



# Data-intensive approaches to digitized museum collections

---

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

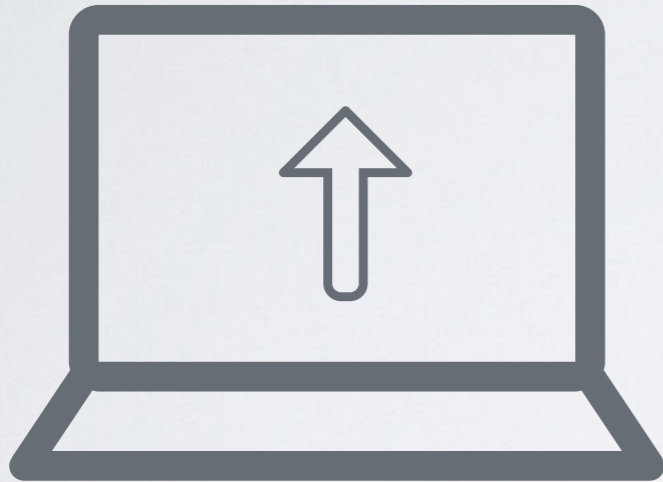


[datascience.si.edu](https://datascience.si.edu)  
[@SIDataScience](https://twitter.com/SIDataScience)





**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

field books  
illustrations  
observations  
scientific publications  
taxonomic descriptions



[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](#)

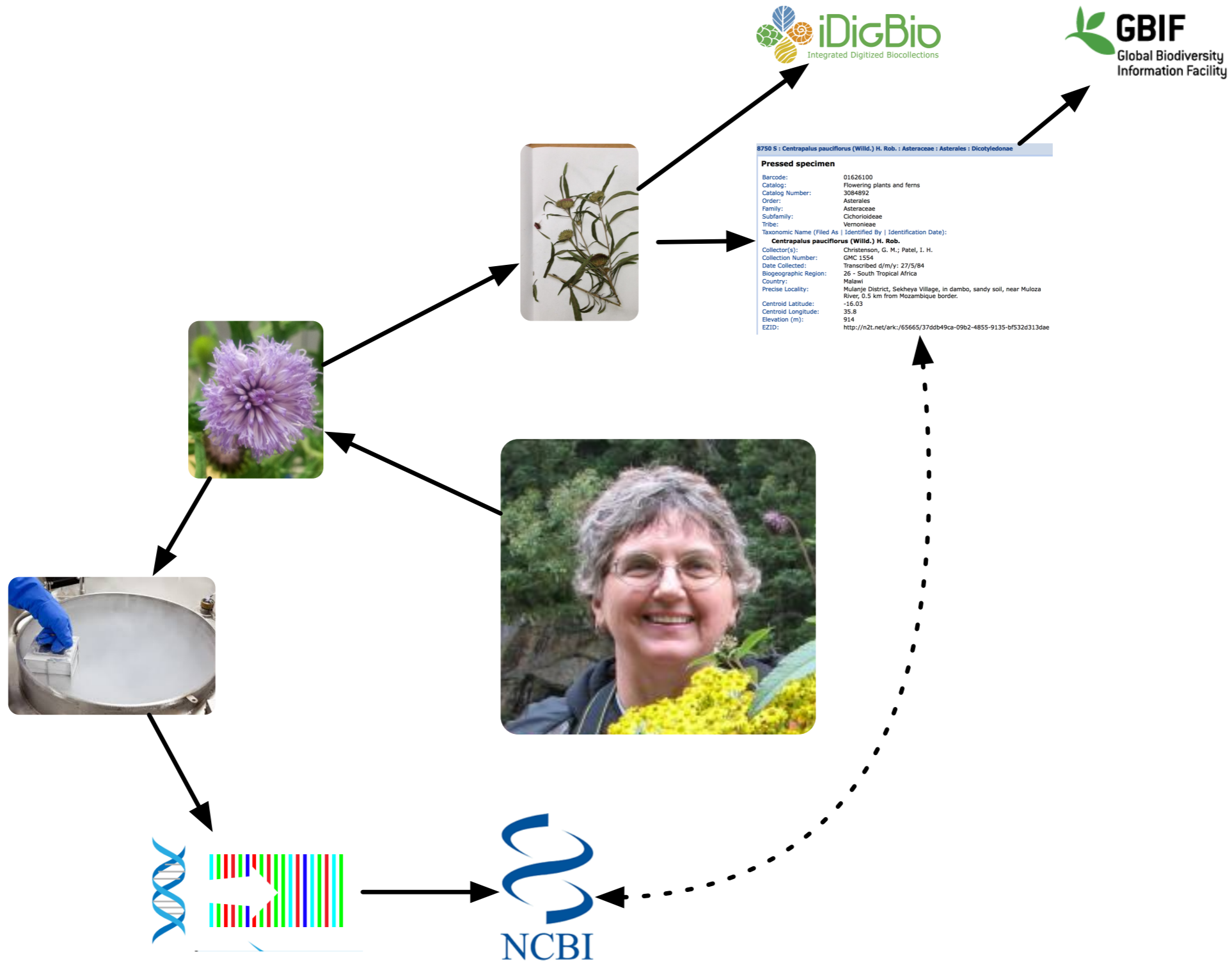
Australia: Western Australia: Wandoo National Park, off Kent Road, 1.6 km S of Deefor Road, *Eucalyptus-Banksia* woodland, on *Verticordia* flowers (Myrtaceae), 32°00'12"S 116°31'43"E, 269 m, 09.xii.2011, T. Dikow J. and F. Hort

