# Introduction to machine learning using digitized collections

Rebecca B. Dikow
Data Science Lab
Office of the Chief Information Officer
Smithsonian Institution

@rdikow
@SIDataScience

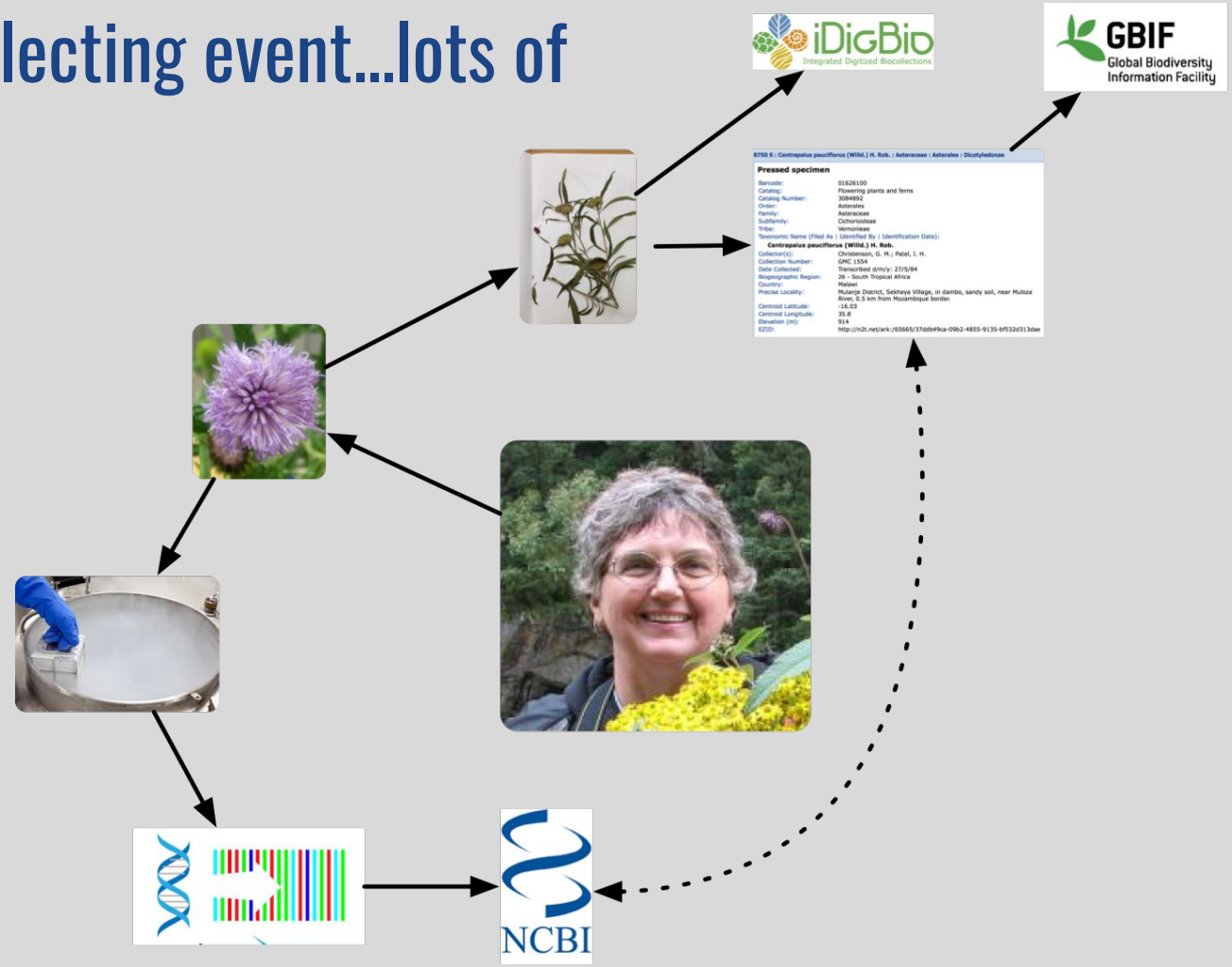

# What are digitized collections?

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

# One specimen collecting event…lots of potential data



# Digitizing the US National Herbarium

## Scaling from thousands to millions:

Smithsonian Institution
NATIONAL MUSEUM of NATURAL HISTORY

Smithsonian
DIGITIZATION

# GLAM data types suited for ML:

- **Tabular data**
- **Images** (computer vision)
- **Text** (natural language processing)

# What does it mean to LEARN in machine learning?

"A computer program is said to learn from experiance E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experiance E."

