Integrating Machine Learning into Geographic Research

Machine Learning for Geography

What is machine learning?

- Machine Learning is the science (and art) of programming computers so they can learn from data.
  - Giving computers the ability to learn from data (empirical observations) without explicitly programing every computing step.

- A classic machine learning example: email spam filter, i.e., given an email how can a computer model decide if it is a spam or not (or "ham")

- A geospatial example: given the RS image of a piece of land, whether it is agriculture land or not

What is machine learning?

- A classic machine learning example: email spam filter, i.e., given an email how can a computer program decide if it is a spam or not (or "ham")

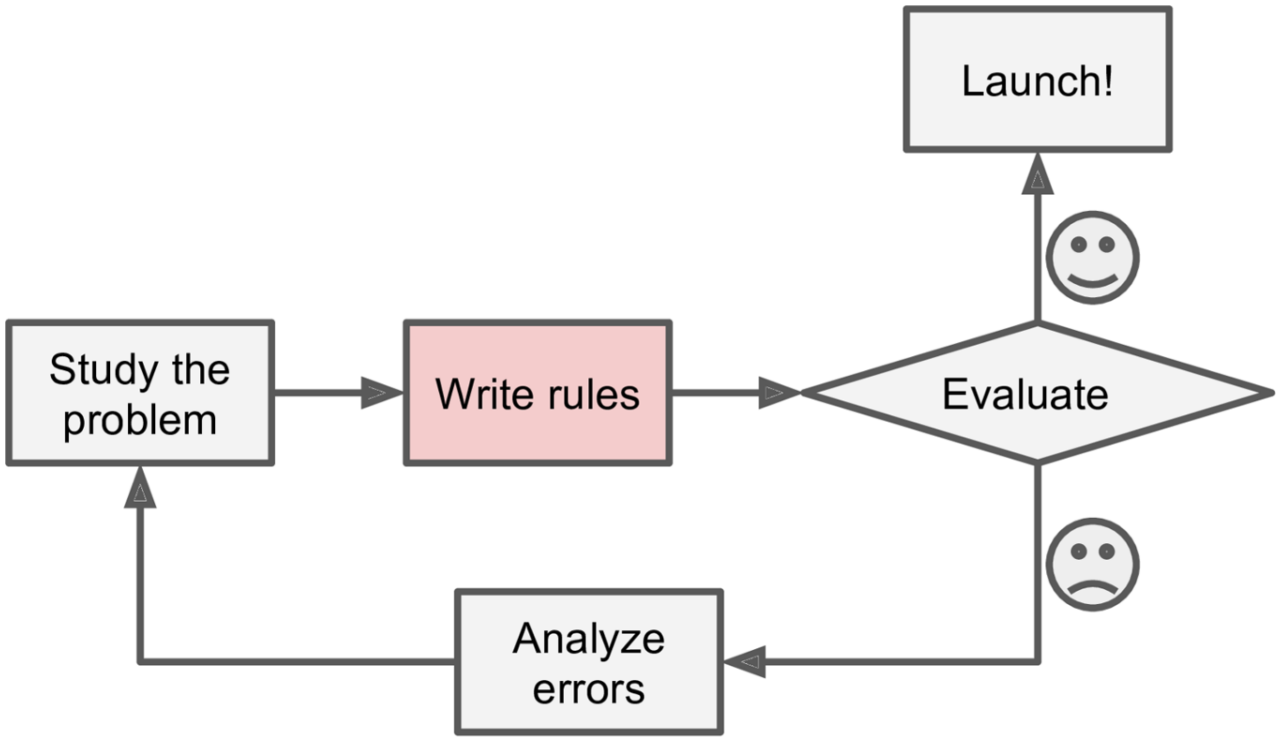
- In traditional programming, we study this problem and then program the computer to do so
  - We manually read 100 spam emails, and noticed that spam emails often contain certain words, such as "free", "amazing", "sale", …
  - We may then write a program: if email contains {"free", "amazing", "sale", …}, then email is spam; else email is ham

What is machine learning?

- You then put your program on a test, and noticed it filtered some spam emails but missed some others
- You study the missed spam emails again, and noticed that they contain some additional words, such as "credit"
- You add this word in your program
  - if email contains {"free", "amazing", "sale", "credit", …}, then email is spam; else email is ham
- You test it again and noticed some other missing words

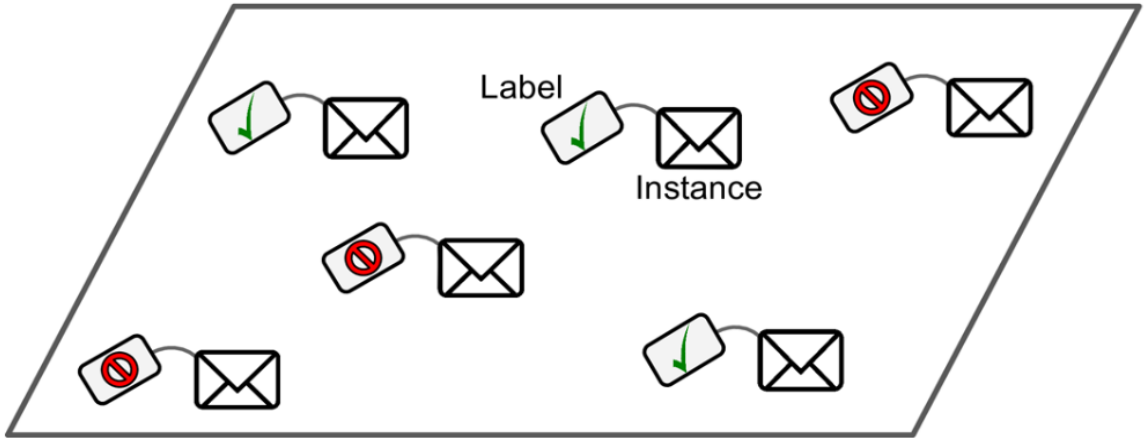
What is machine learning?

- Keeping doing this process, we eventually get a gigantic rule-based program that is inflexible and hard to maintain



What is machine learning?

