Building Capacity to Use Earth Observations in Addressing Environmental Challenges in Bhutan

Day 3 – Land Cover Mapping and Monitoring

Land Cover Mapping and Monitoring – Urban Growth

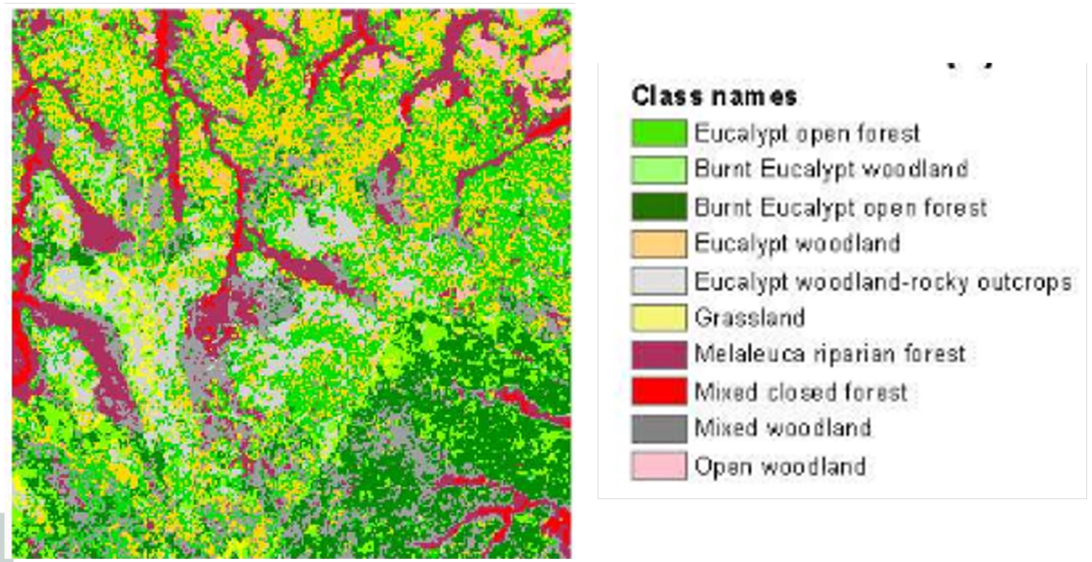
Land Cover Classification Overview

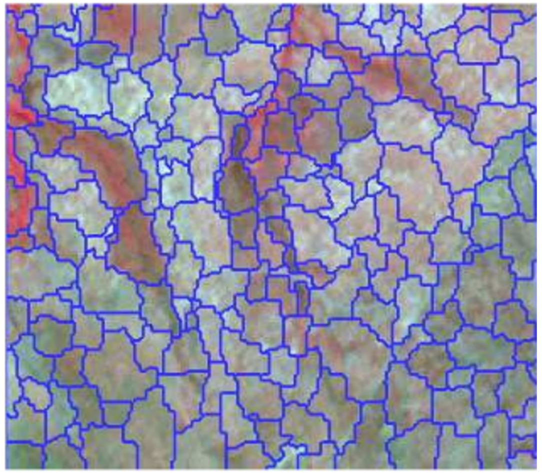
- **Land cover classification** is the process of grouping spectral classes and assigning them informational class names.
- Spectral Classes:
  – Groups of pixels that are uniform with respect to their pixel values in several spectral bands.
- **Informational Classes:**
  – Categories of interest to users of the data (like water, forest, urban, agriculture, etc.).

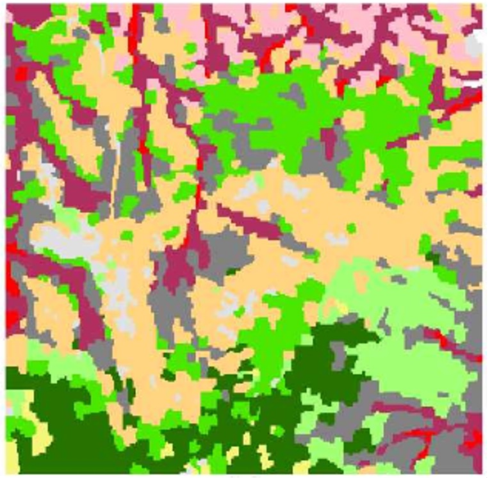
Classic example of a land cover assessment from the USGS National Land Cover Database. Image Credit: USGS

Image Classification

- **Pixel-Based**
  - Each pixel is grouped in a class
  - Useful for multiple changes in land cover within a short period of time
  - Best for complete data coverage and ensuring time series consistency at the pixel level
- **Object-Based**
  - Pixels with common spectral characteristics are first grouped together (segmentation)
  - Useful for reducing speckle noise in radar images and high-resolution imagery

