Imageomics: FAIR ML Products for Biological Knowledge Discovery

Elizabeth G. Campolongo and Matthew J. Thompson

On behalf of the Imageomics Institute

Abstract

A broad goal of the Imageomics Institute is to inspire ML innovation while increasing biological knowledge extraction from images. In furtherance of this goal, we create large and diverse datasets, processing and data exploration tools, and models—big and small—to aid in biological discovery. In this poster we outline many of these open-source tools (on which the poster authors have worked to various degrees) to engage with the broader research community.

Processing Tools

LepidopteraLens — Coming Soon!

Pipeline incorporating preprocessing tools (Imageomics' & other open-source) used to analyze images of butterflies:

- Object detection and segmentation (based on YOLOv8),
- Color standardization (with and without colorchecker),
- Automated Landmarking (using ml-morph by A. Porto),
- Quantification of several phenotypic values.

Outputs: PCA of color and pattern variation using the recolorize and patternize workflow by H. Weller and S. Van Belleghem.

LatLonCover

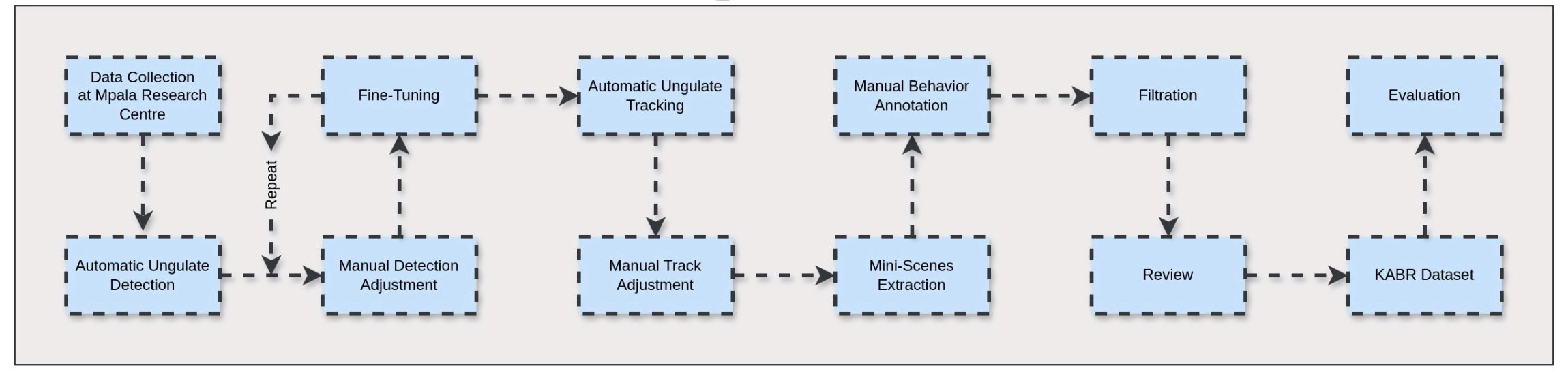
- Generates distribution of landcover types within small and large neighborhoods of given decimal latitude and longitude.
- Uses CropScape land cover designation to determine percentage of each of 7 categories (forest, water, etc.) around given location.
- Available as command line tool, Python API, and HF instance.
- Developed at our Image Datapalooza 2023 workshop.

Video Annotation

KABR Tools

Collection of tools created to annotate animal behavior for the KABR Dataset, generalized to an installable package & CLI tool.

- KABR data: Collected by flying drones over animals in the Mpala Research Centre in Kenya, providing high-quality video footage of their natural behaviors (giraffes and zebras).
  - Resolution: 5472 x 3078 pixels; frame rate: 29.97 frames/sec.


