

Strengths of machine learning

GLAM data types suited for ML:

Tabular data

Images (computer vision)

Text (natural language processing)

Tabular data

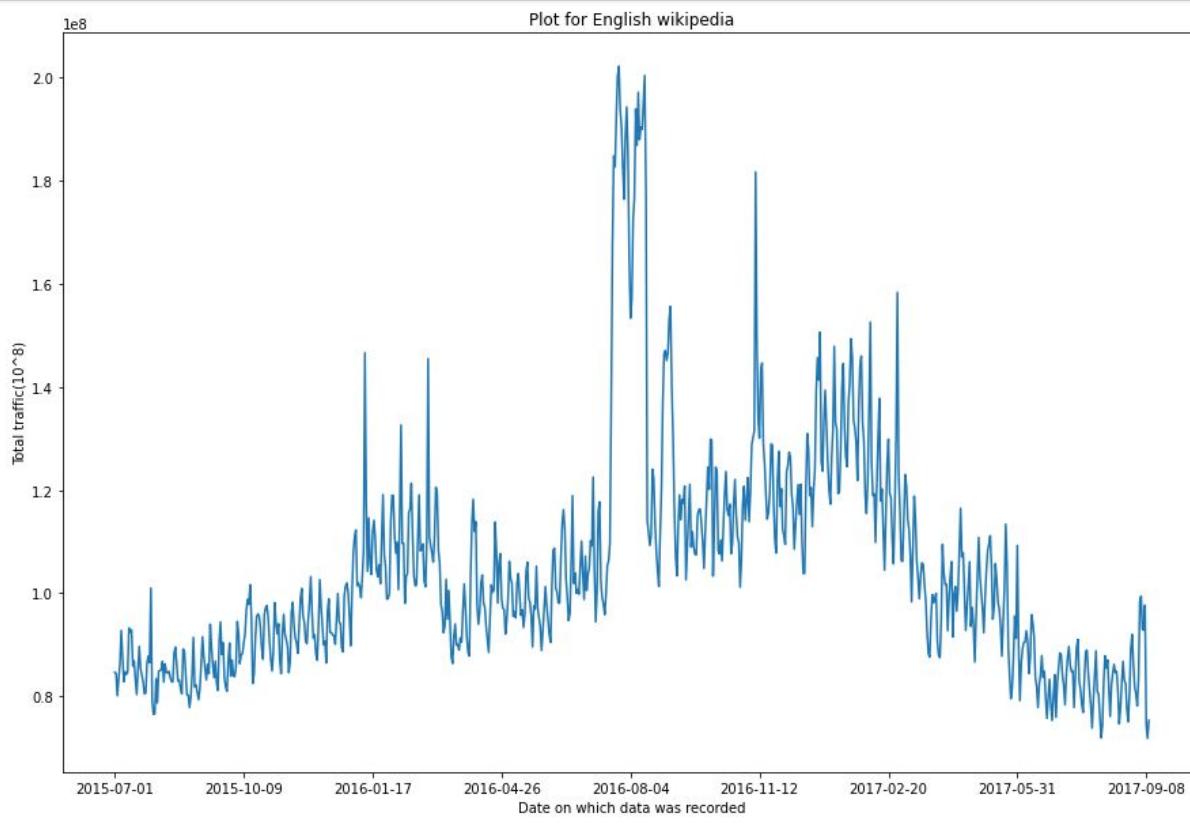
<https://www.kaggle.com/c/titanic>

	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.**

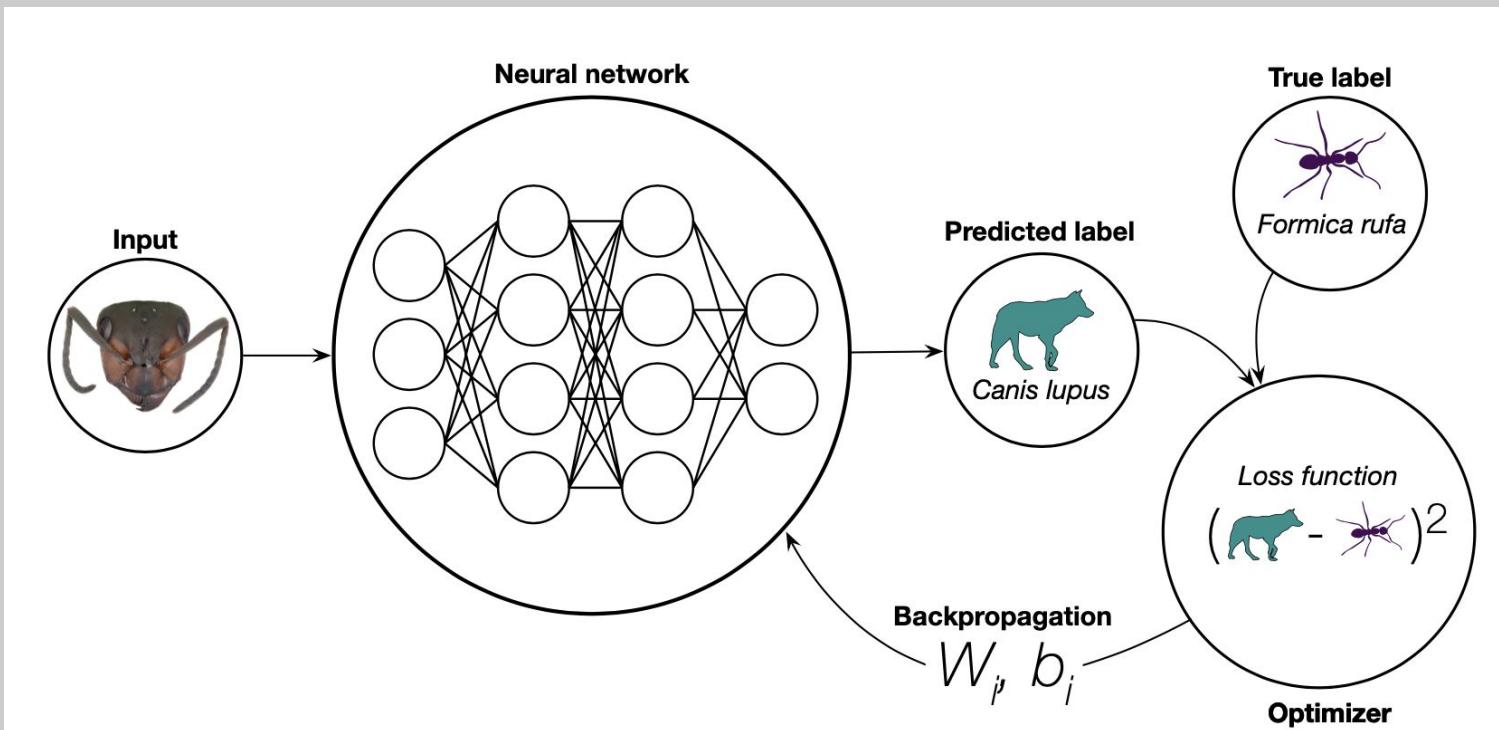
Tabular data

Time series data



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

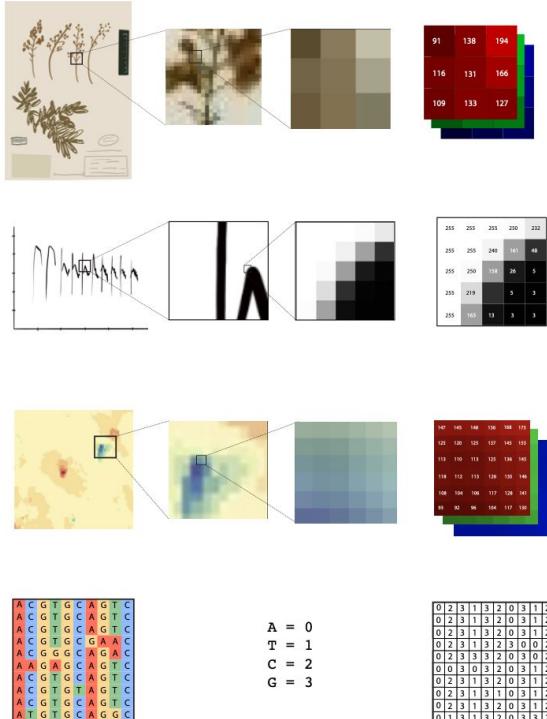
Images (computer vision)



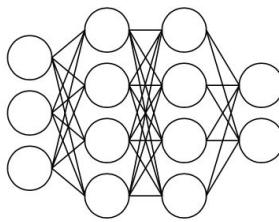
Step 1:
Data Collection



Step 2:
Transform digital data
