

International Journal of Modern Physics D
© World Scientific Publishing Company

DATA MINING AND MACHINE LEARNING IN ASTRONOMY

NICHOLAS M. BALL

*Herzberg Institute of Astrophysics, National Research Council, 5017 West Saanich Road,
Victoria, BC V9E 2E7, Canada
nick.ball@nrc-cnrc.gc.ca*

ROBERT J. BRUNNER

*Department of Astronomy, University of Illinois at Urbana-Champaign, 1002 West Green Street,
Urbana, IL 61801, USA
bigdog@illinois.edu*

Received Day Month Year
Revised Day Month Year
Communicated by Managing Editor

We review the current state of data mining and machine learning in astronomy. *Data Mining* can have a somewhat mixed connotation from the point of view of a researcher in this field. If used correctly, it can be a powerful approach, holding the potential to fully exploit the exponentially increasing amount of available data, promising great scientific advance. However, if misused, it can be little more than the black-box application of complex computing algorithms that may give little physical insight, and provide questionable results. Here, we give an overview of the entire data mining process, from data collection through to the interpretation of results. We cover common machine learning algorithms, such as artificial neural networks and support vector machines, applications from a broad range of astronomy, emphasizing those where data mining techniques directly resulted in improved science, and important current and future directions, including probability density functions, parallel algorithms, petascale computing, and the time domain. We conclude that, so long as one carefully selects an appropriate algorithm, and is guided by the astronomical problem at hand, data mining can be very much the powerful tool, and not the questionable black box.

Keywords: Keyword1; keyword2; keyword3.

1. Introduction

In its broadest sense, data mining is simply the act of turning raw data from an observation into useful information. This information can be interpreted by hypothesis or theory, and used to make further predictions. This scientific method, where useful statements are made about the world, has been widely employed to great effect in the West since the Renaissance, and even earlier in other parts of the world. What has changed in the past few decades is the exponential rise in available computing power, and, as a related consequence, the enormous quantities of observed data, primarily in digital form. The exponential rise in the amount of available data is

now creating, in addition to the natural world, a digital world, in which extracting new and useful information from the data already taken and archived is becoming a major endeavor in itself. This action of *knowledge discovery in databases* (KDD), is what is most commonly inferred by the phrase data mining, and it forms the basis for our review.

Astronomy has been among the first scientific disciplines to experience this flood of data. The emergence of data mining within this and other subjects has been described^{1,2,3} as the *fourth paradigm*. The first two paradigms are the well-known pair of theory and observation, while the third is another relatively recent addition, computer simulation. The sheer volume of data not only necessitates this new paradigmatic approach, but the approach must be, to a large extent, automated. In more formal terms, we wish to leverage a computational machine to find patterns in digital data, and translate these patterns into useful information, hence *machine learning*. This learning must be returned in a useful manner to a human investigator, which hopefully results in human learning.

It is perhaps not entirely unfair to say, however, that scientists in general do not yet appreciate the full potential of this fourth paradigm. There are good reasons for this of course: scientists are generally not experts in databases, or cutting-edge branches of statistics, or computer hardware, and so forth. What we hope to do in this review, primarily for the data mining skeptic, is to shed light on why this is a useful approach. To accomplish this goal, we emphasize either algorithms that have or could currently be usefully employed, and the actual scientific results they have enabled. We also hope to give an interesting and fairly comprehensive overview to those who do already appreciate this approach, and perhaps provide inspiration for exciting new ideas and applications. However, despite referring to data mining as a whole new paradigm, we try to emphasize that it is, like theory, observation, and simulation, only a part of the broader scientific process, and should be viewed and utilized as such. The algorithms described are *tools* that, when applied correctly, have vast potential for the creation of useful scientific results. But, given that it is only part of the process, it is, of course, not the answer to everything, and we therefore enumerate some of the limitations of this new paradigm.

We start in §1.1 with a summary of some of the advantages of this approach. In §2, we summarize the process from the input of raw data to the visualization of results. This is followed in §3 by the actual application of data mining tools in astronomy. §2 is arranged algorithmically, and §3 is arranged astrophysically. It is likely that the expert in astronomy or data mining, respectively, could infer much of §3 from §2, and vice-versa. But it is unlikely (we hope) that the combination of the two sections does not have new ideas or insights to offer to either audience. Following these two sections, in §4, we combine the lessons learned to discuss the future of data mining in astronomy, pointing out likely near-term future directions in both the data mining process and its physical application. We conclude with a summary of the main points in §5.

1.1. Why Data Mining?

Of course, what astronomers care about is not a fashionable new computational method for ever more complex data analysis, but the *science*. A fancy new data mining system is not worth much if all it tells you is what you could have gained by the judicious application of existing tools and a little physical insight⁴. We therefore summarize some of the advantages of this approach:

- *Getting anything at all*: upcoming datasets will be almost overwhelmingly large. When one is faced with Petabytes of data, a rigorous, automated approach that intelligently extracts pertinent scientific information will be the only one that is tractable.
- *Simplicity*: despite the apparent plethora of methods, straightforward applications of very well-known and well-tested data mining algorithms can quickly produce a useful result. These methods can generate a model appropriate to the complexity of an input dataset, including nonlinearities, implicit prior information, systematic biases, or unexpected patterns. With this approach, *a priori* data sampling of the type exemplified by elaborate color cuts, is not necessary. For many algorithms, new data can be trivially incorporated as they become available.
- *Prior information*: this can be either fully incorporated, or the data can be allowed to completely ‘speak for themselves’. For example, an unsupervised clustering algorithm can highlight new classes of objects within a dataset that might be missed if a prior set of classifications were imposed.
- *Pattern recognition*: an appropriate algorithm can highlight patterns in a dataset that might not otherwise be noticed by a human investigator, perhaps due to the high dimensionality. Similarly, rare or unusual objects can be highlighted.
- *Complimentary approach*: although there are numerous examples where the data mining approach demonstrably exceeds more traditional methods in terms of scientific return. Even when the approach does not produce a substantial improvement, it still acts as an important complementary method of analyzing data, because different approaches to an overall problem help to mitigate systematic errors in any one approach.

2. Overview of Data Mining and Machine Learning Methods

In this section, we review the data mining process. Specifically, as described in §1, this data mining review focuses on knowledge discovery in databases (KDD), although our definition of a ‘database’ is somewhat broad, essentially being any machine-readable astronomical data. As a result, this section is arranged algorithmically. To avoid overlap with §3 on the astronomical uses, we defer most of the application examples to that section. Nevertheless, all algorithms we describe have been, or are of sufficient maturity that they could immediately be applied to astronomical data. The reader who is expert in astronomy but not in data mining is advised to read this section to gain the full benefit from §3. As in any specialized

subject, a certain level of jargon is necessary for clarity of expression. Terms likely to be unfamiliar to astronomers not versed in data mining are generally explained as they are introduced, but for additional background we note that there are other useful reviews of the data mining field^{5,6,7}. Another recent overview of data mining in astronomy by Borne has also been published⁸.

2.1. *Data Collection*

The process of data collection encompasses all of the steps required to obtain the desired data in a digital format. Methods of data collection include acquiring and archiving new observations, querying existing databases according to the science problem at hand, and performing as necessary any cross-matching or data combining, a process generically described as *data fusion*.

A common motivation for cross-matching is the use of multiwavelength data, i.e., data spanning more than one of the regions of the electromagnetic spectrum (gamma ray, X-ray, ultraviolet, optical, infrared, millimeter, and radio). A common method in the absence of a definitive identification for each object spanning the datasets is to use the object's position on the sky with some astrometric tolerance, typically a few arcseconds. Cross-matching can introduce many issues including ambiguous matches, variations of the point spread function (resolution of objects) within or between datasets, differing survey footprints, survey masks, and large amounts of processing time and data transfer requirements when cross-matching large datasets.

A major objective of the *Virtual Observatory* (VO, §4.5) is to make the data collection process more simple and tractable. Future VO webservices are planned that will perform several functions in this area, including cross-matches on large, widely distributed, heterogeneous data.

Common astronomical data formats include FITS⁹, a binary format, and plain ASCII, while an emerging format is VOTable¹⁰. Commonly used formats from other areas of data mining, such as attribute relation file format (ARFF)^a, are generally not widely used in astronomy.

2.2. *Preprocessing of Data*

Some data preprocessing may necessarily be part of the data collection process, for example, sample cuts in database queries. Preprocessing can be divided into steps that make the data to be read meaningful, and those that transform the data in some way as appropriate to a given algorithm. Data preprocessing is often problem-dependent, and should be carefully applied because the results of many data mining algorithms can be significantly affected by the input data. A useful overview of data preprocessing is given by Pyle¹¹.

^a<http://weka.wiki.sourceforge.net/ARFF>

Algorithms may require the object *attributes*, i.e., the values in the data fields describing the properties of each object, to be numerical or categorical, the latter being, e.g. ‘star’, or ‘galaxy’. It is possible to transform numerical data to categorical and vice versa.

A common categorical-to-numerical method is scalarization, in which different possible categorical attributes are given different numerical labels, for example, ‘star’, ‘galaxy’, ‘quasar’ labeled as the vectors $[1,0,0]$, $[0,1,0]$, and $[0,0,1]$, respectively. Note that for some algorithms, one should *not* label categories as, say, 1, 2 and 3, if the output of the algorithm is such that if it has confused an object between 1 and 3 it labels the object as intermediate, in this case, 2. Here, 2 (galaxy) is certainly not an intermediate case between 1 (star) and 3 (quasar). One common algorithm in which such categorical but not ordered outputs could occur is a decision tree with multiple outputs.

Numerical data can be made categorical by transformations such as binning. The bins may be user-specified, or can be generated optimally from the data¹². Binning can create numerical issues, including comparing two floating point numbers that should be identical, objects on a bin edge, empty bins, values that cannot be binned such as *NaN*, or values not within the bin range.

Object attributes may need to be *transformed*. A common operation is the differencing of magnitudes to create colors. These transformations can introduce their own numerical issues, such as division by zero, or loss of accuracy.

In general, data will contain one or more types of *bad values*, where the value is not correct. Examples include instances where the value has been set to something such as -9999 or *NaN*, the value appears correct but has been flagged as bad, or the value is not bad in a formatting sense but is clearly unphysical, perhaps a magnitude of a high value that could not have been detected by the instrument. They may need to be removed either by simply removing the object containing them, ignoring the bad value but using the remaining data, or interpolating a value using other information. Outliers may or may not be excluded, or may be excluded depending on their extremity.

Data may also contain *missing values*. These values may be genuinely missing, for example in a cross-matched dataset where an object is not detected in a given waveband, or is not in an overlapping region of sky. It is also possible that the data should be present, but are missing for either a known reason, such as a bad camera pixel, a cosmic ray hit, or a reason that is simply not known. Some algorithms cannot be given missing values, which will require either the removal of the object or interpolation of the value from the existing data. The advisability of interpolation is problem-dependent.

As well as bad values, the data may contain values that are correct but are outside the desired range of analysis. The data may therefore need to be *sampled*. There may simply be a desired range, such as magnitude or position on the sky, or the data may contain values that are correct but are outliers. Outliers may be included, included depending on their extremity (e.g., n standard deviations),

downweighted, or excluded. Alternatively, it may be more appropriate to generate a random subsample to produce a smaller dataset.

Outside any normalization of the data prior to its use in the data mining algorithm, for example, calibration using standard sources, input or target attributes of the data will often be further normalized to improve the *numerical conditioning* of the algorithm. For example, if one axis of the n -dimensional space created by n input attributes encompasses a range that, numerically, is much larger than the other axes, it may dominate the results, or create conditions where very large and small numbers interact, causing loss of accuracy. Normalization can reduce this, and examples include linear transformations, like scaling by a given amount, scaling using the minimum and maximum values so that each attribute is in a given range such as 0–1, or scaling each attribute to have a mean of 0 and a standard deviation of 1. The latter example is known as *standardization*. A more sophisticated transformation with similar advantages is *whitening*, in which the values are not only scaled to a similar range, but correlations among the attributes are removed via transformation of their covariance matrix to the identity matrix.

2.3. *Attribute Selection*

In general, a large number of attributes will be available for each object in a dataset, and not all will be required for the problem. Indeed, use of all attributes may in many cases worsen performance. This is a well-known problem, often called the *curse of dimensionality*. The large number of attributes results in a high-dimensional space with many low density environments or even empty voids. This makes it difficult to generalize from the data and produce useful new results. One therefore requires some form of *dimension reduction*, in which one wishes to retain as much of the information as possible, but in fewer attributes. As well as the curse of dimensionality, some algorithms work less well with noisy, irrelevant, or redundant attributes. An example of an irrelevant attribute might be position on the sky for a survey with a uniform mask, because the position would then contain no information, and highly redundant attributes might be a color in the same waveband measured in two apertures.

The most trivial form of dimension reduction is simply to use one's judgement and select a subset of attributes. Depending on the problem this can work well. Nevertheless, one can usually take a more sophisticated and less subjective approach, such as principal component analysis (PCA)^{13,14,15}. This is straightforward to implement, but is limited to linear relations. It gives, as the principal components, the eigenvectors of the input data, i.e., it picks out the directions which contain the greatest amount of information. Another straightforward approach is *forward selection*, in which one starts with one attribute and selectively adds new attributes to gain the most information. Or, one can perform the equivalent process but starting with all of the attributes and removing them, known as *backward elimination*.

In many ways, dimension reduction is similar to classification, in the sense that

a larger number of input attributes is reduced to a smaller number of outputs. Many classification schemes in fact directly use PCA. Other dimension reduction methods utilize the same or similar algorithms to those used for the actual data mining: an ANN can perform PCA when set up as an autoencoder, and kernel methods can act as generalizations of PCA. A binary genetic algorithm (§2.4.4) can be used in which each individual represents a subset of the training attributes to be used, and the algorithm selects the best subset. The epsilon-approximate nearest neighbor search¹⁶ reduces the dimensionality of nearest neighbor methods. Other methods include information bottleneck¹⁷, which directly uses information theory to optimize the tradeoff between the number of classes and the information contained, Fisher Matrix¹⁸, Independent Component Analysis¹⁹, and wavelet transforms. The curse of dimensionality is likely to worsen in the future for a similar reason to that of missing values, as more multiwavelength datasets become available to be cross-matched. Classification and dimension reduction are not identical of course: a classification algorithm may build a model to represent the data, which is then applied to further examples to predict their classes.

2.4. Selection and Use of Machine Learning Algorithms

Machine learning algorithms broadly divide into *supervised* and *unsupervised* methods, also known as predictive and descriptive, respectively. These can be generalized to form *semi-supervised* methods. Supervised methods rely on a *training set*^b of objects for which the target property, for example a classification, is known with confidence. The method is trained on this set of objects, and the resulting mapping is applied to further objects for which the target property is not available. These additional objects constitute the *testing set*. Typically in astronomy, the target property is spectroscopic, and the input attributes are photometric, thus one can predict properties that would normally require a spectrum for the generally much larger sample of photometric objects. The training set must be representative, i.e., the parameter space covered by the input attributes must span that for which the algorithm is to be used. This might initially seem rather restrictive, but in many cases can be handled by combining datasets. For example, the zCOSMOS redshift survey²⁰, at one square degree, provides spectra to the depth of the photometric portion of the Sloan Digital Sky Survey (SDSS)²¹, $r \sim 22$ mag, which covers over 8000 square degrees. Since SDSS photometry is available for zCOSMOS objects, one can in principle use the 40,000 zCOSMOS galaxies as a training set to assign photometric redshifts to over 200 million SDSS galaxies.

In contrast to supervised methods, unsupervised methods do not require a training set. This is an advantage in the sense that the data can speak for themselves without preconceptions such as expected classes being imposed. On the other hand,

^bFor many astronomical applications, one might more properly call it a training *sample*, but the term training set is in widespread use, so we use that here to avoid confusion.

if there is prior information, it is not necessarily incorporated. Unsupervised algorithms usually require some kind of initial input to one or more of the adjustable parameters, and the solution obtained can depend on this input.

Semi-supervised methods attempt to allow the best-of-both-worlds, and both incorporate known priors while allowing objective data interpretation and extrapolation. But given their generality, they can be more complex and difficult to implement. They are of potentially great interest astronomically because they could be used to analyze a full photometric survey beyond the spectroscopic limit, without requiring priors, while at the same time incorporating the prior spectroscopic information where it is available.

2.4.1. *Supervised Methods*

The most widely used and well-known machine learning algorithm in astronomy to-date, referred to as far back as the mid 1980s,²² is the *artificial neural network* (ANN, Fig. 1)^{23,24,25}. This consists of a series of interconnected nodes with weighted connections. Each node has an activation function, perhaps a simple threshold, or a sigmoid. Although the original motivation was that the nodes would simulate neurons in the brain,^{26,27} the ANNs in data mining are of such a size that they are best described as nonlinear extensions of conventional statistical methods.

The supervised ANN takes parameters as input and maps them on to one or more outputs. A set of parameter vectors, each vector representing an object and corresponding to a desired output, or target, is presented. Once the network is trained, it can be used to assign an output to an unseen parameter vector. The training uses an algorithm to minimize a cost function. The cost function, c , is commonly of the form of the mean-squared deviation between the actual and desired output:

$$c = \frac{1}{N} \sum_{k=1}^N (o_k - t_k)^2,$$

where o_k and t_k are the output and target respectively for the k th of N objects.

In general, the neurons could be connected in any topology, but a commonly used form is to have an $a : b_1 : b_2 : \dots : b_n : c$ arrangement, where a is the number of input parameters, $b_{1,\dots,n}$ are the number of neurons in each of n one dimensional ‘hidden’ layers, and c is the number of neurons in the final layer, which is equal to the number of outputs. Each neuron is connected to every neuron in adjacent layers, but not to any others. Multiple outputs can each give the Bayesian *a posteriori* probability that the output is of that specific class, given the values of the input parameters.

The weights are adjusted by the training algorithm. In astronomy this has typically been either the well-known backpropagation algorithm^{28,29,30} or the quasi-Newton algorithm²³, although other algorithms, such as Levenberg-Marquardt^{31,32} have also been used.

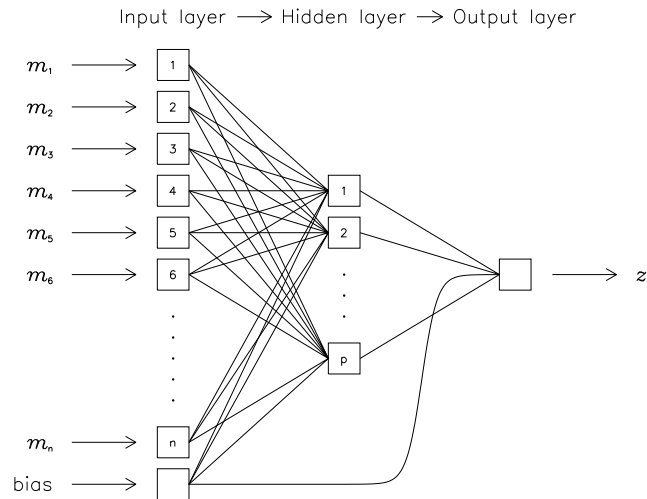


Fig. 1. Schematic of an artificial neural network for an object with n attributes, a hidden layer of size p , and a single continuously-valued output, in this case, the redshift, z . From Firth, Lahav & Somerville³³.

Another common method used in data mining is the *decision tree* (DT, Fig. 2)^{34,35,36,37,38}. Decision trees consist first of a root node which contains all of the parameters describing the objects in the training set population along with their classifications. A node is split into child nodes using the criterion that minimizes the classification error. This splitting subdivides the parent population group into children population groups, which are assigned to the respective child nodes. The classification error quantifies the accuracy of the classification on the test set. The process is repeated iteratively, resulting in layered nodes that form a tree. The iteration stops when specific user-determined criteria are reached. Possibilities include a minimum allowed population of objects in a node (the minimum decomposition population), the maximum number of nodes between the termination node and the root node (the maximum tree depth), or a required minimum decrease resulting from a population split (the minimum error reduction). The terminal nodes are known as the leaf nodes. The split is tested for each input attribute, and can be axis-parallel, or oblique, which allows for hyperplanes at arbitrary angles in the parameter space. The split statistic can be the midpoint, mean, or median of the attribute values, while the cost function used is typically the variance, as with ANN.

