The new secret sauce for data

WHY DATA CATALOGS ARE YESTERDAY'S RECIPE

Section 1:

Let's cook up a feast...

Delivering first class products and services is a lot like cooking a meal.

You take your raw ingredients, prepare them with care, apply your expertise to the cooking process, and then serve up the finished product to your customers. **Bon appétit.**

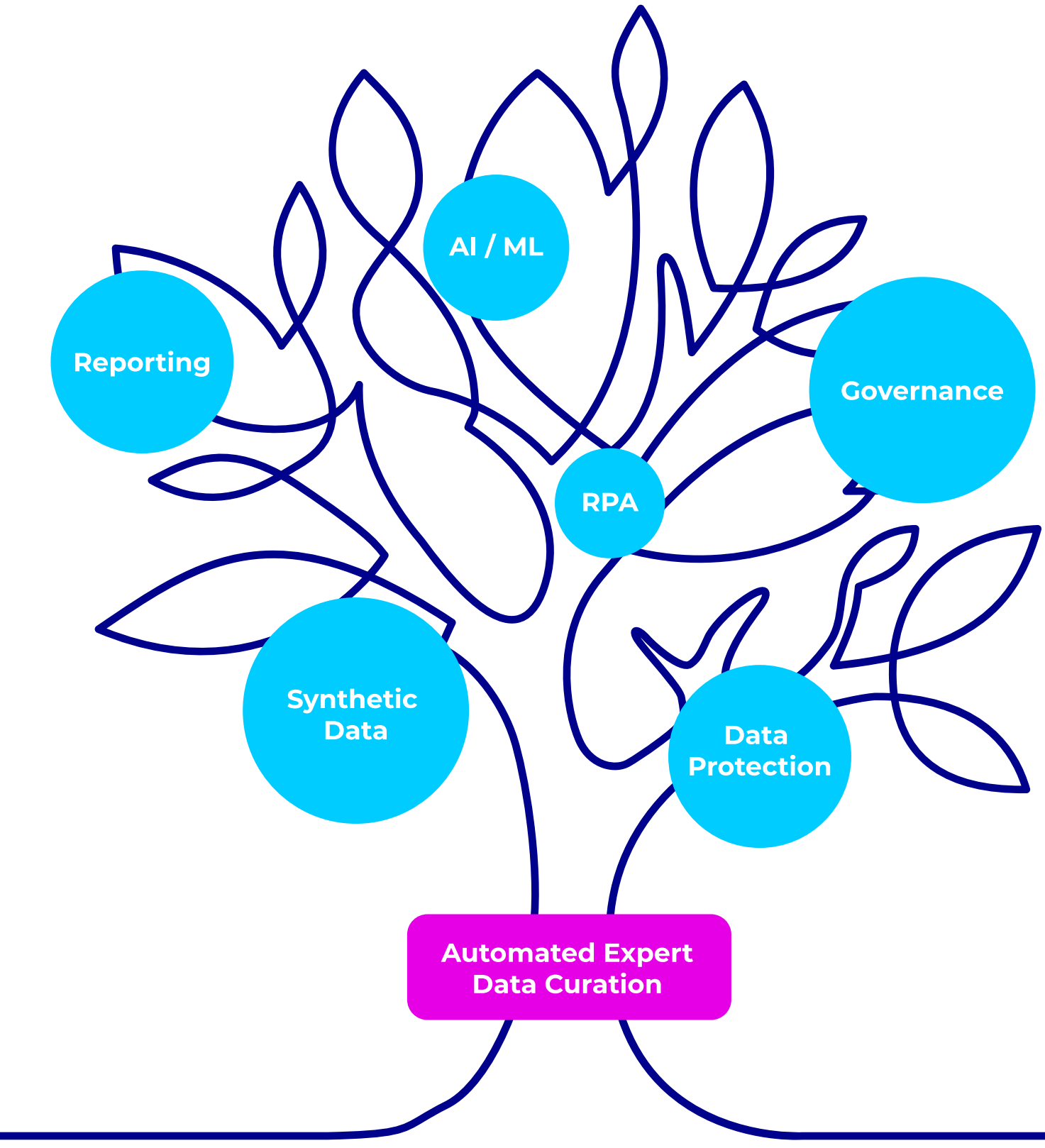
As you already know, it's a lot more complicated than that. In this ebook we've brought together a group of leading data experts to explore the fundamental data challenge that every ambitious enterprise faces, the solutions that have been tried, and what comes next.