Data Collection at Mpala Research Centre
Fine-Tuning
Automatic Ungulate Tracking
Manual Behavior Annotation
Filtration
Evaluation
Repeat
Automatic Ungulate Detection
Manual Detection Adjustment
Manual Track Adjustment
Mini-Scenes Extraction
Review
KABR Dataset

- Detect objects with Ultralytics YOLO detections, apply SORT tracking and convert tracks to CVAT format.
  - Optional: Fine-tune YOLO for your data.
- Create mini-scenes from your raw footage with these tracks.
- Use the KABR model to label behaviors in the mini-scenes.
- Calculate time budgets.
- Save video with tracking (bounding boxes in player).
- Convert CVAT annotations to Charades format.

THE OHIO STATE UNIVERSITY

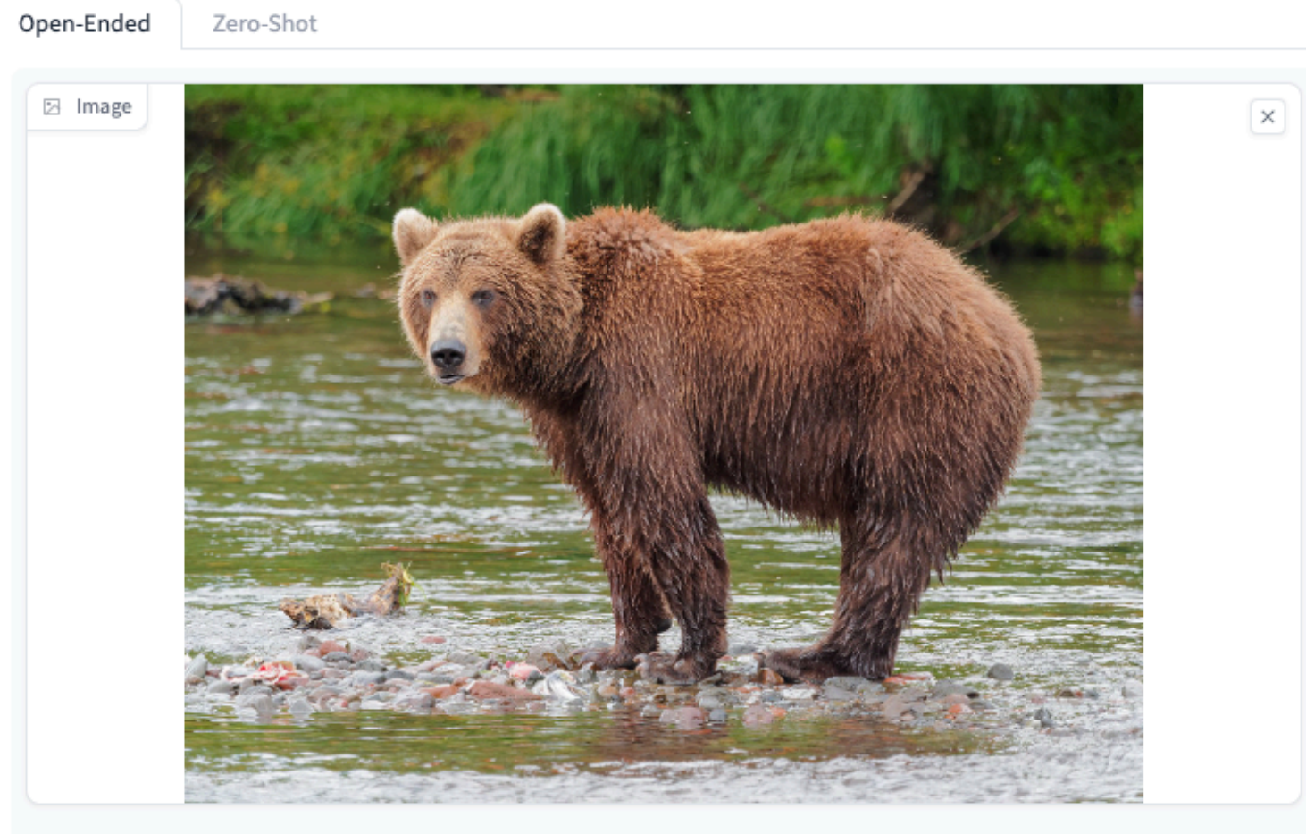
NSF

Biological Foundation Model

BioCLIP-demo

Interactive implementation of BioCLIP model.

