Data Curation: Fueling the Future of Responsible AI

Insights from our expert panel

December 2024

PRAXI.AI

Data Curation: Fueling the Future of Responsible AI

Praxi hosted a webinar on 1st October, 2024 featuring a panel of experts. Talking to the panel members, prior to the webinar amplified the need to have a discussion about what's going on across the world in relation to responsible AI, and their perspective on data curation to support the organizational transformation. It became clear, the key questions are related to the foundational items that you need address to drive business value from AI in a responsible way.

Introducing our Expert Panel

Roberto Maranca

Data Officer
Schneider Electric

Jon Hammant

UK & Ireland Specialist Lead
Amazon Web Services

Maxwell Rebo

CEO & Co-Founder
Straylight

Andrew Ahn

CEO & Co-Founder
Praxi

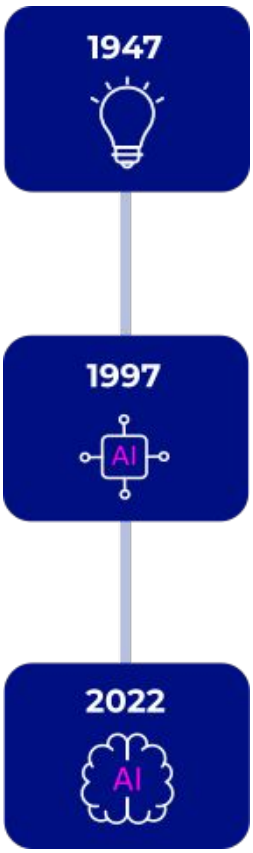
Andrew Turner

GM & Advisory Board Chair
Praxi

PRAXI.AI

What we covered:

AI: How it all started



The history of AI is rooted in human ambition, to create machines capable of learning and evolving. British scientist Alan Turing's pioneering vision laid the foundation in 1947 delivering one of the earliest public lectures in London on the topic of computer intelligence.

Turing boldly stated, “What we want is a machine that can learn from experience,” highlighting the groundbreaking idea that machines could adapt, by altering their own instructions.

While early AI systems relied on brute-force computing, today’s AI, including powerful models like ChatGPT, operates at previously unimaginable scales. This evolution opens new opportunities, particularly in data curation and management, where AI enables businesses to transform vast amounts of data into actionable insights.

In 1997, IBM's Deep Blue achieved a groundbreaking milestone that changed the game forever. It became the first computer to outsmart a reigning world chess champion in a match governed by standard tournament rules.

In the last 20 years, technology and its processing capabilities has progressed at an astounding pace. Within the technology industry, the measure of FLOPs or simultaneous operations per second is used to assess computer’s capabilities. In comparison with Deep Blue’s 11 billion FLOPs (11^9), as we entered the Age of AI in 2022, ChatGPT was trained using a processing power of 10^25 FLOPs, which is 10 quadrillion times more than the processing power of Deep Blue.

Across industries, but particularly where regulation is an important characteristic e.g. insurance, banking , healthcare, telecoms and government, managing data is more critical than ever. With AI, companies can streamline their data processes, ensure regulatory compliance, and enhance decision-making, setting the stage for significant revenue growth and operational efficiency.

This eBook explores the opportunity of data curation as an enabling capability for fueling the future of AI transformations, and how it underpins organisations being responsible in their use of data along this path. A number of organisations have either embarked on, or are embarking on their AI transformations, with senior leadership support. We enclose a summary of the conversation with our expert panel during our recent webinar.

Click here to watch the webinar in full at your convenience.

Very best regards
The Praxi Team

PRAXI.AI

Data Curation is Essential for Responsible AI Development

Data curation isn't just about organizing information - it's about ensuring that AI models are built on reliable, structured, and relevant data. For insurance and finance businesses, responsible AI development is crucial for regulatory compliance, reducing risk, and ensuring ethical decision-making.

High-quality data curation can help businesses mitigate bias, avoid incorrect predictions, and protect brand reputation in the face of regulatory scrutiny.

Q: What's the connection between data curation and responsible AI from a perspective of an operational role within a large multinational company?

