Biodiversity & Big Data: Potential for AI

Annika Smith

iDigBio
Florida Museum of Natural History
University of Florida

Biodiversity & Big Data: Potential for AI

- What are natural history collections?
- What is iDigBio?
- What kinds of biodiversity data are available?
- What kinds of questions do researchers ask with specimen data?
- How can specimen data be linked with other types of data?
- How has AI been used with specimen data, and what are potential future uses?
- What are current challenges to AI and specimen data?

Museum Collections: The Library of Life

~1,600 natural history collections in the US

1-2 billion specimens in the US
3-4 billion specimens worldwide

Systematics & Taxonomy

Linnea (twinflower)

Carl Linné, aka Carolus Linnaeus

Museum Collections: The Library of Life

Genetics
Genomics
Chemistry…

Species interactions
Phenology
Biogeography
More!

Museum Collections: The Library of Life

Most specimens locked away in cabinets, unavailable for general use.

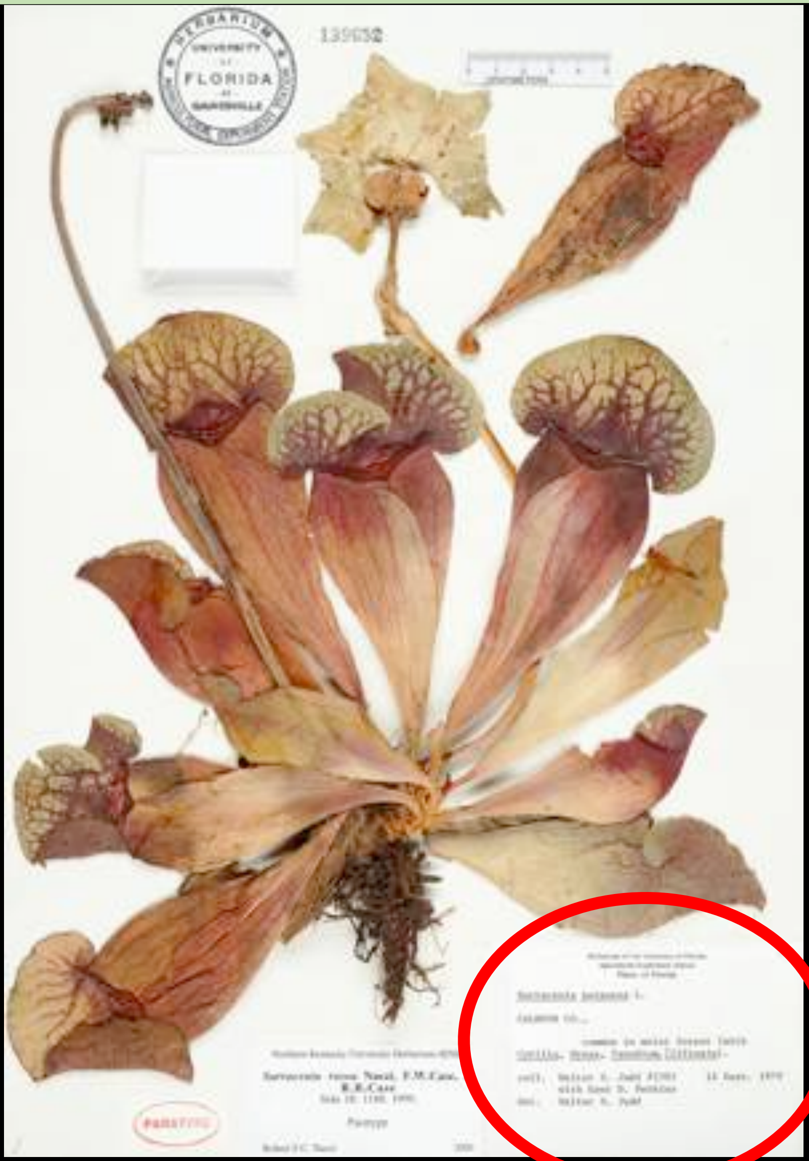
Museum Collections: The Library of Life

Most specimens locked away in cabinets, unavailable for general use.

DIGITIZATION!!!!

Label Data from Herbarium Specimens

- Scientific name – including authority
- Date
- Collector
- Location – state, county, specific site, GPS coordinates
- Associated species
- Notes



iDigBio: www.idigbio.org

iDigBio
Integrated Digitized Biocollections

National Coordinating Center
For Digitization of Biodiversity Collections

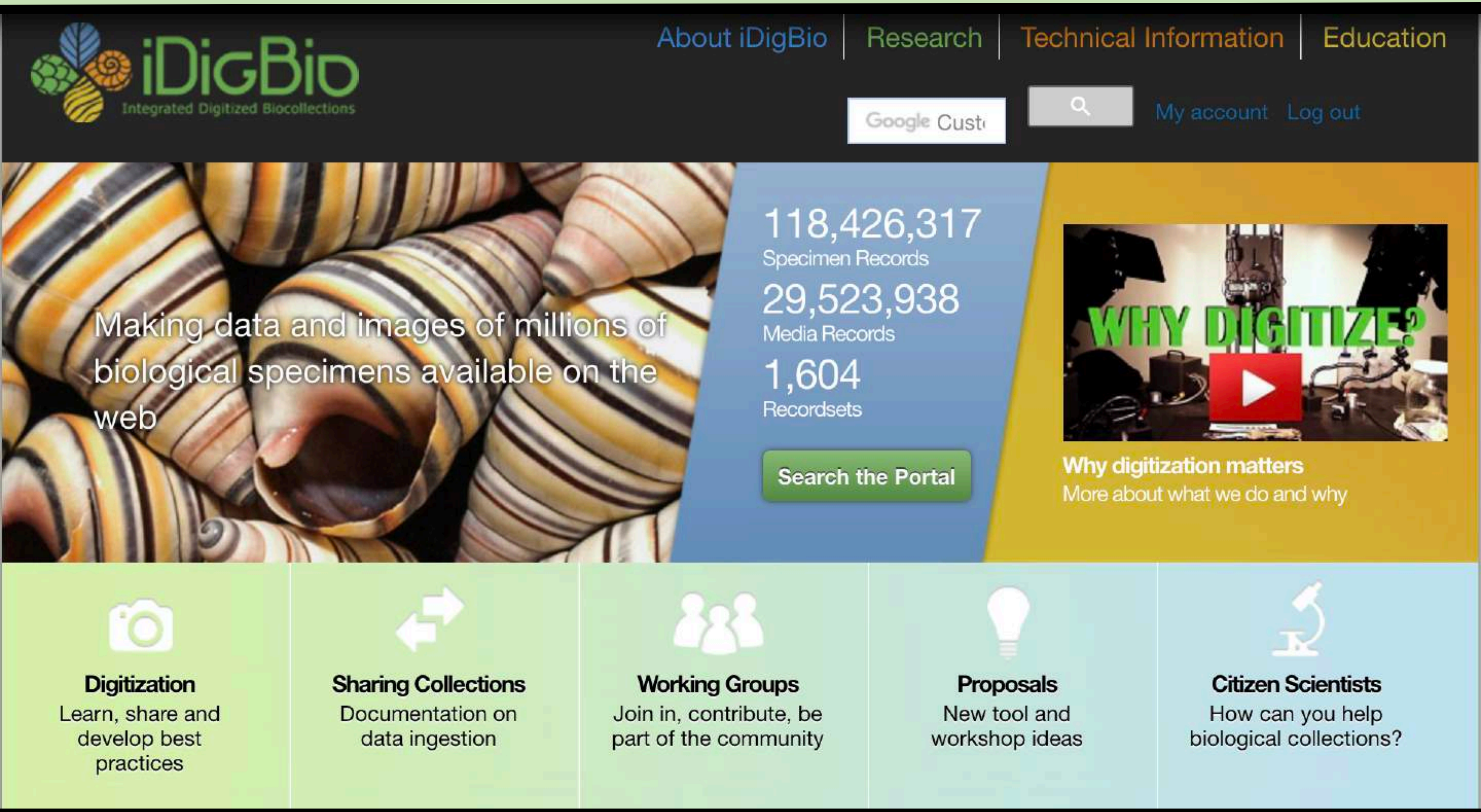
Ingest, serve, integrate data:

Localities

Dates

Images

Digitized Data & Biodiversity Research



Making data and images of millions of biological specimens available on the web

118,426,317
Specimen Records

29,523,938
Media Records

1,604
Recordsets

Why digitization matters
More about what we do and why

Digitization
Learn, share and develop best practices

Sharing Collections
Documentation on data ingestion

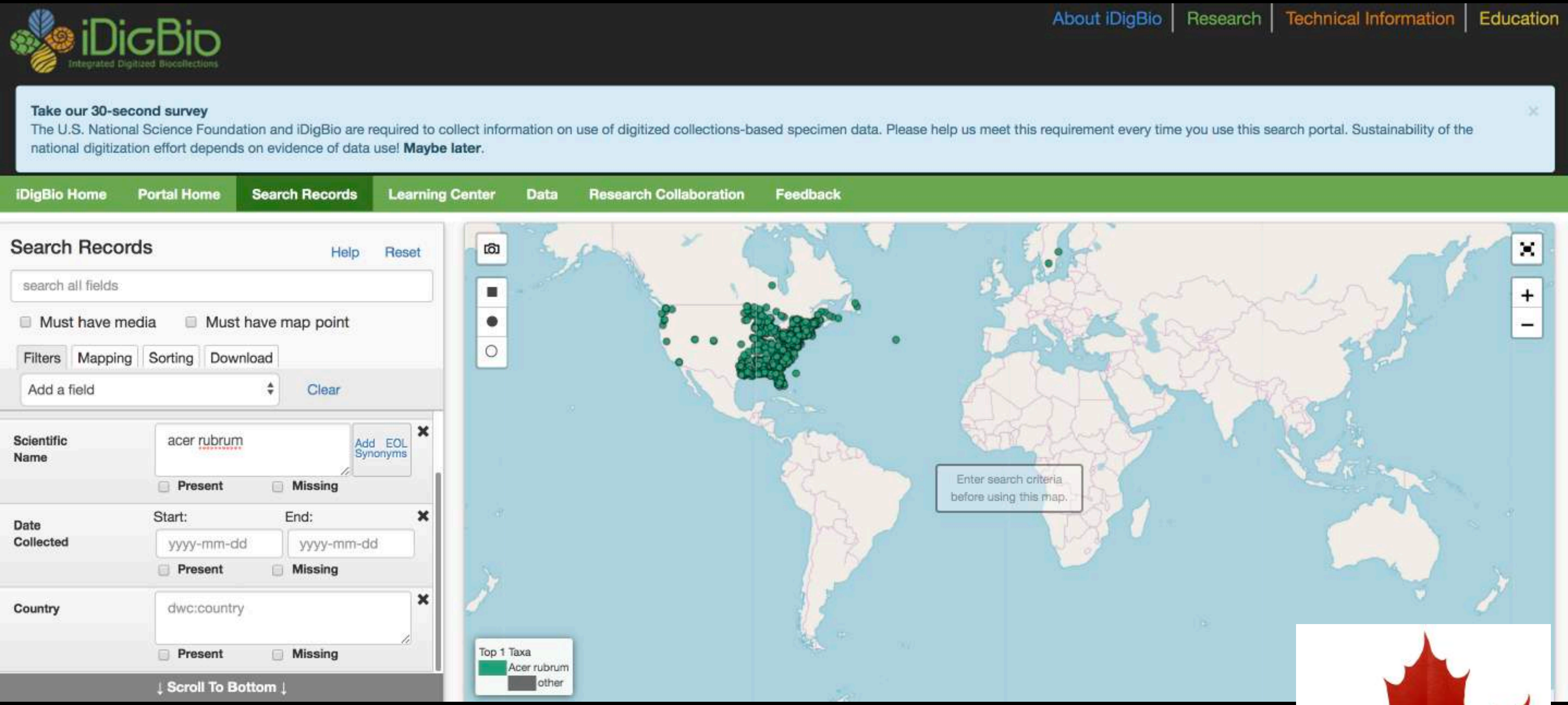
Working Groups
Join in, contribute, be part of the community

Proposals
New tool and workshop ideas

Citizen Scientists
How can you help biological collections?

www.idigbio.org

Search Specimen Records: idigbio.org



Take our 30-second survey
The U.S. National Science Foundation and iDigBio are required to collect information on use of digitized collections-based specimen data. Please help us meet this requirement every time you use this search portal. Sustainability of the national digitization effort depends on evidence of data use! **Maybe later**.

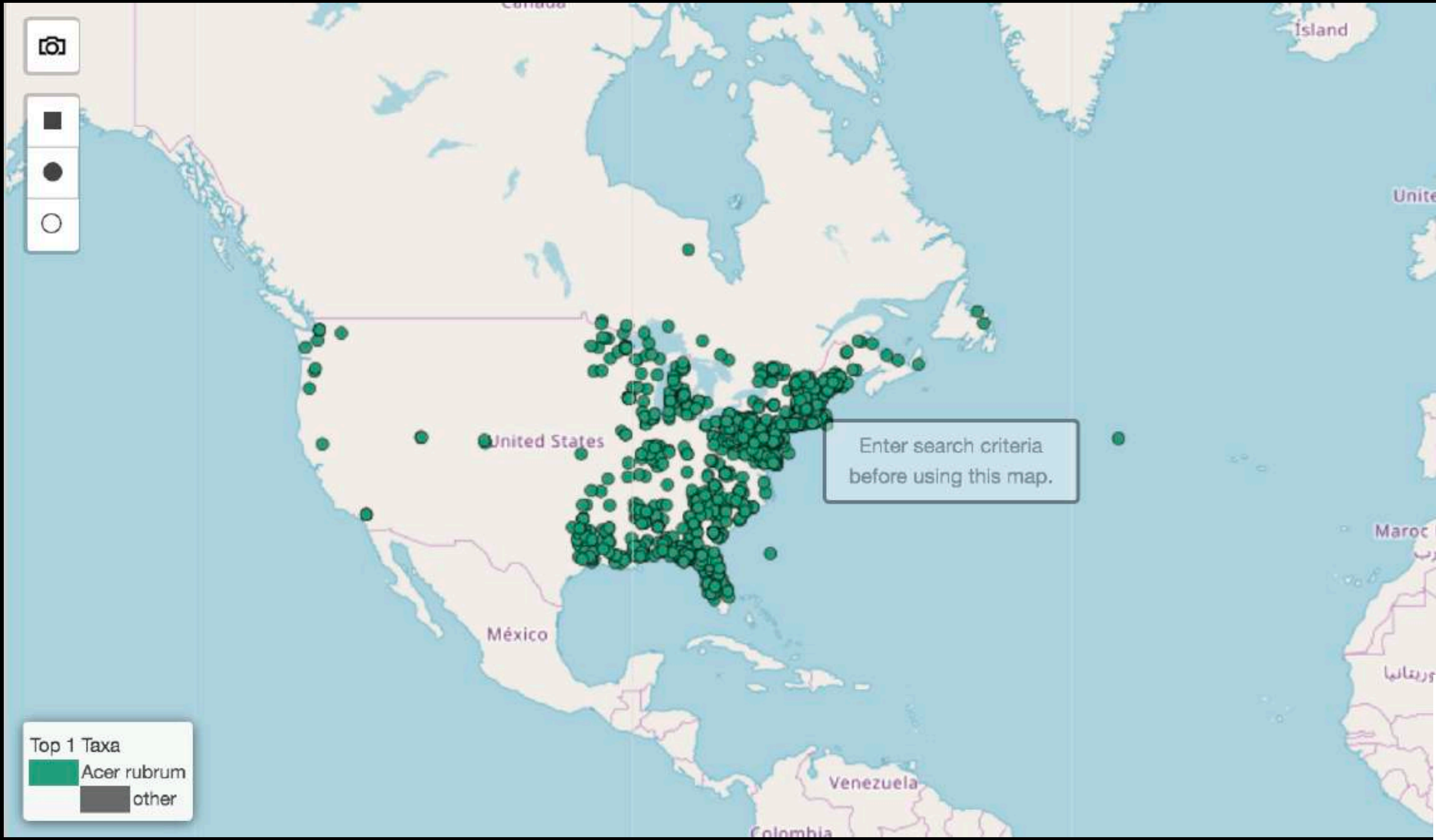
Scientific Name: acer rubrum
□ Present □ Missing

Date Collected: Start: yyyy-mm-dd End: yyyy-mm-dd
□ Present □ Missing

Country: dwc:country
□ Present □ Missing

www.idigbio.org

Search Specimen Records: idigbio.org

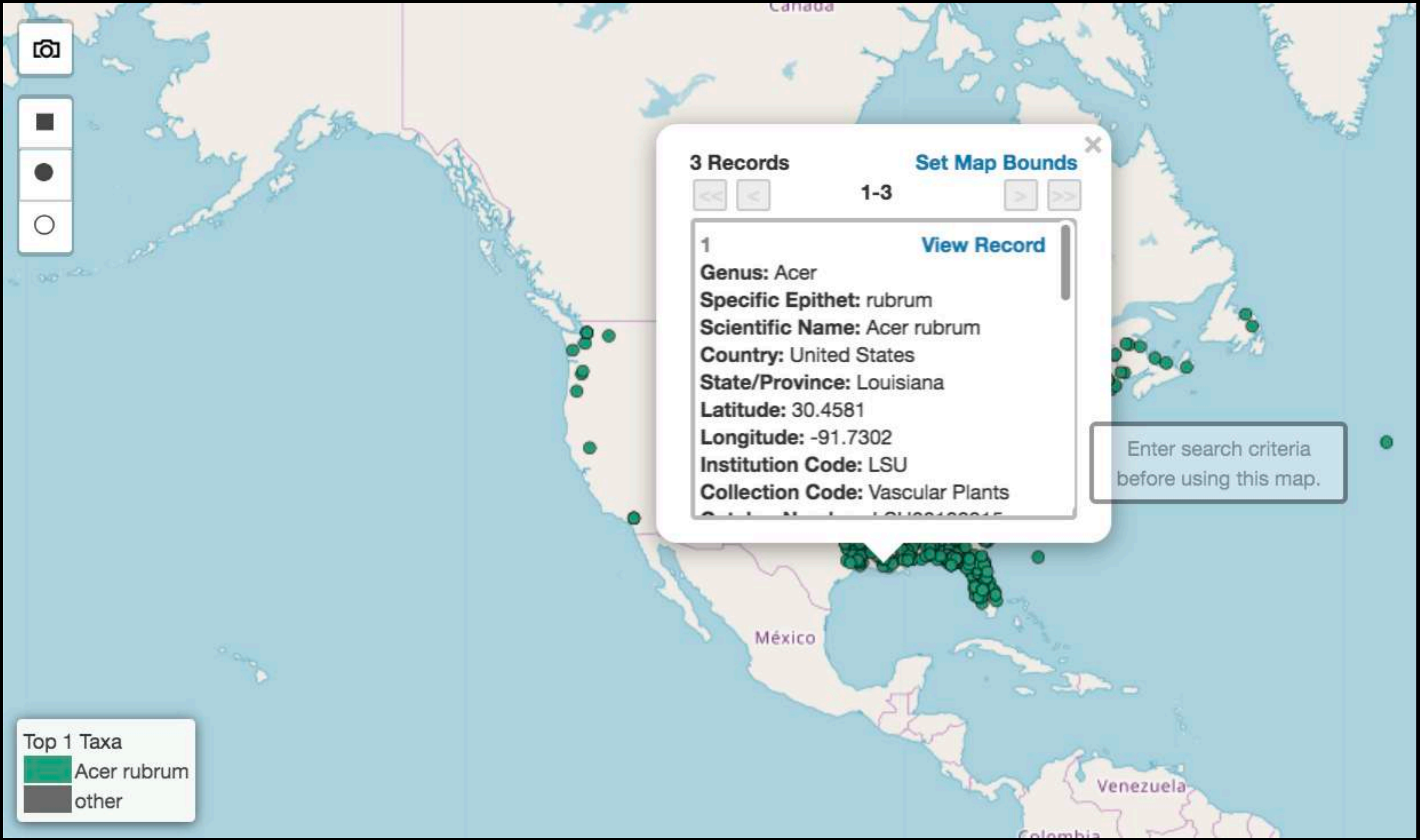


Enter search criteria before using this map.

