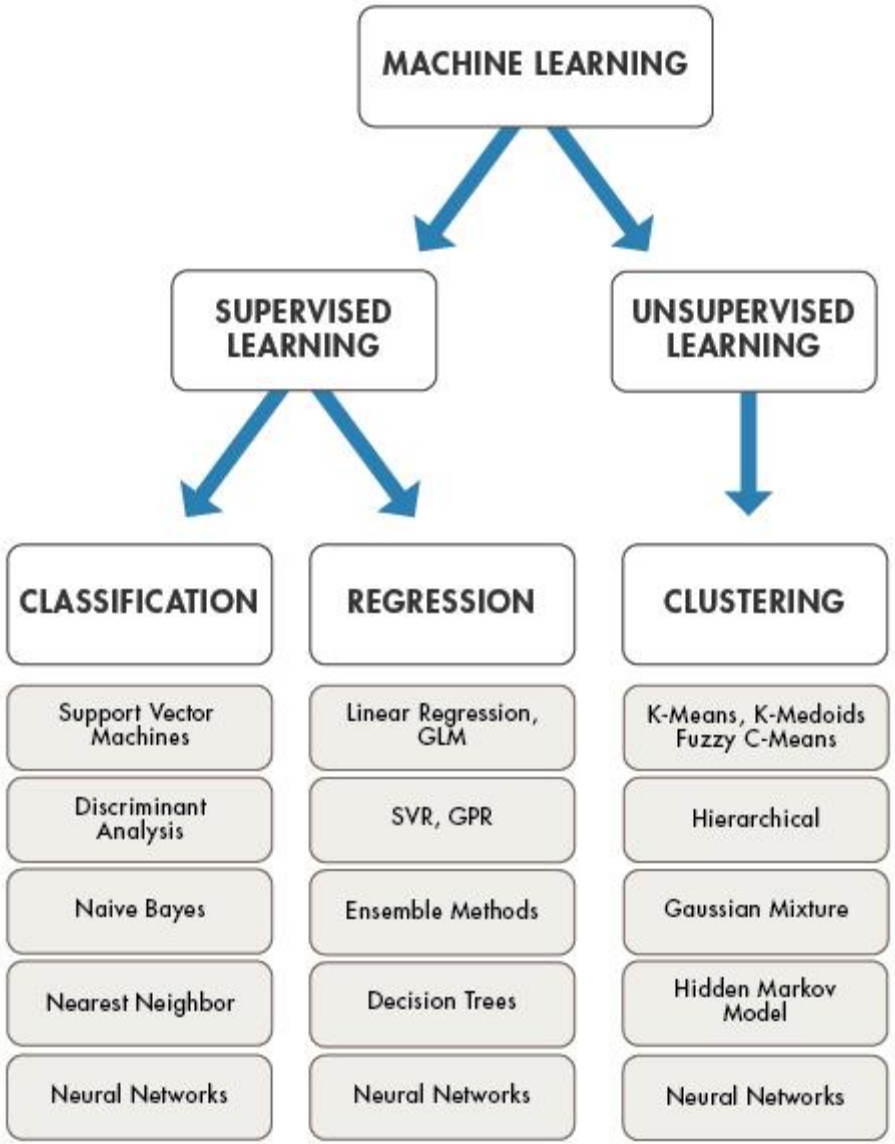
Image and Object Clustering and Classification

GEOG 457/657

Alex Bevington Feb 26, 2024

The ESRI 2020 Land Cover map, created in partnership with Impact Observatory and Microsoft.



https://www.mathworks.com/help/stats/machine-learning-in-matlab.html

Threshold classification

- Classify a single band using fixed thresholds (not machine learning!)
- Very simple, typically only for 2 classes
- OK first approximation
- Hard to manage multiple classes/bands
- Other examples:
  Veg = NDVI > 0.2
  Water = NDWI > 0

Automatic image thresholding: ‘Otsu’s method’

- Automatically separates two classes by iterating thresholds to maximise interclass variance
- E.g. Snowline within a glacier polygon, Lake area within a lake polygon buffer

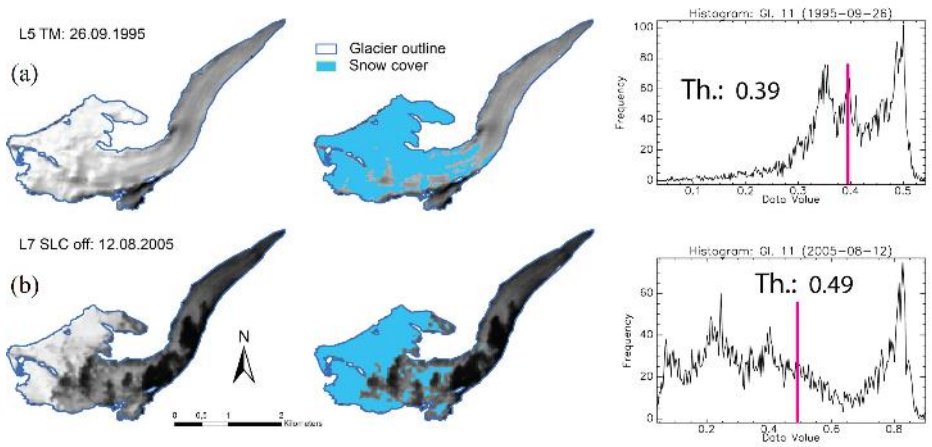
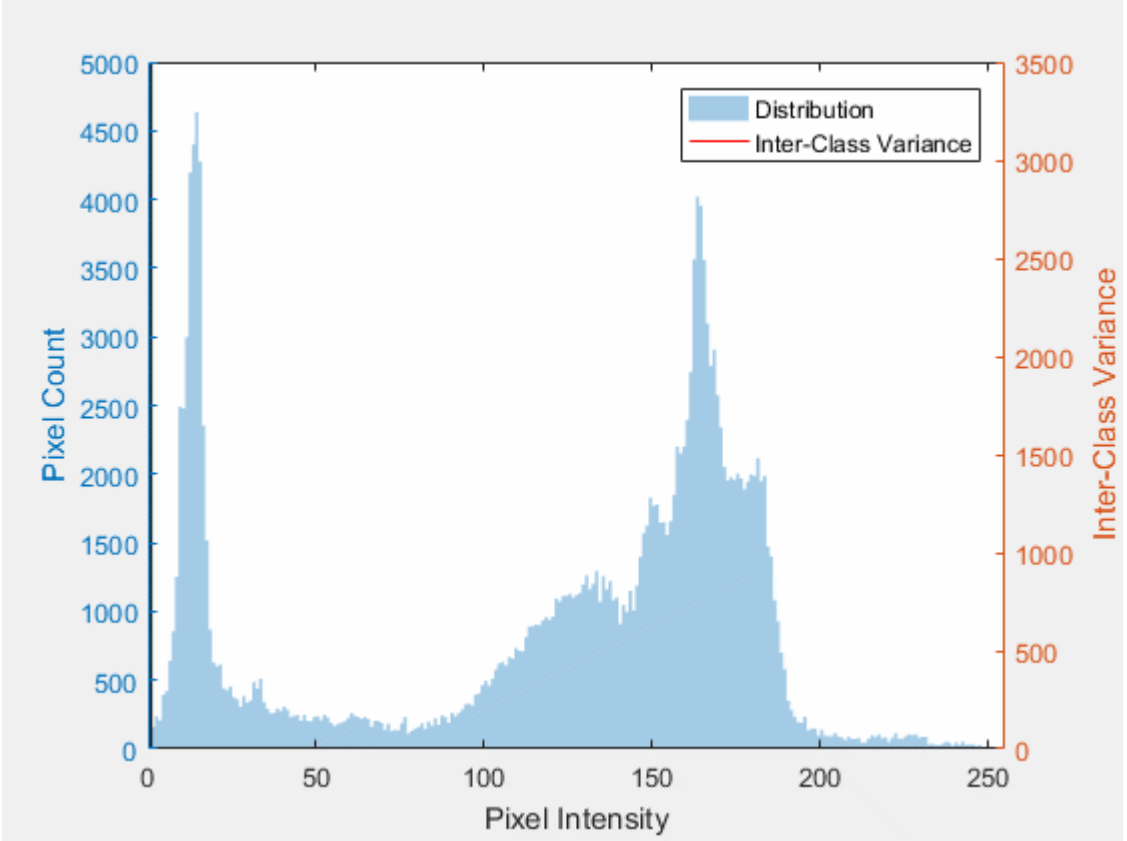


Figure 11. Snow cover mapping performance for two problematic cases. (a) HEF with fresh snow cover and its associated wrong mapping. (b) HEF with clouds and its related shadows, which also cause the algorithm to fail. Right panels show the respective histograms of the Ekstrand corrected image. A clear selection of a threshold is difficult in these cases.

