ONS-UNECE Machine Learning Group 2022

UNECE Statistical Division
Workshop on the Modernisation of Official Statistics

Project Report

Alison Baily
International Programme Manager, ONS

23 November 2022

Data Science Campus

Contact: ML2022@ons.gov.uk

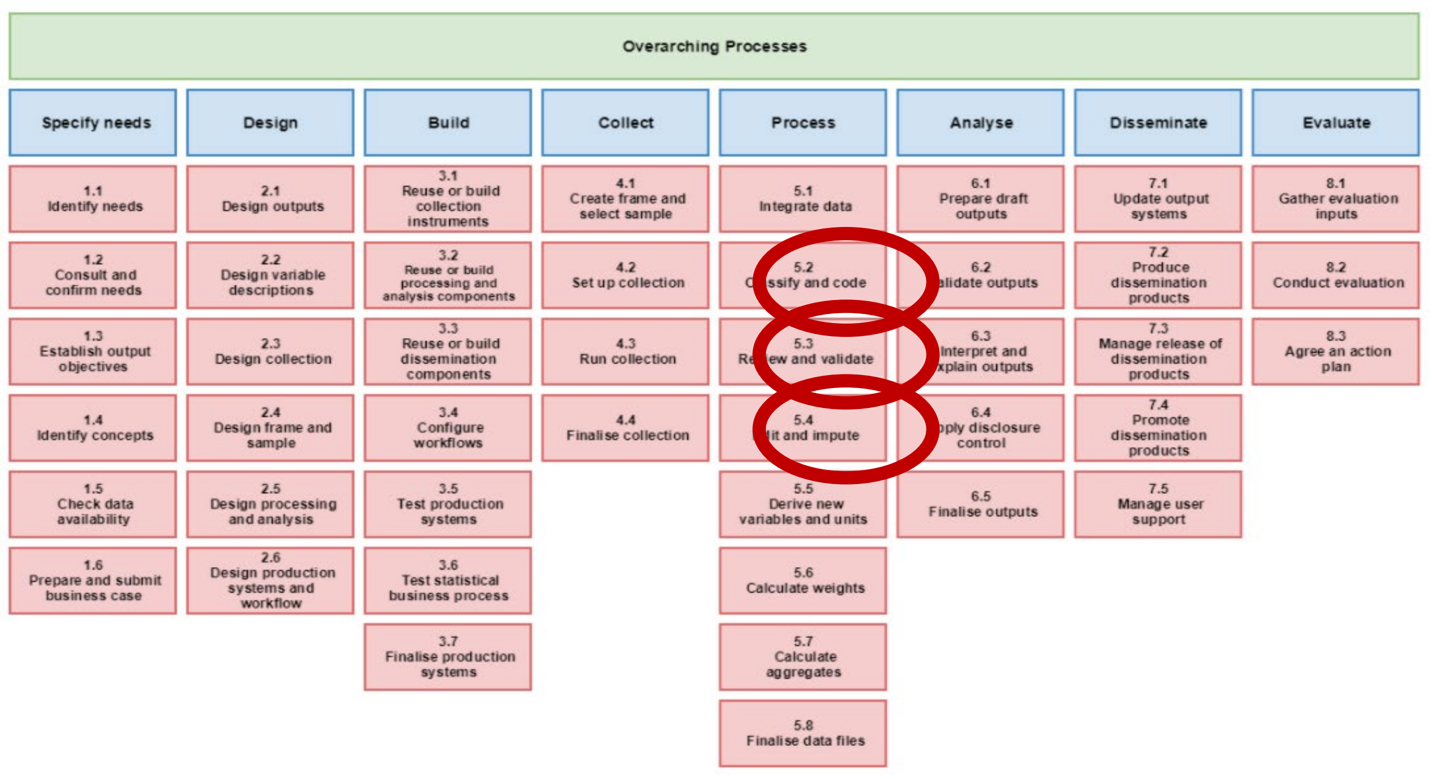
Overview

1. Background
2. Programme Development & Objectives
3. Activity
4. Impact and Lessons Learned
5. Future plans

Background

Machine Learning: “a field of study that gives computers the ability to learn without explicitly being programmed”

Machine learning – application areas



Overarching Processes

Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysis components	4.2 Set up collection	5.2 C ssify and code	6.2 alidate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Re ew and validate	6.3 Interpret and xplain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 lit and impute	6.4 pply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing and analysis	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production systems		5.7 Calculate aggregates			
				5.8 Finalise data files			

Areas with manual, repetitive tasks can be automated with the help of machine learning

Objectives

Platform to facilitate

- Research to modernise official statistics
- Building capacity in machine learning
- Sharing knowledge (data, methods, use cases)

Community driven

- Members design programme and provide content
- Interaction and collaboration is key
- Every contribution is welcome!

