UC Love Data Week

February 13 - 17, 2023

Data De-Identification In Practice

Renata Curty & Greg Janée
Research Data Services, UCSB Library
rds@library.ucsb.edu

UC SANTA BARBARA

Plan for this Session

01

Introduction/
Key Concepts

02

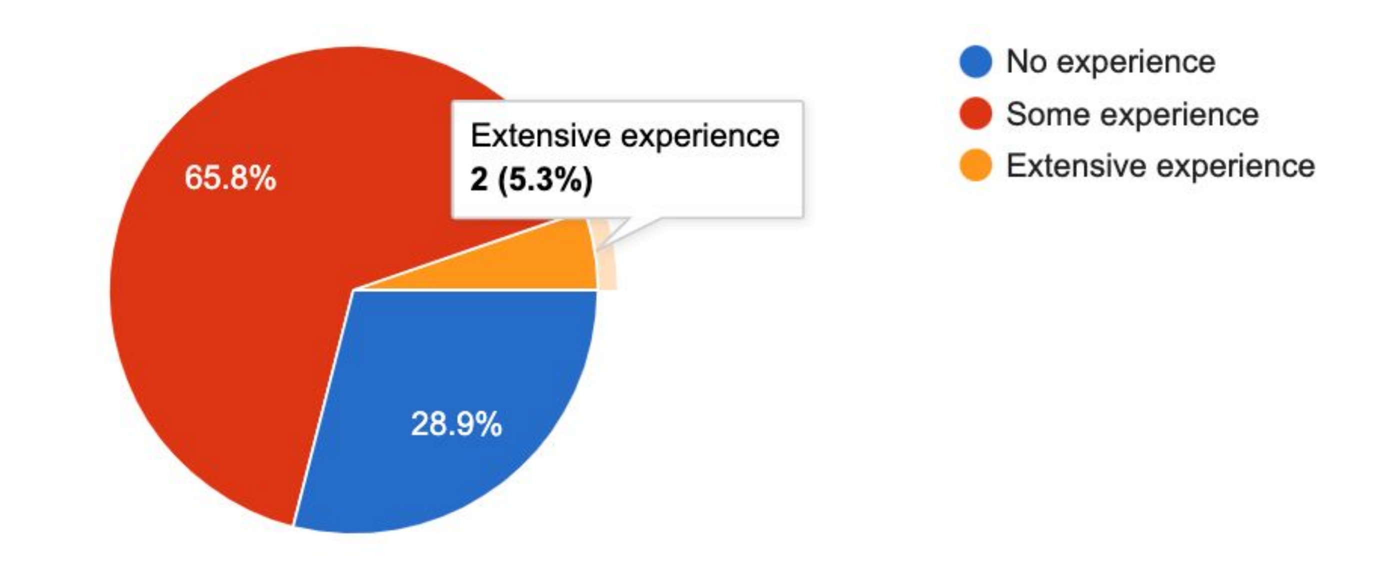
De-identification
Methods

03

Tools
(sdcMicro/sdcApp)

Please rate your prior experience with the workshop topic:

(n=38)



UC SANTA BARBARA

Motivation

- Handle sensitive data ethically and responsibly
- Compliance with data sharing mandates
- Maximize data reuse while preserving individuals' privacy

Sensitive Data

Data that must be protected against unwanted disclosure and which access should be safeguarded.

