Fish-Climate Dataset

Where did it come from? What does it include?

Before we get started…

- Download the dataset called

  `FishClimQuestAll.csv`

- For this presentation, the data may be saved in any format. E.g., excel, google sheets, numbers…
- Open the dataset and follow along as we discuss it.

Outline

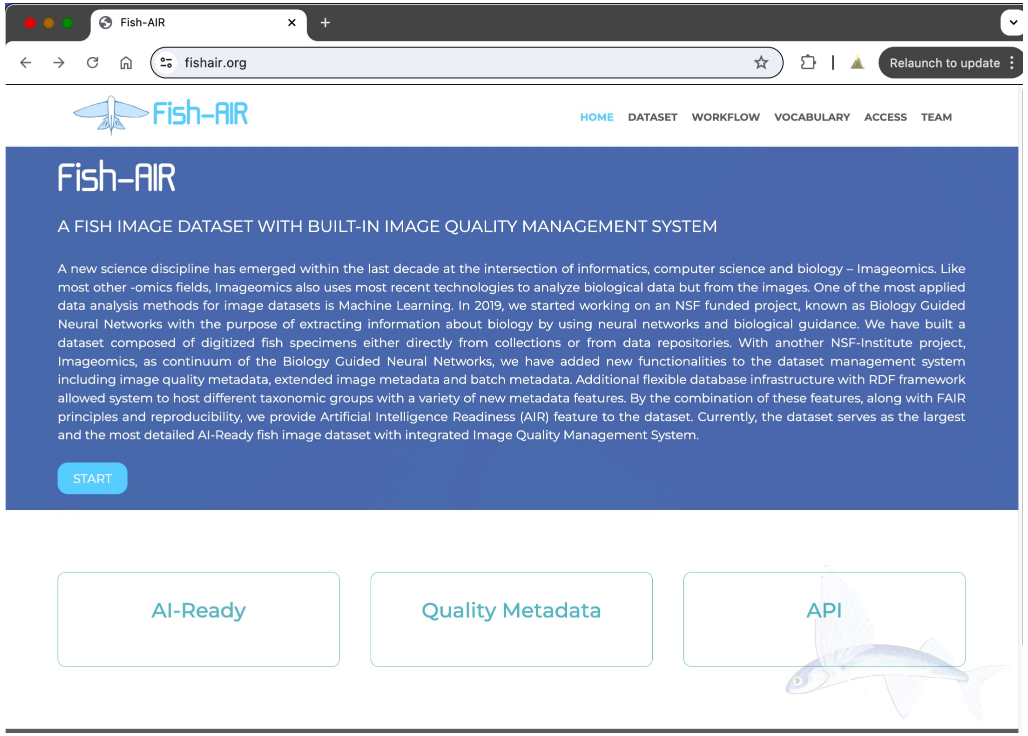
- Where did the data come from?
- What are the variables?

Where did the data come from?

`fishClimQuestAll.csv` combines two databases:

FishAIR (fishair.org)

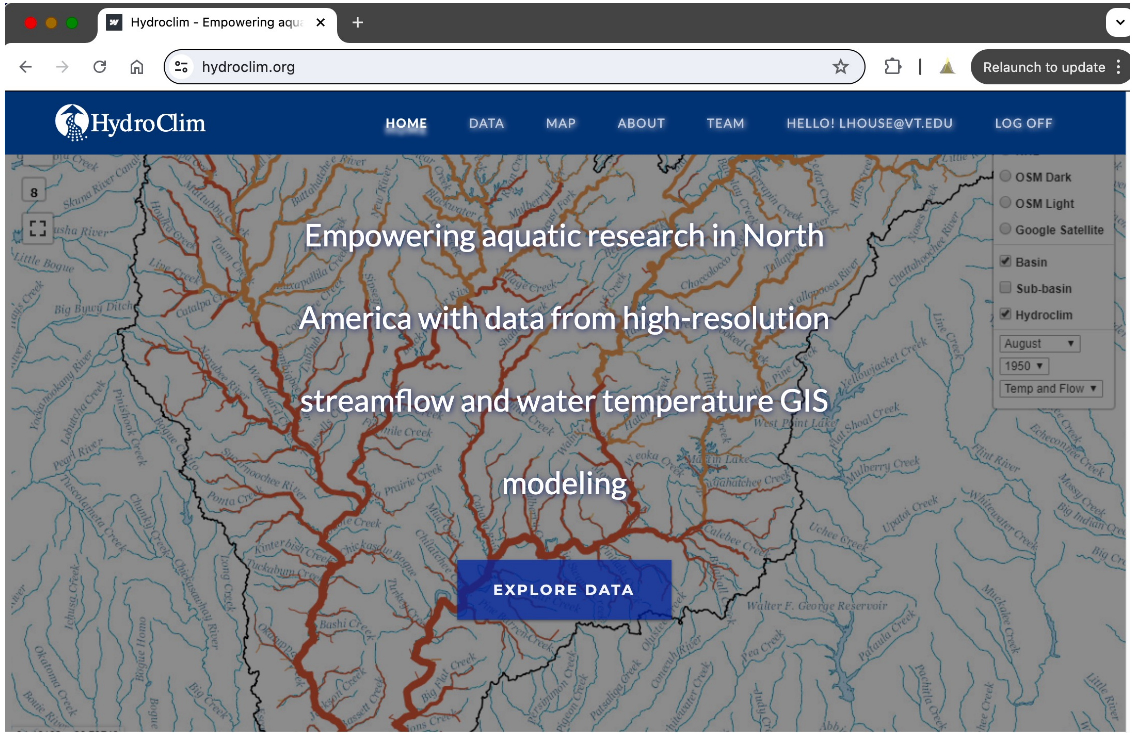
Hydroclim (hydroclim.org)



