Managing Information

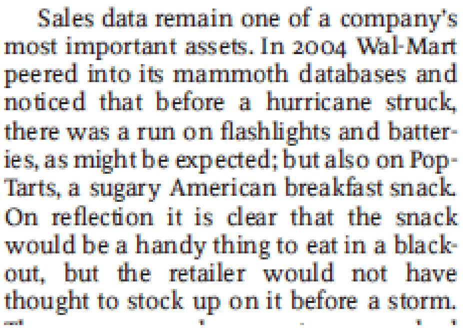
Data, data everywhere

Special report on managing information

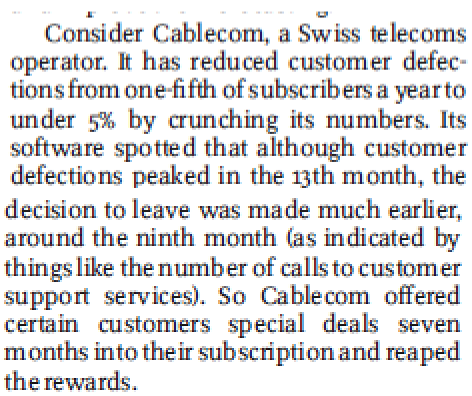
Economist - 2010

A different game

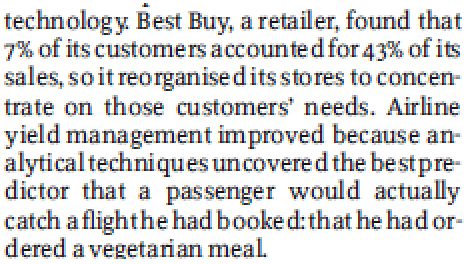
Information is transforming traditional businesses



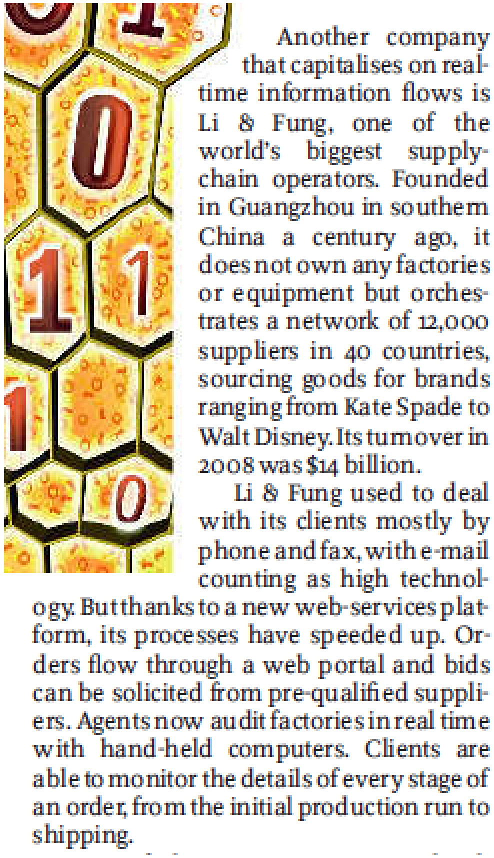
Sales data remain one of a company's most important assets. In 2004 Wal-Mart peered into its mammoth databases and noticed that before a hurricane struck, there was a run on flashlights and batteries, as might be expected; but also on Pop-Tarts, a sugary American breakfast snack. On reflection it is 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.



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.



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.



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.

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.

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.

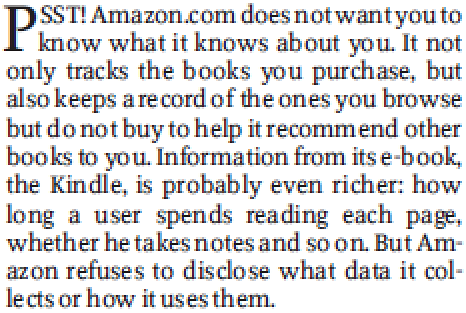
technology. Best Buy, a retailer, found that 7% of its customers accounted for 43% of its sales, so it reorganised its stores to concentrate on those customers' needs. Airline yield management improved because analytical techniques uncovered the best predictor that a passenger would actually catch a flight he had booked: that he had ordered a vegetarian meal.

