Bangkok | Manila | Singapore

Use of Data Analysis and AI Application for ANR Sector Projects in Cambodia

January 10, 2023

Thinking Machines
Data Science

Thinking Machines is a technology consultancy building AI & data platforms to solve high-impact problems

WHAT WE OFFER

Fully customizable enterprise data solutions

- **Core Product Solutions**
  Data Platforms, Location Intelligence, Document Intelligence, Customer Intelligence
- Building enterprise data platforms and customizing AI solutions

SELECT CLIENTS AND PARTNERS

UNDP · unicef INNOVATION · Globe · NDRRMC · ADB · UNILAB · eastwest

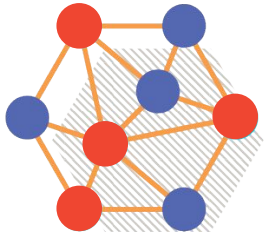
WHAT WE DO

We support clients in their data transformation journeys

We make our clients successful by offering full stack data support; from setting up enterprise-grade data platforms, data system integration, AI/data strategy consulting, building custom AI, to offering capacity building solutions.

Data Platforms

Enterprise Data Warehouse that democratize data access and lay the foundation for AI



Custom AI

Operationalization of leading edge AI through frameworks and leveraging our GeoML, DocAI, Customer Analytics product suite

Capacity Building

Organizational development and scaling of workforce fluency through consulting, hands-on training, and coaching

GEOSPATIAL INTELLIGENCE

Building **Custom Maps** and **Data Platforms** to support Data-Driven Decisions

We specialize in building custom maps using satellite imagery, open data, and computer vision

Infrastructure

Natural Resources

Socioeconomic

Session Facilitators

Ren Flores
Geospatial Intelligence Analyst

Joshua Cortez
Machine Learning Consultant

About Our Instructors

Ren Flores
Geospatial Intelligence Analyst

- Develops geospatial data management and analysis on open data for planning and decision making
- Developed and productionalized geospatial machine learning models for social development and sustainability applications
- Pursuing a Masters degree in Geomatics Engineering

About Our Instructors

Joshua Cortez
Machine Learning Consultant

- Developed and productionalized geospatial machine learning models for social development and sustainability applications
- Built geospatial data pipelines for sustainability and telecommunication organizations
- Worked on customer analytics for banks

Workshop Overview

1. Project Intro: Rice Yield Estimation
2. Open Data for Social Good
3. Geospatial Analysis in Python
4. Introduction to Machine Learning
5. Introduction to Computer Vision

Introduction to GeoML

Day 1 Overview

Instructor: Ren Flores

1. Project Intro: Rice Yield Estimation with Satellite Imagery and Machine Learning
    a. Background
    b. Methods
    c. Results and Learnings
2. Open Data for Social Impact
    a. What is Open Data
    b. Access and Sources
    c. Google Earth Engine
3. Geospatial Analysis in Python
    a. GIS in Python
    b. Vectors
    c. Rasters

Workshop Objectives

By the end of the workshop, participants should:

- Understand GeoML is through its application on Rice Yield Estimation
- Learn about Open Data and how it can be used
- Contextualize areas where open data can help with participant's work
- Familiarize with the basics of GIS
- Able to navigate the basic spatial predicates in Python
- Understand how to interpret and process rasters

01

Rice Yield Estimation with Satellite Imagery and Machine Learning

Project Background

Thinking Machines developed an ML model to estimate the crop yield data from satellite imagery based on survey data conducted on 390 households in Cambodia in 2020.

Challenge

Monitoring agricultural productivity through surveying households across large areas can **take months to complete** and can **cost millions of dollars**

Solution

We used o**pen-source geospatial data** and **satellite imagery** features that represent plant health and different environmental factors to assess how we can **predict rice yield using ML**

Impact

50x

MORE PLOTS WITH RICE YIELD DATA

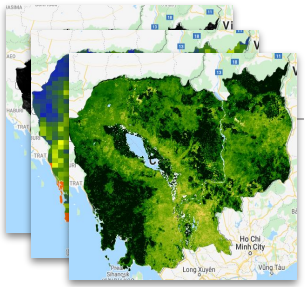
Extended rice yield data to over 50x more plots using machine learning, with an **accuracy of +/- 0.56 tons per ha**

Data Acquisition | Data Preparation | Modeling | Final Output

Target Features

Satellite Features

Plant Health Features



Create ML ready training data from the household survey and farm plot data

Collect open data that can help the model infer crop yield

Experiment on different model algorithms

Benchmark model performance and compare results using the training data

Prediction on unsurveyed households using the best performing model

Result analysis to understand limitations and possible future improvements

Overview of Model Development

Our goal is to train a model using household survey and open satellite data that predicts rice yield at a plot level

The model is built on data composed of...

Rice Yield

Target Feature

- Household rice yield per season in tons per hectare
- Farm plots of each household

Next Step:

Match each farm plot to the correct yield per season

Open Data

Independent Features

- Sentinel-2 images and vegetation indices
- Environmental indicators

Next Step

Get the value of each dataset for each plot

Machine learning models use a dataset with a target label: the value being predicted, and independent features:

End Goal: Create a plot level dataset with crop yield and independent features

Satellite Data

Independent features

12 Sentinel-2 Bands
7 Vegetation indices

Environmental Data

Independent features

Evapotranspiration
Slope
Soil Surface Moisture
Soil Subsurface Moisture
Land Surface Temperature
Total Precipitation

- The model is trained on a total of 92 features
- The model used mean, minimum, maximum and standard deviation of each dataset per season

End Goal: Create a plot level dataset with crop yield and independent features

Target Features

Create ML ready training data from the household survey and farm plot data

Satellite Features

Collect open data that can help the model infer crop yield

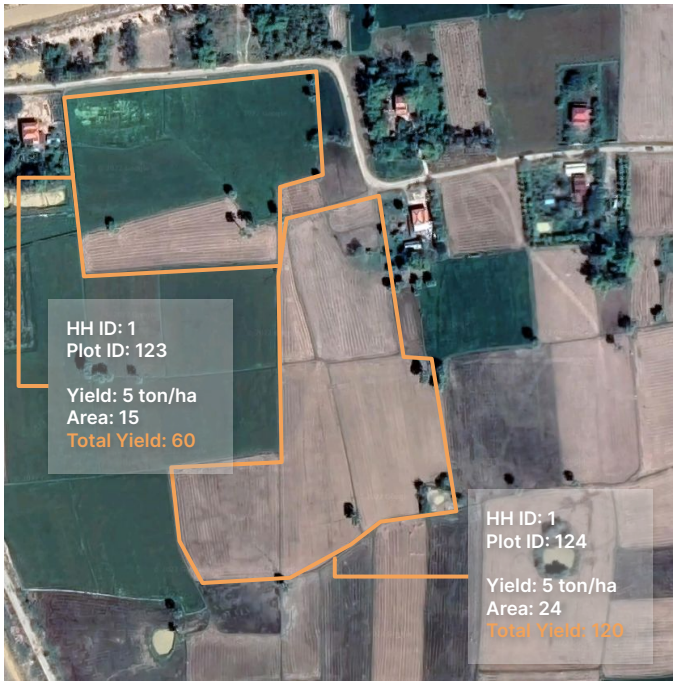
Plant Health Features

Collect open data that can help the model infer crop yield

Plot ID	NDVI	...	Temperature	Soil Moisture
1	0.8		35	20
2	0.7		38	21

All plots of a household are assigned a yield value in the dataset, including *uncultivated plots*

Multiple plots belonging to the same household have the same yield per hectare despite not being cultivated



- Uses household ID to match the two datasets
- Household level survey had only one yield size per season despite each household owning multiple plots
- There are instances where only parts of the plot are cultivated which leads to overestimating total yield
- There are plots that were not cultivated but would still have a yield size

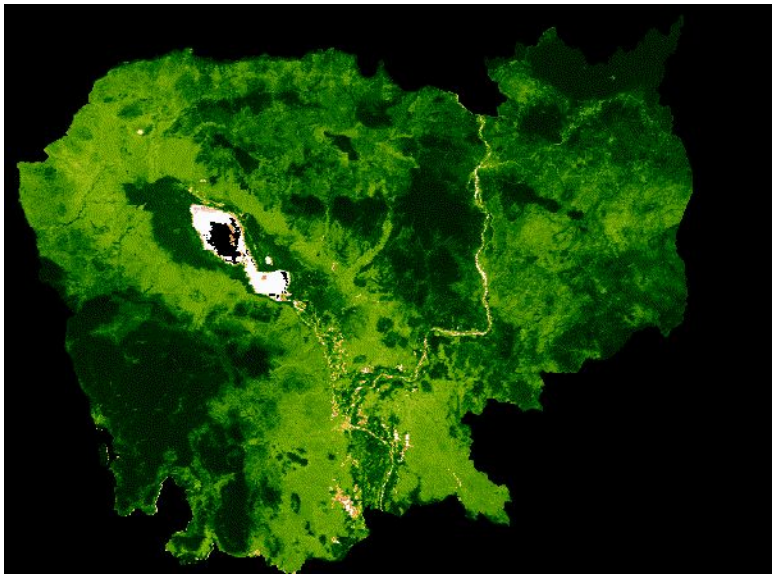
We used vegetation indices to identify uncultivated plots

Vegetation indices are satellite derived measures to indicate presence of plants and was used to identify uncultivated plots.

This minimized the amount of plots with erroneous yield labels

We used a series of satellite images to indicate crop growth and health for modelling

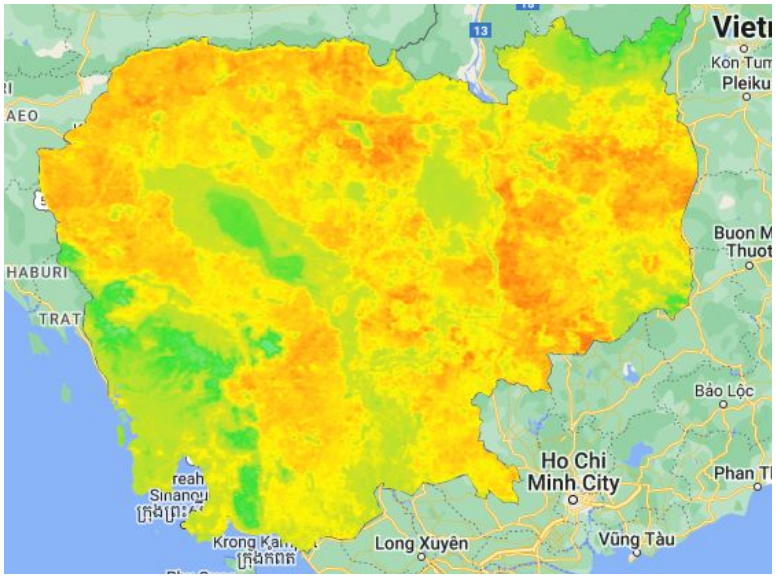
The model uses Sentinel-2 bands and the derived vegetation indices as independent features



- We get statistical aggregates per season to indicate plant growth progression
- Vegetation indices (VI) derived from satellite bands indicate the presence of plants, it's age, health and stress.
- i.e. a higher VI corresponds to a healthy plant

Adding environmental indicators that possibly impact rice productivity

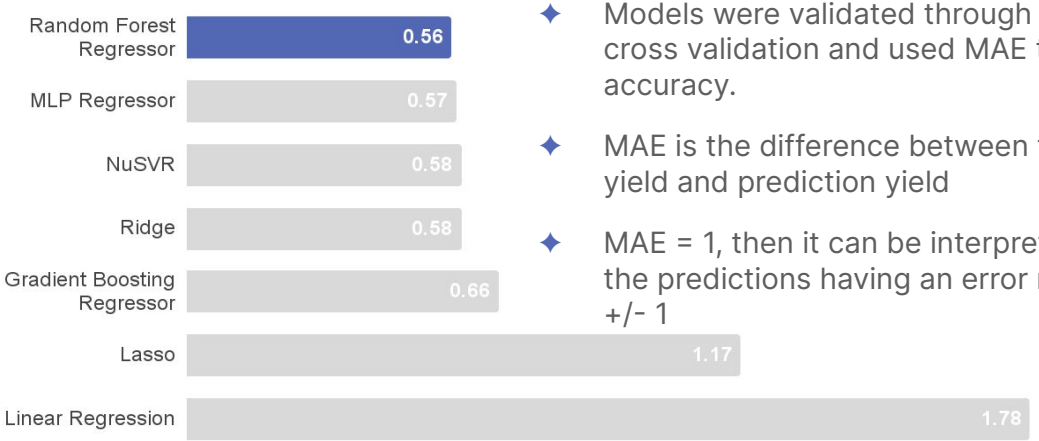
These factors are based on related literature and existing similar work



- We get statistical aggregates per season to indicate the environmental shifts in temperature, soil moisture, etc.
- This helps us identify the growing conditions of the area
- i.e. we have the minimum, maximum, average and standard deviation of the surface temperature for the dry season

Random Forest Regressor is the best model based on multiple experiments with an MAE of 0.56 tons/ha

We testing and fine-tuned different ML algorithms to identify which has the best performance



- Models were validated through grouped cross validation and used MAE to assess accuracy.
- MAE is the difference between the actual yield and prediction yield
- MAE = 1, then it can be interpreted as the predictions having an error margin of +/- 1

We used the best model to **predict yield for plots without survey data**

We used the **trained and fine-tuned model** on unsurveyed plots since it is the **most accurate** as well as the **most efficient** based on the It is also the most efficient model.

Random Forest Regressor is the best model based on multiple experiments with an MAE of 0.56 tons/ha

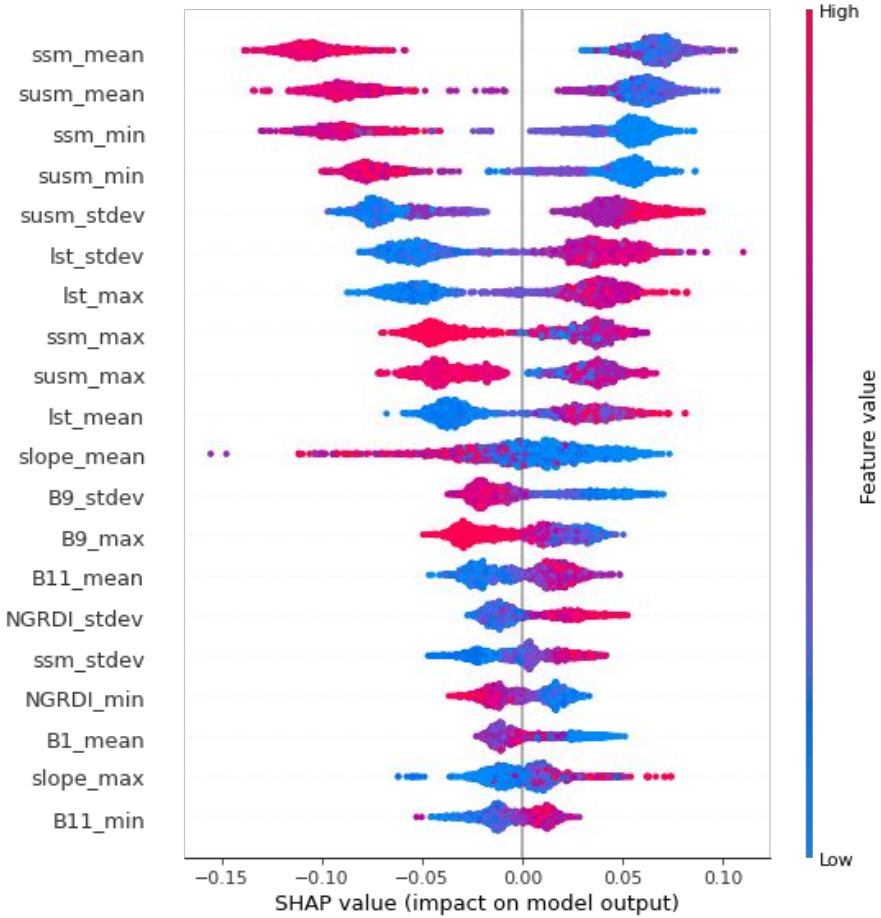
MAE is the difference between the actual yield and prediction yield, and R^2 shows how much data the model can explain or predict/

0.563
Mean Absolute Error

0.468
R^2

The model can predict **values close to the actual yield** but still **has room for improvement**

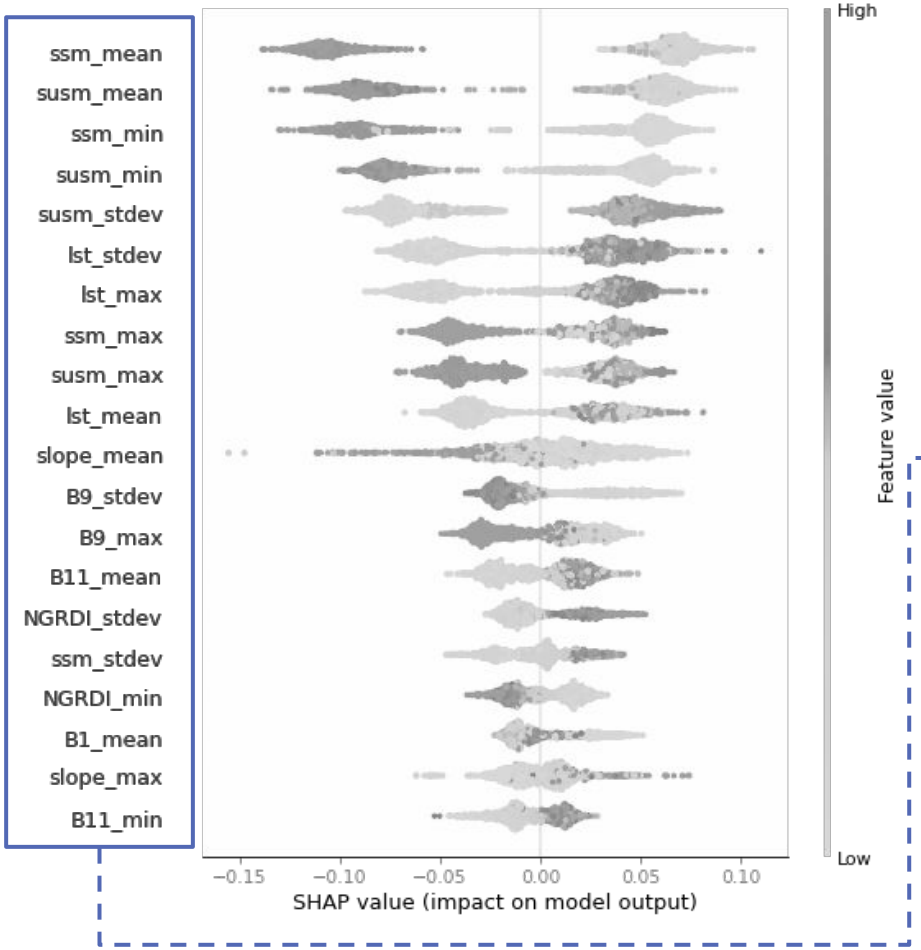
- Yield predictions are +/-0.563 tons per hectare and the average yield per plot is 3-5 tons per hectare.
- The model features are able to explain 46.8% of the variation in rice yield.



