Using Satellite Imagery to Identify Harmful Algal Blooms and Protect Public Health

Emily Dorne
Lead Data Scientist

DRIVENDATA

Data Science + Social Impact

Machine Learning Competitions • Direct Consulting • Open Source Projects

drivendata.org

drivendata.co

github/drivendataorg

Outline

- What is cyanobacteria and why it matters
- CyFi: an open-source package that uses ML to estimate cyanobacteria levels
- Three takeaways for those developing machine learning models with satellite imagery

What is cyanobacteria

Cyanobacteria is a type of microscopic algae, also known as blue-green algae. It is the most common source of harmful algal blooms (HABs) in freshwater environments.

HABs occur when excessive algae growth produces toxins that are harmful to human health, dangerous for other mammals like pets, and damage aquatic ecosystems.

HABs pose a significant risk to inland water bodies – lakes, rivers, and reservoirs – which are commonly used for recreational activities or drinking water.

Current approach

Monitoring for HABs is typically done via manual water sampling, where samples are collected and then sent off to a lab for toxin analysis.

Water quality managers make decisions around issuing public health warnings and/or implementing closures when blooms are detected.

Manual water sampling is accurate, but is too time and resource intensive to perform continuously at scale.

The potential of ML

Machine learning is particularly well-suited to this task because indicators of cyanobacteria are visible in free, routinely collected satellite imagery.

Machine learning models can generate estimates in seconds, making it feasible to provide daily estimates of cyanobacteria for water bodies all across the United States.

Introducing CyFi: Cyanobacteria Finder

CyFi is a python package that uses satellite imagery and machine learning to estimate cyanobacteria levels in small, inland water bodies.

The goal of CyFi is to help water quality managers better allocate resources for in situ sampling, and make more informed decisions around public health warnings for critical water bodies like lakes and reservoirs.

CyFi uses high-resolution Sentinel-2 satellite imagery (10-30m) to focus on smaller water bodies with rapidly changing blooms.

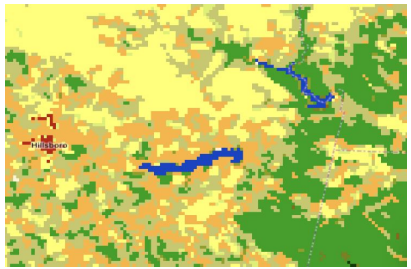
This is a key difference from **existing tools which use Sentinel-3 imagery.** Its resolution of 300-500m is sufficient for oceans and coastal areas, but is often too coarse for small, inland water bodies.

Input | Feature data sources | Decision-tree model | Output

date	latitude	longitude
2015-06-29	41.424144	-73.206937
2013-07-25	36.045000	-79.091942
2017-08-21	35.884524	-78.953997
2019-08-28	41.392490	-75.360700

Sentinel-2 imagery

Land cover classification



LightGBM

density_cells_per_ml	severity
57,433	moderate
83,609	moderate
5,733	low
3,684,003	high

Input

Feature data sources

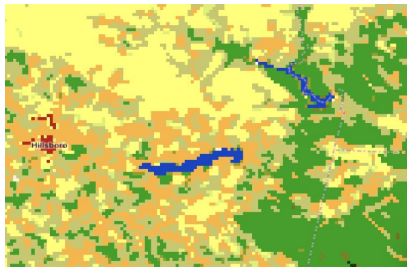
Decision-tree model

Output

date	latitude	longitude
2015-06-29	41.424144	-73.206937
2013-07-25	36.045000	-79.091942
2017-08-21	35.884524	-78.953997
2019-08-28	41.392490	-75.360700

Sentinel-2 imagery

Land cover classification



LightGBM

density_cells_per_ml	severity
57,433	moderate
83,609	moderate
5,733	low
3,684,003	high

Feature generation from satellite imagery

Each observation is a combination of date + lat/lon

Sample points

uid	date	latitude	longitude
bmdk	2015-06-29	41.424144	-73.206937
obdp	2013-07-25	36.045000	-79.091942
fmjb	2017-08-21	35.884524	-78.953997
xyht	2019-08-28	41.392490	-75.360700
gstw	2013-07-11	38.305600	-122.026000

