Statistical De-identification of Public Use Data Sets

Daniel C. Barth-Jones, M.P.H., Ph.D.
Assistant Professor of Clinical Epidemiology,
Mailman School of Public Health
Columbia University
E-mail: db2431@Columbia.edu

State Public Use File Re-identification Attacks

- Repeated health data re-identification attacks against State public use data resources and the associated negative publicity can generate considerable fear of data re-identification risks, which can put in jeopardy the many important public benefits created by State health data resources.
- All three States (WA, ME, VT) data sets attacked were **not** de-identified according to the HIPAA de-identification standard.
- All three data sets were **not** protected by an effective data use agreement prohibiting re-identification attempts.

WA State Hospital Discharge Attack

Bloomberg Our Company | Professional | Anywhere

HOME QUICK NEWS OPINION MARKET DATA PERSONAL FINANCE TECH POLITICS SUSTA

Frustrated Republicans Pressure Boehner to End Shutdown

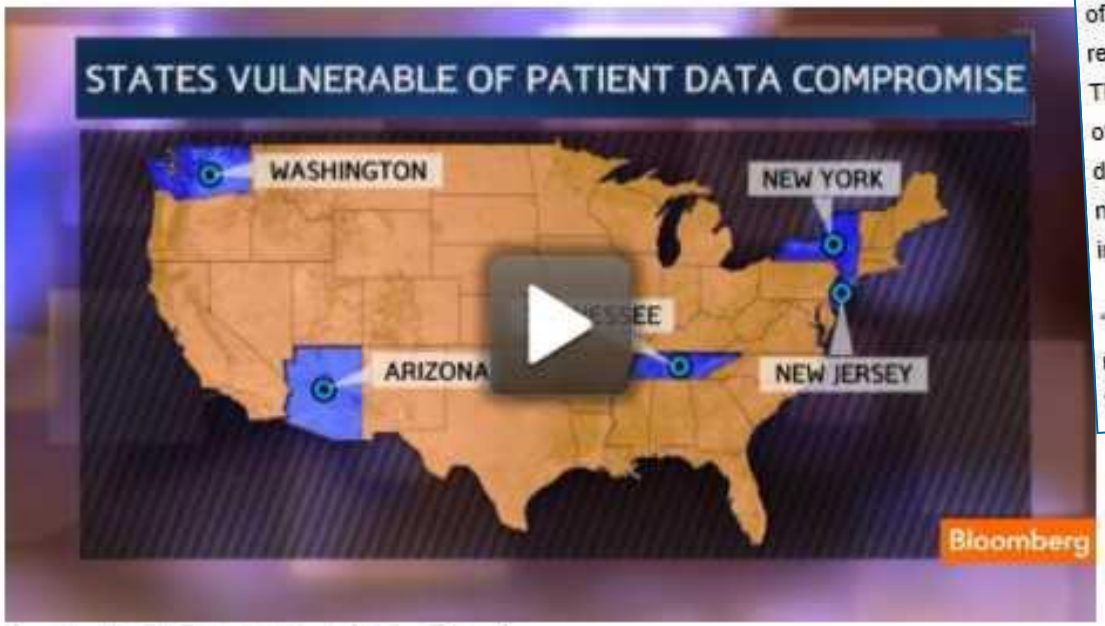
Shutdown Jokes, Day 3: Letterman, Colbert, Stewart

BREAKING NEWS Telecom Italia Ceo Bernabe Is Said to Resign

States' Hospital Data for Sale Puts Privacy in Jeopardy

By Jordan Robertson - Jun 5, 2013 12:01 AM ET

113 COMMENTS + QUEUE



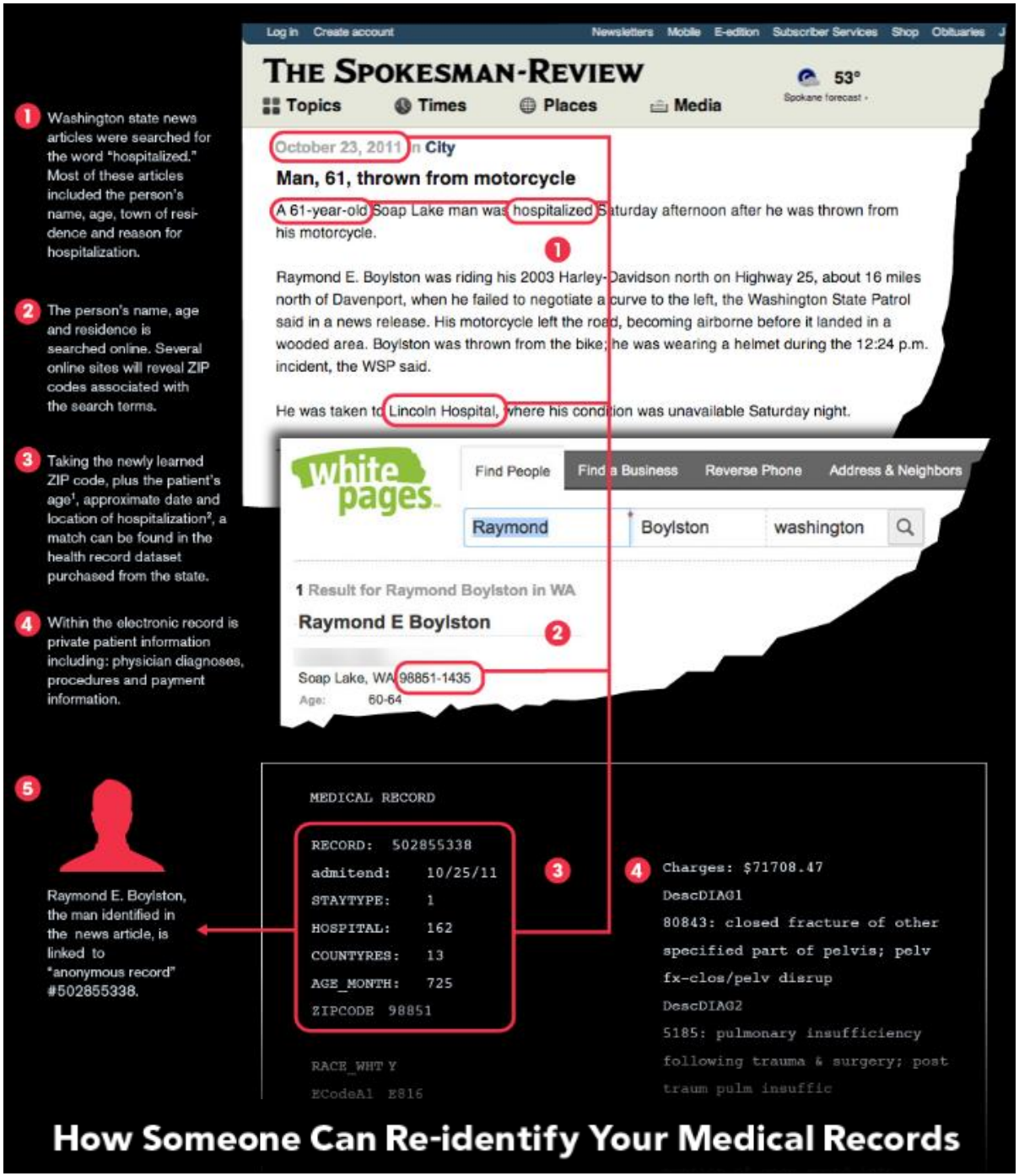
Your Health Data for Sale: Who's Selling, Buying?

Consider Ray Boylston, who went into diabetic shock while riding his motorcycle in rural Washington in 2011. He careened off the road and was thrown into the woods, an accident that was covered only briefly, in the local newspaper. Boylston disclosed his medical condition and history to a handful of loved ones and the hospital that treated him.

After Boylston's discharge, Washington collected the paperwork of his week-long stay from Providence Sacred Heart Medical Center in Spokane and added it to a database of 650,000 hospitalizations for 2011 available for sale to researchers, companies and other members of the public. The data was supposed to remain anonymous. Yet because of state exemption from federal regulations governing discharge information, Boylston could be identified and his medical background exposed using only publicly available information.

"I don't really feel that the public has a right to read up on my medical history," said Boylston, who is 62 and a veteran. "I feel I've been violated."

40/648,384
= 1/16,200


1 Washington state news articles were searched for the word "hospitalized." Most of these articles included the person's name, age, town of residence and reason for hospitalization.
2 The person's name, age and residence is searched online. Several online sites will reveal ZIP codes associated with the search terms.
3 Taking the newly learned ZIP code, plus the patient's age[1], approximate date and location of hospitalization[2], a match can be found in the health record dataset purchased from the state.
4 Within the electronic record is private patient information including: physician diagnoses, procedures and payment information.
5
Raymond E. Boylston, the man identified in the news article, is linked to "anonymous record" #502855338.
Log in Create account Newsletters Mobile E-edition Subscriber Services Shop Obituaries
THE SPOKESMAN-REVIEW
53°
Spokane forecast
Topics Times Places Media
October 23, 2011 n City
Man, 61, thrown from motorcycle
A 61-year-old Soap Lake man was hospitalized Saturday afternoon after he was thrown from his motorcycle.
Raymond E. Boylston was riding his 2003 Harley-Davidson north on Highway 25, about 16 miles north of Davenport, when he failed to negotiate a curve to the left, the Washington State Patrol said in a news release. His motorcycle left the road, becoming airborne before it landed in a wooded area. Boylston was thrown from the bike; he was wearing a helmet during the 12:24 p.m. incident, the WSP said.
He was taken to Lincoln Hospital, where his condition was unavailable Saturday night.
white pages
Find People Find a Business Reverse Phone Address & Neighbors
Raymond Boylston washington
1 Result for Raymond Boylston in WA
Raymond E Boylston
Soap Lake, WA 98851-1435
Age: 60-64
MEDICAL RECORD
RECORD: 502855338
admitend: 10/25/11
STAYTYPE: 1
HOSPITAL: 162
COUNTYRES: 13
AGE_MONTH: 725
ZIPCODE 98851
RACE_WHT Y
ECodeA1 E816
Charges: $71708.47
DescDIAG1
80843: closed fracture of other specified part of pelvis; pelv fx-clos/pelv disrup
DescDIAG2
5185: pulmonary insufficiency following trauma & surgery; post traum pulm insuffic
How Someone Can Re-identify Your Medical Records

Yoo J, Thaler A, Sweeney L, and Zang J. Risks to Patient Privacy: a re-identification of patients in Maine and Vermont statewide hospital data. Technology Science. 2018100901. October 9, 2018.
http://techscience.org/a/2018100901

Risks to Patient Privacy: A Re-identification of Patients in Maine and Vermont Statewide Hospital Data

Ji Su Yoo, Alexandra Thaler, Latanya Sweeney, and Jinyan Zang

Highlights

- We used newspaper data to match names to anonymized patient records in statewide hospital data from Maine and Vermont
- We found that 28.3 percent of names from Maine news stories and 34 percent of names from Vermont news stories uniquely matched to one hospitalization in the Maine and Vermont hospital data.
- When redacted to the HIPAA Safe Harbor standard, the Maine data allowed for a 3.2 percent re-identification rate and Vermont data allowed for a 10.6 percent re-identification rate.
- Our results suggest that states should revisit de-identification practices and reassess risks to patient privacy when determining data sharing protocol

Sweeney L. Only You, Your Doctor, and Many Others May Know. *Technology Science*. 2015092903. September 29, 2015.http://techscience.org/a/2015092903

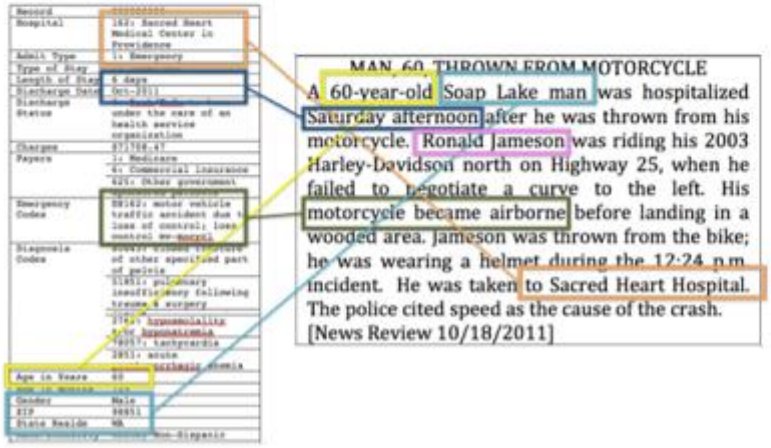
Technology Science

Only You, Your Doctor, and Many Others May Know

Latanya Sweeney

Highlights

- Washington State is one of 33 states that share or sell anonymized health records
- I conducted an example re-identification study by showing how newspaper stories about hospital visits in Washington State leads to identifying the matching health record 43% of the time
- This study resulted in Washington State increasing the anonymization protocols of the health records including limiting fields used for the re-identification study



Matching public medical information to news stories to identify patients.

