



# **Integrating EO with Official Statistics using Machine Learning in Mexico**

**(Work in progress)**

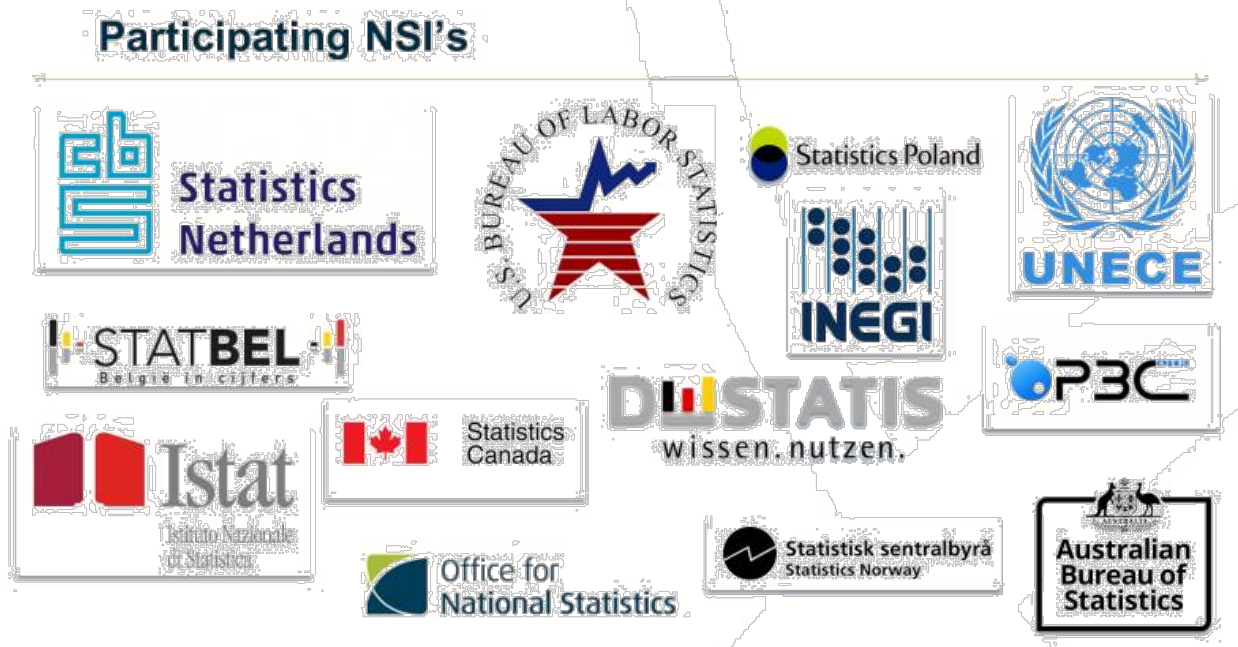
**April 9, 2020**

**INEGI**

**[abel.coronado@inegi.org.mx](mailto:abel.coronado@inegi.org.mx)**

## UNECE Machine Learning Project

This is one of the pilot projects in the **Machine Learning Project** of the **UNECE High-Level Group on Modernization of Official Statistics**.



## UNECE Machine Learning Project

### Objectives

- Investigate and demonstrate the value added of ML in the production of official statistics, where "value added" is increase in relevance, better overall quality or reduction in costs.
- 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.



