The Role of Data in AI Quality

Direct Digital Holdings

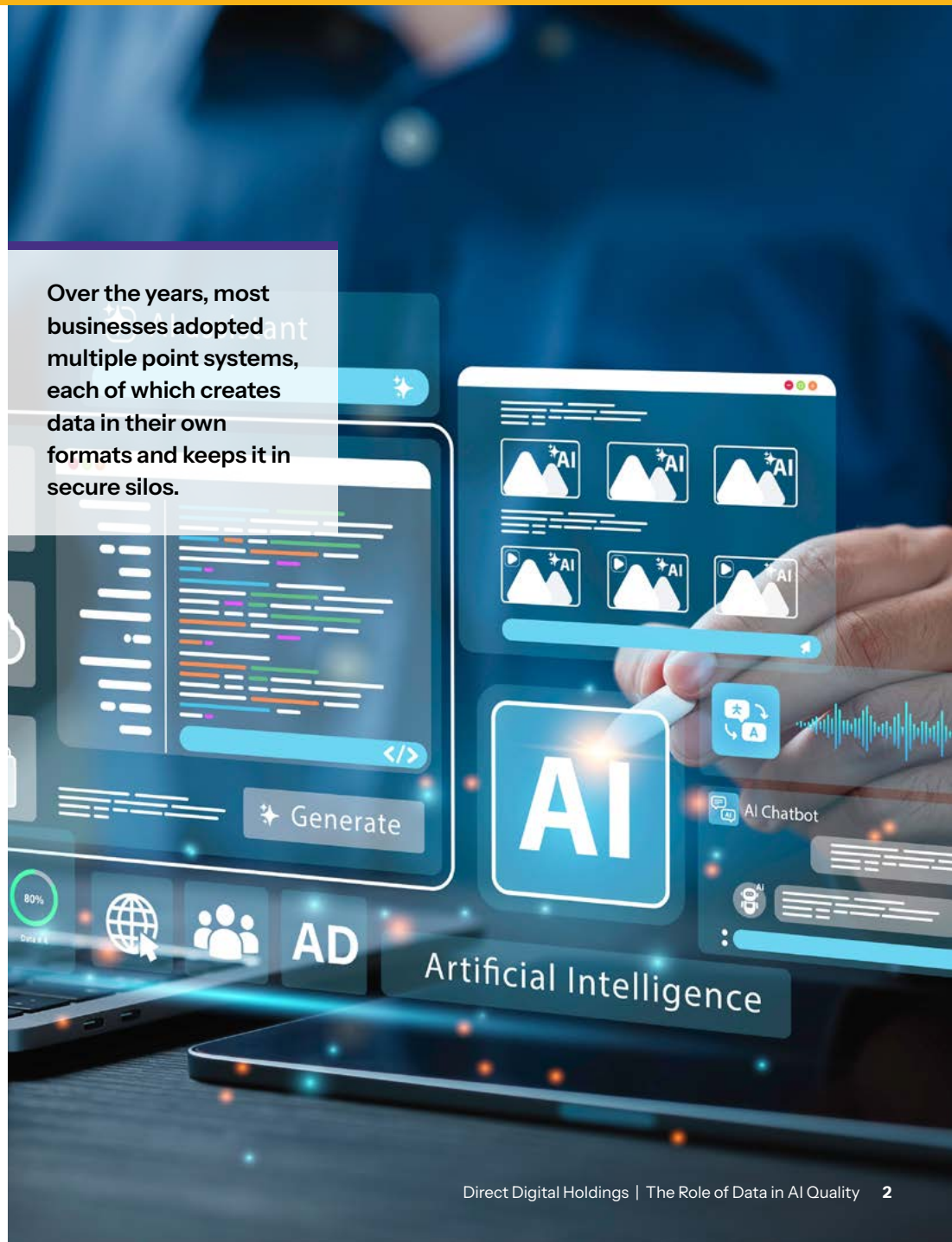
Colossus SSP® Orange 142®

Forward

All AI systems need high-quality data to work properly. As the saying goes: garbage in, garbage out. But how do business leaders know the quality of their data? This guide offers an overview of the issues organizations face in terms of the data. Over the years, most businesses adopted multiple point systems, each of which creates data in their own formats and keeps it in secure silos.

Getting data ready for AI takes a bit of heavy lifting. The goal of this guide isn't to provide data scientists with a step by step roadmap for getting your data AI-ready. It's to provide business leaders with an overview of the scope of work necessary so that you can set reasonable expectations for the road ahead.

With this understanding, the DDH AI Council seeks to help you make informed decisions about resource allocation, timelines, and investment in data preparation, all of which are critical first steps before embarking on any AI initiative.



Direct Digital Holdings
Colossus SSP · Orange 142

Important Note

The information and examples provided here are designed to serve as starting points for organizations navigating the evolving landscape of generative AI. They are provided for illustrative purposes only and do not constitute legal, regulatory, or operational advice.

Organizations should tailor these materials to align with their specific needs, industry requirements, and legal obligations. We strongly recommend consulting legal, compliance, and data privacy experts to ensure adherence to applicable regulations and the safeguarding of organizational interests. All information is provided “as is,” without warranties of any kind, and should be used at your discretion.

