modernstats
BY HLG - MOS

Machine Learning Project

’ML I’ Progress and ‘ML II’ Proposal

Claude Julien – UNECE Project Manager

Outline of ML I and ML II

Initiate *Identify* *Ideas*	**ML I** Background
Create *Generate* *Start*	**ML I** Project launch and planning
Develop *Expand* *Interlink*	**ML I** Project participants and progress **ML I** Integration challenges and ML II proposed review
Share *Support* *Maintain*	**ML I** initial communications ML II proposed outputs and timelines
Transfer *Complete* *Abandon*	ML II project closure and recommendations on future work

Initiate , Identify Ideas

Background

- Risk of irrelevancy
- Stiff competition
- New data sources and methods (new, really?)
- Integration of Machine Learning
- Quality and transparency
- Individual and collective expertise

Initiate , Identify Ideas

2018 BSTN

- Position paper proposing an ML project
  - Interest in ML is rapidly growing
  - Essential for processing some data sources
  - Might offer added value in traditional context
  - Limited experience with concrete applications
- ML project with four work packages
  - WP0. Refine scope
  - WP1. Pilot studies
  - WP2. Quality aspects
  - WP3. Lessons learned (vaguely defined)
- Highly supported by HLG-MOS

Create, Generate, Start

Launch and plan

- March – Hire project manager and create project team (11 participants)
- April – (23 participants) First planning sprint
- May – (27 participants) Planning sprint held in UK
- Limited knowledge and experience in ML
- Most common objective is to learn and share
- Very few integration of ML within the team
- Confirmed the general direction suggested by the BSTN
- Agreement on the overall objectives of the project

Create, Generate, Start

Objectives

Advance the research, development and application of machine learning techniques to add value to the production of official statistics

- Demonstrate the value added of ML in the production of official statistics, where "value added" is increase in relevance, timeliness, better overall quality or reduction in costs.
- Advance the capability of ML to add value to the production of official statistics.
- Advance the capability of national statistical organisations to use ML in the production of official statistics.
- Enhance collaboration between statistical organisations in the development and application of ML.

Create, Generate, Start

Participants

- 38 participants - 18 organisations - 14 countries
- A very good variety of expertise
- A very common high level of engagement
- 31 other persons have access to our work
  - Connect with other international activities
  - Support the studies
  - Identify opportunities for collaboration

more information on participants

list of international activities

Create, Generate, Start

WP and PS Leaders

- WP 1 – Pilot Studies (**Eric Deeben** - UK)
  - Coding and classification lead : **Claus Sthamer** (UK)
  - Edit and imputation: **Florian Dumpert** (Germany)
  - Imagery: **Abel Coronado** & **Jimena Juarez** (Mexico)
- WP 2 – Quality (**Wesley Yung -** Canada)
- WP 3 - ??? (**Alex Measure** USA)
- Invaluable assistance and contributions throughout the project – **InKyung Choi**

Develop, Expand, Interlink

Pilot studies (PS)

- Relevance to the participating organisations
  - Business value
- Relevance to other organisations
  - Demonstrations on added value
  - Recommendations on applicability
  - Access to ML solutions / scripts
  - Best practices
  - Inform and test quality framework and practices

Develop, Expand, Interlink

Communication

- Numerous collaborations within WP and PS
- Monthly progress/update meetings
- FtoF Sprint held in Serbia in September
- One study is discussed at each monthly meeting
- Considerable documentation on **ML wiki space**

See collaborations

Develop, Expand, Interlink

C&C – BLS system

- Five characteristics on injury and illness narratives https://www.bls.gov/iif/autocoding.htm
- Development started in 2012 and gradually implemented since 2014
- Integrated in manual operation; facilitate, not replace
- Clearly demonstrated and monitored improvement in accuracy
- ML automation has moved human resources from coding to review
- Importance and time needed for expert coding
- Needed high-level senior management support to make it happen