Washington, Maine, Vermont Re-identification Risks

Washington Headline: "43% Re-identified"

Reality: 35 / 648,384 = 0.0054% or ***1 in 18,525***
(or 40 / 648,384 = 0.0062% or 1 in 16,210)

Maine Headline: "28% Re-identified"

Reality: 69 / 105,808 = 0.065% or ***1 in 1,533***

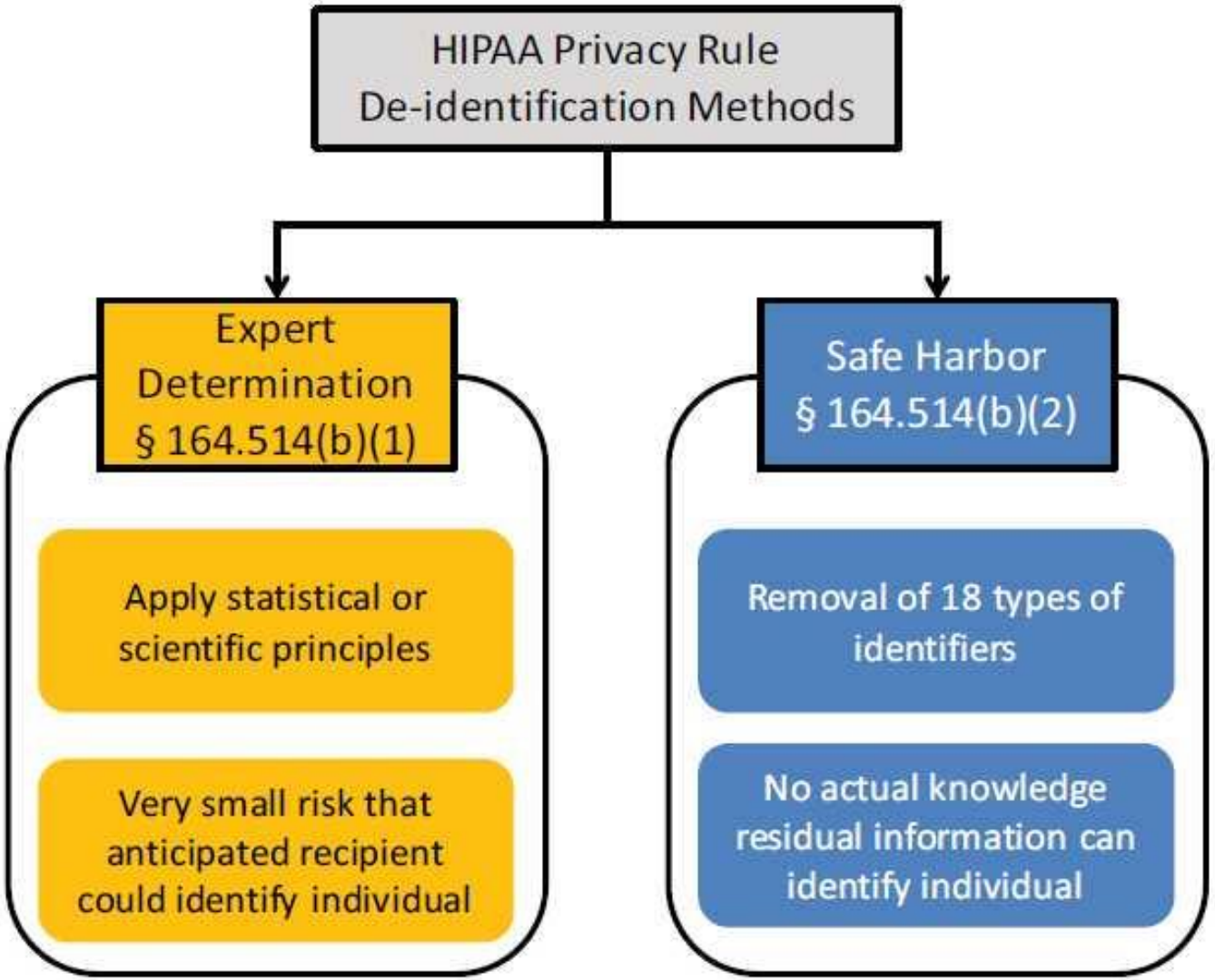
Vermont Headline: "34% Re-identified"

Reality: 16 / 268,984 = 0.008% or ***1 in 16,811***

Overall Public Use Sets

Reality: (40+69+16)/1,023,176 = 0.0012% or ***1 in 8,185***

Two Methods of HIPAA De-identification

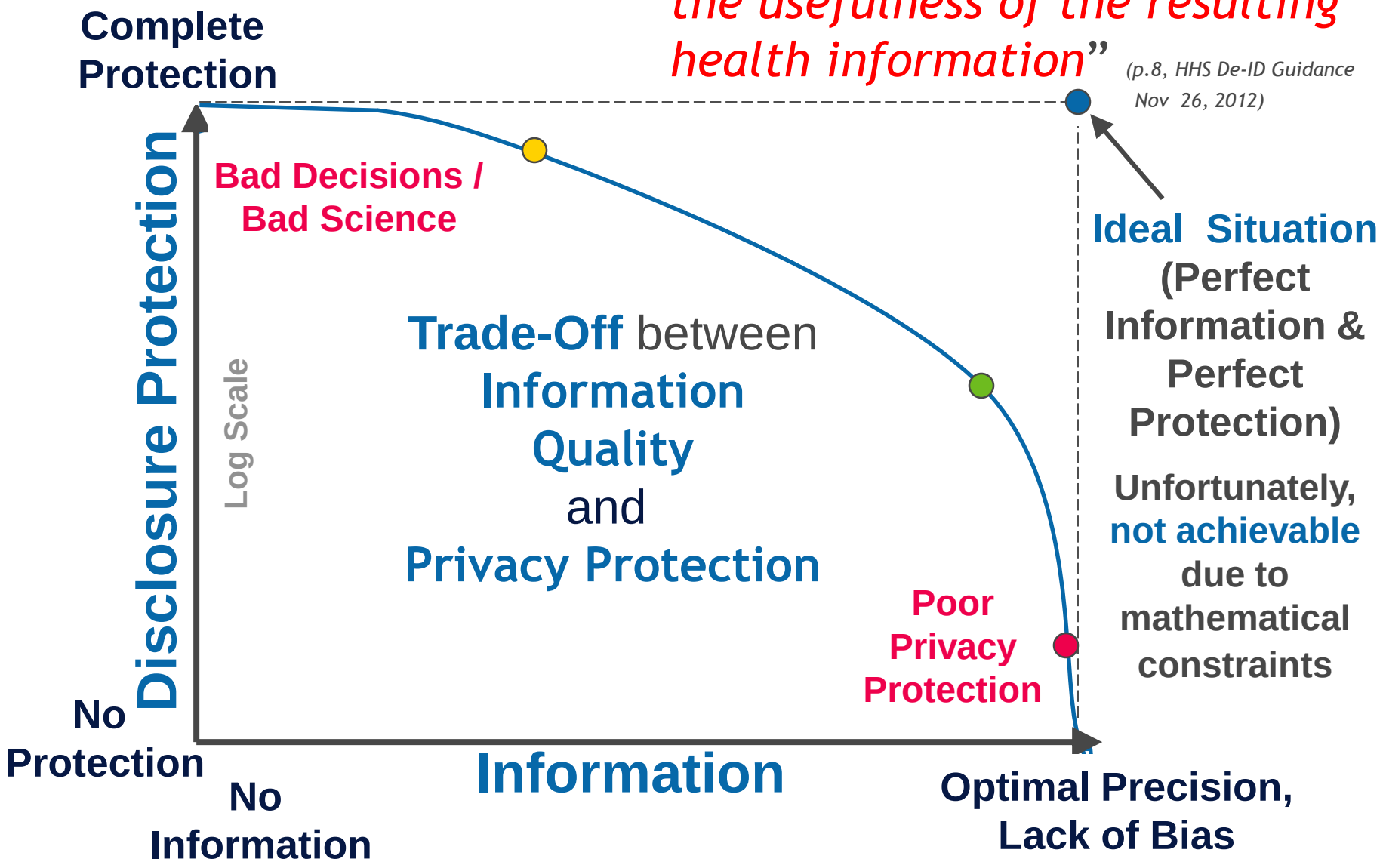


HIPAA Expert Determination Conditions

- "Risk is ***very small…***"
  - "that the ***information could be used***"…
  - "alone or ***in combination with other reasonably available information***"…,
  - "***by an anticipated recipient***"…
  - "***to identify an individual***"…

The Inconvenient Truth:

"De-identification leads to information loss which may limit the usefulness of the resulting health information" (p.8, HHS De-ID Guidance Nov 26, 2012)



Essential Re-identification Concepts

- Essential Re-identification and Statistical Disclosure Concepts
  - Record Linkage
  - Linkage Keys (Quasi-identifiers)
  - *Sample Uniques* and *Population Uniques*
- Straightforward Methods for Controlling Re-identification Risk
  - Decreasing Uniques:
    - by Reducing Key Resolutions
    - by Increasing Reporting Population Sizes

Quasi-identifiers

While individual fields may not be identifying by themselves, the contents of several fields in combination may be sufficient to result in identification, the set of fields in the Key is called the set of ***Quasi-identifiers***.

Fields that should be considered part of a Quasi-identifier are those variables which would be likely to exist in "reasonably available" data sets along with actual identifiers (names, etc.).

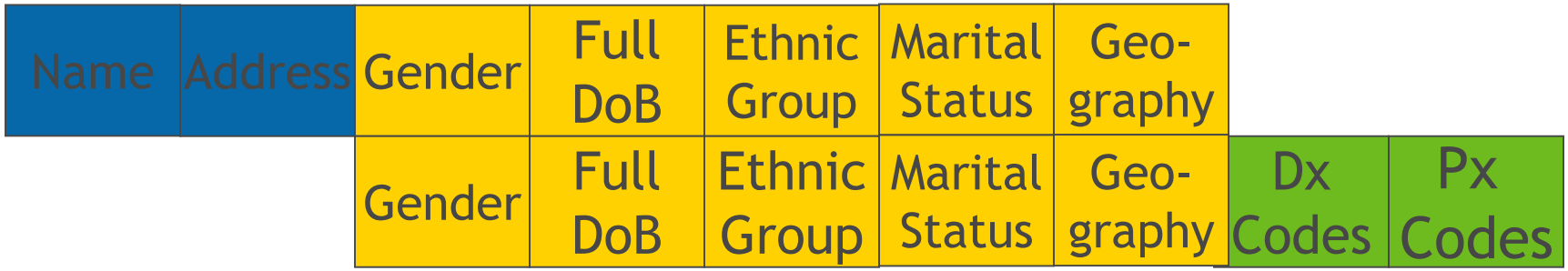
Note that this includes even fields that are not "PHI".

Key Resolution

Key "*resolution*" increases with:

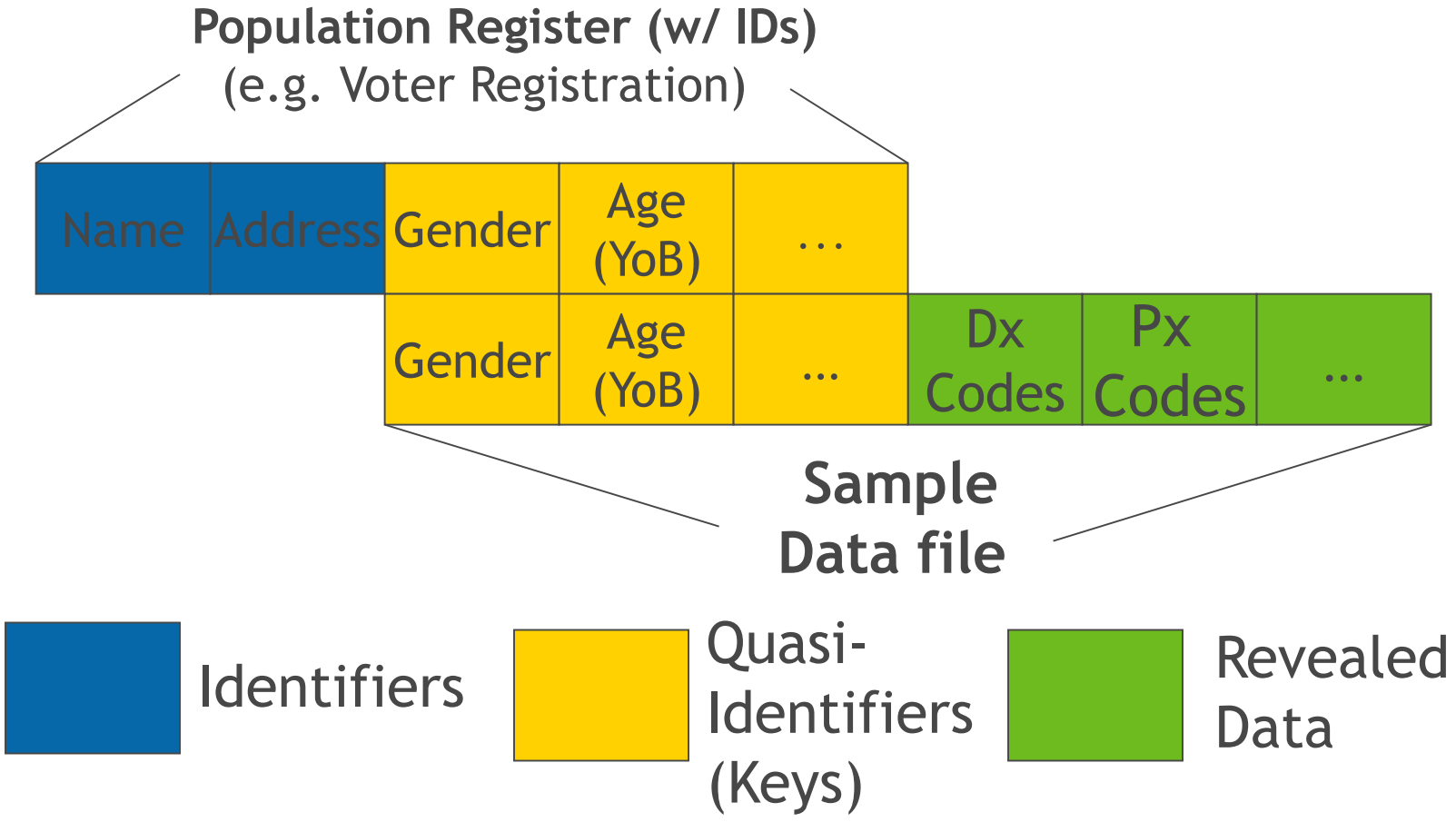
1) the number of matching fields available

