Table 1. Parallels between genome sequencing and genetic network discovery.

Genome sequencing	Genome semantics
Physical maps	Graphical model
Contigs	Low-level functional models
Contig reassembly	Module assembly
Finished genome sequence	Comprehensive model

DNA to be sequences into distinct pieces, parcel out the detailed work of sequencing, and then reassemble these independent efforts at the end. It is not quite so simple in the world of genome semantics.

Despite the differences between genome sequencing and genetic network discovery, there are clear parallels that are illustrated in Table 1. In genome sequencing, a physical map is useful to provide scaffolding for assembling the finished sequence. In the case of a genetic regulatory network, a graphical model can play the same role. A graphical model can represent a high-level view of interconnectivity and help isolate modules that can be studied independently. Like contigs in a genomic sequencing project, low-level functional models can explore the detailed behavior of a module of genes in a manner that is consistent with the higher level graphical model of the system. With standardized nomenclature and compatible modeling techniques, independent functional models can be assembled into a complete model of the cell under study.

To enable this process, there will need to be standardized forms for model representation. At present, there are many different modeling technologies in use, and although models can be easily placed into a database, they are not useful out of the context of their specific modeling package. The need for a standardized way of communicating computational descriptions of biological systems extends to the literature. Entire conferences have been established to explore ways of mining the biology literature to extract semantic information in computational form.

Going forward, as a community we need to come to consensus on how to represent what we know about biology in computational form as well as in words. The key to postgenomic biology will be the computational assembly of our collective knowledge into a cohesive picture of cellular and organism function. With such a comprehensive model, we will be able to explore new types of conservation between organisms and make great strides toward new therapeutics that function on well-characterized pathways.

References

1. S. K. Kim *et al.*, *Science* **293**, 2087 (2001).
2. A. Hartemink *et al.*, paper presented at the Pacific Symposium on Biocomputing 2000, Oahu, Hawaii, 4 to 9 January 2000.
3. D. Pe'er *et al.*, paper presented at the 9th Conference on Intelligent Systems in Molecular Biology (ISMB), Copenhagen, Denmark, 21 to 25 July 2001.
4. H. McAdams, A. Arkin, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 814 (1997).
5. A. J. Hartemink, thesis, Massachusetts Institute of Technology, Cambridge (2001).

VIEWPOINT

Machine Learning for Science: State of the Art and Future Prospects

Eric Mjolsness* and Dennis DeCoste

Recent advances in machine learning methods, along with successful applications across a wide variety of fields such as planetary science and bioinformatics, promise powerful new tools for practicing scientists. This viewpoint highlights some useful characteristics of modern machine learning methods and their relevance to scientific applications. We conclude with some speculations on near-term progress and promising directions.

Machine learning (ML) (*1*) is the study of computer algorithms capable of learning to improve their performance of a task on the basis of their own previous experience. The field is closely related to pattern recognition and statistical inference. As an engineering field, ML has become steadily more mathematical and more successful in applications over the past 20 years. Learning approaches such as data clustering, neural network classifiers, and nonlinear regression have found surprisingly wide application in the practice of engineering, business, and science. A generalized version of the standard Hidden Markov Models of ML practice have been used for ab initio prediction of gene structures in genomic DNA (*2*). The predictions correlate surprisingly well with subsequent gene expression analysis (*3*). Postgenomic biology prominently features large-scale gene expression data analyzed by clustering methods (*4*), a standard topic in unsupervised learning. Many other examples can be given of learning and pattern recognition applications in science. Where will this trend lead? We believe it will lead to appropriate, partial automation of every element of scientific method, from hypothesis generation to model construction to decisive experimentation. Thus, ML has the potential to amplify every aspect of a working scientist's progress to understanding. It will also, for better or worse, endow intelligent computer systems with some of the general analytic power of scientific thinking.

Machine Learning Systems Group, Jet Propulsion Laboratory/California Institute of Technology, Pasadena, CA, 91109, USA.

*To whom correspondence should be addressed. E-mail: mjolsness@jpl.nasa.gov

Machine Learning at Every Stage of the Scientific Process

Each scientific field has its own version of the scientific process. But the cycle of observing, creating hypotheses, testing by decisive experiment or observation, and iteratively building up comprehensive testable models or theories is shared across disciplines. For each stage of this abstracted scientific process, there are relevant developments in ML, statistical inference, and pattern recognition that will lead to semiautomatic support tools of unknown but potentially broad applicability.