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.

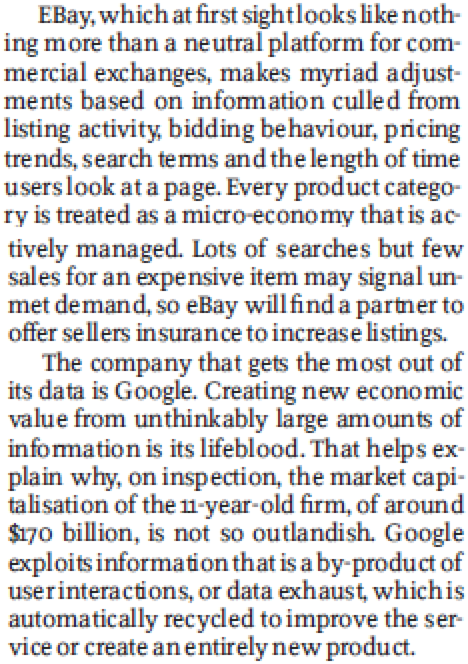
Li & Fung used to deal with its clients mostly by phone and fax, with e-mail counting as high technology. But thanks to a new web-services platform, its processes have speeded up. Orders flow through a web portal and bids can be solicited from pre-qualified suppliers. Agents now audit factories in real 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.

Clicking for Gold

How internet companies profit from data on the web

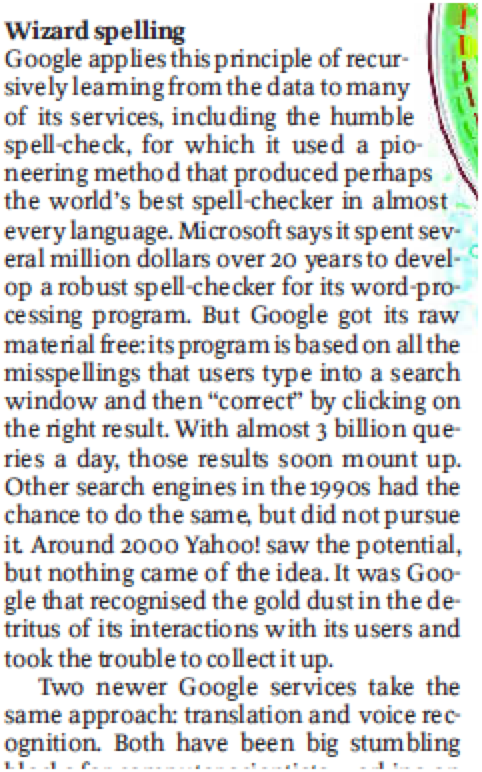


PSST! Amazon.com does not want you to know what it knows about you. It not only tracks the books you purchase, but also keeps a record of the ones you browse but do not buy to help it recommend other books to you. Information from its e-book, the Kindle, is probably even richer: how long a user spends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



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.