Feature generation from satellite imagery

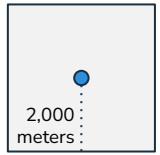
Each observation is a combination of date + lat/lon

- Specify **bounding box** around point
  - 2,000 m

Sample points

uid	date	latitude	longitude
bmdk	2015-06-29	41.424144	-73.206937
obdp	2013-07-25	36.045000	-79.091942
fmjb	2017-08-21	35.884524	-78.953997
xyht	2019-08-28	41.392490	-75.360700
gstw	2013-07-11	38.305600	-122.026000

Satellite imagery



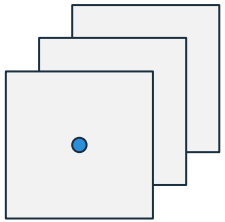
Feature generation from satellite imagery

Each observation is a combination of date + lat/lon

- Specify bounding box around point
  - 2,000 m
- Specify **time window** of imagery prior to sample date
  - 30 day window range

Sample points

uid	date	latitude	longitude
bmdk	2015-06-29	41.424144	-73.206937
obdp	2013-07-25	36.045000	-79.091942
fmjb	2017-08-21	35.884524	-78.953997
xyht	2019-08-28	41.392490	-75.360700
gstw	2013-07-11	38.305600	-122.026000



Satellite imagery

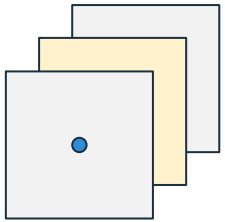
Feature generation from satellite imagery

Each observation is a combination of date + lat/lon

- Specify bounding box around point
  - 2,000 m
- Specify time window of imagery prior to sample date
  - 30 day window range
- Select **most recent, least cloudy** image
  - Calculate % of pixels that are clouds in bbox
  - Use most recent, least cloudy image
  - If all images have more than 5% of clouds, no prediction will be made

Sample points

uid	date	latitude	longitude
bmdk	2015-06-29	41.424144	-73.206937
obdp	2013-07-25	36.045000	-79.091942
fmjb	2017-08-21	35.884524	-78.953997
xyht	2019-08-28	41.392490	-75.360700
gstw	2013-07-11	38.305600	-122.026000



Satellite imagery

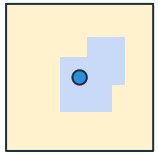
Feature generation from satellite imagery

Each observation is a combination of date + lat/lon

- Specify bounding box around point
  - 2,000 m
- Specify time window of imagery prior to sample date
  - 30 day window range
- Select most recent, least cloudy image
  - Calculate % of pixels that are clouds in bbox
  - Use most recent, least cloudy image
  - If all images have more than 5% of clouds, no prediction will be made
- Filter to **water area**
  - Using scene classification band

Sample points

uid	date	latitude	longitude
bmdk	2015-06-29	41.424144	-73.206937
obdp	2013-07-25	36.045000	-79.091942
fmjb	2017-08-21	35.884524	-78.953997
xyht	2019-08-28	41.392490	-75.360700
gstw	2013-07-11	38.305600	-122.026000



Satellite imagery

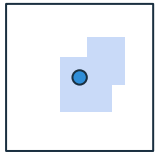
Feature generation from satellite imagery

Each observation is a combination of date + lat/lon

- Specify bounding box around point
  - 2,000 m
- Specify time window of imagery prior to sample date
  - 30 day window range
- Select most recent, least cloudy image
  - Calculate % of pixels that are clouds in bbox
  - Use most recent, least cloudy image
  - If all images have more than 5% of clouds, no prediction will be made
- Filter to water area
  - Using scene classification band
- **Calculate features** from imagery bands in water area
  - Summary stats (mean, max, min)
  - Ratios (NDVI, etc.)

