

# Machine Learning Group 2021

## Work Streams and Activities

**WS1 – Pilot studies: from Idea to Valid solutions**

### Theme 1: Coding and Classification

- 1.1. Apply ML techniques to classification and aggregation web scraped price data, IBGE Brazil - [Vladimir Miranda](#)
- 1.2. Using Big Data Tools and Machine Learning Techniques to Assign Classification of Individual Consumption by Purpose (COICOP) Categories, Turkey Statistical Institute (TURKSTAT) - [Mustafa Karamavus](#)
- 1.3. Code Classification using Multilingual Transformer-based Models, Danish Business Authority - [Casper Eriksen](#)
- 1.4. Coding and Classification: Automated coding of classifiers as a shared service, INE Chile - [Klaus Lehmann](#)

[Wiki-page](#)

### Theme 2: Imputation

- 1.5. Multiple imputation through machine learning, Statistical Office in Rzeszów, Statistics Poland - [Sebastian Wójcik](#)

[Wiki-page](#)

### Theme 3: Imagery Analysis

- 1.6. Estimating Malaysia Rubber Plantation Area Productivity Using Satellite Imagery and Machine, Department of Statistics Malaysia - Rajkumar a/l V.Nagarethinam
- 1.7. Feasibility study of Satellite Imagery Analysis for Wealth Index Development in Indonesia, BPS Statistics Indonesia - Arie Wahyu

[Wiki-page](#)

### Theme 4: Modelling for Estimation

- 1.8. State level expenditure estimates based on ML techniques, U.S. Bureau of Labour Statistics - [Yezzi Lee](#); [Clayton Knappenberger](#)

[Wiki-page](#)

### Theme 5: Transferring Knowledge and Experience

- 1.9. Replicating successful data science projects across NSOs, Statistics Flanders - [Michael Reusens](#)

[Wiki-page](#)

## ML2021 Presentations

| Date     | Speaker   | Presentation   |
|----------|---|--|
| 26 April | Kate Burnett-Isaacs (Statistics Canada)               | HLG-MOS Synthetic Data Project ( <a href="#">presentation slides</a> )                                       |
| 22 March | Sigrid van Hoek (Statistics Netherlands)              | Fair algorithms project ( <a href="#">presentation slides</a> )  |
|          | Lily O'Flynn and Simon Whitworth (UK ONS)             | UK SA Data Ethics ( <a href="#">presentation slides</a> )  |
| 23 Feb   | Riitta Piela and Rok Platinovsek (Statistics Finland) | Best practices in maintaining the quality of data in ML developments ( <a href="#">presentation slides</a> ) |
|          | Casper Eriksen (Danish Business Authority)            | Multilingual Classification of Economic Activities ( <a href="#">presentation slides</a> )                   |
|          | Michael Reusens (Statistics Flanders)                 | WS1 Theme 5: Transferring Knowledge and Experience ( <a href="#">presentation slides</a> )                   |

## How to join the Group ?

Do you have any ML topic you are interested in working together with peers? Do you have any issue (technical, strategic, organisational) you want to discuss with other NSOs? Contact [ML2021](#) (ML2021 at ons dot gov dot uk), copying [UNECE](#) (choii at un dot org) if you want to join the Machine Learning 2021 Group!

[ML 2021 Group Structure and Workstreams](#)

[Terms of Reference](#)

If you are a member of the [Global Network of Data Officers and Statisticians](#), you can follow us from the [ML for Official Statistics group in the Network](#) (see [quick guide](#) on how to join the Global Network).

## Progress Update

### January 2021

Just under 100 members joined for the inaugural Machine Learning 2021 (ML2021) meeting on 29 January 2021.

The meeting ratified the groups governance structure and schedule for 2021. Submitted activity proposals were grouped into Workstreams in advance and confirmed with members.

In anticipation of the February meeting, members have signed up to workstreams and leads have been asked to confirm their output and objectives.

