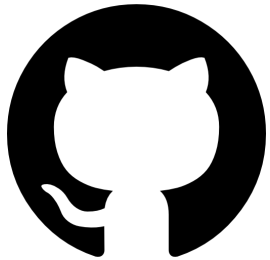
HDR Institute Imageomics

Charles Stewart
Professor, Department of Computer Science,
Rensselaer Polytechnic Institute

This presentation represents the combined efforts of the members of the Imageomics team.

Data/cyberinfrastructure best practices: Code

- GitHub
  - For code, documents and docker images
- Culture shift for biologists
- Goal is full adoption
  - Imageomics course

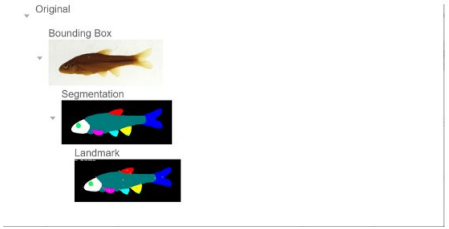


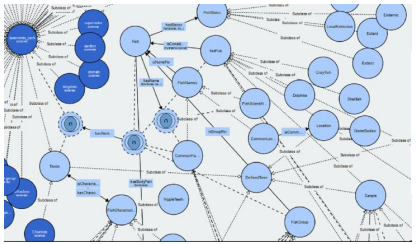
Data/cyberinfrastructure best practices: Data

- Raw images and associated metadata
  - Extensive initial organization
  - ARK Id's
- Biological knowledge:
  - Ontologies
  - Phylogenetic relationships (trees)
  - Knowledge graphs
  - Taxonomic keys
- Intermediate results
  - Curated image sets
  - Results of preprocessing / derived metadata
  - Model training splits
  - Output at each stage of eachpipeline
- ML models
- "Final" ML results
  - Trait predictions, species classifications, distance metrics
  - These become *inputs to biological trait analysis*

Multimedia	Image Quality Metadata	Extended Image Metadata
Ark ID	6j82m58s	
AccessURI	https://bgnn.tulane.edu//ark:/89609/6j82m58s.jpg	
BatchName	GLIN-INHS	
CreateDate	2020-03-02	

Related Images

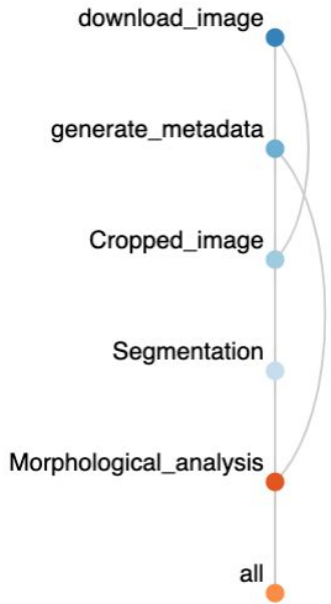




Data/cyberinfrastructure best practices: Reproducibility Via Workflows

- Goals:
  - End-to-end reproducibility
  - Drop-in replacements of algorithms / tuned parameters
- Based on Snakemake
  - Components are docker containers
    - Hosted on github
    - Rebuilt automatically
- Prototype example: fishes analysis from our pre-cursor BGNN project

snakemake.github.io



Data/cyberinfrastructure best practices: Computation

- Local GPUs
  - One GPU machine with 8 NVIDIA A6000 GPUs
  - For quick debugging and initial model training
- Ohio Supercomputer Center
  - 2 GPU clusters (160 NVIDIA Tesla P100; 164 NVIDIA Volta V100)
  - One more (~100 NVIDIA A100 GPUs) being commissioned
- AWS for large scale training
  - Foundation model

Data/cyberinfrastructure best practices: Models

- HuggingFace
- Docker images
- Auto-configured and auto-generated
- …
- First deployment in progress

HUGGING FACE

Impact: HDR and larger cyberinfrastructure

- Workflow / metadata management for
  - Complex scientific systems from heterogeneous and multiple data sources
  - Reproducibility
- Data / model products
  - Biology-focused foundation models and knowledge-guided models
- Biological products
  - Ontologies
  - Trait / phylogeny and other predictions / hypotheses

Current data/cyberinfrastructure needs

- Access to more compute power for training
- Versioning of intermediate data results
- Easier access to large-scale data sets
- Continued cultural improvements: ethos of reproducibility
  - Commits
  - Workflows
  - Data storage and data versioning
  - Maintenance

  Sharing of curated data and problems to expand community

Open Q&A/Discussion and Conclusion