In recent years, another algorithm, the *support vector machine* (SVM, Fig. 3)^{40,41,42,43,44,45,46,47,48}, has gained popularity in astronomical data mining. SVM aims to find the hyperplane that best separates two classes of data. The input data are viewed as sets of vectors, and the data points closest to the classification boundary are the support vectors. The algorithm does not create a model

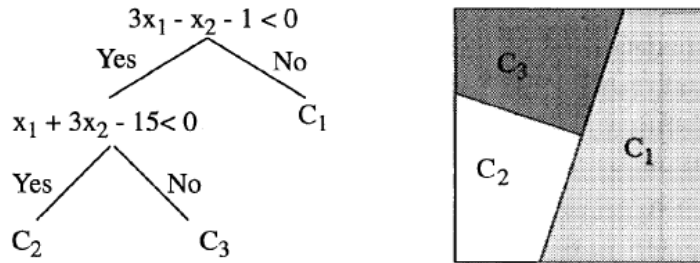


Fig. 2. As Fig. 1, but showing a decision tree. The oblique planes specified by the division criteria on the input attributes x_1 and x_2 at the nodes in this case divide the input parameter space into three regions. From Salzberg *et al.*³⁹.

of the data, but instead creates the decision boundaries, which are defined in terms of the support vectors. The input attributes are mapped into a higher dimensional space using a kernel so that nonlinear relationships within the data become linear (the ‘kernel trick’)⁴⁹, and the decision boundaries, which are linear, are determined in this space. Like ANN and DT, the training algorithm minimizes a cost function, which in this case is the number of incorrect classifications. The algorithm has two adjustable hyperparameters: the width of the kernel, and the regularization, or cost, of classification error, which helps to prevent *overfitting* (§2.5) of the training set. The shape of the kernel is also an adjustable parameter, a common choice being the Gaussian radial basis function. As a result, an SVM has fewer adjustable parameters than an ANN or DT, but because these parameters must be optimized, the training process can still be computationally expensive. SVM is designed to classify objects into two classes. Various refinements exist to support additional classes, and to perform regression, i.e., to supply a continuous output value instead of a classification. Classification probabilities can be output, for example, by using the distance of a data point from the decision boundary.

Another powerful but computationally intensive method is *k nearest neighbor* (*k*NN)^{51,52,53,54,55}. This method is powerful because it can utilize the full information available for each object, with no approximations or interpolations. The training of *k*NN is in fact trivial: the positions of each of the objects in the input attribute space are simply stored in memory. For each test object, the same attributes are compared to the training set and the output is determined using the properties of the nearest neighbors. The simplest implementation is to output the properties of the single nearest neighbor, but more commonly the weighted sum of *k* nearest neighbors is used. The weighting is typically the inverse Euclidean distance in the attribute space, but one can also use adaptive distance metrics. The main drawback of this method is that it is computationally intensive, because for each testing object the entire training set must be examined to determine the nearest neighbors. This

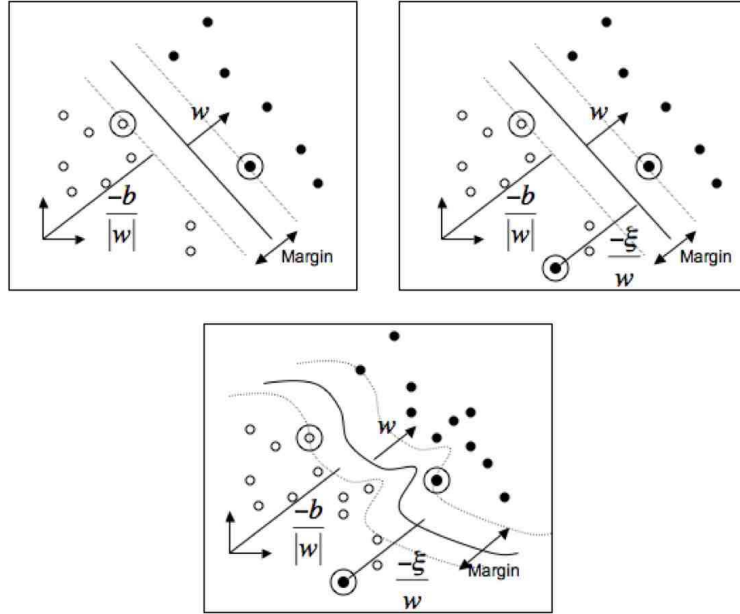


Fig. 3. As Fig. 1, but showing a support vector machine. The circled points are the support vectors between the two classes of objects, represented by open and filled circles. The cases shown are separable and non-separable data with linear and nonlinear boundaries. w is the normal to the hyperplane, and b is the perpendicular distance. From Huertas-Company *et al.*⁵⁰, to which the reader is referred for details of ξ .

requires a large number of distance calculations, since the test datasets are often much larger than the training datasets. The workload can be mitigated by storing the training set in an optimized data structure, such as a kd-tree.

However, in the past few years, novel supercomputing hardware (which is discussed in more detail in §4.7) has become available that is specifically designed to carry out exactly this kind of computationally intensive work, including applications involving a large number of distance calculations. The curve of growth of this technology exceeds that of conventional CPUs, and thus the direct implementation of k NN using this technology has the potential to exceed the performance of conventional CPUs.

2.4.2. Unsupervised Methods

Kernel density estimation (KDE)^{56,57,58,59,60,61,62} is a method of estimating the probability density function of a variable. It is a generalization of a histogram where the kernel function is any shape instead of the top-hat function of a histogram bin. This has the advantages that it avoids the discrete nature of the histogram and does not depend on the position of the bin edges, but the width of the kernel must still be

chosen so as not to over- or under-smooth the data. A Gaussian kernel is commonly utilized. In higher numbers of dimensions, common in astronomical datasets, the width of the kernel must be specified in each dimension.

K-means clustering^{63,64} is an unsupervised method that divides data into clusters. The number of clusters must be initially specified, but since the algorithm converges rapidly, many starting points can be tested. The algorithm uses a distance criterion for cluster membership, such as the Euclidean distance, and a stopping criterion for iteration, for example, when the cluster membership ceases to change.

Mixture models^{65,66} decompose a distribution into a sum of components, each of which is a probability density function. Often, the distributions are Gaussians, resulting in Gaussian mixture models. They are often used for clustering, but also for density estimation, and they can be optimized using either expectation maximization or Monte Carlo methods. Many astronomical datasets consist of contributions from different populations of objects, which allows mixture modeling to disentangle these population groups. Mixture models based on the EM algorithm have been used in astronomy for this purpose^{67,68}.

Expectation maximization (EM)^{69,70,71} treats the data as a sum of probability distributions, which each represent one cluster. This method alternates between an expectation stage and a maximization stage. In the expectation stage, the algorithm evaluates the membership probability of each data point given the current distribution parameters. In the maximization stage, these probabilities are used to update the parameters. This method works well with missing data, and can be used as the unsupervised component in semi-supervised learning (§2.4.3) to provide class labels for the supervised learning.

The *Kohonen self-organizing map* (SOM)^{72,73} is an unsupervised neural network that forms a general framework for visualizing datasets of more than two dimensions. Unlike many methods which seek to map objects onto a new output space, the SOM is fundamentally topological. This is neatly illustrated by the fact that one astronomical SOM application⁷⁴ is titled ‘Galaxy Morphology Without Classification’. A related earlier method is learning vector quantization⁷⁵.

Independent component analysis (ICA)^{76,77,19,78,79}, an example of *blind source separation*, can separate nonlinear components of a dataset, under the assumption that those components are statistically independent. The components are found by maximizing this independence. Related statistical methods include principal component analysis (§2.3), singular value decomposition, and non-negative matrix factorization.

2.4.3. *Semi-Supervised*

The semi-supervised approach^{80,81} has been somewhat underused to-date, but holds great potential for the upcoming, large, purely photometric surveys. Supervised methods require a labeled training set, but will not assign new classes. On the other hand, unsupervised methods do not require training, but do not use existing

known information. Semi-supervised methods aim to capture the best from both of these methods by retaining the ability to discover new classes within the data, and also incorporating information from a training set when available. An example of a dataset amenable to the approach is shown in Fig. 4.

This is particularly relevant in astronomical applications using large amounts of photometric and a more limited subsample of spectroscopic data, which may not be fully representative of the photometric sample. The semi-supervised approach allows one to use the spectral information to extrapolate into the purely photometric regime, thereby allowing a scientist to utilize all of the vast amount of information present there.

Semi-supervised learning represents an entire subfield of data mining research. Given the nontrivial implementation requirements, this field is a good area for potential fruitful collaborations between astronomers, computer scientists, and statisticians. As one example of a possible issue, a lot of photometric data are likely to be a direct continuation in parameter space of spectroscopic data, with, therefore, a highly overlapping distribution. This means that certain semi-supervised approaches will work better than others, because they contain various assumptions about the nature of the labeled and unlabeled data.

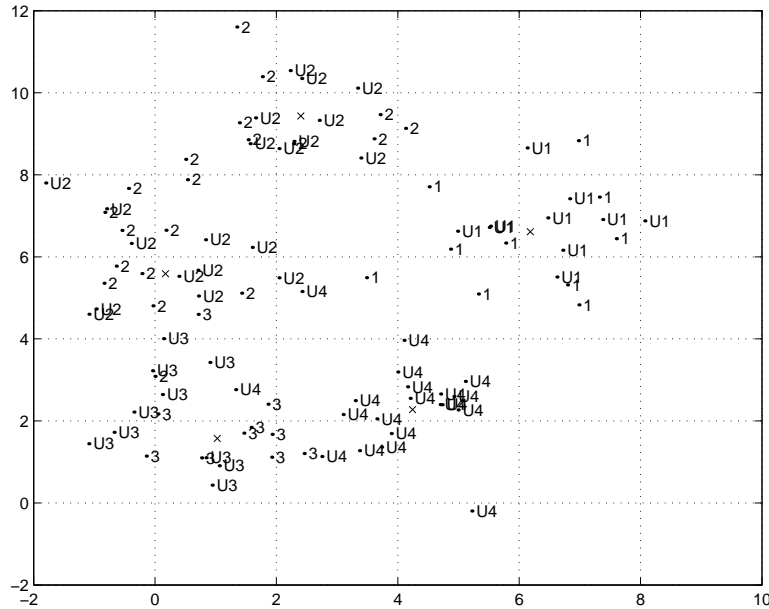


Fig. 4. Dataset amenable to semi-supervised learning, showing labeled and unlabeled classes, denoted by 1–4 and U1–U4, respectively. The axes are arbitrary units. The crosses result from a mixture model applied to the data. From Bazell & Miller⁸².

2.4.4. *Other Algorithms*

In §§2.4.1–2.4.2 above, we described the main data mining algorithms used to date in astronomy, however, there are numerous additional algorithms available, which have often been utilized to some extent. These algorithms may be employed at more than one stage in the process, such as attribute selection, as well as the classification/regression stage.

While neural networks in some very broad sense mimic the learning mechanism of the brain, *genetic algorithms*^{83,84,85,86,87,88} mimic natural selection, as the most successful individuals created are those that are best adapted for the task at hand.

The simplest implementation is the binary genetic algorithm, in which each ‘individual’ is a vector of ones and zeros, which represent whether or not a particular attribute, e.g., a training set attribute, is used. From an initial random population, the individuals are combined to create new individuals. The fitness of each individual is the resulting error in the training algorithm run according to the formula encoded by the individual. This process is repeated until convergence is found, producing the best individual.

A typical method of combining two individuals is one-point crossover, in which segments of two individuals are swapped. To more fully explore the parameter space, and to prevent the algorithm from converging too rapidly on a local minimum, a probability of mutation is introduced into the newly created individuals before they are processed. This is simply the probability that a zero becomes a one, or vice-versa. An approximate number of individuals to use is given by $n_{in} \sim 2n_f \log(n_f)$, where n_f is the number of attributes. The algorithm converges in $n_{it} \sim \alpha n_f \log(n_f)$ iterations, where α is a problem-dependent constant; generally $\alpha > 3$.

Numerous refinements to this basic approach exist, including using continuous values instead of binary ones, and more complex methods for combining individuals. Further possibilities for the design of genetic algorithms exist⁸⁹, and it is possible in principle to combine the optimization of the learning algorithm and the attribute set.

The *Information bottleneck* method¹⁷ is based directly on information theory and is designed to achieve the best tradeoff between accuracy and compression for the desired number of classes. The inputs and outputs are probability density functions. *Association rule* mining^{90,91} is a method of finding qualitative rules within a database such that a rule derived from the occurrence of certain variables together implies something about the occurrence of a variable not used in creating that rule. The *false discovery rate*⁹² is a method of establishing a significant discovery from a smaller set of data than the usual n sigma hypothesis test.

This list could continue, broadening into traditional statistical methods such as least squares, and regression, as well as Bayesian methods, which are widely used in astronomy. For brevity we do not consider these additional methods, but we do note that *graphical models*⁶ are a general way of describing the interrelationships between variables and probabilities, and many of the data mining algorithms

described earlier, such as ANNs, are special cases of these models.

2.4.5. *Choice of Algorithm*

Unfortunately, there is no simple method to select the optimal algorithm to use, because the most appropriate algorithm can depend not only on the dataset, but also the application for which it will be employed. There is, therefore, no single best algorithm. Likewise, the choice of software is similarly non-trivial. Many general frameworks exist, for example WEKA⁵ or Data to Knowledge⁹³, but it is unlikely that one framework will be able to perform all steps necessary from raw catalog to desired science result, particularly for large datasets. In Table 1, we summarize some of the advantages and disadvantages of some of the more popular and well-known algorithms used in astronomy. We do not attempt to summarize available software. Various other general comparisons of machine learning algorithms exist⁷, as well as numerous studies comparing various algorithms for particular datasets, a field which itself is rather complex⁹⁴.

Table 1. Advantages and disadvantages of well-known machine learning algorithms in astronomy. These algorithms, and others, are described in more detail in §§2.4.1–2.4.4.

Algorithm	Advantages	Disadvantages
Artificial Neural Network	<ul style="list-style-type: none"> Good approximation of nonlinear functions Easily parallelized Good predictive power Extensively used in astronomy Robust to irrelevant or redundant attributes 	<ul style="list-style-type: none"> Black-box model Local minima Many adjustable parameters Affected by noise Can overfit Long training time No missing values
Decision Tree	<ul style="list-style-type: none"> Popular real-world data mining algorithm Can input and output numerical or categorical variables Interpretable model Robust to outliers, noisy or redundant attributes Good computational scalability 	<ul style="list-style-type: none"> Can generate large trees that require pruning Generally poorer predictive power than ANN, SVM or kNN Can overfit Many adjustable parameters
Support Vector Machine	<ul style="list-style-type: none"> Copes with noise Gives expected error rate Good predictive power Popular algorithm in astronomy Can approximate nonlinear functions Good scalability with number of attributes Unique solution (no local minima) 	<ul style="list-style-type: none"> Harder to classify > 2 classes No model is created Long training time Poor interpretability Poor at handling irrelevant attributes Can overfit Some adjustable parameters
Nearest Neighbor	<ul style="list-style-type: none"> Uses all available information Does not require training Easily parallelized Few or no adjustable parameters Good predictive power 	<ul style="list-style-type: none"> Computationally intensive No model is created Can be affected by noise and irrelevant attributes
Expectation Maximization	<ul style="list-style-type: none"> Gives number of clusters in the data Fast convergence Copes with missing data Can give class labels for semi-supervised learning 	<ul style="list-style-type: none"> Can be biased toward Gaussians Local minima

2.5. Improving Results

Many of the algorithms previously described involve ‘greedy’ optimization. In these cases, the cost function, which is the measure of how well the algorithm is performing in its classification or prediction task, is minimized in a way that does not allow the value of the function to increase much if at all. As a result, it is possible for the optimization to become trapped in a local minimum, whereby nearby configurations are worse, but better configurations exist in a different region of parameter space. Various approaches exist to overcome local minima. One approach is to simply run the algorithm several times from different starting points. Another approach is *simulated annealing*^{95,96,97,98}, where, in following the metallurgical metaphor, the point in parameter space ‘heats up’, thus perturbing it and allowing it to escape from the local minimum. The point is allowed to ‘cool’, thus having the ability to find a solution closer to the global minimum.

Models produced by data mining algorithms are subject to a fundamental limitation common to many systems in which a predictive model is constructed, the *bias-variance tradeoff*. The bias is the accuracy of the model in describing the data, for example, a linear model might have a higher bias than a higher order polynomial. The variance is the accuracy of this model in describing new data. The higher order polynomial might have a lower bias than a linear model, but it might be more strongly affected by variations in the data and thus have a higher variance. The polynomial has *overfit* the data. There is usually an optimal point between minimizing bias and minimizing variance. A typical way to minimize overfitting is to measure the performance of the algorithm on a test set, which is not part of the training set, and adjusting the stopping criterion for training to stop at an appropriate location.

To help prevent overfitting, training can also be *regularized*, in which an extra term is introduced into the cost function to penalize configurations that add complexity, such as large weights in an ANN. This complexity can cause a function to be less smooth, which increases the likelihood of overfitting. As is the case with supervised learning, unsupervised algorithms can also overfit the data, for example, if some kind of smoothing is employed but its scalelength is too small. In this case, the algorithm will ‘fit the noise’ and not the true underlying distribution.

Another common approach to control overfitting and improve confidence in the accuracy of the results is *cross-validation*, where subsets of the data are left out of the training and used for testing. The simplest form is the holdout method, where a single subset of the training data is kept out of the training, and the algorithm error is evaluated by running on this subset. However, this can have a high bias (see bias-variance tradeoff, above) if the training set is small, due to a significant portion of the training information being left out. K -fold cross-validation improves on this by subdividing the data into K samples and training on $K - 1$ samples, or alternatively using K random subsets. Typically, $K = 5$ or $K = 10$, as small K could still have high bias, as in the holdout method, but large K , while being less

biased, can have high variance due to the testing set being small. If K is increased to the size of the dataset, so that each subsample is a single point, the method becomes leave-one-out cross-validation. In all instances, the estimated error is the mean error from those produced by each run in the cross-validation.

Another important refinement to running one algorithm is the ability to use a *committee* of instances of the algorithm, each with different parameters. One can allow these different instantiations to vote on the final prediction, so that the majority or averaged result becomes the final answer. Such an arrangement can often provide a substantial improvement, because it is more likely that the majority will be closer to the correct answer, and that the answer will be less affected by outliers. One such committee arrangement is *bootstrap aggregating*, or *bagging*^{99,5}, where random subsamples with replacement (bootstrap samples) are taken, and the algorithm trained on each. The created algorithms vote on the testing set. Bagging is often applied to decision trees with considerable success, but it can be applied to other algorithms. The combination of bagging and the random selection of a small subset of features for splitting at each node is known as a Random Forest^{TM100}.

*Boosting*⁷ uses the fact that several ‘weak’ instances of an algorithm can be combined to produce a stronger instance. The weak learners are iteratively added and misclassified objects in the data gain higher weight. Thus boosting is not the same as bagging because the data themselves are weighted. Boosted decision trees are a popular approach, and many different boosting algorithms are available.

As well as committees of the same algorithm, it is also possible to combine the results of more than one different algorithm on the same dataset. Such a *mixture of experts* approach often provides an optimal result on real data. The outcome may be decided by voting, or the output of one algorithm can form the input to another, in a chaining approach.

For many astronomical applications, the results are, or would be, significantly improved by utilizing the full probability density function (PDF) for a predicted property, rather than simply its single scalar value. This is because much more information is retained when using the PDF. Potential uses of PDFs are described further in §4.1.

2.6. *Application of Algorithms and Some Limitations*

The purpose of this review is not to uncritically champion certain data mining algorithms, but to instead encourage scientific progress by exploiting the full potential of these algorithms in a considered scientific approach. We therefore end this section by outlining some of the limitations of and issues raised by KDD and the data mining approach to current and future astronomical datasets. Several of these problems might be ameliorated by increased collaboration between astronomers and data mining experts.

- *Extrapolation*: In many astronomical applications, it is common for data with less information content to be available for a greater number of objects over a larger

parameter space. The classic example is in surveys where photometric objects are typically observed several magnitudes fainter than spectroscopic objects. For a supervised learning algorithm, it is usually inappropriate to extrapolate beyond the parameter space for which the training set (e.g., the spectroscopic objects) is representative.

- *Non-intuitive results:* It is very easy to run an implementation of a well-known algorithm and output a result that appears reasonable, but is in fact either statistically invalid or completely wrong. For example, randomly subsampled training and testing sets from a dataset will overlap and produce a model that overfits the data.
- *Measurement error:* Most astronomical data measurements have an associated error, but most data mining algorithms do not take this explicitly into account. For many algorithms, the intrinsic spread in the data corresponding to the target property is the measurement of the error.
- *Adjustable parameters:* Several algorithms have a significant number of adjustable parameters, and the optimal configuration of these parameters is not obvious. This can result in large parameter sweeps that further increase the computational requirement.
- *Scalability:* Many data mining algorithms scale, for n objects, as n^2 , or even worse, making their simple application to large datasets on normal computing hardware intractable. One can often speed up a naïve implementation of an algorithm that must access large numbers of data points and their attributes by storing the data in a hierarchical manner so that not all the data need to be searched. A popular hierarchical structure for accomplishing this task is the kd-tree¹⁰¹. However, the implementation of such trees for large datasets and on parallel machines remains a difficult problem¹⁰².
- *Learning Curve:* Data mining is an entire field of study in its own right, with strong connections to statistics and computing. The avoidance of some of the issues we present, such as the selection of appropriate algorithms, collaboration where needed, and the full exploitation of their potential for science return, require overcoming a substantial learning curve.
- *Large datasets:* Many astronomical datasets are larger than can be held in machine memory. The exploitation of these datasets thus requires more sophisticated database technology than is currently employed by most astronomical projects.
- *“It’s not science”:* The success of an astronomical project is judged by the science results produced. The time invested by an astronomer in becoming an expert in data mining techniques must be balanced against the expected science gain. It is difficult to justify and obtain funding based purely on a methodological approach such as data mining, even if such an approach will demonstrably improve the expected science return.
- *It does not do the science for you:* The algorithms will output patterns, but will not necessarily establish which patterns or relationships are important scientifically, or even which are causal. The truism ‘correlation is not causation’ is apt

here. The successful interpretation of data mining results is up to the scientist.

