SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

Occupation Coding in Online Self-Completion Surveys: Evidence, Practice and Challenges

Matt Brown, Lisa Calderwood, Cristian Domarchi, Sebastian Kocar, **Helena Koerber** and Olga Maslovskaya

Survey Futures Conference | 18 June | City St George's, University of London

University of Essex

University of Southampton

UKRI Economic and Social Research Council

CITY University of London EST 1894

Ipsos

LSE The London School of Economics and Political Science

Manchester 1824 The University of Manchester

National Centre for Social Research

NCRM National Centre for Research Methods

Office for National Statistics

UCL

Ulster University

University of Leeds

Unil | Université de Lausanne

verian

Warwick The University of Warwick

Background

- Occupation is a core measurement in social surveys:
  - Indicator of socio-economic status
  - Strongly linked to income, health, and lifestyle
- Traditional Measurement Approach:
  1. **Interviewers** asking **open questions** to collect job title and a description of duties (Lyberg & Dean, 1992)
  2. **Interviewers** ensure the necessary information is provided (Conrad et al., 2016)
  3. **Software assisted** manual coding by specialist **office-based coders** to a **standard classification** (e.g., SOC 2020)

Background

- Methodological challenges:
  - Job titles/ descriptions are highly diverse
  - Same job may be described in multiple ways
  - In **self-completion surveys**, the absence of interviewers can have a negative impact on the quality of the collected data for coding (Conrad et al., 2016)
- Ongoing research aims to improve validity and reliability of occupation coding in surveys

Survey Futures Evidence Review

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

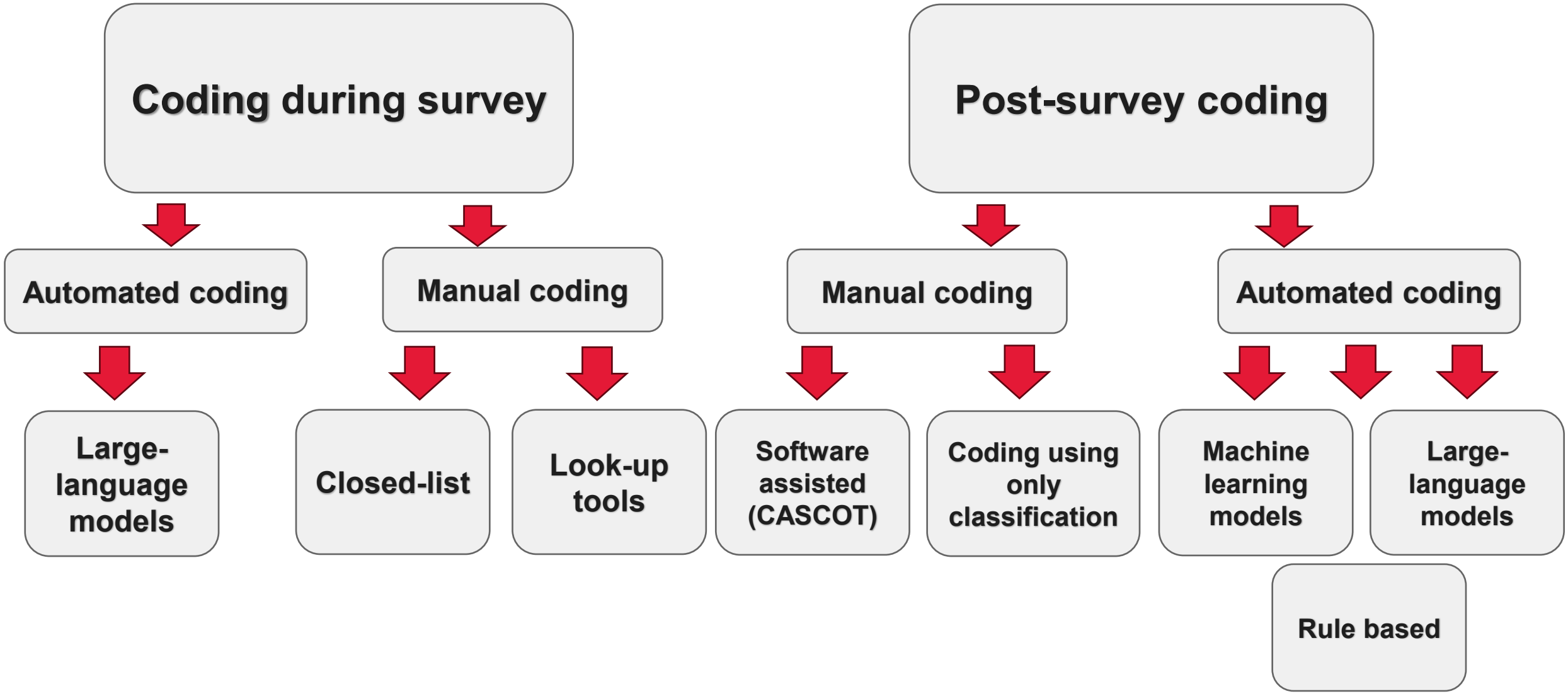
Report 3: Occupation coding in self-completion surveys: Evidence Review

