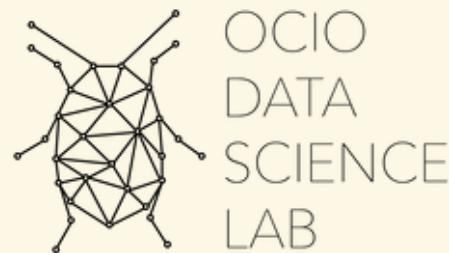


# DATA SCIENCE AT THE SMITHSONIAN



Smithsonian



Mike Trizna

Smithsonian OCIO Data Science Lab

November 1, 2021 | *University of Richmond*

# WHAT IS THE SMITHSONIAN INSTITUTION?

Yes, there are the museums (19 of them, mostly in Washington, DC), but we also have 21 libraries and archives, 9 research centers ... and a zoo.



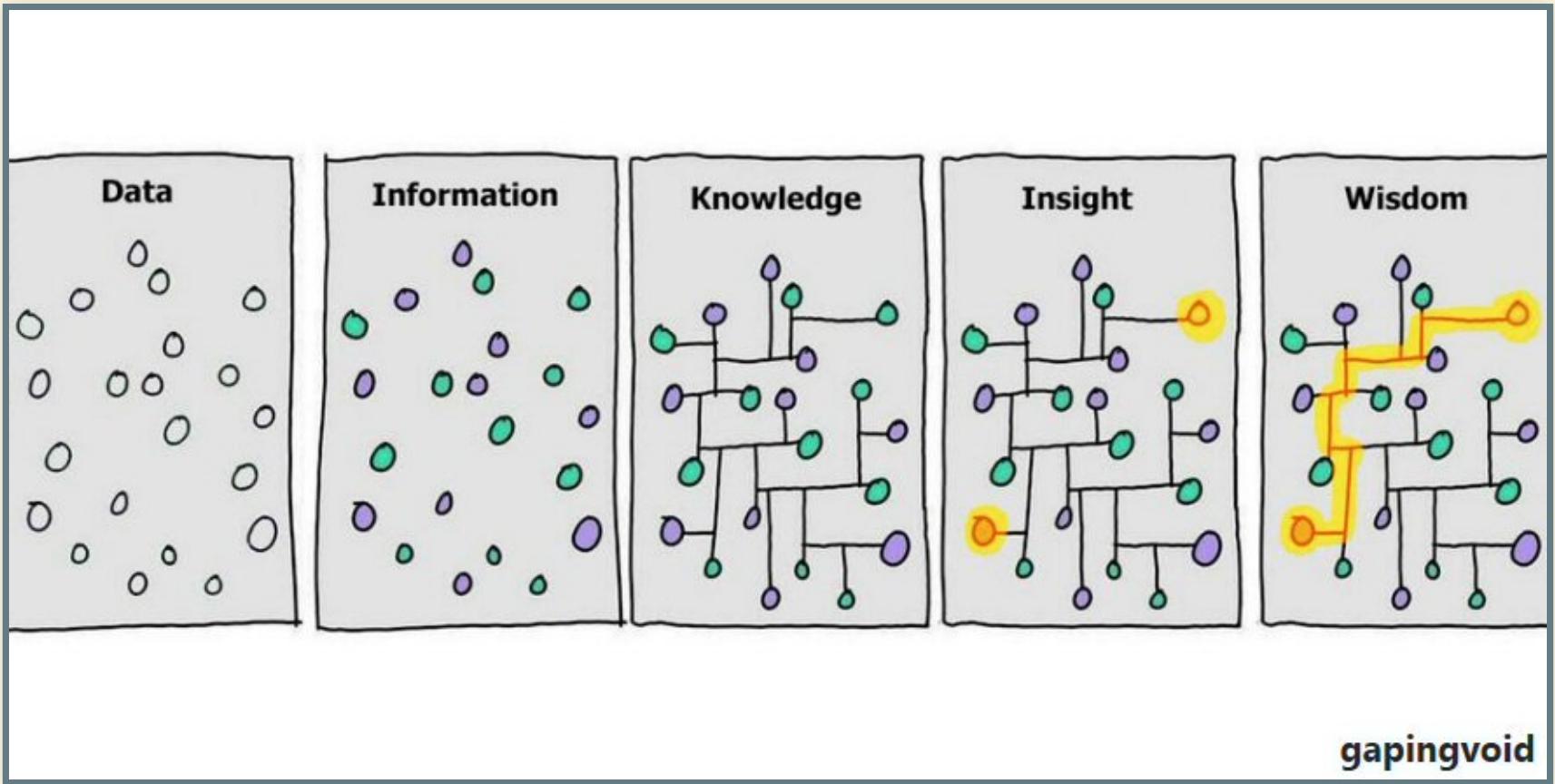
# SMITHSONIAN MISSION

Founded in 1846 from the bequest of Englishman James Smithson with the condition:

*"under the name of the Smithsonian Institution, an establishment for the increase and diffusion of knowledge."*



The researchers of the Smithsonian have been increasing and diffusing Knowledge since 1846.