— Tom Mitchell (1997)

Chris Albon

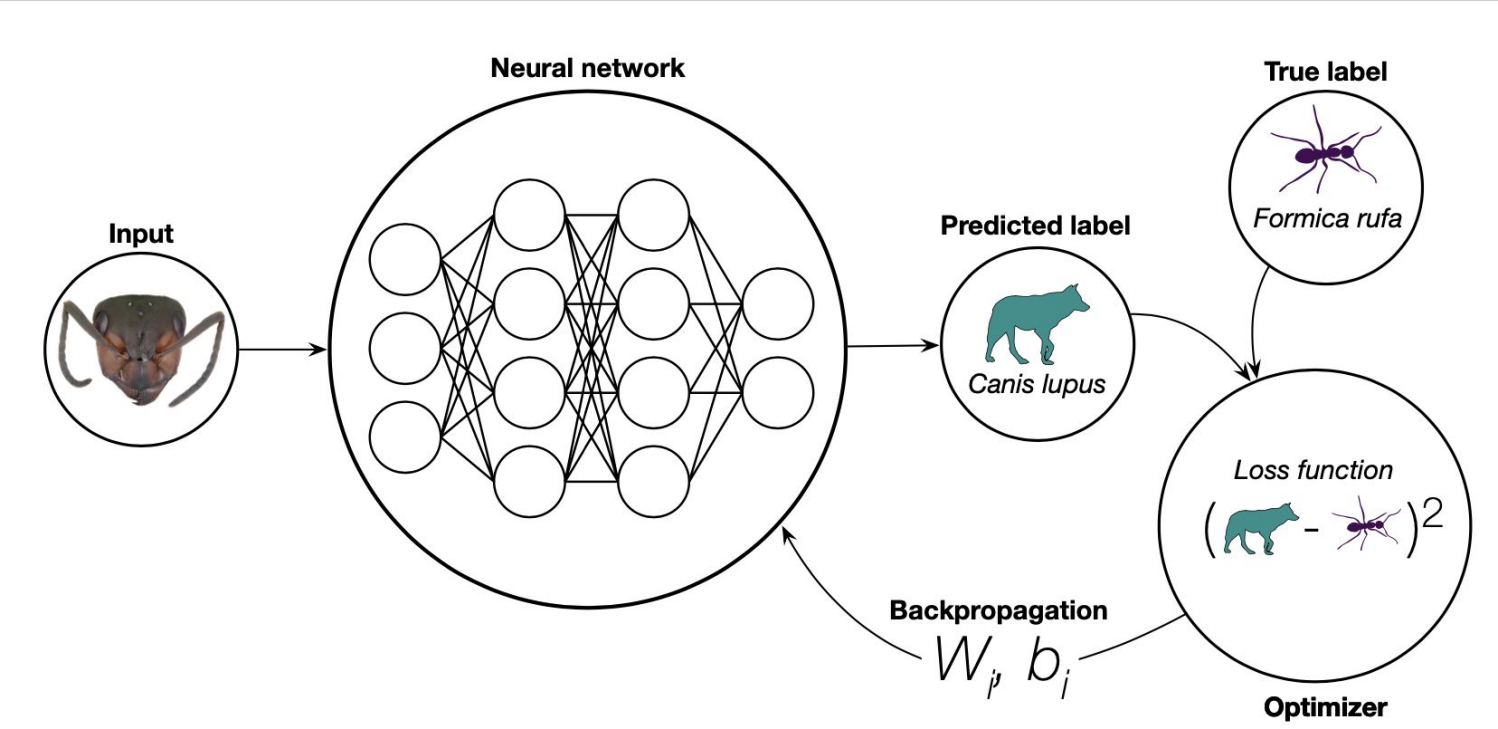
# Tabular data

https://www.kaggle.com/c/titanic

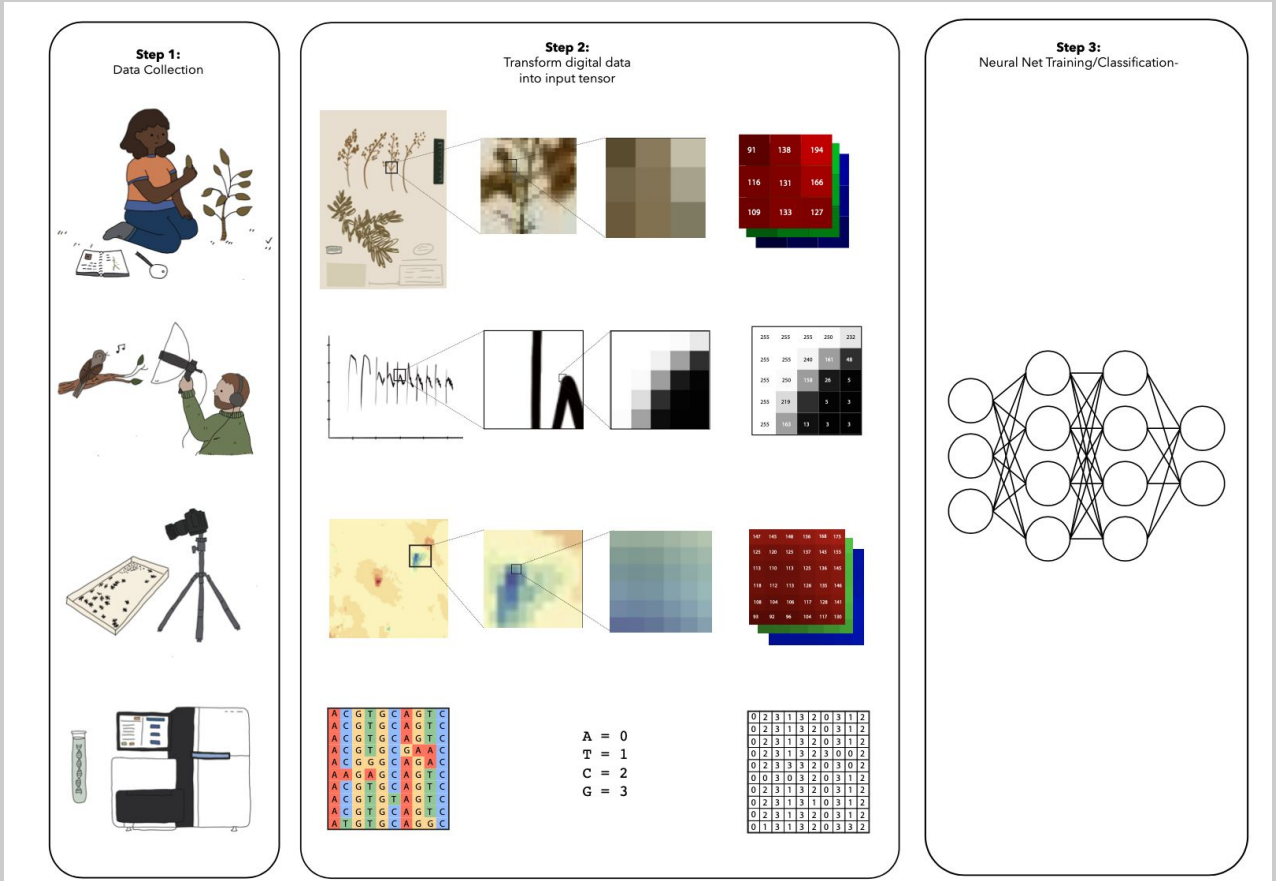
| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embar |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | NaN | S |
| 1 | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 2 | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | NaN | S |