Sebastian Kocar, Centre for Longitudinal Studies, UCL
Matt Brown, Centre for Longitudinal Studies, UCL
Lisa Calderwood, Centre for Longitudinal Studies, UCL

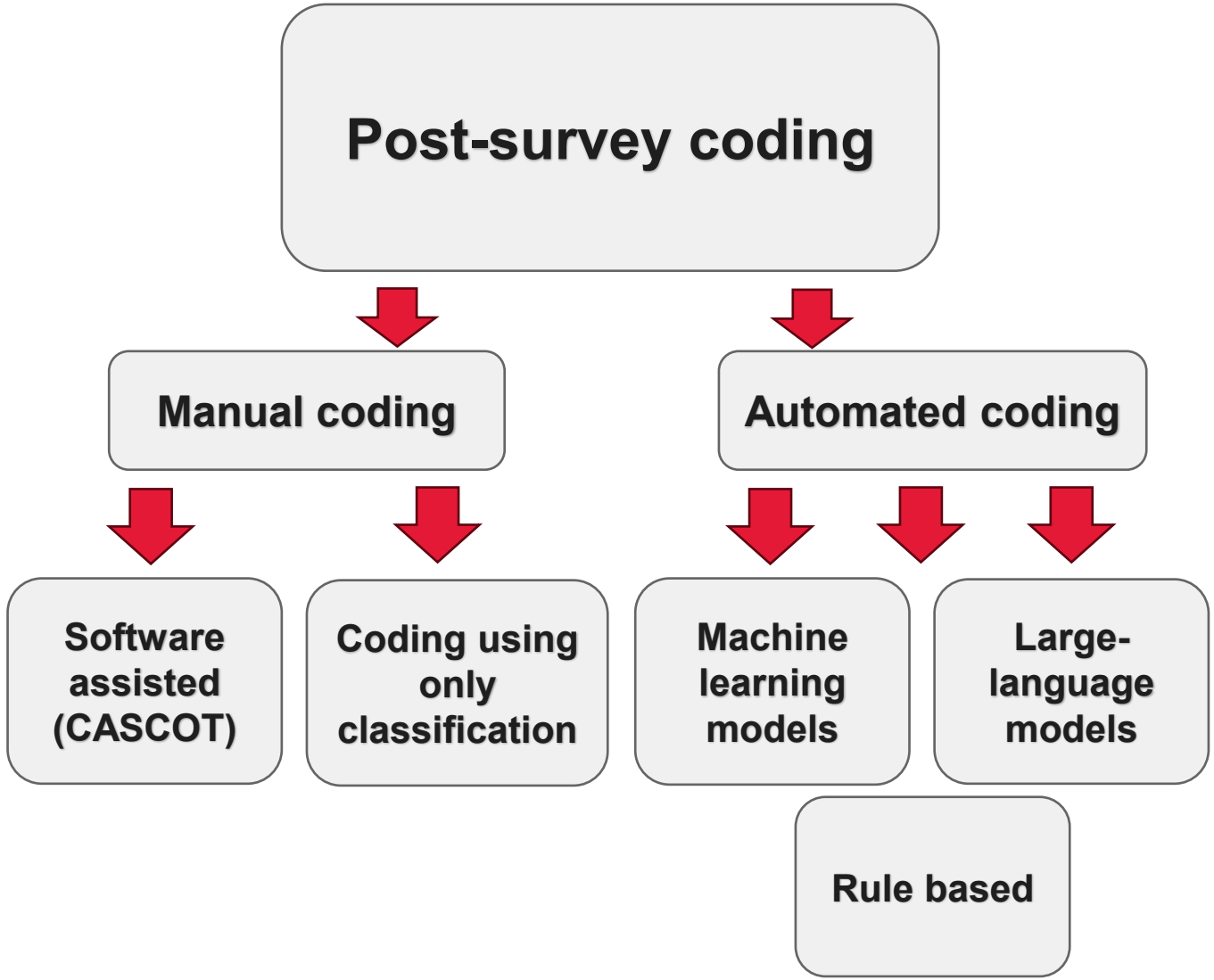
February 2025

University of Essex
University of Southampton
UKRI Economic and Social Research Council

Coding Approaches