Increasingly, the early elements of scientific method—observation and hypothesis generation—face high data volumes, high data acquisition rates, or requirements for objective analysis that cannot be handled by human perception alone. This has been the situation in experimental particle physics for decades. There automatic pattern recognition for significant events is well developed, including Hough transforms, which are foundational in pattern recognition. A recent example is event analysis for Cherenkov detectors (*8*) used in neutrino oscillation experiments. Microscope imagery in cell biology, pathology, petrology, and other fields has led to image-processing specialties. So has remote sensing from Earth-observing satellites, such as the newly operational Terra spacecraft with its ASTER (a multispectral thermal radiometer), MISR (multiangle imaging spectral radiometer), MODIS (imaging

spectrometer), and other high–data-rate instruments that will require pattern recognition for signal extraction and change detection to reach their full potential. The Mars Global Surveyor spacecraft and its MOC (2- to 10-m/pixel visible light camera), MOLA (laser altimeter), and TES (thermal emission spectrometer) instruments, along with three new spectrometers (THEMIS, GRS, and MARIE) to arrive shortly aboard the 2001 Mars Odyssey spacecraft, are now providing comprehensive views of another planet at high data volume, which will also stretch human analytic capabilities.

Step 1: Observe and explore interesting phenomena. Visualization and exploration of high-dimensional vector data are the focus of much current ML research. For example, one promising class of approaches involves dimensionality reduction: reducing data from many original dimensions (e.g., thousands of gene expression measurements) to just a few dimensions in some new space (e.g., two or three dimensions, for easy visualization on computer displays). This includes classic methods, such as principal component analysis (PCA) and multidimensional scaling, as well as many promising new ones, such as kernel principal component analysis (*5*), independent component analysis (ICA) (*6*), and locally linear embedding (*7*). The wide variety of new dimensionality reduction and visualization methods, along with general trends toward ever larger and higher dimensional data sets, suggests that this area of ML will continue to grow in relative importance. This wide variety, coupled with varying domain specificity, close collaboration between scientists and ML researchers.

In bioinformatics, sequence and microarray mRNA gene expression data are ramping up in volume as instrumentation improves, and the scientific understanding of many fundamental life processes is at stake. Data vector clustering methods have played a historic role in the early analyses of microarray data (*4*). They have also been used to develop new molecular classifiers of cancer types (*9*). Cutting-edge classification algorithms such as support vector machines have been used to predict gene function annotations from expression data (*10*). Frustratingly, we do not really understand the mathematical mapping of such cluster structures to causal models of underlying gene expression circuitry [but see (*11*)]. Another biological example is in neuroscience, where functional magnetic resonance imaging brain images have been decomposed into meaningful, statistically independent components of localized activity with the ICA method (*12*).

A good example of how ML classifiers can inform an overall scientific effort, even when those classifiers are not based on scientific theories per se, is the NASA work on learning detectors of planetary features such as volcanoes and impact craters. The scientific need for geological feature catalogs has led to multiyear human surveys of Mars orbital imagery yielding tens of thousands of cataloged, characterized features including impact craters, faults, and ridges. If the tedious aspects of this work could be vastly accelerated and made objective by automation, then feature relationships and impact crater counts could be used to order and date geological units with much finer spatial and temporal resolution than is now possible. In fact, the stratigraphic record could be objectively reanalyzed at high resolution. Recent steps toward this goal involve learning and pattern recognition. Trainable classifiers for geomorphological features were initially based on simple Gaussian models for orbital image data (*13*) and later improved with PCA (*14*) and support vector machines (*15*). So far the most accurate feature detector models bare little resemblance to the process models describing the formation of those geological features.

Simulation observations are also a fruitful but largely untapped source of data for ML techniques. For example, high-quality particle simulators of planetary and comet formations [e.g., (*16*)] generate vast amounts of data, for which careful manual examination is usually infeasible. Semiautomated exploration of this data, such as detecting outlier behaviors significantly and interestingly different from previous simulation runs, could help guide scientific investigation and drastically improve overall throughput for such increasingly important "science by simulation" work.

