Applications in Plant Sciences

Botanical Society of America
leading scientists and educators since 1893

APPLICATION ARTICLE

INVITED SPECIAL ARTICLE

For the Special Issue: Machine Learning in Plant Biology: Advances Using Herbarium Specimen Images

Applying machine learning to investigate long-term insect–plant interactions preserved on digitized herbarium specimens

Emily K. Meineke[1,5], Carlo Tomasi[2], Song Yuan[3], and Kathleen M. Pryer[4]

Manuscript received 1 October 2019; revision accepted 4 March 2020.

[1] Department of Entomology and Nematology, University of California, Davis, California 95616, USA

[2] Department of Computer Science, Duke University, Durham, North Carolina 27708, USA

[3] Department of Mechanical Engineering and Materials Science, Duke University, Durham, North Carolina 27708, USA

[4] Department of Biology, Duke University, Durham, North Carolina 27708, USA

[5] Author for correspondence: emily.meineke@gmail.com

Citation: Meineke, E. K., C. Tomasi, S. Yuan, and K. M. Pryer. 2020. Applying machine learning to investigate long-term insect–plant interactions preserved on digitized herbarium specimens. *Applications in Plant Sciences* 8(6): e11369.

doi:10.1002/aps3.11369

PREMISE: Despite the economic significance of insect damage to plants (i.e., herbivory), long-term data documenting changes in herbivory are limited. Millions of pressed plant specimens are now available online and can be used to collect big data on plant–insect interactions during the Anthropocene.

METHODS: We initiated development of machine learning methods to automate extraction of herbivory data from herbarium specimens by training an insect damage detector and a damage type classifier on two distantly related plant species (*Quercus bicolor* and *Onoclea sensibilis*). We experimented with (1) classifying six types of herbivory and two control categories of undamaged leaf, and (2) detecting two of the damage categories for which several hundred annotations were available.

RESULTS: Damage detection results were mixed, with a mean average precision of 45% in the simultaneous detection and classification of two types of damage. However, damage classification on hand-drawn boxes identified the correct type of herbivory 81.5% of the time in eight categories. The damage classifier was accurate for categories with 100 or more test samples.

DISCUSSION: These tools are a promising first step for the automation of herbivory data collection. We describe ongoing efforts to increase the accuracy of these models, allowing researchers to extract similar data and apply them to biological hypotheses.

KEY WORDS Anthropocene; climate change; herbarium; insects; machine learning; species interactions.

More than 390,000,000 pressed plant specimens are stored in the world's 3100 herbaria (Thiers, 2020). Collected over the past four centuries, herbarium specimens provide the most comprehensive view of Earth's vegetation and how it has changed over time. In most cases, herbarium specimens are used to document plant diversity (Heberling et al., 2019); however, they are now being applied to scientific efforts well beyond taxonomy and systematics, and in particular to global change biology. Millions of herbarium specimens were collected prior to the intensification of human influence on the planet, including the acceleration of climate change. The unique, long-term data preserved within herbarium collections can now help us understand the past and predict the future of global change (Heberling and Isaac, 2017; Meineke et al., 2018a, b, c; Lang et al., 2019).

Until recently, the world's herbarium specimens were under lock and key, accessible only to relatively few scientific specialists. Today, the digitization of herbaria is a global enterprise, and millions of high-resolution specimen images and their associated metadata are newly available in online public databases (Page et al., 2015; Soltis and Soltis, 2016). In their most widespread application outside of taxonomy and systematics, herbarium specimens have accelerated the study of plant phenological change. Reproductive structures such as flowers, buds, and fruits, along with the collection dates and locations associated with the specimens, can provide information on how phenological timing has shifted with climate through time (Willis et al., 2017). Notably, scientists have established that warmer temperatures are associated with widespread, earlier flowering times in temperate North America (Primack et al., 2004; Panchen et al., 2012). Although phenology has been a key focus of global change research using herbarium collections, it encompasses only a small amount of the long-term data that could be mined from digitized specimens.

In particular, herbaria are unmatched repositories for data documenting interactions between plants and associated species. Species interactions, such as those between plants and insects, are notoriously difficult to monitor over time, and data on species interactions spanning the Anthropocene are severely limited (Meineke and Davies, 2018). Plants have been engaged in a reciprocal war with the rasping, sucking, and chewing insects that feed on them for more than 400 million years. This long-term coevolutionary arms race between insects and plants is the basis for Ehrlich and Raven's (1964) hypothesis that plant adaptations and defenses to insect attack have stimulated the diversification of insects. In turn, through their fitness effects on plants, herbivores have elevated plant speciation rates (Futuyma and Agrawal, 2009). As such, relationships between plants and insect herbivores are central to ecology and evolutionary biology.

Herbarium specimens preserve signatures of insect damage (i.e., herbivory) through time on their leaves (Beaulieu et al., 2018; Meineke and Davies, 2018; Meineke et al., 2018a, b). Herbivory encompasses diverse types of damage triggered by a wide range of insect taxa (for examples, see Fig. 1). Interestingly, similar signatures of insect damage have been documented on fossilized leaves. The paleobotanical community has already initiated critical studies that use and interpret these data preserved on fossilized leaves to understand species interactions over deep time between plants and insects (Wilf and Labandeira, 1999; Wilf et al., 2001). This series of studies has revealed that warming over epochs increased the diversity and abundance of insect damage recorded in fossilized leaves (e.g., Currano et al., 2010). Herbarium specimens offer an opportunity to conduct analogous studies that analyze how plant–insect associations may have shifted during the Anthropocene in response to warming, and also to other key drivers of biodiversity change: pollution, harvesting by humans, habitat loss, and invasive species (Meineke et al., 2018a).

To date, the only study to use herbarium specimens to quantify how herbivory has shifted over the past ≥100 years was published by Meineke et al. (2018b). In this study, over a period of two years, a single researcher painstakingly overlaid a physical grid of 5 × 5-cm cells on almost 600 herbarium specimens, developing novel methods for manually identifying and quantifying herbivory. They demonstrated that insect damage to four distantly related, woody angiosperm species in New England increased over the past 112 years by 23%, a pattern attributed to increased winter warming that promoted overwintering survival and/or range expansion of herbivorous insects. Meineke et al. (2018b) is one of the first studies to use data captured from herbarium specimens to investigate hypotheses about the ecological and evolutionary mechanisms driving herbivory. However, the amount of herbivory data that can become available for future study is constrained by how much data individual researchers can manually collect from physical herbarium specimens.

