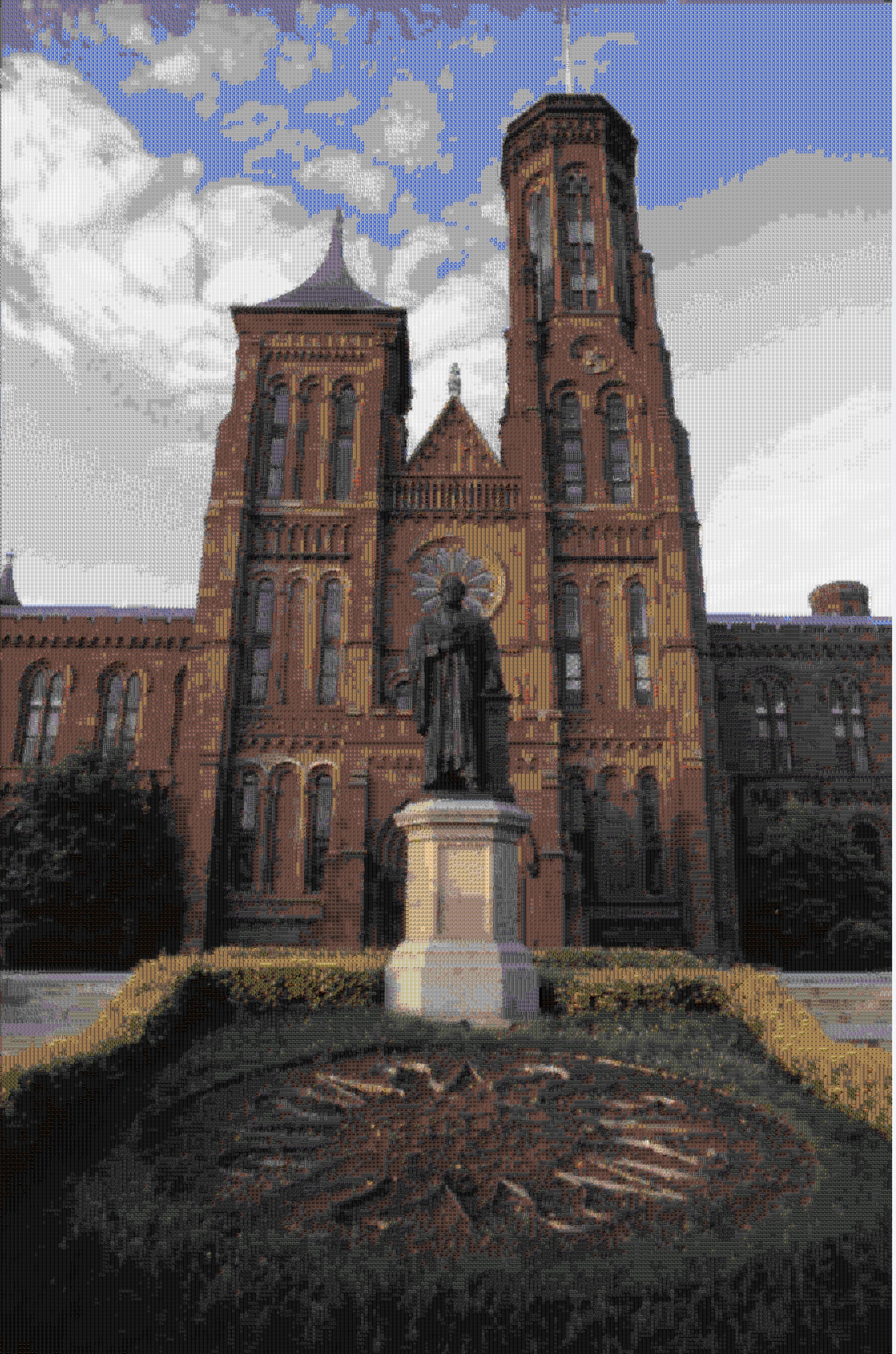


Data-intensive approaches to digitized museum collections

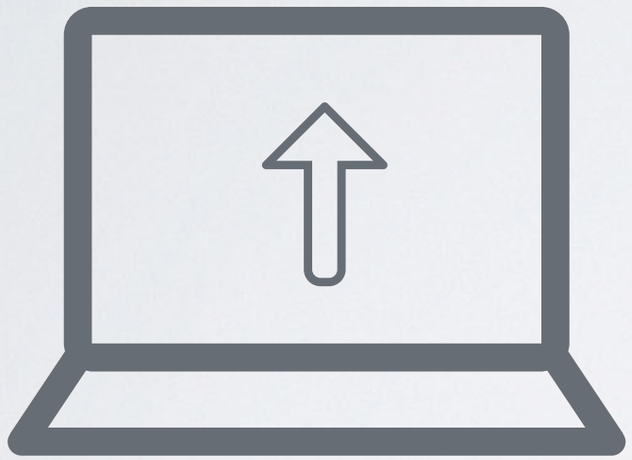
Rebecca B. Dikow
Data Science Lab
Office of the CIO
Smithsonian Institution

DATA SCIENCE LAB

datascience.si.edu
@SIDataScience

Smithsonian

A diversity of locations



Heterogeneous digital data

Lack of purpose-built software tools

Smithsonian

19 museums, 9 research centers, and a zoo

Smithsonian Collections Holdings

155.5M Objects and Specimens
163.3K Archival Cubic Feet
2.2M Library Volumes

Smithsonian

Smithsonian Collections Digitization

32M Objects and Specimens with Digital Record
125K Archival Cubic Feet with Digital Record
1.5M Library Volumes with Digital Record
4.9M Objects and Specimens with Digital Image
56.9K Library Volumes with Digital Image