Roberto Maranca:

To put people at ease, and confirm that I'm not a deep fake, I'm going to connect two things that a deep fake will never connect namely Latin and Engineering. Going back to the roots of the word curation which is "Curare" and means to "take care".

I think it's important that we think about this concept of care, because when we look at how AI works from a system theory point of view:

- you give it an input
- you have some process in the middle
- and you get an output

I think care is important because the difference between something that I know what it does, and something I don't know what it's going to do, is that I can responsibly say I know what this system is creating as an output.

Now if you want to look more from an engineering point of view, we can say "this is an engine". To maintain nominal performance of the engine we need to take care with the fuel, the air composition, the maintenance has to be done, the lubrication and so on. I mean we can go in all kinds of directions, but the reality is that data is a fundamental component of the functioning of our systems.

So I need to take care of what data will go into the system, not only at a point in space, but also at a point in time and so for me it's the key to guarantee the safety, resilience, function and performance of what is an important product for us i.e. to manage decisions, or help people to make effective decisions.

Data Quality and Traceability is Crucial for Accurate AI Decision-Making

Inaccurate or incomplete data leads to poor decision-making, directly affecting business outcomes. For insurers and financial institutions, decisions based on AI - from underwriting to fraud detection - are only as good as the data fed into the system. By prioritizing data quality, businesses can ensure that their AI solutions provide actionable insights, enabling better pricing models, reduced risks, and enhanced customer experience.

Andrew Ahn:

The things that make it really critical for responsible systems is the notion of traceability or chain of custody - a physical trace showing how data moves through various stages so when something happens - a law case or you get put in front of the Department of Justice, you know you need to pull on that thread and you need the chain of custody.

Something that looks backwards that you have to establish to maintain veracity, to maintain legitimacy. Where did you get the source? And this has always been a problem but it's even more of a problem with generative AI. There's a lot of things that can happen where things can fall through the cracks so that's one of the critical aspects where good data curation, and good classification really helps.

Answering authoritatively where the data originated from and what's inside of these documents. While these things may not be at the forefront, it is actually really critical to making sure that the outcomes you want are actually usable, trustworthy and authoritative so that you can build the rest of your data stack on top of it.

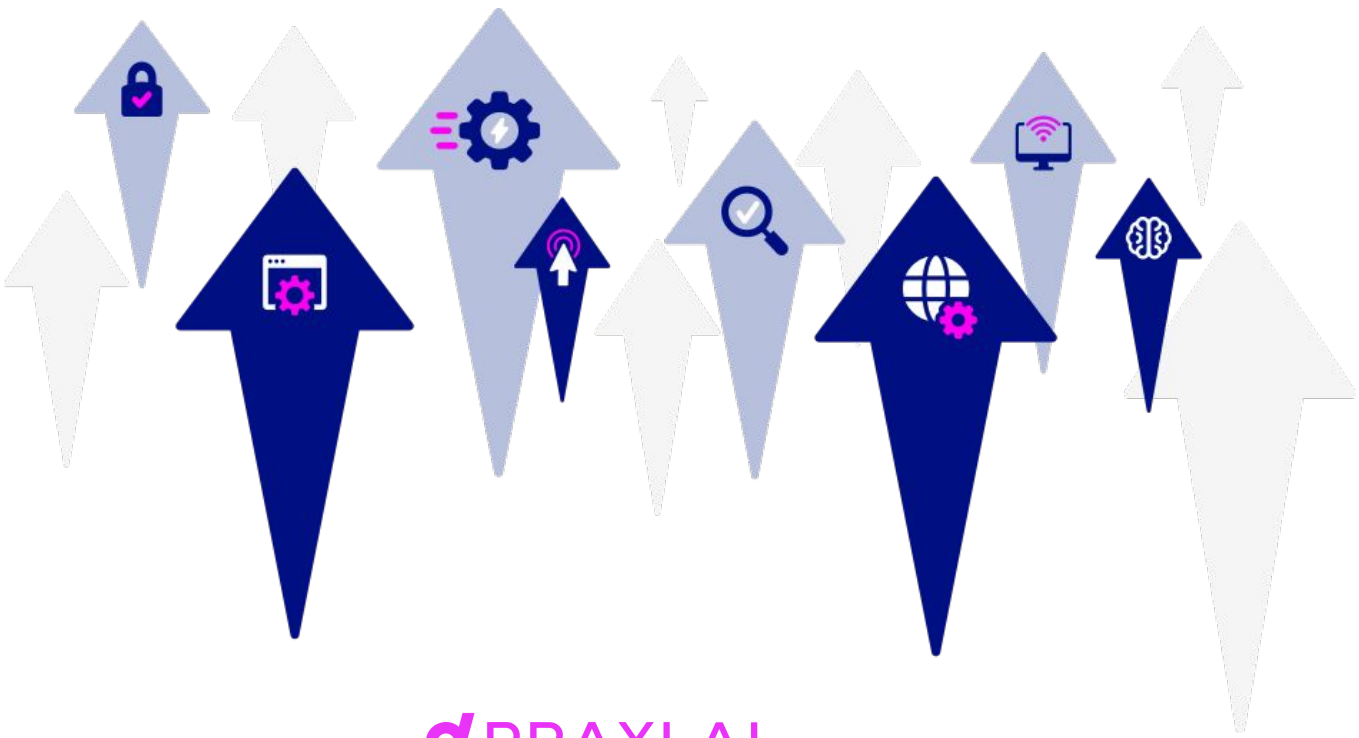
Roberto Maranca:

The data quality issue from the macro perspective is effectively a deviation between the thing is in reality, and what you're having that thing represented digitally. A deviation between what it is and what it should be.

When you are trying to build an AI model to have an intelligence to make a decision, then if the decision is based on something that's actually not fresh in time or not valid or not exactly stipulated on the right reference, the decision might be deviated from the ideal outcome. The challenges for me are really how do I maintain my data's digital representation in sync with reality.

If you don't have a recipe for data maintenance and alignment, you can't tame the change. The world changes all the time and if you can't come up with a recipe, sometimes you're much better off not even start doing AI and data because you know you're going to end up in a bit of a hurt if it goes wrong.

So I think the real challenge for transformation is I can't scale because I don't know what's out there. I can't fix or I can't expand my language model.



Jon Hammant:

If I look back 10 - 15 years, everyone decided to embrace this new concept of data lakes. Everyone who worked in tech was building a data lake, and I don't think people quite knew why they were building these data Lakes. There was a customer who worked with a very large UK bank that said “we don't have data lakes, we have data swamps.”

So these swamps of data were lying around and no-one quite knew what it was, and you kind of tried to stay away from it all. What I find really interesting is how the tech world imitates the natural world. What's happened over time is all those data lakes have been pushed further and further down and now people start talking about this expression that the data is the new oil. Where does AI fit into this new concept?

It may sound weird, but where AI for me starts to fit into this is that it’s a relatively simplistic algorithm so what's come out of it is almost like we've invented the combustion engine, that suddenly starts consuming all of that oil. We have suddenly realized oil can be processed in useful ways by these relatively simple very large scale but standard AI models.

Where I think data curation comes into this is if you look at the process of oil refining, stuff that comes out of the ground - the black gold - starts spurting out. At first it's not really that usable. It's kind of very generic. There's no effective way of finessing down to include privacy and security in Data Management.

Important that we need to take that data and we need to treat it for specific needs. How do we know it's the quality that's needed? How do we know it's the security that's needed? It’s important to make sure that what we're training LLMs [Large Language Models] with is what it should be trained with, and how do we not just forget the 20 - 30 years of good quality computer science security. That's why for me data curation without privacy and security doesn't spin.

It's absolutely key to everything we do both in terms of segmentation of the data and the quality life cycle.

Automation Can Significantly Enhance Data Curation Processes

Manual data curation is time-consuming and prone to errors. By automating the curation process, companies can save time, reduce labor costs, and improve data accuracy.

Automation tools can quickly process large datasets, clean them, and apply them to AI models, allowing insurance firms to respond more quickly to market changes and customer demands. Automation also frees up skilled professionals to focus on high-level strategy and far more valuable work, rather than data cleansing.

