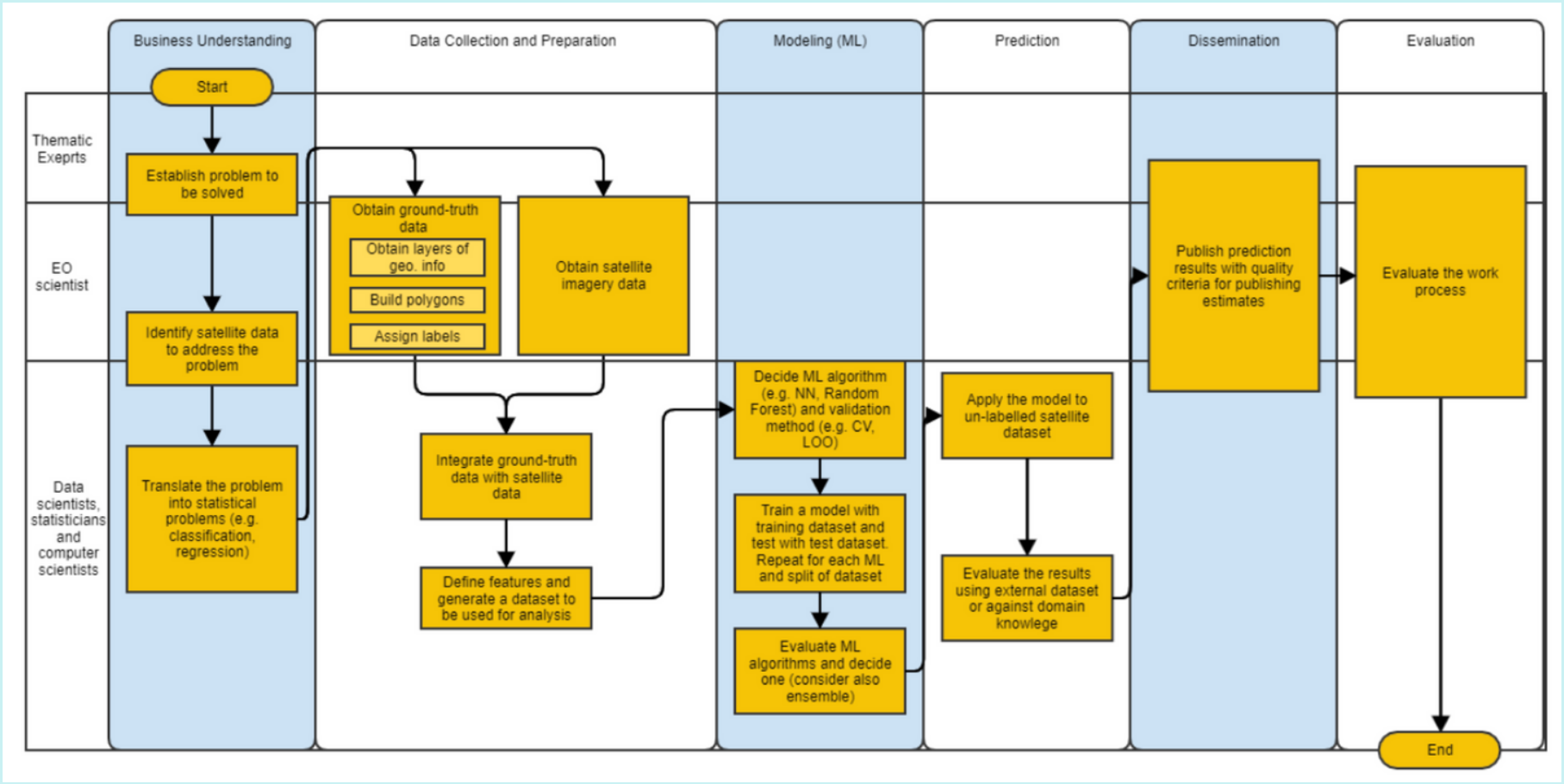
GENERIC PIPELINE FOR PRODUCTION OF OFFICIAL STATISTICS USING SATELLITE DATA AND MACHINE LEARNING

InKyung Choi (UNECE)

UNECE

modernstats BY HLG - MOS

WHAT IS IT?