About the DDH AI Council

The DDH AI Council was founded to address a growing concern: the widening divide between organizations that embrace generative AI and those that are hesitant to adopt it. Generative AI is rapidly reshaping how we work, raising the overall caliber while enabling teams to innovate faster. We understand that for many business leaders, generative AI is still an unknown technology that comes with many risks. Our goal is to demystify generative AI, and to provide the education and insights business leaders need to build a roadmap for its adoption, with full confidence that its use will be safe and transformative.

Table of Contents

Direct Digital Holdings is a fast growing, efficiency-focused solutions provider in the digital marketing and advertising sector. We are a family of brands serving direct advertisers, agencies, publishers, and marketers.

Direct Digital Holdings
Colossus SSP · Orange 142

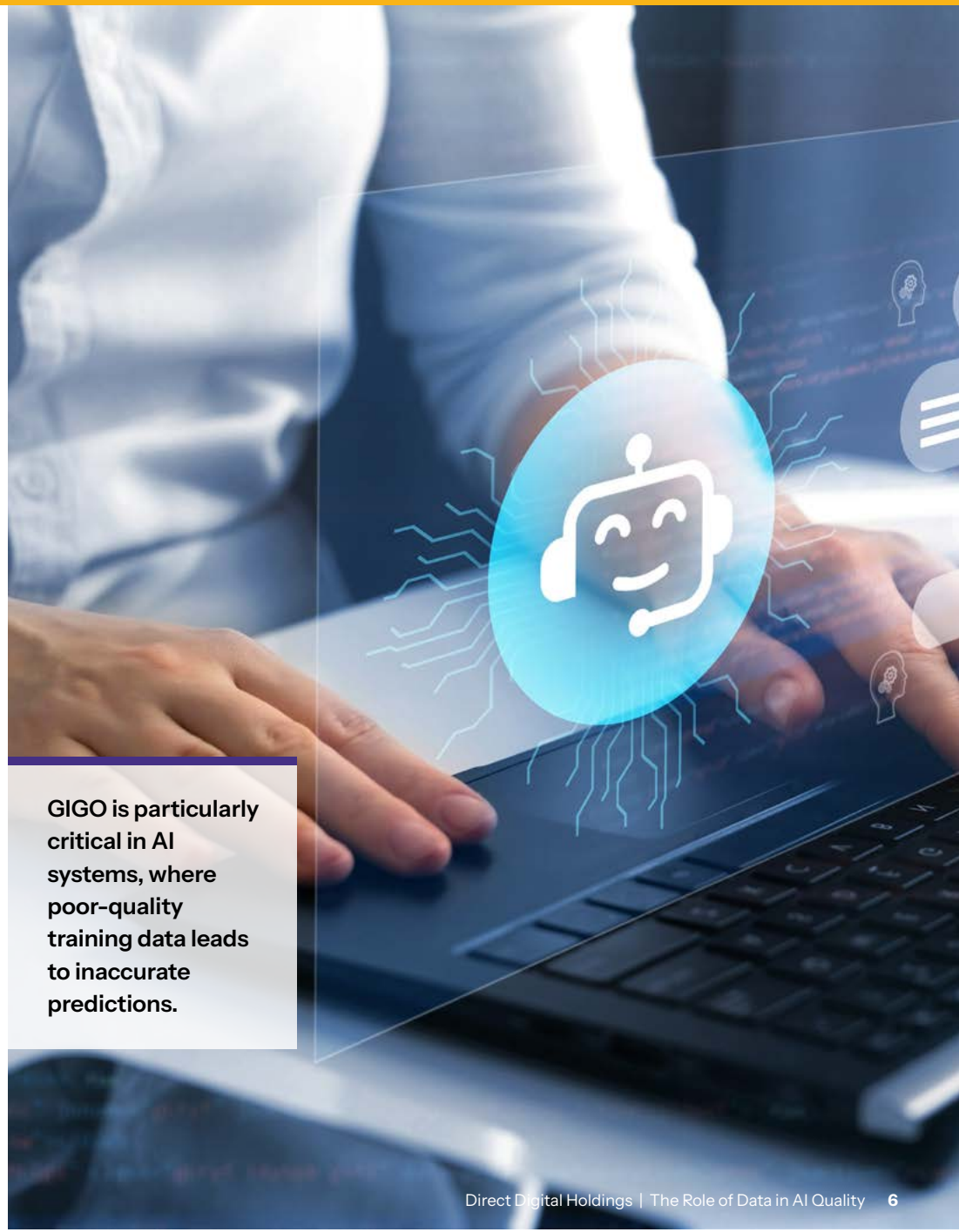
Data Quality Fundamentals

The GIGO Principle

There is a truism in data science, especially in computer science: garbage in, garbage out (GIGO). Coined by IBM programmer George Fuechsel in the 1950s, GIGO highlights a stark reality: computers process data exactly as input, regardless of its accuracy.

GIGO is particularly critical in AI systems, where poor-quality training data leads to inaccurate predictions, biased outcomes, and unreliable models. Consider Amazon's hiring algorithm, which was trained on historically male-dominated resumes, leading to discrimination against women applicants. After many attempts to eliminate the bias, Amazon opted to scrap the project.

