Agricultural Crop Classification with Synthetic Aperture Radar and Optical Remote Sensing

Part 4: Operational Crop Classification Roadmap

Tereza Roth (RUS Copernicus) - October 14, 2021

Training Outline

October 5, 2021

Synthetic Aperture Radar (SAR) Refresher

October 7, 2021

Optical Remote Sensing Refresher and Introduction to SNAP

October 12, 2021

Operational Crop Classification Roadmap using Optical and SAR Imagery (Part 1)

October 14, 2021

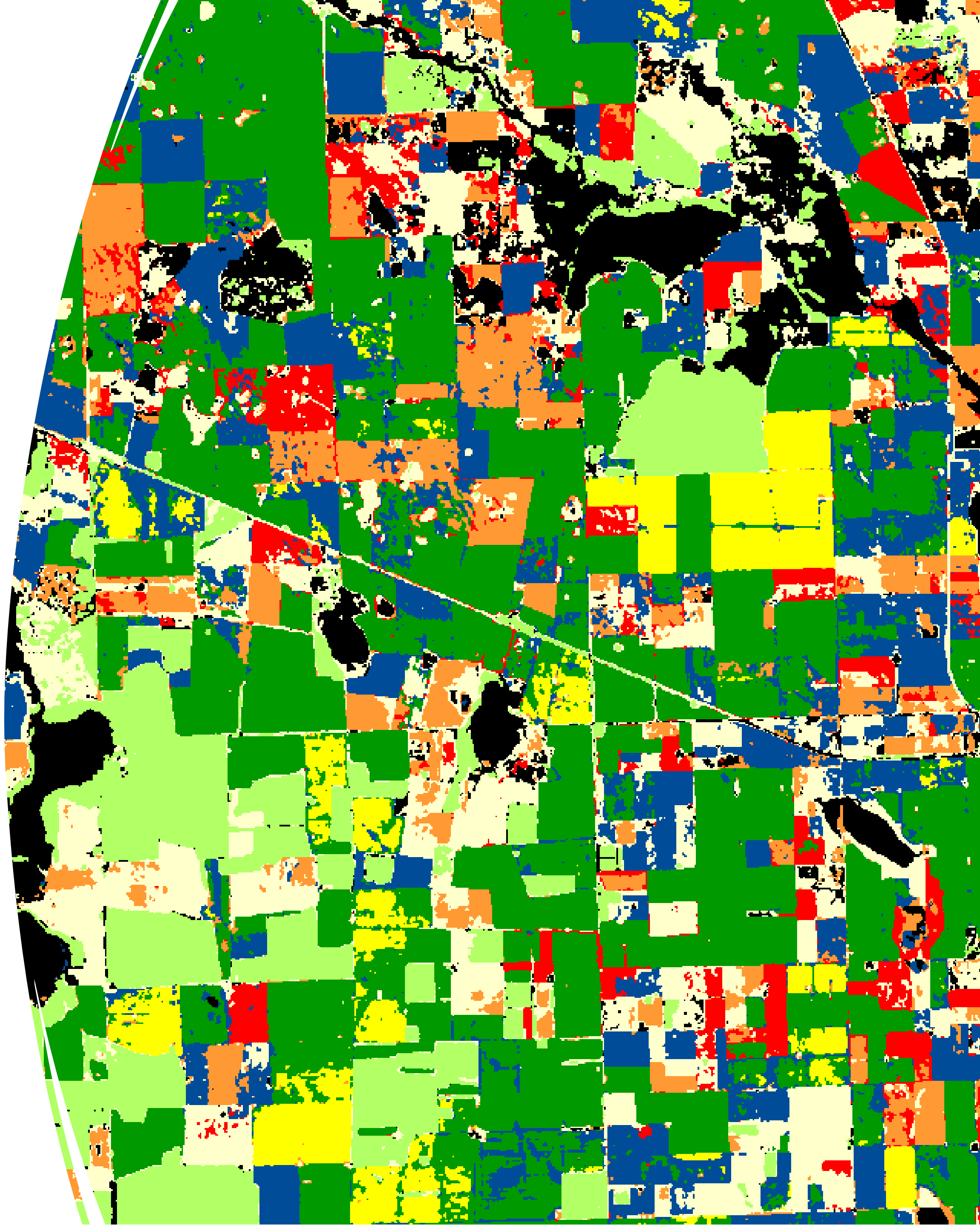
Operational Crop Classification Roadmap using Optical and SAR Imagery (Part 2)

October 19, 2021

Biophysical Variable Retrieval using Optical Imagery to Support Agricultural Monitoring Practices

Part 4 Overview

- Machine Learning for Crop Classification
- Random Forest Introduction
- Support Vector Machine Introduction
- K-Means Introduction
- Demonstration – Python for Crop Classification
  - Selected Python Libraries
  - Satellite and Training/Validation Data
  - Jupyter lab and Crop Classification
- Question and Answer Session



Machine Learning for Crop Classification

Machine Learning (ML)

A branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn.

Supervised ML Methods

- Use labeled datasets to train algorithms into classifying data or predicting outcomes accurately.
- Used for Classification (e.g., linear classifiers, decision trees, **Random Forest**, **Support Vector Machines**) and Regression (linear regression, logistic regression, polynomial regression).

Unsupervised ML Methods

- Use machine learning algorithms to analyze and cluster unlabeled datasets.
- Three main tasks: clustering (**K-Means**), association, and dimensionality reduction.

ML for Crop Classification

- ML is most widely used for operational crop classification.
- **Random Forest** is the most popular algorithm, followed by **Support Vector Machine** algorithms.
- Any errors in the reference dataset will cause degradation of accuracy in our model → **quality control is critical** (min. 3,000 to 10,000+ samples).
- ML models usually are not well transferable between seasons or studies.
- Random Forest (RF) is best for large datasets with many classes.
- Support Vector Machine (SVM) is suitable when the number of classes is small and is sensitive to a balanced share of samples between classes.

Accuracy Assessment

Overall accuracy misleading → **F-Score Metrics**

$$F - Score = 2 \cdot \frac{precision \cdot recall}{precission + recall}$$

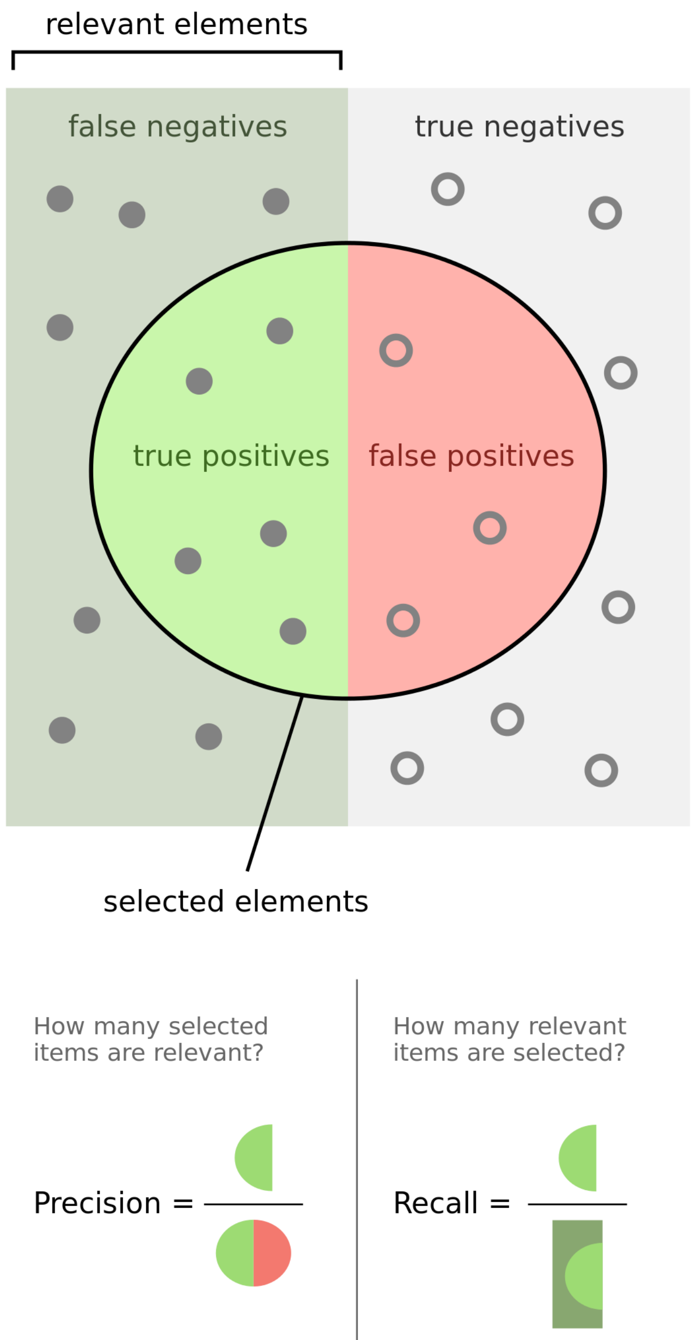
Criteria for a valid accuracy assessment:

- Statistically-rigorous, quality-assured, reliable, transparent, reproducible

Resources:

P. Olofsson, G. M. Foody, M. Herold, S. V. Stehman, C. E. Woodcock, and M. A. Wulder, 'Good practices for estimating area and assessing accuracy of land change', *Remote Sensing of Environment*, vol. 148, pp. 42–57, May 2014, doi: 10.1016/j.rse.2014.02.015.

S. V. Stehman and G. M. Foody, 'Key issues in rigorous accuracy assessment of land cover products', *Remote Sensing of Environment*, vol. 231, p. 111199, Sep. 2019, doi: 10.1016/j.rse.2019.05.018.



Credits: Walber [CC BY-SA]

Random Forest Introduction

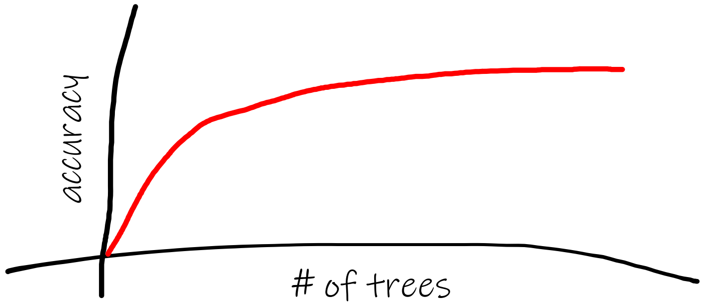
Random Forest

- Random forest is a **Supervised Machine Learning Algorithm** that is used widely in Classification and Regression problems.
- **Ensemble Classifier** - Consists of a large number of individual decision trees. Each tree provides a prediction; the class with the most votes is the model's prediction.
- **Bagging (Bootstrap Aggregation)** - Each individual tree randomly samples from the dataset with replacement.
- **OOB (Out-Of-Bag) Score** - Can be used to get an idea about the performance of the model as early as during the training phase.

