

# Data Privacy & Ethics

# Relevance of Privacy



## Golden State Killer suspect traced using genealogy websites

🕒 27 April 2018

f t m e Share

Golden State Killer



# What is Privacy?

- Privacy is the protection of an individual's personal information.
- Privacy is the rights and obligations of individuals and organizations with respect to the collection, use, retention, disclosure and disposal of personal information.
- Privacy  $\neq$  Confidentiality
- No clear definition, legal or otherwise

# Outline

- Data Consumer
  - Use of data
  - Data location
  - Secure storage
- Data Producer
  - K-anonymity
  - L-diversity
  - T-closeness
  - Generalization attacks
  - Differential Privacy

# OECD Privacy Principles

## 1. Collection Limitation Principle

There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject.

## 2. Data Quality Principle

Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date.

# OECD Privacy Principles

## 3. Purpose Specification Principle

The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfilment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.

## 4. Use Limitation Principle

Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with Principle 3 except:

- a) with the consent of the data subject; or
- b) by the authority of law.

# OECD Privacy Principles

## 5. Security Safeguards Principle

Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorized access, destruction, use, modification or disclosure of data.

## 6. Openness Principle

There should be a general policy of openness about developments, practices and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.

# OECD Privacy Principles

## 7. Individual Participation Principle

An individual should have the right:

- a) to request to know whether or not the data controller has data relating to him;
- b) to request data relating to him, ...
- c) to be given reasons if a request is denied; and
- d) to request the data to be rectified, completed or amended.

## 8. Accountability Principle

A data controller should be accountable for complying with measures which give effect to the principles stated above.



# US - Medical Data - HIPAA

- Comprehensive regulations with respect to treating medical data
- US law broadly adopted by healthcare providers
- Adoption subsidized for years now lack of adoption is penalized
- Broadly used outside US as well for releasing medical data

# HIPAA Privacy Rule

"Under the safe harbor method, covered entities must remove all of a list of 18 enumerated identifiers and **have no actual knowledge that the information remaining could be used, alone or in combination, to identify a subject of the information.**"

"The identifiers that must be removed include direct identifiers, such as name, street address, social security number, as well as other identifiers, such as birth date, admission and discharge dates, and five-digit zip code. The safe harbor requires removal of geographic subdivisions smaller than a State, except for the initial three digits of a zip code if the geographic unit formed by combining all zip codes with the same initial three digits contains more than 20,000 people. In addition, age, if less than 90, gender, ethnicity, and other demographic information not listed may remain in the information. The safe harbor is intended to provide covered entities with a simple, definitive method that does not require much judgment by the covered entity to determine if the information is adequately de-identified."

# Pseudonymization

- Remove identifying fields and replace with artificial identifiers
- Trade-off between statistical utility and anonymization
- Mapping kept for incidental findings
- Oftentimes also carried out as de-identification

# European Union Data Protection Directive

The European Union Data Protection Directive of 1995 **establishes common rules for data protection among Member States of the European Union**. The Directive was created in the early 1990s and formally adopted in 1995. The EU is now in the process of replacing it with a General Regulation on Data Protection (Proposed Regulation). The Commission introduced the Proposed Regulation in 2012, and the Parliament passed an amended version of it in 2014. Once enacted, the Proposed Regulation will replace the Directive and be directly binding on all Member States.”

# EU Data Protection Directive: Implementation of Directives

“Directives are a form of EU law that is **binding for Member States**, but **only as to the result** to be achieved. They allow the national authorities to choose the form and the methods of their implementation and generally fix a deadline for it. Therefore, the rules of law applicable in each Member State are the national laws implementing the directives and not the directive itself. However, the directive has a ‘direct effect’ on individuals: **it grants them rights that can be upheld by the national courts in their respective countries if their governments have not implemented the directive** by the set deadline.

A directive thus grants *rights* rather than creates obligations, and they are enforceable by *individuals* rather than by public authorities.”

## EU Data Protection Directive: Article 25

“The Directive extends privacy safeguards to personal data that are transferred outside of the European Union. Article 25 of the Directive states that data can only be transferred to third countries that provide an ‘adequate level of data protection.’ As a result, implementation focuses on both the adoption of national law within the European Union and the adoption of adequate methods for privacy protection in third party countries.”

## EU Data Protection Directive: International Data Transfers

- Cross-border information flow v. individual privacy
- “**Article 25** governs when Member States may permit the flow of personal data to other countries. This provision has particular relevance for the United States, because it **governs the level of privacy protections other countries must have in place for data transfers to occur**”.

# EU Data Protection Directive:

## Derogations/Exceptions

- “Transfers of personal data to a third-party country that does not ensure an adequate level of protection under Article 25(2) may still take place on condition that the **data subject has unambiguously consented, the transfer of data is ‘necessary in order to protect the vital interests of the data subject,’ or the transfer serves ‘important public interest grounds.’** There are several additional exceptions.
- A Member State may also authorize transfers of personal data to third countries without an adequate level of protection where protection of the privacy and individual freedoms ‘result **from appropriate contractual clauses.**’”



# EU Data Protection Directive: Consent

## Derogation 2 of 4: Consent of Data Subject

- Consent must be a clear and unambiguous indication of wishes, given freely, specific and informed.
- Very high threshold

# EU Data Protection Directive: Privacy Shield

- “Safe Harbor 2.0”
- The EU-U.S. Privacy Shield Framework was designed by the U.S. Department of Commerce and European Commission to provide companies on both sides of the Atlantic with a mechanism to comply with EU data protection requirements when transferring personal data from the European Union to the United States in support of transatlantic commerce.

