Despite practical challenges, we are hopeful that informed discussions among policy-makers and the public about data and the capabilities of machine learning, will lead to insightful designs of programs and policies that can balance the goals of protecting privacy and ensuring fairness with those of reaping the benefits to scientific research and to individual and public health. Our commitments to privacy and fairness are evergreen, but our policy choices must adapt to advance them, and support new techniques for deepening our knowledge.

## REFERENCES AND NOTES


1. M. De Choudhury, S. Counts, E. Horvitz, A. Hoff, in *Proceedings of International Conference on Weblogs and Social Media* [Association for the Advancement of Artificial Intelligence (AAAI), Palo Alto, CA, 2014].
2. J. S. Brownstein, C. C. Freifeld, L. C. Madoff, *N. Engl. J. Med.* **360**, 2153–2155 (2009).
3. G. Eysenbach, *J. Med. Internet Res.* **11**, e11 (2009).
4. D. A. Broniatowski, M. J. Paul, M. Dredze, *PLOS ONE* **8**, e83672 (2013).
5. A. Sadilek, H. Kautz, V. Silenzio, in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (AAAI, Palo Alto, CA, 2012).
6. M. De Choudhury, S. Counts, E. Horvitz, in *Proceedings of the* SIGCHI *Conference on Human Factors in Computing Systems* (Association for Computing Machinery, New York, 2013), pp. 3267–3276.
7. R. W. White, R. Harpaz, N. H. Shah, W. DuMouchel, E. Horvitz, *Clin. Pharmacol. Ther.* **96**, 239–246 (2014).
8. Samaritans Radar; www.samaritans.org/how-we-can-help-you/supporting-someone-online/samaritans-radar.
9. Shut down Samaritans Radar; http://bit.ly/Samaritans-after.
10. U.S. Equal Employment Opportunity Commission (EEOC), 29 Code of Federal Regulations (C.F.R.), 1630.2 (g) (2013).
11. EEOC, 29 CFR 1635.3 (c) (2013).
12. M. A. Rothstein, *J. Law Med. Ethics* **36**, 837–840 (2008).
13. Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values* (White House, Washington, DC, 2014); http://1.usa.gov/1TSOhiG.
14. Letter from Maneesha Mithal, FTC, to Reed Freeman, Morrison, & Foerster LLP, Counsel for Netflix, 2 [closing letter] (2010); http://1.usa.gov/1GCFyXR.
15. In re Facebook, Complaint, FTC File No. 092 3184 (2012).
16. FTC Staff Report, *Mobile Privacy Disclosures: Building Trust Through Transparency* (FTC, Washington, DC, 2013); http://1.usa.gov/1eNz8zr.
17. FTC, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers* (FTC, Washington, DC, 2012).
18. Directive 95/46/ec of the European Parliament and of The Council of Europe, 24 October 1995.
19. L. Sweeney, Online ads roll the dice [blog]; http://1.usa.gov/1KgEcYg.
20. FTC, "Big data: A tool for inclusion or exclusion?" (workshop, FTC, Washington, DC, 2014); http://1.usa.gov/1SR65cv
21. FTC, *Data Brokers: A Call for Transparency and Accountability* (FTC, Washington, DC, 2014); http://1.usa.gov/1GCFoj5.
22. J. Podesta, "Big data and privacy: 1 year out" [blog]; http://bit.ly/WHsePrivacy.
23. White House Council of Economic Advisers, *Big Data and Differential Pricing* (White House, Washington, DC, 2015).
24. Executive Office of the President, *Big Data and Differential Processing* (White House, Washington, DC, 2015); http://1.usa.gov/1eNy7qR.
25. Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values* (White House, Washington, DC, 2014); http://1.usa.gov/1TSOhiG.
26. President's Council of Advisors on Science and Technology (PCAST), *Big Data and Privacy: A Technological Perspective* (White House, Washington, DC, 2014); http://1.usa.gov/1C5ewNv.
27. European Commission, Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), COM(2012) 11 final (2012); http://bit.ly/1Lu5POv.
28. *M. Schrems* v. *Facebook Ireland Limited*, §J. Unlawful data transmission to the U.S.A. ("PRISM"), ¶166 and 167 (2013); www.europe-v-facebook.org/sk/sk_en.pdf.


10.1126/science.aac4520

REVIEW

# Machine learning: Trends, perspectives, and prospects

**M. I. Jordan[1]* and T. M. Mitchell[2]***

**Machine learning addresses the question of how to build computers that improve automatically through experience. It is one of today's most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science. Recent progress in machine learning has been driven both by the development of new learning algorithms and theory and by the ongoing explosion in the availability of online data and low-cost computation. The adoption of data-intensive machine-learning methods can be found throughout science, technology and commerce, leading to more evidence-based decision-making across many walks of life, including health care, manufacturing, education, financial modeling, policing, and marketing.**

Machine learning is a discipline focused on two interrelated questions: How can one construct computer systems that automatically improve through experience? and What are the fundamental statistical-computational-information-theoretic laws that govern all learning systems, including computers, humans, and organizations? The study of machine learning is important both for addressing these fundamental scientific and engineering questions and for the highly practical computer software it has produced and fielded across many applications.