| 3 | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C123 | S |
| 4 | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | NaN | S |

A classic Kaggle challenge: predicting the fates of passengers on the Titanic.

# Images (computer vision)



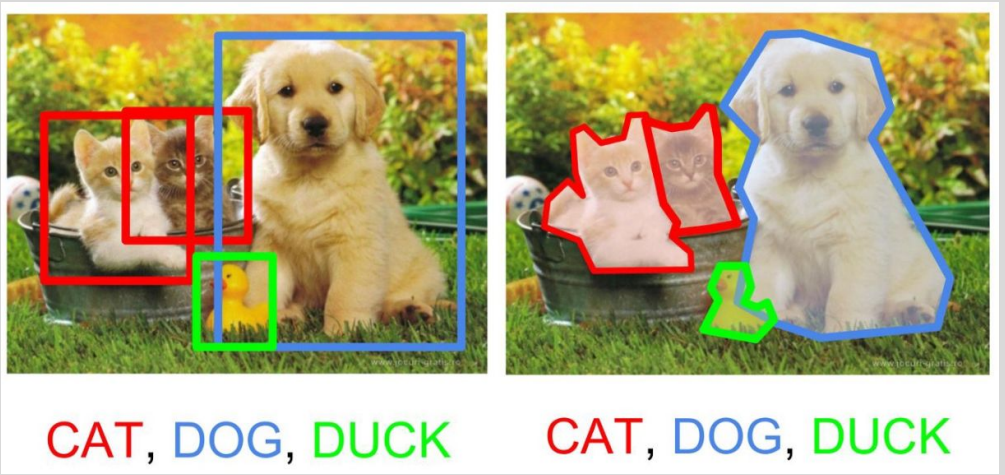

Borowiec et al., 2021. Deep learning as a tool for ecology and evolution.
DOI: 10.32942/osf.io/nt3as




Lots of different data types can be converted to images for ML model building.

