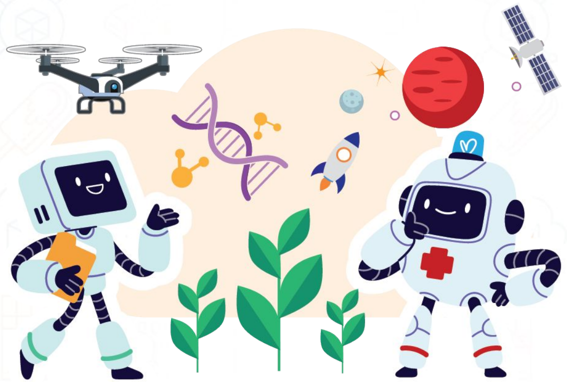


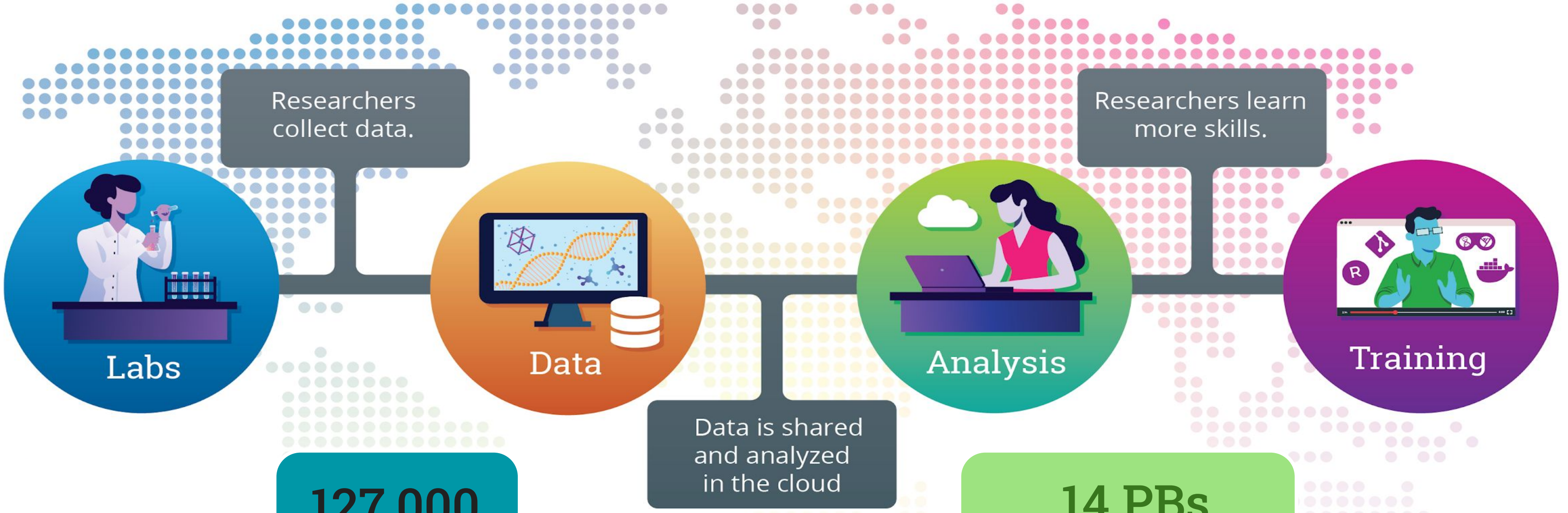
Data Management for ML & AI Projects Using CyVerse

Nirav Merchant (nirav@arizona.edu)
PI CyVerse
Dir. Data Science Institute
University of Arizona

THE UNIVERSITY OF ARIZONA CYVERSE
Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442
RESEARCH, INNOVATION & IMPACT Data Science Institute

CYVERSE®

Delivering the cyberinfrastructure and training to make the cloud easy, collaborate across the world, and scale analyses



