Not that kind of tree: Assessing the potential for decision tree–based plant identification using trait databases

Brianna K. Almeida1,3, Manish Garg2, Miroslav Kubat2, and Michelle E. Afkhami1

PREMISE: Advancements in machine learning and the rise of accessible “big data” provide an important opportunity to improve trait-based plant identification. Here, we applied decision-tree induction to a subset of data from the TRY plant trait database to (1) assess the potential of decision trees for plant identification and (2) determine informative traits for distinguishing taxa.

METHODS: Decision trees were induced using 16 vegetative and floral traits (689 species, 20 genera). We assessed how well the algorithm classified species from test data and pinpointed those traits that were important for identification across diverse taxa.

RESULTS: The unpruned tree correctly placed 98% of the species in our data set into genera, indicating its promise for distinguishing among the species used to construct them. Furthermore, in the pruned tree, an average of 89% of the species from the test data sets were properly classified into their genera, demonstrating the flexibility of decision trees to also classify new species into genera within the tree. Closer inspection revealed that seven of the 16 traits were sufficient for the classification, and these traits yielded approximately two times more initial information gain than those not included.

DISCUSSION: Our findings demonstrate the potential for tree-based machine learning and big data in distinguishing among taxa and determining which traits are important for plant identification.

KEY WORDS: decision tree; information gain; machine learning; plant identification; TRY plant trait database.
widely used methods for species distribution modeling (Vayssières et al., 2000; Elith et al., 2006). Machine learning has also been employed for the identification of lung cancer cell types, bird and frog calls, and tree species based on leaf shape outline (Zhou et al., 2002; Acevedo et al., 2009; Kumar et al., 2012). These applications have all yielded encouraging results with high identification accuracy; however, one limitation of these studies is that they focused on rather narrow taxonomic groups and did not consider the growing online availability (even abundance) of morphological data.

Importantly, not all data are created equal. Existing plant databases show promise for improving plant identification but do not always contain information about the same set of traits for all species, as many entries come from disparate studies. We posit that machine learning can be used to improve plant specimen identification and inform which traits are important for achieving this goal and therefore should be collected in future studies. To accomplish this, a team of plant ecologists and machine learning researchers developed decision trees using morphological traits and descriptions of species based on specimen data from the TRY plant trait database (TRYdb; Kattge et al., 2011). This study (1) informs how the plant sciences community can invest in future research and data entry efforts to increase the database value for plant identification and (2) provides a template for future investigations combining decision trees and big data for trait-based identification.

METHODS

Trait data collection and curation

From TRYdb (Kattge et al., 2011), we selected genera that contained categorical traits that could be used in plant identification, such as flower color, fruit type, and flower sexual syndrome. In total, we compiled data for 35 traits, both vegetative and floral, to create a character matrix (Appendix S1); this initial data set included all species in TRYdb with data on any of these traits. We then filtered our data set to exclude genera that contained fewer than 25 species or species that contained less than 10 of the 35 traits. Finally, we removed traits with greater than 65% missing data, leaving a total of 16 traits available for the construction of our final model (for the final character matrix, see Appendix S2). The final set of plant taxa included 20 genera that spanned diverse species, including trees, grasses, herbs, and sedges.

Decision tree induction

Decision trees rank among the most popular paradigms in machine learning and are frequently used for the automated classification of examples into classes. Formally, a decision tree is a partially ordered set of attribute-value tests, reminiscent of biological keys. In this application of decision trees, our goal was to place species into genera using a series of trait-based tests (e.g., “what is the fruit type of species x?”). However, before the decision tree can be used for classification, it must be induced from a training set consisting of examples (here, plant species), each described by a vector of attribute values (here, species-specific plant traits) and labeled with a class (here, the genus to which that species belongs).

This induction is a recursive process in which a trait to be tested is chosen based on the amount of information it contains for distinguishing between classes or groups of classes (i.e., the information gain). Each decision constitutes an internal node of the tree, where the possible outcomes of the decision divide the training set into two or more subsets, each defined by one outcome. For example, for the test “what is the fruit type of species x?”, the training set would be divided into species with each fruit type (e.g., species with nuts, capsules, schizocarps, etc.). In cases where all examples in the given subset belong to the same class, a leaf node is created and assigned the corresponding class label (here, a genus). If, alternatively, the subset contains examples of two or more classes (genera), the process described above is applied recursively to the subset, creating additional internal nodes.

