DM,
Autumn, 2026

Understanding & Preparing Data

Faculty of DS & AI
Autumn semester, 2026

Trong-Nghia Nguyen

Business AI Lab

Content

- **Recap**
- Data Understanding
- Data Preparation: Data Scrubbing
- Ethics & Privacy

Recap

Let take some **Quizzes!**

Recap

Limitation of Data mining

- **Limit 1: Existence**

  Data mining only works on data that has actually been collected. It cannot analyze what is missing from the system.

- **Limit 2: Representation**

  Data can be deeply unrepresentative of the actual population or the target market.

- **Limit 3: Relevance**

  Data is often incorrectly collected, stored with formatting errors, or simply outdated over time.

Content

- Recap
- **Data Understanding**
- Data Preparation: Data Scrubbing
- Ethics & Privacy

Data Understanding

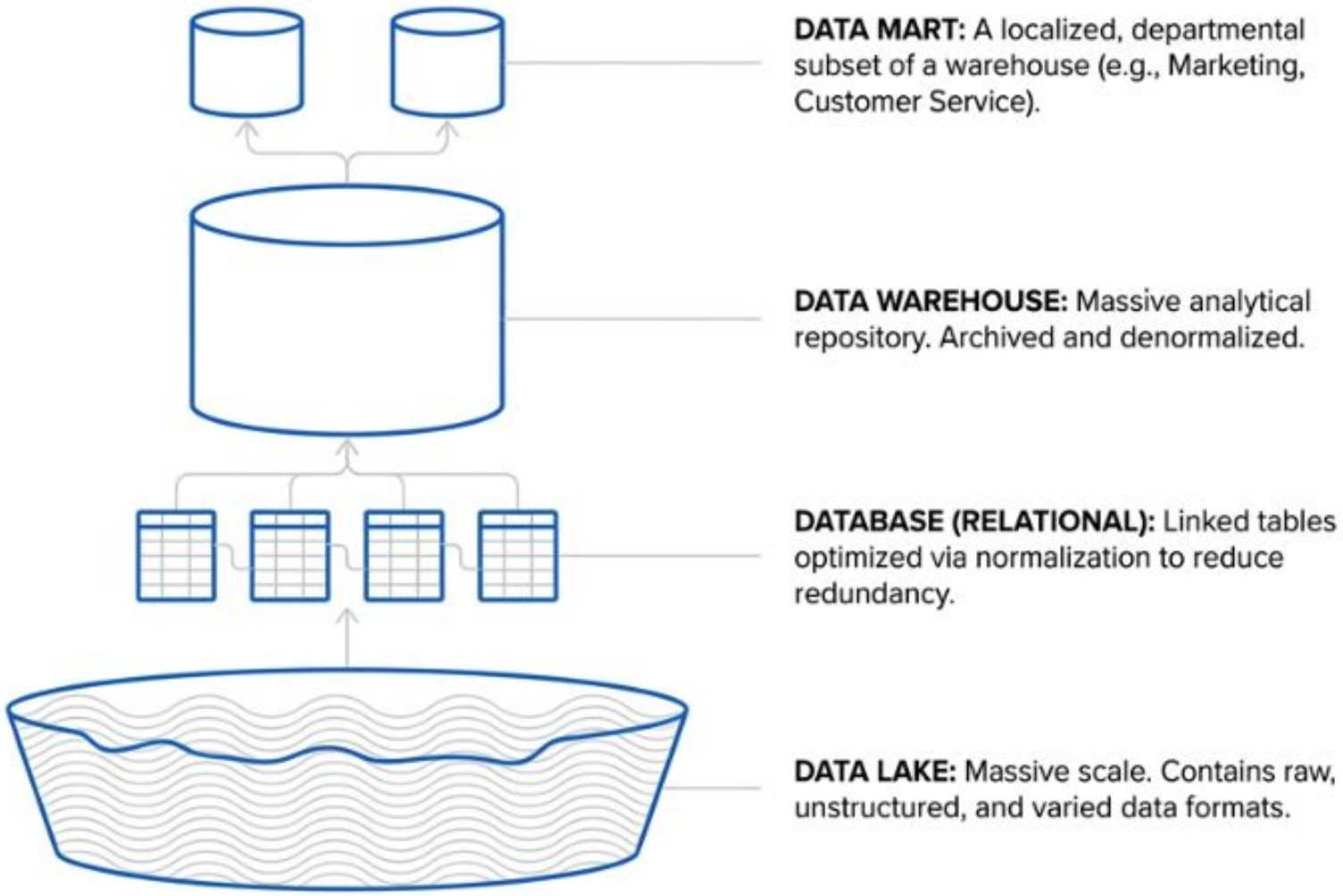
Standardize row and column terminology

- Data is always organized into **rows** and **columns**, but different environments use different terminology. We must standardize these terms to avoid confusion.

Environment / Tool	Row	Column
Database (Traditional RDBMS)	Tuples / Records	Fields
Data Warehouse	Observations / Cases	Variables / Attributes
RapidMiner (Original Textbook Tool)	Examples	Attributes
Orange (Course Practice Tool)	Instances / Rows	Features / Columns
North Textbook (Our Standard)	**Observations**	**Attributes**

Data Understanding

Data scales



Data Understanding

Data scales

Database

- **Structure:** Stores data in highly organized relational tables.
- **Normalization:** Data is split into multiple tables to minimize redundancy and save space.
- **System Type: OLTP** (Online Transaction Processing).
- **Optimization:** Built for fast read/write of daily transactions (e.g., barcode scans). Joining tables for analysis is very slow.

Data Warehouse

- **Structure:** Massive collection of archived data gathered from multiple operational sources.
- **Denormalization:** Tables are merged. Duplication is accepted to reduce the need for JOIN operations.
- **System Type: OLAP** (Online Analytical Processing).
- **Optimization:** Built specifically for extremely fast querying and analysis of massive volumes of data.

Data Understanding

Data scales

Data Mart

A smaller, focused data repository extracted for a **specific department** (e.g., Marketing, HR). Acts as a convenient "one-stop shop" for departmental staff.

Data Lake

A massive reservoir containing **raw, unstructured, and structured** data. Highly flexible but requires extensive processing and cleaning before extraction.

"Out-of-synch" Risk

If the operational Database changes but the Warehouse/Mart isn't updated in time, the analytical data is skewed. Mining out-of-sync data leads to flawed business decisions.

