Protecting the Privacy of Healthcare Data While Preserving the Utility of Geographic Location Information for Epidemiologic Research

Daniel C. Barth-Jones, M.P.H., Ph.D.
Center for Healthcare Effectiveness Research
Wayne State University
(dbjones@med.wayne.edu)

Alternative Titles:

What are the practical implications of the HIPAA privacy rules for epidemiologic and health services research?

and,

What are the practical implications for the healthcare information industry?

Research under HIPAA

- Research can be conducted with Individual Authorizations.
- Research can be conducted with IRB or Privacy Board Wavier.
- Research can be conducted with Statistically De-identified data.
- Research can be conducted with Limited Data Sets.

"Quasi-Research" and the Healthcare Information Industry

- The healthcare information vendors supply administrative data for a broad range of purposes which might be classifiable as research or healthcare operations or could be achieved with data aggregation :
  - Normative data for healthcare quality and costs
  - Actuarial studies
  - Health systems planning (Where should we place our Doc-in-a-box?).

Logistics of Tracking Data Use

- These activities require that data be shared between healthcare providers and generally have important societal benefits.
- However, the complexity of tracking the myriad uses of administrative data to assure use with HIPAA approved purposes and procedures is a serious logistical challenge.
- De-identification is an attractive alternative because the data can be used for any purpose without restrictions.

Problem with "Safe Harbor" De-identification

The vast majority of data elements specified for deletion under the safe harbor method of de-identification are unimportant for health services research with two exceptions:

- All geographic subdivisions smaller than a state
  - street address, city, county, zipcode, equivalent geocodes (Exception: 3 digit zipcode with >20K population)
- All elements of dates (except year) for dates related directly to an individual, including birth date, admission date, discharge date, date of death, ages over 89 years old.

- Elimination of dates and geographic information destroys a great deal of the utility of PHI for many purposes.

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;"

Limited Data Set

- The Limited Data Set approach permits uses and disclosures of a limited data set which does not include ***facially identifiable information*** (i.e., direct identifiers) for research, public health and healthcare operations, conditional on there being a data use agreement in which the data recipient agrees to:

  a) limit the data use to those purposes permitted in the privacy rule,

  b) limit who can use or receive the data, and

  c) not re-identify the data.

Limited Data Set

- The limited data set may include:
  - Admission, Discharge and Service Dates,
  - Date of Death,
  - Age (including age 90 or over), and
  - 5 Digit Zip
  - Any other geographic subdivision, such as State, county, city, precinct and their equivalent geocodes, (except for street address or prohibited postal information).

Standing Question:

The limited data set retains most of the data elements needed for conducting analyses with administrative healthcare data.

So,

What is the risk of identification for the limited data set (or very similar data sets)?

In particular,

Is it reasonable to retain the 5 Digit Zipcode in combination with other important demographic characteristics (e.g., age, gender, family key) or is this level of geographic specificity responsible for too great a level of disclosure risk?

Sweeny Results:*

Zip code & Birthdate = 69% uniquely identified

Sweeny L. J Law Med Ethics. 1997; 25:98-110

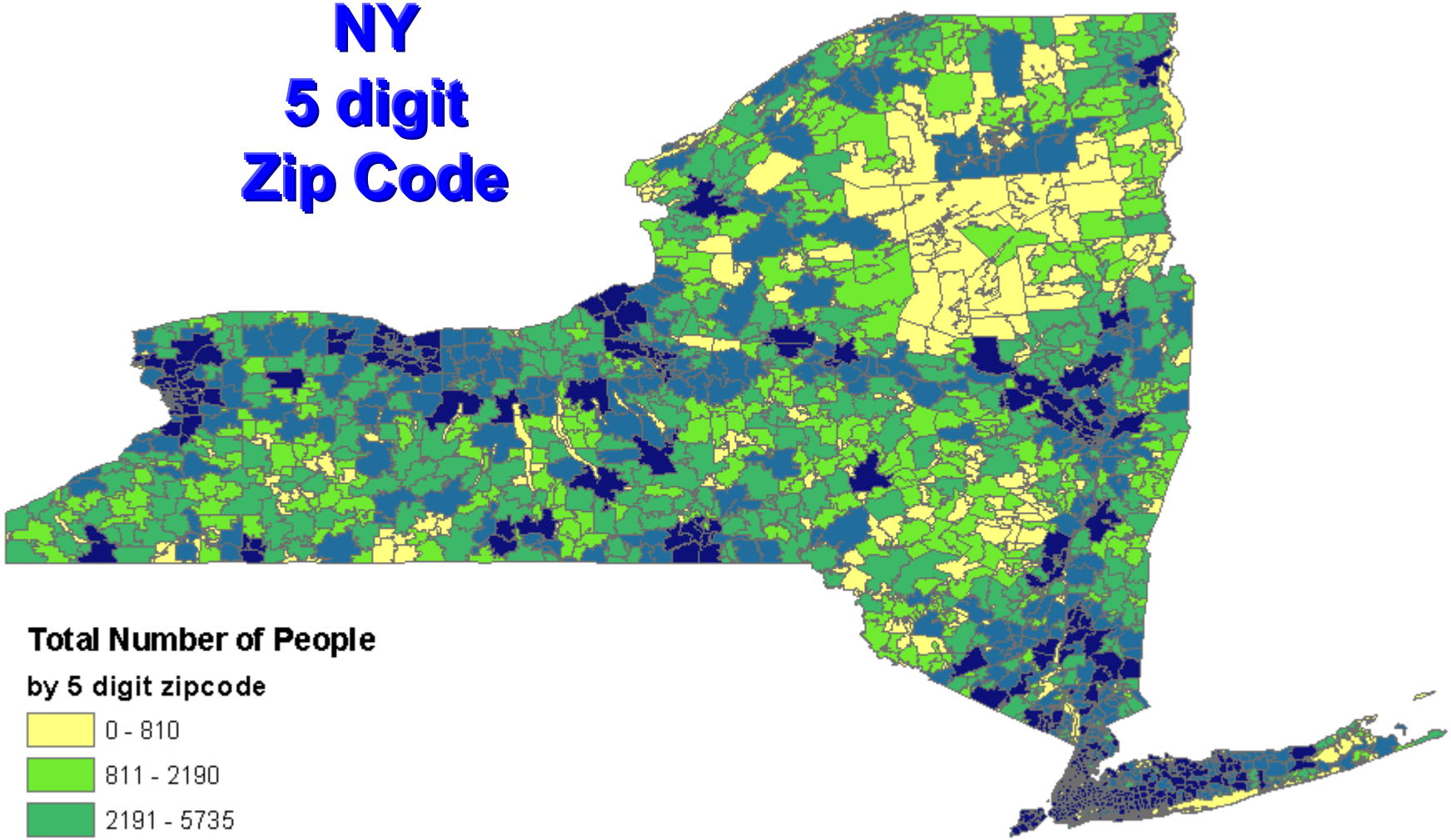
Thinking Critically about Zip Codes

- Zip codes were created for the purpose of mail delivery and follow street routes.
- Zip codes are subject to frequent updating by the postal service.
- Zip codes do not have a neat relationship to city or administrative boundaries.
  - Multiple Zip codes per city
  - Zip codes can divide census blocks.

Thinking Critically about Zip Codes

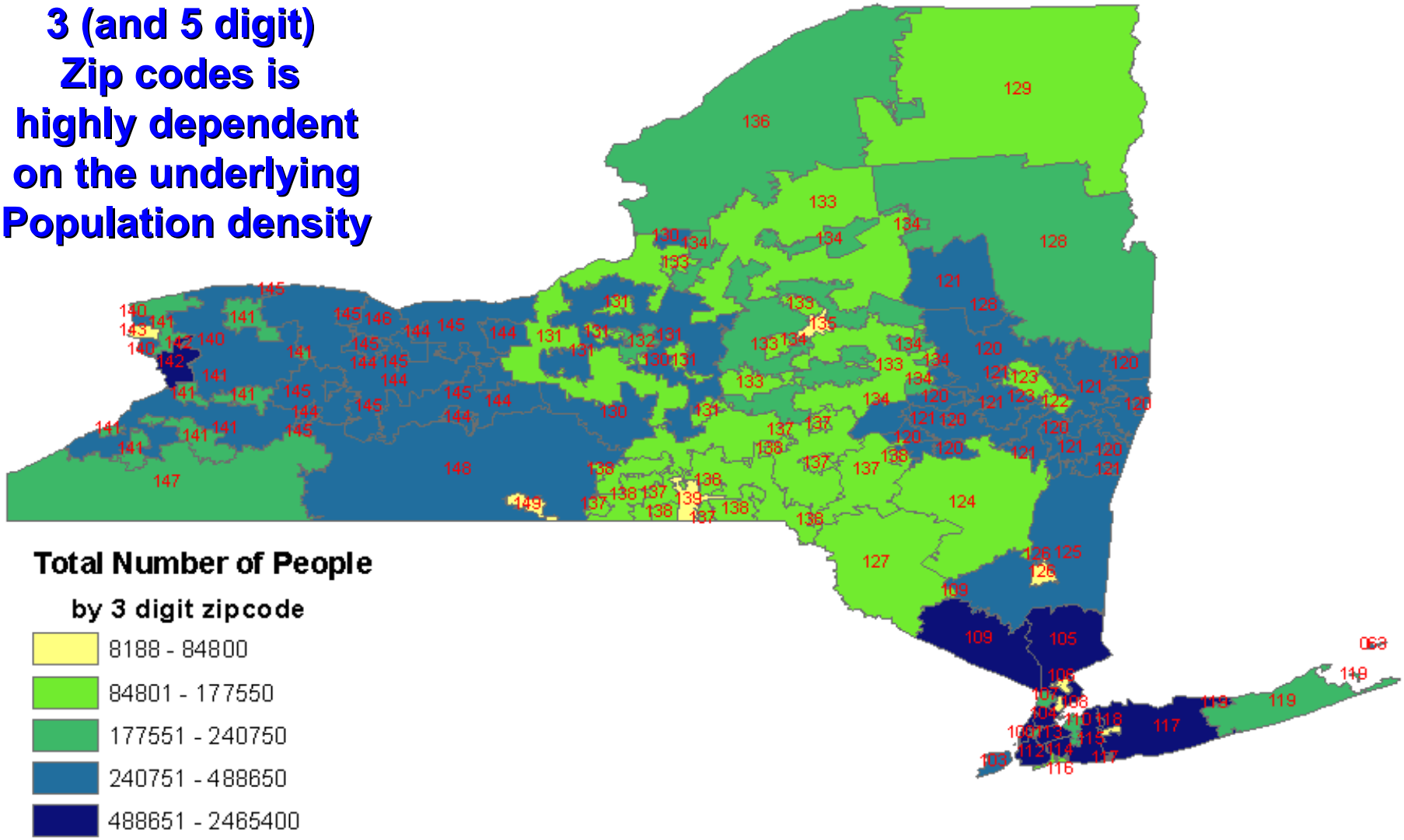
- However, Zip codes are the smallest geographic subdivision routinely collected (aside from the possibility of geocoding street address information).
- Desire to retain smallest geographic units available for flexibility for possible analyses.
- 3 digit zip code roll up is thought to provide aggregation too large for analyses addressing disease clusters or location of health facilities.
- Demand for 5 digit Zip code data is strong.

