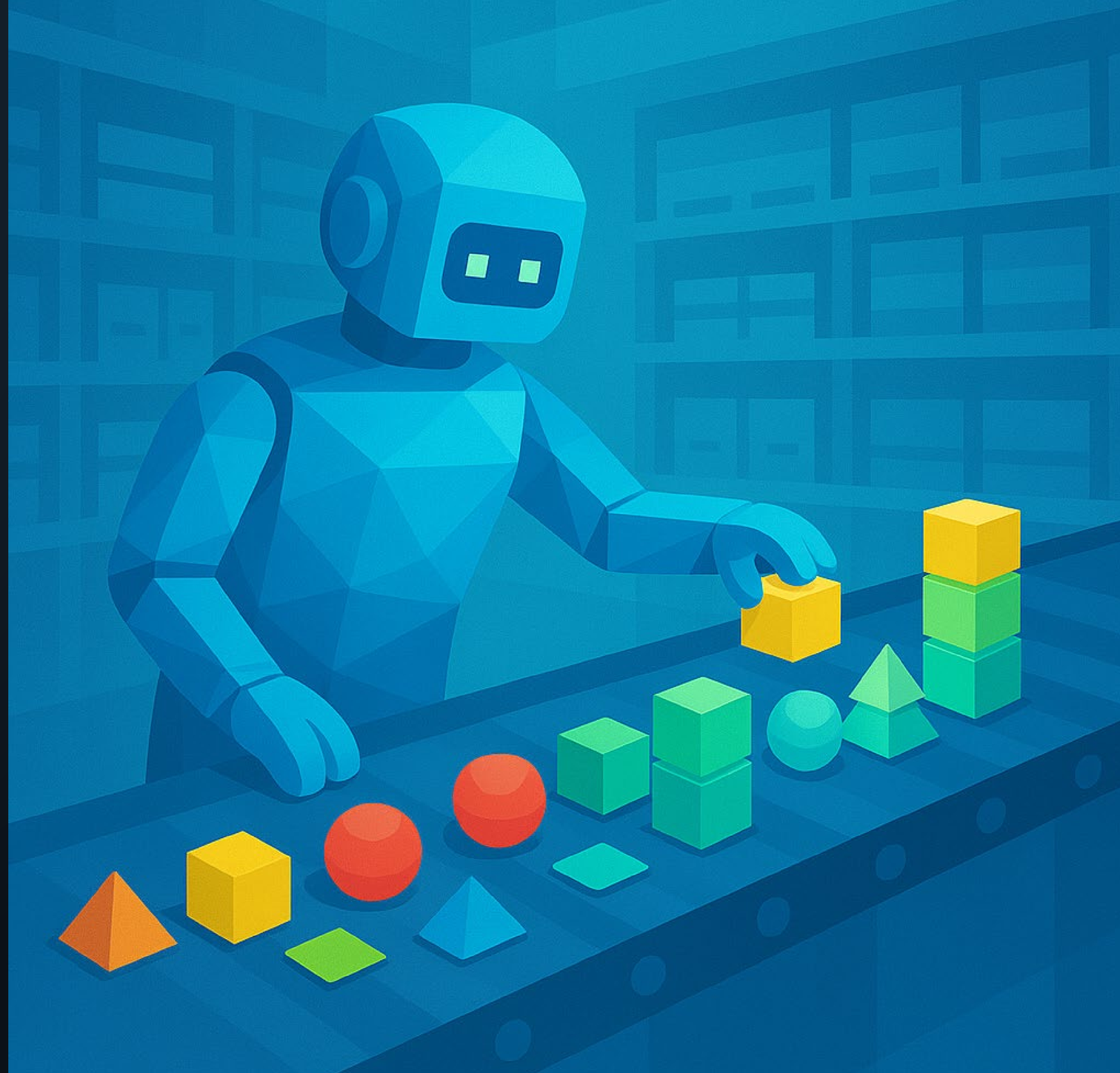
How Agentic Data Curation is Transforming Data Mastering

Perspectives from the Tamr Co-Founders

Tamr



Foreword

Sundar Pichai, CEO of Google, said, "The future of AI is not about replacing humans, it's about replacing human capabilities." To me, these words capture the essence of the seismic shift happening in the world of data management. GenAI—and AI agents—are proving to be a transformative force in the way organizations worldwide approach data management.

Take messy data as an example. It's a pervasive problem. For many, the issue stems from the inability to scale. Imagine for a moment that your task is to compare 10 million customer records to find the duplicates. Doing so would require you to make 50 trillion comparisons, a task that is extremely difficult and too time-consuming and expensive for humans to tackle.

