Science in an exponential world

The amount of scientific data is doubling every year. **Alexander Szalay** and **Jim Gray** analyse how scientific methods are evolving from paper notebooks to huge online databases.

Scientists are trained early to keep careful records in their laboratory notebooks — recording both experimental procedures and observations, so that they can analyse their results and so that others can replicate what they have done. Galileo did it, Mendel did it, Darwin did it, and we are supposed to do it. This worked fine when small amounts of data were entered into notebooks and the analysis was computed alongside them. But data volumes are doubling every year in most areas of modern science and the analysis is becoming more and more complex, exceeding the capacity of the paper notebook. With data correlated over many dimensions and millions of points, none of the old steps — do experiment, record results, analyse and publish — is straightforward. Many predict dramatic changes to the way science is done, and suspect that few traditional processes will survive in their current form by 2020 (ref. 1).

Today, most scientists have replaced or enhanced their notebooks with desktop computers that record their results, provide a portal to the scientific literature, and link them to collaborators via e-mail. These computers also perform data analysis; Matlab, Mathematica and Excel are popular analysis tools. But none of these programs scale up to handle millions of data records — and they are primitive by most standards. As data volumes grow, it is increasingly arduous to extract knowledge. Scientists must labour to organize, sort and reduce the data, with each analysis step producing smaller data sets that eventually lead to the big picture. Analysing terabytes of data (one terabyte is 1,000 gigabytes) is a challenge; but petabyte data sets (of more than 1,000 terabytes) are on the horizon. One petabyte is equivalent to the text in one billion books, yet many scientific instruments, including the Large Synoptic Survey Telescope, will soon be generating several petabytes annually.

In response to this data deluge, the systematic use of databases has become an integral part of the scientific process. Databases provide tools to organize large data sets, find objects that match certain criteria, compute statistics about the data, and analyse them to find patterns. Many experiments today load their data into databases before attempting to analyse them. But there are few tools to properly visualize data across multiple scales and data sets. If we can no longer examine all the data on a single piece of paper, how can we 'see' a new pattern or find a data point that does not fit a hypothesis? Fortunately there are database tools, such as data cubes, that we believe can fulfil this role (see 'Data cubes' overleaf).

The same language

Experiments are themselves becoming electronic as computers become essential parts of scientific instruments; they are used not only to manage and analyse vast data sets, but also to acquire them in the first place. Procedures already involve instruments and software with myriad parameters. It is difficult to capture all the model numbers, software revisions, parameter settings and process steps in an enduring format. For example, imagine a measurement taken using a DNA-sequencing machine. The output is cross-correlated with a sequence archive (GenBank) and the results are analysed with Matlab. Fully documenting these steps would be arduous, and there is little chance that someone could repeat the exact procedure 20 years from now; both Matlab and GenBank will change enormously in that time. As experiments yield more data, and analysis becomes more complex, data become increasingly difficult to document and reproduce.

One might argue that complex biological experiments have always been difficult to reproduce, as there are so many variables. But we believe that with current trends it is nearly impossible to reproduce experiments. We do not have a solution for this problem, but it is important to recognize it as such, and to do what is possible to capture the workflows and to develop protocols for documenting instruments, procedures and measurements in ways that will be usable in several decades' time.

Increasingly, scientists are analysing complex systems that require data to be combined from several groups and even several disciplines. There are collaborations sharing data across departments and time zones, and important discoveries are made by scientists and teams who combine different skill sets — not just biologists, physicists and chemists, but also computer scientists, statisticians and data-visualization experts. It is important to realize that today's graduate students need formal training in areas beyond their central discipline: they need to know some data management, computational concepts and statistical techniques.

A collaboration involving hundreds of Inter-

A. SZALAY & J. GRAY

JOHNS HOPKINS UNIV.

