Leaf mysteries revealed through the computer's eye
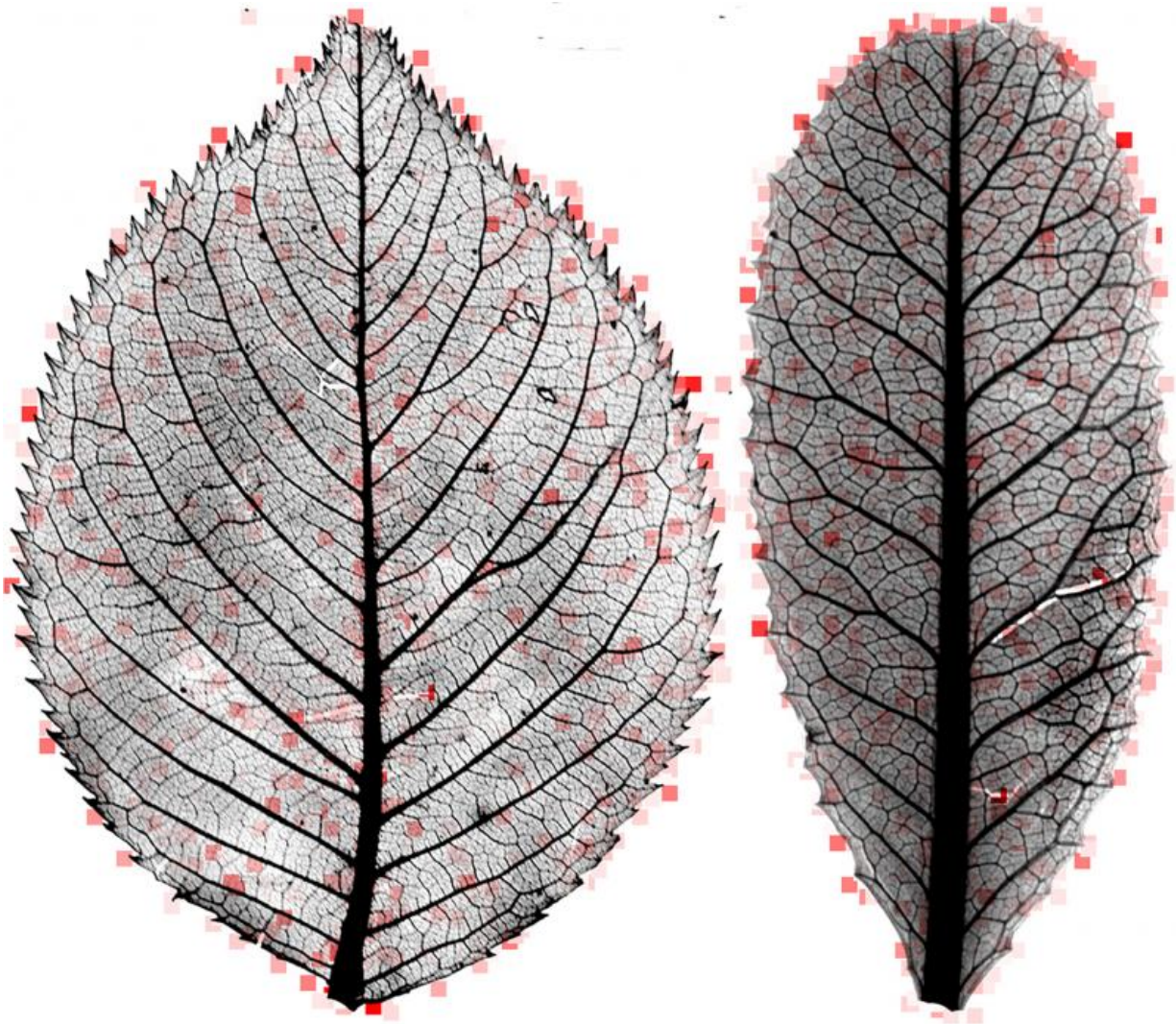
March 7 2016

Image of leaves in the Rosaceae family that is made up of herbs, shrubs and trees including rose, peach, strawberry, plums, cherries, apricots and others. These

cleared leaves have "heat mapping" by the computer vision system that shows leaf attributes important to characterization and classification of the leaves into the Rosaceae family. Depth of color indicates importance of attributes. The heat map shows that many locations on the leaf edges are important as are many large vein intersections. Credit: Shengping Zhang

A computer program that learns and can categorize leaves into large evolutionary categories such as plant families will lead to greatly improved fossil identification and a better understanding of flowering plant evolution, according to an international team of researchers.