# EU Data Protection Directive: Privacy Shield

## Privacy Shield Principles

1. Notice
2. Choice
3. Accountability for Onward Transfer
4. Security
5. Data Integrity and Purpose Limitation
6. Access
7. Recourse, Enforcement and Liability

# GDPR - What's changing?

- Many GDPR principles are similar to those in current the Data Protection Act.
- There are also new and strengthened requirements for how we protect people's data.
- Changes include:
  - new rights (e.g. 'right to be forgotten')
  - greater emphasis on transparency and record-keeping
  - mandatory data breach reporting
  - much larger fines for when organisations get things wrong

# Data Releases

- Legal compliance is one aspect
- Equally important are ethical regulations
- Ethical compliance does not mean legal compliance and vice versa

# SECURE STORAGE

# CryptDB

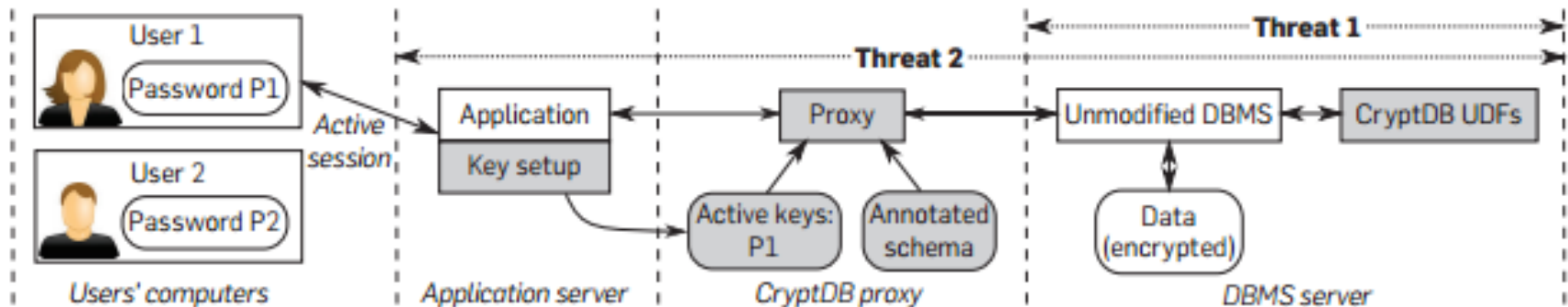
## Passive Attacks

- Compromised hardware
- System Administrators
- Cloud solutions

# Implementation

## Three Components

- Application
- Proxy
- DBMS





# Database Structure

- Table Names
- Column Names

# Encryption Types

- Random (RND)
  - Maximum security
- Deterministic (DET)
  - Plaintext results in consistent ciphertext
- Order-Preserving Encryption (OPE)
  - $100 < 200$    |    $4\text{ex}5\text{d} < 7\text{gfa}3$

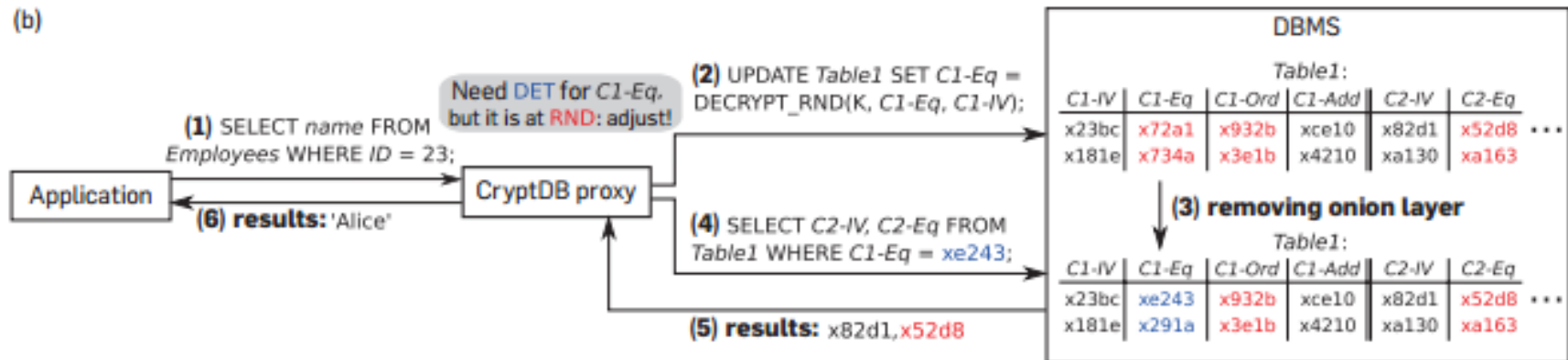
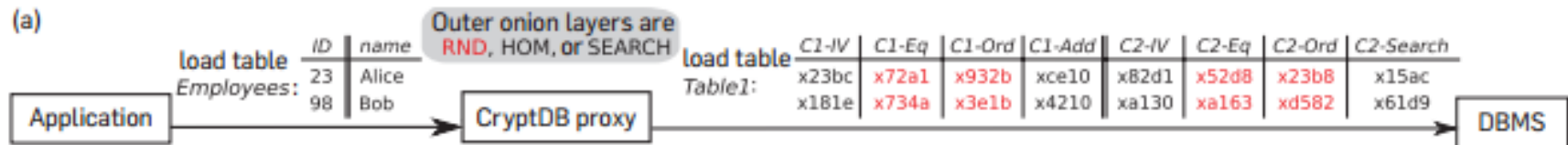
# Encryption Types

