The Economist

SAP

Data, data everywhere

A special report on managing information

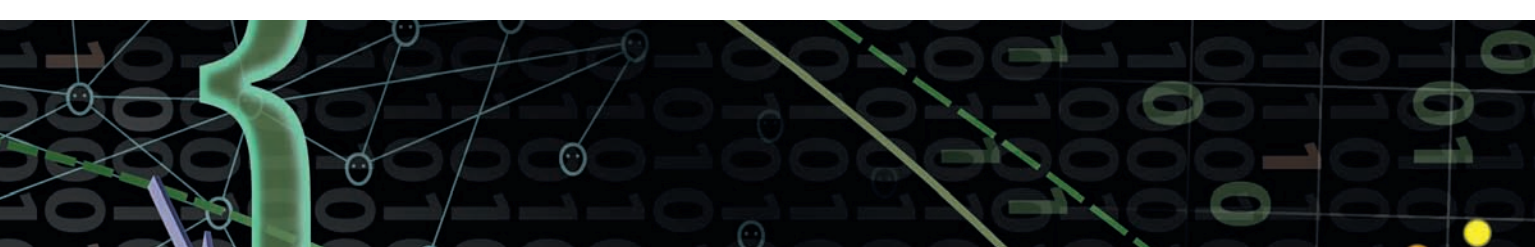
FIND KEY INSIGHTS IN JUST A FEW KEYWORDS

They're in your company. They're in your data. Hidden insights that can support your critical business decisions. Introducing **SAP® BusinessObjects™ Explorer**, an intuitive, information-discovery tool that lets you search vast amounts of data in seconds. Discover the answers you need to know, as well as the questions you didn't know to ask. It's how business gets done in a clear new world. **Visit sap.com/insights or call 866-255-0692.**

THE BEST-RUN BUSINESSES RUN SAP™

SAP®

Contents



The Economist

Reprinted from The Economist February 27th 2010

Technology

The data deluge

Businesses, governments and society are only starting to tap its vast potential

EIGHTEEN months ago, Li & Fung, a firm that manages supply chains for retailers, saw 100 gigabytes of information flow through its network each day. Now the amount has increased tenfold. During 2009, American drone aircraft flying over Iraq and Afghanistan sent back around 24 years' worth of video footage. New models being deployed this year will produce ten times as many data streams as their predecessors, and those in 2011 will produce 30 times as many.

Everywhere you look, the quantity of information in the world is soaring. According to one estimate, mankind created 150 exabytes (billion gigabytes) of data in 2005. This year, it will create 1,200 exabytes. Merely keeping up with this flood, and storing the bits that might be useful, is difficult enough. Analysing it, to spot patterns and extract useful information, is harder still. Even so, the data deluge is already starting to transform business, government, science and everyday life. It has great potential for good—as long as consumers, companies and governments make the right choices about when to restrict the flow of data, and when to encourage it.

Plucking the diamond from the waste

A few industries have led the way in their ability to gather and exploit data. Credit-card companies monitor every purchase and can identify fraudulent ones with a high degree of accuracy, using rules derived by crunching through billions of transactions. Stolen credit cards are more likely to be used to buy hard liquor than wine, for example, because it is easier to fence. Insurance firms are also good at combining clues to spot suspicious claims: fraudulent claims are more likely to be made on a Monday than a Tuesday, since policyholders who stage accidents tend to assemble friends as false witnesses over the weekend. By combining many such rules, it is possible to work out which cards are likeliest to have been stolen, and which claims are dodgy.

Mobile-phone operators, meanwhile, analyse subscribers' calling patterns to determine, for example, whether most of their frequent contacts are on a rival network. If that rival network is offering an attractive promotion that might cause the subscriber to defect, he or she can then be offered an incentive to stay. Older industries crunch data with just as much enthusiasm as new ones these days. Retailers, offline as well as online, are masters of data mining (or "business intelligence", as it is now known). By analysing "basket data", supermarkets can tailor promotions to particular customers' preferences. The oil industry uses supercomputers to trawl seismic data before drilling wells. And astronomers are just as likely to point a software query-tool at a digital sky survey as to point a telescope at the stars.

