Ethics of Data Science

Chapter 10 – Data Provenance , Aggregation
Ethics of Data Scraping and Storage

Dr. Dhiran Kumar Mahto
Assistant Professor
Department of Computer Science and Engineering
Indian Institute of Information Technology (IIIT) Ranchi
Email- dhiran.mahto@iiitranchi.ac.in

Data Provenance (Origination)

1. Data Provenance (Origination)

Definition

Data provenance, also referred to as **origination**, is the complete record of the **history, source, lineage, and context** of data. It documents:

- **Where the data came from** (source)
- **How it was collected** (method/tools)
- **Who collected it** (institution or researcher)
- **When it was collected** (time period)
- **Under what conditions and assumptions** (context, environment)
- **Any transformations** applied afterward (cleaning, aggregation, annotation, filtering)

It is similar to "traceability" in supply chains—just as we track where food or materials come from, data provenance tracks the journey of a dataset.

Why Data Provenance Matters

Provenance is the cornerstone of **reliable, ethical, and responsible data science**. Without provenance:

You cannot assess **trustworthiness** or **credibility** of the dataset.
You cannot identify **sampling bias, errors, or limitations**.
You cannot check **informed consent** or **legal permissions**.
Replication becomes difficult because others cannot reconstruct how the data was obtained.
Decisions based on such data may cause **unfair or unethical outcomes**.

If we do not know where data came from, we cannot responsibly use it.

Data Provenance (Origination)

Key Elements of Data Provenance

1. Source Information

•Origin of data: sensors, surveys, social media, satellites, hospitals, etc.
•Institution or platform responsible for collection.
•Whether data is original or third-party.

2. Collection Method

•Manual entry, automatic sensors, web scraping, crowdsourcing, IoT, laboratory experiments.
•Tools and technologies used: cameras, drones, biometric scanners, etc.

3. Context of Collection

Context may include:
•**Location** (hospital vs. home, urban vs. rural).
•**Population characteristics** (age, gender distribution).
•**Temporal conditions** (season, emergency situations, special events).
•**Policies or constraints** (COVID-19 lockdown, disaster relief period).
Context is important because the same data collected under different contexts may represent *very different populations*.

4. Consent and Legal Framework

•Was the data collected ethically?
•Was consent informed, explicit, or implied?
•Are there copyright or data ownership issues?

Data Provenance (Origination)

5. Data Transformations Over Time

Provenance also tracks how data changes:

•Cleaning, removal of outliers
•Normalization, aggregation
•Labeling or annotation
•Model-based filtering
•Merging with other datasets

Understanding transformations is crucial to evaluate **accuracy** and detect **manipulation**.

Sampling Bias

Sampling bias occurs when the data collected **does not accurately represent the broader population** it is intended to model or generalize to.
This happens when **some groups are systematically included or excluded**, either intentionally or unintentionally.
It makes even "big data" unreliable if the sampling process itself is flawed.

How Sampling Bias Happens
Sampling bias can arise due to:
Limited access to certain communities
Incomplete data collection tools
Historical social inequalities
Convenience sampling (easy-to-reach participants)
Overreliance on online data
Lack of diversity in data sources (e.g., datasets from only one country)
Importantly:
Large datasets do NOT eliminate sampling bias. They only amplify it.

Sampling Bias

Detailed Types of Sampling Bias

1. Selection Bias

Occurs when the method of choosing participants favors certain groups.

Example:

A "national health survey" conducted mostly in **urban hospitals**.

Rural populations are under-represented.

Results generalize poorly.

Policies derived from this data may misallocate resources.

Impact:

Urban lifestyles dominate the dataset → AI predicts health risks inaccurately for rural citizens.

2. Non-Response Bias

Arises when individuals who choose not to participate are **systematically different** from those who do.

Example:

In an income survey:

Wealthy people avoid disclosing income.

Low-income workers may not participate due to lack of time.

Impact:

The dataset may capture mostly middle-class respondents → misleading economic predictions.

