Strengths of machine learning

GLAM data types suited for ML:

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

Tabular data

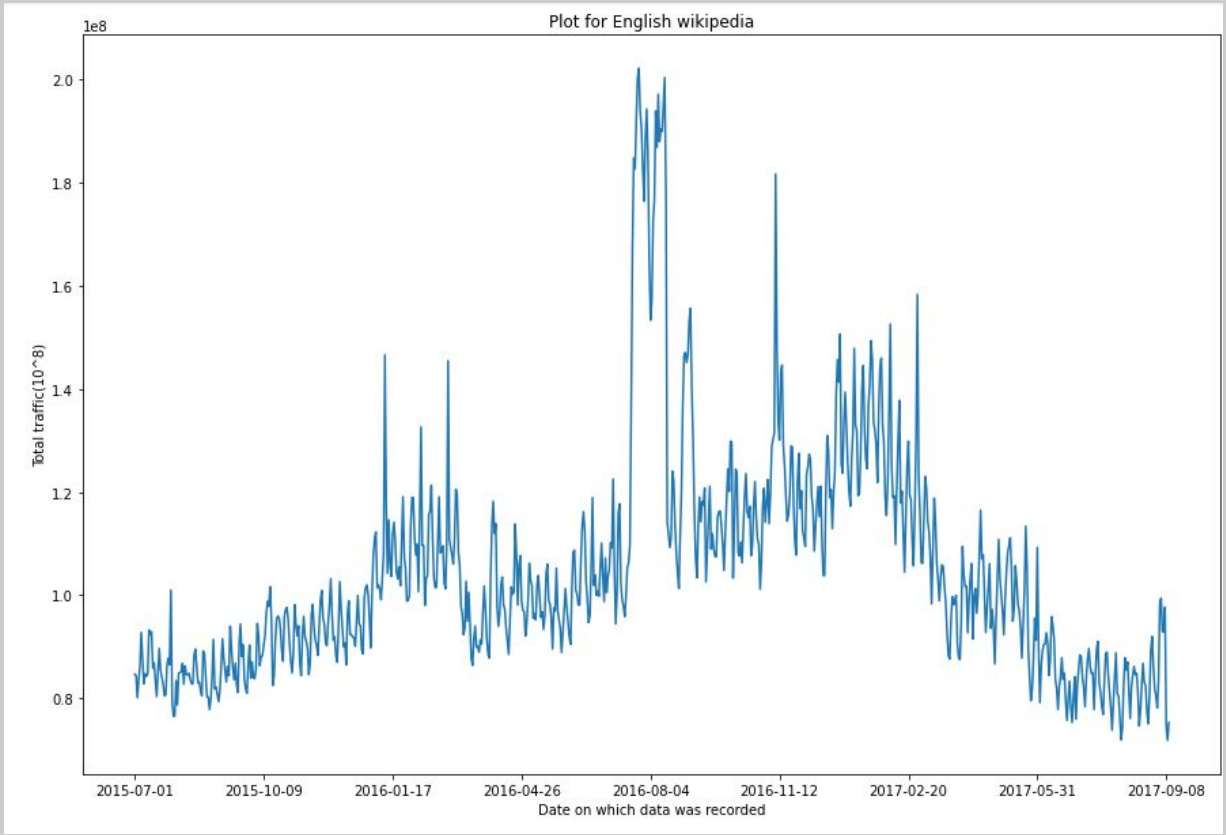
https://www.kaggle.com/c/titanic

	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.

