Computational Intelligence Challenges and Applications on Large-Scale Astronomical Time Series Databases

Pablo Huijse
Millennium Institute of Astrophysics, CHILE

Pablo A. Estévez
*Department of Electrical Engineering,
Universidad de Chile
and the Millennium Institute of Astrophysics, CHILE*

Pavlos Protopapas
*School of Engineering and Applied Sciences,
Harvard University, USA*

José C. Príncipe
*Computational NeuroEngineering Laboratory,
University of Florida, USA*

Pablo Zegers
*Facultad de Ingeniería y Ciencias Aplicadas,
Universidad de los Andes, CHILE*

IMAGE LICENSED BY INGRAM PUBLISHING

Abstract—Time-domain astronomy (TDA) is facing a paradigm shift caused by the exponential growth of the sample size, data complexity and data generation rates of new astronomical sky surveys. For example, the Large Synoptic Survey Telescope (LSST), which will begin operations in northern Chile in 2022, will generate a nearly 150 Petabyte imaging dataset of the southern hemisphere sky. The LSST will stream data at rates of 2 Terabytes per hour, effectively capturing an unprecedented movie of the sky. The LSST is expected not only to improve our understanding of time-varying astrophysical objects, but also to reveal a plethora of yet unknown faint and fast-varying phenomena. To cope with a change of paradigm to data-driven astronomy, the fields of astroinformatics and astrostatistics have been created recently. The new data-oriented paradigms for astronomy combine statistics, data mining, knowledge discovery, machine learning and computational intelligence, in order to provide the automated and robust methods needed for the rapid detection and classification of known astrophysical objects as well as the unsupervised characterization of novel phenomena. In this article we present an overview of machine learning and computational intelligence applications to TDA. Future big data challenges and new lines of research in TDA, focusing on the LSST, are identified and discussed from the viewpoint of computational intelligence/machine learning. Interdisciplinary collaboration will be

Digital Object Identifier 10.1109/MCI.2014.2326100
Date of publication: 14 July 2014

> **The computational intelligence and machine learning fields provide methods and techniques to deal with these problems [...] The correct utilization of these methods is key to dealing with the deluge of available astronomical data.**

required to cope with the challenges posed by the deluge of astronomical data coming from the LSST.

I. Introduction

Time domain astronomy (TDA) is the scientific field dedicated to the study of astronomical objects and associated phenomena that change through time, such as pulsating variable stars, cataclysmic and eruptive variables, asteroids, comets, quasi-stellar objects, eclipses, planetary transits and gravitational lensing, to name just a few. The analysis of variable astronomical objects paves the way towards the understanding of astrophysical phenomena, and provides valuable insights in topics such as galaxy and stellar evolution, universe topology, and others.

Recent advances in observing, storage, and processing technologies have facilitated the evolution of astronomical surveys from observations of small and focused areas of the sky (MACHO [1], EROS [2], OGLE [3]) to deep and extended panoramic sky surveys (SDSS [4], Pan-STARRS [5], CRTS [6]). Data volume and generation rates are increasing exponentially, and instead of still images, future surveys will be able to capture digital "movies of the sky" from which variability will be characterized in ways never seen before.

Several new grand telescopes are planned for the next decade [7], among which is the Large Synoptic Survey Telescope (LSST [8], [9]) under construction in northern Chile and expected to begin operations by 2022. The word "synoptic" is used here in the sense of covering large areas of the sky repeatedly, searching for variable objects in position and time. The LSST will generate a 150 Petabyte imaging database, and a 40 Petabyte worth catalog associated with 50 billion astronomical objects during 10 years [10]. The resolution, coverage, and cadence of the LSST will help us improve our understanding of known astrophysical objects and reveal a plethora of unknown faint and fast-varying phenomena [9]. In addition, the LSST will issue approximately 2 million alerts nightly related to transient events, such as supernovae, for which facilities around the world can follow up.

To produce science from this deluge of data the following open problems need to be solved [11]: a) real-time mining of data streams of ~2 Terabytes per hour, b) real-time classification of the 50 billion followed objects, and c) the analysis, evaluation, and knowledge extraction of the 2 million nightly events.

The big data era is bringing a change of paradigm in astronomy, in which scientific advances are becoming more and more data-driven [12]. Astronomers, statisticians, computer scientists and engineers have begun collaborations towards the solution of the previously mentioned problems, giving birth to the scientific fields of astrostatistics and astroinformatics [13]. The development of fully-automated and robust methods for the rapid classification of what is known, and the characterization of emergent behavior in these massive astronomical databases are the main tasks of these new fields. We believe that computational intelligence, machine learning and statistics will play major roles in the development of these methods [10], [12].

The remainder of this article is organized as follows: In section II the fundamental concepts related to time-domain astronomy are defined and described. In section III an overview of current computational intelligence (CI) and machine learning (ML) applications to TDA is presented. In section IV future big data challenges in TDA are exposed and discussed, focusing on what is needed for the particular case of the LSST from an ML/CI perspective. Finally, in section V conclusions are drawn.

II. Astronomical Background

In this section we describe the basic concepts related to astronomical time series analysis and time-domain astronomical phenomena. Photometry is the branch of astronomy dedicated to the precise measurement of visible electromagnetic radiation from astronomical objects. To achieve this, several techniques and methods are applied to transform the raw data from the astronomical instruments into standard units of flux or intensity. The basic tool in the analysis of astronomical brightness variations is the **light curve**. A light curve is a plot of the magnitude of an object's electromagnetic radiation (in the visible spectrum) as a function of time.

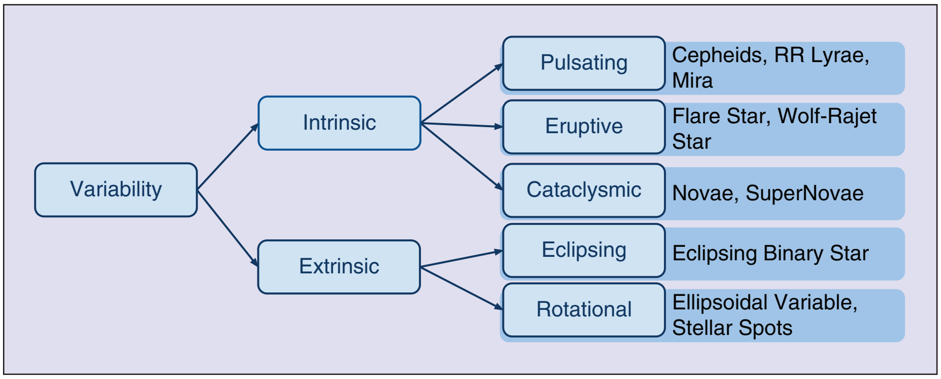

FIGURE 1 Variable star topological classification.

Light curve analysis is challenging, not only because of the sheer size of the databases, but also due to the characteristics of the data itself. Astronomical time series are unevenly sampled due to constraints in the observation schedules, telescope allocations and other limitations. When observations are taken from Earth the resulting light curves will have periodic one-day gaps. The sampling is randomized because observations for each object happen at different times every night. The cycles of the moon, bad weather conditions and sky visibility impose additional constraints which translate into data gaps of different lengths. Space observations are also restricted as they are regulated by the satellite orbits. Discontinuities in light curves can also be caused by technical factors: repositioning of the telescopes, calibration of equipment, electrical, and mechanical failures, etc.

The resolution, coverage, and cadence of the LSST will help us improve our understanding of known astrophysical objects and reveal a plethora of unknown faint and fast-varying phenomena.

Astronomical time series are also affected by several noise sources. These noise sources can be broadly categorized into two classes. The first class is related to observations, such as the brightness of closer astronomical objects, and atmospheric noise due to refraction and extinction phenomena (scattering of light due to atmospheric dust). On the other hand, there are noise sources related to the instrumentation, in particular to the CCD cameras, such as sensitivity variations of the detector, and thermal noise. In general, errors in astronomical time series are non-Gaussian and heterocesdastic, i.e., the variance of the error is not constant, and changes along the magnitude axis.

Other common problematic situations arising in TDA are the sample-selection bias and the lack of balance between classes. Generally the astrophysical phenomena of interest represents a small fraction of the observable sky, hence the vast majority of the data belongs to the "background class". This is especially noticeable when the objective is to find unknown phenomena, a task known as novelty detection. Sufficient coverage and exhaustive labeling are required in order to have a good representation of the sample, and to assure capturing the rare objects of interests.

In the following we briefly describe several time-domain astronomical phenomena emphasizing their scientific interest. We focus on phenomena that vary in the optical spectrum. Among the "observable stars" there is a particular group called the **variable stars** [14]–[16]. Variable stars correspond to stellar objects whose brightness, as observed from Earth, fluctuates in time above a certain variability threshold defined by the sensitivity of the instruments. Variable star analysis is a fundamental pivot in the study of stellar structure and properties, stellar evolution and the distribution and size of our Universe. The major categories of variable stars are briefly described in the following paragraphs with emphasis on the scientific interest behind each of them. For a more in-depth definition of the objects and their mechanisms of variability, the reader can refer to [15]. The relation between different classes of variable stars is summarized by the tree diagram shown in Fig. 1 [16], [17].

