


Carnegie Mellon
DATA PRIVACY LAB

Privacy Technology *The Frontier*

"provable guarantees of privacy protection while allowing information to be widely shared"



Latanya Sweeney

privacy.cs.cmu.edu

Carnegie Mellon
DATA PRIVACY LAB













Privacy Technology

1. Privacy is here to stay.
2. Computer scientist must help solve this problem.
3. Selective Revelation
4. Example: video surveillance
5. Example: bio-terrorism surveillance
6. Example: identity theft
7. Example: distributed surveillance
8. Example: privacy-preserving surveillance
9. Example: DNA privacy
10. Example: Identity theft protections
11. Example: k-Anonymity
12. Example: Webcam surveillance
13. Example: Text de-identification
14. Example: Policy specification and enforcement
15. Example: Scam Spam

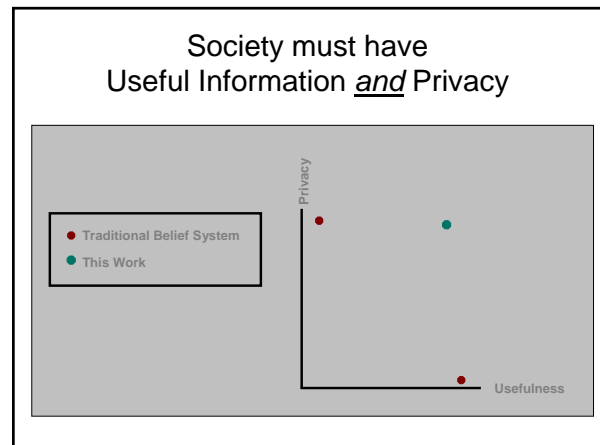
privacy.cs.cmu.edu

Carnegie Mellon
DATA PRIVACY LAB

Some People from the Lab

| | |
|--|--|
|  Edoardo Airoldi |  Latanya Sweeney |
|  Ralph Gross |  Michael Shamos |
|  Yiheng Li |  William Gronim |
|  Bradley Malin |  Rolf Holzer |
|  Brian Carini |  Kishore Madhava |
|  Samuel Edho-Eket |  Sherice Livingston |

...and more...



Surveillance Systems Today are not privacy-preserving

 CAPPS II

- Given: birth date, SSN, home phone, address
- System assigns you a color coded rating according the level of security risk that you pose

 Carnivore

- Gives FBI access to online/e-mail activities of suspected criminals

 TIA

- Use data-mining tools to sort through communication, medical, travel, and financial records
- **Congress canceled due to privacy concerns**

Privacy Trust Survey of The U.S. Gov

| Rank | Privacy Trust Rankings of Selected Federal Agencies | PTS |
|--|---|-----|
| 1 | United States Postal Service | 78% |
| 2 | Department of Veteran Affairs (VA) | 76% |
| 3 | Internal Revenue Service | 75% |
| 4 | Social Security Administration | 70% |
| 4 | Federal Trade Commission (FTC) | 70% |
| ... | ... | ... |
| 28 | National Security Administration (NSA) | 29% |
| 29 | Department of Homeland Security | 27% |
| 29 | Central Intelligence Agency (CIA) | 27% |
| 30 | Department of Justice | 22% |
| 31 | Office of the Attorney General | 22% |
| Overall Average Over 44 Federal Agencies | | 52% |

Survey Conducted by the Ponemon Institute and sponsored by the CIO Institute at CMU in January 2004.

Carnegie Mellon
DATA PRIVACY LAB

Privacy Technology

1. Privacy is here to stay.

privacy.cs.cmu.edu

Definition. Privacy

Privacy reflects the ability of a person, organization, government, or entity to control its own *space*, where the concept of space (or "privacy space") takes on different contexts

- Physical space, against invasion
- Bodily space, medical consent
- Computer space, spam
- Web browsing space, Internet privacy

Privacy Is Essential for Survival

Required for strategy, to compete over limited resources

Privacy reflects the autonomy and free will of the individual

Privacy provides a mechanism for "forgetting" or not knowing of some forms of indiscretions.

Example: play a game while revealing to other players his hand or strategy!

Example: in Christianity, individual choice between good and evil.

Example: "go west young man"

Privacy is Not Just For Individuals "confidentiality"

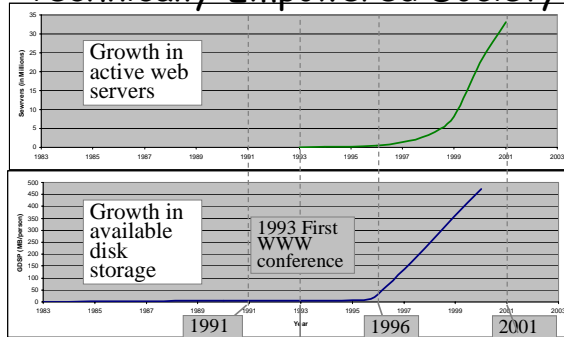
The ability to have private spaces or to limit or control what can be known is as crucial to businesses, governments, and other organizations as to individuals.

Example: FBI witness protection program

Example: US Military in Iraq

Example: VISA clearinghouse and Citibank

Technically-Empowered Society



L. Sweeney, Information Explosion, Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, L. Zayatz, P. Doyle, J. Theeuwes and J. Lane (eds), Urban Institute, Washington, DC, 2001.

Typical Birth Certificate Fields, post 1925

| Field name |
|--|
| Child's first name |
| Child's middle name (sometimes or initial) |
| Child's last name |
| Day, month and year of birth |
| City and/or County of birth (sometimes hospital) |
| Father's name |
| Mother's name (including maiden name) |
| Place of birth (address and town/city) |
| Mother's age and address |
| Mother's birthplace (town/city, state, county) |
| Mother's occupation |
| Mother, number of previous children |
| Father's age and address |
| Father's birthplace (town/city, state, county) |
| Father's occupation |

Typical Electronic Birth Certificate Fields
in 1999 -starting fields 1-15

| Field# | Size | Field name |
|--------|------|----------------------|
| 1 | 1 | File Status |
| 2 | 50 | Baby's First Name |
| 3 | 50 | Baby's Middle Name |
| 4 | 50 | Baby's Last Name |
| 5 | 1 | Baby's Suffix Code |
| 6 | 3 | Baby's Suffix Text |
| 7 | 8 | Baby's Date of Birth |
| 8 | 5 | Baby's Time of Birth |
| 9 | 1 | AM/PM Indicator |
| 10 | 1 | Baby's Sex |
| 11 | 3 | Blood Type |
| 12 | 1 | Born Here? |
| 13 | 40 | Place of Birth |
| 14 | 1 | Facility Type |

Typical Electronic Birth Certificate Fields
in 1999 -starting fields 16-30

| Field# | Size | Field name |
|--------|------|---------------------------------|
| 16 | 20 | County of Birth |
| 17 | 6 | Certifier's Code |
| 18 | 30 | Certifier's Name |
| 19 | 1 | Certifier's Title |
| 20 | 30 | Attendant's Name |
| 21 | 1 | Attendant's Title |
| 22 | 23 | Attendant's Address |
| 23 | 19 | Attendant's City |
| 24 | 2 | Attendant's State |
| 25 | 10 | Attendant's Zip Code |
| 26 | 50 | Mother's First Name |
| 27 | 50 | Mother's Middle Name |
| 28 | 50 | Mother's Last Name |
| 29 | 9 | Mother's Social Security Number |
| 30 | 8 | Mother's Date of Birth |

