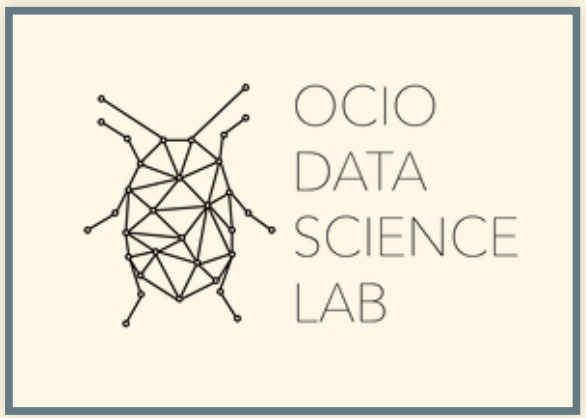
# DATA SCIENCE AT THE SMITHSONIAN


**Mike Trizna**

Smithsonian OCIO Data Science Lab

November 1, 2021 | *University of Richmond*

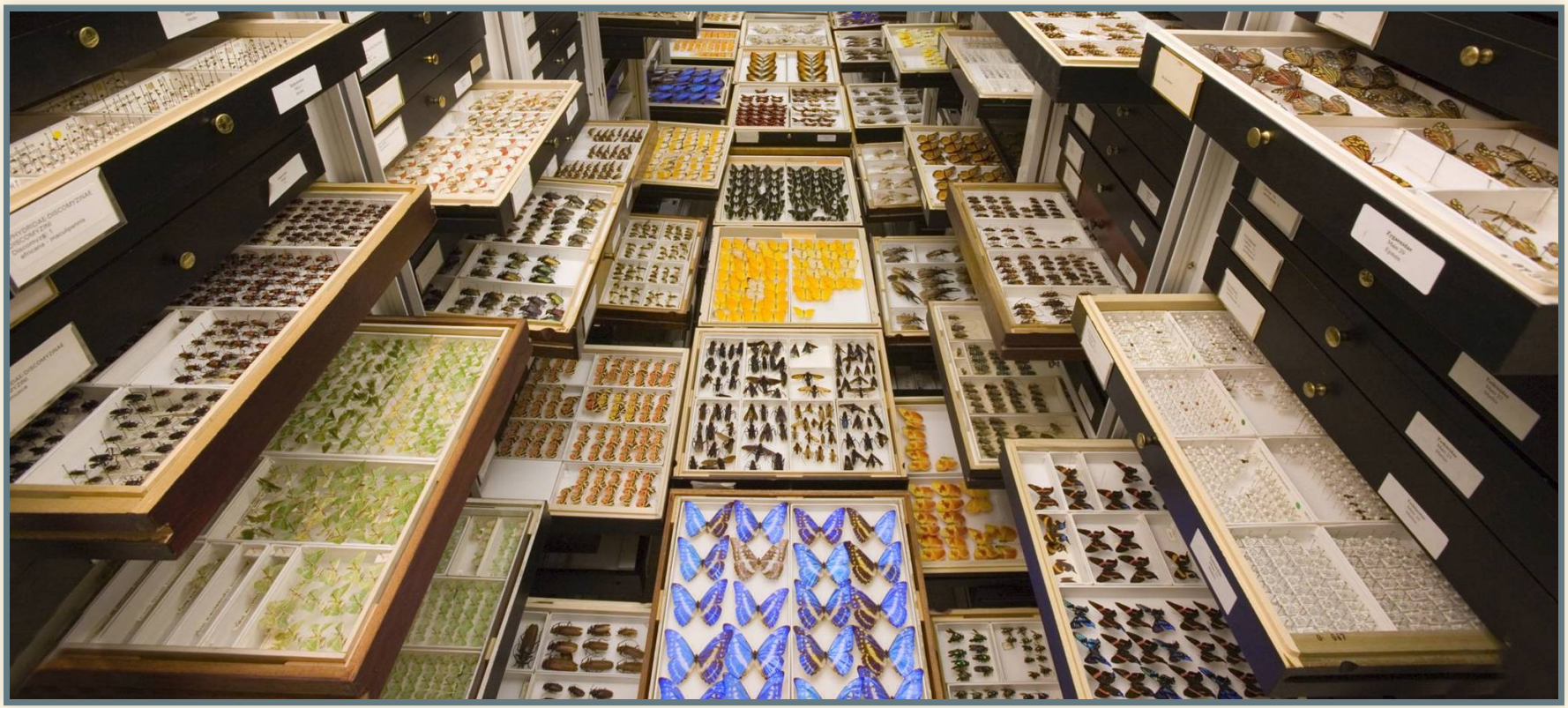
# WHAT IS THE SMITHSONIAN INSTITUTION?