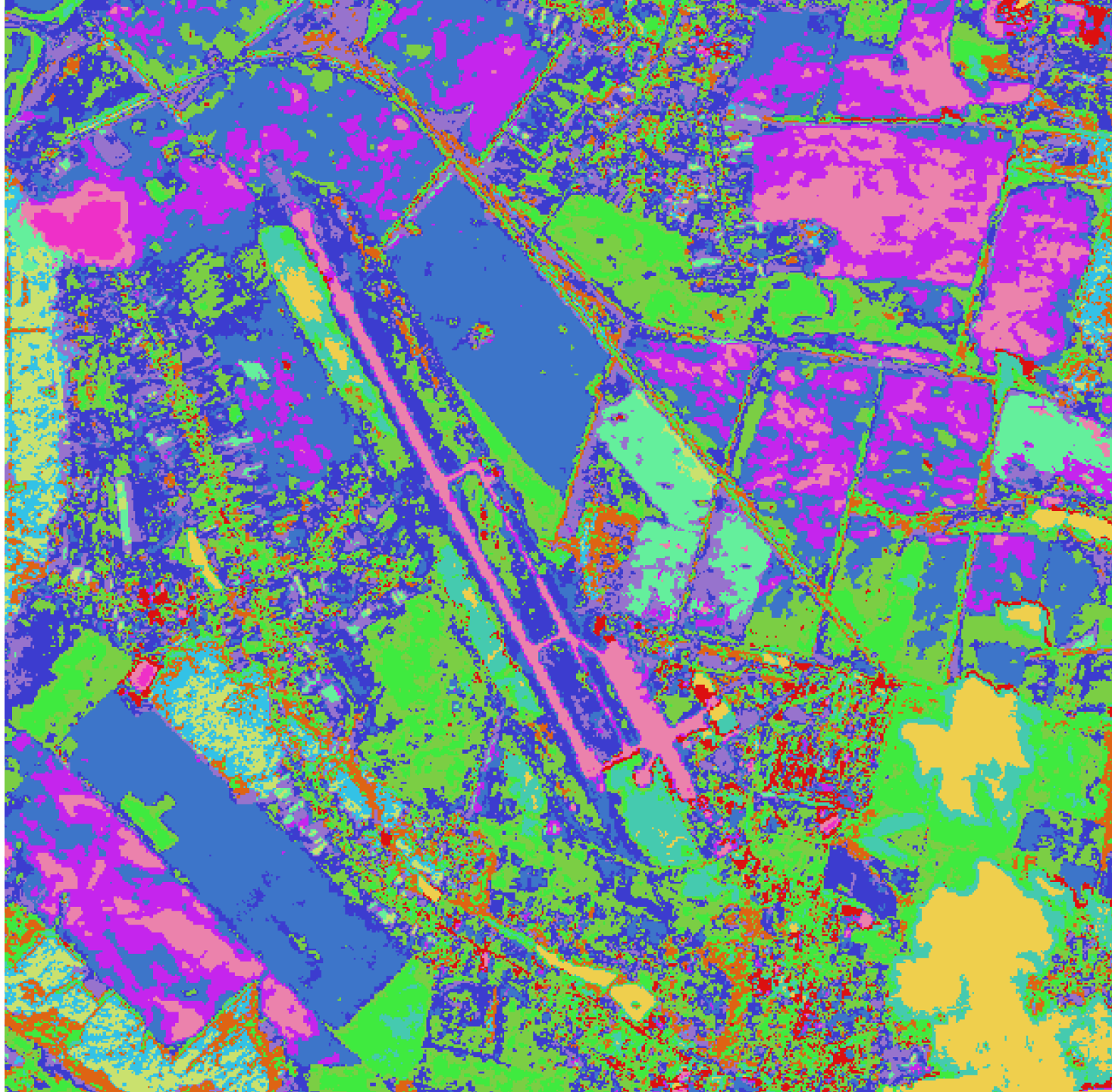
Rastner et al. 2019



Visualization of Otsu's Thresholding Algorithm, Wikipedia CC BY-SA 4.0

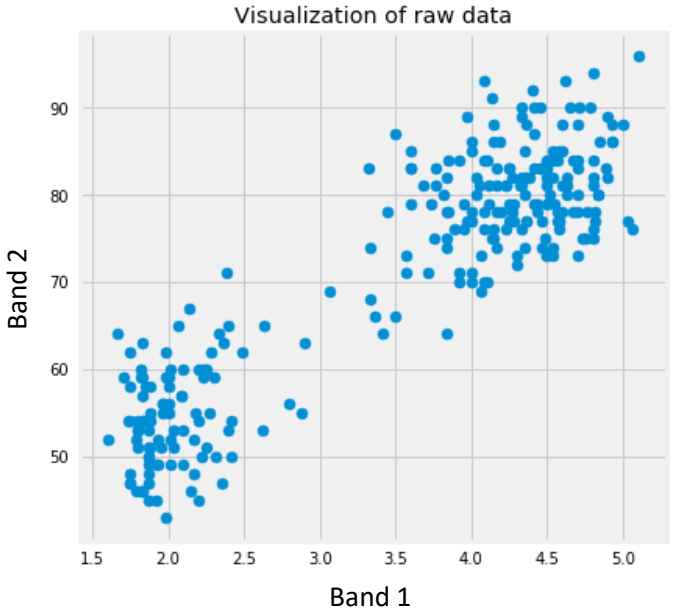
Unsupervised classification

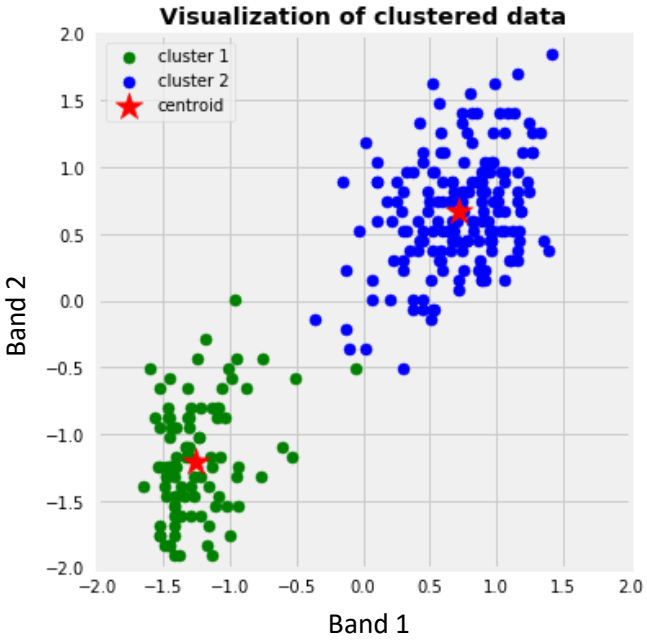
- Useful when you do not know much about your data
- User selects the number of output classes
- Manually assign class names to classes
- Common clustering algorithms
  - K-means
  - ISODATA
  - Fuzzy clustering
  - …