USNM ENT

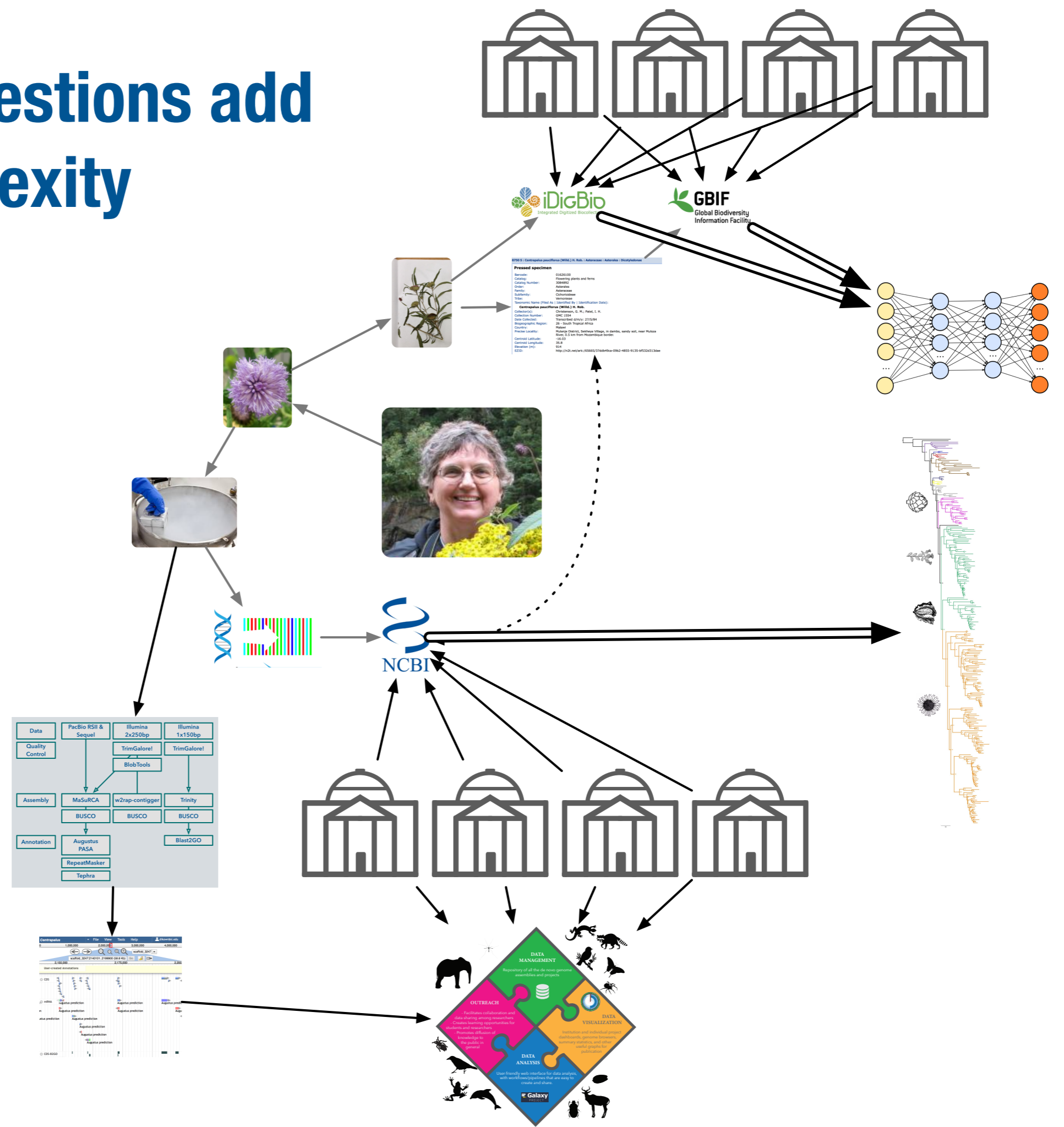


00832111

# Specimen collecting event → digital data



# Research questions add complexity

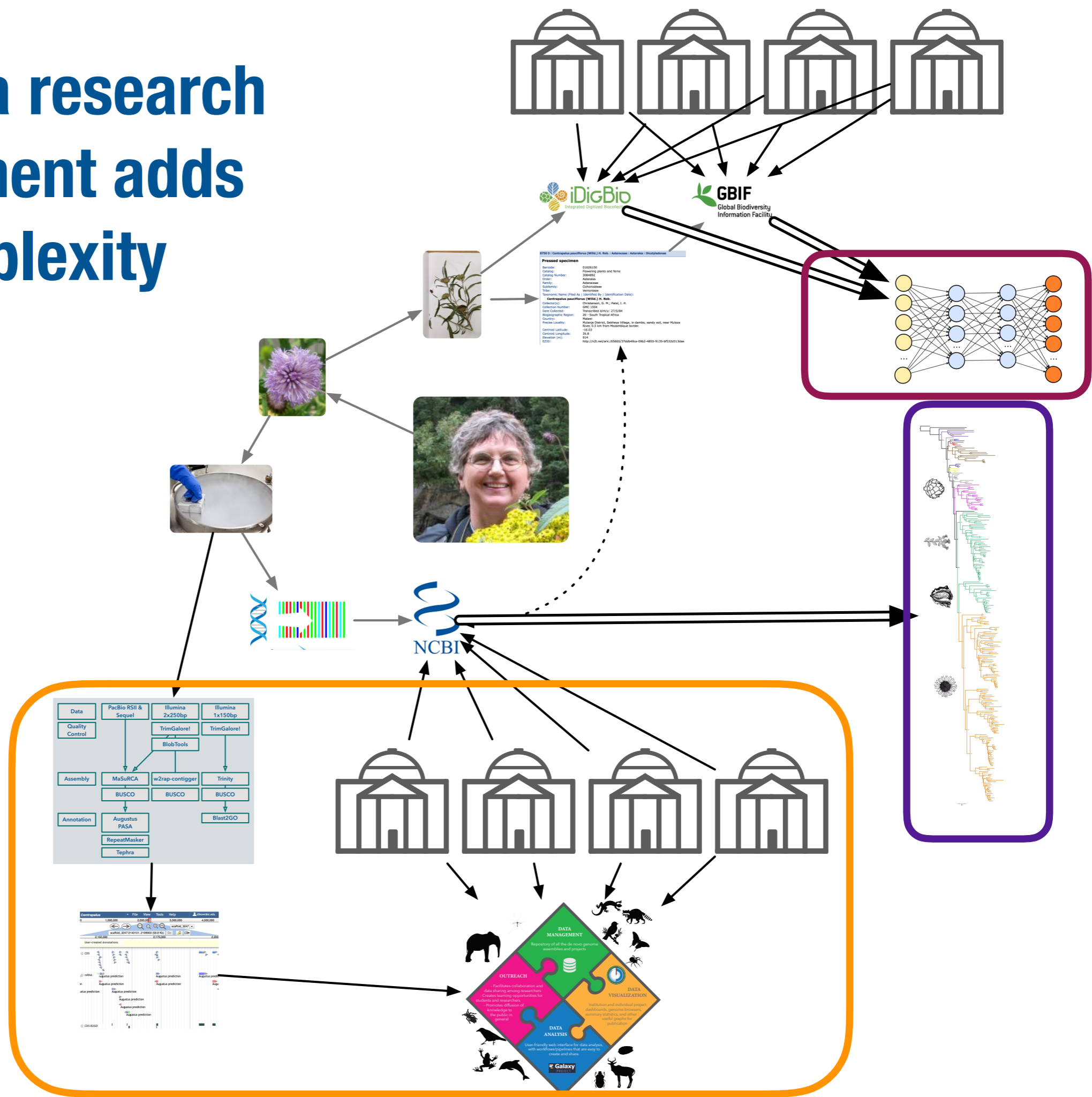


# Adding a research component adds complexity

deep learning

phylogeny

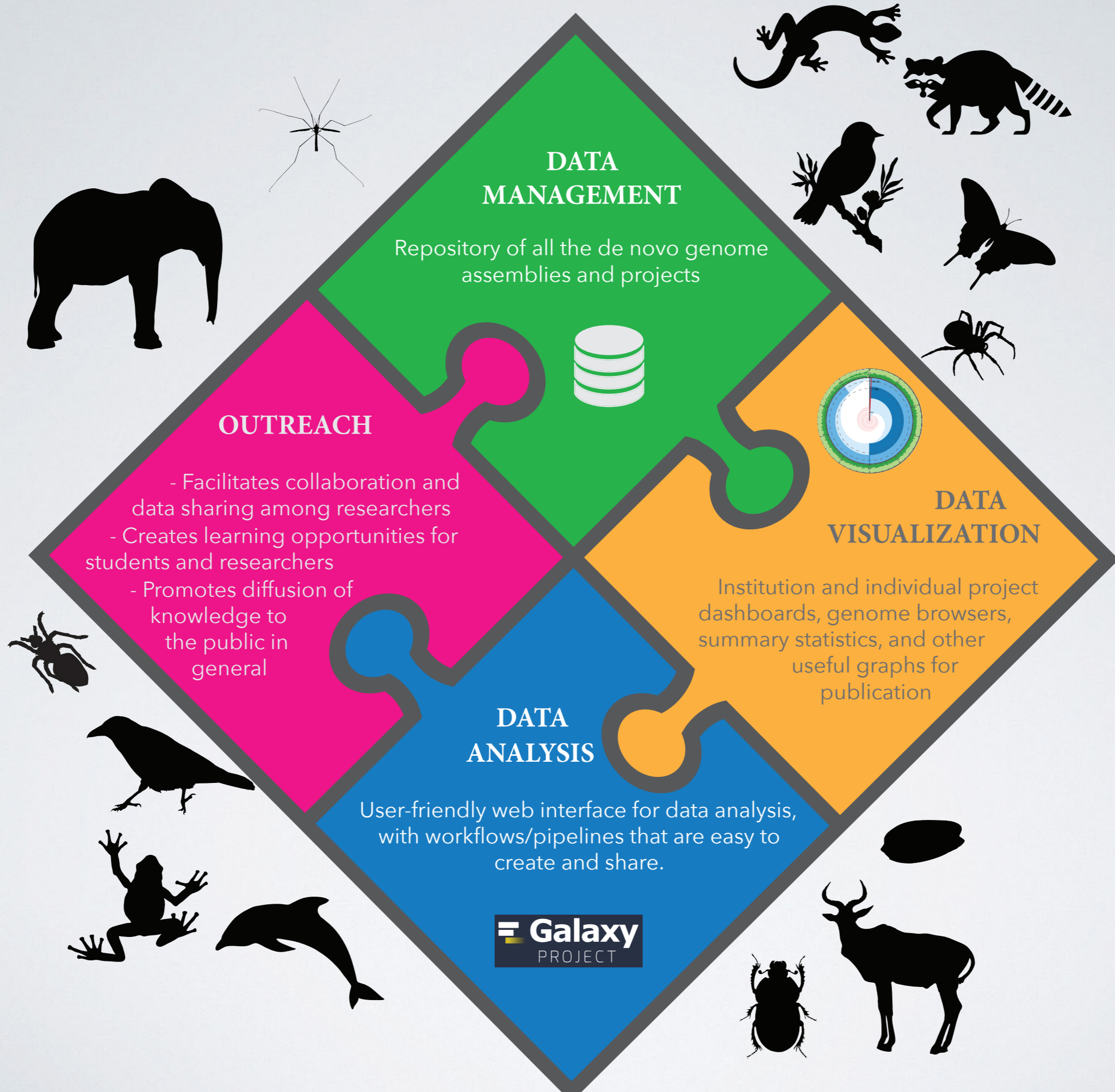
genomics



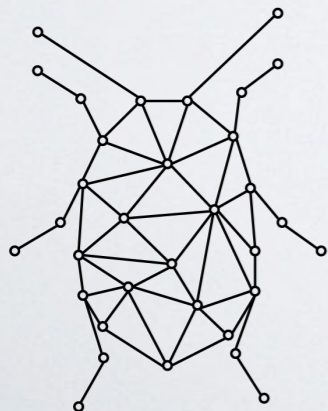


# Smithsonian Biodiversity Genome Hub: under construction

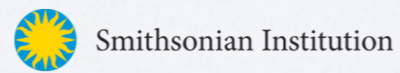




# Botany



DATA  
SCIENCE  
LAB



NATIONAL  
MUSEUM of  
NATURAL  
HISTORY

**NVIDIA®**





# Mercury staining

		PREDICTED	
		unstained	stained
ACTUAL	unstained	882	46
	stained	77	682

# Family ID

		PREDICTED	
		clubmoss	spikemoss
ACTUAL	clubmoss	858	59
	spikemoss	23	901

# Bumblebees





Clerke Wm.  
No. 2384  
July 21, 1970

A. W. Mason  
