The Economist

Home   World   Business & finance   Science & technology   Markets & data   Culture   Site Index   Print Edition

A special report on managing information

Data, data everywhere

Information has gone from scarce to superabundant. That brings huge new benefits, says Kenneth Cukier (interviewed here)—but also big headaches

Feb 25th 2010 | From *The Economist* print edition

Jac Depczyk
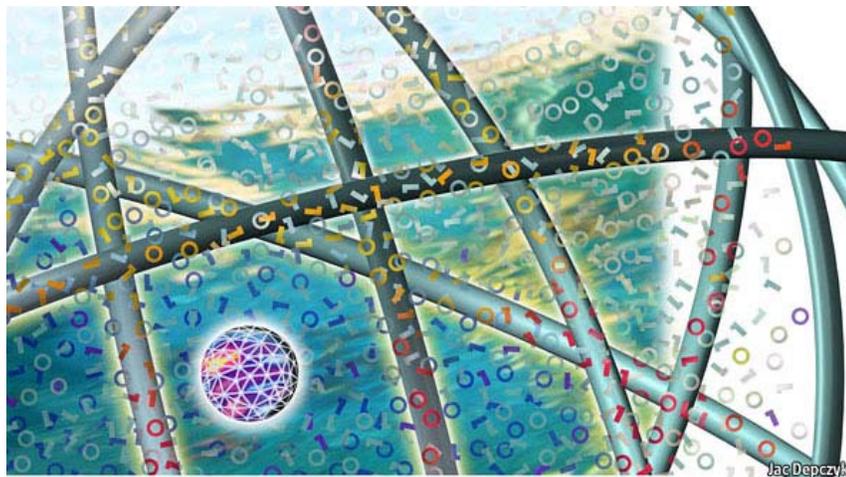
WHEN the Sloan Digital Sky Survey started work in 2000, its telescope in New Mexico collected more data in its first few weeks than had been amassed in the entire history of astronomy. Now, a decade later, its archive contains a whopping 140 terabytes of information. A successor, the Large Synoptic Survey Telescope, due to come on stream in Chile in 2016, will acquire that quantity of data every five days.
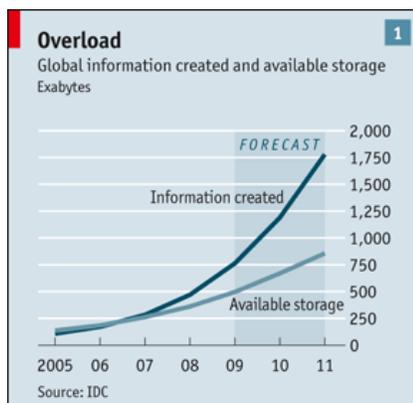
Such astronomical amounts of information can be found closer to Earth too. Wal-Mart, a retail giant, handles more than 1m customer transactions every hour, feeding databases estimated at more than 2.5 petabytes—the equivalent of 167 times the books in America's Library of Congress (see article for an explanation of how data are quantified). Facebook, a social-networking website, is home to 40 billion photos. And decoding the human genome involves analysing 3 billion base pairs—which took ten years the first time it was done, in 2003, but can now be achieved in one week.

All these examples tell the same story: that the world contains an unimaginably vast amount of digital information which is getting ever vaster ever more rapidly. This makes it possible to do many things that previously could not be done: spot business trends, prevent diseases, combat crime and so on. Managed well, the data can be used to unlock new sources of economic value, provide fresh insights into science and hold governments to account.

But they are also creating a host of new problems. Despite the abundance of tools to capture, process and share all this information—sensors, computers, mobile phones and the like—it already exceeds the available storage space (see chart 1). Moreover, ensuring data security and protecting privacy is becoming harder as the information multiplies and is shared ever more widely around the world.

Alex Szalay, an astrophysicist at Johns Hopkins University, notes that the proliferation of data is making them increasingly inaccessible. "How to make sense of all these data? People should be worried about how we train the next generation, not just of scientists, but people in government and industry," he says.

"We are at a different period because of so much information," says James Cortada of IBM, who has written a couple of dozen books on the history of information in society. Joe Hellerstein, a computer scientist at the University of California in Berkeley, calls it "[the industrial] revolution of data". The effect

Overload ⬛ 1
Global information created and available storage
Exabytes

FORECAST
2,000
1,750
1,500
Information created   1,250
1,000
750
500
Available storage   250
0
2005 06 07 08 09 10 11
Source: IDC

is being felt everywhere, from business to science, from government to the arts. Scientists and computer engineers have coined a new term for the phenomenon: "big data".

Epistemologically speaking, information is made up of a collection of data and knowledge is made up of different strands of information. But this special report uses "data" and "information" interchangeably because, as it will argue, the two are increasingly difficult to tell apart. Given enough raw data, today's algorithms and powerful computers can reveal new insights that would previously have remained hidden.

The business of information management—helping organisations to make sense of their proliferating data—is growing by leaps and bounds. In recent years Oracle, IBM, Microsoft and SAP between them have spent more than $15 billion on buying software firms specialising in data management and analytics. This industry is estimated to be worth more than $100 billion and growing at almost 10% a year, roughly twice as fast as the software business as a whole.

Chief information officers (CIOs) have become somewhat more prominent in the executive suite, and a new kind of professional has emerged, the data scientist, who combines the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data. Hal Varian, Google's chief economist, predicts that the job of statistician will become the "sexiest" around. Data, he explains, are widely available; what is scarce is the ability to extract wisdom from them.

