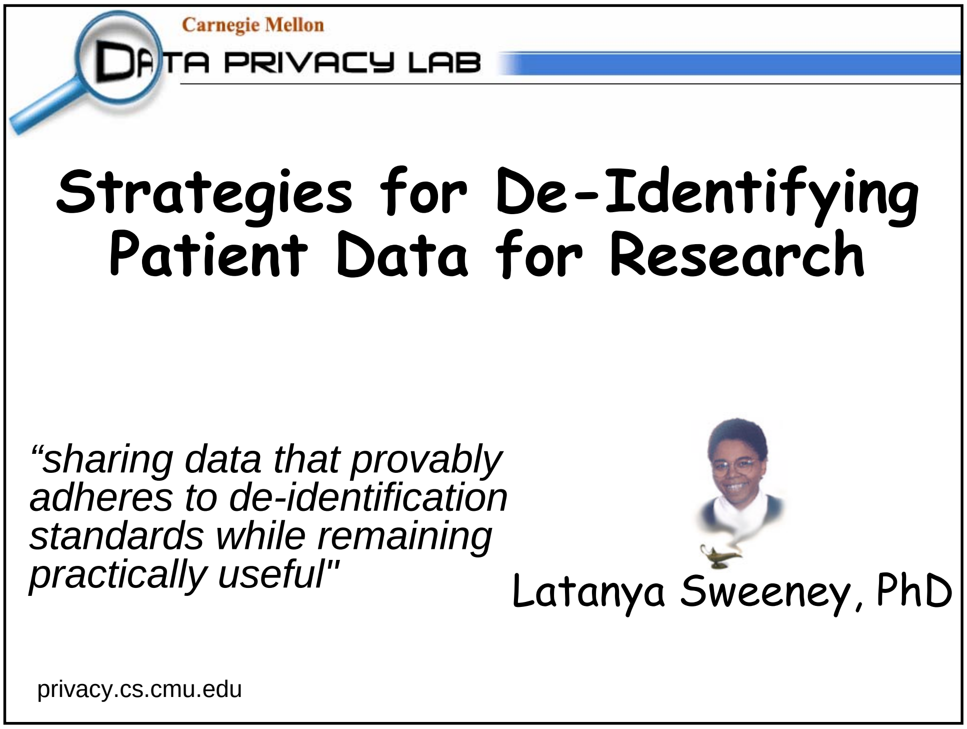


Carnegie Mellon

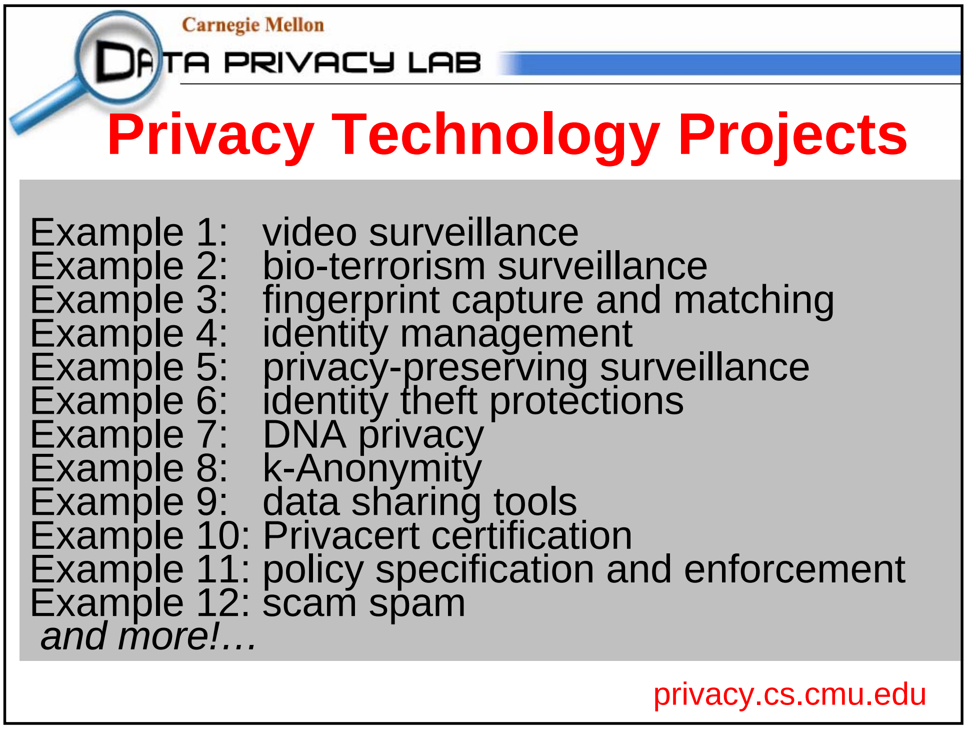
DATA PRIVACY LAB

Strategies for De-Identifying Patient Data for Research

"sharing data that provably adheres to de-identification standards while remaining practically useful"

Latanya Sweeney, PhD

privacy.cs.cmu.edu



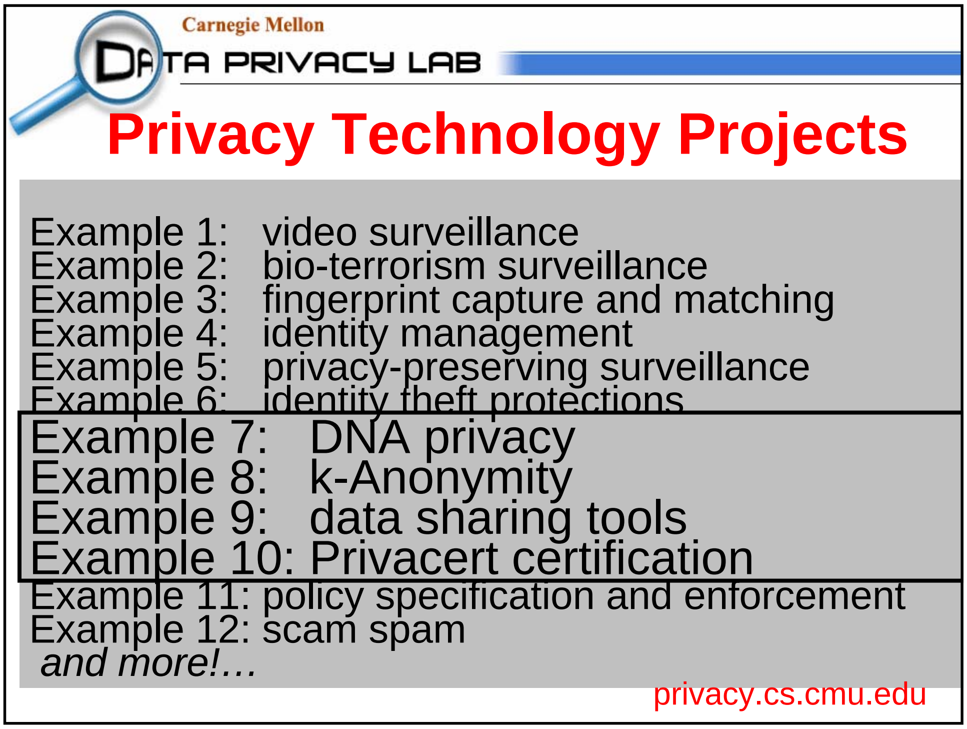
Carnegie Mellon

DATA PRIVACY LAB

Privacy Technology Projects

Example 1: video surveillance
Example 2: bio-terrorism surveillance
Example 3: fingerprint capture and matching
Example 4: identity management
Example 5: privacy-preserving surveillance
Example 6: identity theft protections
Example 7: DNA privacy
Example 8: k-Anonymity
Example 9: data sharing tools
Example 10: Privacert certification
Example 11: policy specification and enforcement
Example 12: scam spam
and more!…

privacy.cs.cmu.edu

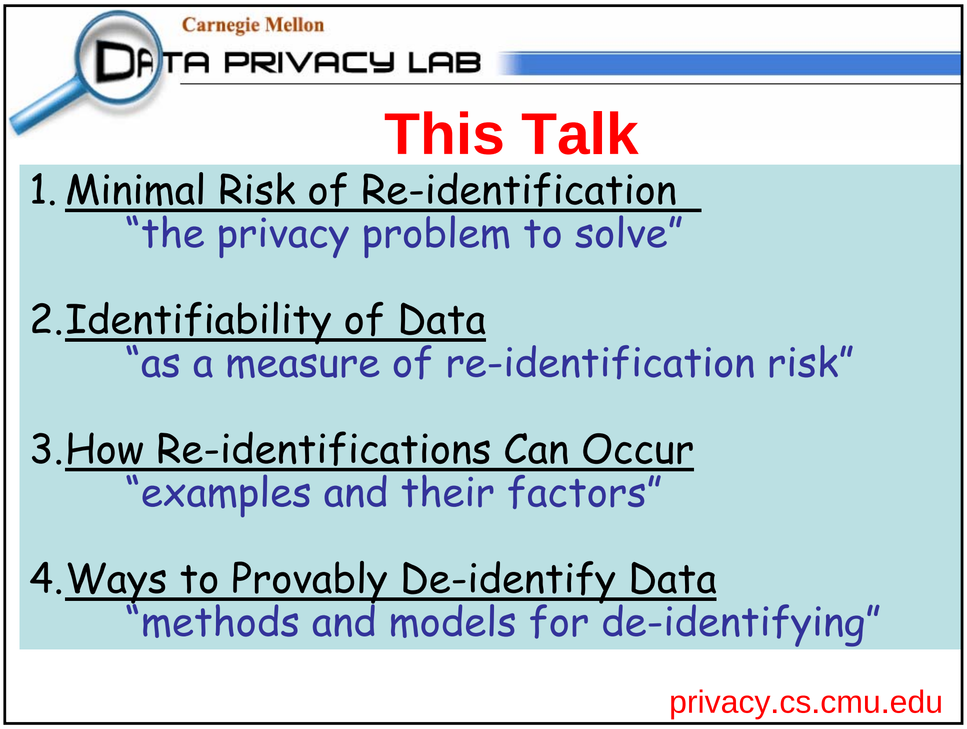


Carnegie Mellon
DATA PRIVACY LAB

Privacy Technology Projects

Example 1: video surveillance
Example 2: bio-terrorism surveillance
Example 3: fingerprint capture and matching
Example 4: identity management
Example 5: privacy-preserving surveillance
Example 6: identity theft protections
Example 7: DNA privacy
Example 8: k-Anonymity
Example 9: data sharing tools
Example 10: Privacert certification
Example 11: policy specification and enforcement
Example 12: scam spam
and more!…

