OPPLANE

Unraveling the Black Box:

Using Automated Data Lineage for Explainability in AI and ML

Whitepaper
2023

Online:
contact@opplane.com
www.opplane.com

Phone:
+1 833-OPPLANE

Date of Issue:
07 // 09 // 2023

Table of contents

1. Introduction

Over the last ten years, we have witnessed the widespread introduction of Artificial Intelligence (AI) and Machine Learning (ML) models in sensitive decision-making processes.

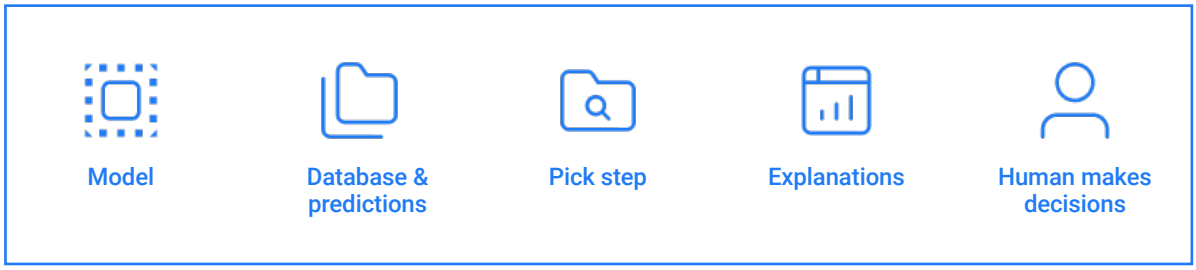
This trend results from the development of computational models capable of emulating a series of cognitive processes inherent to human beings with increasing reliability. At the same time, the vertiginous increase in computational power has led to a profusion of new algorithms that were previously impossible to implement.

As a result, AI and ML systems have become much more complex, capable of integrating a much larger number of parameters and running much faster. This trend has led to a substantial improvement in the ability of algorithms at the basis of these systems to predict future outcomes with a high degree of accuracy.

Governments and companies are incorporating AI systems into high-risk areas such as fraud detection, risk assessment, and natural disaster prediction, with the expectation that the predictions made by these systems are trustworthy. But as algorithms have become more complex, there has been a decrease in the ability to explain the decisions made by these algorithms to all stakeholders. The growth in predictive power has come at the cost of transparency and the ability to get explanations of why the models performed the way they did.

More and more often, developers and data scientists have access to the set of outputs generated in response to a given set of inputs given to a model without understanding the inner workings of the model. Unlike mathematical models that have an inherent structure, ML models can learn the mapping between inputs and outputs directly from the data. Hence, many ML models, such as deep learning models, behave as "black boxes" that do not reveal their internal mechanisms and the reasons behind their predictions (Kamath & Liu, 2021, p. 2).

The inability of AI and ML systems to be transparent and understandable by even their creators can significantly hinder the levels of trust and adoption by stakeholders. In the case of high-risk disease detection, an incorrect diagnosis can make the difference between life and death for a patient. Other activities where unintelligible predictions are equally intolerable include credit loaning and bail and parole assessments.



It was to address the need for explainability that Explainable AI (also known as XAI) emerged. It consists of a set of techniques, concepts, and processes aimed at helping stakeholders understand the behavior of ML models. The goal is that model creators can explain how a model reached a specific prediction. Knowing how a model behaves and how the training data set influences it can help developers and data scientists identify and correct the causes of flawed or inaccurate decisions. Most importantly, the provided explanations can help all stakeholders evaluate the model's decisions more critically (Munn & Pitman, 2022, p. 4 and pp.13-14).

One of the critical processes for achieving explainable AI is automated data lineage, a powerful tool for tracking and understanding any data path. You can get a visual representation of your organization's entire data environment using specialized automation tools. Solutions focused on data lineage thus allow you to determine the path your data has taken from source to destination and the transformations it has undergone. You can then obtain explanations that are easy to understand, access, and share with others in your organization.

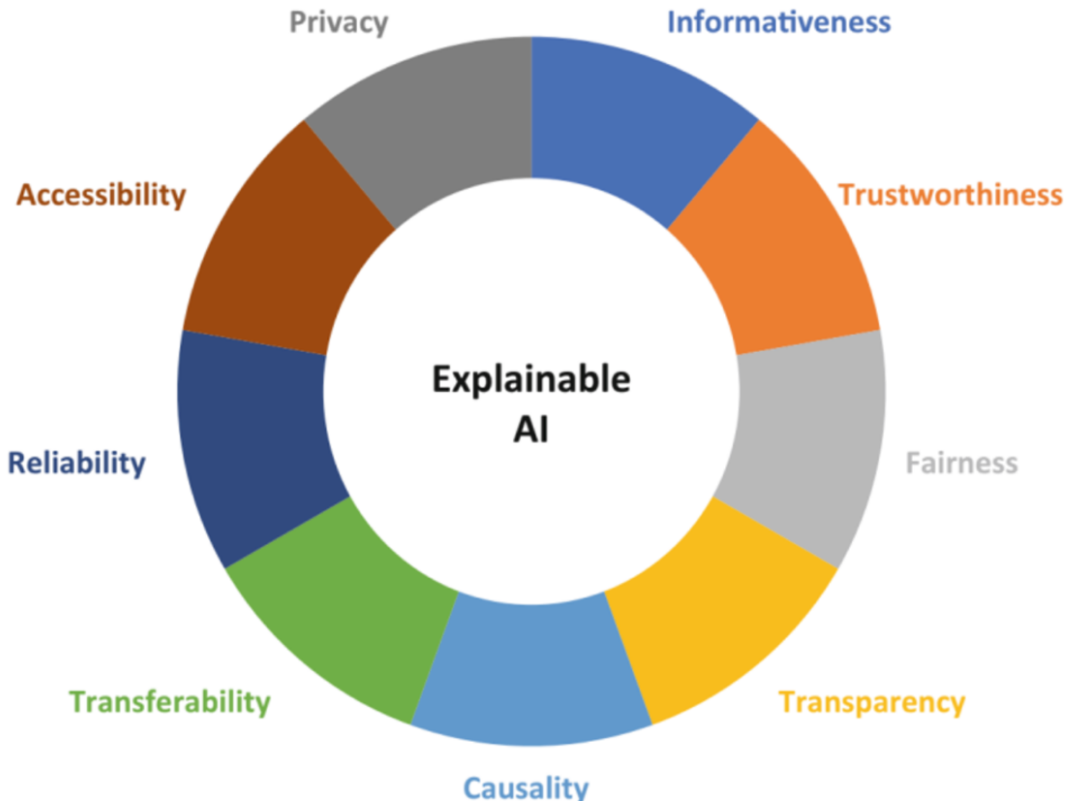
The goal of this white paper, presented by Opplane, is to provide insights to data practitioners, decision-makers, and stakeholders on leveraging the potential of automated data lineage to achieve transparent, accountable, and trustworthy AI and ML systems. We are sure you will find in this white paper valuable principles on tackling the "black box" challenge and providing explicability to your AI and ML-centric services and products.

More specifically, this white paper provides a detailed analysis of the following topics:

- *The significance of explainability in AI and ML systems.*
- *How automated data lineage functions*, its *primary advantages* and *limitations*, and *how it compares to manual techniques*.
- *The role of data lineage in achieving explainability.*
- *Future trends and implications* for achieving explainability via automated data lineage.

