A&A 500, 491–492 (2009)
DOI: 10.1051/0004-6361/200912200

Astronomy & Astrophysics
Special issue

Commentary on: Bertin E. and Arnouts S., 1996, A&AS, 117, 393

Modern astronomical surveys and the development of informatics

S. N. Shore

Dipartimento di Fisica "Enrico Fermi", Università di Pisa and INFN-Sezione di Pisa, largo B. Pontecorvo 3, Pisa 56127, Italy
e-mail: shore@df.unipi.it

The descriptive cataloging of the natural world is one of the oldest intellectual exercises. Begun for use in natural magic, botanicals, and astrological prognostication in virtually all civilizations, it gradually changed into a taxonomic scheme for distinguishing the components of Nature. In biology this was easier since there is a relatively limited set of criteria that can be used to separate species, but the concept of hierarchies was an innovation of the late Renaissance that is best represented in the systems of Linnaeus and Cuvier, mainly in the 18th century (see Foucault 1973). It was the basis of the other achievement celebrated in this year, 2009: the publication by Charles Darwin of his *Origin of Species* in 1859. In astrophysics, morphological classification was exploited by Messier (as a very broad criterion for distinguishing comets from fixed, nebulous objects) and the heroic visual cataloging efforts of William, Caroline, and John Herschel from the end of the 18th through the first third of the 19th centuries. Photometric and astrometric catalogs, such as the BD and CD, contained no discriminating criteria other than that the object was *not* nebulous, and the catalogs were generally of that sort. Instead, taxonomy began to play a fundamental role when new instruments were introduced, particularly spectrographs and photometers.

The basic principle of taxonomy, in any form, is to find a limited set of attributes that suffice to distinguish individual samples within a hierarchy. In biology, this can be qualitative, for example when it is based on a limited set of characteristics that are judged by a human classifier on the basis of some "similarity" scale, or quantitative, using *phylogenetics* (also known as *cladistics*). In this scheme, the classification is based on features that are chosen in advance and may be continuously graded according to some numerical metric (see e.g. Sokal & Sneath 1963). When faced with a completely new and unexplored class of objects, the difficult goal is to find a proper description. Taxonomy can, however, be made precise. A finite set of attributes is chosen that serves as the basic distinguishing features of any sample, and these can be assigned numerically, but these criteria can (and frequently do) evolve with time as experience accumulates. Of course we use such techniques all the time in astronomy for relatively small samples, but it is fatiguing if the sample is large and may not be reliable when based on personal evaluation. Spectroscopic classification is a paradigmatic case. A relatively dense set of standards is required unless the system is very coarse (the difference, for instance, between the Secchi's four – later five – subclasses and the Harvard system, as described by Maury in the fundamental papers of classification at moderate spectral resolution, Cannon in the preface to the Henry Draper catalog for low resolution, and later Morgan and Keenan in their preliminary remarks to the MKK atlas). We know that spectra are not like flowers, they are part of a (at least two-dimensional) continuum, and the discretization of the scheme is artificial. But if dense enough, the classes can capture the physics and provide sufficient discriminating information for the individual objects (the taxa) for further analysis; the simplest groupings for spectra, for instance, were based on either the complexity (qualitative) of the spectra or some set of measures such as the approximate ratios of specific line intensities and widths. Any scheme must be self-consistent and not rely on information that does not come from the data itself. It can, however, be "informed" by other criteria, for instance photometry. In a Bayesian sense, this means the information must be from the data itself, although the establishment of the taxa may have made use of other information to form the set of priors.

The first all-sky survey designed to systematically discover and describe nebulous objects was William Herschel's series of catalogs, published between 1786 and 1802. These were visual studies, limited by the ability of the observer and quite general in the descriptions. One of the first attempts at morphological quantification, by Wolf in 1908, was extended and refined for nebular objects in Hubble's Ph.D. Thesis in 1917. Once they were finally identified as extragalactic, this served as the basis for the minimal set of criteria forming the Hubble classification and its later extensions (see Sandage 2005), an example of what I meant by the classification being modified by information external to the sample that then can be discarded when the scheme is actually applied. If you have the additional information that something is physically different, a galaxy or a Galactic object, then you can build this into the criteria. Zwicky (1948) seems to have been the first to summarize the *general* problem of morphology in the general astronomical context.