The analysis of intrinsic variable stars is of great importance for the study of stellar nuclei and evolution. Some classes of intrinsic variable stars can be used as distance markers to study the distribution and topology of the Universe. Cepheid and RR Lyrae stars [15] (Fig. 2a) are considered standard candles because of the relation between their pulsation period and their absolute brightness. It is possible to estimate the distance from these stars to Earth with the period and the apparent

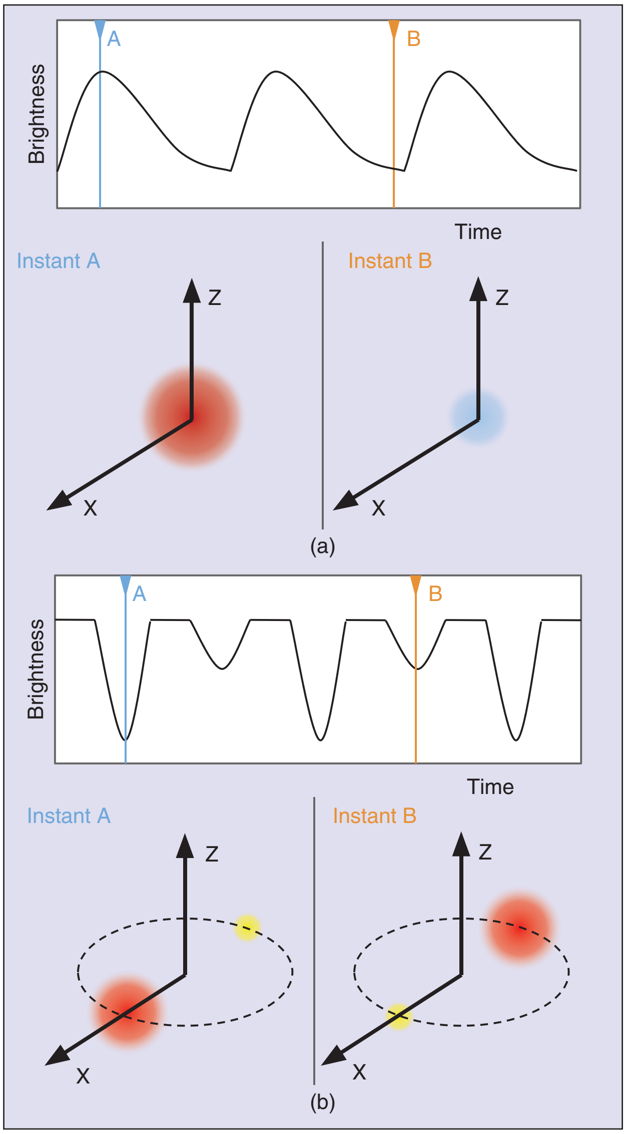

FIGURE 2 (a) Light curve of a pulsating variable star (upper left panel), such as a Cepheid or RR Lyrae. The star pulsates periodically changing in size, temperature and brightness which is reflected on its light curve. (b) Light curve of eclipsing binary star (upper right panel). The lower panels show the geometry of the binary system at the instants where the eclipses occur. The periodic pattern in the light curve is observed because the Earth (X axis) is aligned with the orbital plane of the system (Z axis).

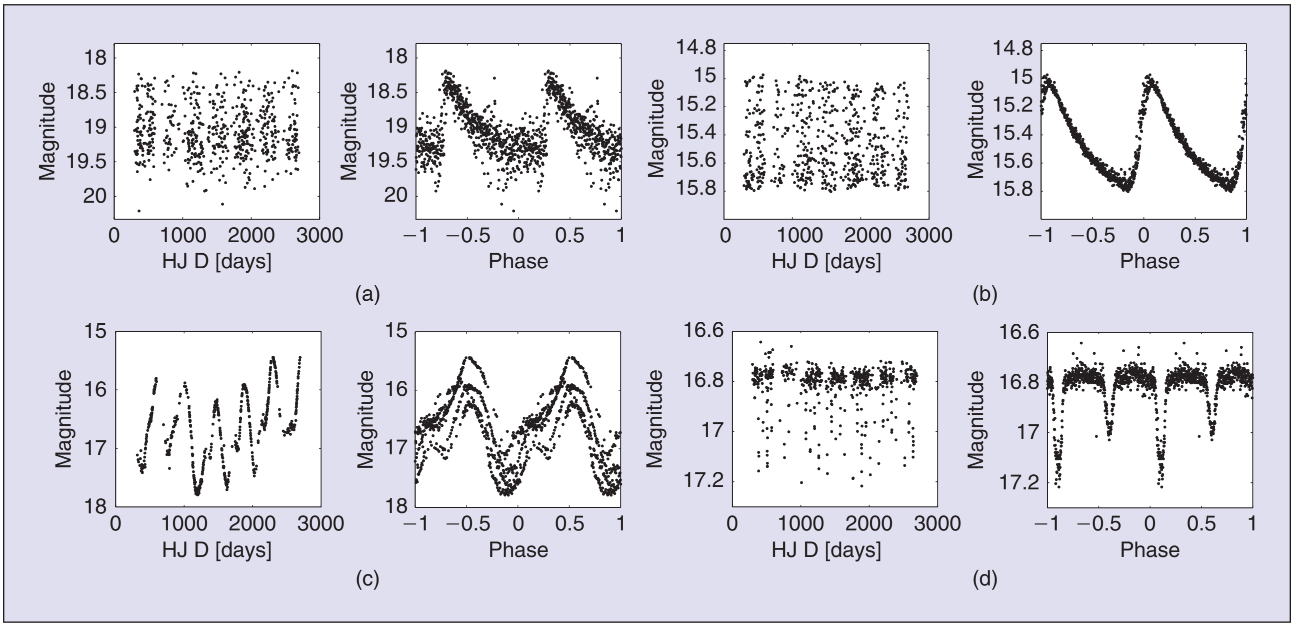

FIGURE 3 Light curve and phase diagram of an RR Lyrae (a), Cepheid (b), Mira (c) and eclipsing binary star (d), respectively. The phase diagram is obtained using the underlying period of the light curves and the epoch folding transformation [15]. If the folding period is correct a clear profile of the periodicity will appear in the phase diagram.

brightness measured from the telescope [18]. Type 1A Supernovae [15] are also standard candles, although they can be used to trace much longer distances than Cepheids and RR Lyrae [19]. The period of eclipsing binary stars [15] (Fig. 2b) is a key parameter in astrophysics studies as it can be used to calculate the radii and masses of the components [20]. Light curves and phase diagrams of periodic variable stars are shown in Fig. 3.

III. Review of Computational Intelligence Applications in TDA

Time-domain astronomers are faced with a wide array of scientific questions that are related to the detection, identification and modeling of variable phenomena such as those presented in the previous Section. We may classify these problems broadly as follows:

1) Extract information from the observed time series in order to understand the underlying processes of its source.
2) Use previous knowledge of the time-varying universe to classify new variable sources automatically. How do we characterize what we know?
3) Find structure in the data. Find what is odd and different from everything known. How do we compare astronomical objects? What similarity measure do we use?

The computational intelligence and machine learning fields provide methods and techniques to deal with these problems in a robust and automated way. Problem 1) is a problem of modeling, parametrization and regression (kernel density estimation). Problem 2) corresponds to supervised classification (artificial neural networks, random forests, support vector machines). Problem 3) deals with unsupervised learning, feature space distances and clustering (k nearest neighbors, self-organizing maps). The correct utilization of these methods is key to dealing with the deluge of available astronomical data. In the following section we review particular cases of computational intelligence based applications for TDA.

A. Periodic Variable Star Discrimination

We begin this review with a case of parameter estimation from light curves using information theoretic criteria. Precise period estimations are fundamental in the analysis of periodic variable stars and other periodic phenomena such as transiting exoplanets. In [21] the correntropy kernelized periodogram (CKP), a metric for period discrimination for unevenly sampled time series, was presented. This periodogram is based on the correntropy function [22], an information theoretic functional that measures similarity over time using statistical information contained in the probability density function (pdf) of the samples. In [21] the CKP was tested on a set of 5,000 light curves from the MACHO survey [1] previously classified by experts. The CKP achieved a true positive rate of 97% having no false positives and outperformed conventional methods used in astronomy such as the Lomb-Scargle periodogram [23], ANOVA and string length.

In [24] the CKP was used as the core of a periodicity discrimination pipeline for light curves from the EROS-2 survey [2]. The method was calibrated using a set of 100,000 synthetic light curves generated from multivariate models constructed following the EROS-2 data. Periodicity thresholds and rules to adapt the kernel parameters of the CKP were obtained in the calibration phase. Approximately 32 million light curves from the Large and Small Magellanic clouds were tested for periodicity. The pipeline was implemented for GPGPU architectures taking 18 hours to process the whole EROS-2 set on a cluster with 72 GPUs. A catalog of 120 thousand periodic variable stars was obtained and cross-matched with existing catalogs for the Magellanic clouds for validation. The main contributions of [24] are the procedure used to create the training database using the available survey data, the fast implementation geared towards large astronomical databases, the large periodic light curve catalog generated from EROS-2, and the valuable inference on the percentage of periodic variable stars.

Another example of information theoretic concepts used for periodicity detection in light curves can be found in [25]. In this work the Shannon's conditional entropy of a light curve is computed from a binned phase diagram obtained for a given period candidate. The conditional entropy is minimized in order to find the period that produces the most ordered phase diagram. The proposed method was tested using a training set of periodic light curves from the MACHO survey and the results show that it is robust against systematic errors produced by the sampling, data gaps, aliasing and artifacts in phase space.

In Tagliaferri et al. [26], neural networks are used to obtain the parameters of the periodogram of the light curve. These parameters are then fed into the MUSIC (Multiple Signal Classification) to generate a curve whose peaks are located in the periods sought. Interestingly, this work also shows the relation between the presented method and the Cramer-Rao bound, thus posing absolute practical limits to the performance of the proposed procedure.