Data Understanding

Data set

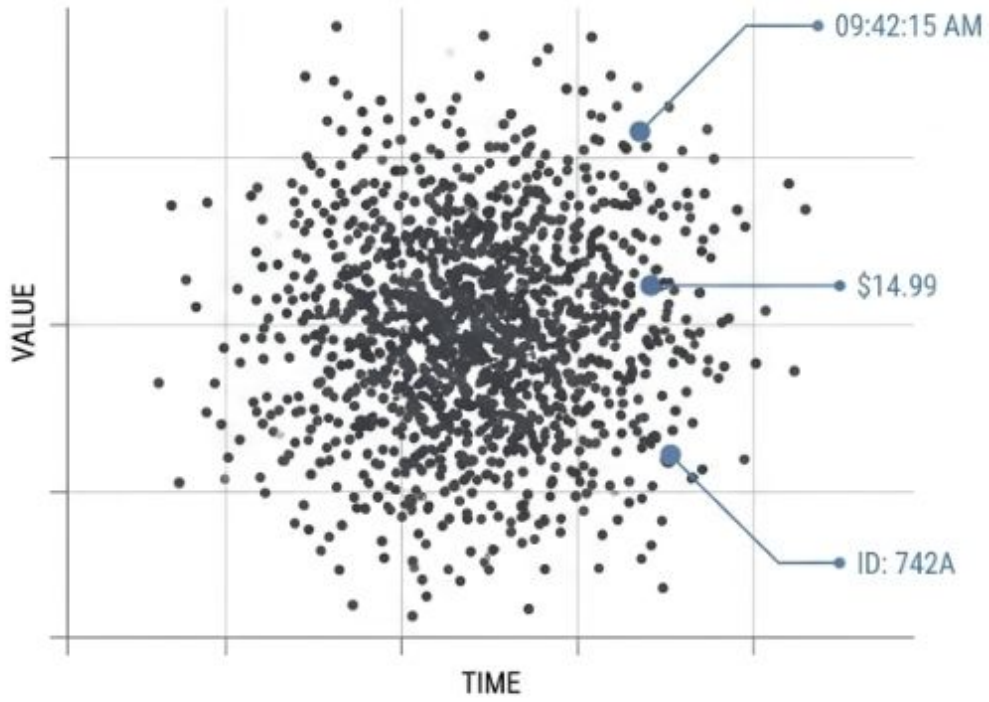
- **Definition:** A subset of data extracted from a Database or Data Warehouse.
- **Format:** Usually "denormalized" into a **single flat table**. Complex formats are simplified (e.g., standardizing datetime strings).
- **Role:** This is the **final form** of data that is actually loaded into algorithms and mining tools (like Orange or Python).
- **The Golden Rule:** Data Mining yields the best performance when run on Data Sets extracted from **OLAP (Warehouses)**. Never run intensive mining algorithms directly on an active OLTP database.

Data Understanding

Data source type

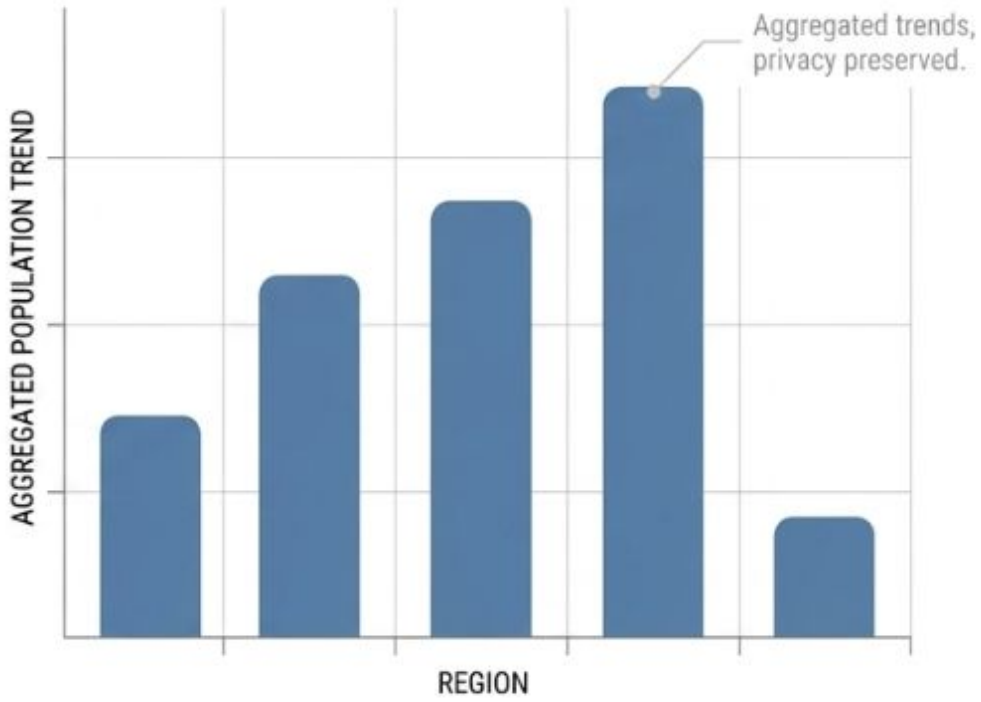
OPERATIONAL DATA

Raw, highly granular daily transactions (e.g., flight check-ins, barcode scans).



ORGANIZATIONAL DATA

Summarized and aggregated datasets (e.g., government census, CDC statistics). Inherently protects privacy by obscuring granular details.



Data Understanding

Data source type

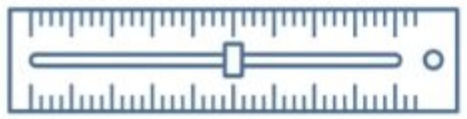
- **Operational data:** the most raw form of data, captured by transaction systems to record **daily activities—**such as purchasing fuel, online shopping, or airport check-ins. It includes transaction-level details (timestamp, price, and product description).

- **Organizational data: aggregated** data typically generated and published by government agencies, academic institutions, or non-profit organizations. For example, the U.S. CDC compiles nationwide influenza data annually. It offers better privacy protection due to its aggregated nature.

- **A note on secondary data:** since it is **derived** from primary sources, it may lack descriptive documentation and vary in rigor. **Always seek permission** before using it—data obtained from the Internet is **not** automatically in the public domain, and internal organizational data is **not** automatically free to use.

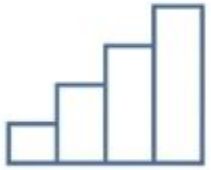
Data Understanding

Data types and variable roles



Numeric (Quantitative): Computable numbers (e.g., Age, Income).

Categorical / Nominal: Unordered labels (e.g., Gender, City). Requires mathematical encoding for models.