Public good

- Open to all official statistical organisations
- Accessible to different levels of expertise
- Resources shared with wider community

What we do

Knowledge Sharing

- Monthly meetings with expert presentations
- External engagement at international conferences
- Regular updates of ML news and opportunities

Research Collaboration

- Research projects explore issues from design to implementation
- Findings shared on public website

Capacity Building

- Coffee and Coding sessions
- Learning and training resources

Membership

Public

- Public Website
- Final report + webinar
- Coffee and Coding Sessions

All members

- Monthly meeting
- Newsletter
- Members website
- Catalogue
- Contribute input where possible

Themes

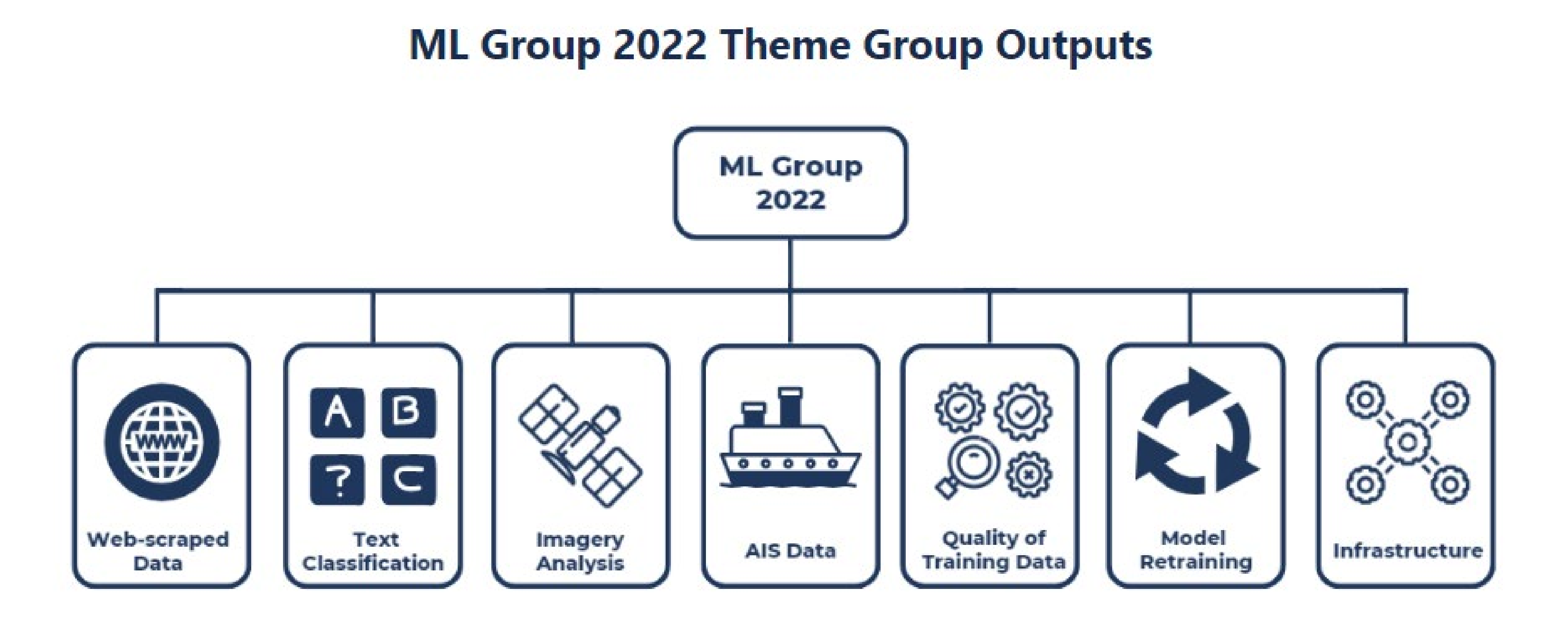
- Research projects
- Study groups
- External presentations
- Regular collaboration

