# Data Mining and Analysis

## Data Privacy

CSCE 676 :: Fall 2019
Texas A&M University
Department of Computer Science & Engineering
Prof. James Caverlee

# Resources

Li Xiong Data Privacy and Security course @Emory

http://www.mathcs.emory.edu/~lxiong/cs573_f16/schedule.html

Privacy-preserving data publishing: A survey of recent developments; k-anonymity (Sweeney)

A firm foundation for private data analysis, CACM, 2011; Designing statistical privacy for your data, CACM, 2015

Differential Privacy in the Wild. SIGMOD 2017 Tutorial. Machanavajjhala, Hay, and He.

# Data Anonymization

# Data Anonymization

Remove "personally identifying information" (PII)

Name, Social Security number, phone number, email, address... what else?

Problem: PII has no technical meaning

- Defined in disclosure notification laws
  - If certain information is lost, consumer must be notified
- Defined in sectoral laws such as HIPAA
  - 18 identifiers
- Any information can be personally identifying
  - Many de-anonymization examples: GIC, AOL dataset, Netflix Prize dataset

# Massachusetts GIC Incident

Massachusetts GIC released "anonymized" data on state employees' hospital visits

Then Governor Weld assured the public on privacy

GIC

| Name | SSN | Age | 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

| Age | Zip | Diagnosis |
|---|---|---|
| 44 | 48202 | AIDS |
| 44 | 48202 | AIDS |
| 44 | 48201 | Asthma |
| 55 | 48310 | Asthma |
| 55 | 48310 | Diabetes |

# Massachusetts GIC Incident

Then graduate student Sweeney linked the data with voter registration data in Cambridge and identified Governor Weld's record!

| Name | SSN | Age | 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 |

| Age | Zip | Diagnosis |
|---|---|---|
| 44 | 48202 | AIDS |
| 44 | 48202 | AIDS |
| 44 | 48201 | Asthma |
| 55 | 48310 | Asthma |
| 55 | 48310 | Diabetes |

Voter roll for Cambridge

| Name | Age | Zip |
|---|---|---|
| Alice | 44 | 48202 |
| Charley | 44 | 48201 |
| Dave | 55 | 48310 |

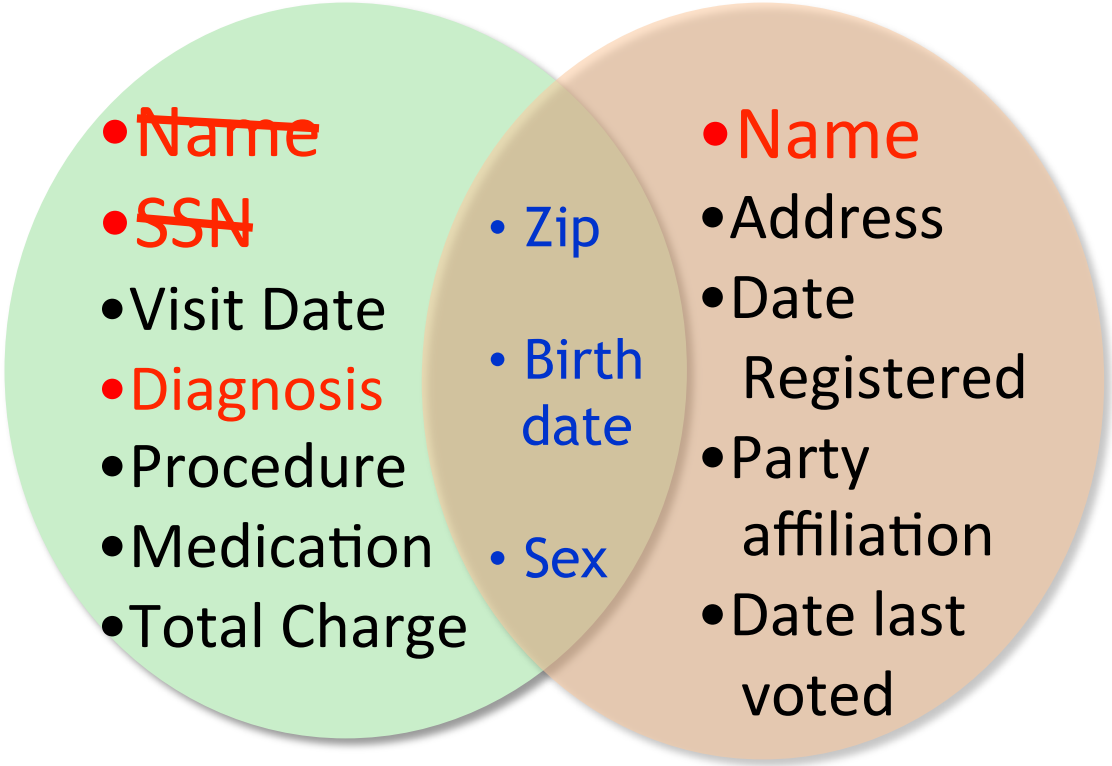
# The Massachusetts Governor Privacy Breach [Sweeney IJUFKS 2002]