127,000 users

14 PBs user data

845 training events

160 countries

46,000 people trained

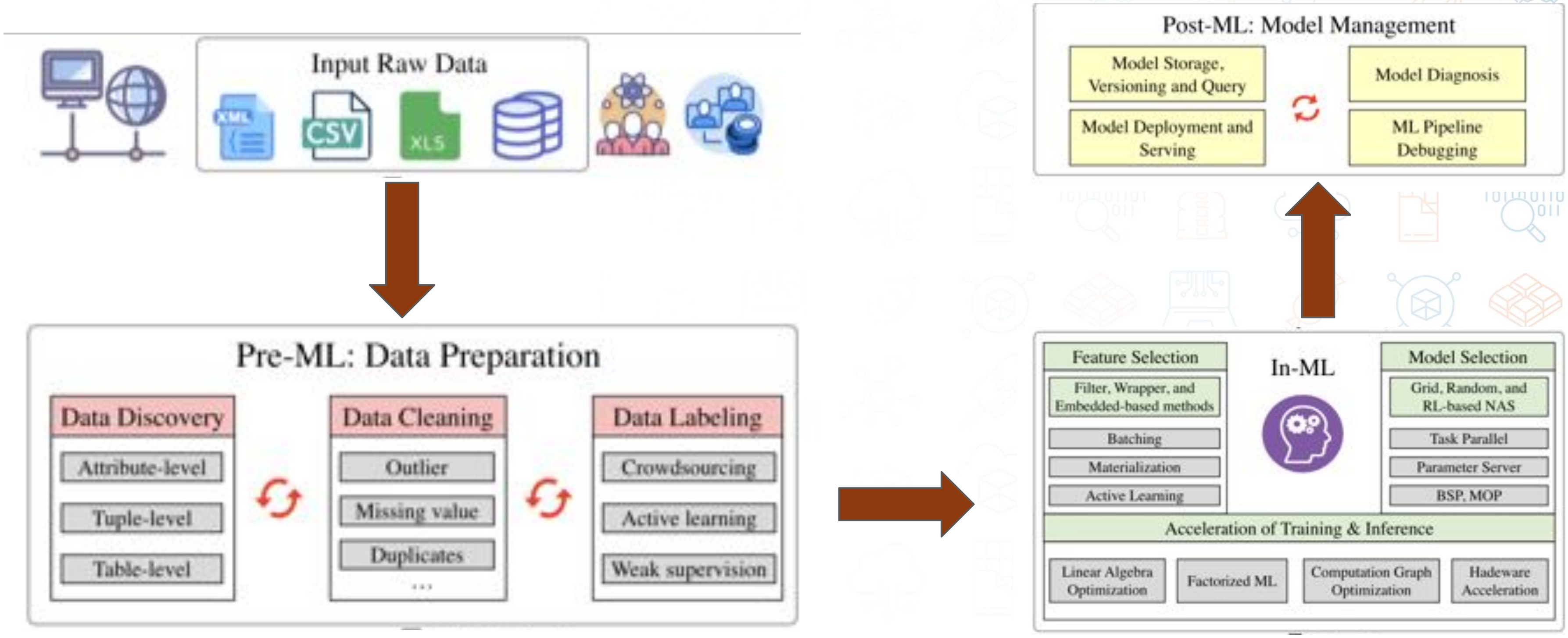
1,800 publications citing CyVerse

What we'll cover today

- What is unique about AI/ML projects ?
- Data Engineering+Data Management
- Tools for Data campaigns in Academia
- Sharing and Publishing

The University of Arizona
CyVerse
Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442
Research, Innovation & Impact Data Science Institute

Data Lifecycle for ML/AI applications



Data Management for Machine Learning: A Survey *IEEE Transactions on Knowledge and Data Engineering May 2023 doi: 10.1109/TKDE.2022.3148237*

THE UNIVERSITY OF ARIZONA
CYVERSE
Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442
RESEARCH, INNOVATION & IMPACT
Data Science Institute

Working with Data Sources & Building Data Sets

Katherine Scott
@kscottz

One of the biggest failures I see in junior ML/CV engineers is a complete lack of interest in building data sets. While it is boring grunt work I think there is so much to be learned in putting together a dataset. It is like half the problem.