Programme Development and Goals

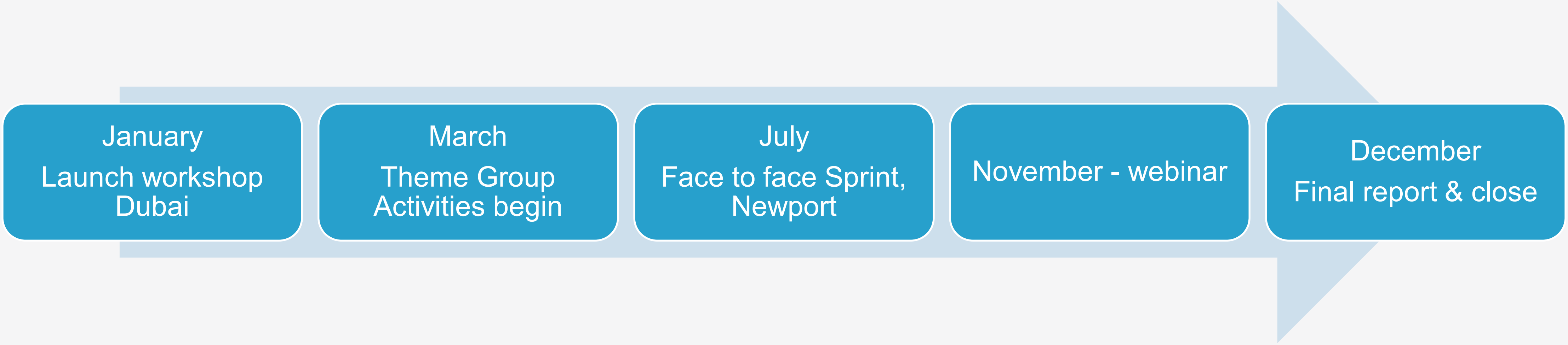
Goals for 2022

- Focus on moving from proof of concept to production
- Other key areas: C&C, ethics, quality of training data.
- Research collaboration & capability
- A hub for ML news and networking
- Increase membership and active participants

Priority themes for 2022



2022 Timeline



January

Launch workshop Dubai

March

Theme Group Activities begin

July

Face to face Sprint, Newport

November - webinar

December

Final report & close

Activities and Results

Research Collaboration & Knowledge Exchange

- Web Scraping Data Theme Group
  - Implementation of experimental statistics using web scraped data for identifying companies active in particular sectors
  - Platform for sharing use cases + discussion best practice

- AIS Modelling Theme Group
  - Exploring methods to identify berth areas using ML and AIS data
  - Testing methods at a larger scale + collating guidance for SOs

Research Collaboration & Knowledge Exchange

- Imagery
  - Research Group – exploring papers on CNN architecture, class imbalance, explainable AI
  - Study Group – building core skills through courses and discussion
  - Platform for sharing use cases + discussion of best practice

- Text Classification
  - Platform for sharing use cases + discussion of best practice

Research Collaboration & Knowledge Exchange

- Quality of Training Data
    - Focus on understudied area
    - Identified sources of error that affect model quality

- Model Retraining
    - Examined key concepts of drifts (model, concept, data)
    - Identified factors that enable monitoring and retraining

Research collaboration & knowledge exchange

- Infrastructure Group
  - Exploring supporting environment and infrastructure for scaling up ML projects
  - Discussion of organisation experience on cross-cutting issues such as:
    - Linking ML processes with traditional production processes
    - Generic patterns for ML deployment & servers
    - Roles and capabilities

Monthly Forum

- Main meeting point for the community
- Expert presentations from statistical organisations and academia
- 13 presentations and 7 meetings
- C. 100 members attend each meeting

Sprint @ ONS, July

- 21 members from 14 different organisations
- Model retraining, quality of training data and web scraping data
- Networking with other international data science groups and national statistical organisations