The stakes are even higher in healthcare. A study by the Leonard Davis Institute of Health Economics found that predictive models trained on incomplete electronic health data performed poorly for patients with lower access to care, disproportionately affecting Black patients. By contrast, Mount Sinai Hospital in New York City used structured, high-quality patient data to create its Sepsis Surveillance System. Sepsis, a deadly condition that kills over 200,000 people annually, is notoriously difficult to detect. Mount Sinai's AI model monitors patients in real-time to identify early warning signs, reducing mortality rates by 20%.



GIGO is particularly critical in AI systems, where poor-quality training data leads to inaccurate predictions.

Poor data quality isn't just a moral issue; it's a financial one. A study by Actain found that low-quality AI training data significantly increases costs, causing inaccurate predictions, inefficiencies, lost revenue, reputational damage, and the need for manual corrections.

It's fair to say that every business leader understands the risks of GIGO. The challenge lies in answering two critical questions: How do you know if your data is "garbage"? And, if it is, what steps do you need to take to clean it?

Separating the Trash from the Treasure

Six key metrics can help you assess and fix any data quality issues. They are:

Metric	Description	Impact
Accuracy	Ensure data reflects real-world values (e.g., customer addresses match postal records).	Prevents errors and enhances reliability.
Completeness	Fill in missing critical fields to create a complete dataset (e.g., names and emails).	Provides a full picture for decision-making.
Consistency	Maintain uniform data across systems (e.g., CRM IDs match billing records).	Improves integration and reduces errors.
Timeliness	Keep data current and updated regularly (e.g., inventory updated daily).	Avoids outdated insights and increases efficiency.
Relevance	Use only data tied to your AI goals (e.g., demographics for campaigns).	Supports meaningful AI-driven insights.
Format Quality	Follow consistent formats for easier processing (e.g., YYYY-MM-DD for dates).	Ensures data can be processed effectively.

Measuring Data Quality

Measuring data quality requires a systematic approach to tracking and evaluation. Key metrics to look at are the error rates, completion percentages, and duplicate counts of your dataset. There are specialized tools to help data scientists check your data against those metrics, including data profiling software and validation scripts.

Benchmarking is another way to assess your data. Benchmarking scoring systems from companies like Qualtrics let you compare your data quality against industry standards or specific companies in your space. These systems provide a composite rate quality to help you track improvement. You can monitor improvement continuously through real-time checks, alerts, and regular audits.

Success depends on choosing the right combination of tools and metrics for your AI use case and consistently tracking and acting on results.

Common Data Quality Challenges

Data quality issues plague even the best organizations, and they are more common than many realize, including:

- Missing values that create incomplete pictures
- Duplications that muddy analysis
- Format inconsistencies make data hard to process
- Outdated information that leads to bad decisioning
- A biased manner in collecting data, which skews results
- Random errors that introduce noise, leading to bad AI decisions

Catching these issues early saves headaches later.

Now that we've established the foundational metrics for data quality, the following sections outline the practical steps required to ensure your data is ready for AI. Each section begins with a high-level summary for decision-makers, followed by actionable details for your team or data partner you've engaged to clean up your data.

Data Assessment

Data assessment starts with understanding your AI goals. Before diving into data preparation, map your data against your use case requirements. What insights do you need? Where does your current data fall short? This helps you answer a fundamental question: Do you have the data you need for your AI to generate reliable predictions?

Data assessment is all about evaluating the availability and quality of data across your systems. Look for gaps that could affect AI performance. For example, do you have enough historical purchase data if you want your AI to predict customer behavior? Is it reliable?

Sometimes you may find that you need additional data or to enhance what you already have. The key is identifying these needs early to avoid AI project delays.

Data steps include:

Step	Purpose	Implementation
Data Inventory	The first step in determining AI use case suitability and ensuring compliance.	Catalog and identify all available data sources.
Data Quality Audit	Essential for AI applications to function effectively.	Evaluate data for completeness, accuracy, consistency, and relevance.
Gap Analysis	Ensures the AI tool has the necessary inputs for valuable insights.	Identify missing or insufficient data for AI use cases.
Data Classification	Crucial for compliance and security, often legally required.	Categorize data based on sensitivity, type, and purpose.
Data Enrichment	Provides critical context needed for meaningful AI insights.	Add external sources or combine datasets to fill identified gaps.

Identifying data gaps is just the start. Ensuring your AI systems handle data responsibly and build trust among its users demands privacy and compliance are built into the system.