AI Impact:

If used for credit scoring, it may produce unfair loan approval patterns.

Sampling Bias

3. Survivorship Bias
Occurs when only those who "survived" a process are included in the data.
Example:
Studying startup success using only companies that are currently profitable.
Failed startups are ignored.
The analysis falsely assumes success is easy and common.
AI Impact:
Investment prediction models become overly optimistic.

4. Convenience Bias
Using easily available data rather than representative data.
Example:
Training a medical AI model using only data from one hospital because it's convenient.
This hospital may serve a specific demographic (e.g., elderly population).
Impact:
Model performs poorly in other hospitals.

5. Temporal Bias (Time-Based Sampling Bias)
Data collected during a specific season or event may not generalize.
Example:
Collecting shopping behavior during **festive season** only.
AI thinks people buy more luxury goods than usual.
Impact:
Retail forecasting becomes inaccurate.

Sampling Bias

3. Survivorship Bias
Occurs when only those who "survived" a process are included in the data.
Example:
Studying startup success using only companies that are currently profitable.
Failed startups are ignored.
The analysis falsely assumes success is easy and common.
AI Impact:
Investment prediction models become overly optimistic.

4. Convenience Bias
Using easily available data rather than representative data.
Example:
Training a medical AI model using only data from one hospital because it's convenient.
This hospital may serve a specific demographic (e.g., elderly population).
Impact:
Model performs poorly in other hospitals.

5. Temporal Bias (Time-Based Sampling Bias)
Data collected during a specific season or event may not generalize.
Example:
Collecting shopping behavior during **festive season** only.
AI thinks people buy more luxury goods than usual.
Impact:
Retail forecasting becomes inaccurate.

Sampling Bias

6. Geographic Bias

Diverse regions are not represented equally.

Example:

Training a weather model only on data from South India → poor predictions in North India.

7. Digital Divide Bias

People with limited internet access are excluded.

Example:

Online surveys often miss rural communities, elderly, or low-income groups.

Impact:

Models may favor digitally active populations.

Aggregation

Definition:
Aggregation is the process of combining individual data points into categories, groups, or higher-level summaries.
Explanation:
While aggregation simplifies data analysis, it can hide significant variations, leading to misunderstanding or unfair outcomes.
Problems with aggregation:
•Loss of nuance
•Masking inequalities
•Incorrect policy decisions
Example:
Aggregating crime data by city may hide neighborhood-level violence. A city may appear "safe," while specific localities face severe risk.

Why Aggregation Matters
Aggregation is useful for:
Reducing complexity
Making data interpretable
Identifying broad trends
Supporting policymakers
Summarizing large datasets
But the **simplification comes at a cost**.
Aggregated data can **hide differences**, **erase minority experiences**, and **distort conclusions**, especially in diverse populations.
If used without caution, aggregation becomes a source of **statistical deception** and **algorithmic unfairness.**

Data Retention

Definition

Data retention refers to the **duration, schedule, and policies** that determine **how long data is stored** by an organization before it is deleted or archived.

Retention policies specify:

What data is stored

How long it is stored

Why it is stored

Where and how it is stored

When and how it is deleted

Explanation

Ethical Principle of Retention

Ethical data practice requires that **data must be stored only for as long as it is needed**, strictly for the **original purpose** for which it was collected.

If the purpose expires, the data should be:

Deleted, or

Anonymized, or

Archived with safeguards

Keeping data longer than necessary increases risk and violates the principle of **purpose limitation**.

Data Retention

Why Excessive Retention is Harmful

Prolonged or unnecessary retention creates major ethical, social, and legal problems.

1. Privacy Breaches

The longer data is stored, the more opportunities exist for:

Cyberattacks

Insider misuse

Accidental exposure

Database leaks

Older systems are often poorly secured, making retained data vulnerable.

2. Unintended Analysis (Inference Risks)

With advancements in AI, **data stored today can be analyzed tomorrow in new ways**, revealing information never intended at the time of collection.