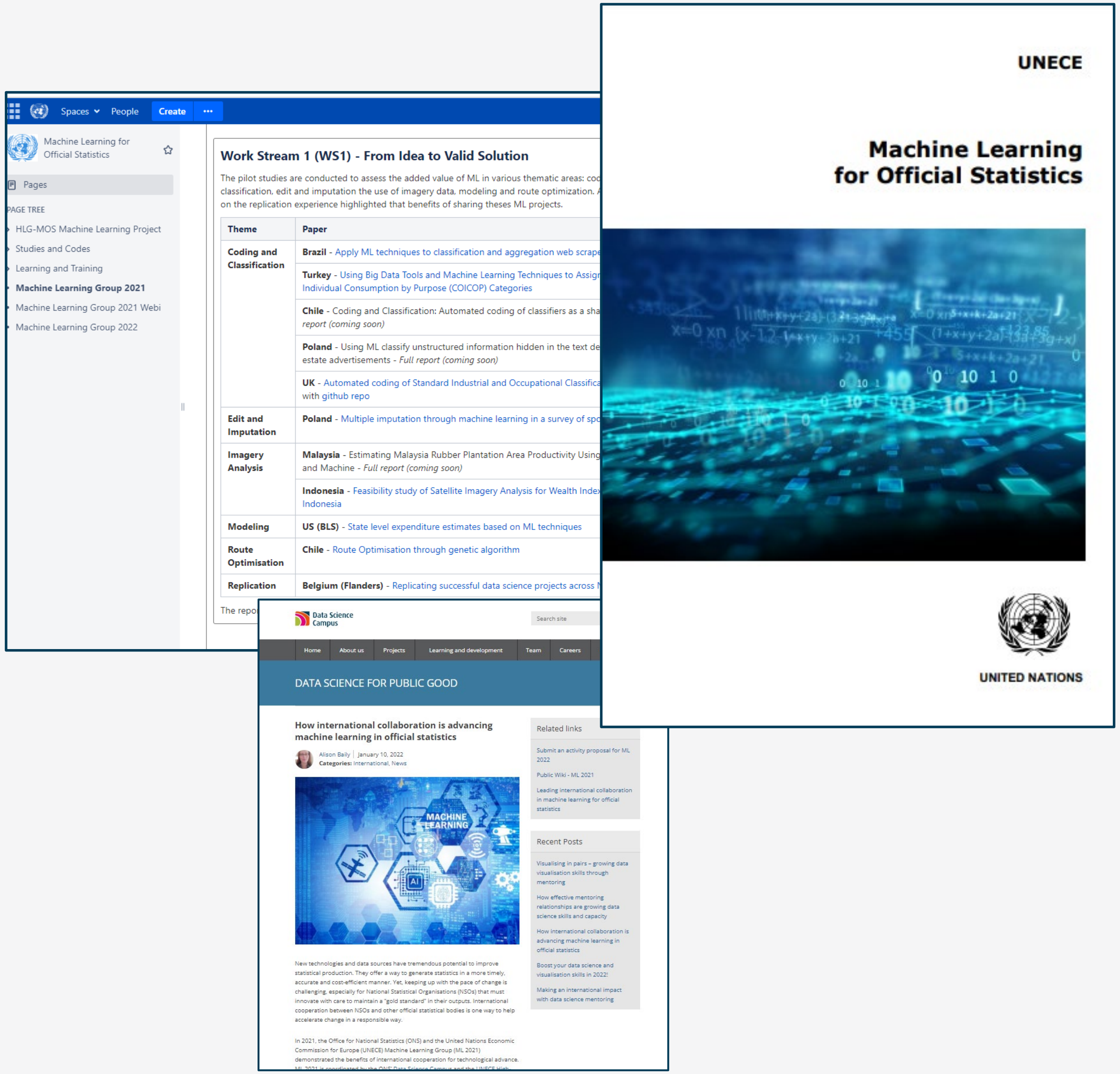
Capacity Building

- Coffee and Coding
  - ML Fundamentals & ML Applications Deep Dive
  - ML Foundations for Non-Programmers
  - Git

- ML strategy workshops
  - UN Regional Hubs for Big Data
  - Middle East, Latin America, Indonesia.

Communications

- Website
- Discussion forum
- Conference presentations
- Guides
- Papers
- Youtube channel
- ML Group video (forthcoming)
- Webinar November 30th



Spaces
People
Create
•••
Machine Learning for Official Statistics
Pages
PAGE TREE
HLG-MOS Machine Learning Project
Studies and Codes
Learning and Training
Machine Learning Group 2021
Machine Learning Group 2021 Webi
Machine Learning Group 2022

Work Stream 1 (WS1) - From Idea to Valid Solution

The pilot studies are conducted to assess the added value of ML in various thematic areas: cod classification, edit and imputation the use of imagery data, modeling and route optimization. A on the replication experience highlighted that benefits of sharing theses ML projects.