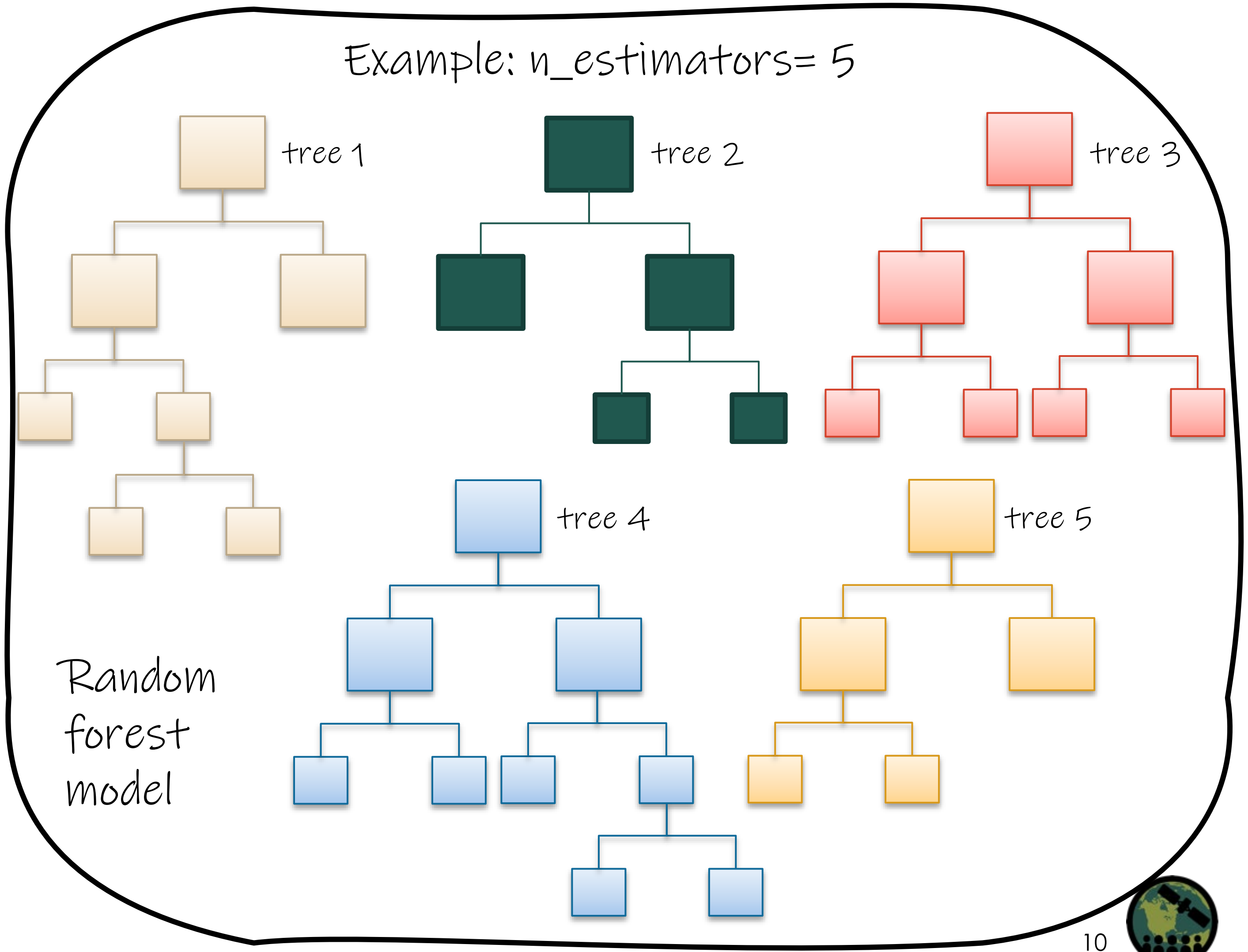
Random Forest - Parameters

Number of Trees (number of estimators, n_estimators = 100)

- ↑ number of trees = ↑ accuracy



- ↑ number of trees = ↑ computational cost
- For EO classification, **500** trees commonly used
- Optimal number can be determined using OOB error rate



Random Forest - Parameters

Criterion (criterion = gini)

- The function to measure the quality of a split
- *Gini* – [0,5] less computationally demanding
- *Entropy* – [0,1]

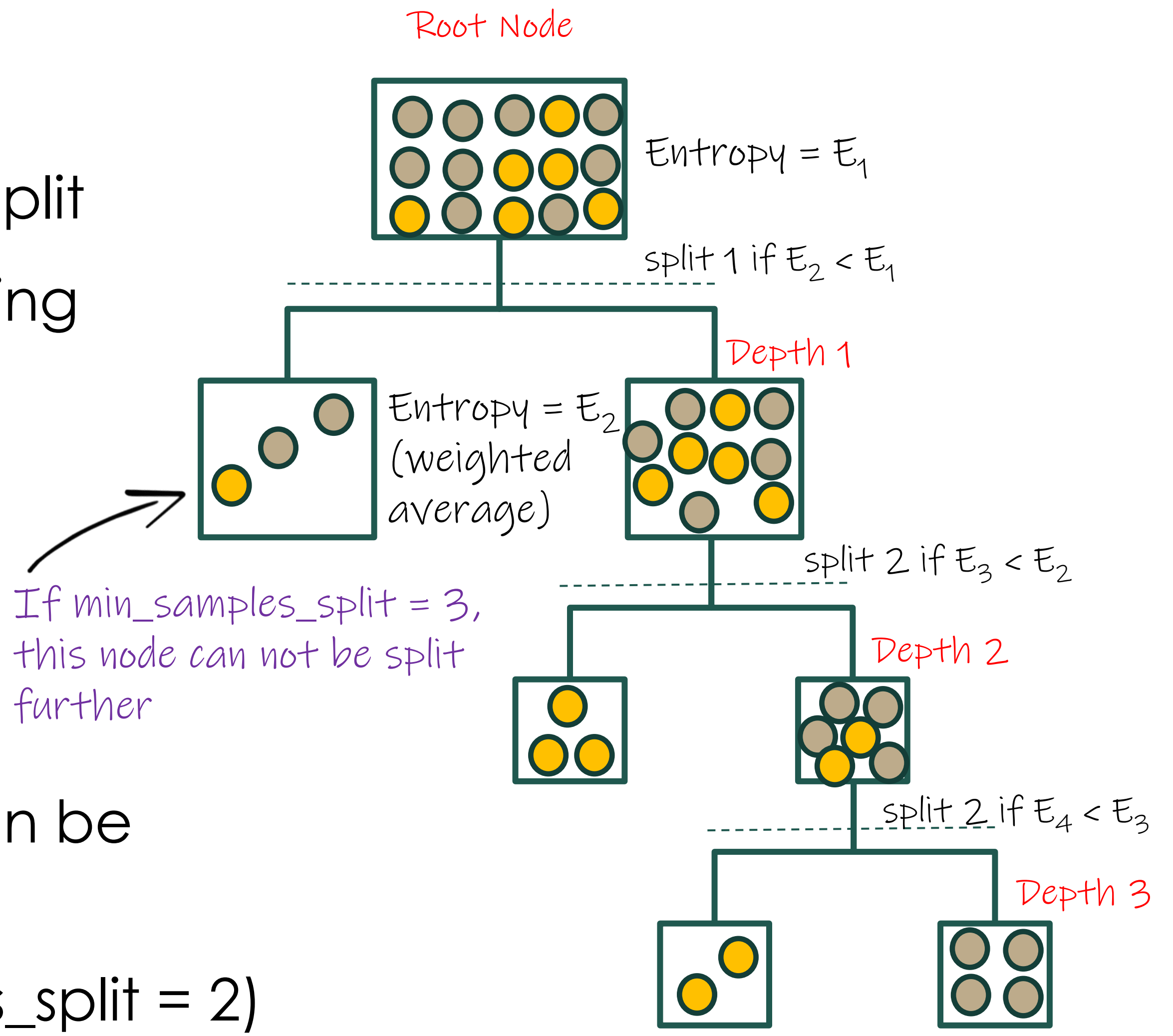
Maximum Features (max_features = “auto”)

- Maximum features tested at every split

Maximum Tree Depth (max_depth = None)

- Measure of how much further the tree can be expanded

Minimum Number of Elements (min_samples_split = 2)



Random Forest - Parameters

Bootstrap (bootstrap = True)

- If False, the whole dataset is used to build each tree.

Sample Size (max_samples = None)

- None → same size as original training dataset
- Int → specific number of samples
- Float [0, 1] → percentage of training dataset

Larger sample size ➔ higher accuracy ➔ higher computational cost

Many other parameters …

Training data of size N = 6

[1, 2, 3, 4, 5, 6]

⬇

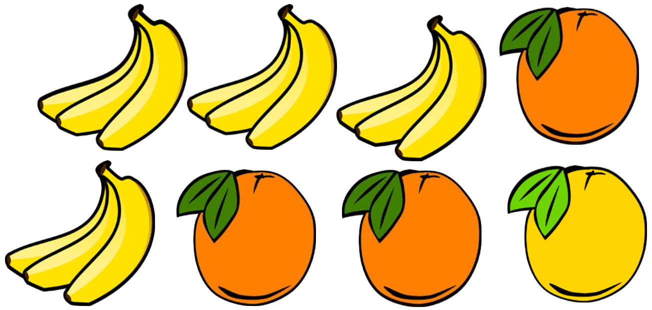
Bootstrap sample of size N = 6

[1, **2, 2**, 3, **6, 6**]

Random Forest - Example

n_estimators = 5
bootstrap = True
max_features = 3
min_samples_split = 2
max depth = None
max_samples = None

Training Data (size = 8)