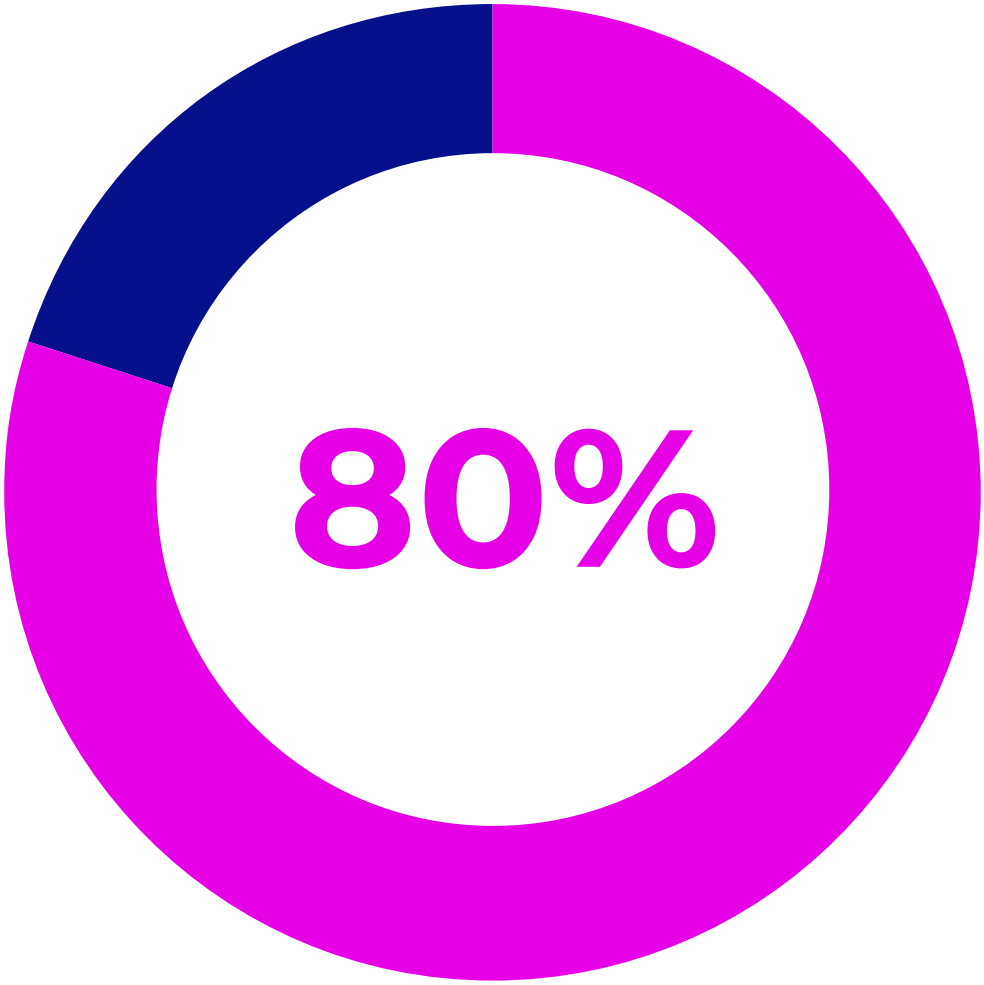
AN OLD PROBLEM

For a 21st century business, the raw ingredients are data. We've built up tried-and-tested methods for preparation and processing that feed directly into decision-making. This includes dynamic pricing strategies, risk profiles, recommendation models and all the basics of analyzing customer and client behavior. All of these key components rely on the quality and freshness of those ingredients.

"How do you get the insights out of the data that's available that you're authorized to use from a security perspective?"

"We're facing the same challenges that have been around for a while: data management, data governance, data protection," says Mitch Schussler, Field CTO in Financial Services at AWS, with a background in application development, infrastructure and security. "You can also add the expense management associated with data, and the resilience, reliability, and quality of the data. Those things haven't changed. How do you get the insights out of the data that's available that you're authorized to use from a security perspective?"


Reporting
AI / ML
RPA
Governance
Synthetic Data
Data Protection
Automated Expert Data Curation



Analyst time spent on getting the right data

-Gartner, ZDNet, IDC

"The biggest challenge is the heterogeneity of information," says Alasdair Anderson, GM EMEA of leading data security firm Protegrity. "The information is stored all over the place. That was a problem back when people just had mainframes. Now, you have to multiply that problem by legacy technology, and the move to the cloud, the move to the edge, and IoT. You have to be able to cater for all of that, in real time with an ever expanding universe. What was a difficult problem is on the verge of becoming an impossible problem."

FINDING THE RIGHT INGREDIENTS

The fact that the data is distributed across so many systems means that filtering out insights is an even more onerous task.

“There are a bunch of studies that say you'll only use about 15% to 20% of your data,” says Andrew Ahn, CEO and founder of Praxi Data. “And that's probably true. But the question is, what is the 15% to 20%? How do you pick that out? And the only way you're going to pick that out is to look at everything to figure out if it is the one-in-five that you need.”

“The right data is the only path to modern analytics”

Andrew Ahn, CEO

Brad Currin is Chief Data Officer at Shell, looking after the Upstream & Integrated Gas businesses, and focusing on safety and asset management excellence across the organization. “Within my own organization, I hear stories of analysts spending 70% to 80% of their time just finding the data they need”, he says. “This is a very common headache. One of the biggest challenges is silos. In a perfect world, data would flow seamlessly from one

part of the organization to another, aligned through central master data. But what often happens is that one part of the organization needs to move faster at a certain point than another, and they may end up creating their own copy of data sets to achieve their specific goals. Over time, the organisation, as a whole, ends up with snowballing duplication of data, requiring significant effort to link up the disparate data sets. As a result, you cannot get a single picture of a business process or the performance of an asset."

"The biggest impediment for a number of organizations on their journey in being data driven is the fact that they don't know what they have," says Akhil Lalwani, Chief Data Officer at Convex Insurance, and has also worked in data roles in telecoms, banking and law enforcement. "You cannot trust what you don't know. Everybody talks about data-driven decisions and being 'different by data'. But if you don't know what you have, then how are you ever going to be different?"

"The biggest impediment for a number of organizations on their journey in being data driven is the fact that they don't know what they have"

Also, with the proliferation and expansion of data comes a growing threat: dark data.

