







EDM Webinar

Leveraging Machine Learning for Data Management

A conversation with

Nicolas Vaillant
Data & Analytics Strategy Lead
Arrayo

Arrayo

EDM Council

Today's Panel

Moderator

Jim Halcomb
Head of Product Management
EDM Council

Nicolas Vaillant
Data & Analytics Strategy Lead
Arrayo

Data Management: Current State

Priority

Data Management is the top priority of **60%** of Data Leaders. (cio.com, 2023)

Strategy

Only **30%** of organizations have a comprehensive data management strategy. (Gartner, 2023)

Activity

70% of data management tasks are still performed manually. (McKinsey, 2023)

Cost

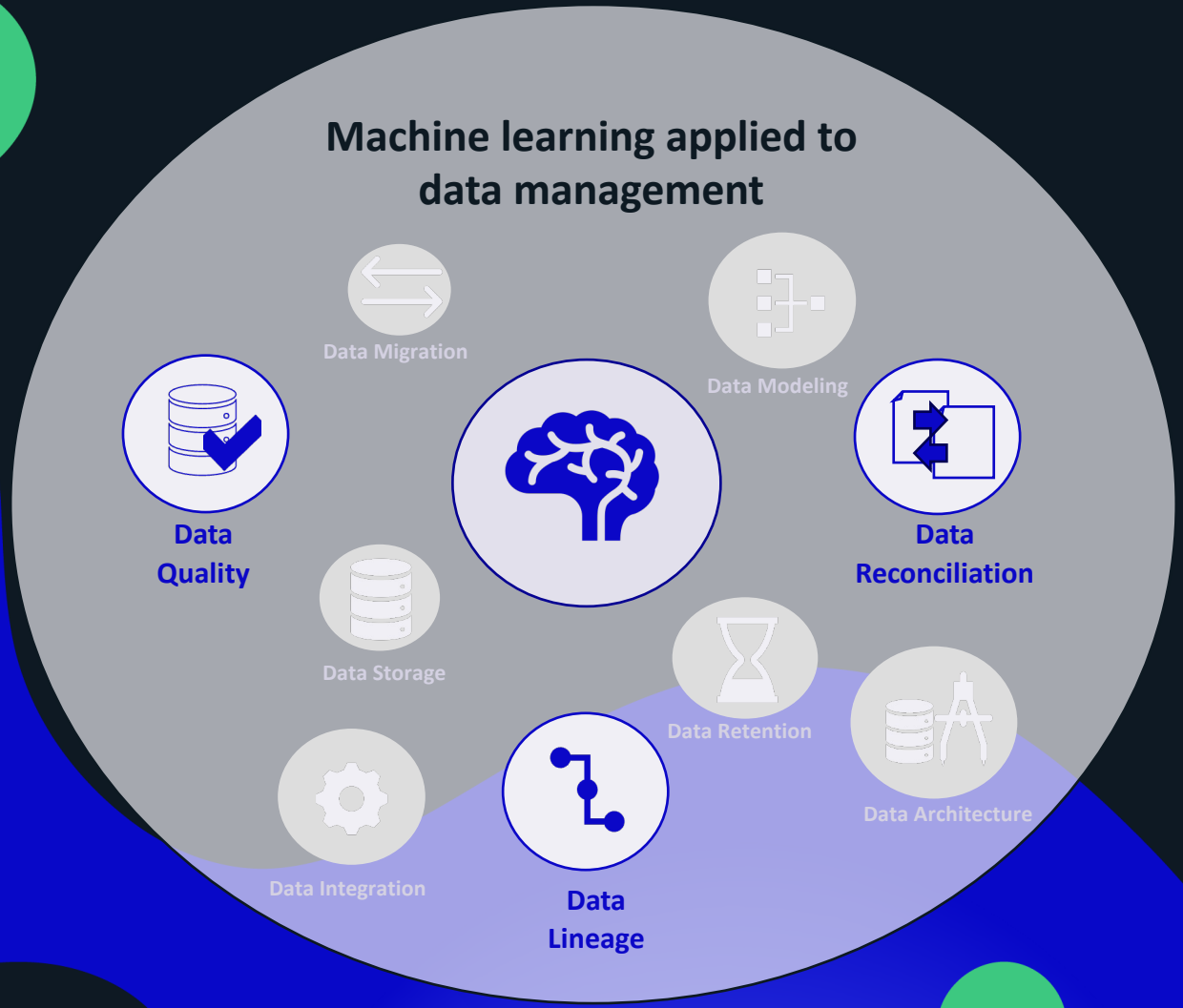
The cost of manual data management is estimated to be **$3** trillion per year. (Ponemon Institute, 2023)

