The Network for Public Health Law

Data Sharing Strategies to Advance Health Equity

May 16, 2019 1:00 p.m. EST

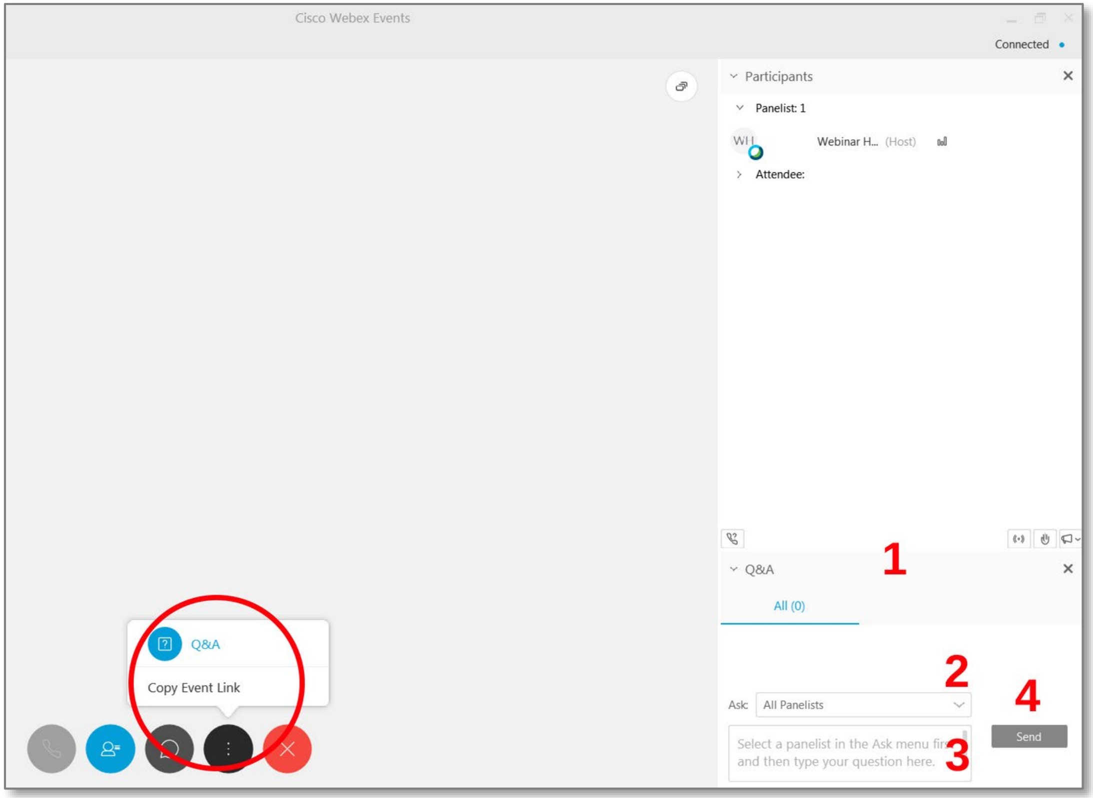
Co-sponsored by:

The Network for Public Health Law

ASLME
AMERICAN SOCIETY OF LAW, MEDICINE & ETHICS

How to Use WebEx Q & A

1. Open the Q&A panel
2. Select “All Panelists”
3. Type your question
4. Click “Send”



The Network for Public Health Law

Presenter

Jennifer Bernstein, Deputy Director, the Network for Public Health Law – Mid-States Region Office

- J.D., M.P.H., University of Iowa
- Research interests/areas of expertise:
  - Public Health Data
  - Adults at Risk
  - Maternal and Child Health
  - Statutory and Regulatory Public Health

The Network for Public Health Law

Presenter

Sallie Milam, Deputy Director, the Network for Public Health Law – Mid-States Region Office

- J.D., University of Richmond School of Law
- C.I.P.P./U.S./G., International Association of Privacy Professionals
- Research interests/areas of expertise:
  - HIPAA Privacy Laws
  - Health Information and Data Sharing
  - De-identification

The Network for Public Health Law

Presenter

Daniel Barth-Jones, Assistant Professor of Clinical Epidemiology, Columbia University Mailman School of Public Health

- Ph.D., M.P.H., University of Michigan
- Research interests/areas of expertise:
  - Epidemiology of HIV and Sexually Transmitted Diseases
  - Health Economic Evaluations of Public Health Policies for Vaccination and Preventative Intervention
  - HIPAA Privacy Regulations

Data sharing for public health

- Faster response to emerging public health threats
- Easier evaluation to design interventions and determine best practices
- Greater collaboration between sectors and across jurisdictions (state, tribal, international)
- Coordination of care
- Identify and address upstream social determinants of health

The Network for Public Health Law

Ethical principles for data sharing

» **Common, foundational considerations**

- Autonomy
- Privacy
- Individual rights

» **Must balance against WHO principles**

- Justice
- Beneficence
- Common good
- Equity
- Reciprocity

Public Health 3.0

- » Requires access to timely, reliable, granular data (i.e. sub-county) and actionable data
- » Depends on data from many and diverse sources – including sources and types of data relevant to social determinants
- » Should have data that are accessible to communities throughout the country that are shared, linked, and synthesized while protecting data security and individual privacy
- » Needs clear metrics to assess impact and document success

The Network for Public Health Law

Public health 3.0 envisions

» Public health leaders as Chief Health Strategist for their communities

- Public health special legal status - broad authority to collect data to prevent and control disease, protect public health, and promote wellness

» Multiple sector public and private partners

» Health and non-health sectors

» Partners that explicitly address ”upstream” social determinants of health

Data Sharing: Using De-Identification to Advance Health Equity.

What does the law require?

Sallie Milam, JD, CIPP/US/G
Deputy Director, Network for Public Health Law
Mid-States Region

Network for Public Health Law De-Identification Toolkit

Network for Public Health Law →Topics & Resources → Health Information and Data Sharing → De-Identification Toolkit

https://www.networkforphl.org/resources/topics__resources/health_information_and_data_sharing/de-identification_toolkit/

-Highlights traditional, non-traditional and emerging data sources

-Provides tools and resources to better understand de-identification

-Provides tools and resources for sharing de-identified data legally and safely

What is de-identification?

- It is an important tool to make data available to communities.
- Law may offer an approach to change those aspects of the data set that identify an individual or lead to the identification of an individual.
- Law may also be silent as to method and require or permit the data to be disclosed, but remain confidential.
- De-identification requires the data steward to remove data elements that directly identify an individual, such as name and Social Security number, as well as data elements that indirectly identify an individual, such as date of birth and address.

The Network for Public Health Law

When public health receives a request to share data, where do you start?

The Network for Public Health Law

Collect Factual Information

Accountability?

ons?

Checklist of Factual Information Needed for Public Health Agencies to Address Proposed Data Collection, Access and Sharing

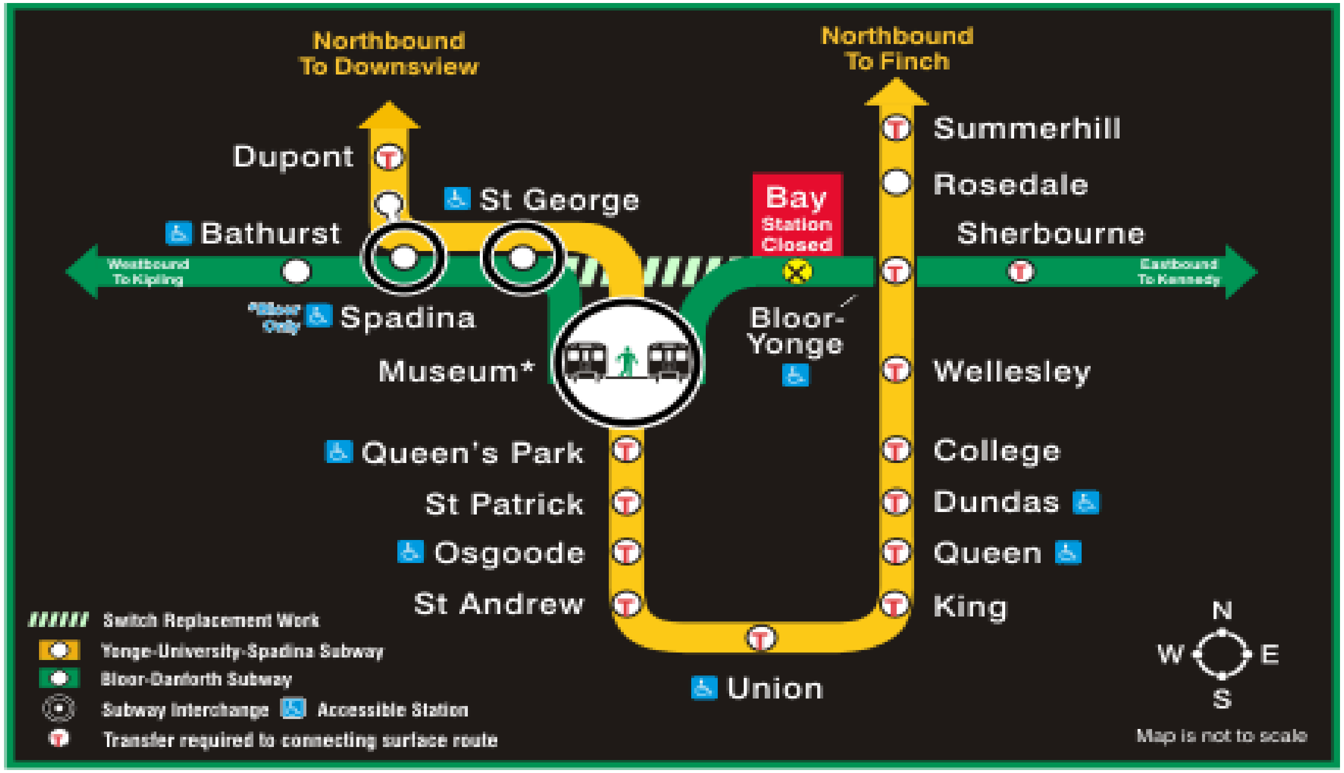
From whom

How? Where?

Vith whom?

Evaluate review criteria

Checklist of Review Criteria for Public Health Agencies to Evaluate Proposed Collection, Access and Sharing of De-identified Data



De-Identification: As Described by Federal Statutes

- HIPAA
- FERPA
- 42 Part 2
- And many more

- *See also*, another new resource: Federal Privacy Laws which include of de-identification provisions. This resource may be found within the Health Information and Data Sharing topic on the Network's website.

De-Identification Table: Guidance from the Courts

- **Courts have addressed the adequacy of specific de-identification methods**
- **Courts examine whether the information, in combination with other information and factors, would identify or tend to reveal the identity of the data subject**
- **Courts balance competing interests of public access to information against the risk of invasion of privacy**
- **Courts interpret relevant law in light of facts:**
  - Specificity of the PHI, such as date of diagnosis, age at diagnosis, sex, race, religion, family medical history, diagnostic information, treatment and vital statistics
  - Denominator or number of cases*, e.g*,, small number results in a greater risk
  - Other readily available information, including community knowledge
  - Whether the identities of the data subjects are already known

De-Identification under HIPAA

» **Two methods:**

(1) **Expert Determination**

(2) **Safe Harbor**

De-Identification Toolkit provides Quick References for both methods

The Network for Public Health Law

De-Identification – Expert Determination

» Person with appropriate knowledge and experience

» Applies statistical or scientific principles

» Determines very small risk that anticipated recipient could identify individual

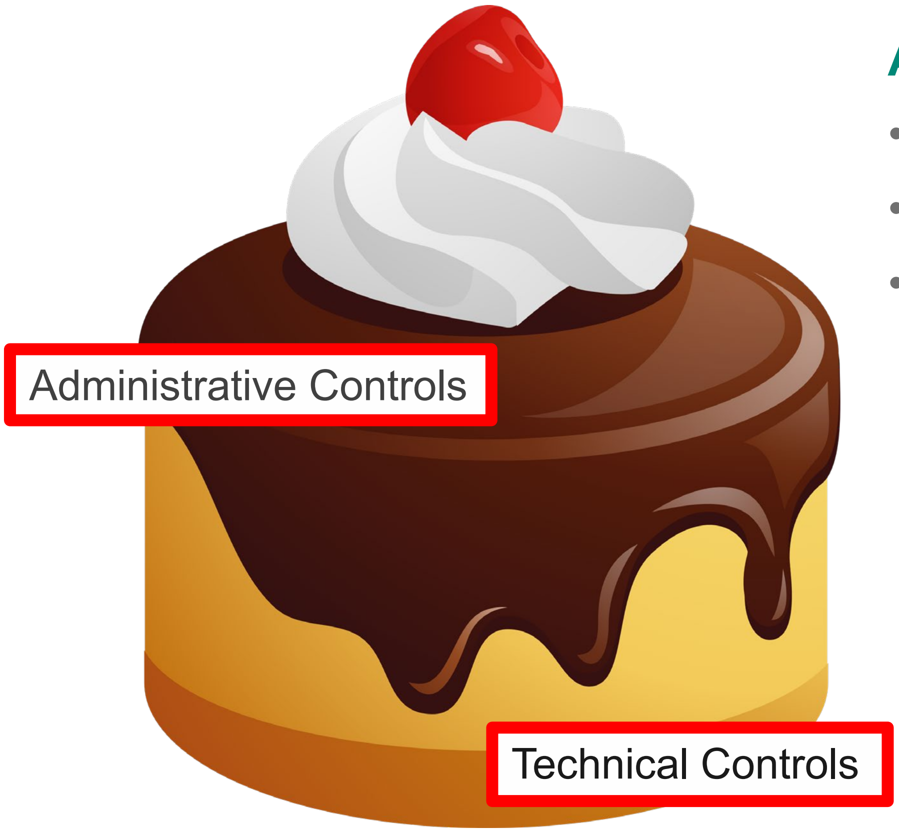
» May use mitigation strategies to reduce risk

» Documents methods and results of analysis

De-Identification – Safe Harbor

- HIPAA lists 18 identifiers that must be removed
- Of the individual and of relatives, employers, or household members of the individual
- And, the covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

De-Identification = Technical Controls + Administrative Controls



Administrative Controls:

- **Data Use Agreements**
- **Corresponding remedies**
- **Auditing and monitoring**

Source: Lagos, Y. & Polonetsky, J., *Public vs. Nonpublic data: The Benefits of Administrative Controls*. 66 Stan. L. Rev. Online 103 (Sept. 3, 2013).

De-Identification of Health Data: Law and Practice

OCR's Website	De-ID Resources and Tools	Sample Policies
Research Resources and Tools	Open Records/Freedom of Information Laws	Guidance on Data Release (ASTHO/NACCHO/AHCJ)

Sallie Milam, JD, CIPP/US/G, Deputy Director
Mid-States Region
smilam@networkforphl.org

The Network for Public Health Law –
National Office
7101 York Avenue South, Suite 270
Edina, MN 55435
Tel (952) 452-9700 X 108

Supporters

Robert Wood Johnson Foundation

The Network for Public Health Law is a national initiative of the Robert Wood Johnson Foundation. Please note that the Network for Public Health Law provides information and technical assistance on issues related to public health. The legal information and assistance provided in this document or within the webinar do not constitute legal advice or legal representation. For legal advice, please consult specific legal counsel.

The Network for Public Health Law

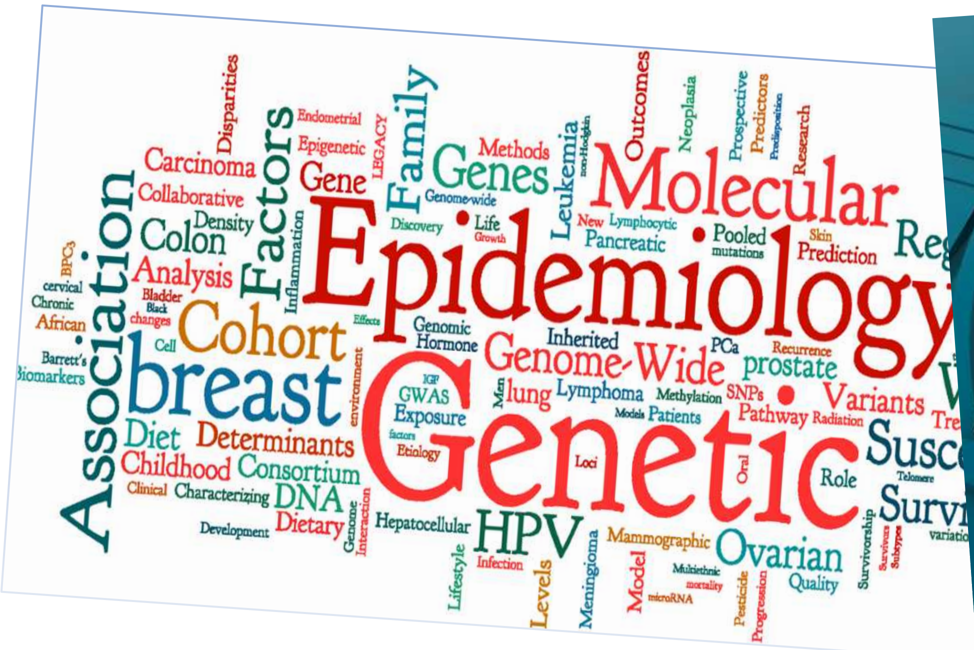
Data De-Identification Toolkit Webinar

De-identifying Public Health Data to Protect Privacy and Assure Public Good

May 16, 2019

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

A Historic and Important Societal Debate is underway…



Public Policy Collision Course

The Research Value of De-identified Health Data

57 UCLA LAW REVIEW 1701 (2010)

BROKEN PROMISES OF PRIVACY: RESPONDING TO THE SURPRISING FAILURE OF ANONYMIZATION

Paul Ohm*

Computer scientists have recently undermined our faith in the privacy-protecting power of anonymization, the name for techniques that protect the privacy of individuals in large databases by deleting information like names and social security numbers. These scientists have demonstrated that they can often "reidentify" or "deanonymize" individuals hidden in anonymized data with astonishing ease. By understanding this research, we realize we have made a mistake, labored beneath a fundamental misunderstanding, which has assured us much less privacy than we have assumed. This mistake pervades nearly every information privacy law, regulation, and debate, yet regulators and legal scholars have paid it scant attention. We must respond to the surprising failure of anonymization, and this Article provides the tools to do so.

Misconceptions about HIPAA De-identified Data:

"It doesn't work…" "easy, cheap, powerful re-identification" (Ohm, 2009 "*Broken Promises of Privacy*")

****Pre-HIPAA*** Re-identification Risks {Zip5, Birth date, Gender} able to identify 87%?, 63%, 28%? of US Population (Sweeney, 2000, Golle, 2006, Sweeney, 2013)

- Reality: HIPAA compliant de-identification provides important privacy protections
  - Safe harbor re-identification risks have been estimated at 0.04% (4 in 10,000) (Sweeney, NCVHS Testimony, 2007)
- Reality: Under HIPAA de-identification requirements, re-identification is expensive and time-consuming to conduct, requires substantive computer/mathematical skills, is rarely successful, and usually uncertain as to whether it has actually succeeded

Misconceptions about HIPAA De-identified Data:

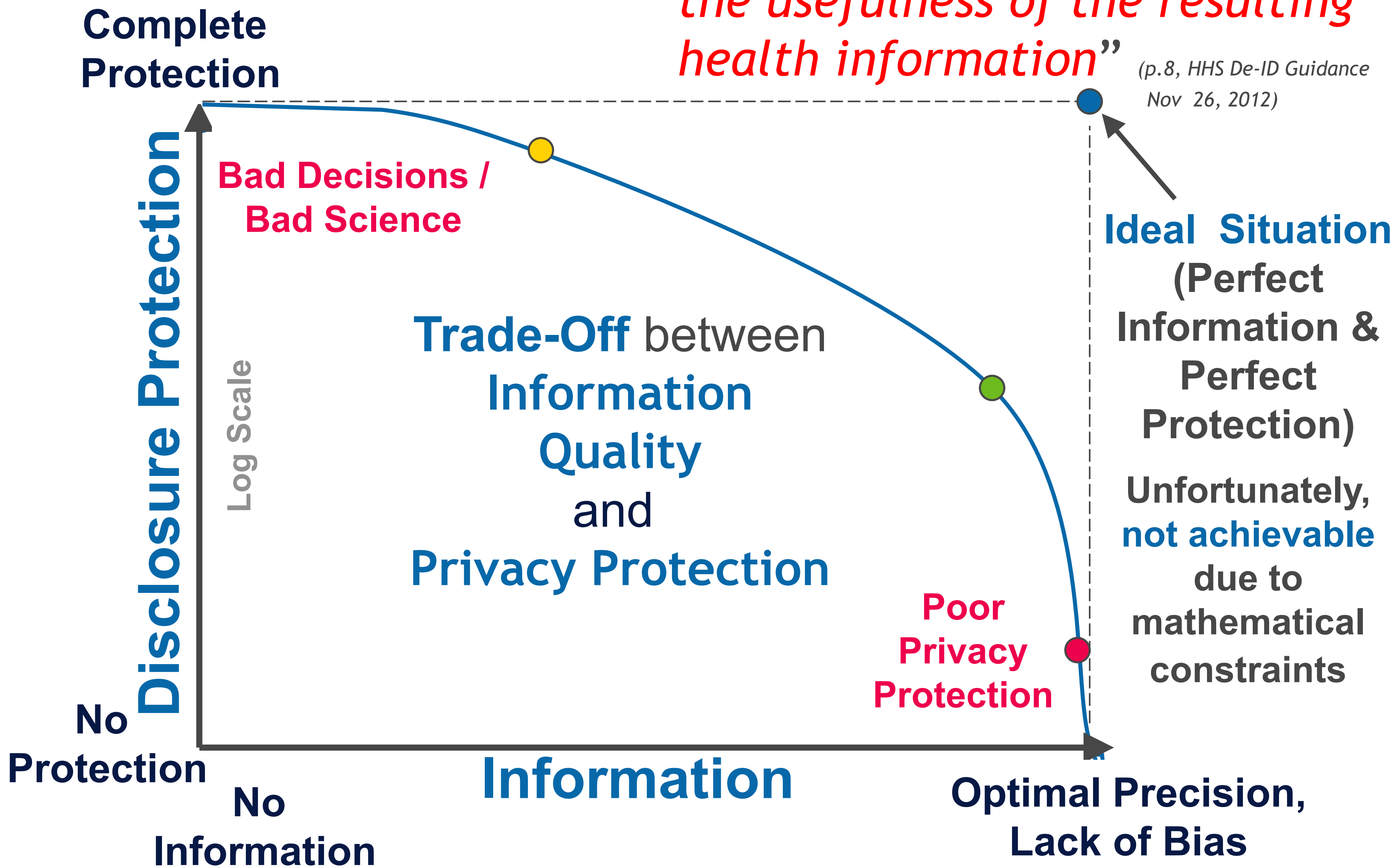
"It works perfectly and permanently…"