Depending on the trait tested at a given internal node, diverse decision trees can be created, ranging from large trees with many internal nodes to much smaller trees that reach classes more quickly (with many or few steps to genera classification, respectively). To create a small tree, machine learning employs formulas from Shannon’s information theory at each trait-selection step, which identify traits offering maximum information about the example’s class (Kubat, 2017). As a result, the classifier model then usually contains only tests on the most relevant and non-redundant traits, with a high level of information gain (in other words, those traits that are best for distinguishing among plant groups). Here, we induced the decision trees using the open-access software Weka (version 3.8; Eibe et al., 2016), whose J48 module is a Java version of C4.5, a fundamental decision tree software (Quinlan, 1993). A non-zero rate of error in classification from the unpruned tree would indicate that two or more species are described by the same set of traits but are labeled with different genera.

Pruning decision trees

Once the tree has been induced, pruning can be used to improve its classification performance in terms of portability and generalization by reducing model overfitting to the training data. This process is crucial if we want a tree capable of classifying future examples (species) that were not used during tree induction. In an unpruned decision tree, a leaf node will only be created if all the species reaching it belong to the same genus; otherwise, a new test (i.e., a new internal node) is created that further splits the set of remaining species. A tree is pruned by the removal of lower-level nodes that are unlikely to contribute to this classifier’s performance. Node termination is determined by the confidence factor and minimum number of objects. The confidence factor is a node’s contribution to the tree’s overall error. For example, if the confidence factor is set to 0.25, the node splits if the error rate at that node without splitting is 25% higher than if it does split. The minimum number of objects determines how many examples need to reach a node for splitting to occur. For example, if the minimum number of objects is 2, then at least two species must reach a node before it can be split. Although our results were robust to a wide range of settings (see Appendices S3–S19), here we report the decision tree built with the complete trait data set, a confidence factor of 0.25, and a minimum number of objects of 2, which returned a low error rate.

10-fold cross-validation

We used a routine statistical evaluation technique, 10-fold cross-validation, to test how well the classifier model can place species into genera. To perform the 10-fold cross-validation, the entire data set
was randomly partitioned into 10 equally sized subsets (“folds”). We then constructed a tree based on 90% of the data (nine folds) and tested with the remaining 10% of the data (one fold). We repeated this process holding out each of the 10 unique folds in turn. From each of the 10 trees, we gained an estimate of the classifier model’s ability to place species that were not included in the training set into genera (included in Table 1). By testing with 10 folds, we minimized the possibility that the traits identified as important in the decision tree induction may be due only to the idiosyncrasies of the random divisions in the training and testing sets. To gain an estimate of the future classification behavior of our pruned tree, we determined the average accuracy of these 10 tests (Table 1).

RESULTS

Curated character matrix and decision tree topology

We identified 20 genera with at least 25 species each (mean ± SE = 35 ± 3.75 species/genera). After filtering, the character matrix contained 16 traits and 689 species across these 20 genera. Despite filtering, data were still sparse within the matrix, with each genus having an average of 60.1% (±3.5%) of data available (range: 29–81%). The pruned decision tree built from this data matrix contained 240 leaves (Appendix S20), where each “leaf” is a genus and multiple paths can be taken to the same genus. Similar to polyclave keys, this flexibility allows the decision tree to properly classify species when within-genus trait variation occurs. The pruned tree used seven of the 16 traits we collected from TRY db (i.e., ~40% of the traits) to distinguish among the genera (Table 2). The unpruned decision tree had 401 leaves and used 11 of the 16 traits (Appendix S21), which corresponds to 67% more leaves and 57% more traits than that of the pruned tree. To classify a species into a genus, the pruned tree used three traits on average, which was 16% fewer steps compared with the unpruned tree.