Here, we move beyond the manual procedures advanced by Meineke et al. (2018b) by developing novel automated methods to replace them, with the goal of allowing future studies on plant–insect species interactions to harness data from millions of herbarium specimen images available online. Should these large-scale data become available for research, they would allow ecologists to tackle long-standing hypotheses that have not been adequately addressed because data collection has proved prohibitive (Table 1). Similar to other studies that focus on plant reproductive parts and phenology (Lorieul et al., 2019), we develop machine learning algorithms that

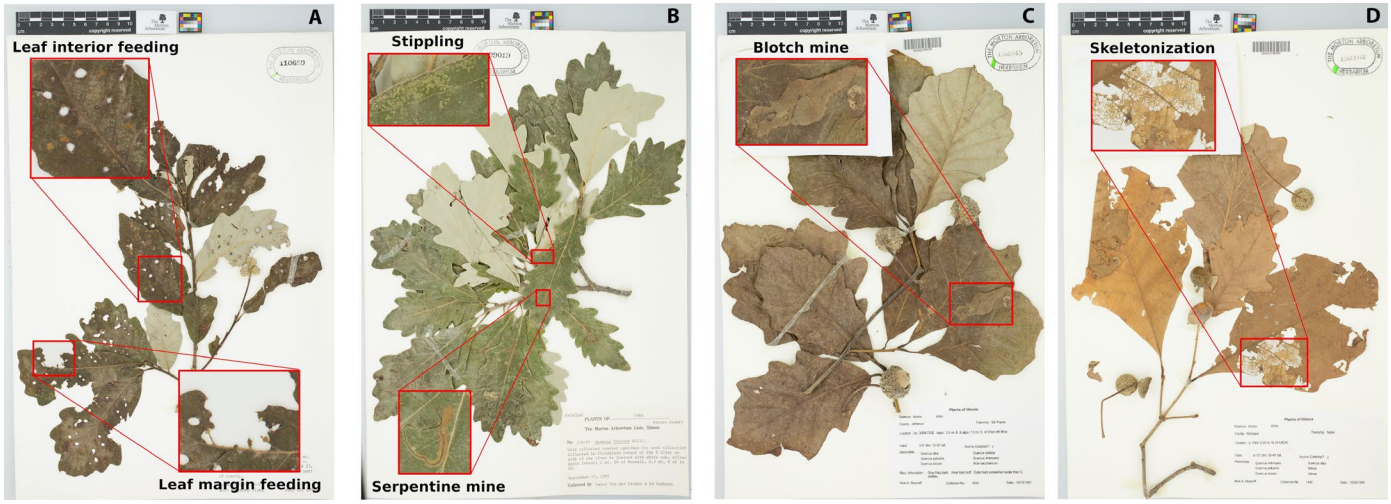

FIGURE 1. Herbarium specimens exhibiting a range of herbivory types made by different insect taxa for which recognition was automated in this study, including examples of leaf interior feeding and leaf margin feeding (A), stippling and serpentine mines (B), blotch mines (C), and skeletonization (D).

TABLE 1. Sampling of a priori hypotheses that are of broad interest in ecology. From left to right, we list predictions made from these hypotheses using limited available data, the data gap that could be filled by large-scale herbivory data sets derived using machine learning algorithms applied to herbarium specimens, and relevant publications pointing to the need for big data to more fully assess predictions.

Hypothesis	Prediction(s)	Data gap to be filled	Relevant publication(s)
Herbivory rates depend on latitude	Herbivory is elevated at lower latitudes.	Limited herbivory data across latitudes	Moles et al., 2011
Herbivory results in a major transfer of energy and nutrients from primary producers to consumers	Herbivores consume about 5% of all leaf tissue, representing a small transfer of energy and nutrients. vs. Herbivores consume 10–20% of all leaf tissue, representing a large transfer of energy and nutrients.	Limited herbivory data worldwide and across the plant phylogeny	Turcotte et al., 2014b vs. Coley et al., 1985 Cyr and Face., 1993 Cebrian and Lartigue, 2004
Herbivory rates vary among plant lineages	Ferns incur less herbivory than angiosperms.	Limited herbivory data across the plant phylogeny	Cooper-Driver, 1978
Herbivory rates depend on plant growth form and size	Large plants are more "apparent" to herbivores and are thus are eaten at higher rates than smaller plants.	Limited standardized herbivory data across plant growth forms	Feeny, 1976
Herbivory intensity has changed due to climate change	Herbivory has increased where winters are warming. Herbivory has decreased where temperatures newly exceed insect thermal maxima.	Spotty monitoring of insects/herbivory before and after the acceleration of climate change in the 1970s	Meineke et al., 2018b
Herbivory intensity has changed due to urbanization	Effects of urban warming on insect herbivory/diversity depend on latitude. In general, urbanization reduces damage by chewing herbivores.	Poor long-term monitoring of how building cities affects insects/herbivory	Diamond et al., 2015 Kozlov et al., 2017 Meineke and Davies, 2018
Invasive plants experience "natural enemy release"	Invasive species escape herbivory in introduced habitats but accumulate herbivores in novel habitats over time.	Limited data on how much introduced species are damaged by herbivores throughout the invasion process	Zangerl and Berenbaum, 2005

can recognize and quantify insect damage on digitized herbarium specimens. We then discuss ongoing improvements on these methods to facilitate the broadscale extraction of species interactions data.

METHODS

Focal species

Meineke et al. (2018b) manually collected herbivory data from ca. 600 herbarium specimens of four woody angiosperm species distributed in eastern North America: *Quercus bicolor* Willd. (Fagaceae), *Vaccinium angustifolium* Aiton (Ericaceae), *Carya ovata* (Mill.) K. Koch (Juglandaceae), and *Desmodium canadense* (L.) DC. (Fabaceae). We chose *Q. bicolor* for this study because it exhibited the most diverse herbivory (i.e., the most types of herbivory per specimen). The other focal species we chose was *Onoclea sensibilis* L. from the fern lineage that is sister to seed plants because we wanted to test similar machine learning classifiers on plant species with diverse leaf morphologies. Both species are native to parts of eastern and central North America.

A primer on machine learning

Machine learning involves the study and construction of computer algorithms that can learn and make predictions based on data. Predictors most relevant to our study are in the form of *classifiers*, i.e., algorithms that predict the category to which some input data belong. For example, the input may be a rectangle within an image, and the output category may be one of several, pre-specified types of insect damage. To train a classifier, another algorithm called a *trainer* is given a *training set* consisting of many *training samples* (examples of inputs, together with the correctly annotated corresponding outputs). The trainer then adjusts the parameters of the classifier such that it will generate the correct output for as many of the training sample inputs as possible.

A classifier will typically perform more poorly on previously unseen data than it does on the training set. When this occurs, the classifier is said to have *overfit* the training set. In contrast, a classifier whose performance on new data is similar to its performance on the training set is said to *generalize* well. Good generalization is easier to achieve when the number of parameters of the classifier is small compared to the number of examples in the training set. In other words, good generalization calls for simple classifiers and large amounts of data.

Detection versus classification

Only the relatively small parts of a specimen image that contain insect damage are relevant to damage type classification, and a machine learning system must also learn how to detect these parts. Thus, an automatic analyzer must address two problems that are conceptually distinct: (1) find the damage and (2) determine its type. Finding the damage is called a *detection* problem, and determining

damage type is a *classification* problem. A detector takes an entire specimen image as its input and outputs a collection of boxes that are likely to contain insect damage, as described in greater detail below. These boxes are similar to those that a human annotator would draw. A classifier then takes each of these boxes in turn and determines which type of damage it contains.