privacy.cs.cmu.edu



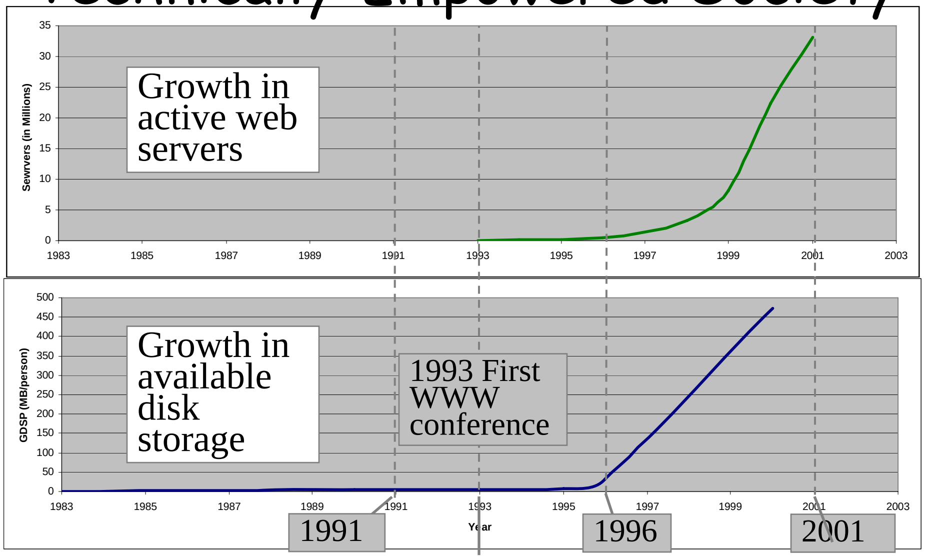
Carnegie Mellon
DATA PRIVACY LAB

This Talk

1. Minimal Risk of Re-identification
   "the privacy problem to solve"
2. Identifiability of Data
   "as a measure of re-identification risk"
3. How Re-identifications Can Occur
   "examples and their factors"
4. Ways to Provably De-identify Data
   "methods and models for de-identifying"

privacy.cs.cmu.edu

Technically-Empowered Society



Growth in active web servers

Growth in available disk storage

1993 First WWW conference

1991

1996

2001

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*

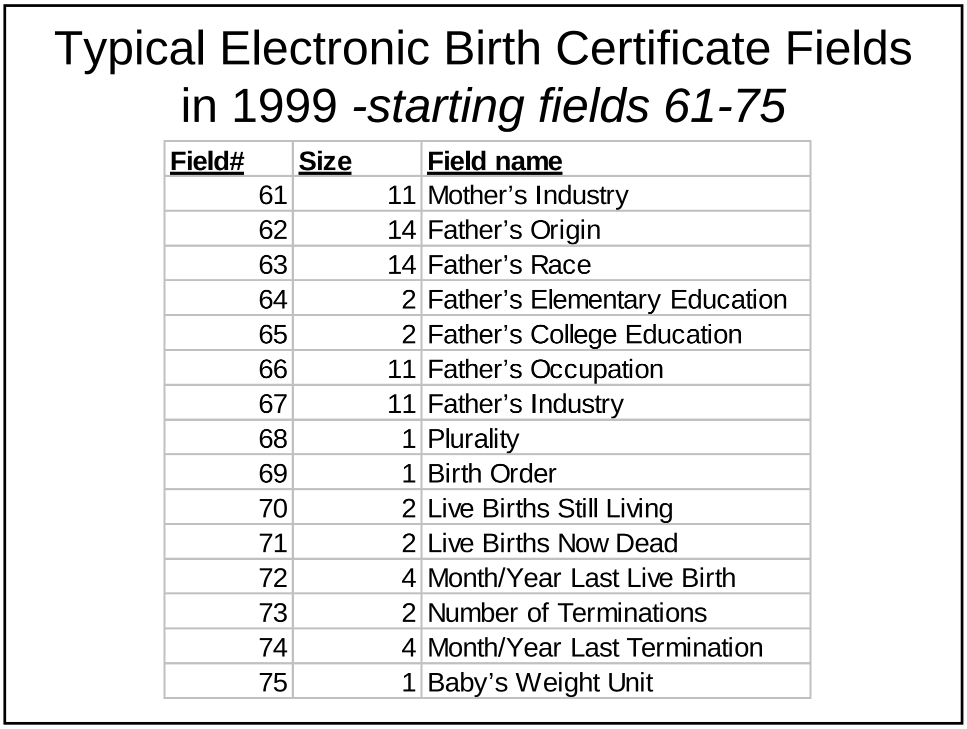
Field#	**Size**	**Field name**
16	20	County of Birth
17	6	Certifier’s Code
18	30	Certifier’s Name
19	1	Certifier’s Title
20	30	Attendant’s Name
21	1	Attendant’s Title
22	23	Attendant’s Address
23	19	Attendant’s City
24	2	Attendant’s State
25	10	Attendant’s Zip Code
26	50	Mother’s First Name
27	50	Mother’s 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

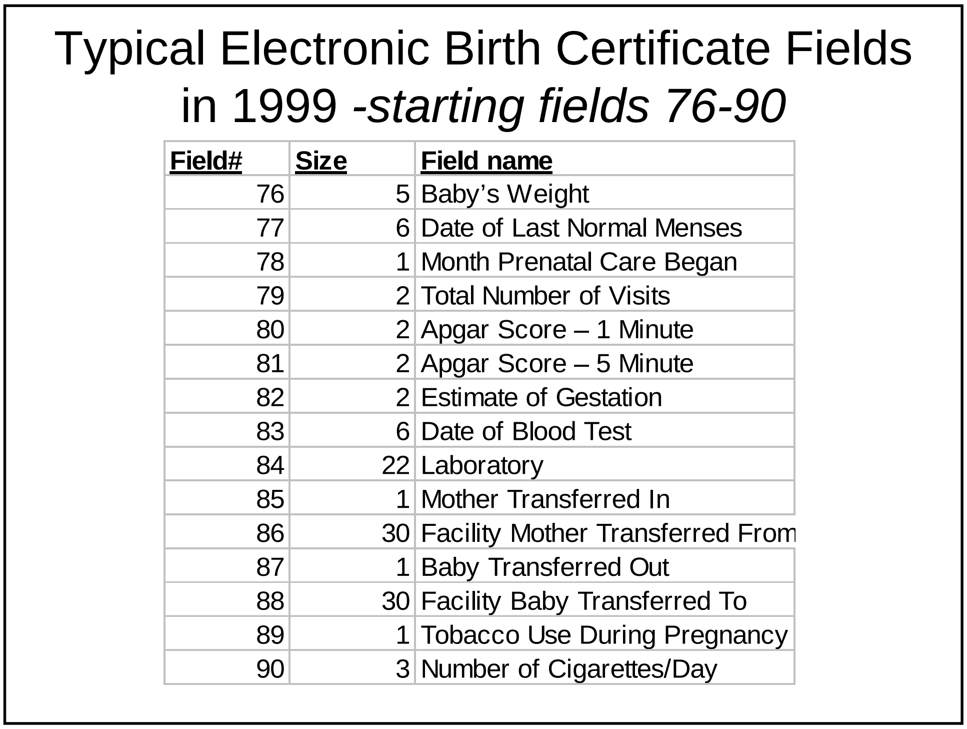
Typical Electronic Birth Certificate Fields in 1999 *-starting fields 46-60*

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



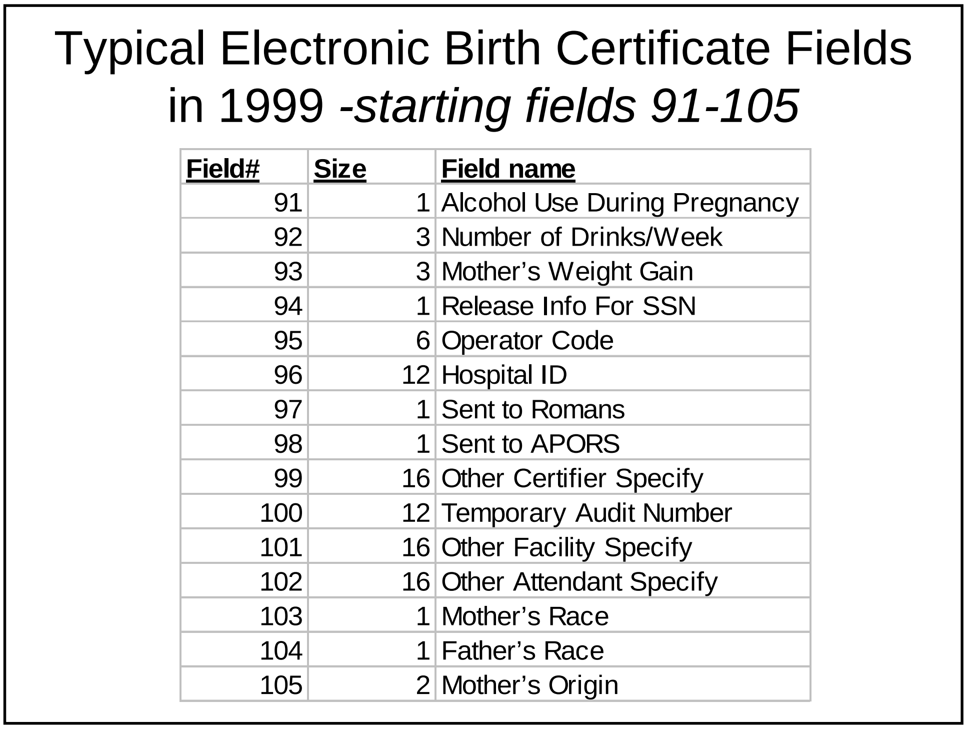
Typical Electronic Birth Certificate Fields in 1999 *-starting fields 61-75*

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



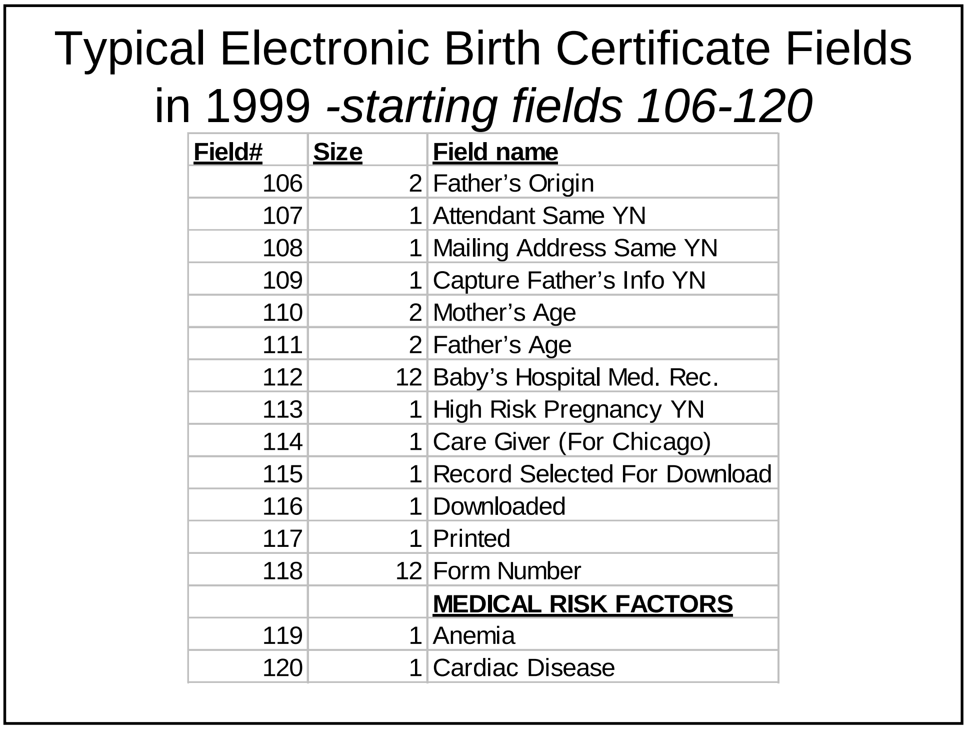
Typical Electronic Birth Certificate Fields in 1999 *-starting fields 76-90*