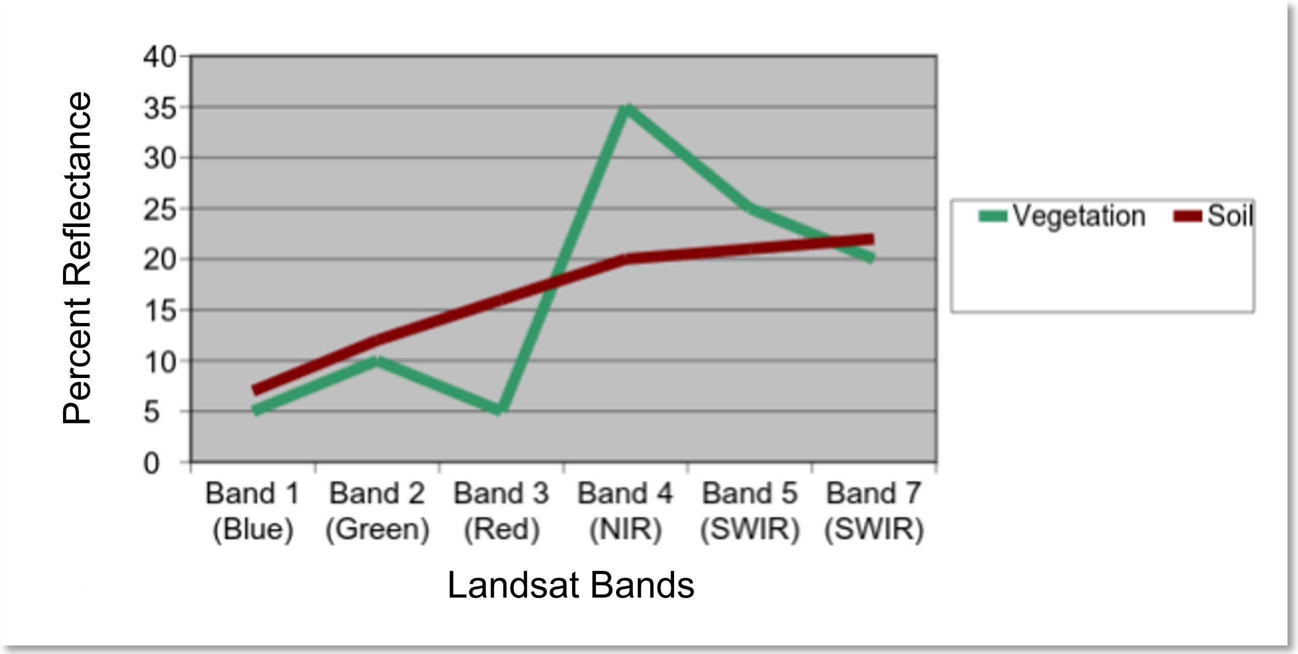






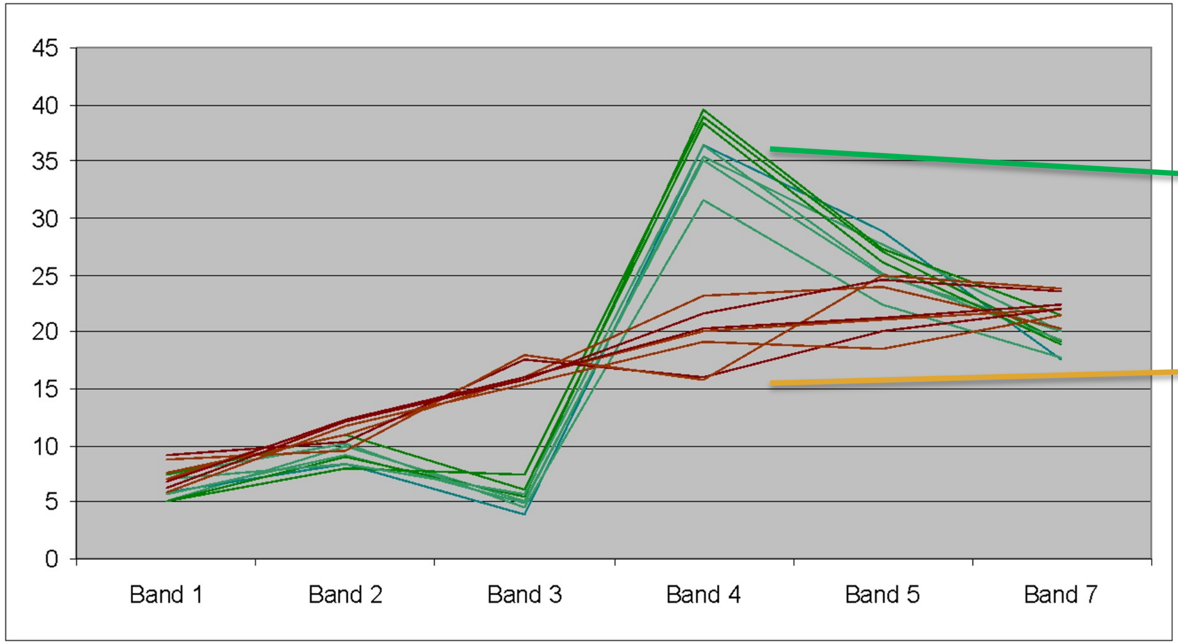
Pixel-Based Classification

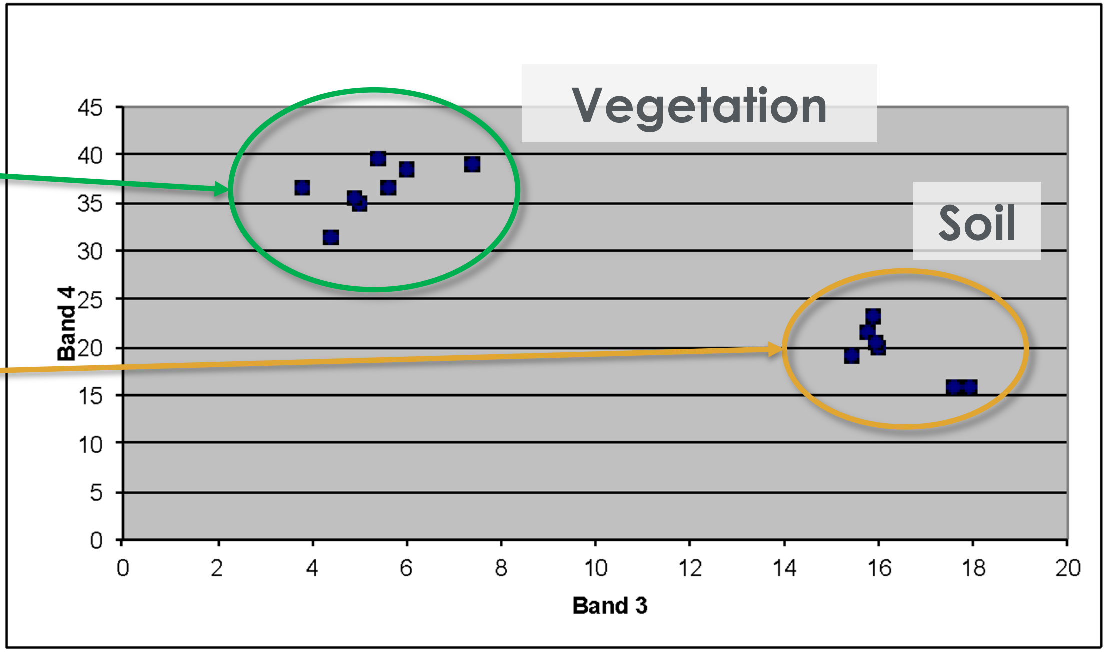
- Pixel-based classification uses the spectral information represented by the digital numbers in sensor data spectral bands and attempts to classify each individual pixel based on this spectral information.
- Spectral Signature:
  - Objects on the ground reflect electromagnetic radiation differently in different wavelengths.
- Example: Green vegetation absorbs red wavelengths but reflects near-infrared (NIR) wavelengths.



Spectral Variation

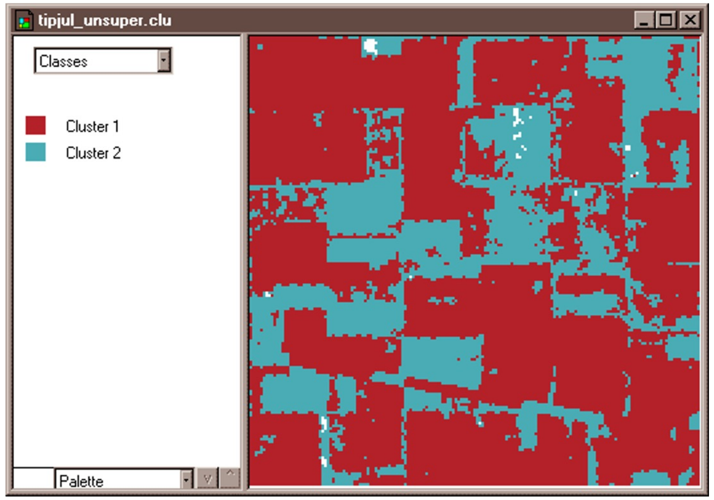
- It’s easier to distinguish **between** broad classes.
  - E.g., forest, agriculture, bare soil
- It’s more difficult to distinguish **within** broad classes.
  - E.g., vegetation species



