Privacy in Data Publication and Release

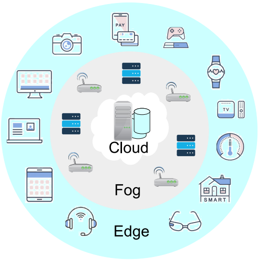
Pierangela Samarati
Dipartimento di Informatica
Università degli Studi di Milano
pierangela.samarati@unimi.it

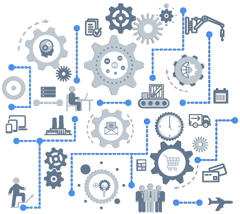
Summer School Cyber In Normandy

Caen, France – July 1, 2024

ICT ecosystem

- Advancements in the ICT and networks have changed our society
- 5G and beyond, infrastructures and services are more powerful, efficient, and complex





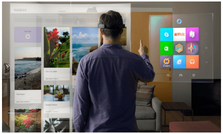
- ICT and network advancements are enabling factors for a smart society ...

. . . Everything is getting smart


Smart car


Museum and exhibitions


Health Care


Augmented reality


Smart e-commerce


Intelligent shops


Smart entertainment systems


Smart governance


Smart toothbrush

Smart society

Smart society - Advantages

Smart services and security – Advantages

+ Better protection mechanisms
+ Business continuity and disaster recovery
+ Prevention and response

. . . but . . .

Smart services and security – Disadvantages

- More complexity …

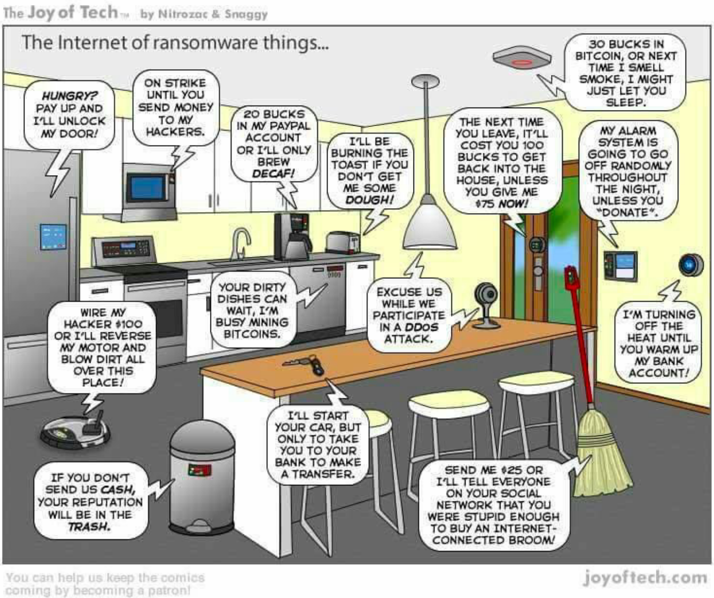
  … weakest link becomes a point of attack

  - system hacking
  - improper information leakage
  - data and process tampering

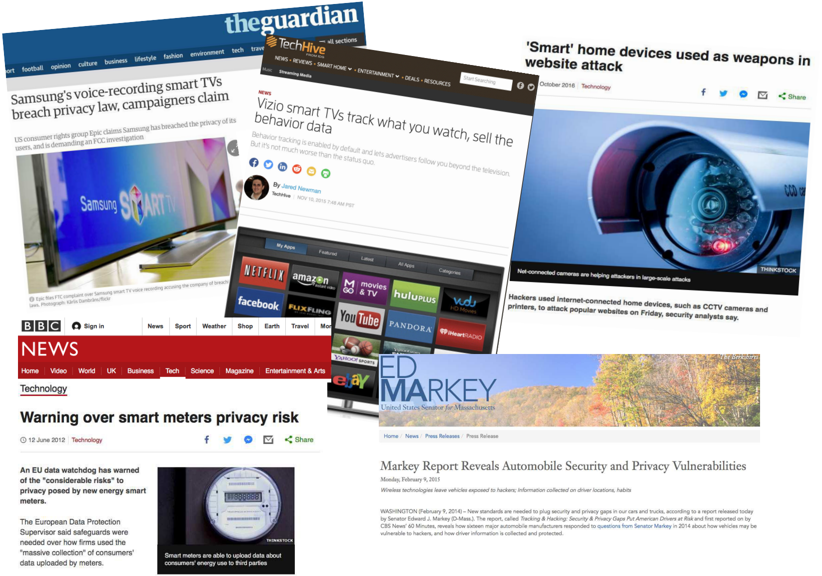
- Explosion of damages and violations

- Loss of control over data and processes

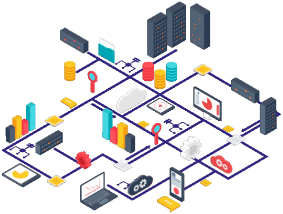
Maybe too smart? – 1



Maybe too smart? – 2



Security . . . a complex problem



Protection of infrastructure

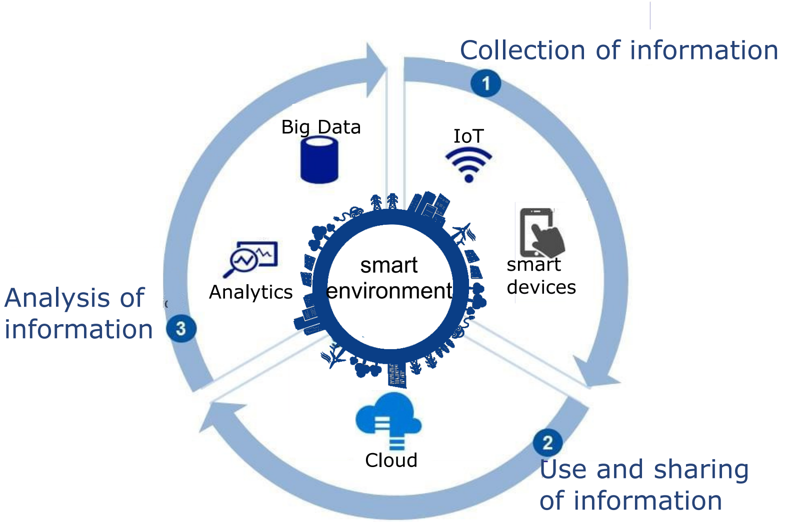
Protection of communication

Protection against malware and attacks

Protection of devices

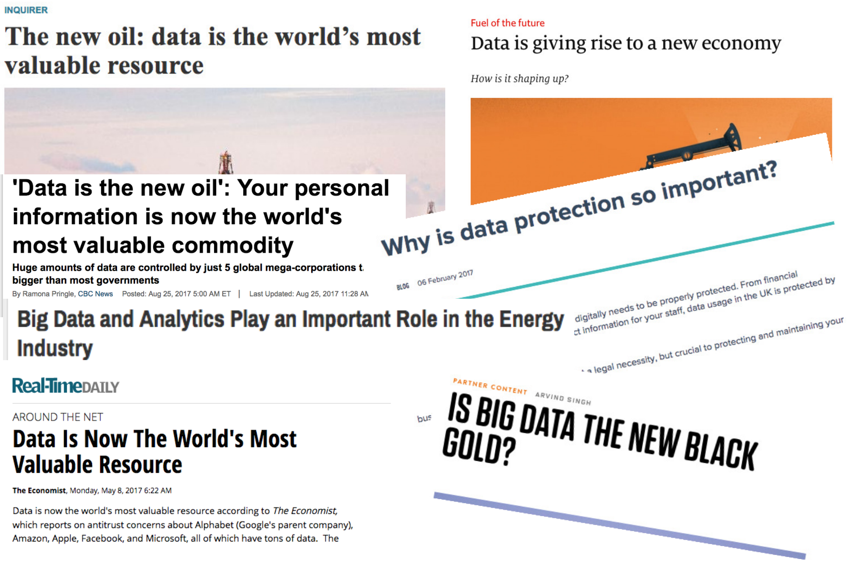
Protection of data

The role of data in a smart environment



$\Longrightarrow$ better governance and intelligent systems

The most valuable resource - Data



INQUIRER

The new oil: data is the world's most valuable resource

'Data is the new oil': Your personal information is now the world's most valuable commodity

Huge amounts of data are controlled by just 5 global mega-corporations t
bigger than most governments

By Ramona Pringle, CBC News | Posted: Aug 25, 2017 5:00 AM ET | Last Updated: Aug 25, 2017 11:28 AM

Fuel of the future

Data is giving rise to a new economy

How is it shaping up?

Why is data protection so important?

BLOG 06 February 2017

digitally needs to be properly protected. From financial
ct information for your staff, data usage in the UK is protected by

a legal necessity, but crucial to protecting and maintaining your
bus

Big Data and Analytics Play an Important Role in the Energy Industry

Real-Time DAILY

AROUND THE NET

Data Is Now The World's Most Valuable Resource

The Economist, Monday, May 8, 2017 6:22 AM

Data is now the world's most valuable resource according to *The Economist*, which reports on antitrust concerns about Alphabet (Google's parent company), Amazon, Apple, Facebook, and Microsoft, all of which have tons of data. The

PARTNER CONTENT ARVIND SINGH

IS BIG DATA THE NEW BLACK GOLD?

Impact on data protection and privacy

Uber reveals 2.7 million UK users of its app were affected by a mass data breach that saw names, emails and phone numbers stolen

- Uber has revealed 2.7m UK users of its app were affected by a 2016 data brea
- The taxi-hailing firm then tried to cover up the breach for more than a year
- It was also found Uber had paid two hackers £75,000 to delete the data

By TH
PUBLI

Computer Scientists Develop a Simple Tool to Tell If Websites Suffered a Data Breach

Published: December 12, 2017.

Uber says data breach compromised 380K users in Singapore

Ride-sharing company reveals 380,000 in Singapore were affected by the massive data breach that compromised 57 million accounts globally, but says no fraud or misuse has been tied to these users.

By Eileen Yu for By The Way | December 16, 2017 -- 10:07 GMT (10:07 GMT) | Topic: Security

The Dutch Data Protection Authority accidentally leaked its employees' data

by MIX — 4 weeks

Approx. 9,000 Penn students affected by security breach that released their private information

By Kelly Heinzerling | 03/12/18 6:50pm

1 DEC 2017 NEWS
Over 100GB of Secret Consumer Credit Data Leaked Online

A collection of 1.4 Billion Plain-Text leaked credentials is available online

December 12, 2017 By Pierluigi Paganini

My Page | Like 2 | G+

NEWS
63,500 records breached by misconfigured database

by Jessica Davis | April 12, 2018

SECURITY
MASSIVE
Personal Data of Over 143 Million Americans Stolen from a Credit Reporting Firm

18 SHARES | SHARE | TWEET | SUBMIT

MyFitnessPal breach affects millions of Under Armour users

19 DEC 2017 NEWS
Californian Voters Suffer Major Data Breach

A 41-gigabyte archive containing **1.4 Billion** credentials in clear text was found in dark web, it had been updated at the end of Novembe

LOCAL
Photo: Jimmy Bernhard, KSDK-TV

Former nursing home employee admits stealing residents' credit card numbers

Shaniece Borney, 29, will be forced to pay the victims back and could face an additional $250,000 fine, 10 years in prison or both.

Mar 01 2018
Equifax discovers another 2.4 million customers hit by data breach

Posted by Dissent at 11:02 am | Business Sector, Hack, U.S.

Deloitte hit by cyber-attack revealing clients' secret emails

Exclusive: hackers may have accessed usernames, passwords and personal details of top accountancy firm's blue-chip clients

NEWS

Facebook admits to far higher number of data breaches

Facebook has said personal data on 87 million users was shared with Cambridge Analytica, millions more than it admitted earlier. The social media giant also unveiled new privacy rules, but the whiff of scandal lingers.

Privacy
Carphone Warehouse Breach: 'Striking' Failures Trigger Fine

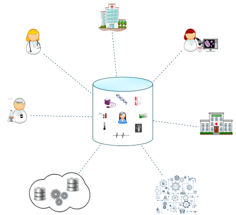
Mathew J. Schwartz • January 10, 2018

Mobile phone retailer Carphone Warehouse has been hit with one of the largest fines ever imposed by Britain's data privacy watchdog

Outline

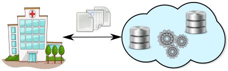
- Privacy in data publication

  $\Longrightarrow$ data release/dissemination



- Privacy in data outsourcing

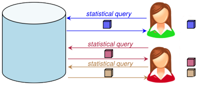
  $\Longrightarrow$ third parties collect, store, process, and manage data



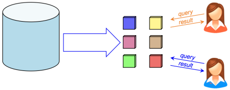
Privacy in Data Publication

Data sharing/publication

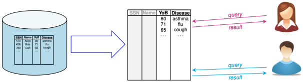
- Statistical DBMS
  - the DBMS responds only to statistical queries (e.g., avg, sum, count, ...)
  - need run-time checking to control information (indirectly) released



- Statistical data (macrodata)
  - publish statistics (e.g., count/frequency or magnitude tables)
  - control on indirect release performed before publication



- Microdata: individual records are released



Information disclosure

Need to protect privacy, i.e., ensure no improper:

- identity disclosure: record in a protected dataset can be linked with a respondent's identity
- attribute disclosure: the value of a confidential attribute of a respondent can be determined or closely estimated with some confidence

The Anonymity Problem

Anonymization

- Datasets truly anonymized are not subject to privacy regulations

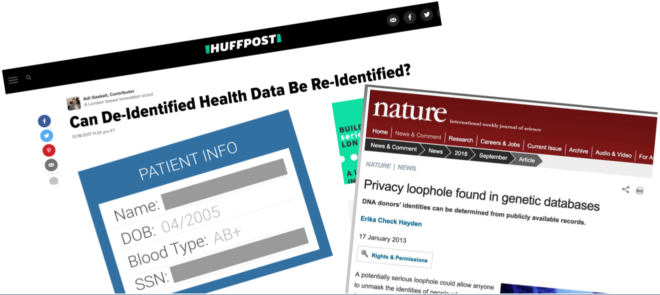
Anonymization

- Datasets truly anonymized are not subject to privacy regulations

> *The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.*
>
> *-EU GDPR, Recital 26*

Anonymization is a complex problem . . .

- Anonymization $\neq$ de-identification
- Correlation among different data sources
- Indirect exposure of sensitive information
- Even pseudonyms can expose users



The anonymity problem

- The amount of privately owned records that describe each citizen's finances, interests, and demographics is increasing every day
- These data are de-identified before release, that is, any explicit identifier (e.g., SSN) is removed
- De-identification is not sufficient
- Most municipalities sell population registers that include the identities of individuals along with basic demographics
- These data can then be used for linking identities with de-identified information $\Longrightarrow$re-identification

The anonymity problem – Example

SSN	Name	Race	DoB	Sex	ZIP	Marital status	Disease
		asian	64/04/12	F	94142	divorced	hypertension
		asian	64/09/13	F	94141	divorced	obesity
		asian	64/04/15	F	94139	married	chest pain
		asian	63/03/13	M	94139	married	obesity
		asian	63/03/18	M	94139	married	short breath
		black	64/09/27	F	94138	single	short breath
		black	64/09/27	F	94139	single	obesity
		white	64/09/27	F	94139	single	chest pain
		white	64/09/27	F	94141	widow	short breath

The anonymity problem – Example

SSN	Name	Race	DoB	Sex	ZIP	Marital status	Disease
		asian	64/04/12	F	94142	divorced	hypertension
		asian	64/09/13	F	94141	divorced	obesity
		asian	64/04/15	F	94139	married	chest pain
		asian	63/03/13	M	94139	married	obesity
		asian	63/03/18	M	94139	married	short breath
		black	64/09/27	F	94138	single	short breath
		black	64/09/27	F	94139	single	obesity
		white	64/09/27	F	94139	single	chest pain
		white	64/09/27	F	94141	widow	short breath

Name	Address	City	ZIP	DOB	Sex	Status
....	....	....	....	....	....	....
....	....	....	....	....	....	....
Sue J. Doe	900 Market St.	San Francisco	94142	64/04/12	F	divorced
....	....	....	....	....	....	....

