

Handle With Care

Ethical and social issues around data collection



What does the app really need?

PAUSE AND TRY

Which of these does a free weather app need?

THE APP ASKS FOR

1. Your city or ZIP code
2. Your contact list
3. Your photos
4. Your birthday
5. Your exact location, every minute

Talk with a partner. Which items does a forecast need? What could happen to the rest?

By the end of class you can...

- 1 Explain how collecting and storing personal data puts privacy at risk, and how a programmer can protect it.
- 2 Define algorithmic bias and tell it apart from a one-time glitch.
- 3 Explain how the way data was collected can make conclusions unfair or wrong.
- 4 Explain how incomplete or inaccurate data makes a program give wrong results.
- 5 Decide whether a data set fits the question you want to answer.

A large white square with rounded corners, tilted slightly to the right, containing a large teal number '1'.

1

PART 1

Personal data and privacy

Stored personal data is at risk



Personal data is information about a specific person. **Privacy** is a person's ability to control who knows it. When a computer stores personal data, privacy is at risk.



Seen

A file is copied, a device is lost, or someone breaks in.



Misused

Emails collected for reminders get sold to advertisers.



Combined

Harmless pieces, put together, point to one person.



Kept

Data stored for years keeps the risk alive for years.

None of these needs a bug. A program that works perfectly still holds the data.

Harmless pieces add up

Born on May 9 matches about 3 students in a school of 1,200

... and in 11th grade matches maybe 1 or 2 students

... and on the chess team matches exactly 1 student



"Each field is harmless" is not the same as "the data set is harmless." Combined, the pieces can **identify a specific person.**

How programmers safeguard privacy

Habit	What it means
Collect only what is needed	Ask for a field only if the program's purpose requires it
Tell users and ask first	Say what is collected and why, and get agreement
Use data only for its purpose	Data collected for reminders is used for reminders
Show as little as possible	Display only the part of a value a person needs to see
Limit who can reach it	Only the people and programs that need the data can open it
Delete it when done	Data that no longer exists cannot be exposed

Strongest habit: data that was never collected can never be exposed.

Show only what the user needs

A CHECKOUT SCREEN

```
String card = "6011223344556677";  
String last = card.substring(card.length() - 4);  
System.out.println("Gift card ending in " + last);
```

OUTPUT

Gift card ending in 6677



`card.length()` is 16, so **`card.length() - 4`** is 12, and `substring(12)` keeps the last four characters. Printing the whole number would show far more than the shopper needs.

A bus app that shows parents where the bus is

PAUSE AND TRY

Which design best protects privacy?

- A. Record every rider's name and stop, and keep the records forever, in case they are useful.
- B. Share the bus's location only, and only while it is running its route.
- C. Track each child's phone location all day, so the bus is never missed.
- D. Store each child's home address in a private instance variable.

A large, white, rounded square containing the number '2' in a bold, teal, sans-serif font.

2

PART 2

Algorithmic bias

Bias is not bad luck



Algorithmic bias is a **systemic** and **repeated** error in a program that creates **unfair outcomes** for a **specific group** of users.

A RANDOM GLITCH

The server goes down for ten minutes.
Whoever was online loses their work. Next
time it could be anyone.

A reliability problem, not bias.

ALGORITHMIC BIAS

A form rule turns away everyone with a
certain kind of name, every time they try.

Systemic, repeated, and aimed at one group.

A rule that turns people away

THE RULE

```
public static String check(String lastName) {  
    if (lastName.indexOf("-") == -1) {  
        return lastName + ": accepted";  
    }  
    return lastName + ": rejected";  
}
```

MAIN PRINTS EACH CALL

```
check("Nguyen")  
check("Smith-Jones")  
check("Patel")  
check("Garcia-Lopez")
```

OUTPUT

```
Nguyen: accepted  
Smith-Jones: rejected  
Patel: accepted  
Garcia-Lopez: rejected
```



No crash, no error message. Every hyphenated name is rejected, **every time**.

"It worked for everyone I tested"



If every test case looks like the programmer, a biased rule **passes every test**. The unfair result shows up only when real users arrive.

TESTS THE TEAM RAN

Nguyen: accepted
Patel: accepted
Brown: accepted

All pass. The rule looks fine.

TESTS THE TEAM NEEDED

Smith-Jones
Garcia-Lopez
a very long name, a one-letter name

Names from every group the program serves.

A large white square with rounded corners, tilted slightly to the right, containing the number '3' in a bold, dark teal font.

3

PART 3

Can you trust the data?

Who did the data come from?

ASKED AT THE BIKE RACK

```
int asked = 20;  
int bike = 18;  
System.out.println(bike * 100 / asked  
    + "% bike to school");
```

OUTPUT

90% bike to school

ASKED IN EVERY HOMEROOM

```
int asked = 400;  
int bike = 60;  
System.out.println(bike * 100 / asked  
    + "% bike to school");
```

OUTPUT

15% bike to school



Before you draw conclusions or **extrapolate** from data, ask how it was collected. A **collection method** that reaches some people more than others gives **biased** data.

More answers do not fix a biased method



A survey posted where only one kind of person sees it is still biased, **no matter how many people answer.**

BIGGER, STILL BIASED

Ask 2,000 students, all at the bike rack.

Every answer comes from someone who bikes.

SMALLER, FAIRER

Ask 400 students, spread over every homeroom.

Bikers, walkers, riders, and drivers all answer.

Missing values and wrong values

QUIZ SCORES

```
int total = 0;  
total += 90;  
total += 85;  
total += 0;    // absent, not a real 0  
total += 95;  
System.out.println("Average: " + total / 4);
```

OUTPUT

Average: 67

Incomplete

Some values are missing, like the absent quiz.

Inaccurate

Some values are wrong, like 95 typed as 9.

Either one can make a correct program give wrong results, or waste work.

What went wrong with the data?

PAUSE AND TRY

Biased collection, incomplete, or inaccurate?

What happened	Problem
A survey on library hours is given only to people inside the library.	
A sign-up sheet lists a student's age as 61 instead of 16.	
A rain gauge was broken for a week, so that week has no readings.	
A poll about bedtimes is posted online at midnight.	

A large, bold, teal-colored number '4' centered within a white rounded square with a slight drop shadow.

4

PART 4

The right data for the question

Does this data set fit?



A data set is gathered for a **specific question**. It might not give correct answers to a different question.

The data: every bike-share ride last year, with the start station, the end station, and the date and time.

Question	Yes, no, or not reliably?
1. Which station has the most rides that start at 8 a.m.?	
2. Which month had the fewest rides?	
3. How many people in the city own a bike?	
4. Do riders enjoy using bike-share?	

Words to know

Personal data

Information about a specific person.

Privacy

A person's control over who knows information about them.

Algorithmic bias

A systemic, repeated error that is unfair to a specific group.

Collection method

How data was gathered: who, where, when, and how.

Extrapolate

Use data to predict beyond what was measured.

Incomplete data

Data with missing values. (Inaccurate data has wrong values.)

Five questions to ask about data



Pick an app you use. With a partner, ask these five questions about its data.

1. Should it collect this at all?
2. How was the data gathered, and from whom?
3. Is anything missing or wrong?
4. Is the rule fair to every group of users?
5. Does this data answer the question being asked?

Next: the Handle With Care puzzle worksheet.