From Promise to Practice: Reimagining Resilient Agriculture Through AI

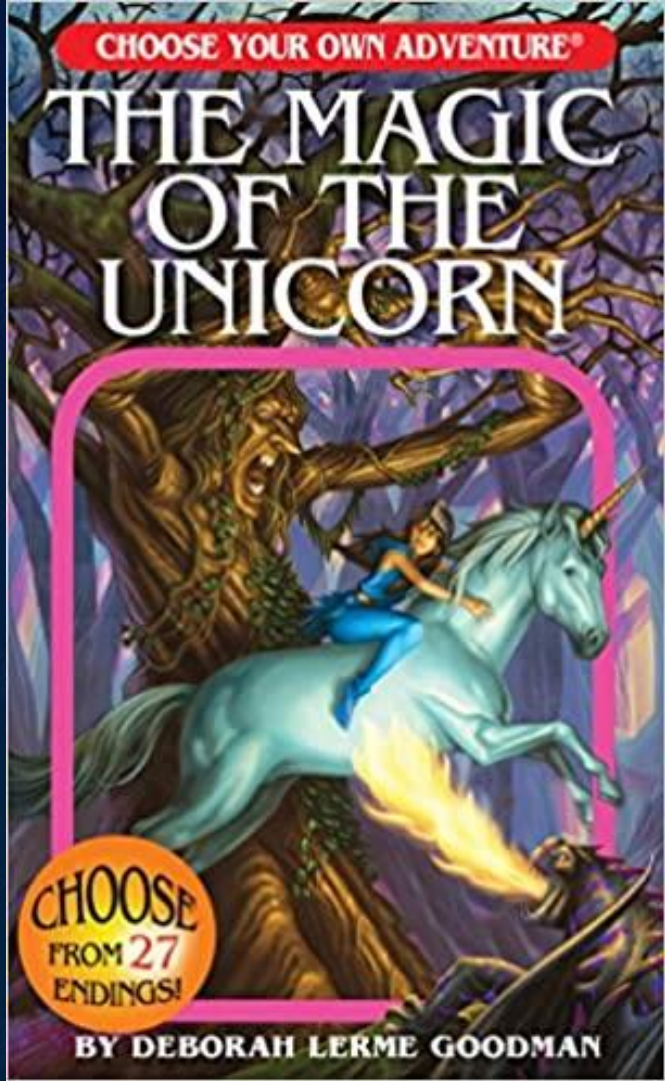
Nirav Merchant (nirav@arizona.edu)
University of Arizona, CyVerse, Jetstream2
AI Institute for Resilient Agriculture (AIIRA)
NSF USDA NIFA #2021-67021-35329

NSF ACCESS Regional AI Workshop
SoCal Edition
University of Southern California
22 Jan 2026

Topic Coverage

- Background for NSF/USDA AIIRA
- Why we need open foundation models
- NSF & NAIRR to our rescue
- From concepts to field deployments

- Deriving maximum value from your NAIRR allocations
- Building your own AI Toolbox: Useable AI Models, Agents & Tools
- Building AI Communities of Practice



Sir Arthur Clarke
(1917 –2008)

Clarke's 3rd law

Any sufficiently advanced technology is indistinguishable from magic

Augmented Intelligence (AI)

If we design machines to complement us, then we are free to emphasize features that are most useful

AIIRA Vision

- AI-driven tools – integrating **knowledge and data**
- Explore **foundational questions** in AI, through the lens of agricultural applications
- Provide tools, products and value to **agriculture** (growers, breeders, consumers)
- Working towards several **moonshots** that can transform agriculture

