Intermediate Data Science

Data Management and Ethics

Joanna Bieri DATA201

Important Information

- Email: joanna_bieri@redlands.edu
- Office Hours take place in Duke 209 – Office Hours Schedule
- Class Website
- Syllabus

Data Management and Ethics - Overview

In my mind data management and ethics go hand in hand, even though they are two very different sides of the same coin.

Data Management

Data management is the process of collecting, storing, organizing, maintaining, and using data efficiently and securely throughout its lifecycle.

Data Management

- Ensuring accuracy, availability, security, and consistency of data.
- Using proper tools, systems, and policies to manage data infrastructure.

Data Management

Examples of data management tasks

- Creating and maintaining databases.
- Implementing backup and recovery systems.
- Ensuring data quality and standardization.
- Controlling access and permissions.
- Complying with data governance and privacy laws

The goal is to make data reliable, accessible, and usable for decision-making and analysis.

Data Ethics

Data ethics deals with the moral principles and values guiding how data is collected, used, shared, and interpreted.

Data Ethics

- Ensuring fairness, transparency, accountability, and respect for privacy.
- Considering the human and social impact of data practices.

Data Ethics

Examples of data ethics questions

- Is it ethical to collect data without informed consent?
- Are algorithms being used in a way that's fair and unbiased?
- Who benefits or is harmed by a data-driven decision?
- Are individuals rights and dignity being respected?

To ensure data is used in ways that are responsible, just, and aligned with societal values.

Where They Meet

Privacy and Security

- *Data management* sets up secure systems to protect personal data.
- *Data ethics* asks whether collecting or storing that personal data is appropriate in the first place.
- Both aim to protect peoples privacy and prevent harm.

Where They Meet

Data Governance and Responsibility

- *Data management* creates rules for who can access data.
- *Data ethics* ensures those rules are fair and respect individual rights. - Both promote responsible use of data.

Where They Meet

Accuracy and Honesty

- *Data management* ensures data is clean and correct.
- *Data ethics* emphasizes truthful use and reporting of that data.
- Both care about trust and integrity in data work.

Where They Meet

Compliance and Transparency

- *Data management* ensures organizations follow data laws and policies.
- *Data ethics* asks whether those laws are enough to protect people and encourage openness.
- Both value transparency and accountability.

Where They Meet

Example

Imagine your school collects data about students grades and attendance:

- Good **data management** keeps that information safe, organized, and accurate.
- Good **data ethics** makes sure its only used for fair reasons - like helping students succeed - not for unfair judgment or discrimination.

The Research Data Lifecycle

Research data goes through several stages, from planning a project to sharing your results. As a Data Scientists you might be involved in any of these stages.

Each stage has best practices and tools that help you keep your data organized, ethical, and reusable.

The Research Data Lifecycle

Plan & Design

Goal: Set up your project and data management strategy before you start collecting data.

The Research Data Lifecycle

Best Practices:

- Create a Data Management Plan (DMP) that explains how you'll collect, store, and share your data. It is important to think about this ahead of time!
- Organize your folders, files, and naming systems clearly.
  - Avoid spaces in your file and folder names.
  - Name folders and files so they are easily findable.
  - Add a README.md file to your folders describing what is contained.

The Research Data Lifecycle

In this class we won't work as much with experimental design. These ideas should have been covered in MATH 111 and are often dependent on your area of expertise.

The Research Data Lifecycle

Collect & Create

Goal: Gather or generate high-quality data responsibly.

The Research Data Lifecycle

Best Practices: - Learn how to find existing datasets you can reuse. - Cite any data or code you use from others to give proper credit.

The Research Data Lifecycle

Proper citation is key! We will often use public data sets for analysis in this class. Be sure to know who who owns the data, what license protects the data, and any rules for data use and dissemination. You should reference the site where you collected the data:

The Research Data Lifecycle

Sometimes citations are discipline specific, but in general you should follow APA or MLA guidelines. For example:

Format

"Title of Webpage." Website Name, Publisher (if different), Publication Date, URL. Accessed Date.

Example

"UN Data Portal." United Nations, 2024, https://data.un.org. Accessed 7 Oct. 2025.

The Research Data Lifecycle

Analyze & Collaborate

Goal: Work with data ethically and transparently while analyzing results.

The Research Data Lifecycle

Best Practices:

- Keep your data and files well-organized so others can understand your work.
- Use the Open Science Framework (OSF) or similar tools for collaboration.
- Include documentation to describe your data. (README.md)
- Protect confidential or sensitive data following ethical guidelines.