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



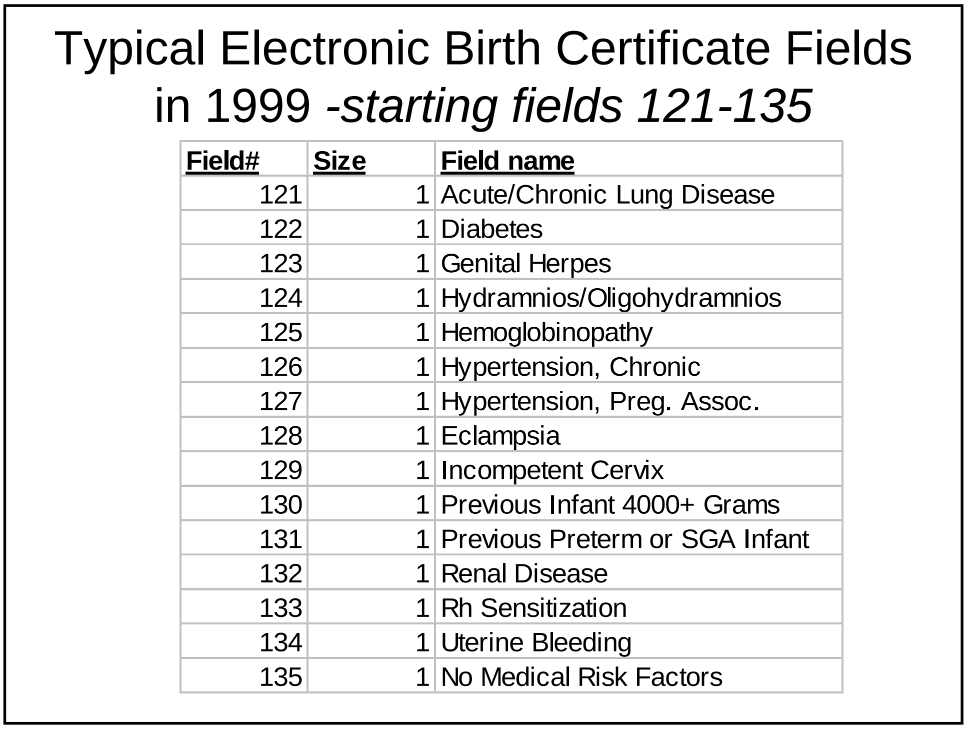
Typical Electronic Birth Certificate Fields in 1999 *-starting fields 91-105*

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



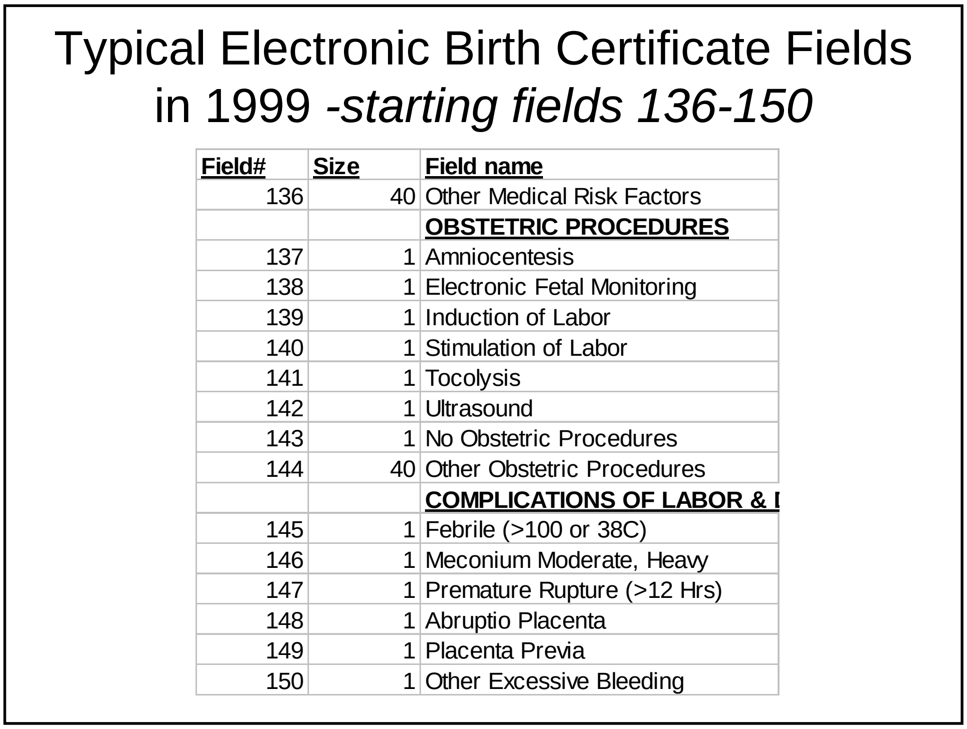
Typical Electronic Birth Certificate Fields in 1999 *-starting fields 106-120*

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



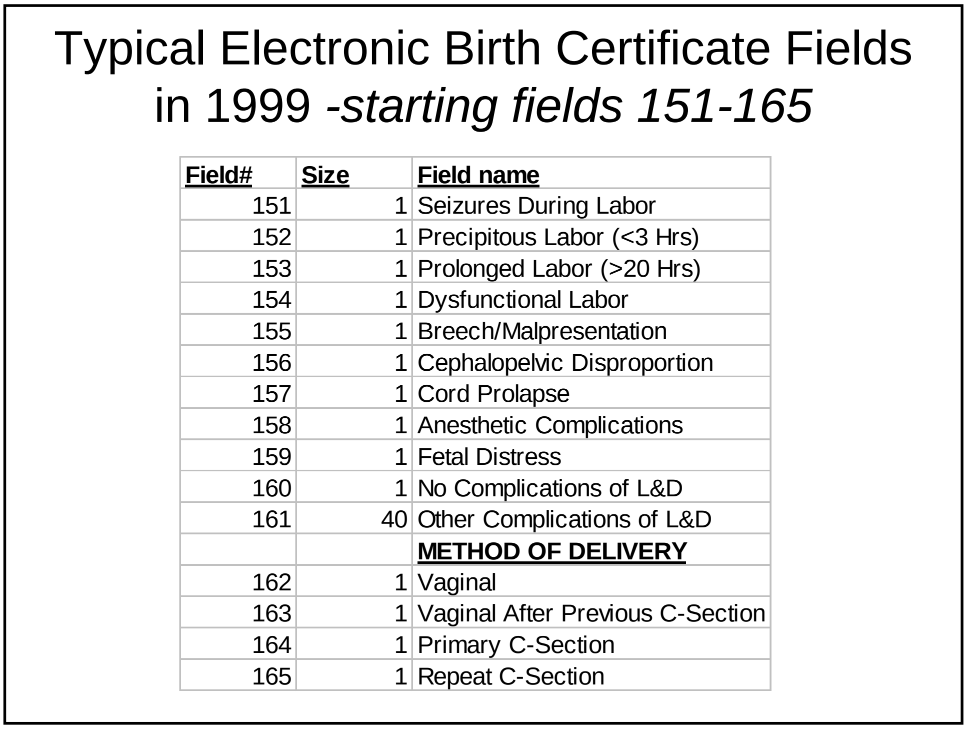
Typical Electronic Birth Certificate Fields in 1999 *-starting fields 121-135*

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



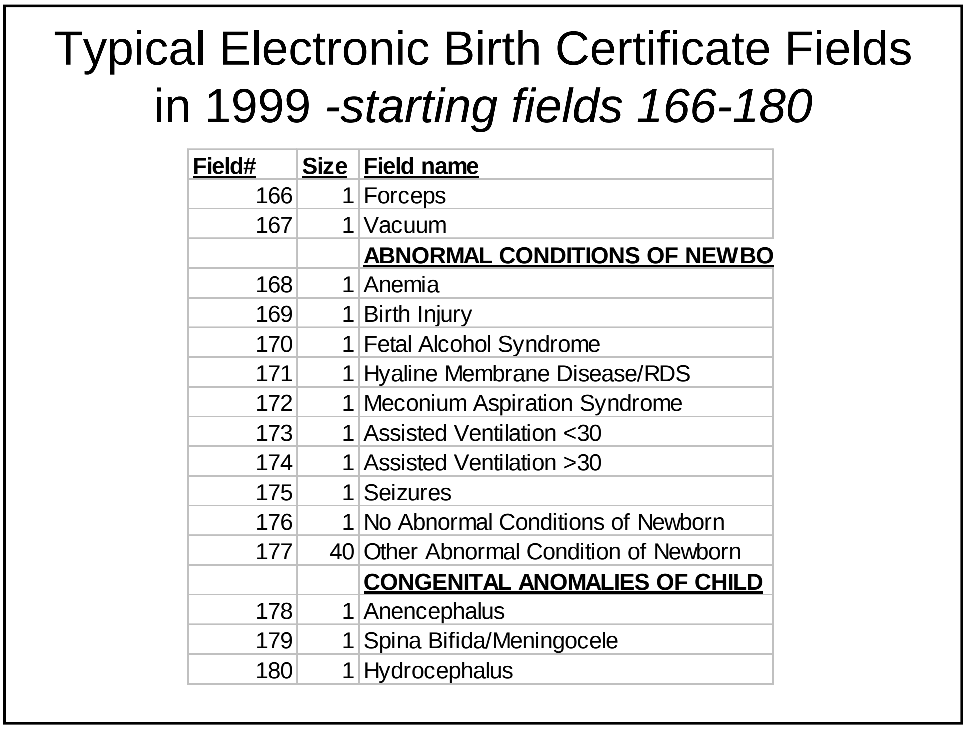
Typical Electronic Birth Certificate Fields in 1999 *-starting fields 136-150*

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 &
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



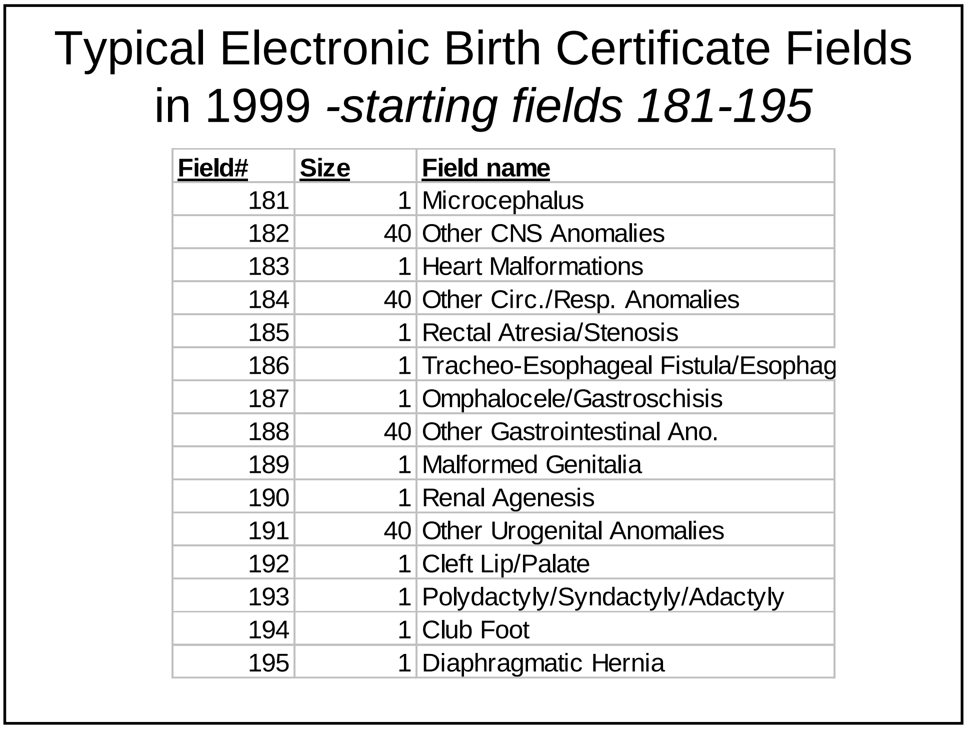
Typical Electronic Birth Certificate Fields in 1999 *-starting fields 151-165*

