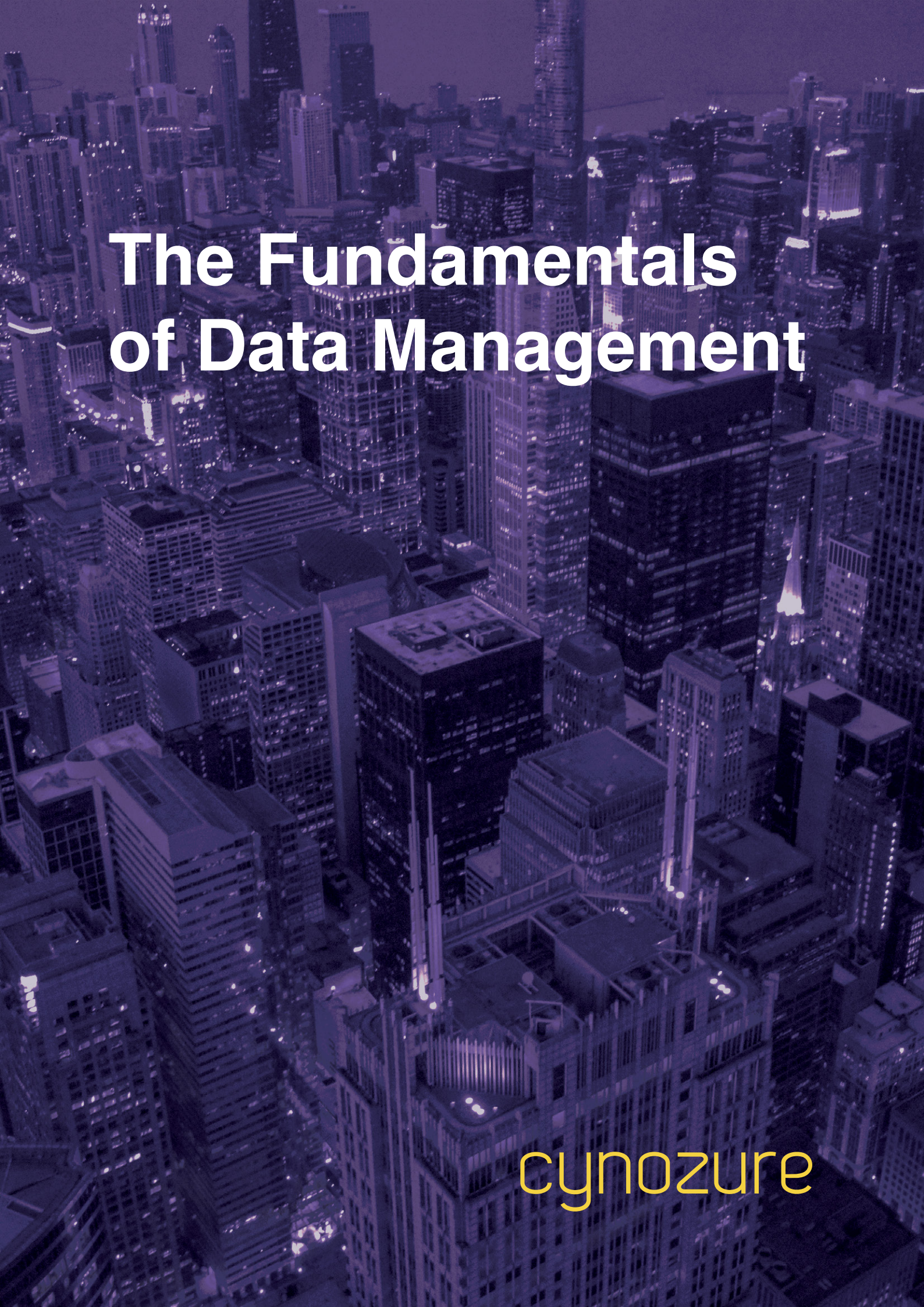
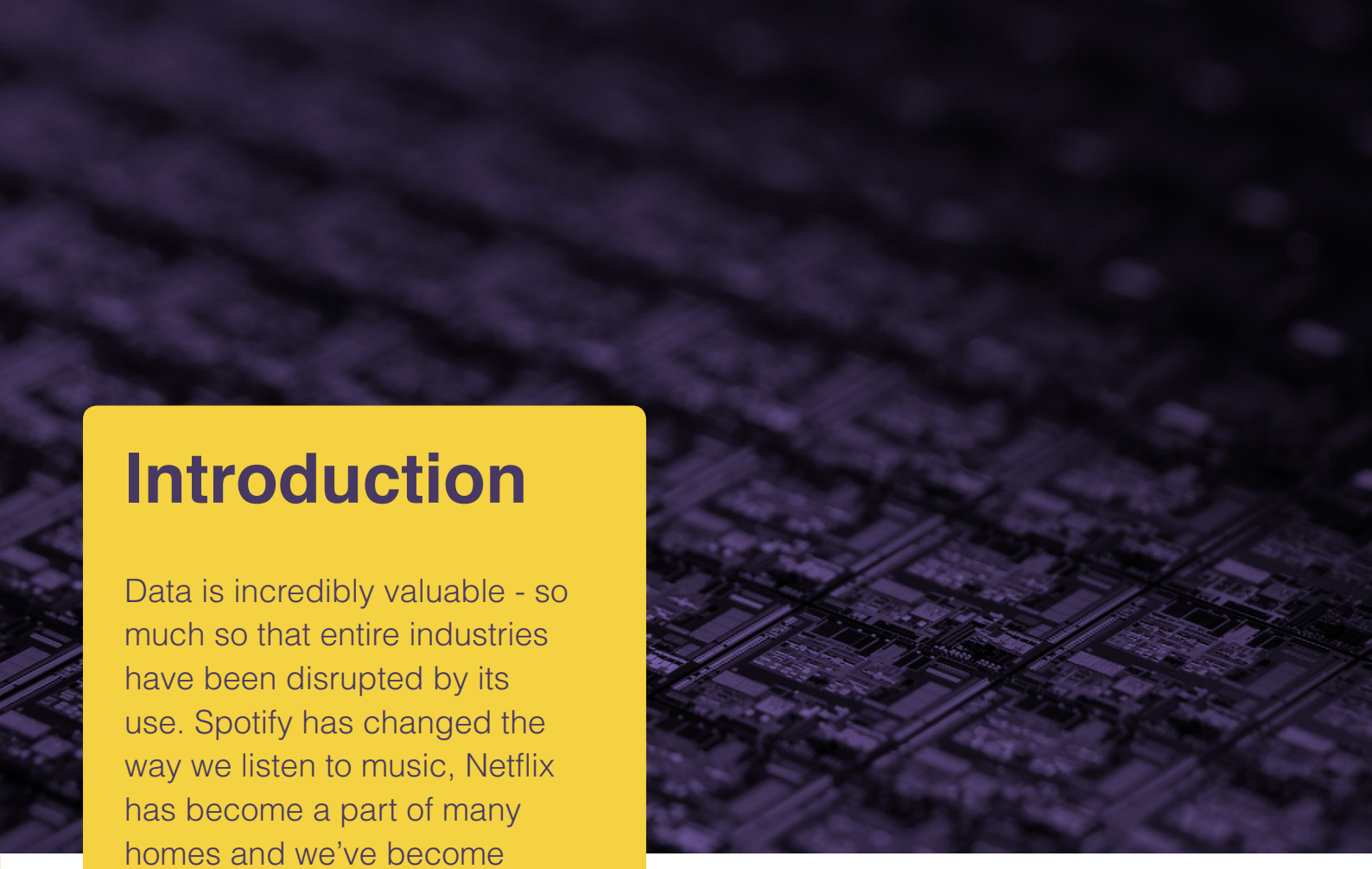


