Medical Data Privacy and Ethics in the Age of Artificial Intelligence

Lecture 10: Re-identification Risk of Health Records

Zhiyu Wan, PhD (wanzhy@shanghaitech.edu.cn)
Assistant Professor of Biomedical Engineering
ShanghaiTech University

November 4, 2025

Learning Objectives of This Lecture

- Know the difference between de-identification and anonymity
  - AOL case
  - Hospital discharge record case
- Know how to estimate uniqueness bounds
  - Threshold approach
  - Probabilistic approach

Terminology

- Explicit identifier: features that permit a direct communication with an individual / entity
- Quasi-identifier: features that, in combination, permit the indirect recognition of an individual / entity

Adapted from Dr. Malin's slides.

The AOL Search Log Case of 2006

Goal: Support web information retrieval research

- 650k customers, 20 mil. queries, 3 mo. period
- Names replaced with persistent pseudonyms

Pseudonym	Name	Query	Date	Time
1	[redacted]	Books	1/2/05	16:52
2	[redacted]	Payscale	1/4/05	23:41
1	[redacted]	Porn	1/8/05	03:15

Queries

User 2178
foods to avoid when breast feeding

User 3505202
depression and medical leave

User 47122
Child porno

User 31350
How to kill oneself with gas

3482401
calorie counting

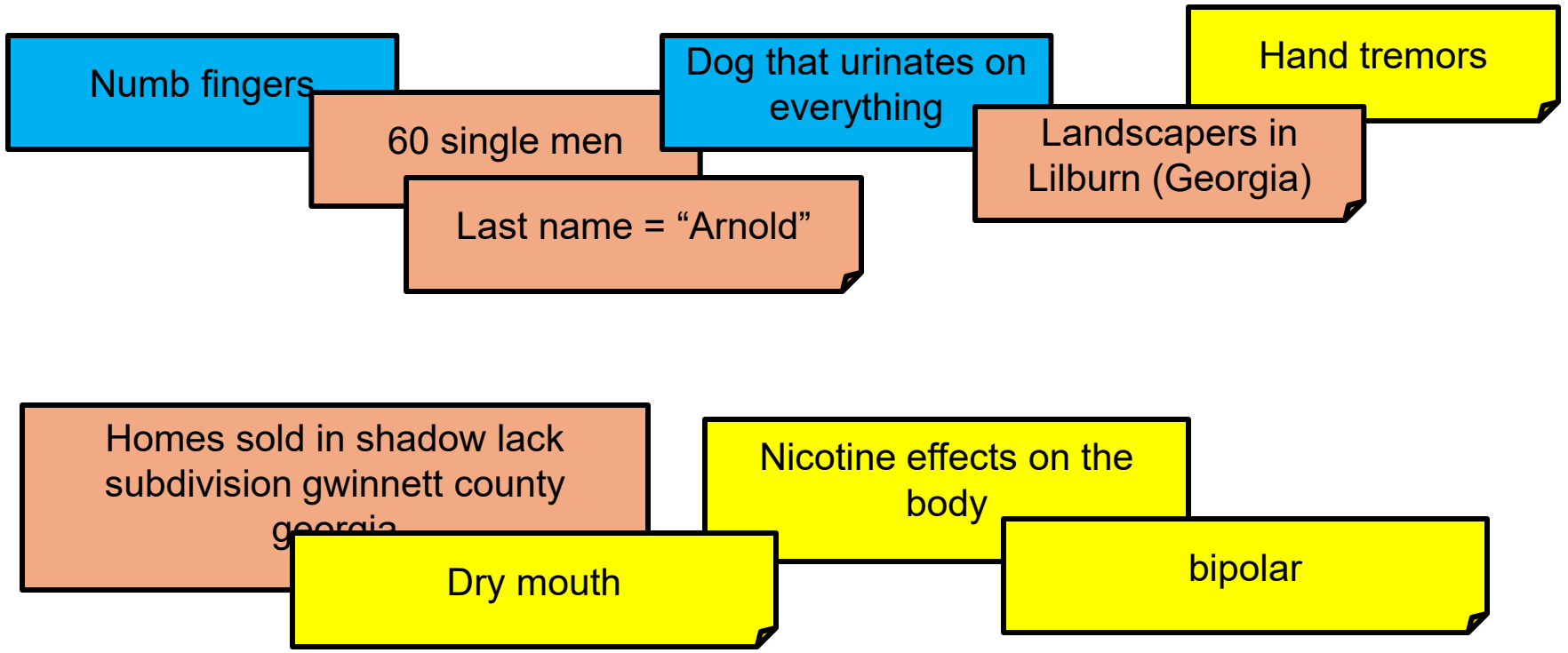
7268042
fear that spouse contemplating cheating

User 3483689
Time after time

User 3483689
Wind beneath my wings

Barbaro & Zeller. A face exposed for AOL searcher no. 4417749.
New York Times. Aug 9, 2006.

User 4417749 issued hundreds of searches



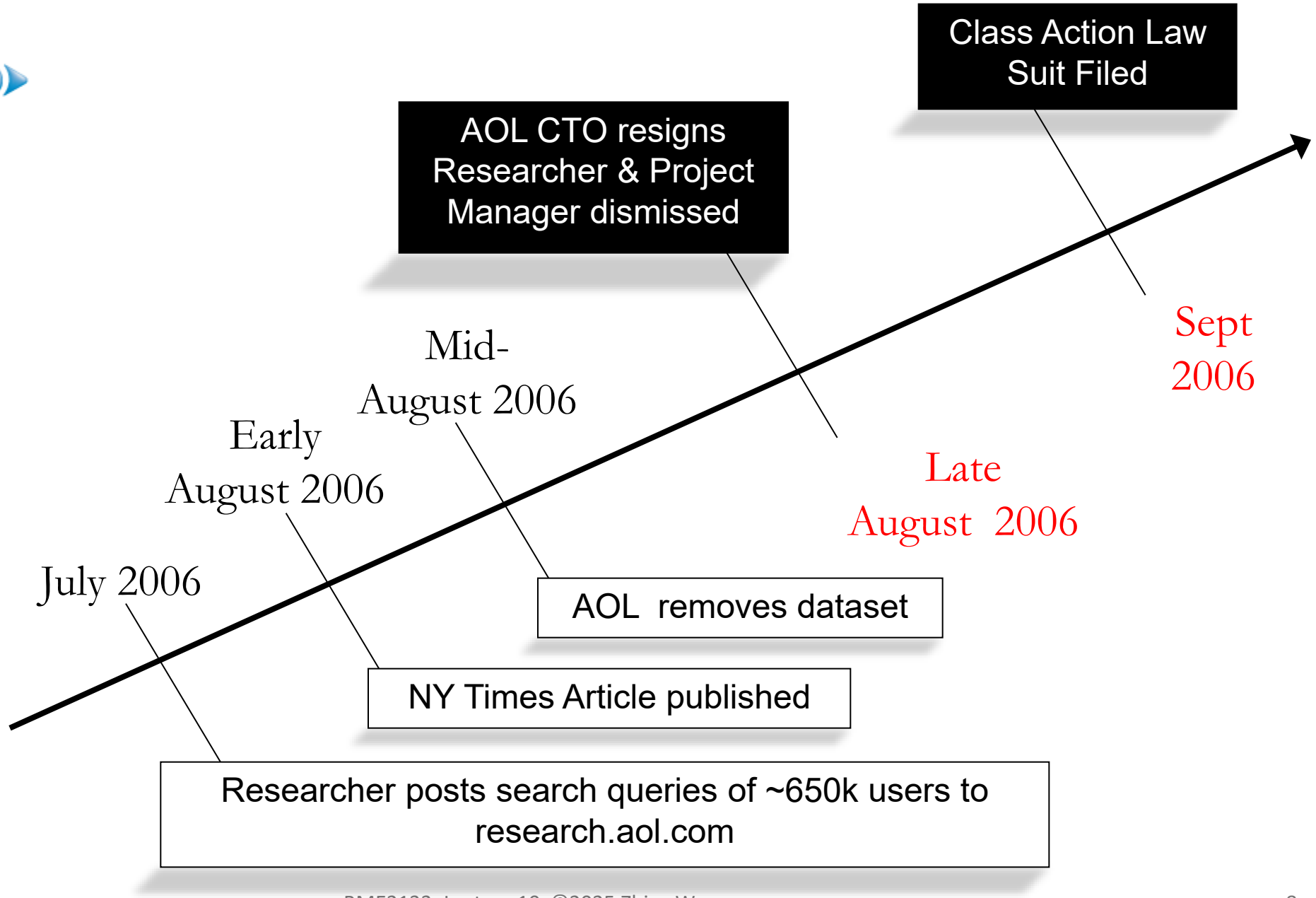
Barbaro & Zeller. A face exposed for AOL searcher no. 4417749.
New York Times. Aug 9, 2006.

User 4417749 issued hundreds of searches

Numb fingers

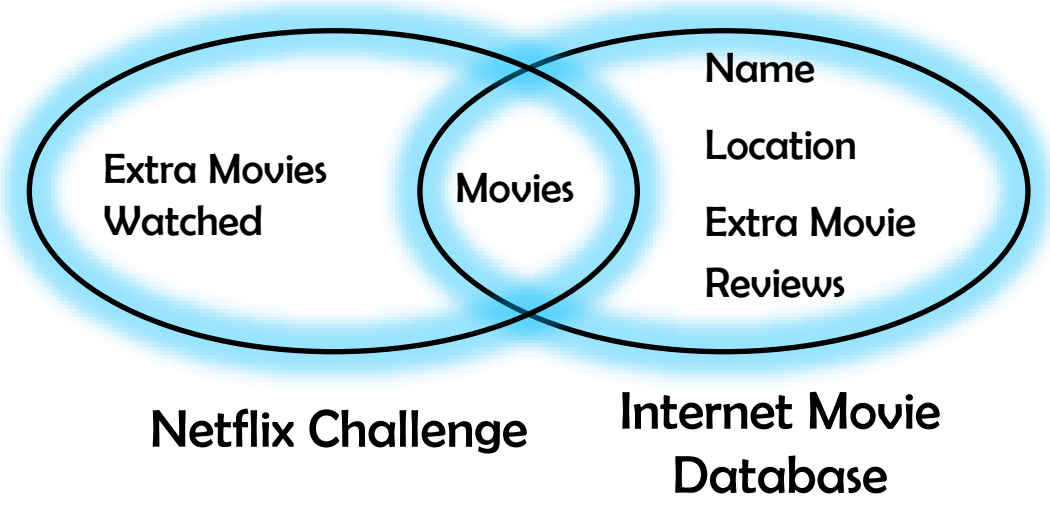
Homes sold
subdivision
ge

Thelma Arnold
& Dudley


Class Action Law Suit Filed
AOL CTO resigns Researcher & Project Manager dismissed
Sept 2006
Mid-August 2006
Early August 2006
Late August 2006
July 2006
AOL removes dataset
NY Times Article published
Researcher posts search queries of ~650k users to research.aol.com

The Netflix Challenge (2008-2009)

- Netflix published movie selections of ~450,000 pseudonymized subscribers
- Re-identification via uniqueness of movie combinations
- Class action filed December 2009



A. Narayanan & V. Shmatikov. IEEE Security and Privacy Conference. 2008.

- How would you measure identifiability of the AOL dataset?
- How would you protect the identities of individuals in the AOL dataset?

Forbes.com
U.S. EUROPE ASIA
Home Lists Business Tech
Breakth

Welcome Google User
Here are more stories related to your searc
• Netflix Settles Privacy Lawsuit, Canc
See all related stories >

The Firewall
Filtering ideas in the world of security.

Netflix Settles Privacy Lawsuit, Cancels Prize Sequel
March 12, 2010 - 12:35 pm

Taylor Buley Bio | Email
Taylor Buley is a staff writer and editorial developer for Forbes

Share 8
67 retweet

On Friday, Netflix announced on its corporate blog that it has settled a lawsuit related to its Netflix Prize, a $1 million contest that challenged machine learning experts to use Netflix's data to produce better recommendations than the movie giant could serve up themselves.

