18734: Foundations of Privacy

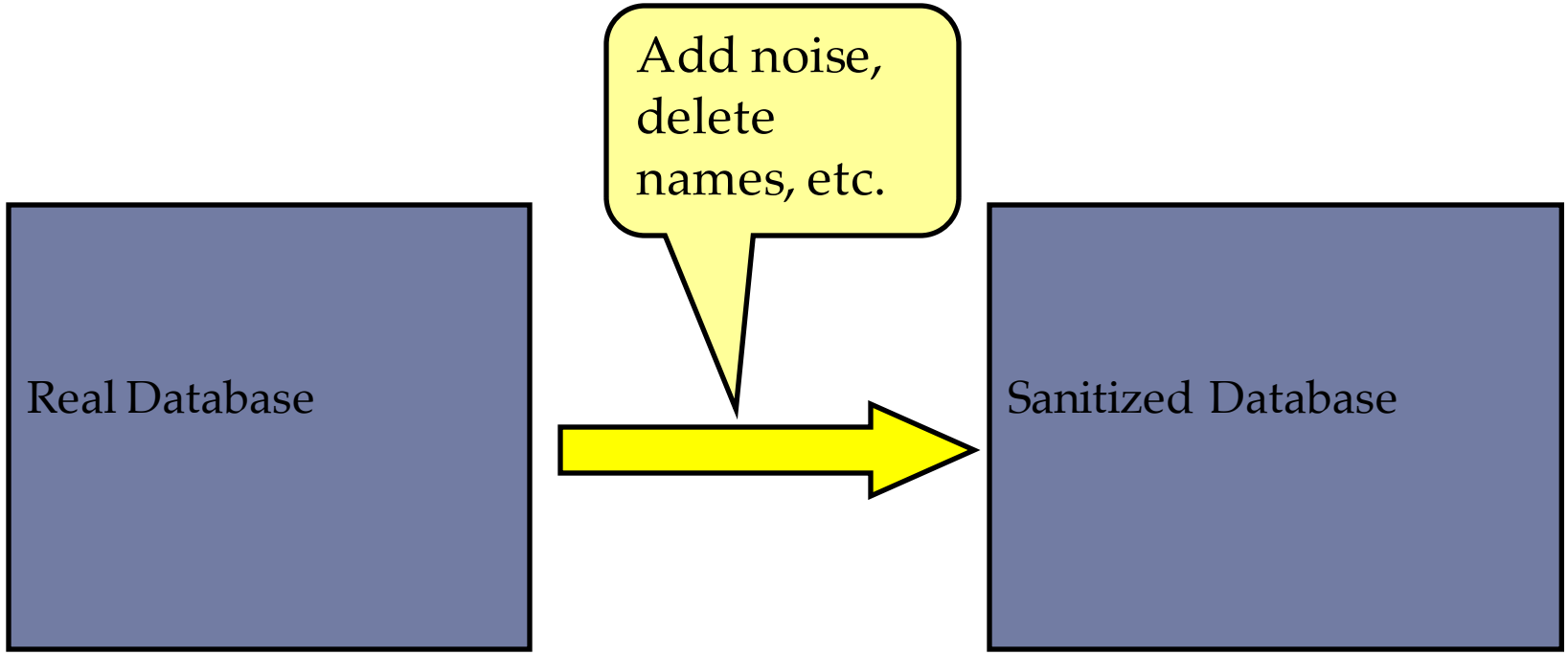
Database Privacy: k-anonymity and de-anonymization attacks

Piotr Mardziel or Anupam Datta
CMU
Fall 2018

Publicly Released Large Datasets

- Useful for improving recommendation systems, collaborative research
- Contain personal information
- Mechanisms to protect privacy, e.g. anonymization by removing names
- Yet, private information leaked by attacks on anonymization mechanisms

Sanitization of Databases



Health records

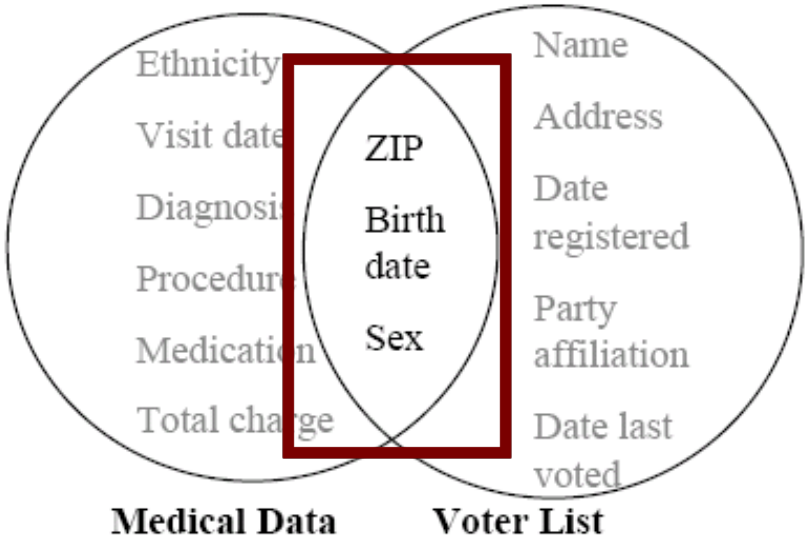
Census data

Protect privacy

Provide useful information (utility)

Re-identification by linking

Linking two sets of data on shared attributes may uniquely identify some individuals:



87 % of US population uniquely identifiable by 5-digit ZIP, gender, DOB

Netflix-IMDb Empirical Attack [Narayanan et al 2008]

Anonymized Netflix DB

	Gladiator	Titanic	Heidi
r_1	4	1	0
r_2	2	1.5	1
r_3	0.5	1	1

Publicly available IMDb ratings (noisy)

	Titanic	Heidi
Bob	2	1

Used as auxiliary information

Weighted Scoring Algorithm

Isolation Attack!

r_1	4	1	0