Data Privacy and Compliance

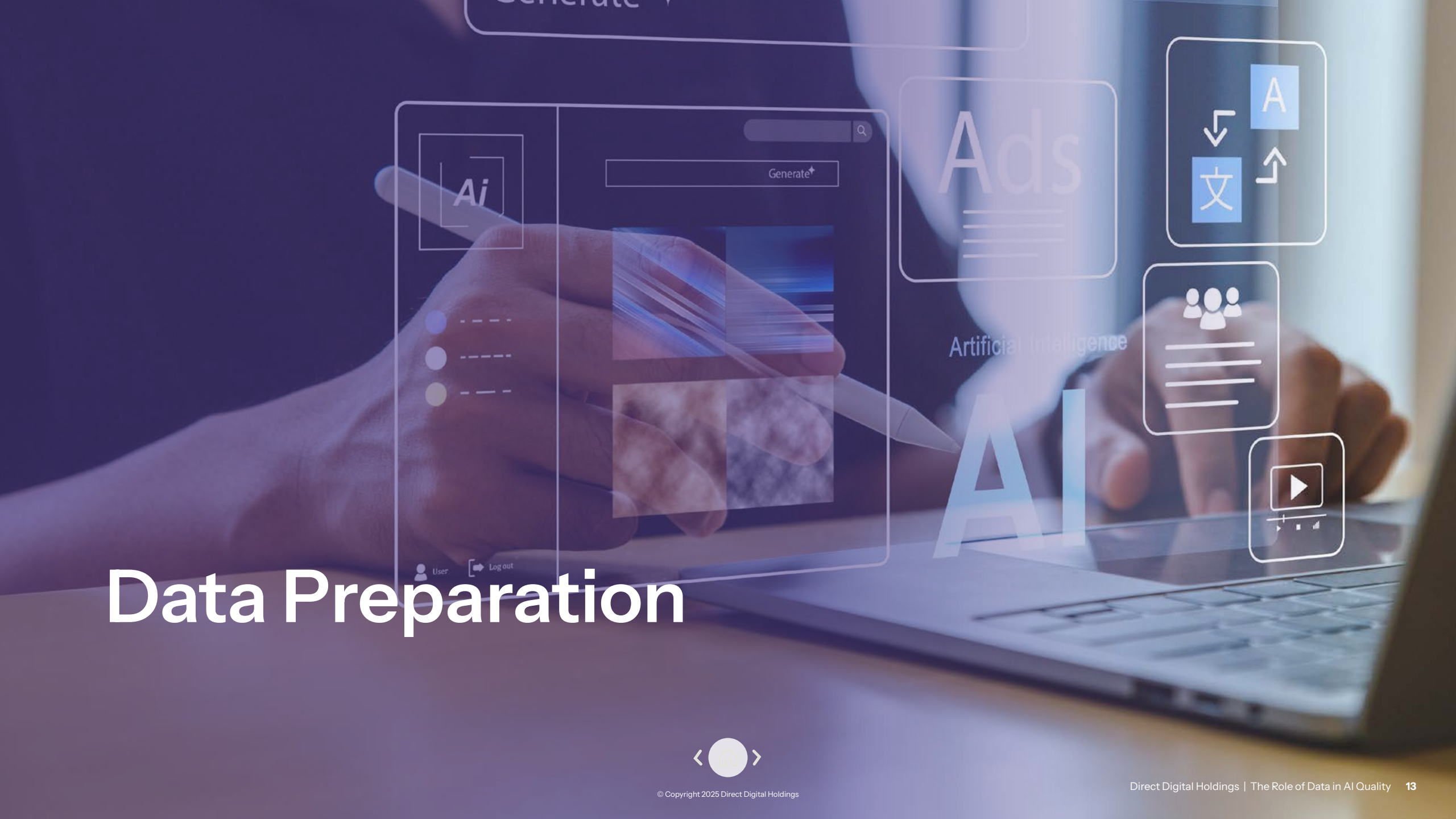
Data privacy and compliance are fundamental to all AI implementations. You need to understand what data you may use to train your AI tool and how to protect it. Regulatory requirements vary by data type and jurisdiction (think: HIPAA's restrictions around electronic healthcare records in the US or non-consented data requirements of GDPR for European citizens).

But proper data handling isn't just about compliance; it's also about trust. You need a governance model that includes clear protocols for data access and anonymization to ensure your AI is used responsibly (see the DDH Responsible AI for further reading).

Data steps include:

Step	Purpose	Implementation
Permission Assessment	Prevents unauthorized data use and legal issues.	Evaluate and document what data can be used, how it can be used, and what consents are required.
Protected Data Handling	Maintains security and regulatory compliance.	Implement security protocols, including encryption, access controls, and secure storage.
Anonymization	Enables data use while protecting privacy.	Apply techniques to remove or encode identifying information while maintaining data usefulness.
Compliance Monitoring	Ensures ongoing adherence to regulations.	Track and document compliance with regulations through systematic audits and reports.
Data Governance	Establishes control and accountability.	Establish and enforce policies for data access, usage, and management organization-wide.

Once you've tackled privacy and compliance, the next step is to ensure your data is properly cleaned and structured so your AI systems can work effectively.



Data Preparation

Data preparation transforms raw information into a format your AI systems can understand and use effectively.

While basic cleaning is a starting point, the process goes deeper — requiring standardization, validation, and integration of multiple data sources. These steps ensure your AI can deliver accurate and meaningful results.

Steps include:

Step	Purpose	Implementation
Standardization	Create uniform data formats.	Convert data to consistent structures, units, and naming.
Cleaning	Remove errors and inconsistencies.	Fix or remove invalid entries, duplicates, and outliers.
Integration	Combine data from multiple sources.	Merge databases while maintaining relationships and quality.
Validation	Ensure data meets requirements for AI use cases.	Test data against quality rules and use case needs.
Documentation	Tracks changes and processes.	Record data transformations, sources, and quality metrics.

Once your data is prepared, the next challenge is managing it effectively — keeping it secure, accessible, and adaptable as your AI needs to evolve.