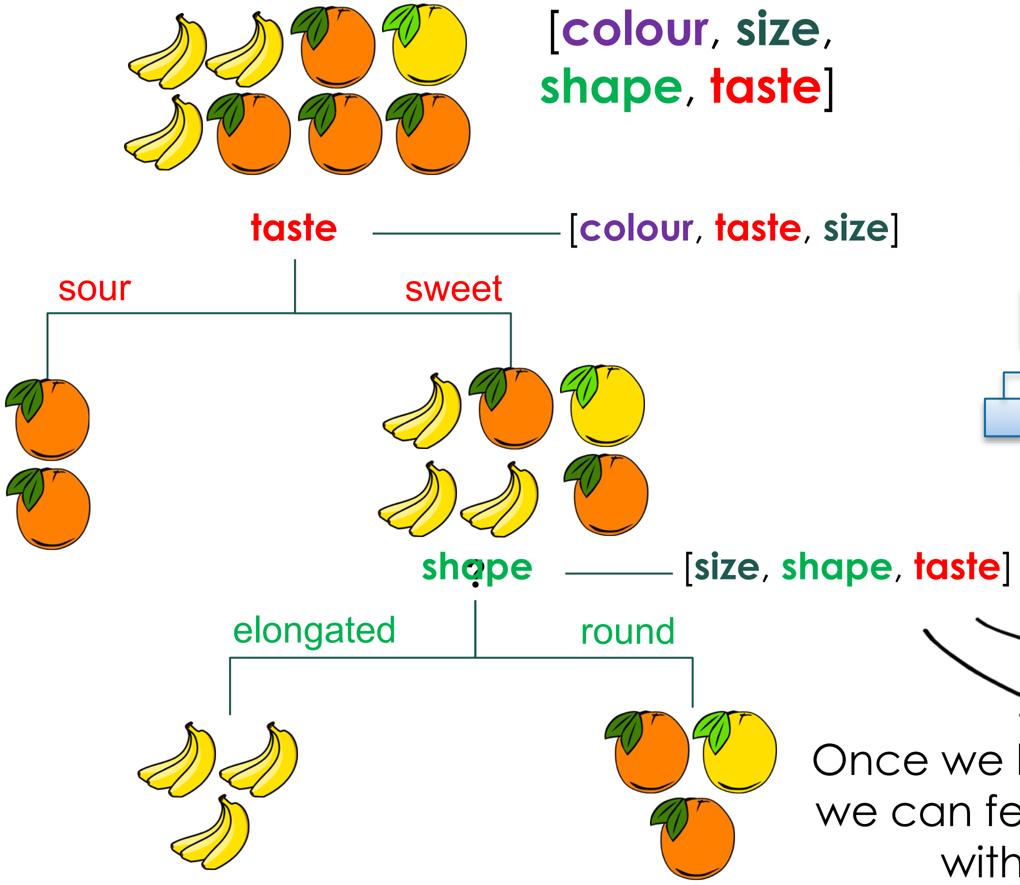


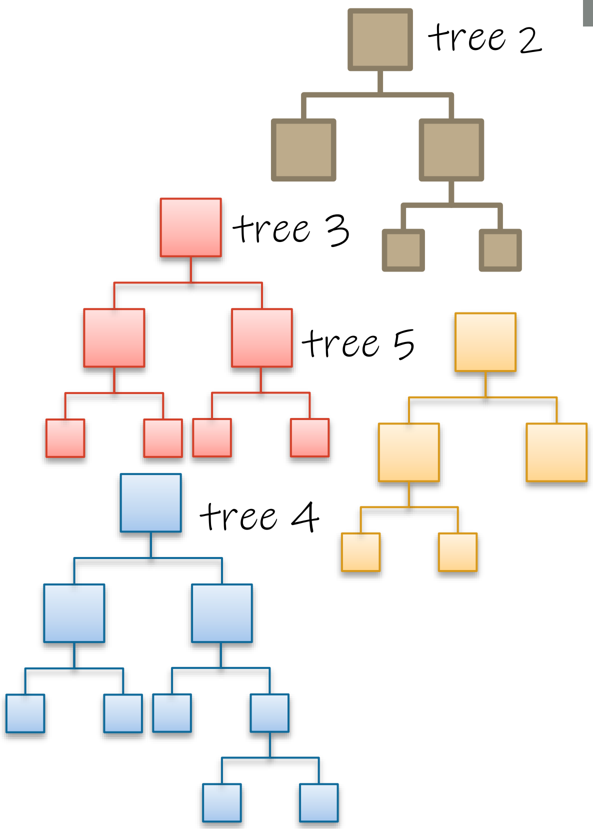
Random Sample 1 with Replacement (sample size = 8)

Tree 1

Features

[colour, size, shape, taste]





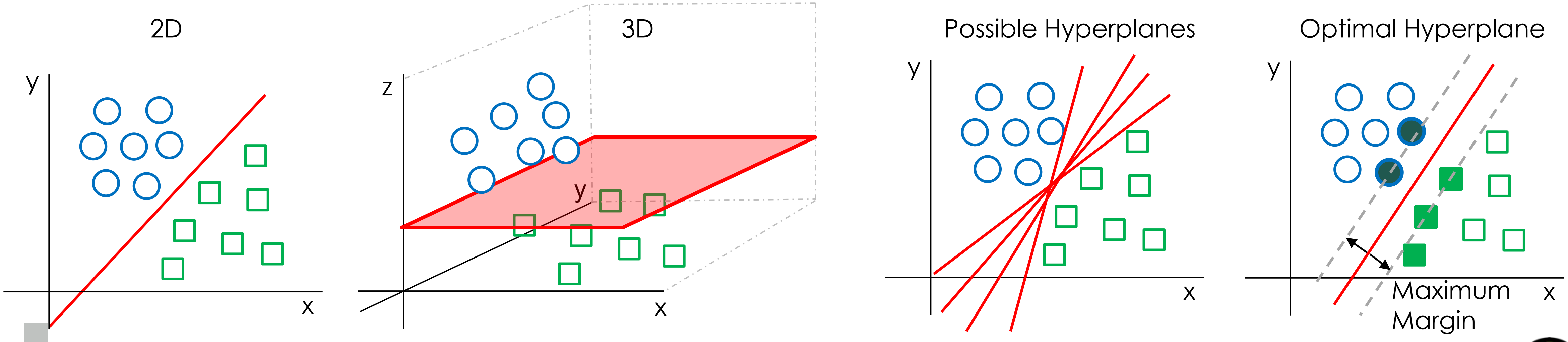
OOB score estimation

Once we have the model trained, we can feed it a new observation without knowing class

Support Vector Machine
Introduction

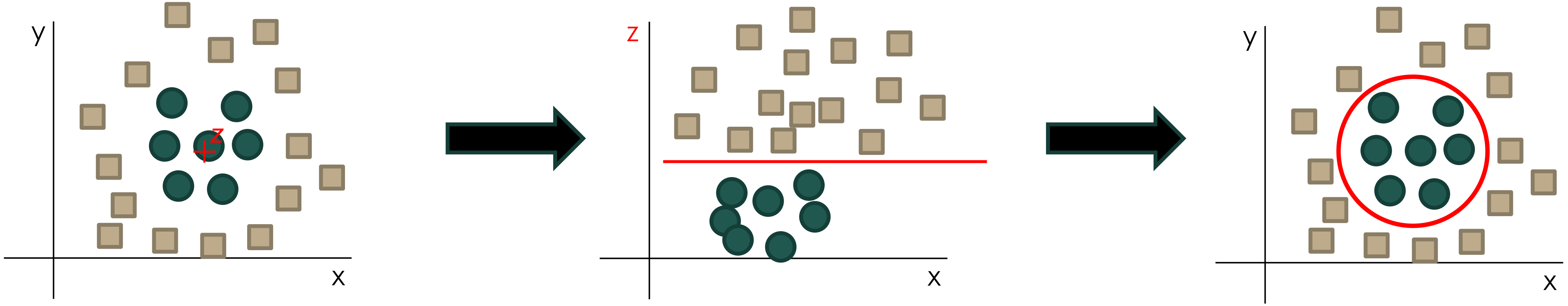
Support Vector Machine (SVM)

- Support vector machine (SVM) is a **Supervised Machine Learning Algorithm** used for classification and regression.
- The objective of the SVM algorithm is to find a **hyperplane** that distinctly classifies the data points.
- The dimension of the hyperplane depends upon the number of features.
- **Support Vectors** - Data points that are closer to the hyperplane and influence the position and orientation of the hyperplane.



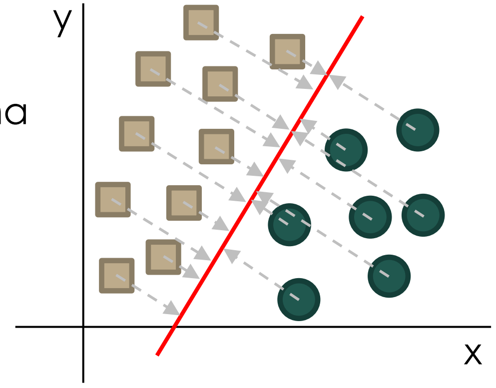
SVM – Parameters

Kernel (kernel = ‘rbf’) – Mathematical function that transforms the input data to a required form, such that a hyperplane can be fitted.

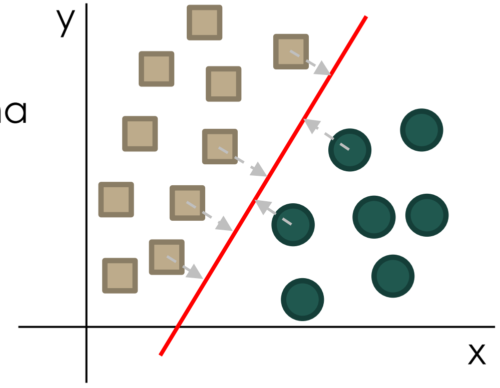


Gamma (gamma = ’scale’, only used if kernel polynomial, rbf or sigmoid) – Controls the distance of influence of a single training point.

Low Gamma Value



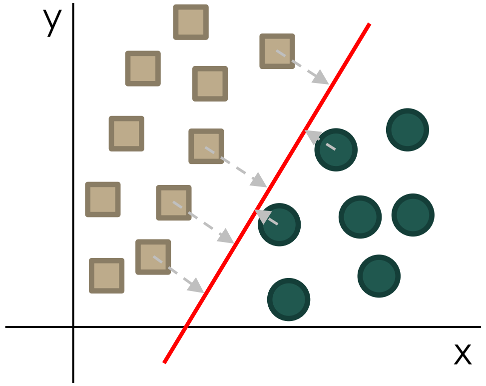
High Gamma Value



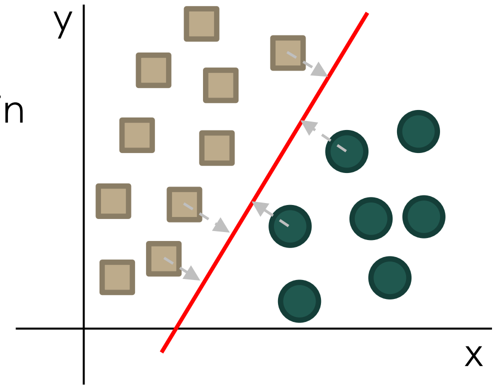
SVM – Parameters

Margin – A separation (distance) of the line/hyperplane to the closest class points.

Bad Margin

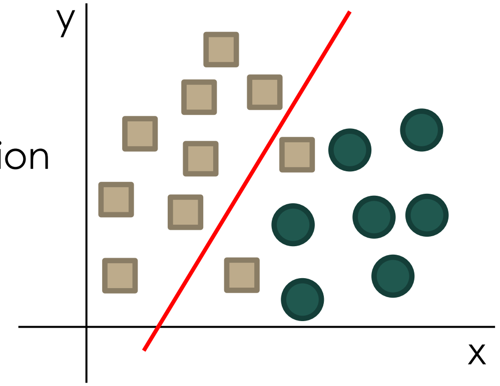


Good Margin

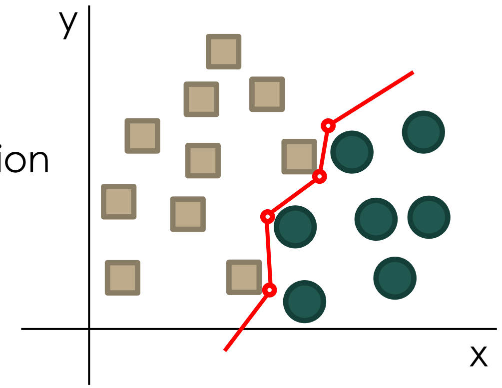


Regularization (C) – Tells the SVM optimization how much you want to avoid misclassifying each training example.