A comprehensive analysis of period finding algorithms for light curves can be found in [27]. In this work classical methods for period discrimination in astronomy such as the Lomb-Scargle periodogram and Phase Dispersion minimization are compared to novel information theoretic criteria [25]. The authors note that the accuracy of each individual method is dependent on observational factors and suggest that an ensemble approach that combines several algorithms could mitigate this effect and provide a more consistent solution. How to combine the output of different methods and the increased computational complexity are key issues to be solved.

B. Automated Supervised Classification for Variable Objects

After obtaining the period, supervised methods can be used to discriminate among the known classes of periodic variable stars. In supervised learning, prior information in the form of a training dataset is needed to classify new samples. The creation and validation of these training sets are complex tasks in which human intervention is usually inevitable. This is particularly challenging in the astronomical case due to the vast amounts of available data. If the data do not initially

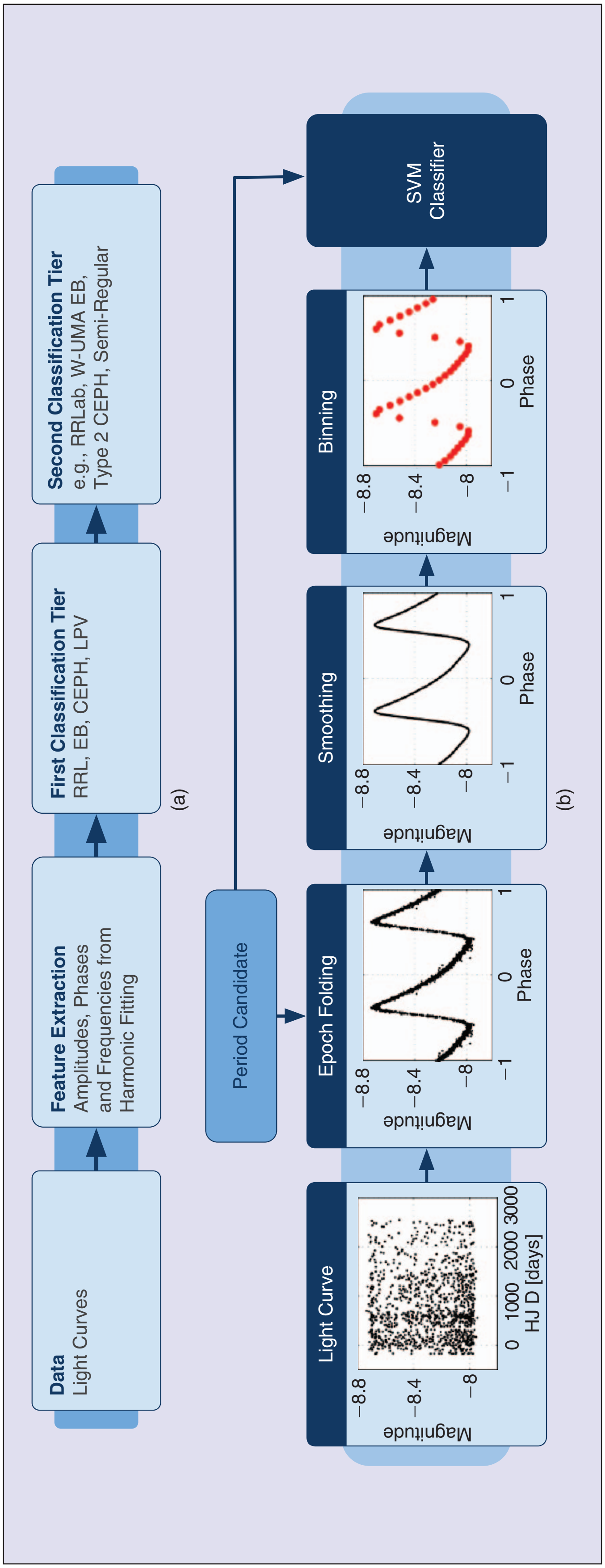

FIGURE 4 (a) Classification scheme used in [28] for variable star classification. (b) Light curve processing pipeline used in [30]. A candidate period is used to obtain a phase diagram of the light curve which is then smoothed, interpolated to a fixed grid, and binned. The period, magnitude, color and binned values are used as features for the SVM classifier.

There is a natural synergy between active learning and astronomy, because the astronomer is [...] able to follow up a certain target in order to obtain additional information.

represent the population well, then scientific discovery may be hindered. In addition, due to the differences between observational surveys, it is very difficult to reuse the training sets. In the following paragraphs several attempts of supervised classification for TDA are reviewed, with emphasis on the classification scheme and the design of the training databases.

Gaussian mixture models (GMMs), and artificial neural networks trained through Bayesian model averaging (BAANN) were used to discriminate periodic variable stars from their light curves in [28]. Classification schemes using single multi-class and multi-stage (hierarchical) classifiers were compared. This work is relevant not only because of the application and comparison between the algorithms, but also because of the extended analysis performed in building the training dataset. First, well-known class prototypes were recovered from published catalogs and reviewed. Data from nine astronomical surveys were used, although the vast majority came from the Hipparcos and OGLE projects. A diagram of the classification pipeline for the hierarchical classifier is shown in Fig. 4a. The selected variability types were parametrized using harmonic fitting (Lomb-Scargle periodogram). The classes were organized in a tree-like structure similar to the one shown in Fig. 1. The final training set contained 1,732 samples from 25 well-represented classes. For the single stage classifier the GMM and BAANN obtained correct classification rates of 69% and 70%, respectively. Only the BAANN was tested using the multi-stage scheme obtaining a correct classification rate of 71%. According to the authors, the GMM provides a simple solution with direct astrophysical interpretation. On the other hand, some machine learning algorithms may achieve lower misclassification rates but their interpretability is reduced. The authors also state the need for higher statistical knowledge, which can be provided through interdisciplinary cooperation, in order to use the machine learning approach. A more recent version of this method can be found in [29]. In this work 26,000 light curves from TrES and Kepler surveys were classified using a multi-stage tree of GMMs. The main difference from the previous version is the careful selection of significant frequencies and overtone features which reduces confusion between classes.

In [30], 14,087 periodic light curves from the OGLE survey were used to train and test supervised classifiers based on k-NN and SVM. The periods and labels were obtained directly from the OGLE survey. The following periodic variable classes were considered: Cepheids, RR Lyrae and Eclipsing Binaries. The period, average brightness and color of the light curves were used as features for the classifier. In addition, the authors included the phase diagram of the light curve as a feature and proposed a kernel to compare time series, which is plugged into the SVM. The phase diagram is obtained using the underlying period of the light curves and the epoch folding transformation [15]. The light curve processing pipeline is shown in Fig. 4b. The proposed kernel takes care of the possible difference in phase between time series. Using the shape of the light curve and the proposed kernel, correct classification rates close to 99% were obtained. Intuitively, the shape of the periodicity (phase diagram) should be a strong feature for periodic variable star classification. The authors note that a complete pipeline would require first discriminating whether the light curve is periodic, and estimating its period with high accuracy. Wrongly estimated periods would alter the folded light curve, affecting the classification performance.

A Random Forest (RF) classifier for periodic variable stars from the Hipparcos survey was presented in [31]. A set of 2,000 reliable variable sources found in the literature was used to train the classifier to discriminate 26 types of periodic variable stars. Non-periodic variables were also added to the training set. Light curves were characterized using statistical moments, periods and Fourier coefficients. The performance of the classifier is consistent with other studies [28]. The authors found that the most relevant features in decreasing order are the period, amplitude, color and the Fourier coefficients (light curve model). The authors also found that the major sources of misclassification are related to the reliability of the estimated periods and the misidentification of non-periodic light curves as periodic.

Statistical classifiers work under the assumption that class probabilities are equivalent for the training and testing sets. According to [32] this assumption may not hold when the number of light curve measurements is different between sets. In [32] this problem is addressed via noisification and denoisification, *i.e.*, trying to modify the pdf of the training set so that it mimics the test set, and to infer the class of a poorly sampled time series according to its most probable evolution. This scheme is tested on light curves from the OGLE survey. Results show that noisification and denoisification improve the classification accuracy for poorly sampled time series by 20%. The authors note that the proposed method may help overcome other systematic differences between sets such as varying cadences and noise properties.

Classification of non-periodic variable sources is less developed than periodic source classification. Non-periodic sources in general are more difficult to characterize, which is why only a few studies do general Active Galactic Nuclei (AGN) classification, instead focusing on discriminating a particular type of quasi-stellar object (QSO). In [33] a supervised classification scheme based on SVM was used to discriminate AGN from their light curves. The objects were characterized using 11 features including amplitude, color, autocorrelation function, variability index and period. A training set of $\sim$5,000 objects including non-variable and variable stars and 58 known quasars from the MACHO survey was used to train the classifier. This work [33] differs from previous attempts at AGN discrimination in its thoughtful study of the efficiency and false positive rates of the model and classifier. The classifier was tested on the full 40 million MACHO light curves finding 1,620 QSO candidates that were later cross-matched with external quasar catalogs, confirming the validity of the candidates.

