Working With Data

Critical Skills and Practice

Week 9, Lecture 2

17th November 2026

What is data?

- Linguistically
  - The Latin word data is the plural of datum, "(something) given"
- Colloquially
  - Values that convey information
  - Data is existing information or knowledge represented or coded in some form suitable for better usage or processing

Etymology Online - Data

Information types

- Nominal
  - No order
  - e.g. Blood Type A, B
- Ordinal
  - Ordered but the difference between states cannot be quantified
  - e.g. Satisfaction Levels
- Interval
  - Difference between values can be quantified but has no fixed 0 value
  - e.g. Time of day
- Ratio
  - Difference between values can be quantified and has a true 0 value
  - e.g. Age, Height, Weight

Sources of Data

- Primary Sources
  - Data collected directly for a specific purpose.
  - E.g. Surveys, observations, sensor measurements
- Secondary Sources
  - Data originally collected by someone else and reused for another purpose.
  - E.g. Census data, health records, commercial databases, open data repositories
- Tertiary Sources
  - A summary or collection of published primary and secondary sources without synthesising new conclusions.
  - E.g. Bibliography, field guides, Wikipedia, AI summaries

Typical Data Life Cycle

1. Data collection – take measurements
2. Data storage - e.g., ".csv", ".json", online or locally
3. Data cleaning and preparation – e.g. Removing outliers, decoding
4. Data analysis – e.g., median, student-t test
5. Visualisation and interpretation – e.g., matplotlib in Python
6. Decision-making – e.g., Which conclusion to support
7. Preservation, sharing, or disposal – e.g., Uploading to UoE DataShare

Raw Data

What do people mean when they say "raw data"?

- A set of values that still have have human, machine or instrument errors
- It may contain outliers
- It may be encoded for ease of collection – e.g. E for Edinburgh and G for Glasgow

Data is never "raw"

- Human experience defines
  - Collection methods and instruments
  - Annotation categories
  - Measured variables

This influences what becomes available for analysis

"Data" is never an unbiased representation of information

Example: User Testing

Question: do audio stimuli alongside visual stimuli improve response time?

Task: Response times recorded for visual stimuli on a screen

- Move mouse left or right
- Listen to beep in left or right ear

Audience: University students on an Informatics course

Collection Method: Post test survey – Questions like "How easy was this for you?"

- Ranked response 1 to 5

Another university is trying to repeat these experiments, but their raw data is showing very different results. What variables might be causing this?

Turn to the person next to you and discuss? (1 minute)

Potential Reasons for Different Results

1. Different audience – less familiarity with using a computer mouse
2. Perceived ease – people have different internal scales of ease
3. Different computer hardware - mouse speed response, wired vs Bluetooth headphones
4. Hearing loss – Did the initial experiment account for different levels of hearing in different ears between subjects?
5. Different experimental conditions – One completed in a quiet room, and another in a shared office

A dataset is almost never perfectly reproducible

Data Ownership and Responsibility

You have just started your thesis project...

- The Professor says that there is a dataset from a few years ago they have for you to analyse, measuring the performance of a facial recognition system
- They swear that the data is on a drive in the shared office
- You find the hard drive, but the data isn't in a file format you are familiar with
- There isn't any documentation on what information is included in each file in the drive

What went wrong here?

What went wrong?

- It wasn't specific where in the lab the drive is, is it labelled? Should it be encrypted?
- The student who collected the data may well have written metadata for the files, but not stored them in the same location
- The student has since left the lab group, without leaving a guide
- No-one took responsibility to look at the dataset before handing it over to check that it is in a fit state to use

What should be changed?

FAIR Guidelines

Generally, a good dataset is:

- Findable
- Accessible
- Interoperable
- (Legally) Reusable
    - As a result of being Documented

GoFair Foundatiton - The FAIR Guiding Principles

Activity

Which principle of FAIR is most threatened by each of these aspects of the example scenario?

- The file is on a drive in the office (somewhere?)
- The metadata is nowhere to be found
- Is in an outdated file format
- The dataset has had outliers removed already (without declaring this)

Talk to the person on the other side of you (1 minute)

FAIR = Findable, Accessible, Interoperable, (Legally) Reusable

Types of Responsibility

Legal responsibility

Organisational responsibility

Moral responsibility

Legal Responsibility

If you are working with "personal information" in the UK, you must comply with GDPR

- Personal information is any data that relates to an identified or identifiable individual.

GDPR ensures we have the:

- Right to be informed
- Right of access
- Right to rectification
- Right to erasure
- Right to restrict processing
- Right to data portability
- Right to object

UK GDPR guidance and resources - ICO
Data Protection Principles - European Parliament

Organisation Responsibility

- An individual researcher is responsible for mishandled data resulting in a security breach
- Their team leader is also responsible for that security breach
- And the manager above them, up to the owner of the business
- A contractor of that business is also responsible for that breach of their users data

Moral Responsibilities: Consent

Is the user information being given consensually?

Consent - FIRES:

- Freely given – No negative consequences if the information is not shared
- Informed – What information is the person sharing
- Retractable – Their data can be removed from the dataset if requested
- Enthusiastic – A firm and certain "Yes"
- Specific – Where is the information being used

The following slides contain reference to text and images that are sexual or violent in nature.

Case Study: Sama Source and Meta smart-glasses

Sama Source (Sama) is a Nairobi based company subcontracted by both Meta and OpenAI

- Employees at Sama manually annotated video and images recorded using Meta's Ray-Ban to train a model
  - Video and images are recorded whenever using voice commands
    - This revealed credit card information, sex and naked people, unaware they are being recorded.

Is this consensual?

- It may have been in the terms and conditions, but was it clear to the wearer or someone in the vicinity that they are being recorded?

"Meta's AI smart-glasses and data privacy concerns: Workers say 'We see everything'" - Svenska Dagbladet (Feb 2026)

Moral Responsibility

- Video data can be collected in unethical ways
- So can the annotations of that data

Case Study: Meta, OpenAI and Sama Source

- Employees at Sama also acted as content moderators for Meta
  - This consisted of being exposed to images and videos of murders, self-harm, suicides, sexual violence, explicit sexual content, and physical and sexual abuse of children.
- Employees also annotated text to train a model for detecting toxicity in the output of ChatGPT
  - These text sections were equally as explicit as the above.
- In February 2022, Sama were paid to collect sexual and violent images, some of them illegal under U.S. law, to deliver to OpenAI.

"More than 140 Kenya Facebook moderators diagnosed with severe PTSD" - Guardian (December 2024)
"OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic" - TIME (January 2023)

Is working with unethically collected data unethical?

The employees at Sama Source were heavily affected by this work

- The company supposedly offered counselling but how easy was this to access?
- Is it reasonable to ask employees to collect data deemed too shocking to be legal in the parent company's country?
- Are these employees able to withdraw their consent without negative consequence? (Loss of income)

Tasks for you

- Make sure you are aware of your legal responsibilities in handling and maintaining any experimental data
- Think about what information you are sharing throughout your interaction with the world

Learning Outcomes

- Know what types of information are represented in data
- Understand the ways that data fails to represent information
- Understand where poor data management causes issues in the future
- Know the different ways that you hold responsibility in data management
- Understand the ways that unethical practices show up in the use of AI