Understanding De-identification, Limited Data Sets, Encryption and Data Masking under HIPAA/HITECH: Implementing Solutions and Tackling Challenges

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

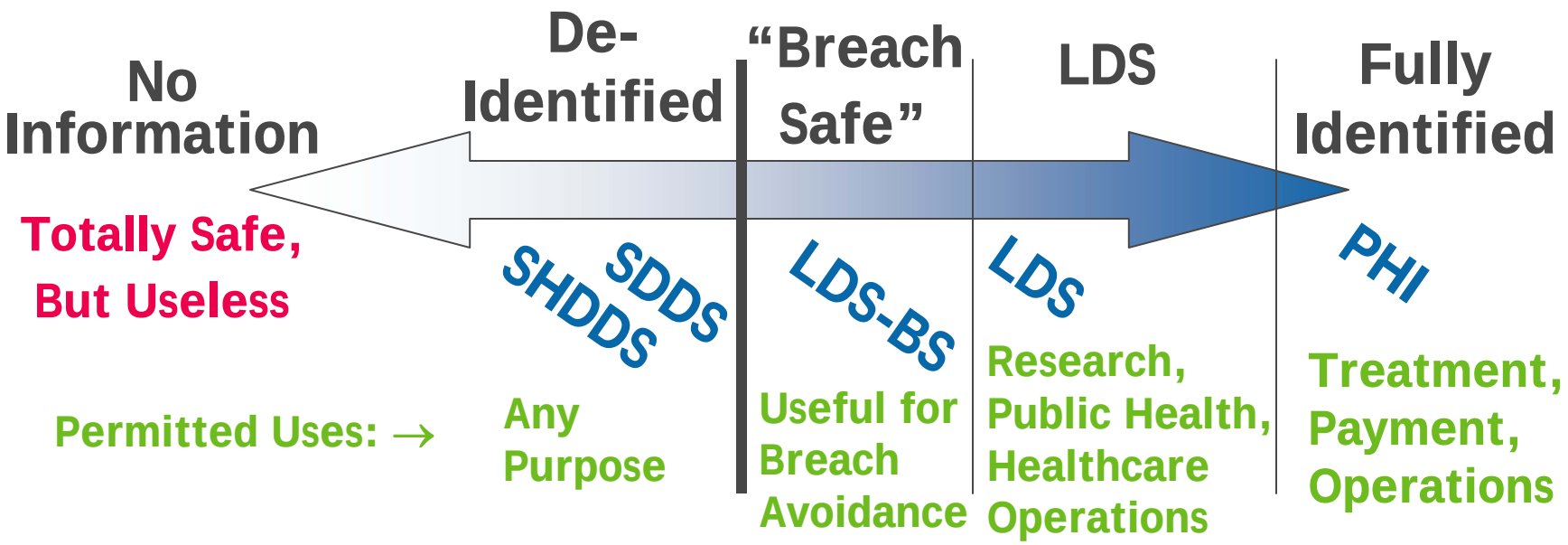
Essential HIPAA/HITECH Solutions:

Understanding LDS, Safe Harbor/Statistical De-identification

For:

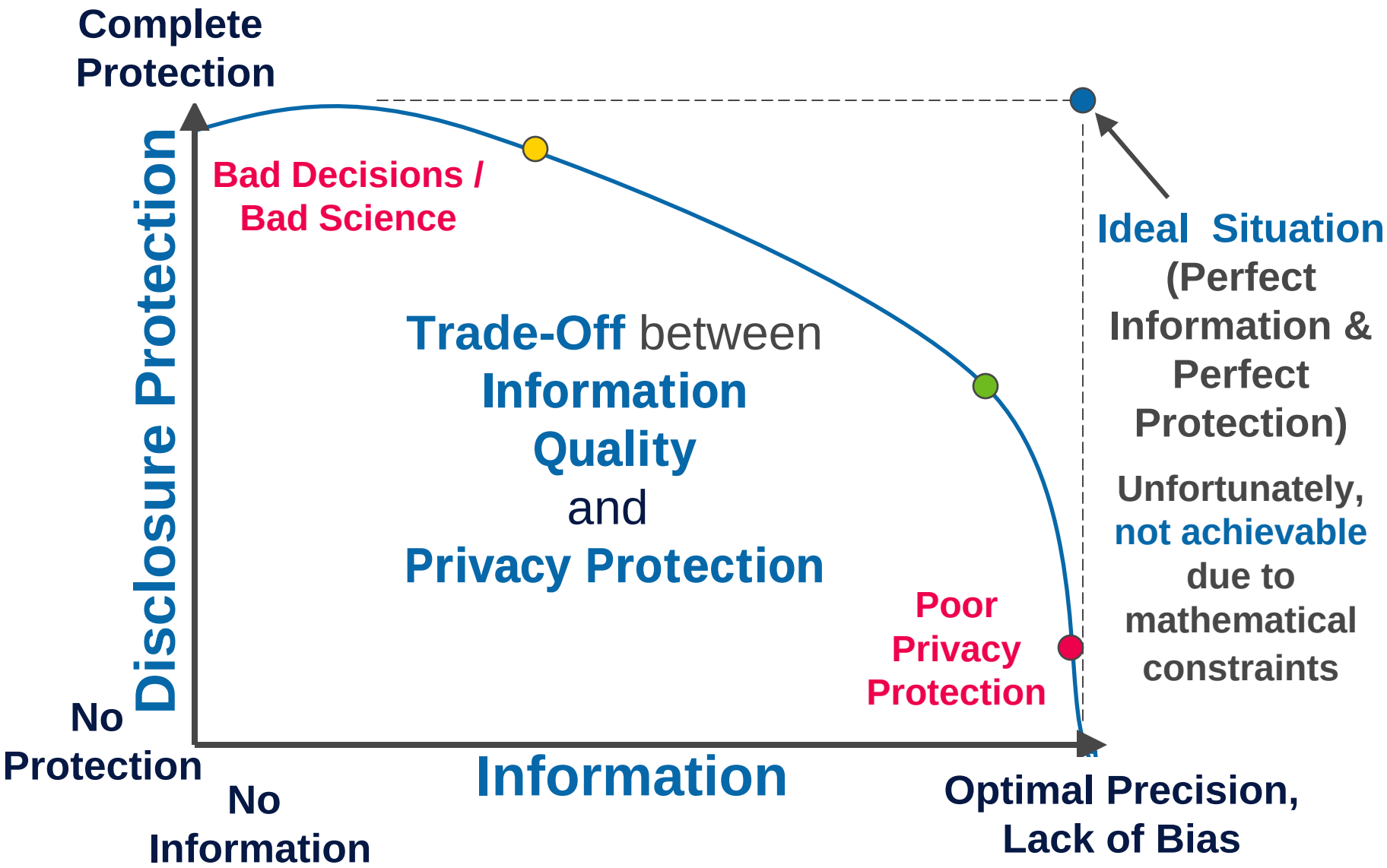
- Avoiding HITECH Breach Notification
- Supporting
  - Research,
  - Public Health,
  - Healthcare Operations
- Permitting Sale of Critical PHI elements
  - Date Information
  - Geographic Information
- Supporting Data Masking
  - Software Testing/Development/Demo