Borowiec et al., 2021. Deep learning as a tool for ecology and evolution.
DOI: 10.32942/osf.io/nt3as

# Computers can learn by looking at many, many examples.


# 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% |

# Smithsonian AI Values Statement

https://datascience.si.edu/ai-values-statement

- Actions
  - Documentation of dataset biases.
  - Documentation of intent in methodologies, datasets, collections, algorithms or tools.
  - Creator positionality statements.
  - Documentation of known or potential risks.
  - Solicitation and inclusion of feedback from relevant members of the community.
- Partnerships
  - Seek projects and partnerships that adhere to our institutional values.
  - Partnerships should avoid using tools with unspecified or undisclosed methods and biases.
  - Partnerships involving AI and machine learning tools should explicitly evaluate and state if the datasets or data descriptions used in these tools was collected without consent, or contains offensive or racist descriptions before we agree to use these tools.
- Community
  - If any community or individual is harmed by the use of a technology, then that is one too many.

# One aspect of our AI Values Statement implementation: Dataset cards

https://github.com/Smithsonian/dataset-cards

https://github.com/huggingface/datasets/blob/main/templates/README_guide.md

Why should you use dataset cards?

- A mechanism to communicate and provide context on a dataset
  - Original intent for gathering dataset
  - Context
  - Assumptions
  - Changes to the data, normalizations, transformations
  - Known biases and social impact
- Can be used for both internal only as well as external/public datasets
- Allows creators to state information up front about dataset, prior to any use in AI or other algorithms

# Machine learning in GLAM* Carpentries lesson
# *Galleries, libraries, archives, and museums

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/

# Examples of machine learning tasks using herbarium specimens:

- Segmentation
- Species identification
- Morphospace exploration
- Trait extraction


## Segmentation:

Can include removing background, labels, isolating/labeling plant pixels, or multiple tissue types (e.g. fruits, flowers, leaves).

Is often the first step needed before building additional models to identify species or extract traits.

# White et al., 2020 built a custom segmentation model.
# First, high-resolution masks were produced as training data.

# 400 of these “ground-truth masks” were used to train a U-Net:

## U-Net: Convolutional Networks for Biomedical Image Segmentation

Olaf Ronneberger, Philipp Fischer, and Thomas Brox

Computer Science Department and BIOSS Centre for Biological Signalling Studies,
University of Freiburg, Germany
ronneber@informatik.uni-freiburg.de,
WWW home page: http://lmb.informatik.uni-freiburg.de/

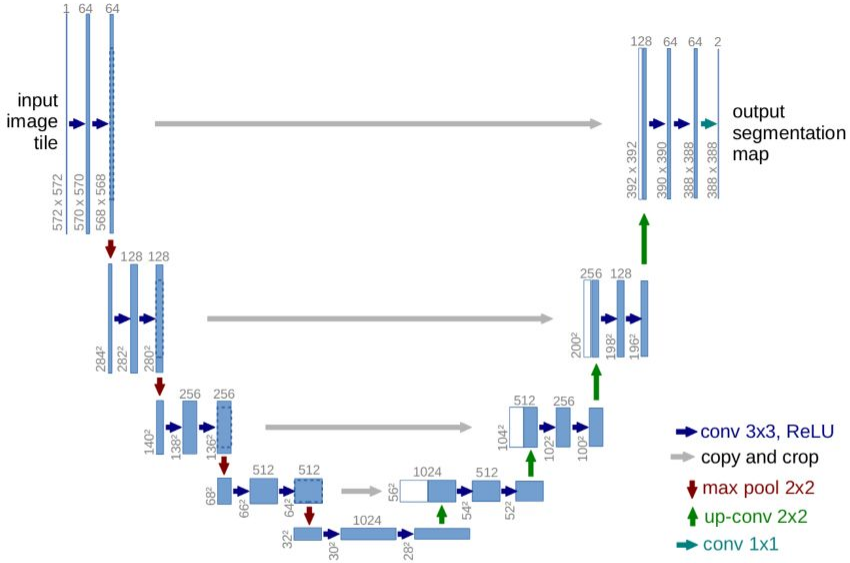
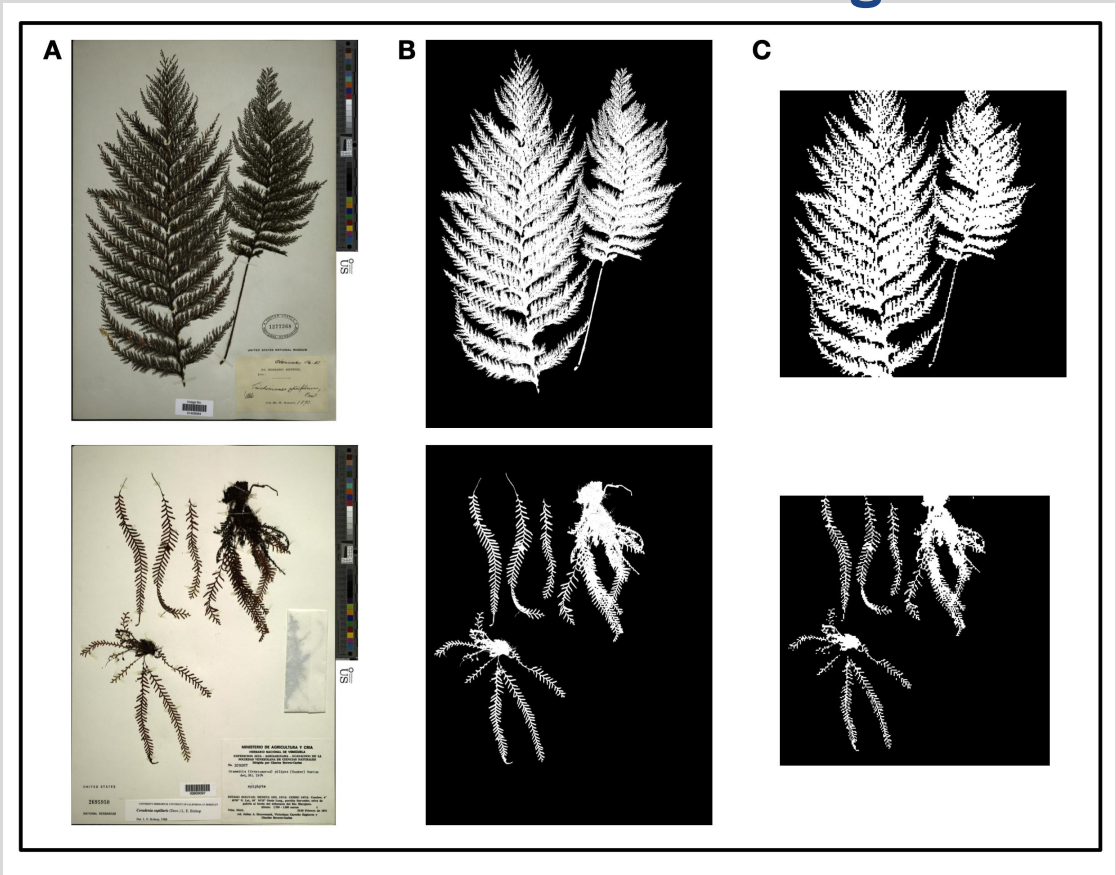



