

# From Citizen-Science to Your Phone: Using iNaturalist's 70M+ Images to Build an ML Insect Identification Mobile App

**Zi Deng**

PhD Student  
Electrical Engineering  
University of Arizona  
Data Science Institute

**Shivani Chiranjeevi**

PhD Student  
Mechanical Engineering  
Iowa State University





Start Using the Discovery Environment Now!

[de.cyverse.org](https://de.cyverse.org)



### DATA & ANALYSIS

- 5 GB free data storage
- 100s open-source scientific apps
- Containers and notebooks
- Visualize & interact with data

### YOUR WORKSPACE

- Manage and share data
- Perform analysis with your own or community datasets
- Do large-scale science from your web browser
- Build and customize apps

### COLLABORATION

- Secure, shared workspace for your team
- Reproducibility
- Manage the data lifecycle
- Make data more FAIR
- Open science

### LEARN, TEACH & TRAIN

- Tutorials and documentation
- Webinars
- Workshops
- Teach using CyVerse

### COMMUNITY

- Join 95K+ users
- In-app chat support
- Find publicly available data
- Share data and analyses
- Deploy your own CyVerse



**Sign Up**  
[user.cyverse.org](https://user.cyverse.org)

**Learn More**  
[www.cyverse.org](https://www.cyverse.org)



**CYVERSE®**

The Open Science Workspace for Collaborative, Data-driven Discovery



National Institute of Food and Agriculture  
U.S. DEPARTMENT OF AGRICULTURE



Award number: 2021-67021-35329



# AIIRA: AI Institute for Resilient Agriculture

[aiira.iastate.edu](http://aiira.iastate.edu)

Director: Baskar Ganapathysubramanian, Iowa State



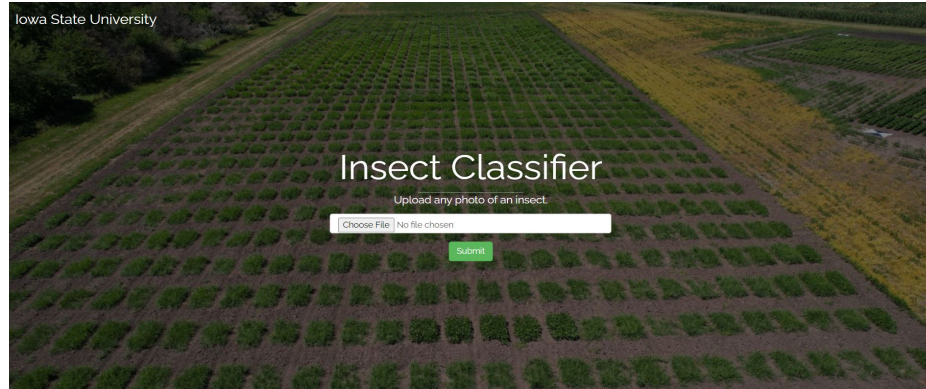
# AIIRA Vision



- AI-driven digital twin – virtual representation integrating knowledge & data
- Explore foundational questions in AI
- USDA Science Blueprint drives our applications
- Social science focus to ensure adoption and long-term payoffs
- Creating a diverse, AI-aware workforce
- Building a resilient US agricultural system



# Mobile Insect Identification App



Mobile web app to identify 142 agriculturally important insect-pest species (extended to 2526 species)