The anonymity problem – Example

SSN	Name	Race	DoB	Sex	ZIP	Marital status	Disease
		asian	64/04/12	F	94142	divorced	hypertension
		asian	64/09/13	F	94141	divorced	obesity
		asian	64/04/15	F	94139	married	chest pain
		asian	63/03/13	M	94139	married	obesity
		asian	63/03/18	M	94139	married	short breath
		black	64/09/27	F	94138	single	short breath
		black	64/09/27	F	94139	single	obesity
		white	64/09/27	F	94139	single	chest pain
		white	64/09/27	F	94141	widow	short breath

Name	Address	City	ZIP	DOB	Sex	Status
....	....	....	....	....	....	....
....	....	....	....	....	....	....
Sue J. Doe	900 Market St.	San Francisco	94142	64/04/12	F	divorced
....	....	....	....	....	....	....

The anonymity problem – Example

SSN	Name	Race	DoB	Sex	ZIP	Marital status	Disease
	Sue J. Doe	asian	64/04/12	F	94142	divorced	hypertension
		asian	64/09/13	F	94141	divorced	obesity
		asian	64/04/15	F	94139	married	chest pain
		asian	63/03/13	M	94139	married	obesity
		asian	63/03/18	M	94139	married	short breath
		black	64/09/27	F	94138	single	short breath
		black	64/09/27	F	94139	single	obesity
		white	64/09/27	F	94139	single	chest pain
		white	64/09/27	F	94141	widow	short breath

Name	Address	City	ZIP	DOB	Sex	Status
....	....	....	....	....	....	....
....	....	....	....	....	....	....
Sue J. Doe	900 Market St.	San Francisco	94142	64/04/12	F	divorced
....	....	....	....	....	....	....

The anonymity problem – Example

SSN	Name	Race	DoB	Sex	ZIP	Marital status	Disease
	Sue J. Doe	asian	64/04/12	F	94142	divorced	hypertension
		asian	64/09/13	F	94141	divorced	obesity
		asian	64/04/15	F	94139	married	chest pain
		asian	63/03/13	M	94139	married	obesity
		asian	63/03/18	M	94139	married	short breath
		black	64/09/27	F	94138	single	short breath
		black	64/09/27	F	94139	single	obesity
		white	64/09/27	F	94139	single	chest pain
		white	64/09/27	F	94141	widow	short breath

Name	Address	City	ZIP	DOB	Sex	Status
....	....	....	....	....	....	....
....	....	....	....	....	....	....
Sue J. Doe	900 Market St.	San Francisco	94142	64/04/12	F	divorced
....	....	....	....	....	....	....

Classification of attributes in a microdata table

The attributes in the original microdata table can be classified as:

- identifiers: attributes that uniquely identify a microdata respondent (e.g., SSN uniquely identifies the person with which is associated)
- quasi-identifiers: attributes that, in combination, can be linked with external information to reidentify all or some of the respondents to whom information refers or reduce the uncertainty over their identities (e.g., DoB, Sex, and ZIP)
- confidential: attributes of the microdata table that contain sensitive information (e.g., Disease)
- non confidential: attributes that the respondents do not consider sensitive and whose release does not cause disclosure

Re-identification

A study of the 2000 census data reported that the US population was uniquely identifiable by:

- gender, year of birth, 5-digit ZIP code: 0.2%
- gender, year of birth, county: 0.0%
- gender, year and month of birth, 5-digit ZIP code: 4.2%
- gender, year and month of birth, county: 0.2%
- gender, year, month, and day of birth, 5-digit ZIP code: 63.3%
- gender, year, month, and day of birth, county: 14.8%

Disclosure risk

Factors contributing to increase the disclosure risk:

- existence of high visibility records (i.e., rare jobs or incomes)
- possibility of matching the microdata table with external sources

Factors contributing to decrease the disclosure risk:

- a microdata table often contains a subset of the whole population
- information in the microdata table or in the external sources may be not up-to-date
- information in the microdata table or in external sources may contain errors/noise

Measures of disclosure risk

Disclosure risk depends on several factors:

- the target respondent is represented in both the microdata table and some external source
- the matching variables are recorded in a linkable way in the microdata table and in the external source
- the respondent is unique (or peculiar) in the population of the external source

Each population unique is a sample unique; the vice-versa is not true

Some microdata protection approaches

- k-anonymity: protects identity of respondents by confusing it in a set of at least k respondents
- ℓ-diversity: builds on k-anonymity adding condition that every computed group of respondents be associated with at least ℓ diverse occurrences of sensitive attributes
- t-closeness: builds on k-anonymity adding condition that distribution of sensitive attributes in every computed group of respondents be close to the one to be expected
- differential privacy: no respondent should make a difference on the result (adds noise to data)
- ...

k-Anonymity

k-anonymity – 1

- k-anonymity, together with its enforcement via generalization and suppression, aims to protect respondents' identities while releasing truthful information

- k-anonymity tries to capture the following requirement:
  - the released data should be indistinguishably related to no less than a certain number of respondents

- Quasi-identifier: set of attributes that can be exploited for linking (whose release must be controlled)

k-anonymity – 2

- Basic idea: translate the k-anonymity requirement on the released data
  - each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k respondents
- In the released table the respondents must be indistinguishable (within a given set) with respect to quasi-identifying attributes
- k-anonymity requires that each quasi-identifier value appearing in the released table must have at least k occurrences
  - sufficient condition for the satisfaction of k-anonymity requirement

Generalization and suppression

- Generalization. The values of a given attribute are substituted by using more general values. Based on the definition of a generalization hierarchy
  - **Example**: consider attribute ZIP code and suppose that a step in the corresponding generalization hierarchy consists in suppressing the least significant digit in the ZIP code
    With one generalization step: 20222 and 20223 become 2022*; 20238 and 20239 become 2023*
- Suppression. Protect sensitive information by removing it
  - the introduction of suppression can reduce the amount of generalization necessary to satisfy the k-anonymity constraint

Generalized table with suppression – Example

Race	DOB	Sex	ZIP
asian	64/04/12	F	94142
asian	64/09/13	F	94141
asian	64/04/15	F	94139
asian	63/03/13	M	94139
asian	63/03/18	M	94139
black	64/09/27	F	94138
black	64/09/27	F	94139
white	64/09/27	F	94139
white	64/09/27	F	94141

PT

Race	DOB	Sex	ZIP
asian	64/04	F	941**
asian	64/09	F	941**
asian	64/04	F	941**
asian	63/03	M	941**
asian	63/03	M	941**
black	64/09	F	941**
black	64/09	F	941**
white	64/09	F	941**
white	64/09	F	941**

$GT_{[0,1,0,2]}$

Generalized table with suppression – Example

Race	DOB	Sex	ZIP
asian	64/04/12	F	94142
asian	64/09/13	F	94141
asian	64/04/15	F	94139
asian	63/03/13	M	94139
asian	63/03/18	M	94139
black	64/09/27	F	94138
black	64/09/27	F	94139
white	64/09/27	F	94139
white	64/09/27	F	94141

PT

Race	DOB	Sex	ZIP
asian	64/04	F	941**
asian	64/04	F	941**
asian	63/03	M	941**
asian	63/03	M	941**
black	64/09	F	941**
black	64/09	F	941**
white	64/09	F	941**
white	64/09	F	941**

$GT_{[0,1,0,2]}$

Achieving k-anonymity

- Need to balance generalization vs suppression
- Need to maintain utility: generalize/suppress as needed not more $\Longrightarrow$ minimal solution (do not overdo)
- Different preference criteria can be applied to choose among minimal solutions
- Different granularity of application (e.g., attribute vs cell)
- Different approaches to generalization (e.g., pre-defined generalization hierarchies or dynamically computed clustering)

Generalization vs suppression – Example

suppression			
Race	**DOB**	**Sex**	**ZIP**
asian	64/04	F	941**
asian	64/04	F	941**
asian	63/03	M	941**
asian	63/03	M	941**
black	64/09	F	941**
black	64/09	F	941**
white	64/09	F	941**
white	64/09	F	941**

$GT_{[0,1,0,2]}$

no suppression			
Race	**DOB**	**Sex**	**ZIP**
asian	64	F	941**
asian	64	F	941**
asian	64	F	941**
asian	63	M	941**
asian	63	M	941**
black	64	F	941**
black	64	F	941**
white	64	F	941**
white	64	F	941**

$GT_{[0,2,0,2]}$

Minimal generalization – Example

MaxSup=0 (no suppression) wished k=2

Race	**ZIP**
asian	94142
asian	94141
asian	94139
asian	94139
asian	94139
black	94138
black	94139
white	94139
white	94141

PT

Minimal generalization – Example

MaxSup=0 (no suppression) wished k=2

Race	ZIP
asian	94142
asian	94141
asian	94139
asian	94139
asian	94139
black	94138
black	94139
white	94139
white	94141

PT

Minimal generalization – Example

MaxSup=0 (no suppression) wished k=2

Race	ZIP
asian	94142
asian	94141
asian	94139
asian	94139
asian	94139
black	94138
black	94139
white	94139
white	94141

PT

Race	ZIP
person	9414*
person	9414*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9414*

$GT_{[1,1]}$

Minimal generalization – Example

MaxSup=0 (no suppression) wished k=2

Race	ZIP
asian	94142
asian	94141
asian	94139
asian	94139
asian	94139
black	94138
black	94139
white	94139
white	94141

PT

Race	ZIP
person	9414*
person	9414*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9414*

$GT_{[1,1]}$

Race	ZIP
asian	941**
asian	941**
asian	941**
asian	941**
asian	941**
black	941**
black	941**
white	941**
white	941**

$GT_{[0,2]}$

Minimal generalization – Example

MaxSup=0 (no suppression) wished k=2

Race	ZIP
asian	94142
asian	94141
asian	94139
asian	94139
asian	94139
black	94138
black	94139
white	94139
white	94141

PT

Race	ZIP
person	9414*
person	9414*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9414*

$GT_{[1,1]}$

Race	ZIP
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**

$GT_{[1,2]}$

Minimal generalization – Example

MaxSup=0 (no suppression) wished k=2

Race	ZIP
asian	94142
asian	94141
asian	94139
asian	94139
asian	94139
black	94138
black	94139
white	94139
white	94141

PT

Race	ZIP
person	9414*
person	9414*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9414*

$GT_{[1,1]}$

Race	ZIP
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**

$GT_{[1,2]}$

Preference criteria – Example

Which one to prefer?

Race	ZIP
asian	94142
asian	94141
asian	94139
asian	94139
asian	94139
black	94138
black	94139
white	94139
white	94141

PT

Race	ZIP
person	9414*
person	9414*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9414*

$GT_{[1,1]}$

Race	ZIP
asian	941**
asian	941**
asian	941**
asian	941**
asian	941**
black	941**
black	941**
white	941**
white	941**

$GT_{[0,2]}$

minimum distance (absolute/relative), maximum distribution, minimum suppression, greater utility for intended use, . . .

Granularity of application – Example

wished k=2

Race	ZIP
asian	94142
asian	94141
asian	94139
asian	94139
asian	94139
black	94138
black	94139
white	94139
white	94141

PT

Granularity of application – Example

wished k=2

Race	ZIP
asian	94142
asian	94141
asian	94139
asian	94139
asian	94139
black	94138
black	94139
white	94139
white	94141

PT

attribute

Race	ZIP
asian	941**
asian	941**
asian	941**
asian	941**
asian	941**
black	941**
black	941**
white	941**
white	941**

$\mathrm{GT}_{[0,2]}$

Granularity of application – Example

wished k=2

attribute, cell

Race	ZIP
asian	94142
asian	94141
asian	94139
asian	94139
asian	94139
black	94138
black	94139
white	94139
white	94141

PT

attribute

Race	ZIP
asian	941**
asian	941**
asian	941**
asian	941**
asian	941**
black	941**
black	941**
white	941**
white	941**

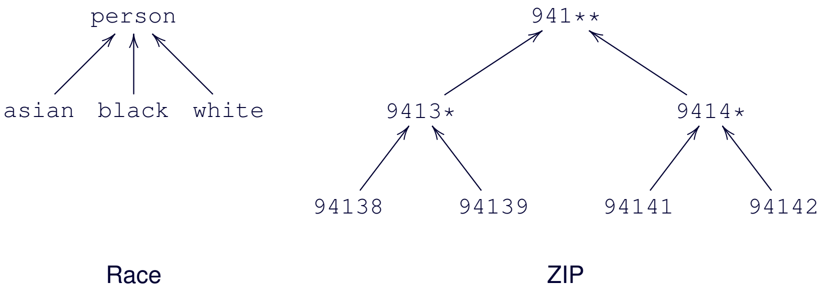
$GT_{[0,2]}$

cell

Race	ZIP
asian	9414*
asian	9414*
asian	94139
asian	94139
asian	94139
black	9413*
black	9413*
white	941**
white	941**

GT

Pre-defined vs dynamic clustering – Example



Pre-defined vs dynamic clustering – Example

wished k=3

Race	ZIP
asian	94142
asian	94141
asian	94139
asian	94139
asian	94139
black	94138
black	94139
white	94139
white	94141

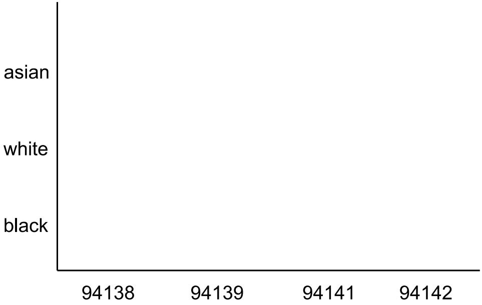
PT

Pre-defined vs dynamic clustering – Example

wished k=3

Race	ZIP
asian	94142
asian	94141
asian	94139
asian	94139
asian	94139
black	94138
black	94139
white	94139
white	94141

PT

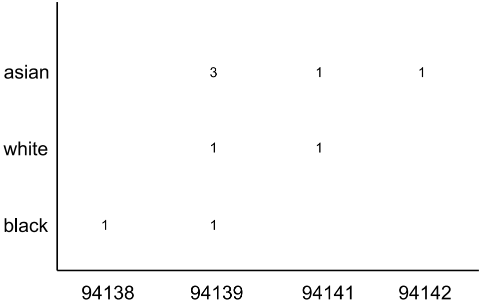


Pre-defined vs dynamic clustering – Example

wished k=3

Race	ZIP
asian	94142
asian	94141
asian	94139
asian	94139
asian	94139
black	94138
black	94139
white	94139
white	94141

PT

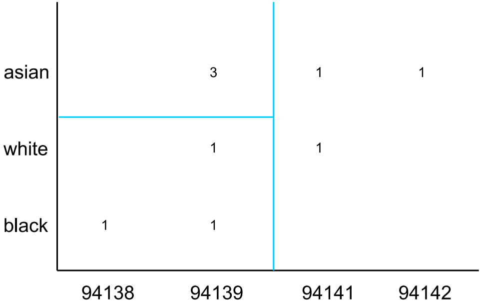


Pre-defined vs dynamic clustering – Example

wished k=3

Race	ZIP
asian	94142
asian	94141
asian	94139
asian	94139
asian	94139
black	94138
black	94139
white	94139
white	94141