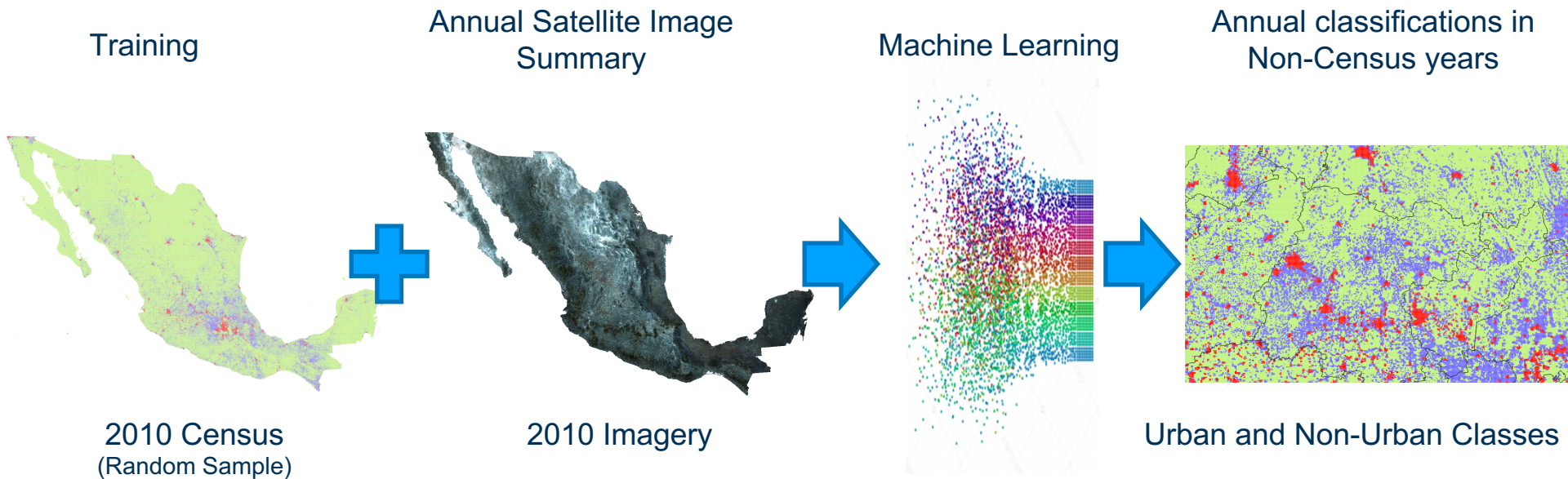


# | Imagery Pilot Project



## Objective of this Imagery Pilot Project (Practical Application)

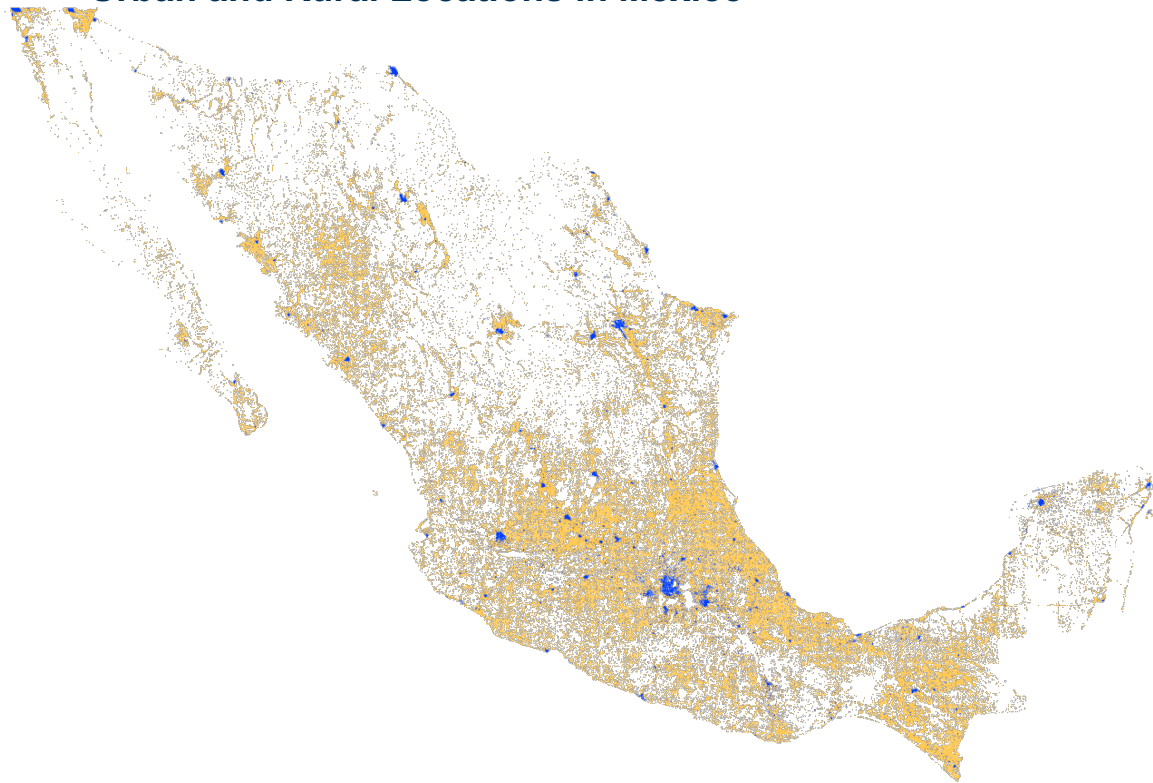
Expand the use of imagery data in the production of official statistics through the further development of knowledge and sharing of ML solutions and practices.



## Imagery Pilot Project

- This pilot project proposes the use of Landsat satellite data for the mapping of urban areas in non-census years, using as ground truth an urban density grid of 1km x 1km generated from the field data of the 2010 Population Census at block level and the updated Georeferenced Business Register to date 2010.
- This pilot project seeks to take advantage of the knowledge, data, and technologies available to show an example of the use of satellite images in the generation of geographic information that can be used to monitor the change in urban density and given the final product meets the appropriate quality, explore the path to incorporating it into the continuous urban monitoring processes of INEGI.
- Quarterly monitoring the growth of the country's urban areas could allow adjusting models and samples from household and business surveys. As the application matures and improves, it could even produce new sets of official statistics on its own.

## Urban and Rural Locations in Mexico



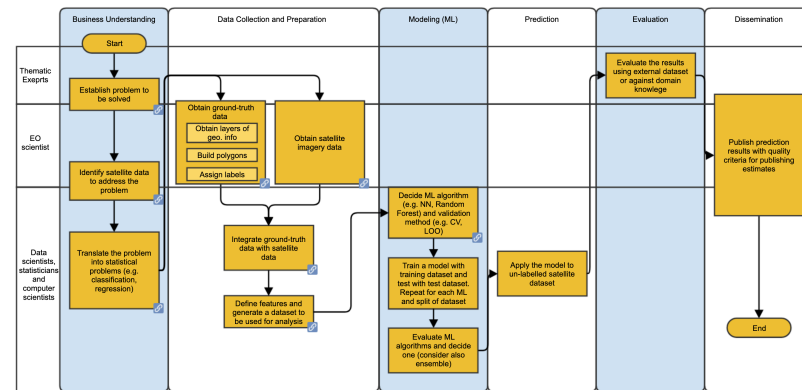
# | **Draft of Imagery Pipeline**

# Imagery Pipeline

In order to have a road map.  
We build an abstract machine learning pipeline for Imagery