Key Features of the OSF

1. Organize Projects

- Keep all your files, data, code, and notes in one place.
- Use folders and tags to stay organized.

Key Features of the OSF

2. Collaborate

- Work with classmates or lab partners.
- Control who can see or edit different parts of your project.

Key Features of the OSF

3. Track Progress

- Document your research steps.
- Keep a clear record of what you did and when, which is great for reports or theses.

The Research Data Lifecycle

Key Features of the OSF

4. Share and Publish

- Make your work publicly available whenever possible. Its ok to keep it private until ready or if the data is private.
- Share data, code, or preprints with the scientific community.

Key Features of the OSF

5. Support Reproducibility

- Anyone can see your workflow, data, and methods.
- Helps ensure research is transparent and trustworthy.

Key Features of the OSF

6. Integrations

- Connect with GitHub, Dropbox, Google Drive, and other tools you might already use.

The Research Data Lifecycle

Evaluate & Archive

Goal: Safely store your data for future use and make it accessible when appropriate.

The Research Data Lifecycle

Best Practices:

- Use secure storage and backup systems.
- Archive your final datasets in trusted repositories.
- Understand intellectual property (IP) and copyright issues.

The Research Data Lifecycle

Take advantage of websites like GitHub, https://zenodo.org/, https://osf.io/, and ReDATA (institutional Repos).

The Research Data Lifecycle

Share & Disseminate

Goal: Share your findings and data so others can learn from and build on your work.

The Research Data Lifecycle

Best Practices:

- Respect copyright and licensing rules.
- Publish your data in repositories to support reproducibility.
- Provide clear metadata and documentation so others can use your data.

*This might also include writing articles to share your knowledge or tutorials for other data scientists to follow?

The Research Data Lifecycle

https://data.library.arizona.edu/data-management/best-practices

The FAIR Principles

FAIR stands for **Findable, Accessible, Interoperable, and Reusable**. These are guidelines for handling data so that it can be used and shared responsibly. Think of it as a checklist for good data practices.

The FAIR Principles

Findable

- Data should be easy to find for humans and computers.
- Assign a unique identifier (like a DOI) to each dataset.
- Provide metadata (descriptions, keywords, author info).

Example for students:

- Naming your file `student_survey_2025.csv` instead of just `data.csv`.
- Adding a description like "Survey of 9th-grade study habits, 2025."

The FAIR Principles

Accessible

- Data should be easy to access using standard tools.
- Specify how it can be obtained, including permissions if it's restricted.

Example for students:

- Sharing a dataset on OSF or GitHub.
- Adding instructions if login is required to access it.

The FAIR Principles

Interoperable

- Data should work well with other datasets and software.
- Use standard formats and controlled vocabularies.

Example for students:

- Saving spreadsheets as `.csv` instead of a proprietary format like `.xlsx`.
- Using standard labels like “Age” or “Score” instead of abbreviations only you understand.

The FAIR Principles

Reusable

- Data should be well-documented so others can use it.
- Include license information, clear descriptions, and methods used.

Example for students:

- Providing a README file explaining your survey questions, coding, and any transformations.
- Adding a license like CC-BY to indicate others can reuse it with credit.

The FAIR Principles

In Short

FAIR = Make your data **Findable, Accessible, Interoperable, and Reusable**.
Following FAIR principles helps your work stay **organized, ethical, and useful** for others.

Data Ethics

Why is data ethics such a big deal? When governments, companies, and even individuals have the power to collect and analyze huge amounts of data - they also have the power to impact lives, for better or for worse.

Data Ethics

How does this apply to you? When you interact with data you should keep moral and ethical issues that you could come across as a data scientist:

Data Ethics

- Your boss tells you to do an analysis that will be used for decision making even after you have warned about strong bias in the data or algorithm. Should you give your boss the analysis? What if your job is on the line?

Data Ethics

- You notice that people you work with are not being careful with access to very private data because following the rules impacts their work flow. Should you say/do something? Even if the likelihood of a hack is low?

Data Ethics

- You are doing academic research and have struggled to get the data you need when suddenly hackers publish the exact data you need. Should you use it? Is all of it okay to use?

Data Ethics

- You found an awesome data set online and using it developed an awesome state of the art algorithm. The data is protected with a Share-Alike license that says anything created from the original work (derivatives) must also be released under the same license. But you did so much work and could really make a lot of money from your idea. Should you try to sell it anyway?