There's much further to go. Despite years of effort, law-enforcement and intelligence agencies' databases are not, by and large, linked. In health care, the digitisation of records would make it much easier to spot and monitor health trends and evaluate the effectiveness of different treatments. But large-scale efforts to computerise health records tend to run into bureaucratic, technical and ethical problems. Online advertising is already far more accurately targeted than the offline sort, but there is scope for even greater personalisation. Advertisers would then be willing to pay more, which would in turn mean that consumers prepared to opt into such things could be offered a richer and broader range of free online

services. And governments are belatedly coming around to the idea of putting more information—such as crime figures, maps, details of government contracts or statistics about the performance of public services—into the public domain. People can then reuse this information in novel ways to build businesses and hold elected officials to account. Companies that grasp these new opportunities, or provide the tools for others to do so, will prosper. Business intelligence is one of the fastest-growing parts of the software industry.

Now for the bad news

But the data deluge also poses risks. Examples abound of databases being stolen: disks full of social-security data go missing, laptops loaded with tax records are left in taxis, credit-card numbers are stolen from online retailers. The result is privacy breaches, identity theft and fraud. Privacy infringements are also possible even without such foul play: witness the periodic fusses when Facebook or Google unexpectedly change the privacy settings on their online social networks, causing members to reveal personal information unwittingly. A more sinister threat comes from Big Brotherishness of various kinds, particularly when governments compel companies to hand over personal information about their customers. Rather than owning and controlling their own personal data, people very often find that they have lost control of it.

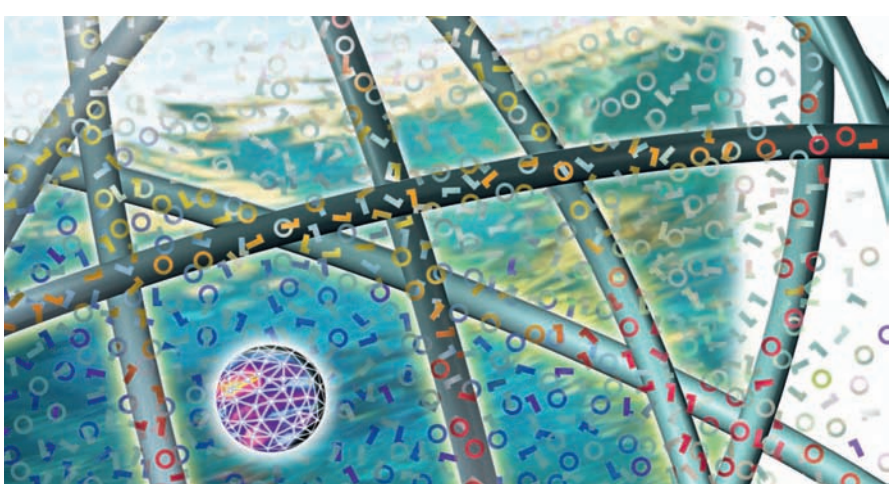
The best way to deal with these drawbacks of the data deluge is, paradoxically, to make more data available in the right way, by requiring greater transparency in several areas. First, users should be given greater access to and control over the information held about them, including whom it is shared with. Google allows users to see what information it holds about them, and lets them delete their search histories or modify the targeting of advertising, for example. Second, organisations should be required to disclose details of security breaches, as is already the case in some parts of the world, to encourage bosses to take information security more seriously. Third, organisations should be subject to an annual security audit, with the resulting grade made public (though details of any problems exposed would not be). This would encourage companies to keep their security measures up to date.

Market incentives will then come into play as organisations that manage data well are favoured over those that do not. Greater transparency in these three areas would improve security and give people more control over their data without the need for intricate regulation that could stifle innovation. After all, the process of learning to cope with the data deluge, and working out how best to tap it, has only just begun. ■

A special report on managing information

Data, data everywhere

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



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.

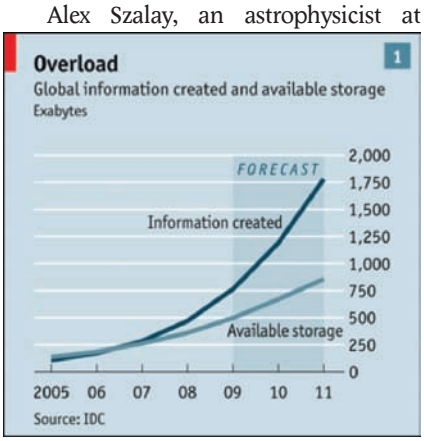
Such astronomical amounts of information can be found closer to Earth too. Wal-Mart, a retail giant, handles more than 1m customer transactions every hour, feeding databases estimated at more than 2.5 petabytes—the equivalent of 167 times the books in America's Library of Congress (see article for an 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 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 Microsoft's search engine Bing, can advise customers whether to buy an airline ticket now or wait for the price to come down 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. ■