## Key Features:

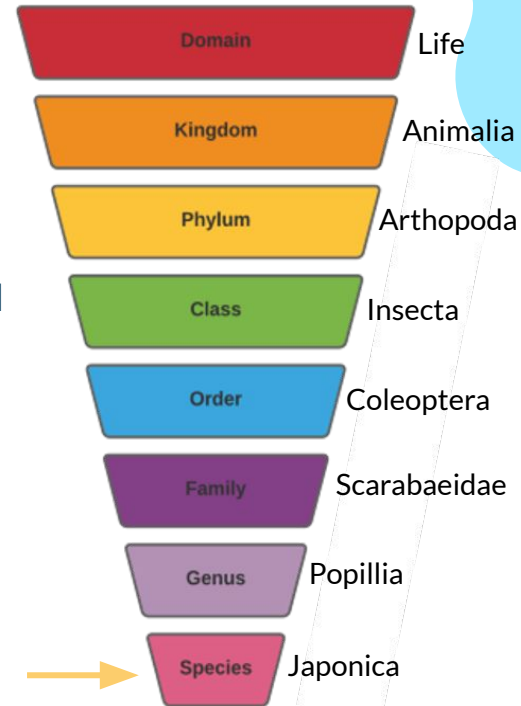
- Differentiate beneficial and harmful insects
- Ability to detect insect-pest species at different stages
- Ability to differentiate fine-grained species
- Provides scientific to common name mapping

# Dataset Organization

- **Taxonomy** : Area of biology that focuses on classifying and naming organisms
- 7 levels of taxa in the hierarchy
- Studies evolutionary relationships between organisms
- Species-level classification requires images at last level in the tree
- Nesting of folders in the taxonomic order up to species-level



Japanese Beetle



# 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



## 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. They are able to acquire high quality labeled images of organisms for research or any other purpose.

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





# iNaturalist

iNaturalist is a citizen science platform that aims to share observations of biodiversity across the world.

It results in one of the largest public datasets of labeled images of organisms in the world

Using the resource for research unfeasible prior to 2021

iNaturalist

Explore

Community

More

Log In or Sign Up

Observations

Species

Location

Go

Filters

The World

120,867,672  
OBSERVATIONS

402,288  
SPECIES

279,843  
IDENTIFIERS

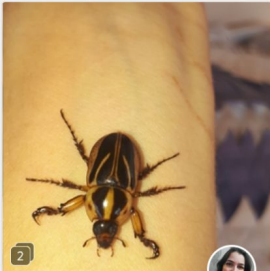
2,465,721  
OBSERVERS

Map

Grid

List

2

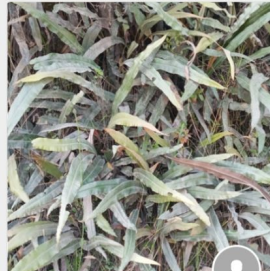


Unknown

(Kingdom Plantae)

11h

2



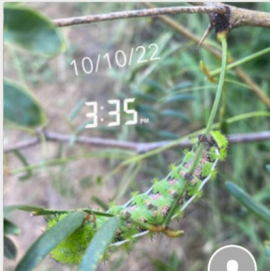
Plants

(Kingdom Plantae)

12h

10/10/22

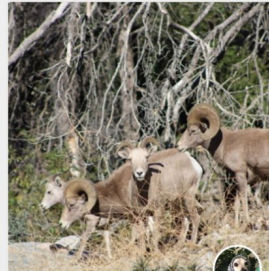
3:35 PM



Emperor, Royal, Moon, ...

(Family Saturniidae)


1mo



Bighorn Sheep

(Ovis canadensis)

8h

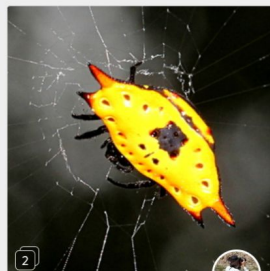


Bighorn Sheep

(Ovis canadensis)

8h

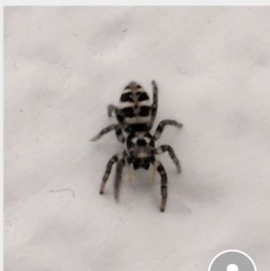
2



Four-spined Spiny Orbw...