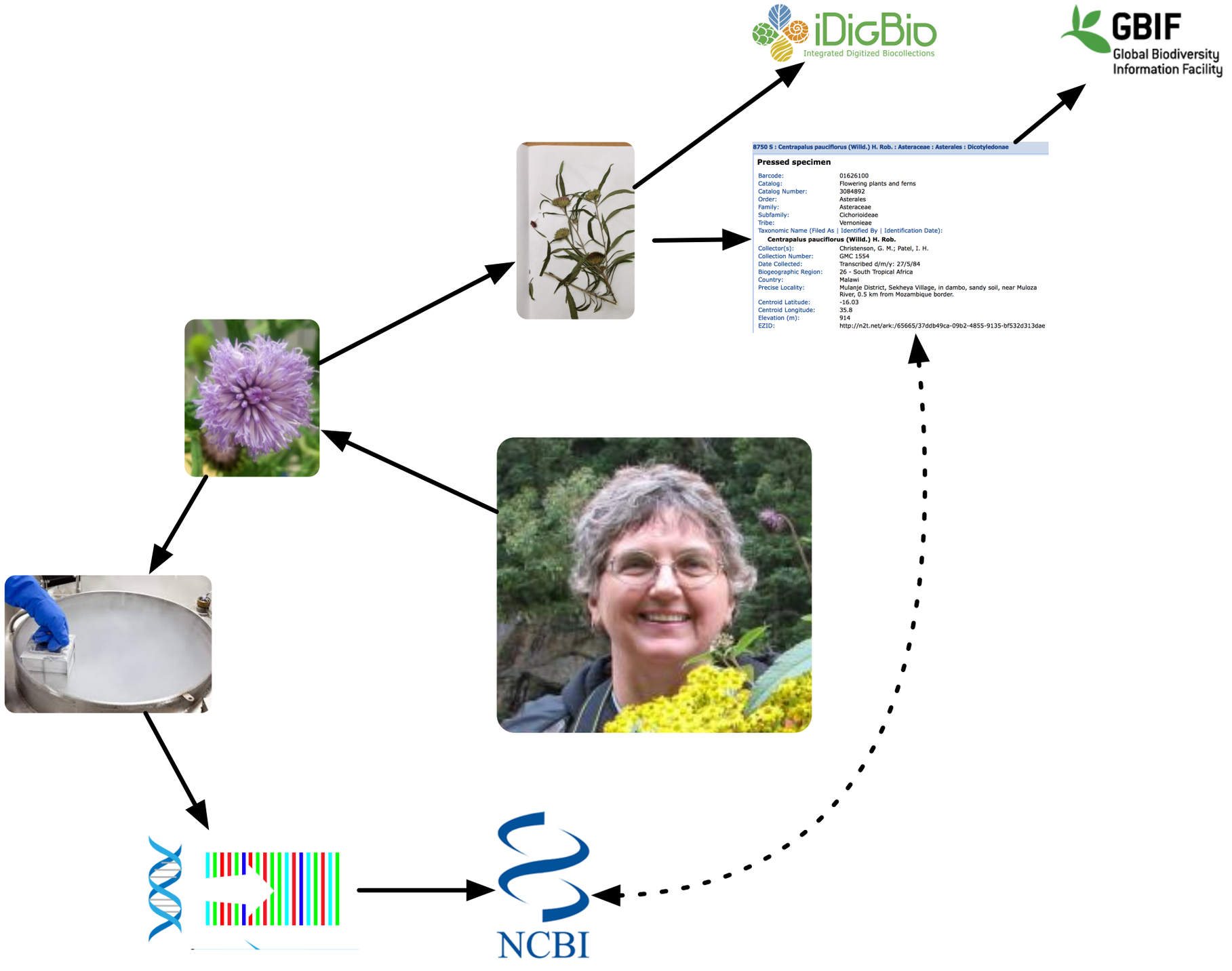
Digitized collections

photos
taxonomic names
specimen records
genomic sequences
geo-referenced localities

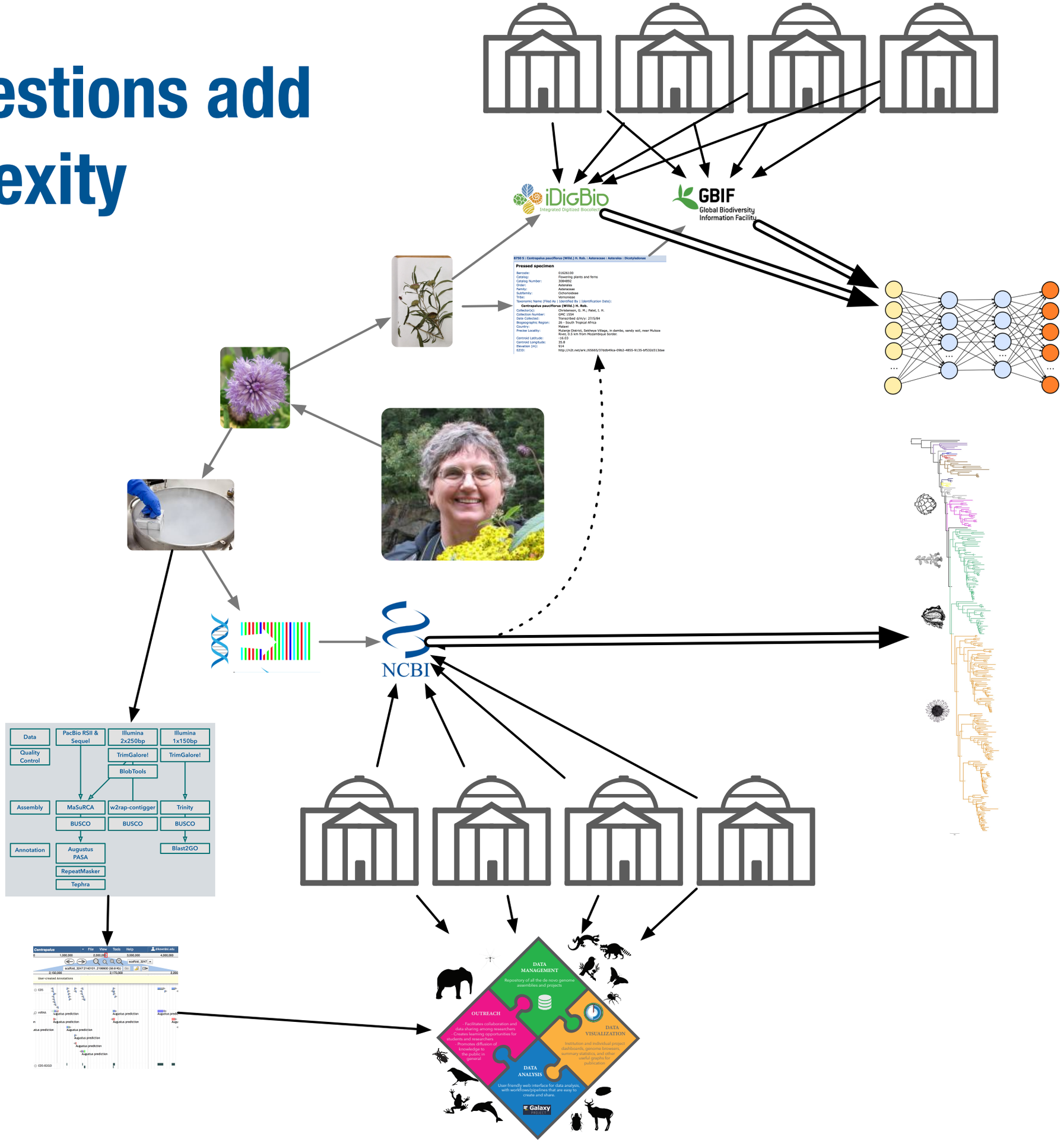
Apiocera pica voucher USNM:ENT:00914599 cytochrome oxidase subunit 1 (COI) mitochondrial
658 bp linear DNA
Accession: KT733539.1 GI: 931147206
BioProject Protein Taxonomy
GenBank FASTA Graphics