- *The result can only be as good as the data*: Related to this, the single largest factor in the success of any data mining algorithm is the quality of the input data. If the data are not sufficient for the task, or are poorly collected or incorrectly treated, the result will not be useful.

3. Uses in Astronomy

We now turn to the use of data mining algorithms in astronomical applications, and their track record in addressing some common problems. Whereas in §2, we introduced terms for the astronomer unfamiliar with data mining, here for the non-expert in astronomy we briefly put in context the astronomical problems. However, a full description is beyond the scope of this review. Whereas §2 was subdivided according to data mining algorithms and issues, here the subdivision is in terms of the astrophysics. Throughout this section, we abbreviate data mining algorithms that are either frequently mentioned or have longer names according to the abbreviations introduced in §2: PCA, ANN, DT, SVM, k NN, KDE, EM, SOM, and ICA.

Given that there is no exact definition of what constitutes a data mining tool, it would not be possible to provide a complete overview of their application. This section therefore illustrates the wide variety of actual uses to date, with actual or implied further possibilities. Uses which exist now but will likely gain greater significance in the future, such as the time domain, are largely deferred to §4. Several other overviews of applications of machine learning algorithms in astronomy exist, and contain further examples, including ones for ANN^{103,104,105,106,107}, DT¹⁰⁸, genetic algorithms¹⁰⁹, and stellar classification¹¹⁰.

Most of the applications in this section are made by astronomers utilizing data mining algorithms. However, several projects and studies have also been made by data mining experts utilizing astronomical data, because, along with other fields such as high energy physics and medicine, astronomy has produced many large datasets that are amenable to the approach. Examples of such projects include the Sky Image Cataloging and Analysis System (SKICAT)¹¹¹ for catalog production and analysis of catalogs from digitized sky surveys, in particular the scans of the second Palomar Observatory Sky Survey; the Jet Propulsion Laboratory Adaptive Recognition Tool (JARTool)¹¹², used for recognition of volcanoes in the over 30,000 images of Venus returned by the Magellan mission; the subsequent and more general Diamond Eye¹¹³; and the Lawrence Livermore National Laboratory Sapphire project¹¹⁴. A recent review of data mining from this perspective is given by Kamath in the book *Scientific Data Mining*¹¹⁵. In general, the data miner is likely to employ more appropriate, modern, and sophisticated algorithms than the domain scientist, but will require collaboration with the domain scientist to acquire knowledge as to which aspects of the problem are the most important.

3.1. Object classification

Classification is often an important initial step in the scientific process, as it provides a method for organizing information in a way that can be used to make hypotheses and to compare with models. Two useful concepts in object classification are the *completeness* and the *efficiency*, also known as recall and precision. They are defined in terms of true and false positives (TP and FP) and true and false negatives (TN and FN). The completeness is the fraction of objects that are truly of a given type that are classified as that type:

$$\text{completeness} = \frac{TP}{TP + FN},$$

and the efficiency is the fraction of objects classified as a given type that are truly of that type

$$\text{efficiency} = \frac{TP}{TP + FP}.$$

These two quantities are astrophysically interesting because, while one obviously wants both higher completeness and efficiency, there is generally a tradeoff involved. The importance of each often depends on the application, for example, an investigation of rare objects generally requires high completeness while allowing some contamination (lower efficiency), but statistical clustering of cosmological objects requires high efficiency, even at the expense of completeness.

3.1.1. Star-Galaxy Separation

Due to their small physical size compared to their distance from us, almost all stars are unresolved in photometric datasets, and thus appear as point sources. Galaxies, however, despite being further away, generally subtend a larger angle, and thus appear as extended sources. However, other astrophysical objects such as quasars and supernovae, also appear as point sources. Thus, the separation of photometric catalogs into stars and galaxies, or more generally, stars, galaxies, and other objects, is an important problem. The sheer number of galaxies and stars in typical surveys (of order 10^8 or above) requires that such separation be automated.

This problem is a well studied one and automated approaches were employed even before current data mining algorithms became popular, for example, during digitization by the scanning of photographic plates by machines such as the APM¹¹⁶ and DPOSS¹¹⁷. Several data mining algorithms have been employed, including ANN^{118,119,120,121,122,123,124}, DT^{125,126}, mixture modeling¹²⁷, and SOM¹²⁸, with most algorithms achieving over 95% efficiency. Typically, this is done using a set of measured morphological parameters that are derived from the survey photometry, with perhaps colors or other information, such as the seeing, as a prior. The advantage of this data mining approach is that all such information about each object is easily incorporated. As well as the simple outputs ‘star’ or ‘galaxy’, many

of the refinements described in §2 have improved results, including probabilistic outputs and bagging¹²⁶.

3.1.2. *Galaxy Morphology*

As shown in Fig. 5, galaxies come in a range of different sizes and shapes, or more collectively, morphology. The most well-known system for the morphological classification of galaxies is the Hubble Sequence of elliptical, spiral, barred spiral, and irregular, along with various subclasses^{129,130,131,132,133,134}. This system correlates to many physical properties known to be important in the formation and evolution of galaxies^{135,136}. Other well-known classification systems are the Yerkes system based on concentration index^{137,138,139}, the de Vaucouleurs¹⁴⁰, exponential^{141,142}, and Sérsic index^{143,144} measures of the galaxy light profile, the David Dunlap Observatory (DDO) system^{145,146,147}, and the concentration-asymmetry-clumpiness (CAS) system¹⁴⁸.

Because galaxy morphology is a complex phenomenon that correlates to the underlying physics, but is not unique to any one given process, the Hubble sequence has endured, despite it being rather subjective and based on visible-light morphology originally derived from blue-biased photographic plates. The Hubble sequence has been extended in various ways, and for data mining purposes the T system^{149,150} has been extensively used. This system maps the categorical Hubble types E, S0, Sa, Sb, Sc, Sd, and Irr onto the numerical values -5 to 10.

One can, therefore, train a supervised algorithm to assign T types to images for which measured parameters are available. Such parameters can be purely morphological, or include other information such as color. A series of papers by Lahav and collaborators^{152,153,154,155,104,156} do exactly this, by applying ANNs to predict the T type of galaxies at low redshift, and finding equal accuracy to human experts. ANNs have also been applied to higher redshift data to distinguish between normal and peculiar galaxies¹⁵⁷, and the fundamentally topological and unsupervised SOM ANN has been used to classify galaxies from Hubble Space Telescope images⁷⁴, where the initial distribution of classes is not known. Likewise, ANNs have been used to obtain morphological types from galaxy spectra.¹⁵⁸

Several authors study galaxy morphology at higher redshift by using the Hubble Deep Fields, where the galaxies are generally much more distant, fainter, less evolved, and morphologically peculiar. Three studies^{159,160,161} use ANNs trained on surface brightness and light profiles to classify galaxies as E/S0, Sabc and Sd/Irr. Another application¹⁶² uses Fourier decomposition on galaxy images followed by ANNs to detect bars and assign T types.

Bazell & Aha¹⁶³ uses ensembles of classifiers, including ANN and DT, to reduce the classification error, and Bazell¹⁶⁴ studies the importance of various measured input attributes, finding that no single measured parameter fully reproduces the classifications. Ball *et al.*¹⁶⁵ obtain similar results to Naim *et al.*¹⁵⁵, but updated for the SDSS. Ball *et al.*¹⁶⁶ and Ball, Loveday & Brunner¹⁶⁷ utilize these classifications

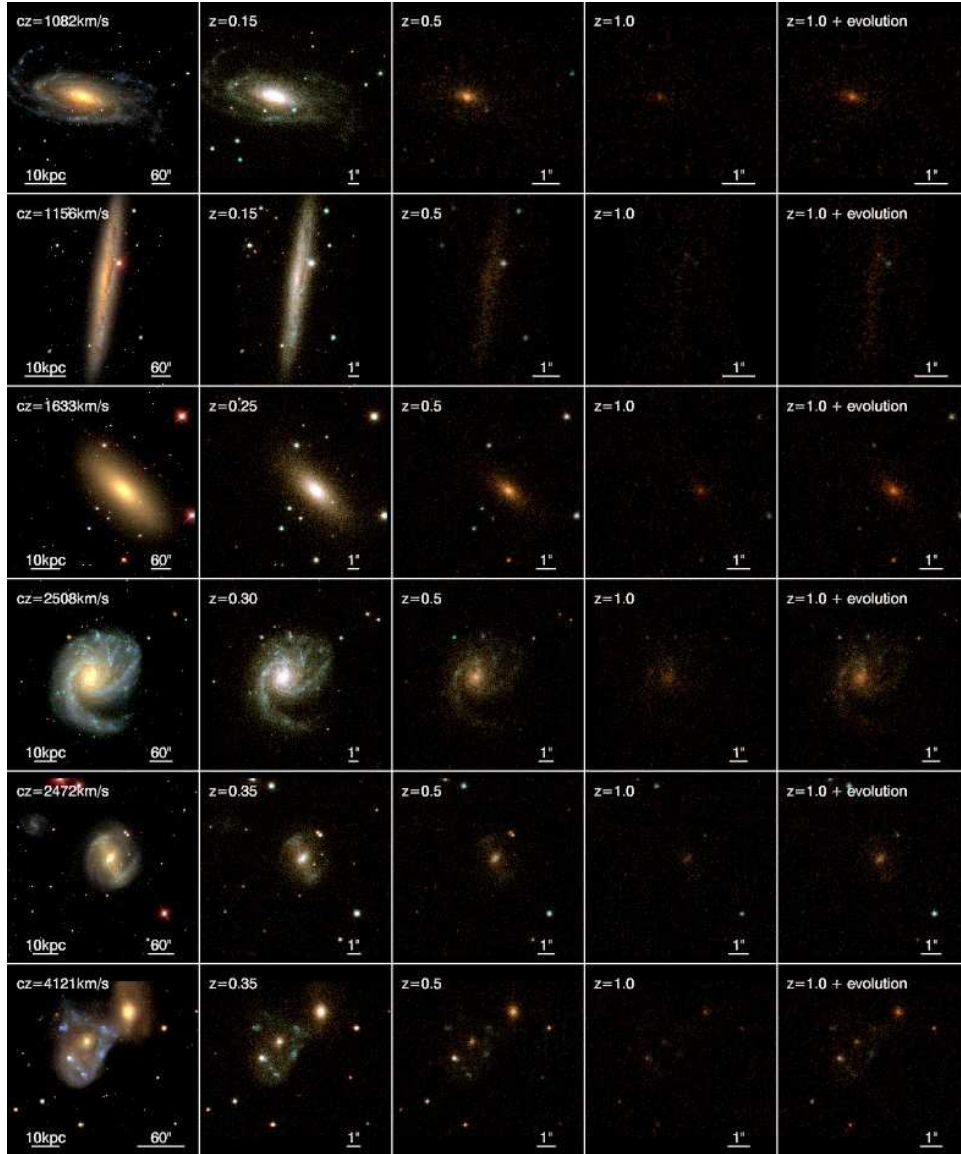


Fig. 5. Examples of galaxy morphology showing many aspects of the information available to, and issues to be aware of for, a data mining process. These include galaxy shape, structure, texture, inclination, arm pitch, color, resolution, exposure, and, from left to right, redshift, in this case artificially constructed. From Barden, Jahnke & Häußler¹⁵¹.

in studies of the bivariate luminosity function and the morphology-density relation in the SDSS, the first such studies to utilize both a digital sky survey of this size and detailed Hubble types.

Because of the complex nature of galaxy morphology and the plethora of available approaches, a large number of further studies exist: Kelly & McKay¹⁶⁸ (Fig. 6) demonstrate improvement over a simple split in $u - r$ using mixture models, within a schema that incorporates morphology. Serra-Ricart *et al.*¹⁶⁹ use an encoder ANN to reduce the dimensionality of various datasets and perform several applications, including morphology. Adams & Woolley¹⁷⁰ use a committee of ANNs in a ‘waterfall’ arrangement, in which the output from one ANN formed the input to another which produces more detailed classes, improving their results. Molinari & Smareglia¹⁷¹ use an SOM to identify E/S0 galaxies in clusters and measure their luminosity function. de Theije & Katgert¹⁷² split E/S0 and spiral galaxies using spectral principal components and study their kinematics in clusters. Genetic algorithms have been employed^{173,174} for attribute selection and to evolve ANNs to classify ‘bent-double’ galaxies in the FIRST¹⁷⁵ radio survey data. Radio morphology combines the compact nucleus of the radio galaxy and extremely long jets. Thus, the bent-double morphology indicates the presence of a galaxy cluster. de la Calleja & Fuentes¹⁷⁶ combine ensembles of ANN and locally weighted regression. Beyond ANN, Spiekermann¹⁷⁷ uses fuzzy algebra and heuristic methods, anticipating the importance of probabilistic studies (§4.1) that are just now beginning to emerge. Owens, Griffiths & Ratnatunga¹⁷⁸ use oblique DTs, obtaining similar results to ANN. Zhang, Li & Zhao¹⁷⁹ distinguish early and late types using k-means clustering. SVMs have recently been employed on the COSMOS survey by Huertas-Company *et al.*^{50,180}, enabling early-late separation to $K_{AB} = 22$ mag twice as good as the CAS system. SVMs will also be used on data from the Gaia satellite¹⁸¹.

Recently, the popular *Galaxy Zoo* project¹⁸² has taken an alternative approach to morphological classification, employing *crowdsourcing*: an application was made available online in which members of the general public were able to view images from the SDSS and assign classifications according to an outlined scheme. The project was very successful, and in a period of six months over 100,000 people provided over 40 million classifications for a sample of 893,212 galaxies, mostly to a limiting depth of $r = 17.77$ mag. The classifications included categories not previously assigned in astronomical data mining studies, such as edge-on or the handedness of spiral arms, and the project has produced multiple scientific results. The approach represents a complementary one to automated algorithms, because, although humans can see things an algorithm will miss and will be subject to different systematic errors, the runtime is hugely longer: a trained ANN will produce the same 40 million classifications in a few minutes, rather than six months.

3.1.3. *Other Galaxy Classifications*

Many of the physical properties, and thus classification, of a galaxy are determined by its stellar population. The spectrum of a galaxy is therefore another method for classification^{183,184}, and can sometimes produce a clearer link to the underlying

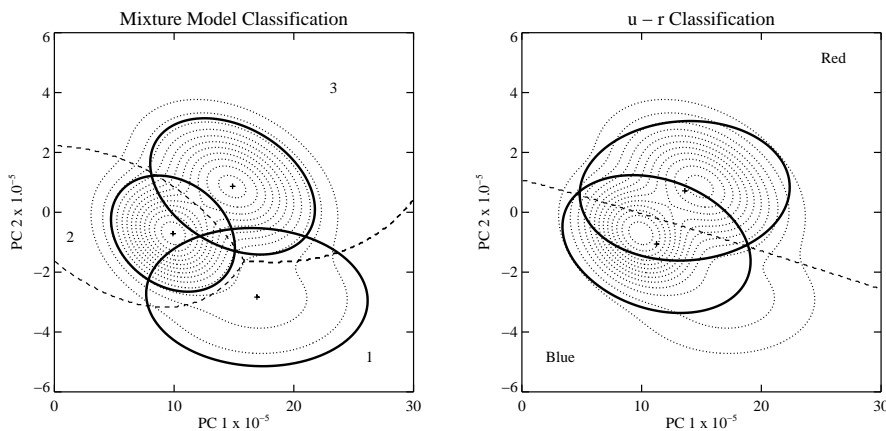


Fig. 6. Improvement in classification using a mixture model over that derived from the u and r passbands ($u - r$ color). In this case, the mixture model clearly delineates the third class, which is not seen using $u - r$. The axes are the first two principle components of the spectro-morphological parameter set (shapelet coefficients in five passbands) describing the galaxies. The light contours are the square root of the probability density from the mixture model fit, and the dark contours are the 95% threshold for each class, in the right-hand panel fitted to the two classes by quadratic discriminant analysis. From Kelly & McKay¹⁶⁸.

physics than the morphology. Spectral classification is important because it is possible for a range of morphological types to have the same spectral type, and vice versa, because spectral types are driven by different underlying physical processes.

Numerous studies^{185,186,187,188} have used PCA directly for spectral classification. PCA is also often used as a preprocessing step before the classification of spectral types using an ANN¹⁸⁹. Folkes, Lahav & Maddox¹⁹⁰ predict morphological types for the 2dF Galaxy Redshift Survey (2dFGRS)¹⁹¹ using spectra, and Ball *et al.*¹⁶⁵ directly predict spectral types in the SDSS using an ANN. Slonim *et al.*¹⁹² use the information bottleneck approach on the 2dFGRS spectra, which maximally preserves the spectral information for the desired number of classes. Lu *et al.*¹⁹³ use ensemble learning for ICA on components of galaxy spectra. Abdalla *et al.*¹⁹⁴ use ANN and locally weighted regression to directly predict emission line properties from photometry.

Bazell & Miller⁸² applied a semi-supervised method suitable for class discovery using ANNs to the ESO-LV¹⁹⁵ and SDSS Early Data Release (EDR) catalogs. They found that a reduction of up to 57% in classification error was possible compared to purely supervised ANNs. The larger of the two catalogs, the SDSS EDR, represents a preliminary dataset about 6% of the final data release of the SDSS, clearly indicating the as-yet untapped potential of this approach. The semi-supervised approach also resembles the hybrid empirical-template approach to photometric redshifts (§3.2), as both seek to utilize an existing training set where available even if it does not

span the whole parameter space. However, the approach used by Bazell & Miller is more general, because it allows new classes of objects to be added, whereas the hybrid approach can only iterate existing templates.

3.1.4. *Quasars/AGN*

Most of the emitted electromagnetic radiation in the universe is either from stars, or the accretion disks surrounding supermassive black holes in active galactic nuclei (AGN). The latter phenomenon is particularly dramatic in the case of quasars, where the light from the central region can outshine the rest of the galaxy. Because supermassive black holes are thought to be fairly ubiquitous in large galaxies, and their fueling, and thus their intrinsic brightness, can be influenced by the environment surrounding the host galaxy, quasars and other AGN are important for understanding the formation and evolution of structure in the universe.

The selection of quasars and other AGN from an astronomical survey is a well-known and important problem, and one well suited to a data mining approach. It is well-known that different wavebands (X-ray, optical, radio) will select different AGN, and that no one waveband can select them all. Traditionally, AGN are classified on the Baldwin-Phillips-Terlevich diagram¹⁹⁶, in which sources are plotted on the two-dimensional space of the emission line ratios $[\text{OIII}] \lambda 5007 / \text{H}\beta$ and $[\text{NII}] / \text{H}\alpha$, that is separated by a single curved line into star-forming and AGN regions. Data mining not only improves on this by allowing a more refined or higher dimensional separation, but also by including passive objects in the same framework (Fig. 7). This allows for the probability that an object contains an AGN to be calculated, and does not require all (or any) of the emission lines to be detected.

Several groups have used ANNs^{197,198,199} or DTs^{200,201,126,202,203,204,205} to select quasar candidates from surveys. White *et al.*²⁰⁰ show that the DT method improves the reliability of the selection to 85% compared to only 60% for simpler criteria. Other algorithms employed include PCA²⁰⁶, SVM and learning vector quantization²⁰⁷, kd-tree²⁰⁸, clustering in the form of principal surfaces and negative entropy clustering²⁰⁹, and kernel density estimation²¹⁰. Many of these papers combine multiwavelength data, particularly X-ray, optical, and radio.

Similarly, one can select and classify candidates of all types of AGN²¹¹. If multiwavelength data are available, the characteristic data mining algorithm ability to form a model of the required complexity to extract the information could enable it to use the full information to extract more complete AGN samples. More generally, one can classify both normal and active galaxies in one system, differentiating between star formation and AGN. As one example, DTs have been used¹²⁶ to select quasar candidates in the SDSS, providing the probabilities $P(\text{star, galaxy, quasar})$. $P(\text{star formation, AGN})$ could be supplied in a similar framework. Bamford *et al.*²¹² combine mixture modeling and regression to perform non-parametric mixture regression, and is the first study to obtain such components and then study them versus environment. The components are passive, star-forming, and two types

of AGN.

3.1.5. Other Classifications

Often, the first component of classification is the actual process of object detection, which often is done at some signal-to-noise threshold. Several statistical data mining algorithms have been employed, and software packages written, for this purpose, including the Faint Object Classification and Analysis System (FOCAS)²¹³, DAOPHOT²¹⁴, Source Extractor (SExtractor)²¹⁵, maximum likelihood, wavelets, ICA²¹⁶, mixture models²¹⁷, and ANNs¹²¹. Serra-Ricart *et al.*²¹⁸ show that ANNs are able to classify faint objects as well as a Bayesian classifier but with considerable computational speedup.