The Fundamentals of Data Management

cynozure

The Fundamentals of Data Management



Introduction

Data is incredibly valuable - so much so that entire industries have been disrupted by its use. Spotify has changed the way we listen to music, Netflix has become a part of many homes and we've become used to asking Siri and Alexa to help with basic tasks. Like any valuable asset, data needs to be governed, managed and secured effectively.

Increasingly, companies are falling foul of poor data governance and data management. Not just in the form of high profile hacks, but also fiascos like TSB's online and mobile banking crisis[1]. Organisations that fail to govern data well are setting themselves up for failure. Good data governance underpins every other data function and is critical for gaining trust in data. Without that, data projects will fail, results will be second-guessed, artificial intelligence models might be trained with incorrect data and ultimately the data strategy will be undermined.

Data governance is not static either. It is always evolving according to business and industry needs - a good data governance strategy will reflect this by being agile and responsive to change.

There are new challenges posed by the rise of real-time data and the Internet of Things (IoT), and new requirements from legislation such as the General Data Protection Regulation (GDPR). Today's technology and data leaders must respond to these, as well as manage current-day challenges surrounding data governance.

Effective data governance is a cross-disciplinary activity, not just the remit of a Chief Technology Officer, a Chief Information Officer or Chief Data Officer. Futureproofing data governance and data management is a cross-team endeavour.

The definition of data management

Data management involves the policies, processes and architecture that has been put in place to manage data at all stages of its life-cycle. From ingestion and storage to processing and analysis.

Through the implementation of a new data management strategy, a medical retail company[1a] managed to gain supply chain efficiencies, a reduced time-to-market, a 360-degree view on reporting and fewer errors in shipping handling. The IT team also spent less time addressing data quality issues.

Data management for the modern day

Today's data management is under tremendous pressure. It has to evolve to meet new, complex data sources such as the IoT, smart cities and wearable data. The volume of data that has to be managed is also growing. The data market is worth some $33.5 billion with this figure expected to double over the next four years[2]. The amount of data generated through consumer Internet use alone stands at 9059 petabytes a month[3]. To put this in context, the 10 billion photos currently on Facebook take up approximately 1.5 petabytes[4].

The foundations of good data management need to be laid, now more than ever before, because of the sheer amount and complexity of data. The increasing ambition of business use cases is adding to this need. Organisations are seeking ways of using real-time data for quick decision-making and IoT technology for remote monitoring of the supply chain, for example.

But before any innovative use of data can be explored, put the right building blocks in place. Those that fail to take data governance and management seriously will find their data strategy and data projects come tumbling down.

Data management versus data governance

Data governance and data management are closely linked. However, whereas data management deals with the practical implementation of governance through architecture and policies, data governance is about setting the rules and defining responsibility. Without effective data governance in place, any data management efforts will not work.

Data legislation placing emphasis on better management

Changing data legislation is placing new emphasis on the value and risk involved with data. The much-publicised General Data Protection regulation in Europe has punitive fines for any breach of the Regulation - including data hacks and leaks. Privacy-by-design is now imperative when building data management processes.

GDPR as a catalyst for change - case study for a major online retailer

An e-commerce company with over 40% of its customers based in Europe, needed to prepare for GDPR before the May 2018 deadline. It took the deadline as an opportunity to move towards a more customer-centric data model.

It communicated with customers about the value of their data and the promises it made in relation to this - to protect data, to never spam, and to deliver any technical information in clear language. Customers were reassured with the knowledge that all data use was in line with ICO guidance and secured with a privacy-by-design system.

The technical side to getting ready for GDPR was more of a challenge. The business had personal customer data stored across multiple systems that all had to be consolidated for GDPR. It prioritised the data sets and storage that needed transferring first, in order to get the highest risk data sorted before the tight deadline. An audit of all data allowed the retailer to also remove any data sets that it did not need - duplicate personal information gained through social media, for instance.

Throughout the process, it kept all stakeholders informed and set up regular meetings between IT, legal, and marketing. Internal communications were also important, with company-wide meetings arranged where all employees learnt the value of their personal data, and how to keep customer data GDPR compliant.

This customer-centric approach, combined with cross-organisation collaboration and clear leadership from the CIO and other board members, enabled the company to hit the compliance deadline - but also changed how data is viewed and used in the organisation. This mindset and culture change means that the company can ensure GDPR compliance over the long term.