field books
illustrations
observations
scientific publications
taxonomic descriptions

Specimen collecting event → digital data



Research questions add complexity

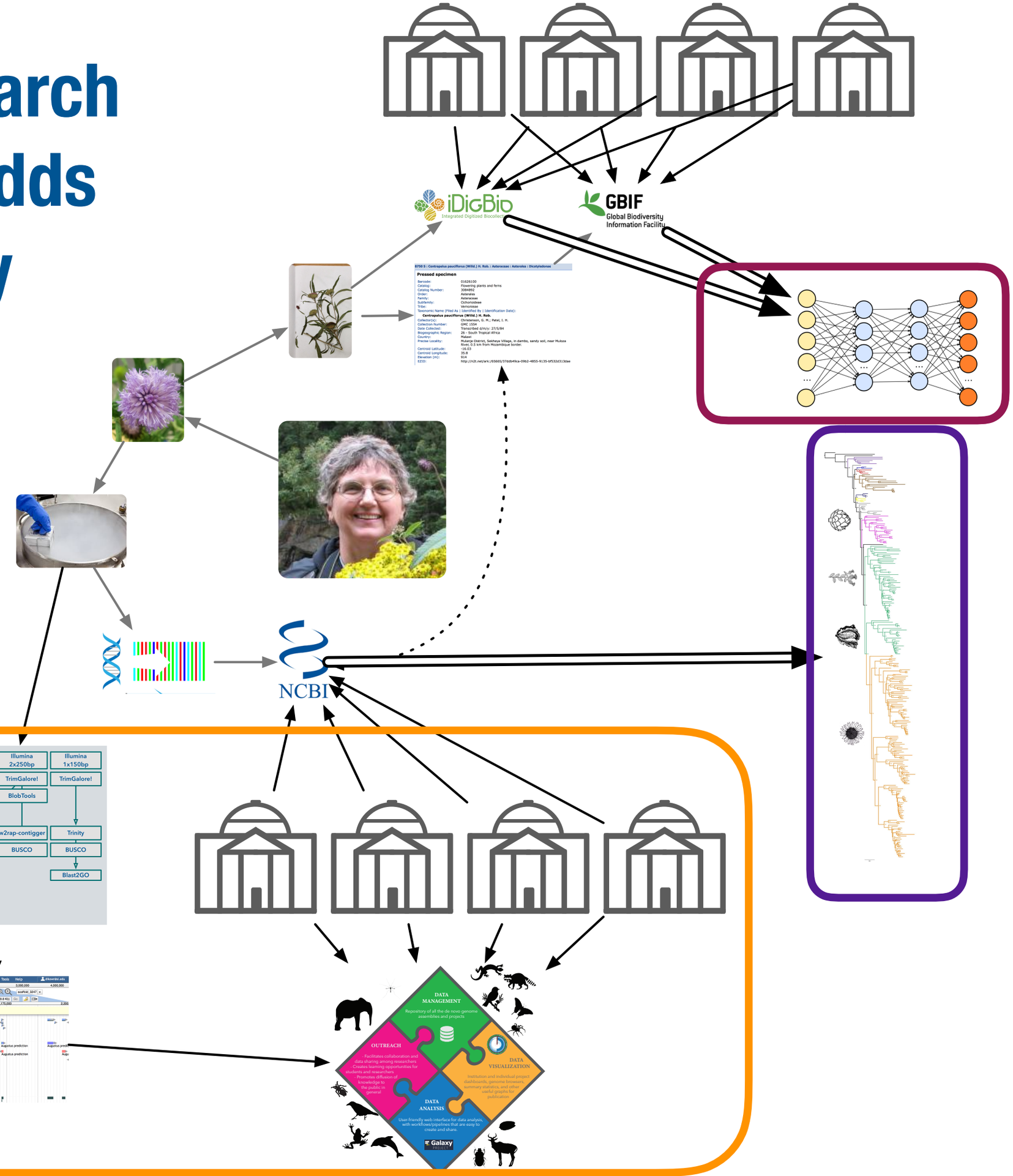


Adding a research component adds complexity

deep learning

phylogeny

genomics



Smithsonian Biodiversity Genome Hub: under construction

DATA MANAGEMENT

Repository of all the de novo genome assemblies and projects

OUTREACH

- Facilitates collaboration and data sharing among researchers
- Creates learning opportunities for students and researchers
- Promotes diffusion of knowledge to the public in general

DATA VISUALIZATION

Institution and individual project dashboards, genome browsers, summary statistics, and other useful graphs for publication

DATA ANALYSIS

User-friendly web interface for data analysis, with workflows/pipelines that are easy to create and share.