Looking at feature impact gives a better understanding of model results

We use a **beeswarm plot** to analyze how the features impact the predicted rice yield. It shows:

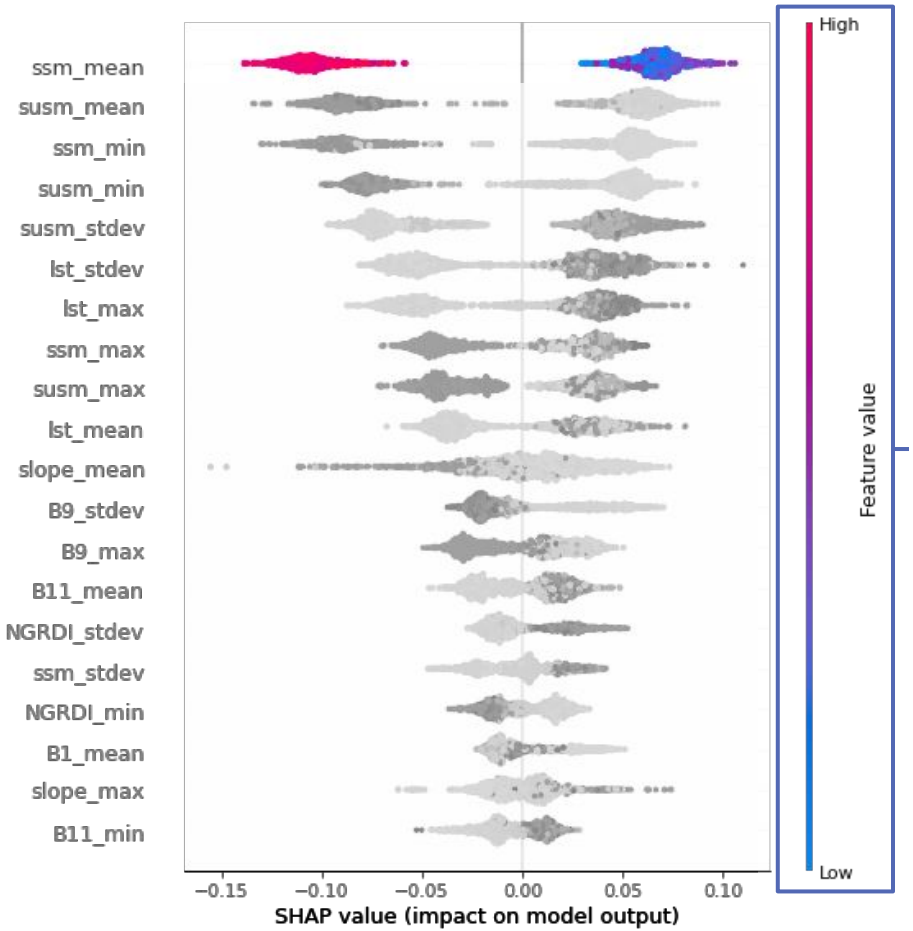
1. Feature importance
2. Feature value
3. Feature impact



Looking at feature impact gives a better understanding of model results

We use a **beeswarm plot** to analyze how the features impact the predicted rice yield. It shows:

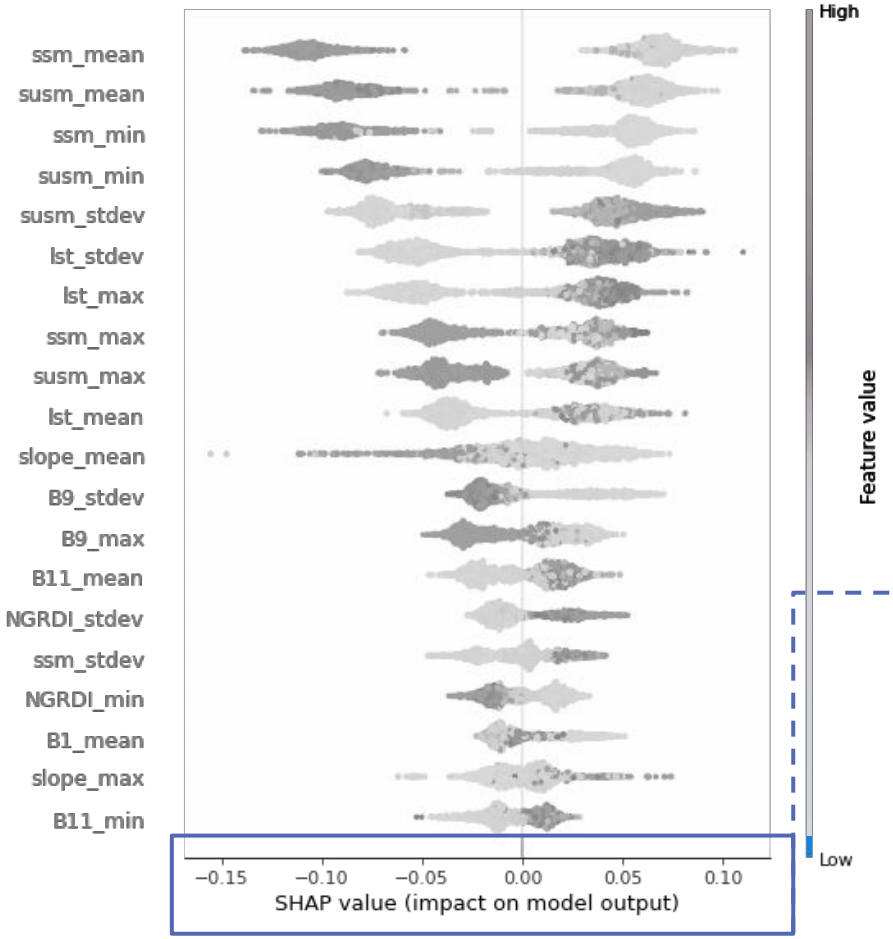
1. **Feature importance** - the features are ordered from most influential to predicted yield to the least
2. Feature value
3. Feature impact



Looking at feature impact gives a better understanding of model results

We use a **beeswarm plot** to analyze how the features impact the predicted rice yield. It shows:

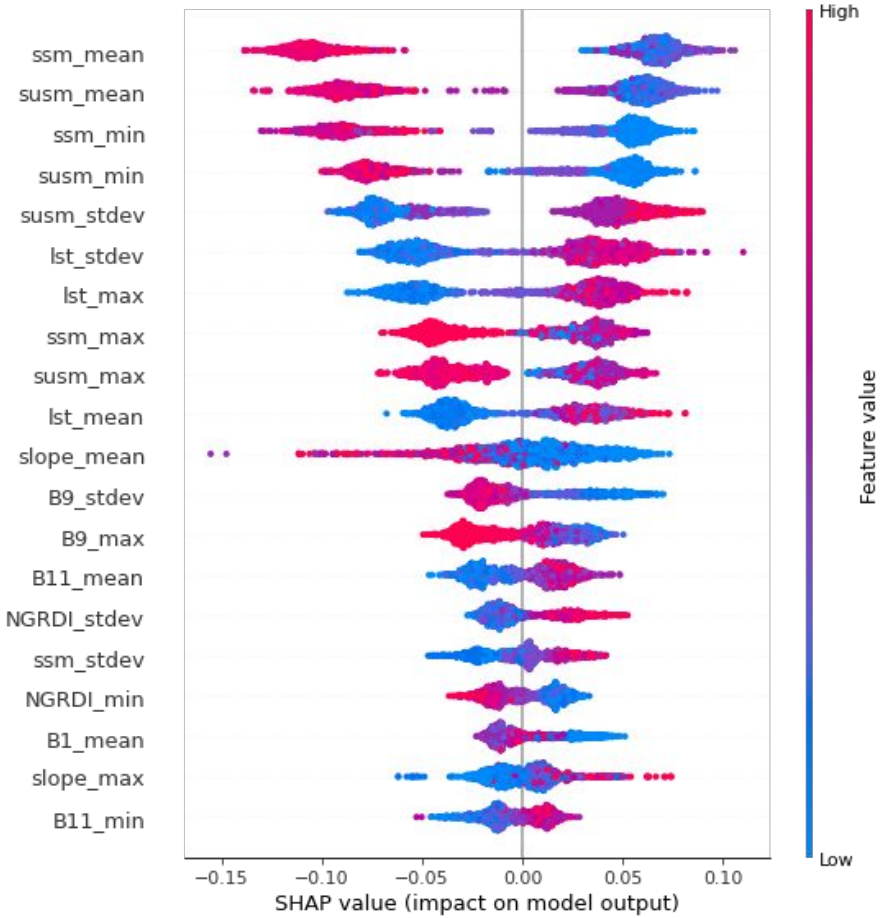
1. Feature importance
2. Feature value - the color of each dot shows if the value of the feature is high or low
3. Feature impact



Looking at feature impact gives a better understanding of model results

We use a **beeswarm plot** to analyze how the features impact the predicted rice yield. It shows:

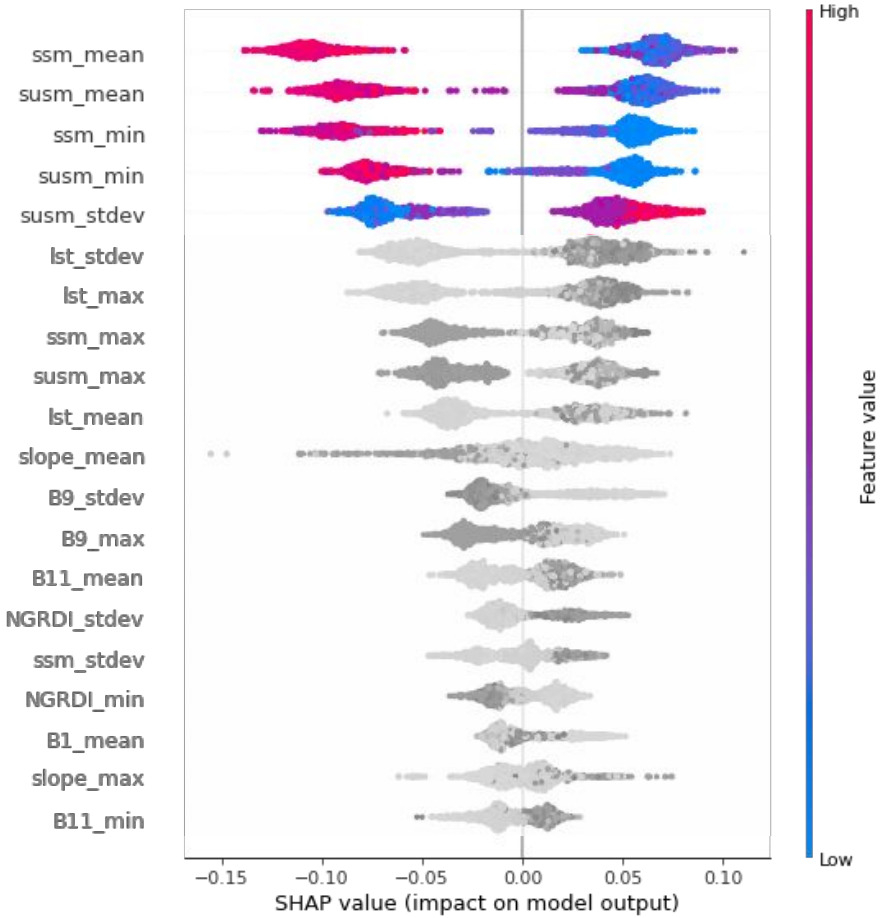
1. Feature importance
2. Feature value
3. Feature impact - the position of the dot shows if the feature translates to higher or lower predicted rice yield



We see that the top features are from the environmental data

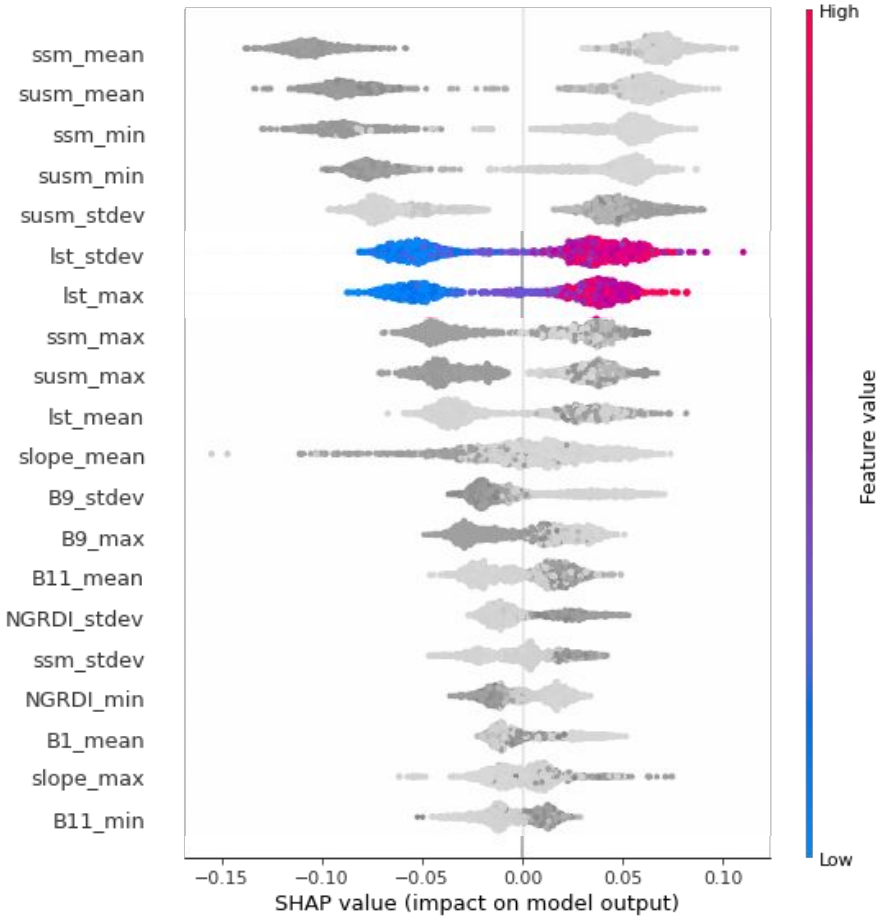
Recurring values are:

1. Soil surface moisture (ssm)
2. Soil subsurface moisture (susm)
3. Evapotranspiration (ET)
4. Land Surface Temperature (lst)
5. Slope
6. Sentinel Band 9
7. Band 11
8. Band 1



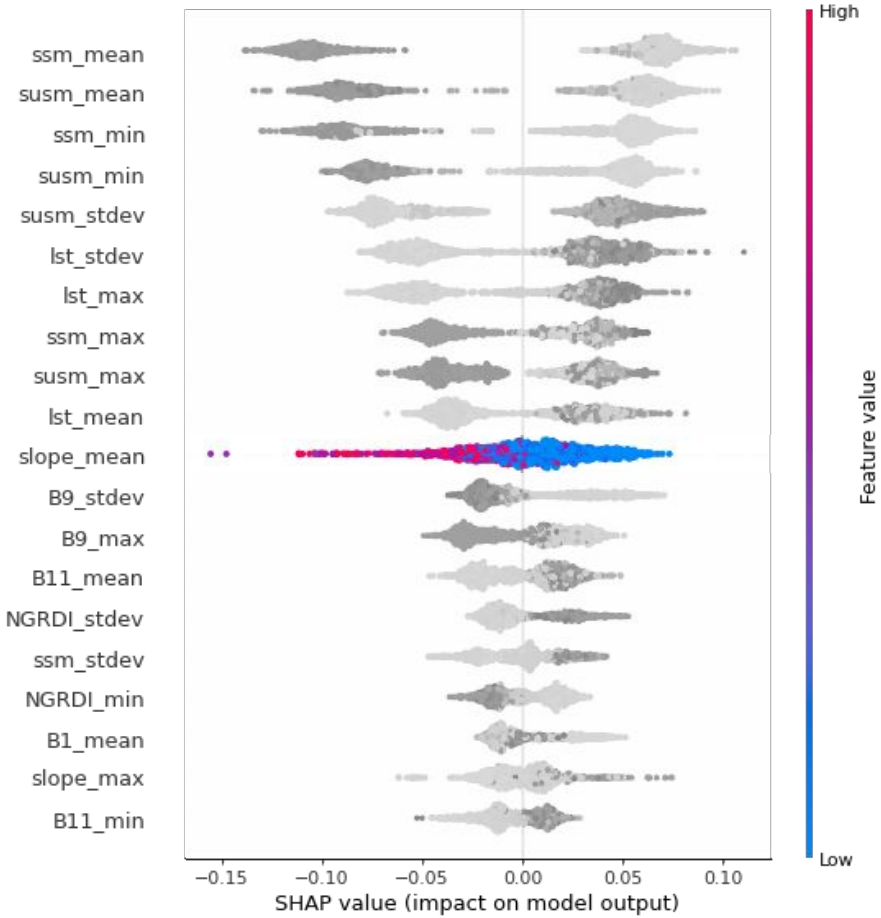
High soil moisture negatively influences the yield which can be attributed to flooded crops

- The data shows that most of the yield were lower during the wet season which can be **attributed to drop damage due to flooding**
- Soil moisture data is derived from remotely sensed data combined with ground measurements generated by NASA