https://aiira.iastate.edu

AI Agents for Pest ID & Mitigation

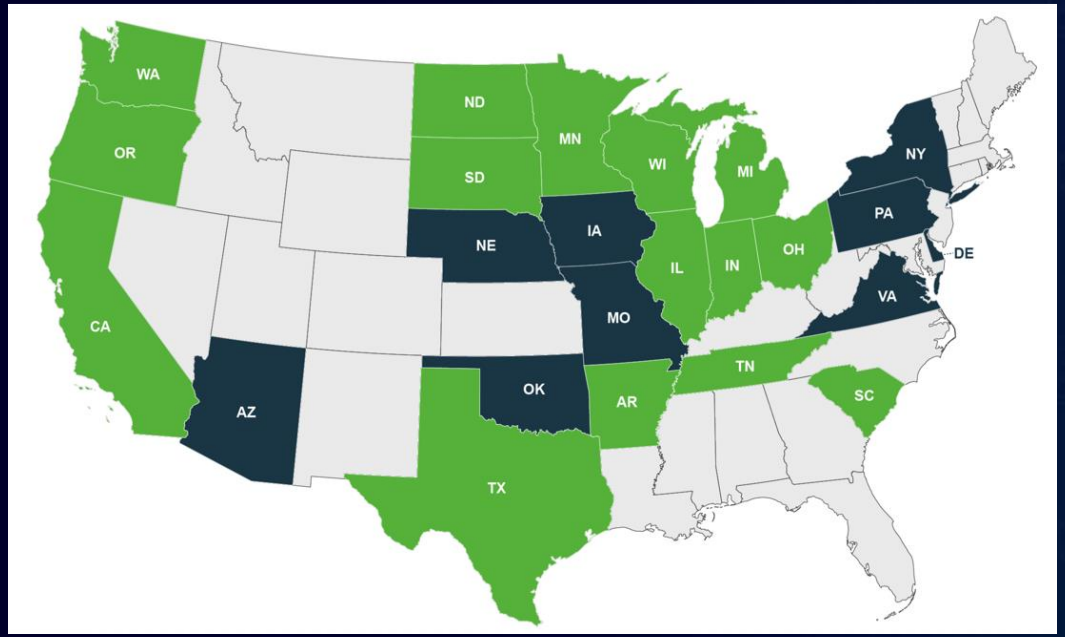
Multi-modal ag foundational models

Multi agent adaptive sensing for field ops

Ag data co-operatives

3D Geometric Twins

AIIRA: Core team & Stakeholders



Iowa State University

Carnegie Mellon University

The University of Arizona

Mizzou University of Missouri

George Mason University

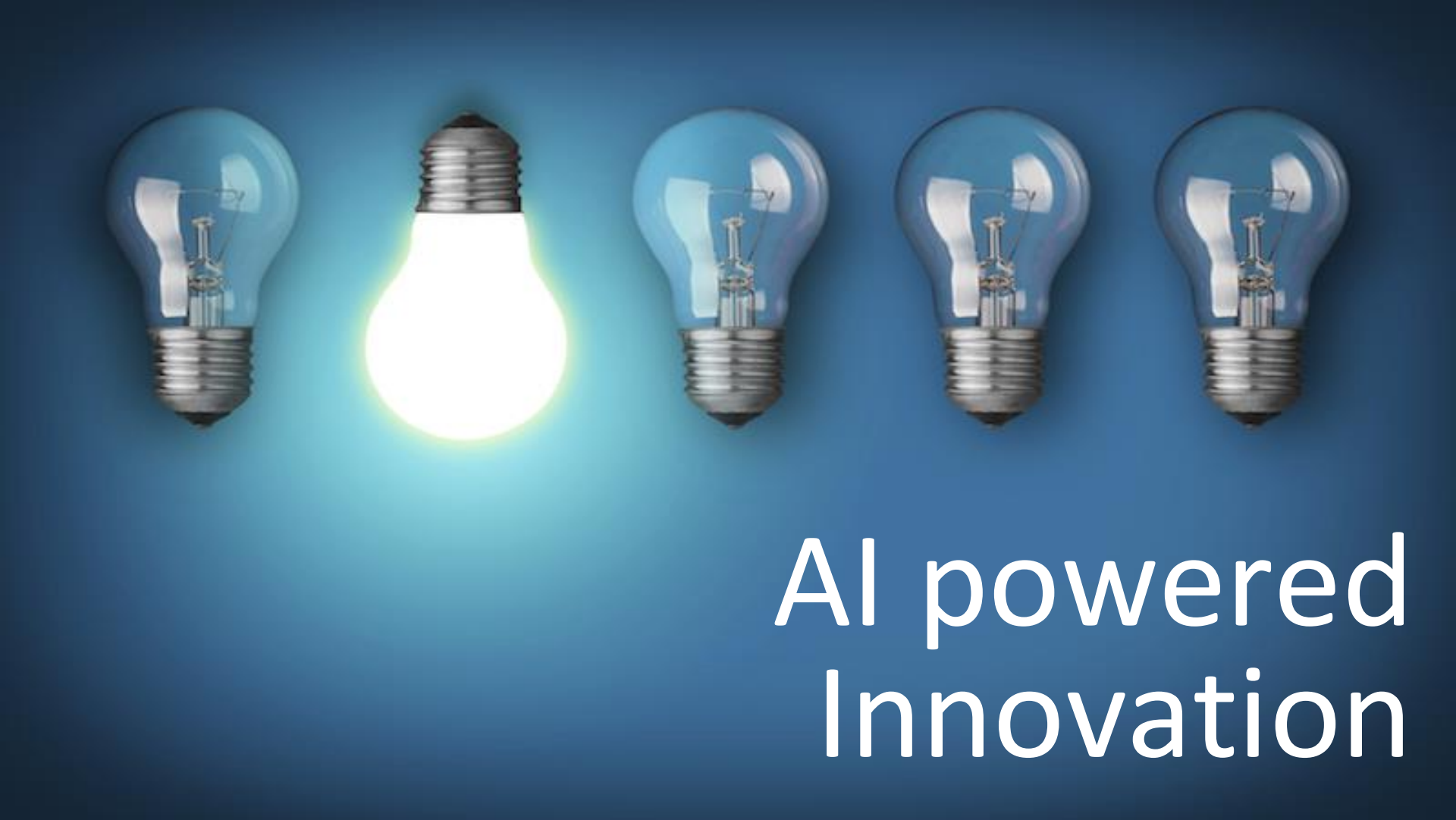
University of Nebraska Lincoln

NYU

Iowa Soybean Association

OSU

University of Delaware


AI powered
Innovation

Pests: Real Loss, Rising Risk !!