THE SIMPLE PROCESS OF RE-IDENTIFYING PATIENTS IN PUBLIC HEALTH RECORDS

In late 2016, doctors' identities were decrypted in an open dataset of Australian medical billing records. Now patients' records have also been re-identified - and we should be talking about it

By Dr Vanessa Teague, Dr Chris Culnane an

ENGINEERING & TECHNOLOGY

Featured

Dr Vanessa Teague
School of Computing and Information Systems, Melbourne School of Engineering, University of Melbourne

Dr Chris Culnane
School of Computing and Information Systems, Melbourne School of Engineering, University of Melbourne

Dr Benjamin Rubinstein
Senior Lecturer, School of Computing and Information Systems, Melbourne School of Engineering, University of Melbourne

In August 2016, Australia's federal Depart
records of about 2.9 million Australians c
Medicare Benefits Scheme (MBS) and the
containing 1 billion lines of historical hea
cent of the population.

These longitudinal records were de-iden
person's identity from being connected w
the government's open data website as pa

RE-IDENTIFYING PATIENTS

We found that patients can be re-identified, without decryption, through a process of linking the unencrypted parts of the record with known information about the individual.

Our findings replicate those of similar studies of other de-identified datasets:

- A few mundane facts taken together often suffice to isolate an individual.
- Some patients can be identified by name from publicly available information.
- Decreasing the precision of the data, or perturbing it statistically, makes re-identification gradually harder at a substantial cost to utility.

A Data Detective Method: Direct Linkage

1. Uses the combination of attributes to determine the uniqueness of an entity in a dataset

2. Second dataset with identified subjects is used to make the re-identification by drawing inferences between the two datasets on the related attributes

3. The attributes do not have to be equal, but there must exist some ability for inference of between attributes.

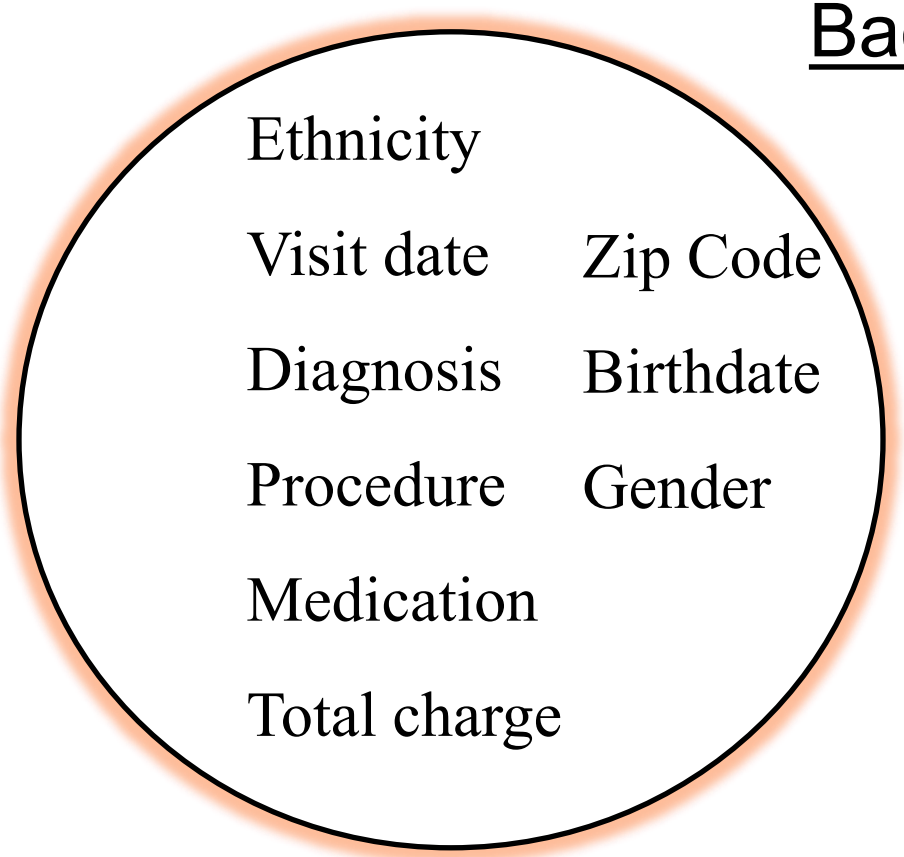
Healthcare Reform At Work

- In 1997, 44 of 50 states collected and disseminated hospital discharge data
- In 2019, its 49 of 50 states “ “
- Attributes recommended by *National Association of Health Data Organizations* for disclosure
  - ❑Patient Zip Code
  - ❑Patient Birth Date
  - ❑Patient Gender
  - ❑Patient Racial Background
  - ❑Patient Number
  - ❑Visit Date
  - ❑ Principle Diagnosis Codes (ICD-9)
  - ❑ Procedure Codes
  - ❑ Physician ID Number
  - ❑ Physician Zip Code
  - ❑ Total Charges

J. Schoenman et al. The value of hospital discharge databases. NORC & NAHDO. 2005.http://www.hcup-us.ahrq.gov/reports/final_report.pdf

Case Study – “Quasi-identifier”

Back in the ‘90s

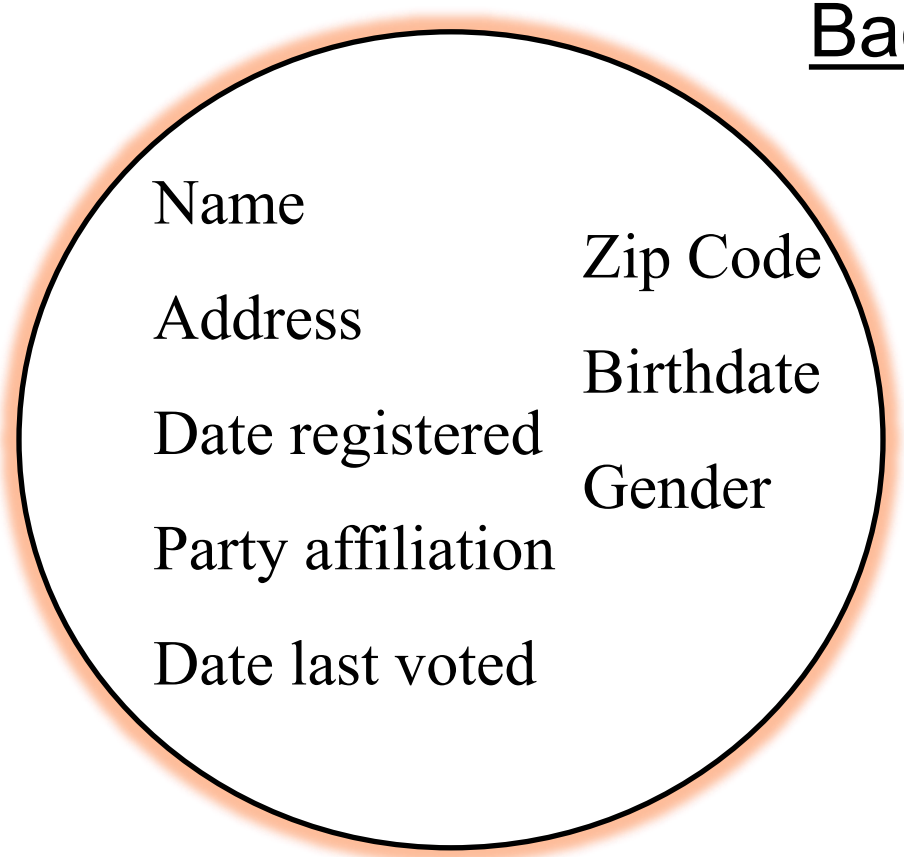


Hospital Discharge Data

L. Sweeney. Journal of Law, Medicine, and Ethics. 1997.

Case Study – “Quasi-identifier”

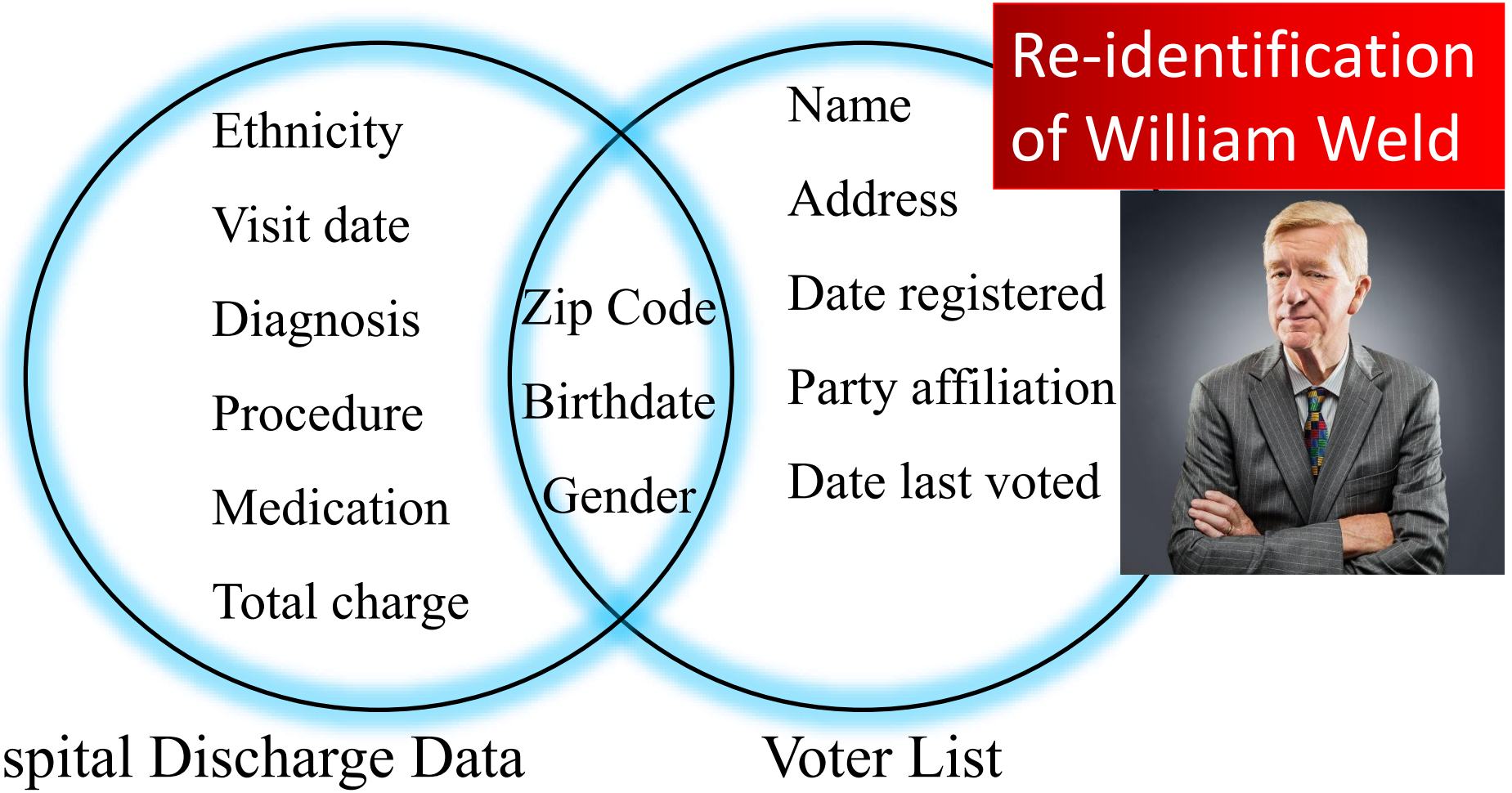
Back in the ‘90s



City of Cambridge, MA Voter Registration Records

L. Sweeney. Journal of Law, Medicine, and Ethics. 1997.

Case Study – “Quasi-identifier”



Re-identification of William Weld

L. Sweeney. Journal of Law, Medicine, and Ethics. 1997.

Discharge Database Users

- Users are diverse
  - Government agencies
  - Provider associations and individual health care providers
  - Health care insurers and large health care purchasers (e.g., self-insuring companies)
  - Policymakers
  - *Researchers*
  - Private-sector (e.g., data consolidators / brokers)
  - Policy deliberations

J. Schoenman et al. The value of hospital discharge databases. NORC & NAHDO. 2005.http://www.hcup-us.ahrq.gov/reports/final_report.pdf

Discharge Database Findings

- Public safety and injury surveillance / prevention
- Public health, disease surveillance, and disease registries
- Public health planning & community assessments
- Public reporting for informed purchasing & comparative reports
- Quality assessment & performance improvement
- Health services & health policy research
- Private sector analysis

J. Schoenman et al. The value of hospital discharge databases. NORC & NAHDO. 2005.http://www.hcup-us.ahrq.gov/reports/final_report.pdf

HCUP

- Agency for Healthcare Research & Quality (AHRQ)
- Sponsors the Healthcare Cost and Utilization Project (HCUP)
  - Integrates state-level data collections
  - creates a uniformly formatted national information resource of discharge-level health care data
- HCUP Training
  - https://www.hcup-us.ahrq.gov/tech_assist/dua.jsp

- Modeling the identifiability of Nationwide Inpatient Sample (NIS)
  - Do not re-identify people.
  - Calculate the re-identification risk given population estimates

How much Identification is There?