Top 1 Taxa
🟩 Acer rubrum
▨ other

www.idigbio.org

Search Specimen Records: idigbio.org



3 Records

1-3

1

Genus: Acer

Specific Epithet: rubrum

Scientific Name: Acer rubrum

Country: United States

State/Province: Louisiana

Latitude: 30.4581

Longitude: -91.7302

Institution Code: LSU

Collection Code: Vascular Plants

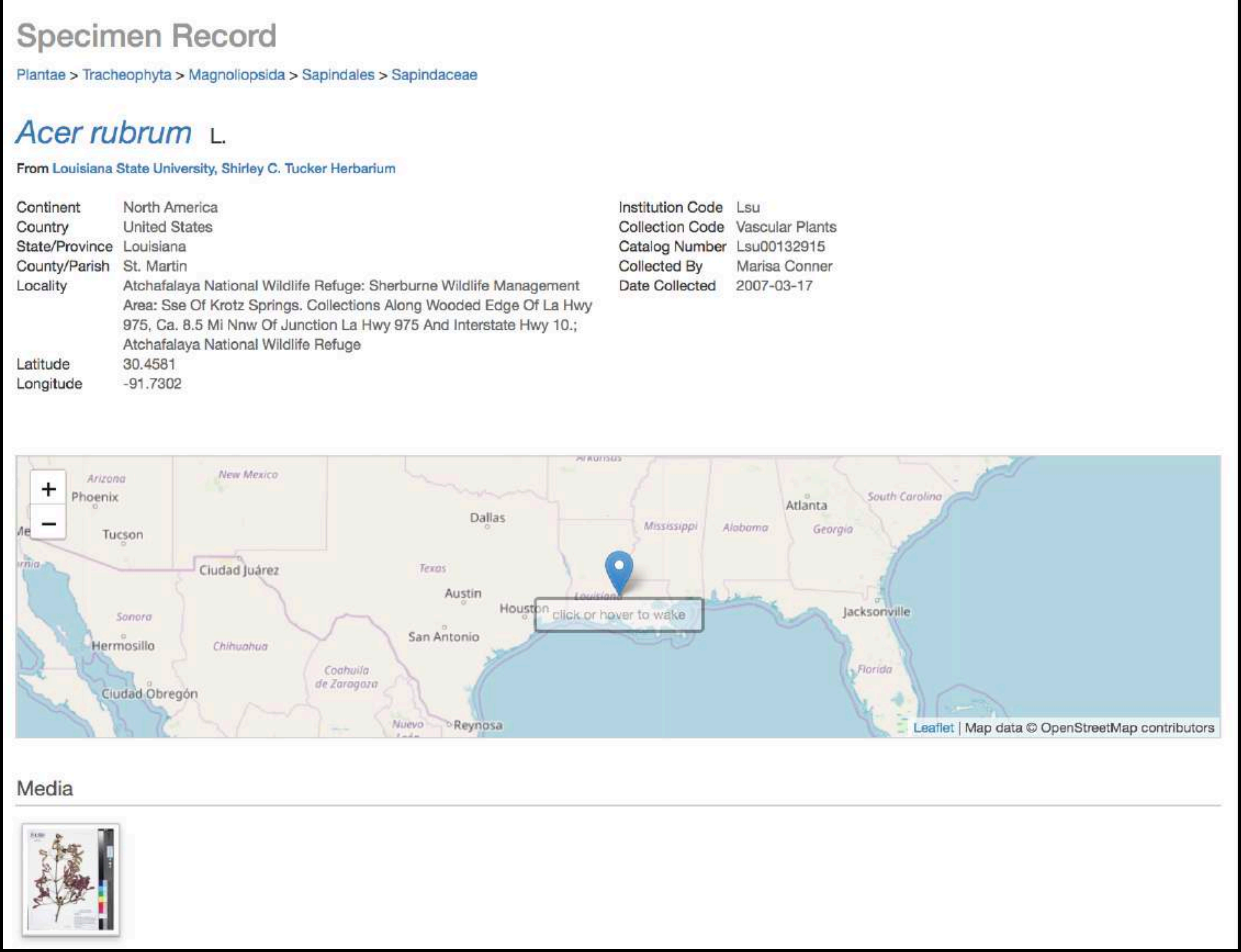
Enter search criteria before using this map.

Top 1 Taxa

🟩 Acer rubrum

⬛ other

Search Specimen Records: idigbio.org



Specimen Record

Plantae > Tracheophyta > Magnoliopsida > Sapindales > Sapindaceae

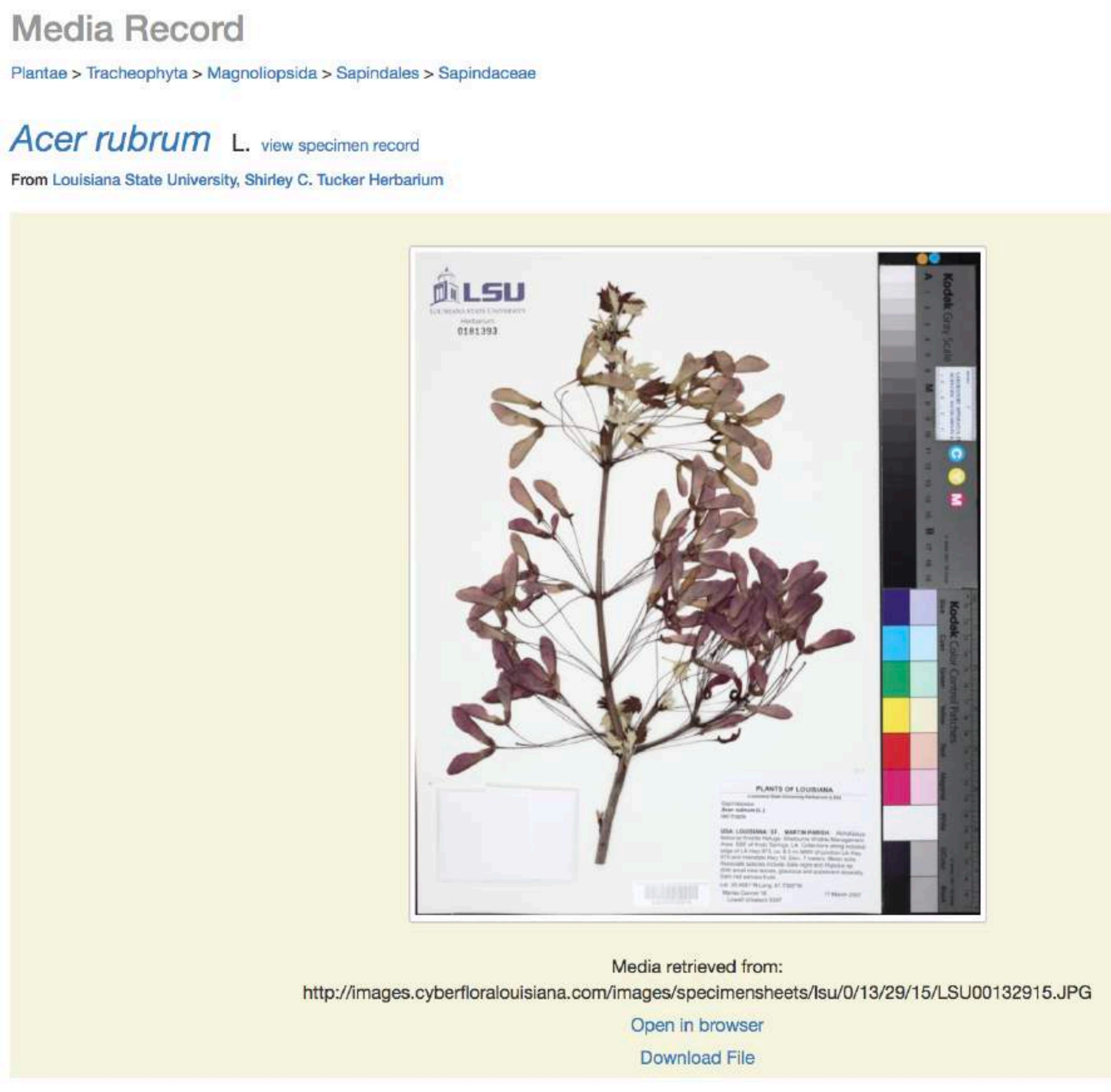
Acer rubrum L.

From Louisiana State University, Shirley C. Tucker Herbarium

Continent	North America	Institution Code	Lsu
Country	United States	Collection Code	Vascular Plants
State/Province	Louisiana	Catalog Number	Lsu00132915
County/Parish	St. Martin	Collected By	Marisa Conner
Locality	Atchafalaya National Wildlife Refuge: Sherburne Wildlife Management Area: Sse Of Krotz Springs. Collections Along Wooded Edge Of La Hwy 975, Ca. 8.5 Mi Nnw Of Junction La Hwy 975 And Interstate Hwy 10.; Atchafalaya National Wildlife Refuge	Date Collected	2007-03-17
Latitude	30.4581		
Longitude	-91.7302		

Media

Search Specimen Records: idigbio.org



Media Record

Plantae > Tracheophyta > Magnoliopsida > Sapindales > Sapindaceae

Acer rubrum L. view specimen record

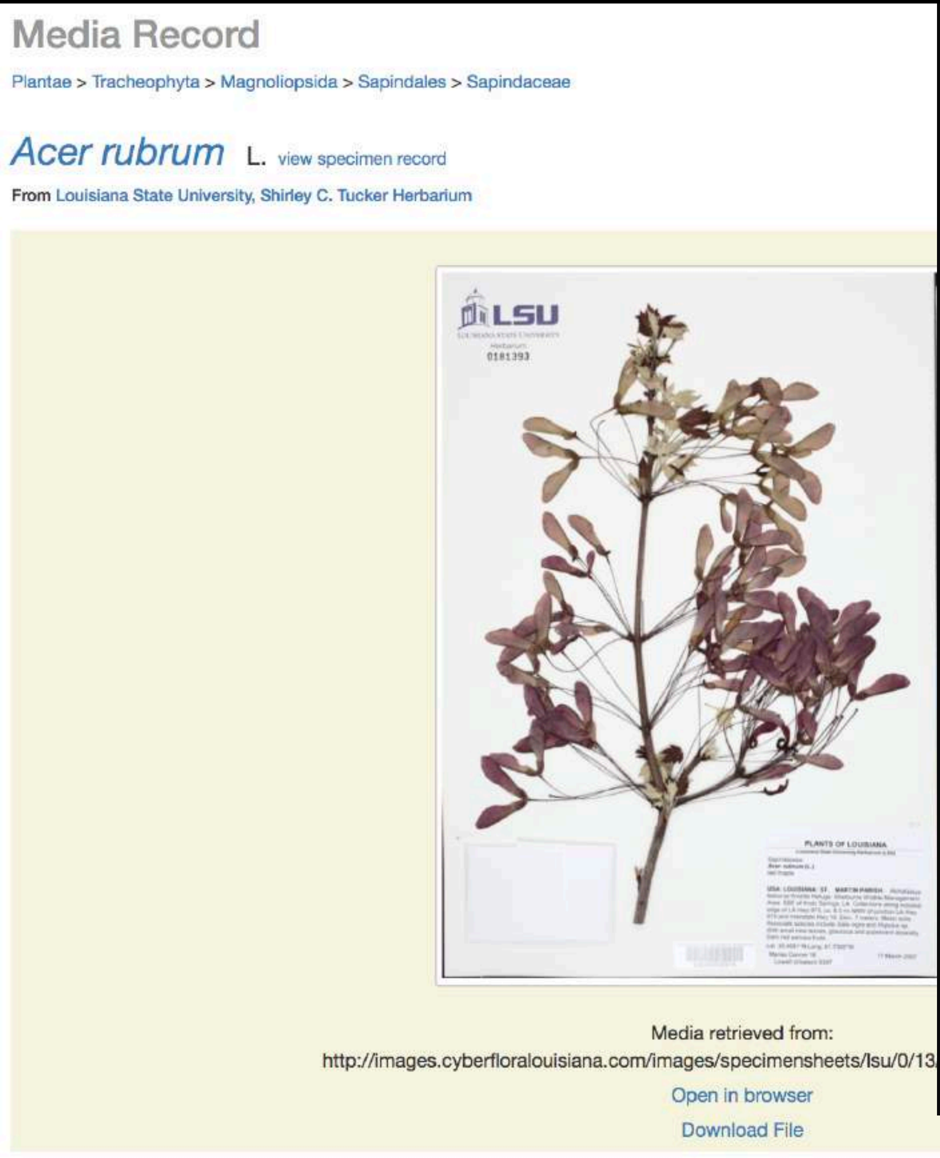
From Louisiana State University, Shirley C. Tucker Herbarium

Media retrieved from:
http://images.cyberfloralouisiana.com/images/specimensheets/lsu/0/13/29/15/LSU00132915.JPG