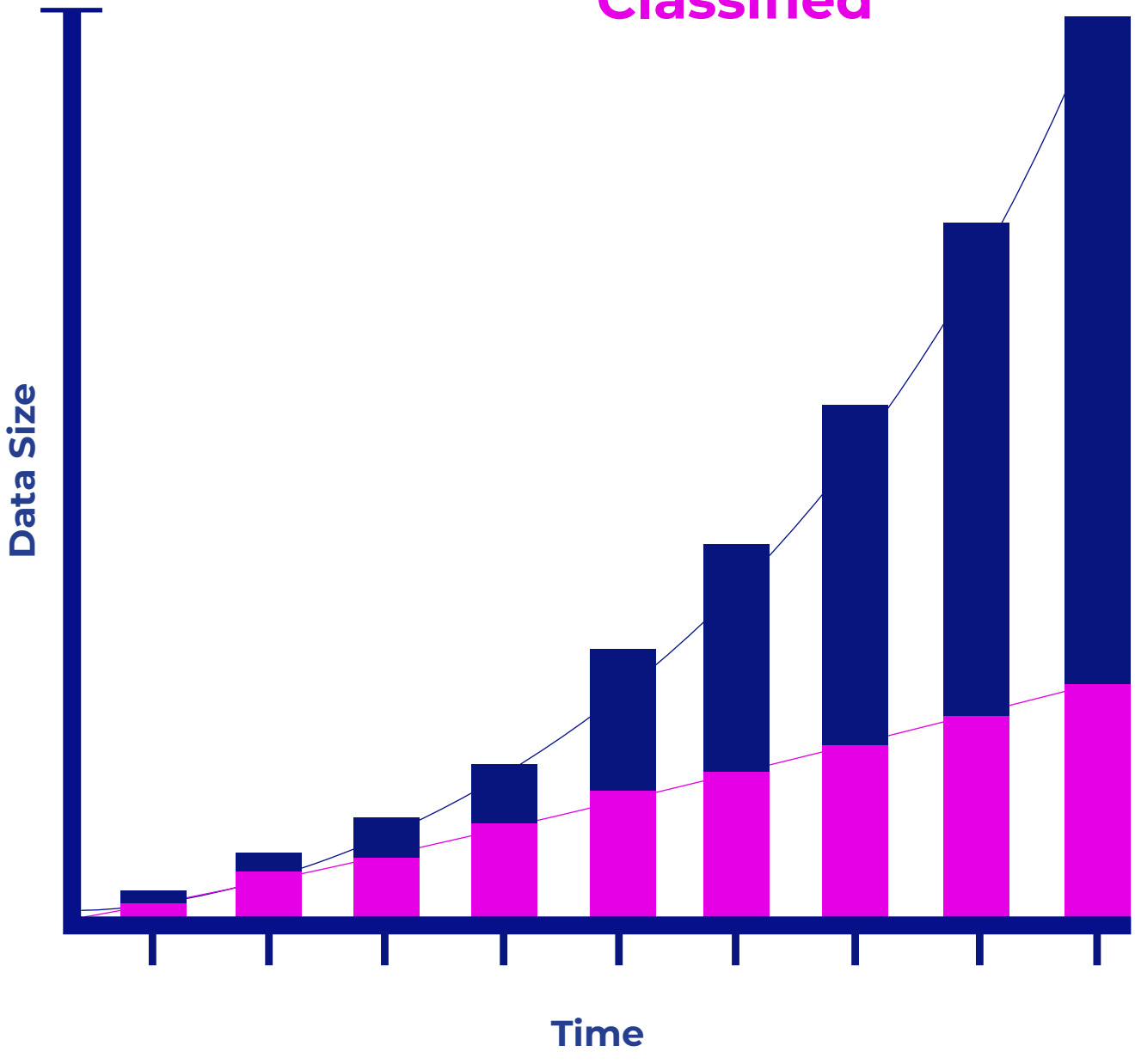
80% Unknown

Data Growth

Data Understood

80 Zettabytes

Classified



TOXIC INGREDIENTS

Up to 80% of business data is dark data – stuff that completely unknown or only partially understood. It's captured and stored, often at great expense, but from a business perspective it's unusable. It's like the unlabeled Tupperware lurking at the back of the freezer – you don't know what it is, where it came from, how long it's been there, or where it can safely be combined with.

Now imagine a warehouse full of those unlabeled containers, with new containers being brought in all the time...

"You see many organizations now taking more of a security stance in regards to data privacy"

This creates business risks. It's impossible to apply effective governance to this disorganized, duplicated mass of datasets. Somewhere in the pile may be an ingredient that's about to turn toxic, causing contamination, or that might provide a vital missing link.

Either scenario carries the potential to expose the organization to severe penalties. Furthermore, some data needs to be stored for several years (eg for financial audits), other data needs to be disposed of as soon as possible (eg for GDPR compliance). If you don't know which is which, that's a major issue.

The pressure to deal with dark data is both internal and external to the business. "You see many organizations now taking more of

a security stance with regards to data privacy," says Schussler. "When you're collecting more and more data, how do you handle that privacy? Because you need to maintain trust with customers. You don't want to do anything that affects your brands or creates any regulatory issues. Part of this journey is about finding the data that's not valuable. Are you collecting or distributing too much data?"

The dark data problem is not only big – it's getting bigger all the time. This means that any fix needs to be strategic, not a one-off project. So for enterprises that are serious about removing this headache once and for all, what is the best approach?

"The dark data problem is not only big – it's getting bigger all the time. This means that any fix needs to be strategic, not a one-off project"

Section 2:

Kitchen nightmares

The solution to these problems is always going to lie in the structure and labeling of business data. Robust, well-governed metadata not only makes our raw ingredients findable – it gives us provenance (where they came from), traceability (how 'fresh' they are), quality (how reliable they are), and security (whether they're safe to throw into the cooking pot).

"You can't manage data directly, you have to manage it through metadata"

"You can't manage data directly, you have to manage it through metadata, it's just too cumbersome and too wiggly, and it is all over the place," says Ahn. "So you need some proxy, some method by which to manage it. And definitely it's through metadata."

“You need some proxy, some method by which to manage it”

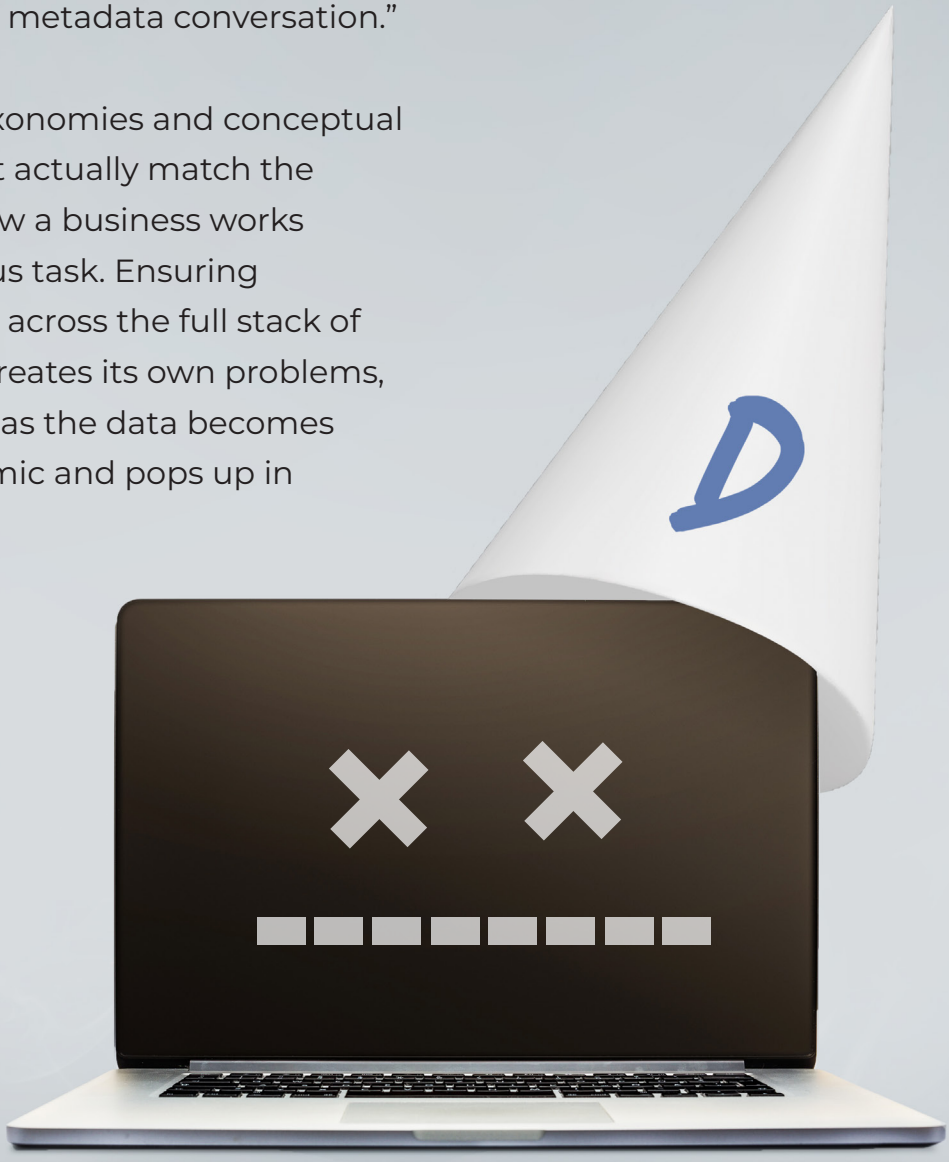
But metadata is hard, particularly when you're working with a mix of structured and unstructured data. Mike Fishwick has worked in senior data roles across the private and public sector since the 1990s, and was working with machine-learning models when the field was considered bleeding-edge. He has a deep understanding of the challenges. "Historically, we've been looking at exploiting structured data. That is, hard, quantifiable, fact-based data," says Fishwick. "Now we're in a digital world where a proportion of our data is electronic versions of paper documents. The big question is, how do we manage that unstructured data? Can we bring analytical exploitation capability to that unstructured data? The way to do that is to create a taxonomy or dictionary that's specific to the particular genre that we're working in, and apply that back to all of the documents."

