cnrs
international magazine
n° 20
Quarterly
January 2011

THE FUTURE OF Computing Science

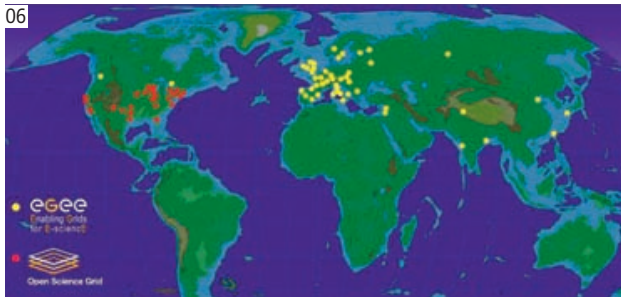
Gérard Férey
Recipient of the
CNRS 2010 Gold Medal

cnrs
advancing the frontiers

The Future of Computing Science

Navigating the Datasphere

Whether you're a tourist looking for the cheapest airline ticket, a physicist analyzing data from a particle accelerator, or an employee at a temp agency sifting through applications, you all have something in common," says Amedeo Napoli from the IT specialized LORIA[1] laboratory in Vandœuvre-lès-Nancy. "You are trying to extract specific information from a huge amount of data." In principle, solving this problem couldn't be simpler: prepare the initial data, feed it to a data mining algorithm, and wait for the system to provide the results in the required format. But in a world where that volume of data is increasing relentlessly, extracting pertinent knowledge becomes a seemingly impossible task.



06 Location of sites involved in the world's two largest grid infrastructures: Egee in Europe (in yellow) and OSG in the US (in red).

Looking for a holiday flight, hotel, and rental car, all at the lowest possible price, is a good example. As Michel Beaudouin-Lafon, from the LRI in Orsay[2] puts it: "Mathematically, we know that the complexity of this type of problem makes it impossible to find an exact solution in reasonable time, given the massive amount of input data." Therefore, in practice, programmers must find clever ways of obtaining the most accurate result within reasonable time. In fact, the burgeoning field of data mining brings together specialists from fields as different as computer science, of course, but also machine architecture, linguistics, and mathematics. These specialists use artificial intelligence, databases, learning techniques, and statistical methods.

OPTIMIZING DATA SIFTING

One thing is certain: every field needs to develop efficient methods to avoid being flooded with unusable data often impossible to store. Take the French Midas project,[3] for example. It brings together, among others, CNRS labs and companies that have to deal with complex sets of data, like the telecommunications company Orange or the French energy provider EDF. Its goal is to develop an algorithm able to condense a large amount of data generated in real time so that it can be stored in a limited central memory for later use. "This is typically the type of situation that France Télécom, EDF, or the French national railway company SNCF have to deal with every day," says Pascal Poncelet of the LIRMM,[4] in Montpellier. "For example, a TGV high-speed train records 250 data points per carriage every five minutes to anticipate maintenance operations. But such a huge amount of data is impossible to store. Events must therefore be sorted by order of importance, which changes over time."

Scientists themselves are heavy users of data mining techniques. The LHC, CERN'S giant particle collider in Geneva, is a prime example. When it reaches its full capacity, 40 million proton collisions will occur every second. Yet physicists estimate that just 100 of those will be of interest and will need to be recorded. Such events will have to be selected in real time using specialized algorithms. "These are typically learning algorithms where the computer's performance improves as it processes the new data to be kept or rejected," explains Beaudouin-Lafon, whose laboratory is involved with the LAL,[5] to elaborate ways of analyzing the