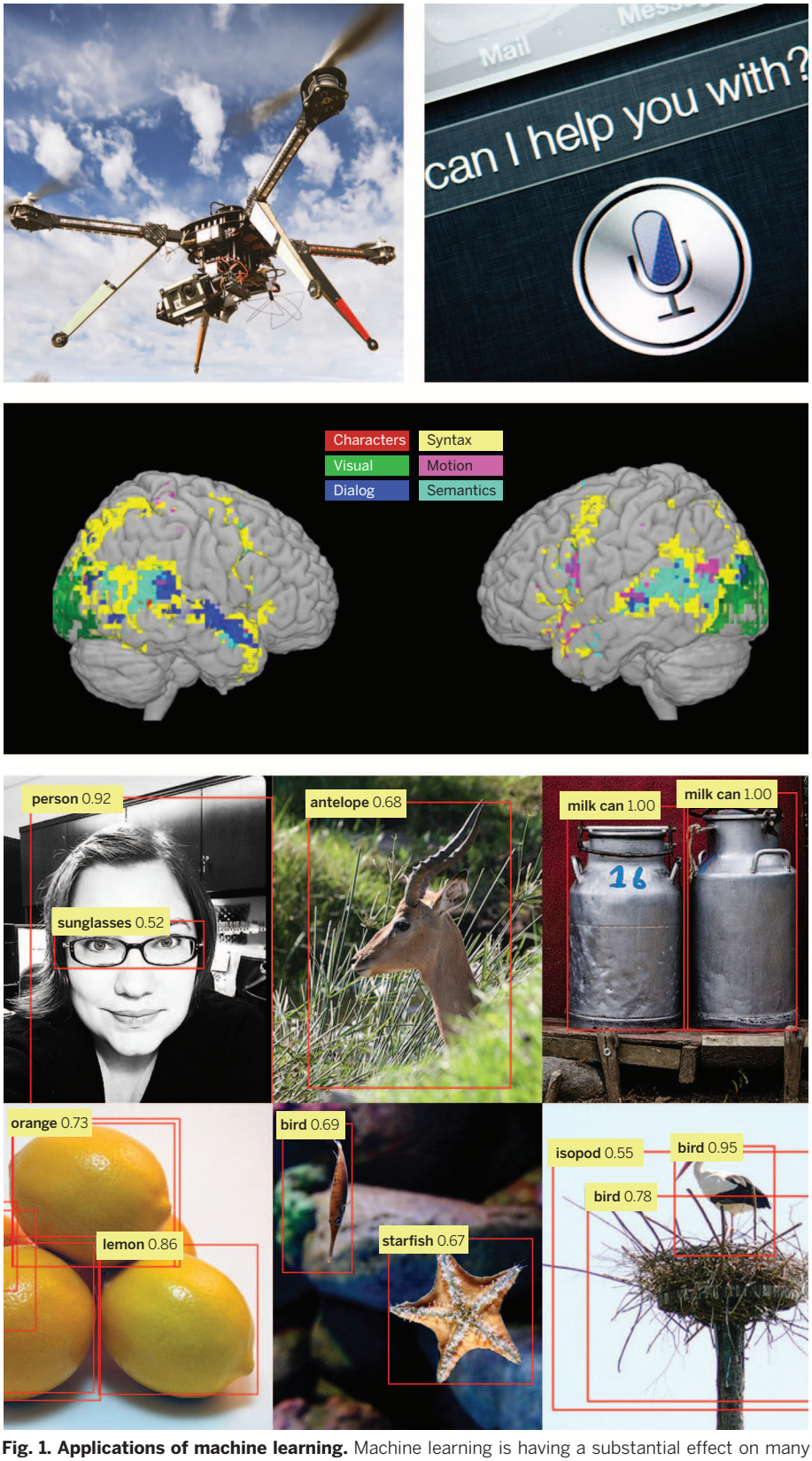
Machine learning has progressed dramatically over the past two decades, from laboratory curiosity to a practical technology in widespread commercial use. Within artificial intelligence (AI), machine learning has emerged as the method of choice for developing practical software for computer vision, speech recognition, natural language processing, robot control, and other applications. Many developers of AI systems now recognize that, for many applications, it can be far easier to train a system by showing it examples of desired input-output behavior than to program it manually by anticipating the desired response for all possible inputs. The effect of machine learning has also been felt broadly across computer science and across a range of industries concerned with data-intensive issues, such as consumer services, the diagnosis of faults in complex systems, and the control of logistics chains. There has been a similarly broad range of effects across empirical sciences, from biology to cosmology to social science, as machine-learning methods have been developed to analyze high-throughput experimental data in novel ways. See Fig. 1 for a depiction of some recent areas of application of machine learning.

A learning problem can be defined as the problem of improving some measure of performance when executing some task, through some type of training experience. For example, in learning to detect credit-card fraud, the task is to assign a label of "fraud" or "not fraud" to any given credit-card transaction. The performance metric to be improved might be the accuracy of this fraud classifier, and the training experience might consist of a collection of historical credit-card transactions, each labeled in retrospect as fraudulent or not. Alternatively, one might define a different performance metric that assigns a higher penalty when "fraud" is labeled "not fraud" than when "not fraud" is incorrectly labeled "fraud." One might also define a different type of training experience—for example, by including unlabeled credit-card transactions along with labeled examples.

A diverse array of machine-learning algorithms has been developed to cover the wide variety of data and problem types exhibited across different machine-learning problems (*1*, *2*). Conceptually, machine-learning algorithms can be viewed as searching through a large space of candidate programs, guided by training experience, to find a program that optimizes the performance metric. Machine-learning algorithms vary greatly, in part by the way in which they represent candidate programs (e.g., decision trees, mathematical functions, and general programming languages) and in part by the way in which they search through this space of programs (e.g., optimization algorithms with well-understood convergence guarantees and evolutionary search methods that evaluate successive generations of randomly mutated programs). Here, we focus on approaches that have been particularly successful to date.

