AI Magazine Volume 17 Number 2 (1996) (© AAAI)

From Digitized Images to Online Catalogs

Data Mining a Sky Survey

Usama M. Fayyad, S. G. Djorgovski, and Nicholas Weir

■ The value of scientific digital-image libraries seldom lies in the pixels of images. For large collections of images, such as those resulting from astronomy sky surveys, the typical useful product is an online database cataloging entries of interest. We focus on the automation of the cataloging effort of a major sky survey and the availability of digital libraries in general.

The SKICAT system automates the reduction and analysis of the three terabytes worth of images, expected to contain on the order of 2 billion sky objects. For the primary scientific analysis of these data, it is necessary to detect, measure, and classify every sky object. SKICAT integrates techniques for image processing, classification learning, database management, and visualization. The learning algorithms are trained to classify the detected objects and can classify objects too faint for visual classification with an accuracy level exceeding 90 percent. This accuracy level increases the number of classified objects in the final catalog threefold relative to the best results from digitized photographic sky surveys to date. Hence, learning algorithms played a powerful and enabling role and solved a difficult, scientifically significant problem, enabling the consistent, accurate classification and the ease of access and analysis of an otherwise unfathomable data set.

In astronomy and space sciences, we currently face a data-glut crisis. The problem of dealing with the huge volume of data accumulated from a variety of sources, correlating the data, and extracting and visualizing the important trends is now fully recognized. This problem will rapidly become more acute because of the advent of new telescopes, detectors, and space missions, with the data flux measured in terabytes. We face a critical need for information-processing technology and methodology with which to manage this data avalanche to produce interesting scientific results quickly and efficiently. The fields of knowledge discovery in databases and data mining (Fayyad et al. 1996) are mainly concerned with the extraction of higher-level knowledge from low-level data. This article presents a data-mining approach based on machine-learning classification methods that represents a good example of how this new generation of automated analysis tools can offer novel and effective solutions to classical problems in the analysis of large data sets in science.

Across a variety of disciplines, two-dimensional digital-image data are now a fundamental component of routine scientific investigation. The proliferation of image-acquisition hardware such as multispectral remote-sensing platforms, medical-imaging sensors, and high-resolution cameras has led to the widespread use of image data in fields such as oceanography, atmospheric studies, planetary science, agriculture, glaciology, forestry, astronomy, and diagnostic medicine. Across all these fields, the pixel image is but a means to an end. The investigator is interested in using the image data to infer some conclusion about the physical properties of the target being imaged. Image data rely on the human visual system's ability to aid in abstraction and recognition.

In the past, both in planetary science and astronomy, images were painstakingly analyzed by human inspection, and much investigative work was carried out using hard-copy photographs or photographic plates. However, the image data sets that are currently being acquired are so large that simple manual cataloging is no longer practical, especially if all the available data are to be used. We focus on one such digital image set that results from

the Second Palomar Observatory Sky Survey (POSS-II). This digital image library defies manual-visual analysis capabilities and illustrates the need for automated cataloging tools to allow users to gain access to its content. Thus, we target the problem of turning a digital-image data set into a true digital library—one that can be queried by content and used for scientific investigation.

In science image libraries, the typical most fundamental operation is that of cataloging content for later retrieval and large-scale statistical analysis. Cataloging and indexing often involve recognition of objects. We use an approach that is based on classification learning algorithms, where the user (astronomer) trains the system to perform classification tasks by providing it with training examples. An example is represented as a vector of features. A *feature* (also called attribute or variable) is a dimension along which some property of the example is measured. Features can be either numeric or continuous (for example, temperature, intensity) or can be categorical, with no ordering on the values (for example, shape that can take the values of circle, triangle, quadrilateral, and so on). The dimensions define a space, called *feature space*.

We believe there is an important and crucial problem that needs to be addressed before collections of digital images can be turned into useful digital libraries, namely, the query-formulation problem.

As an example, suppose the image of an object is represented by 50 × 50 pixels. One choice of feature space is the pixels. In this case, an example would be a feature vector consisting of 2500 numeric values. This low-level representation is often referred to as *pixel space*. Clearly, pixel space is high dimensional and contains many highly correlated dimensions. Hence, it is not a convenient or compact representation of the information contained in the pixels. In problems where the goal is to learn from examples, high dimensionality is a big problem. If a problem has 2500 variables, then an algorithm would need a much higher number of examples to infer anything about the problem. However, if one were able to re-express the problem in a much smaller number of variables (a lower-dimensional feature space) without losing essential information, a dramatically smaller number of examples would be needed to support sufficient statistics for inference and induction.

As every pattern-recognition practitioner knows, two familiar issues are at the heart of the problem of inferring a model out of data: (1) transforming (reducing) the data from pixels to meaningful or useful features and (2) recognizing (classifying) the detected objects in feature space. In our case, rather than requiring the user to design and implement a classification algorithm to achieve the second step, a machine-learning approach can be used to automatically construct the classifier based on training examples provided by the user. Not only does this eliminate the burden of programming for the user, it also provides a mechanism for tackling the often difficult problem of recognizing objects in feature space.

The Query-Formulation Problem

Work on techniques for digital libraries has focused mainly on digitization techniques, storage and retrieval mechanisms, search mechanisms (especially for text), and database issues dealing with efficient indexing and query execution. We believe there is an important and crucial problem that needs to be addressed before collections of digital images can be turned into useful digital libraries, namely, the query-formulation problem. Users would mainly like to be able to use a digital-image library to search for particular targets for cataloging or investigative purposes. A typical query would be something like, "In how many images does this object occur?" Another would be, "Catalog all occurrences and properties-observations of objects in images satisfying certain conditions." Unfortunately, unlike dealing with a relational database or the text of a book, there is no easy way for the user to formulate the required query. This poses a potentially difficult bottleneck that stands in the way of making the notion of a digital-image library a reality.

We propose an approach for developing a system that learns from examples. Hence, rather than issuing queries, the user simply provides training examples. This approach promises to bypass a crucial bottleneck in the way humans currently interact with large databases: query formulation. For most interesting image-analysis tasks, formulating queries to specify a set of target objects (regions) requires solving difficult problems that often involve effectively translating human visual intuition into pixel-level algorithmic constraints. This task is fairly challenging in its own right. In many cases, formulating the query can be impractical for a user to do. Querying a database by providing examples and counterexamples forms a novel and powerful basis for a new generation of intelligent database interface tools. Such tools could enable order-of-magnitude improvements in both the quantity and the quality of analyses of digitized image libraries.

Encoding Knowledge in Pixel-to-Feature Projections

Although it would be convenient to have a system that can be trained directly from training data given only pixel values as input (for example, the work on recognizing volcanoes on Venus (Burl et al. 1994) or some approaches to face recognition (Turk and Pentland 1991), we recognize that domain-specific knowledge is important and often crucial to a recognition task. In many cases, significant domain knowledge can effectively be provided to a learning algorithm in the form of transformations from pixel space to feature space. Hence, a user might be able to define a large number of features that are likely to contain the necessary information to perform the recognition task. The features serve to transform the problem from the noisy, high-dimensional, and highly correlated pixel space to a much–lower-dimensional space. In the process, noise and random variation are greatly reduced. Note that although the users might have good features to measure about targets of interest, they might not necessarily have effective recognizers (classifiers) in feature space. This is exactly the case in our application in astronomy, but this phenomenon holds true across many applications—in medicine, engineering, diagnosis, process control, and so on. In the application we are concerned with in this article, astronomers have a large set of robust features to measure for each object but no good classifiers that can distinguish objects of interest (say, stars from galaxies) in the resulting feature space.

Note that the transformation from pixel to feature space represents an effective way of encoding domain-specific prior knowledge about the problem. Generally speaking, humans tend to find it easier to define features to measure about objects of interest than to encode recognizers (classifiers) for these objects. In a strong sense, this represents an effective way to decompose the difficult (intuitive) recognition–decision-making strategy that is implicitly performed by the human brain. Nevertheless, simply measuring such features does not give a solution: One still needs to design a classifier that can distinguish between classes of interest. This is still typically difficult because although the problem has been transformed into a low-dimensional space (say, 20–80 dimensions), it is a space in which humans can no longer "visualize" solutions. We think that the recognition step is an appropriate stage for using a supervised learning approach to solve the classification problem. We use techniques that can cope with high-dimensional feature spaces, such as decision trees. Note that many traditional classification learning algorithms, for example, *K*-nearest neighbor (Dasarathy 1991), fitting mixtures of Gaussians, or linear discriminate analysis (Fukunaga 1990; Duda and Hart 1973), still have difficulties in these relatively high dimensions. Hence, in general, the recognition task being tackled is still fairly difficult and has no classical solutions.