- Linkage model provides a route to re-identify people
- It does not indicate the number of people that are at risk for re-identification
- This requires quantifiable methods

Counting

- Quasi-identifier $Q = \{q_1, \ldots, q_n\}$
  - Ex: {Date of Birth, Gender, 5-Digit Zip Code}

- Each attribute q_i has a set of associated values
  - Cardinality to represent set size: $|q_i|$
  - Ex: 5-digit Zip Code has maximum-sized value set {00000, …, 99999} and thus |5-digit Zip| = 100000

- Maximum number of quasi-identifying values in a population is:

$$\prod_{i=1}^{n} |q_i|$$

Naïve Risk Analysis

- Dirichlet drawer principle (a.k.a.“Pigeon-hole principle”)
- Population has size n
- Number of quasi-ID values is m
- Principle: There is at least one quasi-ID value with $\lceil n/m \rceil$ individuals from the population
- Proof by contradiction

Application to Privacy

- Imagine you have a population of 100,001 individuals
- The age range is [0,99]
- There are 500 zip codes in the area
- There are two genders {Male, Female}
- Is there a portion of the population that can NOT be unique?
- Quasi-ID size: 100*500*2 = 100,000
- Dirichlet → $\lceil 100{,}001 / 100{,}000 \rceil = 2$
  - At least one quasi-id value with 2 assigned person

Application to Privacy

- Imagine you have a population of 100,001 individuals
- The age range is [0,50]
- There are 50 zip codes in the area
- There are two genders {Male, Female}
- Is there a portion of the population that can NOT be unique?
- Quasi-ID size: 50*50*2 = 5,000
- Dirichlet → $\lceil 100{,}001 / 5{,}000 \rceil = 21$
  - At least one quasi-id value with 21 assigned persons

Bounds

- *n* people
- *m* quasi-id values
- What is the **maximum** number of people that can be uniquely identified?

Threshold Approach (Sweeney 2000)

- Subdivide a population by its quasi-identifier
- Calculate the number of “uniques” in the population
- Given *n* attributes in the QID, uniques is # of QID values with totals equal to 1
- Quasi-identifier = {DOB, Zip, Gender}
- 3-dimensional contingency table (compressed view)

		DOB / Gender							
		Male				Female			
		dob 1	dob 2	…	dob n	dob 1	dob 2	…	dob n
ZIP	zip1	Cell = 1?							
	zip2								
	…								
	zip m								

Estimation

- Sometimes, you don’t have the exact details of what adversary has access to
  - e.g., you disclose a sample with {dob, gender, zip}, but don’t know the full population values
- But you may have access to aggregates
  - E.g., Census counts for {year of birth, gender, county}…
- One way to measure the identifiability is to estimate detailed data from the aggregates
- Year of birth → Date of Birth
  - Uniform distribution of dates
  - Equal likelihood of 365 days

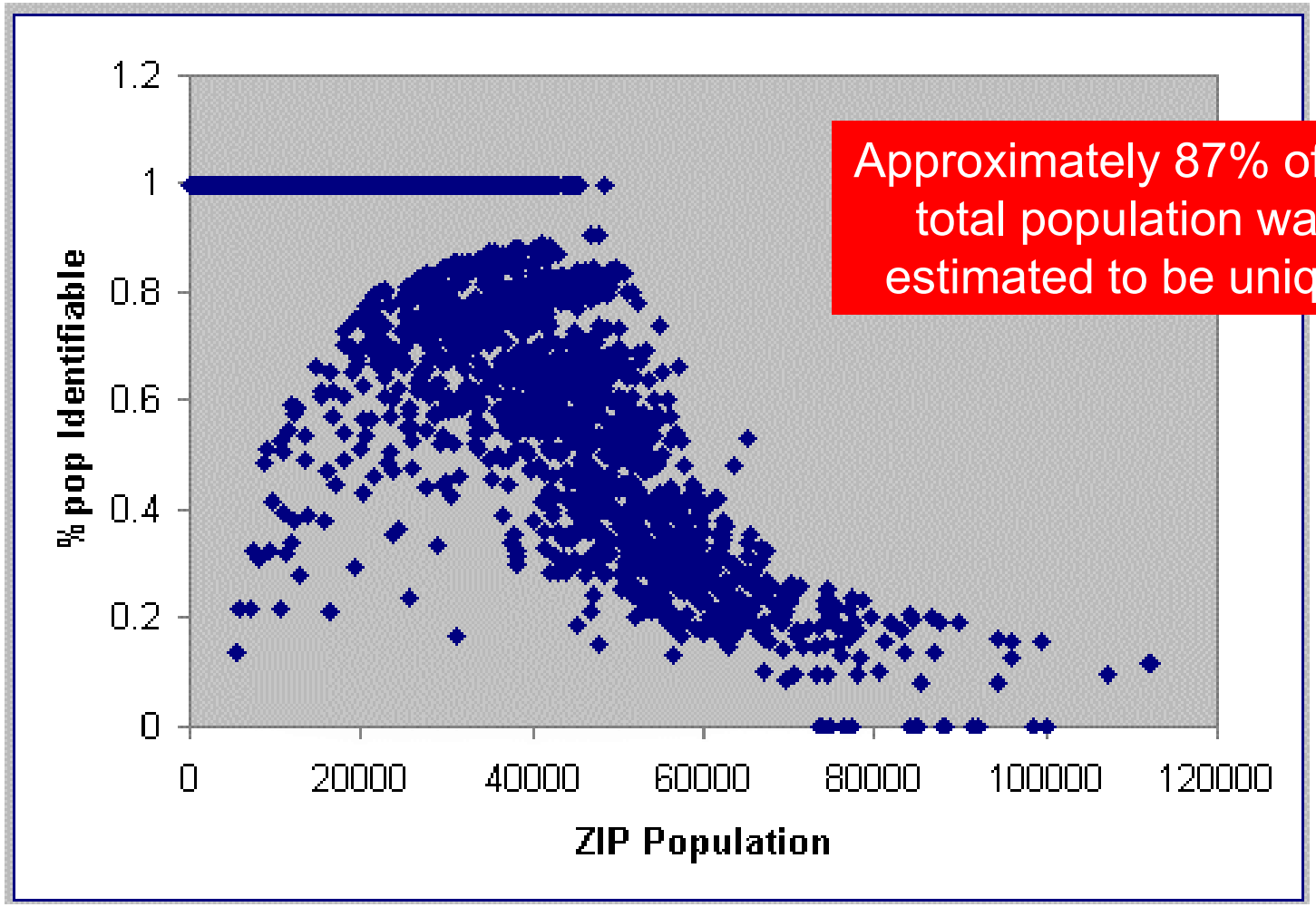
Conversion

- $\{Year\ of\ Birth,\ ZIP\} \rightarrow \{Date\ of\ Birth,\ ZIP\}$
- One option: Equal distribution of values across the cells

		Birth Year
		1980
ZIP	zip1	12000
	zip2	50000
	…	
	zip m	10000

		Birthdate			SUM
		1/1/80	…	12/31/80	
ZIP	zip1	12000 / 365		12000/365	12000
	zip2	50000 / 365		50000/365	50000
	…				
	zip m	10000 / 365		10000 /365	10000

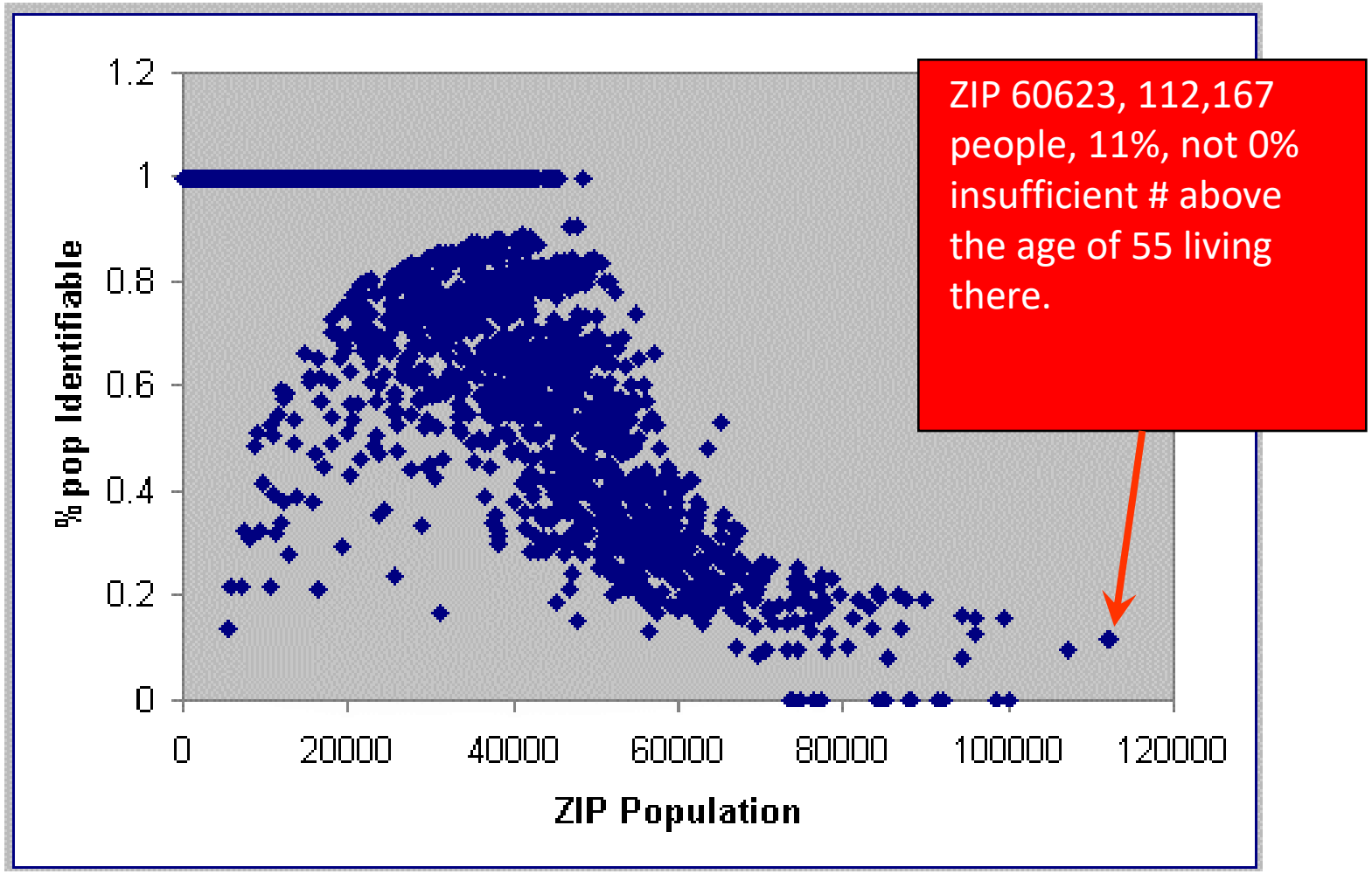
{date of birth, gender, 5-digit ZIP} (1990 US Census)



Approximately 87% of the total population was estimated to be unique

L. Sweeney. 2000.

{date of birth, gender, 5-digit ZIP} (1990 US Census)



L. Sweeney. 2000.

60623 Zip Code in Chica

www.neighborhoodlink.com/zip/60623

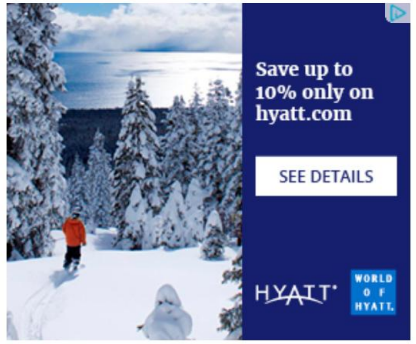
Welcome, visitor! | Sign In

USA | Illinois | Chicago | 60623 | Neighborhoods

60623 Zip Code Profile

Find Neighborhoodss, Home Values, Schools, Demographics, Local Discussions, Maps, & much more.