- BioCLIP is based on OpenAI's CLIP, trained on TreeOfLife-10M.
- Available on Hugging Face (snapshot below), for anyone to use with no installation or coding required.
- Takes in an image and predicts:
  1. From a given list of classes (Zero Shot), or
  2. To the chosen taxonomic rank from all available taxa in TreeOfLife-10M (Open-Ended).
     - Returns a random sample image from training data of the predicted taxa and a link to the associated page on EOL.



Taxonomic Rank
Which taxonomic rank to predict. Fine-grained ranks (genus, species) are more challengi
Species

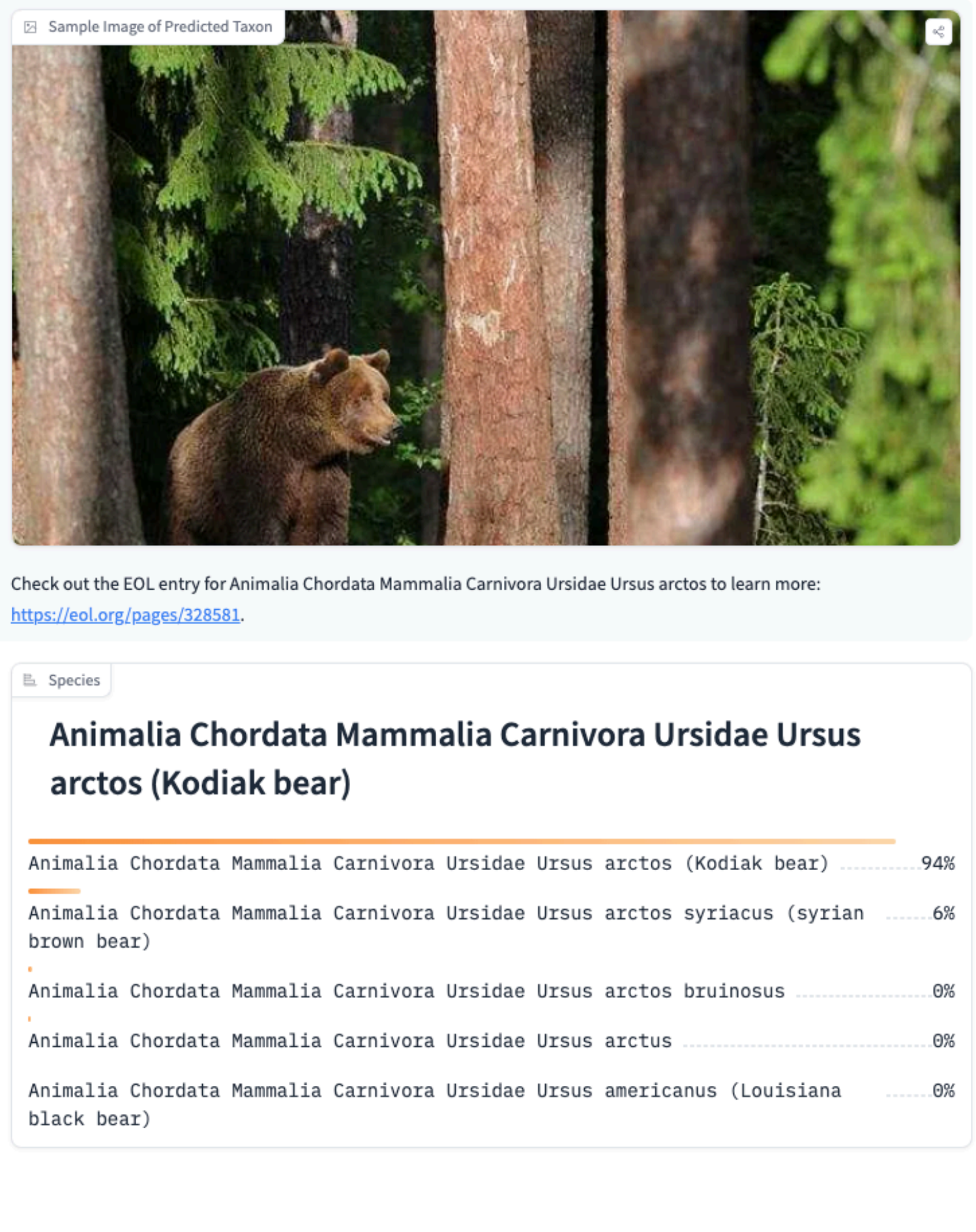
Submit

← Input

- Select image to classify.
- Select taxonomic rank to which to classify from dropdown: species, genus, family, etc.
- Submit for classification.

Output →

- Image is classified by BioCLIP to desired taxonomic rank.
- Top five predicted taxa are displayed.
- Random sample image from that taxonomic rank is pulled from EOL portion of training data (TreeOfLife-10M).
- Link to EOL page for that taxonomic rank is provided.



Check out the EOL entry for Animalia Chordata Mammalia Carnivora Ursidae Ursus arctos to learn more: https://eol.org/pages/328581.

Animalia Chordata Mammalia Carnivora Ursidae Ursus arctos (Kodiak bear)

```
Animalia Chordata Mammalia Carnivora Ursidae Ursus arctos (Kodiak bear) .......... 94%
Animalia Chordata Mammalia Carnivora Ursidae Ursus arctos syriacus (syrian ...... 6%
brown bear)
Animalia Chordata Mammalia Carnivora Ursidae Ursus arctos bruinosus ............. 0%
Animalia Chordata Mammalia Carnivora Ursidae Ursus arctus ....................... 0%
Animalia Chordata Mammalia Carnivora Ursidae Ursus americanus (Louisiana ........ 0%
black bear)
```

pybioclip

Python package and CLI tool designed to make BioCLIP more programmatically accessible.

- Provides access to both predictions and embedding.
- Predict over all available labels (open-ended classification, in the demo) or provide a custom list (or file) of labels from which to predict (Zero-shot, in the demo).
- Custom label prediction extended to other CLIP-based models and checkpoints (e.g., fine-tuned BioCLIP or open_clip B16).
- Example notebooks with implementation available for reference.

Sample Command

```
bioclip predict Ursus-arctos.jpeg
```

Data Validation

sum-buddy ✓∑

Simple and flexible checksum calculation for a dataset.