Direct Digital Holdings
Colossus SSP · Orange 142

Data Management

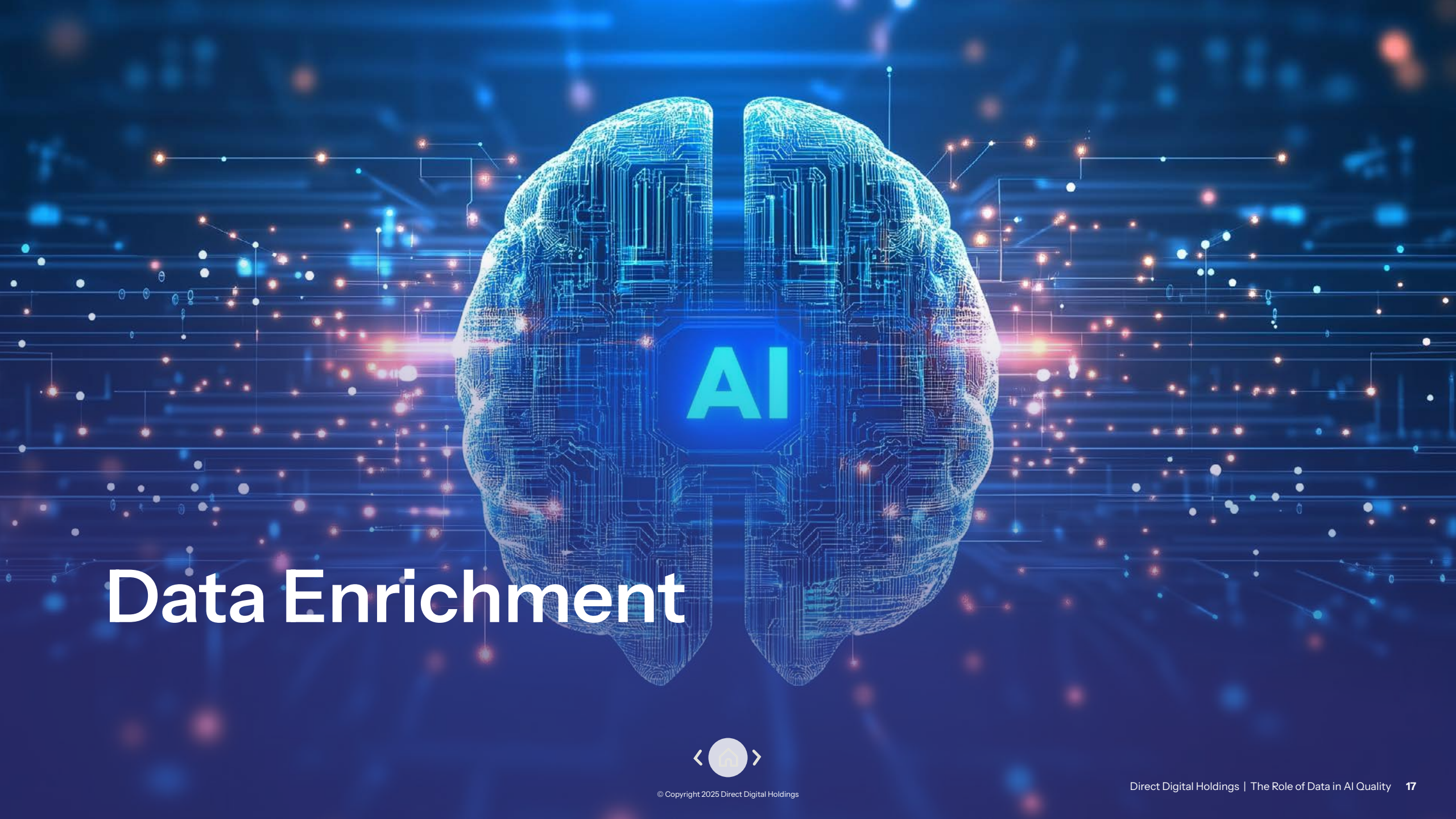
Data is a powerful asset (remember when British mathematician Clive Humby said, “data is the new oil?”). However, it can also be a liability, especially in this era of increasing regulations and frequent data breaches. Keeping unnecessary data is both costly and risky. That’s why many organizations today opt for a lean data strategy, i.e., keeping only what’s essential and managing it securely.

Preparing for AI offers a perfect opportunity to review your data. AI systems thrive on clean, relevant datasets. By eliminating unused or redundant data — such as outdated leads or duplicate records — you will streamline your AI processes, get more accurate predictions, as well as reduce your organization’s exposure to risk. Effective data management involves examining the entire lifecycle of your data, from creation to storage, and implementing strict controls at every step.

Steps include:

Step	Purpose	Implementation
Storage Solutions	Ensure reliable, scalable, and secure storage for AI data.	Choose storage systems like cloud storage (e.g., AWS S3, Azure Blob) or on-premise databases based on data size and access needs.
Access Controls	Protect sensitive data and prevent unauthorized usage.	Implement role-based access controls (RBAC), encryption, and multi-factor authentication to secure data access.
Versioning	Maintain data integrity and track changes over time.	Use tools or platforms that support version control for datasets, allowing rollback and auditing.
Maintenance	Ensure data remains clean, relevant, and up-to-date for AI applications.	Regularly review and update datasets, addressing data rot, redundant entries, or outdated information.
Documentation	Provide clarity on data origin, handling, and transformations.	Record metadata, data processing steps, and usage guidelines in a centralized and accessible location.