Low Regularization Value



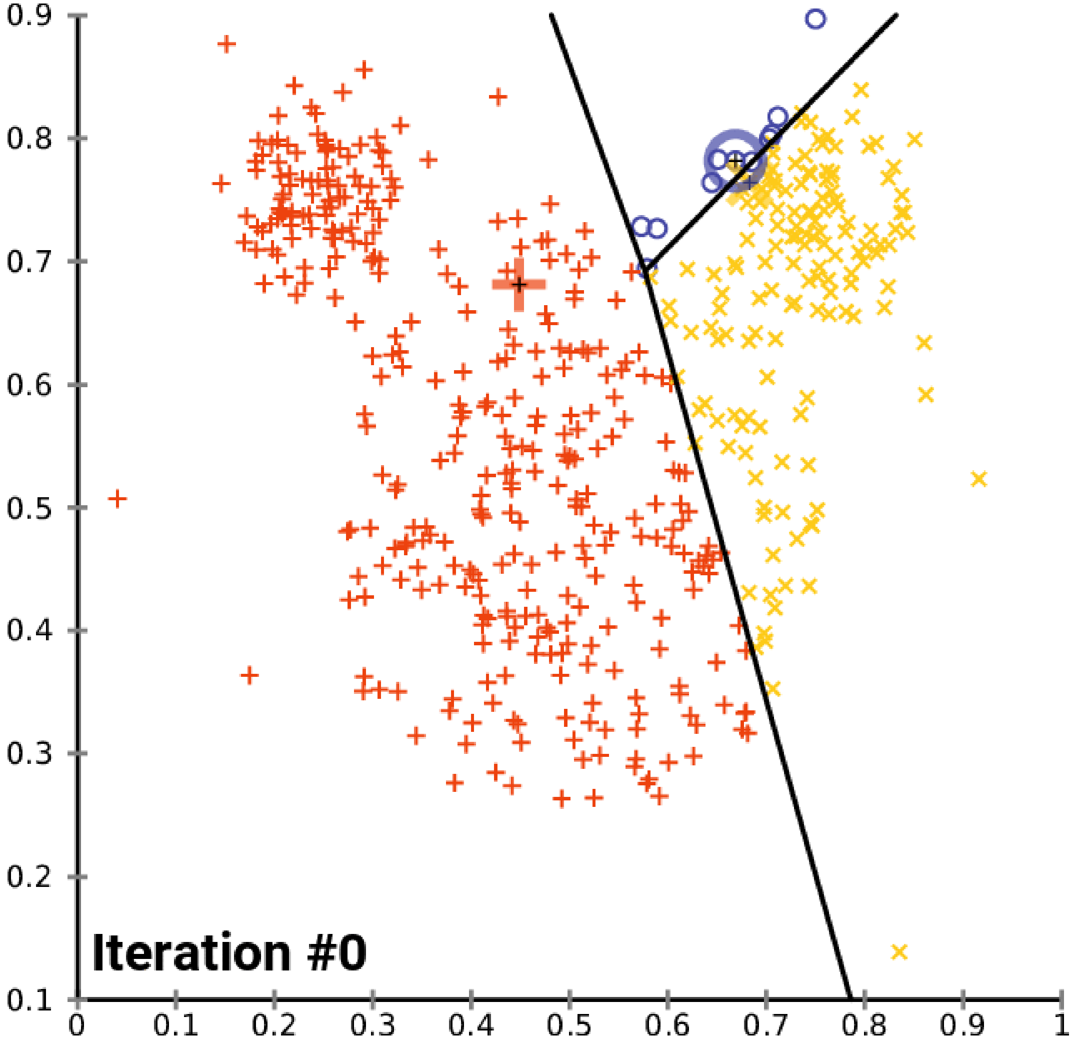
High Regularization Value



K-Means Introduction

K-Means

- Unsupervised clustering algorithm
- No training data needed
- Iterative algorithm that partitions the dataset into *K* pre-defined distinct non-overlapping subgroups (clusters)
- Minimizes the sum of the squared distance between the data points and the centroid
- Data normalization recommended



Credits: Chire [CC BY-SA]

K-Means - Parameters

Number of Clusters (n_clusters=8)

- Number of clusters we expect, or optimal number of clusters derived by, for example, the elbow method and others.

Initial Cluster Centroids (init = default='k-means++')

- If known, we can provide approximate locations of cluster centroids.

Number of Initial Centroid Selections (n_init = 10)

- The algorithm will initialize the centroids x times and will pick the most converging value as the best fit.

Random State (random_state)

- This is setting a random seed. It is useful if we want to reproduce exact clusters over and over again.

Demonstration – Python for Crop Classification

Selected Python Packages for ML

Scikit Learn

- Includes easy integration with different ML programming libraries like NumPy and Pandas.

Tensor Flow

- End-to-end, open-source platform for machine learning. Advanced functionality for deep learning models and neural networks.

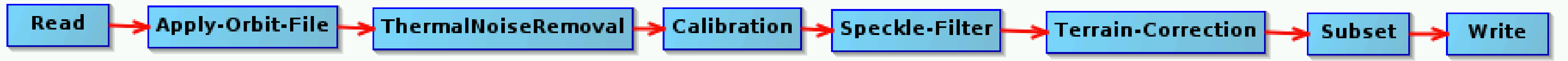
Keras

- One of the most popular open-source neural network libraries for Python. It extends the usability of TensorFlow with additional features for ML and DL programming.

Optical and Radar Input Dataset

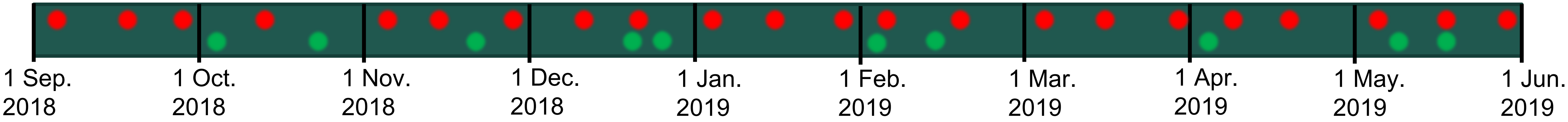
Sentinel-1 IW GRDH - 22 Images

- Preprocessed as shown in Part 3 of this webinar series



Sentinel-2 L1C - Images 10

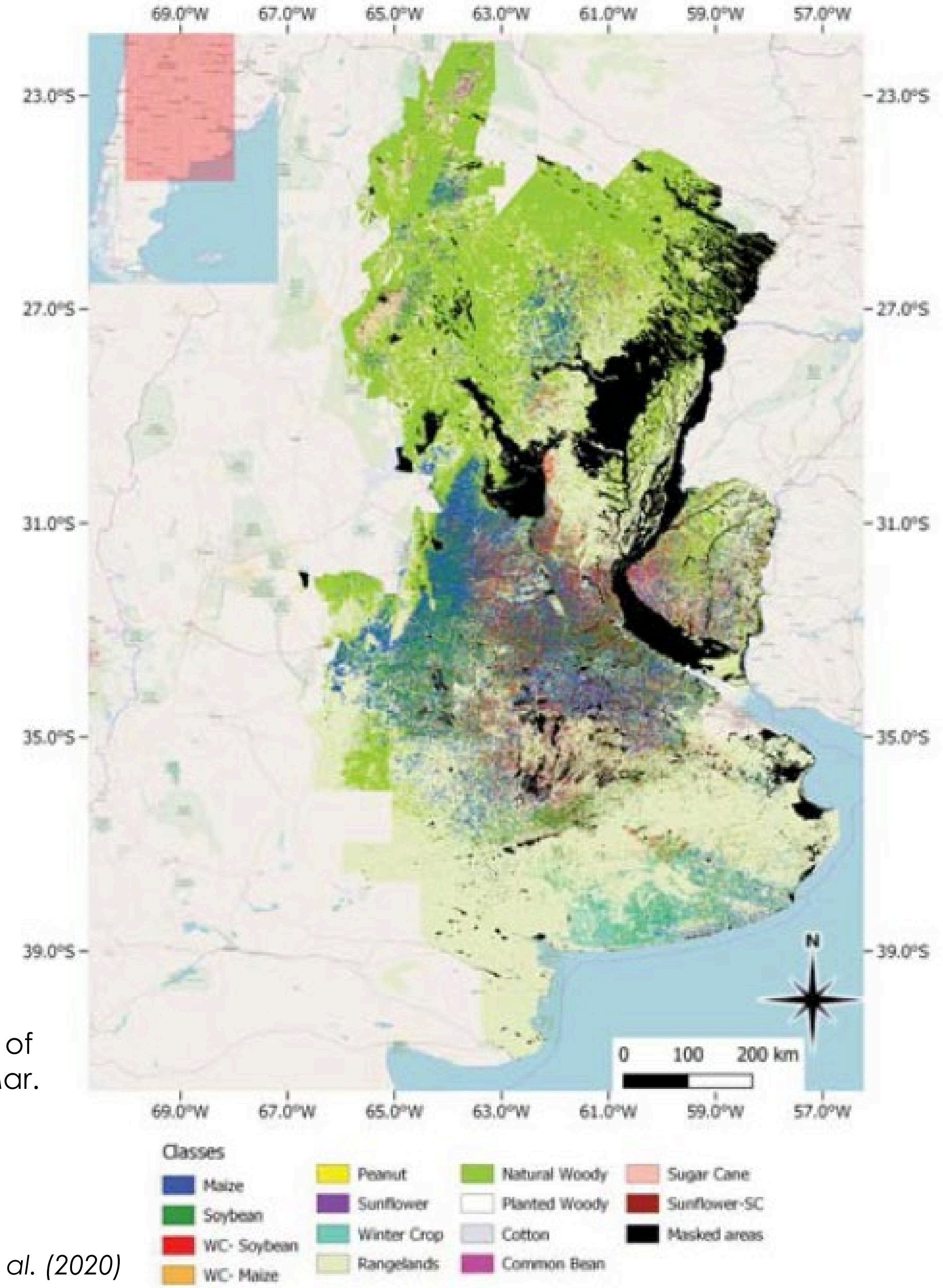
- Bands Used: B2 (blue), B3 (green), B4 (red), B8 (NIR), B11(SWIR 1), B12 (SWIR 2)
- Resampled to 10m and subset to study area



Training Data

- Map based on random forest classification using Landsat-8 and field samples along 14 agricultural zones with overall accuracy of 90% in zone X (used here)
- The data for our study area was vectorized, small polygons were dropped, and the number of polygons for the major classes was reduced.
- 8 classes present
- Split 70/30 to training and validation data
- 1,000 (200) points per class were extracted randomly from the training (validation) data.