Several studies are more general than star-galaxy separation or galaxy classification, and assign classifications of varying detail to a broad range of astrophysical objects. Goebel *et al.*²¹⁹ apply the AutoClass Bayesian classifier to the IRAS LRS atlas, finding new and scientifically interesting object classes. McGlynn *et al.*²²⁰ use oblique DTs in a system called ClassX to classify X-ray objects into stars, white dwarfs, X-ray binaries, galaxies, AGN, and clusters of galaxies, concluding that the system has the potential to significantly increase the known populations of some rare object types. Suchkov, Hanisch & Margon²⁰¹ use the same system to classify objects in the SDSS. Bazell, Miller & Subbarao²²¹ apply semi-supervised learning to SDSS spectra, including those classified as ‘unknown’, finding two classes of objects consisting of over 50% unknown.

Stellar classifications are necessarily either spectral or based on color, due to the pointlike nature of the source. This field has a long history and well established results such as the HR diagram and the OBAFGKM spectral sequence. The latter is extended to a two-dimensional system of spectral type and luminosity classes I–V to form the two-dimensional MK classification system of Morgan, Keenan & Kellman²²². Class I are supergiants, through to class V, dwarfs, or main-sequence stars. The spectral types correspond to the hottest and most massive stars, O, through to the coolest and least massive, M, and each class is subdivided into ten subclasses 0–9. Thus, the MK classification of the sun is G2V.

The use of automated algorithms to assign MK classes is analogous to that for assigning Hubble types to galaxies in several ways: before automated algorithms, stellar spectra were compared by eye to standard examples; the MK system is closely correlated to the underlying physics, but is ultimately based on observable quantities; the system works quite well but has been extended in numerous ways to incorporate objects that do not fit the main classes (e.g., L and T dwarfs, Wolf-Rayet stars, carbon stars, white dwarfs, and so on). Two differences from galaxy classification are the number of input parameters, in this case spectral indices, and the number of classes. In MK classification the numbers are generally higher, of order 50 or more input parameters, compared to of order 10 for galaxies.

Given a large body of work for galaxies that has involved the use of artificial neu-

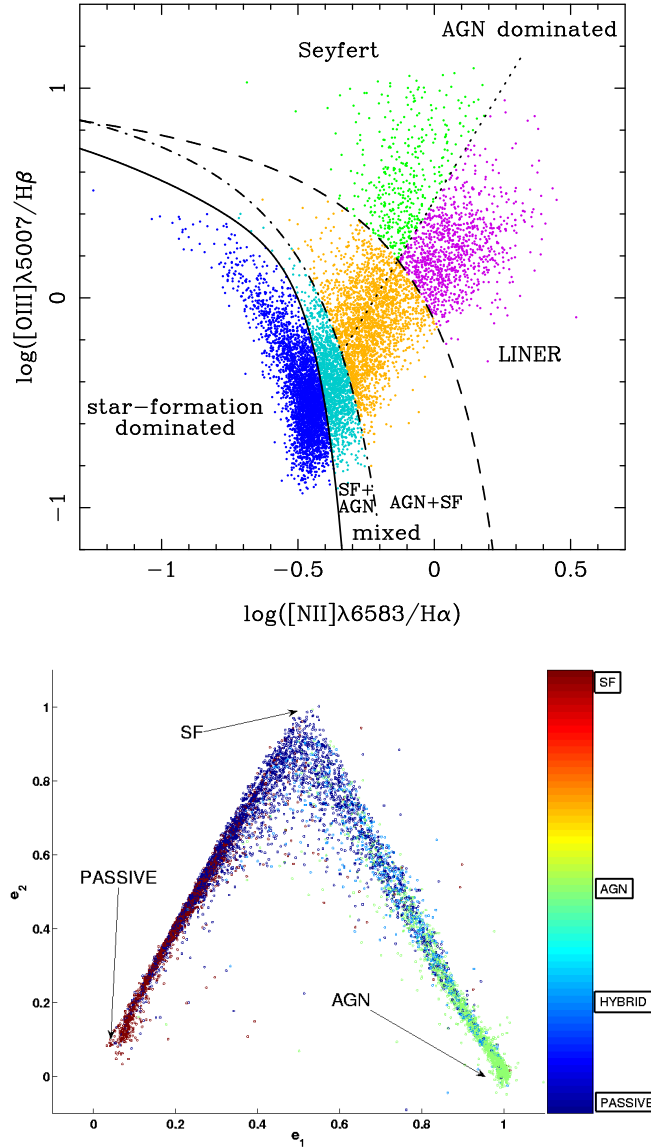


Fig. 7. Upper panel: Baldwin-Philips-Terlevich diagram, which classifies active galactic nuclei (AGN) and star-forming galaxies but requires all four emission lines to be present in the spectrum. From Bamford *et al.*²¹² (although it should be noted that the use of this diagram is not the basis of their study). The axes are the diagnostic emission line ratios from the spectra. Lower panel: AGN/star-forming/passive classification using an ANN, which has no such requirement. The axes are the two outputs from the ANN, e_1 and e_2 mapped onto $(e_1, e_2) = (e_1 + e_2/2)\mathbf{i} + e_2\mathbf{j}$, where passive, AGN, star-forming, and hybrid are $(0,0)$, $(1,0)$, $(0,1)$, and $(0.5,0.5)$, respectively. From Abdalla *et al.*¹⁹⁴.

ral networks, and the similarities just outlined, it is not surprising that similar approaches have been employed for stellar classification^{223,224,225,226,227,228}, with a typical accuracy of one spectral type and half a luminosity type. The relatively large number of object attributes and output classes compared to the number of objects in each class does not invalidate the approach, because the efforts described generally find that the number of principal components represented by the inputs is typically much lower. A well-known property of neural networks is that they are robust to a large number of redundant attributes (§2.4.5).

Neural networks have been used for other stellar classifications schemes, e.g. Gupta *et al.*²²⁹ define 17 classes for IRAS sources, including planetary nebulae and HII regions. Other methods have been employed; a recent example is Manteiga *et al.*²³⁰, who use a fuzzy logic knowledge-based system with a hierarchical tree of decision rules. Beyond the MK and other static classifications, variable stars have been extensively studied for many years, e.g., Wozniak *et al.*²³¹ use SVM to distinguish Mira variables.

The detection and characterization of supernovae is important for both understanding the astrophysics of these events, and their use as standard candles in constraining aspects of cosmology such as the dark energy equation of state. Bailey *et al.*²³² use boosted DTs, random forests, and SVMs to classify supernovae in difference images, finding a ten times reduction in the false-positive rate compared to standard techniques involving parameter thresholds (Fig. 8).

Given the general nature of the data mining approach, there are many further classification examples, including cosmic ray hits^{39,233}, planetary nebulae²³⁴, asteroids²³⁵, and gamma ray sources^{236,237}.

3.2. Photometric redshifts

An area of astrophysics that has greatly increased in popularity in the last few years is the estimation of redshifts from photometric data (photo-zs). This is because, although the distances are less accurate than those obtained with spectra, the sheer number of objects with photometric measurements can often make up for the reduction in individual accuracy by suppressing the statistical noise of an ensemble calculation.

Photo-zs were first demonstrated in the mid 20th century^{238,239}, and later in the 1980s^{240,241}. In the 1990s, the advent of the Hubble Space Telescope Deep fields resulted in numerous approaches^{242,243,244,245,246,247,248}, reviewed by Koo²⁴⁹. In the past decade, the advent of wide-field CCD surveys and multifiber spectroscopy have revolutionized the study of photo-zs to the point where they are indispensable for the upcoming next generation surveys, and a large number of studies have been made.

The two common approaches to photo-zs are the template method and the empirical training set method. The template approach has many complicating issues²⁵⁰, including calibration, zero-points, priors, multiwavelength performance

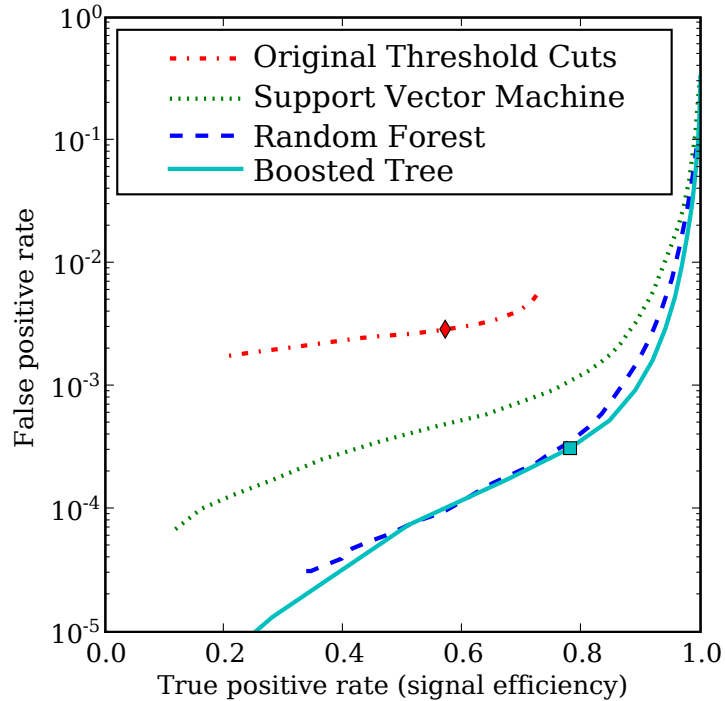


Fig. 8. Improvement in the classification of supernovae using support vector machine and decision tree, compared to previously used threshold cuts. From Bailey *et al.*²³².

(e.g., poor in the mid-infrared), and difficulty handling missing or incomplete training data. We focus in this review on the empirical approach, as it is an implementation of supervised learning. In the future, it is likely that a hybrid method incorporating both templates and the empirical approach will be used, and that the use of full probability density functions will become increasingly important. For many applications, knowing the error distribution in the redshifts is at least as important as the accuracy of the redshifts themselves, further motivating the calculation of PDFs.

3.2.1. *Galaxies*

At low redshifts, the calculation of photometric redshifts for normal galaxies is quite straightforward due to the break in the typical galaxy spectrum at 4000\AA . Thus, as a galaxy is redshifted with increasing distance, the color (measured as a difference in magnitudes) changes relatively smoothly. As a result, both template and empirical photo- z approaches obtain similar results, a root-mean-square deviation of ~ 0.02 in redshift, which is close to the best possible re-

sult given the intrinsic spread in the properties²⁵¹. This has been shown with ANNs^{33,165,156,252,253,254,124,255,256,257,179}, SVM^{258,259}, DT²⁶⁰, k NN²⁶¹, empirical polynomial relations^{262,251,247,263,264,265}, numerous template-based studies, and several other methods. At higher redshifts, obtaining accurate results becomes more difficult because the 4000Å break is shifted redward of the optical, galaxies are fainter and thus spectral data are sparser, and galaxies intrinsically evolve over time. The first explorations at higher redshift were the Hubble Deep Fields in the 1990s, described above (§3.2), and, more recently, new infrared data have become available, which allow the 4000Å break to be seen to higher redshift, which improves the results. Template-based algorithms work well, provided suitable templates into the infrared are available, and supervised algorithms simply incorporate the new data and work in the same manner as previously described.

While supervised learning has been successfully used, beyond the spectral regime the obvious limitation arises that in order to reach the limiting magnitude of the photometric portions of surveys, extrapolation would be required. In this regime, or where only small training sets are available, template-based results can be used, but without spectral information, the templates themselves are being extrapolated. However, the extrapolation of the templates is being done in a more physically motivated manner. It is likely that the more general hybrid approach of using empirical data to iteratively improve the templates,^{266,267,268,269,270,271} or the semi-supervised method described in §2.4.3 will ultimately provide a more elegant solution. Another issue at higher redshift is that the available numbers of objects can become quite small (in the hundreds or fewer), thus reintroducing the curse of dimensionality by a simple lack of objects compared to measured wavebands. The methods of dimension reduction (§2.3) can help to mitigate this effect.

3.2.2. Quasars/AGN

Historically, the calculation of photometric redshifts for quasars and other AGN has been even more difficult than for galaxies, because the spectra are dominated by bright but narrow emission lines, which in broad photometric passbands can dominate the color. The color-redshift relation of quasars is thus subject to several effects, including degeneracy, one emission line appearing like another at a different redshift, an emission line disappearing between survey filters, and reddening. In addition, the filter sets of surveys are generally designed for normal galaxies and not quasars. The assignment of these quasar photo- z s is thus a complex problem that is amenable to data mining in a similar manner to the classification of AGN described in §3.1.4.

The calculation of quasar photo- z s has had some success using SDSS data^{272,273,274,275,276,277}, but they suffer from *catastrophic failures*, in which, as shown in Fig. 9, the photometric redshift for a subset of the objects is completely incorrect. However, data mining approaches have resulted in improvements to this situation. Ball *et al.*²⁷⁸ find that a single-neighbor k NN gives a similar result to the

templates, but multiple neighbors, or other supervised algorithms such as DT or ANN, pull in the regions of catastrophic failure and significantly decrease the spread in the results. Kumar²⁷⁹ also shows this effect. Ball *et al.*²⁶¹ go further and are able to largely eliminate the catastrophics by selecting the subset of quasars with one peak in their redshift probability density function (§4.1), a result confirmed by Wolf²⁸⁰. Wolf *et al.*²⁸¹ also show significant improvement using the COMBO-17 survey, which has 17 filters compared to the five of the SDSS, but unfortunately the photometric sample is much smaller.

Beyond the spectral regime, template-based results are sufficient²⁸², but again suffer from catastrophics. Given our physical understanding of the nature of quasars, it is in fact reasonable to extrapolate in magnitude when using colors as a training set, because while one is going to fainter magnitudes, one is not extrapolating in color. One could therefore quite reasonably assign empirical photo-*z*s for a full photometric sample of quasars.

3.3. Other Astrophysical Applications

Typically in data mining, information gathered from spectra has formed the training set to apply a predictive technique to objects with photometry. However, it is clear from this process that the spectrum itself contains a large amount of information, and data mining techniques may be used directly on the spectra to extract information that might otherwise remain hidden. Applications to galaxy spectral classification were described in §3.1.3. In stellar work, besides the classification of stars into the MK system based on observable parameters, several studies have directly predicted physical parameters of stellar atmospheres using spectral indices. One example is Ramirez, Fuentes & Gulati²⁸³, who utilize a genetic algorithm to select the appropriate input attributes, and predict the parameters using *k*NN. The attribute selection reduces run time and improves predictive accuracy. Solorio *et al.*²⁸⁴ use *k*NN to study stellar populations and improve the results by using active learning to populate sparse regions of parameter space, an alternative to dimension reduction.

Although it has much potential for the future (§4.2), the time domain is a field in which a lot of work has already been done. Examples include the classification of variable stars described in §3.1.5, and, in order of distance, the interaction of the solar wind and the Earth's atmosphere, transient lunar phenomena, detection and classification of asteroids and other solar system objects by composition and orbit, solar system planetary atmospheres, stellar proper motions, extrasolar planets, novae, stellar orbits around the supermassive black hole at the Galactic center, microlensing from massive compact halo objects, supernovae, gamma ray bursts, and quasar variability. A good overview is provided by Becker²⁸⁵. The large potential of the time domain for novel discovery lies within the as yet unexplored parameter space defined by depth, sky coverage, and temporal resolution²⁸⁶. One constraining characteristic of the most variable sources beyond the solar system is that they are

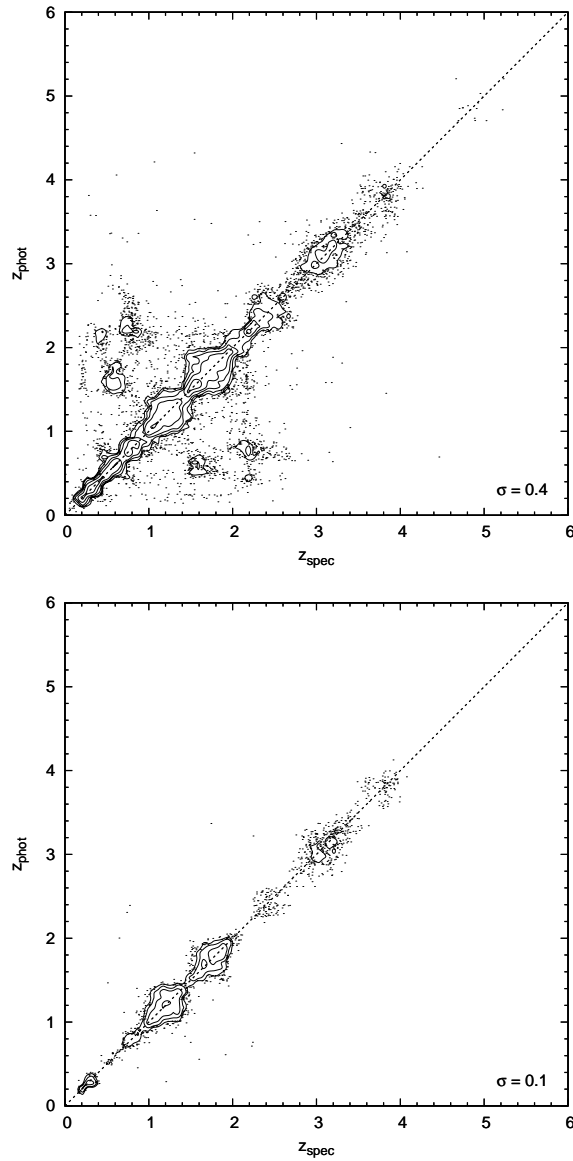


Fig. 9. Photometric redshift, z_p , vs. spectroscopic redshift, z_s , for quasars in the Sloan Digital Sky Survey, showing, in the upper panel, catastrophic failures in which z_p is very different from z_s . Each individual point represents one quasar, and the contours indicate areas of high areal point density. σ is the root-mean-square dispersion between z_p and z_s . The use of data mining techniques, including assigning full probability density functions in photometric redshift, enables the reduction or elimination of these catastrophics, as shown in the lower panel. Data based on that from Ball *et al.*²⁶¹.

generally point sources. As a result, the timescales of interest are constrained by the light crossing time for the source.

The analysis of the cosmic microwave background (CMB) is amenable to several techniques, including Bayesian modeling, wavelets, and ICA. The latter, in particular via the FastICA algorithm²¹⁶, has been used in removal of CMB foregrounds²⁸⁷, and cluster detection via the Sunyaev-Zeldovich effect²⁸⁸. Phillips & Kogut²⁸⁹ use a committee of ANNs for cosmological parameter estimation in CMB datasets, by training them to identify parameter values in Monte Carlo simulations. This gives unbiased parameter estimation in considerably less processing time than maximum likelihood, but with comparable accuracy.

One can use the fact that objects cross-matched between surveys will likely have correlated distributions in their measured attributes, for example, similar position on the sky, to improve cross-matching results using pattern classifiers. Rohde *et al.*²⁹⁰ combine distribution estimates and probabilistic classifiers to produce such an improvement, and supply probabilistic outputs.

Taylor & Diaz²⁹¹ obtain empirical fits for Galactic metallicity using ANNs, whose architectures are evolved using genetic algorithms. This method is able to provide equations for metallicity from line ratios, mitigating the ‘black box’ element common to ANNs, and, in addition, is potentially able to identify new metallicity diagnostics.

Bogdanos & Nesseris²⁹² analyze Type Ia supernovae using genetic algorithms to extract constraints on the dark energy equation of state. This method is non-parametric, which minimizes bias from the necessarily a priori assumptions of parametric models.

Lunar and planetary science, space science, and solar physics also provide many examples of data mining uses. One example is Li *et al.*²⁹³, who demonstrate improvements in solar flare forecasting resulting from the use of a mixture of experts, in this case SVM and k NN. The analysis of the abundance of minerals or constituents in soil samples²⁹⁴ using mixture models is another example of direct data mining of spectra.

Finally, data mining can be performed on astronomical simulations, as well as real datasets. Modern simulations can rival or even exceed real datasets in size and complexity, and as such the data mining approach can be appropriate. An example is the incorporation of theory²⁹⁵ into the Virtual Observatory (§4.5). Mining simulation data will present extra challenges compared to observations because in general there are fewer constraints on the type of data presented, e.g., observations are of the same universe, but simulations are not, simulations can probe many astrophysical processes that are not directly observable, such as stellar interiors, and they provide direct physical quantities as well as observational ones. Most of the largest simulations are cosmological, but they span many areas of astrophysics. A prominent cosmological simulation is the Millennium Run²⁹⁶, and over 200 papers

have utilized its data^c.

4. The Future

We now turn to the future of data mining in astronomy. Several trends are apparent that indicate likely fruitful directions in the next few years. These trends can be used to make informed decisions about upcoming, very large surveys. This section assumes that the reader is somewhat familiar with the concepts in both §§2 and 3, namely, with both data mining and astronomy. We once again arrange the topics by data mining algorithm rather than by astronomical application, but we now interweave the algorithms with examples.