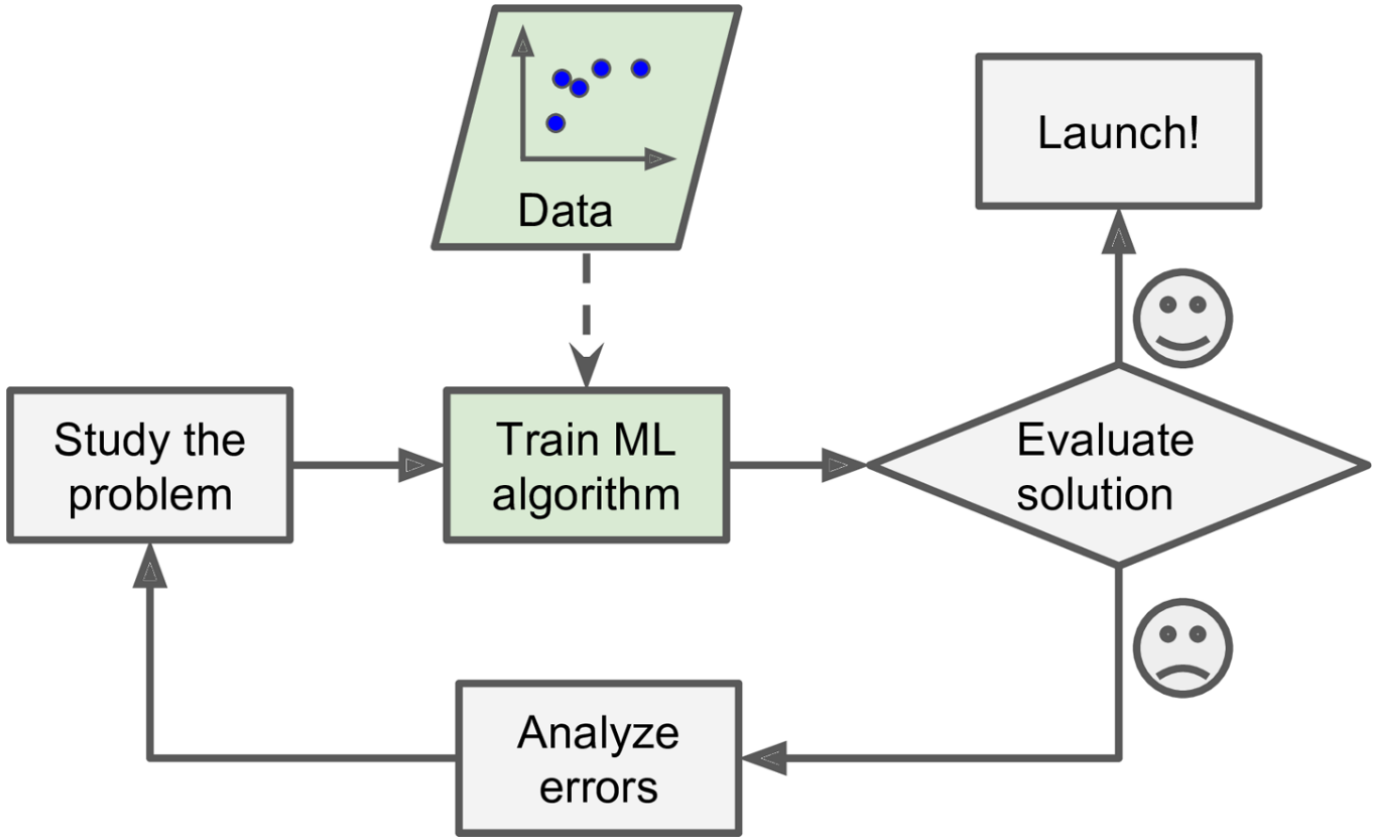
- In machine learning, a computer program learns to classify spam emails by observing data
- We feed the computer with training data, i.e., data with labels
  - E.g., a dataset of 200 emails, with 100 emails labeled as spam, and 100 emails labeled as ham



By observing the words used in these emails and their labels, the computer figure out by itself which words suggest spam and which suggest ham

What is machine learning?

- In machine learning, a computer program learns to classify spam emails by observing data



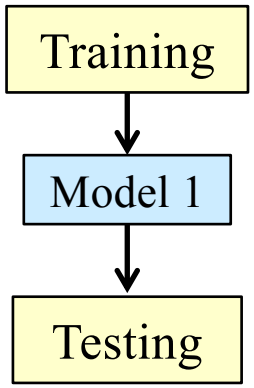
What is machine learning?

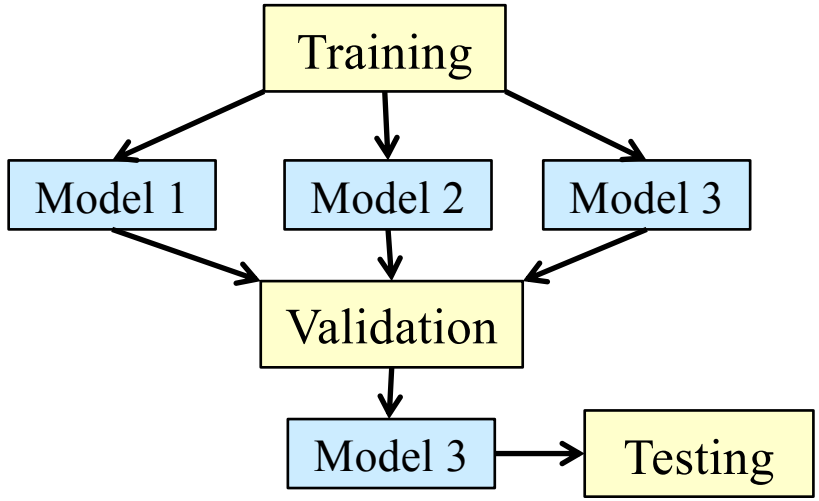
- What we will obtain is a machine learning model that is usually more concise than a traditional rule-based program, and easier to maintain and update

- E.g., as time passes, spam emails may use certain new words; a machine learning model can dynamically catch up with these new words as the user flags some emails as spam

Training data, validation data, and test data

- Training data is used for learning the parameters of the models
- Validation data is used to decide which model to employ
- Test data is used to get a final, unbiased estimate of how well the model works
- One single model usually only uses training and test data; validation dataset is typically used when we have multiple models to train





It is common to use 80% of the data for training and *hold out* 20% for testing.

Classification of machine learning models

- Different ways to classify machine learning models:
  - Whether or not they are trained with human supervision (supervised, unsupervised, semisupervised, and reinforcement learning)
  - Whether or not they can learn incrementally on the fly (online versus offline learning)
  - Whether they work by simply comparing new data points to known data points, or instead detect patterns in the training data and build a predictive model (instance-based versus model-based learning)

Integrating ML into Geographic Research

- A number of typical steps:
  - Define your geospatial problem
  - Identify the current approaches
  - Get geospatial data
  - Prepare and explore data
  - Select and train ML models
  - Evaluate the performance of your models
  - Report and paper writing

Define your geospatial problem

- The goal of this step is to clearly and explicitly formalize the geospatial problem you are trying to solve
  - Study the problem by reading geography literature, discussing with your advisor, reading reports, …
  - What are the input data and desired result of the problem? What might be YOUR contribution to this problem?
  - Can you formalize it into a classification problem (to classify the output into a category), a regression problem (to predict a numeric value), or a clustering problem (to identify clusters in the data)?