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



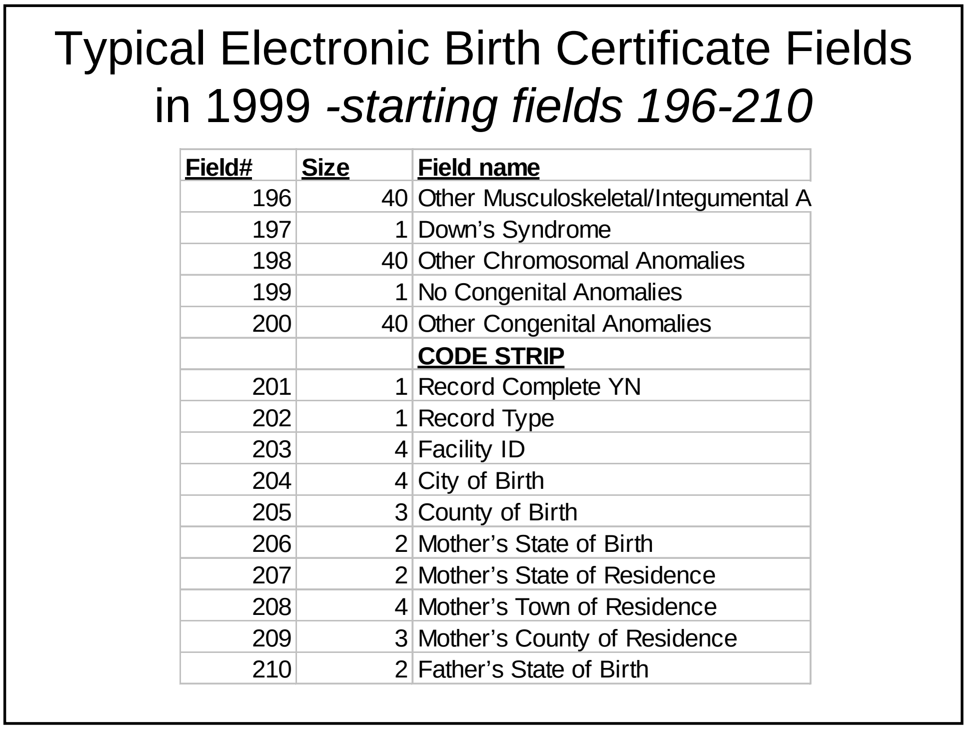
Typical Electronic Birth Certificate Fields in 1999 *-starting fields 166-180*

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



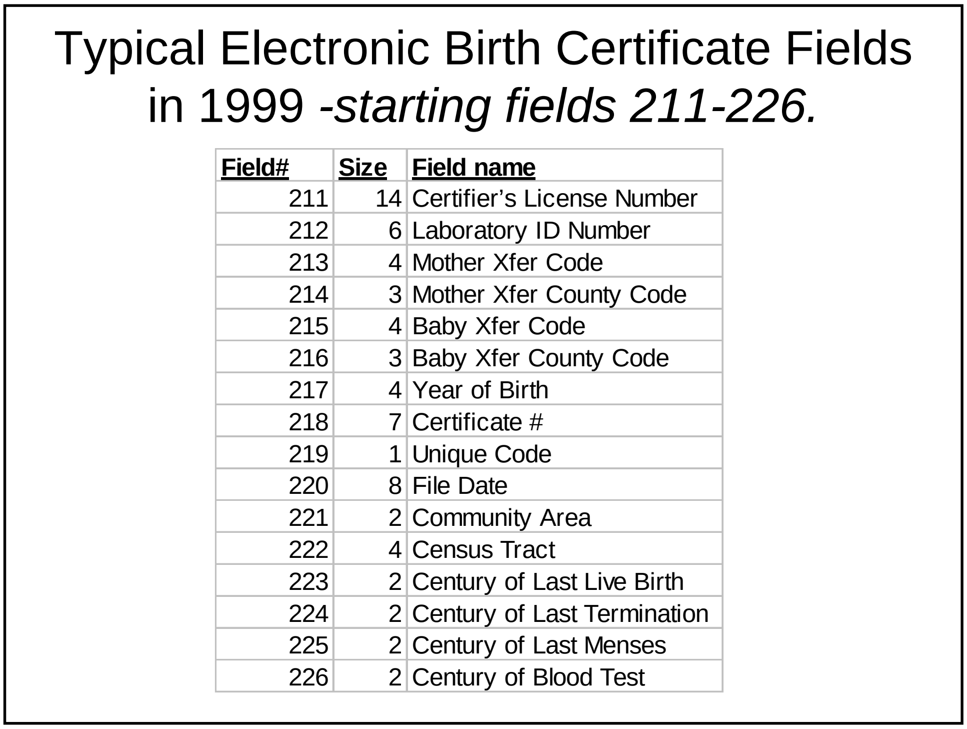
Typical Electronic Birth Certificate Fields in 1999 *-starting fields 181-195*

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	Polydactyly/Syndactyly/Adactyly
194	1	Club Foot
195	1	Diaphragmatic Hernia



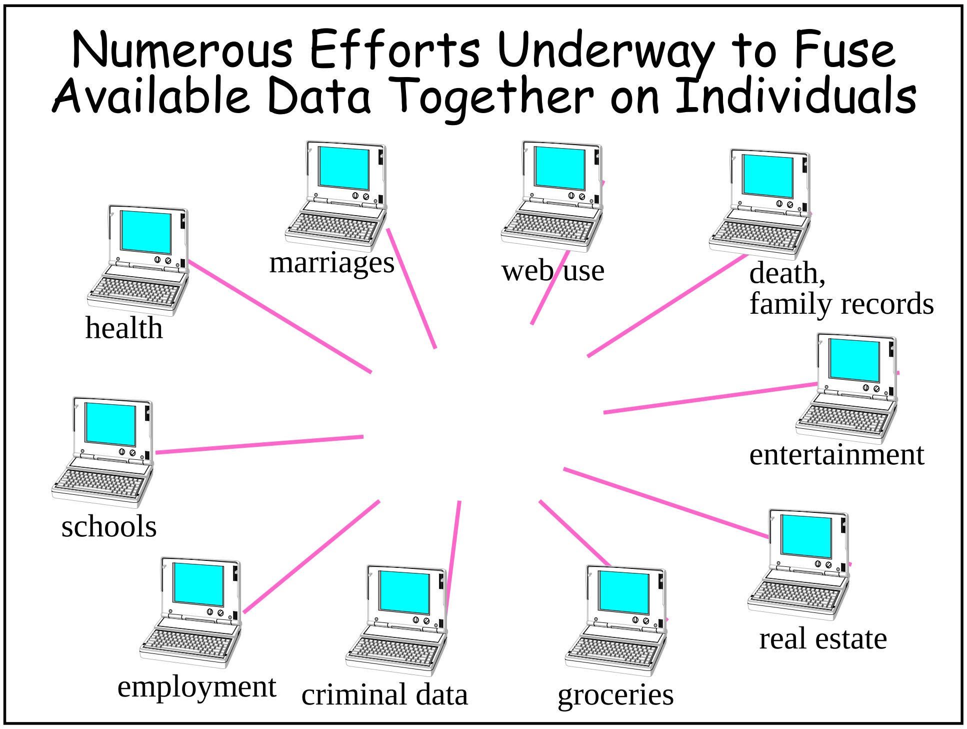
Typical Electronic Birth Certificate Fields in 1999 *-starting fields 196-210*