Assignment of species to genera and information gain of traits

A total of 98% of the species in our complete data set were correctly assigned to genera using the unpruned tree. This result indicates that, in the given database, a few species from different genera have the same trait values, which makes it difficult to classify these few species unless additional traits are added that are able to distinguish among these taxa. The unpruned tree is also likely to be overfit to the training data, and thus while this type of tree may be useful for identification of taxa used in its construction, it is unlikely to be able to identify new species. The pruned tree may be better suited to classify new species (those not used in its induction) into the genera present in the tree. Cross-validation estimated that our pruned decision trees can correctly assign 89.1% ± 1.35% of the species not used in its induction into genera (Table 1), suggesting that nine out of 10 new species (i.e., species not included in the induction of the decision tree) are likely to be classified into their genera. The 89% value is likely a conservative estimate of the classification accuracy for this final pruned tree, because the estimate is based on the testing of trees constructed with only nine folds of the data. The frequency of misassignments of species into genera was fairly low (mean across genera ± SE = 14% ± 3%). However, there was substantial variation in the misassignment rates among genera (Fig. 1), with five genera showing misassignment rates of 26–41%. When we compared the proportion of missing information for each genus to the rate of misassignment by the tree using a linear regression, we found that the amount of trait information missing from a genus did not explain a significant amount of the variation in the misassignment frequency among genera ($F_{1,19} = 1.36, P = 0.26$). However, we note that genera with low misassignment rates often had high agreement among species for the value of traits. For example, the genera *Alchemilla*, *Galium*, and *Hieracium* were misassigned 0% of the time and had three to five traits with 100% agreement for their values, while the genera *Ranunculus* and *Silene* were misassigned ~40% of the time and lacked any trait with 100% agreement.

Ultimately, the machine learning algorithm selected seven traits to be included in the final pruned tree. These traits were the flower sexual system, seed morphology, fruit type, apomixis, leaf distribution, species reproduction type, and a component of leaf shape. Details on these traits can be found in Table 2 and Appendix S22. Traits that were included in the pruned decision tree had approximately twice as much initial information gain as the traits that were excluded (Wilcoxon signed-rank test = 10, $P = 0.02$; information gain of included traits = 1.5 ± 0.23, gain of excluded traits = 0.81 ± 0.02), indicating that initial information content may be a reasonable proxy for assessing the value of traits for distinguishing among groups of taxa. However, it is worth noting that two traits with high initial information gain (i.e., plant growth form and flower color) were excluded from the final pruned tree. This is likely due to their redundancy with other traits of high information gain, meaning they differentiate among taxa in a similar way as other traits that have an even higher information gain and thus come earlier in the tree. In addition, when we regressed trait information gain on the proportion of missing information of that trait, we found that the proportion of missing information explained a significant amount of the variation in the initial information gain (Fig. 2; $t_{19} = -4.867, P = 0.0002$), suggesting that more training data could be useful in improving decision tree–based identification.

DISCUSSION

In this study, we assembled a trait data set using readily accessible online data and machine learning to induce a decision tree with the goal of plant identification. The unpruned tree distinguished among the plant genera in our data set (98% of species were correctly placed into their genus), and a pruned version of the tree was capable of
placing most species into correct genera (89% of test species were properly classified). Although using a pruned tree resulted in a small amount of additional misclassification compared to the unpruned tree, it has several advantages. For example, it required fewer steps for classification (three for the pruned vs. five for the unpruned) and is better suited for placing new species into the correct genera because the pruning avoids overfitting. We found that seven traits were sufficient to properly distinguish most of the species in the data set, indicating their importance in identification. Our results showed that the traits with the highest initial information gain were those with the least missing data, suggesting that the amount of missing data can be important for determining the value of different traits when distinguishing among genera, and for emphasizing the gaps where additional data collection may improve the model quality in the future. Missing information is a common challenge in approaches that use big data to address questions for which the data were not originally collected (Wu et al., 2013; Zhang, 2016).

Compared with traditional polyclave keys, machine learning–derived identification could benefit the plant science community in several ways. First, generating traditional keys requires more expert knowledge of the groups being characterized in order to distinguish among them (Austen et al., 2016) than is required for machine learning methods. Machine learning takes advantage of existing databases and computer algorithms, which require a generalized methodological knowledge (e.g., programming skills for manipulating big data and running algorithms) that can be applied to many different taxonomic groups rather than specialized, group-specific knowledge (Thessen, 2016). While we prize taxonomic and natural history knowledge, the level of expertise needed for effectively distinguishing taxa requires a substantial time investment to acquire.

Second, the computer algorithm determines the importance of traits for species identification using machine learning–based methods, which can save researchers time on identification by focusing their efforts on the most influential traits. Note that the human valuation of traits is not completely removed from this form of identification, as the algorithm picks among the traits researchers have decided to collect and enter into the database (Hortal et al., 2007). Because the traits in these databases are collected for many purposes, it is also possible for the algorithms to pull from traits that are not typically used for identification.

Third, the number of steps required for an identification could be decreased to improve efficiency compared with the use of traditional trait-based keys. For example, our proof-of-principle case study pruned tree was able to classify 89% of species, spanning a large part of the plant phylogeny, in five or fewer steps for all genera. Although improvements are definitely needed to increase the accuracy of this approach, the relatively few steps required to go from “species x” to a genus identity in this study highlights the potential of decision tree–based identification.