Define your geospatial problem

- Some examples:
  - Classifying whether a remote sensing image contains airports: classification problem; input is the spectral information of a remote sensing image; output is a binary classification
  - Predicting the housing price at a geographic area: regression problem; input is the neighborhood environment and demographic information of the area; output is a numeric value
  - Identifying the activity zones of an endangered species: clustering problem; input is the locations of the species; output is the identified location clusters

Identify the current approaches

- The goal of this step is to identify baselines whose performances can be compared against YOUR model in a later step

- You are typically not the first person to look at this problem. If you find yourself in such a situation, dig deeper into the literature or ask for suggestions from your advisors

- Sometimes, it is possible you are the first person studying a problem when you are working with the newest technology, e.g., a technology that just came out this year

Identify the current approaches

- How do people address this geospatial problem currently?
  - Doing this task manually or semi-automatic (low efficiency)
  - Automatic but low accuracy
  - Automatic, high accuracy, but long computing time
  - Automatic, high accuracy, short computing time, but can't generalize to a slightly changed geographic location
  - …
- This step can help you identify the limitations of existing approaches, select the performance aspect to be evaluated, and clarify the contribution of your work.

Get geo data

- The goal of this step is to obtain training (including validation) and test data for your ML model
- The performance of your model can be directly affected by the quality of the data that you feed to the model
- Some sources for geospatial data:
  - Public domain data from US federal governments
  - State or city data are generally OK for research purposes but be cautious for commercial applications
  - Data from existing publications
  - Collecting the data yourself
  - …

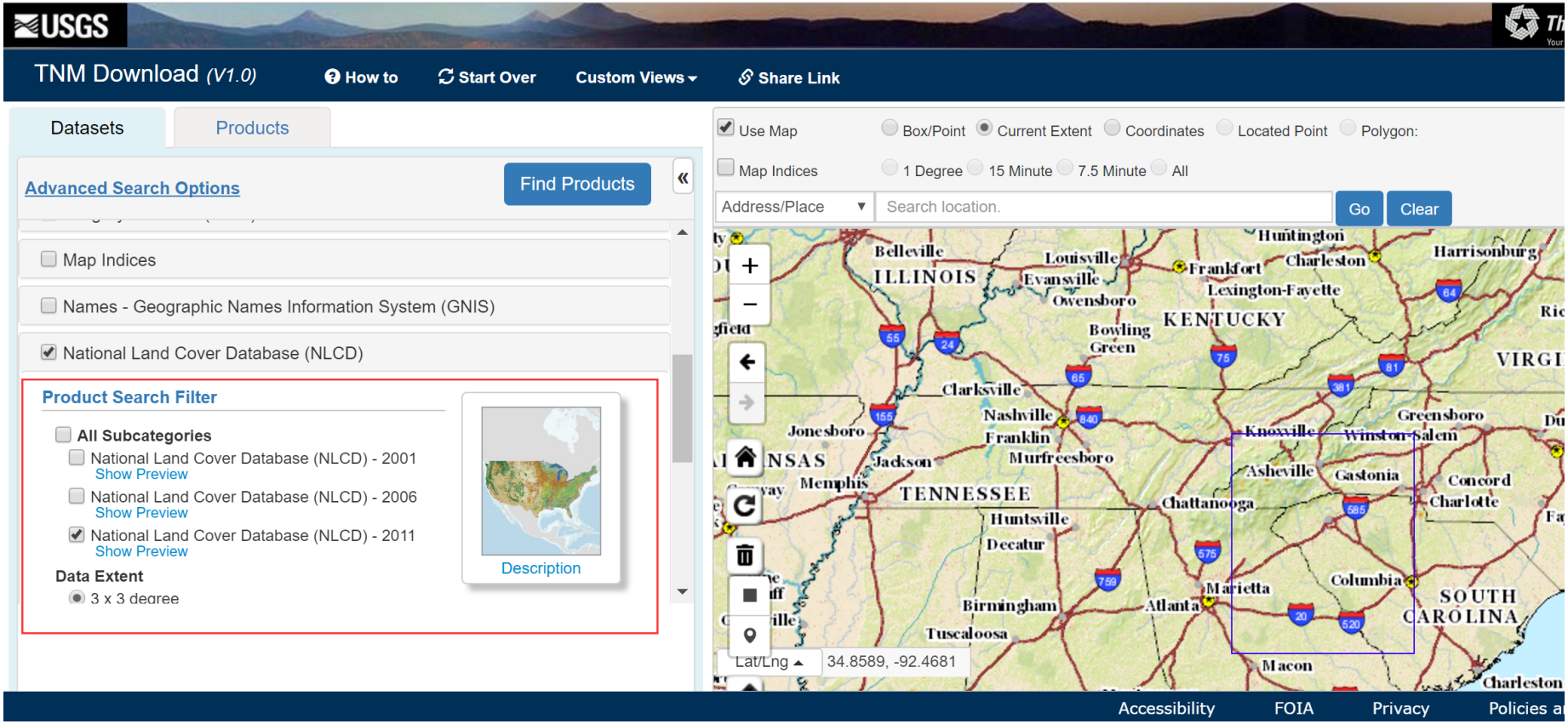
Get geo data

- Public domain data from US federal governments: public domain data are data without copyrights (you can use them for any purposes). All data from the US federal governments are in the public domain.

- Data about natural environment: USGS, NASA, NOAA, ..
- Data about human society: US Census, DOT, …

National Land Cover Dataset (NLCD)

