March 2017

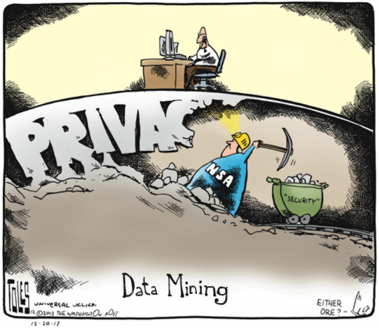
Data anonymisation and privacy

Raul Santos-Rodriguez

Applied Data Science

In this lecture we will discuss different aspects of data sharing and privacy, including:

- What is privacy?
- When data releases go wrong
- Methods and practices for tabular data and databases
- Software tools



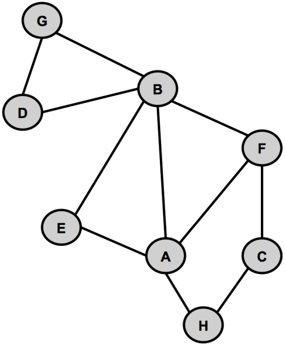
What is privacy?

’If you want to keep a secret, you must also hide it from yourself’

A game

How many people can you identify if Jack and Lucy’s list of contacts is leaked?

Class list:
- Alice
- Dan
- Jack
- Kate
- Lucy
- Poppy
- Tom
- Zara



Jack:
- Dan
- Kate
- Lucy
- Tom
- Zara

Lucy:
- Alice
- Zara
- Jack

https://bits.blogs.nytimes.com/2013/07/10/a-game-that-deals-in-personal-data/?_r=0

Aol.

On August 4, 2006, AOL Research, released a text file on one of its websites containing twenty million search keywords for over 650,000 users over a 3-month period intended for research purposes.

AOL did not identify users in the report ...
... personally identifiable information was present in many of the queries!

https://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/

Netflix Prize

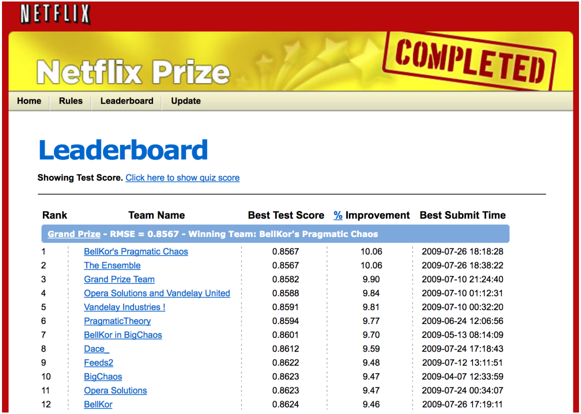
In 2007, Netflix offered a $1 million prize for a 10% improvement in its recommendation system.

Netflix released a training dataset for the developers to train their systems.

$$< user, movie, date, grade >$$

'To protect customer privacy, all personal information identifying individual customers has been removed and all customer ids have been replaced by randomly assigned ids.'

Netflix Prize



Leaderboard

Showing Test Score. Click here to show quiz score

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace_	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11

http://www.netflixprize.com/leaderboard.html

Netflix Prize

But ...

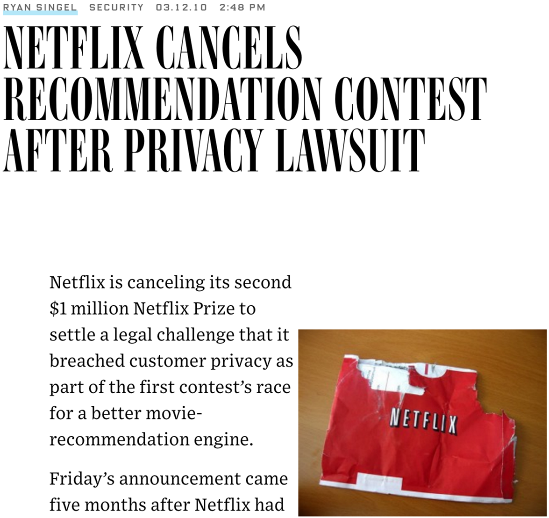
... Netflix is not the only movie-rating portal on the web; e.g., IMDb.

And ...

