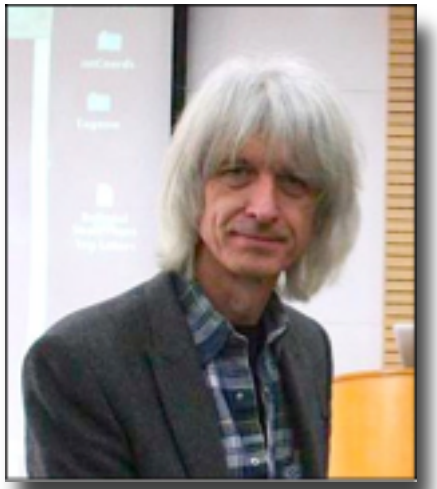
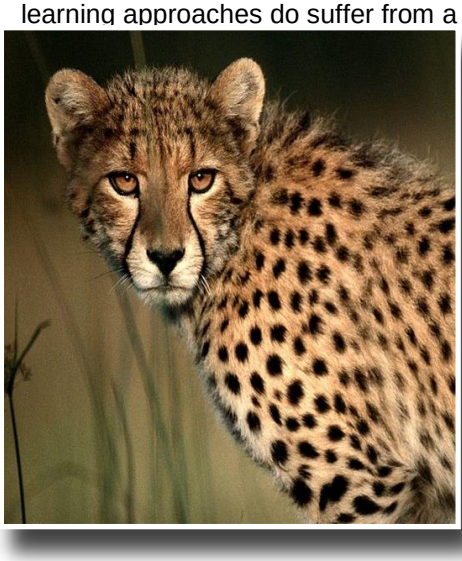
Comparison of Morphometric and Machine-Learning Approaches to Automated Taxon Identification (With Examples from the Vertebrates)



Norman MacLeod
The Natural History Museum, Cromwell Road, London, UK; Department of Earth Sciences, University College London, Gower Street, London, UK; Nanjing Institute of Geology & Palaeontology, Chinese Academy of Sciences, 39 Beijing, Donglu, Nanjing, China; Faculty of Life Sciences, The University of Manchester, Carys Bannister Building, Dover Street, Manchester.

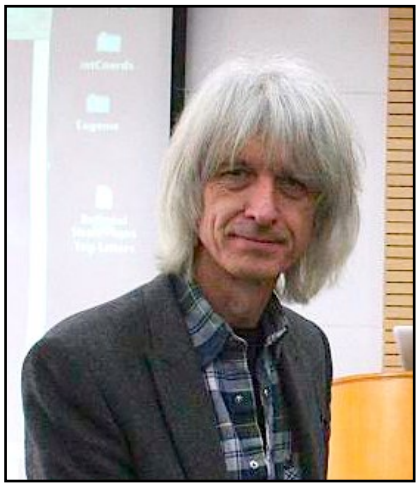
One approach to addressing long-standing concerns associated with the taxonomic impediment and the low reproducibility of taxonomic data is through development of automated species identification systems. Two generalized approach categories are considered relevant in this context: morphometric systems based on measurements taken from 2D images or 3D scans and analyzed by some form of discriminant analysis and machine learning systems (e.g., artificial neural networks, self-organizing maps, support-vector machines, random forest classifiers, deep learning) that analyze the pixel brightness values of digital images. The former category is generally familiar to many systematists, but has rarely been used for taxonomic group-identification. The latter is less familiar, but is employed increasingly in various sorts of mathematical research, information technology, and security-related contexts. Use of either category to augment the performance of human experts is highly desirable in order to (1) raise the quality of taxonomic identifications on which so many scientific results and interpretations depend, (2) stabilize species concepts, and (3) deliver high-quality taxonomic identifications to those who need them in academic, educational, industrial, agricultural, resource management/conservation, government, and cultural (museum) sectors of the economy. Comparisons between these two approaches are needed in order to establish appropriate roles for each and to identify the limitations of each for resolving taxonomic problems in all spheres of human activity.

