USE CASE 1: DATA QUALITY

◆A transition in data quality process is noticeable from static rule-based approach to a dynamic, self-adapting, learning-based ML approach in various domains. ML has the potential to assess the quality of data assets, predict missing values, and provide cleansing recommendations, thereby reducing the complexity and efforts spent by data quality experts and scientists. Some businesses already use ML models to identify and eliminate fake customer records to target their marketing at genuine customers through reliable data. Also, ML partially substitutes the role of data stewards by flagging the data points based on probabilistic ratings as per learning from training set of past data steward decisions and categorizing duplicate, vacant, incorrect or suspicious entries. This reduces manual effort and governance activities. It is always beneficial for the organization to have a good hold of what data is being procured and consumed, and the business purpose it would solve. ML provides assistance in deriving a data quality index score to assess data sets' quality and reliability in real time based on deviation from predicted parameter values. It also has a marked ability to predict trends and identify outliers if trained properly and can make suggestions or take actions on the go. ML algorithms can learn from human decision labels in the training datasets and replicate the scenarios in real-time. However, ML algorithms are also prone to biases that may reflect in these data sets and are learnt through fresh data sets. These biases could lead to erosion of data quality. External validity testing and audits on a regular basis will help in avoiding such situations.

◆Many firms have faced a time lag in making data-driven decisions – by the time the data is located, tidied, sorted and applied, it is virtually out of date and no longer relevant. Firms can run into significant issues – both regulatory and business-related – if their data quality is not up to scratch. Indeed, in a pre-conference survey of delegates heading to the 2017 North American Financial Information Summit, just over half (51%) cited data quality as their biggest immediate hurdle. The countdown to the extended Markets in Financial Instruments Directive II (MiFID II) compliance deadline on 3 January 2018 had many firms particularly focused on data reliability and integrity.

◆To illustrate this, imagine a large bank that regularly deals with NatWest (National Westminster Bank). Across different business units, databases and spreadsheets, there can be many variations on the same client name – perhaps simply appearing as County NatWest, Nat West or National Westminster" and so on. Reconciling all of these entries would take significant manual work. But a computer program can theoretically scan and process data from across the bank and deliver all of the matches in a matter of hours. "Suddenly the bank can see instantly, at a corporate level, its entire exposure to NatWest," explains Rawlings. "This enables faster, better decision-making," he added. This process, or name-identity recognition, is just one of the areas where machine learning is capable of making a radical difference. And the process improves over time. In the NatWest example, the original scan may flag say 10 percent or 15 percent false positive matches on its first attempt. Through continuous feedback, it is then capable of learning from the false positives and applying the adjusted rules to the next set of data. This constant evolution is what makes machine-learning technology so effective at scrubbing and verifying data at speeds previously thought impossible.

◆Data governance is all about getting better data while sharing the wealth with everyone in the company. Essentially, governance is about focusing on the availability, usability, security and integration of the data. Newer businesses, especially, can implement data governance so the whole industry can use the same data. Governance allows the company to reduce cost and improve security while also giving better compliance overall. The data has more quality and insight to implement, so projects have a much lower risk of failure.

◆**AMEX DataQC**: The process of building a successful machine learning (ML) application hinges on the ability of a data scientist to develop a detailed understanding of the data and its attributes, such as variable type, range, outliers, and relationship with the dependent variable, thereby ensuring the data quality. **The successful application of this model depends on the ability to obtain data in the same format and structure over and over again without significant changes in the statistical distribution of any of the attributes**. **Changes in the data require frequent model refreshes or can result in subpar model performances.** For applications that rely on artificial intelligence (AI) to make decisions, these requirements can make it a significant challenge for a single or a group of human data scientists to ensure repeatable, high levels of data quality. **A system that is capable of detecting a wide range of data issues while being fully automated is key to ensuring accurate and reliable models**. Archana Anandakrishnan offers an overview of DataQC Studio, American Express's automated system built to identify data issues and data anomalies and create an exhaustive snapshot of the data. The tool has been built with Python and Spark to be able to scale to large datasets and is completely built with open source tools such the MLlib random forest classifier and the t-SNE implementation in scikit-learn. By combining a variety of methods, DataQC Studio learns a confident quality score for any dataset that can be used to assess the integrity of a dataset before using in a model. The tool solves a fundamental problem of data quality management and its potential far exceeds any manual data quality management process. Archana demonstrates the power of the ML-powered DataQC pipeline built with open source software by showcasing its extensive use at American Express. Since ML models power many critical decisions at American Express, the accuracy of these models are important for managing risk and delivering a superior customer experience. The methods described here are modular and adaptable to any domain where accurate decisions from ML models are critical.

Archana Anandakrishnan, American Express: Archana Anandakrishnan is a senior data scientist in the Decision Science Organization at American Express, where she works on developing data products that accelerate the modeling lifecycle and adoption of new methods at American Express. She is currently a lead developer and contributor to DataQC Studio.