Sky-Object Cataloging

We target the automation of the tasks of cataloging and analyzing objects in digitized sky images. The sky-image cataloging and analysis tool (SKICAT) (Djorgovski, Weir, and Fayyad 1994) was developed to perform a comprehensive analysis of the Second Palomar Observatory Sky Survey (POSS-II) conducted by the California Institute of Technology (Caltech). See Reid et al. (1991) for a detailed description of the POSS-II effort. The photographic plates collected from the survey are digitized at the Space Telescope Science Institute. This process will result in about 3,000 digital images of 23,040 × 23,040 sixteen-bit pixels each, totaling over 3 terabytes of data. When complete, the survey will cover the entire northern sky in three colors, detecting virtually every sky object to an equivalent B-magnitude object intensity of 22.0.[1] This magnitude is at least one magnitude fainter than previous comparable photographic surveys. We estimate that there are at least 5×10^7 galaxies and 2×10^9 stellar objects (including over 10^5 quasars) detectable in this survey. This data set will be the most comprehensive large-scale imaging survey produced to date and will not be surpassed in scope until the completion of a fully digital all-sky survey in the next decade.

There are three basic functional components to SKICAT, serving the purposes of sky-object catalog construction, catalog management, and high-level statistical and scientific analysis. In this article, we emphasize sky-object catalog construction, with a special focus on the use of a supervised classification learning algorithm to automate object recognition based on training data provided by the astronomers.

The first step in analyzing the results of a sky survey is to identify, measure, and catalog the detected objects in the image into their respective classes (for example, stars versus galaxies). Once the objects have been

… the transformation from pixel to feature space represents an effective way of encoding domain-specific prior knowledge about the problem.

classified, further scientific analysis can proceed. For example, the resulting catalog can be used to test models of the formation of large-scale structure in the universe; probe galactic structure from star counts as in Weir, Djorgovski, and Fayyad (1995); perform automatic identifications of radio or infrared sources; and so forth (Weir, Djorgovski, and Fayyad 1995; Djorgovski, Weir, and Fayyad 1994; Weir 1994; Weir et al. 1994). Reducing the images to catalog entries is an overwhelming task that inherently requires an automated approach. The goal of our project is to automate this process, providing a consistent and uniform methodology for reducing the data sets. This will provide the means for objectively performing tasks that formerly required subjective and visually intensive manual analysis. Another goal of this work is to classify objects whose brightness (isophotal magnitude) is too faint for recognition by inspection, thus requiring an automated classification procedure. We do this by using a limited set of high-resolution CCD images in which it is possible for astronomers to assign classes to faint objects. The learning algorithm's job is to find a classifier that can predict classes of faint objects (which are the majority of objects on any plate) based only on measurements from the lower-resolution images (see the Classifying Faint Objects section).

Decision Trees and Rules

A classification learning algorithm is given as input a set of examples that consist of vectors of attribute values (feature vectors) and a class. Hence, an example is a point in feature space. The goal is to output a classification scheme, known as a *classifier*, that will predict the class variable based on the values of the attributes. When the class variable is continuous, the problem is a regression problem. In the case of a categorical class variable, the problem is a classification problem. A particularly efficient method for producing classifiers from data is to generate a decision tree. A decision tree consists of nodes that are tests on the attributes. The outgoing branches of a node correspond to all the possible outcomes of the test at the node. The examples at a node in the tree are thus partitioned along the branches, and each child node gets its corresponding subset of examples.

Decision tree–based approaches to classification learning are typically preferred because they are efficient and, thus, can deal with large training data sets. In addition, the final classifier produced is symbolic and, therefore, not difficult for domain experts to interpret (as opposed to a neural network or a pattern-recognition–based approach).

In brief, a top-down, nonbacktracking decision tree algorithm works as follows (Quinlan 1986; Breiman et al. 1984): Assume we are given a data set of classified examples expressed in terms of a set of attributes. The attributes can be nominal (discrete, categorical) or continuous valued (numeric). The algorithm first discretizes the continuous-valued attributes by partitioning the range of each into at least two intervals (Fayyad and Irani 1992a). For each discrete (or discretized) attribute, the algorithm first formulates a logical test involving the attribute. The test partitions the data into several subsets. For example, in ID3 (Quinlan 1986) and C4.5 (Quinlan 1992), the value of the attribute is tested, and a branch is created for each value of the attribute.

A selection criterion is then applied to select the attribute that induces the best partition on the data. Once selected, a branch for each outcome of the test involving the attribute is created, resulting in at least two child nodes to the parent node. The algorithm is applied recursively to each child node. The algorithm refrains from further partitioning of a given node when all examples in it belong to one class or when no more tests for partitioning it can be formulated. Thus, a leaf node predicts a classification (sometimes probabilistically).

Greedy Tree Generation

Because a large number of possible trees are consistent with the training data, a greedy search is used. The tree starts at a single root node containing all the training data. The algorithm makes a local determination of the best choice of attribute along which the data are to be split. The data are then partitioned along the values of the selected attribute creating the children. The algorithm is then applied recursively to each child node. It takes four rules to specify a greedy tree-growing algorithm:

DRule1 selects the best attribute to be used in splitting a node.

DRule2 decides how the data are to be split along the values of the attribute selected by DRule1.

DRule3 is a stopping rule that determines whether a node should not be split any further and, hence, be deemed a leaf node.

DRule4 assigns a class prediction to be associated with each leaf node.

In addition to these four rules, numeric

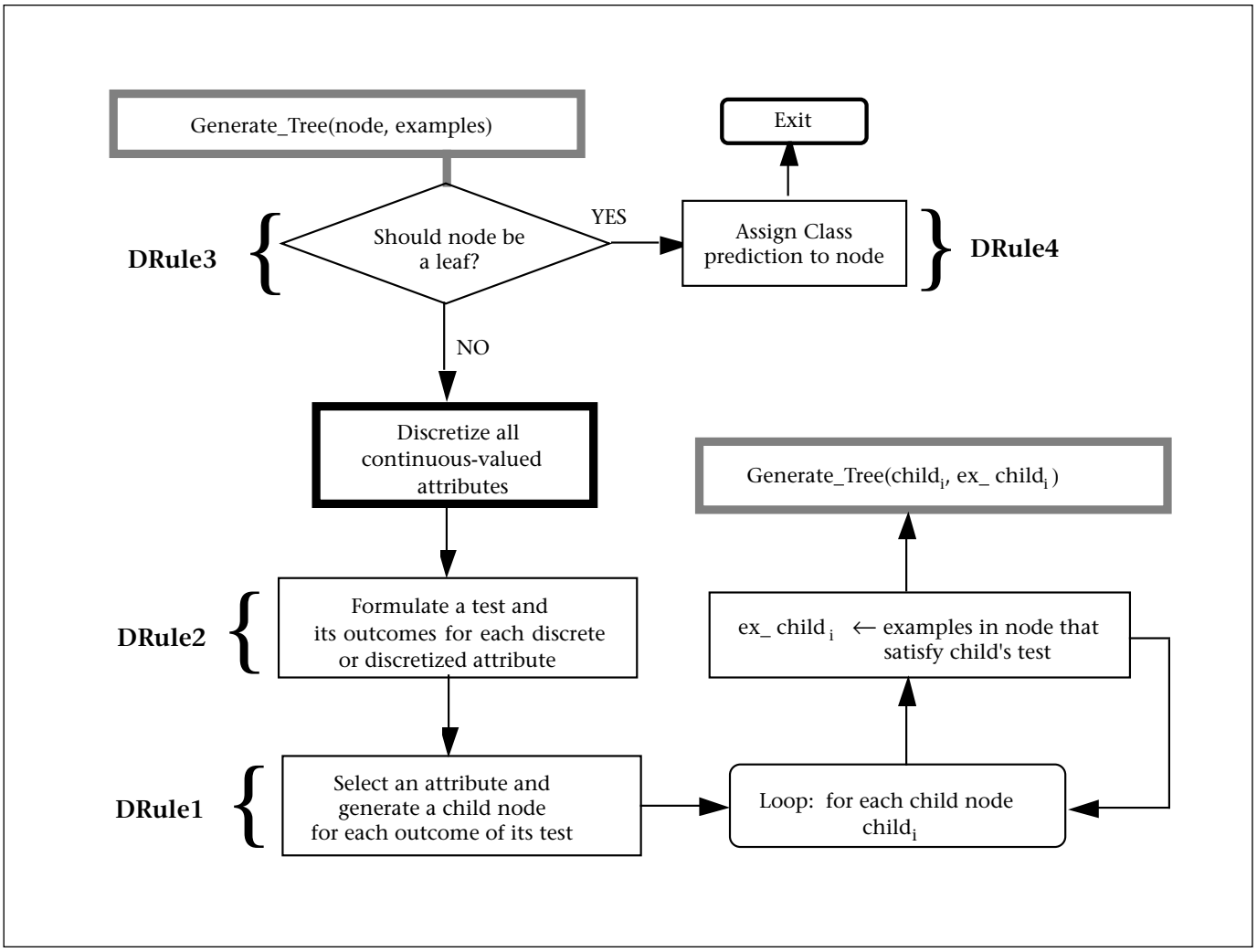

Figure 1. Greedy Tree Generation: Four Decision Rules Specify an Algorithm.

attributes need some special handling, which is done by discretizing the attributes based on the data at each node (Fayyad and Irani 1992a). Discretization can be viewed as a way to extract a symbolic condition involving the attribute. The simplest such condition is to test against a threshold value in the range of an attribute, thus turning it into a binary-valued attribute. Discretization can be viewed as part of DRule2. Figure 1 gives a flowchart for greedy tree growing and shows where the four rules fit.