... on IMDb individuals can register and rate movies and they have the option of **not keeping their details anonymous!**

A. Narayanan and V. Shmatikov linked the Netflix anonymised training database with the IMDb database (using the date of rating by a user) to partially de-anonymize the Netflix training database, compromising the identity of some users.

`http://www.stoweboyd.com/post/882278313/the-limits-of-anonymity-the-netflix-prize-undone`



RYAN SINGEL SECURITY 03.12.10 2:48 PM

NETFLIX CANCELS RECOMMENDATION CONTEST AFTER PRIVACY LAWSUIT

Netflix is canceling its second $1 million Netflix Prize to settle a legal challenge that it breached customer privacy as part of the first contest's race for a better movie-recommendation engine.

Friday's announcement came five months after Netflix had

https://www.wired.com/2010/03/netflix-cancels-contest/

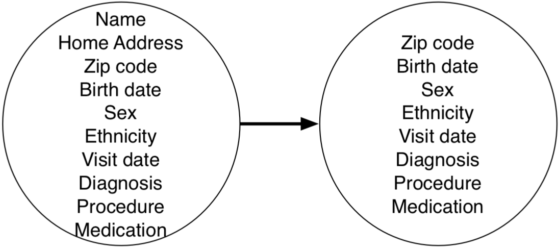
GIC medical encounter database

Massachusetts Group Insurance Commission (GIC) medical encounter database

”Anonymised” data on state employees that showed every single **hospital visit**, removing all obvious identifiers such as name, address, and Social Security number.

At the time GIC released the data, William Weld, then Governor of Massachusetts, assured the public that GIC had protected patient privacy by deleting identifiers. In response, then-graduate student Latanya Sweeney started hunting for the Governor’s hospital records in the GIC data. She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 residents and seven ZIP codes. For twenty dollars, she purchased the complete voter rolls from the city of Cambridge, a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter. By combining this data with the GIC records, Sweeney found Governor Weld with ease. Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code. In a theatrical flourish, Dr. Sweeney sent the Governor’s health records (which included diagnoses and prescriptions) to his office.

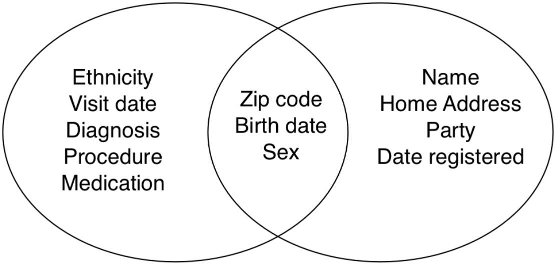
GIC medical encounter database



"Anonymization"

http://www.bitbybitbook.com/en/ethics/dilemmas/info-risk/

GIC medical encounter database



Ethnicity
Visit date
Diagnosis
Procedure
Medication

Zip code
Birth date
Sex

Name
Home Address
Party
Date registered

"Anonymized"
medical records

Voting records

http://www.bitbybitbook.com/en/ethics/dilemmas/info-risk/

Statistical databases

Statistical databases contain **statistical** information

- Healthcare organisations (epidemiology)
- Private organisations (consumer surveys)

Formats

Tabular data tables with counts or magnitudes.

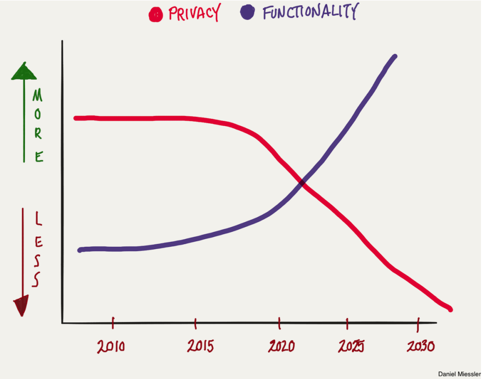
Queryable databases on-line databases which accept statistical queries (sum, average, max, min).

Microdata files where each record contains information on an individual.

The privacy trade-off

Statistical databases must provide useful statistical information ...

... and must also preserve the privacy of respondents



Daniel Miessler

Identifiers and quasi-identifiers

Identifiers

Attributes that unambiguously identify the respondent (e.g. passport no., social security no., name-surname, etc.).

Quasi-identifiers

They identify the respondent with some ambiguity, but their combination may lead to unambiguous identification (e.g. address, gender, age, telephone no., etc.).