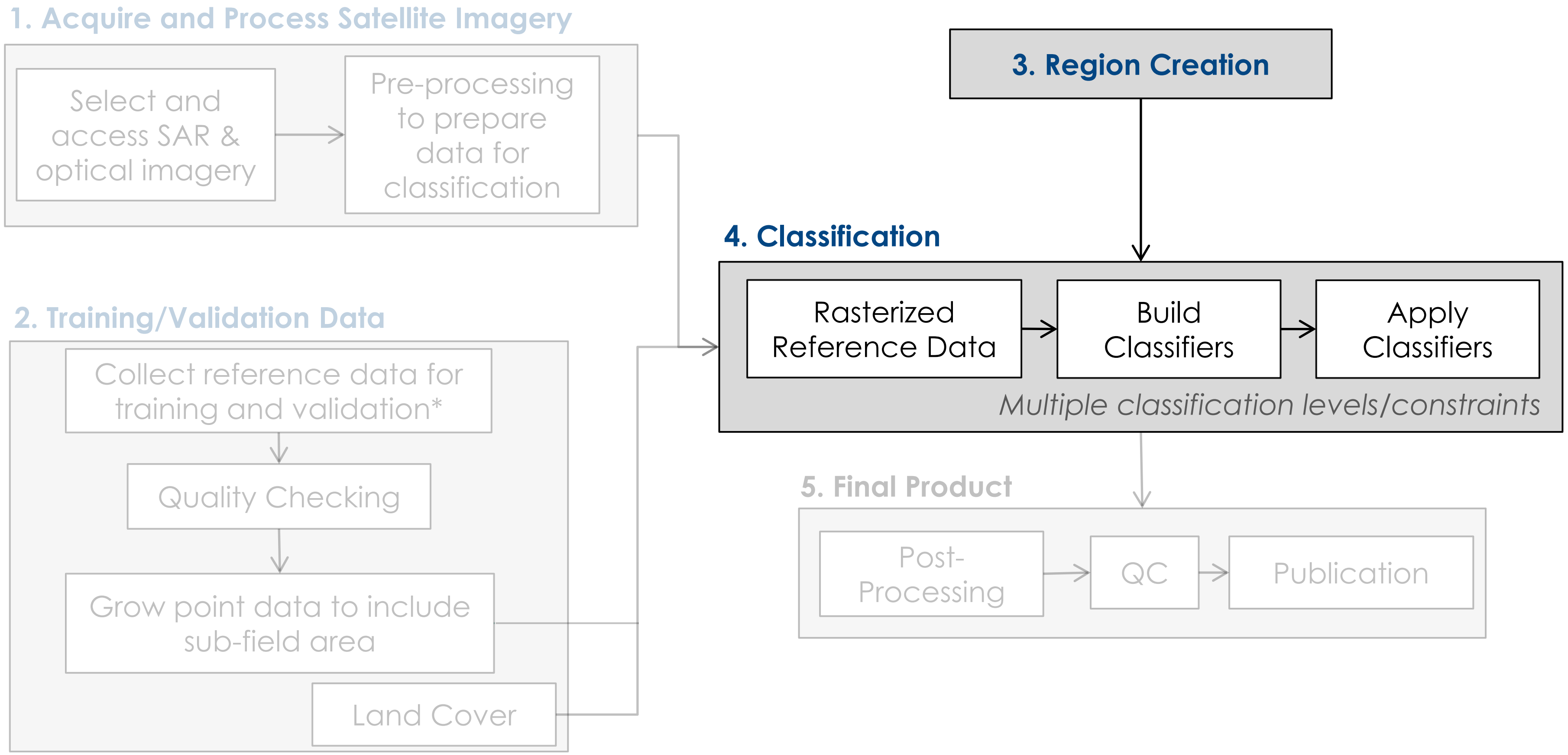
D. de Abelleyra *et al.*, 'First Large Extent and High Resolution Cropland and Crop Type Map of Argentina', in *2020 IEEE Latin American GRSS ISPRS Remote Sensing Conference (LAGIRS)*, Mar. 2020, pp. 392–396. doi: 10.1109/LAGIRS48042.2020.9165610.