It is beyond the scope of this article to cover the details of the algorithms. For details relating to the algorithms used in the application covered here, refer to Fayyad, Djorgovski, and Weir (1996). Commercially available algorithms for tree generation use impurity measures such as GINI in CART (Breiman et al. 1984) or mutual information entropy between the attribute and the class variable (used also by CART [Breiman et al. 1984] and in ID3-C4.5 [Quinlan 1992]). We use variants of these algorithms that avoid some of their problems.

For example, rather than splitting the data along all values of a selected attribute, as is customary, the GID3* algorithm (Fayyad 1994) can branch on arbitrary individual values of an attribute and lump the rest of the values in a single default branch representing a subset of the values of an attribute. Unnecessary subdivision of the data can thus be reduced. See Fayyad (1994) for more details. We also use the O-BTREE algorithm (Fayyad and Irani 1992b), which is designed to overcome problems with the information-entropy selection measure itself. O-BTREE creates strictly binary trees and uses a measure from a different fam-

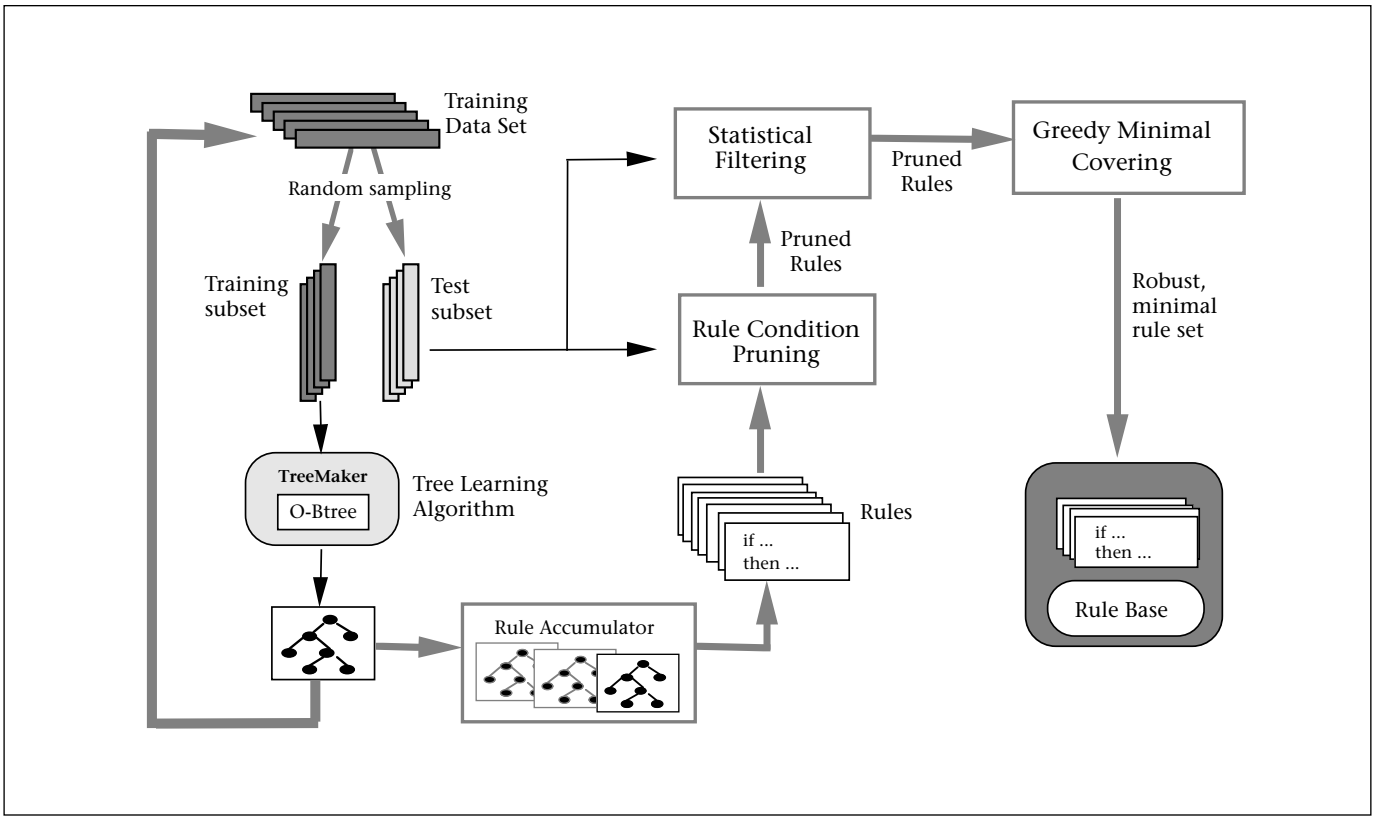

Figure 2. Overview of the RULER Learning System.

ily of measures that detect class separation rather than class impurity. For details on problems with entropy measures and empirical evaluation of O-BTREE, refer to Fayyad and Irani (1992b) and Fayyad (1991).

Both O-BTREE and GID3* differ from ID3 and C4.5 in one additional aspect: the discretization algorithm used at each node to locally discretize continuous-valued attributes. Whereas CART and C4.5 use a binary interval discretization algorithm, we use a generalized version of the algorithm that derives multiple intervals rather than strictly two. For details and empirical tests showing that this algorithm does indeed produce better trees, see Fayyad (1991) and Fayyad and Irani (1993). We have found that this capability improves performance considerably in several domains.

Optimization of Rules from Trees

The very reason that makes decision tree generation efficient (the fact that data are quickly partitioned into ever smaller subsets) is also the reason why overfitting and incorrect classification occur. As data are divided, chance correlations in attribute values begin to appear significant to the algorithm, leading to generation trees that overfit the data (that is, are too specific because they used irrelevant conditions). Typically, to overcome overfit in decision trees, the tree is pruned (Quinlan 1986; Breiman et al. 1984).

We use an approach, called RULER, that is based on extracting multiple trees from a training set and then pruning the rules extracted from the trees. A single tree represents a set of rules. Each path from the root node to a leaf is a classification rule whose conditions are the branches traversed and whose prediction is the class assignment associated with the leaf. In multiple passes, RULER randomly partitions a training set into a training subset and a test subset. A decision tree is generated from the training set, and its rules are tested on the corresponding test set. Using Fisher's exact test (Finney et al. 1963) (the exact hypergeometric distribution), RULER evaluates each condition in a given rule's preconditions for relevance to the class predicted by the rule. Conditions that are deemed to be irrelevant are pruned away. This process results in a large number of redundant rules obtained from the multitude of (similar) trees. The basic idea is to pick the best rules (pruned leaves) from each tree and discard the majority of the rules that were the result of weakly