- USGS national map viewer: https://viewer.nationalmap.gov/basic

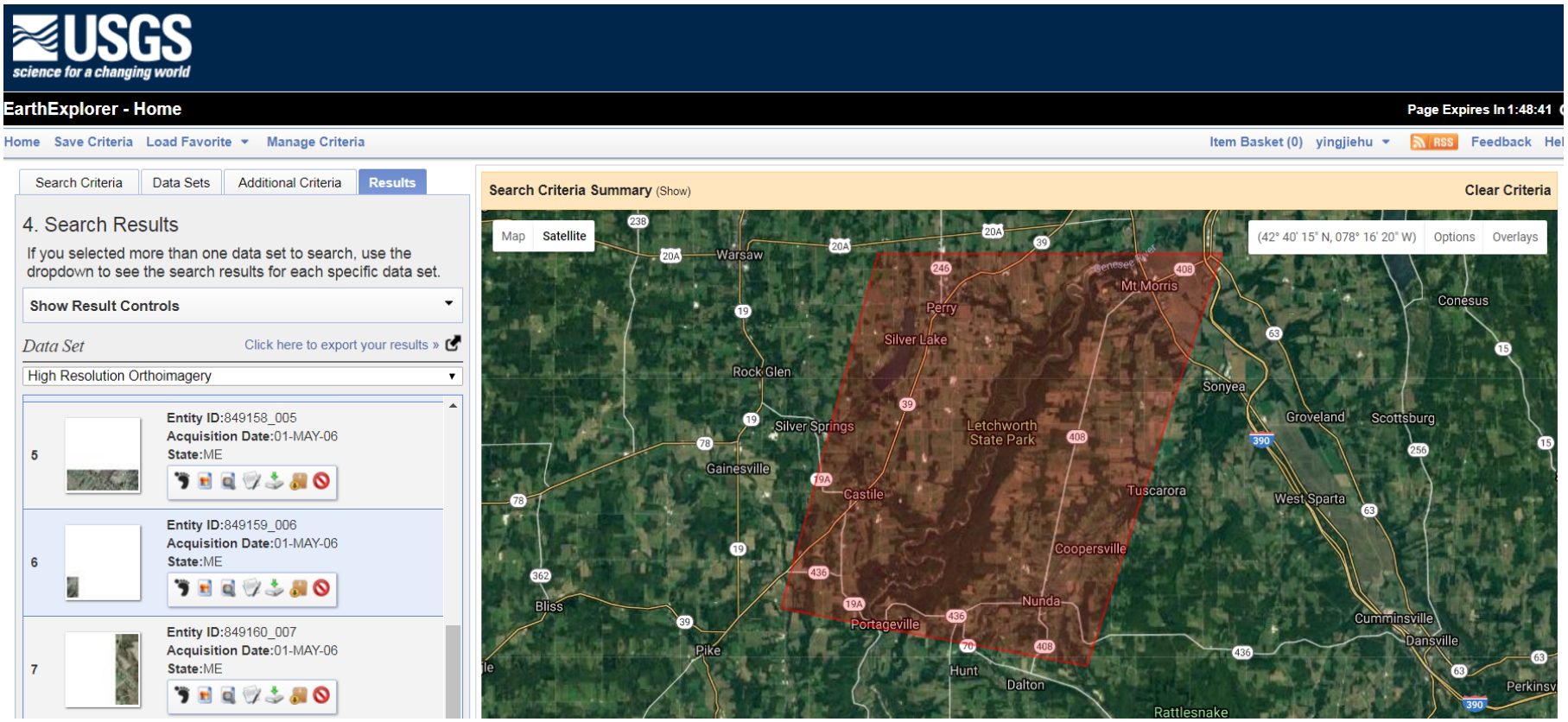


Remote sensing images

- USGS High Resolution Orthophotograph (HRO) images has 0.3 meter (1 foot) spatial resolution

High resolution orthophotograph

- Obtaining HRO: https://earthexplorer.usgs.gov/

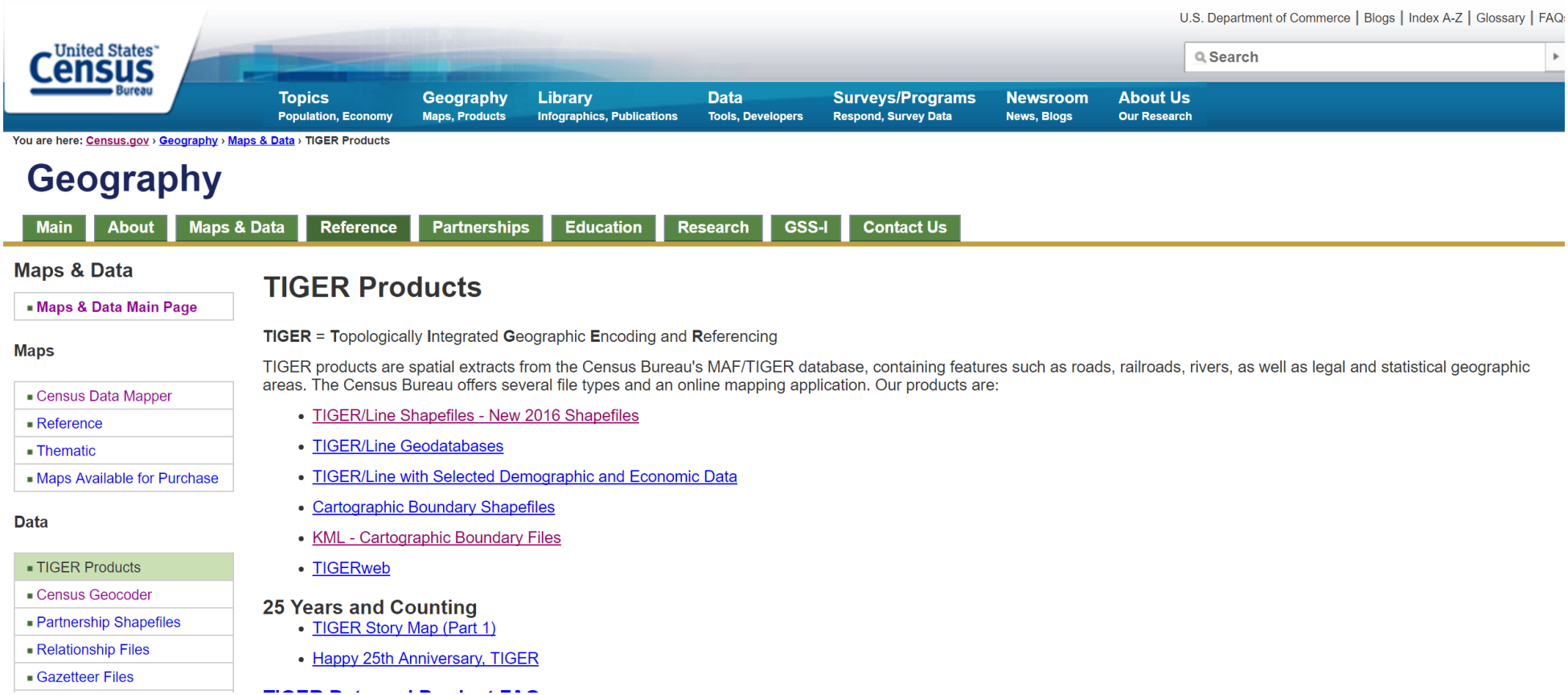


US Census TIGER Data

- Topologically Integrated Geographic Encoding and Referencing
- Developed by the US Census Bureau
- Contain geographic features including roads, railroads, rivers, as well as legal and statistical geographic areas.
- Also called TIGER/Line
- Esri shapefiles