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.

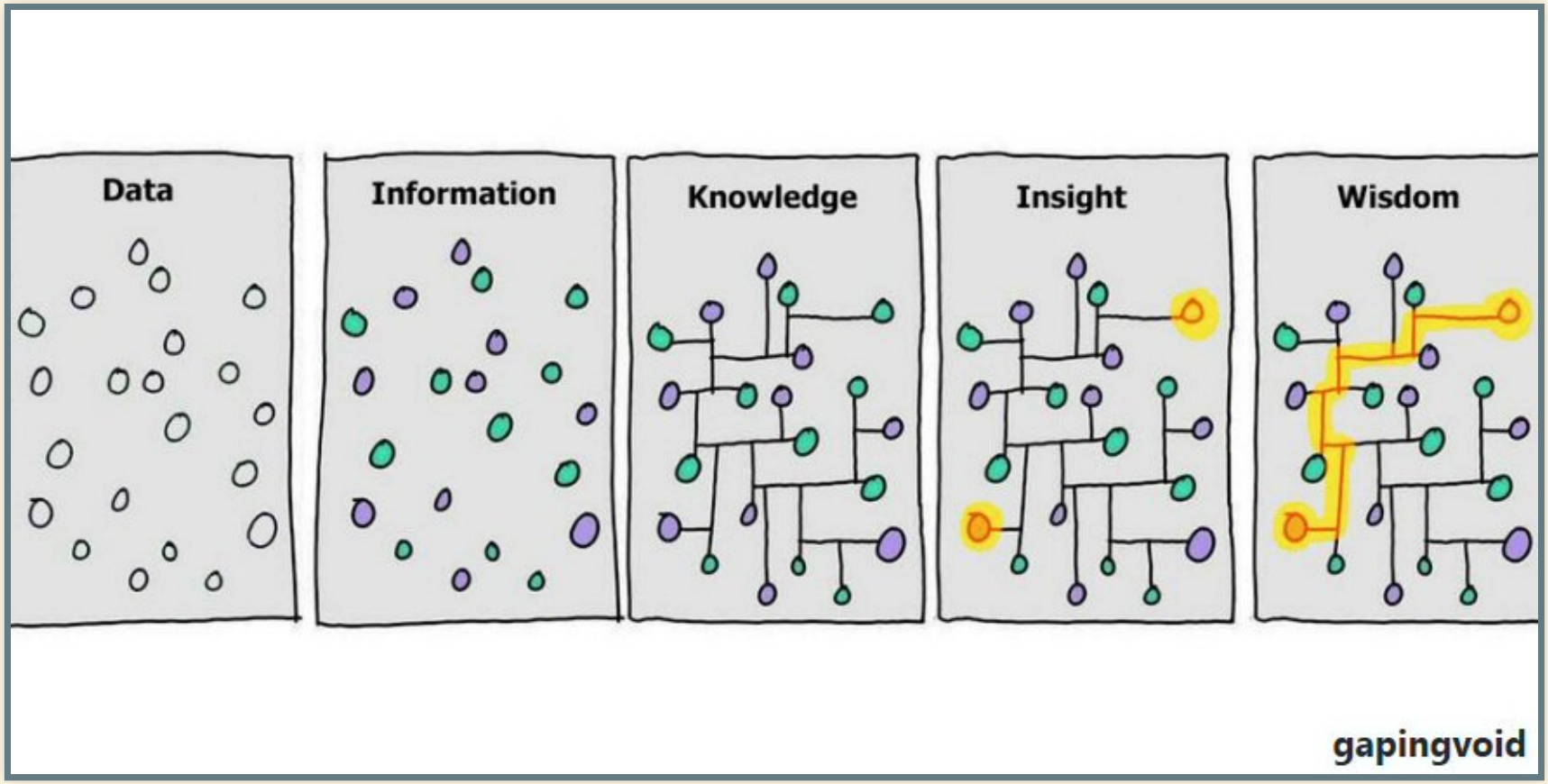
# SMITHSONIAN MISSION

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

Detection of genome contamination using machine learning tools

Genomics and demographic history of clouded leopards

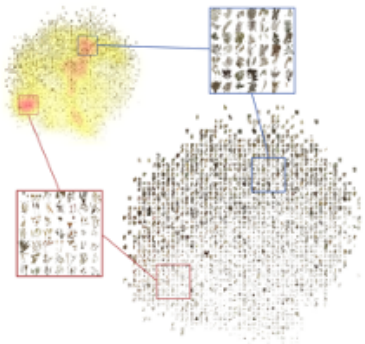