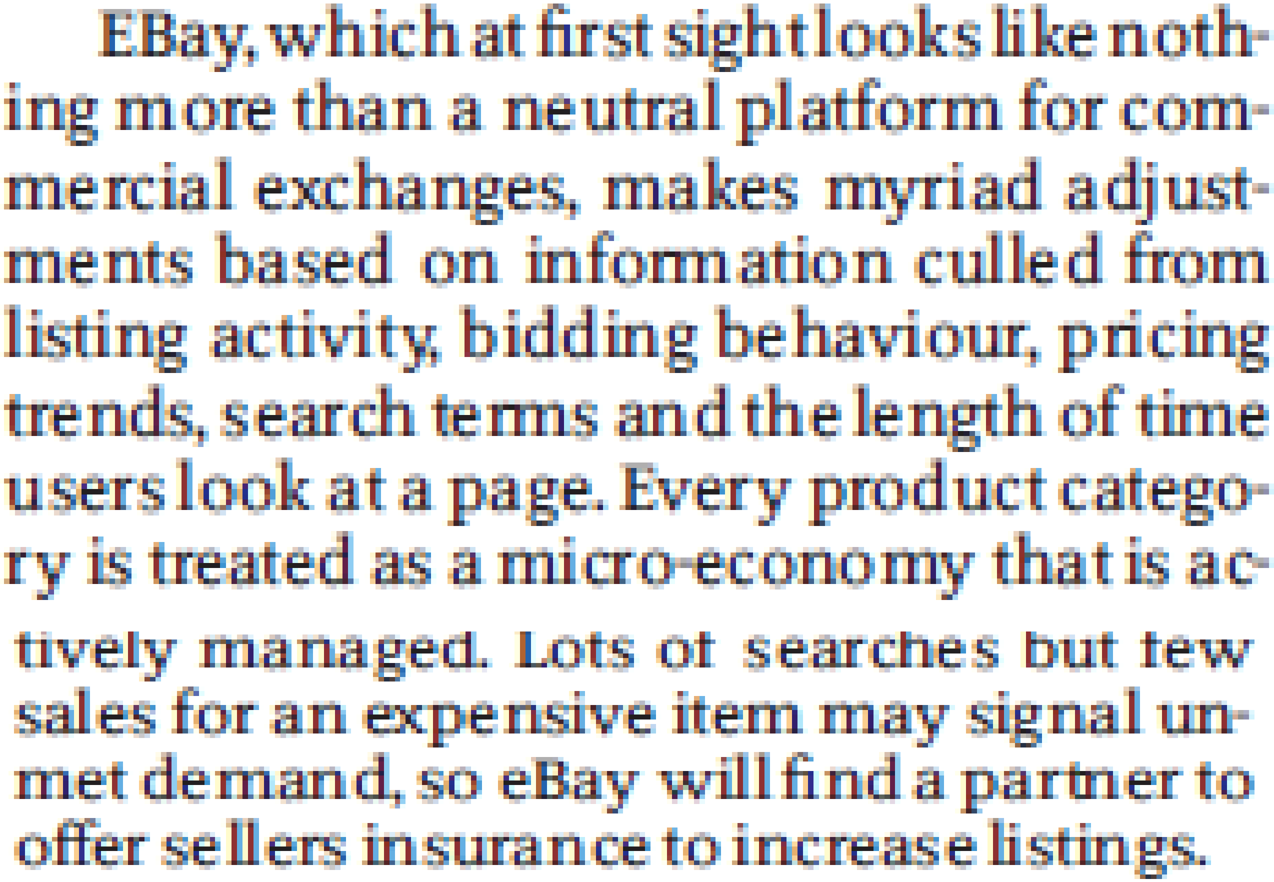


Wizard spelling

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

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.



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.

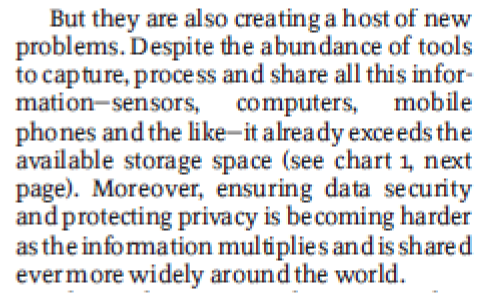
The company that gets the most out of its data is Google. Creating new economic value from unthinkably large amounts of information is its lifeblood. That helps explain why, on inspection, the market capitalisation of the 11-year-old firm, of around $170 billion, is not so outlandish. Google exploits 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.