♡ 571 11:50 AM - Feb 1, 2019

mat kelcey
@mat_kelcey

for my last few ML projects the complexity hasn't been in the modelling or training; it's been in input preprocessing. find myself running out of CPU more than GPU & in one project i'm actually unsure how to optimise the python further (& am considering c++ for one piece)

♡ 130 2:01 PM - Feb 11, 2019

Vicki Boykis
@vboykis

Have been extremely curious about this for a while now, so I decided to create a poll.
"As someone titled 'data scientist' in 2019, I spend most of (60%+) my time:"
("Other") also welcome, add it in the replies.

♡ 189 8:17 AM - Jan 28, 2019

- 6% Picking features/models
- 67% Cleaning data/Moving data
- 4% Deploying models in prod
- 23% Analyzing/presenting data

2,116 votes • Final results

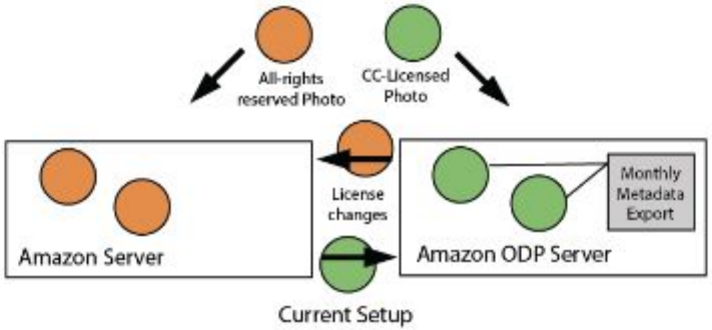
https://fullstackdeeplearning.com/course/2022/lecture-4-data-management/

Reality of Raw Data to Data Sets

- Your data will come from many sources
- Most likely it will not have the compute resources you need in that location
- Be prepared to be a GPU scavenger
- Move your data to where you can get the GPU's
- You need a common place to collect data from multiple sources
- You will have multiple versions of the Data Set
- You need an automation (pipeline) to reliably get the data at scale (and be prepared to repeat it)
- Once you have a template you can repeat the steps, but every project will have its unique need

The University of Arizona CYVERSE NSF Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442 Research, Innovation & Impact Data Science Institute

So what does a Data foraging workflow look like ?



- https://github.com/inaturalist/inaturalist-open-data
- Observation Access: http://inaturalist-open-data.s3.amazonaws.com/photos/[photo_idl]/[size].jpg

Original	Large	Medium	Small	Thumb	Square
2048px	1024px	500px	240px	100px	75px x 75px

Metadata Columns

1. Observations
    a. observation_uuid
    b. observer_id
    c. latitude
    d. longitude
    e. positional_accuracy
    f. taxon_id
    g. quality_grade
    h. observed_on
2. Observers
    a. observer_id
    b. login
    c. name
3. Photos
    a. photo_uuid
    b. photo_id
    c. observation_uuid
    d. observer_id
    e. extension
    f. license
    g. width
    h. height
    i. position
4. Taxa
    a. taxon_i
    b. ancestr
    c. rank_lev

AIIRA
CYVERSE

So what does a Data foraging workflow look like ?

Data Extraction for Classification

Challenges:

- Depth of hierarchy varies for different species, e.g. some levels are missing in the phylogenetic tree for certain species
- Image-by-image querying from iNaturalist website
    - very time consuming, could potentially take months to years
    - Not feasible for dataset size in the range of millions of images
- **Bulk download by species is not available**

AIIRA
CYVERSE®

So what does a Data foraging workflow look like ?

iNaturalist Scalable Download (iNatSD)

This tool allows users to easily download species-level images under the hierarchy of a specific taxon in the iNaturalist format. You are able to acquire high quality labeled images of organisms for research or any other purpose.

Snakemake workflow combined with Python allows for easy to access pipelines that can download customizable datasets.