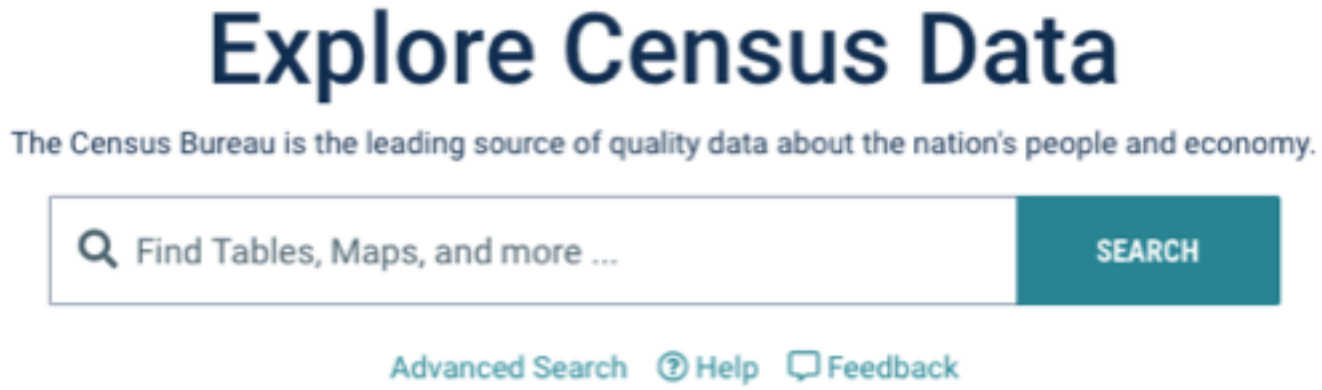
Obtaining TIGER Data

- https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html



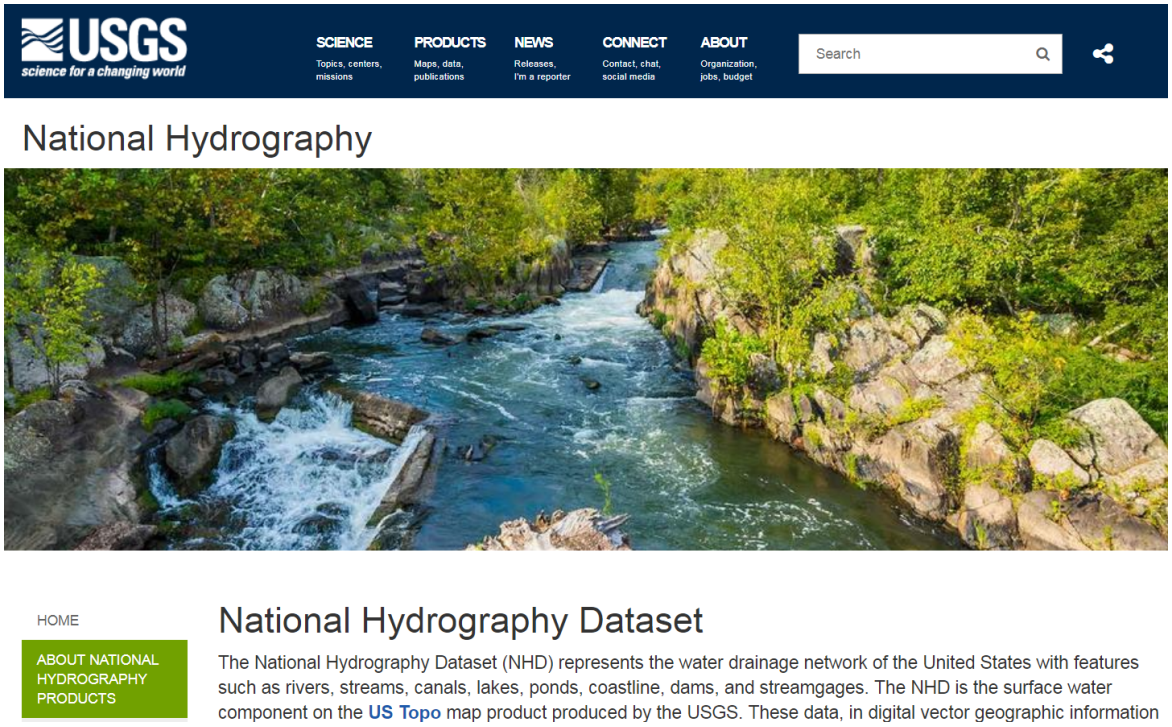
US Census American Community Survey (ACS data)

https://data.census.gov/



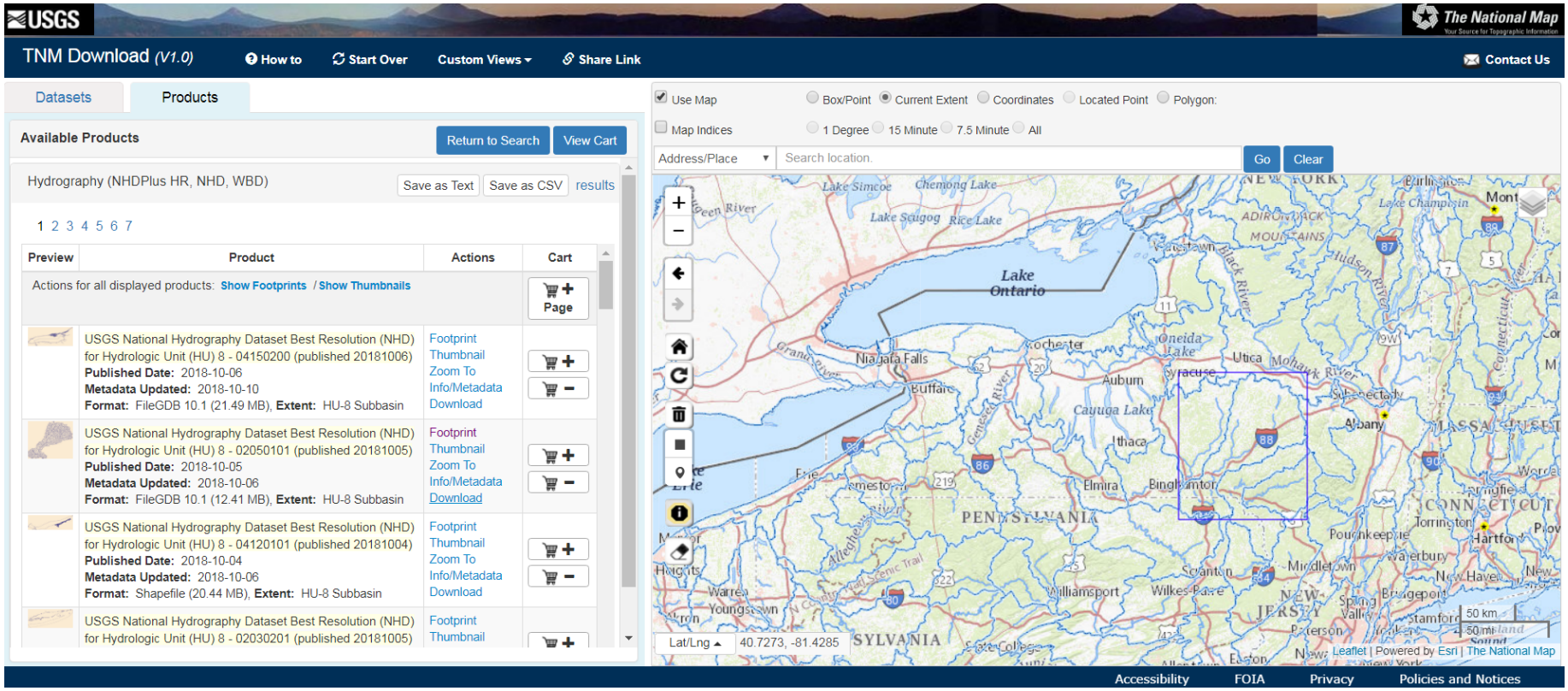
National Hydrography Dataset

- NHD: National Hydrography Dataset
- Data about the water drainage network of the United States, with features such as rivers, streams, canals, lakes, ponds, coastlines, dams, and streamgages.



