
The Economist
Data, data everywhere
A special report on managing information

Special Report | Data, data everywhere

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

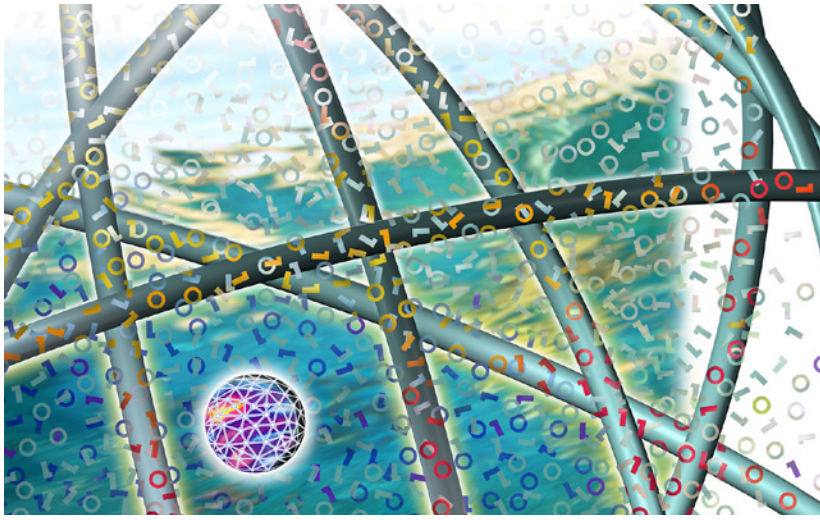
But they are also creating a host of new problems. Despite the abundance of tools to capture, process and share all this information—sensors, computers, mobile phones and the like—it already exceeds the available storage space (see chart 1). Moreover, ensuring data security and protecting privacy is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

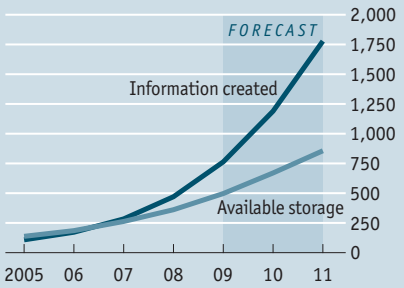
Moreover, there are now many more people who interact with information. Between 1990 and 2005 more than 1 billion people worldwide entered the middle class. As they get richer they become more literate, which fuels information growth, notes Mr Cortada. The results are showing up in politics, economics 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.

Overload 1

Global information created and available storage

Exabytes



Source: IDC

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. ■

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, amount-

ing 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 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. ■

Data inflation 2

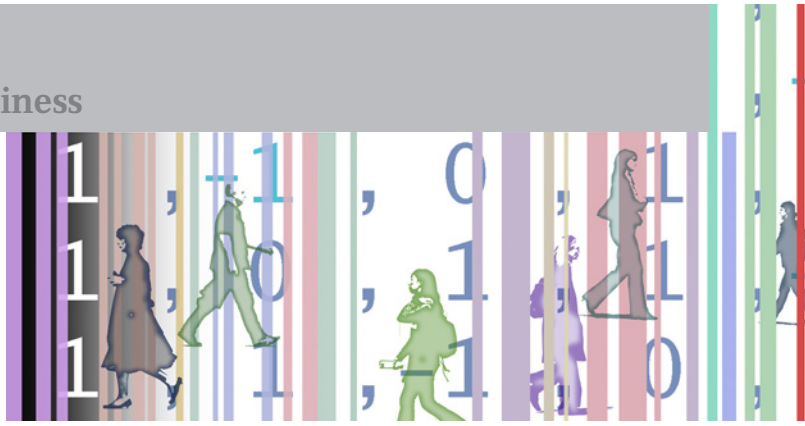
Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "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.

A different game

Information is transforming traditional business



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

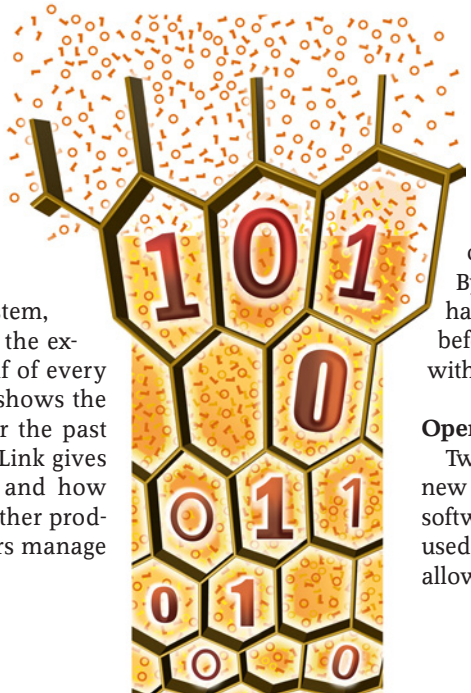
Over the past ten years Nestlé has been overhauling its IT system, using SAP software, and improving the quality of its data. This enabled the firm to become more efficient, says Chris Johnson, who led the initiative. For just one ingredient, vanilla, its American operation was able to reduce the number 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. ■

Clicking for gold

How internet companies profit from data on the web

PSST! Amazon.com does not want you to know what it knows about you. It not only tracks the books you purchase, but also keeps a record of the ones you browse but do not buy to help it recommend other books to you. Information from its e-book, the Kindle, is probably even richer: how long a user spends reading each page, whether he takes notes and so on. But Amazon refuses to disclose what data it collects or how it uses them.

