

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 programming 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?

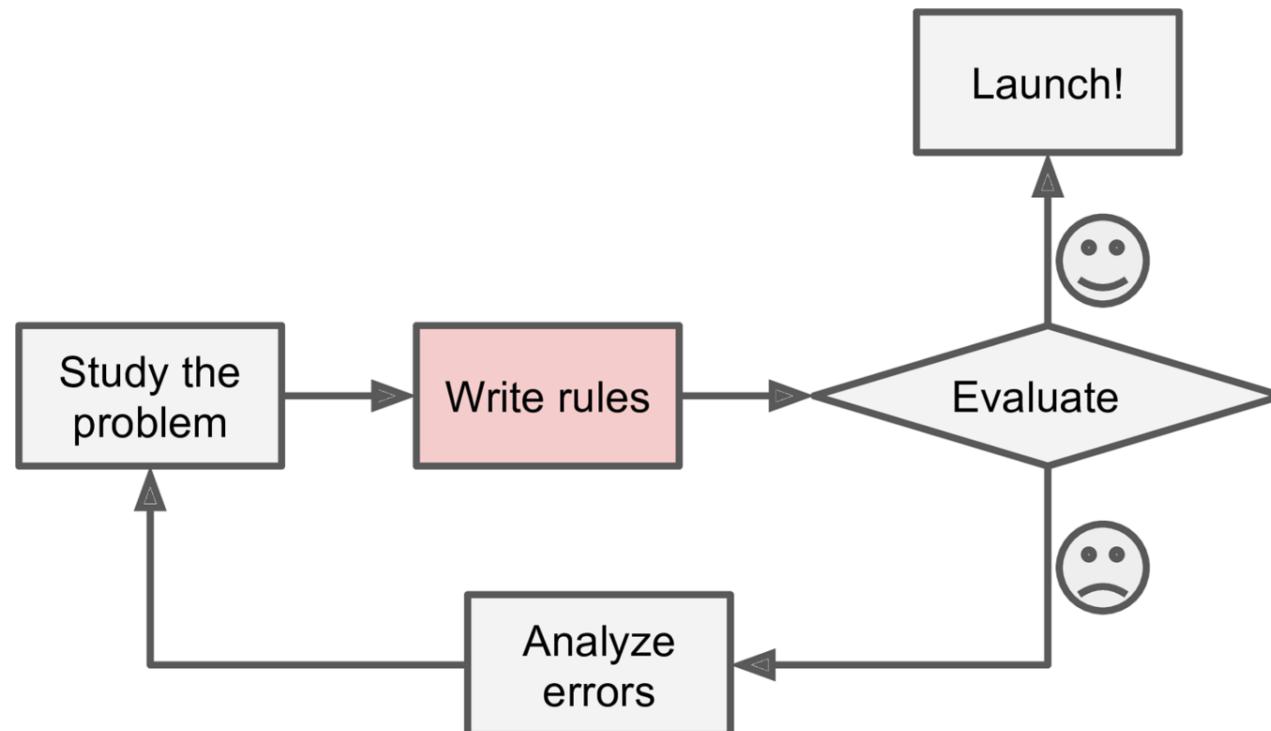
- A classic machine learning example: email spam filter, i.e., given an email how can a computer program decide if it is a spam or not (or “ham”)
- In traditional programming, we study this problem and then program the computer to do so
 - We manually read 100 spam emails, and noticed that spam emails 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

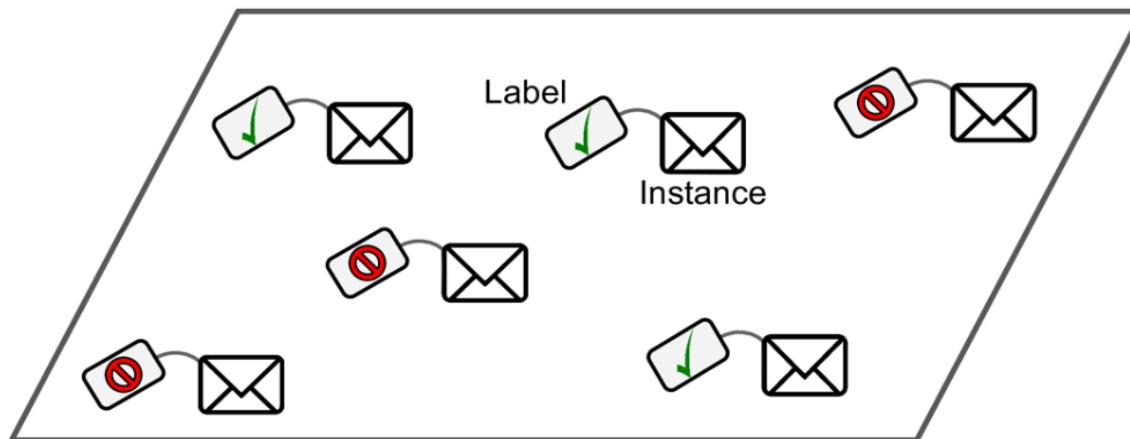
What is machine learning?