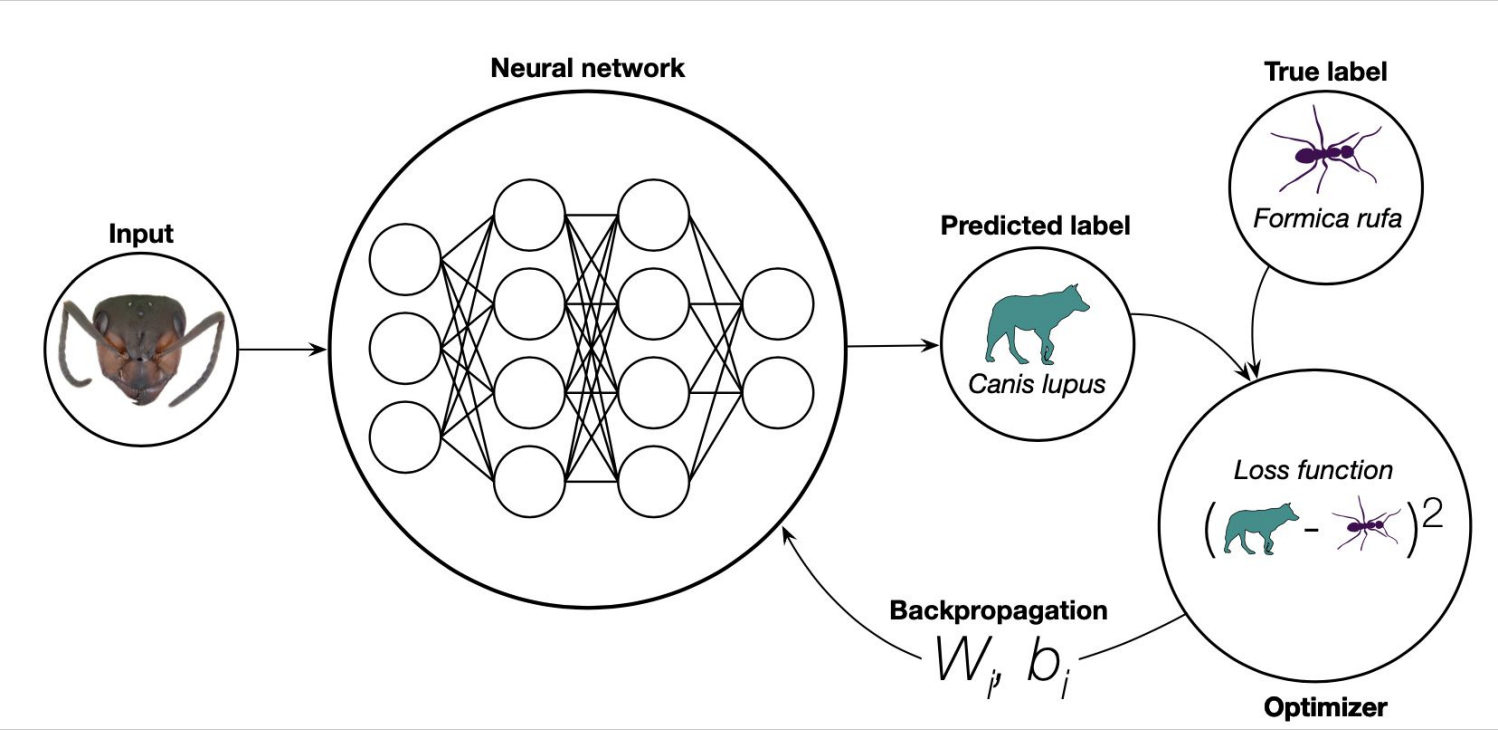
Tabular data

Time series data

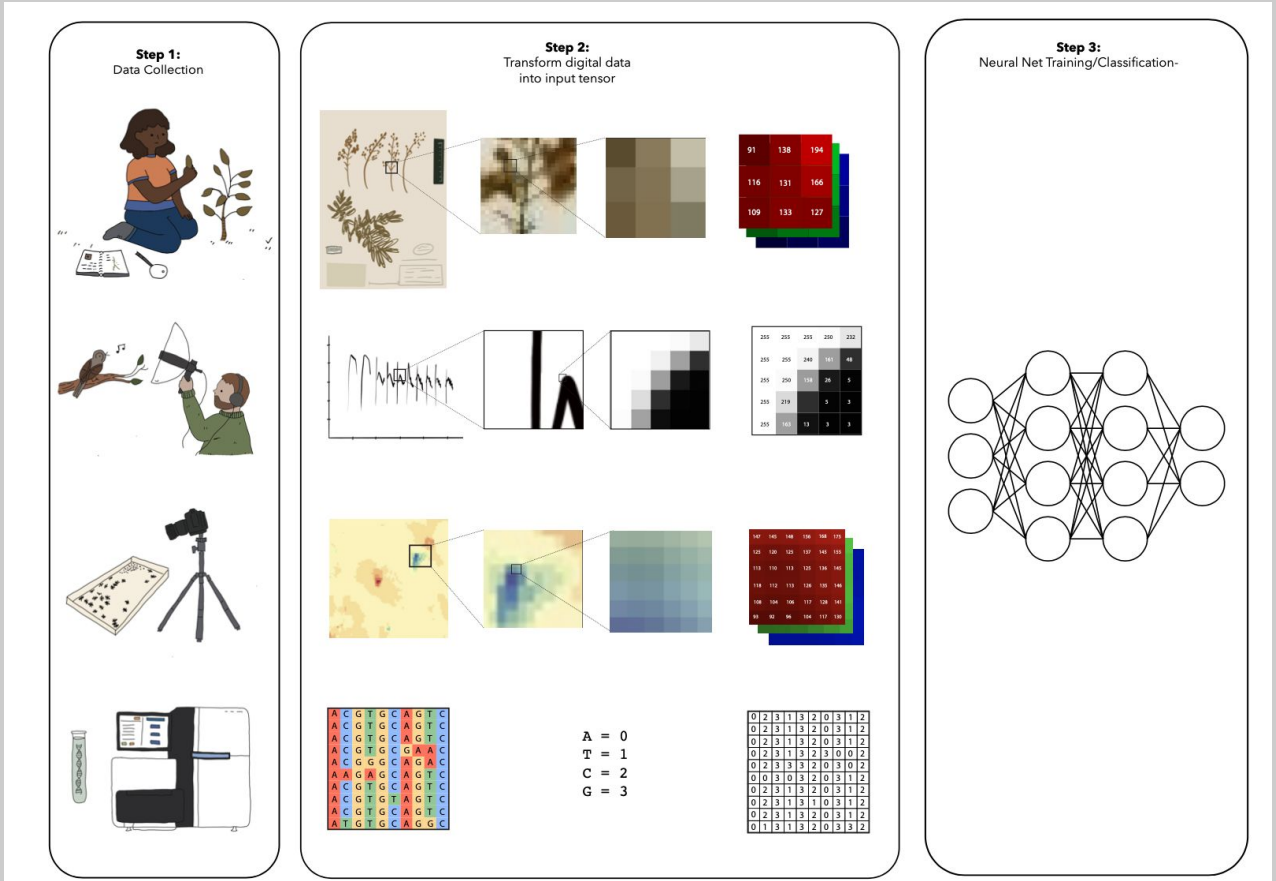


A classic Kaggle challenge: predicting future web traffic from Wikipedia.

Images (computer vision)



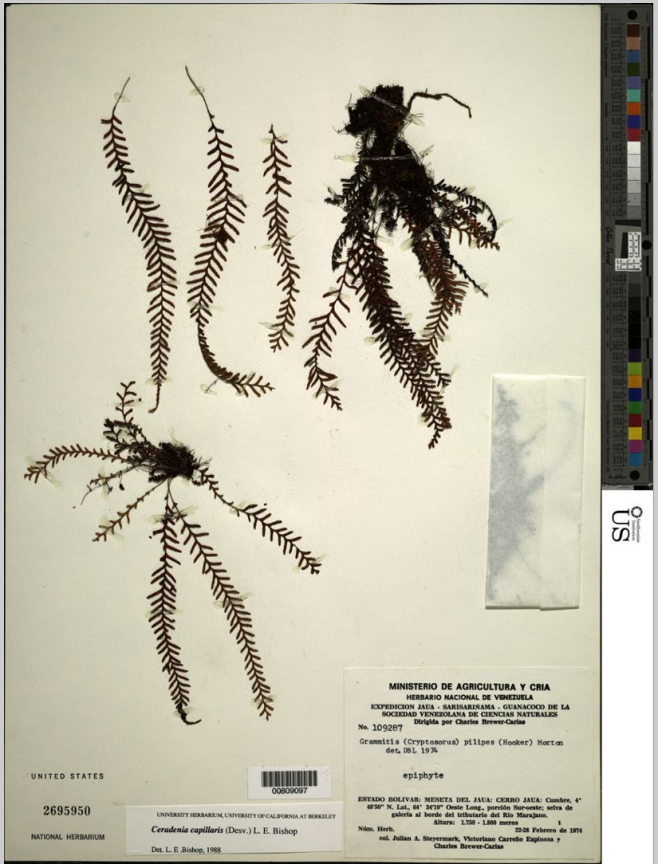
Borowiec et al., 2021. Deep learning as a tool for ecology and evolution.
DOI: 10.32942/osf.io/nt3as



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

Images (computer vision)



Smithsonian Institution

NATIONAL MUSEUM of NATURAL HISTORY

Smithsonian DIGITIZATION

Images (computer vision)

Examples of applications of ML in Botany:

-classification

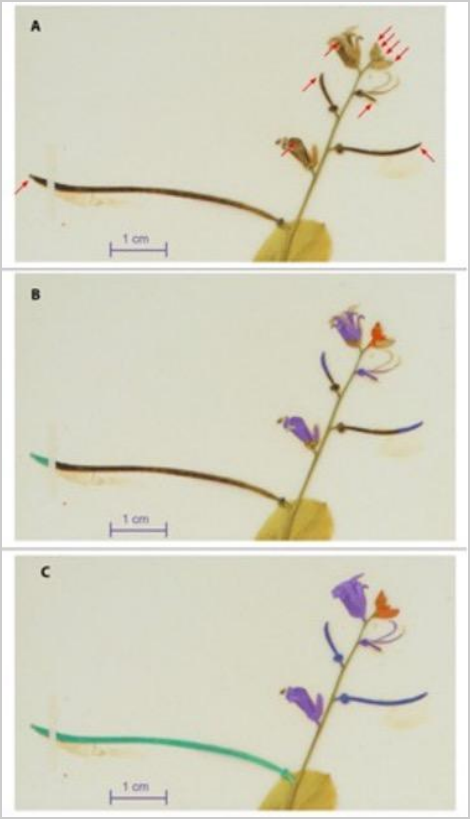
-which species is this?

