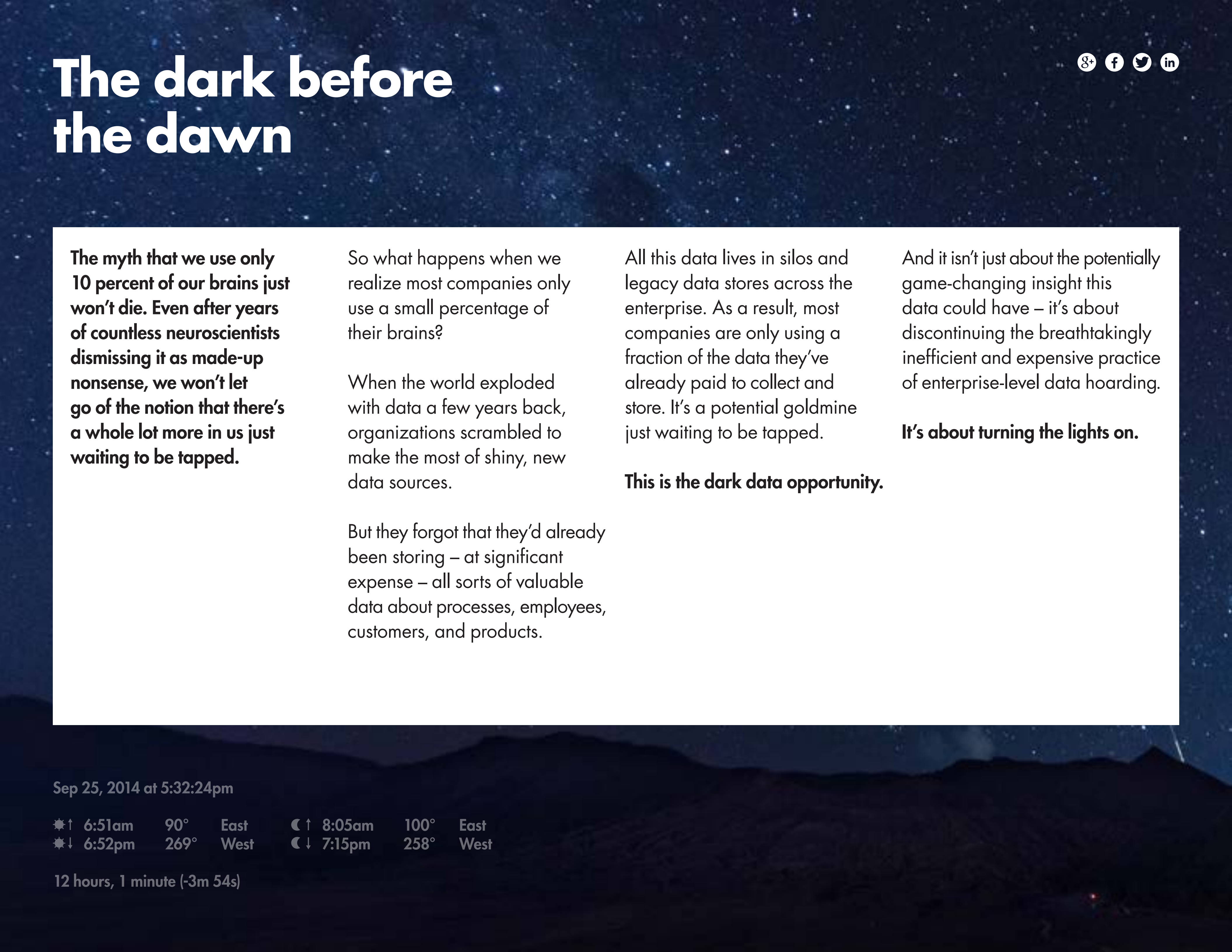
informatica

The Dark Data Imperative

You're sitting on
a data goldmine
– and you don't
even know it.



The dark before the dawn

The myth that we use only 10 percent of our brains just won’t die. Even after years of countless neuroscientists dismissing it as made-up nonsense, we won’t let go of the notion that there’s a whole lot more in us just waiting to be tapped.

So what happens when we realize most companies only use a small percentage of their brains?

When the world exploded with data a few years back, organizations scrambled to make the most of shiny, new data sources.

But they forgot that they’d already been storing – at significant expense – all sorts of valuable data about processes, employees, customers, and products.

All this data lives in silos and legacy data stores across the enterprise. As a result, most companies are only using a fraction of the data they’ve already paid to collect and store. It’s a potential goldmine just waiting to be tapped.

This is the dark data opportunity.

And it isn’t just about the potentially game-changing insight this data could have – it’s about discontinuing the breathtakingly inefficient and expensive practice of enterprise-level data hoarding.

It’s about turning the lights on.

Sep 25, 2014 at 5:32:24pm

☀↑ 6:51am	90°	East	☾↑ 8:05am	100°	East
☀↓ 6:52pm	269°	West	☾↓ 7:15pm	258°	West

12 hours, 1 minute (-3m 54s)

Defining dark data

Dark data refers to all the data that your organization has already paid for, collected, and stored in various systems and data stores, but isn't actually using, analyzing, or even accessing right now.

Machine logs, emails, application data, metadata, social data.

When you acquired the data or deployed the applications it came from.

The most outdated of which are probably still burning a hole in your bottom line because no one knows what data is actually on them.

The fact that you can't even access or profile this data means you have no idea how important or sensitive it is.

Defining dark data

It's an unsightly oversight that:

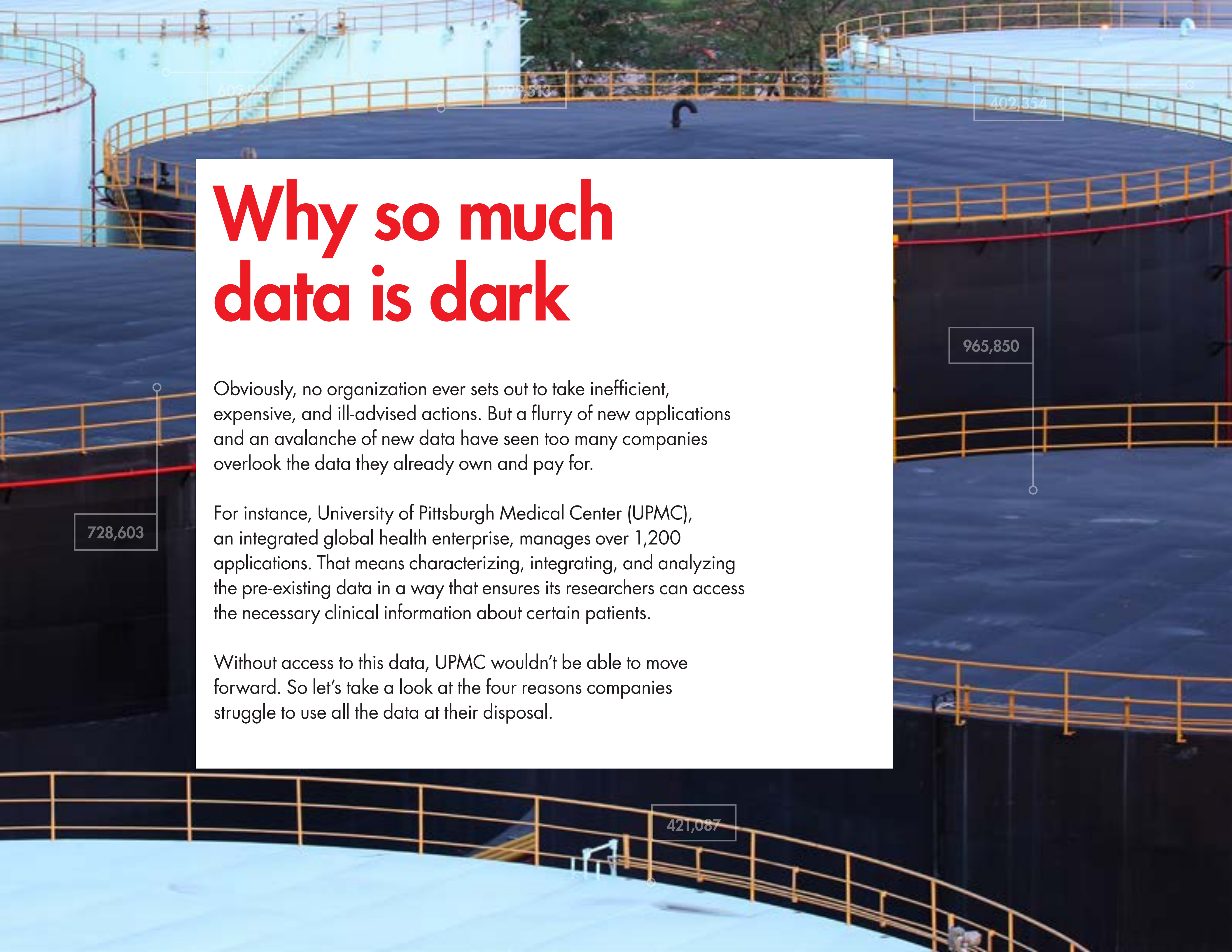
- Costs you an as-yet-undetermined amount more than it should
- Leaves you vulnerable to the mismanagement of sensitive data
- Prevents you from gaining a deeper understanding of your customers, your employees, your processes, and your products