Using traditional, rules-based data management techniques may get you 40-60% of the way there, but at a significant cost and with great complexity. Tamr's machine-learning-centric approach, on the other hand, gets you 90-95% of the way there quickly and cost-effectively with pre-built models. And because the models are probabilistic, Tamr can escalate the 5-10% of cases that would most benefit from human feedback.

But today, all that's beginning to change. Generative AI-based agents are rewriting the rules of data curation and data mastering. They're learning, adapting, and bringing clarity to even the most complex data environments, enabling what once required extensive manual effort to happen in real time and at scale.

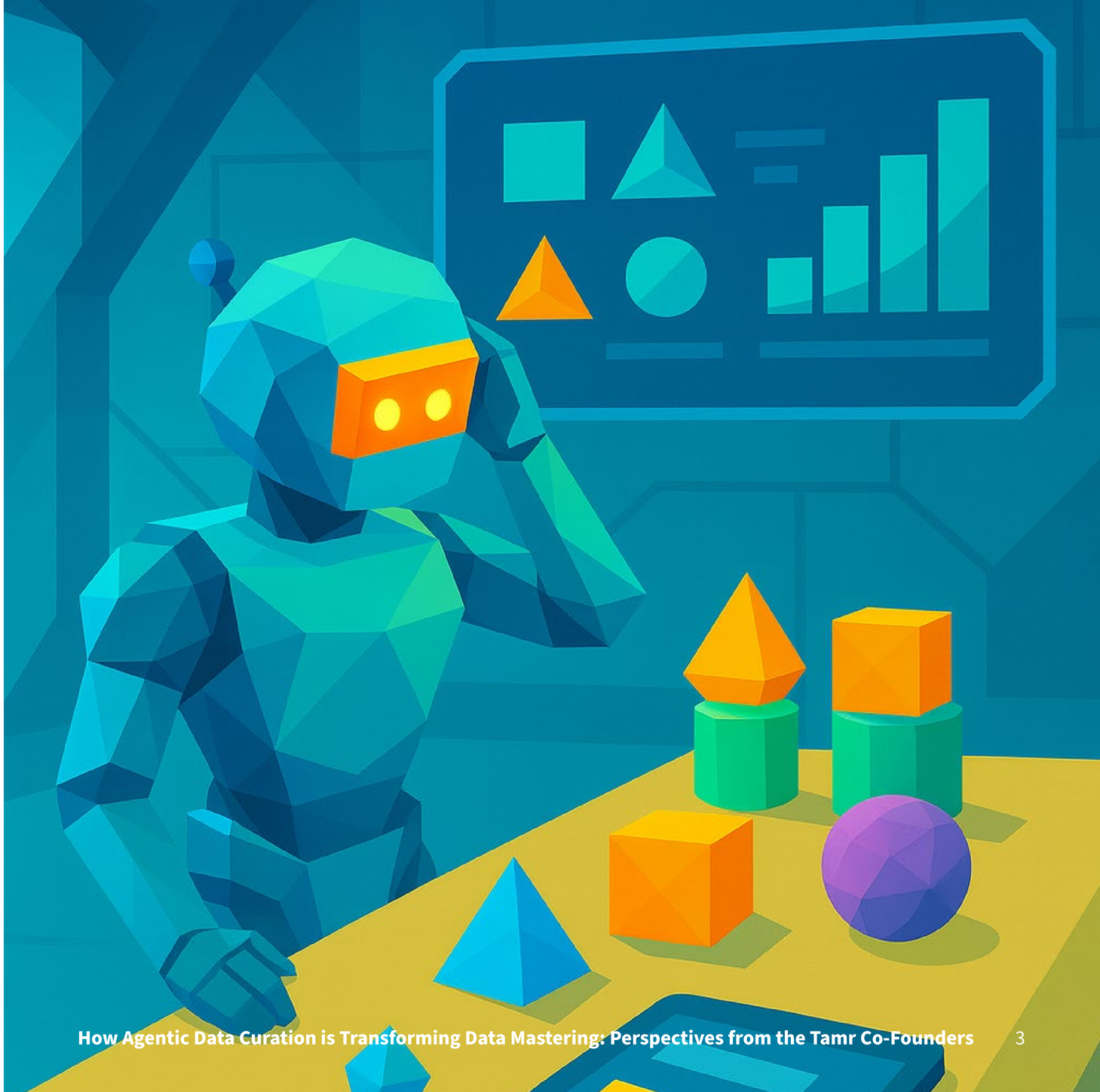
With an extensive portfolio of patents, Tamr has long been a pioneer in data mastering, employing AI to address difficult data management challenges. Now, we're taking the next steps toward using AI agents to solve some of the data curation challenges that have continued to require human intervention. As we reflect on these exciting times and what lies ahead, who better to share their perspectives than the original Tamr founders? Not only do their insights remind us how far we've come, but they also inspire us to keep innovating with the same vision and passion that started it all.