-object detection

-how many flowers/fruits are there and where are they?

-is there evidence of insect damage on this specimen?

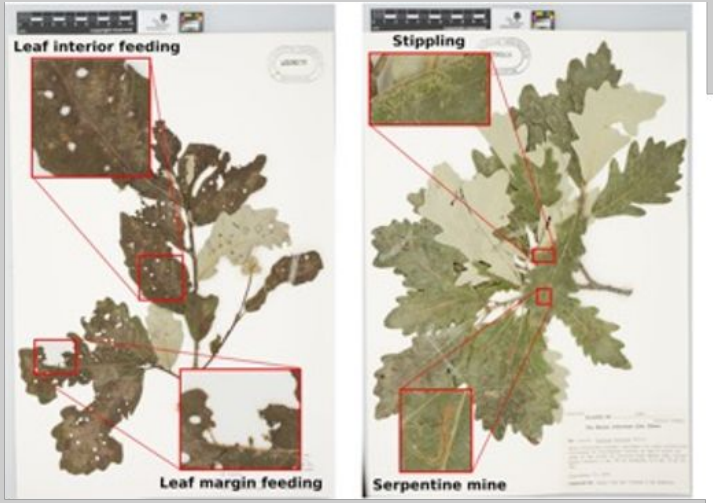
Images (computer vision)

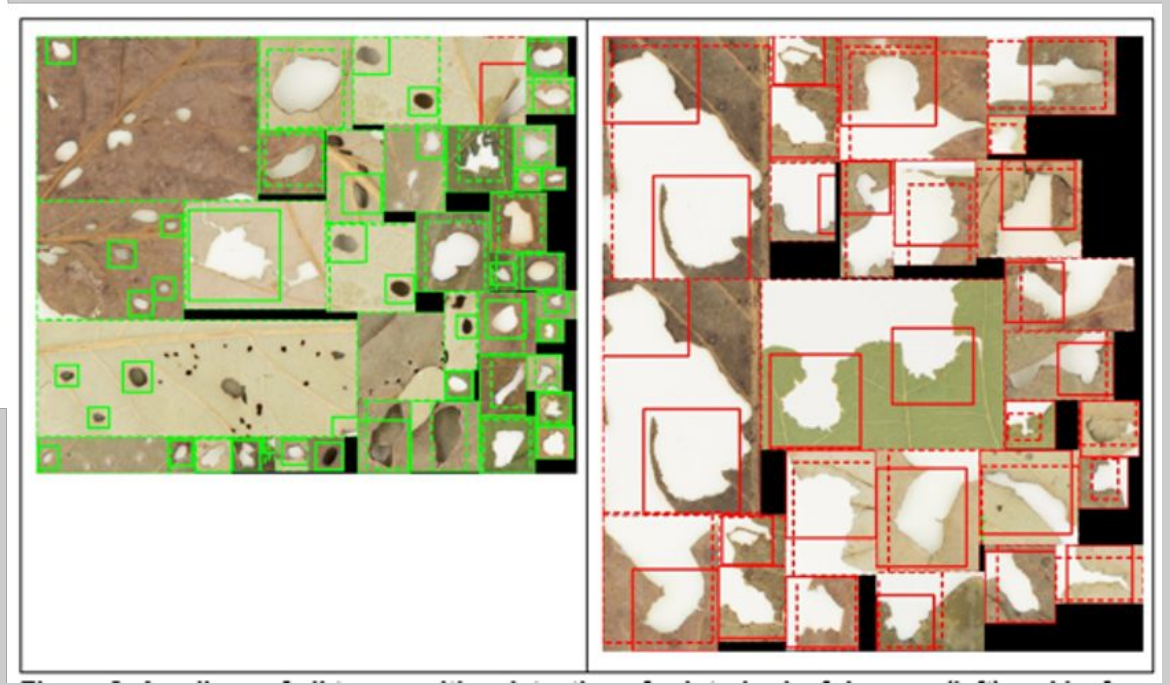


Automate the detection, segmentation, classification of reproductive structures flower buds, flowers, immature fruits, and mature fruits.

Goeau et al. 2020 *APPS*

Images (computer vision)

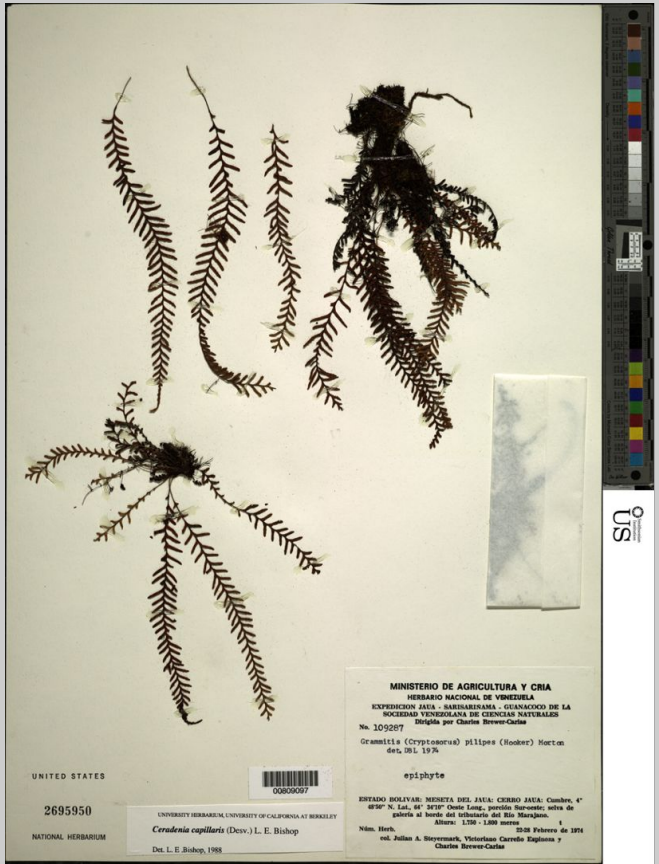




Detect the type and extent of herbivory.