"But metadata is hard, particularly when you're working with a mix of structured and unstructured data"

Andrew Turner is GM of EMEA for Praxi Data, and has spent decades working on data platforms with companies including General Electric and Tesco, launching the latter's mobile telecoms arm in the UK. “Tools like Hadoop allowed us to bring lots of structured and unstructured data into a data lake, and that was a big shift”, says Turner. “But now you need to put a magnifying glass, you could say, on top of that data. How do you search for that data? And that's where you get into the metadata conversation.”

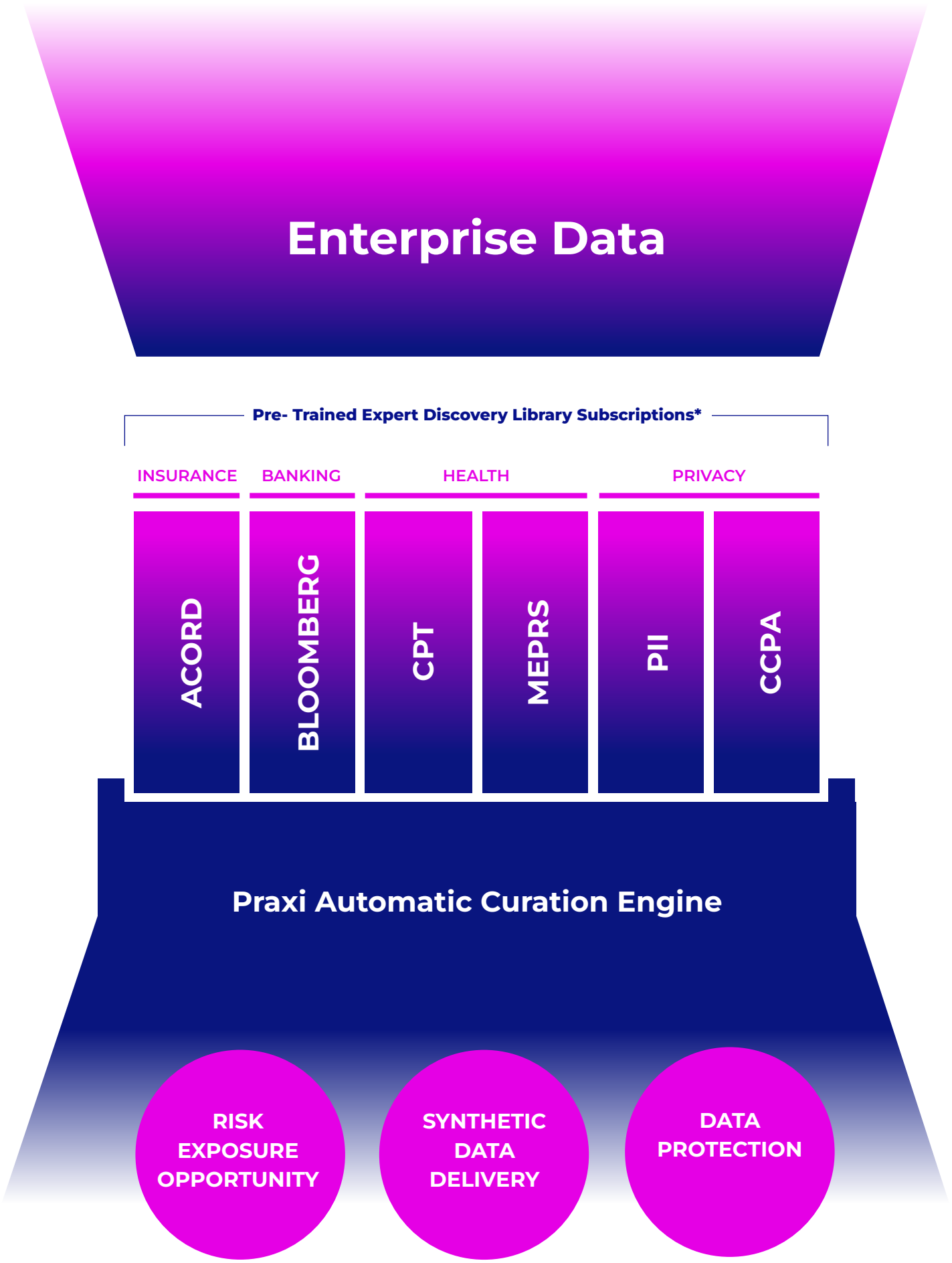
Building taxonomies and conceptual models that actually match the reality of how a business works is an onerous task. Ensuring consistency across the full stack of data tools creates its own problems, particularly as the data becomes more dynamic and pops up in different places. Generating and maintaining a reliable data catalog is an exhausting and expert job. Moreover, traditional data catalogs and BI tools have been general purpose, dumb, and, by the nature of the beast, always out of date.