- ~~Name~~
- ~~SSN~~
- Visit Date
- Diagnosis
- Procedure
- Medication
- Total Charge

- Zip
- Birth date
- Sex

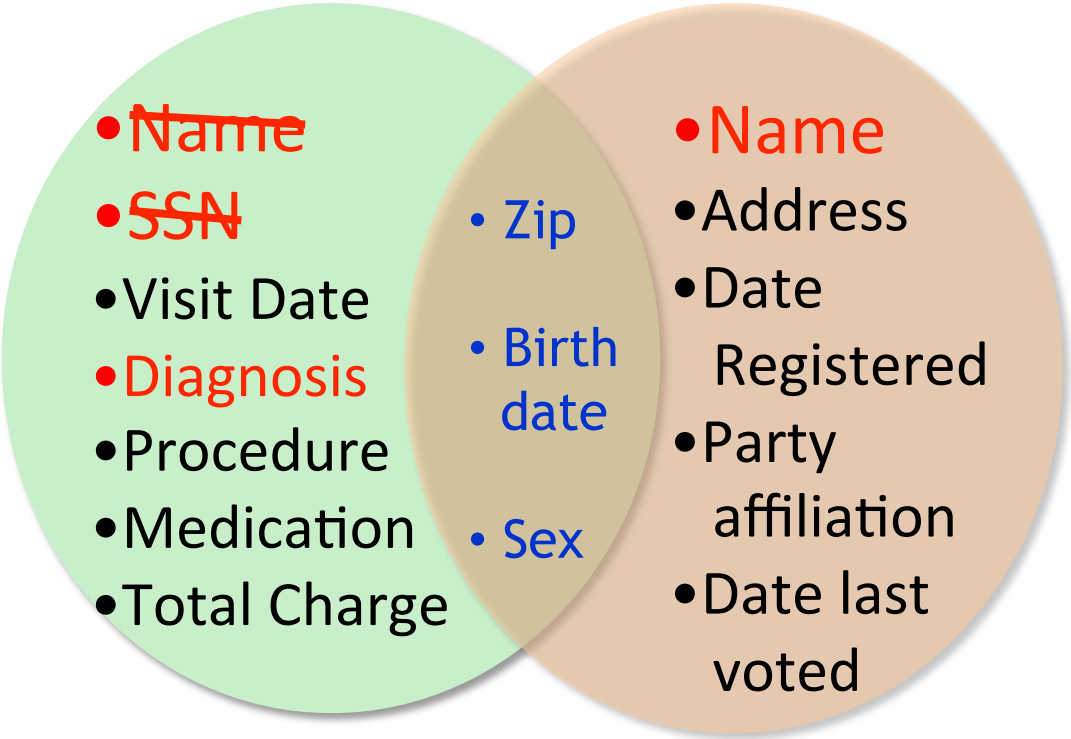
**Medical Data**

# The Massachusetts Governor Privacy Breach [Sweeney IJUFKS 2002]




**Medical Data** **Voter List**

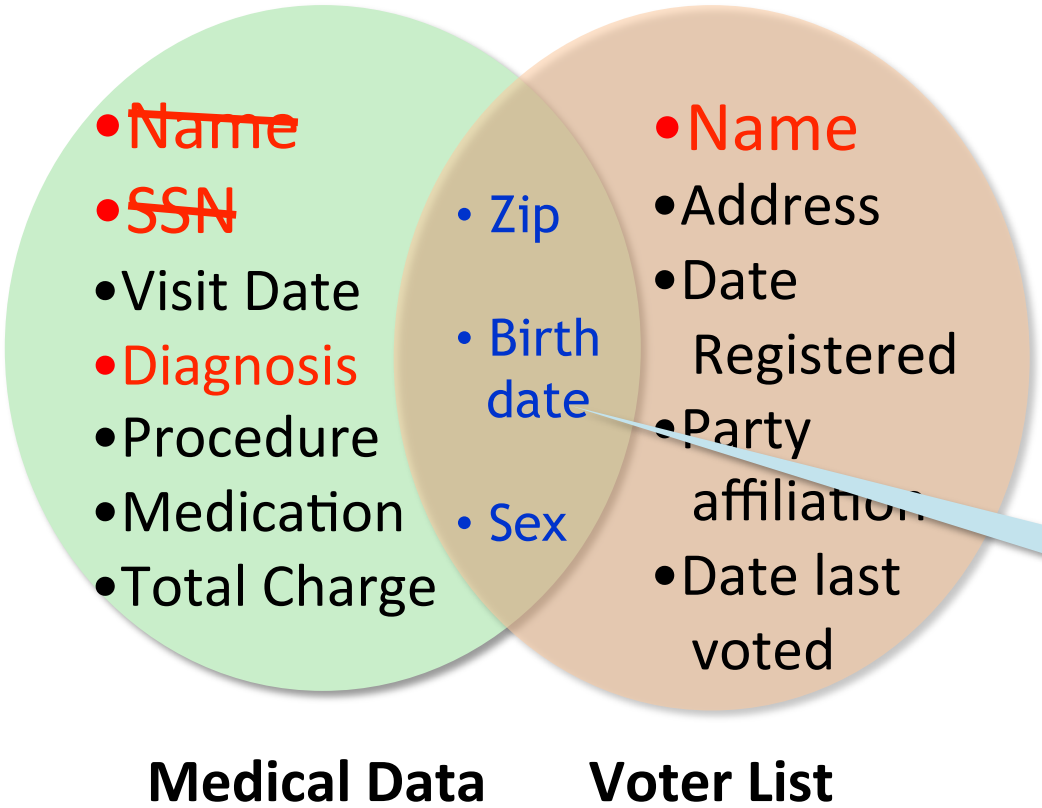
# The Massachusetts Governor Privacy Breach [Sweeney IJUFKS 2002]




- Governor of MA **uniquely identified** using ZipCode, Birth Date, and Sex.

**Name linked to Diagnosis**

# The Massachusetts Governor Privacy Breach [Sweeney IJUFKS 2002]




- 87 % of US population **uniquely identified** using ZipCode, Birth Date, and Sex.

**Quasi Identifier**

# AOL Query Log Release

In August 2006, AOL released a query log containing 21 million search queries for 650,000 users

No names or user identities included