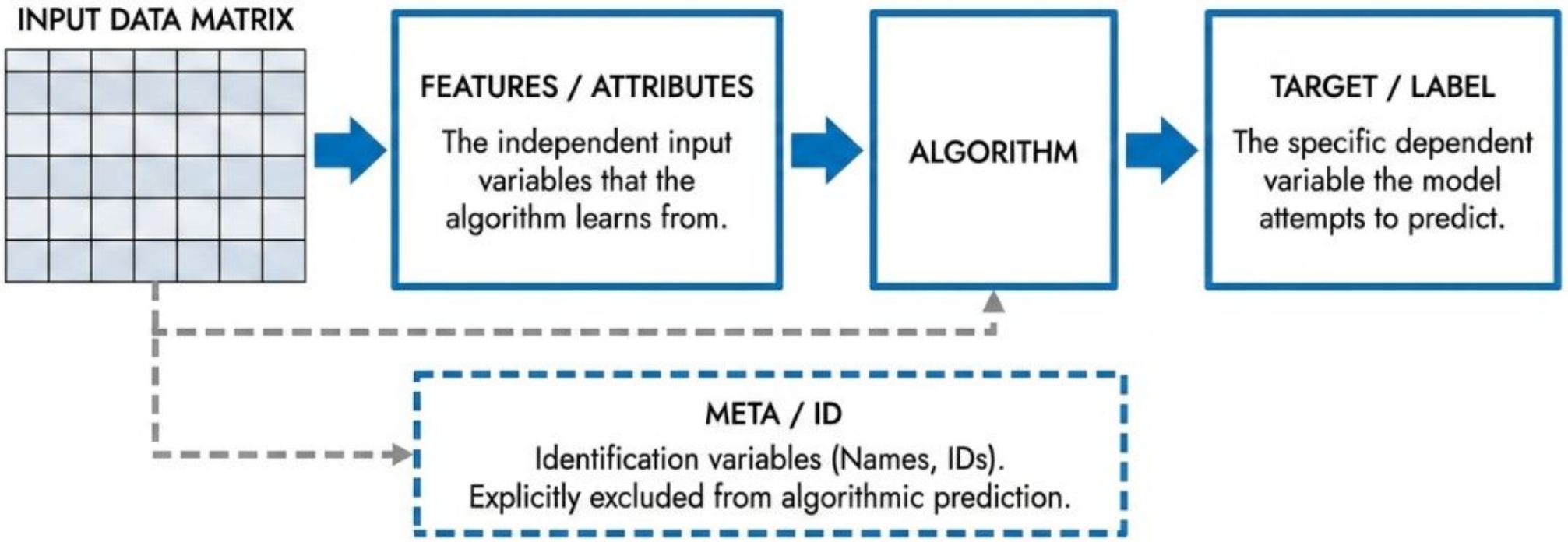
Ordinal: Categorical labels with an inherent logical order (e.g., Low, Medium, High).

Datetime: Temporal markers. Can be split into granular units (e.g., Day/Month) to isolate seasonal patterns.

Data Understanding

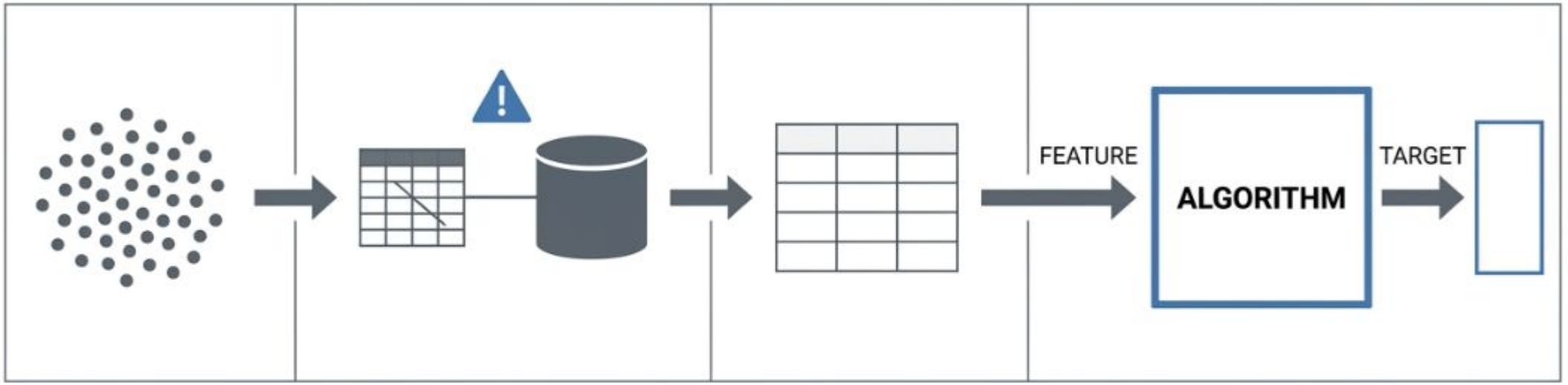
Data types and variable roles

- Variables must be assigned specific algorithmic roles before model training: **ID**, **features,** and **target/label.**
- Never discard variables during the initial data load.



Data Understanding

Lifecycle of Data Understanding



Phase 1: Operational Source Data

Phase 2: Stored in OLTP DB -> Extracted to OLAP Warehouse (Risk: Sync Lag)

Phase 3: Exported as a targeted Data Set (Rows/Cols)

Phase 4: Variables assigned Types (Numeric/Nominal) and ML Roles (Feature/Target)