Example:

Old browsing logs could be used to infer mental health, political preferences, or personal habits.

3. Mission Creep

Mission creep occurs when data collected for **one purpose** is later used for **another purpose** without consent.

Examples:

Using academic records to predict employability

Using location data for advertising

Using medical data for insurance scoring

Mission creep violates **user autonomy** and **trust**.

Data Retention

4. Re-identification Risks

Even anonymized data becomes identifiable over time due to:
Additional datasets
Better algorithms
New linking techniques
Thus, long-term retention increases the chance of re-identification.

5. Legal Non-Compliance

Many data protection laws prohibit indefinite retention.
For example:
GDPR (Europe)
Requires *storage limitation* — keep data only as long as necessary.
India's DPDP Act (2023)
Organizations must delete personal data once the purpose is fulfilled.
HIPAA (Health Data – US)
Requires specific retention periods but mandates secure disposal.
Failure to follow retention schedules leads to fines and penalties.

Data Retention

Common Reasons Organizations Wrongly Retain Data

1. “Storage is cheap” mindset

Developers assume keeping everything forever is harmless.

2. Future value speculation

Organizations think old data might be useful later for:

AI training

Analytics

Marketing

Profit-making

3. Poor governance

No retention policy or unclear data ownership.

4. Lack of deletion systems

Deleting data safely is technically harder than collecting it.

Data Disposition

Definition

Data disposition refers to the **policies, methods, and processes** used to:

- **Archive**
- **Anonymize**
- **Encrypt**
- **Destroy**
- **Delete**

data after it has reached the **end of its intended lifecycle**.

Disposition ensures that data is handled **safely and ethically** when it is no longer needed.

Explanation

Data disposition is a critical part of **responsible data governance**.

Once data has fulfilled its purpose (e.g., after a project ends, a service is discontinued, or a legal retention period expires), organizations must:

Dispose of it securely

Prevent unauthorized use

Ensure the data cannot be reconstructed

Avoid mission creep

Reduce long-term cybersecurity exposure

Proper disposition protects:

Privacy

Confidentiality

Security

Organizational reputation

Data Disposition

Why Data Disposition Is Important

1. Prevents Data Leakage

Old or unused data is vulnerable to:
Hacks
Insider threats
Accidental exposure
The more data you store, the larger the "attack surface."

2. Reduces Storage and Compliance Costs

Cloud storage is cheaper today, but long-term retention increases:
Maintenance costs
Backup costs
Compliance obligations
Audit complexity

3. Maintains Ethical Responsibility

Organizations have a duty to respect user privacy by ensuring data is not kept forever or misused later

4. Supports Legal Compliance

Many data protection laws require secure destruction:
GDPR – "right to be forgotten"
India's DPDP Act – delete data after purpose is fulfilled
HIPAA – requires physical destruction of medical data
PCI-DSS – mandates deletion of financial details
Failure to dispose of data properly can result in fines.

Data Disposition

Key Methods of Data Disposition

1. Deletion (Soft and Hard)

Soft Deletion

Data is marked as deleted but still exists in the database.
Reversible (e.g., trash bin, recycle bin).
Suitable for systems requiring restoration.

Hard Deletion

Data is permanently removed from the database.
Cannot be recovered using normal tools.
Required for sensitive personal data.

2. Secure Destruction

Physically or digitally destroying storage media.

Methods include:

Shredding hard drives
Degaussing magnetic tapes
Physical destruction of USB drives
Overwriting disks multiple times ("wiping")
Used in defense, government, and healthcare where data is extremely sensitive.

Data Disposition

3. Archiving
Storing data in long-term, low-access systems.
Used for:
Legal compliance
Historical analysis
Research
Organizational memory
Archives require:
Restricted access
Encryption
Documentation of purpose
Important: **Archived data must not be used for new purposes without consent.**

4. Anonymization
Modifying data so individuals cannot be identified.
Techniques include:
Removing personal identifiers (names, IDs)
Generalizing values (age → age group)
Adding noise
Aggregation
Suppression
Good anonymization is irreversible.