2. The role of explainability in AI and ML

For an AI/ML system to be deemed trustworthy, it must be auditable, governable, and explainable. Explainability is crucial in achieving this third pillar of trustworthy AI. It helps us comprehend the system's reasoning behind its predictions. The objective is to present the model's structure in an easily comprehensible way for humans, aiming to promote its adoption and provide a safer user experience.



In practice, explainable AI encompasses several types of aspects that must be considered by practitioners to ensure the explainability of solutions. Explainable AI methods aim to mitigate or even solve many of the challenges posed by these aspects in various ways. As the complexity of algorithms increases, it is increasingly certain that the success of any AI/ML system will depend on the correct implementation of these methods. Here is a brief explanation of the essential aspects of explainable AI (Kamath & Liu, idem, pp. 7-10):

1. Information

When an AI system openly shares the inner workings of its underlying model, it enables us to accurately evaluate whether and when the model's objectives contribute to improve the decision-making capabilities of the consumers of that model.

2. Trust

This principle relates to how secure we are in our interactions with AI applications and whether we feel the focus of these solutions fulfills our interests.

3. Fairness

The principle of fairness is crucial in explainable AI because it helps us detect bias in models. By fairness, we mean treating all stakeholders impartially and without discrimination or favoritism.

4. Transparency

This factor helps us identify the input variables influencing decision-making.

5. Casuality

Explanability allows us to identify the causal relationships in the data, which helps overcome a critical issue with AI and ML models - their overreliance on correlation at the expense of investigating causes. However, it's essential to remember that you often need to have significant prior knowledge and expertise in the relevant domain to prove causality.

6. Generalization

Explainable AI helps us understand a model's internal structure and learning process, which allows us to apply that model to other tasks more easily.

7. Reliability

The more often an AI system can arrive at the same decision under the same circumstances, the more that system will be considered by its consumers.

8. Accessibility

By making complex AI models easier to understand for ordinary consumers of the solutions that incorporate those models, explainable AI can help popularize the technology to more users.

9. Privacy

By focusing on model explainability, we can more quickly assess whether the privacy of algorithms or encryption schemes is at risk.

Why should you care about explainablity?

Businesses prioritizing explainability in their AI and ML systems are helping to promote wide-spread acceptance of these technologies. More specifically, the importance of explainable AI lies in several key topics, such as ensuring organizational safety, complying with regulations, being socially responsible, promoting fairness and preventing bias in decision-making:

1. Safety

When it comes to applications that could potentially cause harm to humans, like financial, healthcare, medical diagnosing, or self-driving systems, the developers of these applications must make sure that algorithms make safe decisions. Explainability is crucial in ensuring these decisions are not unintentionally or maliciously harmful. Since it's impossible to test all possible scenarios in the real world, explainable AI is necessary to identify weaknesses in the model.

2. Complying with regulations

In addition to industry-specific regulations - regulations specific to healthcare, such as the Health Insurance Portability and Accountability Act (HIPAA) in the USA and finance, for instance -, explainability is also crucial for meeting general legal requirements like the General Data Protection Regulation (GDPR) in the EU. In particular, Article 15 of the GDPR provides the right to explanation, which means that AI developers must provide explanations for decision-making algorithms with legal or significant effects in applications that process personal data.

3. Social accountability:

If your AI or ML system is customer-based and you want your customers to trust you, you need to have explainability in mind. Human beings value explanations not only in scenarios involving critical decisions but in all sorts of scenarios, such as listening to music recommendations on a streaming platform such as Spotify. Providing clear explanations thus improves the overall user experience, enhances customer satisfaction, and promotes long-term customer loyalty.

4. Promoting fairness and preventing bias:

Social biases have existed for a long time and can seep into the data sources used to train AI models. In just a few years since the introduction of AI and ML systems in high-risk decision-making processes, there have been notable instances of gender and racial bias that highlight the importance of fairness. While it's possible to prevent bias at the start of the data quality process, explainability can assist in identifying which data points may indicate bias coming from human sources. This is particularly important if your data models have social consequences and depend on human-generated data.

3. What is automated data lineage?

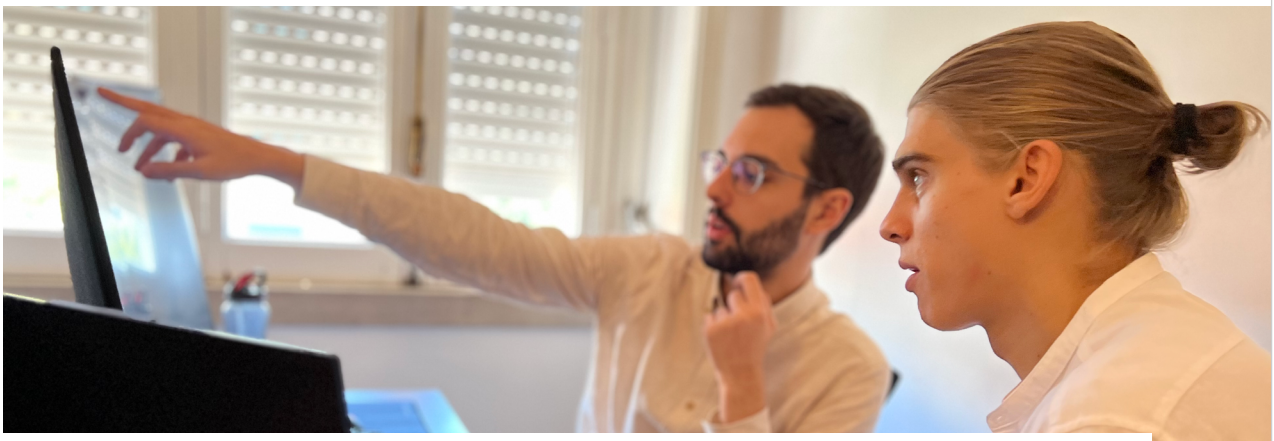
Once collected, data travels down a path along which it goes from its origin to its usage point. While on this path, data is moved back and forth and subjected to various transformations (DAMA International, 2017, p. 28). Data lineage thus describes the data flow from source to consumption within an organization. Understanding this flow provides an overall picture capable of dealing with the growing complexity of data storage systems containing increasingly diverse data sets in different formats and from various sources (Southekal, 2023, p. 88).

In this sense, data lineage helps data-driven organizations answer a series of increasingly essential questions, such as:

- Where did the data come from?
- How was the data transformed or processed?
- What calculations or algorithms were applied to the data?
- Who used the data and for what purpose?

This comprehensive view of how data is collected, manipulated, and analyzed is indispensable to ensure consistency, reliability, transparency, and trust in an organization's data sets by detecting and correcting any flaws. Data lineage thus enables data explainability in AI and ML systems by providing a clear data audit trail, as we will see more closely in a later chapter.

Due to this context of increasing data complexity, organizations tend to use cloud-based tools and services to automatically track the origins and potential changes to the data while in its flow. These options facilitate a task that would be practically unfeasible when done manually.



Applying Data Lineage

Most data lineage software solutions visually represent the path taken by an organization's data using an entity (such as a point, rectangle, or node) and its connecting lines. The entity can represent a particular point of data, a collection of data elements, or even the data source. At the same time, the lines are the flows and transformations the data undergoes as it is processed for consumption throughout the organization.