New York Population - 2000



Note that the geographic area of 3 (and 5 digit) Zip codes is highly dependent on the underlying Population density

New York Population - 2000



Thinking Critically about Zip Codes

- The arbitrary formulation of Zip codes with respect to the proxy variables they substitute for (SES, income, education, housing, geographic location and distance, etc.) most likely means that use of zip codes will aggregate many of these characteristics into heterogeneous groupings.
- Utility of zip code data needs to be more critically evaluated.

Analyzing HHS Rationale for Permitting 3 Digit Zip Code

(FedReg Dec 28, 2002 p.82711)

- "This will result in an average 3-digit zip code area population of 287,858 which should result in an average of about ***4% unique records*** using the 6 variables described above from the Census Short Form. ***Although this level of unique records will be much higher in the smaller geographic areas, the actual risk of identification will be much lower because of the limited availability of comparable data in publicly available, identified databases, and will be further reduced by the low probability that someone will expend the resources to try to identify records when the chance of success is so small and uncertain.***"

Analyzing HHS Rationale

- Probability of Disclosure Potential

  is only the first part of the previous equation for statistical disclosure risk assessment in the justification HHS provides for choosing the 3 digit zip code roll-up.

- The remainder of the equation is: ***"...will be further reduced by the low probability that someone will expend the resources to try to identify records..."***

  i.e., …. Probability of External Data Availability for Record Linkage, Probability of Necessary Computing Resource, Probability of Expertise Needed to Conduct Record Linkage, etc.

Analyzing HHS Rationale

- The actual risk of identification is dependent on:
  - Probability of disclosure potential,
  - Availability and expense of external data for record linkage,
  - Expertise needed to conduct record linkage
  - Necessary computing resources,
  - Time required for conducting data intrusion,
  - Personal risk involved in conducting data intrusion.

Classifying Variables

- Identifying Variables
  - Name, SSN, Address etc. (*Presumably these are already removed from the sample data*)
- Key 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.) for any significant number of individuals.

Conceptualizing Data Intrusion

- What is the “Data Intruder” trying to do?
  - **Looking for a specific “Target” Person**
  - **On a “Fishing Exposition” to identify whomever can be identified.**
- What does the “Data Intruder” know about the sample to population relationship?
  - **Target Person(s) exists in the Population**
  - **Target Person(s) in the Sample Data**
  - **Intruder knows which record(s) in the sample belong to the Target Person(s).**

Conceptualizing Data Intrusion

- *Healthcare data can not be made totally free of identification risk* and still be useful,but it is possible to make most disclosures so difficult to achieve that it isn't worth the bother.
- Part of your "Due Diligence" is finding out what key variables exist in data sets that are available for your data population:
  - Census Data
  - Voter Registration
  - Driver's License
  - Government Surveys
  - Marketing Data
  - Etc.

Conceptualizing Data Intrusion

- Data Intrusion Scenario Example
  - We conservatively estimate the number of persons for which each data intruder might possess information held in confidential variables and how many confidential variables for which they might have information.
    - Example: Each data intruder is assumed to know exactly at most *x* confidential variables (Hospital, Service dates, Dxs, Pxs, etc.) for at most *y* people.

Conceptualizing Data Intrusion

- Because Confidential variables are not typically known for very many target persons in a dataset, and the majority of data intruders are technically capable of only simple query intrusions (or, more rarely, exact record linkage intrusions), *Confidential variables typically pose a reasonably small risk of identification* in large data sets.

Conceptualizing Data Intrusion

- A reasonable and realistic assessment of your statistical disclosure risks will include:
  - Conducting Statistical Disclosure Risk Analyses
  - Formulating a comprehensive set of Data Intrusion Scenarios
  - Estimating (conservatively) the “costs and availability” of the required data intrusion resources
  - Calculating the “real” risk of disclosure given the associated costs, etc.
  - Providing a well-reasoned and clear justification of your case that the risk of identification is “reasonably small”.

Key Variables

- Because our focus is on external data that is "*Reasonably Available*" to data intruders, our disclosure risk analyses focus on demographic variables in public datasets such as:
  - Voter Registration Lists,
  - Department of Motor Vehicle Registration Data,
  - Marriage License Data,
  - Birth Records,
  - Death Records.

Key Variables

- Based on the variables that are commonly found in these public datasets, the following variables were identified as key variables that should be analyzed in Disclosure Control Analyses:
  - Date of Birth/Age
  - Gender
  - Zip Code
  - Family or Household code?
  - Physician or Facility codes?

Exact versus Probabilistic Record Linkage

- Because record linkages made by a data intruder using probabilistic record linkage are subject to uncertainty, it is reasonable to base disclosure limitation analyses on probability models describing Exact record linkage methods.

Estimating Disclosure Risks

Bin Analysis					
	Age Groups	Gender	Zip Code	Bins	Persons Per Bin
5 digit Zip code & Gender	36,500	2	32,038	2,338,774,000	0.12
Age in Yrs Up to 90 & 90+, 5 digit	91	2	32,038	5,830,916	48
Safe Harbor	91	2	887	161,434	1,747
37 Age Groups, 5 digit Zipcode	37	2	32,038	2,370,812	119

Sample Uniques and Population Uniques

Voter Registration Record								
Name	Address	City	State	Full Zipcode	Birth Date	Gender		
Richard Notreal	23 Someware Blvd.	Decatur	GA	30033-5637	12/4/1963	M		
				Medical Record Data (Stripped of Obvious Identifiers)				
				Full Zipcode	Birth Date	Gender	Admission Date	Principle Dx Code
				30033-5637	12/4/1963	M	8/18/2002	042

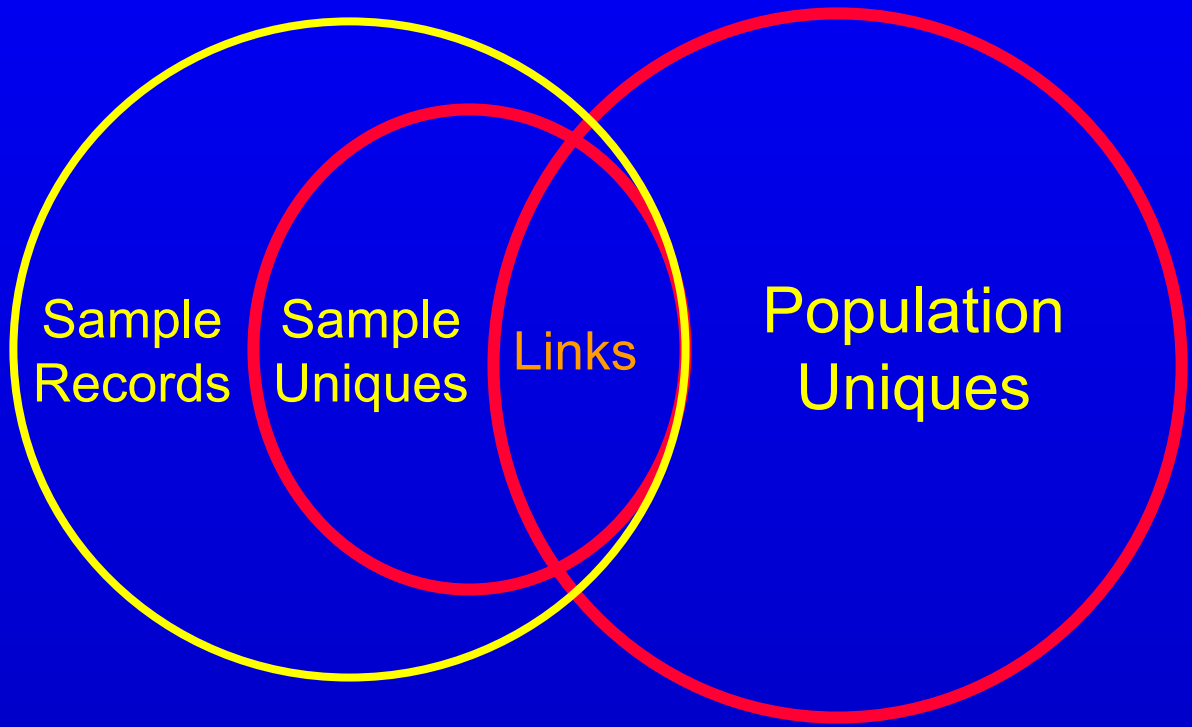
- Exact record linkage is possible only when a set of key variables for an individual combines uniquely to identify the individual in both the sample database and the population database.
- Furthermore, the key variable data must not have ***errors*** due to ***time dynamics*** or ***recording errors*** that will cause the link to fail.