**Fig. 1.** U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

# Results of U-net training




White et al., 2020, *Applications in Plant Sciences*

## Paper, code, model, and data available:

White et al., 2020: https://doi.org/10.1002/aps3.11352

https://github.com/sidatasciencelab/fern_segmentation

Original images (https://doi.org/10.25573/data.9922148)
Curated masks (https://doi.org/10.25573/data.9922232)
Metadata (https://doi.org/10.25573/data.11771004)

# How can we scale this work across collections?

- With U-net masked specimen images, we can now include specimens from any herbarium as training data

- We have built a training data set of >800,000 images from fern and fern ally collections across the world to ask questions about global geographic patterns

Alex White


# Species identification:

Images with their species labels can be used to train a supervised CNN, which can then label images of unknown species with high accuracy.

# Method:

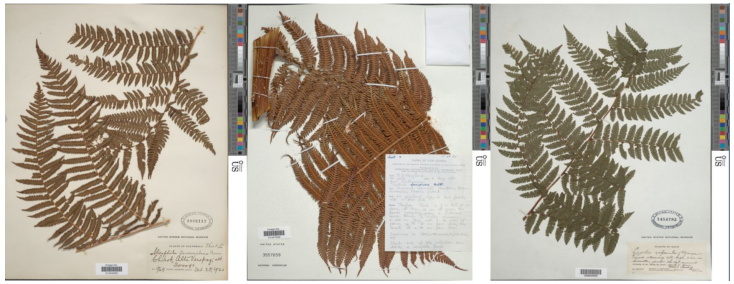
Build a convolutional neural net and train it to label fern specimens:

- by genus
  - >500 specimens (86 genera)
  - >50 specimens (269 genera)
- by species
  - >50 specimens (1425 species)

80% of data used for training neural network
20% of data set aside for validation