Identification Spectrum



- **Limited Data Set** (LDS) §164.514(e)
  - Eliminate 16 Direct Identifiers (Name, Address, SSN, etc.)
- **LDS w/o 5-digit Zip & Date of Birth** (LDS-"Breach Safe") **8/24/09 FedReg**
  - Eliminate 16 Direct Identifiers and Zip5, DoB
- **Safe Harbor De-identified Data Set (SHDDS)** §164.514(b)(2)
  - Eliminate 18 Identifiers (including Geo < 3 digit Zip, All Dates except Yr)
- **Statistically De-identified Data Sets (SDDS)** §164.514(b)(1)
  - **Verified "very small" Risk of Re-identification**

The Inconvenient Truth:



Inadequacies of Safe Harbor De-identification

- Challenging in complex data sets
  - Safe Harbor rules prohibiting Unique codes (§164.514(2)(i)(R)) unless they are not "derived from or related to information about the individual" (§164.514(c)(1)) can create significant complications for:
    - Preserving referential integrity in relational databases
    - Creating longitudinal de-identified data
- Encryption does not equal de-identification
  - Encryption of PHI, rather than its removal - as required under safe harbor, will not necessarily result in de-identification
- Not suitable for "Data Masking"
  - Removal requirement in 164.514(b)(2)(i)
  - Software development requires realistic "fake" data which can pose re-identification risks if not properly managed

Statistically De-identified Data Sets (SDDSs)

- *Statistical De-identification* often can be used to release some of the safe harbor "prohibited identifiers" provided that the risk of re-identification is *"very small"*.
- For example, more detailed *geography*, *dates of service* or *encryption* codes could possibly be used within statistical de-identified data based on statistical disclosure analyses showing that the risks are very small.
- However, disclosure analyses must be conducted to assess risks of re-identification

  (e.g., encrypted data with strong statistical associations to unencrypted data can pose important re-identification risks)

HIPAA Statistical De-identification Conditions

- "Risk is *very small…*"