Identifying and quantifying pest pressure is important

Why

- In U.S. growers lose an estimated 10 to 35 % of their crops

Why now

- Expansion of geographic range, increased over-wintering survival
- Feed a growing population, food security is national security

Impact

- Farmers need targeted, early mitigation
- Reduce chemical usage
- Increase profitability
- Limits impact on beneficial species

Insects

Weeds

Diseases

Why is this a tough problem?

1. Variation within a species (e.g., Bean leaf beetle)

2. Different insects, exhibit similar color and pattern

Southern Corn Rootworm

Bean Leaf beetle

3. Multiple insect stages per class (e.g., Fall armyworm)

4. Insect detection with overlapping objects

Several more challenges:

- Large number of classes, fine grained classification
- Data imbalance
- Need for robust classification
- Diverse backgrounds, camouflaging, multiple life cycle stages, intra-class dissimilarity

Hypothesis:
A vision model (pre)trained using **large, unlabeled dataset** can serve as a **foundational model** for subsequent finetuning using **limited labeled data**

Sir Issac Newton
(1643-1727)

“If I have seen further it is by standing on the shoulders of Giants”

Citizen Science: Data that makes the difference

Citizen Science: How does it work ?

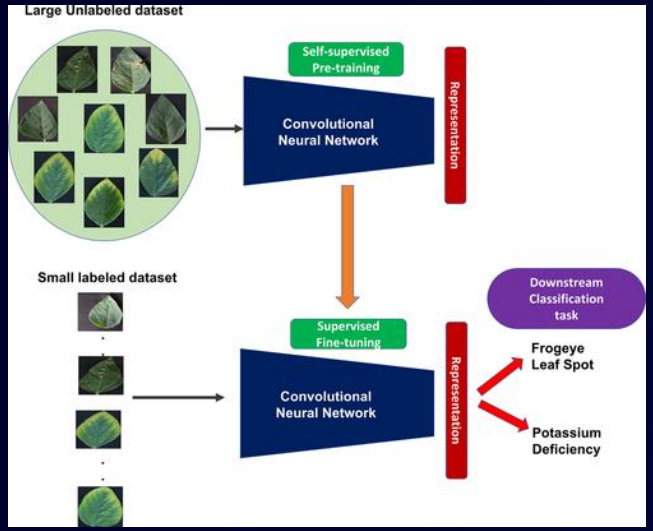
What began as a student project grew into one of the world's largest biodiversity platforms, connecting millions of people to their local nature while contributing critical data to science and conservation

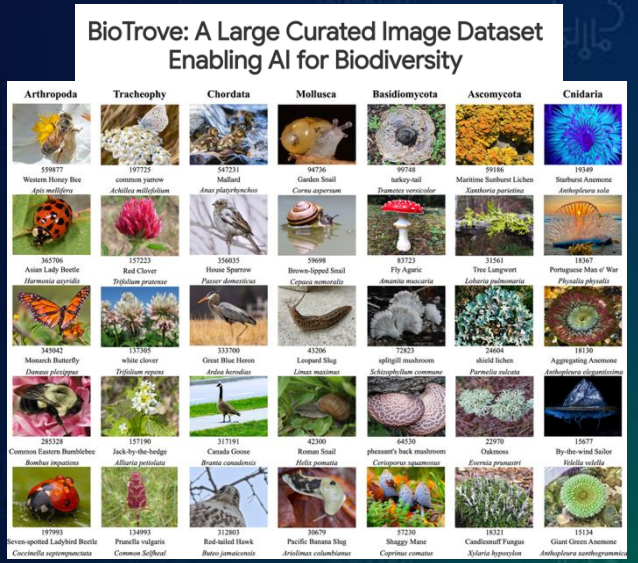

If Data is The New Oil
AI is The Ultimate Refinery

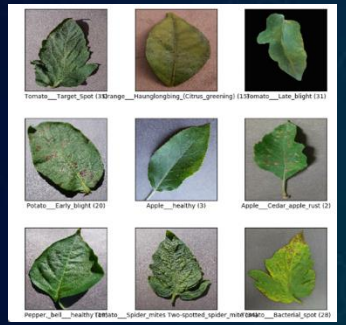
Foundation vision models for agriculture: Self supervised learning

Data collection, curation, and dissemination

- High throughput data acquisition from multiple sources (~100 TB)
- Citizen size data repository with 230+ million images (iNaturalist, PlantVillage, ISU Insect Disease Dataset, BugWood, Australian Weed)
- Total Image Dataset Size: ~42M







Van Horn, Grant, et al. "The inaturalist species classification and detection dataset." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

What changed for citizen science projects ?

Registry of Open Data on AWS

iNaturalist Licensed Observation Images

biodiversity | bioinformatics | conservation | earth observation | life sciences

Description

iNaturalist is a community science effort in which participants share observations of living organisms that they encounter and document with photographic evidence, location, and date. The community works together reviewing these images to identify these observations to species. This collection represents the licensed images accompanying iNaturalist observations.

Update Frequency

Images are posted in real time, and we are currently copying over images from existing observations. Metadata is updated monthly. More information on the metadata can be found in the documentation

License

Creative Commons or Public Domain (CC0), varying by image. More information on how to query and properly treat licenses can be found in the documentation

Documentation

Documentation can be found here. You can learn more about iNaturalist here.

Resources on AWS

Description
Image files (e.g. JPEG) associated with metadata describing the observation associated with the image file.