Sample points

uid	date	latitude	longitude
bmdk	2015-06-29	41.424144	-73.206937
obdp	2013-07-25	36.045000	-79.091942
fmjb	2017-08-21	35.884524	-78.953997
xyht	2019-08-28	41.392490	-75.360700
gstw	2013-07-11	38.305600	-122.026000



Satellite imagery

uid	B02_mean	B02_min	B02_max	B03_mean	B03_min	B03_max	B04_mean
bmdk	290.260418	112.5	1934.0	458.530614	168.0	2158.0	347.846475
obdp	290.260418	112.5	1934.0	458.530614	168.0	2158.0	347.846475
fmjb	418.988123	175.0	2686.0	604.710812	267.0	2934.0	509.557734
xyht	161.532712	50.0	1182.0	312.350417	69.0	1382.0	186.135216
gstw	290.260418	112.5	1934.0	458.530614	168.0	2158.0	347.846475

Sample features

Simply run one line of code to generate predictions

```
$ cyfi predict list_of_points.csv
```

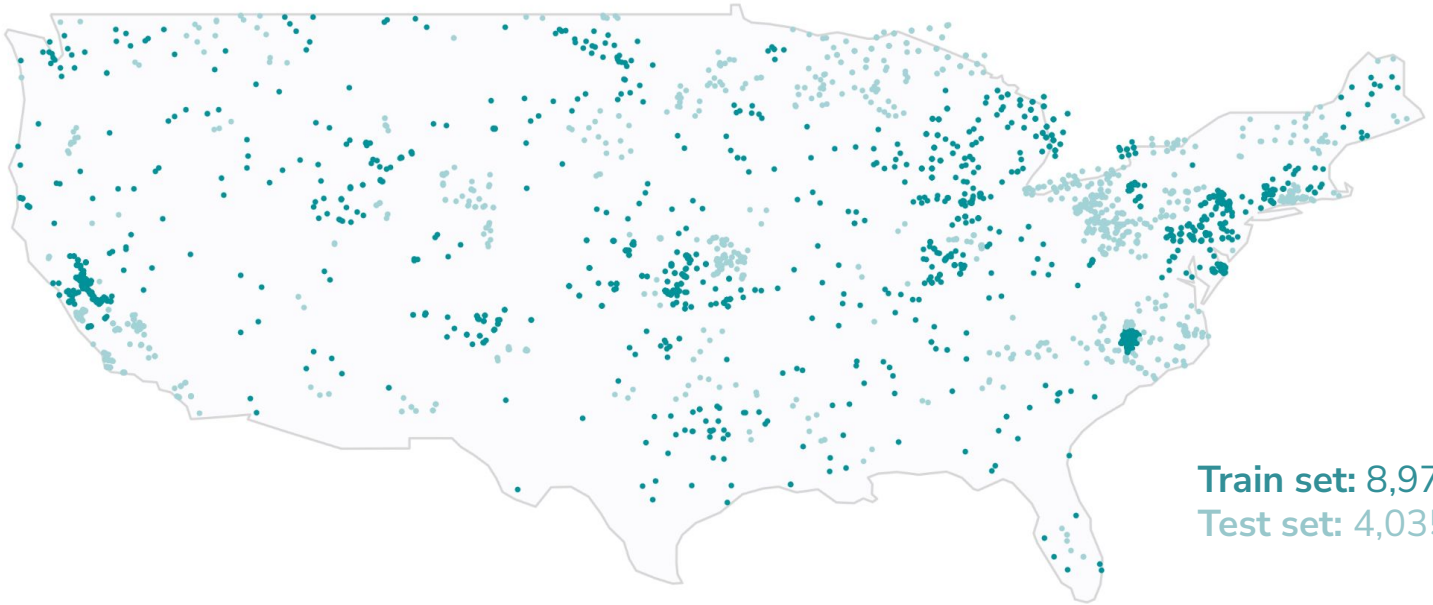
```
SUCCESS  | Loaded 5 sample points (unique combinations of date, latitude, and longitude) for prediction
SUCCESS  | Downloaded satellite imagery
SUCCESS  | Cyanobacteria estimates for 4 sample points saved to preds.csv
```

Or estimate cyanobacteria for a single point rather than providing a file

```
$ cyfi predict-point --lat 35.6 --lon -78.7 --date 2023-09-25
```

```
SUCCESS  | Estimate generated:
date                    2023-09-25
latitude                      35.6
longitude                    -78.7
density_cells_per_ml        22,836
severity                  moderate
```