Typical Electronic Birth Certificate Fields
in 1999 -starting fields 31-45

| field# | Size | Field name |
|--------|------|-----------------------------------|
| 31 | 3 | Mother's State of Birth |
| 32 | 7 | Mother's Residence Address |
| 33 | 2 | Mother's Residence Direction |
| 34 | 20 | Residence Street Address |
| 35 | 10 | Residence Type |
| 36 | 2 | Residence Extension |
| 37 | 10 | Residence Apartment # |
| 38 | 20 | Mother's Town of Residence |
| 39 | 1 | Mother's Residence in City Limits |
| 40 | 14 | Mother's County of Residence |
| 41 | 3 | Mother's State of Residence |
| 42 | 10 | Mother's Residence Zip Code |
| 43 | 38 | Mother's Mailing Address |
| 44 | 19 | Mother's Mailing City |
| 45 | 2 | Mother's Mailing State |

Typical Electronic Birth Certificate Fields
in 1999 -starting fields 46-60

| Field# | Size | Field name |
|--------|------|---------------------------------|
| 46 | 10 | Mother's Mailing Zip Code |
| 47 | 1 | Mother Married? |
| 48 | 50 | Father's First Name |
| 49 | 50 | Father's Middle Name |
| 50 | 50 | Father's Last Name |
| 51 | 1 | Father's Suffix Code |
| 52 | 9 | Father's Suffix Text |
| 53 | 9 | Father's Social Security Number |
| 54 | 8 | Father's Date of Birth |
| 55 | 3 | Father's State of Birth |
| 56 | 14 | Mother's Origin |
| 57 | 14 | Mother's Race |
| 58 | 2 | Mother's Elementary Education |
| 59 | 2 | Mother's College Education |
| 60 | 11 | Mother's Occupation |

Typical Electronic Birth Certificate Fields
in 1999 -starting fields 61-75

| Field# | Size | Field name |
|--------|------|-------------------------------|
| 61 | 11 | Mother's Industry |
| 62 | 14 | Father's Origin |
| 63 | 14 | Father's Race |
| 64 | 2 | Father's Elementary Education |
| 65 | 2 | Father's College Education |
| 66 | 11 | Father's Occupation |
| 67 | 11 | Father's Industry |
| 68 | 1 | Plurality |
| 69 | 1 | Birth Order |
| 70 | 2 | Live Births Still Living |
| 71 | 2 | Live Births Now Dead |
| 72 | 4 | Month/Year Last Live Birth |
| 73 | 2 | Number of Terminations |
| 74 | 4 | Month/Year Last Termination |
| 75 | 1 | Baby's Weight Unit |

Typical Electronic Birth Certificate Fields
in 1999 -starting fields 76-90

| Field# | Size | Field name |
|--------|------|----------------------------------|
| 76 | 5 | Baby's Weight |
| 77 | 6 | Date of Last Normal Menses |
| 78 | 1 | Month Prenatal Care Began |
| 79 | 2 | Total Number of Visits |
| 80 | 2 | Apgar Score – 1 Minute |
| 81 | 2 | Apgar Score – 5 Minute |
| 82 | 2 | Estimate of Gestation |
| 83 | 6 | Date of Blood Test |
| 84 | 22 | Laboratory |
| 85 | 1 | Mother Transferred In |
| 86 | 30 | Facility Mother Transferred From |
| 87 | 1 | Baby Transferred Out |
| 88 | 30 | Facility Baby Transferred To |
| 89 | 1 | Tobacco Use During Pregnancy |
| 90 | 3 | Number of Cigarettes/Day |

Typical Electronic Birth Certificate Fields
in 1999 -starting fields 91-105

| Field# | Size | Field name |
|--------|------|------------------------------|
| 91 | 1 | Alcohol Use During Pregnancy |
| 92 | 3 | Number of Drinks/Week |
| 93 | 3 | Mother's Weight Gain |
| 94 | 1 | Release Info For SSN |
| 95 | 6 | Operator Code |
| 96 | 12 | Hospital ID |
| 97 | 1 | Sent to Romans |
| 98 | 1 | Sent to APORS |
| 99 | 16 | Other Certifier Specify |
| 100 | 12 | Temporary Audit Number |
| 101 | 16 | Other Facility Specify |
| 102 | 16 | Other Attendant Specify |
| 103 | 1 | Mother's Race |
| 104 | 1 | Father's Race |
| 105 | 2 | Mother's Origin |

Typical Electronic Birth Certificate Fields
in 1999 -starting fields 106-120

| Field# | Size | Field name |
|--------|------|------------------------------|
| 106 | 2 | Father's Origin |
| 107 | 1 | Attendant Same YN |
| 108 | 1 | Mailing Address Same YN |
| 109 | 1 | Capture Father's Info YN |
| 110 | 2 | Mother's Age |
| 111 | 2 | Father's Age |
| 112 | 12 | Baby's Hospital Med. Rec. |
| 113 | 1 | High Risk Pregnancy YN |
| 114 | 1 | Care Giver (For Chicago) |
| 115 | 1 | Record Selected For Download |
| 116 | 1 | Downloaded |
| 117 | 1 | Printed |
| 118 | 12 | Form Number |
| | | MEDICAL RISK FACTORS |
| 119 | 1 | Anemia |
| 120 | 1 | Cardiac Disease |

Typical Electronic Birth Certificate Fields
in 1999 -starting fields 121-135

| Field# | Size | Field name |
|--------|------|--------------------------------|
| 121 | 1 | Acute/Chronic Lung Disease |
| 122 | 1 | Diabetes |
| 123 | 1 | Genital Herpes |
| 124 | 1 | Hydramnios/Oligohydramnios |
| 125 | 1 | Hemoglobinopathy |
| 126 | 1 | Hypertension, Chronic |
| 127 | 1 | Hypertension, Preg. Assoc. |
| 128 | 1 | Eclampsia |
| 129 | 1 | Incompetent Cervix |
| 130 | 1 | Previous Infant 4000+ Grams |
| 131 | 1 | Previous Preterm or SGA Infant |
| 132 | 1 | Renal Disease |
| 133 | 1 | Rh Sensitization |
| 134 | 1 | Uterine Bleeding |
| 135 | 1 | No Medical Risk Factors |

Typical Electronic Birth Certificate Fields
in 1999 -starting fields 136-150

| Field# | Size | Field name |
|--------|------|---------------------------------------|
| 136 | 40 | Other Medical Risk Factors |
| | | OBSTETRIC PROCEDURES |
| 137 | 1 | Amniocentesis |
| 138 | 1 | Electronic Fetal Monitoring |
| 139 | 1 | Induction of Labor |
| 140 | 1 | Stimulation of Labor |
| 141 | 1 | Tocolysis |
| 142 | 1 | Ultrasound |
| 143 | 1 | No Obstetric Procedures |
| 144 | 40 | Other Obstetric Procedures |
| | | COMPLICATIONS OF LABOR & I |
| 145 | 1 | Febrile (>100 or 38C) |
| 146 | 1 | Meconium Moderate, Heavy |
| 147 | 1 | Premature Rupture (>12 Hrs) |
| 148 | 1 | Abruptio Placenta |
| 149 | 1 | Placenta Previa |
| 150 | 1 | Other Excessive Bleeding |

