Data Anonymization – Introduction

Li Xiong

CS573 Data Privacy and Security

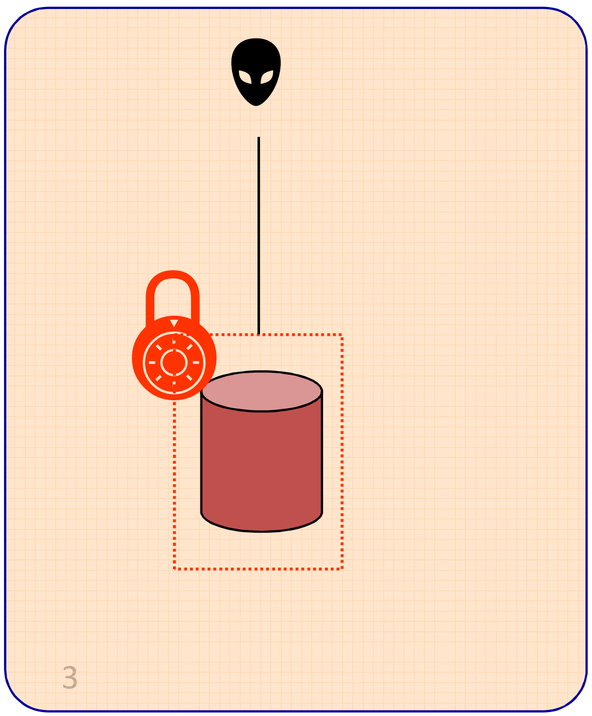
Outline

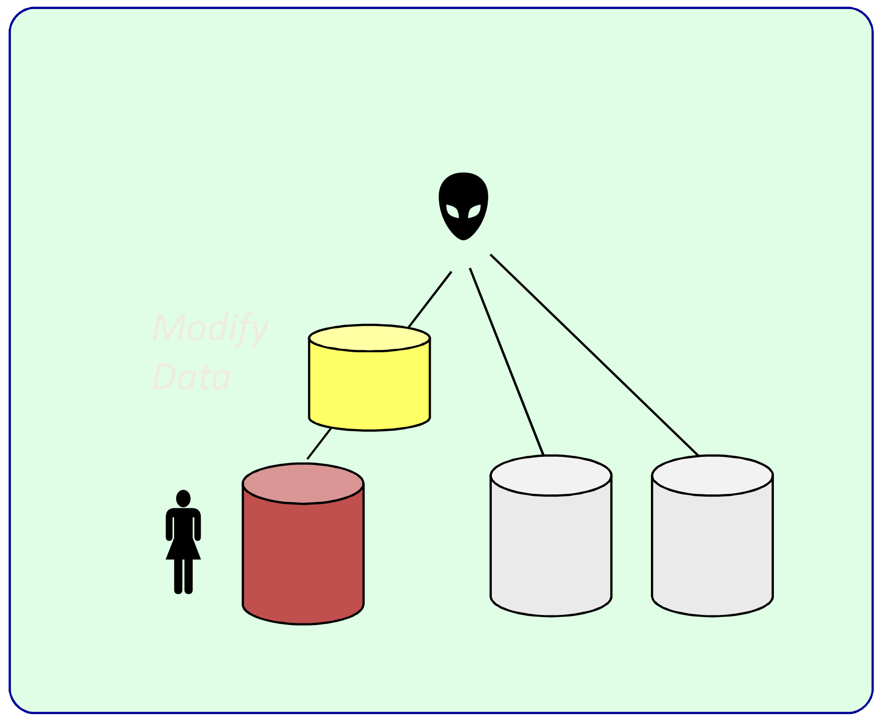
- **Problem definition**
- Principles
- Disclosure Control Methods

Inference Control

- Access control: protecting information and information systems from unauthorized access and use.
- Inference control: protecting private data while publishing useful information

NO FOUL PLAY





Problem: Disclosure Control

- **Disclosure Control** is the discipline concerned with the modification of data, containing confidential information about individual entities such as persons, households, businesses, etc. in order to prevent third parties working with these data to recognize individuals in the data
- Privacy preserving data publishing, anonymization, de-identification

Types of disclosure

- **Identity disclosure** - identification of an entity (person, institution)
- **Attribute disclosure** - the intruder finds something new about the target person
- **Disclosure** – identity, attribute disclosure or both.

Microdata and External Information

- **Microdata** represents a series of records, each record containing information on an individual unit such as a person, a firm, an institution, etc
  - In contrast to computed tables (Macrodata)