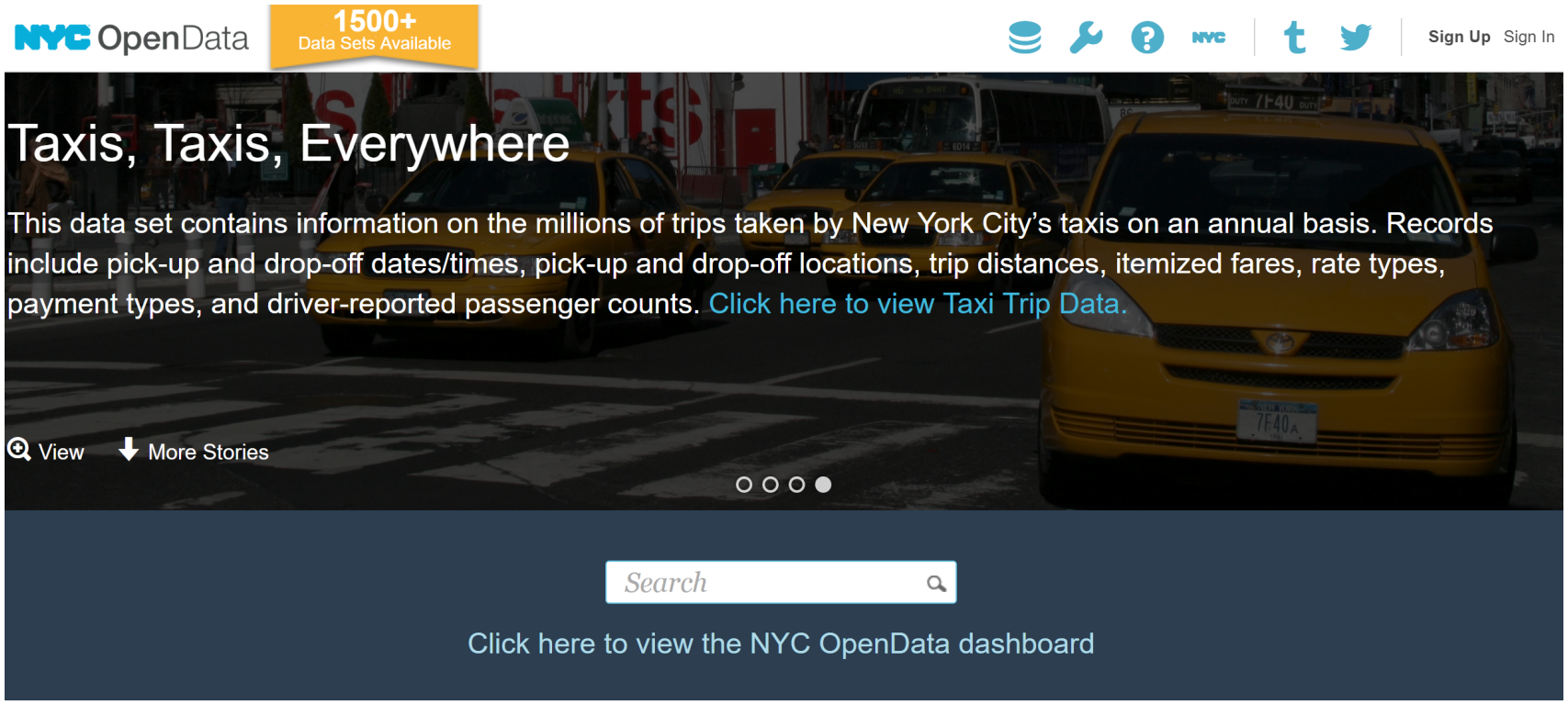
Obtaining NHD

- Downloading from USGS: https://viewer.nationalmap.gov/basic/



State or city data

- NYC data portal (https://opendata.cityofnewyork.us/)



Get geo data

- Data from publications
- Some authors have kindly shared their data on their websites or GitHub

4.1.1 Processing steps

Starting with approximately 10 M entries in the GeoNames database, the following steps were used to create the final corpus:

1. Locations with the same name had to be separated by a large (1000 km) distance to be distinct and to remove duplicates.
2. The most ambiguous locations (ones with the highest count in step 1) were processed first. For instance, the five most ambiguous GeoNames locations are Santa Maria (26 entries), Santa Cruz (25), Victoria (23), Lima (19) and Santa Barbara (19), although Wikipedia does not have a page for each.

[23] https://github.com/milangritta/WhatsMissingInGeoparsing.

[24] http://www.geonames.org/export/.

[25] http://www.geonames.org/wikipedia/.

[26] https://www.mediawiki.org/wiki/API:Main_page.

[27] https://pypi.python.org/pypi/wikipedia.

Prepare and explore data

- The goal is to enhance your understanding of the data and increase the quality of the data
  - Data cleaning and imputation
  - Exploratory data analysis (histograms, min, max, median, …)
- Having a better sense of data can help you find out the reasons when your model does not perform well
- Increasing the quality of data can help you train better models

Prepare and explore data

- Divide your data into training and test in a representative manner

- Hold your test data in a secure vault, and don't look at it
  - Our brain can cheat us if we peek into the testing data, which eventually lead to overfitting

- You may further divide your training data into training and validation
  - K-fold cross-validation

Select and train ML models

- The goal is to produce the best model for your problem
  - You need to have some knowledge about different models: what results do they produce and what assumptions do they make? This is a major focus of this course
  - You need to use the models that can best capture the particular aspects of your problem
    - Co-location patterns
    - Sequential data
    - 2D remote sensing images
    - …

Select and train ML models

- Train one or multiple candidate models using the training data
- Iterate the training process for cross-validation
- Select the best model based on their performances on the validation data

Evaluating model performance

- The goal is to demonstrate that your model is better than existing approaches
  - Adopt one or several existing and commonly used approaches (baselines)
  - Apply these baselines to the same test dataset
  - Compare your model with existing baselines based on the performance aspect you have selected (e.g., accuracy, speed, time, …)
  - If your model performs better than the existing ones, congrats!
  - If not, find out the reason and repeat the training of models with necessary changes (e.g., with new features, other models, better data, …)

Report and paper writing

- The goal is to make your work into tangible outcome and get the reward you deserve

- The structure of a typical paper
  - Introduction (what is the problem? why is it important?)
  - Related Work (baselines)
  - (Study region and data)
  - Methodology (how does your model work?)
  - Experiments (data used, experiment procedure, evaluations)
  - Conclusions and future work