PT

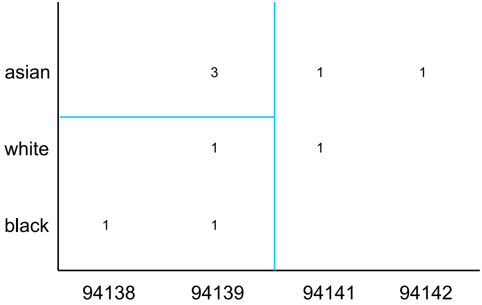


Pre-defined vs dynamic clustering – Example

wished k=3

Race	ZIP
asian or white	9414*
asian or white	9414*
asian	94139
asian	94139
asian	94139
black or white	9413*
black or white	9413*
black or white	9413*
asian or white	9414*

GT



Classification of k-anonymity techniques – 1

Generalization and suppression can be applied at different levels of granularity

- Generalization can be applied at the level of single column (i.e., a generalization step generalizes all the values in the column) or single cell (i.e., for a specific column, the table may contain values at different generalization levels)

- Suppression can be applied at the level of row (i.e., a suppression operation removes a whole tuple), column (i.e., a suppression operation obscures all the values of a column), or single cells (i.e., a k-anonymized table may wipe out only certain cells of a given tuple/attribute)

Classification of k-anonymity techniques – 2

Generalization	Suppression: *Tuple*	*Attribute*	*Cell*	*None*
Attribute	**AG_TS**	**AG_AS** $\equiv$ AG_	**AG_CS**	**AG_** $\equiv$ AG_AS
Cell	**CG_TS** not applicable	**CG_AS** not applicable	**CG_CS** $\equiv$ CG_	**CG_** $\equiv$ CG_CS
None	**_TS**	**_AS**	**_CS**	_ not interesting

2-anonymized tables wrt different models – 1

Race	DOB	Sex	ZIP
asian	64/04/12	F	94142
asian	64/09/13	F	94141
asian	64/04/15	F	94139
asian	63/03/13	M	94139
asian	63/03/18	M	94139
black	64/09/27	F	94138
black	64/09/27	F	94139
white	64/09/27	F	94139
white	64/09/27	F	94141

PT

Race	DOB	Sex	ZIP
asian	64/04	F	941**
asian	64/04	F	941**
asian	63/03	M	941**
asian	63/03	M	941**
black	64/09	F	941**
black	64/09	F	941**
white	64/09	F	941**
white	64/09	F	941**

AG_TS

2-anonymized tables wrt different models – 2

Race	**DOB**	**Sex**	**ZIP**
asian		F	
asian		F	
asian		F	
asian	63/03	M	9413*
asian	63/03	M	9413*
black	64/09	F	9413*
black	64/09	F	9413*
white	64/09	F	
white	64/09	F	

AG_CS

Race	**DOB**	**Sex**	**ZIP**
asian	64	F	941**
asian	64	F	941**
asian	64	F	941**
asian	63	M	941**
asian	63	M	941**
black	64	F	941**
black	64	F	941**
white	64	F	941**
white	64	F	941**

AG_ $\equiv$AG_AS

2-anonymized tables wrt different models – 3

Race	DOB	Sex	ZIP
asian	64	F	941**
asian	64	F	941**
asian	64	F	941**
asian	63/03	M	94139
asian	63/03	M	94139
black	64/09/27	F	9413*
black	64/09/27	F	9413*
white	64/09/27	F	941**
white	64/09/27	F	941**

CG_≡CG_CS

Race	DOB	Sex	ZIP

_TS

2-anonymized tables wrt different models – 4

Race	DOB	Sex	ZIP
asian		F	
asian		F	
asian		F	
asian		M	
asian		M	
black		F	
black		F	
white		F	
white		F	

_AS

Race	DOB	Sex	ZIP
asian		F	
asian		F	
asian		F	
asian		M	94139
asian		M	94139
	64/09/27	F	
	64/09/27	F	94139
	64/09/27	F	94139
	64/09/27	F	

_CS

Attribute Disclosure

Limitation of k-anonymity

2-anonymous table

Race	DOB	Sex	ZIP	Disease
asian	64	F	941**	hypertension
asian	64	F	941**	obesity
asian	64	F	941**	chest pain
asian	63	M	941**	obesity
asian	63	M	941**	obesity
black	64	F	941**	short breath
black	64	F	941**	short breath
white	64	F	941**	chest pain
white	64	F	941**	short breath

Homogeneity of the sensitive attribute values

- All tuples with a quasi-identifier value in a k-anonymous table may have the same sensitive attribute value
  - an adversary knows that Carol is a black female and that her data are in the microdata table
  - the adversary can infer that Carol suffers from short breath

Race	DOB	Sex	ZIP	Disease
...	...	...	...	...
black	64	F	941**	short breath
black	64	F	941**	short breath
...	...	...	...	...

Background knowledge

- Based on prior knowledge of some additional external information
  - an adversary knows that Hellen is a white female and she is in the microdata table
  - the adversary can infer that the disease of Hellen is either chest pain or short breath
  - the adversary knows that Hellen runs 2 hours a day and therefore that Hellen cannot suffer from short breath
    $\Longrightarrow$ the adversary infers that Hellen's disease is chest pain

Race	**DOB**	**Sex**	**ZIP**	**Disease**
...	...	...	...	...
white	64	F	941**	chest pain
white	64	F	941**	short breath

ℓ-diversity – 1

- A q-block (i.e., set of tuples with the same value for QI) is ℓ-diverse if it contains at least ℓ different "well-represented" values for the sensitive attribute
  - "well-represented": different definitions based on entropy or recursion (e.g., a q-block is ℓ-diverse if removing a sensitive value it remains (ℓ-1)-diverse)
- ℓ-diversity: an adversary needs to eliminate at least ℓ-1 possible values to infer that a respondent has a given value

ℓ-diversity – 2

- A table is ℓ-diverse if all its q-blocks are ℓ-*diverse*
  $\Longrightarrow$ the homogeneity attack is not possible anymore
  $\Longrightarrow$ the background knowledge attack becomes more difficult

- ℓ-diversity is monotonic with respect to the generalization hierarchies considered for k-anonymity purposes

- Any algorithm for k-anonymity can be extended to enforce the ℓ-diverse property

BUT

ℓ-diversity leaves space to attacks based on the distribution of values inside q-blocks (skewness and similarity attacks)

Skewness attack

- Skewness attack occurs when the distribution in a q-block is different than the distribution in the original population
- 20% of the population suffers from diabetes; 75% of tuples in a q-block have diabetes
  $\Longrightarrow$ people in the q-block have higher probability of suffering from diabetes

Race	DOB	Sex	ZIP	Disease
black	64	F	941**	diabetes
black	64	F	941**	short breath
black	64	F	941**	diabetes
black	64	F	941**	diabetes

Similarity attack

- Similarity attack happens when a q-block has different but semantically similar values for the sensitive attribute

Race	**DOB**	**Sex**	**ZIP**	**Disease**
black	64	F	941**	stomach ulcer
black	64	F	941**	stomach ulcer
black	64	F	941**	gastritis

Group closeness

- A q-block respects t-closeness if the distance between the distribution of the values of the sensitive attribute in the q-block and in the considered population is lower than t

- A table respects t-closeness if all its q-blocks respect t-closeness

- t-closeness is monotonic with respect to the generalization hierarchies considered for k-anonymity purposes

- Any algorithm for k-anonymity can be extended to enforce the t-closeness property, which however might be difficult to achieve

External knowledge modeling

- An observer may have external/background knowledge that can be exploited to infer information

- Knowledge may be about:
  - the target individual
  - others: information about individuals other than the target
  - same-value families: knowledge that a group (or family) of individuals have the same sensitive value (e.g., genomic information)

External knowledge – Example (1)

Name	DOB	Sex	ZIP	Disease
Alice	74/04/12	F	94142	aids
Bob	74/04/13	M	94141	flu
Carol	74/09/15	F	94139	flu
David	74/03/13	M	94139	aids
Elen	64/03/18	F	94139	flu
Frank	64/09/27	M	94138	short breath
George	64/09/27	M	94139	flu
Harry	64/09/27	M	94139	aids

Original table

$\Longrightarrow$

DOB	Sex	ZIP	Disease
74		941**	aids
74		941**	flu
74		941**	flu
74		941**	aids
64		941**	flu
64		941**	short breath
64		941**	flu
64		941**	aids

4-anonymized table

Released table is 4-anonymized but ......

External knowledge – Example (2)

DOB	Sex	ZIP	Disease
74		941**	aids
74		941**	flu
74		941**	flu
74		941**	aids
64		941**	flu
64		941**	short breath
64		941**	flu
64		941**	aids

4-anonymized table

An adversary knows that Harry, born in 64 and living in area 94139, is in the table

External knowledge – Example (2)

DOB	Sex	ZIP	Disease
74		941**	aids
74		941**	flu
74		941**	flu
74		941**	aids
64		941**	flu
64		941**	short breath
64		941**	flu
64		941**	aids

4-anonymized table

$\Longrightarrow$

DOB	Sex	ZIP	Disease
64		941**	flu
64		941**	short breath
64		941**	flu
64		941**	aids

4-anonymized table

An adversary knows that Harry, born in 64 and living in area 94139, is in the table

$\Longrightarrow$ Harry belongs to the second group

$\Longrightarrow$ Harry has aids with confidence 1/4

External knowledge – Example (3)

DOB	Sex	ZIP	Disease
64		941**	flu
64		941**	short breath
64		941**	flu
64		941**	aids

4-anonymized table

From another dataset, the adversary knows that George (who is in the table, is born in 64, and leaves in area 941**) has flu

External knowledge – Example (3)

DOB	Sex	ZIP	Disease
64		941**	flu
64		941**	short breath
64		941**	flu
64		941**	aids

4-anonymized table

$\Longrightarrow$

DOB	Sex	ZIP	Disease
64		941**	short breath
64		941**	flu
64		941**	aids

4-anonymized table

From another dataset, the adversary knows that George (who is in the table, is born in 64, and leaves in area 941**) has flu

$\Longrightarrow$ Harry has aids with confidence 1/3

External knowledge – Example (4)

DOB	Sex	ZIP	Disease
64		941**	short breath
64		941**	flu
64		941**	aids

4-anonymized table

From personal knowledge, the adversary knows that Harry does not have short breath

External knowledge – Example (4)

DOB	Sex	ZIP	Disease
64		941**	short breath
64		941**	flu
64		941**	aids

4-anonymized table

$\Longrightarrow$

DOB	Sex	ZIP	Disease
64		941**	flu
64		941**	aids

4-anonymized table

From personal knowledge, the adversary knows that Harry does not have short breath

$\Longrightarrow$ Harry has aids with confidence 1/2

Multiple releases

- Data may be subject to frequent changes and may need to be published on regular basis

- The multiple release of a microdata table may cause information leakage since a malicious recipient can correlate the released datasets

Multiple independent releases – Example (1)

T_1

DOB	Sex	ZIP	Disease
74		941**	aids
74		941**	flu
74		941**	flu
74		941**	aids
64		941**	flu
64		941**	short breath
64		941**	flu
64		941**	aids

4-anonymized table at time t_1

T_2

DOB	Sex	ZIP	Disease
[70-80]		9414*	hypertension
[70-80]		9414*	gastritis
[70-80]		9414*	aids
[70-80]		9414*	gastritis
[60-70]		9413*	flu
[60-70]		9413*	aids
[60-70]		9413*	flu
[60-70]		9413*	gastritis

4-anonymized table at time t_2

An adversary knows that Alice, born in 1974 and living in area 94142, is in both releases

Multiple independent releases – Example (1)

T_1

DOB	Sex	ZIP	Disease
74		941**	aids
74		941**	flu
74		941**	flu
74		941**	aids

4-anonymized table at time t_1

T_2

DOB	Sex	ZIP	Disease
[70-80]		9414*	hypertension
[70-80]		9414*	gastritis
[70-80]		9414*	aids
[70-80]		9414*	gastritis

4-anonymized table at time t_2

An adversary knows that Alice, born in 1974 and living in area 94142, is in both releases

$\Longrightarrow$ Alice belongs to the first group in T_1

$\Longrightarrow$ Alice belongs to the first group in T_2

Multiple independent releases – Example (1)

T_1

DOB	Sex	ZIP	Disease
74		941**	aids
74		941**	flu
74		941**	flu
74		941**	aids

4-anonymized table at time t_1

T_2

DOB	Sex	ZIP	Disease
[70-80]		9414*	hypertension
[70-80]		9414*	gastritis
[70-80]		9414*	aids
[70-80]		9414*	gastritis

4-anonymized table at time t_2

An adversary knows that Alice, born in 1974 and living in area 94142, is in both releases

$\Longrightarrow$ Alice belongs to the first group in T_1

$\Longrightarrow$ Alice belongs to the first group in T_2

Alice suffers from aids (it is the only illness common to both groups)

Multiple independent releases – Example (2)

T_1

DOB	Sex	ZIP	Disease
74		941**	aids
74		941**	flu
74		941**	flu
74		941**	aids
64		941**	flu
64		941**	short breath
64		941**	flu
64		941**	aids

4-anonymized table at time t_1

T_2

DOB	Sex	ZIP	Disease
[70-80]		9414*	hypertension
[70-80]		9414*	gastritis
[70-80]		9414*	aids
[70-80]		9414*	gastritis
[60-70]		9413*	flu
[60-70]		9413*	aids
[60-70]		9413*	flu
[60-70]		9413*	gastritis

4-anonymized table at time t_2

An adversary knows that Frank, born in 1964 and living in area 94132, is the only patient in T_1 but not in T_2

Multiple independent releases – Example (2)

T_1

DOB	Sex	ZIP	Disease
64		941**	flu
64		941**	short breath
64		941**	flu
64		941**	aids

4-anonymized table at time t_1

T_2

DOB	Sex	ZIP	Disease
[60-70]		9413*	flu
[60-70]		9413*	aids
[60-70]		9413*	flu
[60-70]		9413*	gastritis

4-anonymized table at time t_2

An adversary knows that Frank, born in 1964 and living in area 94132, is the only patient in T_1 but not in T_2

Multiple independent releases – Example (2)

T_1

DOB	Sex	ZIP	Disease
64		941**	flu
64		941**	short breath
64		941**	flu
64		941**	aids

4-anonymized table at time t_1

T_2

DOB	Sex	ZIP	Disease
[60-70]		9413*	flu
[60-70]		9413*	aids
[60-70]		9413*	flu
[60-70]		9413*	gastritis

4-anonymized table at time t_2

An adversary knows that Frank, born in 1964 and living in area 94132, is the only patient in T_1 but not in T_2

$\Longrightarrow$ Frank suffers from short breath

Multiple releases

Multiple (i.e., longitudinal) releases cannot be independent

$\Longrightarrow$ need to ensure multiple releases are safe with respect to intersection attacks

Extended scenarios

k-anonymity, ℓ-diversity, and t-closeness different variations

- Multiple tuples per respondent
- Release of multiple tables, characterized by (functional) dependencies
- Multiple quasi-identifiers
- Non-predefined quasi-identifiers
- Release of data streams
- Fine-grained privacy preferences

k-anonymity in various applications

In addition to classical microdata release problem, the concept of k-anonymity and its extensions can be applied in different scenarios, e.g.:

- social networks
- data mining
- location data
- ...

