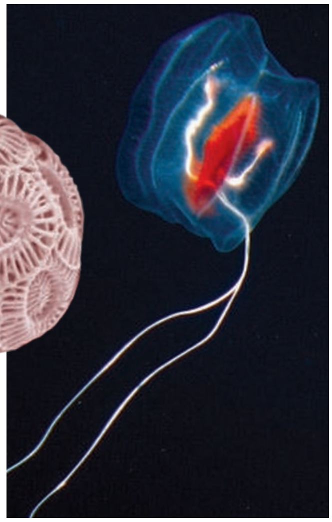
OPINION



LEFT TO RIGHT: NAT. HIST. MUS., LONDON; M. BENFIELD/GULF SERPENT PROJECT; N. CATTLIN/FLPA

Time to automate identification

Taxonomists should work with specialists in pattern recognition, machine learning and artificial intelligence, say **Norman MacLeod**, **Mark Benfield** and **Phil Culverhouse** — more accuracy and less drudgery will result.

An imaging system designed to identify marine zooplankton was recently adopted by scientists working for the US government to monitor the Deepwater Horizon oil spill. By measuring the size of oil droplets produced after chemical dispersants had broken up the oil, modellers could predict the depths at which the plume was accumulating. Only two instruments exist that can measure oil droplets while distinguishing them from other matter suspended in the water column, such as zooplankton, marine snow and gas bubbles at depths of down to 1,500 metres. The deployment, by the US National Oceanic and Atmospheric Administration, of one — the digital holographic imaging (DHI) system, developed jointly by the Massachusetts Institute of Technology in Cambridge and the Woods Hole Oceanographic Institution in Massachusetts — is a working example of something that should be happening on a grand scale: the shared use across diverse disciplines of generalized automated identification technologies.

Taxonomists who identify, describe and name species (who practise alpha taxonomy, as it is known in the trade) are central to many research programmes in applied biology, ecology and conservation. University cuts are shrinking this already small community. What's more, there is no tradition of — much less a requirement for — independent testing and verification of the accuracy of the identifications that taxonomists produce, unlike in virtually every other scientific discipline.

Automating species identification using technologies developed by researchers in pattern recognition, artificial intelligence and machine learning would transform alpha taxonomy from a cottage industry dependent on the expertise of a few individuals to a testable and verifiable science accessible to anyone needing to recognize objects. Indeed, a concerted interdisciplinary research and development effort, within the next decade, could lead to automated systems capable of high-throughput identifications for hundreds or thousands of categories of living as well as non-living specimens.

Human error

Many taxonomists use sophisticated technologies to capture images, sounds, even the smells and tastes of biological specimens. But most routine identifications involve a small group of experts scattered around the world assessing diagnostic data qualitatively — commonly the size, shape or texture of specimens, or the presence or absence of certain features. Surprisingly few blind-test studies have been published to assess the accuracy of taxonomists' findings objectively[1–7]. Those that have been carried out are worrying. For instance, a blind test to resolve a controversy about the pattern of extinction in one marine animal group about 65.5 million years ago[1] resulted in species lists that were so different as to make consensus impossible. Such inconsistencies shouldn't be a surprise given that, in controlled visual-cognition studies, humans frequently miss items presented in a scene, count some objects more than once and misclassify others.

"Humans miss items, count some objects more than once and misclassify others."

Hopes are high among researchers and funding bodies that DNA bar-coding, by which a species is recognized according to a marker in its mitochondrial genome, will increase the accuracy of identifications — and ease bottlenecks resulting from a shortage of trained and experienced taxonomists. But bar codes are generally used to assign organisms to taxonomic categories that have already been defined on the basis of morphological traits. In other words, a bar code isn't useful until the reference species has already been identified by multiple experts. The technique is still relatively expensive, slow and difficult to implement in the field except in certain situations — for example in laboratories on oceanographic research vessels. Moreover, researchers frequently need to identify non-living objects as well as living ones. Ecologists studying plankton, for example, commonly count 'fibres', 'detritus' or 'egg-like particles' that may or may not be alive.

In focusing on bar-coding, stakeholders have overlooked the greater promise of machine-learning to transform taxonomy and the identification of natural objects in general.

DAISY, DAISY, give me your answer, do

Computer systems now exist for classifying objects into between 2 and 30 categories. These systems already deliver faster, more accurate and more consistent semi- or fully automated identifications than any human taxonomist. For instance, a group of entomologists at the Natural History Museum in London have used the Digital Automated Identification System (DAISY) to identify with 100% accuracy 15 species of parasitic wasp from digital images of wings, with each identification taking less than a second[8]. Similarly, oceanographers from the University of Plymouth, UK, have used the Dinoflagellate Categorisation by Artificial Neural Network (DiCANN) system to identify phytoplankton species with about 72% accuracy — the same as experts[7]. People who make a particular assessment routinely, such as counting the dorsal spines on sticklebacks, can return accuracies of 84–95% under test conditions. However, trained personnel often need to deliver one-off species identifications. In tests based on this scenario, people make choices consistent with their own previous selections only 67–83% of the time, and consistency across different identifiers can be as low as 43%[9].

