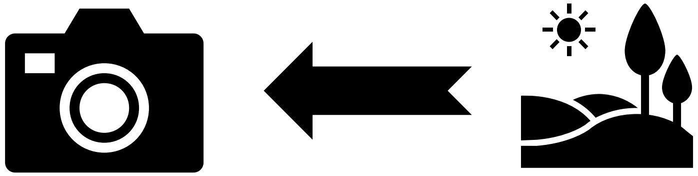
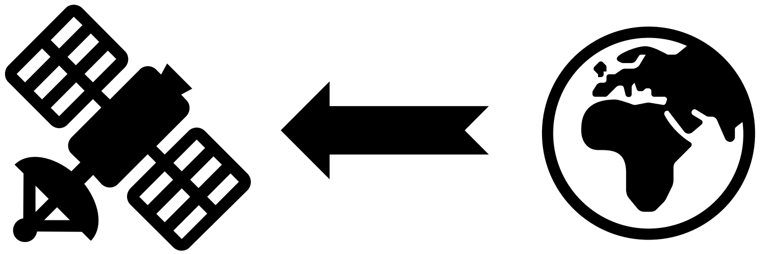
Generic Pipeline for Production of Official Statistics Using Satellite Data and Machine Learning

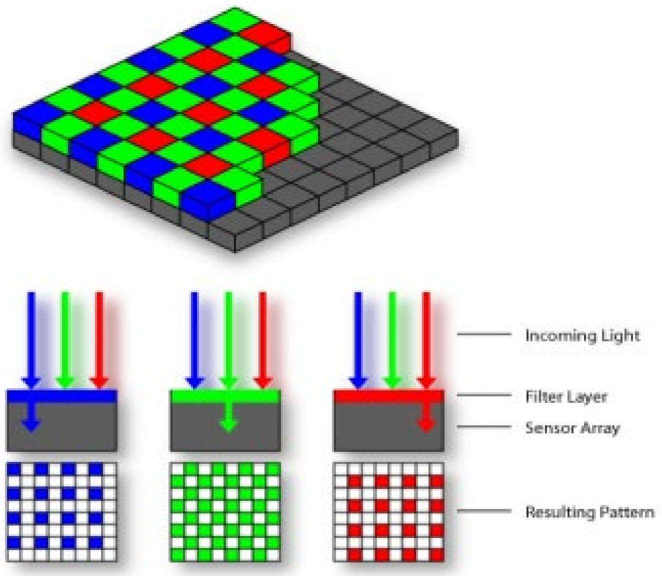
InKyung Choi (UNECE)

HLG-MOS ML Project Webinar (17 November 2020)

Satellite data







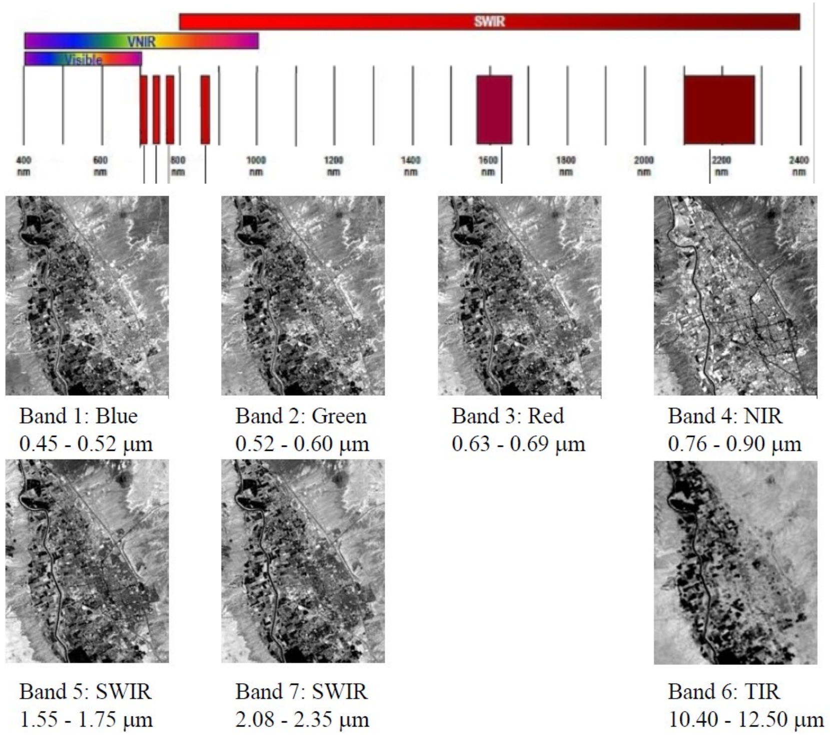

SWIR
VNIR
Visible
400 nm
600 nm
800 nm
1000 nm
1200 nm
1400 nm
1600 nm
1800 nm
2000 nm
2200 nm
2400 nm
Band 1: Blue 0.45 - 0.52 µm
Band 2: Green 0.52 - 0.60 µm
Band 3: Red 0.63 - 0.69 µm
Band 4: NIR 0.76 - 0.90 µm
Band 5: SWIR 1.55 - 1.75 µm
Band 7: SWIR 2.08 - 2.35 µm
Band 6: TIR 10.40 - 12.50 µm