Possible Disclosure Risk Measure

The proportion of sample uniques that are population uniques. (Zayatz 1991, Greenberg & Zayatz 1992)

- Because an individual in the sample can not be a population unique if the individual is not unique in the sample, this measure calculates disclosure risk only among sample uniques.
- Note that this measure does not reflect the disclosure risk for the sample, but rather the disclosure risk for the sample uniques.

Proposed Disclosure Risk Measures



- In other notation, Links / Sample Uniques can be denoted as:

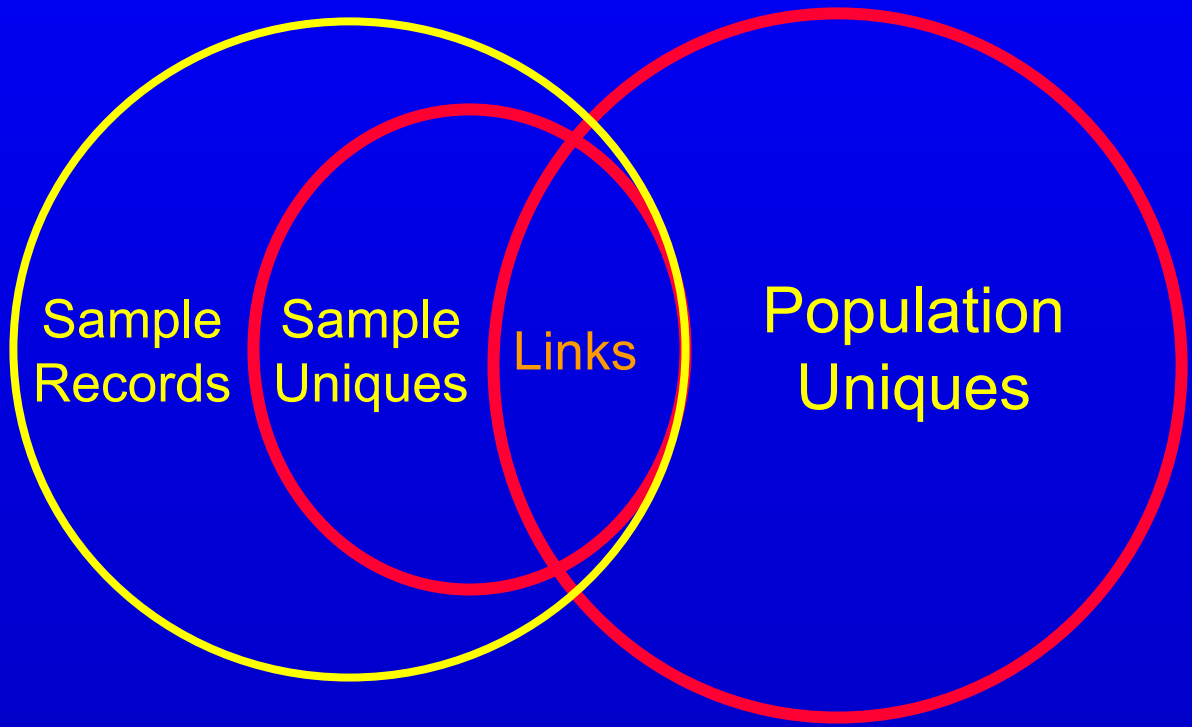
  P(PU | SU) the probability of a record being a population unique, given that it is a sample unique.

Disclosure Risk Measure

The proportion of sample records that are population uniques. (Bethlehem et al. 1990)

- Because the percentage of sample records that can be linked to population uniques indicates the risk of record linkage for a sample record, the percentage of population uniques in the sample most accurately indicates the identification disclosure risk for a sample.

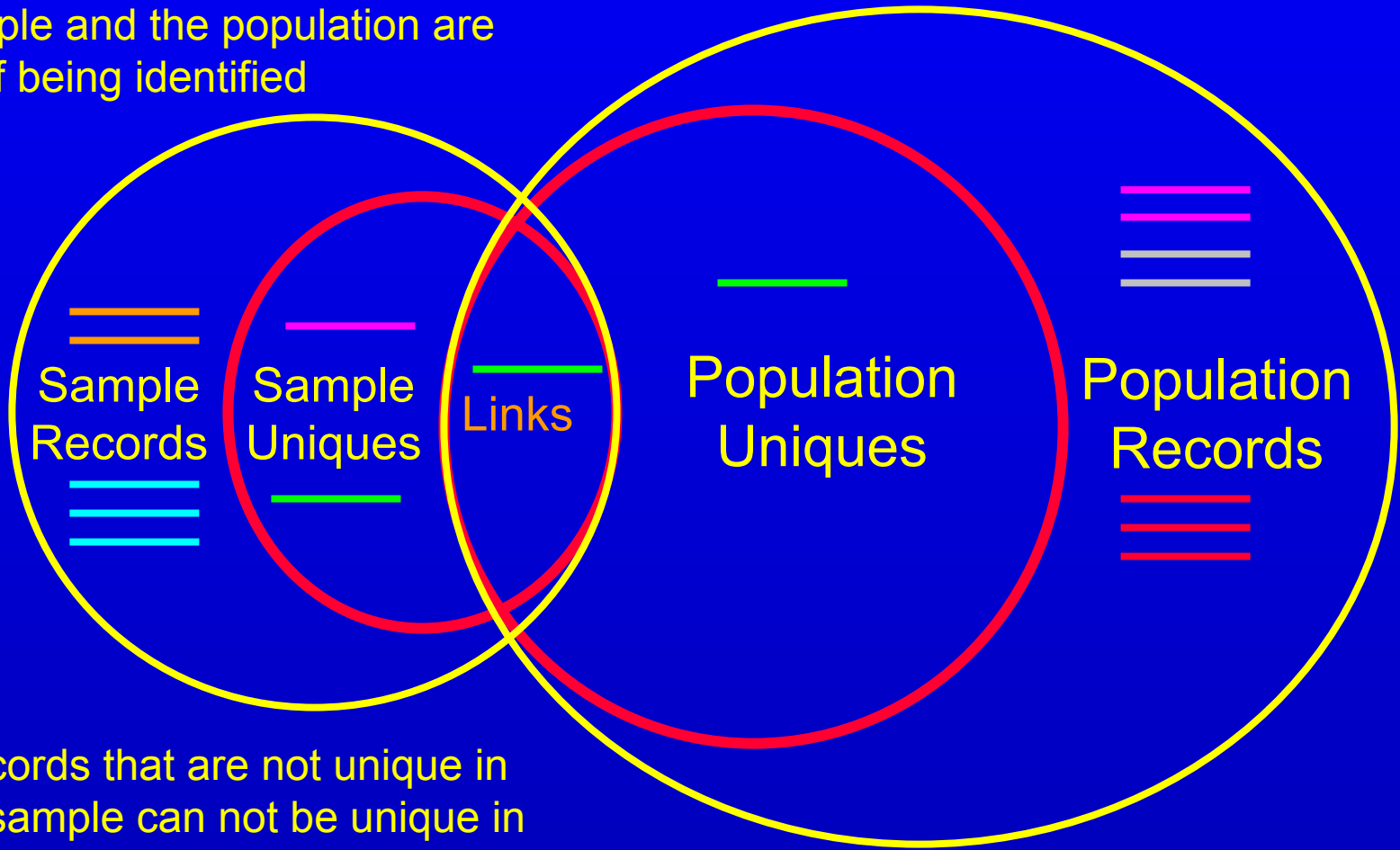
Proposed Disclosure Risk Measures



- The percentage of sample records that can be linked to population uniques is an ideal disclosure risk measure for record linkage risks: Links / Sample Records indicates the risk of record linkage for a sample record.

Records that are unique in the sample but which aren't unique in the population, and, therefore, would match with more than one record in the population, also aren't at risk of being identified

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



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

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

Measuring Disclosure Risks

- For the moment, we will ignore the complicating issues of real world record linkage:
  - Our sample will frequently not have been drawn from the population using probabilistic mechanisms resulting in question about the representativeness of the sample for the population
  - Errors due to Time Dynamics will affect matching
  - Recording Errors due to will affect matching
  - It is not usually possible to get complete census data, so incomplete data is used to attempt record linkage.