Maxwell Rebo:

Part of the challenge in wider adoption of AI automation, is a lot of folks have been conditioned to believe that you have to have terabytes and terabytes of data to start doing anything interesting or anything even remotely useful, and certainly sometimes that's true.

Sometimes you do see an impetus for absolutely massive volumes of data but a lot of the time the diversity of data and the veracity of data is really critical because what ends up happening is - at the end of a successful project we find that a lot of teams come to realize that they can actually automate much more complex tasks from these systems with relatively small or let's say moderate amounts of data.

Much more than they would have ever believed at the beginning of the project, so I think that's really the big message that I would continue to hammer on for - a lot of times less is more but you've got to have that creation process in place that creates quality data pipelines.

Understanding the Multi-V Model as a Framework for Managing Data at Scale

It's also important how we talk about bias in AI systems in the context of diversity of data and how the Multi V model comes into play.

The multi-V model is a framework that describes data at scale, characterized by multiple V's

Volume: The amount of data.

Velocity: The speed at which data is created and moves.

Variety: The diversity of data types.

Veracity: The quality and accuracy of data.

Variability: The degree to which data changes in structure, type, or meaning over time.

Validity: The accuracy and quality of data, ensuring it is clean & reliable.

Volatility: The lifespan and relevance of data.

Value: The value the data provides.

Bias in AI Can Be Both a Challenge and an Opportunity

Andrew Turner:

AI models are often criticized for bias, but recognizing and addressing bias can create opportunities. Insurers and financial firms can differentiate themselves by developing AI models that are fairer and more transparent, which can increase customer trust and loyalty.

Identifying and mitigating bias not only helps in compliance but also leads to better, more inclusive decision-making in areas like loan approvals or insurance underwriting.

The reason we mention it is because obviously if you look at responsible AI, the whole practice of developing it within the context of ethics, safety and society, sometimes it can be handled in the right way but it can also pose danger when handled in the inappropriate way.

By 2025, there is going to be 175 zettabytes of data with over 80% of it unstructured. How do you manage this huge growing mountain of data?

What are we seeing in terms of automation as a leg up for this whole data curation problem on the background of an ever increasing amount of data?

Maxwell Rebo:

There's a lot of fear around data bias because, of course you know every so often you see something pop up in the news about some business who messed up and were facing fines or other issues. Dealing with regulators can be very challenging but a lot of times what we find is that the bias issue really only shows up strongly in some domains - particularly when you're talking about something that makes decisions that really affects individuals.

Actually relating to other types of data that are more innocuous in terms of their outcomes, bias can actually itself be a signal that can be used effectively and sometimes even monetized.

A simple example we found in a cyber security use case for a fintech company, a lot of their previously dark data had surfaced and it was biased in the sense that it only applied to certain types of analysts. There was a demographic segment within the dark data that was telling a story about this particular demographic.

They said we need to go get other data that covers the other analysts. It turned out the segment was previously pretty well understood. The bias told us something really useful about what was going on with that type of analyst so that's the main thing I would highlight - bias is not always as bad as it's made out to be. Sometimes it can actually be useful.

Andrew Ahn:

I have a slightly different take or a variation on that, which is - everybody's concerned with the results in terms of KPIs and metrics but when you're dealing with this amount of scale and complexity, you can't bolt on compliance and governance at the end of a process. It has to be incorporated into the whole pipeline at the beginning as you're classifying and labeling the data.

Doing this in the right sequence, will yield huge amounts of benefits later on. In fact, what happens is the resolution of the data (similarly to resolution of digital imaging) - once it's summarized and aggregated, you can't ever get back to that higher level of resolution that you started with. Otherwise it's just like painting, 90% of a painting.

That follows through with a lot of different things like cooking and everything else as well as analytic work - you really need to start off with the right stuff, the right ingredients. This is really the answer for the garbage-in-garbage-out problem - it has to start with the right data.

Shift Left Strategies Can Improve Data Quality from the Outset

A popular concept in DevOps circles, “shift left” approach has recently made itself present in the data management sector as well. By integrating data quality checks and governance early in the process, insurance and finance companies can prevent costly errors that might affect AI performance later. This proactive approach reduces the need for data rework and ensures that AI models are built on reliable foundations.