Resource type
S3 Bucket

Amazon Resource Name (ARN)
`arn:aws:s3:::inaturalist-open-data`

AWS Region
`us-east-1`

AWS CLI Access (No AWS account required)
```
aws s3 ls --no-sign-request s3://inaturalist-open-data/
```

Dataset curation and training of large vision models

Without NSF NAIRR and NVIDIA support Biotrove would not be possible

- **Biotrove**: AI-ready dataset, 136 million images, and counting!
- Language (Latin, English) image pairs.
- NeurIPS 2024 Spotlight! Extending to image descriptions for multimodal training
- Data acquisition from multiple sources (~60 TB)

Compute Resources:

- 8 Nodes (x8)= 64 A100 80GB GPUS on NVIDIA DGX Cloud
- Running for 4 weeks for ViT-B model ~ 50,000 GPU hours!

Democratizing usage: Easy to use web app

Spotted Lantern Fly Nymph

Spotted Lantern Fly Adult

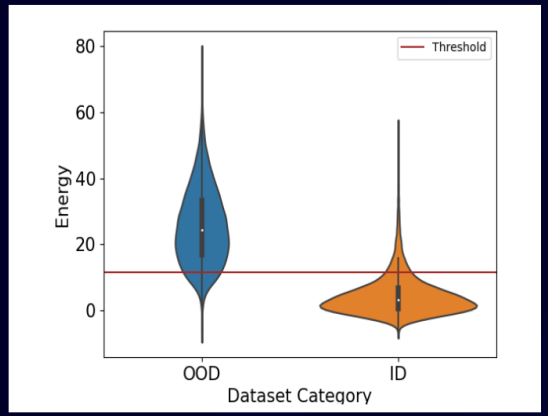
Enhancing trustworthiness

Knowing when to say "**I don't know**" important for robust deployment

Uncertainty quantification: Integrated multiple approaches that serve as filters (OOD, conformal, ensemble)

Out of distribution (OOD) detection:

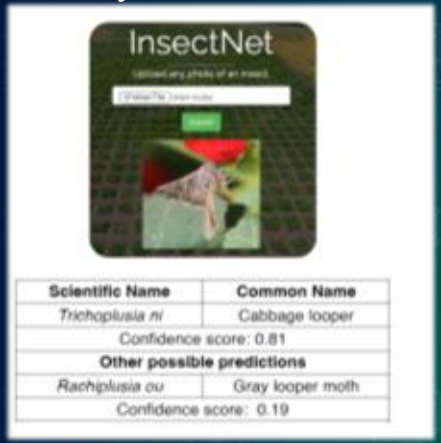
- Checks to see if image is within its distribution, and abstains from predictions
- Use notion of energy of latent representation



Saadati, Mojdeh, et al. "Out-of-Distribution Detection Algorithms for Robust Insect Classification." *Plant Phenomics* 6 (2024): 0170.

Conformal predictions:

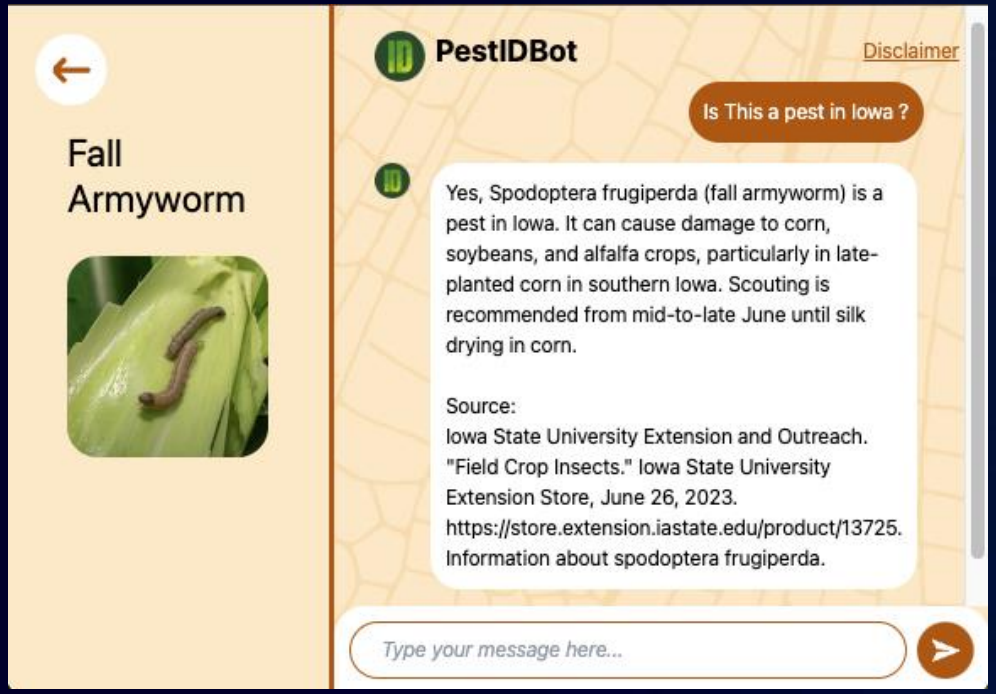
Instead of making a single class prediction, make predictions of conformal set of classes with probability that true class will belong to this set