Step 2: Generate hypotheses. Many data-clustering algorithms may be treated as fitting vector data to a mixture of simpler probability distributions such as Gaussians, one per cluster. Figure 1 shows an example of a muscle cluster of 81 genes culled from hierarchical Expectation Maximization clustering analysis [as in (*11*)] of a C2C12 mouse muscle differentiation time course microarray, involving about 9000 genes (*13*). This cluster was hand-picked by browsing a cluster hierarchy at the top level to pick out genes up-regulated over the time course and at the second level to pick out genes whose pattern of induction was tightly correlated. Biological inspection of this gene list reveals known transcriptional regulators of muscle differentiation as well as their downstream target genes, which include classic markers of muscle differentiation. Clones for which no function could be attributed (of which there were 15) become candidates for a role in regulation or execution of muscle differentiation. On the basis of the assumption that genes that fall into the same expression cluster are particularly likely to share function, this provides a basis for guiding model formulation for genes with no known function. Thus, such clustering provides a basis for bootstrapping the process of understanding possible functions of poorly understood genes from better understood ones.

From unsupervised learning methods such as clustering or dimensionality reduction, there emerge patterns in data that demand explana-

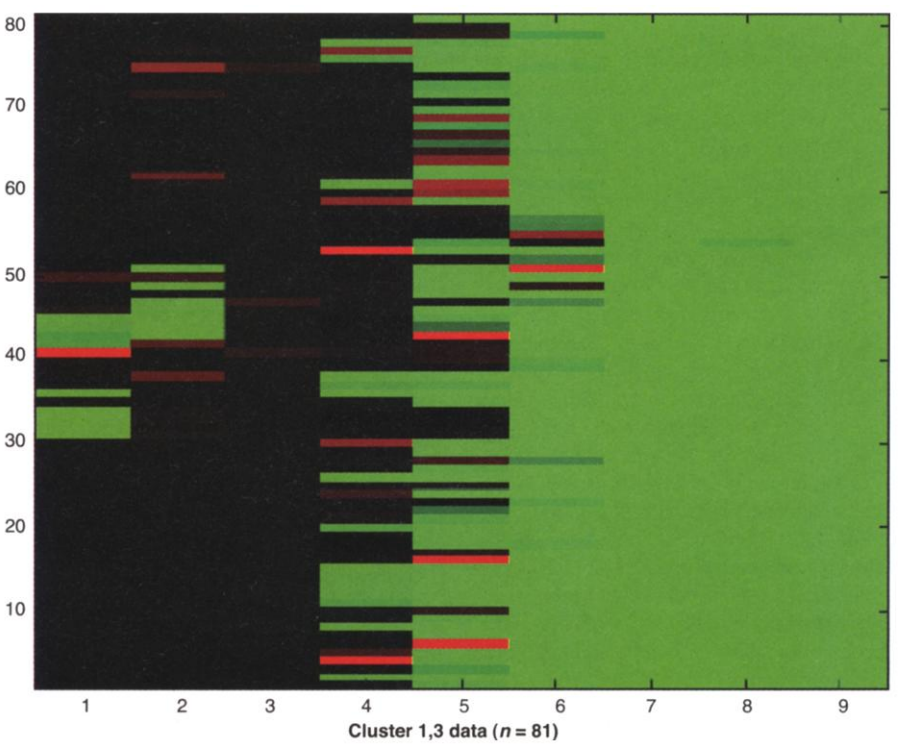

Fig. 1. Example of red-green activity display of one muscle gene cluster. The *x* axis indicates the stage of muscle development. The *y* axis indicates gene number. Colors represent discretized level of expression: Green is up-regulated, red is down-regulated, and black shows no change relative to the expression level of a reference sample.

tion. In expression bioinformatics, one observes clusters of coexpressed genes, which may suggest hypotheses of direct or indirect coregulation. Observed data may be also fit numerically to a relatively generic but predictive, causal model, such as a fully recurrent analog neural network model for gene expression data. This has been done for morphogenetic gene regulation networks in the early *Drosophila melanogaster* embryo, resulting in predictive models (*18*). From these learned models, specific hypotheses were derived about which gap gene controls each boundary of each modeled stripe of even-skipped gene expression. This kind of model inference falls into the fundamental ML category of nonlinear regression. It is still largely up to the imagination of the human scientist to transform observed patterns into testable hypotheses, but the model provides a mathematical language for doing so. Thus, model inversion methods, like unsupervised learning methods, have the potential to formalize or automate some aspects of hypothesis generation, particularly if coupled with Bayesian inference to ensure that the inverse problem is well posed.