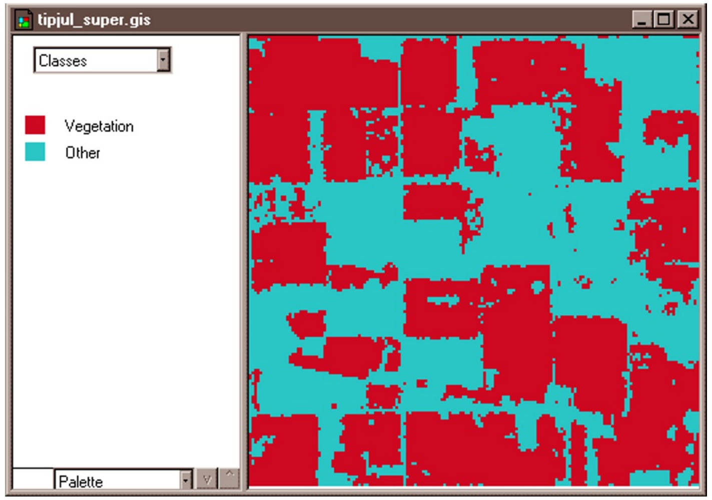


Types of Land Cover Classifications

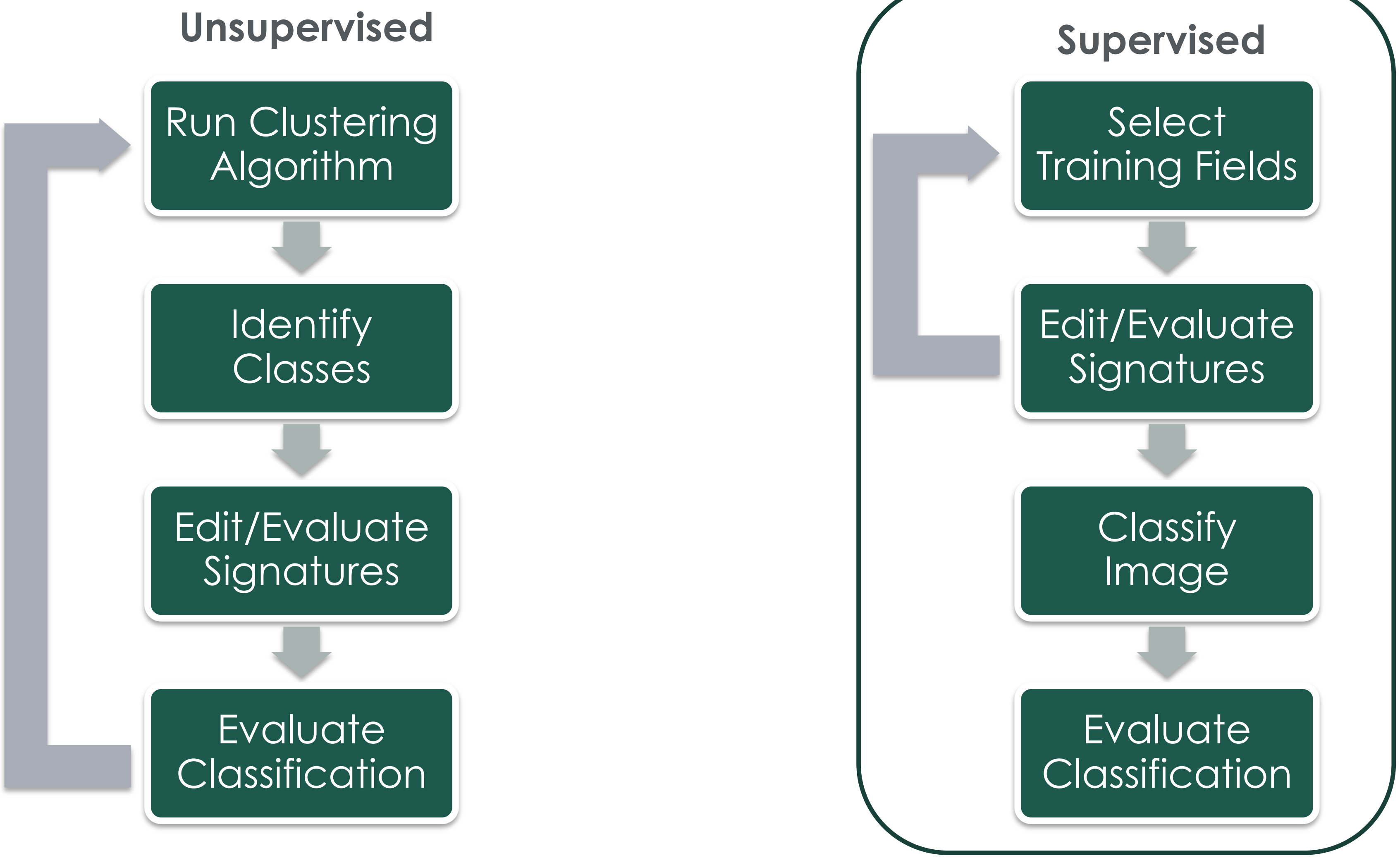
- **Unsupervised Classification**
  - Uses classification algorithms to assign pixels into one of a number of user-specified class groupings.
  - Interpreters assign each of the groupings of pixels a value corresponding to a land cover class.
- **Supervised Classification**
  - Uses expert-defined areas of known vegetation types (training areas) to tune parameters of classification algorithms.
  - The algorithm then automatically identifies and labels areas similar to the training data.



Simple example of supervised and unsupervised classifications. Supervised classifications assign a class through the use of training data. Unsupervised classifications cluster similar pixels for later classification. Image Credits: [PSU Department of Geography]