- Reality:
  - Perfect de-identification is not possible.
  - De-identifying does not free data from all possible subsequent privacy concerns.
  - Data is never permanently "de-identif<u>ied"</u>…

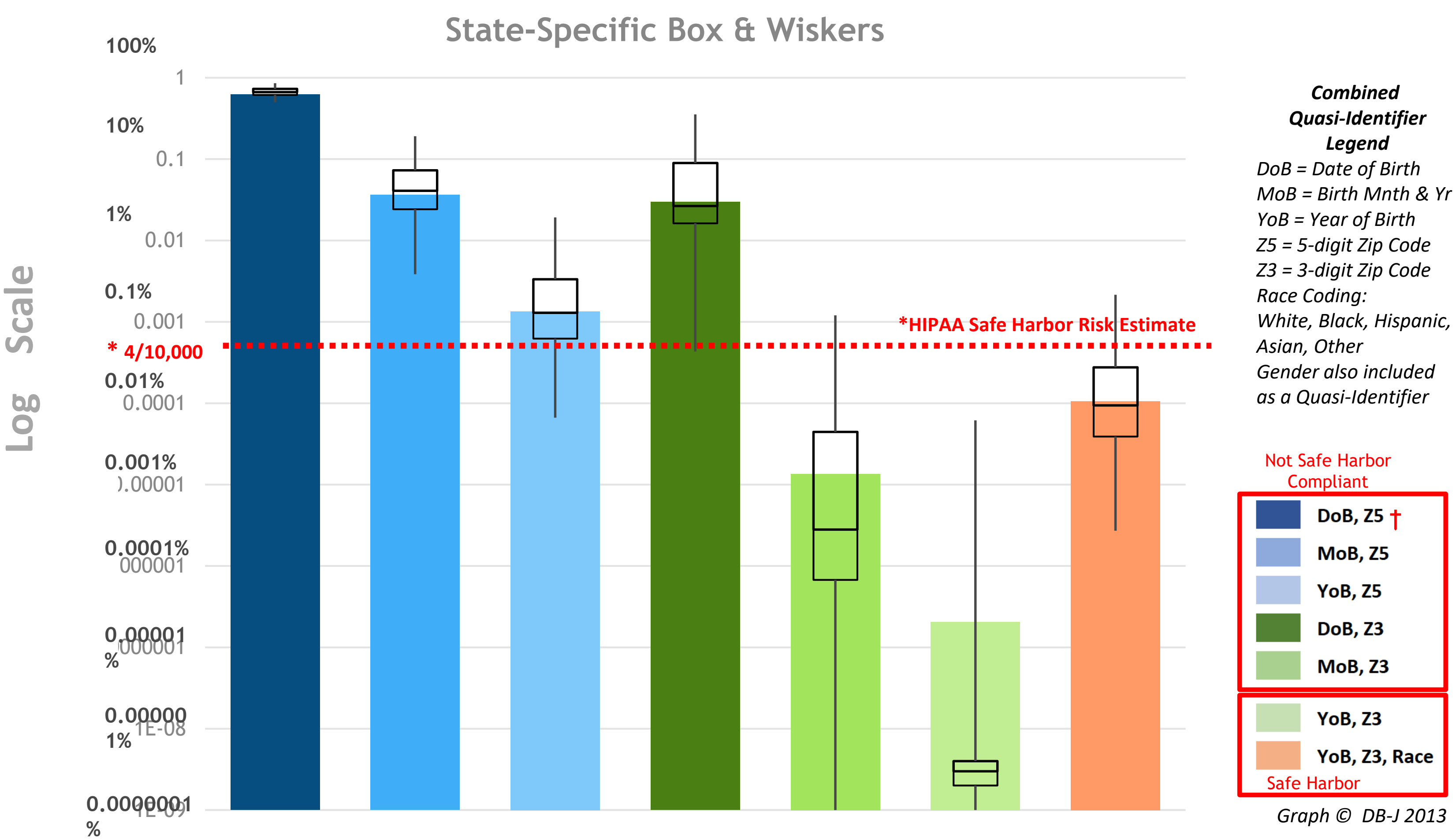
    There is no 100% guarantee that de-identified data will remain de-identified regardless of what you do with it after it is de-identified.

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)*



Re-identification Risks: Population Uniqueness



Data Source: 2010 U.S. Decennial Census

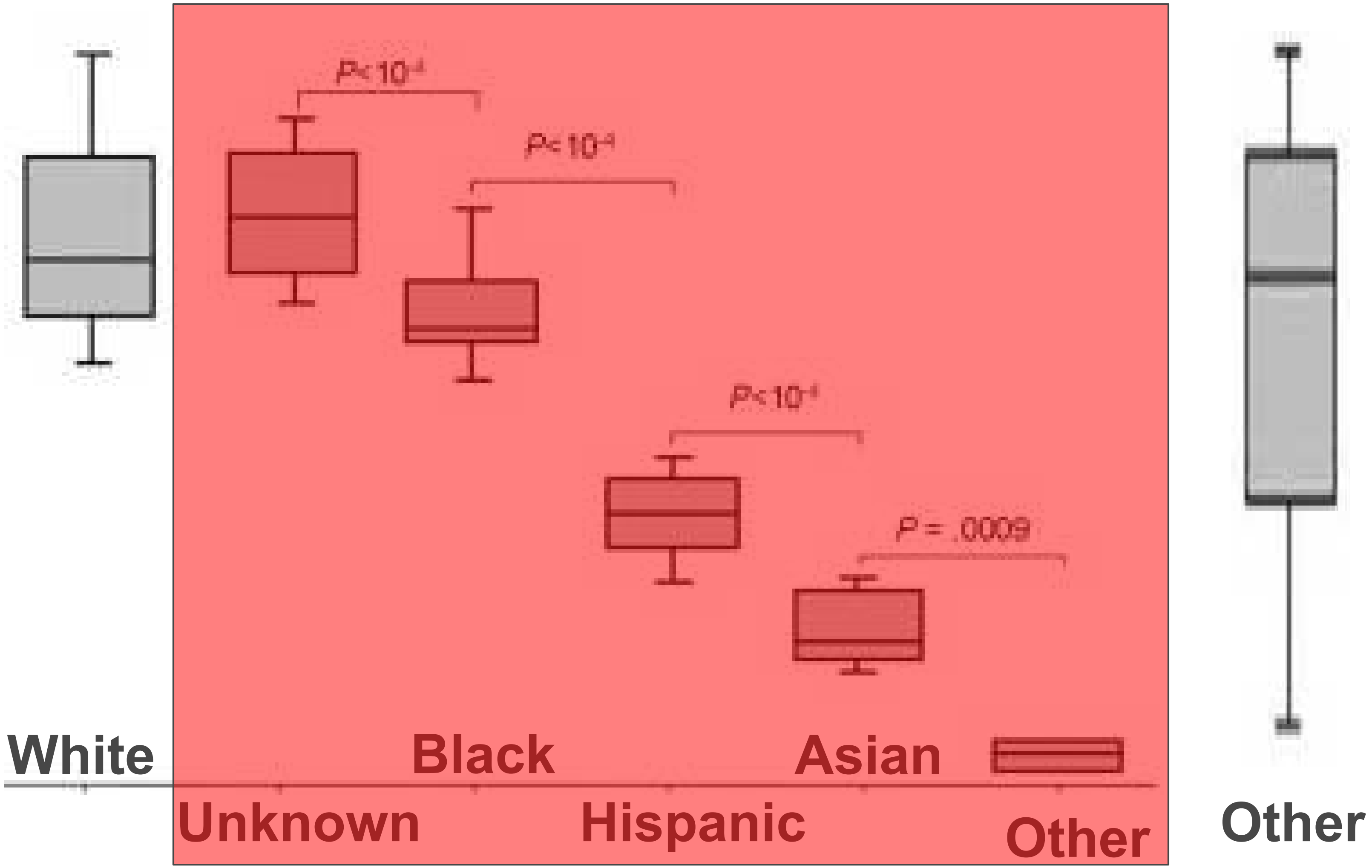
† *HIPAA Safe Harbor does not permit any Dates more specific than the year, or Geographic Units smaller than 3-digit Zip Codes (Z3).*

Balancing Disclosure Risk/Statistical Accuracy

- Balancing disclosure risks and statistical accuracy is essential because some popular de-identification methods (e.g. k-anonymity) can unnecessarily, and often undetectably, degrade the accuracy of de-identified data for multivariate statistical analyses or data mining (distorting variance-covariance matrixes, masking heterogeneous sub-groups which have been collapsed in generalization protections)
- This problem is well-understood by statisticians, but not as well recognized and integrated within public policy.
- Poorly conducted de-identification can lead to "bad science" and "bad decisions".

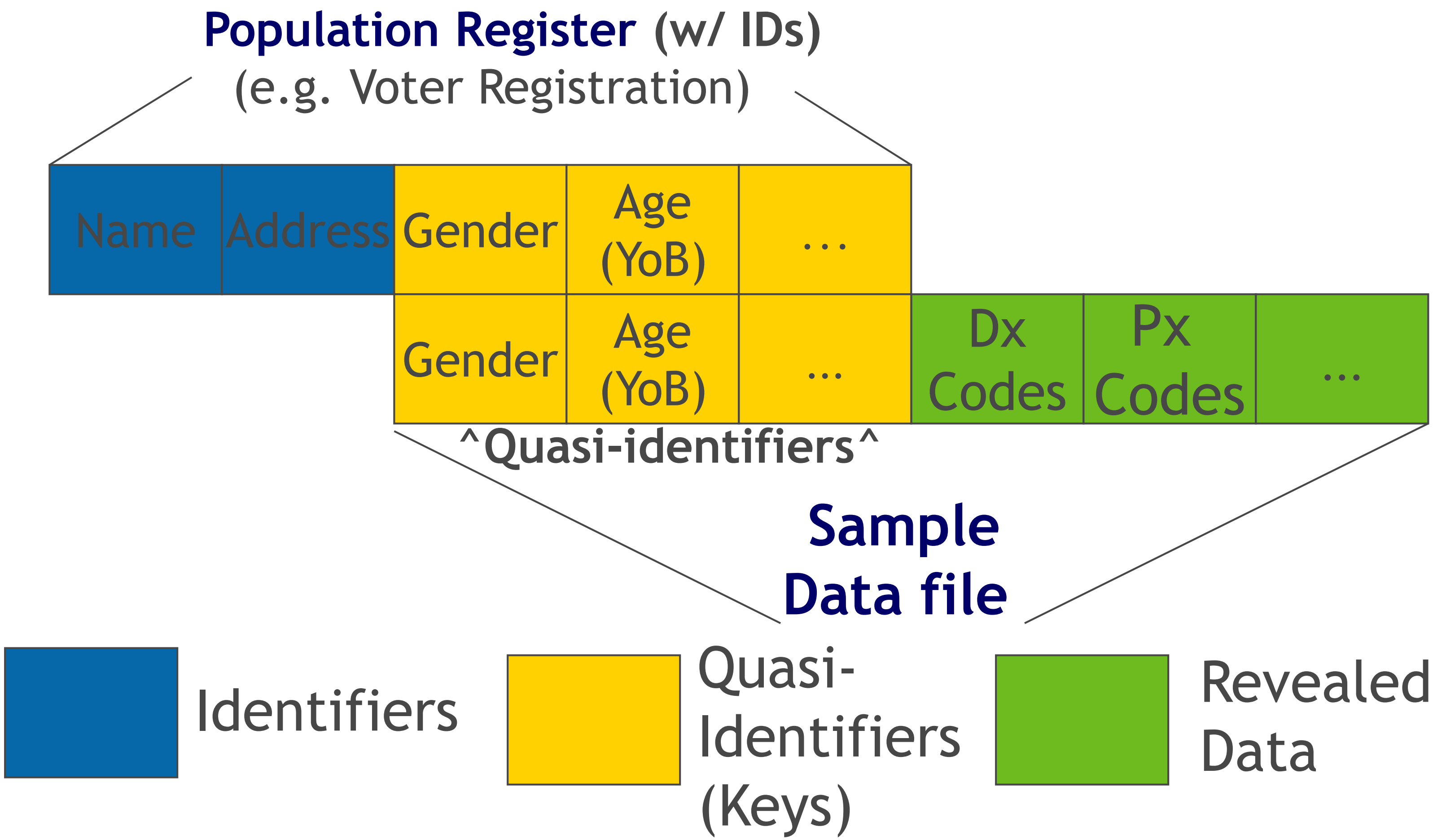
Reference: C. Aggarwal http://www.vldb2005.org/program/paper/fri/p901-aggarwal.pdf

De-identification Can Hide Important Differences



Record Linkage

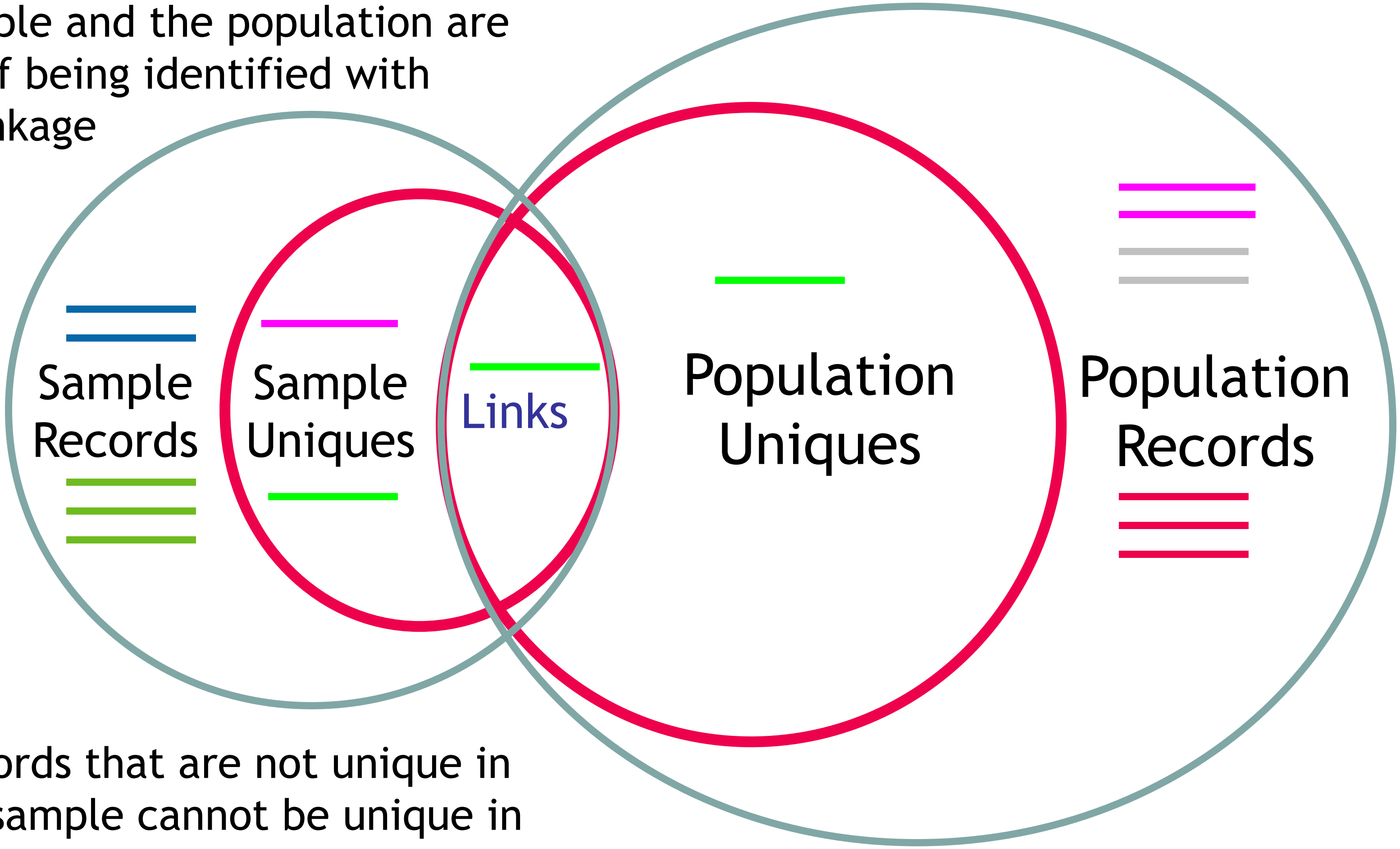
Record Linkage is achieved by matching records in separate data sets that have a common “Key” or set of data fields.



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

Percent of Regression Coefficients which changed Significance:

T.S. Gal et al./Journal of Biomedical Informatics xxx (2014) xxx–xxx

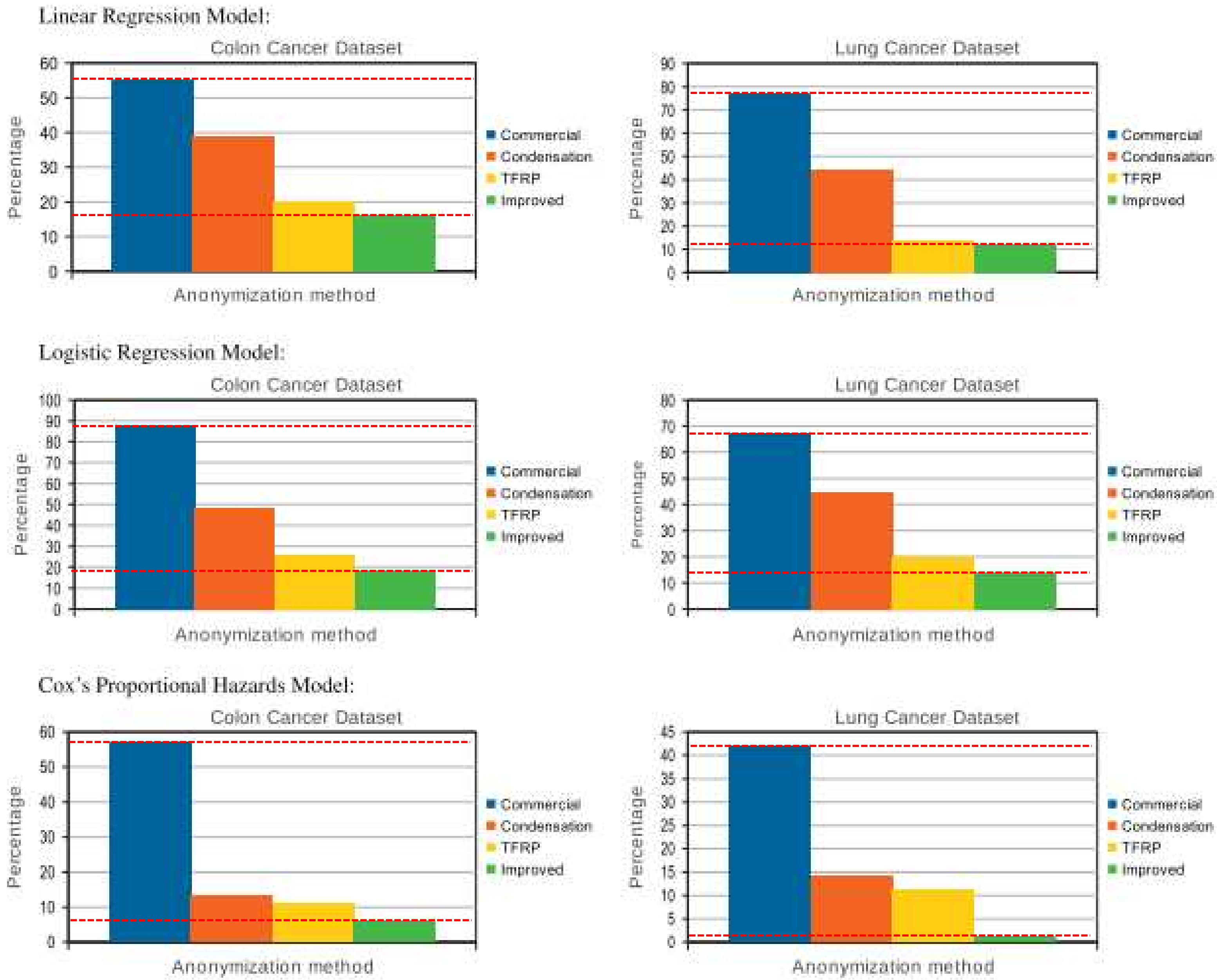


Fig. 1. Coefficients changed significance.

If this is what we are going to do to our ability to conduct accurate research - then… we should all just give up and go home.