Galaxy PROJECT

Botany

DATA SCIENCE LAB

Smithsonian Institution

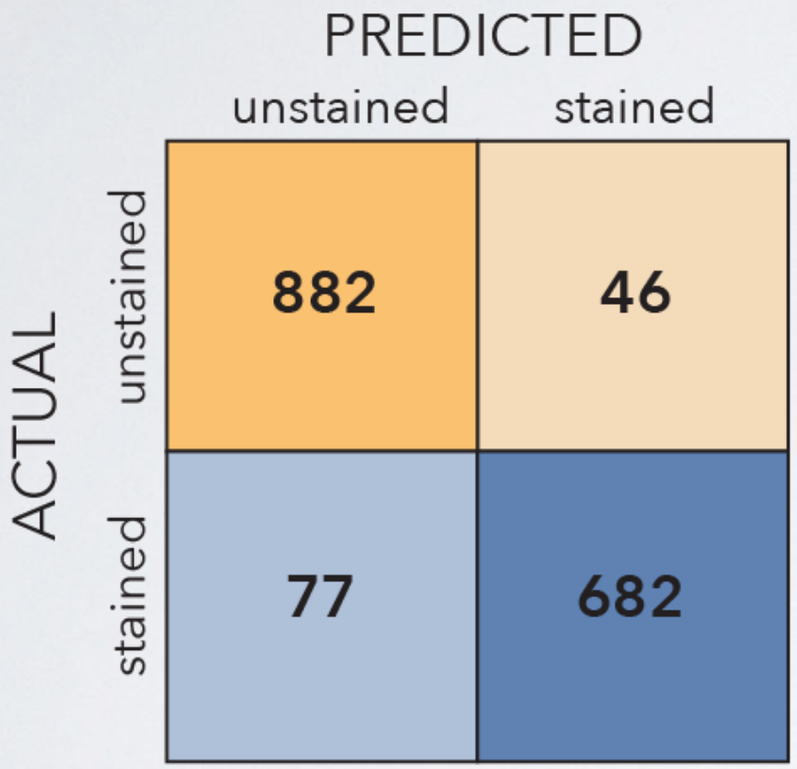
NATIONAL MUSEUM of NATURAL HISTORY

NVIDIA

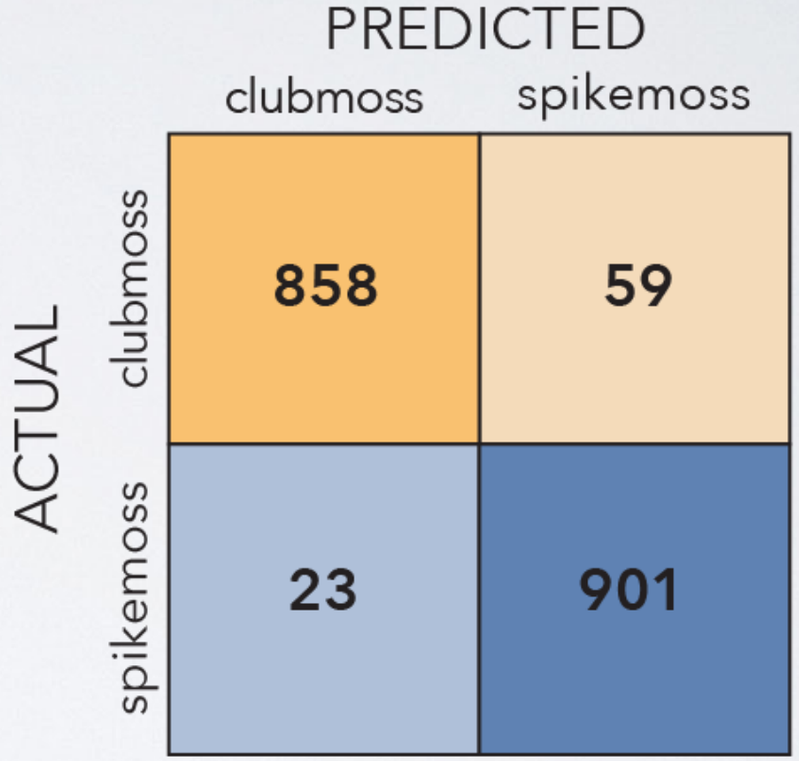
Smithsonian DIGITIZATION

First pilot projects: detecting mercury staining and family ID

Mercury staining



Family ID



Bumblebees


Racine Co., Wis.
No. 4241
Coll. L.W. Macior
MAY 12, 1965
Walworth Co., Wis.
No. 2626
Coll. L.W. Macior
AUG 6, 1964
Racine Co., Wis.
Coll. L.W. Macior
Bombus bifarius
bifarius
Cresson
det. R.B. Miller 1967

Crowdsourced transcription

transcription.si.edu

Smithsonian

HOME PROJECTS SEARCH ABOUT TIPS NEWS

SMITHSONIAN DIGITAL VOLUNTEERS: TRANSCRIPTION CENTER

SIGNUP LOGIN

HOME > PROJECTS > NMNH - DEPARTMENT OF ENTOMOLOGY

THE BUMBLEBEE PROJECT - SET 2

About the Project

Please help us create digital records for the United States National Entomological Collection! We will be transcribing the labels of specimens of bumblebees. Bumblebees are found in the Bombus genus (Hymenoptera: Apidae). They are social insects that feed on nectar and collect pollen to feed their young. Bumblebees are very important pollinators! Learn how to transcribe this project and get started.

The digitization of this project has been made possible with the generous support of Pixel Acuity, LLC. Please contact Jessica Bird (birdj@si.edu), Department of Entomology, for any questions or comments about the transcriptions and thanks to all of you for your help!

Read More

Download PDF | Go to Page Go

feedback

Completed!

Project Progress (details)

