WEBINAR Q&A:

Where do we start and what is the business case?

The journey might have different twists and turns. Begin the journey, don't be afraid. The first thing is don't be afraid. That is sort of where the puck is moving. Let's begin the journey, your data with cloud infrastructure to really help you get better insights. And also remember the journey is towards real-time insights. Don't lose sight of that. Be aware of where things are headed and the tools available for you.

Don't be afraid to get your hands dirty and start trying. There's no wrong way of starting. It's a journey. Don't think it's going to happen overnight, we will make mistakes, but we are here to share all the lessons learned.
How do you see regulated industries being able to overcome natural resistance in sensitive data issues to moving from predictable and reproducible rules-based quality to having to trust a probabilistic model?

Anthony Deighton: I think the challenge is one that’s broader than just regulated industries. When you move from a predictable rules-based approach to a probabilistic approach, that’s sort of a challenge for everyone. The first answer to your question is simply the quality of the results. What we find is that in these rules-based systems, they simply don’t scale. When you have 10, 20, 30, 40 sources and it’s running on the cloud, it falls apart. It becomes sort of possible versus impossible. But I think the root of the question is that in a rules-based system, I can sort of backtrack and say, well, how did I get to this result?

Is that true in a probabilistic system? It is. Why did the machine make the choice that it made? The real sort of intervention point is having the human validate the results. Commonly, the machine processes enormous amounts of data, but then a human checks and says, “Yep, I agree or Nope, disagree.”

And the key there is that becomes feedback into the model. So not only do you get a reproducible answer to be able to say, yes, I validate this result, but you actually use that as training into the model itself. We see this in quite a few of Tamr’s customers, SocGen as an example, are in the banking industry. SocGen, which is using Tamr in conjunction with Google, are in a regulated industry. They’re looking to avoid regular regulatory fines, et cetera. Risk is a big driver for banks to get a handle on their data. They absolutely have to have this. It’s not that the probabilistic approach doesn’t allow it. They come at it from a different perspective.

Evren Eryurek: I come from the healthcare industry, and I remember the first day mobile phones became the thing. Well, these healthcare providers are wanting to use these tools. We got to make them work on the environment and everybody was up in arms. Oh no, the regulators won’t let us use it. And I remember having this conversation with my team. I said: “did we even talk to the regulators that were actually objecting to this?”. We didn’t. Guess what? They understand the technology is moving and no, they were not. We just needed to do the validation. That’s the key. V and V, verification and validation.

The other aspect was we were putting more and more algorithms into the decision-making part of healthcare. Now, how do you make a decision? How do we know that it’s not a black box up approach? Again, you go with the data that you’re generating. You show that it’s repeatable, you show that it is validated by humans. You show that it is a common outcome, and in the end, there’s still the expert looking at it and making the final call. We were able to overcome many of the mythical resistance from regulators by working with them. And that’s exactly what we’re doing these days as well.
When you mention machine learning for data quality, are you talking about building the solution in-house or leveraging a commercial solution?

**Anthony Deighton:** Tamr is a commercial product that is built from the ground up with a machine learning based approach specifically to solve this problem. I could imagine building from first principles, but I suspect it would be quite difficult. I know it’d be quite difficult because it’s difficult to build a software company that’s doing the same thing. Again, I’m biased. So, I think you should look to Tamr as a solution to that challenge.

**Evren Eryurek:** I’m biased too but having been on the other side of not having these tools and having to deal with rule-based systems and so forth, if you can look on our platform, we make it relatively easy and free. Come and try us, Tamr is a very nicely integrated, managed service. There will be free opportunity for you to really learn and improve what you can do and what you can achieve in a very quick manner.

**What are the key takeaways from this webinar?**

**Anthony Deighton:** Number one is to clean your house before you move into it. The move to the cloud is a fantastic opportunity to think about what you want to take with you and what you want to leave behind. The second is aligning that with your business strategy. Having a clear picture of why and what data aligns to the key business questions that you have. It’s very likely that the customer is at the center of that business strategy.

99 times out of 100, starting with customer data as a kind of key driver often connects your business strategy to what data really matters. Starting anything today that isn’t built natively on the cloud and takes advantage of the managed services existing on the cloud is a fool’s errand. You want to start on the cloud. The final takeaway is to start with a machine learning first approach. Trying to do a rules-based data quality MDM initiative is simply not going to work. Building from a machine learning based approach is the right approach.