With your data properly managed, the next step is adding context and depth through enrichment, unlocking new layers of insight.



Data Enrichment

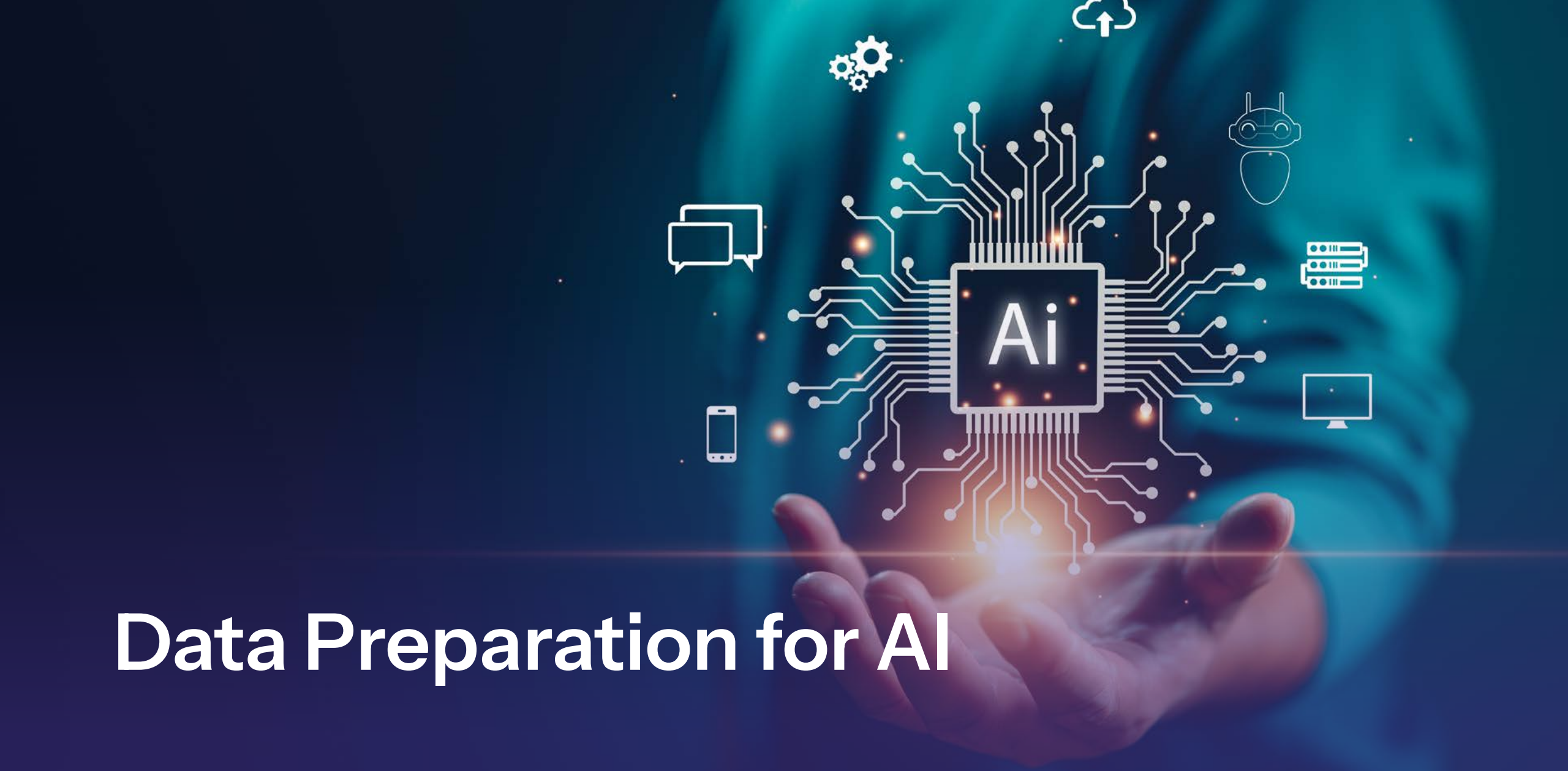
Data enrichment is all about enhancing the value of your existing data by combining it with outside sources or creating different layers of context (e.g., adding demographic insights from the Census Bureau to your customer data). The purpose is to generate more relevant insights from your AI model.

For example, you can enrich your existing data with lifestyle segmentation (e.g., marital status) to better understand your customers. However, more data isn't always better. Any data you add must be high-quality, affordable, and relevant to the answers you want your AI to provide.

Steps include:

Step	Purpose	Implementation
External Data Sources	Expand the dataset with additional relevant information.	Identify and acquire third-party data sources, such as demographic, weather, or industry-specific datasets.
Augmentation Methods	Enhance data variety and depth to improve model performance.	Use techniques like merging external data, creating synthetic data, or filling missing data gaps with reliable sources
Quality Validation	Ensure added data meets quality standards and aligns with use case needs.	Conduct thorough checks for the enriched data's consistency, accuracy, and relevance to prevent errors.
Costs Consideration	Balance data acquisition costs against potential benefits.	Evaluate the cost of data procurement or generation against its expected impact on model performance.
ROI Assessment	Measure the effectiveness of enrichment efforts.	Analyze the impact of enriched data on model outputs, business KPIs, and overall decision-making improvements.