There are extreme cases in which the automation of hypothesis generation is especially important. In robotic planetary exploration, speed-of-light communication delays and bandwidth limitations make robotic autonomy valuable or essential. Future reconnaissance geology robots such as Mars rovers would benefit from the use of supervised and unsupervised learning to classify visible rock surfaces. They would also benefit from having a comprehensive library of preprogrammed geological models and the ability to tune, instantiate, or recombine them to fit locally available evidence. These capabilities could be used to autonomously acquire and send back the most significant data available at each site. The in situ spacecraft, like a human field geologist, could maintain multiple working hypotheses and look for discriminating observations. Very early steps in this direction are taken in (*19*) and (*20*). Similar requirements for autonomy will apply to solar system orbital missions, in which onboard analysis of survey observations may suggest detailed follow-up observations. An early opportunity to test on-board science software in Earth orbit, with many extraterrestrial analogs, may arise with the Autonomous Sciencecraft Constellation experiment (*21*).

Step 3: Formulate model to explain phenomena. Learning good models from data is a central goal of ML, so the field offers a wide variety of powerful tools for this critical stage. Mixture models for clustering and recurrent analog neural net models for nonlinear regression have effective parameter-inference algorithms as described above (for step 2). Like unsupervised data-clustering algorithms, supervised learning algorithms have their own, equally generic statistical interpretations. At the other extreme in model specificity stand detailed simulatable mathematical models of particular systems, as practiced in computational physics, chemistry, and more recently computational biology. An important direction in ML research is to create automatically models of intermediate generality that can incorporate successively more domain expertise. One example is the method of trainable Markov Random Fields (MRFs), which have been applied to images of solar active regions with imagery from the Michelson Doppler Imager (MDI) instrument aboard the Solar and Helioseismic Observatory (SOHO) spacecraft (*22*, *23*) (see Fig. 2). Also of intermediate generality is the influential Bayes Net or "graphical model" formalism to describe interacting causal chains of hidden and observable stochastic variables (*24*). These models generalize Hidden Markov models. Frontier research in these areas addresses the inference of graphical model structure (connections between variables) and probability distribution parameters by optimization from data [e.g., (*25*, *26*)]. Future research will have to address the problem of variable, data-dependent graph structure such as arises in biological development, fluid physics, or the representation of abstract networks of interrelated hypotheses and concepts.

When the observed data can be labeled by a scientist as "positive" and "negative" examples of the phenomena of interest, supervised classifiers can be learned. Specific classifier methods tend to fall into one of two groups: (i) Generative models, which strive to capture the joint probability of the variables in the physical system. These approaches can include the learning of causal mechanisms (e.g., Bayesian networks or graphical models). (ii) Discriminative models, which strive only to capture the ability to distinguish pos-