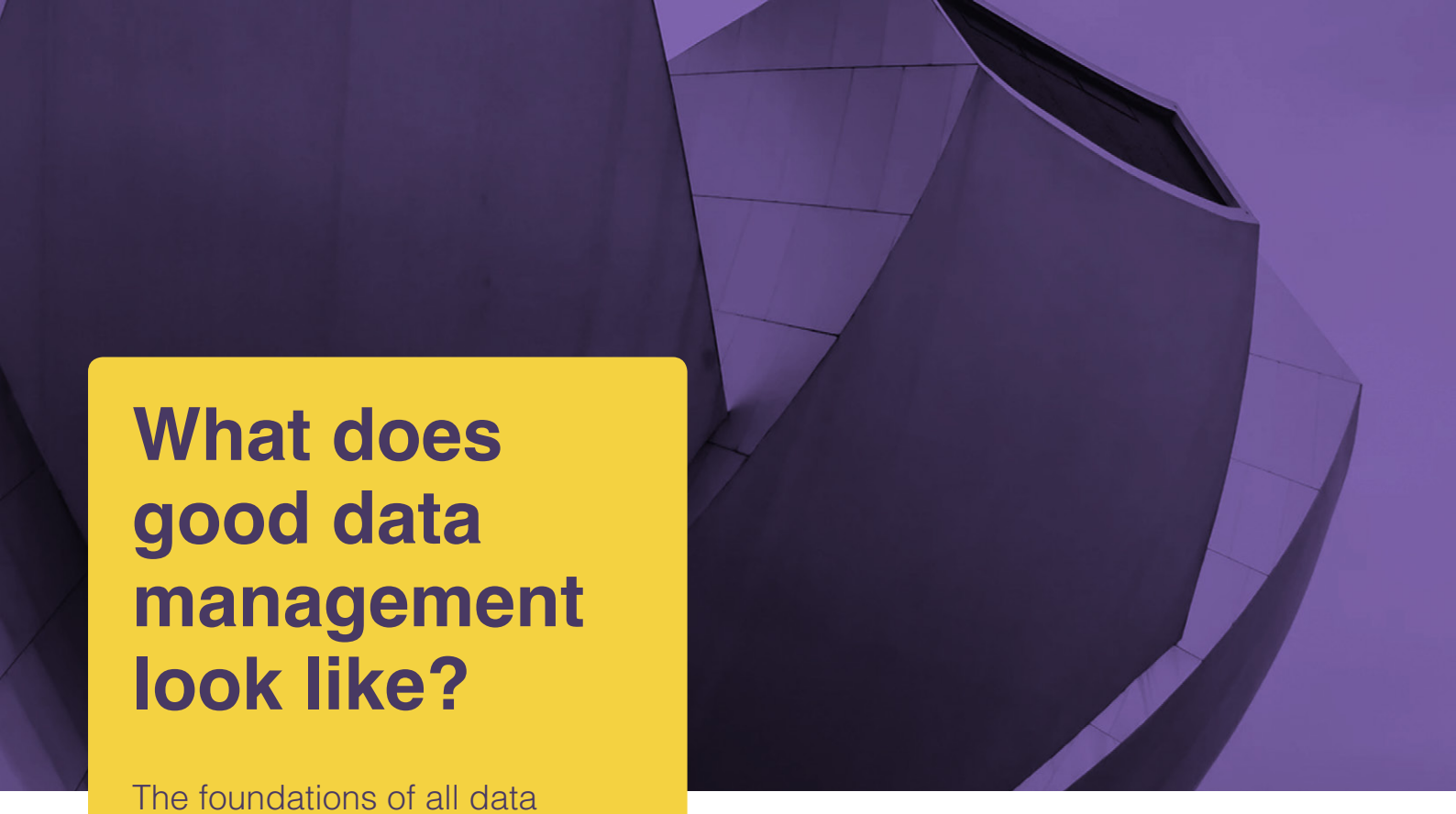
GDPR in the long term

The EU's General Data Protection Regulation (GDPR) was enforced on the 25th May 2018. It is the toughest data law to date, with fines for non-compliance running up to 20 million or 4% of a business' global revenue. Under GDPR, Google has been hit with a fine of £44 million[4a].

In the run-up to enforcement, there was a flurry of activity to prepare. However, work cannot stop after the initial preparation. The requirements of GDPR are ongoing. As a result, organisations need a long term strategy in place to monitor, store and refresh customer consent for their personal data use. Privacy-by-design data management and security are essential. We have found that organisations that do this well have evolved their culture surrounding data use to a more customer-first, privacy-first ethos.

Plus, there's the associated cost to reputation if a breach or hack does occur. Consumer trust in data is fragile and any company found with poor data management (and governance) leading to a loss of data, will struggle to regain that trust.

Legislators are reflecting this new realisation[5]. More are waking up to the need for stringent data controls and protection. Organisations must now follow suit by prioritising data governance; it's the best way to keep abreast of changing legislation and prepare for the future.



What does good data management look like?

The foundations of all data projects lie in good data governance and management. It's important to walk before you run with data projects. Lay the foundations first, and start with more 'basic' data projects, before you move onto more complex use cases such as using wearable biometric data.

Whilst the process of implementing good data management practices is not identical in every organisation, there is a framework and set of principles that tend to apply universally.

Identify and use data you need

Everything, from data strategy to data governance, begins with use cases. Identify and prioritise which ones to implement first, along with future use cases. This will determine what data you will have to collect and store, which in turns tells you the type of data management you'll require.

How to identify your use cases

When starting your data strategy, you might have a long list of projects that you want to do. However, budget and time constraints make prioritising a necessity. Plus, prioritising 'quick win' projects that will offer the speediest returns on your investment can help you demonstrate value and gain buy-in for future, more ambitious projects.

Alignment: To do this, consider the alignment between your business strategy (and goals) and the different data projects. If your business goal is to increase revenue from ticket sales, then a data project that can analyse when tickets are most in demand, and that can assist with dynamic pricing, would be useful to your organisation. By aligning data projects directly with business goals, you can quickly prove data's value to the stakeholders in your organisation.

Investment: Look at the projects that offer significant returns on investment in the shortest time frame. Pick your 'quick win' projects first and leave longer term, more intensive projects for further down the line when your organisation's culture is more data-driven, you have more buy-in and greater data literacy.

Independence: Sometimes, the projects that are being championed by your peers aren't always the best ones for the organisation at that time. Don't just go for the pet-projects of the loudest or highest-powered people in your meetings; follow a set framework for identifying projects.

With two or three clear use cases planned, you can put your data governance and management processes in place.

Processes and infrastructure

Storing data

When choosing a data storage solution, consider not just your current needs but also future plans. Therefore, at Cynozure we prefer to use the term 'data platform' instead of data lake or warehouse. It encompasses all layers of a modern data platform including raw data and allows businesses to store a wide range of data - and to access and use it in many different ways.

Most organisations will be well-served by a core, tabular style, relational database because the majority of business data will be in this format. Plus, it'll be relatively easy to hire people to work with it as the required skills are widely available in the industry. It makes a good starting point, then as your use cases become more advanced you can invest in other storage solutions. Don't forget that the more complex storage solutions will require additional team members with specialist skills to work with it.

However, a different type of storage is required in some scenarios. For example, when mapping out relationships between people, such as when an organisation wishes to map out its employees to better understand the communication between departments and common failure points. In this case, a graphical database that can easily visualise this data is a good choice.

There is no one-size-fits-all for organisations. Luckily, there are a host of different off-the-shelf solutions, so finding the right fit for your use cases shouldn't be too difficult.

Data Ingestion

Hand-in-hand with your data storage comes your data ingestion processes. These are vital to get right as getting them wrong will impact how quickly you can deliver the right data to recipients and the trust in the accuracy of that data.

As a starting point:

- **Build frameworks:** Create a standard for delivery across the company. It also makes it easier to go through a few different iterations before getting to your final process.
- **Go through a few iterations:** Don't expect to get your process completely right the first time. Expect to go through a few versions before finding the process that works best for your organisation and use case.
- **Don't focus too much on data format:** Begin with your use case and goals. From there, work out what data sets, whether structured or unstructured, you'll need.
- **Manage your metadata:** Enrich your raw data with a standard set of metadata. include information

about where the data came from, when it came in, its process ID and customer consent for data use (if applicable).

» **Ensure accurate error handling:** If something goes wrong it's much easier to unpick if you have accurate error handling and logging.

» **Audit data as it is ingested:** A lot of issues could arise when data is ingested, so it's vital that you validate your data with records of its distinct values in specific columns, for instance. This will be priceless if the quality of your data is called into question.

» **Ingest data that aligns with your goals:** Ingest the data that you can predict a use for in the future or that will help with your current plans. Keep adding new data sources, as you never know what insights you might uncover.

» **Secure your data:** Data security is covered in more detail later on, however, as a start, consider what data needs to be encrypted, and how your organisation handles sensitive data.

» **Don't transform as you ingest:** It's no longer efficient to extract, transform and load data in the traditional way. Instead, ingest data in its raw format and transform it afterwards. That also helps from a trust and quality standpoint.