Open in browser

Download File

Search Specimen Records: idigbio.org



Media Record

Plantae > Tracheophyta > Magnoliopsida > Sapindales > Sapindaceae

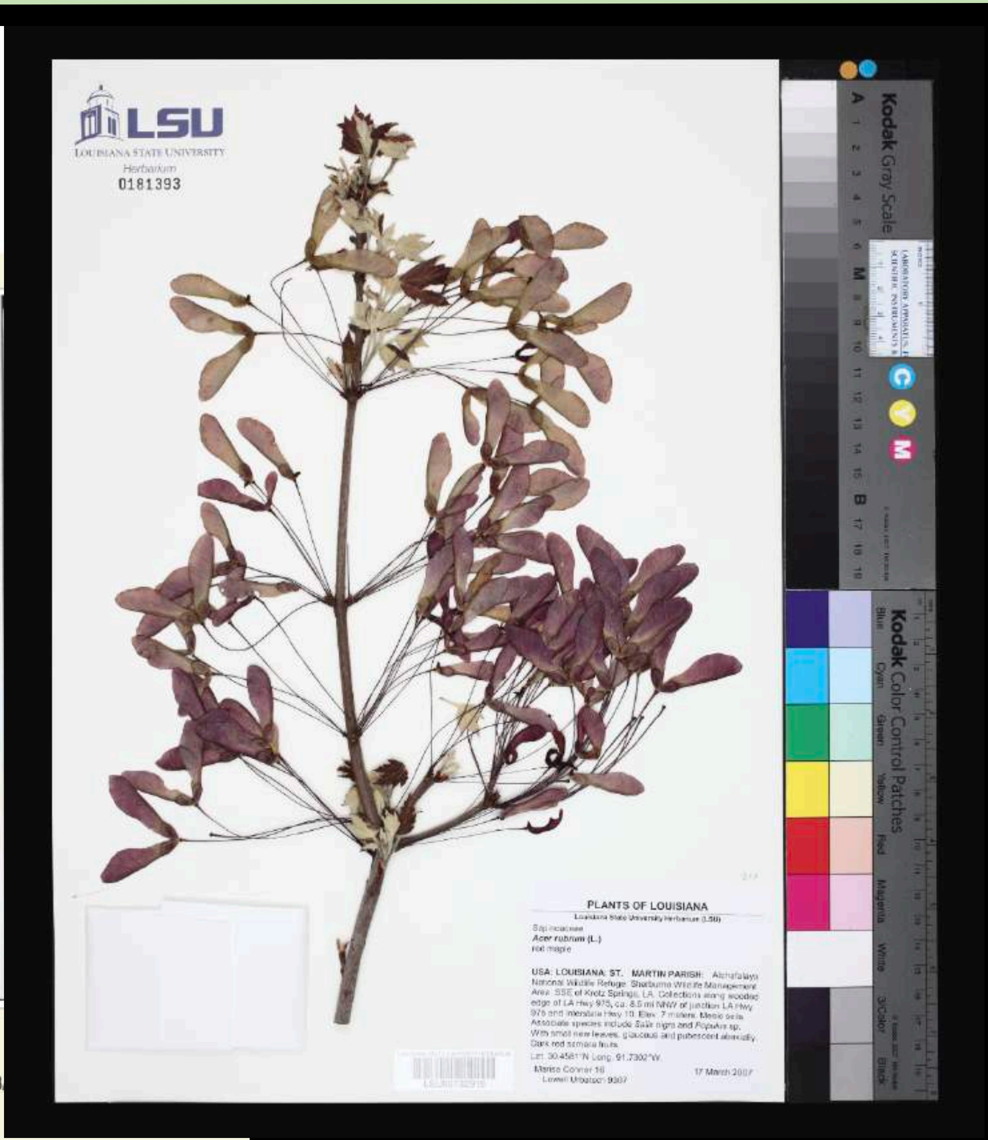
Acer rubrum L. view specimen record

From Louisiana State University, Shirley C. Tucker Herbarium

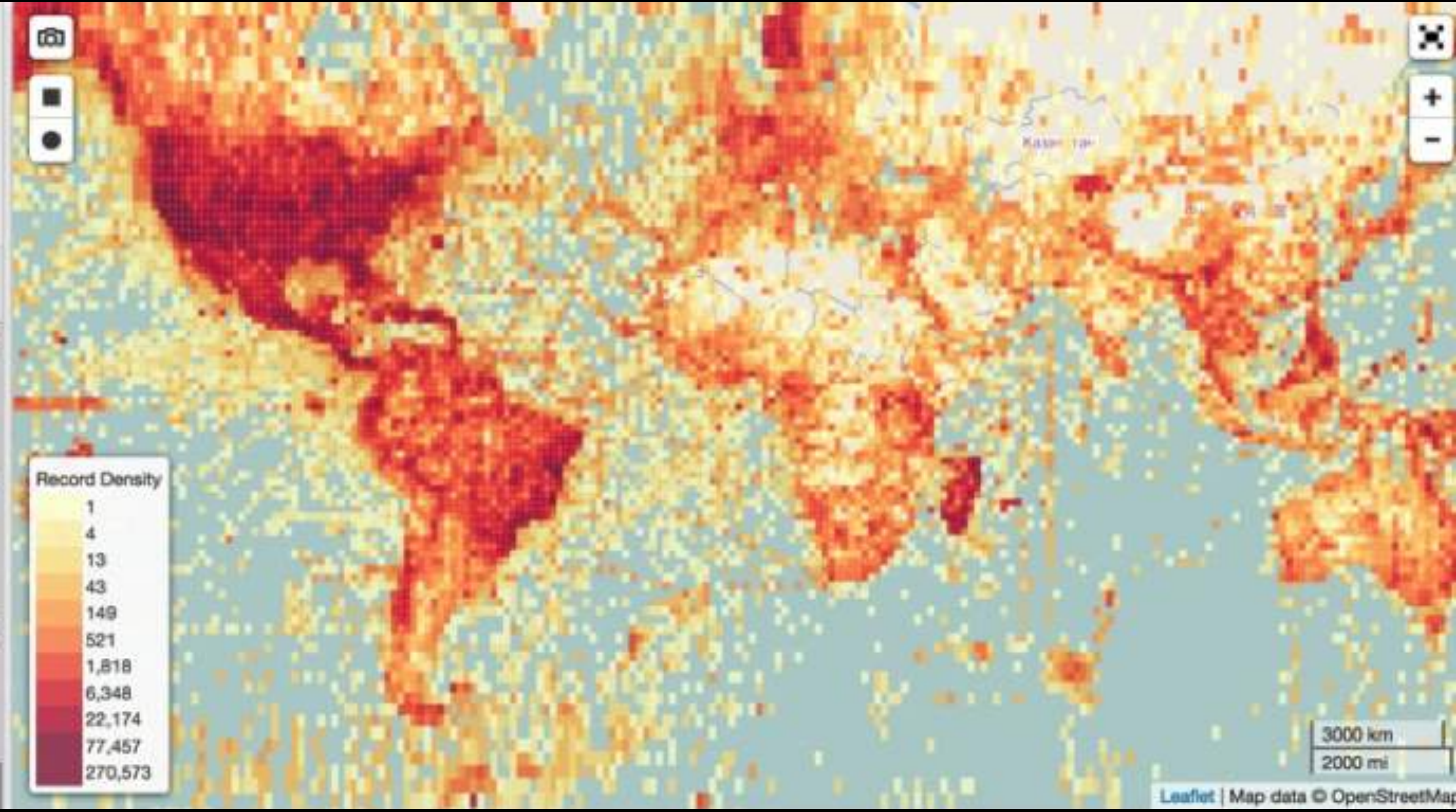
Media retrieved from:
http://images.cyberfloralouisiana.com/images/specimensheets/lsu/0/13

Open in browser

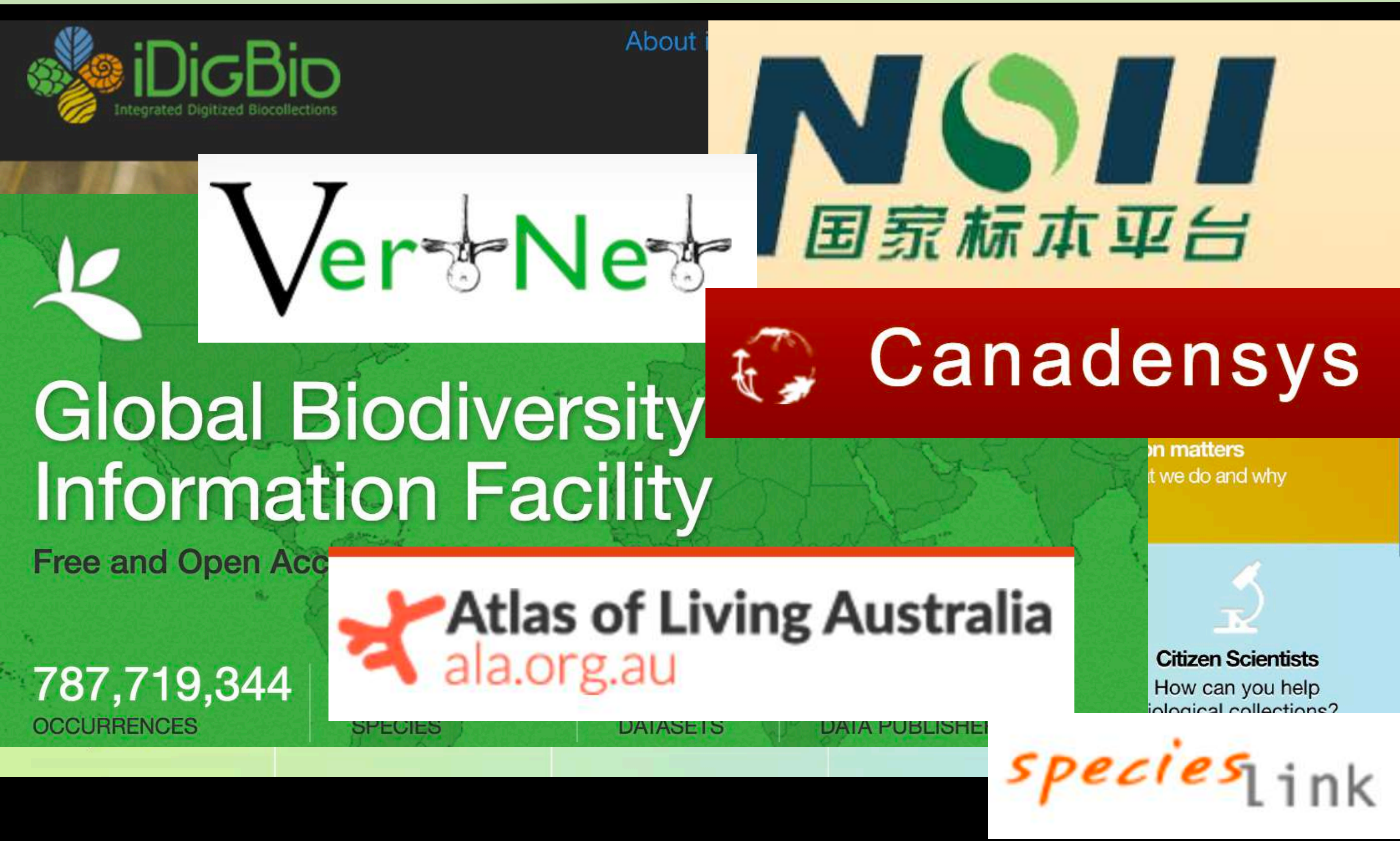
Download File


LSU
LOUISIANA STATE UNIVERSITY
Herbarium
0181393
Kodak Gray Scale
A 1 2 3 4 5 6 M 8 9 10 11 12 13 14 15 B 17 18 19
C
Y
M
Kodak Color Control Patches
Blue
Cyan
Green
Yellow
Red
Magenta
White
3/Color
Black
PLANTS OF LOUISIANA
Louisiana State University Herbarium (LSU)
Sapindaceae
Acer rubrum (L.)
Red maple
USA: LOUISIANA: ST. MARTIN PARISH: Atchafalaya National Wildlife Refuge: Sherburne Wildlife Management Area. SSE of Krotz Springs, LA. Collections along wooded edge of LA Hwy 975, ca. 8.5 mi NNW of junction LA Hwy 975 and Interstate Hwy 10. Elev. 7 meters. Mesic soils. Associate species include Salix nigra and Populus sp. With small new leaves, glaucous and pubescent abaxially. Dark red samara fruits.
Lat. 30.4581°N Long. 91.7302°W.
Marisa Conner 16
Lowell Urbatsch 9307
17 March 2007

Specimen Localities in iDigBio

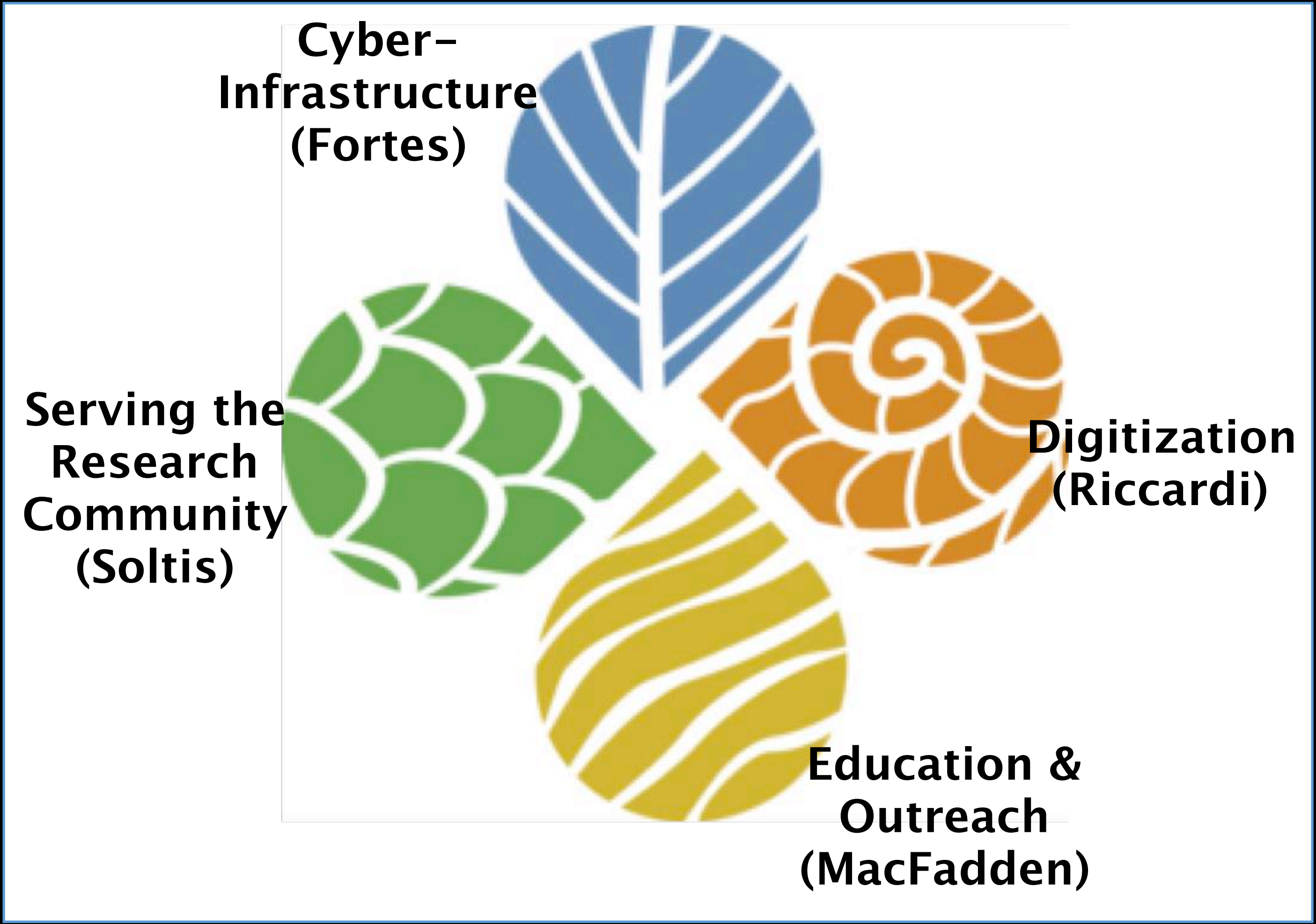


Other Data Aggregators



Components of iDigBio

G. Nelson, Director



Cyber-Infrastructure (Fortes)

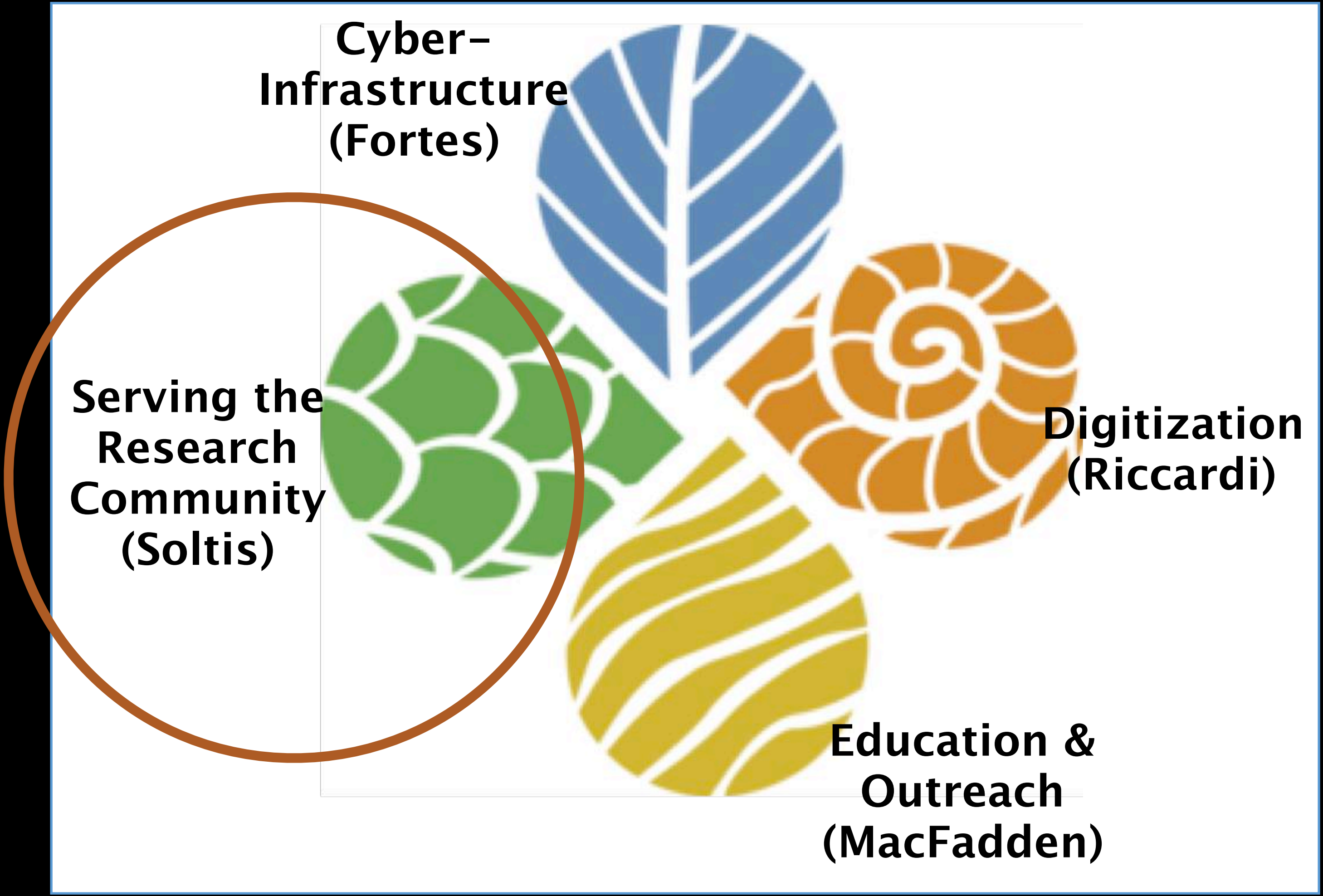
Serving the Research Community (Soltis)

Digitization (Riccardi)

Education & Outreach (MacFadden)