"Paleobotanists have collected many millions of fossil leaves and placed them in the world's museums," said Peter Wilf, professor of geosciences, Penn State. "They represent one of the most underused resources for understanding plant evolution. Variation in leaf shape and venation, whether living or fossil, is far too complex for conventional botanical terminology to capture. Computers, on the other hand, have no such limitation."

When botanists identify modern plants, they look at the leaves, but rely mostly on the associated fruits, seeds and flowers to categorize the specimens. In fossil collections, fruits, seeds and flowers are usually much less common than leaves. Even with modern leaves it is a slow process figuring out which features are botanically informative. If a computer vision approach works on modern leaves, it could help in the classification of fossil leaves as well.

"Leaf characterization builds on an 1800's system of description that we call leaf architecture," said Wilf. "It looks at leaf teeth, margins, lobes, and venation patterns and uses specialized terminology to describe them. For the most part, this procedure tells us how to describe a leaf, not how

to identify one and place it on the tree of life. Cracking the leaf code and accessing the evolutionary information in leaf architecture is the central problem I feel I must try to solve in my career as a paleobotanist."
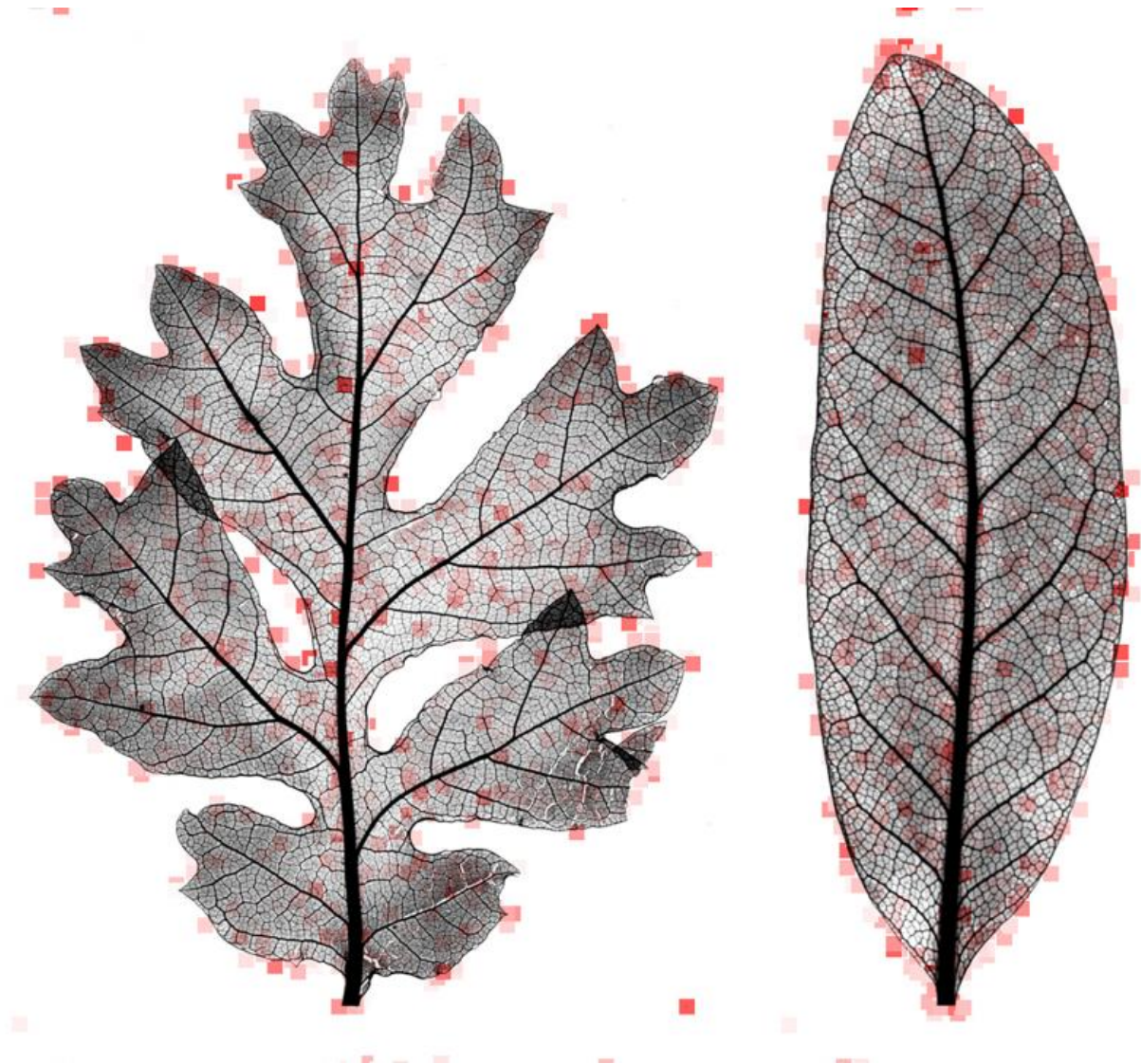