| 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 |

# Query Log Applications

Implicit feedback for web search ranking

Query spelling correction

Query suggestion and refinement

Automatic monitoring and evaluation

Web search personalization

# AOL User 311045

how to change brake pads on scion xb
2005 us open cup florida state champions
how to get revenge on a ex
how to get revenge on a ex girlfriend
how to get revenge on a friend who f---ed you over
replacement bumper for scion xb
florida department of law enforcement
crime stoppers florida

# AOL User 4417749

numb fingers

60 single men

dog that urinates on everything

landscapers in Lilburn, Ga

homes sold in shadow lake subdivision gwinnett county georgia

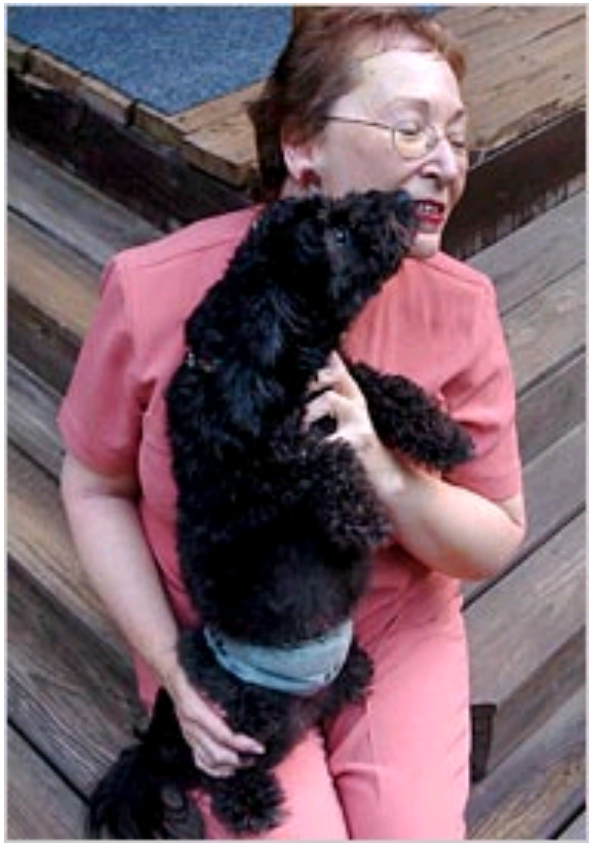
*Plus Several names with last name Arnold*