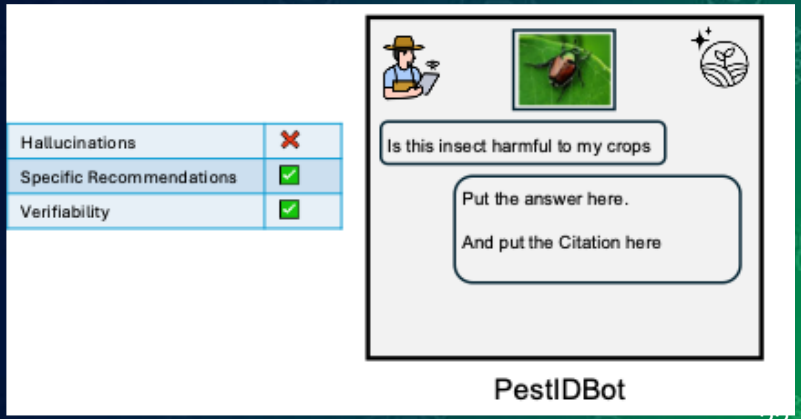
Angelopoulos, Anastasios N., and Stephen Bates. "A gentle introduction to conformal prediction and distribution-free uncertainty quantification." *arXiv preprint arXiv:2107.07511* (2021).

From models to acres: apps, robots and drones

Integrating with LLMs for decision support



- AgLLM: Retrieval Augmented Generation for information retrieval.
- Supports multiple language models (GPT-4, Claude 3, Llama-3).
- Contextually relevant and grounded responses to agricultural queries.
- Interesting opportunities for reasoning



Navigating and Scaling with ACCESS, NAIRR, Campus HPC

FLOW-DC: Distributed Data Collection at Scale

Flexible Large-scale Orchestrated Workflow for Data Collection

THE CHALLENGE

Many modern ML models are capable of processing massive datasets (TB-scale). Research groups may lack the infrastructure and bandwidth of enterprise data centers. Simultaneously, aggressive download attempts can rate throttling, vast errors, and IP bans.

OUR SOLUTION

FLOW-DC is a distributed downloader that combines adaptive request control with flexible orchestration. It enables research groups to acquire web-scale datasets while respecting server limits.

Two Core Components:

TaskVine Orchestration

Manager-worker paradigm for distributed task scheduling. Workers can join/leave dynamically from any platform.

PAARC Rate Controller

Policy-Aware Adaptive Request Controller. BBR-inspired algorithm that adjusts concurrency based on observed latency and HTTP responses.

https://ccl.cse.nd.edu/softwares/taskvine/

Applications and Results

Biotrove-Train Dataset

38.7M images | **14 TB** total size | **~3 hrs** total time

- We developed Biotrove as a benchmark biodiversity benchmark dataset sourced from iNaturalist.
- It was downloaded using 15 Jetstream m3.medium workers. Biotrove is used to train Biotrove-CLIP (91.1% accuracy).
- The dataset was partitioned into ~1400 10gb parquet files. Each file is a TaskVine task and worked on by a Jetstreams worker.

