Machine Learning:
Driving Financial Industry Transformation

Live Date: April 28, 2020

Featuring: Aidar Orunkhanov (Solutions Director, Tamr) & Walid Kara (Sales Engineer, Tamr)

Recording: bit.ly/2Wt4K4z

Presentation: bit.ly/2KVxNbE

EDM Council Homepage: edmcouncil.org **Tamr Homepage:** tamr.com

WEBINAR Q&A:

My organization is falling behind on the digital transformation curve, where should I invest my time to achieve quick wins and avoid getting overwhelmed?

The most obvious answer here is to start with the low-hanging fruit - a challenge that will deliver a tangible return the quickest and will require the least amount of effort. Data is at the root of many challenges' organizations face today, even when this connection is not obvious. There are quite a few examples that can deliver a tangible ROI. For example, spend management is an easy pick: figuring out where the money comes from and goes to will always be one of the most visible tasks. Spend management is a challenge that many procurement managers have been struggling with, not only in industries with supply chains and inventory, but in financial services as well. Another very prevalent challenge in financial services is reference data management. Reference data is used everywhere today - from enriching market data to onboarding customers to sales and sales outreach. It is surprising how few organizations are able to fully tame all of their reference data and put it into centralized locations. Applying machine learning towards data mastering opens up new possibilities to solve data problems that cause business inefficiencies - and machine learning allows organizations to prioritize solving tough projects like reference data management, freeing up resources and decreasing costs in the long run.

For the Canadian bank case study - what tool did you use, and did you utilize any external data set, i.e. D&B to validate the source of truth?

In this particular example, we cannot disclose the exact sources used by this bank. However, based on our experience, the most common external data sources utilized by our customers are Bloomberg, Refinitive, BvD, Moody's, GLEIF, Capital IQ, and D&B. I think OFAC lists and other government sources are other frequent external sources. Quite often enterprises use two or three of these sources, and a few others are on the "wish list". Unfortunately, bringing in any additional sources is a resource intensive task - both due to data quality and an exponential increase in new rules and exceptions. In the case of the Canadian bank, it was a similar situation where there were a few of these reference data sources and additional ones needed to be brought in. In this case, and in other cases where we needed to master external data sources and reference data, we applied Tamr and our machine learning algorithms to master existing sources and quickly onboard new ones. The main benefit of using a machine learning solution for data mastering is its ability to scale rapidly, as any new sources that bring their own intricacies do not require rewriting rules or algorithms.

How can I address and deal with model risk introduced from ML applications?

This is something that we can definitely see a lot in the financial services industry. What Tamr does is it's really giving you the metrics associated with the model development so you're always able to see what the accuracy looks like and you can go deeper into those metrics as well. Tamr's trying to make sure that we're keeping as much information as available to you. Essentially, when you're using this in an operational use case, if you ever get audited for what's going on in your database, Tamr has a full history of events of what's going on in your records and databases. You'll be able to pinpoint when things have changed, records have been clustered together, and removed. That usually satisfies that model risk question we get a lot in the financial services industry.

Is data that lives on mainframes considered by Tamr?

Tamr allows you to easily upload any data sources into the platform to be a part of any project you'd like. If it's a table with rows and columns, you can upload it to Tamr.

It is very difficult to keep track of the original source of a given Data point used in Analytics or Reports. How do you deal with Data Lineage?

While ability to master data is very important and is probably one of the main features that organizations look for in a data solution, data lineage is as equally important. As enterprises generate more and more

data, it is vital to understand how data pipelines work and what information is being fed into downstream systems. Knowing where any of the data comes from, how it was used, how it was modified, and where it went can help organizations streamline their data operations and make better data-driven decisions. Tamr requires primary keys for all data being ingested, automatically adding them in situations where a primary key did not exist in the original data. As the data is being mastered, Tamr further assigns permanent id's to clusters of entities, while maintaining an audit trail of how records enter each cluster. Each individual record within a cluster can then be traced back to the original source. Tamr's schema mapping capability allows easy mapping of datasets into a unified schema, while keeping a visual representation of the mapping by attribute and source.

The 'law of large numbers' does provide machine learning approaches with a sufficient amount of data needed to tune pattern recognition, but it does not necessarily solve the level of discrimination needed for unique identification and de-duplication. Doesn't matching require significant human review?