Image source: https://web.nmsu.edu/~aulery/docs/Lab_9_appendix_e.pdf

Satellite data

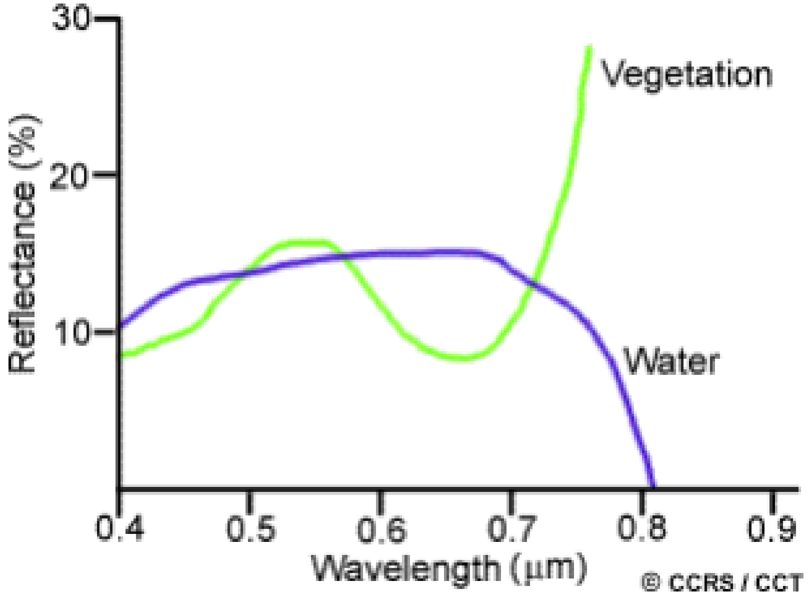
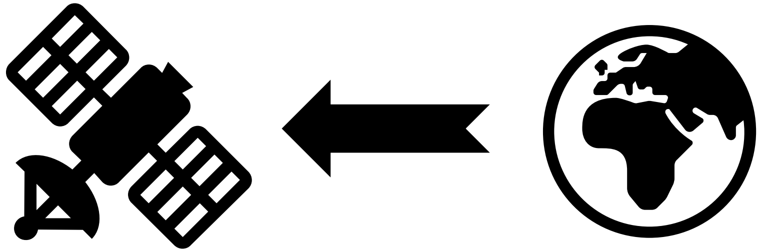
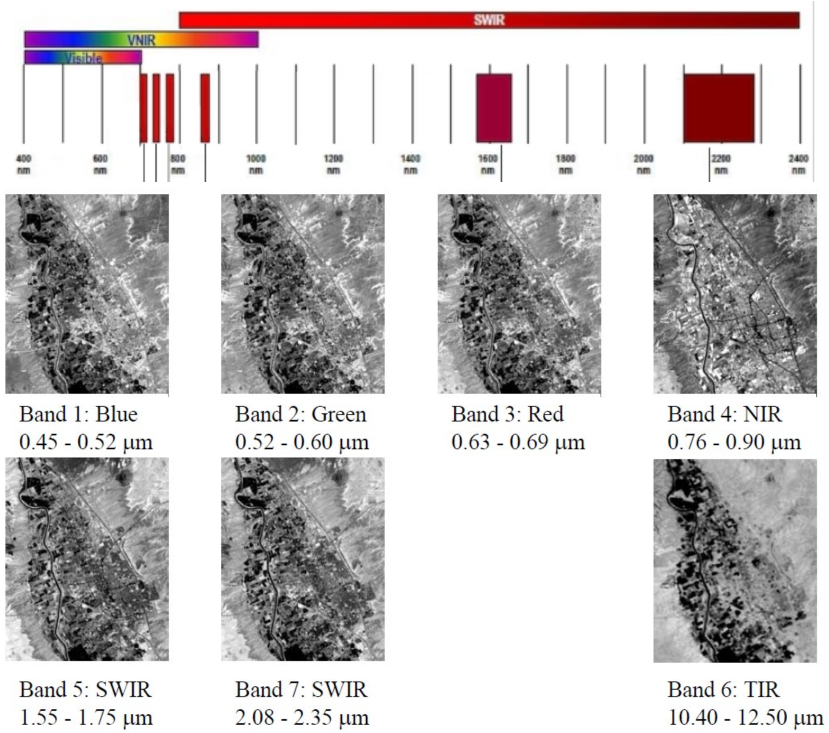


Image source: Natural Resources Canada - Fundamentals of Remote Sensing





Band 1: Blue
0.45 - 0.52 µm

Band 2: Green
0.52 - 0.60 µm

Band 3: Red
0.63 - 0.69 µm

Band 4: NIR
0.76 - 0.90 µm

Band 5: SWIR
1.55 - 1.75 µm

Band 7: SWIR
2.08 - 2.35 µm

Band 6: TIR
10.40 - 12.50 µm

Image source: https://web.nmsu.edu/~aulery/docs/Lab_9_appendix_e.pdf

Satellite data - benefits

LandSat 8 Program

Temporal resolution: Every 16 days

Spatial resolution: 30m

Sentinel Program

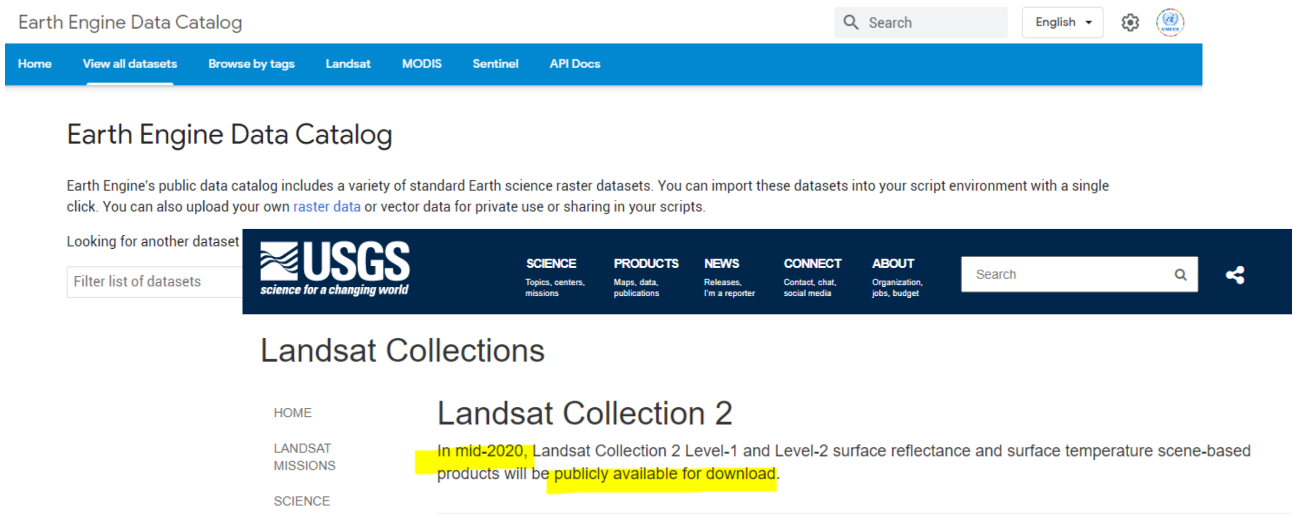
Temporal resolution: Every 5 days

Spatial resolution: 10-60m

- Global coverage with
- High (temporal) resolution
- Barrier is getting lower

Satellite data - benefits

The entire USGS Landsat data archive, dating from 1972 to present, is open NOW and is FREE to all users. http://earthexplorer.usgs.gov



Earth Engine Data Catalog

Earth Engine's public data catalog includes a variety of standard Earth science raster datasets. You can import these datasets into your script environment with a single click. You can also upload your own raster data or vector data for private use or sharing in your scripts.

Looking for another dataset

Filter list of datasets

Landsat Collections

Landsat Collection 2

In mid-2020, Landsat Collection 2 Level-1 and Level-2 surface reflectance and surface temperature scene-based products will be publicly available for download.