Here is a list of the most popular techniques and methods for creating a data lineage:

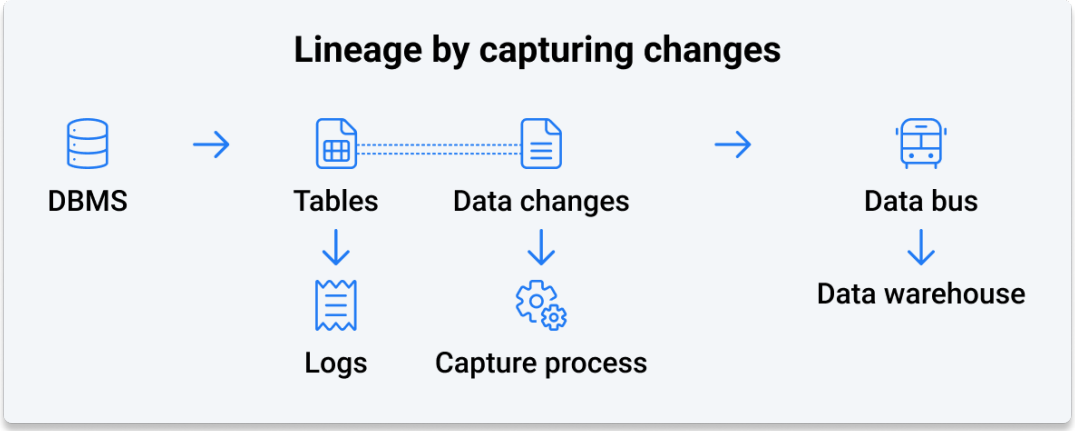
Data tagging:

Some transformation engines can tag each existing data version and create links between them. These tools can then read these tags and visually represent the data lineage. The main drawback of this technique is that it can only track a single internal environment, meaning that the tool will ignore all external data transformations.

Search	Tags
Meditation	Health
Green tea	Health
Novels	Media
Stories	Media
Screeen plays	Media
Documentaries	Media
Yoga	Health
Sunlight	Nature

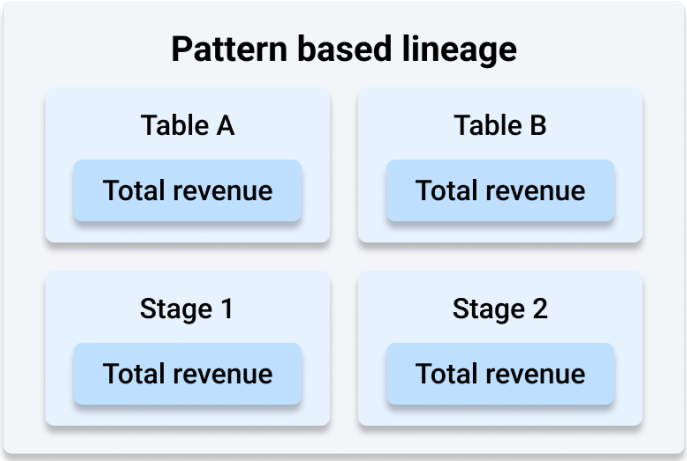
Code parsing:

This is an advanced form of data lineage tracking that can automatically read the logic behind each transformation and interpret it to track all the changes made to the data. The parsing is done by reverse engineering the programming logic behind the data transformations. The most significant disadvantage of this method is that it requires deep knowledge of the programming logic implemented throughout the data lifecycle.



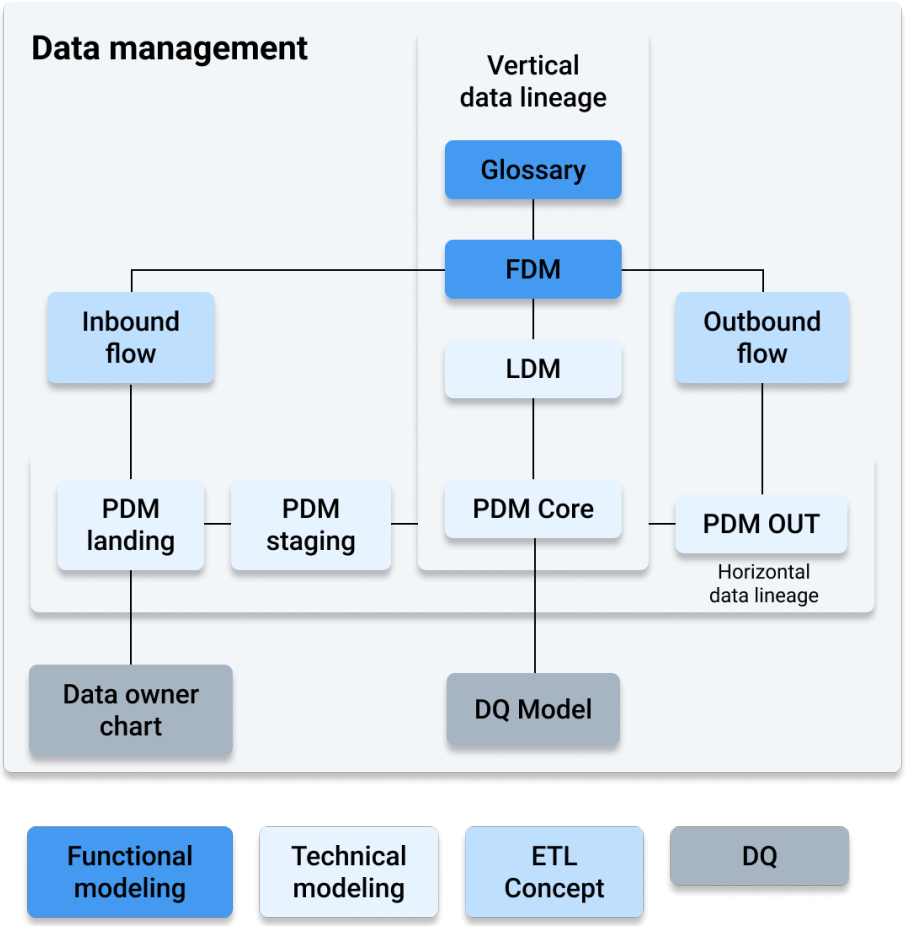
Pattern-based lineage:

This approach relies on the metadata of the tables and databases. Most automated data lineage tools can extract metadata from various sources to provide additional information about the data. Pattern-based lineage is technology-agnostic since it does not deal with the software code used for data transformations but rather with the changes made to the metadata. For instance, you can use it across several database systems (MySQL, SQL Server, or Oracle, among others). The main disadvantage of this method is that sometimes it can provide inaccurate results due to its simpleness.



Different Data Lineage Perspectives

Most data lineage software solutions visually represent the path taken by an organization's data using an entity (such as a point, rectangle, or node) and its connecting lines. The entity can represent a particular point of data, a collection of data elements, or even the data source. At the same time, the lines are the flows and transformations the data undergoes as it is processed for consumption throughout the organization.



One way to better analyze the different paths taken by data within an organization is to distinguish between two major types of data lineage. The first is horizontal data lineage, which represents the physical path data takes from its point of origin to its point of use across different systems or modules. Models focused on horizontal data lineage thus allow visualizing how data moves from one system to another - for example, from a database in production to a data warehouse.