Fish-AIR

A FISH IMAGE DATASET WITH BUILT-IN IMAGE QUALITY MANAGEMENT SYSTEM

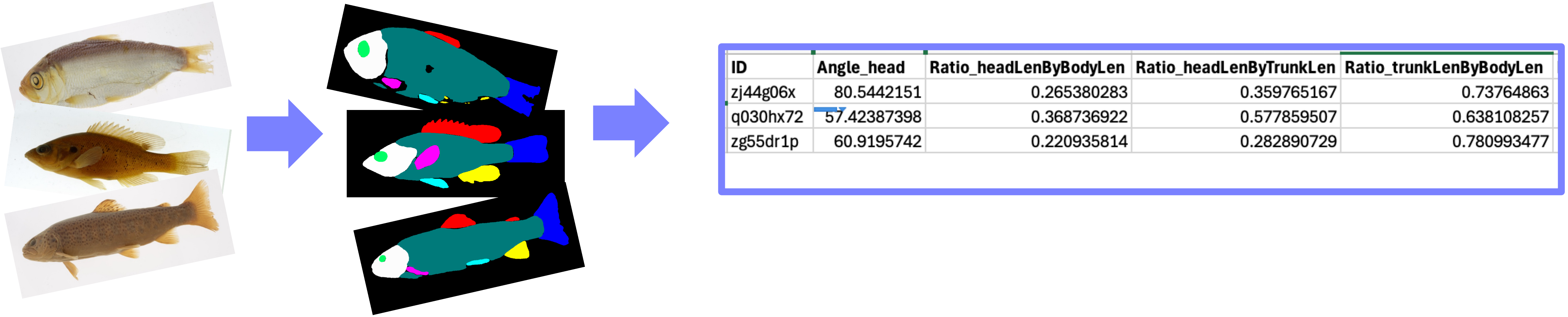
A new science discipline has emerged within the last decade at the intersection of informatics, computer science and biology – Imageomics. Like most other -omics fields, Imageomics also uses most recent technologies to analyze biological data but from the images. One of the most applied data analysis methods for image datasets is Machine Learning. In 2019, we started working on an NSF funded project, known as Biology Guided Neural Networks with the purpose of extracting information about biology by using neural networks and biological guidance. We have built a dataset composed of digitized fish specimens either directly from collections or from data repositories. With another NSF-Institute project, Imageomics, as continuum of the Biology Guided Neural Networks, we have added new functionalities to the dataset management system including image quality metadata, extended image metadata and batch metadata. Additional flexible database infrastructure with RDF framework allowed system to host different taxonomic groups with a variety of new metadata features. By the combination of these features, along with FAIR principles and reproducibility, we provide Artificial Intelligence Readiness (AIR) feature to the dataset. Currently, the dataset serves as the largest and the most detailed AI-Ready fish image dataset with integrated Image Quality Management System.



Empowering aquatic research in North America with data from high-resolution streamflow and water temperature GIS modeling

FishAIR

- Includes AI-ready (AIR) images for 27,082 fish specimen.
- I.e., includes image data of 27,082 fish that have been *structured*
- Tulane: Developed FishAIR to
  Access, Coordinate, Clean, Process, Extract features from, and Share
  fish images from the internet, museums, academic labs, …



ID	Angle_head	Ratio_headLenByBodyLen	Ratio_headLenByTrunkLen	Ratio_trunkLenByBodyLen
zj44g06x	80.5442151	0.265380283	0.359765167	0.73764863
q030hx72	57.42387398	0.368736922	0.577859507	0.638108257
zg55dr1p	60.9195742	0.220935814	0.282890729	0.780993477

FishAir Includes

- **Specimen identifier:** ImageID
- **Meta data:** Specimen Genus, Family, Scientific Name, Location, State, Basin, BasinID, Subbasin, image urls
- **Specimen features**: Angle_head, Ratio_bodyWidByBodyLen, … , Loc_eyeOnHeadHoriz, …
- **Proxy for image quality:** Score

Sample size

- For this dataset, we aimed to have as many species as possible with at least 40 images per species.
- Thus,
  - species were deleted for this dataset that didn’t have enough specimens in FishAIR
  - 40 images were selected from FishAIR, when the species had more than 40 images.
- Notably, our final dataset resulted in species having fewer than 40 images for two reasons:
  - Unclear images (according to the score variable) were removed.
  - Images that lacked lat-lon coordinates for merging with hydroclim were removed.
- At the end of this presentation final dataset summaries are provided

HydroClim

- Includes
  - monthly observed and predicted (from climate models) streamflow & water temperature for stream sections in all major watersheds as defined by regions (basins and subbasins) across the United States from 1950-2099.
- Researchers from Tulane downloaded and processed observed monthly data from 1950, 1960, …, 2010, 2020 to create season (sp, fa) summaries.