In the period after the second world war, a number of surveys were undertaken to extend the same scope of the Cartes du Ciel (from the early 20th century) to the domain of extragalactic objects. The 1950s was particularly fruitful in contacts between statisticians and astronomers (see the volumes of the Berkeley Statistical Conferences, the tradition has been renewed with the continuing *astrostatistics* meetings founded by Eric Feigelson and collaborators). The Shane-Wirtanen survey proceeded in a close collaboration between the UC-Berkeley Statistical Laboratory (e.g. Neyman & Scott 1958) and the Lick Observatory. In particular, the Lick Survey and many of the resulting statistical critiques and analyses were published in the Lick Observatory Bulletin. This, and other large scale structure surveys, were conducted by human classifiers (e.g. the Zwicky catalogs, the Morphological Galactic Catalog, the Jagellonian survey). The development of automated plate scanners, e.g. the PDS microdensitometer and specially designed machines, and advances in computer control and data processing over the next decades finally ushered in a new era, making possible the massive automated APM (e.g. Maddox et al. 1990) and COSMOS (e.g. Beard et al. 1990) surveys. The difficulty faced by large scale surveys is to separate Galactic from extragalactic sources (stellar from merely starlike objects) by some distinguishing features (generally either photometry or spectroscopy) and then, for the separated classes, to perform statistical analyses and select candidates for follow-up studies[1]. These were photographic, the predecessors of the Sloan Digital Sky Survey (SDSS) that includes spectrophotometry and therefore extends the available characteristics for any classifier.

As machine accuracy improved, memory expanded, and processors became faster, progressively larger samples could be handled in a reasonable time frame. Automatic photometric classifiers were introduced early in the PDS era when large surveys could be compiled from multifilter plate scans (e.g. Kron 1980). These were essential for constructing the Guide Star Catalog since the purpose was to have a vast ensemble of possible sources on which HST could be aligned and with which pointing could be maintained (a six-level taxonomy, see Lasker et al. 1990). The surveys capitalized on the accuracy of two-dimensional scanning densitometers and the increasing sensitivity and resolution of photographic emulsions and the availability of multi-epoch all sky coverage. But more significant was the rapid expansion in computer power, with the ability to handle progressively larger samples at ever faster rates. The FOCAS algorithms (Jarvis & Tyson 1981) used a series of essentially geometric parameters to separate taxa. Moving from large surveys to crowded field photometry using CCD detectors, DAOPHOT (Stetson 1987) was developed to extract object photometry using information selected in the field to determine the point spread function, and it permitted impersonal processing of very large samples and provided full statistical information (such as completeness) about the results.

Principal component methods seek to reduce the characteristics of the taxa to a set of relations between the measured quantities and statistically evaluate the combinations that produce a minimal dispersion around some "vector" directions. The technique has been widely used in astronomy (one of the most successful results, for galaxies, being the Fundamental Plane), and it is particularly successful in elucidating inter-relationships in measured quantities that can then be used to separate the contributions from a number of physical mechanisms (such as spectral line profile decompositions). Cluster analysis is similar in taking a set of measurements of observables and combining them in ways that distinguish subgroups. The weighting schemes attempt to capture the expertise of the classifier (see, e.g., the remarkable experiment by Lahav et al. 1995). Static forms are sets of rules that codify a set of criteria, almost like interviews with experts who produce self-consistent classifications. These rules also include, consequently, all of the biases. Artificial neural networks (ANN) are similar in the sense that they use a fixed sample of test objects that have a finite set of attributes that are prejudged as distinct enough to create a dynamical, self-correcting "table" of weights (e.g. Rumhart et al. 1986). These are the so-called training set, which can be real or simulated samples (e.g. real images of stars and galaxies or simulated data based on the statistical properties of real objects). Once these have been processed by the algorithm, the measures of success of the scheme can be quantified and the algorithm can be turned loose to evaluate a sample of unknowns with a reliability that has been determined by the tests. Instead of having a yes or no response, however, because these are attempting to mimic what a human would do, the algorithm produces a set of statistical "maybes" that, together, classify the object with some calibrated uncertainty, the "fuzzy" part of fuzzy logic. An ANN also seeks to duplicate the expertise and experience accumulated by a human observer *during* a survey (see Odenwald 1995; Lahav 1995, for overviews).

