Smithsonian Institution

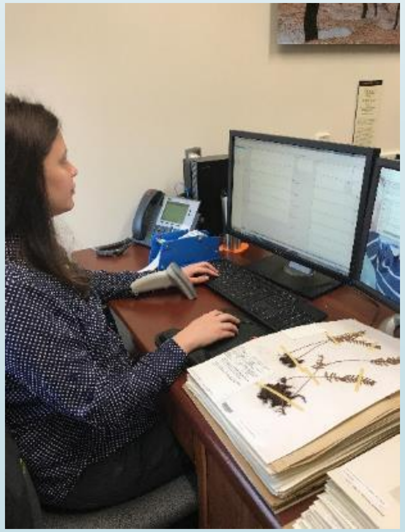
Deep Learning with Botanical Specimen Images

A Voyage into Neural Networks

Sylvia Orli
orlis@si.edu @sylviaorli
Department of Botany
National Museum of Natural History

Digitization at US Herbarium

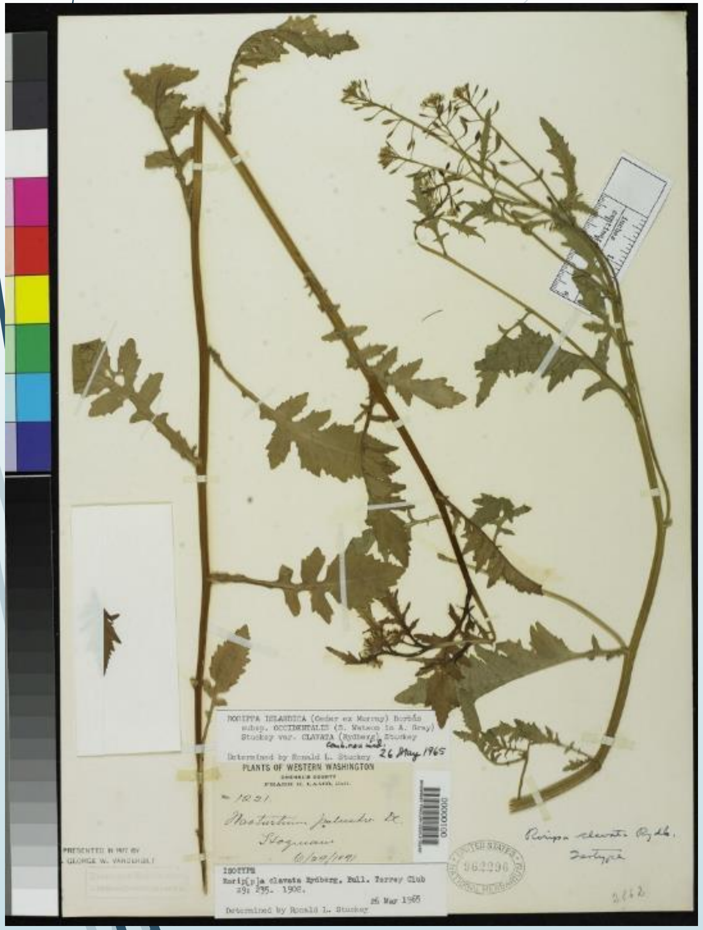
- 1970 – First digitization initiative
- 2001 – Images included in digital record
- 2015 - Digitization through conveyor
- Summer 2017 - 2.2 million inventory records, 1.4 million specimen images total



What information does a specimen image hold?

- Image Metadata
- Collection information
- Taxonomic determinations
- Plant material
- Paper