into input tensor



Step 3:
Neural Net Training/Classification-



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

Images (computer vision)



 Smithsonian Institution
NATIONAL MUSEUM of
NATURAL HISTORY


Smithsonian
DIGITIZATION

Images (computer vision)

Examples of applications of ML in Botany:

- classification

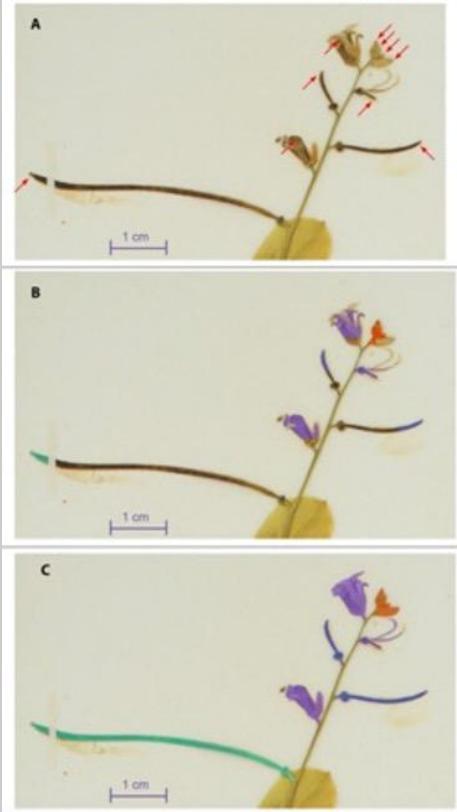
- which species is this?

- object detection

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

- is there evidence of insect damage on this specimen?

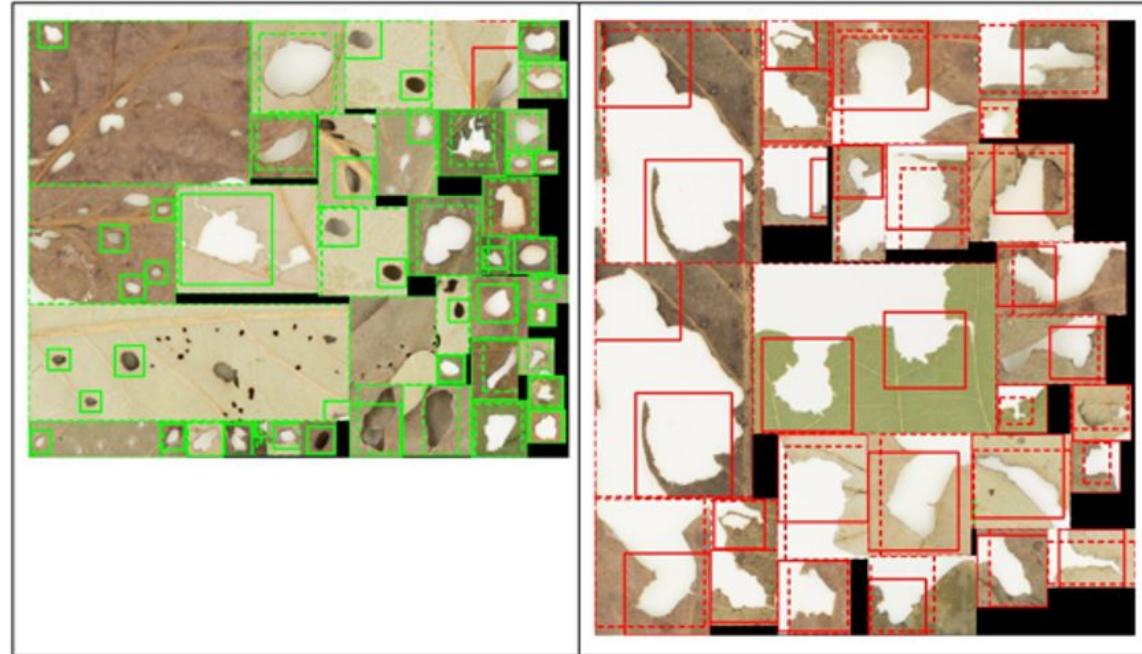
Images (computer vision)