In practice, most current research programmes will require a considerable scale up of DAISY's and DiCANN's capacities. Biodiversity assessment teams, biostratigraphers guiding drilling operations and zooplankton ecologists on oceanographic research vessels all need tools capable of identifying hundreds or even thousands of different species. Automated systems for this level of diversity are not currently available, but neither are they the stuff of science fiction as many researchers imagine.

As a first step, taxonomists should team up with specialists working in pattern recognition, machine learning and artificial intelligence — as well as technology engineers, software designers and mathematicians. (A good example of such a team is the Scientific Committee on Oceanic Research's Working Group 130; www.scor-wg130.net.) Artificial-intelligence researchers are using information from experimental psychology and computational neuroscience to design computer systems that recognize human faces or household objects based on their visual properties. Engineers are developing systems to acquire high-resolution, *in situ*, 'hyperspectral' images of organisms in three dimensions, by sensing visible, ultraviolet and infrared light. Machine learning and pattern-recognition scientists are developing algorithms that incorporate various characteristics of objects — many of which may not be detectable by humans — into identification programs. For example, bat echolocation calls that are outside the range of the human auditory system. Finally, software designers are improving the user interfaces of classification programs.

Currently, grant applications for such interdisciplinary projects are falling between the boundaries defined by funding bodies in engineering and the life sciences. Funding specifically for collaborations on automated species identification should be supplied by the European Union's framework programmes (Europe's main instrument for funding research), and national research councils such as the US National Science Foundation and the UK Natural Environment Research Council. Charitable organizations, such as the Wellcome Trust in London, and the Alfred P. Sloan Foundation in New York, should follow suit.

NAT. HIST. MUS., LONDON

Automated systems such as DAISY (screenshot, bottom) can be more reliable than experts when it comes to identifying specimens.

We estimate that a modest level of investment in these kinds of projects — roughly US$1 million to $2 million per annum over 5–10 years — would encourage the development of large-scale automated identification systems, at least for some high-profile groups across the taxonomic spectrum (for protists, plants and animals). In addition, a public competition, in which researchers are asked to develop an automated system capable of identifying species within a standard set of complex but generalized scenes, such as in photographs of a coral reef, would attract public interest and encourage diverse groups of scientists to explore the technologies available. This would be similar to the 'visual object classes' challenge recently set up by the EU-funded Network of Excellence on Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL), which promotes the development of computer systems that recognize types of common objects.

This investment would pay huge dividends across a range of disciplines. In the past 50 years, academic centres worldwide have cut back or discontinued many taxonomic training and research programmes. As a result, there are only about 4,000–6,000 professional taxonomists worldwide[10], only a subset of whom are routinely engaged in species identifications. Meanwhile, the demand for identifying natural objects has escalated. Agriculturalists and border security staff, for example, increasingly need to identify potential pest species where and when they encounter them. Even areas of research not directly concerned with species identification stand to benefit — as the Deepwater Horizon oil droplets example demonstrates.

Far from making alpha taxonomists obsolete, automated identification systems would free them from the drudgery of routine identifications. This would allow them to focus on the more conceptually difficult issues of discovering, describing and revising species concepts, and establishing how species function within natural systems and fit into higher taxonomic and ecological groups. ■

Norman MacLeod is in the Palaeontology Department of the Natural History Museum, Cromwell Road, London SW7 5BD, UK. **Mark Benfield** is in the Department of Oceanography and Coastal Sciences, School of the Coast and Environment, Louisiana State University, Baton Rouge, Louisiana 70803, USA. **Phil Culverhouse** is at the Centre for Robotics and Neural Systems, University of Plymouth, Plymouth PL4 8AA. UK.
e-mails: n.macleod@nhm.ac.uk; mbenfie@lsu.edu; pculverhouse@plymouth.ac.uk

1. Colquhoun, W. P. *Ergonomics* **2,** 367–372 (1959).
2. Zachariasse, W. J. *et al. Utrecht Micropaleontological Bulletins* **17,** 1–265 (1978).
3. Simpson, R., Culverhouse, P. F., Ellis, R. & Williams, R. *IEEE Conf. Neural Networks Ocean Eng.* 223–230 (IEEE, 1991).
4. Ginsburg, R. N. *Mar. Micropaleontol.* **29,** 67–68 (1997).
5. Kelly, M. G. *Water Res.* **35,** 2784–2788 (2001).
6. Gobalet, K. W. *J. Archaeol. Sci.* **28,** 377–386 (2001).
7. Culverhouse, P. F., Williams, R., Reguera, B., Herry, V. & González-Gil, S. *Mar. Ecol. Prog. Ser.* **247,** 17–25 (2003).
8. Gauld, I. D., O'Neill, M. A. & Gaston, K. J. in *Hymenoptera: Evolution, Biodiversity and Biological Control* (eds Austin, A. D. & Dowton, M.) 303–312 (CSIRO, 2000).
9. Culverhouse, P. F. in *Automated Taxon Identification in Systematics: Theory, Approaches and Applications* (ed. MacLeod, N.) 25 (CRC Press, 2007).
10. The Global Taxonomy Initiative; go.nature.com/fhA3T2