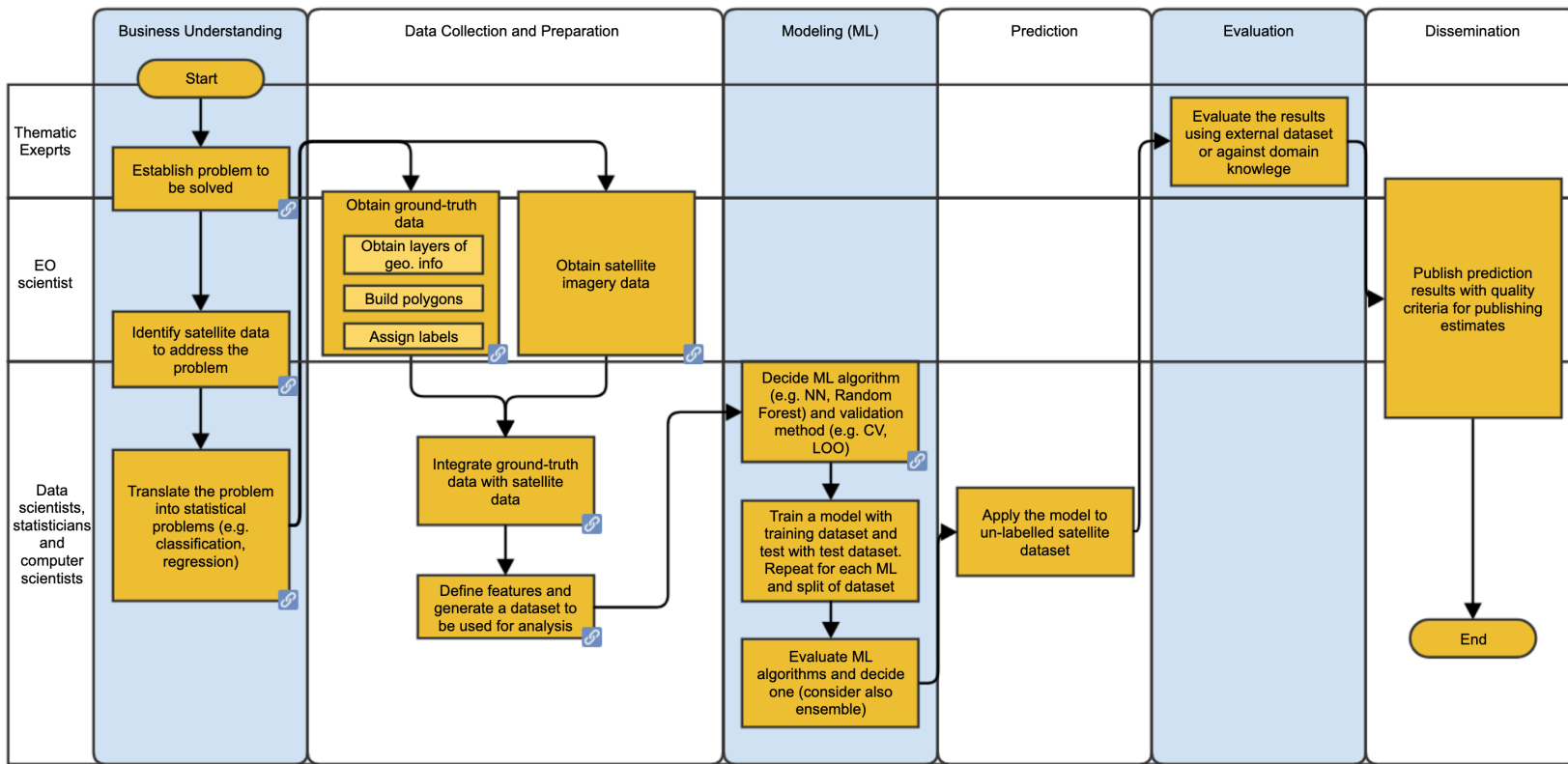
Based on:

IBM CRISP-DM  
Cross Industry Standard Process for Data Mining  
&  
Microsoft Data Science lifecycle



InKyung Choi; Refactoring

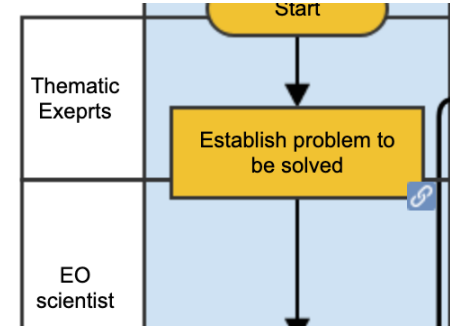
# Imagery Pipeline





# | Pilot Project

## Establish the problem to be solved

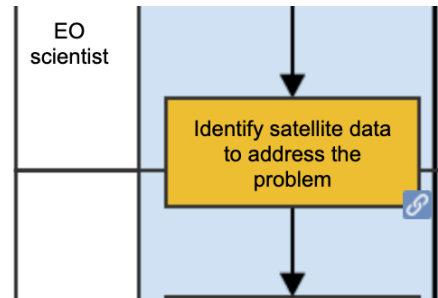


Monitor the growth of urban locations of Mexico, which would generate a more timely input for :

- Cartography update
- Estimation of the population in non-census years
- Related with:
  - **SDG Indicator 11.3.1:** Ratio of land consumption rate to population growth rate
  - **SDG Indicator 15.3.1:** Proportion of land that is degraded over total land area



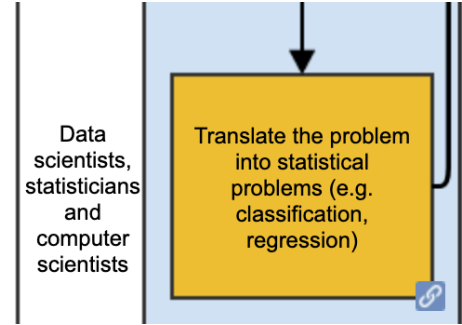
## Identify satellite data to address the problem



In our case, the data to monitor the growth of cities can be Landsat images (NASA & USGS):

- They are open data.
- There is a constant record since the 70s, although they are available from 1985 to date.
- The spatial resolution of Landsat images is 30 meters.
- Temporal resolution is 16 days and 8 days with combined satellites.
- Spectral resolution, in this pilot project we use 6 spectral bands
- All the data we use is Analysis Ready Data (ARD)
- We take advantage that we have just built the Mexican Geospatial Data Cube, with all this information.

## Translate the problem into statistical problem



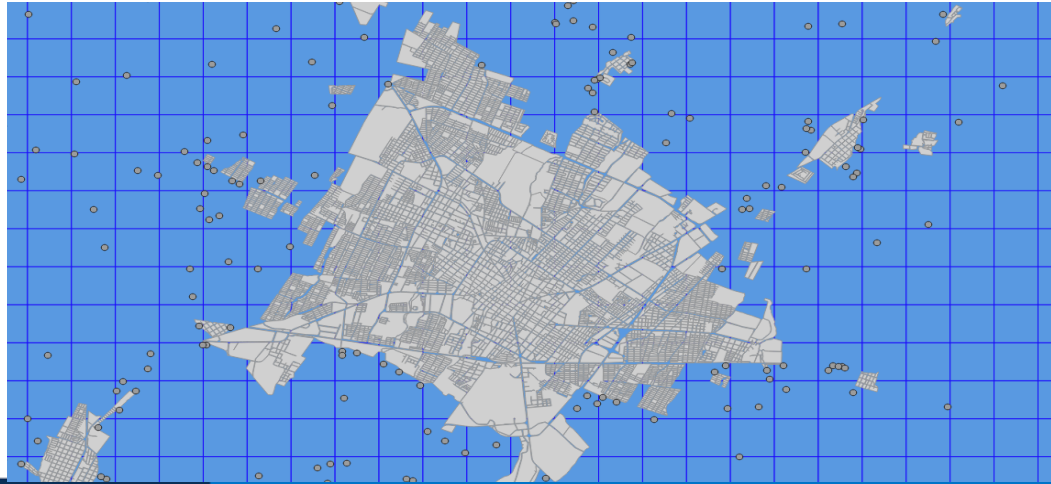
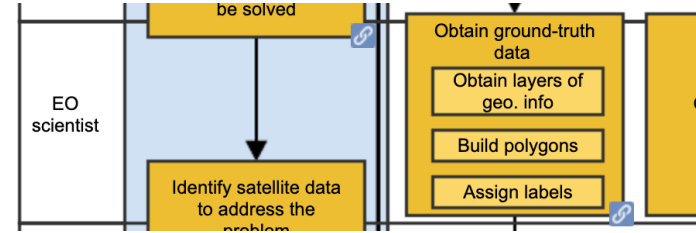
We define that is a classification problem:

- Unit of analysis: 1km x 1km squares covering the whole country: 1' 975, 719
- 2 classes were designated:
  - Urban
  - Non-Urban

## Obtain ground-truth data

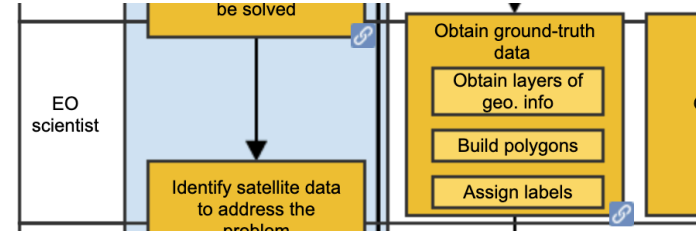
Obtain layers of geographical information:

- Georeferenced Population Census 2010 (Block Level Aggregation)
- Georeferenced Economic Census (Economic Unit Level)

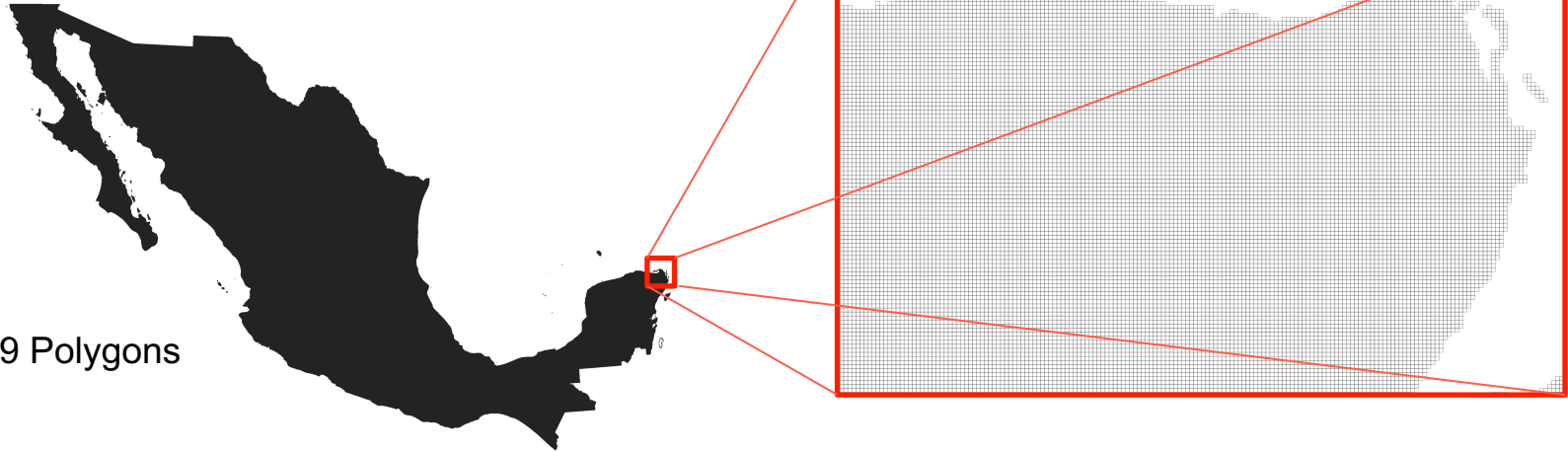


## Obtain ground-truth data

Build the polygons (1 km – 1 km)

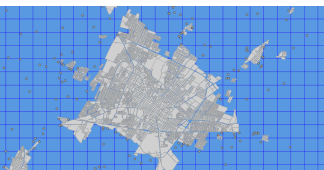
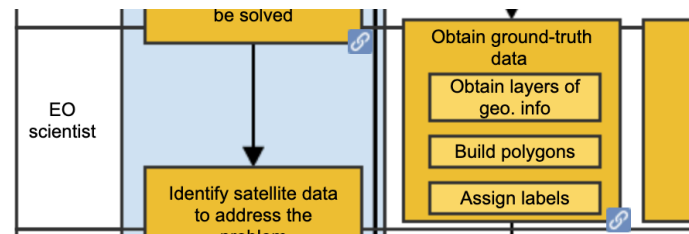


1' 975, 719 Polygons

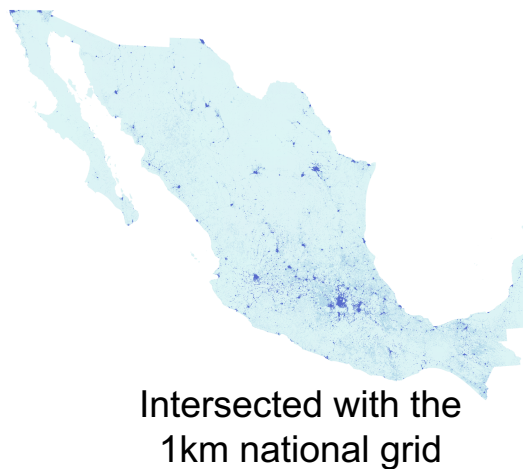


## Obtain ground-truth data

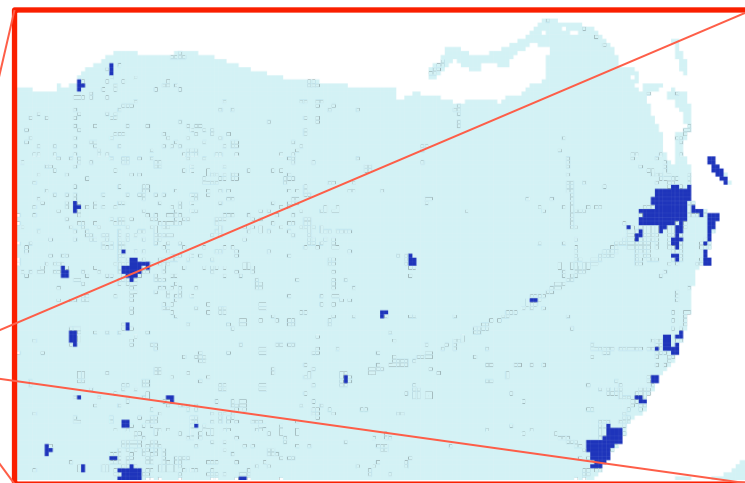
### Assign Labels



Georeferenced  
Census Results  
(Population and  
Economic)