2) the level of detail within these fields. (e.g. Age in Years versus complete Birth Date: Month, Day, Year)



Record Linkage

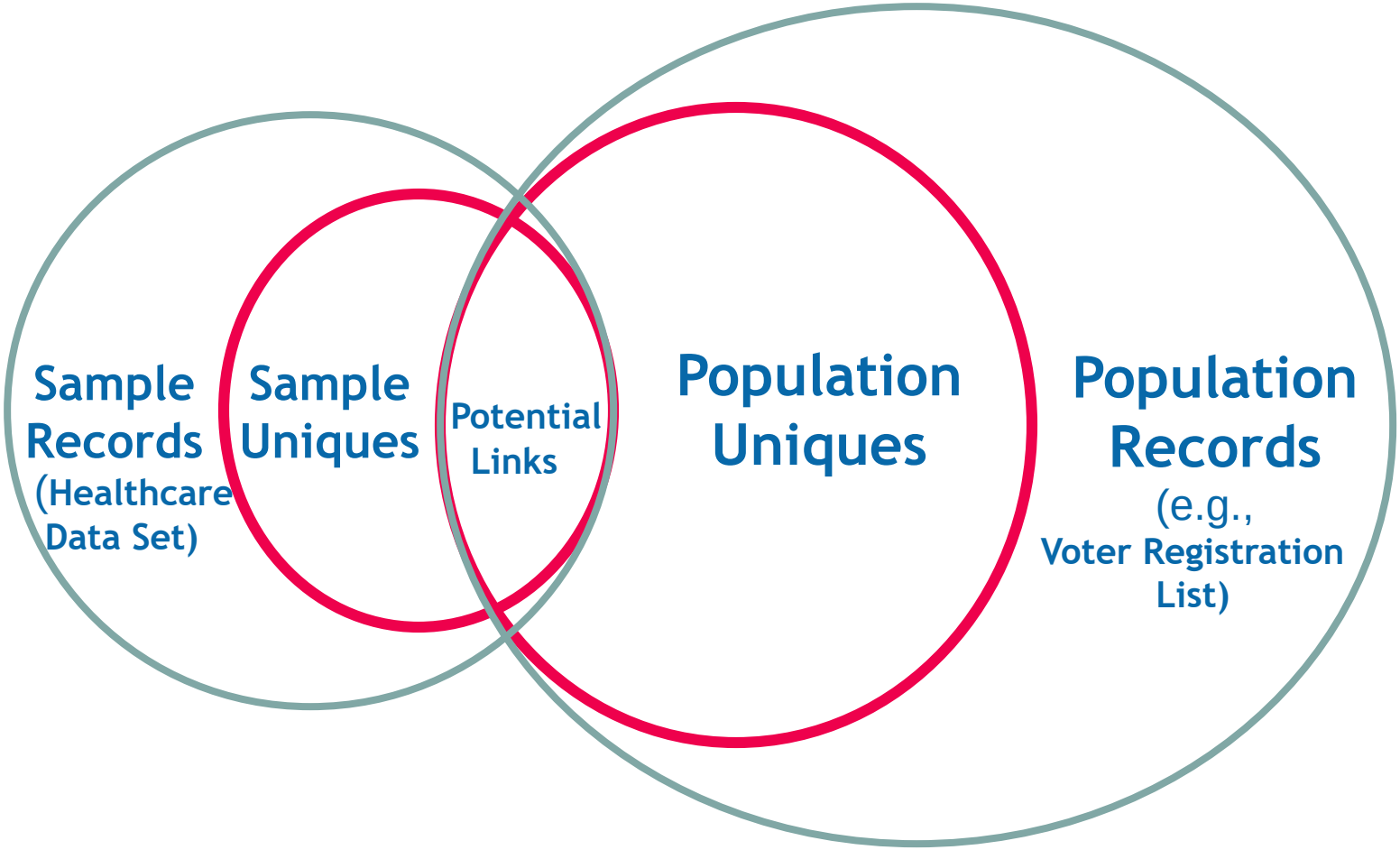
Record Linkage is achieved by matching records in separate data sets that have a common “Key” or set of data fields.



Sample and *Population Uniques*

- When only one person with a particular set of characteristics exists within a given data set (typically referred to as the ***sample*** data set), such an individual is referred to as a "***Sample Unique***".
- When only one person with a particular set of characteristics exists within the entire population or within a defined area, such an individual is referred to as a "***Population Unique***".

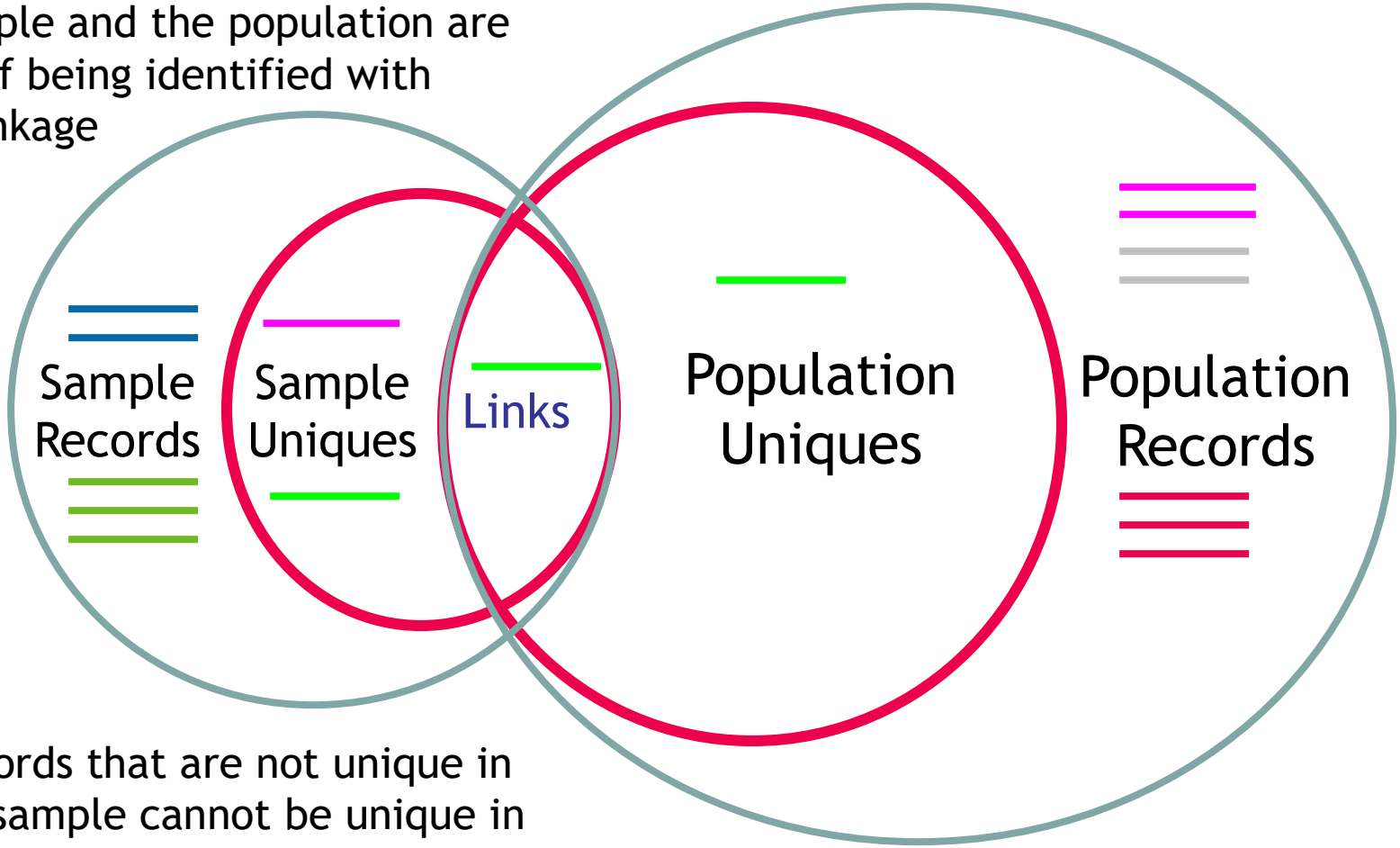
Measuring Disclosure Risks



Linkage Risks

Records that are unique in the sample but which aren't unique in the population, would match with more than one record in the population, and only have a probability of being identified

Only records that are unique in the sample and the population are at risk of being identified with exact linkage



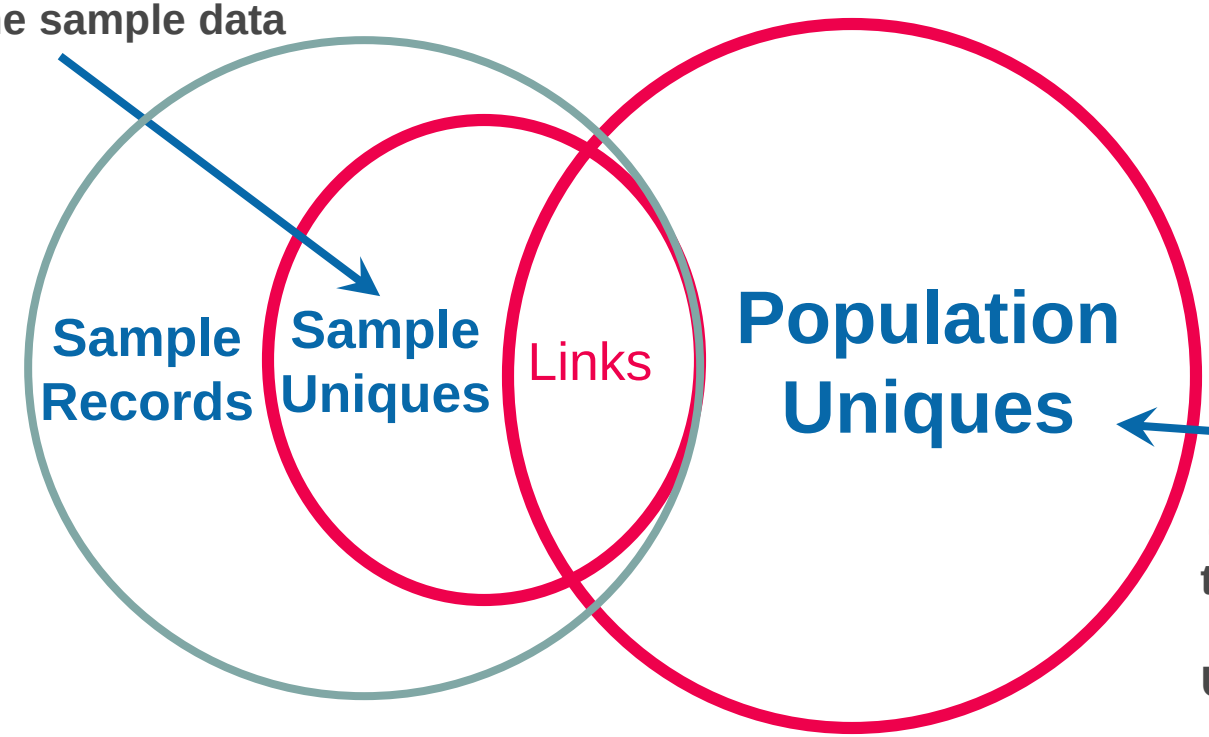
Records that are not unique in the sample cannot be unique in the population and, thus, aren't at definitive risk of being identified