Typical Electronic Birth Certificate Fields
in 1999 -starting fields 151-165

| Field# | Size | Field name |
|--------|------|----------------------------------|
| 151 | 1 | Seizures During Labor |
| 152 | 1 | Precipitous Labor (<3 Hrs) |
| 153 | 1 | Prolonged Labor (>20 Hrs) |
| 154 | 1 | Dysfunctional Labor |
| 155 | 1 | Breech/Malpresentation |
| 156 | 1 | Cephalopelvic Disproportion |
| 157 | 1 | Cord Prolapse |
| 158 | 1 | Anesthetic Complications |
| 159 | 1 | Fetal Distress |
| 160 | 1 | No Complications of L&D |
| 161 | 40 | Other Complications of L&D |
| | | METHOD OF DELIVERY |
| 162 | 1 | Vaginal |
| 163 | 1 | Vaginal After Previous C-Section |
| 164 | 1 | Primary C-Section |
| 165 | 1 | Repeat C-Section |

Typical Electronic Birth Certificate Fields
in 1999 -starting fields 166-180

| Field# | Size | Field name |
|--------|------|--------------------------------------|
| 166 | 1 | Forceps |
| 167 | 1 | Vacuum |
| | | ABNORMAL CONDITIONS OF NEWBO |
| 168 | 1 | Anemia |
| 169 | 1 | Birth Injury |
| 170 | 1 | Fetal Alcohol Syndrome |
| 171 | 1 | Hyaline Membrane Disease/RDS |
| 172 | 1 | Meconium Aspiration Syndrome |
| 173 | 1 | Assisted Ventilation <30 |
| 174 | 1 | Assisted Ventilation >30 |
| 175 | 1 | Seizures |
| 176 | 1 | No Abnormal Conditions of Newborn |
| 177 | 40 | Other Abnormal Condition of Newborn |
| | | CONGENITAL ANOMALIES OF CHILD |
| 178 | 1 | Anencephalus |
| 179 | 1 | Spina Bifida/Meningocele |
| 180 | 1 | Hydrocephalus |

Typical Electronic Birth Certificate Fields in 1999 -starting fields 181-195

| Field# | Size | Field name |
|--------|------|------------------------------------|
| 181 | 1 | Microcephalus |
| 182 | 40 | Other CNS Anomalies |
| 183 | 1 | Heart Malformations |
| 184 | 40 | Other Circ./Resp. Anomalies |
| 185 | 1 | Rectal Atresia/Stenosis |
| 186 | 1 | Tracheo-Esophageal Fistula/Esophag |
| 187 | 1 | Omphalocele/Gastroschisis |
| 188 | 40 | Other Gastrointestinal Ano. |
| 189 | 1 | Malformed Genitalia |
| 190 | 1 | Renal Agenesis |
| 191 | 40 | Other Urogenital Anomalies |
| 192 | 1 | Cleft Lip/Palate |
| 193 | 1 | Polydactyl/Syndactyl/Adactyl |
| 194 | 1 | Club Foot |
| 195 | 1 | Diaphragmatic Hernia |

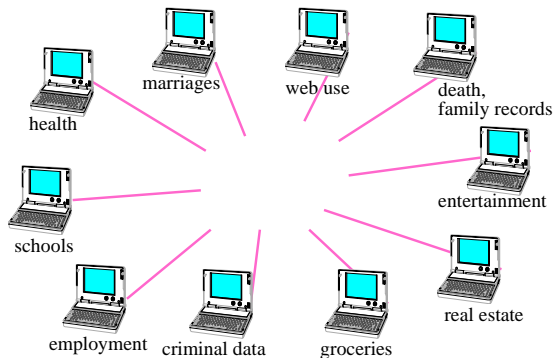
Typical Electronic Birth Certificate Fields in 1999 -starting fields 196-210

| Field# | Size | Field name |
|--------|------|--------------------------------------|
| 196 | 40 | Other Musculoskeletal/Integumental A |
| 197 | 1 | Down's Syndrome |
| 198 | 40 | Other Chromosomal Anomalies |
| 199 | 1 | No Congenital Anomalies |
| 200 | 40 | Other Congenital Anomalies |
| | | CODE STRIP |
| 201 | 1 | Record Complete YN |
| 202 | 1 | Record Type |
| 203 | 4 | Facility ID |
| 204 | 4 | City of Birth |
| 205 | 3 | County of Birth |
| 206 | 2 | Mother's State of Birth |
| 207 | 2 | Mother's State of Residence |
| 208 | 4 | Mother's Town of Residence |
| 209 | 3 | Mother's County of Residence |
| 210 | 2 | Father's State of Birth |

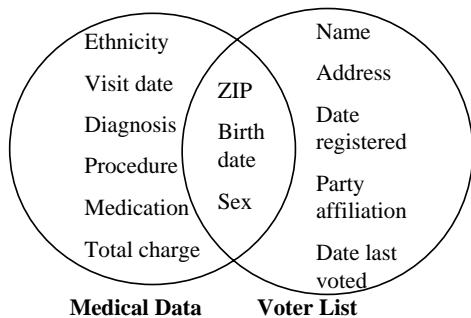
Typical Electronic Birth Certificate Fields in 1999 -starting fields 211-226.

| Field# | Size | Field name |
|--------|------|-----------------------------|
| 211 | 14 | Certifier's License Number |
| 212 | 6 | Laboratory ID Number |
| 213 | 4 | Mother Xfer Code |
| 214 | 3 | Mother Xfer County Code |
| 215 | 4 | Baby Xfer Code |
| 216 | 3 | Baby Xfer County Code |
| 217 | 4 | Year of Birth |
| 218 | 7 | Certificate # |
| 219 | 1 | Unique Code |
| 220 | 8 | File Date |
| 221 | 2 | Community Area |
| 222 | 4 | Census Tract |
| 223 | 2 | Century of Last Live Birth |
| 224 | 2 | Century of Last Termination |
| 225 | 2 | Century of Last Menses |
| 226 | 2 | Century of Blood Test |