Genomics of Neotropical Electric Fish across the Isthmus of Panama

Phylogenomics of mosses in extreme environments

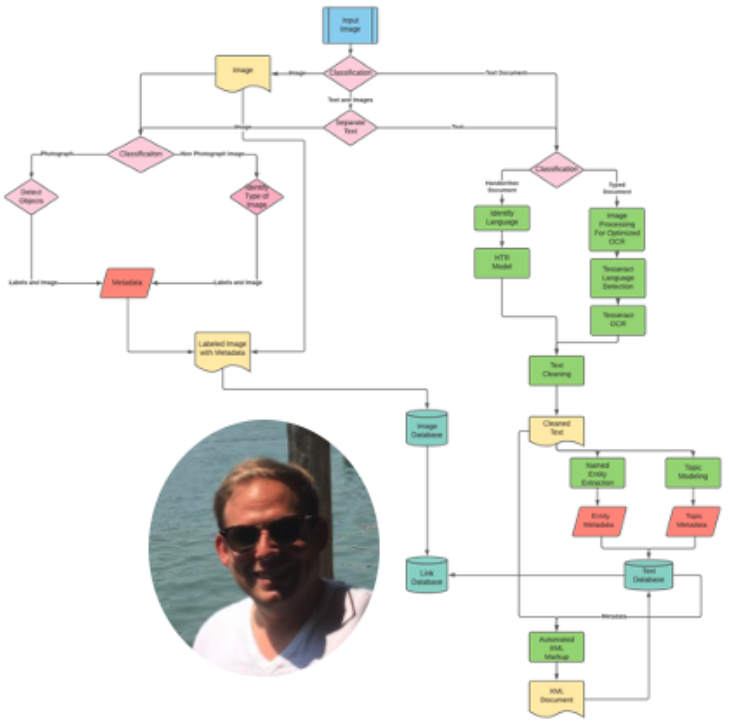
# RESEARCH: COLLECTIONS AND ARCHIVES

## Collections and Archives Machine Learning Projects

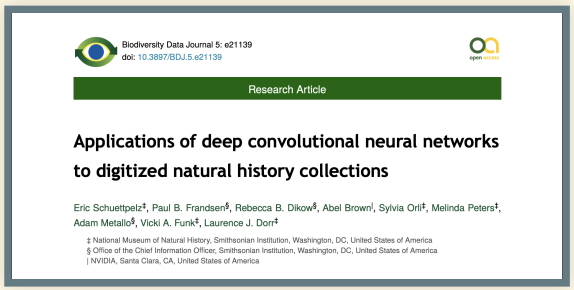
Building machine learning tools to understand fern shape and biogeography