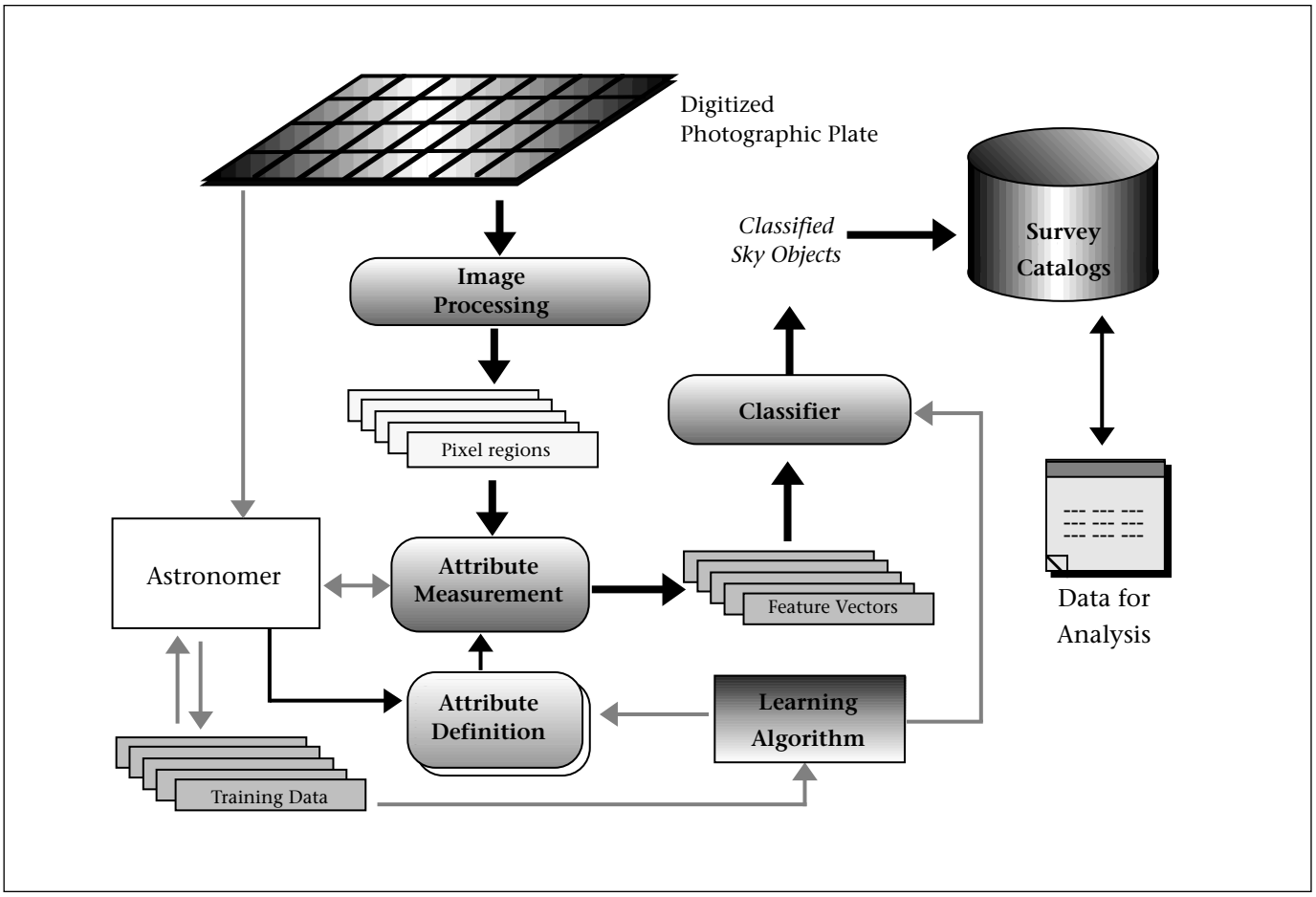

Figure 3. An Overview of the SKICAT Plate-Cataloging Process.

supported correlations in the data. Figure 2 gives an overview of the RULER system.

A greedy-covering algorithm is then employed to select a minimal subset of rules that covers the examples. Using RULER, we can typically produce a robust set of rules that has fewer rules than any of the original decision trees used to create it. The fact that decision tree algorithms constitute a fast and efficient method for generating a set of rules allows us to generate many trees, without requiring extensive amounts of time and computation.

The Cataloging Process

Existing computational methods for classifying the images would preclude the identification of the majority of objects in each image because they are at levels too faint for traditional recognition algorithms or even manual inspection-analysis approaches. Each of the 3,000 digitized plates, consisting of $23{,}040^2$ pixels, is subdivided into a set of partially overlapping frames. Each frame represents a small part of the plate that is small enough to be manipulated and processed conveniently. Figure 3 depicts the overall architecture of the SKICAT catalog construction and classification process.

Low-level image processing and object separation are performed by a modified version of the FOCAS image-processing public-domain software (Valdes 1982; Jarvis and Tyson 1981). The FOCAS image-processing steps detect contiguous pixels in the image that are to be grouped as one object. The grouping is done using a low-level region-growing algorithm to perform segmentation (object versus sky). Some specialized segmentation algorithms are then applied to decide whether an object needs to be split into two (for example, binary stars, stars that are close on the image, and problems in region growing). Attributes are then measured based on this segmentation. Based on the pixel group constituting a single detected object, FOCAS produces basic attributes describing the object. In the section Normalizing Attributes, we explain the arrow

going from the learning algorithm to the attribute definition box in figure 3, indicating the fact that we used learning in the attribute-measurement process. The goal is to classify objects into four major categories, following the original scheme in FOCAS: (1) star (*s*), (2) star with fuzz (*sf*), (3) galaxy (*g*), and (4) artifact (*long*).

Feature Extraction and Normalization

A total of 40 attributes for each detected object are measured automatically. These base-level attributes are generic quantities typically used in astronomical analyses, including the following FOCAS-defined attributes: (1) isophotal, aperture, core, and asymptotic total magnitudes; (2) isophotal and total areas; (3) sky brightness and sigma (variance); (4) peak, intensity-weighted, and intensity-unweighted positions: *xc, yc, icx, icy, cx, cy*; (5) intensity-weighted and intensity-unweighted image moments: *ir*1*, ir*2*, ir*3*, ir*4*, r*1*, r*2*, ixx, iyy, ixy, xx, yy, xy*; and (6) ellipticity and position angle (orientation).

The base-level attributes are not sufficient for accurate classification of the fainter objects that constitute the majority of all detected objects. Furthermore, the base-level attributes do not exhibit desirable invariances that would allow a classifier trained on one plate to make accurate predictions on a different plate that was photographed on a different night with different sky conditions. Hence, a difficult feature-extraction problem needs to be addressed before we can proceed with automated classification.

In classification learning, the choice of attributes used to define examples is by far the single most determining factor of the success or failure of the learning algorithm. Because the base-level features do not provide a suitable feature space in which to perform object-accurate classification, it was necessary to derive additional attributes that have sufficient invariance within a plate (that is, along the borders versus in the center) and across plates.

Low-accuracy classifiers and simple analysis of the value distributions across plates indicated the need for new invariances. For example, we determined that the base-level measurements, such as background sky level, area, and average intensity, are image dependent and, thus, inherently sensitive to plate-to-plate and even frame-to-frame variation. For the learning algorithms to be able to produce robust classifiers, new attributes had to be derived from the base-level attributes.

Normalizing Attributes

Using the following approach, we compute four new normalized attributes based on four base-level attributes: (1) core magnitude; (2) log of the isophotal area; (3) intensity-weighted first-moment radius; and (4) *S*, which is a function of area, core luminosity, and isophotal intensity:

$$S = \frac{\text{Area}}{\log[L_{\text{core}}/(9 \times \text{Ispht})]} .$$

First we derive a nonlinear curve (the stellar locus) in the two dimensions defined by magnitude versus the original base-level attribute for each frame within a plate. We define the new attribute to be the distance of each object from the stellar locus for the plate. We essentially subtract out the stellar locus to normalize the attributes. The quantities described are used by astronomers, and many of them have physical interpretations.

The result of this process is a set of features that exhibit a good degree of invariances across plates and within different regions on a plate. For example, a constant shift in background sky brightness, resulting in differences in intensity observations would be removed by such processing. An example of this normalization process is shown in figure 4, where we can see the stellar locus curve fit both before and after normalization. In each figure, a point for every object is plotted in the two-dimensional space defined by the total magnitude versus log(area).

In addition to the four normalized attributes just described, we compute two additional attributes that are particularly stable across images. However, the computation of these additional attributes requires an empirical measurement based on a selection of stars from each frame. This process was achieved through a second application of the learning algorithms during the attribute-measurement process; this process is depicted at the bottom of figure 4 by the arrow going from "learning" to "attribute definition."

Because of turbulence in the earth's atmosphere, point sources in the sky (stars) appear as blurred, quasi-Gaussian intensity distributions. By selecting some of the objects on a frame that are obviously resolved (sure-thing stars), one can hope to model this effect and compensate for it when classifying. To this end, we fit the pixel values of these sure-thing stars to define a point-spread–function (PSF) template. Using the PSF template, the FOCAS *resolution routine* determines the best-fitting scale (α) and fraction (β) values, which parameterize the fit of a blurred (or sharp-