More of everything

There are many reasons for the information explosion. The most obvious one is technology. As the capabilities of digital devices soar and prices plummet, sensors and gadgets are digitising lots of information that was previously unavailable. And many more people have access to far more powerful tools. For example, there are 4.6 billion mobile-phone subscriptions worldwide (though many people have more than one, so the world's 6.8 billion people are not quite as well supplied as these figures suggest), and 1 billion-2 billion people use the internet.

Moreover, there are now many more people who interact with information. Between 1990 and 2005 more than 1 billion people worldwide entered the middle class. As they get richer they become more literate, which fuels information growth, notes Mr Cortada. The results are showing up in politics, economics and the law as well. "Revolutions in science have often been preceded by revolutions in measurement," says Sinan Aral, a business professor at New York University. Just as the microscope transformed biology by exposing germs, and the electron microscope changed physics, all these data are turning the social sciences upside down, he explains. Researchers are now able to understand human behaviour at the population level rather than the individual level.
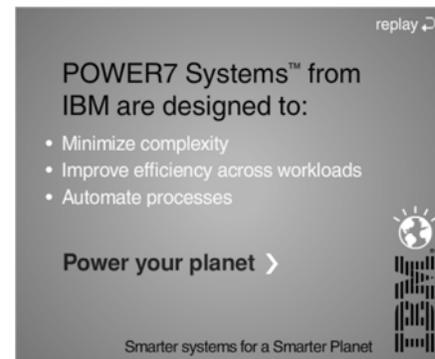
The amount of digital information increases tenfold every five years. Moore's law, which the computer industry now takes for granted, says that the processing power and storage capacity of computer chips double or their prices halve roughly every 18 months. The software programs are getting better too. Edward Felten, a computer scientist at Princeton University, reckons that the improvements in the algorithms driving computer applications have played as important a part as Moore's law for decades.

A vast amount of that information is shared. By 2013 the amount of traffic flowing over the internet annually will reach 667 exabytes, according to Cisco, a maker of communications gear. And the quantity of data continues to grow faster than the ability of the network to carry it all.

People have long groused that they were swamped by information. Back in 1917 the manager of a Connecticut manufacturing firm complained about the effects of the telephone: "Time is lost, confusion results and money is spent." Yet what is happening now goes way beyond incremental growth. The quantitative change has begun to make a qualitative difference.

This shift from information scarcity to surfeit has broad effects. "What we are seeing is the ability to have economies form around the data—and that to me is the big change at a societal and even macroeconomic level," says Craig Mundie, head of research and strategy at Microsoft. Data are becoming the new raw material of business: an economic input almost on a par with capital and labour. "Every day I wake up and ask, 'how can I flow data better, manage data better, analyse data better?" says Rollin Ford, the CIO of Wal-Mart.

Sophisticated quantitative analysis is being applied to many aspects of life, not just missile trajectories or financial hedging strategies, as in the past. For example, Farecast, a part of Mic[...] [...]dvise customers whether to buy an airline ticket now or wai[...]own by examining 225 billion flight and price records. The same

idea is being extended to hotel rooms, cars and similar items. Personal-finance websites and banks are aggregating their customer data to show up macroeconomic trends, which may develop into ancillary businesses in their own right. Number-crunchers have even uncovered match-fixing in Japanese sumo wrestling.

Dross into gold

"Data exhaust"—the trail of clicks that internet users leave behind from which value can be extracted—is becoming a mainstay of the internet economy. One example is Google's search engine, which is partly guided by the number of clicks on an item to help determine its relevance to a search query. If the eighth listing for a search term is the one most people go to, the algorithm puts it higher up.

As the world is becoming increasingly digital, aggregating and analysing data is likely to bring huge benefits in other fields as well. For example, Mr Mundie of Microsoft and Eric Schmidt, the boss of Google, sit on a presidential task force to reform American health care. "Early on in this process Eric and I both said: 'Look, if you really want to transform health care, you basically build a sort of health-care economy around the data that relate to people'," Mr Mundie explains. "You would not just think of data as the 'exhaust' of providing health services, but rather they become a central asset in trying to figure out how you would improve every aspect of health care. It's a bit of an inversion."

To be sure, digital records should make life easier for doctors, bring down costs for providers and patients and improve the quality of care. But in aggregate the data can also be mined to spot unwanted drug interactions, identify the most effective treatments and predict the onset of disease before symptoms emerge. Computers already attempt to do these things, but need to be explicitly programmed for them. In a world of big data the correlations surface almost by themselves.

Sometimes those data reveal more than was intended. For example, the city of Oakland, California, releases information on where and when arrests were made, which is put out on a private website, Oakland Crimespotting. At one point a few clicks revealed that police swept the whole of a busy street for prostitution every evening except on Wednesdays, a tactic they probably meant to keep to themselves.

But big data can have far more serious consequences than that. During the recent financial crisis it became clear that banks and rating agencies had been relying on models which, although they required a vast amount of information to be fed in, failed to reflect financial risk in the real world. This was the first crisis to be sparked by big data—and there will be more.

The way that information is managed touches all areas of life. At the turn of the 20th century new flows of information through channels such as the telegraph and telephone supported mass production. Today the availability of abundant data enables companies to cater to small niche markets anywhere in the world. Economic production used to be based in the factory, where managers pored over every machine and process to make it more efficient. Now statisticians mine the information output of the business for new ideas.

"The data-centred economy is just nascent," admits Mr Mundie of Microsoft. "You can see the outlines of it, but the technical, infrastructural and even business-model implications are not well understood right now." This special report will point to where it is beginning to surface.

Readers' comments

The Economist welcomes your views.

View all comments (1)   Add your comment

Next article »

Want more? Subscribe to *The Economist* and get the week's most relevant news and analysis.

Classified ads