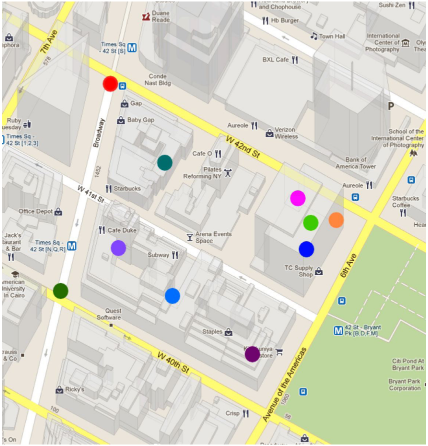
k-anonymity in location-based services

Protect identity of people in locations by considering always locations that contain no less than k individuals:

k-anonymity in location-based services

Protect identity of people in locations by considering always locations that contain no less than k individuals:

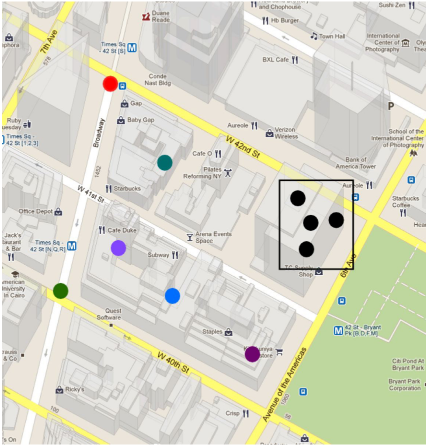
- enlarge the area to include at least other k-1 users (k-anonymity)



k-anonymity in location-based services

Protect identity of people in locations by considering always locations that contain no less than k individuals:

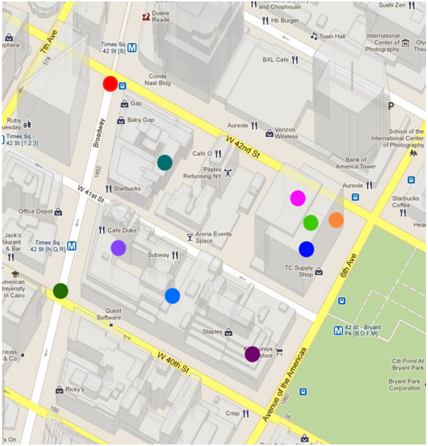
- enlarge the area to include at least other k-1 users (k-anonymity)



Privacy in location-based applications

Protect identity of people in locations by considering always locations that contain no less than k individuals:

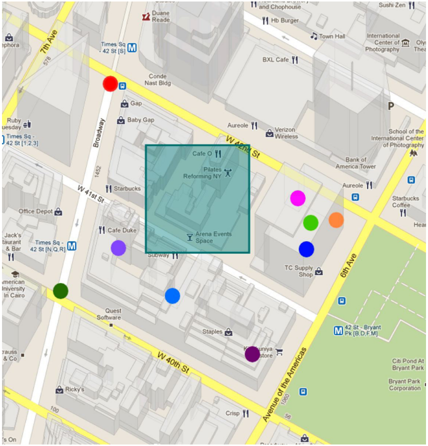
- enlarge the area to include at least other k-1 users (k-anonymity)
- protect the location of users (location privacy)



Privacy in location-based applications

Protect identity of people in locations by considering always locations that contain no less than k individuals:

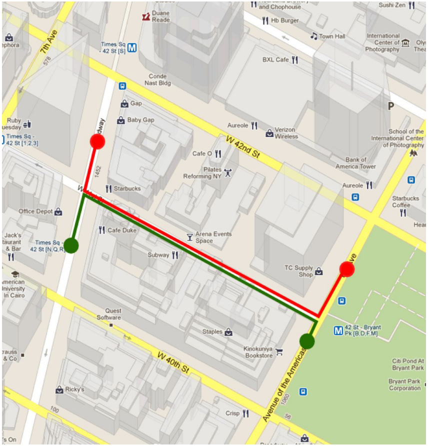
- enlarge the area to include at least other k-1 users (k-anonymity)
- protect the location of users (location privacy)
  $\Longrightarrow$ obfuscate the area so to decrease its precision or confidence



Privacy in location-based applications

Protect identity of people in locations by considering always locations that contain no less than k individuals:

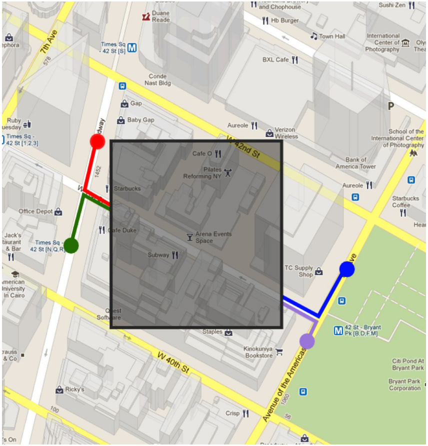
- enlarge the area to include at least other k-1 users (k-anonymity)
- protect the location of users (location privacy)
  $\Longrightarrow$ obfuscate the area so to decrease its precision or confidence
- protect the location path of users (trajectory privacy)



Privacy in location-based applications

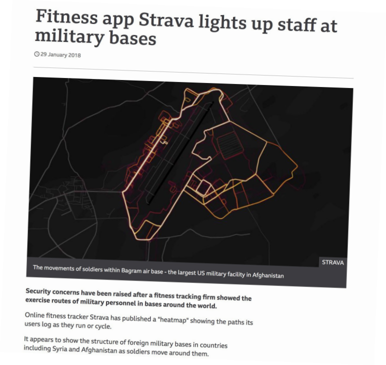
Protect identity of people in locations by considering always locations that contain no less than k individuals:

- enlarge the area to include at least other k-1 users (k-anonymity)
- protect the location of users (location privacy)
  $\Longrightarrow$ obfuscate the area so to decrease its precision or confidence
- protect the location path of users (trajectory privacy)
  $\Longrightarrow$ block tracking by mixing/ modifying trajectories



Fitness app

Maps showing the whereabouts of people who use fitness devices can expose highly sensitive information (location, identity)


Fitness app Strava lights up staff at military bases

29 January 2018

The movements of soldiers within Bagram air base - the largest US military facility in Afghanistan

Security concerns have been raised after a fitness tracking firm showed the exercise routes of military personnel in bases around the world.

Online fitness tracker Strava has published a "heatmap" showing the paths its users log as they run or cycle.

It appears to show the structure of foreign military bases in countries including Syria and Afghanistan as soldiers move around them.


Fitness app Polar exposed locations of spies and military personnel

Location data revealed the home addresses of intelligence officers -- even when their profiles were set to private.

Polar's map based on individualized data, showing exercises done by one person in the Middle East, and the United States.

Anonymization is a complex problem . . .

- Actions/logs can help re-identification
- Even pseudonyms can expose users
  - AOL
  - Netflix
- Multiple sources
- Multiple releases

Re-identification with any information

- Any information can be used to re-identify anonymous data

  $\Longrightarrow$ ensuring proper privacy protection is a difficult task since the amount and variety of data collected about individuals is increased

- Two examples:
  - AOL
  - Netflix

AOL data release – 1

In 2006, to embrace the vision of an open research community, America OnLine publicly posted queries to AOL's search engine

- 20 million search queries for 658,000 users summarizing 3 months of activity
- obviously identifying information (AOL username, IP address) was removed
- usernames replaced with unique identification numbers

AOL data release – 2

User 4417749:

- numb fingers
- 60 single men
- dog that urinates on everything
- hand tremors
- nicotine effects on the body
- dry mouth
- bipolar
- several people with last name Arnold
- landscapers in Lilburn, Ga
- homes sold in shadow lake subdivision Gwinnett county, Georgia

AOL data release – 2

User 4417749:

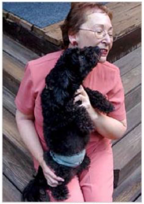
- numb fingers
- 60 single men
- dog that urinates on everything
- hand tremors
- nicotine effects on the body
- dry mouth
- bipolar
- several people with last name Arnold
- landscapers in Lilburn, Ga
- homes sold in shadow lake subdivision Gwinnett county, Georgia

Thelma Arnold, a 62-year-old widow living in Lilburn, Ga

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.



Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

AOL data release – 3

What about user 17556639?

- how to kill your wife
- how to kill your wife
- wife killer
- how to kill a wife
- poop
- dead people
- pictures of dead people
- killed people
- dead pictures
- dead pictures
- dead pictures
- murder photo
- steak and cheese
- photo of death
- photo of death
- death
- dead people photos
- photo of dead people
- www.murderdpeople.com
- decapatated photos
- decapatated photos
- car crashes3
- car crashes3
- car crash photo

Netflix prize data release – 1

In 2006: “Netflix Prize” of USD 1 million for a movie recommendation algorithm that improved Netflix’s algorithm by 10%

- 100 million records (movie rated, rating, date) for 500,000 users from Oct.’98 to Dec.’05
- only a sample (one tenth) of the database was released
- some ratings were perturbed (but not much, not to alter statistics)
- identifying information (usernames) removed, but a unique user identifier was assigned to preserve rating-to-rating continuity

Netflix prize data release – 2

Netflix Prize dataset $+$ IMDb:

- with 6 movie ratings and dates ($\pm$ 2 weeks), 99% of records uniquely identified
- with 2 movie ratings and dates ($\pm$ 3 days), 68% of records uniquely identified
- 84% of subscribers in the dataset uniquely identified by knowing 6 obscure (outside the top 500) movies

Netflix prize data release – 2

Netflix Prize dataset + IMDb:

- with 6 movie ratings and dates ($\pm$ 2 weeks), 99% of records uniquely identified
- with 2 movie ratings and dates ($\pm$ 3 days), 68% of records uniquely identified
- 84% of subscribers in the dataset uniquely identified by knowing 6 obscure (outside the top 500) movies

THREAT LEVEL | privacy

Netflix Spilled Your *Brokeback Mountain* Secret, Lawsuit Claims

BY RYAN SINGEL 12.17.09 4:29 PM

Follow @rsingel

Share 174 | Tweet 18 | +1 0 | Share 5

An in-the-closet lesbian mother is suing Netflix for privacy invasion, alleging the movie rental company made it possible for her to be outed when it disclosed insufficiently anonymous information about nearly half-a-million customers as part of its $1 million contest to improve its recommendation system.

The suit known as Doe v. Netflix (.pdf) was filed in federal court in California on Thursday, alleging that Netflix violated fair-trade laws and a federal privacy law protecting video rental records, when it launched its popular contest in September 2006.

The suit seeks more than $2,500 in damages for each of more than 2 million Netflix customers.

Privacy and genomic data

Genomic information is an opportunity for medicine but there are several privacy issues to be addressed

E.g., human genome:

- identifies its owner
- contains information about ethnic heritage, predisposition to several diseases, and other phenotypic traits
- discloses information about the relatives and descendants of the genome's owner

Privacy and genomic data – Example

The 1000 Genomes Project (2008): to establish a catalogue of human genetic variation

- Re-identification of five men involved in the 1000 Genomes Project and a study on Utah Mormon families
  - their identities determined
  - identities of their male and female relatives discovered
- Cross-reference analysis by WIBR, Cambridge (MA)
  1. extract the haplotypes of short tandem repeats on the donor's Y chromosome (only for males)
  2. enter the haplotypes into genealogical databases to find possible surnames of the donor
  3. enter the surnames into demographic databases

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

News & Comment > News > 2013 > June > Article

NATURE | NEWS

Privacy loophole found in genetic databases

DNA donors' identities can be determined from publicly available records.

Erika Check Hayden

17 January 2013

A potentially serious loophole could allow anyone to unmask the identities of people who contribute their DNA sequences to some research projects, researchers report today.

This is the latest in a series of findings over the past five years that have highlighted privacy vulnerabilities in public databases containing genetic data. The US National Institute of General Medical Sciences (NIGMS), part of the National Institutes of Health (NIH) in Bethesda, Maryland, reacted to the study by removing some data from public view. Some geneticists however question that step, although they acknowledged that the research community must respond to the genetic privacy issue.

Sifting through DNA databases can lead to identify some male subjects that were supposed to be anonymous.

GREG PEASE/GETTY

print

Syntactic vs semantic privacy definitions

- Syntactic privacy definitions capture the protection degree enjoyed by data respondents with a numerical value

  E.g., each release of data must be indistinguishably related to no less than a certain number of individuals in the population

- Semantic privacy definitions are based on the satisfaction of a semantic privacy requirement by the mechanism chosen for releasing the data

  E.g., the result of an analysis carried out on a released dataset must be insensitive to the insertion or deletion of a tuple in the dataset

Differential Privacy

Syntactic vs semantic privacy definitions

- Syntactic privacy definitions capture the protection degree enjoyed by data respondents with a numerical value

  E.g., each release of data must be indistinguishably related to no less than a certain number of individuals in the population

- Semantic privacy definitions are based on the satisfaction of a semantic privacy requirement by the mechanism chosen for releasing the data

  E.g., the result of an analysis carried out on a released dataset must be insensitive to the insertion or deletion of a tuple in the dataset

Differential privacy

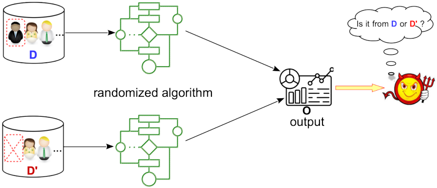
Informally:

- Differential privacy requires the probability distribution on the published results of an analysis to be “essentially the same” independent of whether an individual is represented or not in the dataset

Formally:

- An algorithm A is ε-differentially private if for all pairs of datasets D and D' differing on at most one row, and for all outputs o:

$$\mathrm{P}[A(D) = o] \leq e^{\varepsilon}\ \mathrm{P}[A(D') = o]$$



The privacy budget ε

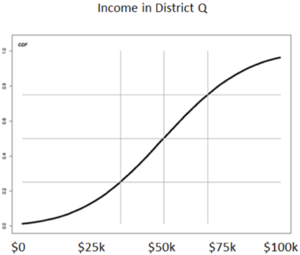
- Determine how much noise is added to the computation
  $\Longrightarrow$ trade-off between privacy and accuracy

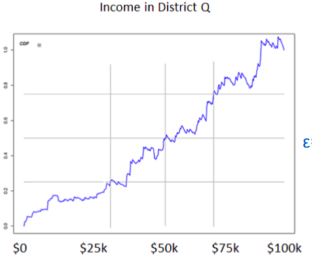
- The smaller (larger) the ε the more (less) the noise
  - small $\varepsilon \Longrightarrow$ more privacy, less utility
  - large $\varepsilon \Longrightarrow$ less privacy, more utility

EXAMPLE

- $\varepsilon = 0 \Longrightarrow$ an analysis could not provide any meaningful output
- $\varepsilon = 0.1 \Longrightarrow$ it provides strong privacy guarantees and useful statistics
- $\varepsilon = 1 \Longrightarrow$ it provides high accuracy but low privacy

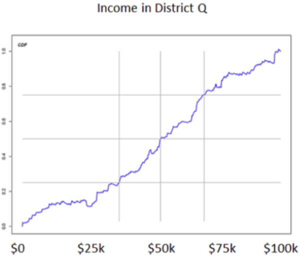
Differential privacy and accuracy

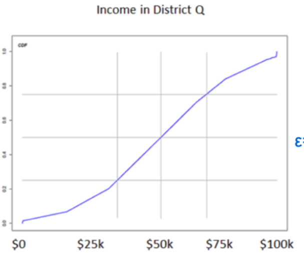




ε=0.005

ε=0.01





ε=0.1

How to achieve differential privacy

- Need to calibrate the noise to the influence an individual can have on the result

- Global sensitivity: characterizes the scale of the influence of one individual (worst case), and hence how much noise we must add

Global sensitivity – Examples (1)

Database D of patients

Sex	Height	DoB	Disease	Drug X
M	6'2"	1960-03-25	Obesity	3.5
F	5'3"	2001-05-05	Diabetes	2.3
F	5'9"	1998-11-13	Healthy	1.0
M	5'3"	2000-10-05	Flu	3.7
M	6'7"	1995-02-22	Flu	2.2
...	...	...	...	...