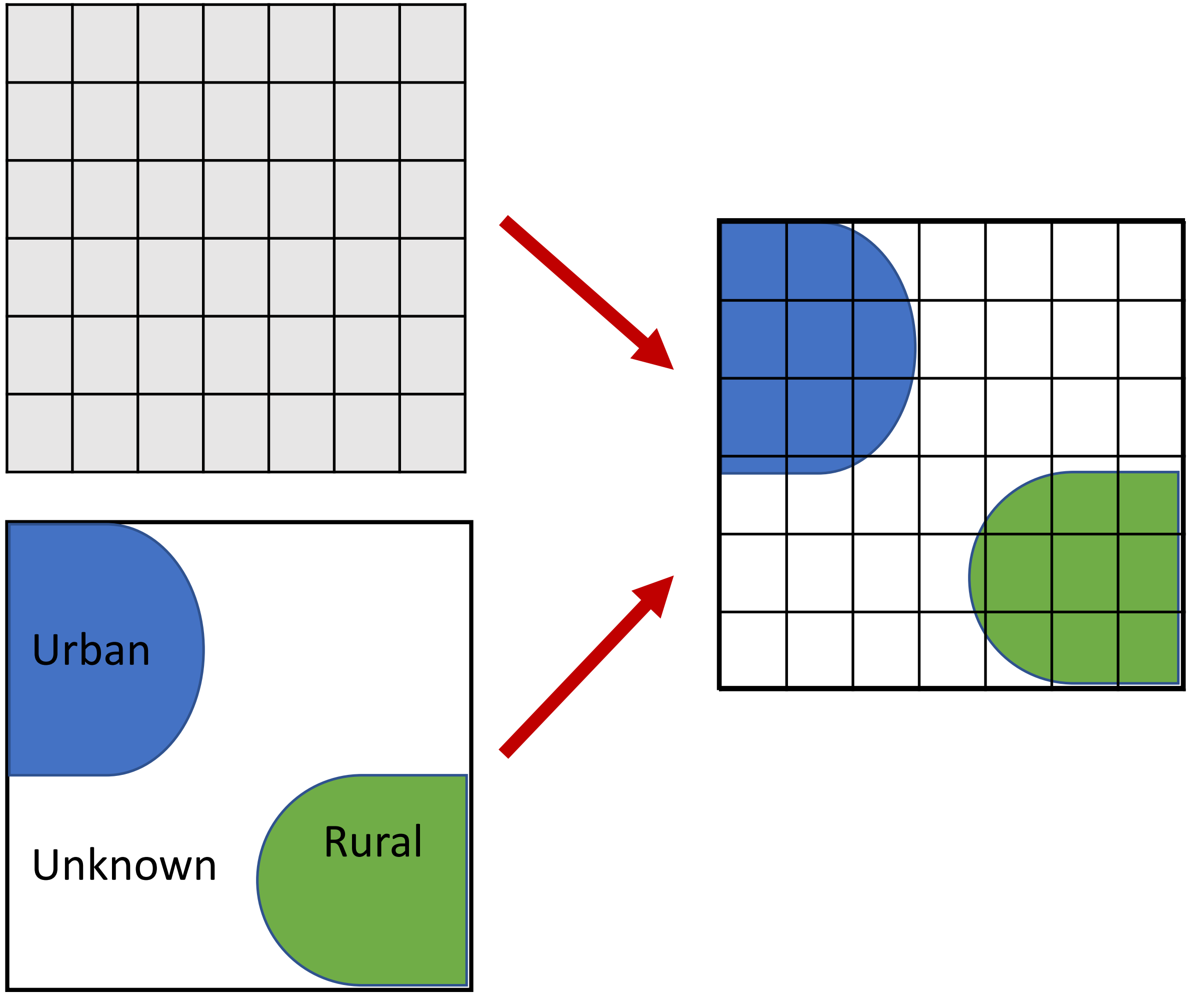
- Global coverage with
- High (temporal) resolution
- Barrier is getting lower

Need of National Statistics Offices (NSOs) to produce statistics at more (spatial) disaggregated at higher (temporal) frequency with lower cost