Collector  
June 10, 1971

Bombus bifarius  
♀  
Cresson  
Det. R. B. Miller 1967  
Fraser  
B. Tolland, Colo.  
July 1975  
L.A. Koenig

Bombus  
No. 12, 209  
July 12, 1976

Bombus  
No. 48, 487  
July 27, 1977

J. C. Robinson  
Collector  
August  
V. 22

Bombus bifarius  
♂  
Cresson  
Det. R. B. Miller 1967

J. C. Robinson  
Collector  
Marion Co  
Arkansas

Pierre  
No. 27, 275  
July 1, 1972

Bombus  
No. 43, 438  
July 10, 1977

Bombus  
No. 78, 788  
det. Franklin  
Jan. 1968  
B. Franklin  
Dallas

Bombus  
No. 10, 100  
July 10, 1970

Bombus  
No. 3, 314  
July 21, 1970

Bombus  
No. 14  
Collector  
C. S. Stiles

Racine Co., Wis.  
No. 4, 44  
Coll. L. W. Mason  
MAY 12, 1965

L. W. Mason  
Collector  
July 10, 1977

Mal King Co.  
Seattle-Arboretum  
30 April 1972  
E. M. Stiles

Bombus  
No. 23, 239  
July 21, 1970

Walworth Co., Wis.  
No. 2, 279  
Coll. L. W. Mason  
July 27, 1965

Bombus  
No. 118  
Coll. L. W. Mason  
July 12, 1965

el. 11-1200  
No. 118-118  
W. S. Stiles  
COLORADO  
Cherokee  
Garfield

viraculatus ♀

Bombus  
Collector  
Vassar

Trifolium repens  
Bombus  
No. 27, 275  
July 1, 1972

Soligo  
elongata

Vicia villosa

Walworth Co., Wis.  
No. 2, 226  
Coll. L. W. Mason  
Aug 6, 1964

Bombus pleuralis  
♂  
Cresson  
Det. R. B. Miller 1967

Bombus  
No. 1862  
Det. Miltron 1862  
USNM  
C. W. Mason  
Collector  
Martha V. Vond  
0-10-56

Trifolium repens  
Walgrave, Barber Co.  
No. 21, 214  
Coll. L. W. Mason  
21 June 1971, 1128hrs  
E. M. Stiles

Pyrobombus  
bifarius (Gr.)  
Det. Miltron 1962  
USNM

MONARCHIA BUB  
No. 2776  
July 1, 1972

Colas  
Cassini  
A. S.