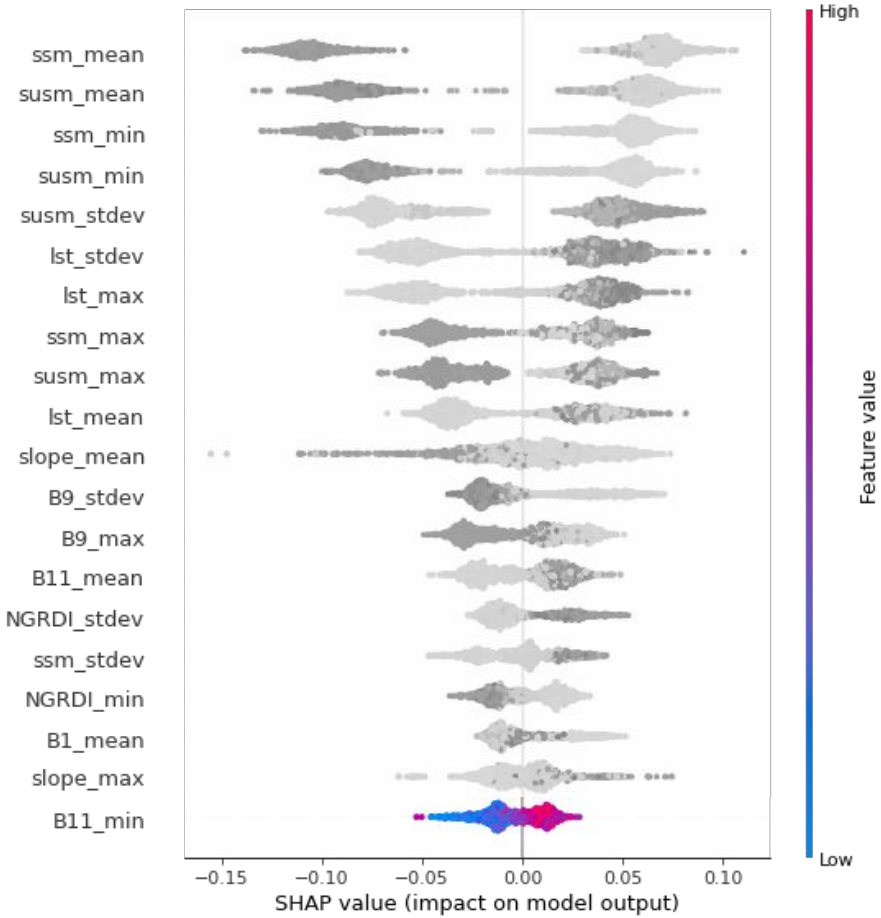
Higher land surface temperature negatively impacts yield prediction

- Higher temperature can induce plant stress and affect crop production
- Temperature is derived from MODIS daily land surface temperature generated by NASA



Lower slope (flatter land) positively impacts yield since these are areas that are easier to irrigate

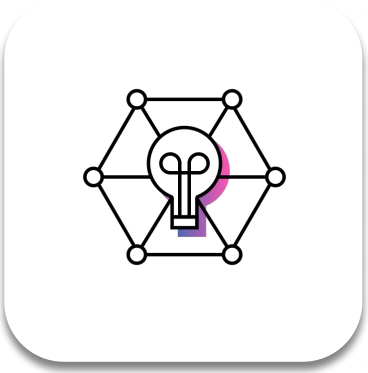
- Slope determines the rate of water run-off hence, flatter plots can be easier to flood and can retain water better
- Slope is derived from ALOS 30m DSM generated by JAXA



Band 11 is used in monitoring crop health is also indicative of higher yield

- Band 11 in Sentinel 2 is sensitive to the chlorophyll level in plants and is mainly used to detect plants and monitor its health

Next Step: Roll Out



Feature Engineering

- Clean up plots for unsurveyed rice fields
- Generate the features for the plots

Model Prediction

- Run model on unsurveyed plots

Output

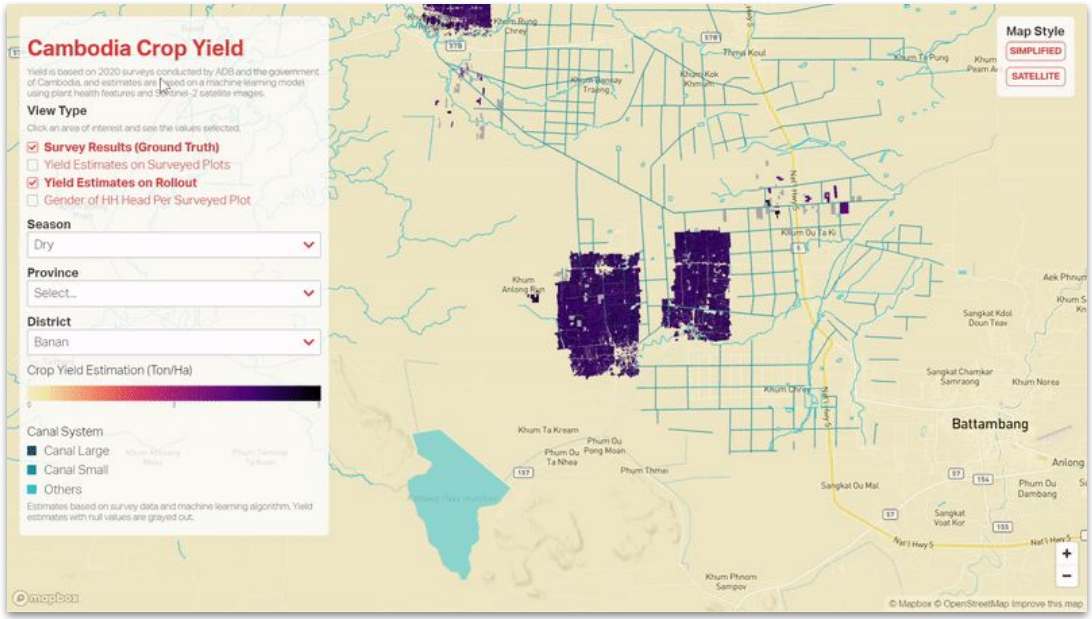
- Predicted yield for each plot in tons per hectare

The model extended survey to ~67,000 plots across three provinces

Number of plots with data increased 50x from 378 to 67,196

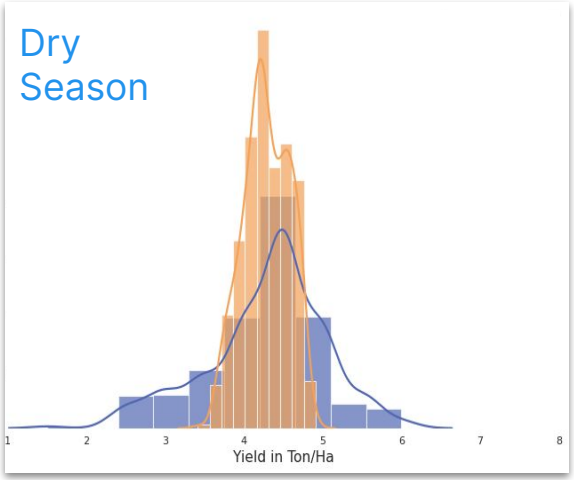
The project concluded by creating a web map to help stakeholders visualize the survey data as well as the yield prediction

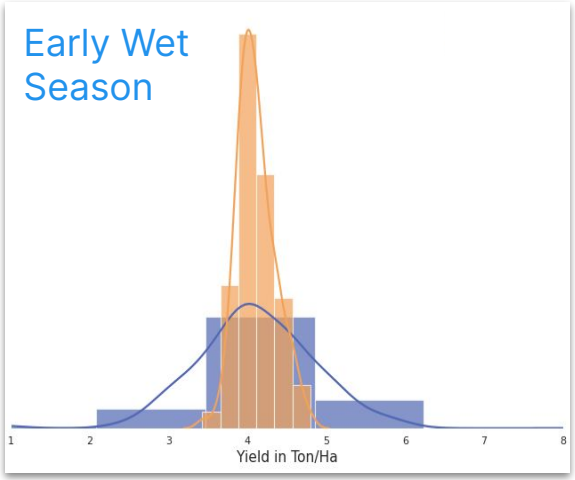
- Visualize data across provinces, filter by district and season
- Access granular and timely data with remote sensing
- Extend survey data and fill spatial gaps with ML estimates

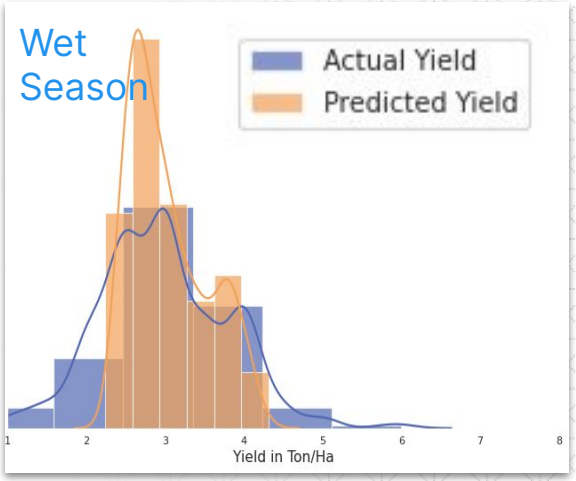


Visit the web map here: https://adb-crop-yield.web.app/

Rice Yield Predictions Perform Best in Estimating Mid-Range Yield







- Majority of samples fall within 3.5-5 tons per hectare, the same range the model predicts.
- The model has difficulty predicting outliers, this can be attributed to the data assumptions.

The project highlights the potential of ML-derived ride yield but also shows areas for improvement

We can improve performance by...

- Minimizing roll-out overestimation by identifying uncultivated plots
- Improving training data by managing the following assumptions used in processing
  - Yield information is not plot specific and repeats across multiple plots belonging to the same household
  - No information on how big the cultivated area is
  - Planting and harvest dates are arbitrary

Any questions?

Thinking Machines
Data Science

Data Stories
stories.thinkingmachin.es

Press
thinkingmachin.es/press-room

Follow us
/thinkdatasci
@thinkdatasci

Bangkok | Manila | Singapore

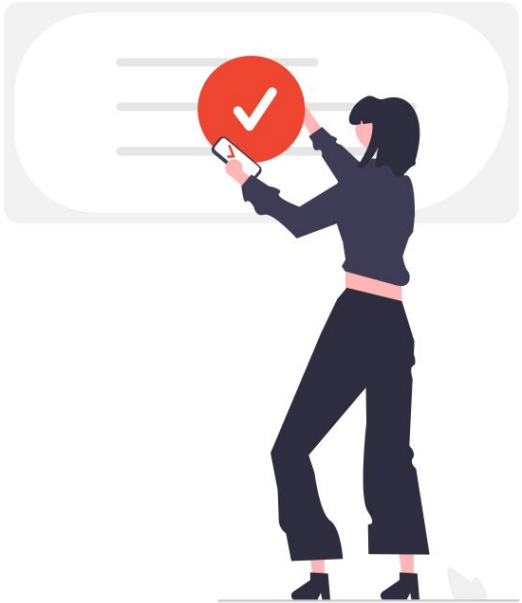
15-Minute Break

If you haven’t yet, please complete the pre-workshop set up. Scan the QR for instructions or visit this link https://bit.ly/adb-pre-workshop-setup

02 Gmail Account Set-up | Create a Google account

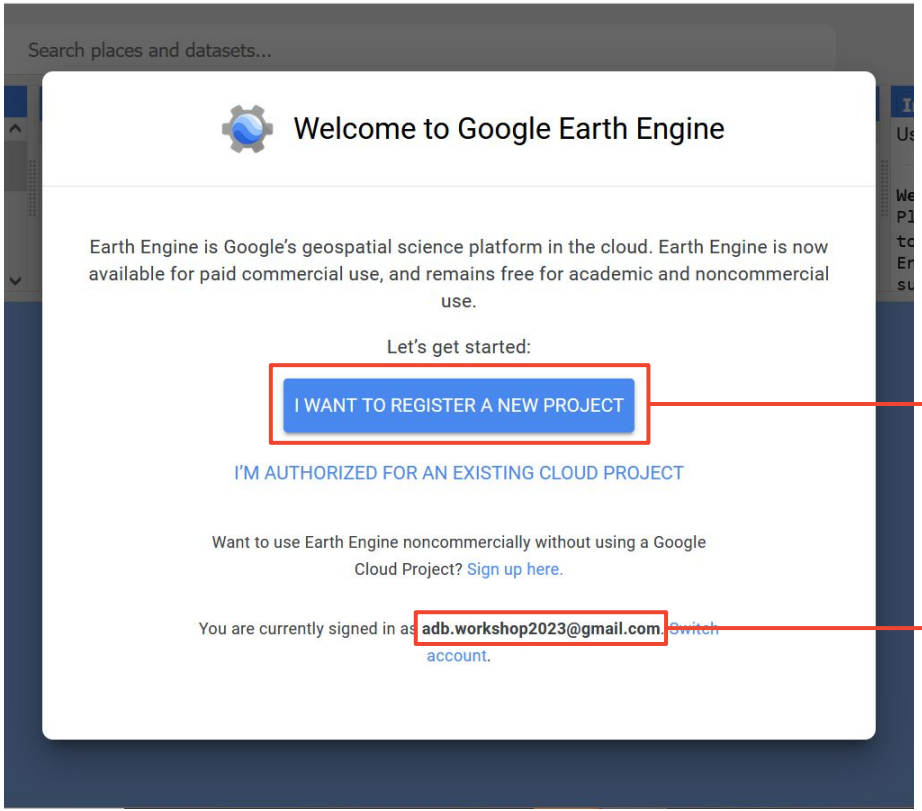
Create your own Google account here

We recommend creating an account dedicated for the workshop but you can skip this step, if you want to use your personal account.



Go to this link and select “I Want To Register A New Project”

Creating an Earth Engine Project

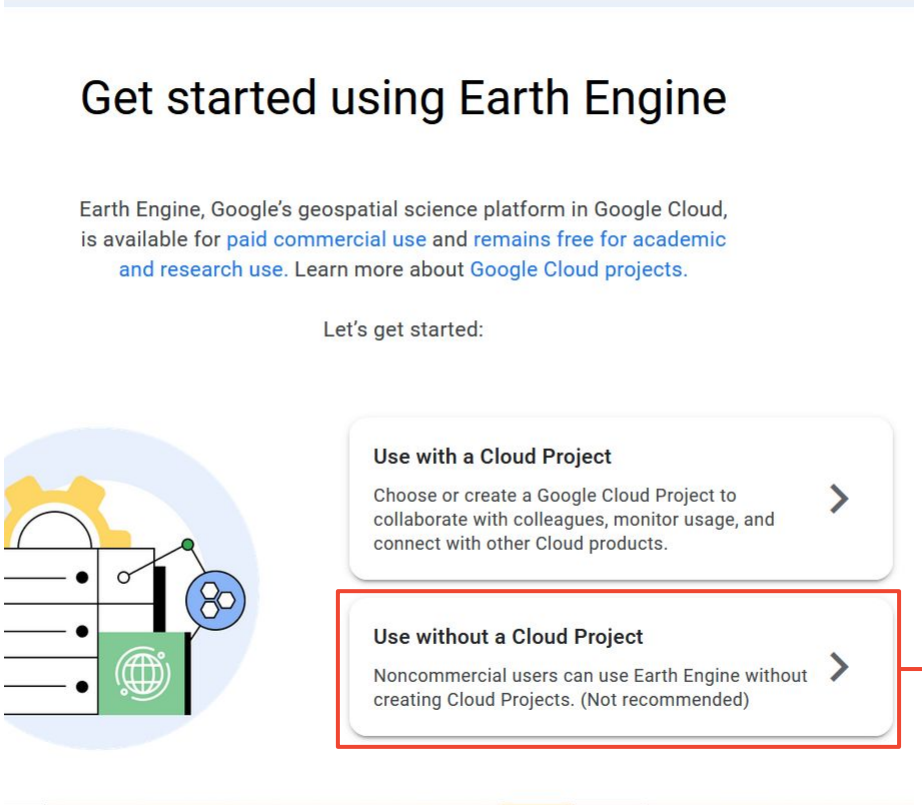


1. Select *I want to register a new project*

2. Make sure you are using the correct gmail account.

   If not you can, click switch account and select the correct one.

On the next page select “Use Without a Cloud Project”

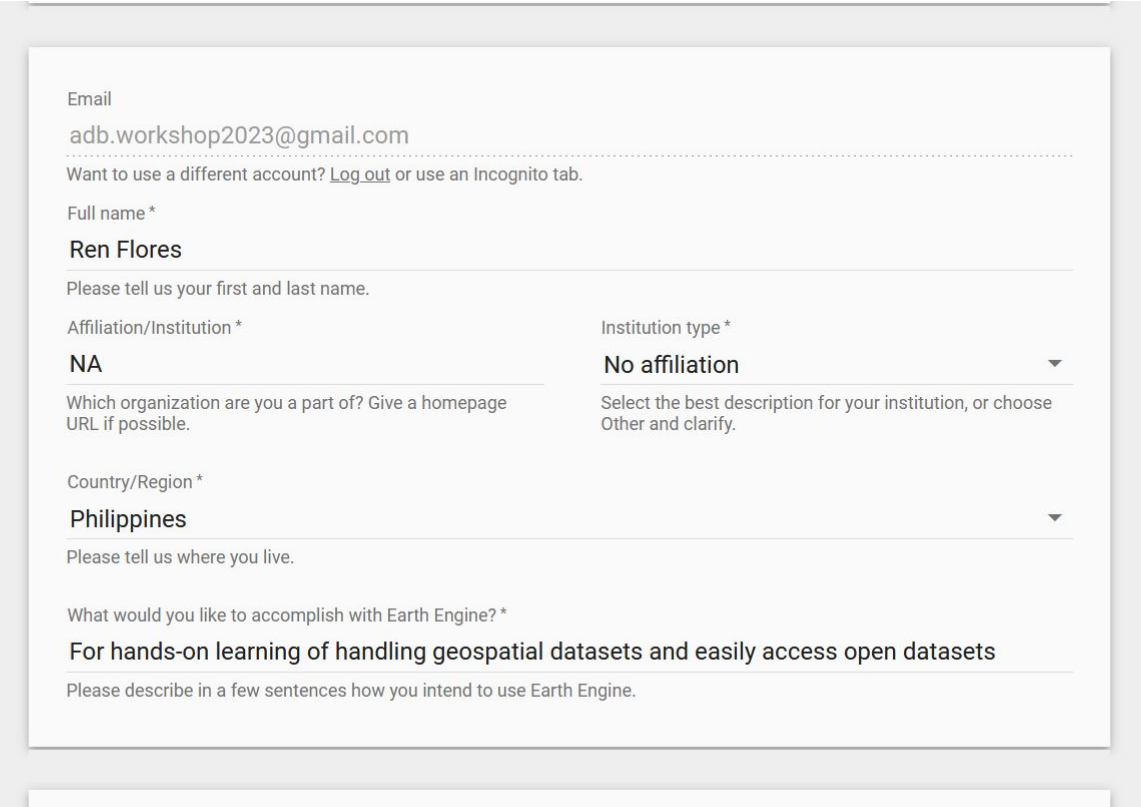


3

Select *Use Without a Cloud Project*

This allows us to use GEE without the overhead of configuring a Google Cloud Project. This is ideal only for educational purposes.