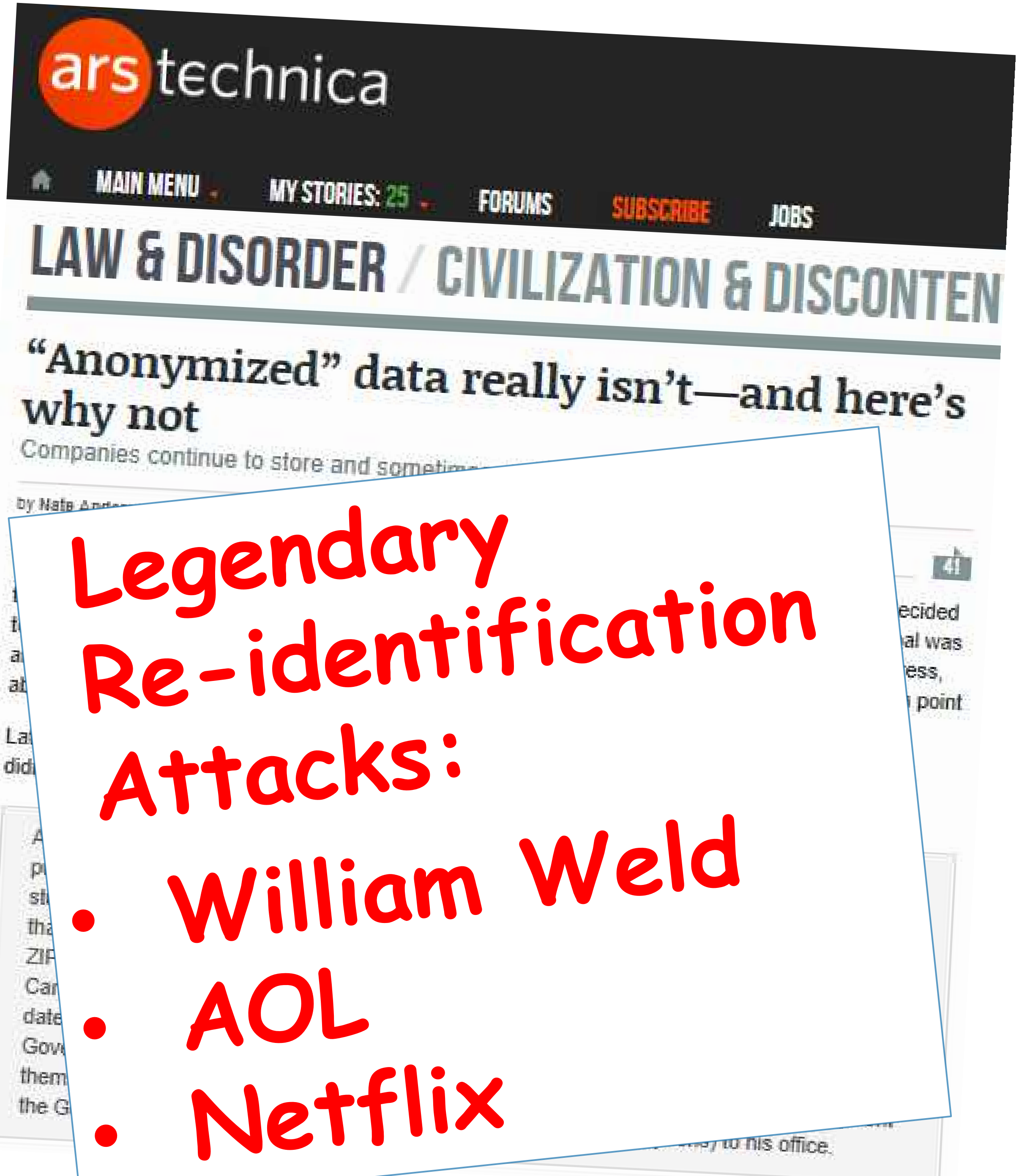
- Although poorly conducted de-identification can distort our ability to learn what is true leading to "bad science/decisions", this does not need to be an inevitable outcome.
- Well-conducted de-identification practice always carefully considers both the re-identification risk context and examines and controls the possible distortion to the statistical accuracy and utility of the de-identified data to assure de-identified data has been appropriately and usefully de-identified.
- But doing this requires a firm understanding/grounding in the extensive body of the statistical disclosure control/limitation literature.

Data Privacy Concerns are Far Too Important (and Complex) to be summed up with Catch Phrases or "Anecdata"

Eye-catching headlines and twitter-buzz announcing "***There's No Such Thing as Anonymous Data***" might draw the public's attention to broader and important concerns about data privacy in this era of "Big Data",

but such statements are essentially meaningless, even misleading, for further generalization without consideration of the specific de/re-identification contexts -- including the precise data details (e.g., number of variables, resolution of their coding schemas, special data properties, such as spatial/geographic detail, network properties, etc.) de-identification methods applied, and associated experimental design for re-identification attack demonstrations.

Good Public Policy demands reliable scientific evidence…



Legendary Re-identification Attacks:

- William Weld
- AOL
- Netflix

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

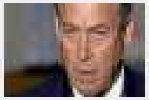
Re-identification Attacks	Quasi-Identifers (w/ HIPAA Safe Harbor exclusion data in Red)	Vulnerable Subgroup Targeted?	Used Stat. Sampling	Individuals w/ Alleged/Verified Re-identification	At-Risk Sample Size	Notable Headlines & Quotes	Attack Against HIPAA Compliant (or SDL Protected) Data?	Demonstrated 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.*

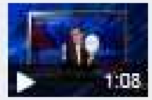
WA State Hospital Discharge Attack

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


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 in 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
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.
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

HERITAGE PROVIDER NETWORK
HEALTH PRIZE

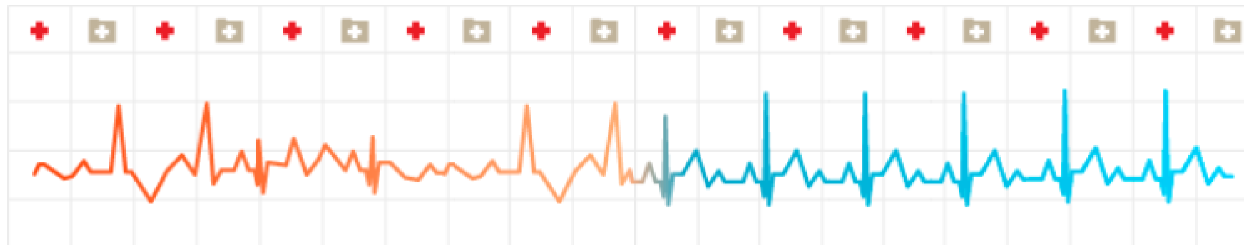
Sign Up In the News Judging Panel Visit HPN

Dashboard

- Home
- Data

Information

- Description
- Evaluation
- Rules



Improve Healthcare, Win $3,000,000.

N=113,000 Individuals

Identify patients who will be admitted to a hospital within the next year using historical claims data. (Enter by 06:59:59

Data de-identified with HIPAA Expert Determination method requiring very small risk

"No Evidence"?: Narayanan was engaged for Heritage Prize re-identification attack attempt. He was unable to re-identify anyone.

n = 0 were Re-identified

103 (18%) of the persons in study had their names embedded within their data files.

These "anonymous" names were used to help re-identify.

Without names only 28% could be re-identified by Zip5, Sex & DoB.

Adam Tanner, Contributor
I write about the business of personal data.
+ Follow (120)

Used Zip5, Sex, DoB & embedded Names

TECH | 4/25/2013 @ 3:47PM | 13,065 views

Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study

"Personal Genome Project" Attack

5 comments, 5 called-out + Comment Now + Follow Comments

A Harvard professor has re-identified the names of more than 40% of a sample of anonymous participants in a high-profile DNA study, highlighting the dangers that ever greater amounts of personal data available in the Internet era could unravel personal secrets.

From the onset, the Personal Genome Project, set up by Harvard Medical School

adamtanner/

Re-identification Demonstration Attack Summary

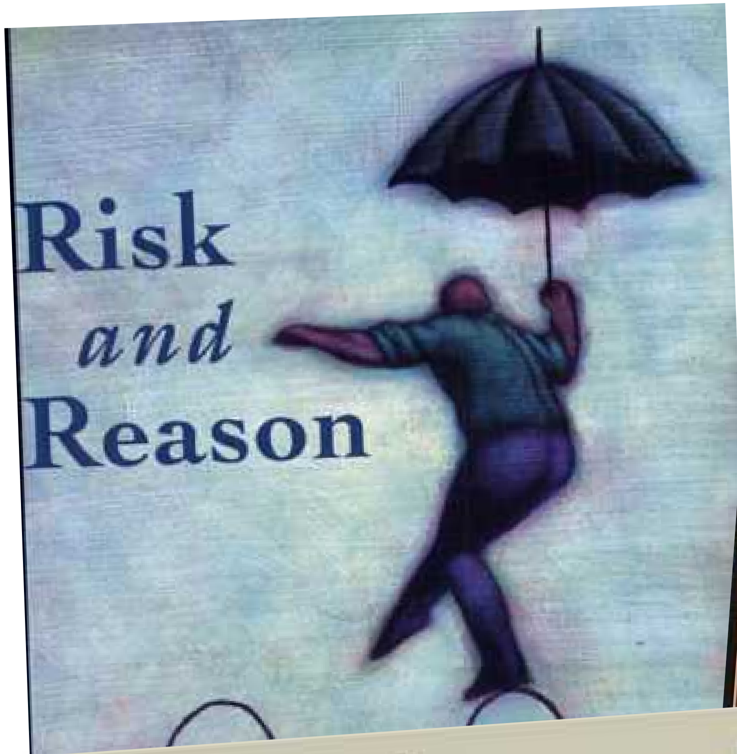
- For Ohm's famous "Broken Promises" attacks (Weld, AOL, Netflix) a total of n=4 people were re-identified out of 1.25 million.
- For attacks against HIPAA de-identified data (ONC, Heritage*), a total of n=2 people were re-identified out of 128 thousand.
  - ONC Attack Quasi-identifers: Zip3, YoB, Gender, 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.

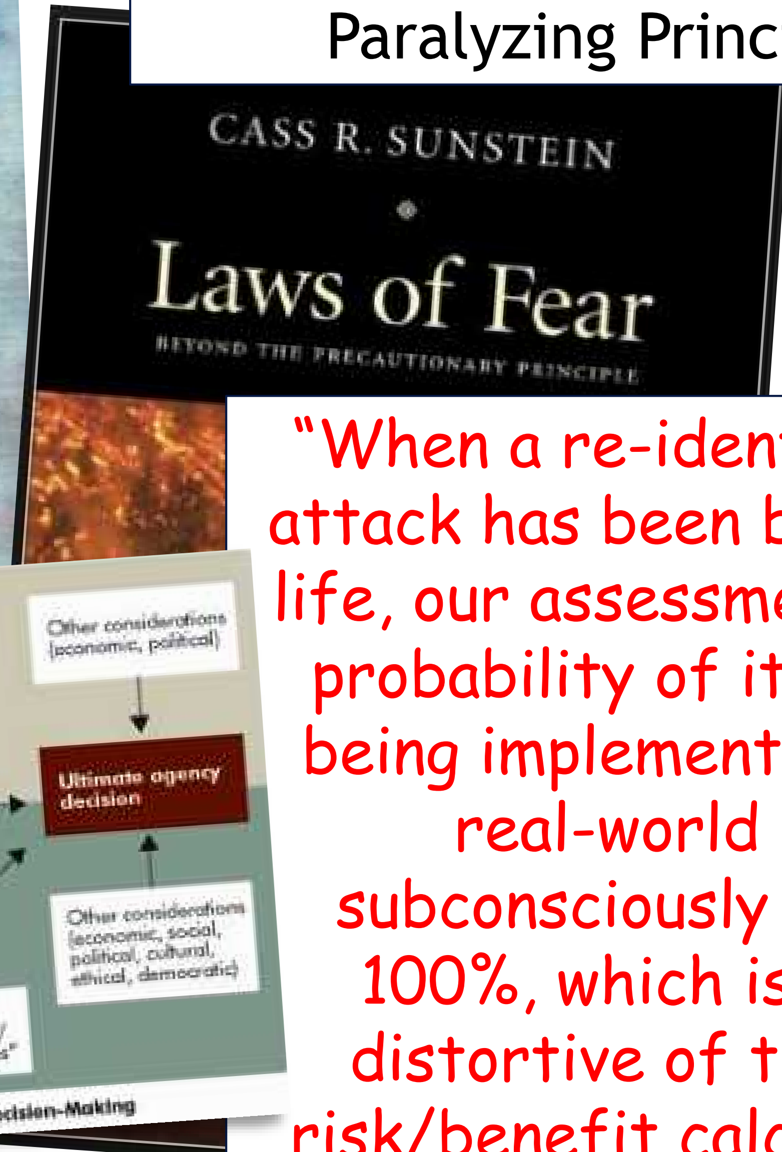
Let's get some perspective on this…

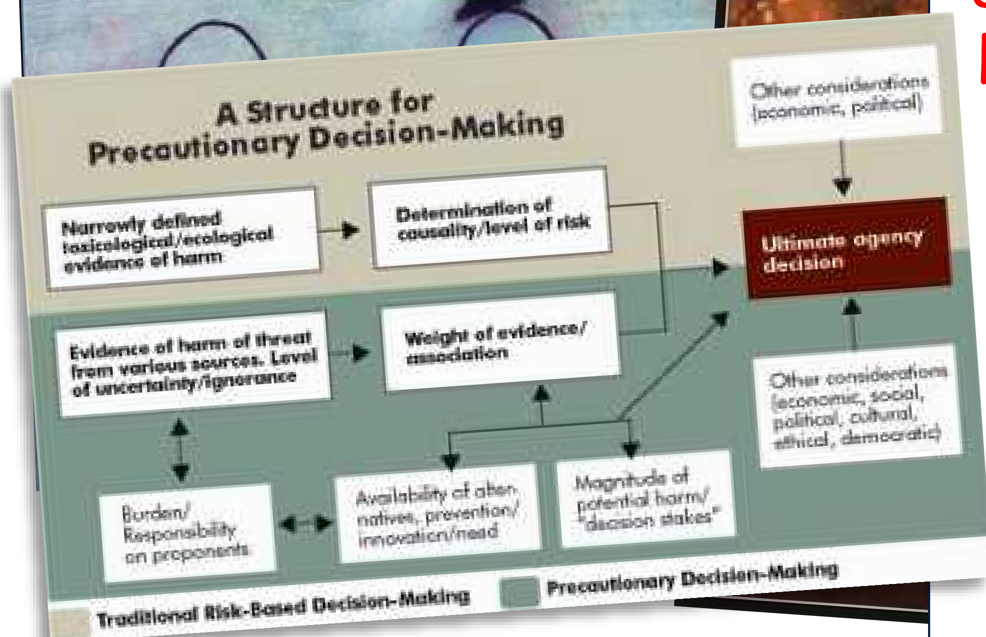
Obviously, This slide is **BLACK**

So clearly, De-identification Doesn't Work.

Precautionary Principle or Paralyzing Principle?







"When a re-identification attack has been brought to life, our assessment of the probability of it actually being implemented in the real-world may subconsciously become 100%, which is highly distortive of the true risk/benefit calculus that we face." – DB-J

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.

Re-identification Science Policy Short-comings:

6 ways in which "Re-identification Science" has (thus far) typically failed to best support sound public policies:

1. Attacking only trivially "straw man" de-identified data, where modern statistical disclosure control methods (like HIPAA) weren't used.
2. Targeting only especially vulnerable subpopulations and failing to use statistical random samples to provide policy-makers with representative re-identification risks for the entire population.
3. Making bad (often worst-case) assumptions and then failing to provide evidence to justify assumptions.

   Corollary: Not designing experiments to show the boundaries where de-identification finally succeeds.

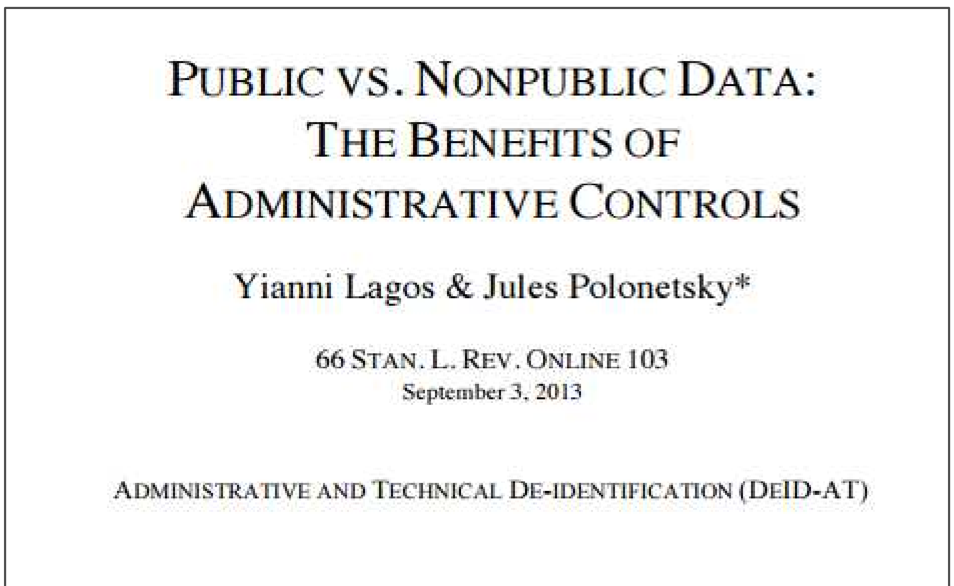
Re-identification Science Policy Short-comings:

6 ways in which "Re-identification Science" has (thus far) typically failed to support sound public policies (Cont'd):

4. Failing to distinguish between sample uniqueness, population uniqueness and re-identifiability (i.e., the ability to correctly link population unique observations to identities).
5. Failing to fully specify relevant threat models (using data intrusion scenarios that account for all of the motivations, process steps, and information required to successfully complete the re-identification attack for the members of the population).
6. Unrealistic emphasis on absolute "Privacy Guarantees" and *failure to recognize unavoidable trade-offs between data privacy and statistical accuracy/utility*.

Supplementing Technical Data De-identification with Legal/Administrative Controls

However, in many cases, because of the possibility of highly-targeted demonstration attacks, arriving at solutions which will appropriately preserve the statistical accuracy and utility will also require that we supplement our statistical disclosure limitation “technical” data de-identification methods with additional legal and administrative controls.



PUBLIC VS. NONPUBLIC DATA:
THE BENEFITS OF
ADMINISTRATIVE CONTROLS

Yianni Lagos & Jules Polonetsky*

66 STAN. L. REV. ONLINE 103
September 3, 2013

ADMINISTRATIVE AND TECHNICAL DE-IDENTIFICATION (DEID-AT)

Data Intrusion Scenarios:

- Prob(Re-identification) =
  Prob(Re-ident|Attempt)*Prob(Attempt)
- Note that Prob(Attempt) & Prob(Reident|Attempt) are actually not likely to be independent - higher re-identification probabilities are likely to increase re-identification attempts.
- Some very useful frameworks exist for characterizing Data Intrusion Scenarios:
  - Elliot & Dale, 1999, Duncan & Elliot Chapter 2, 2011
- We can frame the Prob(Attempt) in terms of: Motivation, Resources, Data Access, Attack Methods, Quasi-identifier Properties and Sets, Data Divergence Issues, and Probability of Success, Consequences and Alternatives for Goal Achievement

Recommended De-identified Data Use Requirements

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 within the data, or their relatives, family or household members.
2) Not link any other data elements to the data without obtaining 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

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.

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

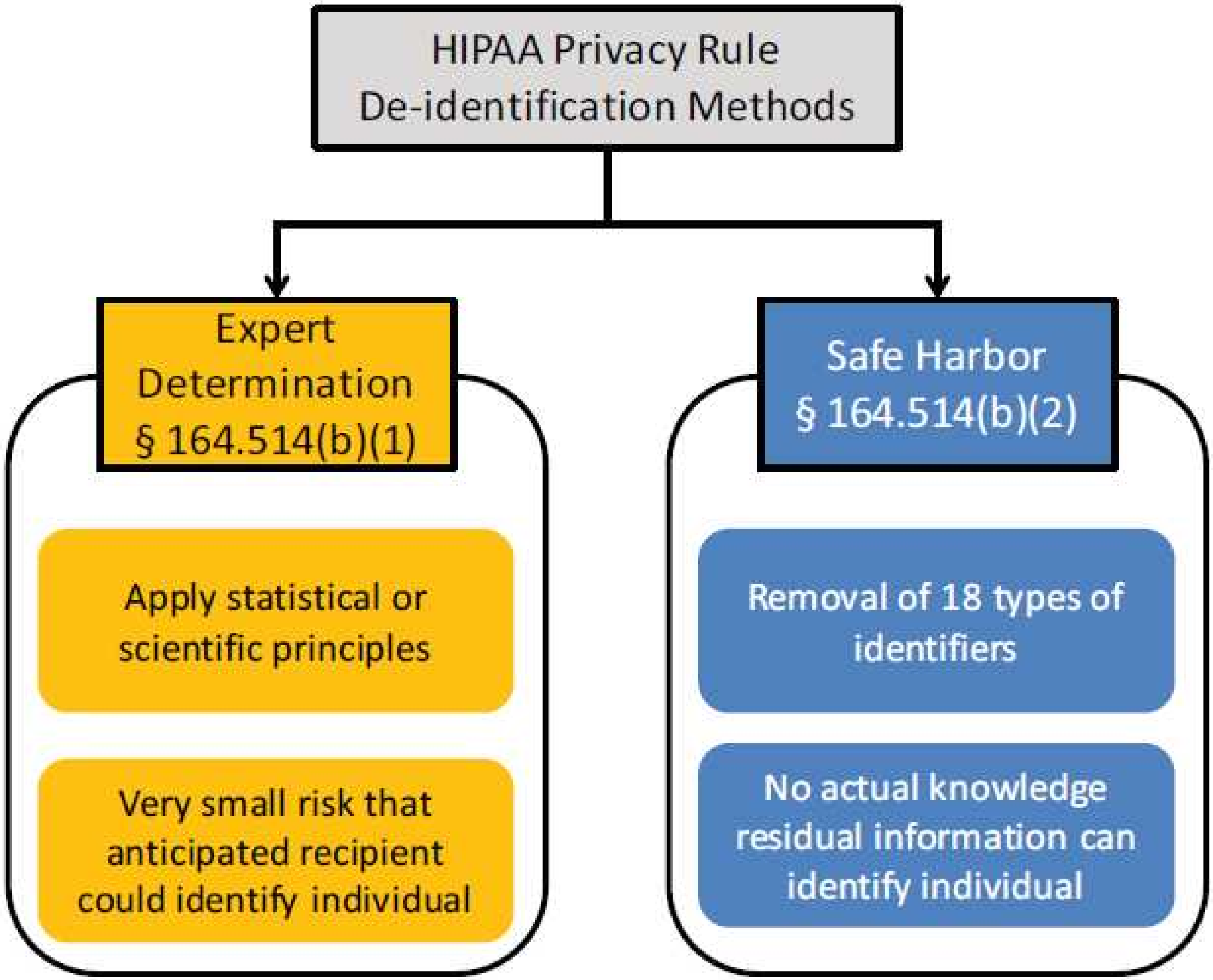
- http://blogs.law.harvard.edu/billofhealth/2013/05/29/public-policy-considerations-for-recent-re-identification-demonstration-attacks-on-genomic-data-sets-part-1-re-identification-symposium/

- 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/

Reserve Slides for Questions

Two Methods of HIPAA De-identification



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;

Why Privacy Science Must Become A "Systems Science"

- Paul Ohm described a dystopic vision that all information is effectively PII and that the failure of perfect de-identification would lead us through cycles of accretive re-identification toward a universal "database of ruin".