Anthony Deighton
CEO, Tamr

Every organization has experienced it—disconnected systems, duplicate records, and data that never quite tells the full story. The cost of this chaos runs deep, slowing decisions, obscuring insights, and even jeopardizing day-to-day operations. And as data continues to proliferate, the chaos will only intensify.

For decades, organizations have attempted to unify, curate, and clean their data using traditional master data management (MDM) solutions. But these solutions come at a cost: rules upon rules that are difficult and costly to maintain and impossible to scale.

Solutions like Tamr take a different approach, employing machine learning (ML) to clean and curate data, and these solutions have made a big difference. ML models can identify duplicates, fill in missing values, resolve inconsistencies, and produce golden records far better and faster than any team of humans or traditional MDM can.



With its power, intelligence, and efficiency, ML generally takes organizations 80-90% of the way toward achieving fully clean, trustworthy data—something rules-based systems alone can never achieve. Incorporating some select organization-specific rules alongside ML can resolve certain additional situations, but it's the last mile—**the 5-10% of data that remains unresolved after the mastering process**—that still requires a higher level of knowledge, precision, and data preparation. And even though the last mile represents a small portion of the data, it typically consumes upwards of 80% of the data team's time to resolve.

Agentic data curation is a new concept that has the potential to revolutionize data management. Agentic data curation uses LLM-based AI agents to automate more of the data curation process by capturing and acting on the contextual insights needed to make confident curation decisions.

AI agents have the capacity to intelligently clean, curate, manage, and refine the last mile of enterprise data—the part that addresses the idiosyncrasies and complex edge cases that are close to consumption and difficult to decipher—with minimal human intervention. By comparing outputs of entity matches and explaining the reasoning behind why records are (or are not) a match, AI agents can provide the preliminary analysis humans need to determine if they trust the AI's output or if they need to tune the model further.

To explore the rise in agentic AI and its potential to transform MDM, we sat down with Tamr's six co-founders, which include renowned academics from MIT, to hear their perspectives on these topics. Throughout our conversations, we examined what agentic data curation is, why it matters, and how it will transform modern data management. We also explored what's ahead for the next 12–18 months, as well as predictions and insights on the future of data management.

Learn from these experts.

Tamr Co-Founders

Andy Palmer
Entrepreneur, Investor, and Tamr Chairperson

Ihab Ilyas
Co-Founder and CEO, hiddenweights

Alex Pagan
Chief Architect, Tamr

George Beskales
Distinguished Software Engineer

Daniel Bruckner
Chief Technology Officer, Tamr

Michael Stonebraker
MIT Professor of Computer Science and Turing Award Winner

The Insurmountable Becomes Surmountable

Andy Palmer

“I believe that agents are going to replace most—if not all—of the self-service data prep that’s out there.”

For years, companies have been trying to make sense of the vast amounts of disparate data within their organizations. And the problem was so big—and the data was so diverse—that it caused, what Andy Palmer calls, “a dumpster fire of epic proportions” that was impossible to solve without a machine.

Andy explained that when the team founded Tamr, they built the solution with machine learning (ML) at the core. And it had one goal: to make an organization’s corpus of data better by automating three key functions:

1. **Schema mapping:** Define how data in one table or database corresponds to data in another.
2. **Record matching:** Determine which records are the same or different.
3. **Values selection:** Choose the right values in each field for each record.

And while machine learning has worked well for the broader corpus of enterprise data, the last mile has still remained unresolved. Humans needed to step in and address idiosyncratic changes and tune the data for their specific needs. This problem—solving the last mile—was always so overwhelming and complex that it felt insurmountable.

Andy went on to share that AI agents, on the other hand, are poised to play a significant role in the mastering process, especially when it comes to self-service data preparation and data curation. Using agentic data curation, AI agents can bridge the gap between data curation and last-mile data preparation by automating the adjustments and preferences users typically make, ultimately reducing the amount of manual work. By observing the actions humans take in data preparation, AI agents learn about the idiosyncratic changes users apply and can then automate these changes, acting very much like a macro on steroids. And because agents will begin to automate some of the manual prep work previously done by humans, data teams will be able to shift their focus to spend time closer to the business.

The time is right for AI agents. And as they evolve, Andy believes it will be very natural for agents to take on more specialized and complex tasks such as updating outdated addresses in a large database like Snowflake or Databricks using public, external data sources like Dun & Bradstreet or ZoomInfo. Agents will have the ability to resolve attributes, deduplicate data, and complete fields with null values. And if the agent doesn't know or isn't sure what update to make, it will be programmed to engage directly with the data owners to validate values that remain unclear.