Records that are not in the sample also aren't at risk of being identified

Estimating Disclosure Risks

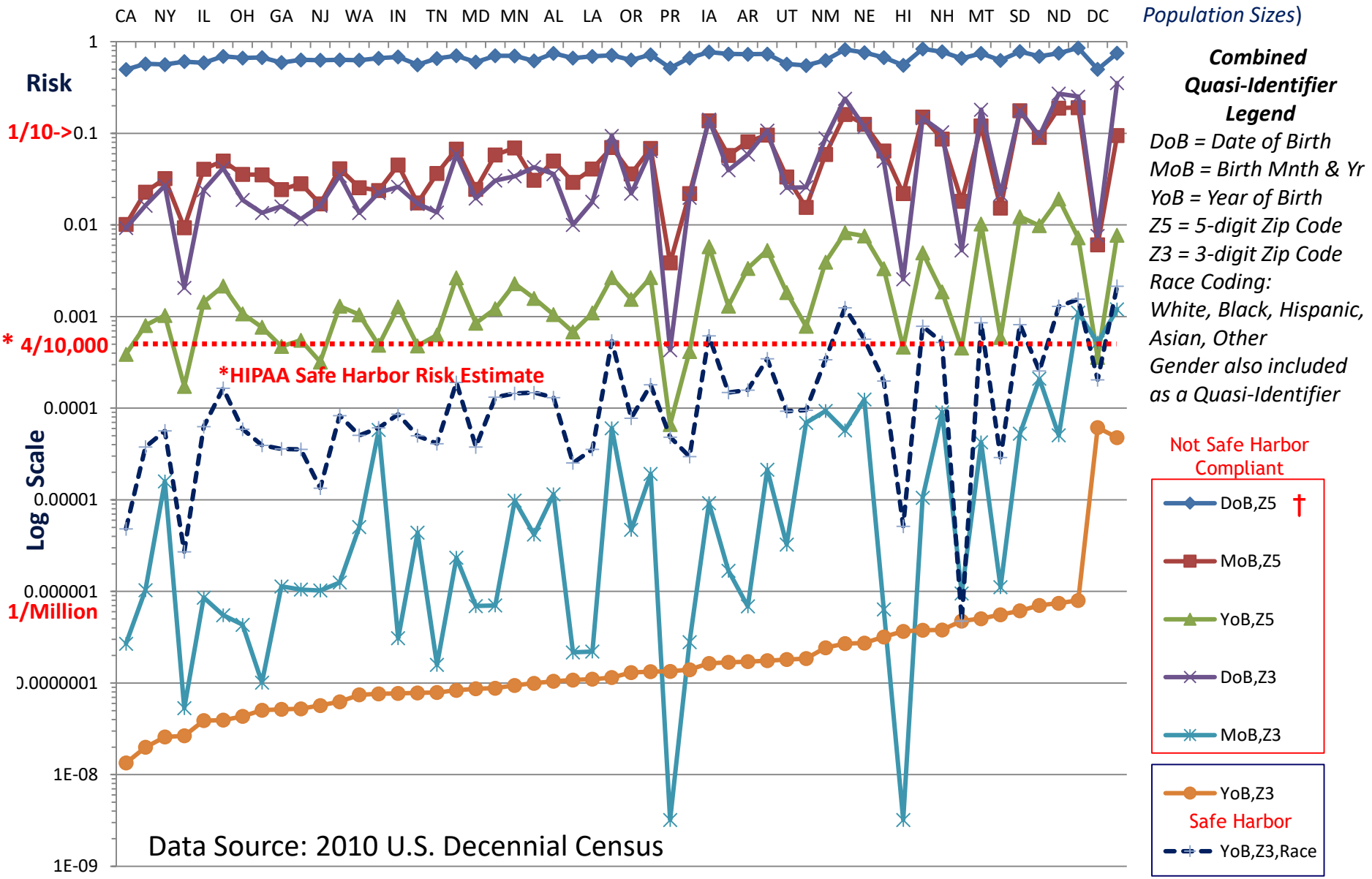
We can determine the Sample Uniques quite easily from the sample data

Links / Sample Records indicates the risk of record linkage.



For many characteristics, the likelihood of Population Uniqueness can be estimated from statistical models of the US Census data

U.S. State Specific Re-identification Risks: Population Uniqueness



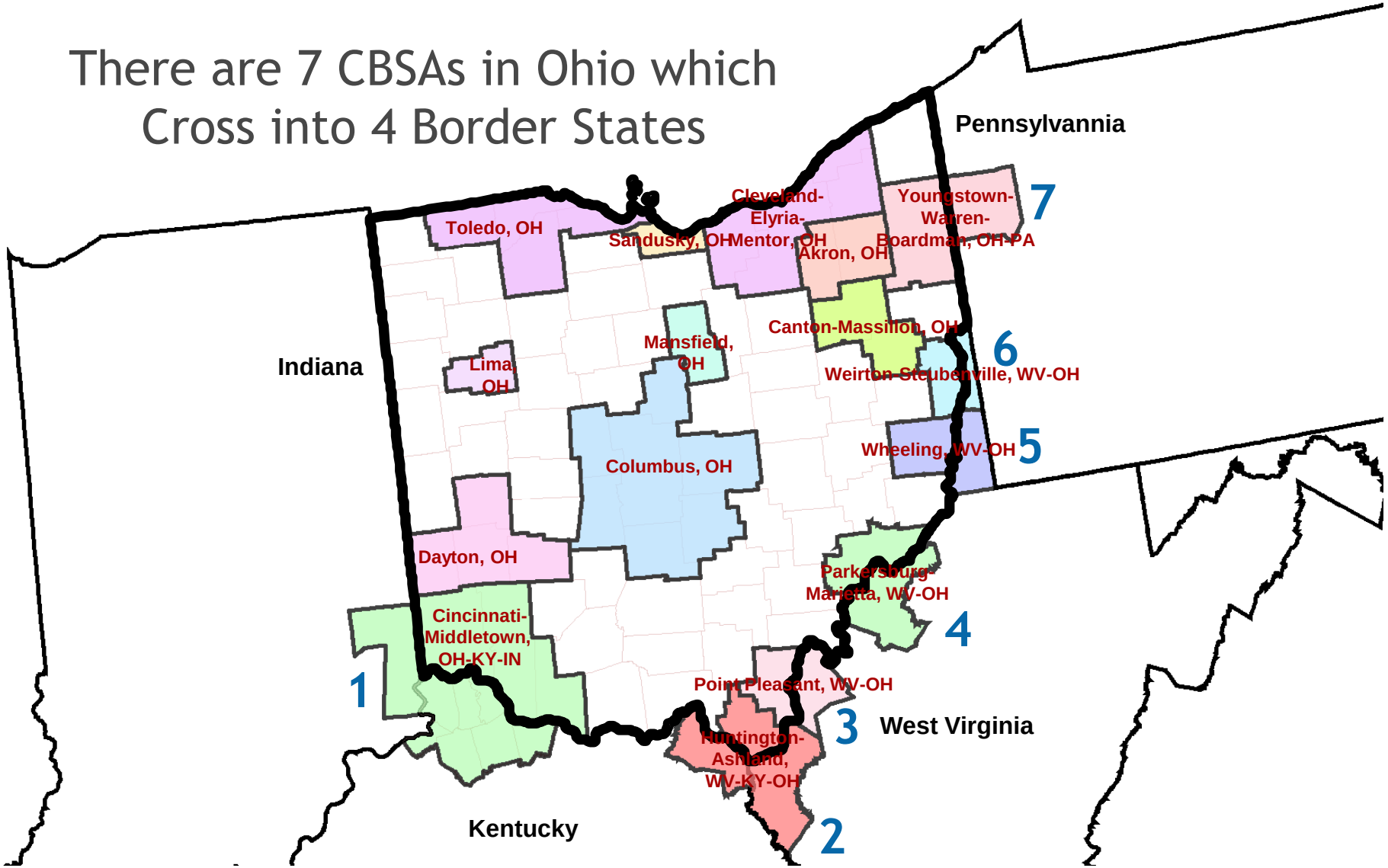
† *HIPAA Safe Harbor does not permit any Dates more specific than the year, or Geographic Units smaller than 3-digit Zip Codes (Z3).*

Challenge: Subtraction Geography (i.e., Geographical Differencing)

- Challenge: Data recipients often request reporting on more than one geography (e.g., both State and 3 digit Zip code).
- ***Subtraction Geography*** creates disclosure risk problems when more than one geography is reported for the same area and the geographies overlap.
- Also called ***geographical differencing***, this problem occurs when the multiple overlapping geographies are used to reveal smaller areas for re-identification searches.

Example: OHIO Core-based Statistical Areas

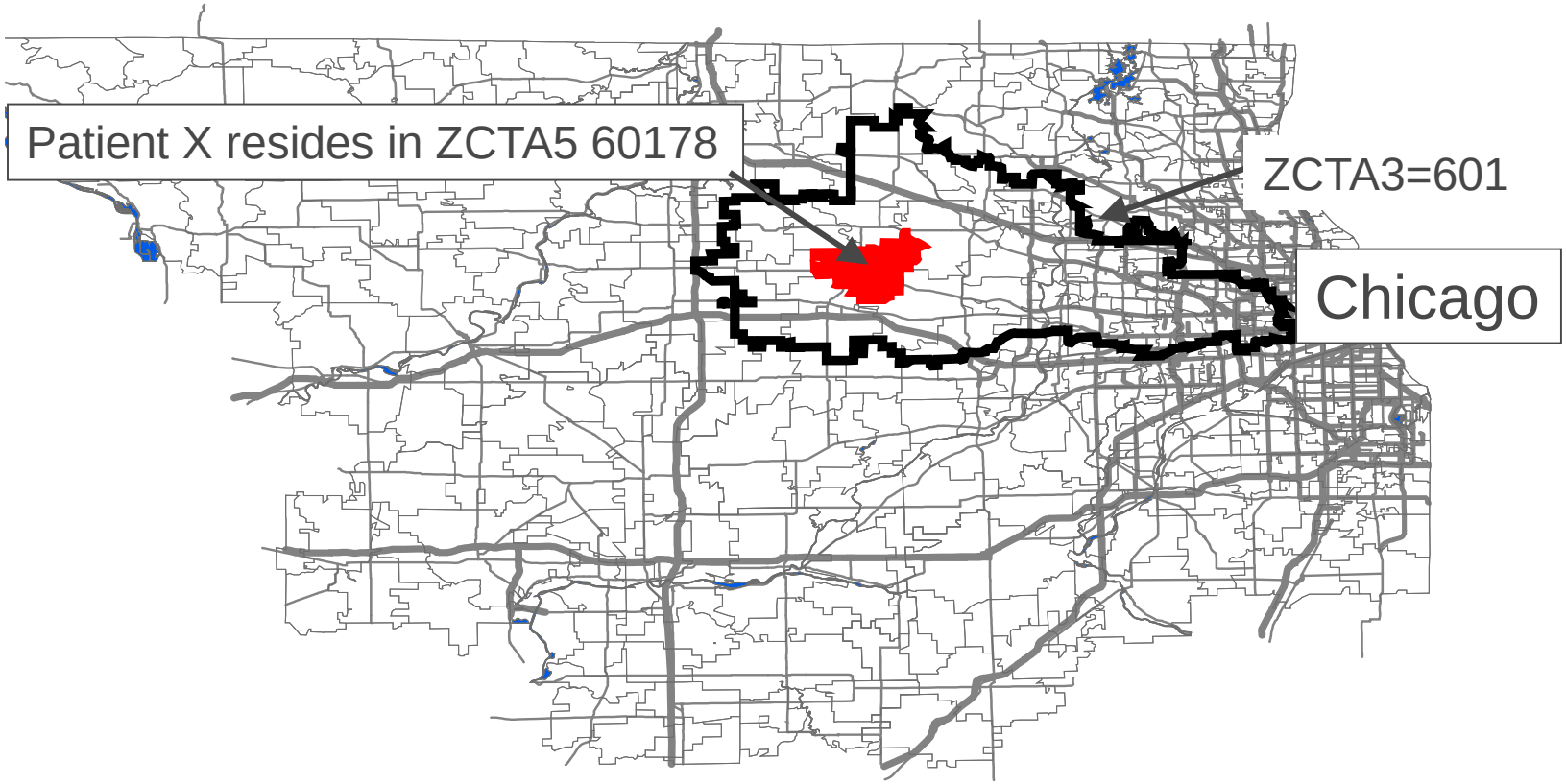
There are 7 CBSAs in Ohio which Cross into 4 Border States



Challenge: "Geoproxy" Attacks

- Challenge: Data intruders can use Geographic Information Systems (GIS) to determine the likely locations of patients from the locations of their healthcare providers
  - Retail Pharmacy Locations
  - Physician or Healthcare Provider Locations
  - Hospital Locations
- ***Geoproxy attacks have become much easier to conduct using newly available tools (***e.g., Web 2.0 mapping "Mash-up" technology) ***on the internet.***