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



What is machine learning?

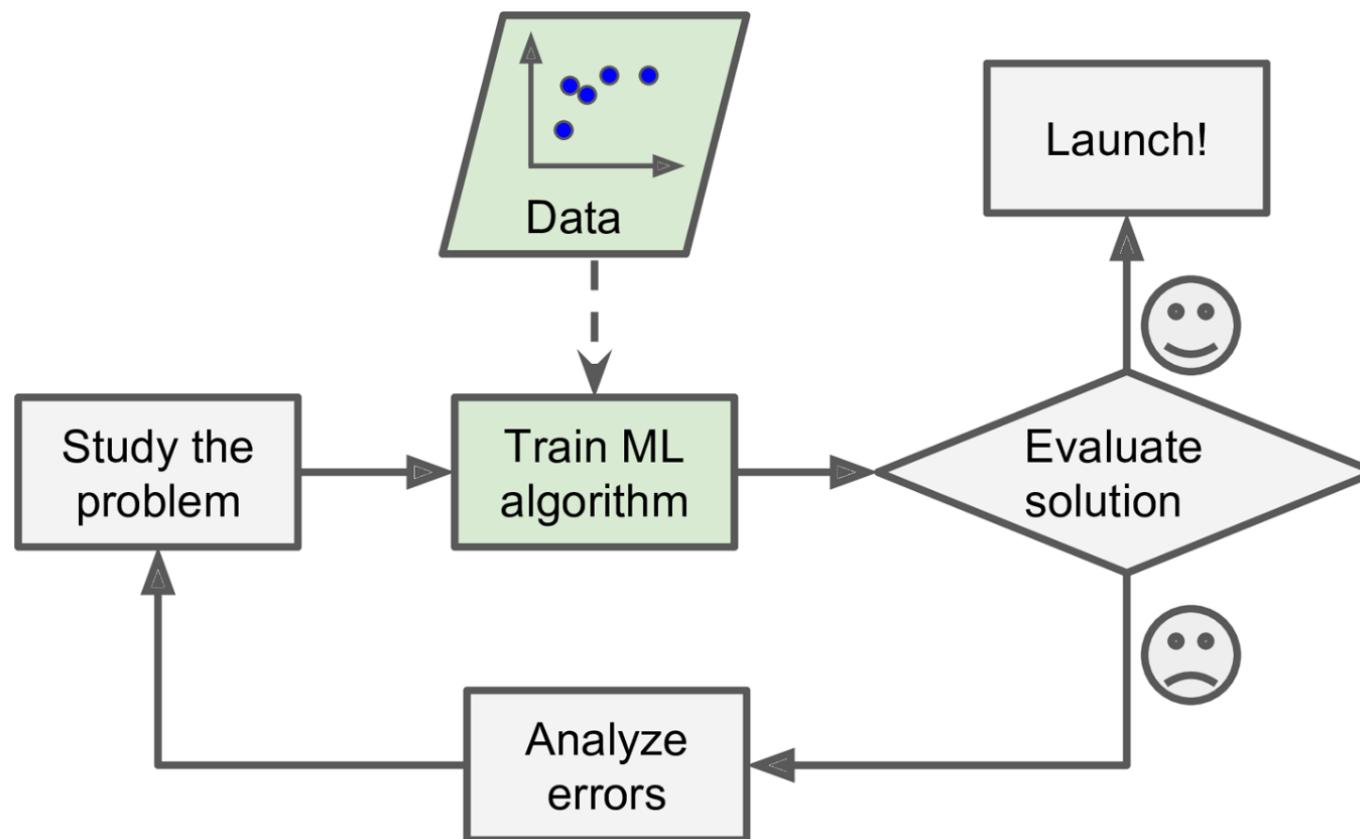
- In machine learning, a computer program learns to classify spam emails by observing data
- We feed the computer with **training data**, i.e., data with labels
 - E.g., a dataset of 200 emails, with 100 emails labeled as spam, and 100 emails labeled as ham