"We're in the next phase of the evolution of artificial intelligence where we can look at things in a way that is more specialized and built-for-purpose. We're evolving to where the LLMs have done everything they can do generally, and now we need to go deep into specific areas so that the output of the systems and agents can be highly contextual and the quality much higher."

Andy Palmer

Redefining “Human in the Loop”

Ihab Ilyas

“Certain tasks require larger context. And traditionally, these tasks have been hard for humans to solve. Now, because of their ability to munch on bigger context such as descriptions and feedback, AI agents can handle harder curation, mastering, and cleaning tasks. And that’s where we’re seeing a breakthrough.”

Machine learning has been at the core of data mastering for many years with Tamr, replacing manual tasks previously done by humans. But as AI's ability to process larger contexts grows, so too does its ability to perform harder curation, mastering, and cleaning tasks. And this has the potential to shift what humans do.

Ihab Ilyas shared that when the team founded Tamr, “human in-the-loop” was a core principle. Humans were in the driver’s seat. They had control. But this control came at the expense of scalability and automation—at least partially.

Ihab went on to discuss that now, agentic AI is giving us the opportunity to redefine “human in-the-loop.” AI agents can handle popular or familiar cases that are known but highly complex, which requires humans to remain involved and maintain a level of control while still doing less.

So what does this look like? Ihab explained that if we take a look at the curation pipeline, it comprises three distinct processes: data prep, data mastering, and post-processing. AI can help at every point, which presents a great opportunity to rethink the end-to-end pipeline. For data prep, AI embodies a data scientist skillset providing capabilities such as pivoting, removing anomalies, and standardizing data. Within data mastering, it handles specific cases that require specific context. And in post-processing, AI supports answer generation, transformation, and generative AI capabilities through AI agents.

By using agentic AI, the data pipeline will become more accurate, more scalable, and more efficient, removing any excuse for inaccuracy.

“We can no longer be satisfied with 85% accuracy. With AI agents, we can now reach ridiculously high levels of accuracy. And this increases the bar of the customer's expectations of what you can do.”

Ihab Ilyas

The Right Tool for the Right Job

Alex Pagan

“Having an AI-native platform that can deploy the right tool at the right time is key to being successful.”

AI agents are beginning to play a pivotal role in data mastering and data curation, backfilling a human’s work to enhance data quality, particularly when it comes to resolving quarantined records that fail to meet baseline criteria within a traditional MDM solution. However, the hardest problem with data curation is knowing what you should look at and where the problems exist. As Alex Pagan discussed, first, you must start with some indicator of what requires attention and then make a decision and suggest some corrective action for the record. And while agents often replace simple tasks like the manual processing of exceptions, this may be an example of where the tool is too complex for the job at hand.

Instead, the greatest leverage point for AI agents is their role in cases that are specific to an organization, such as their data quality standards. From an outside perspective, these standards may look subjective, but inside the organization, they have a very specific meaning that requires judgment, context, and a broader perspective on what the data entails. AI agents can tap into a broad array of tools to improve data quality, including existing rule sets, external data, and purpose-built models. More importantly, they can synthesize this context to make judgements even in ambiguous or conflicting cases.

Alex went on to share that, for example, it’s very easy for a human to spot when a first name and last name are transposed but very difficult to write a rule to detect it. Yet because AI agents have the ability to make judgements, they can spot a transposed name and flag it for resolution.

Another example involves the question of how to handle records with incomplete or conflicting information. In this instance, users need to apply judgement and conduct research to determine how to resolve this specific case. AI agents can add value by doing the legwork of synthesizing relevant data from internal and external sources, presenting the findings, and then streamlining the process to get those changes incorporated. It's these types of scenarios where companies get the greatest uplift.

“If you want to be realistic about clean data, you need to understand success isn’t black or white. Apply the right tools to get to the right level of quality.”

Alex Pagan

AI Agents as a Judge

George Beskales

“I don’t imagine a large language model replacing the [machine learning] we have today, but there is room for distilling the information from an LLM into conventional machine learning, placing agentic AI or GenAI as a judge or a curator.”

AI agents have the potential to play a role across multiple stages of the data mastering process, from data acquisition and preparation to curation, verification, and publishing. Currently, humans provide much of the feedback and reasoning used to train AI models and validate the outputs. But as George Beskales shared, with the genesis of AI agents, we now have the ability to supplement human feedback with common patterns identified by AI agents, as well as apply reasoning about why records do—or do not—match, which allows us to speed up the validation process and identify which areas to tune.

George went on to emphasize that this is especially important in the last mile of data mastering. AI agents can serve as “the man in the middle,” or the layer between human interaction and MDM data products. Agents can take all the information and all the rules and come up with a way to train the underlying machine learning so it can surface edge cases and provide recommended actions to resolve them.