A reliable view of the enterprise

When you can't actually use the data you're paying for, you end up without a coherent view of what your enterprise is doing as a whole. That means you can't analyze your processes, rationalize your resources, or find the data that represents proprietary advantage in your systems.

One of our customers, Interstate Batteries, a $1 billion battery marketer and distributor, found itself struggling to manage and forecast inventories. That meant the company could neither predictably meet the retail market demand nor optimize its pricing decisions.

That's just one example showing you can't optimize what you can't analyze, and you can't innovate when you don't know what works.



Why so much data is dark

Obviously, no organization ever sets out to take inefficient, expensive, and ill-advised actions. But a flurry of new applications and an avalanche of new data have seen too many companies overlook the data they already own and pay for.

For instance, University of Pittsburgh Medical Center (UPMC), an integrated global health enterprise, manages over 1,200 applications. That means characterizing, integrating, and analyzing the pre-existing data in a way that ensures its researchers can access the necessary clinical information about certain patients.

Without access to this data, UPMC wouldn't be able to move forward. So let's take a look at the four reasons companies struggle to use all the data at their disposal.

1 They don't even realize the data exists

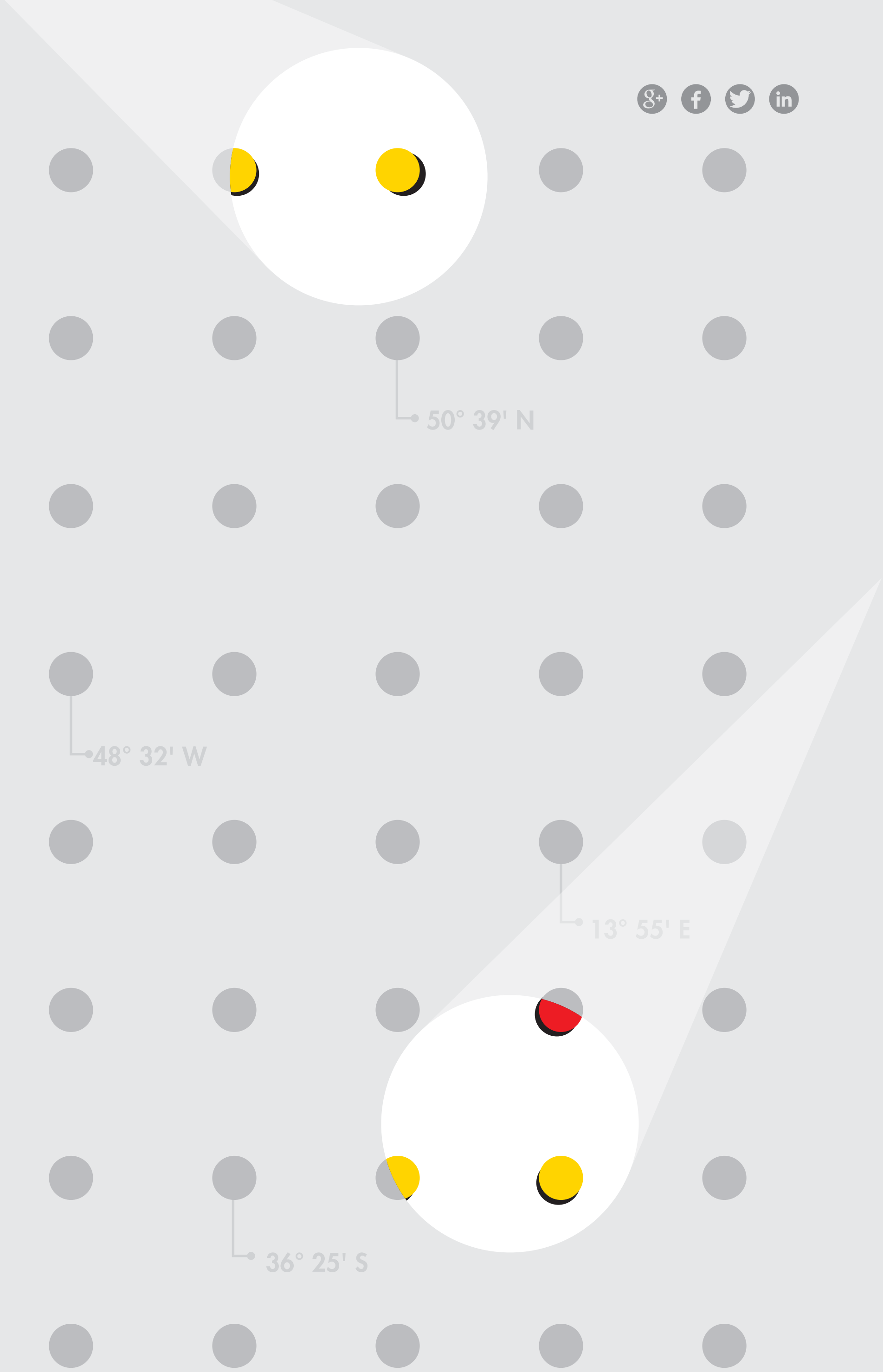
Too often, when managers and whole teams are attempting to answer difficult questions or improve the way they work, they avoid the challenge of seeking out and analyzing datasets they aren't familiar with. Unfortunately, it's common for a lack of skills, time, or bandwidth to preclude them from bringing the right data to light.

If this data were stored in a way that was accessible to the whole organization, more teams would be empowered to make better-informed decisions and test more hypotheses.

2 They realize it exists, but don’t know where

It’s particularly difficult to access your data when either your organizational structure or your data architecture gets in the way. When departments work in silos and data is kept in legacy data stores, even the most curious teams come up against an unnecessarily impenetrable brick wall.