Develop, Expand, Interlink

C&C – Pilots

- Starting point: ML should be very good for C&C
- Good variety of topics and data sources
- Results are demonstrating good to very good accuracy
- Potential cost reductions are assumed
  - Quicker setup to introduce automation
  - Lower maintenance cost
- ML scripts, coding expertise and tips are shared
- Dataset shared, translated in different languages and coded it with different ML approaches

see example of C&C pipeline

Develop, Expand, Interlink

E&I

- Starting point: Added value of ML is uncertain
- Good variety of data sources and applications
- Results are better than expected, even counterintuitive
- Potential cost reductions are assumed
  - Quicker setup
  - Lower maintenance cost
- Expertise and tips are shared; results are challenged
- Looking into an opportunity to share a dataset from publicly available sources to generate further collaboration

Develop, Expand, Interlink

Imagery

- Starting point: Imagery is increasingly accessible and ML is essential to exploit it
- Challenge to address:
  - Using imagery and integrating it with other sources is a complex process
  - Not clear to new users who and what is needed, and where, including ML
- Drafted a process pipeline
- Measure urban growth/density
  - Produce new information related to SDGs
  - Share ML scripts
  - Apply and improve the process pipeline
  - Test quality assurance and measurement approaches

to see the process pipeline

Develop, Expand, Interlink

Quality

- Expand on standard quality frameworks to include ML aspects on processes and outputs
- Identify and standardize quality indicators
- Best practices in maintaining and demonstrating quality in production
- Concrete examples through collaboration with pilot studies

Develop, Expand, Interlink

Integration

- Individually learn and demonstrate ML
- Collectively identify best practices to implement ML in production processes
- Recent discussions have raised common organisational challenges
  - Integrating ML in organisations
  - Bringing the required expertise together
  - Management support
- WP3 is now clear to us, i.e. based on a business need
- More on this in the proposal for ML II

recent comments from participants

Share, Support, Maintain

Sharing

- Access to ML wiki space
- Presentations outside the project
  - ModernStats workshop
  - Bureau of Labor Statistics
  - Statistics Canada
  - Destatis
  - Istat Advisory Committee
  - International Conference on Establishment Statistics
  - GEO Week panel on Earth Observation for Official Statistics

Summary

- Progress in-line with BSTN and HLG-MOS
- Knowledge and ability to use ML within the team
- Added value of ML
- Identification of implementation practices
- Collaboration within participating organisations
- Sustained interest in the project
- Opportunities for collaboration keep arising

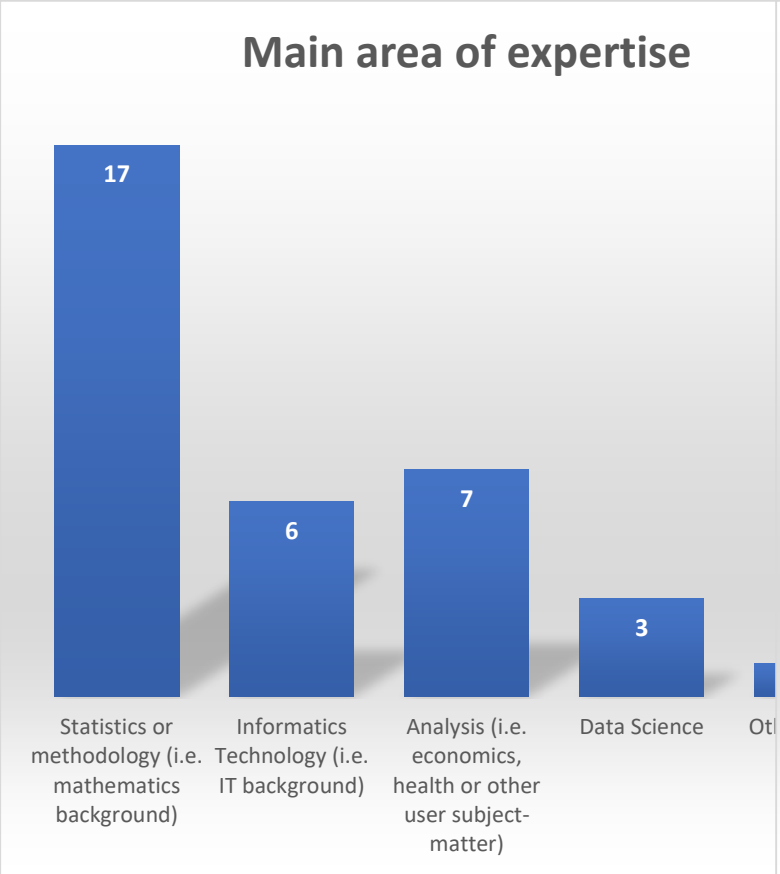
Thank you!