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

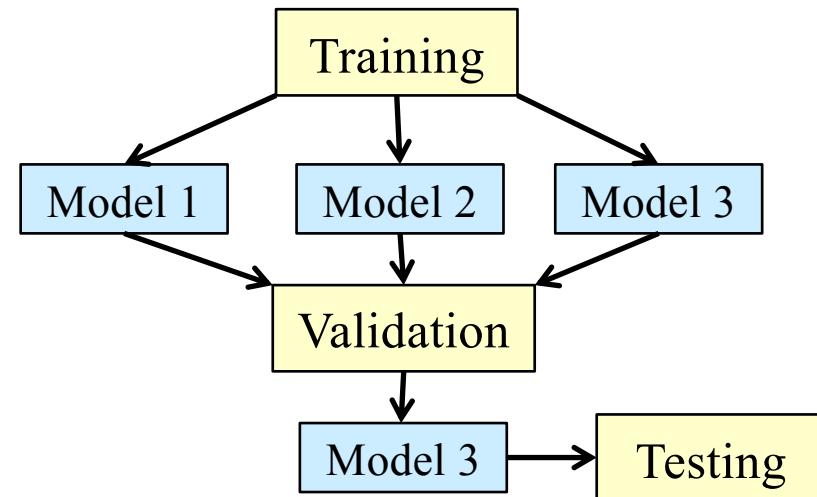
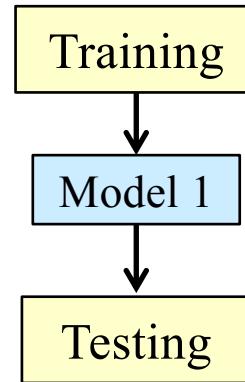


What is machine learning?

- 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>

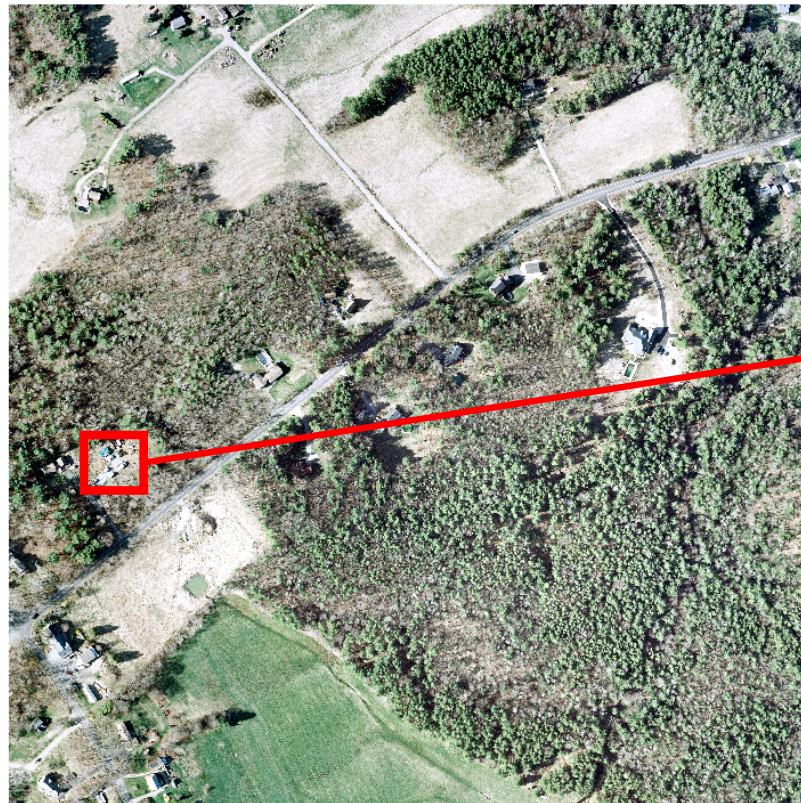
The screenshot shows the USGS TNM Download (V1.0) interface. On the left, there's a sidebar with 'Advanced Search Options' and a 'Product Search Filter' section. The 'Product Search Filter' section is highlighted with a red border and contains the following options:

- All Subcategories
- National Land Cover Database (NLCD) - 2001
Show Preview
- National Land Cover Database (NLCD) - 2006
Show Preview
- National Land Cover Database (NLCD) - 2011
Show Preview

Below this, under 'Data Extent', is a radio button for '3 x 3 degreee'. To the right of the sidebar is a map of the Southeastern United States, showing states like Illinois, Kentucky, Tennessee, and Georgia. Major cities and interstate highways are labeled. A blue rectangular box highlights a specific area on the map, likely indicating the search extent. At the bottom of the map, there are links for Accessibility, FOIA, Privacy, and Policies.

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/>

USGS
science for a changing world

EarthExplorer - Home

Page Expires In 1:48:41

Home Save Criteria Load Favorite Manage Criteria

Item Basket (0) yingjiehu RSS Feedback Help

Search Criteria | Data Sets | Additional Criteria | **Results** | Clear Criteria

4. Search Results

If you selected more than one data set to search, use the dropdown to see the search results for each specific data set.

Show Result Controls

Data Set Click here to export your results »

High Resolution Orthoimagery

5 Entity ID:849158_005 Acquisition Date:01-MAY-06 State:ME

6 Entity ID:849159_006 Acquisition Date:01-MAY-06 State:ME

7 Entity ID:849160_006 Acquisition Date:01-MAY-06 State:ME

Map Satellite (42° 40' 15" N, 078° 16' 20" W) Options Overlays

Search Criteria Summary (Show)

Map Satellite (42° 40' 15" N, 078° 16' 20" W) Options Overlays

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>

The screenshot shows the official website of the United States Census Bureau. At the top left is the logo 'United States Census Bureau'. To the right is a search bar with the placeholder 'Search' and a magnifying glass icon. Above the search bar are links for 'U.S. Department of Commerce | Blogs | Index A-Z | Glossary | FAQ'. Below the search bar is a horizontal menu bar with the following items: Topics (Population, Economy), Geography (Maps, Products), Library (Infographics, Publications), Data (Tools, Developers), Surveys/Programs (Respond, Survey Data), Newsroom (News, Blogs), and About Us (Our Research). The 'Geography' item is highlighted in blue.

You are here: [Census.gov](#) > [Geography](#) > [Maps & Data](#) > TIGER Products

Geography

Main About Maps & Data Reference Partnerships Education Research GSS-I Contact Us

Maps & Data

▪ [Maps & Data Main Page](#)

Maps

▪ [Census Data Mapper](#)
▪ [Reference](#)
▪ [Thematic](#)
▪ [Maps Available for Purchase](#)

Data

▪ [TIGER Products](#)
▪ [Census Geocoder](#)
▪ [Partnership Shapefiles](#)
▪ [Relationship Files](#)
▪ [Gazetteer Files](#)

TIGER Products

TIGER = Topologically Integrated Geographic Encoding and Referencing

TIGER products are spatial extracts from the Census Bureau's MAF/TIGER database, containing features such as roads, railroads, rivers, as well as legal and statistical geographic areas. The Census Bureau offers several file types and an online mapping application. Our products are:

- [TIGER/Line Shapefiles - New 2016 Shapefiles](#)
- [TIGER/Line Geodatabases](#)
- [TIGER/Line with Selected Demographic and Economic Data](#)
- [Cartographic Boundary Shapefiles](#)
- [KML - Cartographic Boundary Files](#)
- [TIGERweb](#)

25 Years and Counting

- [TIGER Story Map \(Part 1\)](#)
- [Happy 25th Anniversary, TIGER](#)

US Census American Community Survey (ACS data)

<https://data.census.gov/>

Explore Census Data

The Census Bureau is the leading source of quality data about the nation's people and economy.

 Find Tables, Maps, and more ...

SEARCH

[Advanced Search](#) [② Help](#) [Feedback](#)

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.