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



# PILOT: HERBARIUM SHEET MERCURY DETECTOR



Biodiversity Data Journal 5: e21139
doi: 10.3897/BDJ.5.e21139

Research Article

**Applications of deep convolutional neural networks to digitized natural history collections**

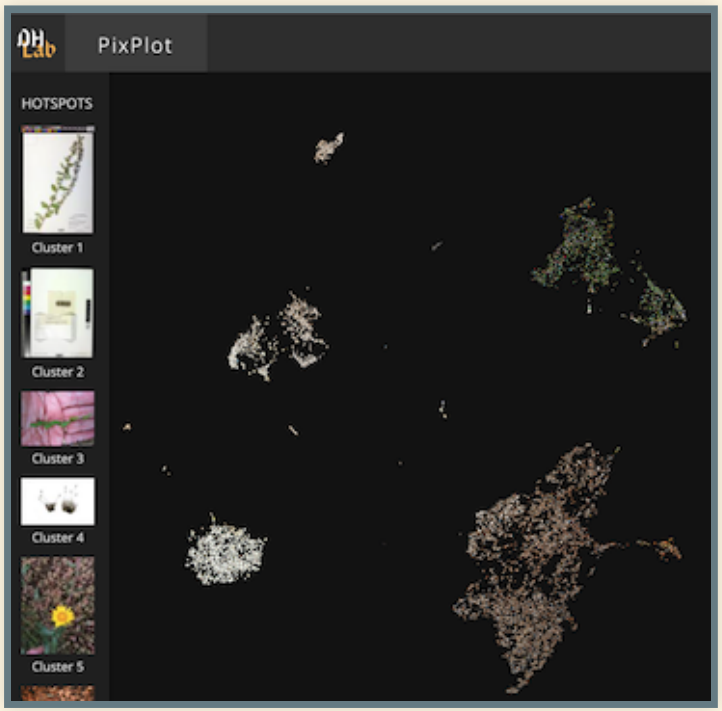
Eric Schuettpelz‡, Paul B. Frandsen§, Rebecca B. Dikow§, Abel Brown|, Sylvia Orli‡, Melinda Peters‡, Adam Metallo§, Vicki A. Funk‡, Laurence J. Dorr‡

‡ National Museum of Natural History, Smithsonian Institution, Washington, DC, United States of America
§ Office of the Chief Information Officer, Smithsonian Institution, Washington, DC, United States of America
| NVIDIA, Santa Clara, CA, United States of America