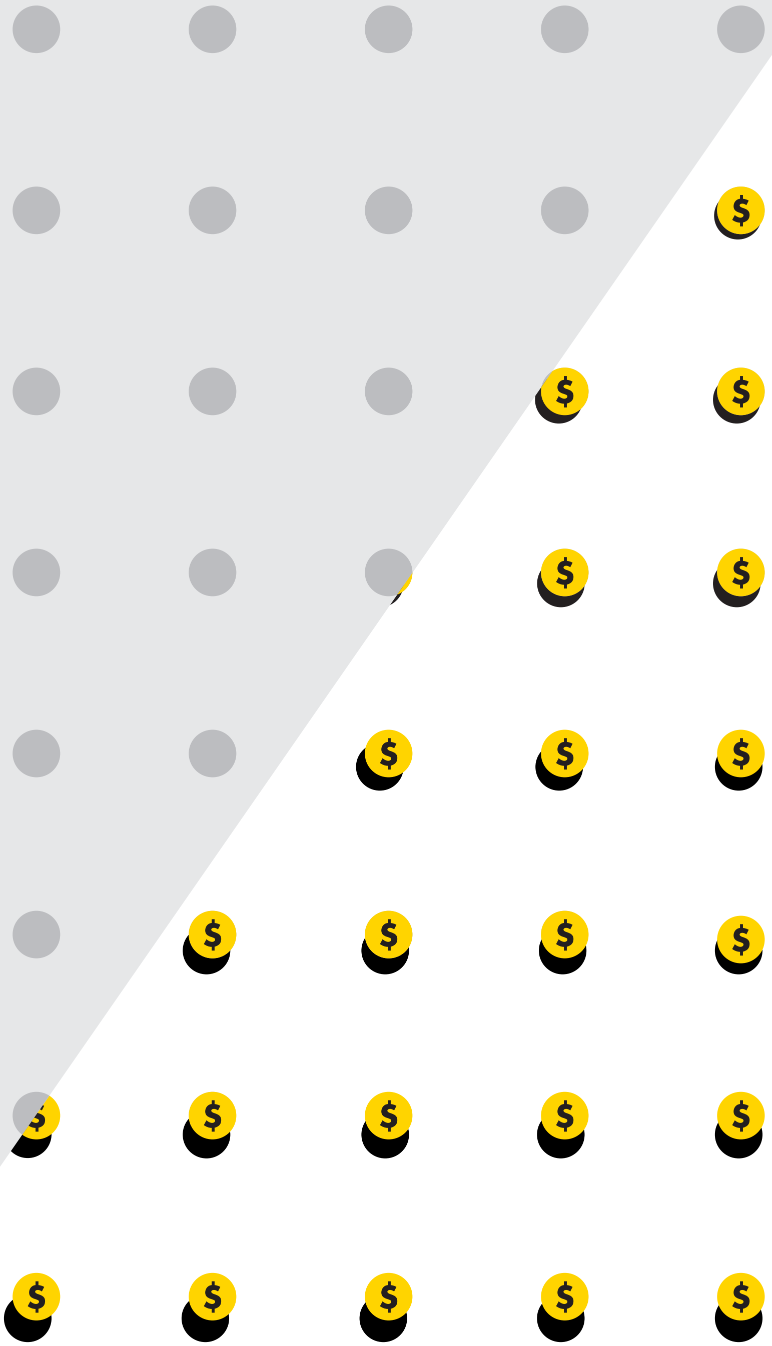
Without an enterprise-wide strategy to store and manage all this data, the quality of your organization’s decision making will remain hamstrung by internal politics and outdated technology.



3 It’s too expensive to actually use the data

Even when organizations discover the dark data they need, they usually have to confront a litany of costs associated with processing it on legacy systems (costs for additional processing units and operations). Even if they try to avoid these costs by replicating the data on cheaper hardware using new software frameworks like Hadoop, the initial costs associated with migrating processes and acquiring new skills are usually too high for a single project.

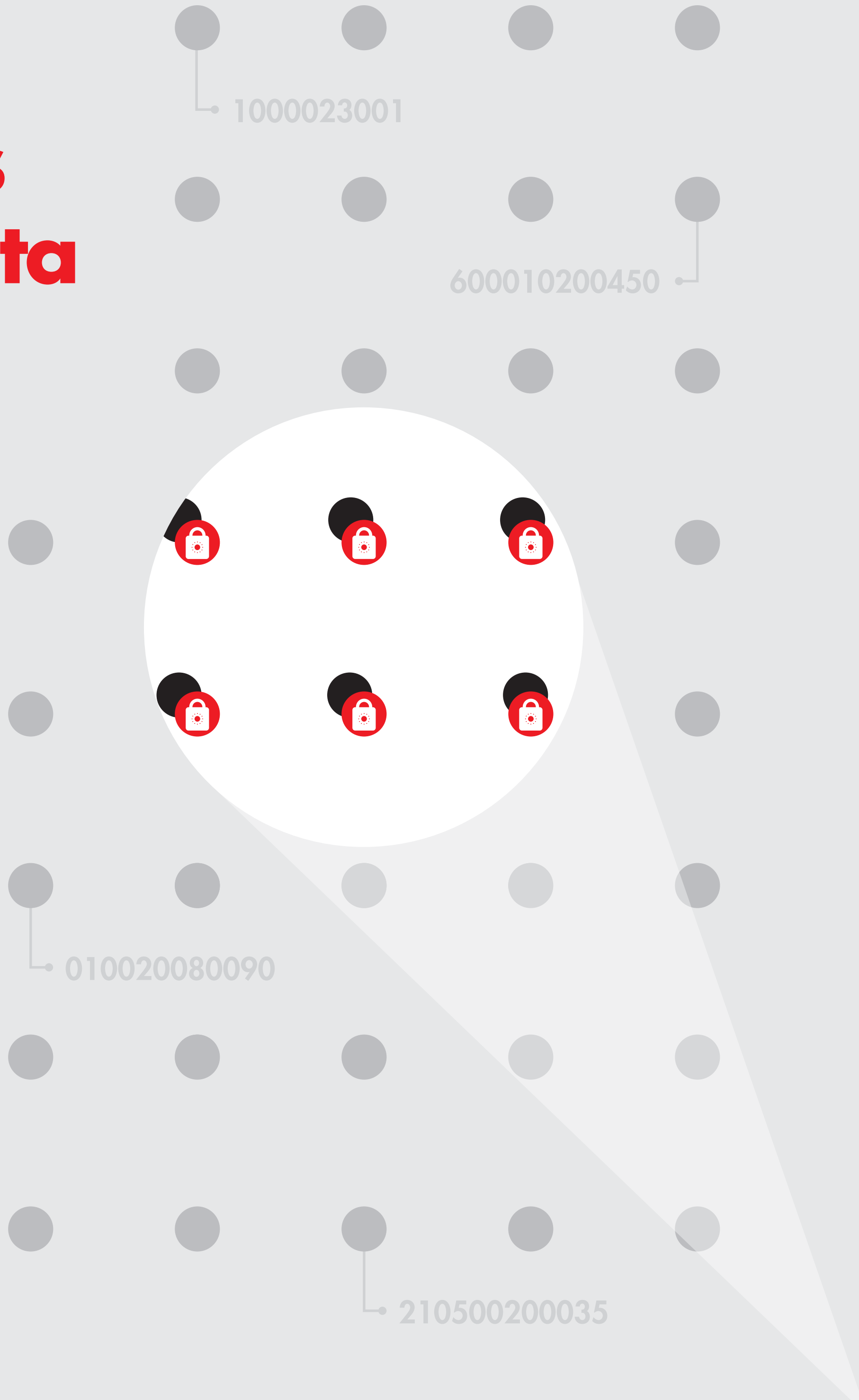
To make the most of the data you already own, the basic foundations for a more modern data architecture need to be put in place. Otherwise, you’ll continue to pay too much for data you can’t afford to analyze.