The “shift left” strategy involves improving data quality early in the data lifecycle, rather than fixing issues downstream.

Roberto Maranca:

Having worked in financial services taught me that when you talk about regulation or in general when you talk about trust, and you want to guarantee an outcome is compliant and resilient, it has to be built in the pipeline and the only way to do that is to employ “shift left”. Meaning that when you think about something that needs to be delivered, everything has to be done more or less at the design stage so you need to design for better quality, you need to design for responsibility, you need to design for whatever you need to be part of your outcome.

Otherwise you're never going to achieve the ideal outcome at the end point of consumption. That’s what “shift left” means to me, so instead of remembering about data when you’re actually going to a testing stage and realizing the data is not good, you should have designed for that in the first place.

“**Federated learning** allows businesses to train AI models on decentralized data without moving it to a central server, addressing privacy concerns. This is especially valuable for industries like insurance and finance, where data security is paramount. By using federated learning, businesses can collaborate on AI projects, while keeping sensitive customer data secure, helping them develop more advanced models without violating privacy regulations.”

Maxwell Rebo:

One of the things I'm really excited about and have been tracking for a while is Federated learning right so quick primer for the audience so that's basically where instead of bringing the data to the model you basically bring the model to the data. What that does is then you know you've got a bunch of data in different areas, different networks that can never ever leave that network.

In highly regulated and high security environments that can be really helpful. It can alleviate a lot of logistical challenges around data movement and getting sign-offs to send data to a certain system or to move it off this network.

Jon Hammant:

I find it really interesting when you start to think about how intelligence works at scale. Fundamentally, fractals are a useful part of nature.

Nature is really good at inventing things which are very efficient and very good at scaling. It doesn't really invent them, it just discovers them as weird as it sounds. If you look at how nature has been optimized within intelligence and how intelligence has been optimized within nature, distributed centers of intelligence.

Whether you look at that within your cells, within your individual neurons and then it also happens at a far larger scale, at societal level.

Here the most intelligent thing that we're aware of is human society. It's not any one piece within the society, and I think similar things to that'll probably happen in the evolution of neural networks.

I'm a massive non-believer in this idea of Winner Takes All LLM. I think that'd be a really bad thing to happen because it would seem like a hideous dystopian future of one LLM dominating. Fundamentally, where all of this is going is different LLMs trained with different amounts of data that are good at different things leading to collaboration.

There's a number of data privacy challenges, that sit within the context of this conversation. Privacy is one of the key things that people are looking for today and in the future.

If you use an LLM, Zephyr, Mistal or Llama or whichever one's coming next and you want to make it special for your company, the reality is if you're shipping an LLM product with a slightly custom prompt, I don't think it's a huge value.

All of the value in AI/ML comes with the data that you add into it, so you know what you're doing is you're taking advantage of these large model providers who are spending very large amounts of money to invest to train their LLMs, and then you're effectively supplementing all of that data within GraphRAG or alternatively you're fine-tuning that LLM and you're depending on what your use case demands.

One of the real challenges that we definitely see is making sure that you don't accidentally leak financial reports or other sensitive data stored somewhere in your knowledge base that realistically only a very small amount of people are supposed to see.

"**GraphRAG** (Graph Retrieval-Augmented Generation) is an invaluable capability for data management that handles vast amounts of customer data. By integrating graph databases and AI-powered data retrieval, GraphRAG enables the company to organize and access interconnected data more efficiently. This approach not only ensures that customer information and related data points are easily retrievable, but also enhances decision-making by providing comprehensive insights in real-time."

This new AI world enables the business value to be driven out of best practice in data classification, and data security because the more you can understand about how data is classified, the more it really helps drive the value.

We're increasingly seeing people with almost a hybrid approach - it's when you look at the different ways of training these AIs - so they will fine-tune some of it to almost that generic level depending on the data classification. Then they'll use a graphRAG on top of it for the more granular control they're looking for. I think that's just going to become increasingly key to all of our customers.