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

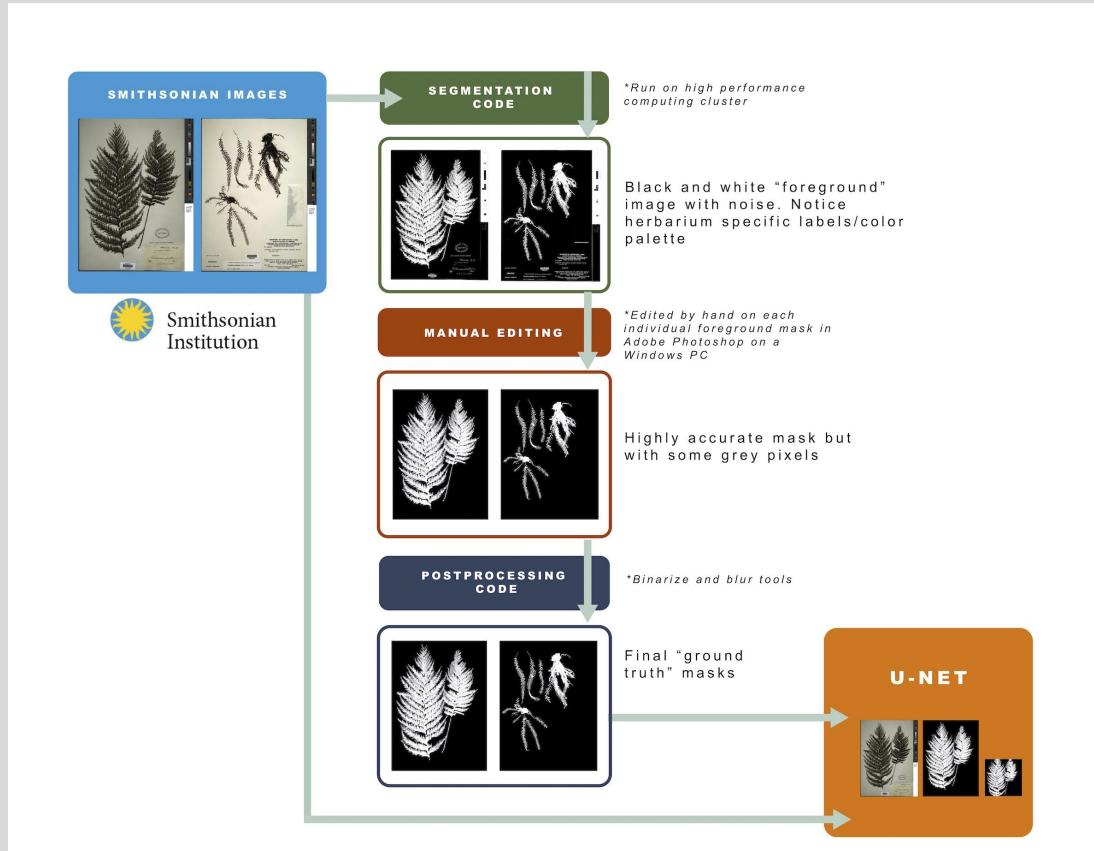


 Smithsonian Institution

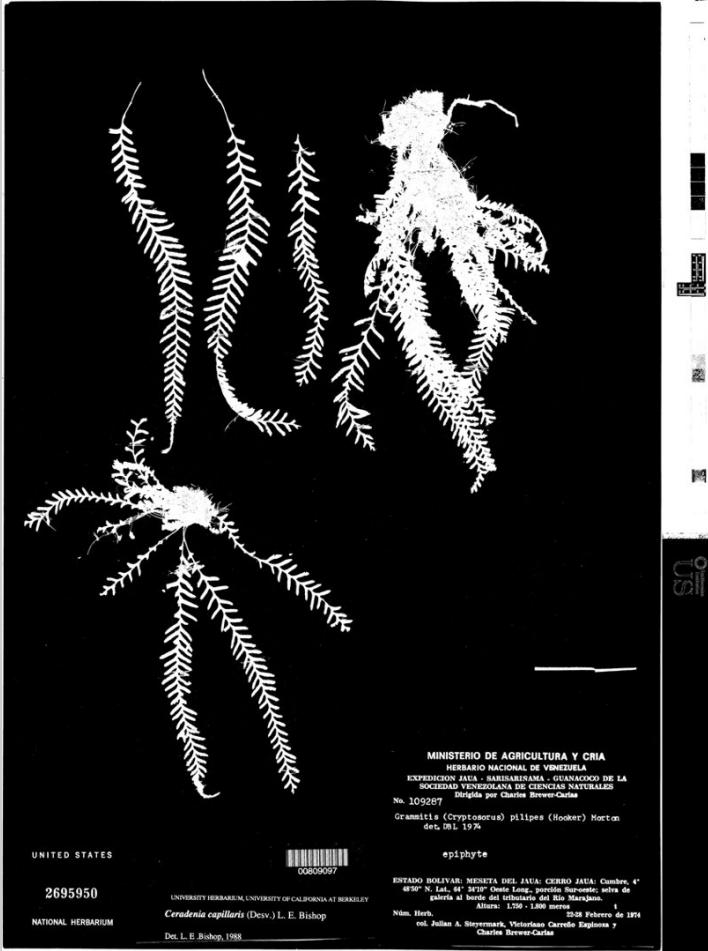
NATIONAL
MUSEUM of
NATURAL
HISTORY



Workflow to produce U-net to mask herbarium images