“But now you need to put a magnifying glass, you could say, on top of that data”



“Most of the data catalogs out there are primed by human knowledge and human intervention”, says Anderson. “What stops that from succeeding? Two problems. First, there’s almost a mirage within the data management community that at some point, you can provide a single view of data with a single view of the truth. It’s very difficult to get a consensus on definitions of information that is actually driven by the usage, and the context. The second problem is the velocity of information. By the time you actually go and build some infrastructure, and place information into it, and then draw a report out of it, you're probably a year later than the question you've asked.”

“A mirage within the data management community that at some point, you can provide a single view of data with a single view of the truth”

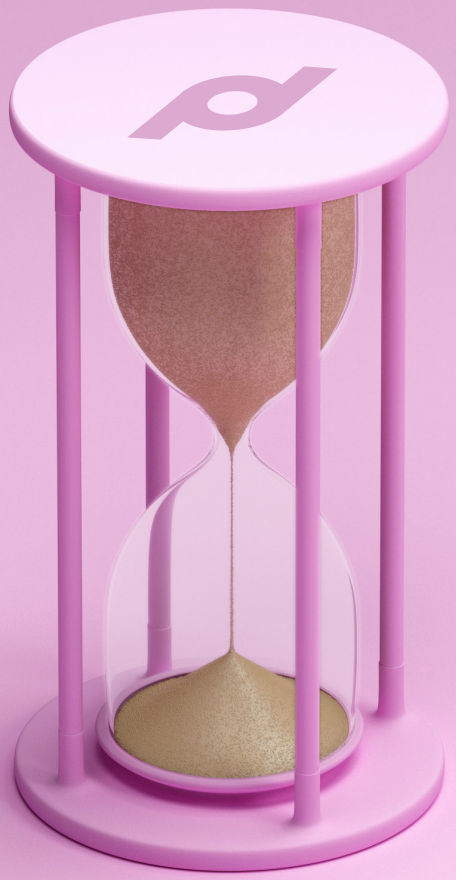

Enterprise Data
Pre- Trained Expert Discovery Library Subscriptions*
INSURANCE
BANKING
HEALTH
PRIVACY
ACORD
BLOOMBERG
CPT
MEPRS
PII
CCPA
Praxi Automatic Curation Engine
RISK EXPOSURE OPPORTUNITY
SYNTHETIC DATA DELIVERY
DATA PROTECTION

NOT ENOUGH COOKS

The human approach doesn't scale – because more data is pouring in all the time, manual processes (patched together with manual scripts and ad hoc code) will never clear the backlog. Our expert chefs can't keep up with the orders that keep coming in. Outsourcing is prohibitively expensive – that's why the deployment of data tools often costs two to three times more than the software itself.

And humans are prone to making mistakes. "If I'm entering 1000 pages of information every day online, I'm going to get fat fingers, and I'm going to press another key which adds another zero or another decimal to the whole equation," says Lalwani. "So, we are moving to a point where we need to leverage things like machine learning and AI to automatically extract what we think is valuable, and make sure that we can avoid those challenges around the quality of the data and the accuracy of the data."

"Even if we were able to scale, at some point in time, it's just not even possible because we generate so much data," says Schussler. "And we also duplicate and repurpose so much data. You may use data for analytics, but then it's repurposed for machine learning. It's very difficult to stay on top of that and manage that data strategy overall."



“The diversity of data has gotten 100,000 times more complex”

“The diversity of data has gotten 100,000 times more complex,” says Turner. You’ve got web data, you’ve got structured data, you’ve got unstructured, semi-structured, and so on. We need a capability that actually brings order to the chaos. From all this data we have – which could be light data, or dark data – there’s stuff you don’t even know exists that you could apply to your business processes. How do you get from that data lake or data store to something you can actually make a business decision with? And then continually refine it, and iterate and get feedback on it.”

Sheer volume of data makes it impossible to scale up manual processes. Meanwhile, the business is seeing no benefit. “The thing you want from your digital transformation, you want it to be faster,” says Anderson. “So you need to understand that time to market and speed is being held up by the lack of automation and data curation process.”

Section 3:

Can’t stand the heat?

By this point the kitchen is on fire, and your cooks need help. This depressing situation will be all too recognizable to CIOs and CDOs who have been on this journey. You've invested in tools, you've brought in systems integrators, you've put in potentially years of work – and you're no closer to solving the problems. The rest of the C-suite is losing faith, and morale is low.

We know that smart automation is the way to go, but how to implement it? Many of the tools on the market try to be all things to all people – solutions that can fit any industry or sector. But this inevitably means a ton of configuration and customization work. So you're back to paying for services from vendors, or hiring in other specialists – and you've still got the human bottleneck.

"In the past, companies tried to hand-carve their metadata models over a number of years. But you can't start with a blank canvas"

"If we can give you 80% of what you need out of the box, then you can refine it and customize it for your particular need"

"In the past, companies tried to hand-carve their metadata models over a number of years. But you can't start with a blank canvas", says Turner. "What if, instead, we can effectively crowdsource the knowledge and build that model. If we can give you 80% of what you need out of the box, then you can refine it and customize it for your particular need."