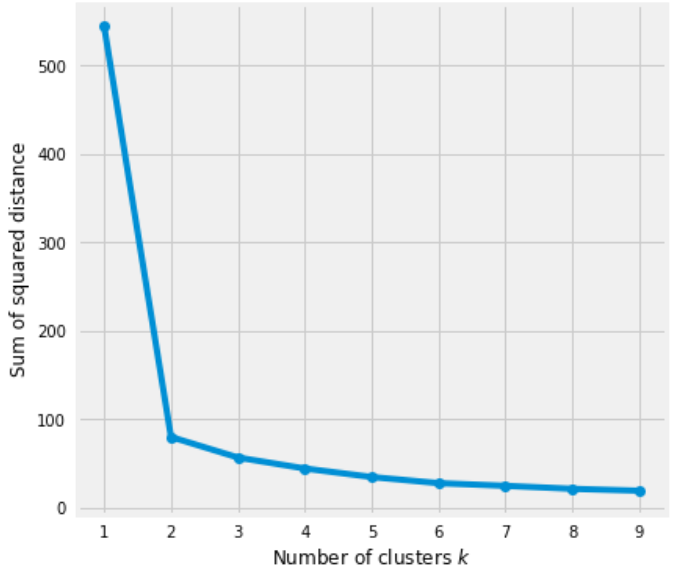


K-means

- This unsupervised algorithm is very common
- Cluster data from *n* observations (pixels) into *k* clusters (classes)
- User sets number of classes
- Iterates the algorithm until non-overlapping clusters/classes are found
- Minimizes within group variance while maximizing between group variance
- Variance calculated as the sum of squared Euclidian distance from the cluster mean for each band



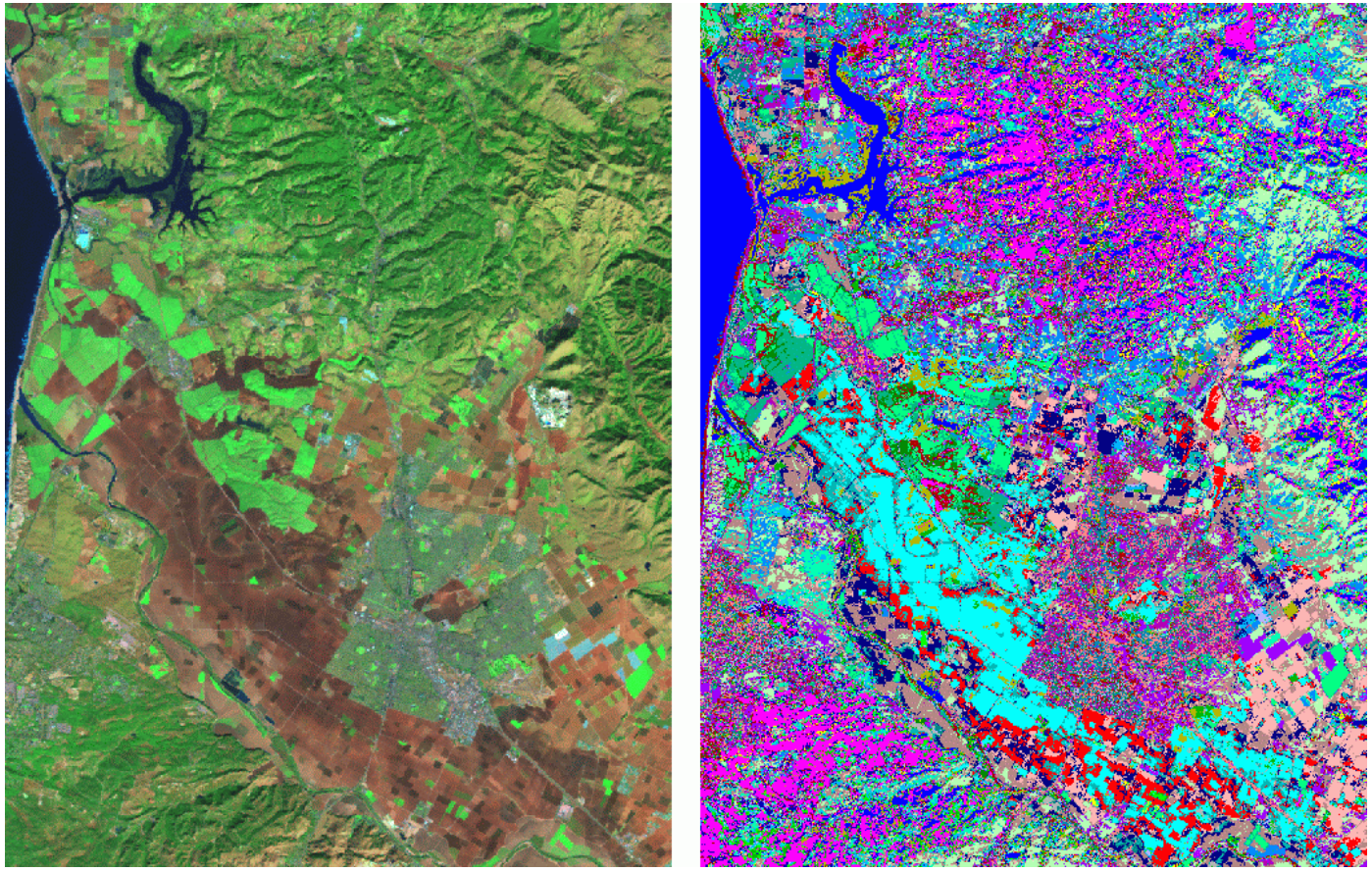




https://imaddabbura.github.io/post/kmeans-clustering/

ISODATA

- Iterative Self Organizing Data Analysis Technique
- Similar to k-means but can split and merge clusters to improve class separation
- User sets minimum number of pixels per cluster, approximate number of clusters (or range of clusters), parameters for splitting/merging classes



Unsupervised ISODATA classification of a portion of Landsat 7 scene. A display of image bands 7-4-2 as R-G-B is shown on the left, and the 30-class ISODATA class raster is shown on the right.

https://www.microimages.com/gallery/imageClassification.htm

Fuzzy C-Means (FCM)