Partnerships



Smithsonian Institution

NATIONAL MUSEUM *of* NATURAL HISTORY

Smithsonian

OCIO DATA SCIENCE LAB

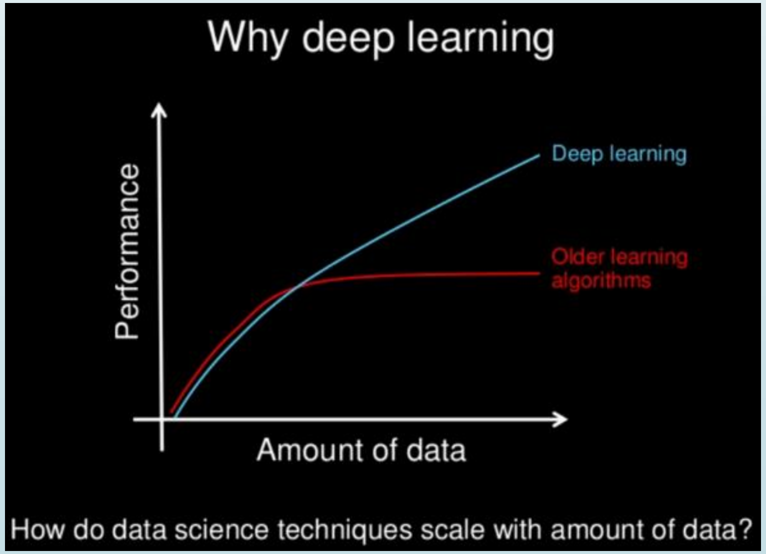
Smithsonian DIGITIZATION

NVIDIA

Artificial Neural Networks

“a computer system modeled on the human brain and nervous system.”

- Computational model used in machine learning
- Used for common every day applications
- GPU and image resource requirements
- A “deep learning” complex model is constructed between raw inputs and the resulting outputs

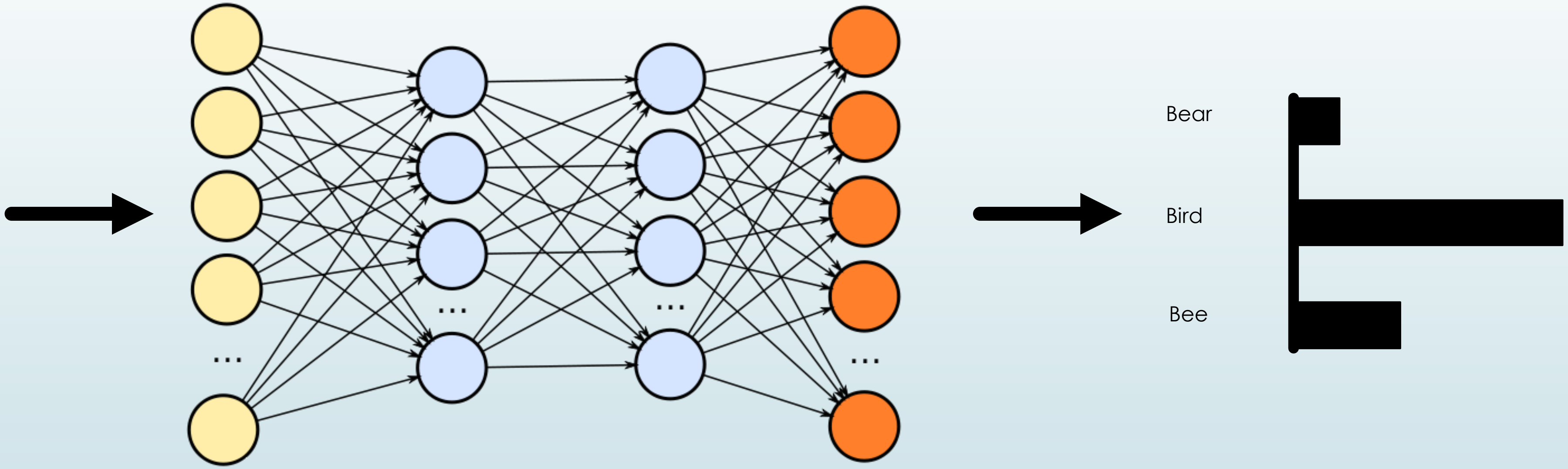


Graphic by Andrew Ng

Deep Learning

Can botanical specimen images be analyzed using this model?



Mercury contaminants

- Mercury added to specimens as insecticide
- Visibility markings show in older specimens
- Estimated 2-5% of specimens affected
- Hot spots in herbarium

Challenges in identifying contaminated samples

- Requires large sample size
- Contaminated vs. non-contaminated samples need to be randomly selected
- Contamination can be unevenly dispersed on specimen
- Contamination is not always contamination
- Contamination not always oxidized and visible