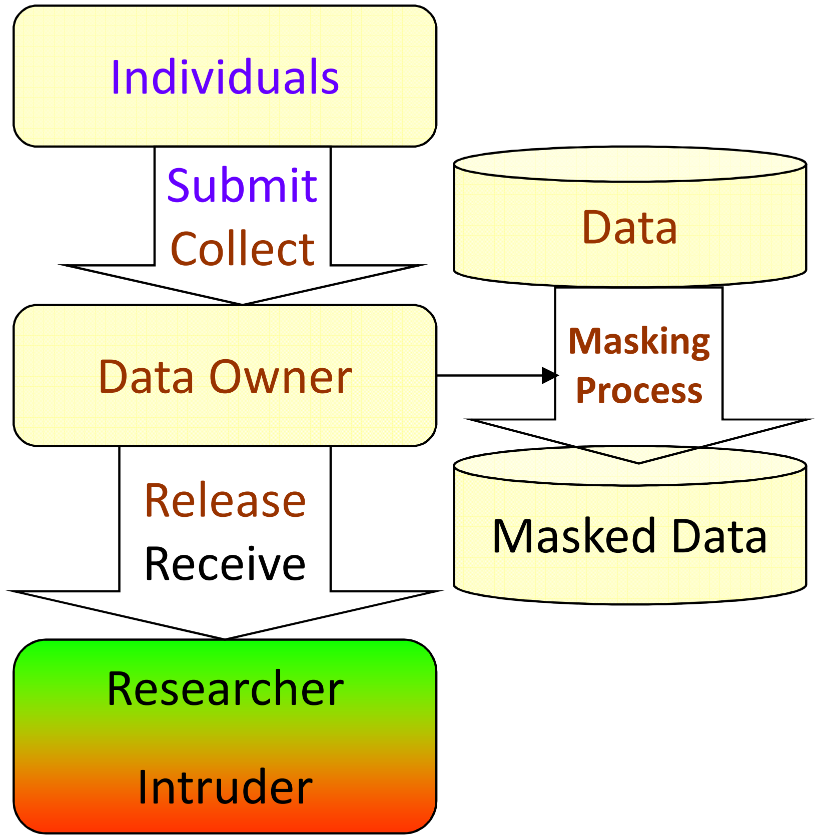
- **Masked Microdata** names and other identifying information are removed or modified from microdata

- **External Information** any known information by a presumptive intruder related to some individuals from initial microdata

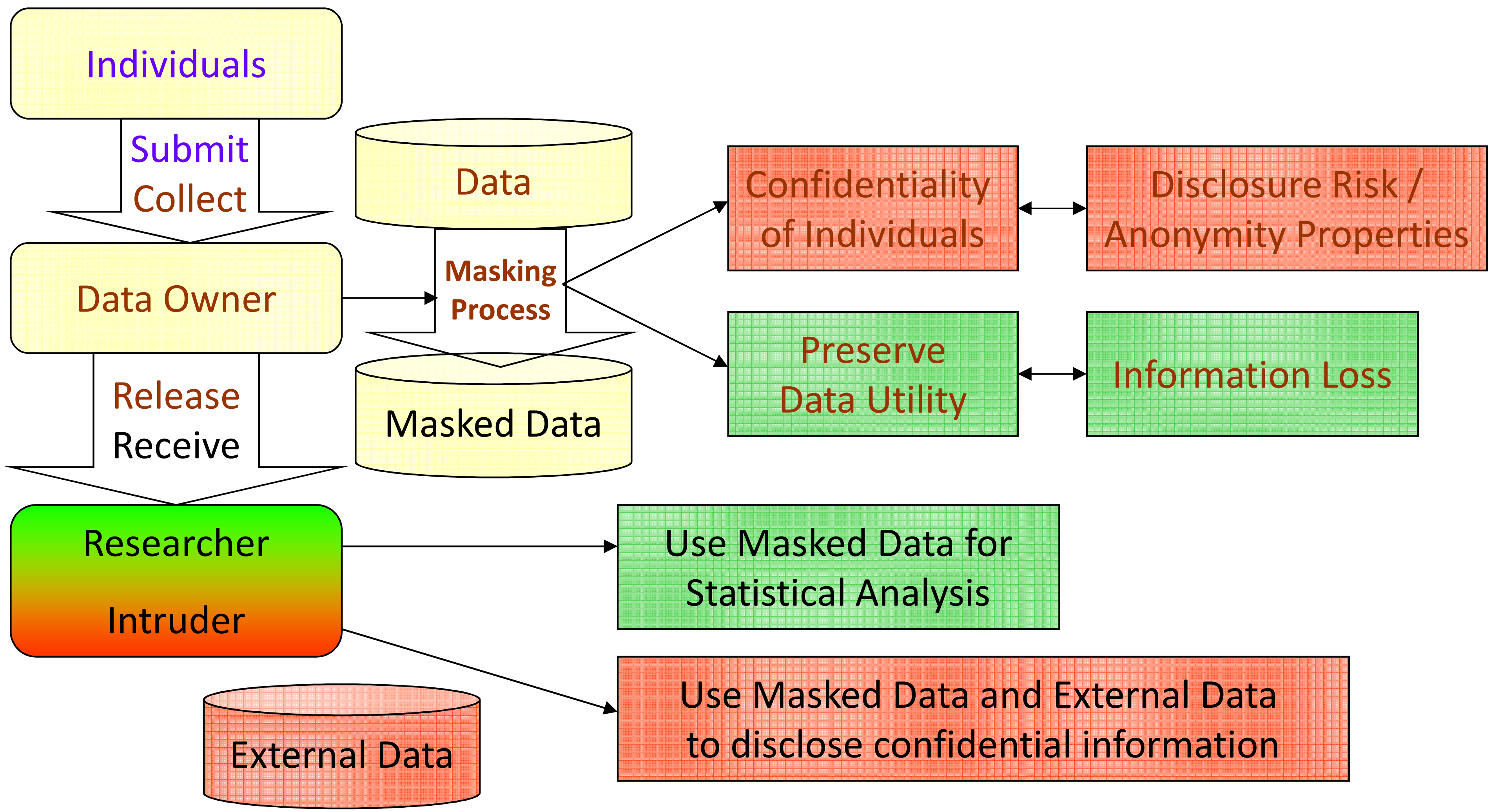
Disclosure Risk and Information Loss

- **Disclosure risk** - the risk that a given form of disclosure will arise if a masked microdata is released
- **Information loss** - the quantity of information which exist in the initial microdata but not in masked microdata due to disclosure control methods