(Gasteracantha quadrispinosa)

8y

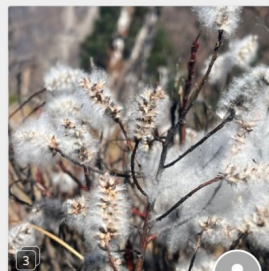


Zebra Jumping Spider

(Salticus scenicus)

1mo

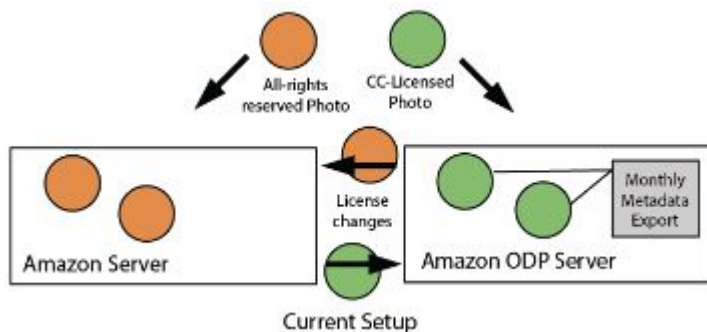
3



Unknown

18d

# iNaturalist Open Dataset



- <https://github.com/inaturalist/inaturalist-open-data>
- Observation Access:  
[http://inaturalist-open-data.s3.amazonaws.com/photos/\[photo\\_id\]/\[size\].jpg](http://inaturalist-open-data.s3.amazonaws.com/photos/[photo_id]/[size].jpg)

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

## Metadata Columns

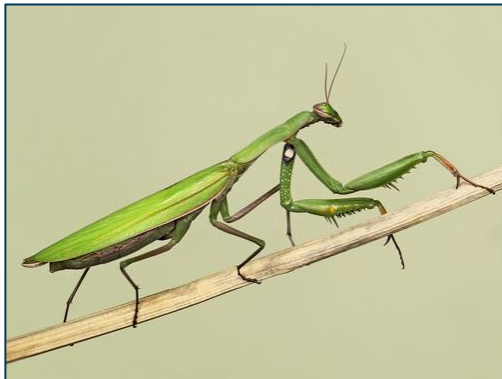
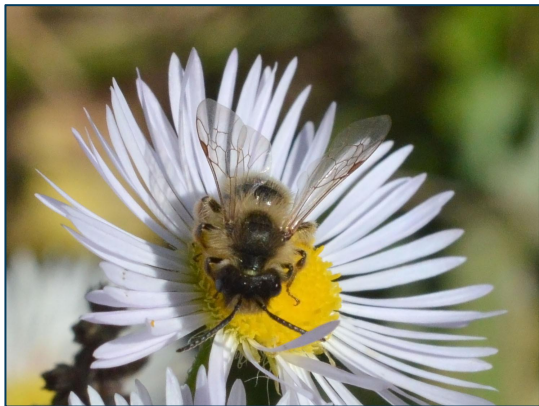
- Observations
  - observation\_uuid
  - observer\_id
  - latitude
  - longitude
  - positional\_accuracy
  - taxon\_id**
  - quality\_grade**
  - observed\_on
- Observers
  - observer\_id
  - login
  - name
- Photos
  - photo\_uuid
  - photo\_id**
  - observation\_uuid
  - observer\_id
  - extension**
  - license**
  - width
  - height
  - position
- Taxa
  - taxon\_id**
  - ancestry**
  - rank\_level
  - rank**
  - name**
  - active



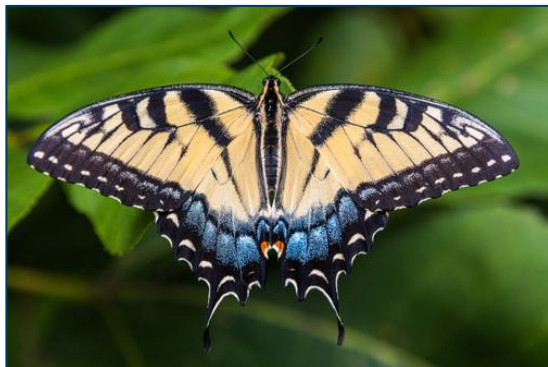
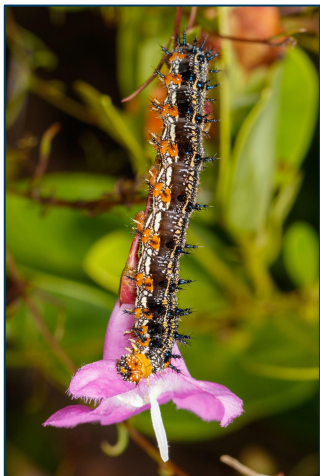
# iNaturalist Scalable Download Demo



