Free the Data,
Manage the Risks

Sponsored by California HealthCare Foundation

Moderator: Andy Krackov
Senior Program Officer, Market and Policy Monitor Program, CHCF

Khaled El Emam, *CEO*, Privacy Analytics

Daniel Barth-Jones, *Assistant Professor*, Columbia University

Barbara Rudolph, *Consultant*, NAHDO

California HealthCare Foundation

Free the Data
Manage the Risks

Khaled El Emam (PhD)

NAHDO 29th Annual Conference
9th October 2014

www.privacyanalytics.ca | 855.686.4781
info@privacyanalytics.ca

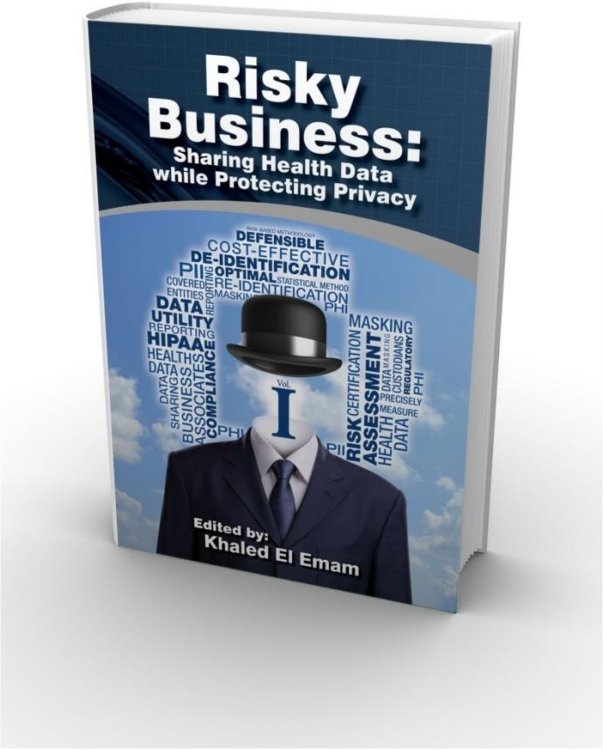
251 Laurier Avenue W, Suite 200
Ottawa, ON Canada K1P 5J6

- It is possible to share data and:
  - Protect privacy
  - Meet regulatory and legal requirements
  - Get high quality data

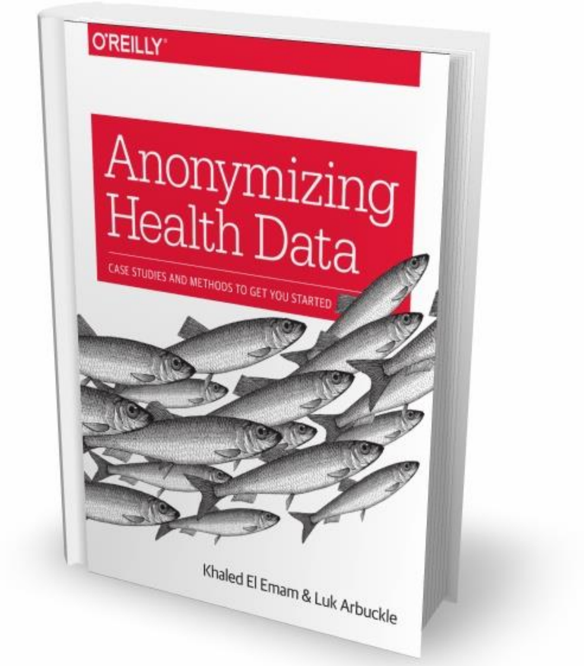
PRIVACYANALYTICS
Data Anonymization Solutions

There are some standards and more detailed (operational) standards are in development for de-identification

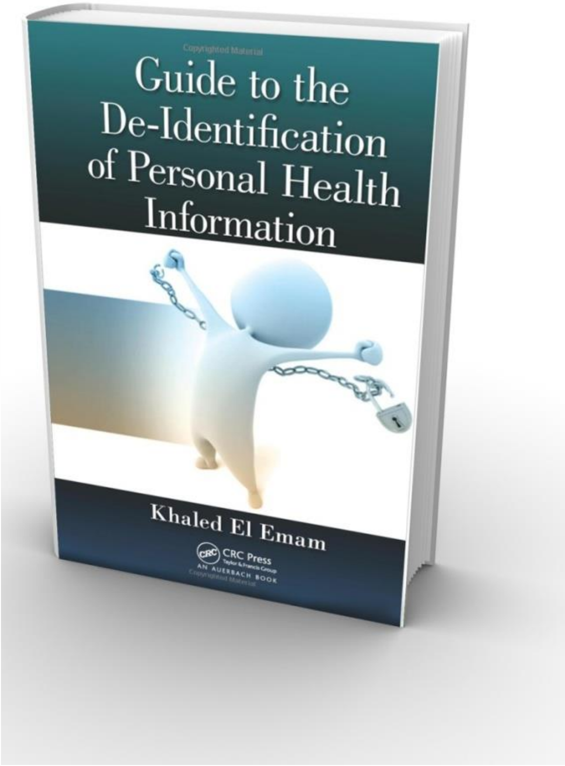
Resources



Risky Business:
Sharing Health Data while Protecting Privacy
Edited by:
Khaled El Emam



O'REILLY
Anonymizing Health Data
CASE STUDIES AND METHODS TO GET YOU STARTED
Khaled El Emam & Luk Arbuckle



Guide to the De-Identification of Personal Health Information
Khaled El Emam
CRC Press

PRIVACYANALYTICS
Data Anonymization Solutions

- There are different levels of data release, and the de-identification scheme needs to match that:
  - Public
  - Quasi-public
  - Non-public

PRIVACYANALYTICS
Data Anonymization Solutions

De-identification is a risk management exercise that takes into account the context of the data release

PRIVACYANALYTICS
Data Anonymization Solutions

There are many precedents on what is acceptable risk – this is not something we should be debating (there are bigger issues)

More sophisticated methods (computational and statistical methods) are needed to be able to generate high quality data and meet the risk thresholds

There is a real need for multi-disciplinary education and certification around de-identification practices to start building a community of practice

PRIVACYANALYTICS
Data Anonymization Solutions

The discourse on re-identification attacks is of deep concern from a scientific, ethical, and integrity perspective

PRIVACYANALYTICS
Data Anonymization Solutions

Our biggest challenge is transitioning good de-identification methodologies into practice

PRIVACYANALYTICS
Data Anonymization Solutions

The existing legal framework is fine as it is and is quite robust – we can do a lot within it to free the data

Contact

kelemam@privacyanalytics.ca

@kelemam

www.privacyanalytics.ca

PRIVACYANALYTICS
Data Anonymization Solutions

NAHDO 29th Annual Conference

Oct 8-9 2014

Free the Data, Manage the Risks: Reality Check for Data Privacy Risks

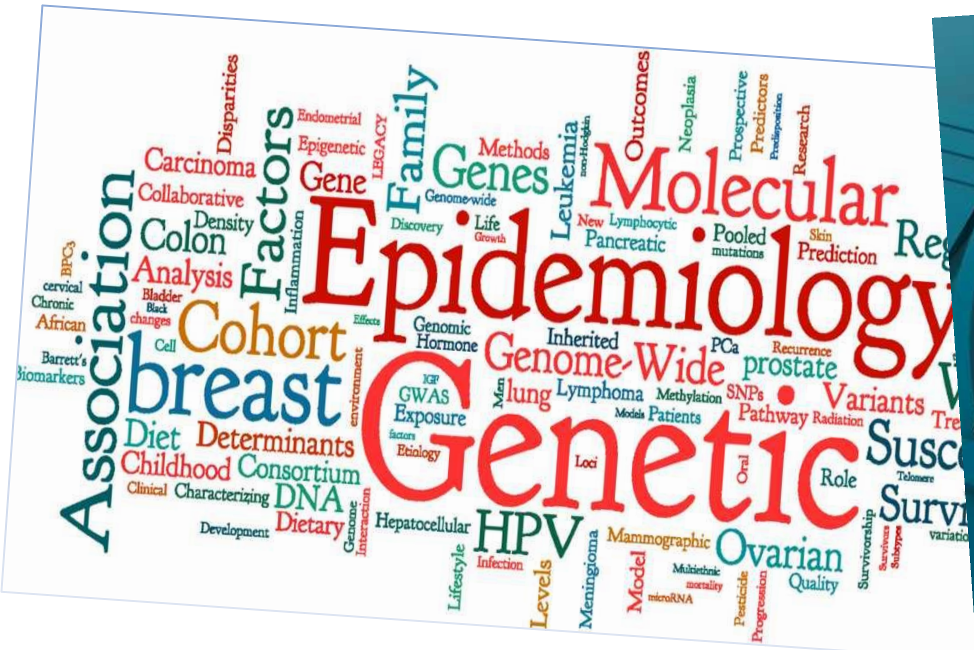
Daniel C. Barth-Jones, M.P.H., Ph.D

Assistant Professor of Clinical Epidemiology,
Mailman School of Public Health
Columbia University

E-mail: *db2431@columbia.edu*

On Twitter : @dbarthjones

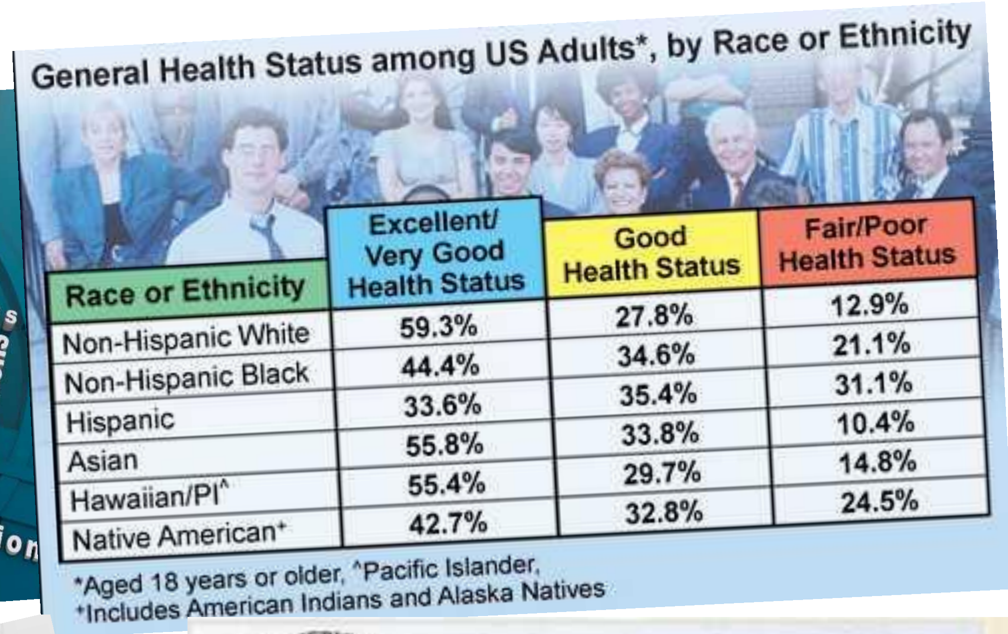
A Historic and Important Societal Debate is underway…



Public Policy Collision Course

The Societal Value of De-identified Data

- Properly de-identified health data is an ***invaluable "public good"***. *The broad availability of de-identified data is an essential tool for society supporting scientific innovation and health system improvement and efficiency.*
- De-identified data does and can serve as the engine driving forward innumerable essential health systems improvements: quality improvement, health systems planning, healthcare fraud, waste and abuse detection, and medical/public health research (e.g. comparative effectiveness research, adverse drug event monitoring, patient safety improvements and reducing health disparities).
- De-identified health data greatly benefits our society and provides strong privacy protections for the individuals. As the promise of EHRs and Health IT yields richer de-identified clinical data, the progress of our nation's healthcare reform will likely be built on a foundation of such de-identified health data.


General Health Status among US Adults*, by Race or Ethnicity

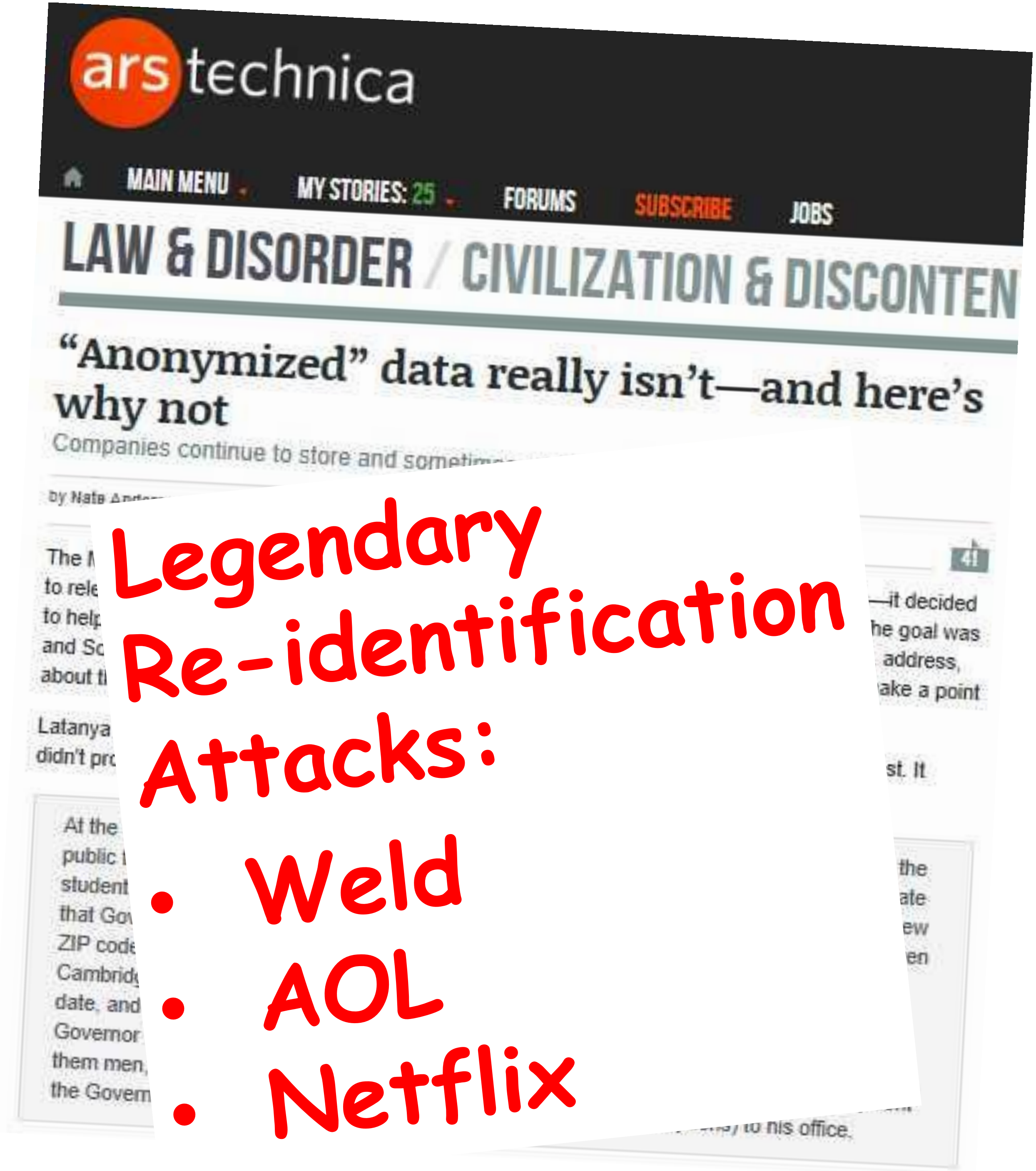
Race or Ethnicity	Excellent/ Very Good Health Status	Good Health Status	Fair/Poor Health Status
Non-Hispanic White	59.3%	27.8%	12.9%
Non-Hispanic Black	44.4%	34.6%	21.1%
Hispanic	33.6%	35.4%	31.1%
Asian	55.8%	33.8%	10.4%
Hawaiian/PI^	55.4%	29.7%	14.8%
Native American+	42.7%	32.8%	24.5%

*Aged 18 years or older, ^Pacific Islander,
+Includes American Indians and Alaska Natives

Counting and Tabulating is Essential to Public Health and Population Science…

- The foundational acts of counting and tallying individual characteristics underlie our ability to analyze the population distributions and determinants of disease—which is essential to medical and population health science.
- But some risk of re-identification exists with every characteristic that we collect and report.
- Thus, the important ongoing debate about health data de-identification and the ethical and public policy implications for research conducted with de-identified data.

