

Unsolved problems in observational astronomy. II. Focus on rapid response – mining the sky with “thinking” telescopes

W.T. VESTRAND, J. THEILER, and P.R. WOZNIAK

Los Alamos National Laboratory, MS B244, Los Alamos, New Mexico, USA

Received 10 September 2004; accepted 20 September 2004; published online 31 October 2004

Abstract. The existence of rapidly slewing robotic telescopes and fast alert distribution via the Internet is revolutionizing our capability to study the physics of fast astrophysical transients. But the salient challenge that optical time domain surveys must conquer is mining the torrent of data to recognize important transients in a scene full of normal variations. Humans simply do not have the attention span, memory, or reaction time required to recognize fast transients and rapidly respond. Autonomous robotic instrumentation with the ability to extract pertinent information from the data stream in real time will therefore be essential for recognizing transients and commanding rapid follow-up observations while the ephemeral behavior is still present. Here we discuss how the development and integration of three technologies: (1) robotic telescope networks; (2) machine learning; and (3) advanced database technology, can enable the construction of smart robotic telescopes, which we loosely call “thinking” telescopes, capable of mining the sky in real time.

Key words: instrumentation:miscellaneous—methods:observational—surveys

©2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

1. Introduction

The global time variability of the night sky is essentially unknown (e.g. Paczynski 2000). Nevertheless, limited surveys of time variability have already had profound scientific impact by revealing the existence of dark energy in the universe (e.g. Riess et al. 1998, Perlmutter et al. 1999). That discovery has led the world’s leading scientific research organizations to plan major investments in extensive variability surveys that will revolutionize time domain astronomy. But the unsolved problem for these giant surveys is how to mine the torrent of data—which for the Large Synoptic Survey Telescope (LSST) will be a 2 Giga-pixel image every 10 seconds (Tyson 2002)—and to recognize important variations in a scene full of normal variations and command follow-up observations in real time. The surveys threaten to drown us in a flood of data, but leave us thirsty for knowledge.

The knowledge extraction and discovery techniques employed in astronomy have not progressed very far from those employed by Tycho Brahe when, on 11 November 1572, he started modern time domain astronomy by discovering a bright new star that was not in his “mental catalog” of the night sky. Typically humans still screen the reduced data even from robotic telescopes and, based on their knowledge and

memory, identify candidates for follow-up observations. But modern data sets are simply becoming too large. Recognition of ephemeral changes of persistent sources in huge data streams and identification of fast celestial transients in the forest of non-celestial transients cannot be left to human analysts. Humans simply do not have the attention span, memory, or reaction time required to monitor huge volumes of data, recognize the important variations, and promptly respond with follow-up observations. The ability of modern instrumentation to collect data at dazzling rates has pushed knowledge extraction in astronomy to a tipping point. The *process* of discovery must fundamentally change. In this paper, we argue that the solution is the integration of robotic telescopes with modern artificial intelligence techniques to construct discovery engines that are capable of autonomously mining the sky in real time.

2. “Thinking” telescopes

The human brain, through a process we loosely call thinking, integrates data collection, pattern recognition, object classification, and memory to obtain a higher understanding of what action needs to be taken and promptly takes action to respond to a threat or an opportunity. To mine the night sky effectively in the new era of time domain astronomy, we must construct

Correspondence to: vestrand@nis.lanl.gov

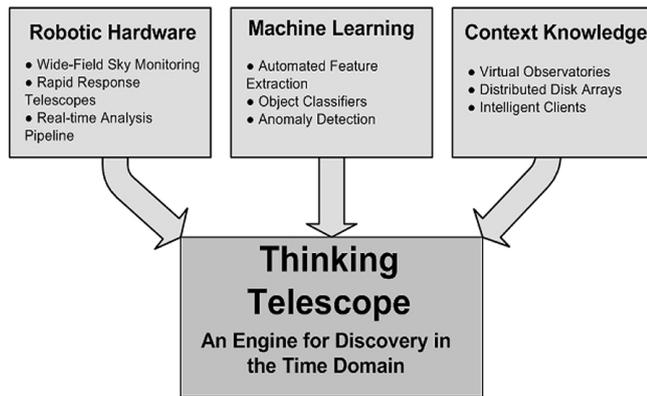


Fig. 1. Three technologies must be integrated in order to create autonomous robotic telescope systems capable of finding and making more detailed follow-up observations of ephemeral source anomalies in real time.