Credits: de Abelleyra *et al. (2020)*

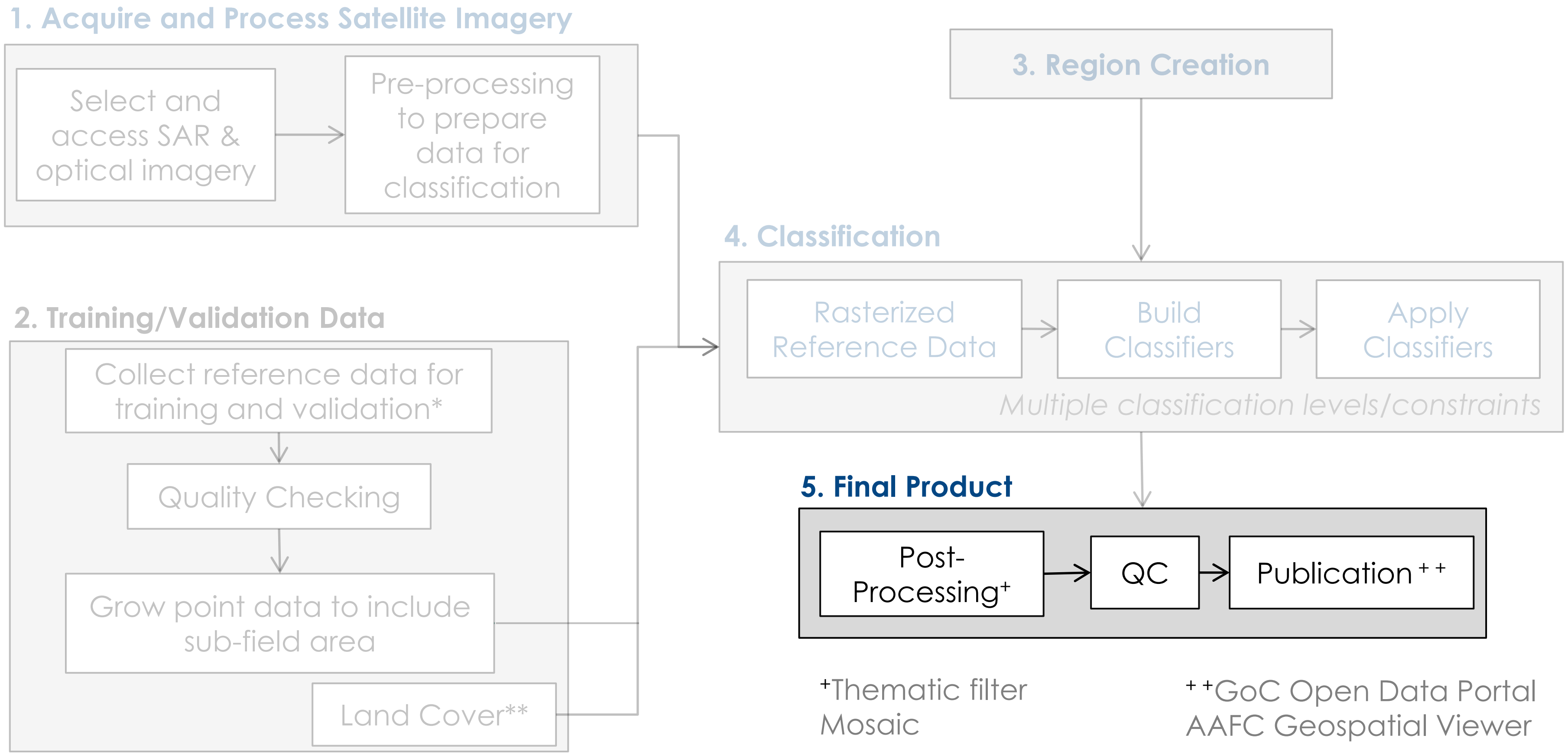
Classification with Python

Crop Inventory Operational Methodology



Creating the Final Product

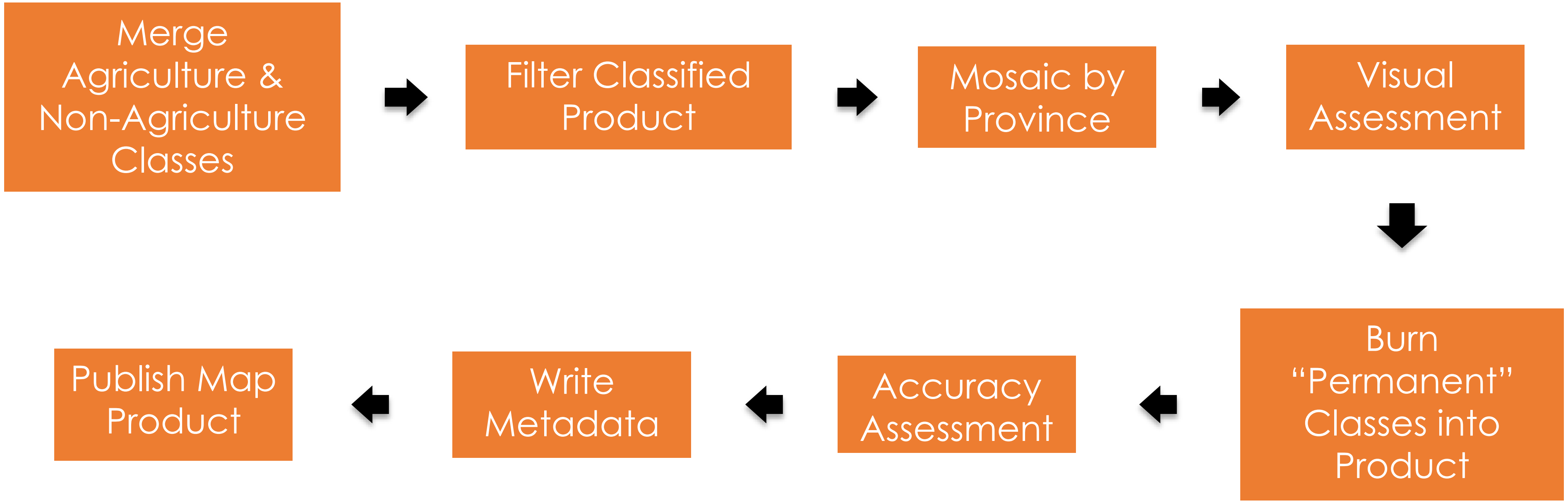
Crop Inventory Operational Methodology



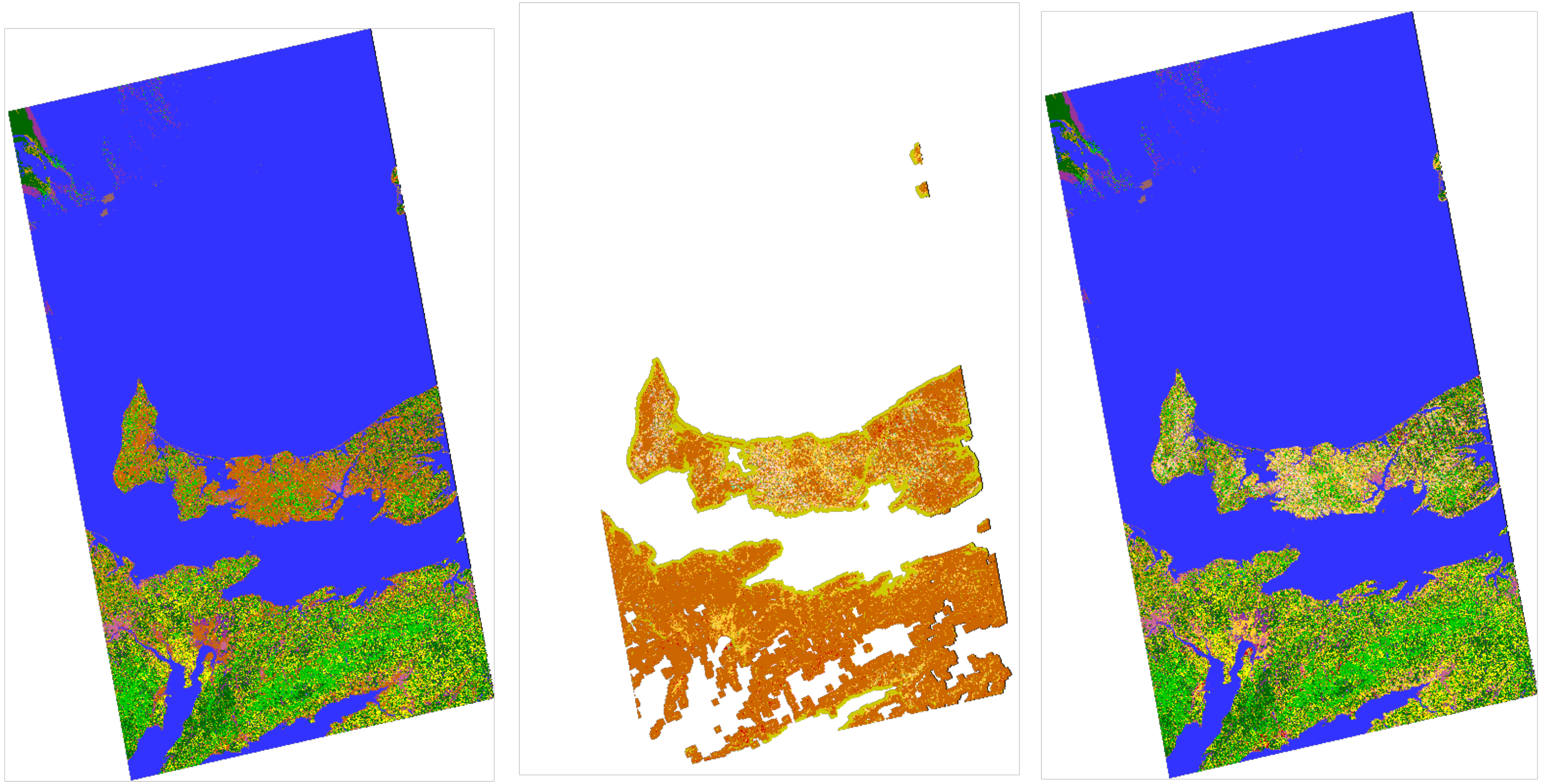
+Thematic filter
Mosaic

++GoC Open Data Portal
AAFC Geospatial Viewer

Steps for Post-Processing



Merge Agriculture and Non-Agriculture Classes



Courtesy: Agriculture and Agri-Food Canada

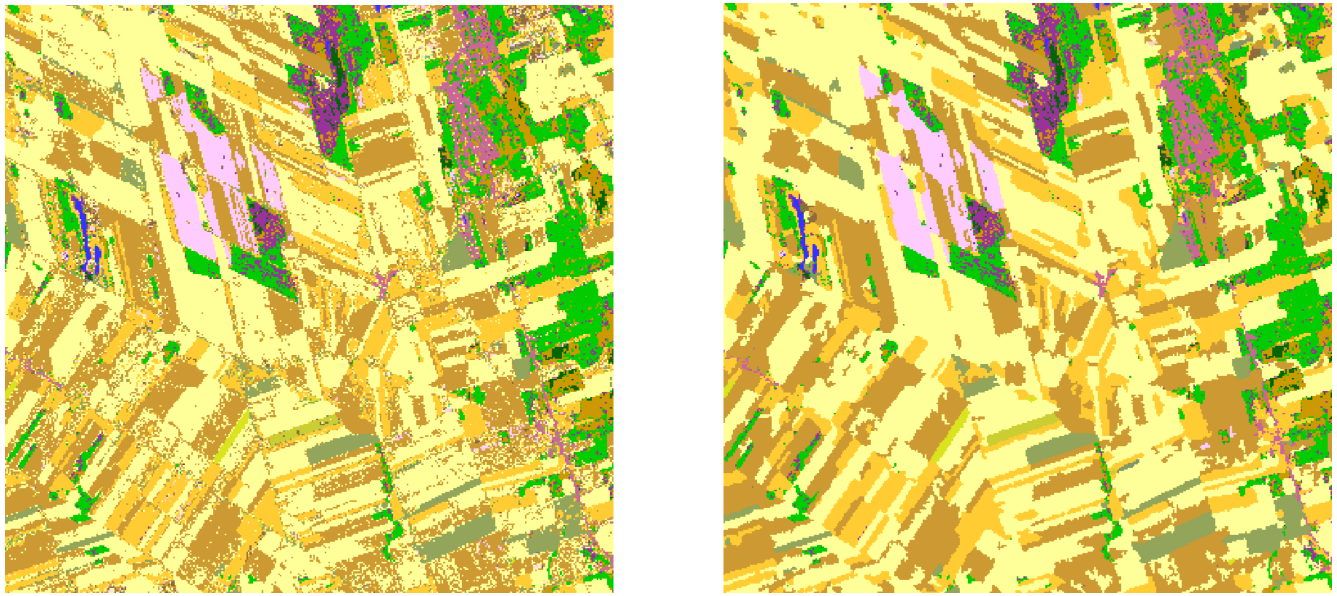
Land cover (left) and crop (middle) classifications are merged to form final classification (right).

Filter Map Product

AAFC uses a pixel-based classification which can result in “orphan” pixels within a classified field.

A mode filter is applied to reassign these orphan pixels to the majority class.

- A mode filter computes the mode of the center pixel values (the most frequently occurring value) within a particular filter window.
- Remember to consider field sizes when selecting kernel size for filtering.
- Think about instances where orphan pixels make sense (inter-cropping; trees within fields; water within fields etc.).

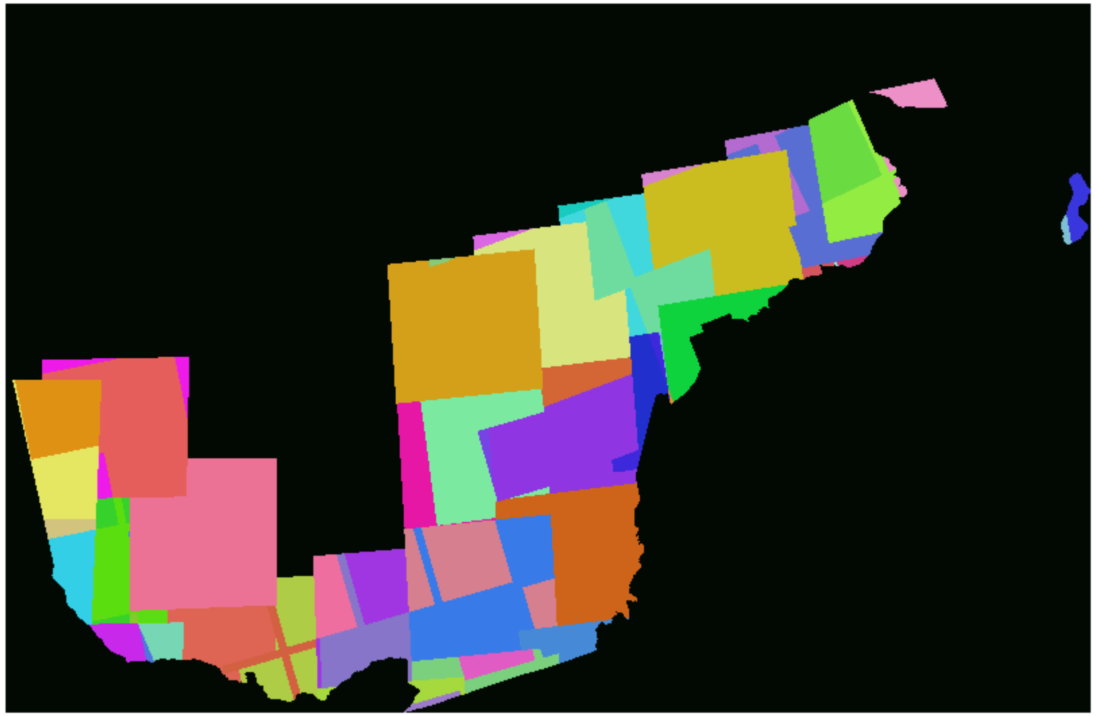


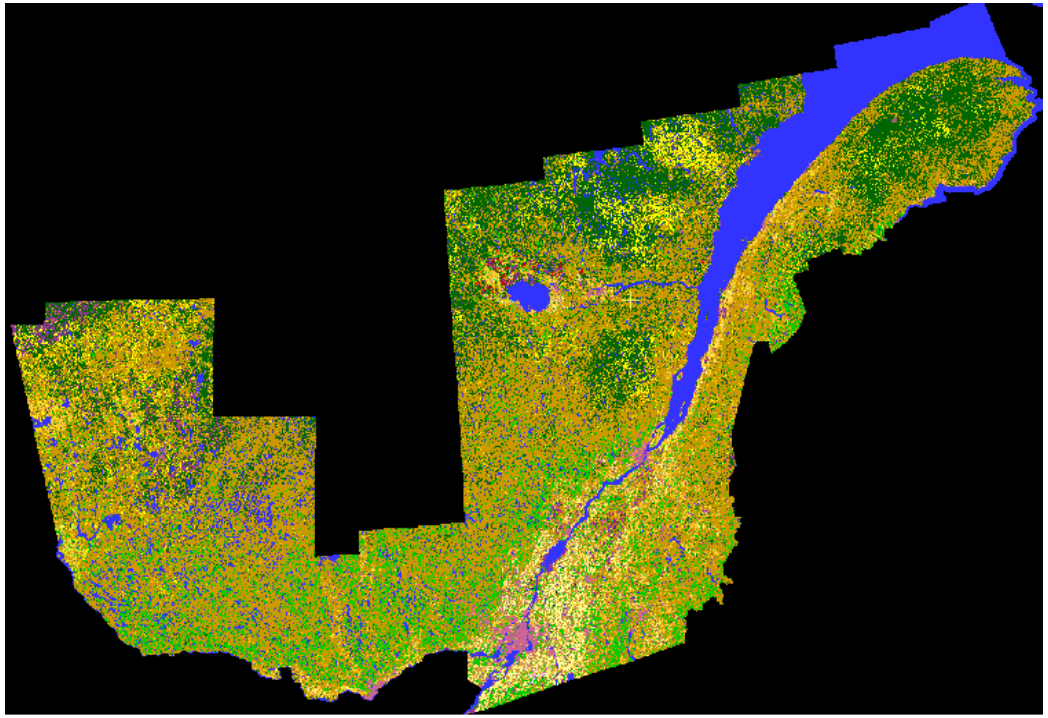
Raw classification (left) vs. filtered classification (right)

Mosaicking

AAFC mosaics within a province and then mosaics between provinces.

It is important to quality check the seams and overlaps.