Evolution of iNatSD to LDAWT

- Became progressively challenging to support multiple data campaigns
- iNatSD to Large Data Acquisition Workflow Template (LDAWT)
- LDAWT decreased the acquisition time by a significant margin
- Able to download the Biotrove dataset (40TB) in roughly 8 hours by utilizing distributed downloading.
- LDAWT is portable and can be utilized on any HPC, NSF ACCESS resource with sufficient bandwidth

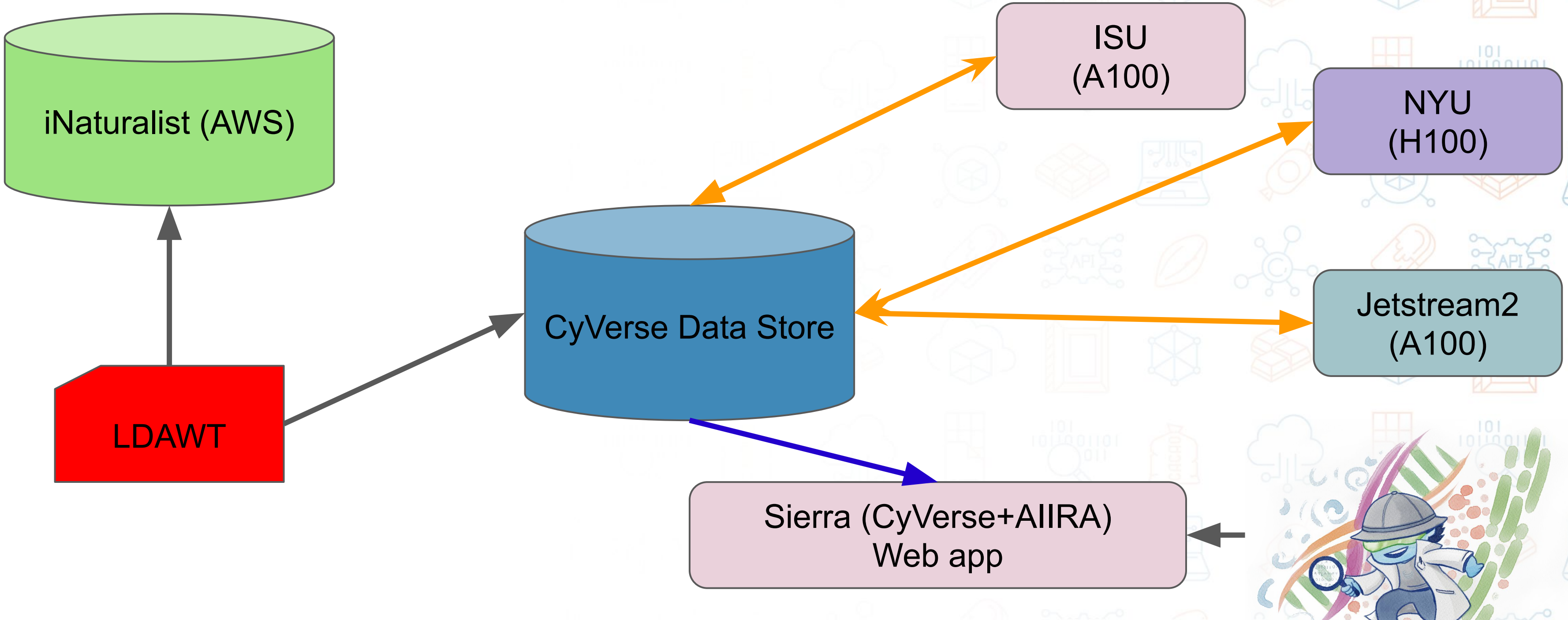
- **Biotrove**- A 134.6 million dataset of image-language pairs for biodiversity assessment and agricultural research.

- **InsectNet** - Utilizes a curated 6 million image dataset of 2526 pest species achieve 96.4% accuracy on pest images.