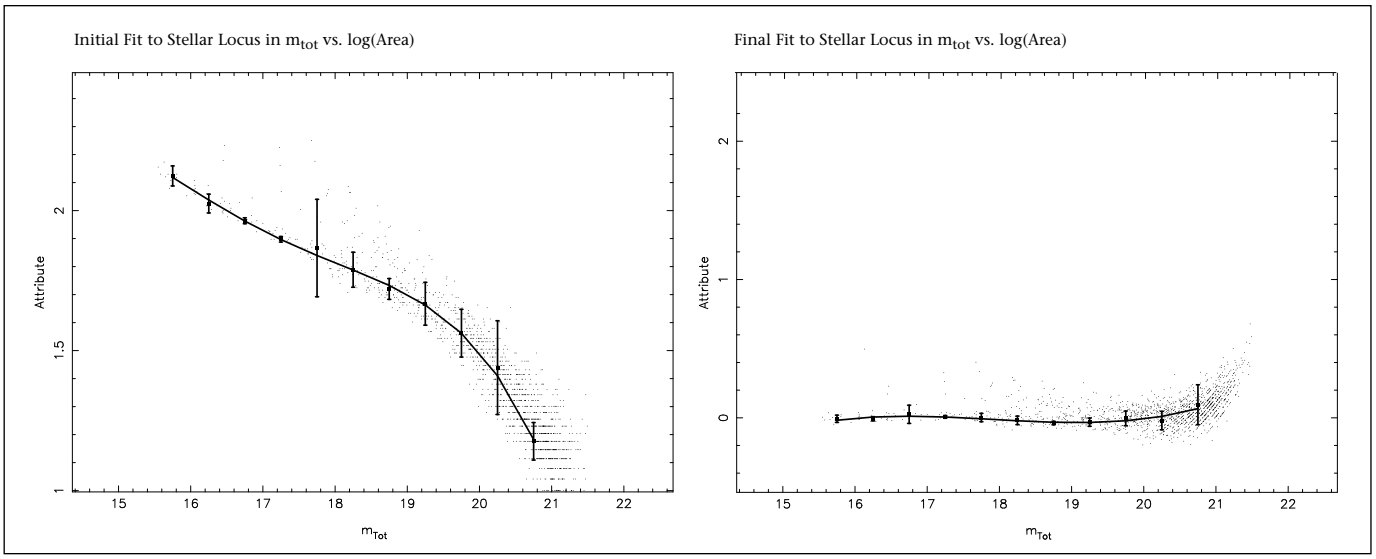

Figure 4. The log(area) Attribute before and after Normalization.

ened) version of the PSF to each object (Valdes 1982). The template used to model each object is of the form

$$t(r_i) = \beta s(r_i/\alpha) + (1 - \beta)\, s(r_i) \quad ,$$

where r_i is the position of pixel i, α is the broadening (sharpening) parameter, β is the fraction of broadened PSF, and $s(r_i)$ represents the pixel value at position r_i.

To form the PSF template, the sure-thing stars would normally be hand selected from an image by the astronomer. We refer to this process as the *star-selection subproblem*. To automate the measurement of these additional attributes, we trained a classifier to detect the sure-thing stars in each frame using the four normalized attributes described previously. We have achieved 98-percent accuracy in detecting the sure-thing stars used to determine the PSF template. Once a template is formed, the resolution attributes are measured automatically for each object on the frame. See the Classification Results section for a discussion of the impact of adding these derived attributes.

Classifying Faint Objects

How can a learning algorithm learn to classify objects too faint for humans to classify? In addition to the scanned photographic plates, we have access to CCD images that span several small regions in some of the frames. CCD images are obtained from a separate telescope. The main advantage of a CCD image is higher resolution and higher signal-to-noise ratio at fainter levels. Hence, many of the objects that are too faint to be classified by inspection on a photographic plate are easily classifiable in a CCD image.

To produce a classifier that classifies faint objects correctly, the learning algorithm needs training data consisting of faint objects labeled with the appropriate class. The class label is therefore obtained by examining the CCD frames. This process is illustrated in figure 5. Once trained on properly labeled objects, the learning algorithm produces a classifier that is capable of properly classifying objects based on the values of the attributes measured from the lower-resolution plate image. Hence, in principle, the classifier will be able to classify objects in the photographic image that are simply too faint for an astronomer to classify by inspection. With the class labels, the learning algorithms are basically being used to solve the more difficult problem of separating the classes in the multidimensional space defined by the set of attributes derived by image processing. This method is expected to allow us to classify objects that are at least one magnitude fainter than objects classified in photographic all-sky surveys to date.

Classification Results

To assess classifier accuracy, we used data consisting of objects collected from four different plates from regions for which we had CCD image coverage. CCD plates provide us with the "ground truth" because these are the only

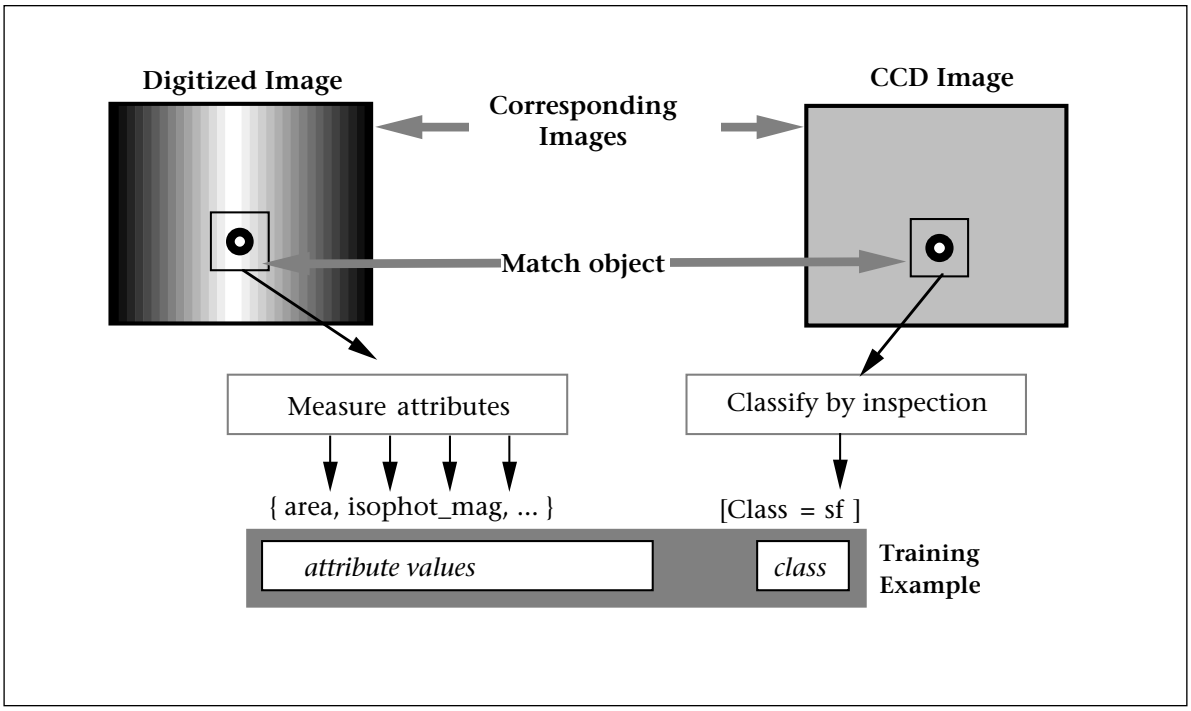

Figure 5. Constructing Training Examples for Faint Objects.

data for which true accurate classifications are available. Each of the learning algorithms is trained on a data set from three plates and tested on data from the remaining plate for cross-validation. This estimates our accuracy in classifying objects across plates. Note that the plates cover different regions of the sky and that CCD frames cover multiple small portions of each plate. The training data consisted of 1688 objects that were classified manually by Nicholas Weir by examining the corresponding CCD frames. It is noteworthy that for the majority of these objects, the astronomer would not be able to reliably determine the classes by examining the corresponding survey (digitized photographic) images. All attributes used by the learning algorithms are derived from the survey images and not, of course, from the higher-resolution CCD frames.

The accuracy for RULER averaged 94.2 percent. In comparison, GID3* and O-BTREE achieved 90.1 percent and 91.2 percent, respectively. These are average accuracy results obtained using cross-validation over the four images. Within each cross-validation fold, we sample a training set 10 times and evaluate each resulting decision tree on the remaining subset. The results for RULER quoted here are with O-BTREE as the decision tree generation component and were obtained by cycling through tree generation and rule merging 10 times.

As a baseline comparison to give the reader a feel for the degree of difficulty of the problem, ID3 achieved only 75.6-percent accuracy on average. If one adds tree pruning and other optimizations (as in C4.5), improved results can be obtained. Note that GID3* and O-BTREE results do not involve any pruning of the trees. Results with CART, which performs a significant amount of pruning using cross-validation, compare favorably with GID3* and O-BTREE results.

For details of data and results and for a detailed breakdown of accuracy results as a function of object brightness, the reader is referred to Weir, Fayyad, and Djorgovski (1995). Results are provided as magnitude gets fainter, and accuracy measurements are broken down into completeness versus contamination for both stars and galaxies (Weir, Fayyad, and Djorgovski 1995).

To emphasize the importance of selecting the right attributes, we report the effect of not computing the two attributes described in the section Normalizing Attributes. When the same experiments were conducted without using the resolution scale and resolution-fraction attributes, the results were significantly worse. The error rates jumped above 20 percent for O-BTREE, above 25 percent for GID3*,

and above 30 percent for ID3. The respective sizes of the trees grew significantly as well, which we took as evidence that the resolution attributes are important for the classification task. The strong dependence on the presence of all relevant features is a facet of the classification problem that makes it particularly difficult for humans to solve: If one fails to include one or two critical attributes, the problem suddenly becomes impossible. However, a priori one has no idea which subset of the attributes is the critical one for accurate classification.