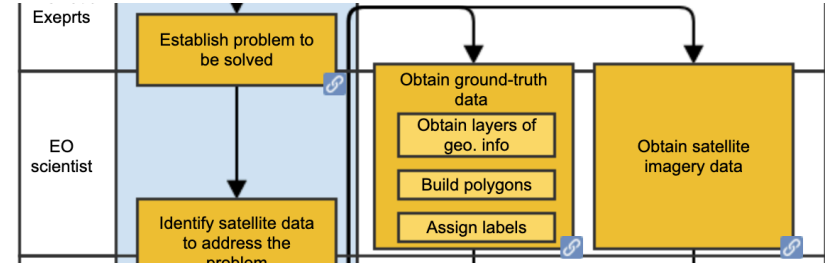
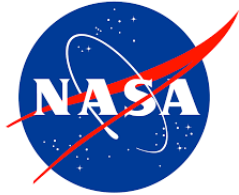
Intersected with the  
1km national grid



Urban	36,759	1km by 1km regions
Non-Urban	1,938,960	

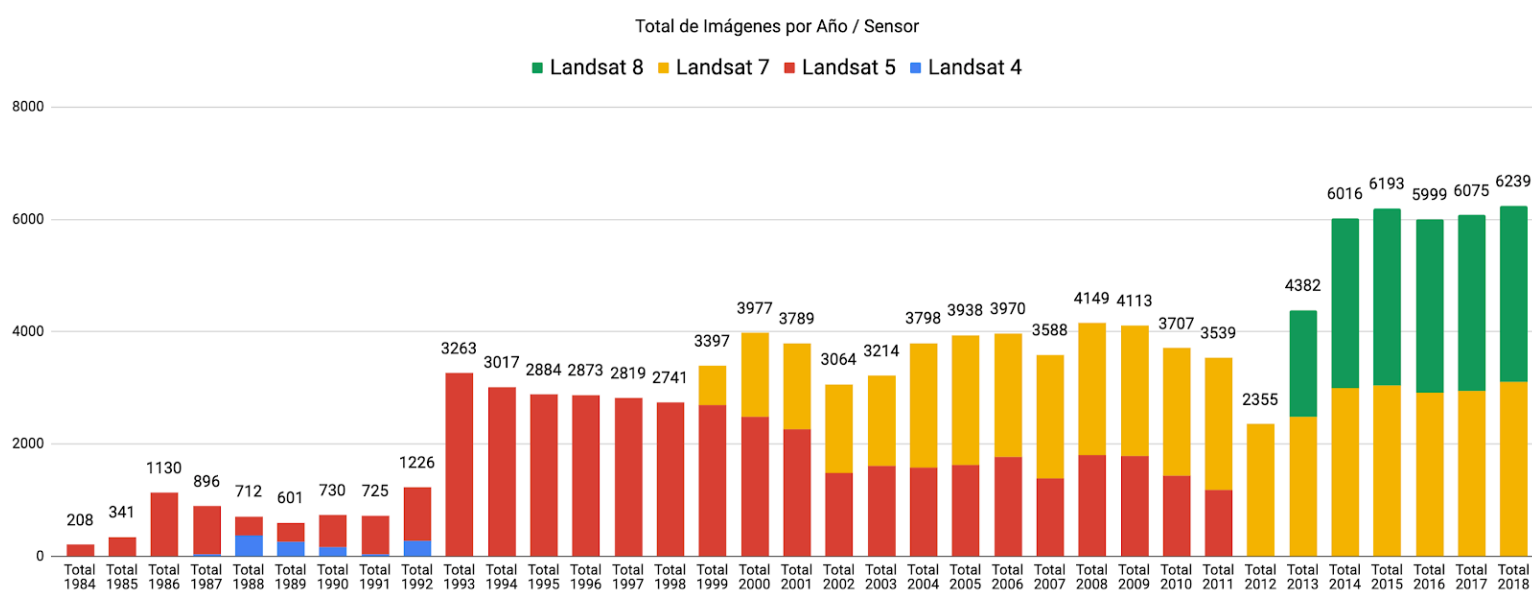
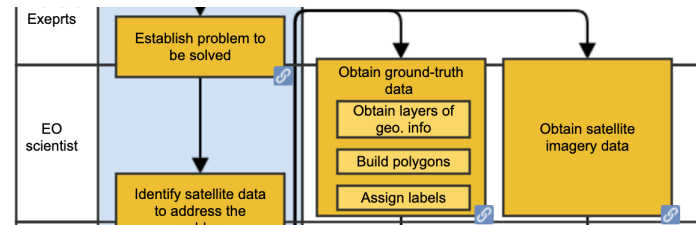
## Obtain satellite imagery data

- 32 TB of Images in external discs.
- 90 TB decompressed
- March 4, 2019
- The images are ARD, Analysis Ready Data.
- In essence ARD means that the pixels of the entire time series are aligned and comparable.