Merging FishAIR and HydroClim

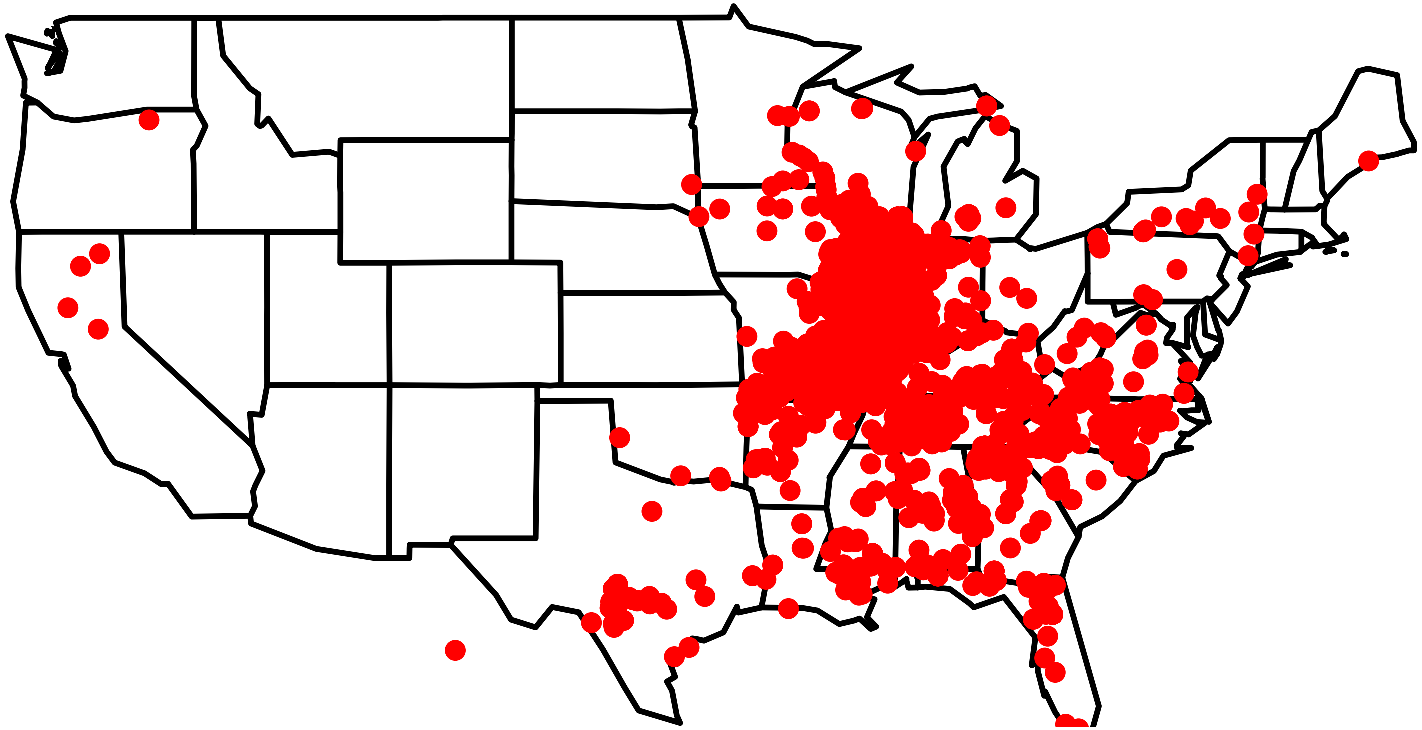
- Data from FishAIR were merged with data summaries from Hydroclim based on latitude and longitude.
- Specifically,
  - FishAIR: meta data of images from FishAIR included latitude and longitude for where the specimens were collected.
  - HydroClim included data for subbasins within basins that are defined by geographical regions.
  - When latitude and longitude of Fish images fell within a subbasin, the data were merged.
- When meta data of images didn’t report latitude and longitude, the image was dropped from the data.

Data Dictionary

- `fishClimQuestAll.csv` includes
  - 2370 observations (i.e., images of fish specimen)
  - 130 Variable
- Data dictionaries define variables in the dataset.
- The data dictionary for `fishClimQuestAll.csv` is called `fishClimateDictionary` or an alphabetized version is called `fishClimateDictionaryAlph`

Data Summaries of `fishClimQuestAll`

- 130 Variables
- 2370 Observations
- 79 Species: 1-40 with average of 30 images per
- Specimens collected from 1882-1916; 33 images missing collection year.



Specimen location marked in red.

Have fun with `fishClimQuestAll.csv`!