**Alex White**

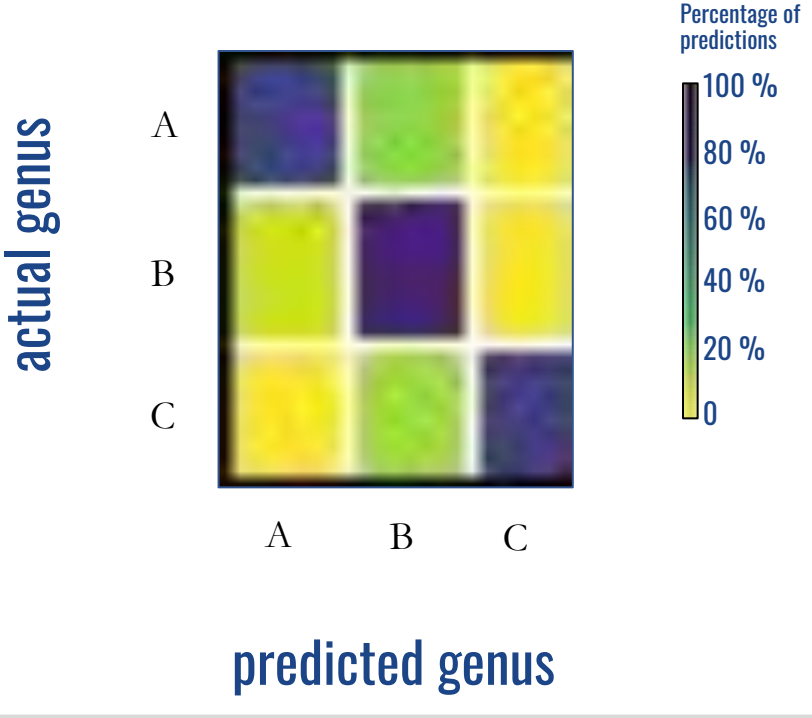
# We validate the model by feeding images with known labels through the network