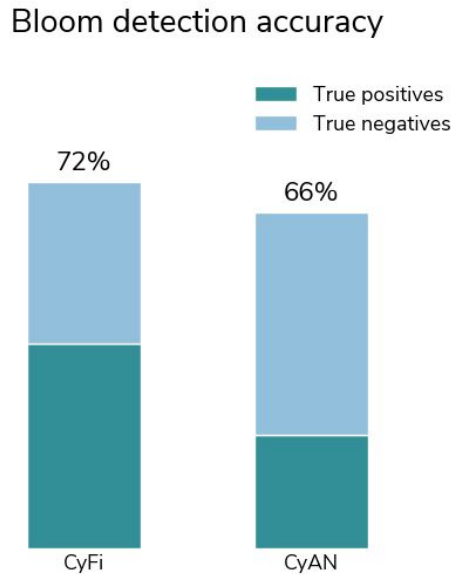
CyFi was trained and evaluated using in-situ measurements of cyanobacteria density from across the U.S.

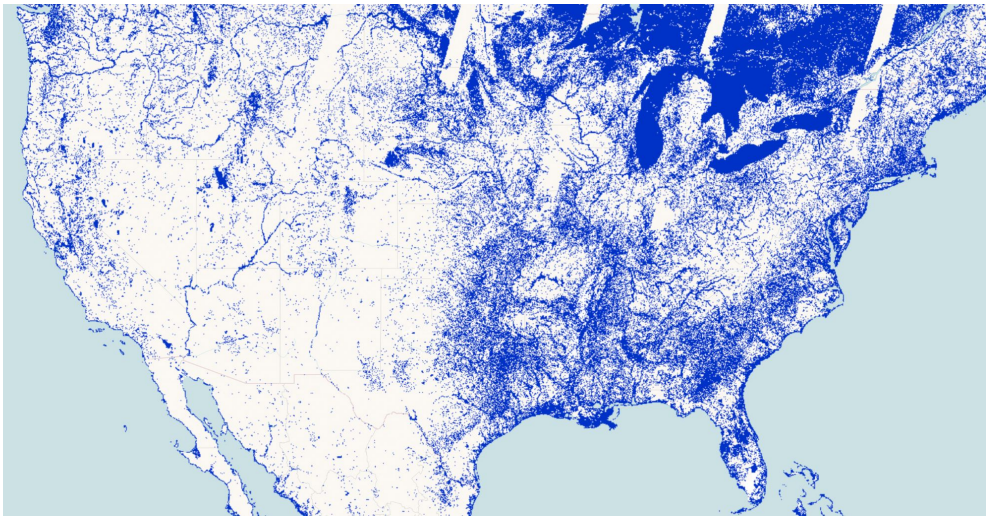


Train set: 8,979 observations
Test set: 4,035 observations

CyFi performs at least as well as Sentinel-3 based tools

And has **10x greater coverage of lakes** across the U.S. thanks to Sentinel-2 imagery!





A true positive (bloom presence) is where cyanobacteria density > 10,000 cells/mL. Uses a dataset of 756 ground measurement observations from across the U.S.

Water bodies detected by Sentinel-2 across the U.S.
Source: Global Water Bodies Product

Primary use cases for CyFi

Low severity

Better allocate ground sampling resources by deprioritizing water bodies where blooms are likely absent

High severity

Support public health interventions by flagging water bodies where severe blooms are likely present

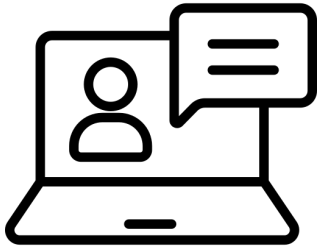
Learnings from developing CyFi

Development of CyFi

Competition

Machine learning competition to assess feasibility and top approaches

User interviews

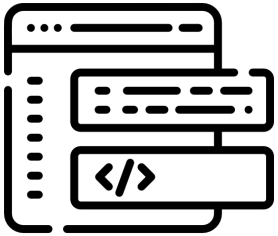


Interviews with water quality managers to understand real-world user needs and decision making workflows