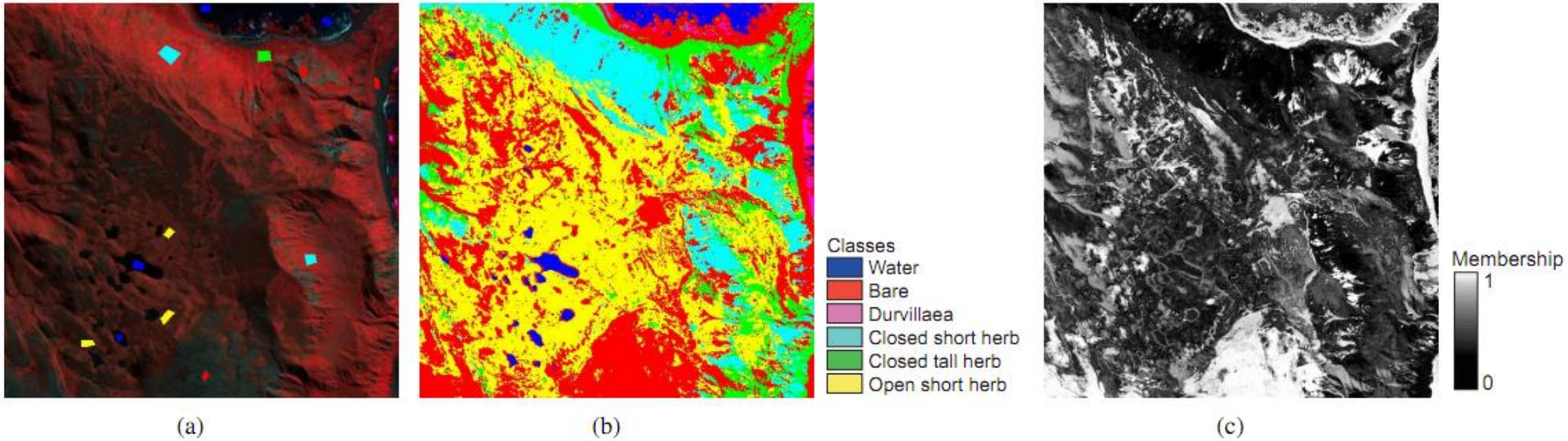


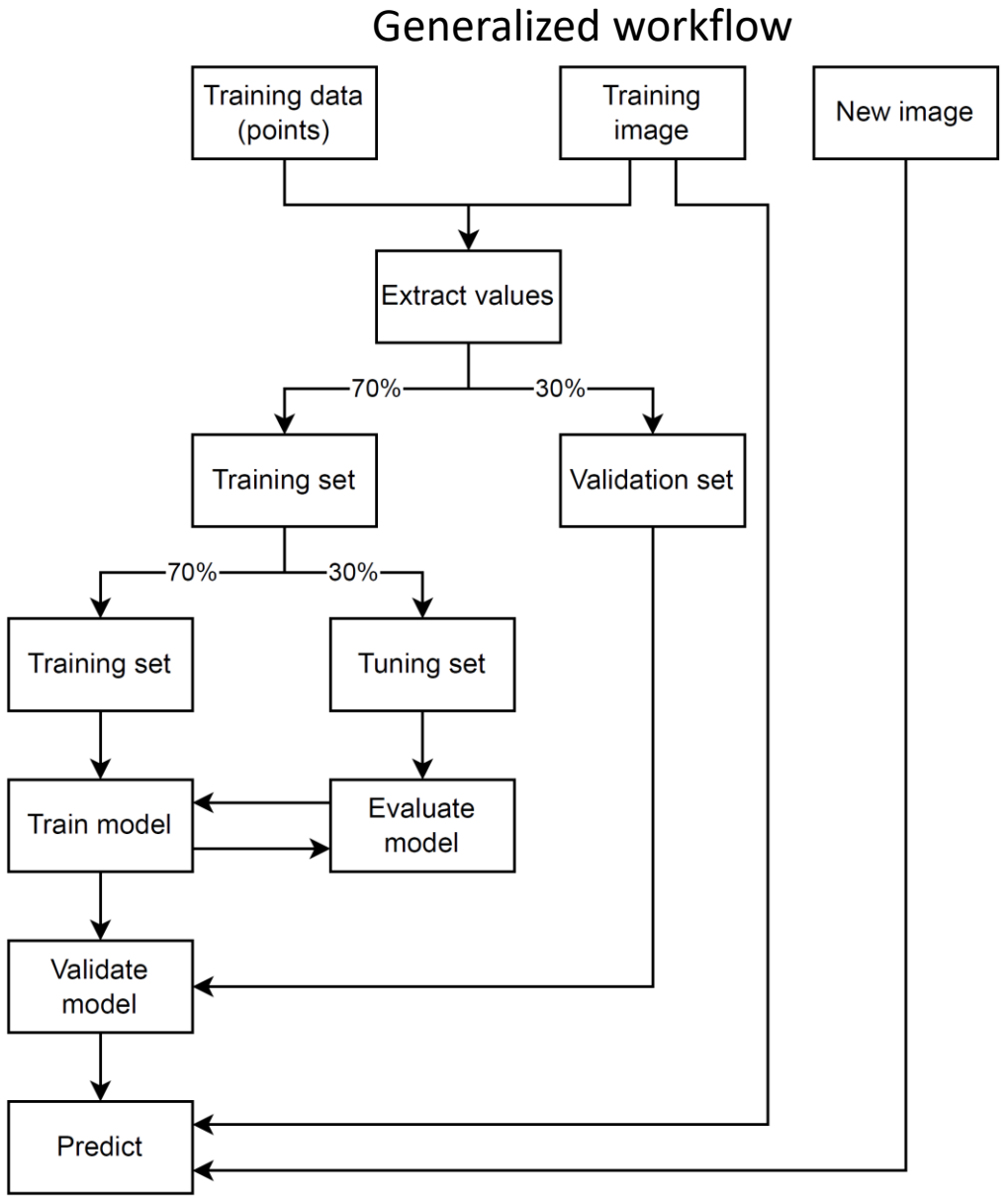
Fig. 1. Supervised fuzzy classification of Quickbird image of Macquarie: (a) Color composite of bands 4,3 and 2 of a Quickbird image of Macquarie Island with the reference areas shown as colored polygons; (b) Defuzzified classification result based on maximum membership values; (c) Image with membership values for class Bare.

Lucier 2006

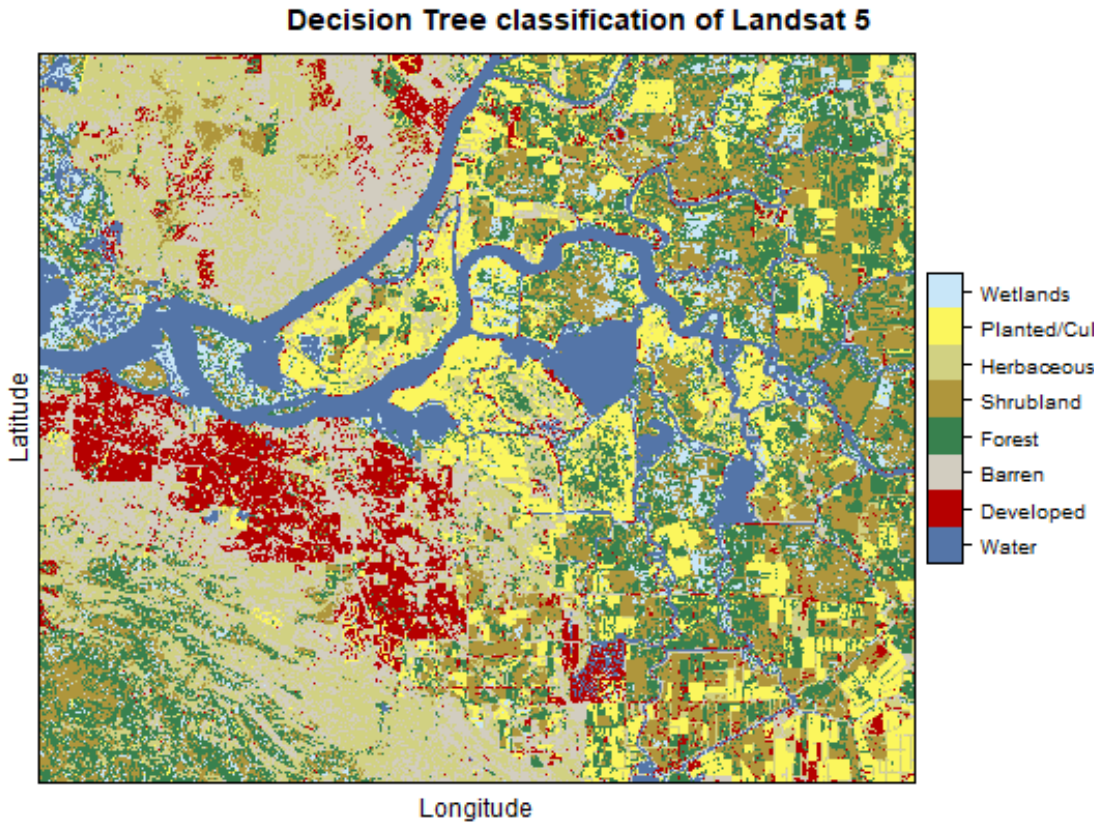
- Each element has a degree of belonging to clusters, where for any given element, the sum of its memberships across all clusters is 1.
- Minimize the distance from any given data point to a cluster center weighted by that data point's membership in the cluster.

