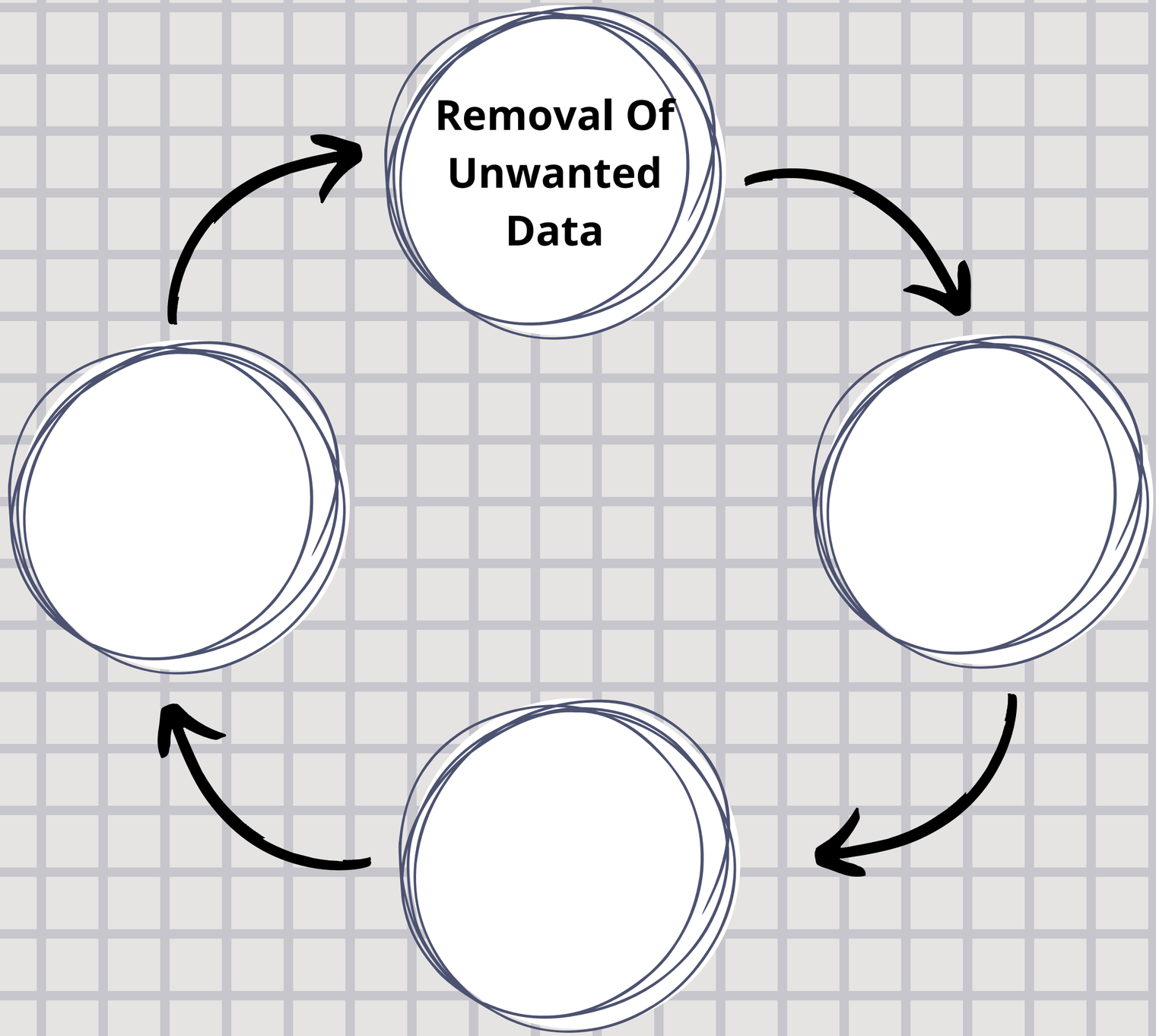
Jenise Bowling, Rhea Jaxon, and Bryan Peavey