Global sensitivity – Examples (1)

Database D of patients

Sex	Height	DoB	Disease	Drug X
M	6'2"	1960-03-25	Obesity	3.5
F	5'3"	2001-05-05	Diabetes	2.3
F	5'9"	1998-11-13	Healthy	1.0
M	5'3"	2000-10-05	Flu	3.7
M	6'7"	1995-02-22	Flu	2.2
...	...	...	...	...

- $A(D)$: COUNT(patients who suffer from flu)

$\mathbf{A(D)}$
50

Global sensitivity – Examples (1)

Database D of patients

Sex	Height	DoB	Disease	Drug X
M	6'2"	1960-03-25	Obesity	3.5
F	5'3"	2001-05-05	Diabetes	2.3
F	5'9"	1998-11-13	Healthy	1.0
~~M~~	~~5'3"~~	~~2000-10-05~~	~~Flu~~	~~3.7~~
M	6'7"	1995-02-22	Flu	2.2
...	...	...	...	...

- $A(D)$: COUNT(patients who suffer from flu)

$\mathbf{A(D)}$
50

Global sensitivity – Examples (1)

Database D of patients

Sex	Height	DoB	Disease	Drug X
M	6'2"	1960-03-25	Obesity	3.5
F	5'3"	2001-05-05	Diabetes	2.3
F	5'9"	1998-11-13	Healthy	1.0
~~M~~	~~5'3"~~	~~2000-10-05~~	~~Flu~~	~~3.7~~
M	6'7"	1995-02-22	Flu	2.2
...	...	...	...	...

- $A(D)$: COUNT(patients who suffer from flu)

$\mathbf{A(D)}$
50

$\mathbf{A(D')}$
49

GS(A)=1

Global sensitivity – Examples (2)

Database D of patients

Sex	**Height**	**DoB**	**Disease**	**Drug X**
M	6'2"	1960-03-25	Obesity	3.5
F	5'3"	2001-05-05	Diabetes	2.3
F	5'9"	1998-11-13	Healthy	1.0
M	5'3"	2000-10-05	Flu	3.7
M	6'7"	1995-02-22	Flu	2.2
...	...	...	...	...

Global sensitivity – Examples (2)

Database D of patients

Sex	**Height**	**DoB**	**Disease**	**Drug X**
M	6'2"	1960-03-25	Obesity	3.5
F	5'3"	2001-05-05	Diabetes	2.3
F	5'9"	1998-11-13	Healthy	1.0
M	5'3"	2000-10-05	Flu	3.7
M	6'7"	1995-02-22	Flu	2.2
...	...	...	...	...

- $A(D)$: SUM(usage of drug X) (suppose all values x are in [1,4])

$\mathbf{A(D)}$
33

Global sensitivity – Examples (2)

Database D of patients

Sex	**Height**	**DoB**	**Disease**	**Drug X**
M	6'2"	1960-03-25	Obesity	3.5
F	5'3"	2001-05-05	Diabetes	2.3
F	5'9"	1998-11-13	Healthy	1.0
~~M~~	~~5'3"~~	~~2000-10-05~~	~~Flu~~	~~3.7~~
M	6'7"	1995-02-22	Flu	2.2
…	…	…	…	…

- $A(D)$: SUM(usage of drug X) (suppose all values x are in [1,4])

$\mathbf{A(D)}$
33

Global sensitivity – Examples (2)

Database D of patients

Sex	Height	DoB	Disease	Drug X
M	6'2"	1960-03-25	Obesity	3.5
F	5'3"	2001-05-05	Diabetes	2.3
F	5'9"	1998-11-13	Healthy	1.0
~~M~~	~~5'3"~~	~~2000-10-05~~	~~Flu~~	~~3.7~~
M	6'7"	1995-02-22	Flu	2.2
...	...	...	...	...

- $A(D)$: SUM(usage of drug X) (suppose all values x are in [1,4])

$\mathbf{A(D)}$
33

$\mathbf{A(D')}$
29

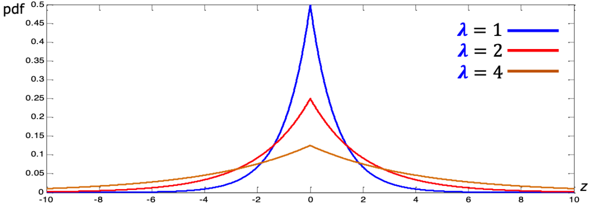
GS(A)=4

Laplace Mechanism with Sensitivity

- Result R is sampled from a Laplace distribution with mean the true result and some scale λ (determined by ε and the global sensitivity of the computation)

$$R = A(D) + z$$

z is a random variable drawn from the Laplace distribution

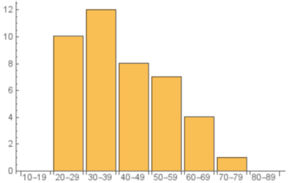


$$\text{Lap}(z, \lambda) = P(z \mid \lambda) = \frac{1}{2\lambda} e^{\frac{-|z|}{\lambda}}, \ \lambda = \frac{GS(A)}{\varepsilon}$$

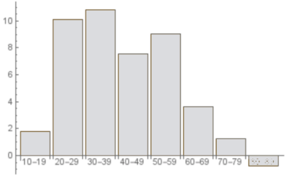
Properties of Differential Privacy

Closure under post-processing

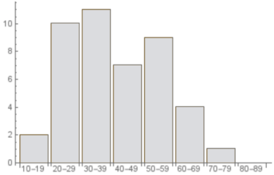
- Differential privacy is resilient to post-processing
  $\Longrightarrow$ the computation of a function over the result of a differentially private computation cannot make it less differentially private



number of users depending on their age ranges ...



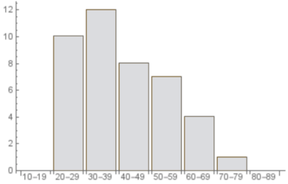
...after the addition of Laplace noise ...



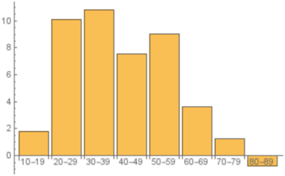
...after rounding all counts and replacing negative numbers with 0

Closure under post-processing

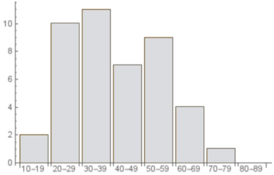
- Differential privacy is resilient to post-processing
  $\Longrightarrow$ the computation of a function over the result of a differentially private computation cannot make it less differentially private



number of users depending on their age ranges ...



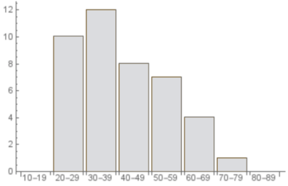
...after the addition of Laplace noise ...



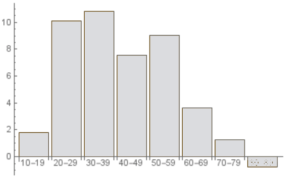
...after rounding all counts and replacing negative numbers with 0

Closure under post-processing

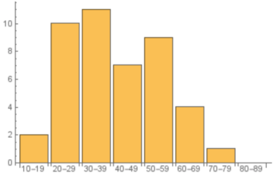
- Differential privacy is resilient to post-processing
  $\Longrightarrow$ the computation of a function over the result of a differentially private computation cannot make it less differentially private



number of users depending on their age ranges ...



...after the addition of Laplace noise ...



...after rounding all counts and replacing negative numbers with 0

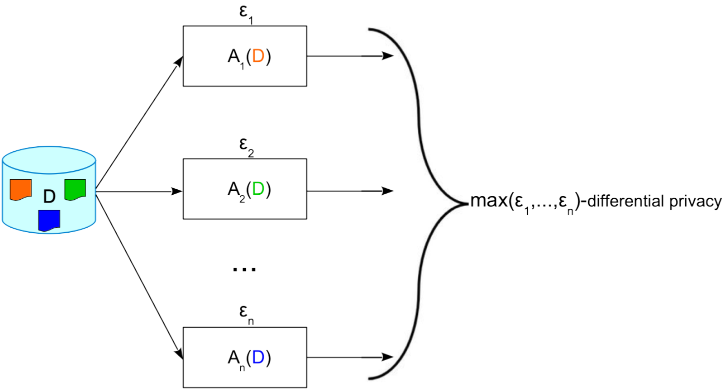
Parallel composition

Differential privacy composes well with itself. But what does it mean?

Parallel composition

Differential privacy composes well with itself. But what does it mean?

- Parallel composition: sequence of m computations over disjoint subsets of a database D



Parallel composition – Example

- $A_1(D)$: COUNT(read hair & left-handed)
- $A_2(D)$: COUNT(blond hair & left-handed)
- $A_3(D)$: COUNT(read hair & right-handed)
- $A_4(D)$: COUNT(blond hair & right-handed)

$\Longrightarrow A_1, A_2, A_3, A_4$ are disjoint

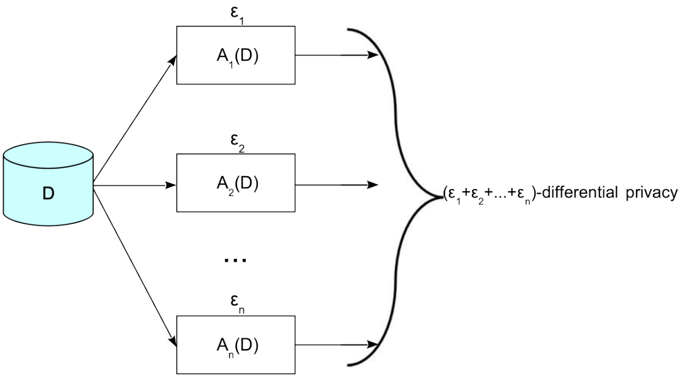
Sequential composition

Differential privacy composes well with itself. But what does it mean?

Sequential composition

Differential privacy composes well with itself. But what does it mean?

- Sequential composition: sequence of m computations over database D with overlapping results



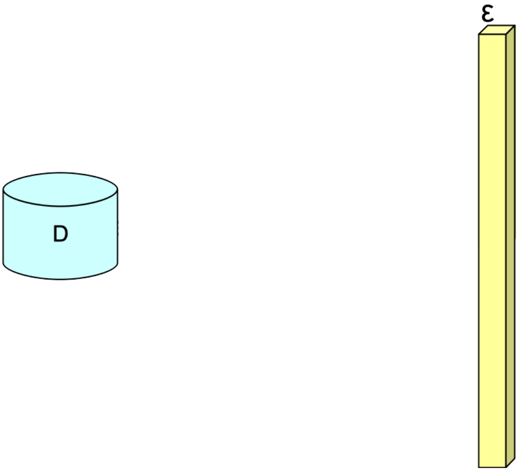
Sequential composition – Example

- $A_1(D)$: COUNT(female patients)

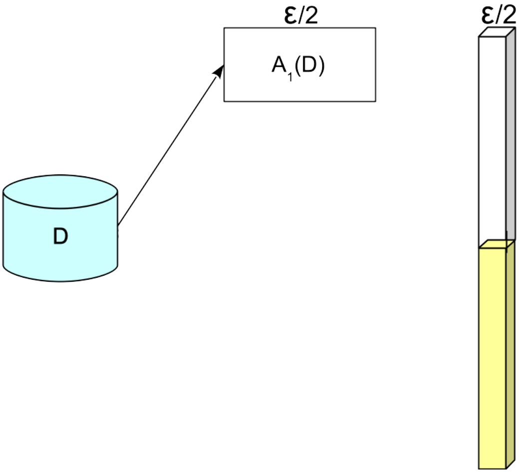
- $A_2(D)$: COUNT(patients suffering from flu)

$\Longrightarrow A_1$ and A_2 can be overlapping (e.g., a female who suffers from flu)

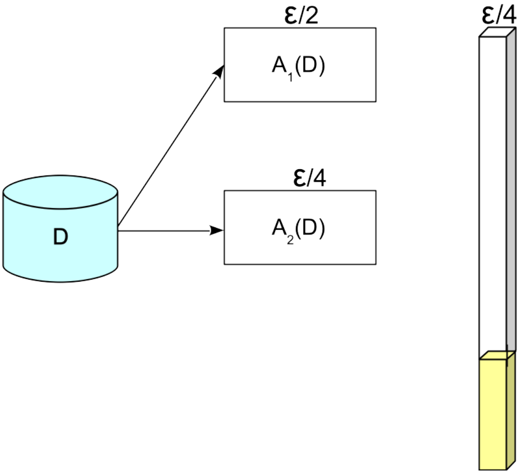
Why ε is called privacy budget?



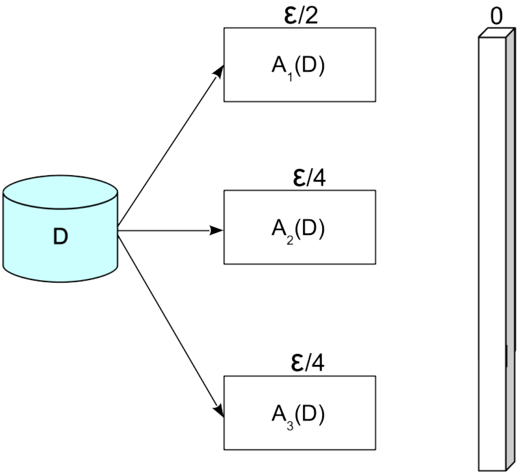
Why ε is called privacy budget?



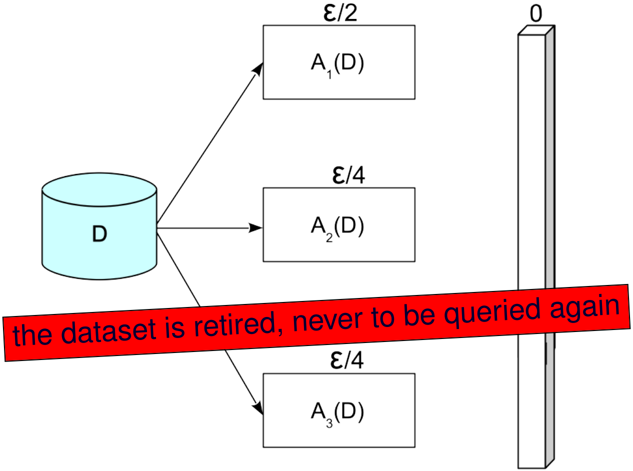
Why ε is called privacy budget?



Why ε is called privacy budget?



Why ε is called privacy budget?