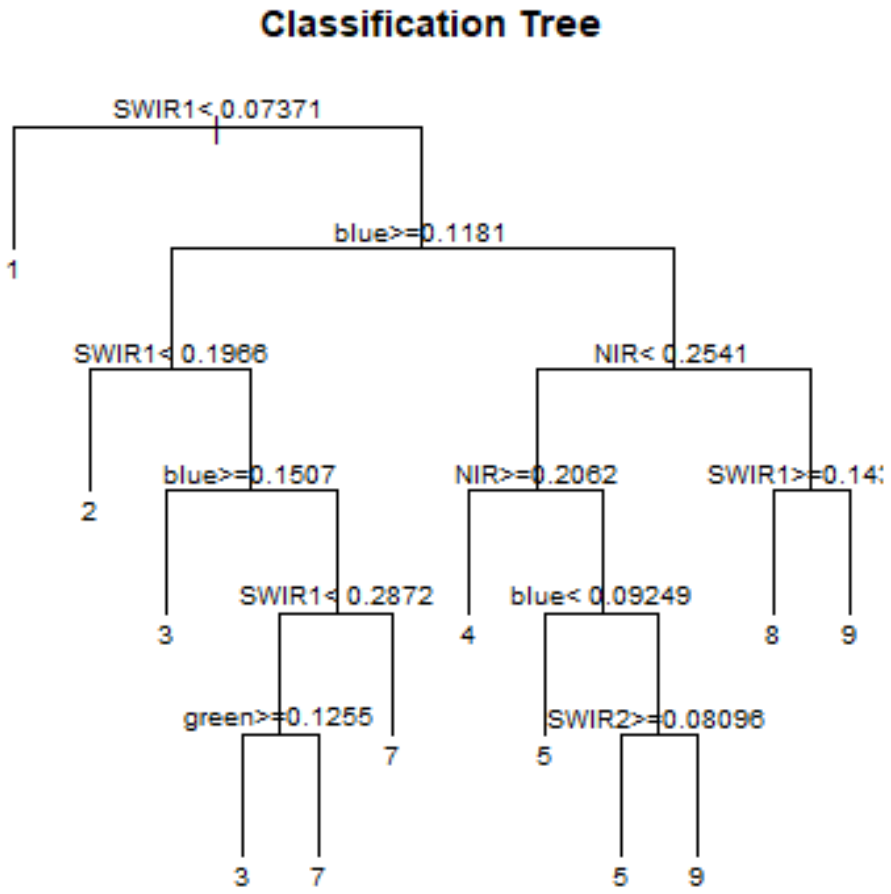
Supervised classification

- Classification and regression
- Workflow:
  - Training dataset
    - E.g. an image with labeled points
  - Train model
  - Run model
  - Accuracy assessment
  - (Optional) Tune model
- Common algorithms
  - Single regression models (linear or non)
  - K-nearest neighbours (KNN)
  - **Decision trees / CART**
  - **Random forest**
  - **Support vector machine (SVM)**
  - Spectral unmixing
  - Gradient descent / boost
  - Neural networks



Classification and Regression Trees

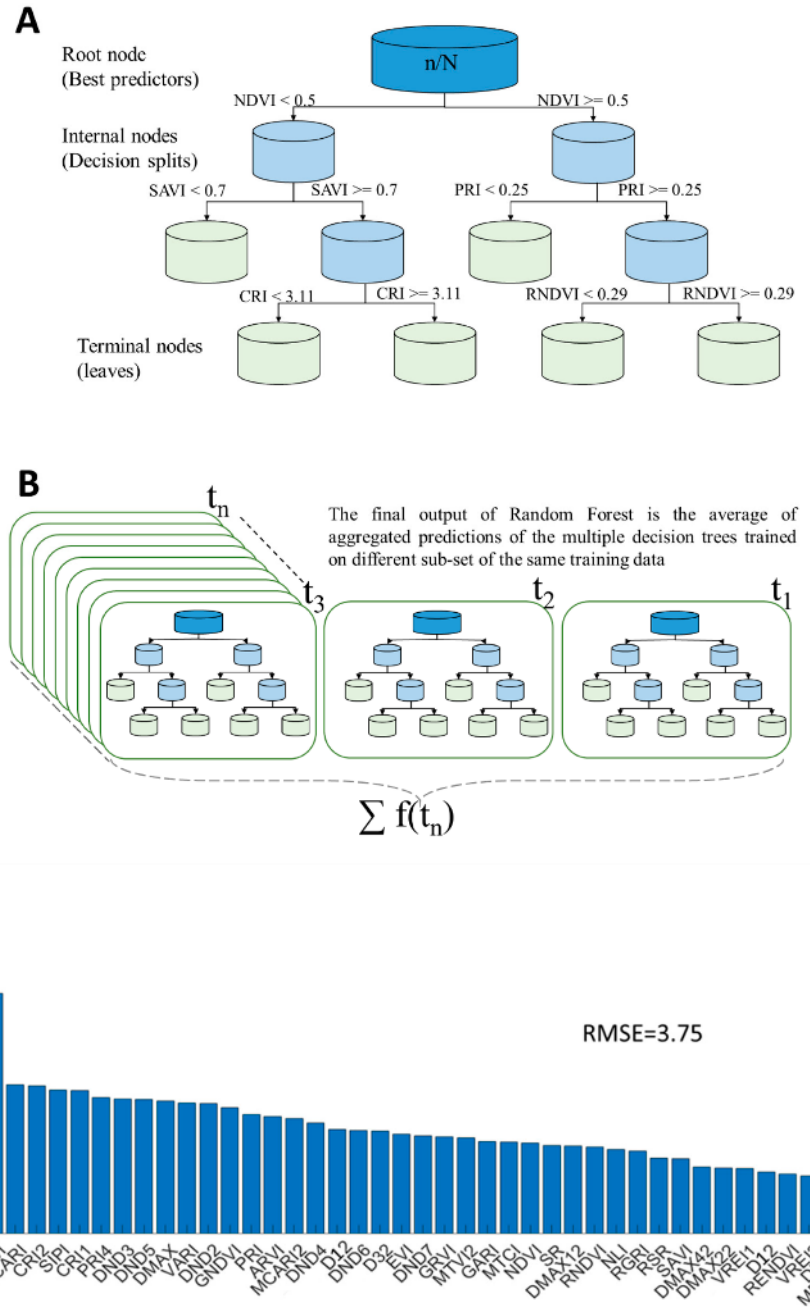




https://rspatial.org/raster/rs/5-supclassification.html

Random forest

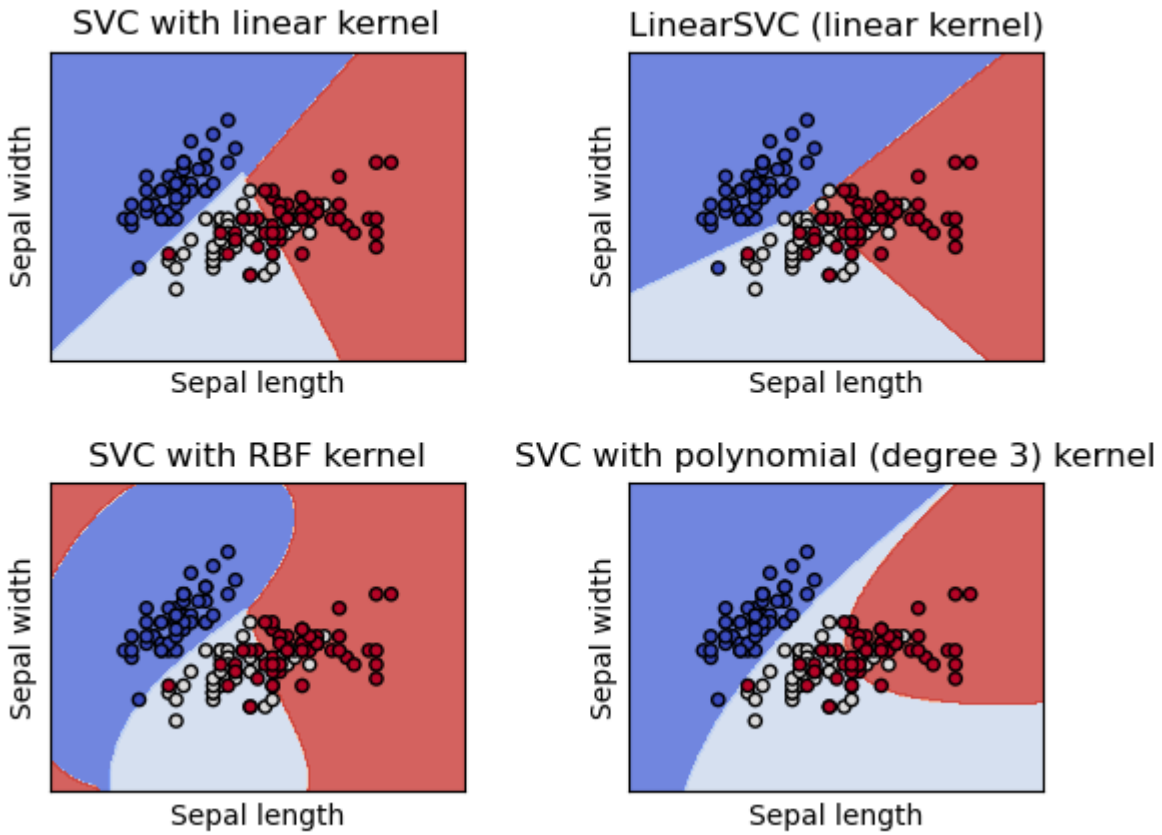
- RF is an ensemble classifier that produces multiple decision trees, using a randomly selected subset of training samples and variables
- Very popular, very accurate
- RF can handle high data dimensionality and multicollinearity, it is fast and insensitive to overfitting, data does not need to be normalized/standardized
- Sensitive to the sampling design
- RF outputs the variable importance



Shah et al. 2019

Support Vector Machines (SVM)

- Optimizes the distance to a hyperplane that best separates different classes in the feature space
- Hyperplane can be non-linear



Model assessment – confusion matrix

- Test model against validation data, where the right classes are already known
- It allows easy identification of confusion between classes, as well as the accuracy of the classification. Here's a breakdown of what it typically includes:
- For binary classification there can be:
  - **True Positives (TP)**
  - **True Negatives (TN)**
  - **False Positives (FP)** also known as Type I error
  - **False Negatives (FN)** also known as Type II error
- In the example:
  - 40 true positives (TP): Cats correctly identified as cats.
  - 45 true negatives (TN): Dogs correctly identified as dogs.
  - 5 false positives (FP): Dogs incorrectly identified as cats.
  - 10 false negatives (FN): Cats incorrectly identified as dogs.