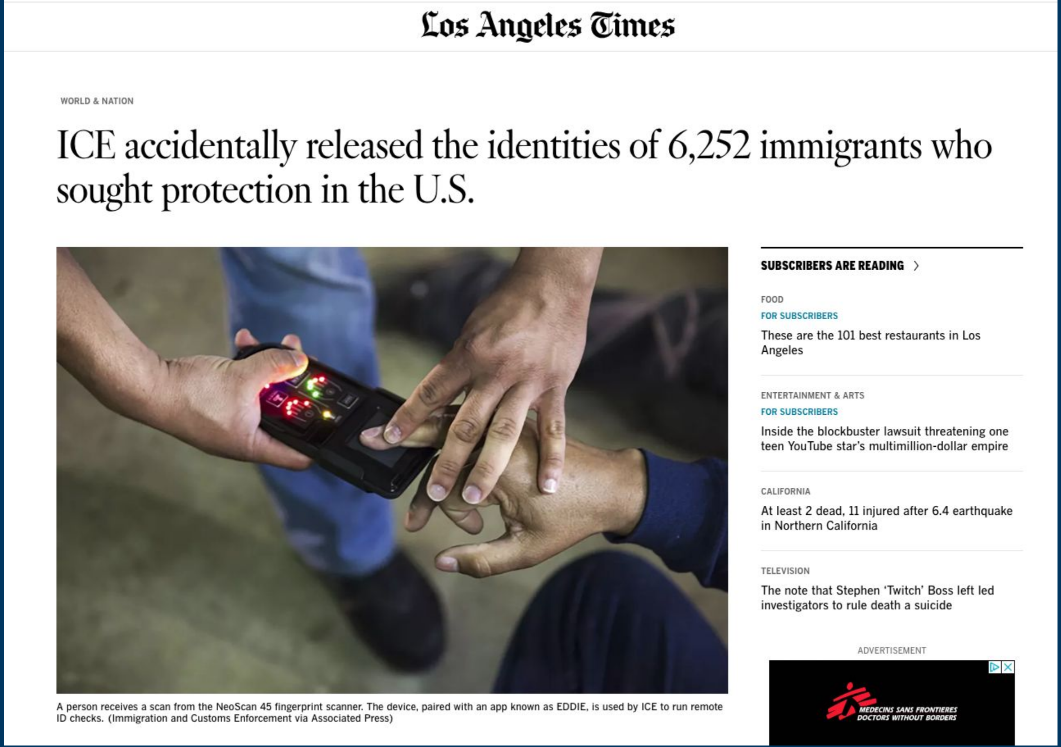
Protection of sensitive data may be required for legal or ethical reasons, for issues pertaining to personal privacy, or for proprietary considerations.

Whose protection?

Any human subject data that can potentially disclose people's identity and damage individual or collective reputations, rights, safety or best interests.

It also includes data, which, if disclosed without precaution, may infringe upon ethical agreements and threaten the ownership, representation, and existence of vulnerable communities, protected lands and species.

Not always the case…



Los Angeles Times

WORLD & NATION

ICE accidentally released the identities of 6,252 immigrants who sought protection in the U.S.

A person receives a scan from the NeoScan 45 fingerprint scanner. The device, paired with an app known as EDDIE, is used by ICE to run remote ID checks. (Immigration and Customs Enforcement via Associated Press)

SUBSCRIBERS ARE READING ›

FOOD
FOR SUBSCRIBERS
These are the 101 best restaurants in Los Angeles

ENTERTAINMENT & ARTS
FOR SUBSCRIBERS
Inside the blockbuster lawsuit threatening one teen YouTube star's multimillion-dollar empire

CALIFORNIA
At least 2 dead, 11 injured after 6.4 earthquake in Northern California

TELEVISION
The note that Stephen 'Twitch' Boss left led investigators to rule death a suicide

Source: https://www.latimes.com/california/story/2022-11-30/ice-released-names-6252-immigrants-persecution

Microdata

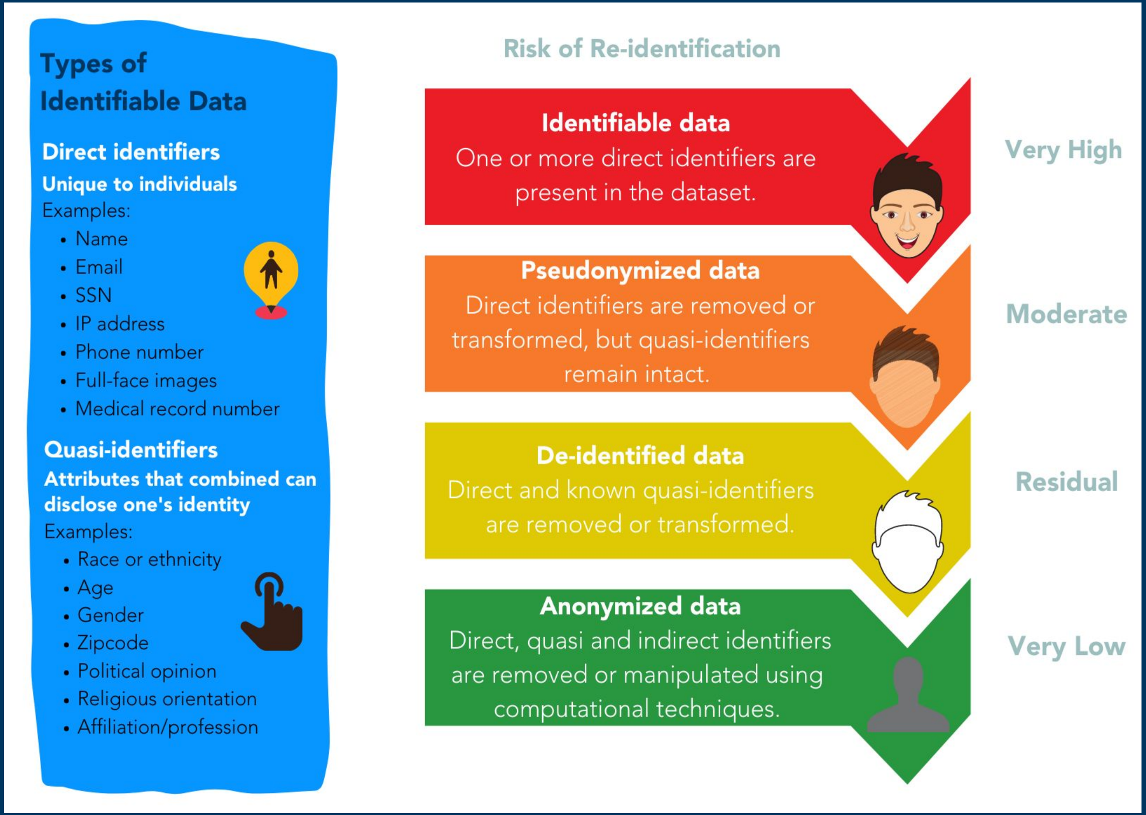
Unit-level data obtained from sample surveys, censuses, and administrative systems.