How can we leverage *Insight* and *Wisdom* recorded by experts to make sense of growing amounts of *Data*?

# WHAT IS THE DATA SCIENCE LAB?

Located in the Smithsonian Office of the Chief Information Officer (OCIO), the OCIO Data Science Lab acts as a "Lab" in the sense that it is:

- an environment producing high-quality scholarship and training new researchers
- a place to pilot new technologies and techniques with Smithsonian data and processes

# WHO IS THE DATA SCIENCE LAB?



*(Image of OCIO Data Science Lab participants circa  
Summer 2019)*

# RESEARCH: GENOMICS

## Genomics Projects

Genomics of Neotropical Electric Fish across the Isthmus of Panama



Detection of genome contamination using machine learning tools



Genomics and demographic history of clouded leopards



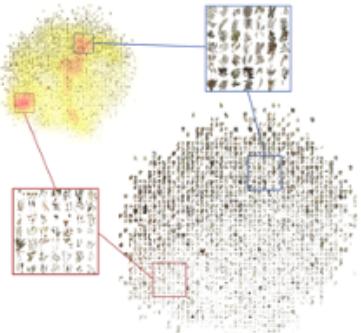
Phylogenomics of mosses in extreme environments





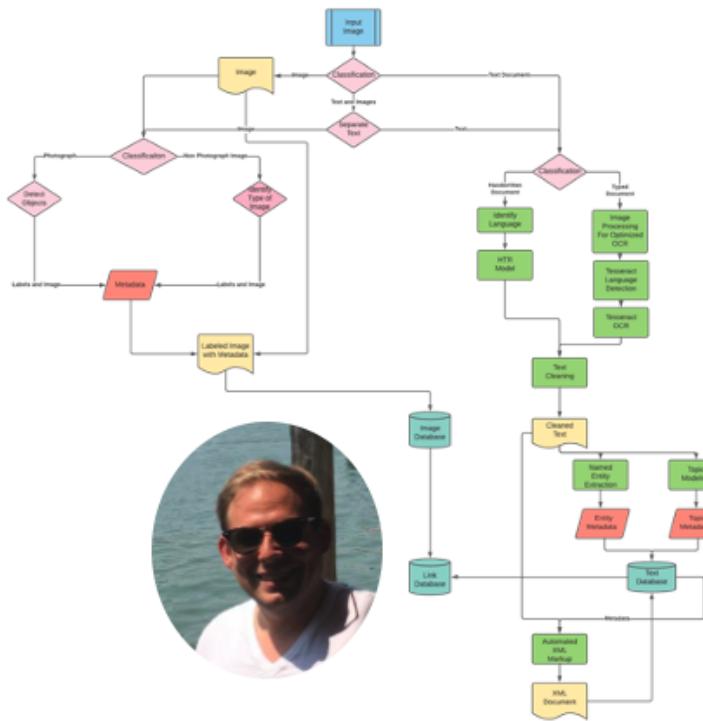
# RESEARCH: COLLECTIONS AND ARCHIVES

## Collections and Archives Machine Learning Projects



Building  
machine learning  
tools to  
understand fern  
shape and  
biogeography

Building machine learning tools to make  
archives and collections discoverable  
(USHMM and AWHI)