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



Bloomberg Our Company | Professional | Anywhere

HOME QUICK NEWS OPINION

Hospitals in the U.S. pledge to keep a patient's health background confidential. Yet states from Washington to New York are putting privacy at risk by selling records that can be used to link a person's identity to medical conditions using public information.

Frustrated Republic
Pressure Boehner
Shutdown

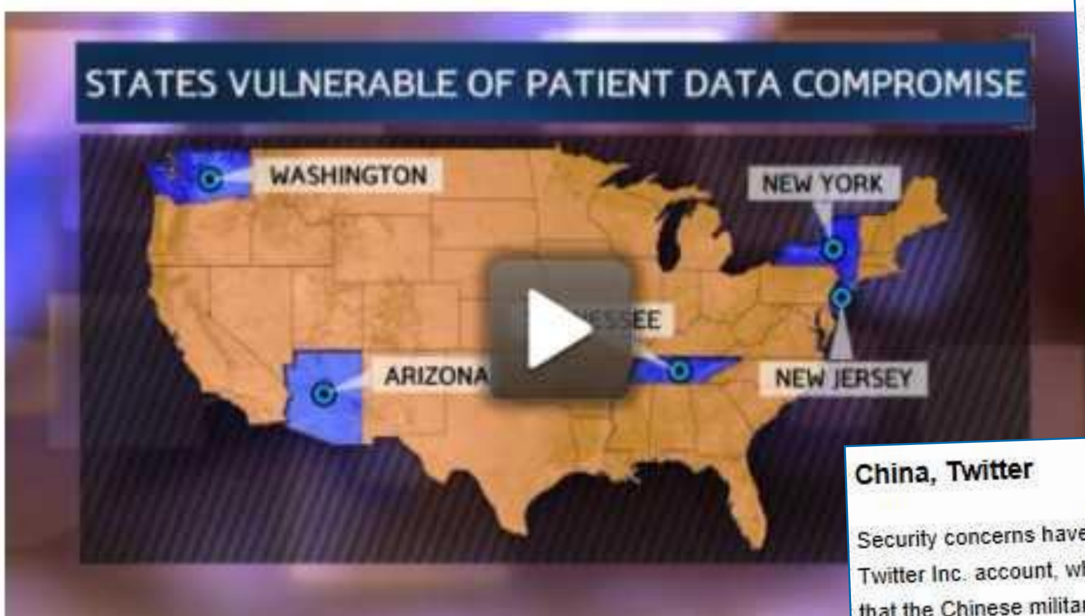
BREAKING NEWS Telecom Italia Ceo Bernabe Is Said to Resign

States' Hospital Data for Sale Puts Privacy in Jeopardy

40 / 648,384 = 1/16,200

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."

China, Twitter

Security concerns have been heightened recently by the breach of the Associated Press Twitter Inc. account, which resulted in a temporary stock-market decline, U.S. accusations that the Chinese military is engaged in a cyber espionage campaign and attacks on financial

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%?, 27%? 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 more recently estimated at 0.04% (4 in 10,000) (Sweeney, NCVHS Testimony, 2007)
  - ***Safe Harbor*** de-identification provides protections that have been estimated to be a minimum of ***400 to 1000 times more protective*** of privacy than permitting direct PHI access.

    (Benitez & Malin, JAMIA, 2010)
- Reality: Under HIPAA de-identification requirements, re-identification is expensive and time-consuming to conduct, requires serious computer/mathematical skills, is rarely successful, and 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*ied”*… (There is no guarantee that de-identified data will remain de-identified regardless of what you do to it after it is de-identified.)
  - Simply collapsing your coding categories until the data is “k-anonymous” can make the data unsuitable for many statistical analyses

Myth of the "Perfect Population Register"

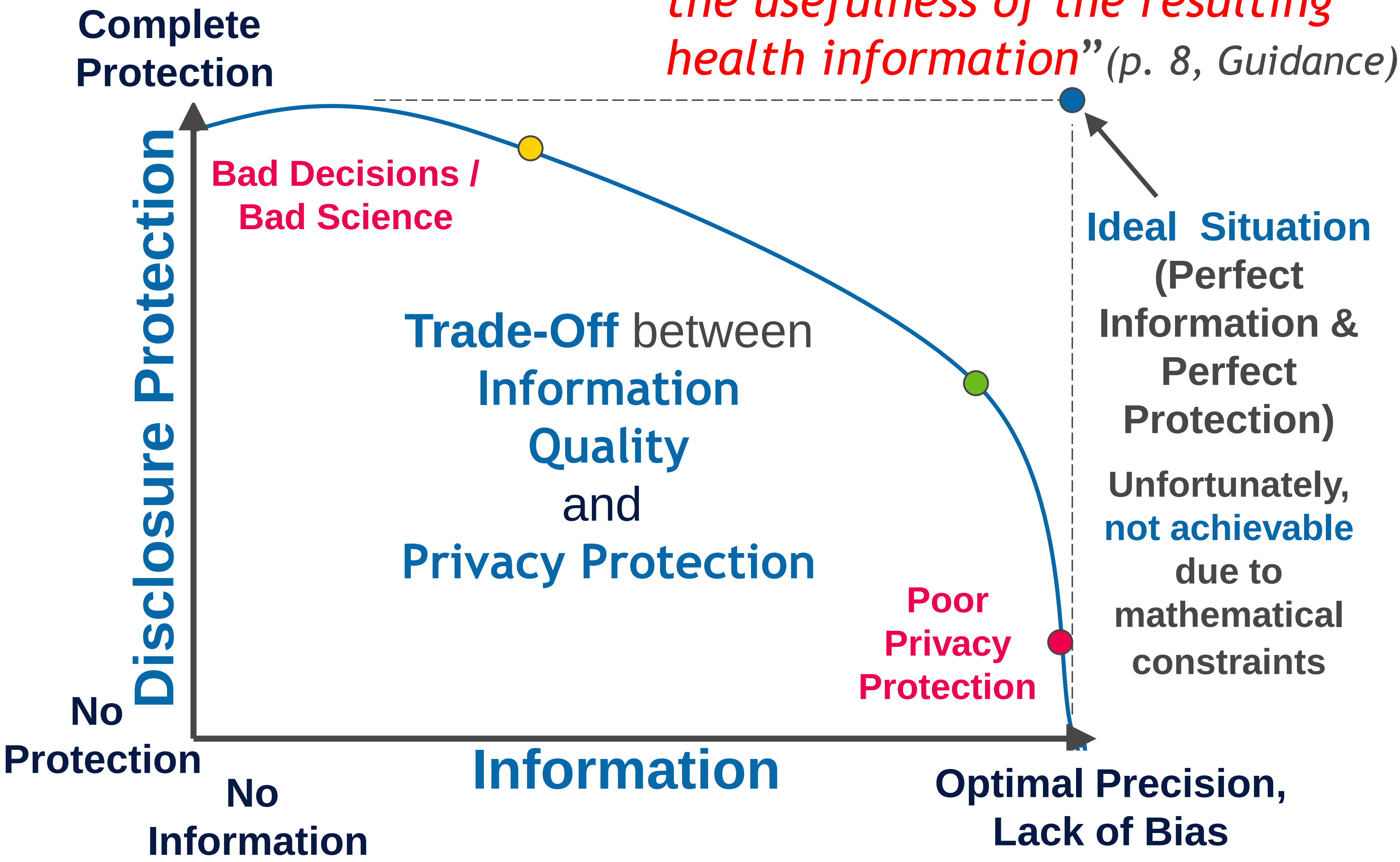
- The critical part of many re-identification efforts that is often assumed by disclosure scientists is the ***assumption of a perfect population register***.
- All Population registers will have data errors and be incomplete to some extent. (e.g. Nationwide voter registration levels typically are about 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.

Importance of "Data Divergence"

- Errors and inconsistencies in the linking data between the sample and the population create "**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 even probabilistic record linkage methods, which can help address such challenges, are subject to uncertainty. The data intruder is never really certain that the correct persons have been re-identified.
- The recent Personal Genome Project re-identification attack using {Zip5, Gender and DoB} was able to achieve only a 27% re-identification rate (not 87%) due to these issues.

The Inconvenient Truth:

"*De-identification leads to information loss which may limit the usefulness of the resulting health information*" *(p. 8, Guidance)*



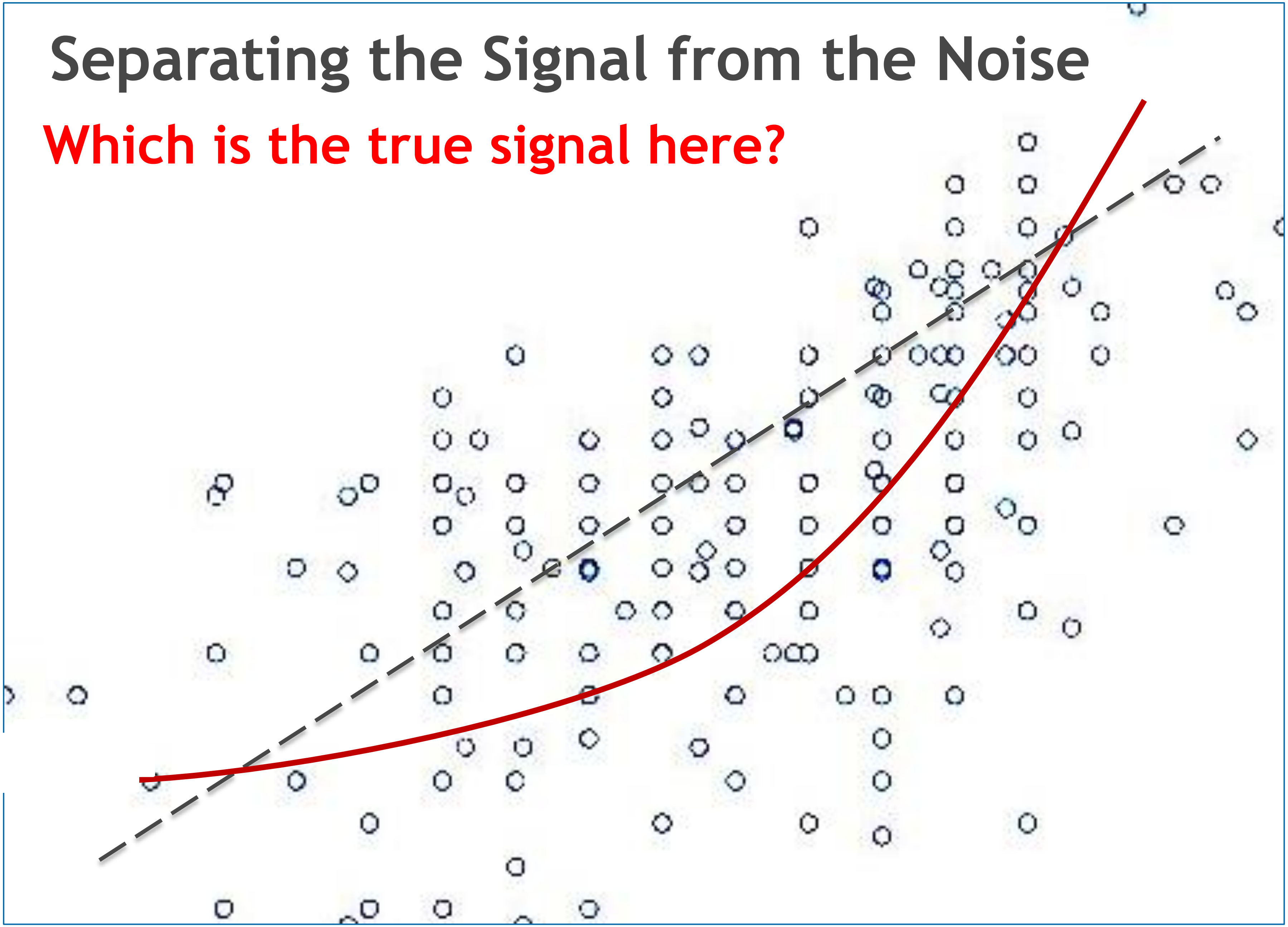
Balancing Disclosure Risk/Statistical Accuracy

- 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

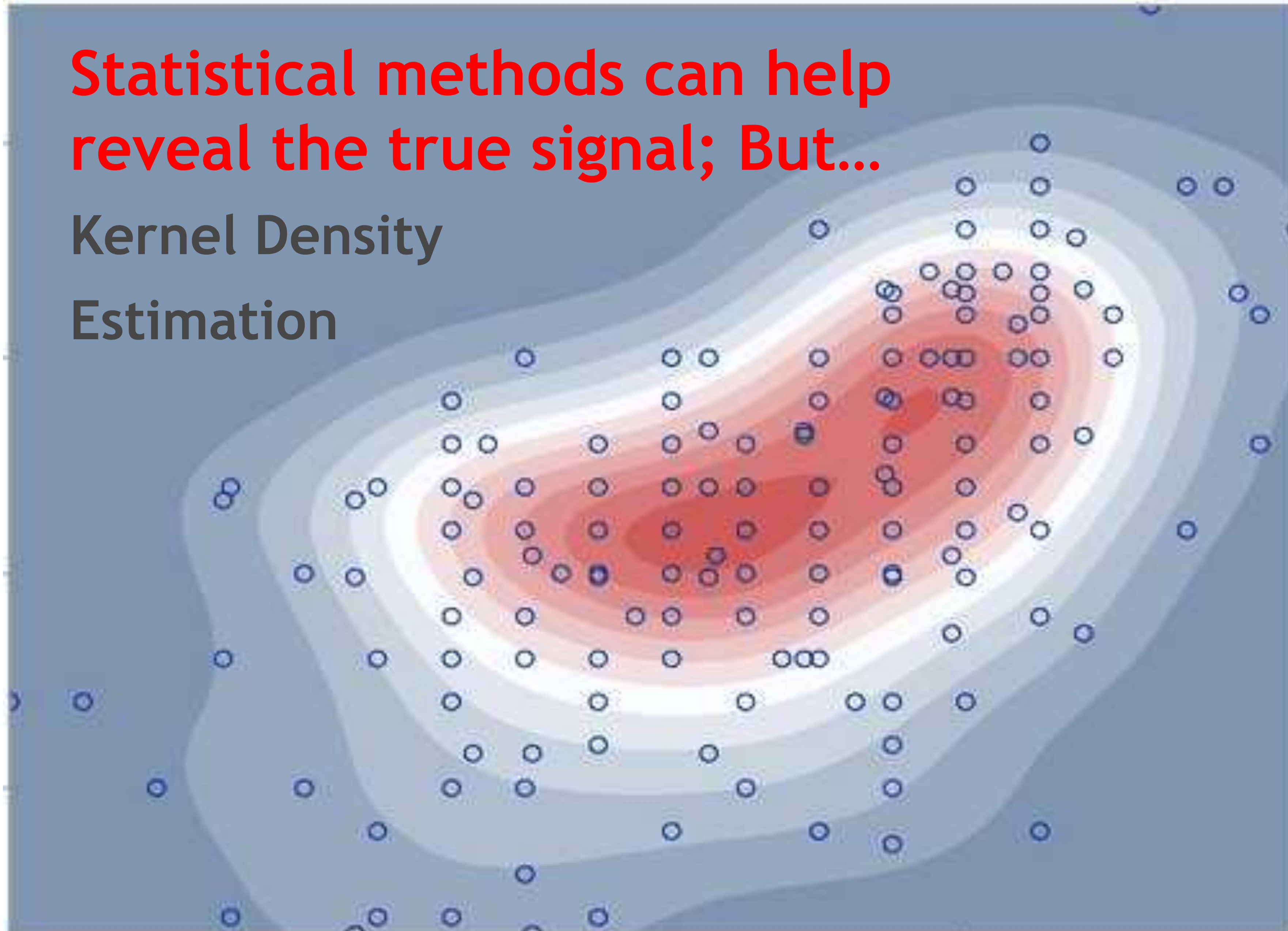
Separating the Signal from the Noise

Which is the true signal here?