  —"that the *information could be used*"…

  —"alone or *in combination with other reasonably available information*"…,

  —"*by an anticipated recipient*"…

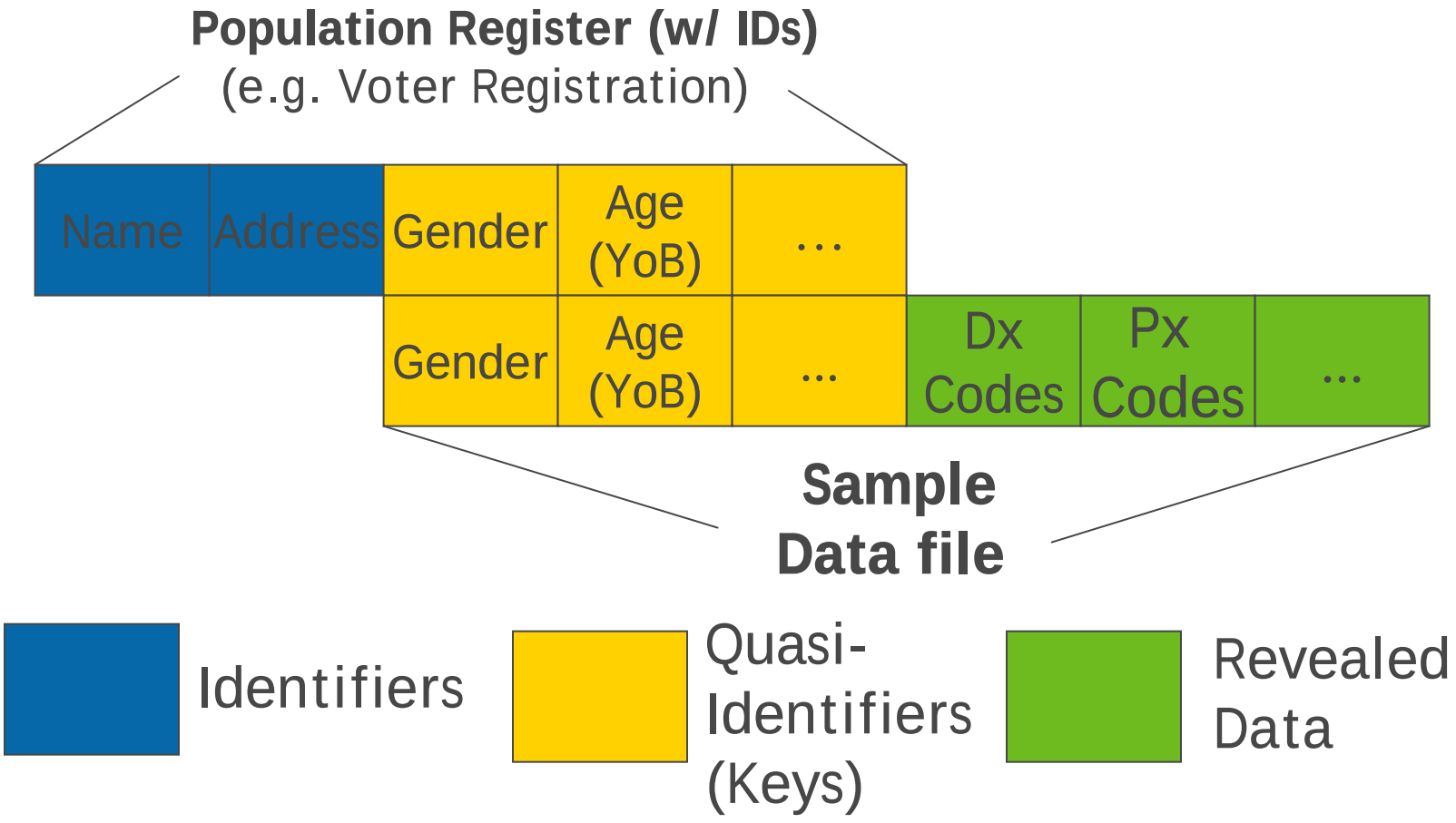
  —"*to identify an individual*"…

Essential Re-identification Concepts

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

Record Linkage

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



Quasi-identifiers

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

^------- Quasi-identifiers ---------^

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

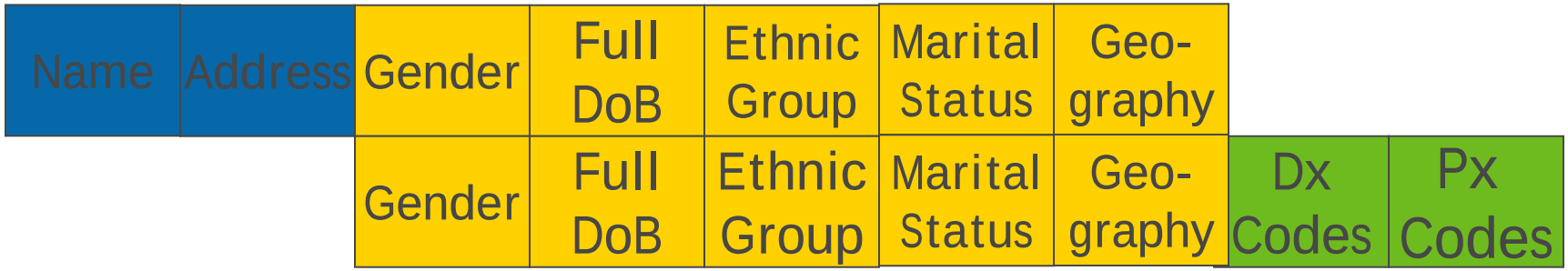
Note that this includes even fields that are not "PHI".

Key Resolution

Key "*resolution*" increases with:

1) the number of matching fields available

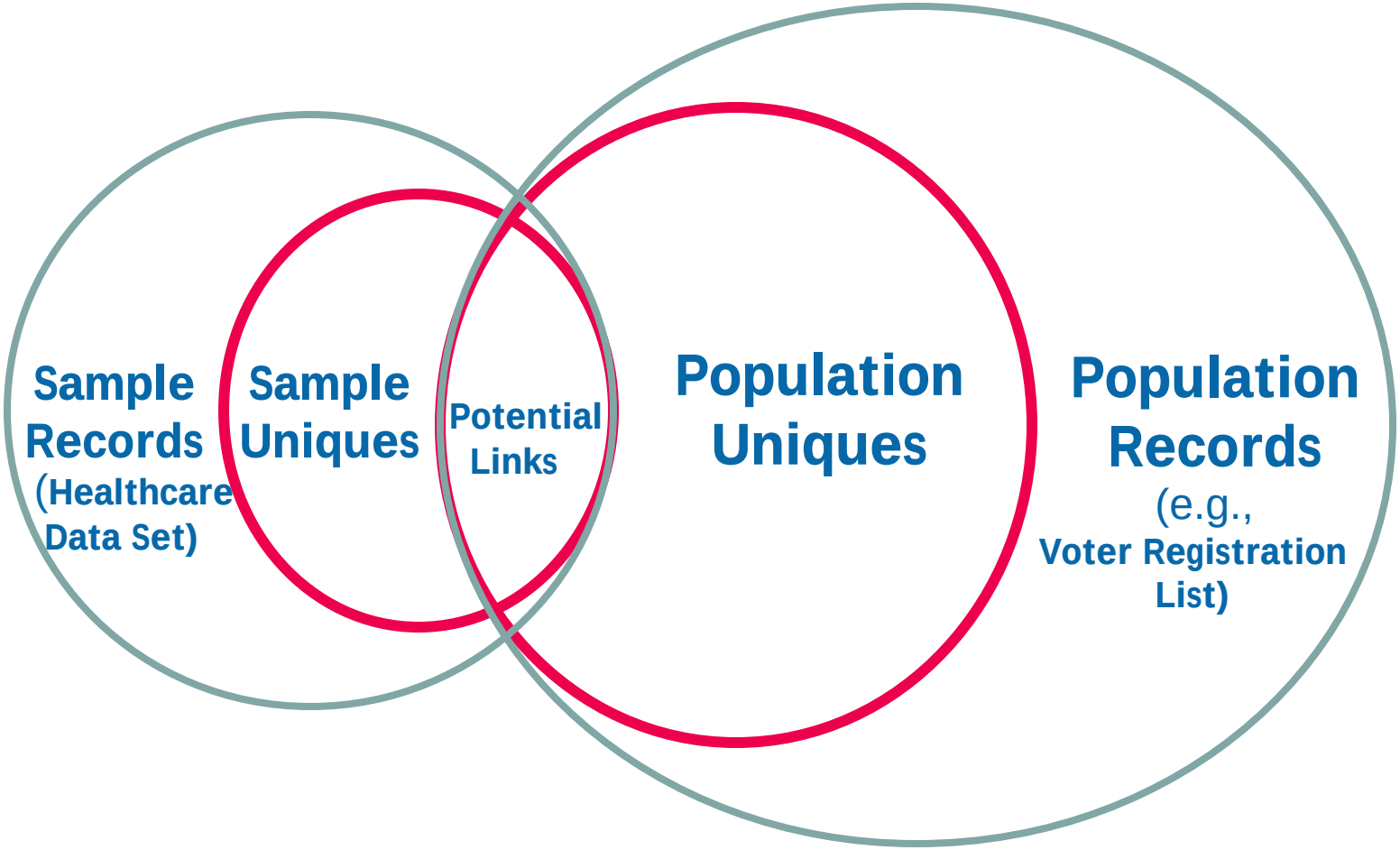
2) the level of detail within these fields. (e.g. Age in Years versus complete Birth Date: Month, Day, Year)