To date I have direct experience with the application of these procedures to pollen, foraminifera, insects, centipedes, birds, mammals, hominins and plant leaves, encompassing modern, fossil, and modern + fossil data sets.[1] Results obtained demonstrate that both approaches are capable of delivering identifications that are over 90 percent accurate for small datasets. The performance of 2D Euclidean distance and landmark-based linear discriminant analysis systems is enhanced substantially through application of generalized least-squares superposition methods that normalize form data for variations in size, position, and orientation. However, these morphometric procedures are practically limited to the detailed analysis of small numbers of groups and small samples sizes by a variety of factors, including the complexity of the morphologies under investigation and low data-collection rates. In many cases performance of these algorithms also improves dramatically when 3D data are used as the basis for morphological comparisons. Machine-learning alternatives usually deliver better performance, are much less labor-intensive to implement, and are capable (at least in principle) of being scaled up to incorporate 100s, and even 1,000s of different groups. Machine learning approaches do suffer from a lack of post hoc interpretability with regard to specification of the relative weightings of the taxonomic characters used to achieve the identification. However, this is not an important consideration if delivery of high-quality identifications is the primary goal.



Both approaches classes should be considered valid within their own analytic domains and both can deliver identification speeds, consistencies, and reproducibilities that far outstrip the abilities of human experts to make such identifications. While additional research is needed to validate the results obtained in small-scale trials and to test the resultant discriminant functions for stability under 'real-world' conditions, the fact that this general approach to biological group identification has now been demonstrated to work in a variety of taxonomic and disciplinary contexts suggests we may be on the threshold of realizing an order-of-magnitude improvement in the scope, speed, and accuracy of taxonomic identifications using morphological data.

[1] Vertebrate datasets will serve as examples for this presentation.



Professor Norman MacLeod

N.MacLeod@nhm.ac.uk

The Natural History Museum (London)

Norman MacLeod (BSc, MSc, PhD, FGS, FLS) is Dean of Post-Graduate Education and Training at the Natural History Museum (London). Honorary Professor at University College London and Visiting Professor at the Nanjing Institute of Geology and Palaeontology, Chinese Academy of Sciences. He is the editor of *Automated Taxon Identification in Systematics: Theory, Approaches, and Applications* (CRC Press, Taylor & Francis Group, 2007, http://www.crcpress.com/product/isbn/9780849382055), the author of *The Great Extinctions: What Causes Them and How They Shape Life* (Natural History Museum, 2013, http://www.nhm.ac.uk/business-centre/publishing/books/earth/great-extinctions/index.html), the Editor-in-Chief of *Grzimek's Animal Life Encyclopedia: Extinctions*, 2nd ed., 2 vols. (Gale-Cengage, 2013, http://www.cengage.com/search/productOverview.do;jsessionid=4EBB42963ECF0CFB547FCB0A490E4003?N=197&Ntk=P_EPI&Ntt=48554724514685656651605721223196204384 2&Ntx=mode+matchallpartial#Overview) which received an Honorable Mention Professional and Scholarly Excellence (PROSE) Award: Multivolume Reference/Science category in 2014 by the Association of American Publishers (see http://www.proseawards.com/current-winners.html) and the co-editor of *Issues in Palaeobiology: A Global View—Interviews and Essays* (with Marcelo R. Sánchez-Villagra; Scidinge Hall, 2014, http://www.amazon.com/Issues-Palaeobiology-Global-Interviews-Essays/dp/3905923173). He is currently writing a book on the mathematical analysis of morphology. Professor MacLeod also serves as Co-Chief Editor of *Palaeoworld* (a Chinese palaeontology journal), Associate Editor of the journal *Systematic Biology*, and an Editorial Board Member of the journal *Royal Society Proceedings B (Biological Sciences)*.

Professor MacLeod has a wide range of research interests. He is perhaps best known for his work documenting patterns and understanding the causes of Phanerozoic extinctions, especially the end-Cretaceous mass extinction event where he is a leading proponent of the multiple-cause model. Equal in terms of output and prominence is his theoretical, methodological, and applied work in the field of morphometrics where he was an early proponent of geometric morphometrics, the use of outline semilandmarks to characterize form and shape, the morphometric characterization of 3D surfaces, and most recently the application of computer vision and machine learning methods to the analysis of morphology. Other research interests include macroevolution, evolutionary rates, quantitative biostratigraphy (esp. graphic correlation), applied statistics, and quantitative data analysis (esp. multivariate ordination, discriminant analysis, Monte Carlo simulation, bootstrapping, and jackknifing). Most recently Prof. MacLeod has developed minor intellectual sidelines in art history picture analysis (esp. with regard to scientific images) and the effect of human evolutionary biology and climate change on human history.