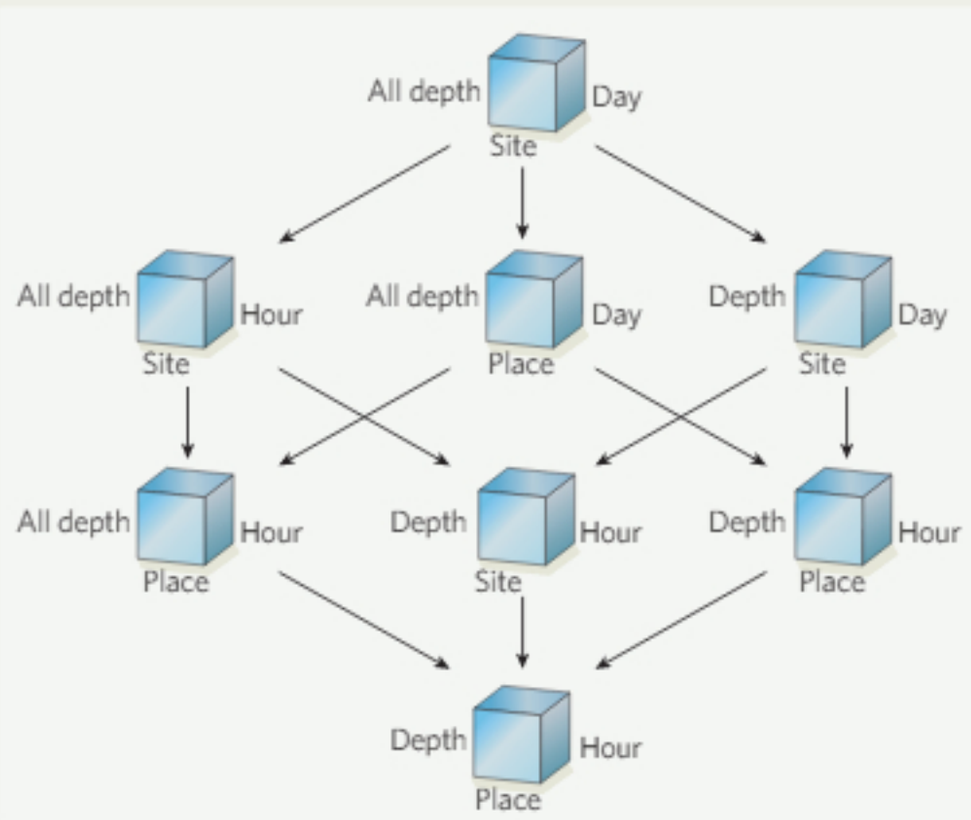
Automated systems will transform data collections, from astronomy (left) to sampling soil properties under our feet.

DATA CUBES

Traditional notebook and analysis tools are being challenged not just by data volumes, but also by data complexity. For example, the three-dimensional structural representation of a complex protein is not easily transcribed into a notebook.

Complex scientific data are often organized as a collection of independent variables and their dependent measurements. For example, meteorological data (temperature, pressure, humidity, wind velocity and direction) are collected at various times and locations (latitude, longitude and altitude). These data can be thought of as a multidimensional cube in which time and place exist in four dimensions and the measurements are shown by vectors at each point in the cube (pictured right).

A meteorologist may ask: show me the minima, maxima and average winds for Australia aggregating over times (hour, day, week, year) and volumes (per square kilometre of the atmosphere). The data cube makes it easy to express such user queries and to compute the answers, even to seemingly complex questions.



C. STOLTE & P. HANRAHAN, STANFORD UNIV./TABLEAU

net-connected scientists raises questions about standards for data sharing. Too much effort is wasted on converting from one proprietary data format to another. Standards are essential at several levels: in formatting, so that data written by one group can be easily read and understood by others; in semantics, so that a term used by one group can be translated (often automatically) by another without its meaning being distorted; and in workflows, so that analysis steps can be executed across the Internet and reproduced by others at a later date.