Disclosure Control Problem



Disclosure Control Problem



Disclosure Control for Tables vs. Microdata

- Microdata
- Macrodata - precomputed statistics tables

Name	Age	Diagnosis	Income
Wayne	44	AIDS	45,500
Gore	44	Asthma	37,900
Banks	55	AIDS	67,000
Casey	44	Asthma	21,000
Stone	55	Asthma	90,000
Kopi	45	Diabetes	48,000
Simms	25	Diabetes	49,000
Wood	35	AIDS	66,000
Aaron	55	AIDS	69,000
Pall	45	Tuberculosis	34,000

Initial Microdata

Age	Diagnosis	Income
44	AIDS	50,000
44	Asthma	40,000
55	AIDS	70,000
44	Asthma	20,000
55	Asthma	90,000
45	Diabetes	50,000
-	Diabetes	50,000
-	AIDS	70,000
55	AIDS	70,000
45	-	30,000

Masked Microdata

Disclosure Control For Microdata

Disclosure Control for Tables

Name	Age	Diagnosis	Income
Wayne	44	AIDS	45,500
Gore	44	Asthma	37,900
Banks	55	AIDS	67,000
Casey	44	Asthma	21,000
Stone	55	Asthma	90,000
Kopi	45	Diabetes	48,000
Simms	25	Diabetes	49,000
Wood	35	AIDS	66,000
Aaron	55	AIDS	69,000
Pall	45	Tuberculosis	34,000

Initial Microdata

Count	Diagnosis
4	AIDS
3	Asthma
2	Diabetes
1	Tuberculosis

Table 1 - Count Diagnosis

Count	Age	Income
1	<= 30	49,000
1	31- 40	66,000
5	41 - 50	188,200
3	51-60	226,000
0	> 60	0

Table 2 - Total Incoming

Tables

Count	Diagnosis
4	AIDS
3	Asthma

Masked Table 1

Count	Age	Income
5	31 - 40	188,200
3	41 - 50	226,000

Masked Table 2

Masked Tables from Tables

Anonymization

- Microdata release
  - Guidelines
  - Cases and controversies
  - Current research
- Macrodata release

HIPAA Privacy Regulation

- De-identification Standards for Health Information in Research
  - a. Safe Harbor
  - b. Statistician Method
  - c. Limited Data Set

HIPAA

- **Protected health information (PHI):**
  - Individually identifiable health information (IIHI = Health Information + Identifier) that is transmitted or maintained electronically, or transmitted or maintained in any other form or medium
- **De-identified Health Information**: health information that does **not** identify an individual and with respect to which there is **no reasonable basis** to believe that the information can be used to identify an individual
  - Once de-identified, the data is out of the Privacy Rule.

HIPAA De-identification Standards

- Two methods for the de-identification of health information:
  - "**Safe Harbor**" -- remove 18 specified identifiers - intended to provide a simple, definitive method for de-identifying health information with protection from litigation
  - "**Statistician Method**" -- retain some of the 18 safe harbor's specified identifiers and demonstrate the standard is met if person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods, e.g., a Biostatistician, makes and documents that the risk of re-identification is very small.

Limited Data Set

- Final rule: added another method requiring removal of facial identifiers -- "Limited Data Set"
  - Under confidentiality agreements - for research, public health, and health care operations
  - Regarded as PHI - NOT de-identified
    - therefore, still subject to Privacy Rule requirements such as minimum necessary rule.

Safe Harbor's 18 Identifiers

- **Names**
- **All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes**
  - **Except for the initial three digits of a zip code if according to the currently available data from the Bureau of the Census:**
    - **The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and**
    - **The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people are changed to 000;**
- **All elements of dates (except year) or dates directly relating to an individual, including:**
  - **birth date, admission date, discharge date, date of death;**
  - **and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;**
- **Telephone numbers;**
- **Fax numbers;**
- **Electronic mail addresses;**
- **Social security numbers;**
- **Medical record numbers;**
- **Health plan beneficiary numbers;**
- **Account numbers;**
- **Certificate/license numbers;**
- **Vehicle identifiers and serial numbers, including license plate numbers;**
- **Device identifiers and serial numbers;**
- **Web Universal Resource Locators (URLs);**
- **Internet Protocol (IP) address numbers;**
- **Biometric identifiers, including finger and voice prints;**
- **Full face photographic images and any comparable images; and**
- **Any other unique identifying number, characteristic, or code.**