A special report on managing information

All too much

Monstrous amounts of data

QUANTIFYING the amount of information that exists in the world is hard. What is clear is that there is an awful lot of it, and it is growing at a terrific rate (a compound annual 60%) that is speeding up all the time. The flood of data from sensors, computers, research labs, cameras, phones and the like surpassed the capacity of storage technologies in 2007. Experiments at the Large Hadron Collider at CERN, Europe's particle-physics laboratory near Geneva, generate 40 terabytes every second—orders of magnitude more than can be stored or analysed. So scientists collect what they can and let the rest dissipate into the ether.

According to a 2008 study by International Data Corp (IDC), a market-research firm, around 1,200 exabytes of digital data will be generated this year. Other studies measure slightly different things. Hal Varian and the late Peter Lyman of the University of California in Berkeley, who pioneered the idea of counting the world's bits, came up with a far smaller amount, around 5 exabytes in 2002, because they counted only the stock of original content.

What about the information that is actually consumed? Researchers at the University of California in San Diego (UCSD) examined the flow of data to American households. They found that in 2008 such households were bombarded with 3.6 zettabytes of information (or 34 gigabytes per person per day). The biggest data hogs were video games and television. In terms of bytes, written words are insignificant, amounting to less than 0.1% of the total. However, the amount of reading people do, previously in decline because of television, has almost tripled since 1980, thanks to all that text on the internet. In the past information

Data inflation 2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

Source: *The Economist*

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

consumption was largely passive, leaving aside the telephone. Today half of all bytes are received interactively, according to the UCSD. Future studies will extend beyond American households to quantify consumption globally and include business use as well.

March of the machines

Significantly, "information created by machines and used by other machines will probably grow faster than anything else," explains Roger Bohn of the UCSD, one of the authors of the study on American households. "This is primarily 'database to database' information—people are only tangentially involved in most of it."

Only 5% of the information that is created is "structured", meaning it comes in a standard format of words or numbers that can be read by computers. The rest are things like photos and phone calls which are less easily retrievable and usable. But this is changing as content on the web is increasingly "tagged", and facial-recognition and voice-recognition software can identify people and words in digital files.

"It is a very sad thing that nowadays there is so little useless information," quipped Oscar Wilde in 1894. He did not know the half of it. ■

A special report on managing information

A different game

Information is transforming traditional businesses

IN 1879 James Ritty, a saloon-keeper in Dayton, Ohio, received a patent for a wooden contraption that he dubbed the "incorruptible cashier". With a set of buttons and a loud bell, the device, sold by National Cash Register (NCR), was little more than a simple adding machine. Yet as an early form of managing information flows in American business the cash register had a huge impact. It not only reduced pilferage by alerting the shopkeeper when the till was opened; by recording every transaction, it also provided an instant overview of what was happening in the business.

Sales data remain one of a company's most important assets. In 2004 Wal-Mart peered into its mammoth databases and noticed that before a hurricane struck, there was a run on flashlights and batteries, as might be expected; but also on Pop-Tarts, a sugary American breakfast snack. On reflection it is clear that the snack would be a handy thing to eat in a blackout, but the retailer would not have thought to stock up on it before a storm. The company whose system crunched Wal-Mart's numbers was none other than NCR and its data-warehousing unit, Teradata, now an independent firm.

A few years ago such technologies, called "business intelligence", were available only to the world's biggest companies. But as the price of computing and storage has fallen and the software systems have got better and cheaper, the technology has moved into the mainstream. Companies are collecting more data than ever before. In the past they were kept in different systems that were unable to talk to each other, such as finance, human resources or customer management. Now the systems are being linked, and companies are using data-mining techniques to get a complete picture of their operations—"a single version of the truth", as the industry likes to call it. That allows firms to operate more efficiently, pick out trends and improve their forecasting.

Consider Cablecom, a Swiss telecoms operator. It has reduced customer defections from one-fifth of subscribers a year to under 5% by crunching its numbers. Its software spotted that although customer defections peaked in the 13th month, the decision to leave was made much earlier, around the ninth month (as indicated by things like the number of calls to customer support services). So Cablecom offered certain customers special deals seven

months into their subscription and reaped the rewards.

