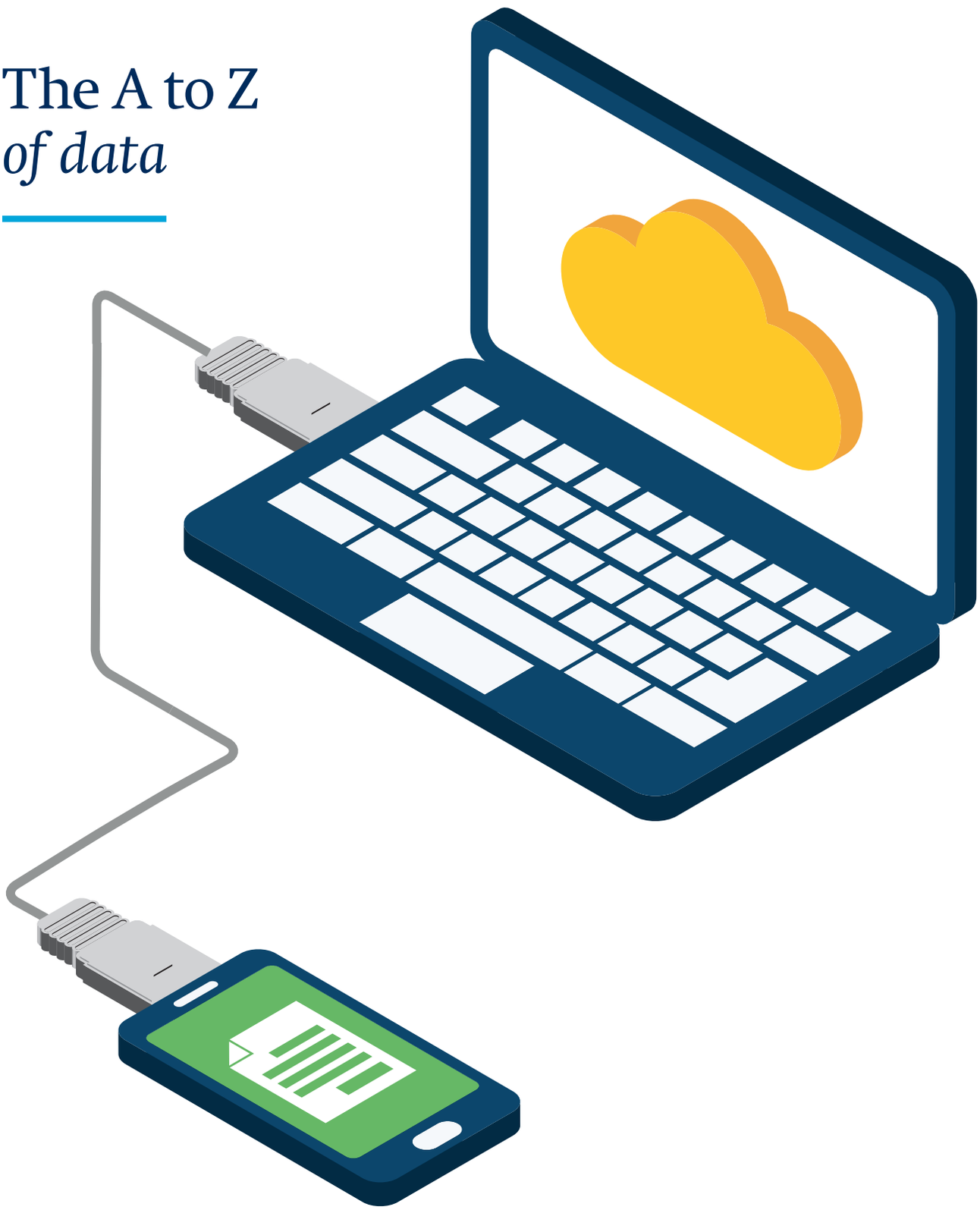
Freshfields Bruckhaus Deringer

The A to Z *of data*



The A to Z of data

A is for antitrust

The collection and use of data is a major area of focus for antitrust regulators. Authorities in Europe and the US are concerned that consumer-facing businesses are collecting so much data that they are becoming dominant. They are also examining data in the context of M&A and considering whether the combination of data sets in deals leads to dominance, or even reduced privacy for consumers.

B is for BEPS

Our survey shows that businesses expect enhanced use of data to shift profit centres. Businesses therefore need to think about how they structure their data analytics in light of the OECD's BEPS recommendations, which are designed to align tax with value creation. Data sets often create value only in combination with other information, so where this takes place could be affected by BEPS.