Many algorithms focus on function approximation problems, where the task is embodied in a function (e.g., given an input transaction, output a "fraud" or "not fraud" label), and the learning problem is to improve the accuracy of that function, with experience consisting of a sample of known input-output pairs of the function. In some cases, the function is represented explicitly as a parameterized functional form; in other cases, the function is implicit and obtained via a search process, a factorization, an optimization

[1]Department of Electrical Engineering and Computer Sciences, Department of Statistics, University of California, Berkeley, CA, USA. [2]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA.
***Corresponding author. E-mail: jordan@cs.berkeley.edu (M.I.J.); tom.mitchell@cs.cmu.edu (T.M.M.)**




CREDIT: ISTOCK/AKINBOSTANCI

**Fig. 1. Applications of machine learning.** Machine learning is having a substantial effect on many areas of technology and science; examples of recent applied success stories include robotics and autonomous vehicle control (top left), speech processing and natural language processing (top right), neuroscience research (middle), and applications in computer vision (bottom). [The middle panel is adapted from (*29*). The images in the bottom panel are from the ImageNet database; object recognition annotation is by R. Girshick.]

procedure, or a simulation-based procedure. Even when implicit, the function generally depends on parameters or other tunable degrees of freedom, and training corresponds to finding values for these parameters that optimize the performance metric.

Whatever the learning algorithm, a key scientific and practical goal is to theoretically characterize the capabilities of specific learning algorithms and the inherent difficulty of any given learning problem: How accurately can the algorithm learn from a particular type and volume of training data? How robust is the algorithm to errors in its modeling assumptions or to errors in the training data? Given a learning problem with a given volume of training data, is it possible to design a successful algorithm or is this learning problem fundamentally intractable? Such theoretical characterizations of machine-learning algorithms and problems typically make use of the familiar frameworks of statistical decision theory and computational complexity theory. In fact, attempts to characterize machine-learning algorithms theoretically have led to blends of statistical and computational theory in which the goal is to simultaneously characterize the sample complexity (how much data are required to learn accurately) and the computational complexity (how much computation is required) and to specify how these depend on features of the learning algorithm such as the representation it uses for what it learns (*3*–*6*). A specific form of computational analysis that has proved particularly useful in recent years has been that of optimization theory, with upper and lower bounds on rates of convergence of optimization procedures merging well with the formulation of machine-learning problems as the optimization of a performance metric (*7*, *8*).

As a field of study, machine learning sits at the crossroads of computer science, statistics and a variety of other disciplines concerned with automatic improvement over time, and inference and decision-making under uncertainty. Related disciplines include the psychological study of human learning, the study of evolution, adaptive control theory, the study of educational practices, neuroscience, organizational behavior, and economics. Although the past decade has seen increased crosstalk with these other fields, we are just beginning to tap the potential synergies and the diversity of formalisms and experimental methods used across these multiple fields for studying systems that improve with experience.

## Drivers of machine-learning progress

The past decade has seen rapid growth in the ability of networked and mobile computing systems to gather and transport vast amounts of data, a phenomenon often referred to as "Big Data." The scientists and engineers who collect such data have often turned to machine learning for solutions to the problem of obtaining useful insights, predictions, and decisions from such data sets. Indeed, the sheer size of the data makes it essential to develop scalable procedures that blend computational and statistical

considerations, but the issue is more than the mere size of modern data sets; it is the granular, personalized nature of much of these data. Mobile devices and embedded computing permit large amounts of data to be gathered about individual humans, and machine-learning algorithms can learn from these data to customize their services to the needs and circumstances of each individual. Moreover, these personalized services can be connected, so that an overall service emerges that takes advantage of the wealth and diversity of data from many individuals while still customizing to the needs and circumstances of each. Instances of this trend toward capturing and mining large quantities of data to improve services and productivity can be found across many fields of commerce, science, and government. Historical medical records are used to discover which patients will respond best to which treatments; historical traffic data are used to improve traffic control and reduce congestion; historical crime data are used to help allocate local police to specific locations at specific times; and large experimental data sets are captured and curated to accelerate progress in biology, astronomy, neuroscience, and other data-intensive empirical sciences. We appear to be at the beginning of a decades-long trend toward increasingly data-intensive, evidence-based decision-making across many aspects of science, commerce, and government.

With the increasing prominence of large-scale data in all areas of human endeavor has come a wave of new demands on the underlying machine-learning algorithms. For example, huge data sets require computationally tractable algorithms, highly personal data raise the need for algorithms that minimize privacy effects, and the availability of huge quantities of unlabeled data raises the challenge of designing learning algorithms to take advantage of it. The next sections survey some of the effects of these demands on recent work in machine-learning algorithms, theory, and practice.

## Core methods and recent progress

The most widely used machine-learning methods are supervised learning methods (*1*). Supervised learning systems, including spam classifiers of e-mail, face recognizers over images, and medical diagnosis systems for patients, all exemplify the function approximation problem discussed earlier, where the training data take the form of a collection of $(x, y)$ pairs and the goal is to produce a prediction $y^*$ in response to a query $x^*$. The inputs $x$ may be classical vectors or they may be more complex objects such as documents, images, DNA sequences, or graphs. Similarly, many different kinds of output $y$ have been studied. Much progress has been made by focusing on the simple binary classification problem in which $y$ takes on one of two values (for example, "spam" or "not spam"), but there has also been abundant research on problems such as multiclass classification (where $y$ takes on one of $K$ labels), multilabel classification (where $y$ is labeled simultaneously by several of the $K$ labels), ranking problems (where $y$ provides a partial order on some set), and general structured prediction problems (where $y$ is a combinatorial object such as a graph, whose components may be required to satisfy some set of constraints). An example of the latter problem is part-of-speech tagging, where the goal is to simultaneously label every word in an input sentence $x$ as being a noun, verb, or some other part of speech. Supervised learning also includes cases in which $y$ has real-valued components or a mixture of discrete and real-valued components.