It is not alone. Across the internet economy, companies are compiling masses of data on people, their activities, their likes and dislikes, their relationships with others and even where they are at any particular moment—and keeping mum. For example, Facebook, a social-networking site, tracks the activities of its 400m users, half of whom spend an average of almost an hour on the site every day, but does not talk about what it finds. Google reveals a little but holds back a lot. Even eBay, the online auctioneer, keeps quiet.

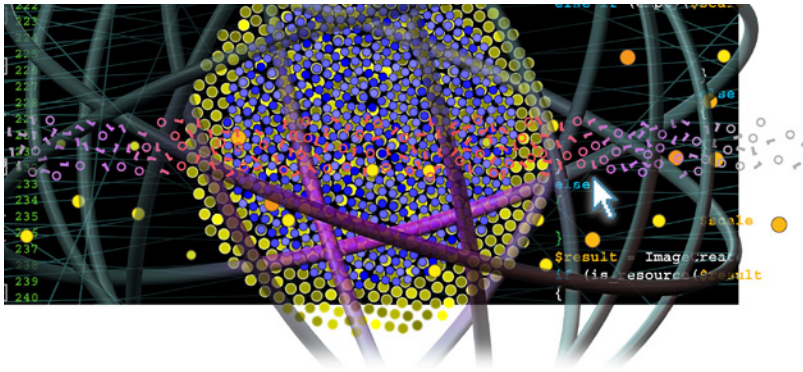
"They are uncomfortable bringing so much attention to this because it is at the heart of their competitive advantage," says Tim O'Reilly, a technology insider and publisher. "Data are the coin of the realm. They have a big lead over other companies that do not 'get' this." As the communications director of one of the web's biggest sites admits, "we're not in a position to have an in-depth conversation. It has less to do with sensitive considerations like privacy. Instead, we're just not ready to tip our hand." In other words, the firm does not want to reveal valuable trade secrets.

The reticence partly reflects fears about consumer unease and unwelcome attention from regulators. But this is short-sighted, for two reasons. First, politicians and the public are already anxious. The chairman of America's Federal Trade Commission, Jon Leibowitz, has publicly grumbled that the industry has not been sufficiently forthcoming. Second, if users knew how the data were used, they would probably be more impressed than alarmed.

Where traditional businesses generally collect information about customers from their purchases or from surveys, internet companies have the luxury of being able to gather data from everything that happens on their sites. The biggest websites have long recognised that information itself is their biggest treasure. And it can immediately be put to use in a way that traditional firms cannot match.

Some of the techniques have become widespread. Before deploying a new feature, big sites run controlled experiments to see what works best. Amazon and Netflix, a site that offers films for hire, use a statistical technique called collaborative filtering to make recommendations to users based on what other users like. The technique they came up with has produced millions of dollars of additional sales. Nearly two-thirds of the film selections by Netflix's customer come from the referrals made by computer.

EBay, which at first sight looks like nothing more than a neutral platform for commercial exchanges, makes myriad adjustments based on information culled from listing activity, bidding behaviour, pricing trends, search terms and the length of time users look at a page. Every



product category is treated as a micro-economy that is actively managed. Lots of searches but few sales for an expensive item may signal unmet demand, so eBay will find a partner to offer sellers insurance to increase listings.

The company that gets the most out of its data is Google. Creating new economic value from unthinkably large amounts of information is its lifeblood. That helps explain why, on inspection, the market capitalisation of the 11-year-old firm, of around $170 billion, is not so outlandish. Google exploits information that is a by-product of user interactions, or data exhaust, which is automatically recycled to improve the service or create an entirely new product.

Vote with your mouse

Until 1998, when Larry Page, one of Google's founders, devised the PageRank algorithm for search, search engines counted the number of times that a word appeared on a web page to determine its relevance—a system wide open to manipulation. Google's innovation was to count the number of inbound links from other web pages. Such links act as "votes" on what internet users at large believe to be good content. More links suggest a webpage is more useful, just as more citations of a book suggests it is better.

But although Google's system was an improvement, it too was open to abuse from "link spam", created only to dupe the system. The firm's engineers realised that the solution was staring them in the face: the search results on which users actually clicked and stayed. A Google search might yield 2m pages of results in a quarter of a second, but users often want just one page, and by choosing it they "tell" Google what they are looking for. So the algorithm was rejigged to feed that information back into the service automatically.

From then on Google realised it was in the data-mining business. To put the model in simple economic terms, its search results give away, say, $1 in value, and in return (thanks to the user's clicks) it gets 1 cent back. When the next user visits, he gets $1.01 of value, and so on. As one employee puts it: "We like learning from large, 'noisy' data sets."