- This misconception ignores the underlying mathematical realities which indicate that when modern statistical disclosure limitation (SDL) methods can be used to effectively de-identify data, we will have resulting increases in "false positive" re-identifications.

- Such false positive linkages will practically prevent the ability of such systemic "crystallization" of iteratively linked de-identified data into accurate dossiers for the very vast majority of the population.

- Because of this de-identification, although imperfectly protective, is critical for reaching reasonable solutions which can continue to offer pragmatic and sustainable data obscurity in the evolving era of big data.

Why Privacy Science Must Become A "Systems Science"

- Modern SDL-based de-identification essential protections for preventing mass re-identification at scale and positions advocating for wholesale abandonment of de-identification due to less-than-perfect efficacy discard one of data privacy's most effective tools for an idealistic hope of perfect privacy protections makes "perfect the enemy of the good".
- Systems perspective using uncertainty analyses can help to apply consistent and rigorous probabilistic methods accounting for our uncertainty about the efficacy of various technical, administrative and legal protections at different stages in data intrusion scenarios to demonstrate that combining these methods can lead to useful assurance that (admittedly less than perfect) de-identification can still provide useful protections without resorting to only worst case scenarios about data intruder's knowledge.

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.

NETFLIX

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.

No silver bullet: De-identification still doesn't work

Arvind Narayanan
arvindn@cs.princeton.edu

Edward W. Felten
felten@cs.princeton.edu

July 9, 2014

Paul Ohm's 2009 article "Broken Promises of Privacy" spurred a debate in legal and policy circles on the appropriate response to computer science research on re-identification.[1] In this debate, the empirical research has often been misunderstood or misrepresented. A new report by Ann Cavoukian and Daniel Castro is full of such inaccuracies, despite its claims of "setting the record straight."[2]

We point out eight of our most serious points of disagreement with Cavoukian and Castro. The thrust of our arguments is that (i) there is no evidence that de-identification works either in theory or in practice[3] and (ii) attempts to quantify its efficacy are unscientific and promote a false sense of security by assuming unrealistic, artificially constrained models of what an adversary might do.

[3] At the risk of being pedantic, when we say that de-identification doesn't work we mean that it isn't effective at resisting adversarial attempts at re-identification.

No silver bullet: De-identification still doesn't work

Arvind Narayanan Edward W. Felten

2. Computing re-identification probabilities based on proof-of-concept demonstrations is silly.

Turning to the Netflix Prize re-identification study,[6] Cavoukian and Castro say: "the researchers re-identified only two out of 480,189 Netflix users, or 0.0004 per cent of users, with confidence."

This is an unfortunate misrepresentation of the results considering that the Netflix paper explicitly warns against this: "Our results should thus be viewed as a proof of concept. They do not imply anything about the percentage of IMDb users who can be identified in the Netflix Prize dataset."

Cautious interpretation is appropriate for simulated re-identification demonstrations in which no empirical evidence or justification is provided for the information requirements needed to actually accomplish re-identification. They often make worst-case assumptions and are don't design experiments to show the boundaries where de-identification finally succeeds.

No silver bullet: De-identification still doesn't work

Arvind Narayanan Edward W. Felten

2. Computing re-identification probabilities based on proof-of-concept demonstrations is silly.

Turning to the Netflix Prize re-identification study,[6] Cavoukian and Castro say: "the researchers re-identified only two out of 480,189 Netflix users, or 0.0004 per cent of users, with confidence."

This is an unfortunate misrepresentation of the results considering that the Netflix paper explicitly warns against this: "Our results should thus be viewed as a proof of concept. They do not imply anything about the percentage of IMDb users who can be identified in the Netflix Prize dataset."

Cavoukian and Castro seem to fundamentally miss the point of proof-of-concept demonstrations. By analogy, if someone made a video showing that a particular car security system could be hacked, it would be an error to claim that there is nothing to worry about because only one out of 1,000,000 such cars had been compromised.

To disclosure control statisticians and social scientists, it is equally nonsensical to suggest that the joint multivariate statistical distribution of quasi-identifiers has any uniformity comparable to a "car security system". This "*proof-of-concept*", as Narayanan acknowledges, says nothing about the re-identification risk beyond that it is not zero.

Identifying Personal Genomes by Surname Inference

Science
AAAS

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, ca[...] the identity of the target. A key feature of this technique is that it entirely re[...]es. We quantitatively analyze the [...]
id[...]
w[...]

nature International weekly journal of science
Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Aud[...]
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 'anonymo[...]

Erika Check Hayden

08 May 2013

PDF | Rights & Permissions

Our analysis projects a success rate of ~12% (SD = 2%) in recovering surnames of U.S. Caucasian males (Fig. 1B and fig. S2). This rate can be accomplished with a conservative threshold that would return a wrong surname in 5% of cases and label 83% of cases as unknown. Higher success rates of up to 18% can be achieved at the price of increased probability to recover an incorrect surname. Because our input cohort is based

TG
TG
ACTACTACTACTACTACTA
7 repeats

"Y-STR Surname" Attack Headlines

CSO | SECURITY AND RISK   Newsletters   Dashboard   RSS   Research Centers   White Pap

Identity & Access

News | Blogs | Tools & Templates | Security Jobs | Basics | Data Protection | Identity & Access | Business C

Home » Identity

"Your Biggest Genetic Secrets Can Now Be Hacked, Stolen, and Used for Target Marketing"

IN DEPTH

DNA hack could make medical privacy impossible

Researchers could find your name by taking samples from a distant cousin

» 1 Comment   Share 17   +1   Like 33   More

By Kevin Fogarty

March 11, 2013 — CSO —

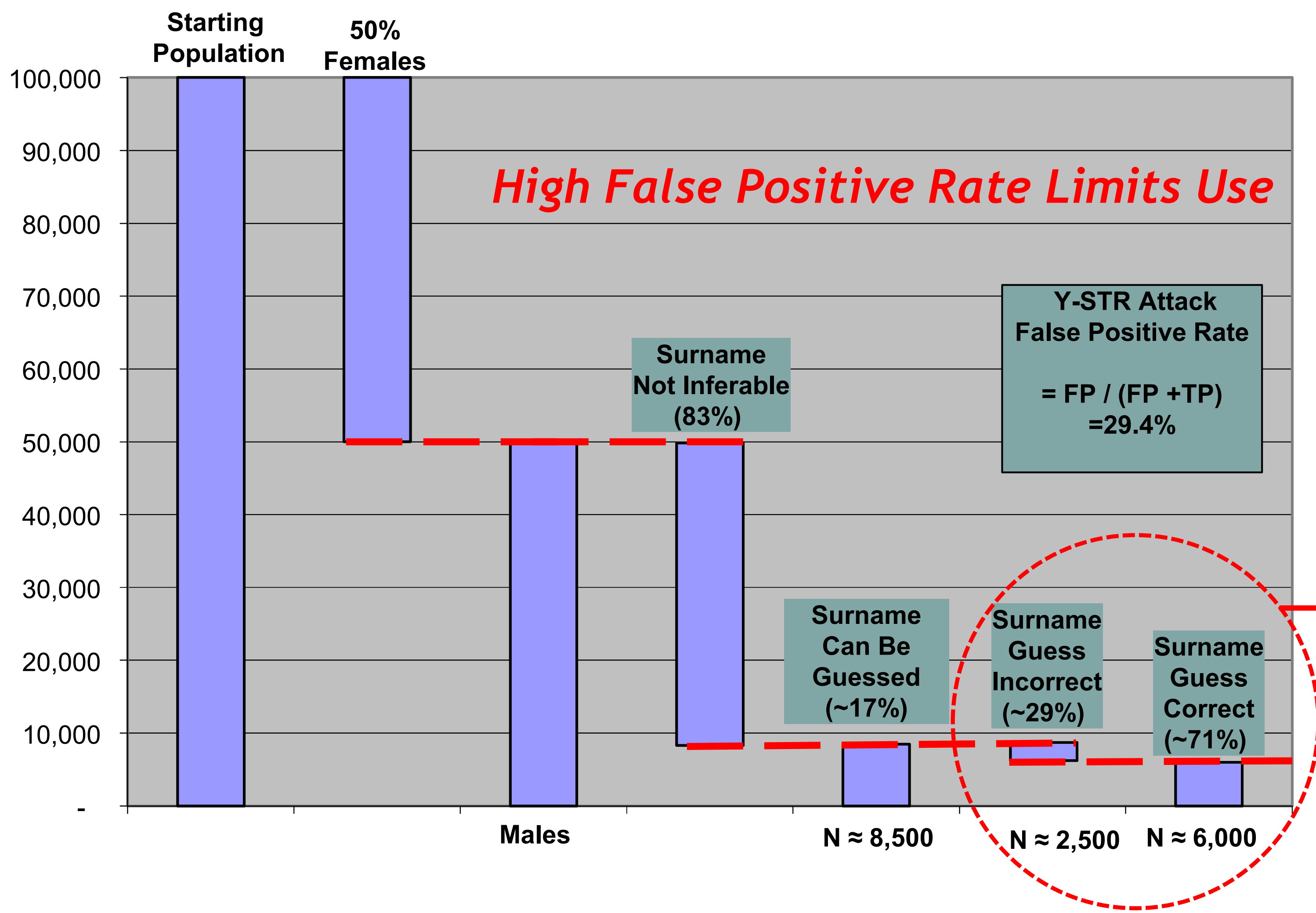
It may now be possible for anyone, even if they follow rigorous privacy and anonymity practices, to be identified by DNA data from people they do not even know.

Question 1: Is Y-STR Attack Economically Viable?

Probably not -- unclear whether it eventually could be.

Question 2: Is "De-identification" pointless?

No, removing State, Grouping YoB would help importantly.



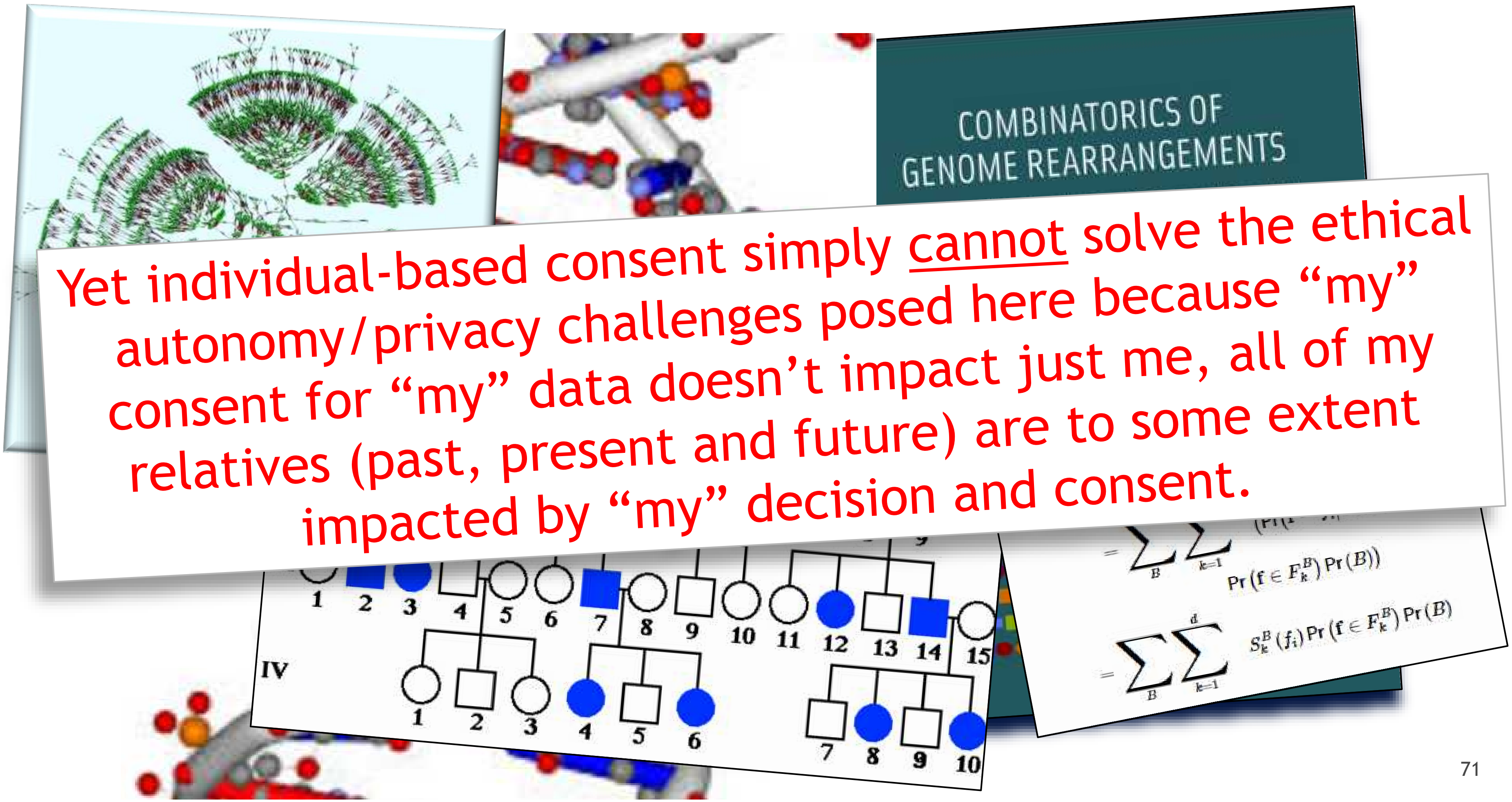
Re-ID isn't achieved by Surname Guess.

So what's the Threat Model?

Surname Guess Could Serve as a (Faulty) Quasi-identifier (e.g., w/ YoB & State)

But Will Produce Substantive Re-identification Errors

Given the inherent extremely large combinatorics of genomic data nested within inheritance networks which determine how genomic traits (and surnames) are shared with our ancestors/descendants, the degree to which such information could be meaningfully "de-identified" are non-trivial.



Yet individual-based consent simply cannot solve the ethical autonomy/privacy challenges posed here because "my" consent for "my" data doesn't impact just me, all of my relatives (past, present and future) are to some extent impacted by "my" decision and consent.

William Weld Re-identification

Dateline: May 18, 1996

- Massachusetts Governor William Weld was about to receive an honorary doctorate degree from Bentley College and give the keynote graduation address.
- Unbeknownst to him, he would instead make a critical contribution to the privacy of our health information. As he stepped forward to the podium, it wasn't what Weld said that now protects your health privacy, but rather what he did:
- Weld teetered and *collapsed unconscious* before a shocked audience. Weld's contribution to this story essentially ended here.

In the News: 1996

Massachusetts Governor William Weld Collapses During Commencement

By Martin Finucane AP (as run in Seattle Times) **May 21, 1996**
WALTHAM, Mass. - **Massachusetts Gov. William Weld collapsed yesterday during commencement at Bentley College**, but doctors said they found nothing seriously wrong with him. The **50-year-old governor** had just received an honorary doctorate of law when he fainted. "He fell headfirst (toward the podium), but they caught him," said Bill Petras, a graduating senior who sat five rows back from the stage. Weld was briefly unconscious, but was alert by the time he was lifted onto a stretcher and taken to an ambulance. The crowd applauded and Weld waved. Moments before fainting, Weld had started shaking as he approached the podium, Petras said.

Weld, a Republican who is challenging U.S. Sen. John Kerry for his Senate seat in November, had been scheduled to give the keynote address at Bentley's undergraduate commencement, but never got a chance to speak. "Right now, it **looks like maybe the flu,"** said Pam Jonah, one of Weld's press aides, adding that he **would stay in Deaconess-Waltham Hospital for 24 hours of observation. Doctors said they performed an electrocardiogram, a chest X-ray and blood tests, but found no immediate cause for concern.**

Ohm's Account of Weld Re-identification Attack

"At the time GIC released the data, William Weld, then Governor of Massachusetts, assured the public that GIC had protected patient privacy by deleting identifiers. In response, then-graduate student Sweeney started hunting for the Governor's hospital records in the GIC data. She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 residents and seven ZIP codes…"

Paul Ohm, 2010 Broken Promises of Privacy, UCLA Law Rev.

Ohm's Account of Weld Re-identification Attack

"…For twenty dollars, she purchased the complete voter rolls from the city of Cambridge, a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter. By combining this data with the GIC records, Sweeney found Governor Weld with ease. Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code. In a theatrical flourish, Dr. Sweeney sent the Governor's health records (which included diagnoses and prescriptions) to his office."

Paul Ohm, 2010 Broken Promises of Privacy, UCLA Law Rev.

Reality Check

U.S. Census Data Comparison for 1990 & 2000

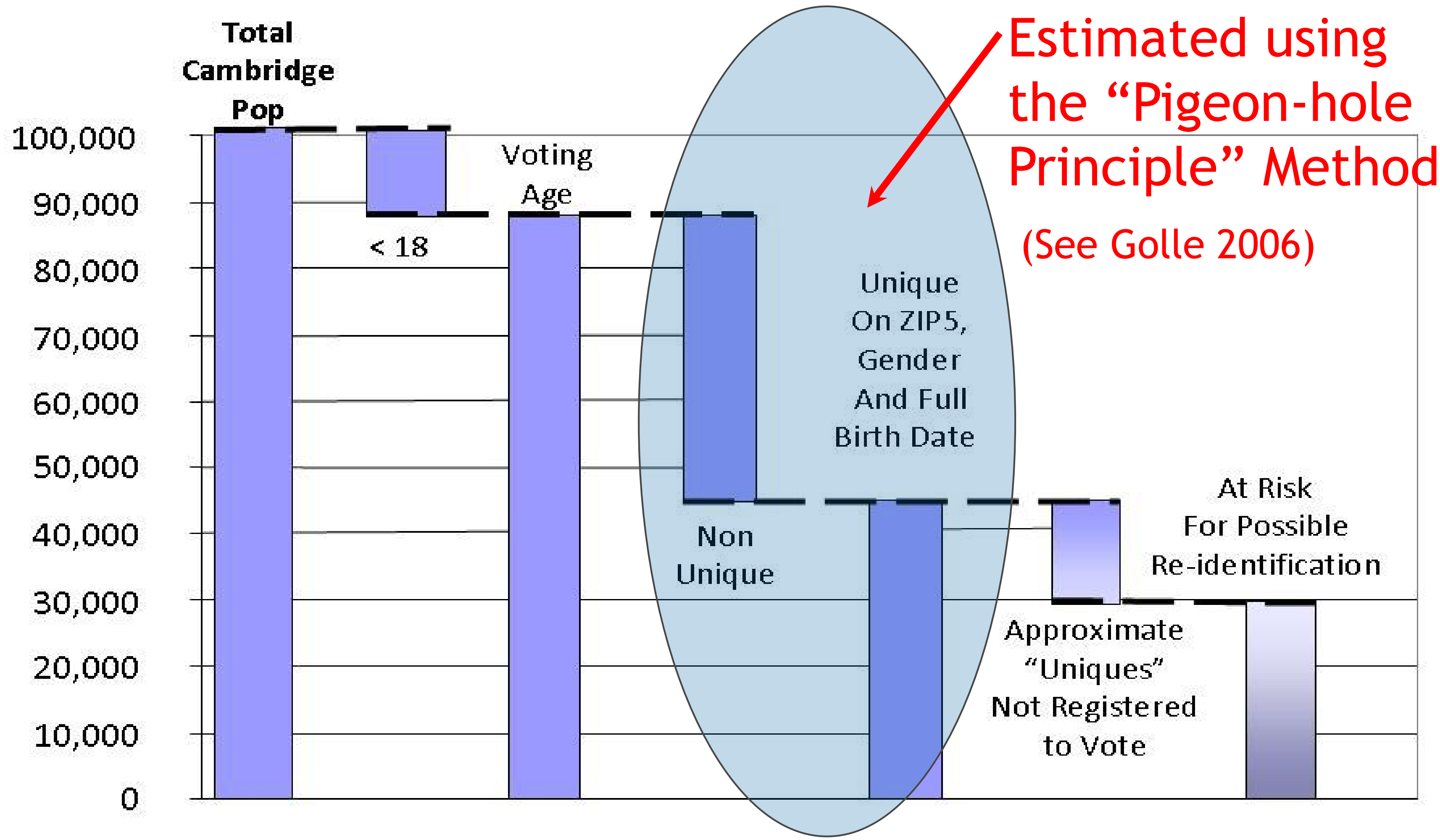
U.S. Census Population Counts and Estimated 1996-97 Total Population for Cambridge, MA		Percent
Total Cambridge, MA Population in 2000 Census	101,391	
Total Cambridge, MA Population 1996-1997*	99,435	100%
Total Cambridge, MA Population in 1990 Census	95,802	
Individuals in 1997 List Used for Weld Attack	54,805	55%
Estimated Unlisted Population	44,630	45%

Cambridge, MA Population and "Registered Voters" at Time of 1996-97 Weld /Cambridge Attack

Almost half of the Cambridge population could not have possibly been re-identified with the voter registration list.

Weld/Cambridge Attack

Estimated Proportion of the Cambridge Population subject to potential re-identification Risk



How Typical was Weld's Re-identification?

- Weld was extremely easy to re-identify within the GIC hospitalization data for Massachusetts employees for several reasons.
  - He was state employee and publicly known to have been hospitalized, so one could expect that Weld's hospital billing data **would be within the GIC hospital data** set.
    - This foreknowledge would not likely exist for random re-identification targets unknown to an imagined "data intruder".
    - For a randomly selected target, a data intruder would be unlikely to know whether any chance target individual was a state employee or had been recently hospitalized.
  - Weld was also sure to be registered to vote and publicly known to reside in Cambridge so he **could be found in the Cambridge Voter Registration list**.
    - This foreknowledge would not exist for random re-identification targets.

Myth of the "Perfect Population Register"

- The critical part of many re-identification efforts that is often assumed by disclosure scientists is the ***assumption of a perfect population register***.
- All Population registers will have data errors and be incomplete to some extent. (e.g. Nationwide voter registration levels typically are about 70%)
  - However, some types of data errors are more critical than others.
  - Persons who are not included in population registers will not have identifiers which can be linked to identify them.
    - Persons who are not in a population register can not re-identified, but they also indirectly reduce the probability of correct re-identification for others.
    - If only one person within a quasi-identifier set is missing from the population register, then the probability of correct re-identification drops to 50%; if two persons are missing, then the probability of correct re-identification is 33% , and so on.