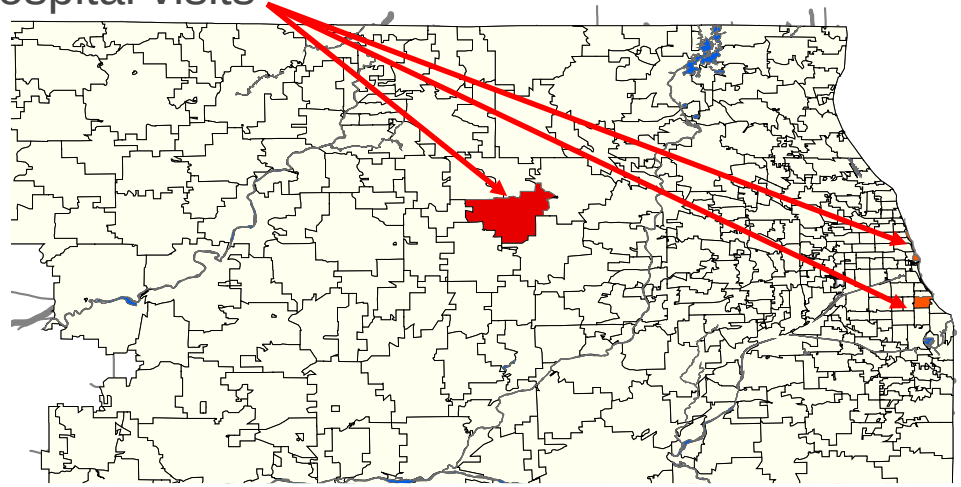
Challenge: Geoproxy Attacks



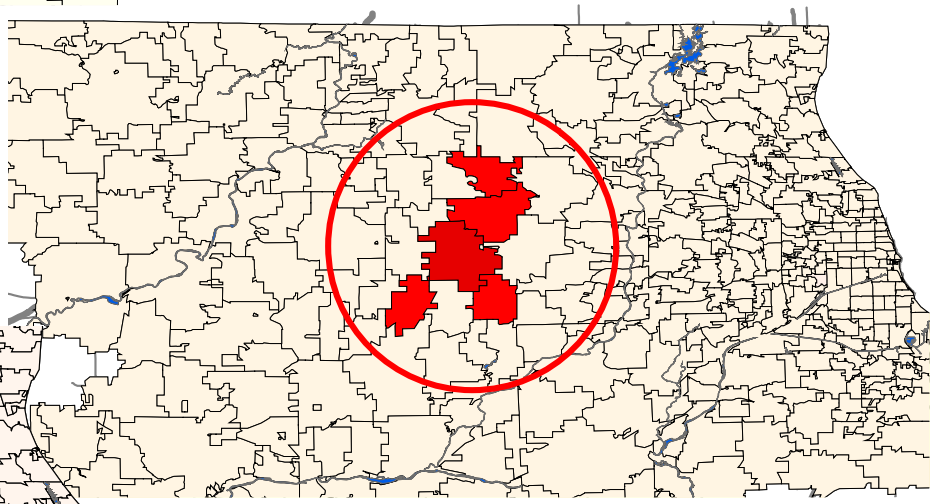
Example: Patient location as revealed within data set, but further narrowed to probable “hotspots” by using healthcare provider location data

Challenge:
Geoproxy Attacks

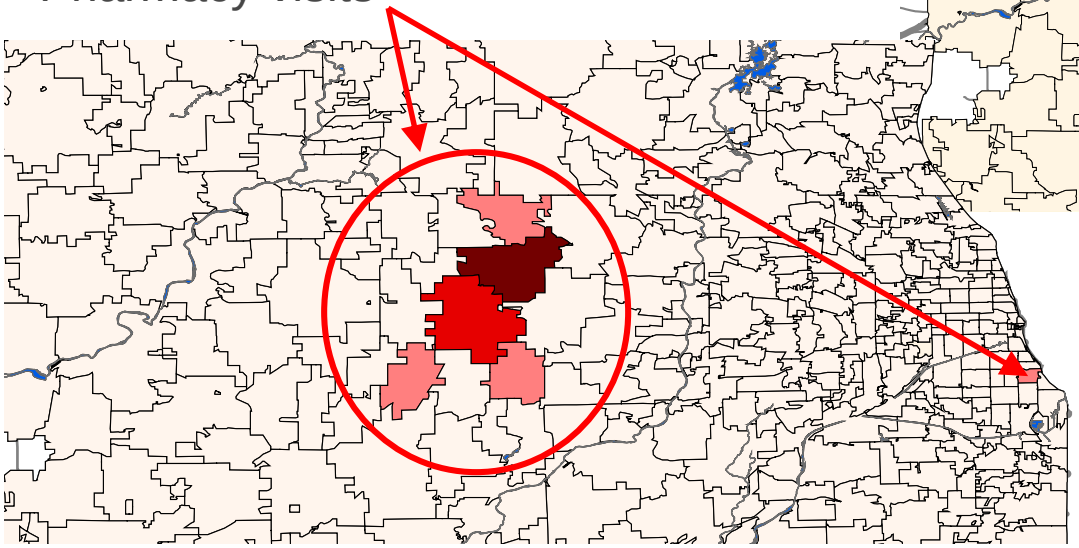
Hospital visits



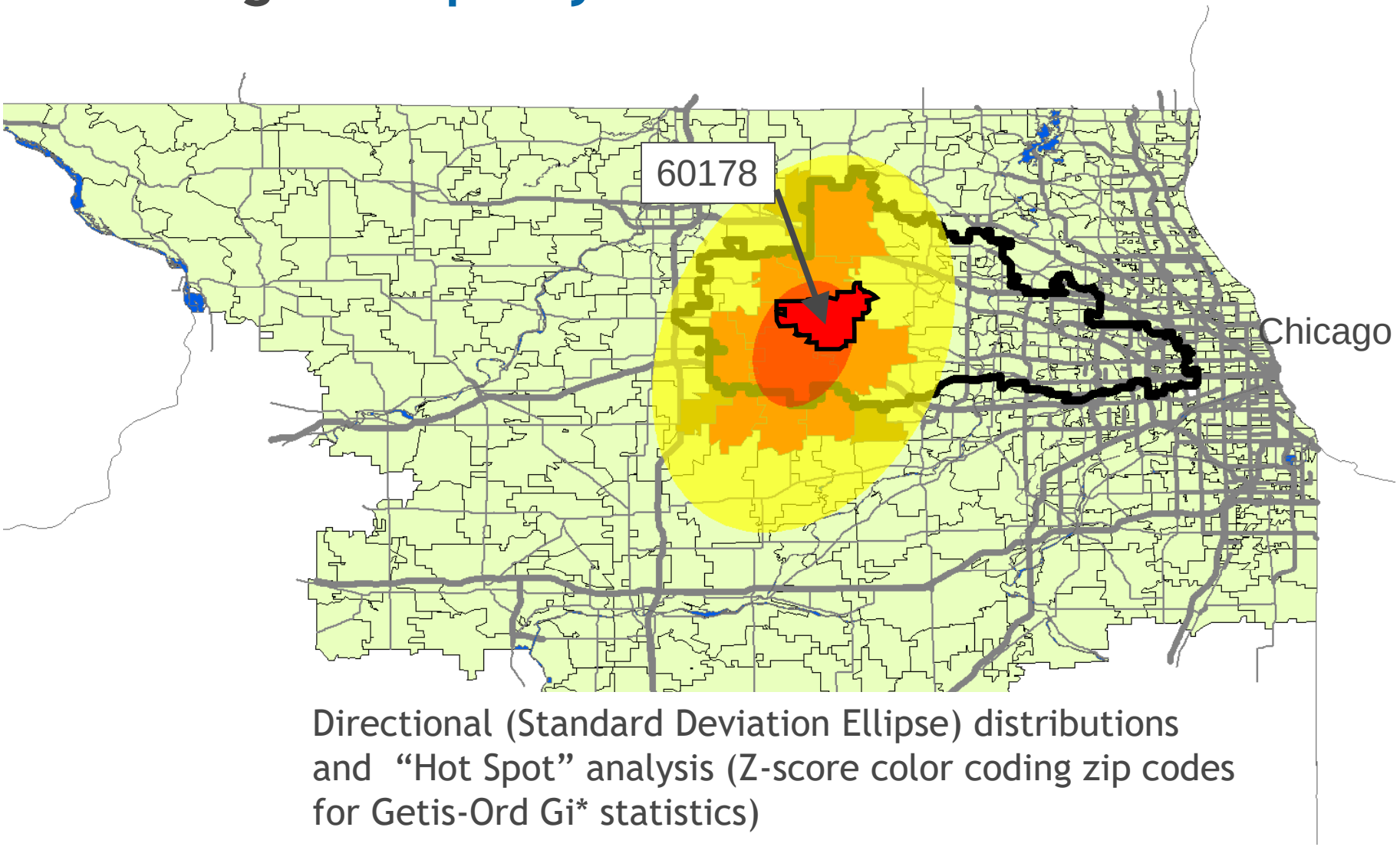
Outpatient/Office visits



Pharmacy visits



Challenge: Geoproxy Attacks

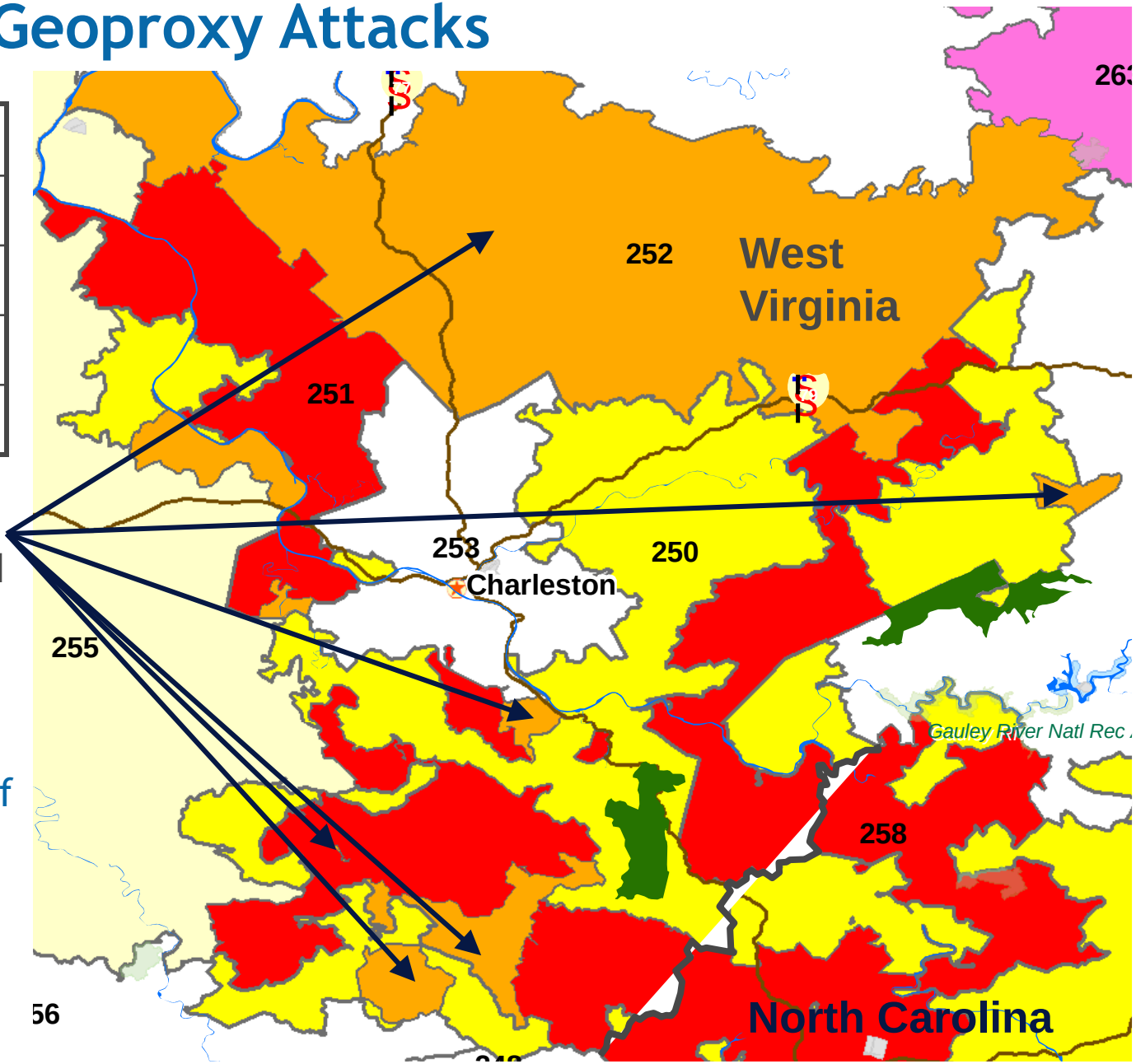


Directional (Standard Deviation Ellipse) distributions and "Hot Spot" analysis (Z-score color coding zip codes for Getis-Ord Gi* statistics)