Agony and torture

Such data-mining has a dubious reputation. "Torture the data long enough and they will confess to anything," statisticians quip. But it has become far more effective as more companies have started to use the technology. Best Buy, a retailer, found that 7% of its customers accounted for 43% of its sales, so it reorganised its stores to concentrate on those customers' needs. Airline yield management improved because analytical techniques uncovered the best predictor that a passenger would actually catch a flight he had booked: that he had ordered a vegetarian meal.

The IT industry is piling into business intelligence, seeing it as a natural successor of services such as accountancy and computing in the first and second half of the 20th century respectively. Accenture, PricewaterhouseCoopers, IBM and SAP are investing heavily in their consulting practices. Technology vendors such as Oracle, Informatica, TIBCO, SAS and EMC have benefited. IBM believes business intelligence will be a pillar of its growth as sensors are used to manage things from a city's traffic flow to a patient's blood flow. It has invested $12 billion in the past four years and is opening six analytics centres with 4,000 employees worldwide.

Analytics—performing statistical operations for forecasting or uncovering correlations such as between Pop-Tarts and hurricanes—can have a big pay-off. In Britain the Royal Shakespeare Company (RSC) sifted through seven years of sales data for a marketing campaign that increased regular visitors by 70%. By examining more than 2m transaction records, the RSC discovered a lot more about its best customers: not just income, but things like occupation and family status, which allowed it to target its marketing more precisely. That was of crucial importance, says the RSC's Mary Butlin, because it substantially boosted membership as well as fund-raising revenue.

Yet making the most of data is not easy. The first step is to improve the accuracy of the information. Nestlé, for example, sells more than 100,000 products in 200 countries, using 550,000 suppliers, but it was not using its huge buying power effectively because its databases were a mess. On examination, it found that of its 9m records of vendors, customers and materials around half were obsolete or duplicated, and of the remainder about one-third were inaccurate or incomplete. The name of a vendor might be abbreviated in one record but spelled out in another, leading to double-counting.

Plainer vanilla

Over the past ten years Nestlé has been overhauling its IT system, using SAP software, and improving the quality of its data. This enabled the firm to become more efficient, says Chris Johnson, who led the initiative. For just one ingredient, vanilla, its American operation was able to reduce the number of specifications and use fewer suppliers, saving $30m a year. Overall, such operational improvements save more than $1 billion annually.

Nestlé is not alone in having problems with its database. Most CIOs admit that their data are of poor quality. In a study by IBM half the managers quizzed did not trust the information on which they had to base decisions. Many say that the technology meant to make sense of it often just produces more data. Instead of finding a needle in the haystack, they are making more hay.

Still, as analytical techniques become more widespread, business decisions will increasingly be made, or at least corroborated, on the basis of computer algorithms rather than individual hunches. This creates a need for managers who are comfortable with data, but statistics courses in business schools are not popular.

Many new business insights come from "dead data": stored information about past transactions that are examined to reveal hidden correlations. But now companies are increasingly moving to analysing real-time information flows.

Wal-Mart is a good example. The retailer operates 8,400 stores worldwide, has more than 2m employees and handles over 200m customer transactions each week. Its revenue last year, around $400 billion, is more than the GDP of many entire countries. The sheer scale of the data is a challenge, admits Rollin Ford, the CIO at Wal-Mart's headquarters in Bentonville, Arkansas. "We keep a healthy paranoia."

Not a sparrow falls

Wal-Mart's inventory-management system, called Retail Link, enables suppliers to see the exact number of their products on every shelf of every store at that precise moment. The system shows the rate of sales by the hour, by the day, over the past year and more. Begun in the 1990s, Retail Link gives suppliers a complete overview of when and how their products are selling, and with what other products in the shopping cart. This lets suppliers manage their stocks better.

The technology enabled Wal-Mart to change the business model of retailing. In some cases it leaves stock management in the hands of its suppliers and does not take ownership of the products until the moment they are sold. This allows it to shed inventory risk and reduce its costs. In essence, the shelves in its shops are a highly efficiently managed depot.

Another company that capitalises on real-time information flows is Li & Fung, one of the world's biggest supply-chain operators. Founded in Guangzhou in southern China a century ago, it does not own any factories or equipment but

orchestrates a network of 12,000 suppliers in 40 countries, sourcing goods for brands ranging from Kate Spade to Walt Disney. Its turnover in 2008 was $14 billion.