# PILOT: HERBARIUM SHEET MERCURY DETECTOR



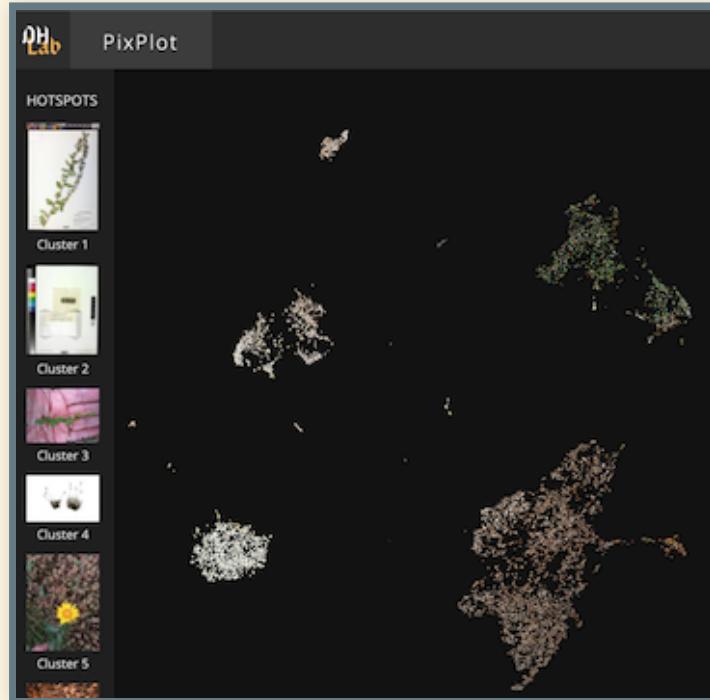
2017 paper that described building a machine learning model to detect herbarium sheets that had been stained with mercury.



<https://doi.org/10.3897/bdj.5.e21139>



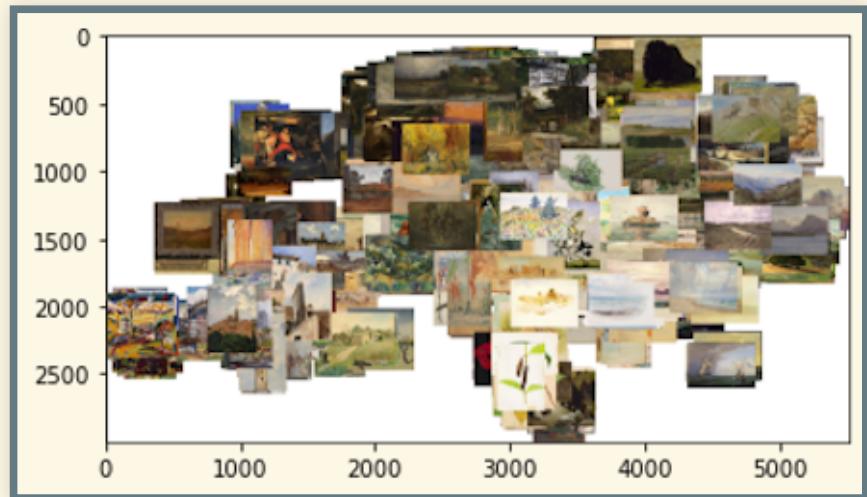
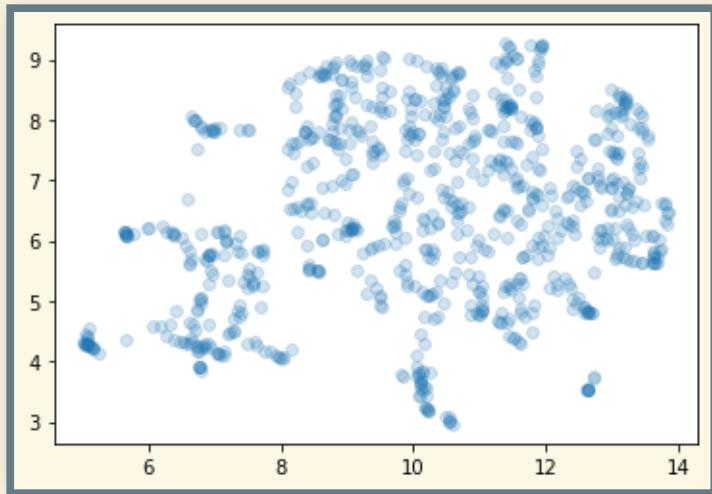
# PILOT: USING YALE DHL PIXPLOT FOR IMAGE DATASET EXPLORATION



[https://sidatascienceLab.github.io/mercury\\_sheets/](https://sidatascienceLab.github.io/mercury_sheets/)