Data Disposition

5. Tokenization / Pseudonymization

Replacing identifying fields with random tokens or pseudonyms.

Example:

Replace “Amit Sharma” with “User12345”.

This protects identity while allowing analysis continuity.

Unlike anonymization, pseudonymization is **reversible** with a key—so it must be protected.

Data Scraping

Definition

Data scraping—also called web scraping—is the automated process of collecting information from websites or digital platforms using tools like bots, scripts, or crawlers. Scraping extracts structured or unstructured data such as text, images, user profiles, reviews, or metadata

Why Data Scraping Matters in Data Ethics

Data scraping sits at the intersection of **technical capability**, **legal constraints**, and **ethical responsibility**. Just because information is publicly visible does **not** mean it is ethically free to collect, store, or reuse.

Organizations often assume that public = permissible, but ethical data science demands consideration of:

User expectations
Context in which data was shared
Consent and awareness
Power imbalance between platforms and scrapers
Risk of privacy breaches or harm

Data Scraping

Key Ethical Considerations

1. Consent

Users generally do **not** explicitly consent to having their data harvested by third-party bots. Even if the data is public, users expect humans—not automated systems—to view it.

Why it matters:

Consent is fundamentally about respecting user autonomy and control. Large-scale scraping bypasses user choice.

Example:

A research team scrapes 10 million Twitter posts to predict mental health patterns without informing users. Even though tweets are public, users didn't consent to their emotional content being used in a psychological study

2. Terms of Service (ToS) Violations

Most websites—Facebook, LinkedIn, Instagram, TikTok—explicitly prohibit scraping in their ToS agreements.

Ignoring ToS:

Risks legal action

Damages trust

Violates the platform's governance rules

Example:

Scraping LinkedIn user profiles using bots violates LinkedIn's ToS. LinkedIn has won multiple legal cases against companies that mined its data without permission.

Data Scraping

3. Privacy Risks

Public data can still reveal **sensitive information** when aggregated or combined.

Scraping can inadvertently expose:

Location histories

Personal identities

Behavioral patterns

Political or religious views

Example:

A group scrapes public Instagram posts to study tourist behavior and accidentally reveals individual users' home addresses and daily routines.

4. Context Collapse

Data posted in one context may not be appropriate in another.

Example:

A teenager's YouTube comment about depression may be harmless in a social context, but using it in a mental health prediction model without consent is unethical and disrespectful

5. Power Imbalance

Large corporations or well-resourced teams can gather massive amounts of user-generated data and use it for profit or surveillance in ways the individual never intended.

Why this is a concern:

Scraping can amplify digital inequality by transforming user data into commercial or predictive tools—without compensation or acknowledgment.

Data Scraping

6. Potential for Harm

Scraped data may be used for:

Identity theft

Facial recognition systems

Political profiling

Misinformation

Discriminatory algorithms

Example:

A company scrapes millions of images from social media to train facial recognition AI, which is later used in law enforcement. Users had no idea their photos were being used for surveillance technolog

Mosaic Data

Mosaic Data

Definition:

Mosaic data refers to combining fragments of information from multiple sources to create a detailed profile of a person or group.

Explanation:

Even when individual pieces are harmless, together they can reveal sensitive information (mosaic effect).

This raises severe privacy concerns because people may not be aware of how their scattered data can be assembled.

Example:

A user's:

- Instagram photos
- Public tweets
- Shopping history

can be combined to infer mental health, income, and daily habits.

Found Data

Definition:

Found data refers to data collected unintentionally or passively, not through planned scientific design (e.g., GPS logs, social media posts, browser history).

Ethical challenges:

•Lack of informed consent
•Unclear provenance
•Unknown sampling bias
•High risk of re-identification

Example:

Using Twitter posts to infer political sentiment without consent may violate user expectations and national regulations.

Thank You