Verification of Results

As mentioned earlier, in addition to using the CCD frames to derive training data for the machine-learning algorithms, we also use them to verify and estimate the performance of our classification technique. Testing is performed on data sets that are drawn independently from the training data. An additional source of internal consistency checks comes from the fact that the plates, and the frames within each plate, are partially overlapping. Hence, objects inside the overlapping regions will be classified in more than one context. By measuring the rate of conflicting classifications, we can obtain further estimates of the statistical confidence in the accuracy of our classifier. For the purposes of the final catalog production, a method is being designed for resolving conflicts on objects within regions of overlap. We have not yet collected reportable results on this aspect of the problem.

To demonstrate the difficulty and significance of the classification results presented to this point, consider the example shown in figure 6. This figure shows four image patches, each centered about a faint sky object that was classified by SKICAT. These images were obtained from a plate that was not provided to SKICAT in the training cycle, and the objects are part of a region in the sky containing the Abell 1551 cluster of galaxies near the North Galactic Pole. SKICAT correctly classified the top two objects as stars and the bottom two as galaxies. According to astronomers, the objects shown in figure 6 are too faint for reliable classification. As a matter of fact, an astronomer visually inspecting these images would be hard pressed to decide whether the object in the lower right-hand corner is a star or a galaxy. The object in the upper right-hand corner appears as a galaxy based on visual inspection. On retrieving the corresponding higher-resolution CCD images of these objects, it was clear that the SKICAT classification was indeed correct. Note that SKICAT produced the prediction based on the lower-resolution survey images (shown in figure 6). This example illustrates how the SKICAT classifier can correctly classify the majority of faint objects that even the astronomers cannot classify. Indeed, the results indicate that SKICAT has a better than 90-percent accuracy identifying objects that are one full magnitude below the comparable magnitude limit in previous automated Schmidt plate surveys.

Unsupervised Learning and New Scientific Discoveries

An additional form of independent confirmation of these results comes from the use of the SKICAT catalog in deriving new science results. For example, using the accurate classification of faint objects given by SKICAT, we were able to help a group of astronomers using SKICAT to discover 16 new high–red-shift quasars in the universe (Kennefick et al. 1995).[2] The search for quasars is an expensive operation requiring many observations. Because SKICAT provides accurate classifications of faint stars, the astronomers were able to use the classes to significantly narrow the search. By combining classes and information from various color attributes, the new quasars were discovered using at least one order of magnitude fewer observations than were required by a comparable effort conducted by Schmidt, Schneider, and Gunn (1995). The accurate classes translated into a small number of false alarms that astronomers had to cope with. The results after the first five quasars were discovered are detailed in Kennefick et al. (1995).

We have also begun exploring the application and implementation of unsupervised classification techniques such as AUTOCLASS, a Bayesian clustering technique that models the data using mixtures of Gaussians (Cheeseman and Stutz 1996). Unlike the supervised classification that we have described to this point, where the algorithm learns how to distinguish user-specified classes within the data, unsupervised classification consists of identifying the statistically significant classes within the data itself. For example, one could use this type of method to try to systematically detect new classes of objects within astronomical catalogs.

Our own initial experiments in applying AUTOCLASS to POSS-II appear to confirm the validity and usefulness of this approach. After

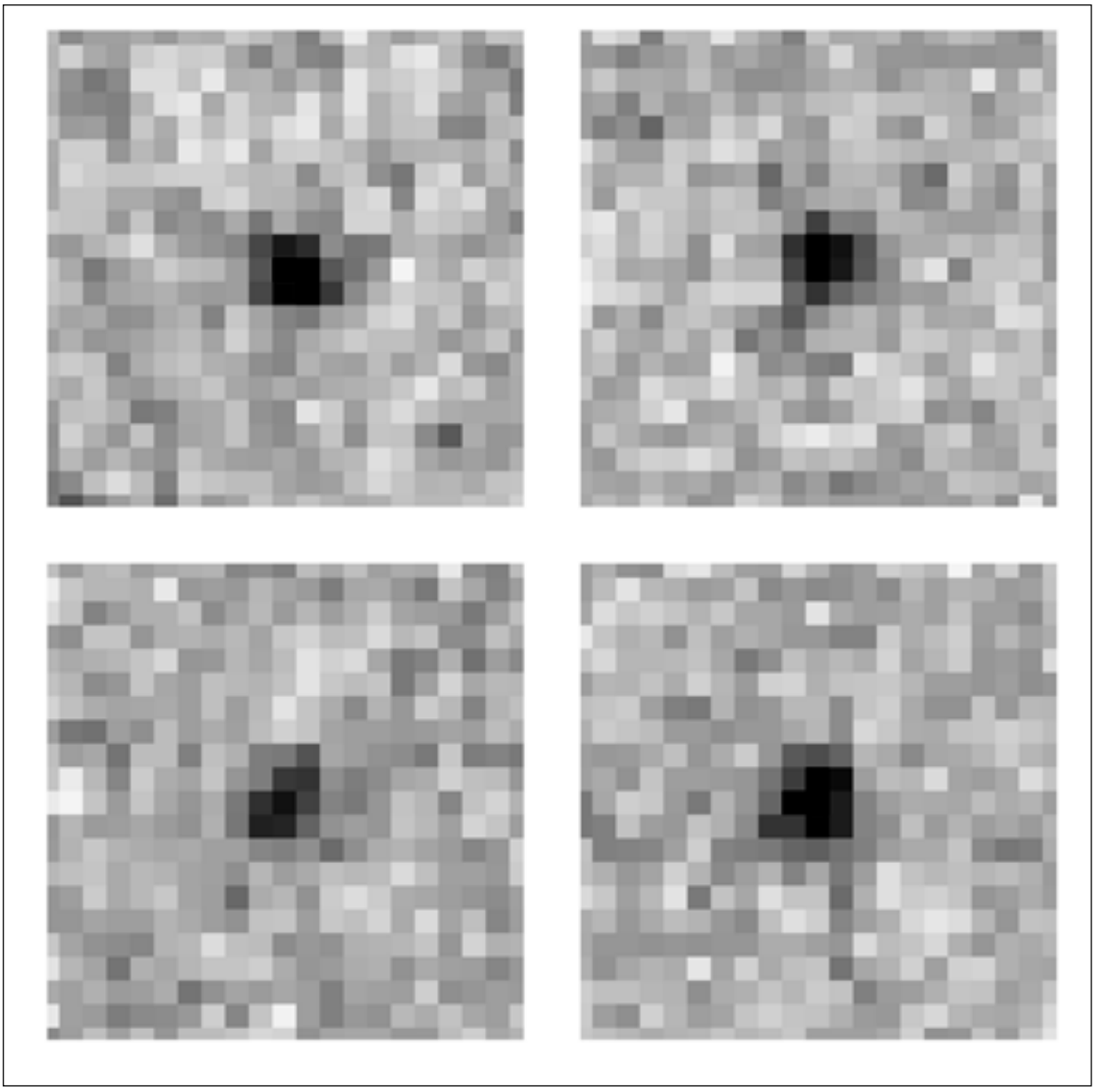

Figure 6. An Illustrative Example: Four Faint Sky Objects.

supplying AUTOCLASS with eight-dimensional feature vectors from a sample of several hundred objects from four fields, it analyzed the distribution of the objects in this parameter space and suggested four distinct classes within the data. Representative objects from these four classes are presented in figure 7. Visually, the classes seem to divide into stellar objects, stellarlike objects with a low–surface-brightness halo, and diffuse or irregular objects with and without a central core (DeCarvalho et al. 1995). The two classes represented by the top two rows are, in reality, stars. The bottom two rows represent classes that consist entirely of galaxies. Note that in its classification, AUTOCLASS did not mix stars with galaxies in this well-understood data set. This result is significant considering that no class information was given to the program.