National Hydrography



HOME

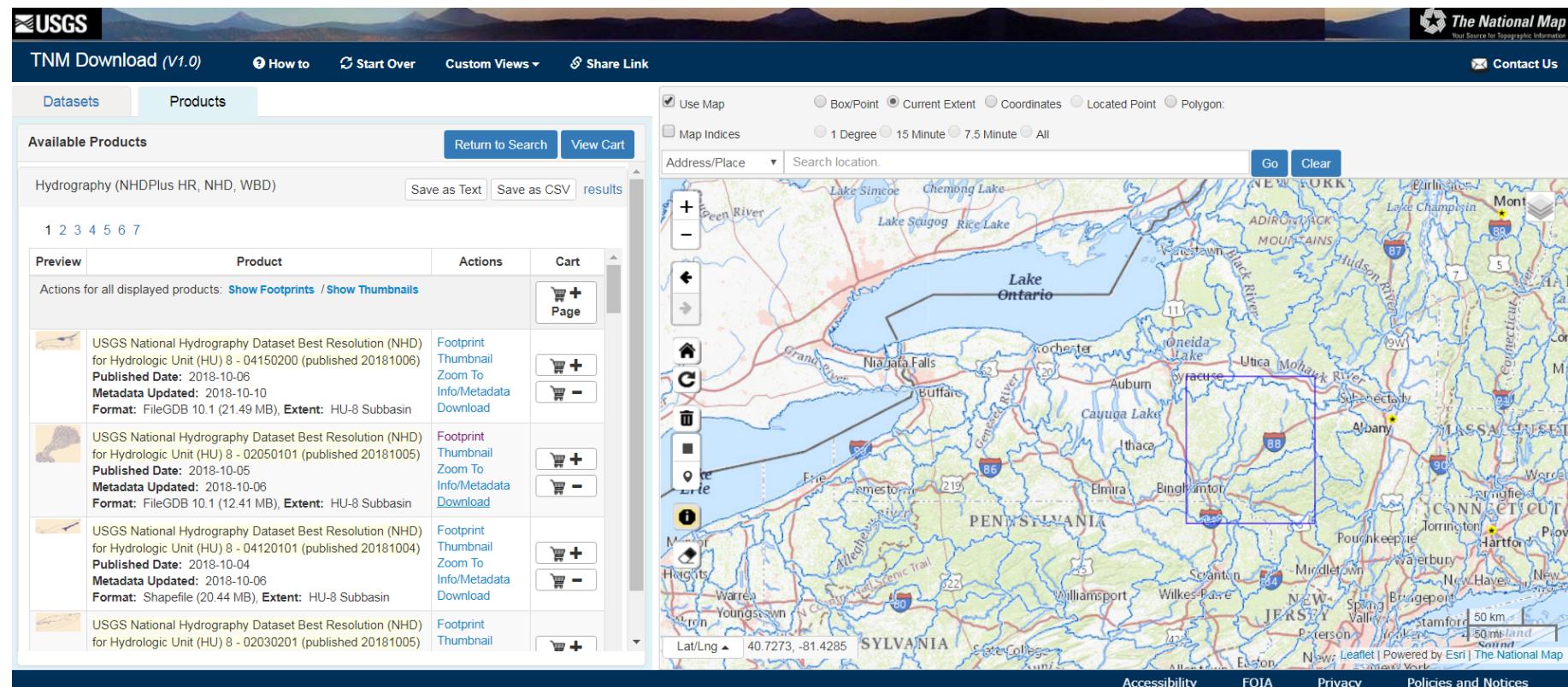
ABOUT NATIONAL
HYDROGRAPHY
PRODUCTS

National Hydrography Dataset

The National Hydrography Dataset (NHD) represents the water drainage network of the United States with features such as rivers, streams, canals, lakes, ponds, coastline, dams, and streamgages. The NHD is the surface water component on the [US Topo](#) map product produced by the USGS. These data, in digital vector geographic information

Obtaining NHD

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



The screenshot shows the USGS TNM Download (V1.0) interface. At the top, there are links for "TNM Download (V1.0)", "How to", "Start Over", "Custom Views", and "Share Link". On the right, there is a "Contact Us" link and the "The National Map" logo. The main area has tabs for "Datasets" and "Products". Below that is a section titled "Available Products" with a "Hydrography (NHDPlus HR, NHD, WBD)" category. A "results" button is visible next to "Save as Text" and "Save as CSV". A navigation bar below shows page numbers 1 through 7. To the right is a map of the Great Lakes region, specifically the Lake Ontario area, with various rivers, lakes, and roads labeled. A legend at the bottom left of the map indicates coordinates (Lat/Lng) and a scale of 50 km. The map also features a "Zoom To" button and a "Footprint" button. The overall interface is designed for searching and downloading hydrographic datasets.