This will redirect you to a sign-up page. Complete the fields accordingly and submit.

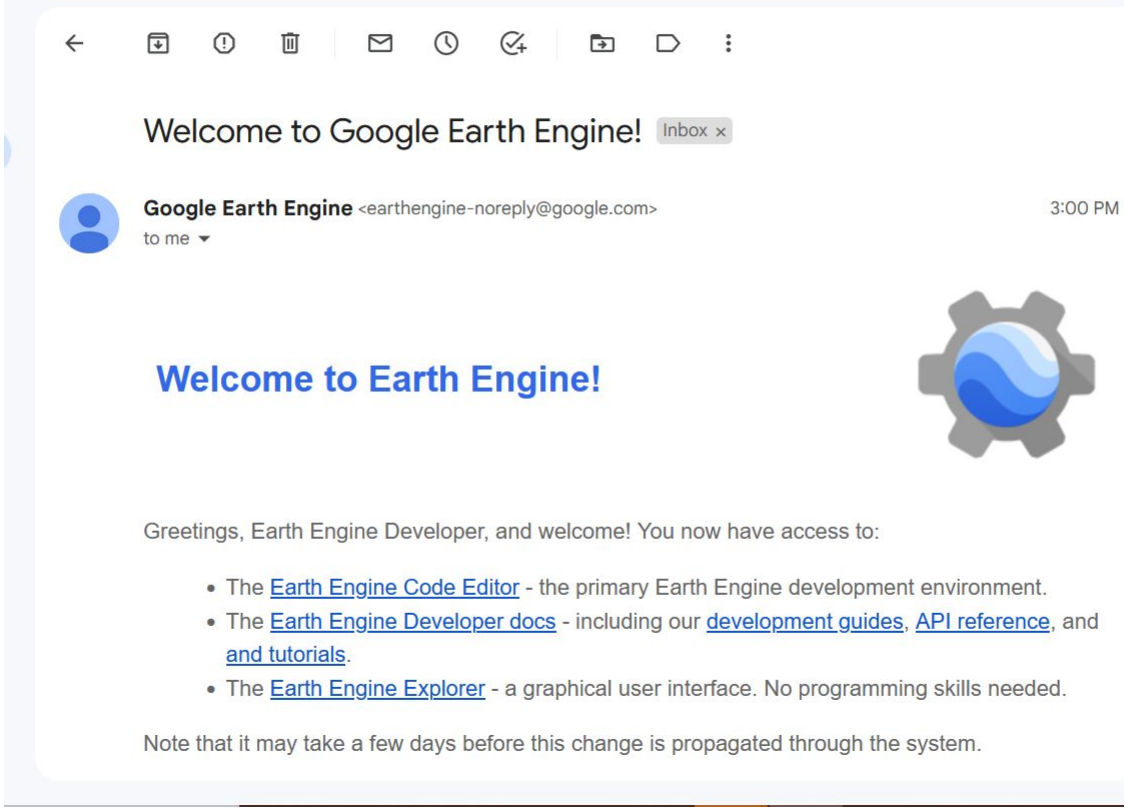


4

Sample Form for GEE Sign-up

Fill up the required fields and submit the form.

Check email for confirmation



Welcome to Google Earth Engine! Inbox ×

Google Earth Engine <earthengine-noreply@google.com> 3:00 PM
to me

Welcome to Earth Engine!

Greetings, Earth Engine Developer, and welcome! You now have access to:

- The Earth Engine Code Editor - the primary Earth Engine development environment.
- The Earth Engine Developer docs - including our development guides, API reference, and and tutorials.
- The Earth Engine Explorer - a graphical user interface. No programming skills needed.

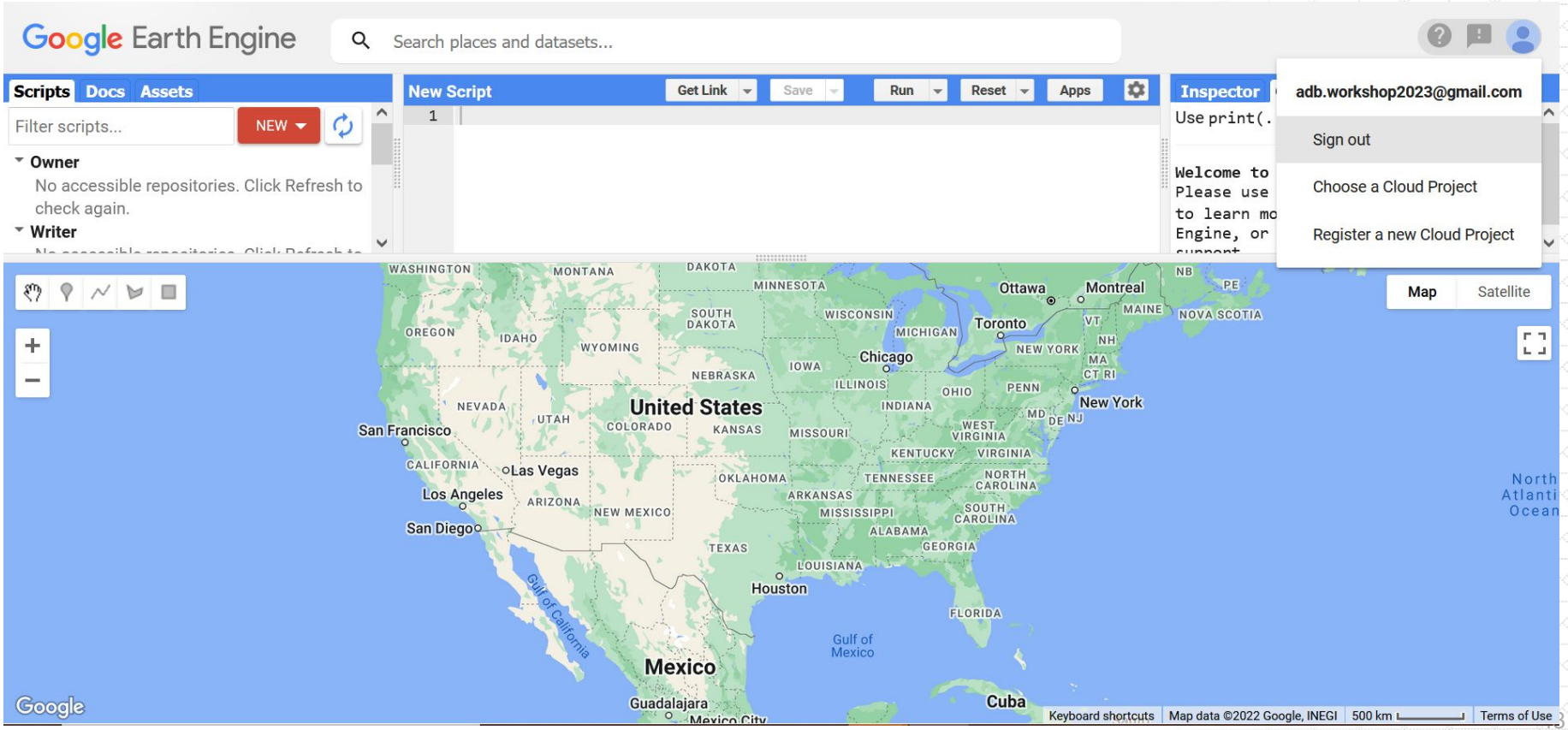
Note that it may take a few days before this change is propagated through the system.

5

You should receive a confirmation email

Check this to make sure you have access to GEE

Go back to the GEE Code Editor, and you should be able to access the page



Accessing Google Earth Engine (GEE)

GEE requires an additional registration process

The following slides provide step by step instructions to creating a GEE account but overall it is:

1. Go to this link and select *“I Want To Register A New Project”*
2. On the next page select *“Use Without a Cloud Project”*
3. This will redirect you to a sign-up page. Complete the fields accordingly and submit.
4. Check email for confirmation
5. Go back to the GEE Code Editor, and you should be able to access the page.

02

Open Data for Social Impact

Open data is data that anyone can access, use, and share

Governments, businesses and individuals can use open data to bring about social, economic and environmental benefits

Accessibility

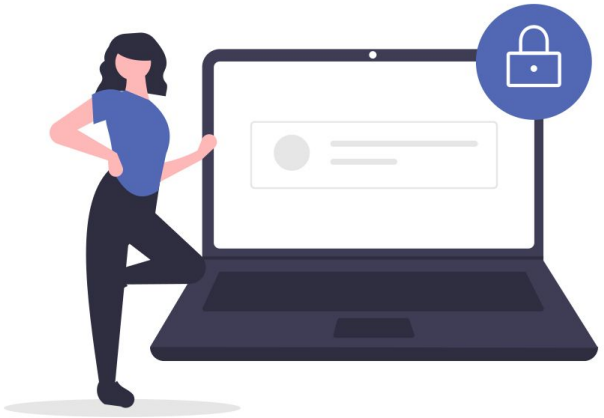
For data to be open, it should have no limitations that prevent it from being used in any particular way. Anyone should be free to use, modify, combine and share the data, even commercially

Format and Cost

Available in common, machine-readable format that can easily be integrated into an analysis and is free *to use*

Licensing

The data should be accompanied by a license that describes conditions for distribution, transformation and sharing. Open data is measured by what it can be used for, not by how it is made available.



Open data can help bring diverse benefits to governments, businesses and civil society.

Open data builds trust and transparency, support planning and decision making, and encourage community involvement

Transparency

Open Data promotes transparency, accountability and value creation by making data available to all, especially from governments and community building institutions

Decision Making

Data-driven decision making allows us to vet our understanding of the problem first before deploying valuable resources

Community

Easy access to data drives development within communities and contribute to the development of innovative services and the creation of new intellectual models

How can we use open data?

The goal is to turn data into information and information **into insight**

Raw Data

Example:
Population Density
Wealth Estimates

Context

Example:
Disaster Recovery
Assistance Program

Insight

Example:
Of all areas affected by the disaster, x district has the most population and lowest wealth and should be prioritized

Data comes to life when given context under the hands of an expertise

Open data only becomes valuable and influential when used. Its importance lies on how it impacts society and not just it's availability

Open Data Resources

Some common and credible data sources

Socioeconomic Data

Socio Economic survey data and other proxy indicators

Demographic and Health Survey Program

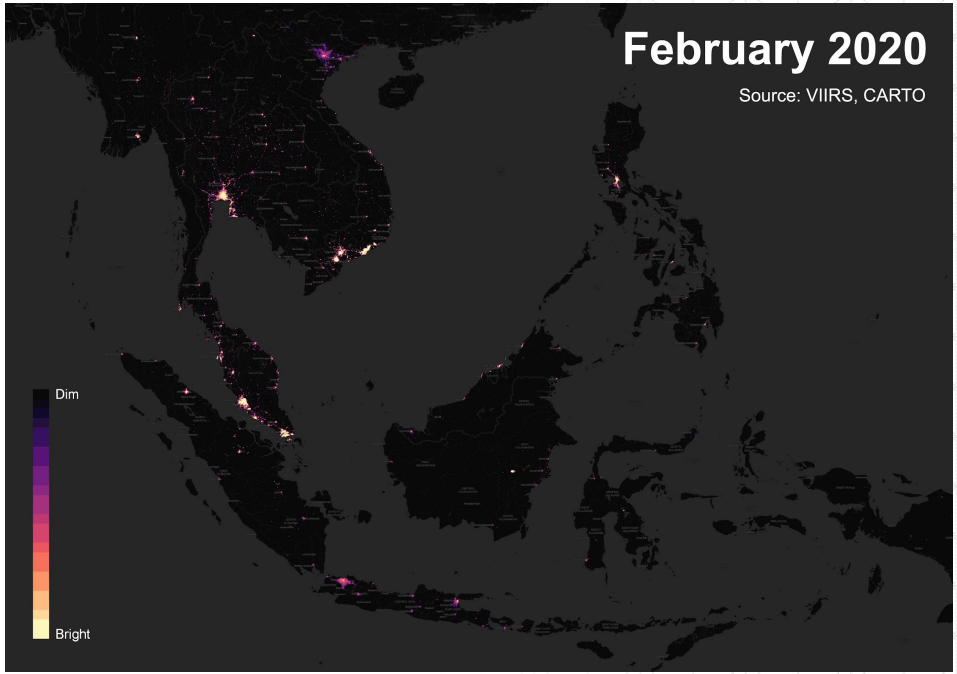
The DHS Program has collects, analyzes, and disseminates accurate and representative data on population, health, HIV, and nutrition through more than 400 surveys in over 90 countries.

Internet Speed Test

M-Lab and Ookla collect internet speed data across the globe. These are good proxy indicators for area wealth based on existing research

Night time Lights

Night time lights show the brightness of areas caused by human activity as seen from space. Multiple studies show this is a good indicator of economic activity and wealth



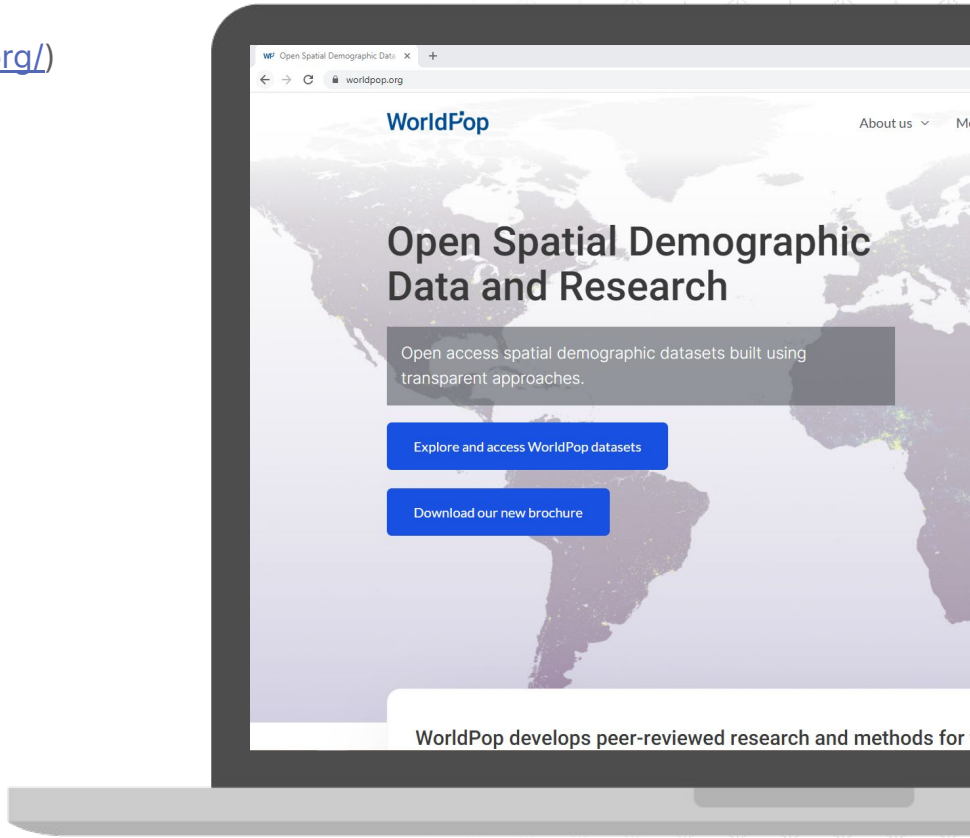
Nighttime lights in Southeast Asia during the height of COVID-19 Pandemic

WorldPop

Open Spatial Demographic Data (https://www.worldpop.org/)

WorldPop has over 45,000 dynamic, high-resolution datasets that complement traditional population data sources around the world, including:

- Administrative areas
- Development indicators
- Migration flows
- Population density
- Urban change



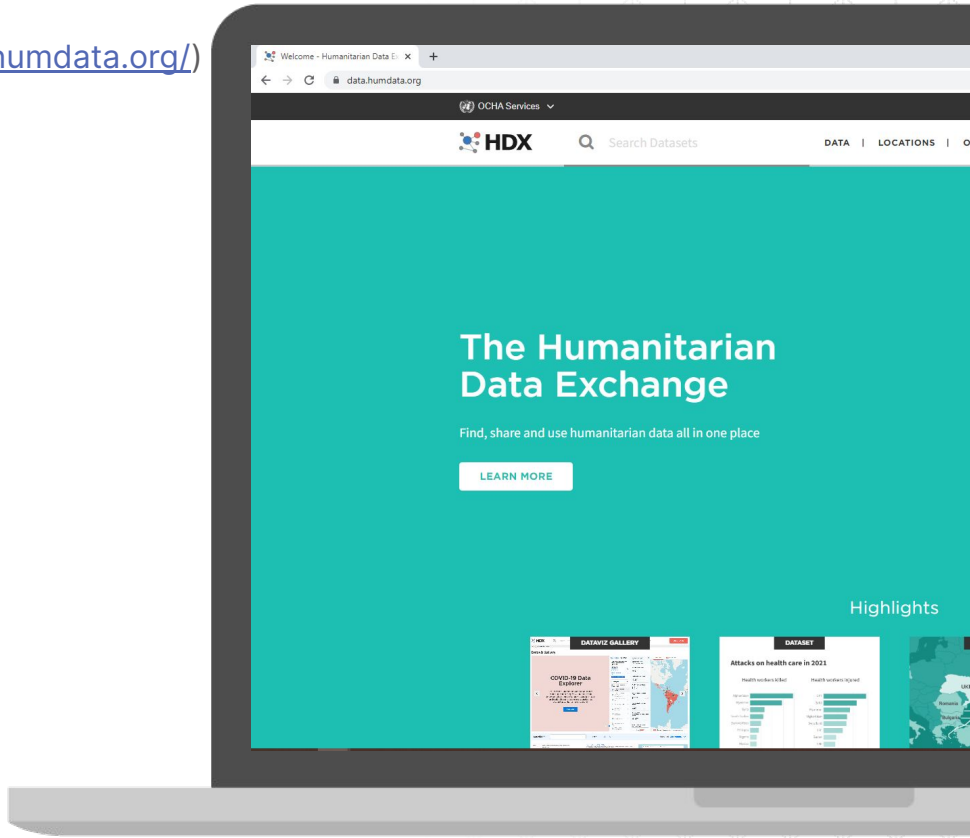
HDX
Humanitarian
Data Exchange

Data sharing platform across organisations (https://data.humdata.org/)

Managed by UN OCHA, HDX is an open platform of analysis-ready data on crises and development. Datasets cover:

- Baseline development data
- Damage assessments
- Geospatial data
- Humanitarian response

Caveat: This is sourced from different government agencies and data is formatted differently per country



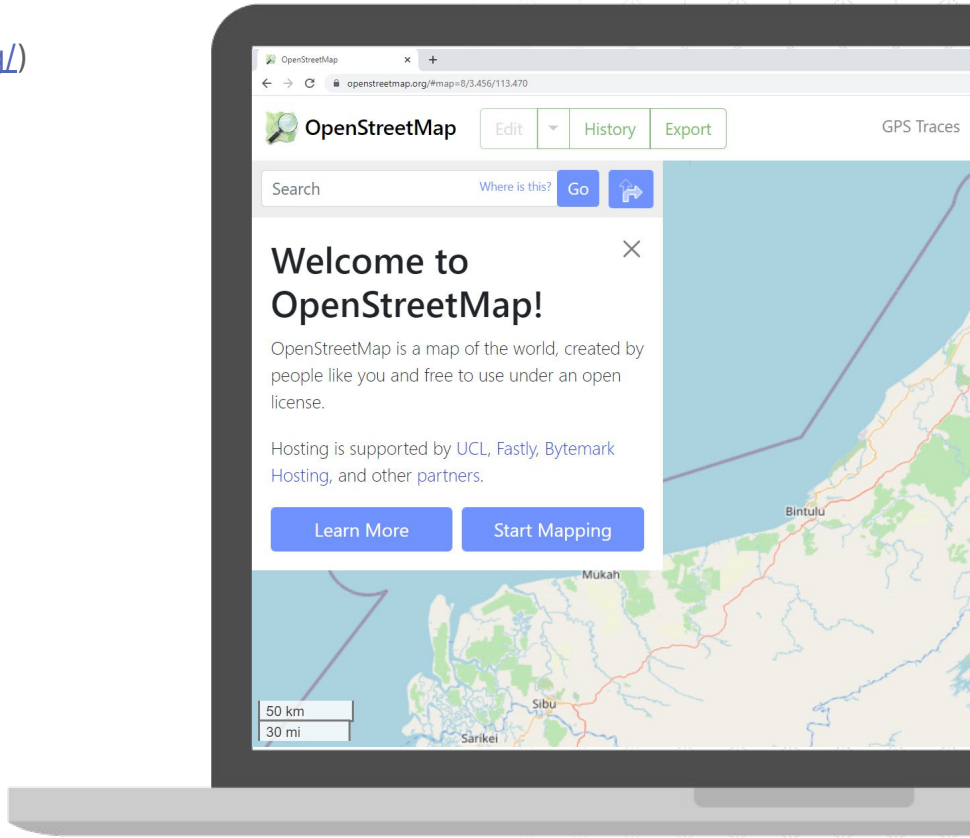
OpenStreetMap

Free, open geodatabase (https://www.openstreetmap.org/)

OSM is a crowd-sourced geospatial database of map data from around the world, covering points of interests like:

- Road networks
- Buildings and other infrastructure
- Points of interests like malls, banks, hospitals, schools, etc.