Model iteration

Model experimentation and testing to create a generalizable and accurate model

Python package



Deployable code package capable of generating predictions on new input data

Just because there's an image involved,
doesn't mean you need a neural network

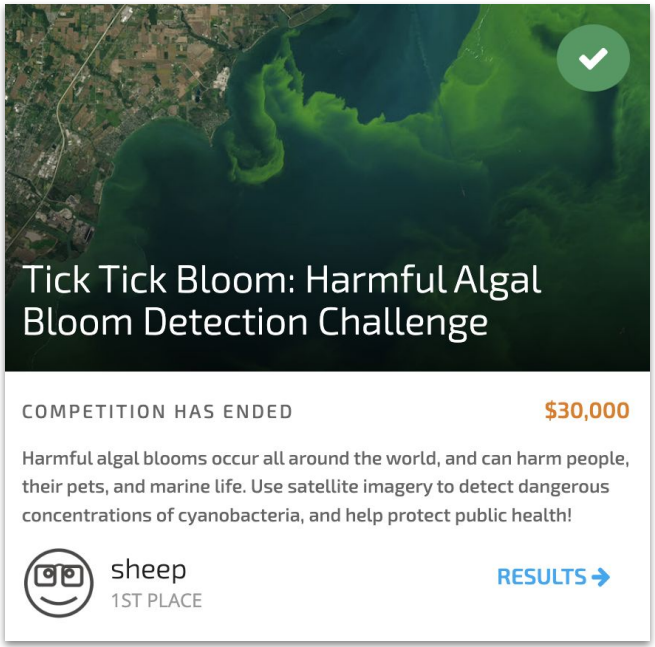
The power of decision trees

All winning approaches used decision trees

Decision trees are a leading approach for satellite-imagery based tasks where the output is point estimation rather than segmentation.

Key benefits

- Quick training
- Quick inference
- Doesn't require a GPU
- Feature importances provide insight into rationale for decision making



Use an out of sample dataset to see if you're getting the right answer for the wrong reasons

The importance of out of sample data

Testing for generalizability

We ran CyFi on newly collected data from summer 2023 in California...and it did not do well.

What went wrong?

Model was estimating high levels of cyanobacteria in most cases

Reviewing images showed this was happening even in cases where water bodies looked dark

→

The culprit: land pixels

The pipeline

- Draw a 200m bounding box around point
- Get images within 7 days prior to sample point
- No cloud filter
- Calculate features from all pixels in an image
- Average predictions across multiple images

The importance of out of sample data

Getting the right answer for the wrong reason

Model had been correctly estimating high levels of cyanobacteria in our test set...but not because of what it saw in the water.

In our test set, points near land tended to have high values of cyanobacteria so our model had incorrectly learned “near land = high severity.”

In the out of sample set, points near land were no longer correlated with high levels of cyanobacteria.

→ The fix: filter to water pixels

Cascading impacts of introducing a water mask

The pipeline

- Draw a 200m bounding box around point
- Get images within 7 days prior to sample point
- No cloud filter
- Calculate features from **only water pixels** in an image
- Average predictions across multiple images

Cascading impacts of introducing a water mask

The pipeline

- Draw a **2,000m bounding box** around point
- Get images within 7 days prior to sample point
- No cloud filter
- Calculate features from **only water pixels** in an image
- Average predictions across multiple images

Cascading impacts of introducing a water mask

The pipeline

- Draw a **2,000m bounding box** around point
- Get images within 7 days prior to sample point
- Only use images with **fewer than 5% cloud pixels**
- Calculate features from **only water pixels** in an image
- Average predictions across multiple images

Cascading impacts of introducing a water mask

The pipeline

- Draw a **2,000m bounding box** around point
- Get images within **30 days** prior to sample point
- Only use images with **fewer than 5% cloud pixels**
- Calculate features from **only water pixels** in an image
- Average predictions across multiple images

Cascading impacts of introducing a water mask

The pipeline

- Draw a **2,000m bounding box** around point
- Get images within **30 days** prior to sample point
- Only use images with **fewer than 5% cloud pixels**
- Calculate features from **only water pixels** in an image
- Use a **single, most recent image** per sampling point