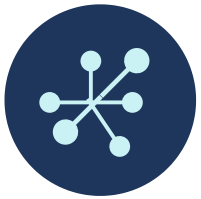
* In collaboration with INEGI

** Work in progress

CHALLENGES

Data is
big and different

Capability not in
place



Institutional
arrangement

Lack of
generalized
approach

CHALLENGES

Data is
big and different

Capability not in
place

Institutional
arrangement

Lack of
generalized
approach

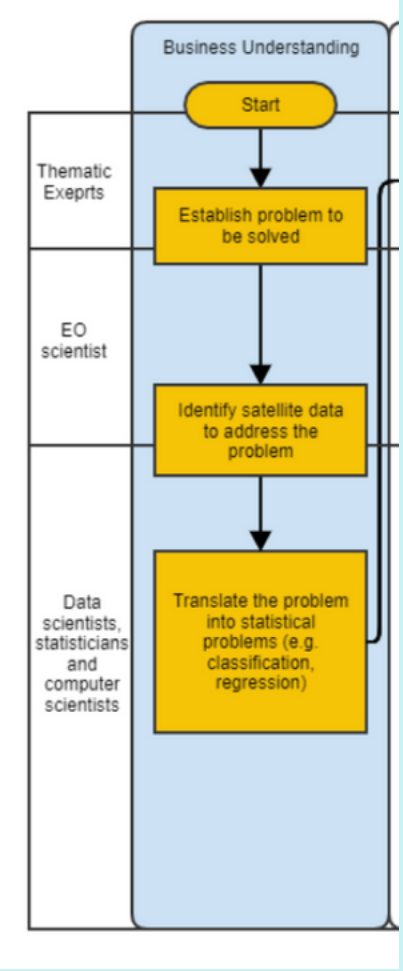
- Lack of understanding about process needed to use satellite data for statistical production
- Unclear scope and boundary of works
- No common reference points to link

GENERIC PROCESS MODEL

Generic process model describes high-level activities that need to be followed to achieve a certain objective

Provides **common language** that facilitate communication and sharing knowledge within and between organisations

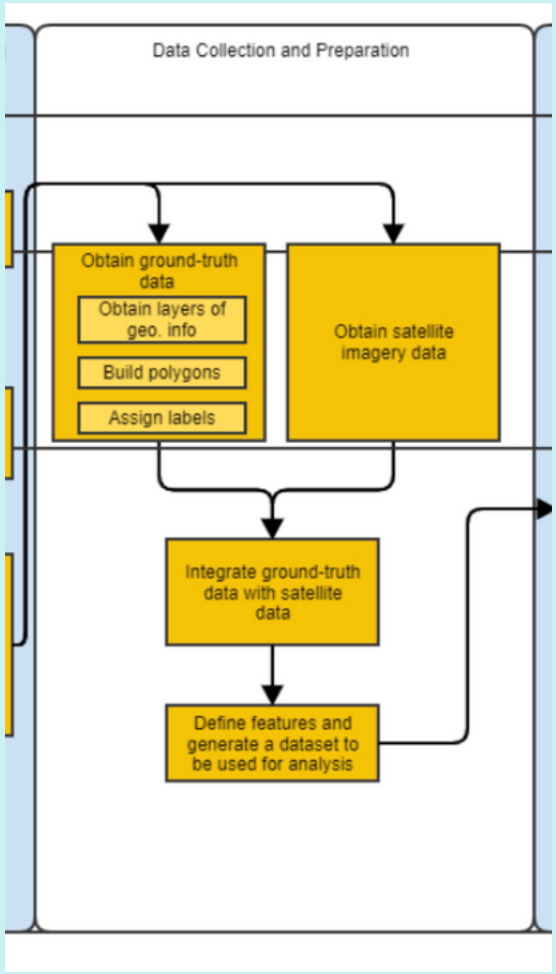
PIPELINE



Business understanding

- Establish problems to be solved
- Identify satellite and ground-truth data to address the problem
- Translate the problem into statistical problem