“thinking” robotic telescopes. These next generation robotic systems will not only have to be autonomous, they must also have a continuously evolving knowledge of normal behavior and be capable of recognizing subtle anomalies in a torrent of data. They must respond in real time by formulating queries and priorities, by commanding follow-up observations, and by learning to optimize the response. Just as the Internet and search engines have revolutionized the way we search for and collect information, discovery in time domain astronomy will rely on advanced data mining, visualization tools, and smart algorithms to automate the extraction of knowledge from the observations in real time.

The technological building blocks for constructing robotic thinking systems exist: autonomous data collection, robotic hardware control, database construction and query, pattern recognition, object classification, and other forms of automated knowledge extraction. The challenge is to integrate these building blocks into autonomous thinking systems for observational science. To construct thinking telescopes for astronomy, we must develop and integrate technology in three key areas (see Fig. 1): (1) distributed networks of robotic monitoring and response telescopes; (2) machine learning techniques for automated knowledge extraction in real time; and (3) virtual observatories employing advanced database technology that provide context information in real time. Here we discuss briefly the application of these three enabling technologies to robotic astronomy.

3. Networks of robotic telescopes

Networks of autonomous robotic response telescopes already exist. The best known example is the informal world-wide network of ground-based telescopes that promptly respond to real time Gamma Ray Burst (GRB) alerts generated by High-Energy satellites and distributed by the GRB Coordinates Network (GCN; Barthelmy et al. 2000). Here the network operates in an “open loop” manner. The network telescopes receive the alert and respond, but all the real time information flows in one direction.

Effective deployment of thinking technology in robotic telescope system will require the capability to operate in a “closed loop” manner. In this mode of operation, the robotic network telescopes not only respond to alerts, but also extract information from the observations in real time and autonomously send it back to the network. Given the feedback of knowledge from other instruments, autonomous decision units (either central or at each telescope) can then command a modified response to optimize the information extraction (e.g. change exposure cadence, change filters or pointing direction, etc.). This feedback of real time information allows a distributed network of telescopes to perform more efficiently than as individual instruments.

An example of a first generation robotic telescope network operating as a closed loop system is provided by the RAPTOR (RAPid Telescopes for Optical Response) system at Los Alamos National Laboratory (Vestrand et al. 2002). RAPTOR links together eleven small telescopes (eight wide-field monitoring and three narrow-field response telescopes) located at two spatially separated sites to act as an autonomous system capable of finding optical transients and making follow-up observations in real time. Four wide-field telescopes at each site simultaneously mosaic the same 1300 square-degree patch of sky. All of the telescopes in the network have dedicated control computers that run a photometry/astrometry pipeline capable of identifying optical transients in real time. The sky-monitoring task is therefore divided into pieces and distributed to a network of telescopes that run in parallel. Network intercommunication synchronizes the operation of the telescopes and passes information about the detection of transients back to the network server. A central decision unit then decides if an alert should be generated and, if the answer is yes, distributes the alert back to the telescope network and closes the information loop by commanding real time follow-up observations of the optical transient.

Intercommunication between telescopes is the key to reliable real time identification of celestial optical transients by the RAPTOR system. To identify celestial transients robustly, the triggering software requires: (1) simultaneous identification of the transient by both spatially separated telescope arrays and (2) the absence of a measurable parallax. The telescope arrays are separated by 38 kilometers and the wide-field imagers have a single pixel resolution of 34 arcseconds. So any non-celestial transient generated out to the distance of the moon will have a parallax that is measurable by the wide-field monitors. This no-parallax requirement reduces the number of false triggers by several orders of magnitude. Real time intercommunication therefore makes the distributed network much more powerful than the sum of its parts.