- **WeedsNet** - Utilizes a curated 13 million image dataset of 1581 weed species and achieves an accuracy of 86.7% on weed images

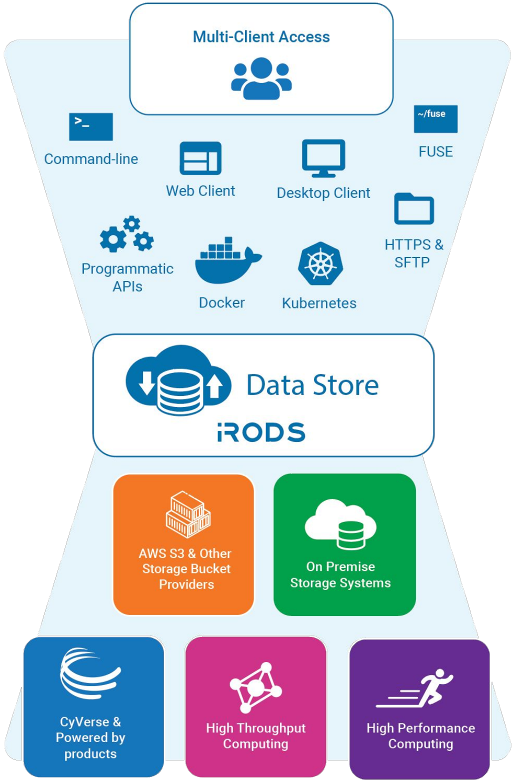
THE UNIVERSITY OF ARIZONA
CYVERSE®
NSF
Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442
RESEARCH, INNOVATION & IMPACT
Data Science Institute

What happens behind the scene ?



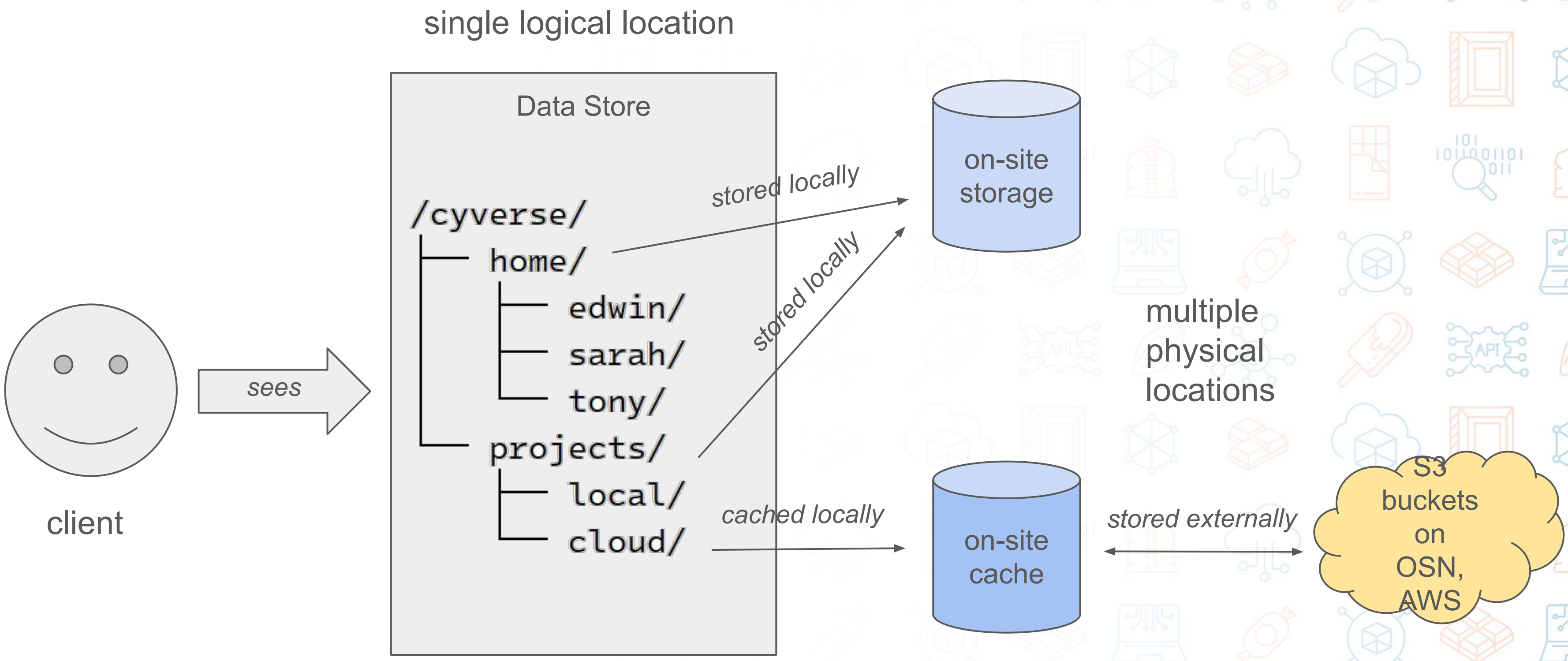
The University of Arizona
CYVERSE
Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442
Research, Innovation & Impact
Data Science Institute