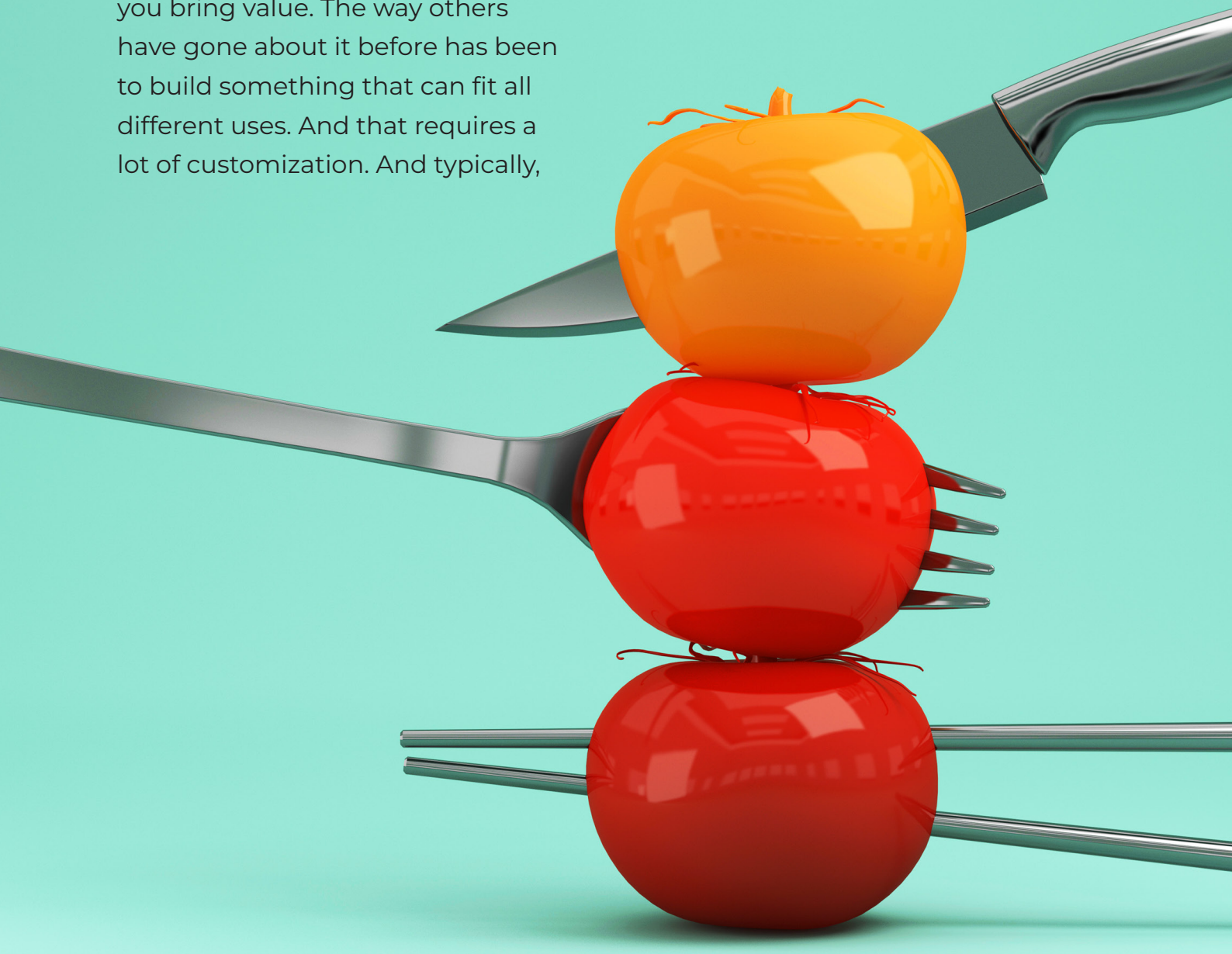
INSURANCE:
ACORD
(Property &
Casualty)

HEALTH:
Procedure,
Codes
(CPT, MEPRS)

Models that have been pre-trained on enough quality data in a specific sector (like healthcare, finance or insurance) can find and identify data more quickly and more accurately. They can also save two or three years of in-house training.

“The idea is not to build a general-purpose tool – it's to create something very narrow, but very deep", says Ahn. "That's the way you bring value. The way others have gone about it before has been to build something that can fit all different uses. And that requires a lot of customization. And typically, that means services – and that's the wrong combination. That's just a bigger time sink."

This type of tool doesn’t need to replace anything in your existing data stack – in fact, it’s the missing link that enables you to start generating ROI from all of your previous tech investments. You can index and optimize across all of the tools you’re using, even with highly dynamic data.



THE MISSING INGREDIENT – CONTEXT

Anything that gets value out of the data more quickly is going to benefit the business – and win hearts and minds along the way. “We definitely need to focus on the culture element, because we forget that we are working with people ultimately,” says Lalwani. “They will be using the assets that we build, and if they are not convinced in the answers they are getting, they will not use it. Also, we always talk about data in terms of value – but how many of us can actually quantify what that means? We need to be able to quantify the value that we are adding back to the business in truly quantifiable terms. That is the game changer.”

This approach not only ensures you’re getting value out of your existing data, it also creates new opportunities. As the nature and usage of data changes, the models can identify new relationships. “"Thirty years back, your phone number was just a number that maybe ten people had, so it was captured and in the form of notes or texts. Today, the number is the most important thing next to a person's name or social security number. Data changes shape, and relevance. Not a lot of time is spent understanding what the relative importance of the data is, hence it lies hidden somewhere that cannot be utilized.

“They will be using the assets that we build, and if they are not convinced in the answers they are getting, they won’t use it”

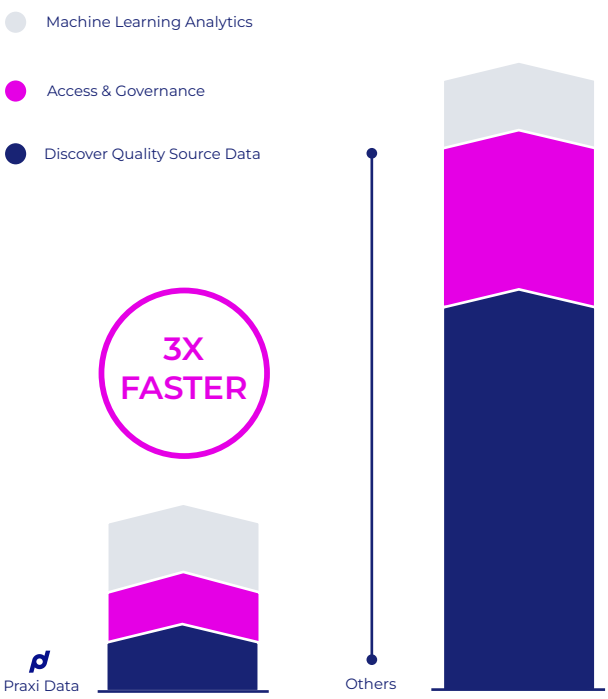
But if new tools can help bring that information to the surface, that can potentially be very, very useful."

And as anyone who is building ML or AI models for their own organization knows, you need both quality and quantity of data to achieve success. Over and under-fitting frequently goes back to the raw materials. As

well as eliminating stale data, the approach outlined above allows you to confidently generate synthetic data.

"If you want the ability to train AI models, you're probably going to need more data than you actually have," says Anderson. "We're definitely seeing a trend for the need for synthetic data – and that dovetails into data management. You can only produce really good representative synthetic data, if you have really well-managed data, where your data is labeled correctly, with its semantic context. Which brings us back to this point of intelligent automation."