Innovation

Organizations that use machine learning for data management can save up to **30%** on their data management costs. (Forrester, 2023)

Arrayo

- **Data quality assessments, data reconciliation, and data lineage** are business-critical tasks performed by many business and data management professionals.
- **Challenges with traditional methods**: most data management tasks require manual effort.
- **Machine Learning** can help automate and enhance data quality, reconciliation, and lineage tasks.



Data Quality Management – Traditional Approach

- Data is an **important asset** therefore its quality and reliability is paramount. We need accurate and reliable data for effective decision making.
- Traditional data quality management approaches **involve comprehensive, time-consuming analysis** across the entire data process.
- The main **challenges** revolve around the issues of scalability, manual error, and time consumption associated with traditional approaches.
- **Machine Learning** techniques can be utilized to automate the identification of potential data quality criteria, enabling the detection of anomalies, and improving data accuracy.

Enhancing Data Quality with Machine Learning

ML application

- No labeled data
- Uses unsupervised and supervised algorithms
- Explanation of identified patterns

Benefits

- Faster processing
- Reduced errors
- Improved scalability
- Reduced manual effort

Data quality assessment

Data collection

Gather and pre-process data. The goal is to prepare data before running algorithms.

Outliers detection

Once data is pre-processed, and given there is no labeled data, we identify pattern with unsupervised algorithms.

Outliers interpretability

Once we've identified all outlier data, it's necessary to interpret the results. We label the data and employ supervised learning algorithms for better identification and detailing of the patterns.

Data quality rules identification

Data collection

We gather and pre-process the data with the objective of preparing it for the subsequent execution of algorithms.

Association mining

Once the data is processed, we employ mining algorithms to identify associations between columns.

