

Privacy

Ethical Data Science

Dr Zak Varty

Privacy and Data Science

What is personal data?

- Name, national insurance number, passport number
- Contact details: address, phone number, email address
- Medical history
- Online activity, GPS data, finger print or face-ID,

Should not be collected, analysed or distributed without consent.

Privacy as a Human Right

Article 12 of the Universal Declaration of Human Rights

No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks. - [UN General Assembly](#), 1948

Data Privacy and the European Union

General Data Protection Regulation (2018)

‘Consent’ of the data subject means any freely given, specific, informed and unambiguous indication of the data subject’s wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her; - GDPR [Article 4](#)

Privacy: Key Terms

Pseudonimisation: processing data so that it does not relate to an identifiable person.

Re-identification: relating a pseudonymised data entry to an identifiable person.

Anonymisation: A pseudonimisation method that precludes re-identification.

Measuring Privacy: Pseudo-identifiers and k -anonymity

Pseudo-identifiers: Attributes that can also be observed in public data. For example, someone's name, job title, zip code, or email.

For the set of quasi-identifiers A_1, \dots, A_p , a table is **k -anonymous** if each possible value assignment to these variables (a_1, \dots, a_n) is observed for either 0 or at least k individuals.

k -anonymity example

	Post Code	Age	Drug Use	Condition
1	OX1****	<20	*	Herpes
2	OX1****	<20	*	Herpes
3	OX2****	>=30	*	Chlamydia
4	OX2****	>=30	*	Herpes
5	OX1****	<20	*	Gonorrhea
6	OX2****	>=30	*	Gonorrhea
7	OX1****	<20	*	Gonorrhea
8	LA1****	2*	*	Chlamydia
9	LA1****	2*	*	Chlamydia
10	OX2****	>=30	*	Gonorrhea
11	LA1****	2*	*	Chlamydia
12	LA1****	2*	*	Chlamydia

k -anonymity example (2)

	Post Code	Age	Drug Use	Condition	Equivalence Class
1	OX1****	<20	*	Herpes	1
2	OX1****	<20	*	Herpes	1
3	OX2****	>=30	*	Chlamydia	2
4	OX2****	>=30	*	Herpes	2
5	OX1****	<20	*	Gonorrhea	1
6	OX2****	>=30	*	Gonorrhea	2
7	OX1****	<20	*	Gonorrhea	1
8	LA1****	2*	*	Chlamydia	3
9	LA1****	2*	*	Chlamydia	3
10	OX2****	>=30	*	Gonorrhea	2
11	LA1****	2*	*	Chlamydia	3
12	LA1****	2*	*	Chlamydia	3

Improving Privacy

There are three main ways that you can improve the privacy within a dataset:

- Redaction (of columns or rows)
- Aggregation (Continuous -> discrete or combining discrete groups)
- Corruption / Noise

Breaking k -anonymity

Lack of diversity in private attributes within an equivalence class

	Post Code	Age	Drug Use	Condition
8	LA1****	2*	*	Chlamydia
9	LA1****	2*	*	Chlamydia
11	LA1****	2*	*	Chlamydia
12	LA1****	2*	*	Chlamydia

Also vulnerable to external data-linkage attacks.

Cautionary tale: Massachusetts Medical Data



- Release of public servant health data to help speed up medical research.
- William Weld, gave public assurances that privacy would not be compromised.
- Latanya Sweeney, then an PhD student at MIT dramatically disproved this claim.

Latanya Sweeney Speaking in New York, 2017.

Image CC-4.0 from [Parker Higgins](#).

Cautionary Tale: Netflix Competition



User ID	Film ID	Rating	Date
000001	548782	5	2001-01-01
000001	549325	1	2001-01-01
...

Wrapping Up

- Privacy is a fundamental concern.
- Privacy is hard to measure and hard to ensure.
- Also a model issue, since models are trained on data.
- No universal answers, but an exciting area of ongoing research.