Vertical data lineage, on the other hand, offers a more comprehensive analysis of the links between different processes and systems. Models centered on vertical data lineage thus provide a more abstract view of data flow, one centered on a business or conceptual data model that underlies the implementation of the physical data model (Freche, Heijer, and Wormuth, 2021, p. 10).

How data lineage helps solving several business problems

While many organizations consider data lineage a business burden, the strategic benefits offered to organizations that invest in detailed and transparent documentation of their data flows can be significant. Here is a list of some of those benefits:

A better understanding of data flows:

The knowledge gained from implementing data lineage techniques helps members of an organization in data discovery and understanding the potential of that data. The overview provided by data lineage encompasses both the flows but also the metadata. At the same time, data lineage also enables all members to access and understand each other's responsibilities and dependencies on IT systems, helping to better cope with technological developments and fluctuations in staff.

Better error detection and resolution:

When your data flow contains dozens of modification points, it becomes challenging to detect potential errors and determine who was responsible for the modification at the source of the error. With data lineage, it is possible to visualize each data set resulting from each modification made throughout the flow, thus making it easier to identify any incorrections.

Impact analysis:

Understanding the lineage of your data is indispensable for assessing the potential impacts of changes that may occur along the data flow. With effective mapping of change points, your organization can assess how a particular action affects the entire data lifecycle and what the consequences of this action are at the database level.

Strengthened information security:

If there is a security threat, utilizing data lineage tools can help your organization determine if there has been any data leakage or deletion and pinpoint where the compromised data was located within the data flow.

Better data governance:

If there is a security threat, utilizing data lineage tools can help your organization determine if there has.

Limitations in Implementing Data Lineage Solutions

Despite the clear advantages indicated above, practical implementation of any data lineage tool can raise several issues, especially for large enterprises that have to deal with a very complex data landscape driven by a diversity of data types (structured and unstructured, sent in real-time, near-real-time or asynchronously):

Complex data environments:

If an organization's data is submitted to a set of complex transformations, for example, when it is subject to software code, it is necessary to parse the code to trace the path of that data. If even in the case of relatively simple query languages like SQL, this parsing can be somewhat complex, in the case of object-oriented programming languages, we are talking about an insurmountable obstacle for most companies.

Third-party proprietary code applications:

In the case of closed-source applications that function like black boxes, the input data sent to an application is subject to several transformations, resulting in the output data returned by an application. However, without the ability to see the source code, it is impossible to determine the mapping between input and output fields. This opacity results in a lower level of detail in the analysis provided by the data lineage.

Changes in data regulations:

Constantly evolving national and international data regulations, such as GDPR in the EU or CCPA in California, make it difficult for organizations to comply. These regulations often presuppose that companies produce a detailed lineage of their data. However, adapting an organization's IT systems to be compliant can be expensive and time-consuming.

When Manual is Better than Automated

We wrote most of this chapter from the assumption that your organization is using automated data lineage tools provided by vendors such as Collibra, Talend, Informatica, and IBM, and because that is the standard approach nowadays to track your data flow. Most of these tools provide real-time tracking of your data and can trace a data source in just a few seconds, regardless of the transformations the data was subjected to and its destination.

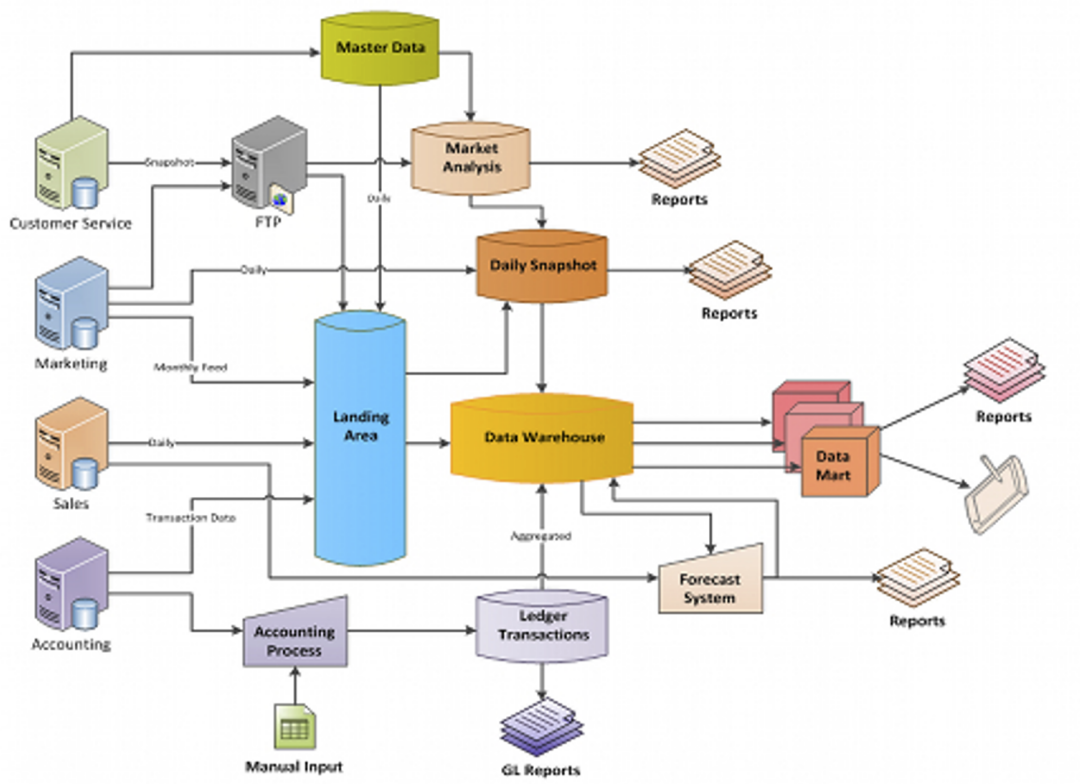
While these automated solutions facilitate data ingestion and enable a better understanding of all the physical elements of the data flow, they can be expensive and require specialized skills. Furthermore, the challenges posed by automated data lineage are way more significant if you must support many legacy systems.

On the other hand, manual approaches to data lineage are more suitable for describing your data lineage's conceptual and logical levels. Manual data lineage is, however, much more time-consuming, and inadequate for large organizations. Some popular applications for manually describing data lineage are Microsoft Excel and Visio.

Having said that, if you want to track the data lineage of your organization to enhance the explainability of your AI and ML solutions, we recommend that you adopt a full-fledged automated software solution.

4. How Can Data Lineage Help Achieve Explainability?

In the previous chapter, we mentioned that automated data lineage enables explainability in AI and ML systems by providing a clear audit trail of the data. In this chapter, we will explain in more detail why automated data lineage is indispensable in achieving explainable AI. In this chapter, we will explain in more detail why automated data lineage is indispensable in achieving explainable AI.



By creating a comprehensive map of how data is transported and modified throughout a system, data lineage helps data consumers track changes, understand the relationships between data elements, and detect inconsistencies or errors. These features ensure traceability and transparency of data throughout an organization:

Data flow tracking:

The ability to capture and document the transformations, manipulations, and interactions with various systems or processes throughout the data lifecycle is one of the critical features of automated data lineage tools that directly contribute to improving the explainability of AI and ML systems. This traceability ensures that all stakeholders clearly understand data usage within an organization.