Sample **and** *Population Uniques*

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

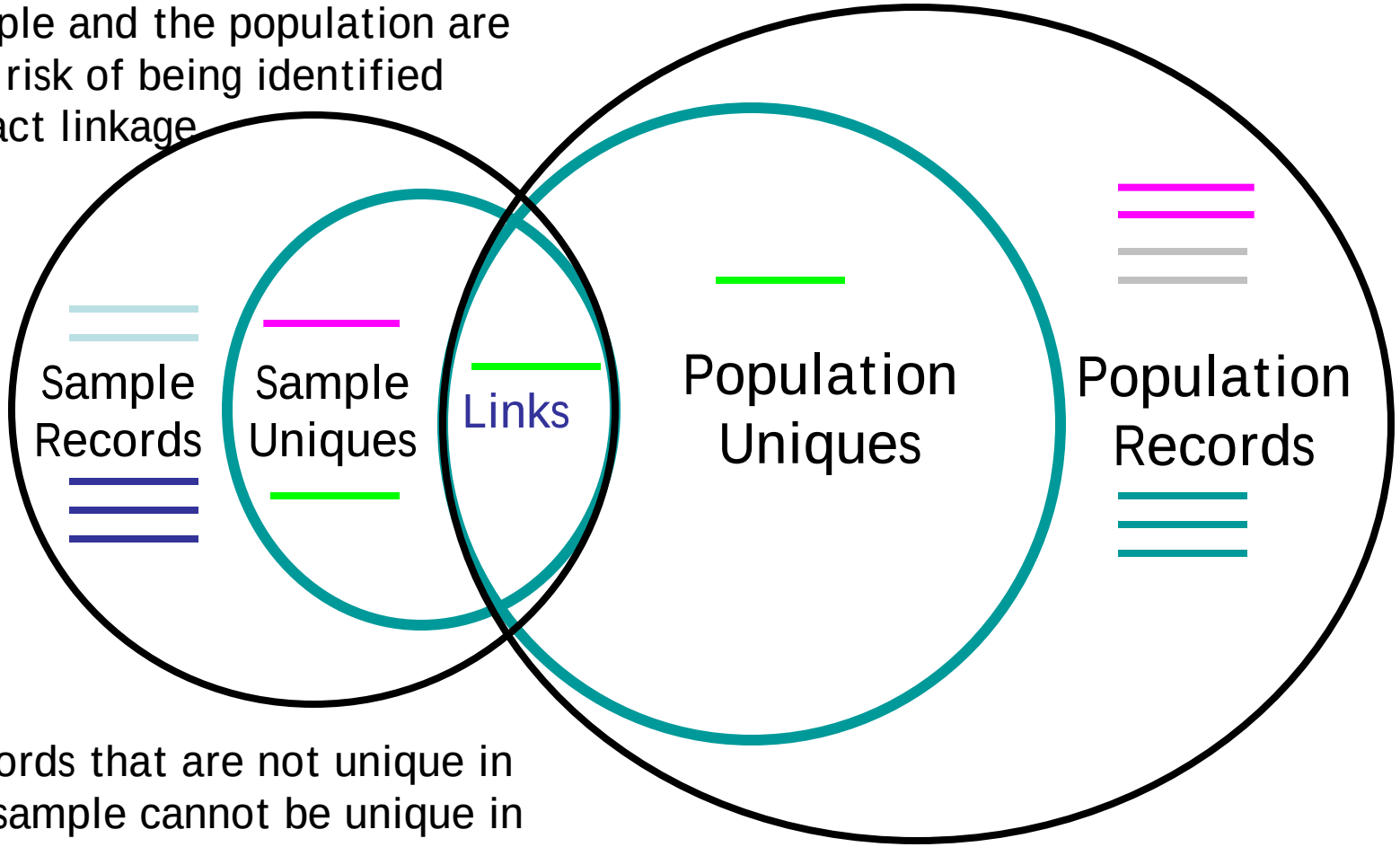
Measuring Disclosure Risks



Linkage Risks

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

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



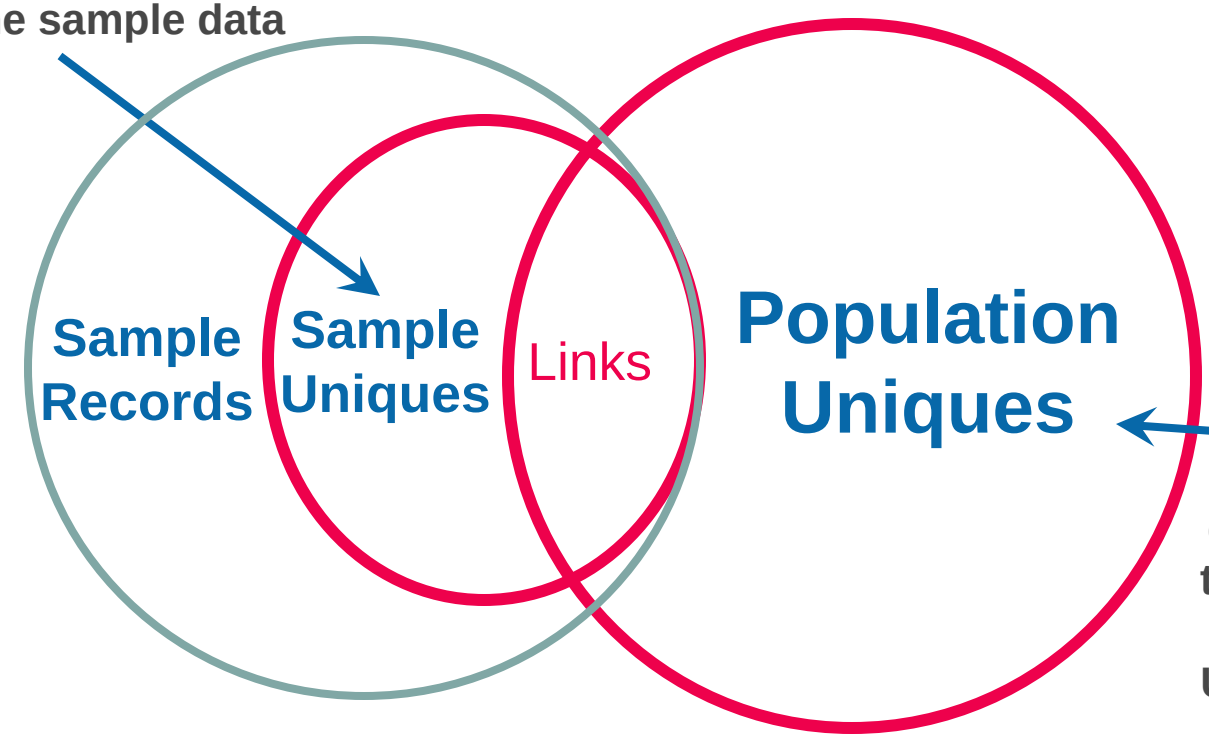
Records that are not unique in the sample cannot be unique in the population and, thus, aren't at definitive risk of being identified

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

Estimating Disclosure Risks

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

Links / Sample Records indicates the risk of record linkage.



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

Successful Solutions:

Balancing Disclosure Risk and Statistical Accuracy

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

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 and computer scientists, but not as well recognized and integrated within public policy.
- Poorly conducted de-identification can lead to "bad science" and "bad decisions".