Image of leaves in the Fagaceae family that includes beech, oak, chestnut and other trees. These cleared leaves have "heat mapping" by the computer vision system that shows leaf attributes important to characterization and classification of the leaves into the Fagaceae family. Depth of color indicates importance of attributes. The heat map shows that many locations on the leaf edges are

important as are some vein intersections. Credit: Shengping Zhang

About nine years ago, Wilf learned of an article in the *Proceedings of the National Academy of Sciences* on a computer vision program that could determine whether or not an animal was in a photograph.

"A bell rang in my head," said Wilf. "Instead of an animal, tell me if the image is of an oak leaf or not, or pick among several categories."

He contacted Thomas Serre, now Manning Assistant Professor of Cognitive, Linguistic and psychological science, Brown University, who, as a graduate student at the Massachusetts Institute of Technology, was lead author of that work. The method worked well right off the bat, and after nine years of development and experiments using different vision algorithms, the team published their first paper from this work today (Mar. 7), also in the *Proceedings of the National Academy of Sciences*.

More than two of those years were required for a Penn State undergraduate team to vet and prepare the final dataset of more than 7500 images of cleared leaves, which are specimens that have been chemically bleached, stained and mounted on slides to reveal venation patterns. The largest collection they used is in the Smithsonian Institution's National Museum of Natural History.
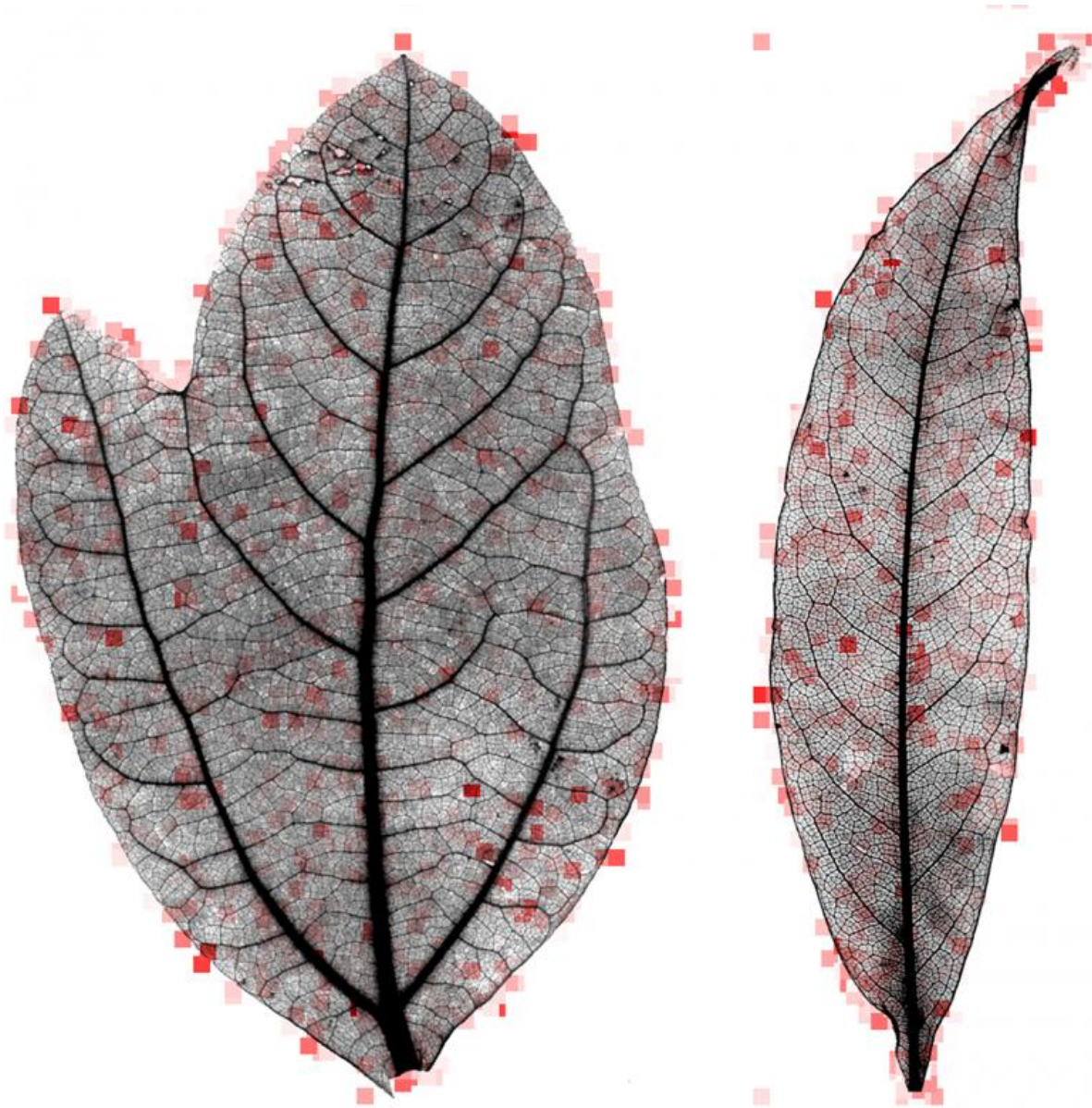