Components of iDigBio

G. Nelson, Director

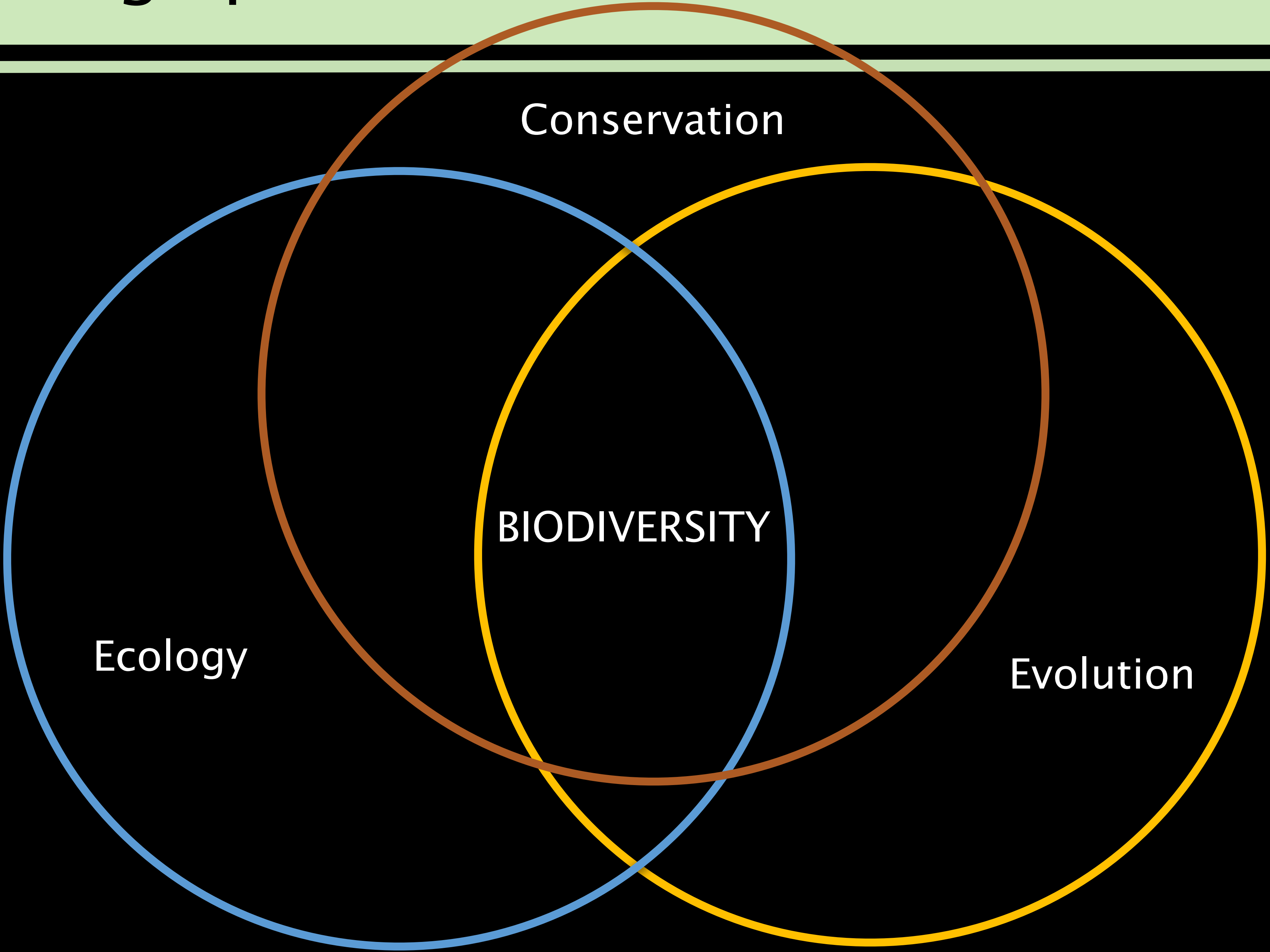


Using Specimen Data for Research

Big questions in biodiversity research:

- How many species are there?
- Why are species where they are?
- How does habitat change affect species?
- How does invasion by exotic species affect the stability of communities of species?
- How will a changing climate affect species— extinction rates, distribution, evolution?

Using Specimen Data for Research



What is **biodiversity**?

“Biodiversity defies easy definition, but we value it nonetheless, much the same way we value justice, freedom, and nature, similarly difficult terms to define.”

Naeem S et al. 2016, Proc. R. Soc. B

Different Measures of Biodiversity

- Species richness— How many species are in the area?

- Richness of endemic species- How many endemic species are there?

- Functional diversity - How similar are the functional traits of species in one area to one another?

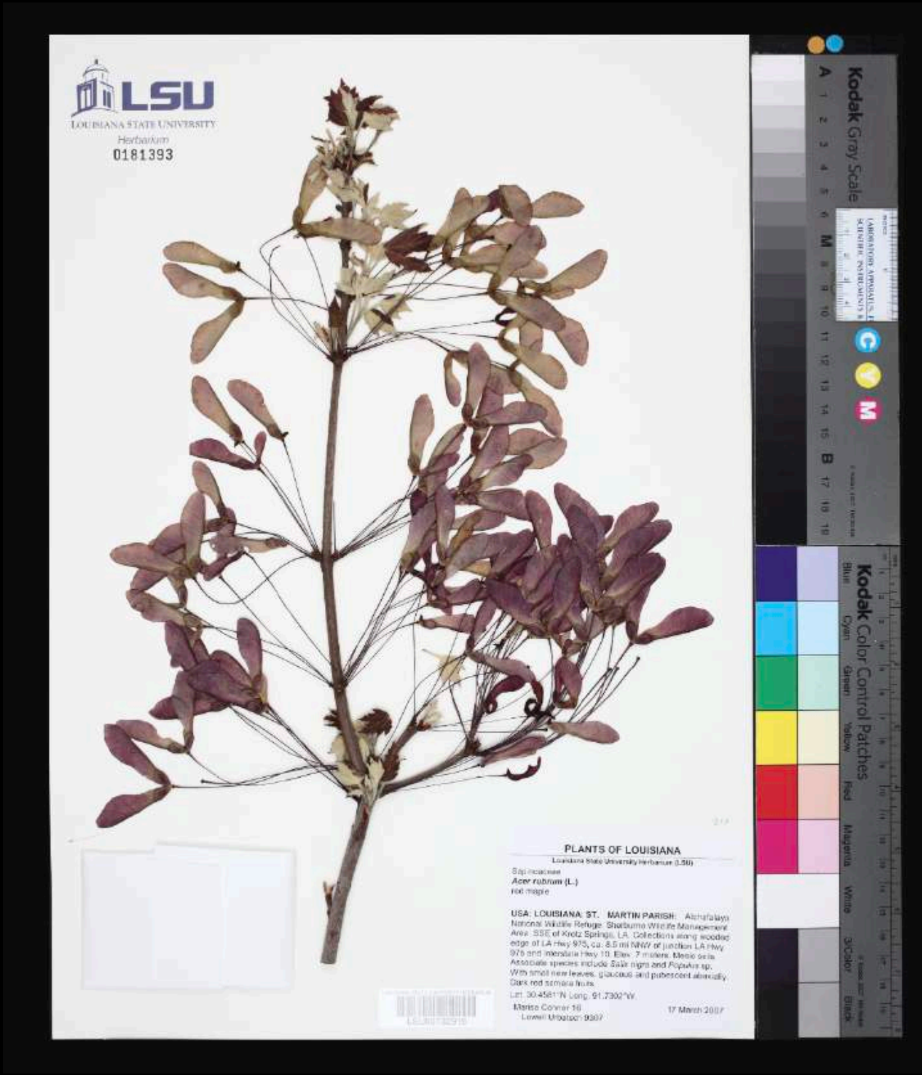
- Phylogenetic diversity- How closely related are the species in an area?

Naeem S et al. 2016, Proc. R. Soc. B

Using AI to generate data for biodiversity research

- Can AI be used to identify plants?

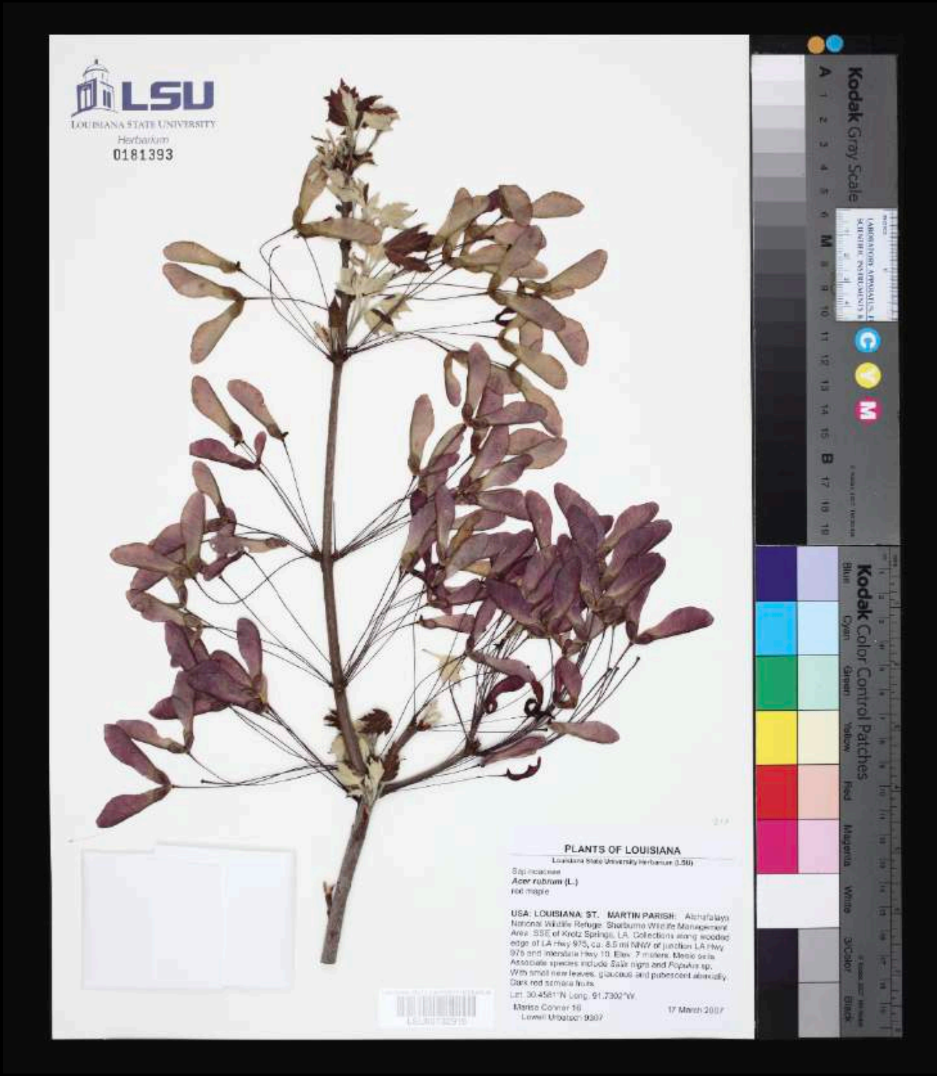
- Can AI be used to label phenological stages of plants?


LSU
LOUISIANA STATE UNIVERSITY
Herbarium
0181393
Kodak Gray Scale
Kodak Color Control Patches
PLANTS OF LOUISIANA
Louisiana State University Herbarium (LSU)
Acer rubrum (L.)
red maple
USA: LOUISIANA: ST. MARTIN PARISH:
Lat. 30.4581°N Long. 91.7302°W
17 March 2007

Using AI to generate data for biodiversity research

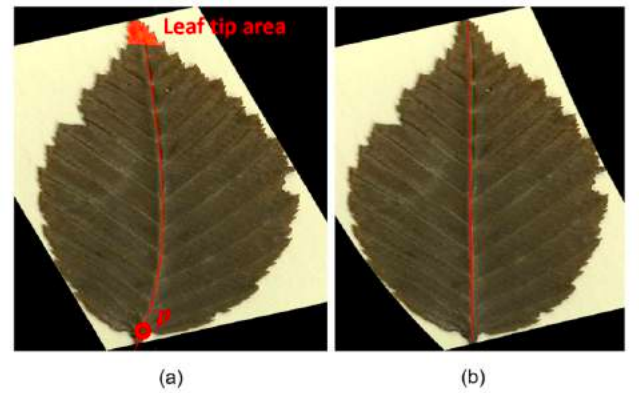
- Can AI be used to identify plants?
- Can AI be used to label phenological stages of plants?

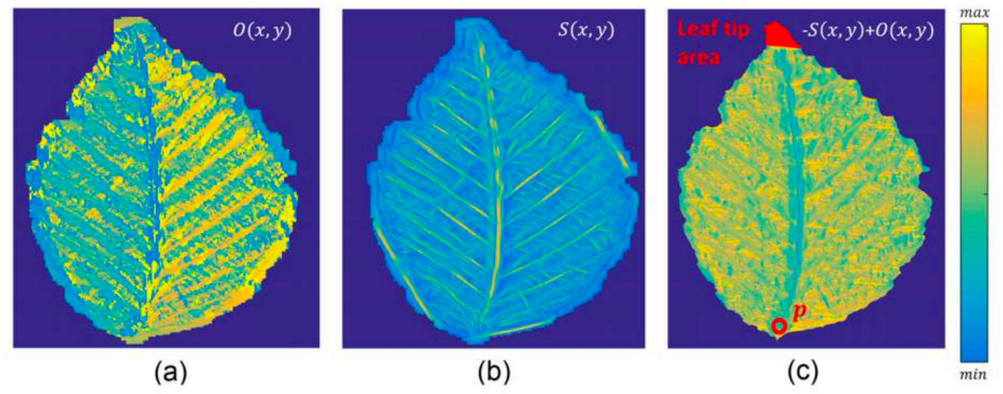
YES!

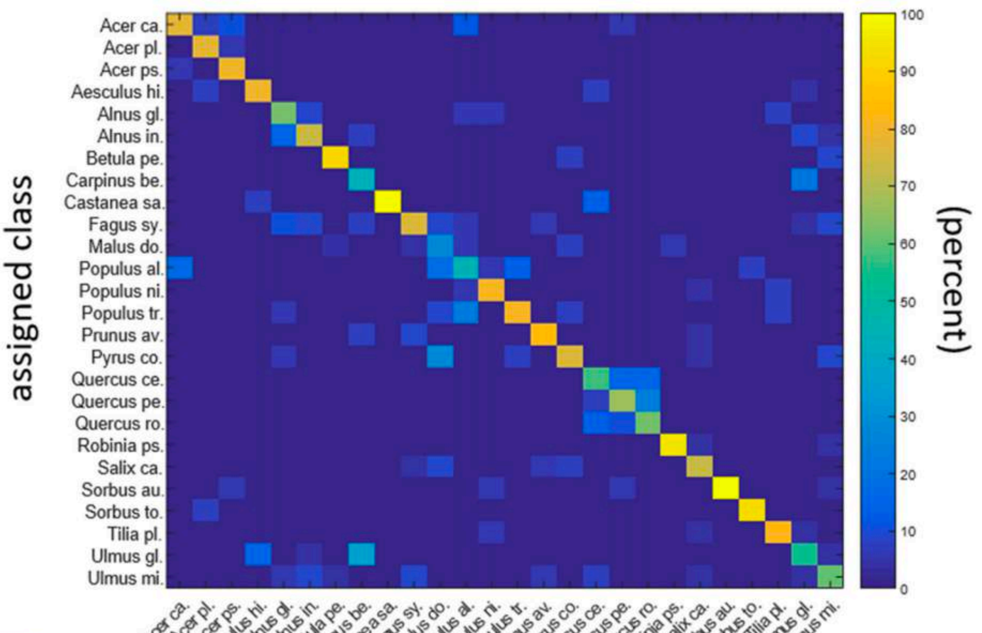

LSU
LOUISIANA STATE UNIVERSITY
Herbarium
0181393
Kodak Gray Scale
A 1 2 3 4 5 6 M 8 9 10 11 12 13 14 15 B 17 18 19
C Y M
Kodak Color Control Patches
Blue Cyan Green Yellow Red Magenta White 3/Color Black
PLANTS OF LOUISIANA
Louisiana State University Herbarium (LSU)
Sapindaceae
Acer rubrum (L.)
red maple
USA: LOUISIANA: ST. MARTIN PARISH: Atchafalaya National Wildlife Refuge: Sherburne Wildlife Management Area. SSE of Krotz Springs, LA. Collections along wooded edge of LA Hwy 975, ca. 8.5 mi NNW of junction LA Hwy 975 and Interstate Hwy 10. Elev. 7 meters. Mesic soils. Associate species include Salix nigra and Populus sp. With small new leaves, glaucous and pubescent abaxially. Dark red samara fruits.
Lat. 30.4581°N Long. 91.7302°W
Marisa Conner 16
Lowell Urbatsch 9307
17 March 2007

Machine Learning: Plant ID

Machine Learning: Herbarium specimens
Classifying German trees to species
Leaf shape, venation
85% accuracy