Re-identification Failure and Success Conditions

	HOSPITAL DATA SET (Found In Data Set)	VOTER DATA SET (Found in Data Set)	NON-VOTERS (in Population)
1	**Not in Hospital Data**	Male 1/1/1945 02138 **Can't Re-identify (No Match)**	
2	Male 1/2/1945 02138	**Not in Voter Data**	Male 1/2/1945 02138 **Can't Re-identify (No Match)**
3	Male 1/3/1945 02138 Male 1/3/1945 02138	Male 1/3/1945 02138 **Can't Re-identify (> 1 Match)**	
4	Male 1/4/1945 02138 **Can't Re-identify (> 1 Match)**	Male 1/4/1945 02138 Male 1/4/1945 02138	
5	Male 1/5/1945 02138 ***Presumed Re-identification*** **(Has Only 50% Chance of Being a Correct Match)**	Male 1/5/1945 02138	Male 1/5/1945 02138 **Directly Protected From Re-identification**
6	Male 1/6/1945 02138 ***Correct Re-identification***	Male 1/6/1945 02138	

Note:
Figure illustrates only those limited cases where only one or two persons with shared "quasi-identifier" characteristics exist in either the healthcare data set or in the voter registration list.

Myth of the "Perfect Population Register"

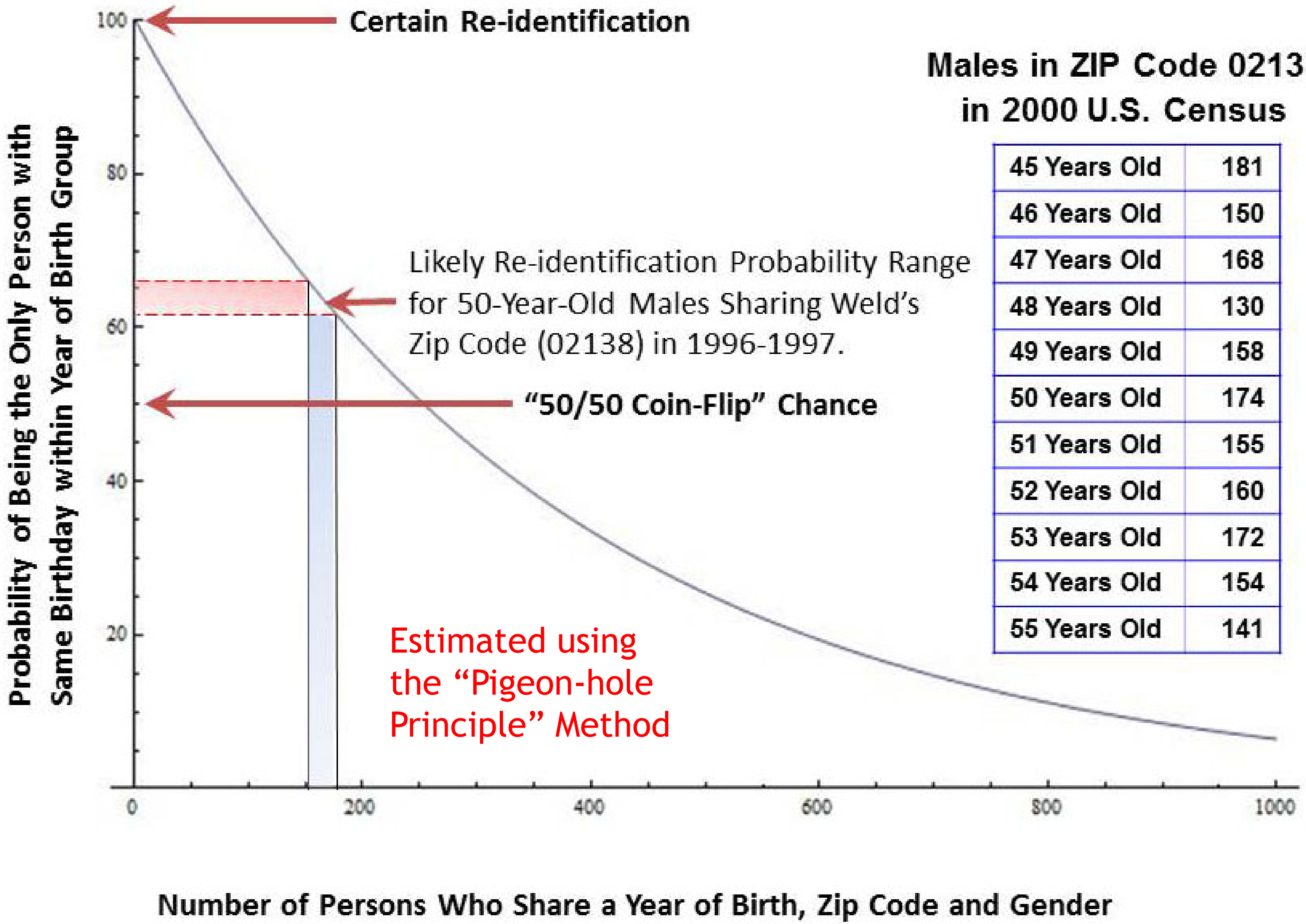
Note that in Row 5 on previous slide:

- Every person not within the voter list is directly protected from re-identification.
- Furthermore, their absence from the population register also reduces the probability that others who share their quasi-identifier set would be correctly re-identified.
- This is an extremely important limitation on re-identification when imperfect population registers are used.

Myth of the "Perfect Population Register"

- Without the important advantage of the public information regarding Weld's hospitalization, a data intruder would have had to go through a daunting process of making sure that there were not any other males living in the ZIP code 02138 at the time of Weld's collapse who were born on Weld's birthday in order to be certain that Weld was correctly re-identified using such a voter list attack method.
- There were approximately 35,000 persons living in ZIP code 02138 in 1997.
- It is difficult to imagine how a lone data intruder would have had the ability to complete this essential step in the re-identification process.

Weld/Cambridge Attack



Males in ZIP Code 02138 in 2000 U.S. Census

45 Years Old	181
46 Years Old	150
47 Years Old	168
48 Years Old	130
49 Years Old	158
50 Years Old	174
51 Years Old	155
52 Years Old	160
53 Years Old	172
54 Years Old	154
55 Years Old	141

Weld “Re-identified” with Voter List?

- While somewhat better than a flip of a coin, this 62-66% probability of accurate re-identification yields little confidence that Weld could actually be "re-identified" on the basis of the voter linkage attack.
- There was apparently about a 35% chance that the alleged re-identification was incorrect.
- Most people reading that Weld was re-identified using voter data are likely to assume that this "re-identification" was made with certainty and had been definitively accomplished via the linkage with voter data.

Weld "Re-identified" with Voter List?

- Even if we take Weld's "re-identification" as a probabilistic statement, a 35% chance for error greatly exceeds the usual p-value standards of 1% percent (or even 5%) for "statistical significance".
- Raises a important question – How we should define re-identification?
- Without the news coverage regarding Weld's public collapse and hospitalization, his "re-identification" might have never become the touchstone for privacy reform that it has become today.

Influence of Weld Re-identification on HIPAA

- It's difficult to overstate the influence of the Weld/ Cambridge voter list attack on U.S. health privacy policy - it had a clear impact on the development of the de-identification provisions within HIPAA Privacy Rule.
- The Weld re-identification has served an important illustration of privacy risks that were not adequately controlled prior to the advent of the HIPAA Privacy Rule in 2003.
- It is now quite clear that simple combinations of high -resolution variables (like birthdates and ZIP codes) can put an unacceptable portion of the population at risk for potential re-identification.

AOL Re-identification Attack

TECHNOLOGY

The New York Times

A Face Is Exposed for AOL Searcher No. 4417749

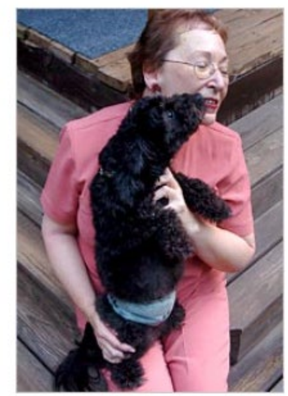
By MICHAEL BARBARO and TOM ZELLER Jr. AUG. 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold,



Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

Full Heritage Prize Data Elements

A. Members Table:
  1. MemberID (a unique member ID)
  2. AgeAtFirstClaim (member's age when first claim was made in the Data Set period)
  3. Sex

B. Claims Table:
  1. MemberID
  2. ProviderID (the ID of the doctor or specialist providing the service)
  3. Vendor (the company that issues the bill)
  4. PCP (member's primary care physician)
  5. Year (the year of the claim, Y1, Y2, Y3)
  6. Specialty
  7. PlaceSvc (place where the member was treated)
  8. PayDelay (the delay between the claim and the day the claim was paid for)
  9. LengthOfStay
  10. DSFS (days since first service that year)
  11. PrimaryConditionGroup (a generalization of the primary diagnosis codes)
  12. CharlsonIndex (a generalization of the diagnosis codes in the form of a categorized comorbidity score)
  13. ProcedureGroup (a generalization of the CPT code or treatment code)
  14. SupLOS (a flag that indicates if LengthOfStay is null because it has been suppressed)

C. Labs Table, contains certain details of lab tests provided to members.

D. RX Table, contains certain details of prescriptions filled by members.

E. DaysInHospital Tables, contains the number of days of hospitalization for each eligible member during Y2 and Y3 and includes:
  1. MemberID
  2. ClaimsTruncated (a flag for members who have had claims suppressed. If the flag is 1 for member xxx in DaysInHospital_Y2, some claims for member xxx will have been suppressed in Y1).
  3. DaysInHospital (the number of days in hospital Y2 or Y3, as applicable).

Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin

Results of de-anonymization. We carried out the experiments summarized in the following table:

Fig	Ratings	Dates	Type	Aux selection
4	Exact	$\pm 3 / \pm 14$	Best-guess	Uniform
5	Exact	$\pm 3 / \pm 14$	Best-guess	Uniform
6	Exact	$\pm 3 / \pm 14$	Entropic	Uniform
8	Exact	No info.	Best-guess	Not 100/500
9	± 1	± 14	Best-guess	Uniform
10	± 1	± 14	Best-guess	Uniform
11	Exact	No info.	Entropic	Not 100/500
12	± 1	± 14	Best-guess	Uniform

Where's experiment with ±1 Ratings, No Dates, Uniform movie selection, and a movie error allowance appropriate for watched vs. rated distinction?

Robust De-anonymization of Large Sparse Datasets

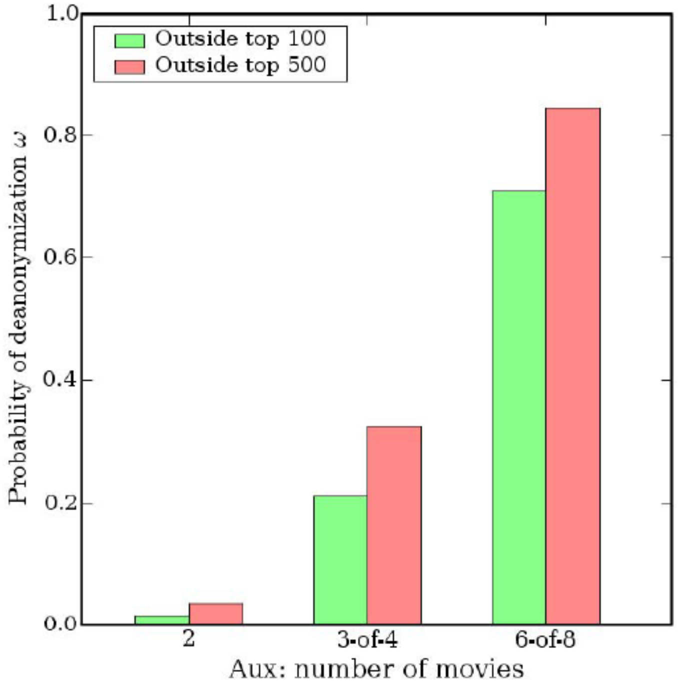


Figure 8. Adversary knows exact ratings but does not know dates at all.

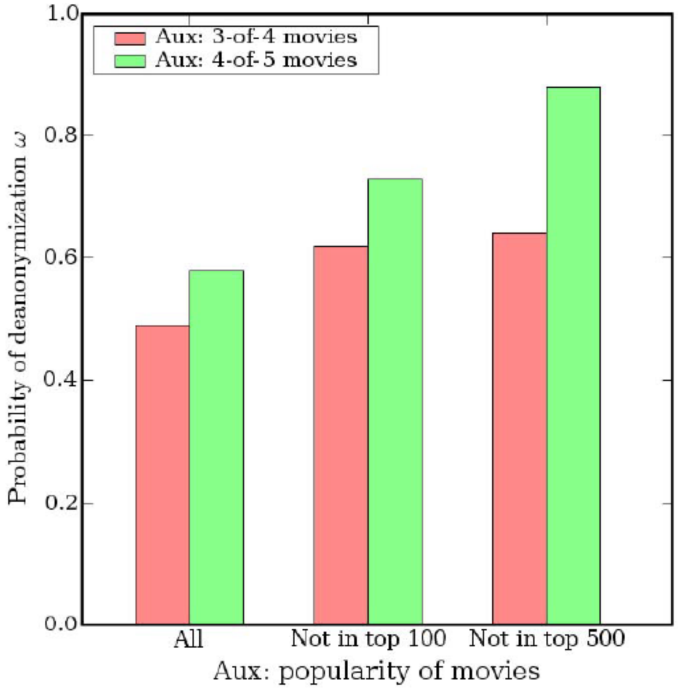


Figure 9. Effect of knowing less popular movies rated by victim. Adversary knows approximate ratings (± 1) and dates (14-day error).

SCIENTIFIC REPORTS

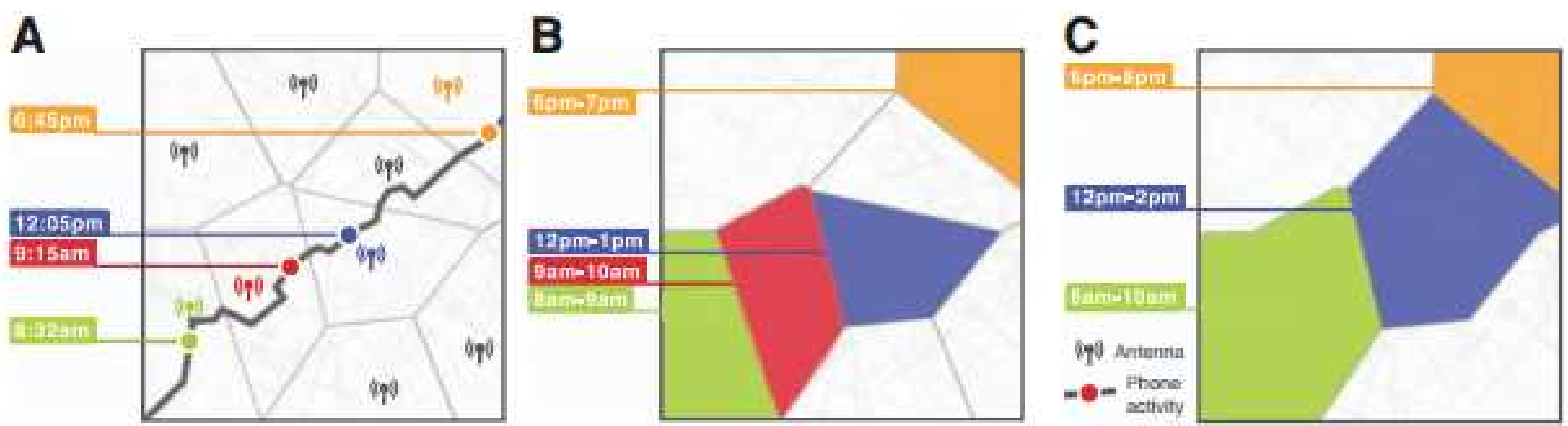
Published
25 March 2013

Unique in the Crowd: The privacy bounds of human mobility

Yves-Alexandre de Montjoye[1,2], César A. Hidalgo[1,3,4], Michel Verleysen[2] & Vincent D. Blondel[2,5]

Cell Data Uniqueness

We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. We coarsen the data spatially and temporally to find a formula for the uniqueness of human mobility traces given their resolution and the available outside information. This formula shows that the uniqueness of mobility traces decays approximately as the 1/10 power of their resolution. Hence, even coarse datasets provide little anonymity. These findings represent fundamental constraints to an individual's privacy and have important implications for the design of frameworks and institutions dedicated to protect the privacy of individuals.



Sample Unique ≠ Re-identifiable

Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset

SEPTEMBER 15, 2014 BY ATOCKAR 55 COMMENTS

NYC Taxi Data Attack

Violating Privacy

Let's consider some of the different ways in which this dataset can be exploited. If I knew an acquaintance or colleague had been in New York last year, I could combine known information about their whereabouts to try and track their movements for my own personal advantage. Maybe they filed a false expense report? How much did they tip? Did they go somewhere naughty? This can be extended to people I don't know – a savvy paparazzo could track celebrities in this way, for example.

There are other ways to go about this too. Simply focusing the search on an embarrassing night spot, for example, opens the door to all kinds of information about its customers, such as name, address, marital status, etc. Don't believe me? Keep reading…

Stalking celebrities

Unsalted Crypto-Hash

Firs… can use any combination of known characteristics that

INFO/LAW

The Antidote for "Anecdata": A Little Science Can Separate Data Privacy Facts from Folklore

Posted on November 21st, 2014 by jyakowitz

Guest post by Daniel Barth-Jones

NYC Taxi Data Attack

For anyone who follows the increasingly critical topic of data privacy closely, it would have been impossible to miss the remarkable chain reaction that followed the New York TLC's (Taxi and Limousine Commission) recent release of data on more than 173 million taxi rides in response to a FOIL (Freedom of Information Law) request by Urbanist and self-described "Data Junkie" Chris Whong. It wasn't long at all after the data went public that the sharp eyes and keen wit of software engineer Vijay Pandurangan detected that taxi drivers' license numbers and taxi plate (or medallion) numbers hadn't been anonymized properly and could

http://blogs.law.harvard.edu/infolaw/2014/11/21/the-antidote-for-anecdata-a-little-science-can-separate-data-privacy-facts-from-folklore/

Stars: Passenger Privacy in the NYC Taxicab Dataset

introducing the concept of "differential privacy" and announcing Neustar's

Harvard Business Review

REGULATION

There's No Such Thing as Anonymous Data

January 2015

DATA PRIVACY DAY

ENT TEXT SIZE PRINT

About a decade ago, a hacker said to me, flatly, "Assume every card in your wallet is compromised

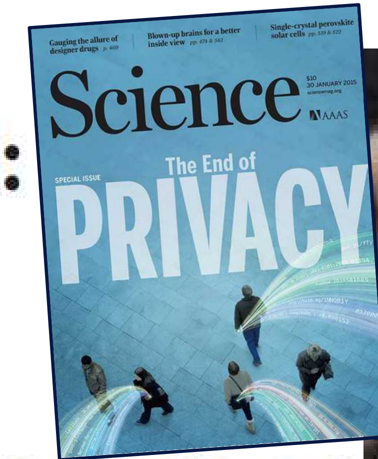
For scientists, the vast amounts of data that people shed every day offer great new opportunities but new dilemmas as well. New computational techniques can identify people or trace their behavior by combining just a few snippets of data. There are ways to protect the private information hidden in big data files, but they limit what scientists can learn; a balance must be struck. Some medical researchers acknowledge that keeping patient data private is becoming almost impossible;

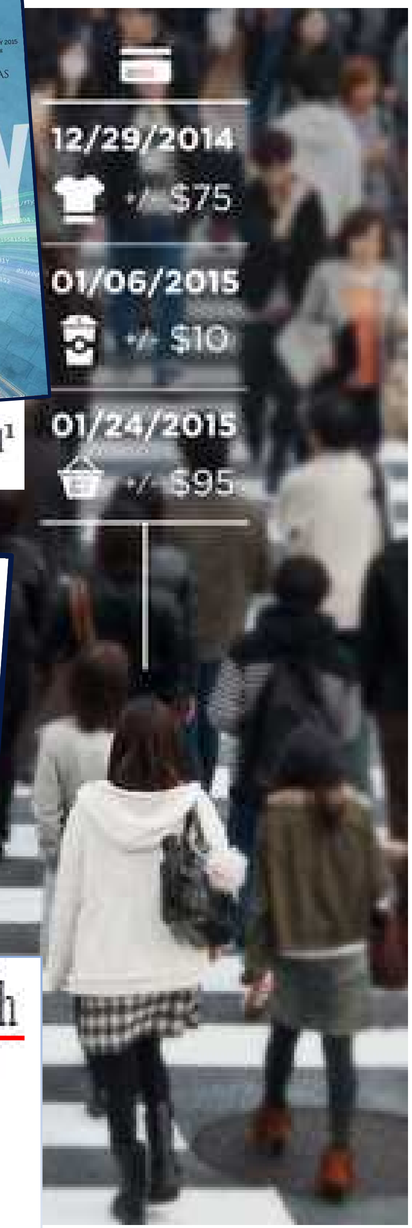
IDENTITY AND PRIVACY

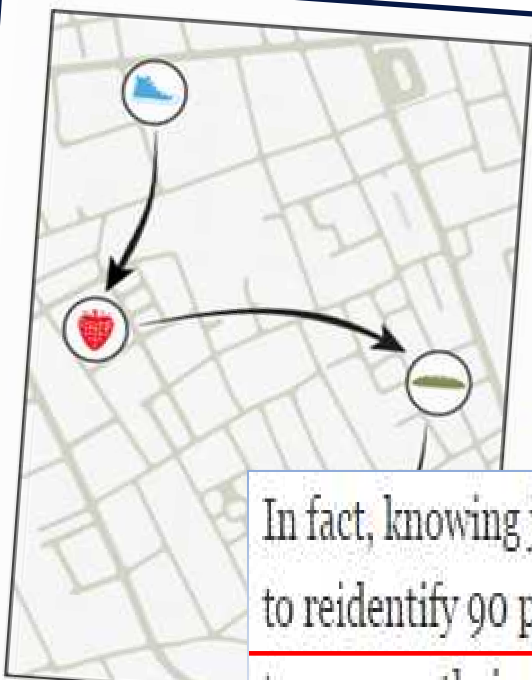
Credit Card Data Uniqueness

Unique in the shopping mall: On the reidentifiability of credit card metadata

Yves-Alexandre de Montjoye,[1*] Laura Radaelli,[2] Vivek Kumar Singh,[1,3] Alex "Sandy" Pentland[1]







shop	user_id	time	price	price_bin
	7abc1a23	09/23	$97.30	$49 – $146
	7abc1a23	09/23	$15.13	$5 – $16
	3092fc10	09/23	$43.78	$16 – $49
	7abc1a23	09/23	$4.33	$2 – $5

In fact, knowing just four random pieces of information was enough to reidentify 90 percent of the shoppers as unique individuals and to uncover their records, researchers calculated.