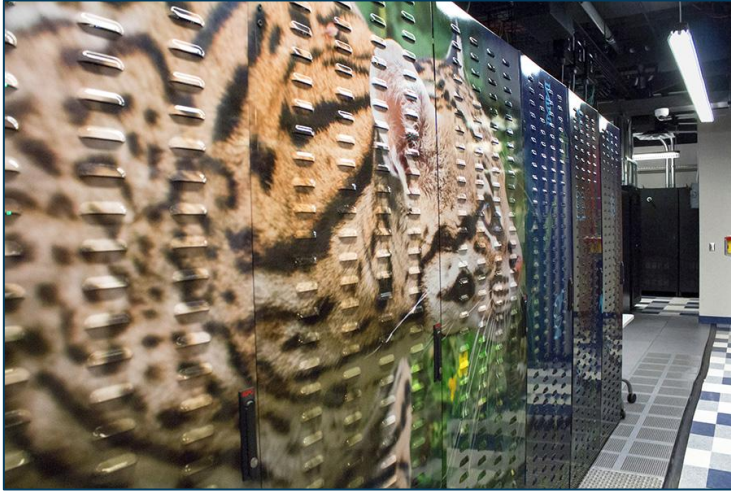
# Insecta Dataset



- The insecta dataset totals around ~14 million images over ~95,000 species, totaling ~5.7 terabytes.
- Acquisition of this dataset would not have been feasible without iNatSD.



# Scaling and Parallelization



- Each species download considered a separate job
- Parallelization through Snakemake

- Scalable Design
- AWS Open Data Sponsorship covers all costs
- Limited only by self architecture

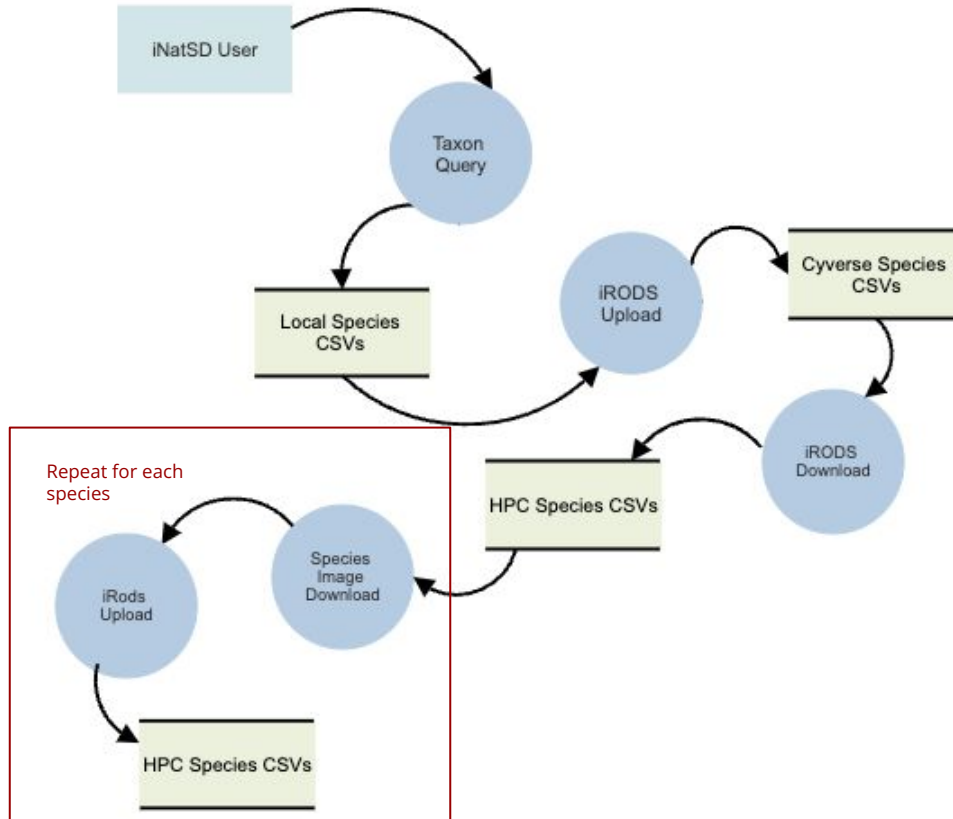
```
Job stats:
job
-----
all_species_download    157      1      1
download_imgs_all       1        1      1
total                   158      1      1

Reasons:
(check individual jobs above for details)
input files updated by another job:
  download_imgs_all
missing output files:
  all_species_download, download_imgs_all

This was a dry-run (flag -n). The order of jobs does not reflect the order of execution.
```