23 CONTRIBUTING MEMBERS

366 TOTAL PAGES

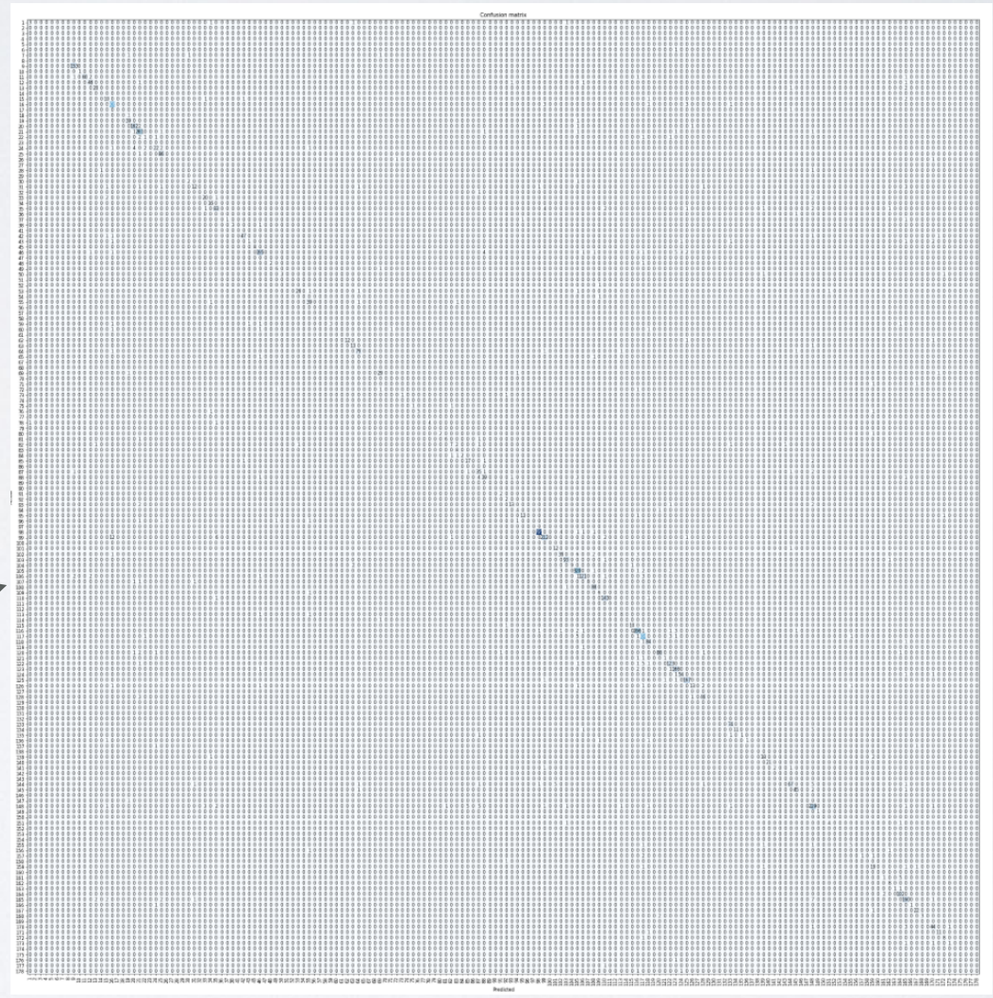
Models in progress

Training data: 33,347 images with subgenus/species labels

We also have images for >10,000 unidentified specimens

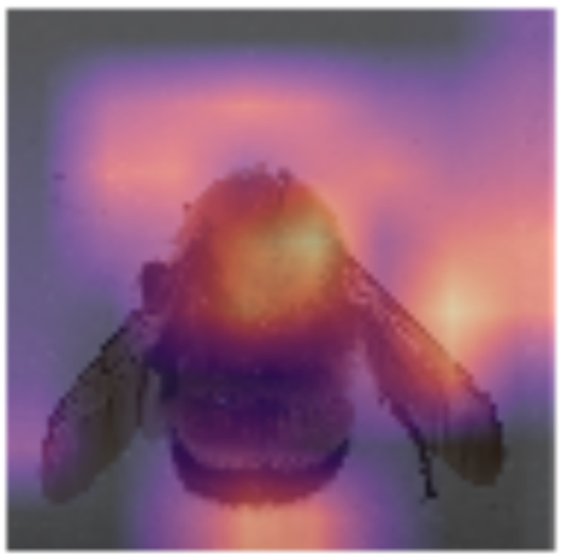
Subgenus model:
15 classes
Overall accuracy: 93.8 %

Species model:
178 classes
Overall accuracy: 92.5%



Models are built in PyTorch using fastai (https://github.com/fastai/fastai).
The backbone is a 101 layer deep residual convolutional network (ResNet-101; He et al., 2015)

Activation heat maps allow us to start exploring the models in more detail.



How can we broaden this work to Smithsonian history, art, and culture digital collections and archives?

Morse Daguerreotype Camera

Sound Box	86%
Audio Equipment	79%
Electronics	79%
Loudspeaker	75%
Wood	74%
Electronic Instrument	63%
Computer Speaker	59%
Antique	58%
Subwoofer	56%