Pierre  
No. 27, 275  
July 1, 1972

J. C. Robinson  
Collector  
Corvallis  
Or. Jun 09

Bombus  
No. 21, 214  
Coll. L. W. Mason  
21 June 1971, 1128hrs  
E. M. Stiles

Epilobium  
angustifolium  
W. S. Stiles  
Walworth Co.  
Tel. W. Garfield  
1 August 1971 0900hrs  
E. M. Stiles

Racine Co., Wis.  
No. 4, 425  
Coll. L. W. Mason  
July 27, 1965

Bombus  
No. 118  
Coll. L. W. Mason  
July 12, 1965

Walworth Co.  
No. 4, 425  
Coll. L. W. Mason  
23 June 1965  
E. M. Stiles

Bombus  
No. 34, 346  
June 13, 1968

Bombus  
No. 34, 346  
June 13, 1968

Bombus  
No. 21, 214  
Coll. L. W. Mason  
21 June 1971, 1128hrs  
E. M. Stiles

Pyrobombus  
bifarius (Gr.)  
Det. Miltron 1962  
USNM

Bombus  
No. 118  
Coll. L. W. Mason  
July 12, 1965

Cassini  
Cassini

Bombus  
No. 34, 346  
June 13, 1968

Trifolium pratense  
Walworth Co.  
Tel. W. Garfield  
10 July 1970  
E. M. Stiles

# Crowdsourced transcription

The screenshot shows a web browser window with the URL transcription.si.edu. The page features a teal header with the Smithsonian logo and navigation links: HOME, PROJECTS, SEARCH, ABOUT, TIPS, NEWS. Below the header is a black bar with the text 'SMITHSONIAN DIGITAL VOLUNTEERS: TRANSCRIPTION CENTER' and links for SIGNUP and LOGIN. The main content area has a breadcrumb trail: HOME > PROJECTS > NMNH - DEPARTMENT OF ENTOMOLOGY, followed by the title 'THE BUMBLEBEE PROJECT - SET 2'. On the left, there is a section titled 'About the Project' with a paragraph of text and a 'Read More' link. At the bottom left of this section is a 'Download PDF | Go to Page' form with an input field and a 'Go' button. On the right, a 'Completed!' box is followed by a 'Project Progress (details)' section with a green progress bar. Below this are two statistics boxes: '23 CONTRIBUTING MEMBERS' and '366 TOTAL PAGES'. A vertical 'feedback' button is located on the far left edge of the page.

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](mailto: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

## Completed!

Project Progress ([details](#))

