

Unit 4: Data Collections

Ethical and Social Issues Around Data Collection

Aligned to:

AP Computer Science A Course and Exam Description
Effective Fall 2025

Topic 4.1

This work is licensed under the
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

<https://longbaonguyen.github.io>

Before Any Code

- Everything in this unit (arrays, ArrayLists, files, searching, sorting) exists to work with data.
- So before the syntax, one question worth asking:
- Where did this data come from, and what did collecting it cost?
- That question is now part of the course and can be assessed.

Personal Data

- Data about people is different from data about temperatures or dice rolls.
- Personal data that is collected and stored can be used to identify someone, to follow what they do, or to reveal something they did not choose to share.
- Stored data also lasts. Data collected today can be combined with data collected years from now, for a purpose nobody has thought of yet.
- Deciding to store data is a decision with long-lasting effects.

Not Really Anonymous

- A fitness app published a global heat map of popular running routes. It contained no names and no accounts, only aggregated activity.
- It showed the perimeter of military bases whose locations were not public.
- And a single user's runs formed a loop that started and ended at their house.
- Removing names does not remove people. Aggregate data can still point at an individual, and often does.

Consent

- The runners in that story did agree to something. They tapped "share my activity."
- Did they agree to a public map? To their base being visible? To a stranger finding their home?
- Consent to collect is not consent to every use. A user agreeing to one purpose has not agreed to all future purposes.
- When you design a system that collects data, you are deciding what people are agreeing to, usually without them reading it.

Data Quality

- A data set is a record of what somebody managed to measure, which is not the same as a record of what is true.
- Values can be missing, entered wrong, out of date, or measured differently in different places.
- Before you compute anything from a data set, examine it. An average of inaccurate values is still inaccurate, but it looks authoritative.

Who Is Missing?

- The heat map showed where people with expensive smartwatches run.
- Read as a map of where people run, it is misleading. No value in it is false, but a great many people were never included.
- This is bias by omission, and it is hard to spot, because missing data leaves no trace in the data you do have.
- Always ask: who would be in this data set if it were complete, and who is not?

Algorithmic Bias

- The CED names this one, so learn the phrase:
- **ALGORITHMIC BIAS**: systemic and repeated errors in a program that create unfair outcomes for a specific group of users.
- This does not describe a prejudiced programmer. It describes a program that repeatedly produces worse results for one group, usually because of the data it was built from or tested on.
- Before you use a data set to extrapolate new information or draw a conclusion, know **HOW** it was collected and what bias that method could introduce.
- Incomplete or inaccurate data can also make the program itself work incorrectly or inefficiently.

Choosing a Data Set

- Given a question, not every data set can answer it. Check four things:
- Relevance: does it actually measure the thing you are asking about?
- Completeness: who or what is missing?
- Currency: how old is it, and has the world moved on?
- Permission: are you allowed to use it for this?
- A data set that fails any one of these will still produce an answer, which is what makes it risky.

Your Turn

- A school logs every hall-pass swipe: student, time, origin, destination. Someone proposes using it to flag students who may be struggling.
- Discuss:
 - 1) What could this get right?
 - 2) Which students would this system misread, and why?
 - 3) The students agreed to the swipe system for safety. Have they agreed to this?
 - 4) Who should be allowed to see the flags, and for how long should the data be kept?

References

- 1) Building Java Programs: A Back to Basics Approach by Stuart Reges and Marty Stepp
- 2) Runestone CSAwesome Curriculum:
<https://runestone.academy/runestone/books/published/csawesome/index.html>

For more tutorials/lecture notes in Java, Python, game programming, artificial intelligence with neural networks:

<https://longbaonguyen.github.io>