Making improvements on the back of a big data set is not a Google monopoly, nor is the technique new. One of the most striking examples dates from the mid-1800s, when Matthew Fontaine Maury of the American navy had the idea of aggregating nautical logs from ships crossing the Pacific to find the routes that offered the best winds and currents. He created an early variant of a "viral" social network, rewarding captains who submitted their logbooks with a copy of his maps. But the process was slow and laborious.

Wizard spelling

Google applies this principle of recursively learning from the data to many of its services, including the humble spell-check, for which it used a pioneering method that produced perhaps the world's best spell-checker in almost every language. Microsoft says it spent several million dollars over 20 years to develop a robust spell-checker for its word-processing program. But Google got its raw material free: its program is based on all the misspellings that users type into a search window and then "correct" by clicking on the right result. With almost 3 billion queries a day, those results soon mount up. Other search engines in the 1990s had the chance to do the same, but did not pursue it. Around 2000 Yahoo! saw the potential, but nothing came of the idea. It was Google that recognised the gold dust in the detritus of its interactions with its users and took the trouble to collect it up.

Two newer Google services take the same approach: translation and voice recognition. Both have been big stumbling blocks for computer scientists working on artificial intelligence. For over four decades the boffins tried to program computers to "understand" the structure and phonetics of language. This meant defining rules such as where nouns and verbs go in a sentence, which are the correct tenses and so on. All the exceptions to the rules needed to be programmed in too. Google, by contrast, saw it as a big maths problem that could be solved with a lot of data and processing power—and came up with something very useful.

For translation, the company was able to draw on its other services. Its search system had copies of European Commission documents, which are translated into around 20 languages. Its book-scanning project has thousands of titles that have been translated into many languages. All these translations are very good, done by experts to exacting standards. So instead of trying to teach its computers the rules of a language, Google turned them loose on the texts to make statistical inferences. Google Translate now covers more than 50 languages, according to Franz Och, one of the company's engineers. The system identifies which word or phrase in one language is the most likely equivalent in a second language. If direct translations are not available (say, Hindi to Catalan), then English is used as a bridge.

Google was not the first to try this method. In the early 1990s IBM tried to build a French-English program using translations from Canada's Parliament. But the system did not work well and the project was abandoned. IBM had only a few million documents at its disposal, says Mr Och dismissively. Google has billions. The system was first developed by processing almost 2 trillion words. But although it learns from a big body of data, it lacks the recursive qualities of spell-check and search.

The design of the feedback loop is critical. Google asks users for their opinions, but not much else. A translation start-up in Germany called Linguee is trying something different: it presents users with snippets of possible translations and asks them to click on the best. That provides feedback on which version is the most accurate.

Voice recognition highlights the importance of making use of data exhaust. To use Google's telephone directory or audio car navigation service, customers dial the relevant number and say what they are looking for. The system repeats the information; when the customer confirms it, or repeats the query, the system develops a record of the different ways the target word can be spoken. It does not learn to understand voice; it computes probabilities.

To launch the service Google needed an existing voice-recognition system, so it licensed software from Nuance, a leader in the field. But Google itself keeps the data from voice queries, and its voice-recognition system may end up performing better than Nuance's—which is now trying to get access to lots more data by partnering with everyone in sight.

Re-using data represents a new model for how computing is done, says Edward Felten of Princeton University. "Looking at large data sets and making inferences about what goes together is advancing more rapidly than expected. 'Understanding' turns out to be overrated, and statistical analysis goes a lot of the way." Many internet companies now see things the same way. Facebook regularly examines its huge databases to boost usage. It found that the best single predictor of whether members would contribute to the site was seeing that their friends had been active on it, so it took to sending members information about what their friends had been up to online. Zynga, an online games company, tracks its 100m unique players each month to improve its games.

"If there are user-generated data to be had, then we can build much better systems than just trying to improve the algorithms," says Andreas Weigend, a former chief scientist at Amazon who is now at Stanford University. Marc Andreessen, a venture capitalist who sits on numerous boards and was one of the founders of Netscape, the web's first commercial browser, thinks that "these new companies have built a culture, and the processes and the technology to deal with large amounts of data, that traditional companies simply don't have."

Recycling data exhaust is a common theme in the myriad projects going on in Google's empire and helps explain why almost all of them are labelled as a "beta" or early test version: they truly are in continuous development. A service that lets Google users store medical records might also allow the company to spot valuable patterns about diseases and treatments. A service where users can monitor their use of electricity, device by device, provides rich information on energy consumption. It could become the world's best database of household appliances and consumer electronics—and even foresee breakdowns. The aggregated search queries, which the company makes available free, are used as remarkably accurate predictors for everything from retail sales to flu outbreaks.