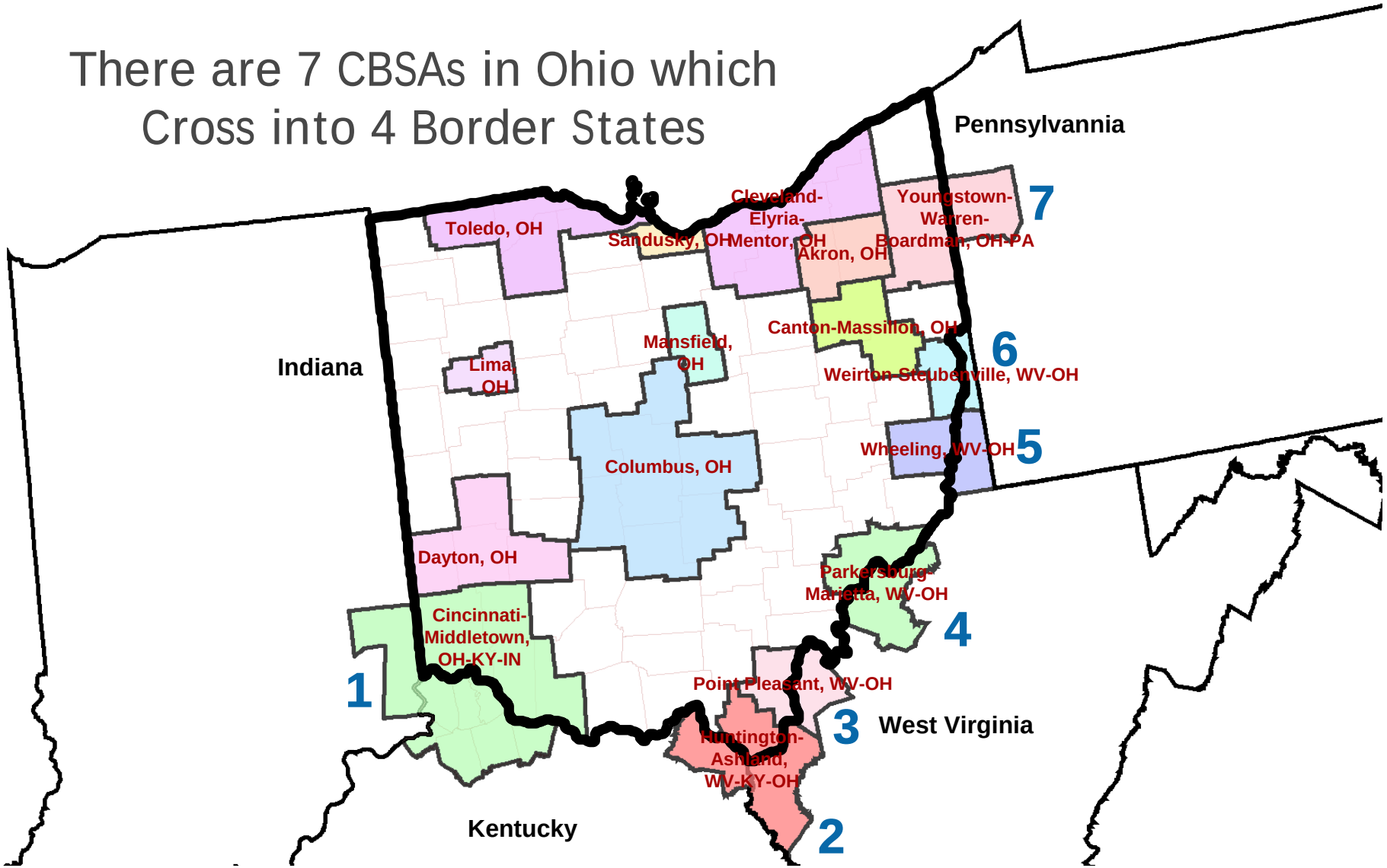
Reference: "*On k-Anonymity and the Curse of Dimensionality*" by C. Aggarwal

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

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.

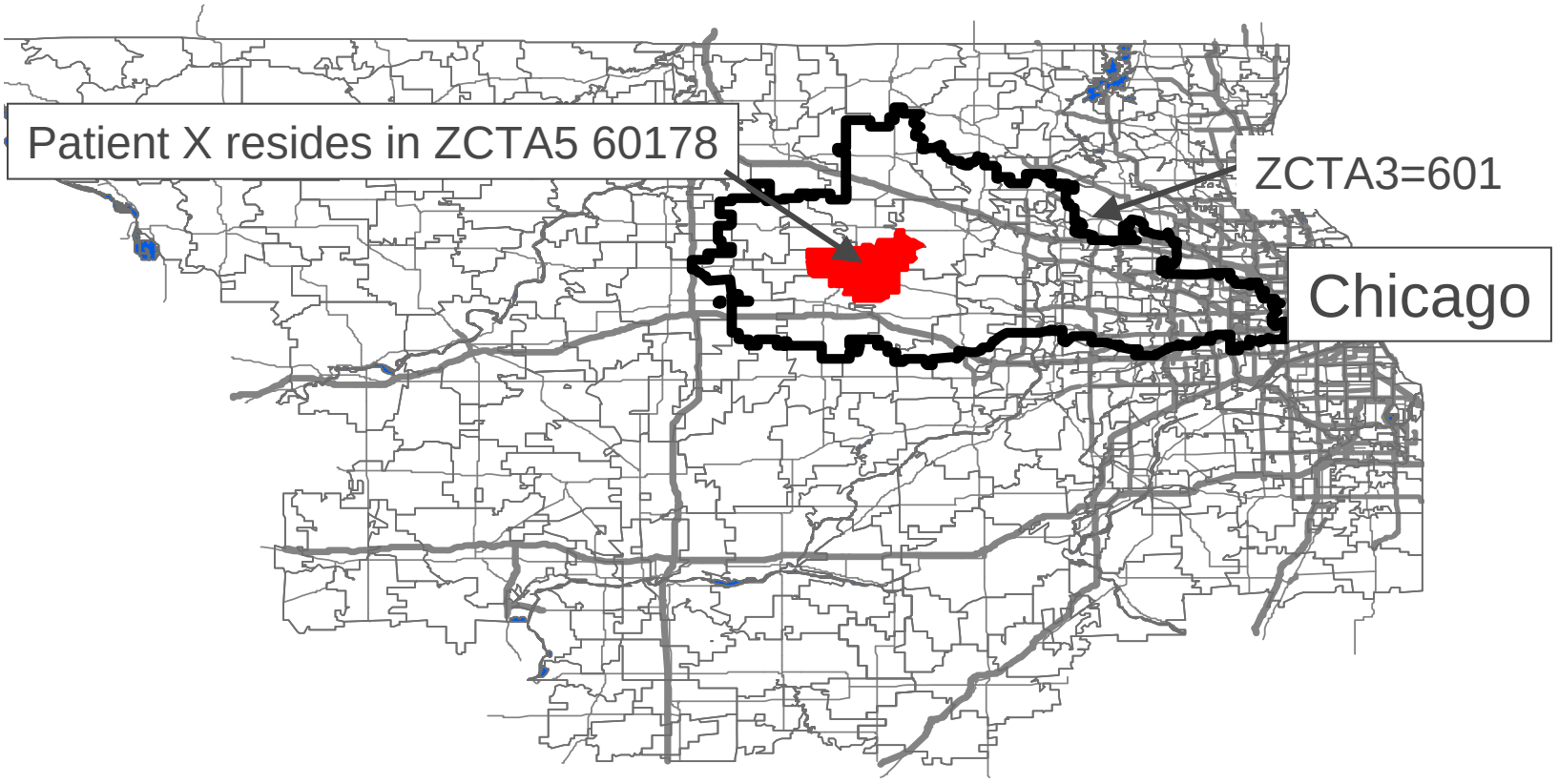
Example: OHIO Core-based Statistical Areas