INFO/LAW

INFORMATION, LAW, AND THE LAW OF

Science

AAAS

LETTERS

Assessing data intrusion threats

Barth-Jones, et.al.

Y.-A. DE MONTEJOYE *et al.*'s Report "Unique in the shopping mall: On the reidentifiability of credit card data" (special section on The End of Privacy, 30 January, p. 536) led to a widespread media sensation proclaiming that reidentification is easy with only a few pieces of credit card data (*1–3*). Although we agree with de Montejoye *et al.* that data disclosure practices must be responsibly balanced with data privacy and utility, we are concerned that the study's findings reflect unrealistic data intrusion threats. Making policy deci...

Is De-Identification Dead Again?

...sted on April 28th, 2015 by jyakowitz

...rlier this year, the journal Science published a study called "Unique in ... Shopping Mall: On the Reidentifiability of Credit Card Metadata" by ...es-Alexandre de Montjoye et al. The article has reinvigorated claims that ...identified research data can be reidentified easily. These claims are not ...w, but their recitation in a vaunted science journal led to a new round of ...nic in the popular press.

Sample Unique ≠ Re-identifiable

1.1 Million = small sample fraction

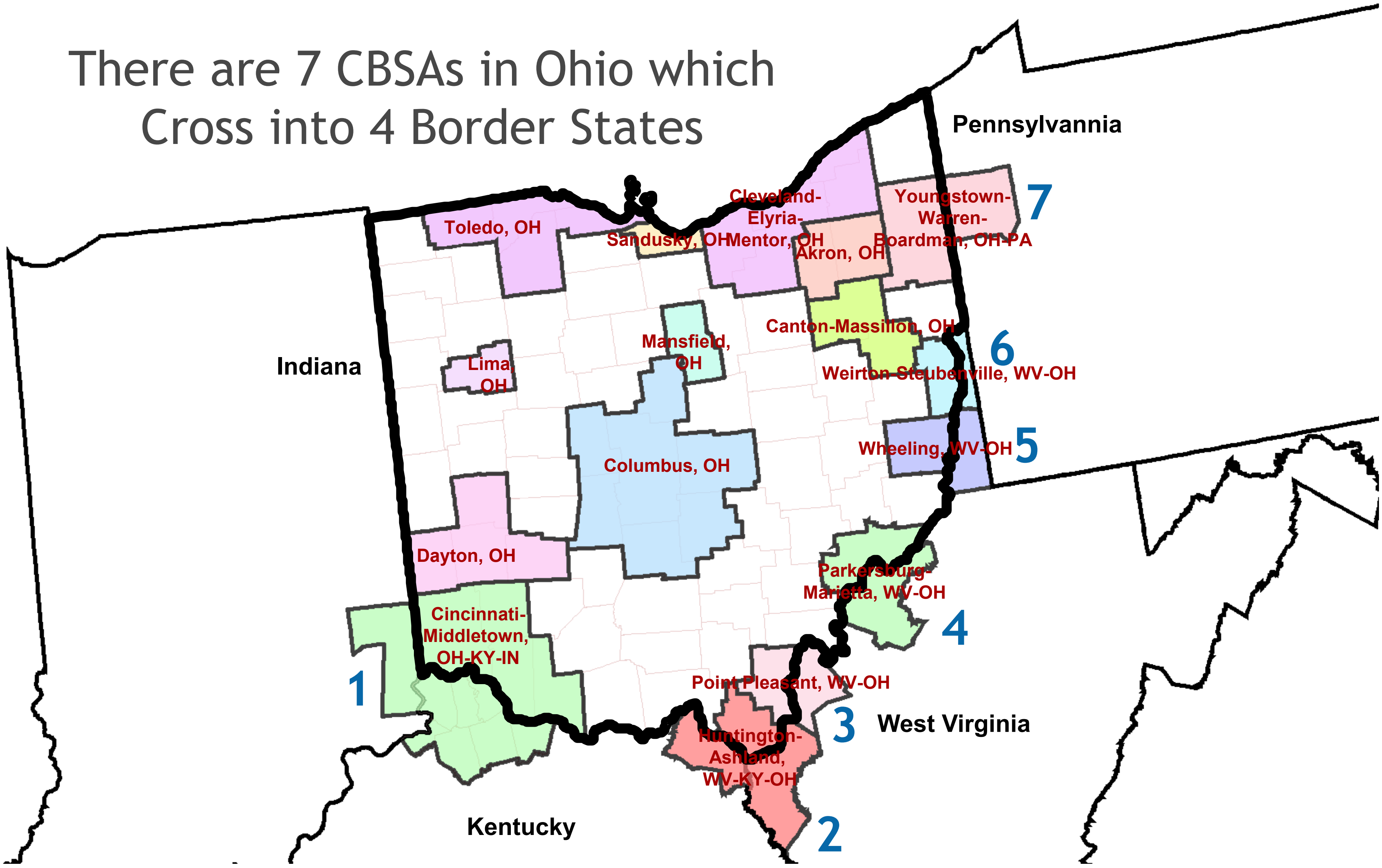
https://blogs.law.harvard.edu/infolaw/2015/04/28/is-de-identification-dead-again/

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

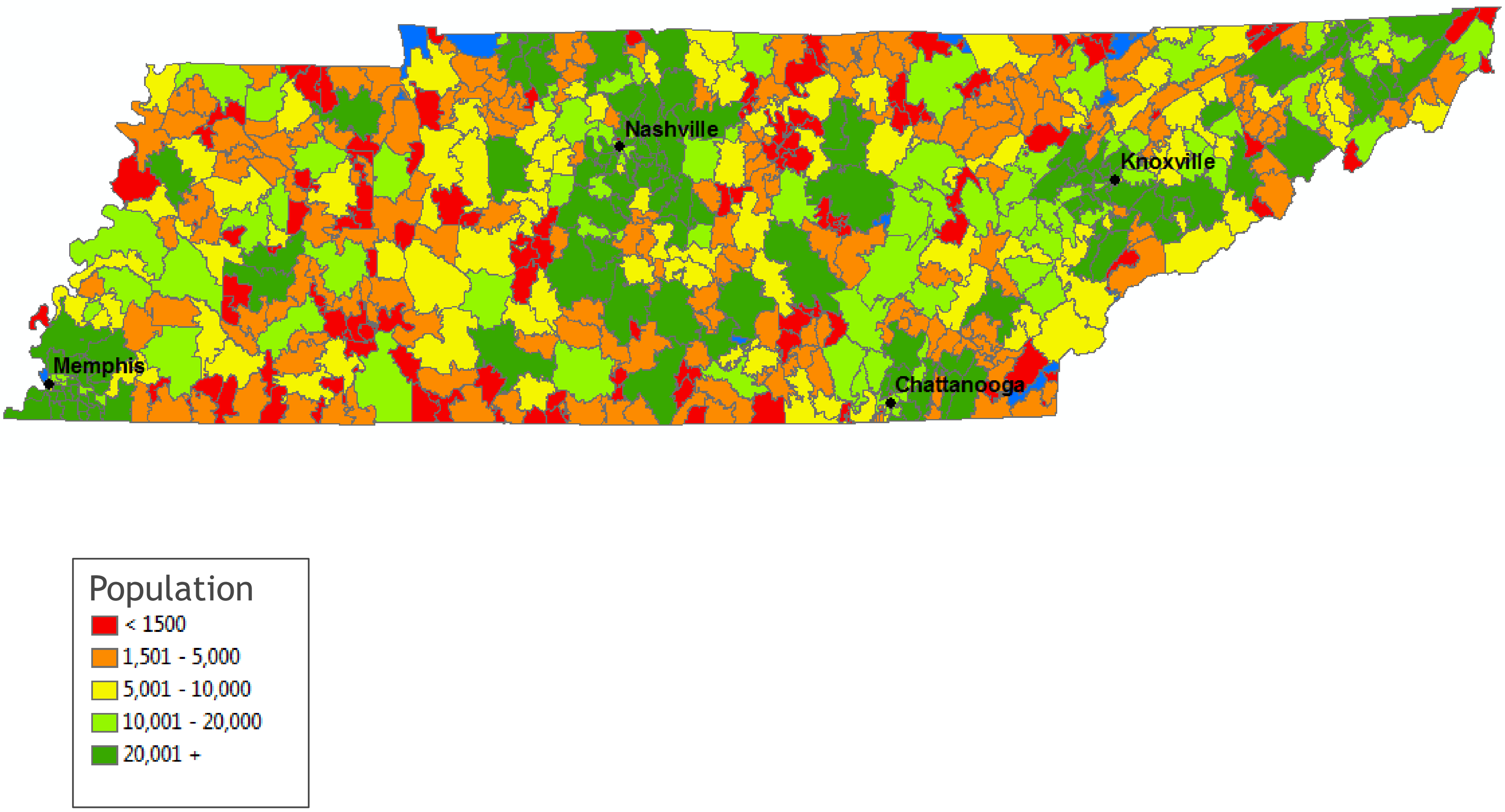
- Challenge: Data recipients often request reporting on more than one geography (e.g., both State and 3 digit Zip code).
- ***Subtraction Geography*** creates disclosure risk problems when more than one geography is reported for the same area and the geographies overlap.
- Also called ***geographical 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

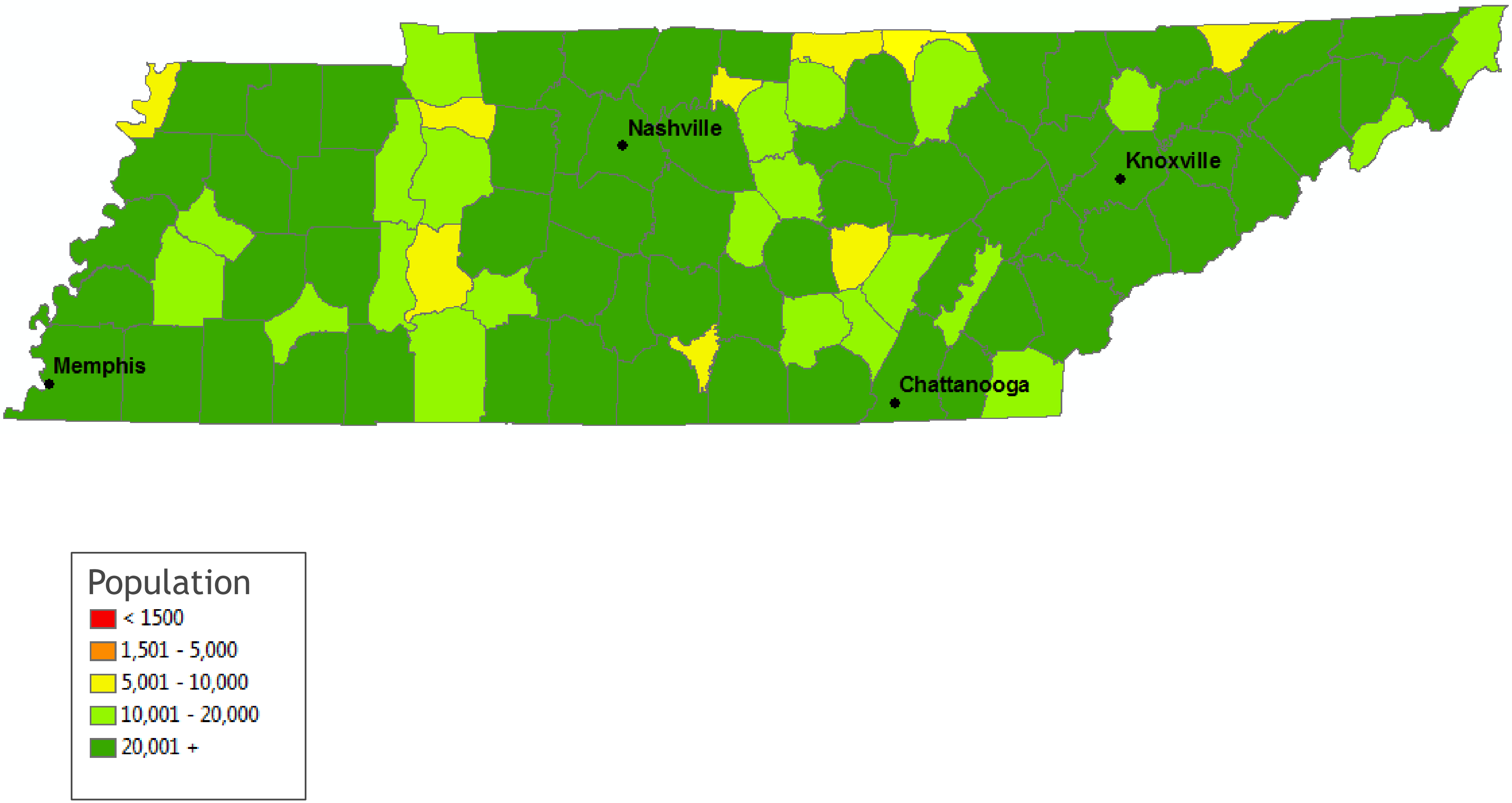
There are 7 CBSAs in Ohio which Cross into 4 Border States