Meineke et al. 2020 *APPS*

USNM Herbarium Project

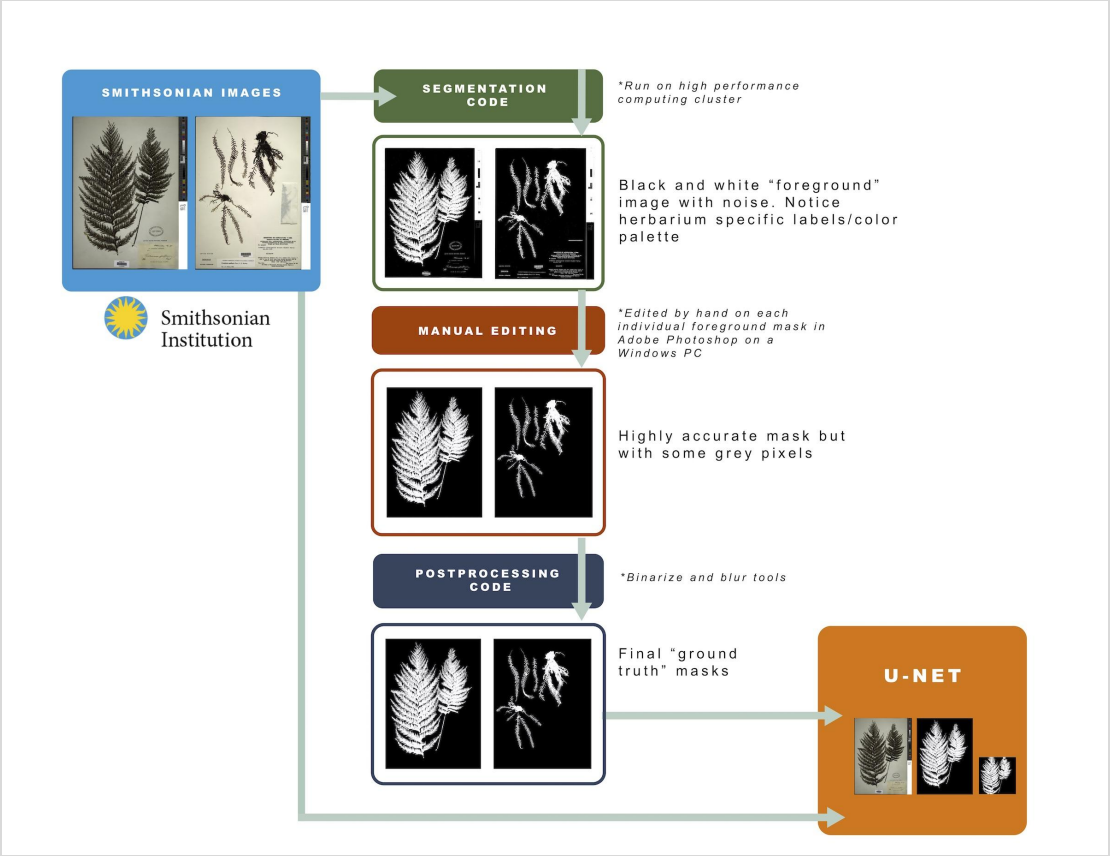


Smithsonian Institution

NATIONAL MUSEUM of NATURAL HISTORY

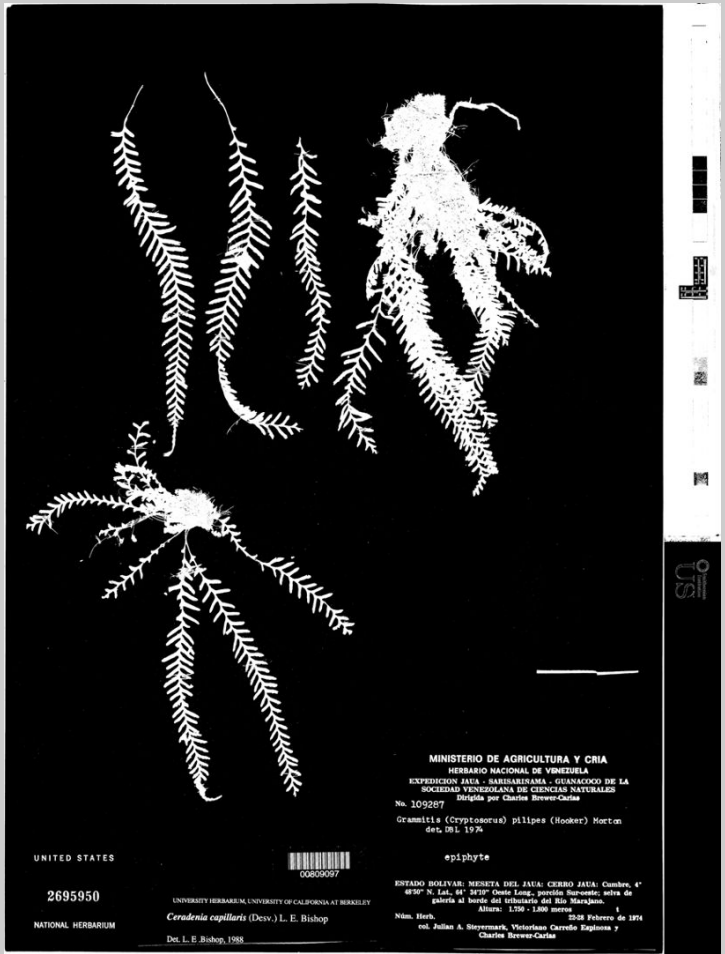
Smithsonian DIGITIZATION

Workflow to produce U-net to mask herbarium images



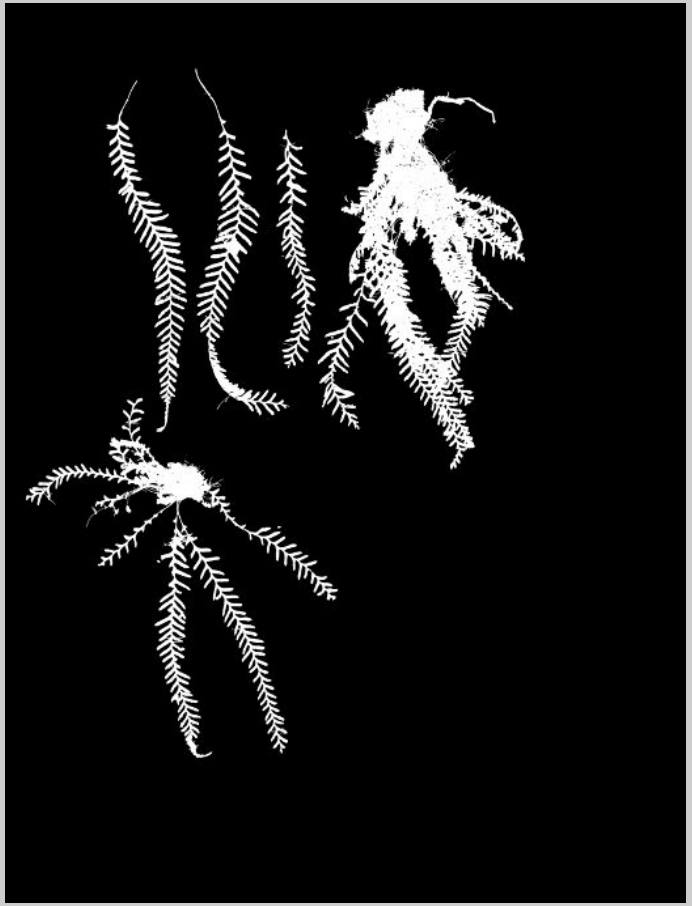
White et al., 2020, *Applications in Plant Sciences*

After running segmentation code (built using PlantCV and OpenCV):



After manual processing to remove any residual non-plant material:

These processed images are called masks: images of identical resolution that define the identity of each pixel in the original image.