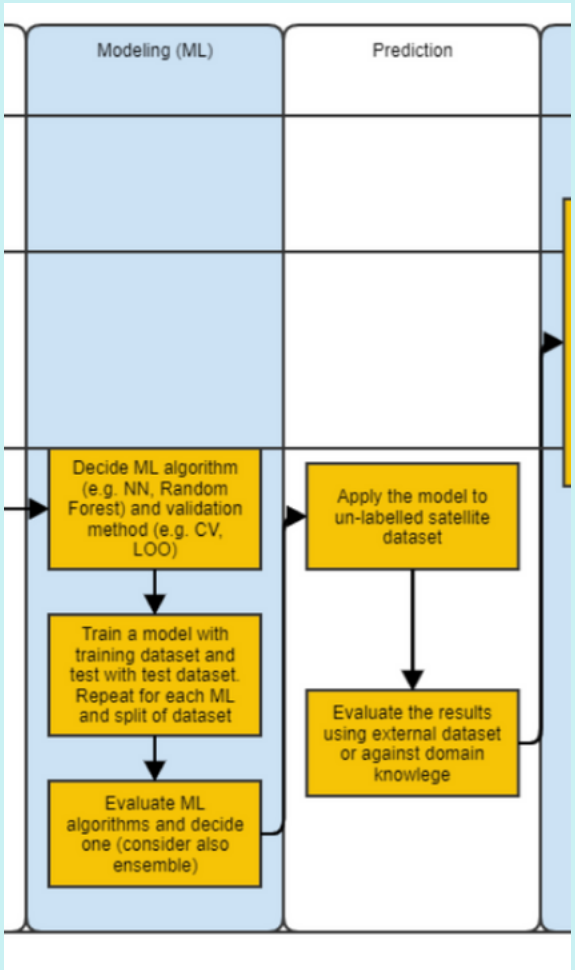
PIPELINE



Data collection and processing

- Obtain ground-truth data
- Obtain satellite data
- Integrate ground-truth data with satellite imagery data
- Define features and generate a dataset to be used for analysis

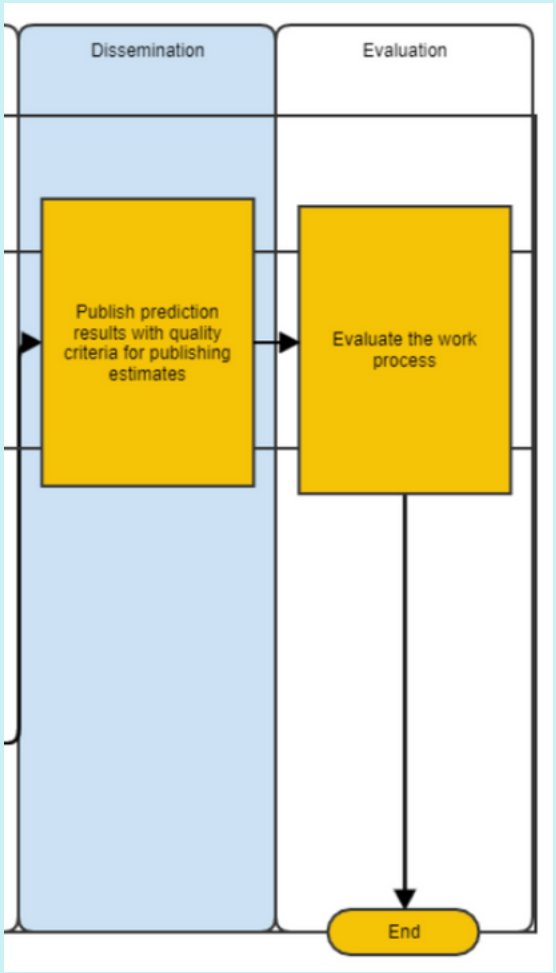
PIPELINE



Modeling and prediction

- Decide ML methods and validation method
- Train ML and test
- Evaluate
- Apply model to un-labelled data
- Evaluate using external source