C is for consent

The new EU Data Protection Regulation – which governs how businesses use European citizens' personal data – creates a harmonised regime across the EU. Among other things it imposes stricter rules on the type of consents needed to work with personal information.

D is for datafication

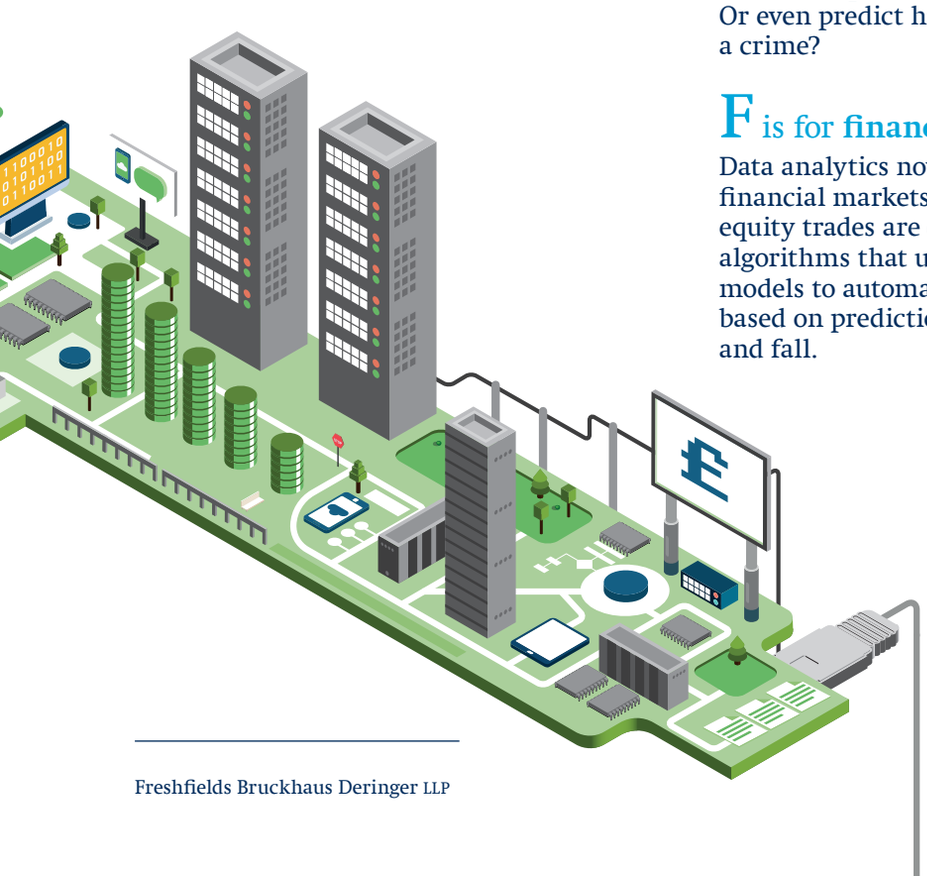
A phenomenon identified by Viktor Mayer-Schonberger and Kenneth Cukier in *Big data – a revolution that will transform how we live, work and think*. Datafication is the process by which activity is measured and tabulated in a way that can be analysed. Amazon's Kindle datafies books; Facebook datafies our relationships; Twitter datafies our emotions.

E is for ethics

There is much debate about the ethics of big data analytics. Should we allow algorithms to decide our insurance premiums based on individual predictions of life expectancy? Or even predict how likely we are to commit a crime?

F is for finance

Data analytics now dominate the world's financial markets. Around two-thirds of US equity trades are carried out by computer algorithms that use advanced mathematical models to automatically buy and sell stocks based on predictions of when they will rise and fall.



G is for **Google**

A big data pioneer. Google builds many of its products via smart data analytics and machine learning, from its search engine to Google Translate (which was developed by pouring in millions of existing translations from across the internet and using algorithms to spot the correlations between languages). Parent company Alphabet's executive chairman Eric Schmidt is a major political player as tech adviser to Barack Obama and an influential voice in the debate about data privacy and net neutrality.

H is for **Hadoop**

An open-source software framework that chops massive data sets into smaller pieces, distributes them among multiple machines and allows the entire set to be analysed simultaneously.

I is for **intellectual property**

IP rights are a vital consideration of a holistic data strategy. Confidential business secrets are protected by IP law in both Europe and the US, while European business data may also qualify for database rights if the company has invested significant sums in Europe collecting the data.