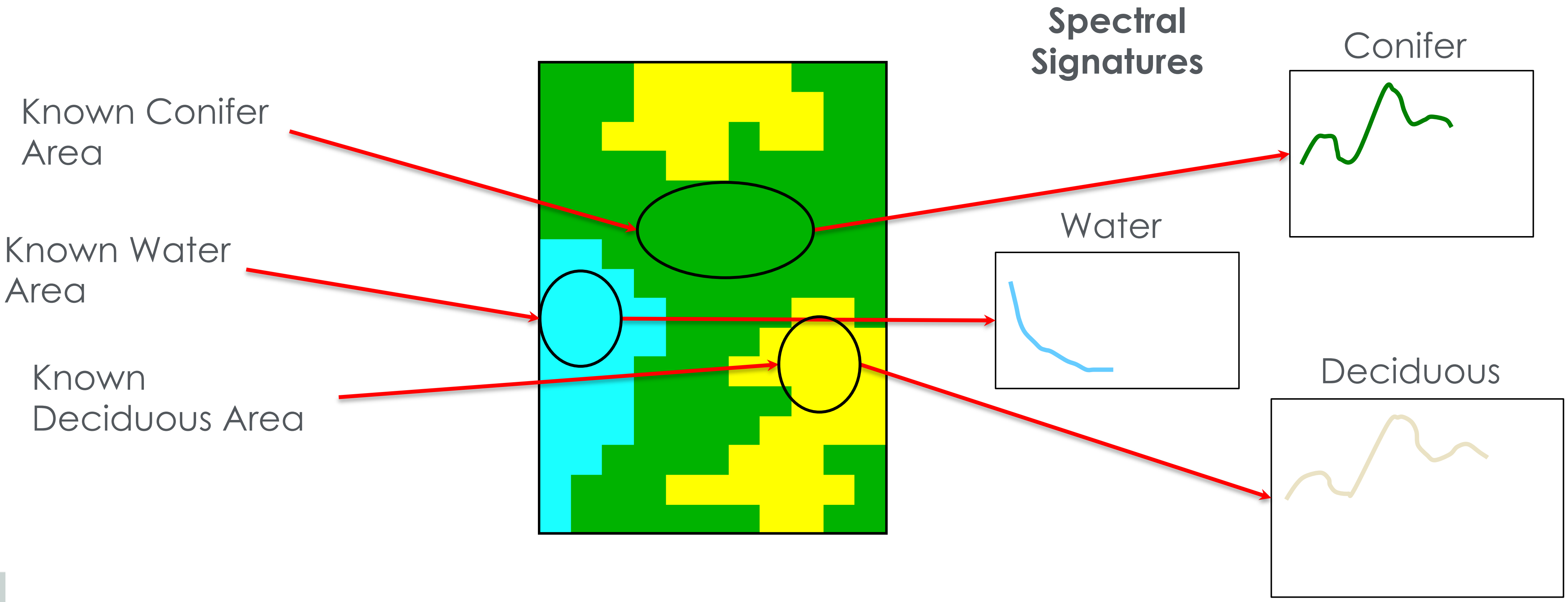


Supervised vs. Unsupervised Classification Workflow



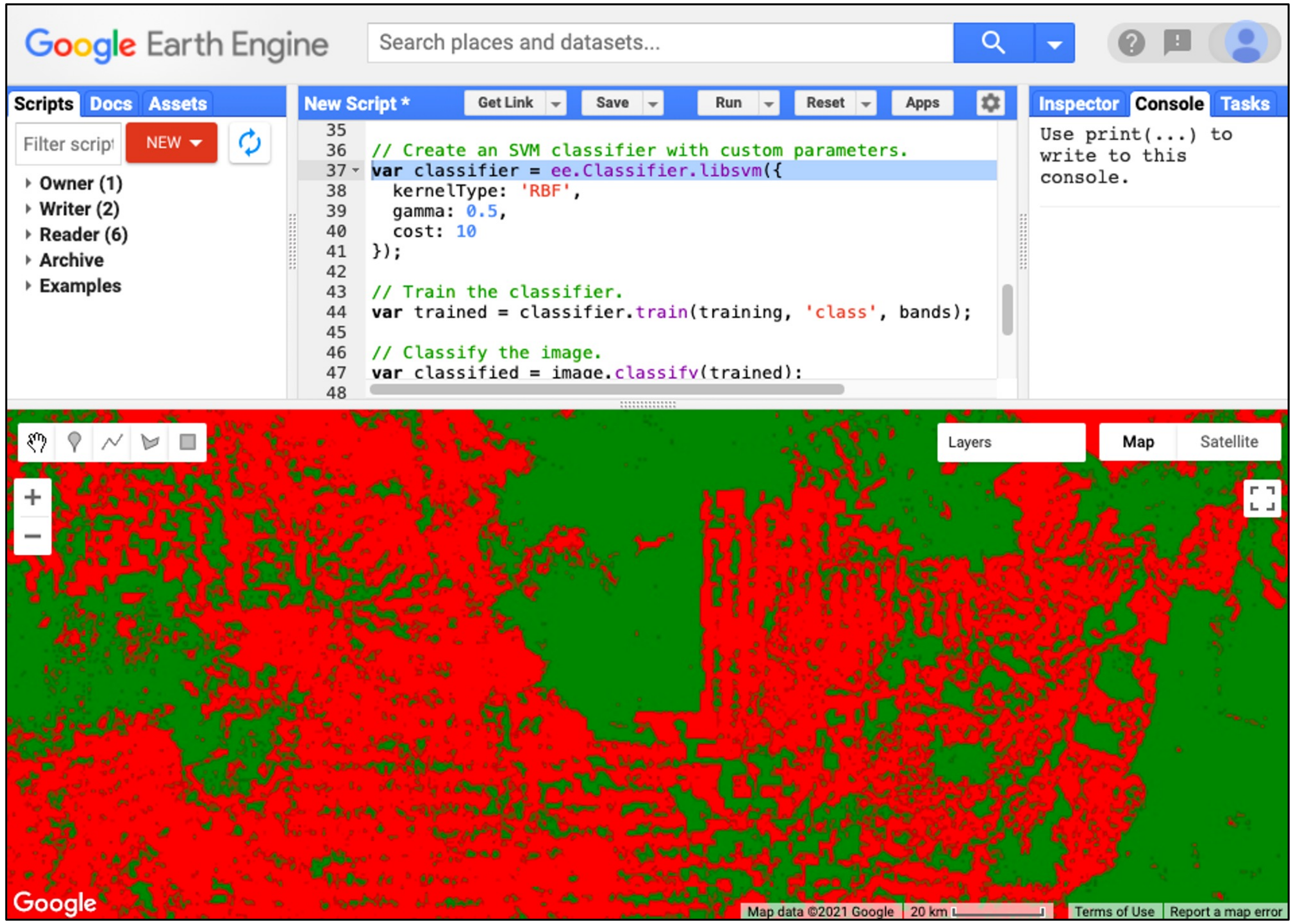
Training a Supervised Land Classification

- Supervised classification requires the analyst to select training areas or points where the land cover class on the ground is known.
- These reference points are used by an algorithm to classify remaining pixels.



Supervised Classifications Algorithms Available in GEE

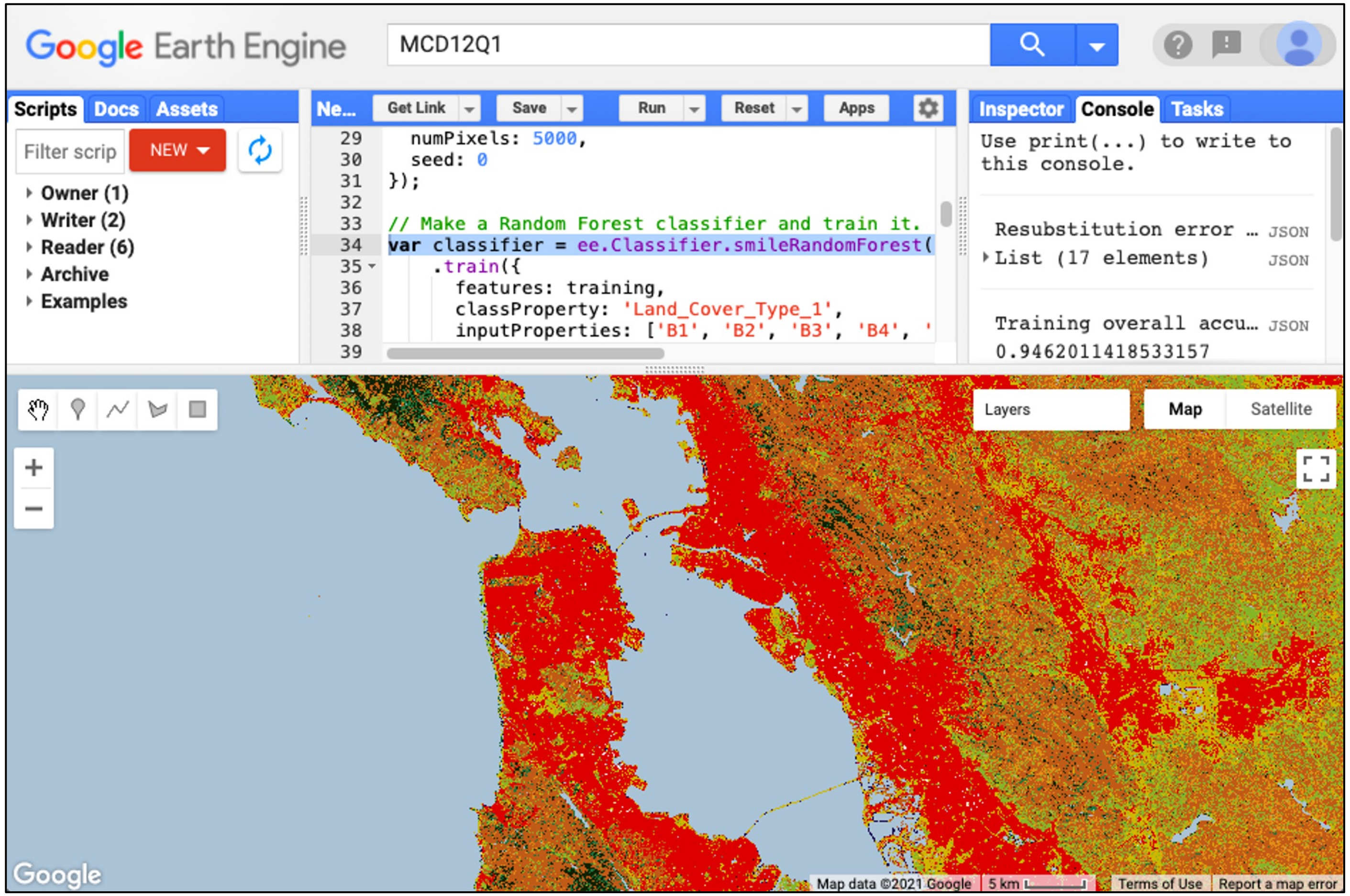
- Classification and Regression Trees (CART)
  - ee.Classifier.smileCart
- Naive Bayes
  - ee.Classifier.smileNaiveBayes
- Support Vector Machine (SVM)
  - ee.Classifier.libsvm
- RandomForest
  - ee.Classifier.smileRandomForest
- In this session we will be using the **RandomForest** algorithm.