However, to achieve these results, we had to bin the values of one of the parameters (isophotal magnitude) before presenting AUTOCLASS with the data. Thus, we partitioned the data by meaningful magnitude ranges before running AUTOCLASS on each subset. We also selected the eight-dimensional subspace by hand. Nevertheless, AUTOCLASS's success at distinguishing these apparently physically rele-

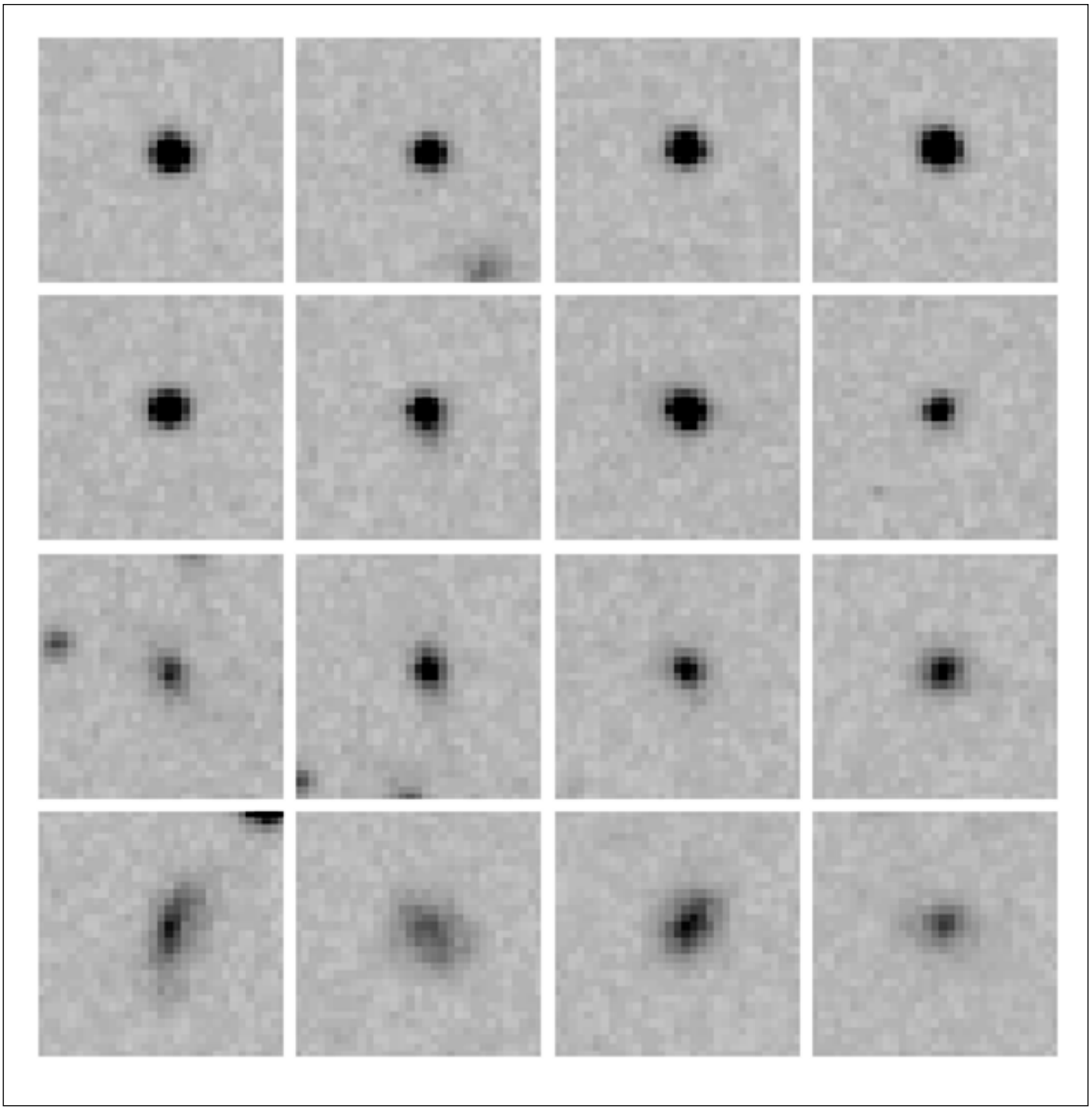

Figure 7. A Sampling of the Four Classes Found by AUTOCLASS.

vant classes based just on eight image parameters suggests that far richer and innovative results might be in store when one matches multiple catalogs together, increasing the informational dimensionality of the data set manyfold. Problems of extending clustering algorithms to high-dimensional spaces and large data sets still need to be addressed.

Concluding Remarks and Future Directions

With SKICAT, classification learning algorithms proved to a be useful and powerful tool in the automation of a significant scientific data analysis task, producing tangible new scientific results (Weir, Djorgovski, and Fayyad

1995; Weir, Fayyad, and Djorgovski 1995). SKICAT can catalog and classify objects that are at least one magnitude fainter than objects cataloged in previous surveys. We exceeded our initial accuracy target of 90 percent. This level of accuracy is required for the data to be useful in testing or refuting theories on the formation of a large structure in the universe and on other phenomena of interest to astronomers. The SKICAT tool is now being used to both process and analyze the survey images as they arrive from the digitization instrument.

By effectively defining robust features, we were able to derive classifiers with an accuracy exceeding that of humans for faint objects. Because faint objects constitute the majority of objects on any plate, the number of classified objects available for further scientific analysis is dramatically increased. In effect, the pixels contained important information that was difficult for the human visual system to extract. Projection of the high-dimensional pixel space onto a suitable lower-dimensional feature space allowed us to transform the problem into one solvable by a supervised learning algorithm. By defining additional normalized image-independent attributes, we were able to obtain high-accuracy classifiers within and across photographic plates.

The catalog generated by SKICAT will eventually contain about two billion entries, representing hundreds of millions of sky objects.

The implications of a tool such as SKICAT for astronomy might indeed be profound. One could reclassify any portion of the survey using alternative criteria better suited to a particular scientific goal (for example, star catalogs versus galaxy catalogs). The catalogs will also accommodate additional attribute entries in the event that other pixel-based measurements are deemed necessary. The catalog generated by SKICAT will eventually contain about two billion entries, representing hundreds of millions of sky objects. Unlike the traditional notion of a static printed catalog, our target is the development of a new generation of scientific analysis tools that render it possible to have a constantly evolving, improving, and growing catalog. Without the availability of these tools for the first survey (POSS-I) conducted over four decades ago, no objective and comprehensive analysis of the data was possible. In contrast, we are targeting a comprehensive sky catalog that will be available online for use by the scientific community.

Future directions for this work are being pursued along two fronts: The first targets the automated scientific discovery problem using clustering techniques, as described in the Unsupervised Learning and New Scientific Discoveries section. This involves overcoming two challenges: (1) developing efficient clustering algorithms that can process millions to hundreds of millions of data points efficiently and (2) developing algorithms than can search for very–low-probability classes in data rather than treating such occurrences as noise or negligible outliers.

The second point is particularly important because new discoveries in astronomy are likely to be rare objects. For example, high–red-shift quasars occur with a frequency of 1 every 10 million. Classical approaches to clustering would ignore such minority classes. Random sampling would completely miss them. We are pursuing directions along the lines of specialized iterative sampling schemes for homing in on a sample that is likely to contain objects that are different than the rest of the data.

The second front of future research is to pursue tools for searching large image collections where the user only labels examples. Unlike the case of SKICAT, where astronomers provided a rich set of attributes to measure, we would like to address problems where no such knowledge is available. An example is the Jet Propulsion Laboratory (JPL) adaptive recognition tool (JARTOOL) (Fayyad et al. 1996; Burl et al. 1994) being developed to catalog an estimated 1 million small volcanoes in 30,000 synthetic aperture radar images of the surface of Venus. This image set collected by the Magellan spacecraft represents a situation that is becoming commonplace in science and many other fields. The data are simply too large to examine manually, and the user cannot invest resources to develop a recognition system to automate the task. Our long-term goal is to develop a tool that can be trained by example to perform object recognition in large image libraries. Because the user might not know all the details of the data, we cannot expect the system to be given much background knowledge about the data (for example, in SKICAT in the form of pixel-to-feature transformations). Other applications at JPL involve earthquake measurement, atmospheric modeling, sunspot classification, and time-seris data analysis. Information can be obtained by visiting the World Wide Web URL: http://www-aig.jpl.nasa.gov/mls, the home page of the Machine-Learning Systems Group at JPL.

Should the training-by-example approach advocated in this article prove to be successful and general, the applications would be truly wide ranging. Finding objects of interest in large digital-image libraries can range from

finding a family member in a digital photo album to searching video libraries for a particular target to inspecting manufacturing, surveillance, and remote sensing applications. In medicine, with the proliferation of digital medical imagers and digitized historical image libraries, many opportunities exist, for example, if a medical researcher notices a new pattern and would like the libraries of several hospitals searched for the new pattern and results correlated with treatments and outcomes. A tool that can be trained by example would make such an operation practical and convenient to execute. Of course, we remain far from this long-term goal. We hope the directions we are pursuing will take us closer to such general adaptive search and information-gathering tools. This, of course, is the tempting promise of the new field of data mining and knowledge discovery in databases.

Acknowledgments

The majority of the funding for SKICAT was provided by the National Aeronautics and Space Administration (NASA) Office of Space Access and Technology (Code XS): We thank Dr. M. Montemerlo for his support and program management. The Jet Propulsion Laboratory (JPL) SKICAT team included Joe Roden, John Loch, Scott Burleigh, Maureen Burl, and Jennifer Yu. This work was supported by a National Science Foundation graduate fellowship (NW), Caltech President's Fund, NASA contract NAS5-31348 (SD and NW), and the NSF Presidential Young Investigator Award AST-9157412 (SD).

