Machine Learning for Science: State of the Art and Future Prospects
Eric Mjolsness, et al.
Science 293, 2051 (2001);
DOI: 10.1126/science.293.5537.2051

The following resources related to this article are available online at www.sciencemag.org (this information is current as of April 15, 2009):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
http://www.sciencemag.org/cgi/content/full/293/5537/2051

This article cites 16 articles, 6 of which can be accessed for free:
http://www.sciencemag.org/cgi/content/full/293/5537/2051#otherarticles

This article has been cited by 27 article(s) on the ISI Web of Science.

This article has been cited by 5 articles hosted by HighWire Press; see:
http://www.sciencemag.org/cgi/content/full/293/5537/2051#otherarticles

This article appears in the following subject collections:
Computers, Mathematics
http://www.sciencemag.org/cgi/collection/comp_math

Information about obtaining reprints of this article or about obtaining permission to reproduce this article in whole or in part can be found at:
http://www.sciencemag.org/about/permissions.dtl
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**


---

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

<table>
<thead>
<tr>
<th>Genome sequencing</th>
<th>Genome semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical maps</td>
<td>Graphical model</td>
</tr>
<tr>
<td>Contigs</td>
<td>Low-level functional models</td>
</tr>
<tr>
<td>Contig reassembly</td>
<td>Module assembly</td>
</tr>
<tr>
<td>Finished genome</td>
<td>Comprehensive model</td>
</tr>
</tbody>
</table>

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 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 (multispectral imaging spectroradiometer), MODIS (imaging
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. Semi-automated 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.

**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.
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 Heliospheric 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.
C O M P U T E R S A N D S C I E N C E

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.

References and Notes

15. M. C. Burl et al., paper presented at 5th International
15. D. DeCoste, B. Schölkopf, Machine Learning, in press.
17. B. Williams, S. Damle, B. Wold, unpublished data.