As in the past, it is likely that cross-fertilization with other fields will continue to be beneficial to astronomy, and of particular relevance here, the data mining efforts made by these fields. Examples include high energy physics, whose most obvious spinoff is the World Wide Web from CERN, but the subject has an extensive history of extremely large datasets from experiments such as particle colliders, and has provided well-known and commonly used data analysis software such as ROOT²⁹⁷, designed to cope with these data sizes and first developed in 1994. In the fields of biology and the geosciences, the concepts of *informatics*, the study of computer-based information systems, have been extensively utilized, creating the subfields of bio- and geoinformatics. The official recognition of an analogous subfield within astronomy, *astroinformatics*, has recently been recommended⁸.

4.1. Probability Density Functions

A *probability density function* (PDF, Fig. 10) is a function such that the probability that the value, x , is in the interval $a < x < b$, is the definite integral over the range:

$$P(a < x < b) = \int_a^b f(x)dx.$$

Thus the total area under the function is one. PDFs are of great significance for data mining in astronomy because they retain information that is otherwise lost, and because they enable results with improved signal-to-noise from a given dataset. One can think of a PDF as a histogram in the limit of small bins but many objects. Approaches such as supervised learning are in general taking as input the information on objects and providing as output a prediction of properties. The most general way to do this is to work with the full PDFs at each stage. The formalism has recently been demonstrated in an astronomical context by Budavári²⁷¹, and it is applicable to the prediction of any astronomical property. For inputs a, b, c, \dots , the output probabilities of a set of properties, $P(x, y, z, \dots)$ can be predicted. Fully probabilistic cross-matching of surveys has also been implemented by the same author²⁹⁸.

^c<http://www.mpa-garching.mpg.de/millennium>

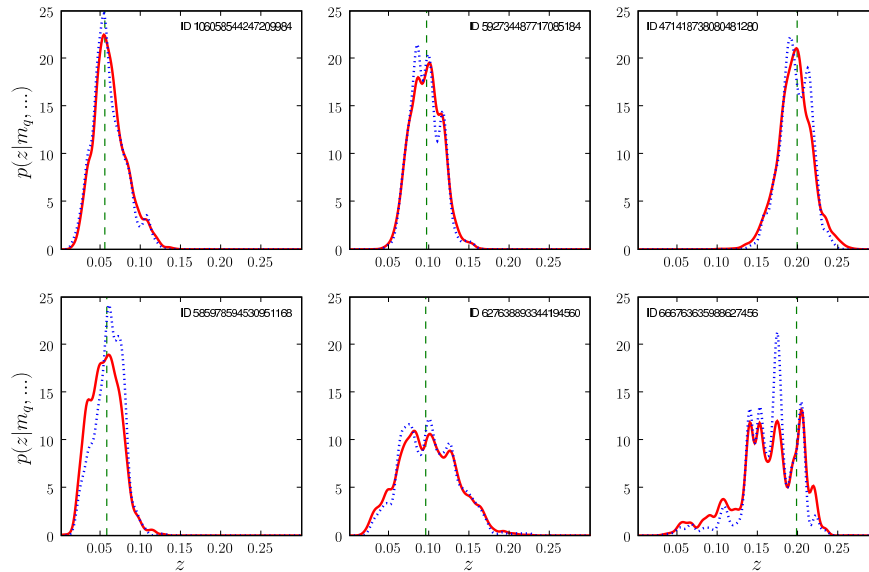


Fig. 10. Example photometric redshift probability density functions (PDFs) for galaxies, showing the rich content of extra information compared to a single value, or value plus Gaussian error. The horizontal axes, z , are the photometric redshifts, and the vertical are the probability densities. The solid red and dotted blue lines are the PDF with and without the photometric uncertainties, respectively, and the vertical dashed green lines are, in these cases, the true distances. From Budavári²⁷¹.

Results with PDFs in photo- z s are starting to appear, either with single values and a spread, or the full PDF. Cunha *et al.*²⁹⁹ show that full PDFs help reduce bias. Margoniner & Wittman³⁰⁰ show that they enable subsamples with improved signal-to-noise, and Wittman³⁰¹ also demonstrates reduction in error. Ball *et al.*²⁶¹ show that generating full photo- z PDFs for quasars allows subsection of a sample virtually free of catastrophic failures, the first time this has been demonstrated, and an important result for their use as tracers of the large scale structure in the universe. Wolf²⁸⁰ confirms a similar result. Myers, White & Ball³⁰² show that using the full PDF for clustering measurements will improve the signal-to-noise by four to five times for a given dataset without any alteration of the data (Fig. 11). This method is applicable to the clustering of any astronomical object. Full PDFs have also been shown to improve performance in the photometric detection of galaxy clusters³⁰³, again due to the increased signal-to-noise ratio. Several further efforts use a single photo- z and a spread, but not the full PDF. However, the method of Myers, White & Ball shows that it is the full PDF that will give the most benefit. PDFs will also be important for weak lensing³⁰⁰.

As well as photo-*z*s, predicting properties naturally incorporates probabilistic classification. Progress has been made, e.g., the SDSS has been classified according to $P(\text{galaxy, star, neither})$ ¹²⁶. Similar classifications that could be made are $P(\text{star formation, AGN})$ and $P(\text{quasar, not quasar})$. Bailer-Jones *et al.*³⁰⁴ implement probabilistic classification that emphasizes finding very rare objects, in this case quasars among the stars that will be seen by Gaia.

Ball *et al.*²⁶¹ generate a PDF by perturbing inputs for a single-neighbor *k*NN. The idea of perturbing data has been studied in the field of Privacy Preserving Data Mining^{305,306}, but here the aim is to generate a PDF using the errors on the input attributes in a way that is computationally scalable to upcoming datasets. The approach appears to work well despite the fact that at present, survey photometric errors are generally poorly characterized³⁰⁷. Proper characterization of errors will be of great importance to future surveys as the probabilistic approach becomes more important. Scalability may be best implemented either by using kd-tree like data structures, or by direct implementation on novel supercomputing hardware such as FPGA, GPU, or Cell processors (§4.7), which can provide enormous performance benefits for applications that require a large number of distance calculations.

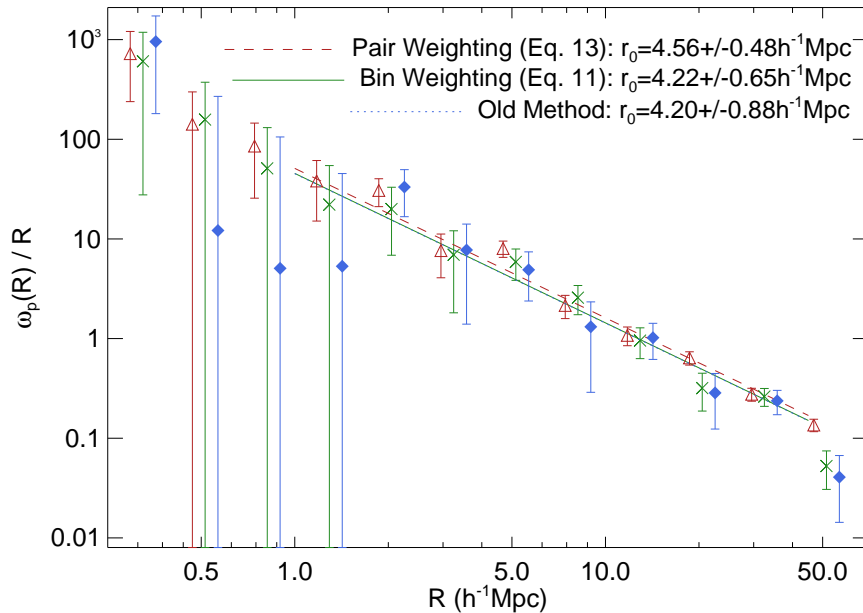


Fig. 11. Improvement in the signal-to-noise ratio of the clustering signal of quasars enabled by PDFs. The improvements to the projected correlation function (vertical axis) enabled by utilizing PDFs are shown by the green crosses and red triangles, compared to the old method, based on single-valued photometric redshifts, shown by blue diamonds. The horizontal axis is the projected radial distance between objects. The diagonal lines are power-law fits, with scale length r_0 , to the correlation function. The points are offset for clarity. From Myers, White & Ball³⁰².

4.2. *Real-Time Processing and the Time Domain*

The time domain is already a significant area of study and will become increasingly important over the next decade with the advent of large scale synoptic surveys such as the Large Synoptic Survey Telescope (LSST)³⁰⁸. A large number of temporal resolved observations over large areas of the sky remains an unexplored area, and the historical precedent suggests that many interesting phenomena remain to be discovered²⁸⁶.

However, as one might expect, this field presents a number of challenges not encountered in the data mining of static objects. These include (i) how to handle multiple observations of objects that can vary in irregular and unpredictable ways, both intrinsic and due to the observational equipment, (ii) objects in difference images (the static background is subtracted, leaving the variation), (iii) the necessarily extremely rapid response to certain events such as gamma ray bursts where physical information can be lost mere seconds after an event becomes detectable, (iv) robust classification of large streams of data in real time, (v) lack of previous information on several phenomena, and (vi) the volume and storage of time domain information in databases. Other challenges are seen in static data, but will assume increased importance as real-time accuracy is needed. For example, the removal of artifacts³⁰⁹ that might otherwise be flagged as unusual objects and incur expensive follow-up telescope time. Variability will be both photometric, a change in brightness, and astrometric, because objects can move. While some astronomical phenomena, such as certain types of variable stars, vary in a regular way, others vary in a nonlinear, irregular, stochastic, or chaotic manner, and the variability itself can change with time (heteroskedasticity)³¹⁰. Time series analysis is a well developed area of statistics, and many of these techniques will be useful.

The combination of available information, but incomplete coverage of the possible phenomena suggests that a probabilistic (§4.1) approach³¹¹, either involving priors, or semi-supervised (§2.4.3) will in general be the most appropriate. This is because the algorithms can use the existing information, but objectively interpret new phenomena. Supervised learning will perform better for problems where more information and larger datasets are available, and unsupervised or Bayesian priors will perform better when there are fewer observations. Many events will still require followup observations, but since there will be far more events than can ever be followed up in detail, data mining algorithms will help ensure that the observations made are optimal in terms of the targeted scientific results.

As a confederation of data archives and interoperable standards of many of the world's existing telescopes, the Virtual Observatory (VO, §4.5) will be crucial in meeting the challenge of the time domain, and significant infrastructure for the VO already exists. The VOEventNet³¹² is a system for the rapid handling of real time events, and provides an online federated data stream of events from several telescopes. It can be followed by both human observers and robotic telescopes.

Numerous next-generation wide-field surveys in the planning or construction

stages will be synoptic. The largest such survey in the optical is the LSST, which will observe the entire sky, visible from its location, every three nights. These observations will provide a data stream exceeding one petabyte per year, and, as a result, they anticipate many of the challenges described here³¹³. Like LSST³¹⁴, the Gaia satellite³¹⁵ has working groups dedicated to data mining. The Classification Working Group has employed several data mining techniques, and developed new approaches^{316,304} to be used when the survey comes online. Other ongoing or upcoming synoptic surveys include Palomar-Quest³¹⁷, the Catalina Real-Time Transient Survey³¹⁸, Pan-STARRS³¹⁹, and those at other wavelengths such as instruments leading up to and including the Square Kilometer Array³²⁰.

The time domain will not only provide challenges to existing methods of data mining, but will open up new avenues for the extraction of information, such as using the variability of objects for classification³²¹ or photometric redshift³²². Because they are due to a relatively compact source in the center of galaxies, active galactic nuclei vary on much shorter timescales than normal galaxies. This variability has been proposed as a mechanism to select quasar and other AGN candidates. Other events are suspected theoretically but have not been observed³²³. But given the dataset sizes, automated detection of such events at some level is clearly required. The computational demands of real time processing of the enormous data streams from these surveys is significant, and will likely be met by the use of newly emerging specialized computing hardware (§4.7).

4.3. Petascale Computing

The current state of the art in supercomputing consists of terabyte-sized files and teraflop computing speeds, which is conveniently encapsulated in the term *terascale computing*. Following Moore's law³²⁴, in which computer performance has increased exponentially for the last several decades, the coming decade will feature the similarly-derived *petascale computing*³²⁵. Much of the performance increase in the past decade has been driven by increases in processor (CPU) clock frequency, but this rate has now slowed due to physical limitations on the sizes of components, and more importantly power consumption and energy (heat) dissipation. It has therefore become more economical to manufacture chips with multiple processor cores.

The typical supercomputer today is a cluster, which consists of a large number of conventional CPUs connected by a specialized interconnect system, a distributed or shared memory, a shared filesystem, and hosting the Linux operating system. Many systems are heterogeneous because this is scalable and cost-effective, but coordinating and making effective such a system can be challenging. In particular, it will be vital that the system is properly balanced between processing power and disk input/output (I/O) to supply the data. Combined with the increasing number of processor cores, this means that *parallel and distributed computing* is rapidly increasing in importance.

A useful set of ‘rules of thumb’ for parallel and other aspects of computing were formulated by Amdahl in the 1960s³²⁶, and they remain true today. One of these is that roughly 50,000 CPU cycles are required per byte of data. Most scientific datasets require far fewer cycles than this, and it is thus likely that future performance will be I/O limited, unless sufficient disks are provided in parallel. Bell, Gray & Szalay¹ estimate that a petascale system will require 100,000 one TB disks. The exact details of how to distribute the data for best performance are likely to be system-dependent⁶⁸. The available CPU speed should scale to the data size, although it will not scale to most naïvely implemented data mining algorithms (§4.4).

An example of an upcoming petascale system whose uses will include astronomical data mining is the *Blue Waters*^d system at the National Center for Supercomputing Applications (NCSA), which is due to come online in 2011. Specifications include 200,000 compute cores with 4 GHz 8 core processors, 1 PB of main memory, 10 PB of user disk storage, 500 PB of archival storage, and 400 GB s⁻¹ bandwidth connectivity to provide sustained petascale compute power. It will implement the IBM PERCS (Productive, Easy-to-use, Reliable Computer System)³²⁷, which will integrate their CPU, operating system, parallel programming, and file systems. This provides a method of addressing the issues of running real-world applications at the petascale by balancing the CPU, I/O, networking, and so on. Similarly, a considerable investment of effort is being carried out in the years leading up to deployment in 2011 on the development of applications for the system, in consultation with the scientists who will run them. Several astronomical applications are included, mostly simulations, but also data mining in the form of the analysis of LSST datasets.

Not all petascale computing will be done on systems as large as Blue Waters. In the US, the National Science Foundation Office of Cyberinfrastructure has been advised¹ to implement a power-law type system, with a small number of very large systems, of order ten times more regional centers, and ten times more local facilities (Tiers 1–3). Such local facilities, for example Beowulf clusters, are already common in university departments, and consist of typically a few dozen commodity machines. A recent trend matching the increasing requirements for data-intensive as opposed to CPU-intensive computing is the GrayWulf cluster³²⁸, which implements the idea of data ‘storage bricks’: cheap, modular, and portable versions of a balanced system which when added together provide petascale computation.

4.4. *Parallel and Distributed Data Mining*

As indicated in §4.3 above, because of the slowing increase in raw speed of individual CPUs, processors are becoming increasingly parallelized, both in terms of the number of processor cores on a single chip, and increasing amounts of these chips being deployed in parallel on supercomputing clusters. Providing appropriately scaled sys-

^d<http://www.ncsa.uiuc.edu/BlueWaters>

tems (CPU, I/O, etc.) is one challenge, but most data mining algorithms not only will be required to run on petascale data, but their naïve implementations scale as N^2 , or worse. It has been suggested³²⁹ that any algorithm that scales beyond $N \log N$ will rapidly be rendered infeasible.

McConnell and Skillicorn³³⁰ have promoted parallel and distributed data mining^{331,332,333,334}, which is well-known in the data mining field, but virtually unused in astronomy. In this approach, the algorithms explicitly take advantage of available parallelism. The simplest example is task-farming, or the embarrassingly parallel approach, in which a task is divided into many mutually-independent sub-tasks, each of which is allocated to a single processor. This can be done on an array of ordinary desktop machines as well as a supercomputer. A more complex challenge is when many parts of the data must be accessed, or when an algorithm relies on the outputs from calculations distributed across multiple compute nodes. For a large dataset the hardware required likely includes shared memory (§4.3), thus shared memory parallelization³³⁵ can be important. Many algorithms exist for the implementation of data mining on parallel computer systems beyond simple task farming, but these are not widely used within science, as compared to the commercial sector. The application programming interfaces MPI and OpenMP have been widely used on distributed and shared memory systems, respectively, for simulation and some data analysis, but they do not offer the semantic capabilities³³⁶ needed for data mining, i.e., the metadata describing the meaning of the data being processed and the results produced are not easily incorporated.

Parallel data mining is challenging, as not only must the algorithm be implemented on the hardware, but many algorithms simply cannot be ported as-is to such a system. Instead, parallelization requires that the algorithm itself, as encapsulated in the code, must often be fundamentally altered at the pseudocode level. This can be a time-consuming and counterintuitive process, especially to scientists who are generally not trained or experienced in parallel programming. Progress is slowly being made in astronomy, including a parallel implementation of kd-trees¹⁰², cosmological simulations requiring datasets larger than the node memory size³³⁷, and parallelization of algorithms³³⁸.

An alternative approach is grid computing, in which the exact resource used is unimportant to the user, although not all data mining algorithms lend themselves to this paradigm. A variant of grid computing is crowdsourcing, in which members of the public volunteer their spare CPU cycles to process data for a project. The most well-known project of this type is SETI@Home, and more recently, the Galaxy Zoo project, which employed large numbers of people to successfully classify galaxies in SDSS images. Such crowdsourcing is likely to become even more important in the future, particularly in combination with greatly improved outreach via astronomical applications on social networking sites such as Facebook³³⁹.

Scalability is also helped on conventional CPUs by the employment of tree structures, such as the kd-tree, which partition the data. This enables a search

to access any data value without searching the whole dataset. Kd-trees have been used for many astronomical applications, including speeding up N-point correlation functions³⁴⁰; cross-matching, classification, and photometric redshifts³⁴¹. They can be extended to more sophisticated structures, for example, the multi-tree³⁴². However, implementation of such tree structures on parallel hardware or computational accelerators (§4.7) remains difficult¹⁰².

4.5. *The Virtual Observatory*

The Virtual Observatory (VO) is an analogous concept to a physical observatory, but instead of telescopes, various centers house data archives. The VO consists of numerous national-level organizations, and the International Virtual Observatory Alliance. Within the national organizations there are various data centers that house large datasets, computing facilities to process and analyze them, and people with considerable expertise in the datasets stored at that particular center.

Common data standards and web services are necessary for the VO to work. Such standards have emerged, including web services using XML and SOAP, a data format, VOTable¹⁰, a query language based on SQL, the Astronomical Data Query Language³⁴³, image access protocols for images (SIAP³⁴³), and spectra (SSAP)^e, VOEventNet³¹² for the time domain, plus various standards of interoperability and ways of describing resources such as the Unified Content Descriptor³⁴⁴. Large numbers of high level tools for working with data are also available^f.

An example of the emerging data standards for archiving is the Common Archive Observation Model³⁴⁵ (CAOM) of the Canadian Astronomical Data Center (CADC). Given that it is likely that the future VO will continue to consist of a number of data centers like the CADC, this model represents a useful and realistic way in which data can be made meaningfully accessible, but not so rigidly presented as to prevent the desired analysis of future researchers with as yet unforeseen science goals. This model consists of the components Artifact, Plane, SimpleObservation, and CompositeObservation, which describe logical parts of the data from individual files to logical sets of observations such as spectra, and forms the basis of all archiving activity at the CADC.

The increasing immobility of large datasets as described in §4.3 will render it uneconomical in terms of time and money to download large datasets to local machines. Rather than bringing the data to the analysis, it will become more sensible to take the analysis to the data³⁴⁶. To be able to perform complicated data mining analyses, it is necessary that the data be organized well enough to make this tractable, and that the center archiving the data must have sufficient computing power and web services to perform the analyses. The organizational requirement means that the data must be stored as a database with the sophistication found in

^e<http://www.ivoa.net/Documents>

^f<http://www.us-vo.org>

the commercial sector, where mining of terascale databases is routine. Commercial software and computer science expertise will help, but the task is non-trivial because astronomical data analysis can require particular data types and structures not usually found in commercial software, such as time series observations. An example of such a database already in place is the SDSS, and its underlying schema³⁴⁷ has been used and copied by other surveys such as GALEX.

Nevertheless, it is likely that considerable analyses will continue to be carried out on smaller subsets of the data, and this data may well continue to be downloaded and analyzed locally, as it has been to date. If one anticipates working exclusively with one survey, it may still be more efficient to implement a GrayWulf-like cluster locally and download the complete dataset.