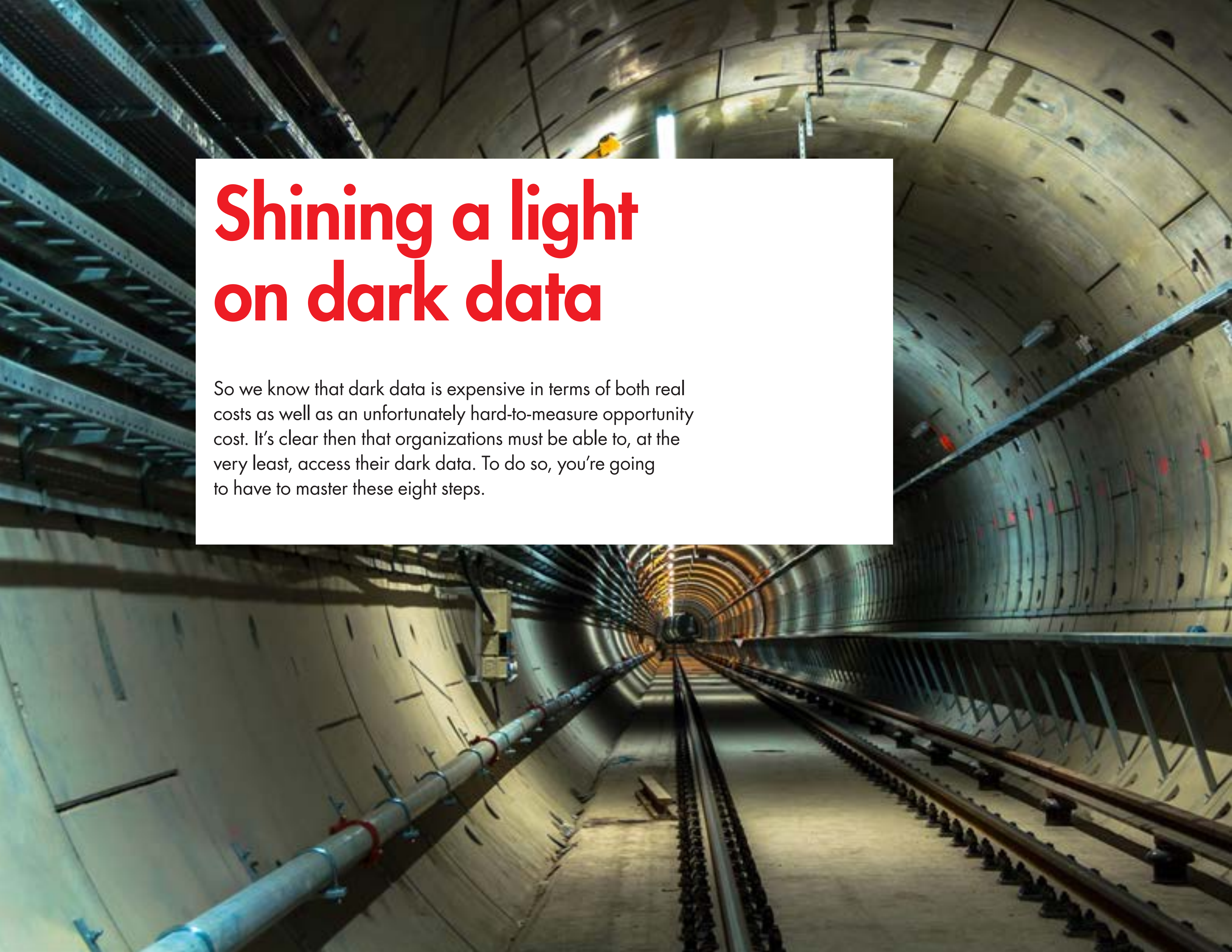


4 There are legitimate compliance concerns about some of the data

One of the most important reasons certain data isn't made available to analytics tools is the fear of security breaches and concerns about regulatory compliance with important privacy legislation (like HIPAA in the healthcare industry). As legitimate as these concerns are, it's hard to justify the cost of storing this data when you lack the ability to analyze it.

For your organization to make sense of the data it owns, clearly defined processes and tools need to be put in place to guarantee the security – and in some cases anonymity – of that data.





Shining a light on dark data

So we know that dark data is expensive in terms of both real costs as well as an unfortunately hard-to-measure opportunity cost. It's clear then that organizations must be able to, at the very least, access their dark data. To do so, you're going to have to master these eight steps.

Shining a light on dark data

Identify business benefits

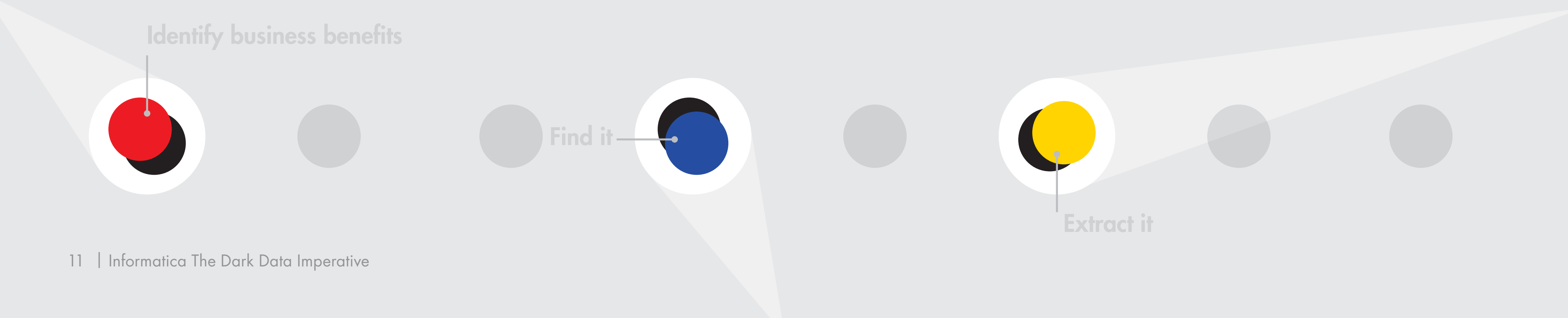
If you don't have a project that can bring innovation, optimize processes, reduce costs, or enable competitive advantage, you'll justifiably struggle to get the buy-in and budget you need. You'll need to be able to outline clearly and comprehensively the business benefits you'll gain from accessing dark data.

Find it

The best way to start your search is to tie it to a specific project and identify the data needed to achieve its goals. Work backward from that point and list the source systems, applications, and even departments that would offer or contain the datasets you need.

Extract it

The most important feature of big data is its variety. So make sure you don't underestimate the variety of the dark data you'll be attempting to extract. That means learning how to collect data in a real-time stream as efficiently as you extract the data from a database. Additionally, different legacy data stores will offer different extraction challenges.



Shining a light on dark data

Ingest it

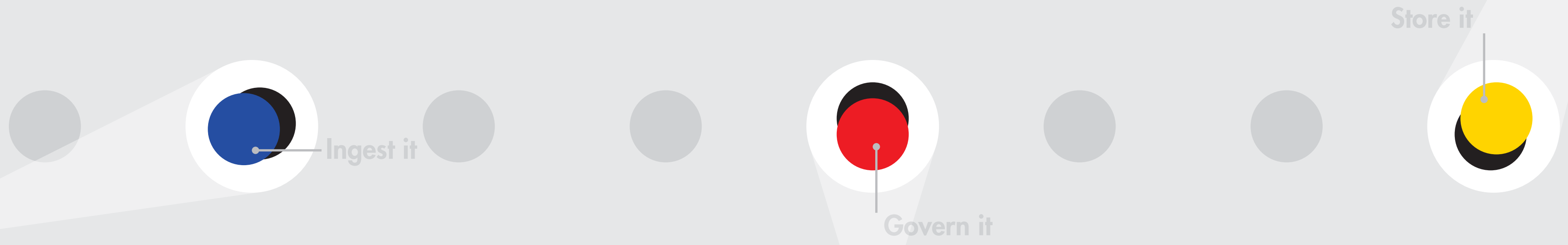
When you're pouring the data into your systems (sometimes after having transformed it), you need to be able to document what you've done and ensure you record as much metadata as you need. It's also important at this stage that you start establishing – and maintaining – standards for semantic consistency between different data formats.

Govern it

Rather than avoiding the chance to analyze your dark data at all, it's important that you commit to the mandates in the legislations with which you're trying to comply. Define the data quality needed for your project to really deliver. And then set up a policy of access control that can be enforced using encryption and masking technology to protect the data at the source. The aim is to manage the permissions you grant to different tools and business users.

Store it

Decentralized data storage is one of the main reasons so much data is dark to your enterprise. So you need to think long term and ensure you store all this data in a logical (centralized or federated), accessible, managed environment. The aim is to avoid unnecessary duplication and proliferation of data across systems. That way you can manage your storage costs and avoid the risk of a security breach. To that end, a managed data lake is your best bet. (More on this in a moment.)



Shining a light on dark data

Prepare it

There are many types of dark data hidden throughout your systems in a variety of structures and formats. These datasets will need to be reformatted, parsed, filtered, standardized, combined, and refined in ways that prepare them for input into your analytics tools and applications.

Deliver it

The final step you'll have to master pertains to the actual ease of use and accessibility of all your data. Having maintained an appropriate level of semantic consistency, security, and the required metadata about its lineage and characteristics, your business users should now be able to serve up this data to any analytic tool they choose. Based on the applications you're feeding and the nature of your project, you'll know if the data needs to be delivered in real time, in batches, or based on events.