Confidential vs non-confidential

Confidential attributes contain sensitive respondent information (e.g. salary, religion, diagnosis, etc.). **Non-confidential** attributes contain non-sensitive respondent info.

Identifiers and quasi-identifiers

Identifiers → suppressed in anonymised data sets.

Disclosure risk comes from **quasi-identifiers**:

- QIs cannot be suppressed because they often have high analytical value.
- QIs can be used to link anonymised records to external non-anonymous databases → re-identification!!!

Disclosure types

Attribute disclosure the value of a confidential attribute of an individual can be determined more accurately with access to the released statistics than without.

Identity disclosure a record in the anonymised data set can be linked with a respondent’s identity.

Membership disclosure whether or not data about an individual is contained in a dataset.

Attacks in tables

External attack

Let a table $Ethnicity$ x $Town$ contain a single respondent for ethnicity E_i and town T_j . If a table is released with the blood pressure for each ethnicity and town, the exact blood pressure of the respondent with ethnicity E_i in town T_j is publicly disclosed.

Internal attack

If there are only two respondents for ethnicity E_i and town T_j , the blood pressure of each of them is disclosed to the other.

Dominance attack

If one (or few) respondents dominate in the contribution to a cell in a magnitude table, the dominant respondent can upper-bound the contributions of the rest.

Techniques for tables

Non-perturbative do not modify the values in the cells, but they may suppress or recode them: **cell suppression**, **recoding of categorical attributes**

Perturbative modify the values in the cells: **controlled rounding**

Techniques for databases

Query perturbation perturbation can be applied to records on which queries are computed (input perturbation) or to the query result after computing it on the original data (output perturbation).

Query restriction the database refuses to answer certain queries.

Camouflage deterministically correct non-exact answers (small interval answers) are returned by the database.

Question

Suppose you have access to a database that allows you to compute the total income of all residents in a certain area. If you knew that Mr. Smith was going to move to another area, simply querying this database before and after his move would allow you to deduce his income.

What could you do to stop this?

Differential privacy

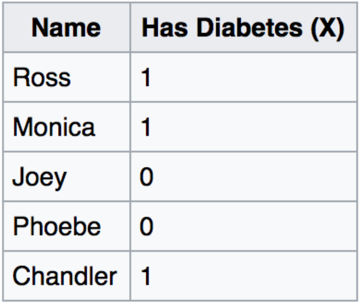
- A learner implements a summary statistic called **A**.
- Adversary proposes two datasets S and S' that differ by only one row or example, and a test set Q.

Differential privacy

A is called epsilon-differentially private iff

$$|\log(Prob(A(S) \in Q)/Prob(A(S') \in Q))| \leq \epsilon$$

- DP is a condition on the release mechanism and **not on the dataset** itself.
- Presence or absence of an individual will not affect the final output.



Name	Has Diabetes (X)
Ross	1
Monica	1
Joey	0
Phoebe	0
Chandler	1

User wants to find whether Chandler has diabetes or not.
User also knows in which row Chandler resides.
User is only allowed to use a query A_i that returns the partial sum of the first i rows of column.

$$A_5(S_1) - A_4(S_1)$$

Wikipedia

Social sciences *Do you own the attribute A?*

1. Throw a coin.
2. If head, then answer honestly.
3. If tail, then throw the coin again and answer ”Yes” if head, ”No” if tail.

Idea → add noise

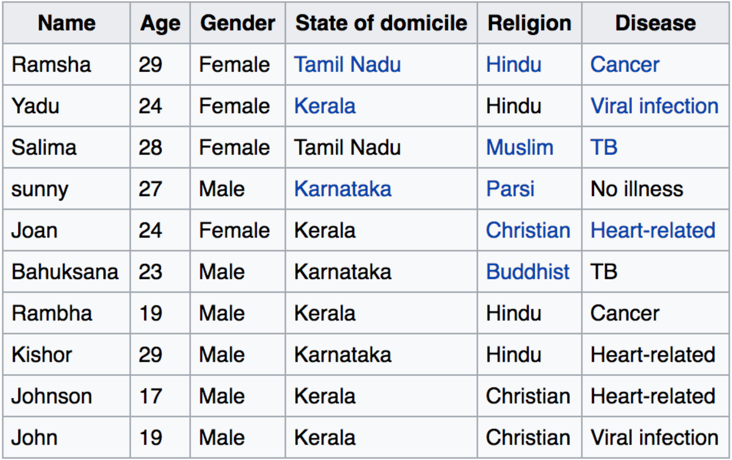
k-anonymity [Samarati & Sweeney, 1998]

A dataset is said to satisfy k-anonymity if each combination of values of the quasi-identifier attributes in it is shared by at least k records.