To our knowledge, this study is the first to use decision tree methods in this way; however, recent efforts to improve plant identification have applied other machine learning techniques, such as deep learning and hierarchical learning, to plant images (Lee et al., 2017; Singh et al., 2018). For example, Fan et al. (2015) developed a structural learning algorithm that identified plants from images using a coarse to fine subset of plant features. The convenience of image-based methods, in which the computer analyzes the picture to “collect” trait data based on pixels, can save researchers time; however, image-based machine learning methods also face challenges, as shown in a study by Lee et al. (2015), where poor leaf condition

### TABLE 2. Description and information content of the 16 plant traits from the TRY plant trait database (TRYdb).

| Traita | Information gain | Descriptionb |
|--------|------------------|--------------|
| Leaf shape 1: outline | 2.3554 | Describes the leaf shape (linear, ovate, reniform). |
| Fruit type | 2.3318 | Describes the fruit type (dry, fleshy). |
| Plant growth form | 2.0645 | Describes the plant growth form (annual, biennial, perennial). |
| Flower color | 1.8353 | Describes the flower color (red, pink, blue). |
| Flower sexual system | 1.7827 | Describes the flower sexual system (perfect, imperfect). |
| Leaf distribution along the axis (arrangement) | 1.5035 | Describes the leaf distribution along the axis (submersed, emergent). |
| Seed morphology | 1.2794 | Describes the seed morphology (smooth, wrinkled). |
| Apomixis | 0.9106 | Describes the presence of apomicts (true, false). |
| Species reproduction types | 0.8188 | Describes the species reproduction types (sporadic, continuous). |
| Leaf shape 5: leaf base | 0.6729 | Describes the leaf shape (oblong, elliptic, lanceolate). |
| Leaf shape 6: petiole type | 0.6571 | Describes the petiole type (sessile, submersed). |
| Dicliny (monoeocious, dioecious, hermaphroditic) | 0.6374 | Describes the dicliny (monoeocious, dioecious, hermaphroditic). |
| Leaf shape 2: outline | 0.5969 | Describes the leaf shape (ovate, reniform, lanceolate). |
| Shoot growth form | 0.507 | Describes the shoot growth form (ascending, prostrate, erect). |
| Fertilization | 0.448 | Describes the fertilization (self-fertile, cross-fertile). |
| Leaf shape 3: pointed/round | 0.3384 | Describes the leaf shape (rounded, cuneate, submersed). |

---

*a The traits shaded in gray were included in the final decision tree by the machine learning algorithm.

*b TRYdb uses trait definitions from Garnier et al. (2017).

*c Three categories of each trait are shown in parentheses after the definition. In cases with more than three possible categories, please see Appendix S22 for additional categories. For the definitions of these categories, consult the TRY website (https://www.try-db.org/TryWeb/Data.php [accessed July 2019]).
(e.g., damaged or wrinkled leaves) caused the misclassification of tree species from leaf pictures. It can also be difficult to capture certain types of information from images, including features that require dissection or direct measurement (Wäldchen et al., 2018). While the trait-based method we employed in this study requires time to measure the traits, challenging features such as these may be more accurately and/or efficiently collected through their direct measurement. The different strengths of direct- and image-based approaches to using machine learning suggest it may be valuable to use them in tandem for improved identification.

Finally, the availability of online data is rapidly increasing (Stajich et al., 2012; Chaudhary et al., 2016; National Ecological Observatory Network [https://www.neonscience.org/]), and recent work has followed TRYdb's lead, establishing new trait-focused databases for other groups of organisms (e.g., Funran for fungal functional trait data; Zanne et al., 2020). Excitingly, in addition to traditional measurements of trait data, new efforts to automate trait data collection from herbarium specimens (see Ott et al., 2020; Weaver et al., 2020 in this collection) will provide a richer and more complete database of trait information for future research. Overall, as the quantity and quality of this type of big data increase for more plant taxa, we expect that the integration of trait-based approaches and machine learning will improve the efficiency of plant identification, facilitating a wide range of plant biology research.

ACKNOWLEDGMENTS

This study was supported by the TRY initiative on plant traits (http://www.try-db.org). The TRY initiative and database is hosted, developed, and maintained by J. Kattge and G. Bönisch (Max Planck Institute for Biogeochemistry, Jena, Germany). TRY is currently supported by DIVERSITAS/Future Earth and the German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Germany. The authors thank D. Revillini, D. Hernandez, and K. Kiesewetter for their feedback on this manuscript.