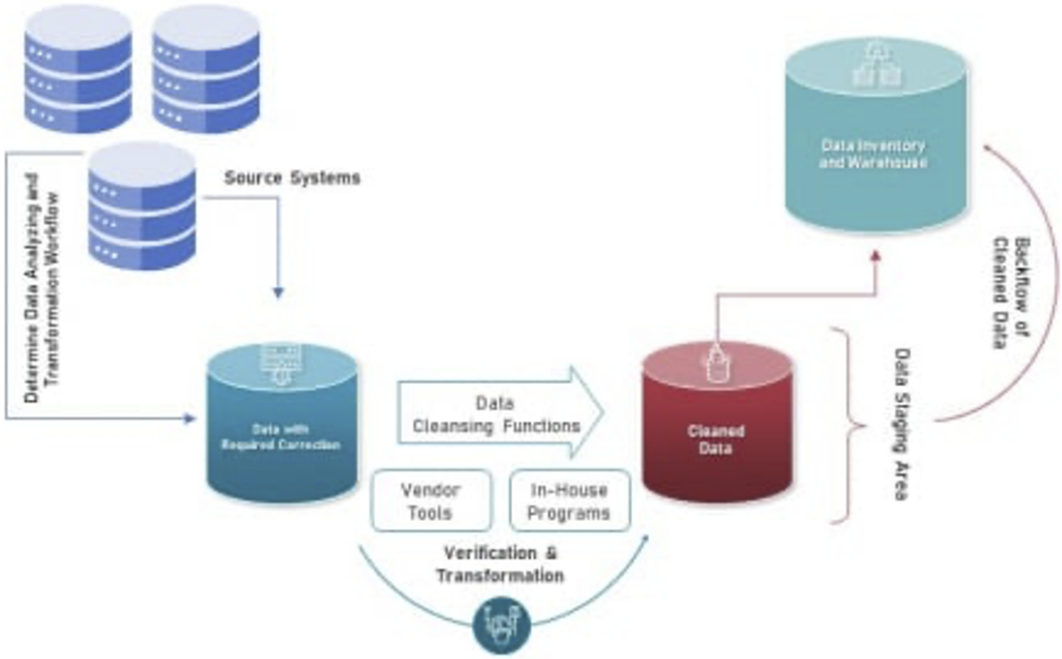
Data flow visualization:

Another critical advantage of automated data lineage solutions is the ability to produce visual representations such as flow diagrams or graphs that help to detect the source, transformations, and destination of data, thus contributing to greater transparency in data movements.

Validation of data sources and transformations

By providing access to data sources, automated data lineage solutions allow consumers to check and validate data sources. In addition, these tools often provide clues to determine the reliability and trustworthiness of data used to train AI and ML models. This type of validation ensures that the input data is reliable and fit for purpose.

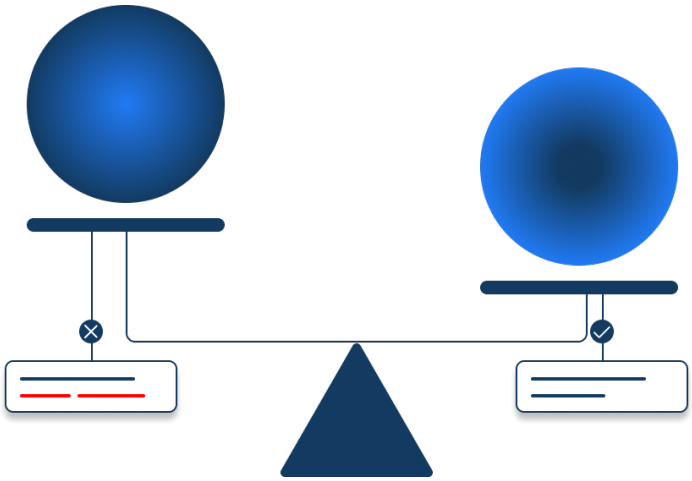
Data validation process with source system



Another validation ability provided by data lineage is the ability to visualize the transformations applied to the data during processing, which helps stakeholders understand how data manipulation, aggregation, and cleaning occurs. This understanding of the transformation processes enhances the interpretability and reliability of IA and ML models.

Ultimately, these validations help improve your data quality by looking out for defects, such as missing values, incorrect data, or wrong data types, that can lead to inaccurate results.

Ensuring fairness: detection and correction of biases



Since AI applications are prone to biases that can affect their performance and fairness, organizations responsible for them must detect and remove these biases from the data used to train the base models. Automated data lineage solutions can quickly identify and remedy these biases through powerful error analysis tools. With automated data lineage, organizations can now quickly complete tasks that used to take hours or even days. The system can detect and eliminate biases that previously required a manual process of reviewing records individually, which saves time and reduces the need to start over with the entire dataset.

This way, automated data lineage helps to achieve fairness in AI and ML models, allowing an organization to assess whether these models show discriminatory behavior or unintended biases.

5. Implementing Automated Data Lineage for Explainability

In today's market, there are many data lineage tools to choose from. Each tool has unique features, supports different integrations, and offers specific scalability options. It's important to evaluate these tools based on your organization's specific requirements and business needs.

To ensure explainability, automated data lineage solutions should have specific essential components. Firstly, a comprehensive solution should be able to automatically display the connections between data from various sources.

Secondly, the solution should track all data flows automatically. Manual lineage definition is not only resource-intensive but also prone to errors. Additionally, manually documented lineage can become outdated and difficult to maintain due to frequent changes in data flows.

To meet current market demands, a complete solution should automatically derive end-to-end lineage from a diverse range of fragmented data sources, whether cloud-based or on-premises. Ideally, the solution should also automate the capture of detailed data transformation details from sources such as SQL scripts, stored procedures, business intelligence reports, and Extract, Transform, and Load (ETL) jobs.

Here is a list of some of the leading automated data lineage tools in the market:

- **The Alation Data Catalog** is a helpful tool for businesses looking to improve their data governance. It offers features such as data lineage tracking and searchability to help organize data effectively. Regardless of the storage solution adopted, whether on-premises, cloud-based, or a hybrid approach, organizations can easily manage all their data. With it, users can better understand every data source, transformation, and dependency.
- **Apache Atlas** is an open-source data governance and metadata framework for organizations with data-intensive applications. Despite being primarily built for Hadoop clusters, it also supports metadata sharing with other frameworks and tools. Its data lineage capabilities include automatically tracking and visualizing any changes made to the data as it moves through various processes along its lifecycle.
- **Collibra Data Lineage** is a data lineage tracking solution that is part of Collibra's comprehensive data governance platform. This solution provides end-to-end data visibility within your organization, from origin to source. Its main feature is the automatic mapping of relationships between systems and processes. It also includes a user-friendly interface and easily integrates with other data tools and platforms, such as Collibra's own Data Governance tool, to link data lineage information with data governance policies and rules.
- **IBM InfoSphere Information Governance Catalog** is a metadata management solution that also provides data lineage features allowing you to track and document the data flow of your organization. The solution uses advanced algorithms to automatically discover and capture data lineage information, thus reducing manual effort and ensuring accuracy. Besides providing comprehensive visibility into the entire data flow, it enables users to understand how data is processed and transformed across the whole data landscape.