Data Store Overview



- **Data Accessibility**
- Data Sharing
- Data Discovery
- **Data Virtualization**
- Policy Automation

Data Store Storage Virtualization



The University of Arizona
CYVERSE
Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442
Research, Innovation & Impact Data Science Institute

CyVerse Data Store Tools for Accessing Data (automation)

All major computing environments supported

arm

IOT

GUI

iDrop

Cloud Browser

metalnx

SFTP

Transfer Optimizations Coverage

- large sets of small files
- very large files
- fast networks
- unreliable networks

File System Mount

WebDAV/HTTPS

FUSE

NFS

Programmatic

GO

Python

Java

REST APIs

Command Line

GoCommands

iCommands

The University of Arizona

CYVERSE

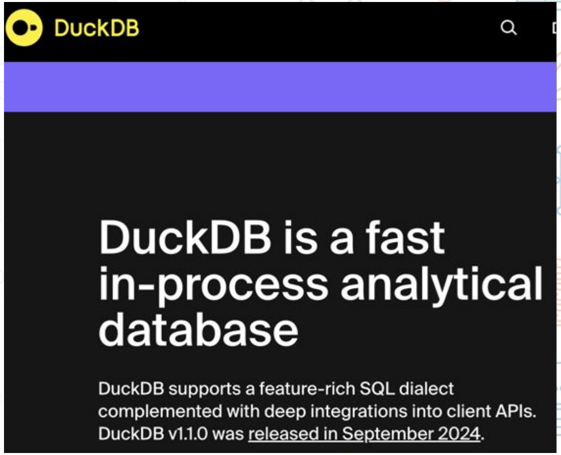
NSF

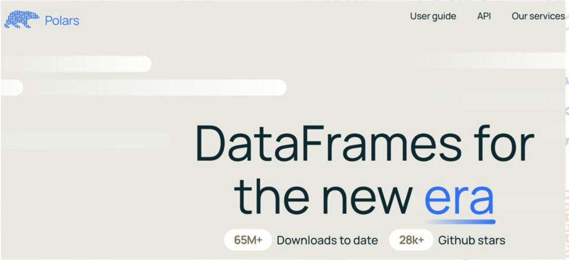
Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442

