WEBINAR Q&A:

You point out that "the bigger the data, the more accurate the results." Is ML a viable approach for a specialized lender with tens of thousands of customers, not millions or more?

Yes, it is. Machine learning will also work on small datasets. We have worked on smaller datasets in the past where there has been an important business case driving innovation. Machine learning does really shine with high data volume and variety, as Dan mentioned, whereas with rules you often see a deterioration of data quality as data volume increases, machine learning can even improve as data volumes get higher. That being said, machine learning normally out-performs rules at small scale too.

What is the difference between Master Data (MDM) and Mastery Data?

I think what Dan calls "Data Mastery" is really just doing MDM well. But on a related note, at Tamr, we distinguish between organizations looking for a full-blown MDM system vs. organizations looking for what we call agile mastering - agile mastering for us means using machine learning to boost the quality
of entity resolution in a way that is complementary to the incumbent MDM system. Tamr can provide both MDM and agile mastering.

**Have you been able to use machine learning to derive data lineage and metadata as well?**
As Dan mentioned, data lineage tends to be pretty straightforward for us, we know which source system records come from and we keep track of that as we cluster records together and merge them into golden records. However, at Tamr we do use machine learning quick often for inferring metadata. One example of that is our classification platform which allows you to classify records into a taxonomy. Another example is the way we use ML for ETL: we infer which columns in source datasets should be mapped onto a target schema.

**MDM ensures non duplication or many replications of entities, and data elements?**
Yes, that is the goal, to ensure that all the data that an organization has about a particular entity, whether it be a customer, product, security etc. is linked and that there is a single trusted view of that entity that can be shared throughout the organization.

**Dan, how did you sell the idea of moving to ML vs rules to the stakeholders?**
Ultimately it all comes down to the business value: is this initiative going to improve the customer experience or help acquire new customers? Will it lower operational expenses or reduce risk? On the technical proof side, I often find that once people see a few examples of the improvements in data quality that Tamr can produce, they're convinced.

**To what extent was this initiative you led at your previous employer considered part of a 'regulatory remediation project'? ... As opposed to a 'new business initiative'? ... Is this a case of leveraging funding assigned for Reg/Compliance purposes to provide business value, beyond 'check box' Compliance?**
It wasn't just a regulatory initiative - but that was certainly one output from it, and certainly one of the elements of the business case leveraged in order to get funding. In large banks, regulatory initiatives often get significantly more capital and priority - so using that as a lightning rod helped get resources faster.
Are there any key requirements to enable this approach? E.g., Data Catalog/Metadata? Any items where if you do not have, you could not achieve this approach?

I'd say storage and data movement tools are always essential if you're doing MDM/data unification. Having a data catalog or dedicated governance tool may be less necessary - depending on what you're trying to achieve. For example, Tamr has some data governance tools - the tools that you need for governing mastering - so you may be able to get away with governance 'light' in some scenarios.

This blog post from Tamr's CEO is a great discussion of that question: [https://www.tamr.com/blog/the-key-components-of-a-dataops-ecosystem/](https://www.tamr.com/blog/the-key-components-of-a-dataops-ecosystem/)

What strategy did you use to get all this mastering done in 6 months? What would you do different if you did this again?

The strategy to get the mastering done in such a short timeframe was multi-faceted. Most of the work was done by the ML engine - but there was a fair bit of work needed to sort out strange data. Making sure the right subject-matter experts were available for questions was key.

Does the Data Governance process differ anyway in ML based approach for mastery or golden source? Does results will have the same rigor/quality compared to proven rule-based approach?

I don't think ML should have a significant impact on how you do governance. However, using software that's designed to be highly interoperable with other systems should help you to achieve automated implementation of data governance policies.

The results for ML should be higher quality and have a higher level of rigor. Why? The improvement in data quality really comes from being data driven - learning from the statistics of the data. The rigor comes from having a testing process that's also based on the statistics of the data - our testing process looks at the statistics of the data to ensure we've tested all the important scenarios.

Would love AI to help me out. Are there still humans in the loop? If so, what are the roles they need to do?

There can be humans in the loop - it depends on the use case. There are three areas where humans can be in the loop. First is during development when training a model, humans provide answers to questions to train the model, in our case: is this a match or not? You can avoid having humans in the
loop at this stage by using a pre-trained model, a model that's been trained on similar data for the same use case. The second part is in model testing or validation, here it's important that humans review a small sample of the data to verify that the results are good. A good product will guide you on what to review so that you make sure that review process is unbiased. The third part is reviewing edge cases that crop up from time to time. What I mean by an edge case is an example where the data looks quite different from anything the model has seen before. This is again optional, and the decision is really do you want to automate the handling of edge cases or would you prefer that someone looks at edge cases and decides what to do. In summary, the amount of human involvement is quite flexible. We encounter some use cases where there is no human involvement and some where people are very involved with interacting with the data.

What would be your advice to bridge the gap between the business and the analytics departments in order to capture commercial opportunities?
Communication is the most important way to bridge the gap. The best way to make sure you're tackling a high priority business problem is to go and talk to your business partners. I think a good question for analytics departments to ask is: what's the number one question you wish you had a data-driven answer for?