# PILOT: EXAMPLE OF USING SI OPEN ACCESS MATERIALS

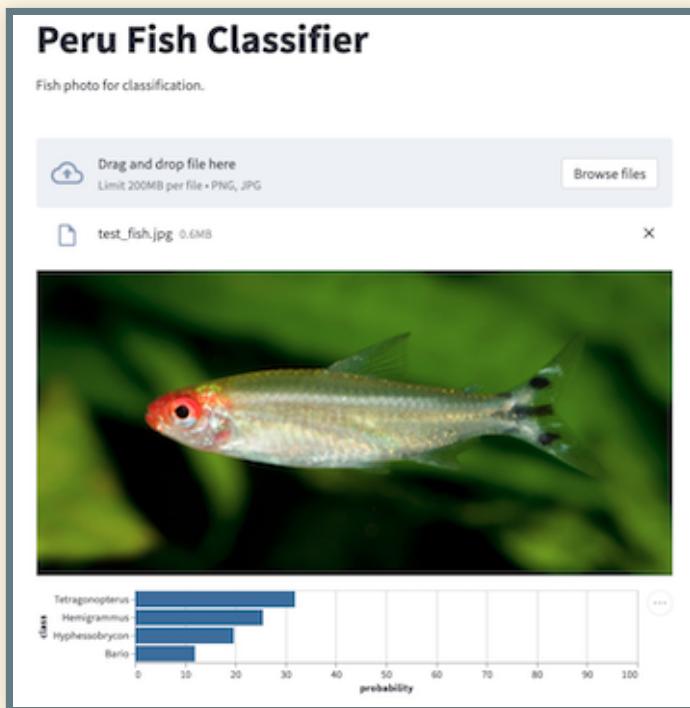
## Semantic Clustering American Art Paintings



Full interactive notebook (through Binder) available at  
<https://github.com/sidatascience/siopenaccess>.

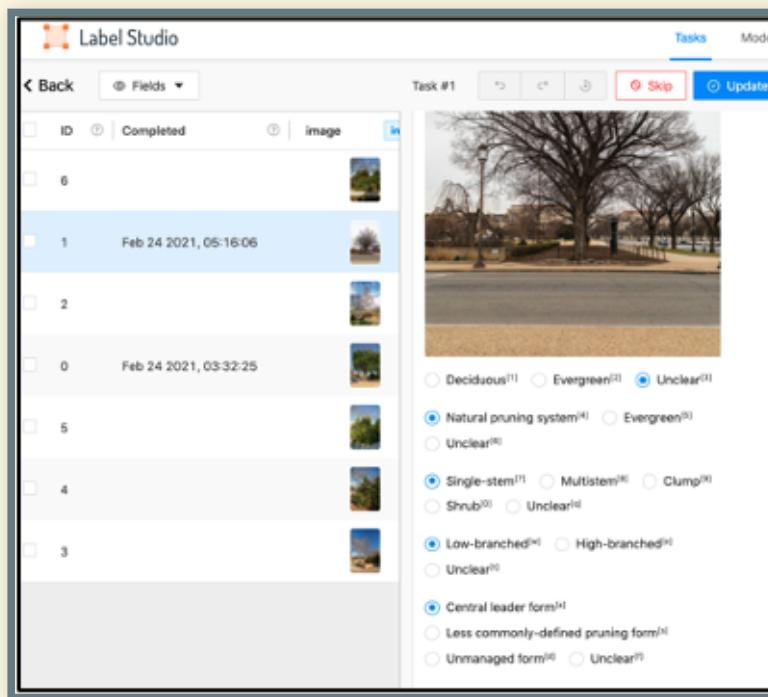
# PILOT: AMAZONIAN FISH CLASSIFIER

Automated fish species classifier, supplemented with digitized museum specimens.