In [34], a multi-class classifier for variable stars was proposed and tested. This implementation shares the features, data, and the 25-class taxonomy used in [28], allowing direct comparison. An important distinction with respect to [28] is that the classification is extended to include non-periodic eruptive variables[1]. Two classification schemes are tested: In the first, pairwise comparisons between two-class classifiers are used, and in the second, a hierarchical classification scheme based on the known taxonomy of the variable stars is employed (Fig. 1). The best performance is obtained by the RF with pairwise class comparisons, achieving a correct classification rate of 73.3% when using the same features as [28], and 77.2% when only the more relevant features are used. In a taxonomical sense, a mistake committed in the first tier of the classification hierarchy (catastrophic error) is more severe than a mistake in the final tiers (sub-type classifiers). The hierarchical RF implementation obtains a slightly worse overall performance (1%) and a smaller catastrophic error (8%) than the pairwise RF. Although there is a considerable improvement with respect to [28], the accuracy is not high enough for fully automated classification. The taxonomical classification of variable stars and their multiple subtypes is still an open problem.

Using information from several catalogs may improve the characterization of the objects under study, but joining catalogs is not a trivial task as different surveys use different instruments and may be installed in totally different locations. Intersecting the catalogs, i.e., removing columns/rows with missing data may result in a database that is smaller than the original single catalogs. A variable star classifier for incomplete catalogs was proposed in [35]. In this work the structure of a Bayesian network is learned from a joined catalog with missing data from several surveys. The joined catalog is "filled" by estimating the probability distributions and dependencies between the features through the Bayesian network. The resulting training set has 1,833 samples including non-variables, non-periodic variables (quasars and Be stars), and periodic variable stars (Cepheids, RR Lyrae, EB and Long Period Variables). An RF classifier is trained with the joined catalog. The Bayesian network is compared to a second approach for filling missing data based on GMM, obtaining better classification accuracy for most of the selected classes. An additional test on 20 million light curves from the MACHO survey (a catalog with no missing data) was performed. From this test a set of 1,730 quasar candidates was obtained which corresponds to a 15% improvement with respect to previous quasar lists in the literature [33].

C. Unsupervised and Semi-Supervised Learning in Time-Domain Astronomy

In some cases previous information on the phenomena might be insufficient or non-existent, hence a training set cannot be constructed. In these cases one may need to go back one step and obtain this information from the data using unsupervised methods. In a broad sense the objective of unsupervised methods is to estimate the density function that generated the data revealing its structure. One of the first references for unsupervised learning in astronomy is found in [36] where self-organizing maps (SOMs) were used to discriminate clusters of stars in the solar neighborhood. One hundred thousand stars from the Hipparcos catalog, mixed with synthetic data, were used. The synthetic stars were modeled with particular characteristics of known stellar populations. The Hipparcos catalog provides information about the position, magnitude (brightness), color[2], spectral type and variability among other features. The SOM was trained on a 10×10 grid with the additional constraint that each node should have at least one synthetic star. Using the synthetic stars as markers, clusters of stellar populations were recognized in the visualization of the SOM. In this case the SOM was used not only to find population clusters but also to validate the theoretical models used to create the synthetic stars.

SOM was also used in [37] in order to learn clusters of mono-periodic variable stars. The main objective was to classify periodic variable stars in the absence of a training set, which was the reason the SOM was selected. The feature in this case was an N-dimensional vector obtained from the folded light curves. Each light curve was folded with its period which was known a priori. The folded light curves were then normalized in scale and discretized in N bins. The number of bins and SOM topology parameters were calibrated using five thousand synthetic light curves representing four classes of periodic variables. The SOM was then tested on a relatively small set of 1,206 light curves. The clusters for each class were discriminated using a U-matrix visualization and density estimations. Clusters associated with Eclipsing Binaries, Cepheids, RR Lyrae and δ-scuti populations were identified. Although restricted in size, this application shows the potential of the SOM for class discovery in time-domain astronomy.

In [38] a density based-approach for clustering was used to find groups of variable stars in the OGLE and CoRoT surveys. The light curves are characterized using the features described in [28]. Each point in feature space is assigned to one of the clusters using Modal Expectation Maximization. This work addresses the need of tying up astronomical knowledge with the outcome of the computational intelligence algorithms. The manner in which astronomers have classified stellar objects and

[1] Simple statistical moments from the flux distribution are used as features for the eruptive variables.

[2] The color corresponds to the difference in average brightness between two different spectra.

Astronomy has a long history of serendipitous discovery [...] Computational intelligence and machine learning may provide the means for facilitating the task of novelty detection.

events does not necessarily correspond to that produced by automated systems. Interestingly, this study establishes that there is another problem as well: the same computational intelligence algorithms working on different databases produced distinct classification structures, showing that even though these databases have large numbers of examples, they have inherent biases and may not be sufficiently large to allow the discovery of general rules. This problem has also been reported in other fields, specifically artificial vision [39]. The work in [39] showed that in order to produce consistent classification performances, one could not simply use databases with hundreds of thousands of examples, it was necessary to use close to 80 million images, far exceeding what was traditionally considered enough by the practitioners of the field.

Kernel Principal Component Analysis (KPCA) was used in [40] to perform spectral clustering on light curves from the CoRoT survey. The light curves were characterized using three different approaches: Fourier series, autocorrelation functions, and Hidden Markov Models (HMMs). Then, dimensionality was reduced with KPCA using the Gaussian kernel. Finally, the eigenvalues were used to find clusters of variable stars. This novel characterization of light curves permits identifying not only periodic variable stars correctly (Fourier and autocorrelation features), but also irregular variable stars (HMM features).

Unsupervised learning can also be used for novelty detection, i.e., finding objects that are statistically different from everything that is known and hence cannot be classified in one of the existing categories. Astronomy has a long history regarding serendipitous discovery [41], i.e., to find the unexpected (and unsought). Computational intelligence and machine learning may provide the means for facilitating the task of novelty detection.

One may argue that the first step for novelty detection is to define a similarity metric for astronomical time series in order to compare time-varying astronomical objects. This is the approach found in [42] where a methodology for outlier light curve identification in astronomical catalogs was presented. A similarity metric based on the correlation coefficient is computed between every pair of light curves in order to obtain a similarity matrix. Intuitively, the outlier variable star will be dissimilar to all the other variables. Before any distance calculation, light curves are interpolated and smoothed in order to normalize the number of points and time instants. For each pair the lag of maximum cross-correlation is found in Fourier space, which solves the problem of comparison between light curves with arbitrary phases. Finally the outliers correspond to the light curves with the lowest cross correlations with respect to each row of the distance matrix. The method was tested on ~34,500 light curves from early-stage periodic variable star catalogs originated from the MACHO and OGLE [3] surveys. The results of this process were lists of mislabeled variables with careful explanations of the new phenomena and reasons why they were misclassified. Calculating the similarity matrix scales quadratically with the number of light curves in the survey. The authors discuss this issue and provide an approximation of the metric that reduces the computational complexity to $O(N)$. Also, an exact and efficient solution for distance-based outlier detection can be found in [43], which uses a discord metric that requires only two linear searches to find outlier light curves as well.

A different approach for novelty detection was given in [44], where an anomaly detection technique dubbed PCAD (Periodic Curve Anomaly Detection) was proposed and used to find outlier periodic variables in large astronomical databases. PCAD finds clusters using a modified k-means algorithm called phased k-means (pk-means). This modification is required in order to compare asynchronous time series (arbitrary phases). By using a clustering methodology the authors were able to find anomalies in both a global and local sense. Local anomalies correspond to periodic variables that lie in the frontier of a given class. Global anomalies on the other hand differ from all the clusters. Approximately 10,000 periodic light curves from the OGLE survey were tested with PCAD. The pre-processing of the light curves, the selection of features, and the computation of the cross-correlations follow the work of [42]. The cross-correlation is used as a distance metric for the pk-means. The results obtained by the method were then evaluated by experts and sorted as noisy light curves, misclassified light curves and interesting outliers worthy of follow-up.

A problem with purely unsupervised methods is that prior knowledge, when available, is not necessarily used. Semi-supervised learning schemes deal with the case where labels (supervised information) exist although not for all the available data. Semi-supervised methods are able to find the structure of the data distribution, learn representations and then combine this information with what is known. Semi-supervised methods can also be used for novelty detection, with the benefit that they may improve their discrimination by automatically incorporating the newly extracted knowledge. The semi-supervised approach is particularly interesting in the astronomical case where prior information exists, although scarce in comparison to the bulk of available unlabeled data. In [45] a semi-supervised scheme for classification of supernova subtypes was proposed. In this work the unlabeled supernovae data are used to obtain optimal low-dimensional representations in an unsupervised way. A diagram of the proposed implementation is shown in Fig. 5a. In general, features are extracted from supernovae light curves following fixed templates. The data-driven feature extraction proposed in [45]