Challenge: "Geoproxy" Attacks

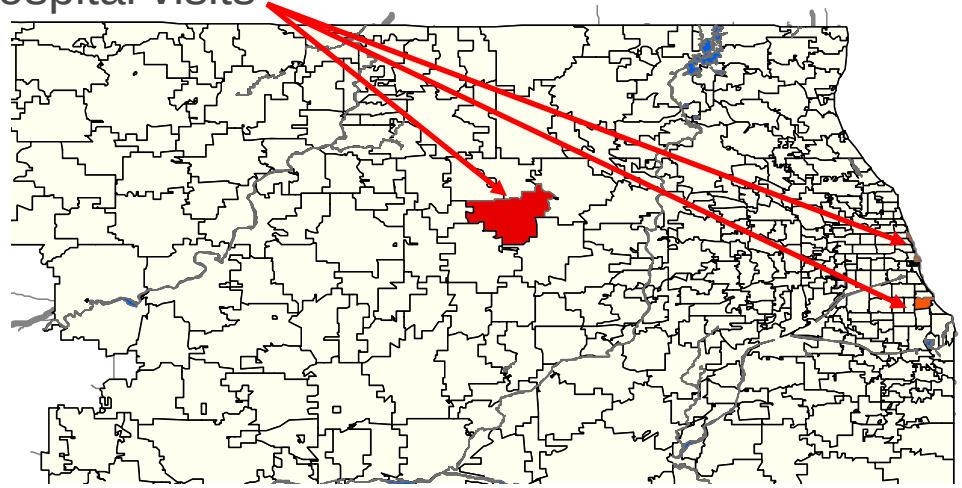
- Challenge: Data intruders can use Geographic Information Systems (GIS) to determine the likely locations of patients from the locations of their healthcare providers
  - Retail Pharmacy Locations
  - Physician or Healthcare Provider Locations
  - Hospital Locations
- *Geoproxy attacks have become much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

Challenge: Geoproxy Attacks



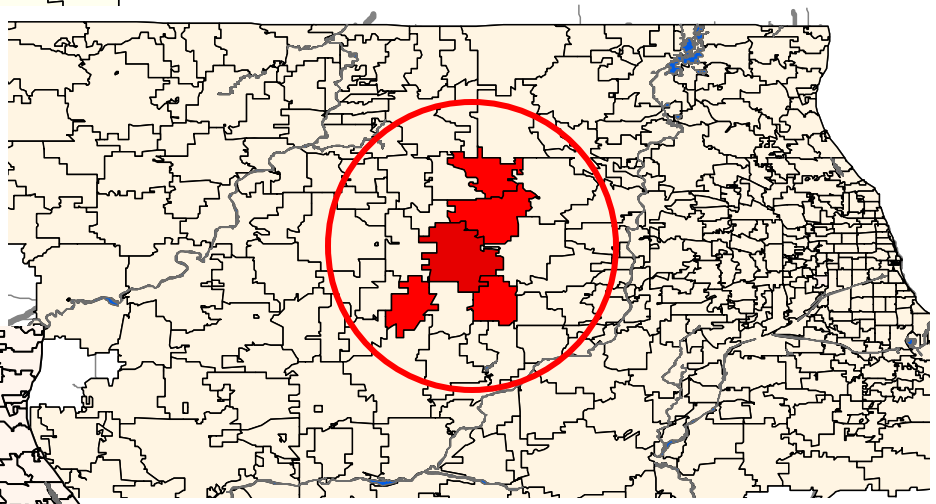
Example: Patient location as revealed within data set, but further narrowed to probable “hotspots” by using healthcare provider location data