Spatially distributed robotic telescopes linked together with Internet communication can therefore operate in a manner similar to the GRID paradigm that is employed for distributed high-performance computing (Foster et al. 2001). Infrastructure software that will allow the formation of intercommunicating global networks of autonomous robotic telescopes is already being developed (e.g. White et al. 2004,

Allan et al. 2004). This networking will allow alerts to be passed among the telescopes and, when coupled with intelligent client software at the network telescopes, enable each telescope to configure a real time response based upon its capabilities, schedule, and priorities.

By closed-loop networking of even small aperture monitoring telescopes with rapid response telescopes into a worldwide distributed system, one can expect to extract a broad range of discoveries. Such a system could conduct the first comprehensive global census of stellar flaring, find unpredicted close encounters with nearby solar system objects, search for "orphan" gamma-ray bursts, discover novae, and find the nearby supernovae needed to calibrate observational estimates of dark energy in the universe (see, e.g. Paczynski et al. 2001).

4. Machine Learning

Even without "thinking", distributed networks of rapidly responding robotic telescopes can produce important science. But the full power of robotic systems will only be unleashed if we build them with an ability to recognize not only transients but also important variations in persistent sources.

An important step in this direction has already been taken by the microlensing surveys. The OGLE and Macho projects, for example, developed a real time alert system capable of identifying on-going microlensing events with minimal human screening (e.g. Udalski et al. 1994, Alcock et al. 1997, Udalski 2003). These systems look for the unique lightcurve shape and achromaticity indicative of microlensing events. But the understanding of most astronomical objects is too incomplete to predict the properties of important changes.

An alternative approach is to create a record of the observed variations for every source and to employ Machine Learning (ML) techniques to train the system to recognize important changes as they are happening. ML techniques have been developed for knowledge extraction both from images and time series data. But can one automate some of those ML techniques and integrate them with robotic telescopes to enable real-time follow-up observations? Speed, sensitivity, and the suppression of false positives are important challenges.

Various real world artifacts, from airplane lights and clouds to sensitivity variations for individual pixels and other non-celestial phenomena that cannot be predicted, will inevitably contaminate the observations. Automated identification of these unwanted artifacts is essential for the construction of autonomous robotic systems. But experience has shown us that there is a broad continuum of extraneous artifacts and it is very difficult to deal with them on a piecemeal basis with hard-wired code. The advantage of ML is that it can provide techniques that allow the system to be trained to identify anomalies and artifacts, and learn to adapt, based on actual in-the-field data. This approach can enable the robotic system to become smarter and to continuously improve the overall quality of the data collection and system response.

Another significant advantage of the ML approach is that it can provide software with an interface that allows as-

tronomers to interact with the autonomous telescope system and to construct queries such as "find more like this" or if you ever detect something like "this" notify me and make it a high priority for prompt follow-up observations. The robotic system can then act as an evolving search engine continuously monitoring a dynamic database—the night sky—and responding when a new observing opportunity arises.

Here we briefly discuss three broad classes of Machine Learning applications that are likely play an important role in robotic astronomy.

4.1. Machine Learning and image operators

Construction of thinking astronomical telescopes will require the development of flexible software capable of rapid and reliable pattern recognition in imaging data. The evolution of living systems has inspired a technique called genetic programming for developing optimal code by competing a population of individuals through successive reproduction with modification followed by selection based on a fitness measure (Fogel et al. 1966). A software package called GENIE (Genetic Imagery Exploration; Theiler et al. 1999) uses this genetic programming technique to construct imaging processing tools for feature extraction from low-level image operators (e.g. dilation, erosion, smoothing, sharpening, etc.). GENIE generates a population of candidate tools, ranks them according to performance on training data, and the highest ranked tools are permitted to reproduce. The process continues until the population converges on an optimal solution or the user accepts the solution and stops the evolution. The user is also able to direct the evolution by modifying or supplementing the training data. The GENIE package has been shown to quite successful in the development of tools for efficient automated feature recognition in remote sensing applications ranging from the identification of craters on Mars (Plesko et al. 2002) to the identification of cancerous cells in bio-medical imaging (Harvey et al. 2003). We believe that the GENIE approach has great potential for developing efficient algorithms for real time feature extraction and classification in astronomical images.