Caveat: Because OSM is crowd-sourced, completeness levels can vary per geography depending on the activity of the local communities



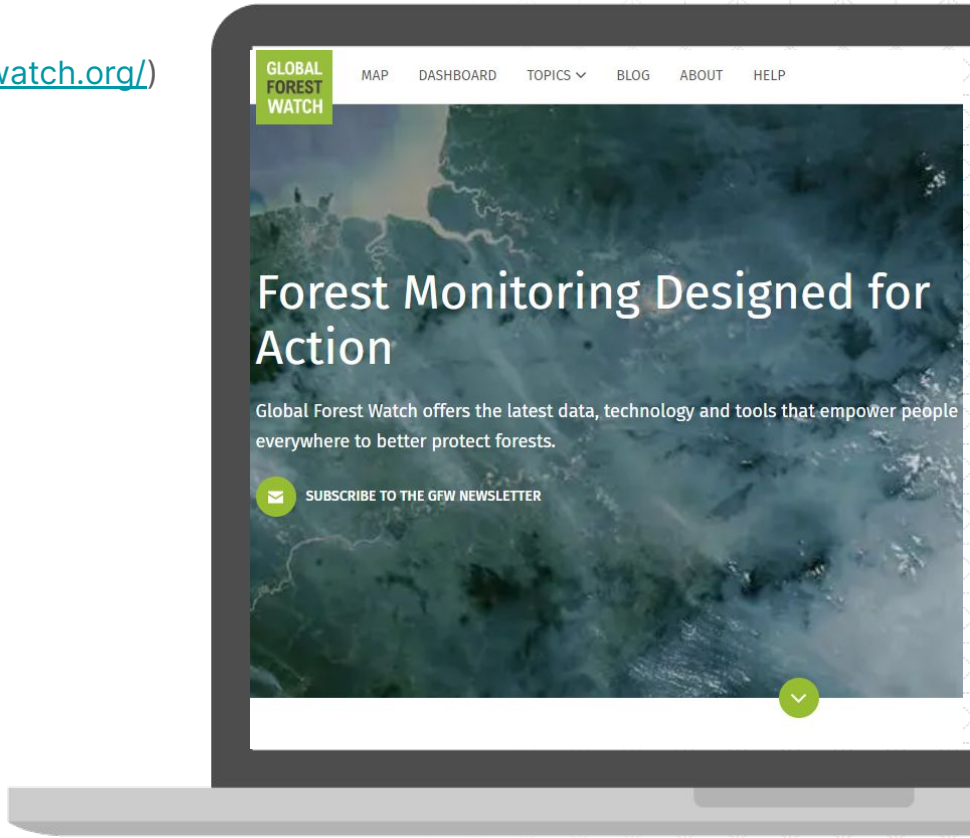
GLOBAL FOREST WATCH

Online platform for forest data (https://www.globalforestwatch.org/)

Global Forest Watch (GFW) is an online platform that provides data and tools for monitoring forests. It includes data on:

- Forest Land Cover
- Forest Change
- Forest Fires

Caveats: GFW doesn't have complete data worldwide, but it is richer in tropical regions. GFW also uses a model to get forest data estimates; it is not 100% based on ground truth data.



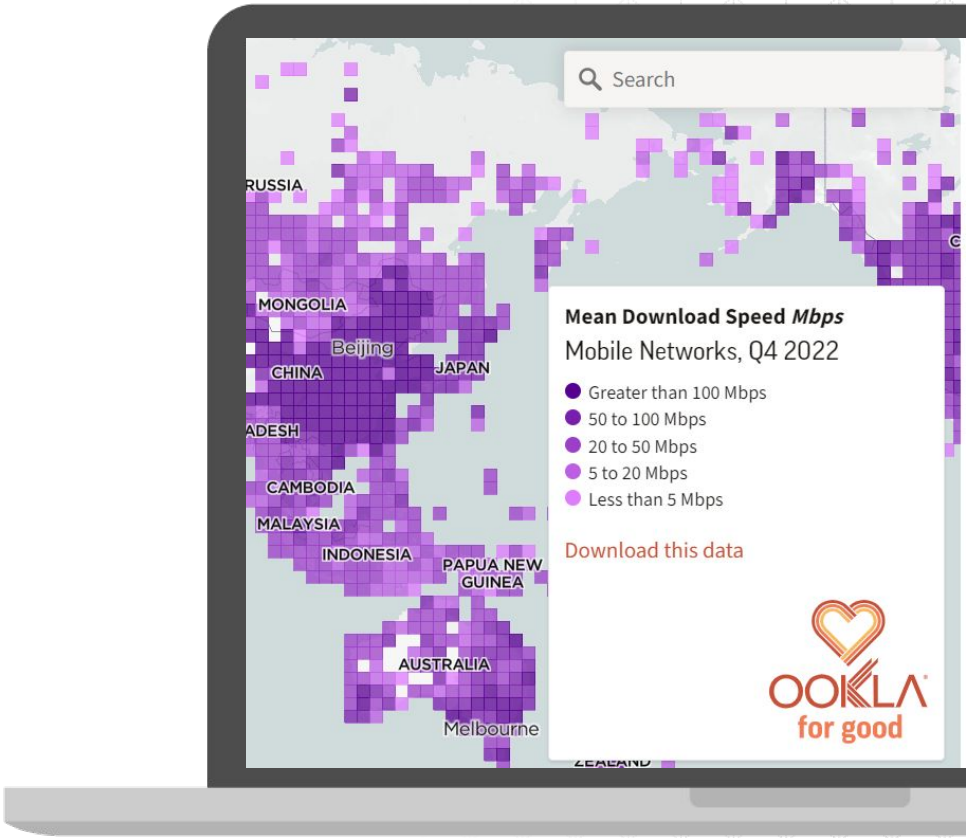
OOKLA

Ookla's Open Data Initiative
(https://www.ookla.com/ookla-for-good/open-data)

Oooka provides open datasets for more informed decisions on internet connectivity, policy, development, etc. It includes data on:

- Global Fixed Broadband and Mobile Network Maps
- Speedtest Global Index
- Ookla 5G Map

Caveats: The data is based on Ookla only, so it doesn't capture speedtests from other services.



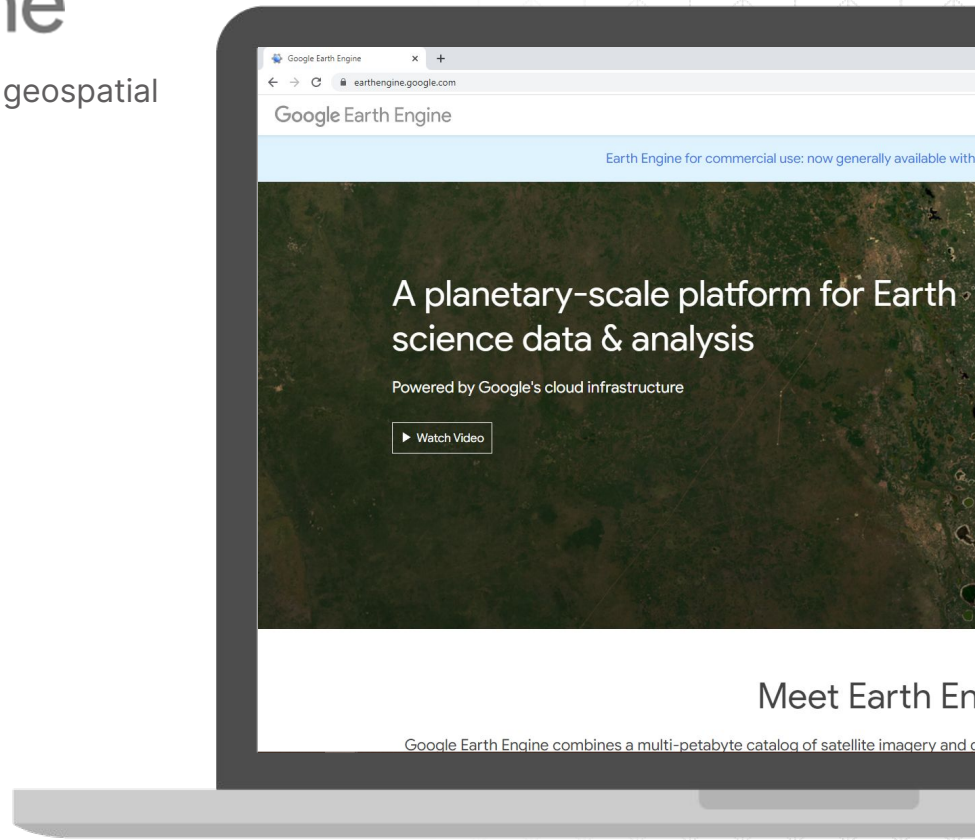
Google Earth Engine

Catalog and processing platform of satellite imagery and geospatial datasets (https://earthengine.google.com/)

Earth Engine is a platform for scientific analysis and visualization of geospatial datasets, that hosts satellite imagery such as:

- Sentinel-2, Landsat, MODIS
- Forest Monitoring datasets
- Weather and climate data

Caveat: It's mainly accessed through code via a web-based code editor that uses javascript or through its python API



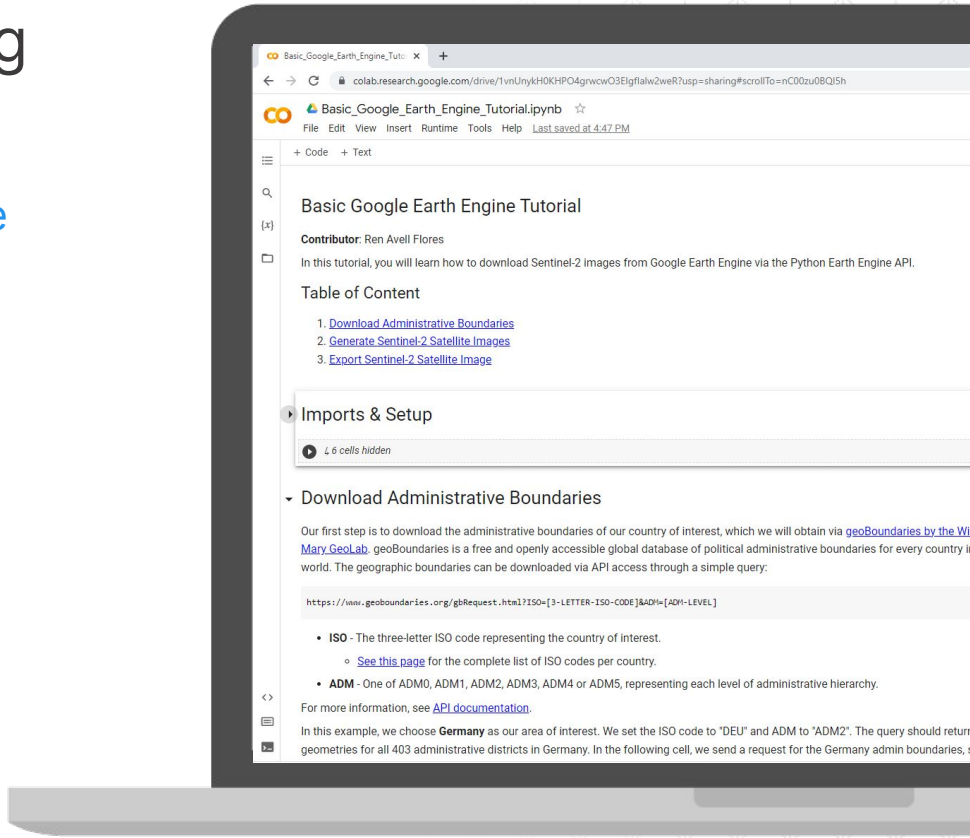
Using Google Earth Engine

Let’s try downloading a satellite image from Earth Engine Using Python and Colaboratory

First, let’s do a quick run-through of Google Colaboratory

It’s a platform that allows the user to write and execute Python in your browser, with:

- Zero configuration required
- Access to GPUs free of charge
- Easy sharing

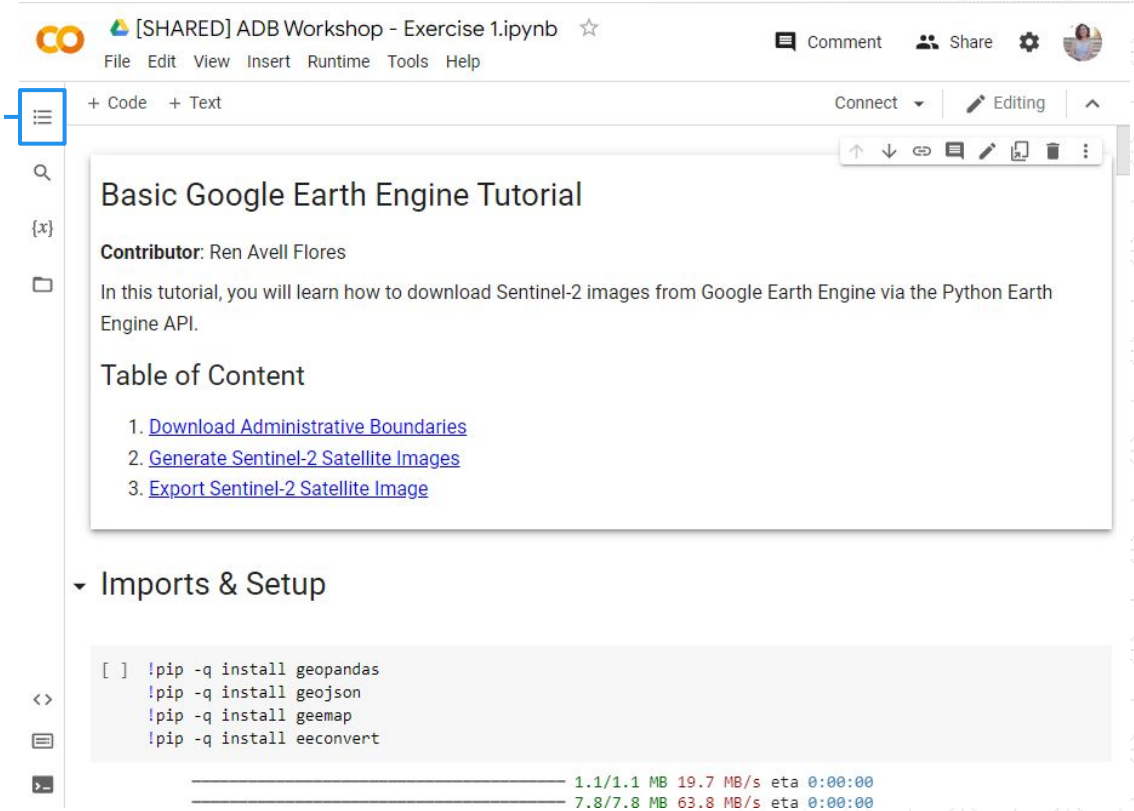


Colaboratory Side Panel

This shows the table of contents of the notebook

Table of Contents

Clicking on this icon will open the side panel displaying the different sections of the notebook