PIPELINE



Dissemination and evaluation

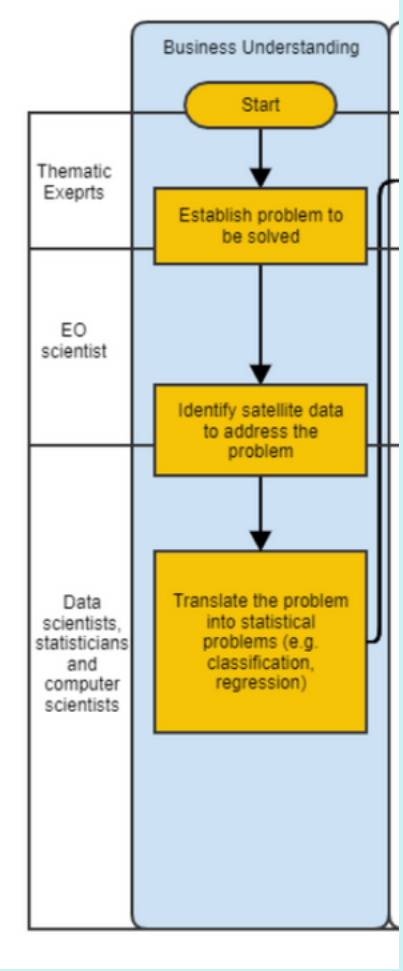
- Publish the output
- Evaluate the work process

EXAMPLE

Modelling crop yield (Statistics Canada)

- Until 2015, the survey was conducted at six time points throughout the year: March, June, July, September, November and December
- From 2012-13, started collaborating with EC) on a model development
- In 2016, September estimates were replaced by this model-based estimates

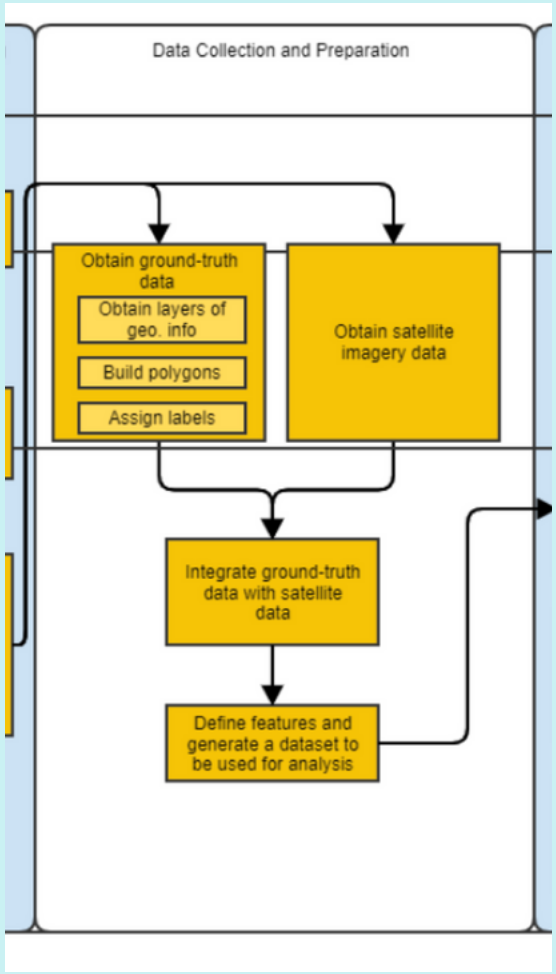
EXAMPLE



Business understanding

- Predict crop yield for province/country level
- Satellite data (NOAA), history crop yield data, agroclimatic data
- Regression problem with Census Agricultural Region (CAR) as unit of analysis