How to achieve k-anonymity

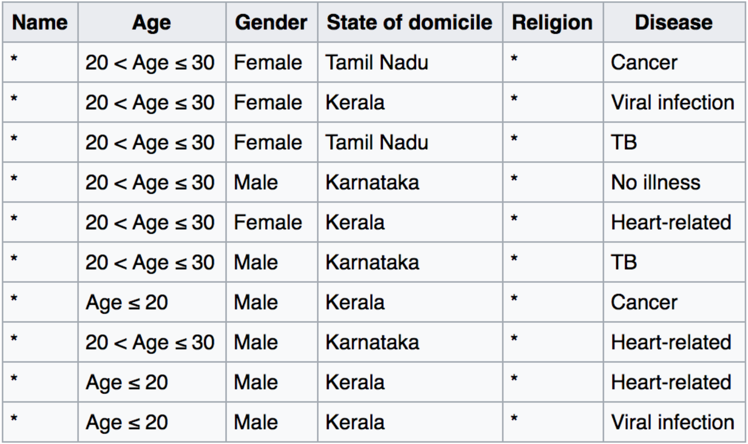
1. Suppression values of the attributes are replaced by an asterisk.

2. Generalization values of attributes are replaced by with a broader category.



Name	Age	Gender	State of domicile	Religion	Disease
Ramsha	29	Female	Tamil Nadu	Hindu	Cancer
Yadu	24	Female	Kerala	Hindu	Viral infection
Salima	28	Female	Tamil Nadu	Muslim	TB
sunny	27	Male	Karnataka	Parsi	No illness
Joan	24	Female	Kerala	Christian	Heart-related
Bahuksana	23	Male	Karnataka	Buddhist	TB
Rambha	19	Male	Kerala	Hindu	Cancer
Kishor	29	Male	Karnataka	Hindu	Heart-related
Johnson	17	Male	Kerala	Christian	Heart-related
John	19	Male	Kerala	Christian	Viral infection

Wikipedia



Name	Age	Gender	State of domicile	Religion	Disease
*	20 < Age ≤ 30	Female	Tamil Nadu	*	Cancer
*	20 < Age ≤ 30	Female	Kerala	*	Viral infection
*	20 < Age ≤ 30	Female	Tamil Nadu	*	TB
*	20 < Age ≤ 30	Male	Karnataka	*	No illness
*	20 < Age ≤ 30	Female	Kerala	*	Heart-related
*	20 < Age ≤ 30	Male	Karnataka	*	TB
*	Age ≤ 20	Male	Kerala	*	Cancer
*	20 < Age ≤ 30	Male	Karnataka	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Viral infection

Wikipedia

k-anonymity: attacks

k-anonymity does not protect against attribute disclosure in general ...

> ... if the **values of a confidential attribute are very similar** in a group of k records sharing quasi-identifier values!

Homogeneity Attack

When all the values for a sensitive value within a set of k records are identical, the sensitive value for the set of k records may be exactly predicted.

Background Knowledge Attack

Association between one or more quasi-identifier attributes with the sensitive attribute to reduce the set of possible values for the sensitive attribute.

Machanavajjhala, Kifer, Gehrke, and Venkitasubramaniam (2007): Japanese heart attack patients.

Extensions of k-anonymity

l-diversity

l-diversity requires that the values of all confidential attributes within a group of k records contain at least l clearly distinct values.

l-diversity: drawbacks

- Not every value shows equal sensitivity. A rare positive indicator for a disease may provide more information than a common negative indicator.

- While l-diversity ensures diversity of sensitive values in each group, it does not recognise that values may be semantically close.

  An attacker could deduce a stomach disease applies to an individual if a sample containing the individual only listed three different stomach diseases.