Wizard spelling

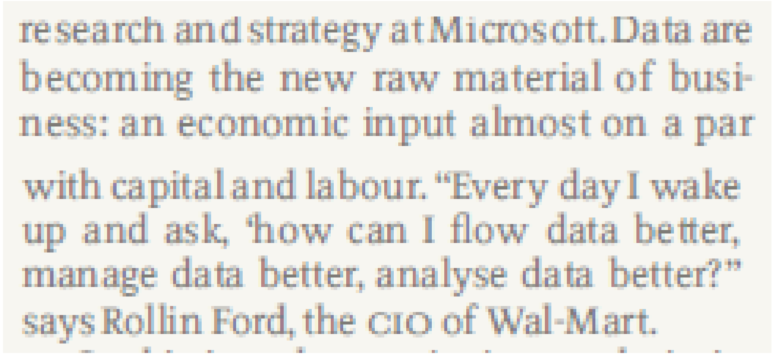
Google applies this principle of recursively learning from the data to many of its services, including the humble spell-check, for which it used a pioneering method that produced perhaps the world's best spell-checker in almost every language. Microsoft says it spent several million dollars over 20 years 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

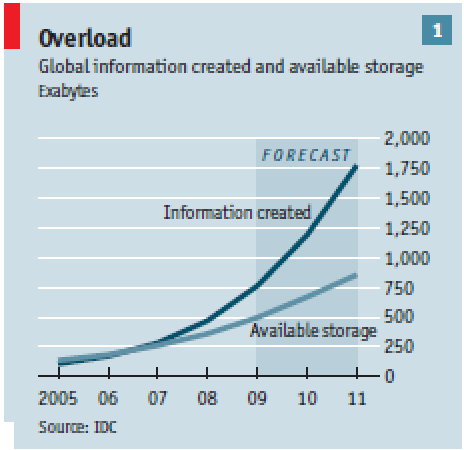
Data, data everywhere



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, next page). Moreover, ensuring data security and protecting privacy is becoming harder as the information multiplies and is shared ever more widely around the world.



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.

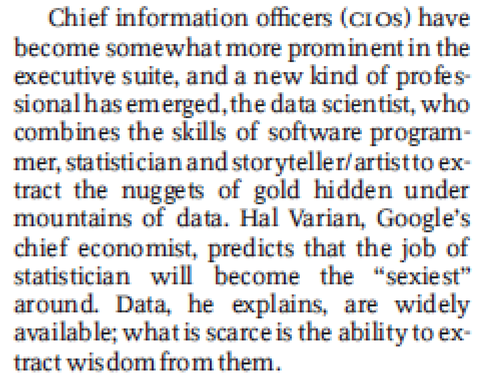


Overload

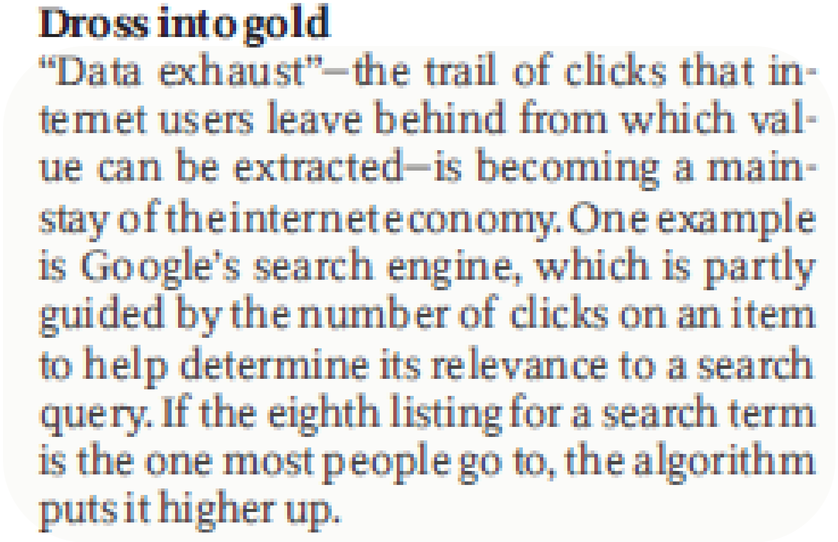
Global information created and available storage