Research, Innovation & Impact
Data Science Institute

Cloud Native Formats & Data Lakes

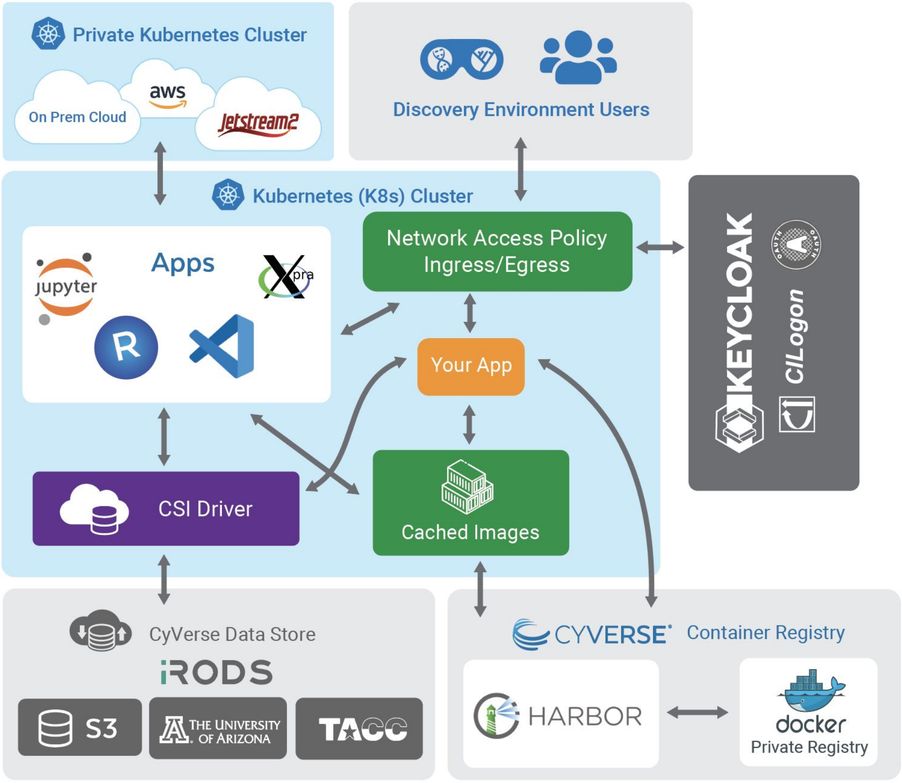
- Binary data (images, audio): find ways to combine/group (tar/gz) into units
- For metadata (labels) / tabular data / text data: include manifest
- Compressed csv/json/txt files are convenient (but not easy to query)
- Cloud native and object store friendly formats are ideal; convert text formats into these formats
- Parquet is a table format that's fast, compact, and widely used
- Polars is a amazingly fast dataframe to convert, etc.
- Use a Duckdb at the analytics engine on top
- **All of this together gives you an economical Data lake**





The University of Arizona
CYVERSE
NSF
Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442
Research, Innovation & Impact
Data Science Institute

Discovery Environment (DE) aka Data Lake House



The University of Arizona
CYVERSE
NSF
Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442
Research, Innovation & Impact
Data Science Institute

Where do you publish the models, weights and data ?

- Respect copyright and permissions before you download/scrape content
- Most journals do not want 10 TB datasets
- Institutional repositories are also not equipped to handle large ML data sets
- Do not abuse free resources (Zenodo, etc.) by chunking data
- Hugging face is a great option (for now) and has great tooling and is not aggressively policing data set limits (but we all know how that movie ends)
- You need a clear (affordable) strategy to make your data public from these campaigns
- Think of forming a data commons (ML commons) for your community/team

The University of Arizona
CyVerse®
NSF
Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442
Research, Innovation & Impact
Data Science Institute

What does an ML-friendly Data Commons look like ?

ML Commons

Mission

Standardize how ML datasets are described to make them easily discoverable and usable across tools and platforms.

Purpose

Data is paramount in machine learning (ML). However, finding, understanding and using ML datasets is still unnecessarily tedious. One reason is the lack of a consistent way to describe ML datasets to facilitate reuse. That's the aim of Croissant.

Croissant is an open community-built standardized metadata vocabulary for ML datasets, including key attributes and properties of datasets, as well as information required to load these datasets in ML tools. Croissant enables data interoperability between ML frameworks and beyond, which makes ML work easier to reproduce and replicate.