[NYTimes 2006]

# A Face Is Exposed for AOL Searcher No. 4417749

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

SIGN IN TO E-
THIS



— IN SOLIDARITY WITH THE MANY AOL USERS WHOSE OFTEN EMBARASSING WEB SEARCHES WERE RELEASED TO THE PUBLIC, I OFFER A SAMPLE OF MY OWN SEARCH HISTORY:

Google

**Web** Images Video New! News Maps **more »**

Advanced Search
Preferences
Language Tools

velociraptors
site:imdb.com "jurassic park"
raptors
dromaeosaurids
utahraptor
"home depot" deadbolts
security home improvement
surviving a raptor attack
robert bakker paleontologist
robert bakker "possible raptor sympathizer"
site:en.wikipedia.org surviving a raptor attack
learning from mistakes in jurassic park
big-game rifles
tire irons
treating raptor wounds
do raptors fear fire
how to make a molotov cocktail
do raptors fear death
can raptors pick locks
how to tell if my neighbors are raptors

# The Genome Hacker

nature International weekly journal of science

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

Archive > Volume 497 > Issue 7448 > News Feature > Article

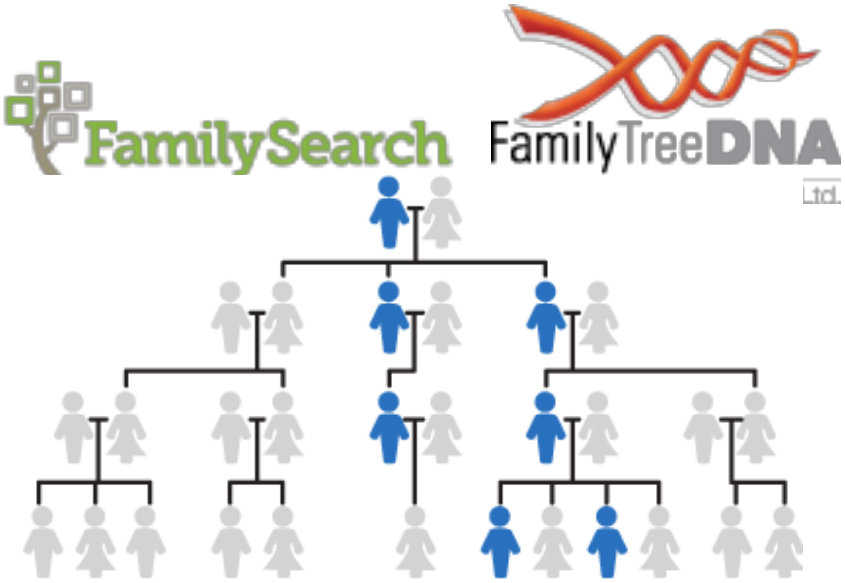
*NATURE* | NEWS FEATURE

Privacy protections: The genome hacker

**Yaniv Erlich shows how research participants can be identified from 'anonymous' DNA.**

**Erika Check Hayden**

08 May 2013



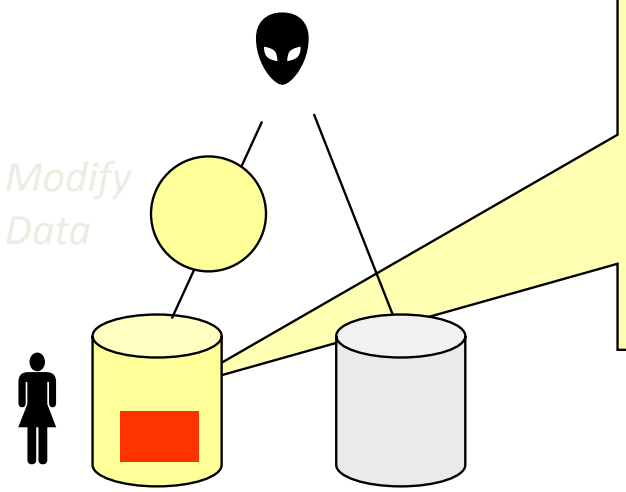
# K-anonymity

Introduced by Samarati and Sweeney in 1998

Hundreds of papers on the topic

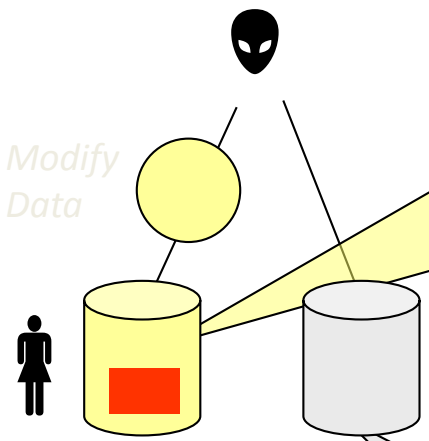
- Theoretical results
- Algorithms achieving k-anonymity
- Improvements

# 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



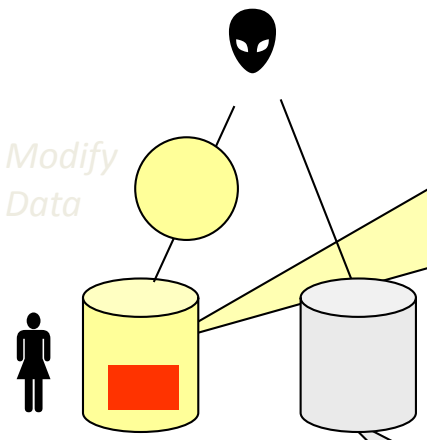

Published Data: Alice publishes data without the Name

| | *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