State or city data

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

The screenshot shows the NYC OpenData homepage. At the top, there's a banner with the NYC OpenData logo and a yellow callout box stating "1500+ Data Sets Available". To the right are social media icons for Facebook, Twitter, and LinkedIn, along with "Sign Up" and "Sign In" links. Below the banner, a large image of several yellow taxis on a city street serves as the background for a news article. The headline reads "Taxis, Taxis, Everywhere" and describes a dataset of taxi trip records. A link "Click here to view Taxi Trip Data." is provided. At the bottom of the page is a search bar and a link to the dashboard.

Taxis, Taxis, Everywhere

This data set contains information on the millions of trips taken by New York City's taxis on an annual basis. Records include pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. [Click here to view Taxi Trip Data.](#)

View More Stories

Search

Click here to view the NYC OpenData dashboard

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

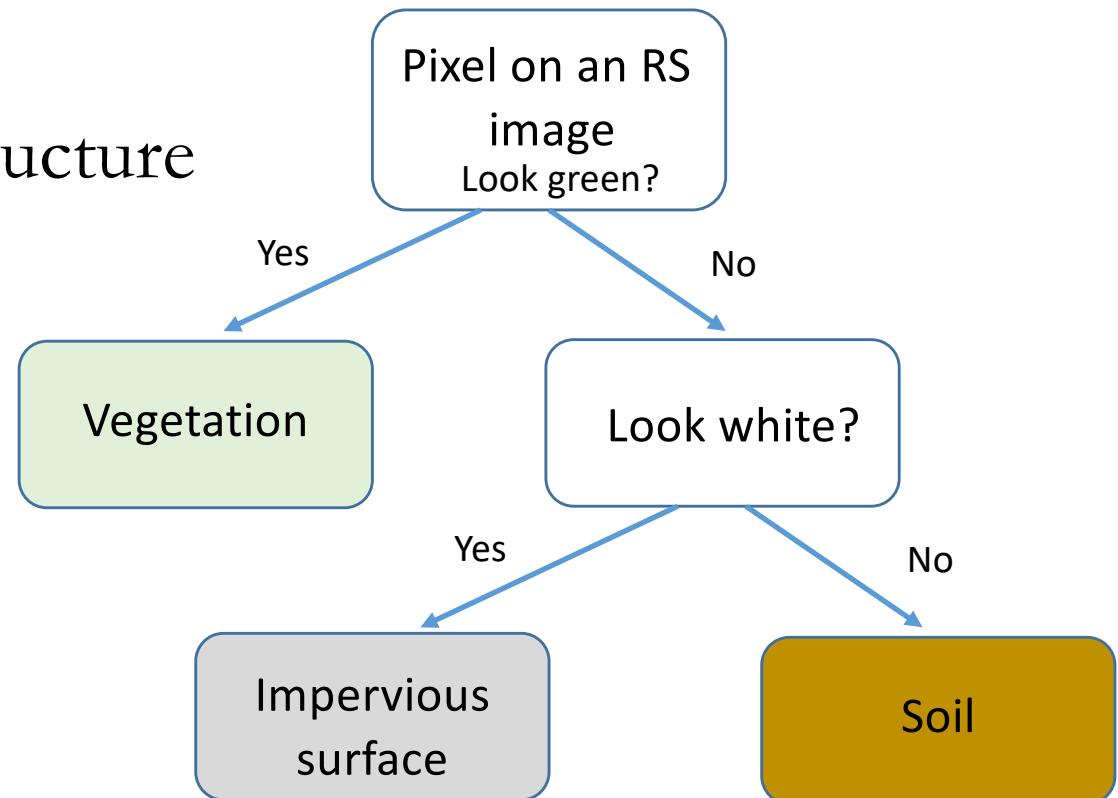
- Developing new models by including spatial relations (e.g., spatial weights); **you are changing the existing model**
- Model development oriented and relatively fewer studies compared with applications
- The challenges are in understanding the core mechanism of a model (understanding some of the math 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