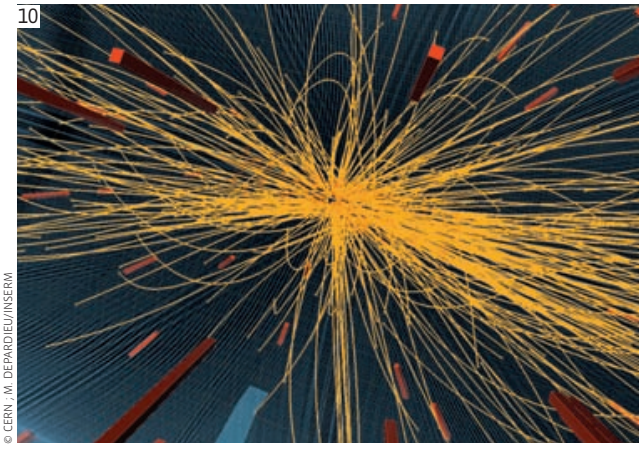
MAKING PICTURES TALK

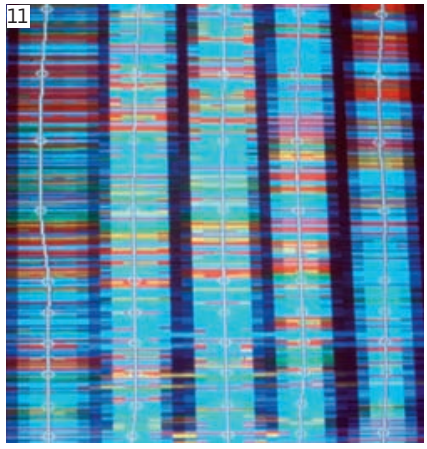
If you think sorting family pictures on your home computer is a hassle, imagine sifting through the largest image databases in existence, which contain millions. Luckily, tools like face recognition software are already available. Yet as Matthieu Cord of the LIP6[1] points out, "the success rate is only 50-60%." Typically, a specialized algorithm can perfectly handle so-called "low-level" information: color, contrast, or pixel movement vectors in a video stream, for example. It is somewhat trickier to transform this data into high-level information making it possible to positively identify a particular object or event. This has not prevented the emergence of increasingly powerful applications, like the one developed by Jenny Benois-Pineau's team at the LaBRI,[2] a laboratory near Bordeaux working in conjunction with the French national medical research center (Inserm). "We film Alzheimer patients at home with wearable cameras, and identify behaviors associated with the disease so that doctors can follow a patient's evolution," Benois-Pineau explains. Cord is working on the iTowns project, a digital map of Paris elaborated from photographs, and modeled after Google Street View—but with an accuracy of just one centimeter. "We are developing tools to automatically detect people and cars in order to blur personal data," he explains. "But we are also working on the recognition of a multitude of objects more or less buried in these images, such as street signs, façades, or vegetation, in order to improve advanced navigation."

01. Laboratoire d'informatique de Paris-6 (CNRS / UPMC).
02. Laboratoire Bordelais de recherche en informatique (CNRS / Université Bordeaux-I / IPB Enseirb-Matmeca Bordeaux / Université Victor-Segalen).

CONTACT INFORMATION:
Jenny Benois-Pineau
› jenny.benois-pineau@labri.fr
Matthieu Cord
› matthieu.cord@lip6.fr

07 08 09 iTowns automatically extracts all types of information contained in an image.


10


11

© CERN ; M. DEPARDIEU/INSERM


12

© C. LEBEDINSKY/INRIA

10 11 Some experiments, such as particle collisions (left) or genome sequencing (right), generate large amounts of data that must be sorted and analyzed.
12 Analyzing scientific data sometimes requires extensive computing resources as well as interconnecting machines through networks, as shown here in the Grid 5000 project.

GRID COMPUTING

Grid computers are virtual infrastructures consisting of a set (or clusters) of computers, including home computers, that are geographically remote but working as a network. These systems, which emerged a few years ago to meet the demands of particle physics experiments, enable research scientists and industrialists to have access to extensive computing resources at lower cost, in sectors ranging from engineering to the study of neurodegenerative diseases or astrophysics. The CNRS's Grids Institute (Institut des grilles) managed by Vincent Breton, has been the leading research center in this field in France for the past three years. Along with Grid 5000, a tool specifically dedicated to grid research, it provides scientists and industry with a production grid comprising around 20,000 processors scattered over some 20 centers at CNRS, CEA,[1] and universities. Last September, this already sizeable system reached new heights with the creation of "France Grilles," involving several research organizations and universities. Its purpose is to coordinate the deployment of a nationwide grid infrastructure, which will eventually be integrated into a European grid. For Breton, who heads the program, its objective is clear-cut: "to double resources and users by 2015."