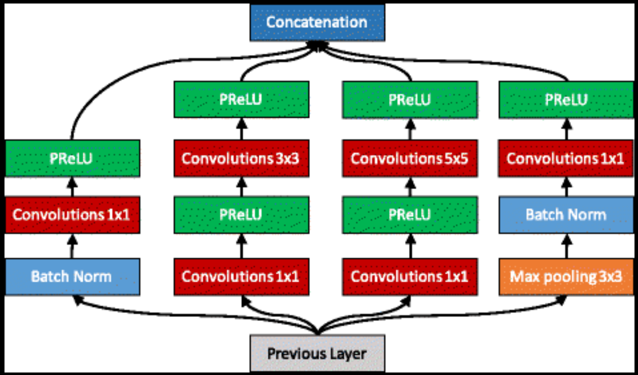


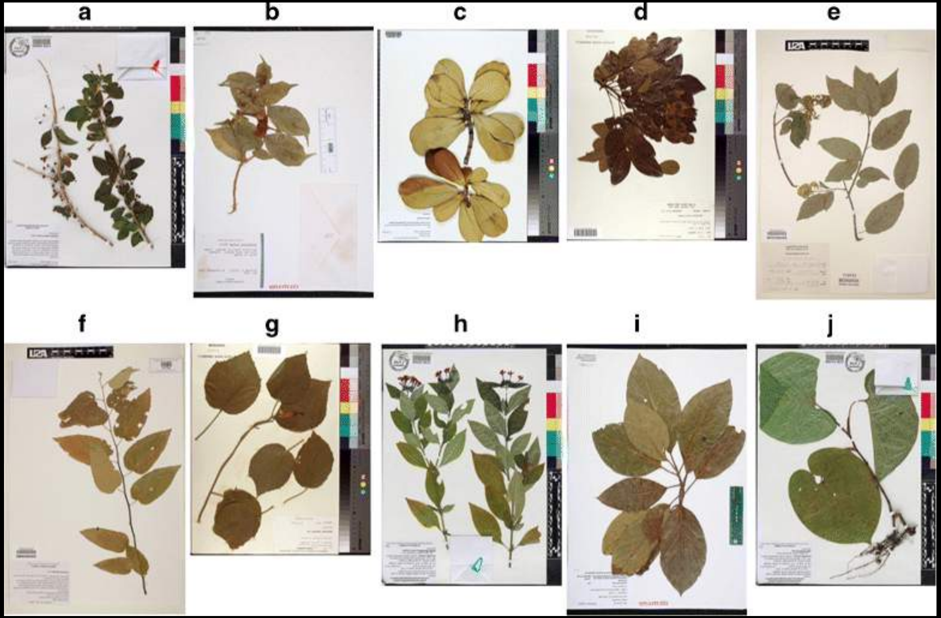


Unger et al. 2016

Machine Learning: Plant ID

Deep Learning: Herbarium specimens (2 data sets)
Classifying plants to species
>1200 species
250,000 images from iDigBio
90% accuracy

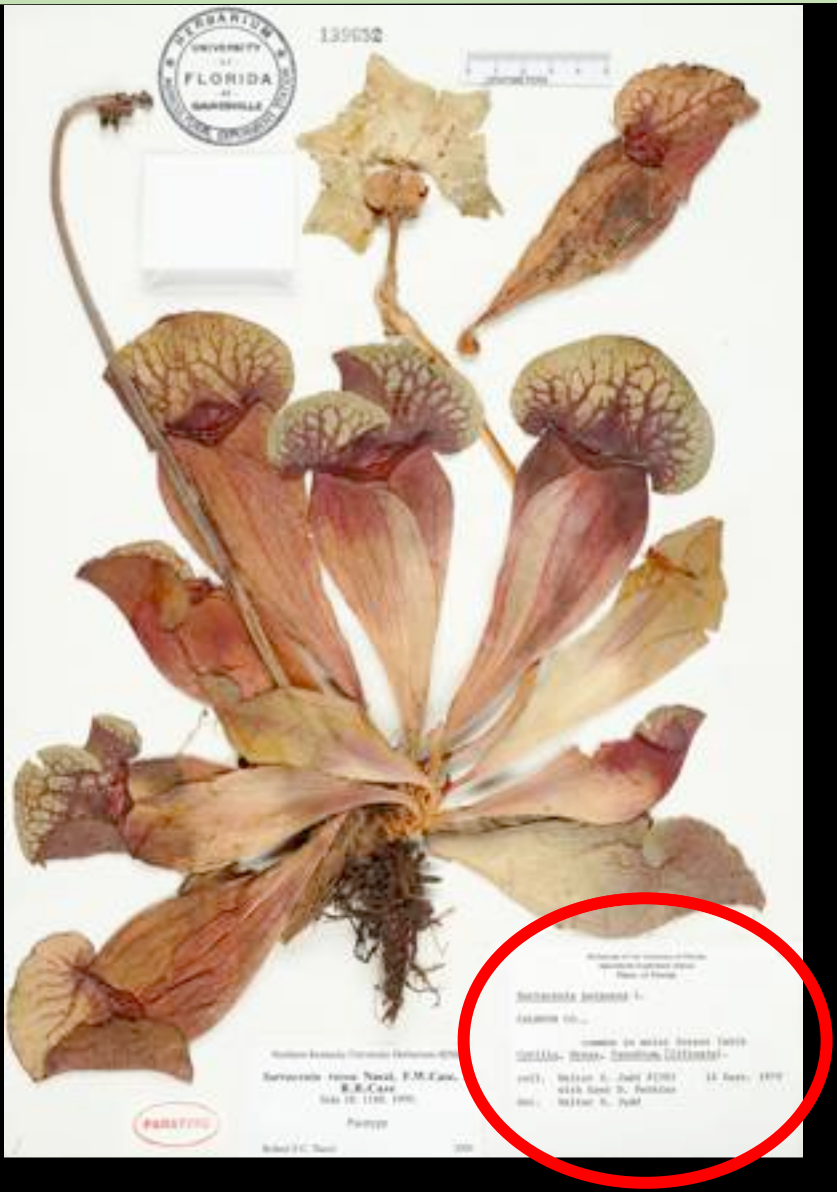




Carranza-Rojas et al. 2017

Label Data from Herbarium Specimens

- Scientific name – including authority
- Date
- Collector
- Location – state, county, specific site, GPS coordinates
- Associated species
- Notes



Phenology: “Nature’s Calendar”
Bud Burst, Flowering, Fruiting

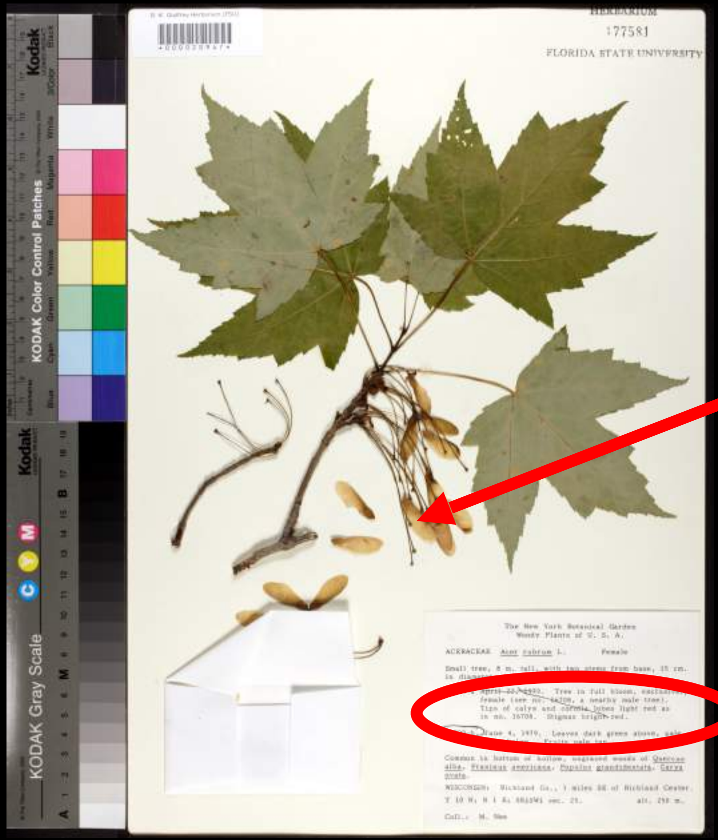
www.sbs.utexas.edu www.sites.psu.edu

Phenological Data from Labels and Images

Phenological data – as described in label notes

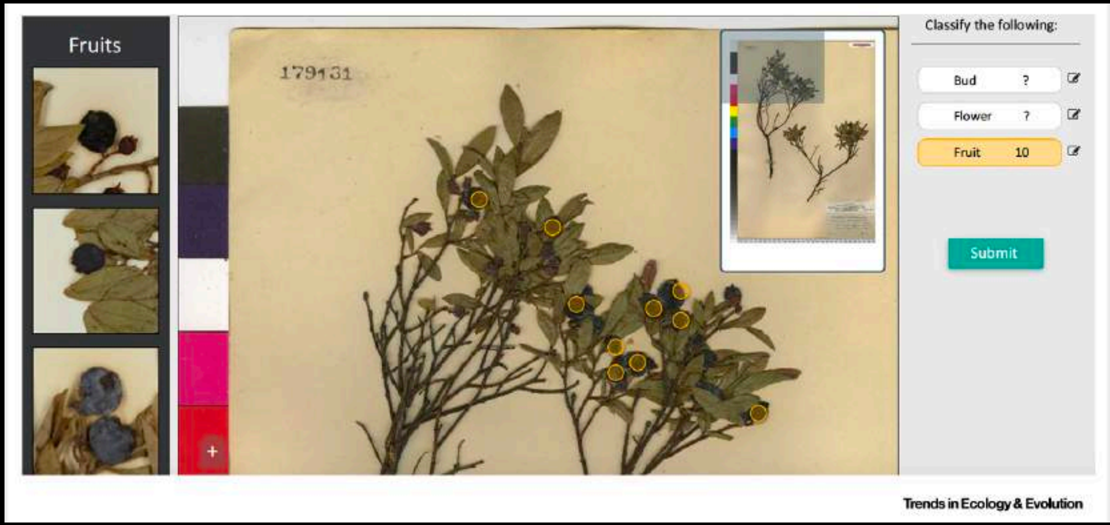
“tree in full fruit…”

or from image itself:



Kodak
KODAK Color Control Patches
KODAK Gray Scale
HERBARIUM
177581
FLORIDA STATE UNIVERSITY
The New York Botanical Garden
Woody Plants of U. S. A.
ACERACEAE Acer rubrum L. Female
Small tree, 8 m. tall, with two stems from base, 35 cm. in diameter.
April 22, 1979. Tree in full bloom, exclusively female (see no. 16708, a nearby male tree). Tips of calyx and corolla lobes light red as in no. 16708. Stigmas bright red.
June 4, 1979. Leaves dark green above, pale below. Fruits pale tan.
Common in bottom of hollow, ungrazed woods of Quercus alba, Fraxinus americana, Populus grandidentata, Carya ovata.
WISCONSIN: Richland Co., 1 miles SE of Richland Center.
alt. 250 m.
Coll.: M. Nee

CrowdCurio, in Willis et al. 2017



Machine Learning and Phenology

Maples (*Acer*): unfolded leaves present or absent?

unfolded leaves present

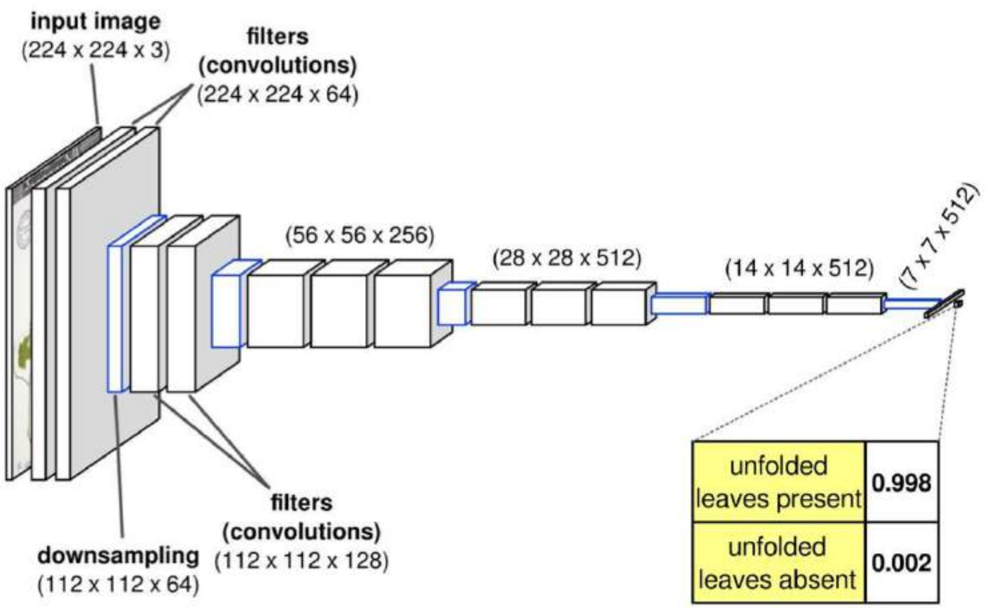
unfolded leaves absent

***Prunus*: flowers present or absent?**

flowers present

flowers absent

Deep convolutional neural network

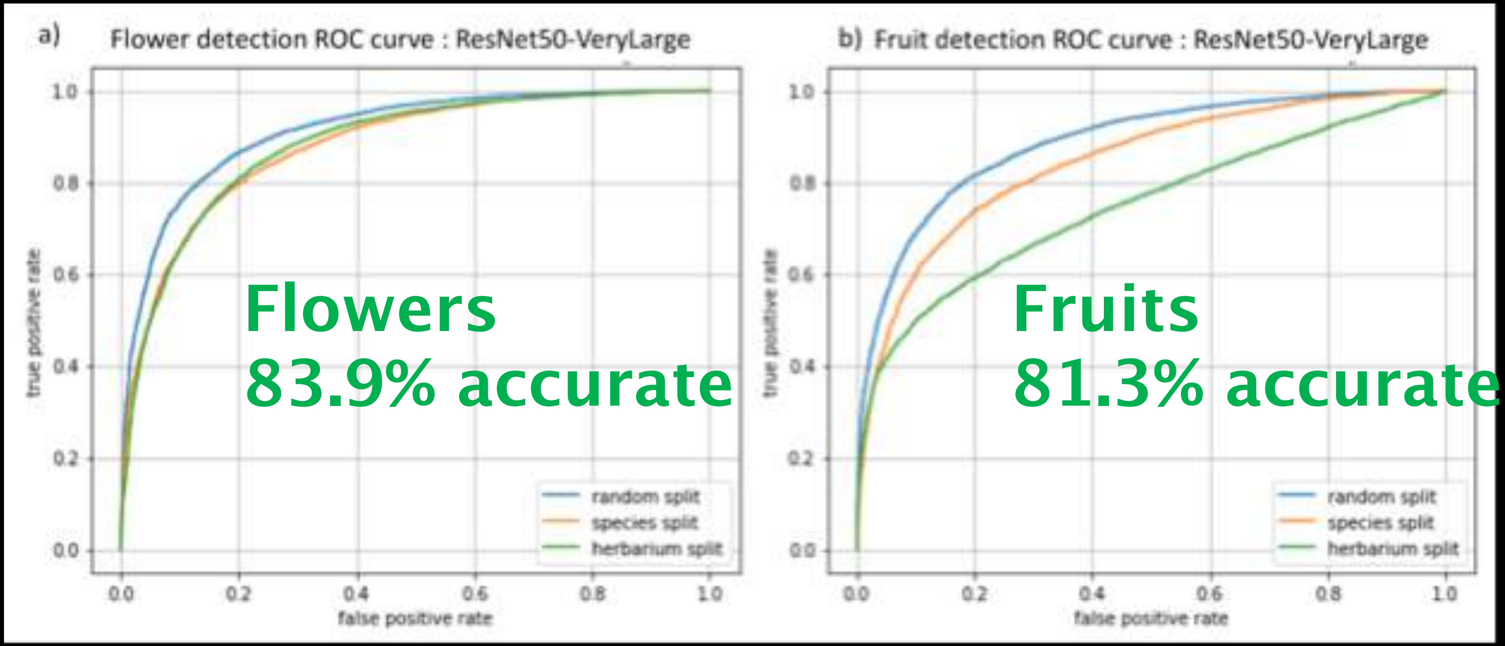


Architecture based on Simonyan and Zisserman (2015), arXiv:1409.1556v6

Brian Stucky

Machine Learning and Phenology

Deep learning annotation
3 herbarium data sets:
  163,233 specimens
  7782 species, 236 families
4th data set: Asteraceae



a) Flower detection ROC curve : ResNet50-VeryLarge

b) Fruit detection ROC curve : ResNet50-VeryLarge

Flowers
83.9% accurate

Fruits
81.3% accurate

Lorieul et al. 2019

Phenological mismatches between species

Insect Emergence, Bird Nesting & Migration, and Phenological Synchrony/Asynchrony

Bartomeus et al. 2011

http://pheno-mismatch.org/

http://budburst.org/phenology_whyphenology

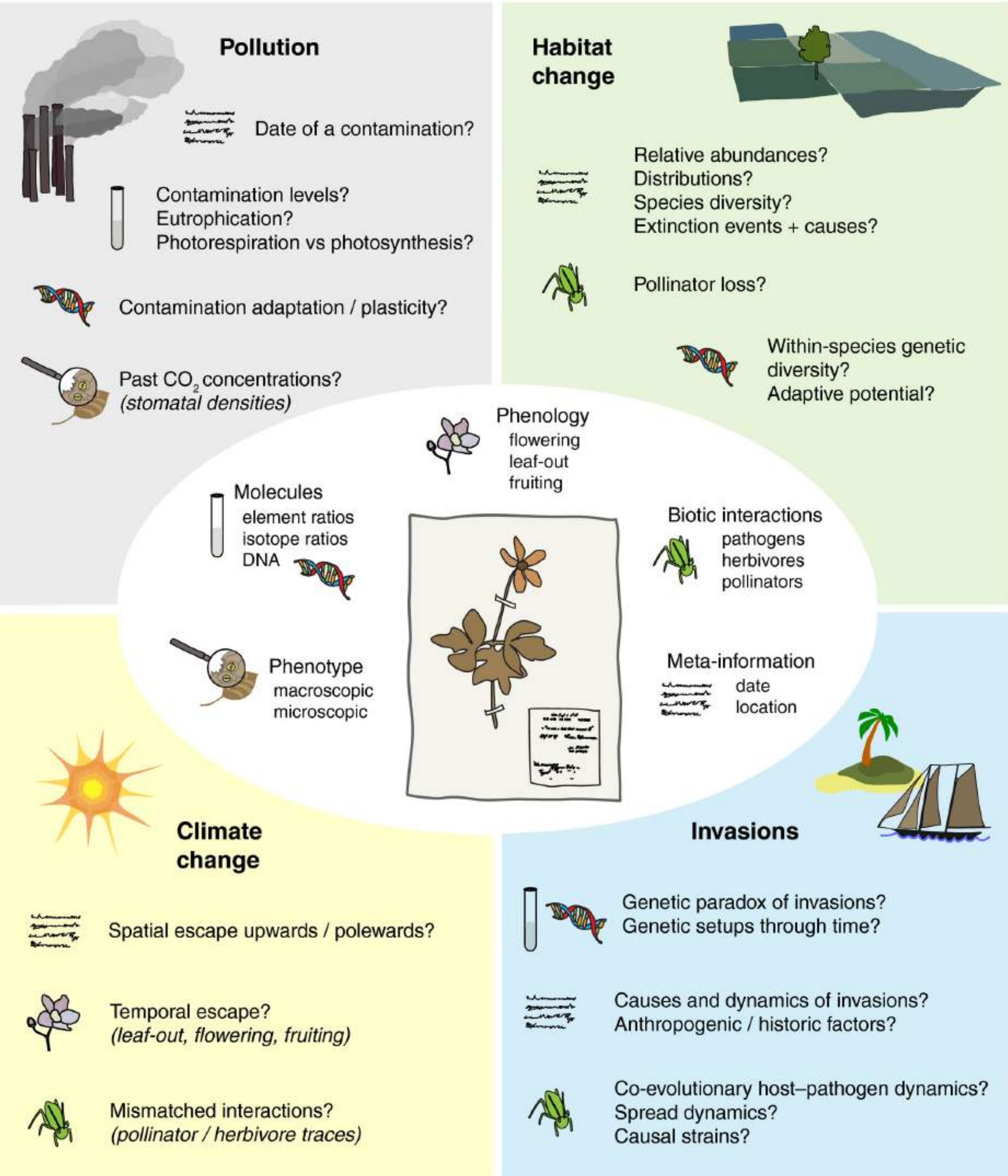
How will organisms respond to climate change?