He also highlighted that data enrichment presents another opportunity for AI agents. By continuously crawling the web and other external sources, AI agents can augment both structured and unstructured text with additional columns or insights from reviews, news articles, and other credible sources, further enhancing the ability to improve the quality of enterprise data at scale.

“Because the 10-20% of entities that remain unresolved after the general mastering process are generally the hardest to solve and consume the majority of a human’s time and effort, reducing the number of entities by even 1 or 2% is a huge gain.”

George Beskales

Ranking the Rankings: How Agentic Data Curation Works

Dan Bruckner

“AI agents fit in between machine learning and human involvement, providing an even more intelligent method for reviewing cases, as well as acting as an assistant and productivity tool for humans.”

Dan Bruckner explained that unlike traditional approaches, AI agents can learn, reason, and act autonomously within defined parameters, both general and organization-specific, allowing them to continuously monitor data flows and apply context so humans can spend less time on initial validation and more time analyzing and resolving complex edge cases. By organizing outliers into a prioritized queue of work, AI agents make it easy for humans to quickly review the records and determine how to move forward.

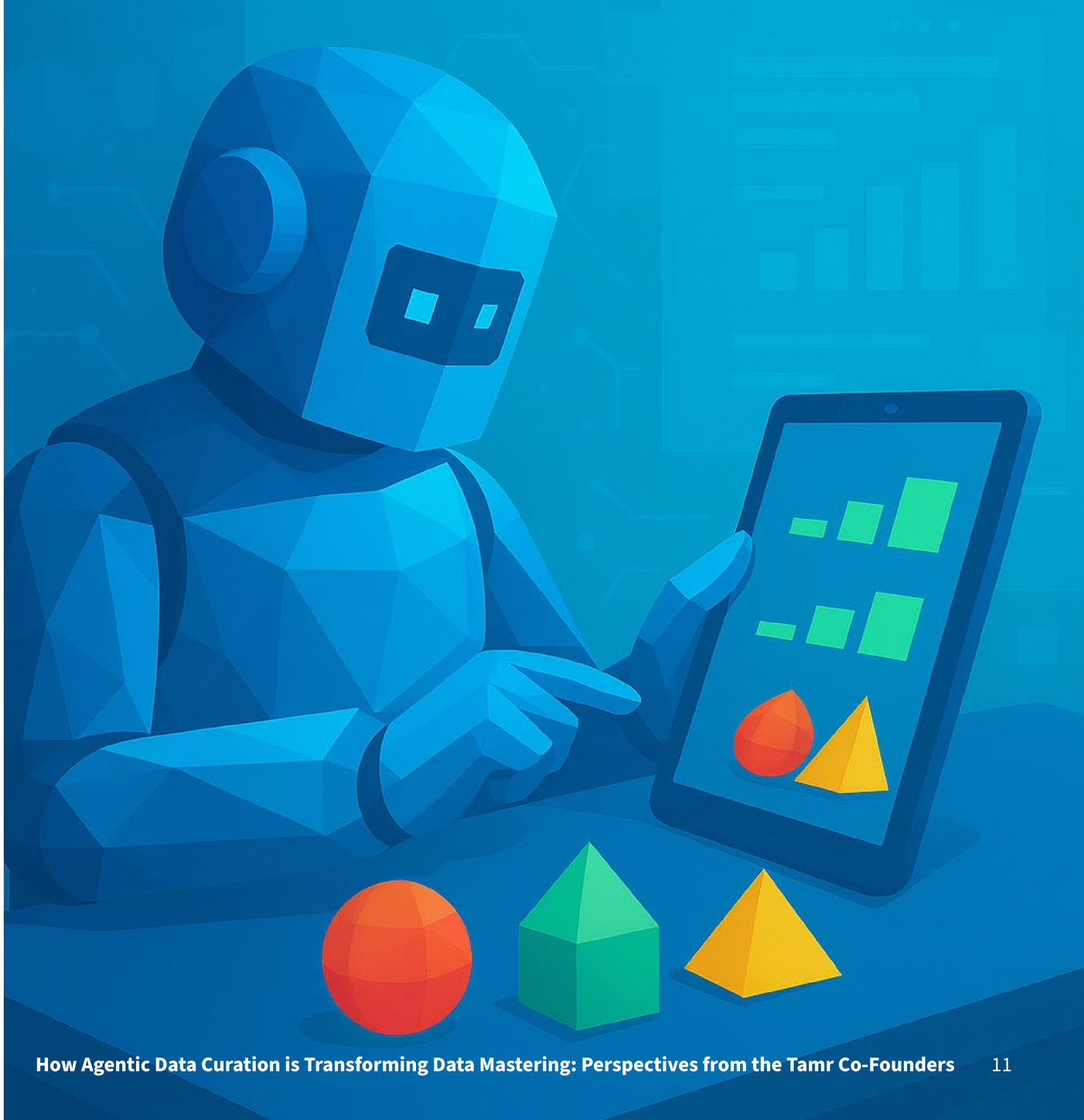
Using insights and KPIs collected by the system, such as real-time metrics like popularity, views, or number of edits, AI agents can flag priority records and move them to the top of the queue, essentially “ranking the rankings,” as Dan described it, so humans can focus on resolving the most important cases first.

