

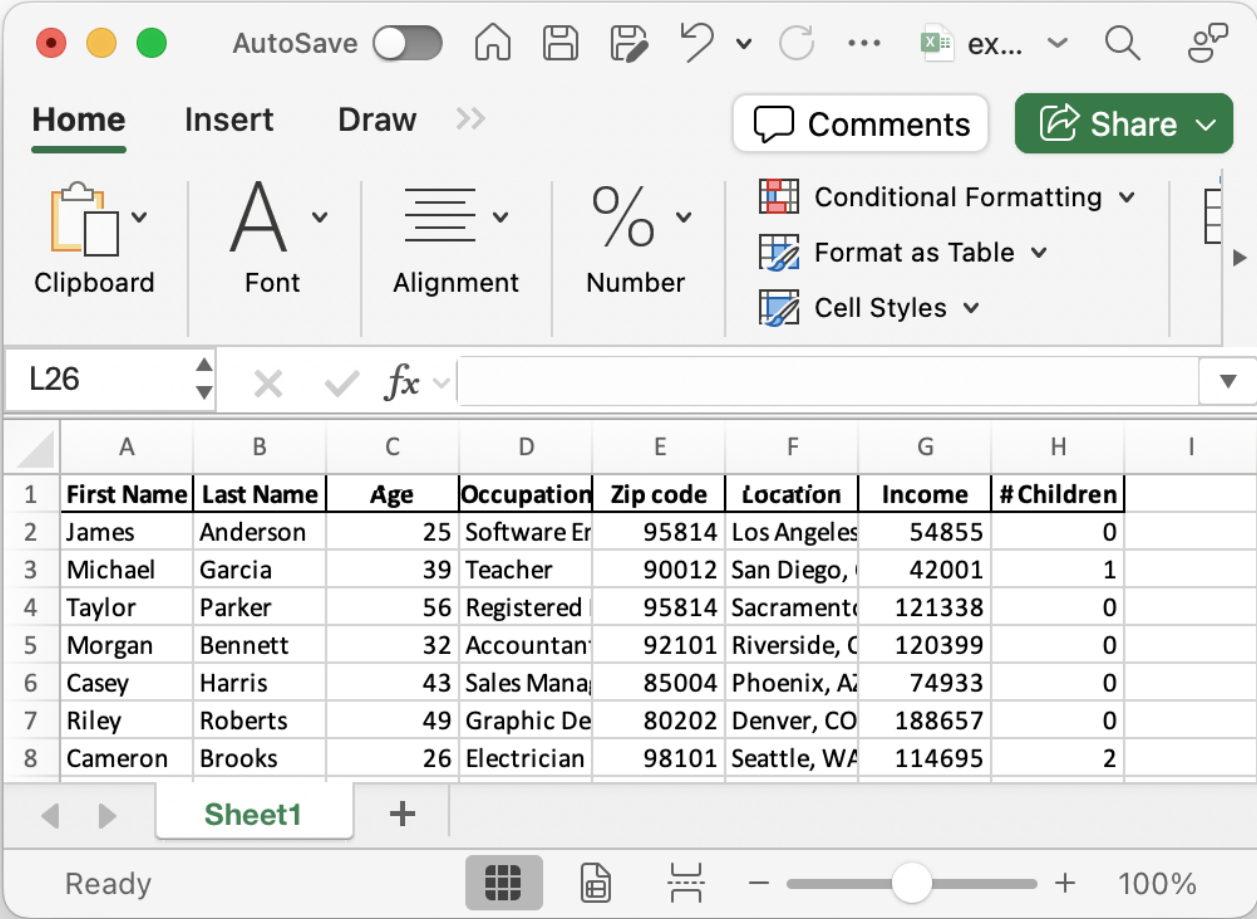
# Lecture 3: Anonymization

---

CS 181E

Fall 2026

# Example: Releasing a Dataset



The screenshot shows a spreadsheet application window. The title bar includes window controls, 'AutoSave' (disabled), and icons for home, save, print, undo, redo, and search. The ribbon has 'Home', 'Insert', and 'Draw' tabs. The 'Home' tab is active, showing groups for Clipboard, Font, Alignment, Number, Conditional Formatting, Format as Table, and Cell Styles. The active cell is L26. The spreadsheet contains a table with 9 columns: First Name, Last Name, Age, Occupation, Zip code, Location, Income, and # Children. The data is as follows:

|   | A          | B         | C   | D                 | E        | F               | G      | H          | I |
|---|------------|-----------|-----|-------------------|----------|-----------------|--------|------------|---|
|   | First Name | Last Name | Age | Occupation        | Zip code | Location        | Income | # Children |   |
| 1 | James      | Anderson  | 25  | Software Engineer | 95814    | Los Angeles, CA | 54855  | 0          |   |
| 2 | Michael    | Garcia    | 39  | Teacher           | 90012    | San Diego, CA   | 42001  | 1          |   |
| 3 | Taylor     | Parker    | 56  | Registered Nurse  | 95814    | Sacramento, CA  | 121338 | 0          |   |
| 4 | Morgan     | Bennett   | 32  | Accountant        | 92101    | Riverside, CA   | 120399 | 0          |   |
| 5 | Casey      | Harris    | 43  | Sales Manager     | 85004    | Phoenix, AZ     | 74933  | 0          |   |
| 6 | Riley      | Roberts   | 49  | Graphic Designer  | 80202    | Denver, CO      | 188657 | 0          |   |
| 7 | Cameron    | Brooks    | 26  | Electrician       | 98101    | Seattle, WA     | 114695 | 2          |   |

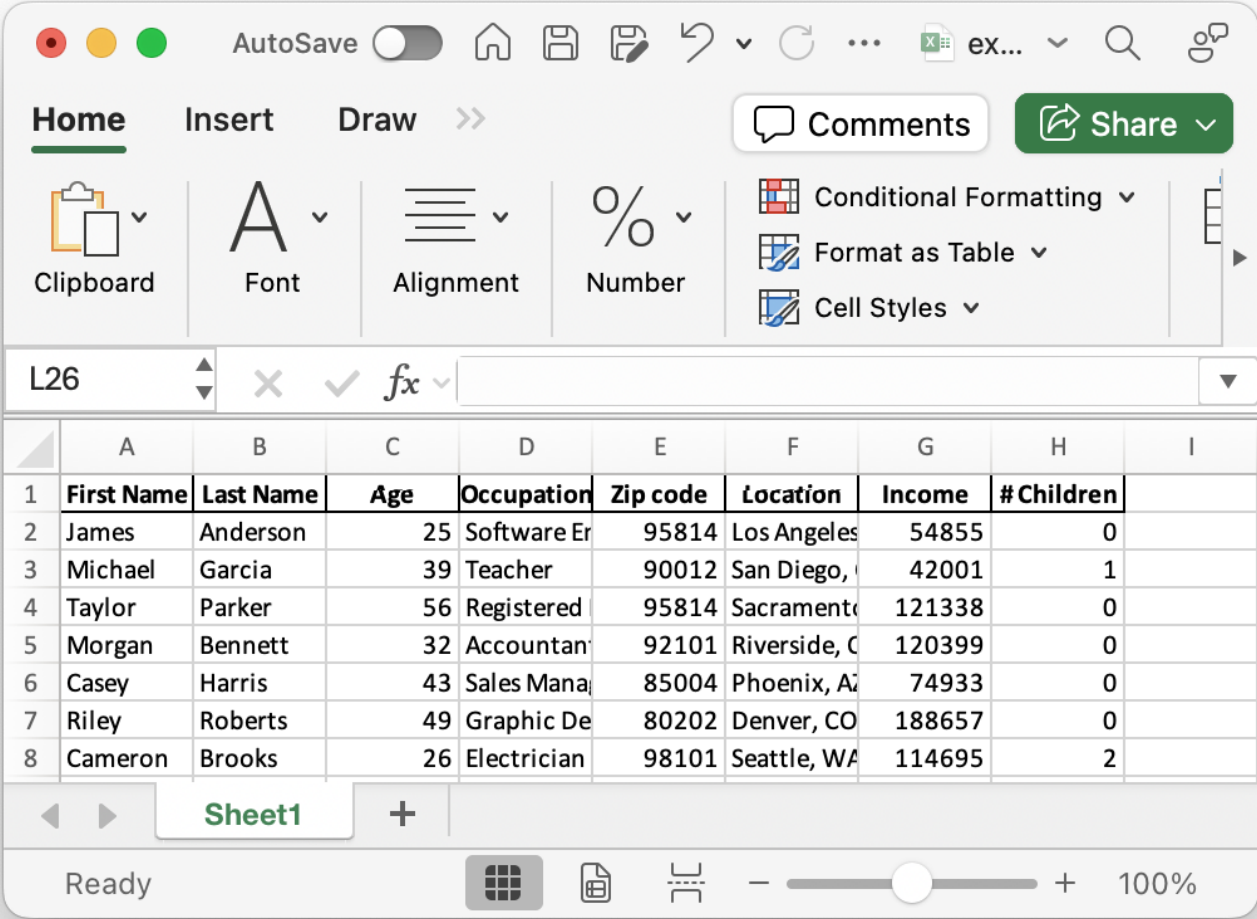
The bottom of the window shows 'Sheet1' as the active sheet, a status bar with 'Ready', and a zoom slider set to 100%.

