Artificial intelligence:
Scientific issues and socioeconomic expectations

Stéphan Clemençon,
professor of applied mathematics, Telecom-ParisTech, Institut Mines-Télécom

Abstract:
Beyond the excitement in the media, apart from the fears and hopes aroused, most of the recent successes of artificial intelligence rely on scientific concepts that were formulated several decades ago. This is even more so in machine learning. This field at the interface of mathematics and computer science seeks to develop techniques for automatically analyzing masses of data, usually for predictive purposes. These successes are being presented as harbingers of a revolution that will spare no field of human activity.

Let us face it: the acceleration of the advances made during recent years in fields such as image or voice recognition or automatic searches mainly stems from the spectacular progress in the infrastructure for doing calculations and storing data.[1] It does not, therefore, come as a surprise to see the giants of the Internet in the lead with applications. These big players in developing new technology are eager to convert the information contained in masses of accumulated data into new products and services.

Through machine learning, a computer automatically accomplishes a task — for example, recognizing a specific object in an image — via a program "learned" by solving the statistical version of an optimization problem. It thus demonstrates the effectiveness of the rules that, in a complex catalog, have been formulated using data from a large number of observations. Sample images associated with labels are used to test whether or not the object figures in a given image — a "supervised learning" characteristic of the applications that recognize shapes. In some cases, the machine itself searches for sample images in order to learn. The samples, stored in the huge databases, offer a nearly exhaustive description of the variability of the image to be analyzed. This availability of samples is combined with the computational power of clusters of modern computers that run programs of optimization with a complex catalog of predictive rules. This combination has made operational the statistical learning methods designed in the late 20th century and theorized by, in particular, the Russian mathematicians V. Vapnik and A. Chervonenkis during the 1960s (a prefiguration being Rosenblatt's Perceptron algorithm in the 1950s, the most elementary version of neural networks). These methods are now applied in several "intelligent systems" for purposes as different as biometrics, self-driving vehicles, automatic medical diagnoses, virtual assistants and the recommendation systems on e-business websites.

The era of big data and generalized artificial intelligence (AI) has opened. It uses technological bricks to automatically store and process in a short time a massive quantity of data of various sorts and formats. The Web giants, increasingly successful in this field, seem to be the first to have grasped the omnipresent function that data are now going to have in the information sciences and technology. The infatuation with machine learning is spreading to nearly all fields (science, transportation, energy, medicine, security, banking, insurance, commerce, etc.) as the Internet of things (IoT) and the widespread use of technology for analytics (*e.g.*, mass spectrometry or the cloud) make more data available with an ever finer granularity.

[1] This article has been translated from French by Noal Mellott (Omaha Beach, France).

Expectations are high. AI is supposed to allow for the development of a personalized medicine that will adapt a treatment to the patient's genetic traits. It is to be used to design systems of predictive maintenance for complex infrastructures, such as electricity grids. It will help make aircraft with systems for the early detection of "weak" signals that announce breakdowns, and will thus serve to plan the replacement of components before their probable failure. The vehicles using AI will be safer and fully autonomous, and be in service longer. There is no denying the opportunities, and we can rightfully hope for operational applications with big data as the input.

However AI will keep its promises only if certain issues are addressed. Before becoming THE solution, machine learning raises several exciting problems, in particular for mathematicians.

The three *v*'s of big data: Variety, volume and velocity

Often enough mentioned in relation to big data, the three *v*'s point toward some of the mathematical problems to be solved. They are, therefore, a good starting point for understanding the scientific challenges of AI.

Data processing is, of course, not a novelty. Statistics started developing along with the social sciences as early as the 18th century. They have grown along with agronomy, epidemiology, clinical tests, quality controls in industry, econometrics and, more recently, the Web. While data (usually collected via questionnaires) were scarce and costly, "traditional" statistics depended on human expertise, which was needed to preprocess data and build variability models.