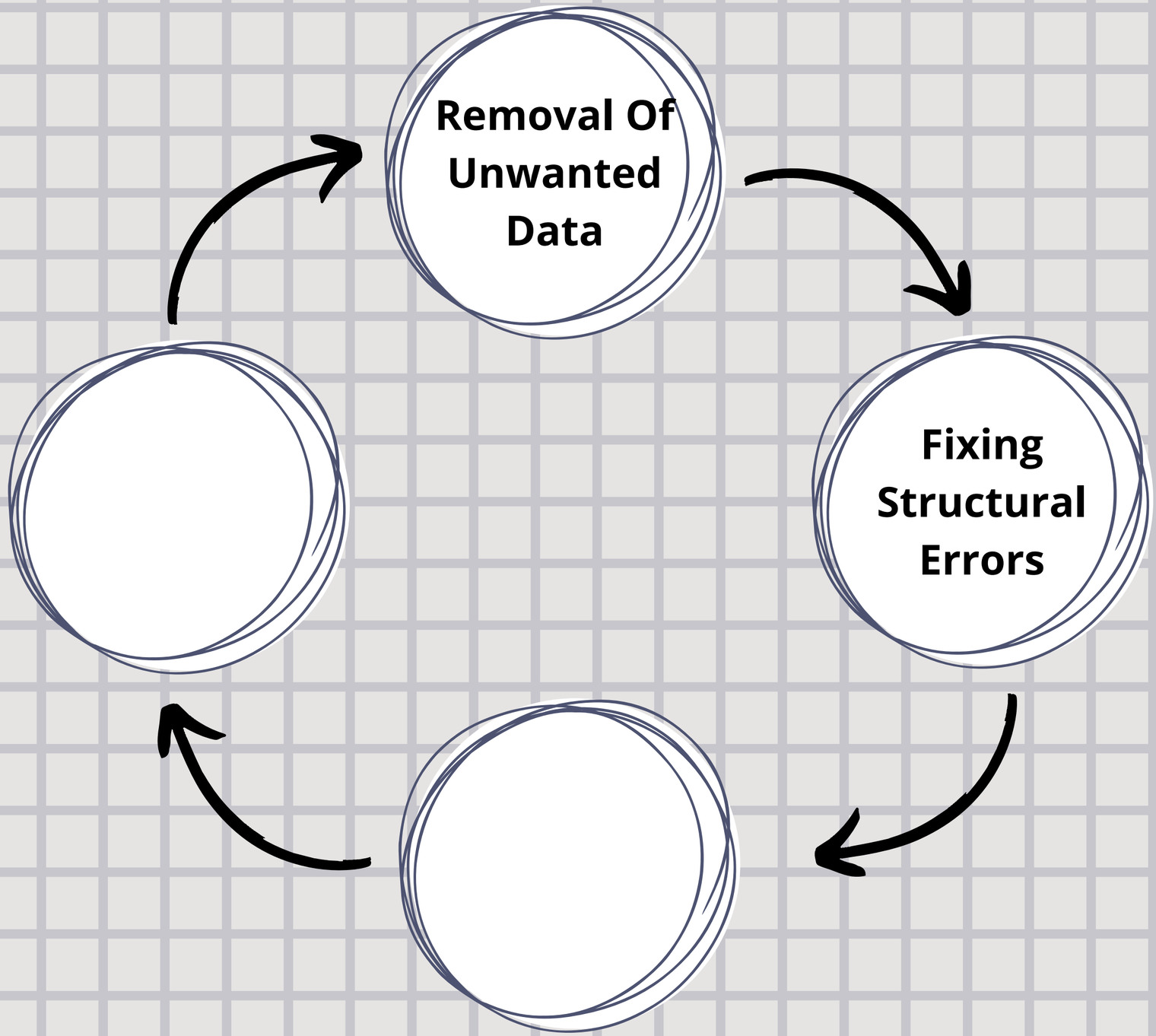
DATA CLEANING AND INTEGRATION

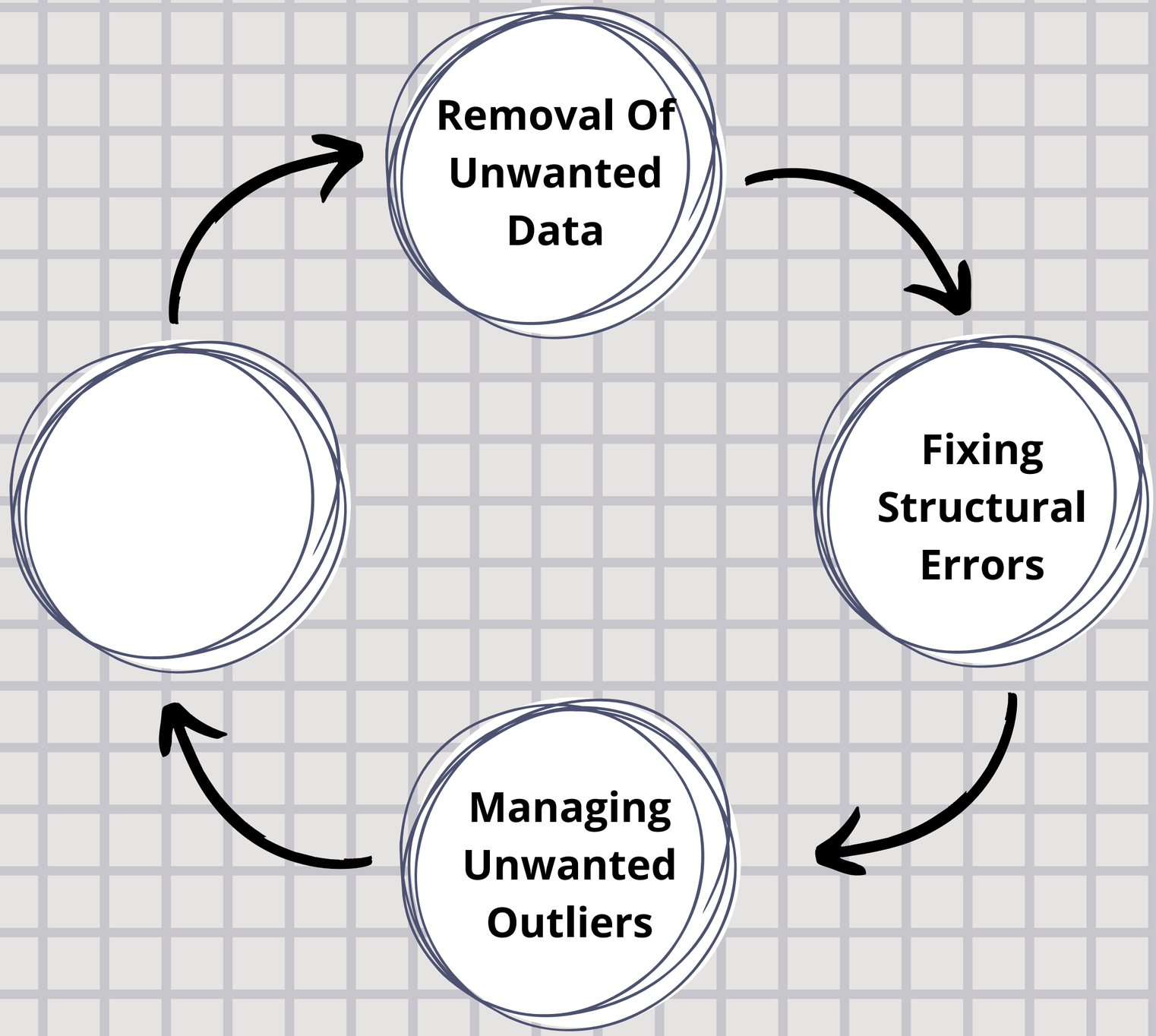
Introduction

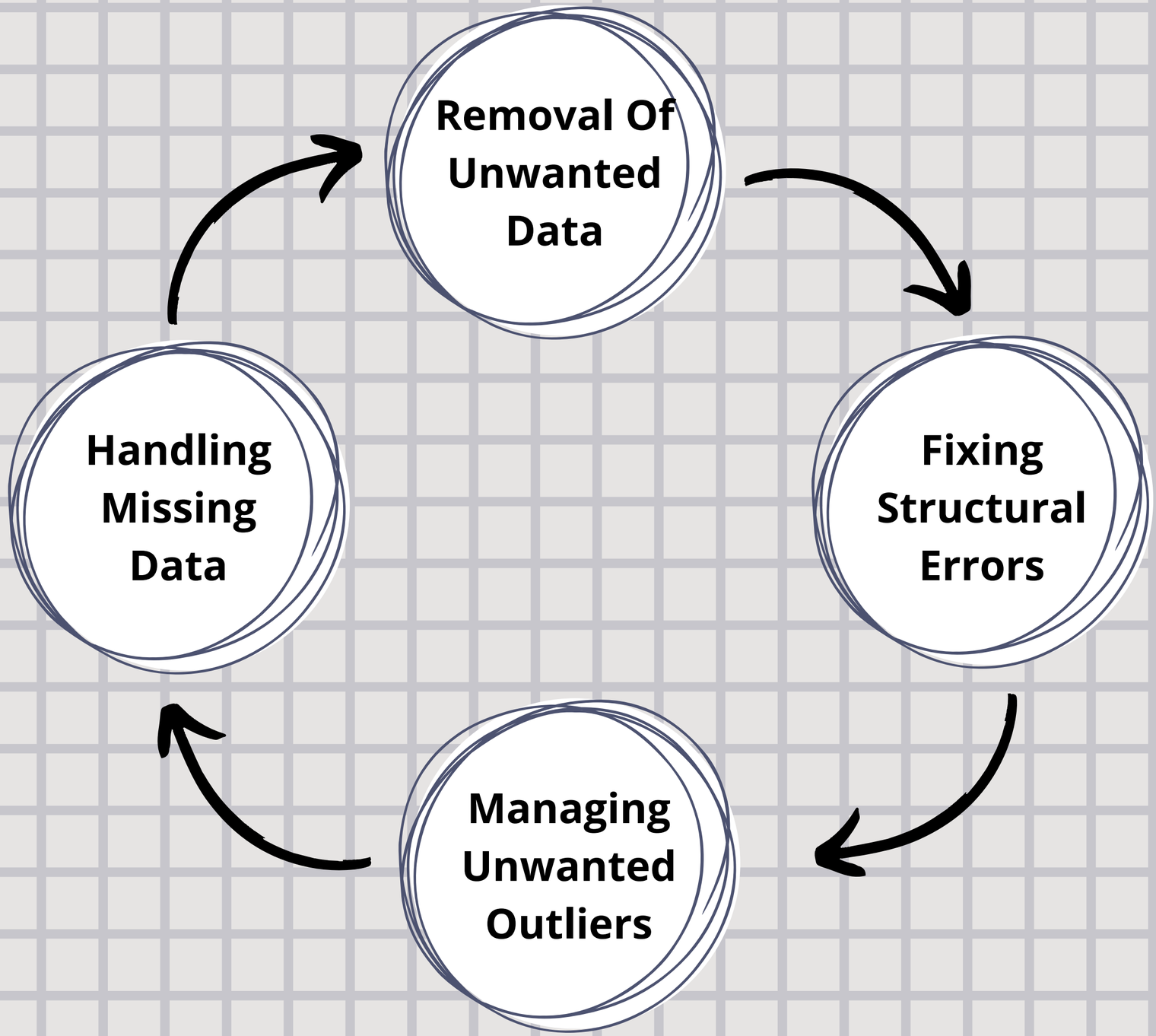
Data Cleaning: Identifying and removing any missing duplicate or irrelevant data

Data Integration: Involves combining data from multiple sources into a single, coherent dataset


Removal Of Unwanted Data


Removal Of Unwanted Data
Fixing Structural Errors


Removal Of Unwanted Data
Fixing Structural Errors
Managing Unwanted Outliers


Removal Of Unwanted Data
Fixing Structural Errors
Managing Unwanted Outliers
Handling Missing Data

Importance of Data Cleaning and Integration

Data cleaning and integration are crucial for ensuring accurate, reliable, and usable data, which is essential for informed decision-making, improved operational efficiency, and compliance with regulations

Reduces Operational Costs:
By preventing errors and inconsistencies, data cleaning can reduce costs associated with data processing, storage, and analysis.