Another difficult problem faced by the VO is that a significant future scientific benefit from large datasets will be in the cross-matching of multiple datasets, in particular, multiwavelength data. But if such data are distributed among different data centers and are difficult to move, such work may be intractable. What can be done, however, is to make available as part of the VO web services, tools for cross-matching datasets at a given center. A common data format and description, combined with the fact that much of the science is done from small subsets of large datasets, means that this is certainly tractable. As a result, it is not surprising that there is significant demand for such tools³⁴⁸.

An important consideration for the VO is that many astronomers, indeed many scientists in general, will want to run their own software on the data, and not simply a higher level tool that involves trusting someone else's code. This will be true even if the source code is available. Or, a scientist might wish to complete an analysis that is not available in a higher level tool. It is thus important that the data are available at a low level of processing so that one can set one's own requirements as needed. NASA has a categorization of data where 0 is raw, 1 is calibrated, and 2 is a derived product, such as a catalog. An ideal data archive would have available well documented and accessible level 2 catalogs, similarly documented and accessible level 1 data, and perhaps not online but stored level 0 data, to enable, for example, a re-reduction.

Data have been released using the VO publishing interfaces³⁴⁹, data mining algorithms such as ANNs have been implemented³⁵⁰, and applications for analyses with web interfaces are online³⁵¹. Multiwavelength analyses are becoming more feasible and useful³⁴⁸, and it is therefore now possible, but still time-consuming, to perform scientific analyses using VO tools³⁵². We expect this will be an area where considerable work will still need to be done, however, in order to fully enable the full exploitation of the archives of astronomy data in the future.

4.6. Visualization

Visualization of data is an important part of the scientific process, and the combination of terascale computing and data mining poses obvious challenges. Common

plotting codes presently in use in astronomy include SuperMongo^g, PGPlot^h, Gnuplotⁱ, and IDL^j ³⁵³, but these are stand-alone codes that do not easily cope with data that cannot be completely loaded into the available memory space. Newer tools, such as TOPCAT³⁵⁴, VisIVO³⁵⁵, and VOMegaPlot³⁵⁶ support the Virtual Observatory standards such as VOTable and PLASTIC³⁵⁷ for interoperability between programs. The full library on which the TOPCAT program is based, STILTS³⁵⁸, is able to plot arbitrarily-sized datasets.

As with hardware, software, and data analysis, collaboration with computer scientists and other disciplines has resulted in progress in various areas of scientific visualization. At Harvard, the AstroMed project at the Initiative for Innovative Computing (IIC) has collaborated with medical imaging teams³⁵⁹. The rendering of complex multi-dimensional volumetric and surficial data is a common desire of both fields, and the medical imaging software was considerably more advanced than was typical in astronomy in terms of graphical capability. As with the creation and curation of databases for large datasets, collaboration with the IT sector has enabled significant progress and the use of tools beyond the scope of those that could be created by astronomers alone, such as Google Sky³⁶⁰. It is likely that such collaboration will continue to increase in importance.

The program S2Plot³⁶¹, developed at Swinburne, is motivated by the idea of making three-dimensional plots as easy to transfer from one medium to another (interchange) as two-dimensional plots. The existing familiar interface of a plotting code, in this case PGPlot, has been extended³⁶² to enable rendering of multi-dimensional data on several media, including desktop machines, PDF files, Powerpoint-style slides, or web pages. Systems in which the user is able to interact directly with the data are also likely to play a significant role. Partiview³⁶³, developed at NCSA, enables the visualization of particulate data and some isosurfaces either on a desktop or in an immersive CAVE system, and several astronomical datasets have been visualized. Szalay, Springel & Lemson³⁶⁴ describe using graphical processing units (§4.7) to aid visualization, in which the data are preprocessed to hierarchical levels of detail, and only rendered to the resolution required to appear to the eye as if the whole dataset is being rendered. Paraview^k is a program designed for parallel processing on large datasets using distributed memory systems, or on smaller data on a desktop.

Finally, in recent years, numerous online virtual worlds have become popular, the most well-known of which is Second Life. Hut³⁶⁵ and Djorgovski^l describe their interaction within these worlds, both with other astronomers in the form of avatars in meetings, and with datasets. While it may initially seem to be just a

^g<http://www.astro.princeton.edu/~rhl/sm>

^h<http://www.astro.caltech.edu/~tjp/pgplot>

ⁱ<http://www.gnuplot.info>

^j<http://idlastro.gsfc.nasa.gov>

^k<http://www.paraview.org>

^l<http://blogs.discovermagazine.com/cosmicvariance/2008/11/03/guest-post-george-djorgovski-a-new-world-overture>

gimmicky way to have a meeting, the interaction with other avatars is described as ‘fundamentally visceral’, much more so than one would expect. This suggests that, along with social networks for outreach, such interaction among astronomers may become more common, as one will be able to attend a meeting without having to travel physically.

4.7. Novel Supercomputing Hardware

For the final part of §4, we turn to novel supercomputing hardware. This is a rapidly developing area, but it has enormous potential to speed up existing analyses, and render previously impossible questions tractable. Specialized hardware has been used in astronomy for many years, but until recently only in limited contexts and applications, such as the GRAPE³⁶⁶ systems designed specifically for n -body calculations, or direct processing of data in instrument-specific hardware. Here, we describe three hardware formats that have emerged in recent years as viable solutions to a more general range of astronomical problems: graphical processing units (GPUs), field-programmable gate arrays (FPGAs), and the Cell processor.

As described in §4.3, the increasing speed of CPU clock cycles has now been largely replaced by increasing parallelism as the main method for continuing improvements in computing power. The methods described there implement *coarse-grained* parallelism, which is at the level of separate pieces of hardware or application processes. The hardware described here implements *fine-grained* parallelism, in which, at the instruction level, a calculation that would require multiple operations on a CPU is implemented in one operation. The hardware forms an intermediary between the previously-used application-specific integrated circuits (ASIC), and the general purpose CPU.

Future petascale machines (§4.3) are likely to include some or all of these three, either as highly integrated components in a cluster-type system, or as part of the heterogeneous hardware making up a distributed grid-like system that has overall petascale performance.

Spurred by the computer gaming industry, the GPUs on graphics cards within desktop-scale computers have increased in performance much more rapidly than conventional processors (CPUs). They are specially designed to be very fast at carrying out huge numbers of operations that are used in the rendering of graphics, by using vector datatypes and streaming the data. Vector processors have been used before in supercomputing, but GPUs have become of great interest to the scientific community due to their commodity-level pricing, which results from their widespread commercial use, and the increasing ease of use for more general operations than certain graphical processes.

At first, GPUs dealt only with fixed-point numbers, but now single-precision floating point and even double-precision are becoming more common. Thus the chips are no longer simply specialized graphics engines, but are becoming much more general-purpose (GPGPUs). Double-precision is required or highly desir-

able for many scientific applications. The ease of use of GPUs has been increased thanks to NVidia's Compute Unified Device Architecture development environment (CUDA)^m for its cards, and will be further aided by the Open Computing Language (OpenCL)ⁿ for heterogeneous environments. These enable the GPU functions to be called in a similar way to a C library, and are becoming a de facto standard. CUDA has also been ported to other higher level languages, including PyCUDA in Python.

GPUs are beginning to be used in astronomy, and several applications have appeared. GPUs can reproduce the functionality of the GRAPE hardware for n-body simulations³⁶⁷, and CUDA implementations have been shown to outperform GRAPE in some circumstances³⁶⁸. GPUs are beginning to be used for real-time processing of data from next generation instruments³⁶⁹ as part of the Data Intensive Science Consortium at the Harvard IIC. Significant speedup has been demonstrated of a k nearest neighbor search on a GPU compared to a kd-tree implemented in C on a CPU³⁷⁰.

FPGAs^{371,372} are another form of hardware that has become viable for somewhat general-purpose scientific computing. While FPGAs have been widely used as specialized hardware for many years, including in telescopes for data processing or adaptive optics, it is only in the past few years that their speed, cost, capacity, and ease of use have made them viable for more general use by non-specialists. As with GPUs, the ability to work with full double precision floating point numbers is also increasing, and their use is via libraries and development environments that enable the FPGA portion of the code to appear as just another function call in C or a C-like language. These tools implement the hardware description language to program the FPGA, which need not be known by the user.

An FPGA consists of a grid of logic gates which must be programmed via software to implement a specific set of functions before running code (hence field-programmable). If the calculation to be performed can be fully represented in this way on the available gates, this enables a throughput speed of one whole calculation of a function per clock cycle, which given a modern FPGA's clock speed of 100 MHz or more, is 100 million per second. In practice, however, the actual speed is often limited by the I/O.

One recent example is the direct mapping of an ANN onto an FPGA³⁷³, which can then in principle classify one object per clock cycle, or 100 million objects per second at 100 MHz. FPGAs will continue to be widely used as specialized components for astronomical systems, for example in providing real-time processing of the next generation synoptic surveys. Brunner, Kindratenko & Myers³³⁸ demonstrated a significant speedup of the N-point correlation function using FPGAs. Freeman, Weeks & Austin³⁷⁴ directly implement distance calculations, such as required by the k NN data mining algorithm, on an FPGA.

Finally, the IBM Cell processor³⁷⁵ is a chip containing a conventional CPU and

^m<http://www.nvidia.com/cuda>

ⁿ<http://www.khronos.org/openc1>

and array of eight more powerful coprocessors for hardware acceleration in a similar manner to the GPU and FPGA. Like the NVidia GPU, it has been widely used in mass-production machines such as the Playstation 3, and is or will be incorporated into several ‘hybrid’ petascale machines, including IBM’s Roadrunner, and possibly Blue Waters. Unfortunately, also like the GPU, it is not yet as easy to use as desired for large scale scientific use, but progress in the area is continuing.

Further novel supercomputing hardware such as ClearSpeed may become viable for science and widely used. It is an area of exciting developments and considerable potential. As with many new developments, however, one must be somewhat careful, in this case because the continued development of the hardware is driven by large commercial companies (NVidia, IBM, etc.), and not the scientific community. Nevertheless, the potential scientific gains are so large that it is certainly worth keeping an eye on.

5. Conclusions

In this review, we have introduced data mining in astronomy, given an overview of its implementation in the form of knowledge discovery in databases, reviewed its application to various science problems, and discussed its future. Throughout, we have tried to emphasize data mining as a tool to enable improved science, not as an end in itself, and to highlight areas where improvements have been made over previous analyses, where they might yet be made, and limitations of this approach.

An astronomer is not a cutting-edge expert in data mining algorithms any more than they are in statistics, databases, hardware, software, etc., but they will need to know enough to usefully apply such approaches to the science problem they wish to address. It is likely that such progress will be made via collaboration with people who are experts in these areas, particularly within large projects, that will employ specialists and have working groups dedicated to data mining. Fully implemented, commercial-level databases will be required since the data will be too big to organize, download, or analyze in any other way.

The available infrastructure should, therefore, be designed so that this data mining approach to research is maximally enabled. The raw or minimally-processed data should be made available in a manner so one can apply user-specific codes either locally or using computational resources local to the data if data size necessitates it. It is unlikely that most researchers will either require or trust the exact resources made available by higher level tools. Instead, they will be useful for exploratory work, but ultimately one must be able to run personal or trusted code on the data, from the level of re-reduction upwards.

A problem arises when one wishes to utilize multiple or distributed datasets, for example in cross-matching data for multi-wavelength studies. Therefore, datasets that can be easily made interoperable via a standard storage schema should be made available. In this manner, a user can bring computing power and algorithms to tackle their particular science question. This problem is particularly acute when

large datasets are held at widely separated sites, because transfer of such data across the network is currently impractical. A great deal of science is done on small subsets of the full data, so data will still be frequently downloaded and analyzed locally, but the paradigm of downloading entire datasets is not sustainable.

Acknowledgments

We thank the referee for a useful and comprehensive report.

The authors acknowledge support from NASA through grants NN6066H156 and NNG06GF89G, from Microsoft Research, and from the University of Illinois.

The authors made extensive use of the storage and computing facilities at the National Center for Supercomputing Applications and thank the technical staff for their assistance in enabling this work.

This research has made use of the SAO/NASA Astrophysics Data System.

References

1. G. Bell, J. Gray and A. Szalay, *IEEE Computer* **39**, 110 (2006).
2. G. Bell, T. Hey and A. Szalay, *Science* **323**, 1297 (2009).
3. T. Hey, S. Tansley and K. Talle (eds.), *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Microsoft Research, Redmond, WA, 2009).
4. D. J. Hand, *Statistical Science* **21**, p. 1 (2006).
5. I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Series in Data Management Systems, 2nd edn. (Morgan Kaufmann, San Francisco, 2005).
6. C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2007).
7. T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, 2nd edn. (Springer, New York, 2009).
8. K. Borne, *Scientific Data Mining in Astronomy*, Data Mining and Knowledge Discovery Series Data Mining and Knowledge Discovery Series, (Taylor & Francis: CRC Press, Boca Raton, FL, 2009), pp. 91–114.
9. D. C. Wells, E. W. Greisen and R. H. Harten, *A&AS* **44**, p. 363 (1981).
10. F. Ochsenbein *et al.*, VOTable: Tabular Data for the Virtual Observatory, in *Toward an International Virtual Observatory*, eds. P. J. Quinn and K. M. Górski (2004).
11. D. Pyle, *Data Preparation for Data Mining*, Morgan Kaufmann Series in Data Management Systems (Morgan Kaufmann, San Francisco, 1999).
12. D. W. Hogg, preprint, [arXiv/0807.4820] (2008).
13. K. Karhunen, *Annales Academiae Scientiarum Fennicae Series A. I. Mathematica-Physica* **37**, 3 (1947).
14. M. M. Loève, *Fonctions Aléatoires de Second Ordre*, in *Processus Stochastiques et Mouvement Brownien*, ed. P. Levy (Hermann, Paris, 1948), Paris.
15. I. T. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics, 2nd edn. (Springer, New York, 2002).
16. S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman and A. Y. Wu, *Journal of the Association for Computing Machinery* **45**, 891 (1998).
17. N. Tishby, F. C. Pereira and W. Bialek, The Information Bottleneck Method, in *The 37th annual Allerton Conference on Communication, Control, and Computing*, 1999.

18. R. A. Fisher, *Annals of Eugenics* **7**, 179 (1936).
19. A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis* (John Wiley & Sons, New York, 2001).
20. S. J. Lilly *et al.*, *ApJS* **172**, 70 (2007).
21. D. G. York *et al.*, *AJ* **120**, 1579 (2000).
22. W. Jeffrey and R. Rosner, *ApJ* **310**, 473 (1986).
23. C. M. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, Oxford, 1995).
24. B. D. Ripley, *Pattern Recognition and Neural Networks* (Cambridge University Press, Cambridge, UK, 2008).
25. R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, 2nd edn. (Cambridge University Press, Cambridge, UK, 2000).
26. W. S. McCulloch and W. H. Pitts, *Bulletin of Mathematical Biophysics* **5**, 115 (1943).
27. J. J. Hopfield and D. W. Tank, *Science* **233**, 625 (1986).
28. P. J. Werbos, Beyond regression: new tools for prediction and analysis in the behavioural sciences, PhD thesis, Harvard, (Cambridge, MA, 1974).
29. D. B. Parker, *Learning Logic*, Tech. Rep. TR-47, Center for Computational Research in Economics and Management Science, MIT (Cambridge, MA, 1985).
30. D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Nature* **323**, 533 (1986).
31. K. Levenberg, *Quarterly of Applied Mathematics* **2**, p. 164 (1944).
32. D. W. Marquardt, *Journal of the Society of Industrial and Applied Mathematics* **2**, p. 431 (1963).
33. A. E. Firth, O. Lahav and R. S. Somerville, *MNRAS* **339**, 1195 (2003).
34. J. N. Morgan and J. A. Sonquist, *Journal of the American Statistical Association* **58**, 415 (1963).
35. L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees* (Wadsworth, 1984).
36. J. R. Quinlan, *Machine Learning* **1**, p. 81 (1986).
37. J. R. Quinlan, *C4.5: Programs for Machine Learning* (Morgan Kaufmann, San Francisco, 1993).
38. L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theory and Applications* (World Scientific, New York, 2008).
39. S. Salzberg, R. Chandar, H. Ford, S. K. Murthy and R. White, *PASP* **107**, 279 (1995).
40. C. Cortes and V. Vapnik, *Machine Learning* **20**, 273 (1995).
41. C. J. C. Burges, *Knowledge Discovery and Data Mining* **2**, 121 (1998).
42. V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd edn. (Springer, New York, 1999).
43. N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* (Cambridge University Press, 2000).
44. V. Kecman, *Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models* (MIT Press, Cambridge, MA, 2001).
45. B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (MIT Press, Cambridge, MA, 2001).
46. S. Abe, *Support Vector Machines for Pattern Classification* (Springer, New York, 2005).
47. L. Wang, *Support Vector Machines: Theory and Applications* (Springer, New York, 2005).
48. I. Steinwart and A. Christmann, *Support Vector Machines* (Springer, New York, 2008).

49. M. A. Aizerman, E. M. Braverman and L. I. Rozonoer, *Automation and Remote Control* **25**, 1175 (1964).
50. M. Huertas-Company, D. Rouan, L. Tasca, G. Soucail and O. Le Fèvre, *A&A* **478**, 971 (2008).
51. E. Fix and J. Hodges Jr., *Discriminatory analysis: non-parametric discrimination: Consistency properties.*, Tech. Rep. Report No. 4, USAF School of Aviation Medicine (Randolph Field, TX, 1951).
52. T. M. Cover and P. E. Hart, *IEEE Transactions on Information Theory* **13**, 21 (1967).
53. D. W. Aha, D. Kibler and M. K. Albert, *Machine Learning* **6**, 37 (1991).
54. B. Dasarathy, *Nearest Neighbor Pattern Classification Techniques* (IEEE Computer Society Press, New York, 1991).
55. G. Shakhnarovich, T. Darrell and P. Indyk (eds.), *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice* (MIT Press, Cambridge, MA, 2006).
56. E. Parzen, *Annals of Mathematical Statistics* **33**, 1065 (1962).
57. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (John Wiley, New York, 1973).
58. B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Monographs on Statistics and Applied Probability (CRC Press, Boca Raton, FL, 1986).
59. D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley Series in Probability and Statistics (Wiley-Interscience, New York, 1992).
60. C. Taylor, *Vistas in Astronomy* **41**, 411 (1997).
61. L. Wasserman, *All of Statistics: a Concise Course in Statistics* (Springer, New York, 2005).
62. J. Klemelä, *Smoothing of Multivariate Data: Density Estimation and Visualization*, Wiley Series in Probability and Statistics (John Wiley & Sons, New York, 2009).
63. H. Steinhaus, *Bull. Acad. Polon. Sci.* **4**, 801 (1956).
64. J. MacQueen, Some Methods for Classification and Analysis of Multivariate Observations, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, eds. L. M. LeCam and J. Neyman (University of California Press, Berkeley, 1967).
65. D. M. Titterton, A. F. M. Smith and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions* (John Wiley, New York, 1985).
66. G. J. McLachlan and D. Peel, *Finite Mixture Models*, Wiley Series in Probability and Statistics (Wiley-Interscience, New York, 2000).
67. A. J. Connolly, C. Genovese, A. W. Moore, R. C. Nichol, J. Schneider and L. Wasserman, preprint, [arXiv:astro-ph/0008187] (2000).
68. J. Dolence and R. J. Brunner, Fast Two-Point Correlations of Extremely Large Data Sets, *The 9th LCI International Conference on High-Performance Clustered Computing*, Urbana-Champaign, IL, (2008).
69. A. Dempster, N. Laird and D. Rubin, *Journal of the Royal Statistical Society B* **39**, 1 (1977).
70. M. Watanabe and K. Yamaguchi (eds.), *The EM Algorithm and Related Statistical Models*, Statistics: a Series of Textbooks and Monographs (CRC Press, Boca Raton, FL, 2003).
71. G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, Wiley Series in Probability and Statistics (John Wiley & Sons, New York, 2008).
72. T. Kohonen, *Biological Cybernetics* **43**, 59 (1982).
73. T. Kohonen, *Self-Organizing Maps, 3rd extended edition*, Springer Series in Information Sciences, Vol. 30, 3rd edn. (Springer, Berlin, 2001).