**23**  
CONTRIBUTING  
MEMBERS

**366**  
TOTAL  
PAGES

feedback



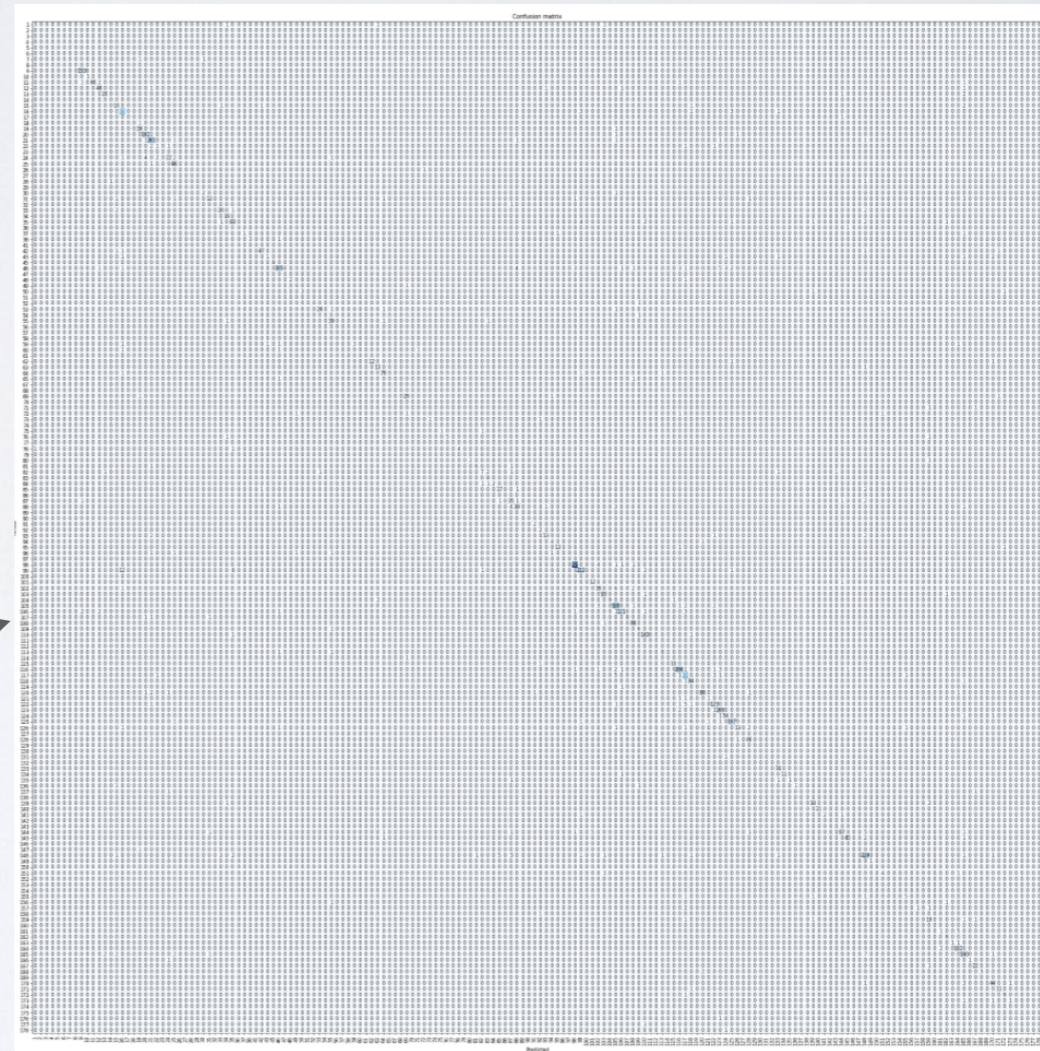
# Models in progress

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 %**