Estimating Disclosure Risks

- We define those categories that have at least one observation as an "Equivalence Class" because all individuals in a equivalence class are equivalent with regard to these variables. (Zayatz 1991)

Estimating Disclosure Risks by Record Linkage

- Disclosure Risk as measured by Links/Sample Records can be estimated by conducting a Record Linkage experiment, replicating the actions that would be undertaken by a data intruder.
- However, conducting record linkage experiments is expensive and time-consuming and, therefore, not feasible for monitoring frequent releases of data.

Estimating Disclosure Risks

- Fortunately, if our sample is representative of the population, then one option is to use statistical estimation methods to estimate the number of population uniques from the sample data. (Chen et. al. 1998,)
- It is useful to distinguish between sample uniques which have a high probability of also being a population unique and sample uniques that are unlikely to be population unique. (Elliot et al., 2001)

Sample Uniques and Population Uniques

- Methods for estimating population uniques from sample data:
  - Equivalence Class Procedure (Zayatz 1991a, 1991b, Greenberg et al 1992)
  - Poisson-Gamma Model (Bethlehem et al. 1990, Keller et al. 1992, Skinner et al. 1994)
  - "Slide Negative Binomial" Method (Chen et al. 1998)
  - Data Intrusion Simulation (Elliot, 2000).

Equivalence Class Method

- An Equivalence Class is simply a non-empty cell with a size equal to the size of the cell.
- Developed under the assumption of simple random sampling.
- According to Bayes' rule, the conditional probability that an observed equivalence class of size one in the sample came from a population equivalence class of size one is:

$$P(1_p|1_s) = \frac{p_1 P(1_s|1_p)}{\sum_{allj} p_j P(1_s|j)}$$

Equivalence Class Method

$$P(1_p|1_s) = \frac{p_1 P(1_s|1_p)}{\sum_{allj} p_j P(1_s|j)}$$

where p_j is the proportion of equivalence classes of size j in the population, and $P(1_s|j)$ follows a hypergeometric distribution for all j's. The total number of population uniques, U_p, can be estimated based on the estimate of this probability. The population proportions p_j's are estimated by the sample proportions c_j/k, for all j's; where c_j is the number of cells of size j in the sample.

Equivalence Class Method

- The Equivalence Class works fairly well for large sampling fractions (i.e., $f > 0.1$), but for small sampling fractions the procedure dramatically overestimates the number of population uniques (thus overestimating the disclosure risks).

Estimating Disclosure Risks

For some obviously good reasons, the Census Bureau does not release exact information on the combination of Date of Birth, Gender and Zip Code…

- They have, however, released Table PCT12 in the Census 2000 100 percent Short Form SF1 data release. This table provides the Age and Gender breakdowns for each ZCTA.
- To protect against data intrusion, a technique called "*Data Swapping*" has been used on the original data before it was released.

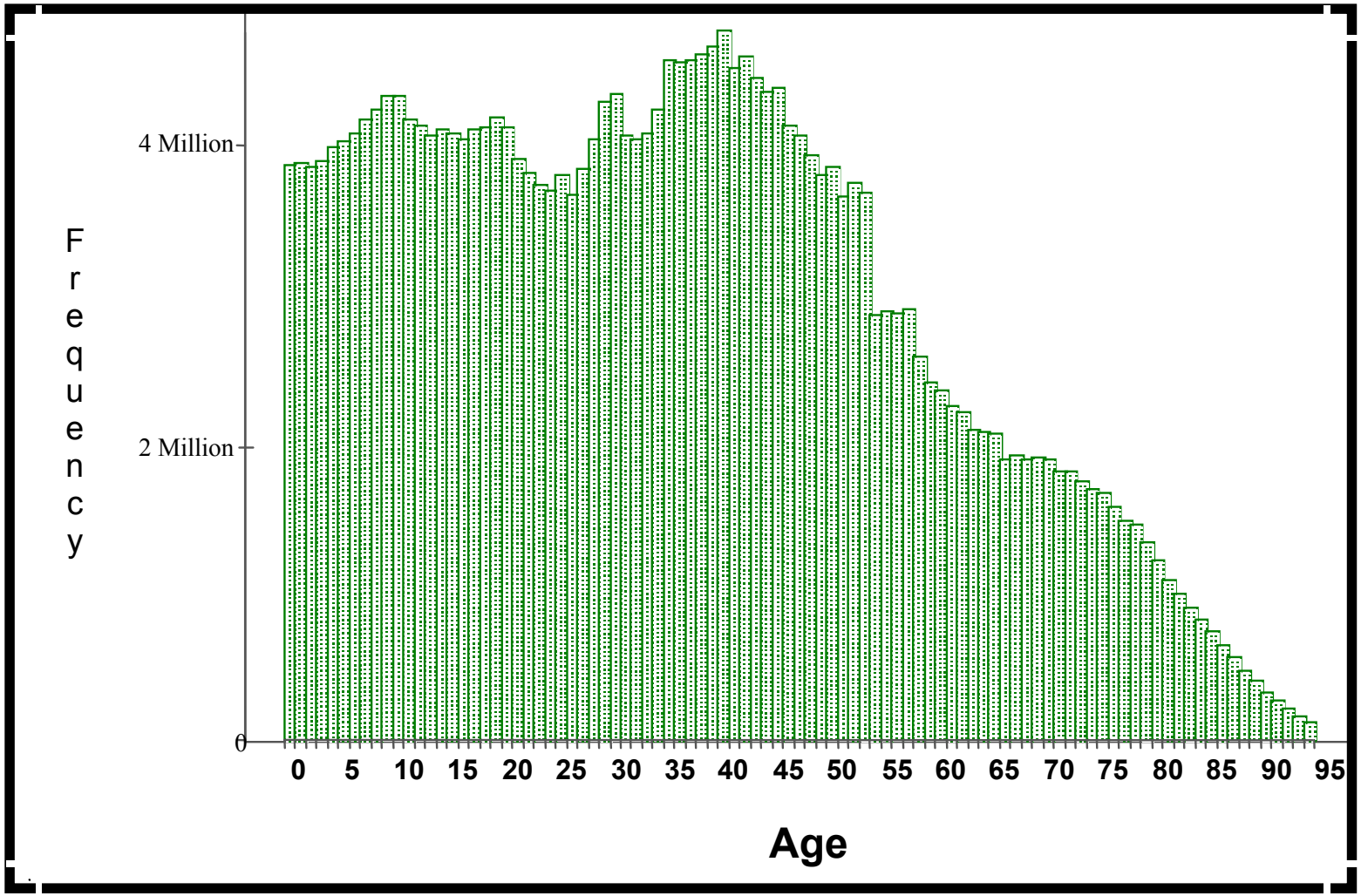
Sidebar: Data Swapping

- We are interested only in the statistical relationships between the variables age, gender and ZCTA and how these combine to create population uniques.
- So, while the specific locations, ages and genders of the population uniques in the Census data may not be precisely accurate, this data is appropriate for our purposes because the manner in which Data Swapping is performed is designed to preserve the marginal distributions and the local associations between age, gender and location.

Estimating Disclosure Risks

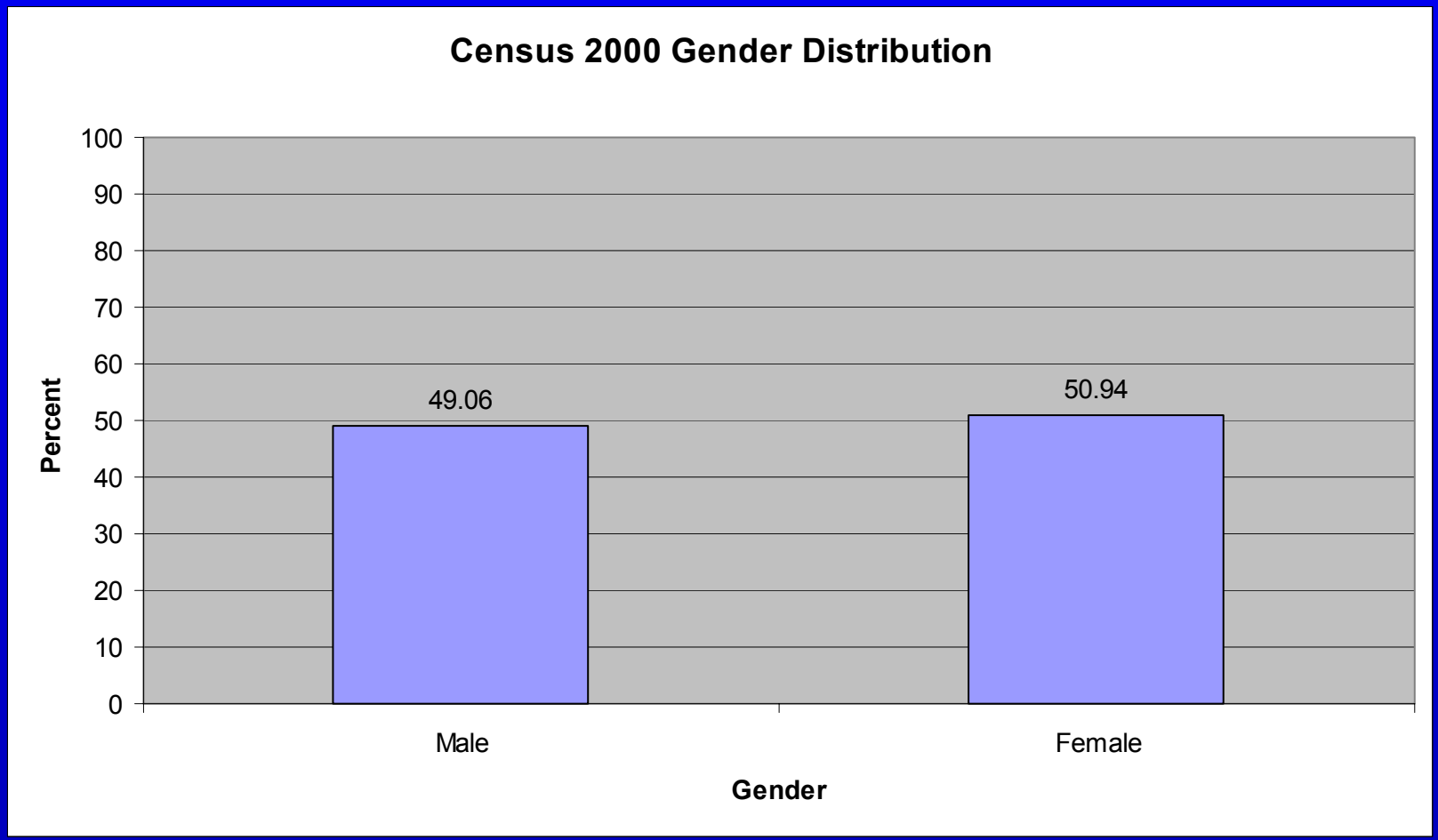
- Because we have the Census data available, one possible approach to estimating the percentage of population uniques for variables collected by the Census is to estimate the expected number of individuals in each potential category from the marginal distributions for the variables.
- This method of estimating the percent of population records that are population uniques treats the number of individuals in each equivalence class as a random variable for which we know the expectations, but not the actual values.