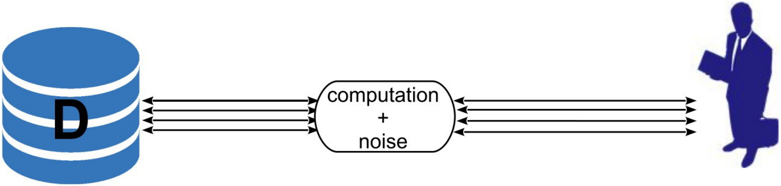
Differential privacy models

- Non-interactive scenario vs interactive

- Global vs local

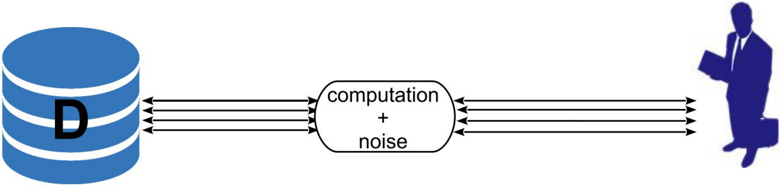
Interactive vs non-interactive

Interactive: run-time evaluation of queries

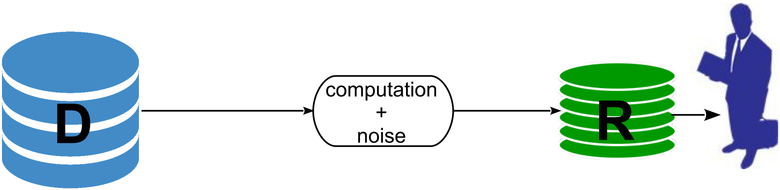


Interactive vs non-interactive

Interactive: run-time evaluation of queries

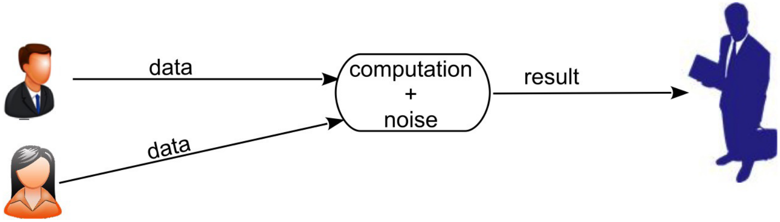


Non-interactive: release of pre-computed macrodata tables



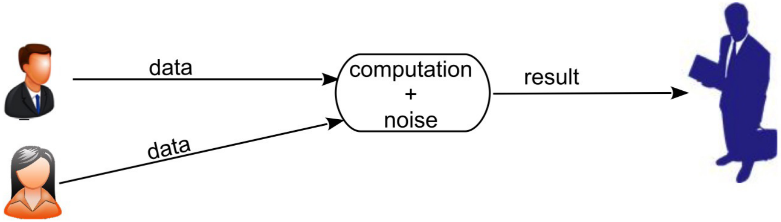
Global vs local differential privacy

Global: applies on the whole dataset comprising all inputs

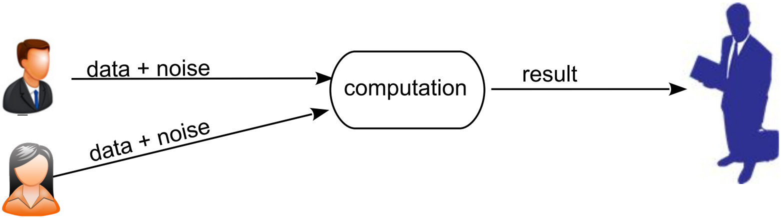


Global vs local differential privacy

Global: applies on the whole dataset comprising all inputs



Local: applies individually to each input before populating the dataset



Local differential privacy definition

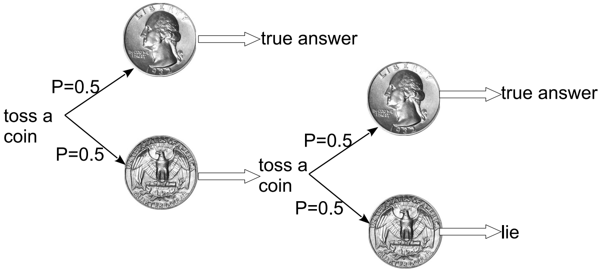
- A randomized algorithm K satisfies ε-local differential privacy iff for all input x, x' and output o of K:

  $$\mathrm{P}[K(x) = o] \leq e^{\varepsilon}\ \mathrm{P}[K(x') = o]$$

  $\Longrightarrow$ any output should not depend on user's secret

(Local) differential privacy in practice

- Differential privacy based on coin tossing is deployed in
  - Google to anonymize data
  - Apple iOS and MacOS to collect typing statistics
- All deployments are based on randomized response



  - P(true answer) = 0.75 = 0.5 + (0.5 $\times$ 0.5)
  - P(lie) = 0.25 = 0.5 $\times$ 0.5

k-anonymity vs differential privacy

Each has its strengths and weaknesses, e.g.,

Syntactic privacy (extending k-anonymity):

+ nice capturing of real-world requirements

– not complete protection

Differential privacy:

+ better protection guarantees

– not easy to understand/enforce, noise can introduce problems, not guaranteeing complete protection either

Still work to be done on both fronts

Some Examples of Other Privacy Issues

Target data mining

In 2012, Target found to mine customers' data for targeted advertising

- Every customer assigned a Guest ID number:
    - tied to credit card, name, email address, . . .
    - stores history of bought goods and other (bought) information

- Purchase history enables mining to
    - infer major life events
    - predict shopping habits
    - target on expected interest

Target data mining

In 2012, Target found to mine customers' data for targeted advertising

- Every customer assigned a Guest ID number:
  - tied to credit card, name, email address, . . .
  - stores history of bought goods and other (bought) information

- Purchase history enables mining to
  - infer major life events
  - predict shopping habits
  - target on expected interest

Forbes

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill Former Staff
Tech
Welcome to The Not-So Private Parts where technology & privacy collide

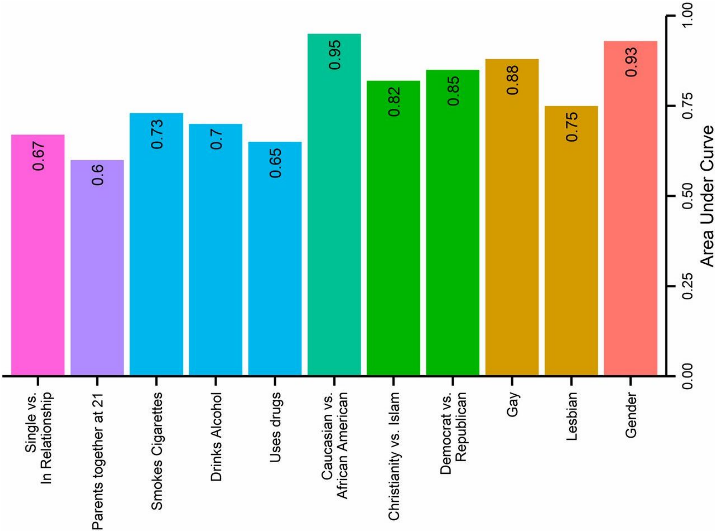
Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Target has got you in its aim

Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into rampant -- and loyal -- buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole -- before Target freaked out and cut off all communications -- about the clues to a customer's impending bundle of joy. Target assigns every customer a Guest ID number, tied to their credit card, name, or email address that becomes a bucket that stores a history of everything they've bought and any demographic information Target has collected from them or bought from other sources. Using that, Pole looked at historical buying data for all the ladies who had signed up for Target baby registries in the past. From the NYT:

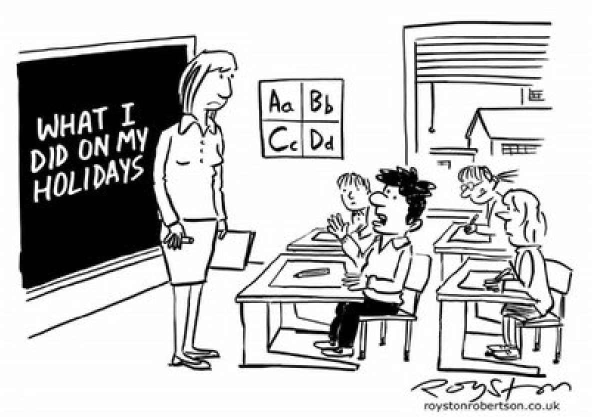
Profiling in social media

Our social media activities and likes may reveal sensitive information



[M. Kosinski, D. Stillwell, T. Graepel, "Digital records of behavior expose personal traits," PNAS, Apr 2013, 110 (15) 5802-5805]

... With the users' help



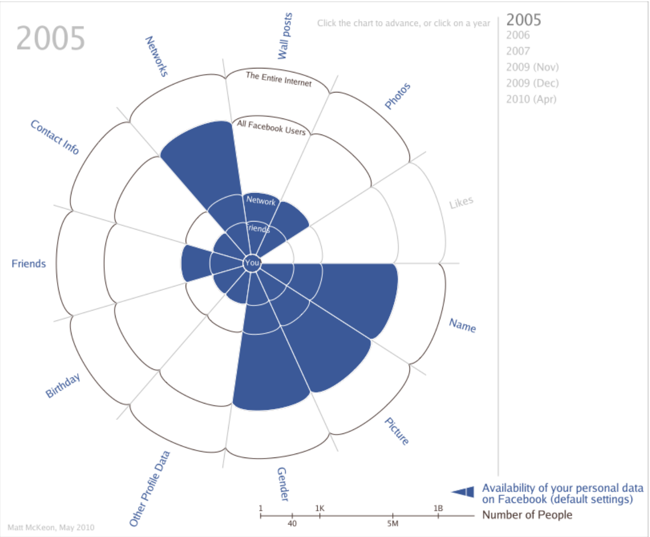
"Can't I just email you a link to my blog, Miss?"

Is information shared with whom?

Facebook default sharing settings from 2005 to 2010

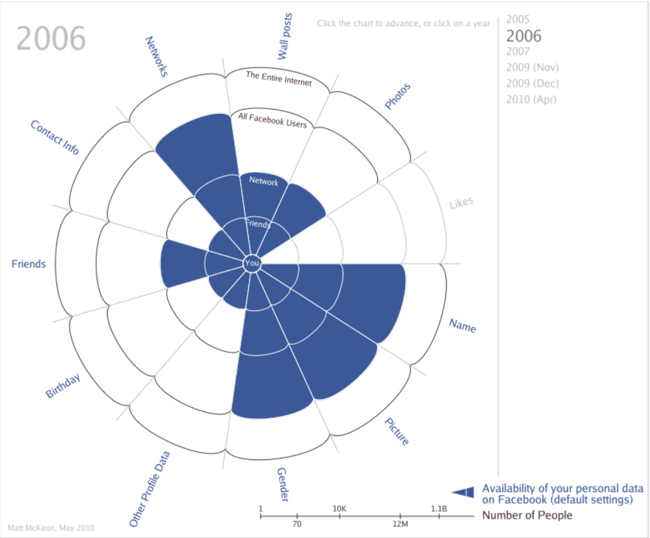
Is information shared with whom?

Facebook default sharing settings from 2005 to 2010



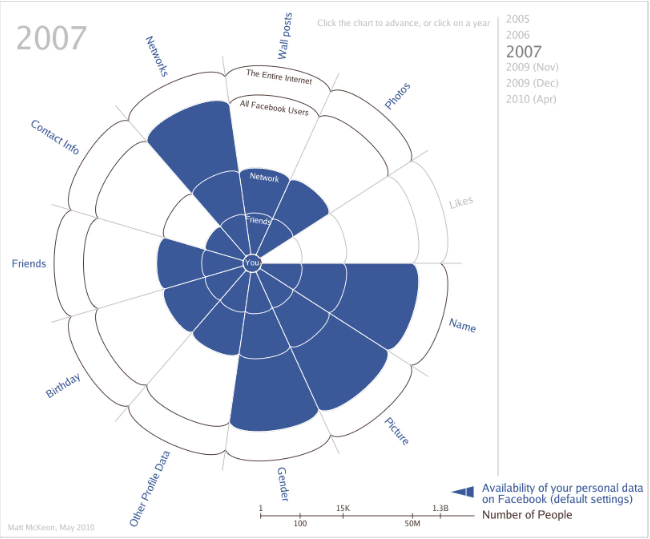
Is information shared with whom?

Facebook default sharing settings from 2005 to 2010



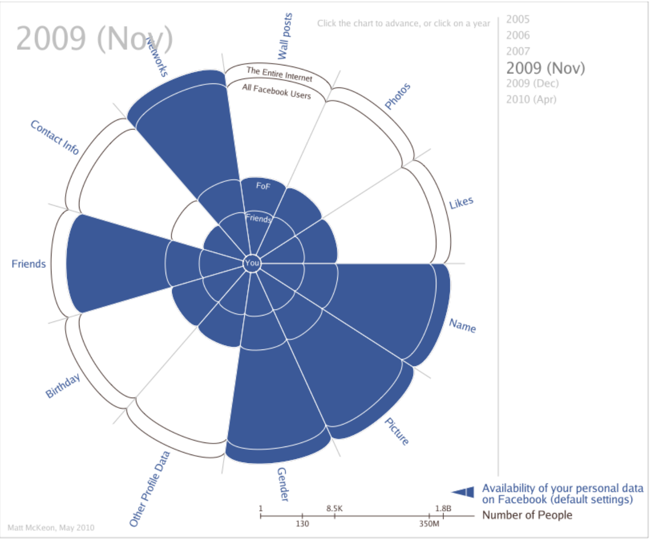
Is information shared with whom?

Facebook default sharing settings from 2005 to 2010



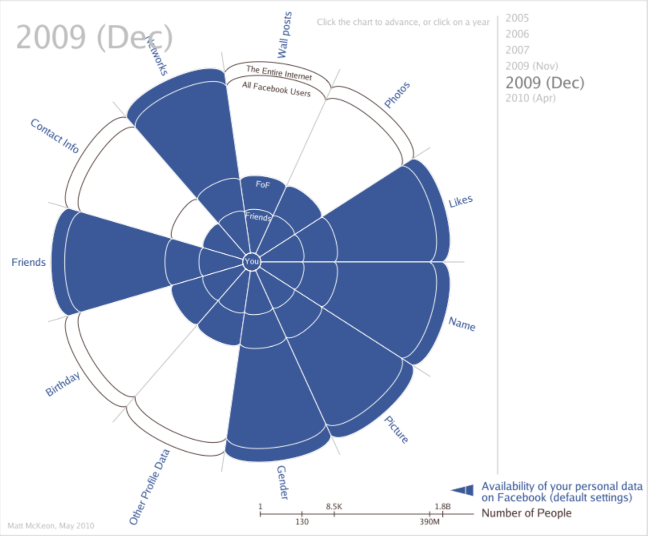
Is information shared with whom?

Facebook default sharing settings from 2005 to 2010



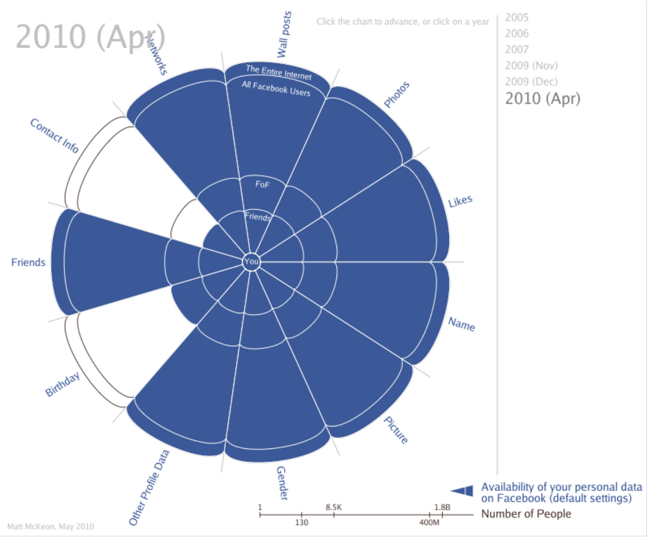
Is information shared with whom?