- **Informatica Metadata Manager** is an enterprise metadata management platform that can provide complete views of data lineage through a visual map of the data flow through the data integration environment. The tool collects this metadata from enterprise applications, repositories, databases, and data modeling/business intelligence tools. By leveraging AI and ML techniques, it can create a knowledge graph of an organization's data assets and their relationships.
- **Manta** is a specialized data lineage platform that tracks and maps all your data flows. It lets users visualize the entire data journey through an intuitive user interface and offers robust search and filtering capabilities for quickly locating particular data items. Moreover, it provides reporting tools for impact analysis and allows you to compare two different times to understand how data flows have changed. Finally, it supports integration capabilities with several other data platforms and tools.
- **Octopai Data Lineage XD** is a data lineage tool with data discovery and lineage tracking. It enables organizations the ability to see their data flow across three layers of automated data lineage:
  - » **Cross-System:** End-to-end lineage at the system level.
  - » **Inner-System:** Column-level lineage within an ETL process, report, or database object.
  - » **End-to-End Column:** Column-to-column-level lineage between systems.
- **Talend Data Catalog** is a data catalog that acts as a metadata repository. The repository can be stored on-premises or via a cloud service. It offers powerful search and discovery tools to extract metadata from various data sources. Its main goal is to automate the data pipeline management process. Besides automatically crawling data sources, it leverages ML techniques for data classification and provides custom user access controls for better security and compliance.

6. Future Trends and Implications

As we saw in the Introduction, explainable AI emerged as an effort to enhance complex AI and ML algorithms with augmented explanations to clarify the decisions made by these models. This attempt to increase the explainability of AI/ML solutions aims to make it easier for all stakeholders to understand and trace back the decision-making process behind those solutions.

Researchers have developed several approaches for achieving explainability in the last few years to illuminate the opaque "black boxes" of the advanced ML models used to make complex predictions and decisions. These techniques offer detailed insights into how the models arrive at their predictions and help understand the impact of various features on the models' decisions.

What follows is a brief description and explanation of the taxonomy made by Christopher Molnar in his book *Interpretable Machine Learning* (2022, chapter 3, section 2) of some of the most advanced methods for machine learning interpretability[1] : Intrinsic versus extrinsic, model-specific versus model agnostic, and local versus global. We add a fourth distinction to Molnar's taxonomy, the one between model-centric and data-centric models (Bhattacharya, 2022, p. 16; Munn and Pitman, 2022, p. 242).

Intrinsic versus extrinsic (or post hoc) models

While some models are designed to be self-explanatory, requiring no additional effort to interpret their decisions, others are more complex, which makes it challenging for human beings to understand them. Explainability can thus only be achieved later by applying methods to them after model training. Simpler models are then generally considered to have intrinsic explainability, while more complex models are said to have extrinsic or post hoc explainability. Examples of intrinsic AI and ML methods include short decision trees, linear models such as linear regressions, logistic regressions, Generalized Additive Models (GAMs), and k-nearest neighbors. Post-hoc methods are generally used to interpret obfuscated models such as deep neural networks. Most of the explainability techniques we will explain below are post hoc methods. They emerged in the last few decades as a response to the increasing complexity of AI and ML models.

[1] For this white paper, we assume that explainability and interpretability are equivalent. However, we recognize that many authors refer to explainability as the process of understanding why a system made a given decision and interpretability as the process of understanding how a system arrived at a decision and what it means.

Model-specific versus model-agnostic methods

Some explainability methods apply only to specific models and provide a deeper understanding of only that model. Because they require an inner knowledge of how the model works, model-specific methods are used in intrinsically interpretable models. For example, a visualization tool for decision tree models is specific to the decision tree algorithm. Another example of a model-specific technique is the analysis of regression weights (Pruksachatkun et al., 2023, chapter 3) in linear models.

On the other hand, model-agnostic methods can be used to explain the predictions and decisions made by any ML model, regardless of its algorithms. Most model-agnostic methods are post hoc methods and are applied after the model has been trained. These methods are more flexible than model-specific methods because they do not have access to the model structure or weights. Recently, the research area has seen the rise of several new promising advanced techniques such as LIME - Local Interpretable Model-agnostic Explanations and SHAP - SHapley Additive exPlanations, which we will briefly describe later.

Local versus global methods

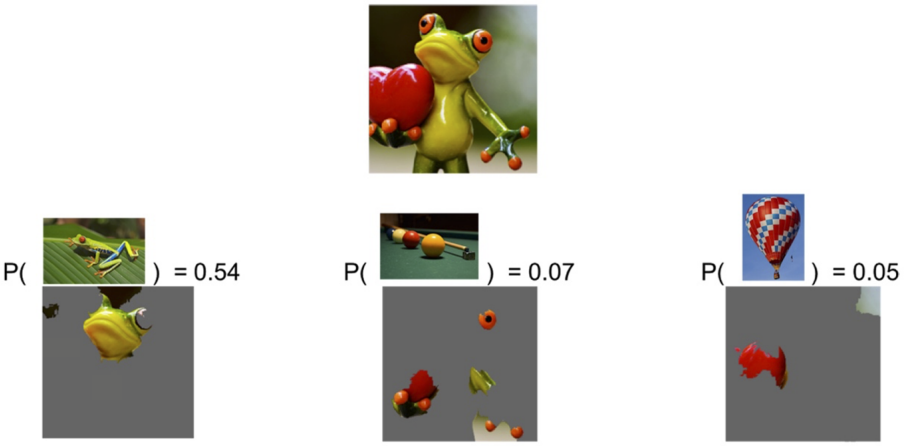
While some explainability methods aim to provide explanations for individual predictions or decisions made by the model by describing how a given data element was weighted in the final prediction or decision, others aim to explain the behavior of the entire model. This is the distinction between local and global methods. In practice, the scope of most models is always somewhere between local and global.

Modern explainability techniques

Typically, most methods used in explainable AI have been model-centric because they focus on the complex elements used by AI and ML models to make predictions and decisions. They aim to optimize the performance of the model by improving its algorithms. However, these model-centric methods may contain biases that can sometimes lead to inaccurate estimates of how the models account for different data elements. Furthermore, if the quality of the data used to train the model is inadequate, this will inevitably affect the quality of the model.

Hence the rise in recent years of data-centric explainable AI methods that emphasize the accuracy, completeness, and appropriateness of the data in light of the business problems being addressed by the system. Some approaches employed to achieve data-centric explainable AI rely on inspecting the data's volume, consistency, purity, and integrity (Bhattacharya, 2023, pp. 61-80). Data lineage tools such as the ones we covered in the last chapter can be here of much help.