Numerous Efforts Underway to Fuse Available Data Together on Individuals

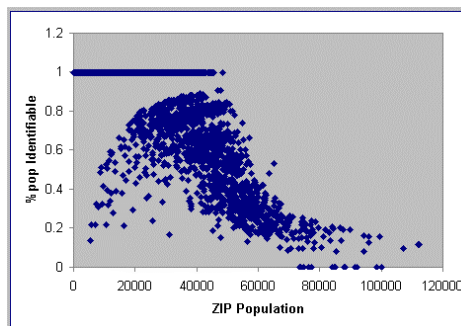


Linking to re-identify data

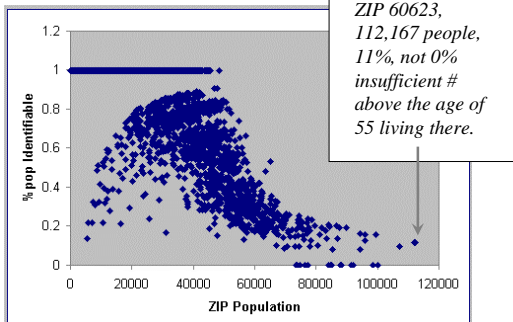


L. Sweeney. Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine and Ethics*. 1997; 25:98-110.

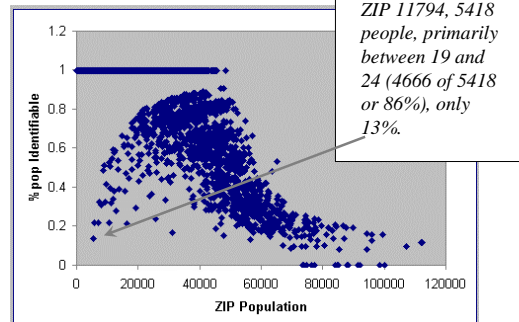
{date of birth, gender, 5-digit ZIP}
uniquely identifies 87.1% of USA pop.



{date of birth, gender, 5-digit ZIP}
uniquely identifies 87.1% of USA pop.



{date of birth, gender, 5-digit ZIP}
uniquely identifies 87.1% of USA pop.



Carnegie Mellon
DATA PRIVACY LAB

Privacy Technology

1. Privacy is here to stay.
2. Computer scientist must help solve this problem.
3. Selective Revelation
4. Example: video surveillance
5. Example: bio-terrorism surveillance
6. Example: identity theft
7. Example: distributed surveillance
8. Example: privacy-preserving surveillance
9. Example: DNA privacy
10. Example: Identity theft protections
11. Example: k-Anonymity
12. Example: Webcam surveillance
13. Example: Text de-identification
14. Example: Policy specification and enforcement
15. Example: Scam Spam

privacy.cs.cmu.edu

Computer Scientists Must Help Save the World

Policy is limited by that which can be expressed in words, tends to be crude descriptions in absence of technical refinement.

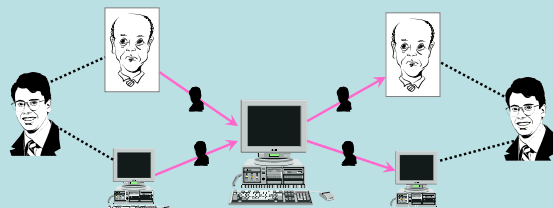
Traditional research based on statistical and economic approaches to policy problems tend to be retrospective and descriptive and assume technology is relatively static.

IS/IT can provide glue technology, but heavily relies on existing technology.

Laws can change and lawyers often lack understanding of ways technology will continue to change. Rivest: technology changes in months, laws change in years

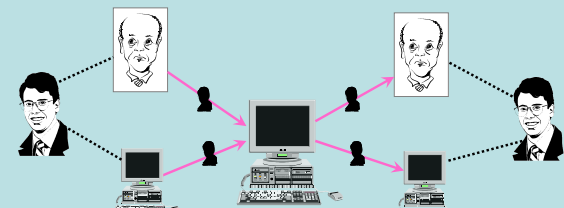
Computer scientists construct tomorrow's machines, and can do so with privacy as part of their problem definition, so that new technology can be deployed and easily adopted.

Privacy > Computer Security



Authentication, Authorization, Encryption can help prevent data stolen from intrusion and break-ins → privacy ala computer security

Privacy > Computer Security



What about data given away? What about data over which the individual has no control? What about all those front page news articles?

Authentication, Authorization, Encryption are not sufficient! → privacy > computer security

Relationships Between Related Areas

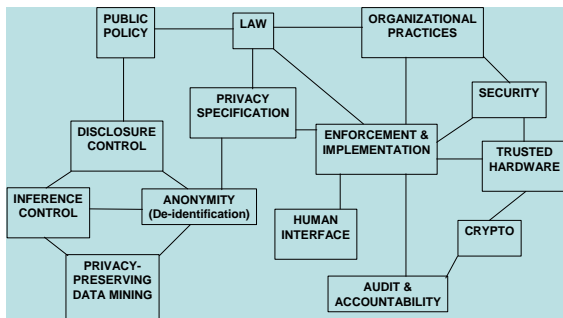


Diagram courtesy of Michael Shamos.

Data Privacy has 2 Basic Areas in 2 Settings

| Data Anonymity (anonymity) | Policy Specification & Enforcement (rights mgt) | Database Security (security) | Distributed & Ubiquitous Environments (distributed) |
|---|---|--|--|
| Methods for detecting and controlling inferences in data. | Methods for language design with automated detection and enforcement. | Methods for controlling access to and protecting database content. | Methods related to having a network of data sources. |

DATA PRIVACY LAB

Privacy Technology

1. Privacy is here to stay.
2. Computer scientist must help solve this problem.
3. Selective Revelation
4. Example: video surveillance
5. Example: bio-terrorism surveillance
6. Example: identity theft
7. Example: distributed surveillance
8. Example: privacy-preserving surveillance
9. Example: DNA privacy
10. Example: Identity theft protections
11. Example: k-Anonymity
12. Example: Webcam surveillance
13. Example: Text de-identification
14. Example: Policy specification and enforcement
15. Example: Scam Spam

DATA PRIVACY LAB

Selective Revelation

| | |
|----------------------------|---------------------|
| Gross overview | |
| Sufficiently anonymous | Normal operation |
| Sufficiently de-identified | Unusual activity |
| Identifiable | Suspicious activity |
| Readily identifiable | Outbreak suspected |
| Explicitly identified | Outbreak detected |

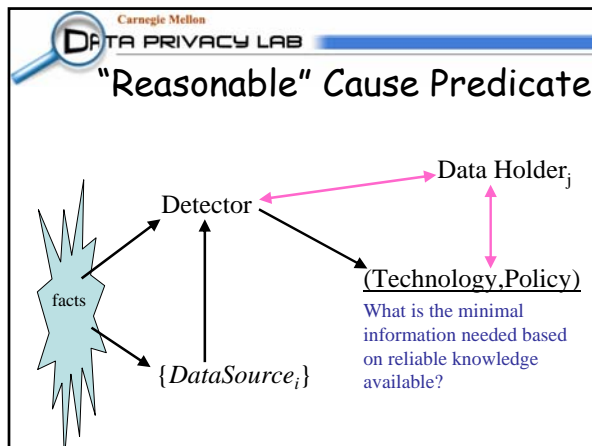
Identifiability 0..1 Detection Status 0..1