Importance of Data Cleaning and Integration

Data cleaning and integration are crucial for ensuring accurate, reliable, and usable data, which is essential for informed decision-making, improved operational efficiency, and compliance with regulations

Reduces Operational Costs:
By preventing errors and inconsistencies, data cleaning can reduce costs associated with data processing, storage, and analysis.

Improves Data Quality:
By removing duplicates, handling missing values, and standardizing formats, data cleaning enhances the overall quality of the data, making it more usable for analysis and decision-making.

Importance of Data Cleaning and Integration

Data cleaning and integration are crucial for ensuring accurate, reliable, and usable data, which is essential for informed decision-making, improved operational efficiency, and compliance with regulations

Reduces Operational Costs:
By preventing errors and inconsistencies, data cleaning can reduce costs associated with data processing, storage, and analysis.

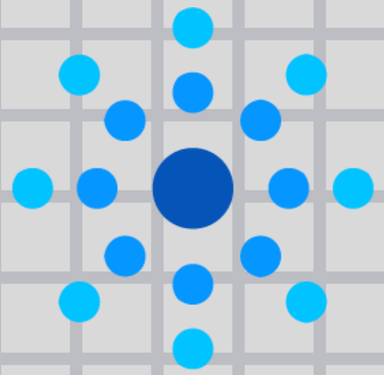
Improves Data Quality:
By removing duplicates, handling missing values, and standardizing formats, data cleaning enhances the overall quality of the data, making it more usable for analysis and decision-making.

Better Analysis of Data:
By integrating data from various sources, organizations can gain a more comprehensive understanding of their business operations and make more informed decisions.

Product Overview

HoloClean

alteryx



Tamr

HoloClean

- A semi-automated data repairing model for cleaning and enriching data
- Creates a probabilistic model based on: quality checks, references, correlations, etc.
- Scales for large size datasets
- Focuses on structured datasets

- Cleans the data with a probabilistic model, functional dependencies, and external info
- Utilizes PostgreSQL
- Average accuracy of about 90% upon release
- Accuracy achieved by combining methods mentioned

alteryx

- A cloud-based data analytics application with data cleaning and integration abilities
- Can integrate with AWS, Tableau
- Data-cleansing tool is a minor component
- Remove null rows/columns, replace nulls, modify case, etc.

- Uses generative AI models for data integration
- A user-friendly drag-and-drop UI for integration
- Utilizes data replication, data virtualization, streaming data
- Data enriched with external data sources and APIs

Tamr

- AI-based MDM software used by companies like Western Union and Santander
- Utilizes AI to clean/integrate data as well as provide dashboards for customers
- Spot incorrect data, validate data, and simplifies the data stack

- Standardize, match internal and external data, add data attributes
- Relies on Google Cloud Vertex AI and OpenAI, multiple third-party data providers
- Combines AI and external data sources for entity resolution
- AI allows for great scalability

	HoloClean	alteryx	Tamr
Data Cleaning	✓	✓	✓
Data Integration		✓	✓
Utilizes AI	✓	✓	✓
Cloud-based		✓	
Probabilistic Inference	✓		

Technical Details

HoloClean

HoloClean is a framework for data cleaning that uses probabilistic inference and combination of several signals in order to clean large volumes of data with accuracy.

Framework:

1. Error Detection Module
2. Compilation Module
3. Repair Module

Architecture:
Python, DeepDive, PostgreSQL

Typical Use Cases:
Dealing with large volumes of data in a variety of contexts (academic, industrial, etc.). Particularly important with user-entered data. The product is marketed towards data practitioners and scientists.

Key Technical Differentiator:
Probabilistic inference is used for both error detection and data repairing. This involves combining signals including integrity constraints, outlier identification, external info, and quantitative statistics into a unified framework.