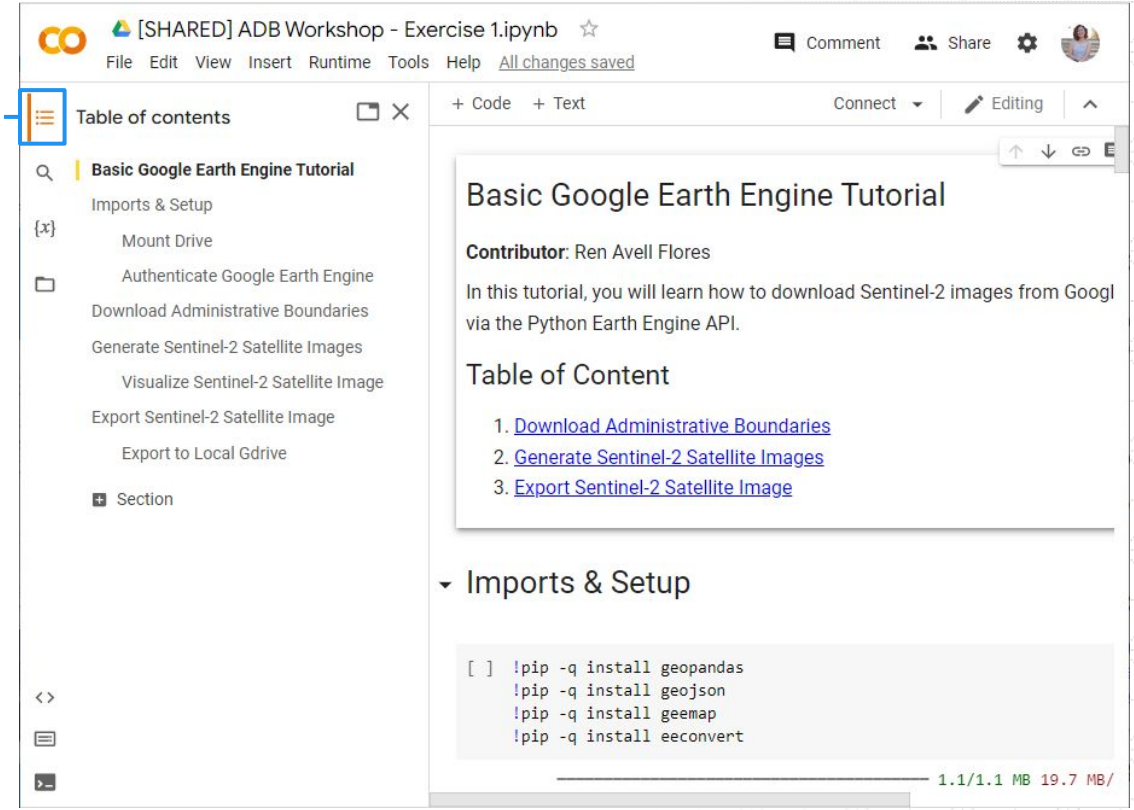


Colaboratory Side Panel

This shows the table of contents of the notebook

Table of Contents

Clicking on this icon will open the side panel displaying the different sections of the notebook

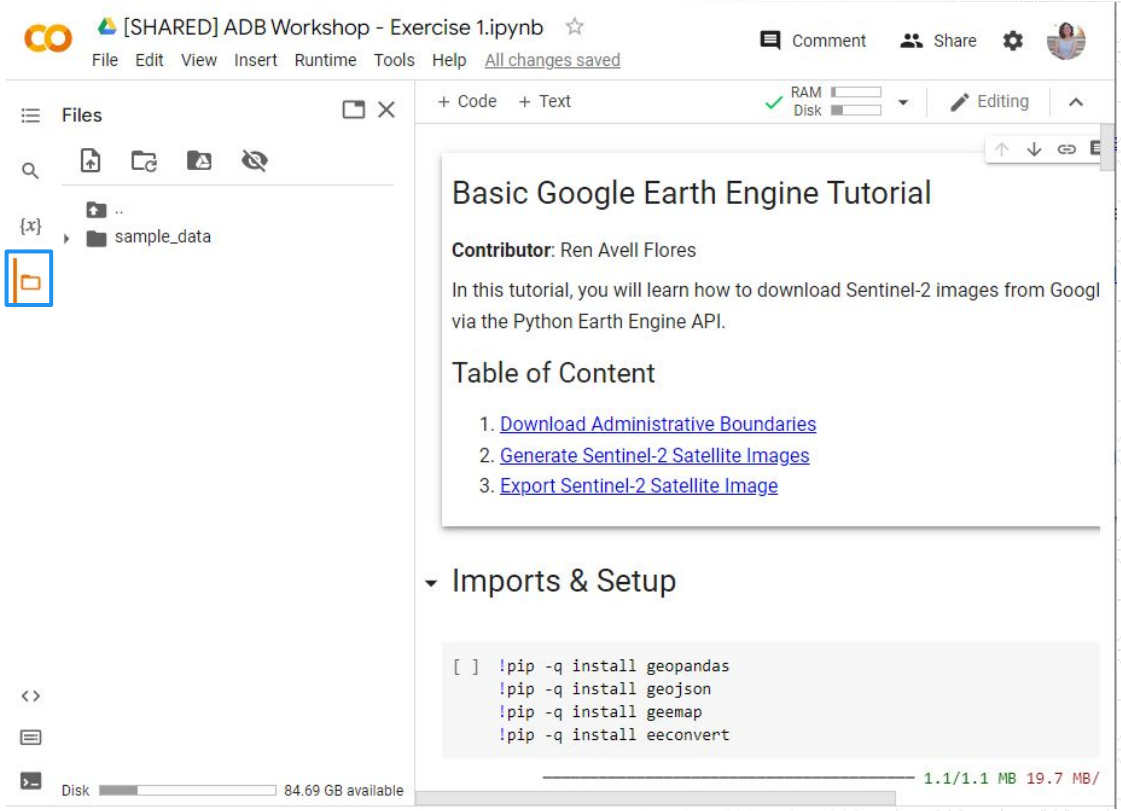


Colaboratory Side Panel

This shows the table of contents of the notebook

Files

Clicking on this icon will show the file directory so you can see what folders and files

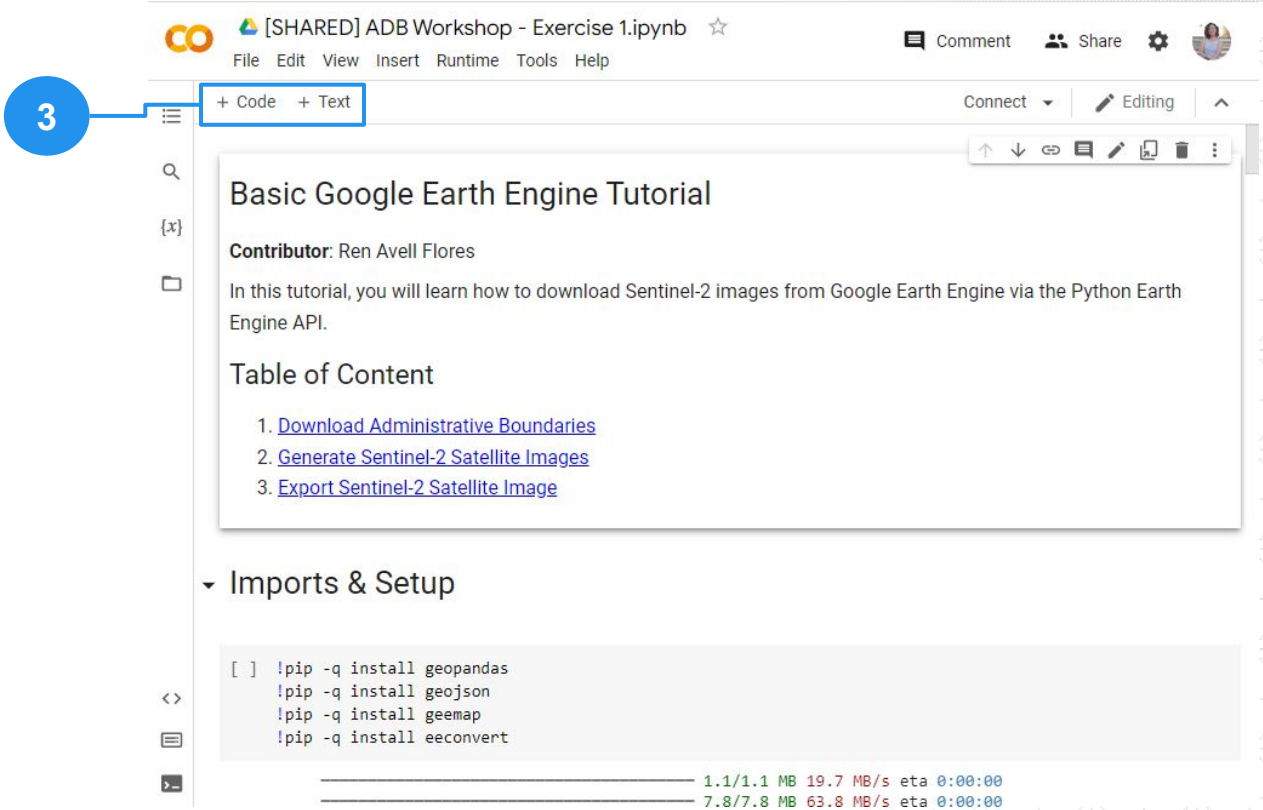


How Code Cells Work

Code cells are where you type the python code

Adding code or text

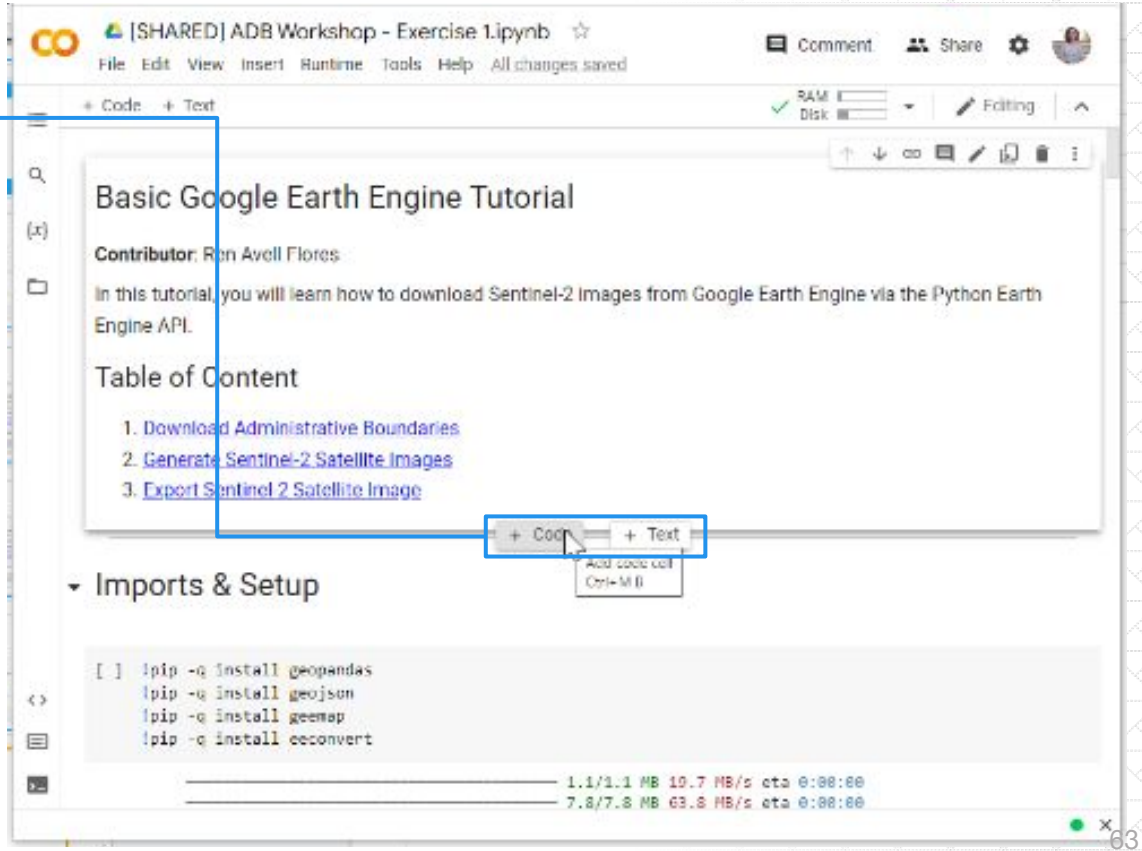
Clicking on this icon will create a new code cell. You can choose between a text cell for descriptions and notes or an executable code block



How Code Cells Work

Code cells are where you type the python code

Adding code or text these icons also appear when you hover at the bottom of each cell



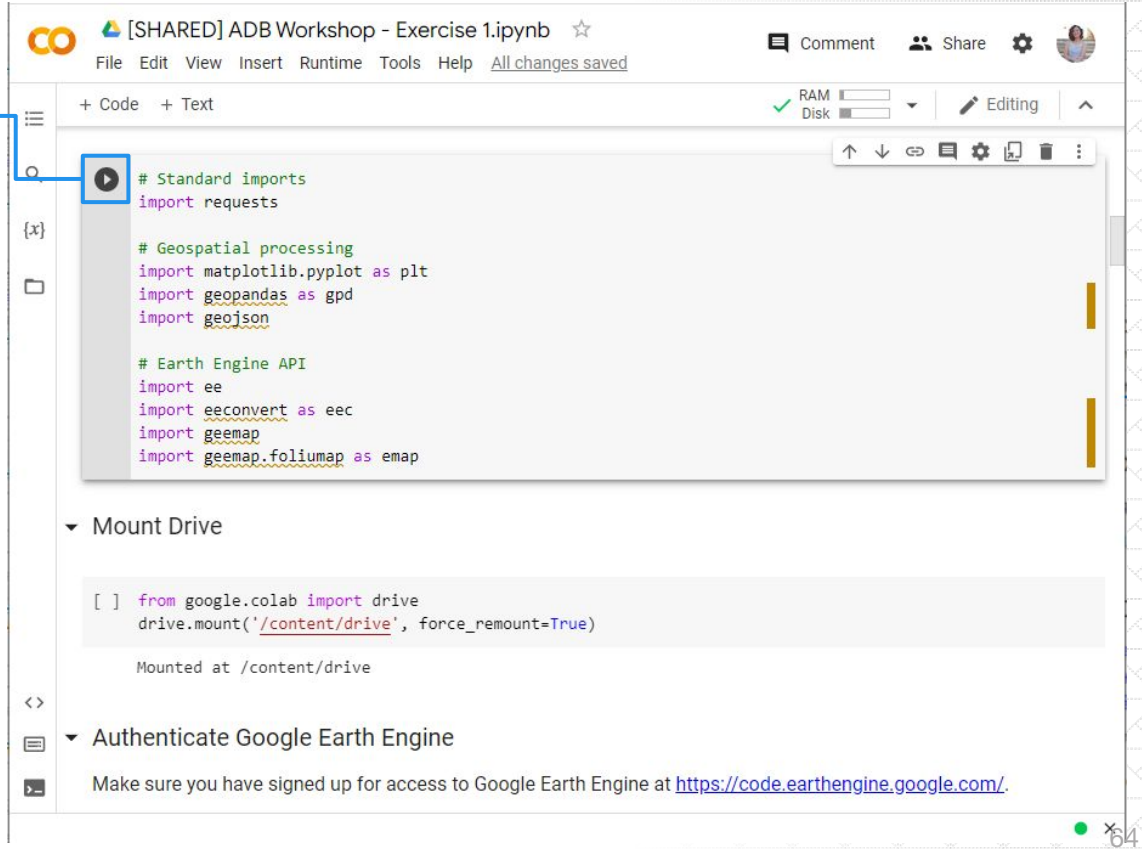
How Code Cells Work

Code cells are where you type the python code

Run Icon

This will execute the code in the cell.

Alternatively you can press CTRL + Enter



Google Colaboratory Walk-through

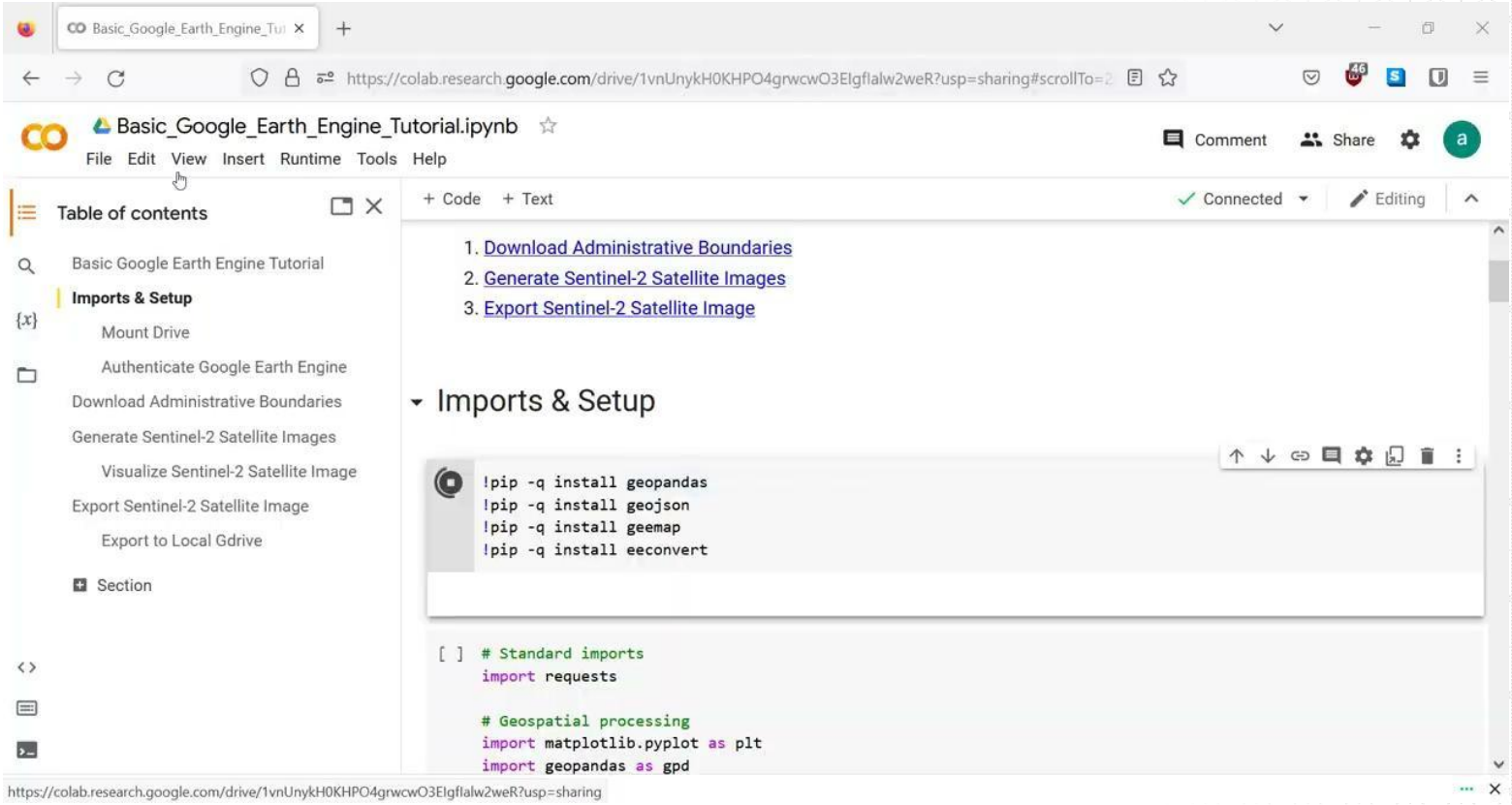
Hands-on Exercise

Exercise 1

Make your own copy by clicking File > Save a Copy in Drive

Enabling GEE on Python

IMPORTANT: Make sure you have a google email registered to GEE, if you have used GEE previously you can create a new account to try out the exercises



03

Introduction to Geospatial Data Analysis

Geospatial Analysis connects data to geography and adds dimension to the analysis

Majority of the datasets we will encounter can have a geospatial aspect to them. GIS can help us:

- Understand patterns, relationships, and geographic context
- Improve communication and data summarization
- Combine multiple datasets for more holistic understanding

How can Geospatial Analysis Help?