- Homomorphic Encryption (HOM)
  - Math functions (ex. Addition)
- Join (JOIN and OPE-JOIN)
  - Equality Joins
- Word Search (SEARCH)
  - LIKE

# Query Processing Steps

- Application issues query, intercepted by proxy and rewritten.
- If necessary, adjust column encryption level.
- Proxy sends encrypted query to DBMS for execution.
- Encrypted result returned, proxy decrypts, returns to application.

# Query Processing



# CryptDB

- CryptDB utilizes several encryption technologies to take steps to secure data within your client/server applications from passive attacks.
- More secure than encryption provided by DBMS. DBMS decrypts data to perform queries.
- Supports most relational queries – not all. Further research is being done here.

# PUBLISHING DATA

# Public Data Conundrum

- Health-care datasets
  - Clinical studies, hospital discharge databases ...
- Genetic datasets
  - \$1000 genome, HapMap, deCode ...
- Demographic datasets
  - U.S. Census Bureau, sociology studies ...
- Search logs, recommender systems, social networks, blogs ...
  - AOL search data, social networks of blogging sites, Netflix movie ratings, Amazon ...



# What About Privacy?

- First thought: anonymize the data
- How?
- Remove “personally identifying information” (PII)
  - Name, Social Security number, phone number, email, address... what else?
  - Anything that identifies the person directly
- Is this enough?

# Re-identification by Linking

Microdata

ID	QID			SA
Name	Zipcode	Age	Sex	Disease
Alice	47677	29	F	Ovarian Cancer
Betty	47602	22	F	Ovarian Cancer
Charles	47678	27	M	Prostate Cancer
David	47905	43	M	Flu
Emily	47909	52	F	Heart Disease
Fred	47906	47	M	Heart Disease

Voter registration data

Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

# Latanya Sweeney's Attack (1997)

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	City	Date Of Birth	Sex	ZIP	Marital Status	Problem
			09/27/64	female	02139	divorced	hypertension
			09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	.....
.....	.....	.....	.....	.....	.....	.....	
Sue J. Carlson	1459 Main St.	Cambridge	02142	9/15/61	female	democrat	.....
.....	.....	.....	.....	.....	.....	.....	

Figure 1 Re-identifying anonymous data by linking to external data

Public voter dataset

# Netflix Movie Rating Data

- Netflix released anonymized movie rating data for its Netflix challenge
  - With date and value of movie ratings
- Knowing 6-8 approximate movie ratings and dates is able to uniquely identify a record with over 90% probability
  - Correlating with a set of 50 users from imdb.com yields two records
- Netflix cancels second phase of the challenge

# Quasi-Identifiers

- Key attributes
  - Name, address, phone number - uniquely identifying!
  - Always removed before release
- Quasi-identifiers
  - (5-digit ZIP code, birth date, gender) uniquely identify 87% of the population in the U.S.
  - Can be used for linking anonymized dataset with other datasets

# Classification of Attributes

## Sensitive attributes

- Medical records, salaries, etc.
- These attributes is what the researchers need, so they are always released directly

Key Attribute	Quasi-identifier			Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

# K-Anonymity: Intuition

- The information for each person contained in the released table cannot be distinguished from at least  $k-1$  individuals whose information also appears in the release
  - Example: you try to identify a man in the released table, but the only information you have is his birth date and gender. There are  $k$  men in the table with the same birth date and gender.
- Any quasi-identifier present in the released table must appear in at least  $k$  records

# K-Anonymity Protection Model

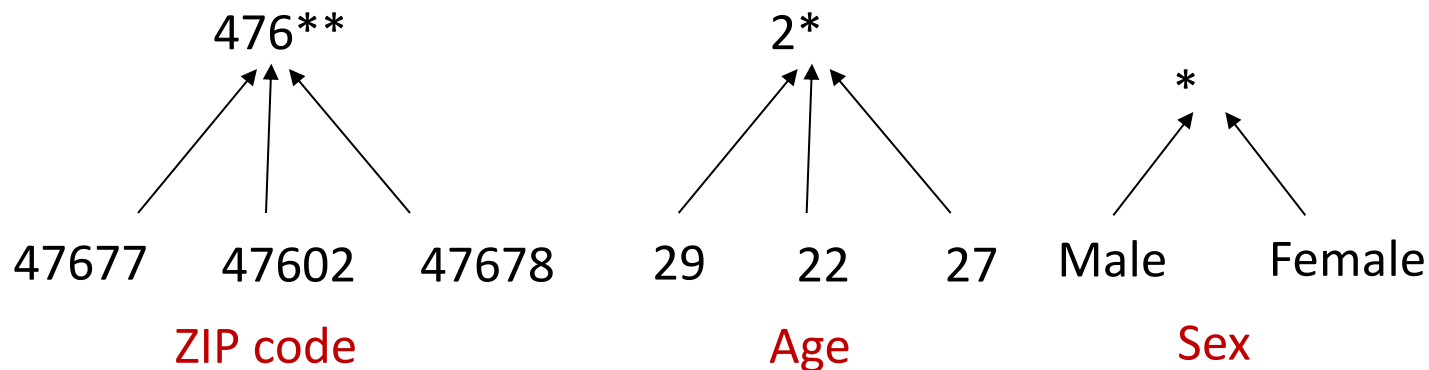
- Private table
- Released table:  $RT$
- Attributes:  $A_1, A_2, \dots, A_n$
- Quasi-identifier subset:  $A_i, \dots, A_j$