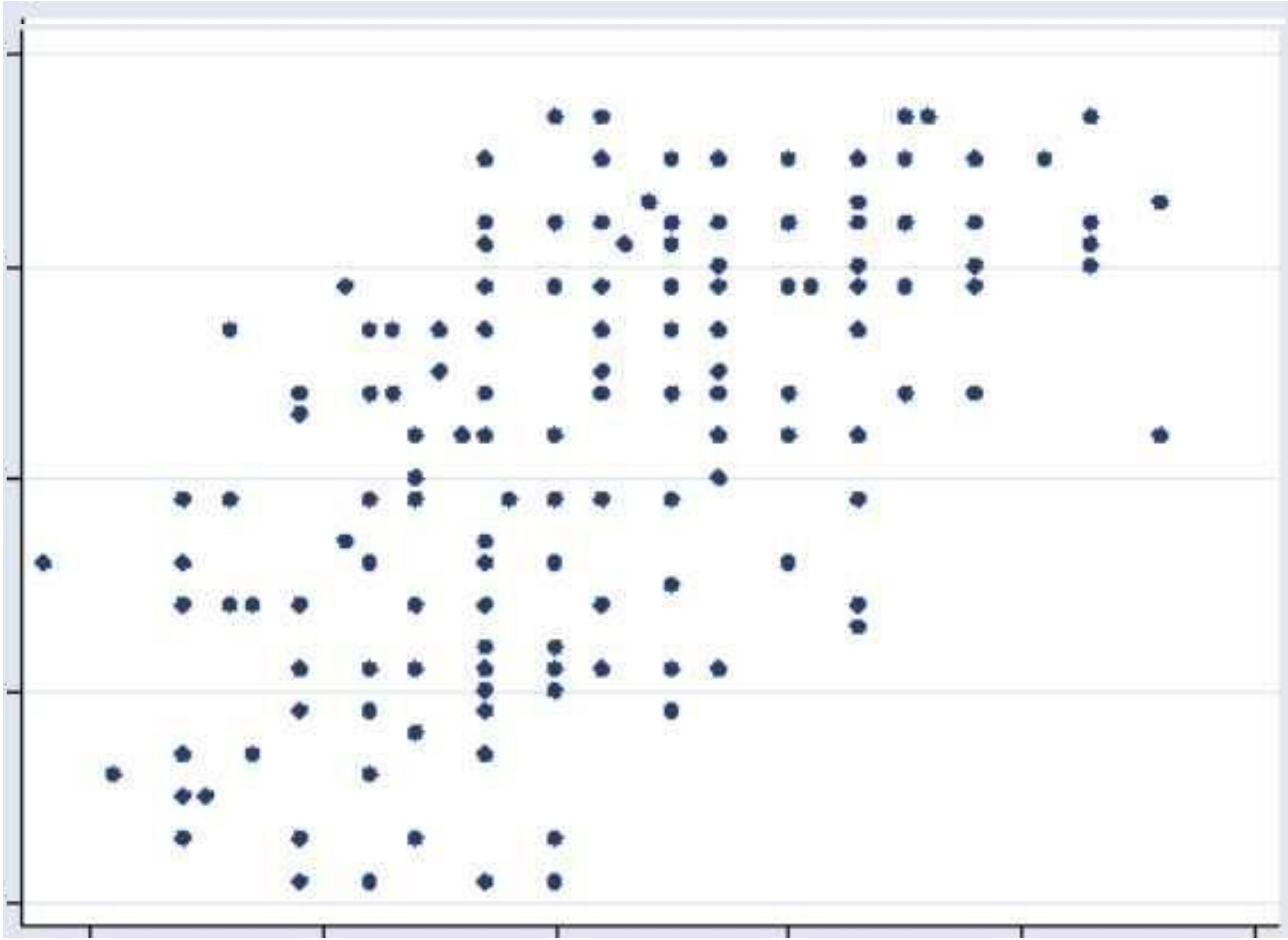


Statistical methods can help reveal the true signal; But…

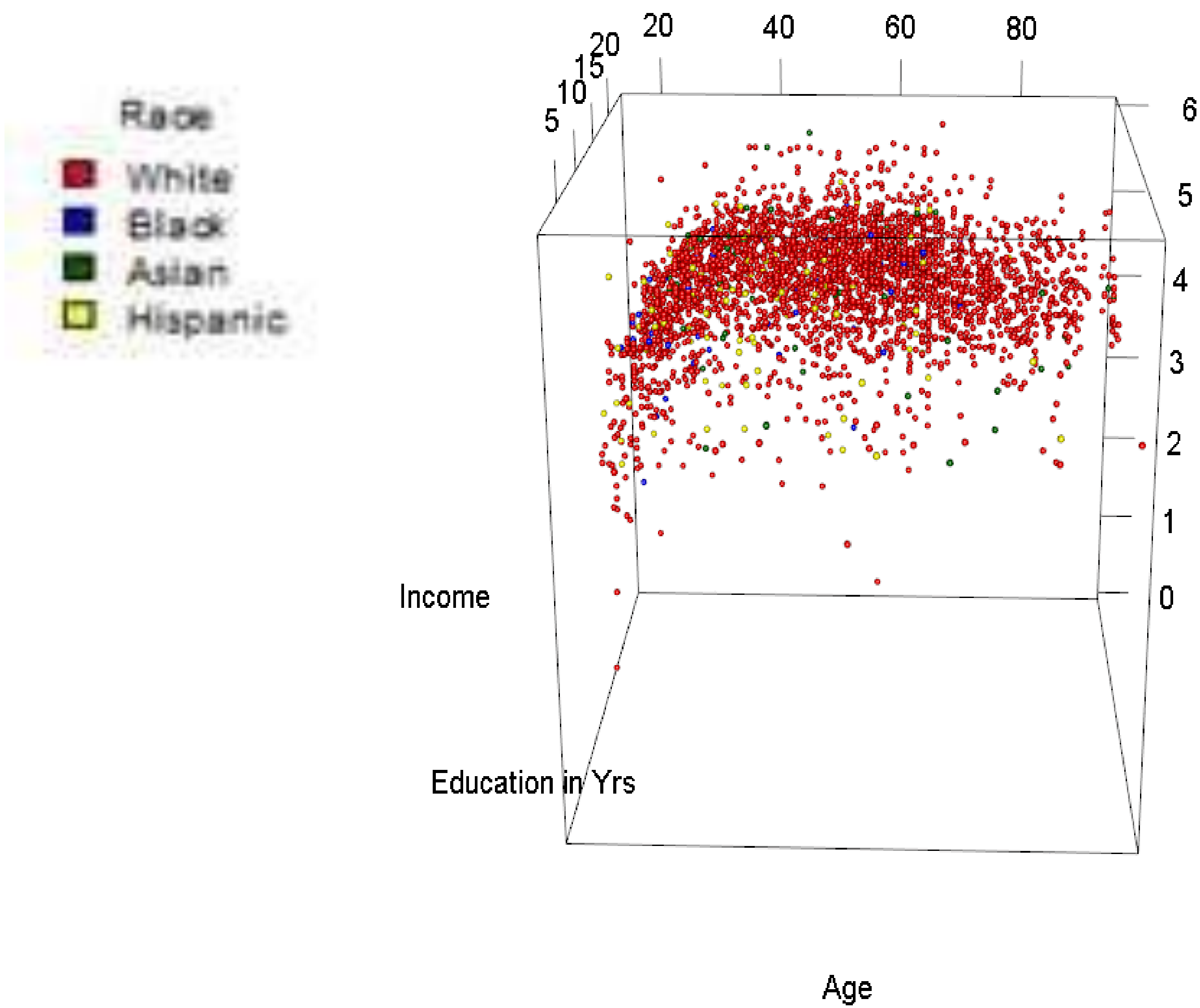
Kernel Density Estimation

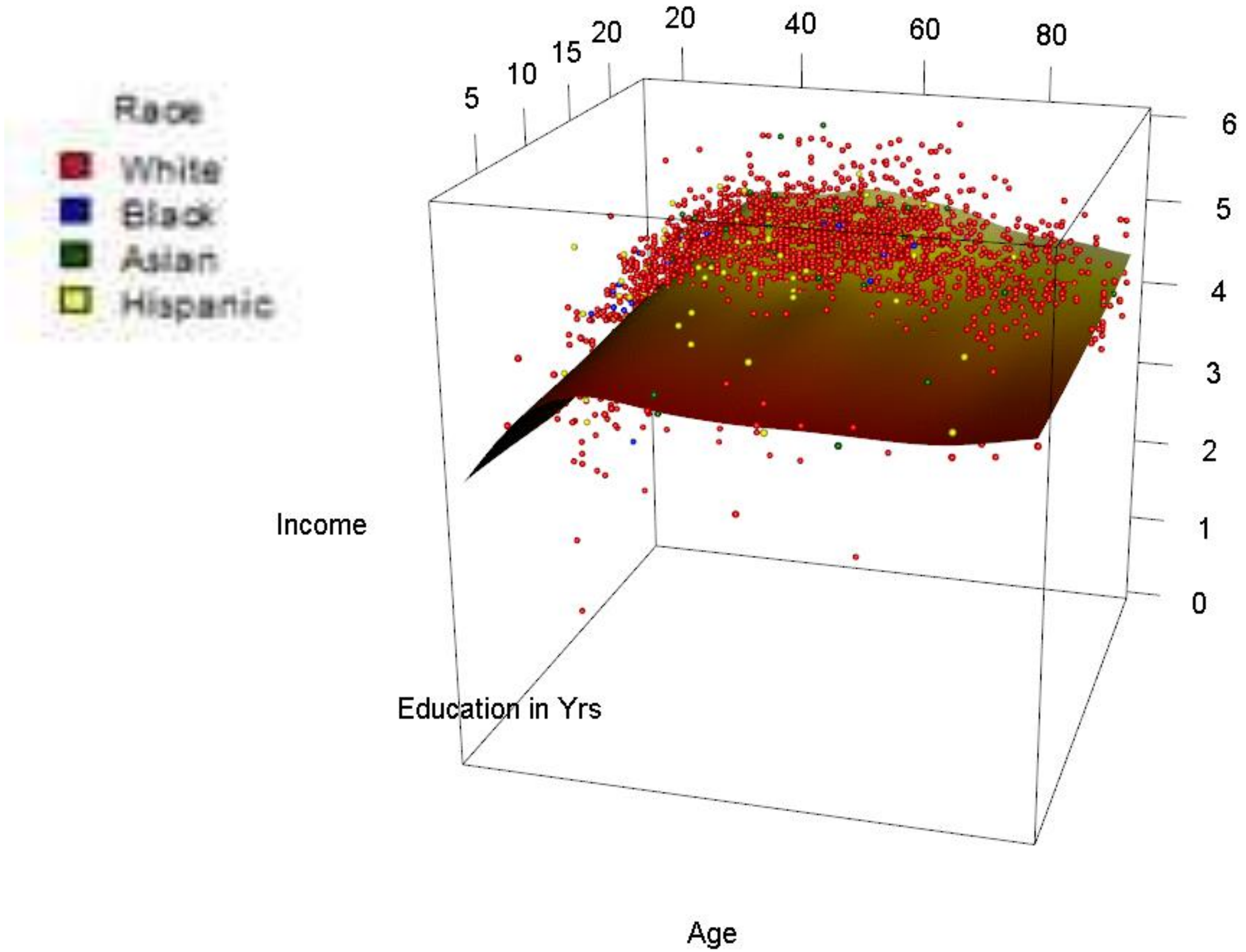


K-anonymity Can Distort Multivariate Relationships

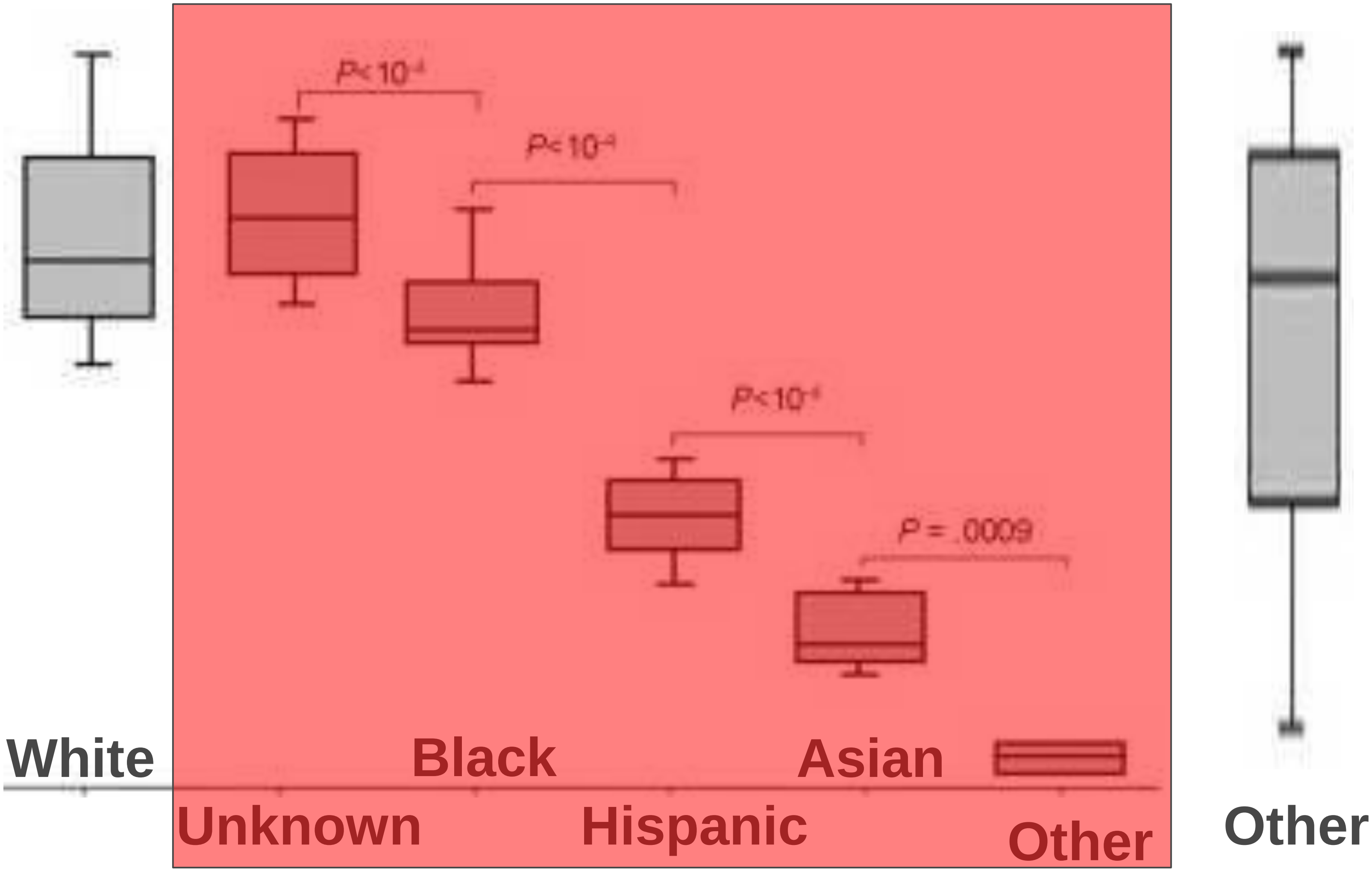


and this problem becomes more severe with with higher multi-dimensional space…




Race
White
Black
Asian
Hispanic
5
10
15
20
20
40
60
80
6
5
4
3
2
1
0
Income
Education in Yrs
Age