Dan went on to discuss that AI agents can standardize data by surfacing customer records that are inconsistent based on company-defined data quality standards. For instance, AI could identify contact records that are missing phone numbers or clusters of records that include people with different last names and flag them for human review. Or, the agent could surface ambiguous cases such as records where the first name and last name are transposed, analyze them, and offer an opinion on what action the human should take. In this case, the human’s role would be to accept or decline the suggestion.

As confidence in AI agents grows, Dan believes that organizations can scale curation efforts by giving the agents permission to apply certain data quality changes directly to the records, especially when those changes are straightforward. But as the use of AI agents continues to rise, he also emphasized that explainability becomes increasingly important. Organizations want to understand not only what decision the agent made, but also why it made it. And because AI agents have the ability to reason and express their rationale in natural language, their decisions become easier to understand.

"AI agents allow organizations to expand the breadth of data under management. Not only will this allow smaller teams to deliver more value to the business faster, but it also enables them to expand the number of use cases."

Dan Bruckner



Machine Learning Isn’t Going Away, But…

Michael Stonebraker

“Will LLMs work on enterprise data integration? I don’t know, but it’s not beyond the realm of possibility.”

According to Michael Stonebraker, when it comes to data mastering, machine learning is the only thing that scales. And it’s not going away.

That said, he also thinks that large language models (LLMs) could play a role as well. However, they have some limits. Michael and his team of researchers at MIT tested an LLM against an MIT database with 1,400 tables. They took 200 queries in SQL, and with the help of real users and database admins, produced the natural language that corresponded to the SQL. Then, they handed those 200 queries to an LLM, which spit out the SQL and ran that against the data warehouse. And the accuracy was zero—it didn’t work at all! As Michael shared, there are a few reasons why.

First, LLMs are trained on, what Michael calls, “the pile” or publicly available web pages, and the MIT data warehouse is not publicly available. So the adage is, if the LLM hasn’t seen it before, it’s not going to get it. And the team proved that was true. In addition, there is a lot of semantic overlap and idiosyncratic data in the data warehouse which confuses the LLM. And finally, the queries are complicated, with multiple joins and aggregates.

So the team tried a different approach. Using a database of 200,000 public, yet idiosyncratic, records from a now-defunct company, they used an LLM to unify two data sources. Initially, the accuracy increased to 75%. And recently, using ChatGPT 5.0, the accuracy increased to 90%+. As Michael acknowledged, that makes things interesting.

But Michael believes this question still remains: How accurate will the AI agents be when working with enterprise data?

> “If the 90% accuracy we saw with a public, yet idiosyncratic, dataset turns into 60% accuracy when we apply an LLM to enterprise data, then it’s not ready for prime time. But maybe it could be at some point in the future.”
>
> **Michael Stonebraker**

The Risks of Agentic Data Curation

While AI agents hold promise when it comes to data mastering and data curation, the Tamr co-founders noted that it's not without some number of risks.

Risk #1

Applying AI agents to your full data set

AI agents are powerful and potentially very beneficial. But beware of turning them loose on your entire data set as they can also potentially do a lot of damage such as incorrectly matching records or overwriting values, fast. And the more data they touch, the harder it is to verify when mistakes occur.

How to overcome:

Instead of unleashing AI agents on your entire data set, determine which tool is right for which part of the process. While AI agents can help to identify and provide recommendations to resolve outliers, ML and rules still play a big role, especially when scalability is a priority.

Risk #2

Losing control of the agents—and your data

AI agents are a smart tool, but it's important to remember that they are still a tool. And without the right governance in place, it becomes increasingly difficult to control their access and verify if the output is correct.

How to overcome:

Treat AI agents as partially trusted users, and give them the access they need to do the job at hand—and nothing more. It's also important to clearly define their role. Are the agents analyzing outliers and presenting findings to humans or making decisions and taking action on the data?

Risk #3

Hallucinations

Agents will get things wrong. Not only will they hallucinate, but different agents will give different responses, making it difficult to know which answer is right.

How to overcome:

The best way to protect against hallucinations and misinformation is to use multiple models and cross-reference outputs, as well as apply human validation when questions arise. Explainability is an advantage with AI agents because they can put their rationale into plain language, and you can interact with them when their rationale is unclear.