Together, all this is in line with the company's audacious mission to "organise the world's information". Yet the words are carefully chosen: Google does not need to own the data. Usually all it wants is to have access to them (and see that its rivals do not). In an initiative called "Data Liberation Front" that quietly began last September, Google is planning to rejig all its services so that users can discontinue them very easily and take their data with them. In an industry built on locking in the customer, the company says it wants to reduce the "barriers to exit". That should help save its engineers from complacency, the curse of many a tech champion. The project might stall if it started to hurt the business. But perhaps Google reckons that users will be more inclined to share their information with it if they know that they can easily take it back. ■

The open society

Governments are letting in the light

FROM antiquity to modern times, the nation has always been a product of information management. The ability to impose taxes, promulgate laws, count citizens and raise an army lies at the heart of statehood. Yet something new is afoot. These days democratic openness means more than that citizens can vote at regular intervals in free and fair elections. They also expect to have access to government data.

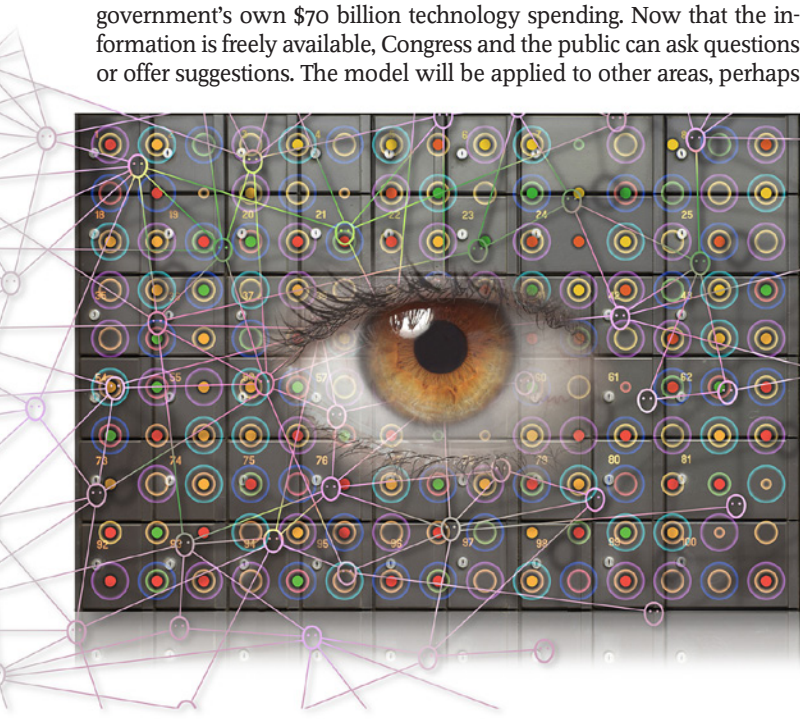
The state has long been the biggest generator, collector and user of data. It keeps records on every birth, marriage and death, compiles figures on all aspects of the economy and keeps statistics on licences, laws and the weather. Yet until recently all these data have been locked tight. Even when publicly accessible they were hard to find, and aggregating lots of printed information is notoriously difficult.

But now citizens and non-governmental organisations the world over are pressing to get access to public data at the national, state and municipal level—and sometimes government officials enthusiastically support them. "Government information is a form of infrastructure, no less important to our modern life than our roads, electrical grid or water systems," says Carl Malamud, the boss of a group called Public.Resource.Org that puts government data online. He was responsible for making the databases of America's Securities and Exchange Commission available on the web in the early 1990s.

America is in the lead on data access. On his first full day in office Barack Obama issued a presidential memorandum ordering the heads of federal agencies to make available as much information as possible, urging them to act "with a clear presumption: in the face of doubt, openness prevails". This was all the more remarkable since the Bush administration had explicitly instructed agencies to do the opposite.

Mr Obama's directive caused a flurry of activity. It is now possible to obtain figures on job-related deaths that name employers, and to get annual data on migration free. Some information that was previously available but hard to get at, such as the Federal Register, a record of government notices, now comes in a computer-readable format. It is all on a public website, data.gov. And more information is being released all the time. Within 48 hours of data on flight delays being made public, a website had sprung up to disseminate them.

Providing access to data "creates a culture of accountability", says Vivek Kundra, the federal government's CIO. One of the first things he did after taking office was to create an online "dashboard" detailing the government's own $70 billion technology spending. Now that the information is freely available, Congress and the public can ask questions or offer suggestions. The model will be applied to other areas, perhaps including health-care data, says Mr Kundra—provided that looming privacy issues can be resolved.



All this has made a big difference. "There is a cultural change in what people expect from government, fuelled by the experience of shopping on the internet and having real-time access to financial information," says John Wonderlich of the Sunlight Foundation, which promotes open government. The economic crisis has speeded up that change, particularly in state and city governments.

"The city is facing its eighth budget shortfall. We're looking at a 50% reduction in operating funds," says Chris Vein, San Francisco's CIO. "We must figure out how we change our operations." He insists that providing more information can make government more efficient. California's generous "sunshine laws" provide the necessary legal backing. Among the first users of the newly available data was a site called "San Francisco Crimespotting" by Stamen Design that layers historical crime figures on top of map information. It allows users to play around with the data and spot hidden trends. People now often come to public meetings armed with crime maps to demand police patrols in their particular area.