Exabytes

Source: IDC

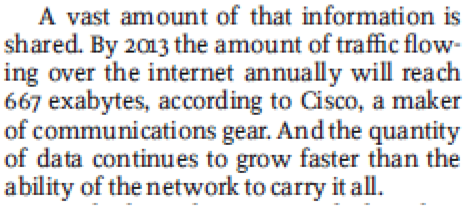


Chief information officers (CIOs) have become somewhat more prominent in the executive suite, and a new kind of professional has emerged, the data scientist, who combines the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data. 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.

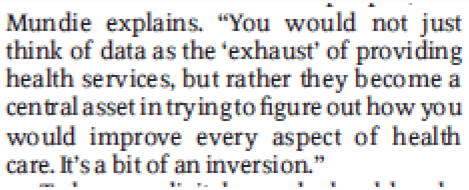


Dross into gold

"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.

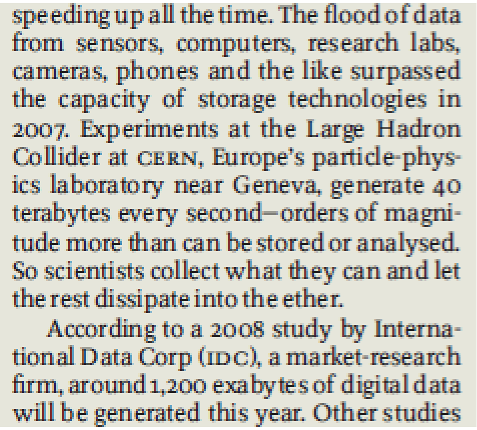


A vast amount of that information is shared. By 2013 the amount of traffic flowing over the internet annually will reach 667 exabytes, according to Cisco, a maker of communications gear. And the quantity of data continues to grow faster than the ability of the network to carry it all.



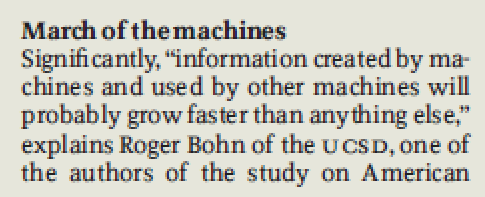
Mundie explains. "You would not just think of data as the 'exhaust' of providing health services, but rather they become a central asset in trying to figure out how you would improve every aspect of health care. It's a bit of an inversion."

All too Much



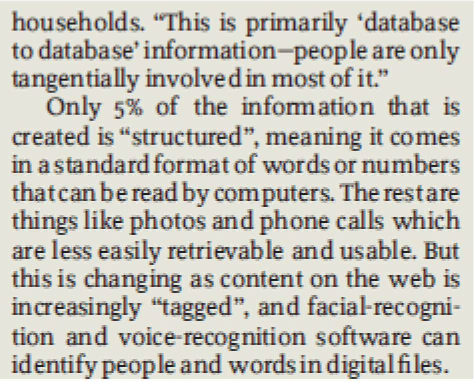
speeding up all the time. The flood of data from sensors, computers, research labs, cameras, phones and the like surpassed the capacity of storage technologies in 2007. Experiments at the Large Hadron Collider at CERN, Europe's particle-physics laboratory near Geneva, generate 40 terabytes every second—orders 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