01. French Atomic Energy and Alternative Energies Commission.

CONTACT INFORMATION:
Vincent Breton
› vincent.breton@idgrilles.fr

huge amount of data provided by particle accelerators.

TRIAL AND ERROR

Particle physicists are far from being the only ones to handle large amounts of data. Pascal Poncelet's team, working in collaboration with researchers from the French medical research center Inserm, has developed an algorithm able to single out the genes involved in various types of breast cancer tumors, based on patient data (genetic information, age, weight and size of the tumor, treatment used, and results obtained). "It gives doctors information on the potential evolution of the tumor," the researcher explains.

In a different field, Amedeo Napoli's team has worked with astronomers to develop data mining software applied to information collected in astrophysics. Researchers hope this type of software will reveal particular characteristics or combinations that might have escaped a human operator.

Can data mining work miracles? Not exactly. It is a relatively new field, first explored at the end of the 1980s, and still in full expansion. For Beaudouin-Lafon, "most methods used today are empirical. Parameters are adjusted manually and when something works, it is not really clear why. In many cases, there are no quantitative criteria for judging the quality of information extracted from a database. That is left up to specialists in the field." Napoli adds: "much work still has to be done to handle very large amounts of data. At present, we can manage a few thousand objects with a few hundred attributes. Beyond that, the hardware's physical limits become apparent."

To overcome this obstacle, two complementary approaches are currently used. First, when a single machine does not have enough computing power for a specific task, several computers can be run in parallel. This is the principle of grid computing (*see box*), which LHC has pushed to the limit: it relies on 50,000 PCs located in various research centers worldwide to analyze the 15 millions Gigaoctets of scientific data (the equivalent of a

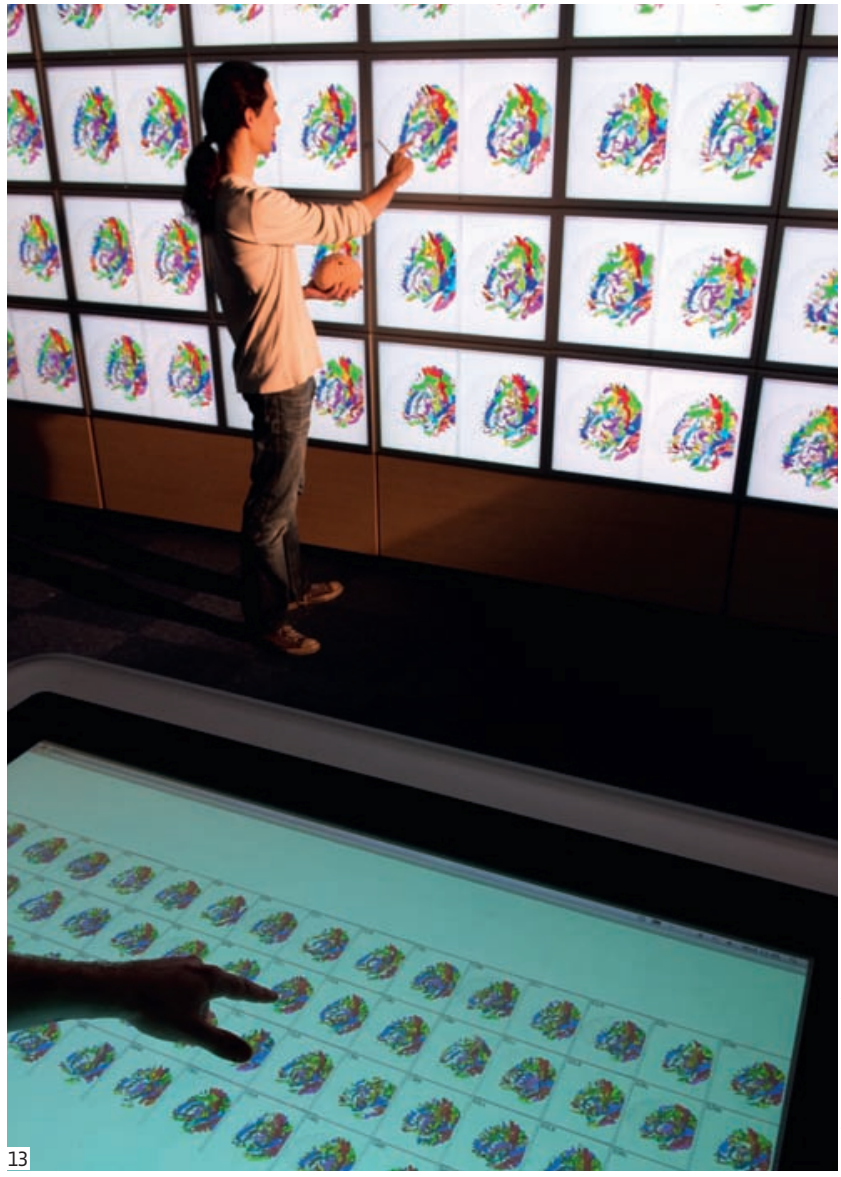
20km-long pile of CDs) that researchers will collect every year. The second approach is based on supercomputers, like the one used since 2008 at CNRS's IDRIS[6]—a monster capable of performing 207 thousand billion FLOPS. "In some cases, such as weather simulation, for which it is difficult to parcel out the data to a network of PCs, supercomputers remain the best solution," explains Beaudouin-Lafon.