Challenge: Geoproxy Attacks

ZCTA3	Population
250	68,890
251	80,077
252	55,954
253	121,609

ZCTA3 252 is highly dispersed

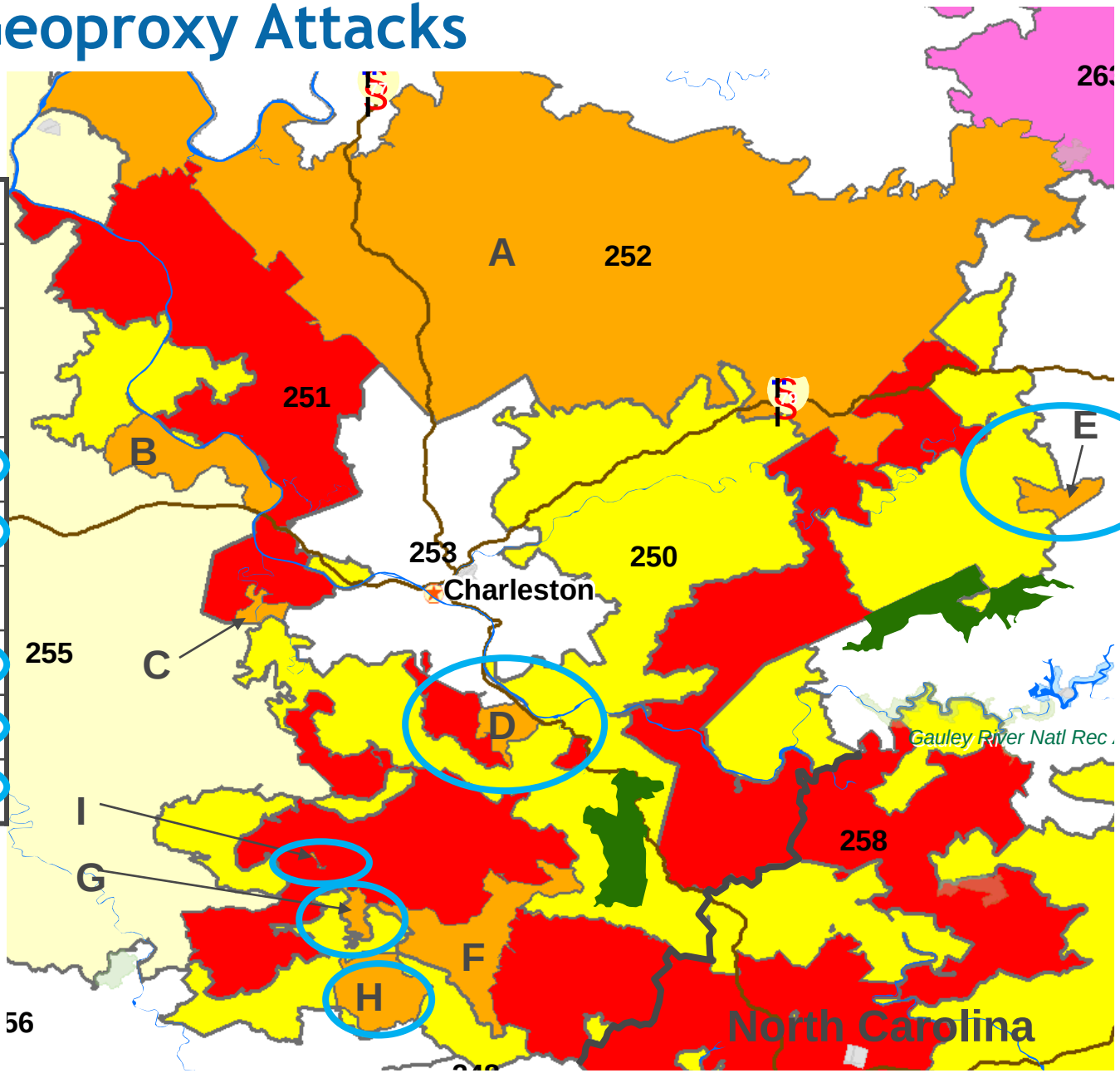


The complexity of 3-digit Zip Code Geography amplifies the threat of Geoproxy attacks

Challenge: Geoproxy Attacks

ZCTA3 252

Area	Population
A	46,076
B	4,754
C	1,254
D	768
E	242
F	1,581
G	649
H	447
I	183



Successful Solutions:

Balancing Disclosure Risk and Statistical Accuracy

- When appropriately implemented, statistical de-identification seeks to protect and balance two vitally important societal interests:
  - 1) Protection of the privacy of individuals in healthcare data sets, (Disclosure or Identification Risk), and
  - 2) Preserving the utility and accuracy of statistical analyses performed with de-identified data (Loss of Information).
- Limiting disclosure inevitably reduces the quality of statistical information to some degree, but the appropriate disclosure control methods result in small information losses while substantially reducing identifiability.

Suggested Conditions for De-identified Data Use

Recipients of De-identified Data should be required to:

1) Not re-identify, or attempt to re-identify, or allow to be re-identified, any patients or individuals who are the subject of Protected Health Information within the data, or their relatives, family or household members.
2) Not link any other data elements to the data without obtaining a determination that the data remains de-identified.
3) Implement and maintain appropriate data security and privacy policies, procedures and associated physical, technical and administrative safeguards to assure that it is accessed only by authorized personnel and will remain de-identified.
4) Assure that all personnel or parties with access to the data agree to abide by all of the foregoing conditions.

Reserve Slides for Questions

References for Re-identification Attack Summary Table

1. Sweeney, L. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.
2. Barth-Jones, DC., The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now (July 2012). http://ssrn.com/abstract=2076397
3. Michael Barbaro, Tom Zeller Jr. A Face Is Exposed for AOL Searcher No. 4417749. New York Times August 6, 2006. www.nytimes.com/2006/08/09/technology/09aol.html
4. Narayanan, A., Shmatikov, V. Robust De-anonymization of Large Sparse Datasets. Proceeding SP '08 Proceedings of the 2008 IEEE Symposium on Security and Privacy p. 111-125.
5. Kwok, P.K.; Lafky,D. Harder Than You Think: A Case Study of Re-Identification Risk of HIPAA Compliant Records. Joint Statistical Meetings. Section on Government Statistics. Miami, FL Aug 2, 2011. p. 3826-3833.
6. El Emam K, et al. De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset. J Med Internet Res 2012;14(1):e33
7. Valentino-DeVries, J. May the Best Algorithm Win… With $3 Million Prize, Health Insurer Raises Stakes on the Data-Crunching Circuit. Wall Street Journal. March 16, 2011. March 17, 2011 http://www.wsj.com/article_email/SB10001424052748704662604576202392747278936-lMyQjAxMTAxMDEwNTExNDUyWj.html
8. Narayanan, A. An Adversarial Analysis of the Reidentifiability of the Heritage Health Prize Dataset. May 27, 2011 http://randomwalker.info/publications/heritage-health-re-identifiability.pdf
9. Narayanan, A. Felten, E.W. No silver bullet: De-identification still doesn't work. July 9, 2014 http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf
10. Melissa Gymrek, Amy L. McGuire, David Golan, Eran Halperin, Yaniv Erlich. Identifying Personal Genomes by Surname Inference. Science 18 Jan 2013: 321-324.
11. Barth-Jones, D. Public Policy Considerations for Recent Re-Identification Demonstration Attacks on Genomic Data Sets: Part 1. Harvard Law, Petrie-Flom Center: Online Symposium on the Law, Ethics & Science of Re-identification Demonstrations. http://blogs.harvard.edu/billofhealth/2013/05/29/public-policy-considerations-for-recent-re-identification-demonstration-attacks-on-genomic-data-sets-part-1-re-identification-symposium/
12. Sweeney, L., Abu, A, Winn, J. Identifying Participants in the Personal Genome Project by Name (April 29, 2013). http://ssrn.com/abstract=2257732

References for Re-identification Attack Summary Table

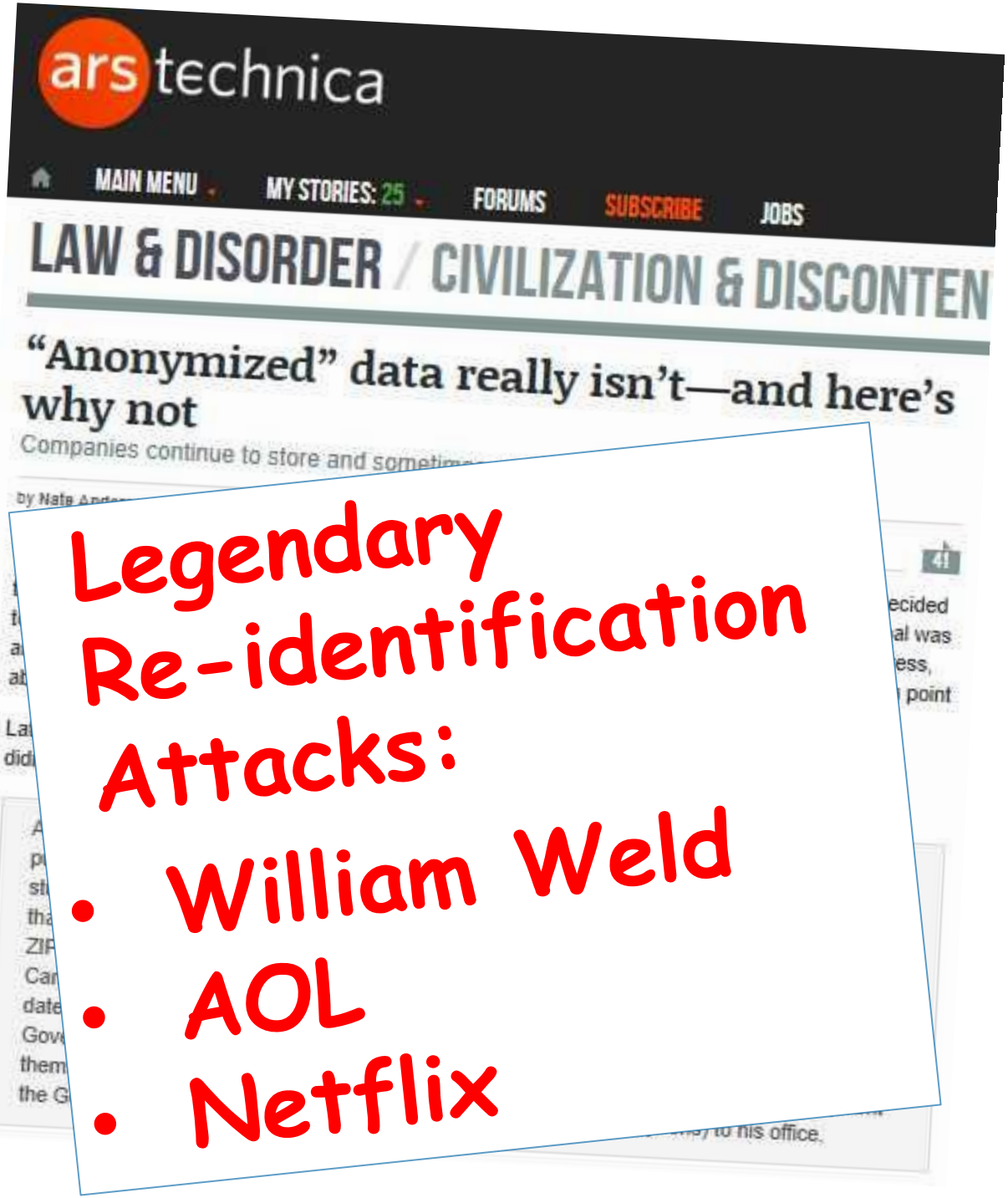
13. Jane Yakowitz. Reporting Fail: The Reidentification of Personal Genome Project Participants May 1, 2013. https://blogs.harvard.edu/infolaw/2013/05/01/reporting-fail-the-reidentification-of-personal-genome-project-participants/
14. Barth-Jones, D. Press and Reporting Considerations for Recent Re-Identification Demonstration Attacks: Part 2. Harvard Law, Petrie-Flom Center: Online Symposium on the Law, Ethics & Science of Re-identification Demonstrations. http://blogs.harvard.edu/billofhealth/2013/10/01/press-and-reporting-considerations-for-recent-re-identification-demonstration-attacks-part-2-re-identification-symposium/
15. Sweeney, L. Matching Known Patients to Health Records in Washington State Data (June 5, 2013). http://ssrn.com/abstract=2289850
16. Robertson, J. States' Hospital Data for Sale Puts Privacy in Jeopardy. Bloomberg News June 5, 2013. https://www.bloomberg.com/news/articles/2013-06-05/states-hospital-data-for-sale-puts-privacy-in-jeopardy
17. Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, Vincent D. Blondel. Unique in the Crowd: The privacy bounds of human mobility. Scientific Reports 3, Article number: 1376 (2013) http://www.nature.com/articles/srep01376
18. Anthony Tockar. Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset. September 15, 2014. https://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/
19. Barth-Jones, D. The Antidote for "Anecdata": A Little Science Can Separate Data Privacy Facts from Folklore. https://blogs.harvard.edu/infolaw/2014/11/21/the-antidote-for-anecdata-a-little-science-can-separate-data-privacy-facts-from-folklore/
20. de Montjoye, et al. . Unique in the shopping mall: On the reidentifiability of credit card metadata. Science. 30 Jan 2015: Vol. 347, Issue 6221, pp. 536-539.
21. Barth-Jones D, El Emam K, Bambauer J, Cavoukian A, Malin B. Assessing data intrusion threats. Science. 2015 Apr 10; 348(6231):194-5.
22. de Montjoye, et al. Assessing data intrusion threats—Response Science. 10 Apr 2015: Vol. 348, Issue 6231, pp. 195
23. Jane Yakowitz Bambauer. Is De-Identification Dead Again? April 28, 2015. https://blogs.harvard.edu/infolaw/2015/04/28/is-de-identification-dead-again/
24. David Sánchez, Sergio Martínez, Josep Domingo-Ferrer. Technical Comments: Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata". Science. 18 Mar 2016: Vol. 351, Issue 6279, pp. 1274.
25. Sánchez, et al. Supplementary Materials for "How to Avoid Reidentification with Proper Anonymization"- Comment on "Unique in the shopping mall: on the reidentifiability of credit card metadata". http://arxiv.org/abs/1511.05957
26. de Montjoye, et al. Response to Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata" Science 18 Mar 2016: Vol. 351, Issue 6279, pp. 1274