Criteria of quality

Ex : Completeness criteria
A is null when C = 00
B is null when E = 1

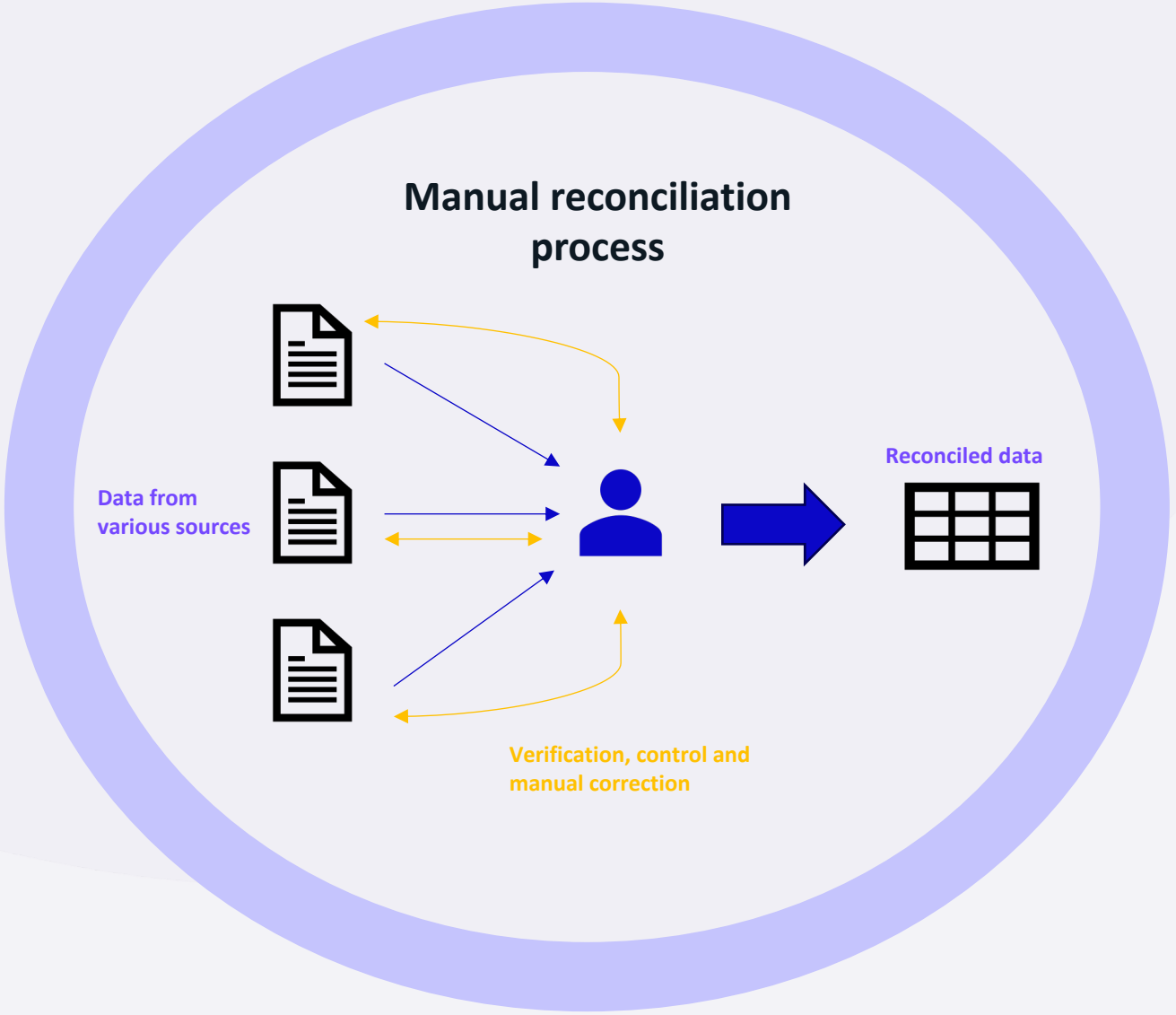
We simplify the results to formalize the findings into understandable rules.

Share with SMEs

The results are shared with the business to assist them in establishing data quality rules.

Data Reconciliation – Traditional Approach

- The crucial function of reconciliation in guaranteeing the **accuracy** of data.
- **Traditional data reconciliation:** manual cross-checking of data from various sources.
- **Challenges:** time-intensive, error-prone, and struggles with large data volumes.
- **Machine learning for data reconciliation:** automate and optimize the data reconciliation process, overcoming challenges of manual cross-checking.



Streamlining Data Reconciliation with Machine Learning

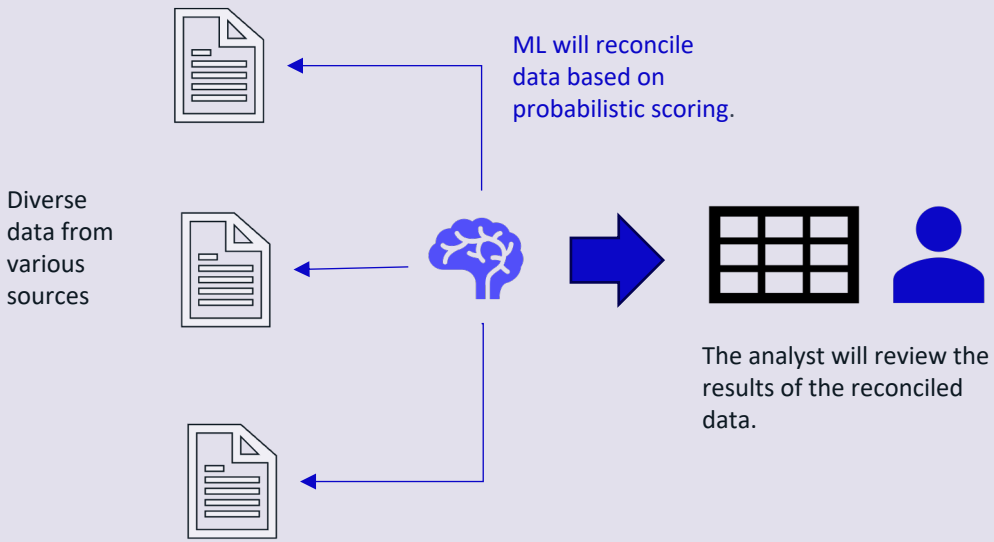
ML application

- No labeled data
- Uses Linkage – data matching algorithm
- The result will be classified by match