Statistician Method

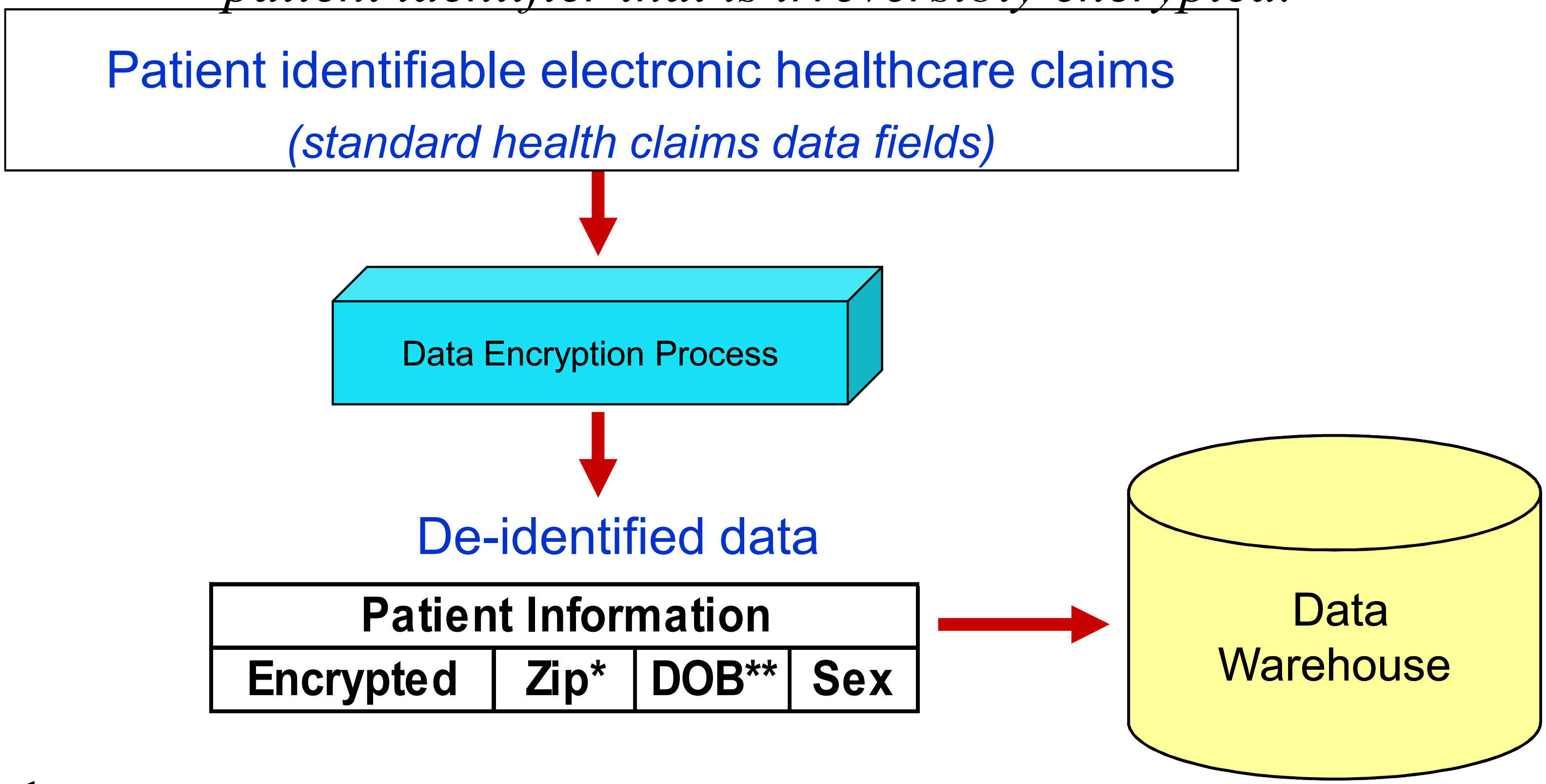
- Statistician must
  - determine that there is a “very small risk” of re-identification
  - after applying “generally accepted statistical and scientific principles and methods for rendering information not individually identifiable”
  - documents the methods and results of the analysis that justify such determination.

Limited Data Set

- For research, public health, or health care operations purposes
- Authorization not required
- A limited data use agreement must be in place between the covered entity and the recipient of limited data set (LDS)

Ensuring HIPAA Compliance

All data handled is de-identified using a unique patient identifier that is irreversibly encrypted.



* zip = 3 digit
** DOB = modified

Upon completion of the de-identification process a unique patient identifier is created, which is irreversibly encrypted.

Anonymization

- Microdata release
  - Guidelines
  - Cases and controversies
  - Current research
- Macrodata release

Massachusetts GIC Incident

- Massachusetts GIC released "anonymized" data on state employees' hospital visit
- Then Governor William Weld assured public on privacy

GIC

Name	SSN	Birth date	Zip	Diagnosis
Alice	123456789	44	48202	AIDS
Bob	323232323	44	48202	AIDS
Charley	232345656	44	48201	Asthma
Dave	333333333	55	48310	Asthma
Eva	666666666	55	48310	Diabetes

Anonymized

Birth date	Zip	Diagnosis
44	48202	AIDS
44	48202	AIDS
44	48201	Asthma
55	48310	Asthma
55	48310	Diabetes

Massachusetts GIC

- Then graduate student Sweeney linked the data with Voter roller in Cambridge and identified Governor Weld's record

Name	SSN	Birth date	Zip	Diagnosis	Income
Alice	123456789	44	48202	AIDS	17,000
Bob	323232323	44	48202	AIDS	68,000
Charley	232345656	44	48201	Asthma	80,000
Dave	333333333	55	48310	Asthma	55,000
Eva	666666666	55	48310	Diabetes	23,000

Birthd ata	Zip	Diagnosis	Income
44	48202	AIDS	17,000
44	48202	AIDS	68,000
44	48201	Asthma	80,000
55	48310	Asthma	55,000
55	48310	Diabetes	23,000