References for Re-identification Attack Summary Table

27. Nate Anderson. “Anonymized” data really isn’t—and here’s why not. Sep 8, 2009 http://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/
28. Sorrell v. IMS Health: Brief of Amici Curiae Electronic Privacy Information Center. March 1, 2011. https://epic.org/amicus/sorrell/EPIC_amicus_Sorrell_final.pdf
29. Ruth Williams. Anonymity Under Threat: Scientists uncover the identities of anonymous DNA donors using freely available web searches. The Scientist. January 17, 2013. http://www.the-scientist.com/?articles.view/articleNo/34006/title/Anonymity-Under-Threat/
30. Kevin Fogarty. DNA hack could make medical privacy impossible. CSO. March 11, 2013. http://www.csoonline.com/article/2133054/identity-access/dna-hack-could-make-medical-privacy-impossible.html
31. Adam Tanner. Harvard Professor Re-Identifies Anonymous Volunteers in DNA Study. Forbes. Apr 25, 2013. http://www.forbes.com/sites/adamtanner/2013/04/25/harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/
32. Adam Tanner. The Promise & Perils of Sharing DNA. Undark Magazine. September 13, 2016. http://undark.org/article/dna-ancestry-sharing-privacy-23andme/
33. Sweeney L. Only You, Your Doctor, and Many Others May Know. Technology Science. 2015092903. September 29, 2015. http://techscience.org/a/2015092903
34. David Sirota. How Big Brother Watches You With Metadata. San Francisco Gate. October 9, 2014. http://www.sfgate.com/opinion/article/How-Big-Brother-watches-you-with-metadata-5812775.php
35. Natasha Singer. With a Few Bits of Data, Researchers Identify ‘Anonymous’ People. New York Times. Bits Blog. January 29, 2015. http://bits.blogs.nytimes.com/2015/01/29/with-a-few-bits-of-data-researchers-identify-anonymous-people/

Additional Re-identification Attack Review References

1. Khaled El Emam, Jonker, E.; Arbuckle, L.; Malin, B. A systematic review of re-identification attacks on health data. PLoS One 2011; Vol 6(12):e28071.
2. Jane Henriksen-Bulmer, Sheridan Jeary. Re-identification attacks - A systematic literature review. International Journal of Information Management, 36 (2016) 1184–1192.



Unfortunately, de-identification public policy has often been driven by largely anecdotal and limited evidence, and re-identification demonstration attacks targeted to particularly vulnerable individuals, which fail to provide reliable evidence about real world re-identification risks

Re-identification Demonstration Attack Summary

Re-identification Attacks	Quasi-Identifers (w/ HIPAA Safe Harbor exclusion data in Red)	Vulnerable Subgroup Targeted?	Used Stat. Sampling	Individuals w/ Alleged/Verified Re-identification	At-Risk Sample Size	Notable Headlines & Quotes	Attack Against HIPAA Compliant (or SDL Protected) Data?	Demonstrated Re-identification Risk
Governor Weld [1,2]	Zip5, Gender, DoB	Yes	No	n=1	99,500	*"Anonymized" Data Really Isn't* [27]	No	0.00001
AOL [3]	Free Text from Search Queries w/ Name, Location, etc	Yes	No	n=1	657,000	*A Face is Exposed* [3]	No	0.0000015
Netflix [4]	Movie Ratings & Dates	Yes	No	n=2	500,000	*"...successfully identified 99% of people in Netflix database"* [28]	No	0.000004
ONC Safe Harbor [5]	Zip3, YoB, Gender, Marital Status, Hispanic Ethnicity	No	N/A	n=2	15,000	[Press Did Not Cover This Study]	Yes	0.00013
Heritage Health Prize [6,7,8,9]	Age, Sex, Days in Hospital, Physician Specialty, Place of Service, CPT Code, Days Since First Claim, ICD-9 Diagnosis	Yes	No	n=0	113,000	*To best of my judgment, reidentification is within realm of possibility* [8] *El Emam estimated < 1% of Pts could be re-identified. Narayanan estimated > 12% of Pts were identifiable.* [29]	Yes	0.0
Y-Chromosome STR Surname Inference [10,11] - Simulation Study Part	Y-STR DNA Sequences* Age in Years & State	No	N/A, Simulation	Not Attempted: Simulated Results	~150 Million US Males	*"nice example of how simple it is to re-identify de-identified samples"* [30]	*No? (Safe Harbor vs. Expert Determination)	.12 (For Males Only), after accounting for 30% False Positive Rate
- CEU Attack Part	Age, Utah State, Genealogy Pedigrees & Mormon Ancestry	Yes, Highly Targeted	No	n=5 w/ Y-STR Alone, (but w/ Geneology Amplification n=50)	?	*DNA Hack Could Make Medical Privacy Impossible* [31]	*Safe Harbor Excludes: *Any unique identifying #, characteristic or code*	Not Clearly Calculable for CEU Attack
Personal Genome Project [12,13,14]	Zip5, Gender, DoB	No	N/A	n=161	579	*"...re-identified names of > 40% anonymous participants"* [32] *re-identified 84 to 97% of sample of PGP volunteers* [33]	No	0.28 (w/ Embedded Names Excluded)
Washington St. Hospital Discharge [15,16]	Hospital Data w/ Diagnoses, Zip5, Month/Yr of Discharge	Yes	No	n=40 (8 verified) from 81 News Reports	648,384	*"...how new stories about hospital visits in Washington State leads to identifying matching health record 43% of the time"* [34]	No	0.000062
Cell Phone "Unicity" [17]	High Resolution Time (Hours) and Cell Tower Location	No	N/A	Not Attempted	1.5 Million	*"four spatio-temporal points enough to uniquely identify 95%"* [17]	No	0.0
NYC Taxi [18,19]	High Resolution Time (Minutes) and GPS Locations	Yes	No	n=11	173 Million Rides	*How Big Brother Watches You With Metadata* [35]	No	0.0000001
Credit Card "Unicity" [20,21,22,23,24,25,26]	High Resolution Time (Days), Location and Approx. Price	No	N/A	Not Attempted	1.1 Million	*With a Few Bits of Data, Researchers Identify 'Anonymous' People* [36]	No	0.0

- *Publicized attacks are on data without HIPAA/SDL de-identification protection.*
- *Many attacks targeted especially vulnerable subgroups and did not use sampling to assure representative results.*
- *Press reporting often portrays re-identification as broadly achievable, when there isn't any reliable evidence supporting this portrayal.*

Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

The Narayan/Shmatikov "Netflix" algorithm is an intelligently designed advance for re-identification methods. However, scrutiny is warranted for the experimental design and associated information assumptions when considering how robust the algorithm really is and other conditions in which it might work well.

Re-identification Demonstration Attack Summary

- For Ohm's famous "Broken Promises" attacks (Weld, AOL, Netflix) a total of n=4 people were re-identified out of 1.25 million.
- For attacks against HIPAA de-identified data (ONC, Heritage*), a total of n=2 people were re-identified out of 128 thousand.
  - ONC Attack Quasi-identifers: Zip3, YoB, Gender, Marital Status, Hispanic Ethnicity
  - Heritage Attack Quasi-identifiers*: Age, Sex, Days in Hospital, Physician Specialty, Place of Service, CPT Procedure Codes, Days Since First Claim, ICD-9 Diagnoses (*not complete list of data available for adversary attack)
  - Both were "adversarial" attacks.
- For all attacks listed, a total of n=268 were re-identified out of 327 million opportunities.

Let's get some perspective on this…

Obviously, This slide is **BLACK**

So clearly, De-identification Doesn't Work.

Re-identification Demonstration Attack Summary

What can we conclude from the empirical evidence provided by these 11 highly influential re-identification attacks?

- The proportion of *demonstrated* re-identifications is extremely small.
- Which ***does not imply data re-identification risks are necessarily very small*** (*especially if the data has not been subject to Statistical Disclosure Limitation methods*).
- But with only 268 re-identifications made out of 327 million opportunities, Ohm's "Broken Promises" assertion that "*scientists have demonstrated they can often re-identify with astonishing ease*" seems rather dubious.
- It also seems clear that the state of "re-identification science", and the "evidence", it has provided needs to be dramatically improved in order to better support good public policy regarding data de-identification.

Bill of Health

Examining the intersection of law and health care, biotech & bioethics
A blog by the Petrie-Flom Center and friends

Online Symposium on the Law, Ethics & Science of Re-identification Demonstrations

- http://blogs.law.harvard.edu/billofhealth/2013/05/29/public-policy-considerations-for-recent-re-identification-demonstration-attacks-on-genomic-data-sets-part-1-re-identification-symposium/
- https://blogs.law.harvard.edu/billofhealth/2013/10/01/press-and-reporting-considerations-for-recent-re-identification-demonstration-attacks-part-2-re-identification-symposium/
- http://blogs.law.harvard.edu/billofhealth/2013/10/02/ethical-concerns-conduct-and-public-policy-for-re-identification-and-de-identification-practice-part-3-re-identification-symposium/

HIPAA §164.514(b)(2)(i) -18 "Safe Harbor" Exclusions

All of the following must be **removed in order** for the information **to be** considered **de-identified.**

(2)(i) The **following identifiers of the individual or of relatives, employers, or household members** of the individual, are removed:

(A) Names;

(B) 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 current publicly available data from the Bureau of the Census: (*1*) The geographic unit formed by combining all zip codes with the same three initial digits contains **more than 20,000 people**; and (*2*) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.

(C) **All elements of dates (except year)** for dates directly related 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;

(D) Telephone numbers;

(E) Fax numbers;

(F) Electronic mail addresses;

(G) Social security numbers;

(H) **Medical record numbers**;

(I) **Health plan beneficiary numbers**;

(J) Account numbers;

(K) Certificate/license numbers;

(L) Vehicle identifiers and serial numbers, including license plate numbers;

(M) **Device identifiers and serial numbers**;

(N) Web Universal Resource Locators (URLs);

(O) Internet Protocol (IP) address numbers;

(P) Biometric identifiers, including finger and voice prints;

(Q) Full face photographic images and any comparable images; and

(R) **Any other unique identifying number, characteristic, or code** except as permitted in §164.514(c)

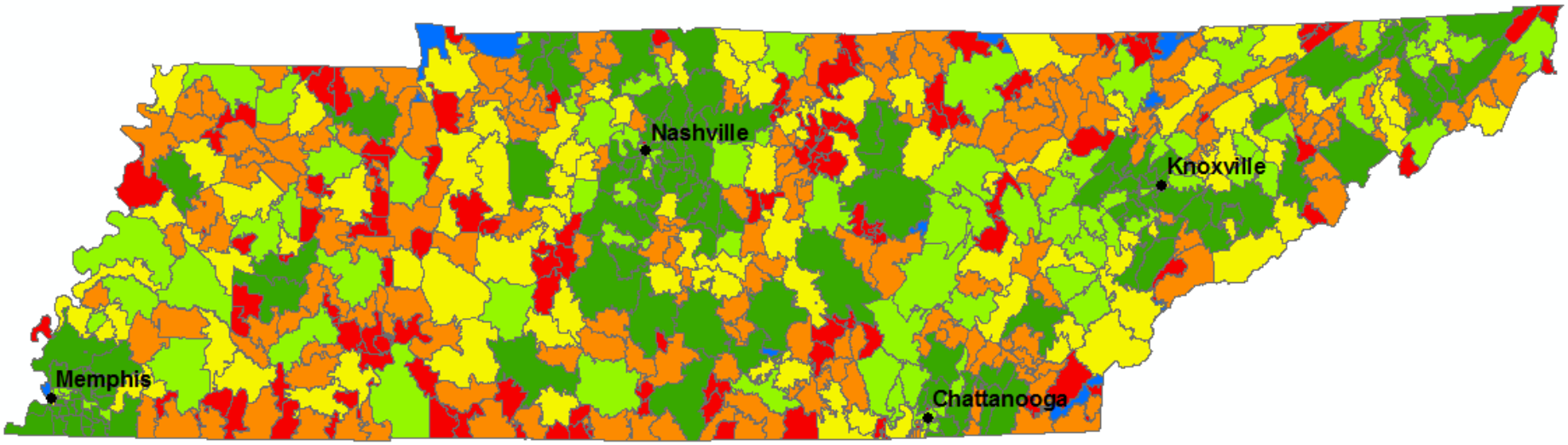
HIPAA §164.514(b)(1) "Expert Determination"