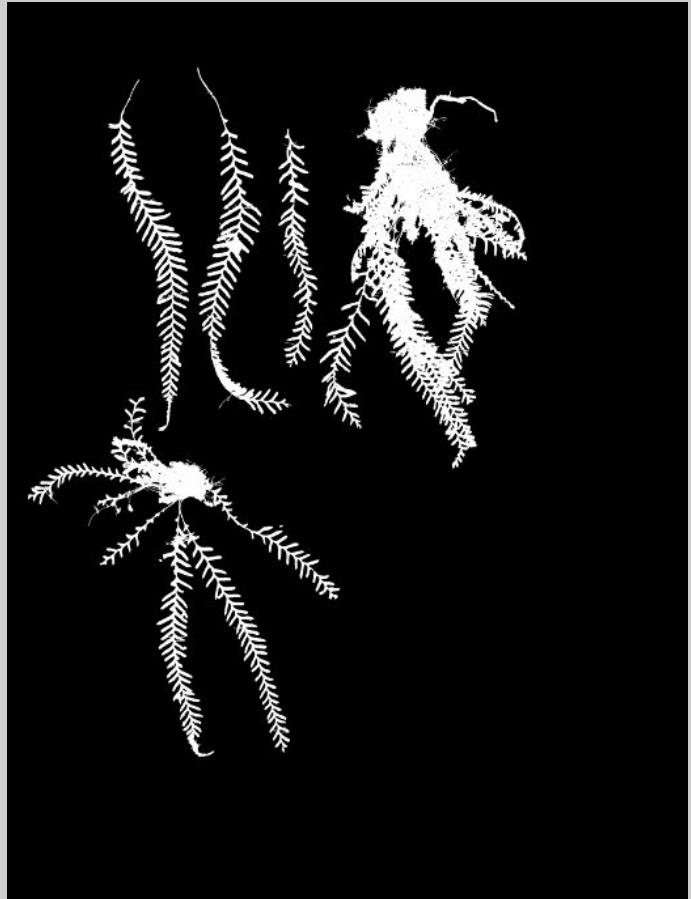


After running segmentation code (built using PlantCV and OpenCV):

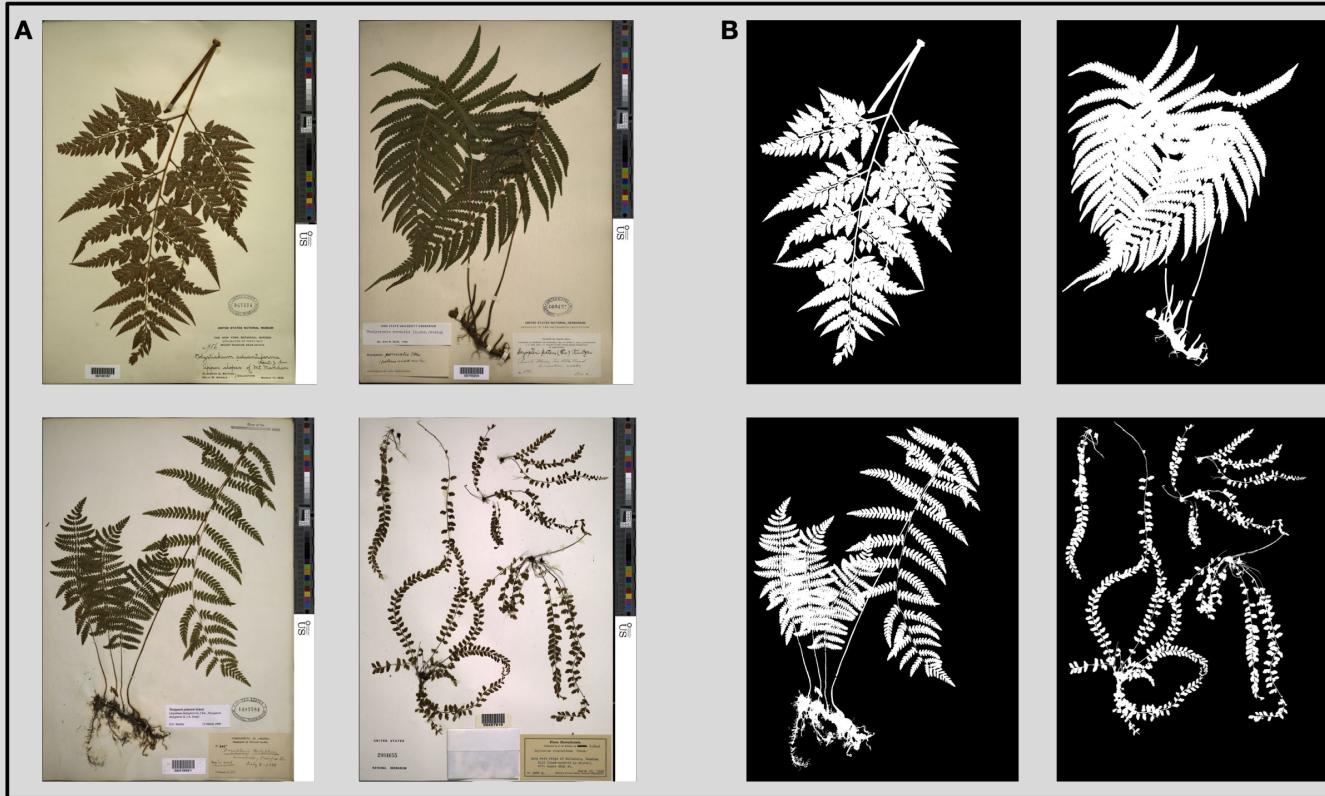


After manual processing to remove any residual non-plant material:

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/>

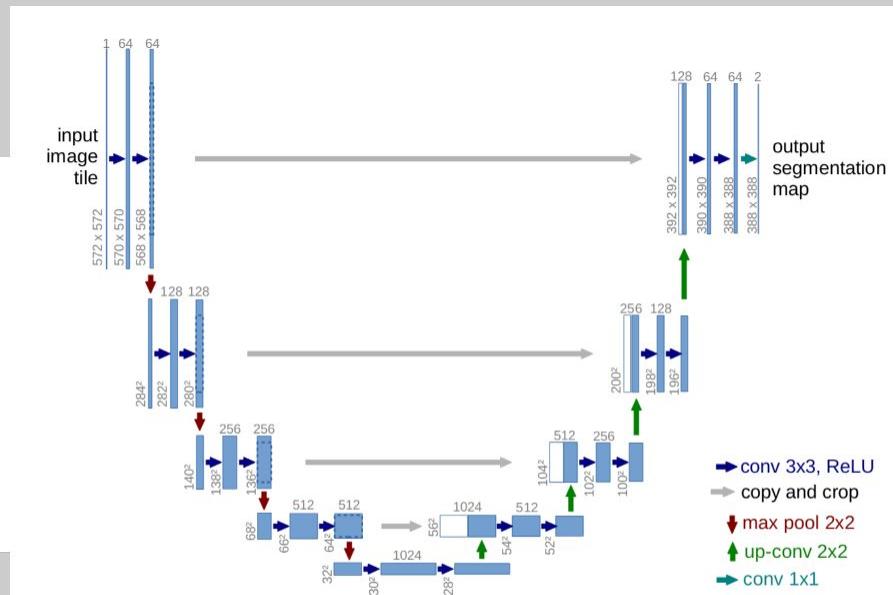
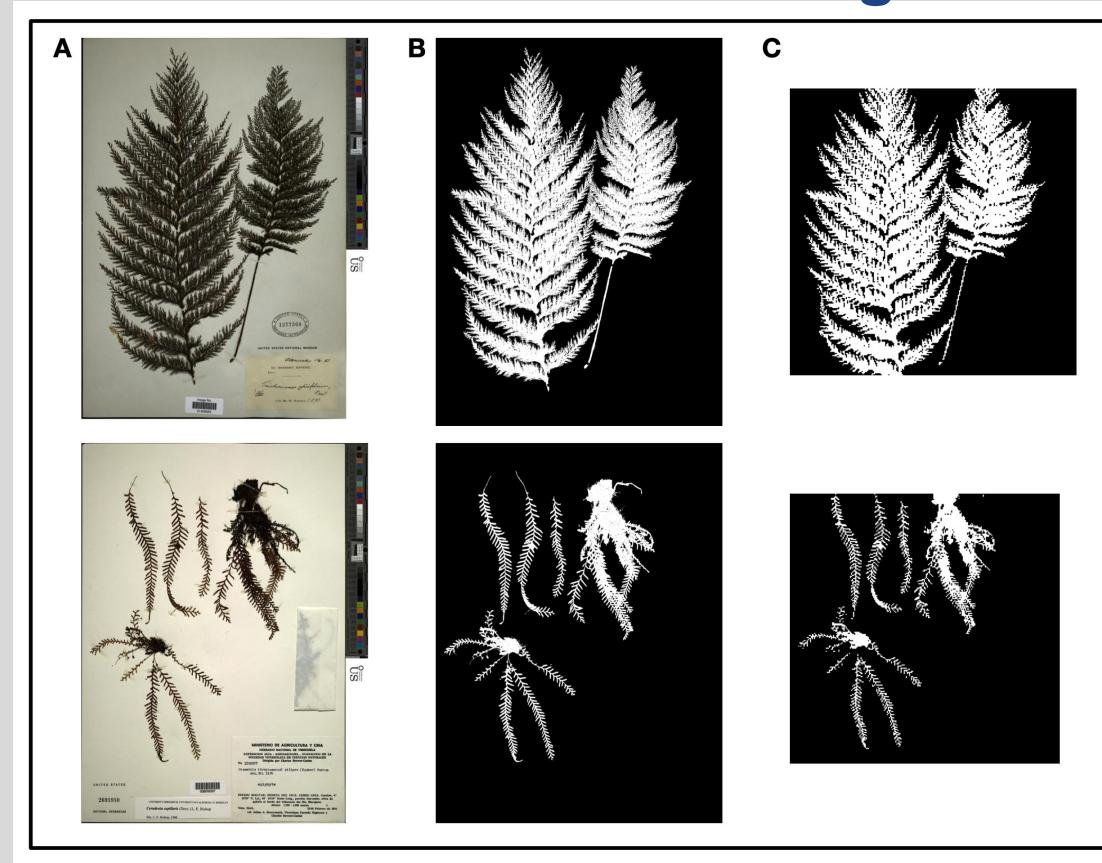


Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows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

