Big data:
Technological issues and scientific effects

Stephan Clémençon,
Professor of applied mathematics, Télécom-ParisTech, Institut Mines-Télécom

Abstract:
The mathematical and algorithmic concepts used for machine learning and predictive analytics are not all that new, but they are now being widely used owing to the exploding quantity of available data. The phenomenon of big data both attracts and frightens. The risks related to it can be controlled only if people (beyond the small circle of data scientists) understand how probability and statistics are used to handle big data.

Mentioning "big data" usually sets off ambivalent reactions.[1] For one thing, the fears usually aroused by the quite real dangers: the automation of decision-making processes along with a loss of control, the negative impact on employment, the dependance on information systems, and the deprivation of privacy. But, definitely too, an enthusiasm for what could be accomplished in science, medicine, commerce, transportation, communications and security by combining the masses of data now available with the thriving information sciences, in particular with machine learning (following the example of the advances made over the past twenty years in fields such as image or voice recognition). Although it is still hard to know how to best organize efficient regulations without hampering the promised advances, controlling the risks entails, among other things, education and training via a wider diffusion of a "culture of data and algorithms".

The fears aroused by automation are not new. To process masses of electronic data, automation is inevitable and desirable. Wrongly perceived as a discipline with the goal of replacing the expertise of human operators with machines for automated data processing, machine learning has, instead, the objective of helping us process the raw data collected by modern sensors (telescopes, mass spectrometers, mobile telephones). These data contain complex information that is absolutely unfathomable unless computer programs perform the appropriate mathematical operations. Machine learning is now used in several fields. Its successes stand out in applications such as video surveillance and monitoring, the predictive maintenance of big systems and infrastructures, and the recommendation systems used on the Web.

We can predict that the body of knowledge and techniques at the interface between mathematics and computer science, which has made constant progress in recent decades, will be at the origin of many an innovation with a strong societal, economic or scientific impact — but only if an ever larger public understands the potential of this knowledge, if a growing number of engineers and technicians control it, and if this body of knowledge addresses the concerns of modern society. The real danger of automating the processing of big data comes, on the contrary, from the lack of expertise and skills for: verifying the conditions under which the data have been collected; guaranteeing the veracity of data, validating the solidity of the statistical models used by modern applications, and interpreting the results.

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

The paradigm of statistical learning

The recognition of patterns is one of the best examples of big data's impact. This application of artificial intelligence is most frequently mentioned to illustrate the efficiency of solutions such as computer vision, or the automatic recognition of speech or handwriting.

The mathematical and algorithmic concepts used to develop intelligent systems are far from new. Most of them date back more than half a century, but have undergone significant improvement since then.

In all these problem areas, the machine's task is to take as input a datum *X* and to automatically recognize (with a minimal margin of error) it as belonging to the category *Y*, specified in advance as being a possibility for *X*. In biometrics for example, *X* might be an image in pixels or a sound signal; and *Y*, the identity of the person on the image or in a recording of the voice. This technology is being deployed for assistance with medical diagnoses (or prognoses) or for the risk management of loans. We realize that, when the input *X* is less of a determinant of the category *Y*, the expected margin of error will be much higher in these fields than in biometrics. Pattern recognition is a predictive problem insofar as the rules have to be both applicable by using the available software databases and capable of minimizing the errors made when using training data that contain past instances (*X*, *Y*). Furthermore, these rules must efficiently predict the label *Y* for any new input *X* of instances that have not yet been observed but come from the same statistical population as the examples in the training database (It is, of course, always easy to "predict" the past). This is the predictive rule's generalization ability.

Formulating the problem of learning a rule relies, naturally, on the language of probability. And solving the problem in practice means selecting a predictive rule using an optimization algorithm that operates on a given class of rules to be tested and minimizes a statistical version of the calculated probability of error using examples stored in the training database. The mathematical theory formulated by Vladimir Vapnik at the end if the 1960s proves the generalization ability of the rules designed in this way, under condition that the classes used for machine learning are of a controlled degree of complexity. The validation framework that this theory has provided to statistical learning methods corresponds to a very active current of research at the interface between several disciplines (mathematics, computer science and, too, the cognitive sciences).

The impact of big data

Even though the fundamental concepts of machine learning and certain algorithms (such as neural networks) had been worked out by the end of the 1970s, machine learning did not start its rise toward the success it now has till approximately a decade ago: the start of the "big data era". At the time, one major impediment was the scarcity of electronic data, since data were usually collected with expensive sampling methods. Another impediment was related to the limitations on memory and on the computational power of machines, which kept us from implementing optimization programs that could operate on vast classes of rules so as to make machine learning efficient.

In many situations, the low predictability of the rules produced through machine learning could be attributed both to the stochastic error stemming from the small number of examples used for the learning process and to the crudeness of the predictive models that formed the classes to which the optimization programs were applied. The technological blocks used to build the Web, such as distributed file systems (using Hadoop or programming languages such as MapReduce), set the conditions for considerable progress in collecting and storing data and in the distributed, parallel, processing of masses of data. Megadata from the Web — the enormous libraries of "labeled" images, sounds or texts accessible via the Web — can be used as a vast set of training data for programs of content recognition. Thanks to the advances made in managing

computer memory and in parallel computing (in particular graphic processes), learning programs have been launched that operate on flexible classes of data (such as neural networks or deep learning) that can, for several problems, efficiently explain how the input information *X* can be used to predict the output *Y*. The ubiquity of sensors and the development of the Internet of things are making it easier to access electronic data; and countless applications are being developed in line with pattern recognition technology.

The infrastructures for collecting and managing masses of data and for making computations are not the only factors conditioning the progress made in machine learning. Nor will the future be limited to enumerating deep learning applications. For example, to embed biometrical recognition technology in a smartphone without compromising the device's autonomy, engineers must understand how to "compress deep networks" so as to limit the energy used without impairing the quality of the pattern recognition technology.

Big data are, for stochastic learning, a sort of nirvana, since the methods will be all the more reliable insofar as they are based on observing a "large number" of instances. However it is indispensable to control the conditions for acquiring data and the hypotheses about the validity of predictive algorithms if the models calculated by computers are to be successful. A probabilistic and statistical education should increasingly be provided in university courses (not just for data scientists, the new specialists of statistics and algorithms). The broader diffusion of an education of this sort would dissipate the fear of a world where big data will be used to predict without error our behavior patterns or our death day… "Large numbers" can be used to estimate the predictive performance of models, to accurately assess risks and to optimize decisions in an uncertain universe, but not to reduce the inherent randomness of certain phenomena.