The true utility of the machine learning techniques employed by GENIE is their flexibility. Traditional approaches require the writing of specialized code for each type of feature that one wants to find and entails careful specification of the features as well as a substantial amount of trial and error when applied to real images. In the ML approach one shows the machine what to find and the software derives classification algorithms directly from examples in real data. So when new classes of objects are identified or instrumental artifacts emerge, one can rapidly evolve new code for identifying them.

These ML techniques will provide powerful tools for developing a flexible pipeline capable of change detection in astronomical images and automated classification of the detected changes in real time. For example, one could apply the Alard-Lupton PSF-matching algorithm (Alard and Lupton 1998) with a spatially varying convolution kernel to difference the new image with an earlier reference image. An

ML program like GENIE could then be used to develop real time classification of the detected changes. In an autonomous robotic system, these classified changes would then be fed back to the central decision unit to determine their priority for follow-up and command the response if action is needed. The changes could be generated by instrumental problems (camera noise, focus, clouds, etc.) that require action to maintain data quality or real astrophysical transients that require alert generation and real time follow-up.

4.2. Machine Learning and object classification

The sky is full of varying celestial sources, so another challenge to synoptic time domain surveys will be the classification of variable objects. Even a shallow survey with a limiting magnitude of 16th would detect about 100,000 variable stars in the full sky. If one allocates only two minutes for an analyst to load the data and manually classify each variable, classifying the entire sample would take a human analyst, working 40-hour weeks, nearly two years. It is just not practical for human analysts to classify the variables in samples that large.

Development of accurate Machine Learning algorithms for classification of astronomical objects will be essential for efficient mining of the next generation of time domain surveys. There are two basic ML approaches that are employed by automated machine classifiers – supervised learning and unsupervised learning. Supervised techniques require a training sample composed of examples with known class membership produce a classifier that can associate these class labels to data not in the training set. Unsupervised techniques operate without information about class membership and cluster the data into distinct classes without providing any particular labels for the different classes. Since they utilize prior domain knowledge, supervised classifiers normally outperform unsupervised classifiers. But unsupervised classifiers are capable of finding previously unknown classes of objects.

Both supervised and unsupervised classifiers have been successfully applied to the classification of variable stars based on light curves (e.g. Wozniak et al. 2002, Brett et al. 2004). For example, Support Vector Machines (SVMs; Vapnik 1995), a modern algorithm employing supervised learning can classify variable stars with accuracies of better than 90% – comparable to the level of agreement typically achieved between independent human experts. We expect a significant effort in the application of the broad range of existing machine learning techniques to the classification of astronomical objects as the volume of data provided by giant surveys make automated classification essential.

4.3. Machine Learning and anomaly detection

Particularly interesting targets for real time follow-up observations are anomalies—unusual image change objects or unusual variability patterns that are not among the specified classes. ML theory has developed techniques for anomaly detection that use unlabelled data (from samples of the archive itself) and through an unsupervised learning procedure establish a "simplest" specification of the data (Theiler and Cai