Let  $RT(A_1, \dots, A_n)$  be a table,  $QI_{RT} = (A_i, \dots, A_j)$  be the quasi-identifier associated with  $RT$ ,  $A_i, \dots, A_j \subseteq A_1, \dots, A_n$ , and  $RT$  satisfy  $k$ -anonymity. Then, each sequence of values in  $RT[A_x]$  appears with at least  $k$  occurrences in  $RT[QI_{RT}]$  for  $x=i, \dots, j$ .



# Generalization

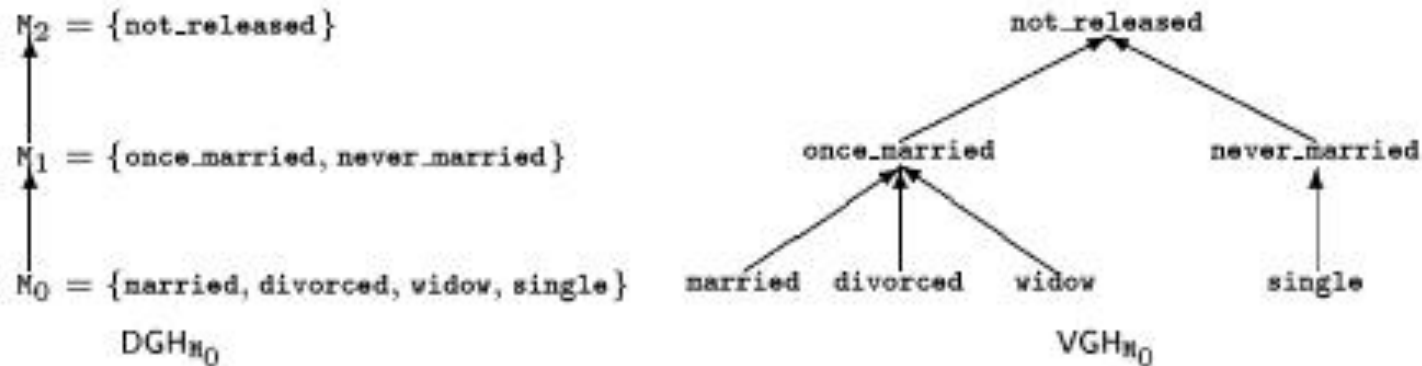
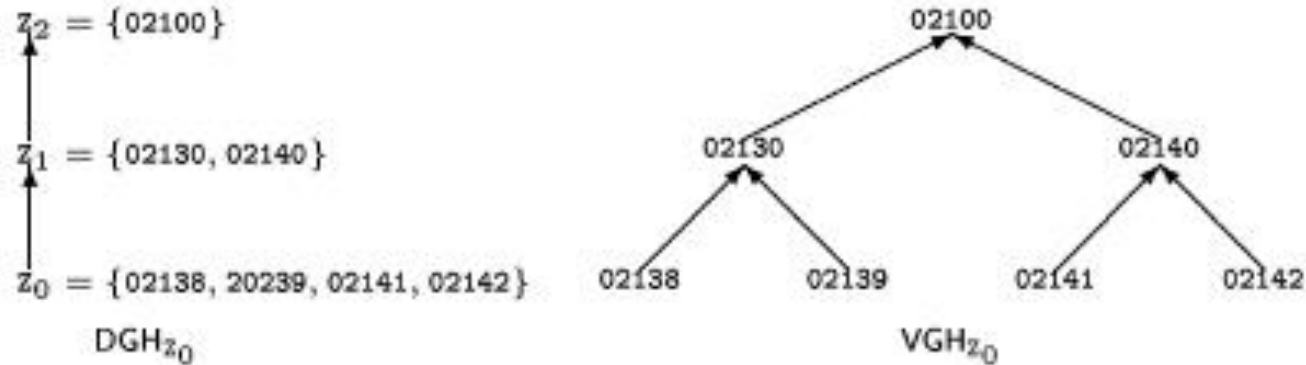
- Goal of k-Anonymity
  - Each record is indistinguishable from at least k-1 other records
  - These k records form an equivalence class
- **Generalization:** replace quasi-identifiers with less specific, but semantically consistent values



# Achieving k-Anonymity

- Generalization
  - Replace specific quasi-identifiers with less specific values until get  $k$  identical values
  - Partition ordered-value domains into intervals
- Suppression
  - When generalization causes too much information loss
    - This is common with “outliers”
- Lots of algorithms in the literature
  - Aim to produce “useful” anonymizations
    - ... usually without any clear notion of utility

# Generalization in Action



# Example of a k-Anonymous Table

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Figure 2 Example of  $k$ -anonymity, where  $k=2$  and  $Q=\{Race, Birth, Gender, ZIP\}$