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

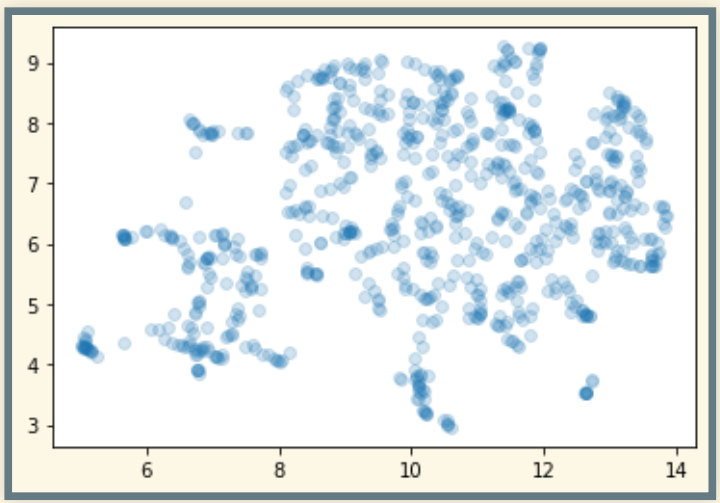
# PILOT: USING YALE DHL PIXPLOT FOR IMAGE DATASET EXPLORATION

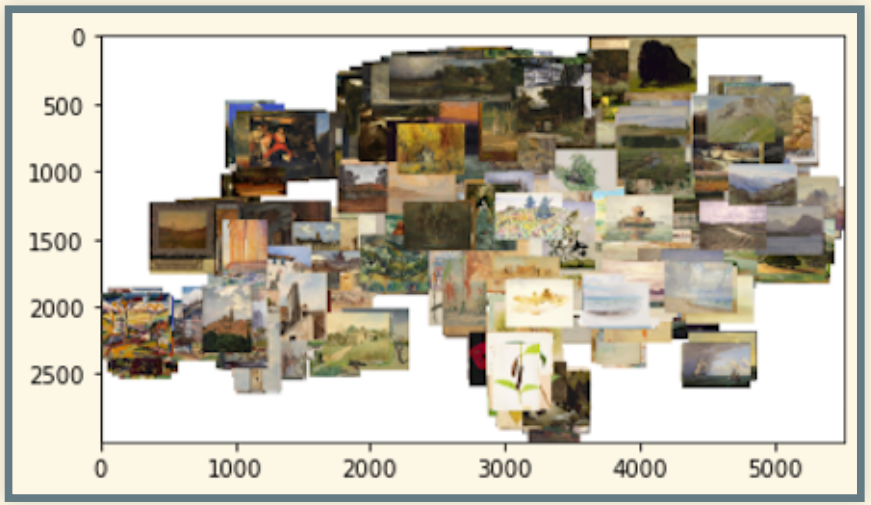



https://sidatasciencelab.github.io/mercury_sheets/

# PILOT: EXAMPLE OF USING SI OPEN ACCESS MATERIALS

## Semantic Clustering American Art Paintings



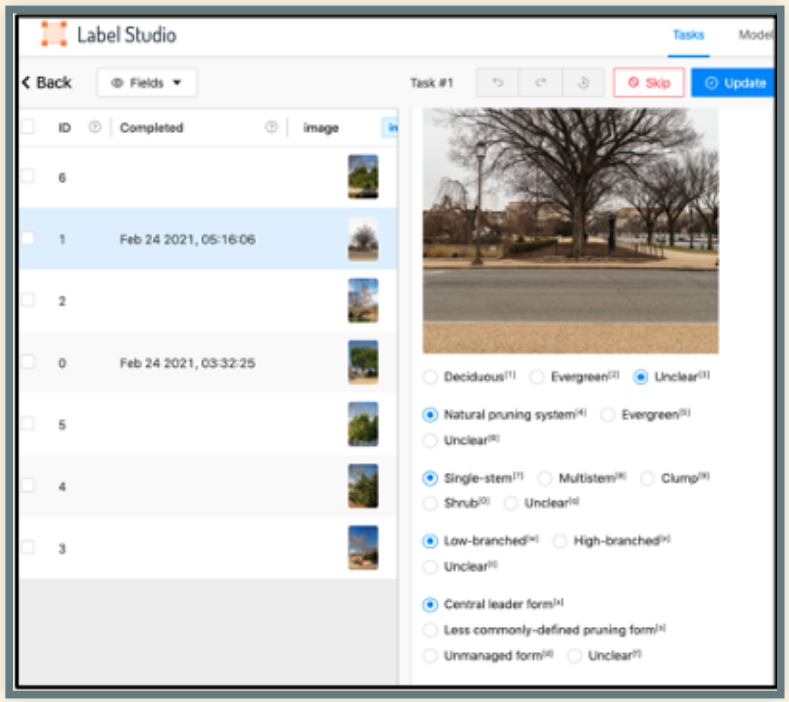

Full interactive notebook (through Binder) available at https://github.com/sidatasciencelab/siopenaccess.

# PILOT: AMAZONIAN FISH CLASSIFIER

Automated fish species classifier, supplemented with digitized museum specimens.