High-resolution masks produced as training data

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

400 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/

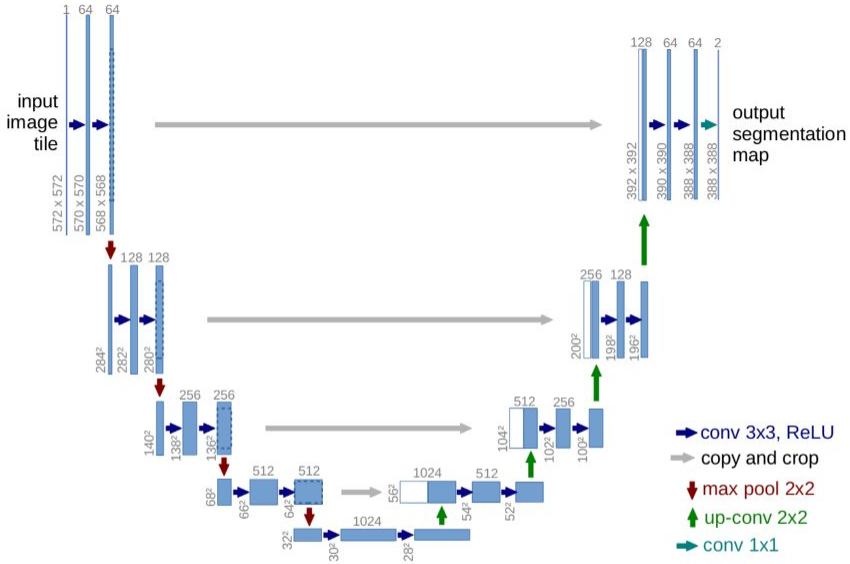
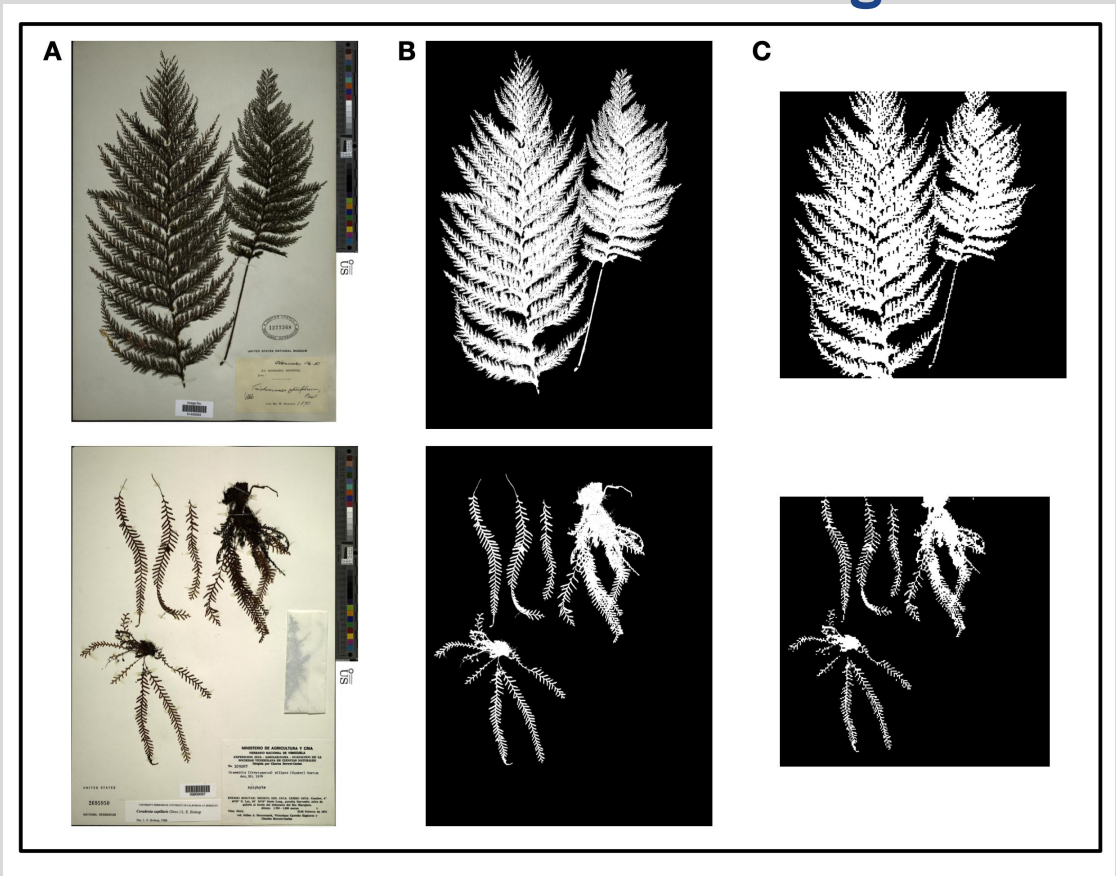


Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows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

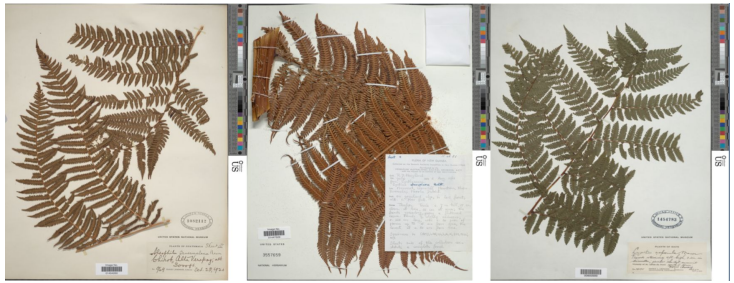
Results of U-net training



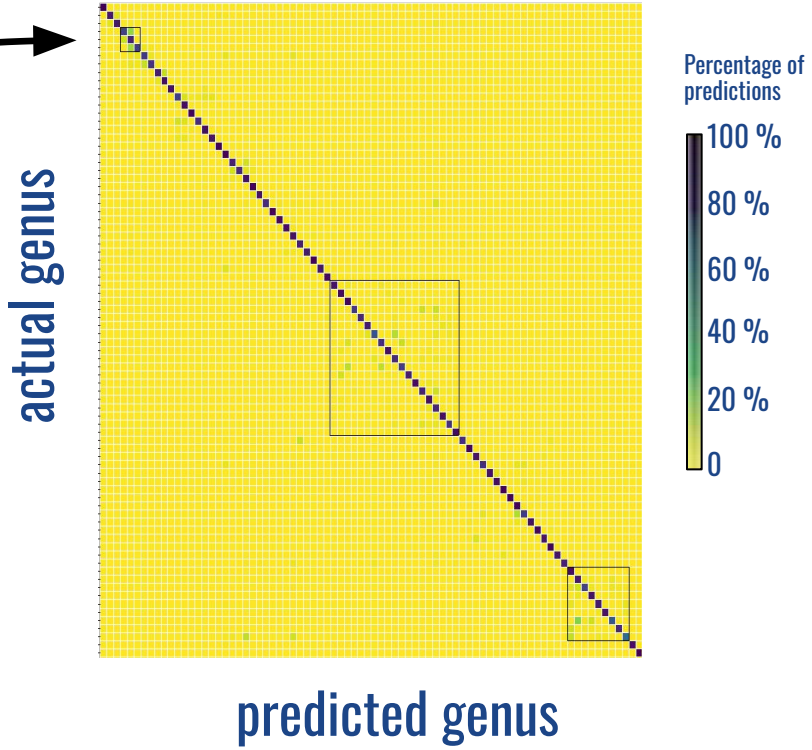
White et al., 2020, *Applications in Plant Sciences*