Dynamically Augment Access As Surveillance Warrants

Lower the privacy threshold when potential attack detected

DATA PRIVACY LAB

Probable Cause Predicate



Carnegie Mellon
DATA PRIVACY LAB

Privacy Technology

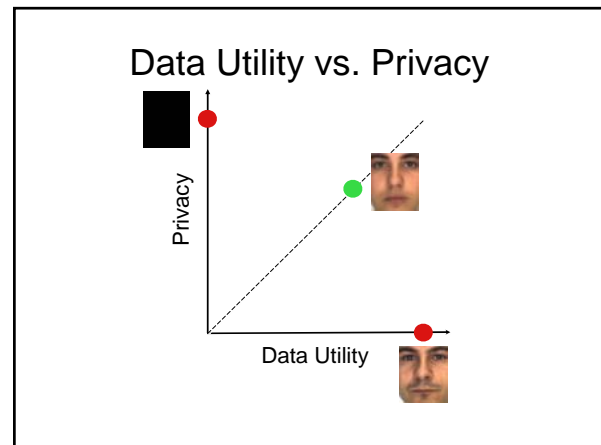
1. Privacy is here to stay.
2. Computer scientist must help solve this problem.
3. Selective Revelation
4. Example: video surveillance
5. Example: bio-terrorism surveillance
6. Example: identity theft
7. Example: distributed surveillance
8. Example: privacy-preserving surveillance
9. Example: DNA privacy
10. Example: Identity theft protections
11. Example: k-Anonymity
12. Example: Webcam surveillance
13. Example: Text de-identification
14. Example: Policy specification and enforcement
15. Example: Scam Spam

privacy.cs.cmu.edu

Probable Cause "catch 22:" something useful MAY be on a video recording related to a crime, but without viewing the video cannot get a search warrant to access the video.

Can we share video with law-enforcement such that no matter how good face recognition software might become, people cannot be re-identified without due process?

Want to retain privacy protections afforded by US Constitution and the need for search warrant yet enable more sharing of video.



Controlling Face Recognition

De-identify faces such that no face recognition software can be successful, even if face recognition software is perfect.

Newton, Sweeney, and Malin. Preserving Privacy by De-identifying Facial Images. *IEEE Transactions on Knowledge and Data Engineering*, Accepted 2004. See also CMU Tech Report

Ad Hoc Schemes that Don't Work I

Single Bar Mask

T-Mask


Mouth Only
✓ Grayscale
✓ Black & White

Ordinal

Threshold

In each of these, face recognition software can be trained on the method and the result yields successful re-identifications!

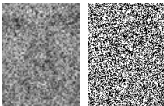
Ad Hoc Schemes that Don't Work II



Pixelation

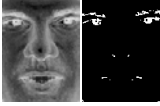
Random

- ✓ Grayscale
- ✓ Black & White




Negative

- ✓ Grayscale
- ✓ Black & White



In each of these, face recognition software can be trained on the method and the result yields successful re-identifications!



Mr. Potato Head

Pixelation – recognition improved!



Face De-Identification: Wrong Way

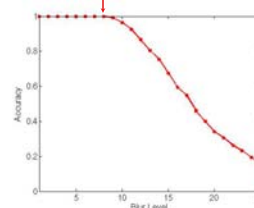
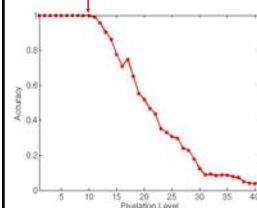
Blurring



Pixelate



Face De-Identification: Wrong Way





PCA Recognition Accuracies

Carnegie Mellon
DATA PRIVACY LAB


**k-Same
anonymizing faces**

Elaine Newton
Latanya Sweeney
Bradley Malin





Thwarts face recognition while many facial details remain!

-Pixel

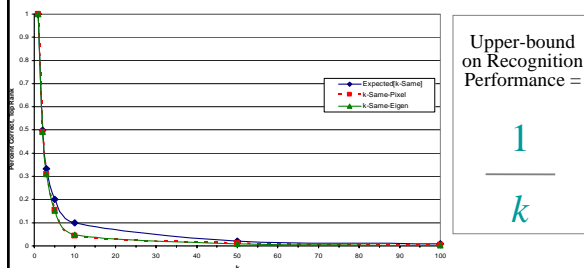


-Eigen



$k =$ 2 3 5 10 50 100

Performance of k -Same Algorithms v. Values of k



k -Same Works even as Face Recognition Software Improves

Theorem. There cannot exist any face recognition software for which a subject's k -Samed image can be correctly recognized better than $1/k$ probability.

Note. Theorem above loosely specified.
 H is the de-identified face set, $|H| \geq k$ and $k > 1$.

Face Tracking



- Model fit at 230 frames per second
- Accurately captures non-rigid facial motions

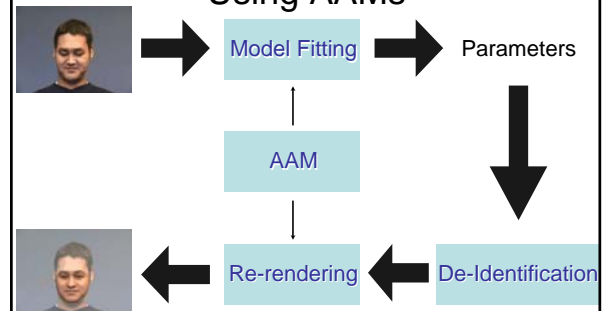
[L.Matthews and S.Baker, Active Appearance Models Revisited, IJCV, 60(2), 2004]

Fitting Models With Occlusion



[R. Gross, L.Matthews and S.Baker, Constructing and Fitting Active Appearance Models with Occlusion, IEEE Workshop on Face Processing in Video, 2004]

Face De-Identification Using AAMs

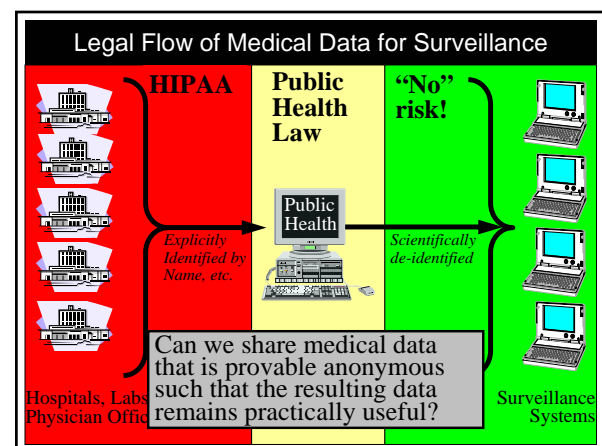


DATA PRIVACY LAB

Privacy Technology

1. Privacy is here to stay.
2. Computer scientist must help solve this problem.
3. Selective Revelation
4. Example: video surveillance
5. Example: bio-terrorism surveillance
6. Example: identity theft
7. Example: distributed surveillance
8. Example: privacy-preserving surveillance
9. Example: DNA privacy
10. Example: Identity theft protections
11. Example: k-Anonymity
12. Example: Webcam surveillance
13. Example: Text de-identification
14. Example: Policy specification and enforcement
15. Example: Scam Spam

privacy.cs.cmu.edu



Fields of the Bio-Surveillance DataStream

| Field# | Description | Name |
|--------|---------------------------------------|-------------|
| 1 | * Date of visit (month, day and year) | Date |
| 2 | Transaction# | Transaction |
| 3 | Unique patient identifier | PatientID |
| 4 | * Patient 5-digit ZIP code | ZIP |
| 5 | * Month, day and Year of Birth | DOB |
| 6 | * Gender | Sex |
| 7 | Unique Provider ID | ProviderID |
| 8 | Provider 5-digit ZIP code | ProviderZIP |
| 9 | ICD9 diagnosis code 1 | Dx1 |
| 10 | * ICD9 diagnosis code 2 | Dx2 |
| 11 | * ICD9 diagnosis code 3 | Dx3 |
| 12 | * ICD9 diagnosis code 4 | Dx4 |
| 13 | * ICD9 diagnosis code 5 | Dx5 |
| 14 | * ICD9 diagnosis code 6 | Dx6 |

Fields ESSENCE II considers important to their ability to conduct bio-terrorism surveillance. Asterisked fields are considered critical.