**Question:** Would it be okay to publish this dataset?

# Personally-Identifiable Data (PII)

- anonymization efforts generally try to remove any information that can uniquely identify an individual
- exact definitions vary
- CCPA:
  - “Personal information” means **information that identifies, relates to, describes, is reasonably capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household.**
  - “Personal information” does not include consumer information that is deidentified
  - “Deidentified” means information that cannot reasonably be used to infer information about, or otherwise be linked to, a particular consumer

# Exercise 1: Anonymizing a Dataset



The screenshot shows a spreadsheet application window. The title bar includes window controls, 'AutoSave' (disabled), and icons for home, save, print, undo, redo, and search. The ribbon has 'Home', 'Insert', and 'Draw' tabs. The 'Home' tab is active, showing groups for Clipboard, Font, Alignment, Number, Conditional Formatting, Format as Table, and Cell Styles. The active cell is L26. The spreadsheet contains a table with 9 columns: First Name, Last Name, Age, Occupation, Zip code, Location, Income, and # Children. The data is as follows:

|   | A          | B         | C   | D                 | E        | F               | G      | H          | I |
|---|------------|-----------|-----|-------------------|----------|-----------------|--------|------------|---|
|   | First Name | Last Name | Age | Occupation        | Zip code | Location        | Income | # Children |   |
| 1 | James      | Anderson  | 25  | Software Engineer | 95814    | Los Angeles, CA | 54855  | 0          |   |
| 2 | Michael    | Garcia    | 39  | Teacher           | 90012    | San Diego, CA   | 42001  | 1          |   |
| 3 | Taylor     | Parker    | 56  | Registered Nurse  | 95814    | Sacramento, CA  | 121338 | 0          |   |
| 4 | Morgan     | Bennett   | 32  | Accountant        | 92101    | Riverside, CA   | 120399 | 0          |   |
| 5 | Casey      | Harris    | 43  | Sales Manager     | 85004    | Phoenix, AZ     | 74933  | 0          |   |
| 6 | Riley      | Roberts   | 49  | Graphic Designer  | 80202    | Denver, CO      | 188657 | 0          |   |
| 7 | Cameron    | Brooks    | 26  | Electrician       | 98101    | Seattle, WA     | 114695 | 2          |   |

The bottom of the window shows 'Sheet1' as the active sheet, a status bar with 'Ready', and a zoom slider set to 100%.

What modifications would you make to this dataset to retain utility without compromising anyone's privacy?

# Technique #1: Pseudonymization

- replacing identifying information (e.g., names) with unique identifiers
- chronological
- hashing
- randomly generated

# Technique #2: Suppression

- delete particular columns or values
- delete columns that are too sensitive
- delete rare or sensitive values

# Technique #3: Generalization

- replace fine-grained values with course-grained values
- Example:
  - replace DOB with age
  - replace age with age range
  - replace GPS coordinates with city or state

# Technique #4: Approximation

- Add random noise to values
- Potential concerns:
  - how much noise?
  - what distribution?
  - does this make the data useless?

# Technique #5: Aggregation

- Instead of raw data, release summary statistics
- Potential concerns:
  - does this make the data useless?

# What do people commonly do?

- focus on particular attributes (e.g., age, gender, location)
- mostly use suppression or generalization of entire values
- remove distinctive values or combinations of values
- informal and social processes for determining when done

Is this good enough?