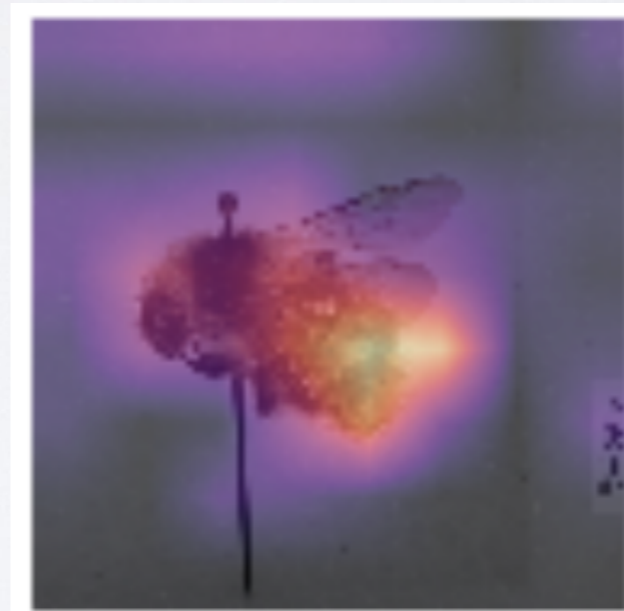
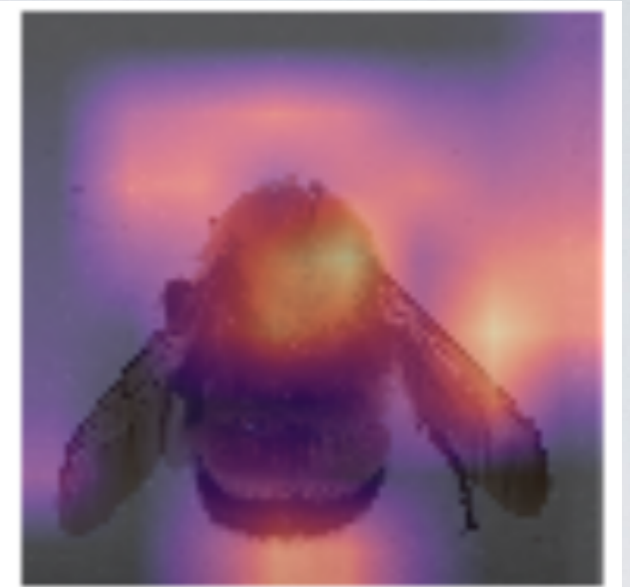
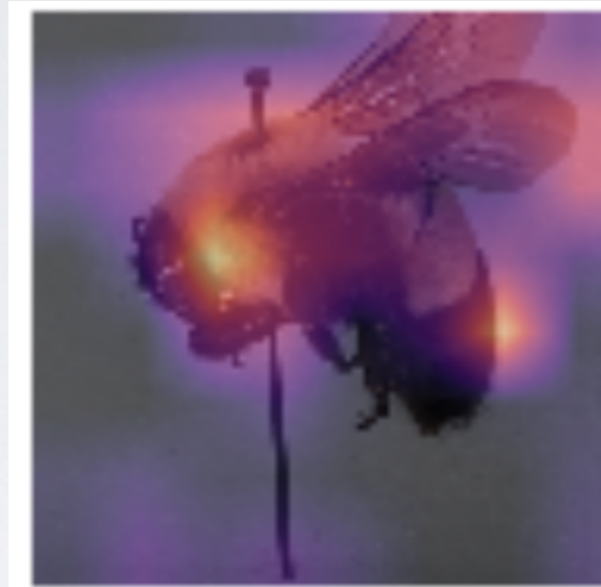
**Species model:**  
**178 classes**  
**Overall accuracy: 92.5%**