Anyone can play

Other cities, including New York, Chicago and Washington, DC, are racing ahead as well. Now that citizens' groups and companies have the raw data, they can use them to improve city services in ways that cash-strapped local governments cannot. For instance, cleanscores.com puts restaurants' health-inspection scores online; other sites list children's activities or help people find parking spaces. In the past government would have been pressed to provide these services; now it simply supplies the data. Mr Vein concedes, however, that "we don't know what is useful or not. This is a grand experiment."

Other parts of the world are also beginning to move to greater openness. A European Commission directive in 2005 called for making public-sector information more accessible (but it has no bite). Europe's digital activists use the web to track politicians and to try to improve public services. In Britain FixMyStreet.com gives citizens the opportunity to flag up local problems. That allows local authorities to find out about people's concerns; and once the problem has been publicly aired it becomes more difficult to ignore.

One obstacle is that most countries lack America's open-government ethos, nurtured over decades by laws on ethics in government, transparency rules and the Freedom of Information act, which acquired teeth after the Nixon years.

An obstacle of a different sort is Crown copyright, which means that most government data in Britain and the Commonwealth countries are the state's property, constraining their use. In Britain postcodes and Ordnance Survey map data at present cannot be freely used for commercial purposes—a source of loud complaints from businesses and activists. But from later this year access to some parts of both data sets will be free, thanks to an initiative to bring more government services online.

But even in America access to some government information is restricted by financial barriers. Remarkably, this applies to court documents, which in a democracy should surely be free. Legal records are public and available online from the Administrative Office of the US Courts (AOUSC), but at a costly eight cents per page. Even the federal government has to pay: between 2000 and 2008 it spent $30m to get access to its own records. Yet the AOUSC is currently paying $156m over ten years to two companies, WestLaw and LexisNexis, to publish the material online (albeit organised and searchable with the firms' tech-

nologies). Those companies, for their part, earn an estimated $2 billion annually from selling American court rulings and extra content such as case reference guides. "The law is locked up behind a cash register," says Mr Malamud.

The two firms say they welcome competition, pointing to their strong search technology and the additional services they provide, such as case summaries and useful precedents. It seems unlikely that they will keep their grip for long. One administration official privately calls freeing the information a "no-brainer". Even Google has begun to provide some legal documents online.

Change agent

The point of open information is not merely to expose the world but to change it. In recent years moves towards more transparency in government have become one of the most vibrant and promising areas of public policy. Sometimes information disclosure can achieve policy aims more effectively and at far lower cost than traditional regulation.

In an important shift, new transparency requirements are now being used by government—and by the public—to hold the private sector to account. For example, it had proved extremely difficult to persuade American businesses to cut down on the use of harmful chemicals and their release into the environment. An add-on to a 1986 law required firms simply to disclose what they release, including "by computer telecommunications". Even to supporters it seemed like a fudge, but it turned out to be a resounding success. By 2000 American businesses had reduced their emissions of the chemicals covered under the law by 40%, and over time the rules were actually tightened. Public scrutiny achieved what legislation could not.

There have been many other such successes in areas as diverse as restaurant sanitation, car safety, nutrition, home loans for minorities and educational performance, note Archon Fung, Mary Graham and David Weil of the Transparency Policy Project at Harvard's Kennedy School of Government in their book "Full Disclosure". But transparency alone is not enough. There has to be a community to champion the information. Providers need an incentive to supply the data as well as penalties for withholding them. And web developers have to find ways of ensuring that the public data being released are used effectively.

Mr Fung thinks that as governments release more and more information about the things they do, the data will be used to show the public sector's shortcomings rather than to highlight its achievements. Another concern is that the accuracy and quality of the data will be found wanting (which is a problem for business as well as for the public sector). There is also a debate over whether governments should merely supply the raw data or get involved in processing and displaying them too. The concern is that they might manipulate them—but then so might anyone else.

Public access to government figures is certain to release economic value and encourage entrepreneurship. That has already happened with weather data and with America's GPS satellite-navigation system that was opened for full commercial use a decade ago. And many firms make a good living out of searching for or repackaging patent filings.

Moreover, providing information opens up new forms of collaboration between the public and the private sectors. Beth Noveck, one of the Obama administration's recruits, who is a law professor and author of a book entitled "Wiki Government", has spearheaded an initiative called peer-to-patent that has opened up some of America's patent filings for public inspection.

John Stuart Mill in 1861 called for "the widest participation in the details of judicial and administrative business...above all by the utmost possible publicity." These days, that includes the greatest possible disclosure of data by electronic means. ■

Show me

New ways to visualising data

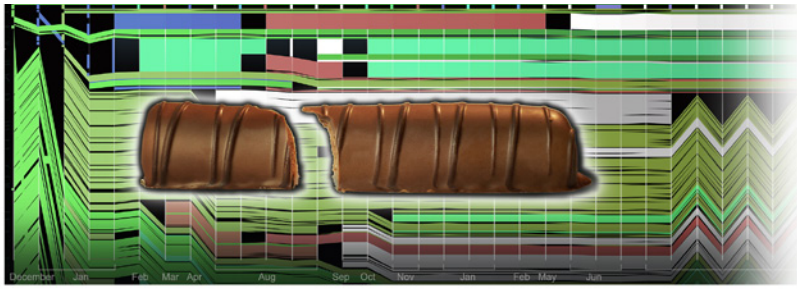
IN 1998 Martin Wattenberg, then a graphic designer at the magazine SmartMoney in New York, had a problem. He wanted to depict the daily movements in the stockmarket, but the customary way, as a line showing the performance of an index over time, provided only a very broad overall picture. Every day hundreds 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. (These and other images are available here.)