Previously, she was a postdoc researcher in particle physics at Cornell University. She is passionate about mentoring and is currently a workplace mentor with Big Brothers Big Sisters, NYC. Archana holds a PhD in physics from the Ohio State University.

How Can Machine Learning Support our Data Management and Help us Improve our Data Quality?

In order to assess the role and potential, the Competence Center Corporate Data Quality (CC CDQ) collected and analyzed ML use cases from academic research, software vendors, and data management experts (Fadler & Legner, 2018). With our study, we aim to identify typical application scenarios that can help data managers find potential areas of application for ML in data management. By now, we developed a taxonomy for classification of uses cases and derived 11 typical application scenarios for machine learning in data management from 44 collected use cases. Our study reveals that ML can be applied in all phases of the data life-cycle to achieve the following:

- To create and enrich data assets in an efficient, user-friendly way
- To maintain high-quality data by supporting proactive and reactive data maintenance as well as for data unification
- To manage the data life-cycle, especially when it comes to sensitive data and retiring data
- To increase the use of data by improving data discovery by users, specifically by data scientists

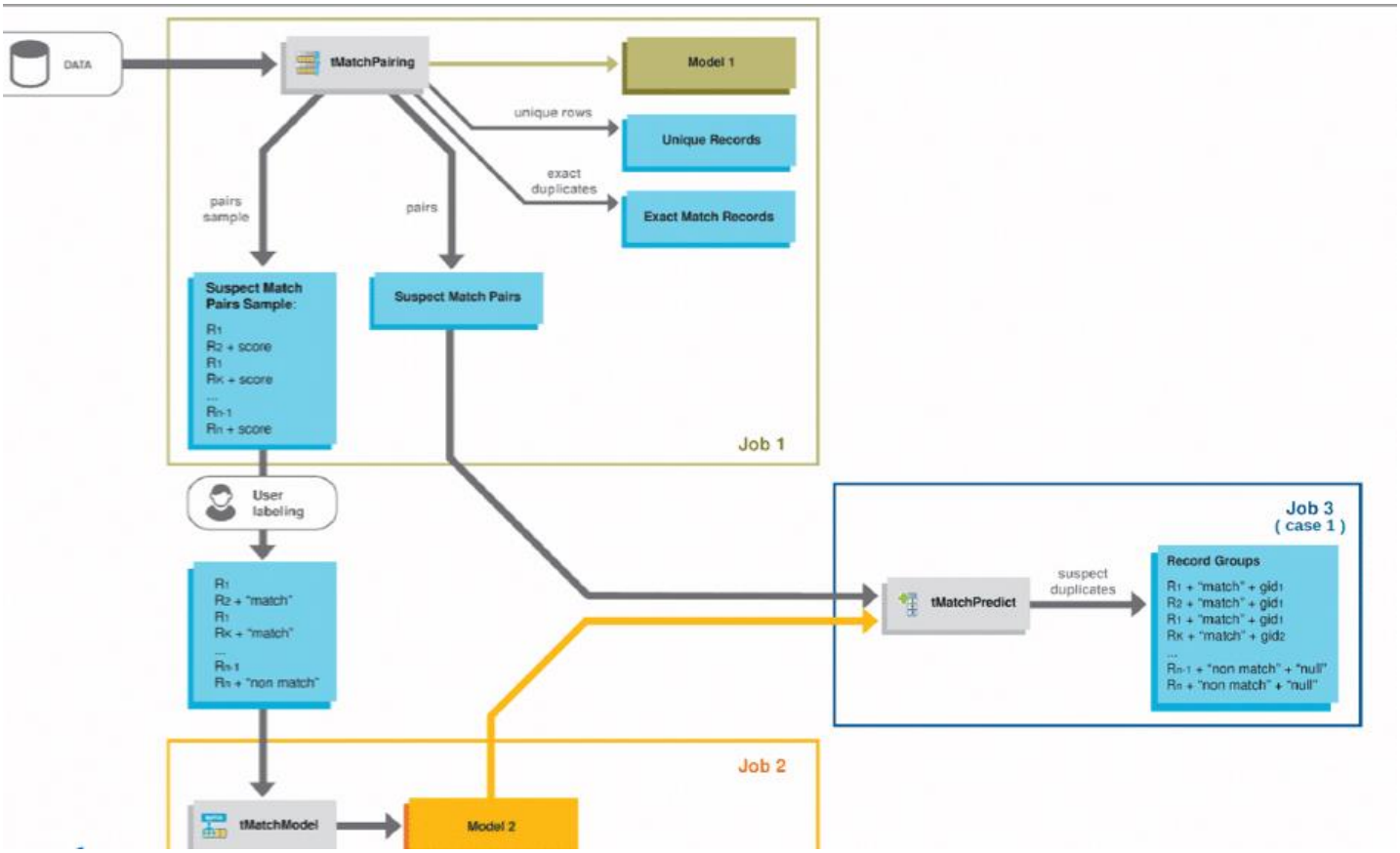
The following overview provides more details about the data life-cycle phases and ML application scenarios.