Standards for sharing data are crucial, for example, in understanding soil ecosystems. We are helping to build a system for measuring long-term environmental trends that affect soil biodiversity (www.lifeunderyourfeet.org; see also News Feature, page 402). This system integrates local environmental data from a sensor network with regional data on hydrology, climate, biodiversity and biogeochemistry. For these data to be useful to others, they must be published using a controlled vocabulary and in standard forms, and the instruments and measurements must be well specified. Fully documenting the sensors and data-collection process is arduous, and there are few standards for us to draw on.

Data gold-mine

Multidisciplinary databases also provide a rich environment for performing science; that is, a scientist may collect new data, combine them with data from other archives, and ultimately deposit the summary data back into a common archive. Many scientists no longer 'do' experiments the old-fashioned way. Instead they 'mine' available databases, looking for new patterns and discoveries, without ever picking up a pipette.

But this data-rich approach to science faces challenges. The speed of the Internet has not kept pace with the growth of scientific data sets. And so large data archives are becoming increasingly 'isolated' in the network sense — one can copy gigabytes across the Internet today, but not petabytes. In the future, working with large data sets will typically mean sending computations to the data, rather than copying the data to your workstation. But the management of distributed computations raises new questions of security, free access to public data and cost. Few data archives address these issues today.

Are we reaching the limits of what one scientist, or one lab, can expect to achieve in data handling and analysis? If so, this will have implications for how we review and publish our work. For example, a data-mining paper needs to include the explicit description (database query) of how the data that were analysed in the paper were collected and filtered, but not the data themselves. In this way, a reviewer with access to public data could reproduce the data sets and analysis procedures. For the analysis to be repeatable in 20 years' time requires archiving both data and tools.

The publication process itself is increasingly electronic, with new ways to disseminate scientific information (such as the preprint repository arXiv.org). But there is, as yet, no standard for publishing large volumes of data. Paper appendices cannot hold all the data needed to reproduce the results. Some disciplines have created their own data archives, such as GenBank; others just let data show up, and then disappear, on individual scientists' websites. Astronomers created the International Virtual Observatory Alliance (www.IVOA.net), integrating most of the world's medium and large astronomy archives. This required new standards for data exchange, and a semantic dictionary that offers a controlled vocabulary of astronomy terms.

To encourage data sharing, it should be rewarded. Public data creators and publishers should be given credit, and archives must be able to automatically provide provenance details. Current databases have a long way to go to achieve this ideal.

For how long will this exponential growth in scientific data continue? Desktop computers today are as powerful as the super-computers of 10 years ago. Similar progress is happening with scientific instruments — they quickly become obsolete and are replaced by better and often cheaper ones. Likely computer-performance improvements by 2011 include tenfold more processing, storage and network bandwidth per dollar. So we can expect ten times more data.

Smaller is faster

However, not all experiments will experience exponential growth. There is reason to believe that it will be the smaller experiments, not the big multibillion-dollar facilities, that will grow the fastest. Exponential growth occurs when a new generation of instruments leapfrogs the previous generation, which become obsolete. There are two trends in science today, scaling up and scaling out. Some scientists are building billion-dollar facilities, such as astronomy's Large Synoptic Survey Telescope or the Large Hadron Collider, which are only affordable as international collaborations. Such facilities are not easily leapfrogged. And once these peta-scale experiments are switched on they will produce roughly the same amount of data each year — merely linear growth. But in the scaling-out model, experiments that deploy an array of small instruments can exploit the coming explosion in cheaper commodity technology. The wireless sensors that were US$300 a year ago are $100 today, and will be $30 next year. A similar phenomenon occurred with DNA chips and gene sequencers. It is important to recognize this pattern; it is universal. And so although some sub-disciplines may reach a plateau in data generation, other technological innovations will take their place. Scientists in 2020 will continue to work in an exponential world. ■

Alexander Szalay is in the Department of Physics and Astronomy, Johns Hopkins University, Baltimore, Maryland 21218, USA.
Jim Gray is at Microsoft Research, San Francisco, California 94105, USA.

1. *Towards 2020 Science* (Microsoft, 2006); http://research.microsoft.com/towards2020science