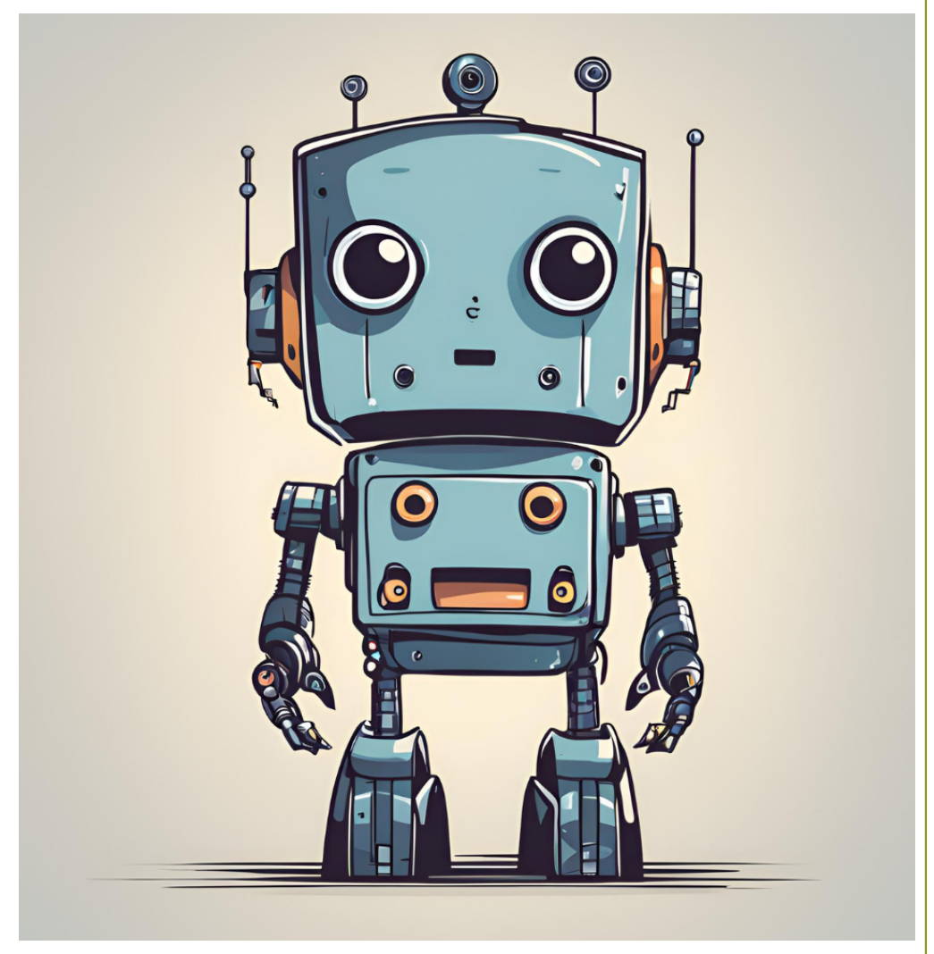
- **Input**: Folder with things to checksum.
- **Output**: CSV or printout of filepaths, filenames, and their checksums.
- **Options**:
  - Ignore subfolders and patterns,
  - Hash algorithm to use,
  - Avoid hidden files and directories.
- **Usage**: Run as a CLI or with exposed Python methods.
- **Use-case**: Duplicate file identification for with flexibility for handling complex hierarchies.

Sample Command

```
sum-buddy --output-file checksums.csv --ignore-file .sbignore data_dir/
```

Data Access

cautious-robot



Simple image from CSV downloader.

- Downloads images from URLs.
  - Configurable wait time & max attempts for retry.
- Names images by given column with unique values.
- Logs all successful responses and errors for review after download.
- Uses sum-buddy to record checksums of all downloaded images.
- Performs minimal check that the number of expected images matches the number sum-buddy counts.

Optional features:

- Organize images into subfolders based on any column in CSV.
- Create square images for modeling:
  - Organizes images in a second directory (same format) with copies of images in specified size.
- Buddy-check: verifies all expected images downloaded intact (compares given checksums with sum-buddy output).

Sample Command

```
cautious-robot --input-file examples.csv --output-dir example_images
```