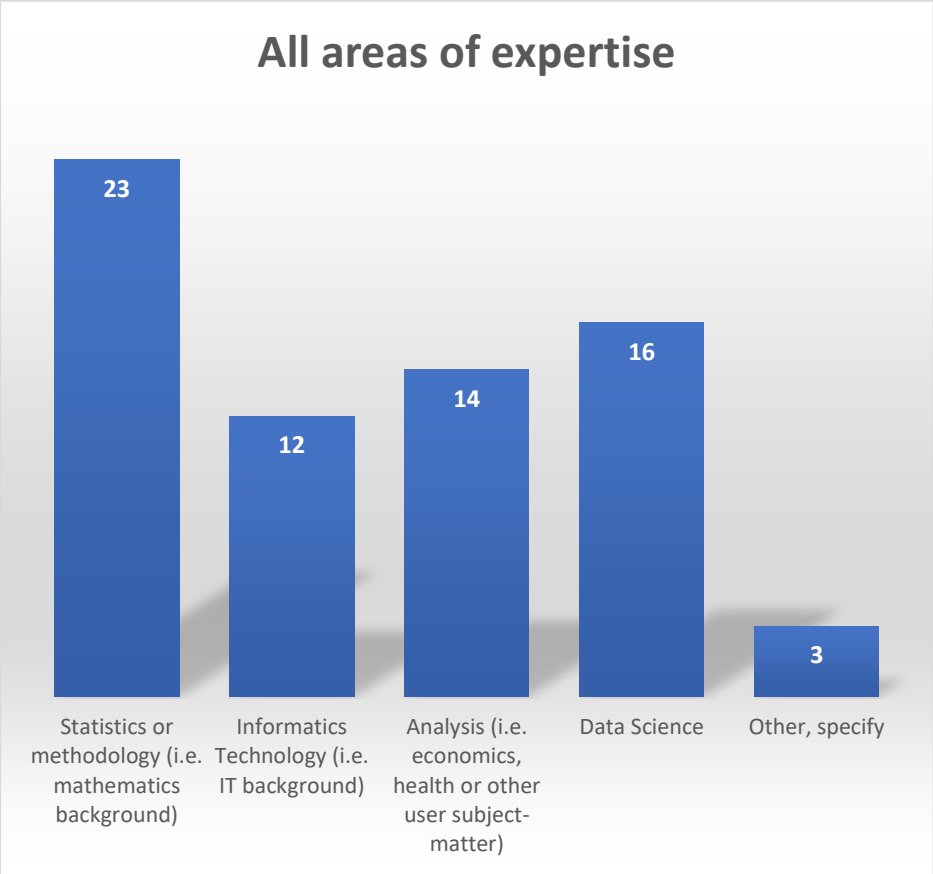
Project participants

- 38 participants from 18 organisations in 14 countries

- 31 followers and additional collaborators from 14 organisations in 9 countries

Project participants
Multi-disciplinary expertise





Back to Participants slide

Develop, Expand, Interlink

Collaborations

- C&C on product descriptions
- C&C on industry
- C&C on Web sentiment
- C&C Quality Assurance
- E&I
- Imagery
- Quality