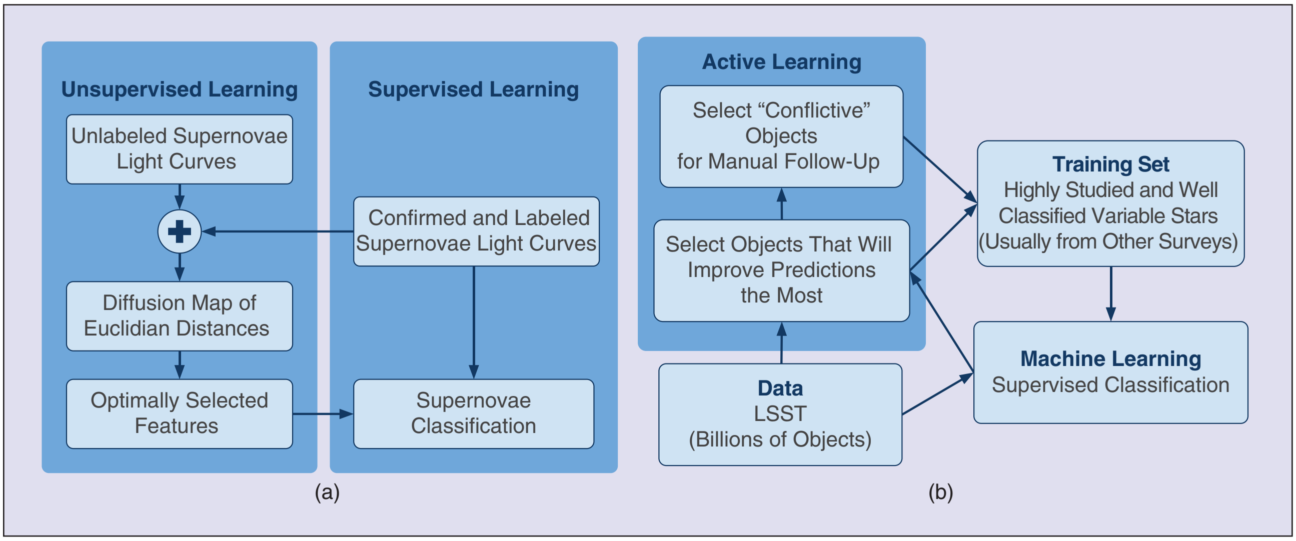
Unsupervised Learning
Unlabeled Supernovae Light Curves
Diffusion Map of Euclidian Distances
Optimally Selected Features
Supervised Learning
Confirmed and Labeled Supernovae Light Curves
Supernovae Classification
(a)
Active Learning
Select "Conflictive" Objects for Manual Follow-Up
Select Objects That Will Improve Predictions the Most
Data
LSST
(Billions of Objects)
Training Set
Highly Studied and Well Classified Variable Stars
(Usually from Other Surveys)
Machine Learning
Supervised Classification
(b)

FIGURE 5 (a) Semi-supervised learning scheme used in [45] for Supernovae classification. (b) An active learning approach to building training sets for supervised classification of variable stars. Samples from the testing set are moved (unsupervised) to the training sets reducing sample selection bias. The expert is queried by the method if more information is needed for obtaining the new labels.

performs better and is more efficient than template methods with respect to data utilization and scaling.

A bias due to sample selection occurs when training and test datasets are not drawn from the same distribution. In astronomical applications, training datasets often come from older catalogs which, because of technological constraints, contain more information on brighter and closer astronomical objects, i.e., the training dataset is a biased representation of the whole. In these cases standard cross-validation procedures are also biased resulting in poor model selection and sub-optimal prediction. The selection bias problem is addressed from an astronomical perspective in [46] through the use of active learning (AL). A diagram of the implementation proposed in [46] is shown in Fig. 5b. In AL, the method queries the expert for manual follow-up of objects that cannot be labeled automatically. There is a natural synergy between AL and astronomy, because the astronomer is, in general, able to follow up a certain target in order to obtain additional information. The AL classifier consistently performed better than traditional implementations.

IV. Future Big Data Challenges in Time Domain Astronomy

In this section we describe the future big data challenges in TDA from the viewpoint of computational intelligence and machine learning using as an example the LSST. The astronomical research problems targeted by the LSST are described in the LSST Science Book [9]. The LSST focuses on time-domain astronomy, and will be able to provide a movie of the entire sky in the southern hemisphere for the first time. Some of the LSST challenges are: detecting faint transient signals such as supernovae with a low false-positive rate, classifying transients, estimating the distance from Earth, and making discoveries and classification in real-time [12]. Facilities such as the LSST will produce a paradigm change in astronomy. On the one hand the new telescope will be entirely dedicated to a large-scale survey of the sky, and individual astronomers will not be allowed to make private observations as they used to do in the past [12]. On the other hand, the data volume and its flow rate would be so large that most of the process should be done automatically using robotic telescopes and automated data analysis. Data volumes from large sky surveys will grow from Terabytes during this decade (e.g., PanSTARRS [5]) to hundreds of Petabytes during the next decade (LSST [8], [9]). The final LSST image archive will be ~150 Petabytes and the astronomical object catalog (object-attribute database) is expected to be ~40 Petabytes, comprising 200 attributes for 50 billion objects [10]. In [11] the following three challenges are identified for the LSST: a) Mining a massive data stream of ~2 Terabytes per hour in real time for 10 years, b) classifying more than 50 billion objects and following up many of these events in real time, c) extracting knowledge in real time for ~2 million events per night. The analysis of astronomical data involves many stages, and in all of them it is possible to use CI techniques to help its automation. In this paper we focused on the analysis of the light curves, but there are other CI challenges in image acquisition & processing [9], dimensionality reduction and feature selection [47], [48], etc. There are also major technical challenges related to the astronomical instruments, storage, networking and processing facilities. The challenges associated with data management and the proposed engineering solutions are described in [8].

The LSST will be dedicated exclusively to the survey program, thus follow-up observations (light curves, spectroscopy, multiple wavelengths), which are scientifically essential, must be done by other facilities around the world [10]. With this goal the LSST will generate millions of event alerts during each night for 10 years. Many of the observed phenomena are

Computational intelligence methods for pattern recognition are essential for the proper exploitation of synoptic surveys, being able to detect and characterize events that otherwise might not even be noticed by human investigators.

transient events such as supernovae, gamma-ray bursts, gravitational microlensing events, planetary occultations, stellar flares, accretion flares from supermassive black holes, asteroids, etc. [49]. A key challenge is that the data need to be processed as it streams from the telescopes, comparing it with the previous images of the same parts of the sky, automatically detecting any changes, and classifying and prioritizing the detected events for rapid follow-up observations [50]. The system should output a probability of any given event as belonging to any of the possible known classes, or as being unknown. An important requirement is maintaining high level of completeness (do not miss any interesting events) with a low false alarm rate, and the capacity to learn from past experience [49]. The classification must be updated dynamically as more data come in from the telescope and the feedback arrives from the follow-up facilities. Another problem is determining what follow-up observations are the most useful for improving classification accuracy, and detecting objects of scientific interest. In [51] maximizing the conditional mutual information is proposed.

Tackling the future challenges in astronomy will require the cooperation of scientists working in the fields of astronomy, statistics, informatics and machine learning/computational intelligence [12]. In fact the fields of astroinformatics [13] and astrostatistics have been recently created to deal with the challenges mentioned above. Astroinformatics is the new data-oriented paradigm for astronomy research and education, which includes data organization, data description, taxonomies, data mining, knowledge discovery, machine learning, visualization and statistics [10].

The characterization (unsupervised learning) and classification (supervised learning) of massive datasets are identified as major research challenges [10]. For time-domain astronomy the rapid detection, characterization and analysis of interesting phenomena and emergent behavior in high-rate data streams are critical aspects of the science [10]. Unsupervised learning and semisupervised learning are believed to play a key role in new discoveries. To deal with big data in TDA in the Peta-scale era the following open problems need to be solved:

1) Developing Very Efficient Algorithms for Large-Scale Astroinformatics/Astrostatistics

Fast algorithms for commonly used operations in astronomy are described in [52]: e.g., all nearest neighbors, n-point correlation, Euclidean minimum spanning tree, kernel density estimation (KDE), kernel regression, and kernel discriminant analysis. N-point correlations are used to compare the spatial structure of two data sets, e.g., luminous red galaxies in the Sloan digital sky survey [53]. KDE is used for comparing the distributions of different kinds of objects. Most of these algorithms involve distance comparisons between all data pairs, and therefore are naively $O(N^2)$ or of even higher complexity. With the goal of achieving linear or $O(N \log N)$ runtimes for pair-distance problems, space-partitioning tree data structures such as kd-trees are used, in a divide and conquer approach. In the KDE problem series expansions for sums of kernels functions are truncated to approximate continuous functions of distance. In [54] it is argued that the algorithms should be efficient in three respects: computational (number of computations done), statistical (number of samples required for good generalization), and human involvement (amount of human labor to tailor the algorithm to a task). The authors state that there are fundamental limitations for certain classes of learning algorithms, *e.g.*, kernel methods. These limitations come from their shallow structure (single layered) which can be very inefficient in representing certain types of functions, and from using local estimators which suffer the curse of dimensionality. Contrarily, deep architectures, which are compositions of many layers of adaptive nonlinear components, *e.g.*, multilayer neural networks with several hidden layers, have the potential to generalize in nonlocal ways. In [55] a layer-by-layer unsupervised learning algorithm for deep structures was proposed, opening a new line of research that is still on-going.

2) Developing Effective Statistical Tools for Dealing with Big Data

The large data sample and high dimensionality characteristics of big data, raise three statistical challenges [56]: i) noise accumulation, spurious correlations, and incidental endogeneity (residual noise is correlated with the predictors), ii) heavy computational cost and algorithmic instability, iii) heterogeneity, statistical biases. Dimension reduction and variable selection are key for analyzing high dimensional data [47], [48], [56]. Noise accumulation can be reduced by using the sparsity assumption.