Image of leaves in the Lauraceae family that includes bay laurel, camphor, various cinnamons, avocado and others. These cleared leaves have "heat mapping" by the computer vision system that shows leaf attributes important to characterization and classification of the leaves into the Lauraceae family. Depth of color indicates importance of attributes. The heat map shows that many locations on the leaf edges are important as are many vein intersections. Credit: Shengping Zhang

"The success of our computer vision approach suggests that this may be one of those tasks that are comparatively easier for computers because of computers' ability to process and analyze large numbers of specimens, to discover novel visual features that may have phylogenetic significance," said Serre.

The researchers currently have a 72 percent accuracy rate over 19 leaf families compared to about 5 percent for random chance. This project is not the first to computerize leaf identification. A popular app, Leafsnap: An Electronic Field Guide, matches the shape of an unknown leaf from a particular region and identifies it down to the species level. However, this current work is the first to analyze cleared leaves or [leaf venation](#) for thousands of species from around the world, to learn the traits of evolutionary groups above the species level such as plant families, or to directly visualize informative new characteristics. The variation among the hundreds to thousands of species in a family is many times greater than within a species, and yet, the computer algorithms could learn a set of features and apply it successfully. Because nearly all leaf fossils are of extinct species, family-level identification is usually the first target for paleobotanists.