Coding Approaches



Post-Survey Coding:
Manual Coding of Open Text

1. Respondent enters job title and duties
2. Professional coders assign SOC code with help of software (CASCOT)

Benefits

✓High coding success

✓High intercoder-reliability

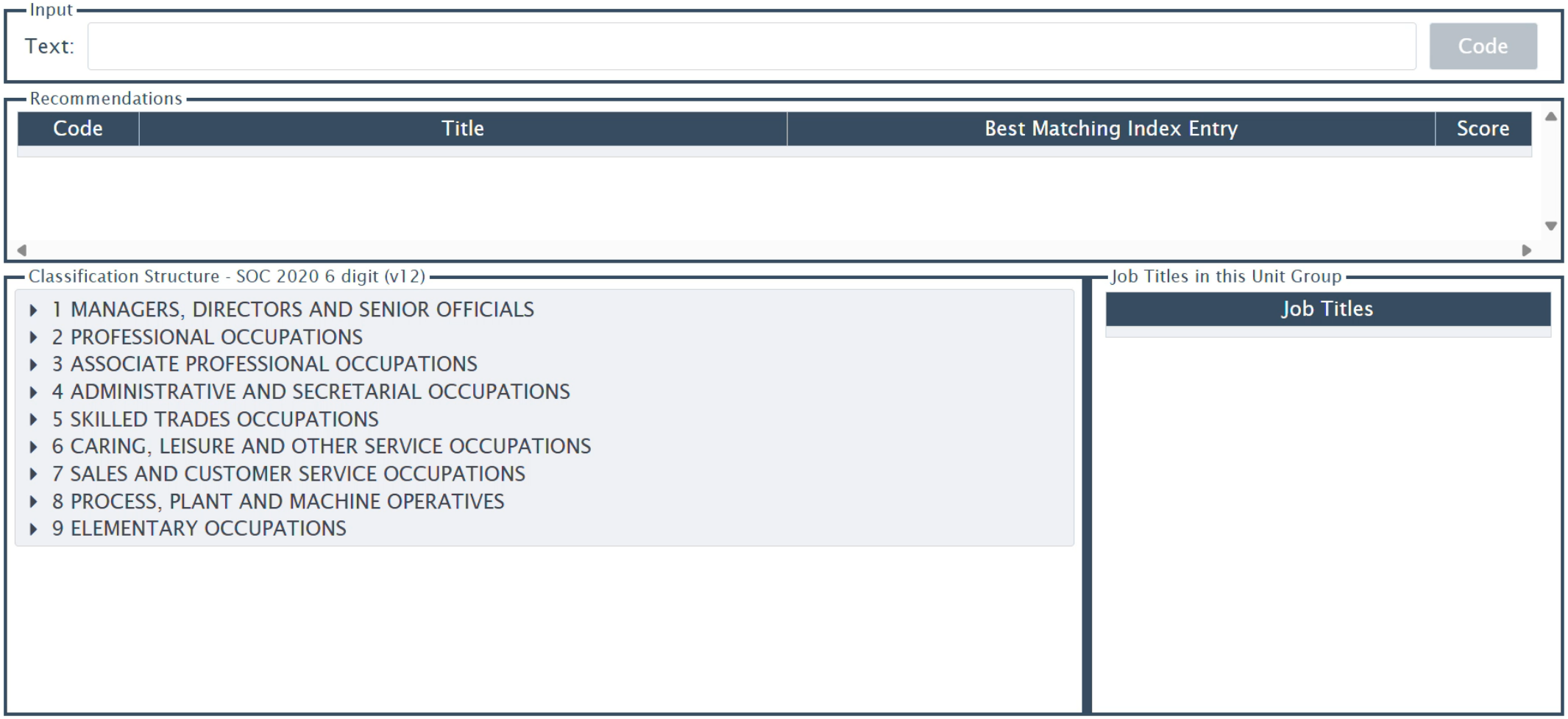
✓Comparatively high coding consistency over time