# Example of Generalization (1)

Released table

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

External data Source

Name	Birth	Gender	ZIP	Race
Andre	1964	m	02135	White
Beth	1964	f	55410	Black
Carol	1964	f	90210	White
Dan	1967	m	02174	White
Ellen	1968	f	02237	White

By linking these 2 tables, you still don't learn Andre's problem

# Example of Generalization (2)

Microdata

QID			SA
Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

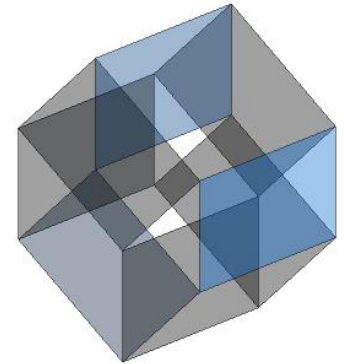
Generalized table

QID			SA
Zipcode	Age	Sex	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Prostate Cancer
4790*	[43,52]	*	Flu
4790*	[43,52]	*	Heart Disease
4790*	[43,52]	*	Heart Disease

- Released table is 3-anonymous
- If the adversary knows Alice's quasi-identifier (47677, 29, F), he still does not know which of the first 3 records corresponds to Alice's record

# Curse of Dimensionality

- Generalization fundamentally relies on **spatial locality**
  - Each record must have  $k$  close neighbors
- Real-world datasets are very sparse
  - Many attributes (dimensions)
    - Netflix Prize dataset: 17,000 dimensions
    - Amazon customer records: several million dimensions
  - “Nearest neighbor” is very far
- Projection to low dimensions loses all info  $\Rightarrow$   $k$ -anonymized datasets are useless



# HIPAA Privacy Rule

"Under the safe harbor method, covered entities must remove all of a list of 18 enumerated identifiers and **have no actual knowledge that the information remaining could be used, alone or in combination, to identify a subject of the information.**"

"The identifiers that must be removed include direct identifiers, such as name, street address, social security number, as well as other identifiers, such as birth date, admission and discharge dates, and five-digit zip code. The safe harbor requires removal of geographic subdivisions smaller than a State, except for the initial three digits of a zip code if the geographic unit formed by combining all zip codes with the same initial three digits contains more than 20,000 people. In addition, age, if less than 90, gender, ethnicity, and other demographic information not listed may remain in the information. The safe harbor is intended to provide covered entities with a simple, definitive method that does not require much judgment by the covered entity to determine if the information is adequately de-identified."



# Two (and a Half) Interpretations

- **Membership disclosure:** Attacker cannot tell that a given person in the dataset
- **Sensitive attribute disclosure:** Attacker cannot tell that a given person has a certain sensitive attribute
- **Identity disclosure:** Attacker cannot tell which record<sup>↑</sup> corresponds to a given person

This interpretation is correct, **assuming the attacker does not know anything other than quasi-identifiers**

**But this does not imply any privacy!**

Example: k clinical records, all HIV+

# Unsorted Matching Attack

- Problem: records appear in the same order in the released table as in the original table
- Solution: randomize order before releasing

Race	ZIP
Asian	02138
Asian	02139
Asian	02141
Asian	02142
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

PT

Race	ZIP
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

GT1

Race	ZIP
Asian	02130
Asian	02130
Asian	02140
Asian	02140
Black	02130
Black	02130
Black	02140
Black	02140
White	02130
White	02130
White	02140
White	02140

GT2

Figure 3 Examples of  $k$ -anonymity tables based on PT

# Complementary Release Attack

Different releases of the same private table can be linked together to compromise k-anonymity

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
person	1965	female	0213*	painful eye
person	1965	female	0213*	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	0213*	short of breath
person	1965	female	0213*	hypertension
white	1964	male	0213*	obesity
white	1964	male	0213*	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

GT1

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1960-69	male	02138	short of breath
white	1960-69	human	02139	hypertension
white	1960-69	human	02139	obesity
white	1960-69	human	02139	fever
white	1960-69	male	02138	vomiting
white	1960-69	male	02138	back pain

GT3

# Linking Independent Releases

Race	BirthDate	Gender	ZIP	Problem
black	9/20/1965	male	02141	short of breath
black	2/14/1965	male	02141	chest pain
black	10/23/1965	female	02138	painful eye
black	8/24/1965	female	02138	wheezing
black	11/7/1964	female	02138	obesity
black	12/1/1964	female	02138	chest pain
white	10/23/1964	male	02138	short of breath
white	3/15/1965	female	02139	hypertension
white	8/13/1964	male	02139	obesity
white	5/5/1964	male	02139	fever
white	2/13/1967	male	02138	vomiting
white	3/21/1967	male	02138	back pain

PT

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	02138	short of breath
white	1965	female	02139	hypertension
white	1964	male	02139	obesity
white	1964	male	02139	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

LT

# Attacks on k-Anonymity

k-Anonymity does not provide privacy if

- Sensitive values in an equivalence class lack diversity
- The attacker has background knowledge