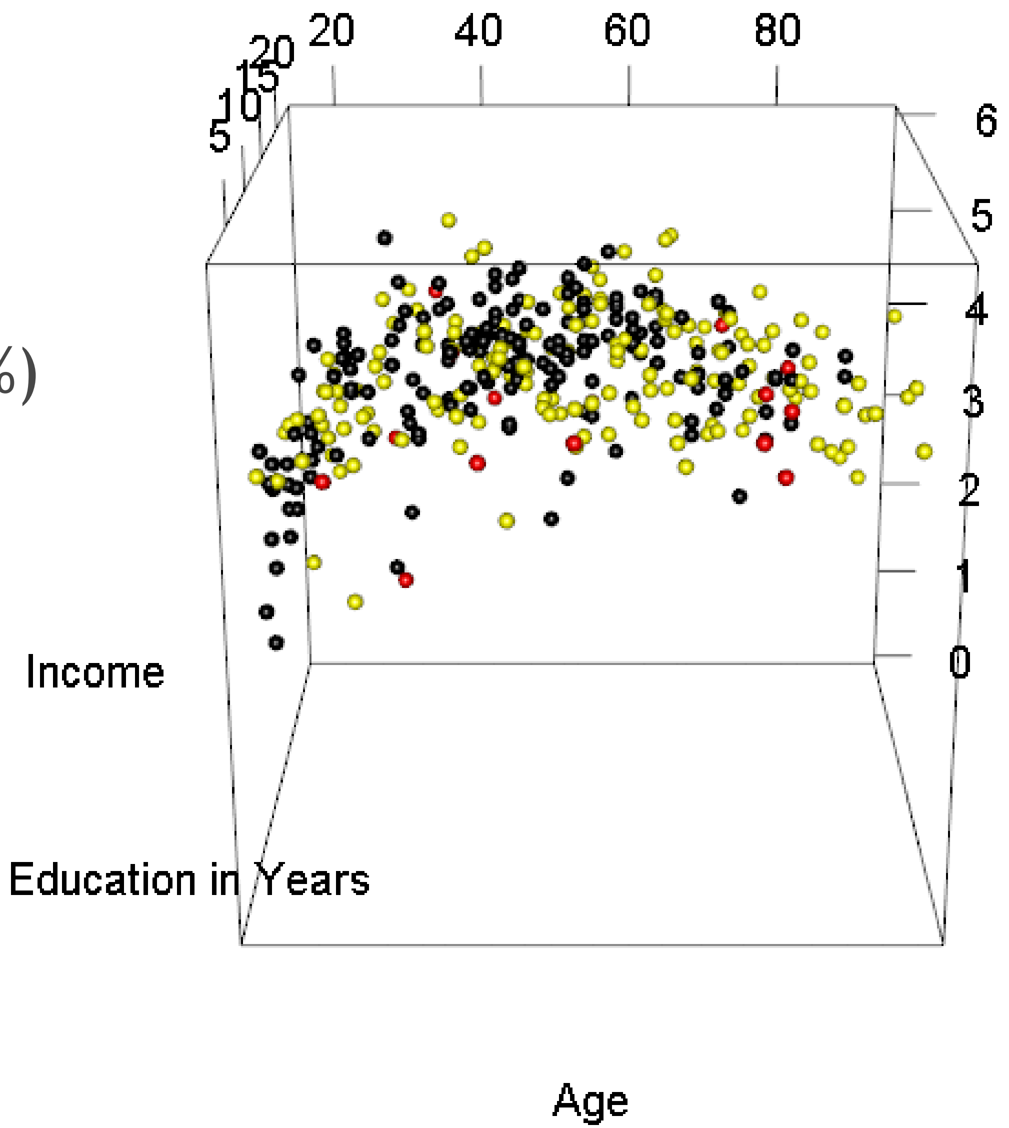
and… K-anonymity Hides Heterogeneities



K-anonymity Can Distort Multivariate Relationships

2 Percent Sample
from Population

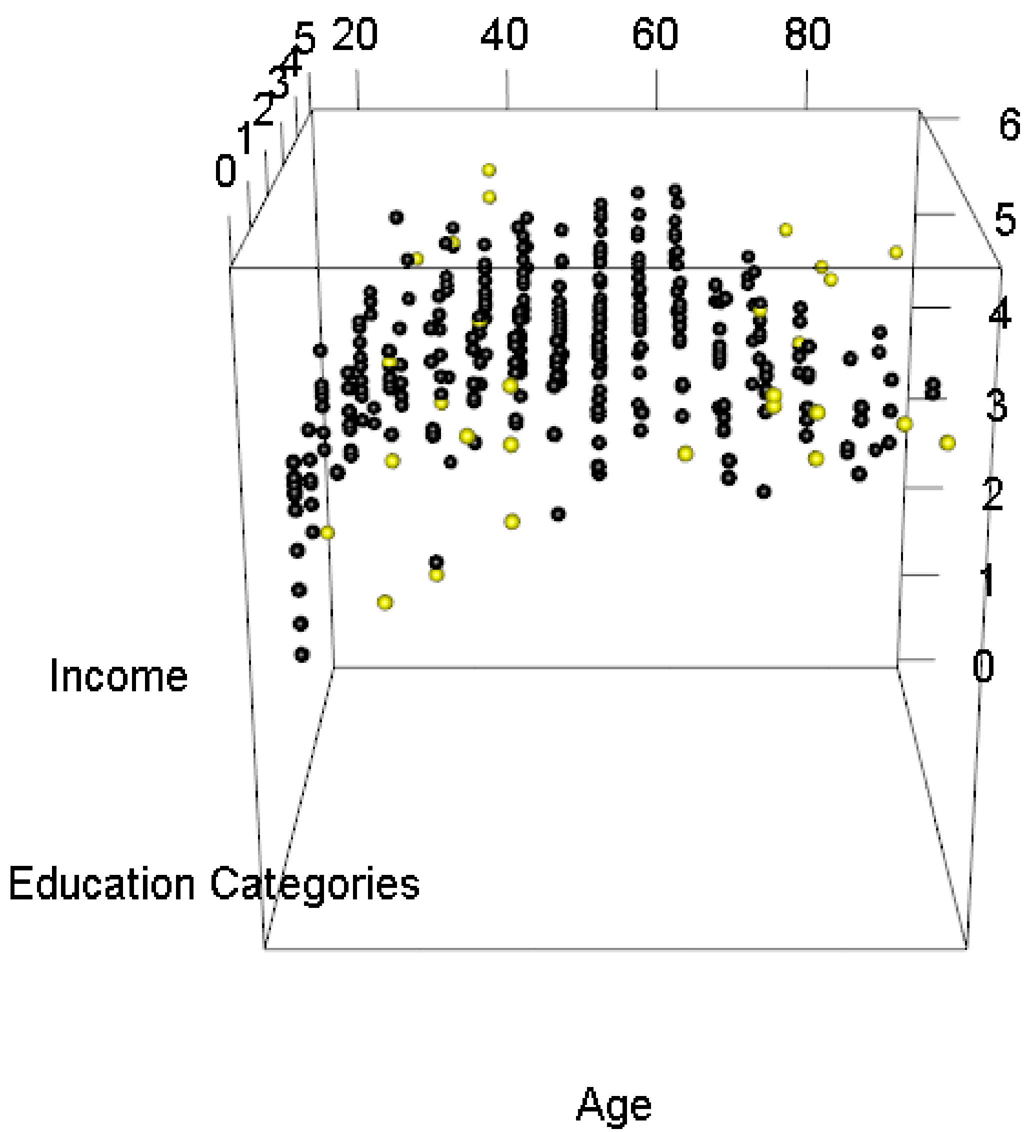
- Not Pop Unique (56%)
- Sample Unique, but not Pop Unique (40.6%)
- Pop Unique (3.5%)



K-anonymity Can Distort Multivariate Relationships

2 Percent Sample from Population

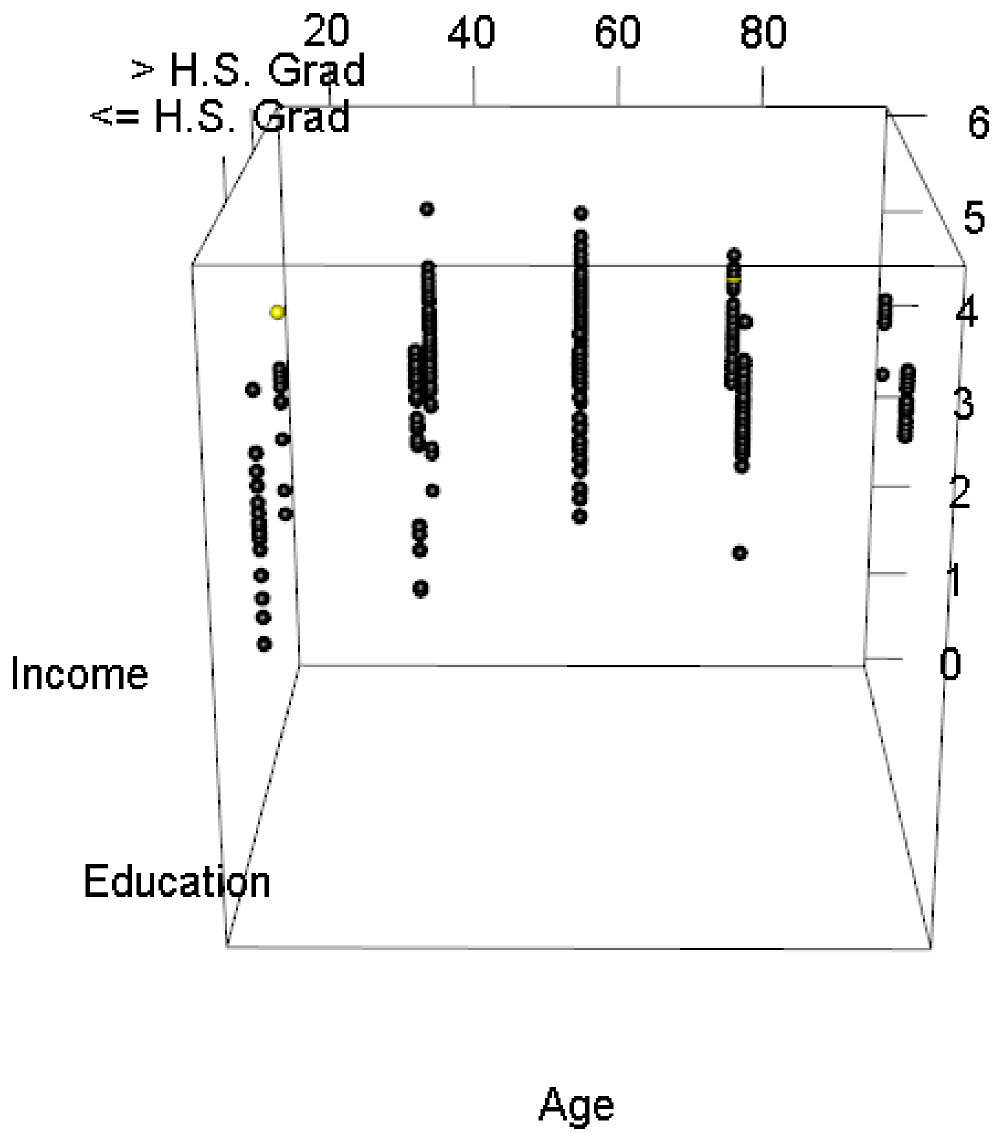
- Not Pop Unique (92%)
- Sample Unique, but not Pop Unique (8%)
- Pop Unique (0%)



K-anonymity Can Distort Multivariate Relationships

2 Percent Sample from Population

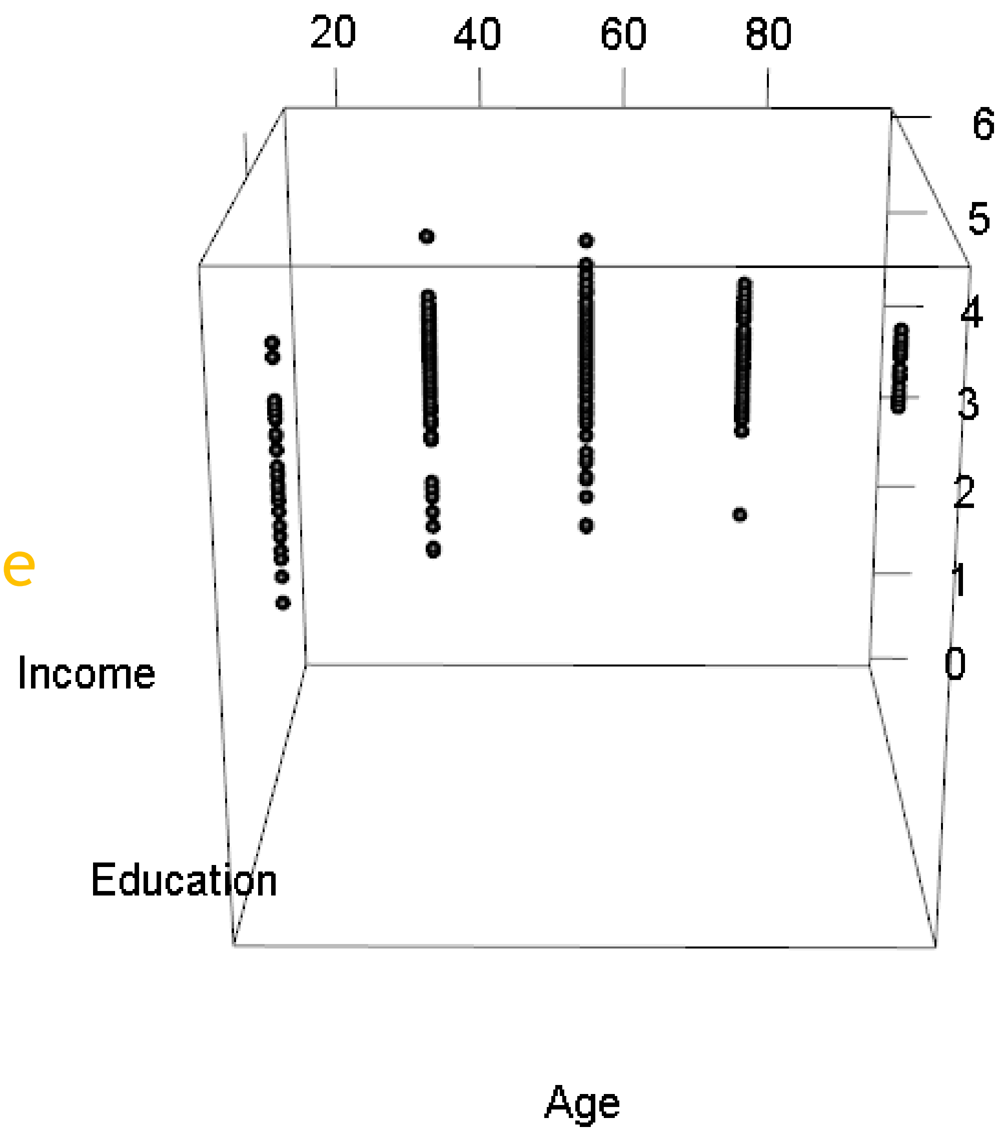
- Not Pop Unique (99.4%)
- Sample Unique, but not Pop Unique (0.6%)
- Pop Unique (0%)



K-anonymity Can Distort Multivariate Relationships

2 Percent Sample from Population

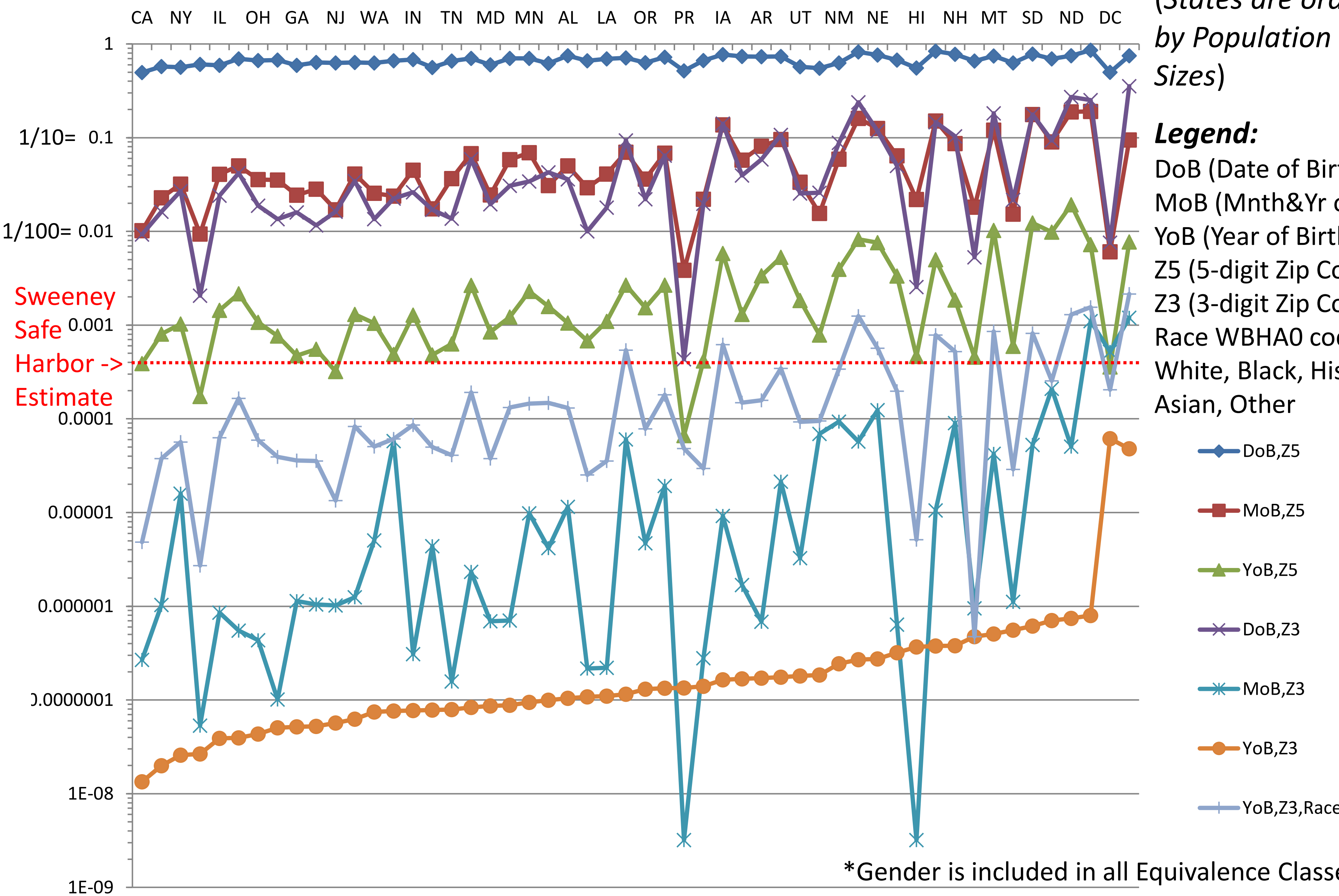
- Not Pop Unique (100%)
- Sample Unique, but not Pop Unique (0%)
- Pop Unique (0%)