There are many different data ingestion tools to decide between. These are split into three options:

Open source: Open-source tools are free, however, this cost saving might be offset by the need to hire the right people who can use the solution. Also, consider the long-term support (or lack thereof) offered.

Buying tools: Most organisations will purchase a tool to do most of the heavy lifting when ingesting data. For specific use cases that require it, they may then consider a custom-built solution.

Custom-built: Building a tool yourself gives great flexibility, however it requires significant investment in time and talent. Which means it's often out of the reach of many smaller organisations.

Whatever your end-process when ingesting data, remember to always begin with your use cases. Otherwise, you could end up with ingestion processes that don't suit the data you are trying to on-board.

Data terminology - the lake versus the warehouse

Data lakes and warehouses are each optimised for different purposes and it's key to use each one for what they were designed to do.

Data Warehouse: A store of data that has been modelled and/or structured, suitable for then creating insights for the organisation.

Data Lake: A type of storage that holds a vast quantity of raw data in its native format, including structured, semi-structured, and unstructured data (such as video). The data format and requirements are then refined when the data is needed.

There are pros and cons to each approach. Generally speaking, a data lake will allow for more flexibility because a data warehouse is highly structured and cannot be altered quickly. However, a data warehouse is usually more secure.

Data swamp: Be aware that a data lake could turn into a data swamp if it's not governed correctly. By putting all sorts of data into a lake, without much thought to its metadata, maintenance or use, it turns into a swamp. Data in a swamp cannot be accessed or used easily or might be used out-of-context.

Picking the right data leader

The right data team can make or break your data governance. A clear line of command and ownership of data is crucial. Therefore, a data leader or Chief Data Officer (CDO) is a requisite. Without somebody in charge, the people in your organisation won't know who is accountable for data, its use and processes, and your governance could easily fall by the wayside.

CDOs can command six-figure salaries, putting them out of the reach of many organisations. However, there are an increasing number of data leaders choosing to work on a freelance or consultant basis, available to businesses to consult on data strategy and governance.

Begin your team by putting strong data leadership in place first. Then, consider who requires access to your data, in what format and what their skill and knowledge level is. This will inform what tools are needed and influence some architecture decisions. If a skills gap is identified, then there are a number of options including hiring a full-time employee, contracting a freelancer or getting third-party help.

Security of data

Data security has come to the forefront of many consumers' minds due to several high-profile data breaches. T-Mobile[8], Superdrug[9], and Butlins[10] are some of the most recent leaks, with over 2 million customers affected to date.

Any business that uses and stores data puts it at risk of misuse. The value of data is increasingly recognised and it's an attractive target for hackers, as seen in the Superdrug breach where hackers contacted the company to hold it to ransom.

There are additional risks in mishandling data, where employees fail to follow governance processes and potentially leak sensitive information. West Ham football club recently leaked 200 email addresses when employees mistakenly sent a round-robin email to ticket holders[11].

When a breach does occur, there's not just the financial penalty involved for the organisation (now a much heftier 20 million under GDPR) but also a loss of reputation. It can take years to rebuild that trust and reputation. Data breaches make headlines, meaning you could find your organisation in the news for all the wrong reasons - over and over again. Many large breaches are repeatedly brought-up when another leak occurs, such as the Dixons Carphone breach[12].

The process of securing data

Effective data security is not difficult to achieve. There are a few key areas that have to be decided on.

01 **Who is responsible?** Assigning ownership of data to an individual or team is vital. Under GDPR, depending on where your data originates from, you're either a data controller or processor. If your organisation is the data controller, then any companies and third parties who use your company data are data processors and you're ultimately responsible for managing consent and access to that data (and communicating it to the data processors).

It's important that everyone understands who to report to, especially if a hack or breach does occur when time is of the essence. Everyone has a responsibility towards the data that they use, however, there should be a single point of contact and escalation. Usually, in large organisations, this falls to a dedicated cybersecurity team. However, in smaller companies it's often the remit of IT. If this is the case, then it's worth investing in training for team members who might just be getting by with their security knowledge - data is too valuable to simply 'muddle through' with your data security.

02 **Start securing data:** Your data security plan outlines who has access to data, how people can request and receive access, physical security, encryption of personal data, and regular penetration testing and auditing.

03 **Pick your tools:** For smaller organisations there are many effective security tools that have out-of-the-box functionality to get you started. You can turn different functions on and off depending on your needs, such as masking, encryption and role-level permissions.

04 **Personally identifiable information and sensitive data:** Generally speaking, your data ingestion should remove any personally identifiable information (PII) before analysis as it's rarely useful for this purpose. Sensitive data obviously needs to be more protected than other data sources (such as day-to-day operational data). It's worth ranking your data by its sensitivity as you don't want to waste resources over-protecting data that doesn't need it, or risk a leak of highly sensitive information because you're not protecting it enough. One size does not fit all. Also, consider data that is not sensitive when separate, but becomes sensitive when combined. An example of this is someone's postcode and date of birth, that could lead them to be identified when combined.

05 **User experience:** Another consideration is the need for data security to work behind-the-scenes and not impact daily operations. It cannot hinder employees' work.

06 **Keep backing up data:** Having regular weekly (or daily) data backups is important as it gives you an opportunity to audit your data and will also come in useful if a breach does occur.

Notable recent data breaches

T-Mobile[8]: Over 2 million customer records were accessed, and details such as their names, post/zip codes, account numbers, email addresses and encrypted passwords were accessed.

Butlins[9]: 34,000 guest details were accessed in this hack, with addresses and contact details included. Payment data wasn't compromised but the company still recommended caution and watching for any fraudulent activity.

Superdrug[10]: Up to 20,000 customers could have been affected by this hack where a hacker actually contacted the company to hold its customer data to ransom. However, some reports have suggested this may, in fact, have been an attempt at credential stuffing, where details from a previous hack are used to access accounts and extort money in the future[12a].

West Ham[11]: A relatively small breach with 200 people affected, however, it is notable because of the way the breach occurred. It was not a hack but employee error that led season ticket holders to be cc'd into a mass email. The breach is also one of the first to occur post-GDPR enforcement and it will be interesting to see whether the football club receives a fine for the breach and how much it is.

Dixons Carphone[12]: One of the largest in recent years, 10 million customers are reported to be impacted by this breach, although the company originally estimated that 1.2 million were affected. Personal information, names, addresses and email addresses were taken along with the records of 5.9 million credit/debit cards (but these were protected from fraud by the chip and pin system). Shares fell more than 6% following the breach report, and the company's troubles don't end there: it's now facing a £400 million fine as the Information Commissioner's Office (ICO) determines whether the breach should be considered under GDPR - despite it happening well before the GDPR enforcement date[12b].

Air Canada[12c]: A security flaw in the aviation company's mobile app caused a breach of 20,000 flyer details, including highly-sensitive passport numbers. The Canadian Government states that the risk of someone filing for a new passport with the details is low, but is urging caution. Cyber-experts have criticised Air Canada's password policy, stating its eight character limit is outdated and insecure.