They provide information about characteristics of individual people or entities such as households, business enterprises, facilities, farms or even geographical areas such as villages or towns.

Data De-identification

The process which removes direct and indirect identifiers from data and mitigates privacy risks, while allowing data to be shared and reused.

Direct and Indirect Identifiers



Types of Identifiable Data

Direct identifiers

Unique to individuals

Examples:

- Name
- Email
- SSN
- IP address
- Phone number
- Full-face images
- Medical record number

Quasi-identifiers

Attributes that combined can disclose one's identity

Examples:

- Race or ethnicity
- Age
- Gender
- Zipcode
- Political opinion
- Religious orientation
- Affiliation/profession

Risk of Re-identification

Identifiable data One or more direct identifiers are present in the dataset. — Very High

Pseudonymized data Direct identifiers are removed or transformed, but quasi-identifiers remain intact. — Moderate

De-identified data Direct and known quasi-identifiers are removed or transformed. — Residual

Anonymized data Direct, quasi and indirect identifiers are removed or manipulated using computational techniques. — Very Low

Source: https://www.library.ucsb.edu/sites/default/files/dls_n3_dataprivacy_navy_0.pdf

An spectrum throughout the project lifecycle!

See: https://osf.io/7fpmw

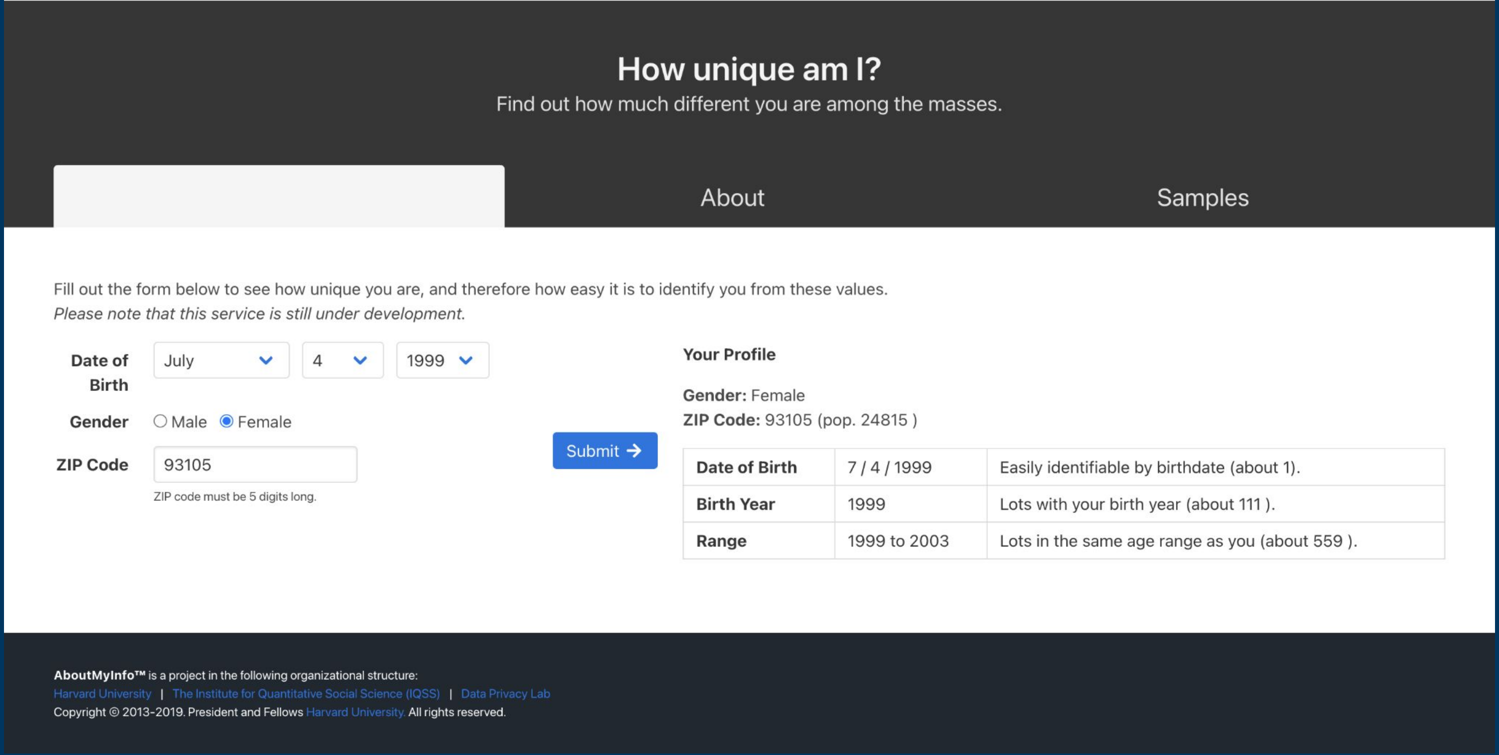
HIPAA (identifiers)
Safe Harbor Methods

- Name
- Address (all geographic subdivisions smaller than state)
- All elements (except years) of dates (e.g., birthdate, admission date, discharge date, date of death)
- Telephone numbers
- Fax numbers
- Email address
- Social Security Number
- Medical record number
- Health plan beneficiary number
- Account number
- Certificate or license number
- Vehicle identifiers
- Device identifiers and serial numbers
- Web URL
- Internet Protocol (IP) Address
- Finger or voice print
- Photographic image
- Any other characteristics that could uniquely identify the individuals