Other ML Models Trained with FLOW-DC Data

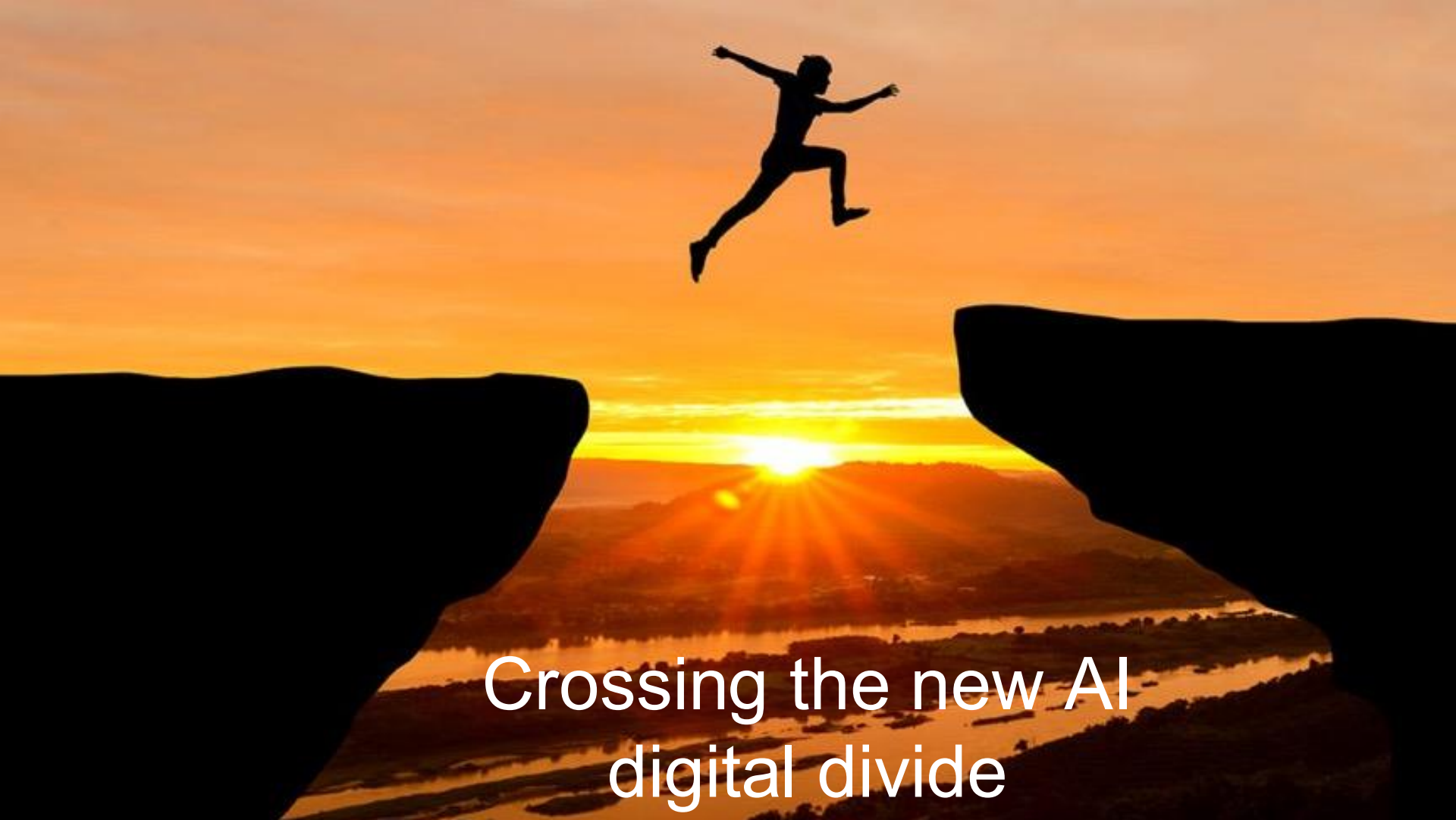
WeedNet 14M images • 1,593 species	**91.0%**
InsectNet 6M images • 2,526 classes	**96.4%**
AraNet 549K images • 1,000 classes	**96.2%**

BioTrove Single Task Timing Breakdown (100000 images, 10gb)

~60s	~27s	~12s
Async. Download	Tar archive	Upload

Jetstream2 (m3.medium instances): up to 128 cores, 30 GB Ram, 100 Gbps
UofA HPC (Puma): Up to 94 cores, Up to 470 GB RAM, 25 Gbps
Local testing: Apple M1 Pro (10-core), 32 GB RAM, Local ISP

Tested data sources: iNaturalist Open Data, GBIF, Hugging Face datasets


Crossing the new AI
digital divide

What should be in your AI workbench ?

- Interactive environment to explore data and code (Jupyter, Collab notebooks)
- A coding environment (VScode, Posit)
- Access to datasets (Hugging Face, your own)
- Data Management tools and cloud native formats (DVC, Croissant, Duckdb, Parquet)
- Compute resources that can scale up (laptop → campus cluster → NSF cloud)
- Experiment tracking (Trackio, MLFlow etc.)
- Access to Models for inference (Frontier & Custom with API and web interfaces)
- Automation and rapid application deployment frameworks (Streamlit, Gradio)
- Be prepare to opportunistically use resources from ANY infrastructure provider (quickly)

This scaffolding for workbench and tools facilitates learning responsibly and understanding what tools to use when

Few additional tools you need access to

- **Desktop Client:** Claude Desktop (many others)
- **Large Language Model:** API keys with generous token limits
- **Data:** Ability for agents to access it programmatically (MCP: Model Context Protocol)
- **Tools:** Programs that can be launched by agents programmatically (API) to analyze data
- **Secure Sandbox:** Disposable computing environment with read only access* (prototyping)
- **Automation:** **P**roduce reproducible and scalable workflows from prototyping
- **MOST IMPORTANT**: Someone that can assess the security risk/exposure for you (see Sandbox)

CACAO: Cloud Automation & Continous Analysis Orchestration

CACAO is an **open-source, multi-cloud software platform** that enables **researchers, educators, research software professionals (RSPs)** to create, use, and share their **data-driven workflows, infrastructure, and software stacks**, while allowing them to **minimize costs by leveraging cloud frameworks.**