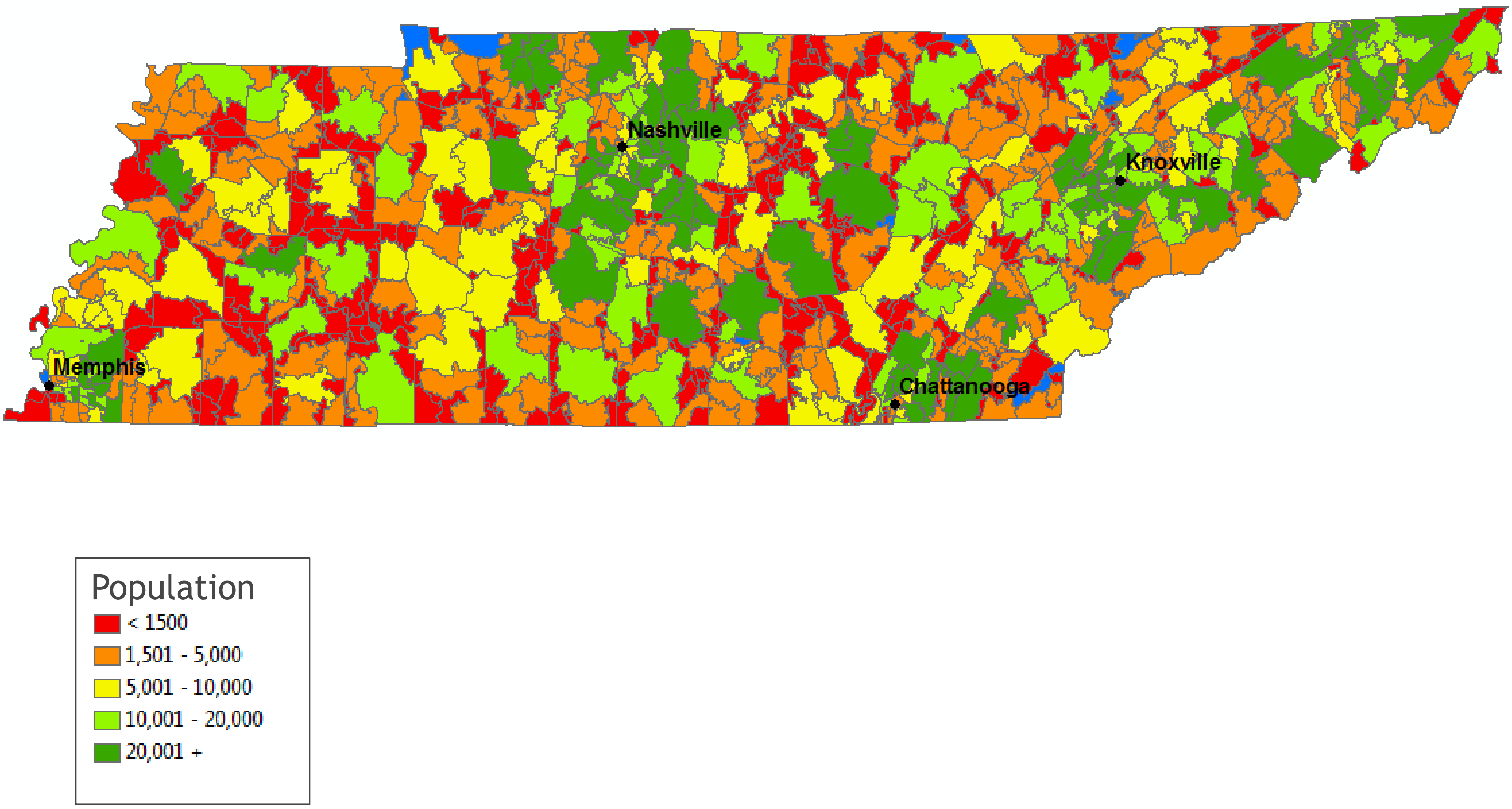
Tennessee – ZCTA5 Populations



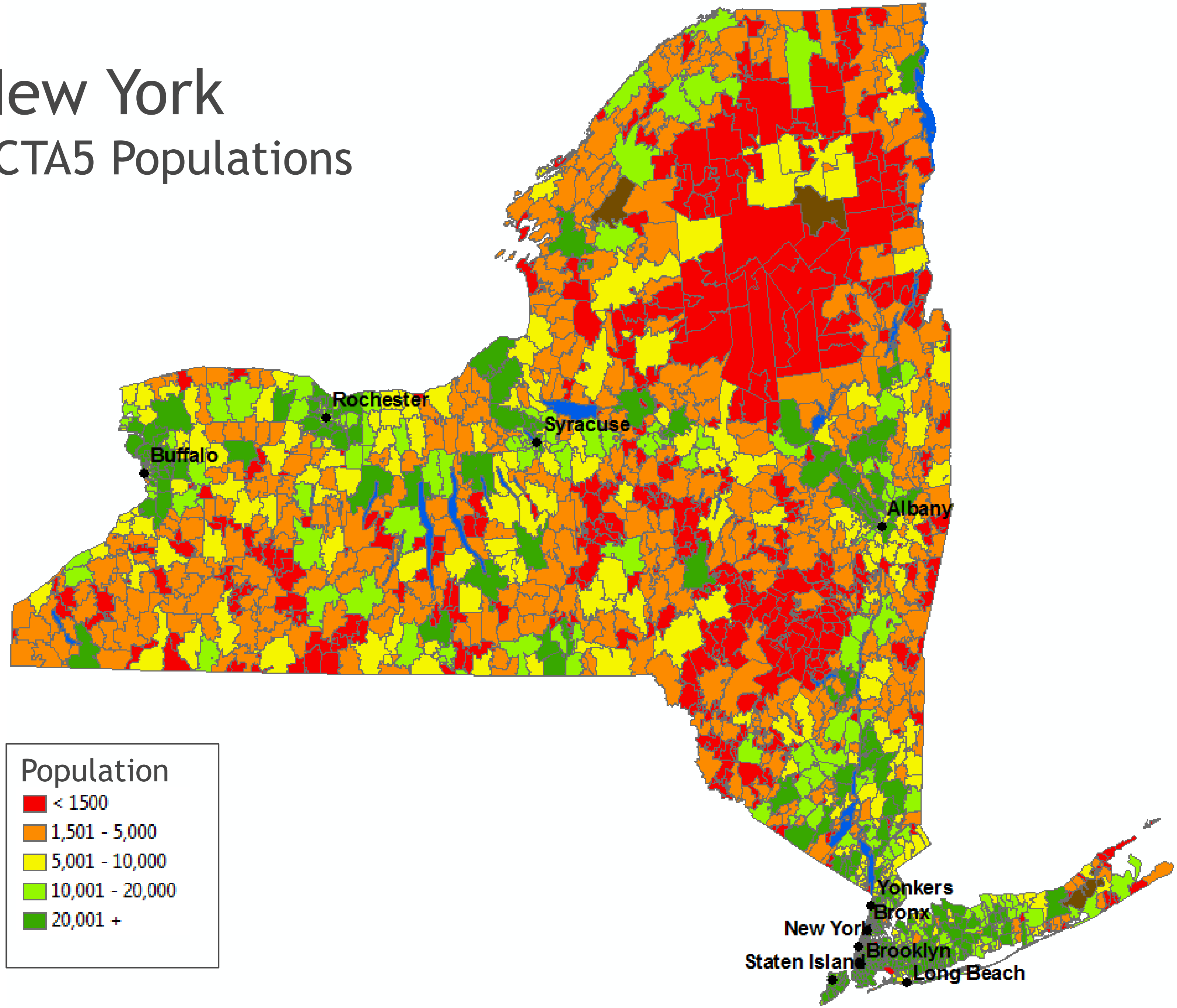
Tennessee – County Populations



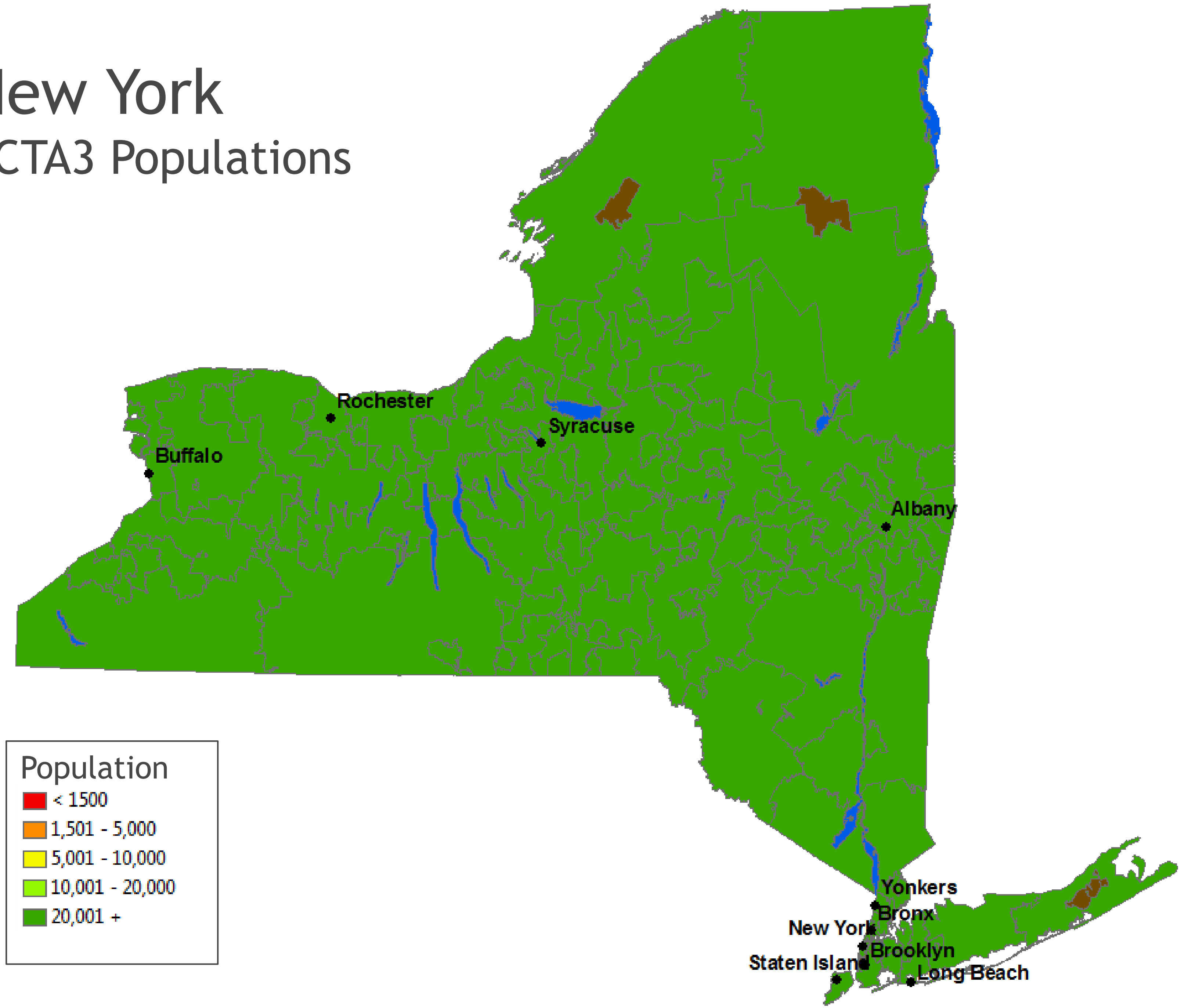
Tennessee – ZCTA5 X County Populations



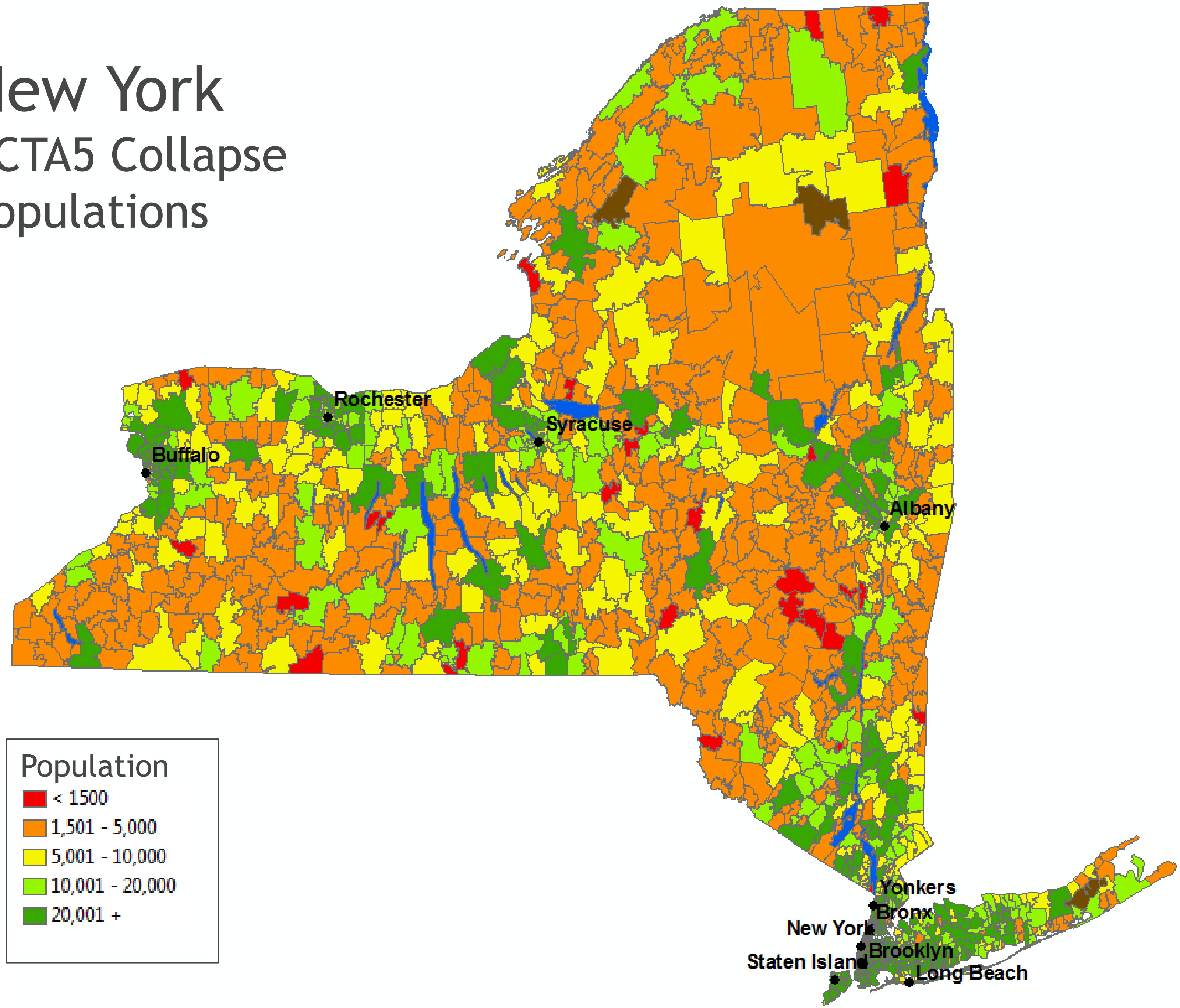
New York
ZCTA5 Populations



New York
ZCTA3 Populations



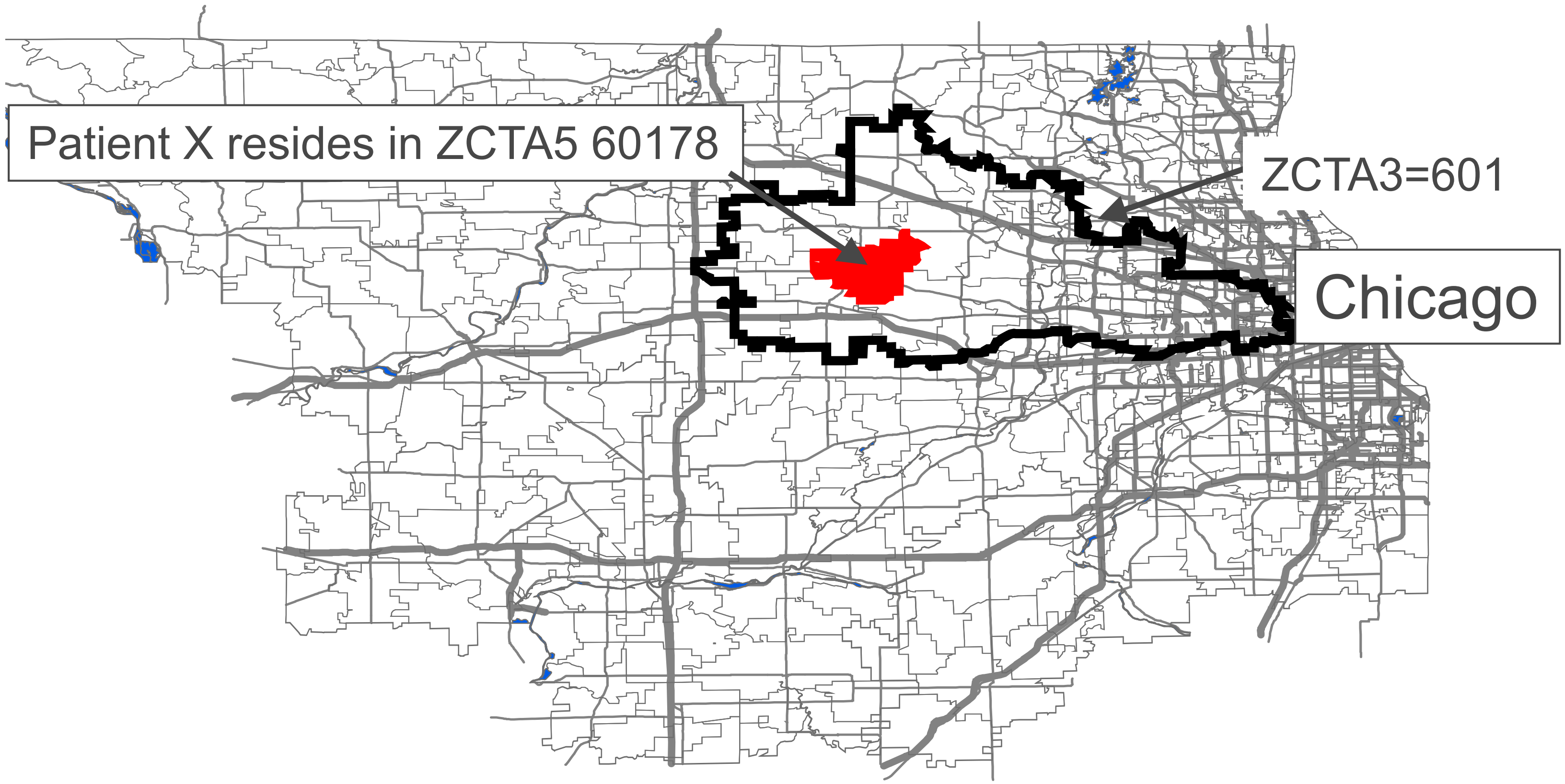
New York
ZCTA5 Collapse Populations



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.***

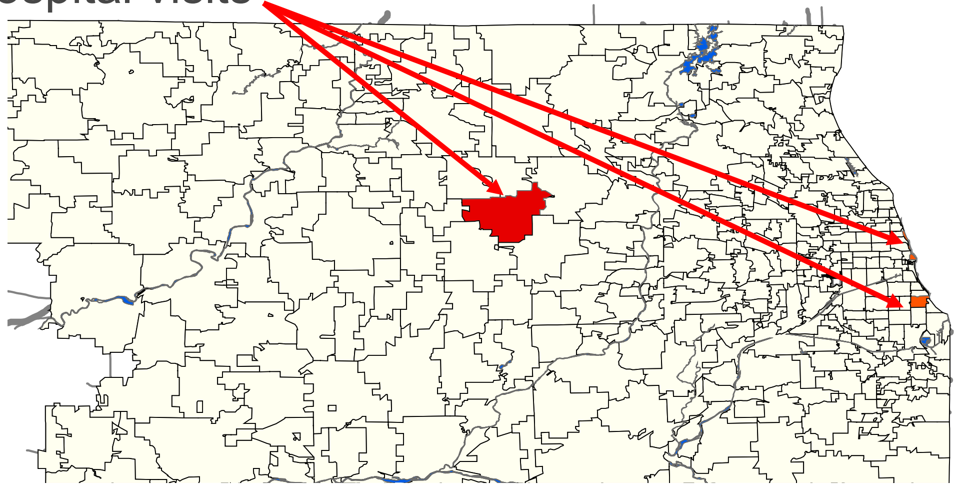
Challenge: Geoproxy Attacks



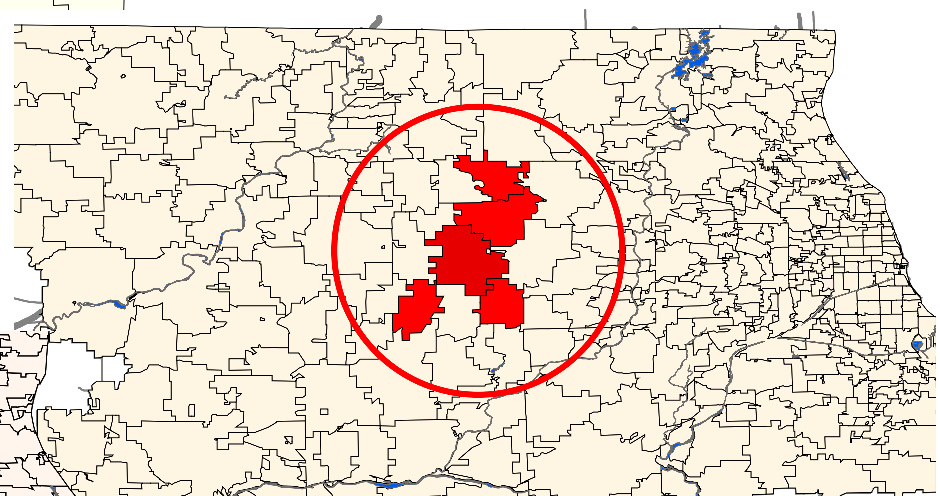
Example: Patient location as revealed within data set, but further narrowed to probable “hotspots” by using healthcare provider location data

Challenge: Geoproxy Attacks

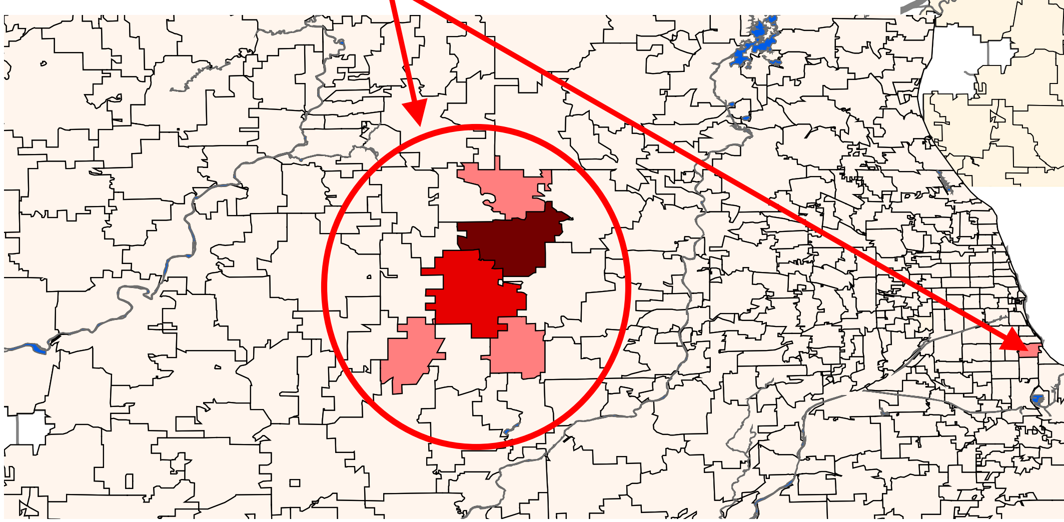
Hospital visits



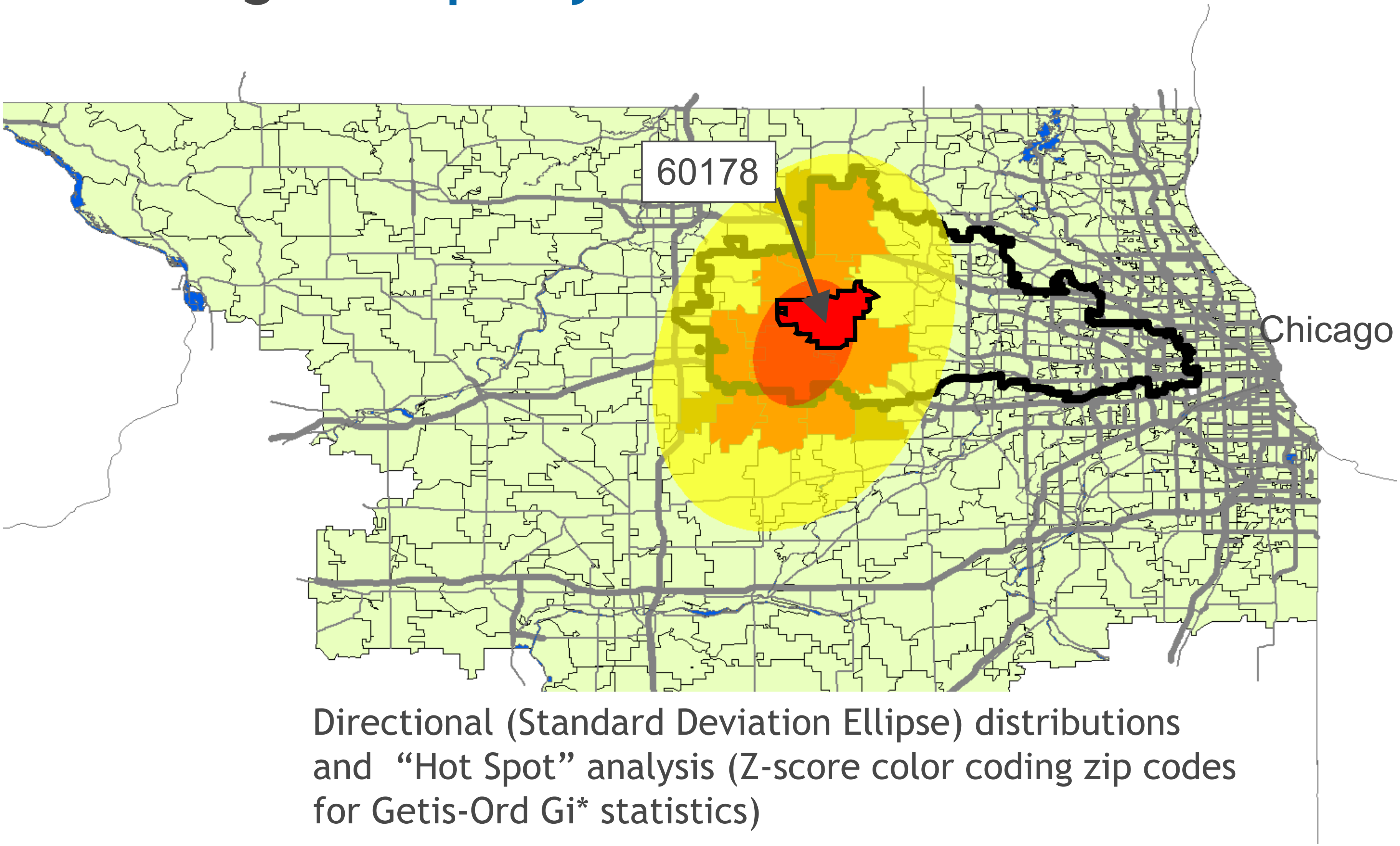
Outpatient/Office visits



Pharmacy visits



Challenge: Geoproxy Attacks



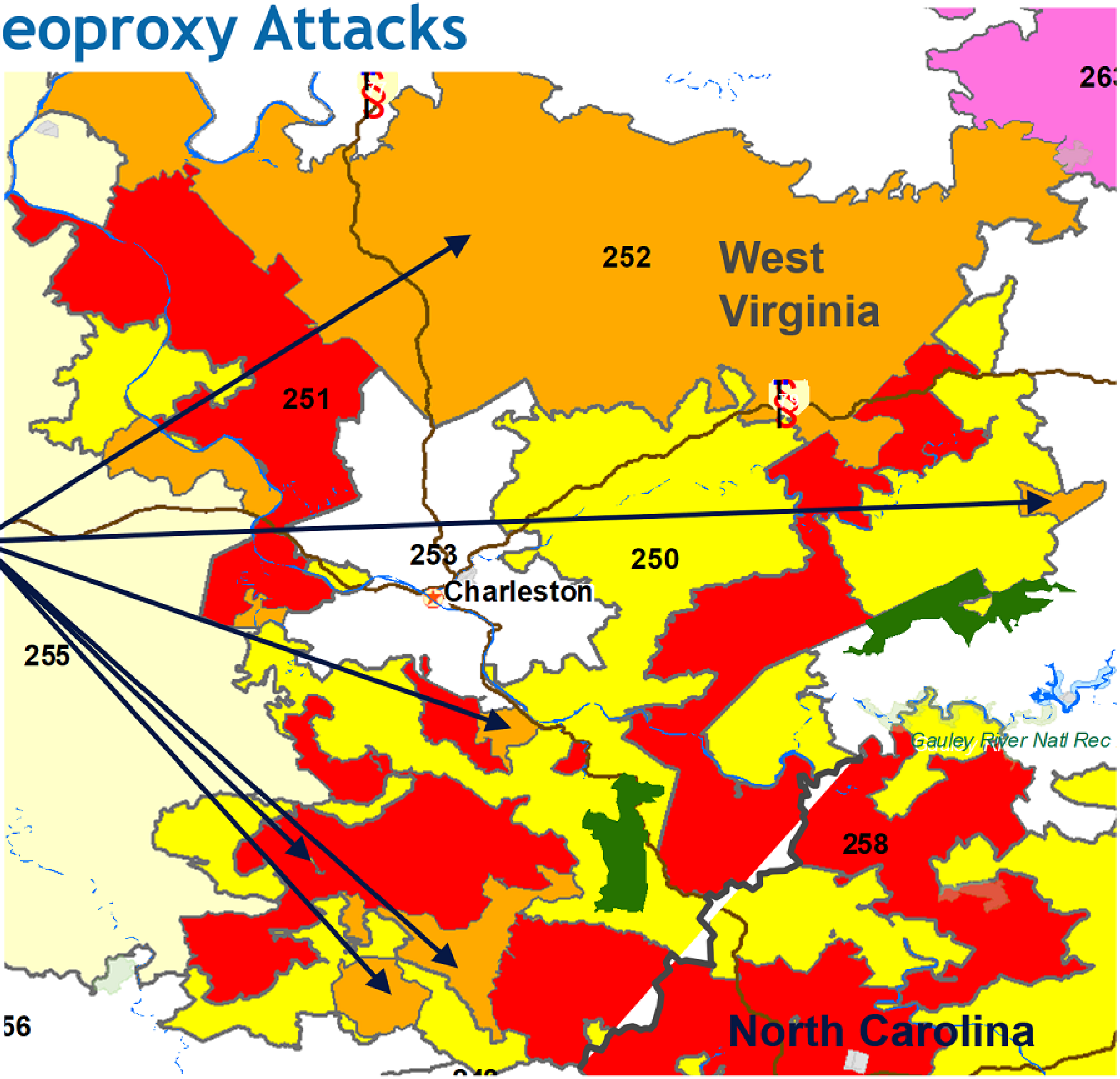
Directional (Standard Deviation Ellipse) distributions and "Hot Spot" analysis (Z-score color coding zip codes for Getis-Ord Gi* statistics)