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. ■

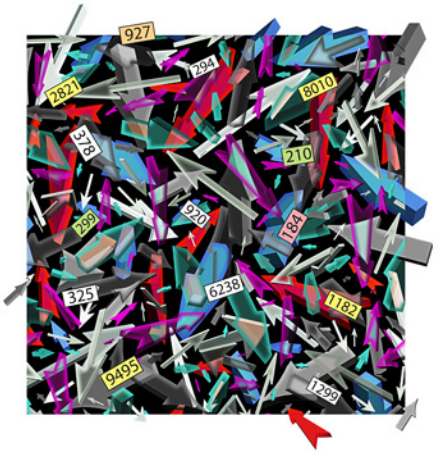
Needle in a haystack

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.

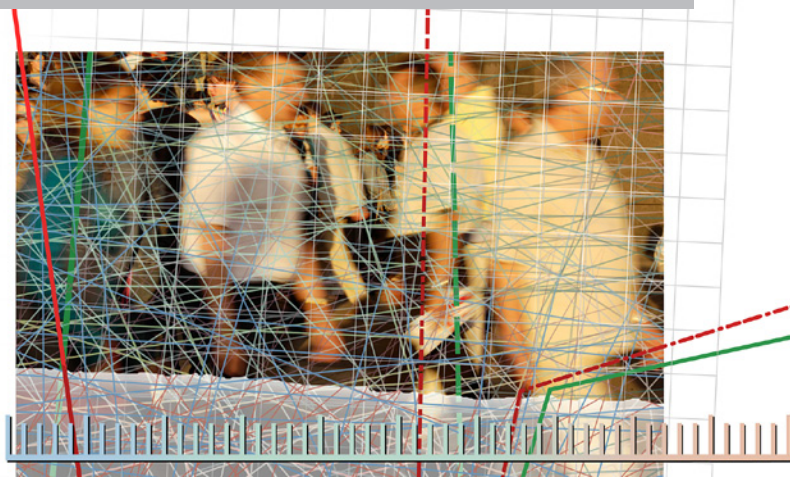
Google handles around half the world's internet searches, answering around 35,000 queries every second. Metadata are a potentially lucrative business. "If you can control the pathways and means of finding information, you can extract rents from subsequent levels of producers," explains Eli Noam, a telecoms 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. ■

New rules for big data

Regulators are having to rethink their brief



TWO centuries after Gutenberg invented movable type in the mid-1400s there were plenty of books around, but they were expensive and poorly made. In Britain a cartel had a lock on classic works such as Shakespeare's and Milton's. The first copyright law, enacted in the early 1700s in the Bard's home country, was designed to free knowledge by putting books in the public domain after a short period of exclusivity, around 14 years. Laws protecting free speech did not emerge until the late 18th century. Before print became widespread the need was limited.

Now the information flows in an era of abundant data are changing the relationship between technology and the role of the state once again. Many of today's rules look increasingly archaic. Privacy laws were not designed for networks. Rules for document retention presume paper records. And since all the information is interconnected, it needs global rules.

New principles for an age of big data sets will need to cover six broad areas: privacy, security, retention, processing, ownership and the integrity of information.

Privacy is one of the biggest worries. People are disclosing more personal information than ever. Social-networking sites and others actually depend on it. But as databases grow, information that on its own cannot be traced to a particular individual can often be unlocked with just a bit of computer effort.

This tension between individuals' interest in protecting their privacy and companies' interest in exploiting personal information could be resolved by giving people more control. They could be given the right to see and correct the information about them that an organisation holds, and to be told how it was used and with whom it was shared.

Today's privacy rules aspire to this, but fall short because of technical difficulties which the industry likes to exaggerate. Better technology should eliminate such problems. Besides, firms are already spending a great deal on collecting, sharing and processing the data; they could divert a sliver of that money to provide greater individual control.

The benefits of information security—protecting computer systems and networks—are inherently invisible: if threats have been averted, things work as normal. That means it often gets neglected. One way to deal with that is to disclose more information. A pioneering law in California in 2003 required companies to notify people if a security breach had compromised their personal information, which pushed companies to invest more in prevention. The model has been adopted in other states and could be used more widely.

In addition, regulators could require large companies to undergo an annual information-security audit by an accredited third party, similar to financial audits for listed companies. Information about vulnerabilities would be kept confidential, but it could be used by firms to improve their practices and handed to regulators if problems arose. It could even be a requirement for insurance coverage, allowing a market for information security to emerge.