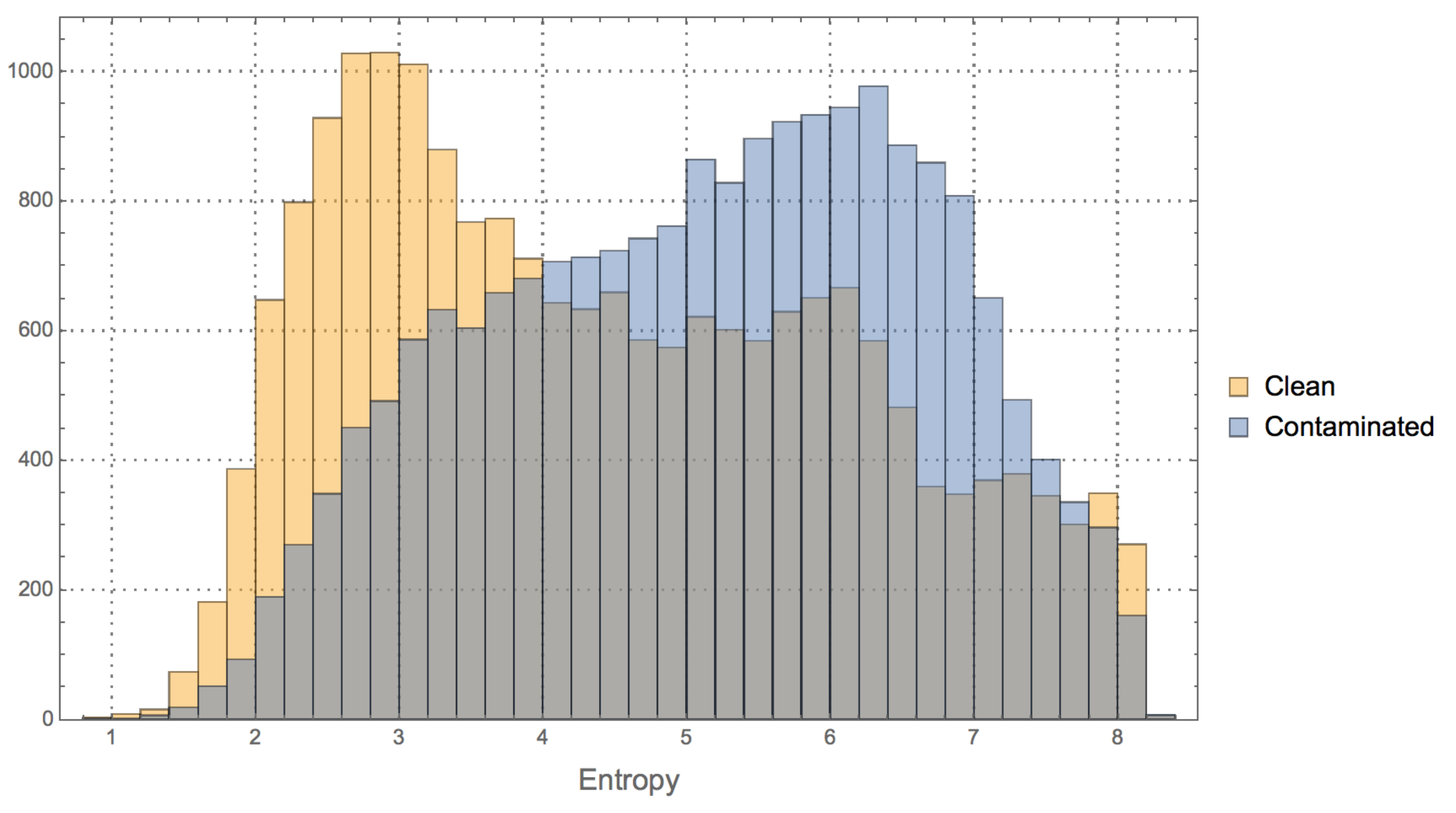
NMNH Deep Learning Approach

- Create neural net model using Mathematica software (Wolfram language)
- Convolutional neural net
- Supervised learning
- Use highly stratified clean / contaminated image set
- Image set: 70% training, 20% validation and 10% test

Initial Approach – 2000 images

- Partition the high-res images into 128x128 px tiles to **inflate training dataset**

Entropy Distribution of Labelled Tiles



Classifying Tile Samples

net[, "Probabilities"]

<|clean → 0.849673, contaminated → 0.150327|>

net[, "Probabilities"]