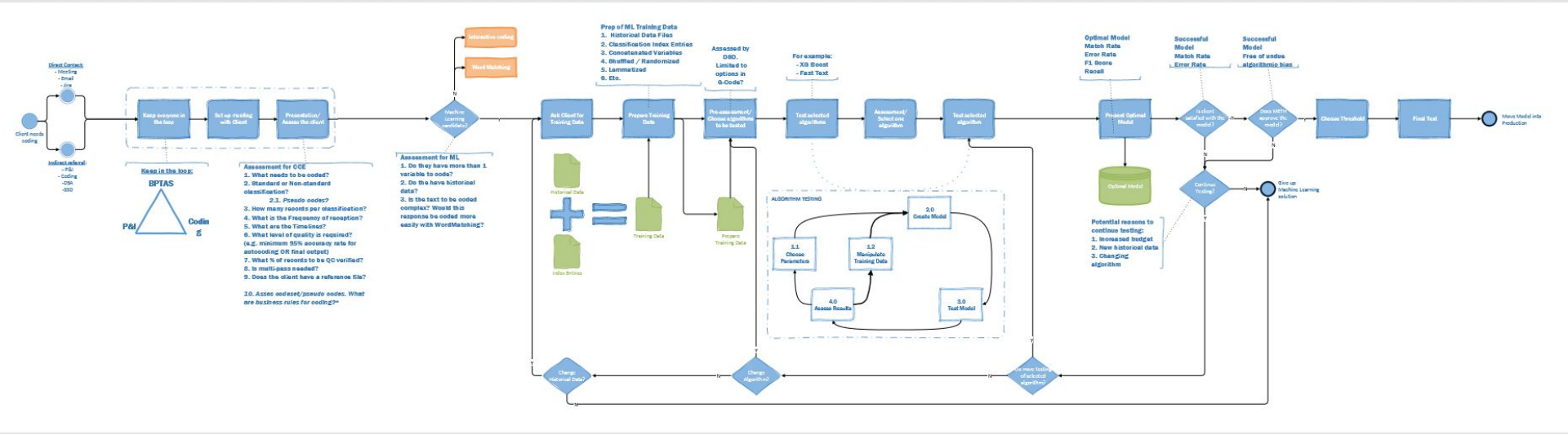
back to Communication

List of international activities in which ML project participants are involved

- UN Global Working Group for Big Data - Satellite Imagery and Geo-spatial data
- ESSNet Big Data: from exploration to exploitation
- Copernicus – Europe's eyes on Earth
- UN-GGIM Inter-Agency and Expert Group on the Sustainable Development Goal Indicators (IEAG-SDGS) Working Group on Geospatial
- CES In-depth review of satellite imagery / earth observation technology in official statistics
- HLG-MOS Specialized topic on Statistical Data Editing
- HLG-MOS project – Strategic Communication (Phase2)
- European Union's Horizon 2020 - MAKSWELL - Work Package 3

