Machine Learning (ML) holds a great potential for statistical organisations. It can make the production of statistics more efficient by automating certain processes or assisting humans to carry out the process. It also allows statistical organisations to use new data types of data such as social media data and imagery. Many national statistical offices (NSOs) are investigating how ML can be used in the context of official statistics which have to maintain rigorous standards for quality. While specific business environment may vary depending on country, NSOs face similar type of problems which can benefit from developing shared understanding and solution within the broad official statistical community.

To address this need, UNECE HLG-MOS created a Machine Learning Project in 2019. The project aims to demonstrate the value addition of ML, advance capability of NSOs to use ML and identify common issues encountered when incorporating ML in the organisation.

The project started in April 2019 with 23 participants from 13 organisations. Within a year, its membership has grown to 39 participants from 19 NSOs or other organisations, as well as being connected to 50 other individuals from 25 NSOs/organisations who either provide assistance or follow the progress of the project.

The project includes the conduct of numerous pilot studies, the development of quality assurance aspects, and the identification of integration challenges and practices to overcome some of them.
Twenty pilot studies are conducted to assess the added value of ML to **statistical business processes** (coding and classification, edit and imputation) and the use of imagery data. No two pilot studies are alike. They are conducted on a wide variety of data sources (survey respondents, administrative, web-scraped, published official statistics, twitter, satellite images, aerial images) and contexts (survey, census, registers, proof of concepts, production).

In the months to come, the project will deliver several reports (demonstrations of added value and recommended practices), ML code, a hands-on application and references. In the meantime, presentations given at a virtual sprint held in April are provided below as interim outputs so that those who start exploring the use of ML can benefit from the knowledge and experience gained by project at the earliest possible. **The final project outputs will be published on this wiki by the end of 2020.**

**Project Proposal (November 2018)**

**Progress Report (November 2019)**

**Presentations from Virtual Sprint (April 2020)**

<table>
<thead>
<tr>
<th>ML Pilot Study Theme</th>
<th>Pilot Study Presentations</th>
</tr>
</thead>
</table>
| **Coding and Classification** | • Pilot Studies Overview  
• Belgium - Flanders: twitter sentiment analysis  
• Canada: industry & occupation coding  
• Iceland: industry & occupation coding  
• Mexico: economic activity & occupation coding  
• Norway: cloud platform  
• Serbia: economic activity coding  
• USA: occupational injuries & illnesses coding  
• IMF: economic time series coding* |
| **Edit and Imputation** | • Pilot Studies Overview  
• Australia: imputation of dwelling occupancy  
• Germany: imputation simulation study  
• Italy: imputation for education level variable  
• Italy: design the editing process*  
• Poland: estimation of tourist expenditure  
• UK: editing of income data |
| **Imagery** | • Pilot Studies Overview  
• Australia: automated imagery recognition  
• Mexico: satellite imagery analysis  
• Netherlands: image analysis  
• UNECE: pipeline for satellite imagery |

* Most of the pilot studies were initiated at the beginning of the project in May 2019; other projects were launched in February 2020 and are at early stages.