Supervised learning systems generally form their predictions via a learned mapping $f(x)$, which produces an output $y$ for each input $x$ (or a probability distribution over $y$ given $x$). Many different forms of mapping $f$ exist, including decision trees, decision forests, logistic regression, support vector machines, neural networks, kernel machines, and Bayesian classifiers (*1*). A variety of learning algorithms has been proposed to estimate these different types of mappings, and there are also generic procedures such as boosting and multiple kernel learning that combine the outputs of multiple learning algorithms. Procedures for learning $f$ from data often make use of ideas from optimization theory or numerical analysis, with the specific form of machine-learning problems (e.g., that the objective function or function to be integrated is often the sum over a large number of terms) driving innovations. This diversity of learning architectures and algorithms reflects the diverse needs of applications, with different architectures capturing different kinds of mathematical structures, offering different levels of amenability to post-hoc visualization and explanation, and providing varying trade-offs between computational complexity, the amount of data, and performance.

One high-impact area of progress in supervised learning in recent years involves deep networks, which are multilayer networks of threshold units, each of which computes some simple parameterized function of its inputs (*9*, *10*). Deep learning systems make use of gradient-based optimization algorithms to adjust parameters throughout such a multilayered network based on errors at its output. Exploiting modern parallel computing architectures, such as graphics processing units originally developed for video gaming, it has been possible to build deep learning systems that contain billions of parameters and that can be trained on the very large collections of images, videos, and speech samples available on the Internet. Such large-scale deep learning systems have had a major effect in recent years in computer vision (*11*) and speech recognition (*12*), where they have yielded major improvements in performance over previous approaches

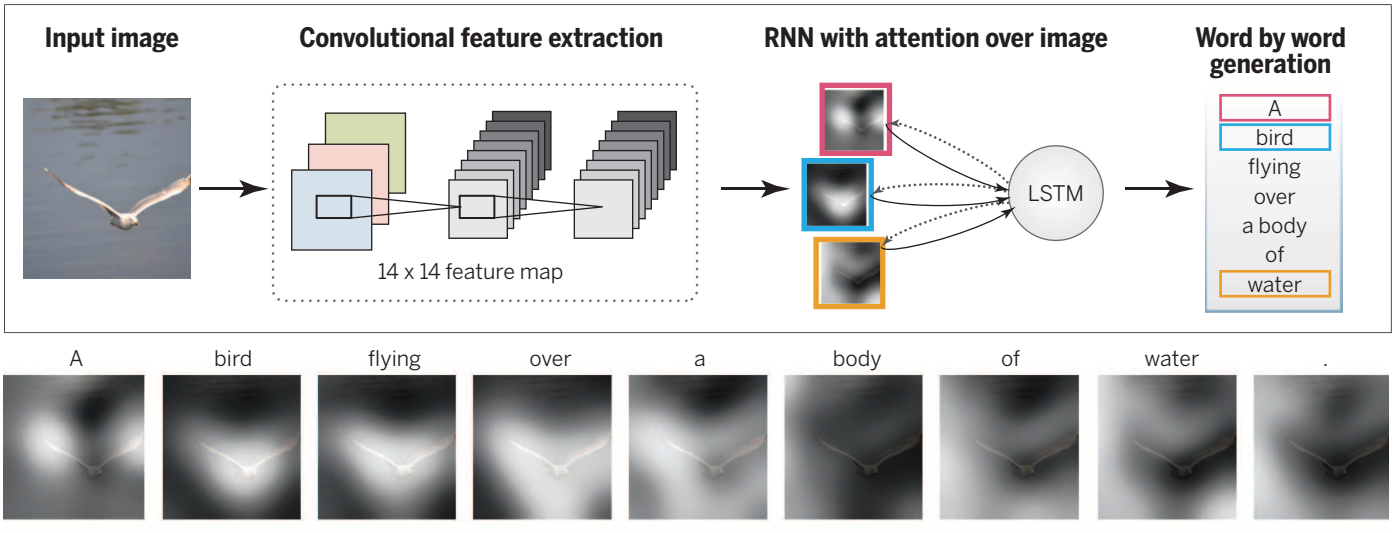



**Fig. 2. Automatic generation of text captions for images with deep networks.** A convolutional neural network is trained to interpret images, and its output is then used by a recurrent neural network trained to generate a text caption (top). The sequence at the bottom shows the word-by-word focus of the network on different parts of input image while it generates the caption word-by-word. [Adapted with permission from (*30*)]