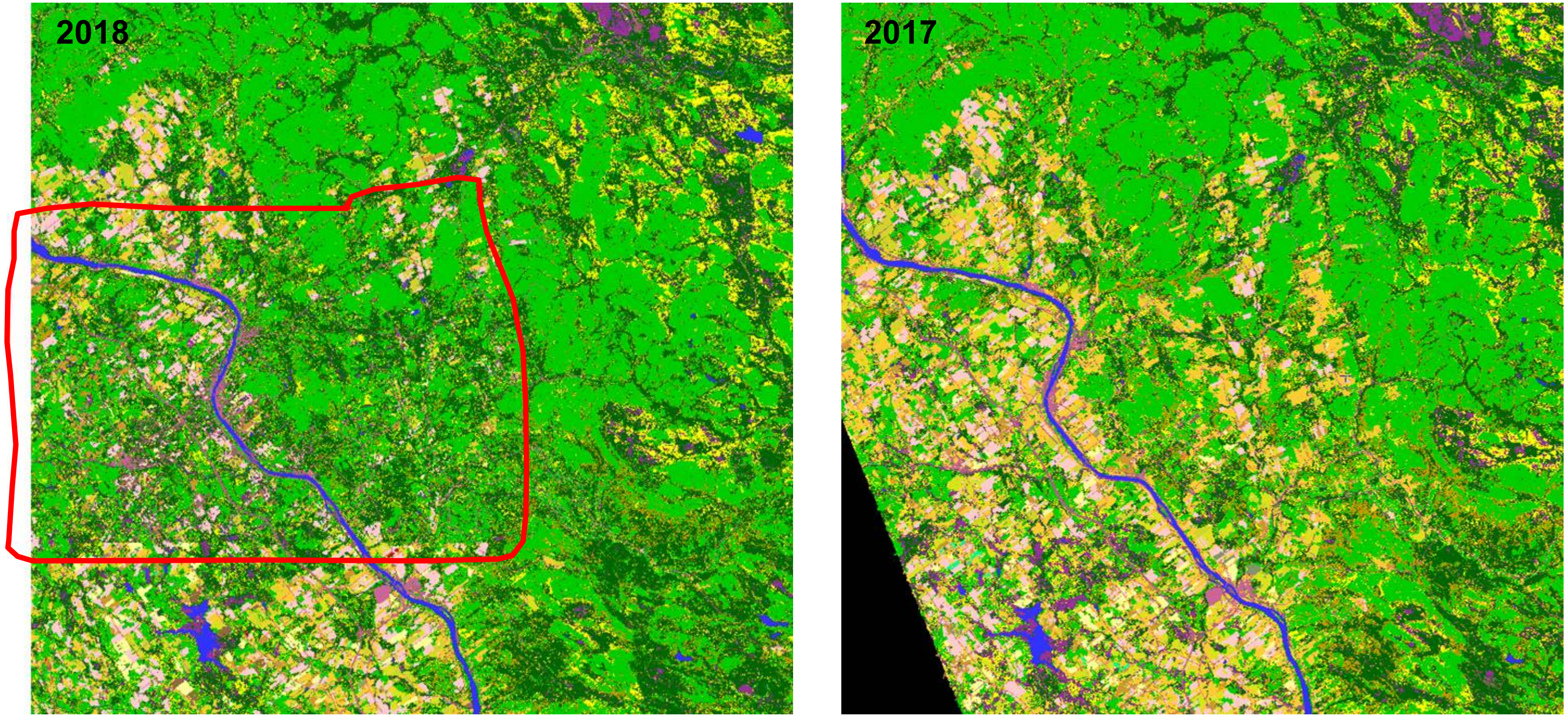
Individual stack overlap areas within the Quebec mosaic process.

Courtesy: Agriculture and Agri-Food Canada

Visual Assessment

AAFC operations undertake a manual visual inspection & revision, but only for major errors.

Given the operational nature of the ACI, it is important to balance map quality vs. resources (analyst time).



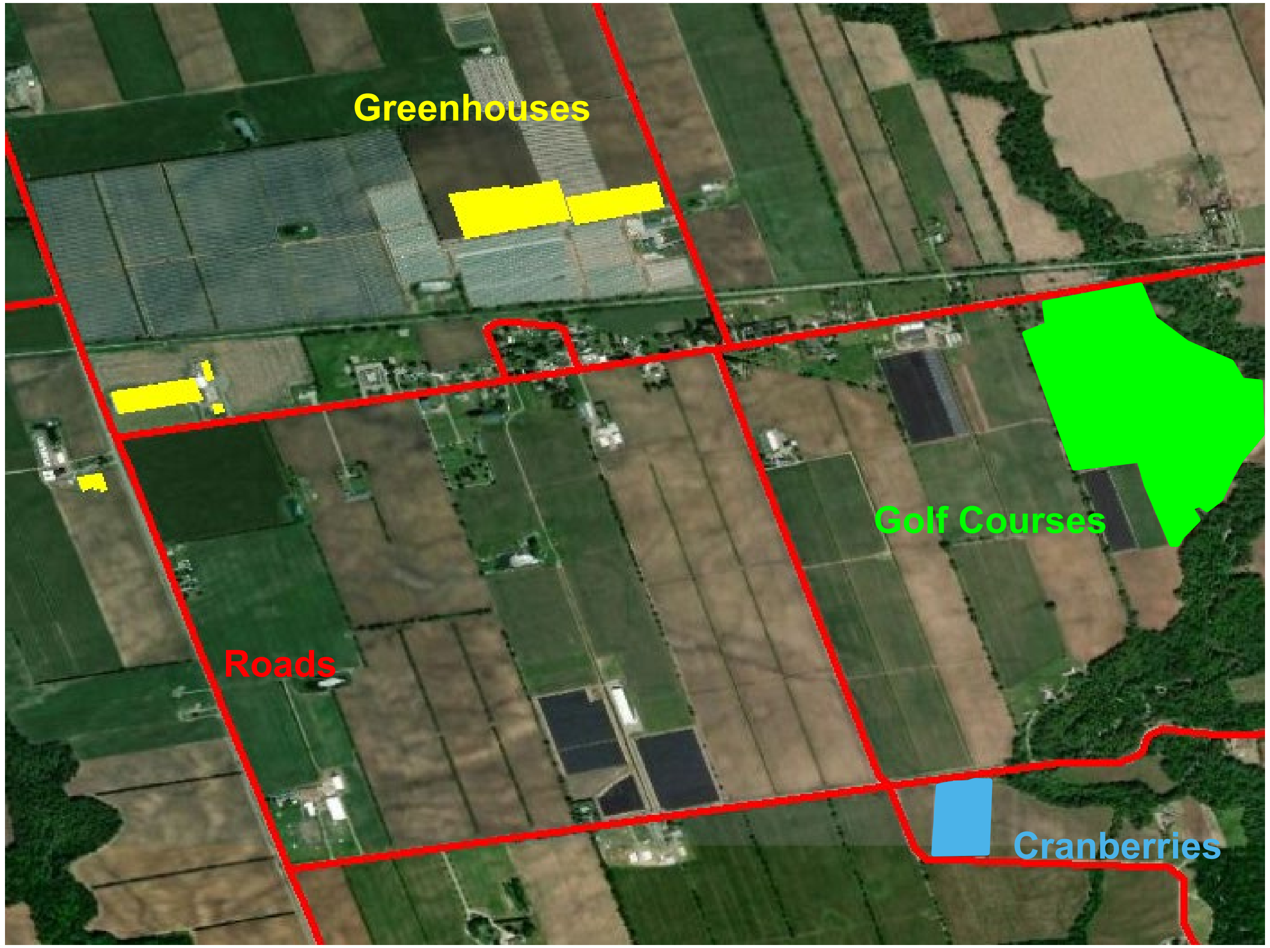
The area inside the red polygon underestimates the number of agricultural fields compared to the previous year. This error was detected by visual inspection and then corrected by the analyst.

Courtesy: Agriculture and Agri-Food Canada

Burn "Permanent" Classes

"Permanent" classes are then burned into the classified product.

These classes include, for example, greenhouses, roads, solar panel farms, golf courses, etc.

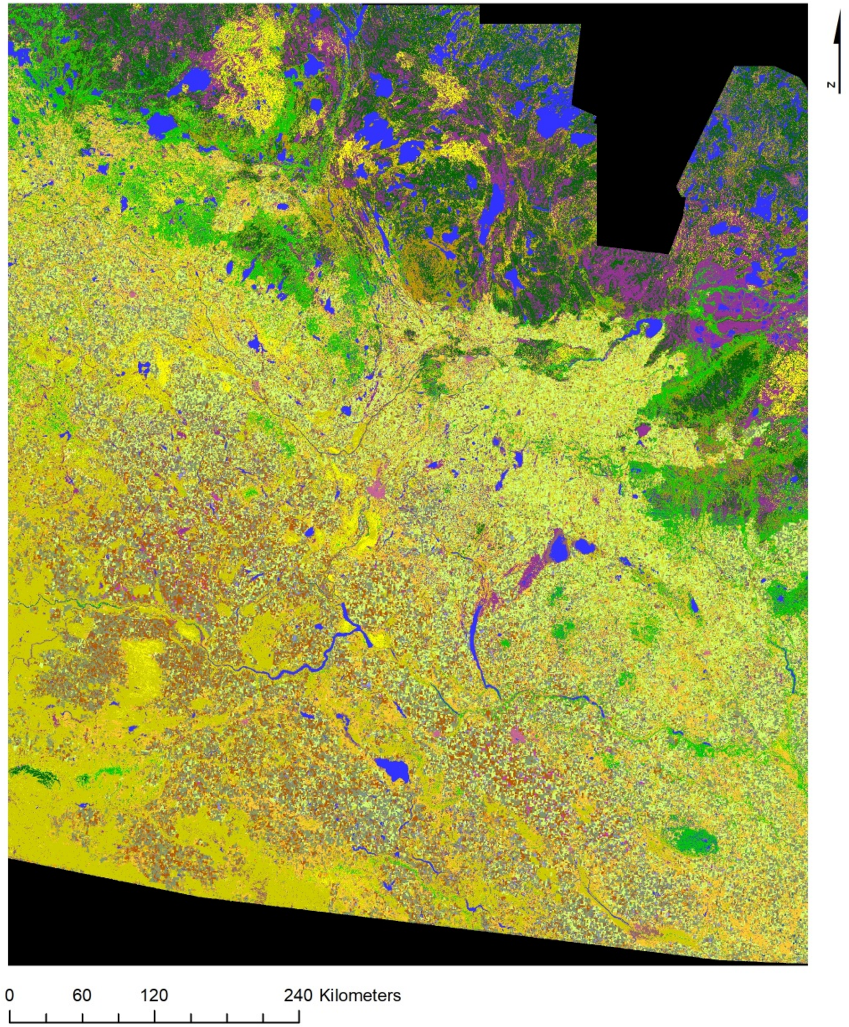


Courtesy: Agriculture and Agri-Food Canada

Accuracy Assessment

Confusion matrix, user's & producer's accuracy (errors of omission & commission); overall accuracy; kappa coefficient

Accuracy Assessment - Saskatchewan 2020		
	User's Accuracy	Producer's Accuracy
Pasture/Forage	85.91	90.70
Too Wet to be Seeded	65.46	33.48
Fallow	83.84	76.91
Barley	91.13	87.56
Millet	0.00	0.00
Oats	87.13	78.86
Rye	92.60	61.36
Spelt	98.02	75.51
Triticale	74.38	24.52
Quinoa	94.47	37.20
Winter Wheat	87.53	59.52
Spring Wheat	91.91	96.49
Corn	76.88	69.65
Camelina	91.36	37.77
Canola	96.60	98.39
Flaxseed	93.19	81.45
Mustard	96.77	73.73
Sunflower	92.19	94.41
Soybeans	95.08	88.92
Peas	96.21	92.28
Chickpeas	96.48	84.94
Beans	79.21	37.46
Fava beans	94.55	83.77
Lentils	94.44	96.26
Potatoes	39.41	88.33
Herbs	98.69	48.56
Buckwheat	93.25	45.41
Canaryseed	96.10	74.00
Hemp	85.43	40.86
Overall Accuracy	93.87	
Kappa:	0.92	