Welcome to CHICAGO, IL 60623



60623 is a densely populated, urban zip code in Chicago, Illinois. Median household income here ($28,203) is significantly lower than US average ($56,604). The population is racially diverse, younger, and about evenly divided between singles and married couples. Housing prices here (average $203,300) are fairly typical for the Chicago-Naperville-Joliet metro area.

The average family income here is $36,547/year. This is less than the national average of $70,000/year. However, the national average number is a little misleading because there are some very wealthy Americans earning $50 million a year or more. (Wouldn't that be nice!) This pushes the national average higher than in would be without those edge cases. The median household income in the US (median, meaning half of US households make more and half make less) is at about $50,000/year. Median family income in 60623 is $29,137.

An intersting fact about income: Men in 60623 earn an average of $22,253/year. Women earn only $18,178/year.

32,172 people in 60623 have jobs. This statistic includes anyone over the age of 16.

Are you thinking about moving to a neighborhood in 60623? You might be interested to know that the average commute time to work for people living in 60623 is 36.3 minutes!

The median age here is 28. There are 46,188 men and 45,920 women. The median age for men is 27 while for women the median age is 28.

To give you a sense of the community, 58,248 people (out of the 92,108 people live here) have lived in their home at least 5 years. The Post Office delivers mail to 17,511 homes, and 1,534 businesses every day. 2,408 people ride bikes or walk to work on a fairly regular basis.

60623 Zip code is located in the Central time zone at 42 degrees latitude (**Fun Fact**: this is the same latitude as Andorra la Vella, Andorra!) and -88 degrees longitude. It has an average elevation of 599 feet above sea levell.

SHARE

60623 Guide

- Overview
- Neighborhoods
- Real Estate & Home Values
- Demographics
- Geography
- Economics
- Schools
- Local Pictures
- Sex Offender Info

Discussion Topics

- Chicago
- Politics
- Crime
- Children
- Lost Pets
- Landscaping

Nearby Zip Codes

- 60624
- 60804
- 60608
- 60612
- 60632
- 60644

Safety

Alertag

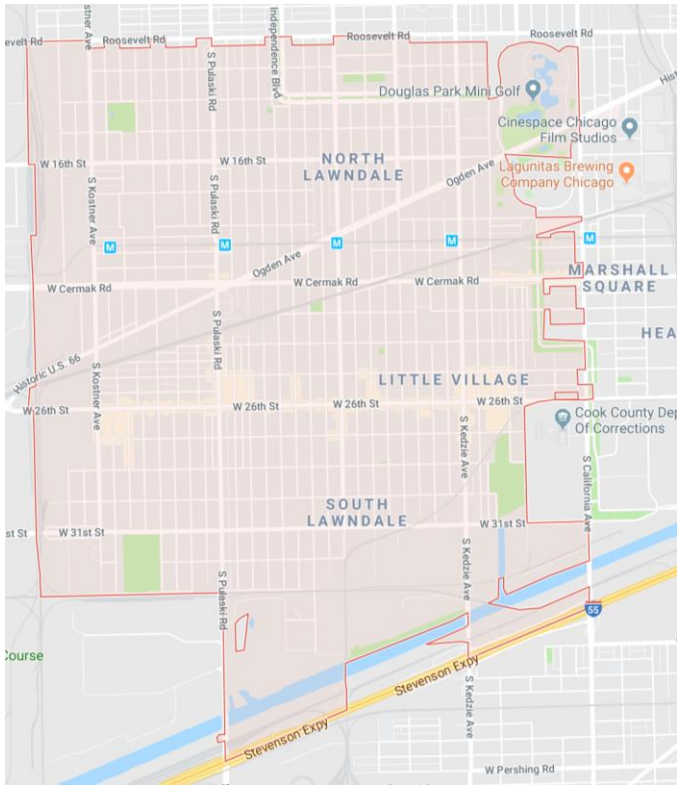
Take the Alertag quiz and find out if an Alertag is right for you

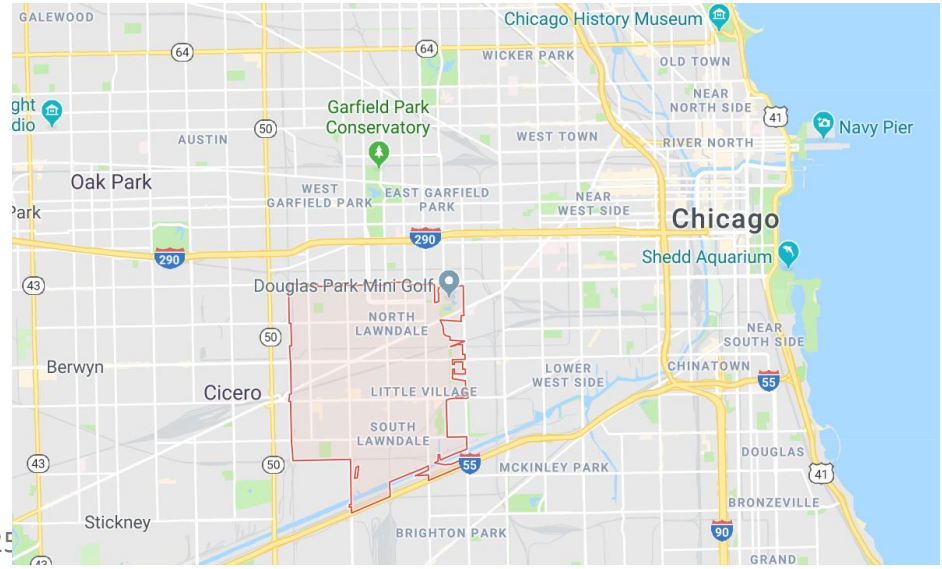
Neighborhoods in Zip Code 60623

☆ Worth Watching ★Premier Neighborhoods How to Earn a Star

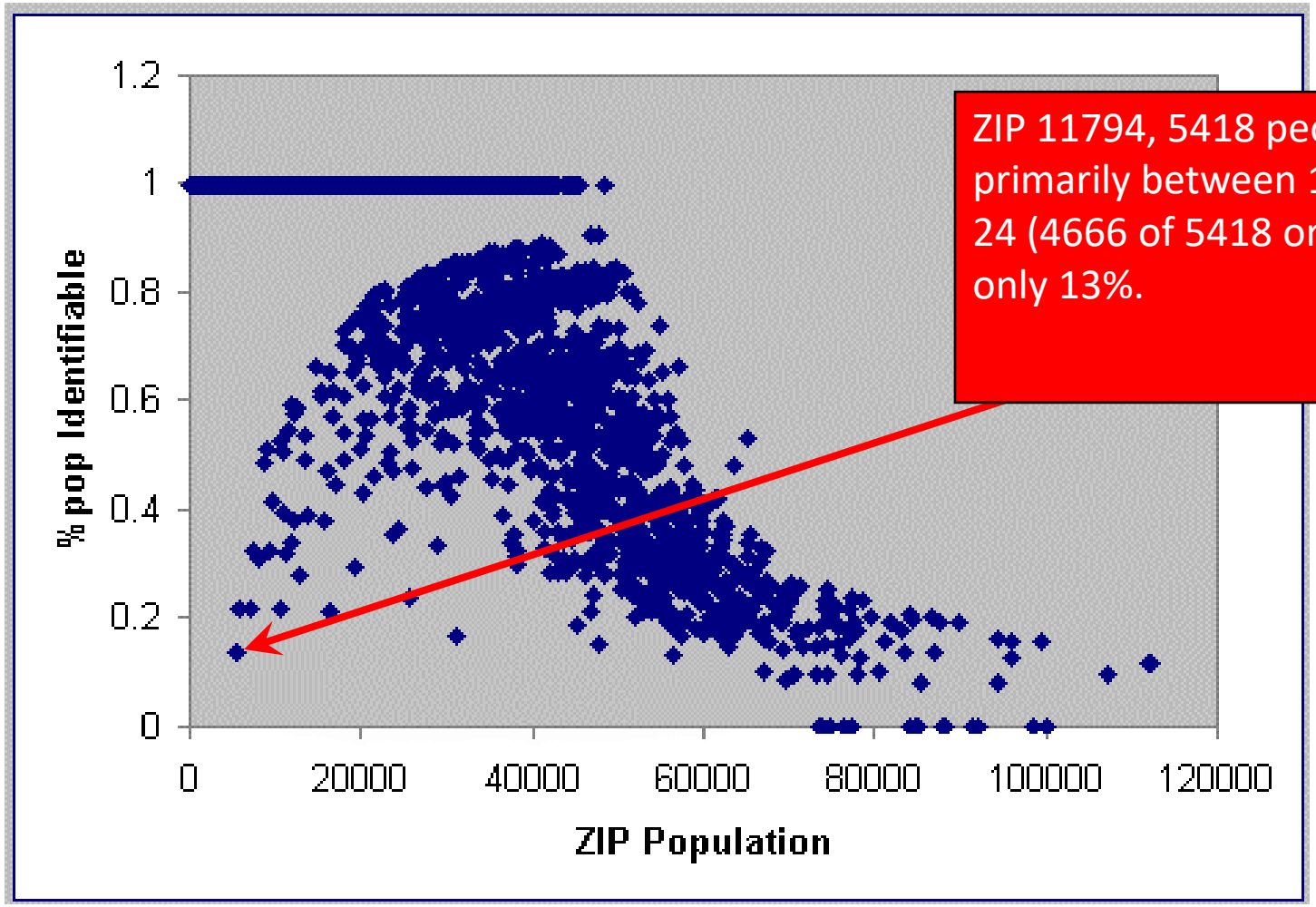
- Harrison Block Club
- Lawndale
- Little Village
- Near Westside Section 8
- Westside Association For Community Action