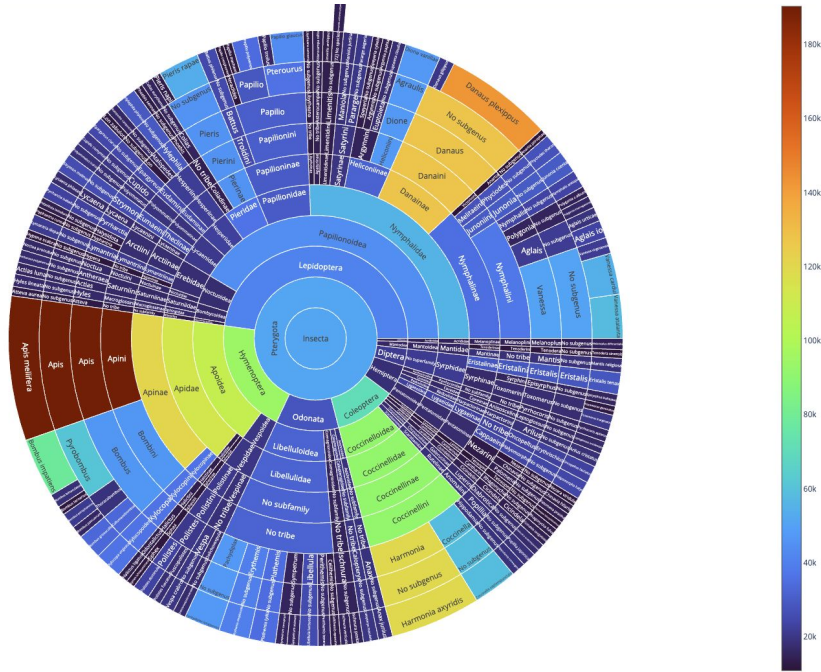
# Dataflow













- For large datasets:
  - Utilize Cyverse for data storage
  - Utilize HPC or other computing resources for computation

# Additional Features

## iNaturalist Insect Top 100 Species Sunburst



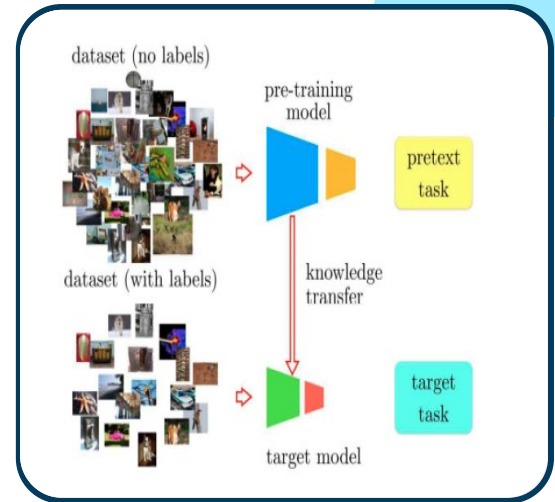
A sunburst plot is a data visualization technique that displays hierarchical data in a radial layout.