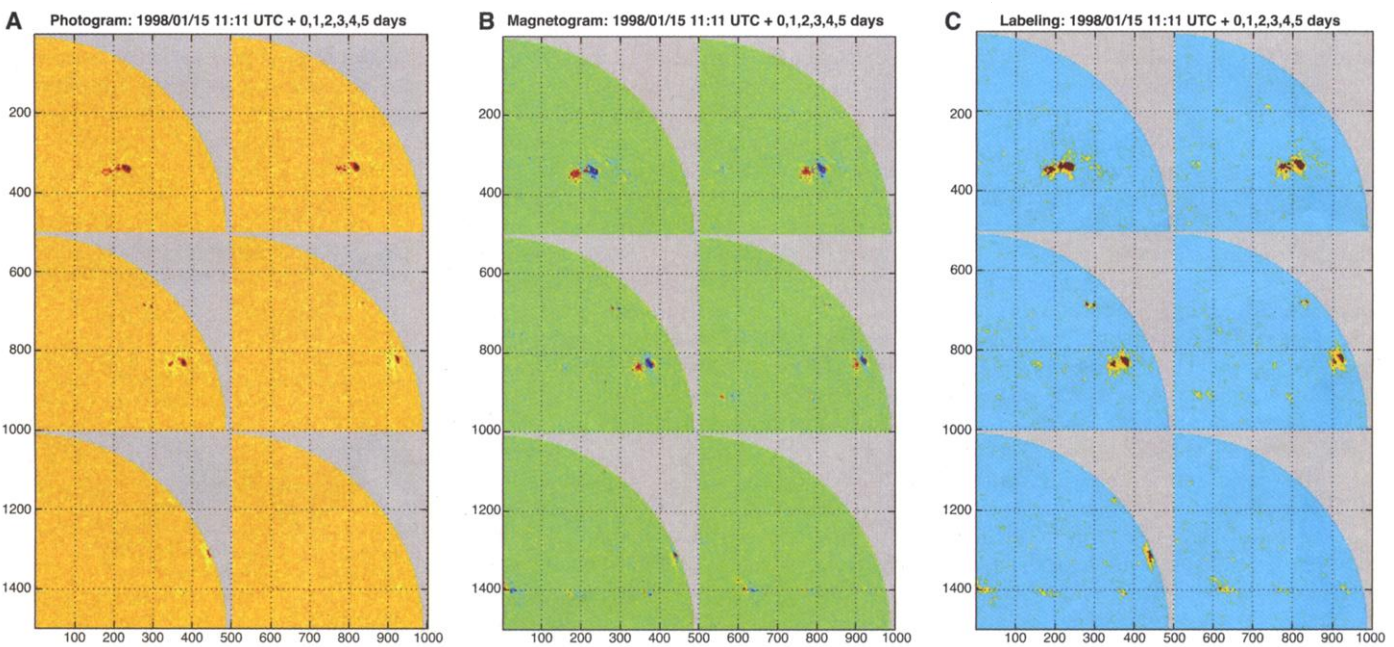

Fig. 2. Solar image analysis. The raw (SOHO/MDI) data consist of temporal sequences of two-dimensional images of intensity and magnetic flux. (**A**) Photograms (over 6 days). (**B**) Magnetograms. (**C**) Labelings given by the learned MRF model. The model assigns each pixel to one of three classes:sunspot (deep red), faculae (yellow), or quiet sun (cyan). Given the class labeling, a pixel's intensity and magnetic flux are assumed to be governed by mixture models appropriate to that class. The mixture model parameters for each class are learned. Images courtesy of M. Turmon, K. Shelton, and the MDI team.

itive from negative examples, while avoiding the expense of learning full joint probability densities. Support vector machines (SVMs) (*27*) are a very popular and successful example of this group.

As an example of discriminative models for robotic geology, Gilmore *et al.* (*19*) demonstrate the ability to discriminate carbonate from noncarbonate minerals using a feed-forward neural network taking inputs from a reflectance spectrometer suitable for use on a Mars rover.

Discriminative models make no attempt to explicitly capture the true underlying physics of the phenomena. Nevertheless, as many recent successful applications of methods such as SVMs have shown [e.g., (*10*)], such classifiers can provide strong insights into the nature of the phenomena, including such aspects as which input dimensions are most useful, which examples are most likely to be outliers, and what new observations might be most worthwhile to gather (e.g., "active learning" methods discussed in the next section).

It is important to realize that even the relatively more complex generative/probabilistic models need not be realistic to be useful. This is a critical point for potential scientific users of ML technologies to realize and try to exploit in practice. For example, one could train ML classifiers to classify atmospheric image data pixels (e.g., "clouds" versus "nonclouds"), using as inputs not only the raw pixel values but also the prediction of complex realistic physics-based models for each pixel. In this way, modern ML methodologies can systematically determine on what sort of images the complex models actually work or fail. Furthermore, active learning methods can suggest new data that would allow the ML classifier to best correct the predictions coming from the realistic model, toward much more accurate overall classifications.

Step 4: Test predictions made by the theory. One of the most expensive and error-prone aspects of ML of classifier models is the need for relatively large volumes of (manually) labeled data. An emerging topic in current ML research is active learning, which provides automated means of determining which potential new data points would be most useful to label. Thus, active learning methods directly address the issue of automating the process of determining what predictions to make and what data to gather to test them. Active learning promises great savings over the more standard [e.g., (*13*)] use of automated classifiers in the scientific process, by radically reducing the amount of manual labeling required from scientists and lab technicians.