Li & Fung used to deal with its clients mostly by phone and fax, with e-mail counting as high technology. But thanks to a new web-services platform, its processes have speeded up. Orders flow through a web portal and bids can be solicited from pre-qualified suppliers. Agents now audit factories in real time with hand-held computers. Clients are able to monitor the details of every stage of an order, from the initial production run to shipping.

One of the most important technologies has turned out to be videoconferencing. It allows buyers and manufacturers to examine the colour of a material or the stitching on a garment. "Before, we weren't able to send a 500MB image—we'd post a DVD. Now we can stream it to show vendors in our offices. With real-time images we can make changes quicker," says Manuel Fernandez, Li & Fung's chief technology officer. Data flowing through its network soared from 100 gigabytes a day only 18 months ago to 1 terabyte.

The information system also allows Li & Fung to look across its operations to identify trends. In southern China, for instance, a shortage of workers and new legislation raised labour costs, so production moved north. "We saw that before it actually happened," says Mr Fernandez. The company also got advance warning of the economic crisis, and later the recovery, from retailers' orders before these trends became apparent. Investment analysts use country information provided by Li & Fung to gain insights into macroeconomic patterns.

Now that they are able to process information flows in real time, organisations are collecting more data than ever. One use for such information is to forecast when machines will break down. This hardly ever happens out of the blue: there are usually warning signs such as noise, vibration or heat. Capturing such data enables firms to act before a breakdown.

Similarly, the use of "predictive analytics" on the basis of large data sets may transform health care. Dr Carolyn McGregor of the University of Ontario, working with IBM, conducts research to spot potentially fatal infections in premature babies. The system monitors subtle changes in seven streams of real-time data, such as respiration, heart rate and blood pressure. The electrocardiogram alone generates 1,000 readings per second.

This kind of information is turned out by all medical equipment, but it used to be recorded on paper and examined perhaps once an hour. By feeding the data into a computer, Dr McGregor has been able to detect the onset of an infection before obvious symptoms emerge. "You can't see it with the naked eye, but a computer can," she says.

Open sesame

Two technology trends are helping to fuel these new uses of data: cloud computing and open-source software. Cloud computing—in which the internet is used as a platform to collect, store and process data—allows businesses to lease computing power as and when they need it, rather than having to buy expensive equipment. Amazon, Google and Microsoft are the most prominent firms to make their massive computing infrastructure available to clients. As more corporate functions, such as human resources or sales, are managed over a network, companies can see patterns across the whole of the business and share their information more easily.

A free programming language called R lets companies examine and present big data sets, and free software called Hadoop now allows ordinary PCs to analyse huge quantities of data that previously required a supercomputer. It does this by parcelling out the tasks to numerous computers at once. This saves time and money. For example, the New York Times a few years ago used cloud computing and Hadoop to convert over 400,000 scanned images from its archives, from 1851 to 1922. By harnessing the power of hundreds of computers, it was able to do the job in 36 hours.

Visa, a credit-card company, in a recent trial with Hadoop crunched two years of test records, or 73 billion transactions, amounting to 36 terabytes of data. The processing time fell from one month with traditional methods to a mere 13 minutes. It is a striking successor of Ritty's incorruptible cashier for a data-driven age. ■