There's the privacy of the response and then there's interacting with these systems - what is happening to the data you're submitting. Increasingly, people will want to know data transparency, and this is where regulation comes into play.

Today it's October 2024 and this is the worst AI is ever going to be within our lifetime. It is only getting better. You see the progress that's happening and that almost reminds of the Moors law - scaling that's happening at the moment is incredible, and I think fundamentally the thing that's going to drive all of that value, is going to be the data that sits underneath.

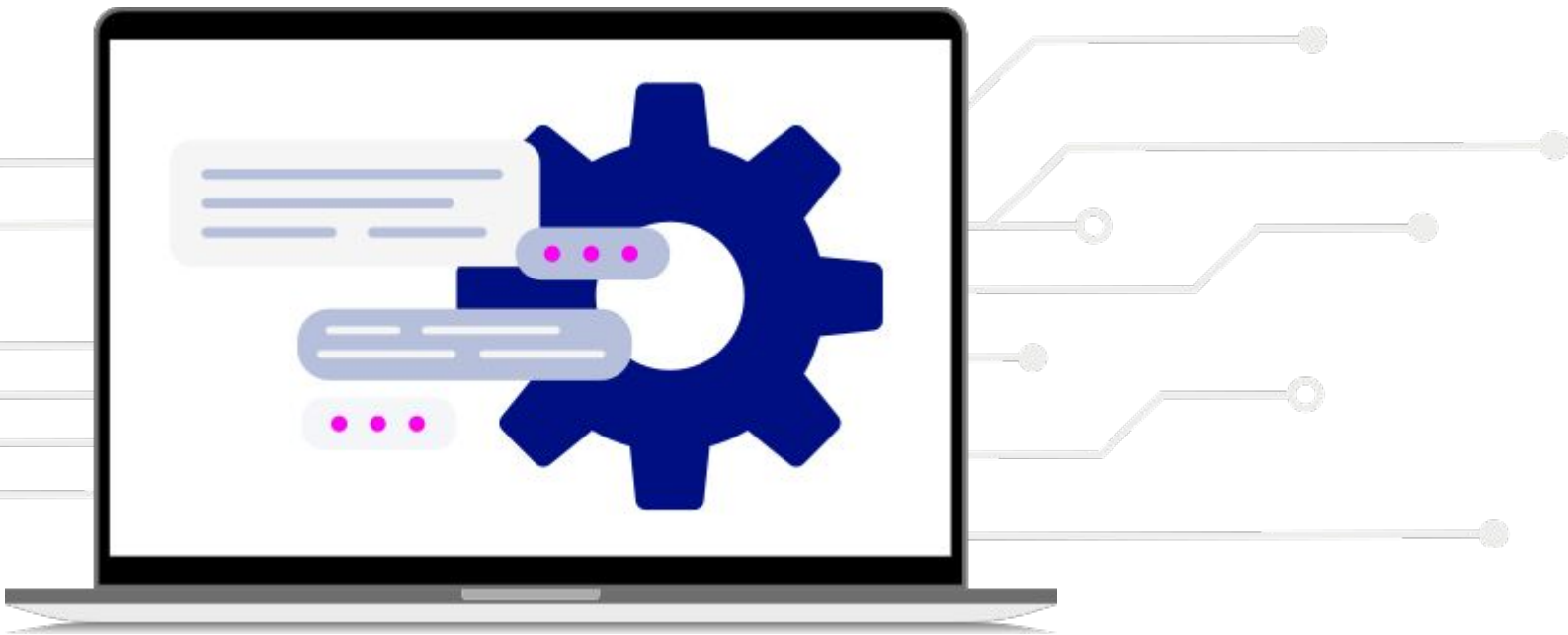
Andrew Ahn:

There is this interesting notion of a complex object. An analyst, a human being would be able to see a bunch of data and know that if you have some transactional data with IP addresses over there and you have a customer list that has last names, if you combine those two, suddenly you've de-anonymized this data and now it's a privacy issue.

It's a big deal, so the problem that we face is this is not a simple thing that you can just do and forget. You need to do it at a very, very large scale so you need to automate the data privacy process and it has to have a score.