Application of ML models to geospatial data and problems

- Leveraging existing ML models to address geospatial problems (you are not changing the model itself)
- Application oriented and more commonly seen in the literature
- The challenges are in formalizing the geospatial problem, obtaining and organizing your geospatial data, and geospatial feature engineering

Developing spatially explicit ML models

- Developing new models by including spatial relations (e.g., spatial weights); you are changing the existing model

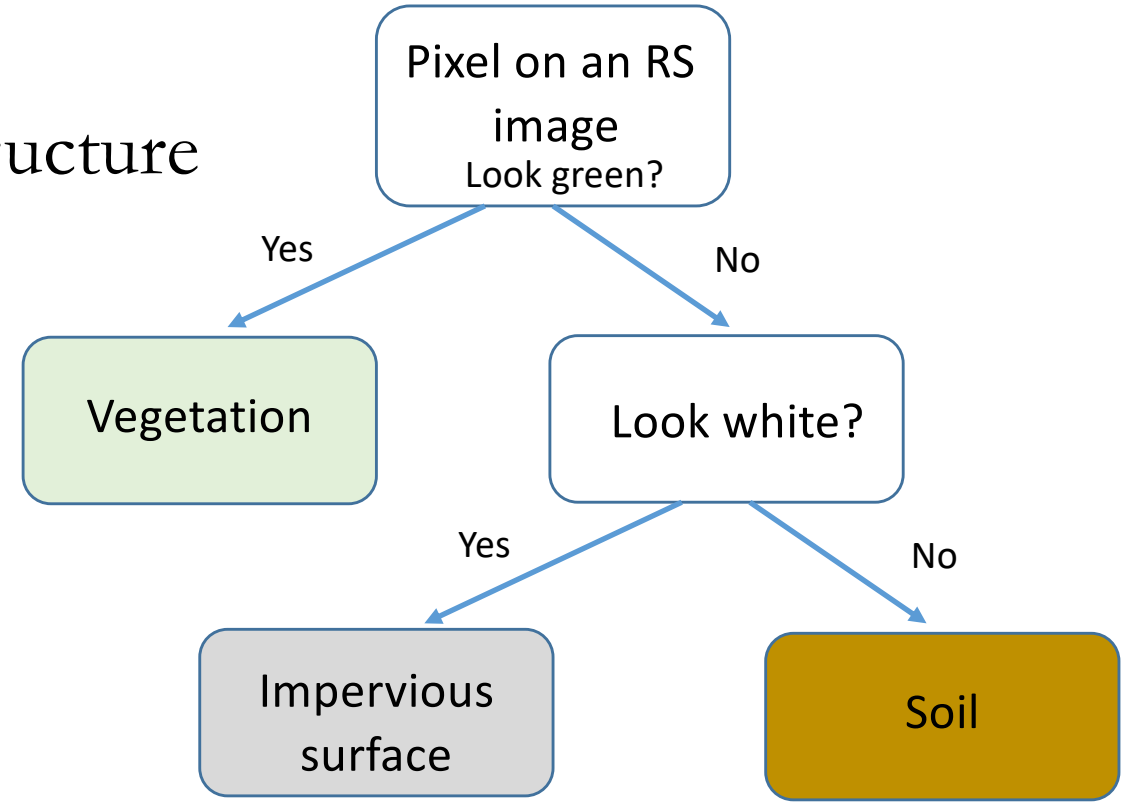
- Model development oriented and relatively fewer studies compared with applications

- The challenges are in understanding the core mechanism of a model (understanding some of the math behind) and revise the mechanism to capture spatial relations

Decision tree

- Decision tree is a versatile machine learning model that can be used for both classification and regression