Modern Explainability Techniques

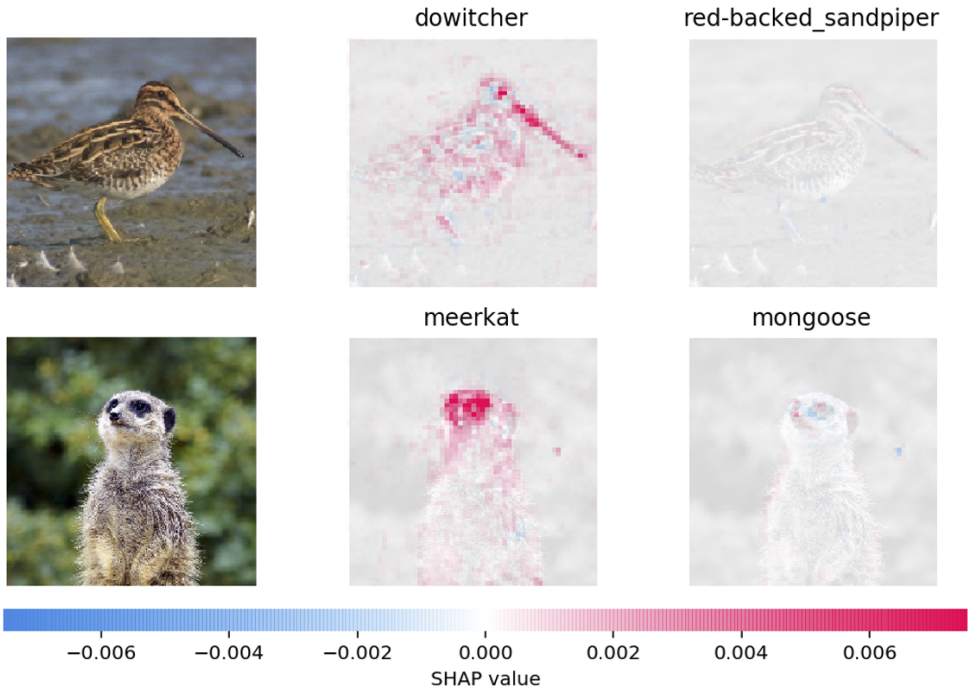


(taken from Ribeiro et al., 2016)

LIME

As the name hints, Local Interpretable Model-agnostic Explanations (LIME) is a local and model-agnostic model. Researchers from the University of Washington developed it to interpret how an ML model works by providing insights that make it easier to understand and trust the predictions. It performs several perturbations in the input data to analyze how these changes in the model components (for example, removing words or hiding parts of an image) impact the predictions and decisions made. As it continues to generate new data sets made of perturbed samples and the corresponding predictions, it trains an interpretable model that approximates the original one locally. Since this local approximation is easier than approximating a global model, LIME focuses on explaining individual predictions (Ribeiro et al., 2016; Molnar, idem, chapter 9.2).

LIME has been applied to various classifiers in text and image domains, such as random forests, support vector machines, and neural networks. The explanations generated by LIME help identify cases where the classifier predicts correctly but for the wrong reasons.



(taken from github.com/shap/shap)

SHAP

Shapley Additive exPlanations (SHAP) is another method used to explain individual predictions (Lundberg & Su-In Lee, 2017; Molnar, ibidem, chapter 9.6). Explanations are provided by computing the importance of each feature for a particular prediction. This method relies on assigning importance values to each prediction. The values are based on the game theory concept of Shapley values. SHAP measures the average marginal contribution of a feature value of all possible coalitions. While coalitions are combinations of features used to estimate the Shapley value of a specific feature, the marginal contribution is the difference between two predictions.

SHAP has been applied to various models, including ensemble models, deep learning models, and other complex models.

Automated Data Lineage for Explainable AI: Privacy and Security Best Practices

Although automated data lineage, as we saw in the fourth chapter, can help provide explanations for the predictions and decisions made by AI and ML systems (particularly by ensuring the fairness of the inputs via the removal of biases), it also may create potential vulnerabilities in terms of privacy and security. Here is a list of best practices organizations should observe if they are already using automated data lineage solutions for achieving explainability in AI and ML solutions:

Enforce data privacy:

Automated data lineage involves collecting and storing information about the data flow within AI/ML systems. This data may contain sensitive or personally identifiable information (PII). To protect data privacy, it's important to implement robust data anonymization techniques or use differential privacy mechanisms (Jarmul, 2023, pp. 27-57) to ensure that third parties cannot trace individual data points back to specific individuals.

Implement data access controls:

Automated data lineage can provide detailed information about the data sources, transformations, and usage. Organizations should restrict this lineage information to authorized personnel only. Implementing proper access controls can help prevent unauthorized users from accessing sensitive data and ensure that only authorized individuals can view and analyze the data lineage.

Secure data storage:

Data lineage information must be stored securely to prevent unauthorized access or tampering. Encrypting the data at rest and in transit can help ensure that even if someone gains access to it, they can only read or understand its contents with the proper decryption keys.

Protect against data leakage:

Data lineage information can provide insight into the underlying data and model architecture. Organizations must be careful not to inadvertently expose proprietary information or trade secrets through the data lineage. Regularly auditing access to the data lineage and monitoring for potential data leaks is critical.

Provide model explainability while preventing leaks:

While data lineage can improve model explainability, it's essential to strike a balance between providing insights and preventing model leakage. Model leakage occurs when the explanations inadvertently reveal too much information about the model's internal parameters or sensitive data used during training.

Establish data retention policies:

Organizations should establish clear data retention policies for the data lineage information. Keeping lineage data longer than necessary increases the risk of data breaches. Establishing and adhering to appropriate retention periods can minimize exposure.

Monitor third-party integration:

When using third-party tools or platforms for automated data lineage, it's critical to evaluate their privacy and security practices. Ensure that these tools comply with relevant data protection regulations and industry standards.

Conclusion

In this white paper, we explored the importance of data lineage in achieving explainability in AI and ML systems. While explainable AI or XAI provides information about the data used in machine learning algorithms, it does not indicate its reliability. As AI gains traction in the enterprise to drive innovation and gain a competitive advantage, it is crucial to have transparent, explainable, and trustworthy AI solutions. One way to address this need is through automated data lineage. Data lineage tools help keep track of the data's journey from its source system to the ML model, showing how it was transformed along the way.

In the *second chapter*, we learned about the importance of explainability in trustworthy AI and ML systems. Explainability refers to the ability of these systems to provide clear and transparent explanations for their predictions and decision-making processes. Practitioners should consider various aspects, including information conveyed, trust, fairness, transparency, causality, generalization, reliability, accessibility, and privacy, to ensure explainability. It is recommended that business leaders prioritize explainability in their organizations' AI and ML solutions to increase adoption, enhance user trust, and promote social responsibility by avoiding bias and ensuring regulatory compliance.

In the *third chapter*, we discussed how automated data lineage enables data tracking throughout an organization, from its source to its point of use. We described the leading data lineage methods, including data tagging, code parsing, and pattern-based lineage. We also distinguished between horizontal data lineage, which focuses on the physical path data takes across systems, and vertical data lineage, which aims to analyze the links between different processes. While data lineage offers several benefits, including a better understanding of data flow, error detection, impact analysis, improved security, and enhanced data governance, it can also present challenges in complex data environments, integrating proprietary third-party applications, or adapting to changing data regulations. Despite the benefits of automated data lineage tools for real-time data tracking - especially in large organizations - there are some use cases where manual approaches are more appropriate than automated ones.