Using pre-trained, deep models to curate your data makes ML initiatives far more likely to succeed. "If we were to look at safety data, for example, then we want predictive analytics: looking at data trends and the outlier patterns which can model likelihood, location and environments where safety incidents might take place and where equipment maintenance is required," says Currin. "If you're able to better predict, for example, where mechanical failure might take place, then you can prevent incidents that might cause harm to people, facilities or the environment. Equally, if you're able to accurately predict when maintenance should take place,



The amount of time spent to set up automatic curation percentage of data

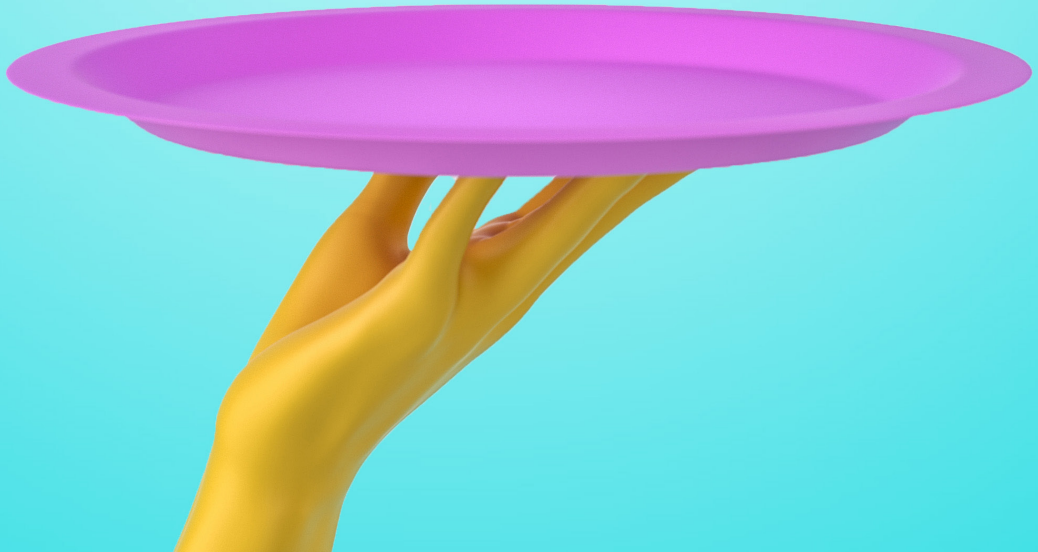
instead of running it on a fixed schedule (eg inspections every week or month), you're being smarter and more efficient by doing it only when it's needed."

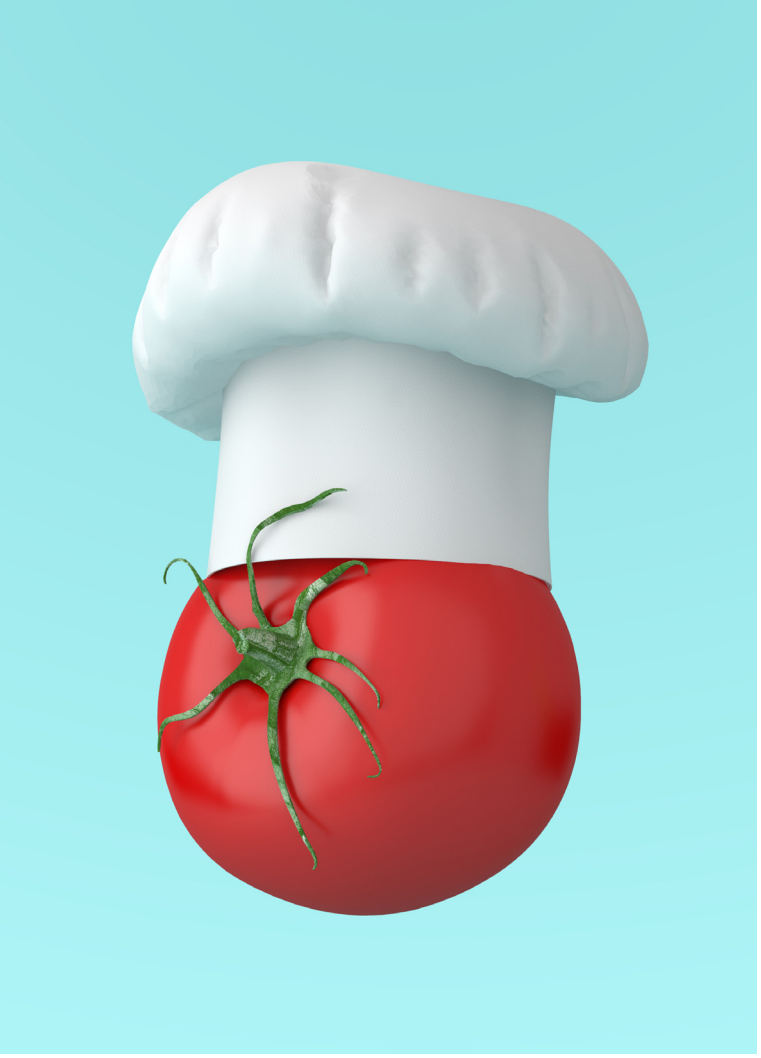
This is the missing recipe book. Creating these narrow, deep models is the only strategy that makes sense to tackle the data challenges of the 21st century.

" If you want the ability to train AI models, you're probably going to need more data than you actually have"

Section 4:

The finishing touches





LET THE COOKS COOK

This approach also frees up your in-house specialists from manual work, so they can focus time and expertise on more complex problems. “What you don’t want is staff using their PhD-qualified brains for basic data classification cataloging work, which at the moment has to be done because they are the ones with the specialist knowledge,” says Currin. “If that piece could be automated, then you’re in the world of properly leveraging those fantastic brains. Once we have rapid curation of that data, that’s when we can derive insights that will drive business value.”

“The more that humans work on these problems, the more data quality goes down”

“The more that humans work on these problems, the more data quality goes down,” says Anderson. “People get bored, they leave their jobs. If you train a machine learning model, your quality goes up, the more information you feed it. It creates automated experts. This is much more efficient, much more able to generate profit, on much lower, with improved quality over time. So you have an ROI that is insanely sharp. It can happen within a matter of weeks, as opposed to the data catalog, which has got an ROI of years.”

And while the models keep getting better, they also keep working through the new data that’s flowing into the system. “This gives you the ability to go and continue doing this discovery,” says Schussler. “The freshness of the data is paramount. This allows you to scale.”

CULINARY DISCOVERY

So your data will now be trustworthy, fresh and easy-to-find. But the tools don't have to stop there. The same vertical expertise that allows the models to map and manage your data will enable them to hunt for further opportunities. Once your kitchen is consistently producing great dishes, you can start to explore what else is possible.

