Big data, Ethics and Privacy

Jisu Kim[1]

[1] Max Planck Institute for Demographic Research

Outline

- Pros and cons
- Ethics and privacy
- GDPR
- How to respect GDPR

Pros and cons

Pros:

- Access
- Open box
- Longitudinal: access to historical data
- Sample size

Cons:

- Bias
- Sample size

Issue:

- **Ethics and Privacy**

Ethics and Privacy

Twitter API is an open-access

Eavesdropping someone's conversation in a bus $\rightarrow$ can I expose this information?

Does this mean that we can do anything with the data?

$\rightarrow$ Quick answer= NO!

Facebook-Cambridge Analytica scandal

What happened?

GDPR

General Data Protection Regulation was enacted in 2016: An EU law that aims to protect and regulate data and privacy.
7 principles[1]:

- **Lawfulness, fairness and transparency**
- Purpose limitation
- Data minimisation
- Accuracy
- **Storage limitation**
- **Integrity and confidentiality (security)**
- Accountability

For more info: `https://gdpr.eu`

[1] `https://www.onetrust.com/blog/gdpr-principles/`

Consent

Article 4(11) defines consent: "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."[2]

Consent also means that "the data subject knows your identity, what data processing activities you intend to conduct, the purpose of it and that they can withdraw their consent at any time."[3]

[2] https://gdpr.eu/gdpr-consent-requirements/
[3] Idem.

Confidentiality

- Pseudonymization: jisu $\rightarrow$ DKjxid73
- Anonymization: jisu $\rightarrow$ ******
- Data transformation: By modifying the format, value, or structure of data

Data storage limitation

How, where and for how long can one store sensitive and personal data?
$\rightarrow$ Before storing the data, think of pseudonymizing, anonymising or transforming the data

$\rightarrow$ Secure data storage: need to store data where it is safe and well protected where it can prevent possible attacks or loss of data

$\rightarrow$ The length of the time that you are planning to store the data is required by the GDPR.

When sharing Tweets

Sharing your Twitter is permitted as long as they are for non-commercial purposes. BUT you should not share entire tweets directly!

So how? $\rightarrow$ By dehydrating your data!

"Dehydrated" data set - each tweet is reduced to its unique ID number, and a list of these IDs is saved as a text document.[4]
Check out also this webpage:`https://covid.dh.miami.edu/2020/06/11/hydrating-tweetsets/`

[4] `https://scholarslab.github.io/learn-twarc/06-twarc-command-basics#dehydrated-and-rehydrated-data-sets`

Now let's collect our own!