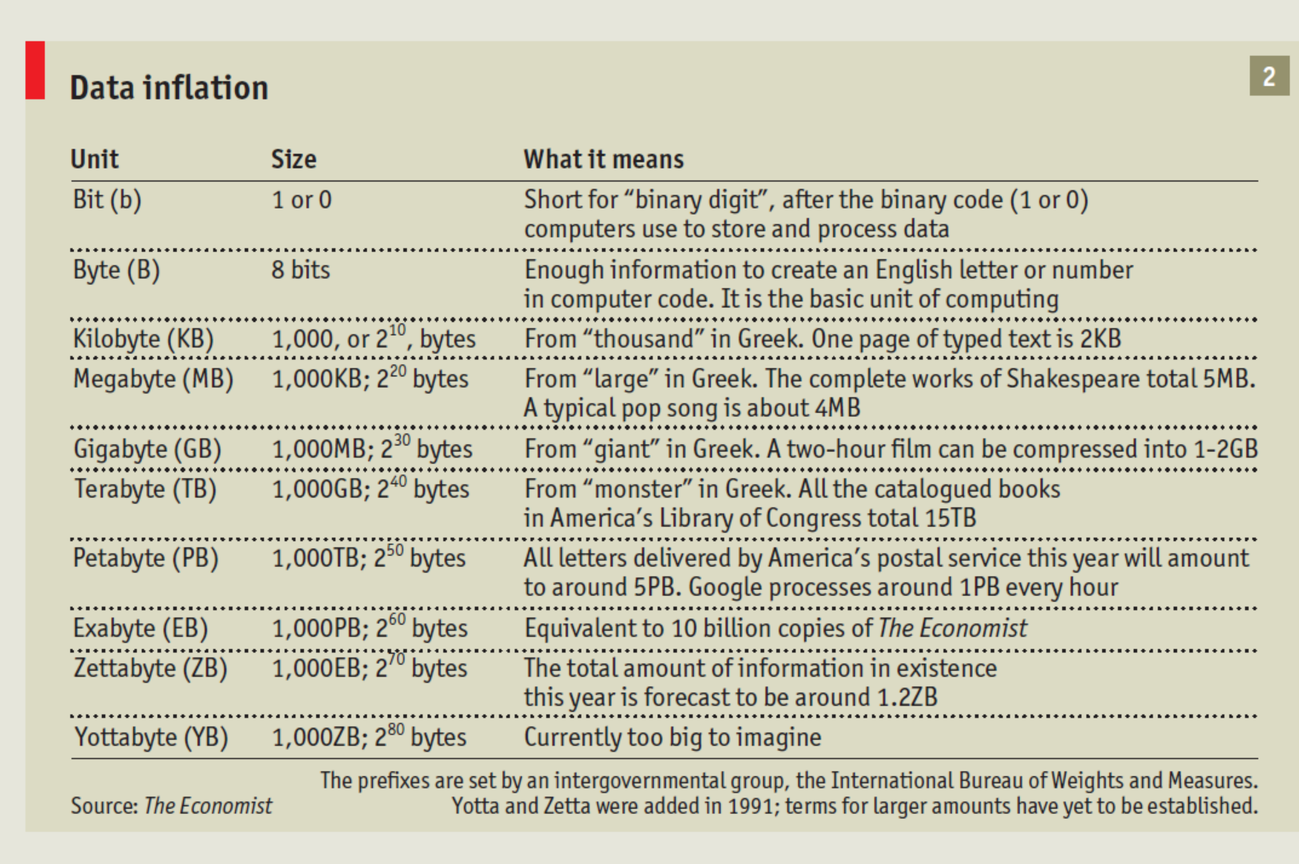
March of the machines

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.