A good example of this emerging area of ML, which also illustrates some of the unique advantages of the discriminative SVM approach, is given by Tong and Koeller (*28*). A common active learning technique is to select as the next data point to label (e.g., by performing an experiment) a point for which the current ML model is least certain of its classification. However, Tong and Koeller (*28*) show that selecting the new data point that would most evenly halve the number of models consistent with the augmented data set, regardless of whether the label turns out to be positive or negative, can be much more effective. They show that the special geometrical nature of the "version space" of SVM models consistent with the data is ideally suited to the active learning task.

Step 5: Modify theory and repeat (at step 2 or 3). ML work on "theory refinement" addresses the issue of how best to update models on the basis of new data. Much early work on neural networks focused on such incremental, online tasks. Both ML and the traditional scientific process approaches face the same fundamental issues, such as when to refine the theory versus doubt the data and how to best seek parsimony in explanation. For parsimony, Occam's razor has a technical counterpart in ML theory. Following Einstein's "everything should be made as simple as possible, but not simpler," ML theory provides formal mathematic bases for encouraging low model complexity. In model selection, one may minimize complexity metrics such as Vapnik-Chervonenkis dimension or minimum description length (*1*), while maintaining good performance on a training data set, in order to obtain good generalization to test data sets drawn from the same or related distributions.

Obstacles to Automation

To achieve its promise to improve the science process across all stages, ML methods face a variety of outstanding obstacles. One is that most ML work to date focuses on vector data, limiting its value for richer, nonvector relations such as graph clusters (*29*) and text data. ML work on bioinformatics and Internet data that addresses these issues is relatively new and immature. Similarly, most ML work assumes relatively fixed model structures, whereas variable structures (such as data-dependent graphical models) would often seem necessary—especially during early stages of investigation, when nothing even close to a unified theory is available to guide the structure.

From a systems perspective, much work is still needed. Standards and methods for model sharing and formal specification, enabling ML methods to communicate with both scientists and other ML methods, are still relatively primitive and incomplete. The Holy Grail of integrating automated reasoning across all relevant representations and processes seems far from current reality. This is in no small part due to our continuing ignorance of the creative human thought processes guiding the art of doing science.

Conclusions

Despite the obstacles, current machine learning (ML) research is producing impressive advances on some fundamental issues. These include scaling up ML methods to large sample and dimension sizes, finding the productive balance between generative and discriminative approaches, and focusing attention on the most useful data (e.g., active learning). We expect steady progress on these core areas of ML research in coming years.

Some core ML research directions are especially likely to further the partial automation of scientific processes. These include learning from nonvector data such as labeled graphs, text, and images; multiscale methods for large-scale optimization in inference; and feature selection to find the most relevant aspects of large data sets.

A particularly exciting recent trend in ML research has been the development of nonlinear Mercer kernel versions of many classic linear methods [e.g., kernel PCA, kernel nearest-neighbor, and kernel Fischer discriminates (*30*)] and domain-specific kernels [e.g., (*31*)]. The promise of such kernel methods is the ability to learn highly accurate models on large feature spaces, while overcoming the traditional "curse of dimensionality." We expect breakthroughs in this area, both for new algorithms and for new powerful domain-specific kernels (*32*).

Although we have focused on semiautomated, human-interactive scientific inference, there is already demand for more fully automatic inference in special situations. Such situations include very short or long time scales or locations inhospitable to human intervention, such as robotic planetary exploration missions. It will be a fascinating and instructive endeavor—requiring contributions across technology, science, and even philosophy—to develop and understand the full spectrum of such scientific inference systems.