Results of U-net training



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

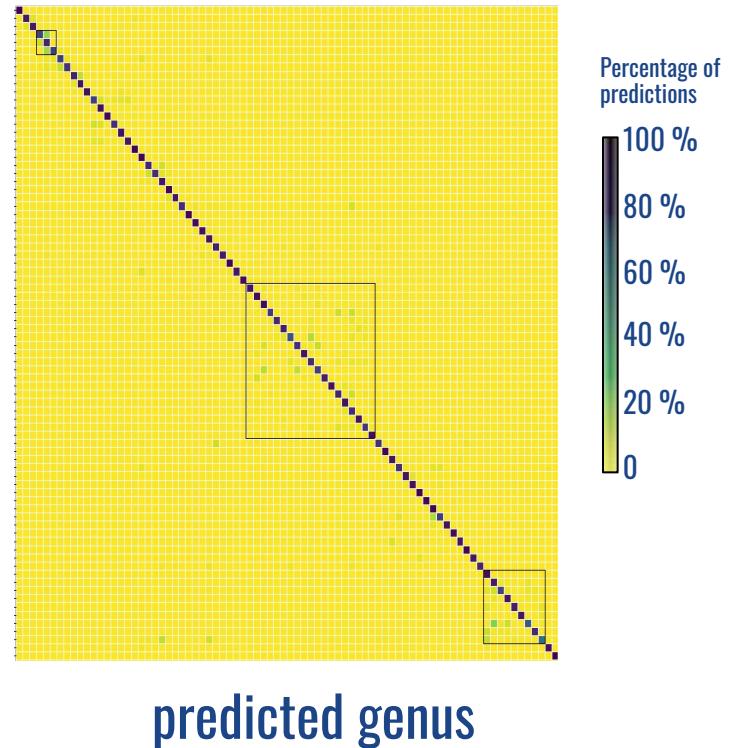
FernNet is 97% accurate at genus ID

3 genera in the tree fern family Cyatheaceae



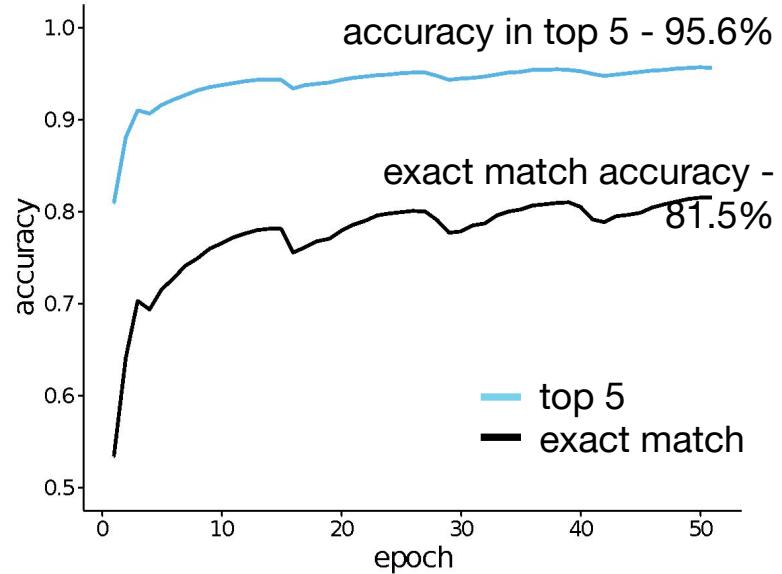
Boxes contain examples of genera within the same family

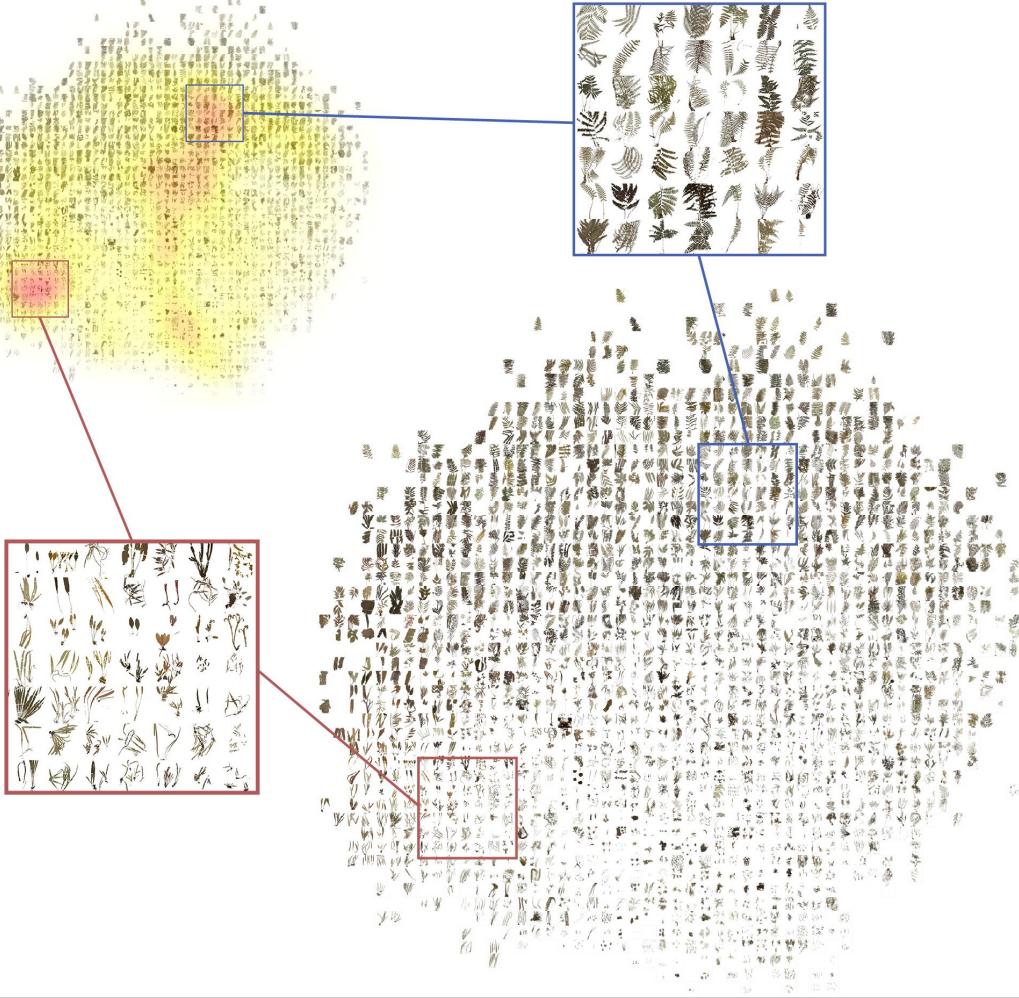
actual genus



Confusion is most often between closely related genera

FernNet is highly accurate for species ID (1425 species)