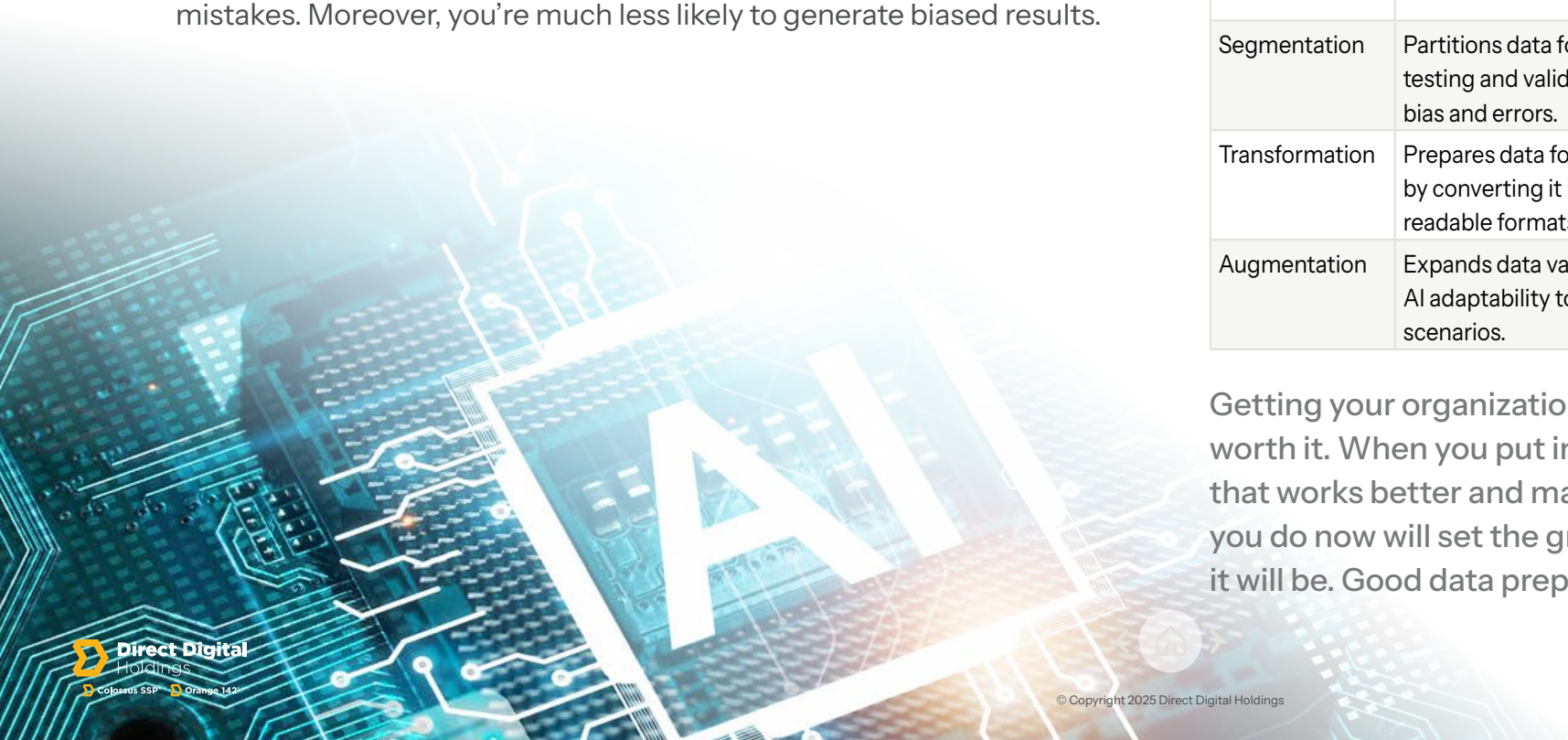
Data enrichment lets you dig deeper into what matters by adding smart layers to what you already know. However, enriched data often needs additional preparation to ensure it's in the right format and context for your AI applications. The next step is transforming this enriched data into AI-ready assets your systems can interpret and learn from effectively.



Data Preparation for AI

Even when your data looks clean and organized, it is more complex to prepare it for AI. AI needs crystal-clear instructions to calculate predictions. That means you need to tell it what each piece of data means and how it should be used. Data scientists spend a lot of time labeling, annotating and dividing data into pieces that AI models can understand.

It's extra work, but the payoff will justify your investments. When you take time with these steps, your AI learns better and makes fewer mistakes. Moreover, you're much less likely to generate biased results.



Steps include:

Step	Purpose	Implementation
Labeling & Annotation	Ensures AI systems learn effectively by accurately tagging supervised learning data.	Tag data (e.g., objects in images) to build precise models tailored to specific use cases.
Feature Engineering	Extracts meaningful patterns from raw data, improving AI performance and insights.	Identify patterns in raw data to create meaningful features, such as aggregating transaction amounts to calculate monthly spending or extracting keywords from the text for sentiment analysis.
Segmentation	Partitions data for thorough testing and validation, reducing bias and errors.	Divide data into training, validation, and testing sets to ensure reliable machine learning workflows.
Transformation	Prepares data for AI analysis by converting it into machine-readable formats.	Apply scaling, encoding, or tokenizing to numerical, categorical, or text data for seamless AI integration.
Augmentation	Expands data variety, enhancing AI adaptability to diverse scenarios.	Incorporate synthetic or additional data sources to supplement limited datasets and improve AI performance.