Mathematically, detection and classification are distinct problems in that they compute different types of results. Specifically, a detector is what is called a *regressor* in machine learning. This means that its output is an element out of a potentially infinite set (or at least an extremely large number) of options. Each output option for a detector is a list of image rectangles, and each rectangle can be specified by the two coordinates of its upper-left corner and those of its lower-right corner. Thus, the detector outputs a list of quadruples of numbers. In contrast, a classifier outputs one of a small number of predefined categories (damage types), so the options it has from which to choose an answer are limited. Because regressors must choose among a much larger set of possible answers than classifiers, they typically require more data to train than classifiers do. In other words, regression is more data-hungry than classification.

Several detectors have been developed recently for a variety of problems (Erhan et al., 2014; Girshick et al., 2014; Girshick, 2015; Ren et al., 2017). An interesting lesson learned from this research is that for problems that involve both detection and classification it is more effective and efficient to compute the solution to these two problems with a single system that simultaneously finds boxes *and* determines the category for each of them (Hariharan et al., 2015; Liu et al., 2016; Redmon et al., 2016), rather than first detecting regions of interest and then classifying them.

However, a single system requires very large amounts of data to be trained well. Table 2 shows that we only had access to modest amounts of annotated data, and this limitation required some compromises in our experiments. Before discussing these compromises, it will be useful to describe the nature of our data annotations.

Herbivory annotation

We downloaded all of the high-resolution, digitized specimens available for *Q. bicolor* and *O. sensibilis* from the SouthEast Regional Network of Expertise and Collections (SERNEC; http://sernecportal.org/portal/; Appendix S1). We then manually annotated 109 images of *Q. bicolor* and 15 images of *O. sensibilis* specimens for all instances of clear insect damage. We used the VGG Image Annotator version 2.0.4 (Dutta and Zissermann, 2019) to draw bounding boxes and assigned a damage category to each box. Table 2 shows the number of instances annotated for each of the six damage types we investigated in this study (leaf margin feeding, leaf interior feeding, skeletonization, stippling, blotch mines, and serpentine mines). The categories "normal margin" and "normal interior" represent boxes drawn on undamaged parts of the specimens and constitute "negative examples" for the training set; that is, examples of what damage regions do *not* look like. For an example of each annotation category, see Fig. 1.

Detection and classification experiments

High-quality automatic detection and classification of leaf-damage boxes require more annotated data than we had. Therefore, we split the detection and classification experiments into two groups and made simplifications to each of them. Specifically, in the first group of experiments, we simultaneously detected and classified only two types of damage, namely, leaf margin feeding and leaf interior feeding, for each of which there are several hundred available annotations (Table 2). In the second group of experiments, we classified each of a set of manually annotated boxes into one out of eight categories (six damage categories and two "no damage" categories, see Table 2). Experiments in the first group demonstrate that joint detection and classification can work, although data scarcity required us to reduce the number of damage categories from eight to two for this complex task. Experiments in the second group show that even the limited amount of data at our disposal is enough to address the eight-category classification problem, but only when classification is performed separately from detection. For both experiments, we extracted a subset of the data to be used for training, and we used the rest for performance evaluation. This data split is described in more detail below for each experiment.

Data split for detection experiments—Of the 120 images that were manually annotated, we retained the 105 images that contained some margin feeding and interior feeding annotations. Of these images, we selected 83 at random for training and kept the remaining 22 for testing. The detection methods we used in our experiments work best with non-overlapping annotation rectangles. Because of this, we discarded a small fraction of annotations when overlaps occurred. In the end, training images contained 348 instances of margin feeding and 444 of interior feeding. Test images contained 106 instances of margin feeding and 124 of interior feeding.

To sharpen the detector's ability to distinguish between these types of damage (margin feeding and interior feeding) and normal parts of a leaf, or to distinguish them from other types of damage, we also provided what are called *hard negative examples* to the detector. Specifically, we added a third category, which can be interpreted as a "null" category (that is, neither interior feeding nor margin feeding) and placed in it all the boxes in the training images that were annotated as different types of damage by the human annotators. The detector was then trained to detect boxes of the three types (margin feeding, interior feeding, and other types of damage), but all detections from the null category were eventually ignored during testing. This mechanism allows the detector to learn a more detailed understanding of the boundary between, e.g., an instance of margin feeding and an instance of a normal margin. Intuitively, negative examples teach the detector not only what margin feeding looks like, but also what it does *not* look like. Negative examples are most useful to a training algorithm when they are *hard*, that is, when they are similar to positive examples, because they are closer to the boundary in question, and therefore help to pin it down. Because of this, we did not include *easy* negative examples, such as

TABLE 2. Number of leaf-damage rectangles annotated for each category and species in our data set. The last two columns denote "no damage" categories. Each rectangle was drawn to enclose the leaf damage tightly.

Species	Margin feeding	Interior feeding	Skeletonization	Stippling	Blotch mines	Serpentine mines	Normal margin	Normal interior
Onoclea sensibilis	39	28	8	0	11	11	46	25
Quercus bicolor	456	616	215	184	28	18	206	223

healthy tissue or healthy leaf margins. This strategy has been found useful in a variety of computer vision systems (Dong et al., 2017).

Data split for classification experiments—Annotation boxes are rectangles of arbitrary sizes that are drawn to enclose the leaf damage as tightly as possible. We used deep neural networks, described in more detail later, as classifiers. Because a typical deep neural network expects an input box of a fixed size, we wrote software that samples squares of 224 × 224 pixels from each of the annotation boxes. This is the smallest image size expected by a wide range of current deep learning architectures. For annotation boxes smaller than this size, a 224 × 224-pixel square was extracted from the original image, with the annotation box at its center. For larger annotation boxes, several 224 × 224-pixel squares were sampled, providing a crude form of data augmentation.

For the margin feeding, interior feeding, and normal margin categories, it was important that each box contain both leaf and background. To this end, we wrote image segmentation software to identify the boundaries between leaf and background, and only sampled boxes for which the leaf covered at least 40% of the area. This software is described in Appendix 1. Although our software is fully adequate for our purposes, recent specimen segmentation methods (White et al., 2020) may produce even more accurate plant/background separation in more demanding scenarios.

Our sampling procedure resulted in a data set with 6616 samples that were split uniformly at random into 4157 samples for training, 1354 for validation, and 1105 for testing. This split was executed at the annotation-box level so that all boxes from a given annotation fall into exactly one of the three sets. A classifier was trained on training samples and evaluated on test samples. Validation samples are used during training to estimate when the classification algorithm has reached its best generalization performance.

Experimental machine learning architectures and pre-training detection network architecture

We used the Single Shot Multibox Detector (SSD) with a VGG16 base classification network (Liu et al., 2016) to simultaneously detect and classify interior feeding and margin feeding. This detector samples the set of all possible boxes and produces thousands or tens of thousands of box hypotheses in the input image. Box hypotheses differ by both their location in the image and their shape and size. The base classification network then classifies each box hypothesis and computes a score for each of them, which measures the network's confidence in the classification result. The numerical values of these scores do not admit an immediate interpretation in isolation: during training, the network is penalized every time it assigns a small score to a correct result or a large score to an incorrect one. As a consequence, the network learns to assign larger scores to correct results than it does to incorrect ones. During testing, each high-score box is output as a detection, together with the category that yielded the maximum score for that box. A box is not output if it has a score that is high, but lower than that of another box with which it overlaps. This criterion prevents overlapping detections to be output.