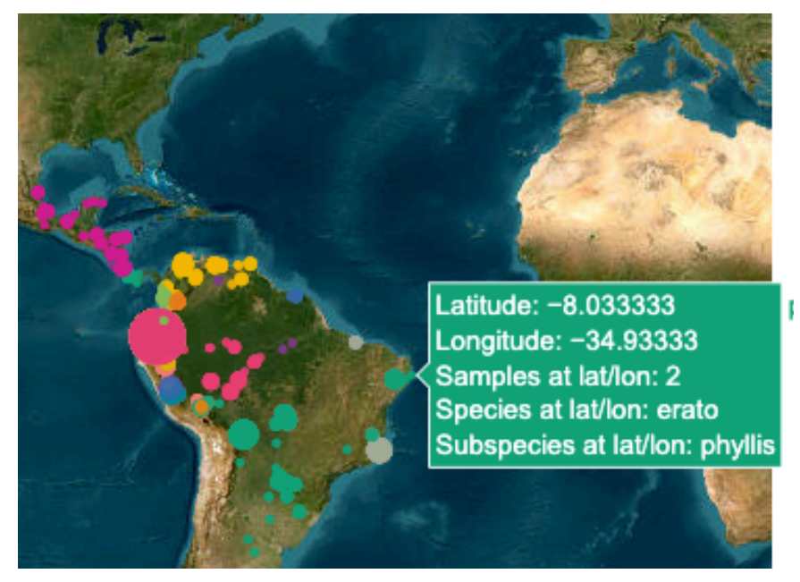
distributed-downloader

Download a huge number of images (on HPC), as quickly as is allowed.

- **Problem**: Have a list of 10s or 100s of millions of images from many sources to download, each with different rate limits:
  - Some sources can handle rapid downloads, others will block an IP address for too many requests.
- **Limited existing solutions:**
  - img2dataset: Fast, but no rate limiting options. Its use could cause disruption or blacklisting.
  - cautious-robot: User-friendly, respectful of data providers, effective for smaller image sets, but it cannot handle massive scales.
- **Solution**: *distributed-downloader* dynamically adjusts download rates to maximize speed while ensuring safe and parallel retrieval, preventing server overload or blacklisting.
  - *Adaptive rate limiting*: Adjusts download rate based on server response times and error codes to stay within acceptable limits.
  - *Scalable*: Distributes workload across nodes, handling large datasets.
  - *Resilient error handling*: Retries for transient issues, logs for persistent issues.
  - *Storage control*: Resizing and checksums possible during the download process.
  - *Detailed logging*: Understand what succeeded, what failed and why, what was filtered out for insufficient resolution or duplication.