Marginal Distribution: Age

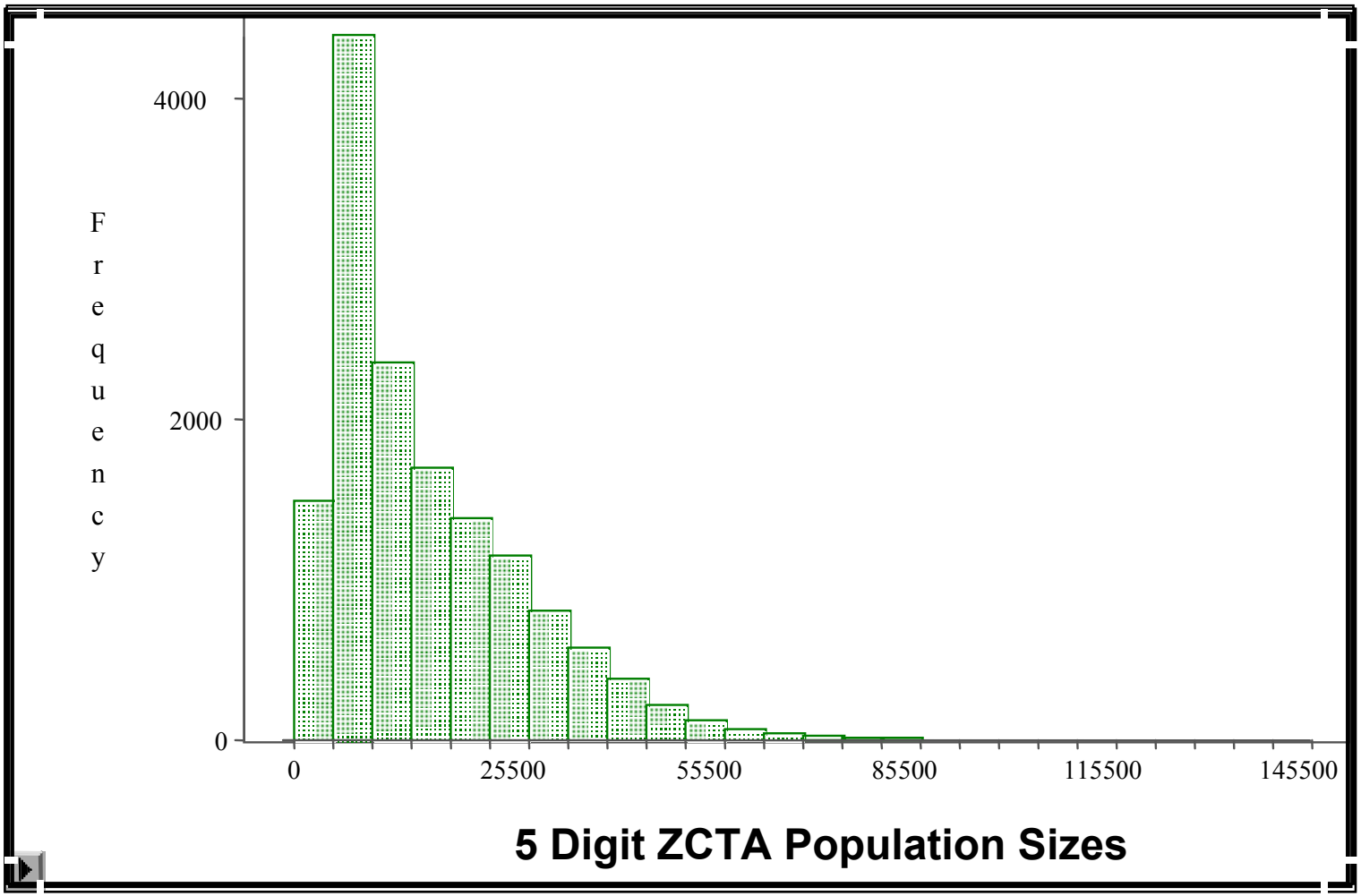


Data Source: U.S. Census 2000 PCT12 Table 100% SF1

Marginal Distribution: Gender



Marginal Distribution: Zip Code Size

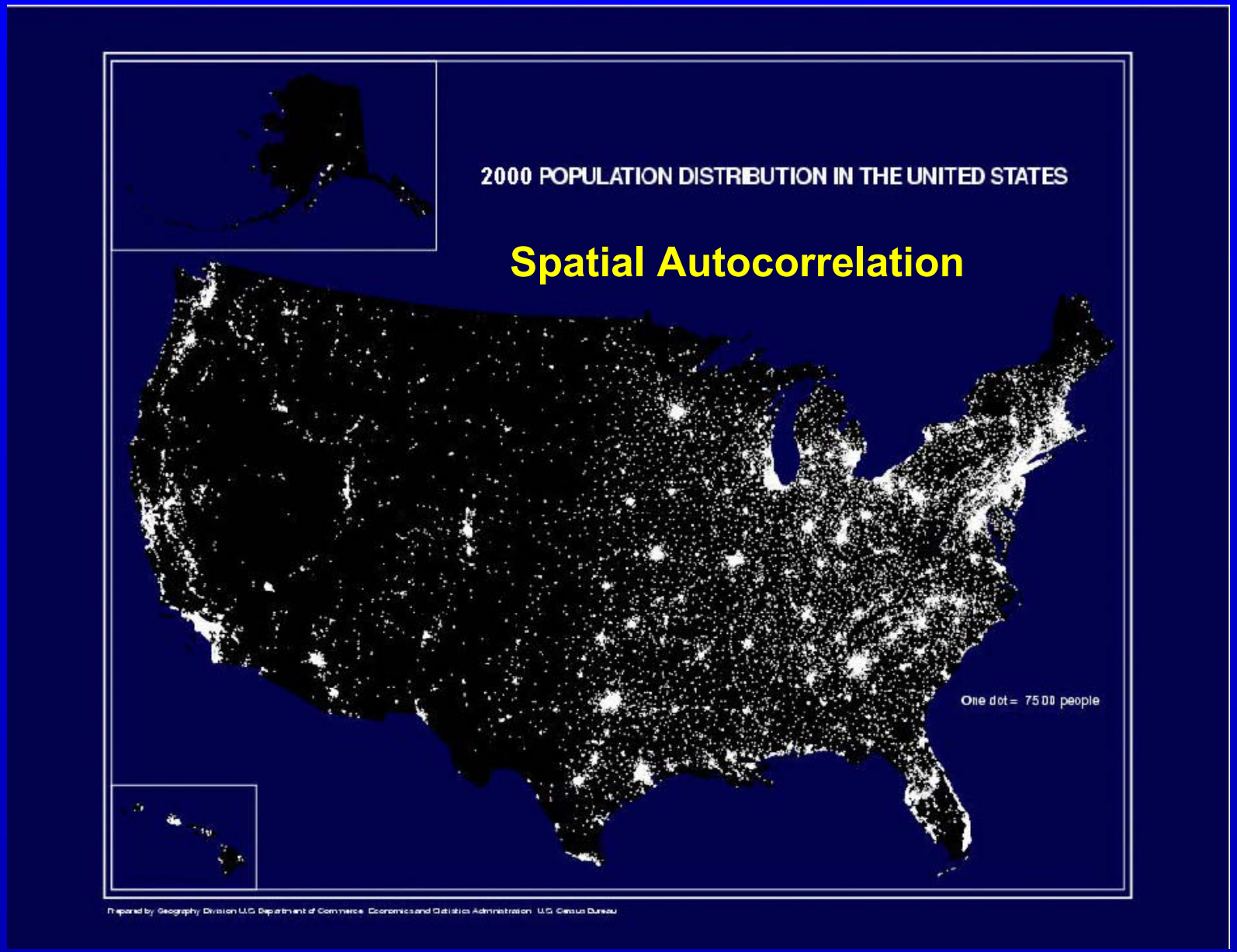


Estimating Disclosure Risks

- Under the assumption that there is no association between these characteristics, we can determine the expected number of individuals in each equivalence class by multiplying the marginal distributions that cross-classify the equivalence classes and the total population size

  - $E[n_{ikm}] = a_i * g_k * z_m * N$

Population Density in the U.S.