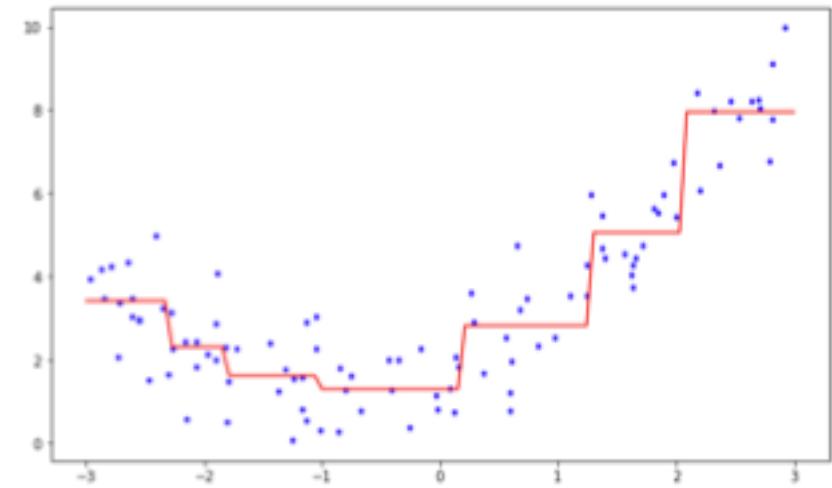
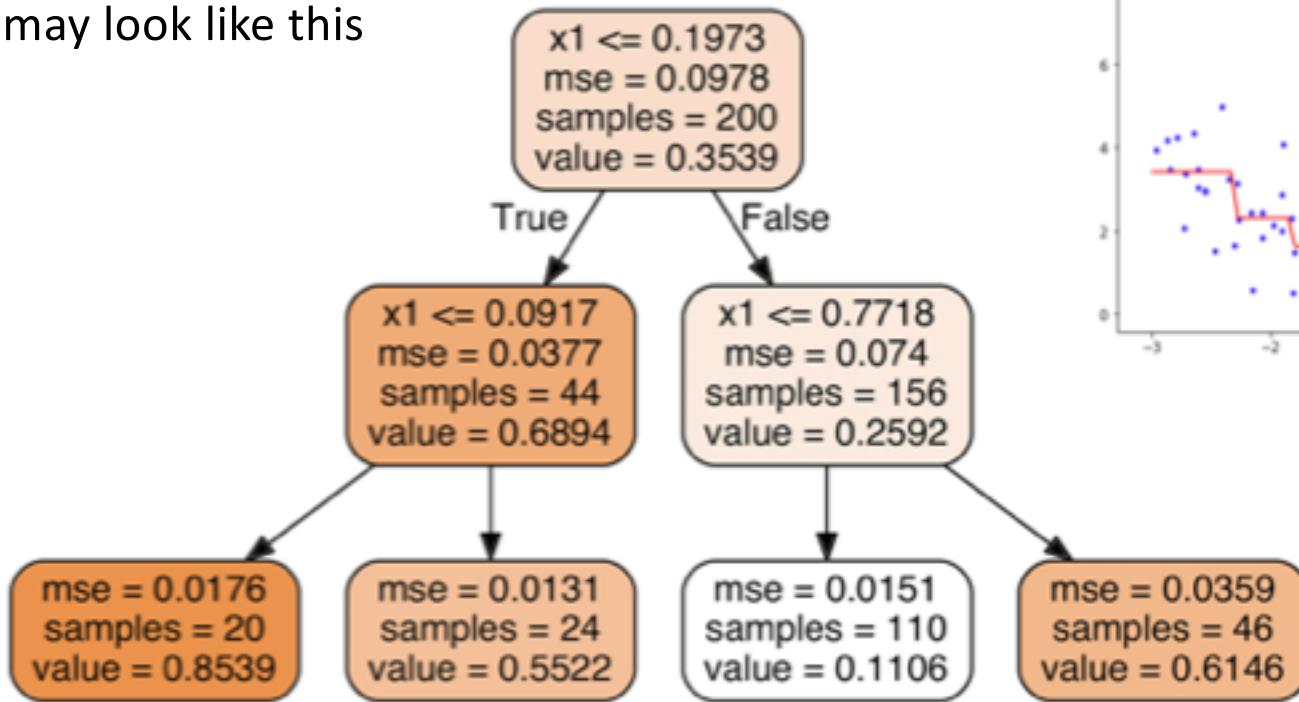
- Training a model means building this tree using training data