# Case Study: Massachusetts Health Data

- Group Insurance Commission would release records with 100 attributes for every state employee's hospital visits
- Anonymized by removing any value that identified the individual (e.g., name, mailing address, insurance ID)
- Researchers found **quasi-identifiers** that were unique for a lot of people
  - 87%: ZIP code + full Date of Birth + Sex
  - 53%: *City* + full Date of Birth + Sex
  - 18%: *County* + full Date of Birth + Sex
- These same values are also in public records (e.g., voter registration)



# k-Anonymity

| Name    | Pronouns  | Grad        | Grade |
|---------|-----------|-------------|-------|
| Alice   | she/her   | Spring 2028 | 95    |
| Bob     | he/him    | Spring 2028 | 80    |
| Charlie | they/them | Spring 2028 | 95    |
| David   | he/him    | Spring 2028 | 60    |
| Edward  | he/him    | Fall 2027   | 80    |
| Flora   | she/her   | Spring 2027 | 99    |
| Georgia | she/her   | Spring 2027 | 60    |

- **Quasi-identifiers (QIs)** are sets of attributes that can be exploited for linking
- A database is **k-anonymous** if each QI maps to at least k different individuals
- Techniques: suppression and generalization

# Exercise 2: k-anonymity

- Modify this dataset to make it 2-anonymous with respect to Zipcode/DOB/Sex

| Zipcode | DOB     | Gender | Marital Status | Health Issues       |
|---------|---------|--------|----------------|---------------------|
| 91711   | 9/27/00 | female | divorced       | hypertension        |
| 91711   | 9/30/00 | female | divorced       | obesity             |
| 91711   | 4/18/00 | male   | married        | chest pain          |
| 91711   | 4/15/00 | male   | married        | obesity             |
| 91767   | 3/13/99 | male   | married        | hypertension        |
| 91767   | 3/18/99 | male   | married        | shortness of breath |
| 91767   | 9/13/00 | female | married        | obesity             |
| 91767   | 9/07/00 | female | married        | obesity             |
| 91101   | 5/14/01 | male   | single         | chest pain          |
| 91101   | 4/08/01 | male   | single         | obesity             |
| 91101   | 9/15/01 | female | married        | shortness of breath |

# $\ell$ -Diversity

- quasi-identifiers form equivalence classes of individuals who share the same values for all attributes in that set
- it is possible that everyone in an equivalence class also has other attributes they all share
- it is possible that an attacker has background knowledge that can further narrow down identity within an equiv class
- idea: multiple values of sensitive attributes should be well represented within all equivalence classes
- a dataset is  $\ell$ -diverse if at least  $\ell$  different values of each sensitive attribute appear about as often for each class

Key assumption: You can correctly identify quasi-identifiers

# Case Study: AOL Search Records

- AOL Research released 20,000,000 search queries for 650,000 users of AOL's search engine (3 months)
- Suppressed AOL username and IP address
  - Replaced them with unique, pseudonymous identifiers



# Case Study: AOL Search Records

## *A Face Is Exposed for AOL Searcher No. 4417749*

By Michael Barbaro and Tom Zeller Jr.