ML Commons

March 6, 2024 News

New Croissant Metadata Format helps Standardize ML Datasets

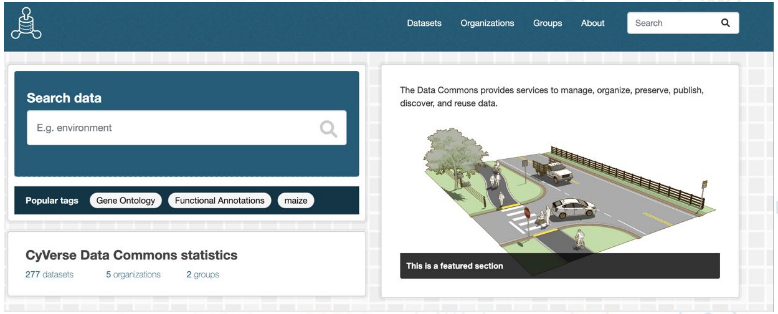
Support from Hugging Face, Google Dataset Search, Kaggle, Open ML, and TFDS, makes datasets easily discoverable and usable.

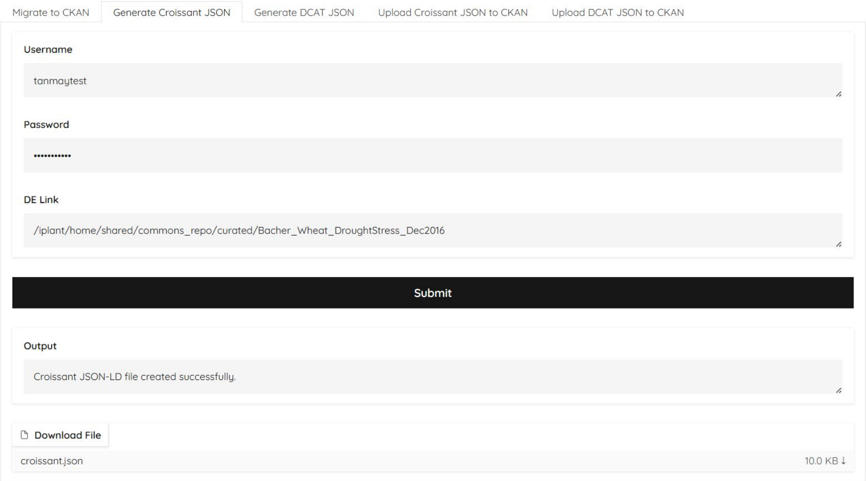
https://mlcommons.org/

The University of Arizona
CYVERSE
Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442
Research, Innovation & Impact
Data Science Institute

What is CyVerse providing for commons ?

- Data Store with http and S3 interface (2025)
- Integration with CKAN to connect data sources from any external provider for creating project specific Data (ML) Commons
- Tools to automatically convert Metadata to Croissant complaint format





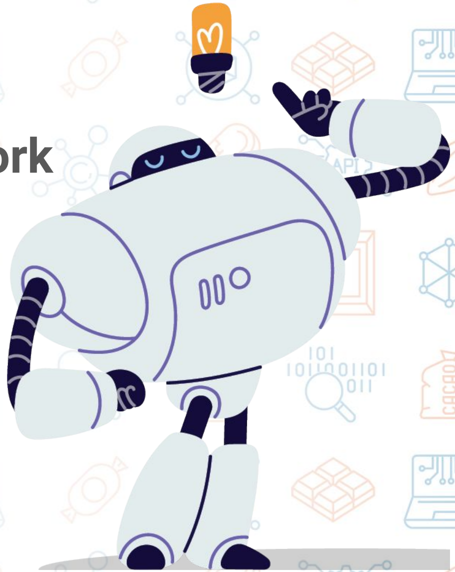
THE UNIVERSITY OF ARIZONA
CYVERSE
Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442
RESEARCH, INNOVATION & IMPACT
Data Science Institute

Thank you to many !

- Special thanks to CyVerse Data Engineering and Cloud Native team and Research Software Engineers
- Collaborators at Iowa State (AIIRA), NYU (AIIRA) , Indiana (Jetstream2) and TACC
- University of Arizona HPC/Research Computing and **Network Ops+Security**

For more information:

Nirav Merchant nirav@arizona.edu



THE UNIVERSITY OF ARIZONA CYVERSE® Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442 RESEARCH, INNOVATION & IMPACT Data Science Institute