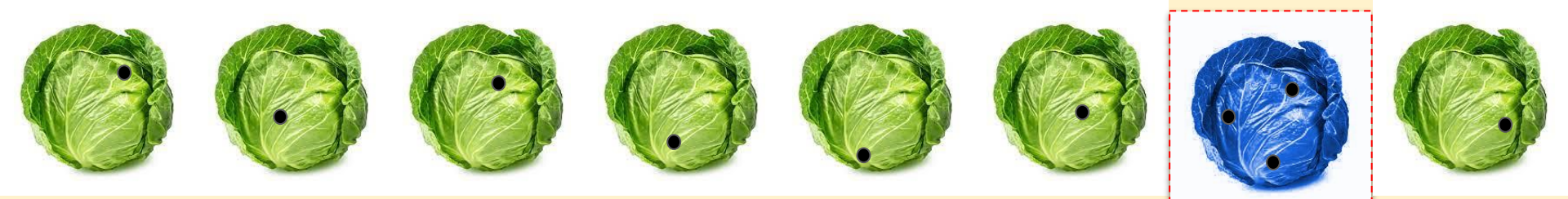


DataBuilder

Dataset builder for training geospatial ML models

Working concept prototype

Brought to you by Cabbage Collectivist Society

Andy Lee, Kenneth Ruan, Kyssen Yu, Dason Wang, Emily Wang

DataBuilder – A basic introduction

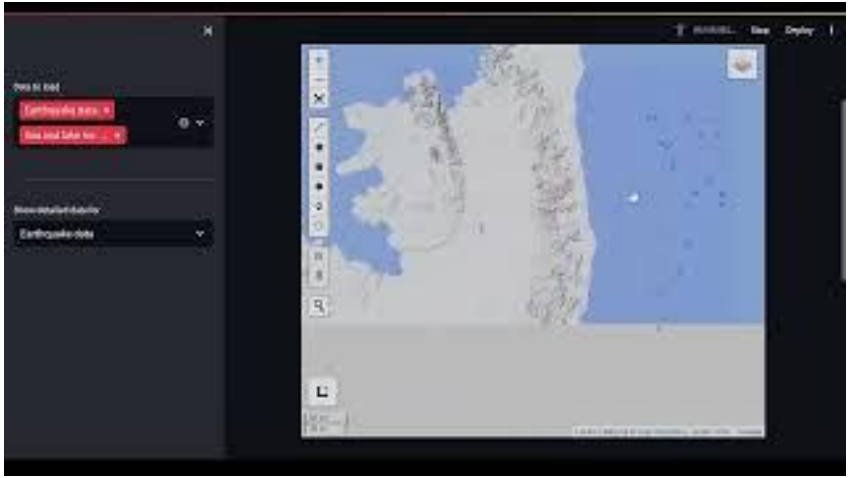
Challenge:

Mapping Data for Societal Benefit

We built a powerful data acquisition platform specifically designed to generate **AI training datasets from** meshes of **publicly available data**, making it the ideal aide for AI researchers and analysts.

We strove to make the design **intuitive for just about anyone to use**, improving the accessibility of the data for the general population while still being capable of **powering modern-day research.**

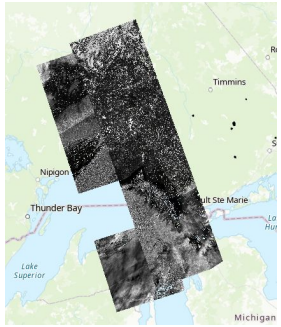
An ambitious future for both **civilian science** as well as **academic research** lies ahead for DataBuilder and the AI community.



The Problem: The struggle of dataset creation

Satellite imagery plotted on a map in ArcGIS ->

Needs to be tilted, adjusted for resolution, rasterized, exported to GeoTIFF format with a vector layer overimposed over the features wherein wildfire burn scars are visible

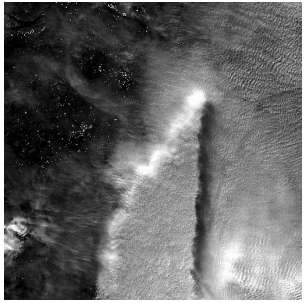


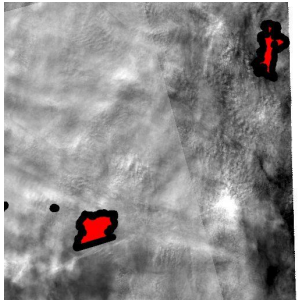
Models train by processing **massive amounts of labelled data sets**; in particular, to develop a geospatial AI model, researchers need to source significant amounts of satellite data.