State Specific Re-identification Risks: Population Uniqueness

(*States are ordered by Population Sizes*)

Legend:
DoB (Date of Birth)
MoB (Mnth&Yr of Birth)
YoB (Year of Birth)
Z5 (5-digit Zip Code)
Z3 (3-digit Zip Code)
Race WBHA0 coding: White, Black, Hispanic, Asian, Other



Data Source: 2010 U.S. Decennial Census

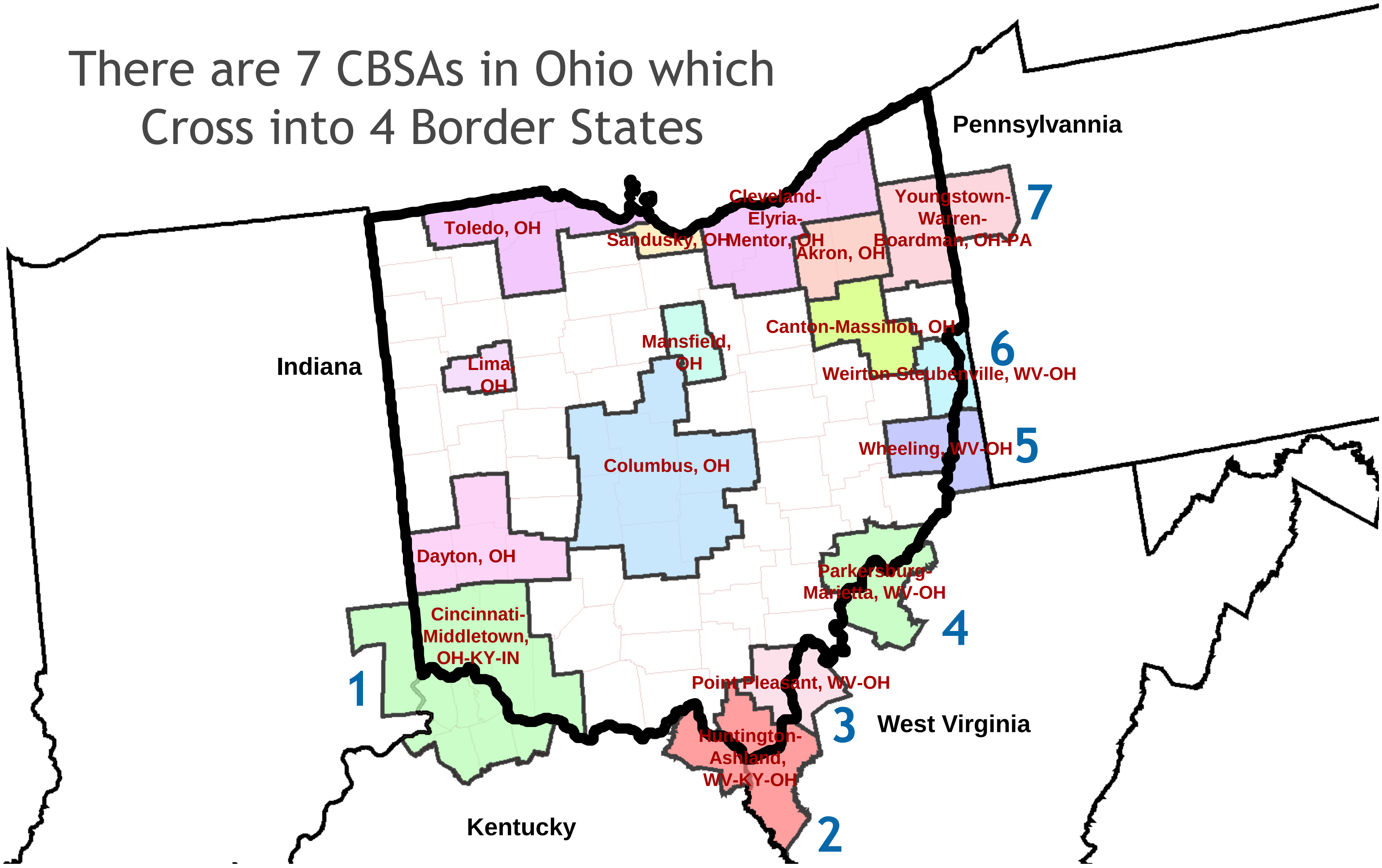
Graph © DB-J, dEpid/dt Consulting Inc.

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

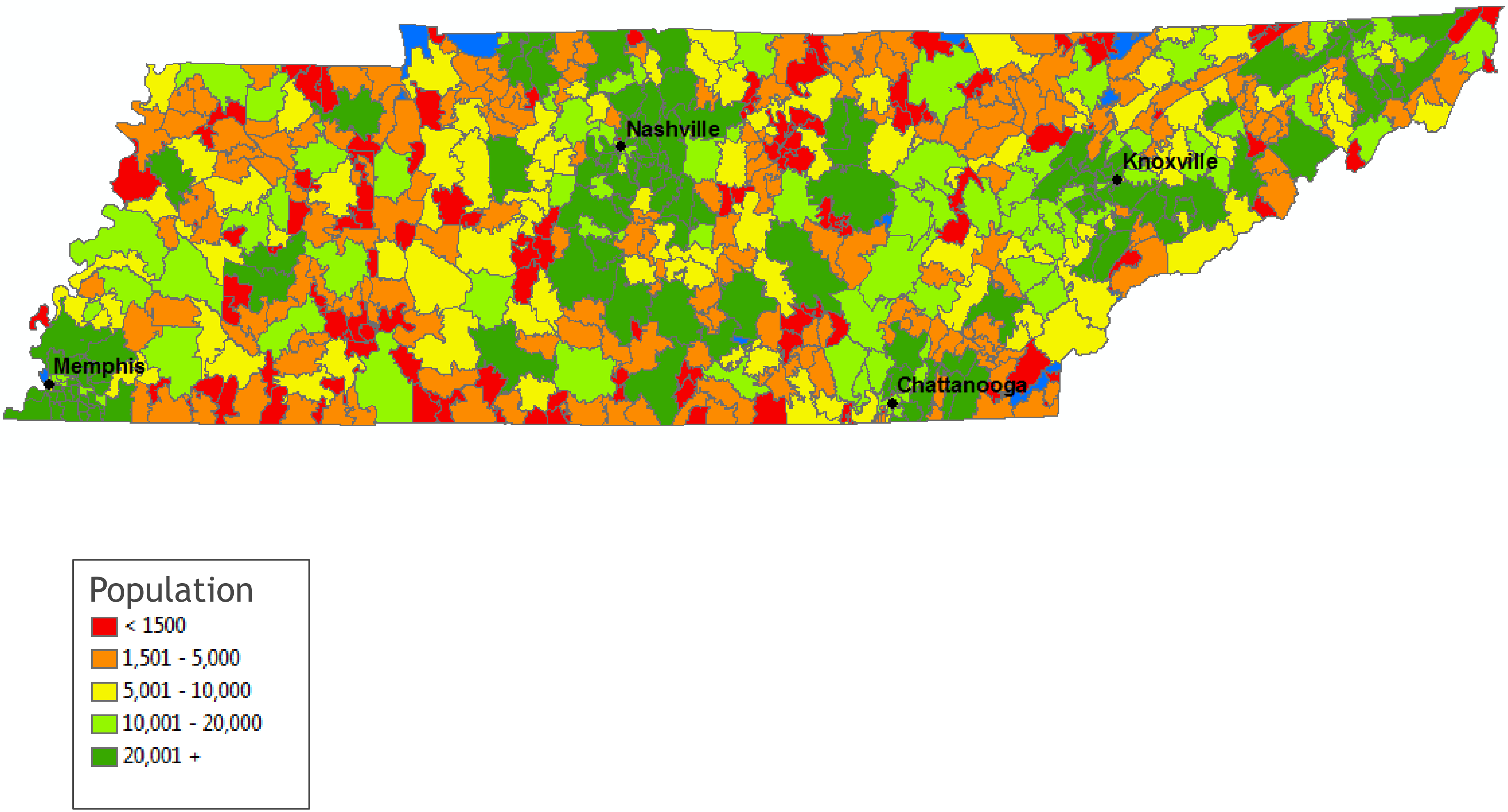
- Challenge: Data recipients often request reporting on more than one geography (e.g., both State and 3 digit Zip code).
- ***Subtraction Geography*** creates disclosure risk problems when more than one geography is reported for the same area and the geographies overlap.
- Also called ***geographical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

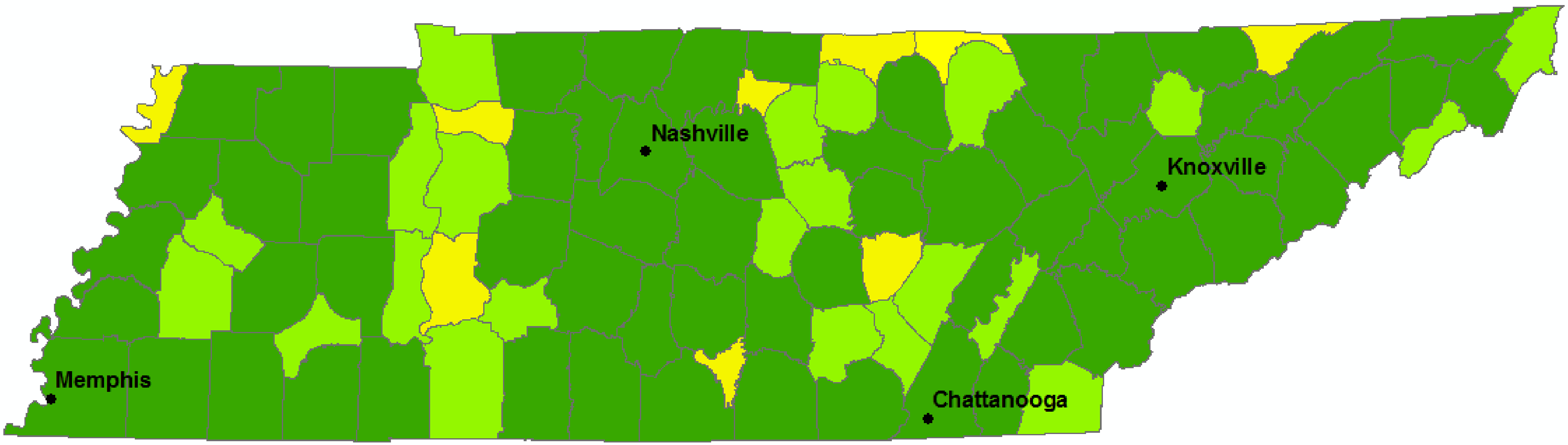
There are 7 CBSAs in Ohio which Cross into 4 Border States