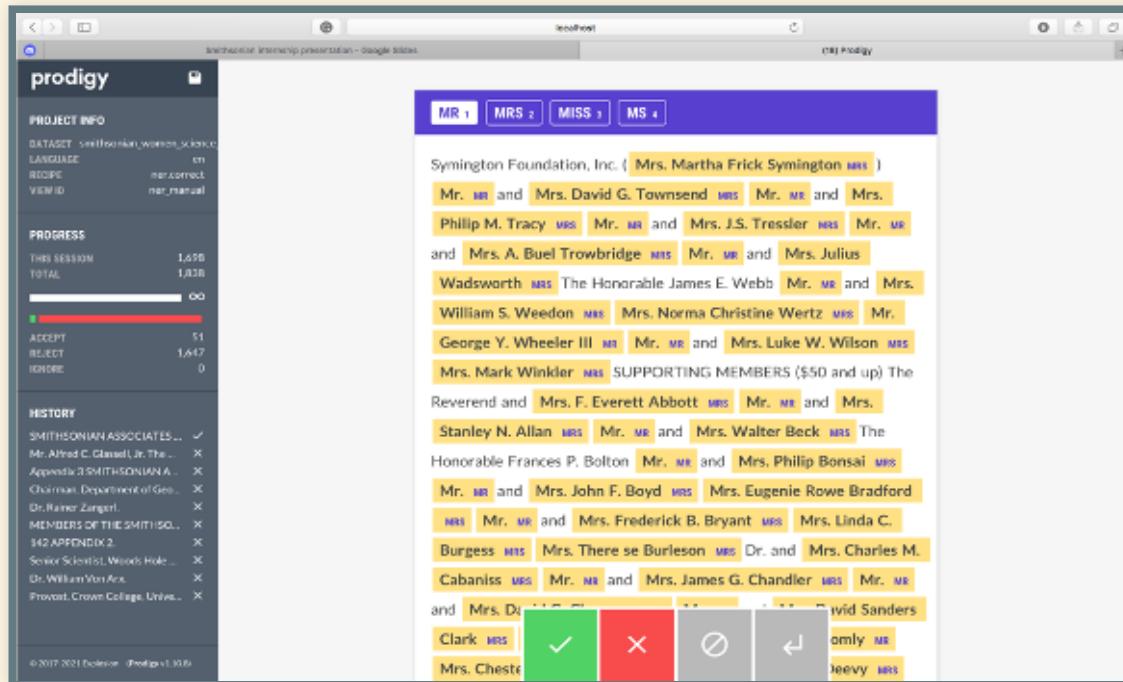
# PILOT: SI GARDENS TREE LABELER

Winter project for SI horticulturalist to train models on tree-less photos of trees



# PILOT: AMERICAN WOMEN'S HISTORY INITIATIVE NLP

Using ML to learn complicated Mr./Mrs./Dr. titles



# INTERNAL TRAINING: THE CARPENTRIES

## Carpentries and Data Science Skills Training

**Learners**  
Over 500 staff, fellows, interns from all units of SI.  
*No prior experience necessary.*



AI/ML

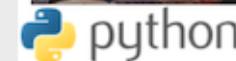
### Workshops

Workshops are entirely hands-on to promote learning by debugging.



### Communities of Practice

Connecting workshop "alumni" and other practitioners to keep learning together.



### Instructors

Workshops organized by peers (~25 SI staff and fellows from DC, Cambridge, Panama)

### Instructor Training

Certified instructors taught evidence-based practices for teaching computational skills.



<https://datascience.si.edu/carpentries>

# INTERNAL TRAINING: CARPENTRIES AI/ML FOR GLAM

The screenshot shows a website header with navigation links: Home, Code of Conduct, Setup, Episodes (dropdown), Extras (dropdown), License, and Improve this page (with a pencil icon). There is also a search bar labeled "Search...". A blue banner at the top states: "This lesson is part of The Carpentries Incubator, a place to share and use each other's Carpentries-style lessons. This lesson has not been reviewed by and is not endorsed by The Carpentries." The main content area features a large blue heading "Intro to AI for GLAM". Below it, a text block explains the lesson's purpose: "This lesson aims to empower GLAM (Galleries, Libraries, Archives, and Museums) staff by providing the foundation to support, participate in and begin to undertake in their own right, machine learning-based research and projects with heritage collections." Another text block states: "After attending, learners will be able to:" followed by a bulleted list of nine items detailing what learners will be able to do after attending.

This lesson is part of The Carpentries Incubator, a place to share and use each other's Carpentries-style lessons. This lesson has not been reviewed by and is not endorsed by The Carpentries.

## Intro to AI for GLAM

This lesson aims to empower GLAM (Galleries, Libraries, Archives, and Museums) staff by providing the foundation to support, participate in and begin to undertake in their own right, machine learning-based research and projects with heritage collections.

After attending, learners will be able to:

- Explain and differentiate key terms, phrases, and concepts associated with AI and Machine Learning in GLAM;
- Describe ways in which AI is being innovatively used in the cultural heritage context today;
- Identify what kinds of tasks machine learning models excel at in GLAM applications;
- Identify weaknesses in machine learning models;
- Reflect on ethical implications of applying machine learning to cultural heritage collections and discuss potential mitigation strategies;
- Summarise the practical, technical steps involved in undertaking machine learning projects.
- Identify additional resources on AI and Machine Learning in GLAM.

<https://carpentries-incubator.github.io/machine-learning-librarians-archivists/>

# QUESTIONS?

Email: [triznam@si.edu](mailto:triznam@si.edu)

Slides:

[https://github.com/MikeTrizna/richmond\\_datascience\\_ta](https://github.com/MikeTrizna/richmond_datascience_ta)