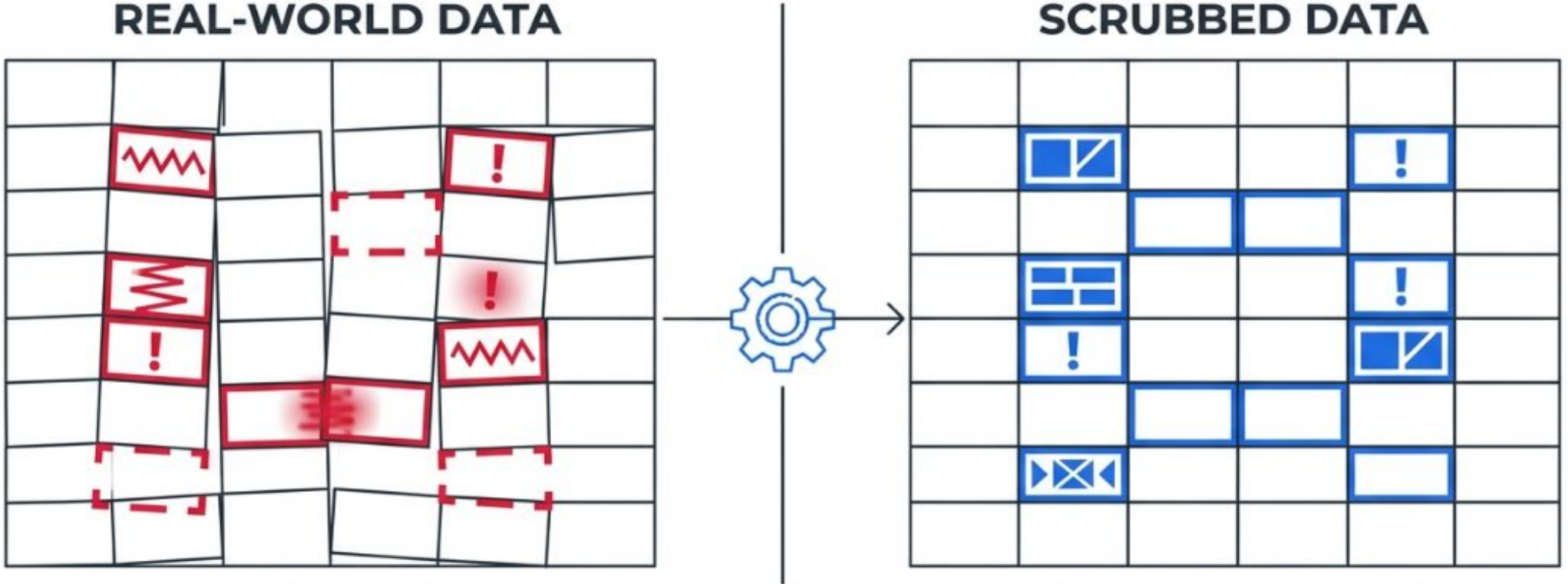
Content

- Recap
- Data Understanding
- **Data Preparation: Data Scrubbing**
- Ethics & Privacy

Data Preparation

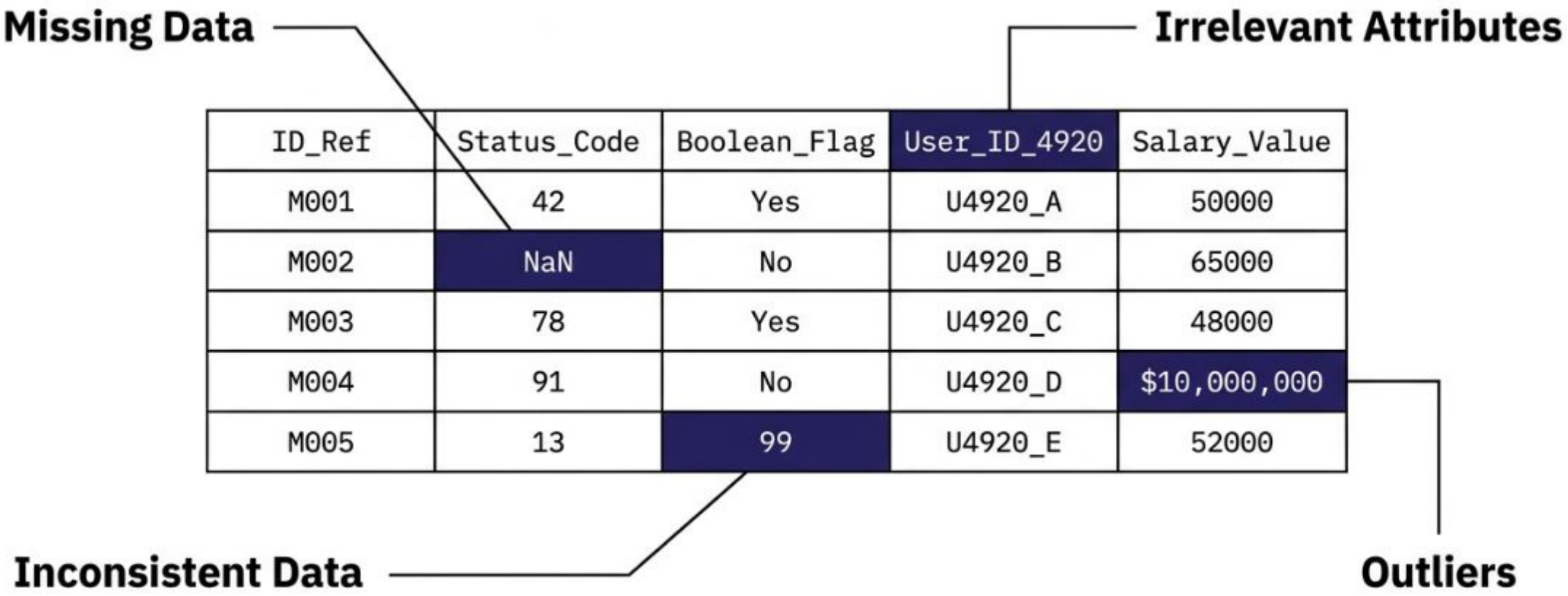
Data Preparation

Data Scrubbing is the systematic identification and resolution of these anomalies to ensure the integrity of the underlying analysis.



