International Machine Learning - with UNECE

"Work Package 1 – ML Classification and Coding working group

Claus Sthamer

Data Science Campus, ONS

2nd April 2020

Claus.Sthamer@ONS.gov.uk

Agenda C&C Workshop 2nd April

1. Opening presentation on the C&C sub work package

1. Presentations on:
   - 15:20 – 15:40 Coding Economic Activity (Serbia)
   - 15:40 – 16:00 Sentiment Analysis of twitter data (Belgium - Flanders)
   - Break – 16:00 – 16:20
   - 16:20 – 16:40 Industry and Occupation Coding (Canada)
   - 16:40 – 17:00 Occupation and Economic activity (Mexico)

2. Re-cap and look at Friday's session, agree on which area's we need to delve deeper and start discussion in our C&C summary report/lessons learned

Eric.Deeben@ONS.gov.uk

Classification & Coding

What is Classification In Machine Learning

Classification is a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories.

Reference: Mohammad Waseem
https://www.edureka.co/blog/classification-in-machine-learning/

Eric.Deeben@ONS.gov.uk

Classification & Coding

Objective:

To test out ML for classification in official statistics

Discussion Point during 1st Sprint in Newport:

- Compare Approaches
- Compare Results
- Common Findings
- Quality Issues → Lessons learned → WP2
- Implementation best Practices → WP3
- Sharing/Transferability

Eric.Deeben@ONS.gov.uk

Web Sentiment

Sentiment analysis labels a body of text as expressing either a positive or negative opinion, as in summarizing the content of an online product review. In this sense, sentiment analysis can be considered the challenge of building a classifier from text. Sentiment analysis can be done by counting the words from a dictionary of emotional terms, by fitting traditional classifiers such as logistic regression to word counts, or, most recently, by employing sophisticated neural networks. These methods progressively improve classification at the cost of increased computation and reduced transparency

Reference: Robert A. Stine
Department of Statistics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; email: stine@upenn.edu

Eric.Deeben@ONS.gov.uk

Proof of Concepts

- **Mexico:** Occupation and Economic activity coding using natural language processing
- **Canada:** Industry and Occupation Coding
- **Belgium:** Sentiment Analysis of twitter data
- **Serbia:** Coding Economic Activity
- **Norway:** Standard Industrial Code Classification by Using Machine Learning
- **USA BLS:** Coding Workplace Injury and Illness
- **Poland:** Production description to ECOICOP

Eric.Deeben@ONS.gov.uk

To be delivered

- **Australia:** [Title tbc]
- **USA IMF:** Automated Coding of IMF's Catalogue of Time Series (CTS)
- **Statistics Iceland:** Automatically coding of SIC and SOC for social surveys

Eric.Deeben@ONS.gov.uk

Progress since September

- **Statistics Canada have put their project into use for two surveys (CCHS & CHMS)**
- **Norway's project is now used in supporting classification.**
- INEGI plan to use more data & explore new methodologies.
- Statistics Poland have presented project with good reception.
- Belgium (Flanders) are to address challenges faced and have laid out next steps they wish to take.
- SORS are in the process of getting better results.

