Anonymised data

Data about people but not specific people

Figure 5: The rise of data science depends on, and incentivises a wealth of publicly available datasets: Kaggle counting more than 52K of them.

Netflix Prize Attack

- Netflix 1 million USD Netflix prize for movie recommendation
- 100 mil ratings, created by 480K users, over 6 years
- *All customer identifying information was removed*

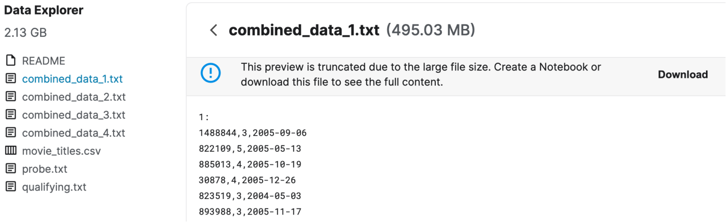

Figure 6: A slice of the Netflix dataset: each data record captures all the movie ratings given by an individual, including their dates.

Imperial College London

Pause to think: so what?

Discussion point

Does it matter if someone knows what movies you watched?

Pause to think: so what?

Discussion point

Does it matter if someone knows what movies you watched?

A movie rating can be perceived as suggestive of:

- Your sexual orientation
- Your political orientation
- Your beliefs about specific things (e.g., conspiracy theories)
- Your opinion about violence in movies

Even consumption of content is sensitive, but ratings even more so.
It is not wise to assume that you can always guess what is harmful.

The Netflix Prize attack

- With 8 movie ratings (of which 2 may be wrong) and dates $\pm$ 14 days, 99% of records can be uniquely identified.
- With 8 ratings (of which 2 may be wrong) and no dates, 84% of records can be identified.

Landmark publication on statistical de-anonymization (Narayanan and Shmatikov, 2008).

> *It is important that the Netflix Prize dataset has been released to support development of better recommendation algorithms. A significant perturbation of individual attributes would have affected cross-attribute correlations and significantly decreased the dataset's utility for creating new recommendation algorithms*

Group Insurance Commission attack

- In the 90s, the Group Insurance Commission released hospital records of state employees, suppressing all identifiable information, but keeping birth year, zip code and gender.
- The Governor of Massachussets publicly affirmed that this was privacy-preserving. Latanya Sweeney, now professor then student, used voter rolls to find his zip code and year of birth, identified his hospital record, and sent it to his office.
- Discussed in "Broken promises of privacy" Ohm, 2009.
- Indeed, date of birth, gender and zip code suffice to uniquely identify $> 80\%$ of US citizens in publicly available databases.

Discussion point

What could the Commission have done to avoid this?

Anonymous or pseudonymous?

Pseudonymisation

A data entry is pseudonymised when it has been processed in a way that it does not relate to an identifiable person.

Re-identification

The act of using processing and/or external information to relate a pseudonymised data entry to an identifiable person and hence make known something about them that was not known before.

Anonymisation

A data entry is anonymised when it is pseudonymised and processed in a way that precludes re-identification.

k-Anonymity

Assume a table where each row represents a personal data record, and each column represents an attribute of that person.

Quasi-identifier

A variable (attribute) that can also be observed in public data. For example, someone's name, job title, zip code, or email.

k-Anonymity

Assume a table where each row represents a personal data record, and each column represents an attribute of that person.

Quasi-identifier

A variable (attribute) that can also be observed in public data. For example, someone's name, job title, zip code, or email.

k-Anonymity

Consider the set of quasi-identifiers $A_1, \ldots, A_n$. A table is k-anonymous if each possible value assignment $v_1, \ldots, v_n$ to these variables is observed for either 0 or at least k individuals (i.e., among observed value assignments, each is shared by k data rows).

A theme in this course is that definitions matter.

An example: anonymising medical records

patient.id	zip.code	date.of.birth	race	disease	age
pid 3	23843	1964-07-03	Native Hawaiian/Pacific Islander	cardiovascular disease	57
pid 5	23843	1987-12-18	Asian American	diabetes	34
pid 11	23843	1986-11-19	American Indian/Alaska Native	diabetes	35
pid 13	23843	1960-02-02	Black or African American	viral infection	62
pid 14	23843	1964-07-28	Native Hawaiian/Pacific Islander	viral infection	57
pid 1	51523	1980-08-01	White or European American	diabetes	41
pid 4	51523	1978-12-10	Black or African American	viral infection	43
pid 6	51523	1975-11-04	Asian American	diabetes	46
pid 7	51523	1963-09-23	Black or African American	viral infection	58
pid 8	51523	1979-01-01	Asian American	cardiovascular disease	43
pid 9	51523	1966-05-21	White or European American	bipolar disorder	55
pid 10	51523	1976-09-13	Black or African American	cancer	45
pid 12	51523	1973-01-13	Asian American	cancer	49
pid 2	62422	1980-09-11	American Indian/Alaska Native	diabetes	41
pid 15	62422	1962-08-03	White or European American	viral infection	59

Table 1: A synthetic table of medical records.

Zip-code and date of birth are quasi-identifiers. Is race?[1]

[1] Here we use the five racial categories employed by the U.S. Census, but we will revisit the complex relationship of race to ethnicity and ancestry.

An example: anonymising medical records

patient.id	zip.code	date.of.birth	race	disease	age_group
pid 3	238**	*	Native Hawaiian/Pacific Islander	cardiovascular disease	41-60
pid 5	238**	*	Asian American	diabetes	21-40
pid 11	238**	*	American Indian/Alaska Native	diabetes	21-40
pid 13	238**	*	Black or African American	viral infection	> 60
pid 14	238**	*	Native Hawaiian/Pacific Islander	viral infection	41-60
pid 1	515**	*	White or European American	diabetes	41-60
pid 4	515**	*	Black or African American	viral infection	41-60
pid 6	515**	*	Asian American	diabetes	41-60
pid 7	515**	*	Black or African American	viral infection	41-60
pid 8	515**	*	Asian American	cardiovascular disease	41-60
pid 9	515**	*	White or European American	bipolar disorder	41-60
pid 10	515**	*	Black or African American	cancer	41-60
pid 12	515**	*	Asian American	cancer	41-60
pid 2	624**	*	American Indian/Alaska Native	diabetes	41-60
pid 15	624**	*	White or European American	viral infection	41-60

Table 2: A k-anonymised version of the table.

- *Suppressed* (censored) DOB
- *Aggregated* (coarsen) zip code and age

An example: anonymising medical records

	Equivalence class	Size	Unique disease values
1	238** AND > 60	1	1
2	238** AND 21-40	2	1
3	238** AND 41-60	2	2
4	515** AND 41-60	8	5
5	624** AND 41-60	2	2

Table 3: Counts of the quasi-identifier *equivalence classes* and unique diseases per class.

- Only one patient > 60. Suppress the patient?
- Both patients in 21-40 have the same disease.

Discussion: anonymity

- Possibility of re-identification depends on your data set as well as third party (public or not) datasets
- Quasi-identifiers are attributes shared between these sources
- k-anonymity tries to ensure that quasi-identifiers can at worst identify a group of k individuals that cannot be told apart
- information can still be revealed by a k-anonymous dataset if the equivalence class has low variability, even if it is non-zero (e.g., 'X either has diabetes or cardiovascular disease')

What if we release the model, but not the data?

Releasing a predictive model can indirectly release data, too! It also presupposes access by data engineers and data scientists, which introduces security risks, but also constitutes disclosure in itself.