References and Notes

1. T. Mitchell, *Machine Learning* (McGraw-Hill, New York, 1997).
2. C. Burge, S. Karlin, *J. Mol. Biol.* **268**, 78 (1997).
3. D. D. Shoemaker *et al.*, *Nature* **409**, 922 (2001).
4. P. T. Spellman *et al.*, *Mol. Biol. Cell* **9**, 3273 (1998). Data analysis discussed by M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863 (1998).
5. B. Schölkopf, A. Smola, K.-R. Müller, in *Advances in Kernel Methods–SV Learning*, B. Schölkopf, C. J. C. Burges, A. J. Smola, Eds. (MIT Press, Cambridge, MA, 1999), pp. 327–352.
6. A. J. Bell, T. J. Sejnowski, *Neural Comput.* **7**, 6 (1995).
7. S. Roweis, L. Saul, *Science* **290**, 2323 (2000).
8. M. Shiozawa and the Super-Kamiokande collaboration, *Nucl. Instrum. Methods Phys. Res. Sect. A* **433**, 240 (1999).
9. T. R. Golub *et al.*, *Science* **286**, 531 (1999).
10. M. Brown *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 1 (2000).
11. E. Mjolsness, T. Mann, R. Castaño, B. Wold, *Adv. Neural Inform. Processing Syst.* **12**, 928 (2000).
12. T.-P. Jung *et al.*, *Proc. IEEE* **89**, 7 (2001).
13. M. C. Burl *et al.*, *Machine Learning* **30**, 165 (1998).
14. M. C. Burl *et al.*, paper presented at 5th International

Symposium on Artificial Intelligence, Robotics, and Automation in Space (i-SAIRAS), Montreal, Canada, June 2001.
15. D. DeCoste, B. Schölkopf, *Machine Learning*, in press.
16. H. F. Levison, L. Dones, M. J. Duncan, *Astron. J.* **121**, 4 (2001).
17. B. Williams, S. Damle, B. Wold, unpublished data.
18. J. Reinitz, D. H. Sharp, *Mech. Dev.* **49**, 133 (1995). Mathematical foundation introduced by E. D. Mjolsness, H. Sharp, J. Reinitz, *J. Theor. Biol.* **152**, 429 (1991).
19. M. S. Gilmore *et al.*, *J. Geophys. Res.* **105**, 29233 (2000).
20. T. Estlin *et al.*, in *Proceedings of the American Association for Artificial Intelligence Conference* (AAAI Press/MIT Press, Orlando, FL, 1999), pp. 613–620.
21. A. Davies *et al.*, Autonomous Sciencecraft Constellation Science Study Report (Jet Propulsion Laboratory, Pasadena, CA, August 2001). Available at http://asc.jpl.nasa.gov, which also describes the technology of the ASC experiment.
22. M. Turmon, S, Mukhtar, J. Pap, in *Proceedings of the Third Conference on Knowledge Discovery and Data Mining* (AAAI Press, Newport Beach, CA, 1997), pp. 267–270.
23. M. Turmon, J. M. Pap, S. Mukhtar, in *Structure and Dynamics of the Interior of the Sun and Sun-like Stars* (ESA SP-418, European Space Agency Publications Division, Noordwijk, Netherlands, 1998), pp. 979–984.
24. M. Jordan, Ed., *Learning in Graphical Models* (Kluwer, Dordrecht, Netherlands, 1998).
25. K. Murphy, Y. Weiss, M. Jordan, in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, K. B. Laskey, H. Prade, Eds. (Kaufmann, San Francisco, CA, 1999), pp. 467–475.
26. Y. Weiss, *Neural Comput.* **12**, 1 (2000).
27. C. J. C. Burges, *Data Mining Knowledge Discov.* **2**, 2 (1998).
28. S. Tong, D. Koller, in *Proceedings of the Seventeenth International Conference on Machine Learning* (Kaufman, San Francisco, CA, 2000), pp. 999–1006.
29. S. Gold, A. Rangarajan, E. Mjolsness, *Neural Comput.* **8**, 4 (1996).
30. S. Mika, G. Rätsch, K.-R. Müller, *Adv. Neural Inform. Processing Syst.* **13**, 591 (2001).
31. A. Zien *et al.*, *Bioinformatics* **16**, 799 (2000).
32. Good Web sites for kernel methods are www.kernel-machines.org/ and www.support-vector.net/. A good site about graphical models is www.auai.org/. The annual Neural Information Processing Systems (NIPS) conference is a good source of the latest results in both areas and many others (with complete online archives at http://nips.djvuzone.org/). The Web site for the JPL Machine Learning Systems Group is www-aig.jpl.nasa.gov/mls.
33. Work supported in part by the Intelligent Data Understanding and Automated Reasoning elements of the NASA Intelligent Systems program, by the Whittier Foundation, by the NASA Applied Information Systems Research Program, by a National Research Service Award from the NIH, and by the NASA Autonomy and Cross Enterprise Technology Development Programs.