Classification network architecture—The damage type classifier was adapted from a residual net architecture (He et al., 2016) with 18 layers. This is a standard, small deep network that has shown good performance in a variety of image recognition tasks. Originally designed to distinguish among 1000 object categories, we adapted it to work with our eight damage type categories instead of the original 1000. The only layer in the network that depends on the number of categories is the last one, so we only had to change the structure of this layer. Details of the classifier's architecture are given in Appendix 2.

Pre-training—The detection network we use in our experiments has 11,180,616 parameters to be estimated during training. These parameters are the coefficients and bias parameters in the convolution kernels listed in Appendix 1. This necessitates large amounts of annotated data for training. In order to address the disparity between the size of our training set and the number of parameters to train, we used what is called "pre-training" in the literature. Specifically, we started our training with neural networks (both for the base classifier in the SSD detector and for the damage type classifier) that had already been trained on a classification task that is entirely different from the target task and for which ample data are available, namely, the ImageNet database (Russakovsky et al., 2015). This data set contains millions of labeled images of ordinary scenes and objects in thousands of categories. This initial phase is called the pre-training phase, and its purpose is to replace random values for the network parameters with values that at least relate to plausible images. We then further trained the networks on our annotated images. This technique, which is commonly used in computer vision, has also proven useful in the domain of plant species identification (Carranza-Rojas et al., 2017).

RESULTS

Damage detection

Simultaneous detection and classification, as performed by the SSD (Liu and Stiling, 2006) or similar detectors (Sermanet et al., 2013; Redmon et al., 2016), requires large amounts of annotated images per class. In our data, only interior feeding and margin feeding damage have several hundred manually annotated boxes, and after the data split these two categories have 392 and 321 training examples, respectively. These numbers are barely enough for detection classification and, even so, we had to use both pre-training and hard negative examples, as described earlier, in order to achieve the performance described in the experiments. Without pre-training, there is just not enough data for the neural network parameter settings to converge to repeatable values during training. Hard negative examples, on the other hand, improve performance more modestly, by increasing mean average precision (reported later) by 13% on average. Because of this scarcity of annotated data, we limited our detection experiments to these two categories, as explained above.

Detailed results—Figures 2 through 4 show all the experimental results for the 10 images for which we asked the annotator to exhaustively annotate all the margin feeding and interior feeding boxes. Specifically, Fig. 2 shows all true-positive detections of interior damage. These are instances where the human annotator and detection algorithm agree, in the sense that their two boxes overlap at least 50% of the size of the smaller of them. Figure 2A shows results for interior damage (green boxes; dashed line for annotations and solid line for predictions) and Fig. 2B shows results for margin damage (red boxes; dashed line for annotations and solid

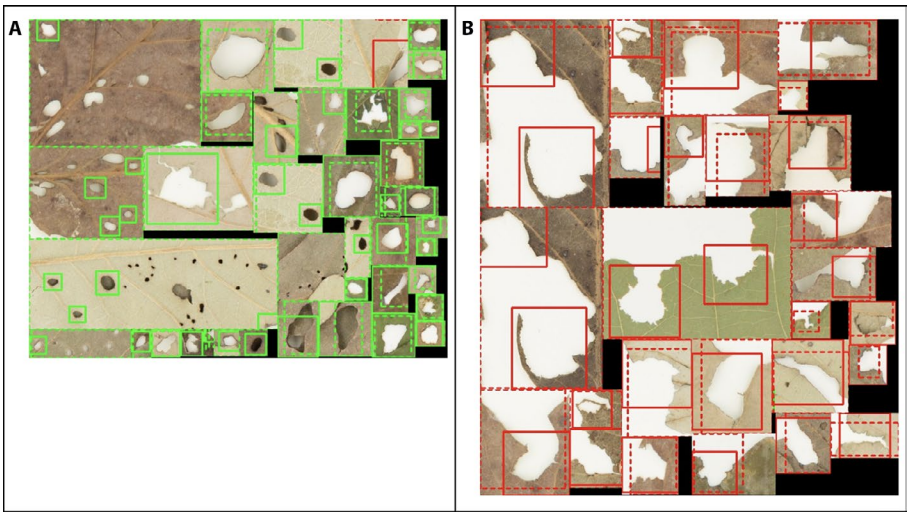

FIGURE 2. A collage of all true-positive detections for interior leaf damage (A) and leaf margin damage (B) in 22 test images. A true positive is an instance of damage that was annotated by a human *and* detected by the detector. In these images, dashed boxes represent human annotations, and solid boxes represent detector results. Green boxes represent interior feeding, and red boxes represent margin feeding.

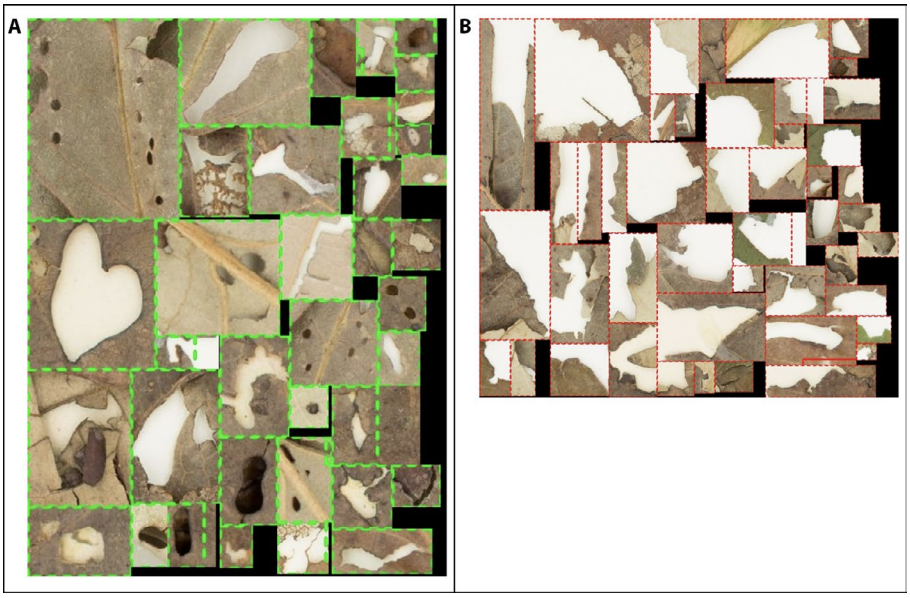

FIGURE 3. A collage of all false-negative detections for interior leaf damage (A) and leaf margin damage (B) in 22 test images. A false negative is an instance of damage that was annotated by a human but was missed by the detector. In these images, dashed boxes represent human annotations, and solid boxes represent detector results. Green boxes represent interior feeding, and red boxes represent margin feeding.

line for predictions). Looking at the example in the top left box in Fig. 2A, we see that a human annotator may annotate an entire area of sparse damage with a single, large box, while the algorithm may detect small, separate sub-areas as distinct instances or, in this particular case, miss several of the sub-areas altogether. This is also the case for margin damage (Fig. 2B).