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)

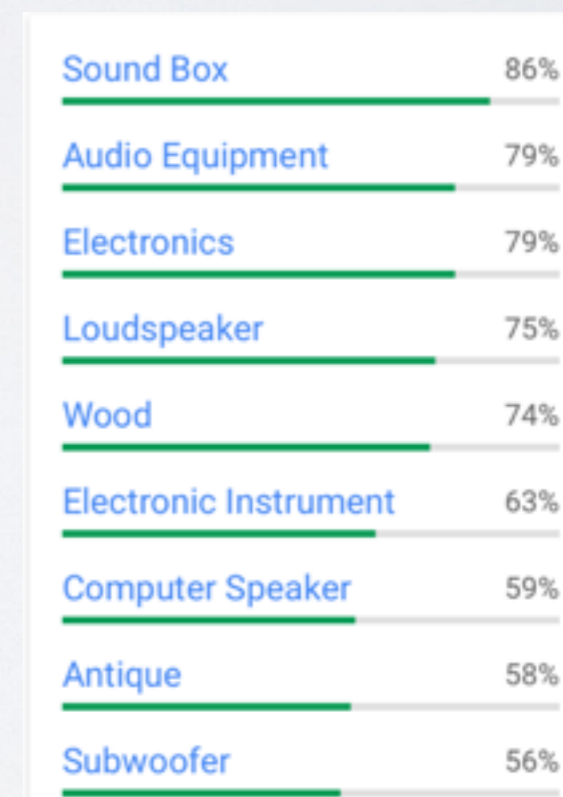
**Activation heat maps  
allow us to start  
exploring the models  
in more detail.**



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



## Morse Daguerreotype Camera



# Archives

UNITED STATES  
**HOLOCAUST**  
MEMORIAL  
**MUSEUM**

<p>BECAUSE OF <b>HER</b> STORY</p>	<p>SMITHSONIAN AMERICAN WOMEN'S HISTORY INITIATIVE</p>
	<p>WOMENSHISTORY.SI.EDU</p>

# 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](https://github.com/SmithsonianWorkshops)

Instructors and schedule: [datascience.si.edu/carpentries](https://datascience.si.edu/carpentries)





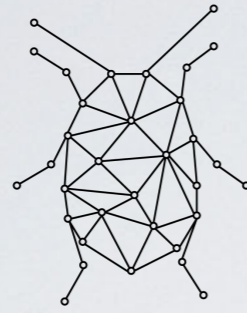
# Takeaways

**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**