Aug. 9, 2006



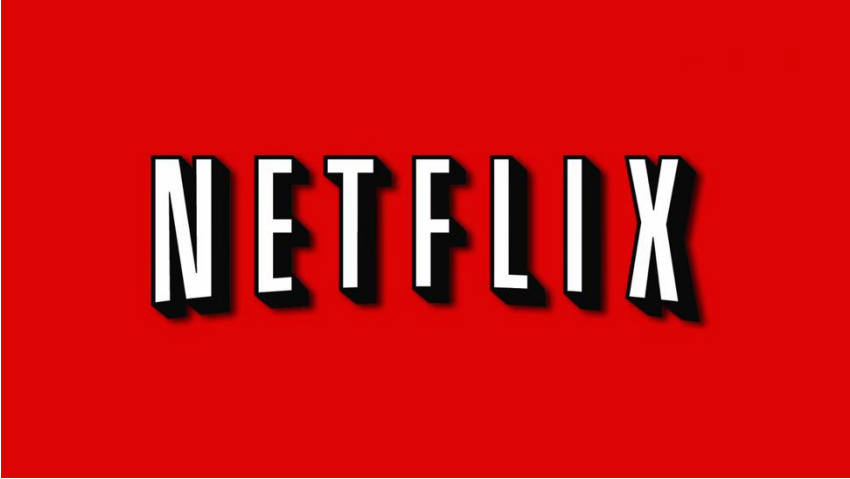
# Case Study: AOL Search Records

- User 4417749 example queries:
  - ‘landscapers in Lilburn, Ga,’
  - ‘homes sold in shadow lake subdivision gwinnett county georgia.’
  - several people with the last name Arnold
- Identified as Thelma Arnold, a 62-year-old from Lilburn, GA
- Arnold acknowledged authoring the searches, including some “mildly embarrassing”
  - ‘numb fingers,’
  - ‘60 single men’
  - ‘dog that urinates on everything’



# Case Study: Netflix Prize Data

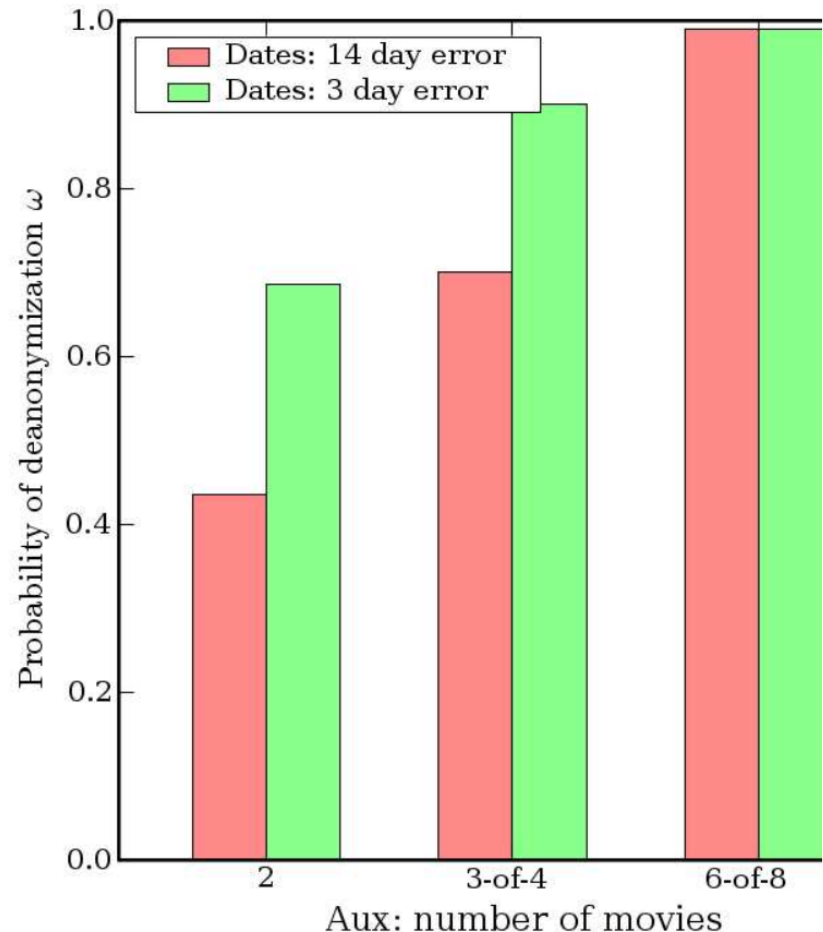
- Netflix released 100,000,000 records from 500,000 users
  - December 1999 to December 2005
  - Assigned a unique pseudonymous identifier to each user
- Each record included the pseudonymous identifier, the movie watched, the rating (1-5 stars), and rating's date

The Netflix logo is displayed on a solid red rectangular background. The word "NETFLIX" is written in a bold, white, sans-serif typeface. Each letter has a subtle 3D effect, with a dark shadow cast to its right and slightly downwards, giving it a floating appearance.

**NETFLIX**

# Case Study: Netflix Prize Data

- used IMBD ratings as external information
- defined two ratings to be similar if both the numeric rating and the date were within a threshold
- quantified similarity between two records as the number of similar ratings divided by the total number of movies rated in the union of the two records



# Why is anonymization so hard?

- Personal data showing up in unexpected places
- Errors suppressing data
- **Thinking only about personal data / PII is not sufficient**
- It's hard to reason about what is/is not identifiable
- Availability of **auxiliary information**
  - Extremely hard to control