An interpretation of the Law of Large Numbers stipulates that with a sufficient number of outcomes, a machine learning model for mastering data will eventually reach an accuracy of 1. This process can be accelerated by having a human guide the learning rather than reviewing all the matching, by providing feedback on specific pairs to kickstart the algorithm tuning, usually through simple yes/no questions. Once an algorithm is trained initially, humans can then decide what level of confidence is required from the machine to perform certain actions. Humans then remain in charge of reviewing scenarios where the machine is not at the desired level of confidence, thus further fine-tuning the algorithm rather than looking at all pairs. Remaining data outliers can be further accommodated with rules to get to 100% accuracy. The main advantage of using machine learning is the fact that algorithms can accommodate small changes in the data - so when new sources are brought in, the algorithms already take differences into account. Whereas with a fully rules-based approach, new rules will have to be defined for all new variations, which may or may not be overlooked by humans, thus introducing the human factor as an important aspect of accurate matching. One of the things that Tamr has patented in its technology is what we call high-impact questions. That's Tamr's ability to surface the questions, when it comes to this deduplication effort, that have the strongest impact on the Tamr model. Tamr looks at how similar the questions are to other questions around similar records and by bringing those up to human experts and getting the feedback on things that Tamr isn't necessarily confident about, then you're able to provide the strongest impact on the model and by answering a few of these you can actually get to a pretty accurate model and fine-tune from there. The key takeaway here is that when you're doing this human

feedback with Tamr, that's replacing a lot of the time you would be spending developing rules to deduplicate your records. So, it might be easy to develop a good number of rules that would help you deduplicate about 50% of your data but when it comes to fine-tuning those rules, you're going to be looking at a lot of edge cases. With Tamr, you're spending your time giving feedback to the model instead of developing these rules. You'll be able to get a higher level of accuracy much faster based on what we've seen with customer implementations.

Using data to train AI models that includes historical bias is a big issue. What tools, techniques exist to test training data for bias?

In the context of mastering data, when we look at training data that comes to train the AI model and implement the solution, we look for specific data that encompasses what an entity means to you. This is going to be based on each use case and may differ even from similar use cases elsewhere. If we're talking about a "customer" entity, for example, internally to your organization, you're going to have a definition of what a "customer" is (for your organization). If this definition defines a "customer" on a legal entity level then so be it; alternatively, this definition can be more granular, such as on an address level - e.g. State Street in Boston at a specific building might be different than State Street in a different city in a different building. Here at Tamr, we designed a company around mastering data, and this is certainly a common use case. At Tamr, we use human-guided machine learning to overcome these challenges and, as we mentioned in some other questions before, we have humans respond to high-impact questions to train the algorithm. Continuing from the prior example, depending on what the definition of a "customer" is, your organization will be training the Tamr model by providing direct feedback in a yes or no form and confirming whether certain entities are the same or whether they are different. When we're looking at having a bias in the training data, we're not really gathering data from the external universe in order to train the model. We're actually taking training data directly from your team based on what your definition of an entity is in order to fix the deduplication problem.

When using AI for EDM how can you provide explainability and transparency into how decisions like clustering have been made?

That's a great question, and this is something we hear often from our customers in the financial services industry. The first aspect that often pops up is the ability to defend machine learning to the regulators. When using a fully rules-based approach, it's easy to defend against a regulator because a rule always does what a rule was designed to do. The primary concern is around machine learning acting like a black-box solution and whether it provides enough transparency to prove its effectiveness. Fortunately,

the regulators are coming around on the machine learning solutions - for example OCC, the FRB and others released a directive a few months ago that enterprises are encouraged to use more machine learning and AI in their normal operations. The key here is to have the machine learning model to be human-guided and have a configurable confidence interval to ensure the accuracy. As long as the machine picks up the grunt work, but was directed to do so by a human, then it is not much different from defining a specific task in typical software used nowadays. Now let's say a machine picks up X amount of work, and there's a configurable confidence level interval to what we think is good. If we configure the model to 99% confidence, there's still 1% that we're not confident about. This one percent is where a human can add a few rules to bring it to a full 100% and have that kind of response defensible to a regulator. This also brings us to the point how Tamr actually operates as an open-box machine learning solution where traceability is a big feature. So, for clustering, what we actually do is we have persistent IDs, which is what we call them in the product, and that's tracked throughout the life of each record and you're able to see when a record enters a specific cluster and for what reason. This actually allows you to track the lineage of a record and figure out when a record has been moved from one cluster to another and be able to understand why these things are happening in Tamr, and to be able to trace the source of the data for every single cluster.