While geospatial data is readily available for public access, it is often locked behind confusing repositories or databases which may be difficult for the novice researcher to navigate. Furthermore, we discovered a **lack of available tools to streamline the process of compiling datasets together.**

This is especially true for any task which necessitates **compositing multiple sets of data:** for example, in wildfire detection, a map of historical fires overlaid atop changes in infrared heat signs can be used to train accurate AI classification/segmentation models to aid disaster relief.

The difficulty of accessing public data also deters casual learners from interacting with it, further exacerbating the issue.

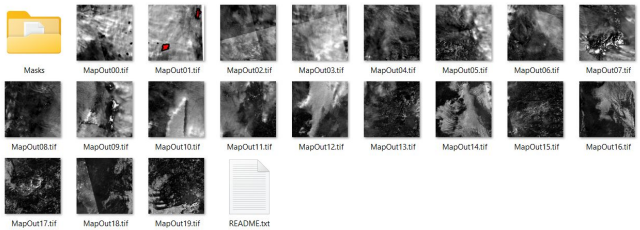




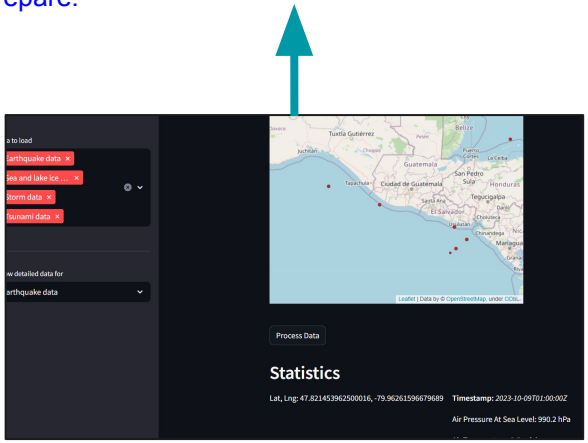
Infrared intensity composited against burn scars
These finished labeled TIFF files (converted to jpg to put here) took 10 hours of research and work to compile!

Left: has no burn scars
Right: has burn scars (highlighted in red for clarity) layered on top as a removable mask. The mask and the underlying image serve as datapoint and label for the AI to learn from

Why is this important? The implications



This dataset is exceedingly small, and yet would take a human a great deal of time to find, fetch, and prepare.



Our solution, DataBuilder, allows you to acquire a set just like above without the time sink.

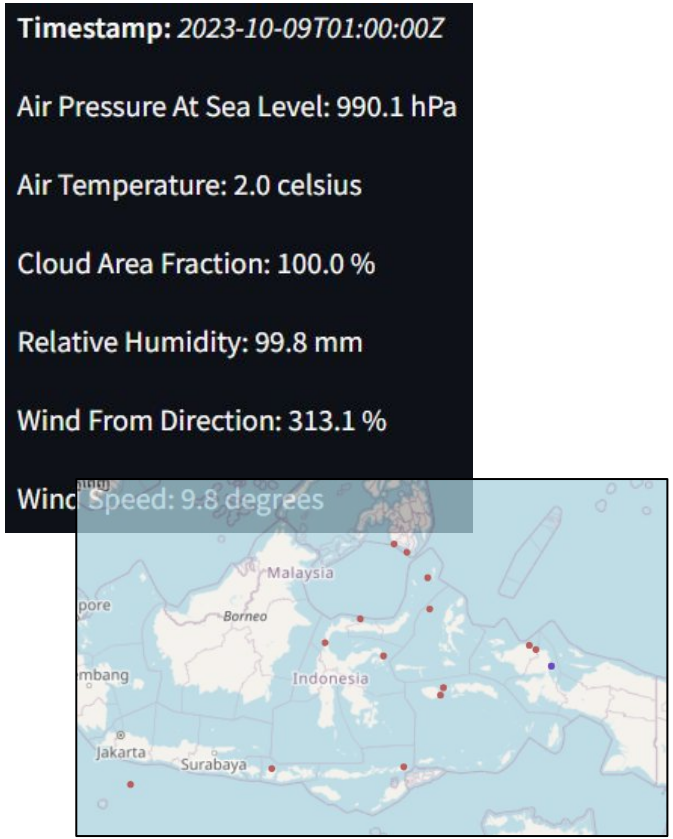
Many hidden correlations can be uncovered by broadly analyzing multiple sources of data. Likewise, compiling these sources together into one product allows greater accessibility for general audiences. Not only does this allow the training of countless ML models, but it also allows for intuitive centralized data visualization: both of which are currently sorely lacking.

Imagine a tool containing multiple sources of geospatial data, constantly being **updated in real time** and **equipped with the capacity to download the data in bulk** to build unique custom datasets with composite layers.