## Obtain satellite imagery data

- 36 Years



## Obtain satellite imagery data

- 2010 same year of the Census

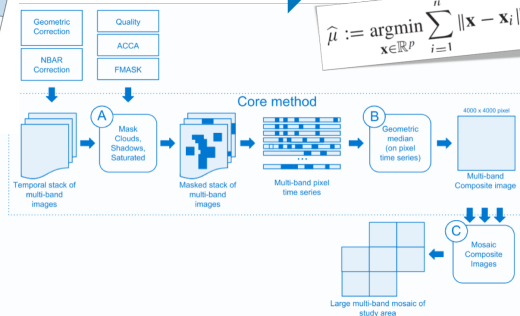


Australian Government

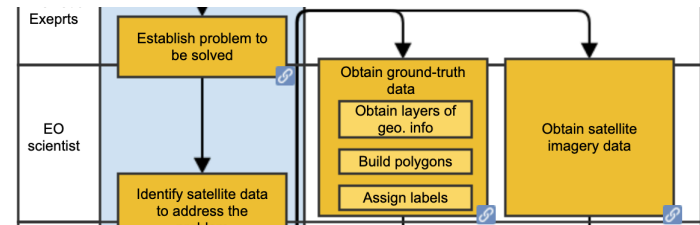
Geoscience Australia



3,707 images:  
3.1 T.B.



$$\hat{\mu} := \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^p} \sum_{i=1}^n \|\mathbf{x} - \mathbf{x}_i\|$$