Learnings from the out of sample set lead to a more generalizable model

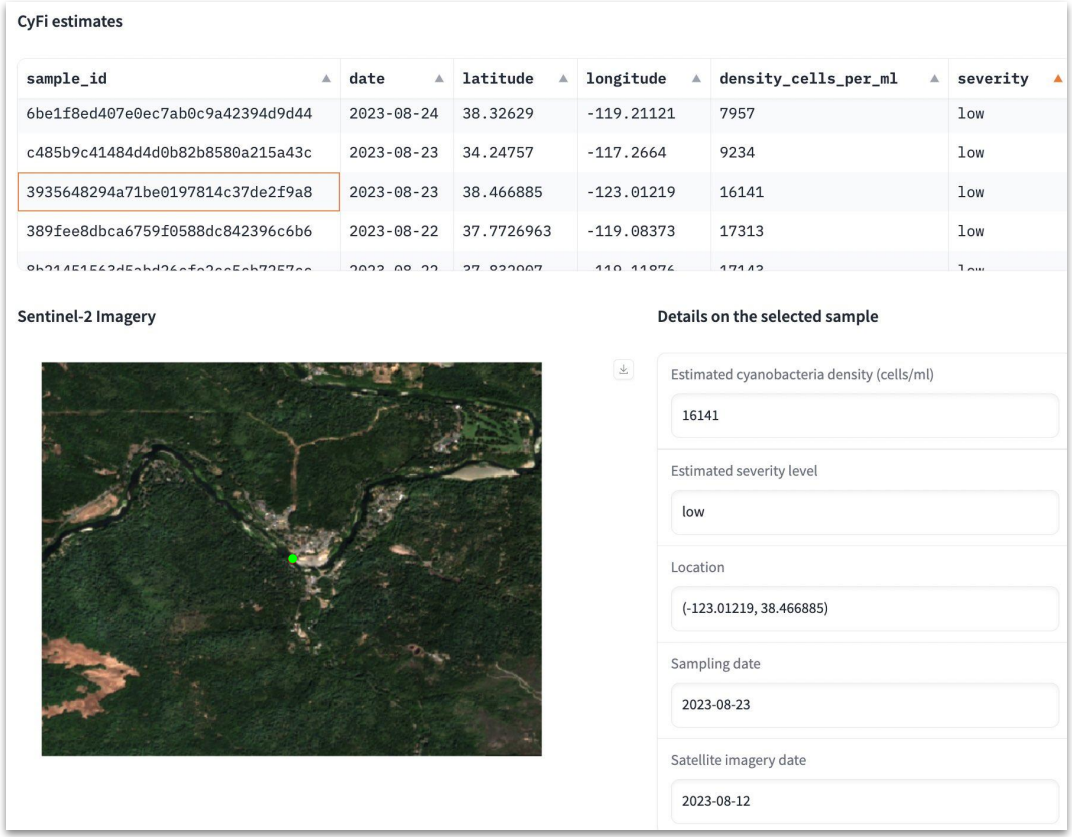
Visual examples help users build confidence and trust in machine learning models

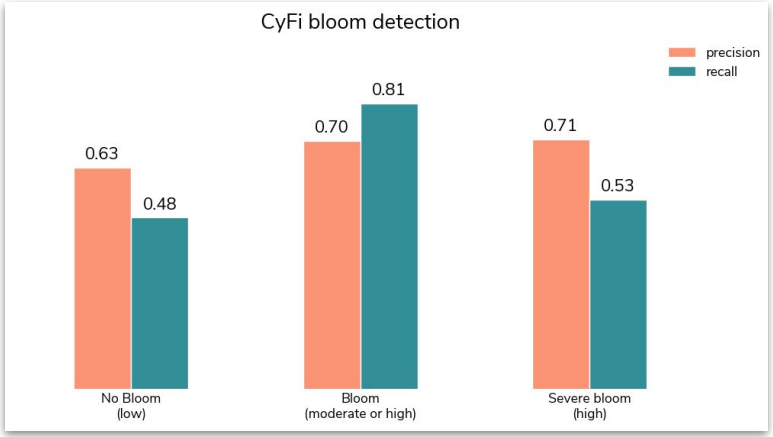
Use visual examples to build trust in the model

Machine learning models can feel like a black box to non-technical users.