3) Creating Implementations Targeted for High-Performance Computing (HPC) Architectures

Traditional analysis methods used in astronomy do not scale to peta-scale volumes of data on a single computer [57]. One could rework the algorithms to improve computational efficiency but even this might prove to be insufficient with the new surveys. An alternative is to decompose the problem into several independent sub-problems. Computations can then proceed in parallel over a shared memory cluster, a distributed memory cluster, or a combination of both. In a shared memory cluster the processes launched by the user can communicate and share data and results through memory. In a distributed environment each processor receives data and instructions, performs the computations and reports the results back to the main server. The number of processors per node, amount of shared memory and network speed have to be taken into account when implementing an algorithm for HPC architectures. Efficiency will ultimately depend on how separable the problem is in the first place. Another parallel computing strategy involves the use of GPUs (graphical processing units) instead or side by side with conventional CPUs (central processing units). GPGPU (general purpose computing in GPU) is a relatively new paradigm for highly parallel applications in which high-complexity calculations are offloaded to the GPU (coprocessor). GPUs are inherently parallel harnessing up to 2,500 processing cores[3]. The processing power and relatively low cost of GPUs have made them popular in the HPC community and their availability has been on the rise [58]. Note that explicit thread and data parallelism must be exploited in order to get the theoretical speed-ups of GPUs over CPUs. Dedicated hardware based on FPGAs may provide interesting speed-ups for TDA algorithms [59]. However due to the advanced technical knowledge required to use them, FPGAs are not as popular in astronomy as the HPC resources already presented. Interdisciplinary collaborations between electrical & computer engineers and astronomers might change this in the near future. An existing non-parallel algorithm can be extended using the MapReduce [60] model for distributed computing, a model inspired from functional programming. Programs written in this functional style are automatically parallelized. In [61] the MapReduce model was used to develop distributed and massively parallel implementations of k-means, support vector machines, neural networks, Naïve Bayes, among others. In the cloud computing paradigm the hardware resources (processors, memory and storage) are almost entirely abstracted and can be increased/decreased by the user on demand. Distributed models such as MapReduce have a high synergy with the cloud computing paradigm. Cloud computing services such as Amazon EC2 provide cost-effective HPC solutions for compute and/or memory bound scientific applications as shown in [62]. As pointed out in [63], one of the biggest advantages of implementing astronomical pipelines using cloud services is that the computing resources can be scaled rather easily and according to changing workloads. The granular computing paradigm [64] is also of interest for astronomical big data applications. In granular computing the information extracted from data is modeled as a hierarchical structure across different levels of detail or scales. This can help to compress the information and reduce the dimensionality of the problem. Another approach is the virtual observatory (VO), a cyberinfrastucture for discovery and access to all distributed astronomical databases [10]. A useful data portal for data mining is OpenSkyQuery [65], which allows users to do multi-database queries on many astronomical object catalogs.

Tackling the future challenges in astronomy will require the cooperation of scientists working in the fields of astronomy, statistics, informatics and machine learning/computational intelligence.

4) Developing Fast Algorithms for Online Event Detection and Discrimination

Several facilities around the world will follow the 2 million events that the LSST will issue each night. These facilities will need to decide which events are most relevant so as not to waste their limited observing and storage resources. In addition, these decisions have to be made as fast as possible to avoid missing important data. Pattern recognition methods to quickly analyze and discriminate interesting phenomena from the streamed data are needed. These methods should update their results online and return an associated statistical confidence that increases as more data is retrieved from the LSST. It is critical not to miss any relevant event while keeping the contamination from false positives as low as possible. Additionally, these methods should learn from past experience and adapt depending on the previously selected events. Designing methods that comply with these requirements is currently an open problem.

V. Concluding Remarks

In a few years the LSST will be fully operational capturing the light of billions of astronomical objects, and generating approximately two million events each night for ten years. The LSST team itself, and multiple external facilities around the world, will follow and study these events. The main objectives are to characterize and classify the transient phenomena arising from the moving sky. Additionally, it is expected that a plethora of scientific discoveries will be made. If the right tools are used, science would be produced at rates without precedent.

Conventional astronomy is not prepared for this deluge of observational data and hence a paradigm shift in TDA has been observed. Astronomy, statistics and machine learning have been combined in order to produce science that can provide automated methods to deal with the soon to come synoptic surveys. Computational intelligence methods for pattern recognition are essential for the proper exploitation of synoptic surveys, being able to detect and characterize events that otherwise might not even be noticed by human investigators.

In this review we have studied several machine learning based implementations proposed to solve current astronomical problems. The particular challenges faced when applying machine learning methods in TDA include 1) the design of representative training sets, 2) the combination and reuse of training databases for new surveys, 3) the definition of feature vectors from domain knowledge, 4) the design of fast and

[3] NVIDIA Tesla K20 module.

scalable computational implementations of the methods in order to process the TDA databases within feasible times, and finally, 5) the sometimes difficult interpretation of the results obtained and the question of how to gain physical insight from them.

The quality of a training set is critical for the correct performance of supervised methods. In astronomy an intrinsic sample selection bias occurs when knowledge gathered from previous surveys is used with new data. Semi-supervised learning and active learning rise as feasible options to cope with large and heterogeneous astronomical data, providing particular solutions to the dilemmas regarding training sets. It is very likely that we will see more semi-supervised applications for astronomy in the near future. The reuse of training sets is critical in terms of scalability and validity of results. The integration with existing databases and the incorporation of data observed at different wavelengths are currently open issues. Feature spaces that are survey-independent may provide an indirect solution to the combination of training sets and the applications of trained classifiers across different surveys.

Although powerful, the sometimes extended calibration required by machine learning methods can be difficult for inexperienced users. The selection of the algorithms, the complexity of the implementations, the exploration of parameter space, and the interpretation of the outputs in physical terms are some of the issues one has to face when using machine learning methods. The learning curve might be too steep for an astronomer to take the initiative, but all the issues named here can be solved by inter-disciplinary collaboration. Teams assembled from the fields of astronomy, statistics, computer science and engineering have everything that is needed to propose solutions for data-intensive TDA. The deluge of astronomical data opens up huge opportunities for professionals with knowledge in computational intelligence and machine learning.

VI. Acknowledgment

This work was funded by CONICYT-CHILE under grant FONDECYT 1110701 and 1140816, and its Doctorate Scholarship program. Pablo Estévez acknowledges support from the Ministry of Economy, Development, and Tourism's Millennium Science Initiative through grant IC12009, awarded to The Millennium Institute of Astrophysics, MAS.

References

[1] C. Alcock, R. A. Allsman, D. R. Alves, T. S. Axelrod, A. C. Becker, D. P. Bennett, K. H. Cook, N. Dalal, A. J. Drake, K. C. Freeman, M. Geha, K. Griest, M. J. Lehner, S. L. Marshall, D. Minniti, C. A. Nelson, B. A. Peterson, P. Popowski, M. R. Pratt, P. J. Quinn, C. W. Stubbs, W. Sutherland, A. B. Tomaney, T. Vandehei, and D. Welch, "The MACHO project: Microlensing results from 5.7 years of LMC observations," *Astrophys. J.*, vol. 542, no. 1, pp. 281–307, 2000.

[2] Y. R. Rahal, C. Afonso, J.-N. Albert, J. Andersen, R. Ansari, E. Aubourg, P. Bareyre, J.-P. Beaulieu, X. Charlot, F. Couchot, C. Coutures, F. Derue, R. Ferlet, P. Fouqué, J.-F. Glicenstein, B. Goldman, A. Gould, D. Graff, M. Gros, J. Haissinski, C. Hamadache, J. de Kat, E. Lesquoy, C. Loup, L. L. Guillou, C. Magneville, B. Mansoux, J.-B. Marquette, E. Maurice, A. Maury, A. Milsztajn, M. Moniez, N. Palanque-Delabrouille, O. Perdereau, S. Rahvar, J. Rich, M. Spiro, P. Tisserand, A. Vidal-Madjar, "The EROS-2 search for microlensing events towards the spiral arms: The complete seven season results," *Astron. Astrophys.*, vol. 500, no. 3, pp. 1027–1044, 2009.

[3] A. Udalski, M. Kubiak, and M. Szymanski, "Optical gravitational lensing experiment. OGLE-II—The second phase of the OGLE project," *Acta Astron.*, vol. 47, pp. 319–344, Oct. 1997.

[4] D. G. York, J. Adelman, J. E. Anderson, Jr., S. F. Anderson, J. Annis, N. A. Bahcall, J. A. Bakken, R. Barkhouser, S. Bastian, E. Berman, W. N. Boroski, S. Bracker, C. Briegel, J. W. Briggs, J. Brinkmann, R. Brunner, S. Burles, L. Carey, M. A. Carr, F. J. Castander, B. Chen, P. L. Colestock, A. J. Connolly, J. H. Crocker, I. Csabai, P. C. Czarapata, J. E. Davis, M. Doi1, T. Dombeck, D. Eisenstein, N. Ellman, B. R. Elms, M. L. Evans, X. Fan, G. R. Federwitz, L. Fiscelli, S. Friedman, J. A. Frieman, M. Fukugita, B. Gillespie, J. E. Gunn, V. K. Gurbani, E. de Haas, M. Haldeman, F. H. Harris, J. Hayes, T. M. Heckman, G. S. Hennessy, R. B. Hindsley, S. Holm, D. J. Holmgren, C.-H. Huang, C. Hull, D. Husby, S.-I. Ichikawa, T. Ichikawa, Ž. Ivezić, S. Kent, R. S. J. Kim, E. Kinney, M. Klaene, A. N. Kleinman, S. Kleinman, G. R. Knapp, J. Korienek, R. G. Kron, P. Z. Kunszt, D. Q. Lamb, B. Lee, R. F. Leger, S. Limmongkol, C. Lindenmeyer, D. C. Long, C. Loomis, J. Loveday, R. Lucinio, R. H. Lupton, B. MacKinnon, E. J. Mannery, P. M. Mantsch, B. Margon, P. McGehee, T. A. McKay, A. Meiksin, A. Merelli, D. G. Monet, J. A. Munn, V. K. Narayanan, T. Nash, E. Neilsen, R. Neswold, H. J. Newberg, R. C. Nichol, T. Nicinski, M. Nonino, N. Okada, S. Okamura, J. P. Ostriker, R. Owen, A. G. Pauls, J. Peoples, R. L. Peterson, D. Petravick, J. R. Pier, A. Pope, R. Pordes, A. Prosapio, R. Rechenmacher, T. R. Quinn, G. T. Richards, M. W. Richmond, C. H. Rivetta, C. M. Rockosi, K. Ruthmansdorfer, D. Sandford, D. J. Schlegel, D. P. Schneider, M. Sekiguchi, G. Sergey, K. Shimasaku, W. A. Siegmund, S. Smee, J. A. Smith, S. Snedden, R. Stone, C. Stoughton, M. A. Strauss, C. Stubbs, M. SubbaRao, A. S. Szalay, I. Szapudi, G. P. Szokoly, A. R. Thakar, C. Tremonti, D. L. Tucker, A. Uomoto, D. V. Berk, M. S. Vogeley, P. Waddell, S.-i. Wang, M. Watanabe, D. H. Weinberg, B. Yanny, and N. Yasuda, "The sloan digital sky survey: Technical summary," *Astron. J.*, vol. 120, no. 3, pp. 1579–1587, 2000.