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 identifiers

There are some fields that may uniquely identify some individual

| | *Non-Sensitive Data* | | | *Sensitive Data* |
|---|---|---|---|---|
| ***#*** | ***Zip*** | ***Age*** | ***Nationality*** | ***Condition*** |
| ... | ... | ... | ... | ... |

Quasi Identifier

The attacker can use these fields to join with other sources to identify the individuals

# Attribute Types

Identifier, Key (Quasi-Identifier), 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 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

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

| 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 |

# Achieving K-Anonymity

**Generalization**

- Replace specific quasi-identifiers with more general values until get k identical values
  - Example: area code instead of phone number
- 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

$Z_3$={*****}

↑

$Z_2$={021**}

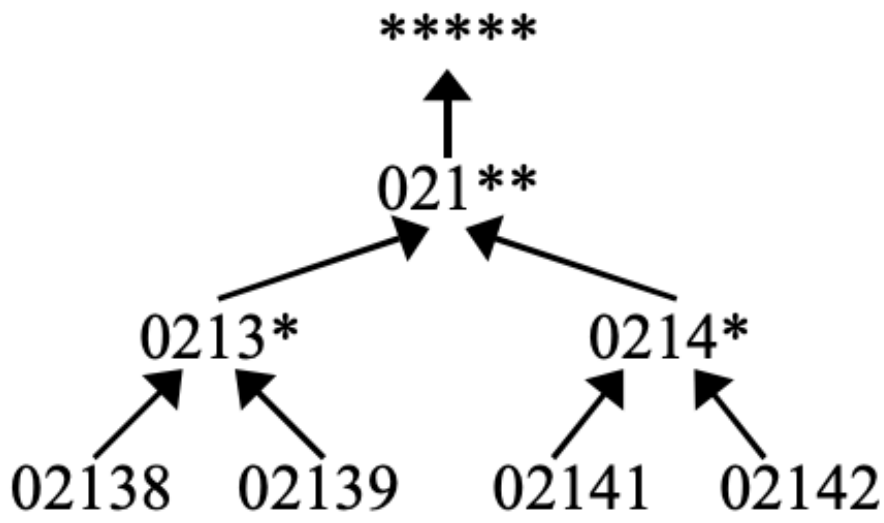
↑

$Z_1$={0213*,0214*}

↑

$Z_0$={02138, 02139, 02141, 02142}

$DGH_{Z0}$




$VGH_{Z0}$

$Z_2$={******}

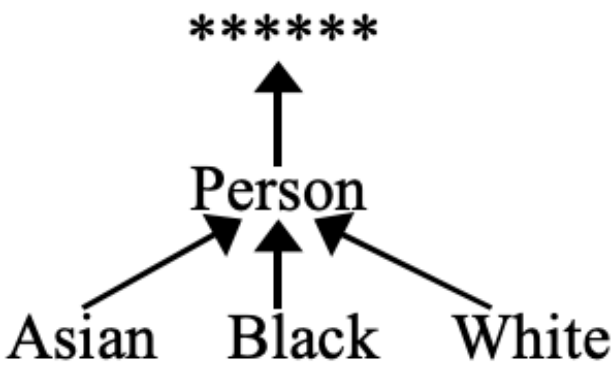
↑

$Z_1$={Person}

↑

$Z_0$={Asian,Black,White}

$DGH_{E0}$




$VGH_{E0}$

| 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* |

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

**Figure 4 Examples of generalized tables for PT**

# Disclosure Control

**Membership disclosure:** Attacker cannot tell that a given person is 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 corresponds to a given person

# Issues: Membership Disclosure

With large probability, quasi-identifier is unique in the population

But generalizing/suppressing quasi-identifiers in the dataset does not affect their distribution in the population (obviously)!