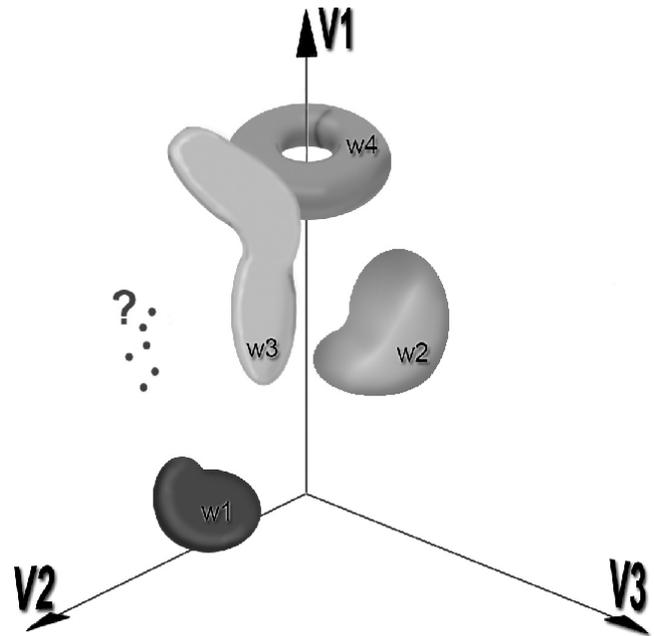


Fig. 2. The clustering of objects in feature space. The question marks indicate the positions of anomalies.

2003). As illustrated in Fig. 2, similar objects cluster in complex regions of n -dimensional parameter space. Objects in the various feature regions can therefore be flagged as instrumental artifacts or as known object families. Data not described by this specification (the question mark points in Fig. 2) are treated as anomalies and identified for follow-up studies to determine their properties.

Since there is no way to define *a priori* what constitutes an interesting anomaly, in the early phases human experts help guide the anomaly detection system toward optimal automatic operation. During training, human experts are offered anomalies by the system and the expert determines whether or not the anomaly is interesting and, if yes, grades the anomaly on importance for follow-up. The system therefore learns to optimize its performance for real time automated operation. Many anomalies will be instrumentation errors; automating their identification will allow the system to recognize problems and take corrective action to maintain data quality. But some of the anomalies will be rare astrophysical objects or events that might otherwise have gone undiscovered.

5. Context, Virtual Observatories, and database technology

Persistent monitoring and construction of a baseline record of temporal variability is essential for distinguishing normal variations from anomalous variations. Without context information the system cannot "think and learn", it can only apply static criteria to newly collected data in the analysis pipeline.

Context information for robotic instrumentation is best provided by a Virtual Observatory (VO) that is constructed from the instrument's own observations. The VO can then

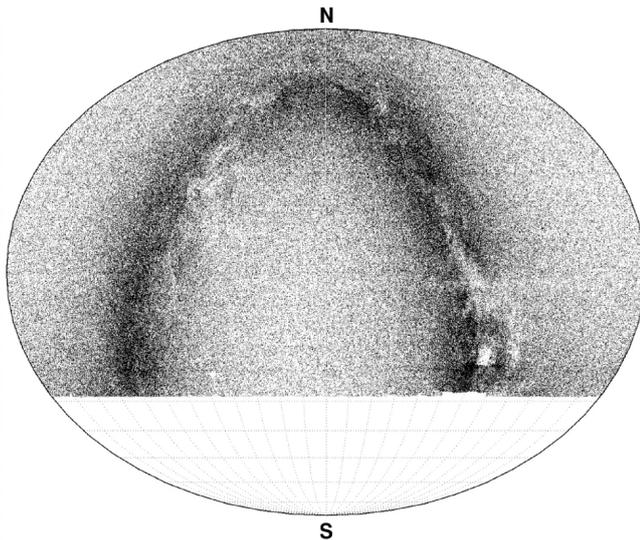


Fig. 3. The positions of objects in the SkyDOT Virtual Observatory brighter than 11^{th} magnitude ($\sim 500,000$ objects) plotted in an equal area projection. Each object has a measured time history composed of typically a few hundred measurements spread over at least a one-year baseline.

store both the measurements and learned experience (derived information and relations found by ML algorithms) as well as act as a platform for mining and knowledge extraction. But to maximize adaptability of the anomaly detection software and scientific utility of the VO, it must have a powerful interface that enables extensive data mining and integration with other VO databases.