Courtesy: Agriculture and Agri-Food Canada

Write Metadata

- [ISO 19131 Annual Crop Inventory – Data Product Specifications]
- Overview of methodology
- Overall accuracies per province; individual crop accuracies can be obtained upon request
- Translation of geospatial products and metadata to French

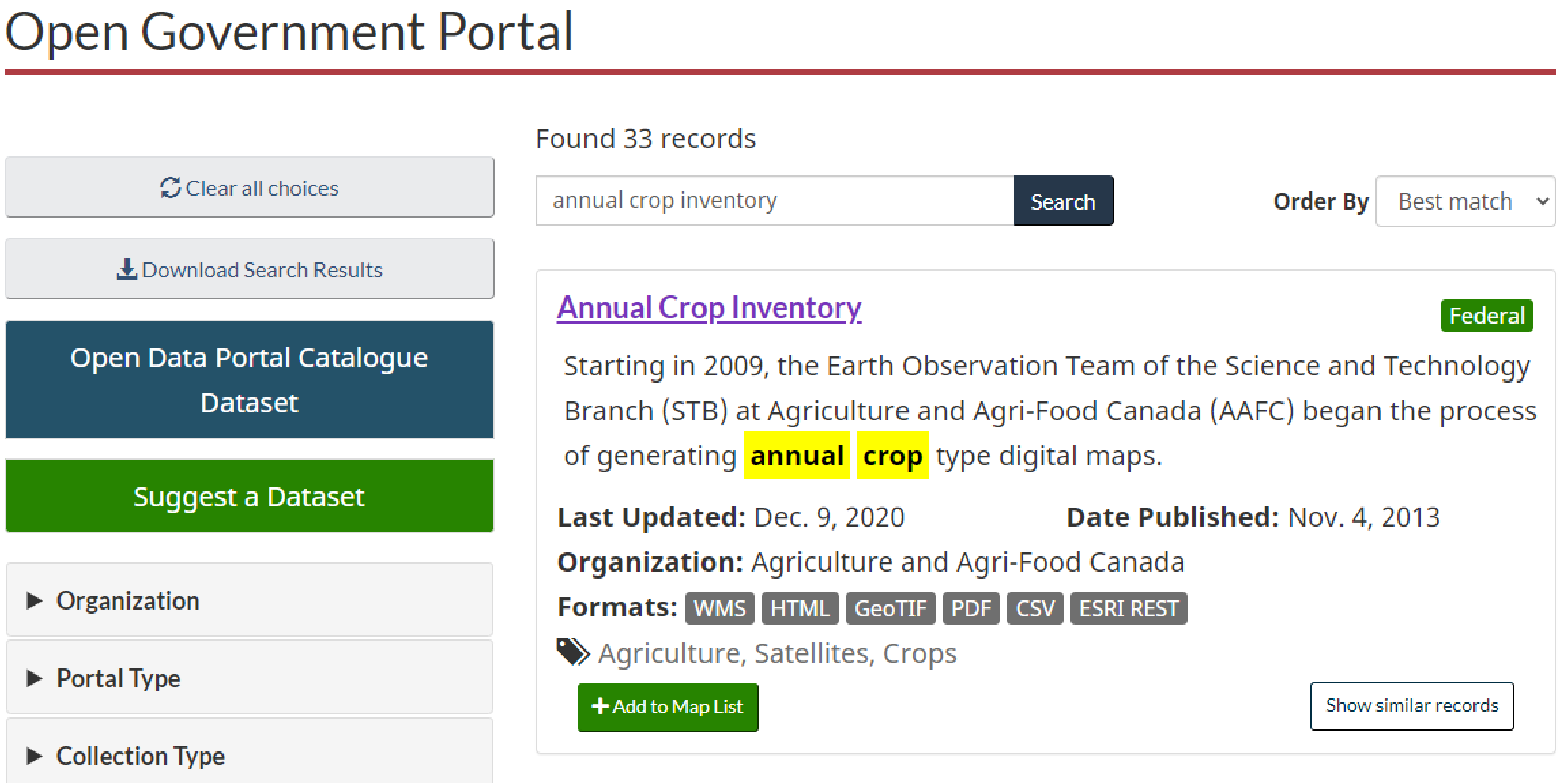
Agriculture and Agri-food Canada — Data Product Specification (ISO 19131)

3.2.12. Annual Crop Inventory, 2020

Title	Annual Crop Inventory, 2020
Alternate Title	AAFC Crop Type Mapping, 2020
Abstract	In 2020, the Earth Observation Team of the Science and Technology Branch (STB) at Agriculture and Agri-Food Canada (AAFC) repeated the process of generating annual crop inventory digital maps using satellite imagery to for all of Canada, in support of a national crop inventory. A Decision Tree (DT) based methodology was applied using optical (Landsat-8, Sentinel-2) based satellite images, and having a final spatial resolution of 30m. In conjunction with satellite acquisitions, ground-truth information was provided by: provincial crop insurance companies in Alberta, Manitoba, & Quebec; point observations from the PEI Department of Environment, Water and Climate Change; the Ontario Ministry of Agriculture, Food and Rural Affairs; and data collection supported by our regional AAFC Research and Development Centres in St. John's, Charlottetown, Fredericton, and Guelph. **Due to COVID-19 travel restrictions, complete sampling coverages in NL, NS, NB and BC were not possible, as a result the general agriculture class (120) is found in these provinces in areas where there was no ground data collected**.
Purpose	An annual national crop type map
Topic Category	Farming; Environment; GeoscientificInformation; imagery; BaseMaps; EarthCover;
Spatial Representation Type	grid
Spatial Resolution	30 m
Geographic Description	Canada
Supplemental Information	Overall provincial accuracies for crop classes are: Newfoundland: 95.08% Prince Edward Island: 85.85% Nova Scotia: N/A New Brunswick: 95.74% Quebec: 91.20% Ontario: 88.26% Manitoba: 93.58% Saskatchewan: 93.87% Alberta: 90.99% British Columbia: 85.16% Overall provincial accuracies for non-agriculture land cover are: Newfoundland: 75.09% Prince Edward Island: 75.74% Nova Scotia: 70.52% New Brunswick: 68.57% Quebec: 72.85% Ontario: 75.84% Manitoba: 68.97% Saskatchewan: 69.93% Alberta: 65.59% British Columbia: 77.09% Citation: Agriculture and Agri-Food Canada, 2020, "Annual Space-Based Crop Inventory for Canada, 2020", Agroclimate, Geomatics and Earth Observation Division, Science and Technology Branch. https://open.canada.ca/data/en/dataset/ba2645d5-4458-414d-b196-6303ac06c1c9
Constraints	Data are subject to the Government of Canada Open Data Licence : http://open.canada.ca
Keywords	Government of Canada Core Subject Thesaurus (2000-02-01) - Remote Sensing, Satellites, Agriculture, Crops, Crop insurance, Farmlands, Forage crops, Land cover, Geomatics, Geographic Information Systems, Geographic data, maps, Geography
Scope Identification	dataset
Feature Attribute Names	Class(Value)

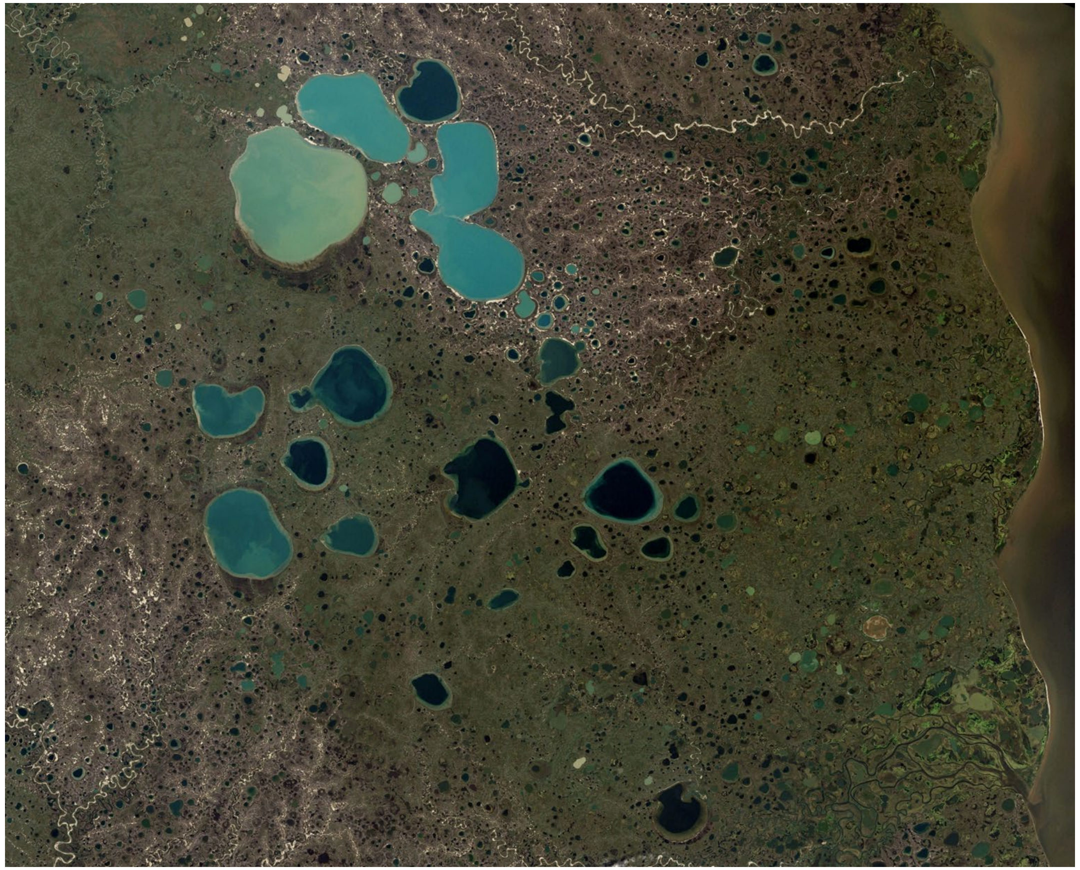
Publish to Open Canada

- [Demo]



Questions?

- Please enter your questions in the Q&A box. We will answer them in the order they were received.
- We will post the Q&A to the training website following the conclusion of the webinar.



https://earthobservatory.nasa.gov/images/6034/pothole-lakes-in-siberia

Contacts

- Trainer:
  - Tereza Roth: eotraining@serco.com
  - Georgia Karadimou: eotraining@serco.com
  - Dr. Laura Dingle Robertson, Agriculture and Agri-food Canada: Laura.Dingle-Robertson@AGR.GC.CA
- Training Webpage:
  - https://appliedsciences.nasa.gov/join-mission/training/english/arset-agricultural-crop-classification-synthetic-aperture-radar-and
- ESA's EO4Society Website:
  - https://eo4society.esa.int/training-education/
- Twitter: @EOOpenScience

Thank you!