Don't See Your Neighborhood? Email Neighborhood Link

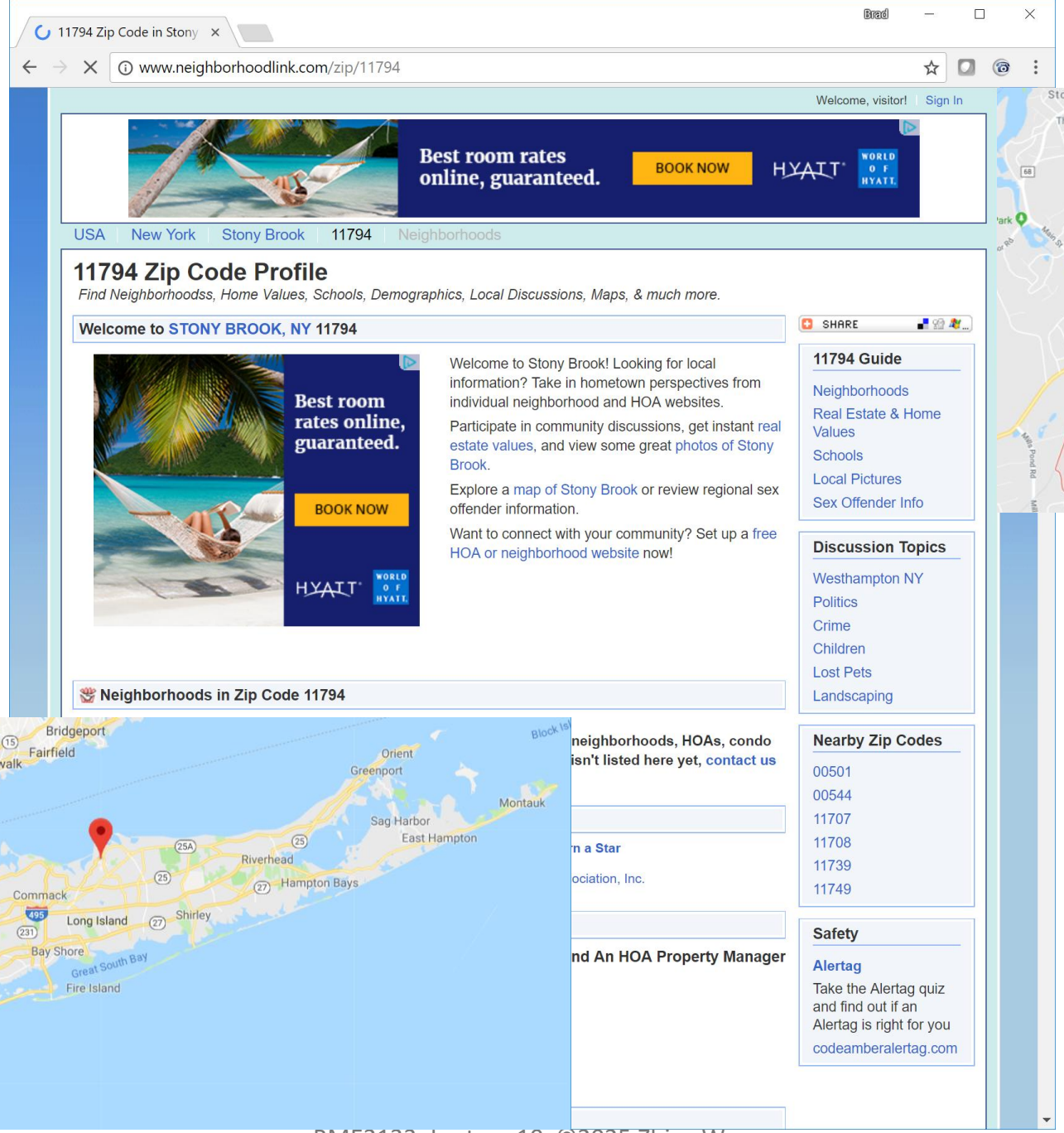


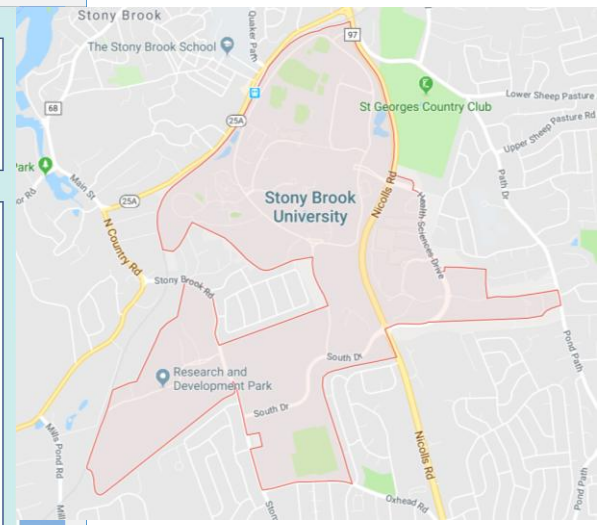


{date of birth, gender, 5-digit ZIP} (1990 US Census)



L. Sweeney. 2000.


11794 Zip Code in Stony
www.neighborhoodlink.com/zip/11794
Welcome, visitor! | Sign In
Best room rates online, guaranteed.
BOOK NOW
HYATT
WORLD OF HYATT
USA | New York | Stony Brook | 11794 | Neighborhoods
11794 Zip Code Profile
Find Neighborhoodss, Home Values, Schools, Demographics, Local Discussions, Maps, & much more.
Welcome to STONY BROOK, NY 11794
Best room rates online, guaranteed.
BOOK NOW
HYATT
WORLD OF HYATT
Welcome to Stony Brook! Looking for local information? Take in hometown perspectives from individual neighborhood and HOA websites.
Participate in community discussions, get instant real estate values, and view some great photos of Stony Brook.
Explore a map of Stony Brook or review regional sex offender information.
Want to connect with your community? Set up a free HOA or neighborhood website now!
SHARE
11794 Guide
Neighborhoods
Real Estate & Home Values
Schools
Local Pictures
Sex Offender Info
Discussion Topics
Westhampton NY
Politics
Crime
Children
Lost Pets
Landscaping
Nearby Zip Codes
00501
00544
11707
11708
11739
11749
Safety
Alertag
Take the Alertag quiz and find out if an Alertag is right for you
codeamberalertag.com
Neighborhoods in Zip Code 11794
neighborhoods, HOAs, condo
isn't listed here yet, contact us
n a Star
ociation, Inc.
nd An HOA Property Manager


Stony Brook
The Stony Brook School
St Georges Country Club
Stony Brook University
Research and Development Park
Nicolls Rd
N Country Rd
Main St
Stony Brook Rd
South Dr
Pond Path
Mills Pond Rd
Oxhead Rd

Bridgeport
Fairfield
Norwalk
Stamford
White Plains
Hudson River
Orient
Greenport
Montauk
Sag Harbor
East Hampton
Riverhead
Hampton Bays
Commack
Long Island
Shirley
Hicksville
Bay Shore
Great South Bay
Fire Island
Hempstead
Newark
New York
BRONX
MANHATTAN
QUEENS
BROOKLYN
Edison

Beyond the 87% Stat

Quasi-Identifier	Uniques
Date of Birth Only	12%

Beyond the 87% Stat

Quasi-Identifier	Uniques
Date of Birth Only	12%
Date of Birth & Gender	29%

Beyond the 87% Stat

Quasi-Identifier	Uniques
Date of Birth Only	12%
Date of Birth & Gender	29%
Date of Birth & 5-Digit Zip Code	69%

Beyond the 87% Stat

Quasi-Identifier	Uniques
Date of Birth Only	12%
Date of Birth & Gender	29%
Date of Birth & 5-Digit Zip Code	69%
Date of Birth & Full Postal Code	97%

A Model for Beyond Uniqueness

- D → the set of demographics
  e.g., $\{Date\ of\ Birth,\ Gender,\ Zip\ Code\}$

- d → a specific value combination over the set of demographics
  e.g., [1/2/1903, M, 65432]

- P → a population of individuals

G. Skinner and M. Elliot. Skinner G, Elliot M. A measure of disclosure risk for microdata. Journal of the Royal Statistical Society (Series B). 2002; 64(Part 4):855-867.

A Model for Beyond Uniqueness

- x_d →the number of people in P with demographic d
- F_i →the number of demographics with $|x_d| = i$.
  - F_1 = the number of demographics that correspond to exactly 1 person
  - F_2 = the number of demographics that correspond to exactly 2 people

Example

Record	Age	City	Gender	Race
1	25	aaa	M	Black
2	26	aaa	M	Black
3	25	aaa	F	White
4	24	zzz	M	White
5	22	zzz	F	Asian
6	22	zzz	F	White

{Age, City}

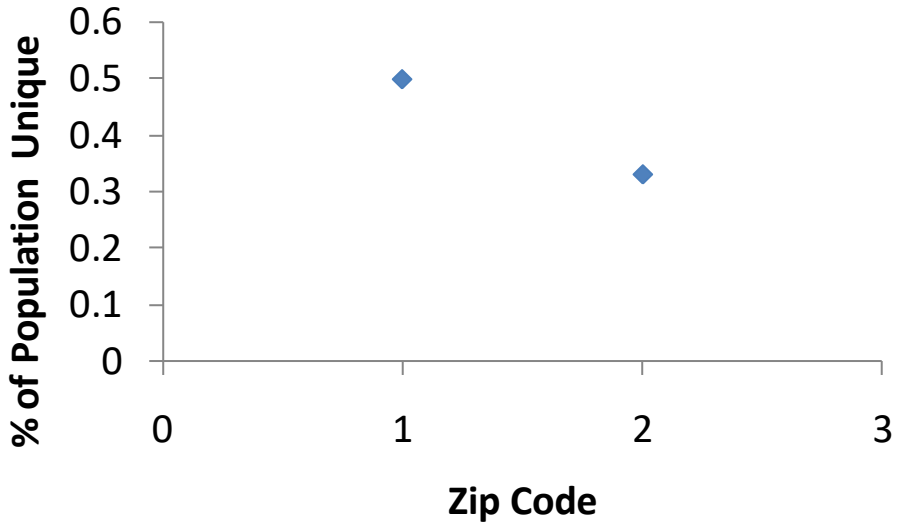
Record	Age	City	Gender	Race
1	25	aaa	M	Black
2	26	aaa	M	Black
3	25	aaa	F	White
4	24	zzz	M	White
5	22	zzz	F	Asian
6	22	zzz	F	White
7	24	aaa	F	White

Age	City	X_d
25	aaa	2
26	aaa	1
22	zzz	2
24	zzz	1
24	aaa	1

Uniqueness Per City

Age	City	X_d
25	aaa	2
26	aaa	1
22	zzz	2
24	zzz	1
24	aaa	1

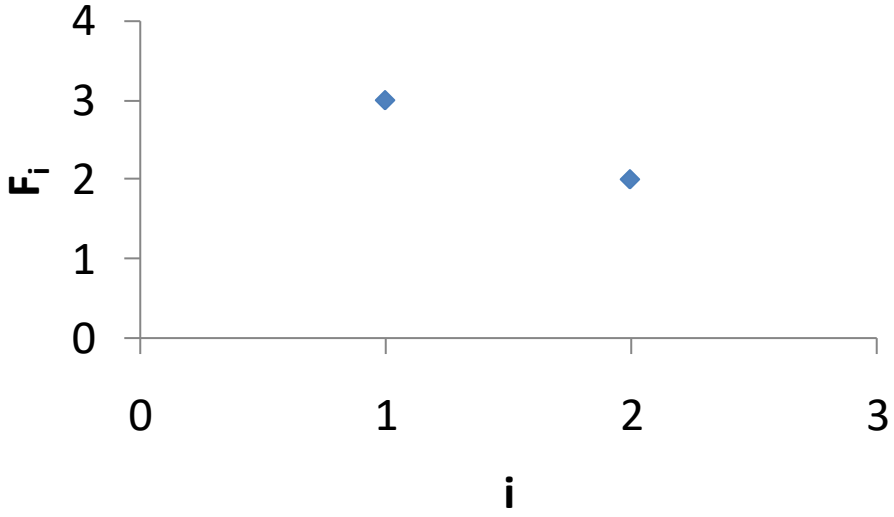
City	Percent Unique
aaa	2 / 4 = 50%
zzz	1 / 3 = 33%



Distribution Model for All Values

Age	City	X_d
25	aaa	2
26	aaa	1
22	zzz	2
24	zzz	1
24	aaa	1

i	F_i
1	3
2	2



Risk Analysis

- Let t be a threshold equal to the minimum number of people to which a demographic should correspond
- If privacy "threshold" is t, then the total number of people at risk is