Employee responsibilities on privacy

Everyone has a responsibility towards using data securely. Ensuring the privacy of data - especially sensitive information - needs to be part-and-parcel of every role. A privacy policy that outlines company expectations (such as always locking computer screens when away from a desk and resetting passwords regularly) is a good start. Privacy best practice can then be reinforced with regular company meetings, lunch-and-learns, or workshops.

This is especially important now that GDPR is enforced. For many employees, it might be seen as a bit of a tick-box exercise. It is vital that you help them understand the high stakes involved - for themselves and the organisation.

How to deal with a data crisis

Even the best-laid plans can go wrong and you can never completely rule out a data breach. Nothing is stopping a rogue employee from taking a photo of a screen full of sensitive information and releasing it. This is where crisis management comes in.

Data breaches are ranked as one of the top three worst events for a brand's reputation[13]. Damage to reputation can be mitigated with a well-thought-out and quickly executed crisis management plan.

01 Consider the impact of a breach

First, you need to identify potential scenarios and determine the risk of each occurring. After this, consider the consequences of this

happening - will it lead to a loss of trust? Is there a monetary impact? How badly will it impact your bottom-line?

In comparison to West Ham's email data leak, in 2016 Dean Street clinic was found to have leaked 700 emails addresses in a similar round-robin style email newsletter[14]. Unfortunately, the details related to users of a confidential HIV service. Therefore, the fallout from the Dean Street leak was potentially higher than the West Ham one, because users lost trust in the anonymity and confidentiality of a highly sensitive service.

Other areas to look at include:

- Whether your organisation collects personal data, such as names and contact details.
- If your organisation collects credit/debit card details and if you follow best practice and all regulatory requirements for this.
- The security of your website and email provider. Also considering the strength of any passwords, how these are distributed and to who, and if these could be written down somewhere.
- Whether your employees bring in their own devices (I.E. smartphones and tablets) to use at work and how secure these are. A bring-your-own-device policy that details security and use protocol is a good idea in this case.

02 Who to notify in a crisis

For each potential scenario identified, create a communications plan. This will set out the type of messages that will need to be sent, who receives them and via what channel. Usually, email is a quick way to notify users, and social media can increase awareness further. In the recent Superdrug hack, it notified affected customers via email, released a statement through social media and began a press campaign as well.

Relevant authorities, such as the ICO, will have to be notified - usually within a certain timeframe (under 72 hours for the ICO).

It's also worth identifying potential spokespeople beforehand. Ideally, someone who is specifically trained in media relations and crisis communications. In the high-stress, high risk scenario that a hack produces, this training will be critical.

Buffer - crisis comms done right

When hacked in 2013, social media company Buffer[15] offered a good example of how to manage a crisis situation. On discovering the hack, which posted spam from social media accounts linked up to Buffer, it quickly sprung into action with its crisis communications plan.

First, Buffer's CEO immediately posted an apology and update to social media, where most of its customers were likely to be most active. It also sent communications to its community and support groups. Then, a blog was posted with details of how they were fixing the situation. Any users on the Buffer platform at that point also received instant notifications of what was happening.

It's this level of communication and care that helped Buffer come out of the situation relatively unscathed. In fact, many people took to social media to praise its response.

A data crisis plan must be communicated company-wide, and include information on who to report to, the lines of communication and escalation points. Plus, details on what to do in out-of-office hours.

It's a good idea to regularly test and run-through your crisis plan, so that if the worst does occur, it will run like clockwork.

03 Prioritise prevention first

Prevention will always be a better route than having to respond to a data crisis. Therefore, ensure your data governance procedures follow best practice and are regularly tested and reviewed. Make sure everyone realises their role in protecting data. It cannot just be left to a cybersecurity or IT team.

Challenges of legacy systems

Meeting modern day needs with yesterday's architecture

For some organisations the challenge of keeping up-to-date with new data management technology, whilst effectively integrating it with legacy systems, is all too familiar. Plus, today's modern architecture is tomorrow's legacy system, meaning at some point every organisation is likely going to combine the old and the new.

When this situation occurs, return to your use cases and planned projects. Use this to inform which systems you can keep and what technology to get rid of in order to succeed with your data strategy. Once in place, it's also important to constantly review your tech stack to determine how well it is performing and meeting your organisation's needs.

Ensure that all decisions, and any new processes or technologies implemented are well-documented. This will help you with future integrations. Unpicking legacy architecture with little-to-no documentation is a long-winded and frustrating process.

Future-proofing Tungsten's Data Management

TUNGSTEN NETWORK

Tungsten specialises in invoicing, financing, invoice flow and procurement. It suffered from a legacy architecture that prevented it from keeping pace with current and future data plans.

Cynozure was tasked with untangling a complicated data architecture where none of its original developers worked for the company any more. This had been built in modules, with no oversight on how the architecture fully works and what processes were driving it. Tungsten needed to adapt the system to fulfil future goals that included improving cross-selling opportunities, growing revenue, and establishing new product lines. It previously tried fixing the architecture in-house but brought Cynozure on-board to help create a conceptual data architecture and standardised terminology across the company.

Cynozure began by developing a strategy exercise that was heavily focussed on creating a data architecture that would set up Tungsten in the future. Several workshops were held with Tungsten employees to better understand how the business worked. This was used to create a data model that covered all the different use cases within the business. From this, an 'ideal' data strategy was developed that took into account how the business operated, how data could support this, without being constrained by the existing architecture.

This model was then imported into Salesforce, where Cynozure also assisted with the technical implementation. In this way, Tungsten benefits from an ideal data model formed in Salesforce, but with its legacy architecture.

Then, there was a need for clear documentation regarding the architecture. An internal wiki was created to help Tungsten's in-house team understand the different processes behind the system. New employees can also access this wiki to get fully up-to-speed with the system.

Now, Tungsten has a fully documented data architecture that will help it make and achieve future data plans. Development of the internal wiki prevents knowledge being lost to the company when a developer leaves.

"Cynozure played a critical role in helping us understand Tungsten's data architecture and setting it up for the future, along with changing the wider data culture. There's been an added benefit in being able to cross-sell and upsell to more of our clients, as well as target parent organisations. There's greater knowledge of Tungsten's system, consistent terminology, and everything runs much more smoothly now."

- David Williams, CFO at Tungsten

Factory vs lab environment

When setting-up or refreshing your architecture, it's worth considering the different needs of your lab and factory environments.

The lab environment is the place for your data team to experiment with different models, coding languages and data. It's where innovation often happens and is, therefore, more flexible compared to the factory. Labs are useful as they provide a single place for people to experiment, with clear data sets, as opposed to everyone working on different projects with no clear oversight on what's happening. Once a use case is proved in the lab (and approved by senior stakeholders) it can then move to the factory.

The factory has more structure, with key processes to follow and a standard language used. A project that has been moved from the lab to the factory will possibly need to be re-coded in the standard language and tweaked to fit its processes.