Researcher
Educator
RSP*

cacao
Cloud Automation & Continuous Analysis Orchestration

Templates
e.g.
Terraform, Ansible, etc

Jetstream2

CSP

Google Cloud

Azure

aws

Edge Clouds

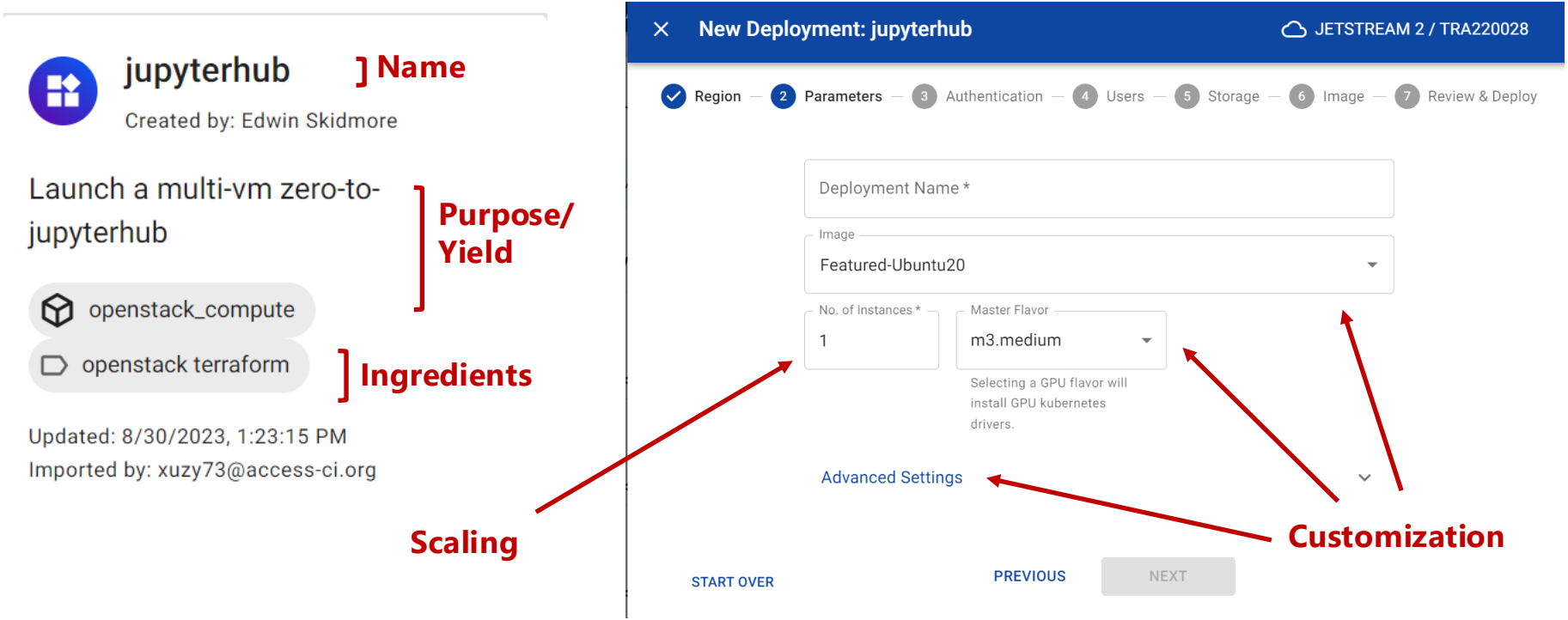
** RSP is a Research Software Professional, including RSEs, RS Devops, etc.*