Natural history collections contain data with….

Temporal Range
Taxonomic Breadth
Geographic Diversity
Intraspecific Diversity

…. invaluable to make predictions
about how species will behave
in the future.

Other uses of specimens



Pollution

Date of a contamination?

Contamination levels?
Eutrophication?
Photorespiration vs photosynthesis?

Contamination adaptation / plasticity?

Past CO_2 concentrations?
(stomatal densities)

Habitat change

Relative abundances?
Distributions?
Species diversity?
Extinction events + causes?

Pollinator loss?

Within-species genetic diversity?
Adaptive potential?

Phenology
flowering
leaf-out
fruiting

Molecules
element ratios
isotope ratios
DNA

Biotic interactions
pathogens
herbivores
pollinators

Phenotype
macroscopic
microscopic

Meta-information
date
location

Climate change

Spatial escape upwards / polewards?

Temporal escape?
(leaf-out, flowering, fruiting)

Mismatched interactions?
(pollinator / herbivore traces)

Invasions

Genetic paradox of invasions?
Genetic setups through time?

Causes and dynamics of invasions?
Anthropogenic / historic factors?

Co-evolutionary host–pathogen dynamics?
Spread dynamics?
Causal strains?

Lang et al. 2019, New Phytologist

Linking Specimen Data to other Data Sources

Climate

- World-Clim

Genetic Data

- Genbank
- 1KP project

Functional Trait Databases

- TRY

Character Traits

- Extract from literature
    - (eg. Biodiversity Heritage Library)
- Extract from specimens (images, etc)

… and more (eg. soil)

Specimen Data linked with Climate Data

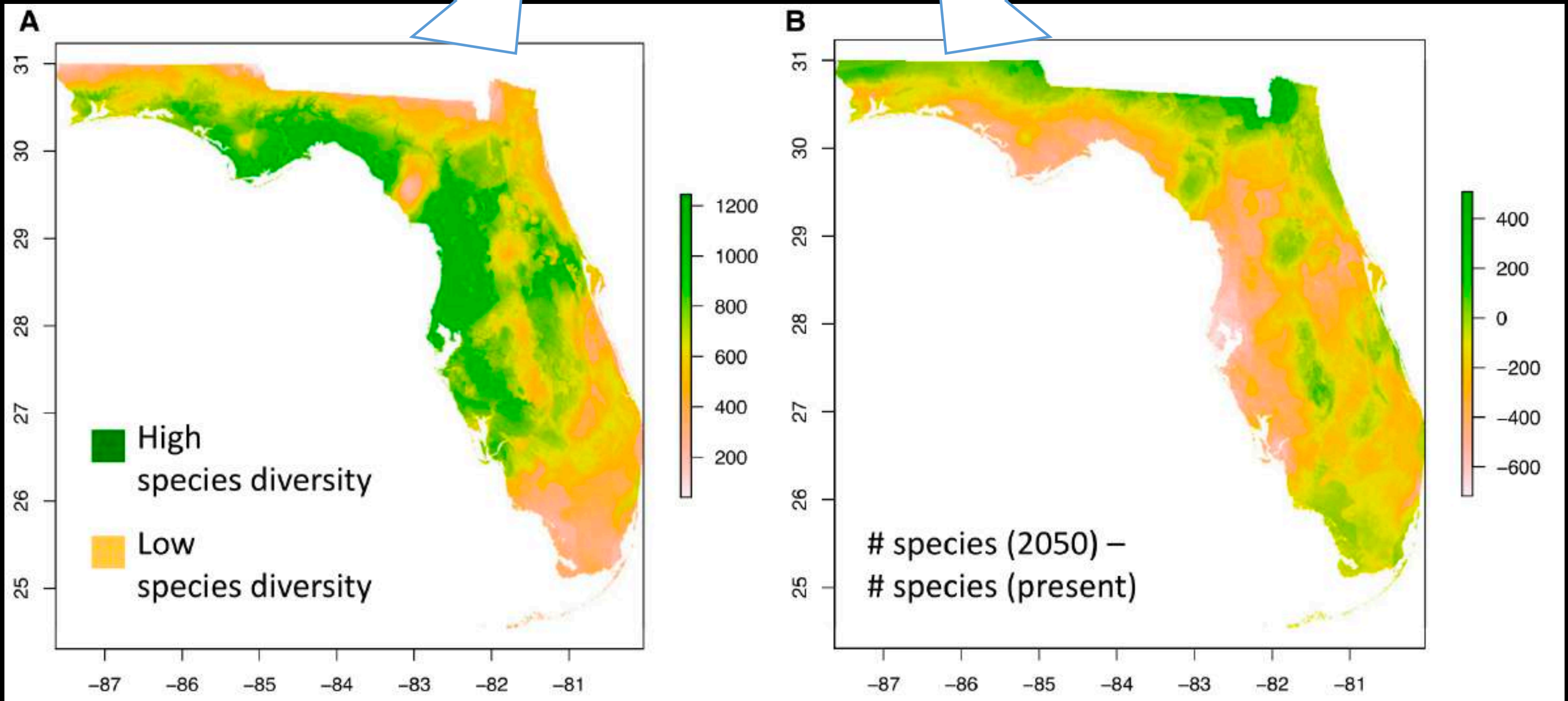
Vascular Plant Diversity of Florida— Ecological Niche Modeling with 1500 species of Florida Plants, based on >500,000 specimens

iDigBio
Integrated Digitized Biocollections

WorldClim - Global Climate Data

Free climate data for ecological modeling and GIS

Soltis P. 2017,
American Journal of Botany



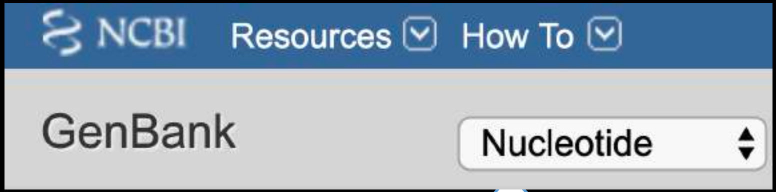
Specimen Data linked with Climate and Genetic Data

WorldClim - Global Climate Data

Free climate data for ecological modeling and GIS

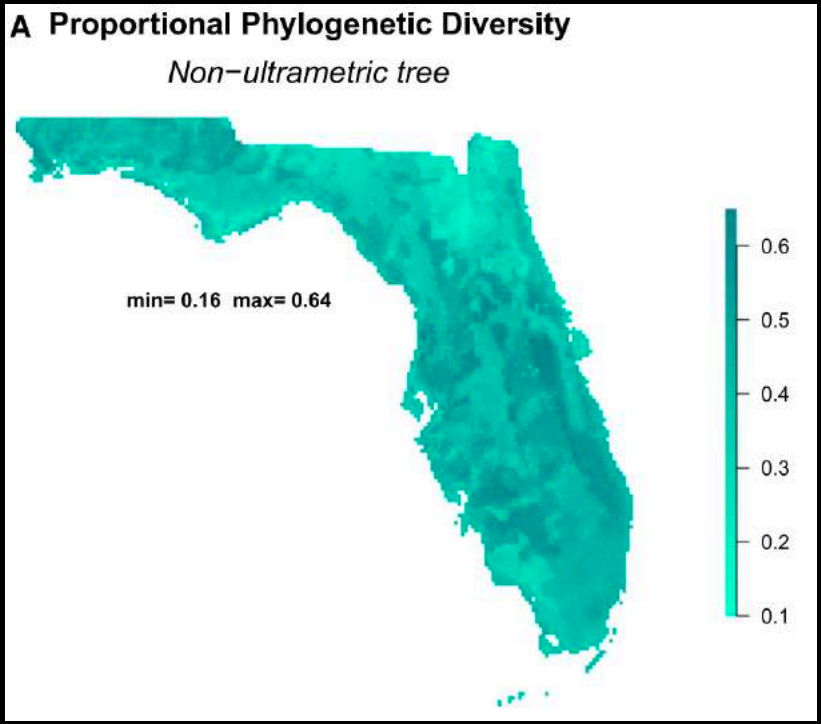
iDigBio

Integrated Digitized Biocollections



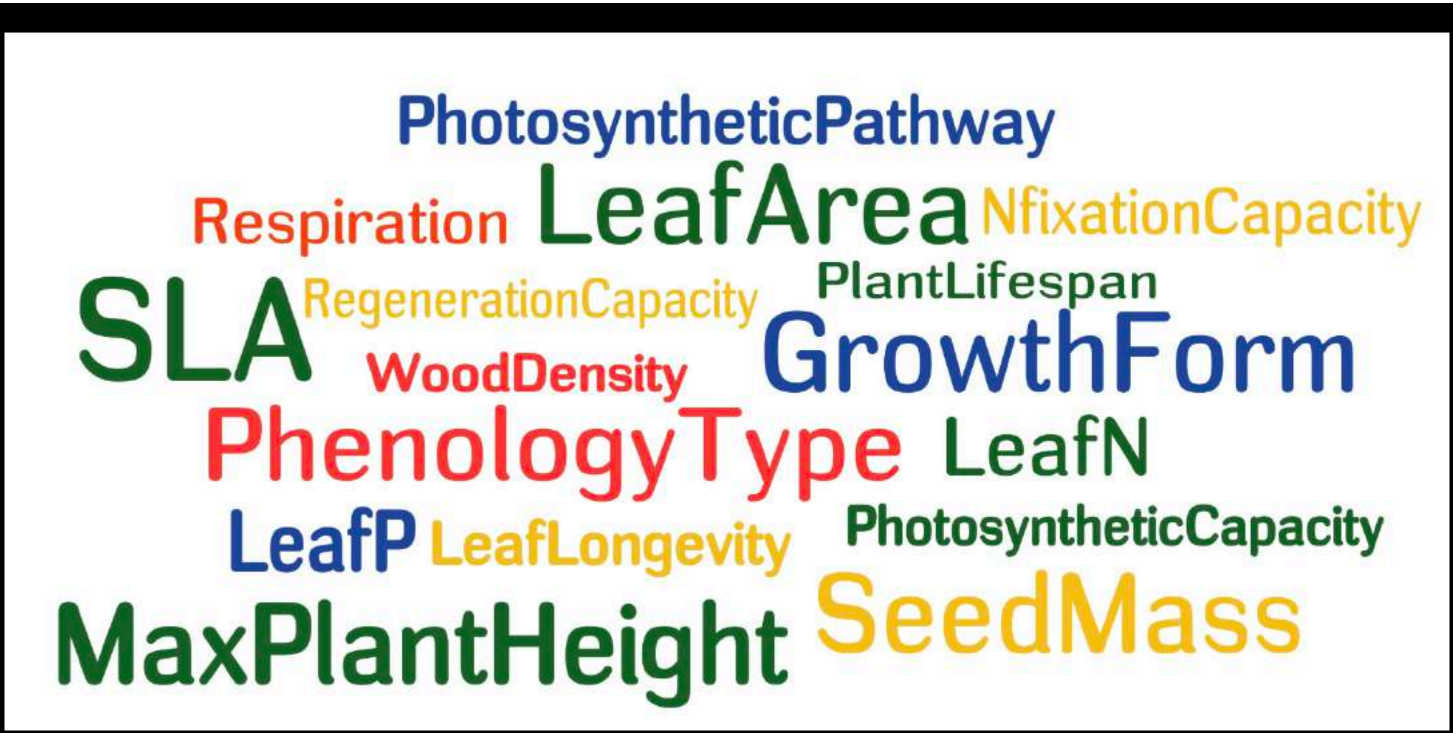
GenBank

Nucleotide



Soltis P. 2017,
American Journal of Botany

Specimen Data Linked with Functional Trait Data



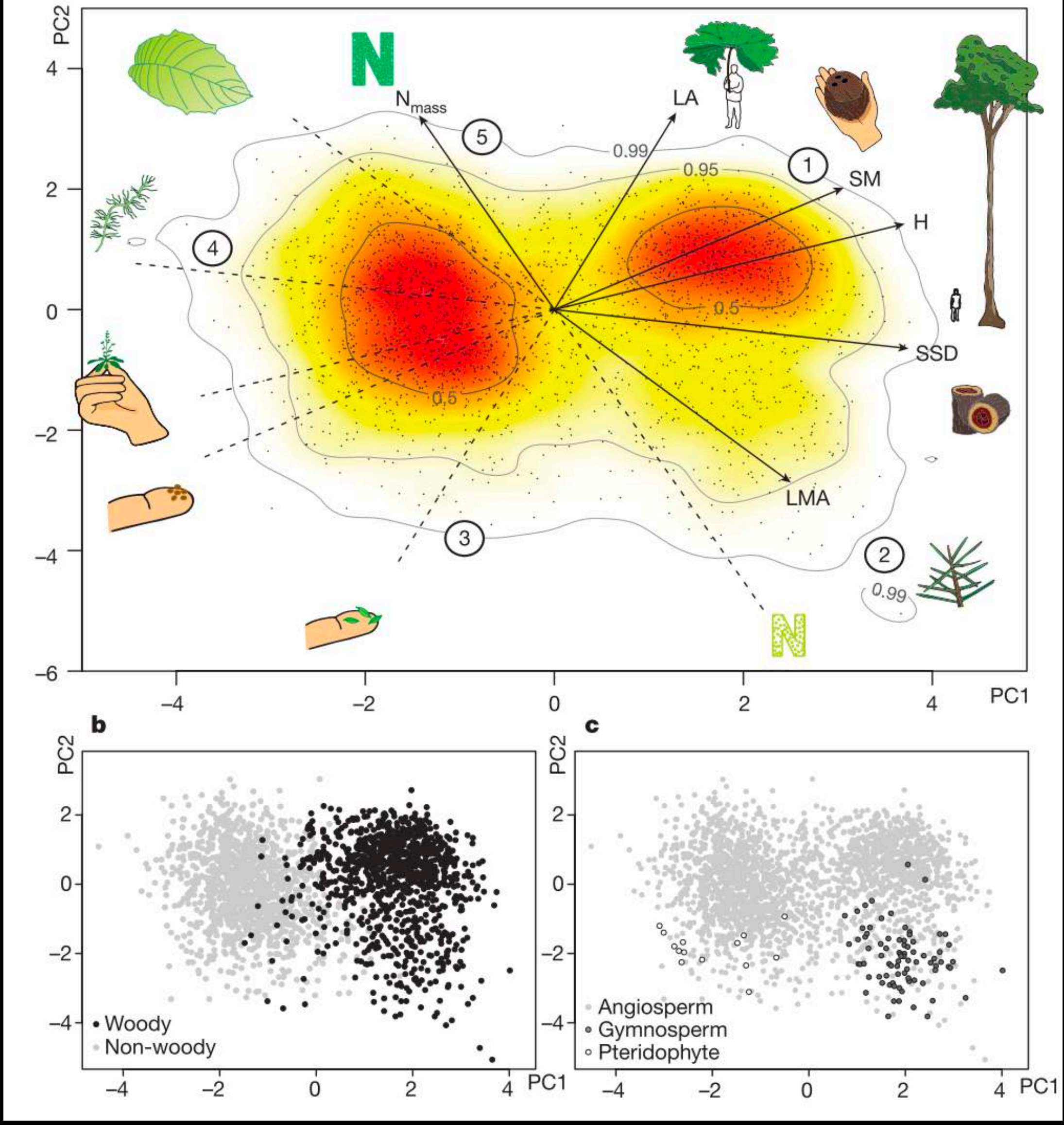
www.try-db.org

Specimen Data Linked with Functional Trait Data

Diaz et al. Nature (2016)

Data from 46,085 plant species from 423 families

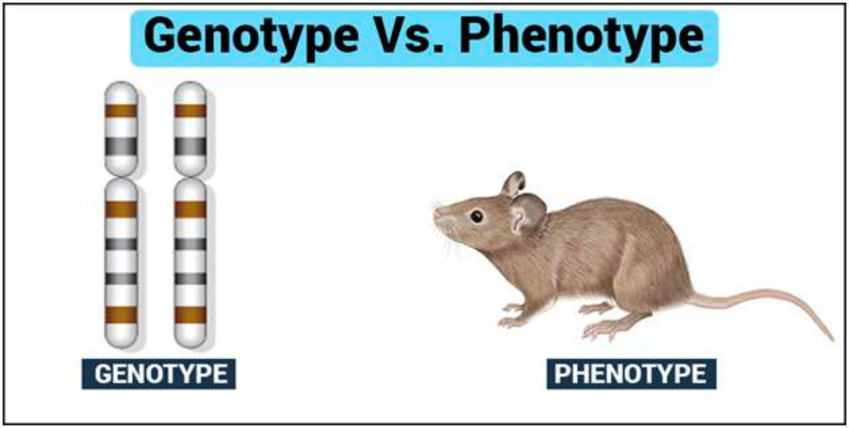
Reveals “hot spots” in the trait-space of vascular plants— some combinations of traits have evolved repeatedly across many different clades



Potential for AI— Phenotype Data

Big data in genomics, ecology (data layers), geography… but what about phenotype?