Use case - prediction

Satellite data
in pixels

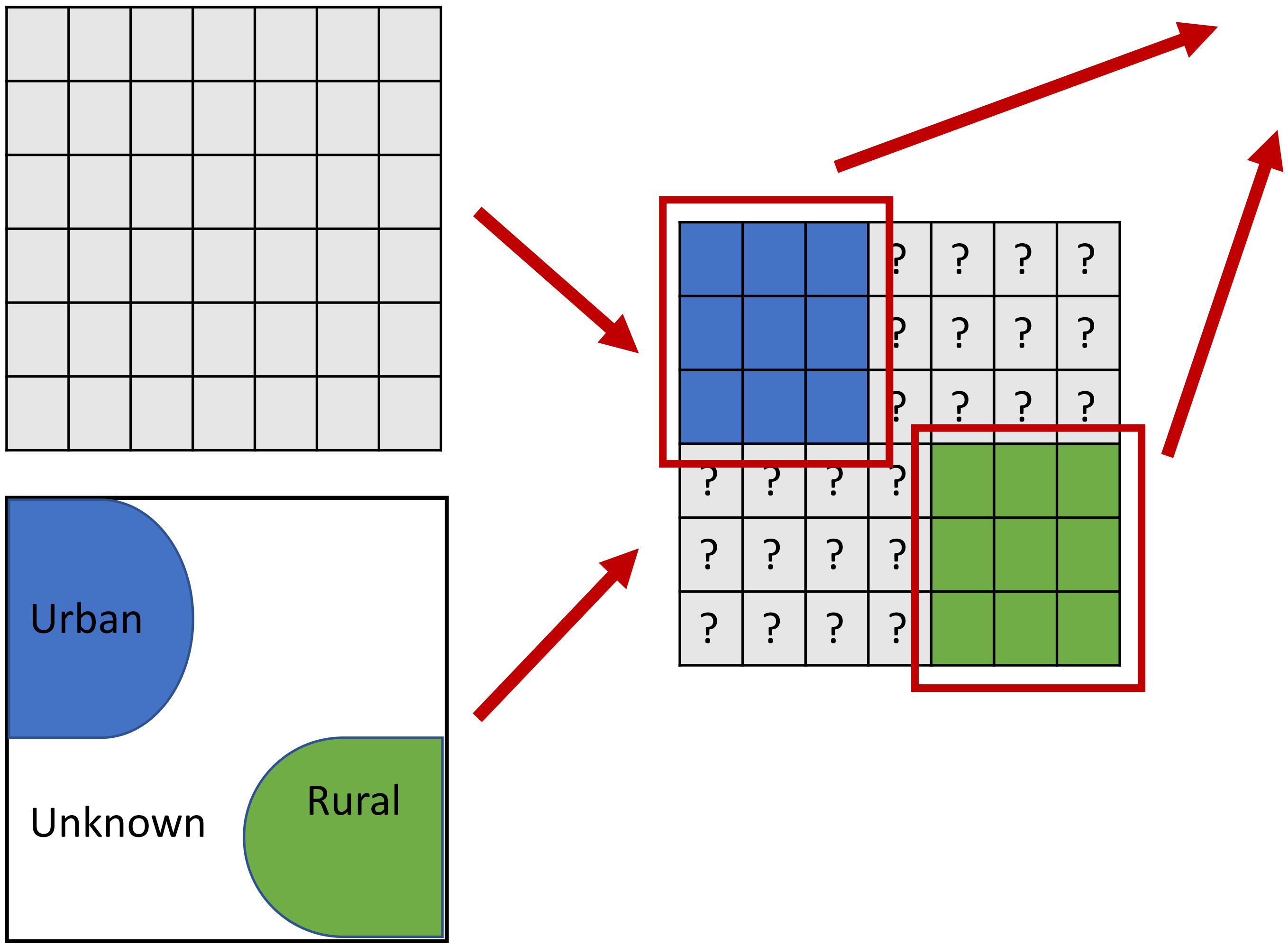
Ground-
truth data in
some shapes



Use case - prediction

Satellite data in pixels

Ground-truth data in *some* shapes

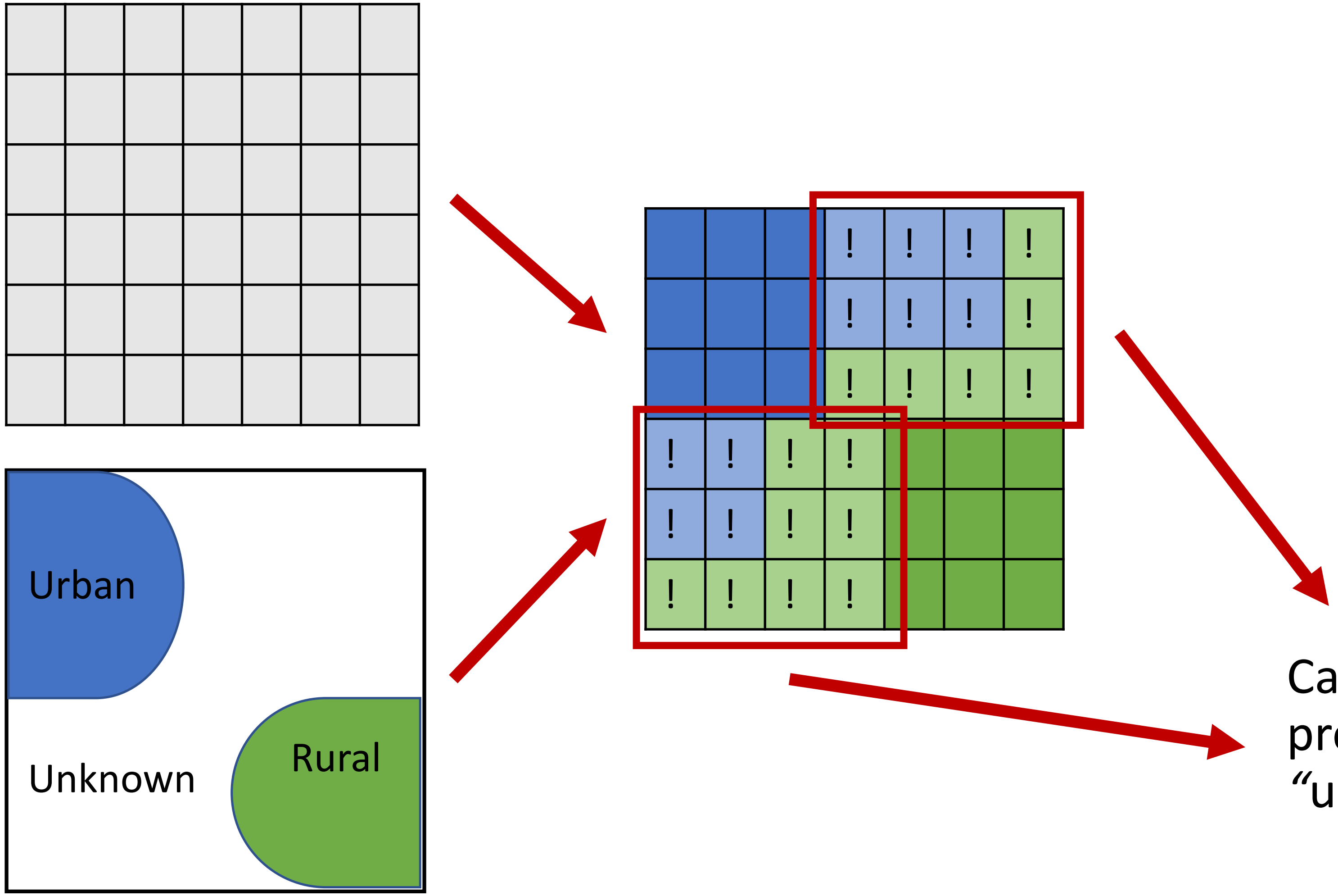


ML to learn relationship between category and satellite signal based on “labelled” pixels