	<b>Grasshopper Sparrow</b> ( <i>Ammodramus saviannarum</i> )		Jun 14, 2022 6:49 AM EST	📍 Montgomery County, PA, USA	Today 8:00 AM EST
	<b>Bobolink</b> ( <i>Dolichonyx oryzivorus</i> )		Jun 14, 2022 6:48 AM EST	📍 Montgomery County, PA, USA	Today 8:00 AM EST
	<b>Oleander Aphid</b> ( <i>Aphis nerii</i> )		Today 2:00 PM CET	📍 Calle del Maestro Miguel Fernández, Archena, Murcia, ES	Today 2:00 PM CET
	<b>Green Dragonail Butterfly</b> ( <i>Lamproptera meges</i> )		Oct 28, 2022 2:49 PM ICT	📍 Chiang Dao, Chiang Dao District, Chiang Mai 50170, Thailand	Today 8:00 PM ICT
	<b>Mitnan</b> ( <i>Thymelaea hirsuta</i> )		Nov 13, 2022	📍 NW of Benalmádena, Calamorro, ~500m, Andalucía, Spain	Today 2:00 PM CET

iNaturalist is constantly updated. iNatSD includes update feature to repeat downloads.

<https://chart-studio.plotly.com/~zkdeng/13/#/>

# App Introduction





# Challenges

Large Number of  
Classes



Fine-grained Classes



Classes with  
Diverse Samples



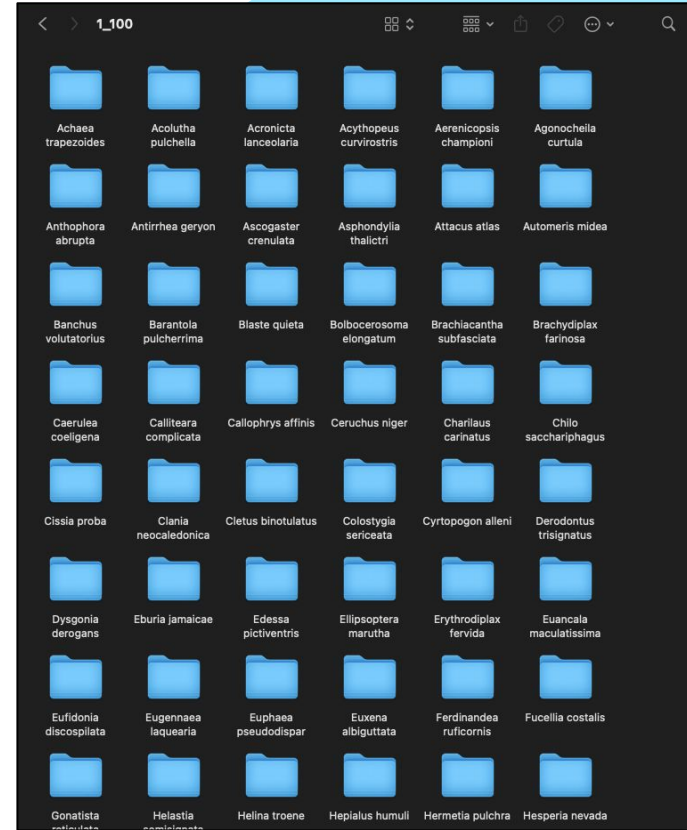
# What is it?



# Species-level Extraction for Classification

## How does iNatSD help:

- Enables downloading of images in a single folder at specified taxonomic rank
  - Eg: Querying for genus “Popillia” would give all images within that genus (groups all species into genus-level), if user is interested in genus-level organization of dataset
- In our application, for species-level classification, folders are organized at species-level
- Varied resolution of images could be downloaded
- Dataset organization compatible with classifier dataset module