> Suppose anonymized dataset contains 10 records with a certain quasi-identifier …
>
> … and there are 10 people in the population who match this quasi-identifier

k-anonymity may not hide whether a given person is in the dataset

# #Hospital discharges in NJ of ovarian cancer patients, 2009

Counts less than k are suppressed achieving k-anonymity

| Age | #discharges | White | Black | Hispanic | Asian/ Pcf Hlnder | Native American | Other | Missing |
|---|---|---|---|---|---|---|---|---|
| #discharges | 735 | 535 | 82 | 58 | 18 | * | 19 | 22 |
| 1-17 | * | * | * | * | * | * | * | * |
| 18-44 | 70 | 40 | 13 | * | * | * | * | * |
| 45-64 | 330 | 236 | 31 | 32 | * | * | 11 | * |
| 65-84 | 298 | 229 | 35 | 13 | * | * | * | * |
| 85+ | 34 | 29 | * | * | * | * | * | * |

# #Hospital discharges in NJ of ovarian cancer patients, 2009

| Age | #discharges | White | Black | Hispanic | Asian/ Pcf Hlnder | Native American | Other | Missing |
|---|---|---|---|---|---|---|---|---|
| #discharges | 735 | 535 | 82 | 58 | 18 | **1** | 19 | 22 |
| 1-17 | **3** | **1** | * | * | * | * | * | * |
| 18-44 | 70 | 40 | 13 | * | | | | * |
| 45-64 | 330 | 236 | 31 | 32 | | | 1 | * |
| 65-84 | 298 | 229 | 35 | 13 | * | * | * | * |
| 85+ | 34 | 29 | * | * | * | * | * | * |

= 535 – (40+236+229+29)

# #Hospital discharges in NJ of ovarian cancer patients, 2009

| Age | #discharges | White | Black | Hispanic | Asian/ Pcf Hlnder | Native American | Other | Missing |
|---|---|---|---|---|---|---|---|---|
| #discharges | 735 | 535 | 82 | 58 | 18 | **1** | 19 | 22 |
| 1-17 | **3** | **1** | [0-2] | [0-2] | [0-2] | [0-2] | [0-2] | [0-2] |
| 18-44 | 70 | 40 | 13 | * | * | * | * | * |
| 45-64 | 330 | 236 | 31 | 32 | * | * | 11 | * |
| 65-84 | 298 | 229 | 35 | 13 | * | * | * | * |
| 85+ | 34 | 29 | * | * | * | * | * | * |

# #Hospital discharges in NJ of ovarian cancer patients, 2009

| Age | #discharges | White | Black | Hispanic | Asian/ Pcf Hlnder | Native American | Other | Missing |
|---|---|---|---|---|---|---|---|---|
| #discharges | 735 | 535 | 82 | 58 | 18 | **1** | 19 | 22 |
| 1-17 | **3** | **1** | [0-2] | [0-2] | [0-2] | [0-2] | [0-2] | [0-2] |
| 18-44 | 70 | 40 | 13 | * | * | * | * | * |
| 45-64 | 330 | 236 | 31 | 32 | * | * | 11 | * |
| 65-84 | 298 | 229 | 35 | 13 | * | * | * | * |
| 85+ | 34 | 29 | [1-3] | * | * | * | * | * |

# Can reconstruct tight bounds on rest of data

In fact, when linked with queries giving other statistics, we can figure out that exactly 1 Native American woman diagnosed with ovarian cancer went to a privately owned, not for profit, teaching hospital in new Jersey with more than 435 beds in 2009. Furthermore, the woman did not pay by private insurance, had a routine discharge, with a stay in the hospital of 33.5 days, with her home residence being in a county with 1 million plus residents (large fringe metro, suburbs), and her age was exactly 75 years.