FernNet is 97% accurate at genus ID

3 genera in the tree fern family Cyatheaceae

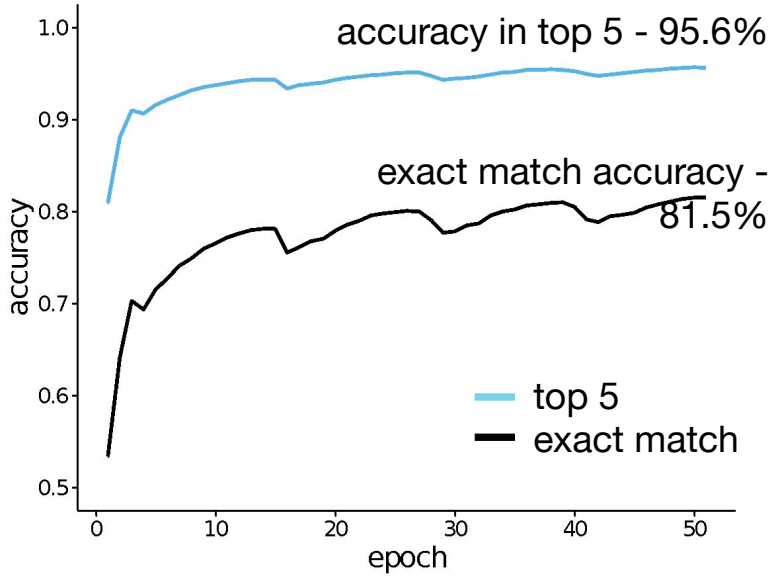


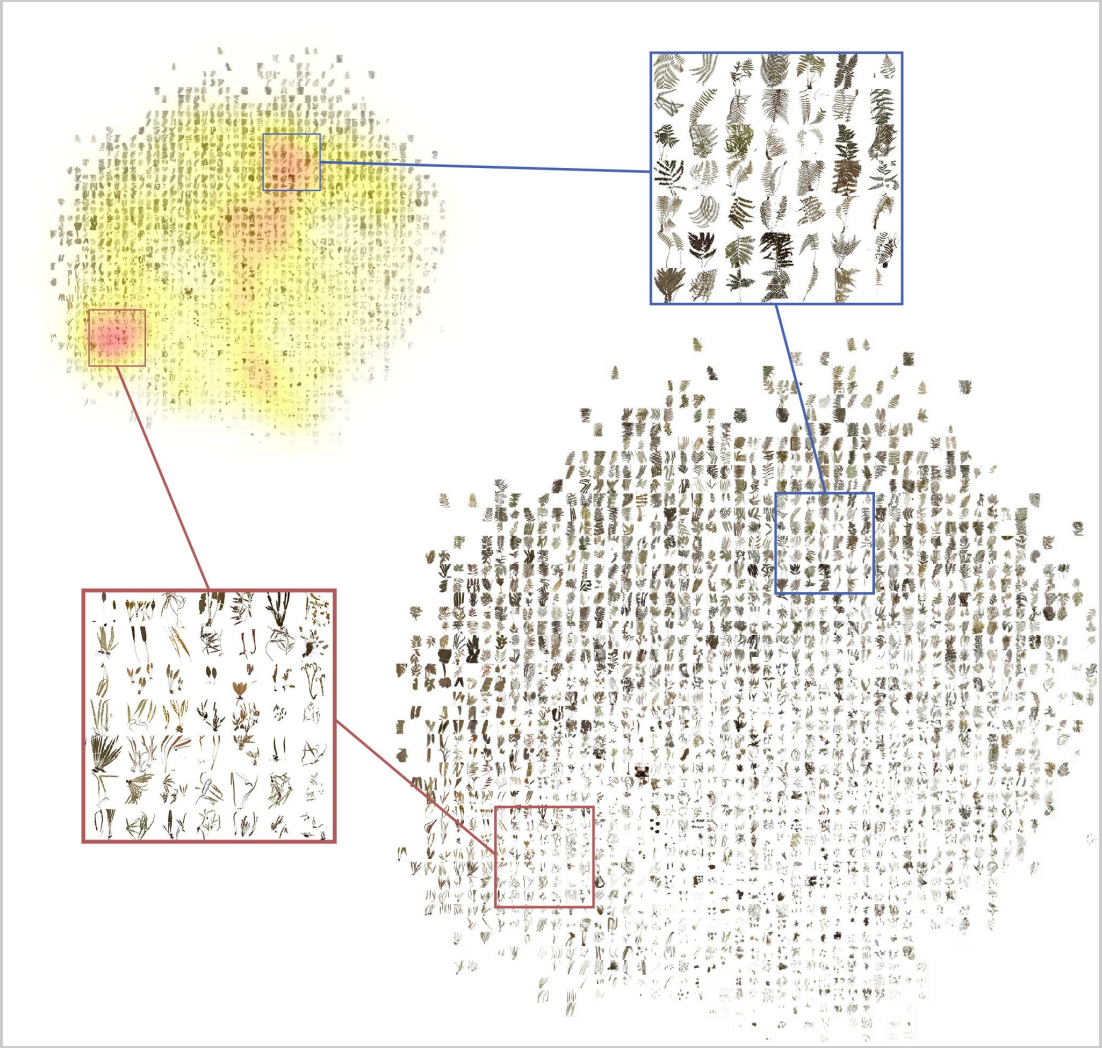
Boxes contain examples of genera within the same family