Early Aberration Reporting System (EARS)

- Aberration detection models identify deviations in current data when compared to a historical mean.
- Cumulative Sums (CUSUM) is a statistical quality control algorithm used to detect shifts away from the mean
- Historical: compares current count to 5-year mean
- Non-historical: compares current count to 7-day mean

Hutwagner, L. Seeman M, Thomson W, Treadwell T. (2002). CDC: Early aberration reporting system (EARS), presentation at National Syndromic Surveillance Conference, New York City, Fall 2002

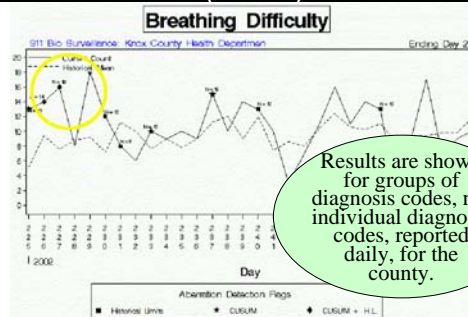
Burkom, H. EARS Java source code, Johns Hopkins University, 2003.

Example. Use of Daily Case Counts (EARS)



Lawson, B., Fitzhugh, E., Hall, S., Garcia, M., Hutwagner, L., Seeman, G., Implementing the CDC Early Aberration Reporting System (EARS): A Front-Line Perspective from the Knox County (TN) Health Department, 2002.

Example. Use of Daily Case Counts (EARS)



Results are shown for groups of diagnosis codes, not individual diagnosis codes, reported daily, for the county.

Lawson, B., Fitzhugh, E., Hall, S., Garcia, M., Hutwagner, L., Seeman, G., Implementing the CDC Early Aberration Reporting System (EARS): A Front-Line Perspective from the Knox County (TN) Health Department, 2002.

De-identification Under HIPAA Safe Harbor, Remove following:

- (A) Names;
- (B) All geographic subdivisions, except first 3 digits of ZIP code (only 2 digits if ZIP population < 20K)
- (C) All elements of dates (except year) for dates
- (D) Telephone numbers; (E) Fax numbers;
- (F) Electronic mail addresses; (G) Social security numbers;
- (H) Medical record numbers; and other numbers
- (N) Web Universal Resource Locators (URLs);
- (O) Internet Protocol (IP) address numbers;
- (P) Biometric identifiers, etc

U.S. Health and Human Services; Standards for Privacy of Individually Identifiable Health Information; Final Rule, 45 CFR Parts 160 and 164. *Federal Register*, vol 67, no 157, August 14, 2002.

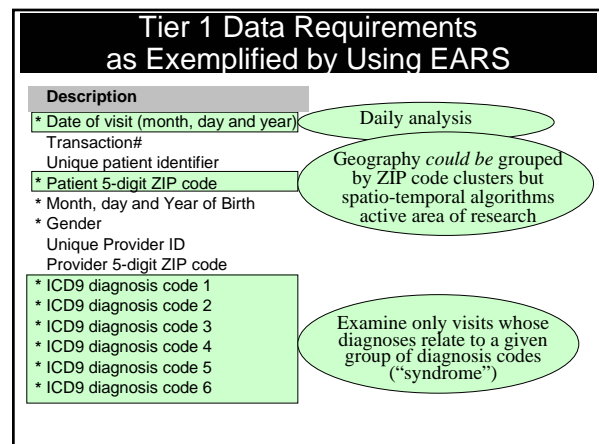
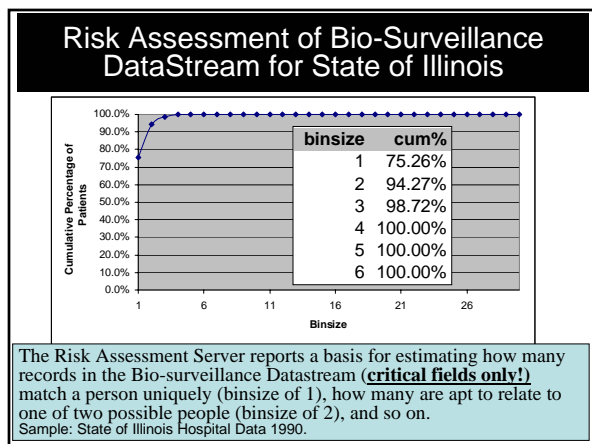
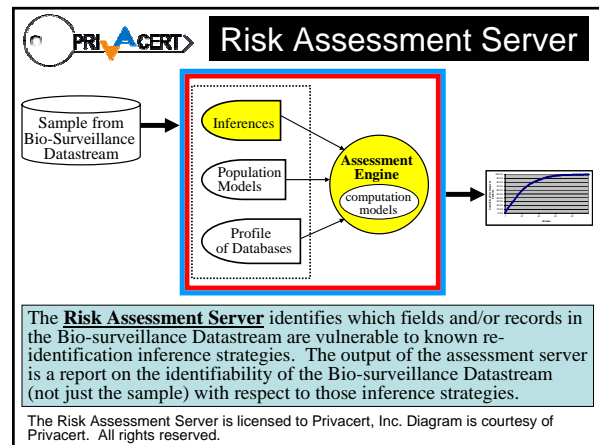
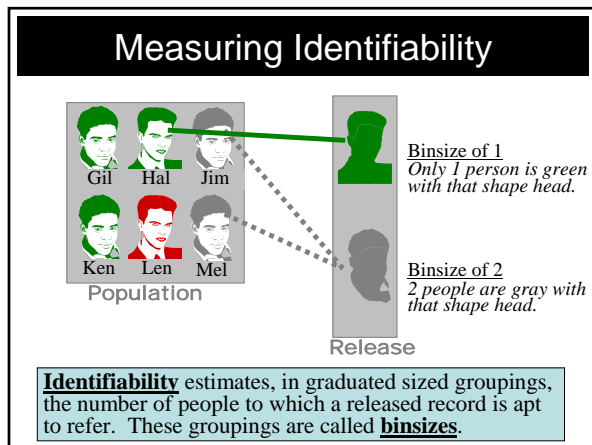
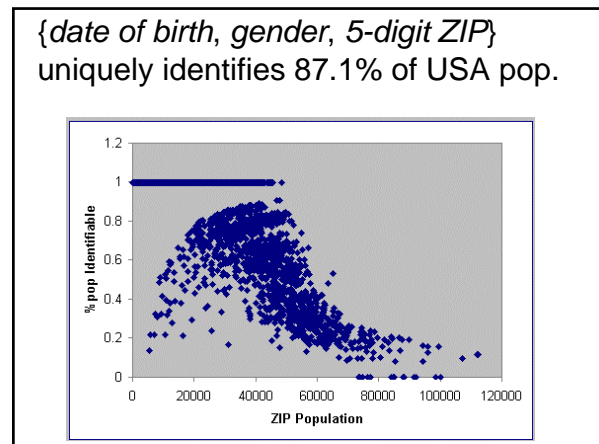
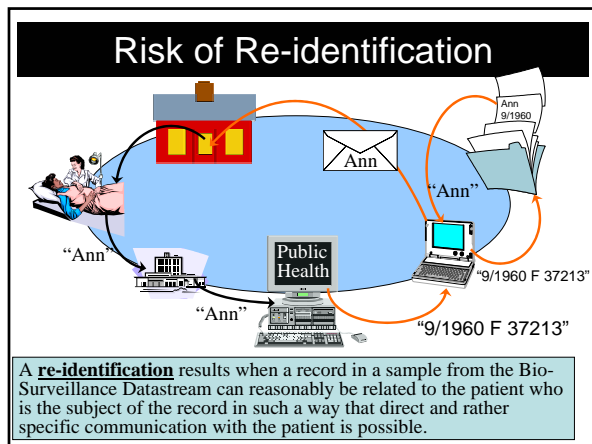
De-Identification Under HIPAA Scientific Standard