The Risks of Agentic Data Curation

While AI agents hold promise when it comes to data mastering and data curation, the Tamr co-founders noted that it's not without some number of risks.

Risk #4

Diagnostic challenges

Knowing an agent made a mistake is one thing. But understanding why the agent did it and how to fix it is another matter entirely.

How to overcome:

Implement clear data quality metrics to gauge if AI agents are improving the data quality problem—or making it worse. That way, you have a clear view into the impact, good or bad, that AI agents are having on the data and can audit their actions against what you expect.

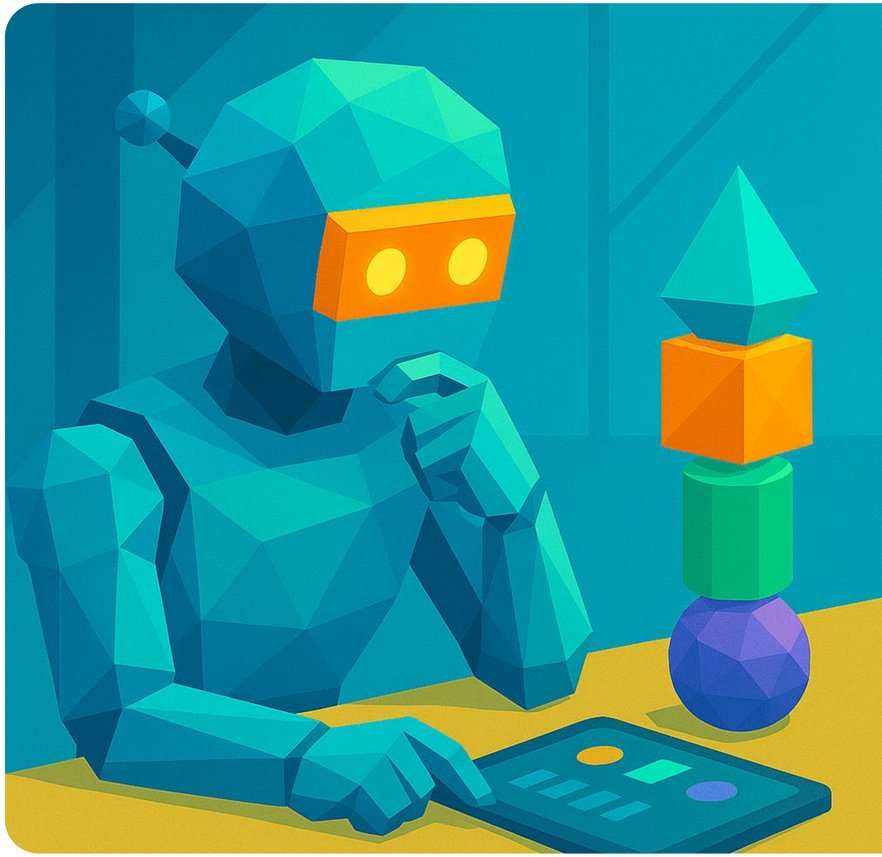
Risk #5

User skepticism

Even if AI agents get the answer right 99% of the time, users may remain skeptical about the information or guidance an agent provides.

How to overcome:

Start by giving the easier, lower-risk use cases to the AI agents. And if the case requires expert knowledge, make sure a human verifies the output before it reaches the users. That way, human curators will steadily gain confidence in the agents' output over time.