Theme	Paper
Coding and Classification	**Brazil** - Apply ML techniques to classification and aggregation web scrape
	Turkey - Using Big Data Tools and Machine Learning Techniques to Assign Individual Consumption by Purpose (COICOP) Categories
	Chile - Coding and Classification: Automated coding of classifiers as a sha *report (coming soon)*
	Poland - Using ML classify unstructured information hidden in the text de estate advertisements - *Full report (coming soon)*
	UK - Automated coding of Standard Industrial and Occupational Classifica with github repo
Edit and Imputation	**Poland** - Multiple imputation through machine learning in a survey of spo
Imagery Analysis	**Malaysia** - Estimating Malaysia Rubber Plantation Area Productivity Using and Machine - *Full report (coming soon)*
	Indonesia - Feasibility study of Satellite Imagery Analysis for Wealth Index Indonesia
Modeling	**US (BLS)** - State level expenditure estimates based on ML techniques
Route Optimisation	**Chile** - Route Optimisation through genetic algorithm
Replication	**Belgium (Flanders)** - Replicating successful data science projects across N

The repo

How international collaboration is advancing machine learning in official statistics

Alison Baily | January 10, 2022
Categories: International, News

New technologies and data sources have tremendous potential to improve statistical production. They offer a way to generate statistics in a more timely, accurate and cost-efficient manner. Yet, keeping up with the pace of change is challenging, especially for National Statistical Organisations (NSOs) that must innovate with care to maintain a "gold standard" in their outputs. International cooperation between NSOs and other official statistical bodies is one way to help accelerate change in a responsible way.

In 2021, the Office for National Statistics (ONS) and the United Nations Economic Commission for Europe (UNECE) Machine Learning Group (ML 2021) demonstrated the benefits of international cooperation for technological advance. ML 2021 is coordinated by the ONS' Data Science Campus and the UNECE High-

Related links

- Submit an activity proposal for ML 2022
- Public Wiki - ML 2021
- Leading international collaboration in machine learning for official statistics

Recent Posts

- Visualising in pairs – growing data visualisation skills through mentoring
- How effective mentoring relationships are growing data science skills and capacity
- How international collaboration is advancing machine learning in official statistics
- Boost your data science and visualisation skills in 2022!
- Making an international impact with data science mentoring