# Dataset Preparation for Classification

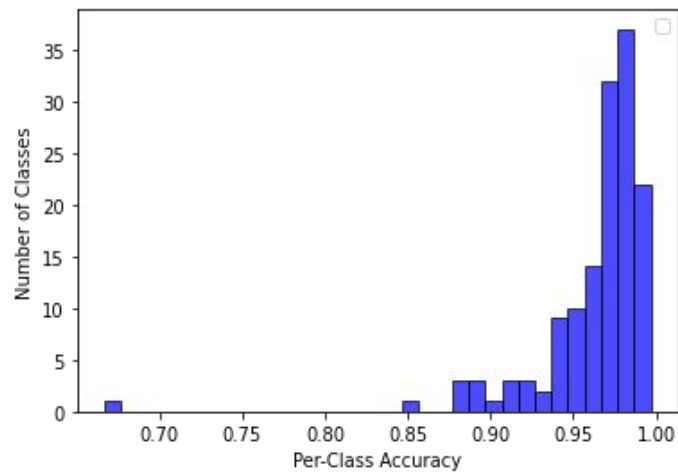
- Dataset Size : 2 Million images
- Number of classes: 142 Agriculturally Important Insect Species (Both beneficial and harmful species)
- Input Resolution: 224 x 224
  - Higher resolution demands more computational resources
  - Lower resolution could potentially affect classifier's accuracy
  - Compute power vs Accuracy trade-off : 224 x 224 is a reasonable resolution and is compatible with popular SOTA architectures like ResNet-50

# Classifier Pre-Training Details

- Two stages in classifier training:
  - Self-Supervised Pre-training
  - Classifier Finetuning
- Self-Supervised Pre-training
  - **Method:** SwAV (Swapping Assignments between multiple Views of the same image)
  - **Dataset Size:** 2 Million Unlabeled images
  - **Computational Resources:** 4 nodes of A100x4 = 16 GPUs
  - **Computation Time:** 6 days
  - **Dataset Size:** 6 Million Unlabeled images
  - **Computational Resources:** 4 nodes of A100x4 = 16 GPUs
  - **Computation Time:** ~30 days

# Best Classifier Training Details of Insect-App

- Classifier Finetuning
  - **Computational Resources:** 1 x A100 GPU
  - **Computation Time:** 3.5 days
  - **Dataset Size:** 2 Million Images
  - **Number of classes:** 142
- Classifier Performance
  - **Top-1 Accuracy:** 97.766
  - **Top-5 Accuracy:** 99.577
  - **Mean-per-class Accuracy:** 96.334



# Comparing with iNaturalist's Insect Classifier

- Test Dataset: iNaturalist Insect Dataset with 2526 classes

Super-Class	Avg Train	Public Test	
		Top1	Top5
Plantae	75.4	69.5	87.1
Insecta	98.4	77.1	93.4
Aves	222.3	67.3	88.0
Reptilia	121.8	45.9	80.9
Mammalia	157.7	61.4	85.1
Fungi	48.1	74.0	92.3
Amphibia	67.9	51.2	81.0
Mollusca	81.0	72.4	90.9
Animalia	67.9	73.8	91.1
Arachnida	87.0	71.5	88.8
Actinopterygii	37.4	70.8	86.3
Chromista	44.2	73.8	92.4
Protozoa	77.0	89.2	96.0