$$\sum_{i=1}^{t} iF_i$$

- Then, the fraction of the population at risk is

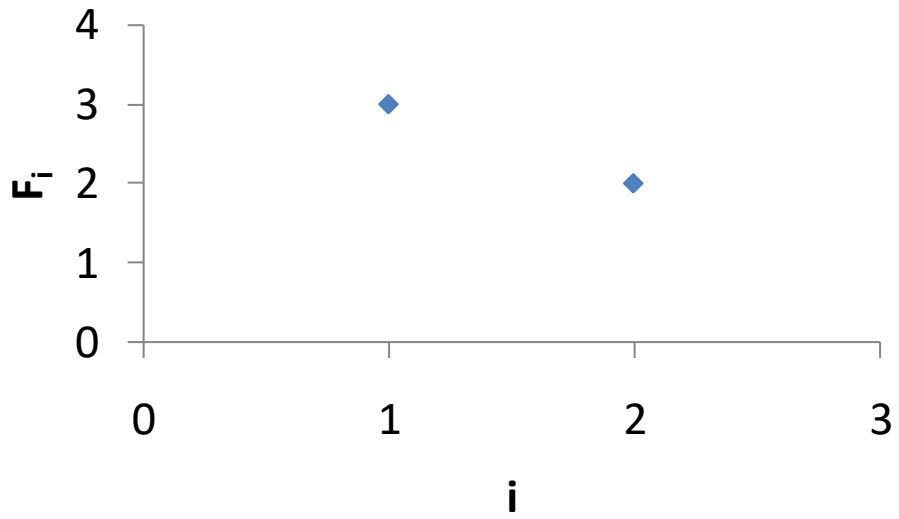
$$\frac{\sum_{i=1}^{t} iF_i}{|P|}$$

Distribution Model

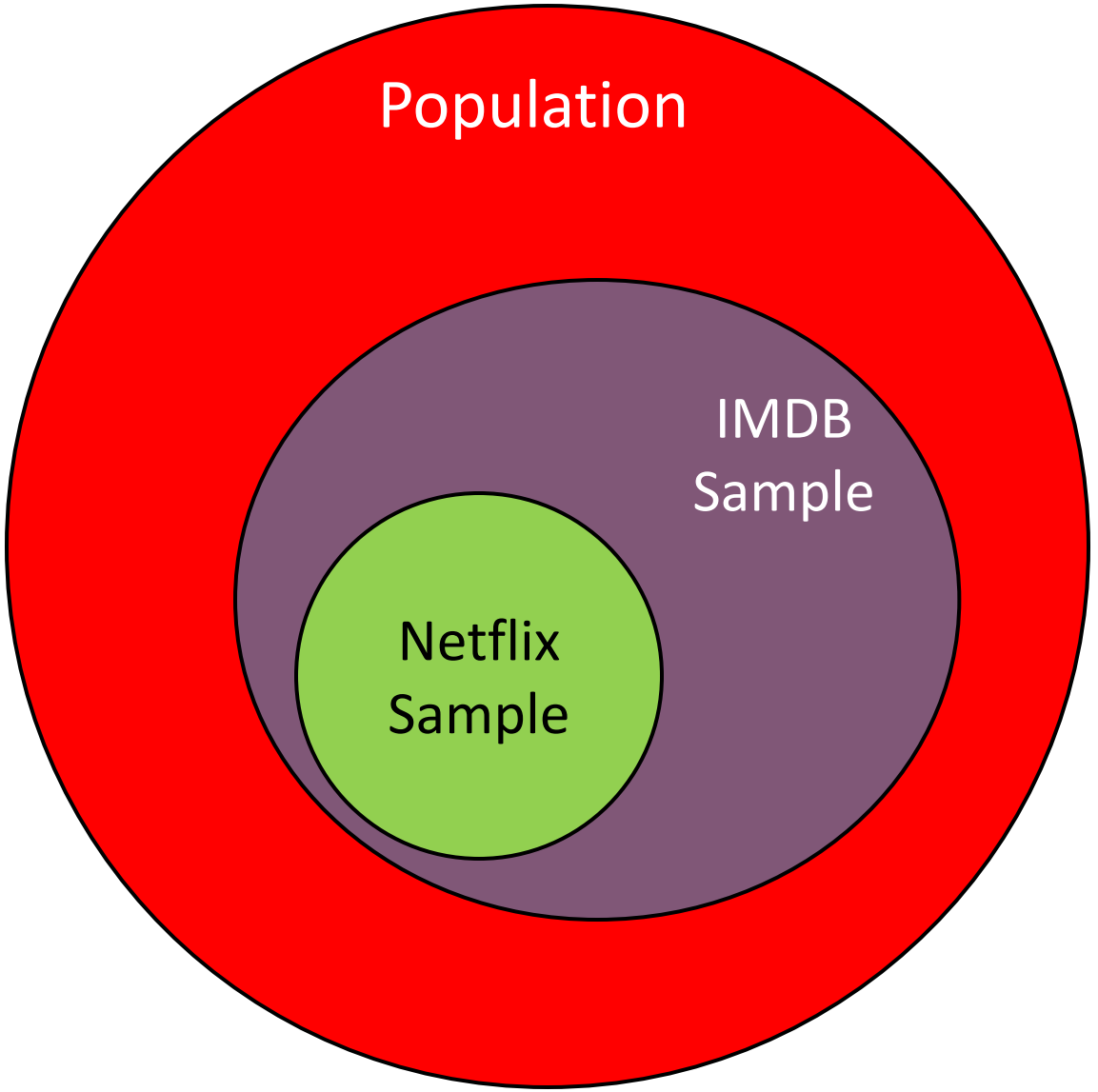
- t = 1:

  3*1 = 3 people at risk
  ~43% of population

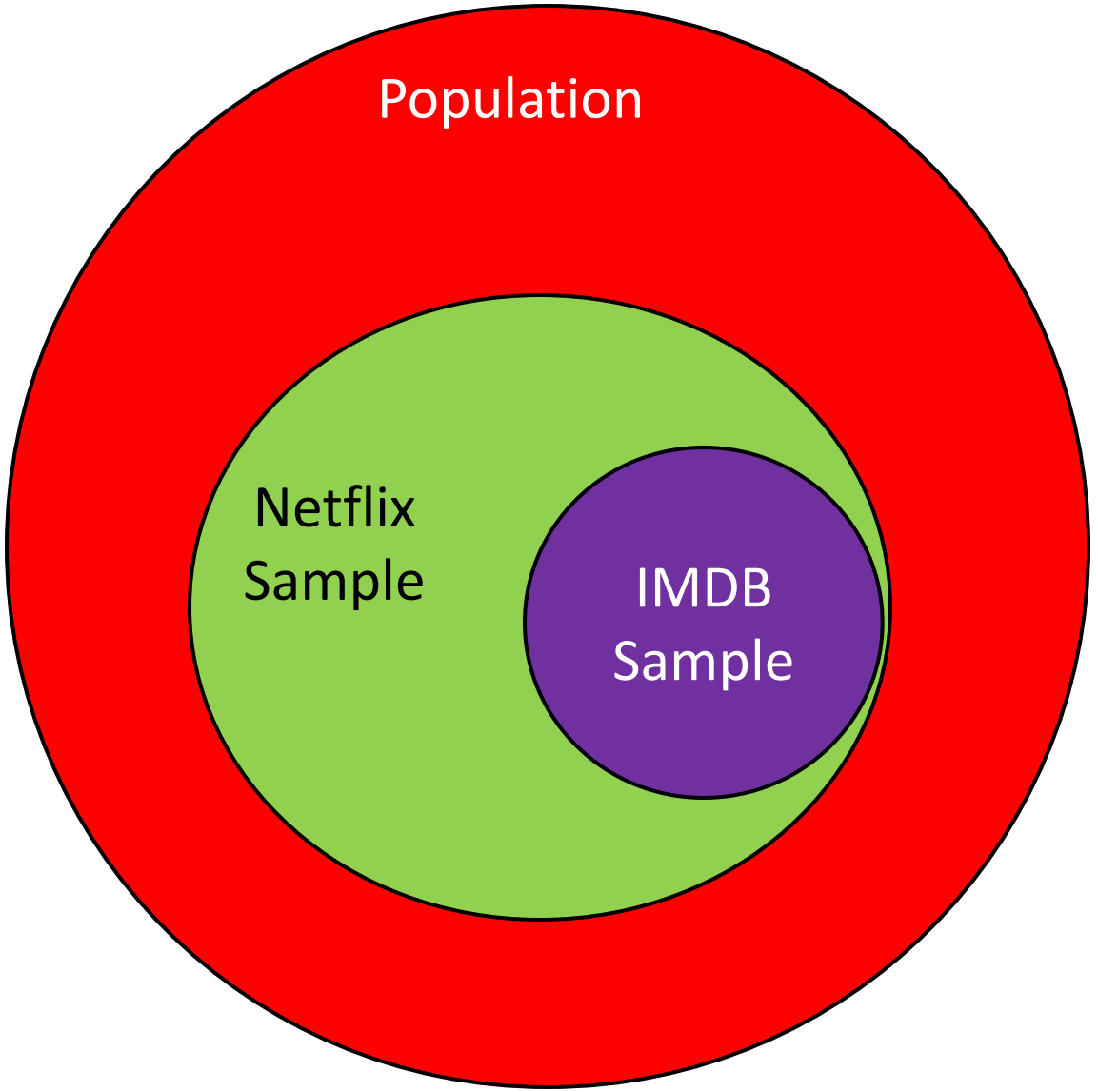
- t = 2, there are 3*1 + 2*2 = 7 people at risk ~100% of population



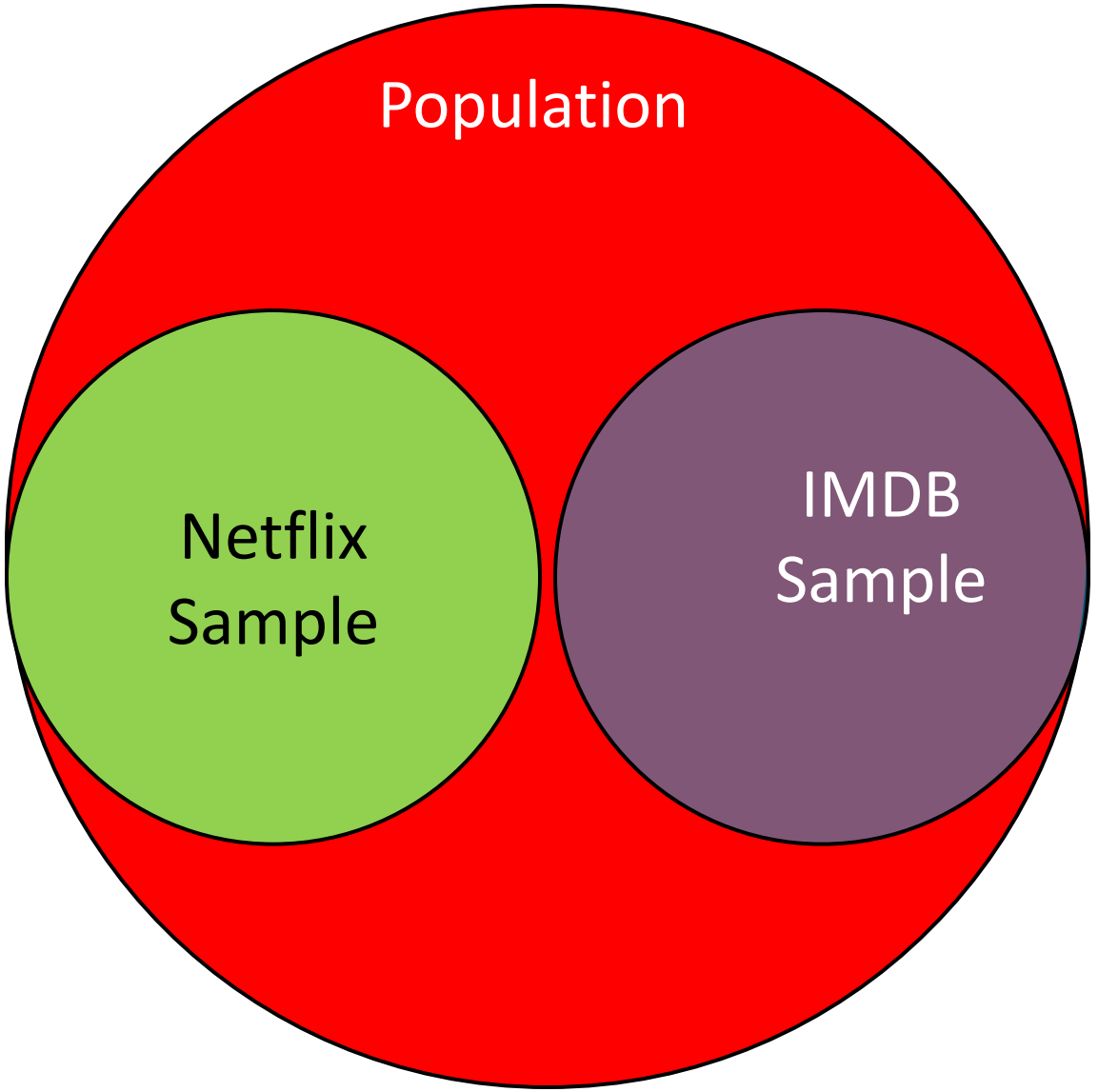
Re-identification ?



Re-identification ?



Re-identification ?