UNECE

Machine Learning
for Official Statistics

UNITED NATIONS

Lessons learned & Impact

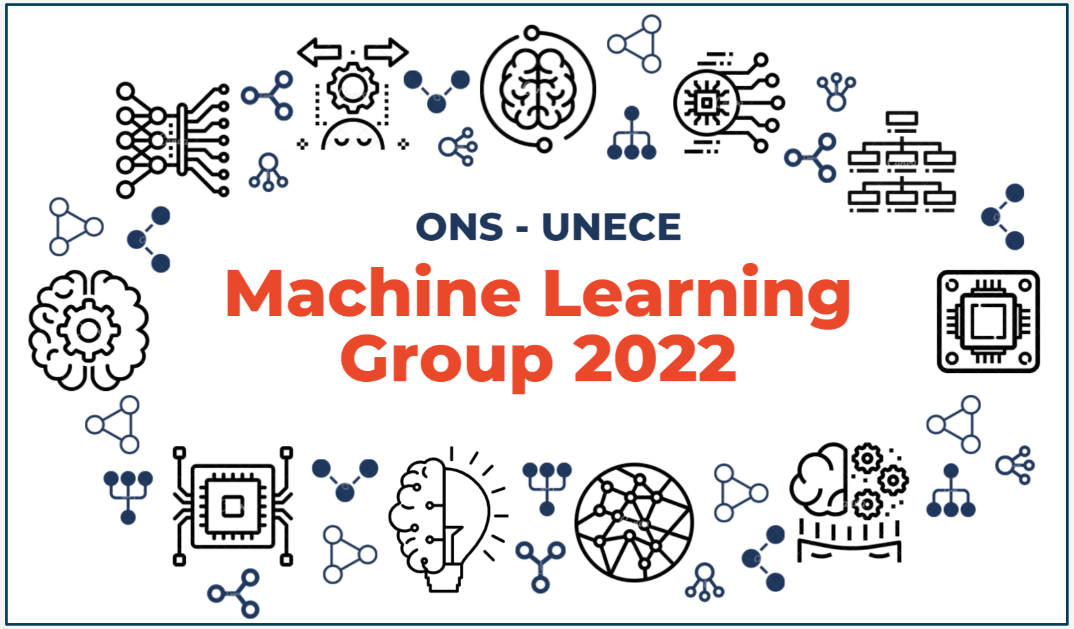
Impact

- A place to explore and test ideas
- Sharing of tried-and tested approaches
- Enabling organisations new to ML to accelerate their development
- Addressing common production challenges
- Raising profile of ML among strategic decision-makers
- Building staff skills

What our members say

“enabled me to access a vast repertoire of experience on the use of ML for the production of official statistics"

“helped me understand ML in the context of official statistics and government data science"



“It motivated my team to increase their ML skills”

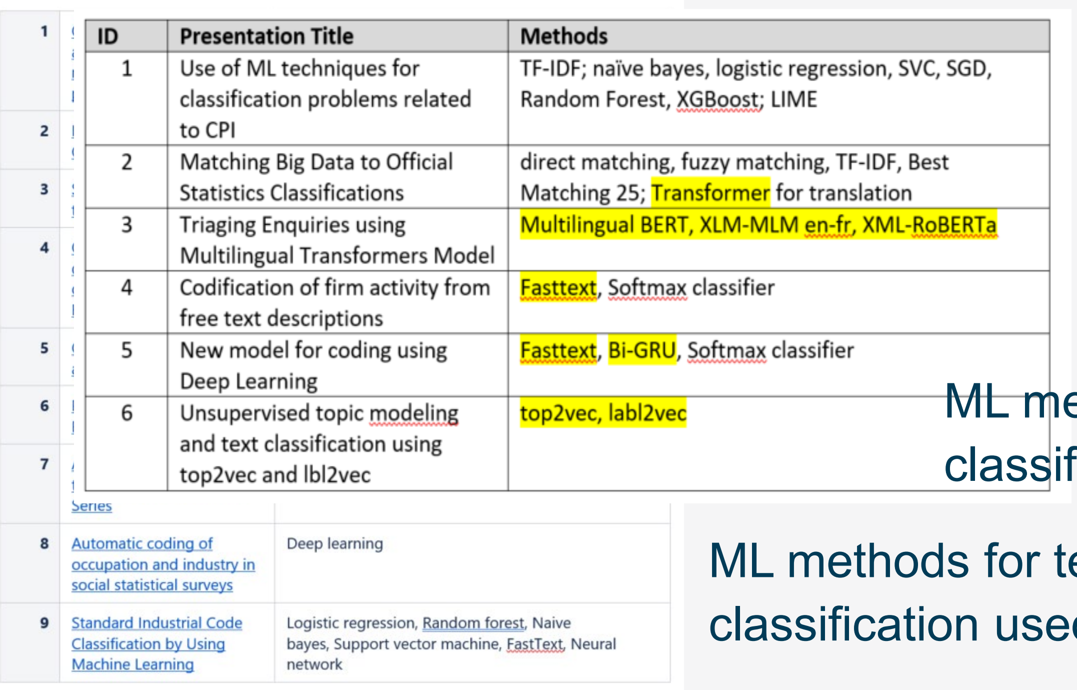
“I built awareness of different uses for ML applications and how ML applications combine with other tools in the statistical processes."

“Learning which projects other organisations are successfully doing helps us allocate our limited ML capacity."