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

health

marriages

web use

death, family records

entertainment

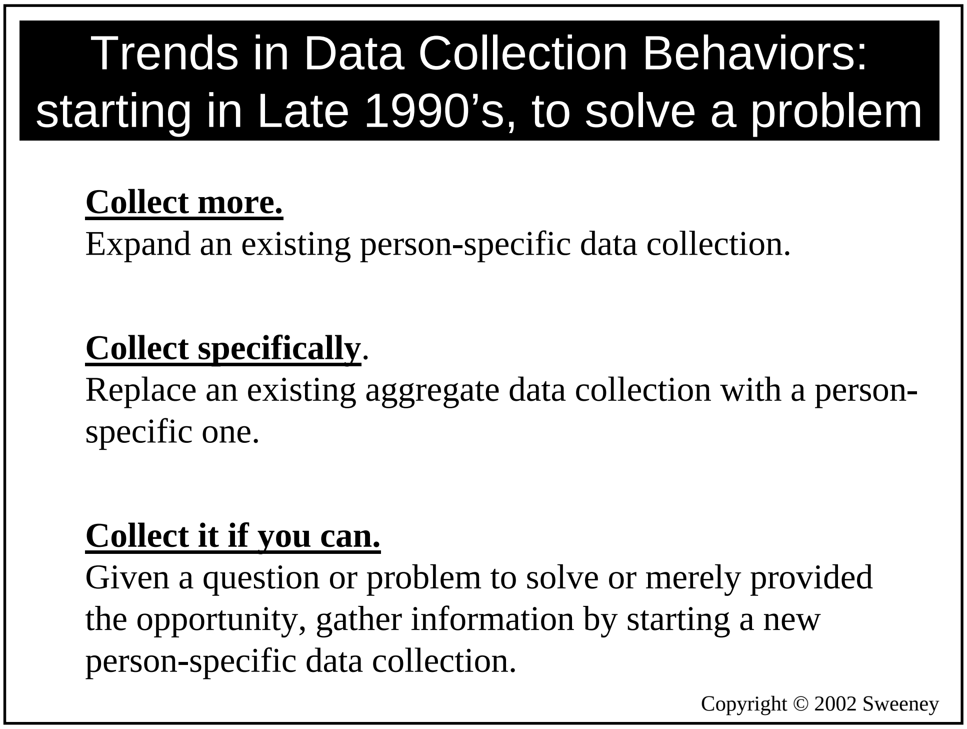
schools

real estate

employment

criminal data

groceries

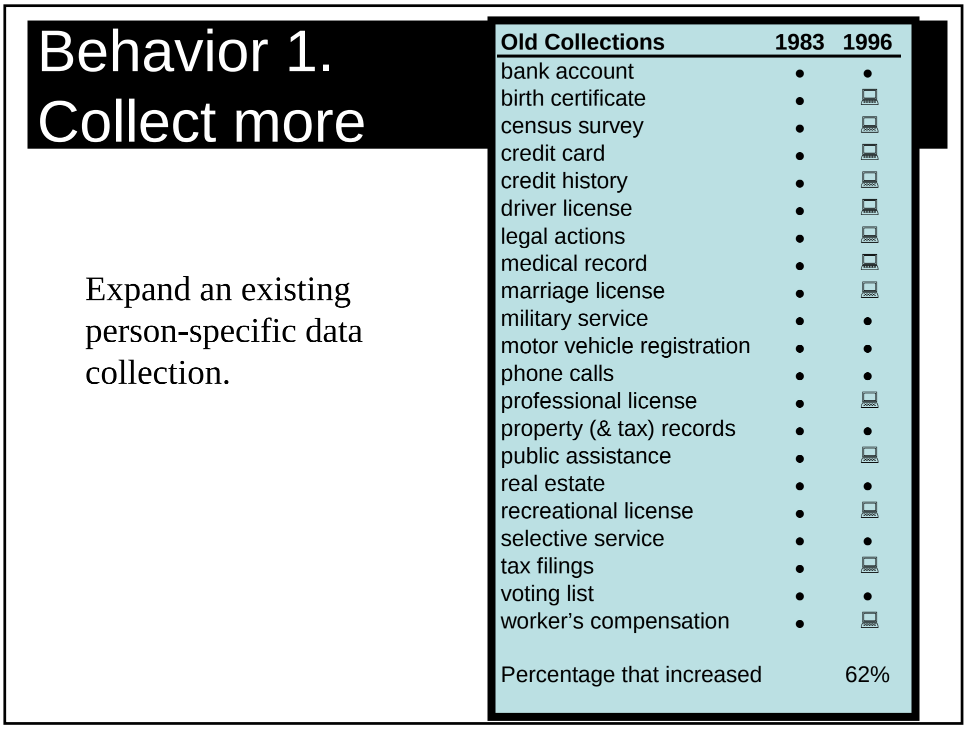


Trends in Data Collection Behaviors: starting in Late 1990's, to solve a problem

Collect more.
Expand an existing person-specific data collection.

Collect specifically.
Replace an existing aggregate data collection with a person-specific one.

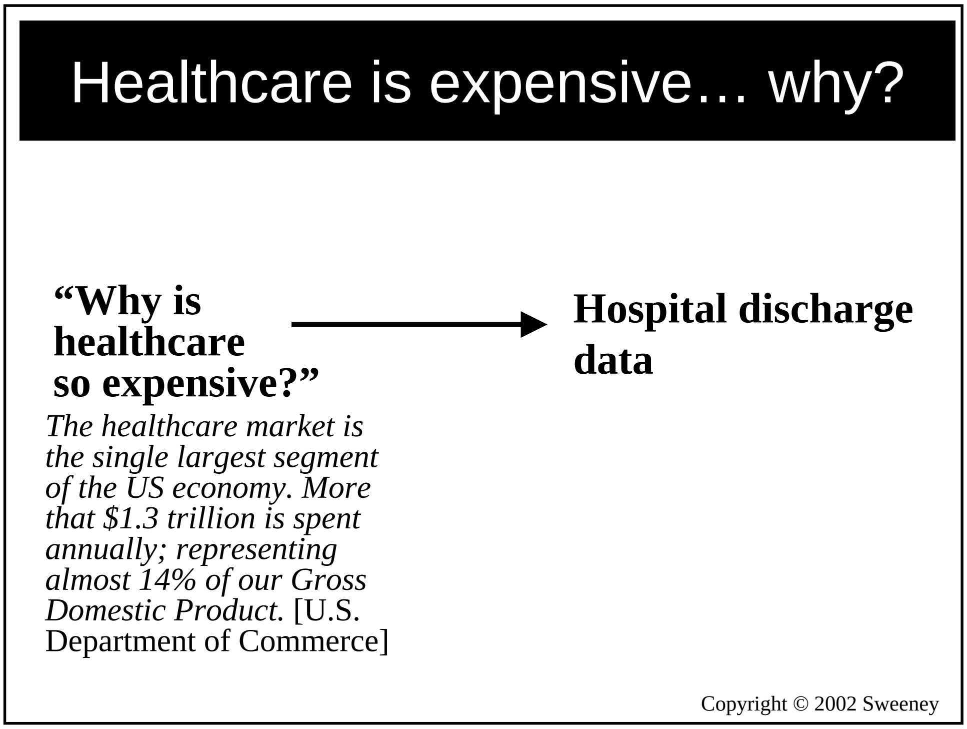
Collect it if you can.
Given a question or problem to solve or merely provided the opportunity, gather information by starting a new person-specific data collection.



Behavior 1. Collect more

Expand an existing person-specific data collection.

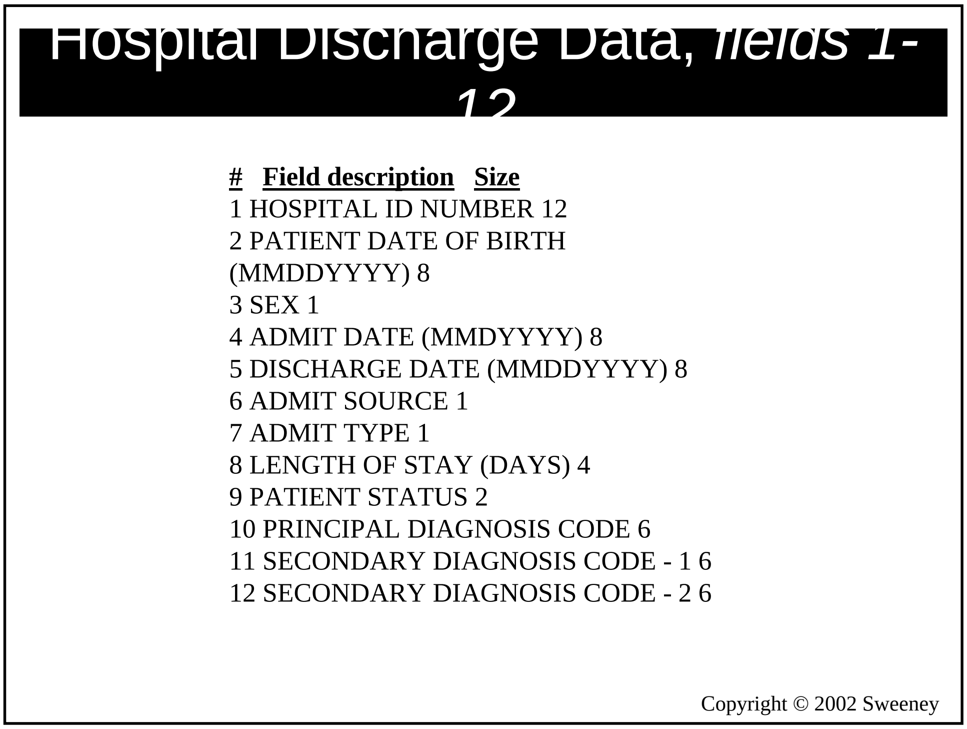
Old Collections	1983	1996
bank account	•	•
birth certificate	•	💻
census survey	•	💻
credit card	•	💻
credit history	•	💻
driver license	•	💻
legal actions	•	💻
medical record	•	💻
marriage license	•	💻
military service	•	•
motor vehicle registration	•	•
phone calls	•	•
professional license	•	💻
property (& tax) records	•	•
public assistance	•	💻
real estate	•	•
recreational license	•	💻
selective service	•	•
tax filings	•	💻
voting list	•	•
worker's compensation	•	💻
Percentage that increased		62%



Healthcare is expensive… why?

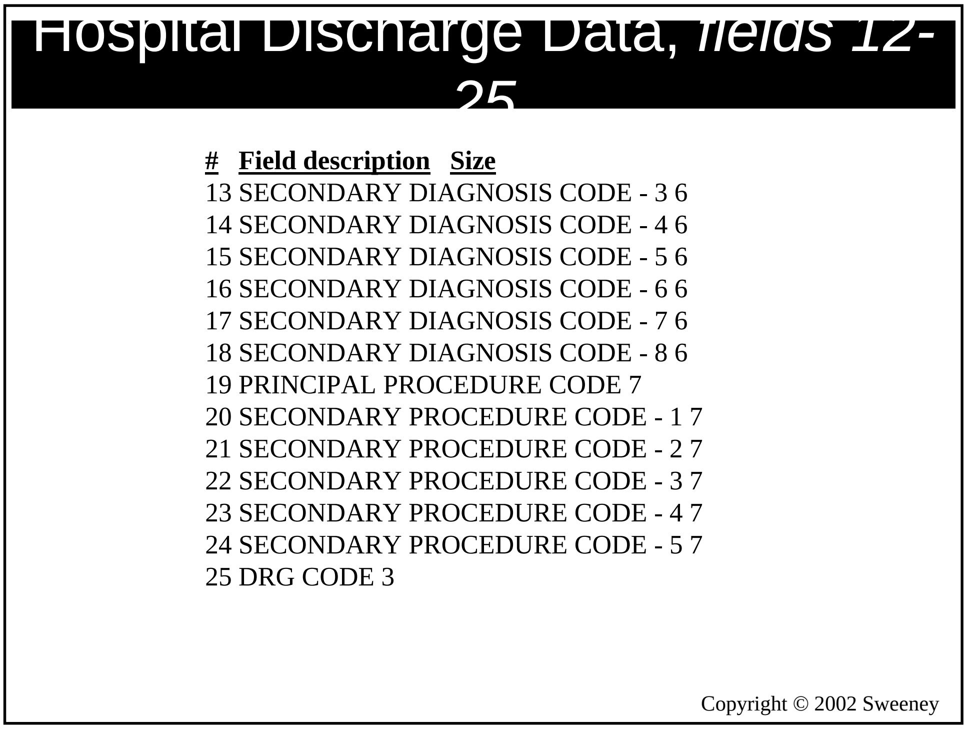
“Why is healthcare so expensive?” → **Hospital discharge data**

The healthcare market is the single largest segment of the US economy. More that $1.3 trillion is spent annually; representing almost 14% of our Gross Domestic Product. [U.S. Department of Commerce]