Confusion is most often between closely related genera

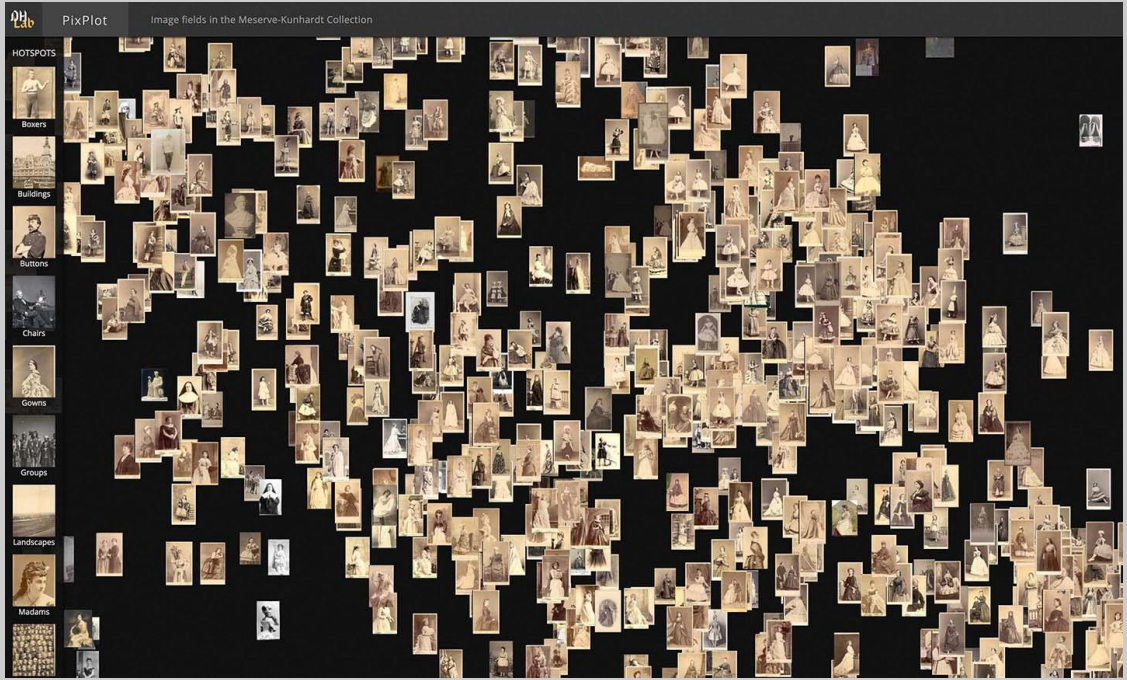
FernNet is highly accurate for species ID (1425 species)





Species classification model can be used to explore shape space occupation.

Alex White, postdoctoral fellow



Feature vectors from pretrained models can be used to cluster new data, e.g. in PixPlot.

PixPlot: https://dhlab.yale.edu/projects/pixplot/

Breakout activity:

Now that you have learned the types of computer vision tasks where machine learning excels, what are some things you might try to do with this image?

Text (natural language processing)

Traditional NLP: “bag of words”

Segment a document into words, count frequency (disregards grammar and word order).

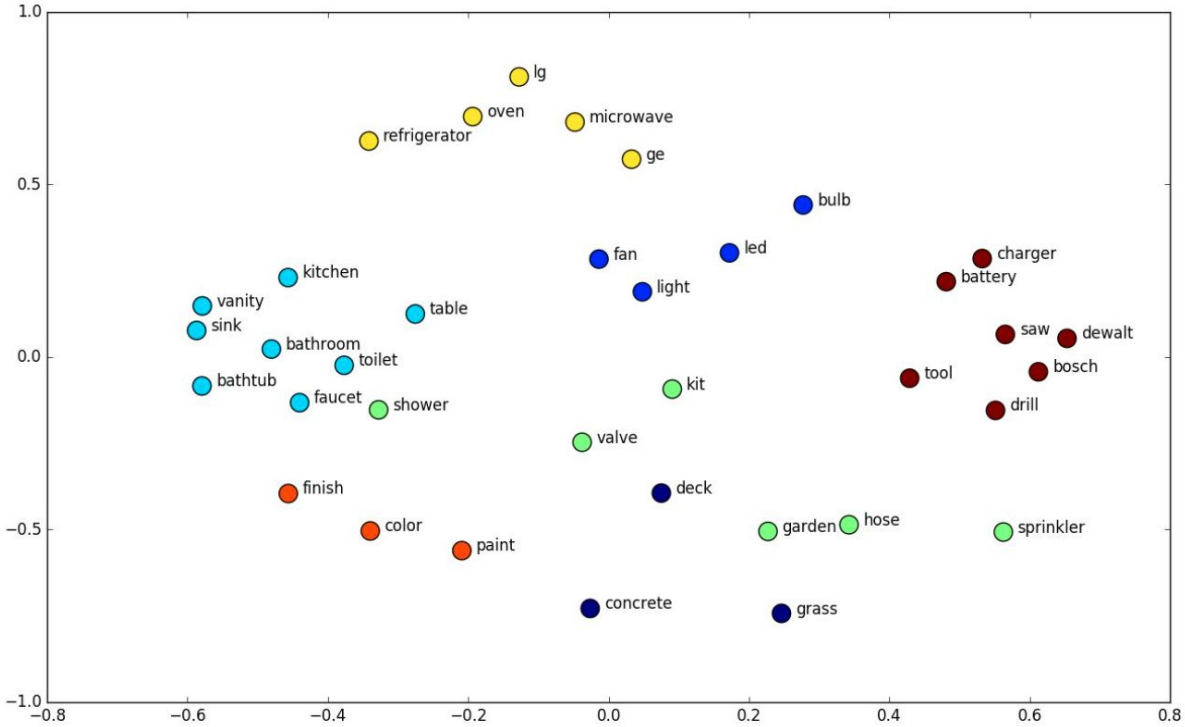

wikipedia
people
foundation-l
wikimedia
one
think
like
foundation
project
projects
community
lists.wikimedia.org
board
language
make
use
need
much
good
work
way
may
see
just
also
time
know
new
even
now
get
free
content
well
many
page
want
wmf
discussion
something
articles
english
pm
first
policy
article
languages
right
really
images
public
wiki
sure
issue
information
thanks
going
help
different
point
erik
copyright
support
wikimedia.org
might
things
anyone
however
name
things

Text (natural language processing)

Deep NLP: e.g. word embeddings

Computers can learn how words are used in context. We can use texts as training data to assign vectors to words. Words closer in the vector space are expected to have similar meanings.

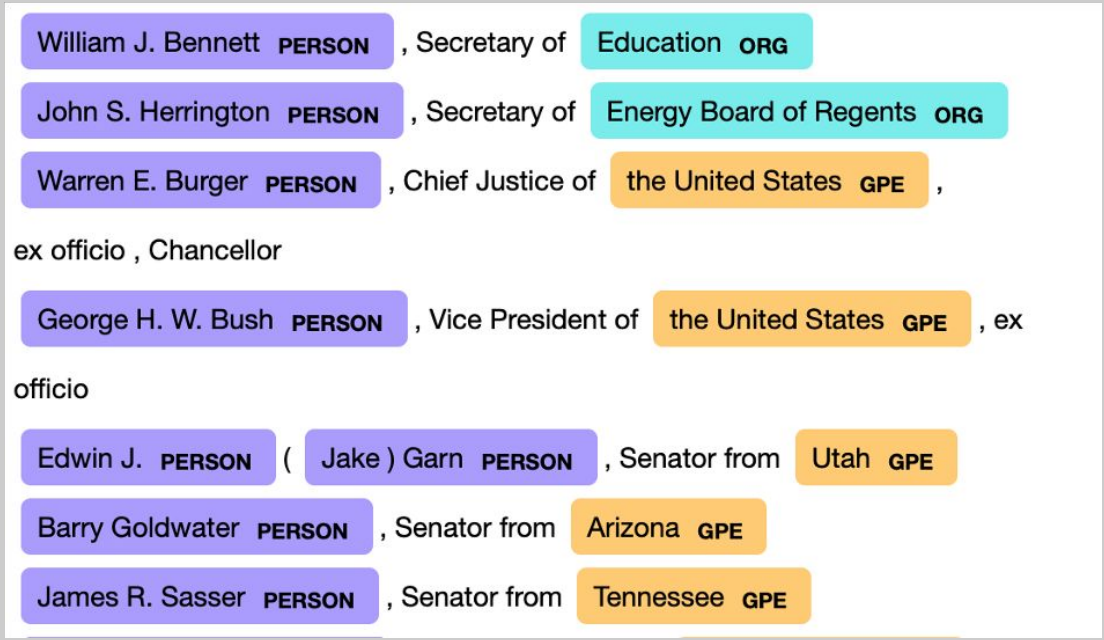
Word embeddings can be built for specific datasets.