<|clean → 0.777121, contaminated → 0.222879|>

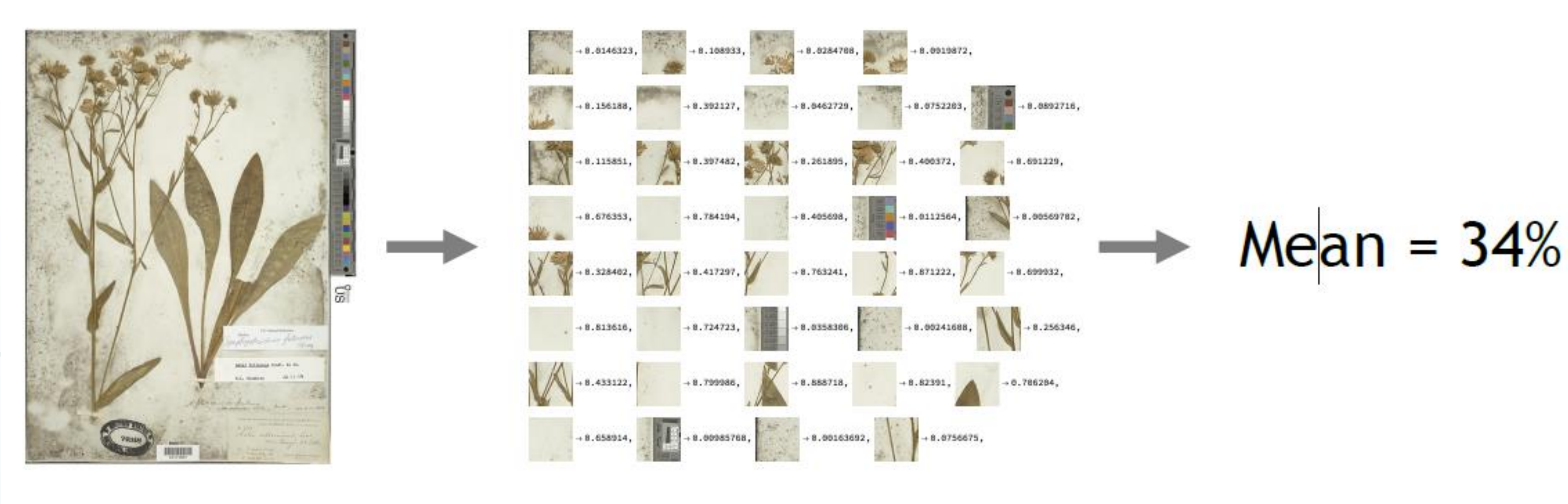
net[, "Probabilities"]

<|clean → 0.71387, contaminated → 0.28613|>

net[, "Probabilities"]

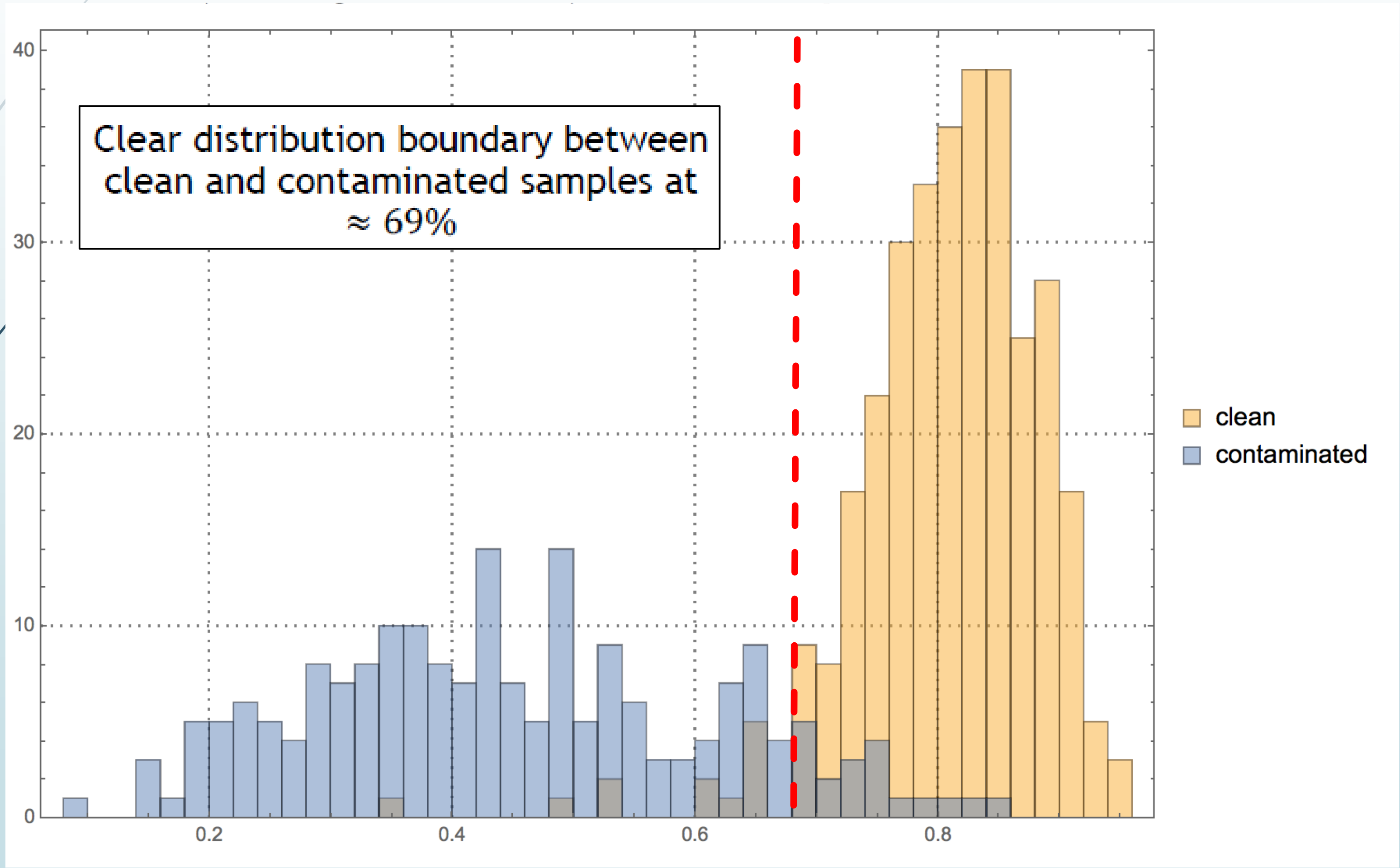
<|clean → 0.251451, contaminated → 0.748549|>