AUTHOR CONTRIBUTIONS

B.K.A. curated the biological trait data set, conducted the statistical analyses, and took the lead on writing the paper. M.G. conducted the machine learning analysis using the Weka software under the supervision of M.K. M.E.A. oversaw the organization/statistical analysis of the trait data and substantially contributed to writing the paper. M.E.A. and M.K. conceptualized the project. All authors contributed to the manuscript writing.

DATA AVAILABILITY STATEMENT

All data used in this paper are available in the supporting information and through the TRY plant trait database (http://www.try-db.org).

SUPPORTING INFORMATION

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

APPENDIX S1. Original character matrix and trait set.
APPENDIX S2. Final character matrix.

APPENDIX S3. Pruned tree: confidence factor 0.2, minimum number of objects 2.

APPENDIX S4. Pruned tree: confidence factor 0.2, minimum number of object 5.

APPENDIX S5. Pruned tree: confidence factor 0.2, minimum number of objects 8.

APPENDIX S6. Pruned tree: confidence factor 0.3, minimum number of objects 2.

APPENDIX S7. Pruned tree: confidence factor 0.3, minimum number of objects 5.

APPENDIX S8. Pruned tree: confidence factor 0.3, minimum number of objects 8.

APPENDIX S9. Pruned tree: confidence factor 0.4, minimum number of objects 2.

APPENDIX S10. Pruned tree: confidence factor 0.4, minimum number of objects 5.

APPENDIX S11. Pruned tree: confidence factor 0.4, minimum number of objects 8.

APPENDIX S12. Pruned tree: confidence factor 0.5, minimum number of objects 2.

APPENDIX S13. Pruned tree: confidence factor 0.5, minimum number of objects 5.

APPENDIX S14. Pruned tree: confidence factor 0.5, minimum number of objects 8.

APPENDIX S15. Pruned tree: confidence factor 0.6, minimum number of objects 2.

APPENDIX S16. Pruned tree: confidence factor 0.6, minimum number of objects 5.

APPENDIX S17. Pruned tree: confidence factor 0.25, minimum number of objects 8.

APPENDIX S18. Pruned tree: confidence factor 0.6, minimum number of objects 8.

APPENDIX S19. Pruned tree: confidence factor 0.25, minimum number of objects 5.

APPENDIX S20. Pruned tree: confidence factor 0.25, minimum number of objects 2.

APPENDIX S21. Unpruned tree.

APPENDIX S22. Trait descriptions and possible categories.

LITERATURE CITED

Acevedo, M. A., C. J. Corrada-Bravo, H. Corrada-Bravo, L. J. Villanueva-Rivera, and T. M. Aide. 2009. Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics* 4: 206–214.

Austen, G. E., M. Bindeemann, R. A. Griffiths, and D. L. Roberts. 2016. Species identification by experts and non-experts: Comparing images from field guides. *Scientific Reports* 6: 33634.

Casaloi, M., R. Mantovani, F. P. Scorzonì, S. Puca, A. Speranza, and B. Tirozzi. 2003. Linear and nonlinear post-processing of numerically forecasted surface temperature. *Nonlinear Processes in Geophysics* 10: 373–383.

Chaudhary, V. B., M. A. Rúa, A. Antoninka, J. D. Bever, J. Cannon, A. Craig, and M. Ha. 2016. MycoDB, a global database of plant response to mycorrhizal fungi. *Scientific Data* 3: 160028.

Edwards, M., D. R. Morse, and A. H. Fielding. 1987. Expert systems: frames, rules or logic for species identification? *Bioinformatics* 3: 1–7.

Eibe, F., M. A. Hall, and I. H. Witten 2016. The WEKA Workbench. In I. H. Witten and E. Frank [eds.], Data mining: Practical machine learning tools and techniques, 4th ed. Morgan Kaufmann Publishers, Burlington, Massachusetts, USA.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R. J. Hijmans, et al. 2006. Novel methods improve prediction of species’ distributions from occurrence data. *Ecography* 29(2): 129–151.

Fan, J., N. Zhou, J. Peng, and L. Gao. 2015. Hierarchical learning of tree classifiers for large-scale plant species identification. *IEEE Transactions on Image Processing* 24: 4172–4184.

Gallo, T., and D. Waitt. 2011. Creating a successful citizen science model to detect and report invasive species. *BioScience* 61: 459–465.

Garnier, E., U. Stahl, M. A. Laporte, J. Kattge, I. Mougenot, I. Kühn, B. Laporte, et al. 2017. Towards a thesaurus of plant characteristics: An ecological contribution. *Journal of Ecology* 105: 298–309.