Data Preparation

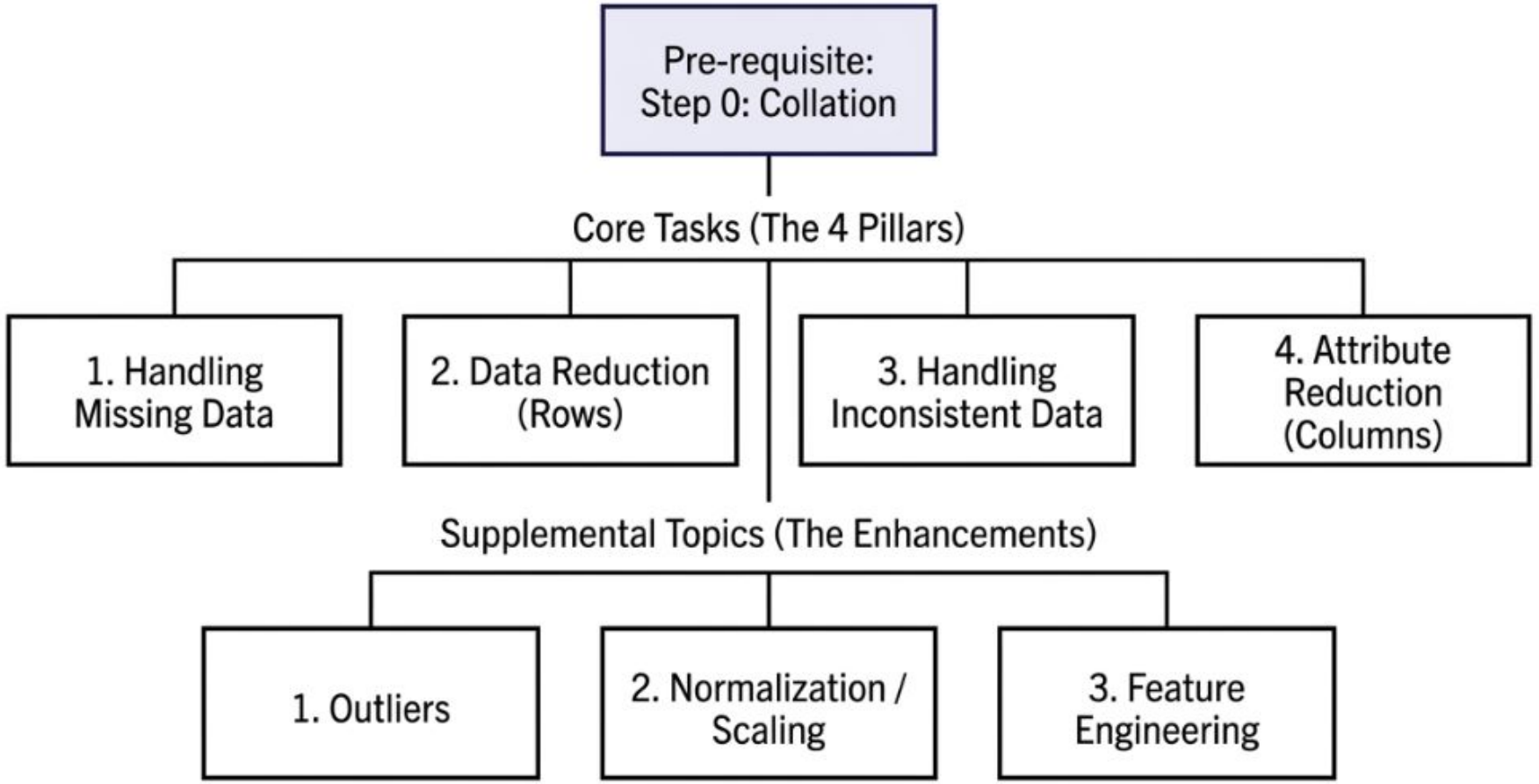
Data Preparation



ID_Ref	Status_Code	Boolean_Flag	User_ID_4920	Salary_Value
M001	42	Yes	U4920_A	50000
M002	NaN	No	U4920_B	65000
M003	78	Yes	U4920_C	48000
M004	91	No	U4920_D	$10,000,000
M005	13	99	U4920_E	52000

Data Preparation

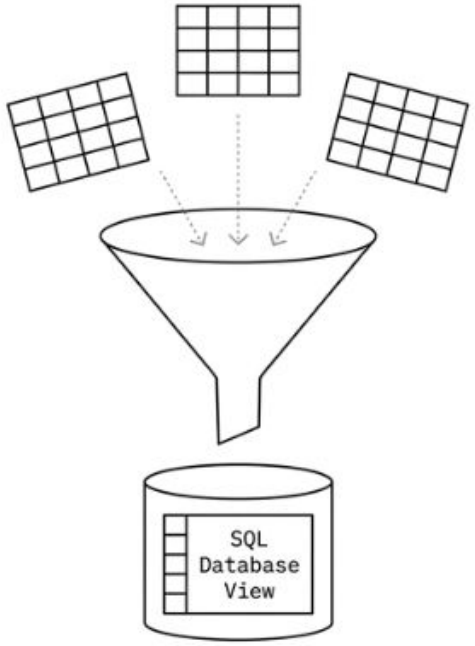
Data Preparation



Data Preparation

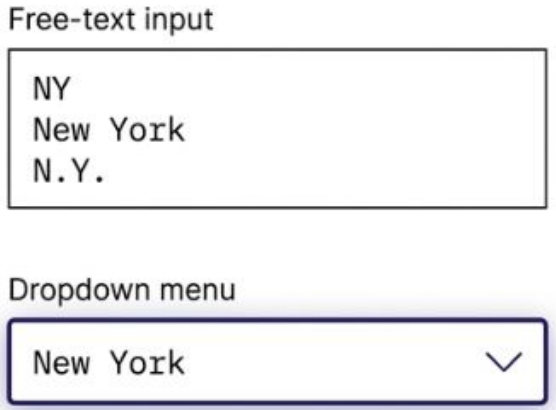
0. Collation

Establishing the Foundation Before Scrubbing Begins



Collation

Gather scattered data into a single, rich, and consistent SQL Database View to establish a single source of truth.

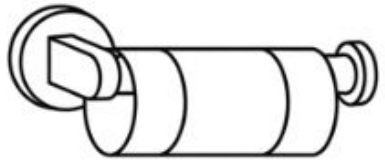


Prevention at Source

Replace free-text fields with dropdown menus during data collection to eliminate spelling errors and structural inconsistencies.

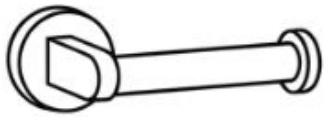
Data Preparation

1. Handling missing data



0

A known value of zero or empty.



NaN

A state of non-existence or unknown.

The Nature of Missing Data

Data represented as null, ?, None, or np.nan indicates a state of non-existence or an unknown value. It is mathematically distinct from the number 0.

The Ethical Imperative

Never force a value where one does not exist. For example, never default a missing Criminal Record to Yes simply because it happens to be the statistical Mode. Subjective variables demand safe handling over blind statistical imputation.

Data Preparation

1. Handling missing data

Strategy	Application	Risk / Warning
Keep	When analytical goals permit missing values (e.g., certain tree-based algorithms).	May cause distance-based algorithms to fail.
Impute (Numerical)	Replace missing quantitative values with the Mean or Median.	Artificial reduction of variance.
Impute (Categorical)	Replace missing qualitative values with the Mode.	High Ethical Risk. Proceed with extreme caution to avoid distorting subjective realities.
Remove (Filter)	Delete rows entirely. Safest for highly subjective variables.	Loss of statistical power if the dataset is small.

Data Preparation

2. Data reduction

Reducing Observations to Isolate Signal and Accelerate Speed

1	~~R1C1: NaN~~	~~R1C2: NaN~~	~~R1C3: NaN~~	~~R1C4: NaN~~	~~R1C5: NaN~~
2	12.45	67.89	45.10	90.21	33.56
3	67.89	45.10	90.21	90.21	33.56
4	~~R1C1: NaN~~	~~R1C2: NaN~~	~~R1C3: NaN~~	~~R1C4: NaN~~	~~R1C5: NaN~~
5	12.45	67.89	45.10	90.21	33.56
6	67.89	45.10	90.21	90.21	33.56
7	~~R1C1: NaN~~	~~R1C2: NaN~~	~~R1C3: NaN~~	~~R1C7: NaN~~	~~R1C7: NaN~~
8	12.45	67.89	45.10	90.21	33.56

Filtering

Systematically purging rows containing missing or unusable data from the analytical pipeline to ensure model quality.