Voter roll for Cambridge

Name	Birth date	Zip
Alice	44	48202
Charley	44	48201
Dave	55	48310

Re-identification

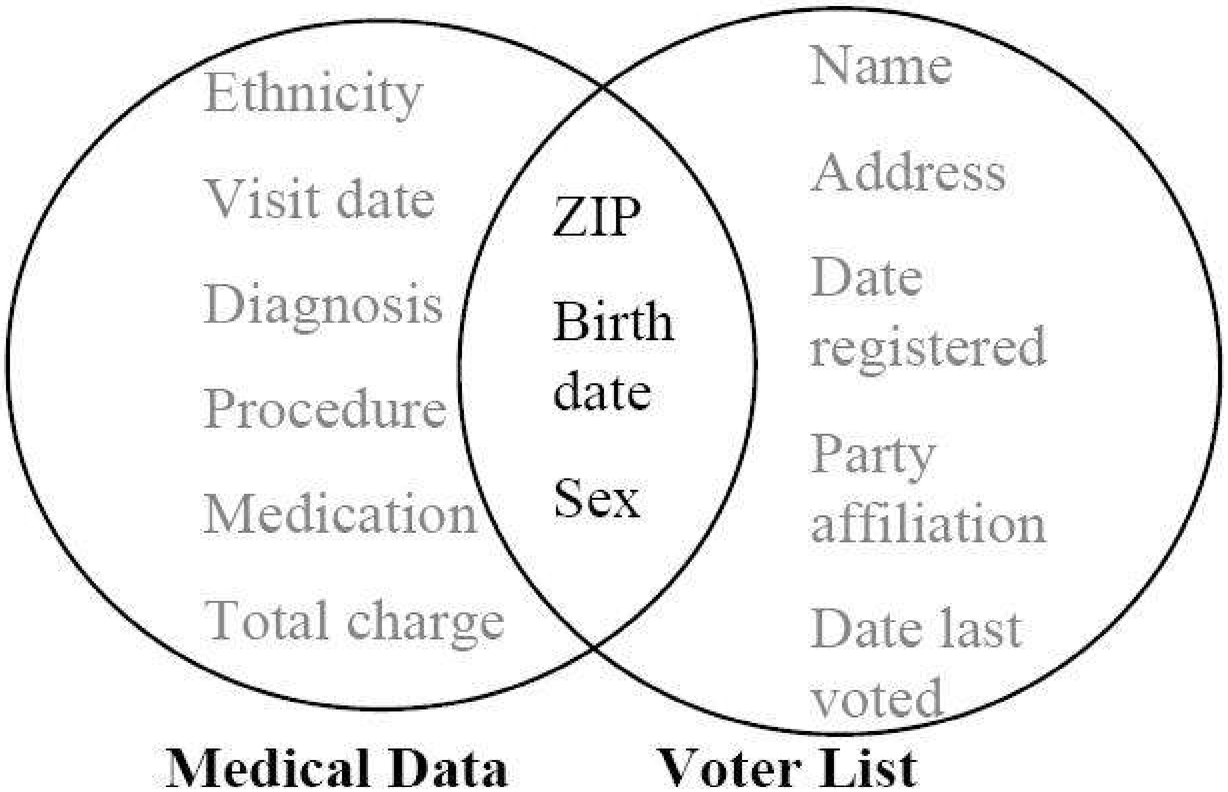


Figure 1 Linking to re-identify data

AOL Query Log Release

20 million Web search queries by AOL

AnonID	Query	QueryTime	ItemRank	ClickURL
217	lottery	2006-03-01 11:58:51	1	http://www.calottery.com
217	lottery	2006-03-27 14:10:38	1	http://www.calottery.com
1268	gall stones	2006-05-11 02:12:51		
1268	gallstones	2006-05-11 02:13:02	1	http://www.niddk.nih.gov
1268	ozark horse blankets	2006-03-01 17:39:28	8	http://www.blanketsnmore.com

(Source: AOL Query Log)

User No. 4417749

- User 4417749
  - “numb fingers”,
  - “60 single men”
  - “dog that urinates on everything”
  - “landscapers in Lilburn, Ga”
  - Several people names with last name Arnold
  - “homes sold in shadow lake subdivision gwinnett county georgia”

User No. 4417749

- User 4417749
  - “numb fingers”,
  - “60 single men”
  - “dog that urinates on everything”
  - “landscapers in Lilburn, Ga”
  - Several people names with last name Arnold
  - “homes sold in shadow lake subdivision gwinnett county georgia”

Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends’ medical ailments and loves her dogs

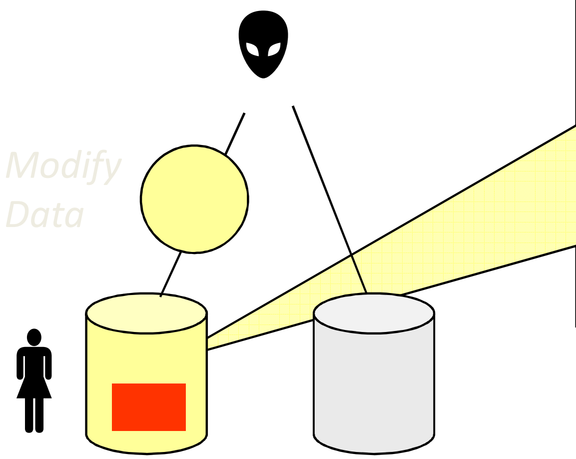
Anonymization

- Microdata release
  - Guidelines
  - Cases and controversies
  - Current research
    - Principles
    - Anonymization methods
- Macrodata release