Extensions of k-anonymity

$t - closeness$

t-closeness requires that the distribution of the confidential attribute within a group of k records is similar to the distribution of the confidential attribute in the entire data set (at most distance t between both distributions).

Anonimisation tools

ARX k-anonymity, l-diversity, t-closeness implementation in Java.

`http://arx.deidentifier.org`

Argus software designed to create safe microdata files.

`http://neon.vb.cbs.nl/casc`

SdcMicro Disclosure control methods for anonymisation and risk estimation.

`http://cran.r-project.org/package=sdcMicro`

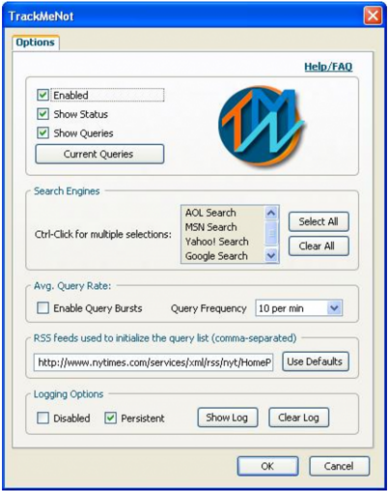
`http://arx.deidentifier.org/overview/related-software/`

Other approaches to data privacy

PPDM Privacy-preserving data mining seeks the data owner's privacy when several owners wish to co-operate without giving away their data to each other.

PIR Private information retrieval seeks user privacy to allow the user of a database to retrieve some information item without the database knowing which item was recovered.

TrackMeNot



☑ Enabled
☑ Show Status
☑ Show Queries

Search Engines

Ctrl-Click for multiple selections: AOL Search, MSN Search, Yahoo! Search, Google Search

Avg. Query Rate:

☐ Enable Query Bursts Query Frequency 10 per min

RSS feeds used to initialize the query list (comma-separated)

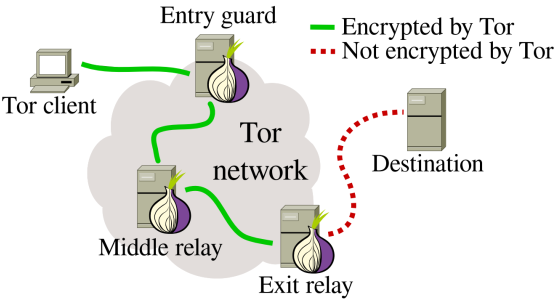
http://www.nytimes.com/services/xml/rss/nyt/HomeP

Logging Options

☐ Disabled ☑ Persistent

https://cs.nyu.edu/trackmenot/

Tor



https://www.torproject.org/

Data protection plan

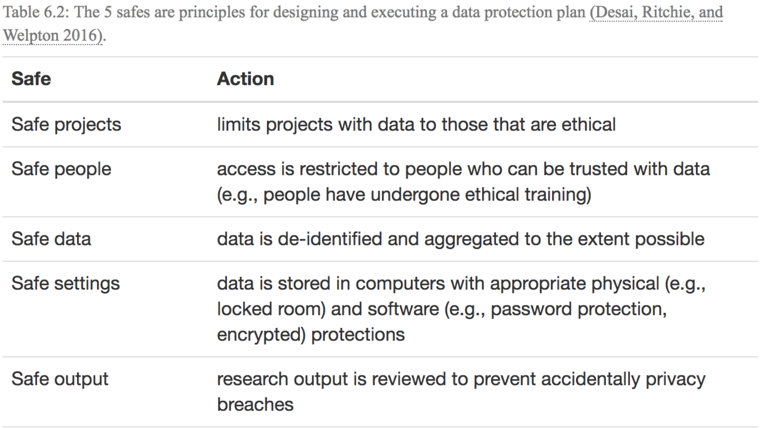


Table 6.2: The 5 safes are principles for designing and executing a data protection plan (Desai, Ritchie, and Welpton 2016).

Safe	Action
Safe projects	limits projects with data to those that are ethical
Safe people	access is restricted to people who can be trusted with data (e.g., people have undergone ethical training)
Safe data	data is de-identified and aggregated to the extent possible
Safe settings	data is stored in computers with appropriate physical (e.g., locked room) and software (e.g., password protection, encrypted) protections
Safe output	research output is reviewed to prevent accidentally privacy breaches

Next time we will discuss the deployment of data science systems!