Example 2D word embedding space, where similar words are found in similar locations. (src: http://suriyadeepan.github.io)

Text (natural language processing)

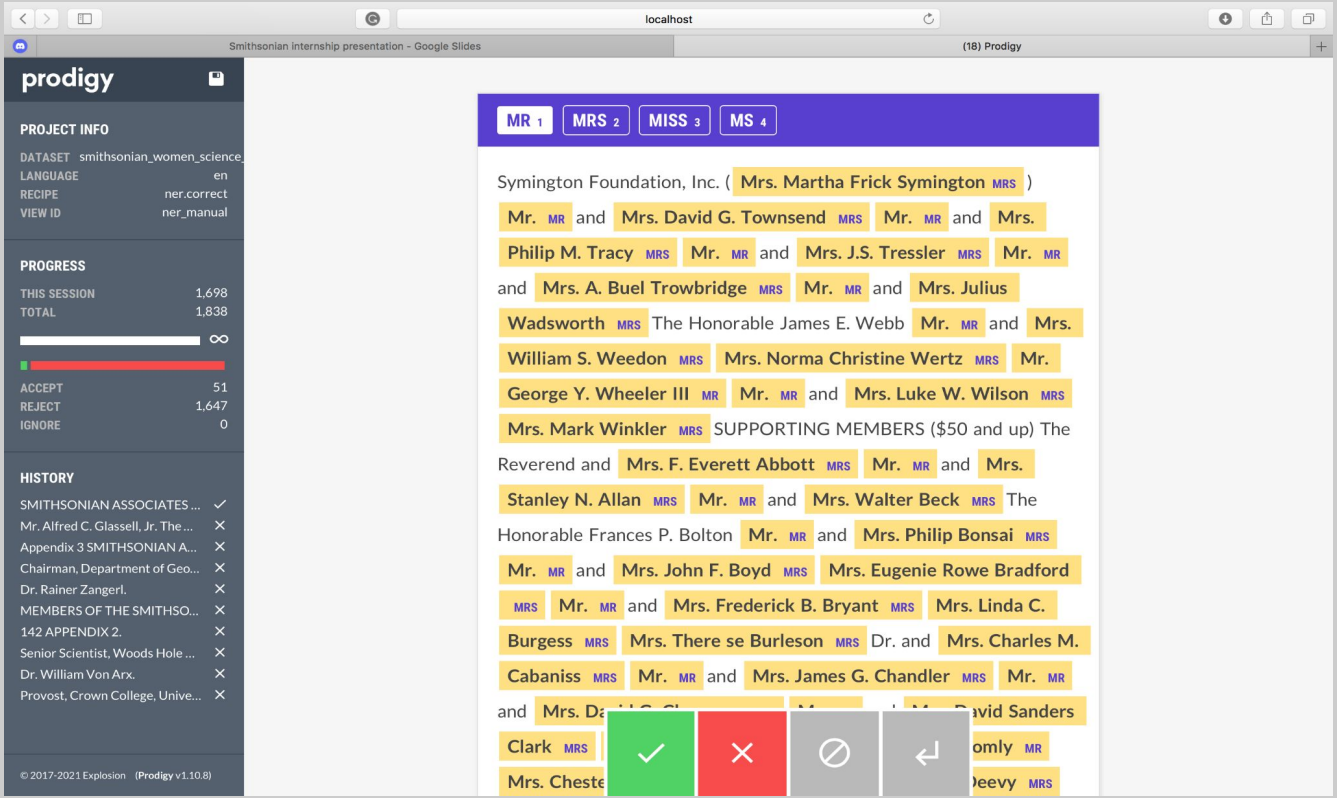
NER - Named Entity Recognition

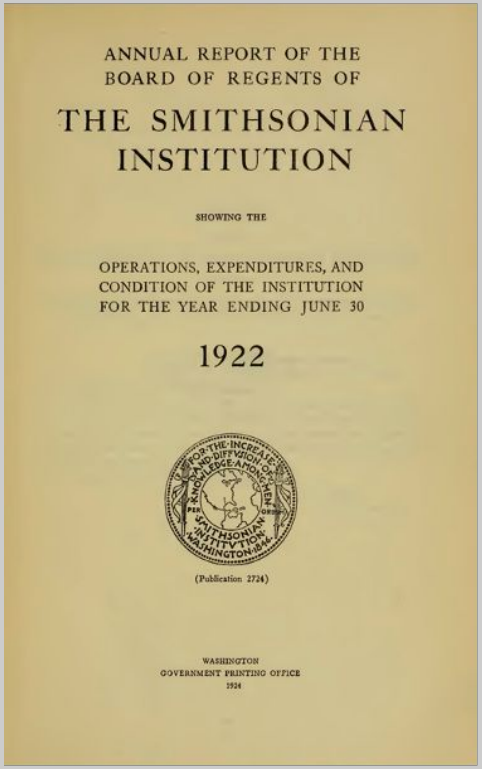


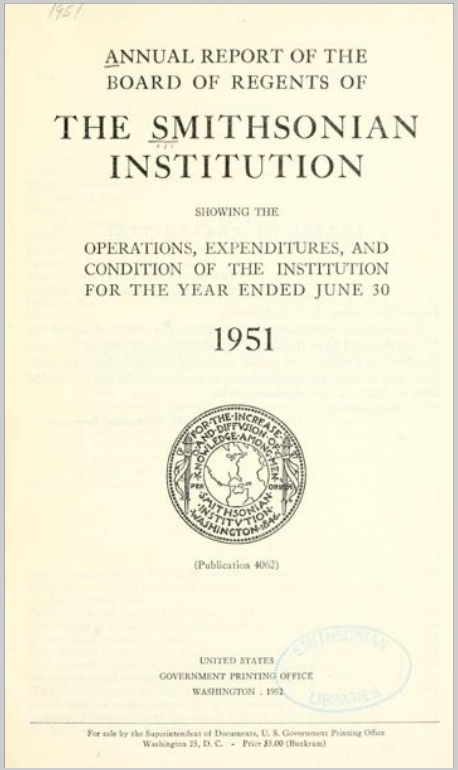
1985 Smithsonian Annual Report - list of the Board of Regents members

Named Entity Recognition

Custom models are often necessary (e.g. the Mrs. problem)







Breakout activity:

Now that you have learned about the strengths and weakness of natural language processing, brainstorm some applications to Smithsonian data.