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

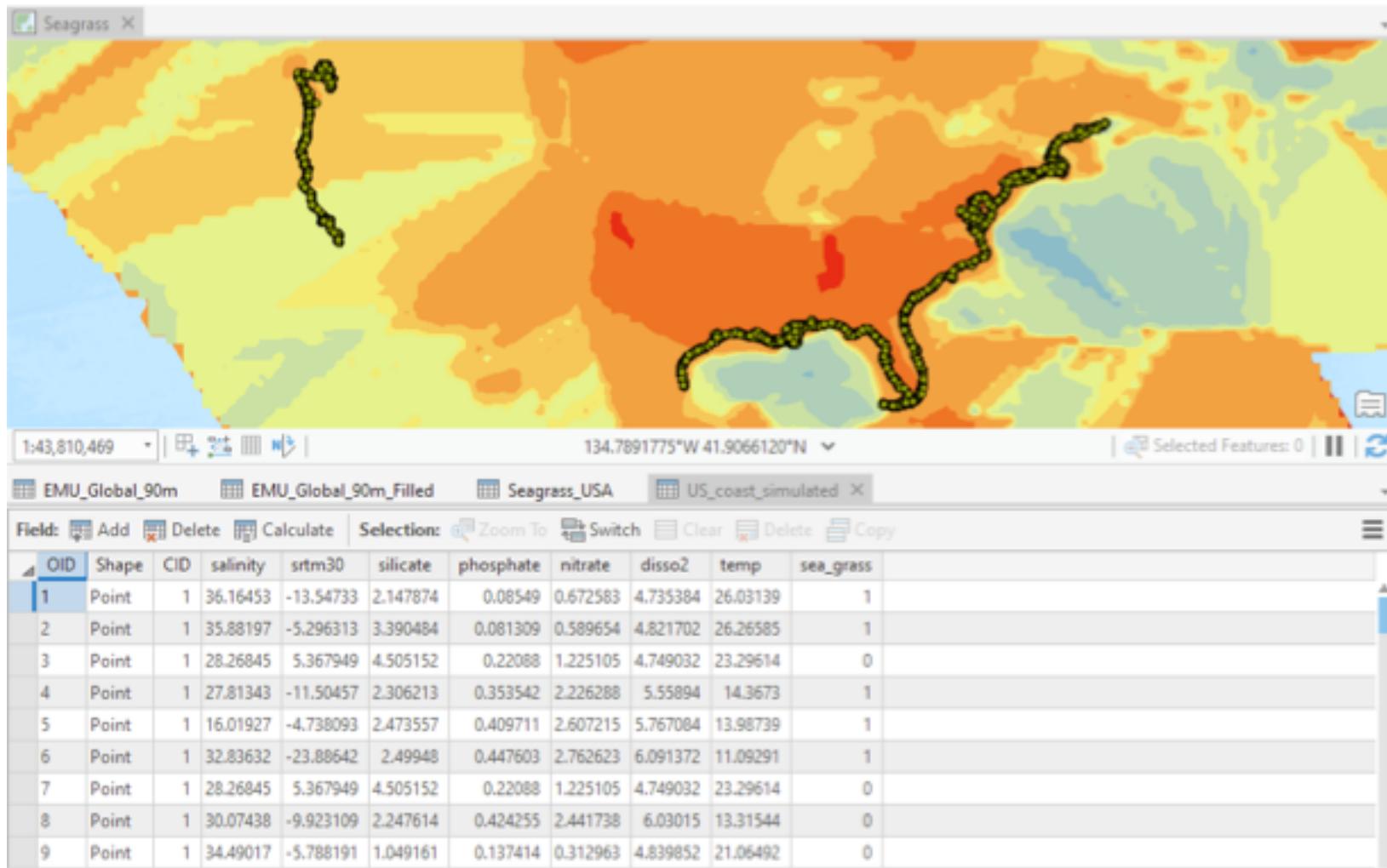
- Ensemble learning: train multiple models for the same problem (based on the same training dataset), and then aggregate their predictions
- The wisdom of the crowd: ask many people about the same question, and then aggregate their answers
- Random forest: an ensemble of decision trees (e.g., 100 decision 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



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