Actual \ Predicted	**Cat**	**Dog**
Cat	40	10
Dog	5	45

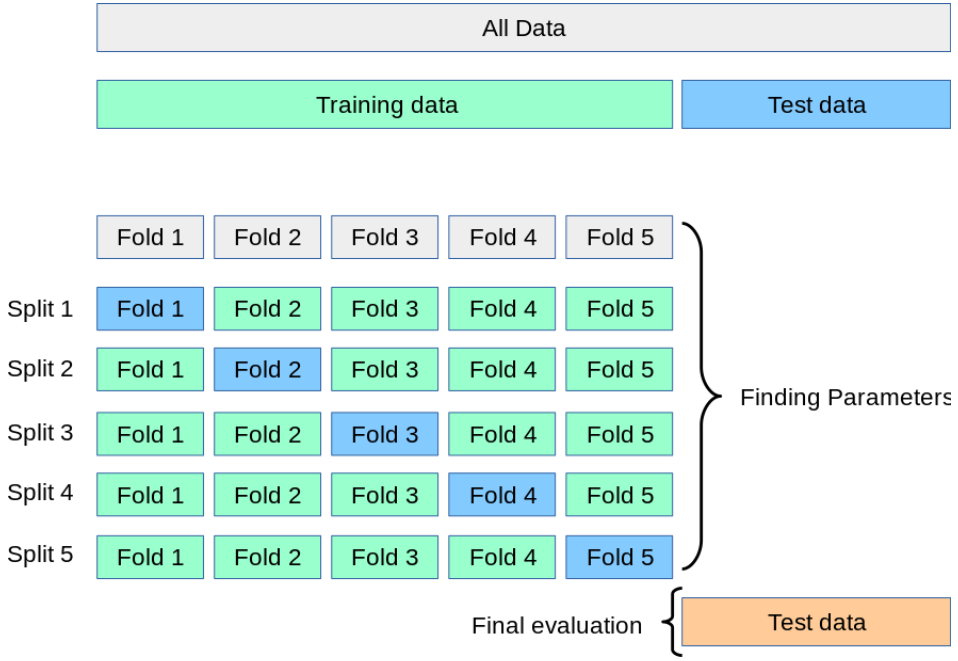
Model assessment – confusion matrix

- **Accuracy**:
  - Overall, how often is the classifier correct?
  - $(TP+TN)/(TP+TN+FP+FN) = 85.71\%$
- **Precision** (Positive Predictive Value):
  - When it predicts positive, how often is it correct?
  - $TP/(TP+FP) = 83.33\%$
- **Recall** (Sensitivity, True Positive Rate):
  - How often does it correctly identify positives?
  - $TP/(TP+FN) = 90.91\%$
- **F1 Score**:
  - A weighted average of precision and recall.
  - $2*(Precision*Recall)/(Precision+Recall) = 86.96\%$

Actual \ Predicted	Cat	Dog
Cat	40	10
Dog	5	45

K-fold cross validation

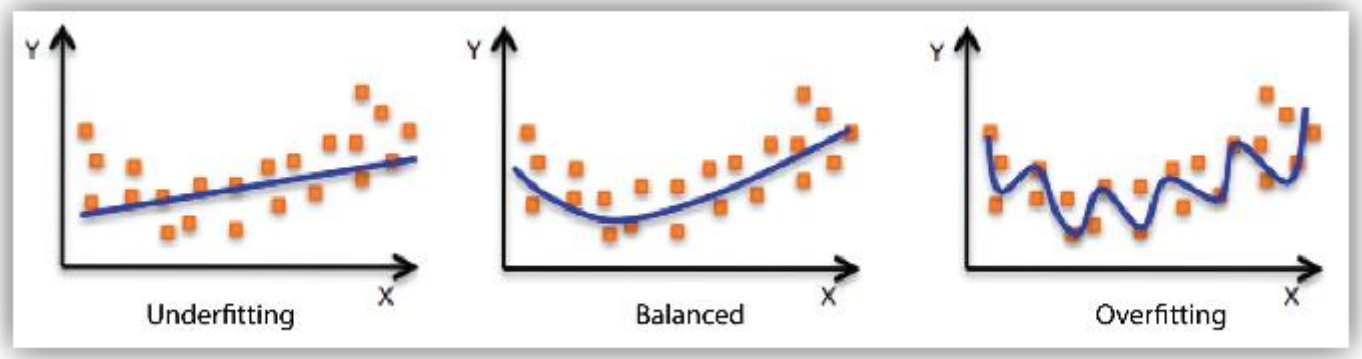
- Method to evaluate the performance of a machine learning model
- Helps ensure its generalizability to an independent dataset
- Divide the Dataset:
  - The entire dataset is randomly divided into k equal-sized folds or subsets.
- Iteratively Train and Test:
  - The model is trained and tested k times. Each time, one of the k subsets is used as the test set (also known as the validation set), and the remaining k−1 subsets are put together to form a training set.
- Average the Results:
  - After k iterations, the model will have been trained and tested across all subsets of the data. Performance metrics are averaged to produce a single estimate.
- If k = n-1, called Leave one out cross validation, typically k = 10



https://scikit-learn.org/stable/modules/cross_validation.html

Overfitting models

- Machine learning models can learn the training data too well, capturing noise or random fluctuations in the training data instead of the underlying distribution.
- **Characteristics of Overfitting:**
  - High training accuracy
  - Poor test accuracy
  - More likely with complex models
- **How to Test for Overfitting**
  - Split your data
  - Cross-validation
  - Learning curves
- **How to Address Overfitting:**
  - Simplify the model
  - More training data
  - Prune the model
  - Early stopping
  - Dropout / Introduce noise



https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

Model selection criteria

- We did not even scratch the surface of available models…
  - List of models: https://scikit-learn.org/stable/supervised_learning.html
- Model selection:
  - Interpretability
    - Simpler models generally are mote interpretable
  - In memory vs out memory
    - Storage vs RAM
  - Number of features and examples
    - Small training dataset with few covariates = CART or KNN
    - Small training dataset with many covariates = SVM or gaussian
    - Large training dataset with many covariates = Neural networks and boosting
  - Categorical vs numerical features
    - Classification or regression?
  - Normality of data
    - Linear relationships or non-linear?
  - Training speed
    - More complex models generally take longer to train
  - Prediction speed
    - Important if real-time predictions are needed
  - https://medium.com/mlearning-ai/brief-guide-for-machine-learning-model-selection-a19a82f8bdcd