Archives

UNITED STATES HOLOCAUST MEMORIAL MUSEUM

BECAUSE OF HER STORY

SMITHSONIAN AMERICAN WOMEN'S HISTORY INITIATIVE

WOMENSHISTORY.SI.EDU

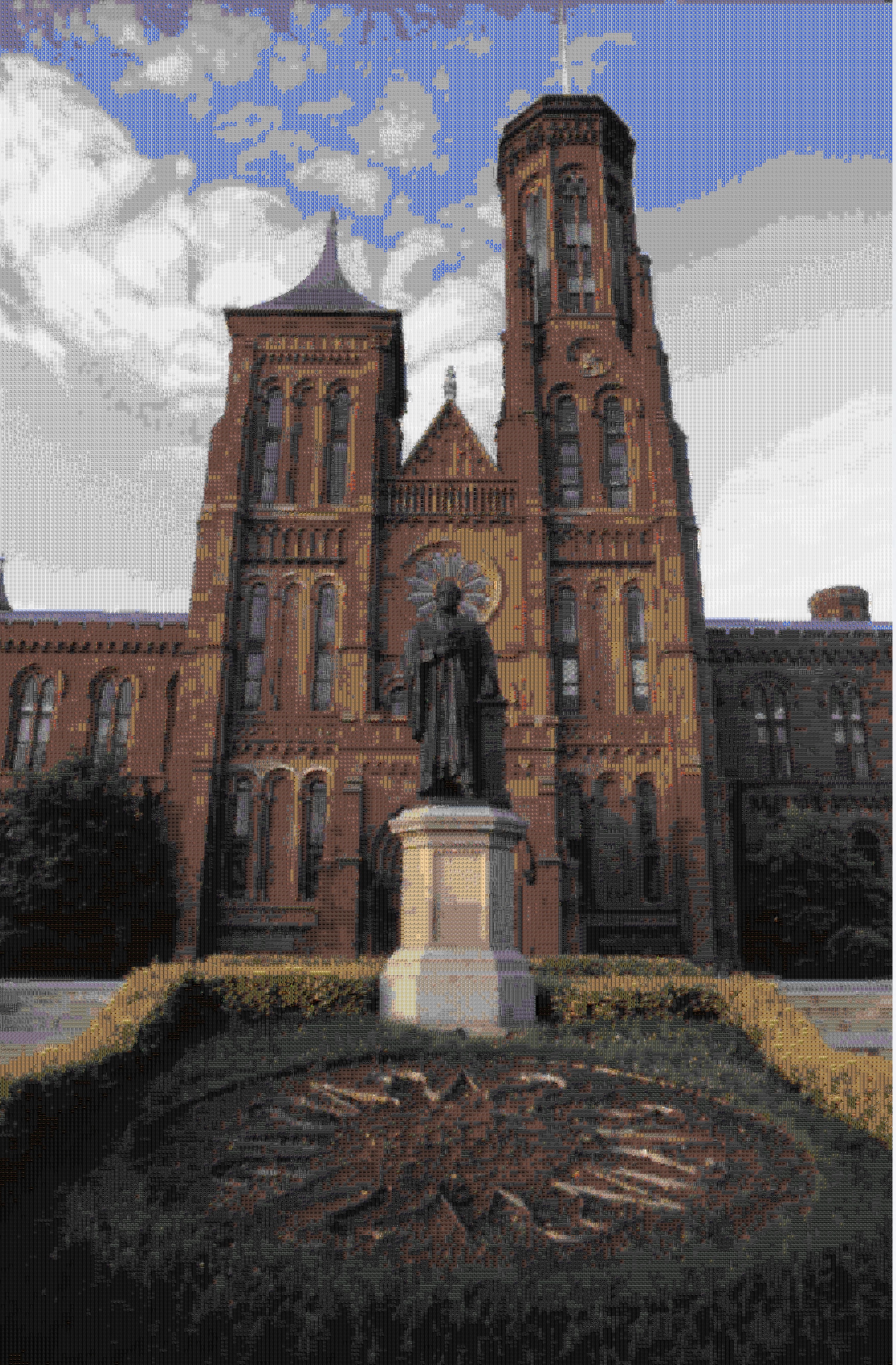
Building capacity across the Smithsonian includes lots of training!

We are building a community of Carpentries instructors across the Smithsonian.

More than 300 Smithsonian researchers have been trained in topics such as Python, R, genome analysis, and data management in the past 3 years.

Workshop materials: github.com/SmithsonianWorkshops

Instructors and schedule: datascience.si.edu/carpentries



Takeaways

We've only just begun!

Right now, every application of machine learning tools is a research project given our diverse, unique, incomplete data.

Let's use these tools to elevate new stories that better represent Smithsonian audiences.

Thank you!

DATA SCIENCE LAB

Data Science Lab:

Mike Trizna
Mirian Tsuchiya
Alex White
Alex Robillard
Maddy Bursell
Alejandro Sanchez

Partners:

NMNH Botany
NMNH Entomology
Smithsonian Conservation Biology Institute
OCIO DPO
OCIO DAMS
Smithsonian Institution Archives
American Women's History Initiative
United States Holocaust Memorial Museum

Funding:

Smithsonian Women's Committee
Smithsonian Office of the Provost
Smithsonian Office of the CIO