K-Anonymity

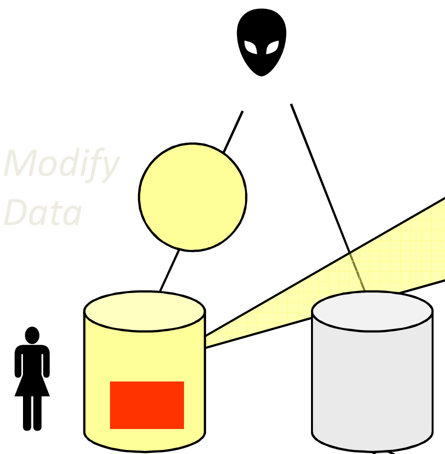
- The term was introduced in 1998 by Samarati and Sweeney.
- Important papers:
  - Sweeney L. (2002), *K-Anonymity: A Model for Protecting Privacy,* International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10, No. 5, 557-570
  - Sweeney L. (2002), *Achieving K-Anonymity Privacy Protection using Generalization and Suppression,* International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10, No. 5, 571-588
  - Samarati P. (2001), *Protecting Respondents Identities in Microdata Release,* IEEE Transactions on Knowledge and Data Engineering, Vol. 13, No. 6, 1010-1027
- Many new research papers in the last 10 years
  - Theoretical results
  - Many algorithms achieving k-anonymity
  - Many improved principles and algorithms

Motivating Example



	Non-Sensitive Data			*Sensitive Data*	
#	**Zip**	**Age**	**Nationality**	**Name**	**Condition**
1	13053	28	Brazilian	Ronaldo	Heart Disease
2	13067	29	US	Bob	Heart Disease
3	13053	37	Indian	Kumar	Cancer
4	13067	36	Japanese	Umeko	Cancer

Motivating Example (continued)



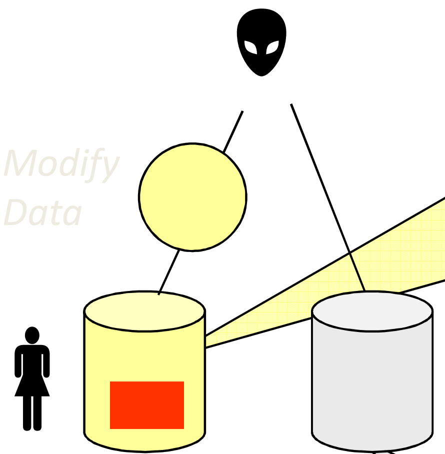
Published Data: Alice publishes data without the Name

	Non-Sensitive Data			*Sensitive Data*
#	**Zip**	**Age**	**Nationality**	**Condition**
1	13053	28	Brazilian	Heart Disease
2	13067	29	US	Heart Disease
3	13053	37	Indian	Cancer
4	13067	36	Japanese	Cancer

Attacker’s Knowledge: Voter registration list

#	**Name**	**Zip**	**Age**	**Nationality**
1	John	13067	45	US
2	Paul	13067	22	US
3	Bob	13067	29	US
4	Chris	13067	23	US

Motivating Example (continued)



Published Data: Alice publishes data without the Name

	Non-Sensitive Data			*Sensitive Data*
#	***Zip***	***Age***	***Nationality***	***Condition***
1	13053	28	Brazilian	Heart Disease
2	13067	29	US	**Heart Disease**
3	13053	37	Indian	Cancer
4	13067	36	Japanese	Cancer

Attacker's Knowledge: Voter registration list

#	***Name***	***Zip***	***Age***	***Nationality***
1	John	13067	45	US
2	Paul	13067	22	US
3	**Bob**	13067	29	US
4	Chris	13067	23	US

Data Leak !

Source of the Problem

Even if we do not publish the individuals:

- There are some fields that may *uniquely* identify some individual

	Non-Sensitive Data			*Sensitive Data*
#	***Zip***	***Age***	***Nationality***	***Condition***
…	…	…	…	…

$\underbrace{\text{Zip} \quad \text{Age} \quad \text{Nationality}}_{\text{Quasi Identifier}}$

- The attacker can use them to *join* with other sources and identify the individuals

Attribute Classification

- $I_1, I_2, \ldots, I_m$ - **identifier** attributes
  - Ex: *Name* and *SSN*
  - Information that leads to a specific entity.
- $K_1, K_2, \ldots., K_p$ - **key** attributes (***quasi-identifiers***)
  - Ex: *Zip Code* and *Age*
  - May be known by an intruder.
- $S_1, S_2, \ldots., S_q$ - **confidential** attributes
  - Ex: *Principal Diagnosis* and *Annual Income*
  - Assumed to be unknown to an intruder.

Attribute Types

- Identifier, Key (Quasi-Identifiers) and Confidential Attributes

RecID	Name	SSN	Age	State	Diagnosis	Income	Billing
1	John Wayne	123456789	44	MI	AIDS	45,500	1,200
2	Mary Gore	323232323	44	MI	Asthma	37,900	2,500
3	John Banks	232345656	55	MI	AIDS	67,000	3,000
4	Jesse Casey	333333333	44	MI	Asthma	21,000	1,000
5	Jack Stone	444444444	55	MI	Asthma	90,000	900
6	Mike Kopi	666666666	45	MI	Diabetes	48,000	750
7	Angela Simms	777777777	25	IN	Diabetes	49,000	1,200
8	Nike Wood	888888888	35	MI	AIDS	66,000	2,200
9	Mikhail Aaron	999999999	55	MI	AIDS	69,000	4,200
10	Sam Pall	100000000	45	MI	Tuberculosis	34,000	3,100

K-Anonymity Definition

- The *k-anonymity property* for a masked microdata (MM) is satisfied if with respect to Quasi-identifier set (QID) if every count in the frequency set of MM with respect to QID is greater or equal to k

K-Anonymity Example

RecID	Age	Zip	Sex	Illness
1	50	41076	Male	AIDS
2	30	41076	Female	Asthma
3	30	41076	Female	AIDS
4	20	41076	Male	Asthma
5	20	41076	Male	Asthma
6	50	41076	Male	Diabetes

- QID = { Age, Zip, Sex }
- **SELECT COUNT(*) FROM Patient GROUP BY Sex, Zip, Age;**
- If the results include groups with count less than k, the relation Patient does not have k-anonymity property with respect to QID.