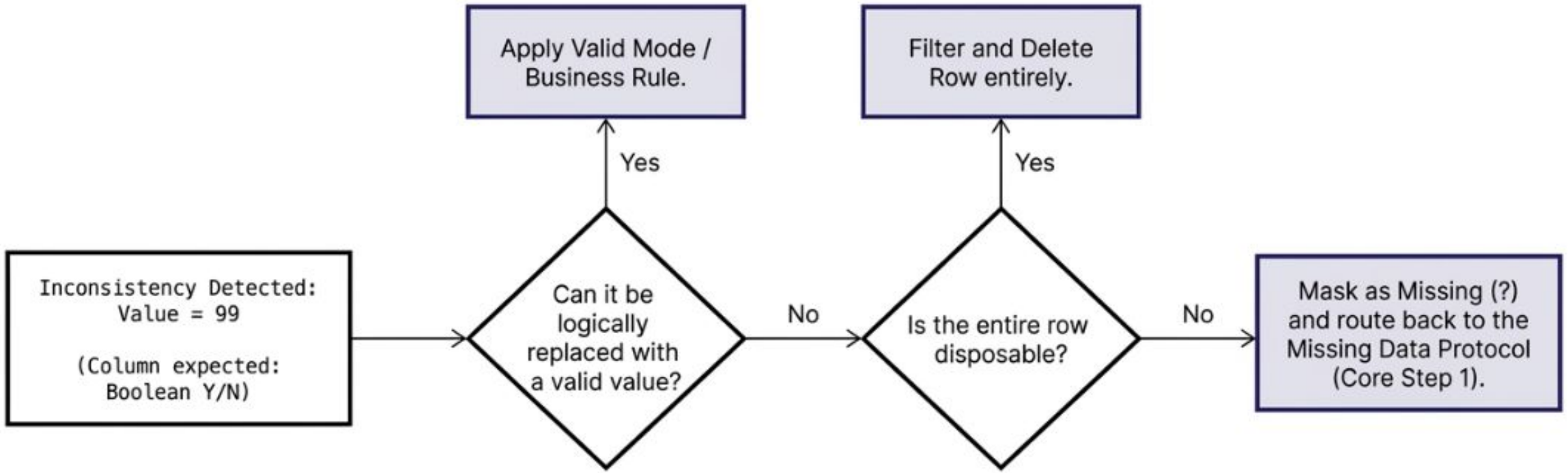
1	R1C1: 23.56	R1C1: 23.56	R1C1: 23.56	R1C1: 23.56	R1C1: 23.56	R1C1: 23.56
2	R2C1: 78.12	R2C2: 78.12	R2C1: 78.12	R2C2: 78.12	R2C2: 78.12	R2C2: 78.12
3	R3C3: 45.10	R3C2: 45.10	R3C3: 45.10	R3C3: 95.10	R3C3: 45.10	R3C3: 45.10
4	R4C1: 45.10	R4C2: 45.10	R4C3: 45.89	R4C4: 45.10	R4C4: 90.21	R4C4: 45.10
5	R5C5: 33.56	R5C5: 33.21	R5C5: 90.21	R5C4: 33.56	R5C5: 90.21	R5C5: 33.56
6	R6C1: 33.56	R5C6: 25.62	R6C6: 24.57	R6C6: 33.21	R6C6: 13.56	R6C6: 25.69
7	R7C7: 71.13	R7C7: 78.12	R7C7: 90.23	R7C7: 84.61	R7C7: 77.28	R7C7: 90.21
8	R8C1: 23.56	R8C8: 45.10	R8C8: 18.33	R8C8: 33.17	R8C8: 50.29	R8C8: 33.56
9	R9C1: 78.12	R9C1: 90.21	R9C9: 83.26	R9C9: 45.56	R9C9: 53.39	R9C9: 81.53
10	R1C1: 23.56	R108: 62.89	R1C0: 45.55	R100: 93.35	R1C0: 69.10	R1C0: 45.10
11	R1C1: 25.56	R1C1: 78.12	R1C1: 78.12	R1C1: 31.58	R1C1: 79.12	R1C1: 45.93
12	R1C1: 82.13	R1C2: 23.56	R1C2: 12.69	R1C2: 33.36	R1C2: 23.83	R1C2: 78.13
13	R1C1: 56.86	R1C2: 80.15	R1C3: 52.87	R1C3: 21.33	R1C3: 16.18	R1C3: 33.56
14	R1C1: 78.12	R1C4: 59.94	R1C4: 78.11	R1C4: 22.56	R1C4: 25.74	R1C4: 95.12
15	R1C1: 23.56	R1C2: 13.63	R1C5: 67.53	R1C5: 41.24	R1C5: 18.32	R1C5: 23.56

Sampling

Extracting a random subset (e.g., 50%) of a massive dataset to enable rapid model testing and iteration before deploying on the full database.

Data Preparation

3. Handling inconsistent data



Data Preparation

4. Attribute reduction

Pruning Attributes for Algorithmic Clarity

Source Database

User_ID	Age	Gender	Income
U_1001	34	M	75000
U_1002	29	F	62000
U_1003	41	M	91000
U_1004	35	F	68000
U_1005	48	M	112000

Analytical Workflow

~~User_ID~~	Age	Gender	Income
~~U_1001~~	34	M	75000
~~U_1002~~	29	F	62000
~~U_1003~~	41	M	91000
~~U_1004~~	35	F	68000
~~U_1005~~	48	M	112000

The Purpose

Remove variables irrelevant to the specific business question (e.g., dropping ID tags) to reduce algorithmic noise.