Facebook default sharing settings from 2005 to 2010



Is information shared with whom?

Facebook default sharing settings from 2005 to 2010



Friends on Facebook? – 1

- In 2011: experiment to study how friendships are created on Facebook
- Implementation of a socialbot
  - software agent simulating human behaviors
  - impersonating a non-existing user
- The socialbot sent friendship requests to unknown users
- Two-step process: no friends in common, and friends of friends

Friends on Facebook? – 2

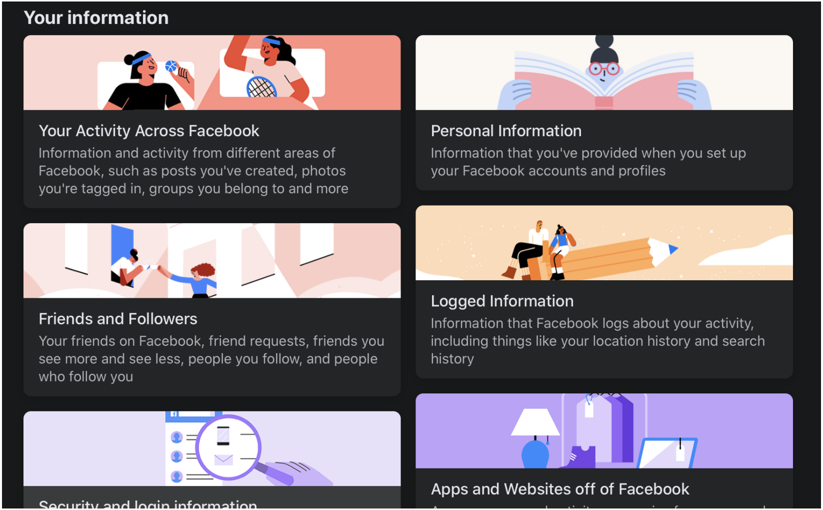
- Accepted requests:
  - 2 out of 10 if no friends in common
  - 6 out of 10 if friends in common
- Three weeks activity, 102 bots:
  - 3,000 friends
  - 46,500 e-mail addresses
  - 14,500 physical addresses

Friends on Facebook? – 2

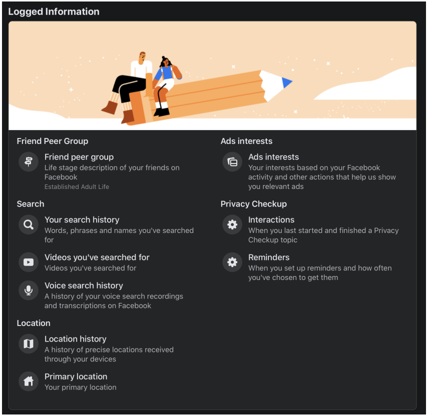
- Accepted requests:
    - 2 out of 10 if no friends in common
    - 6 out of 10 if friends in common
- Three weeks activity, 102 bots:
    - 3,000 friends
    - 46,500 e-mail addresses
    - 14,500 physical addresses

"On the Internet, nobody knows you're a dog."

Facebook: information on you



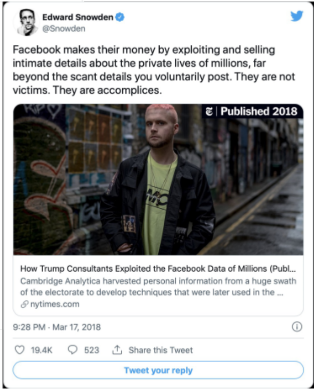
Facebook: information on you



Facebook: information on you

. . . And it’s not only Facebook

Cambridge Analytica scandal – 1



Cambridge Analytica scandal – 2

- Personality quiz app
  - installed by 330,000 Facebook users who gave permission for accessing their data. . .
  - . . . but the app was also collecting data of those users' friends
- Data from 87 million Facebook users retrieved by the app
  - data shared with Cambridge Analytica
  - users profiled through their data

User profiling - Facebook/Cambridge Analytica

OCEAN model

- **O**penness
- **C**onscientiousness
- **E**xtraversion
- **A**greeableness
- **N**euroticism

User profiling - Facebook/Cambridge Analytica

OCEAN model

- **O**penness
  do you enjoy new experiences?
- **C**onscientiousness
  do you prefer plans and order?
- **E**xtraversion
  how social you are?
- **A**greeableness
  do you value others' needs
  and society?
- **N**euroticism
  how much do you tend to worry?

User profiling - Facebook/Cambridge Analytica

OCEAN model

- **O**penness
  do you enjoy new experiences?
- **C**onscientiousness
  do you prefer plans and order?
- **E**xtraversion
  how social you are?
- **A**greeableness
  do you value others' needs and society?
- **N**euroticism
  how much do you tend to worry?

Message to push support for Second Amendment of US Constitution

---

Conscientious individual with high neuroticism:

"The second amendment isn't just a right. It's an insurance policy. Defend the righ to bear arms!"

User profiling - Facebook/Cambridge Analytica

OCEAN model

- **O**penness
  do you enjoy new experiences?
- **C**onscientiousness
  do you prefer plans and order?
- **E**xtraversion
  how social you are?
- **A**greeableness
  do you value others' needs and society?
- **N**euroticism
  how much do you tend to worry?

Message to push support for Second Amendment of US Constitution

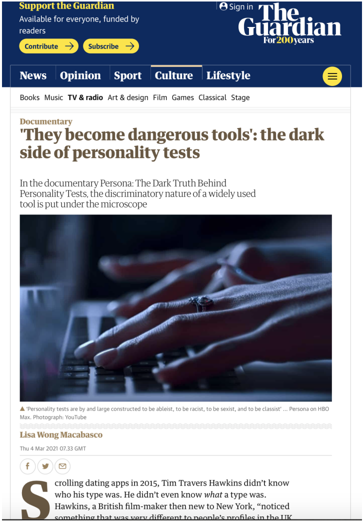
---

Close and agreeable individual:
individual:

"From father to son,
since the birth of our Nation.
Defend the second amendment."

Online quizzes?

- What color are you?
- Which famous historical figure are you?
- Which famous painting are you?
- Who will be your Valentine's Day date?
- . . .
- What will you look like when old?



Support the Guardian
Available for everyone, funded by readers
Contribute → Subscribe →

Sign in
The Guardian
For 200 years

News | Opinion | Sport | Culture | Lifestyle

Books Music **TV & radio** Art & design Film Games Classical Stage

Documentary

'They become dangerous tools': the dark side of personality tests

In the documentary Persona: The Dark Truth Behind Personality Tests, the discriminatory nature of a widely used tool is put under the microscope

▲ 'Personality tests are by and large constructed to be ableist, to be racist, to be sexist, and to be classist' ... Persona on HBO Max. Photograph: YouTube

Lisa Wong Macabasco

Thu 4 Mar 2021 07.33 GMT

Scrolling dating apps in 2015, Tim Travers Hawkins didn't know who his type was. He didn't even know *what* a type was. Hawkins, a British film-maker then new to New York, "noticed something that was very different to people's profiles in the UK

... Is it worth?

"It's this new app – you put in your Social Security Number, and it makes you look like a cat."

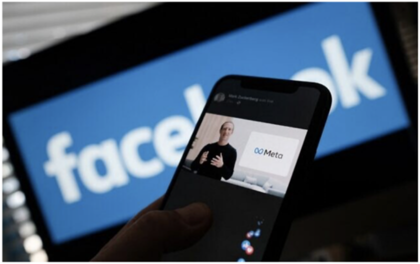
Facebook facial recognition

Facebook to shut down facial recognition system, delete data on 1 billion people

Move by beleaguered company comes amid growing concerns about tech and its misuse by governments, police; parent company Meta appears to be looking at new ways to identify people

By **MATT O'BRIEN** and **BARBARA ORTUTAY**

3 November 2021, 11:22 am |



Facebook facial recognition

Facebook to shut down facial recognition system, delete data on 1 billion people

Move by beleaguered company comes amid growing concerns about tech and its misuse by governments, police; parent company Meta appears to be looking at new ways to identify people

By MATT O'BRIEN and BARBARA ORTUTAY

3 November 2021, 11:22 am |

INSIDER

Login Subscribe

HOME > TECH

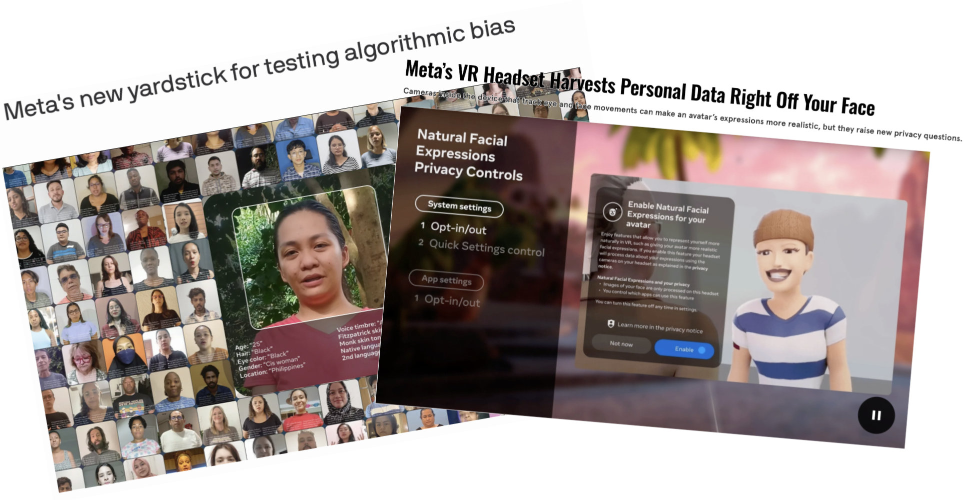
Meta says it's getting rid of facial recognition on Facebook — but that won't apply to the metaverse

Isobel Asher Hamilton Nov 4, 2021, 11:35 AM

Facebook CEO Mark Zuckerberg. Facebook

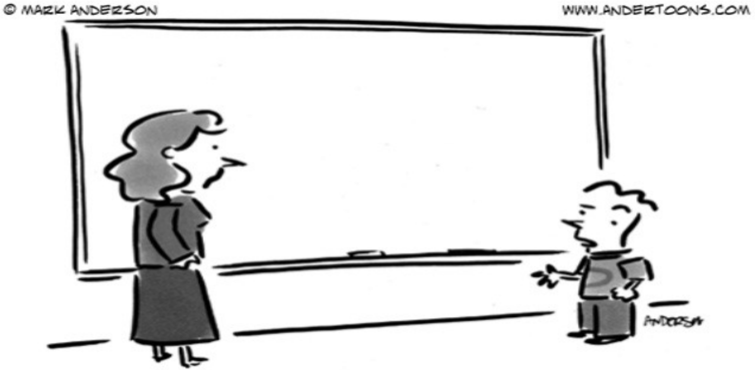
- Facebook announced Tuesday it's shutting down its facial recognition system.
- It said it made the decision because of "growing societal concerns."
- But Meta, Facebook's parent company, isn't ruling out the use of

Biometrics in the Metaverse



Conclusions

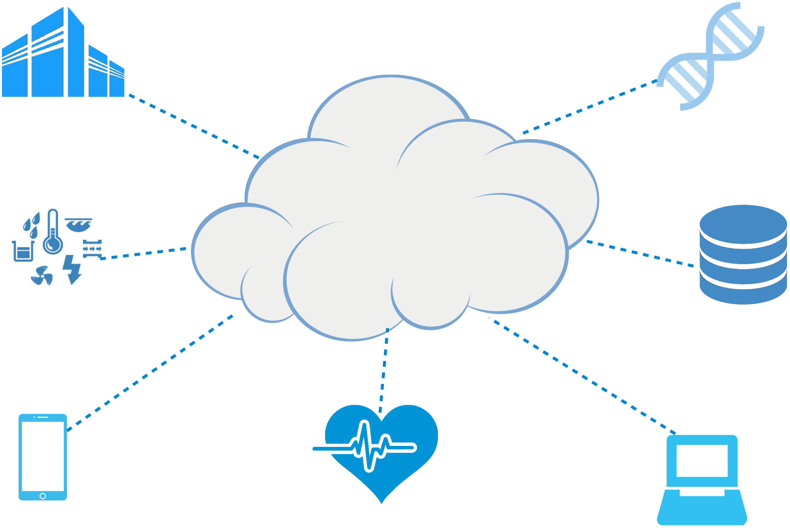
- Technical solutions can provide privacy and data protection
- Legislations demand privacy and data protection
- Privacy and data protection can become assets for ICT players
- … and then there is the user



"Before I write my name on the board, I'll need to know how you're planning to use that data."

Privacy in Data Outsourcing

Huge amount of data stored at external providers

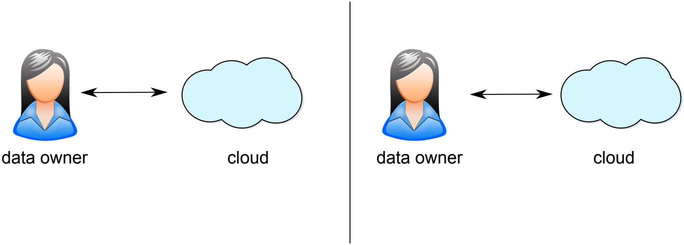


Cloud computing

- The Cloud allows users and organizations to rely on external providers for storing, processing, and accessing their data
  - + high configurability and economy of scale
  - + data and services are always available
  - + scalable infrastructure for applications
- Users lose control over their own data
  - – new security and privacy problems
- Need solutions to protect data and to securely process them in the cloud

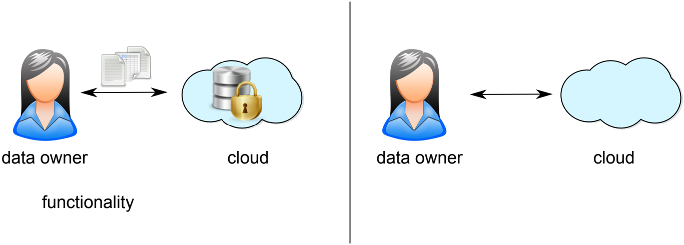
Cloud computing: Today

Cloud Service Providers (CSPs) apply security measures in the services they offer but these measures protect only the perimeter and storage against outsiders



Cloud computing: Today

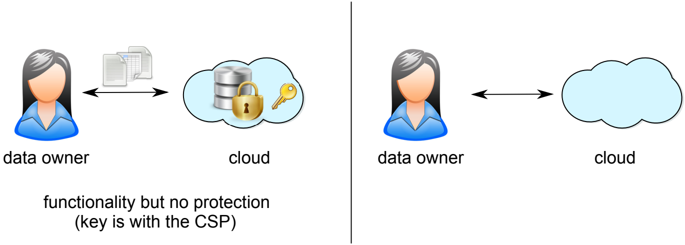
Cloud Service Providers (CSPs) apply security measures in the services they offer but these measures protect only the perimeter and storage against outsiders