J is for **jurisdiction**

Complying with data regulations can be challenging for multinationals given the competing – and often conflicting – requirements in different jurisdictions. Privacy regulations are tightest in Europe, while some countries, including Russia, China and Indonesia, do not allow their citizens' personal data to be exported. In the US, the SEC requires disclosure of cyber incidents and is considering requesting certification of internal controls as a part of financial reporting. Sector-specific regulations further complicate the picture.

K is for **Kafka**

Kafka is an open-source software system developed by LinkedIn that brings together disparate streams of data in real time. Creator Jay Kreps describes it as a 'central nervous system' that takes data feeds, processes them and sends the information where it needs to go. Alongside LinkedIn, Kafka also powers real-time data processing systems at Verizon, Uber and Netflix.

L is for **location**

Location data has huge potential for monetisation. Mobile operators for example – whose traditional income streams have been undermined by the rise of messaging platforms such as WhatsApp and regulatory regimes such as the EU's Telecoms Single Market, which will scrap roaming charges – are beginning to realise the value in the location data their infrastructure collects, which can be licensed for others to analyse.

M is for **machine learning**

A branch of artificial intelligence, machine learning systems improve their performance over time by analysing vast quantities of data and adjusting the way they behave based on the patterns they see. Machine learning is used by Google to improve the performance of its search engine (by correlating search queries with the links that are clicked on) and its Translate system (by analysing millions of passages of translated text to spot the most common connections).

N is for **NoSQL**

Big data analysis relies on NoSQL, a data query language that is able to process information that isn't organised in fields like traditional databases. The mountain of data in existence is like an iceberg with the 5 per cent above the water consisting of structured data and the mass below the surface made up of unstructured videos, audio files and web pages. NoSQL's ability to make sense of this information makes advanced analytics possible.

O is for **Oakland Athletics**

The Oakland A's baseball team are a great example of effective data strategy in action. The problem: poor performance from a well-paid roster. The goal: greater success at lower cost. The strategy: harnessing data analytics to identify effective performers on lower salaries. The result: the A's return to the playoffs with one of the cheapest teams in the league.

P is for **'poison pill'**

Most data privacy regimes require informed and explicit consent to process personal data. In the context of big data, information is often gathered without a pre-defined use case, making such consents almost impossible to obtain. To unlock its potential it must therefore be anonymised. Identifiable personal data can act as a 'poison pill' in data pools and increase regulatory risk.

Q is for **quality**

Robust data sharing arrangements define what type of data a business would allow to flow into a data pool – and how to maintain its quality. While using massive data sets for predictive analytics gives greater tolerance for statistical anomalies, in reality most businesses don't work with such large quantities of data and therefore need to ensure they remove outliers.