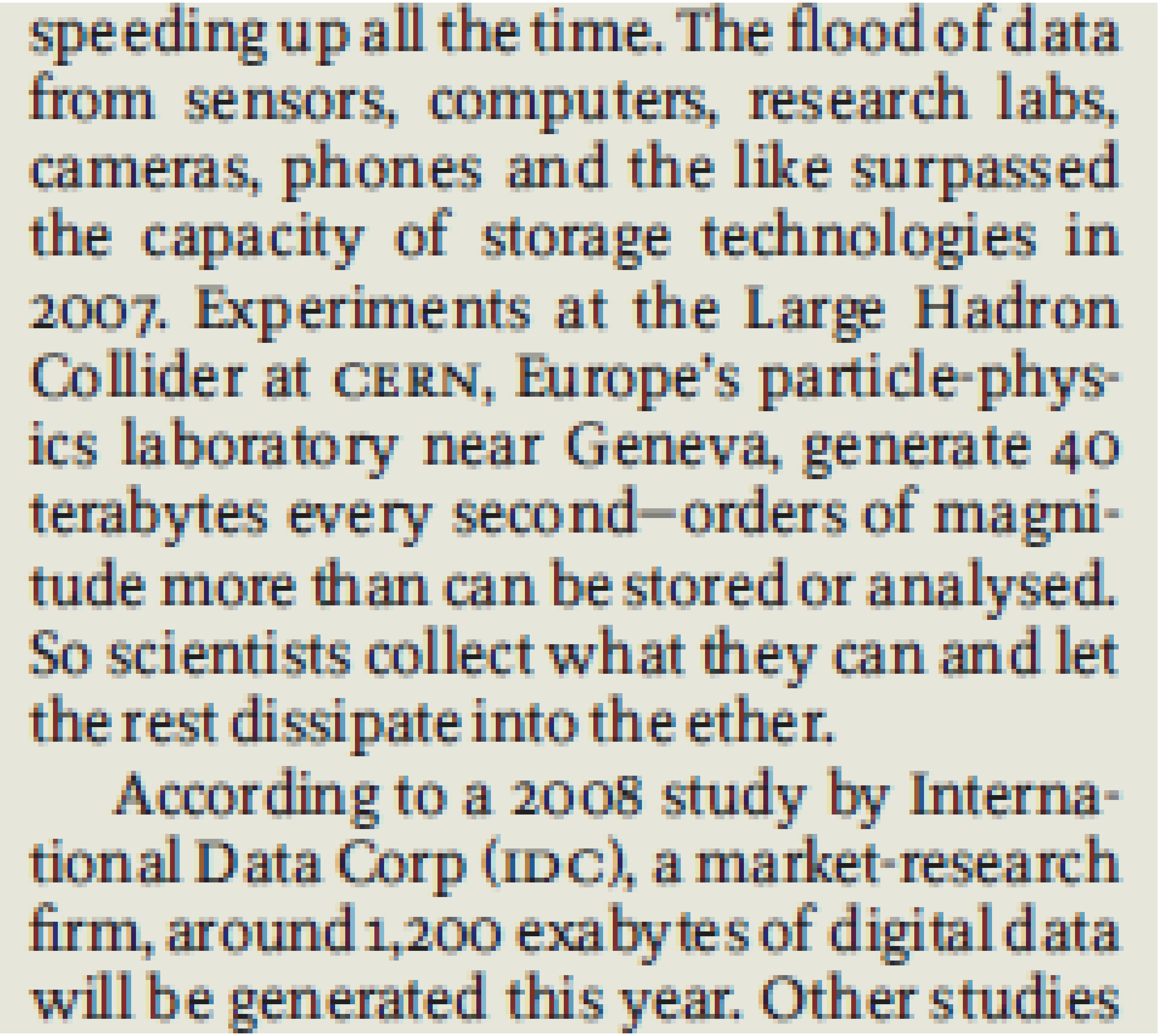
Data inflation

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.