Key limitations

×Time-consuming and labour-intensive

×Expensive for large surveys

CASCOT Assisted Office Coding



Post-Survey Coding:
Manual Coding of Open Text

Examples from UK surveys

- Understanding Society
- National Child Development Study (NCDS)
- 1970 British Cohort Study (BCS70)
- European Social Survey (ESS)

Post-Survey Coding:
Automated Coding

Rule-based	Machine Learning
Expert-defined rules	Learns from data
Matches keywords and phrases	Predicts occupation codes
Transparent process	More flexible and scalable

Key challenge

- Higher accuracy ↔ Higher coding rate
- Not yet widely adopted as evidence suggests they are not as accurate as manual coding (Gweon et al., 2017).

Post-Survey Coding:
Automated Coding

Examples

- SOCcer (United States)
- 2021 Census (England & Wales)
- Transformed Labour Force Survey (TLFS)

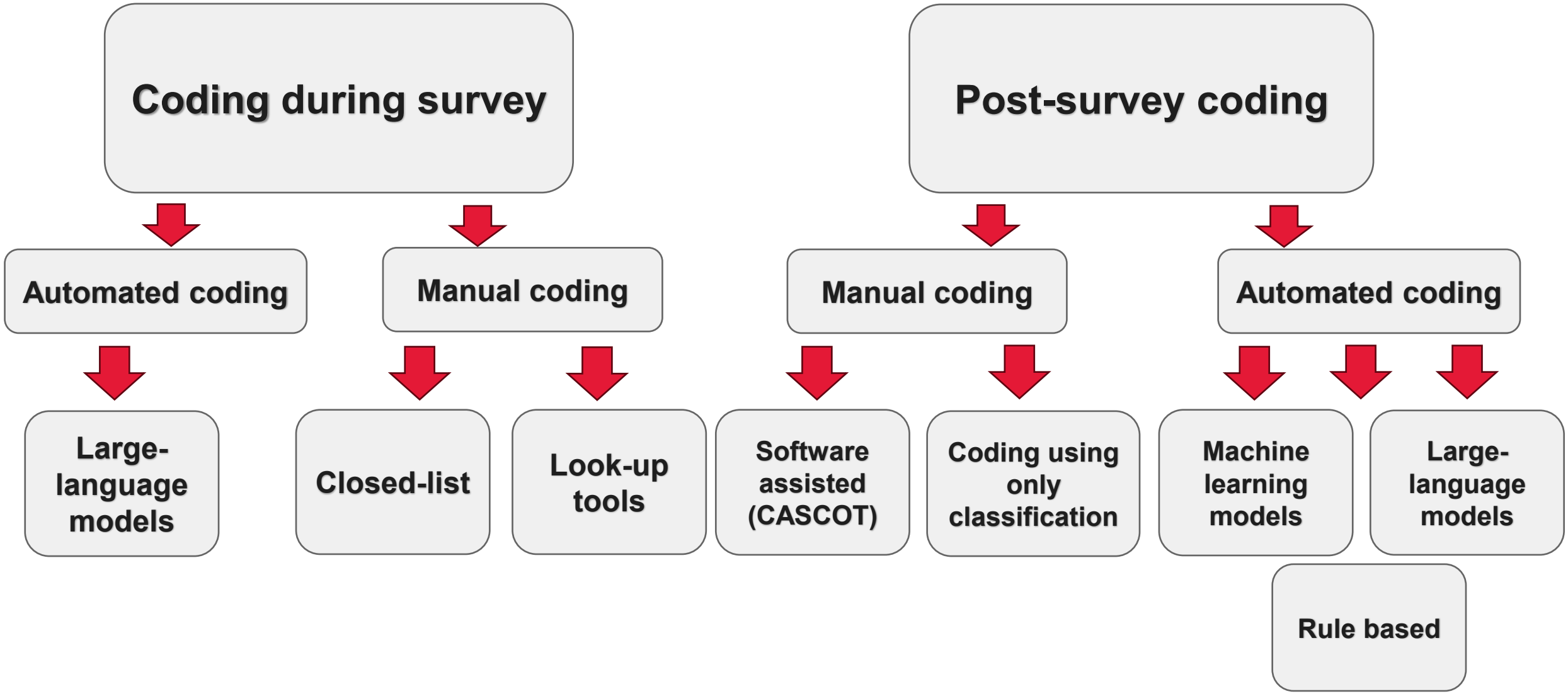
Post-Survey Coding:
Automated Coding: Large Language Models