In the case of UPMC, it was only after concerted effort and investment that researchers could electronically integrate – for the first time ever – clinical and genomic information on 140 patients who had previously been treated for breast cancer.

A repeatable process to harness dark data

The biggest mistake most companies make is to think they'll dip into their dark data only once. This is not just a one-off process. Your data is only going to multiply – in size, variety, and value. It will do so exponentially and continuously. Additionally, the number and type of applications to which you're going to deliver all that data will keep changing.

So instead of solving single dark data projects over and over, you should think about building a repeatable process. That means adopting the technology needed to build a modern infrastructure that can make all your data readily accessible and consistent so it's clean, safe, and connected.

This is why so many leading organizations are turning to a managed data lake solution.

A repeatable process to harness dark data

A managed data lake is a place to manage the supply and demand of data. It gives your data analysts and data scientists a single collaborative environment to discover, explore, relate, and acquire any type of data from any source inside or outside the enterprise, so they can then prepare and deliver it for analysis.

The aim is to avoid creating a data 'swamp' where you just dump all the data you can find into one place without worrying about the fact that it's a multi-structured mess.

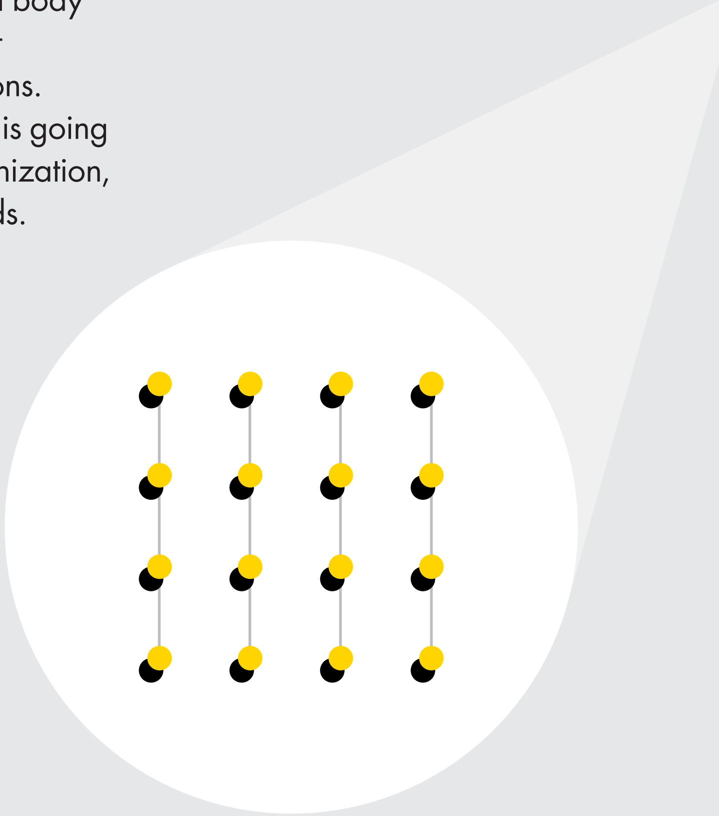
A managed data lake is one that:

Rapidly provisions data

It makes sure all types of data can be on-boarded, prepared for analysis, and delivered to business users' tools in a quick, automated way. In essence, it's making all the data your analysts need available to them.

Abides by data governance operational guidelines

That means it follows rules for accessing and analyzing data that are predefined by a body of people from different departments and functions. If a managed data lake is going to serve the whole organization, it must represent its needs.



A repeatable process to harness dark data

Holds on to descriptive metadata

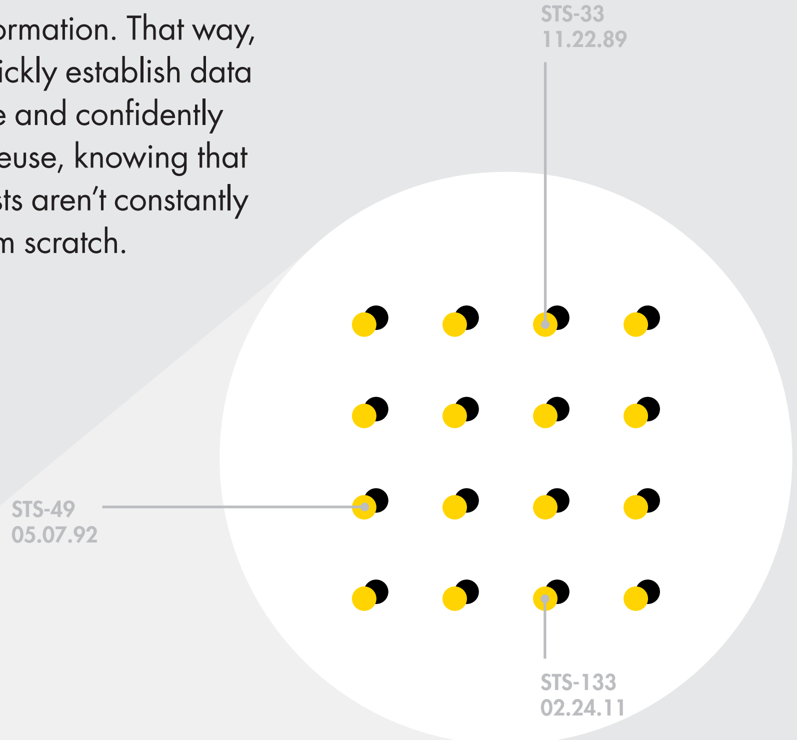
In a managed data lake, the data is mastered by domain and stored with as much contextual metadata as business users will need before querying. Without this catalog of context, your data will remain shrouded in confusion.

Maintains data quality

For your data lake to offer consistent value to the business, you have to be able to rely on the quality of the data. As such, it must maintain standards of data quality that are in keeping with your business users' expectations and the intended use of the data.

Records the data lineage

So that your analysts aren't working backward to understand what transformations have taken place in the making of a dataset, a managed data lake stores data lineage information. That way, you can quickly establish data provenance and confidently maximize reuse, knowing that your analysts aren't constantly starting from scratch.



A repeatable process to harness dark data

Secures your data

The difference between a managed data lake and an unmanaged one usually comes down to the ability of the technology to adequately secure the data. The aim is to control access to different data through layers of security such as access control, encryption, and data masking.

Maintains semantic consistency

So that your business users can conduct self-service analysis of a variety of datasets, you need to make sure those datasets are pre-integrated into your system whenever possible. That way, they're easy to find, simple to understand, and don't require constant transformation.

You should know that at this point, data lake technology doesn't offer all the same capabilities of a data warehouse. It doesn't support transactional consistency or offer a purpose-based model for optimized reporting.

That being said, the data lake offers you a view of your dark data that doesn't tax your data warehouse. It gives you the option to explore your data before you actually process it in your data warehouse and business intelligence systems.

This is a crucial advantage over the current reality that most companies find themselves in, where they can take months to provision data to their business users.