Probability of "Clean"


Mean = 34%


Mean = 85%

Distribution of mean probability (clean) over all test images



Clean Accuracy: 95%

Contaminated Accuracy: 92%

Gathering more images

- 9380 images of contaminated sheets
- 9383 images of clean sheets
- Reduce Full Image dimensions to 256 x 256
- Use full image instead of tile probability

Entropy Distribution of Full Specimen Images

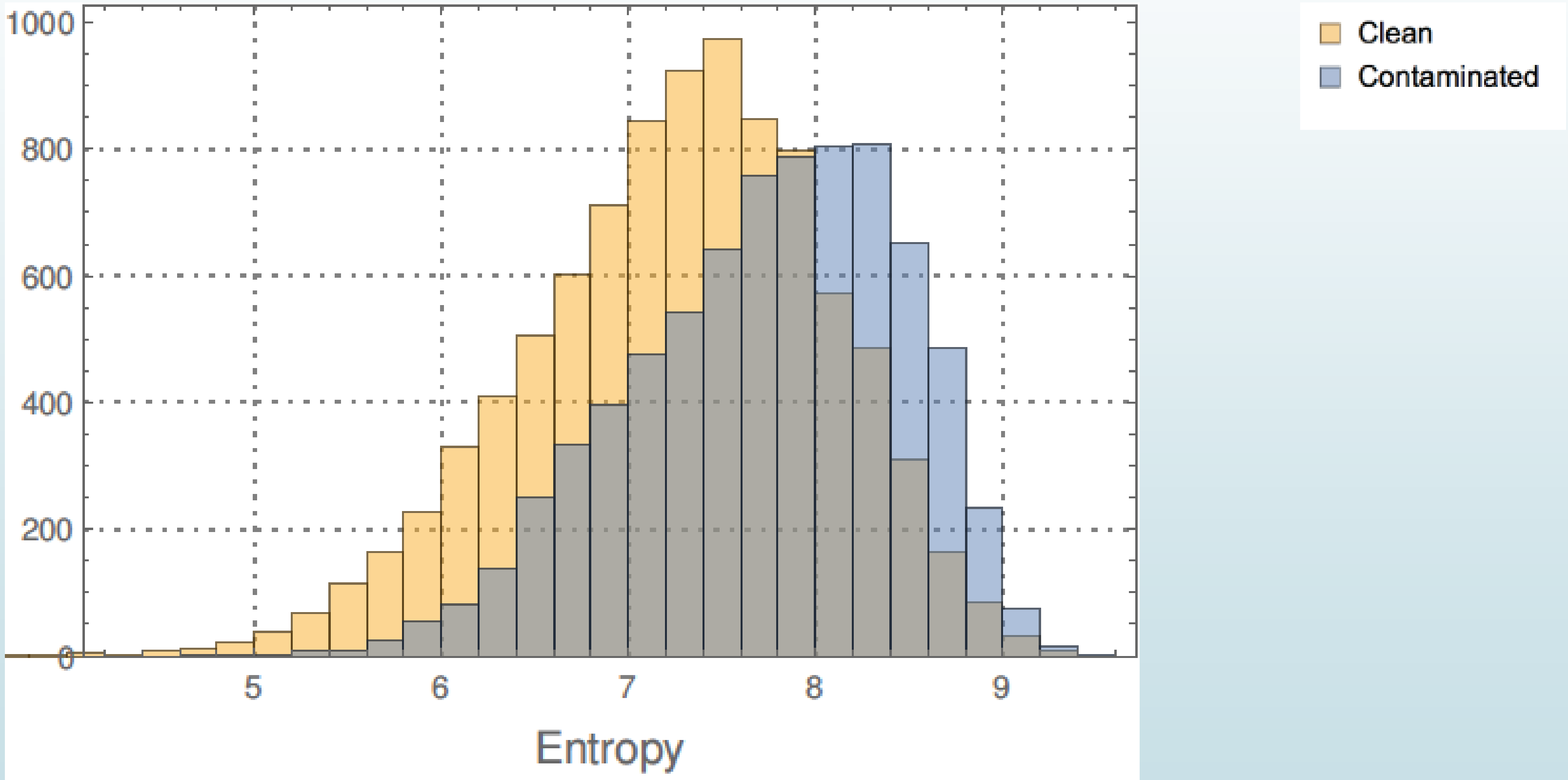
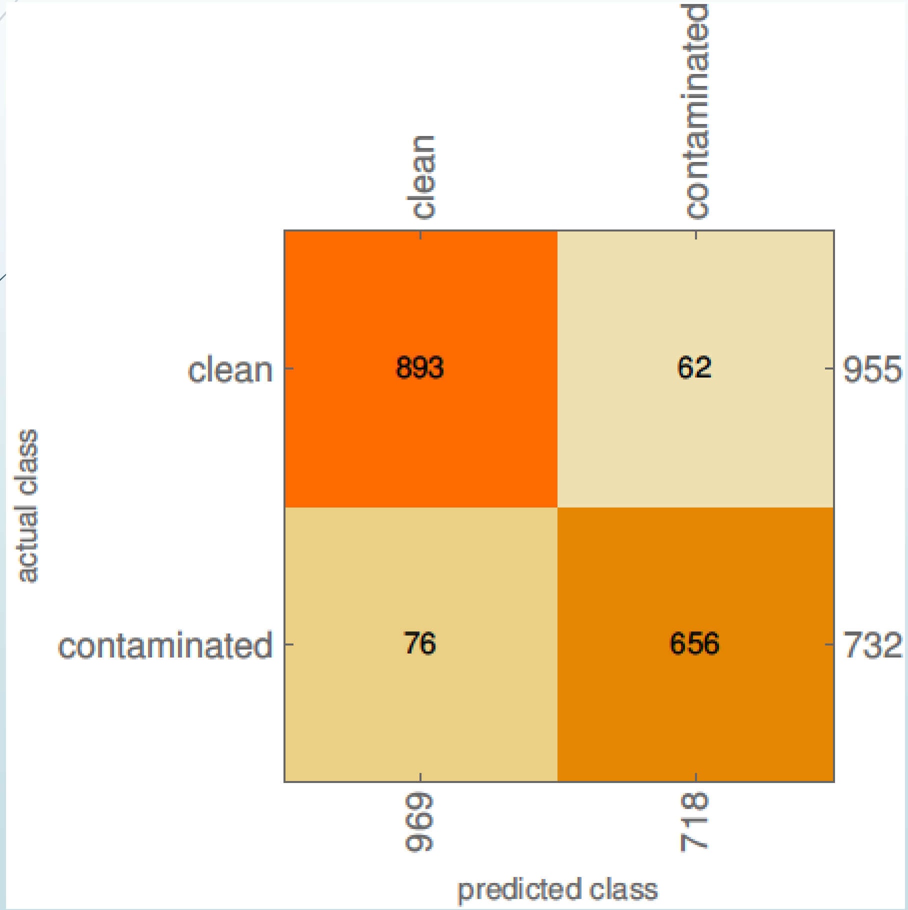
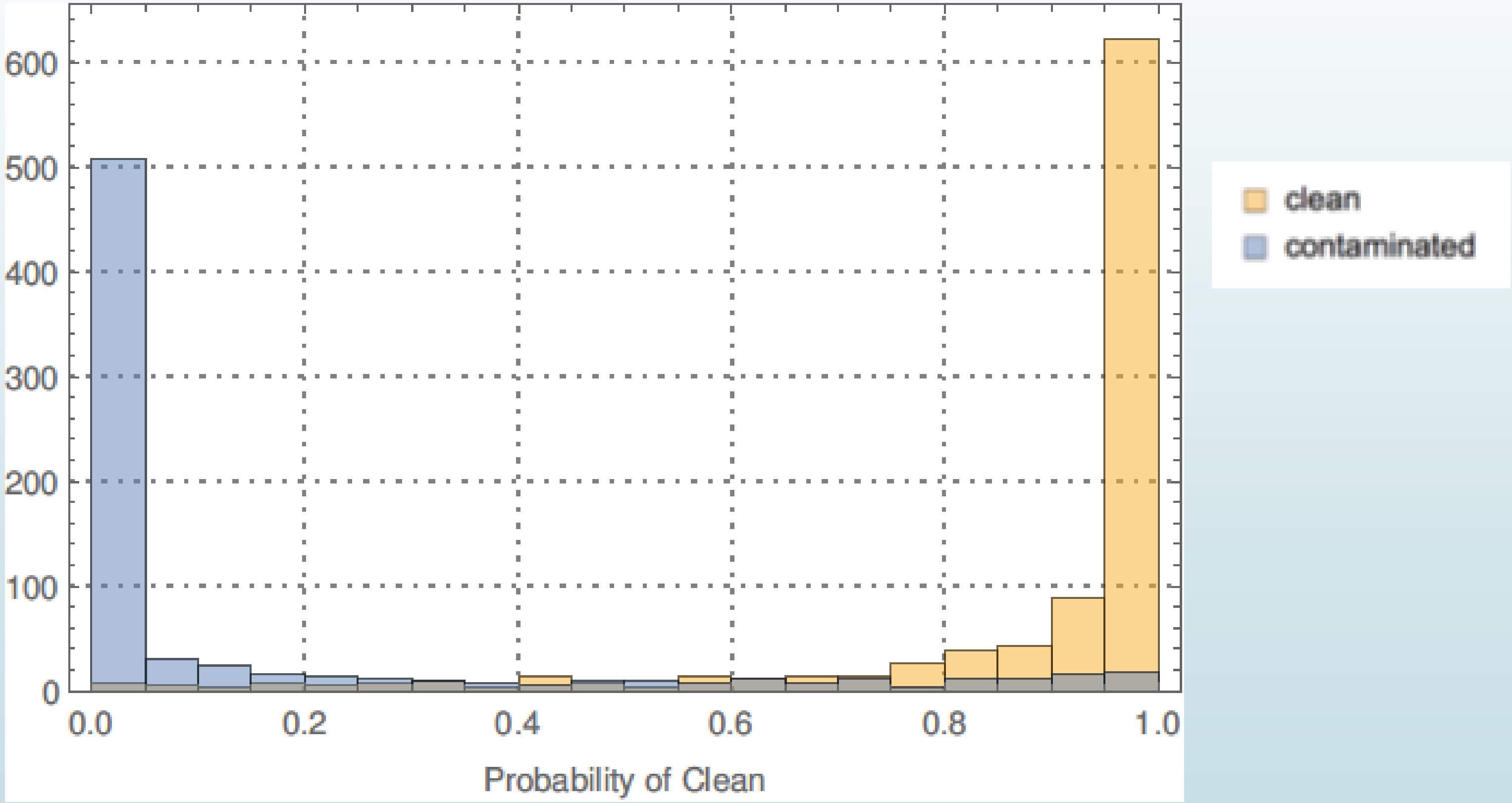


Image classifier (Confusion Matrix)



- 92% accuracy in detecting clean vs. contaminated specimens
- Higher accuracy with further tweaking of neural nets

Distribution of probability (clean) over all test images



Misclassified specimens

Further Deep Learning Uses

- Plant family differences
- Species Identification
- Transcription of specimen labels
- Collaborations?