CREDIT: ISTOCK/CORR

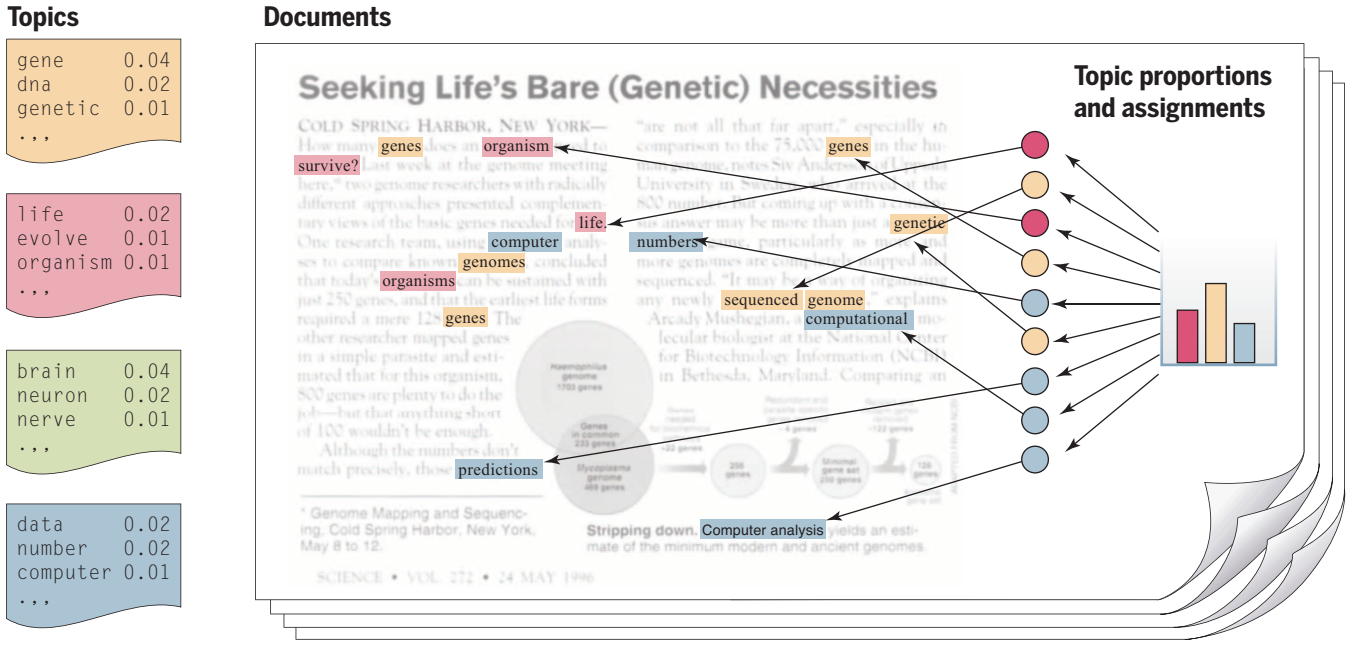



**Fig. 3. Topic models.** Topic modeling is a methodology for analyzing documents, where a document is viewed as a collection of words, and the words in the document are viewed as being generated by an underlying set of topics (denoted by the colors in the figure). Topics are probability distributions across words (leftmost column), and each document is characterized by a probability distribution across topics (histogram). These distributions are inferred based on the analysis of a collection of documents and can be viewed to classify, index, and summarize the content of documents. [From (*31*). Copyright 2012, Association for Computing Machinery, Inc. Reprinted with permission]

(see Fig. 2). Deep network methods are being actively pursued in a variety of additional applications from natural language translation to collaborative filtering.

The internal layers of deep networks can be viewed as providing learned representations of the input data. While much of the practical success in deep learning has come from supervised learning methods for discovering such representations, efforts have also been made to develop deep learning algorithms that discover useful representations of the input without the need for labeled training data (*13*). The general problem is referred to as unsupervised learning, a second paradigm in machine-learning research (*2*).

Broadly, unsupervised learning generally involves the analysis of unlabeled data under assumptions about structural properties of the data (e.g., algebraic, combinatorial, or probabilistic). For example, one can assume that data lie on a low-dimensional manifold and aim to identify that manifold explicitly from data. Dimension reduction methods—including principal components analysis, manifold learning, factor analysis, random projections, and autoencoders (*1*, *2*)—make different specific assumptions regarding the underlying manifold (e.g., that it is a linear subspace, a smooth nonlinear manifold, or a collection of submanifolds). Another example of dimension reduction is the topic modeling framework depicted in Fig. 3. A criterion function is defined that embodies these assumptions—often making use of general statistical principles such as maximum likelihood, the method of moments, or Bayesian integration—and optimization or sampling algorithms are developed to optimize the criterion. As another example, clustering is the problem of finding a partition of the observed data (and a rule for predicting future data) in the absence of explicit labels indicating a desired partition. A wide range of clustering procedures has been developed, all based on specific assumptions regarding the nature of a "cluster." In both clustering and dimension reduction, the concern with computational complexity is paramount, given that the goal is to exploit the particularly large data sets that are available if one dispenses with supervised labels.

A third major machine-learning paradigm is reinforcement learning (*14*, *15*). Here, the information available in the training data is intermediate between supervised and unsupervised learning. Instead of training examples that indicate the correct output for a given input, the training data in reinforcement learning are assumed to provide only an indication as to whether an action is correct or not; if an action is incorrect, there remains the problem of finding the correct action. More generally, in the setting of sequences of inputs, it is assumed that reward signals refer to the entire sequence; the assignment of credit or blame to individual actions in the sequence is not directly provided. Indeed, although simplified versions of reinforcement learning known as bandit problems are studied, where it is assumed that rewards are provided after each action, reinforcement learning problems typically involve a general control-theoretic setting in which the learning task is to learn a control strategy (a "policy") for an agent acting in an unknown dynamical environment, where that learned strategy is trained to chose actions for any given state, with the objective of maximizing its expected reward over time. The ties to research in control theory and operations research have increased over the years, with formulations such as Markov decision processes and partially observed Markov decision processes providing points of contact (*15*, *16*). Reinforcement-learning algorithms generally make use of ideas that are familiar from the control-theory literature, such as policy iteration, value iteration, rollouts, and variance reduction, with innovations arising to address the specific needs of machine learning (e.g., large-scale problems, few assumptions about the unknown dynamical environment, and the use of supervised learning architectures to represent policies). It is also worth noting the strong ties between reinforcement learning and many decades of work on learning in psychology and neuroscience, one notable example being the use of reinforcement learning algorithms to predict the response of dopaminergic neurons in monkeys learning to associate a stimulus light with subsequent sugar reward (*17*).