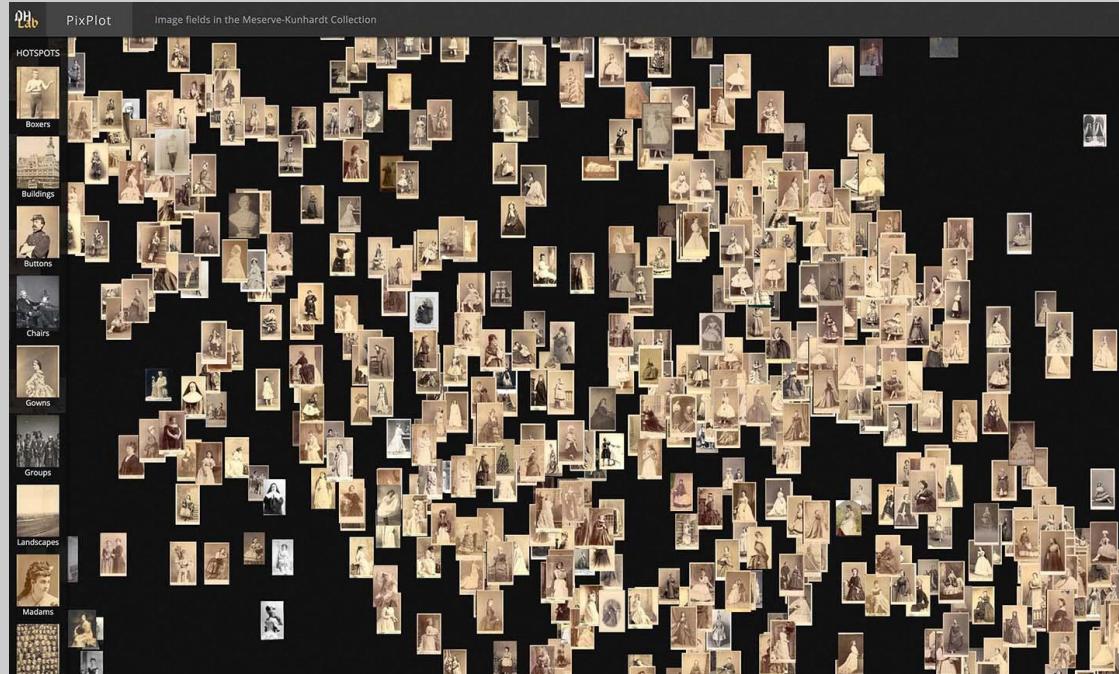




Species classification model can be used to explore shape space occupation.



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”

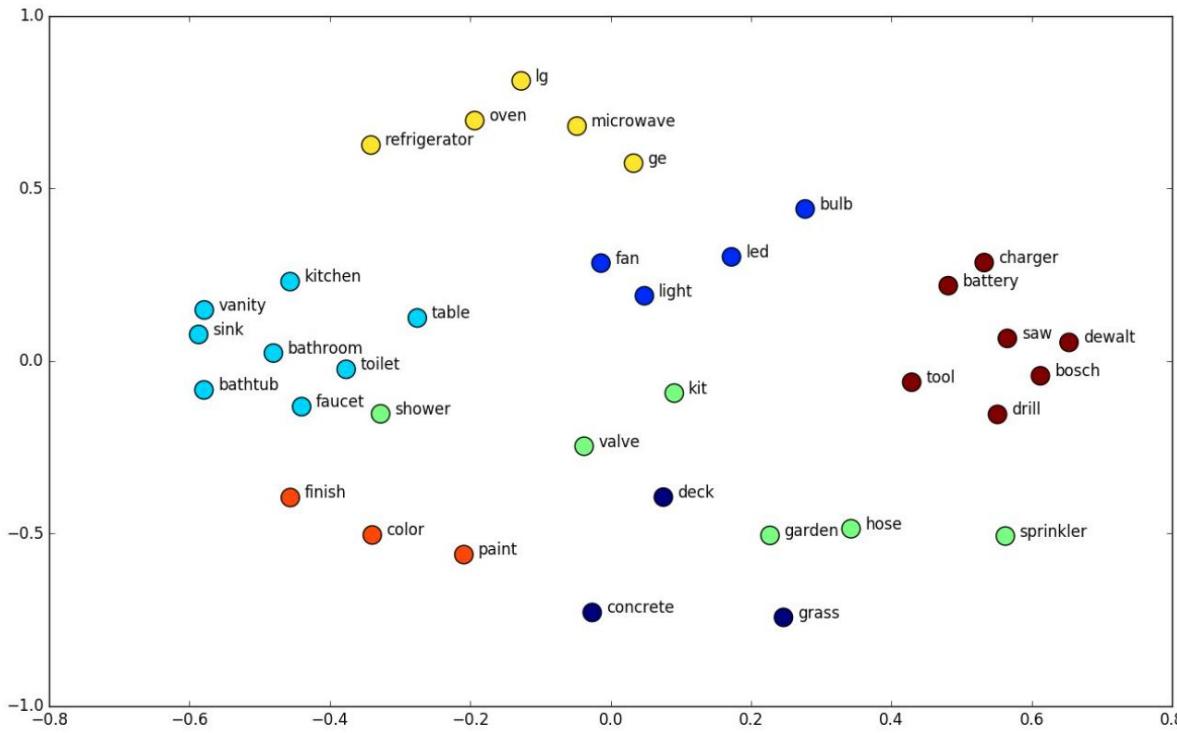
Segment a document into words, count frequency (disregards grammar and word order).