Current rules on digital records state that data should never be stored for longer than necessary because they might be misused or inadvertently released. But Viktor Mayer-Schönberger of the National University of Singapore worries that the increasing power and decreasing price of computers will make it too easy to hold on to everything. In his recent book "Delete" he argues in favour of technical systems that "forget": digital files that have expiry dates or slowly degrade over time.

Yet regulation is pushing in the opposite direction. There is a social and political expectation that records will be kept, says Peter Allen of CSC, a technology provider: "The more we know, the more we are expected to know—for ever." American security officials have pressed companies to keep records because they may hold clues after a terrorist incident. In future it is more likely that companies will be required to retain all digital files, and ensure their accuracy, than to delete them.

Processing data is another concern. Ian Ayres, an economist and lawyer at Yale University and the author of "Super-Crunchers", a book about computer algorithms replacing human intuition, frets about the legal implications of using statistical correlations. Rebecca Goldin, a mathematician at George Mason University, goes further: she worries about the "ethics of super-crunching". For example, racial discrimination against an applicant for a bank loan is illegal. But what if a computer model factors in the educational level of the applicant's mother, which in America is strongly correlated with

race? And what if computers, just as they can predict an individual's susceptibility to a disease from other bits of information, can predict his predisposition to committing a crime?

A new regulatory principle in the age of big data, then, might be that people's data cannot be used to discriminate against them on the basis of something that might or might not happen. The individual must be regarded as a free agent. This idea is akin to the general rule of national statistical offices that data gathered for surveys cannot be used against a person for things like deporting illegal immigrants—which, alas, has not always been respected.

Privacy rules lean towards treating personal information as a property right. A reasonable presumption might be that the trail of data that an individual leaves behind and that can be traced to him, from clicks on search engines to book-buying preferences, belong to that individual, not the entity that collected it. Google's "data liberation" initiative mentioned earlier in this report points in that direction. That might create a market for information. Indeed, "data portability" stimulates competition, just as phone-number portability encourages competition among mobile operators. It might also reduce the need for antitrust enforcement by counteracting data aggregators' desire to grow ever bigger in order to reap economies of scale.

Ensuring the integrity of the information is an important part of the big-data age. When America's secretary of state, Hillary Clinton, lambasted the Chinese in January for allegedly hacking into Google's computers, she used the term "the global networked commons". The idea is that the internet is a shared environment, like the oceans or airspace, which requires international co-operation to make the best use of it. Censorship pollutes that environment. Disrupting information flows not only violates the integrity of the data but quashes free expression and denies the right of assembly. Likewise, if telecoms operators give preferential treatment to certain content providers, they undermine the idea of "network neutrality".

Governments could define best practice on dealing with information flows and the processing of data, just as they require firms to label processed foods with the ingredients or impose public-health standards. The World Trade Organisation, which oversees the free flow of physical trade, might be a suitable body for keeping digital goods and services flowing too. But it will not be quick or easy. ■

Handling the cornucopia

The best way to deal with all that information is to use machines. But they need watching

IN 2002 America's Defence Advanced Research Projects Agency, best known for developing the internet four decades ago, embarked on a futuristic initiative called Augmented Cognition, or "AugCog". Commander Dylan Schmorrow, a cognitive scientist with the navy, devised a crown of sensors to monitor activity in the brain such as blood flow and oxygen levels. The idea was that modern warfare requires soldiers to think like never before. They have to do things that require large amounts of information, such as manage drones or oversee a patrol from a remote location. The system can help soldiers make sense of the flood of information streaming in. So if the sensors detect that the wearer's spatial memory is becoming saturated, new information will be sent in a different form, say via an audio alert instead of text. In a trial in 2005 the device achieved a 100% improvement in recall and a 500% increase in working memory.

Is this everybody's future? Probably not. But as the torrent of information increases, it is not surprising that people feel overwhelmed. "There is an immense risk of cognitive overload," explains Carl Pabo, a molecular biologist who studies cognition. The mind can handle seven pieces of information in its short-term memory and can generally deal with only four concepts or relationships at once. If there is more information to process, or it is especially complex, people become confused.

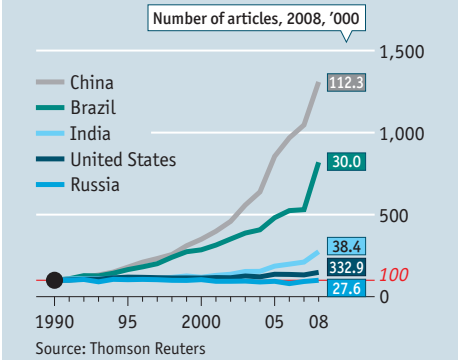
Moreover, knowledge has become so specialised that it is impossible for any individual to grasp the whole picture. A true understanding of climate change, for instance, requires a knowledge of meteorology, chemistry, economics and law, among many other things. And whereas doctors a century ago were expected to keep up with the entire field of medicine, now they would need to be familiar with about 10,000 diseases, 3,000 drugs and more than 1,000 lab tests. A study in 2004 suggested that in epidemiology alone it would take 21 hours of work a day just to stay current. And as more people around the world become more educated, the flow of knowledge will increase even further. The number of peer-reviewed scientific papers in China alone has increased 14-fold since 1990 (see chart 3).