R is for Rolls-Royce

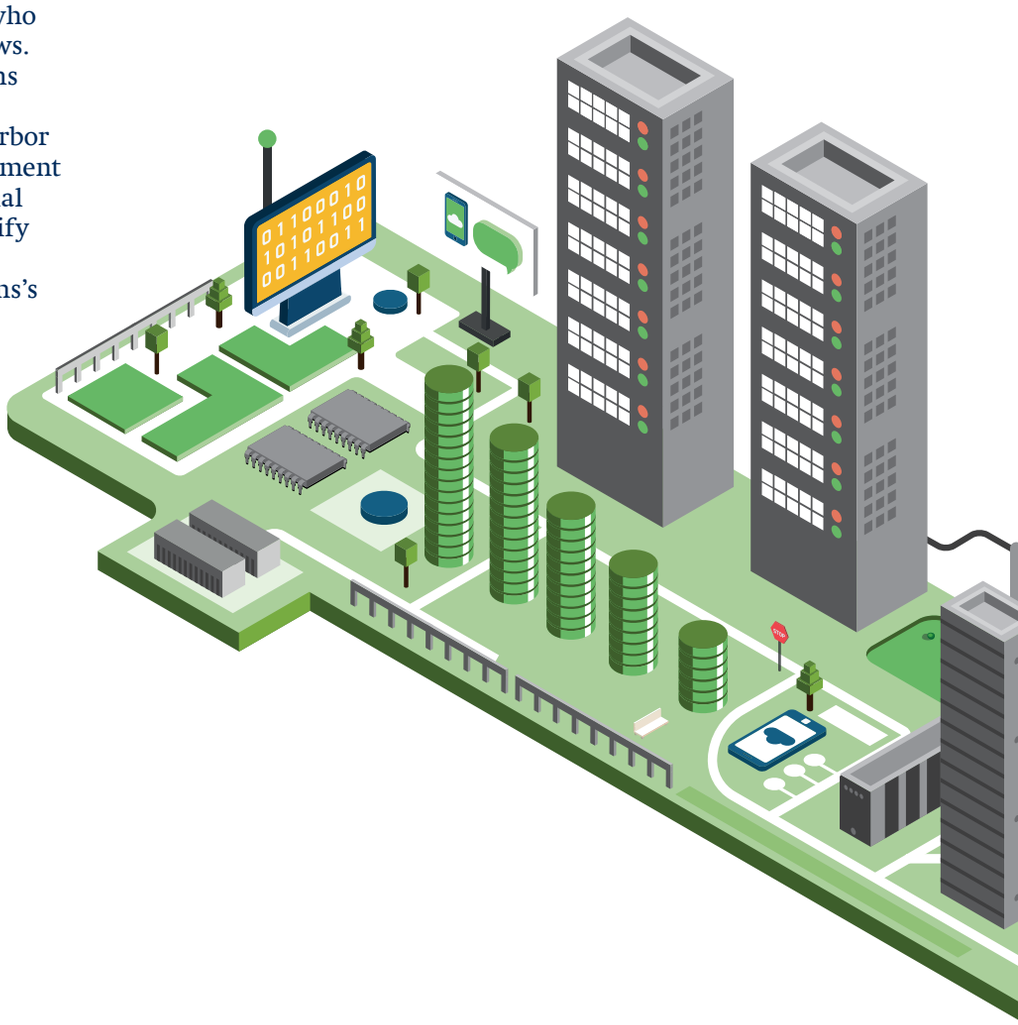
Rolls-Royce has cut maintenance times for its engines using data from embedded sensors. Where once its products would have been serviced at regular intervals, they are now maintained on an 'as necessary' basis. Data on everything from fuel efficiency to potential parts failures is transmitted from the engines via satellite link. Engineers analyse this to spot potential issues and fix them before they escalate, averting the need to take aircraft out of service for extended periods.

S is for Schrems

Max Schrems is the Austrian student who single-handedly altered global data flows. Following Edward Snowden's revelations about mass surveillance by the NSA, Schrems challenged the EU/US Safe Harbor deal in the European courts. The agreement allowed businesses transferring personal data from Europe to the US to self-certify that they offered 'adequate' protection under EU law. The ECJ accepted Schrems's argument and ruled against the deal. Negotiations recently concluded on its successor, the Privacy Shield.

T is for Telefonica

The Spanish telco set up a new business unit – Dynamic Insights – in 2012 to monetise its subscriber data. The company sells anonymised location information to retailers and public sector organisations in the UK, Germany and Brazil, enabling them to see where subscribers in certain age groups are at particular times in the day. The data allows retailers to tailor products and promotions, and monitor in real time whether they are working. It is also used by city councils to assess the impact of free parking pilots, for example.



U is for **unstructured**

Unstructured data cannot be easily indexed or mapped into database fields. It is often user-generated – eg emails or instant messages – and is mined by businesses to create value. Analysing unstructured data from social media, for example, enables companies to assess sentiment around a marketing campaign or product launch.

V is for **volume, velocity** and **variety**

Perhaps the most commonly used definition of big data comes from Gartner analyst Doug Laney, who quantifies it in relation to 'three Vs' – volume, velocity and variety. The higher the volume of data, the faster it arrives and the greater the variety (eg text, video, email, audio), the 'bigger' the data.

W is for **Walmart**

In 2013 the world's biggest retailer began correlating its sales information with weather data to drive its merchandising and digital ads at a hyper-local level. The data shows demand for berries rises when winds are low – particularly when the temperature is below 80F. With the same wind conditions but temperatures above 80F, salad sells. And when winds are low, the temperature is high and the sky is clear, ground beef is popular. Walmart has also started promoting items based on Pinterest trends, and is planning to produce tailored maps of its stores to help consumers find products on their shopping lists saved in Walmart's app.

X is for **xenottabyte**,

Y is for **yottabyte** and

Z is for **zettabyte**

Data storage is quantified in bytes, kilobytes, megabytes, gigabytes, terabytes, petabytes, exabytes, zettabytes, yottabytes, xenottabytes, shilentnobytes and domegemegrottebytes, each unit 1,000 times larger than the last. Data volumes are growing exponentially, and KPMG predicts that by 2020 there will be zettabytes of data in existence. In 2013 the entire World Wide Web comprised just 4ZB.

freshfields.com