You keep using the term 'machine learning' ... can you expand on what you mean by it (i.e. is this a neural network, genetic algorithms?) and what is the lead time in terms of creating and preparing training data sets?

Tamr uses patented proprietary models (that include neural networks among them) in our supervised machine learning engine. While the algorithms are proprietary, our process is guided by human feedback, so humans remain in charge of decision-making on key reconciliation decisions (through simple yes/no questions), which trains the machine learning model. With regards to transparency, the underlying algorithms are not visible, but all inputs from employees on yes/no questions are available, and associated confidence levels from the algorithms are fully transparent. Tamr algorithm requires some lead time for the initial training to reach a high degree of accuracy, and this lead time often depends on the Volume, Variety and Velocity of data, which may be as little as a few days. However, historically our customers have seen a positive result after just a couple training sessions of a few hours each. Once the algorithm is trained initially, humans can then decide what level of confidence is required from the machine to perform certain actions. Humans then remain in charge of reviewing scenarios where the machine is not at the desired level of confidence, thus further fine-tuning the algorithm. Remaining data outliers can be further accommodated with rules.

Have you found that having a Data Governance framework, including ratified standards is integral to mastering your data? How does data ownership play into mastering your data, in terms of defining the rules?

We have observed that there are essentially two camps that the customers we work with fall into. There are camps that try to deal with all of this data management in silos within every department and then there's a more centralized version of overall data governance. That centralized data governance team usually puts everything together in one place and then feeds this information to everyone else downstream. What we have noticed is that a lot of enterprises who have tried to develop the centralized approach did not have a lot of usability of the system because users would identify problems with it or data inconsistencies or quality problems. They would steer away from it and adoption ended up being low, so they went back to the silos. So, while it's important to have a centralized data governance model, it's also important to make sure that everything there is more accurate before being released to downstream users.

Does a machine learning solution work independent of ETL tools?

Tamr can be complementary to existing ETL tools. If you already have ETL tools in your pipeline, Tamr may be added to the pipeline to help master or categorize your data. Tamr can also replace the need for an ETL tool depending on the use case.

Getting organizational buy-in for new projects is tough, especially during times of economic turmoil. What are the top three points you'd make to justify investing in a data? project involving a modern solution like Tamr?

We covered some of the examples during the webinar. As you can see, there's definitely more than three points that I could be making about it. Given that we've seen how long it has taken everyone to move from being a fully on-premise operation to virtual operations, this entire situation has been quite a turmoil. Organizations are now looking how to enable their teams to work independently and when it comes to doing anything that is data-driven it starts with easy access to accurate and up-to-date data.

How are we enabling our team to be fully independent? Another thing we might be looking at here when justifying investing in a solution like this is now more than ever might be important to start cutting costs and actually becoming data-driven or enhancing a data-driven approach on an enterprise-wide level. So,

using something like Tamr would allow you to get access to the insights you might've been looking for in order to make decisions about spend or decisions on your customer base, anything of that nature.

The last point that you'd want to make here is really trying to look for the root of the problem that you're trying to solve. So, if you're having an analytics issue because of bad data, let's move up the chain to figure out where that bad data is coming from and clean it from the source. Let's go all the way upstream to where the data is coming in and find a tool that can really help you clean that up because then every effort downstream of that will become better. Everyone knows that it's garbage in, garbage out (GIGO). Let's fix that problem from the root and get clean data to the teams downstream and have this operational workflow become better and better.

I think bringing up these points would really allow you to move forward, especially in a time like this because a lot of the tools that we might be looking at are nice to have but it is essential for sure to have clean, up-to-date, and accurate data. So, with that, we'd really love to hear more about your experiences in this space, how your teams have gone fully remote, any stories you have about anything we discussed in this call. We'd love to hear more. If you have any initiatives you're trying to brainstorm, Aidar is an awesome solution to help guide you through that.

Thanks for joining us, we hope everyone is safe out there, and we look forward to hearing from you soon!