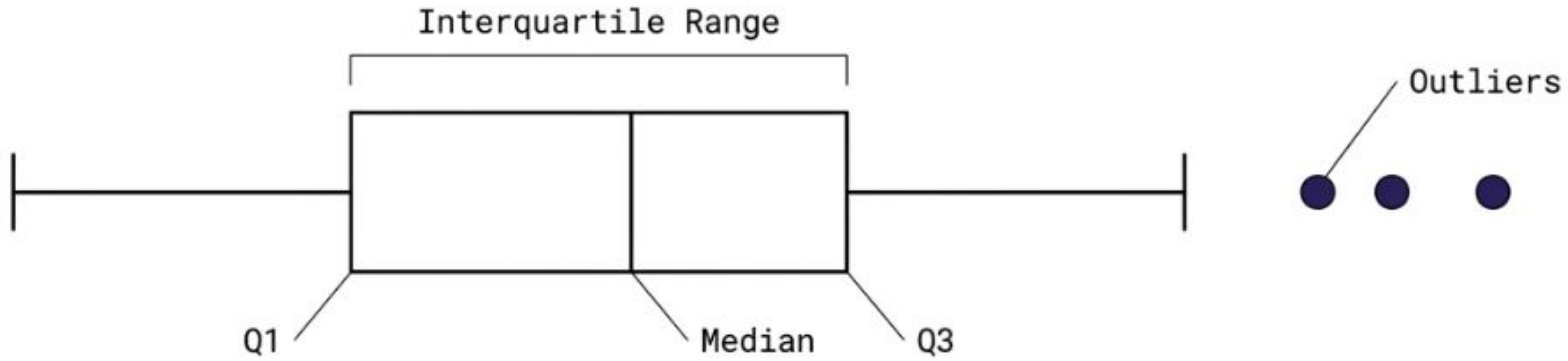
The Golden Rule

Only drop columns within the active software workflow. Never delete them from the foundational source database—today's noise may be tomorrow's critical signal.

Data Preparation

Supp. 1. Outliers

Detecting and Evaluating Extreme Observations



Definition

Data points that are strictly accurate but radically deviate from the norm (e.g., a single customer spending 100x the average amount).

Detection Mechanism

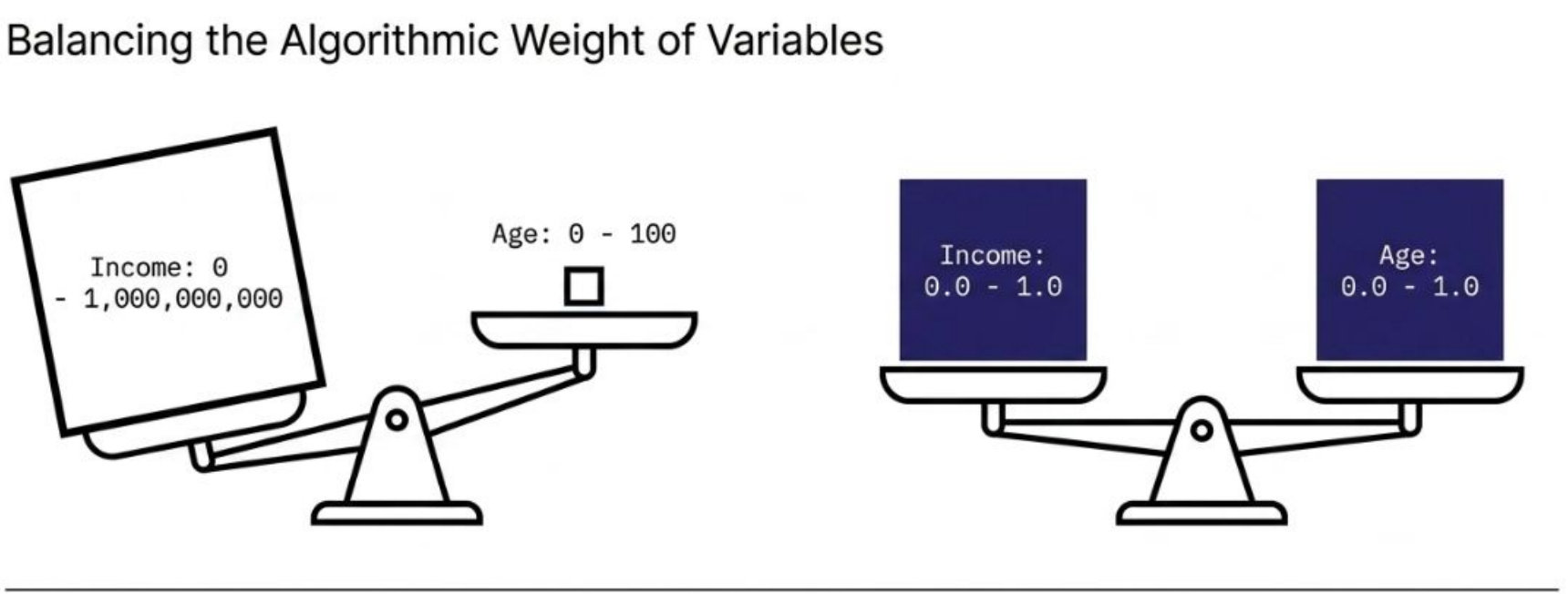
The Box Plot is the primary visual diagnostic tool for isolating extreme values beyond standard statistical variance.

Resolution Strategy

Retain, remove, or transform the outlier based strictly on the business context and the objective of the analysis.