74. A. Naim, K. U. Ratnatunga and R. E. Griffiths, *ApJS* **111**, p. 357 (1997).
75. T. Kohonen, *Self-Organization and Associative Memory*, 3rd edn. (Springer-Verlag, Berlin, 1989).
76. P. Comon, *Signal Processing* **36**, 287 (1994).
77. T. Lee, *Independent Component Analysis - Theory and Applications* (Kluwer Academic Publishers, New York, 1998).
78. S. Roberts and R. Everson (eds.), *Independent Component Analysis: Principles and Practice* (Cambridge University Press, Cambridge, UK, 2001).
79. J. V. Stone, *Independent Component Analysis: A Tutorial Introduction* (MIT Press, Cambridge, MA, 2004).
80. O. Chapelle, B. Schölkopf and A. Zien (eds.), *Semi-Supervised Learning* (MIT Press, Cambridge, MA, 2006).
81. X. Zhu, A. Goldberg, R. Brachman and T. Dietterich, *Introduction to Semi-supervised Learning*, Synthesis Lectures on Artificial Intelligence and Machine Learning (Morgan & Claypool, San Rafael, CA, 2009).
82. D. Bazell and D. J. Miller, *ApJ* **618**, 723 (2005).
83. J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence* (The University of Michigan Press, Ann Arbor, MI, 1975).
84. D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, Reading, MA, 1989).
85. D. A. Coley, *An Introduction to Genetic Algorithms for Scientists and Engineers* (World Scientific, New York, 1997).
86. M. Mitchell, *An Introduction to Genetic Algorithms* (MIT Press, Cambridge, MA, 1998).
87. R. L. Haupt and S. E. Haupt, *Practical Genetic Algorithms*, 2nd edn. (Wiley Inter-Science, New York, 2004).
88. S. N. Sivanandam and S. N. Deepa, *Introduction to Genetic Algorithms* (Springer, New York, 2007).
89. Goldberg, D. E., *Design of innovation: Lessons from and for competent genetic algorithms* (Kluwer Academic Publishers, Boston, MA, 2002).
90. J. M. Adamo, *Data Mining for Association Rules and Sequential Patterns: Sequential and Parallel Algorithms* (Springer, New York, 2000).
91. C. Zhang and S. Zhang, *Association Rule Mining: Models and Algorithms*, Lecture Notes in Computer Science (Springer, New York, 2002).
92. Y. Benjamini and Y. Hochberg, *Journal of the Royal Statistical Society B* **57**, 289 (1995).
93. M. Welge, W. H. Hsu, L. S. Auvil, T. M. Redman and D. Tchong, High-Performance Knowledge Discovery and Data Mining Systems Using Workstation Clusters, in *12th National Conference on High Performance Networking and Computing (SC99)*, 1999.
94. S. L. Salzberg, *Data Mining and Knowledge Discovery* **1**, 1 (1995).
95. S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi, *Science* **220**, 671 (1983).
96. V. Černý, *Journal of Optimization Theory and Applications* **45**, 41 (1985).
97. P. J. van Laarhoven and E. H. Aarts, *Simulated Annealing: Theory and Applications* (Springer, New York, 1987).
98. E. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing* (Wiley, New York, 1989).
99. L. Breiman, *Machine Learning* **26**, 123 (1996).
100. L. Breiman, *Machine Learning* **45**, 5 (2001).

101. J. L. Bentley, *Communications of the ACM* **18**, 509 (1975).
102. J. P. Gardner, A. Connolly and C. McBride, Enabling rapid development of parallel tree search applications, in *CLADE '07: Proceedings of the 5th IEEE workshop on Challenges of large applications in distributed environments*, (ACM, New York, 2007).
103. A. S. Miller, *Vistas in Astronomy* **36**, 141 (1993).
104. O. Lahav, A. Naim, L. Sodré and M. C. Storrie-Lombardi, *MNRAS* **283**, p. 207 (1996).
105. C. A. L. Bailer-Jones, R. Gupta and H. P. Singh, An Introduction to Artificial Neural Networks, in *Automated Data Analysis in Astronomy*, eds. R. Gupta, H. P. Singh and C. A. L. Bailer-Jones (2002).
106. L.-L. Li, Y.-X. Zhang, Y.-H. Zhao and D.-W. Yang, *Progress in Astronomy* **24**, 285 (2006).
107. R. Tagliaferri *et al.*, *Neural Networks*, **16**, 297 (2003).
108. R. L. White, *Astronomical Applications of Oblique Decision Trees*, American Institute of Physics Conference Series Vol. 1082 (2008).
109. P. Charbonneau, *ApJS* **101**, p. 309 (1995).
110. C. A. L. Bailer-Jones, Automated Stellar Classification for Large Surveys: A Review of Methods and Results, in *Automated Data Analysis in Astronomy*, eds. R. Gupta, H. P. Singh and C. A. L. Bailer-Jones (2002).
111. N. Weir, U. M. Fayyad, S. G. Djorgovski and J. Roden, *PASP* **107**, p. 1243 (1995).
112. M. C. Burl, L. Asker, P. Smyth, U. Fayyad, P. Perona, J. Aubele and L. Crumpler, *Machine Learning* **30**, 165 (1998).
113. M. C. Burl, C. Fowlkes, J. Roden, A. Stechert and S. Mukhtar, *Diamond Eye: a distributed architecture for image data mining*, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 3695 (1999).
114. C. Kamath, *Journal of Physics Conference Series* **125**, 012094 (2008).
115. C. Kamath, *Scientific data mining: a practical perspective* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 2009).
116. S. J. Maddox, G. Efstathiou, W. J. Sutherland and J. Loveday, *MNRAS* **243**, 692 (1990).
117. S. G. Djorgovski, R. R. Gal, S. C. Odewahn, R. R. de Carvalho, R. Brunner, G. Longo and R. Scaramella, *The Palomar Digital Sky Survey (DPOSS)*, in *Wide Field Surveys in Cosmology*, eds. S. Colombi, Y. Mellier and B. Raban (1998).
118. S. C. Odewahn, E. B. Stockwell, R. L. Pennington, R. M. Humphreys and W. A. Zmach, *AJ* **103**, 318 (1992).
119. S. C. Odewahn and M. L. Nielsen, *Vistas in Astronomy* **38**, 281 (1994).
120. D. Bazell and Y. Peng, *ApJS* **116**, p. 47 (1998).
121. S. Andreon, G. Gargiulo, G. Longo, R. Tagliaferri and N. Capuano, *MNRAS* **319**, 700 (2000).
122. N. S. Philip, Y. Wadadekar, A. Kembhavi and K. B. Joseph, *A&A* **385**, 1119 (2002).
123. S. C. Odewahn, R. R. de Carvalho, R. R. Gal, S. G. Djorgovski, R. Brunner, A. Mahabal, P. A. A. Lopes, J. L. K. Moreira and B. Stalder, *AJ* **128**, 3092 (2004).
124. A. Collister *et al.*, *MNRAS* **375**, 68 (2007).
125. N. Weir, U. M. Fayyad and S. Djorgovski, *AJ* **109**, p. 2401 (1995).
126. N. M. Ball, R. J. Brunner, A. D. Myers and D. Tchong, *ApJ* **650**, p. 497 (2006).
127. D.-M. Qin, P. Guo, Z.-Y. Hu and Y.-H. Zhao, *Chinese Journal of Astronomy and Astrophysics* **3**, 277 (2003).
128. A. S. Miller and M. J. Coe, *MNRAS* **279**, 293 (1996).
129. E. P. Hubble, *ApJ* **64**, 321 (1926).
130. E. P. Hubble, *Realm of the Nebulae* (Yale University Press, Newhaven, CT, 1936).

131. A. Sandage, *The Hubble atlas of galaxies* (Carnegie Institution, Washington, DC, 1961).
132. A. Sandage and J. Bedke, *The Carnegie atlas of galaxies* (Carnegie Institution of Washington with The Flintridge Foundation, Washington, DC, 1994).
133. S. van den Bergh, *Galaxy morphology and classification* (Cambridge University Press, Cambridge, UK, 1998).
134. A. Sandage, *ARA&A* **43**, 581 (2005).
135. M. S. Roberts and M. P. Haynes, *ARA&A* **32**, 115 (1994).
136. C. Firmani and V. Avila-Reese, Physical processes behind the morphological Hubble sequence, *Revista Mexicana de Astronomia y Astrofisica Conference Series Vol. 17* (2003).
137. W. W. Morgan, *PASP* **70**, p. 364 (1958).
138. W. W. Morgan, *PASP* **71**, p. 394 (1959).
139. G. de Vaucouleurs, Qualitative and Quantitative Classifications of Galaxies., in *Evolution of Galaxies and Stellar Populations*, eds. B. M. Tinsley and R. B. Larson (1977).
140. G. de Vaucouleurs, *Annales d'Astrophysique* **11**, p. 247 (1948).
141. F. S. Patterson, *Harvard College Observatory Bulletin* **914**, 9 (1940).
142. K. C. Freeman, *ApJ* **160**, p. 811 (1970).
143. J. L. Sérsic, *Atlas de galaxies australes* (Observatorio Astronomico, Cordoba, Argentina, 1968).
144. A. W. Graham and S. P. Driver, *PASA* **22**, 118 (2005).
145. S. van den Bergh, *ApJ* **131**, p. 215 (1960).
146. S. van den Bergh, *ApJ* **131**, p. 558 (1960).
147. S. van den Bergh, *ApJ* **206**, 883 (1976).
148. C. J. Conselice, *ApJS* **147**, 1 (2003).
149. G. de Vaucouleurs, *Memoirs of the Commonwealth Observatory, Mount Stromlo* **3** (1956).
150. G. de Vaucouleurs, *Handbuch der Physik* **53**, p. 275 (1959).
151. M. Barden, K. Jahnke and B. Häußler, *ApJS* **175**, 105 (2008).
152. M. C. Storrie-Lombardi, O. Lahav, L. Sodré and L. J. Storrie-Lombardi, *MNRAS* **259**, p. 8P (1992).
153. O. Lahav *et al.*, *Science* **267**, 859 (1995).
154. A. Naim *et al.*, *MNRAS* **274**, 1107 (1995).
155. A. Naim, O. Lahav, L. Sodré and M. C. Storrie-Lombardi, *MNRAS* **275**, 567 (1995).
156. A. A. Collister and O. Lahav, *PASP* **116**, 345 (2004).
157. A. Naim, K. U. Ratnatunga and R. E. Griffiths, *ApJ* **476**, p. 510 (1997).
158. D. S. Madgwick, *MNRAS* **338**, 197 (2003).
159. S. C. Odewahn, R. A. Windhorst, S. P. Driver and W. C. Keel, *ApJ* **472**, p. L13 (1996).
160. R. Windhorst, S. Odewahn, C. Burg, S. Cohen and I. Waddington, *Ap&SS* **269**, 243 (1999).
161. S. H. Cohen, R. A. Windhorst, S. C. Odewahn, C. A. Chiarenza and S. P. Driver, *AJ* **125**, 1762 (2003).
162. S. C. Odewahn, S. H. Cohen, R. A. Windhorst and N. S. Philip, *ApJ* **568**, 539 (2002).
163. D. Bazell and D. W. Aha, *ApJ* **548**, 219 (2001).
164. D. Bazell, *MNRAS* **316**, 519 (2000).
165. N. M. Ball, J. Loveday, M. Fukugita, O. Nakamura, S. Okamura, J. Brinkmann and R. J. Brunner, *MNRAS* **348**, 1038 (2004).
166. N. M. Ball, J. Loveday, R. J. Brunner, I. K. Baldry and J. Brinkmann, *MNRAS* **373**,

- 845 (2006).
167. N. M. Ball, J. Loveday and R. J. Brunner, *MNRAS* **383**, 907 (2008).
 168. B. C. Kelly and T. A. McKay, *AJ* **129**, 1287 (2005).
 169. M. Serra-Ricart, X. Calbet, L. Garrido and V. Gaitan, *AJ* **106**, 1685 (1993).
 170. A. Adams and A. Woolley, *Vistas in Astronomy* **38**, 273 (1994).
 171. E. Molinari and R. Smareglia, *A&A* **330**, 447 (1998).
 172. P. A. M. de Theije and P. Katgert, *A&A* **341**, 371 (1999).
 173. E. Cantú-Paz and C. Kamath, Evolving Neural Networks For The Classification Of Galaxies, in *GECCO '02: Proceedings of the Genetic and Evolutionary Computation Conference*, (Morgan Kaufmann Publishers Inc., San Francisco, 2002).
 174. C. Kamath, E. Cantú-Paz, I. K. Fodor and N. I. Tang, *Computing in Science and Engineering* **4**, 52 (2002).
 175. R. H. Becker, R. L. White and D. J. Helfand, *ApJ* **450**, p. 559 (1995).
 176. J. de la Calleja and O. Fuentes, *MNRAS* **349**, 87 (2004).
 177. G. Spieckermann, *AJ* **103**, 2102 (1992).
 178. E. A. Owens, R. E. Griffiths and K. U. Ratnatunga, *MNRAS* **281**, 153 (1996).
 179. Y. Zhang, L. Li and Y. Zhao, *MNRAS* **392**, 233 (2009).
 180. M. Huertas-Company *et al.*, *A&A* **497**, 743 (2009).
 181. P. Tsalmantza *et al.*, *A&A* **470**, 761 (2007).
 182. C. J. Lintott *et al.*, *MNRAS* **389**, 1179 (2008).
 183. M. L. Humason, *ApJ* **83**, p. 10 (1936).
 184. W. W. Morgan and N. U. Mayall, *PASP* **69**, p. 291 (1957).
 185. A. J. Connolly, A. S. Szalay, M. A. Bershad, A. L. Kinney and D. Calzetti, *AJ* **110**, p. 1071 (1995).
 186. A. J. Connolly and A. S. Szalay, *AJ* **117**, 2052 (1999).
 187. D. Madgwick, O. Lahav, K. Taylor and The 2dFGRS Team, Parameterisation of Galaxy Spectra in the 2dF Galaxy Redshift Survey, in *Mining the Sky*, eds. A. J. Banday, S. Zaroubi and M. Bartelmann (2001).
 188. C. W. Yip *et al.*, *AJ* **128**, 585 (2004).
 189. M. C. Storrie-Lombardi, M. J. Irwin, T. von Hippel and L. J. Storrie-Lombardi, *Vistas in Astronomy* **38**, 331 (1994).
 190. S. R. Folkes, O. Lahav and S. J. Maddox, *MNRAS* **283**, 651 (1996).
 191. M. Colless *et al.*, preprint, [arXiv:astro-ph/0306581] (2003).
 192. N. Slonim, R. Somerville, N. Tishby and O. Lahav, *MNRAS* **323**, 270 (2001).
 193. H. Lu, H. Zhou, J. Wang, T. Wang, X. Dong, Z. Zhuang and C. Li, *AJ* **131**, 790 (2006).
 194. F. B. Abdalla, A. Mateus, W. A. Santos, L. Sodrè, Jr., I. Ferreras and O. Lahav, *MNRAS* **387**, 945 (2008).
 195. A. Lauberts and E. A. Valentijn, *The surface photometry catalogue of the ESO-Uppsala galaxies* (European Southern Observatory, Garching, Germany, 1989).
 196. J. A. Baldwin, M. M. Phillips and R. Terlevich, *PASP* **93**, 5 (1981).
 197. R. Carballo, A. S. Cofiño and J. I. González-Serrano, *MNRAS* **353**, 211 (2004).
 198. J.-F. Claeskens, A. Smette, L. Vandenbulcke and J. Surdej, *MNRAS* **367**, 879 (2006).
 199. R. Carballo, J. I. González-Serrano, C. R. Benn and F. Jiménez-Luján, *MNRAS* **391**, 369 (2008).
 200. R. L. White *et al.*, *ApJS* **126**, 133 (2000).
 201. A. A. Suchkov, R. J. Hanisch and B. Margon, *AJ* **130**, 2439 (2005).
 202. Y.-X. Zhang and Y.-H. Zhao, *Chinese Journal of Astronomy and Astrophysics* **7**, 289 (2007).
 203. Y. Zhang, Y. Zhao and D. Gao, *Advances in Space Research* **41**, 1949 (2008).

204. Y. Zhao and Y. Zhang, *Advances in Space Research* **41**, 1955 (2008).
205. C. Knigge, S. Scaringi, M. R. Goad and C. E. Cottis, *MNRAS* **386**, 1426 (2008).
206. C. W. Yip *et al.*, *AJ* **128**, 2603 (2004).
207. Y. Zhang and Y. Zhao, *PASP* **115**, 1006 (2003).
208. D. Gao, Y.-X. Zhang and Y.-H. Zhao, *MNRAS* **386**, 1417 (2008).
209. R. D'Abrusco, G. Longo and N. A. Walton, *MNRAS* **396**, 223 (2009).
210. G. T. Richards *et al.*, *ApJS* **180**, 67 (2009).
211. M.-F. Zhao, C. Wu, A. Luo, F.-C. Wu and Y.-H. Zhao, *Chinese Astronomy and Astrophysics* **31**, 352 (2007).
212. S. P. Bamford, A. L. Rojas, R. C. Nichol, C. J. Miller, L. Wasserman, C. R. Genovese and P. E. Freeman, *MNRAS* **391**, 607 (2008).
213. J. F. Jarvis and J. A. Tyson, *AJ* **86**, 476 (1981).
214. P. B. Stetson, *PASP* **99**, 191 (1987).
215. E. Bertin and S. Arnouts, *A&AS* **117**, 393 (1996).
216. D. Maino *et al.*, *MNRAS* **334**, 53 (2002).
217. F. Guglielmetti, R. Fischer and V. Dose, *MNRAS* **396**, 165 (2009).
218. M. Serra-Ricart, V. Gaitan, L. Garrido and I. Perez-Fournon, *A&AS* **115**, p. 195 (1996).
219. J. Goebel, J. Stutz, K. Volk, H. Walker, F. Gerbault, M. Self, W. Taylor and P. Cheeseman, *A&A* **222**, L5 (1989).
220. T. A. McGlynn *et al.*, *ApJ* **616**, 1284 (2004).
221. D. Bazell, D. J. Miller and M. SubbaRao, *ApJ* **649**, 678 (2006).
222. W. W. Morgan, P. C. Keenan and E. Kellman, *An atlas of stellar spectra, with an outline of spectral classification* (The University of Chicago press, 1943).
223. T. von Hippel, L. J. Storrie-Lombardi, M. C. Storrie-Lombardi and M. J. Irwin, *MNRAS* **269**, p. 97 (1994).
224. W. B. Weaver and A. V. Torres-Dodgen, *ApJ* **487**, p. 847 (1997).
225. H. P. Singh, R. K. Gulati and R. Gupta, *MNRAS* **295**, 312 (1998).
226. C. A. L. Bailer-Jones, M. Irwin and T. von Hippel, *MNRAS* **298**, 361 (1998).
227. R. K. Gulati and L. Altamirano, *Ap&SS* **273**, 73 (2000).
228. M. Bazarghan and R. Gupta, *Ap&SS* **315**, 201 (2008).
229. R. Gupta, H. P. Singh, K. Volk and S. Kwok, *ApJS* **152**, 201 (2004).
230. M. Manteiga, I. Carricajo, A. Rodriguez, C. Dafonte and B. Arcay, *AJ* **137**, 3245 (2009).
231. P. R. Woźniak, S. J. Williams, W. T. Vestrand and V. Gupta, *AJ* **128**, 2965 (2004).
232. S. Bailey, C. Aragon, R. Romano, R. C. Thomas, B. A. Weaver and D. Wong, *ApJ* **665**, 1246 (2007).
233. W. Waniak, *Experimental Astronomy* **21**, 151 (2006).
234. M. Faundez-Abans, M. I. Ormeno and M. de Oliveira-Abans, *A&AS* **116**, 395 (1996).
235. A. Misra and S. J. Bus, Artificial Neural Network Classification of Asteroids in the Sloan Digital Sky Survey, in *AAS/Division for Planetary Sciences Meeting Abstracts*, 2008.
236. T. Chattopadhyay, R. Misra, A. K. Chattopadhyay and M. Naskar, *ApJ* **667**, 1017 (2007).
237. S. Scaringi, A. J. Bird, D. J. Clark, A. J. Dean, A. B. Hill, V. A. McBride and S. E. Shaw, *MNRAS* **390**, 1339 (2008).
238. J. Stebbins and A. E. Whitford, *ApJ* **108**, p. 413 (1948).
239. W. A. Baum, Photoelectric Magnitudes and Red-Shifts, in *IAU Symp. 15: Problems of Extra-Galactic Research*, 1962.
240. D. C. Koo, *AJ* **90**, 418 (1985).