# 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

# Issues: Sensitive Attribute Disclosure

Intuitive reasoning:

> k-anonymity prevents attacker from telling which record corresponds to which person
>
> Therefore, attacker cannot tell that a certain person has a particular value of a sensitive attribute

This reasoning is fallacious!

# 3-Anonymization

| | | |
|---|---|---|
| Caucas | 78712 | Flu |
| Asian | 78705 | Shingles |
| Caucas | 78754 | Flu |
| Asian | 78705 | Acne |
| AfrAm | 78705 | Acne |
| Caucas | 78705 | Flu |

| | | |
|---|---|---|
| Caucas | 787XX | Flu |
| Asian/AfrAm | 78705 | Shingles |
| Caucas | 787XX | Flu |
| Asian/AfrAm | 78705 | Acne |
| Asian/AfrAm | 78705 | Acne |
| Caucas | 787XX | Flu |

This is 3-anonymous, right?

# Joining with External Data

| | | |
|---|---|---|
| … | … | … |
| Rusty Shackleford | Caucas | 78705 |
| … | … | … |

+

| | | |
|---|---|---|
| Caucas | 787XX | Flu |
| Asian/AfrAm | 78705 | Shingles |
| Caucas | 787XX | Flu |
| Asian/AfrAm | 78705 | Acne |
| Asian/AfrAm | 78705 | Acne |
| Caucas | 787XX | Flu |

Problem: sensitive attributes are not "diverse" within each quasi-identifier group

# Another Attempt: l-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 |

Entropy of sensitive attributes within each quasi-identifier group must be at least L

# However …

Original dataset

| | |
|---|---|
| … | HIV- |
| … | HIV- |
| … | HIV- |
| … | HIV- |
| … | HIV- |
| … | HIV+ |
| … | HIV- |
| … | HIV- |
| … | HIV- |
| … | HIV- |
| … | HIV- |
| … | HIV- |

99% have HIV-

Anonymization A

| | |
|---|---|
| Q1 | HIV+ |
| Q1 | HIV- |
| Q1 | HIV+ |
| Q1 | HIV- |
| Q1 | HIV+ |
| Q1 | HIV- |
| Q2 | HIV- |
| Q2 | HIV- |

Anonymization B

| | |
|---|---|
| Q1 | HIV- |
| Q1 | HIV- |
| Q1 | HIV- |
| Q1 | HIV+ |
| Q1 | HIV- |
| Q1 | HIV- |
| Q2 | HIV- |
| Q2 | HIV- |

99% HIV- ⇒ quasi-identifier group is <u>not</u> “diverse”
…yet anonymized database does not leak anything

50% HIV- ⇒ quasi-identifier group is “diverse”
**This leaks a ton of information**

# 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 |

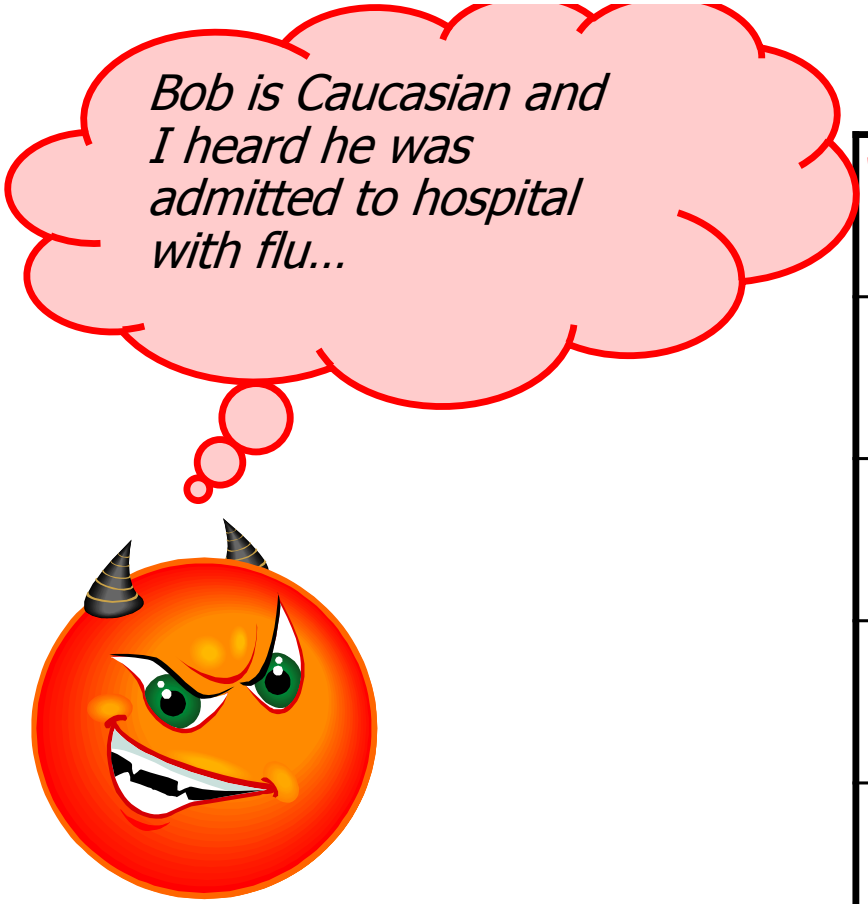
Distribution of sensitive attributes within each quasi-identifier group should be “close” to their distribution in the entire original database