- Traditional machine learning approaches require large amounts of manually coded training data
- LLMs may offer alternative where training data is limited

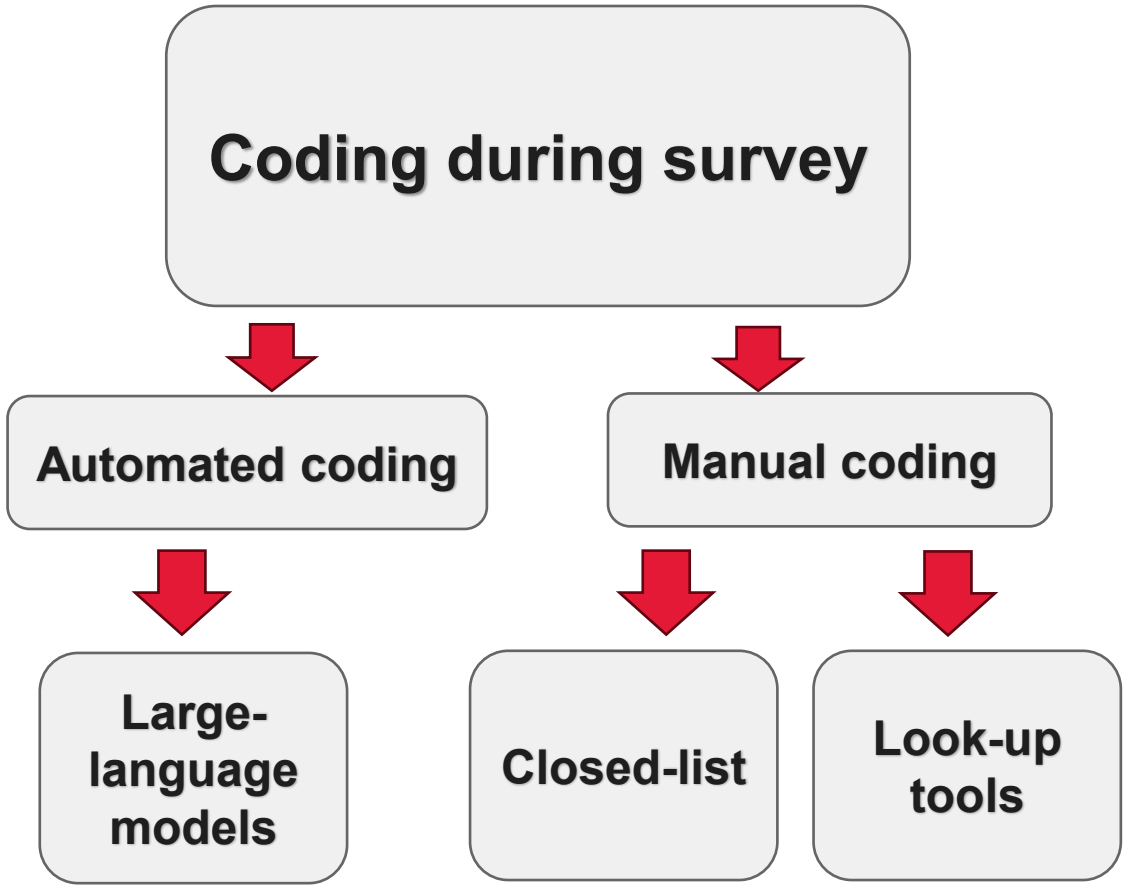
Example approach (Kononykhina & Schierholz, 2026)