Simple Support Vector Machine (SVM) classifier example in the GEE code editor mapping deforestation in red. Credit: GEE Developers

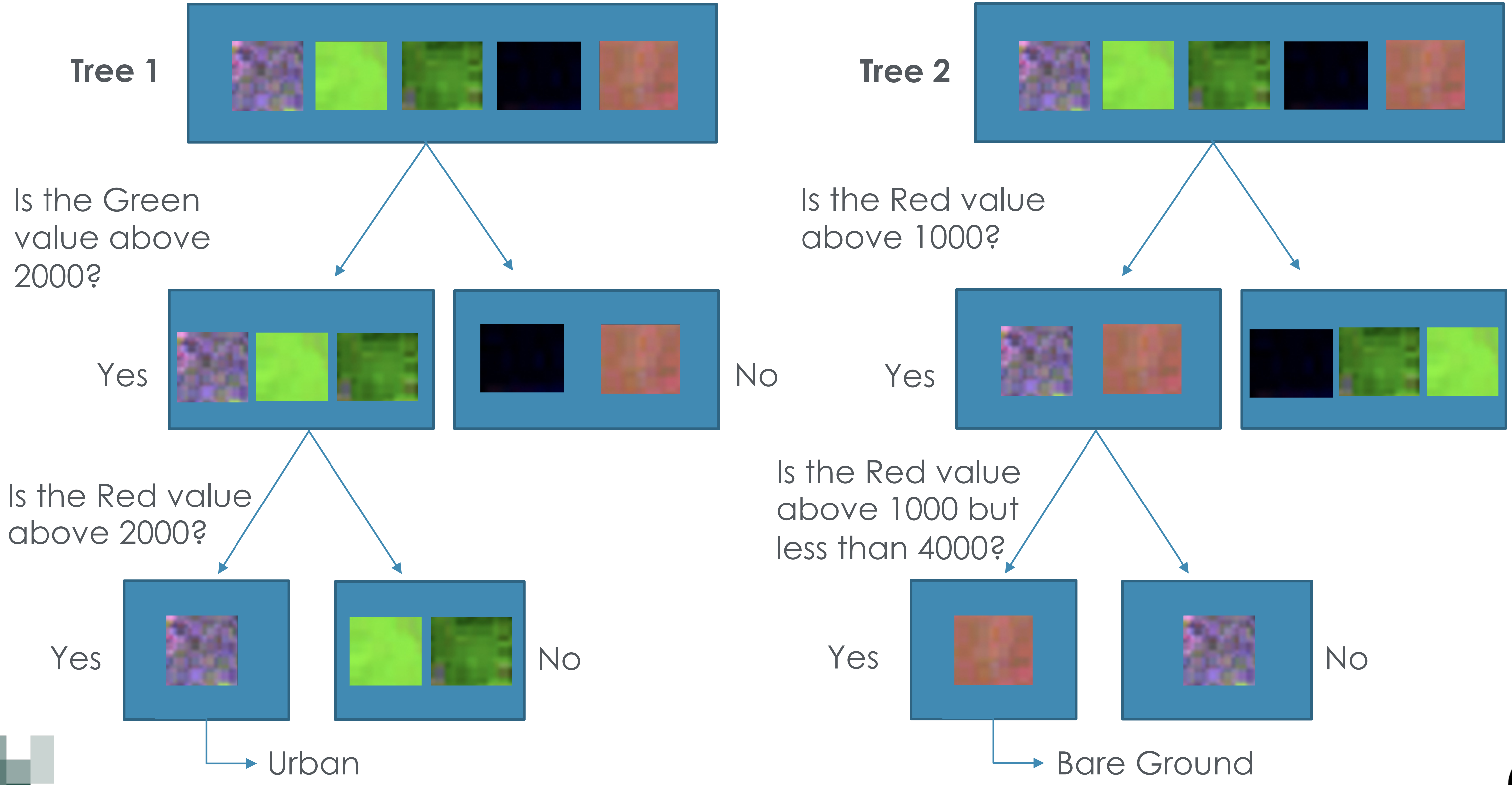
Random Forest Classification

- Learns from training data and identifies statistical patterns in large datasets
- Tree-based machine learning algorithm
  - Uses a series of decision trees to select the best classification for all pixels within imagery
  - Iterative use of decision trees allows algorithm to "vote" for the best solution

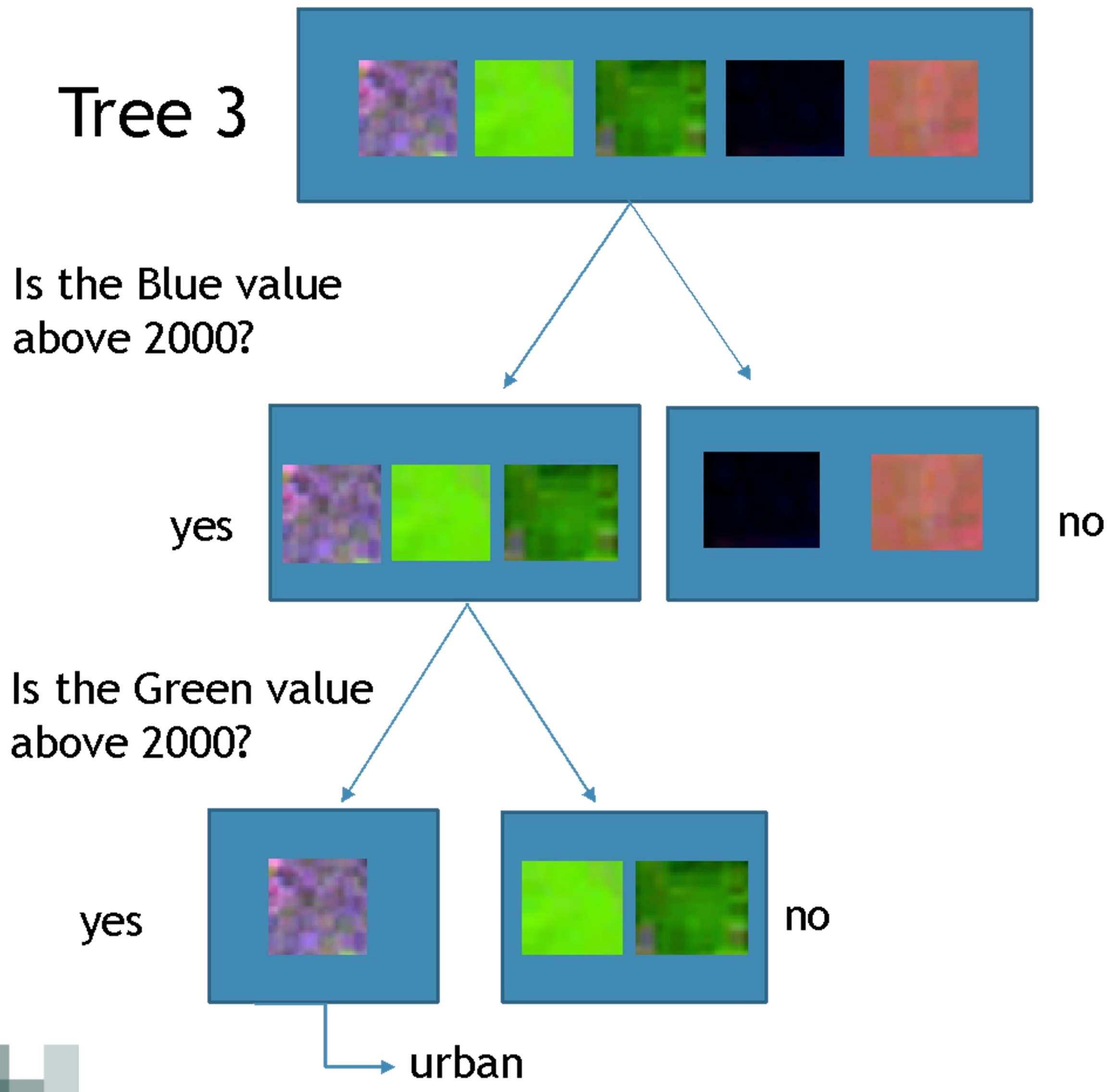


Simple random forest classification completed for the San Francisco Bay Area of California in the GEE interface. Credit: [GEE Developers]