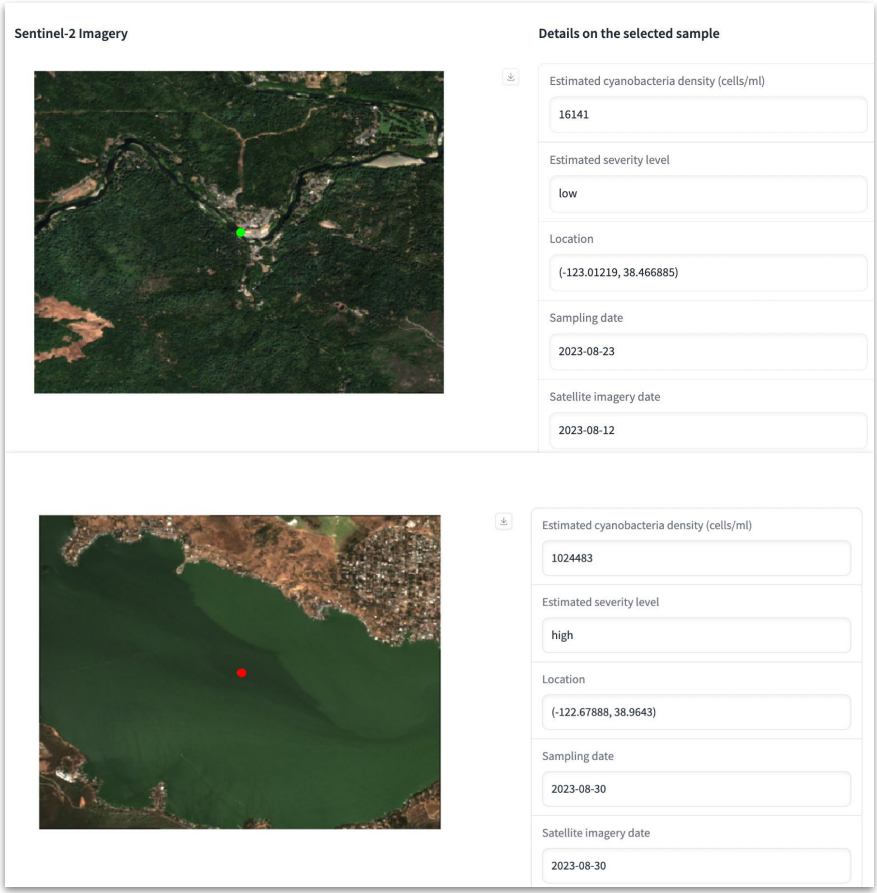
Visual examples are one of the best ways to help users build trust in the model and develop an intuition for cases where the model is more or less reliable.

CyFi includes visualization functionality (Gradio app) which shows you the satellite imagery on which an estimate is based.





Metrics like these are important but they're not the way to help users build trust in a model as they don't build intuition for places where a model does well and does poorly.



Visuals examples, on the other hand, helps users develop an intuition for the decisions that a model makes.

Takeaways

- Just because there's an image involved, doesn't mean you need a neural network
- Use an out of sample dataset to see if you're getting the right answer for the wrong reasons
- Visual examples help users build confidence and trust in machine learning models

To learn more and start using CyFi today, go to:

cyfi.drivendata.org

cyfi Installation Quickstart Visualize Background Accuracy Changelog Search Edit on GitHub

CyFi: Cyanobacteria Finder
Quickstart
About the model

CyFi: Cyanobacteria Finder

CyFi is a command line tool that uses satellite imagery and machine learning to estimate cyanobacteria levels in small, inland water bodies. Cyanobacteria is a type of harmful algal bloom (HAB), which can produce toxins that are poisonous to humans and their pets, and can threaten marine ecosystems.

The goal of CyFi is to help water quality managers better allocate resources for in situ sampling, and make more informed decisions around public health warnings for critical resources like lakes and reservoirs.

Ultimately, more accurate and more timely detection of algal blooms helps keep both the human and marine life that rely on these water bodies safe and healthy.