What about understanding the data and its content before implementing the ML models?
Exploratory data analysis is an important part of the process both before and while implementing ML models. A good way to do this is to ensure you're using tools that allow you to really easily visualize the data i.e., tools with an intuitive UI that allows you to make quick decisions about the data. Being able to iterate fast is also important, we often discover when we start to train a model that actually it would be helpful to transform the data in a certain way before it runs through the model, and because our software is designed to very quickly move between transforming data and training a machine learning model, we can do this.

What is the latency of cloud data synchronization or is it done only with time stamping of data?
Apologies, I'd need a little more information to know which latencies you're referring to. Feel free to email me at alex.batchelor@tamr.com to follow-up.

How do the training sets get modified with expanding data and datasets?
Good question, it's important to add some more feedback for testing when you add a major new data source to make sure your model works well for that new source. After testing, if you see that the results for the new data source are not as good as your existing data sources, then you go back and add additional training feedback for that new data source.

**For clustering it usually works with all or most data.**
There are always exceptions that don't follow the rule. Our goal is to build a model that works for the vast majority of the data. Those special cases that it doesn't work for would either be reviewed by a person or, depending on the use case, it might be ok for occasional mistakes to be made. The nice thing about ML, over rules, is that the confidence score is actually based on the uncertainty of the model which depends on the statistics of the data. So, targeting low confidence data for review is a good way to make sure you capture those rare examples that don't follow the rule.

**You want humas to help explainability?**
What I meant is that a good index of whether your system is explainable is literally "can a person explain it?" i.e., software should provide tools that allow someone to understand what's going on and therefore explain that to another person.

**Can you clarify "Missing data is corrected in most instances" and how this might work? Is there a human in the loop or assumption the ML approach will be accurate enough? (referring to slide #14)**
By matching data across many sources, we're able to fill in values that are missing in some sources but not in others. Matching to third party reference datasets can also assist in filling in missing values.

**How important are data models to understanding the data you are combining and running ML over?**
Most machine learning use cases will benefit from having an agile approach to data modelling, you bring in all the data that you think is relevant to the machine learning model, the machine learning model then learns from the data about which attributes in the data model are important for making a decision. Once you review the results of the model you may realize that bringing in an additional piece of data could enhance your results, in which case you bring in that new data source and rerun your model. The
important piece here is you don't have to specify which attributes to use, the machine learning model learns which attributes are reliable for a making a decision.

What sort of logical modelling tools or modelling steps have proven helpful, prior to the mastering & curation step?
We don't tend to spend extensive time on data modelling prior to mastering. Our approach is designed to take data from any source data model and output whatever target data model is needed. Those data models are normally defined by the needs of data providers and consumers respectively.

To what extent do you suggest a PDE/CDE inventory, to dovetail with the logical modelling/curation process?
It is important to make sure that your source data contains the CDE’s that your model will need for matching - a useful rule of thumb here is: if people were reviewing this data and deciding if these records are a match or not, would they look at that attribute? However, I recommend taking an agile approach, bring in the data you think you need, master it and get to valuable results quickly, i.e., don't let creating a data inventory hold up getting valuable results. If you realize there's an important attribute you need later on, you can also go back and update the source data, with software that takes and agile, incremental approach, this should be easy to do.

Finally, how to guard against what may be manual curation across clusters, relative to steady ML (pair feedback) over time?
If you're finding you have to do a lot of manual review of clusters, that's probably an opportunity to do further training of the model so that you can be confident in the automated results. In Tamr there's an option to learn from any cluster curation that data stewards do so that if they see the same situation again and again, that gets automatically translated into pairs that are used to train the model. The other scenario where you might have to do a lot of review is where there are ambiguous results due to missing data, that's a great opportunity to enrich your data to try and fill in some of the missing data prior to mastering.

I'm a student currently studying for my undergrad in Statistics and Data Science and have worked as an analyst within the financial industry. In wanting to be a part of this type of solution,
what specific algorithms or concepts do you suggest studying, and do you have any recommendations for advocating for these solutions in entry/junior level positions?

If you're a student trying to get your first job in data science, and you've spent some time studying data science maybe taken a few classes then I think it's quite likely that you already know the key algorithms that you'll need for interviews and most jobs. Where classes tend to prepare you less well is in the auxiliary tools that you'd use as a data scientist: e.g., SQL, simple command line tools. If you want to really impress people having some experience of big data technologies like Spark or learning about the software provided by one of the major cloud providers (AWS, Azure, or GCP) is a good start. Another key part of interviewing is demonstrating your knowledge and enthusiasm for a company.

Are you seeing any use of this approach outside of Financial Services?

Yes lots! In addition to our financial services customers, we have customers in life sciences and healthcare, manufacturing, energy, public sector, and high tech.

Where do you see the bleeding edge in data governance? Is it AI/Machine learning in the cloud and/or changes in analytics? Should individual data professionals focus on learning all aspects of this or place our focus other principals first?

I think it's likely that that automation will use AI/ML and analytics in the future. As for which to focus on or whether to focus, I think a mix of generalists and specialists are important in any organization. Do you want to be the go-to person for a particular technology? Or do you want to be someone who can give a big picture view of the overall ecosystem and advise on many different projects?

In this blog post: https://www.tamr.com/blog/the-key-components-of-a-dataops-ecosystem/

Tamr's CEO places his bet on the future of data governance on automation.

What are the key takeaways from this webinar?

Cloud and ML are going to be a big part of data management in financial services going forward. Keep sharing your vision for what good data could look like in your organization.