? Unbekannter Anhang

## Theme 6: Route Optimisation

1.10. Route Optimisation through genetic algorithm, INE Chile - [Jose Bustos](#)

[Wiki-page](#)

### WS2 – From Valid Solution to Production

Activities

2.1. Automated production tool to code IMF member state time series data using ML algorithms, International Monetary Fund - Ayoub Mharzi

2.2. Deployment of a Data Lake architecture to put into production data science projects, INEGI Mexico - [Abel Coronado](#)

2.3. Design and assess a whole workflow to enable Natural Language Processing and Machine Learning methodologies to be integrated into a continuous production process, INEGI Mexico - [Jael Pérez](#); [Alejandro Ruiz](#)

2.4. A technical platform that supports the whole machine learning process and thereby ensures the quality of that process and in the end contribute the overall quality of the statistical output, Statistics Sweden - [Alexander Thorell](#)

### WS3 – Data Ethics and Governance

3.1 The establishment of a set of ethical principles to provide a clear framework to enable ethical use of Machine Learning for research and statistics, UK Statistics Authority - [Lily O'Flynn](#); [Simon Whitworth](#)

### WS4 – On The Quality of Training Data

4.1 Identifying the circumstances under which an ML model should be retrained in order to maintain the predictive power and quality of the model, Statistics Finland - [Riitta Piela](#); [Rok Platinovsek](#)

### WS5 – On The Quality Framework for Statistical Algorithms

5.1 Explore dimensions of QF4SA in a consolidated project to analyse an output based on a set of standard metrics and procedures INEGI Mexico - [Jose Jimenez](#); [Alejandro Ruiz](#)

## Resources

Machine learning competency in context of Big Data training and human resources (from [UN Global Working Group on Big Data Task Team on Skills](#) and it is relevant to the community)

- [Competency Framework for big data acquisition and processing](#) (chapter 4)
- [Big Data Maturity Matrix](#)

## Group 2020 Planning

### Machine learning Group 2021

Following the great interest in continuing the work of the HLG-MOS Machine Learning Project, UK Office of National Statistics (ONS) is launching the Machine Learning Group 2021. The new ML group will focus on developing and implementing ML for official statistics (see slides below for presentation and live survey conducted during the [HLG-MOS ML Project webinar](#))



Webinar Future Directions.pdf



ML Project Webin... Poll Result.pdf

**Table 1: Work packages investigated by the ML 2019/2020 project (hyperlinks) and potential themes to be explored by the ML 2021 project (text in red).**

|                |  |  |   |                                  |   |   |               |
|----------------|--|--|---|----------------------------------|---|---|---------------|
| The Journey    | Moving from idea to valid solution (demonstration)   |  | Moving from valid solution to production (Operationalisation) |                                  | Ensuring production robustness (Maintenance)    |   |               |
|                | All WP1 pilot studies  |  | Some WP1 pilot studies  |                                  | Very few WP1 pilot studies                      |   |               |
|                | Other applications of Machine Learning   |  | Some other applications of Machine Learning                   |                                  | Very few other applications of Machine Learning |   |               |
|                | WP3 Integration (Q5 & Q6)  |  | WP3 Integration (Q5)  |                                  |   |   |               |
|                | Workstream 1: Support current studies towards production; welcome new studies in other processes (e.g. record linkage) and/or data sources (e.g. satellite data)   |  |   |                                  |   |   |               |
| Supported by   | Quality (accuracy, timeliness, efficiency, explainability and reproducibility)   | Good Training Data   | Skills/Competences  | Computing Infrastructure         | Interoperability / Business Process             | Ethics and Legal                                  | Security      |
|                | WP2 Quality  |  | WP3 Integration (Q3 & Q4)                                     |                                  |   |   |               |
|                | Workstream 2: Experiment with practices and methods on some dimensions of QF4SA (WP2);<br><br>Workstream 3: Review and improve the Framework   | Workstream 4: How to get good training data, how to keep it up to date, when to relearn a model, what does 'good' mean, how to measure that? | Workstream 5: What skills? How to learn? Where to find them?  | To be defined                    | To be defined                                   | Workstream 6: Ethics handbook, regulations , etc. | To be defined |
|                |  |  |   |                                  |   |   |               |
| Facilitated by | Organisation   |  |   | Sharing and Collaboration        |   |   |               |
|                | WP3 Integration (Q1 & Q2)  |  |   | HLG-MOS Machine Learning Project |   |   |               |
|                | Initiatives to accelerate the integration of machine learning solutions  |  |   | ML Studies and Codes             |   |   |               |
|                | Workstream 7 : Create/maintain a network of data science unit leaders;<br>Workstream 8: Beyond 2021: How can we better prepare for next 2-5 years? What technology and data sources can we expect? What skills will we need? |  |   | Learning and Training            |   |   |               |
|                |  |  |   | HLG-MOS ML Project webinar       |   |   |               |

## Progress status

Update from ONS (December 15, 2020)



ML 2021 update.pptx