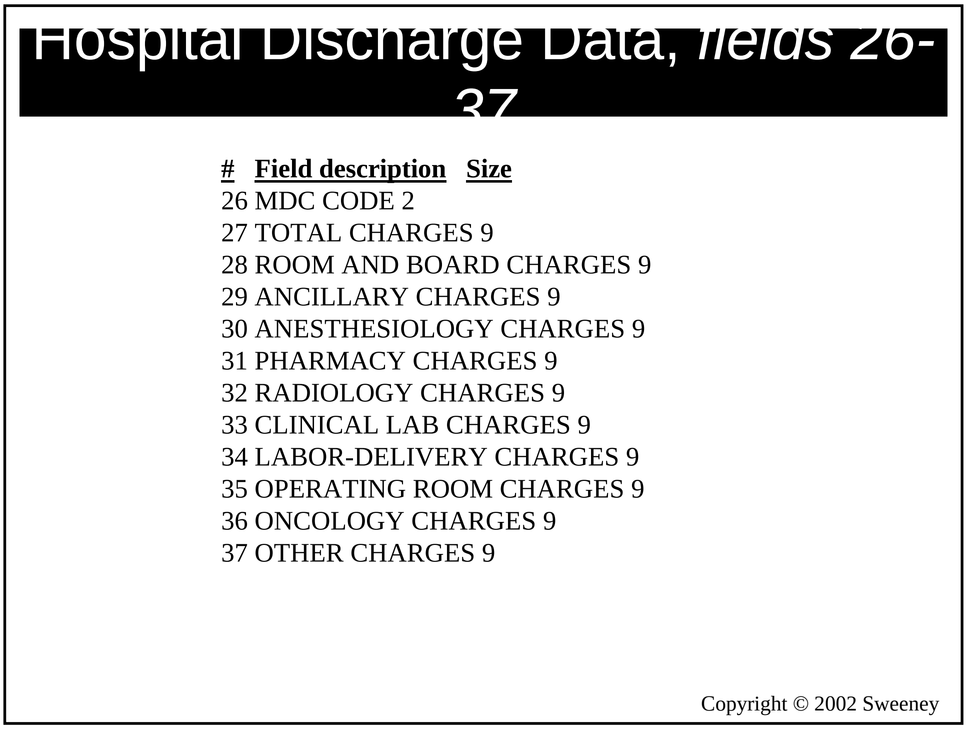
Hospital Discharge Data, *fields 1-12*

# Field description Size
1 HOSPITAL ID NUMBER 12
2 PATIENT DATE OF BIRTH
(MMDDYYYY) 8
3 SEX 1
4 ADMIT DATE (MMDYYYY) 8
5 DISCHARGE DATE (MMDDYYYY) 8
6 ADMIT SOURCE 1
7 ADMIT TYPE 1
8 LENGTH OF STAY (DAYS) 4
9 PATIENT STATUS 2
10 PRINCIPAL DIAGNOSIS CODE 6
11 SECONDARY DIAGNOSIS CODE - 1 6
12 SECONDARY DIAGNOSIS CODE - 2 6



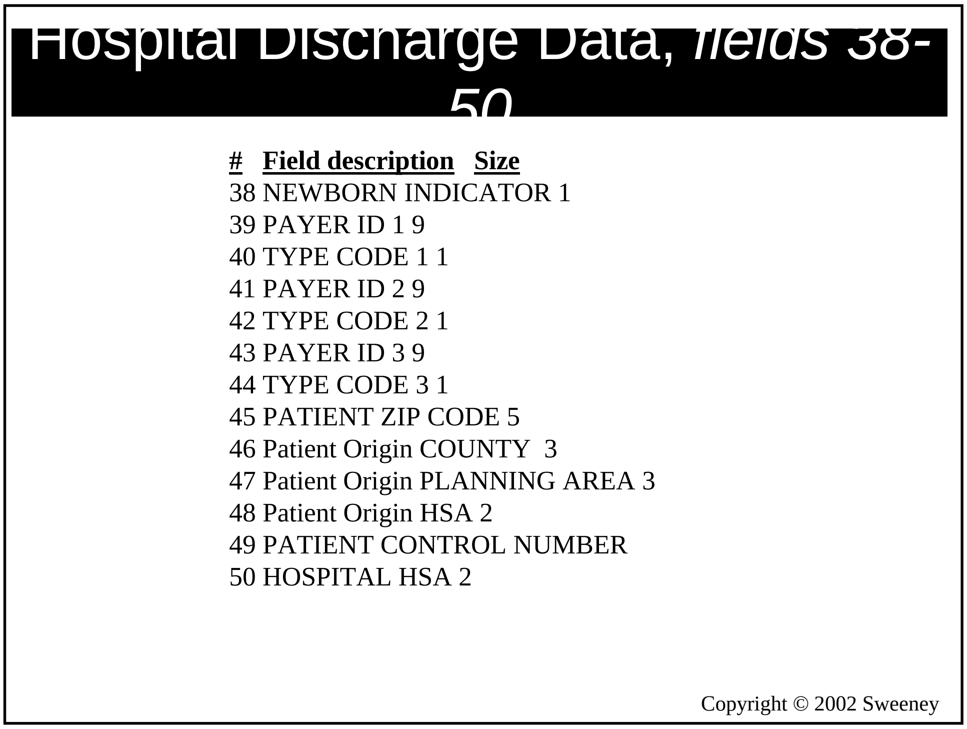
Hospital Discharge Data, *fields 12-25*

#	Field description	Size
13	SECONDARY DIAGNOSIS CODE - 3	6
14	SECONDARY DIAGNOSIS CODE - 4	6
15	SECONDARY DIAGNOSIS CODE - 5	6
16	SECONDARY DIAGNOSIS CODE - 6	6
17	SECONDARY DIAGNOSIS CODE - 7	6
18	SECONDARY DIAGNOSIS CODE - 8	6
19	PRINCIPAL PROCEDURE CODE	7
20	SECONDARY PROCEDURE CODE - 1	7
21	SECONDARY PROCEDURE CODE - 2	7
22	SECONDARY PROCEDURE CODE - 3	7
23	SECONDARY PROCEDURE CODE - 4	7
24	SECONDARY PROCEDURE CODE - 5	7
25	DRG CODE	3



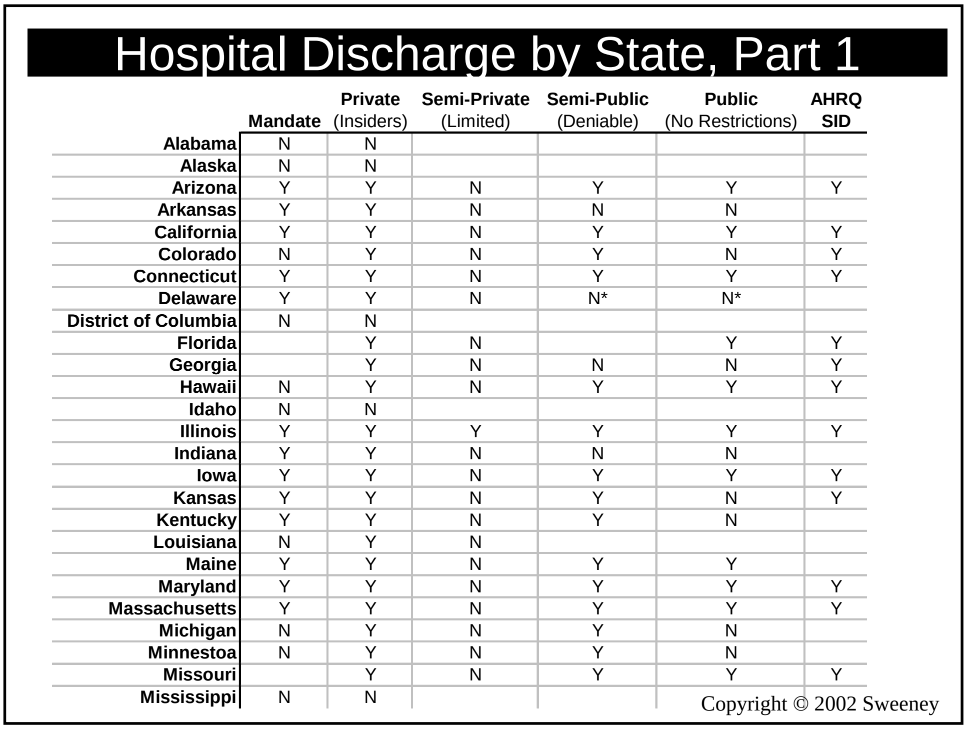
Hospital Discharge Data, *fields 26-37*

#	Field description	Size
26	MDC CODE	2
27	TOTAL CHARGES	9
28	ROOM AND BOARD CHARGES	9
29	ANCILLARY CHARGES	9
30	ANESTHESIOLOGY CHARGES	9
31	PHARMACY CHARGES	9
32	RADIOLOGY CHARGES	9
33	CLINICAL LAB CHARGES	9
34	LABOR-DELIVERY CHARGES	9
35	OPERATING ROOM CHARGES	9
36	ONCOLOGY CHARGES	9
37	OTHER CHARGES	9



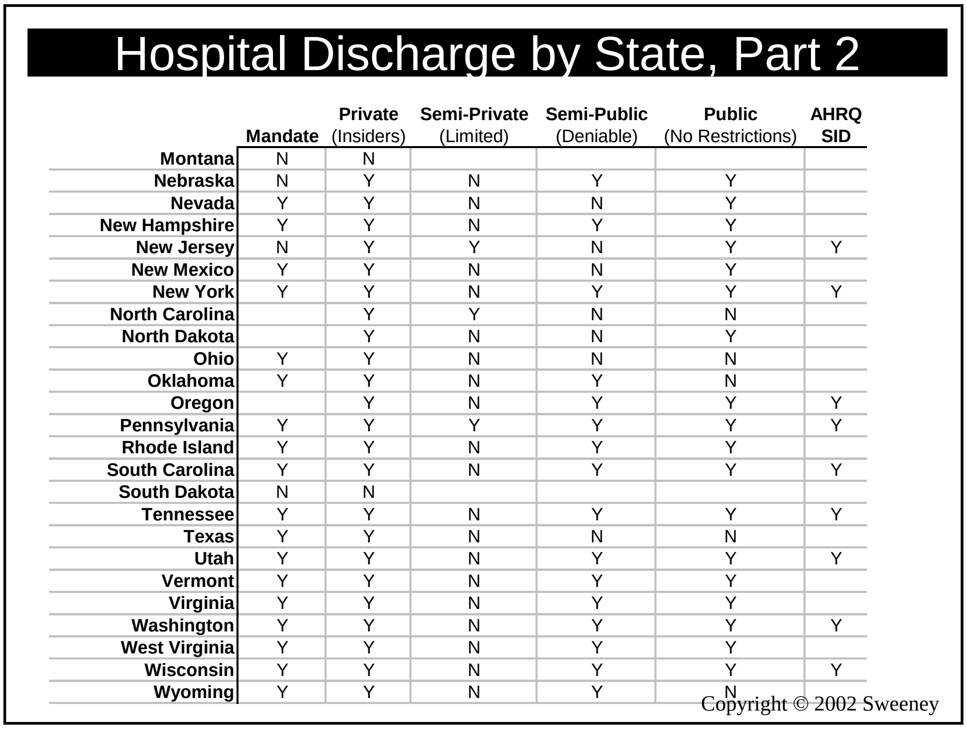
Hospital Discharge Data, *fields 38-50*

#	Field description	Size
38	NEWBORN INDICATOR	1
39	PAYER ID 1	9
40	TYPE CODE 1	1
41	PAYER ID 2	9
42	TYPE CODE 2	1
43	PAYER ID 3	9
44	TYPE CODE 3	1
45	PATIENT ZIP CODE	5
46	Patient Origin COUNTY	3
47	Patient Origin PLANNING AREA	3
48	Patient Origin HSA	2
49	PATIENT CONTROL NUMBER	
50	HOSPITAL HSA	2