A special report on managing information

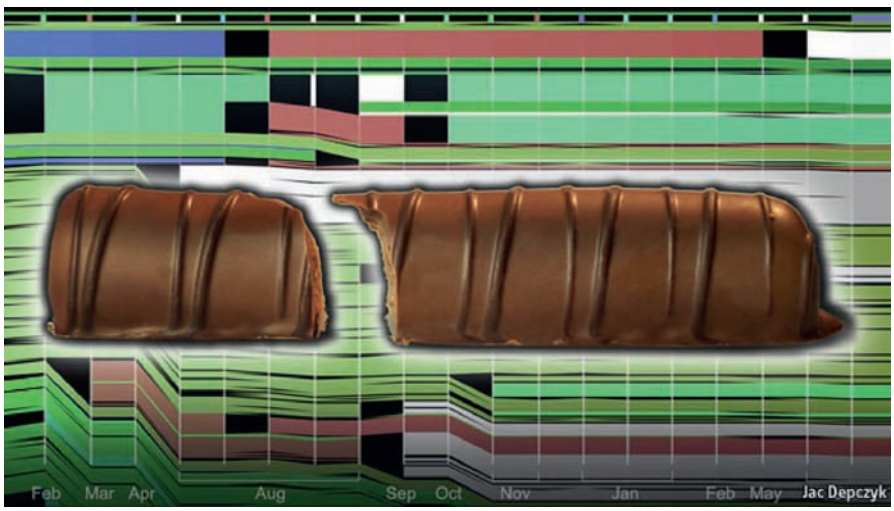
Show me

New ways of visualising data

IN 1998 Martin Wattenberg, then a graphic designer at the magazine SmartMoney in New York, had a problem. He wanted to depict the daily movements in the stockmarket, but the customary way, as a line showing the performance of an index over time, provided only a very broad overall picture. Every day hundreds of individual companies may rise or fall by a little or a lot. The same is true for whole sectors. Being able to see all this information at once could be useful to investors. But how to make it visually accessible?

Mr Wattenberg's brilliant idea was to adapt an existing technique to create a "Map of the Market" in the form of a grid. It used the day's closing share price to show more than 500 companies arranged by sector. Shades of green or red indicated whether a share had risen or fallen and by how much, showing the activity in every sector of the market. It was an instant hit—and brought the nascent field of data visualisation to a mainstream audience.

In recent years there have been



big advances in displaying massive amounts of data to make them easily accessible. This is emerging as a vibrant and creative field melding the skills of computer science, statistics, artistic design and storytelling.

"Every field has some central tension it is trying to resolve. Visualisation deals with the inhuman scale of the information and the need to present it at the very human scale of what the eye can see," says Mr Wattenberg, who has since moved to IBM and now spearheads a new generation of data-visualisation specialists.

Market information may be hard to display, but at least the data are numerical. Words are even more difficult. One way of depicting them is to count them and present them in clusters, with more common ones shown in a proportionately larger font. Called a "word cloud", this method is popular across the web. It gives a rough indication of what a body of text is about.

Soon after President Obama's inauguration a word cloud with a graphical-semiotic representation of his 21-minute speech appeared on the web. The three most common words were nation, America and people. His predecessor's had been freedom, America and liberty. Abraham Lincoln had majored on war, God and offence. The technique has a utility beyond identifying themes. Social-networking sites let users "tag" pages and images with words describing the content. The terms displayed in a "tag cloud" are links that will bring up a list of the related content.

Another way to present text, devised by Mr Wattenberg and a colleague at IBM, Fernanda Viégas, is a chart of edits made on Wikipedia. The online encyclopedia is written entirely by volunteers. The software creates a permanent record of every edit to show exactly who changed what, and when. That amounts to a lot of data over time.

One way to map the process is to assign different colours to different users and show how much of their contribution remains by the thickness of the line that represents it. The entry for "chocolate", for instance, looks smooth until a series of ragged zigzags reveals an item of text being repeatedly removed and restored as an arcane debate rages. Another visualisation looks at changes to Wikipedia entries by software designed to improve the way articles are categorised, showing the modifications as a sea of colour.

Is it art? Is it information? Some data-visual works have been exhibited in places such as the Whitney and the Museum of Modern Art in New York. Others have been turned into books, such as the web project "We Feel Fine" by Jonathan Harris and Sep Kamvar, which captures every instance of the words "feel" or "feeling" on Twitter, a social-networking site, and matches it to time, location, age, sex and even the weather.

For the purposes of data visualisation as many things as possible are reduced to raw data that can be presented visually, sometimes in unexpected ways. For instance, a representation of the sources cited in the journal Nature gives each source publication a line and identifies different scientific fields in different colours. This makes it easy to see that biology sources are most heavily cited, which is unsurprising. But it also shows, more unexpectedly, that the publications most heavily cited include the Physical Review Letters and Astrophysical Journal.

The art of the visible

Resembling a splendid orchid, the Nature chart can be criticised for being more picturesque than informative; but whether it is more art or more information, it offers a new way to look at the world at a time when almost everything generates huge swathes of data that are hard to understand. If a picture is worth a thousand words, an infographic is worth an awful lot of data points.