Virtual Observatories will be powerful tools for discovery in time domain astronomy (e.g. Brunner et al. 2001). Several time domain projects have already created first generation VOs mining their data. For example, the ASAS (All Sky Automated Survey) project (Pojmanski 2002) project has constructed a VO that allow one to explore the variability of sources in the southern hemisphere. We have constructed a northern hemisphere VO called SkyDOT (Sky Database for Objects in the Time Domain; Wozniak et al. 2002) that contains a full year of our observations for ~ 10 million sources covering the full sky visible from Los Alamos, New Mexico (see Fig. 3). These first generation time domain VOs give us valuable databases to use in the development of ML and anomaly detection algorithms as well as give a reasonable temporal baseline to explore with our new algorithms. But these first generation VOs are static or slowly updated.

To be most effective, the next generation of time-domain VOs will have to be dynamic. By dynamic we mean continuously updatable in real, or nearly real, time. This will allow robotic telescopes and astronomers to determine what other telescopes measured earlier in the day or what other instruments are measuring both in real time and in historical context. This is a challenging goal. It will require the development of new metadata standards for temporal databases and astronomical extensions for database management and query languages.

Another challenge for the implementation of dynamic VOs is the development of new techniques to append the

data in real time while maintaining very efficient access to all the data. Some limited capability for incremental indexing exists but query performance rapidly degrades as data is appended. A related issue is that of data locking and the ability to keep running very large queries while new data is being written. Relational database technology has solved some of those problems, but work remains to be done to ensure that it is scalable to the Petabyte regime.

The massive data volumes generated by the new generation of time domain instrumentation and the need to access that information rapidly, will require the development of new, cost effective, massive storage systems. The price/performance advantages of Beowulf clusters built of commodity components changed the landscape for supercomputing. Due to the development of distributed disk array (DDA) technology and the sharply declining price per Terabyte (TB) of commodity disk drives, we are poised for a similar advance in data storage. Currently (summer 2004), single-node RAID storage systems with 2 TB capacity using commodity hard drives cost of less than \$4K. Those "brick" systems, when used in a parallel cluster environment, can form DDAs with hundreds of TB storage capacity and read/write bandwidths that greatly exceed those of existing Gigabit local and wide-area networking technology. Further, the greater CPU/storage ratio and network integration in a DDA offers the possibility of data analysis techniques not possible with traditional storage approaches.

6. Conclusions

Discovery in astronomy is too important to be left to Astronomers (or their graduate students). The massive data flows being generated by modern instrumentation has pushed human capabilities to their limit, so that the process of discovery must change. The problem is most acute in time domain astronomy, where, to be effective, ephemeral and often subtle changes must be recognized and responded to in real time. The integration of robotic telescope technology, machine learning techniques, and virtual observatory context information to form autonomous robotic telescope networks (see Fig. 4) shows promise for solving the challenge of knowledge extraction in real time. We believe scientific instrumentation is entering an era wherein these technologies will be integrated to be build advanced observing systems, which we call "thinking" systems, that will fundamentally change the process of discovery in observational science.

Acknowledgements. This research was supported by Laboratory Directed Research and Development funds at Los Alamos National Laboratory.

References

- Alcock, C., et al.: 1997, ApJ 491, 436
- Allan, A., et al.: 2004, SPIE 5496, in press
- Alard, C., Lupton, R.H.: 1998, ApJ 503, 325
- Barthelmy, S.D., et al.: 2000, Gamma Ray Bursts: 5th Huntsville Symposium, ed. R.M. Kippen et al., AIP: New York, 731
- Brett, D.R., West, R.G., Wheatley, P.J.: 2004, MNRAS 353, 369