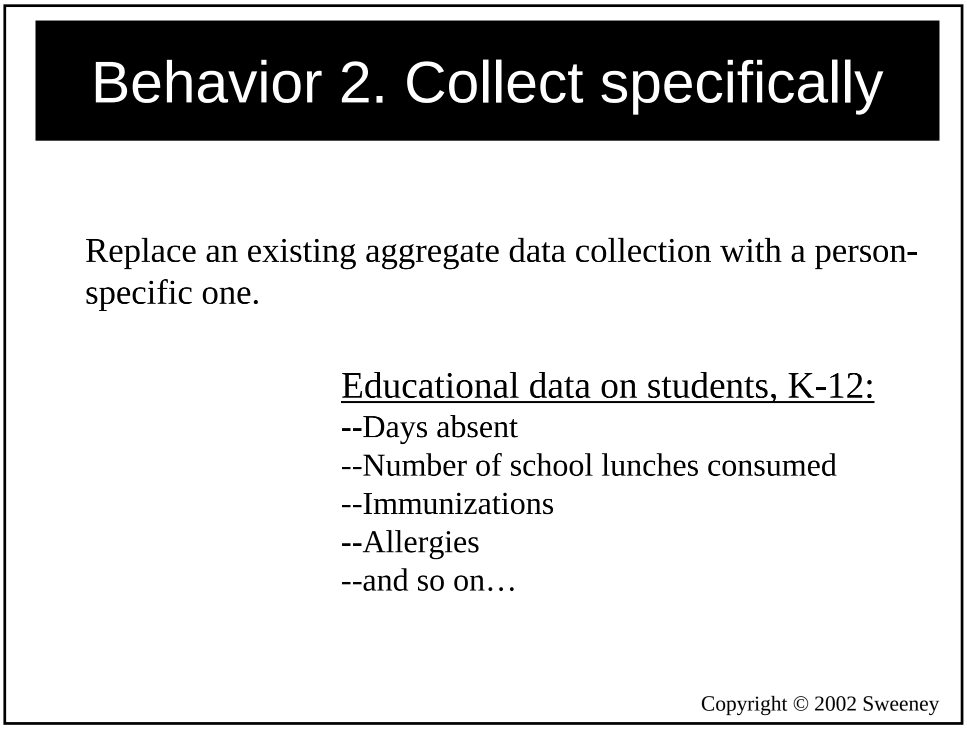
Hospital Discharge by State, Part 1

	Mandate	Private (Insiders)	Semi-Private (Limited)	Semi-Public (Deniable)	Public (No Restrictions)	AHRQ SID
Alabama	N	N				
Alaska	N	N				
Arizona	Y	Y	N	Y	Y	Y
Arkansas	Y	Y	N	N	N	
California	Y	Y	N	Y	Y	Y
Colorado	N	Y	N	Y	N	Y
Connecticut	Y	Y	N	Y	Y	Y
Delaware	Y	Y	N	N*	N*	
District of Columbia	N	N				
Florida		Y	N		Y	Y
Georgia		Y	N	N	N	Y
Hawaii	N	Y	N	Y	Y	Y
Idaho	N	N				
Illinois	Y	Y	Y	Y	Y	Y
Indiana	Y	Y	N	N	N	
Iowa	Y	Y	N	Y	Y	Y
Kansas	Y	Y	N	Y	N	Y
Kentucky	Y	Y	N	Y	N	
Louisiana	N	Y	N			
Maine	Y	Y	N	Y	Y	
Maryland	Y	Y	N	Y	Y	Y
Massachusetts	Y	Y	N	Y	Y	Y
Michigan	N	Y	N	Y	N	
Minnestoa	N	Y	N	Y	N	
Missouri		Y	N	Y	Y	Y
Mississippi	N	N				



Hospital Discharge by State, Part 2

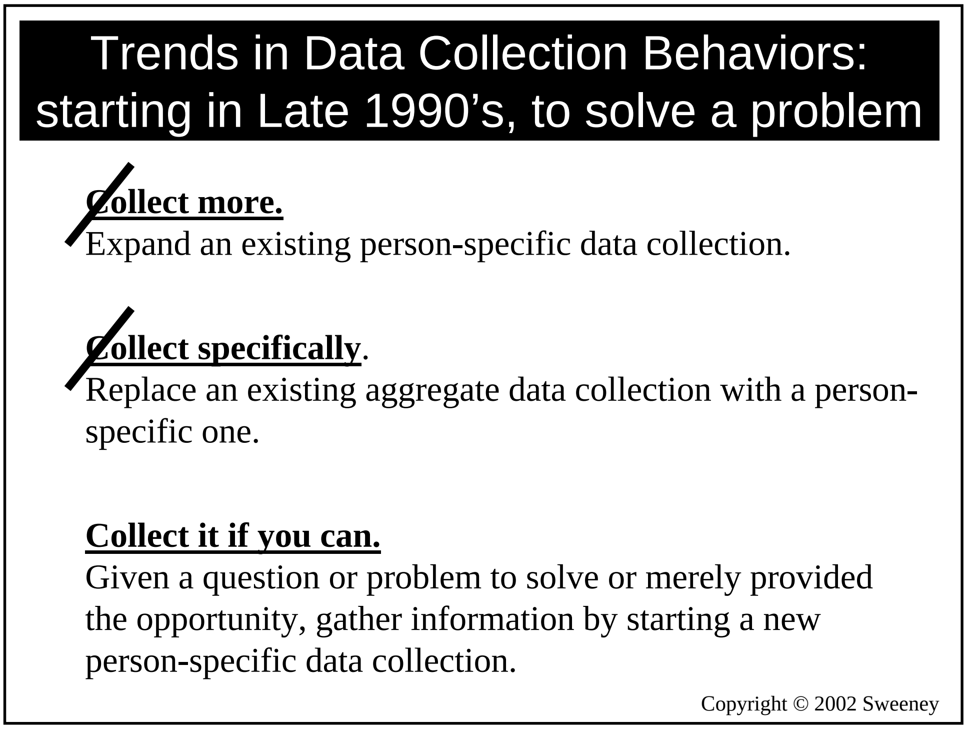
	Mandate	Private (Insiders)	Semi-Private (Limited)	Semi-Public (Deniable)	Public (No Restrictions)	AHRQ SID
Montana	N	N				
Nebraska	N	Y	N	Y	Y	
Nevada	Y	Y	N	N	Y	
New Hampshire	Y	Y	N	Y	Y	
New Jersey	N	Y	Y	N	Y	Y
New Mexico	Y	Y	N	N	Y	
New York	Y	Y	N	Y	Y	Y
North Carolina		Y	Y	N	N	
North Dakota		Y	N	N	Y	
Ohio	Y	Y	N	N	N	
Oklahoma	Y	Y	N	Y	N	
Oregon		Y	N	Y	Y	Y
Pennsylvania	Y	Y	Y	Y	Y	Y
Rhode Island	Y	Y	N	Y	Y	
South Carolina	Y	Y	N	Y	Y	Y
South Dakota	N	N				
Tennessee	Y	Y	N	Y	Y	Y
Texas	Y	Y	N	N	N	
Utah	Y	Y	N	Y	Y	Y
Vermont	Y	Y	N	Y	Y	
Virginia	Y	Y	N	Y	Y	
Washington	Y	Y	N	Y	Y	Y
West Virginia	Y	Y	N	Y	Y	
Wisconsin	Y	Y	N	Y	Y	Y
Wyoming	Y	Y	N	Y	N	



Behavior 2. Collect specifically

Replace an existing aggregate data collection with a person-specific one.

Educational data on students, K-12:
--Days absent
--Number of school lunches consumed
--Immunizations
--Allergies
--and so on…

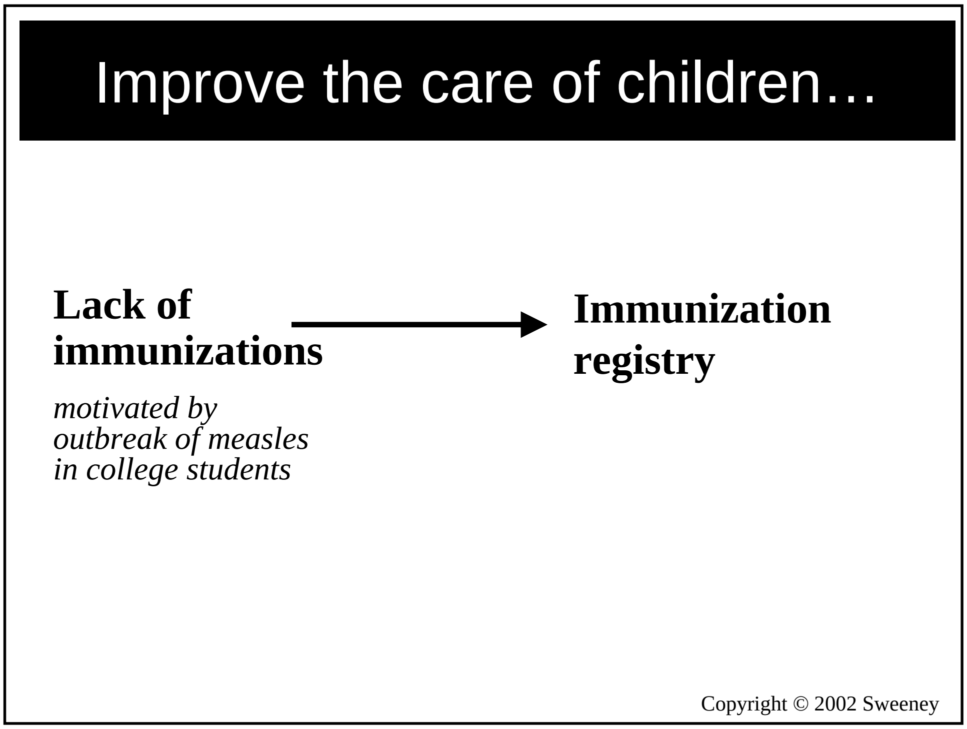


Trends in Data Collection Behaviors: starting in Late 1990's, to solve a problem

~~**Collect more.**~~
Expand an existing person-specific data collection.

~~**Collect specifically**.~~
Replace an existing aggregate data collection with a person-specific one.

Collect it if you can.
Given a question or problem to solve or merely provided the opportunity, gather information by starting a new person-specific data collection.