Random Forest Classification



Random Forest Classification



Tree 1 = Urban
Tree 2 = Bare Ground
Tree 3 = Urban

= **Urban**

Advantages and Limitations of the Random Forest Algorithm

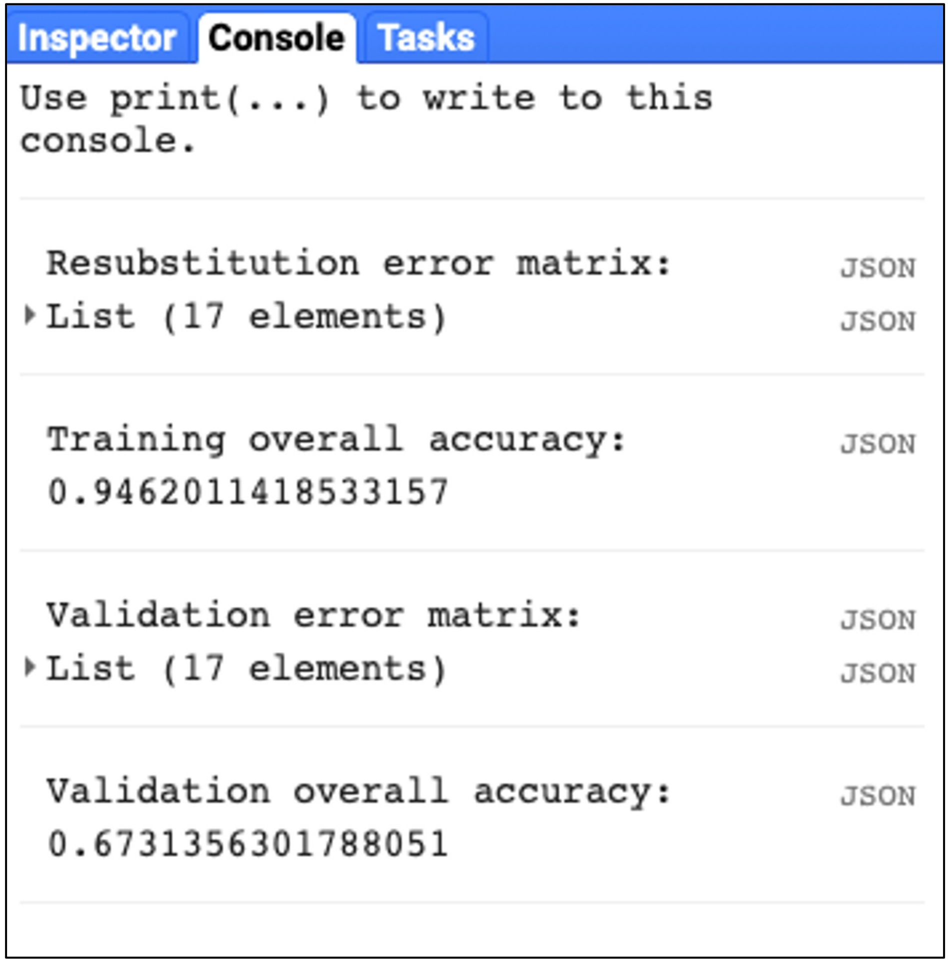
- **Advantages:**
  - Use of multiple trees reduces the risk of overfitting
  - Training time is shorter and not sensitive to outliers in training data
  - Runs efficiently and produces high accuracy for large datasets
  - Easy to parameterize
  - Easy to determine feature importance (Random forest makes it easy to evaluate variable importance, or contribution, to the model)

- **Limitations:**
  - Algorithm cannot predict spectral range beyond training data
  - Training data must capture the entire spectral range
  - More complex (The prediction of a single decision tree is easier to interpret when compared to a forest of them)

What is random forest? – IBM

Accuracy Assessment

- Accuracy refers to the degree of correspondence between classification and reality.
- Accuracy assessment is the process by which the accuracy or correctness of an image classification is evaluated.
- This involves the comparison of the image classification to **reference data** that are assumed to be true.
  - References can include ground data, or a subset of training points withheld for accuracy assessment purposes.



Example of error matrix display in the GEE console for accuracy assessment. Credit: GEE Developers

Determining Classification Accuracy

- Agreement between reference data and algorithm classifications indicates the classifier performed well.
- If there is not agreement, the classifier has incorrectly measured the land class resulting in error.
- Comparison of reference data and classifications is typically done using a confusion (or error) matrix to compile these comparisons.

Plot ID	Class in Reference Source	Class in Classification Map	Agreement
1	Urban	Urban	Yes
2	Bare Ground	Urban	No
3	Forest	Forest	Yes
4	Forest	Agriculture	No
5	….	….	….

Error Matrix

- Table of reference classes to predicted classes:
  - Reference classes are assumed to be correct (columns).
  - Mapped classes are the output of the classification (rows).

Mapped Classes	Classification	Reference Classes				
		Urban	Agriculture	Forest	Bare Ground	Row Total
	Urban	45	4	12	24	85
	Agriculture	6	91	5	8	110
	Forest	0	8	55	9	72
	Bare Ground	4	7	3	55	69
	Column Total	55	110	75	96	**336**

The number of correctly classified pixels is shown along the diagonal.

Error Matrix

- Table of reference classes to predicted classes:
  - Reference classes are assumed to be correct (columns).
  - Mapped classes are the output of the classification (rows).

	Reference Classes				
Classification	**Urban**	**Agriculture**	**Forest**	**Bare Ground**	**Row Total**
Urban	45	4	12	24	85
Agriculture	6	91	5	8	110
Forest	0	8	55	9	72
Bare Ground	4	7	3	55	69
Column Total	55	110	75	96	**336**

Mapped Classes

Column totals are the total number of reference pixels in each class.

Error Matrix

- Table of reference classes to predicted classes:
  - Reference classes are assumed to be correct (columns).
  - Mapped classes are the output of the classification (rows).

Classification	Reference Classes				
	Urban	Agriculture	Forest	Bare Ground	Row Total
Urban	45	4	12	24	85
Agriculture	6	91	5	8	110
Forest	0	8	55	9	72
Bare Ground	4	7	3	55	69
Column Total	55	110	75	96	**336**

Mapped Classes

Row totals are the total number of classified pixels in each class.

Error Matrix

- Table of reference classes to predicted classes:
  - Reference classes are assumed to be correct (columns).
  - Mapped classes are the output of the classification (rows).

		Reference Classes			
Classification	**Urban**	**Agriculture**	**Forest**	**Bare Ground**	**Row Total**
Urban	45	4	12	24	85
Agriculture	6	91	5	8	110
Forest	0	8	55	9	72
Bare Ground	4	7	3	55	69
Column Total	55	110	75	96	**336**

Mapped Classes

Off-diagonal numbers represent errors of commission and omission.

Map Accuracy

$$\text{Overall Accuracy} = \frac{\text{Number of Correctly Classified Pixels (Sum of Diagonal)}}{\text{Number of Total Sampled Pixels}}$$

$$\text{Overall Accuracy} = \frac{45+91+55+55}{336} * 100 = \mathbf{73\%}$$

Mapped Classes		Reference Classes				
	Classification	**Urban**	**Agriculture**	**Forest**	**Bare Ground**	**Row Total**
	Urban	45	4	12	24	85
	Agriculture	6	91	5	8	110
	Forest	0	8	55	9	72
	Bare Ground	4	7	3	55	69
	Column Total	55	110	75	96	**336**

Map accuracy is defined by the percent of correctly classified pixels.

Kappa Statistic

- Considers the possibility of the agreement occurring by chance
- Proportion of agreement after chance agreement has been removed
- Calculated from an error matrix as an additional accuracy check
- A higher kappa value means higher accuracy
- Does not usually provide more information about accuracy than the error matrix overall accuracy calculation

Kappa	Interpretation
< 0	No agreement
0.0 - 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Image Credit: Okwuashi et al. 2012

Kappa Statistic

$$k = \frac{p_0 - p_c}{1 - p_c}$$

p_0= Observed accuracy. $\Sigma\ p_{ii}$ is the sum of relative frequency in the diagonal of the error matrix.

p_c= Chance agreement. $\Sigma\ p_{i+} * p_{+i}$ is the relative frequency of a random allocation of observations to the cells of the error matrix.