Membership survey, 2021

Lessons learned

1. Fast-changing field



ID	Presentation Title	Methods
1	Use of ML techniques for classification problems related to CPI	TF-IDF; naïve bayes, logistic regression, SVC, SGD, Random Forest, XGBoost; LIME
2	Matching Big Data to Official Statistics Classifications	direct matching, fuzzy matching, TF-IDF, Best Matching 25; Transformer for translation
3	Triaging Enquiries using Multilingual Transformers Model	Multilingual BERT, XLM-MLM en-fr, XML-RoBERTa
4	Codification of firm activity from free text descriptions	Fasttext, Softmax classifier
5	New model for coding using Deep Learning	Fasttext, Bi-GRU, Softmax classifier
6	Unsupervised topic modeling and text classification using top2vec and lbl2vec	top2vec, labl2vec

ML methods for text classification used in 2022

1		
2		
3		
4		
5		
6		
7	Series	
8	Automatic coding of occupation and industry in social statistical surveys	Deep learning
9	Standard Industrial Code Classification by Using Machine Learning	Logistic regression, Random forest, Naive bayes, Support vector machine, FastText, Neural network

ML methods for text classification used in 2019-20

Lessons learned

2. More ML use cases for statistical orgnaisations

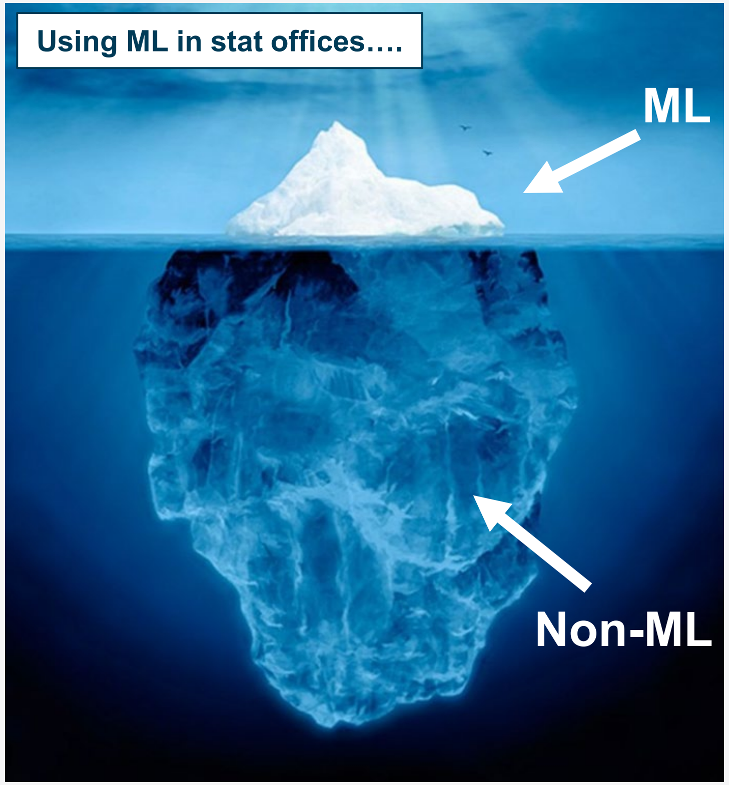
- Text classification and imagery analysis continue to be popular use case
- Use cases outside usual application areas: Predict the respondents to follow up, create a more evidence-based survey frameworks, Triage the multilingual customer inquiries

3. Quality

- Different importance for different stakeholders and different use case
- Unique expertise that statistical organisations (e.g., quality control process) can help

Lessons learned

4. Building infrastructure needed for integration



Model registry

Model serving

Data security ……….

Monitoring and re-training

………..

Data management

Standardisation

Quality control

……… Versioning

Ddocumentation

………

Lessons learned

5. High demand for knowledge exchange and capacity building

- Interest in capacity building for machine learning remains high
- Community is dynamic, enthusiastic and increasingly experienced
- Lack of time resource limits deeper engagement
- Quality and range of activities requires dedicated staff resource
- Partnering with other international groups is beneficial

Future plans

Plans for 2023

- Group to close at end of 2022 as ONS and UNECE redeploy resources to other areas
- Exploring more resource-efficient ways to respond to demand
- Discussions with UN Statistics Division

ML Group 2022 Webinar – 30 November

- Session I 1000-1130 CET
  - Applications of machine learning: web scraping data, text classification, imagery data, AIS data
- Session II 1500-1630 CET
  - Statistical production issues: quality of training data, model retraining, IT infrastructure
- Registration open on Eventbrite
  - Go to the link on the Machine Learning 2022 page here

Discussion

- What progress has ML made in your organisation? What role do you think the UNECE ML Group has played in that?
- What lessons could the ML Group's development have for other modernisation projects?
- What are the most important aspects of the group's work that should be continued? What channels can we consider for this?