Recommended reading: https://doi.org/10.1136/bmj.c181

NOT as anonymous as you think!

Risk of Re-identification



How unique am I?

Find out how much different you are among the masses.

Fill out the form below to see how unique you are, and therefore how easy it is to identify you from these values.
Please note that this service is still under development.

Date of Birth	July / 4 / 1999
Gender	○ Male ● Female
ZIP Code	93105

ZIP code must be 5 digits long.

Your Profile

Gender: Female
ZIP Code: 93105 (pop. 24815)

Date of Birth	7 / 4 / 1999	Easily identifiable by birthdate (about 1).
Birth Year	1999	Lots with your birth year (about 111).
Range	1999 to 2003	Lots in the same age range as you (about 559).

AboutMyInfo™ is a project in the following organizational structure:
Harvard University | The Institute for Quantitative Social Science (IQSS) | Data Privacy Lab

https://aboutmyinfo.org/identity

Statistical Disclosure Control

SDC is a method for risk estimation and adjustment considering the utility of the data, having responsible data sharing in mind.

This is essentially a 3-steps process:

1. Assessing the risk of re-identification
2. Reducing the risk of re-identification
3. Quantifying information loss

A Direct Proportional Relationship

Disclosure Risk $\propto$ Data Utility

- **Non-perturbative methods**: no distortion to the data structure
- **Perturbative methods**: creates uncertainty around the true values

Techniques to Mitigate Re-id [1]

- Aggregation
- Top-coding
- Collapsing or combining variables
- Bracketing/categorization

Techniques to Mitigate Re-id [2]

- Redaction/Suppression
- Swapping/Shuffling
- Pseudonymization/Tokenization/Hashing
- Noising or disturbing

Other Considerations

- Be cautious when using small subgroups or small areas
- Avoid listings of cases with outliers
- Consider using weighted data to generate outputs
- Avoid submitting tables with small cell sizes (i.e., cells with fewer than 5 respondents)
- Restrict cross-tabular analysis to two or three dimensions