Health Information is not individually identifiable if:

A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:

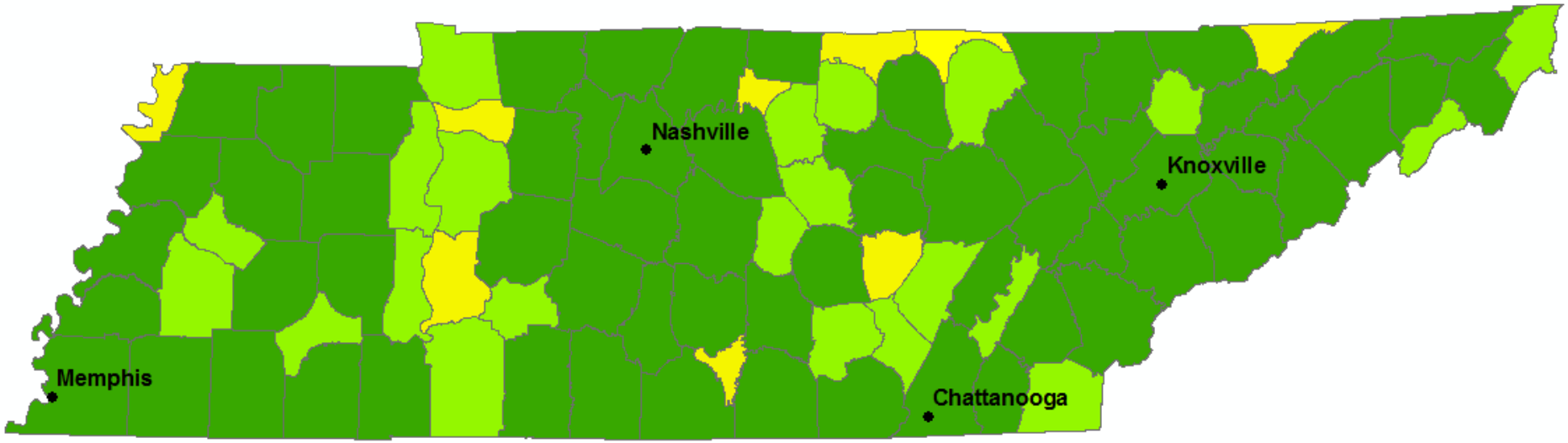
(i) Applying such principles and methods, determines that the *risk is very small* that *the information could be used*, alone or *in combination with other reasonably available information, by an anticipated recipient to identify an individual* who is a subject of the information; and (ii) Documents the methods and results of the analysis that justify such determination;

Tennessee – ZCTA5 Populations



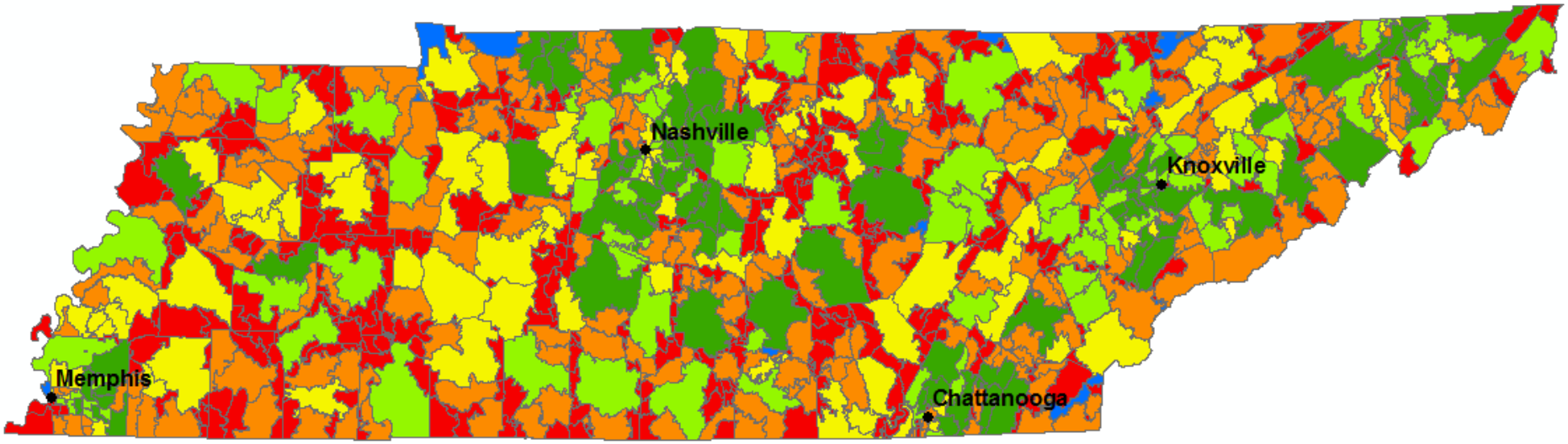
Population
- < 1500
- 1,501 - 5,000
- 5,001 - 10,000
- 10,001 - 20,000
- 20,001 +

Tennessee – County Populations



Population
- < 1500
- 1,501 - 5,000
- 5,001 - 10,000
- 10,001 - 20,000
- 20,001 +

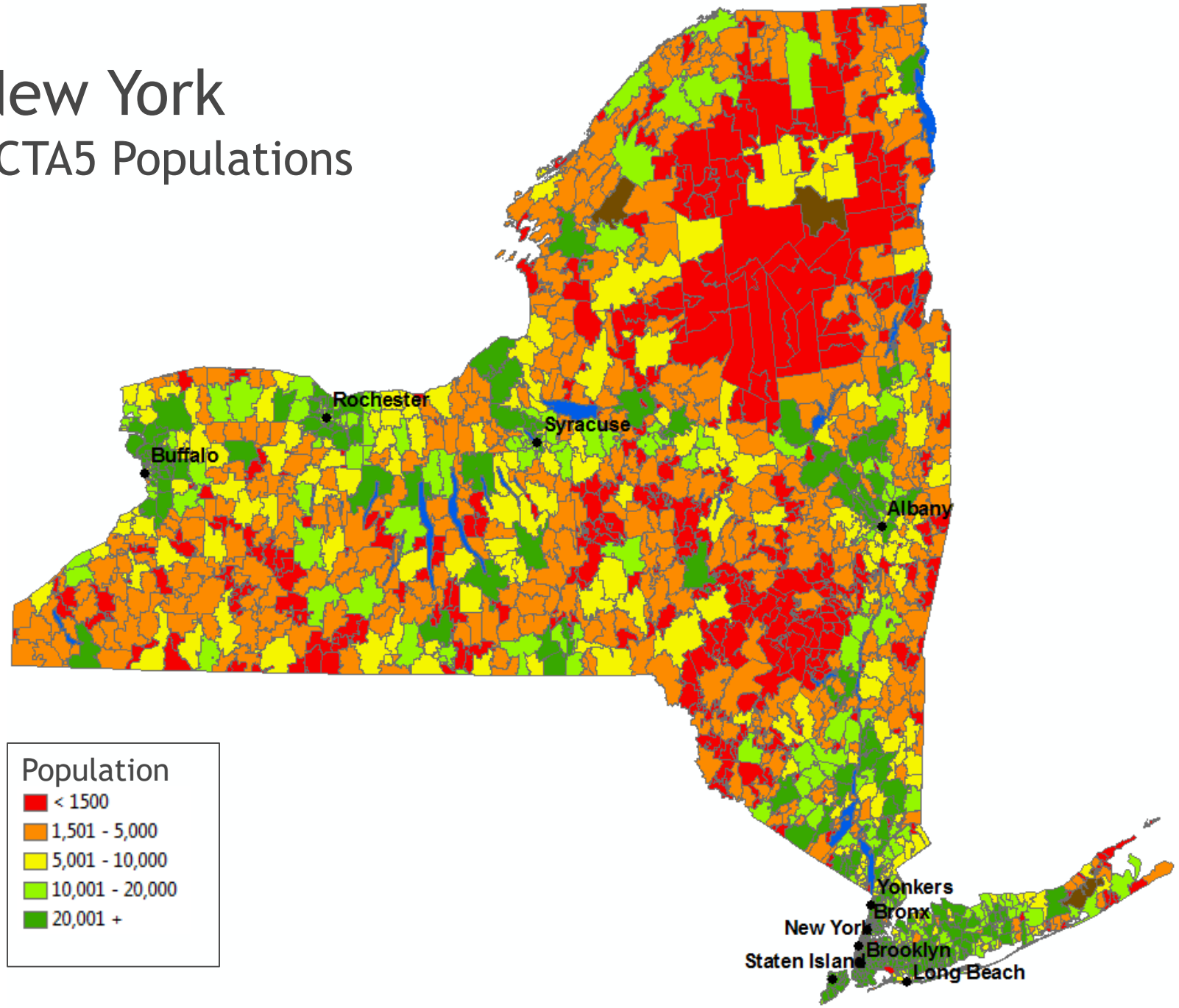
Tennessee – ZCTA5 X County Populations



Population
- < 1500
- 1,501 - 5,000
- 5,001 - 10,000
- 10,001 - 20,000
- 20,001 +

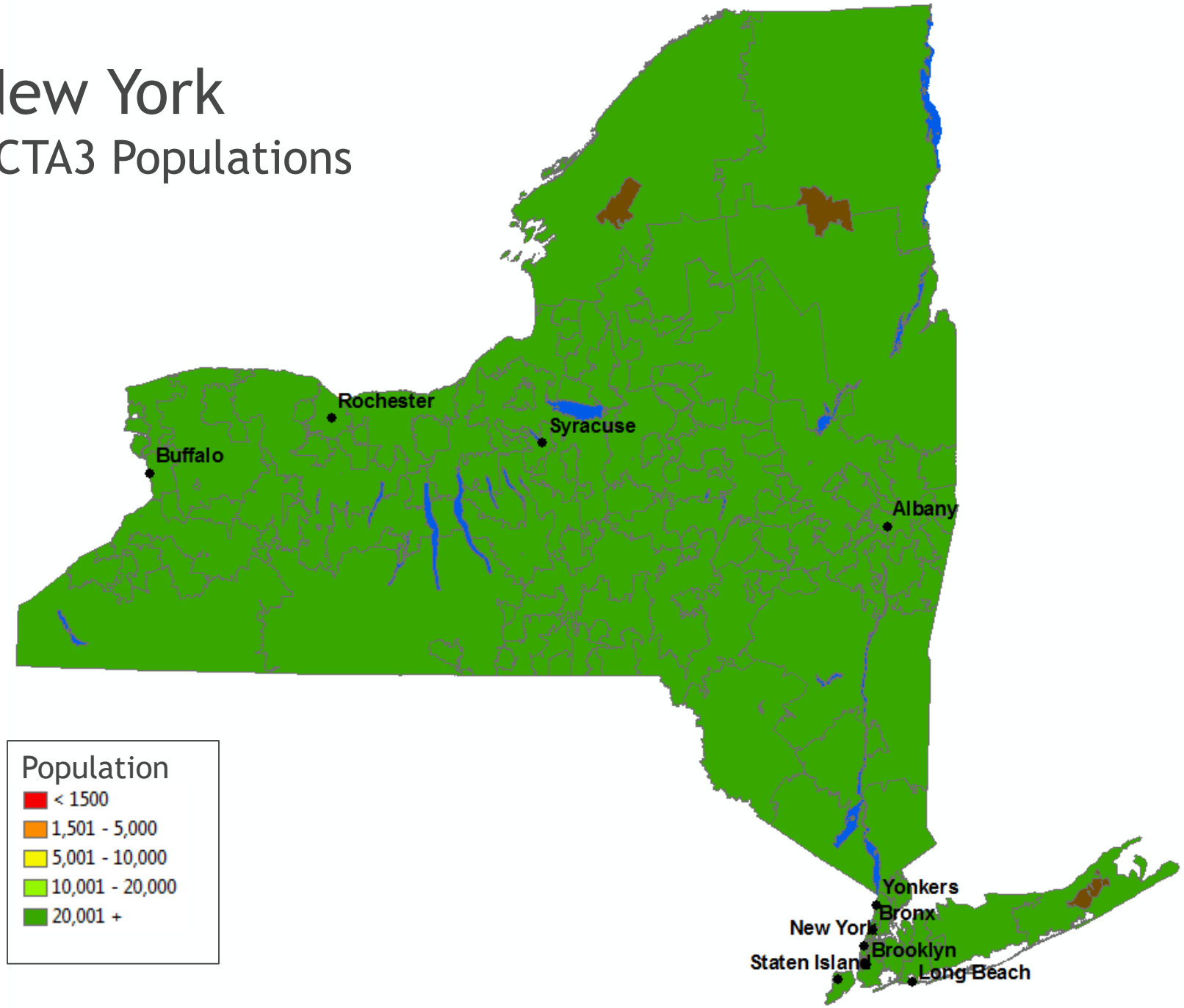
New York

ZCTA5 Populations



New York

ZCTA3 Populations



New York

ZCTA5 Collapse Populations