Randomized (Golle '06)

- Don't always have exact knowledge of what a data recipient has access to
  - Disclose sample with {*dob, gender, zip*}, but don't know the population's values
- May know population counts, such as
  - U.S. Census aggregates for {*year of birth, gender, county*}
- Conversion: {Year of Birth, ZIP} → {Date of Birth, ZIP}
- Alternative option: Randomly allocate 12,000 "people" to 365 cells

		Birth Year
		1980
ZIP	zip1	12000
	zip2	50000
	…	
	zip m	10000

		Birthdate			SUM
		1/1/80	…	12/31/80	
ZIP	zip1	random		random	12000
	zip2	random		random	50000
	…				
	zip m	random		random	10000

It's an Occupancy Problem (Golle '06)

Binomial Distribution $P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$

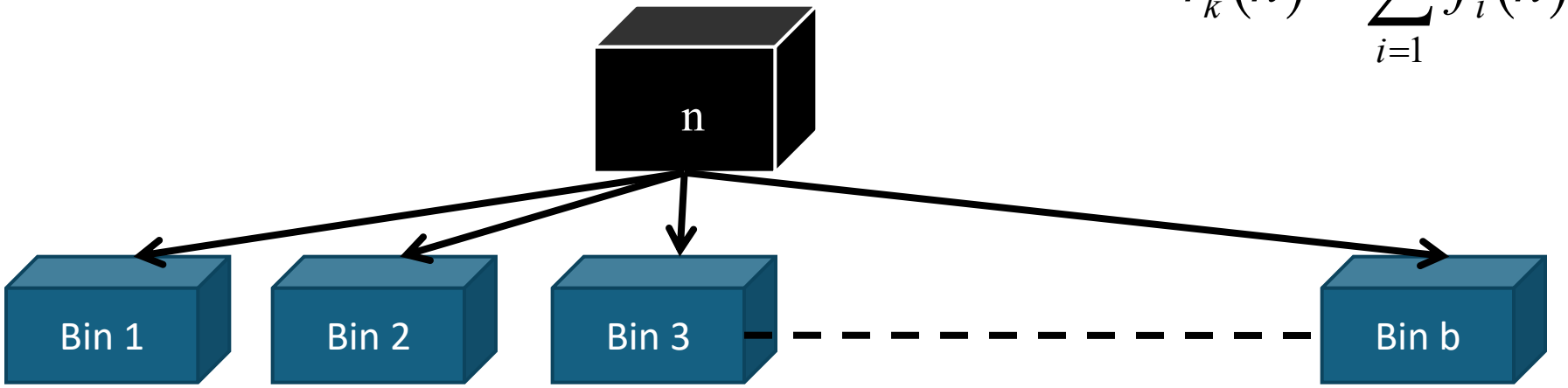
$$\binom{n}{i} = \frac{n!}{i!\,(n-i)!}$$

- n people in aggregated bin
- b disaggregated bins
- the expected # of bins with exactly i people

$$f_i(n) = \binom{n}{i} b^{1-n} (b-1)^{n-i}$$

- Total number of people in a group of size less than k

$$r_k(n) = \sum_{i=1}^{k-1} f_i(n)$$



Sample Calculation

		Quasi-ID values (Bins)					
		2	**4**	**256**	**512**	**1024**	**8192**
Population (Balls)	**2**	0.5	2.25	254.01	510.00	1022.00	8190.00
	4	0.125	1.26	252.03	508.02	1020.01	8188.00
	64	1.08×10^{-19}	4.04×10^{-8}	199.37	451.84	961.96	8128.25
	1024	0.00	0.00	4.69	69.29	376.71	7229.41
	2048	0.00	0.00	0.09	9.38	138.58	6379.94

Expected Number of Quasi-ID values with 0 people

Sample Calculation

Quasi-ID values (Bins)

Population (Balls)	2	4	256	512	1024	8192
2	0.250	0.563	0.992	0.996	0.998	1.000
4	0.063	0.315	0.984	0.992	0.996	1.000
64	0.000	0.000	0.779	0.883	0.939	0.992
1024	0.000	0.000	0.018	0.135	0.368	0.882
2048	0.000	0.000	0.000	0.018	0.135	0.779

Expected Ratio of Quasi-ID values with 0 people

Poisson Approximation

- When $b \to \infty, n \to \infty, \frac{n}{b} = \lambda$

$$f_i(n) = \binom{n}{i} b^{1-n} (b-1)^{n-i}$$

$$\approx b \frac{e^{-\lambda} \lambda^i}{i!}$$

Poisson Distribution

Birthday Problem

- Assume birthday is uniformly distributed at random over the year.
- If n people are born in a year, the expected number of days on which exactly 1 person born is

$$f_1(n) = n * \left(\frac{364}{365}\right)^{n-1}$$

Golle's Approach

- Special case of general equation
- If *n* people are born in a year, the expected # of days on which exactly *k* people born is

$$f_k(n) = \binom{n}{k} 365^{1-n} 364^{n-k}$$

Golle's Findings

- Results with the 2000 Census
  - Table PCT12: year of birth
  - Counties + County Equivalents
  - 33,233 Zip Code Tabulation Areas (ZCTAs), DC, plus Puerto Rico

- Uniqueness Calculations

Variable	5-Digit Zip	County
Birth Year	0.2%	0.0%
Birth Year & Month	4.2%	0.2%
Birth Date	63.3%	14.8%

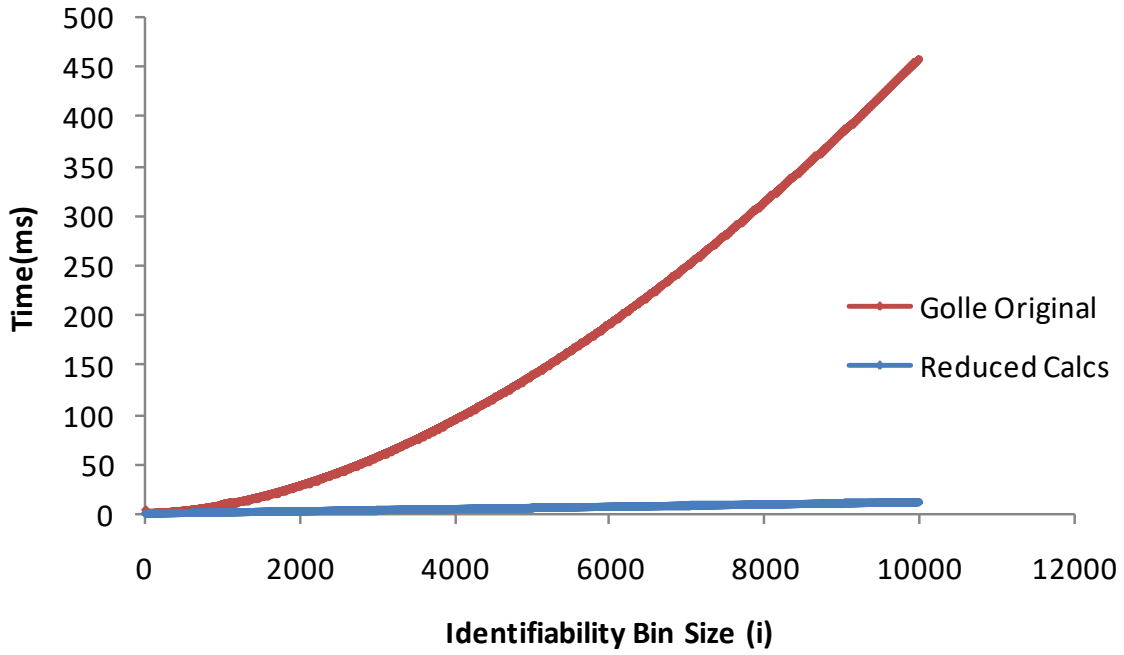
Application Challenges to Computations

- No exact closed form expression for the truncated binomial
- Can apply recursion to speed up exact computation

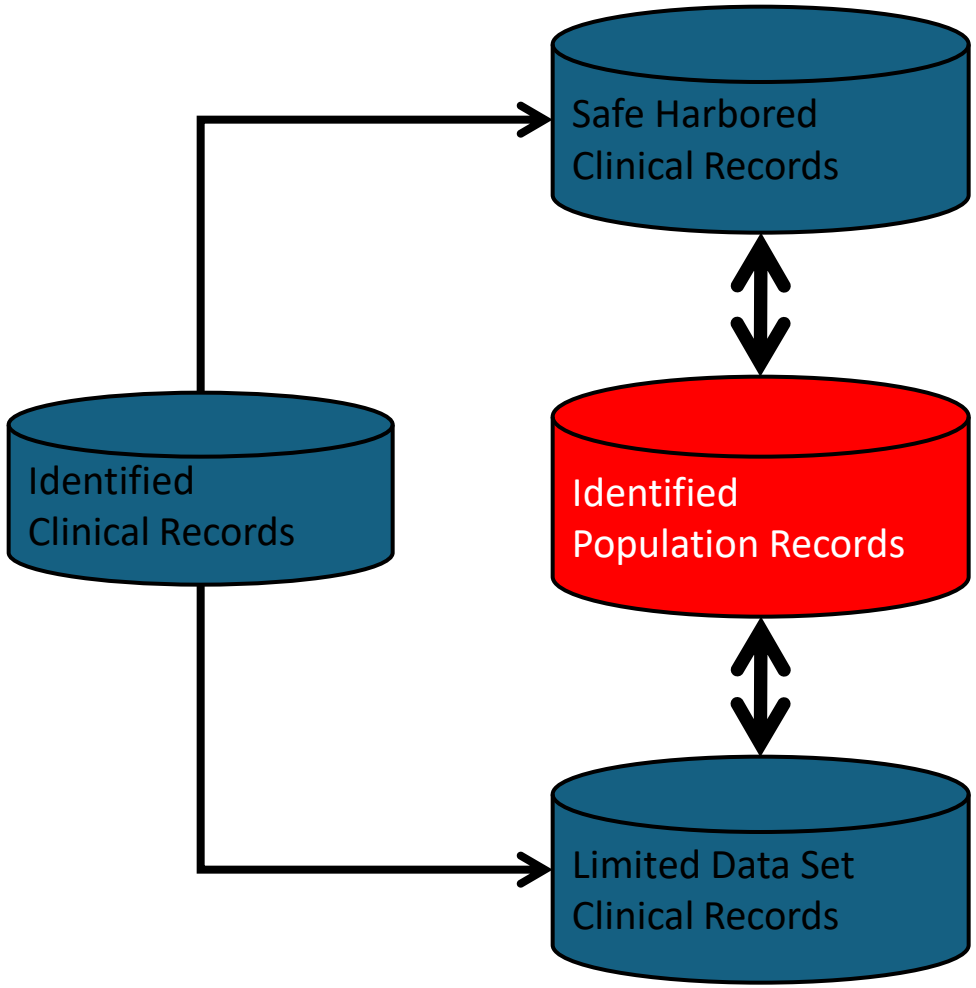
$$r_k(n) = n\left(\frac{b}{b-1}\right)^{1-n} \sum_{i=1}^{k} \prod_{j=1}^{i-1} \frac{n-j}{j(b-1)}$$

Or Just use the Poisson to approximate the result

$$\Pr[X_i = k] \approx \frac{1}{k!}\left(\frac{n}{b}\right)^k e^{-n/b}$$



Attacks on Demographics

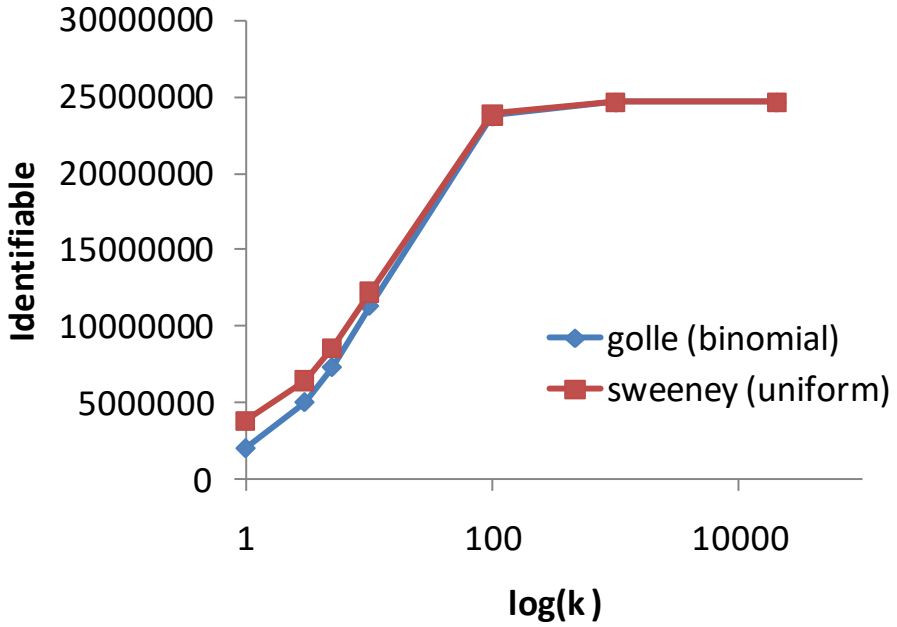


- Use population counts from the US Census
- Apply a statistical model to estimate distribution of disaggregate demographics
- It's not perfect, but it's a start.