Technical Details

HoloClean

Address	City	State	Zip
3465 S Morgan ST	*Chicago*	IL	*60608*
3465 S Morgan ST	Chicago	IL	*60609*
3465 S Morgan ST	Chicago	IL	*60609*
3465 S Morgan ST	*Cicago*	IL	60608

Each cell is a random variable

Constraints introduce correlations

c3: City, State, Address → Zip

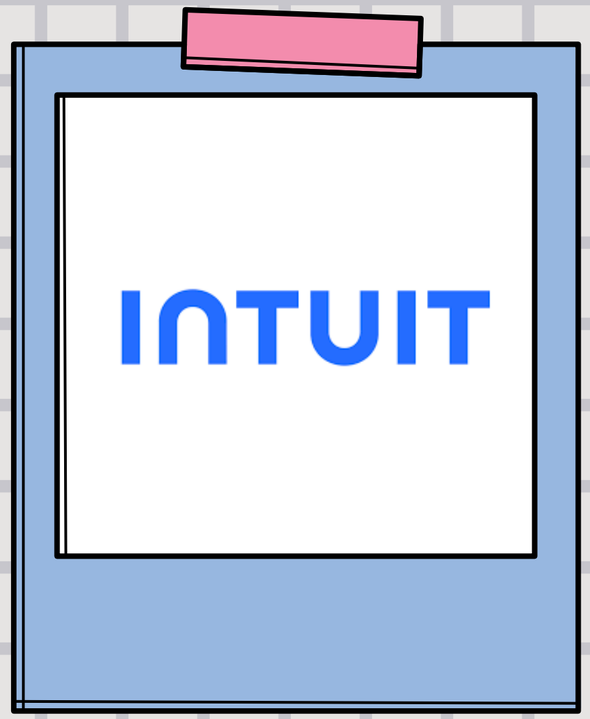
External data introduce evidence

Ext_Address	Ext_City	Ext_State	Ext_Zip
3465 S Morgan ST	Chicago	IL	60608

Image sourced from: http://www.holoclean.io/blog/holoclean.html

Sample Application #1

Case Study



Integration Challenges:

- The acquired platforms operate on different data systems, making integration complex. The main challenge was harmonizing disparate data systems to create a seamless user experience.

Solution – Interoperable Data Systems:

- Instead of merging all data into a single repository, Intuit built interoperable data pipelines. Credit Karma operates on Google Cloud, while Intuit's other services use AWS. Intuit designed data-sharing pipelines to connect these platforms across different cloud environments.

Impact of This Approach:

- Enables seamless data sharing across platforms with user consent.
- Provides personalized financial solutions without requiring a complete infrastructure overhaul.

https://logos-world.net/intuit-logo/

Sample Application #2

Case Study

Integration Challenges:

- Managing such vast amounts of data from diverse sources necessitates harmonizing disparate data systems to create a cohesive user experience and streamline operations.

Solution – Big Data Analytics:

- Walmart employs advanced data integration and analytics tools to process and analyze the massive influx of data. This approach enables the company to optimize inventory management, personalize customer experiences, and enhance supply chain efficiency.

Impact of This Approach:

- Keeps track of enormous amounts of data in inventory or other business fields the company is required to keep track of