- Create embeddings (numerical representations of text) for standard occupational descriptions
- Convert respondents' free-text job descriptions into embeddings
- Calculate **cosine similarity** between respondent descriptions and occupation descriptions
- Return the most likely occupation codes

Coding Approaches



Coding Approaches



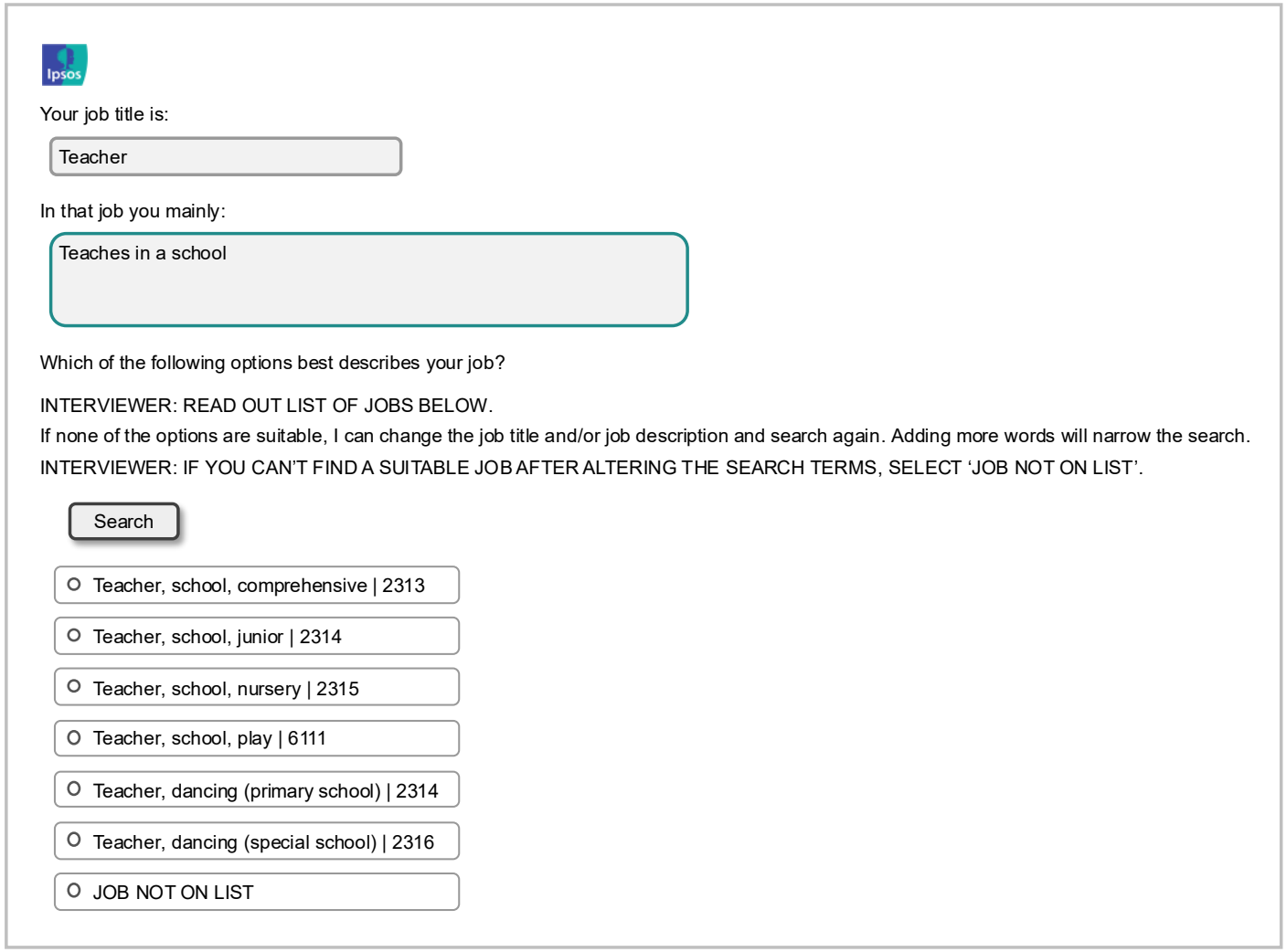
Coding During Survey:
Manual Coding: Look-Up Tool