Geospatial Analysis and GIS improves data management, analysis and in delivering insights

Data management

Geospatial files are designed to be compact and analysis ready. It can easily fit into commonly used analytics software. It also incentivises data management and promotes good data hygiene

Analysis and Decision Making

Multiple information can be overlayed and analyzed together, giving the user capacity to consider different factors that affect planning and decision-making

Communication

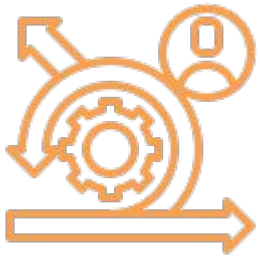
When properly done, maps can easily convey powerful information that is easy to understand even for non-technical people.

Data, Methods and People make up GIS

GIS is more than just software

Data

Data based on geographic location, e.g. poverty estimates, points of interest, population count



Tools and Methods

Data collected by satellites, e.g. building footprints, land cover, nighttime lighting

Experts

Data extracted from connected devices, e.g. signal coverage, internet speed, marketing data

What can we use to do Geospatial Analysis

There's an array of tools from softwares to cloud platforms

GIS Software

ArcGIS, QGIS

Code Libraries

Python and R

Cloud Platforms

Earth Engine, Planetary Computer

Data Types

Vector

Raster

Vector Data

Vector Data represents geographic elements as points, lines and polygons called geometry. Each geometry has attached attributes ie unique identifier, name, classification, etc.

Raster Data

Rasters store data as a matrix of cells or pixels organized into a grid where each cell contains a digital number (DN), that represents information such as temperature or elevation.

The Pros and Cons

Vectors

✔ Graphically more accurate

✔ Works well with discrete values

X Can be processing intensive

X Needs a lot of work and maintenance to ensure that it is accurate and reliable

Rasters

✔ Simple and efficient data model

✔ Works well with continuous values

X Datasets can become very large because they record values for each cell

X Loss in precision

Working with Vectors

Installation and Dependencies

Run the Installation and Dependencies tab on colab

1. Shapely - performs geometric operations
2. Fiona - file access and output
3. Matplotlib - plotting

Vector data comes in may formats

Adding spatial information to data

1. Shapefiles
2. Geopackages
3. GeoJSON
4. CSV
5. Others: Markup Languages, ESRI formats, DLG, etc.

1. GeoPandas extends the datatypes used by pandas to allow spatial operations on geometric types.
2. Combines Pandas with other libraries such as shapely, fiona and matplotlib.

Hands-on Exercise

Exercise 2

Please make your own copy by clicking
File > Save a Copy in Drive

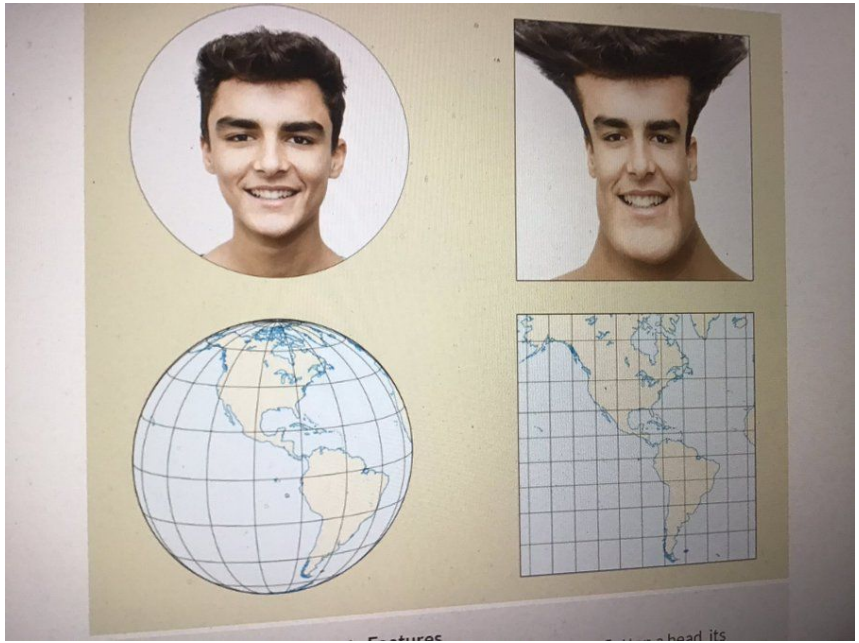
Reading Spatial Data

Loading spatial data into python

1. Read vector file formats
2. Read from CSV

Defining Location - Projection

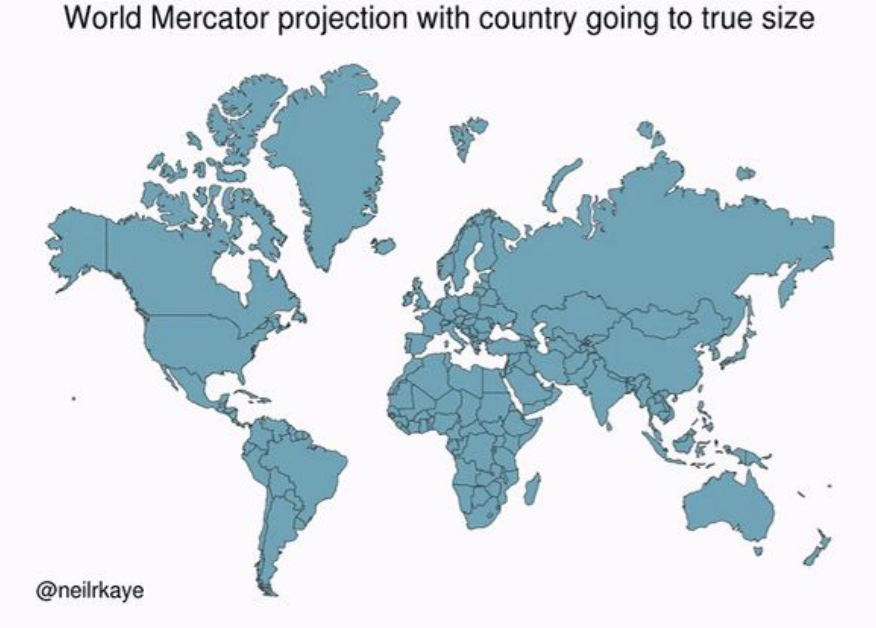
Projections and coordinate reference systems



A projection is the means by which you display a spherical coordinate system on a flat surface and how it relates to real places on the earth.

Defining Location

Distortions



- The default is World Geodetic System 84 (WGS 84)
- Recognized as a global reference system.
- CRS for local areas minimize the distortions

Map animation from @neilrkaye

Spatial Predicates

Spatial Predicates are..

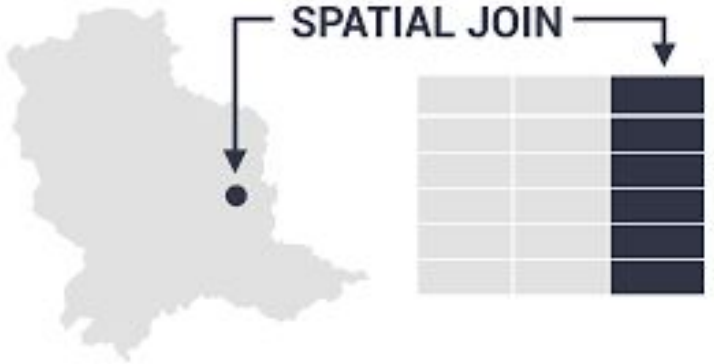
1. Keywords that indicate the type of relationship each piece of geometry has with another.
2. Functions that return TRUE or FALSE for some spatial relationship between two geographies

Examples: Intersects, Disjoint, Equal, etc.

Table Join using Predicates

Just like those table joins but “spatial”

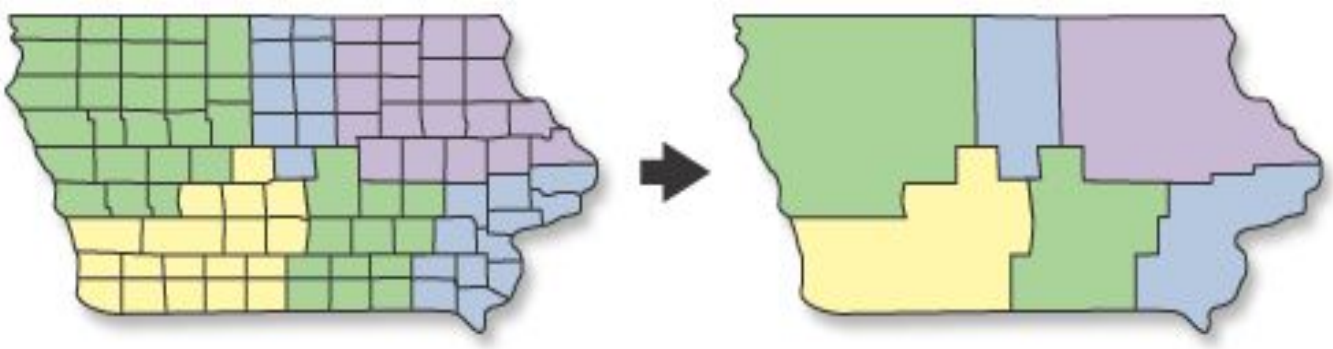
A spatial join matches rows from the Join Features to the Target Features based on their relative spatial locations.
Imagine polygons “inheriting” features based on spatial relation.



Geometric Operations

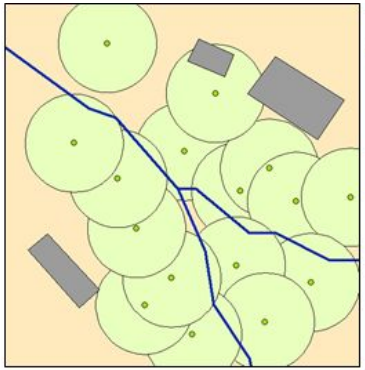
Dissolve

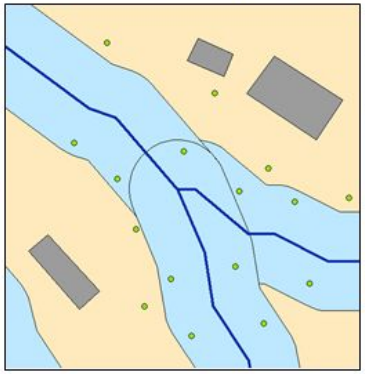
The Dissolve Tool unifies adjacent boundaries based on common attribute values.

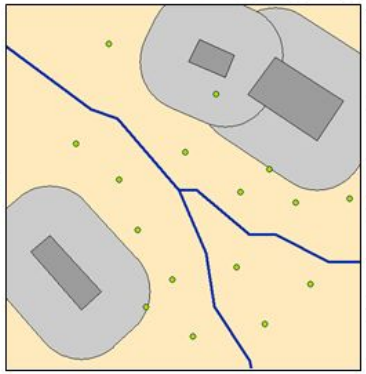


Buffer

Creates a polygon that surrounds any geometry at a given distance.

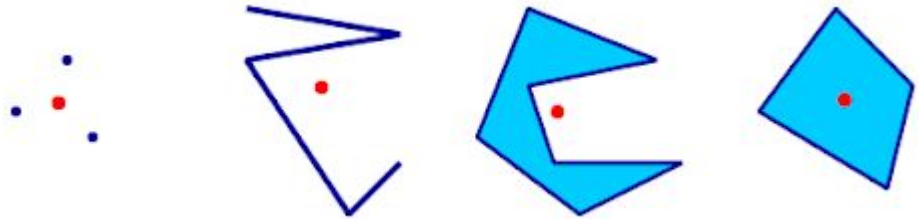






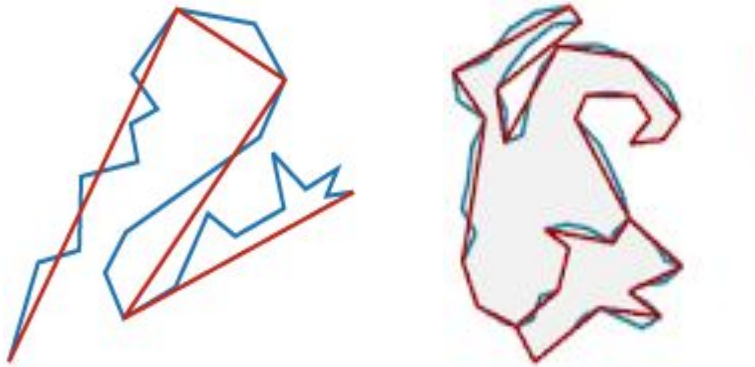
Centroid

Feature analysis tool that finds and generates points from the representative center (centroid) of each input.



Simplify

Cartographic generalization to remove unwanted details.