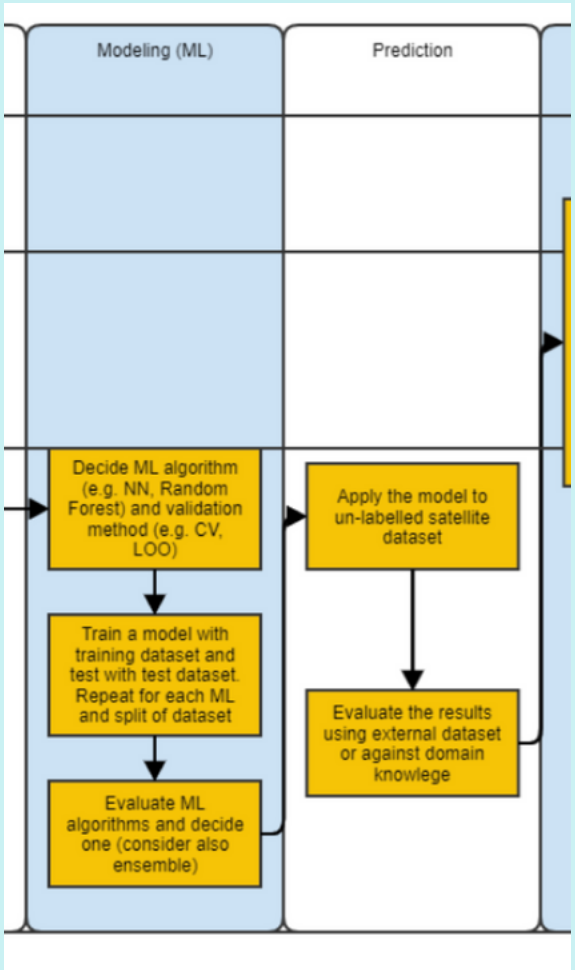
EXAMPLE



Data collection and processing

- Crop yield/weather data (EC)
- Normalized Difference Vegetation Index (NDVI)
- Integrate all at CAR level
- Each CAR has 28 years of data and 80 explanatory variables.

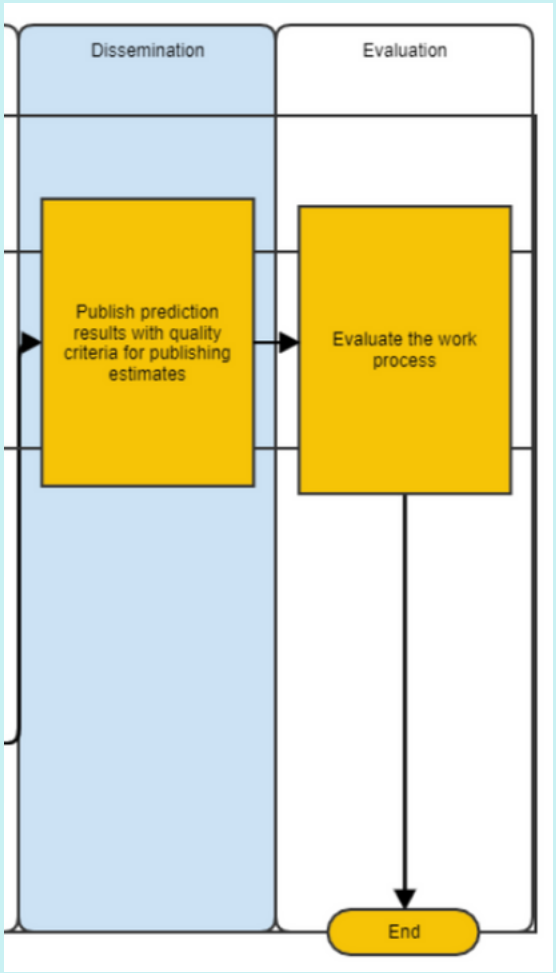
EXAMPLE



Modeling and prediction

- Linear regression/LASSO model
- Relative difference
- LASSO selected
- Subject-matter experts also review the results to identify any questionable estimates.

EXAMPLE



Dissemination and evaluation

- Acceptable level of quality to publish, e.g. minimum of 12 years of historical survey data

THANK YOU!

(MORE DETAILS IN PILOT STUDY REPORT)