Data Exploration Tools

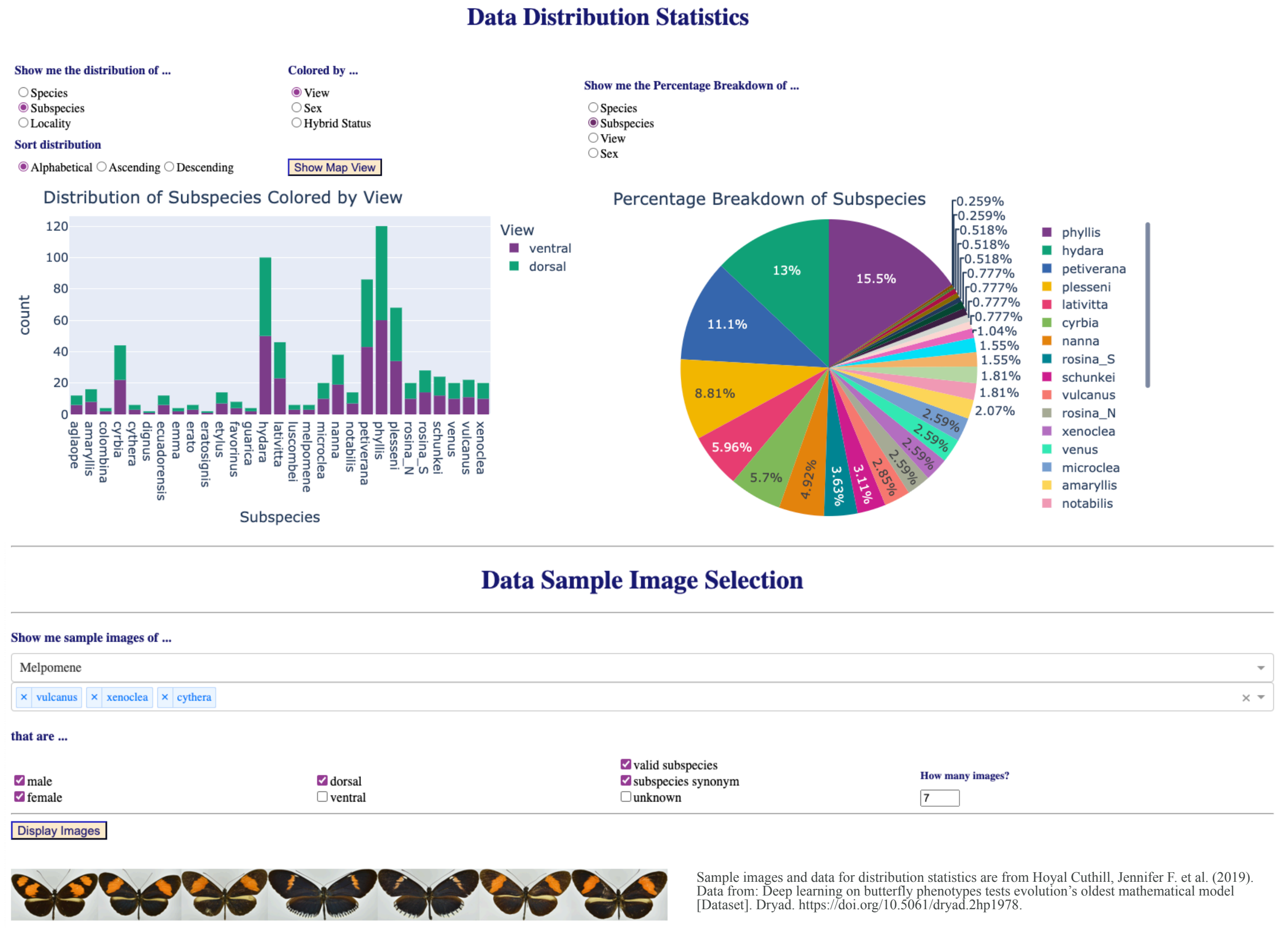
Data Dashboard



Latitude: −8.033333
Longitude: −34.93333
Samples at lat/lon: 2
Species at lat/lon: erato
Subspecies at lat/lon: phyllis

Facilitates data exploration: visualize distribution information and sample images efficiently.

- Telemetry dashboard needs only latitude and longitude, keeps all provided columns.
  - Broader applicability, less organization.
  - Prototype focused on plant and animal images.
- Expanded functionality and scalability in progress.
- Alternate distribution view on ESRI maps, color points by feature.



Data Distribution Statistics

Show me the distribution of ...
○ Species
● Subspecies
○ Locality

Colored by ...
● View
○ Sex
○ Hybrid Status

Sort distribution
● Alphabetical ○ Ascending ○ Descending

Show me the Percentage Breakdown of ...
○ Species
● Subspecies
○ View
○ Sex

Data Sample Image Selection

Show me sample images of ...
Melpomene
× vulcanus × xenoclea × cythera

that are ...
☑ male
☑ female
☑ dorsal
☐ ventral
☑ valid subspecies
☑ subspecies synonym
☐ unknown