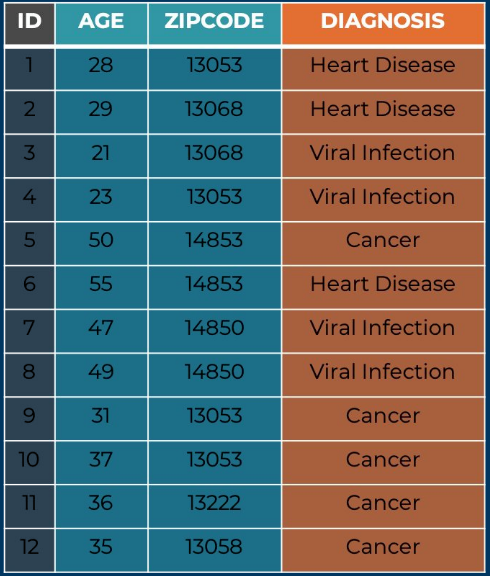
K-Anonymity

RISK ASSESSMENT AND MITIGATION

“Hiding in the crowd”

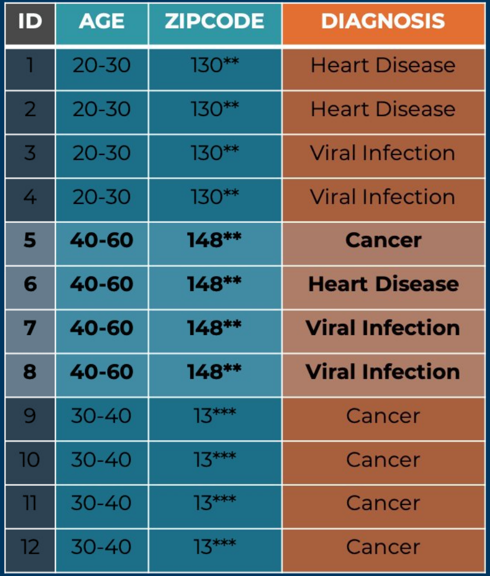
At least *k* individuals in the dataset who share the set of attributes that might become identifying for each individual.

*3 to 5 / 11-20 “matching cases” are desired depending on access permissions



ID	AGE	ZIPCODE	DIAGNOSIS
1	28	13053	Heart Disease
2	29	13068	Heart Disease
3	21	13068	Viral Infection
4	23	13053	Viral Infection
5	50	14853	Cancer
6	55	14853	Heart Disease
7	47	14850	Viral Infection
8	49	14850	Viral Infection
9	31	13053	Cancer
10	37	13053	Cancer
11	36	13222	Cancer
12	35	13058	Cancer

K-anonymization →



ID	AGE	ZIPCODE	DIAGNOSIS
1	20-30	130**	Heart Disease
2	20-30	130**	Heart Disease
3	20-30	130**	Viral Infection
4	20-30	130**	Viral Infection
5	**40-60**	**148****	**Cancer**
6	**40-60**	**148****	**Heart Disease**
7	**40-60**	**148****	**Viral Infection**
8	**40-60**	**148****	**Viral Infection**
9	30-40	13***	Cancer
10	30-40	13***	Cancer
11	30-40	13***	Cancer
12	30-40	13***	Cancer

Suppression + Global recoding/generalization

UC SANTA BARBARA

Restricting Access

as open as possible, as closed as necessary

- Data Use Agreements (DUA)
- Data Enclaves
- Limited sharing
  - Subset of the data
  - Metadata Only

Tools

Many existing tools, supporting different de-identification techniques/methods.

sdcMicro package/sdcApp (GUI)

- R-based (open and free)
- Robust and widely used/cited
- Reproducibility!

De-id tools:
https://dataservices.library.jhu.edu/resources/applications-to-assist-in-de-identification-of-human-subjects-research-data/
https://amnesia.openaire.eu/download.html

sdcMicro/sdcApp

Method	Classification	Data Type
Global recoding	non-perturbative, deterministic	continuous and categorical
Top and bottom coding	non-perturbative, deterministic	continuous and categorical
Local suppression	non-perturbative, deterministic	categorical
PRAM	perturbative, probabilistic	categorical
Micro aggregation	perturbative, probabilistic	continuous
Noise addition	perturbative, probabilistic	continuous
Shuffling	perturbative, probabilistic	continuous
Rank swapping	perturbative, probabilistic	continuous

Let's practice!

Mayor McDaniels and Peter Charles (aka PC Principal) need our help!

Survey with 100 students from South Park Elementary School:

1. What is the level of disclosure risk of this dataset?
2. How the risk of re-identification can be considerably reduced?
3. What would be the utility loss after implementing these strategies?