Changing for the intelligence age

There are many modern-day use cases that require traditional data management (like batch processing) to evolve. Organisations that intend to use real-time data, for example, require technology that can stream and analyse on-the-fly. However, the foundations of good data management and governance still ring true. Make sure your basic data architecture is up to scratch before embarking on more complex use cases.

Real-time data needs really good data management

Real-time data is gaining popularity, and may even become the standard form of analysis in the future. Therefore, it pays to prepare for real-time data when your use cases call for it.

Real-time data is becoming more common because of the IoT. Machine learning and cloud technology have also encouraged its adoption. Gaming company Tapjoy uses real-time analytics to determine the games that players are most likely to buy and then serve adverts for those games. It processes 300,000 requests per minute with a response time of less than 10 milliseconds[16].

Relying on lagging metrics is a common problem for many businesses when using data, hence the rising popularity of real-time. Forecasting based on historical data isn't as accurate as using real-time data.

However, real-time analytics requires a different approach to data management. Time is of the essence when dealing with real-time data, therefore everything from your infrastructure to your skills needs to be fine-tuned. Legacy systems are unable to store a process real-time data which is in a stream format. Extract, Transform and Load (ETL) is designed to run by batch processes, which isn't compatible with real-time data analysis. By the time real-time data reaches a data warehouse it has lost its usefulness.

Instead, two types of data architecture are suited for real-time data.

Lambda uses a hybrid model that analyses historical data and newly arrived data at the same time. The architecture involves two layers in order to process the two types of data – batch and speed. This allows

it to process both, but it is slightly more complex, less responsive, and requires more maintenance compared to Kappa.

Kappa takes in streams of data from feeds such as social media and the stock market. This data is placed in temporary storage where it is then ingested into an engine like Apache Spark or Flink. Each of these engines are slightly different and work for different use cases.

The technology is still largely in its infancy, and skills to deal with real-time are a rarity. More than half of developers are still learning how to use the tools involved[17]. For now, a hybrid approach (where you use traditional batch processing for some analysis and real-time when the use case requires it) is likely the best use of your resources. Before exploring real-time, ensure the rest of your data management is in place and running smoothly.

The rise of IoT and what this means for data management

The IoT has captured the imagination of many business leaders but has progressed little beyond that. Part of the issue with implementing IoT-based use cases is that organisations don't really understand how to go about it. Using IoT data required gold standard data governance and management, but many organisations remain stilted by legacy systems.

But every business is set to benefit from the IoT. From manufacturers monitoring and optimising their supply chains, to retailers measuring footfall in-store. Indeed, by 2020 it's expected that over half of business processes will involve the IoT[18]. The success of any IoT project, however, lies in the data management behind it.

Managing the IoT

First, any data collected through the IoT is subject to the same requirements as other data. That is, it needs to be of good quality, trusted, secure, and in the right format. To obtain the most value from the IoT it's worth integrating it with other data sources. Of course, this can pose problems when trying to integrate it with a legacy system.

IoT data is real-time. That means you need a system that can continuously collect, stream and analyse it.

ETL vs ELT

The traditional process of ETL has recently been giving way to Extract, Load and Transform (ELT). The difference between the two is where and when data is transformed. Plus, with ETL access to the information is limited to when the process has completed.

With ELT, as soon as you have your data you immediately begin moving it into a central data store. Thanks to developments in cloud computing there's the ability to store vast quantities of data and process it rapidly. So ELT can offer a modern alternative to ETL, with greater access for developers and insights provided in a shorter timeframe.

However, it is still evolving so the frameworks and processes involved in its management aren't yet fully developed.

Therefore, a streaming-first architecture (where you get insights from data as soon as it is created) is needed. In order to integrate it with existing data in your warehouse, you'll also have to transform this into data streams.

With IoT data the stakes are a lot higher if a system fails. Everything from your crisis response and management to your technical support feeds into this. Because IoT is real-time, it will need constant 24/7 support. Unlike other data sources you cannot afford an issue for hours.

Data ownership is key. 82% of security leaders state that there's no clear ownership over their IoT data[19]. This underpins all effective data governance. Before using the IoT, ensure there are clear lines of accountability and that everyone understands who has responsibility for the data.

Finally, consider whether you require IoT data in the first place. There is a cost and time investment involved in implementing it. Therefore, any use needs to link back to your data strategy and business goals. What do you ultimately wish to achieve and how does IoT data help with this?

For example, a clear use case for the IoT in agriculture is in increasing yields and lowering production costs associated with farming (such as Intel using the IoT to feed dairy cows the optimum amount and mix of food on 3,000 farms[20]).

Smart cities and smart data management

Smart cities are urban areas that use information technology to share information with citizens, increase operational efficiency and improve government services and public welfare[21]. In order to be successful, they require businesses to be able to share data. However, because many businesses suffer from legacy infrastructure and processes, smart city development is being held back.

For smart cities to work, there is a crucial step in first learning how to manage and store IoT and other real-time data sources, and to share it securely with others who require it. This requires establishing a common format and governance principles across the city and relevant businesses. Data would have to be collected, transferred and analysed in real-time and APIs would need to be set up for data sharing.

An example of this would be a waste removal company connecting with IoT-enabled rubbish bins across the city to better understand what bins to prioritise. It could also use smart traffic management to optimise its collection routes and to feed data back into the city of any obstacles it encounters along the way.

Data management and AI

How to make your data management AI-ready

AI is an attractive proposition for many businesses. So much so, that the market for AI technology is set to increase from $7.35 billion in 2018 to $89.85 billion by 2025[22]. 20% of organisations are using AI technology at scale, with a further 41% experimenting with it[23]. The time is ripe for future-thinking organisations to learn about the potential applications of AI.

It's crucial to lay the groundwork and prepare for AI's adoption into your organisation. AI is heavily reliant on data. So the data has to be trustworthy and the only way to achieve this is through strong data governance. The AI needs timely access to data when required, in the right format, and to log any alterations and uses of that data. Again, clear ownership over the data, AI models and governance processes is key. If you cannot trust the data that an AI model uses, then you cannot trust its outputs.

What is AI?

For the purposes of our current discussion, we're talking about two forms of AI: machine learning and automation. Strictly speaking, automation doesn't always involve AI. It uses software that follows certain rules in order to execute tasks. Usually, these are repetitive and monotonous. Automating certain activities allows people to work on more strategic tasks that often provide higher returns.

Machine learning is a subset of AI where a machine is trained using a specific data set to perform a certain task. This is limited to whatever data it is trained on. An image recognition algorithm cannot be used for voice recognition and vice versa. Importantly, there's a feedback system in machine learning that allows the machine to constantly refine its approach until it has mastered the task.

Automation can be combined with machine learning to get smarter at performing its tasks.

There are two other types of AI. True AI has not been mastered yet, but it aims to mimic human intelligence. It will be able to achieve lots of different activities, including speech and voice recognition, constructing sentences and speaking back, comprehending its environment and predicting danger.

Deep learning is based on the human brain's network of neuron cells. It uses several algorithms to process complex tasks - these are artificial neural networks. Deep learning offers a middle ground between machine learning and true AI.

How AI can help solve your data management challenges

In a somewhat virtuous circle, data management is needed to support AI, but AI can also help solve some of data management's challenges.