Controlling Disclosure Risks

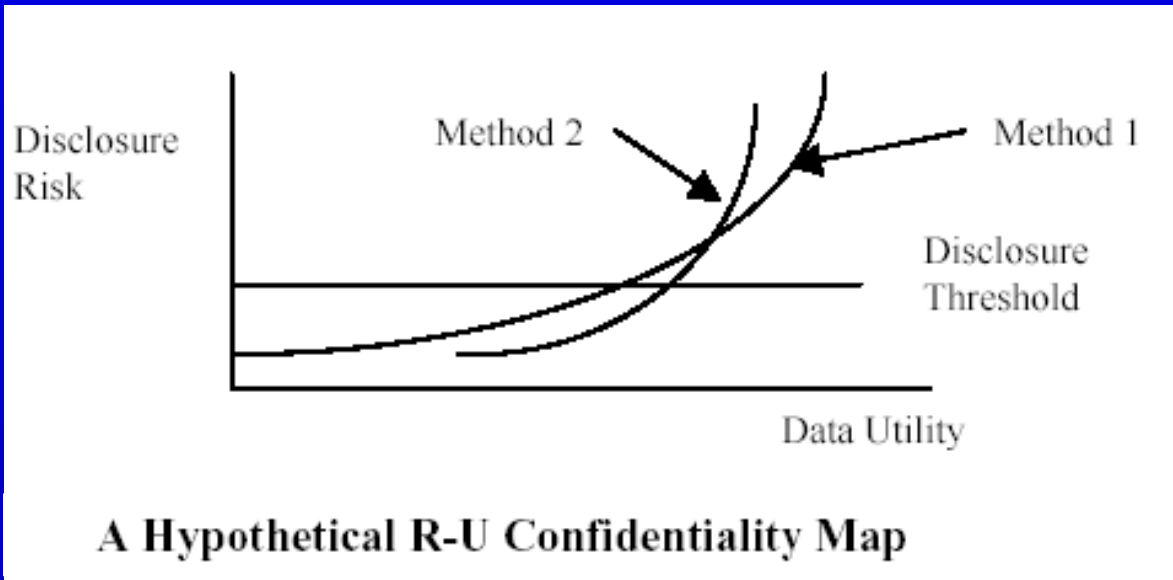
- Once sample uniques with a high probability of being a population unique have been identified, disclosure control measures can be applied to protect high-risk individuals from potential re-identification.
- Such disclosure control measures will inevitably result in some information loss (e.g., increased bias or loss of precision), but disclosure protection can be maximized while information loss is minimized.

Constrained Disclosure Control for Sample Uniques in Zip Codes

- Use Principal Components Analysis to summarize correlations between variables like education, income, housing problems, etc.
- Typically, these variables are highly correlated and a large proportion of the variability can be summarized in a small number of principal components.
- Perform Cluster Analysis with clusters formed from the predominant principal components.
- Perform geographically constrained data swapping to assure that swaps occur within limited distances and between demographically similar zip codes.

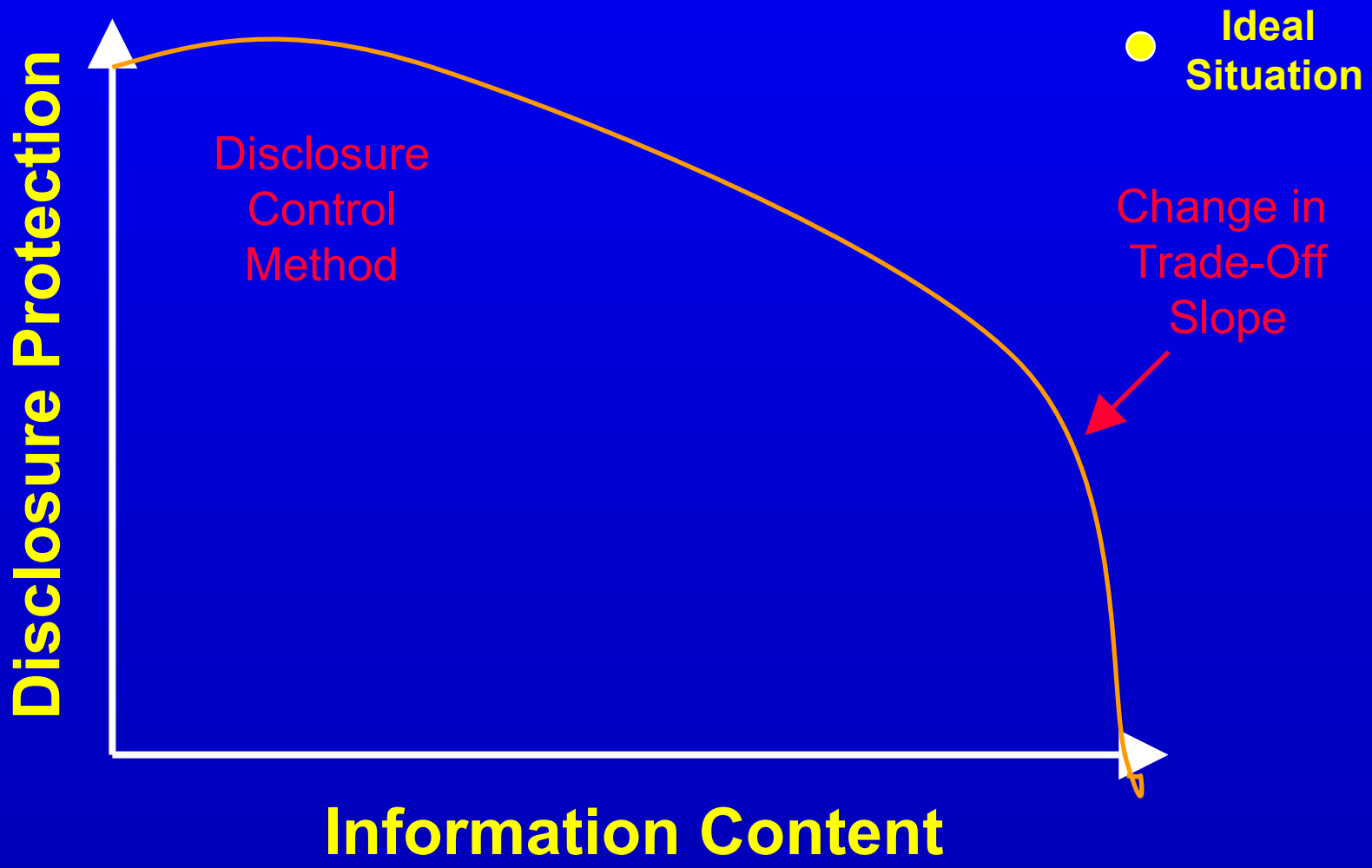
Statistical Disclosure Risk vs. Information Loss

- "R-U" Confidentiality Map proposed by George Duncan, Stephen Fienberg and colleagues. The R stands for (Disclosure) Risk, the U for (Data) Utility.