Figure 3 shows all false negatives, that is, all the human annotation boxes that the detector missed. Comparison of Figs. 2A and 3A suggests that the detector "got used" to identifying interior damage as being more or less oval holes and did not have enough examples of more complex shapes to adapt to these during training. Perusal of Figs. 2B and 3B suggests that the shapes of margin feeding instances are highly varied, and many more examples would be necessary for the network to learn an appropriate model of them.

Figure 4 shows all false positives, that is, image regions that the detector thought to be examples of interior (Fig. 4A) or margin (Fig. 4B) feeding, but that the human annotator did not mark. First, the false-positive rate is 35% (12 out of 34 boxes). Second, some of the detections seem to be genuine damage, although not necessarily of the declared type. For example, many of the oval lesions seen in Fig. 4 could be from disease (perhaps fungal) rather than insects. An additional consideration about Fig. 4 is the misclassification of digits from the printed text in the specimen images as either interior (Fig. 4A) or margin (Fig. 4B) feeding. It would be relatively straightforward to develop image preprocessing routines that exclude these areas from consideration. We left these detections in the figure because they clearly point to overfitting; the digits in Fig. 4A all contain closed ovals, and this reinforces the observation we made for Fig. 3A, that the detector takes anything that looks like an oval and classifies it as interior damage. Similarly, all digits in Fig. 4B contain at least some open digits, whose shape could be interpreted as the profile of part of a leaf's boundary. In this context, overfitting means that the detector formed a naive model of what these two types of damage look like, based on sparse data. More data would be needed to develop a more robust model.

Issues with aggregate measures of performance—It is standard practice in machine learning papers to give aggregate measures of performance such as overall error rates or accuracy. For a detector, one very popular measure is the average *intersection over union* (IoU): A human annotation box H and a detected box D are said to match if they overlap, and the extent of overlap is measured by the area of the overlap region (the intersection) divided by the area of the union of the two regions. If H and D are identical, this measure is equal to 1, and if H and D do not overlap the measure is 0. The average IoU over all boxes in the test set then gives an overall measure of performance.

For damage detection, this measure would be very misleading, and the examples discussed above show why. What constitutes "a single instance" of damage is a poorly defined concept, and even when the human and the detector happen to disagree, both results are sometimes plausible. For instance, even if the detector had

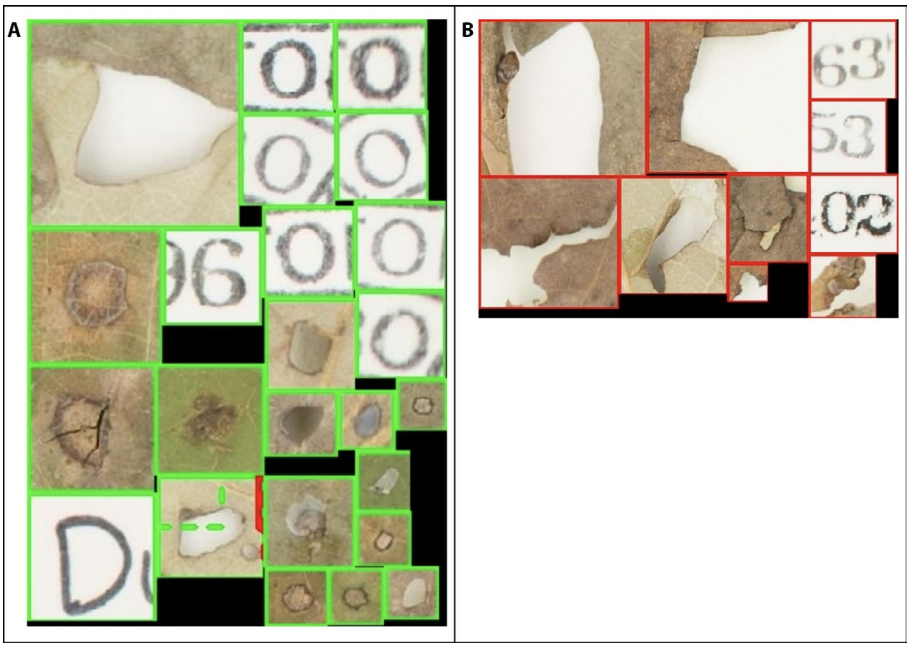

FIGURE 4. A collage of all false-positive detections for interior leaf damage (A) and leaf margin damage (B) in 22 test images. A false positive is an image region that was detected by the algorithm but not marked as either interior feeding or margin feeding by the human annotator. In these images, dashed boxes represent human annotations, and solid boxes represent detector results. Green boxes represent interior feeding, and red boxes represent margin feeding.

found all the small holes in the top-left box of Fig. 2A, the intersection would be the aggregate area of the small holes, and the union would be the area of the human annotation. The ratio of these two quantities is small, suggesting poor performance. However, most people would agree that either interpretation (one large box or several small ones) effectively identifies the same damage. Similar considerations hold for other aggregate measures of performance.

Nonetheless, we report *mean average precision* (mAP), a popular measure of performance for detection algorithms, for the sake of completeness. To this end, we note that a box is detected when the detection score computed by the neural network exceeds some threshold. Increasing this threshold reduces the number of detections and therefore improves precision (fraction of detected boxes that are correct) at the expense of recall (fraction of all existing true boxes that are detected). Decreasing the threshold has the opposite effect.

We then calculate precision and recall on the test set for a sampling of all possible thresholds. The average precision (AP) for one type of damage is the mean of precision values integrated over all recall values, and the mean AP (mAP) is the mean of the average precision values over the two types, weighted by the number of instances in each type. We measured APs of 34% for margin feeding (106 true boxes) and 54% for interior feeding (124), for a mAP value of 45%.

Damage classification

Our damage classifier network returned the correct answer in 81.5% of the 1105 testing samples (those that were *not* used for training or validation). The confusion matrix for our results is shown in Table 3. Each row in the table corresponds to a true answer, and each column corresponds to the answer given by the classifier. Each entry in the table is the number of times an instance from a row category was classified into a column category. For instance, the entry 39 in the first column means that a normal margin was mistaken for an instance of margin feeding in 39 cases. An ideal confusion matrix would have non-zero entries only on the diagonal, which contains the number of correct classifications (boldfaced in Table 3).

The damage classifier was quite accurate for those categories with at least 100 test sample boxes (margin feeding, interior feeding, normal margin, normal interior). These are categories for which the entries in the rows of Table 3 add up to more than 100. As explained earlier, we split the sample boxes in the data set into training, validation, and test sets, and placed four boxes into the training set for each box in the test set. Therefore, the four categories above have at least 400 training samples each.

The classifier error rate on these four categories was 8.8%. The main contributor to this error rate was the confusion between normal margin and margin feeding (39 in the first column and 14 in the next-to-last column). The results for more sparsely represented categories are not very meaningful statistically except for blotch mines, which are very often confused with skeletonization damage (81 cases). Interestingly, the damage that insects make within blotch mines is skeletonization, so the error is not overly surprising and indicates to us that critical refinements will be needed in specifying our damage types moving forward.