Eric.Deeben@ONS.gov.uk

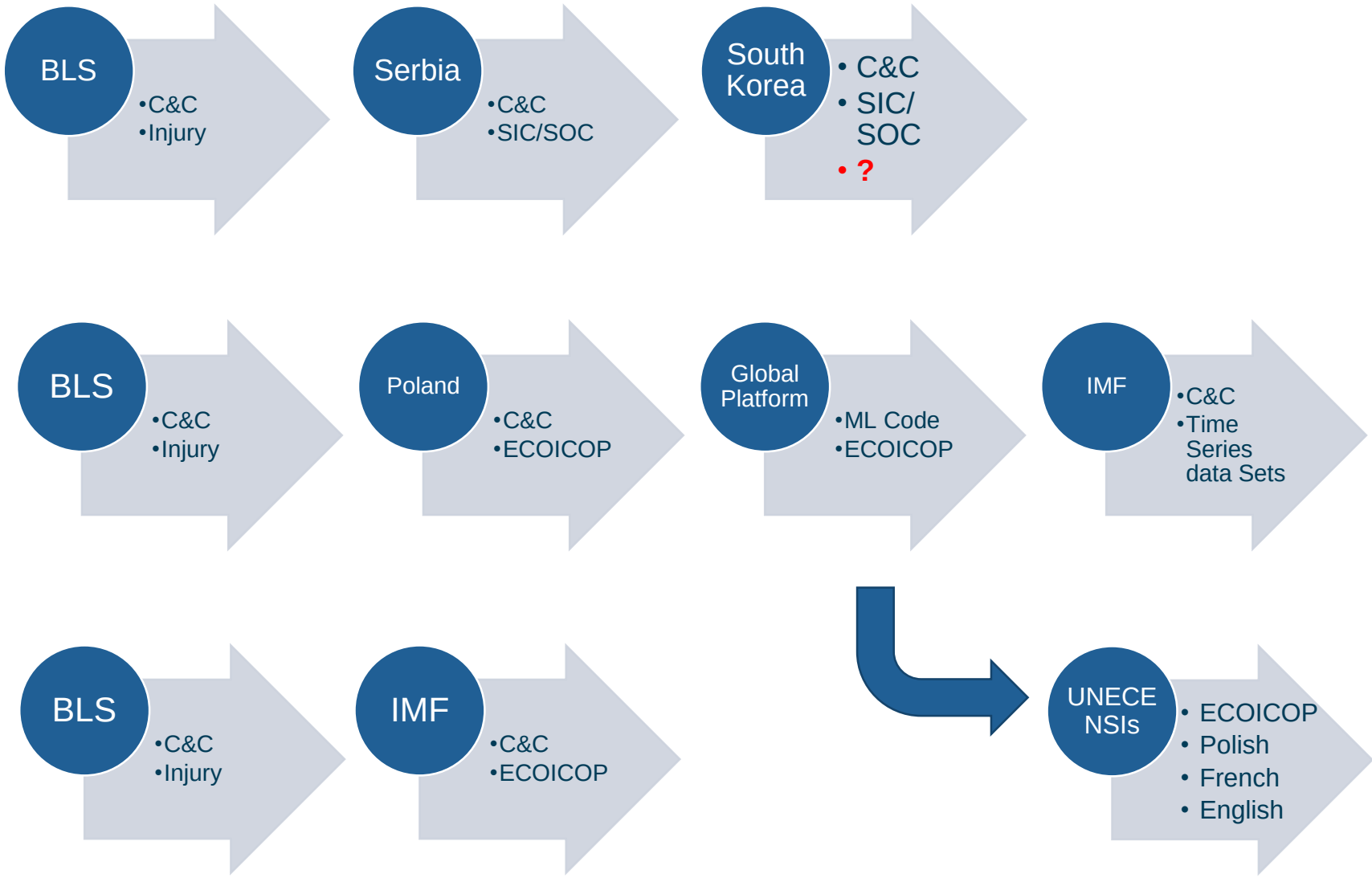
International Collaborations

C&C on product descriptions

C&C on industry

C&C on Web sentiment

Eric.Deeben@ONS.gov.uk



Eric.Deeben@ONS.gov.uk

Classification & Coding

Objective:

To test out ML for classification in official statistics

Discussion Point during 1st Sprint in Newport:

- Compare Approaches
- Compare Results
- Common Findings
- Quality Issues → Lessons learned → WP2
- Implementation best Practices → WP3
- Sharing/Transferability

Eric.Deeben@ONS.gov.uk

Classification & Coding

Q1 : Write down your expectation of the virtual workshop

Q2: Can I ask everyone to read each report in the C&C category

Please respond to Q1 on the WebEx chat.

Eric.Deeben@ONS.gov.uk

C&C Workshop Friday 3rd April

1. Presentation from Statistics Norway - Machine Learning on the Classification of Economic Activities

1. Sub group presentation on:
   - Standard Industrial Code Classification (Norway)
   - Coding Workplace Injury and Illness (USA - BLS)
   - Product description to ECOICOP (Poland)

2. Re-cap two sessions, agree on topics for deeper discussion for sessions for next week

Eric.Deeben@ONS.gov.uk