Although these three learning paradigms help to organize ideas, much current research involves blends across these categories. For example, semi-supervised learning makes use of unlabeled data to augment labeled data in a supervised learning context, and discriminative training blends architectures developed for unsupervised learning with optimization formulations that make use of labels. Model selection is the broad activity of using training data not only to fit a model but also to select from a family of models, and the fact that training data do not directly indicate

which model to use leads to the use of algorithms developed for bandit problems and to Bayesian optimization procedures. Active learning arises when the learner is allowed to choose data points and query the trainer to request targeted information, such as the label of an otherwise unlabeled example. Causal modeling is the effort to go beyond simply discovering predictive relations among variables, to distinguish which variables causally influence others (e.g., a high white-blood-cell count can predict the existence of an infection, but it is the infection that causes the high white-cell count). Many issues influence the design of learning algorithms across all of these paradigms, including whether data are available in batches or arrive sequentially over time, how data have been sampled, requirements that learned models be interpretable by users, and robustness issues that arise when data do not fit prior modeling assumptions.

## Emerging trends

The field of machine learning is sufficiently young that it is still rapidly expanding, often by inventing new formalizations of machine-learning problems driven by practical applications. (An example is the development of recommendation systems, as described in Fig. 4.) One major trend driving this expansion is a growing concern with the environment in which a machine-learning algorithm operates. The word "environment" here refers in part to the computing architecture; whereas a classical machine-learning system involved a single program running on a single machine, it is now common for machine-learning systems to be deployed in architectures that include many thousands or ten of thousands of processors, such that communication constraints and issues of parallelism and distributed processing take center stage. Indeed, as depicted in Fig. 5, machine-learning systems are increasingly taking the form of complex collections of software that run on large-scale parallel and distributed computing platforms and provide a range of algorithms and services to data analysts.

The word "environment" also refers to the source of the data, which ranges from a set of people who may have privacy or ownership concerns, to the analyst or decision-maker who may have certain requirements on a machine-learning system (for example, that its output be visualizable), and to the social, legal, or political framework surrounding the deployment of a system. The environment also may include other machine-learning systems or other agents, and the overall collection of systems may be cooperative or adversarial. Broadly speaking, environments provide various resources to a learning algorithm and place constraints on those resources. Increasingly, machine-learning researchers are formalizing these relationships, aiming to design algorithms that are provably effective in various environments and explicitly allow users to express and control trade-offs among resources.

As an example of resource constraints, let us suppose that the data are provided by a set of individuals who wish to retain a degree of privacy. Privacy can be formalized via the notion of "differential privacy," which defines a probabilistic channel between the data and the outside world such that an observer of the output of the channel cannot infer reliably whether particular individuals have supplied data or not (*18*). Classical applications of differential privacy have involved insuring that queries (e.g., "what is the maximum balance across a set of accounts?") to a privatized database return an answer that is close to that returned on the nonprivate data. Recent research has brought differential privacy into contact with machine learning, where queries involve predictions or other inferential assertions (e.g., "given the data I've seen so far, what is the probability that a new transaction is fraudulent?") (*19*, *20*). Placing the overall design of a privacy-enhancing machine-learning system within a decision-theoretic framework provides users with a tuning knob whereby they can choose a desired level of privacy that takes into account the kinds of questions that will be asked of the data and their own personal utility for the answers. For example, a person may be willing to reveal most of their genome in the context of research on a disease that runs in their family but may ask for more stringent protection if information about their genome is being used to set insurance rates.

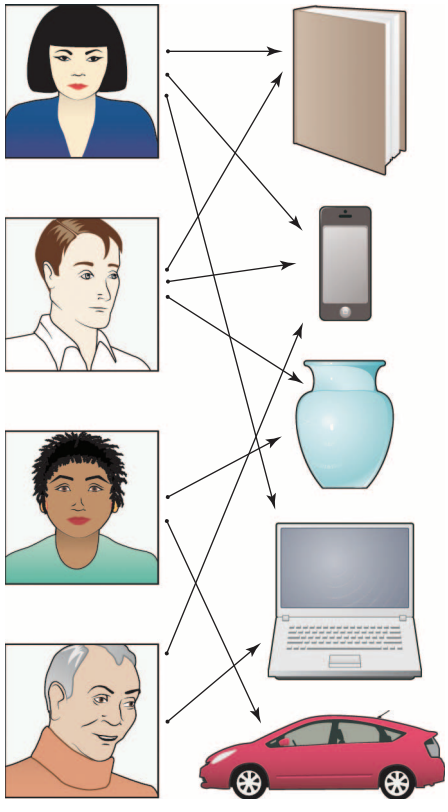


**Fig. 4. Recommendation systems.** A recommendation system is a machine-learning system that is based on data that indicate links between a set of a users (e.g., people) and a set of items (e.g., products). A link between a user and a product means that the user has indicated an interest in the product in some fashion (perhaps by purchasing that item in the past). The machine-learning problem is to suggest other items to a given user that he or she may also be interested in, based on the data across all users.