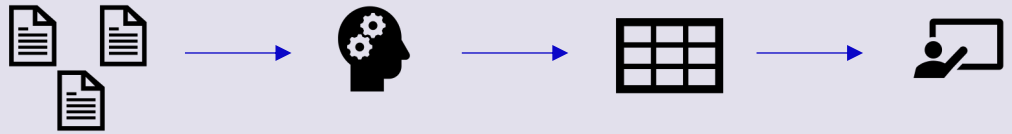
Benefits

- Faster processing
- Improved scalability
- Reduced manual effort

Machine learning solution (overview)



Machine learning framework (overview)



Gathering and pre-processing various datasets.

The model will identify matches among different data across all files.

The results show matches divided into three categories: Perfect, Close, and No Match.

Present the results to the Analyst/Subject Matter Expert for validation.

Example

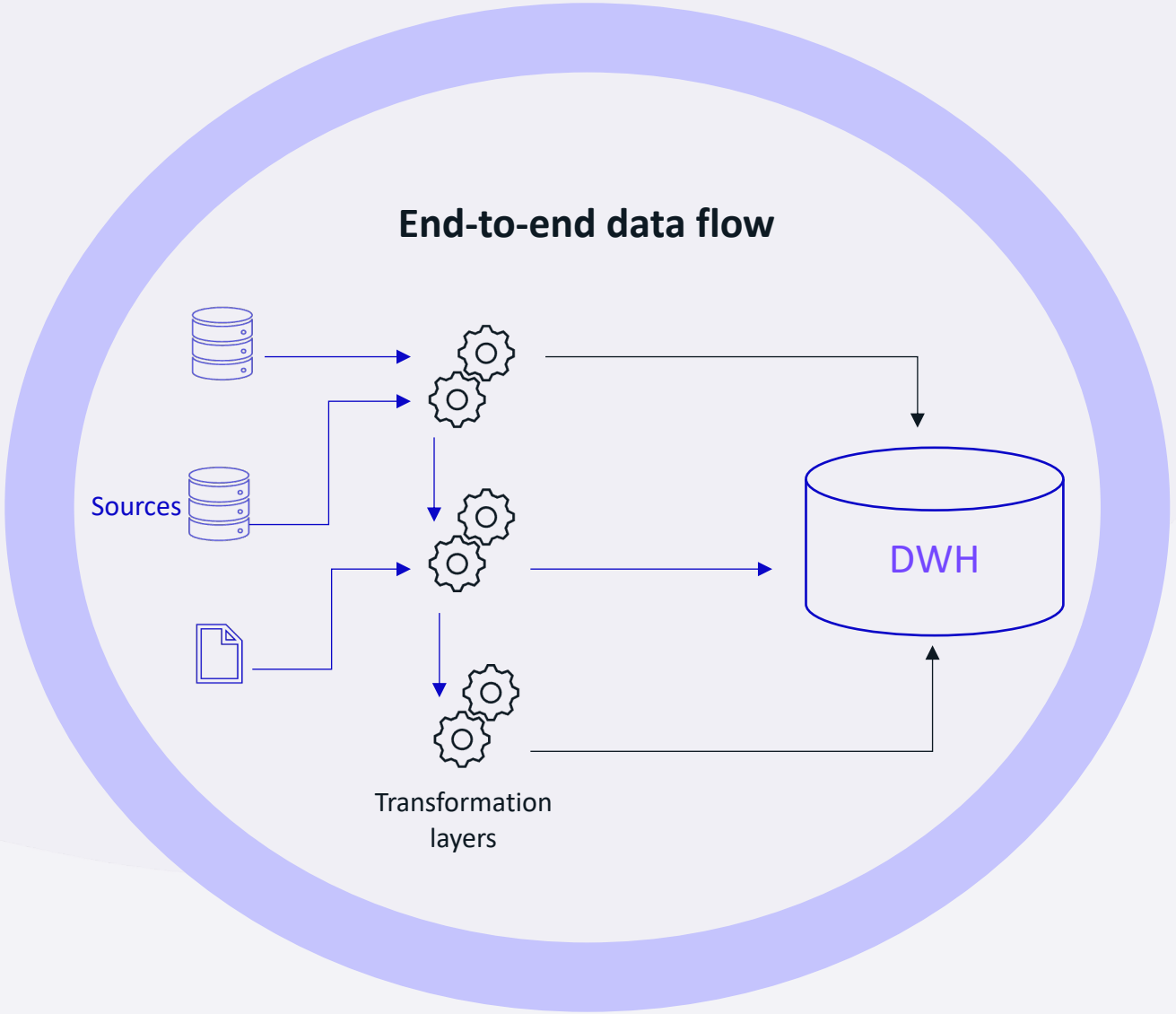
1st dataset
- customer : Nicolas Vaillant, address : 42nd st, Manhattan