The work described in this article was carried out by JPL, California Institute of Technology, under a contract with NASA.

Notes

1. This is a standard astronomical magnitude scale for measuring relative brightness of astronomical sources. It is logarithmic, with 1 mag = –4 db; the brightest stars visible with a naked eye are first magnitude. Magnitudes are usually defined in a particular band pass, given by a combination of a filter and a detector, for example, the blue (*B*) band.

2. At the time this article was written, only 5 objects had been found. As of April 1996, the count stood at 20.

References

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1984. *Classification and Regression Trees*. Monterey, Calif.: Wadsworth and Brooks.

Burl, M. C.; Fayyad, U. M.; Perona, P.; Smyth, P.; and Burl, M. P. 1994. Automating the Hunt for Volcanoes on Venus. In Proceedings of the 1994 Computer Vision and Pattern-Recognition Conference (CVPR-94), 302–309. Washington, D.C.: IEEE Computer Society.

Cheeseman, P., and Stutz, J. 1996. Bayesian Classification (AUTOCLASS): Theory and Results. In *Advances in Knowledge Discovery and Data Mining,* eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 153–180. Menlo Park, Calif.: AAAI Press.

Cheeseman, P.; Self, M.; Kelly, J.; Taylor, W.; Freeman, D.; and Stutz, J. 1988. Bayesian Classification. In Proceedings of the Seventh National Conference on Artificial Intelligence, 607–611. Menlo Park, Calif.: American Association for Artificial Intelligence.

Dasarathy, B. V. 1991. *Nearest Neighbor Norms: NN Pattern Classification Techniques*. Los Alamitos, Calif.: IEEE Computer Society Press.

DeCarvalho, R.; Djorgovski, S. G.; Weir, N.; Fayyad, U.; Cherkauer, K.; Roden, J.; and Gray, A. 1995. Clustering Analysis Algorithms and Their Applications to Digital POSS-II Catalogs. In *Astronomical Data Analysis Software and Systems IV,* eds. R. Hanisch, et al. A. S. P. Conf. Ser 77:272.

Djorgovski, S. G.; Weir, N.; and Fayyad, U. M. 1994. Processing and Analysis of the Palomar—STScI Digital Sky Survey Using a Novel Software Technology. In *Astronomical Data Analysis Software and Systems III,* eds. D. Crabtree, R. Hanisch, and J. Barnes. A. S. P. Conf. Ser. 61:195.

Duda, R. O., and Hart, P. E. 1973. *Pattern Classification and Scene Analysis*. New York: Wiley.

Fayyad, U. M. 1994. Branching on Attribute Values in Decision Tree Generation. In Proceedings of the Twelfth National Conference on Artificial Intelligence, 601–606. Menlo Park, Calif.: American Association for Artificial Intelligence.

Fayyad, U. M. 1991. On the Induction of Decision Trees for Multiple Concept Learning. Ph.D. thesis, Department of Electrical Engineering and Computer Science, University of Michigan at Ann Arbor.

Fayyad, U. M., and Irani, K. B. 1993. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, 1022–1027. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence.

Fayyad, U. M., and Irani, K. B. 1992a. On the Handling of Continuous-Valued Attributes in Decision Tree Generation. *Machine Learning* 8(2).

Fayyad, U. M., and Irani, K. B. 1992b. The Attribute-Selection Problem in Decision Tree Generation. In Proceedings of the Tenth National Conference on Artificial Intelligence, 104–110. Menlo Park, Calif.: American Association for Artificial Intelligence.

Fayyad, U. M.; Djorgovski, S. G.; and Weir, N. 1996. Automating the Analysis and Cataloging of Sky Surveys. In *Advances in Knowledge Discovery and Data Mining,* eds., U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 471–494. Menlo Park, Calif.: AAAI Press.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds. 1996. *Advances in Knowledge Discovery and Data Mining*. Menlo Park, Calif.: AAAI Press.

Fayyad, U. M.; Smyth, P.; Burl, M. C.; and Perona, P. 1996. Learning to Catalog Science Images. In *Early Visual Learning,* eds. S. Nayar and T. Poggio. New York: Oxford University Press. Forthcoming.

Finney, D. J.; Latscha, R.; Bennett, B. M.; and Hsu, P. 1963. *Tables for Testing Significance in a* 2 3 2 *Contingency Table*. Cambridge, U.K.: Cambridge University Press.

Fukunaga, K. 1990. *Introduction to Statistical Pattern Recognition*. San Diego, Calif.: Academic Press.

Jarvis, J., and Tyson, A. 1981. FOCAS: Faint Object Classification and Analysis System. *Astronomical Journal* 86:476.

Kennefick, J. D.; de Carvalho, R. R.; Djorgovski, S. G.; Wilber, M. M.; Dickson, E. S.; Weir, N.; Fayyad, U. M.; and Roden, J. 1996. The Discovery of Five Quasars at $z > 4$ Using the Second Palomar Sky Survey. *Astronomical Journal* 110(1): 78–86.

Mingers, J. 1989. An Empirical Comparison of Selection Measures for Decision-Tree Induction. *Machine Learning* 3(4): 319–342.

Odewahn, S.; Stockwell, E.; Pennington, R.; Humphreys, R.; and Zumach, W. 1992. Automated Star/Galaxy Discrimination with Neural Networks. *Astronomical Journal* 103:318.

Quinlan, J. R. 1992. C*4.5: Programs for Machine Learning*. San Francisco, Calif.: Morgan Kaufmann.

Quinlan, J. R. 1986. The Induction of Decision Trees. *Machine Learning* 1(1).

Reid, I. N.; Brewer, C.; Brucato, R.; McKinley, W.; Maury, A.; Mendenthall, D.; Mould, J.; Mueller, J.; Neugebauer, G.; Phinney, J.; Sargent, W.; Schombert, J.; and Thicksten, R. 1991. The Second Palomar Sky Survey. *Publications of the Astronomical Society of the Pacific* 103(665).

Schmidt, M; Schneider, D.; and Gunn, J. E. 1995. *Astronomical Journal* 110:68.

Turk, M., and Pentland, A. 1991. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience* 3:71–86.

Valdes, F. 1982. The Resolution Classifier. *Instrumentation in Astronomy* 331(4): 465.

Weir, N. 1994. Automated Analysis of the Digitized Second Palomar Sky Survey: System Design, Implementation, and Initial Results. Ph.D. diss., Math, Physics, and Astronomy Division, California Institute of Technology.

Weir, N.; Djorgovski, S. G.; and Fayyad, U. M. 1995. Initial Galaxy Counts from Digitized POSS-II. *Astronomical Journal* 110(1): 1–20.

Weir, N.; Fayyad, U. M.; and Djorgovski, S. G. 1995. Automated Star/Galaxy Classification for Digitized POSS-II. *Astronomical Journal* 109(6): 2401–2412.

Weir, N.; Djorgovski, S. G.; Fayyad, U. M.; Smith, J. D.; and Roden, J. 1994. Cataloging the Northern Sky Using a New Generation of Software Technology. In *Astronomy from Wide-Field Imaging,* eds. H. MacGillivray, 205. Dordrecht, The Netherlands: Kluwer.

Usama Fayyad is a senior researcher at Microsoft Research. Prior to joining Microsoft in 1996, he headed the Machine Learning Systems Group at the Jet Propulsion Laboratory (JPL), California Institute of Technology, where he developed data-mining systems for automated science data analysis. He remains affiliated with JPL as a distinguished visiting scientist. Fayyad received the JPL 1993 Lew Allen Award for Excellence in Research and the 1994 National Aeronautics and Space Administration Exceptional Achievement Medal. His research interests include knowledge discovery in large databases, data mining, machine-learning theory and applications, statistical pattern recognition, and clustering. He was program cochair of KDD-94 and KDD-95 (the First International Conference on Knowledge Discovery and Data Mining). He is general chair of KDD-96, an editor-in-chief of *Data Mining and Knowledge Discovery,* and coeditor of *Advances in Knowledge Discovery and Data Mining* (AAAI Press, 1996).

S. George Djorgovski is an associate professor of astronomy at the California Institute of Technology. He received his M.A. and Ph.D. degrees from the University of California at Berkeley and was a Harvard University junior fellow before joining the Caltech faculty. His professional interests include observational cosmology, origins and evolution of galaxies and quasars, digital sky surveys, novel approaches to data exploration in astronomy and elsewhere, globular star clusters, and gravitational lenses.

Nicholas Weir is currently an associate in fixed-income research at Goldman, Sachs & Co., where he develops and applies quantitative strategies for pricing and hedging bond securities and their derivatives. He received his Ph.D. in astronomy, with additional studies in physics from the California Institute of Technology, concentrating in digital sky surveys, observational cosmology, and image restoration.