Use case - prediction

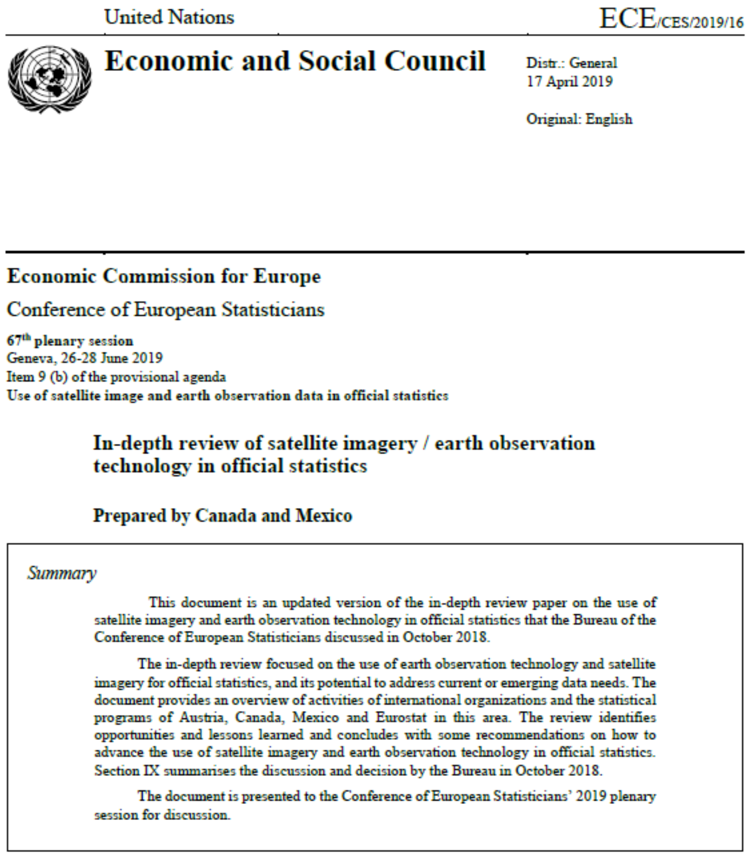
Satellite data in pixels

Ground-truth data in *some* shapes



Can make prediction on the “unlabelled” pixels

Satellite data - issues



United Nations ECE/CES/2019/16

Economic and Social Council

Distr.: General
17 April 2019

Original: English

Economic Commission for Europe

Conference of European Statisticians

67th plenary session
Geneva, 26-28 June 2019
Item 9 (b) of the provisional agenda
Use of satellite image and earth observation data in official statistics

In-depth review of satellite imagery / earth observation technology in official statistics

Prepared by Canada and Mexico

Summary

This document is an updated version of the in-depth review paper on the use of satellite imagery and earth observation technology in official statistics that the Bureau of the Conference of European Statisticians discussed in October 2018.

The in-depth review focused on the use of earth observation technology and satellite imagery for official statistics, and its potential to address current or emerging data needs. The document provides an overview of activities of international organizations and the statistical programs of Austria, Canada, Mexico and Eurostat in this area. The review identifies opportunities and lessons learned and concludes with some recommendations on how to advance the use of satellite imagery and earth observation technology in official statistics. Section IX summarises the discussion and decision by the Bureau in October 2018.

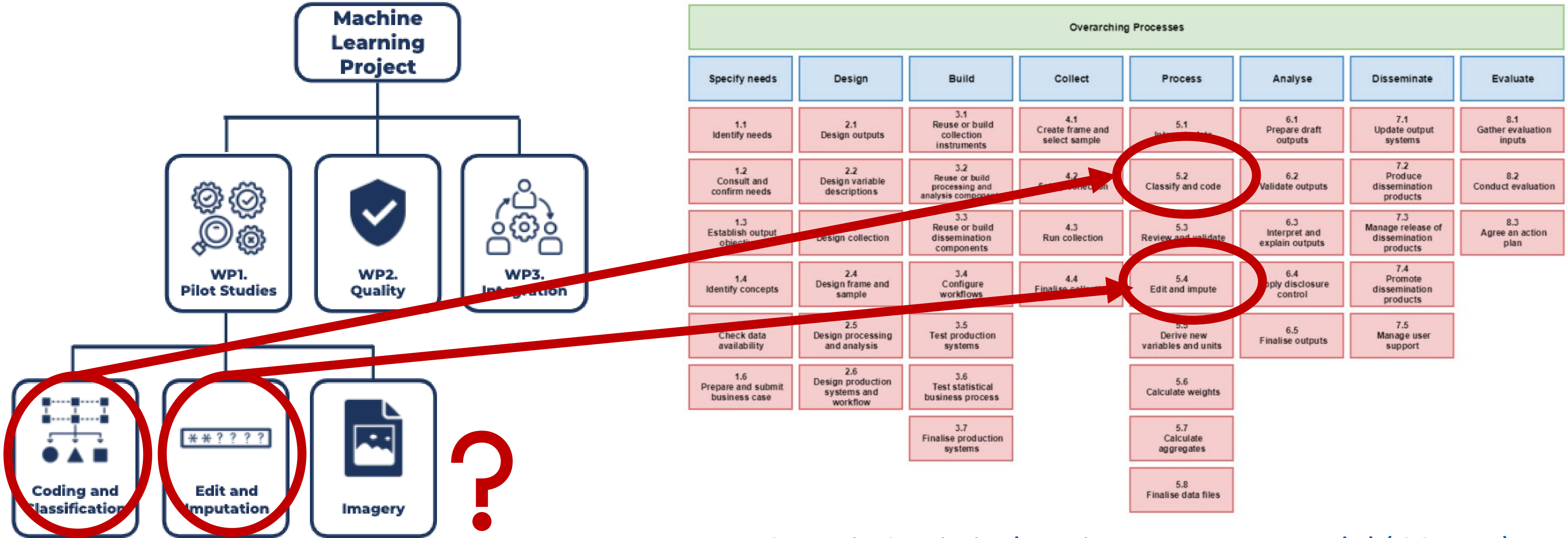
The document is presented to the Conference of European Statisticians' 2019 plenary session for discussion.

CES (2019) In-depth review of satellite imagery / earth observation technology in official statistics

Conference of European Statisticians(CES) 67th Plenary Session (2019) identified issues such as

- Lack expertise handling satellite data
- Satellite data are big
- Satellite data alone cannot produce statistics
- Institutional commitment to satellite data integration

Generic Pipeline – why?