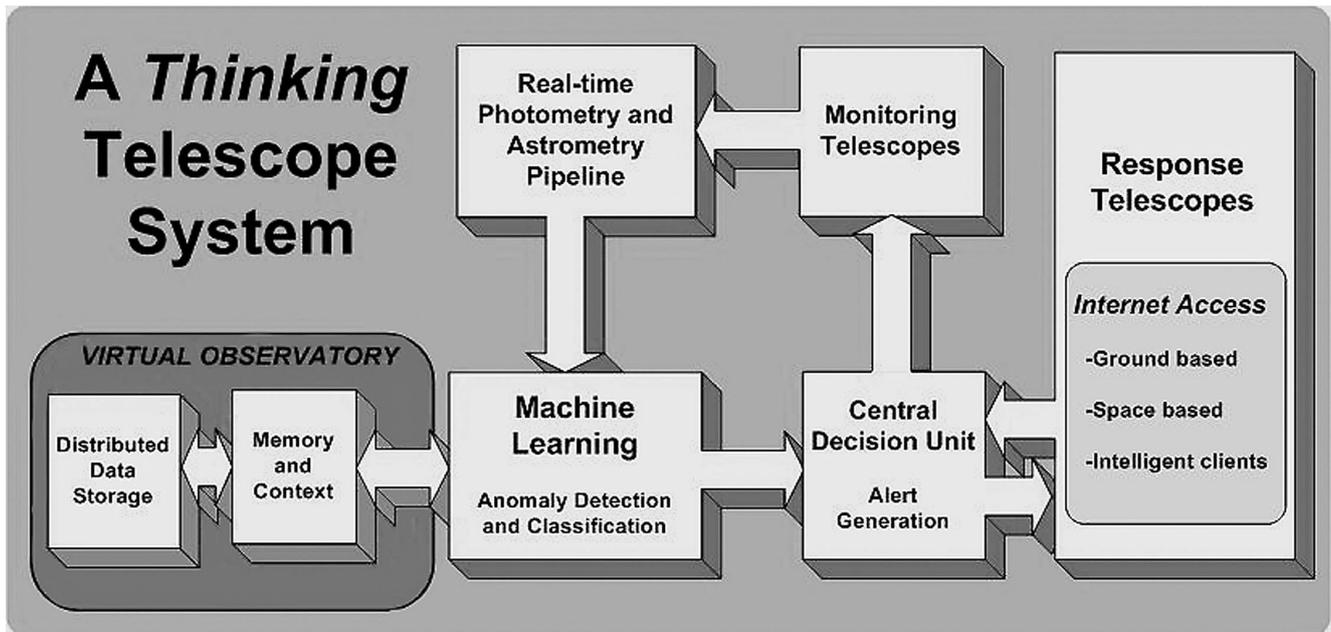


Fig. 4. A block diagram of the basic components in a thinking telescope system constructed for rapid response mining of the night sky.

- Brunner, R.J., Djorgovski, S.G. and Szalay, A.S.: 2001, Virtual Observatories of the Future, ASP Conference Proc., 225, San Francisco: Astron. Soc. Of the Pacific
- Fogel, A., Owens, A. and Walsh, M.: 1966, Artificial Intelligence through Simulated Evolution, Wiley, New York
- Foster, I., Kesselman, C., Tuecke, S.: 2001, Int. Journal of Super-computer Applications 15, 3
- Harvey, N. R., et al.: 2003, SPIE 5032, 557
- Plesko, C.S., Brumby, S., Leovy, C.: 2002, SPIE 4480, 139
- Perlmutter S., et al.: 1999, ApJ 517, 565
- Paczynski, B.,: 2000, PASP 112, 1281
- Paczynski, B., Chen, C.P., and Lemme, C.: 2001, Small Telescope Astronomy on Global Scales, ASP Conf. Proc., 246
- Pojmanski, G.: 2002, Acta Astron. 52, 397
- Riess A.G., et al.: 1998, AJ 116, 1009
- Theiler, J., et al.: 1999, SPIE 3753, 416
- Theiler, J., Cai, M.: 2003, SPIE 5093, 230
- Tyson, J.A.: 2002, SPIE 4836, 10
- Udalski, A., et al.: 1994, Acta Astron. 44, 227
- Udalski, A.: 2003, Acta Astron. 53, 291
- Vapnik, V.: 1995, The Nature of Statistical Learning Theory, New York: Springer-Verlag
- Vestrand, W.T., et al.: 2002, SPIE 4845, 126
- White, R.R., et al.: 2004, SPIE 5496, 302
- Wozniak, P.R., et al.: 2002, SPIE 4846, 147
- Wozniak, P.R., et al.: 2004, AJ 127, 2436