The program *Source Extractor* (*SExtractor*) was the latest (at the time) in a line of development of ANN schemes and the first to be tailored to general use and public release. It was also specifically adapted to the data produced by fixed pixelated detectors, CCDs, that do not suffer from some of the inherent problems associated with photographic emulsion[2]. It is particularly important that the pixels are fixed, along with the sensitivity, since the sky is always changing, and seeing dominates the ability of any classifier to properly render the separations of the sample. That accounts for its popularity and the large number of citations. The advent of CCDs made a simpler set of criteria possible, since in this first release, SExtractor did not include the category of "defects". For instance, cosmic ray hits and other single pixel events were excluded by the requirement that the target image be resolved. The flexibility of such algorithms is that they permit the user to keep pace with the improvements in the technology. The popularity of *SExtractor* derived, and still comes from, its particularly efficient implementation and ease of use.

Acknowledgements. My thanks to Claude Bertout and Thierry Forveille for their comments on the manuscript and to Mario Bertero, Tom Bolton, and Tarcisio del Prete for discussions.

References

Beard, S. M., MacGillivray, H. T., & Thanisch, P. F. 1990, MNRAS, 247, 311
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Foucault, M. 1973, The Nature of Things (NY: Vintage)
Irwin, M. J. 1985, MNRAS, 214, 575
Jarvis, J. F., & Tyson, J. A. 1981, AJ, 86, 476
Kron, R. G. 1980, ApJS, 43, 305
Lahav, O. 1995, [arXiv:astro-ph/9505091v1]
Lahav, O., Naim, A., Buta, R. J., et al. 1995, Science, 267, 859
Lasker, B., Sturch, C., McLean, B., et al. 1990, AJ, 99, 2019
Maddox, S. J., Efstathiou, G., Sutherland, W. J., & Loveday, J. 1990, MNRAS, 243, 692
Neyman, J., & Scott, E. L. 1958, J. Roy. Stat. Soc., 20. 1
Odenwald, S. C. 1995, PASP, 107, 770
Rumelhart, D. E., Hinton, E. E., & Williams, R. J. 1986, Nature, 323, 533
Sandage, A. 2005, ARA&A, 43, 581
Sokal, R. R., & Sneath, P. A. H. 1963, Principles of Numerical Taxonomy (San Francisco: W. H. Freeman)
Stetson, P. B. 1987, PASP, 99, 191
Storrie-Lombardi, L. J. 1992, MNRAS, 259, 8P
Zwicky, F. 1948, Observatory, 68, 121 (see also Zwicky, F. 1957, Morphological Astronomy, Berlin: Springer-Verlag)

[1] The observational task is somewhat easier when distinguishing between stars, galaxies, and "artifacts" in an image than in classifying spectra or assigning a Hubble type. At least at some level of resolution, this is truly discrete, given a point spread function you know what a point source is, and a relatively small set of parameters can be used to separate the classes (e.g. Stetson 1987; Irwin 1985). For spectral classification the discretization is artificial, spectral types do not really exist and are a way of quantifying and binning an actually continuous distribution of characteristics in some convenient way. The same is true for morphological subtypes within a single category, as, for instance, the division between classes of spirals and ellipticals. There, however, the system *can* be made quantitative.

[2] It's useful to recall that each photographic plate was unique, depending on a chemical development process that could only be controlled so far. Instead, a CCD detector, while it may have a weird surface sensitivity, has a fixed flat field that can, in principle, be removed. Thus, when building a classifier based on *any* set of criteria – real images or simulations – the training set is reliable in a way that photographic images cannot be.