example genus A

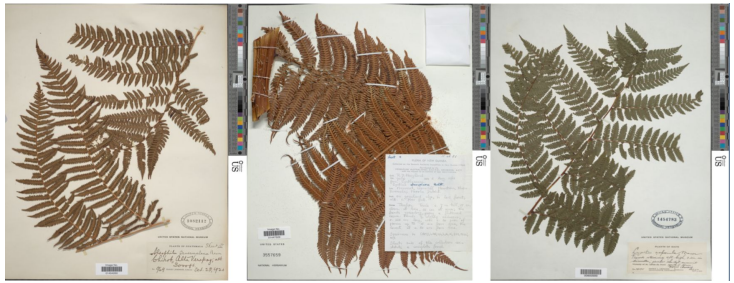
example genus B

example genus C

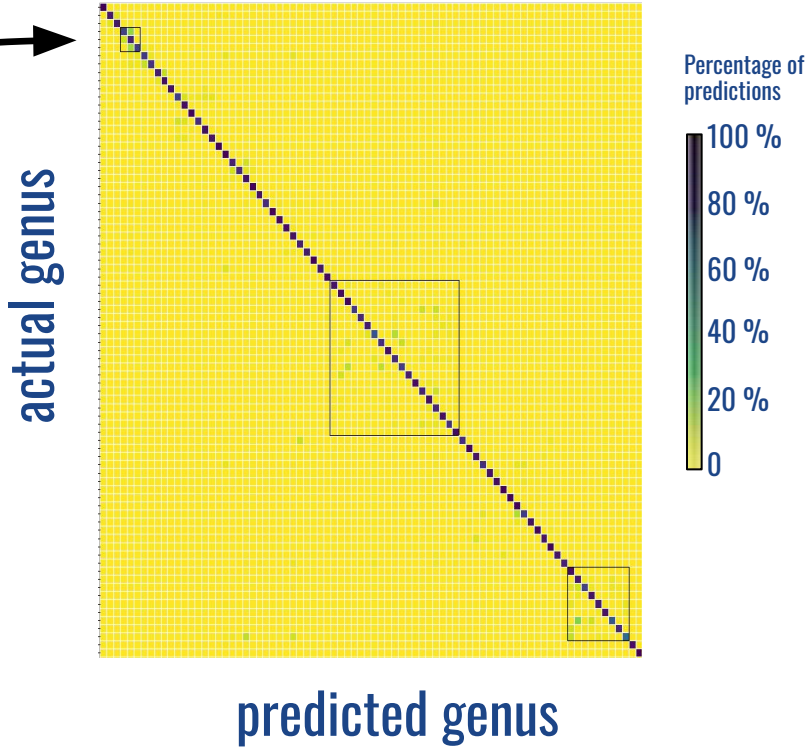


# FernNet is 97% accurate at genus ID

3 genera in the tree fern family Cyatheaceae

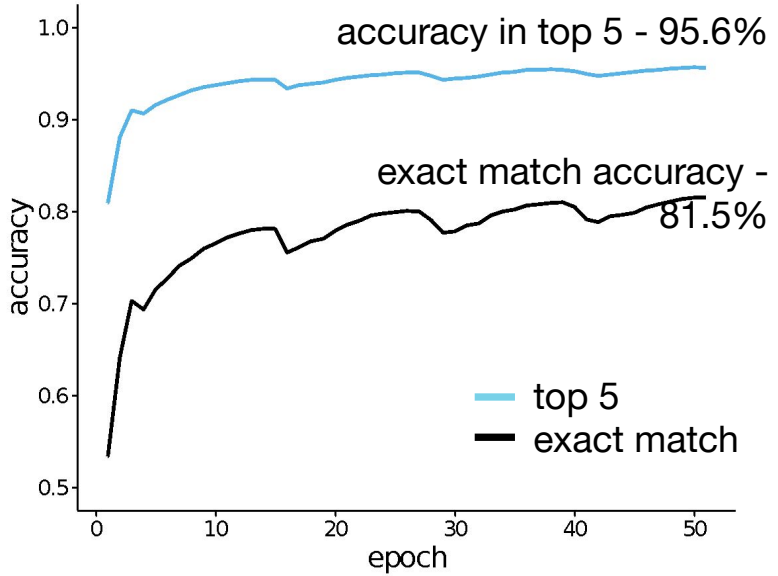


Boxes contain examples of genera within the same family

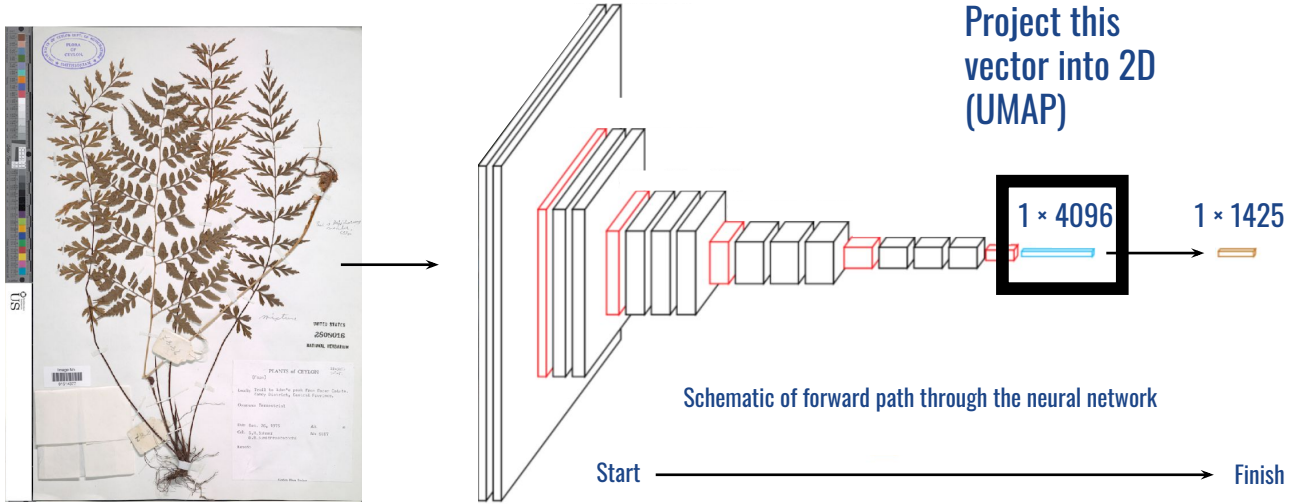



Confusion is most often between closely related genera

# FernNet is highly accurate for species ID (1425 species)

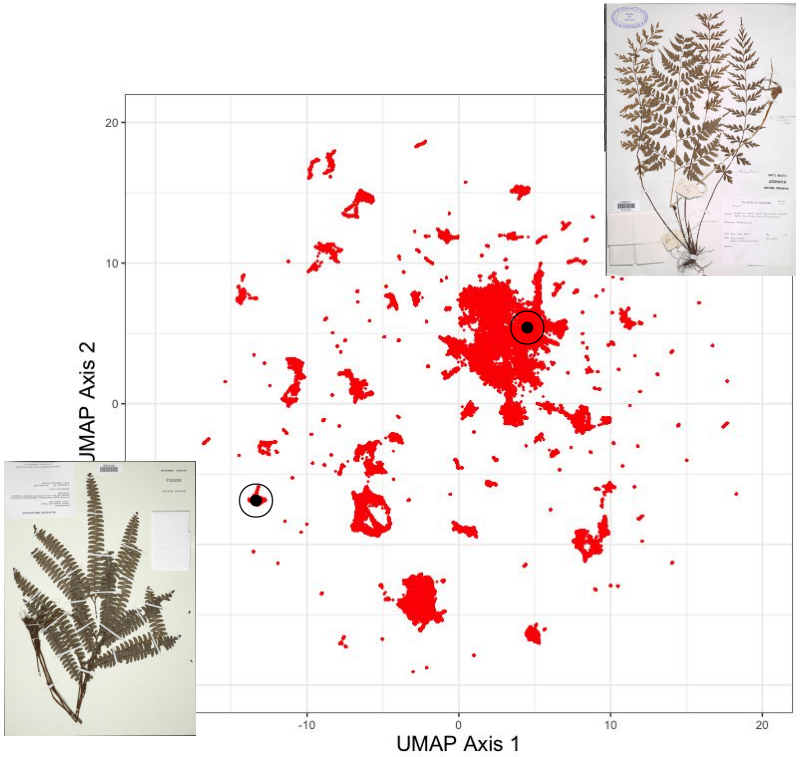


# First approach to dimensionality reduction and visualization of shape space:




Simonyan and Zisserman 2014

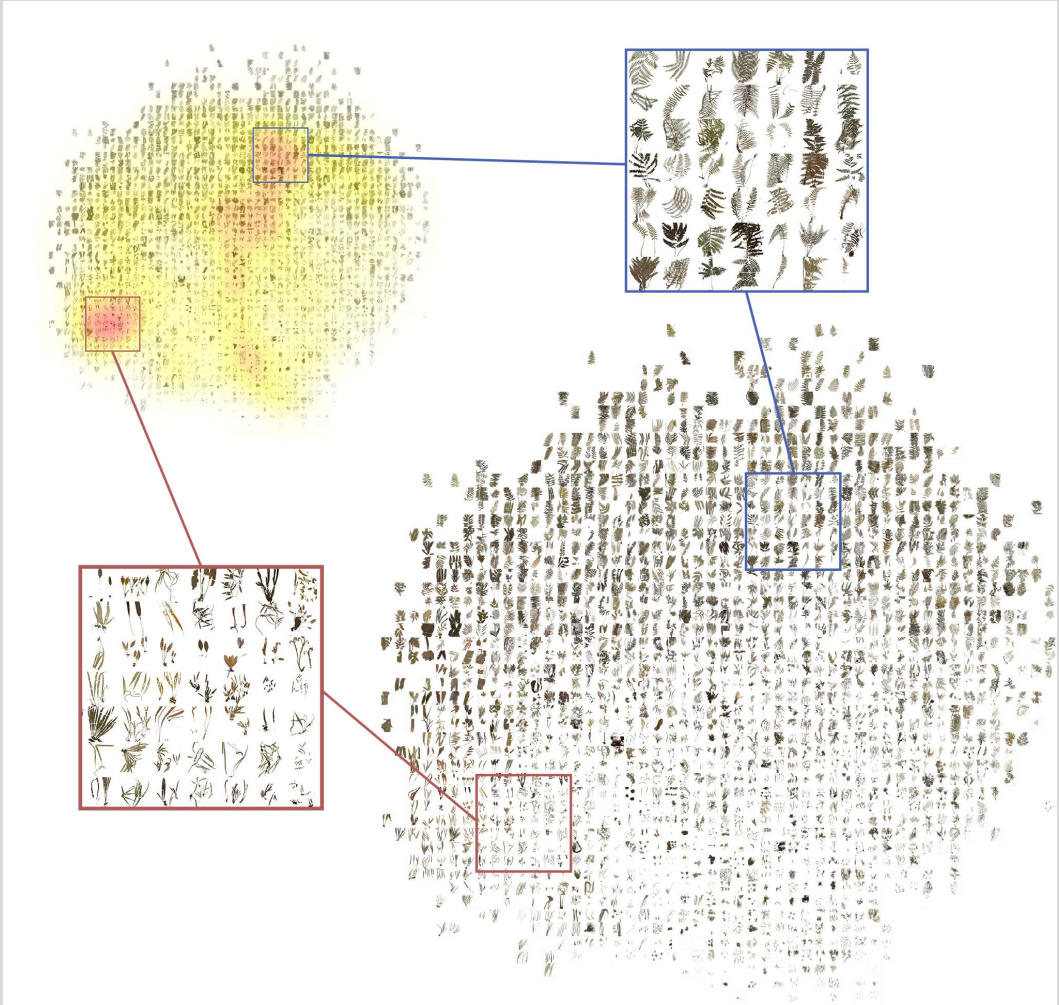
# Each specimen is summarized by a set of bivariate coordinates



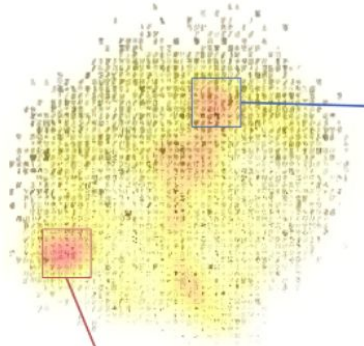
# 2nd approach to DR: ivis (Szubert *et al.* 2019)

- twin neural network approach that preserves local and global distances

# Overlay of species richness heatmap on shape space occupation



*Elaphoglossum* (n = 1763)



**more-divided**
**+**
**larger**

Cyatheaceae (n = 728)

**less-divided**
**+**
**smaller**

# Trait extraction:

Rapid scoring of specimens and measurement of specific structures can allow researchers to study large-scale patterns of phenology and morphological variation that would be impossible if specimens need to be manually scored and measured.

# Trait extraction: check out LeafMachine2

https://github.com/Gene-Weaver/LeafMachine2

https://www.leafmachine.org