What's more, this needs to be constantly refreshed, because the regulatory regimes are constantly being updated, and a lot of companies don't have the resources to constantly update and manage privacy stuff. Yet the same companies now use more and more data. So if you take the small to medium-sized insurance companies, their ability to analyze data is how they get their competitive advantage.

They may not have the revenue or the management of capital that the larger insurance carriers have but they really, really need to leverage that data so this is precisely the crossroads where we need to have better AI, and data curation with automation.



Guidance and Considerations Thinking About Responsible AI

Understanding Data Provenance is Key to Maintaining Trust in AI

In industries that handle sensitive information, knowing where the data comes from (data provenance) is essential for maintaining trust in AI systems. Without transparency, it's impossible to ensure compliance with regulations. insurers and financial firms need a clear chain of data ownership to ensure that their AI-driven decisions are both ethical and legal, especially when handling customer data.

The Historical Context of AI Informs Current Practices and Challenges

AI development has evolved from rule-based systems to deep learning, but the past challenges - such as managing large datasets and handling unstructured data - are still relevant. Today's CEOs should recognize that while AI is more powerful, managing data effectively remains a core challenge. Businesses need to learn to navigate data integration and regulatory hurdles, ensuring AI continues to deliver value.

Data Frugality is Important for Sustainable AI Practices

AI doesn't always need massive amounts of data. With data frugality, companies can train AI models using smaller, high-quality datasets, reducing costs and computational resources. This approach is particularly relevant for businesses handling sensitive data, as it minimizes exposure while still enabling accurate predictions. By focusing on frugality, companies can scale AI more sustainably and reduce the environmental and financial costs.

The Future of AI Will Increasingly Rely on Effective Data Management

As AI becomes more integral to business operations, effective data management will be critical to its success. Data will continue to grow in volume and complexity, but the companies that prioritize data governance, curation, and privacy will be the ones that lead the industry. For CEOs, the future of AI lies not only in technological advancements but in their ability to manage, protect, and leverage data to drive growth.

Fascinating Facts about AI and Data Management

You might think things will calm down, unfortunately not anytime soon. Here are some key insights and facts about AI and how effective AI data management can have an impact on your business.

- AI is expected to contribute **$15.7 trillion** to the global economy by **2030**, driven by innovations in data processing, automation, and machine learning.
- Gen AI is revolutionizing data management, with organizations using AI to enhance product development, risk management, and supply chain processes. These organizations attribute up to **20% of their EBIT** to AI.
- The AI solutions market is projected to reach **$1.81 trillion by 2030**, underscoring the explosive growth potential of AI technologies across all industries sectors, but particularly where regulation is mandatory.
- More than **50%** of businesses, see AI as a critical tool for enhancing customer relationships.
- Federated learning, an AI technique that enables decentralized data management, is gaining traction for its ability to improve data privacy while still training high-performance AI models.
- AI in data management can reduce processing time by up to **60%**, particularly in industries like insurance and finance, where massive datasets need to be curated and analyzed.
- AI-powered data analytics can increase business productivity by up to **40%**, providing deeper insights and enabling faster decision-making for companies dealing with large datasets.
- Only **21%** of organizations using AI have established policies for generative AI, highlighting a gap in the governance of AI systems and potential risks related to data management.

Praxi will continue to research, collaborate and publish information and guides to support you and your businesses effective adoption of these AI technologies that are available today and in the future.

We look forward to your feedback and continuing the conversation.

PRAXI.AI

AI enabling total Global Economic Impact of $15.7Tn by 2030

About Praxi

Focused and Deep - Praxi offers a unique approach to data discovery, profiling and matching by using pre-built industry specific models, in a collection known as a library. This allows businesses to easily search, find and analyze data out of the box, without the extensive training required by traditional methods. Praxi's solutions enable immediate automation and understanding of data from day one.

Sign up for Praxi CaaS [Curation-as-a-service] to try-before-you-buy with your own data

https://www.praxi.ai/curation-as-a-service

Let's keep in touch:

https://www.praxi.ai/learn-more

PRAXI.AI