The result? Researchers will be better able to create data-centric tools for the benefit of society. The datasets created with the help of our platform could be used to build models for forest fire prediction, climate change tracking, or any other model used in land-related applications.

Furthermore, general audiences will be better able to access and learn about these models and the data used to create them, fostering public interest and knowledge in geospatial AI.

Our Solution: **DataBuilder** in depth

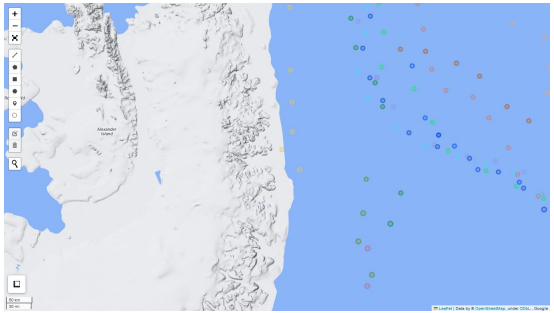


To fill the gap in data compositing avenues, we created a **centralized dynamic** Web App updated in real time via API calls. It reveals **spatial correlations** between events, addresses the issue of manually layering data from multiple sources geospatially, and allows for **easy access to machine learning training data.** It **displays data intuitively** and features a simple user interface, making it suitable for use by anyone.

Databuilder **represents and provides access to data** gathered from NASA and various government institutions on a dynamic map and provides a **platform to access this data**. The **geospatial relationships** between datasets are rendered apparent and can be exported into a format suitable for training machine learning models.

Currently DataBuilder offers **5** different datasets and **2** different export modes, with plans to add more of both. We foresee that increasing the **customizability** of our datasets and export options will lead to a wider variety of **opportunities** for data classification, experimentation, and prediction in Earth Science.

Methodology: Process, resources, and attributions



Visual plot of Iceberg and earthquake data returned by api calls for a certain time

Databuilder's beginnings at this Hackathon encompass five reputable APIs.

From Norway, the **MET Weather API** offers a variety of live climate information. It intuitively operates alongside NASA's very own **EONET**, a natural event tracker reporting on various storms and other phenomena.

Similarly, the three NOAA APIs contribute data on tsunamis, earthquakes, and volcanic activity. The weather-related nature of these APIs encourage data and annotation sets woven between the resources, with even the seismic data correlating with tsunamis and storms.

Met Office

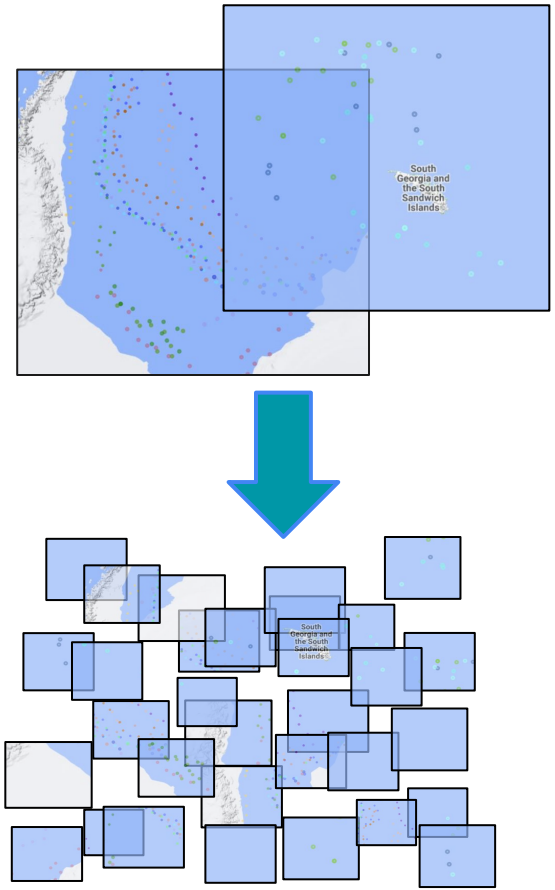
NASA

NOAA

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

U.S. DEPARTMENT OF COMMERCE

Conclusion – The journey and plans for the future



Our motivation to create DataBuilder was **in response to the roadblocks we faced** when trying to acquire appropriate machine learning model training data. We believe that **we've created a solution** - an easily accessible interface to streamline the dataset creation process for ML model training, with applications ranging from the professional sector to personal endeavours.

However, our journey will hopefully not end here. We have big plans to take DataBuilder to the next level, adding features including a **data upload function** to facilitate community data sourcing, as well as more polished bulk data download functionality, with support for a **broader and more versatile range** of data types and formats.

Ultimately, our goal is to create a platform that spatially colocates and displays **any geospatial data**, making it easily accessible and usable for machine learning, to facilitate the creation of better models across a variety of applications.