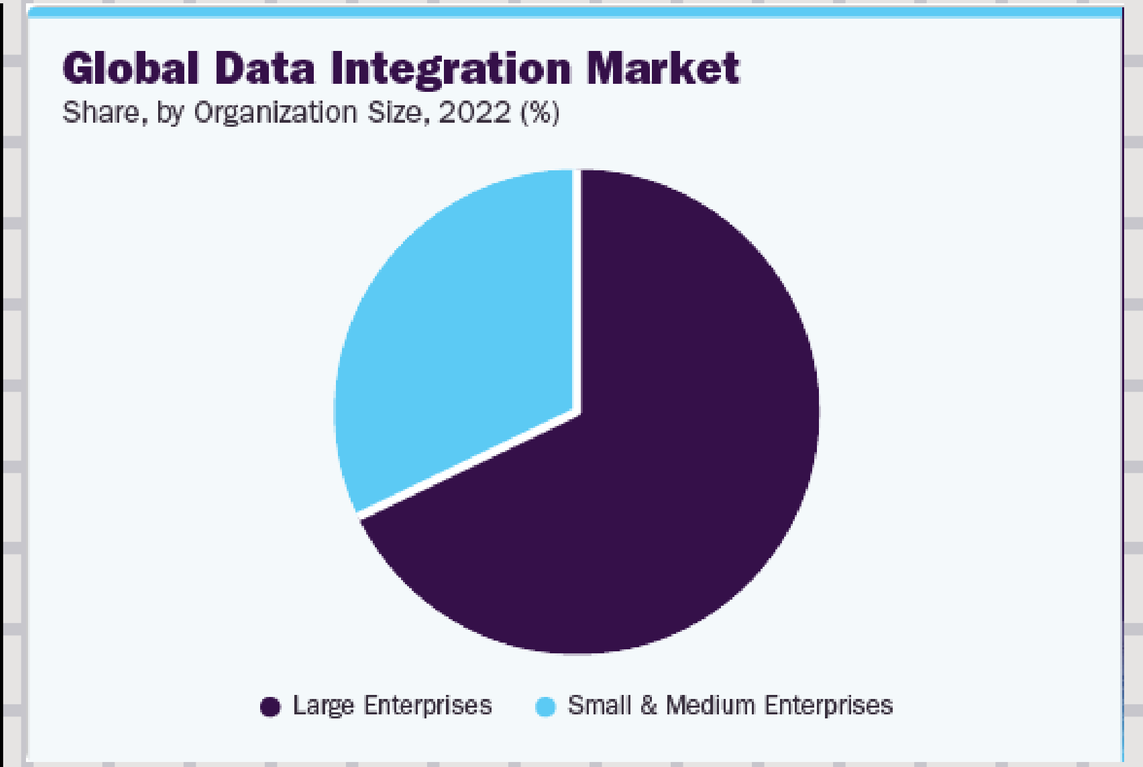
https://logos-world.net/walmart-unveils-new-logo-and-brand-identity/

Market Analysis

The global data market was estimated at $11.91 billion in 2022, and it's expected to reach $30.27 billion by 2030.

The biggest sectors that ultiilze cleaning and integration are IT & Telelcom, Government, Healthcare, Banking, Financial Services, and Retail & Commerce. The biggest sector was Marketing, with over 32% of global revenue. HR is the sector that is expected to grow significantly in the coming years.

This market is only growing due to the demand for using data to improve business operations.



https://www.grandviewresearch.com/industry-analysis/data-integration-market-report

Future Trends

AI and Machine Learning: Used to detect and predict future errors.

Cloud Computing: Allows faster and more efficient access across multiple data sources.

Real-time Data Integration: Moving data as it’s collected will make processing and analyzing data more efficient.

No-code and Low-code Integration: Cleaning and integrating data using less-technical methods results in more people being able to assist in these processes.

References

https://www.geeksforgeeks.org/data-cleansing-introduction/
https://logos-world.net/intuit-logo/
https://logos-world.net/walmart-unveils-new-logo-and-brand-identity/
https://www.actian.com/blog/data-integration/the-future-of-data-integration/#:~:text=AI%20and%20Machine%20Learning,intelligence%20and%20machine%20learning%20capabilities.
https://www.grandviewresearch.com/industry-analysis/data-integration-market-report

References

fhttps://www.vldb.org/pvldb/vol10/p1190-rekatsinas.pdf
https://github.com/HoloClean/holoclean?tab=readme-ov-file
http://www.holoclean.io/
http://www.holoclean.io/blog/holoclean.html
https://help.alteryx.com/current/en/designer/tools/preparation/data-cleansing-tool.html#idp340368
https://www.alteryx.com/about-us/trifacta-is-now-alteryx-designer-cloud
https://www.alteryx.com/products/alteryx-designer
https://www.tamr.com/data-enrichment, https://www.tamr.com/data-quality