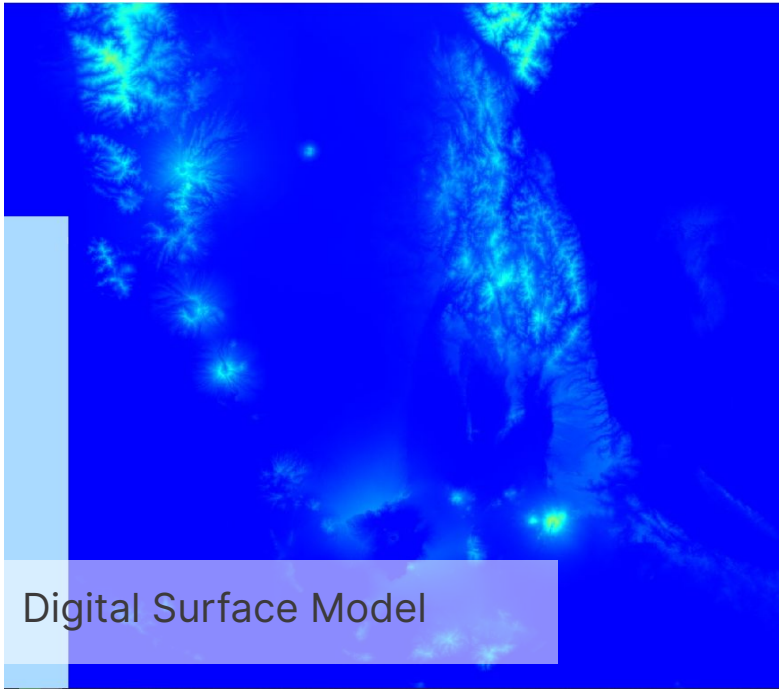
[Extra Materials]
Working with rasters

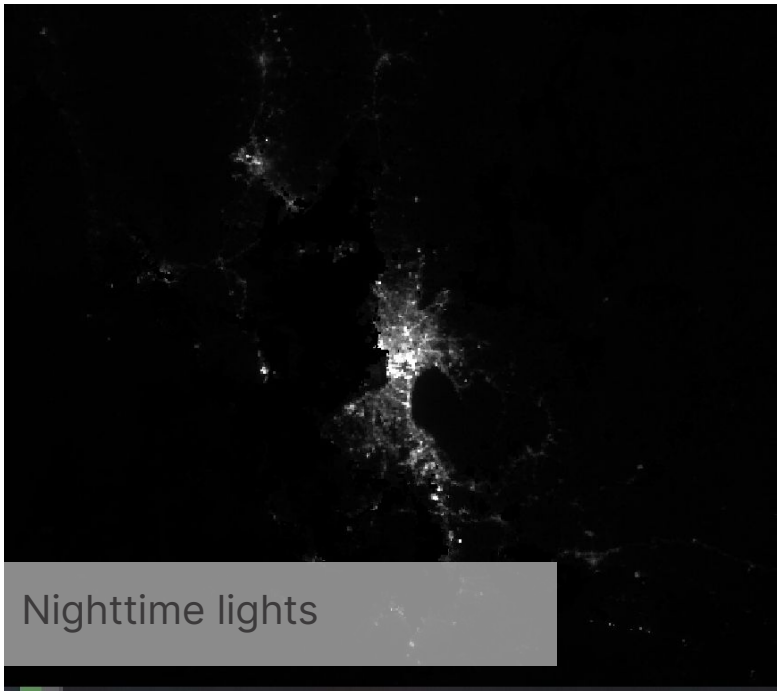
Data Types

Vector

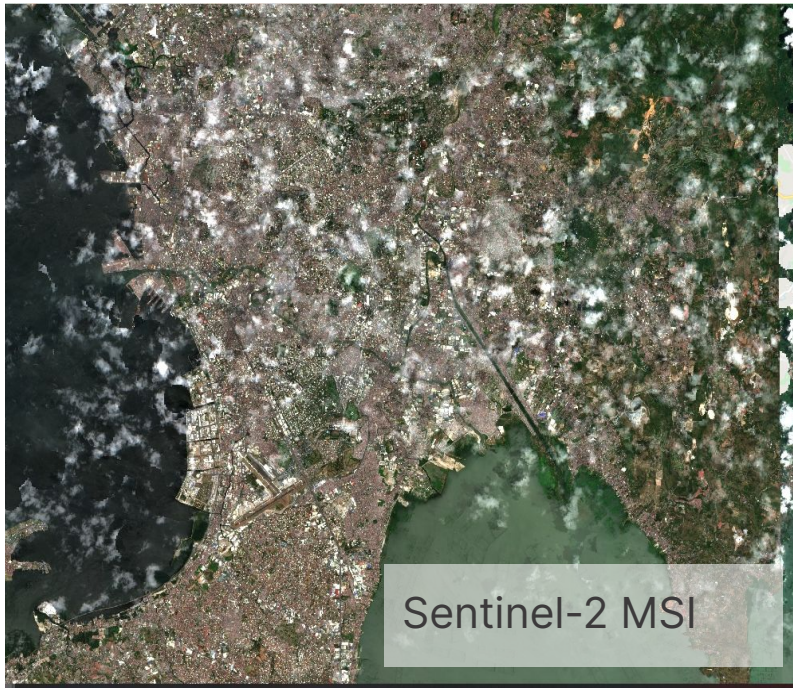
Raster

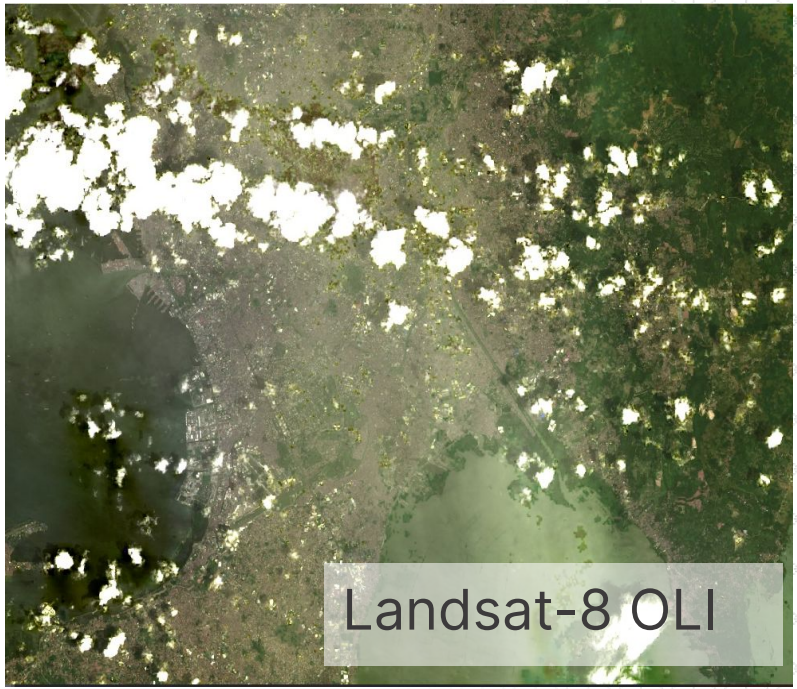
Rasters with Continuous Data





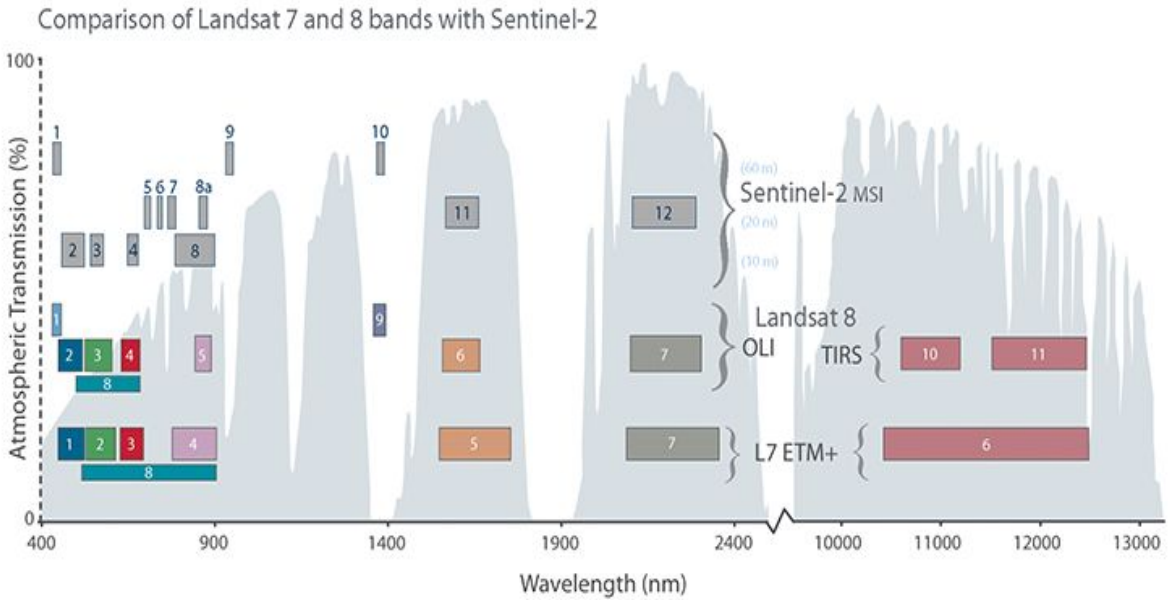
Multi Spectral Images





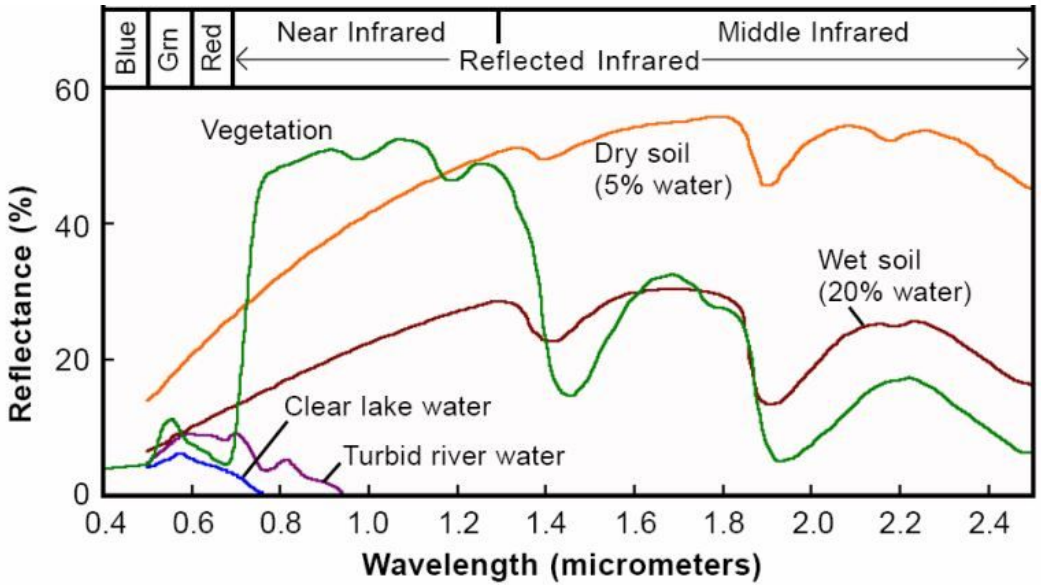
Multi Spectral Images

Multiband or multispectral images are those that capture individual images at specific wave numbers or wavelengths

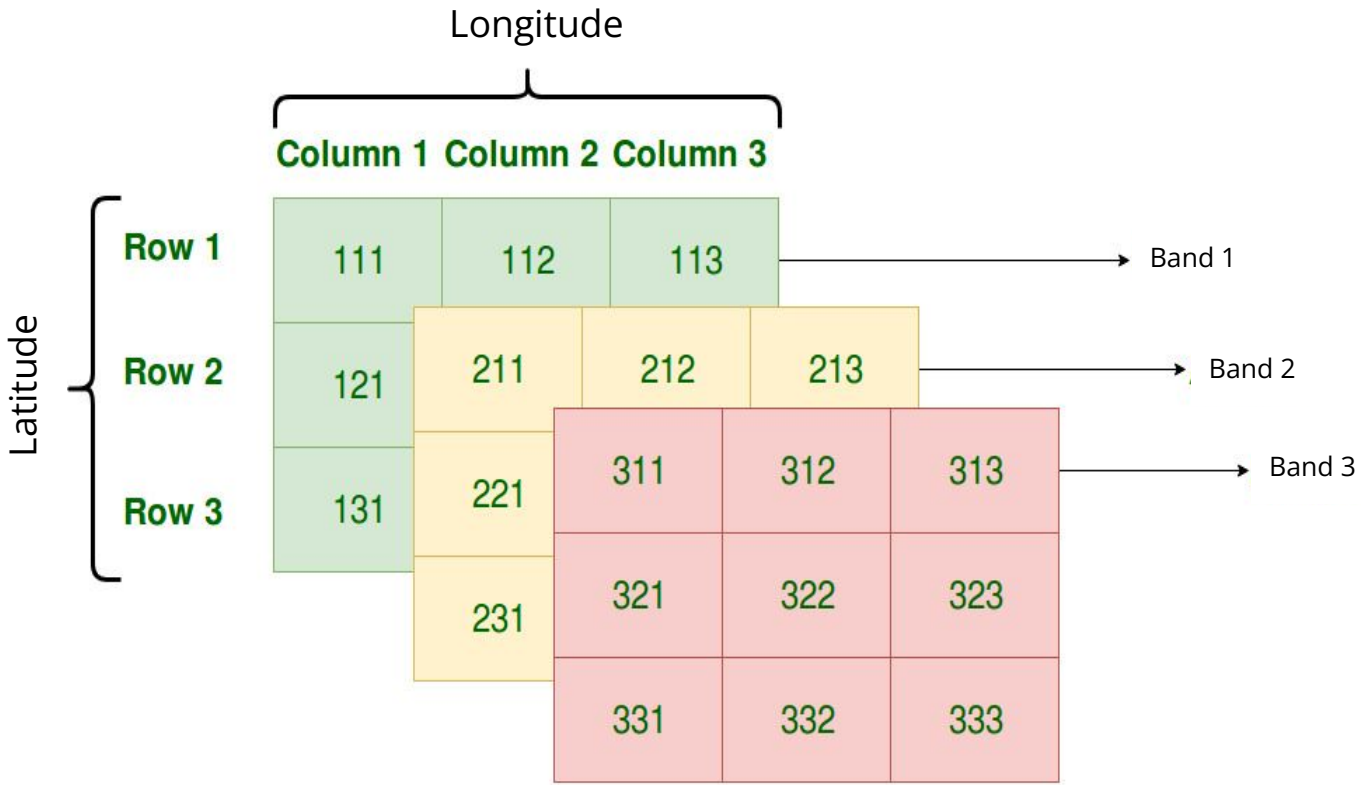


Spectral Signatures

Each object has a unique spectral fingerprint and imaging in different parts of the spectrum helps us in distinguishing one from another



Imagine the dataset as a multidimensional array



RasterIO

1. Rasterio is a highly useful module for raster processing which you can use for reading and writing several different raster formats in Python

2. It provides a Python API based on Numpy N-dimensional arrays and GeoJSON

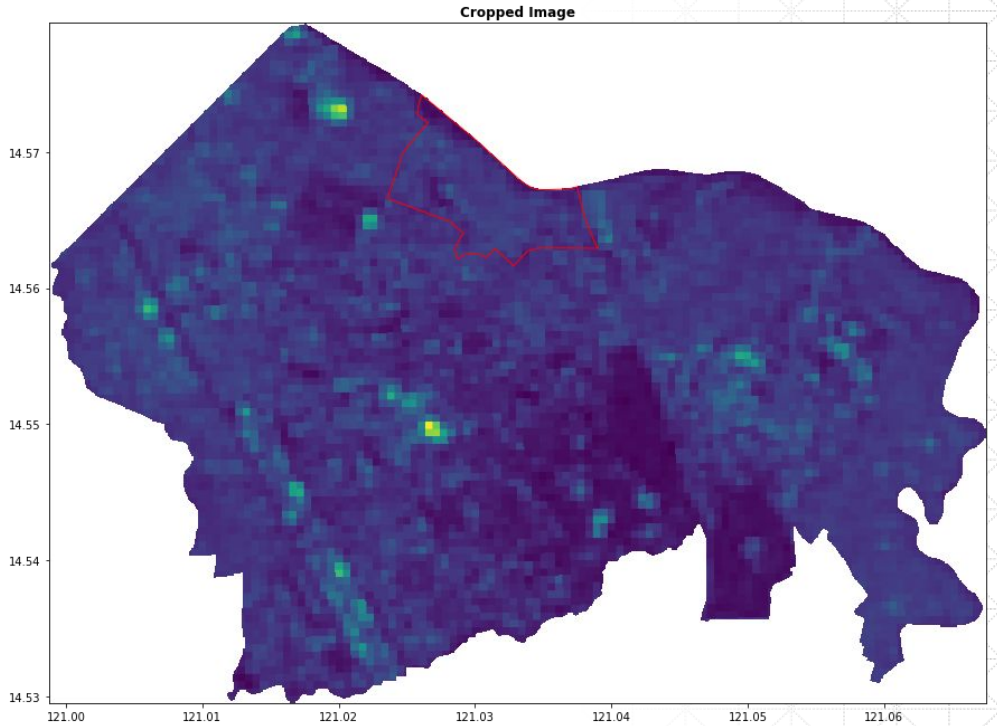
Hands-on Exercise

Exercise 3

Please make your own copy by clicking
File > Save a Copy in Drive

Cropping

Subsetting a smaller part of an image using a vector geometry



Band Math

Band math is using raster band values in algebraic expressions. In python we can do this by converting rasterio objects to numpy arrays and using the array subsets in the expressions.

Installation and Dependencies

Run the Installation and Dependencies tab on colab

1. GDAL - also known as GDAL/OGR, is a library of tools used for manipulating geospatial data. This is also the dependency of almost all geospatial libraries.

Image Geo-transformation

1. The process of transforming the image coordinate space (row, column) to the georeferenced coordinate space (projected or geographic coordinates).

2. Default uses affine transformation; there are other methods but let's skip that for today!

Reprojection... but for images!

Reprojection in GIS consists in changing the coordinates values of a dataset from one coordinate system to another coordinate system.

Satellite Indices using Band Math

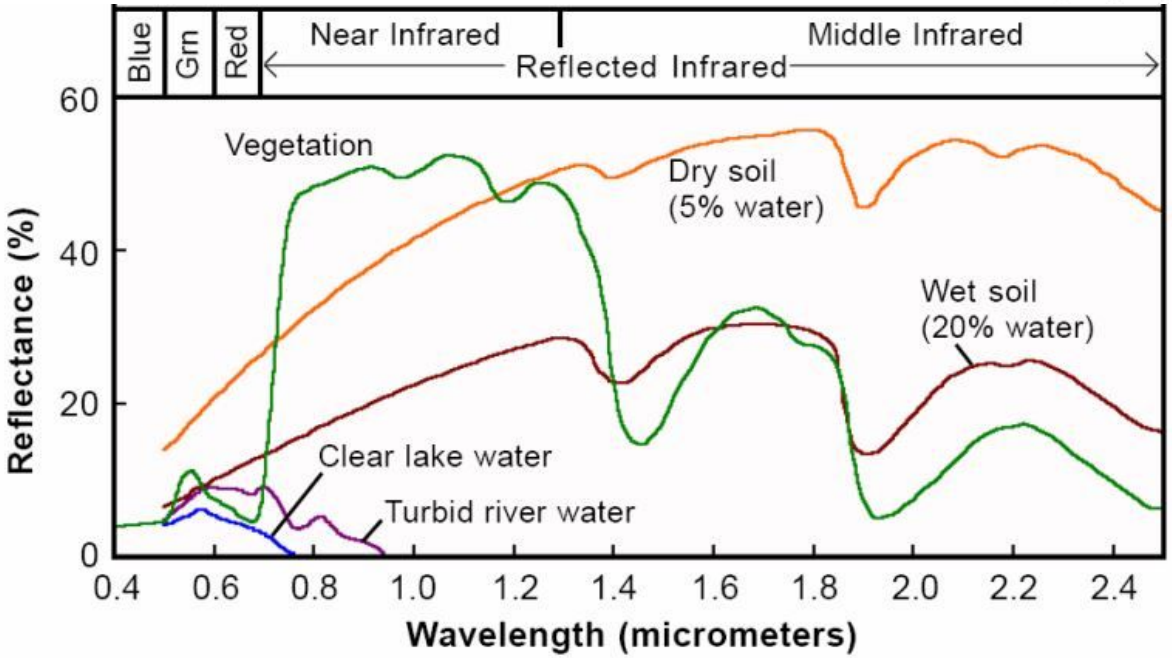
What are satellite indices?

Satellite Indices

An index is basically a ratio of values in different satellite bands to measure how high or low a reflectance of a particular feature is. The most common example is Vegetation Index

Vegetation Index

Ratio between red and near-infrared



Other Operations

1. Vectorization
    a. Reclassification is the process of reassigning a value, a range of values, or a list of values in a raster to new output values.
2. Zonal Statistics
    a. Zonal statistics returns summary statistics of raster values within the specified polygon or polygons.
3. Point Sampling
    a. Point sampling extracts the raster value that intersects or is closest to a coordinate point.

Introduction to GeoML

Recap

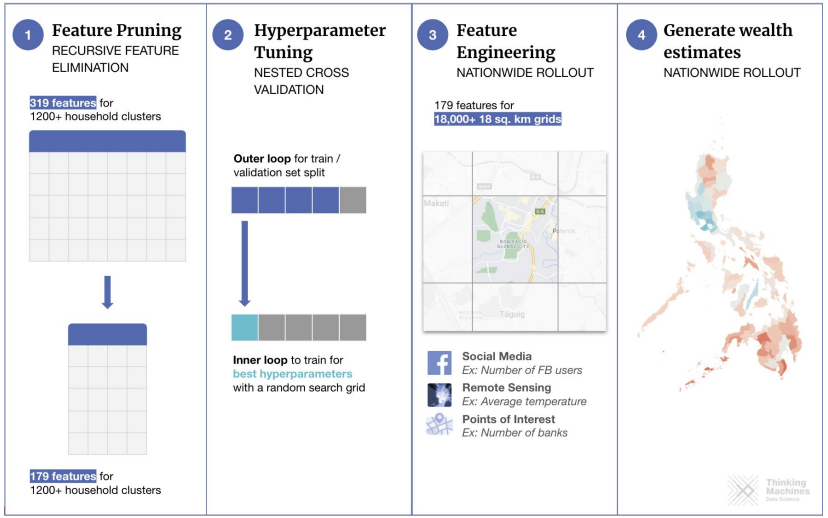
Instructor: Ren Flores

1. Project Intro: Rice Yield Estimation with Satellite Imagery and Machine Learning
    a. Background
    b. Methods
    c. Results and Learnings
2. Open Data for Social Impact
    a. What is Open Data
    b. Access and Sources
    c. Google Earth Engine
3. Geospatial Analysis in Python
    a. GIS in Python
    b. Vectors
    c. Rasters

Introduction to GeoML

Up Next for Day 2: Machine Learning Using Python

1. Classifical Machine Learning
2. Computer Vision



We'd love to hear your feedback!

Feedback Link:
https://forms.gle/iAkb1Lixww3FesGv8

Thinking Machines
Data Science

Data Stories
stories.thinkingmachin.es

Press
thinkingmachin.es/press-room

Follow us
/thinkdatasci
@thinkdatasci

Bangkok | Manila | Singapore