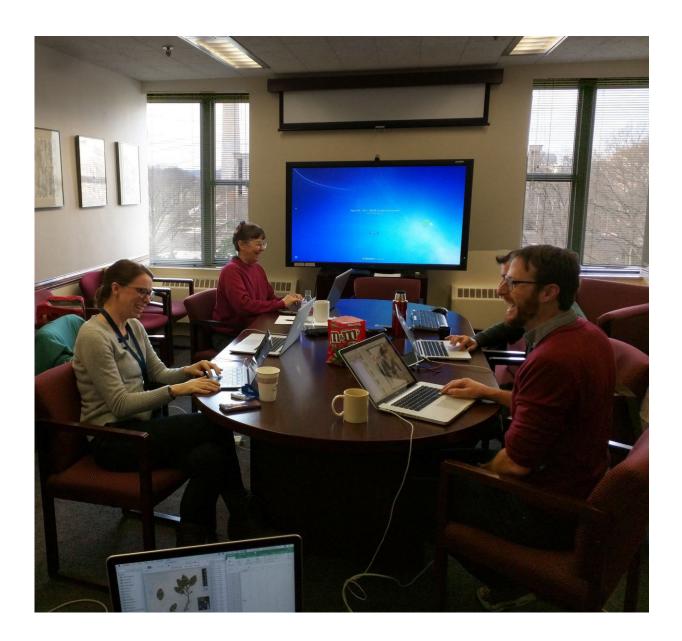# Artificial neural networks could power up curation of natural history collections

November 2 2017

Smithsonian researchers classifying digitized herbarium sheets that have been

stained with mercury to build a training dataset. Credit: Paul B. Frandsen, Smithsonian

Millions, if not billions, of specimens reside in the world's natural history collections, but most of these have not been carefully studied, or even looked at, in decades. While containing critical data for many scientific endeavors, most objects are quietly sitting in their own little cabinets of curiosity.

Thus, mass digitization of natural history collections has become a major goal at museums around the world. Having brought together numerous biologists, curators, volunteers and citizens scientists, such initiatives have already generated large datasets from these collections and provided unprecedented insight.

Now, a study, recently published in the open access *Biodiversity Data Journal*, suggests that the latest advances in both digitization and machine learning might together be able to assist museum curators in their efforts to care for and learn from this incredible global resource.

A team of researchers from the Smithsonian Department of Botany, Data Science Lab, and Digitization Program Office recently collaborated with NVIDIA to carry out a pilot project using deep learning approaches to dig into digitized herbarium specimens.

Their study is among the first to describe the use of deep learning methods to enhance our understanding of digitized collection samples. It is also the first to demonstrate that a deep convolutional neural network—a computing system modelled after the neuron activity in animal brains that can basically learn on its own—can effectively differentiate between similar plants with an amazing accuracy of nearly

100%.

The US National Herbarium at the Smithsonian's National Museum of Natural History in Washington, D.C. Credit: Chip Clark, Smithsonian

In the paper, the scientists describe two different neural networks that they trained to perform tasks on the digitized portion (currently 1.2 million specimens) of the United States National Herbarium.

The team first trained a net to automatically recognize herbarium sheets that had been stained with mercury crystals, since mercury was commonly used by some early collectors to protect the plant collections from insect damage. The second net was trained to discriminate between

two families of plants that share a strikingly similar superficial appearance.

The trained neural nets performed with 90% and 96% accuracy respectively (or 94% and 99% if the most challenging specimens were discarded), confirming that deep learning is a useful and important technology for the future analysis of digitized museum collections.

Digitized collections combined with deep learning will help us to automate an

otherwise human task of identifying an unknown number of stained specimen sheets across a collection of over 5 million. See http://collections.si.edu. Credit: Smithsonian Institution

"The results can be leveraged both to improve curation and unlock new avenues of research," conclude the scientists.

"This research paper is a wonderful proof of concept. We now know that we can apply machine learning to digitized natural history specimens to solve curatorial and identification problems. The future will be using these tools combined with large shared data sets to test fundamental hypotheses about the evolution and distribution of plants and animals," says Dr. Laurence J. Dorr, Chair of the Smithsonian Department of Botany and co-author of the study.

More information: Eric Schuettpelz et al, Applications of deep convolutional neural networks to digitized natural history collections, *Biodiversity Data Journal* (2017). DOI: 10.3897/BDJ.5.e21139

Provided by Pensoft Publishers