TABLE 3. Confusion matrix with correct/incorrect predictions made by our classifier on the 1105 sample boxes (of size 224 × 224 pixels) in our test set. Each rectangle counted in Table 2 produced a variable number of sample boxes (see text for details).[a]

	Margin feeding	Interior feeding	Skeletonization	Stippling	Blotch mines	Serpentine mines	Normal margin	Normal interior
Margin feeding	**114**	13	0	1	0	0	14	0
Interior feeding	4	**128**	2	0	0	0	0	0
Skeletonization	1	1	**31**	2	2	0	0	0
Stippling	1	0	3	**30**	2	0	0	1
Blotch mines	5	3	81	13	**40**	0	0	1
Serpentine mines	0	0	1	0	1	**3**	0	1
Normal margin	39	0	0	0	0	0	**334**	1
Normal interior	1	0	0	1	0	0	6	**224**

[a]Boldfaced numbers show correct classifications.

DISCUSSION

We developed automated techniques that detect insect damage for two categories (leaf interior feeding and leaf margin feeding) and categorize a diverse array of insect herbivory by damage morphology from pre-segmented damage boxes. To the best of our knowledge, this is the first attempt to develop automated techniques that will scan pressed plant specimens and identify clear instances of insect damage. The overall performance observed in this study, which includes only a modest data set, is promising. Specifically, results of detection are mixed (45% mAP and Figs. 2–4), while herbivory classification results are objectively promising, with an accuracy of 81.5% in an eight-class experiment. Future studies can improve the accuracy of the tools developed here.

The detection algorithm performed qualitatively well at detecting damage where a human annotator detected damage. This is a promising initial step in the development of a model that can locate specific types of damage. However, for the detector, it is clear that much more data will be needed for performance that is accurate enough to be used to extract data to test ecological and evolutionary hypotheses. For both detector and classifier, the confusion between damaged and normal margins suggests that it may be useful to introduce a damage mask to focus the damage classifier's attention on the margin itself: most of a margin box contains pixels other than pixels on the margin (that is, it contains background pixels or healthy leaf pixels), and these non-margin pixels are irrelevant to classification. The mask would remove these pixels from the classifier's input.

The confusion between blotch mines and skeletonization indicates that explicit texture and color descriptors may help a classifier make more nuanced distinctions. An example of such a descriptor is a correlogram. For instance, a color correlogram is a function of three variables: two colors (c and c′) and a distance (d). The correlogram specifies for each such triplet the frequency of instances in which colors c and c′ appear in the image separated by distance d. Thus, a correlogram describes the spatial distribution of colors to second order (occurrence of color pairs). Correlograms have been shown to enhance classification and retrieval of images (Huang et al., 1997). In our context, we would provide a correlogram as an additional input to the classifier, to make information about spatial color distribution more explicit. We did not use either damage masks or correlograms in our preliminary experiments.

The issues described earlier concerning aggregate measures of performance are not just a technicality, but rather suggest a possible fundamental shift in the entire approach. To understand this point, consider that a detector is trained by minimizing a measure of the average loss one incurs every time a detection mistake is made. In the SSD system we used for our experiments (Liu et al., 2016), one of the components of this loss penalizes errors in box localization. Because of the complex shapes that regions of damage exhibit in specimen images, the notion of a "true box" is simply not well defined, and the localization error is therefore also ill-defined as a consequence. No amount of data will make this weakness go away.

A better approach may be to cast the damage-detection problem as a problem of *image segmentation*: Dispose of boxes altogether, and instead build a system that classifies each *pixel* of the image into "no damage," "interior damage," or "margin damage" (or more categories, once more data are available). The output from the segmentation system is then a new image, a map that represents region areas in their full complexity, one pixel at a time. The main drawback of a segmentation approach is that manual image annotations would also have to be in the form of images, in which every pixel that belongs to a damage region is "painted" with a label that identifies the damage type. The annotation burden would be much greater, but it may be possible to develop a user interface that allows an annotator to "paint" damage regions efficiently. We plan to investigate related approaches in future work.

Regardless of the exact approach used, we knew that one of the key challenges of this project would be to annotate enough instances of different types of leaf damage to accurately train the damage detector and the damage classifier. The size of the training set is critical. As other researchers have discovered––even in the domain of leaf classification––thousands of images are needed for the simpler task of binary classification of an entire specimen image into mercury-stained or unstained (Schuettpelz et al., 2017), and hundreds of thousands are needed for more complex tasks, such as species classification (Carranza-Rojas et al., 2017). As discussed above, box detection is even harder. We expect that tens of thousands of annotations per damage type will eventually be necessary for high-quality results, and it will take time and effort to accomplish this in follow-up studies.

Anecdotally, it can be assumed that botanists may prefer to collect specimens with little or no insect damage. For this reason, any herbivory quantified on specimens is expected to be a conservative estimate of total herbivory experienced by plants. Thus, it is important to acknowledge and/or account for these biases as machine learning methods continue to be developed. Importantly, the tool we present could be used on any pressed leaves, opening the possibility for automated scoring of percent leaf herbivory in situations where this down-bias is not an issue. For example, our techniques could be applied to leaves from field or greenhouse studies, which is perhaps a more common application in ecology than measuring herbivory on herbarium specimens (e.g., Turcotte et al., 2014a, b; Johnson et al., 2016).

Paleobotanists have used insect damage data from fossils to quantify changing patterns of insect diversity during periods of warming in the fossil record (Wilf and Labandeira, 1999; Labandeira et al., 2002; Wilf et al., 2006). We show here that not only could herbaria offer an analogous resource for examining modern changes in herbivory, but also that we may be able to use automated techniques to extract damage types from specimens, allowing for the possibility of "big data" extraction. This is important because some limited data that are available on insect abundances (Boyle et al., 2019; Wepprich et al., 2019) and biomass (Hallmann et al., 2017; Lister and Garcia, 2018) over time suggest that insects are in decline in the Anthropocene. However, these studies are highly debated today (Ries et al., 2019; Wepprich, 2019; Willig et al., 2019), and biomass studies often represent data collected from only two time points, one before and one after the acceleration of climate change. If the automated techniques described here are developed further and harnessed to their full potential, herbaria will offer an unprecedented opportunity to assess changing insect damage and diversity across broad scales of space, time, and plant phylogeny.

ACKNOWLEDGMENTS

The authors thank the reviewers for their constructive and insightful comments, which helped improve this paper significantly. In particular, one of the reviewers pointed out the limitations of the box detection approach itself, and this observation led to our considerations on detection versus segmentation in the Discussion

section. We would like to thank Conrad Labandeira for guidance on scoring herbivory categories, and iDigBio (http://www.idigbio.org) along with its various Thematic Collections Networks for making herbarium specimens digitally available for study. This work was supported in part by the National Science Foundation under Grant No. 1909821 and a Duke University Arts and Sciences Council Faculty Research Grant to K.M.P.

AUTHOR CONTRIBUTIONS

E.K.M., C.T., and K.M.P. designed the study, performed analyses, and wrote the paper. C.T. and S.Y. designed and implemented machine learning approaches.

DATA AVAILABILITY

All data are published in Appendices 1, 2, and S1. All code are available on GitHub (https://github.com/Oliver-ss/Applying-machine-learning-to-investigate-long-term-insect-plant-interactions)

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

APPENDIX S1. Data associated with herbarium specimens used in this study.