Sorting large quantities of data: The amount of data each organisation has to store and manage is growing rapidly. AI can help by analysing data sets to determine the best way to store it, where, and whether it requires further processing.

Cleaning data: Automation can also cleanse and process data to be in the right format. This can significantly cut down on the amount of preparation that your data team has to do. AI can scale as the amount of data grows, which is more efficient than hiring extra data engineers to work on your data.

Highlighting sensitive data: AI can be trained to recognise certain data formats, such as email addresses, and then bring them to your attention. This helps you prioritise the security of different datasets and ensures that nothing slips through the net.

Actively monitoring for threats: AI can constantly check for any threats or unusual activity around your data. This level of oversight is difficult to achieve manually with large datasets. 58% of IT leaders surveyed expect AI to assist with this, especially given that 48% found their current alert systems to be too noisy and/or high[24].

Finding data: When faced with a vast store of data, manually finding the information that you need can be impossible. Machine learning can help uncover data sets and flag them for your attention.

For example, IBM's Watson has been helping medical researchers sift through thousands of past medical papers that would be impossible to read manually. It spots patterns to help researchers find cures for cancer, as well as other diseases[25].

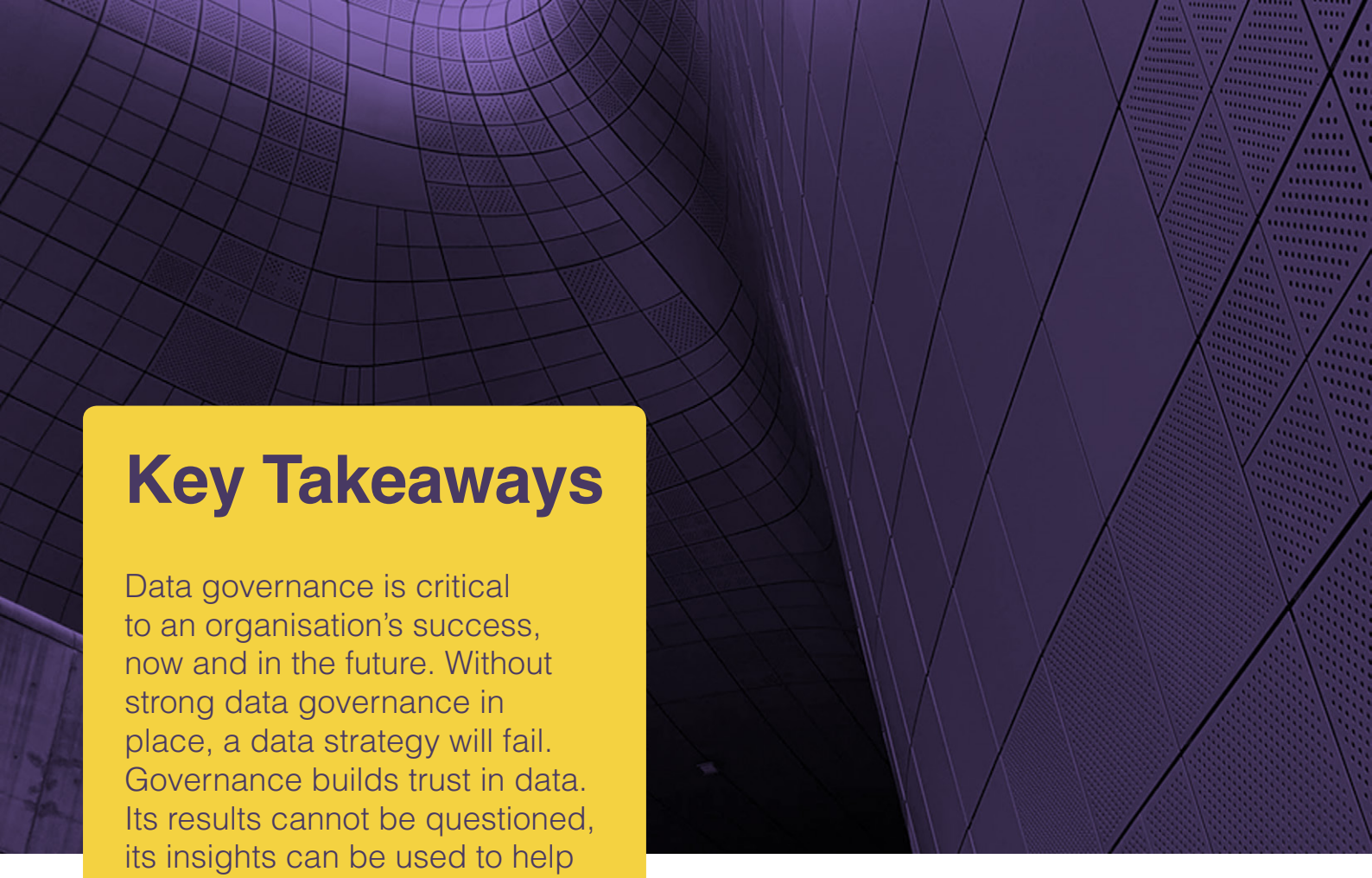
Automated data integration: AI-powered tools to load and ingest data are becoming more common. AI can recommend rules, data transformations, or the next best action after data integration.

The Future of Data Management

The future will be data-driven. Good data governance will be a competitive advantage, setting organisations apart as consumers become more aware of how their data is used and managed.

The amount of data will exponentially increase as well, thanks to increased adoption of the IoT, as well as emerging technology such as autonomous vehicles reaching maturity. Therefore, the pressure on data management to sort, store and secure it will also rise. Because of this, organisations will explore alternative forms of data management that are best suited to their use cases - relational versus non-relational databases, for example. Master data management will also become more popular as a way of creating a single point for all business-critical data, increasing efficiency and reducing errors and duplication.

5G is set to come into the UK by 2020 and this will vastly increase the amount of data that can be streamed via the cloud. Meaning off-site solutions will likely become more common, as will real-time streaming[26]. Data-as-a-service will likely become more popular as well. We may also see consumers begin selling their data in personal data marketplaces, such as Datacoup[27], as they begin to realise its value.



Key Takeaways

Data governance is critical to an organisation's success, now and in the future. Without strong data governance in place, a data strategy will fail. Governance builds trust in data. Its results cannot be questioned, its insights can be used to help transform an organisation's culture into a data-driven one. As technology such as AI becomes more popular, data governance and management is the essential step to get right in order to progress effectively.

01 This involves a commitment to the quality and ownership of data.

02 Data sources must be identified and aligned with use cases.

03 This will inform its storage. Most organisations will benefit from a data platform that allows timely access to data, in the right format, to those who need it (and nobody else).

04 Data ingestion processes are key. Create a framework that can adapt as needed. Audit data when needed and ingest it in its raw format.

05 The security of data has to be efficient. To achieve this, prioritise what data does and doesn't need a high level of protection. Additionally, we often see companies caught out by data that becomes sensitive when combined with other sources.