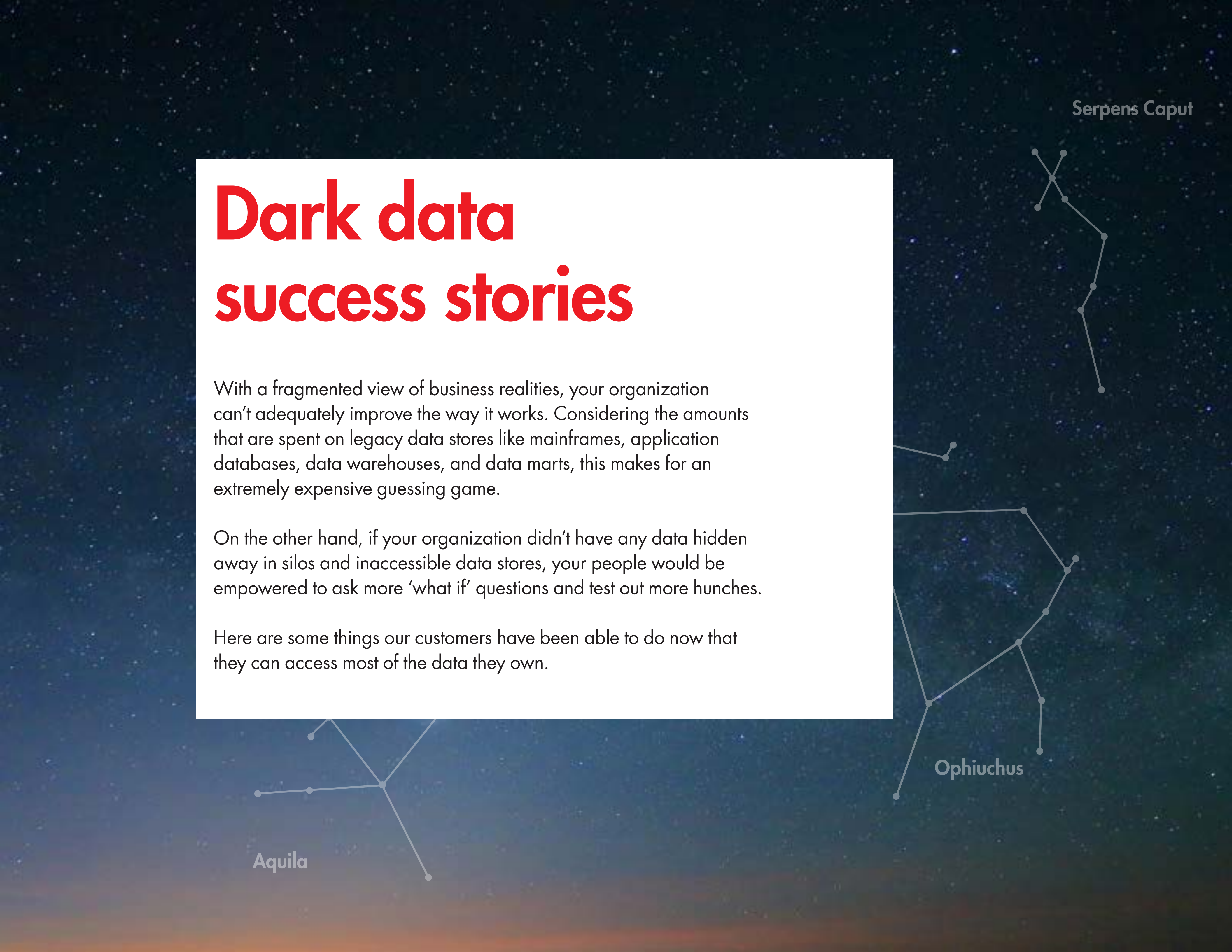
In the next section, we'll look at what our customers have done to make the most of the data hidden away in their systems.

Dark data success stories

With a fragmented view of business realities, your organization can't adequately improve the way it works. Considering the amounts that are spent on legacy data stores like mainframes, application databases, data warehouses, and data marts, this makes for an extremely expensive guessing game.

On the other hand, if your organization didn't have any data hidden away in silos and inaccessible data stores, your people would be empowered to ask more 'what if' questions and test out more hunches.

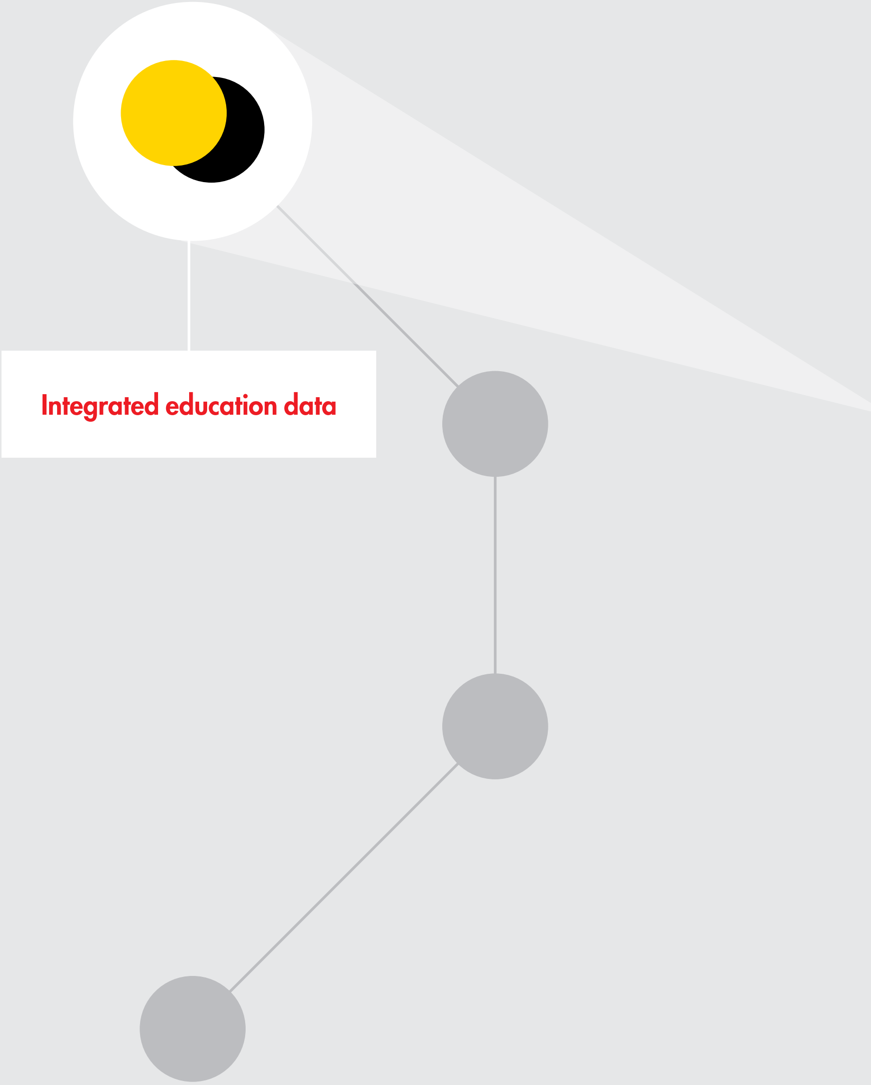
Here are some things our customers have been able to do now that they can access most of the data they own.



Dark data success stories

In 2010, the state of Washington was awarded a federal grant to enhance the statewide integrated education system. It had to pull together data that was sitting in silos across 10 separate departments – including social services, workforce, and education agencies.