Approach:

1. Respondents first provide a free-text description of their job
2. A system automatically generates a list of possible occupation codes or categories
3. The respondent (or interviewer) selects the best match from the list

- Coding responsibility shifts from expert coders to participants
- Promises fast and cost-efficient data

Look-Up Approach



Evidence on a Look-Up Tool

Next Steps Age 32 and mode experiment (Kocar et al. 2026; Koerber et al., 2026)

Strengths

✓ High coding success rates (around 80–90% of respondents selected a code)

✓ Respondents generally reported that the selected occupation matched their job

✓ Performed mostly similarly across web and interviewer-administered modes

Challenges

✗ Only moderate agreement with office-coded occupations (around 65%)

✗ Moderate stability over time: only around 60% selected the same occupation code when re-interviewed two weeks later

Coding During Survey:
Manual Coding: Closed-List Approach

Approach:

- Respondents choose their industry and occupation from **pre-defined categories**
- Closed-list questions are frequently used across surveys that only require higher-level classifications of industries and jobs
- A challenge is that category labels can be difficult to interpret for respondents

Closed-List Approach: Industry

Section	Industry
A	Agriculture, forestry and fishing
B	Mining and quarrying
C	Manufacturing
D	Electricity, gas, steam & air conditioning
E	Water supply, sewerage and waste management
F	Construction
G	Trade (retail sales and wholesales) and vehicle repair
H	Transportation and storage
I	Accommodation and food service
J	Information and communication

Section	Industry
K	Finance and insurance
L	Real estate
M	Professional, scientific and technical
N	Administrative and support services
O	Public administration and defence
P	Education
Q	Human health and social work
R	Arts, entertainment and recreation
S	Other service activities
T	Employed by a household as domestic staff
U	International organisations and bodies

Evidence on a Closed-List Approach

NatCen Panel Survey (UK) (Domarchi et al., 2026)

Main findings:

- Agreement between respondent-selected categories and professional coding was moderate
- Accuracy varied across different industry groups
- Some respondents appeared to interpret categories differently from professional coders
- Approach is unlikely to be suitable if high degree of accuracy is required

Coding During Survey:
Automated Coding: Emerging AI Approaches

Approach:

- Large Language Models (LLMs) assign occupation codes in real time during the survey
- Generate follow-up questions when information is incomplete or ambiguous

Potential advantages

✓ Reduced costs and faster data

✓ Lower respondent burden

✓ Potentially higher coding accuracy through interactive questioning

Examples

- **SOCBot** (LSE; Sturgis et al., 2026)
- **SurveyAssist** (ONS)

Summary and Conclusion

- **Traditional office coding** remains effective, but its status as the “gold standard” is challenged by variation across organisations and collection modes
- **Look-up tools show potential but require further development** to improve usability and better reflect the language respondents use to describe their jobs and industries
- **Closed-list approaches reduce burden and costs**, but respondents often struggle to identify with higher-level classification categories, potentially reducing coding accuracy and detail
- **Respondent experience is critical**: many respondents find it difficult to map their occupation or industry to formal classification systems
- **Large Language Models (LLMs)** offer promising new possibilities for post survey coding of open text as well as real-time coding during the survey, but further validation and testing are needed before widespread adoption

SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

Thank you for your attention!

University of Essex

University of Southampton

UKRI Economic and Social Research Council

CITY UNIVERSITY OF LONDON EST 1894

Ipsos

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

MANCHESTER 1824 The University of Manchester

National Centre for Social Research

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

Office for National Statistics

UCL

Ulster University

UNIVERSITY OF LEEDS

UNIL | Université de Lausanne

verian

WARWICK THE UNIVERSITY OF WARWICK