[5] N. Kaiser, H. Aussel, B. E. Burke, H. Boesgaard, K. Chambers, M. R. Chun, J. N. Heasley, K.-W. Hodapp, B. Hunt, R. Jedicke, D. Jewitt, R. Kudritzki, G. A. Luppino, M. Maberry, E. Magnier, D. G. Monet, P. M. Onaka, A. J. Pickles, Rhoads, H. H. Pui, T. Simon, A. Szalay, I. Szapudi, D. J. Tholen, J. L. Tonry, M. Waterson, and J. Wick, "Pan-STARRS: A large synoptic survey telescope array," in *Proc. Society Photo-Optical Instrumentation Engineers Conf. Series*, 2002, vol. 4836, pp. 154–164.

[6] S. Larson, E. Beshore, R. Hill, E. Christensen, D. McLean, S. Kolar, R. McNaught, and G. Garradd, "The CSS and SSS NEO surveys," in *Proc. AAS/Division for Planetary Sciences Meeting Abstracts #35*, Bulletin of the American Astronomical Society, May 2003, vol. 35, p. 982.

[7] J. Tyson and K. Borne, "Future sky surveys, new discovery frontiers," in *Advances in Machine Learning and Data Mining for Astronomy*, M. Way, J. D. Scargle, K. Ali, and A. Srivastava, Eds. Boca Raton, FL: CRC Press, 2012, ch. 9, pp. 161–181.

[8] Z. Ivezic, J. A. Tyson, E. Acosta, R. Allsman, S. F. Anderson, J. Andrew, R. Angel, T. Axelrod, J. D. Barr, A. C. Becker, J. Becla, C. Beldica, R. D. Blandford, J. S. Bloom, K. Borne, W. N. Brandt, M. E. Brown, J. S. Bullock, D. L. Burke, S. Chandrasekharan, S. Chesley, C. F. Claver, A. Connolly, K. H. Cook, A. Cooray, K. R. Covey, C. Cribbs, R. Cutri, G. Daues, F. Delgado, H. Ferguson, E. Gawiser, J. C. Geary, P. Gee, M. Geha, R. R. Gibson, D. K. Gilmore, W. J. Gressler, C. Hogan, M. E. Huffer, S. H. Jacoby, B. Jain, J. G. Jernigan, R. L. Jones, M. Juric, S. M. Kahn, J. S. Kalirai, J. P. Kantor, R. Kessler, D. Kirkby, L. Knox, V. L. Krabbendam, S. Krughoff, S. Kulkarni, R. Lambert, D. Levine, M. Liang, K-T. Lim, R. H. Lupton, P. Marshall, S. Marshall, M. May, M. Miller, D. J. Mills, D. G. Monet, D. R. Neill, M. Nordby, P. O'Connor, J. Oliver, S. S. Olivier, K. Olsen, R. E. Owen, J. R. Peterson, C. E. Petry, F. Pierfederici, S. Pietrowicz, R. Pike, P. A. Pinto, R. Plante, V. Radeka, A. Rasmussen, S. T. Ridgway, W. Rosing, A. Saha, T. L. Schalk, R. H. Schindler, D. P. Schneider, G. Schumacher, J. Sebag, L. G. Seppala, I. Shipsey, N. Silvestri, J. A. Smith, R. C. Smith, M. A. Strauss, C. W. Stubbs, D. Sweeney, A. Szalay, J. J. Thaler, D. V. Berk, L. Walkowicz, M. Warner, B. Willman, D. Wittman, S. C. Wolff, W. M. Wood-Vasey, P. Yoachim, and H. Zhan. (June 2011). LSST: From science drivers to reference design and anticipated data products. ArXiv e-prints. [Online]. Available: http://www.lsst.org/lsst/overview/

[9] LSST Science Collaborations and LSST Project 2009. (2013). LSST science book. Version 2.0, arXiv:0912.0201. [Online]. Available: http://www.lst.org/lsst/scibook

[10] K. Borne, "Virtual observatories, data mining and astroinformatics," in *Planets, Stars and Stellar Systems. Astronomical Techniques, Software, and Data*, vol. 2, T. Oswalt and H. Bond, Eds. New York: Wiley, 2013, pp. 404–443.

[11] K. Borne. (2012). Learning from big data in astronomy—An overview. [Online]. Available: http://www.samsi.info/sites/default/files/Borne_september2012.pdf

[12] E. Feigelson and G. Babu, "Big data in astronomy," *Significance*, vol. 9, no. 4, pp. 22–25, 2012.

[13] K. Borne, "Astroinformatics: Data-oriented astronomy research and education," *J. Earth Sci. Inform.*, vol. 3, nos. 1–2, pp. 5–17, 2010.

[14] M. Petit, *Variable Stars*. Reading, MA: New York: Wiley, 1987.

[15] J. Percy, *Understanding Variable Stars*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[16] L. Eyer and N. Mowlavi, "Variable stars across the observational HR diagram," *J. Phys.: Conf. Ser.*, vol. 118, no. 1, p. 012010, 2008.

[17] N. Samus, O. V. Durlevich, E. V. Kazarovets, N. N. Kireeva, E. N. Pastukhova, and A. V. Zharova, "General catalogue of variable stars," (Samus+ 2007-2012), VizieR On-line Data Catalog: B/gcvs, 2012.

[18] A. Walker, “Distances to local group galaxies,” in *Stellar Candles for the Extragalactic Distance Scale* (Lecture Notes in Physics, vol. 635), D. Alloin and W. Gieren, Eds. Berlin Heidelberg, Germany: Springer-Verlag, 2003, pp. 265–279.

[19] F. E. Olivares and M. Hamuy, *Core-Collapse Supernovae As Standard Candles*. Germany: Lambert Academic Publishing, 2011.

[20] D. Popper, “Stellar masses,” *Annu. Rev. Astron. Astrophys.*, vol. 18, no. 1, pp. 115–164, 1980.

[21] P. Huijse, P. A. Estevez, P. Protopapas, P. Zegers, and J. C. Principe, “An information theoretic algorithm for finding periodicities in stellar light curves,” *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5135–5145, 2012.

[22] J. Principe, *Information Theoretic Learning: Renyi’s Entropy and Kernel Perspectives*. New York: Springer-Verlag, 2010.

[23] J. Scargle, “Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data,” *Astrophys. J.*, vol. 263, pp. 835–853, Dec. 1982.

[24] P. Protopapas, P. Huijse, P. A. Estévez, J. Principe, and P. Zegers, “A novel, fully automated pipeline for period estimation in the EROS-2 data set,” *Astrophys. J.*, 2014, to be published.

[25] M. J. Graham, A. J. Drake, S. G. Djorgovski, A. A. Mahabal, and C. Donalek, “Using conditional entropy to identify periodicity,” *Mon. Not. Roy. Astron. Soc.*, vol. 434, pp. 2629–2635, Sept. 2013.

[26] R. Tagliaferri, A. Ciaramella, L. Milano, F. Barone, and G. Longo, “Spectral analysis of stellar light curves by means of neural networks,” *Astron. Astrophys. Suppl. Ser.*, vol. 137, no. 2, pp. 391–405, 1999.

[27] M. J. Graham, A. J. Drake, S. G. Djorgovski, A. A. Mahabal, C. Donalek, V. Duan, and A. Maher, “A comparison of period finding algorithms,” *Mon. Not. Roy. Astron. Soc.*, vol. 434, pp. 3423–3444, Oct. 2013.

[28] J. Debosscher, L. M. Sarro, C. Aerts, J. Cuypers, B. Vandenbussche, R. Garrido, and E. Solano, “Automated supervised classification of variable stars. I. Methodology,” *Astron. Astrophys.*, vol. 475, no. 3, pp. 1159–1183, 2007.

[29] J. Blomme, L. M. Sarro, F. T. O’Donovan, J. Debosscher, T. Brown, M. Lopez, P. Dubath, L. Rimoldini, D. Charbonneau, E. Dunham, G. Mandushev, D. R. Ciardi, J. de Ridder, and C. Aerts, “Improved methodology for the automated classification of periodic variable stars,” *Mon. Not. Roy. Astron. Soc.*, vol. 418, no. 1, pp. 96–106, 2011.

[30] G. Wachman, R. Khardon, P. Protopapas, and C. Alcock, “Kernels for periodic time series arising in astronomy,” in *Proc. European Conf. Machine Learning Knowledge Discovery Databases: Part II, (Bled, Slovenia)*, 2009, pp. 489–505.