Ultimately, this requires strong data leadership and clear lines of ownership. So that everyone understands who to turn to in order to request access and to report an issue or breach. If you don't have a data leader or CDO in place, then now is the time to get one. Data governance doesn't work without a leader.

Organisations are realising the need to use data to future proof their operations. However, many are missing the vital step of governing that data well. This is a mistake, especially since data breaches are becoming more common and high-profile. Consumers now have the right to remove access to their data in organisations they do not trust. A data breach will not just impact an organisation immediately, it will have long-lasting effects.

Build strong data management = build trust

Trust is really the defining factor. We are entering the age of trust. With fake news, the public has lost trust in the media. Post-Cambridge Analytica, the public trust in ethical data use is shaken. Consumers are wary of emerging technology, AI and self-driving cars included. It's up to organisations to build trust with the public. To let people know that their data is safe, it is managed well and will be used for their benefit. The only way to achieve this is through a good data management strategy.

References

1. The Guardian “Timeline of trouble: how the TSB IT meltdown unfolded” https://www.theguardian.com/business/2018/jun/06/timeline-of-trouble-how-the-tsb-it-meltdown-unfolded

   1a. Fujitsu “Master Data Management” https://www.fujitsu.com/caribbean/Images/Master-Data-Management-casestudy.pdf

2. Statista “Big data - Statistics & Facts” https://www.statista.com/topics/1464/big-data/

3. Statista “Data volume of global consumer web usage, e-mails and data traffic from 2016 to 2021 (in petabytes per month)” https://www.statista.com/statistics/267181/forecast-of-consumer-internet-traffic-through-e-mail-and-web-usage/

4. ITPRO “What are petabytes and just how big are they?” http://www.itpro.co.uk/storage/30251/what-are-petabytes-and-just-how-big-are-they

   4a. The Guardian “Google fined record £44m by French data protection watchdog” https://www.theguardian.com/technology/2019/jan/21/google-fined-record-44m-by-french-data-protection-watchdog

5. The Verge “After Facebook hearing, senators roll out new bill restraining online data use” https://www.theverge.com/2018/4/10/17221046/facebook-data-consent-act-privacy-bill-markey-blumenthal

6. European Commission “The Digital Skills Gap in Europe” https://ec.europa.eu/digital-single-market/en/news/digital-skills-gap-europe

7. Upwork “New report finds majority of companies are embracing remote teams, yet more than half lack a remote work policy” https://www.upwork.com/press/2018/02/28/future-workforce-report-2018/

8. CNET “T-Mobile hack may have exposed data of 2 million customers” https://www.cnet.com/news/t-mobile-hack-may-have-exposed-2-million-customers-data/

9. Independent “Superdrug hack: Data thieves claim to have information on 20,000 customers” https://www.independent.co.uk/news/uk/home-news/superdrug-hacked-data-information-20000-customer-phone-number-points-date-birth-a8501871.html

10. The Guardian “Butlin’s data hack: up to 34,000 guest details may have been stolen” https://www.theguardian.com/technology/2018/aug/10/butlins-data-hack-guest-details-stolen

11. The Telegraph “Exclusive: West Ham could face investigation after sharing personal data of up to 200 season ticket holders in email error” https://www.telegraph.co.uk/football/2018/08/23/exclusive-west-ham-could-face-investigation-sharing-personal/

12. BBC “Dixons Carphone admits huge data breach” https://www.bbc.co.uk/news/business-45016906

    12a: Computerworld “The curious case of the Superdrug ‘hack’” https://www.computerworlduk.com/security/curious-case-of-superdrug-hack-3682719/

    12b: IT Governance “Dixons Carphone faces £400 million fine following biggest online data breach in UK history” https://www.itgovernance.co.uk/blog/dixons-carphone-faces-400m-fine-following-biggest-online-data-breach-in-uk-history/

    12c: CBC “Air Canada mobile app breach affects 20,000 people” https://www.cbc.ca/news/business/air-canada-mobile-app-1.4802879

13. Raconteur “Data breaches: don’t make a catastrophe out of a crisis” https://www.raconteur.net/risk-management/data-breaches-dont-make-a-catastrophe-out-of-a-crisis

14. BBC “NHS trust fined for 56 Dean Street HIV status leak” https://www.bbc.co.uk/news/technology-36247186

15. Buffer “Buffer security breach has been resolved – here is what you need to know” https://open.buffer.com/buffer-has-been-hacked-here-is-whats-going-on/

16. O’Reilly “Hot data meets big data to make real-time, real-world decisions” https://www.oreilly.com/ideas/hot-data-meets-big-data-to-make-real-time-real-world-decisions

17. TechRepublic “Enterprises still slow to embrace real-time data, survey finds” http://www.techrepublic.com/article/enterprises-still-slow-to-embrace-real-time-data-survey-finds/

18. Gartner “AI Newsroom” https://www.gartner.com/newsroom/id/3185623

19. Forescout “Enterprises risk all in massive IoT timebomb” https://www.forescout.com/company/news/press-release/enterprises-risk-massive-iot-ot-security-compliance-time-bomb/

20. Intel “IoT Farming” https://www.intel.co.uk/content/www/uk/en/it-managers/smart-farming-iot.html

21. https://internetofthingsagenda.techtarget.com/definition/smart-city

22. Statista “Worldwide AI market revenues” https://www.statista.com/statistics/607716/worldwide-artificial-intelligence-market-revenues/

23. Harvard Business Review “A Survey of 3,000 Executives Reveals How Businesses Succeed with AI” https://hbr.org/2017/08/a-survey-of-3000-executives-reveals-how-businesses-succeed-with-ai

24. ZDNet “Most managers want IT operations managed by artificial intelligence” https://www.zdnet.com/article/enter-aiops-artificial-intelligence-guiding-it-operations/

25. IBM “Watson in healthcare” https://www-05.ibm.com/innovation/uk/watson/watson_in_healthcare.shtml

26. 5G “What is 5g?” https://5g.co.uk/guides/what-is-5g/

27. Datacoup http://datacoup.com/

cynozure

A data and analytics strategy consultancy, Cynozure is on a mission to change the way business is done through positive use of data.

In collaboration with forward-thinking organisations, governments, and individuals Cynozure advises - and delivers on - all aspects of data and analytics strategies. This is achieved through advisory services, coaching and mentoring, solution and organisational design, technology implementation, business change programmes, and on-going support services.

Cynozure's team and associates are thought leaders and experts in this space. Many have a background in industry, and frontline experience of what is required to create leading data-driven organisations. Now they have a united goal: equip leaders and their organisations with the ability to understand and leverage their data. Cynozure will help identify the value that exists within data, and how it can be used to transform business strategy, products, services and operations. There is a clear focus on ensuring that incredible business (and social) value is delivered, to maximise the transformational power of data across society.

Organisations that have benefited from Cynozure's approach include The National Trust, Soho House, Tokio Marine Kiln, MSD, The Really Useful Group, Camden Council, Lloyd Webber Theatres, Kondor and Tungsten Network.

cynozure.co.uk

@cynozure_uk

Cynozure

@cynozure_uk

hey@cynozure.co.uk