Homogeneity attack

Bob	
<b>Zipcode</b>	<b>Age</b>
47678	27

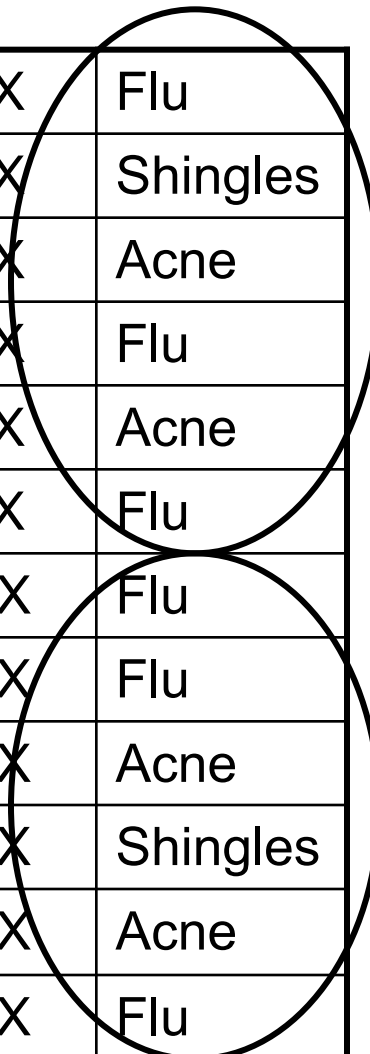
A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

Background knowledge attack

Carl	
<b>Zipcode</b>	<b>Age</b>
47673	36

# I-Diversity



Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Sensitive attributes must be “diverse” within each quasi-identifier equivalence class

# Distinct I-Diversity

- Each equivalence class has at least  $I$  well-represented sensitive values
- Doesn't prevent probabilistic inference attacks

...	<b>Disease</b>
	...
	HIV
	HIV
	...
	HIV
	pneumonia
	bronchitis
	...

10 records

8 records have HIV

2 records have other values

# Other Versions of l-Diversity

- Probabilistic l-diversity
  - The frequency of the most frequent value in an equivalence class is bounded by  $1/l$
- Entropy l-diversity
  - The entropy of the distribution of sensitive values in each equivalence class is at least  $\log(l)$
- Recursive  $(c,l)$ -diversity
  - $r_1 < c(r_l + r_{l+1} + \dots + r_m)$  where  $r_i$  is the frequency of the  $i^{\text{th}}$  most frequent value
  - Intuition: the most frequent value does not appear too frequently



# Neither Necessary, Nor Sufficient

Original dataset

...	Cancer
...	Cancer
...	Cancer
...	Flu
..	Cancer
...	Cancer
...	Cancer
...	Cancer
..	Cancer
...	Cancer
...	Cancer
...	Flu
...	Flu

99% have cancer

Anonymization A

Q1	Flu
Q1	Flu
Q1	Cancer
Q1	Flu
Q1	Cancer
Q1	Cancer
Q2	Cancer

Anonymization B

Q1	Flu
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q2	Cancer

99% cancer  $\Rightarrow$  quasi-identifier group is not “diverse”  
...yet anonymized database does not leak anything

50% cancer  $\Rightarrow$  quasi-identifier group is “diverse”  
**This leaks a ton of information**

# Limitations of l-Diversity

- Example: sensitive attribute is HIV+ (1%) or HIV- (99%)
  - Very different degrees of sensitivity!
- l-diversity is unnecessary
  - 2-diversity is unnecessary for an equivalence class that contains only HIV- records
- l-diversity is difficult to achieve
  - Suppose there are 10000 records in total
  - To have distinct 2-diversity, there can be at most  $10000 * 1\% = 100$  equivalence classes

# Skewness Attack

- Example: sensitive attribute is HIV+ (1%) or HIV- (99%)
- Consider an equivalence class that contains an equal number of HIV+ and HIV- records
  - Diverse, but potentially violates privacy!
- l-diversity does not differentiate:
  - Equivalence class 1: 49 HIV+ and 1 HIV-
  - Equivalence class 2: 1 HIV+ and 49 HIV-

**l-diversity does not consider overall distribution of sensitive values!**

# Sensitive Attribute Disclosure

Similarity attack

Bob	
<b>Zip</b>	<b>Age</b>
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

## Conclusion

1. Bob's salary is in [20k,40k], which is relatively low
2. Bob has some stomach-related disease

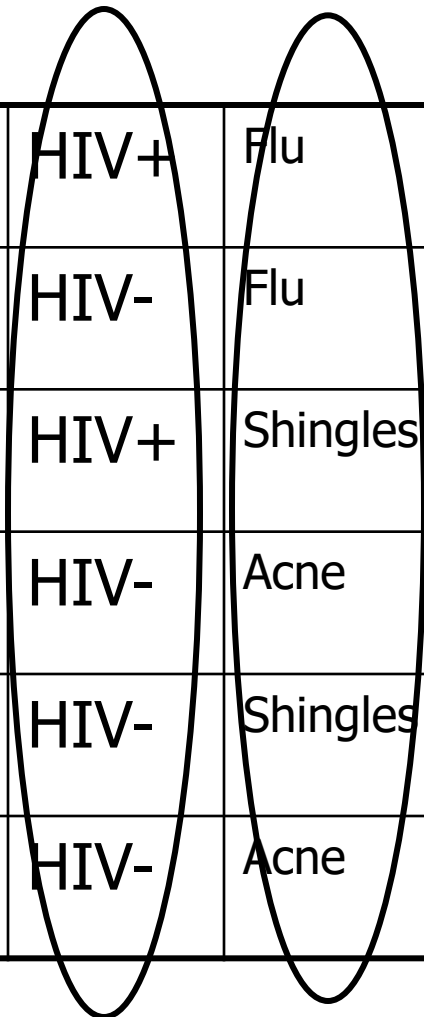
l-diversity does not consider semantics of sensitive values!