A model with deep, specialist knowledge can start to intuit other relationships in the data. "Intelligent technology can actually profile the data automatically," says Turner. "It can look at the record structure and at the attributes, and can inspect the context of that data. It can then start to do matching. So there's a file called 'Cost Center', and another one called 'C Center', but the records are very similar. You can match them, so that you can extrapolate the metadata structure, and you can start to create and expand your metadata catalog. Yeah. So you can use metadata search to find data assets in this massive lake and quickly and provide insights."

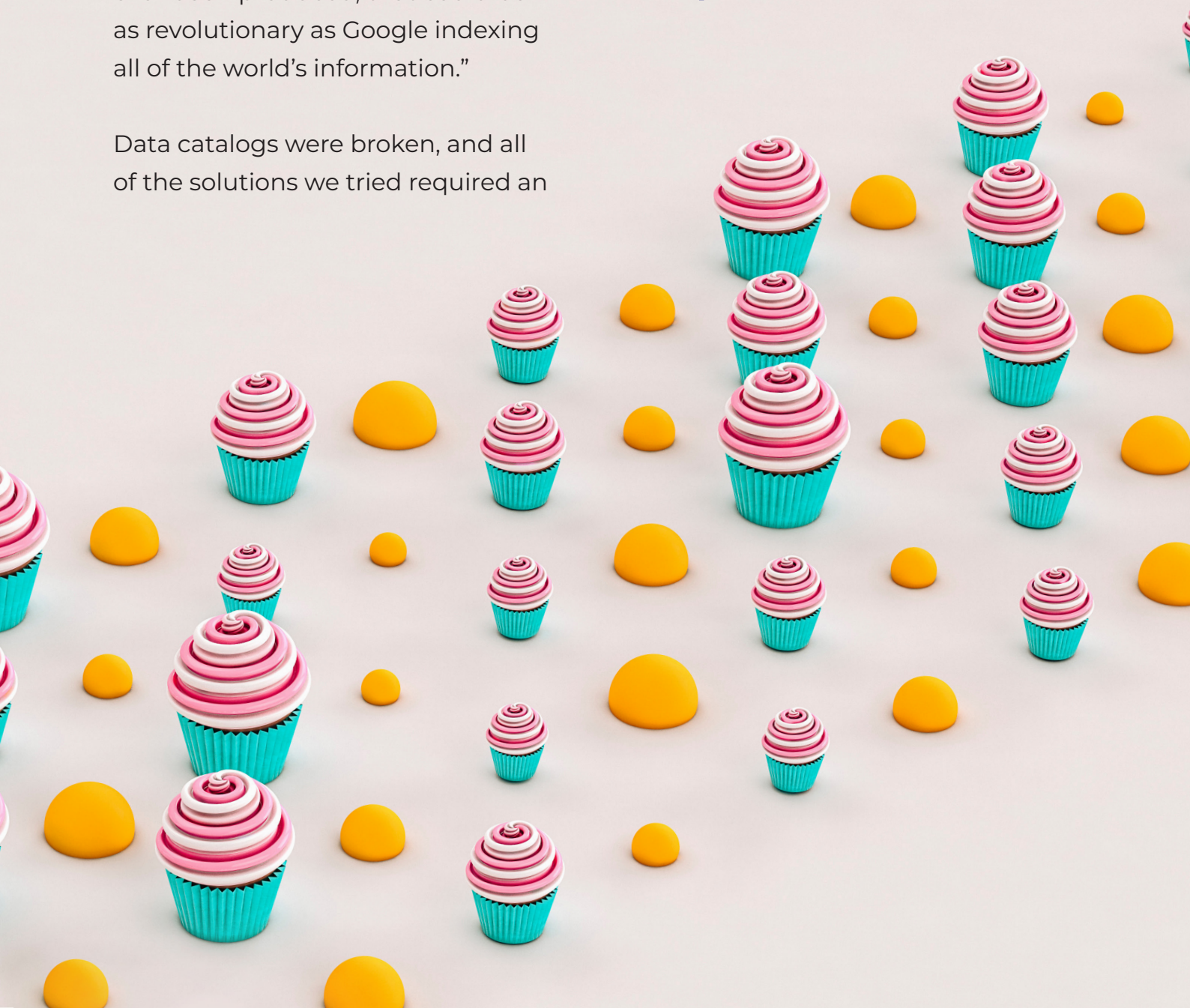
"A model with deep, specialist knowledge can start to intuit other relationships in the data"

New approaches to these data challenges are not only beneficial – they may be essential. "Consider how much data is being generated on a daily basis", says Currin. "The data sets firms generate are only going to get bigger and more complex. To efficiently handle these volumes and complexities of data we need tech companies that are thinking differently, thinking innovatively, and bringing solutions to the market that can help in that space. If not, we're going to drown. A few small tech companies coming up with bright ideas will potentially make a massive difference to organizations and their ability to extract value from their data."

“The opportunities that come from solving this problem are profound – across all kinds of sectors,” says Fishwick. “Actually, my nervousness with this, as with all digital transformation, is that people might want to do what they’ve always done, just a little bit faster. Instead, we might want to do new things, or do things in completely different ways. If we can create metadata for every document that’s ever been produced, that could be as revolutionary as Google indexing all of the world’s information.”

Data catalogs were broken, and all of the solutions we tried required an impossible amount of manual work. Generic, general purpose software wasn’t helping. So leveraging pre-trained models with deep sector knowledge is the only strategy that makes sense.

“The opportunities that come from solving curation problems are profound”



CONTRIBUTORS

Andrew Ahn
CEO, Founder

Extensive experience with Big Data technologies & governance in both highly regulated industries and open-source domains.

Andrew Turner
Advisor

Operational leader with a focus on strategy, growth & scaling. Proud alumni of GE, SAP, Tesco, WANdisco and EE.

Mitch Schussler
Advisor

Executive Manager of Information Technology with extensive experience developing and implementing strategies.

Akhil Lalwani
Advisor

Chief Data Officer at Convex. Experienced technology leader, having worked in insurance, banking and telecom.

Brad Currin
Advisor

Multi-disciplinary information, digital and data technology leader with 24 years of experience across a range of industries.

Mike Fishwick
Advisor

An expert in building data and insights teams from scratch. Also influenced the creation of data management and governance practises inside of organisations.

Alasdair Andersen
Advisor

Recognised industry expert who speaks globally on the issues surrounding the complexity of managing data.

PRAXI DATA

Designed by Phable