Given the nature and content of the health information

and based on generally accepted computational, statistical and scientific principles and methods, a person certifies 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.”

U.S. Health and Human Services; Standards for Privacy of Individually Identifiable Health Information; Final Rule, 45 CFR Parts 160 and 164. *Federal Register*, vol 67, no 157, August 14, 2002.

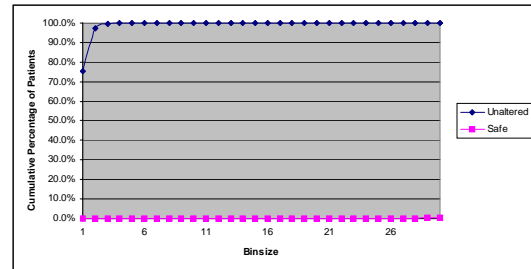


Fields of the Bio-Surveillance DataStream, HIPAA Safe Harbor Version

| Field# | Description | Name |
|--------|--------------------------------|-------------|
| 1 | * Date of visit (Year) | Date |
| 2 | Transaction# | Transaction |
| 3 | Unique patient identifier | PatientID |
| 4 | * Patient 3-digit ZIP Code | ZIP |
| 5 | * Month, day and Year of Birth | DOB |
| 6 | * Gender | Sex |
| 7 | Unique Provider ID | ProviderID |
| 8 | Provider 3-digit ZIP Code | ProviderZIP |
| 9 | * ICD9 diagnosis code 1 | Dx1 |
| 10 | * ICD9 diagnosis code 2 | Dx2 |
| 11 | * ICD9 diagnosis code 3 | |
| 12 | * ICD9 diagnosis code 4 | |
| 13 | * ICD9 diagnosis code 5 | |
| 14 | * ICD9 diagnosis code 6 | |

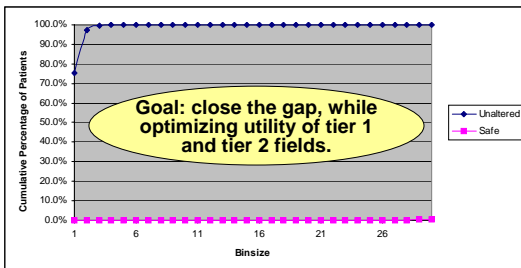
If data provided daily, then full date of visit can be inferred even though only year is provided!

Risk Assessment of Bio-Surveillance DataStream Safe Harbor Version for State of New York



Results from the Risk Assessment Server (provided by Privacert).
Sample: New York Hospital Data for 1990.

Risk Assessment of Bio-Surveillance DataStream Safe Harbor Version for State of New York



Goal: close the gap, while optimizing utility of tier 1 and tier 2 fields.

Results from the Risk Assessment Server (provided by Privacert).
Sample: New York Hospital Data for 2000.

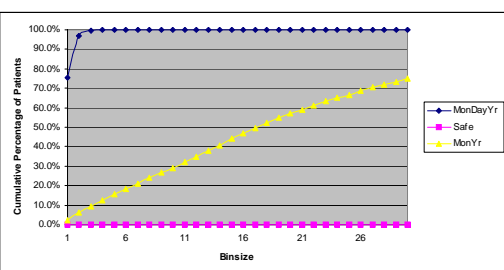
Fields of Bio-Surveillance Datastream Based on Usefulness For Detection -NY

| | Description | Name |
|--------|-------------------------------------|------|
| Tier 1 | Date of visit (month, day and year) | Date |
| | Patient 5-digit ZIP code | ZIP |
| | ICD9 diagnosis code 1 | Dx1 |
| | ICD9 diagnosis code 2 | |
| | ICD9 diagnosis code 3 | |
| | ICD9 diagnosis code 4 | |
| Tier 2 | ICD9 diagnosis code 5 | |
| | ICD9 diagnosis code 6 | |
| | Month, day and Year of Birth | DOB |
| | Gender | Sex |

Decision 1: change DOB to month and year of birth

Results from the Risk Assessment Server (provided by Privacert).
Sample: New York Hospital Data for 2000.

Risk Assessment of Bio-Surveillance DataStream, Change DOB to Report Month and Year of Birth