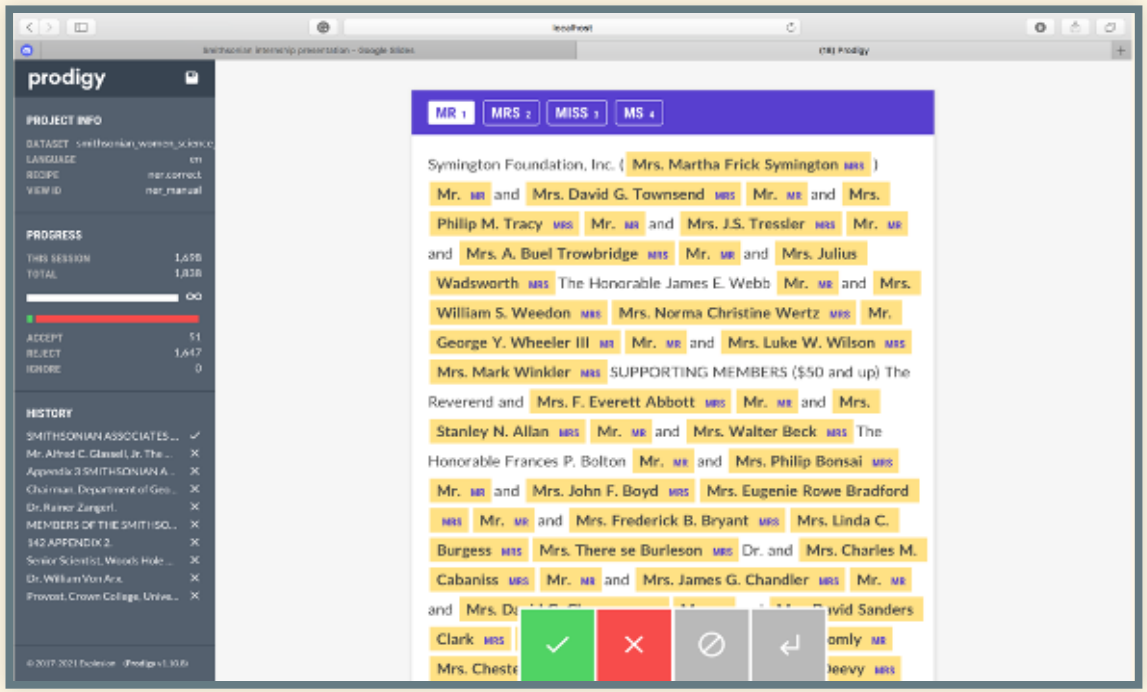
# PILOT: SI GARDENS TREE LABELER

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

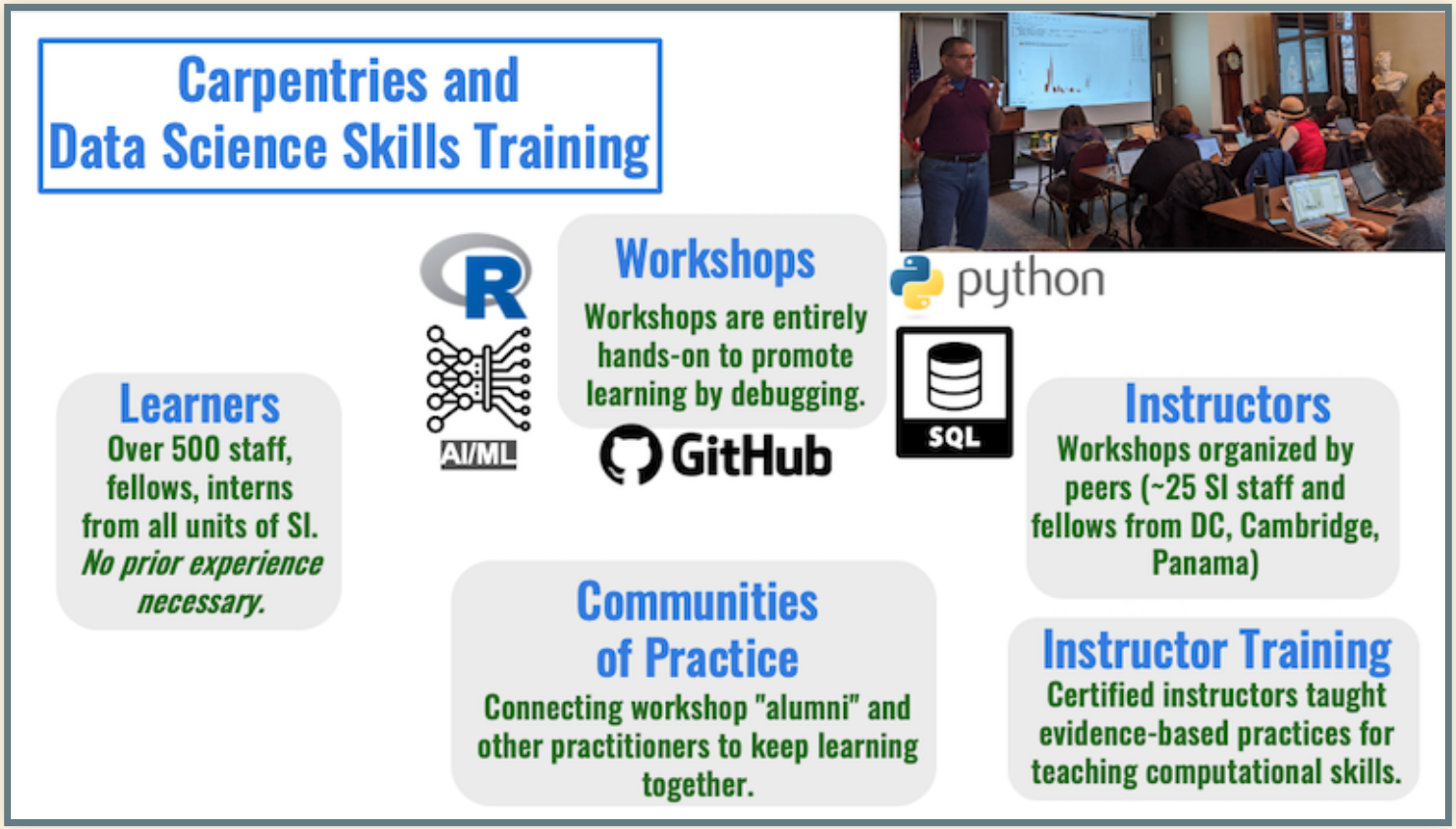


# PILOT: AMERICAN WOMEN'S HISTORY INITIATIVE NLP

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

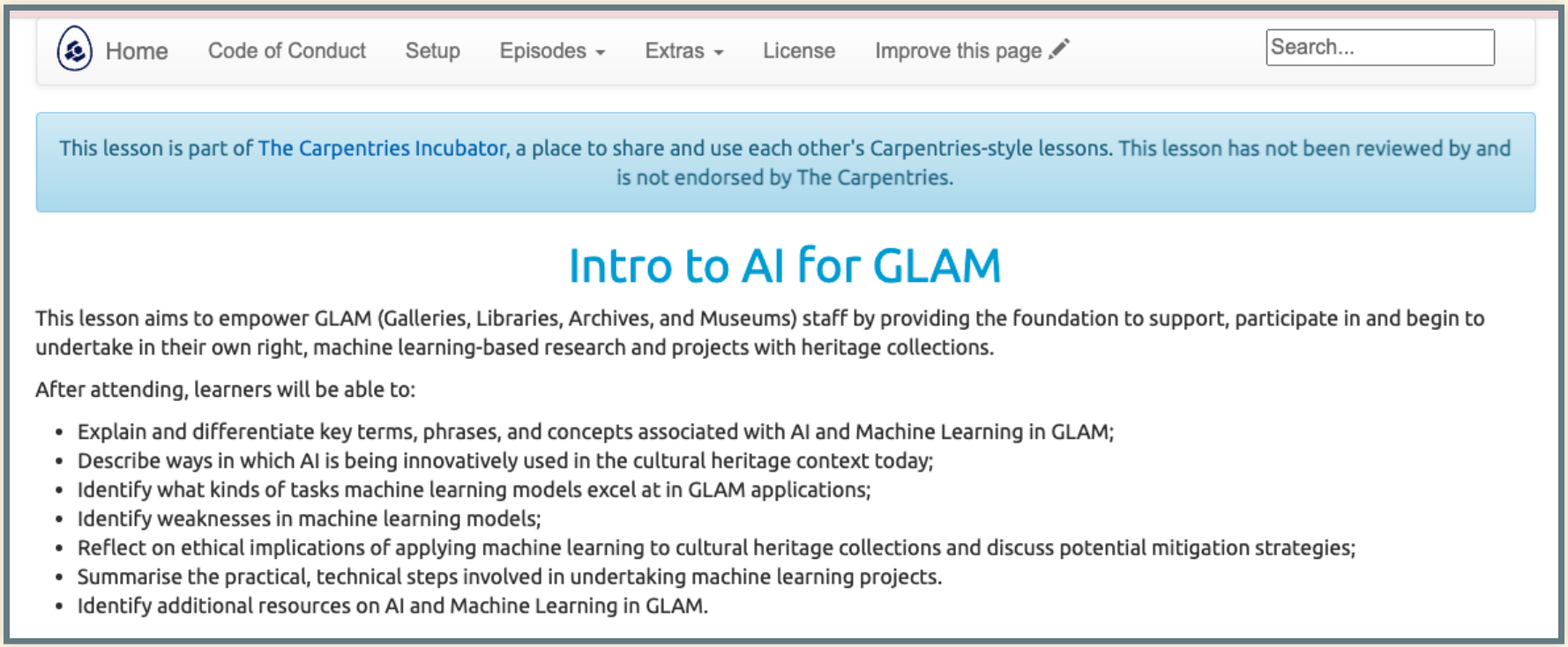


# INTERNAL TRAINING: THE CARPENTRIES




https://datascience.si.edu/carpentries

# INTERNAL TRAINING: CARPENTRIES AI/ML FOR GLAM




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

# QUESTIONS?

Email: triznam@si.edu

Slides:
https://github.com/MikeTrizna/richmond_datascience_ta