[31] P. Dubath, L. Rimoldini, M. Suveges, J. Blomme, M. Lopez, L. M. Sarro, J. de Ridder, J. Cuypers, L. Guy, I. Lecoeur, K. Nienartowicz, A. Jan, M. Beck, N. Mowlavi, P. de Cat, T. Lebzelter, and L. Eyer, “Random forest automated supervised classification of hipparcos periodic variable stars,” *Mon. Not. Roy. Astron. Soc.*, vol. 414, no. 3, pp. 2602–2617, 2011.

[32] J. Long, J. Bloom, N. Karoui, J. Rice, and J. Richards, “Classification of poorly time sampled light curves of periodic variable stars,” in *Astrostatistics and Data Mining* (Springer Series in Astrostatistics, vol. 2), L. M. Sarro, L. Eyer, W. O’Mullane, and J. de Ridder, Eds. New York: Springer-Verlag, 2012, pp. 163–171.

[33] D.-W. Kim, P. Protopapas, Y.-I. Byun, C. Alcock, R. Khardon, and M. Trichas, “Quasi-stellar object selection algorithm using time variability and machine learning: Selection of 1620 quasi-stellar object candidates from MACHO large magellanic cloud database,” *Astrophys. J.*, vol. 735, p. 68, July 2011.

[34] J. W. Richards, D. L. Starr, N. R. Butler, J. S. Bloom, J. M. Brewer, A. Crellin-Quick, J. Higgins, R. Kennedy, and M. Rischard, “On machine-learned classification of variable stars with sparse and noisy time-series data,” *Astrophys. J.*, vol. 733, no. 1, p. 10, 2011.

[35] K. Pichara and P. Protopapas, “Automatic classification of variable stars in catalogs with missing data,” *Astrophys. J.*, vol. 777, no. 2, p. 83, 2013.

[36] M. Hernandez-Pajares and J. Floris, “Classification of the HIPPARCOS input catalogue using the Kohonen network,” *Mon. Not. Roy. Astron. Soc.*, vol. 268, p. 444, May 1994.

[37] D. R. Brett, R. G. West, and P. J. Wheatley, “The automated classification of astronomical light curves using Kohonen self-organizing maps,” *Mon. Not. Roy. Astron. Soc.*, vol. 353, pp. 369–376, Sept. 2004.

[38] L. M. Sarro, J. Debosscher, C. Aerts, and M. López, “Comparative clustering analysis of variable stars in the Hipparcos, OGLE Large Magellanic Cloud, and CoRoT exoplanet databases,” *Astron. Astrophys.*, vol. 506, pp. 535–568, Oct. 2009.

[39] A. Torralba, R. Fergus, and W. Freeman, “80 million tiny images: A large data set for nonparametric object and scene recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, pp. 1958–1970, Nov. 2008.

[40] C. Varón, C. Alzate, J. A. K. Suykens, and J. Debosscher, “Kernel spectral clustering of time series in the CoRoT exoplanet database,” *Astron. Astrophys.*, vol. 531, p. A156, July 2011.

[41] A. C. Fabian, “Serendipity in astronomy,” *ArXiv e-prints*, Aug. 2009.

[42] P. Protopapas, J. M. Giammarco, L. Faccioli, M. F. Struble, R. Dave, and C. Alcock, “Finding outlier light curves in catalogues of periodic variable stars,” *Mon. Not. Roy. Astron. Soc.*, vol. 369, no. 2, pp. 677–696, 2006.

[43] D. Yankov, E. Keogh, and U. Rebbapragada, “Disk aware discord discovery: Finding unusual time series in terabyte sized datasets,” *Knowl. Inform. Syst.*, vol. 17, no. 2, pp. 241–262, 2008.

[44] U. Rebbapragada, P. Protopapas, C. E. Brodley, and C. Alcock, “Finding anomalous periodic time series: An application to catalogs of periodic variable stars,” *Mach. Learn.*, vol. 74, pp. 281–313, May 2009.

[45] J. W. Richards, D. Homrighausen, P. E. Freeman, C. M. Schafer, and D. Poznanski, “Semi-supervised learning for photometric supernova classification,” *Mon. Not. Roy. Astron. Soc.*, vol. 419, pp. 1121–1135, Jan. 2012.

[46] J. W. Richards, D. L. Starr, H. Brink, A. A. Miller, J. S. Bloom, N. R. Butler, J. B. James, J. P. Long, and J. Rice, “Active learning to overcome sample selection bias: Application to photometric variable star classification,” *Astrophys. J.*, vol. 744, p. 192, Jan. 2012.

[47] C. Donalek, S. G. Djorgovski, A. A. Mahabal, M. J. Graham, A. J. Drake, T. J. Fuchs, M. J. Turmon, A. A. Kumar, N. S. Philip, M.T.-C. Yang, and G. Longo, “Feature selection strategies for classifying high dimensional astronomical data sets,” in *Proc. IEEE Int. Conf. Big Data*, Oct. 2013, pp. 35–41.

[48] J. Vergara and P. A. Estévez, “A review of feature selection methods based on mutual information,” *Neural Comput. Applicat.*, vol. 24, no. 1, pp. 175–186, 2014.

[49] S. G. Djorgovski, C. Donalek, A. Mahabal, R. Williams, A. Drake, M. Graham, and E. Glikman, “Some pattern recognition challenges in data-intensive astronomy,” in *Proc. 18th Int. Conf. Pattern Recognition*, 2006, p. 865.

[50] M. J. Graham, S. G. Djorgovski, A. Mahabal, C. Donalek, A. Drake, G. Longo, “Data challenges of time domain astronomy,” *Distrib. Parallel Databases*, vol. 30, no. 5, pp. 371–384, 2012.

[51] S. G. Djorgovski, C. Donalek, A. Mahabal, B. Moghaddam, M. Turmon, M. J. Graham, A. J. Drake, N. Sharma, and Y. Chen, “Towards an automated classification of transient events in synoptic sky surveys,” in *Proc. CIDU*, NASA Ames, 2011, pp. 174–188.

[52] W. B. March, A. Ozakin, D. Lee, R. Riegel, and A. G. Gray, “Multitree algorithms for large-scale astrostatistics,” in *Advances in Machine Learning and Data Mining for Astronomy*, M. J. Way, J. D. Scargle, K. M. Ali, and A. N. Srivastava, Eds. Boca Raton, FL: CRC Press, 2012, pp. 463–483.

[53] G. V. Kulkarni, R. C. Nichol, R. K. Sheth, H.-J. Seo, D. J. Eisenstein, and A. Gray, “The three-point correlation function of luminous red galaxies in the sloan digital sky survey,” *Mon. Not. Roy. Astron. Soc.*, vol. 378, pp. 1196–1206, July 2007.

[54] Y. Bengio and Y. LeCun, “Scaling learning algorithms towards AI,” in *Large Scale Kernel Machines*, L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, Eds. Cambridge, MA: MIT Press, 2007.

[55] G. E. Hinton and S. Osindero, “A fast learning algorithm for deep belief nets,” *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[56] J. Fan, F. Han, and H. Liu, “Challenges of big data analysis,” *ArXiv e-prints*, Aug. 2013.

[57] K. Das and K. Bhaduri, “Parallel and distributed data mining for astronomy applications,” in *Advances in Machine Learning and Data Mining for Astronomy*, M. J. Way, J. D. Scargle, K. M. Ali, and A. N. Srivastava, Eds. Boca Raton, FL: CRC Press, 2012, pp. 595–615.

[58] V. Kindratenko and P. Trancoso, “Trends in high-performance computing,” *Comput. Sci. Eng.*, vol. 13, no. 3, pp. 92–95, 2011.

[59] D. Sart, A. Mueen, W. Najjar, E. Keogh, and V. Niennattrakul, “Accelerating dynamic time warping subsequence search with GPUs and FPGAs,” in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 1001–1006.

[60] J. Dean and S. Ghemawat, “Mapreduce: Simplified data processing on large clusters,” in *Proc. 6th Conf. Symp. Opearting Systems Design Implementation*, USENIX Association, Berkeley, CA, 2004, vol. 6, pp. 10–10.

[61] C. T. Chu, S. K. Kim, Y. A. Lin, Y. Yu, G. R. Bradski, A. Y. Ng, and K. Olukotun, “Map-reduce for machine learning on multicore,” in *Neural Information Processing Systems*, B. Schölkopf, J. C. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2006, pp. 281–288.

[62] G. B. Berriman, E. Deelman, G. Juve, M. Regelson, and P. Plavchan, “The application of cloud computing to astronomy: A study of cost and performance,” *ArXiv e-prints*, Oct. 2010.

[63] K. Wiley, A. Connolly, J. Gardner, S. Krughoff, M. Balazinska, B. Howe, Y. Kwon, and Y. Bu, “Astronomy in the cloud: Using MapReduce for image co-addition,” *Publications Astron. Soc. Pacific*, vol. 123, pp. 366–380, Mar. 2011.

[64] W. Pedrycz, A. Skowron, and V. Kreinovich, *Handbook of Granular Computing*. New York: Wiley, 2008.

[65] G. Greene, T. Budavari, N. Li, M. Nieto-Santisteban, A. Szalay, and A. Thakar, “Chapter 13: Web-based tools—Open SkyQuery: Distributed database queries and crossmatching,” in *Astronomical Society of the Pacific Conference Series*, vol. 382, M. J. Graham, M. J. Fitzpatrick, and T. A. McGlynn, Eds. San Francisco, CA: Astron. Soc. Pacific, 2007, p. 111.