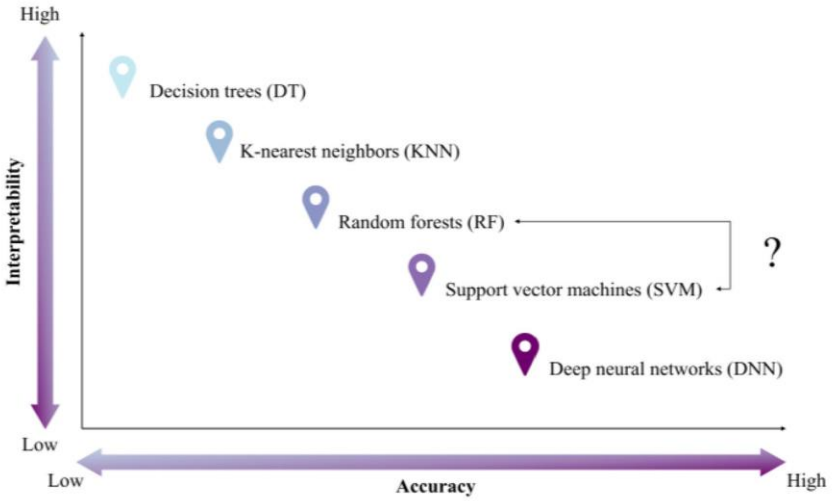
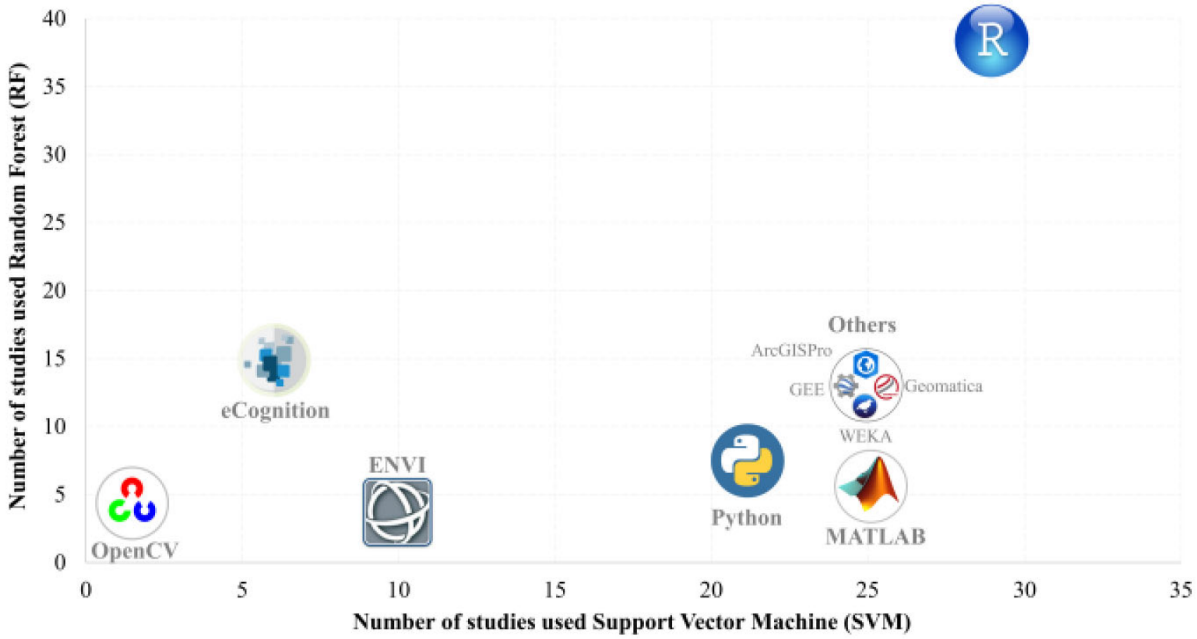
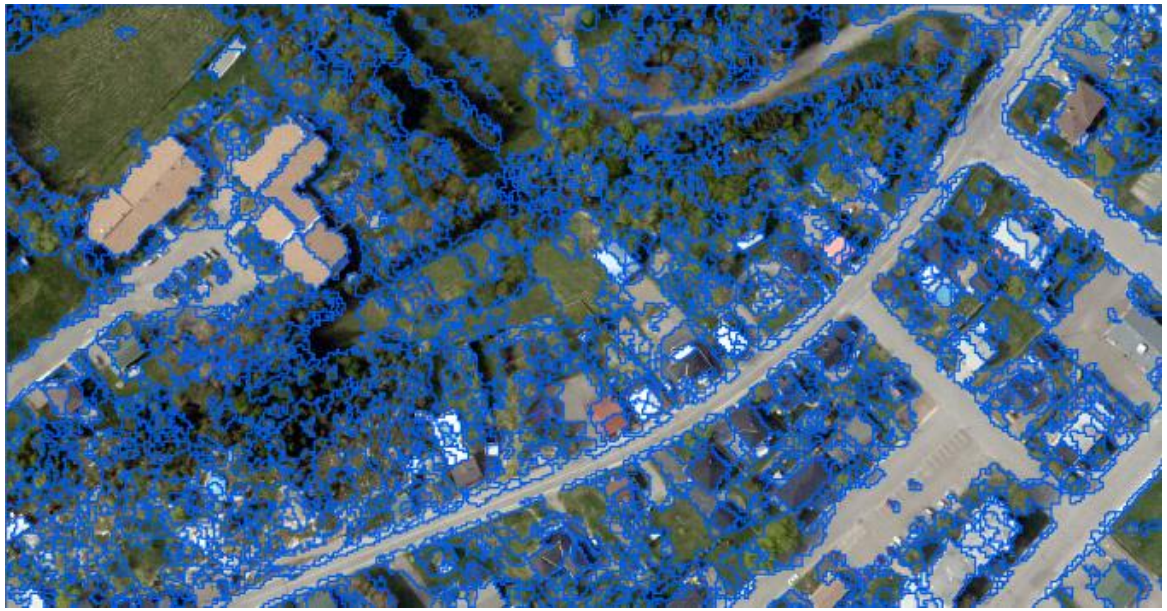


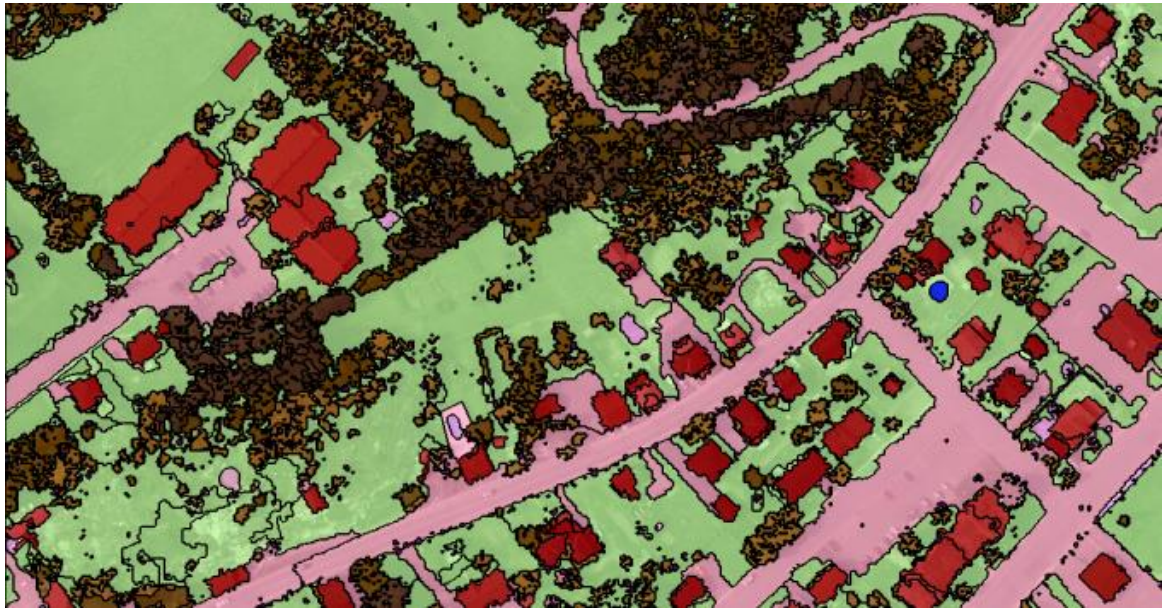
Fig. 1. Interpretability-accuracy tradeoff in machine learning classification algorithms.