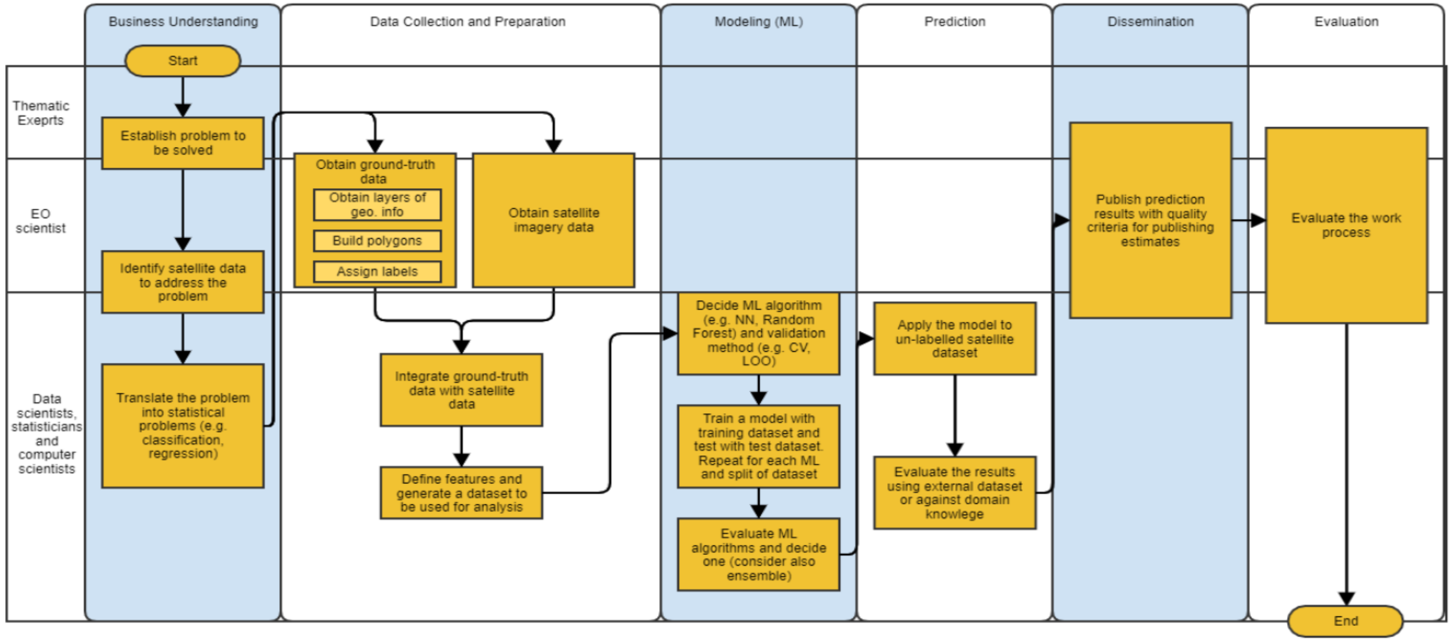


Overarching Processes							
Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation 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 Classify and code	6.2 Validate 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 Review and validate	6.3 Interpret and explain 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 Edit and impute	6.4 Apply 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			

Generic Statistical Business Process Model (GSBPM)

Generic Pipeline - aims

- To improve understanding about workflow needed to use satellite data and ML for statistical production
- To clarify scope and boundary of works
- To serve as common reference points to link increasing body of works

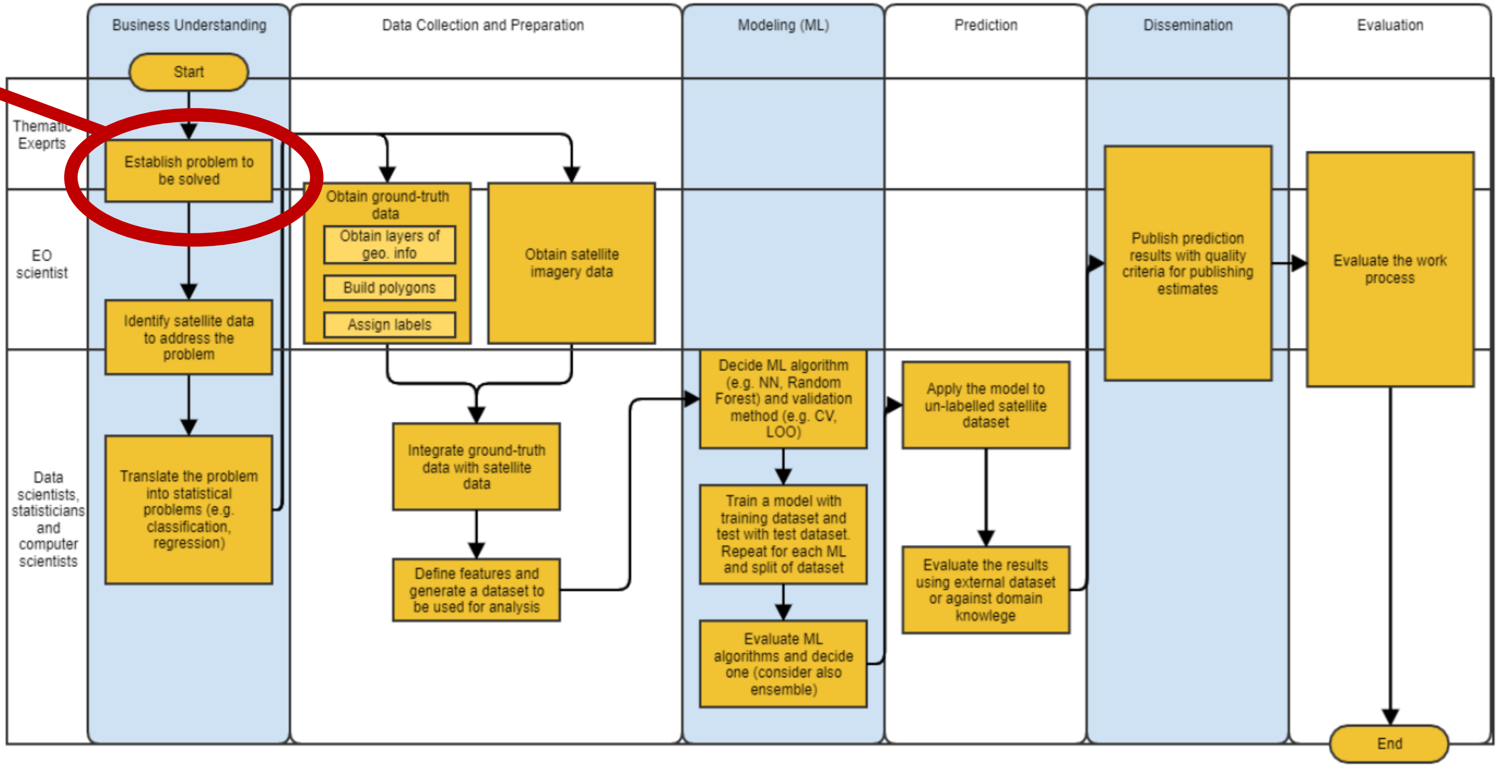

Business Understanding
Data Collection and Preparation
Modeling (ML)
Prediction
Dissemination
Evaluation
Thematic Exeprts
EO scientist
Data scientists, statisticians and computer scientists
Start
Establish problem to be solved
Identify satellite data to address the problem
Translate the problem into statistical problems (e.g. classification, regression)
Obtain ground-truth data
Obtain layers of geo. info
Build polygons
Assign labels
Obtain satellite imagery data
Integrate ground-truth data with satellite data
Define features and generate a dataset to be used for analysis
Decide ML algorithm (e.g. NN, Random Forest) and validation method (e.g. CV, LOO)
Train a model with training dataset and test with test dataset. Repeat for each ML and split of dataset
Evaluate ML algorithms and decide one (consider also ensemble)
Apply the model to un-labelled satellite dataset
Evaluate the results using external dataset or against domain knowlege
Publish prediction results with quality criteria for publishing estimates
Evaluate the work process
End

Generic Pipeline - overview

- What activities should be carried out
- Who should play the leading role?
- What resources are available?