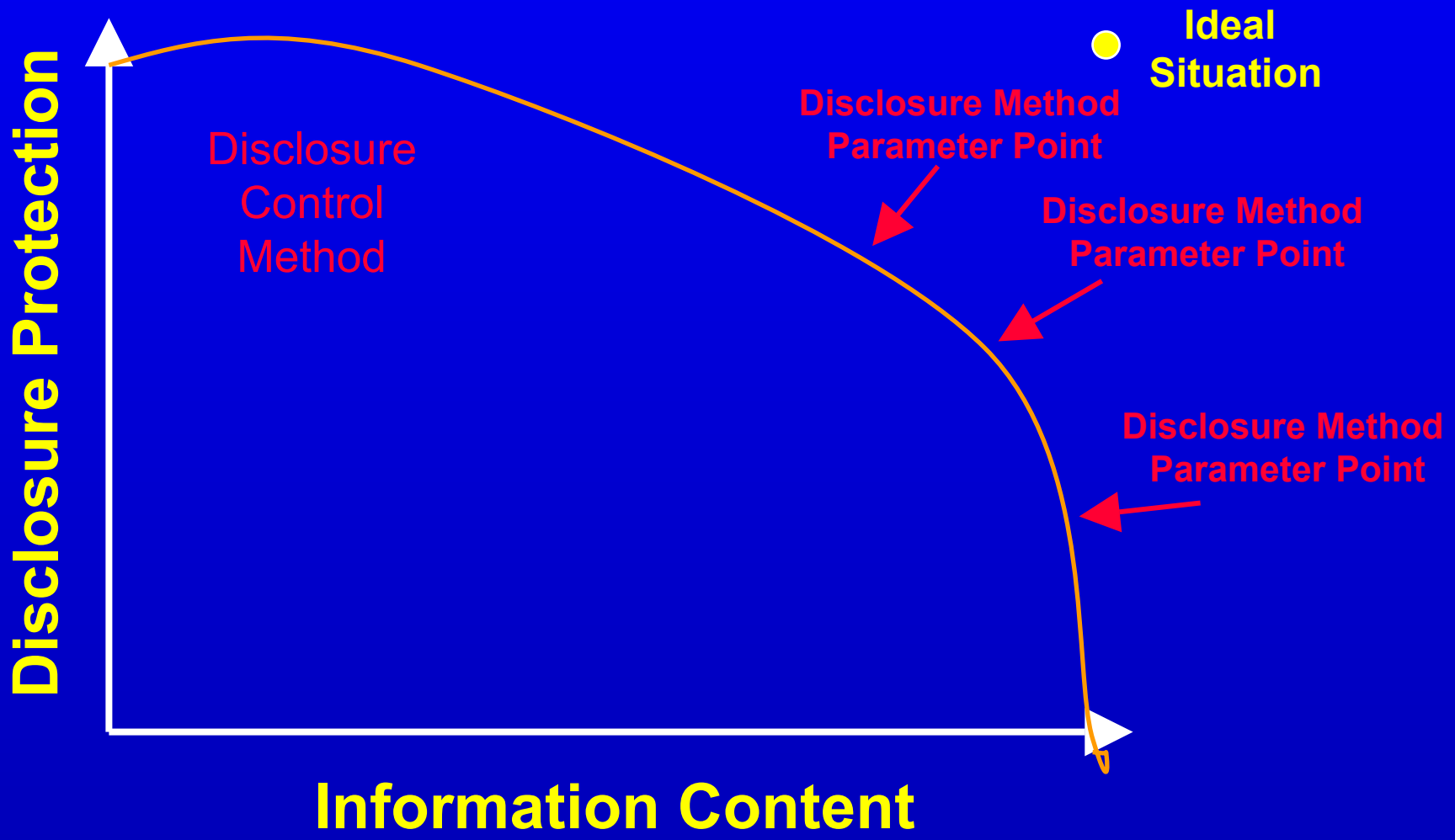


A Hypothetical R-U Confidentiality Map

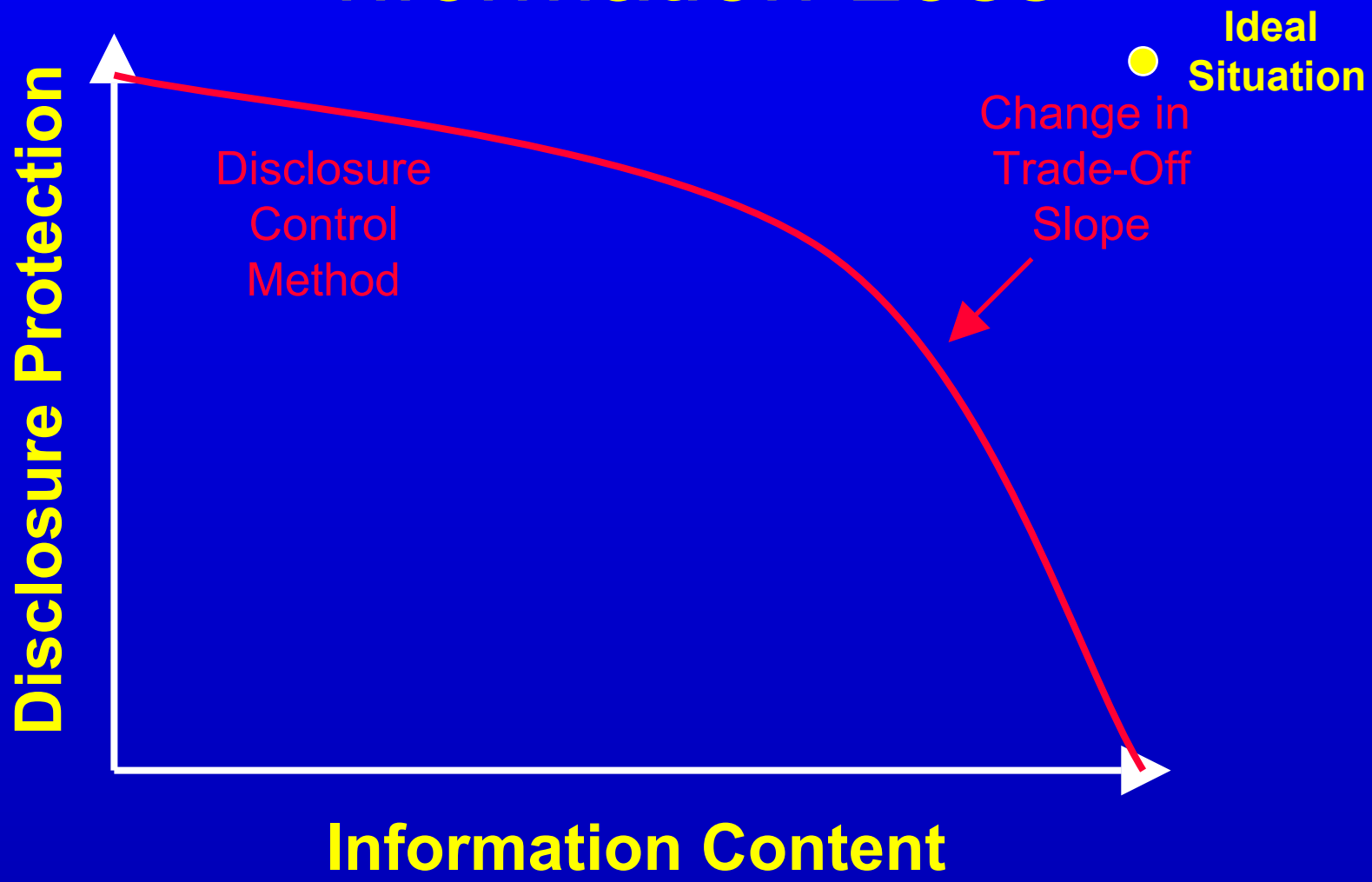
Statistical Disclosure Risk vs. Information Loss



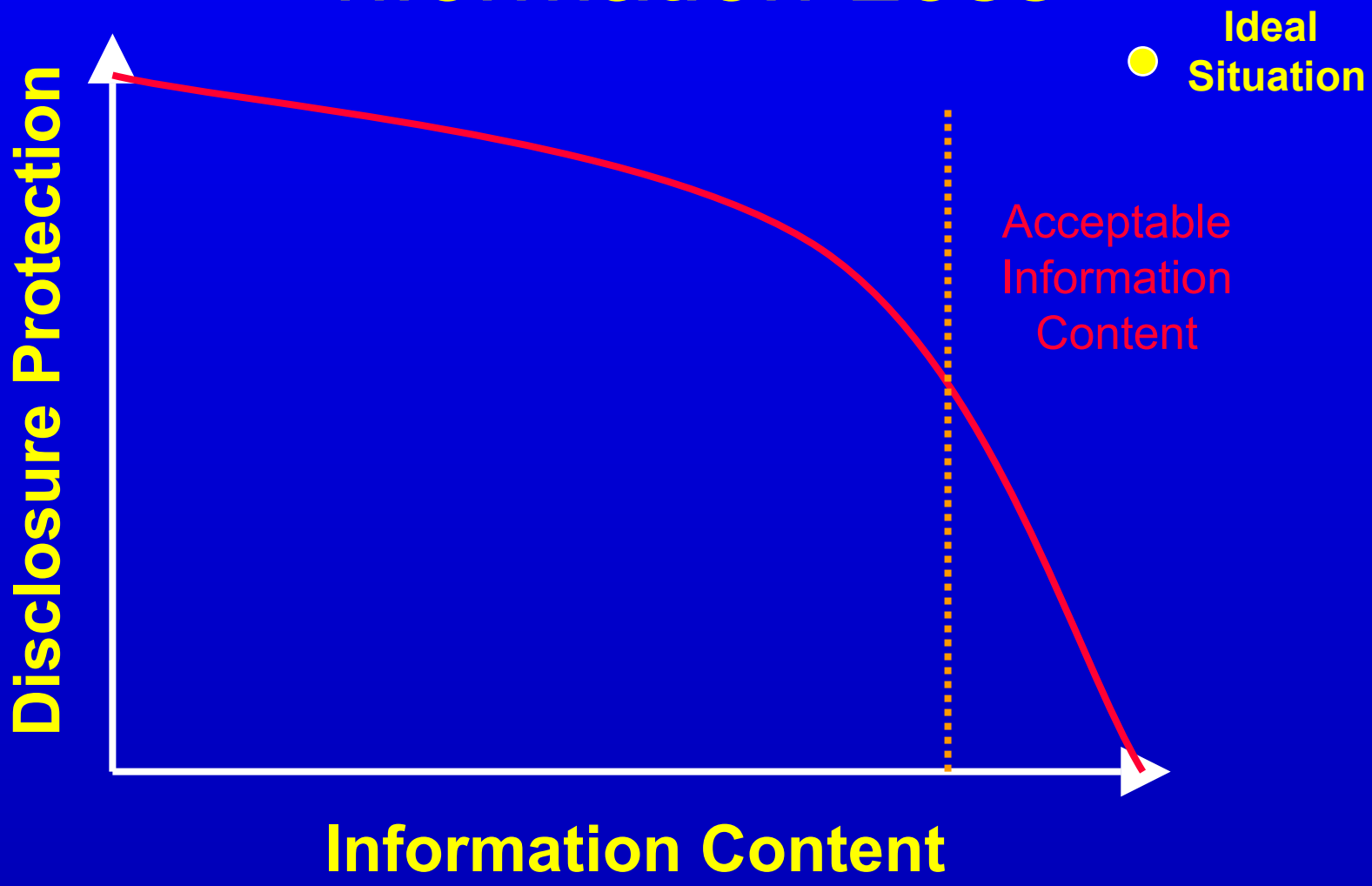
Statistical Disclosure Risk vs. Information Loss