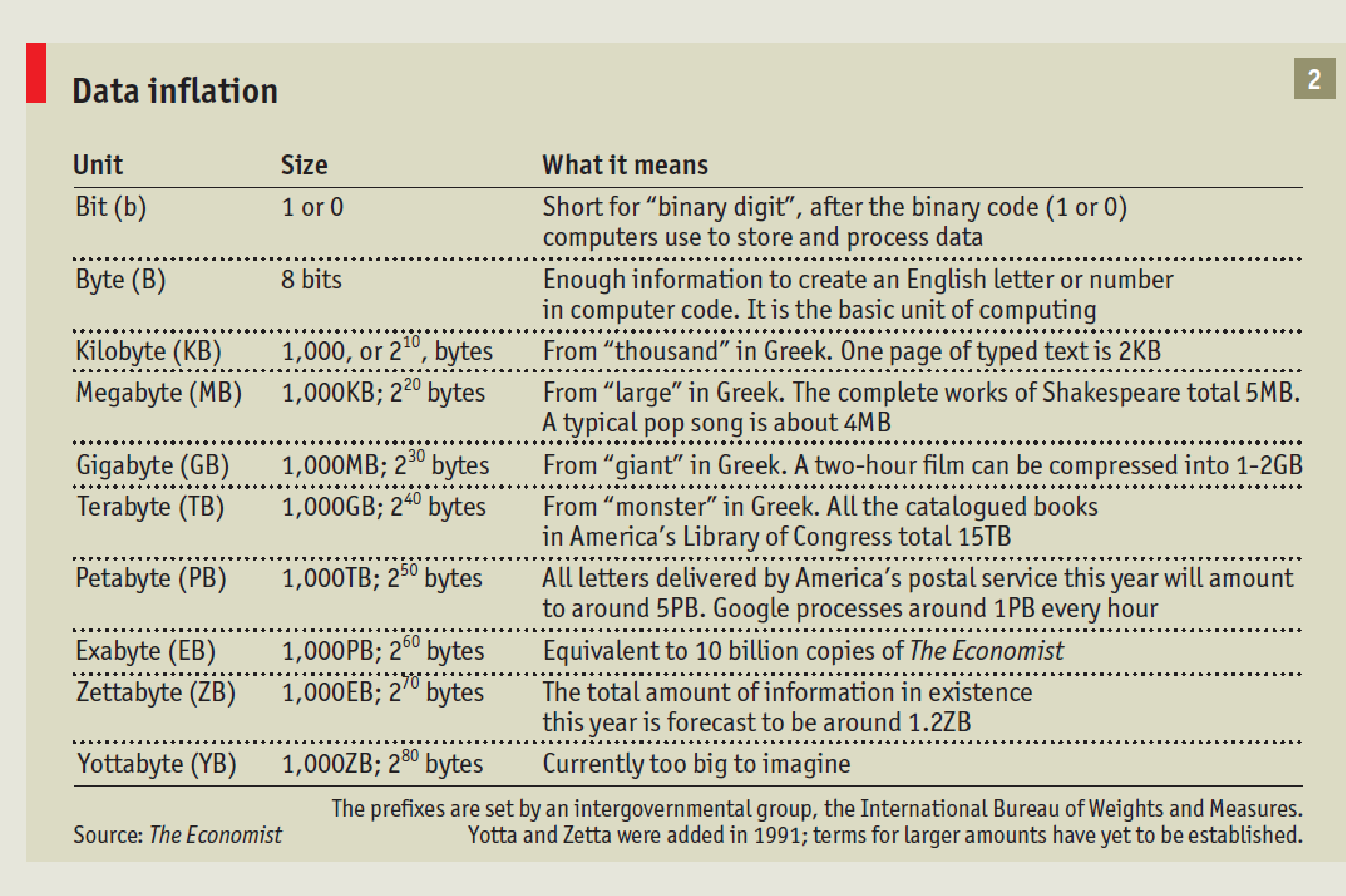
speeding up all the time. The flood of data from sensors, computers, research labs, cameras, phones and the like surpassed the capacity of storage technologies in 2007. Experiments at the Large Hadron Collider at CERN, Europe's particle-physics laboratory near Geneva, generate 40 terabytes every second—orders 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

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.



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

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.

Source: *The Economist*