# t-Closeness

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Distribution of sensitive attributes within each quasi-identifier group should be “close” to their distribution in the entire original database

# Anonymous, “t-Close” Dataset



Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
Asian/AfrAm	787XX	HIV+	Shingles
Caucas	787XX	HIV-	Acne
Caucas	787XX	HIV-	Shingles
Caucas	787XX	HIV-	Acne

This is k-anonymous,  
l-diverse and t-close...  
...so secure, right?

# What Does Attacker Know?

*Bob is Caucasian and  
I heard he was  
admitted to hospital  
with flu...*



This is against the rules!  
"flu" is not a quasi-identifier

Yes... and this is yet another  
problem with k-anonymity

Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
Caucas	787XX	HIV+	Shingles
Caucas	787XX	HIV-	Acne
Caucas	787XX	HIV-	Shingles
Caucas	787XX	HIV-	Acne

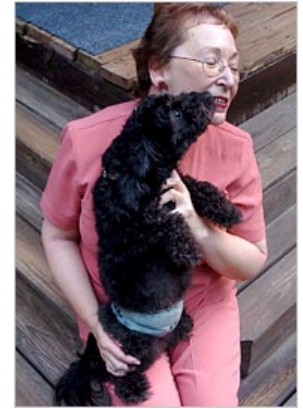
# AOL Privacy Debacle

- In August 2006, AOL released anonymized search query logs
  - 657K users, 20M queries over 3 months (March-May)
- Opposing goals
  - Analyze data for research purposes, provide better services for users and advertisers
  - Protect privacy of AOL users
    - Government laws and regulations
    - Search queries may reveal income, evaluations, intentions to acquire goods and services, etc.



# AOL User 4417749

- AOL query logs have the form  
<AnonID, Query, QueryTime, ItemRank, ClickURL>
  - ClickURL is the truncated URL
- NY Times re-identified AnonID 4417749
  - Sample queries: “numb fingers”, “60 single men”, “dog that urinates on everything”, “landscapers in Lilburn, GA”, several people with the last name Arnold
    - Lilburn area has only 14 citizens with the last name Arnold
  - NYT contacts the 14 citizens, finds out AOL User 4417749 is 62-year-old Thelma Arnold



# Further Examples

- Attacks: many successful attacks identified individual users
  - Ego-surfers: people typed in their own names
  - Zip codes and town names identify an area
- Consequences: CTO resigned, two researchers fired
  - Well-intentioned effort failed due to inadequate anonymization

# k-Anonymity Considered Harmful

- Syntactic
  - Focuses on data transformation, not on what can be learned from the anonymized dataset
  - “k-anonymous” dataset can leak sensitive information
- “Quasi-identifier” fallacy
  - Assumes a priori that attacker will not know certain information about his target
- Relies on locality
  - Destroys utility of many real-world datasets

# Issues with Syntactic Definitions

- What adversary do they apply to?
  - Do not consider adversaries with side information
  - Do not consider probability
  - Do not consider adversarial algorithms for making decisions (inference)
- Any attribute is a potential quasi-identifier
  - External / auxiliary / background information about people is very easy to obtain

# Classical Intuition for Privacy

Dalenius (1977): “If the release of statistics  $S$  makes it possible to determine the value [of private information] *more accurately* than is possible without access to  $S$ , a disclosure has taken place”

- Privacy means that anything that can be learned about a respondent from the statistical database can be learned without access to the database

# Problems with Classic Intuition

- Popular interpretation: prior and posterior views about an individual shouldn't change "too much"
  - What if my (incorrect) prior is that every student has three arms?
- How much is "too much?"
  - Can't achieve cryptographically small levels of disclosure and keep the data useful
  - Adversarial user is supposed to learn unpredictable things about the database

# Absolute Guarantee Unachievable

- Privacy: for some definition of “privacy breach”,  
 $\forall$  distribution on databases,  $\forall$  adversaries  $A$ ,  $\exists A'$   
such that  $\Pr(A(\text{San})=\text{breach}) - \Pr(A'()=\text{breach}) \leq \varepsilon$ 
  - For reasonable “breach”, if  $\text{San}(\text{DB})$  contains information about DB, then some adversary breaks this definition
- Example
  - I know that you are 2 inches taller than the average Russian
  - DB allows computing average height of a Russian
  - This DB breaks your privacy according to this definition... even if your record is not in the database!

# Generalization at Runtime

- Only allow for aggregate queries, e.g., sum, average etc.
- Given the table:

QID			SA
Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

- What happens if I query:
  - SELECT avg(age) from Table WHERE Disease='Flu'
- Or:
  - SELECT avg(age) from Table WHERE Disease in (Ovarian Cancer, Flu)
  - SELECT avg(age) from Table WHERE Disease in (Ovarian Cancer)