Data Management Activity	Acquire and Create Data	Unify and Maintain Data	Protect and Retire Data	Discover and Use Data
Role(s) Involved	Data Collector	Data Manager, Data Integrator	Data Protection Officer, Data Manager	Data Scientist
Problem Areas	Typos; wrong/invalid data entries; blank fields; manual effort	Data integration across multiple systems (leading to inconsistencies); correction of data errors; definition of business rules	Lack of transparency where personally identifiable information (PII) is stored; compliance with data protection regulations	Finding and cleaning relevant data; identification of data relationships
Learning	Data entry patterns; data incidents; data extraction patterns; data creation patterns	Data repairing patterns; association rules; outliers and anomalies; similarities	PII identifiers; fraudulent data access behavior	Data recommendations; linking of datasets
ML Application Scenarios	**ML assisted data creation** (e.g. auto-filling values in forms, automatic extraction of data) and **data enrichment**	**ML assisted data maintenance** (reactive: data correction; proactive: business rules) and **data unification** (matching and deduplication)	**AI-/ML assisted data protection** (e.g. identification of sensitive data, detection of fraudulent behavior) and **data retirement** (end of life)	**AI-/ML assisted data discovery** (e.g. recommendations, linking of datasets)

Although ML's use in data management is only in an early stage, the first implementations are very promising! For instance, Bosch has been able to almost completely automate the manual process of commodity code assignment in product master data creation with the help of machine learning. With this approach, Bosch can fulfill the increasing demand for this assignment task across the enterprise with a scalable solution. Find detailed information in Bosch's winning application for the CDQ Good Practice Award 2018.

The bottom line of our analysis is that ML has the potential to significantly enhance data management practices and improve data quality. ML allows for managing data assets in an intelligent and more scalable way, but also disrupts the way data is managed.

◆**Machine Learning Going Mainstream** According to some studies,22 percent of the companies surveyed have already implemented machine learning algorithms in their data management platforms. NASA, for example, has discovered a lot of applications for machine learning in assessing the quality of scientific data such as detection of unusual data values and anomaly detection. The reason ML is becoming mainstream is because Big Data processing engines such as Spark have made it possible for developers to now use ML libraries to process their code. Each of the ML libraries currently available through Spark are also available for Talend developers. The Winter17 release of Talend Data Fabric also introduced ML components for data matching. They are **tMatchpairing**, **tMatchModel**, and **tMatchPredict**. Below is a high-level overview of the process required to use these components for predicting matching results. **Data matching with machine learning in four easy steps: Step1:** Pre-analyze the data set using the **tMatchpairing** component. This uncovers any suspicious data whose match score is between the threshold and match score. The match scores would also be the part of the data set. **Step2**: Data stewards then label the suspect match record as 'match' and 'non-match'. It is a manual process and the Talend Stewardship console can be leveraged to streamline this labelling. **Step3:** A sample of result set from Step2 is fed into the **tMatchModel** for 'learning' and the output would be a ML classification model. Model validation is automatically done here using the **tMatchPredict** component. **Step4:** The model generated in Step3 is ready to be used to predict matches for new data sources.



NASA Big Data

Machine Learning and Big Data Services

We are expanding our portfolio into new pilot services for big data, machine learning, and data analytics support. For users considering advanced analytics using the latest technologies in machine learning and deep learning to either facilitate scientific understanding or to share complex datasets, our experts can work closely with you to provide guidance on which machine learning techniques are best suited for your use cases. We will also work with you to assist in sharing your data with other NASA or external users through custom-built data portals.

Free Services

The following big data and analytics services are available at no charge to you.

- **Assist with publication and discovery**
  - Support a common access point for public data to assist you in sharing your data with external users
  - Maintain a portal to assist you in sharing your data with other NASA users

To view currently available public datasets, see NAS Data Portal.

- **Develop targeted environments for machine learning and data analytics**
  - Analyze your current project and future goals to advise on whether your problems are a good fit for machine learning
  - Team with our systems experts to get machine learning libraries added to your HECC environment
  - Work with your team and build custom machine learning models
  - Provide container-based solutions for user-specific software stacks

For more information, see HECC Machine Learning Overview.

- **Provide documentation of use cases completed by data analytics**
  - All completed machine learning projects are documented by our team; these documents can be used for reference by future projects

To request help with these free services, please contact the NAS Control Room: (800) 331-8737, (650) 604-4444, support@nas.nasa.gov.

Funded Services

Our machine learning and big data experts can also work closely with you to provide more extensive support. Such services may require external funding based on the level of effort.

For more information on these in-depth services, please contact:

Shubha Ranjan
Big Data and Analytics Group Lead
shubha.ranjan@nasa.gov
(650) 604-1918