Phenotype-- observable traits above a molecular level (anatomy, morphology, behavior)



Potential for AI— Phenotype Data

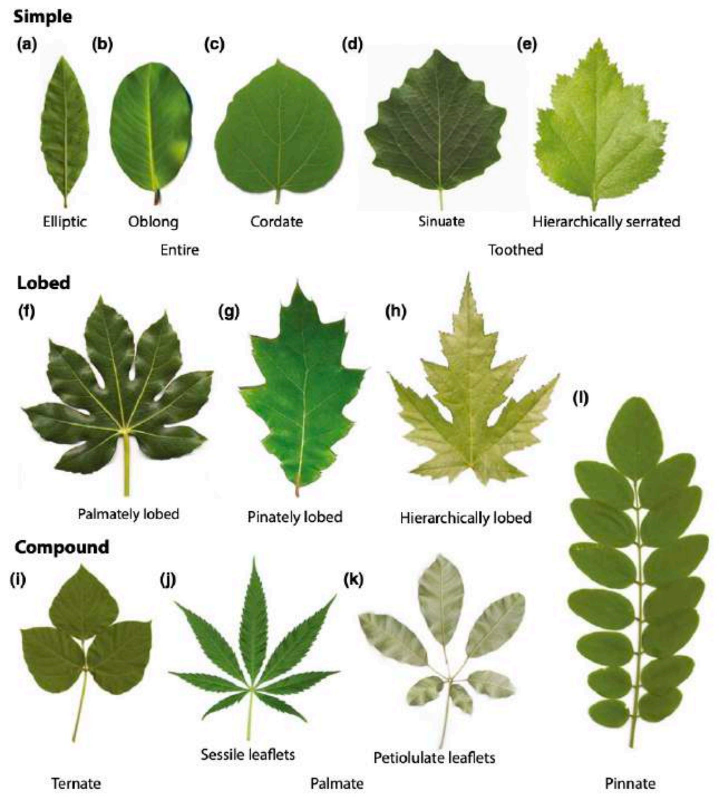
The phenotype bottleneck:

“While phenotype data are as complex, diverse, and nuanced as genomic data, they have not seen data standardization and analyses applied with the same broad strokes as we have seen for genomics.”

–Deans et al. 2015, PLOS Biology

Potential for AI— Phenotype Data

In order to understand the evolution and ecology of biodiversity, it’s necessary to develop ways of quantifying phenotypic diversity in a way that is comparable across species.



Runions, Tsiantis, & Prusinkiewicz
2018, New Phytologist

Why is leaf shape interesting?

Research at the interface of Ecology-Evolution-Development (Eco-Evo-Devo)

- Leaf shape linked to functional diversity, physiology

  eg. dissection, toothineess of leaf linked to thermoregulation of leaves and water balance

- Evolutionary constraints and patterns

- Ecological constraints and patterns (eg. habitat filtering)

- Plasticity and adaptability

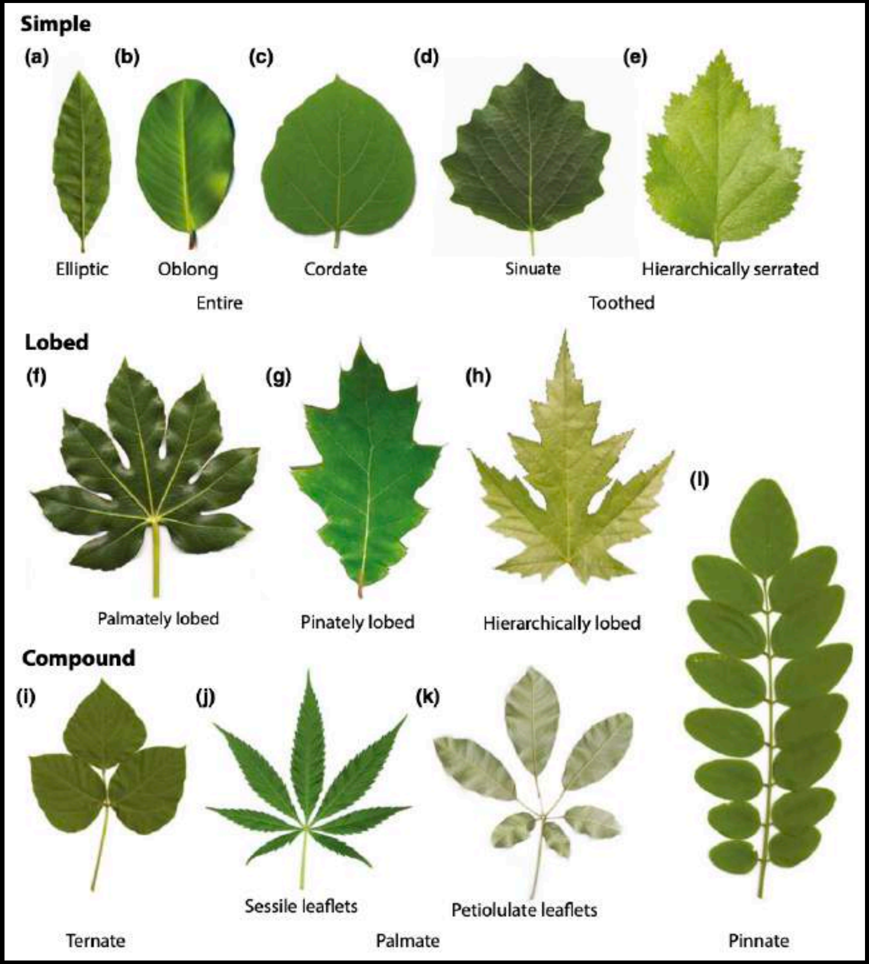
- Developmental and genetic mechanisms of leaf shape variability largely understood for many plant groups.

AI and Images— Future Research Potential

Historically, botanists have described leaf shape in qualitative terms: lobed, compound, pinnate, etc.

Challenging to quantify shape:

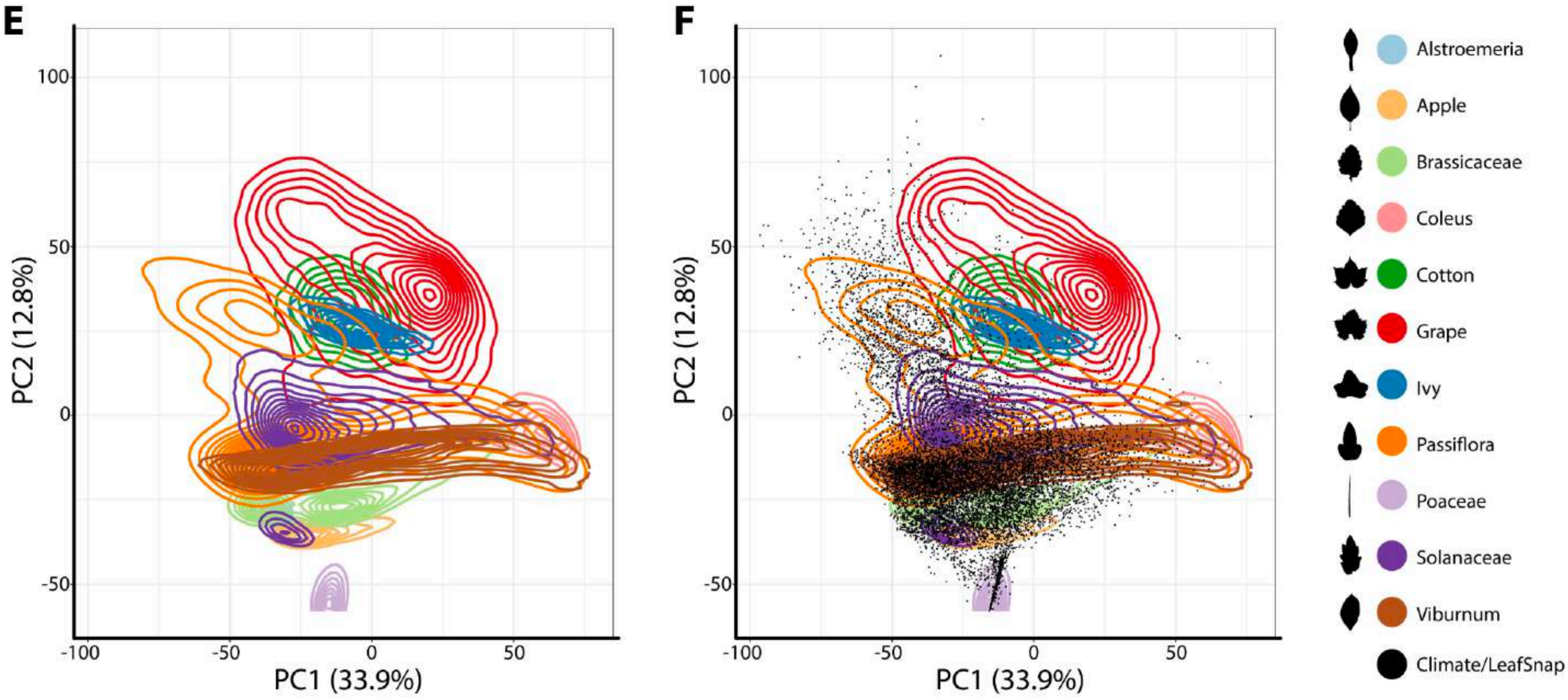
Traditional land-mark based morphometric methods are only useful for closely related plants.



Runions, Tsiantis, & Prusinkiewicz 2018, New Phytologist

Advances in quantifying leaf shape

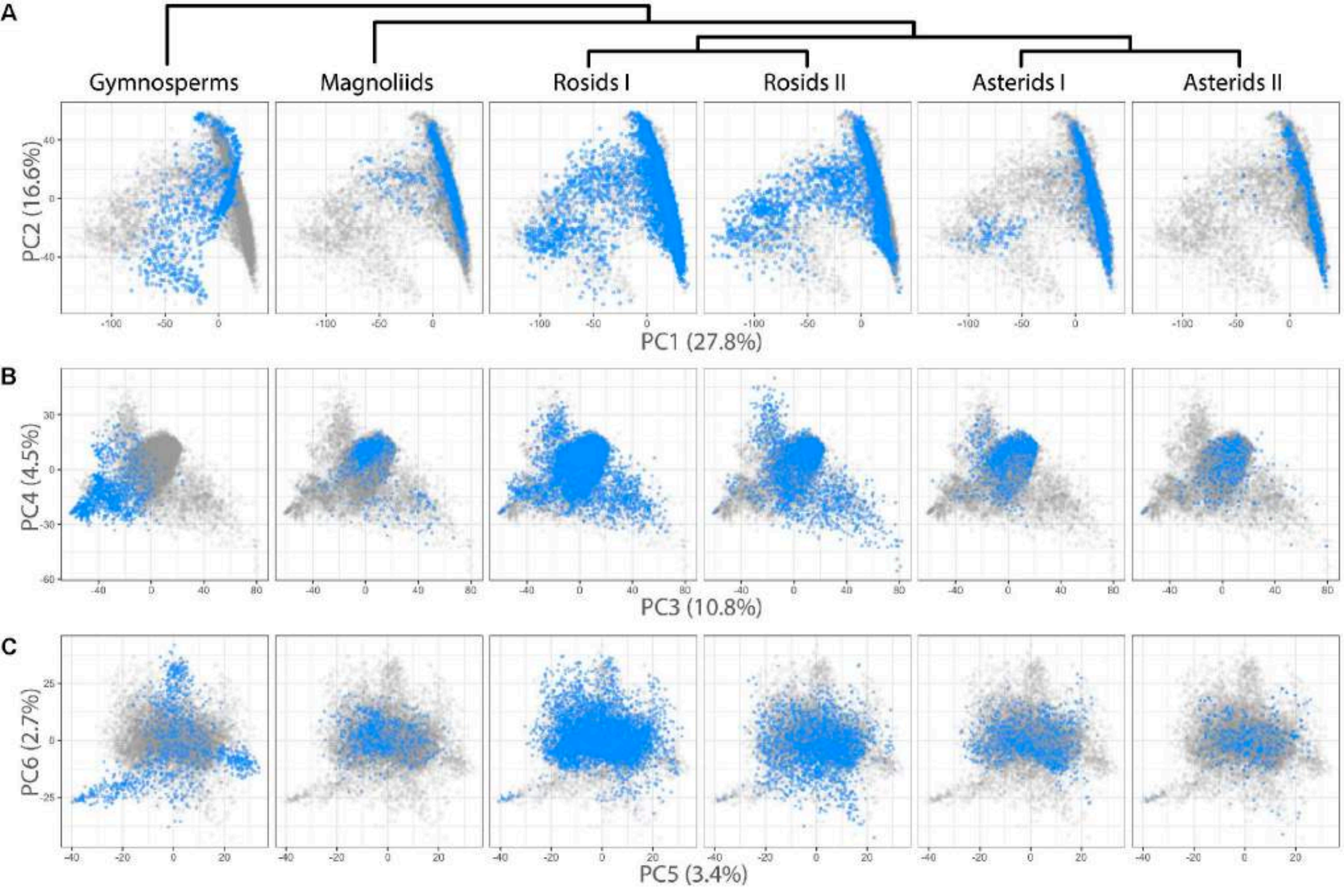
Comparing the morphospace of leaves in different clades using persistent homology



Li et al. 2018, Frontiers in Plant Science

Advances in quantifying leaf shape

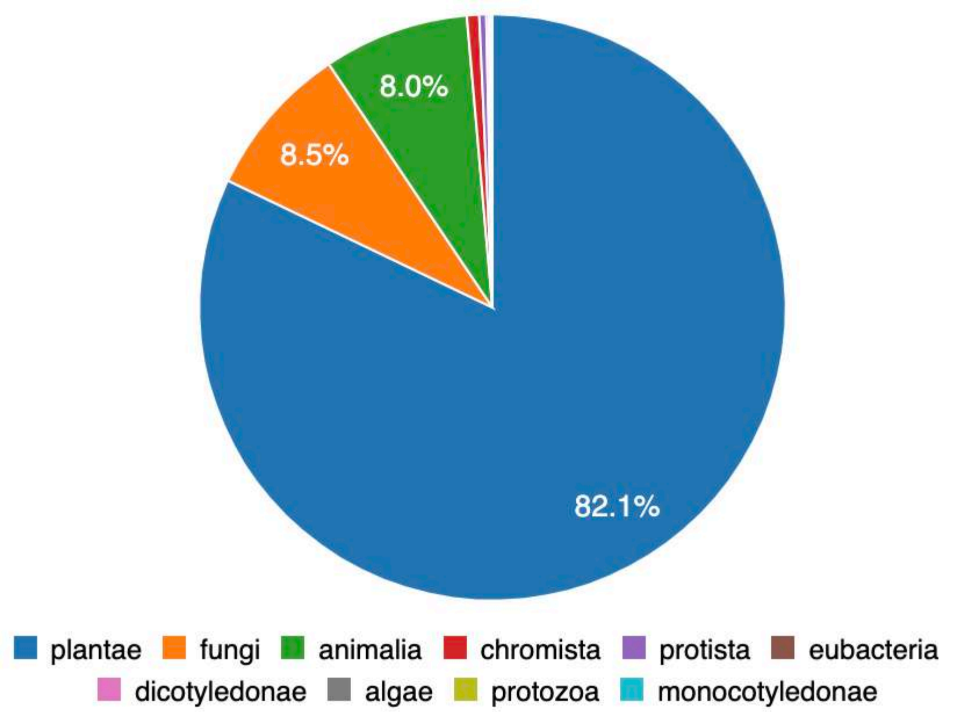
Comparing the morphospace of leaves in different clades using persistent homology— linking shape and phylogeny



Li et al. 2018, Frontiers in Plant Science

Potential for AI— Phenotype Data

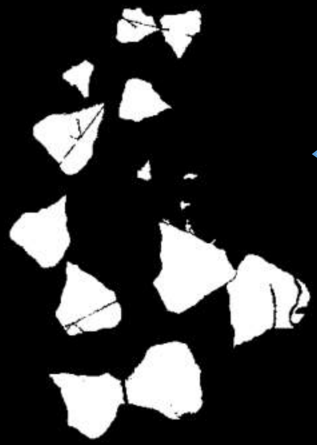
The BIG question: Can AI unlock natural history collections as a resource for phenotype data?