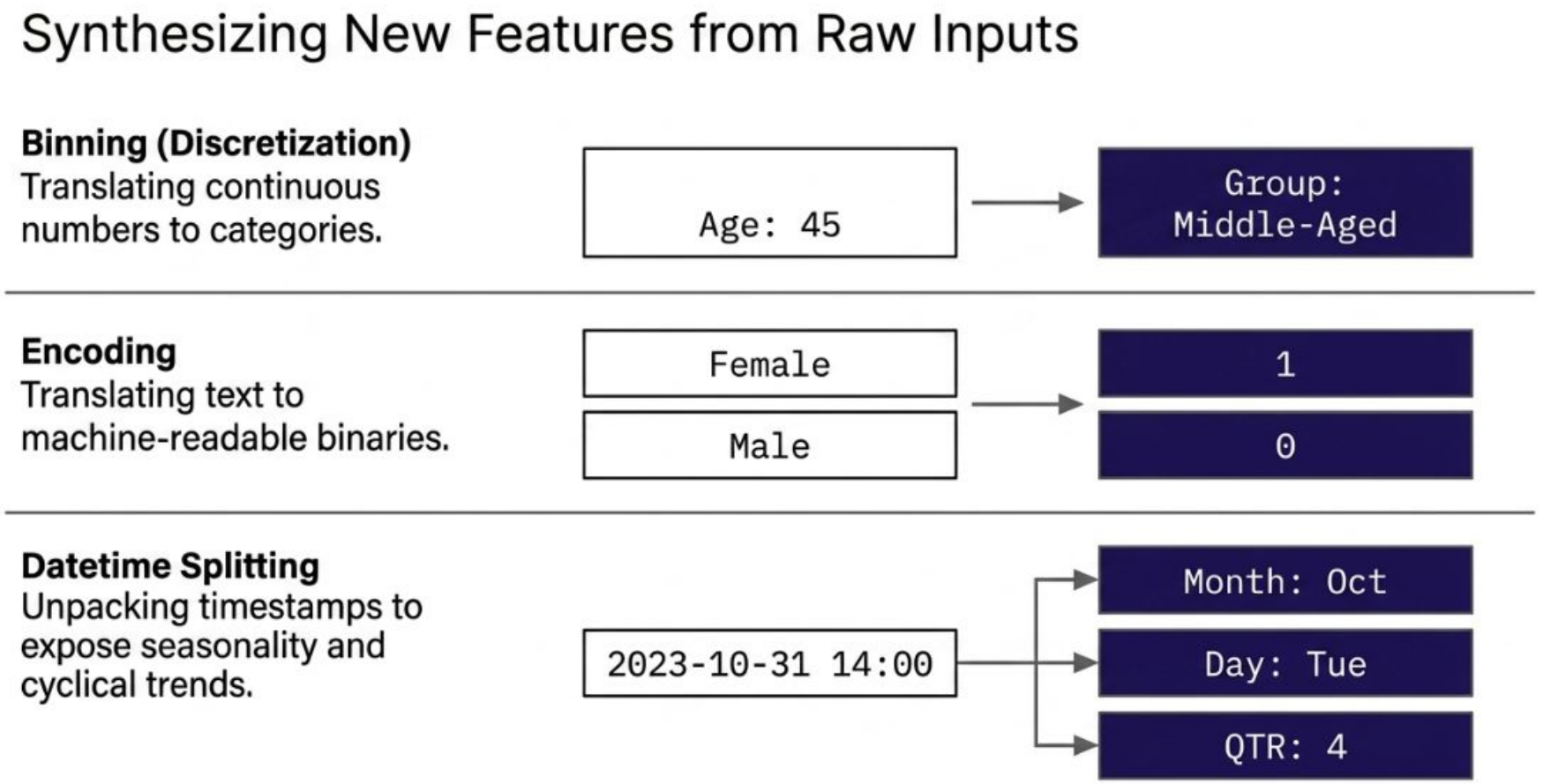
Data Preparation

Supp. 2. Normalize/Scaling



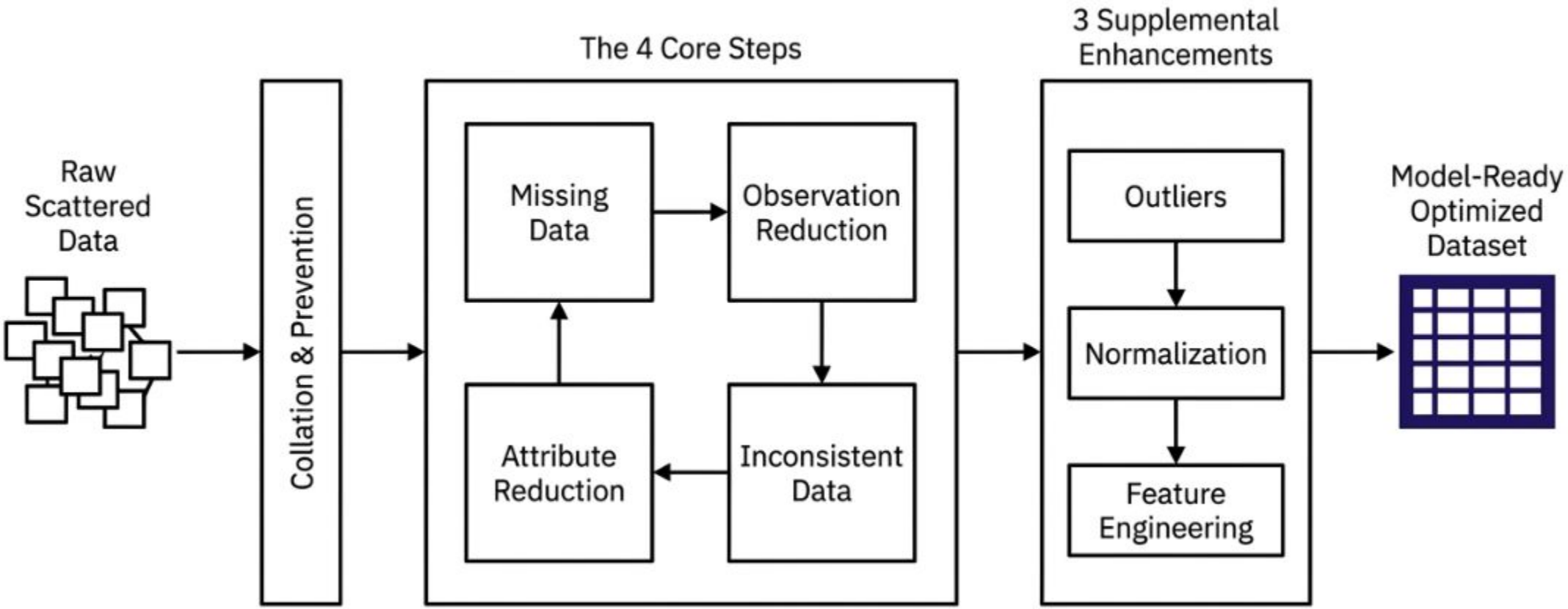
Data Preparation

Supp. 3. Features Engine



Data Preparation

Complete protocol



Content

- Recap
- Data Understanding
- Data Preparation: Data Scrubbing
- **Ethics & Privacy**

Ethics & Privacy

Ethics & Privacy

In data mining, continuous exposure to numbers, algorithms, and software obscures a fundamental truth: **behind every number is a real person.**

The Analyst's Mandate: Ethics and privacy must remain the primary filter at every stage of abstraction.

Ethics & Privacy

Ethics & Privacy

- **JetBlue** provided over a million passenger records to the contractor **Torch Concepts**.
- **Torch** supplements its data with information—such as household size and Social Security numbers—purchased from data broker Acxiom, aiming to build "*potential terrorist profiles*."
- All of this was done **without notifying or seeking permission** from the passengers.
- When the scandal broke, a wave of privacy lawsuits was filed, and senators called for an investigation.

Entity	JetBlue Airways
Volume	**1,000,000** passenger data records.
Recipient	Torch Concepts (Government Contractor)
Fallout	Widespread privacy lawsuits and United States Senate investigations.

Ethics & Privacy

- Data miners have an **ethical obligation** to protect privacy and guard against risks such as identity theft.
- Just because a government agency or contractor requests data does **not** mean it must be provided.
- Consider both technical obligations (information security) and ethical obligations.
- Business objectives does not justify unethical means.

Dimension	Technical Security (What we can do)	Ethical Boundary (What we should do)
Goal	Protecting the system from unauthorized access.	Protecting the user from unauthorized exploitation.
Constraint	Can we securely encrypt and transfer this data?	Did we obtain informed consent to utilize this data?
Example	Implementing robust data storage protocols.	Navigating the supermarket loyalty card coercion dilemma.

Thank you!