Challenge: Geoproxy Attacks

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

ZCTA3 252 is highly dispersed

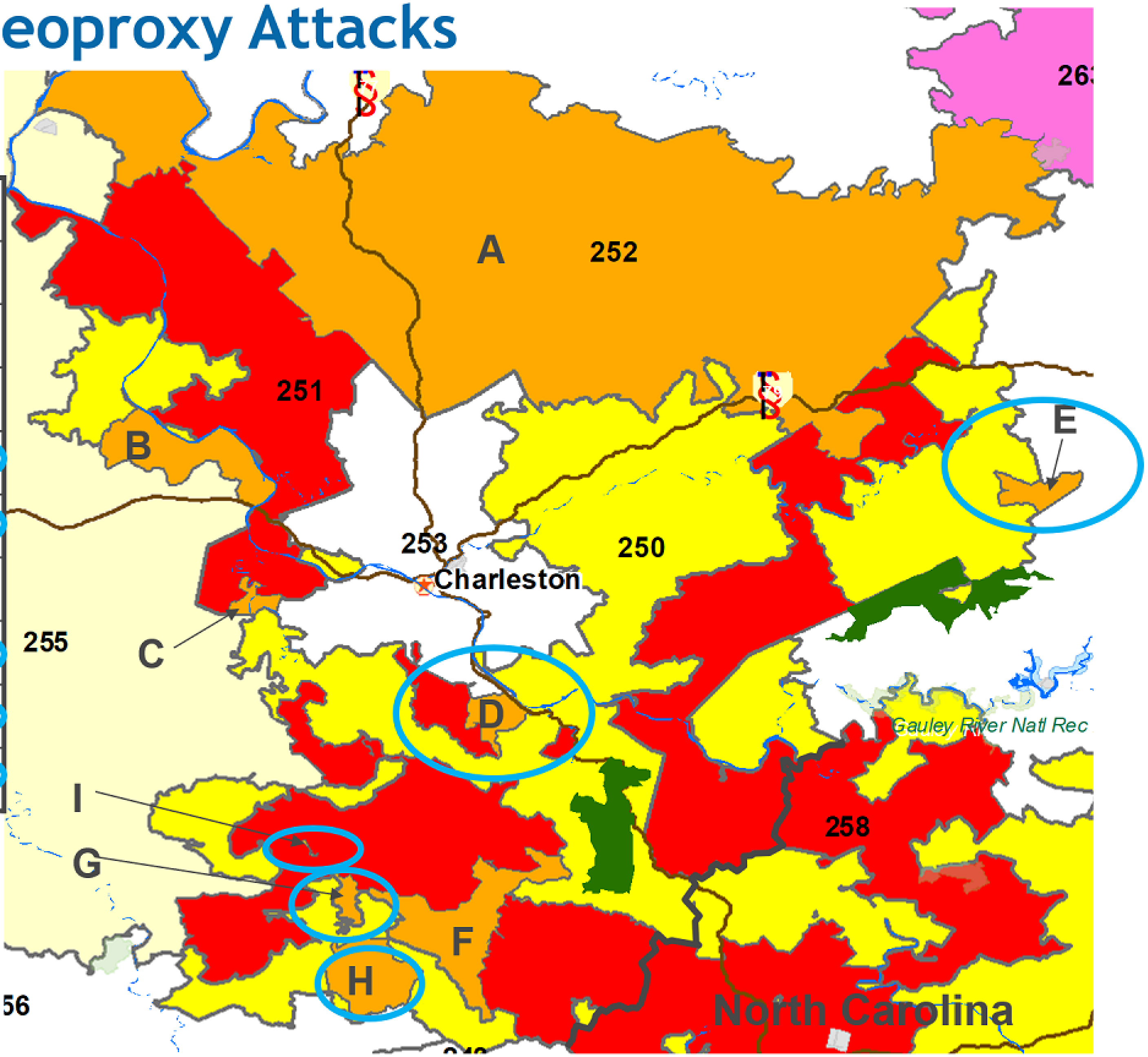
The complexity of 3-digit Zip Code Geography amplifies the threat of Geoproxy attacks



Challenge: Geoproxy Attacks

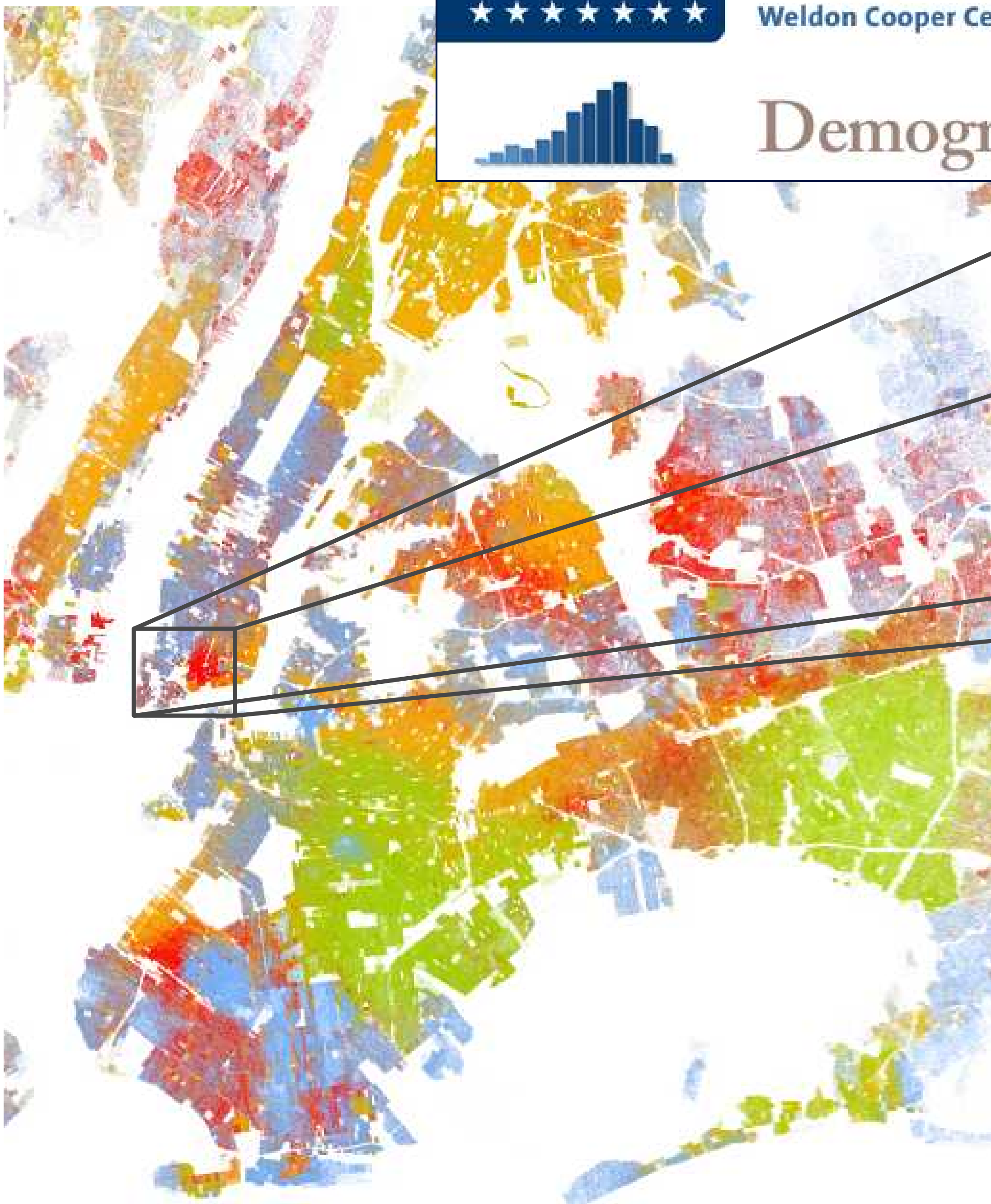
ZCTA3 252

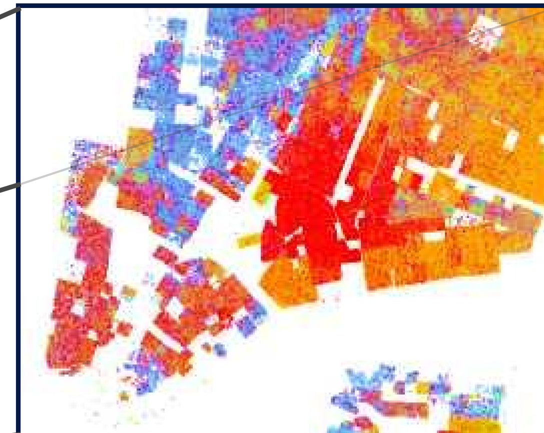
Area	Population
A	46,076
B	4,754
C	1,254
D	768
E	242
F	1,581
G	649
H	447
I	183



Weldon Cooper Center for Public Service • University of Virginia

Demographics Research Group





The Racial Dot Map

One Dot Per Person for the Entire United States

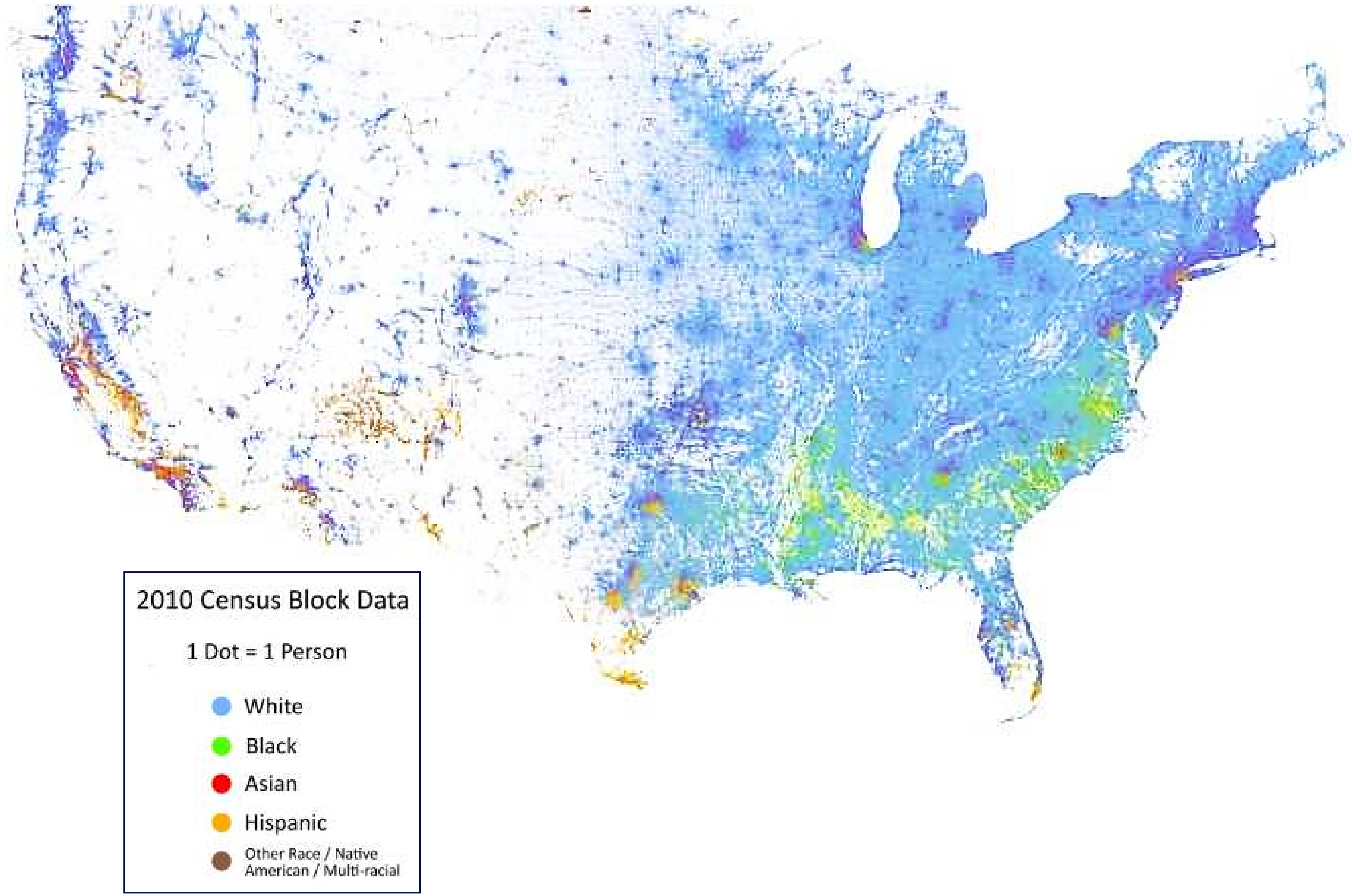
Created by Dustin Cable, July 2013

This is the most comprehensive map of race in America ever created.

[redacted]


8 MILE RD
South Woodward Avenue
8 Mile Road
8 Mile Road
8 Mile Road
8 Mile Road
Woodward Avenue
2010 Census Block Data
1 Dot = 1 Person
White
Black
Asian
Hispanic
Other Race / Native American / Multi-racial

http://demographics.coopercenter.org/DotMap/index.html



Quantitative Policy Analyses for De-identification Policy:

- De-identification policy is the subject of considerable controversy because it must balance important risks and benefits to individuals and societies and both sides of this question are subject to important uncertainties and competing values.
- Essential to recognize that complex social, psychological, economic and political motivations can underlie whether re-identification attempts are made.
- Quantitative Policy Analyses have been used for decades by many government agencies (EPA, Energy Dept.) to help address challenging policy decisions regarding difficult risk management questions.

Data Intrusion Scenarios:

- Prob(Re-identification) =
  Prob(Re-ident|Attempt)*Prob(Attempt)
- Note that Prob(Attempt) & Prob(Reident|Attempt) are actually not likely to be independent - higher re-identification probabilities are likely to increase re-identification attempts.
- Some very useful frameworks exist for characterizing Data Intrusion Scenarios:
  - Elliot & Dale, 1999, Duncan & Elliot Chapter 2, 2011
- We can frame the Prob(Attempt) in terms of: Motivation, Resources, Data Access, Attack Methods, Quasi-identifier Properties and Sets, Data Divergence Issues, and Probability of Success, Consequences and Alternatives for Goal Achievement

Conceptualizing Data Intrusion

- The information assumed about the Data Intruder's state of knowledge and resources is called a "Data Intrusion Scenario".
- We can't protect against every possible scenario, but we can protect against a realistic set of likely scenarios.
- For example, it may be reasonable to assume that there will be multiple data intruders each possessing different confidential knowledge.

Classifying Variables

—Identifying Variables

- Name, SSN, Address etc. (*Should already be removed from the sample data*)

—Key (or Quasi-identifier Variables)

- Variables that in combination can identify and are "*reasonably available*" in databases along with Identifying variables (e.g., Date of Birth, Gender, Zip Code)

—Confidential Variables

- Variables that the intruder might know about a specific target, but which would be very unlikely to be known in general (Hosp. Adm. Date, Diagnoses, etc.)

Conceptualizing Data Intrusion

- A reasonable assessment of statistical disclosure risks should include:
  - Formulating a comprehensive set of Data Intrusion Scenarios
  - Estimating (conservatively) the “costs and availability” of the required data intrusion resources
  - Conducting Statistical Disclosure Risk Analyses
  - Calculating the risk of disclosure given the associated costs, etc.
  - Providing a well-reasoned, clear and probablistically coherent justification for the case that the risk of identification is “very small” (under HIPAA Expert Determination.

Three Main Data Intrusion Scenarios:

- Specific-Target (aka "Nosy Neighbor") Attacks (Have specific target individuals in mind: acquaintances or celebrities)
- Marketing Attacks (Want as many re-identifications as possible in order to market to these individuals, may tolerate a high proportion of incorrect re-identifications, but this can come at the risk of being caught re-identifying)
- Demonstration Attacks (Want to demonstrate re-identification is possible to discredit the practice or to harm the data holder; Doesn't matter who is re-identified so unverified re-identifications may also achieve intended goals)

Data Intrusion Details:

- Motivation: To acquire specific information vs. Discredit/Harm de-identification policies or data holders
- Resources/Data Access: Statistical Skills; Knowledge/Data Access and Data Sources (Matters of Public Record, Commercially Available Data, Personal Knowledge); Computing Skills & Resources; Impediments provided by Computer Security and Governance/Legal controls.
- Attack Methods: Primary Intrusion Scenarios (Specific Target, Marketing, Demonstration), Deterministic vs. Probabilistic matching, Multi-stage Linkage attacks with or without verifications steps.

Data Intrusion Details:

- Quasi-identifier Properties and Sets
  - Key Resolution
  - Skewness
  - Associations between Quasi-identifiers & “Special Unique” Interactions for Combinations of Quasi-identifiers
- Data Divergence Issues
  - Missing Data Rates
    - The “Myth of the Perfect Population Register”
  - Time Dynamic Variables
  - Measurement and Coding Variations and Errors

Importance of “Data Divergence”

- Probabilistic record linkage has some capacity deal with errors and inconsistencies in the linking data between the sample and the population caused by “**data divergence**”:
  - Time dynamics in the variables (e.g. changing Zip Codes when individuals move, Change in Martial Status, Income Levels, etc.),
  - Missing and Incomplete data and
  - Keystroke or other coding errors in either dataset,
- But the links created by probabilistic record linkage are subject to uncertainty. The data intruder is never really certain that the correct persons have been re-identified.

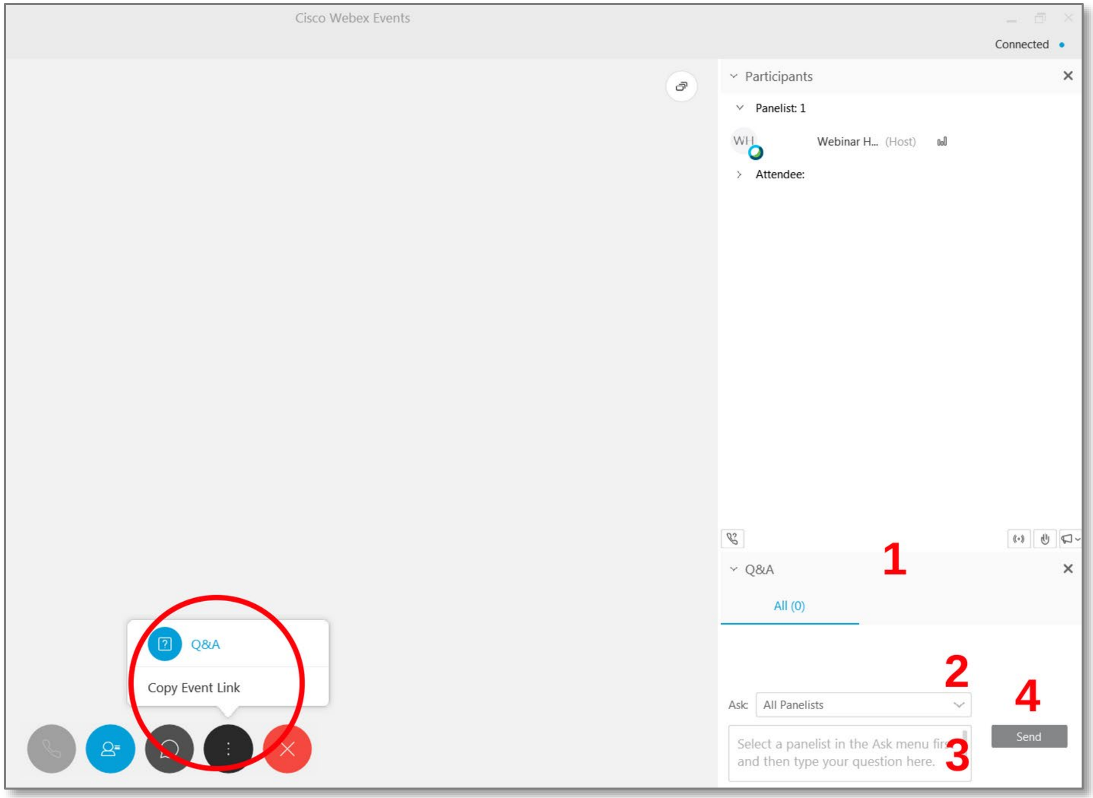
Data Intrusion Details:

- Probability of:
  - Success (Not only information from verifiable re-identifications or economic gains, but also success in terms of desired policy or organizational harm goals)
  - Consequences for Re-identification Attempts (Legal and/or Economic Ramifications for Re-identification Attempts)
- Alternatives for Goal Achievement
  - Are there preferable alternatives for data intruder's goal achievement that have more cost-effective economic incentives or avoid negative consequences of re-identification attempts?

The Network for Public Health Law

How to Use WebEx Q & A

1. Open the Q&A panel
2. Select "All Panelists"
3. Type your question
4. Click "Send"



Thank you for attending

For a recording of this webinar and information about future webinars, please visit networkforphl.org/webinars

2019 Public Health Law Summit
Data Sharing to Improve Community Health
October 3-4 | Plymouth, MI

Measles Outbreak – Public Health Authority, New York City's Immunization Mandate, and the Current Legislative Landscape
June 4, 1:00 – 2:30 p.m. EST

ASLME
AMERICAN SOCIETY OF LAW, MEDICINE & ETHICS

You may qualify for CLE credit. All webinar attendees will receive an email from ASLME, an approved provider of continuing legal education credits, with information on applying for CLE credit for this webinar.