6 Stages

3 Roles



Business Understanding

Data Collection and Preparation

Modeling (ML)

Prediction

Dissemination

Evaluation

Thematic Exeprts

EO scientist

Data scientists, statisticians and computer scientists

Start

Establish problem to be solved

Identify satellite data to address the problem

Translate the problem into statistical problems (e.g. classification, regression)

Obtain ground-truth data

Obtain layers of geo. info

Build polygons

Assign labels

Obtain satellite imagery data

Integrate ground-truth data with satellite data

Define features and generate a dataset to be used for analysis

Decide ML algorithm (e.g. NN, Random Forest) and validation method (e.g. CV, LOO)

Train a model with training dataset and test with test dataset. Repeat for each ML and split of dataset

Evaluate ML algorithms and decide one (consider also ensemble)

Apply the model to un-labelled satellite dataset

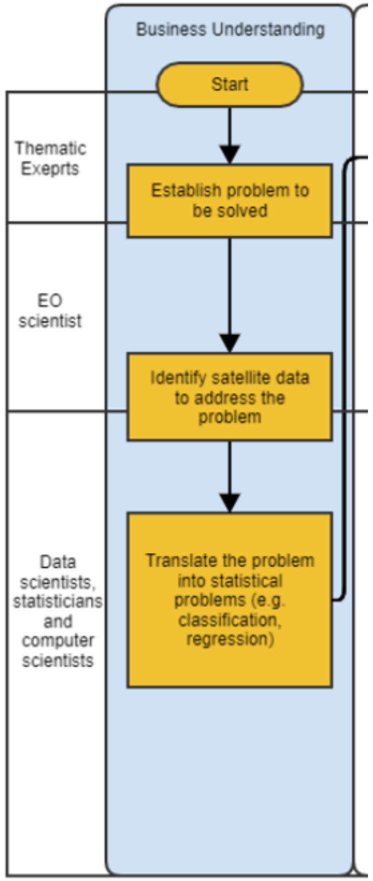
Evaluate the results using external dataset or against domain knowlege

Publish prediction results with quality criteria for publishing estimates

Evaluate the work process

End

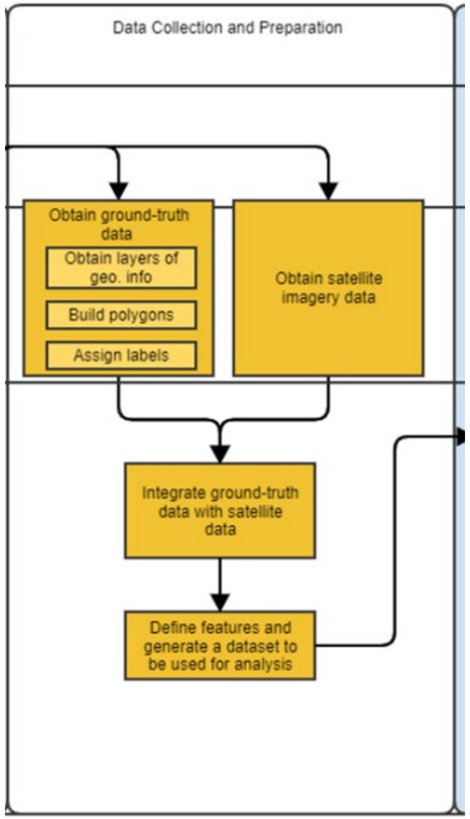
Generic Pipeline – Business understanding


Business Understanding
Start
Thematic Exeprts
Establish problem to be solved
EO scientist
Identify satellite data to address the problem
Data scientists, statisticians and computer scientists
Translate the problem into statistical problems (e.g. classification, regression)

- Establish the problem to be resolved
- Identify satellite data and ground-truth data to address the problem. Factors to consider:
  - Temporal resolution
  - Spectral resolution
  - Spatial resolution
  - Sustainability
  - Easy of use

  ** UN Global Working Group on Big Data (2017) Satellite Imagery and Geospatial Data Task Team Report Chapter 2. Data Sources*
- Translate the problem into statistical problem, typically
  - Regression
  - Classification

Generic Pipeline – Data collection and preparation

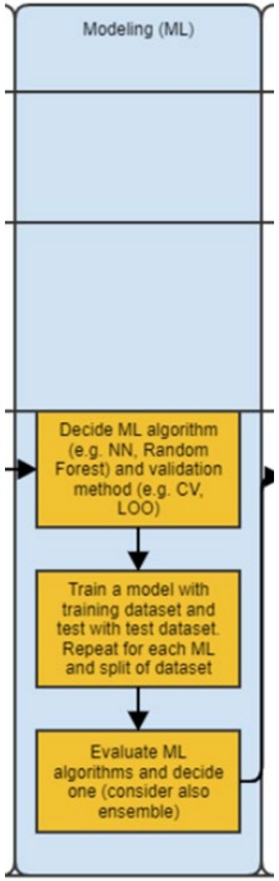

Data Collection and Preparation
Obtain ground-truth data
Obtain layers of geo. info
Build polygons
Assign labels
Obtain satellite imagery data
Integrate ground-truth data with satellite data
Define features and generate a dataset to be used for analysis

- Obtain ground-truth data with attention to geo-referencing
- Obtain satellite data
- Integrate ground-truth data with satellite imagery data

** HLG-MOS Data Integration Project (2016-17)*

- Define variables and generate a dataset to be used for analysis (e.g. Normalised Difference Vegetation Index (NDVI))

Generic Pipeline – Modeling (ML)

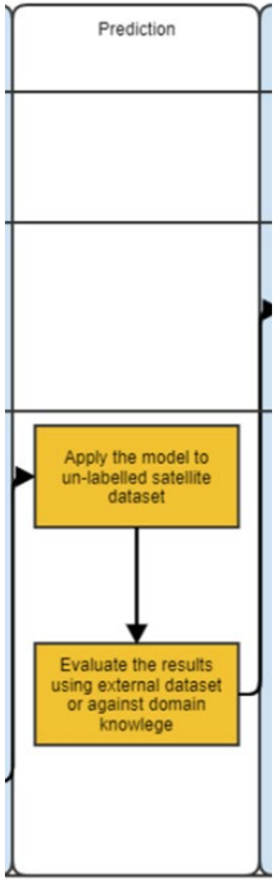

Modeling (ML)
Decide ML algorithm (e.g. NN, Random Forest) and validation method (e.g. CV, LOO)
Train a model with training dataset and test with test dataset. Repeat for each ML and split of dataset
Evaluate ML algorithms and decide one (consider also ensemble)

- Decide ML algorithms (e.g. SVM, Random Forest, NN), measure of performance (e.g. precision, F1) and validation method