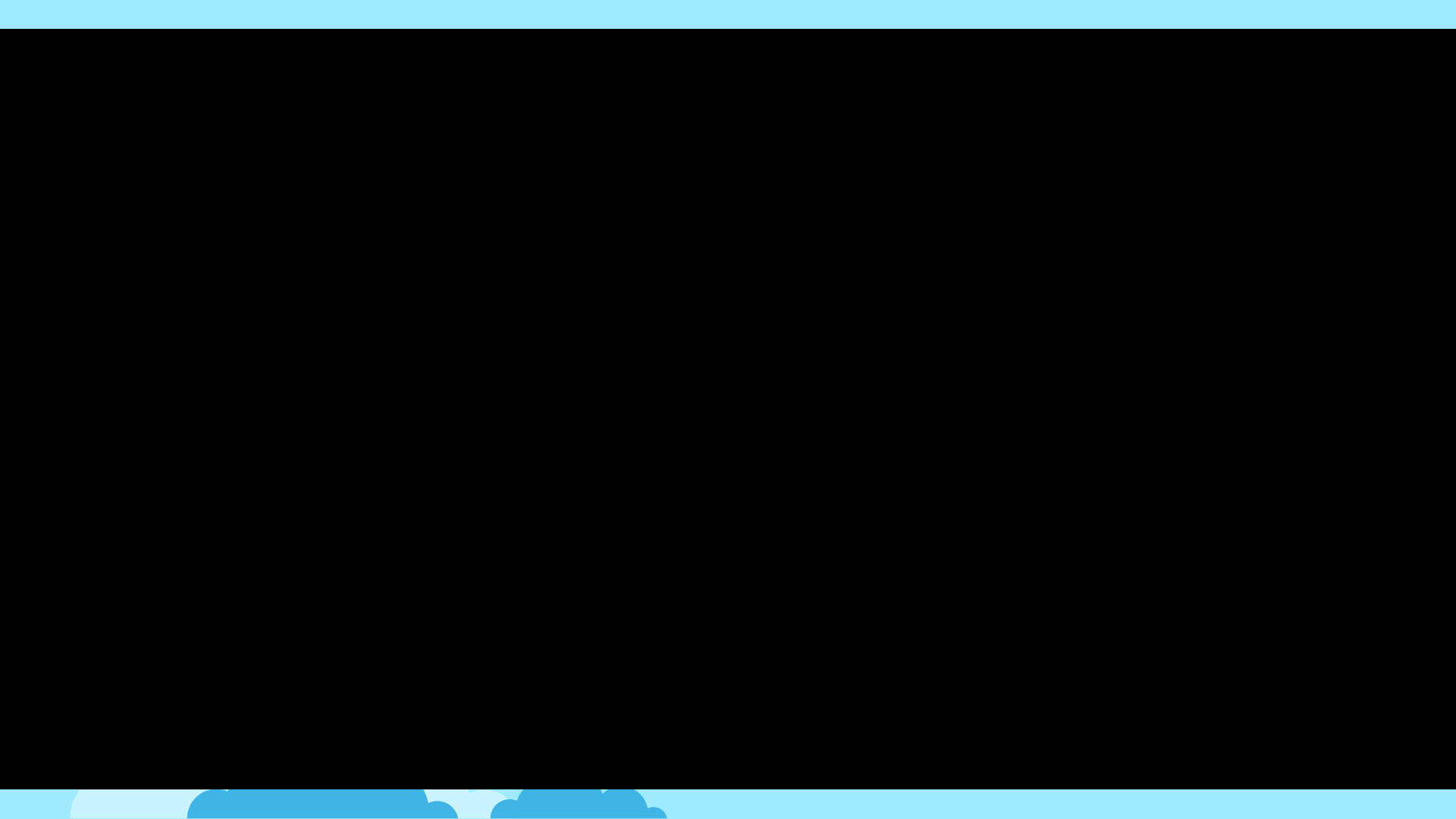
[1] <https://arxiv.org/pdf/1707.06642.pdf>

iNaturalist's Classifier		Our Classifier	
Test Accuracy	Top-1: 77.1	Test Accuracy	Top-1: 92.6
	Top-5: 93.4		Top-5: 98.3

# Insect Mobile Web App Demo







# Future Work

- Dedicated preprocessing options to more easily be plugged into popular architectures
- Integration of locations data representation through heat map visualizations for observations
- Investigation on automatic bounding box labeling of output datasets
- Release insect identification mobile web app for 2526 classes early next year
- Incorporate IPM (Integrated Pest Management) strategies for each species, to be displayed along with classification results on the app

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik** and illustrations by **Stories**

# Thanks!



Do you have any questions?

[arti@iastate.edu](mailto:arti@iastate.edu)

[baskarg@iastate.edu](mailto:baskarg@iastate.edu)

[nirav@arizona.edu](mailto:nirav@arizona.edu)