Hospital visits

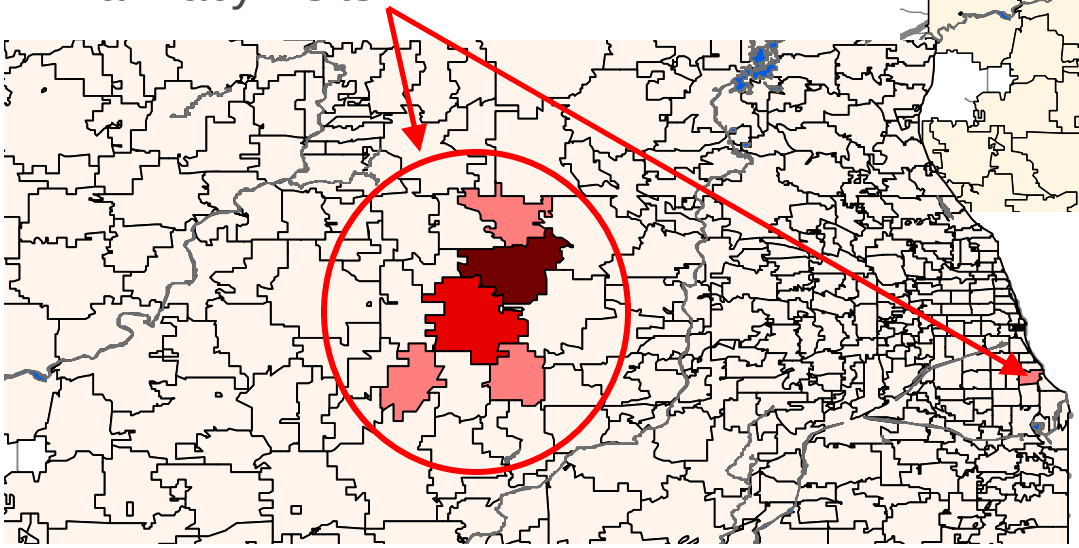


Challenge:
Geoproxy Attacks

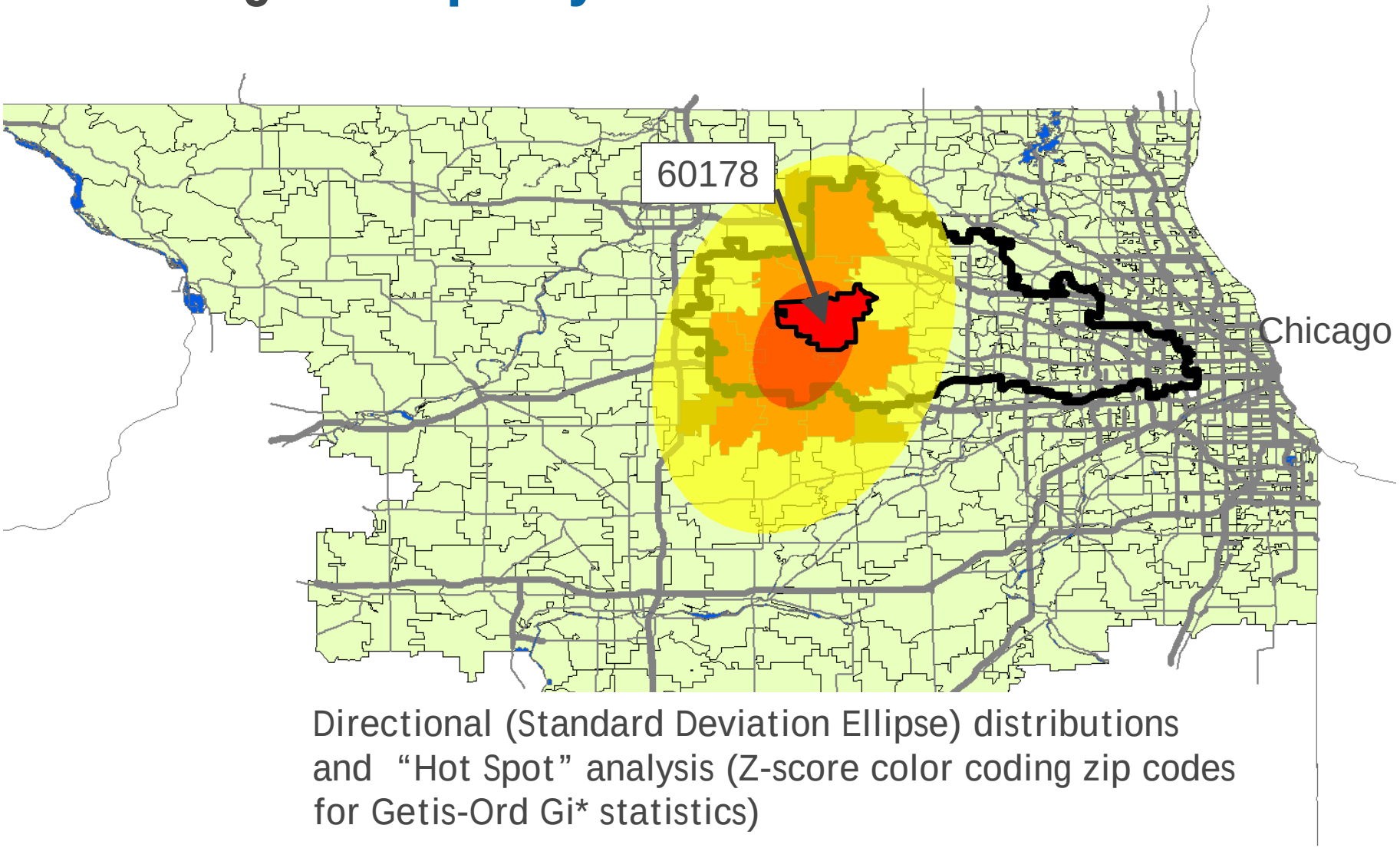
Outpatient/Office visits



Pharmacy visits



Challenge: Geoproxy Attacks

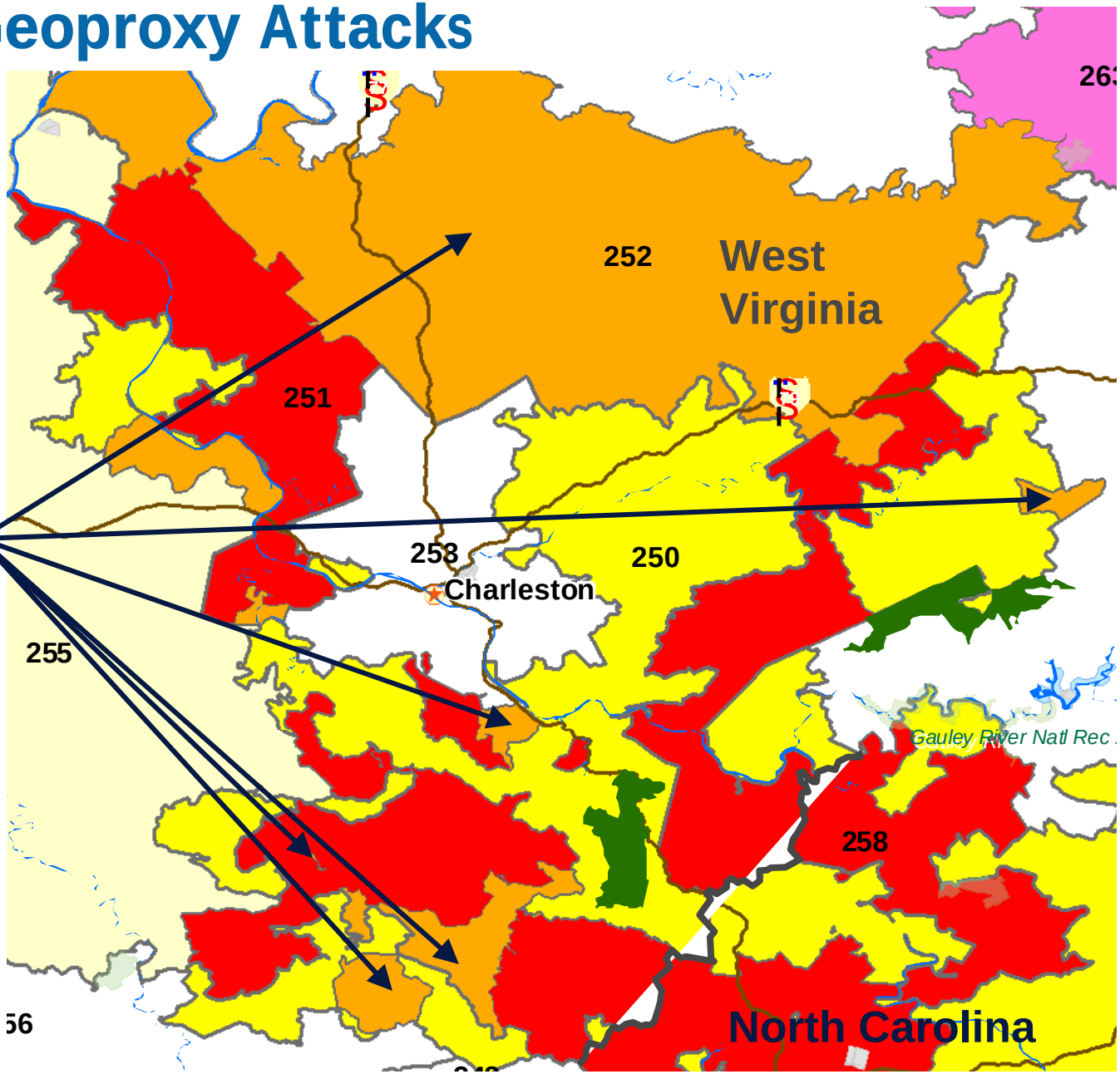


Directional (Standard Deviation Ellipse) distributions and "Hot Spot" analysis (Z-score color coding zip codes for Getis-Ord Gi* statistics)