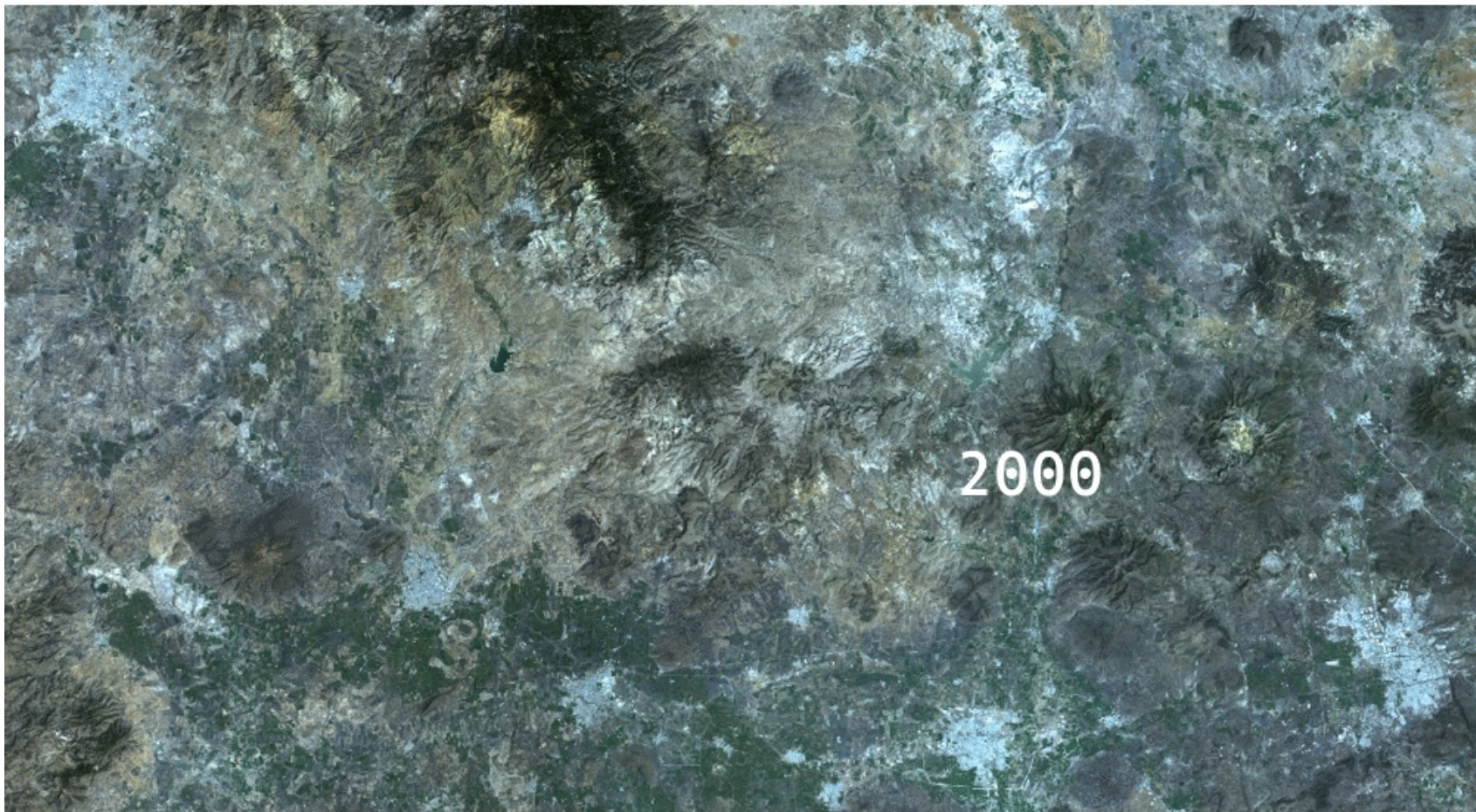


2,737,273,075 pixels – 35 Gb  
National Cloud-Free Mosaic



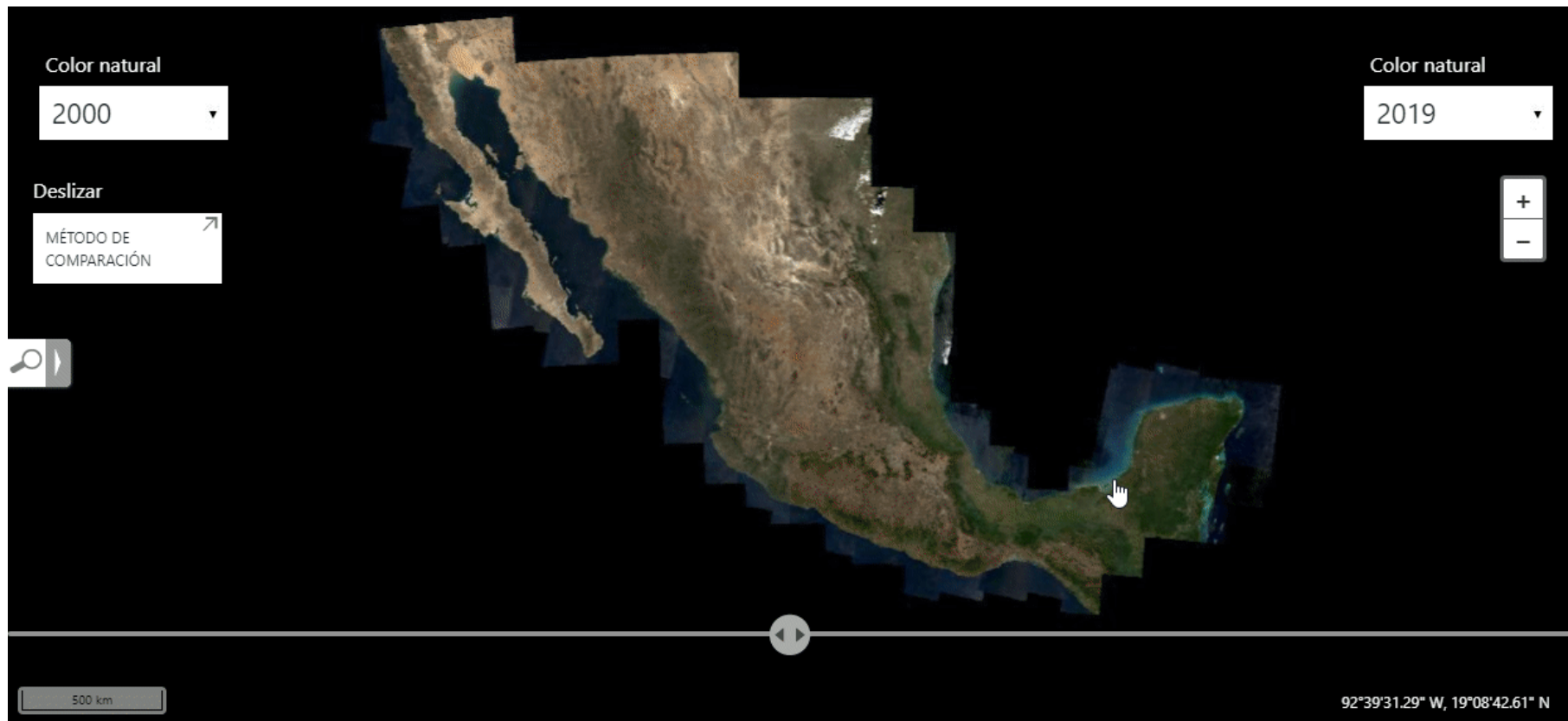




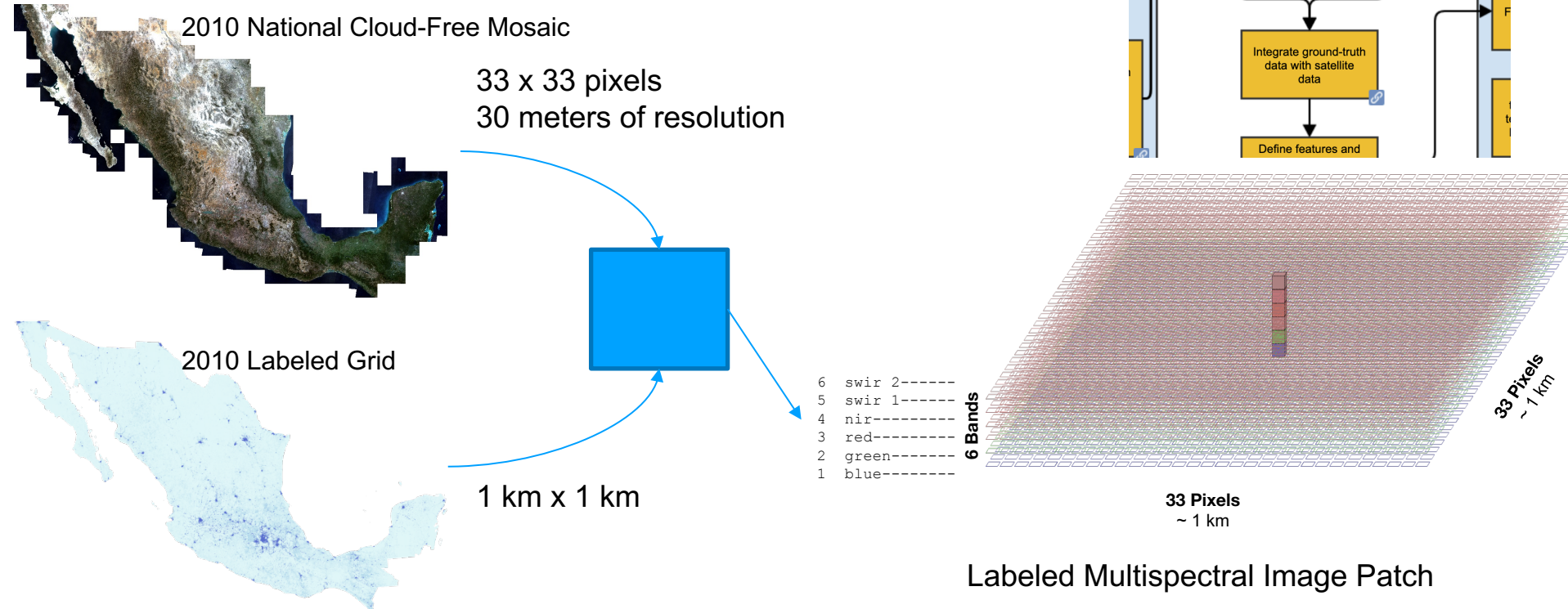








## Integrate ground-truth data with satellite data



## Take a Random Sample

- From national grid with the urban and non-urban labels, a random sample of 40,000 elements was taken: 20,000 for each class. Then, image patches were extracted from the cloud-free mosaic. Resulting in 40,000 images labeled, each one with 33 pixels x 33 pixels, with 6 spectral bands (or layers).



```

graph TD
    A[Define features and generate a dataset to be used for analysis] --> B[Train a model using the dataset]
    B --> C[Evaluate the model's performance]
    C --> D[Deploy the model to a production environment]
    D --> E[Monitor the model's performance]
    E --> F[Retrain the model if necessary]
    F --> A
  
```

Define features and generate a dataset to be used for analysis

The flowchart illustrates the proposed feature extraction pipeline for multi-band image patches. It starts with a multi-band image patch, which is processed through several parallel paths for each band (Blue, Green, Red, NIR, SWIR1, and SWIR2). The main processing steps are:

- Calculate Spectral Indices for each pixel in the patch:** This step leads to **Spectral Features**.
- Calculate 7 statistical measures:** This step leads to **Spectral Statistical Features**.
- Calculate 7 statistical measures:** This step leads to **Plain Statistical Features**.
- Calculate histogram:** This step leads to **Plain Histogram Features**.
- 12 GLCM Matrices:** This step leads to **Plain GLCM Features**.
- 6 Statistical properties for each GLCM calculated:** This step leads to **GLCM Statistical Features**.
- LBP Filter:** This step leads to a series of features:
  - Calculate 7 statistical measures:** Leads to **GLCM Histogram Features**.
  - Calculate Histogram:** Leads to **LBP Histogram Features**.
  - 12 GLCM Matrices:** Leads to **LBP-GLCM Plain Features**.
  - 6 Statistical properties for each GLCM calculated:** Leads to **LBP-GLCM Statistical Features**.
  - Calculate histogram:** Leads to **LBP-GLCM Histogram Features**.
- 40 Gabor Filters:** This step leads to **Gabor Features**.