- The model has a tree structure

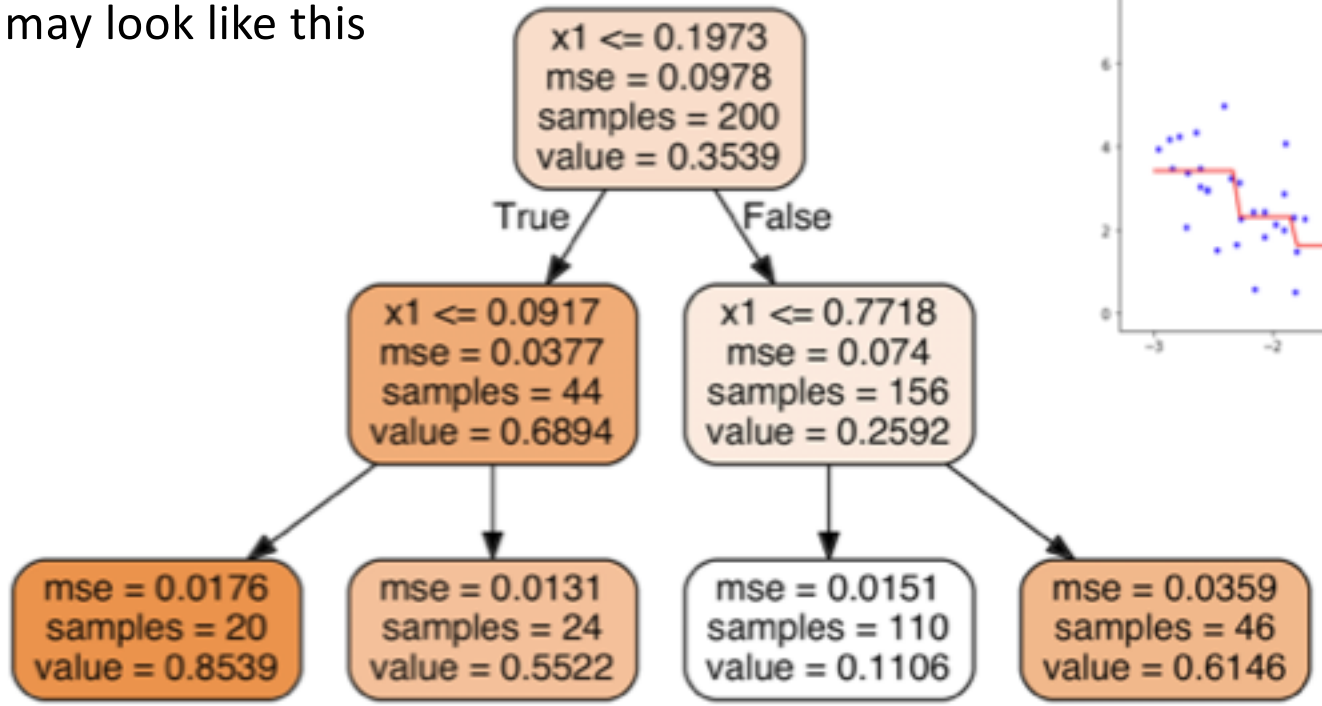


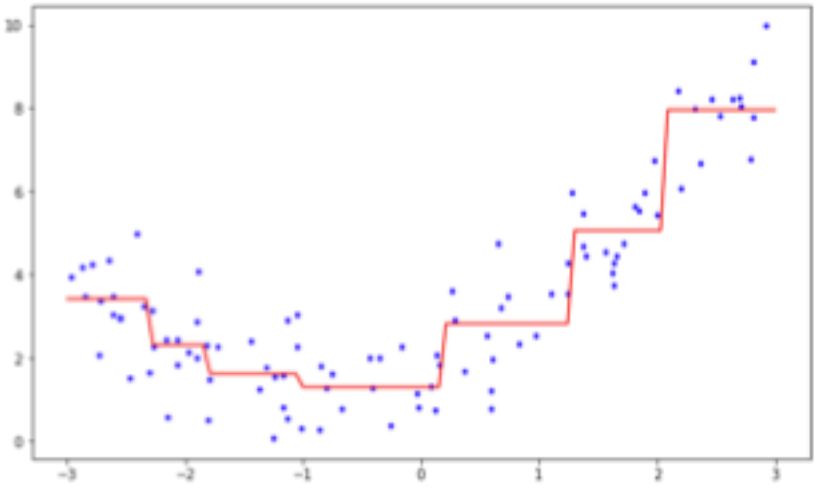
- Training a model means building this tree using training data

Decision tree

- A decision tree can also be used for regression
- When used for regression, the decision tree outputs the average value of a leaf node when given a new data record

A regression decision tree may look like this





Random forest

- Ensemble learning: train multiple models for the same problem (based on the same training dataset), and then aggregate their predictions

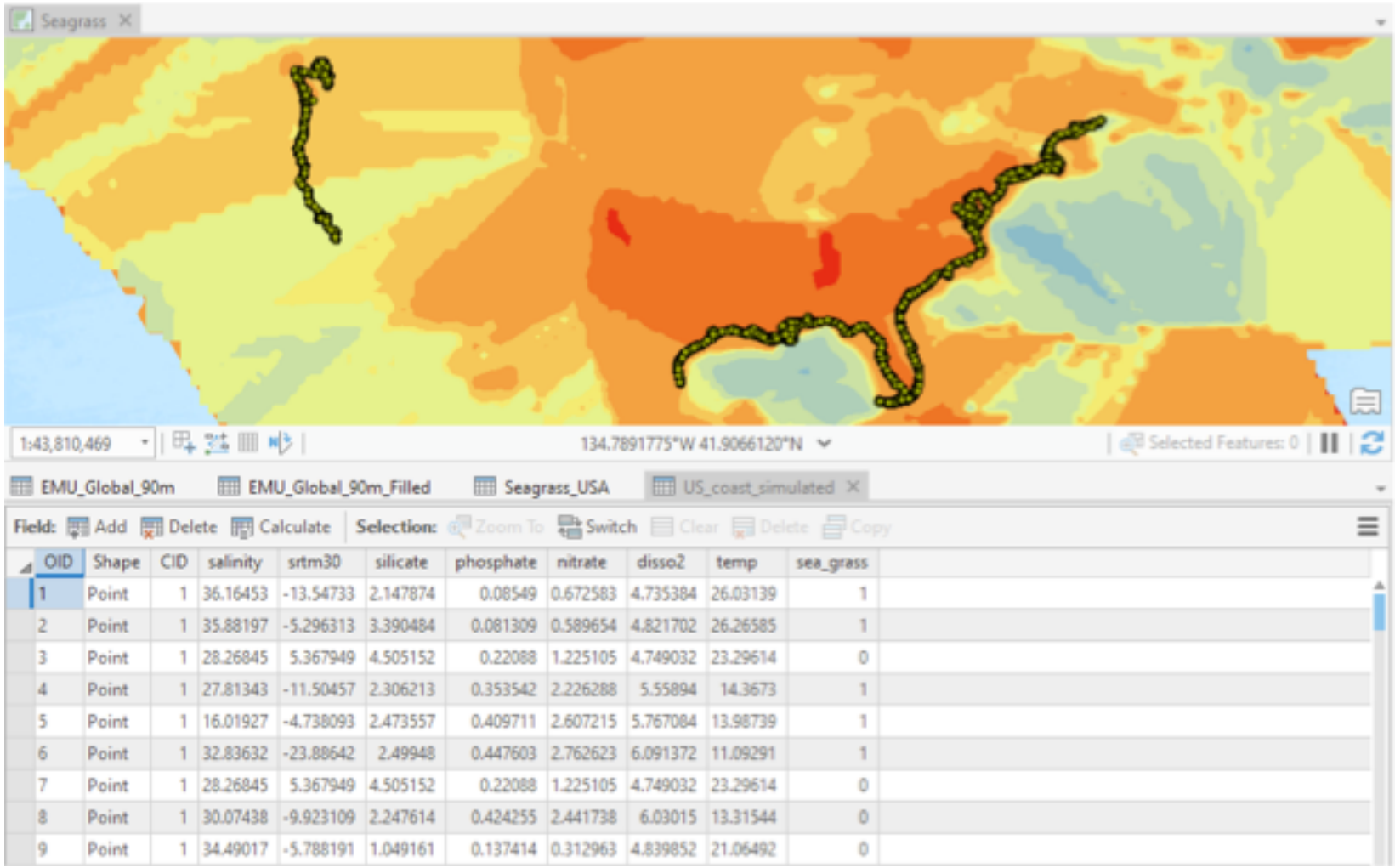
- The wisdom of the crowd: ask many people about the same question, and then aggregate their answers

- Random forest: an ensemble of decision trees (e.g., 100 decision trees) trained on random subsets of the training data

Using decision tree and random forest to predict nitrate concentration along US coastline

Experiment data are revised based on https://learn.arcgis.com/en/projects/predict-seagrass-habitats-with-machine-learning/

Using decision tree and random forest to predict nitrate concentration along US coastline



OID	Shape	CID	salinity	srtm30	silicate	phosphate	nitrate	disso2	temp	sea_grass
1	Point	1	36.16453	-13.54733	2.147874	0.08549	0.672583	4.735384	26.03139	1
2	Point	1	35.88197	-5.296313	3.390484	0.081309	0.589654	4.821702	26.26585	1
3	Point	1	28.26845	5.367949	4.505152	0.22088	1.225105	4.749032	23.29614	0
4	Point	1	27.81343	-11.50457	2.306213	0.353542	2.226288	5.55894	14.3673	1
5	Point	1	16.01927	-4.738093	2.473557	0.409711	2.607215	5.767084	13.98739	1
6	Point	1	32.83632	-23.88642	2.49948	0.447603	2.762623	6.091372	11.09291	1
7	Point	1	28.26845	5.367949	4.505152	0.22088	1.225105	4.749032	23.29614	0
8	Point	1	30.07438	-9.923109	2.247614	0.424255	2.441738	6.03015	13.31544	0
9	Point	1	34.49017	-5.788191	1.049161	0.137414	0.312963	4.839852	21.06492	0

Experiment data are revised based on https://learn.arcgis.com/en/projects/predict-seagrass-habitats-with-machine-learning/