Getting your organization's data into shape takes real work, but it's worth it. When you put in the time to prepare it right, the result is AI that works better and makes decisions you can trust. And, the work you do now will set the groundwork for the next round of AI, whatever it will be. Good data prep today means better results down the road.

Direct Digital Holdings
Colossus SSP · Orange 142

Parting Thoughts

If AI is the new machine driving innovation, then we can easily make the case that data is the new oil. However, like oil, data must be refined for its intended use, whether gasoline, heating oil, or another form of energy.

Your AI use cases will dictate your data requirements, whether predicting your next high-value customers, driving operational efficiency, or creating better customer experiences. The success of AI hinges on the quality of the data it consumes.

In the end, the true power of AI lies not just in processing data but in transforming it into meaningful outcomes. By prioritizing data quality at every stage, you position your organization to solve today's challenges while anticipating tomorrow's opportunities.

Ultimately, AI's impact depends on how well we, as data stewards, harness its potential to solve meaningful problems and create a better world.

Want to Learn More?

Check out our additional AI resources:

- **Demystifying Generative AI**
- **The Generative AI Playbook: Implementation & Best Practices**
- **The Role of Data in AI Quality**
- **Responsible AI**

Click Here

Direct Digital Holdings
Colossus SSP Orange 142

Glossary of Terms

Artificial Intelligence (AI): The simulation of human intelligence in machines, enabling them to perform tasks such as learning, reasoning, and decision-making.

Generative AI: A subset of AI that creates new content—such as text, images, or code—based on patterns learned from existing data, often using models like GPT or DALL-E.

Supervised Learning: A type of machine learning where the model is trained on labeled data, allowing it to make predictions or decisions based on new, unseen inputs.

Data Labeling: The process of tagging raw data—such as images, text, or audio—with meaningful labels that AI models can use to learn and make predictions.

Data Annotation: A specific type of data labeling where detailed metadata is added to training datasets, enhancing AI models' ability to process and understand information.

Feature Engineering: The process of transforming raw data into meaningful features that improve an AI model's accuracy and performance.

Data Segmentation: Dividing data into subsets, such as training, validation, and testing sets, to ensure comprehensive testing and validation of AI models.

Data Transformation: The conversion of raw data into a structured and machine-readable format using techniques like scaling, encoding, or tokenization.

Synthetic Data: Artificially generated data created to supplement real-world datasets, used to address data scarcity or enhance model training.

Bias Mitigation: Efforts to identify and reduce biases in data or AI models to ensure fair and equitable outcomes across diverse groups.

Data Enrichment: The process of enhancing existing data by integrating it with additional sources or context to provide deeper insights and improve AI outputs.

Disclaimer: The responses provided by this artificial intelligence system are generated by artificial intelligence based on patterns in data and programming. While efforts are made to ensure accuracy and relevance, the information may not always reflect the latest data and programming news or developments. This artificial intelligence system does not possess human judgment, intuition, or emotions and is intended to assist with general inquiries and tasks. Always conduct your own independent in-depth investigation and analysis of ANY information provided herein, and verify critical information from trusted sources before making decisions.

Interested parties should not construe the contents of ANY responses and INFORMATION PROVIDED herein as legal, tax, investment or other professional advice. In all cases, interested parties must conduct their own independent in-depth investigation and analysis of ANY responses and information provided herein. In addition, such interested party should make its own inquiries and consult its advisors as to the accuracy of any materials, responses and information provided herein, and as to legal, tax, and related matters, and must rely on their own examination including the merits and risk involved with respect to such materials, responses and information.

We nor any of our affiliates or representatives make, and we expressly disclaim, any representation or warranty (expressed or implied) as to the accuracy or completeness of the materials, responses and information PROVIDED or any other written or oral communication transmitted or made available with respect to such materials, responses and information or communication, and we, nor any of our affiliates or representatives shall have, and we expressly disclaim, any and all liability for, or based in whole or in part on, such materials, responses and information or other written or oral communication (including without limitation any expressed or implied representations), errors therein, or omissions therefrom.

Direct Digital Holdings

Colossus SSP® Orange 142®

For more information:

Direct Digital Holdings
1177 West Loop South | Suite 1310
Houston, TX 77027
marketing@directdigitalholdings.com

Digital advertising **built for everyone.**

About Direct Digital Holdings

Direct Digital Holdings (Nasdaq: DRCT) brings state-of-the-art sell- and buy-side advertising platforms together under one umbrella company. Direct Digital Holdings' sell-side platform, Colossus SSP, offers advertisers of all sizes extensive reach within the general market and multicultural media properties.

The Company's buy-side platform, Orange 142, delivers significant ROI for middle-market advertisers by providing data-optimized programmatic solutions for businesses in sectors ranging from energy to healthcare to travel to financial services. Direct Digital Holdings' sell- and buy-side solutions generate billions of impressions per month across display, CTV, in-app, and other media channels.

To learn more please visit **directdigitalholdings.com**

Direct Digital Holdings

Colossus SSP® Orange 142®