2nd dataset
- customer : Nicolas Drisse V, address : NY

Client : Nicolas Vaillant, address : 42nd st, Manhattan
Client : Nicolas Drisse V, address : NY
Similarity score : 0.7
Predicted ; 1

Data Lineage – Traditional Approach

- The crucial function of **data lineage** is to ensure the transparency and trust in data.
- **Traditional data lineage:** manual tracing of the movement of data from the origin.
- **Challenges:** complexity, manual error, lack of real-time tracking.
- **Machine learning** techniques can be utilized to automate and enhance data lineage tracking, improving efficiency, and monitoring.



Enhancing Data Lineage with Machine Learning

ML application

- Requires to label data
- Needs Feature Engineering
- Supervised algorithm

Benefits

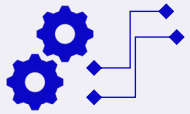
- Faster processing
- Improved transparency
- Reduced manual effort

Machine learning framework (overview)

Data collection

Extraction of required data: The Subject Matter Expert (SME) verifies the lineage of specific data points to establish labels.

Features Engineering

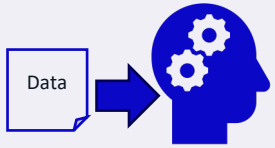


The objective is to analyze the data and identify potential metrics for the training process.

Build the ML model

After identifying the metrics, the algorithm is trained to predict the data source based on the labeled data.

Run the ML model



Once the model is trained and tested, it can be applied to other data for predictions.

DWH Columns : **X1 , X2, Y1, Y2**

Source A columns : **"A1", "A2", "A3"**

Source B columns : **"B1", "B2", "B3"**

Column Name	Starts with X?	Mean Value	Source
X1	True	52.6	0
X2	True	100.0	0
Y1	False	10.0	1
Y2	False	20.0	1

Starts with X?	Mean Value	Source (Target)
True	52.6	0
True	100.0	0
False	10.0	1
False	20.0	1

Column Name	Starts with X?	Mean Value	Predicted Source
X3	True	75.0	0
Y3	False	15.0	1
X4	True	85.0	0
Y4	False	25.0	1