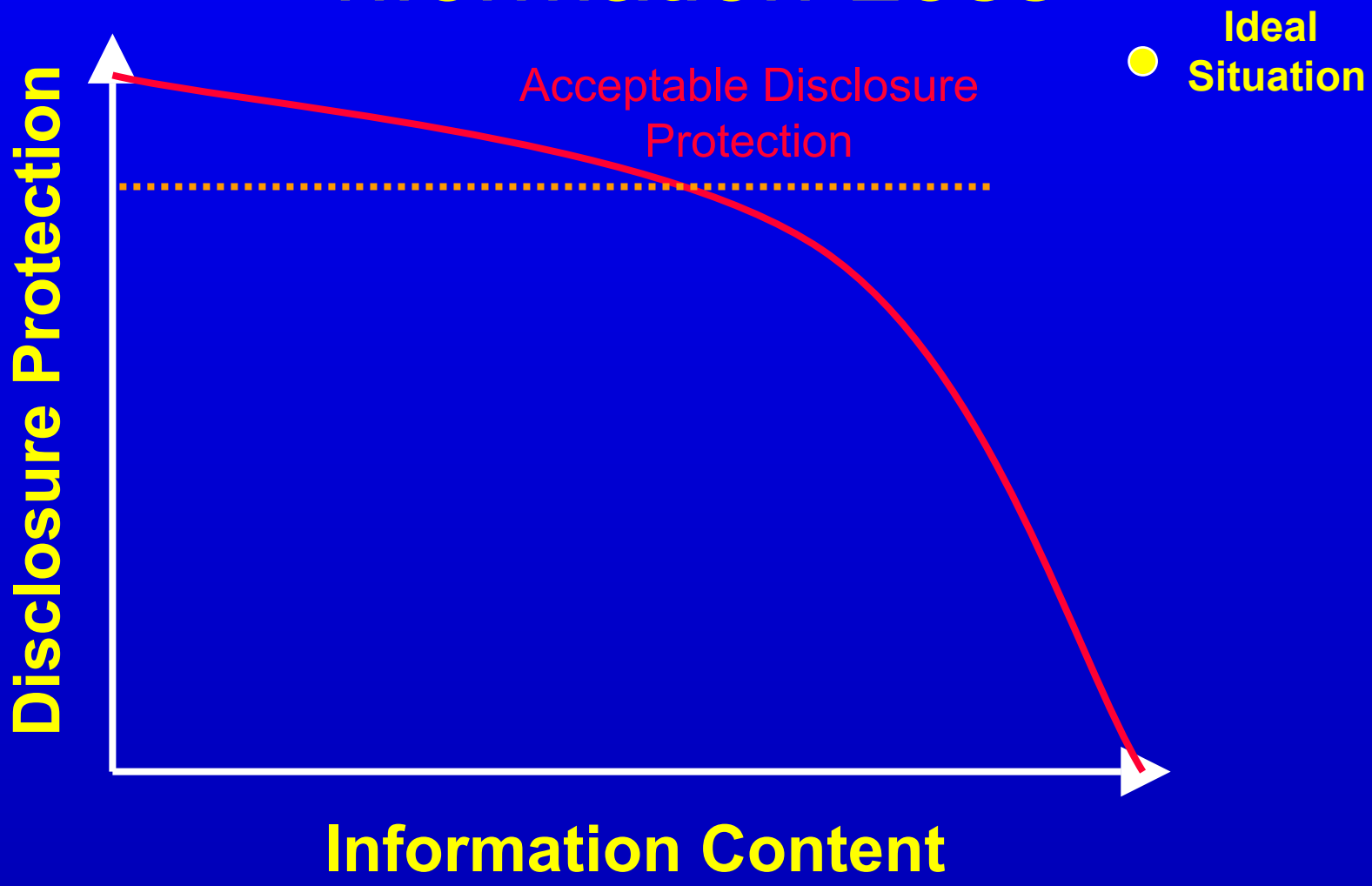
Statistical Disclosure Risk vs. Information Loss



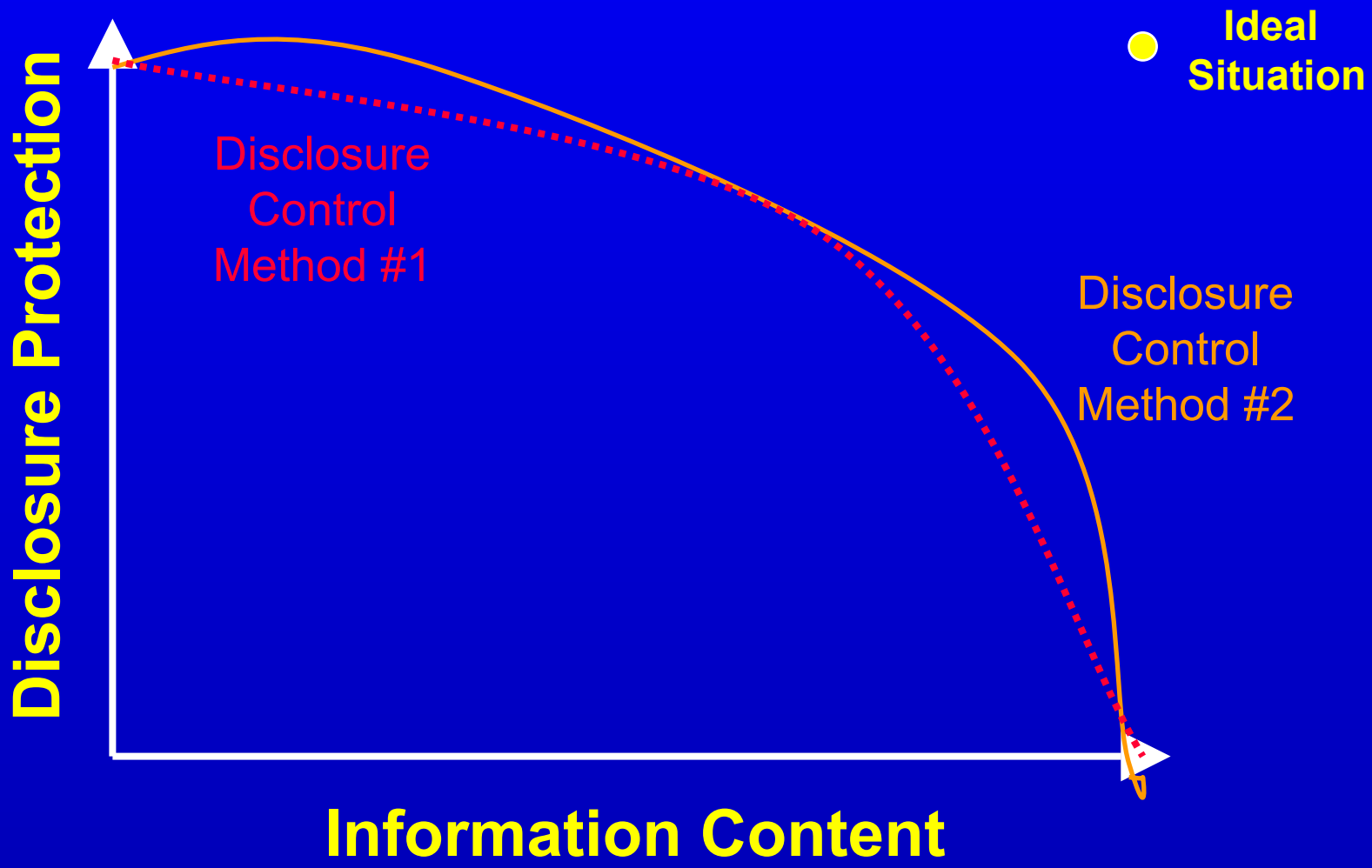
Statistical Disclosure Risk vs. Information Loss



Statistical Disclosure Risk vs. Information Loss



Statistical Disclosure Risk vs. Information Loss



Conclusions

- A comprehensive evaluation of statistical disclosure risks will include:
  - Conducting Statistical Disclosure Risk Analyses
  - Formulating a comprehensive set of Data Intrusion Scenarios
  - Estimating (conservatively) the “costs and availability” of the required data intrusion resources
  - Calculating the “real” risk of disclosure given the associated costs, etc.
  - Providing a well-reasoned and clear justification of your case that the risk of identification is “reasonably small”.
- Results of numerous analyses indicate that considerable disclosure control can be achieved with simple modifications of administrative data sets while preserving important geographic location detail.