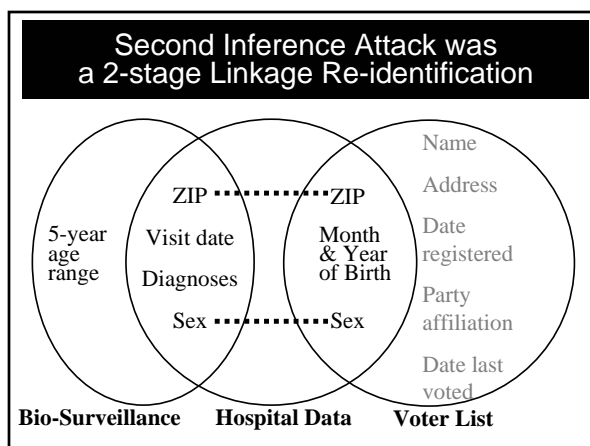
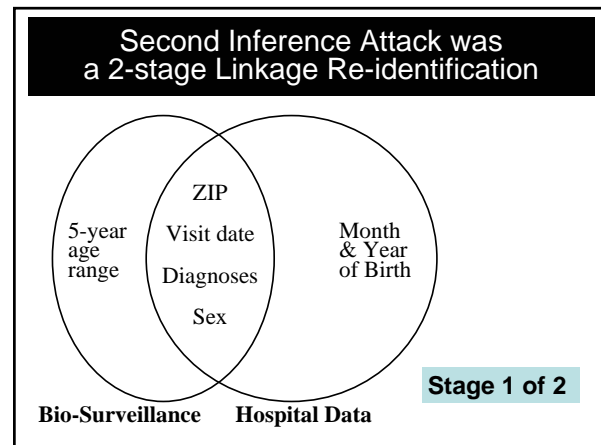
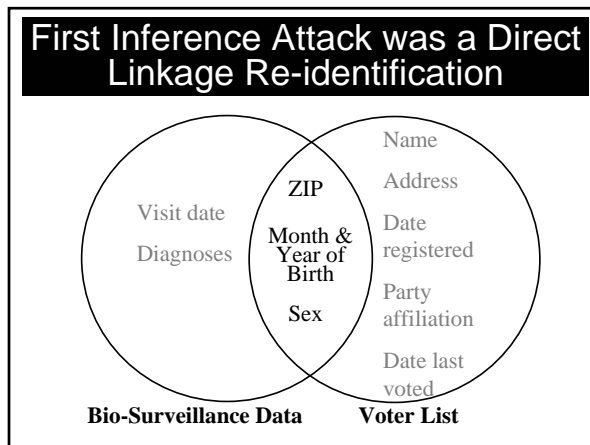
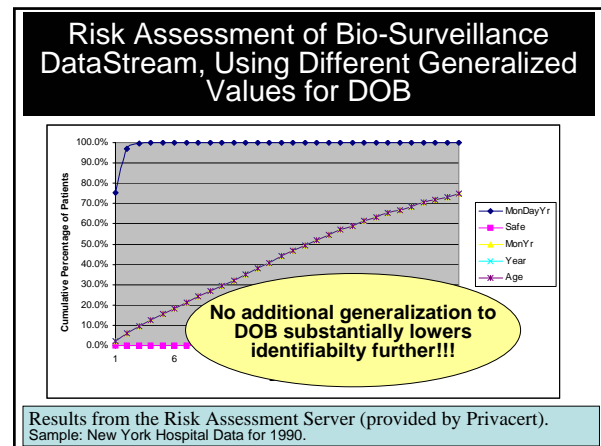
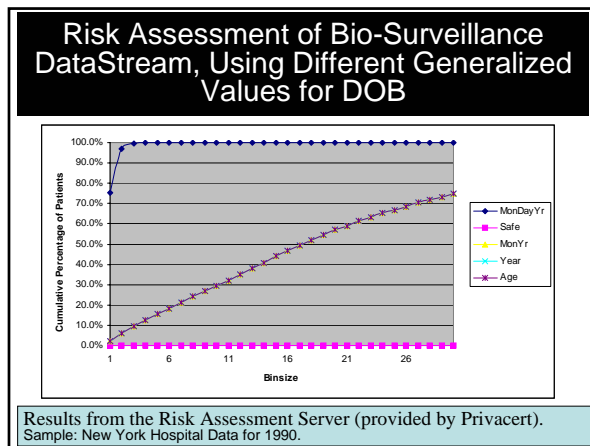
Results from the Risk Assessment Server (provided by Privacert).
Sample: New York Hospital Data for 1990.

Fields of Bio-Surveillance Datastream Based on Usefulness For Detection -NY

| | Description | Name |
|--------|-------------------------------------|------|
| Tier 1 | Date of visit (month, day and year) | Date |
| | Patient 5-digit ZIP code | ZIP |
| | ICD9 diagnosis code 1 | Dx1 |
| | ICD9 diagnosis code 2 | |
| | ICD9 diagnosis code 3 | |
| | ICD9 diagnosis code 4 | |
| Tier 2 | ICD9 diagnosis code 5 | |
| | ICD9 diagnosis code 6 | |
| | Month, day and Year of Birth | DOB |
| | Gender | Sex |

Generalize DOB more?

Given the improvement realized when date of birth was generalized to month and year of birth in in NY data, one might falsely believe generalizing DOB values further to year of birth, age or a 5-year range would provide further improvements -- not so!

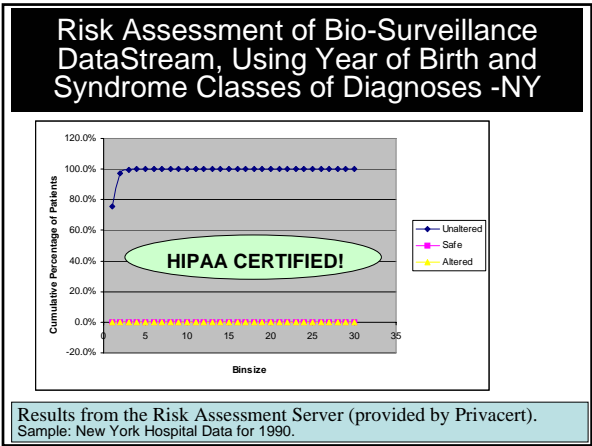


Fields of Bio-Surveillance Datastream Based on Usefulness For Detection -NY

| | Description | Name |
|--------|-------------------------------------|------|
| Tier 1 | Date of visit (month, day and year) | Date |
| | Patient 5-digit ZIP code | ZIP |
| | ICD9 diagnosis code 1 | Dx1 |
| | ICD9 diagnosis code 2 | Dx2 |
| | ICD9 diagnosis code 3 | Dx3 |
| | ICD9 diagnosis code 4 | Dx4 |
| Tier 2 | ICD9 diagnosis code 5 | Dx5 |
| | ICD9 diagnosis code 6 | Dx6 |
| | Month and Year of Birth | DOB |
| | Gender | Sex |

Decision 3. Group diagnosis codes into syndrome or sub-syndrome classes

Results from the Risk Assessment Server (provided by Privacert).
Sample: New York Hospital Data for 2000.



Fields of Bio-Surveillance Datastream Based on Usefulness For Detection -NY

| | Description | Name |
|--------|-------------------------------------|------|
| Tier 1 | Date of visit (month, day and year) | Date |
| | Patient 5-digit ZIP code | ZIP |
| | Syndrome subclass for dx1 | Dx1 |
| | Syndrome subclass for dx2 | Dx2 |
| | Syndrome subclass for dx3 | Dx3 |
| | Syndrome subclass for dx4 | Dx4 |
| Tier 2 | Syndrome subclass for dx5 | Dx5 |
| | Syndrome subclass for dx6 | Dx6 |
| | Year of birth | DOB |
| | Gender | Sex |

Results from the Risk Assessment Server (provided by Privacert).
Sample: New York Hospital Data for 2000.

Carnegie Mellon
DATA PRIVACY LAB

Privacy Technology

1. Privacy is here to stay.
2. Computer scientist must help solve this problem.
3. Selective Revelation
4. Example: video surveillance
5. Example: bio-terrorism surveillance
6. Example: identity theft
7. Example: distributed surveillance
8. Example: privacy-preserving surveillance
9. Example: DNA privacy
10. Example: Identity theft protections
11. Example: k-Anonymity
12. Example: Webcam surveillance
13. Example: Text de-identification
14. Example: Policy specification and enforcement
15. Example: Scam Spam

privacy.cs.cmu.edu