Communication is another resource that needs to be managed within the overall context of a distributed learning system. For example, data may be distributed across distinct physical locations because their size does not allow them to be aggregated at a single site or because of administrative boundaries. In such a setting, we may wish to impose a bit-rate communication constraint on the machine-learning algorithm. Solving the design problem under such a constraint will generally show how the performance of the learning system degrades under decrease in communication bandwidth, but it can also reveal how the performance improves as the number of distributed sites (e.g., machines or processors) increases, trading off these quantities against the amount of data (*21*, *22*). Much as in classical information theory, this line of research aims at fundamental lower bounds on achievable performance and specific algorithms that achieve those lower bounds.

A major goal of this general line of research is to bring the kinds of statistical resources studied in machine learning (e.g., number of data points, dimension of a parameter, and complexity of a hypothesis class) into contact with the classical computational resources of time and space. Such a bridge is present in the "probably approximately correct" (PAC) learning framework, which studies the effect of adding a polynomial-time computation constraint on this relationship among error rates, training data size, and other parameters of the learning algorithm (*3*). Recent advances in this line of research include various lower bounds that establish fundamental gaps in performance achievable in certain machine-learning problems (e.g., sparse regression and sparse principal components analysis) via polynomial-time and exponential-time algorithms (*23*). The core of the problem, however, involves time-data tradeoffs that are far from the polynomial/exponential boundary. The large data sets that are increasingly the norm require algorithms whose time and space requirements are linear or sublinear in the problem size (number of data points or number of dimensions). Recent research focuses on methods such as subsampling, random projections, and algorithm weakening to achieve scalability while retaining statistical control (*24*, *25*). The ultimate goal is to be able to supply time and space budgets to machine-learning systems in addition to accuracy requirements, with the system finding an operating point that allows such requirements to be realized.

## Opportunities and challenges

Despite its practical and commercial successes, machine learning remains a young field with many underexplored research opportunities. Some of these opportunities can be seen by contrasting current machine-learning approaches to the types of learning we observe in naturally

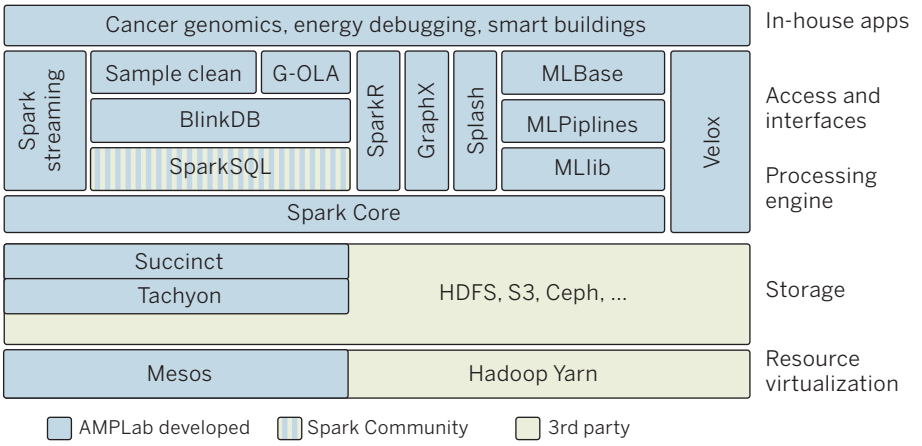



**Fig. 5. Data analytics stack.** Scalable machine-learning systems are layered architectures that are built on parallel and distributed computing platforms. The architecture depicted here—an open-source data analysis stack developed in the Algorithms, Machines and People (AMP) Laboratory at the University of California, Berkeley—includes layers that interface to underlying operating systems; layers that provide distributed storage, data management, and processing; and layers that provide core machine-learning competencies such as streaming, subsampling, pipelines, graph processing, and model serving.

occurring systems such as humans and other animals, organizations, economies, and biological evolution. For example, whereas most machine-learning algorithms are targeted to learn one specific function or data model from one single data source, humans clearly learn many different skills and types of knowledge, from years of diverse training experience, supervised and unsupervised, in a simple-to-more-difficult sequence (e.g., learning to crawl, then walk, then run). This has led some researchers to begin exploring the question of how to construct computer lifelong or never-ending learners that operate nonstop for years, learning thousands of interrelated skills or functions within an overall architecture that allows the system to improve its ability to learn one skill based on having learned another (*26*–*28*). Another aspect of the analogy to natural learning systems suggests the idea of team-based, mixed-initiative learning. For example, whereas current machine-learning systems typically operate in isolation to analyze the given data, people often work in teams to collect and analyze data (e.g., biologists have worked as teams to collect and analyze genomic data, bringing together diverse experiments and perspectives to make progress on this difficult problem). New machine-learning methods capable of working collaboratively with humans to jointly analyze complex data sets might bring together the abilities of machines to tease out subtle statistical regularities from massive data sets with the abilities of humans to draw on diverse background knowledge to generate plausible explanations and suggest new hypotheses. Many theoretical results in machine learning apply to all learning systems, whether they are computer algorithms, animals, organizations, or natural evolution. As the field progresses, we may see machine-learning theory and algorithms increasingly providing models for understanding learning in neural systems, organizations, and biological evolution and see machine learning benefit from ongoing studies of these other types of learning systems.