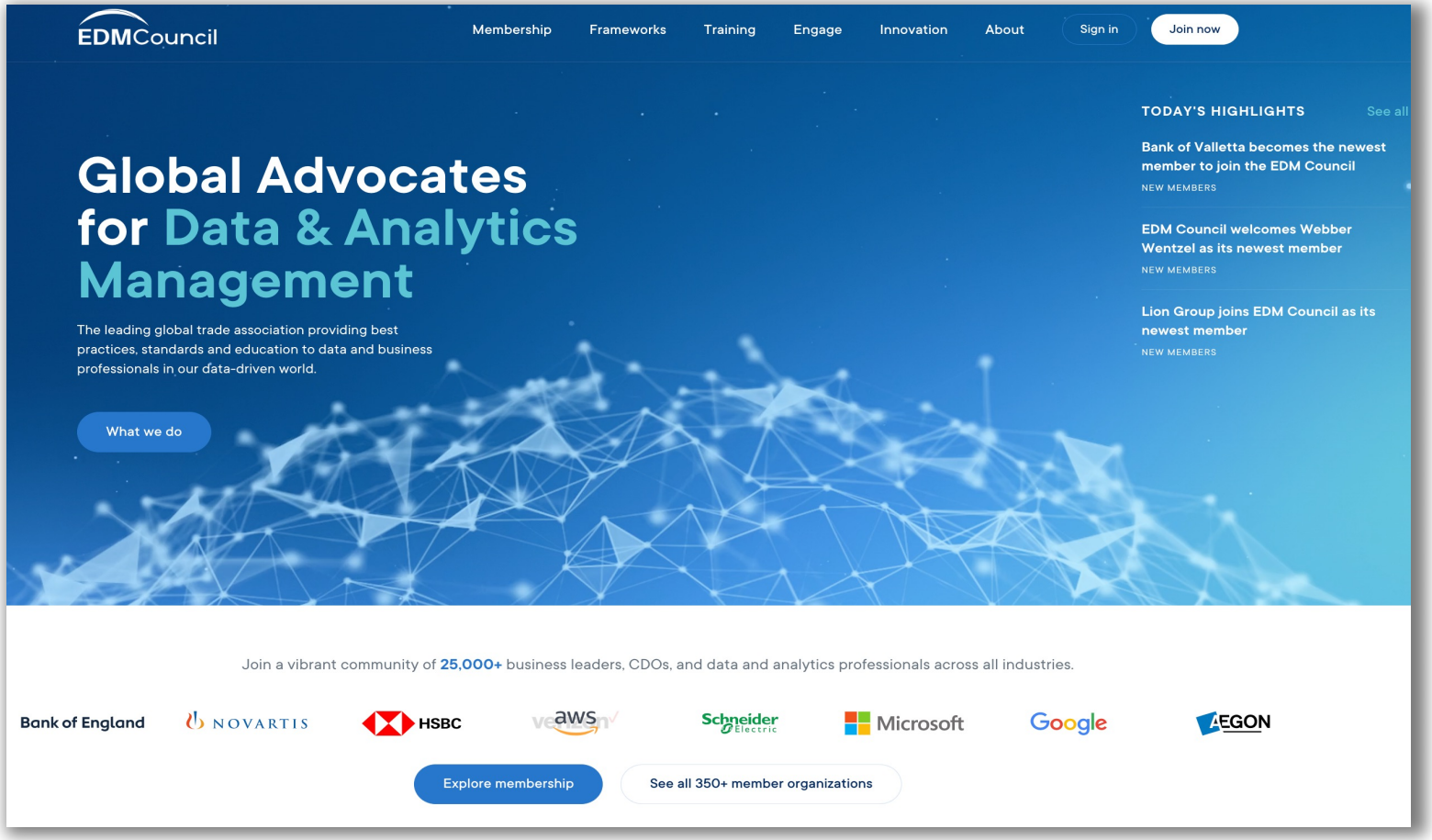
Arrayo

What’s next?

Questions?

EDM Webinar

Join EDM Council and our membership community of companies…



EDMCouncil

350+ Member Firms
Cross-industry,
including Regulators

25,000+
Professionals

Worldwide
Americas, Europe,
Africa, Asia, Australia

edmcouncil.org

DCAM

CDMC EDMCouncil

Data ROI EDMCouncil

ESG EDMCouncil

Open Knowledge Graph INNOVATION LAB

elc eLearningCurve









EDM Webinar

Thank you!

FOR MORE INFORMATION:

Jean-Philippe Michel
Arrayo - Head of Business Development
jpmichel@teamarrayo.com

Nicolas Vaillant
Arrayo – Data & Analytics Strategy Lead
nvaillant@teamarrayo.com

Arrayo

EDM Council