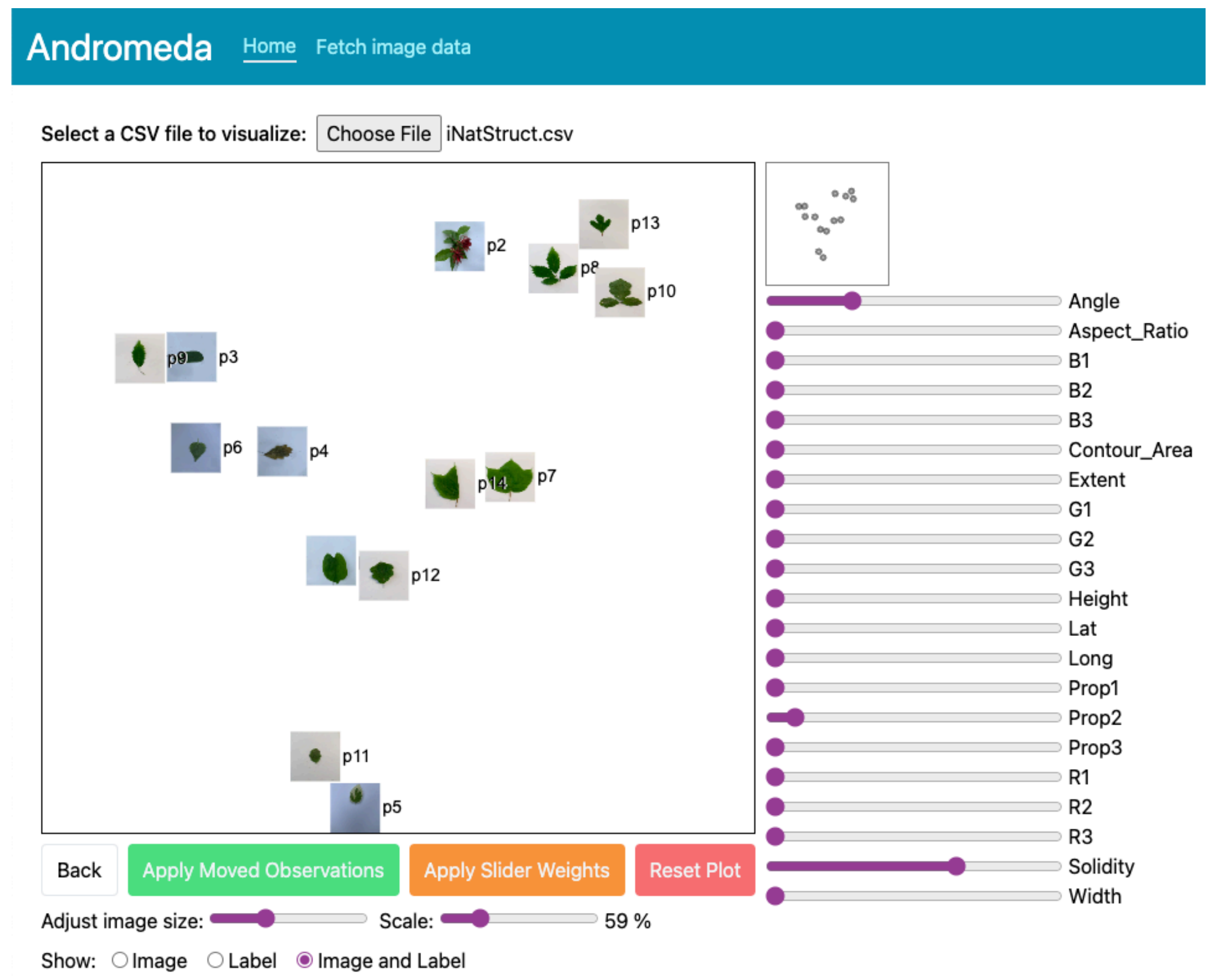
How many images? 7

Sample images and data for distribution statistics are from Hoyal Cuthill, Jennifer F. et al. (2019). Data from: Deep learning on butterfly phenotypes tests evolution's oldest mathematical model [Dataset]. Dryad. https://doi.org/10.5061/dryad.2hp1978.

Andromeda: FAIR high-dimensional data exploration

Interact with high-dimensional data in a 2D plot using WMDS.

- Adjust weights and see how positions change; OR
- Adjust position of data points and see which features are more heavily weighted, then apply those weights to the full projection.
- Hugging Face instance also includes fetch data from iNaturalist option with a LatLonCover integration.



Select a CSV file to visualize: iNatStruct.csv

Adjust image size: Scale: 59 %

Show: ○ Image ○ Label ● Image and Label

Follow us on Hugging Face and GitHub!

Hugging Face

GitHub

This material is based upon work supported by the National Science Foundation under Award No. 2118240.