As with any powerful technology, machine learning raises questions about which of its potential uses society should encourage and discourage. The push in recent years to collect new kinds of personal data, motivated by its economic value, leads to obvious privacy issues, as mentioned above. The increasing value of data also raises a second ethical issue: Who will have access to, and ownership of, online data, and who will reap its benefits? Currently, much data are collected by corporations for specific uses leading to improved profits, with little or no motive for data sharing. However, the potential benefits that society could realize, even from existing online data, would be considerable if those data were to be made available for public good.

To illustrate, consider one simple example of how society could benefit from data that is already online today by using this data to decrease the risk of global pandemic spread from infectious diseases. By combining location data from online sources (e.g., location data from cell phones, from credit-card transactions at retail outlets, and from security cameras in public places and private buildings) with online medical data (e.g., emergency room admissions), it would be feasible today to implement a simple system to telephone individuals immediately if a person they were in close contact with yesterday was just admitted to the emergency room with an infectious disease, alerting them to the symptoms they should watch for and precautions they should take. Here, there is clearly a tension and trade-off between personal privacy and public health, and society at large needs to make the decision on how to make this trade-off. The larger point of this example, however, is that, although the data are already online, we do not currently have the laws, customs, culture, or mechanisms to enable society to benefit from them, if it wishes to do so. In fact, much of these data are privately held and owned, even though they are data about each of us. Considerations such as these suggest that machine learning is likely to be one of the most transformative technologies of the 21st century. Although it is impossible to predict the future, it appears essential that society begin now to consider how to maximize its benefits.

## REFERENCES


1. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, 2011).
2. K. Murphy, *Machine Learning: A Probabilistic Perspective* (MIT Press, Cambridge, MA, 2012).
3. L. Valiant, *Commun. ACM* **27**, 1134–1142 (1984).
4. V. Chandrasekaran, M. I. Jordan, *Proc. Natl. Acad. Sci. U.S.A.* **110**, E1181–E1190 (2013).
5. S. Decatur, O. Goldreich, D. Ron, *SIAM J. Comput.* **29**, 854–879 (2000).
6. S. Shalev-Shwartz, O. Shamir, E. Tromer, Using more data to speed up training time, *Proceedings of the Fifteenth Conference on Artificial Intelligence and Statistics*, Canary Islands, Spain, 21 to 23 April, 2012.
7. S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, in *Foundations and Trends in Machine Learning 3* (Now Publishers, Boston, 2011), pp. 1–122.
8. S. Sra, S. Nowozin, S. Wright, *Optimization for Machine Learning* (MIT Press, Cambridge, MA, 2011).
9. J. Schmidhuber, *Neural Netw.* **61**, 85–117 (2015).
10. Y. Bengio, in *Foundations and Trends in Machine Learning 2* (Now Publishers, Boston, 2009), pp. 1–127.
11. A. Krizhevsky, I. Sutskever, G. Hinton, *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2015).
12. G. Hinton *et al.*, *IEEE Signal Process. Mag.* **29**, 82–97 (2012).
13. G. E. Hinton, R. R. Salakhutdinov, *Science* **313**, 504–507 (2006).
14. V. Mnih *et al.*, *Nature* **518**, 529–533 (2015).
15. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998).
16. E. Yaylali, J. S. Ivy, Partially observable MDPs (POMDPs): Introduction and examples. *Encyclopedia of Operations Research and Management Science* (John Wiley, New York, 2011).
17. W. Schultz, P. Dayan, P. R. Montague, *Science* **275**, 1593–1599 (1997).
18. C. Dwork, F. McSherry, K. Nissim, A. Smith, in *Proceedings of the Third Theory of Cryptography Conference*, New York, 4 to 7 March 2006, pp. 265–284.
19. A. Blum, K. Ligett, A. Roth, *J. ACM* **20**, (2013).
20. J. Duchi, M. I. Jordan, J. Wainwright, *J. ACM* **61**, 1–57 (2014).
21. M.-F. Balcan, A. Blum, S. Fine, Y. Mansour, Distributed learning, communication complexity and privacy. *Proceedings of the 29th Conference on Computational Learning Theory*, Edinburgh, UK, 26 June to 1 July 2012.
22. Y. Zhang, J. Duchi, M. Jordan, M. Wainwright, in *Advances in Neural Information Processing Systems 26*, L. Bottou, C. Burges, Z. Ghahramani, M. Welling, Eds. (Curran Associates, Red Hook, NY, 2014), pp. 1–23.
23. Q. Berthet, P. Rigollet, *Ann. Stat.* **41**, 1780–1815 (2013).
24. A. Kleiner, A. Talwalkar, P. Sarkar, M. I. Jordan, *J. R. Stat. Soc., B* **76**, 795–816 (2014).
25. M. Mahoney, *Found. Trends Machine Learn.* **3**, 123–224 (2011).
26. T. Mitchell *et al.*, *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI-15)*, 25 to 30 January 2015, Austin, TX.
27. M. Taylor, P. Stone, *J. Mach. Learn. Res.* **10**, 1633–1685 (2009).
28. S. Thrun, L. Pratt, *Learning To Learn* (Kluwer Academic Press, Boston, 1998).
29. L. Wehbe *et al.*, *PLOS ONE* **9**, e112575 (2014).
30. K. Xu *et al.*, *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, Lille, France, 6 to 11 July 2015, pp. 2048–2057.
31. D. Blei, *Commun. ACM* **55**, 77–84 (2012).


10.1126/science.aaa8415