- functionality

Cloud computing: Today

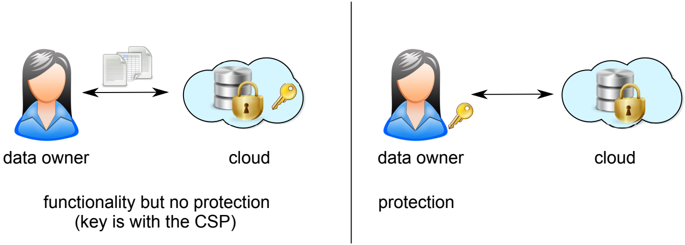
Cloud Service Providers (CSPs) apply security measures in the services they offer but these measures protect only the perimeter and storage against outsiders



- functionality implies full trust in the CSP that has full access to the data (e.g., Google Cloud Storage, iCloud)

Cloud computing: Today

Cloud Service Providers (CSPs) apply security measures in the services they offer but these measures protect only the perimeter and storage against outsiders

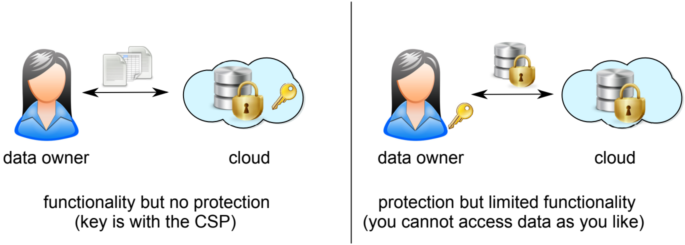


- functionality implies full trust in the CSP that has full access to the data (e.g., Google Cloud Storage, iCloud)

- protection

Cloud computing: Today

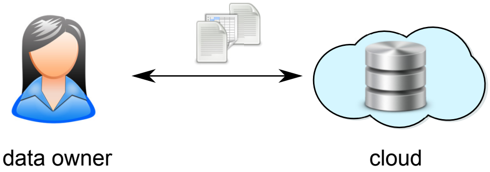
Cloud Service Providers (CSPs) apply security measures in the services they offer but these measures protect only the perimeter and storage against outsiders



- functionality implies full trust in the CSP that has full access to the data (e.g., Google Cloud Storage, iCloud)
- protection but limited functionality since the CSP cannot access data (e.g., Boxcryptor, SpiderOak)

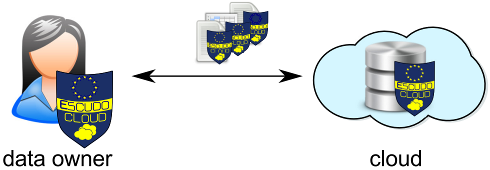
Cloud computing: New vision

Solutions that provide protection guarantees giving the data owners both: full control over their data and cloud functionality over them



Cloud computing: New vision

Solutions that provide protection guarantees giving the data owners both: full control over their data and cloud functionality over them



- client-side trust boundary: only the behavior of the client should be considered trusted

  $\Longrightarrow$ techniques and implementations supporting direct processing of encrypted data in the cloud

Data protection – Base level

Data protection – Base level

a supporter | subscribe | search | jobs | dating | more | International edition

theguardian

otball | opinion | culture | business | lifestyle | fashion | environment | tech | travel

Yahoo hack: 1bn accounts compromised by biggest data breach in history

The latest incident to emerge - which happened in 2013 - is probably distinct from the breach of 500m user accounts in 2014

BUSINESS DAY

Equifax Says Cyberattack May Have Affected 143 Million in the U.S.

TARA SIEGEL BERNARD, TIFFANY HSU, NICOLE PERLROTH and RON LIEBER SEPT. 7, 2017

Technology

Hackers steal 2.5 million PlayStation and Xbox players'

theguardian

Deloitte hit by cyber-attack revealing clients' secret emails

Exclusive: hackers may have accessed usernames, passwo... details of top accountancy firm's blue-chip cli...

Healthcare IT News | TOPIC

Privacy & Security

Even with encryption, EMR data at risk

'While encryption could offer some protections ... it also has serious limitations'

The Register® Biting the hand that feeds IT

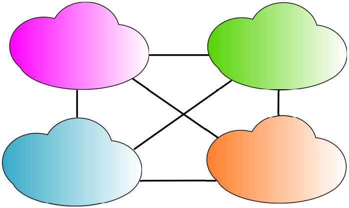
RE | SOFTWARE | SECURITY | TRANSFORMATION | DEVOPS | BUSINESS | PERSONAL TECH

Security

Two million customer records pillaged in IT souk CeX hack attack

Data protection – Regulation

Access and usage control



Selective sharing

General Data Protection Regulation

HIPAA
Health Insurance Portability & Accountability Act

Governance and regulation

Data protection – Confidentiality (1)

- Minimize release/exposition
  - correlation among different data sources
  - indirect exposure of sensitive information
  - de-identification $\neq$ anonymization

Forbes · Real Time · Most Popular · Lists

TECH | 2/16/2012 @ 11:02AM | 1,544,998 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

259 comments, 150 called-out · + Comment now

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are

TECHNOLOGY | UNBOXED

Big Data Is Opening Doors, but Maybe Too Many

By STEVE LOHR MARCH 23, 2013

IN the 1960s, mainframe computers posed a significant technological challenge to common notions of privacy. That's when the federal government started putting tax returns into those giant machines, and consumer credit bureaus began building databases containing the personal financial information of millions of Americans. Many people feared that the new computerized databanks would be put in the service of an intrusive corporate or government Big Brother.

Data protection – Confidentiality (2)

THREAT LEVEL | privacy

Netflix Spilled Your *Brokeback Mountain* Secret, Lawsuit Claims

BY RYAN SINGEL 12.17.09 4:29 PM

nature

International weekly journal of science

NATURE | NEWS

Privacy loophole found in genetic databases

DNA donors' identities can be determined from publicly available records.

Erika Check Hayden

17 January 2013

The Telegraph

HOME » TECHNOLOGY » FACEBOOK

Gay men 'can be identified by their Facebook friends'

Homosexual men can be identified just by looking at their Facebook friends, a… to unpublished research by two students at the Massachusetts Institute of Tec…

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

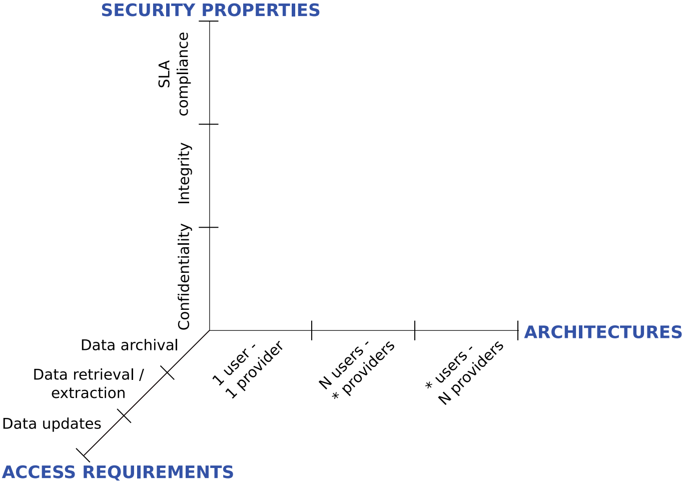
No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake

Characterization of Data Protection Challenges in Cloud Scenarios

Scientific and technical challenges

Three dimensions characterize the problems and challenges



Security properties

Confidentiality
- data externally stored
- users identities
- actions that users perform on the data

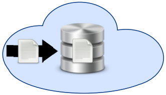
Integrity
- data externally stored
- computation and query results

SLA compliance
- assurance and certification

Access requirements

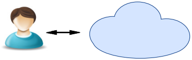
Data archival
- upload/download
- protection of data in storage

Data retrieval/extraction
- support for fine-grained data retrieval and queries
- protection of computations and query results

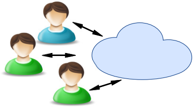


Data update
- support for access retrieval and enforcement of updates
- protection of the actions and of their effects on the data

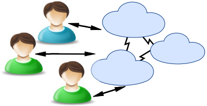
Architectures



1 user - 1 provider
- protection of data at rest
- fine-grained retrieval
- query privacy/integrity



n users - * providers
- authorizations and access control
- multiple writers



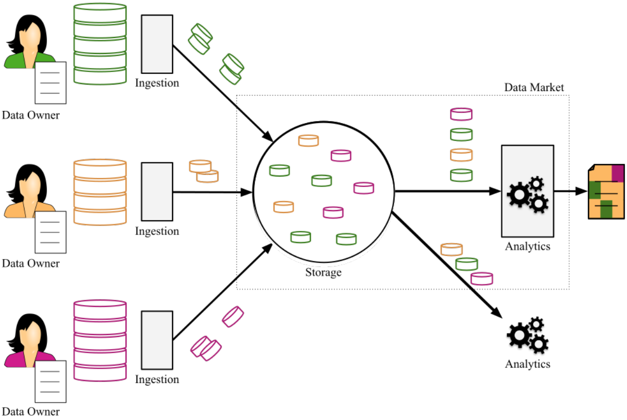
*** users - n providers**
- controlled data sharing and computation

Combinations of the dimensions

- Every combination of the different instances of the dimensions identifies new problems and challenges
- The security properties to be guaranteed can depend on the access requirements and on the trust assumption on the providers involved in storage and/or processing of data
- Providers can be:
  - curious
  - lazy
  - malicious

Digital Data Market

Digital Data Market



Dimensions of the problem and challenges

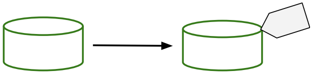
- Requirements capturing and representation

  policies regulating access, sharing, usage and processing

Dimensions of the problem and challenges

- Requirements capturing and representation

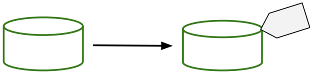
  policies regulating access, sharing, usage and processing



Dimensions of the problem and challenges

- Requirements capturing and representation

  policies regulating access, sharing, usage and processing



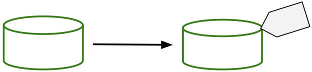
- Enforcing technologies

  data wrapping / sanitization

Dimensions of the problem and challenges

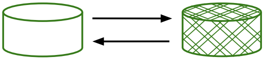
- Requirements capturing and representation

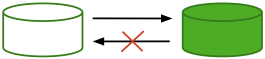
  policies regulating access, sharing, usage and processing



- Enforcing technologies

  data wrapping / sanitization

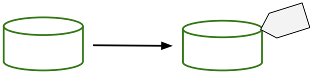




Dimensions of the problem and challenges

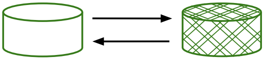
- Requirements capturing and representation

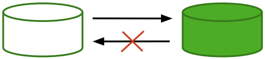
  policies regulating access, sharing, usage and processing



- Enforcing technologies

  data wrapping / sanitization



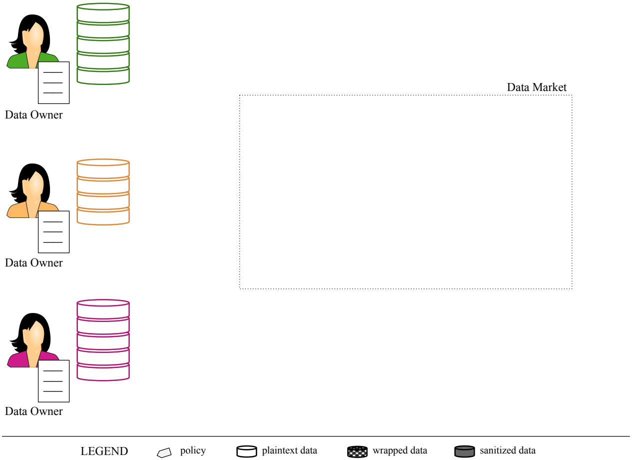


- Enforcement phase

  ingestion / storage / analytics

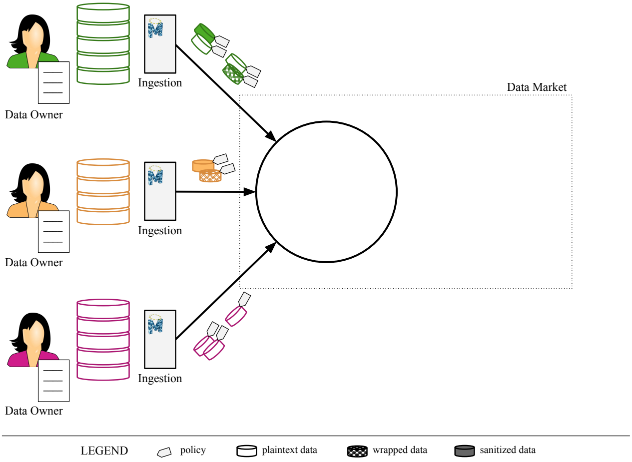
Enforcement phase

- Ingestion / Storage / Analytics



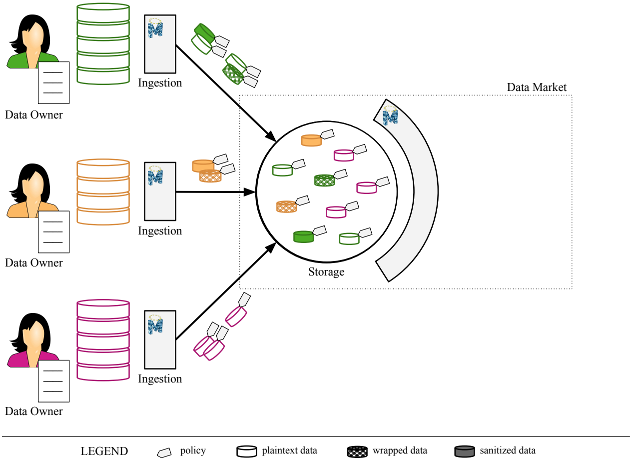
Enforcement phase

- Ingestion / Storage / Analytics



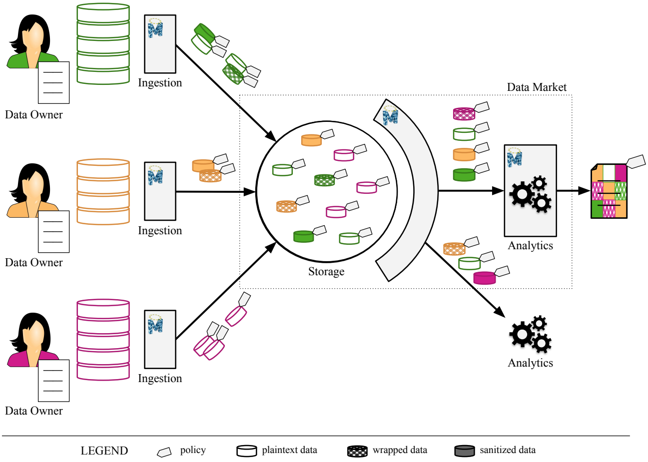
Enforcement phase

- Ingestion / Storage / Analytics



Enforcement phase

- Ingestion / Storage / Analytics



Some open issues

Fine-grained access over encrypted data

Controlled collaborative query execution

Distributed resource allocation and computations

Data/computation integrity

Access confidentiality

Providers/plans selection

User privacy

Security metrics

Query privacy

Secure energy-aware data management

Protection of data at rest

Data publication and utility

Green IT and cybersecurity

Policy definition and modeling

Conclusions

- Advancements in ICT:
  - enable new and better applications and services, bringing social and economic benefits
  - need to address new security and privacy risks and challenges

... towards allowing society to fully benefit from information technology while enjoying security and privacy