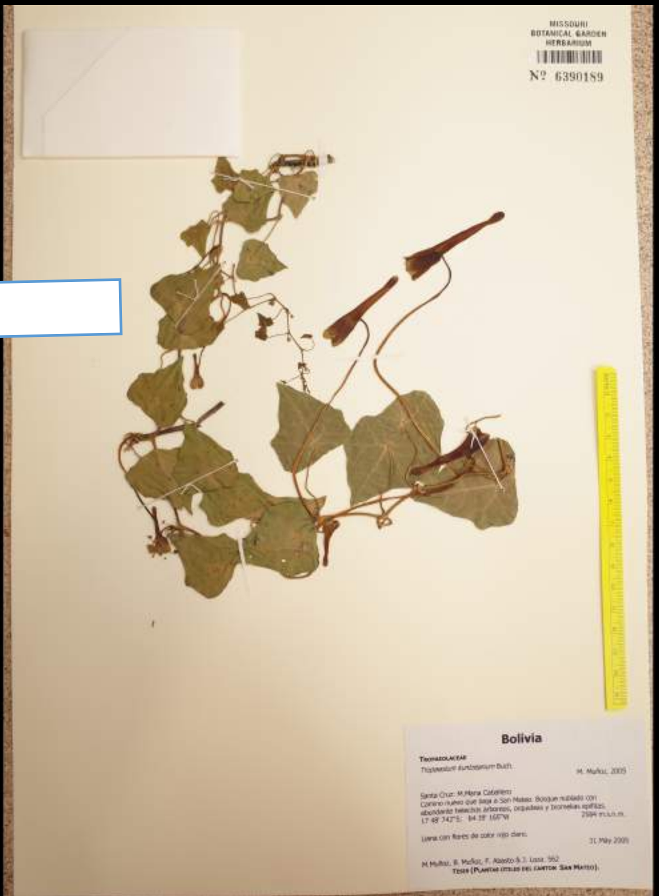


29,523,938 Media Records

Potential for AI— Phenotype Data

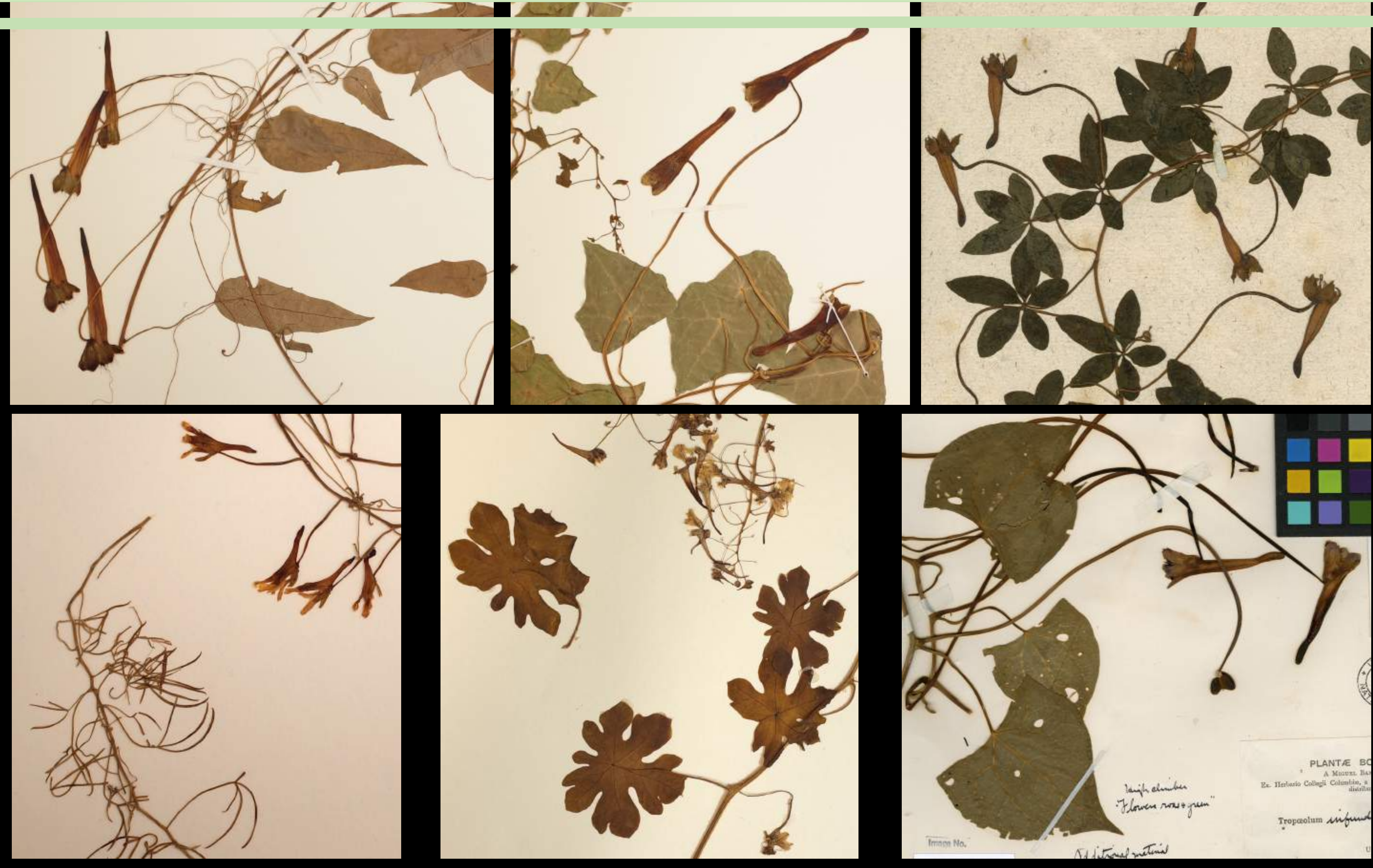
The BIG question: Can AI unlock natural history collections as a resource for phenotype data?



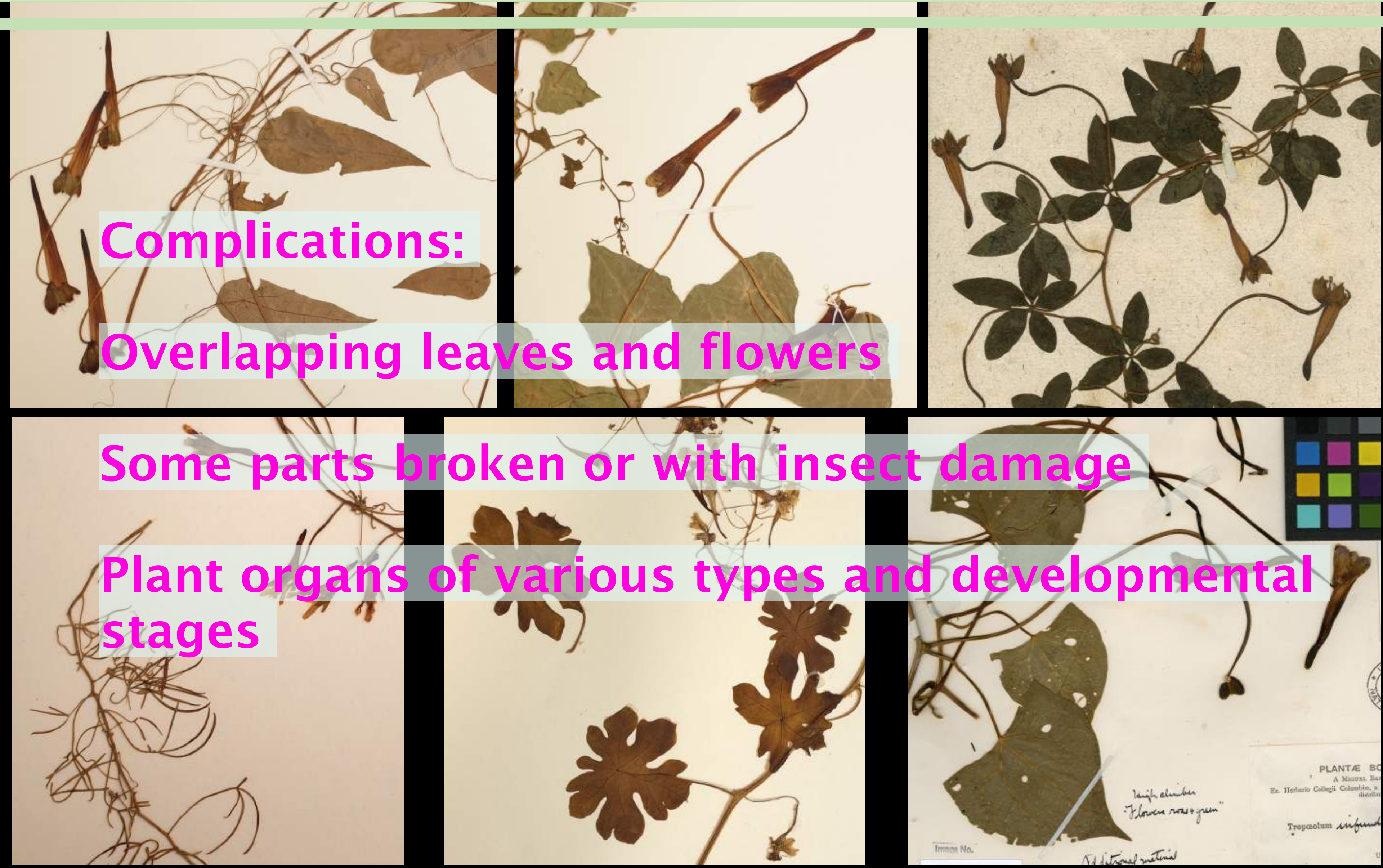


A single image contains a great deal of information about the shape and size of plant organs— could AI methods (eg. semantic segmentation) help extract and quantify this?

Leaf and flower diversity in the nasturtiums (Tropaeolaceae)



Leaf and flower diversity in the nasturtiums (Tropaeolaceae)



Complications:

Overlapping leaves and flowers

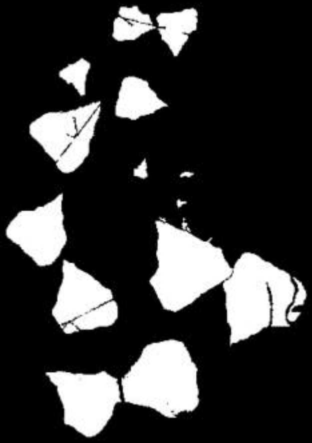
Some parts broken or with insect damage

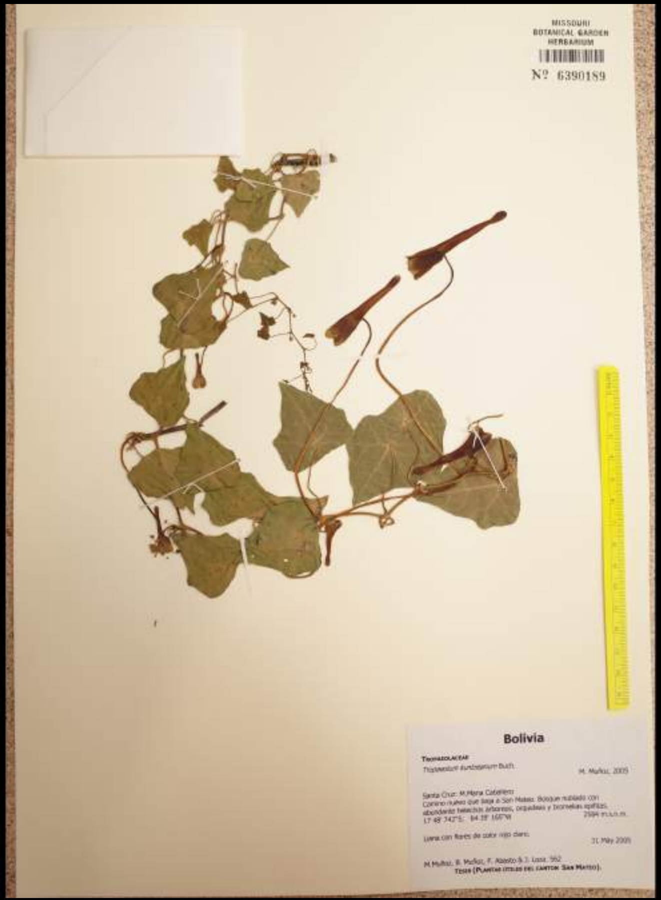
Plant organs of various types and developmental stages

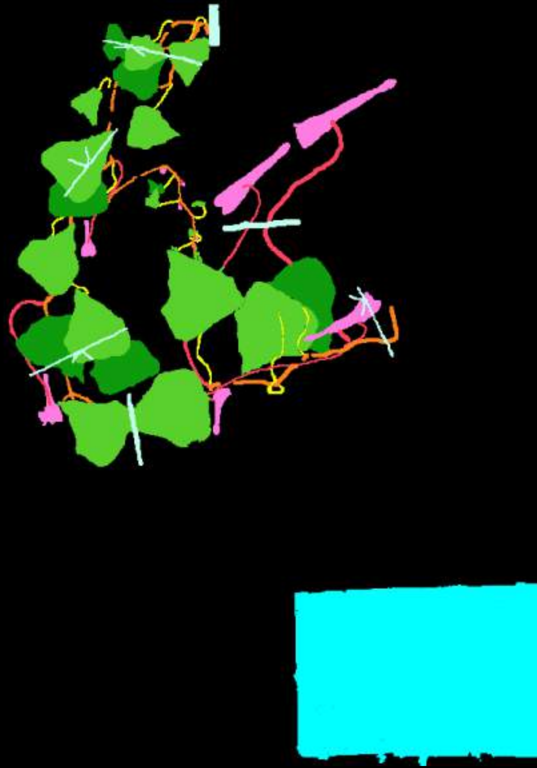
Building a training data-set for semantic segmentation

The Challenge: Labeling individual pixels is exceptionally time consuming….

RectLabel





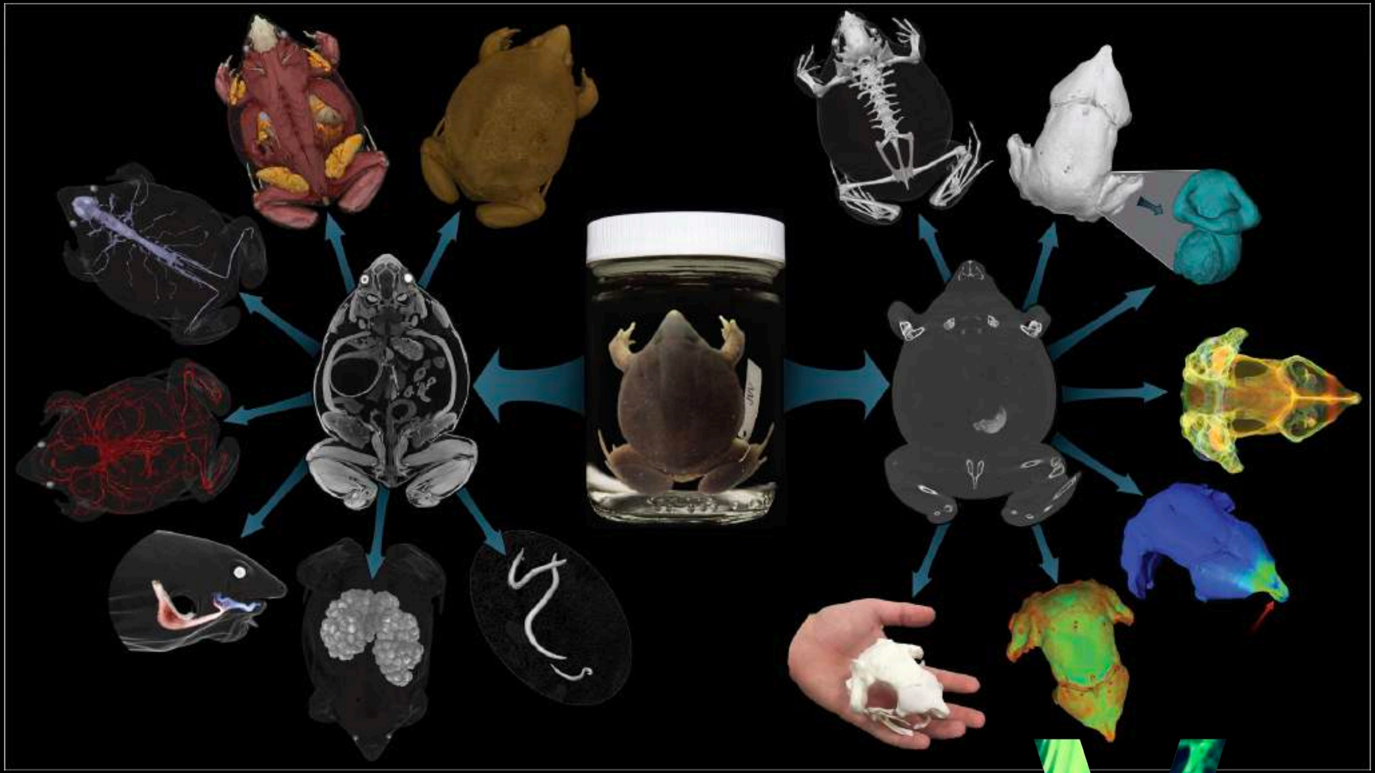


Some Challenges: AI and Trait Data

- Vast diversity of the structure of organisms— need massive training data sets?
- Biases in specimen collection
- Images: Overlapping or damaged structures.
- Some characters too small or hidden.
- Changes between fresh and dry plants?
- Data quality— the balance between the abundance of data vs imperfect data

Vertebrate Phenotype Data– oVert

oVert (iDigBio TCN). CT scans of vertebrate specimens, freely available through Morphosource.



3D printing for research, education, and outreach

oVert

https://www.floridamuseum.ufl.edu/overt/

Acknowledgments

Pamela Soltis

iDigBio
Integrated Digitized Biocollections

FLORIDA MUSEUM

UF UNIVERSITY of FLORIDA

NSF

PRAGMA 36
colocated with CENTRA 4

Thank you! Questions?

Contact:
Annika Smith
annikals@ufl.edu
🐦 @meristemming

iDigBio
Integrated Digitized Biocollections

FLORIDA MUSEUM

UF UNIVERSITY of FLORIDA