Motivated at the origin by problems related to the recognition of shapes and stimulated by the cognitive sciences, machine learning does not serve just to process data and adjust the parameters of a more or less rigid, predefined model. It is used to design algorithms that "learn" automatically, from examples, which model in an immense catalog of models will be the most efficient. The efficiency of learning algorithms (*e.g.*, "neural networks", "support vector machines" or "random forests") and the development of optimized software for implementing them have popularized the use of AI over the past two decades. Machine learning draws from several fields of applied mathematics, in particular and of course, from probability, since it provides the right language for describing data variability, statistically formulating predictive problems and optimizing or processing images and signals. The problems it can be used to tackle are diverse.

Research is stimulated by the nature and format of the data available for analysis, but the data evolve along with the techniques used to collect them. The phrase "big data" mainly brings to mind data from the Internet, browsing preferences and the contents of Web pages (images, videos, sound clips, text, uses and social relations). This is far from the "rectangular matrices" of data used by 20th century statisticians for placing, in the rows, the individuals representative of the population under study and, in the columns, the (few) variables used to describe them. At the start, the semantic analysis of textual data on the Web mainly sought to index documents so that search engines could be run. It was more or less limited to a normalized count of each word from a dictionary on each web page. In contrast, recent applications, such as virtual assistants or the analytics of opinions or feelings, require much more sophisticated mathematical models.

Some web data, in particular from the social media, can also be used in graphs to describe the relations between objects or individuals. The theory of random graphs was worked out in the second half of the 20th century. Studies by mathematicians such as P. Erdös formulated a wide variety of problems about how information circulates over a network and how given parts of the network exert influence. Modern technology presents us with graphs representing an unprecedented wealth of details. To visualize and exhaustively use the information they convey, the complexity of their properties and dimensions calls for new methods to understand their structure and evolution. The graphs will probably remain a field of mathematical investigation for a long time to come.

In the same vein, the analysis of data on preferences is not a new topic; it dates back to the pioneering studies by Condorcet during the 18th century and social choice theory. Once again, the data now being collected via digital applications, whereby users express their preferences for given products, are of quite different sorts. The actually expressed preferences vary depending on the user and involve a very small number of objects compared with the dimensions of the catalog. It would be in vain to build a model of all cataloged preferences, since the number of arrangements of objects by preference explodes with the number of objects. Mathematicians are still looking for an efficient way to present this type of data. Consequent to wavelet theory as developed by Yves Meyer (recipient of the prestigious Abel Prize in 2017) for signals, images and videos, the standard JPEG 2000 (JP2) has been proposed.

A major problem in mathematics is to find a sparse, adaptative representation of information and work out algorithms for quickly calculating it. This is the key to efficient data-processing. A representation will be all the more operational insofar as it eliminates “noise”, highlights “patterns” and augments the predictive capacity of the algorithms of statistical learning that uses the thus formatted data as input. In several cases, the many layers of deep neural networks produce representations that make it easier to adjust the rules of prediction to the right degree of generalization.

As biometrics is deployed on smartphones, the operation of embedded predictive models should not compromise the autonomy of the systems where they are embedded and should take account of the constraints of (nearly) realtime operations. This raises questions about compressing information and the rules for processing it. The IoT is coming. Smart sensors (for applications such as predictive maintenance in transportation systems) connected in a network will be able to come up with the best strategy for sharing information and distributing the tasks of calculation as a function of the data collected and of the tasks to be performed.

Several scientific questions crop up. In some cases, the level of “delegation” to be granted to “smart systems” will heavily depend on how research in methodology will reply to questions of ethics (the processing of personal data while respecting privacy) and reliability. This brings us back to the development of statistical learning techniques that hold up even if part of the data have been “contaminated” (for example due to biases in measurements or the deliberate intention to impair the operation of the automated system, as suspected during the last presidential election in the United States). Not only must they hold up under such circumstances, but they must also yield decisions that end users (human beings) can interpret. The adventure has just started…