Homogeneity Attack

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

k-Anonymity can create groups that leak information due to lack of diversity in sensitive attribute.

Anonymization

- Microdata release
  - Guidelines
  - Cases and controversies
  - Current research
    - Principles
    - Anonymization methods
- Macrodata release

L-diversity

- Each equivalence group must have l “well-represented” sensitive values

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

More attacks and principles

- t-closeness – skewed data
- m-variance – incremental releases
- …

Disclosure Control Techniques

- **Remove Identifiers**
- ***Generalization***
- ***Suppression***
- **Sampling**
- **Microaggregation**
- Perturbation / randomization
- Rounding
- **Data Swapping**
- Etc.

Disclosure Control Techniques

- Different disclosure control techniques are applied to the following initial microdata:

RecID	Name	SSN	Age	State	Diagnosis	Income	Billing
1	John Wayne	123456789	44	MI	AIDS	45,500	1,200
2	Mary Gore	323232323	44	MI	Asthma	37,900	2,500
3	John Banks	232345656	55	MI	AIDS	67,000	3,000
4	Jesse Casey	333333333	44	MI	Asthma	21,000	1,000
5	Jack Stone	444444444	55	MI	Asthma	90,000	900
6	Mike Kopi	666666666	45	MI	Diabetes	48,000	750
7	Angela Simms	777777777	25	IN	Diabetes	49,000	1,200
8	Nike Wood	888888888	35	MI	AIDS	66,000	2,200
9	Mikhail Aaron	999999999	55	MI	AIDS	69,000	4,200
10	Sam Pall	100000000	45	MI	Tuberculosis	34,000	3,100

Remove Identifiers

- Identifiers such as Names, SSN etc. are removed

RecID	Age	State	Diagnosis	Income	Billing
1	44	MI	AIDS	45,500	1,200
2	44	MI	Asthma	37,900	2,500
3	55	MI	AIDS	67,000	3,000
4	44	MI	Asthma	21,000	1,000
5	55	MI	Asthma	90,000	900
6	45	MI	Diabetes	48,000	750
7	25	IN	Diabetes	49,000	1,200
8	35	MI	AIDS	66,000	2,200
9	55	MI	AIDS	69,000	4,200
10	45	MI	Tuberculosis	34,000	3,100

Sampling

- Sampling is the disclosure control method in which only a subset of records is released
- If n is the number of elements in initial microdata and t the released number of elements we call $sf = t / n$ the sampling factor
- Simple random sampling is more frequently used. In this technique, each individual is chosen entirely by chance and each member of the population has an equal chance of being included in the sample

RecID	Age	State	Diagnosis	Income	Billing
5	55	MI	Asthma	90,000	900
4	44	MI	Asthma	21,000	1,000
8	35	MI	AIDS	66,000	2,200
9	55	MI	AIDS	69,000	4,200
7	25	IN	Diabetes	49,000	1,200

Microaggregation

- Order records from the initial microdata by an attribute, create groups of consecutive values, replace those values by the group average
- Microaggregation for attribute Income and minimum size 3
- The total sum for all Income values remains the same.

RecID	Age	State	Diagnosis	Income	Billing
2	44	MI	Asthma	30,967	2,500
4	44	MI	Asthma	30,967	1,000
10	45	MI	Tuberculosis	30,967	3,100
1	44	MI	AIDS	47,500	1,200
6	45	MI	Diabetes	47,500	750
7	25	IN	Diabetes	47,500	1,200
3	55	MI	AIDS	73,000	3,000
5	55	MI	Asthma	73,000	900
8	35	MI	AIDS	73,000	2,200
9	55	MI	AIDS	73,000	4,200

Data Swapping

- In this disclosure method a sequence of so-called elementary swaps is applied to a microdata
- An elementary swap consists of two actions:
  - A random selection of two records i and j from the microdata
  - A swap (interchange) of the values of the attribute being swapped for records i and j

RecID	Age	State	Diagnosis	Income	Billing
1	44	MI	AIDS	48,000	1,200
2	44	MI	Asthma	37,900	2,500
3	55	MI	AIDS	67,000	3,000
4	44	MI	Asthma	21,000	1,000
5	55	MI	Asthma	90,000	900
6	45	MI	Diabetes	45,500	750
7	25	IN	Diabetes	49,000	1,200
8	35	MI	AIDS	66,000	2,200
9	55	MI	AIDS	69,000	4,200
10	45	MI	Tuberculosis	34,000	3,100

Generalization and Suppression

- • Generalization
  - Replace the value with a less specific but semantically consistent value