Jetstream2

THE UNIVERSITY OF ARIZONA

CYVERSE

NSF



Jetstream2

THE UNIVERSITY OF ARIZONA

CYVERSE

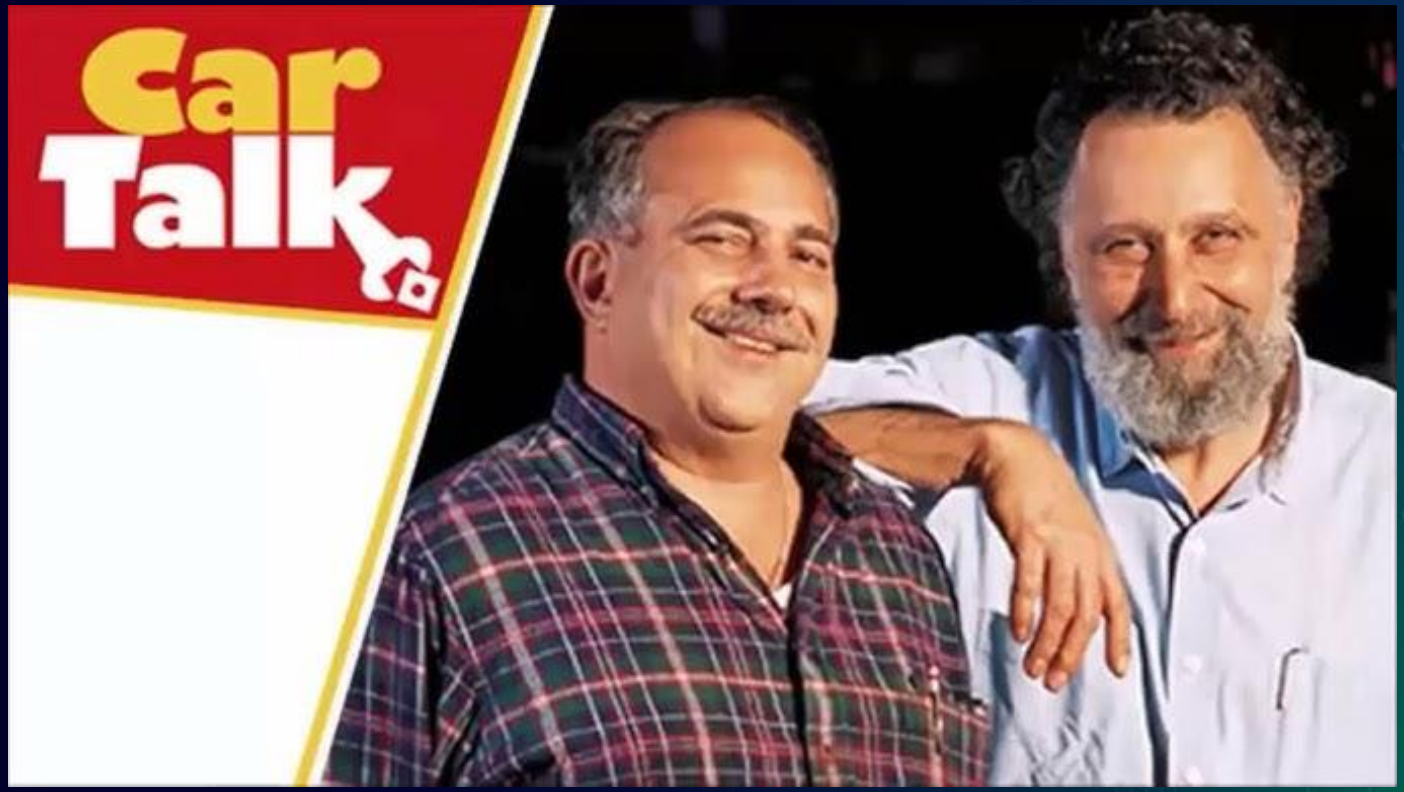
NSF

CyVerse-curated Templates

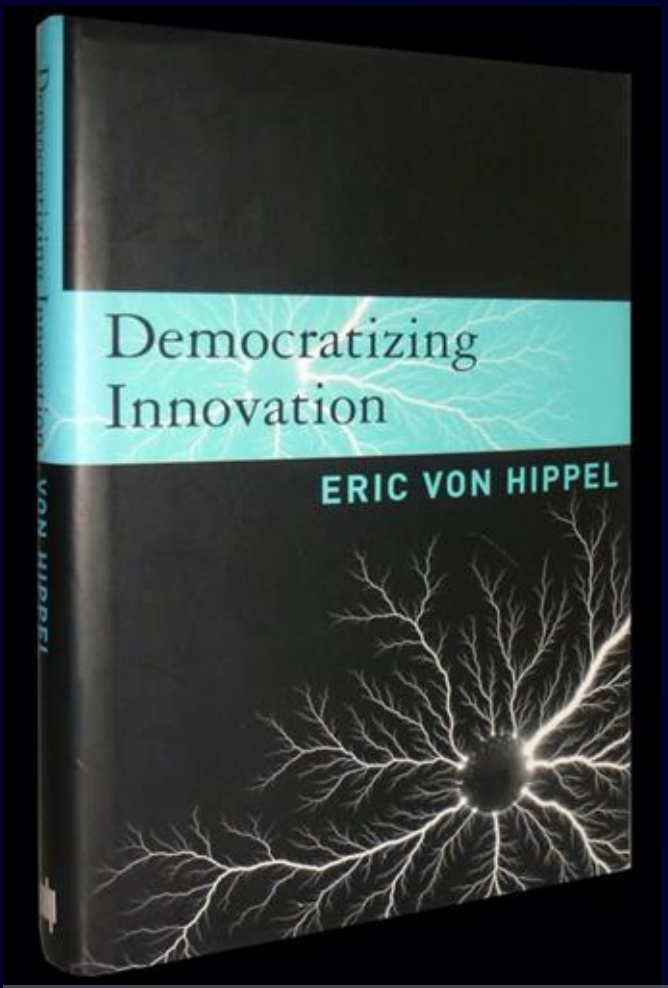
Name	Purpose	Git repo
colab runtime	Deploy Google Colab runtime connected any gpu instance	https://gitlab.com/stack0/jippy
kubeflow	Deploy a Kubeflow on kubernetes	https://gitlab.com/cyverse/cacao-tf-kubeflow.git
vms4workshop	Deploy a set of instances with instructor and separate student instances, with a desktop and ssh	https://gitlab.com/cyverse/cacao-tf-os-ops/-/tree/main/vms4workshop
ray.io clusters	Deploy distributed ray cluster	https://gitlab.com/stack0/cacao-tf-ray
jupyterhub	Deploy a Zero to Jupyterhub with multiple nodes configurable GPUs, Shared Storage, and other configurable options; Dask Gateway as an option	https://gitlab.com/stack0/cacao-tf-jupyterhub
magic castle	Deploy a virtual HPC/Slurm cluster	https://github.com/edwins/magic_castle
text-generation-webui	Gradio webui for LLMs	https://github.com/edwins/text-generation-webui/tree/main/terraform
bag of words	Bag of Words Agentic Analytics Platform	https://gitlab.com/stack0/cacao-tf-bow

Community of Practice: Shared Extensible Infrastructure Just in Time Learning

We need Car Talk for AI tinkerers



Car Talk: Tom and Ray Magliozzi on NPR (1977-2012)



Democratizing Innovation

Innovating users often freely share their innovations with others, creating user-innovation communities and a rich intellectual commons.

Cyberinfrastructure allows these freely shared innovations to be readily used by multiple communities for solving real world problems.

Please support your colleagues and projects that provide open sources resources and infrastructure

Von Hippel, Eric. Democratizing innovation. MIT press, 2005


If you want to go fast, go alone.
If you want to go far, go together.
-African Proverb