241. E. D. Loh and E. J. Spillar, *ApJ* **303**, 154 (1986).
242. S. D. J. Gwyn and F. D. A. Hartwick, *ApJ* **468**, p. L77 (1996).
243. K. M. Lanzetta, A. Yahil and A. Fernandez-Soto, *Nature* **381**, 759 (1996).
244. B. Mobasher, M. Rowan-Robinson, A. Georgakakis and N. Eaton, *MNRAS* **282**, L7 (1996).
245. M. J. Sawicki, H. Lin and H. K. C. Yee, *AJ* **113**, 1 (1997).
246. A. J. Connolly, A. S. Szalay and R. J. Brunner, *ApJ* **499**, p. L125 (1998).
247. Y. Wang, N. Bahcall and E. L. Turner, *AJ* **116**, 2081 (1998).
248. N. Benítez, *ApJ* **536**, 571 (2000).
249. D. C. Koo, Overview - Photometric Redshifts: A Perspective from an Old-Timer[!] on their Past, Present, and Potential, in *Photometric Redshifts and the Detection of High Redshift Galaxies*, eds. R. Weymann, L. Storrie-Lombardi, M. Sawicki and R. Brunner, Astronomical Society of the Pacific Conference Series, Vol. 191 (1999).
250. M. Massarotti, A. Iovino and A. Buzzoni, *A&A* **368**, 74 (2001).
251. R. J. Brunner, A. J. Connolly, A. S. Szalay and M. A. Bershad, *ApJ* **482**, p. L21 (1997).
252. E. Vanzella *et al.*, *A&A* **423**, 761 (2004).
253. L.-L. Li, Y.-X. Zhang, Y.-H. Zhao and D.-W. Yang, *Chinese Journal of Astronomy and Astrophysics* **7**, 448 (2007).
254. R. D'Abrusco, A. Staiano, G. Longo, M. Brescia, M. Paolillo, E. De Filippis and R. Tagliaferri, *ApJ* **663**, 752 (2007).
255. M. Banerji, F. B. Abdalla, O. Lahav and H. Lin, *MNRAS* **386**, 1219 (2008).
256. H. Oyaizu, M. Lima, C. E. Cunha, H. Lin, J. Frieman and E. S. Sheldon, *ApJ* **674**, 768 (2008).
257. M. D. Niemack, R. Jimenez, L. Verde, F. Menanteau, B. Panter and D. Spergel, *ApJ* **690**, 89 (2009).
258. Y. Wadadekar, *PASP* **117**, 79 (2005).
259. D. Wang, Y.-X. Zhang, C. Liu and Y.-H. Zhao, *Chinese Journal of Astronomy and Astrophysics* **8**, 119 (2008).
260. S. Carliles, T. Budavári, S. Heinis, C. Priebe and A. Szalay, Photometric Redshift Estimation on SDSS Data Using Random Forests, in *Astronomical Data Analysis Software and Systems XVII*, eds. R. W. Argyle, P. S. Bunclark and J. R. Lewis, Astronomical Society of the Pacific Conference Series, Vol. 394 (2008).
261. N. M. Ball, R. J. Brunner, A. D. Myers, N. E. Strand, S. L. Alberts and D. Tcheng, *ApJ* **683**, 12 (2008).
262. A. J. Connolly, I. Csabai, A. S. Szalay, D. C. Koo, R. G. Kron and J. A. Munn, *AJ* **110**, p. 2655 (1995).
263. D. Sowards-Emmerd, J. A. Smith, T. A. McKay, E. Sheldon, D. L. Tucker and F. J. Castander, *AJ* **119**, 2598 (2000).
264. B. C. Hsieh, H. K. C. Yee, H. Lin and M. D. Gladders, *ApJS* **158**, 161 (2005).
265. P. A. A. Lopes, *MNRAS* **380**, 1608 (2007).
266. T. Budavári, A. S. Szalay, A. J. Connolly, I. Csabai and M. Dickinson, *AJ* **120**, 1588 (2000).
267. I. Csabai, A. J. Connolly, A. S. Szalay and T. Budavári, *AJ* **119**, 69 (2000).
268. I. Csabai *et al.*, *AJ* **125**, 580 (2003).
269. N. Padmanabhan *et al.*, *MNRAS* **359**, 237 (2005).
270. M. Brodwin *et al.*, *ApJ* **651**, 791 (2006).
271. T. Budavári, *ApJ* **695**, 747 (2009).
272. T. Budavári *et al.*, *AJ* **122**, 1163 (2001).
273. G. T. Richards *et al.*, *AJ* **122**, 1151 (2001).

274. T. S. R. Babbedge *et al.*, *MNRAS* **353**, 654 (2004).
275. M. A. Weinstein *et al.*, *ApJS* **155**, 243 (2004).
276. X.-B. Wu, W. Zhang and X. Zhou, *Chinese Journal of Astronomy and Astrophysics* **4**, 17 (2004).
277. S. Kitsionas, E. Hatziminaoglou, A. Georgakakis and I. Georgantopoulos, *A&A* **434**, 475 (2005).
278. N. M. Ball, R. J. Brunner, A. D. Myers, N. E. Strand, S. Alberts, D. Tchenguiz and X. Llorà, *ApJ* **663**, p. 774 (2007).
279. N. D. Kumar, Machine learning techniques for astrophysical modelling and photometric redshift estimation of quasars in optical sky surveys, Master's thesis, Oxford University (2008).
280. C. Wolf, *MNRAS* **397**, 520 (2009).
281. C. Wolf, L. Wisotzki, A. Borch, S. Dye, M. Kleinheinrich and K. Meisenheimer, *A&A* **408**, 499 (2003).
282. M. Salvato *et al.*, *ApJ* **690**, 1250 (2009).
283. J. F. Ramírez, O. Fuentes and R. K. Gulati, *Experimental Astronomy* **12**, 163 (2001).
284. T. Solorio, O. Fuentes, R. Terlevich and E. Terlevich, *MNRAS* **363**, 543 (2005).
285. A. C. Becker, *Astronomische Nachrichten* **329**, p. 280 (2008).
286. S. G. Djorgovski, A. A. Mahabal, R. J. Brunner, R. R. Gal, S. Castro, R. R. de Carvalho and S. C. Odewahn, Searches for Rare and New Types of Objects, in *Virtual Observatories of the Future*, eds. R. J. Brunner, S. G. Djorgovski and A. S. Szalay, Astronomical Society of the Pacific Conference Series, Vol. 225 (2001).
287. M. Bottino, A. J. Banday and D. Maino, *MNRAS* **389**, 1190 (2008).
288. S. Pires, J. B. Juin, D. Yvon, Y. Moudou, S. Anthoine and E. Pierpaoli, *A&A* **455**, 741 (2006).
289. N. G. Phillips and A. Kogut, preprint, [arXiv:astro-ph/0108234] (2001).
290. D. J. Rohde, M. R. Gallagher, M. J. Drinkwater and K. A. Pimbblet, *MNRAS* **369**, 2 (2006).
291. M. Taylor and A. I. Diaz, On the Deduction of Galactic Abundances with Evolutionary Neural Networks, in *From Stars to Galaxies: Building the Pieces to Build Up the Universe*, eds. A. Vallenari, R. Tantalò, L. Portinari and A. Moretti, Astronomical Society of the Pacific Conference Series, Vol. 374 (2007).
292. C. Bogdanos and S. Nesseris, *Journal of Cosmology and Astro-Particle Physics* **5**, p. 6 (2009).
293. R. Li, Y. Cui, H. He and H. Wang, *Advances in Space Research* **42**, 1469 (2008).
294. J. F. Mustard, L. Li and G. He, *J. Geophys. Res.* **103**, 19419 (1998).
295. G. Lemson and J. Zuther, *Memorie della Societa Astronomica Italiana* **80**, 342 (2009).
296. V. Springel *et al.*, *Nature* **435**, 629 (2005).
297. R. Brun and F. Rademakers, *Nuclear Instruments and Methods in Physics Research A* **389**, 81 (1997).
298. T. Budavári and A. S. Szalay, *ApJ* **679**, 301 (2008).
299. C. E. Cunha, M. Lima, H. Oyaizu, J. Frieman and H. Lin, *MNRAS* **396**, 2379 (2009).
300. V. E. Margoniner and D. M. Wittman, *ApJ* **679**, 31 (2008).
301. D. Wittman, *ApJ* **700**, L174 (2009).
302. A. D. Myers, M. White and N. M. Ball, *MNRAS* **399**, 2279 (2009).
303. C. van Breukelen and L. Clewley, *MNRAS* **395**, 1845 (2009).
304. C. A. L. Bailer-Jones, K. W. Smith, C. Tiede, R. Sordo and A. Vallenari, *MNRAS* **391**, 1838 (2008).
305. J. Vaidya, C. Clifton and M. Zhu., *Privacy Preserving Data Mining* (Springer, New

- York, 2005).
306. C. C. Aggarwal and P. S. Yu (eds.), *Privacy-Preserving Data Mining: Models and Algorithms* (Springer, New York, 2008).
 307. R. Scranton, A. J. Connolly, A. S. Szalay, R. H. Lupton, D. Johnston, T. Budavári, J. Brinkmann and M. Fukugita, preprint, [arXiv:astro-ph/0508564] (2005).
 308. Ž. Ivezić, J. A. Tyson, R. Allsman, J. Andrew, R. Angel and for the LSST Collaboration, preprint, [arXiv/0805.2366] (2008).
 309. C. Donalek, A. Mahabal, S. G. Djorgovski, S. Marney, A. Drake, E. Glikman, M. J. Graham and R. Williams, New Approaches to Object Classification in Synoptic Sky Surveys, American Institute of Physics Conference Series Vol. 1082 (2008).
 310. A. H. Studenmund, *Using Econometrics*, 2nd edn. (Addison-Wesley, New York, 2005).
 311. A. Mahabal, S. G. Djorgovski, M. Turmon, J. Jewell, R. R. Williams, A. J. Drake, M. G. Graham, C. Donalek, E. Glikman and Palomar-QUEST team, *Astronomische Nachrichten* **329**, 288 (2008).
 312. A. J. Drake, R. Williams, M. J. Graham, A. Mahabal, S. G. Djorgovski, R. R. White, W. T. Vestrand and J. Bloom, VOEventNet: An Open Source of Transient Alerts for Astronomers., *Bulletin of the American Astronomical Society* Vol. 38 (2007).
 313. Ž. Ivezić *et al.*, Parametrization and Classification of 20 Billion LSST Objects: Lessons from SDSS, American Institute of Physics Conference Series Vol. 1082 (2008).
 314. K. Borne, J. Becla, I. Davidson, A. Szalay and J. A. Tyson, The LSST Data Mining Research Agenda, American Institute of Physics Conference Series Vol. 1082 (2008).
 315. M. A. C. Perryman *et al.*, *A&A* **369**, 339 (2001).
 316. C. A. L. Bailer-Jones, A Method for Exploiting Domain Information in Astrophysical Parameter Estimation, in *Astronomical Data Analysis Software and Systems XVII*, eds. R. W. Argyle, P. S. Bunclark and J. R. Lewis, Astronomical Society of the Pacific Conference Series, Vol. 394 (2008).
 317. S. G. Djorgovski *et al.*, *Astronomische Nachrichten* **329**, p. 263 (2008).
 318. A. J. Drake *et al.*, *ApJ* **696**, 870 (2009).
 319. K. W. Hodapp *et al.*, *Astronomische Nachrichten* **325**, 636 (2004).
 320. S. Johnston *et al.*, *Publications of the Astronomical Society of Australia* **24**, 174 (2007).
 321. L. Eyer *et al.*, Variability type classification of multi-epoch surveys, American Institute of Physics Conference Series Vol. 1082 (2008).
 322. M. C. Kaczmarczik, G. T. Richards, S. S. Mehta and D. J. Schlegel, *AJ* **138**, 19 (2009).
 323. A. Mahabal *et al.*, Towards Real-time Classification of Astronomical Transients, American Institute of Physics Conference Series Vol. 1082 (2008).
 324. G. E. Moore, *Electronics* **38**, 114 (1965).
 325. D. A. Bader (ed.), *Petascale Computing: Algorithms and Applications*, Computational Science Series (CRC Press, Boca Raton, FL, 2007).
 326. G. Amdahl, Validity of the Single Processor Approach to Achieving Large-Scale Computing Capabilities, in *Spring Joint Computer Conference*, (AFIPS Press, Atlantic City, N.J., 1967).
 327. K. Ebcioğlu, V. Saraswat and V. Sarkar, The IBM PERCS Project and New Opportunities for Compiler-Driven Performance via a New Programming Model, *Compiler-Driven Performance Workshop (CASCON 2004)*, (2004).
 328. A. S. Szalay *et al.*, GrayWulf: Scalable Clustered Architecture for Data Intensive Computing, *Hawaii International Conference on System Sciences* (IEEE Computer Society, Los Alamitos, CA, 2009).

329. A. S. Szalay, J. Gray and Vandenberg, J., Petabyte Scale Data Mining: Dream or Reality?, *SPIE Astronomy Telescopes and Instruments*, Waikoloa, Hawaii, (2002).
330. S. M. McConnell and D. B. Skillicorn, Distributed Data Mining for Astrophysical Datasets, in *Astronomical Data Analysis Software and Systems XIV*, eds. P. Shopbell, M. Britton and R. Ebert, Astronomical Society of the Pacific Conference Series, Vol. 347 (2005).
331. A. A. Freitas and S. H. Lavington, *Mining Very Large Databases with Parallel Processing* (Kluwer Academic Publishers, 1998).
332. H. Kargupta and P. Chan, *Advances in Distributed and Parallel Knowledge Discovery* (AAAI/MIT Press, Cambridge, MA, 2000).
333. M. J. Zaki and C. Ho (eds.), *Large-scale Parallel Data Mining*, Lecture Notes in Artificial Intelligence, State-of-the-Art-Survey, Vol. 1759 (Springer, New York, 2002).
334. K. Bhaduri, K. Liu, H. Kargupta and J. Ryan, Distributed Data Mining Bibliography Online bibliography, (2006).
335. R. Jin, G. Yang and G. Agrawal, *IEEE Transactions On Knowledge and Data Engineering* **17**, 71 (2005).
336. N. Gray, The Fact and Future of Semantic Astronomy, in *Astronomical Data Analysis Software and Systems XVII*, eds. R. W. Argyle, P. S. Bunclark and J. R. Lewis, Astronomical Society of the Pacific Conference Series, Vol. 394 (2008).
337. M. L. Norman, G. L. Bryan, R. Harkness, J. Bordner, D. Reynolds, B. O'Shea and R. Wagner, preprint, [arXiv/0705.1556] (2007).
338. R. J. Brunner, V. Kindratenko and A. D. Myers, *Developing and Deploying Advanced Algorithms to Novel Supercomputing Hardware*, tech. rep., NASA (2007).
339. E. L. Gomez, H. L. Gomez and J. Yardley, preprint, [arXiv/0903.0266] (2009).
340. A. W. Moore *et al.*, Fast Algorithms and Efficient Statistics: N-Point Correlation Functions, in *Mining the Sky*, eds. A. J. Banday, S. Zaroubi and M. Bartelmann (2001).
341. D. Gao, Y. Zhang and Y. Zhao, The Application of kd-tree in Astronomy, in *Astronomical Data Analysis Software and Systems XVII*, eds. R. W. Argyle, P. S. Bunclark and J. R. Lewis, Astronomical Society of the Pacific Conference Series, Vol. 394 (2008).
342. A. G. Gray, A. W. Moore, R. C. Nichol, A. J. Connolly, C. Genovese and L. Wasserman, Multi-Tree Methods for Statistics on Very Large Datasets in Astronomy, in *Astronomical Data Analysis Software and Systems (ADASS) XIII*, eds. F. Ochsenbein, M. G. Allen and D. Egret, Astronomical Society of the Pacific Conference Series, Vol. 314 (2004).
343. Y. Shirasaki, M. Ohishi, Y. Mizumoto, M. Tanaka, S. Honda, M. Oe, N. Yasuda and Y. Masunaga, Structured Query Language for Virtual Observatory, in *Astronomical Data Analysis Software and Systems XIV*, eds. P. Shopbell, M. Britton and R. Ebert, Astronomical Society of the Pacific Conference Series, Vol. 347 (2005).
344. S. Derriere *et al.*, UCD in the IVOA context, in *Astronomical Data Analysis Software and Systems (ADASS) XIII*, eds. F. Ochsenbein, M. G. Allen and D. Egret, Astronomical Society of the Pacific Conference Series, Vol. 314 (2004).
345. P. Dowler, S. Gaudet, D. Durand, R. Redman, N. Hill and S. Goliath, Common Archive Observation Model, in *Astronomical Data Analysis Software and Systems XVII*, eds. R. W. Argyle, P. S. Bunclark and J. R. Lewis, Astronomical Society of the Pacific Conference Series, Vol. 394 (2008).
346. J. Gray, D. T. Liu, M. Nieto-Santisteban, A. S. Szalay, D. DeWitt and G. Heber, *Scientific Data Management in the Coming Decade*, Technical Report MSR-TR-2005-10, Microsoft Research (2005).

347. J. Gray, A. S. Szalay, A. R. Thakar, P. Z. Kunszt, C. Stoughton, D. Slutz and J. vandenBerg, preprint, [arXiv:cs/0202014] (2002).
348. C. Vignali, F. Fiore, A. Comastri, M. Brusa, R. Gilli, N. Cappelluti, F. Civano and G. Zamorani, Multi-wavelength data handling in current and future surveys: the possible role of Virtual Observatory, in *Multi-wavelength Astronomy and Virtual Observatory*, ed. D. Baines & P. Osuna (2009).
349. E. A. González-Solares *et al.*, *MNRAS* **388**, 89 (2008).
350. M. Brescia *et al.*, *Memorie della Societa Astronomica Italiana* **80**, p. 565 (2009).
351. T. Kitching, A. Amara, A. Rassat and A. Refregier, preprint, [arXiv/0901.3143] (2009).
352. I. V. Chilingarian, Virtual Observatory for Astronomers: Where Are We Now?, in *Multi-wavelength Astronomy and Virtual Observatory*, ed. D. Baines & P. Osuna (2009).
353. W. B. Landsman, The IDL Astronomy User's Library, in *Astronomical Data Analysis Software and Systems II*, eds. R. J. Hanisch, R. J. V. Brissenden and J. Barnes, Astronomical Society of the Pacific Conference Series, Vol. 52 (1993).
354. M. B. Taylor, TOPCAT & STIL: Starlink Table/VOTable Processing Software, in *Astronomical Data Analysis Software and Systems XIV*, eds. P. Shopbell, M. Britton and R. Ebert, Astronomical Society of the Pacific Conference Series, Vol. 347 (2005).
355. M. Comparato, U. Becciani, A. Costa, B. Larsson, B. Garilli, C. Gheller and J. Taylor, *PASP* **119**, 898 (2007).
356. N. Urunkar, A. K. Kembhavi, A. Navelkar, J. Pandya, V. Moosani, P. Nair and M. Shaikh, *Highlights of Astronomy* **14**, 620 (2007).
357. J. D. Taylor, T. Boch, M. Comparato, M. Taylor, N. Winstanley and R. G. Mann, Binding Applications Together with PLASTIC, in *Astronomical Data Analysis Software and Systems XVI*, eds. R. A. Shaw, F. Hill and D. J. Bell, Astronomical Society of the Pacific Conference Series, Vol. 376 (2007).
358. M. B. Taylor, STILTS - A Package for Command-Line Processing of Tabular Data, in *Astronomical Data Analysis Software and Systems XV*, eds. C. Gabriel, C. Arviset, D. Ponz and S. Enrique, Astronomical Society of the Pacific Conference Series, Vol. 351 (2006).
359. M. Borkin, A. Goodman, M. Halle and D. Alan, Application of Medical Imaging Software to 3D Visualization of Astronomical Data, in *Astronomical Data Analysis Software and Systems XVI*, eds. R. A. Shaw, F. Hill and D. J. Bell, Astronomical Society of the Pacific Conference Series, Vol. 376 (2007).
360. R. Scranton *et al.*, preprint, [arXiv/0709.0752] (2007).
361. D. G. Barnes, C. J. Fluke, P. D. Bourke and O. T. Parry, *Publications of the Astronomical Society of Australia* **23**, 82 (2006).
362. C. J. Fluke, D. G. Barnes and N. T. Jones, *Publications of the Astronomical Society of Australia* **26**, 37 (2009).
363. S. Levy, Interactive 3-D visualization of particle systems with Partiview, in *Astrophysical Supercomputing using Particle Simulations*, eds. J. Makino and P. Hut, IAU Symposium, Vol. 208 (2003).
364. T. Szalay, V. Springel and G. Lemson, preprint, [arXiv/0811.2055] (2008).
365. P. Hut, Virtual Laboratories and Virtual Worlds, IAU Symposium Vol. 246 (2008).
366. T. Ebisuzaki, J. Makino, T. Fukushige, M. Taiji, D. Sugimoto, T. Ito and S. K. Okumura, *PASJ* **45**, 269 (1993).
367. E. Gaburov, S. Harfst and S. Portegies Zwart, *New Astronomy* **14**, 630 (2009).
368. R. G. Belleman, J. Bédorf and S. F. Portegies Zwart, *New Astronomy* **13**, 103 (2008).
369. S. Ord, L. Greenhill, R. Wayth, D. Mitchell, K. Dale, H. Pfister and R. G. Edgar,

- preprint, [arXiv/0902.0915] (2009).
370. V. Garcia, E. Debreuve and M. Barlaud, preprint, [arXiv/0804.1448] (2008).
 371. S. D. Brown, R. J. Francis, J. Rose and Z. G. Vranesic, *Field-Programmable Gate Arrays*, The Springer International Series in Engineering and Computer Science (Springer, New York, 1992).
 372. D. Buell, T. El-Ghazawi, K. Gaj and V. Kindratenko, *Computer* **40**, 23 (2007).
 373. E. Won, *Nuclear Instruments and Methods in Physics Research A* **581**, 816 (2007).
 374. M. Freeman, M. Weeks and J. Austin, Hardware implementation of similarity functions, in *IADIS AC*, 2005.
 375. M. Scarpino, *Programming the Cell Processor: For Games, Graphics, and Computation* (Prentice Hall PTR, New York, 2008).