"What information consumes is rather obvious: it consumes the attention of its recipients," wrote Herbert Simon, an economist, in 1971. "Hence a wealth of information creates a poverty of attention." But just as it is machines that are generating most of the data deluge, so they can also be put to work to deal with it. That highlights the role of "information intermediaries". People rarely deal with raw data but consume them in processed form, once they have been aggregated or winnowed by computers. Indeed, many of the technologies described in this report, from business analytics to recursive machine-learning to visualisation software, exist to make data more digestible for humans.

Some applications have already become so widespread that they are taken for granted. For example, banks use credit scores, based on data about past financial transactions, to judge an applicant's ability to repay a loan. That makes the process less subjective than the say-so of a bank manager. Likewise, landing a plane requires a lot of mental effort, so the process has been largely automated, and both pilots and passengers feel safer. And in health care the trend is towards "evidence-based medicine", where not only doctors but computers too get involved in diagnosis and treatment.

Too much to read 3

Scientific articles published in peer-reviewed journals by authors' origin, 1990=100



Source: Thomson Reuters

The dangers of complacency

In the age of big data, algorithms will be doing more of the thinking for people. But that carries risks. The technology is far less reliable than people realise. For every success with big data there are many failures. The inability of banks to understand their risks in the lead-up to the financial crisis is one example. The deficient system used to identify potential terrorists is another.

On Christmas Day last year a Nigerian man, Umar Farouk Abdulmutallab, tried to ignite a hidden bomb as his plane was landing in De-

troit. It turned out his father had informed American officials that he posed a threat. His name was entered into a big database of around 550,000 people who potentially posed a security risk. But the database is notoriously flawed. It contains many duplicates, and names are regularly lost during back-ups. The officials had followed all the right procedures, but the system still did not prevent the suspect from boarding the plane.

One big worry is what happens if the technology stops working altogether. This is not a far-fetched idea. In January 2000 the torrent of data pouring into America's National Security Agency (NSA) brought the system to a crashing halt. The agency was "brain-dead" for three-and-a-half days, General Michael Hayden, then its director, said publicly in 2002. "We were dark. Our ability to process information was gone."

If an intelligence agency can be hit in this way, the chances are that most other users are at even greater risk. Part of the solution will be to pour more resources into improving the performance of existing technologies, not just pursue more innovations. The computer industry went through a similar period of reassessment in 2001-02 when Microsoft and others announced that they were concentrating on making their products much more secure rather than adding new features.

Another concern is energy consumption. Processing huge amounts of data takes a lot of power. "In two to three years we will saturate the electric cables running into the building," says Alex Szalay at Johns Hopkins University. "The next challenge is how to do the same things as today, but with ten to 100 times less power."

It is a worry that affects many organisations. The NSA in 2006 came close to exceeding its power supply, which would have blown out its electrical infrastructure. Both Google and Microsoft have had to put some of their huge data centres next to hydroelectric plants to ensure access to enough energy at a reasonable price.

Some people are even questioning whether the scramble for ever more information is a good idea. Nick Bostrom, a philosopher at Oxford University, identifies "information hazards" which result from disseminating information that is likely to cause harm, such as publishing the blueprint for a nuclear bomb or broadcasting news of a race riot that could provoke further violence. "It is said that a little knowledge is a dangerous thing," he writes. "It is an open question whether more knowledge is safer." Yet similar concerns have been raised through the ages, and mostly proved overblown.

Knowledge is power

The pursuit of information has been a human preoccupation since knowledge was first recorded. In the 3rd century BC Ptolemy stole every available scroll from passing travellers and ships to stock his great library in Alexandria. After September 11th 2001 the American Defence Department launched a program called "Total Information Awareness" to compile as many data as possible about just about everything—e-mails, phone calls, web searches, shopping transactions, bank records, medical files, travel history and much more. Since 1996 Brewster Kahle, an internet entrepreneur, has been recording all the content on the web as a not-for-profit venture called the "Internet Archive". It has since expanded to software, films, audio recordings and scanning books.

There has always been more information than people can mentally process. The chasm between the amount of information and man's ability to deal with it may be widening, but that need not be a cause for alarm. "Our sensory and attentional systems are tuned via evolution and experience to be selective," says Dennis Proffitt, a cognitive psychologist at the University of Virginia. People find patterns to compress information and make it manageable. Even Commander Schmorrow does not think that man will be replaced by robots. "The flexibility of the human to consider as-yet-unforeseen consequences during critical decision-making, go with the gut when problem-solving under uncertainty and other such abstract reasoning behaviours built up over years of experience will not be readily replaced by a computer algorithm," he says.

The cornucopia of data now available is a resource, similar to other resources in the world and even to technology itself. On their own, resources and technologies are neither good nor bad; it depends on how they are used. In the age of big data, computers will be monitoring more things, making more decisions and even automatically improving their own processes—and man will be left with the same challenges he has always faced. As T.S. Eliot asked: "Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?" ■