Stylized view of severity estimates for points on a lake with a cyanobacteria bloom.
Base image from NASA Landsat Image Gallery

Quickstart

Install

Install CyFi with pip:

```
pip install cyfi
```

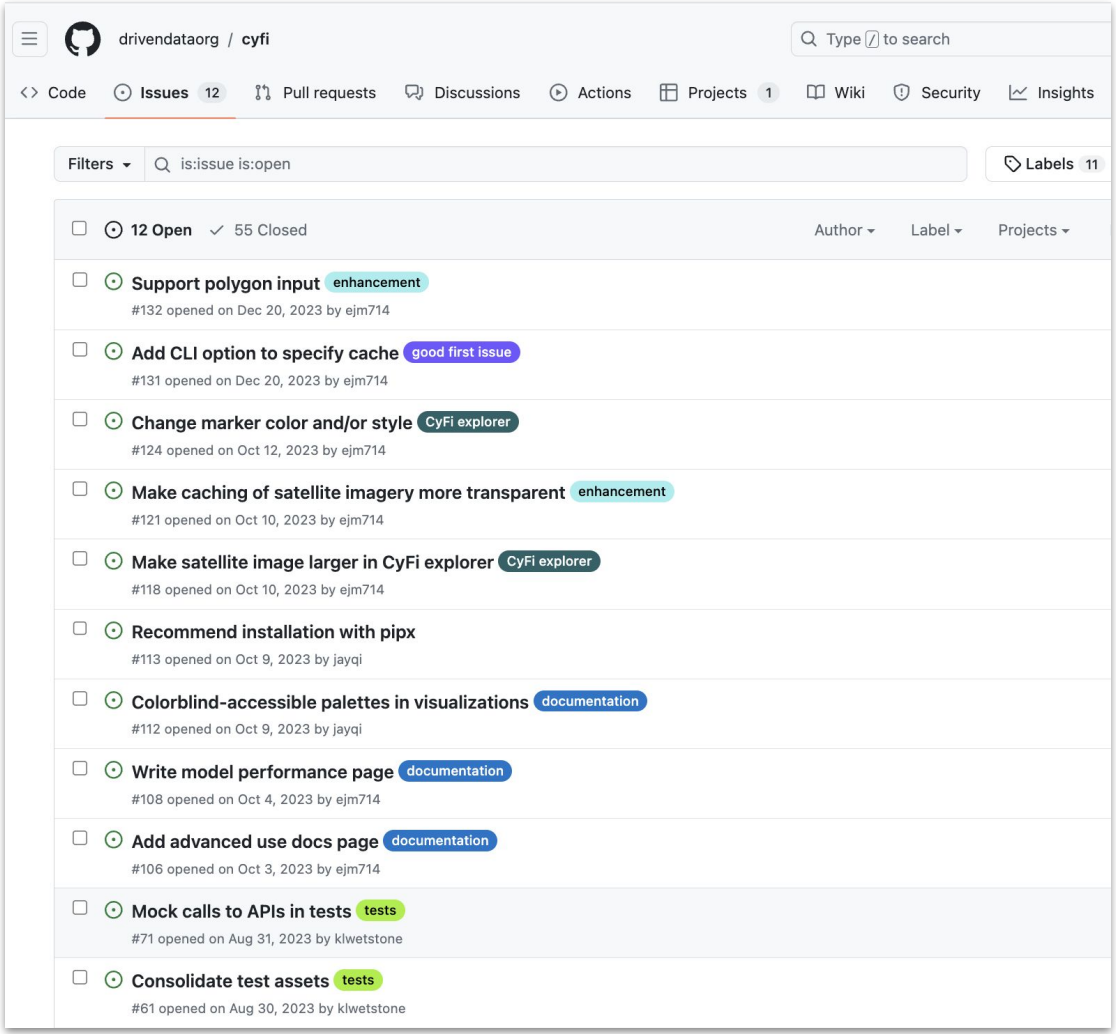
For detailed instructions for those installing python for the first time, see the Installation docs.

Generate batch predictions

Generate batch predictions at the command line with `cyfi predict`.

First, specify your sample points in a csv with the following columns:

CyFi is open source and we welcome contributions!



Thank you!

Questions?
emily@drivendata.org

Check out CyFi:
cyfi.drivendata.org

DRIVENDATA