Visualisation is a relatively new discipline. The time series, the most common form of chart, did not start to appear in scientific writings until the late 18th century, notes Edward Tufte in his classic "The Visual Display of Quantitative Information", the bible of the business. Today's infographics experts are pioneering a new medium that presents meaty information in a compelling narrative: "Something in-between the textbook and the novel", writes Nathan Yau of UCLA in a recent book, "Beautiful Data".

It's only natural

The brain finds it easier to process information if it is presented as an image rather than as words or numbers. The right hemisphere recognises shapes and colours. The left side of the brain processes information in an analytical and sequential way and is more active when people read text or look at a spreadsheet. Looking through a numerical table takes a lot of mental effort, but information presented visually can be grasped in a few seconds. The brain identifies patterns, proportions and relationships to make instant subliminal comparisons. Businesses care about such things. Farecast, the online price-prediction service, hired applied psychologists to design the site's

charts and colour schemes.

These graphics are often based on immense quantities of data. Jeffrey Heer of Stanford University helped develop sense.us, a website that gives people access to American census data going back more than a century. Ben Fry, an independent designer, created a map of the 26m roads in the continental United States. The dense communities of the north-east form a powerful contrast to the desolate far west. Aaron Koblin of Google plotted a map of every commercial flight in America over 24 hours, with brighter lines identifying routes with heavier traffic.

Such techniques are moving into the business world. Mr Fry designed interactive charts for Ge's health-care division that show the costs borne by patients and insurers, respectively, for common diseases throughout people's lives. Among media companies the New York Times and the Guardian in Britain have been the most ambitious, producing data-rich, interactive graphics that are strong enough to stand on their own.

The tools are becoming more accessible. For example, Tableau Software, co-founded in 2003 by Pat Hanrahan of Stanford University, does for visualising data what word-processing did for text, allowing anyone to manipulate information creatively. Tableau offers both free and paid-for products, as does a website called Swivel.com. Some sites are entirely free. Google and an IBM website called Many Eyes let people upload their data to display in novel ways and share with others.

Some data sets are best represented as a moving image. As print publications move to e-readers, animated infographics will eventually become standard. The software Gapminder elegantly displays four dynamic variables at once.

Displaying information can make a difference by enabling people to understand complex matters and find creative solutions. Valdis Krebs, a specialist in mapping social interactions, recalls being called in to help with a corporate project that was vastly over budget and behind schedule. He drew up an intricate network map of e-mail traffic that showed distinct clusters, revealing that the teams involved were not talking directly to each other but passing messages via managers. So the company changed its office layout and its work processes—and the project quickly got back on track. ■

A special report on managing information

Needle in a haystack

The uses of information about information

AS DATA become more abundant, the main problem is no longer finding the information as such but laying one's hands on the relevant bits easily and quickly. What is needed is information about information. Librarians and computer scientists call it metadata.

Information management has a long history. In Assyria around three millennia ago clay tablets had small clay labels attached to them to make them easier to tell apart when they were filed in baskets or on shelves. The idea survived into the 20th century in the shape of the little catalogue cards librarians used to note down a book's title, author, subject and so on before the records were moved onto computers. The actual books constituted the data, the catalogue cards the metadata. Other examples include package labels to the 5 billion bar codes that are scanned throughout the world every day.

These days metadata are undergoing a virtual renaissance. In order to be useful, the cornucopia of information provided by the internet has to be organised. That is what Google does so well. The raw material for its search engines comes free: web pages on the public internet. Where it adds value (and creates metadata) is by structuring the information, ranking it in order of its relevance to the query.

Google handles around half the world's internet searches, answering around 35,000 queries every second. Metadata are a potentially lucrative business. "If you can control the pathways and means of finding information, you can extract rents from subsequent levels of producers," explains Eli Noam, a telecoms economist at New York's Columbia Business School. But there are more benign uses too. For example, photos uploaded to the website Flickr contain metadata such as when and often where they were snapped, as well as the camera model—useful for would-be buyers.

Internet users help to label unstructured information so it can be easily found, tagging photos and videos. But they disdain conventional library classifications. Instead, they pick any word they fancy, creating an eclectic "folksonomy". So instead of labelling a photograph of Barack Obama as "president", they might call it "sexy" or "SOB". That sounds chaotic, but needn't be.

When information was recorded on a tangible medium—paper, film and so on—everything had only one correct place. With digital information the same item can be filed in several places at once, notes David Weinberger, the author of a book about taxonomy and the internet, "Everything Is Miscellaneous". Digital metadata make things more complicated and simpler at the same time. ■