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

William J. Bennett PERSON , Secretary of Education ORG

John S. Herrington PERSON , Secretary of Energy Board of Regents ORG

Warren E. Burger PERSON , Chief Justice of the United States GPE ,
ex officio , Chancellor

George H. W. Bush PERSON , Vice President of the United States GPE , ex
officio

Edwin J. PERSON (Jake) Garn PERSON , Senator from Utah GPE

Barry Goldwater PERSON , Senator from Arizona GPE

James R. Sasser PERSON , Senator from Tennessee GPE

Named Entity Recognition

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

The screenshot shows a web-based interface for the Prodigy NLP platform. The top navigation bar includes links for 'localhost', 'Smithsonian Internship presentation - Google Slides', and '(18) Prodigy'. The main interface has a purple header with tabs: 'MR 1' (selected), 'MRS 2', 'MISS 3', and 'MS 4'. On the left, there's a sidebar titled 'PROJECT INFO' with details: DATASET smithsonian_women_science, LANGUAGE en, RECIPE ner.correct, and VIEW ID ner_manual. Below this is the 'PROGRESS' section showing 'THIS SESSION' (1,698), 'TOTAL' (1,838), and a progress bar. Under 'HISTORY', a list of previous annotations is shown, such as 'SMITHSONIAN ASSOCIATES ... ✓' and 'Mr. Alfred C. Glassell, Jr. The ... ✗'. The main content area displays a list of names and titles from the Smithsonian Internship presentation, each followed by a 'MRS' tag. At the bottom, there are four buttons: a green checkmark, a red X, a grey circle with a slash, and a left arrow.

Smithsonian Internship presentation - Google Slides (18) Prodigy

localhost

Smithsonian Internship presentation - Google Slides (18) Prodigy

prodigy

PROJECT INFO

DATASET smithsonian_women_science

LANGUAGE en

RECIPE ner.correct

VIEW ID ner_manual

PROGRESS

THIS SESSION 1,698

TOTAL 1,838

ACCEPT 51

REJECT 1,647

IGNORE 0

HISTORY

SMITHSONIAN ASSOCIATES ... ✓

Mr. Alfred C. Glassell, Jr. The ... ✗

Appendix 3 SMITHSONIAN A... ✗

Chairman, Department of Geo... ✗

Dr. Rainer Zangerl. ✗

MEMBERS OF THE SMITHSO... ✗

142 APPENDIX 2. ✗

Senior Scientist, Woods Hole ... ✗

Dr. William Von Arx. ✗

Provost, Crown College, Unive... ✗

Symington Foundation, Inc. (Mrs. Martha Frick Symington MRS)

Mr. MR and Mrs. David G. Townsend MRS Mr. MR and Mrs.

Philip M. Tracy MRS Mr. MR and Mrs. J.S. Tressler MRS Mr. MR

and Mrs. A. Buel Trowbridge MRS Mr. MR and Mrs. Julius

Wadsworth MRS The Honorable James E. Webb Mr. MR and Mrs.

William S. Weedon MRS Mrs. Norma Christine Wertz MRS Mr.

George Y. Wheeler III MR Mr. MR and Mrs. Luke W. Wilson MRS

Mrs. Mark Winkler MRS SUPPORTING MEMBERS (\$50 and up) The

Reverend and Mrs. F. Everett Abbott MRS Mr. MR and Mrs.

Stanley N. Allan MRS Mr. MR and Mrs. Walter Beck MRS The

Honorable Frances P. Bolton Mr. MR and Mrs. Philip Bonsai MRS

Mr. MR and Mrs. John F. Boyd MRS Mrs. Eugenie Rowe Bradford

MRS Mr. MR and Mrs. Frederick B. Bryant MRS Mrs. Linda C.

Burgess MRS Mrs. There se Burleson MRS Dr. and Mrs. Charles M.

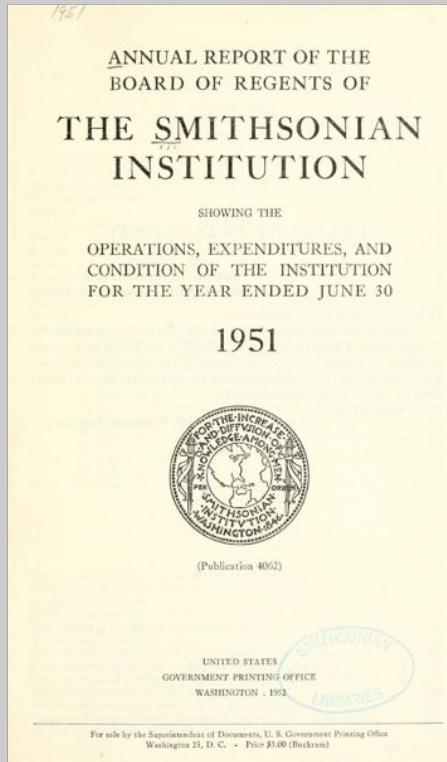
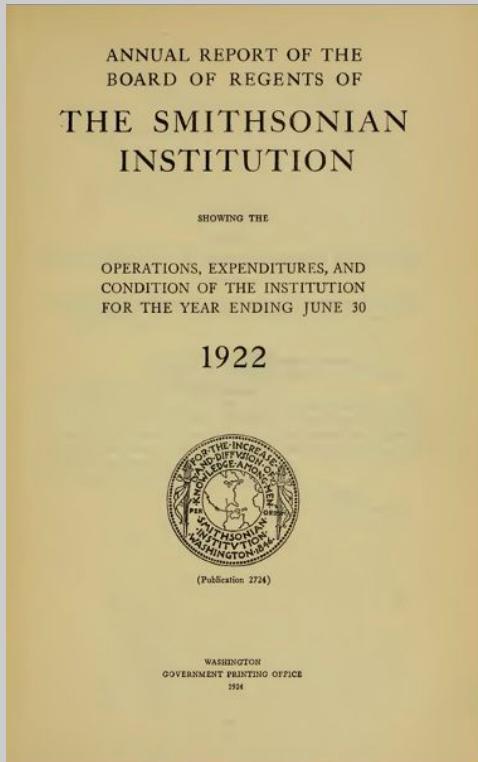
Cabaniss MRS Mr. MR and Mrs. James G. Chandler MRS Mr. MR

and Mrs. Da ILC. Clark MRS Mr. MR and Mrs. David Sanders

Clark MRS ✓ ✗ ✘ ↙ only MR

Mrs. Chester MRS Deevy MRS

© 2017-2021 Explosion (Prodigy v1.10.8)



Breakout activity:

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