- Suppression
  - Do not release a value at all

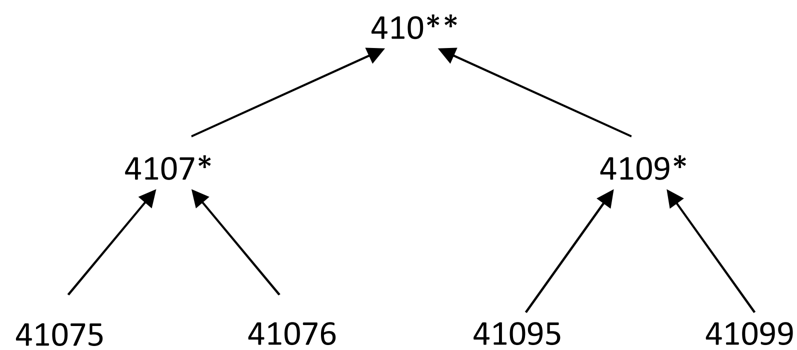
#	Zip	Age	Nationality	Condition
1	41076	< 40	*	Heart Disease
2	48202	< 40	*	Heart Disease
3	41076	< 40	*	Cancer
4	48202	< 40	*	Cancer

Domain and Value Generalization Hierarchies

Z2 = {410**}

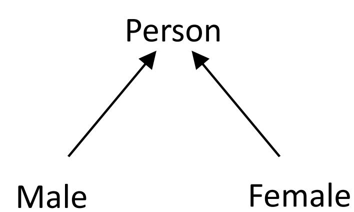
Z1 = {4107*. 4109*}

Z0 = {41075, 41076, 41095, 41099}



S1 = {Person}

S0 = {Male, Female}



Generalization Lattice

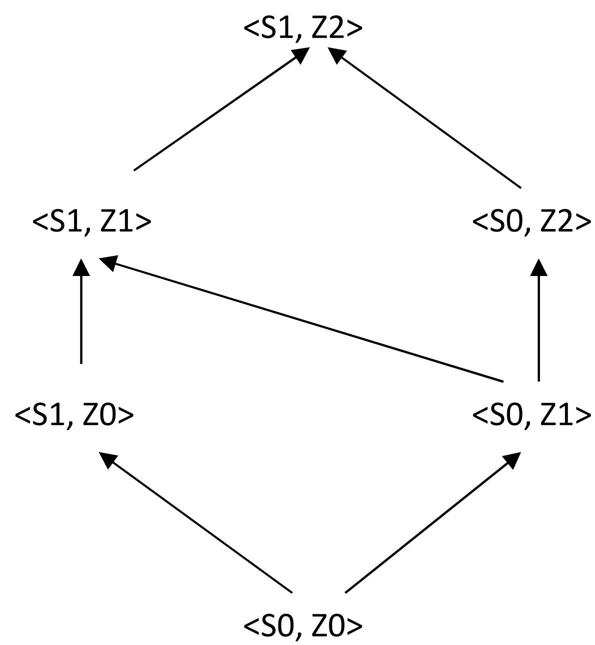
S1 = {Person}

S0 = {Male, Female}

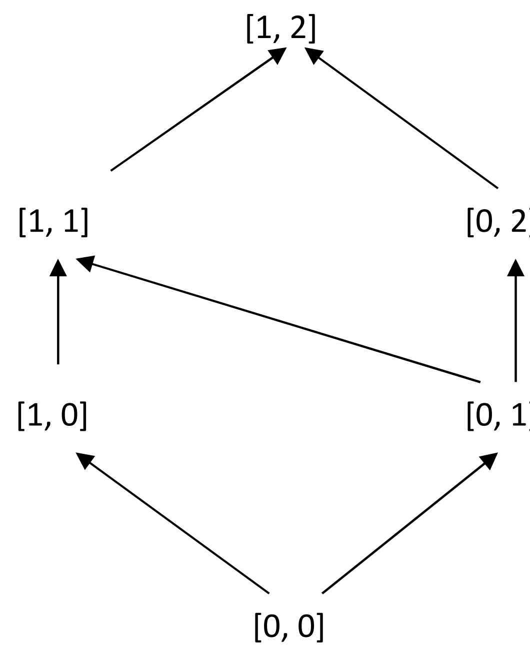
Z2 = {410**}

Z1 = {4107*, 4109*}

Z0 = {41075, 41076, 41095, 41099}



Generalization Lattice



Distance Vector Generalization Lattice

Generalization Tables

Race	Birth	Gender	ZIP	Problem
Black	1965	m	0214*	short breath
Black	1965	m	0214*	chest pain
Black	1965	f	0213*	hypertension
Black	1965	f	0213*	hypertension
Black	1964	f	0213*	obesity
Black	1964	f	0213*	chest pain
White	1964	m	0213*	chest pain
White	1964	m	0213*	obesity
White	1964	m	0213*	short breath
White	1967	m	0213*	chest pain
White	1967	m	0213*	chest pain

Race E_0	ZIP Z_0
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

PT

Race E_1	ZIP Z_0
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

$GT_{[1,0]}$

Race E_1	ZIP Z_1
Person	0213*
Person	0213*
Person	0214*
Person	0214*
Person	0213*
Person	0213*
Person	0214*
Person	0214*

$GT_{[1,1]}$

Race E_0	ZIP Z_2
Black	021**
Black	021**
Black	021**
Black	021**
White	021**
White	021**
White	021**
White	021**

$GT_{[0,2]}$

Race E_0	ZIP Z_1
Black	0213*
Black	0213*
Black	0214*
Black	0214*
White	0213*
White	0213*
White	0214*
White	0214*

$GT_{[0,1]}$

Coming up

- Guest lecture by James Gardner
- Improved principles and anonymization algorithms