Tennessee – ZCTA5 Populations

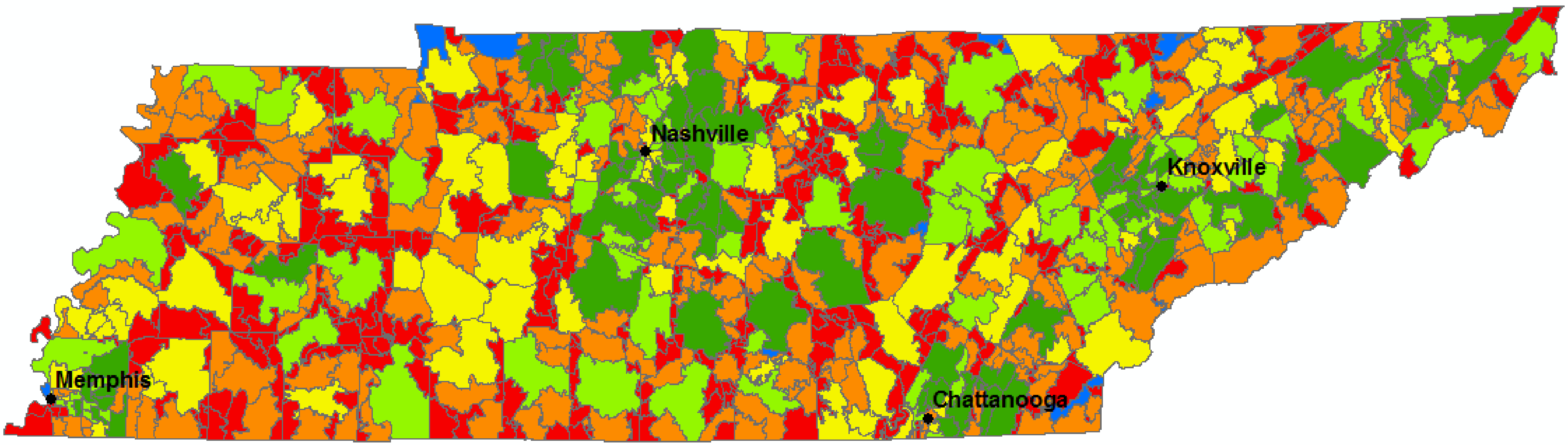


Tennessee – County Populations



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

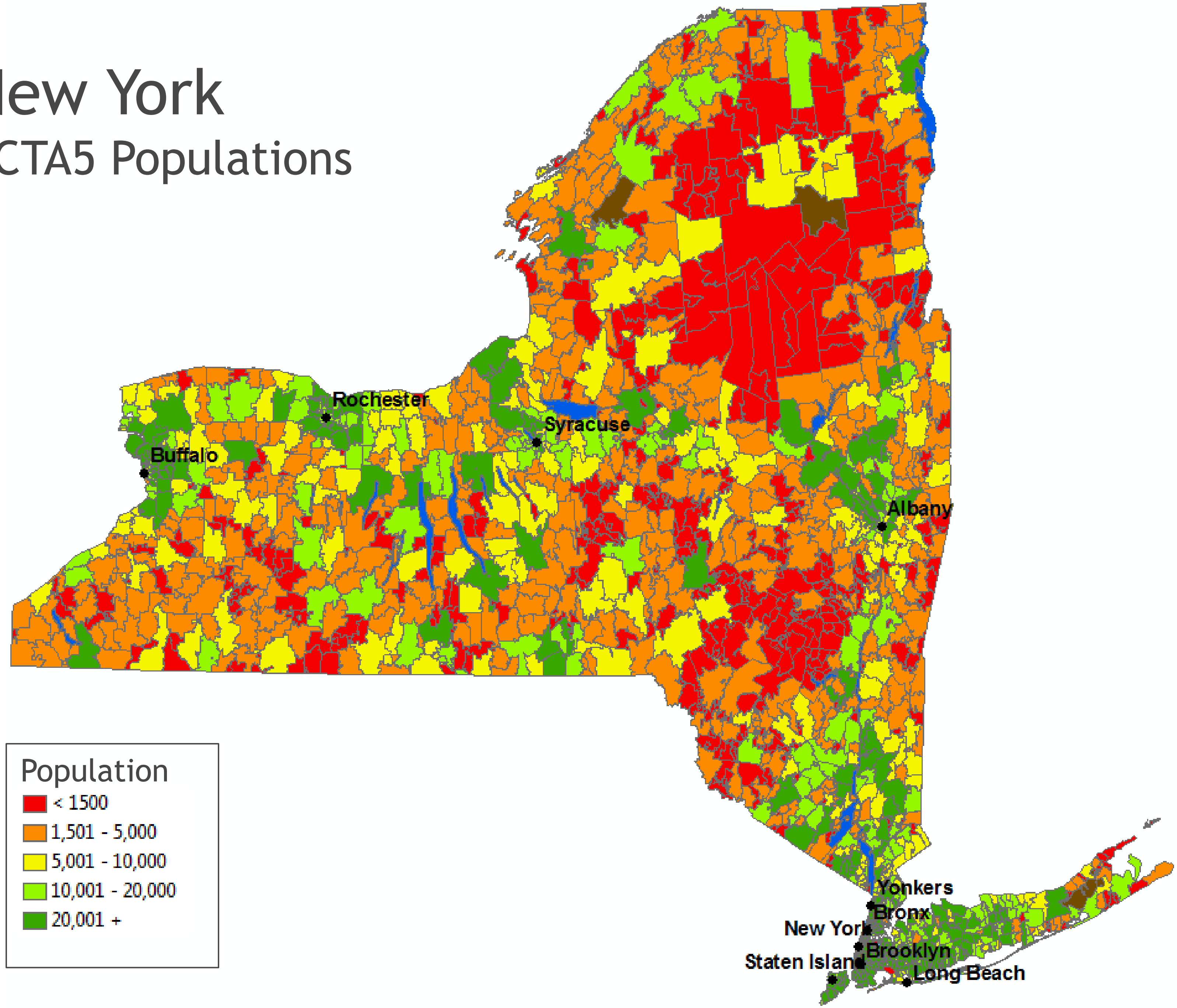
Tennessee – ZCTA5 X County Populations



Population

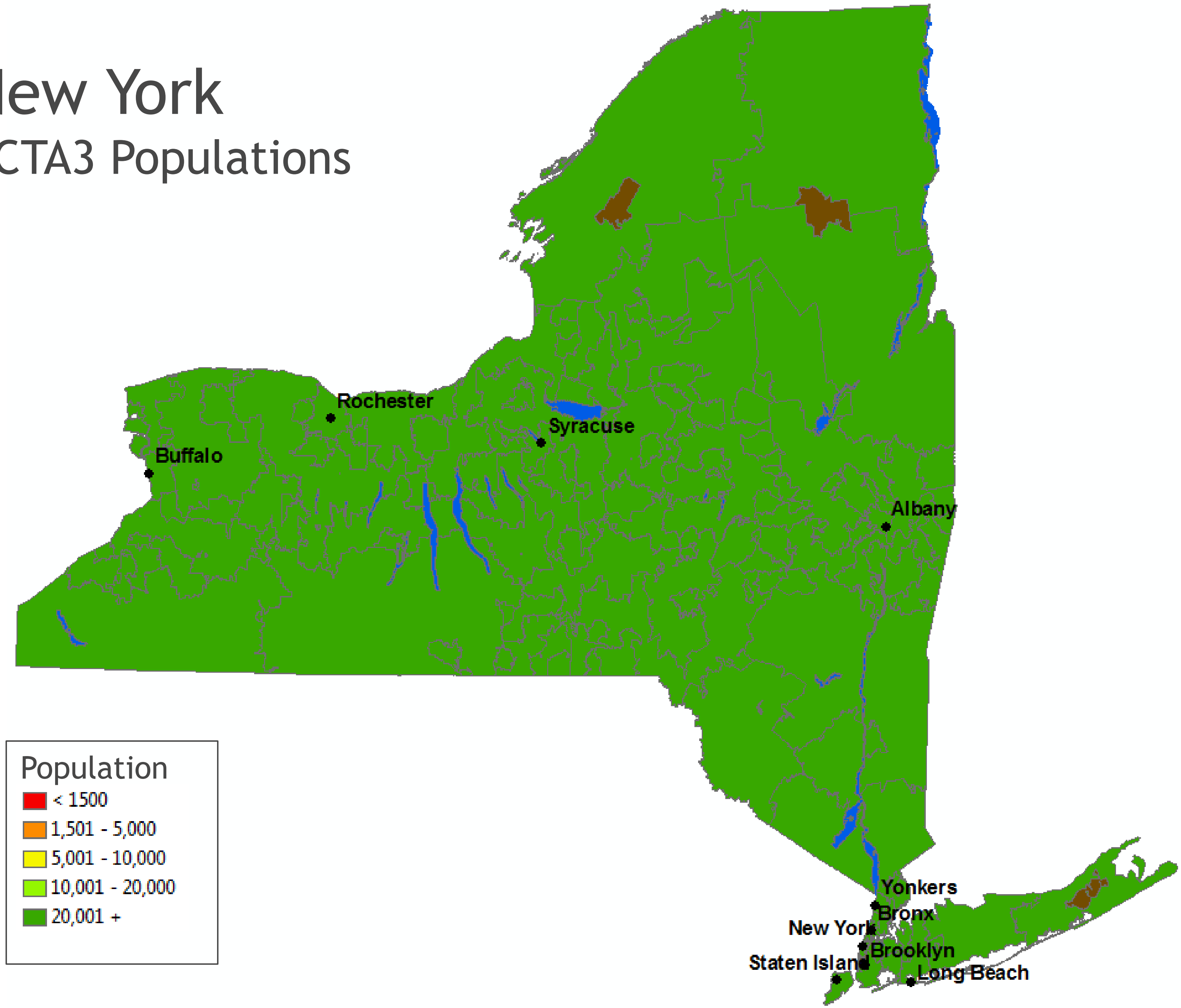
- < 1500
- 1,501 - 5,000
- 5,001 - 10,000
- 10,001 - 20,000
- 20,001 +

New York
ZCTA5 Populations

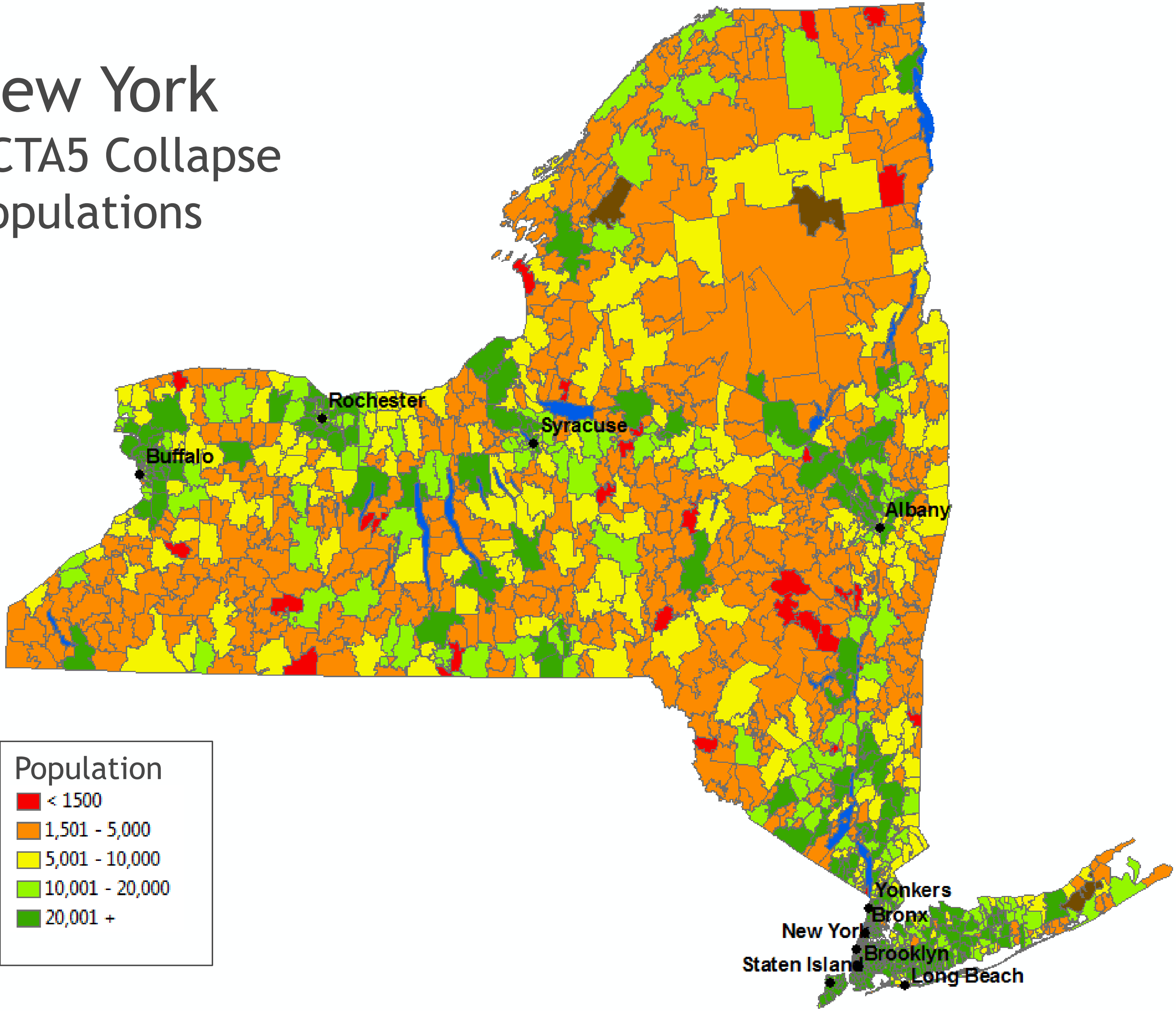


New York

ZCTA3 Populations



New York
ZCTA5 Collapse Populations



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 mapping & “Mash-up” technologies) ***on the internet***.

So, How Do We Move Beyond Anecdotes to a Rigorous, Scientific, Evidence-Based Risk Management Approach for Dealing with Re-identification Risks?

- 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 where considerable uncertainty exists for important risk management questions.

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.

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 Scenarios:

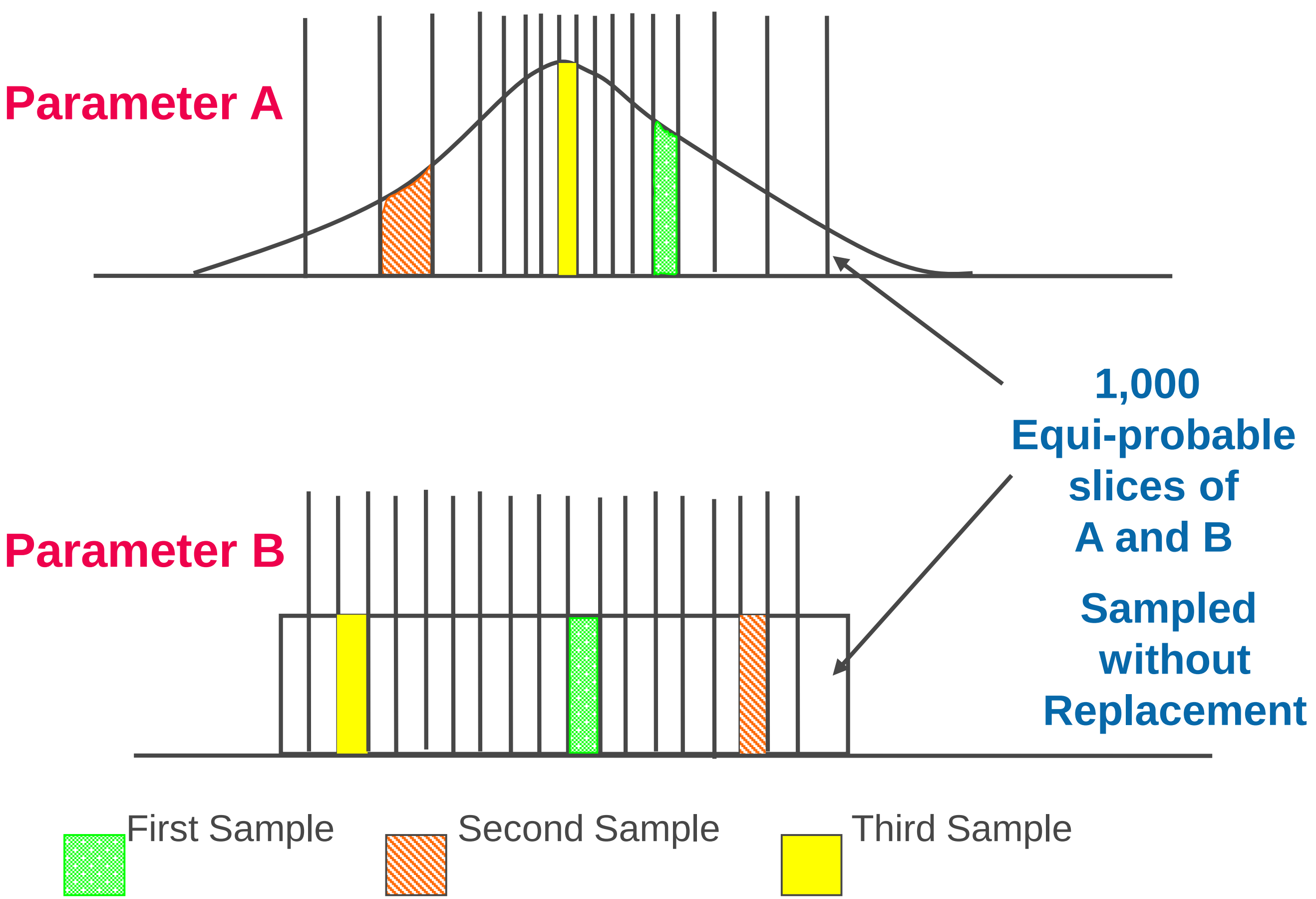
- Prob(Re-identification) =
  Prob(Re-ident|Attempt)***Prob(Attempt)**
- Note that Prob(Attempt) & Prob(Reident|Attempt) are actually not likely to be independent - higher re-identification probabilities are likely to increase re-identification attempts.
- Some very useful frameworks exist for characterizing 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

Quantitative Policy Science

- Conducting systematic quantitative cost-benefit policy analyses using state-of-the-art uncertainty and sensitivity analysis methods (e.g. with Latin-Hypergrid exploration of uncertain parameters) allows us to properly deal with the many important unknowns which could impact whether re-identification attempts under various data intrusion scenarios are likely be economically viable and realistic.

Latin Hypercube Sampling in Uncertainty Analyses

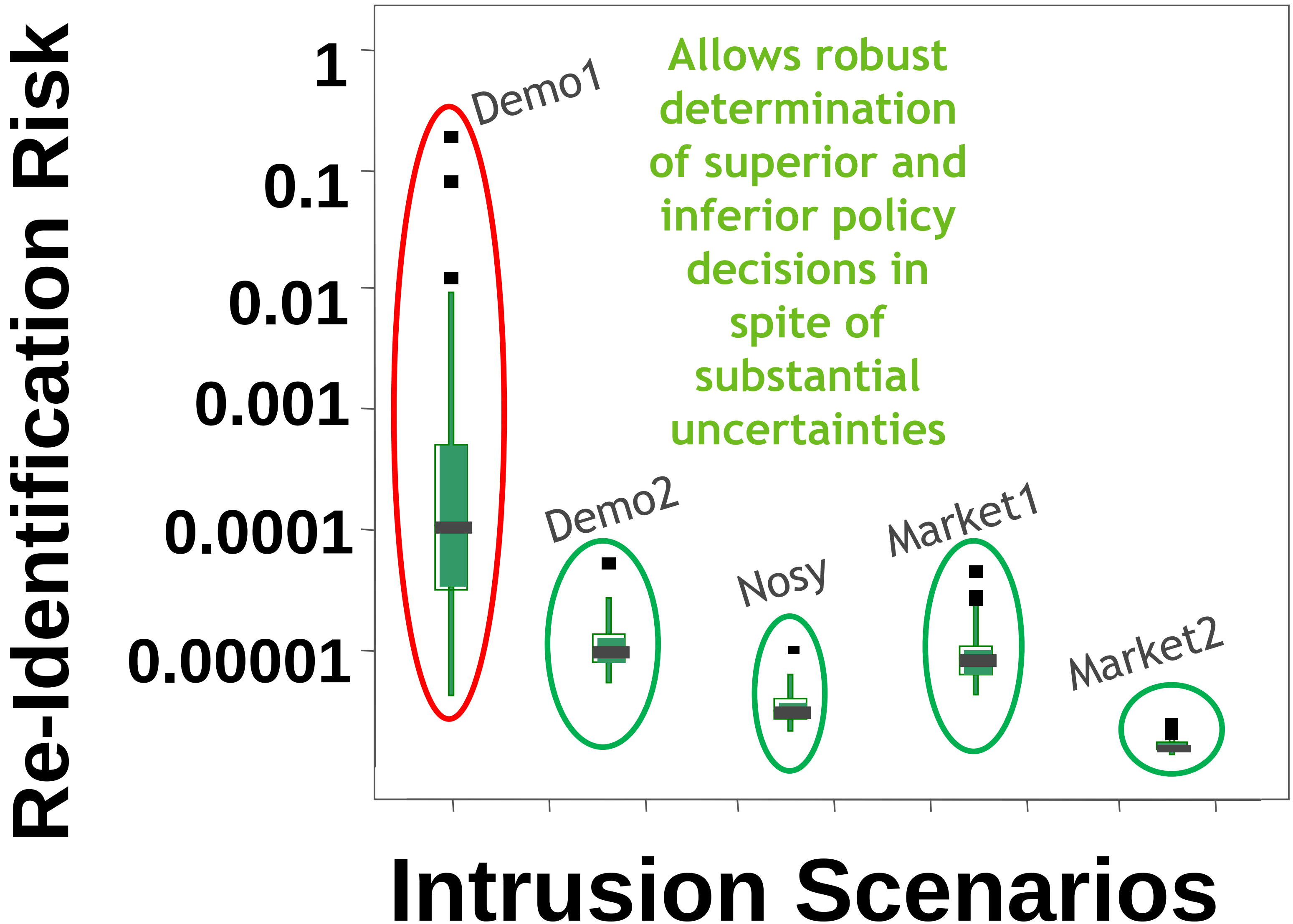
How multivariate uncertainty/sensitivity methods work



Use of Latin Hypercube Sampling:

Assures an **efficient and thorough search** of the plausible parameter space.