Calculation of Kappa

Classification	Reference Classes				
	Urban	Agriculture	Forest	Bare ground	Row total
Urban	45	4	12	24	85
Agriculture	6	91	5	8	110
Forest	0	8	55	9	72
Bare ground	4	7	3	55	69
Column total	55	110	75	96	**336**

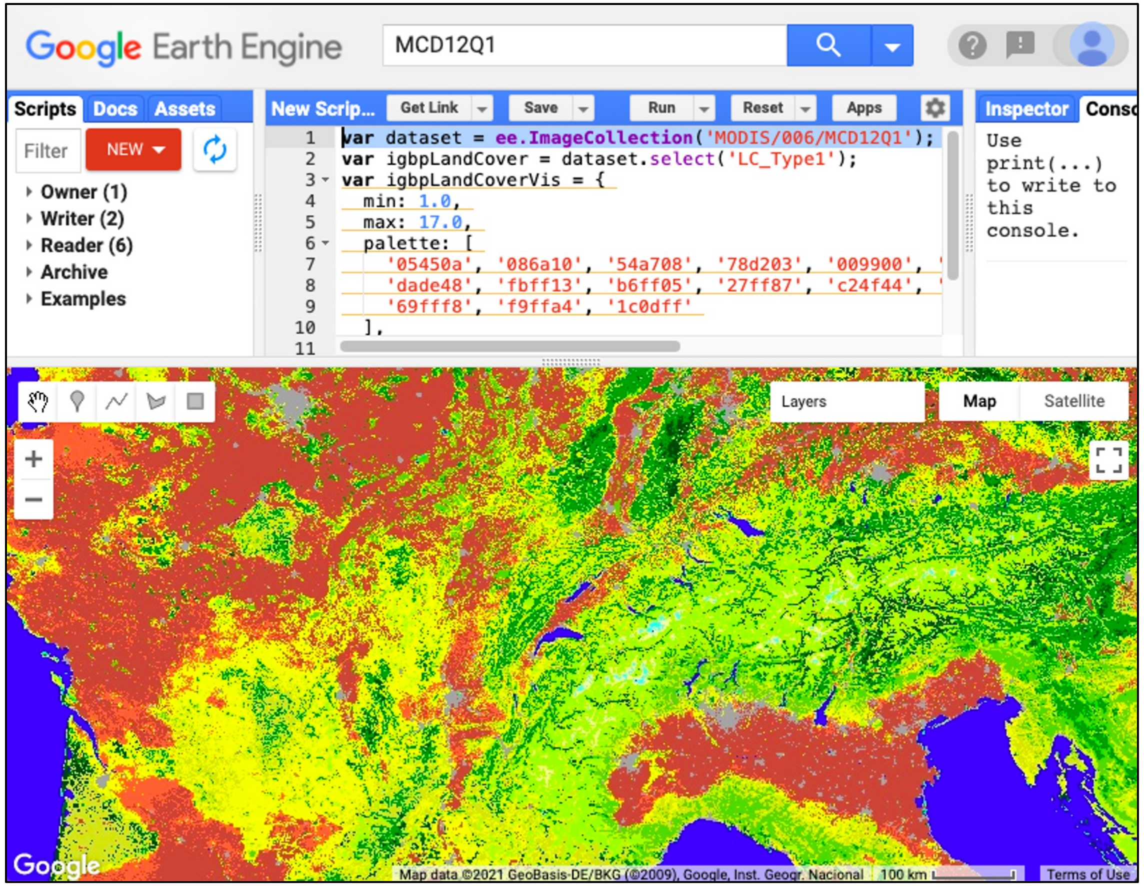
$p_0 = (45+91+55+55)/336 = 0.7321$

$$p_c = \frac{(85*55)/336 + (110*110)/336 + (72*75)/336 + (69*96)/336 = 0.2551}{336}$$

$$k = \frac{0.7321 - 0.2551}{1-0.2551} = \mathbf{0.64}$$

Accuracy Assessment in GEE: Confusion/Error Matrix

- **Confusion/Error Matrix:**
  - Describes how well the classifier was able to correctly label training data the classifier has already seen.
  - Can also compare predicted values to actual values the classifier has not seen.
  - A subset of training data can be withheld as validation data and used to test the ability of the trained classifier to accurately predict land cover class.
  - Other land cover products like MODIS land cover product (MCD12Q1) can be used as validation data.



MODIS annual, global land cover data product that is mentioned in the GEE developer repository as a source of points for accuracy assessment of your own land classification. Credit: GEE Developers

Previous ARSET Trainings

- Past ARSET trainings relevant to land classification and accuracy assessment:
  - Land Cover Classification with Satellite Imagery
  - Accuracy Assessment of a Land Cover Classification
  - Remote Sensing for Mangroves in Support of the UN Sustainable Development Goals

GEE Developer Guides

- Relevant guides to GEE features and JavaScript Code:
  - FeatureCollection Overview
  - Compositing, Masking, and Mosaicking
  - Machine Learning in Earth Engine
  - Supervised Classification (including accuracy assessment)

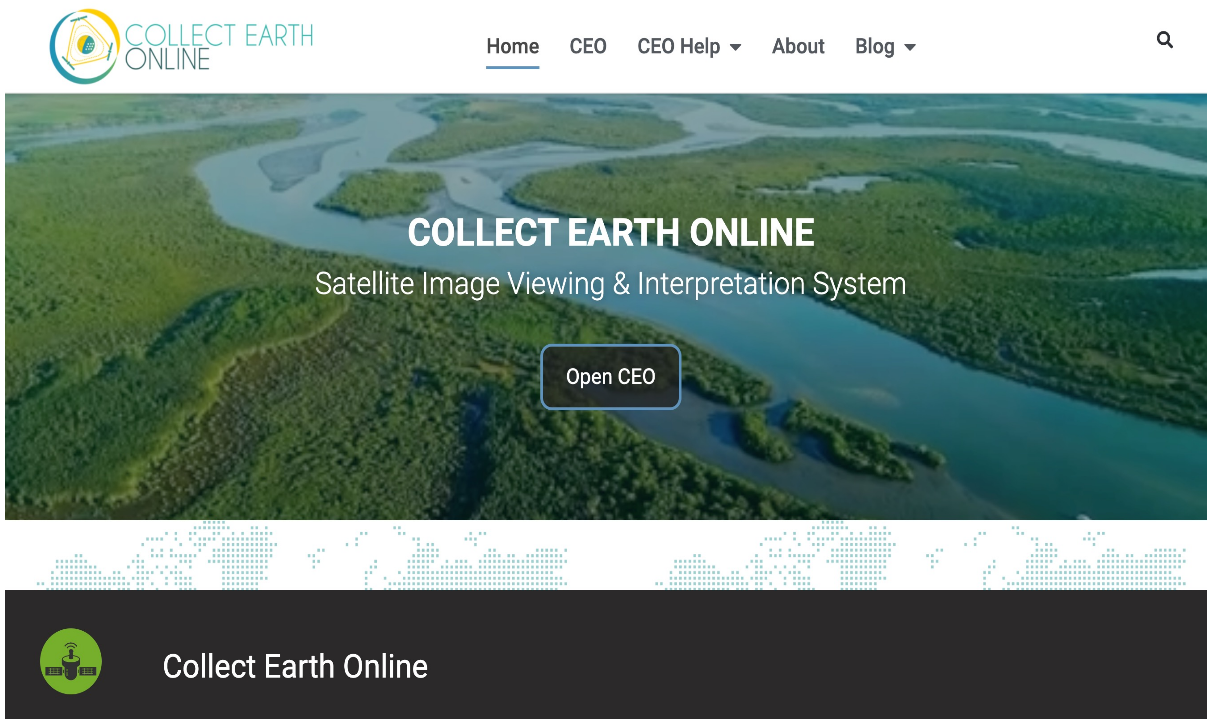
- The full list of guides and tutorials made available by the developers:
  - JavaScript and Python Guides

Image Credit: Google Earth Engine

Collect Earth Online (CEO)