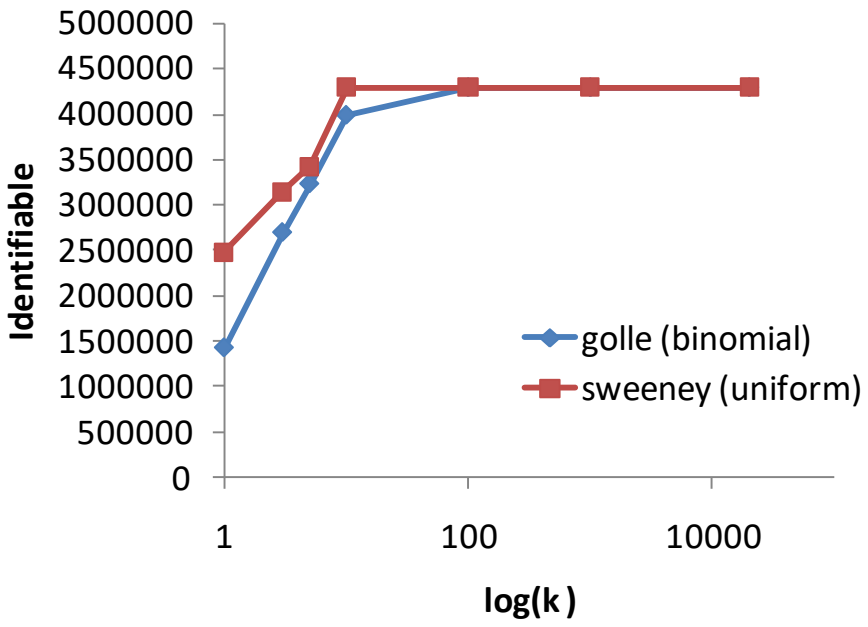
K. Benitez and B. Malin. Evaluating re-identification risks with respect to the HIPAA Privacy Rule. JAMIA. 2010; 17: 169-177.

Comparison

- {Date of Birth, Gender, Race, County}
  - Date predicted from Year

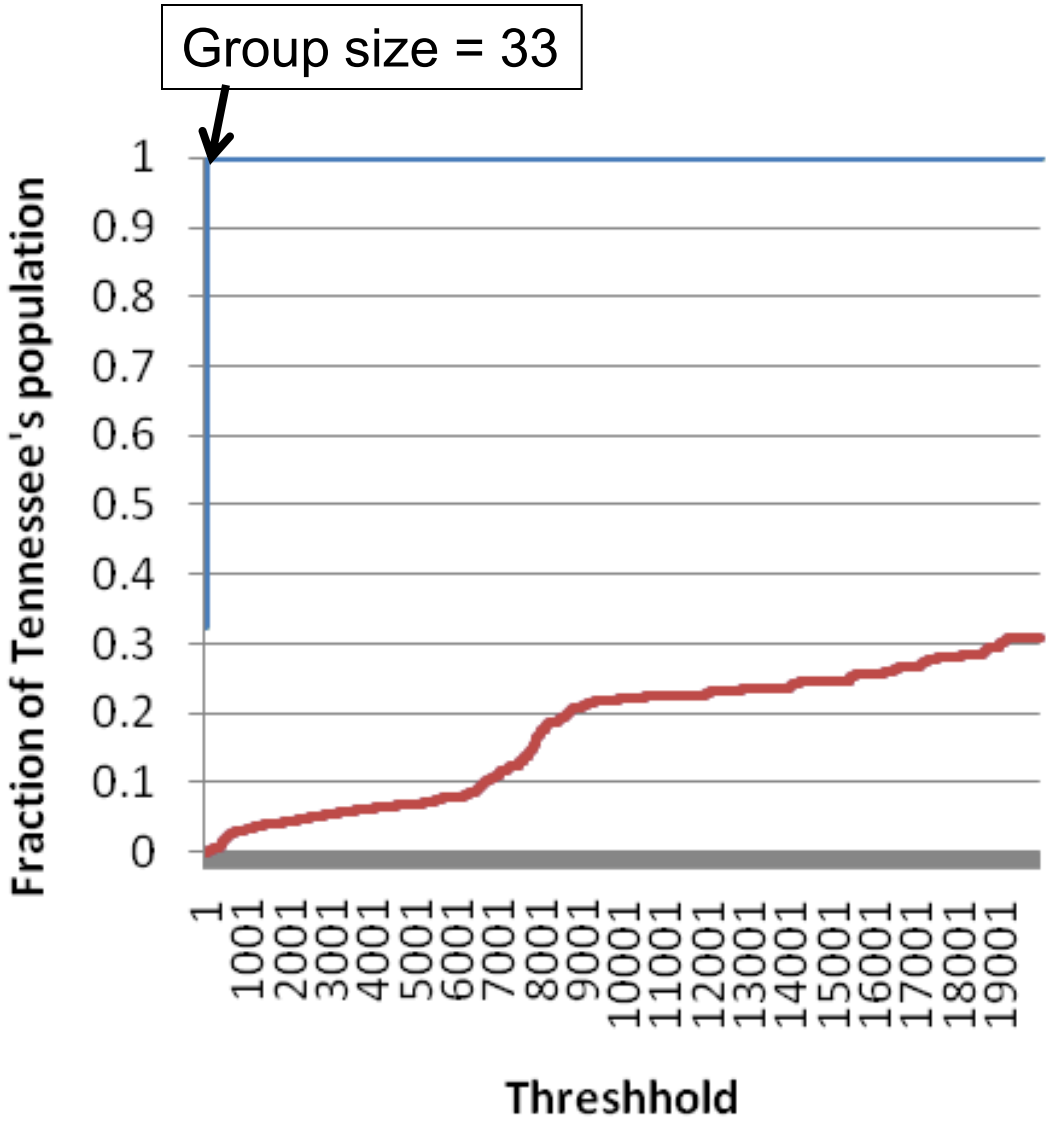


California



Tennessee

Case Study: Tennessee



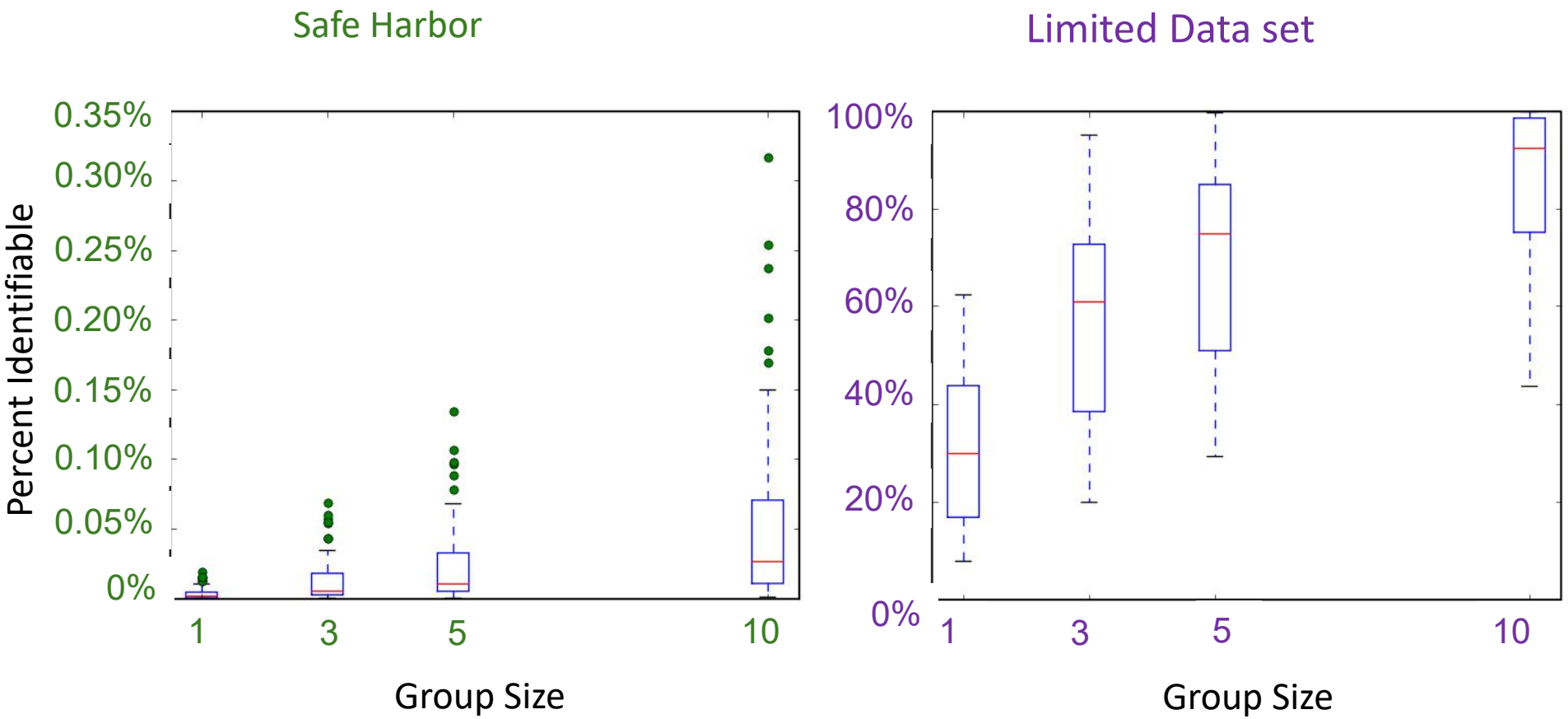
Limited Dataset
{Race, Gender, Date (of Birth), County}

Fraction of population identifiable under Limited Dataset

Fraction of populatiion identifiable under Safe Harbor

Safe Harbor(-like)
{Race, Gender, Year (of Birth), State}

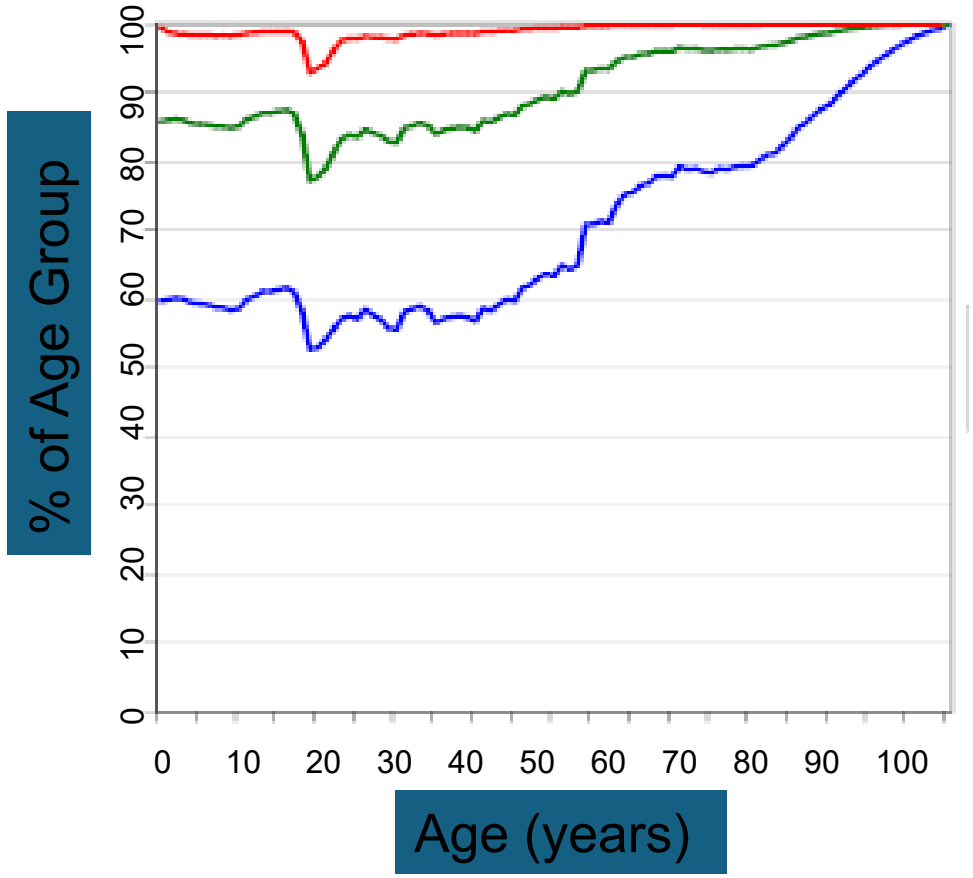
All U.S. States



K. Benitez and B. Malin. Evaluating re-identification risks with respect to the HIPAA Privacy Rule. JAMIA. 2010; 17: 169-177.

Beyond Uniqueness - Golle

- Given {Date of Birth, Gender, Zip}, grouped by age



no more than 5 people

no more than 2 people

unique

Readings due on November 12

- None.
- Optional
  - None.

Feedback Survey

- One thing you learned or felt was valuable from today's class & reading

- Muddiest point: what, if anything, feels unclear, confusing or "muddy"

- https://www.wjx.cn/vm/hX0mIro.aspx