FLOPS
FLOPS stands for floating-point operations per second. It is a measurement of a computer's performance. By comparison, an average handheld calculator can perform around 10 FLOPS.

THE HUMAN FACTOR

But developing fast and high-performing computers is not enough. The data they have sorted still need to be understandable to human users. Take Google for example: the search engine can bring up several thousand addresses for a query, but can only display a dozen or so per page. "It is a shame to have sophisticated data retrieval algorithms and yet not be able to display the results comprehensively," says Beaudouin-Lafon. This raises the question of how search results can best be presented.

To answer this question, the LRI has developed a new type of platform called Wild: a wall of 32 computer screens—over 130 million pixels in total—allowing users to grasp a huge amount of information at a glance. "We are working with eight other laboratories from CNRS and the Saclay Campus, on this project," says Beaudouin-Lafon. For neuroscience specialists, Wild can display 64 brain MRIs, "which offers an indisputable advantage when trying to identify a pathology, considering there are significant variations even among healthy brains," he adds. Similarly, in astrophysics, certain observatories now compile images much too large to be displayed on single computer screens. To view an entire image at its highest resolution, Wild-like tools make all the difference. "I am convinced that this type of approach will expand in the future—not only for research, but also for industry," concludes Beaudouin-Lafon. "Indeed, the amount of data is constantly expanding, and the questions asked are both increasingly vague and complex." In other words, everything must be done to prevent today's information society from drowning in this massive quantity of data.


13

© C. FRESILLON/CNRS PHOTOTHEQUE

13 The "Substance Grise" (Grey Matter) application used on the Wild platform allows users to simultaneously compare 3-D reconstructions of 64 patients' brains.

800,000 petabytes*

was the estimated amount of digital data in the world in 2009. This number was expected to rise to 1.2 million in 2010, and experts predict it to grow by 45% each year between now and 2020.

* 10^{15} bytes.

01. Laboratoire lorrain de recherche en informatique et ses applications (CNRS / Université Henri-Poincaré / Université Nancy-II / Inria).
02. Laboratoire de recherche en informatique (CNRS / Université Paris-Sud-XI).
03. Microwave Data Analysis for petascale computers.
04. Laboratoire d'informatique, de robotique et de microélectronique (CNRS / Université Montpellier-II).
05. Laboratoire de l'accélérateur linéaire (CNRS / Université Paris-Sud-XI).
06. Institut du développement et des ressources en informatique scientifique.

CONTACT INFORMATION:
Michel Beaudouin-Lafon
› michel.beaudouin-lafon@lri.fr
Amedeo Napoli
› amedeo.napoli@loria.fr
Pascal Poncelet
› pascal.poncelet@lirmm.fr