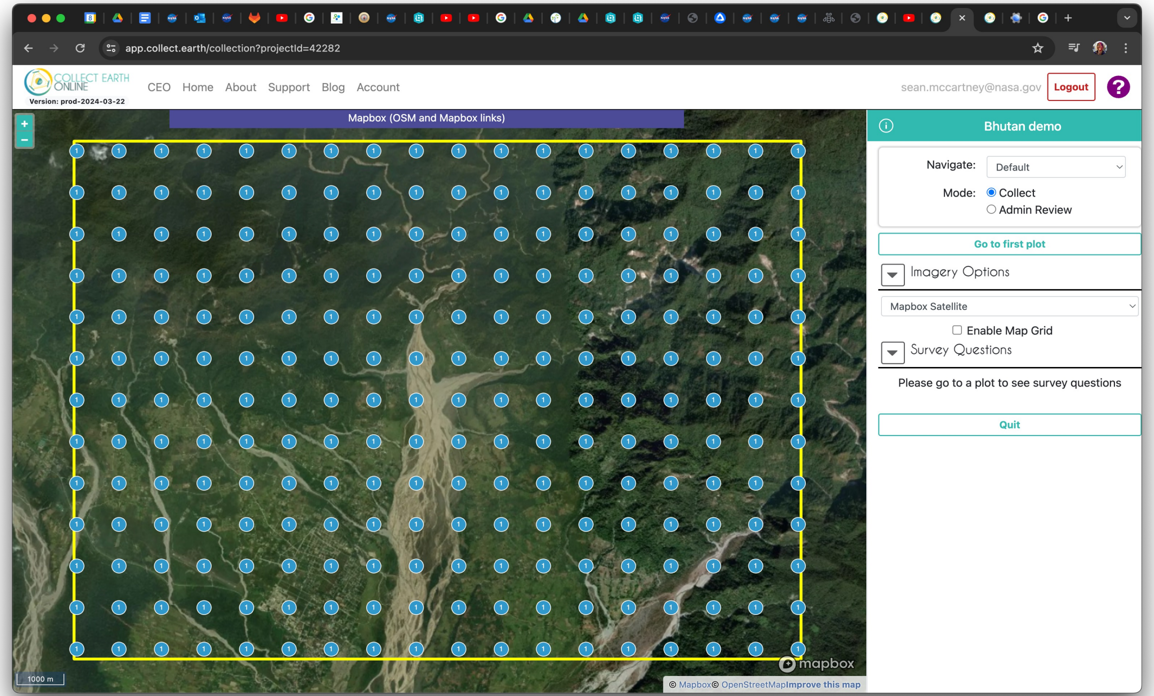
Collect Earth Online (CEO)

- Free, open-source, online system for viewing and interpreting high-resolution satellite imagery.
- Used to detect deforestation and other changes in the landscape.
- Enables multiple users to simultaneously contribute to labeling land cover types.
- Access to multiple sources of satellite imagery (medium [10–30m] to high [<5m] spatial resolution).
- Developed by the Forestry Department of the Food and Agriculture Organization of the United Nations (FAO) and partners.



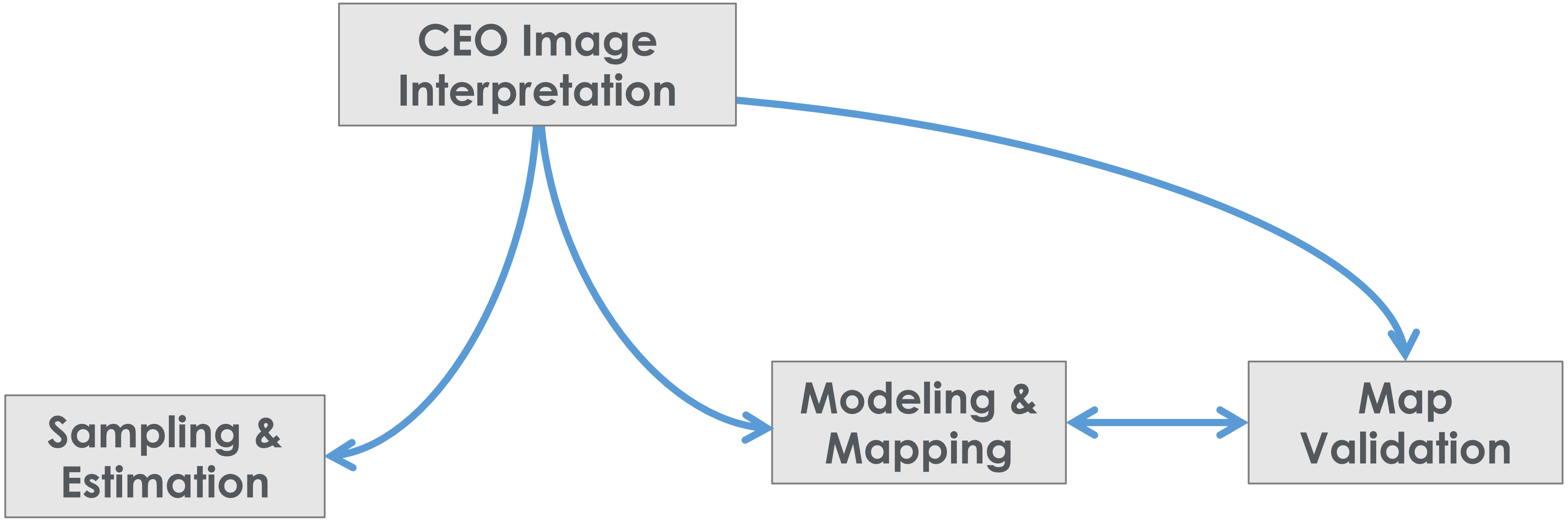
Collect Earth Online (CEO)

- Promotes consistency in locating, interpreting, and labeling reference data.
- For any mapping effort, gathering reference data is essential.
- Relying on local experts to interpret satellite imagery offers a cost-effective way to assess map accuracy, eliminating the need for time-consuming data collection in the field.
- Users can collect data using high-resolution satellite imagery and analyze it with Google Earth Engine.



CEO Uses & Planning

Can be used in multiple ways to support estimation, mapping, and/or validation.



CEO Uses & Planning - Example Applications

- Point-based estimates
- Resulting in statistical summaries
- Single interest LULC
- Estimate or map of target LULC type (e.g., canopy cover)
- Full LULC with complex classification
- LULC area estimates or support creation of a wall-to-wall LULC map
- Map validation

All applications could be done in a single year or multi-year manner to evaluate change or trends…

CEO Sample Design – Determining Sample Size

- **Sample Size:**
  - The number of examples that must be interpreted
  - E.g., the number of plots in CEO for interpretation

- Sample size can have a huge influence on the success of the project!
  - If sample size is too small, land cover types may not all get sampled representatively.
  - If sample size is too large, it may exhaust your project resources.

- This is an extensive topic, and you must do your homework!
  - Different approaches and considerations for different projects
  - More Information:
    - Foody, G. M. (2009). Sample size determination for image classification accuracy assessment and comparison. International Journal of Remote Sensing, 30(20), 5273–5291. http://doi.org/10.1080/01431160903130937

Collect Earth Online (CEO) – Resources

- Website: https://www.collect.earth/
- CEO User Guides: https://www.collect.earth/ceo-guides/
- CEO Lectures & Recordings
- Application: https://app.collect.earth/
- Data Collection Manual
- The code is available under an MIT license on GitHub.

Demonstration:
Collect Earth Online

Demonstration:

Land Classification and Accuracy Assessment in Google Earth Engine

Summary

- Land classification in GEE is a valuable tool for mapping and monitoring land cover.
- To complete a supervised land classification, the user must first establish training points of known land cover classes to train the classifier.
- CEO is a powerful tool for locating, interpreting, and labeling reference data.
- Image interpretation is a complex skill – understanding imagery characteristics is important for selecting appropriate imagery data.
- GEE provides many machine learning classification algorithms built into the API:
  - Such as Classification and Regression Trees, Naïve Bayes, Support Vector Machine, and RandomForest.
- Random Forest is a tree-based machine learning algorithm that uses a series of decision trees to select the best classification for all pixels within imagery.
- Simple accuracy assessment in GEE can be completed using confusion/error matrices to compare predicted classifications to withheld training data and validation data.
- A well-defined QA/QC plan is necessary to ensure data quality