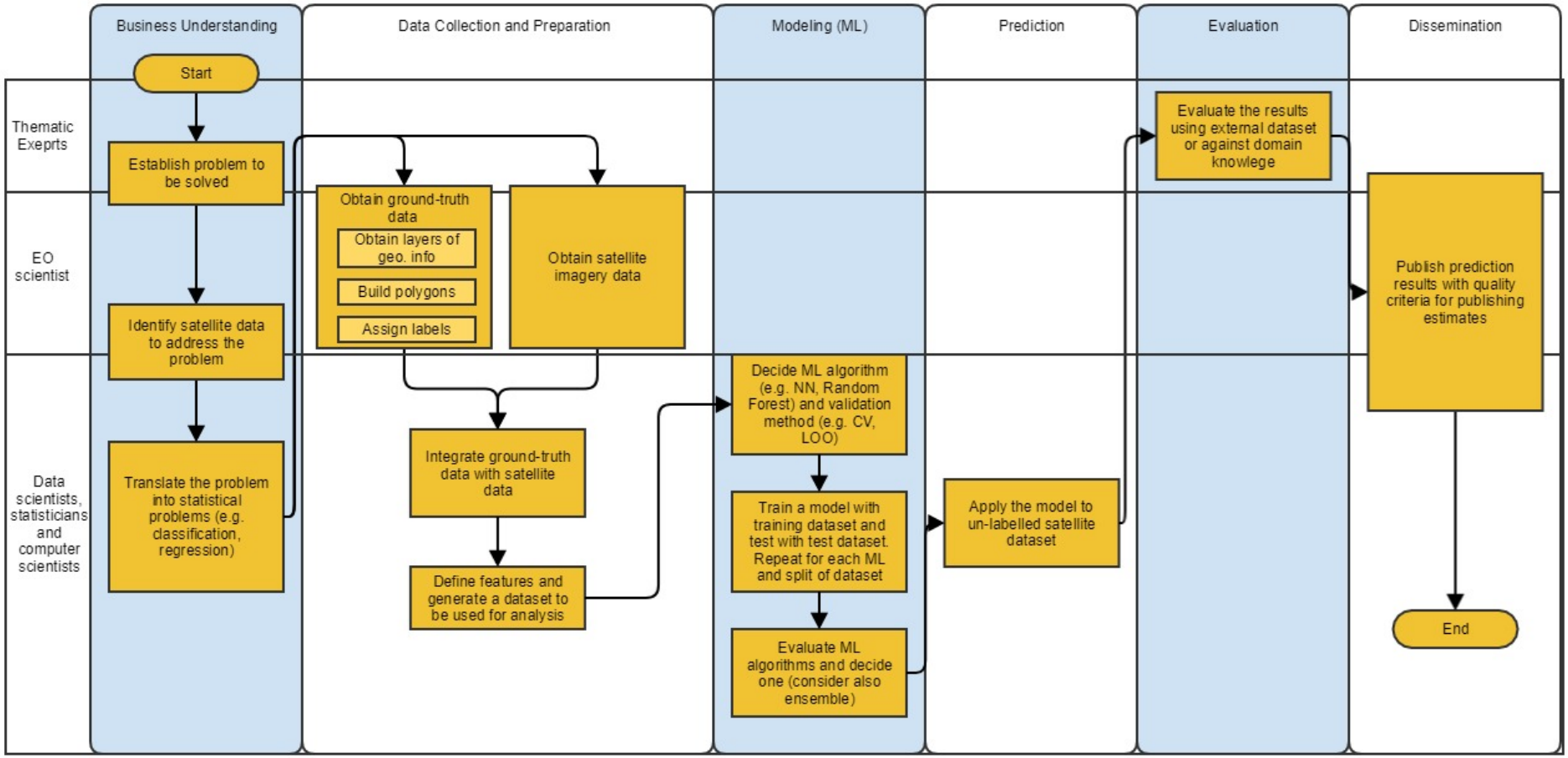
Back to Participants slide

C&C Process Pipeline



return to C&C Pilots

Imagery process pipeline



Back to Imagery slide

Develop, Expand, Interlink

Organisational challenges

- One of the recurring themes from the discussions at the sprints and monthly meetings is that integrating machine learning into official statistics requires more than simply building machine learning systems. In fact, a number of participants noted that they had already developed otherwise successful machine learning solutions, but had been unable to implement them into production processes because of a variety of organizational and structural impediments including uncertainty over who should be responsible for building, evaluating, and maintaining these highly interdisciplinary systems. Another comment observation is that in pursuing their pilot studies, team members are finding out similar or complementary developments within their own organisations.
- Existing work on numerous proof of concepts (PoCs) has shown that ML can add a great deal of value. However, the transition from a PoC or demo application to a ML solution embedded in the production of official statistics is another matter. Too often have I heard that early promising results learned from PoCs have been shelved for a later day. I propose to keep the momentum going and learn from how NSIs overcome blockers.
- Here in … we are working on the concept of an institutionalized Data Science Group. I was wondering if you could ask all the members of the HLG-MOS for any sort of documentation as to how their institutions started with these sort of teams/areas, ... I am sure … would greatly benefit from understanding similar cases so we can design a proper strategy.
- From the perspective of my organization, this project was a very valuable excercise which has enabled us to come up with the solutions which might have not been developed so quickly without an external push. However, although the exercise have been completed and we can present the results, there is a big challenge of how to actually proceed with the implementation and the risk of shelving is high. So, at least from our perspective, focusing on the implementation aspects is very desirable. The implementation challenge is big indeed, as at this stage not only do we have to cope with methodological or scientific issues, but we have to transfer our developments to our colleagues who work with traditional methods, reach the top management, ensure appropriate technology, etc.

back to Integration