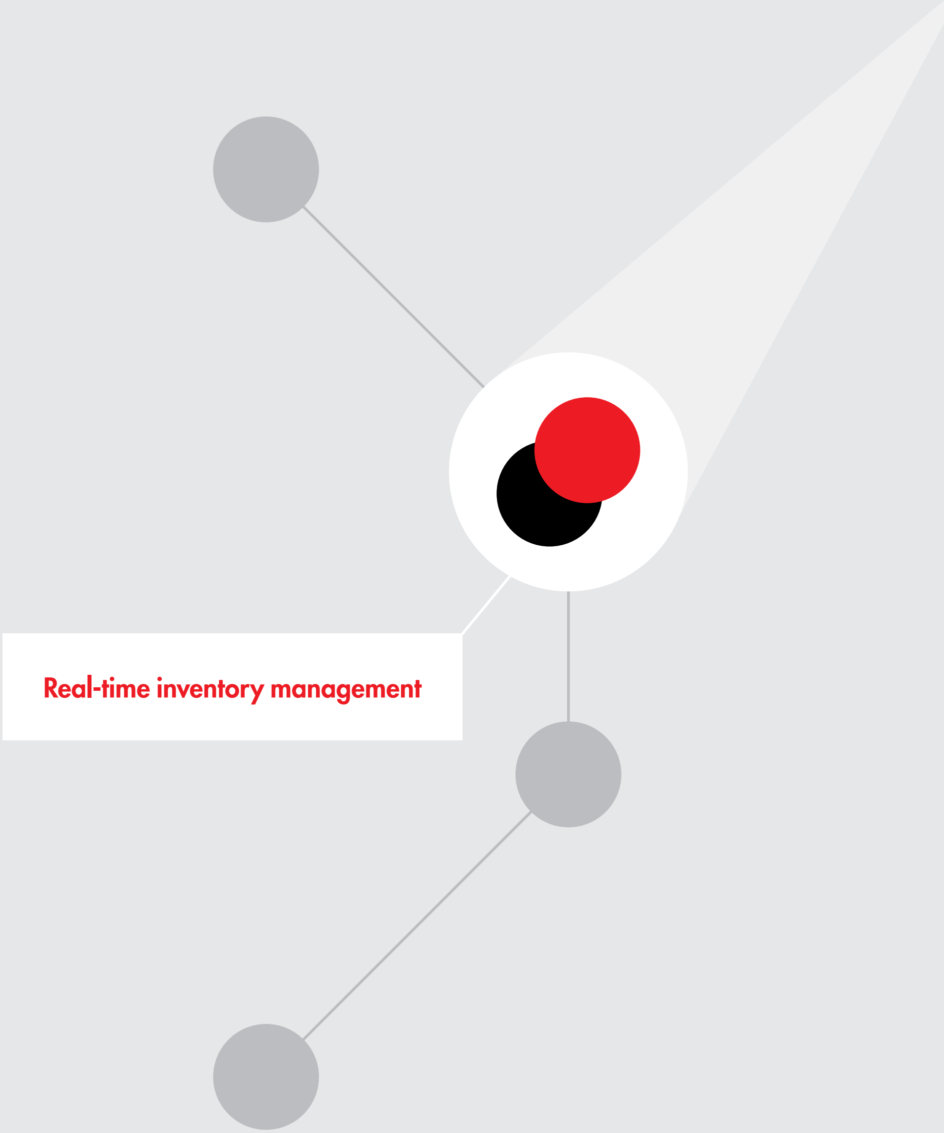
Four years later, the state has pulled together K-12, post-secondary education, and workforce data to gain a single, longitudinal view of students over time. Legislators, researchers, educators, and program managers have improved the quality of the statewide education system. Their policies and programs are now based on a clear understanding of the variables that cause some students to prosper while others struggle once they finish school.



Dark data success stories

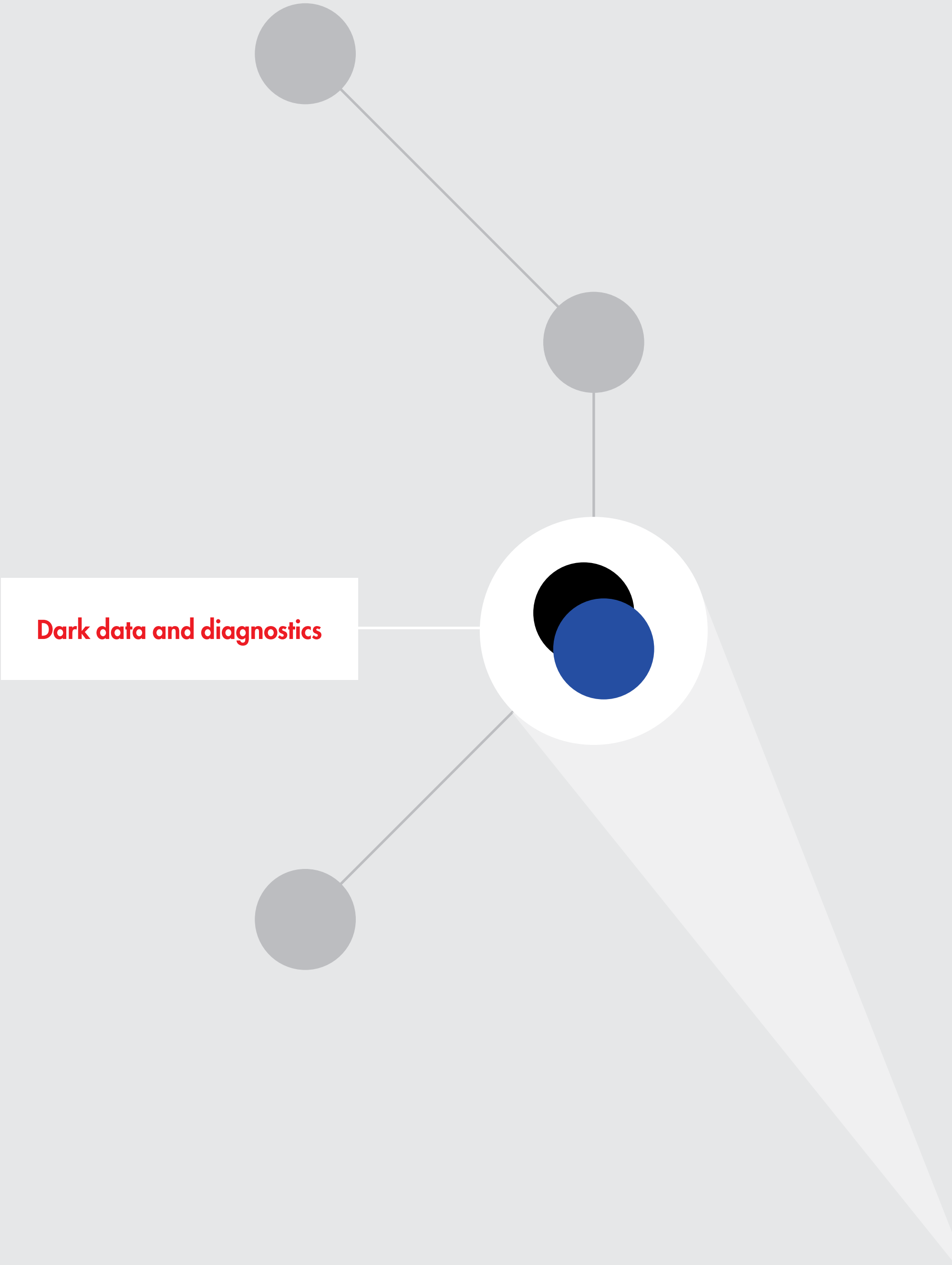
We've already mentioned the struggles Interstate Batteries was facing – managing and forecasting inventories to predictably meet retail market demand and optimize pricing decisions.

The company can now make agile and responsive inventory and pricing decisions by pooling disparate internal datasets, and supplementing them with data from unstructured external data sources such as weather, competitor pricing, auto purchase trends, and social media conversations.



Dark data success stories

We've already told you the story of how UPMC integrated clinical and genomic information on 140 patients previously treated for breast cancer. Watch the video to understand how it used information and standardized, shared datasets to give a more holistic picture of its data, enabling true competitive advantage and the ability to understand the impact of care far before the potential onset of disease.