LITERATURE CITED

Beaulieu, C., C. Lavoie, and R. Proulx. 2018. Bookkeeping of insect herbivory trends in herbarium specimens of purple loosestrife (*Lythrum salicaria*). *Philosophical Transactions of the Royal Society B: Biological Sciences* 374: 20170398.

Boyle, J. H., H. J. Dalgleish, and J. Puzey. 2019. Monarch butterfly and milkweed declines substantially predate the use of genetically modified crops. *Proceedings of the National Academy of Sciences USA* 116: 3006–3011.

Carranza-Rojas, J., H. Goeau, P. Bonnet, E. Mata-Montero, and A. Joly. 2017. Going deeper in the automated identification of herbarium specimens. *BMC Evolutionary Biology* 17: 181.

Cebrian, J., and J. Lartigue. 2004. Patterns of herbivory and decomposition in aquatic and terrestrial ecosystems. *Ecological Monographs* 74: 237–259.

Coley, P. D., J. P. Bryant, and F. S. Chapin. 1985. Resource availability and plant antiherbivore defense. *Science* 230: 895–899.

Cooper-Driver, G. A. 1978. Insect-fern associations. *Entomologia Experimentalis et Applicata* 24: 310–316.

Currano, E. D., C. C. Labandeira, and P. Wilf. 2010. Fossil insect folivory tracks paleotemperature for six million years. *Ecological Monographs* 80: 547–567.

Cyr, H., and M. L. Face. 1993. Magnitude and patterns of herbivory in aquatic and terrestrial ecosystems. *Nature* 361: 148–150.

Diamond, S. E., R. R. Dunn, S. D. Frank, N. M. Haddad, and R. A. Martin. 2015. Shared and unique responses of insects to the interaction of urbanization and background climate. *Current Opinion in Insect Science* 11: 71–77.

Dong, Q., S. Gong, and X. Zhu. 2017. Class rectification hard mining for imbalanced deep learning. *In* Proceedings of the IEEE International Conference on Computer Vision in Venice, Italy, 2017, 1851–1860. Institute of Electrical and Electronics Engineers, Washington, D.C., USA.

Dutta, A., and A. Zissermann. 2019. The VIA annotation software for images, audio, and video. *In* Proceedings of the 27th ACM International Conference on Multimedia, 2276–2279. https://doi.org/10.1145/3343031.3350535.

Ehrlich, P. R., and P. H. Raven. 1964. Butterflies and plants: A study in coevolution. *Evolution* 18: 586–608.

Erhan, D., C. Szegedy, A. Toshev, and D. Anguelov. 2014. Scalable object detection using deep neural networks. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2147–2154. Institute of Electrical and Electronics Engineers, Washington, D.C., USA.

Feeny, P. 1976. Plant apparency and chemical defense. *In* J. W. Wallace *and R. L. Mansell [eds.],* Biochemical interaction between plants and insects, 1–40. Plenum Press, New York, New York, USA.

Futuyma, D. J., and A. A. Agrawal. 2009. Macroevolution and the biological diversity of plants and herbivores. *Proceedings of the National Academy of Sciences USA* 106: 18054–18061.

Girshick, R. 2015. Fast r-cnn. *In* Proceedings of the IEEE International Conference on Computer Vision in Santiago, Chile, 2015, 1440–1448. Institute of Electrical and Electronics Engineers, Washington, D.C., USA.

Girshick, R., J. Donahue, T. Darrell, and J. Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 580–587. Institute of Electrical and Electronics Engineers, Washington, D.C., USA.

Hallmann, C. A., M. Sorg, E. Jongejans, H. Siepel, N. Hofland, H. Schwan, W. Stenmans, et al. 2017. More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLoS ONE* 12: e0185809.

Hariharan, B., P. Arbeláez, R. Girshick, and J. Malik. 2015. Hypercolumns for object segmentation and fine-grained localization. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 447–456. Institute of Electrical and Electronics Engineers, Washington, D.C., USA.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778. Institute of Electrical and Electronics Engineers, Washington, D.C., USA.

Heberling, J. M., and B. L. Isaac. 2017. Herbarium specimens as exaptations: New uses for old collections. *American Journal of Botany* 104: 963–965.

Heberling, J. M., L. A. Prather, and S. J. Tonsor. 2019. The changing uses of herbarium data in an era of global change: An overview using automated content analysis. *BioScience* 69: 812–822.

Huang, J., S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. 1997. Image indexing using color correlograms. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, USA, 1997, 762–768. Institute of Electrical and Electronics Engineers, Washington, D.C., USA.

Johnson, M. T., J. A. Bertrand, and M. M. Turcotte. 2016. Precision and accuracy in quantifying herbivory. *Ecological Entomology* 41: 112–121.

Kozlov, M. V., V. Lanta, V. Zverev, K. Rainio, M. A. Kunavin, and E. L. Zvereva. 2017. Decreased losses of woody plant foliage to insects in large urban areas are explained by bird predation. *Global Change Biology* 23: 4354–4364.

Labandeira, C. C., K. R. Johnson, and P. Wilf. 2002. Impact of the terminal Cretaceous event on plant-insect associations. *Proceedings of the National Academy of Sciences USA* 99: 2061–2066.

Lang, P. L., F. M. Willems, J. Scheepens, H. A. Burbano, and O. Bossdorf. 2019. Using herbaria to study global environmental change. *New Phytologist* 221: 110–122.

Lister, B. C., and A. Garcia. 2018. Climate-driven declines in arthropod abundance restructure a rainforest food web. *Proceedings of the National Academy of Sciences USA* 115: E10397–E10406.

Liu, H., and P. Stiling. 2006. Testing the enemy release hypothesis: A review and meta-analysis. *Biological Invasions* 8: 1535–1545.

Liu, W., D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. 2016. SSD: Single Shot Multibox Detector. *In*B. Leibe*, J. Matas, N. Sebe, and M. Welling [eds.],* Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol. 9905. Springer, Cham, Switzerland.

Lorieul, T., K. D. Pearson, E. R. Ellwood, H. Goëau, J. F. Molino, P. W. Sweeney, J. Yost, et al. 2019. Toward a large-scale and deep phenological stage annotation of herbarium specimens: Case studies from temperate, tropical, and equatorial floras. *Applications in Plant Sciences* 7: e01233.

Meineke, E. K., and T. J. Davies. 2018. Museum specimens provide novel insights into changing plant–herbivore interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences* 374: 20170393.

Meineke, E. K., C. C. Davis, and T. J. Davies. 2018a. The unrealized potential of herbaria in global change biology. *Ecological Monographs* 88: 505–525.

Meineke, E. K., A. T. Classen, N. J. Sanders, and T. J. Davies. 2018b. Herbarium specimens reveal increasing herbivory over the past century. *Journal of Ecology* 107: 105–117.

Meineke, E. K., T. J. Davies, B. H. Daru, and C. C. Davis. 2018c. Biological collections for understanding biodiversity in the Anthropocene. *Philosophical Transactions of the Royal Society B: Biological Sciences* 374: 20170386.