"This approach is a key distinction between what we call image processing, where literally a computer expert programs a computer to see, as opposed to machine learning and computer vision, where the machine is not programmed to exhibit a particular behavior but rather it learns from examples," said Serre. "Here, our examples were leaf images together with category labels corresponding to family and order."
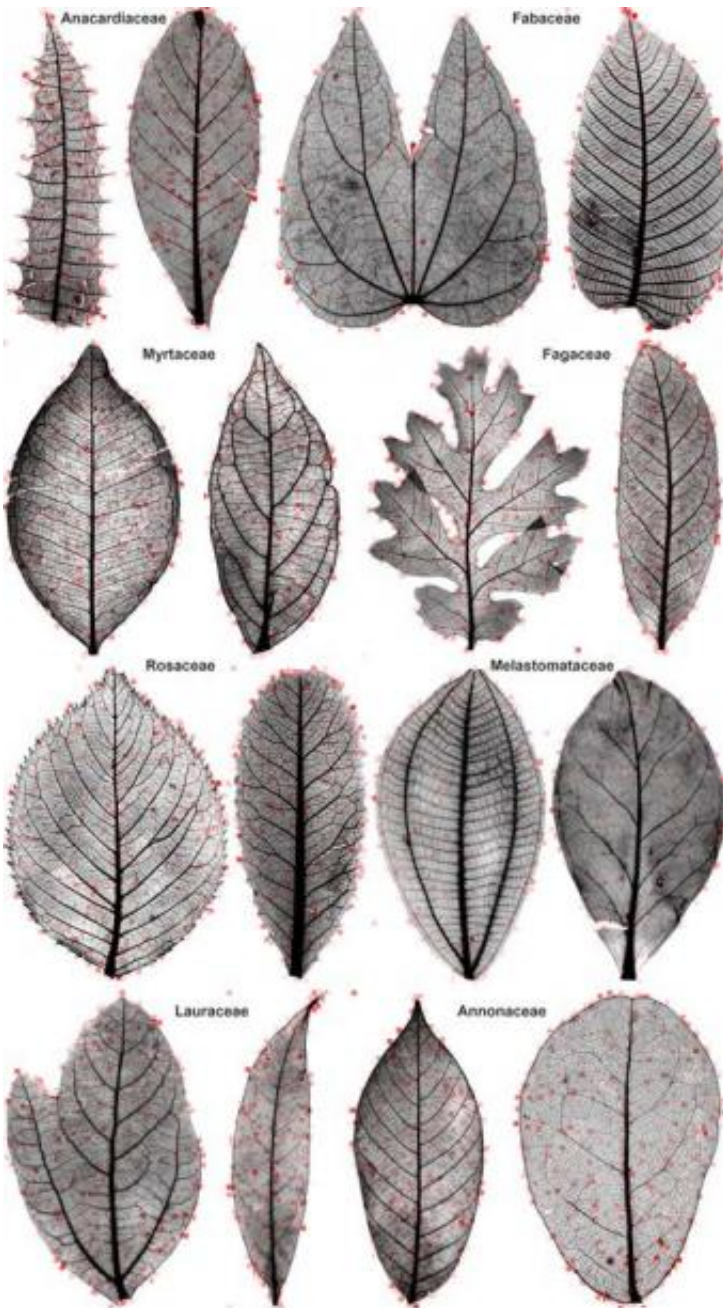
Software learned to classify plants into their proper families and orders, by coming to understand key features of their shape and vein structure. Red spots highlight areas the computer identified as key. Credit: Peter Wilf/Thomas Serre

The researchers provide the computer program with half the photos already identified so that it can automatically learn a dictionary of

special features such as vein intersections and tiny bumps and asymmetries that turn out to matter quite a bit in identifying leaves. The system also learns to disregard the typical problems of low image quality, insect bites and mounting defects. Then the algorithm receives unlabeled test photos and uses its dictionary to identify them. The researchers repeated this procedure 10 times, randomly choosing the training and test images. The results agreed with only 1 percent difference between the runs.

"It normally takes a trained person a few hours to describe one leaf according to the standard protocol, which uses about fifty terms, " said Wilf. "The computer program is thousands of times faster, automatically generates a dictionary of more than 1,000 elements and then actually shows us what parts of the leaf are diagnostic."

Instead of producing only a black box of results, the computer generates a "heat" map directly on the leaf image, identifying and rating areas of importance for correct identification. This approach generates a flood of previously hidden botanical information.

Wilf notes that leaf teeth in the rose family have always been considered distinctive, but the heat maps highlight previously unknown features of their tips. Leaves of the coffee family, with 13,000 living species, are very hard to identify when not attached to twigs, but the computer program found it one of the least problematic at 90 percent accuracy.

The ability of computer vision to classify leaves quickly and to generate vast quantities of new botanical knowledge will allow scientists to develop more accurate evolutionary pedigrees for plants and plant fossils.

 More information: Computer vision cracks the leaf code, *PNAS*, www.pnas.org/cgi/doi/10.1073/pnas.1524473113