# k-anonymous, t-close dataset

This is k-anonymous, l-diverse, and t-close …

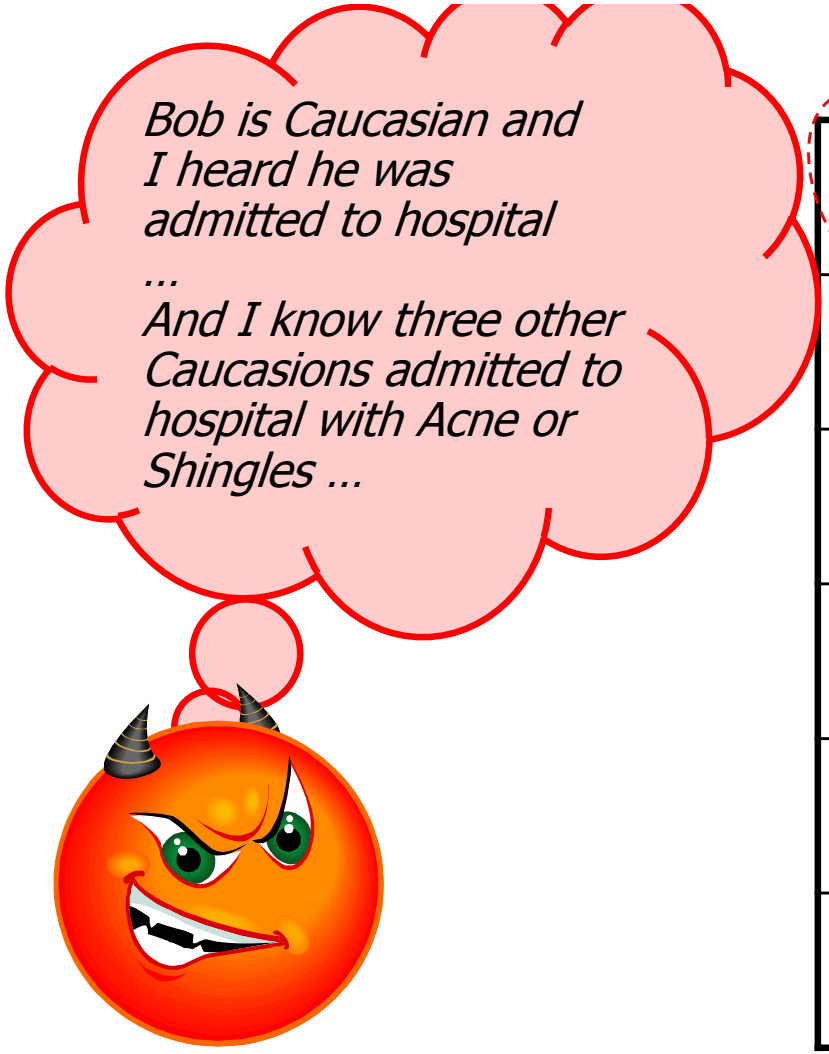
| | | | |
|---|---|---|---|
| Caucas | 787XX | HIV+ | Flu |
| Asian/AfrAm | 787XX | HIV- | Flu |
| Asian/AfrAm | 787XX | HIV+ | Shingles |
| Caucas | 787XX | HIV- | Acne |
| Caucas | 787XX | HIV- | Shingles |
| Caucas | 787XX | HIV- | Acne |

# What does attacker know?




| | | | |
|---|---|---|---|
| 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 |

# What does attacker know?




| | | | |
|---|---|---|---|
| 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 |

# Takeaways

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
- Any attribute can be a potential quasi-identifier (AOL example)

Relies on locality

- Destroys utility of many real-world datasets