Uncertainty Analyses



References:

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

http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2076397

Bill of Health

Examining the intersection of law and health care, biotech & bioethics

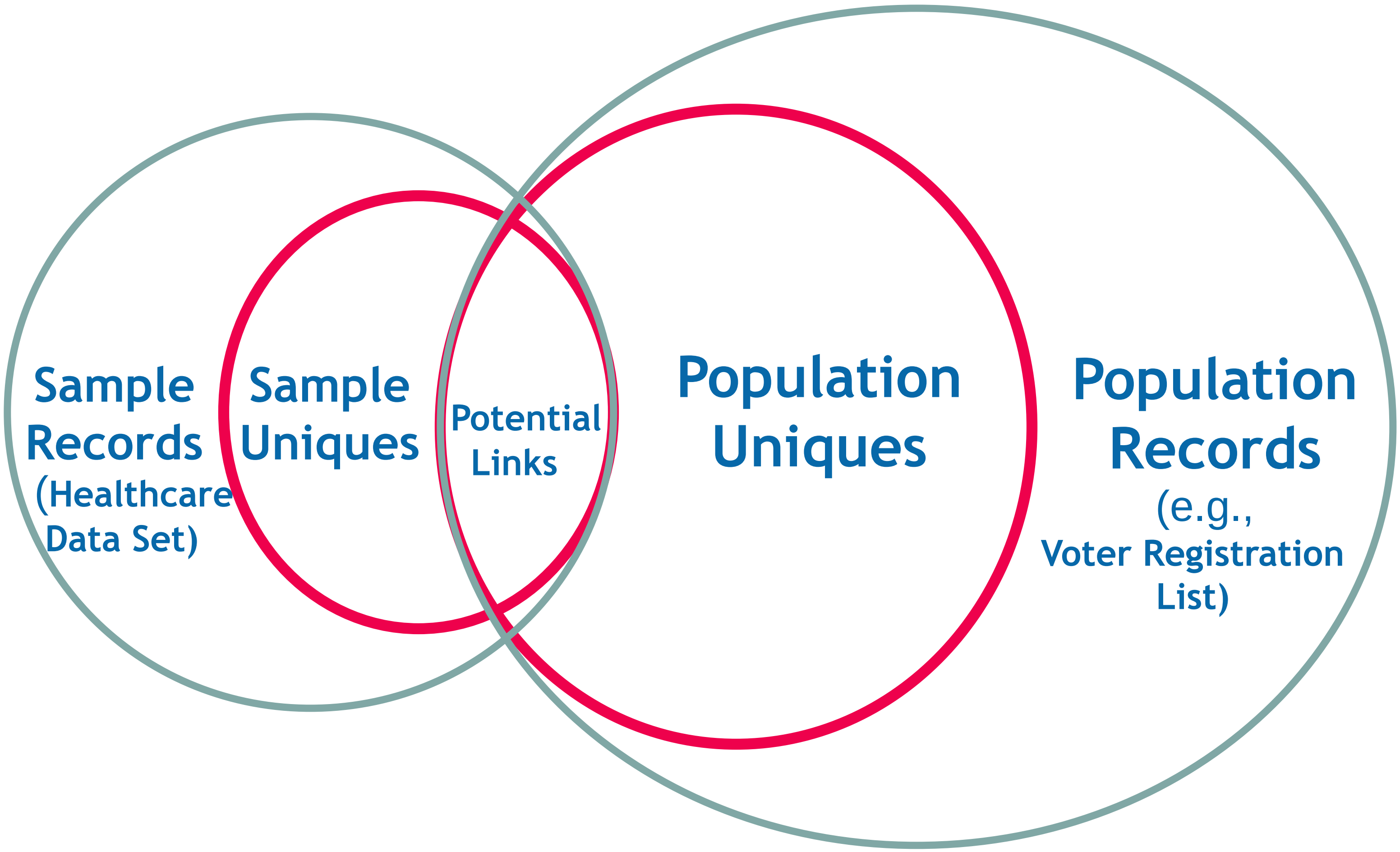
A blog by the Petrie-Flom Center and friends

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

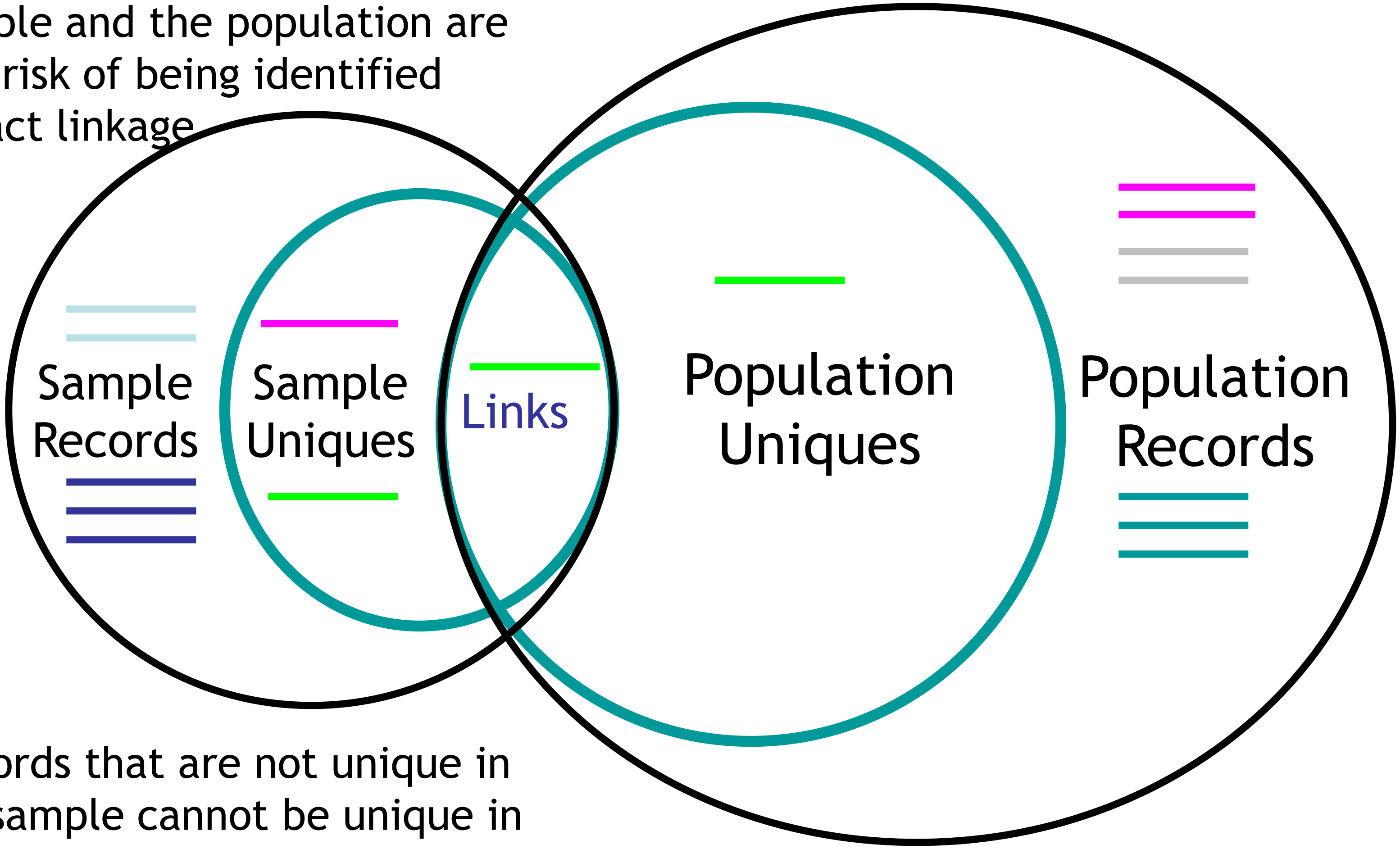
Measuring Disclosure Risks



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 clear 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

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	

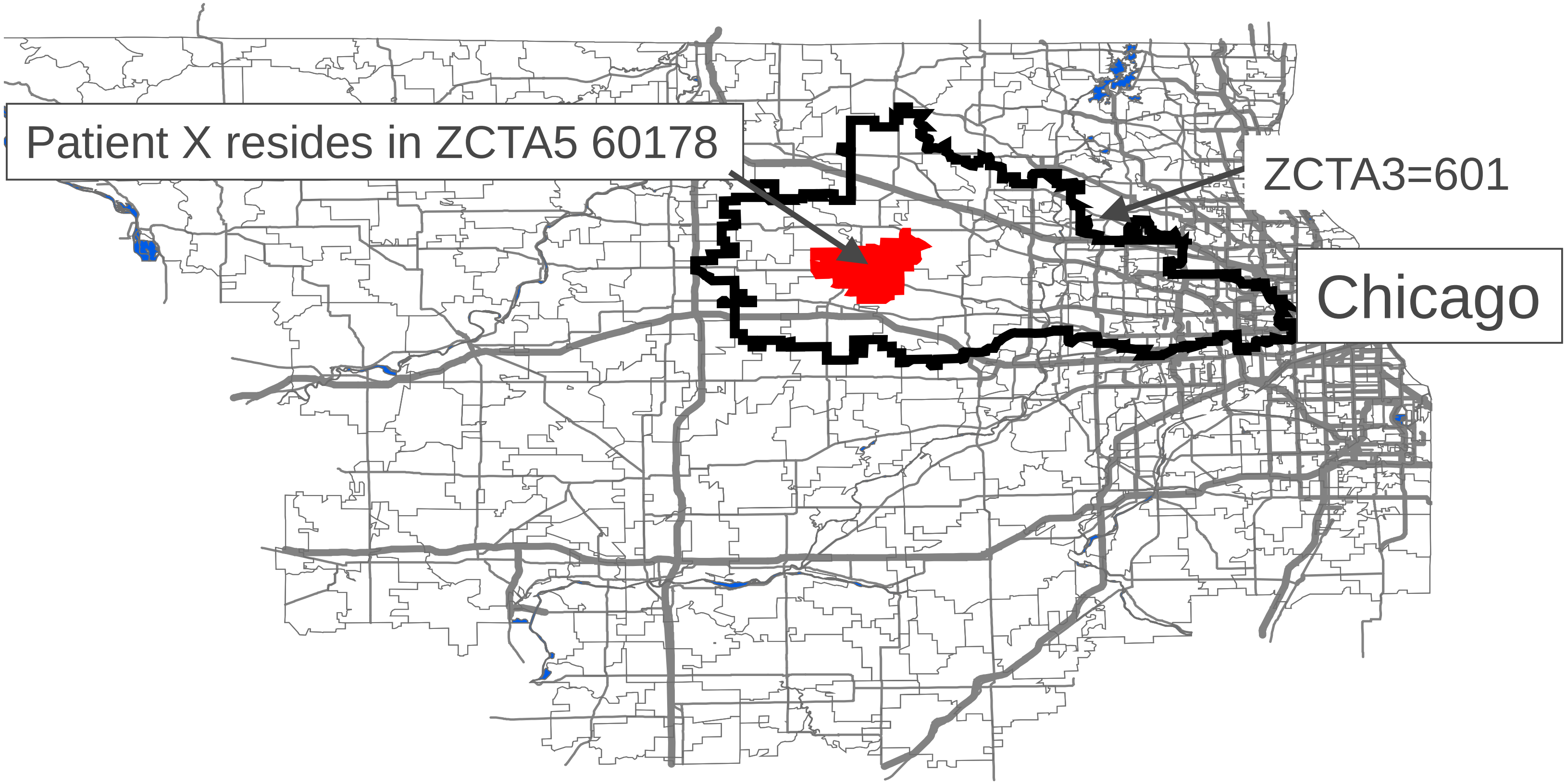
Note:
Figure illustrates only those limited cases where only one or two persons with shared "quasi-identifier" characteristics exist in either the healthcare data set or in the voter registration list.

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.

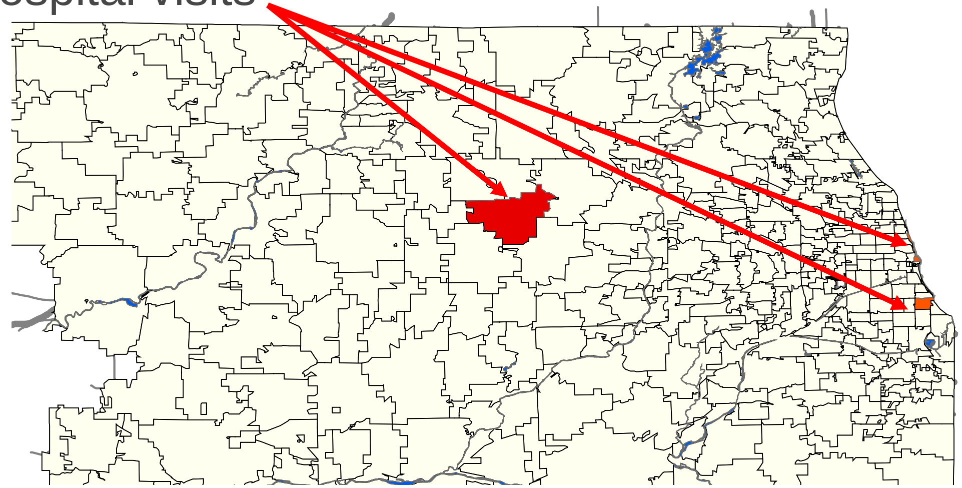
Challenge: Geoproxy Attacks



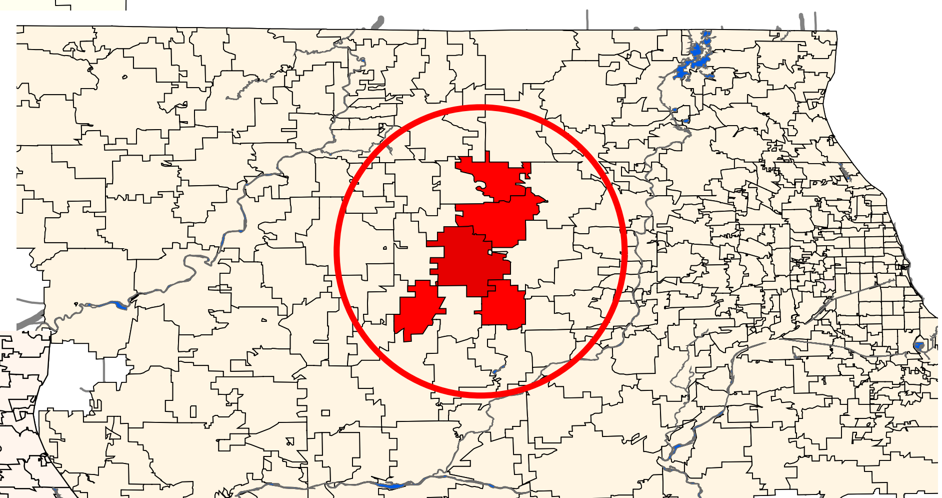
Example: Patient location as revealed within data set, but further narrowed to probable “hotspots” by using healthcare provider location data