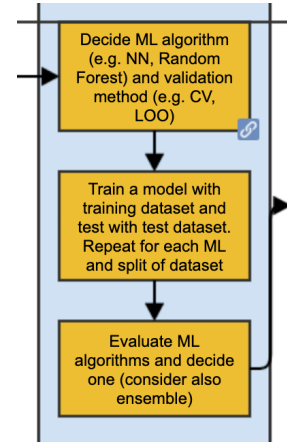
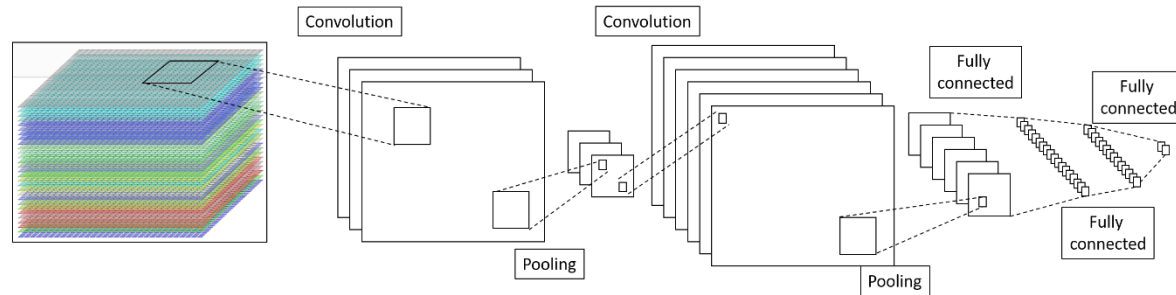
The final output is a set of features extracted from the multi-band image patch, including Spectral, Plain, GLCM, LBP, and Gabor features.

[illegible]

40,000 rows in a data matrix

## Models tried

Two different models were tested, an Extra Trees model also known as Extremely Randomized Trees and a LeNet Convolutional Neural Network





## Results

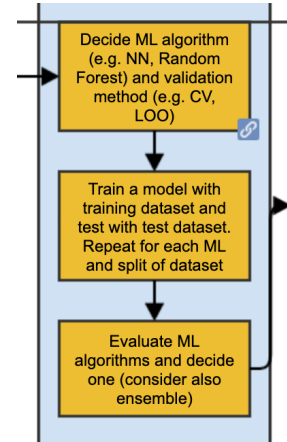
The evaluation with training data was performed 10-fold cross-validation, for both methods.

### Extra Trees

	precision	recall	f1-score
Non-Urban	0.92312	0.93532	0.92916
Urban	0.93438	0.92218	0.92821
O.A.			<b>0.92870</b>
macro avg	0.92873	0.92875	0.92868
weighted avg	0.92882	0.9287	0.92870

### LeNET

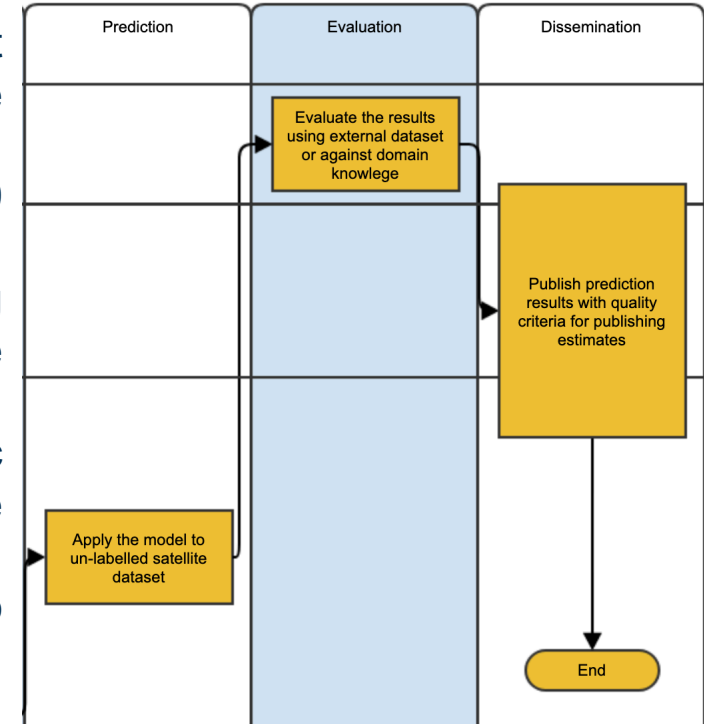
	precision	recall	f1-score
Non-Urban	0.91372	0.90296	0.90808
Urban	0.90465	0.91445	0.90932
O.A.			<b>0.90873</b>
macro avg	0.90919	0.90868	0.90869
weighted avg	0.90917	0.90873	0.90872



## | **Next Steps**

## Next Steps

- Finish other iteration, no later than one 2 months.
- The grid is also an important innovation and is evolving, it is likely that we will have a new version very soon and we should update the classification.
- Apply the model to un-labelled years, for example 2019 first semester.
- Validate a sample of the un-labelled year, requesting support for the area of visual interpretation of the geography division, to have a measure of product quality.
- Hold more meetings with the area of sociodemographic statistics, involve them in the exercise to improve the potential benefit in the population estimate.
- Hold meetings with the cartographic update area, to receive feedback.





**| GRACIAS!**

[abel.coronado@inegi.org.mx](mailto:abel.coronado@inegi.org.mx)

# Conociendo México

01 800 111 46 34

[www.inegi.org.mx](http://www.inegi.org.mx)

[atencion.usuarios@inegi.org.mx](mailto:atencion.usuarios@inegi.org.mx)



**INEGI** Informa