Moles, A. T., S. P. Bonser, A. G. Poore, I. R. Wallis, and W. J. Foley. 2011. Assessing the evidence for latitudinal gradients in plant defence and herbivory. *Functional Ecology* 25: 380–388.

Otsu, N. 1979. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics* 9(1): 62–66.

Page, L. M., B. J. MacFadden, J. A. Fortes, P. S. Soltis, and G. Riccardi. 2015. Digitization of biodiversity collections reveals biggest data on biodiversity. *BioScience* 65: 841–842.

Panchen, Z. A., R. B. Primack, T. Aniśko, and R. E. Lyons. 2012. Herbarium specimens, photographs, and field observations show Philadelphia area plants are responding to climate change. *American Journal of Botany* 99: 751–756.

Primack, D., C. Imbres, R. B. Primack, A. J. Miller-Rushing, and P. Del Tredici. 2004. Herbarium specimens demonstrate earlier flowering times in response to warming in Boston. *American Journal of Botany* 91: 1260–1264.

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. 2016. You only look once: Unified, real-time object detection. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 779–788. Institute of Electrical and Electronics Engineers, Washington, D.C., USA.

Ren, S., K. He, R. Girshick, and J. Sun. 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39: 1137–1149.

Ries, L., E. F. Zipkin, and R. P. Guralnick. 2019. Tracking trends in monarch abundance over the 20th century is currently impossible using museum records. *Proceedings of the National Academy of Sciences USA* 116: 13745–13748.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, et al. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* 115: 211–252.

Schuettpelz, E., P. B. Frandsen, R. B. Dikow, A. Brown, S. Orli, M. Peters, A. Metallo, et al. 2017. Applications of deep convolutional neural networks to digitized natural history collections. *Biodiversity Data Journal* 5: e21139.

Sermanet, P., D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. 2013. OverFeat: Integrated recognition, localization and detection using convolutional networks. arXiv 1312.6229 [Preprint]. Published 21 December 2013 [accessed 29 May 2020]. Available from: https://arxiv.org/abs/1312.6229.

Smith, A. R. 1978. Color gamut transform pairs. *Computer Graphics* 12(3): 12–19.

Soltis, D. E., and P. S. Soltis. 2016. Mobilizing and integrating big data in studies of spatial and phylogenetic patterns of biodiversity. *Plant Diversity* 38: 264–270.

Thiers, B.2020. Index Herbariorum: A global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium. Website http://sweetgum.nybg.org/science/ih/ [accessed 4 March 2020].

Turcotte, M. M., C. J. Thomsen, G. T. Broadhead, P. V. Fine, R. M. Godfrey, G. P. Lamarre, S. T. Meyer, et al. 2014a. Percentage leaf herbivory across vascular plant species. *Ecology* 95: 788–788.

Turcotte, M. M., T. J. Davies, C. J. Thomsen, and M. T. Johnson. 2014b. Macroecological and macroevolutionary patterns of leaf herbivory across vascular plants. *Proceedings of the Royal Society B: Biological Sciences* 281: 20140555.

Wepprich, T. 2019. Monarch butterfly trends are sensitive to unexamined changes in museum collections over time. *Proceedings of the National Academy of Sciences USA* 116: 13742–13744.

Wepprich, T., J. R. Adrion, L. Ries, J. Wiedmann, and N. M. Haddad. 2019. Butterfly abundance declines over 20 years of systematic monitoring in Ohio, USA. *PLoS ONE* 14: e0216270.

White, A. E., R. B. Dikow, M. Baugh, A. Jenkins, and P. B. Frandsen. 2020. Generating segmentation masks of herbarium specimens and a data set for training segmentation models using deep learning. *Applications in Plant Sciences* 8(6): e11352.

Wilf, P., and C. C. Labandeira. 1999. Response of plant-insect associations to Paleocene-Eocene warming. *Science* 284: 2153–2156.

Wilf, P., C. C. Labandeira, K. R. Johnson, P. D. Coley, and A. D. Cutter. 2001. Insect herbivory, plant defense, and early Cenozoic climate change. *Proceedings of the National Academy of Sciences USA* 98: 6221–6226.

Wilf, P., C. C. Labandeira, K. R. Johnson, and B. Ellis. 2006. Decoupled plant and insect diversity after the end-Cretaceous extinction. *Science* 313: 1112–1115.

Willig, M., L. Woolbright, S. J. Presley, T. D. Schowalter, R. B. Waide, T. H. Scalley, J. K. Zimmerman, et al. 2019. Populations are not declining and food webs are not collapsing at the Luquillo Experimental Forest. *Proceedings of the National Academy of Sciences USA* 116: 12143–12144.

Willis, C. G., E. R. Ellwood, R. B. Primack, C. C. Davis, K. D. Pearson, A. S. Gallinat, J. M. Yost, et al. 2017. Old plants, new tricks: Phenological research using herbarium specimens. *Trends in Ecology & Evolution* 32: 531–546.

Zangerl, A. R., and M. R. Berenbaum. 2005. Increase in toxicity of an invasive weed after reassociation with its coevolved herbivore. *Proceedings of the National Academy of Sciences USA* 102: 15529–15532.

APPENDIX 1. Specimen segmentation algorithm.

Our algorithm for separating pixels on the specimen from pixels in the background of specimen images is based on the consideration that the background has an unsaturated color: white or off-white in most images, and perhaps dark, or even black in others. In contrast, the color of leaves is much more saturated, as it is between green and brown in most cases. Saturation measures the "colorfulness" of a color, i.e., how far it is from gray. We used the hue, saturation, and value (HSV) definition of saturation (Smith, 1978): If a pixel has components R (red), G (green), and B (blue), let M be the maximum among R, G, B, and let m be the minimum. Then chroma is defined as $C = M - m$ and saturation as $S = C/M$ if M is nonzero and $S = 0$ otherwise.

Our segmentation algorithm computes saturation (S) for every pixel in the image and then builds a histogram of it. This histogram typically shows a large peak at low-saturation values for the background and another at higher-saturation values for the specimen. We used Otsu's algorithm (Otsu, 1979) to compute an image-specific threshold τ for separating the two peaks. Comparing the saturation (S) of a pixel to this threshold classifies it into background if $S < \tau$ and if the specimen is otherwise.

APPENDIX 2. Classifier and training details.

Our neural net had an input convolutional layer with a 7×7 kernel, stride 2, batch normalization, a rectified linear unit (ReLU), and a max-pooling layer with a 3×3 kernel and stride 2. Four standard residual blocks followed, each with four convolutional layers with 3×3 kernels, batch normalization, and ReLU. After a 1×1 convolution to adjust the output size to a vector of length 512, a final, fully connected layer changed the number of outputs to 8 (i.e., the number of categories in our experiments). Classification

was achieved by identifying the largest entry in a soft-max transformation of the output. The exact architecture can be downloaded and modified with the following PyTorch commands:

```
model = torchvision.models.resnet18(pretrained = True)
model.fc = torch.nn.Linear(model.fc.in_features, 8)
```

The model was trained with stochastic gradient descent with a constant learning rate of 0.0002 and a momentum of 0.9, and the cross entropy loss was used as the risk function to minimize.