Gaylard, A., and G. I. H. Kerley. 1995. The use of interactive identification keys in ecological studies. *South African Journal of Wildlife Research* 25: 35–40.

Hortal, J., J. M. Lobo, and A. Jimenez-Valverde. 2007. Limitations of biodiversity databases: Case study on seed-plant diversity in Tenerife, Canary Islands. *Conservation Biology* 21: 853–863.

Kattge, J., S. Diaz, S. Lavorel, I. C. Prentice, P. Leadley, G. Bönisch, E. Garnier, et al. 2011. TRY—a global database of plant traits. *Global Change Biology* 17: 2905–2935.

Kubat, M. 2017. Introduction to machine learning, 2nd ed. Springer Publishing, New York, New York, USA.

Kumar, N., P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. C. Lopez, and J. V. Soares. 2012. Leafsnap: A computer vision system for automatic plant species identification. *Computer Vision — ECCV 2012* 7573: 502–516.

Lee, S. H., C. S. Chan, P. Wilkin, and P. Remagnino 2015. Deep-plant: Plant identification with convolutional neural networks. In *IEEE International Conference on Image Processing*, 452–456. IEEE, New York, New York, USA.

Lee, S. H., C. S. Chan, S. J. Mayo, and P. Remagnino. 2017. How deep learning extracts and learns leaf features for plant classification. *Pattern Recognition* 71: 1–13.

Ott, T., C. Palm, R. Vogt, and C. Oberprieler. 2020. GinnJinn: An object-detection pipeline for automated feature extraction from herbarium specimens. *Applications in Plant Sciences* 8(6): e11351.

Quinlan, J. R. 1993. C4. 5: Programs for machine learning. Morgan Kaufmann Publishers, San Francisco, California, USA.

Randler, C. 2008. Teaching species identification—A prerequisite for learning biodiversity and understanding ecology. *Eurasia Journal of Mathematics, Science and Technology Education* 4(3): 223–231.

Roy, H. E., E. Baxter, A. Saunders, and M. J. Pocock. 2016. Focal plant observations as a standardised method for pollinator monitoring: Opportunities and limitations for mass participation citizen science. *PLoS ONE* 11(3): e0150794.

Singh, A. K., B. Ganapathysubramanian, S. Sarkar, and A. Singh. 2018. Deep learning for plant stress phenotyping: Trends and future perspectives. *Trends in Plant Science* 23: 883–898.

Stajich, J. E., T. Harris, B. P. Brunk, J. Brestelli, S. Fischer, O. S. Harb, J. C. Kissinger, et al. 2012. FungiDB: An integrated functional genomics database for fungi. *Nucleic Acids Research* 40: D675–D681.
Thessen, A. 2016. Adoption of machine learning techniques in ecology and earth science. *One Ecosystem* 1: e8621.

Tilling, S. 1984. Keys to biological identification: their role and construction. *Journal of Biological Education* 4: 293–304.

Vayssières, M. P., R. E. Plant, and B. H. Allen-Diaz. 2000. Classification trees: An alternative non-parametric approach for predicting species distributions. *Journal of Vegetation Science* 11: 679–694.

Wäldchen, J., M. Rzanny, M. Seeland, and P. Mader. 2018. Automated plant species identification—Trends and future directions. *PLoS Computational Biology* 14(4): e1005993.

Weaver, W. N., J. Ng, and R. G. Laport. 2020. LeafMachine: Using machine learning to automate leaf trait extraction from digitized herbarium specimens. *Applications in Plant Sciences* 8(6): e11367.

Will, K. W., and D. Rubinoff. 2004. Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics* 20: 47–55.

Wu, X., X. Zhu, G. Q. Wu, and W. Ding. 2013. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering* 26: 97–107.

Zanne, A., K. Abarenkov, M. E. Afkhami, C. A. Aguilar-Trigueros, S. Bates, J. M. Bhatnagar, P. E. Busby, et al. 2020. Fungal functional ecology: Bringing a trait-based approach to plant-associated fungi. *Biological Reviews* 95(2): 409–433.

Zhang, Z. 2016. Missing data imputation: Focusing on single imputation. *Annals of Translational Medicine* 4(1): 9.

Zhou, Z.-H., Y. Jiang, Y.-B. Yang, and S.-F. Chen. 2002. Lung cancer cell identification based on artificial neural network ensembles. *Artificial Intelligence in Medicine* 24: 25–36.