** Methodological Approaches for Utilising Satellite Imagery to Estimate Official Crop Area Statistics (2014; ABS)*

- Train a model with training dataset and test with test dataset. Repeat for each ML and split of dataset
- Evaluate ML models and decide one (consider also ensemble)

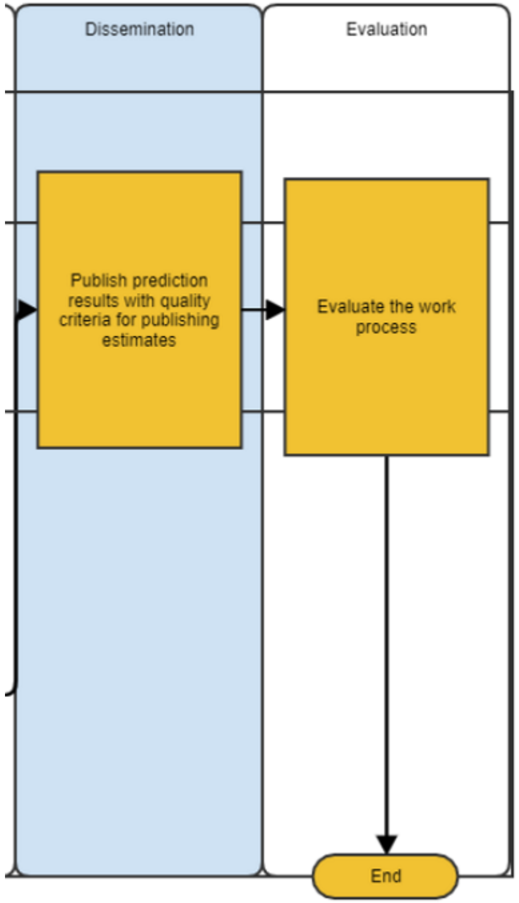
Generic Pipeline – Prediction



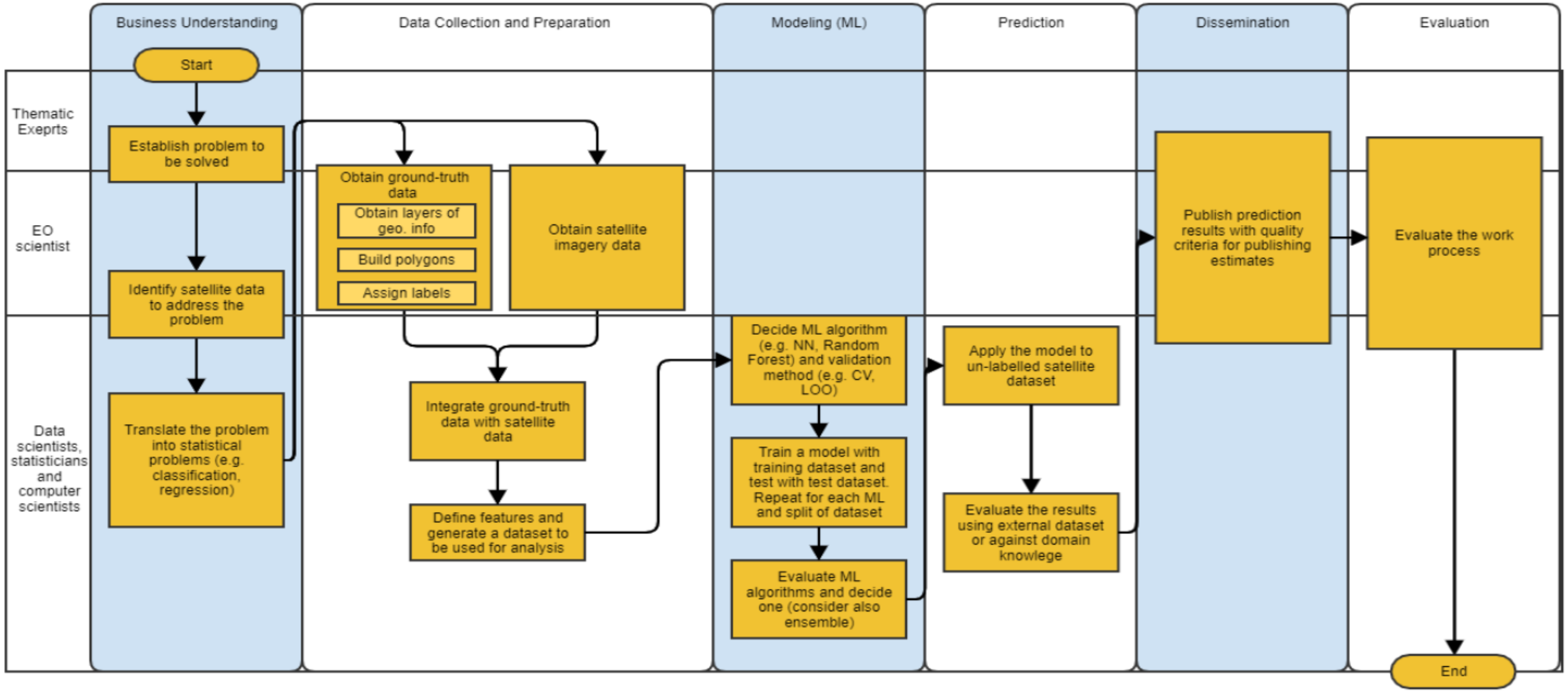
Prediction
Apply the model to un-labelled satellite dataset
Evaluate the results using external dataset or against domain knowlege

- Apply the model to un-labelled satellite dataset
- Evaluate the prediction results using external dataset or against domain knowledge

Generic Pipeline – Dissemination and evaluation


Dissemination
Evaluation
Publish prediction results with quality criteria for publishing estimates
Evaluate the work process
End

- Establish quality criteria for publishing estimates and publish data
- Evaluate the work process



Business Understanding

Data Collection and Preparation

Modeling (ML)

Prediction

Dissemination

Evaluation

Thematic Exeprts

EO scientist

Data scientists, statisticians and computer scientists

Start

Establish problem to be solved

Identify satellite data to address the problem

Translate the problem into statistical problems (e.g. classification, regression)

Obtain ground-truth data

Obtain layers of geo. info

Build polygons

Assign labels

Obtain satellite imagery data

Integrate ground-truth data with satellite data

Define features and generate a dataset to be used for analysis

Decide ML algorithm (e.g. NN, Random Forest) and validation method (e.g. CV, LOO)

Train a model with training dataset and test with test dataset. Repeat for each ML and split of dataset

Evaluate ML algorithms and decide one (consider also ensemble)

Apply the model to un-labelled satellite dataset

Evaluate the results using external dataset or against domain knowlege

Publish prediction results with quality criteria for publishing estimates

Evaluate the work process

End

Thank you!