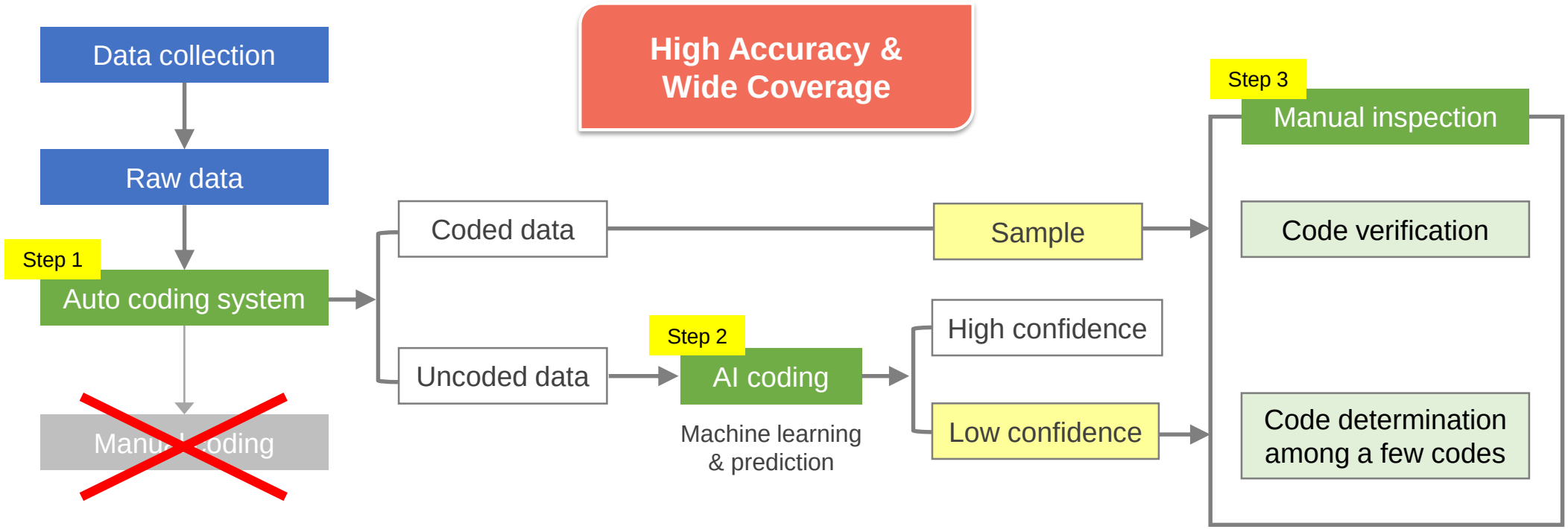

**Occupation**



# A study on automatic classification using AI

## ➢ 3. Suggestions: How to apply AI on coding procedures

### 3.2. Possible scenario for census data processing



# 3 Future plans

## ➢ Applying on 2025 Population & Housing Census 1st Pilot Survey

- 1st Pilot Survey: 2022. 10. 17. ~ 2022. 11. 18.
- There’s no ‘fixed answer’ of the data: a good opportunity to verify AI’s performance
- Comparing…
  - Previous auto coding system's codes for each data
  - AI model's top 3 codes for each data

→ Coders will determine which code is appropriate, so that we can clarify…

- whether AI's performance is good enough or not to predict codes
- if it is good enough, then to which area or phase that AI is applicable

# 3 Future plans

## ➢ Establishing census work process improvement plan

- Based on comprehensive consideration of research report and pilot survey analysis
- Re-establishment of census data processing: applying AI on industry & occupation classification

## ➢ Developing “Next-generation Census Management System“

- Based on derived 'work process improvement plan’
- Developing an AI-based automatic coding system applicable for 2025 census data processing

# Thank you

통계청
Statistics Korea