In the *fourth chapter*, we emphasized the importance of data lineage in achieving explainable AI and ML systems. Automated data lineage has two critical features for explainability: data flow tracking, which documents data transformations and interactions between systems, and data flow visualization, which creates visual representations such as flowcharts to increase data transparency. Automated data lineage also enables validating data sources and transformations, ensuring data reliability and accuracy. It helps identify and correct biases in AI models, promoting fairness. Finally, it saves time by automating bias detection, resulting in more efficient and trustworthy AI and ML models.

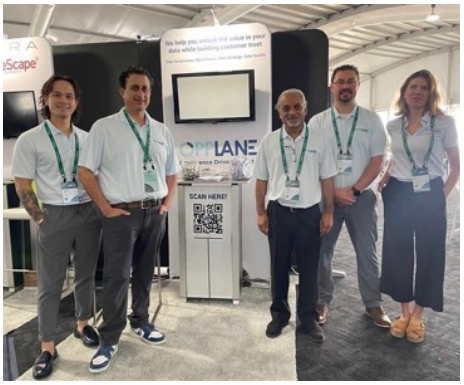
In *Chapter 5*, we discussed how organizations can achieve explainability in AI and ML by leveraging automated data lineage tools. These tools track and display data connections, flows, and transformations, providing end-to-end lineage from multiple sources, including on-premises and cloud-based data. Compared to manual lineage definition, automated solutions are less resource-intensive and less prone to error. There are several automated data lineage tools on the market, including Alation Data Catalog, Apache Atlas, Collibra Data Lineage, IBM InfoSphere Information Governance Catalog, Informatica Metadata Manager, Manta, Octopai Data Lineage XD, and Talend Data Catalog. Choosing the right automated data lineage tool depends on your organization's requirements and business needs. Doing so can improve data management and ensure transparency and traceability of data usage.

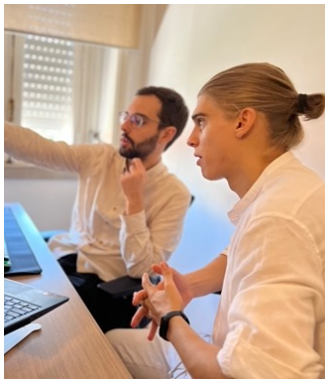
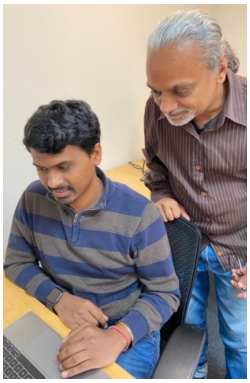
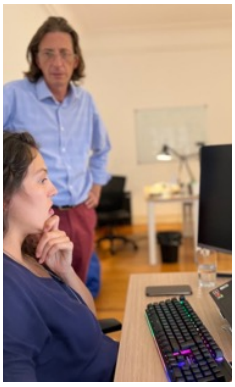
Explainable AI aims to improve complex AI and ML models by providing explanations for their decisions. In *Chapter 6*, we reviewed several approaches to achieving explainability in AI and ML models, including intrinsic and extrinsic models, model-specific and model-agnostic methods, local and global methods, and model-centric and data-centric methods. We also discussed two modern explainable AI techniques: LIME (Local Interpretable Model-agnostic Explanations), which explains individual predictions by modifying the input data, and SHAP (Shapley Additive exPlanations), which calculates the importance of features for individual predictions by using game theory concepts. While automated data lineage can help provide explainability, it also raises privacy and security concerns. Organizations must take steps to ensure data privacy, implement access controls, secure data storage, and prevent data leakage. It is also important to establish data retention policies, monitor third-party integrations, and provide model explainability while avoiding leaks.

A key takeaway from this white paper we want to leave you with is the importance of automated data lineage to ensure the trustworthiness and accountability of AI/ML models. The methods and approaches discussed here allow organizations to make crucial decisions based on reliable insights. Automated data lineage allows tracking data movements, transformations, and connections within AI/ML pipelines, clarifying how decisions are made. It also serves as a valuable tool for demonstrating compliance with data privacy regulations and industry standards, reducing legal risks and potential liabilities.

References

- Bhattacharya, Aditya (2022). Applied Machine Learning Explainability Techniques. Birmingham, UK, Packt Publishing.
- DAMA International (2017). DAMA-DMBOOK: Data Management Body of Knowledge. Basking Ridge, NJ, Technics Publications
- Freche, J., Heijer, M. den, and Wormuth, B. (2021). "Data Lineage" in Liermann, V. & Stegmann, C. The Digital Journey of Banking and Insurance, Volume III: Data Storage, Data Processing and Data Analysis, pp. 5-19. Cham, Switzerland, Springer Nature Switzerland AG.
- Jarmul, Katharine (2023). Practical Data Privacy: Enhancing Privacy and Security in Data. Sebastopol, CA. O'Reilly Media
- Kamath, Uday & Liu, John (2021). Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning. Cham, Switzerland, Springer Nature Switzerland A
- Lundberg, Scott & Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA: *https://arxiv.org/abs/1705.07874*.
- Molnar, Christoph (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Independently published: *https://christophm.github.io/interpretable-ml-book/*.
- Munn, Michael & Pitman, David (2022). Explainable AI for Practitioners: Designing and Implementing Explainable ML Solutions. Sebastopol, CA. O'Reilly Media.
- Pruksachatkun, Y., McAteer, M, and Majumdar, M. (2023). Practicing Trustworthy Machine Learning: Consistent, Transparent, and Safe AI Pipelines. Sebastopol, CA. O'Reilly Media
- Ribeiro. M. T., Singh, S, and Guestrin, C. (2016, August 12). "Local Interpretable Model-Agnostic Explanations (LIME): An Introduction." O'Reilly: *https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/*.
- Southekal, Prashanth H. (2023). Data Quality: Empowering Businesses with Analytics and AI. Hoboken, NJ. John Wiley & Sons.
- Strong, Diane M. and Wang, Richard Y. (1996). "Beyond Accuracy: What Data Quality Means to Data Consumers" in Journal of Management Information Systems, Vol. 12, No. 4, pp. 5-33. *http://mitiq.mit.edu/documents/publications/tdqmpub/14_beyond_accuracy.pdf*.





About Opplane

Opplane is a company led by former PayPal leaders who bring years of leadership experience from various domains and top organizations. Our experienced team specializes in data modernization programs, covering areas such as privacy management, banking, fintech, security, machine learning, risk management, and cloud infrastructure. This expertise ensures comprehensive solutions for data modernization initiatives.

Our company operates globally with offices in Silicon Valley (USA), Western Europe, India, and Singapore. This wide geographic presence brings numerous advantages, including access to diverse talent, continuous support, and customized solutions for regional markets. Leveraging the expertise of our international team, we excel in delivering exceptional results at scale.

We assist organizations with data transformation and digitization, providing comprehensive solutions including end-to-end implementations of cloud-native platforms, data lakes, and privacy tools. Additionally, we utilize machine learning algorithms to gain consumer insights and enhance consumer experiences through personalization.

OPPLANE

Address:

US office:

Opplane Inc suite 300
6200 StoneRidge Mall
Pleasanton , CA 94588

Portugal office:

Rua Camilo Castelo Branco
nr 2 - 3º esquerdo
1150 - 084 Lisboa

India office:

204, Express Arcade
Netaji Subhash Place
Plot No. H10, Pitampura Delhi
North West, Delhi, 110034

Email:

contact@opplane.com

Website:

www.opplane.com