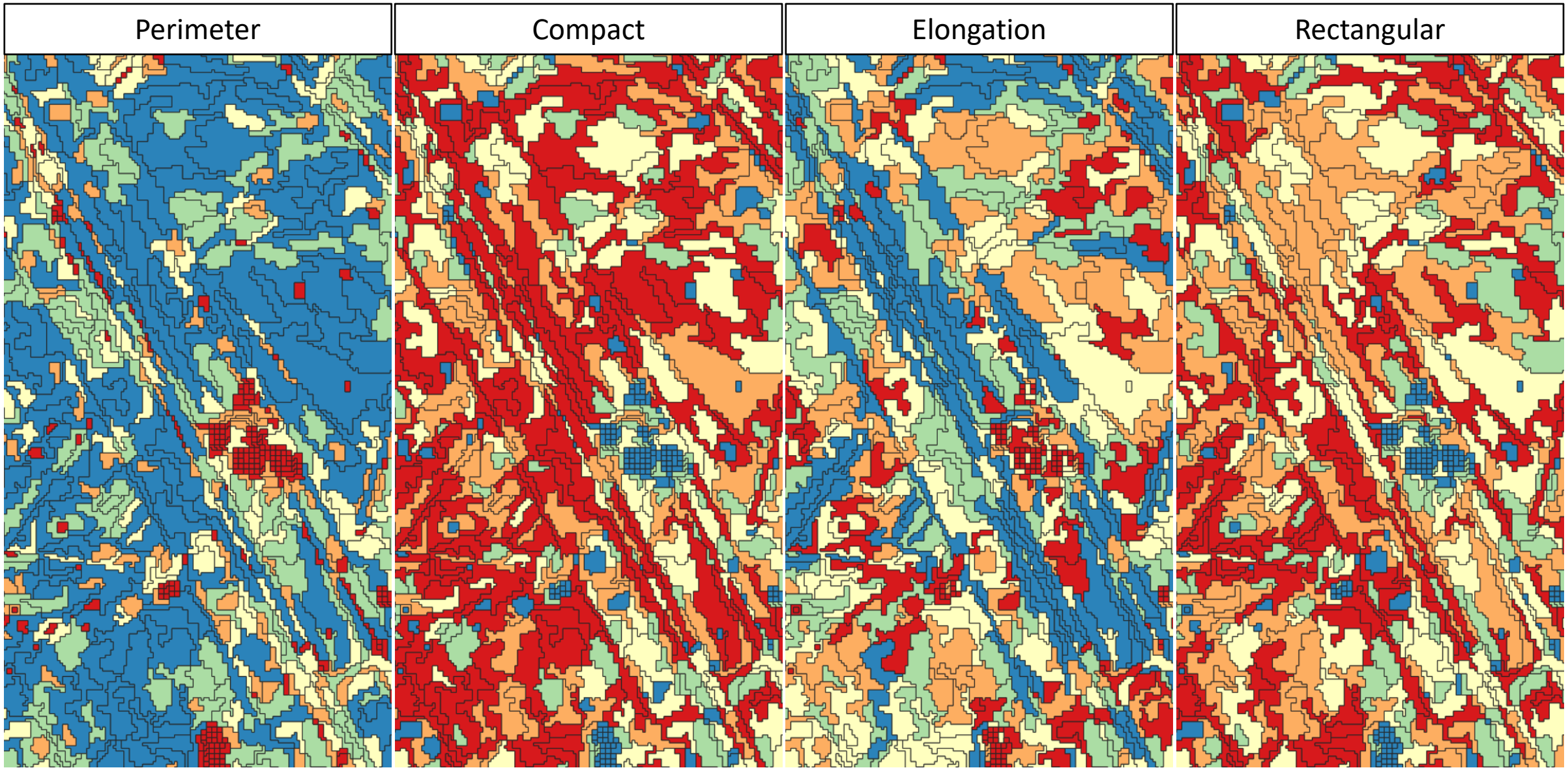


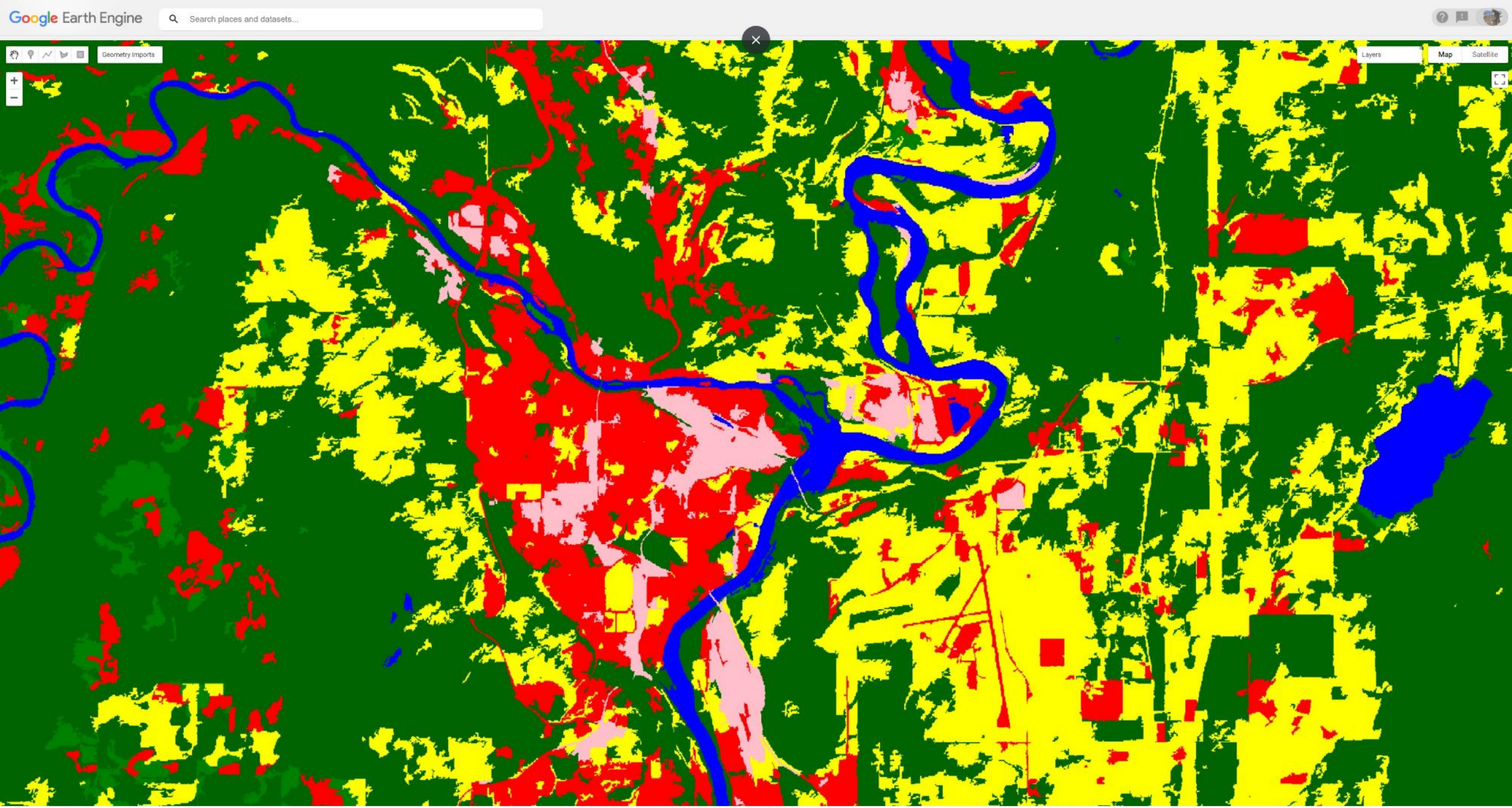
OBIA / Segmentation

- Image segmentation
  - Break up the image into similar polygons
  - Calculate shape attributes (length, width, area, perimeter, adjacency, etc.)
  - Algorithms have parameters: Input bands, scale, shape, compactness
- Classification
  - Same as pixel-based but uses spectral and geometrical polygon statistics
- Grew in popularity with high resolution satellite imagery
- Software
  - Commercial: E-cognition, ArcGIS, PCI
  - Open: GRASS, OrfeoToolbox, R-SuperPixels, Python-SciKit Learn, SAGA






Perimeter
Compact
Elongation
Rectangular



https://code.earthengine.google.com/018f25cf7c7f66041f0168a3e47d32ba

Next steps

- Steps to improve accuracy
  - Reducing dimensionality,
  - Adding terrain and derivatives,
  - Spectral indices,
  - Texture and edges,
  - SAR,
  - Seasons, …
- Other methods
  - Time series classification
  - Spectral unmixing
  - Semantic segmentation
  - Deep learning / AI
- Recent advances
  - Pre-labeled training data https://bigearth.net/#downloads
  - Pre-trained models https://github.com/tchambon/DeepSentinel