Data Ethics

- You work for a company that is making decisions using and AI product. You know how the product was trained and that the underlying data probably produces something with significant bias. Should you say something? Even if this product is making millions?

Some case studies

From: Data Camp - Intro to Data Ethics

https://www.datacamp.com/blog/introduction-to-data-ethics

Some case studies

- In September 2018, hackers injected malicious code into British Airways' website, diverting traffic to a fraudulent replica site. Customers then unknowingly gave their information to fraudsters, including login details, payment card information, address, and travel booking information.

Some case studies

- In 2019, after Apple introduced its credit card to consumers, allegations of an algorithm with gender bias emerged. Several prominent tech executives (including Steve Wozniak, the famous technologist and cofounder of Apple) described receiving exponentially higher credit limits than their wives, with whom they shared assets. Besides gender, no clear factors could suggest such a difference.

Some case studies

- In March 2021, the privacy of over 533 million Facebook users was compromised when their data was posted on an open hackers' forum. It was one of the largest data breaches of all time. The incident raised concerns about how organizations store and secure personal information and whether they should be allowed access to such data in the first place.

Some case studies

https://www.statista.com/chart/24495/apps-sharing-personal-information-with-third-parties/

Principles of Big Data Ethics

Transparency: The people providing data, should understand how their data is being used. It should be communicated in a straightforward way; they should not need to be lawyers to understand.

Accountability: Data scientists and organizations should take full responsibility for the data they collect and store. They are responsible to make sure that the data is not misused or disseminated (breached).

Individual Agency: Individuals should have control over their personal data. You should be able to access, update, or remove their data from a dataset.

Data Privacy Privacy should be the expectation. Data should be protected from unauthorized access and use.

More Data Ethics Examples

https://www.datacamp.com/blog/introduction-to-data-ethics

These are summarized in the class notes!

Data Ethics: Public vs. Private Data

Public Data

Data that is already available to anyone (e.g., government statistics, public tweets, open datasets).

Ethical considerations:

- Even if data is public, consider context — was it shared with consent for public use?
- Avoid re-identifying individuals in datasets that appear anonymous.
- Respect the intent of the data source.

Data Ethics: Public vs. Private Data

Example:

Using public crime statistics for a research project is generally fine, but combining it with other datasets to track individual behavior could violate ethics.

Data Ethics: Public vs. Private Data

Private Data

Data collected from individuals with limited access (e.g., medical records, student grades, social media messages).

Ethical considerations:

- Consent: Individuals must know how their data will be used.
- Confidentiality: Protect the identities of people in the dataset.
- Security: Safeguard against unauthorized access or breaches.
- Purpose limitation: Only use data for the reasons it was collected.

Data Ethics: Public vs. Private Data

Example:

A survey of students' mental health requires secure storage, restricted access, and informed consent.

Data Ethics: Public vs. Private Data

Expectation of Privacy

Individuals have an expectation that private data will remain private, even if it's stored digitally.

Ethical use of data depends on:

- Transparency: Explain how data is collected and used.
- Respect: Avoid using data in ways people wouldn't reasonably anticipate.
- Minimization: Only collect what is necessary.

Key Principle:

Just because data is technically accessible doesn't mean it is ethically okay to use it without consent.

Data Ethics - Data Aggregation

Data aggregation is the process of combining data from multiple sources to generate insights, summaries, or trends. While it's very useful, it can raise several ethical concerns.

Data Ethics - Data Aggregation

Privacy Risks

Aggregated data can sometimes reveal personal information, even if individual records were anonymized.
Re-identification attacks can match aggregated data with other sources to identify individuals.

Data Ethics - Data Aggregation

Example:

Health data aggregated for research might accidentally reveal someone's medical condition if combined with publicly available information.

Data Ethics - Data Aggregation

Consent and Transparency

People may not know or agree that their data will be aggregated. Ethical aggregation requires informed consent or at least clear disclosure of how data will be used.

Data Ethics - Data Aggregation

Example:

Using location data from a fitness app without telling users that it will be combined with other datasets for research.

Data Ethics - Data Aggregation

Bias and Fairness

Aggregated datasets may over-represent or under-represent certain groups, leading to biased insights.
This can amplify inequalities or unfair decision-making when used in AI or policy.

Data Ethics - Data Aggregation

Example:

Combining census data with social media data could undercount marginalized communities if they are less represented online.

Data Ethics could be a WHOLE class!

This is just the tip of the iceburg. These themes will come up as we continue on in DATA201 and hopefully this lecture builds on your thinking from DATA101. However, I strongly encourage you to take a data ethics class!