# Differential Privacy

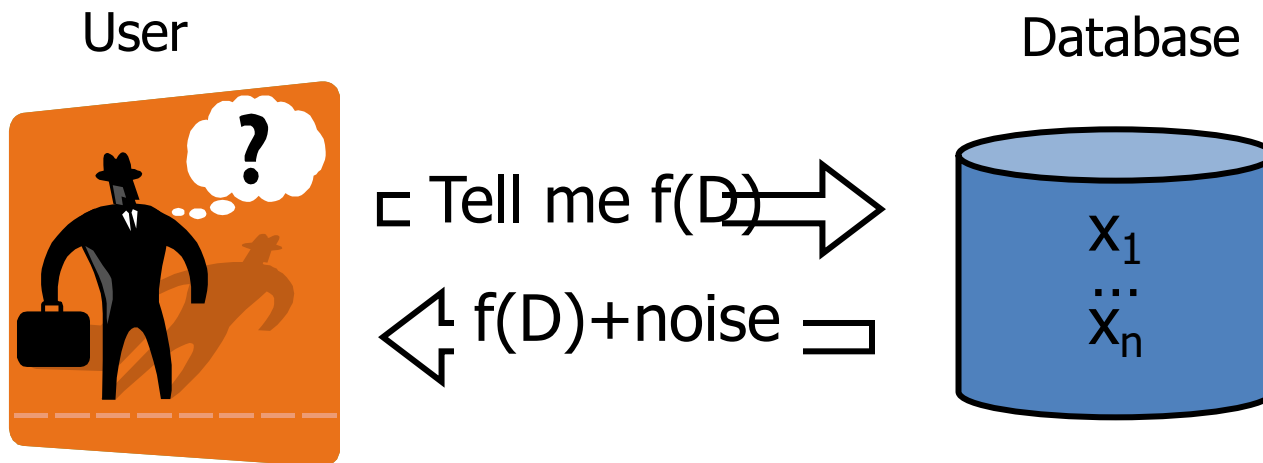
- Limitation of previous privacy notions:
  - Requires identifying which attributes are quasi-identifier or sensitive, not always possible
  - Difficult to pin down due to background knowledge
  - Syntactic in nature (property of anonymized dataset)
    - Not exhaustive in inference prevented
- Differential Privacy
  - Privacy is not violated if one's information is not included
  - Output does not overly depend on any single tuple

# Variants of Differential Privacy

- Bounded Differential Privacy:  $D$  and  $D'$  are neighbors if and only if  $D'$  can be obtained from  $D$  by replacing one tuple with another tuple
  - $D$  and  $D'$  have the same number of tuples
  - Revealing size of dataset does not affect privacy
- Unbounded Differential Privacy:  $D$  and  $D'$  are neighbors if and only if  $D'$  can be obtained from  $D$  by adding or removing one tuple
  - The numbers of tuples in  $D$  and  $D'$  differ by 1
- In most cases, can use either one.

# Add Noise to Output

- Intuition:  $f(D)$  can be released accurately when  $f$  is insensitive to individual entries  $x_1, \dots, x_n$
- Global sensitivity  $GS_f = \max_{\text{neighbors } D, D'} ||f(D) - f(D')||_1$ 
  - Example:  $GS_{\text{average}} = 1/n$  for sets of numbers between 0 and 1
- Theorem:  $f(x) + \text{Lap}(GS_f / \epsilon)$  is  $\epsilon$ -indistinguishable
  - Noise generated from Laplace distribution



# Exponential Mechanism

- The goal is to output  $f(D)$ ;  $f(D) \in R$ 
  - E.g., which item is purchased the most frequently
- Define a quality function  $q(D, r \in R)$ 
  - which gives a real number describing the desirability of outputting  $r$  on input dataset  $D$
- Compute the sensitivity of the quality function
  - $\Delta q = \max_r \max_{D, D'} |q(D, r) - q(D', r)|$
- Returns  $r$  with probability  
proportional to  $\exp(q(D, r) / 2\varepsilon \Delta q)$   
satisfies  $\varepsilon$ -DP

# Adding noise

- The stream of numbers above is applied to the result set.
- While masking the individuals, it allows accurate percentages and trending.
- Presuming the magnitude is small (i.e. small error), the numbers are themselves accurate within an acceptable margin.

Category	Value
A	36
B	22
...	...
N	102

noise

Category	Value
A	34
B	23
...	...
N	108

# Windows Live User Data

- Case study is based on Windows Live user data:
  - 550 million Passport users
  - Passport has web site visitor self-reported data: gender, birth date, occupation, country, zip code, etc.
  - Web data has: IP address, pages viewed, page view duration, browser, operating system, etc.
- Created two groups for this case study to study the acceptability / applicability of differential privacy within the WL reporting context:
  - WL Sampled Users Web Analytics
  - Customer Churn Analytics

# Windows Live Example Report

As per below, you can see the effect on the data

Country	Unknown	Very Low	Low	Moderate	High
afghanistan	121561	11277	3853	3985	18107
albania	557376	70422	30895	30289	117330
algeria	444665	50614	14928	14943	47312
american samoa	36963	3373	1150	1130	5612
andorra	30568	4142	1541	1514	7767
angola	71292	4658	1838	1911	7073
anguilla	9003	981	416	490	2479
antarctica	26340	2549	839	911	4377

Country	Unknown	Very Low	Low	Moderate	High
afghanistan	121559	11281	3852	3984	18107
albania	557374	70420	30896	30289	117329
algeria	444663	50615	14927	14946	47313
american samoa	36962	3373	1149	1131	5612
andorra	30567	4144	1541	1516	7763
angola	71291	4659	1835	1910	7079
anguilla	9004	981	416	492	2478
antarctica	26340	2549	834	909	4376