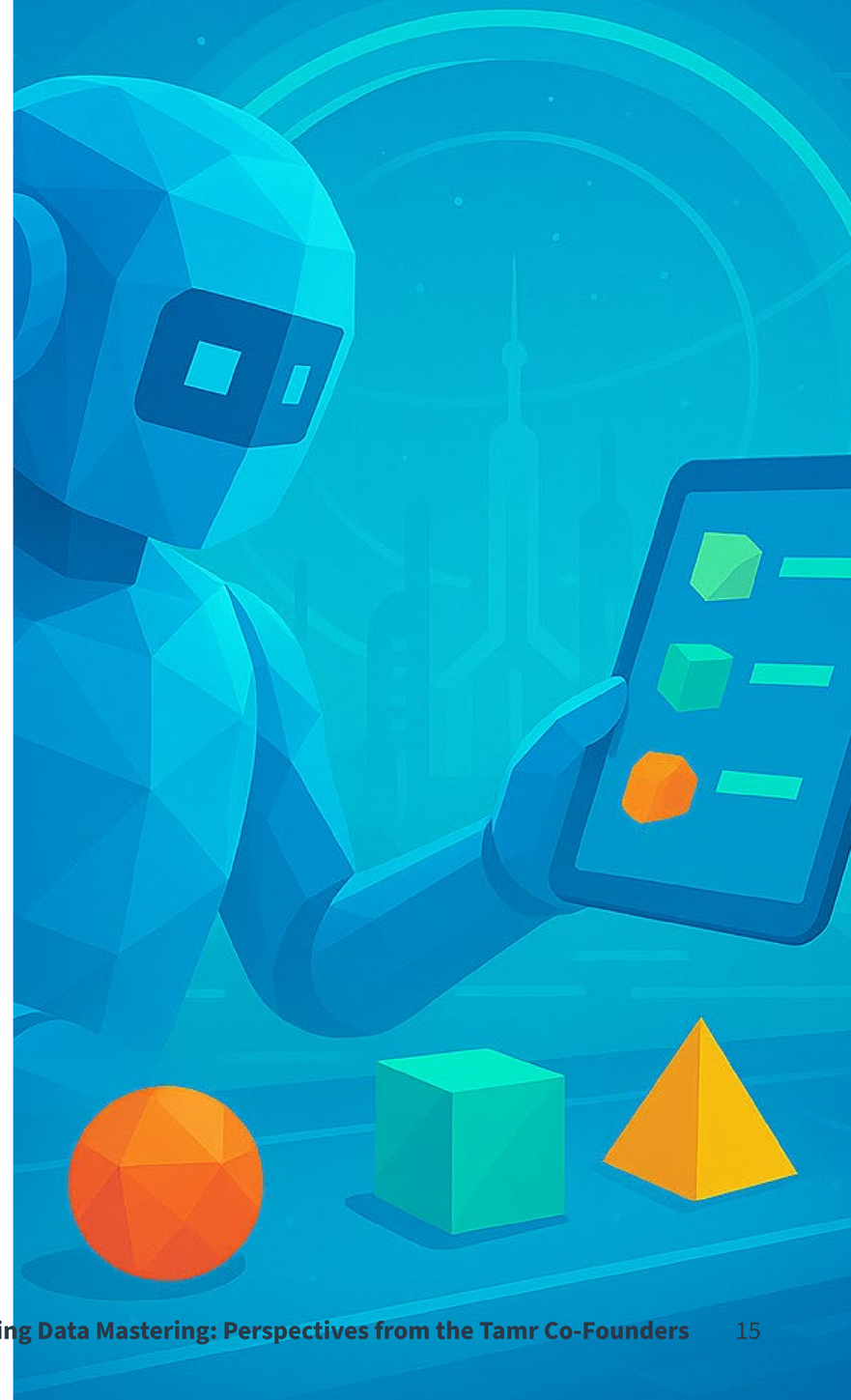


The Future of Agentic Data Curation

As we look to the future of agentic data curation, Tamr's co-founders believe AI will evolve to look at things in more specialized ways. AI agents will play an increasingly important role in the data mastering and data curation process, making it commonplace for them to be integrated into existing frameworks to take on critical data cleaning responsibilities such as address correction, attribute resolution, and data enrichment. The co-founders also predict that we will see improvements in the quality of the agents, but not necessarily an explosion of new features. And the question remains: How much better will LLMs get, given that we're already seeing their accuracy curve begin to flatten?

Further, Tamr's co-founders see AI agents enabling MDM solutions to move beyond structured data, allowing agents to incorporate unstructured data from documents and emails which provide crucial signals and context to help resolve more entities, faster. And as AI agents become more advanced—and explainability and attribution improve—it's likely that adoption and trust will accelerate, too.

However, the co-founders also noted that these advancements in data mastering through the use of agentic AI could come at a significant cost. LLMs are still quite costly. But what's expensive today may be more affordable in the future, prompting companies to rethink the economics of data curation.



Putting AI Agents to Work

Data mastering has always focused on delivering trustworthy, golden records for use in analytics and operations. And while an ML-based MDM solution like Tamr's can deliver 90-95% of the golden records an organization needs, automating the work to resolve the remaining 5-10% has seemed out of reach…until now.

Agentic data curation offers a new, promising approach to data mastering and data curation that enables organizations to improve data quality at scale while shortening the last mile curation of data closest to consumption.

Companies like Tamr have a distinct advantage in the market and are poised to take full advantage of AI agents through its AI-native platform. With Tamr's groundbreaking Curator Hub module, companies are able to deploy and control AI agents, appearing as users in the system, and enable them to play an active role in the curation process. Agents also have the ability to analyze and provide insights about unresolved data, review and resolve items in queues, and identify and escalate exceptions for human users to review. And that's just the beginning!

Want to learn more about Tamr, Curator Hub, and the future AI agents? **Schedule a demo today**.

Learn more at **tamr.com**

Tamr