Dark data and diagnostics

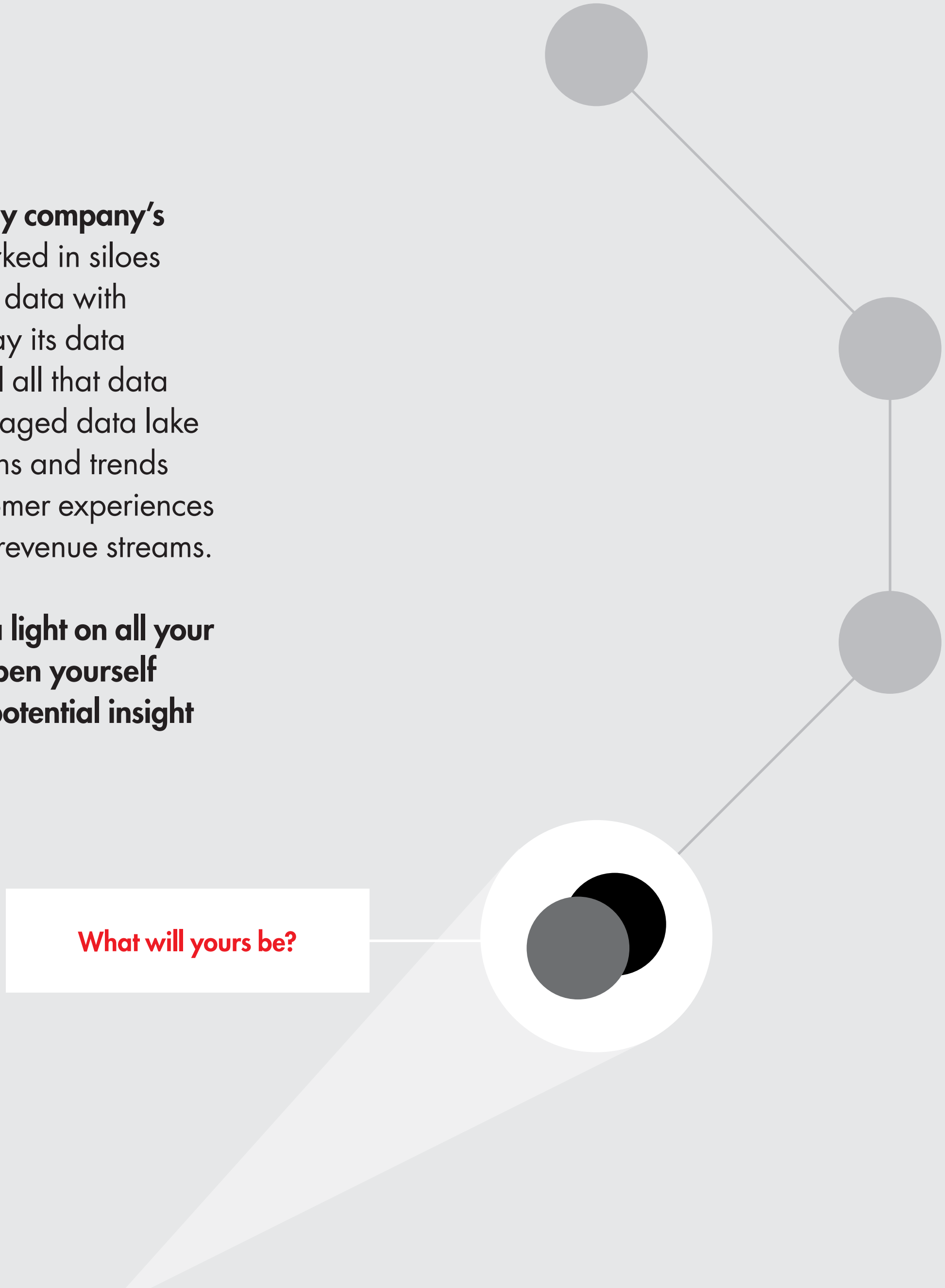
Dark data success stories

A large, global financial services company has built a managed data lake to reduce the cost and improve the effectiveness of its anti-money laundering and fraud prevention processes.

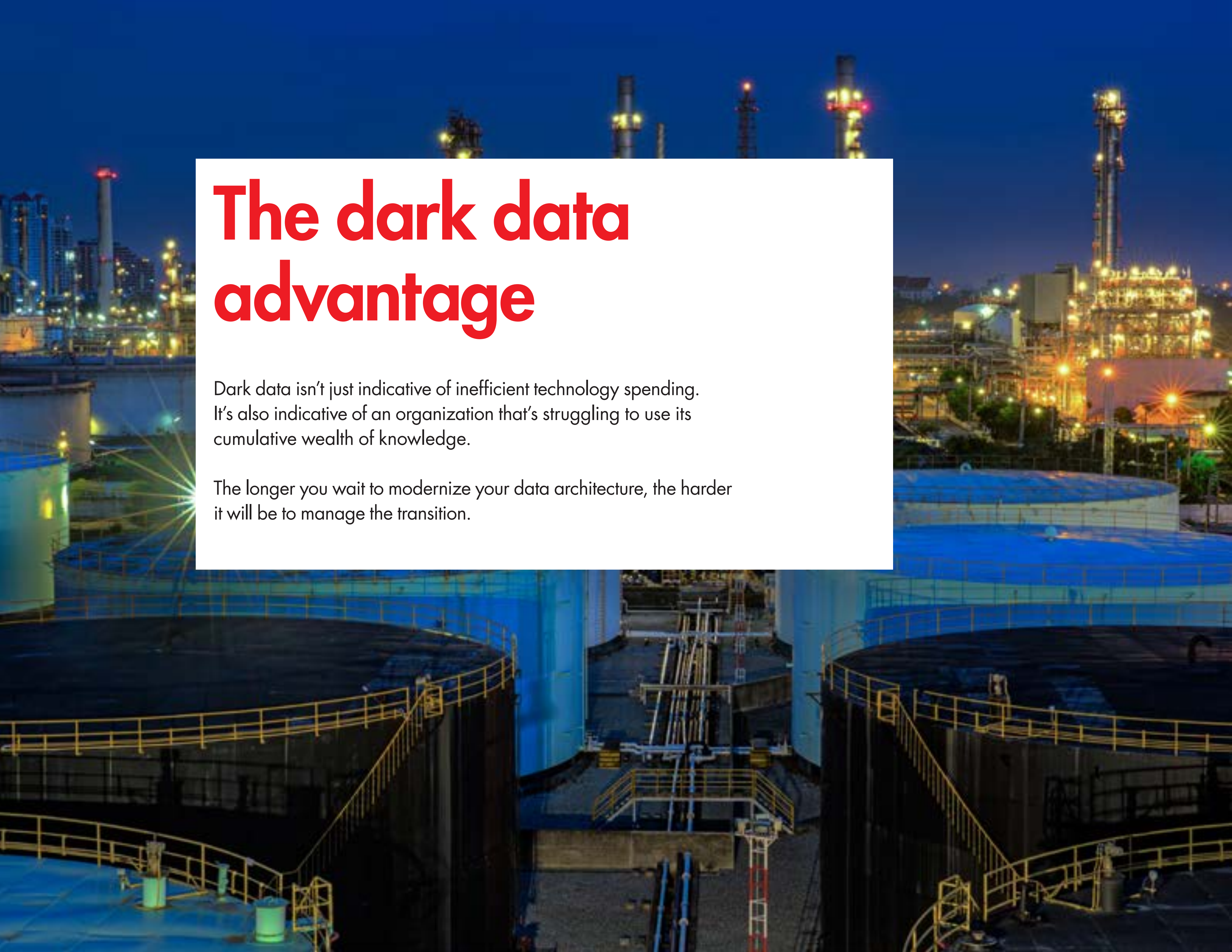
An oil and gas company has built a data lake to store and process data from seismic activity, sensors, devices, and its wells. That way, the company could integrate it with enterprise application data and reduce the cost of service and maintenance while increasing the production efficiency of its wells.

A large technology company's business units worked in siloes and never shared data with one another. Today its data scientists can pool all that data together in a managed data lake to discover patterns and trends that improve customer experiences and identify new revenue streams.

When you shine a light on all your dark data, you open yourself up to a range of potential insight and efficiency.



What will yours be?



The dark data advantage

Dark data isn't just indicative of inefficient technology spending. It's also indicative of an organization that's struggling to use its cumulative wealth of knowledge.

The longer you wait to modernize your data architecture, the harder it will be to manage the transition.

The dark data advantage

The longer you wait to modernize your architecture and make this data available to business units, the harder it's going to be to deliver the information they need.

On one hand, there's no telling when competitors might rise to the challenge and use their dark data to gain competitive advantage. But there's also the issue of enterprise data's unrelenting, exponential growth. More than 90 percent of the world's data was created in the last three years. And it isn't going to slow down any time soon.

For organizations to finally take advantage of all the data they've been paying to collect over the years, the data has to be made visible, accessible, and ready for querying. And the data itself has to be clean, safe, and connected.

Once that's done, you can start to answer the kinds of questions your business users have always asked. Only now you'll have empowered them to base their decisions in solid data you can vouch for. Data that addresses questions like:

- What are the next best actions to reduce churn of my most valuable customers?
- What is the best indicator of product failure?
- How can I proactively seek out the typical causes of maintenance?
- What post-sale activities typically lead to repeat purchases?
- Can we identify internal subject matter experts by the amount of content they produce and store in our systems?
- Do our interview scores tell us which HR employees typically identify candidates who perform well?