Challenge: Geoproxy Attacks

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

ZCTA3 252 is highly dispersed

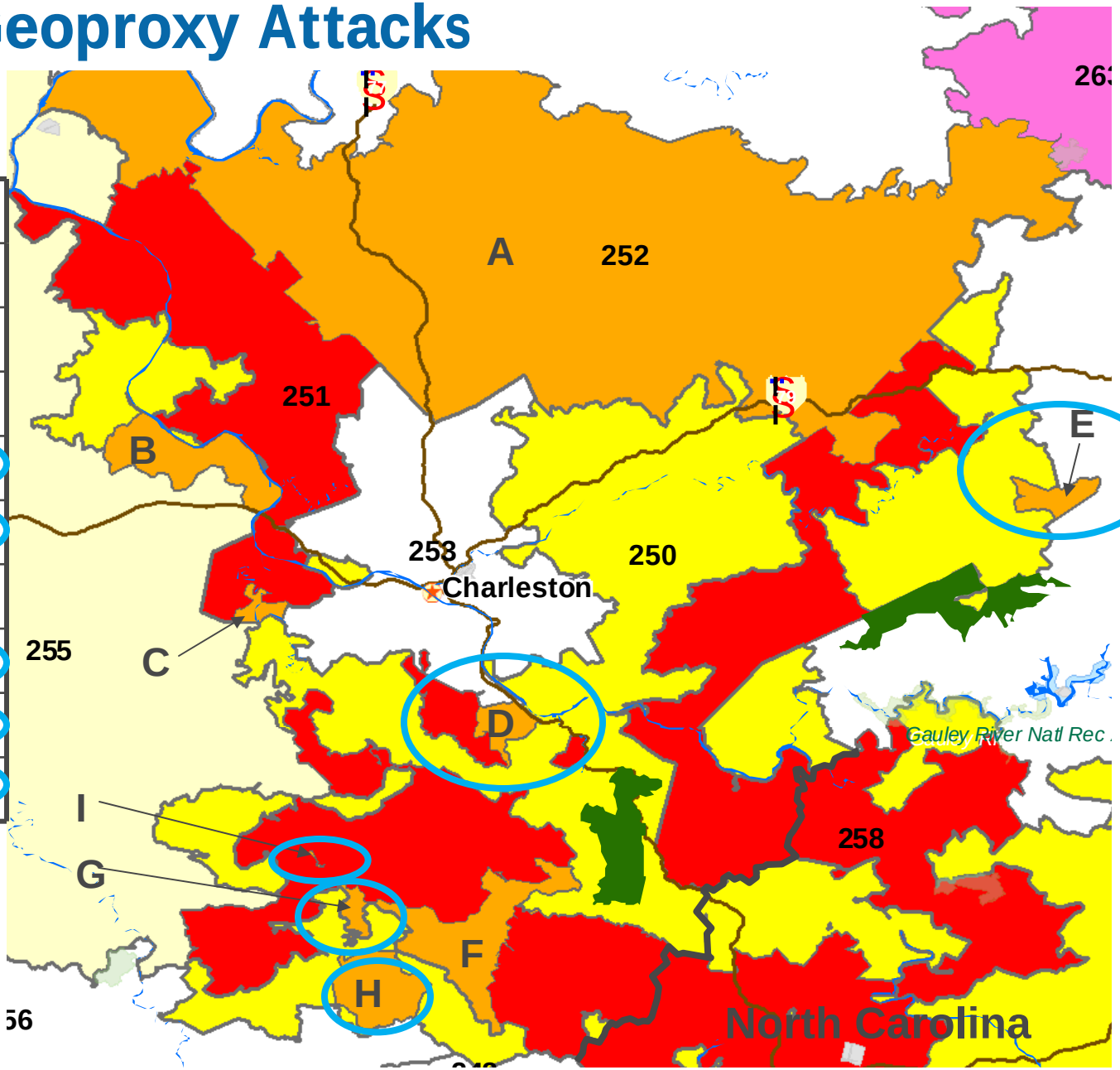


The complexity of 3-digit Zip Code Geography amplifies the threat of Geoproxy attacks

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



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% of US Population (Sweeney, 2000, Golle, 2006)

- 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-identified"… (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" and importance of "Data Divergence"

- The critical part of re-identification efforts that is virtually never tested by disclosure scientists is *assumption of a perfect population register*.
- Probabilistic record linkage has some capacity to dealing 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),
  - —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.

Re-identification Risks in Context:

- The Statistical De-identification provision's *"very small" risk threshold* should *take into account the entire data release context*, including assessment of:
  - —The *anticipated recipients* and the *technical, physical and administrative safeguards and agreements* that help to assure that *re-identification attempts* will be *unlikely, detectable and unsuccessful,*
  - —The *motivations, costs, effort* required and *necessary skills* required to undertake a re-identification attempt.

Suggested Conditions for De-identified Data

Recipients of De-identified Data should be required to:

1) Not re-identify, or attempt to re-identify, or allow to be re-identified, any patients or individuals who are the subject of Protected Health Information within the data, or their relatives, family or household members.
2) Not link any other data elements to the data without obtaining certification 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.

Solutions… Practical and Visionary

- *De-identification offers practical solutions for:*
  - *Avoiding Breaches*
  - *Preserving valuable Date and Geographic Information*
  - *Creating "masked data" for Systems Testing, Development and Demo*
- *The broad availability of de-identified data is an essential tool supporting scientific innovation and health system improvement and efficiency.*
- De-identified data serves 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 while providing strong privacy protections for individuals.

Reserve Slides for Questions

§164.514(b)(2)(i) -18 Safe Harbor Exclusion Elements

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) and..

Safe Harbor Continued…, and §164.514(c)

§164.514(b)(2)(ii) 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.

§164.514(c) A covered entity may assign a code or other means of record identification to allow information de-identified under this section to be re-identified by the covered entity, provided that:

(1) Derivation. The code or other means of record identification is *not derived from or related to* information about the individual and is not otherwise capable of being translated so as to identify the individual; and

(2) Security. The covered entity does not use or disclose the code or other means of record identification for any other purpose, and does not disclose the mechanism for re-identification.

HIPAA §164.514(b)(1) "Statistical De-identification"

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;