Challenge: Geoproxy Attacks

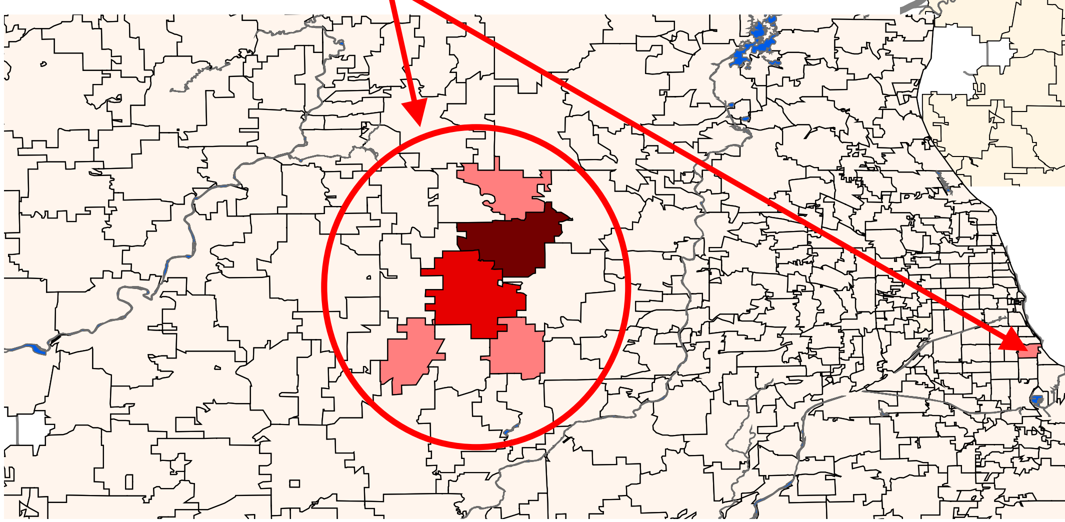
Hospital visits



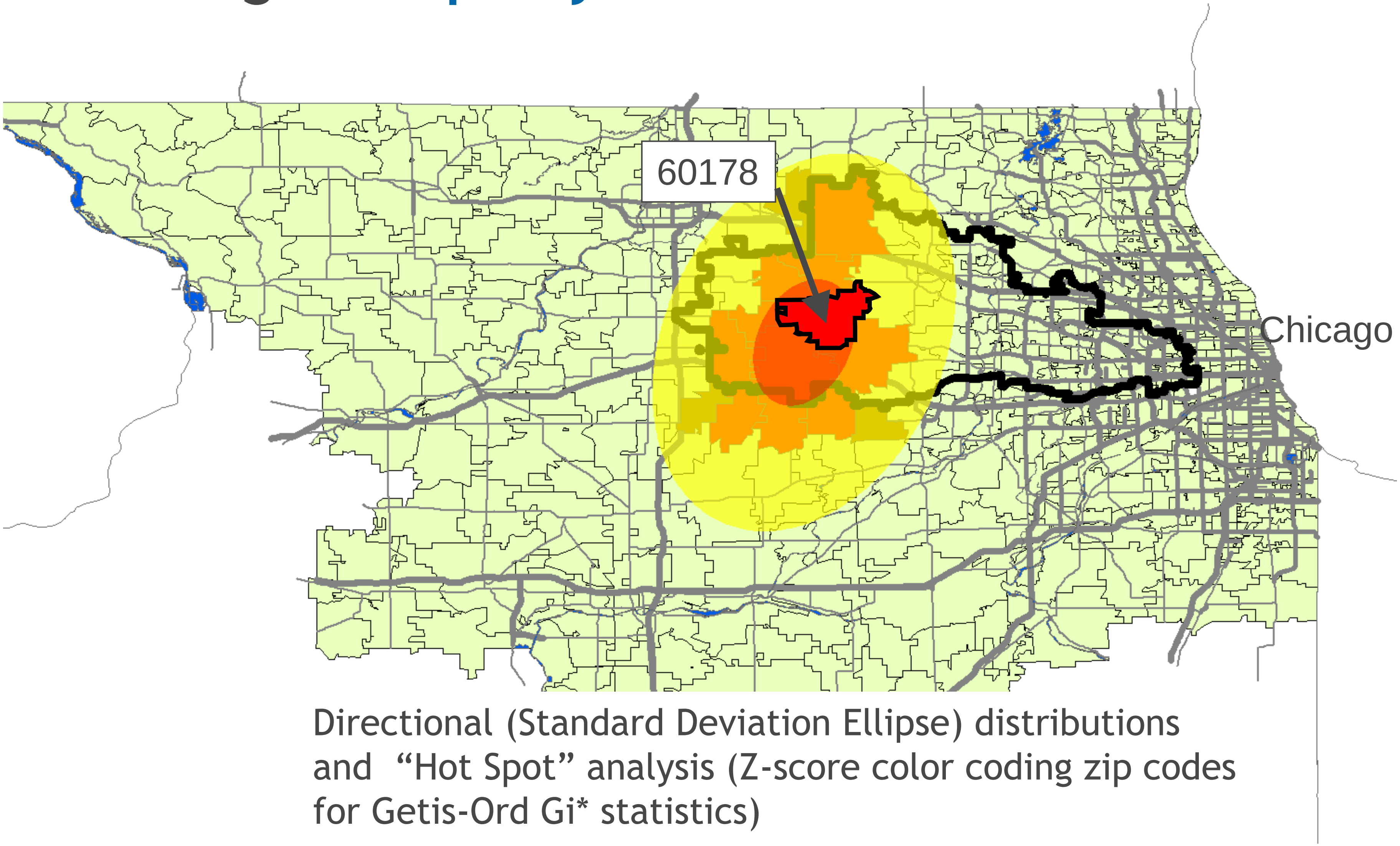
Outpatient/Office visits



Pharmacy visits



Challenge: Geoproxy Attacks

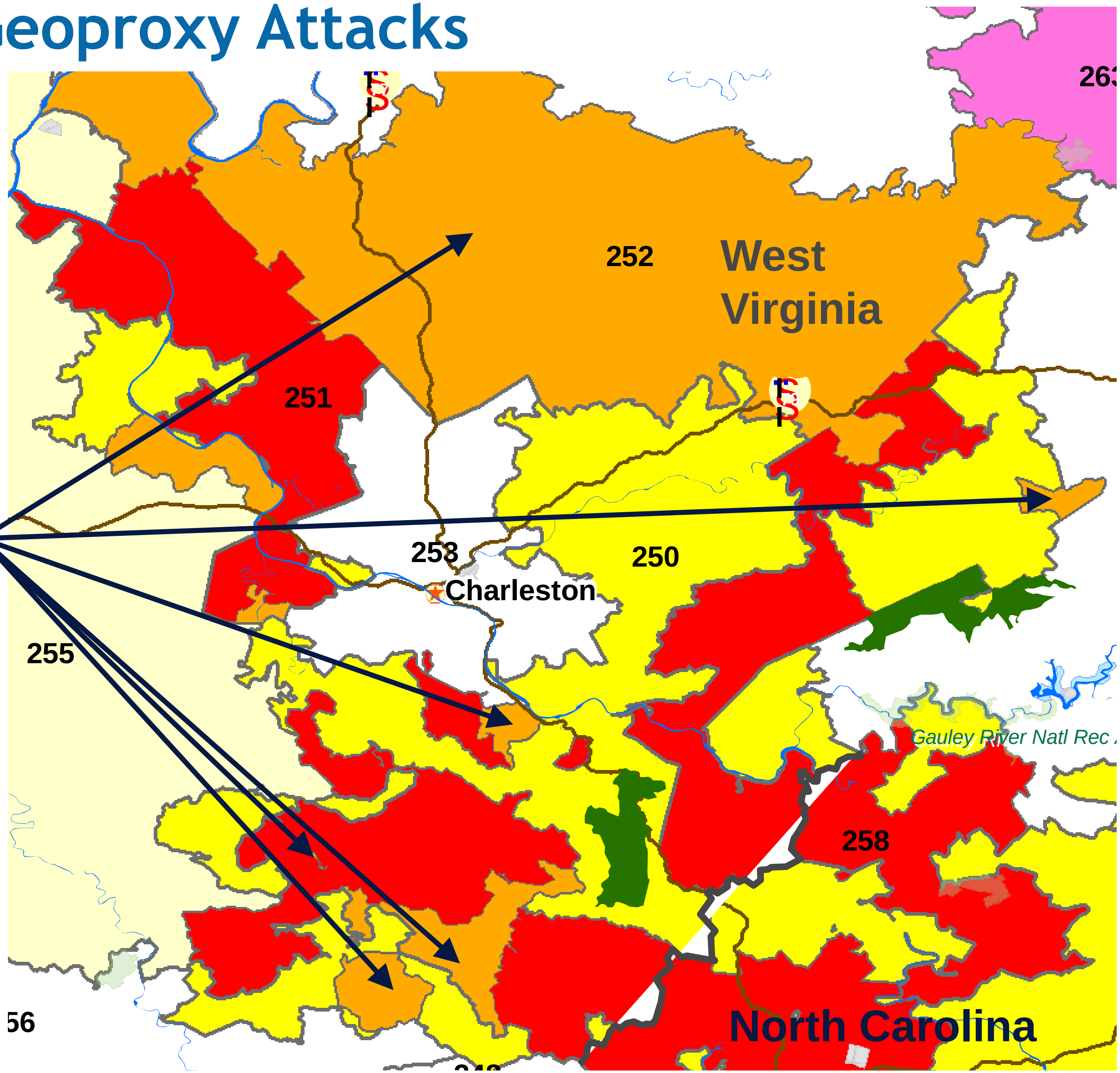


Directional (Standard Deviation Ellipse) distributions and "Hot Spot" analysis (Z-score color coding zip codes for Getis-Ord Gi* statistics)

Challenge: Geoproxy Attacks

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

ZCTA3 252 is highly dispersed

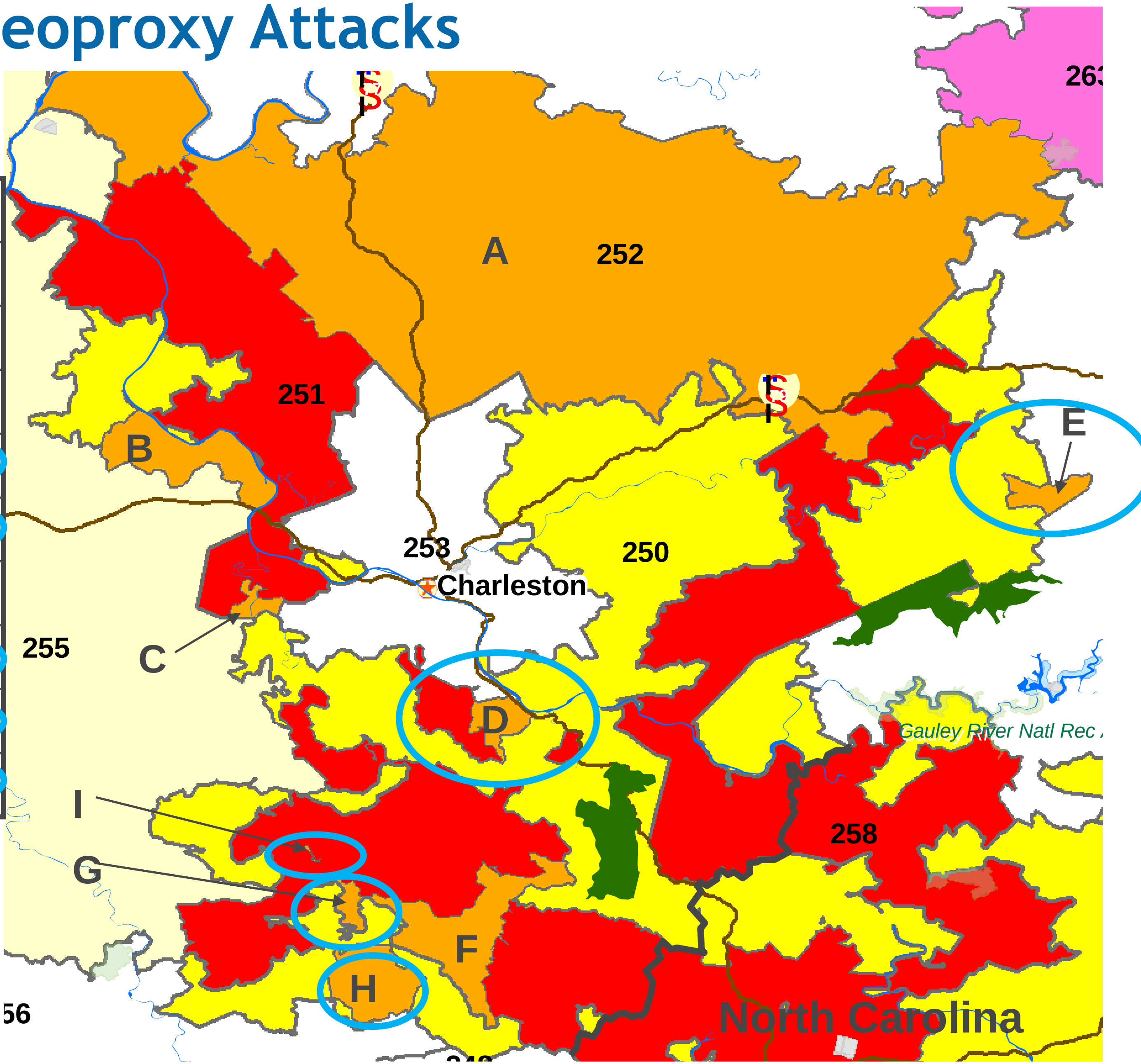


The complexity of 3-digit Zip Code Geography amplifies the threat of Geoproxy attacks

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



»Free the Data, Manage the Risks

Barbara Rudolph PHD, MSSW
UW-Madison, Center for Health Systems Research and Analysis
NAHDO Consultant

History of NAHDO Efforts in Privacy of Data

» Testimony at NCVHS

» White Papers and Contract Project Reports

» Published articles

» Technical Tools—Inventory for Screening (Data) Release

» Administrative Data Committee leading to PHDSC

» Guidance Documents and Technical Assistance (NY)

» Privacy Workgroups

NAHDO Documents Covering Data Management and Release Policies

Rudolph, B. Davis, R. Administrative Data and Disease Surveillance: An Integration Toolkit. NAHDO-CDC Cooperative Agreement Project, CDC Assessment Initiative

Rudolph B., Shah, G., Love D. Small numbers, disclosure risk, security, and reliability issues in web-based data query systems. J Public Health Management Practice, 2006.

Person-level Data: An Inventory for Screening Release, NAHDO, 2008.

PHDSC, PRISM--A Privacy Toolkit for Public Health Professionals.; Glossary.

Guidance Document on Creating and Releasing Hospital and Facility Discharge Data Public Use Files, NAHDO, 2012.

Key Privacy Points From NAHDO Documents

>Define policies for release keeping in mind important societal needs for data and privacy for the individual patient. The utility of data must be preserved as well as the privacy of patient data or collection will cease.

>Establish management controls such as: data use agreements, limited access to files, limited access to data locations, limited data use beyond premises, encryption, etc.

>Promulgate administrative rules on: data release; exceptions to open record statutes; IRBs and/or privacy boards; direct and indirect or sensitive data elements and their release; limits for re-release of data elements by users; limits on attempt to re-identify patients; penalties for misuse and penalty enforcement

> Establish statistical processes, define specific tests, or establish rules for reducing risk of re-identification from release of sensitive data elements, such as detailed geographic identifiers, service or other dates, rare diagnoses, race and ethnicity.

Key Privacy Points from NAHDO Documents

>Establish limits for cell sizes in tables or reports; web-query systems, tables and maps and determine if you must aggregate years of data or geography to address small cell sizes

>Define a responsible person(s) within your organization for each of these tasks in job descriptions to assure internal compliance with the processes

>Provide annual staff training on the processes for maintaining privacy of the patient