Improve the care of children…

Lack of immunizations → **Immunization registry**

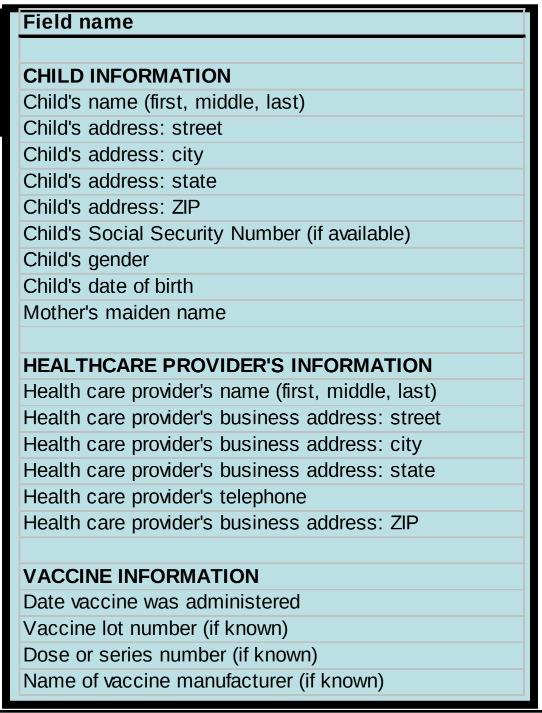
motivated by outbreak of measles in college students

Immunization registries

motivated by outbreak of measles in college students

seeded by electronic birth certificate database

state collections, national database maintained by CDC

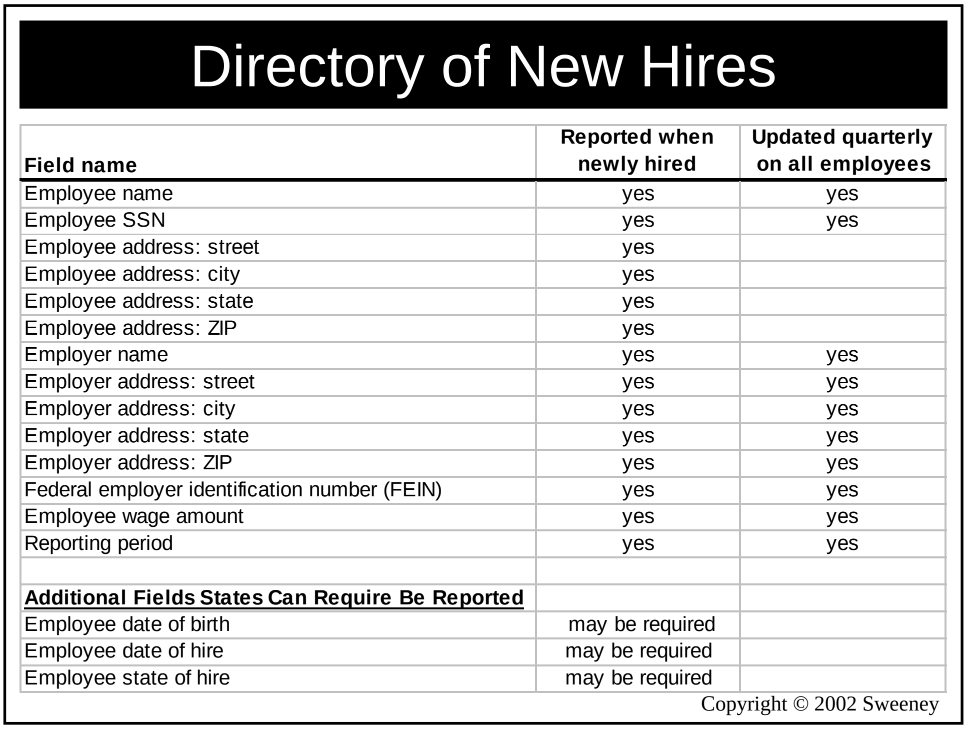


Field name
CHILD INFORMATION
Child's name (first, middle, last)
Child's address: street
Child's address: city
Child's address: state
Child's address: ZIP
Child's Social Security Number (if available)
Child's gender
Child's date of birth
Mother's maiden name
HEALTHCARE PROVIDER'S INFORMATION
Health care provider's name (first, middle, last)
Health care provider's business address: street
Health care provider's business address: city
Health care provider's business address: state
Health care provider's telephone
Health care provider's business address: ZIP
VACCINE INFORMATION
Date vaccine was administered
Vaccine lot number (if known)
Dose or series number (if known)
Name of vaccine manufacturer (if known)

Dead Beat Parents

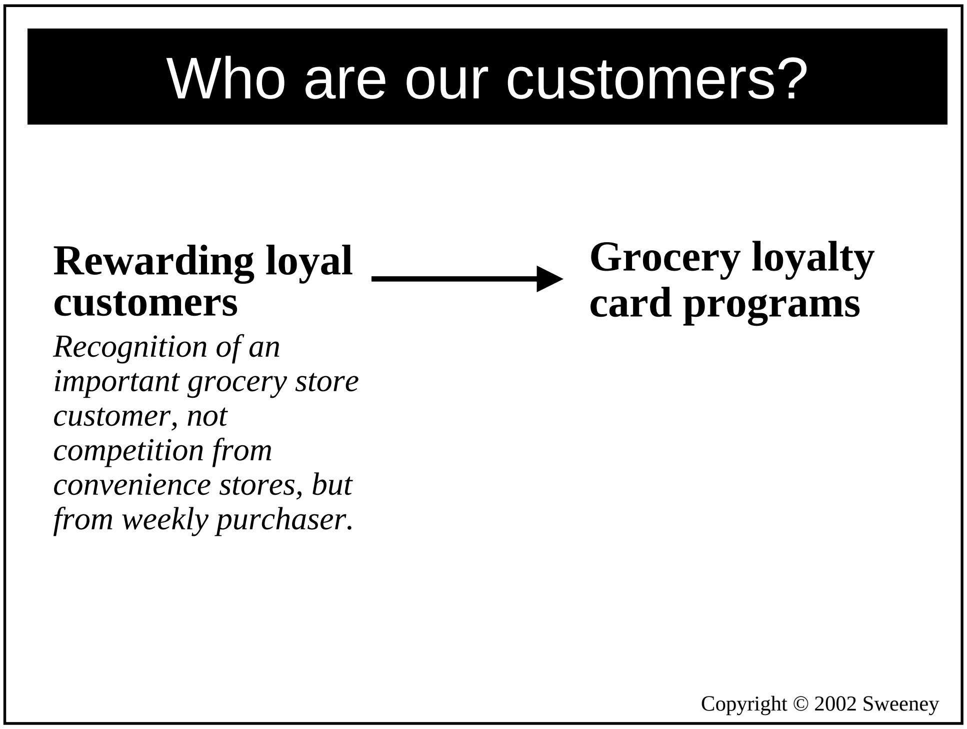
"Dead beat dads" problem → **National Directory of New Hires**

Parents not being financially responsible for their children, costs taxpayers



Directory of New Hires

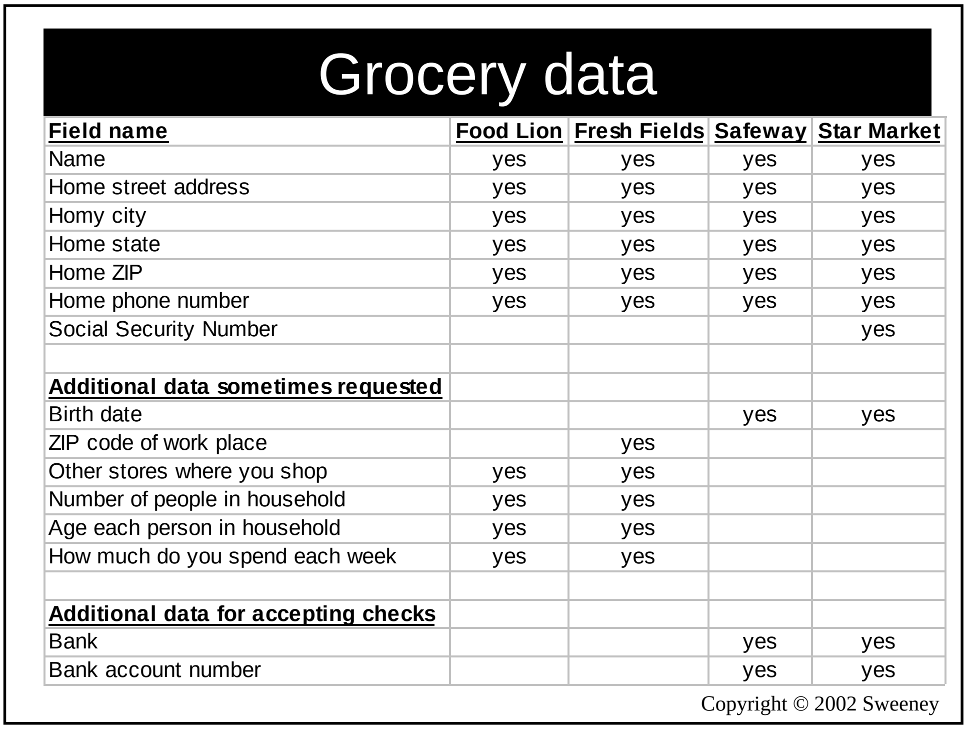
Field name	Reported when newly hired	Updated quarterly on all employees
Employee name	yes	yes
Employee SSN	yes	yes
Employee address: street	yes	
Employee address: city	yes	
Employee address: state	yes	
Employee address: ZIP	yes	
Employer name	yes	yes
Employer address: street	yes	yes
Employer address: city	yes	yes
Employer address: state	yes	yes
Employer address: ZIP	yes	yes
Federal employer identification number (FEIN)	yes	yes
Employee wage amount	yes	yes
Reporting period	yes	yes
Additional Fields States Can Require Be Reported		
Employee date of birth	may be required	
Employee date of hire	may be required	
Employee state of hire	may be required	



Who are our customers?

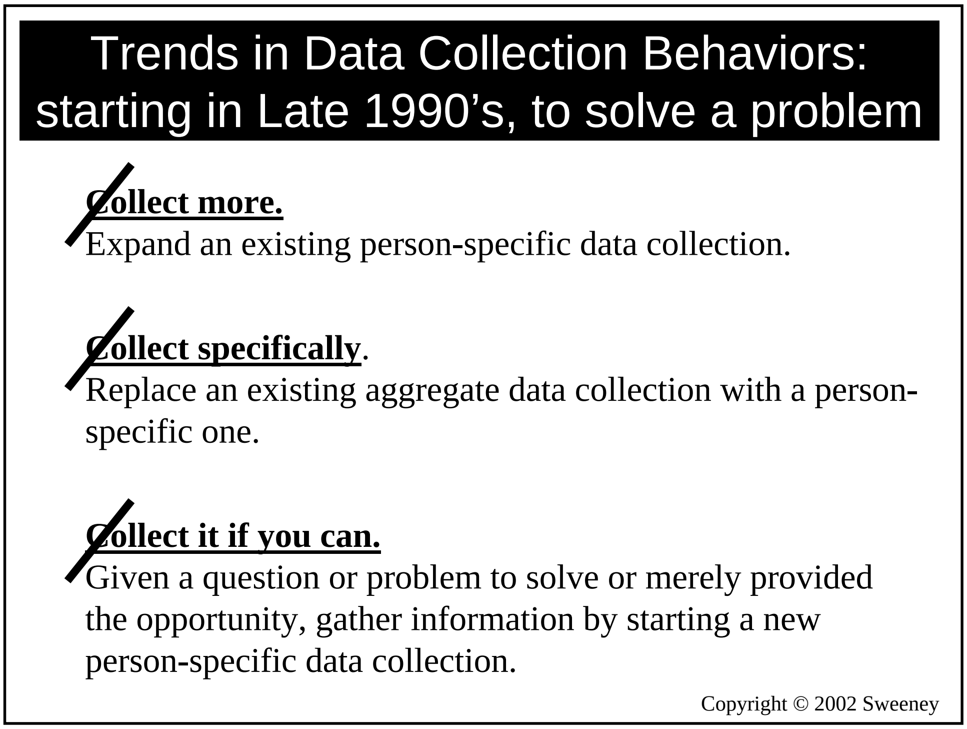
Rewarding loyal customers → **Grocery loyalty card programs**

Recognition of an important grocery store customer, not competition from convenience stores, but from weekly purchaser.



Grocery data

Field name	Food Lion	Fresh Fields	Safeway	Star Market
Name	yes	yes	yes	yes
Home street address	yes	yes	yes	yes
Homy city	yes	yes	yes	yes
Home state	yes	yes	yes	yes
Home ZIP	yes	yes	yes	yes
Home phone number	yes	yes	yes	yes
Social Security Number				yes
Additional data sometimes requested				
Birth date			yes	yes
ZIP code of work place		yes		
Other stores where you shop	yes	yes		
Number of people in household	yes	yes		
Age each person in household	yes	yes		
How much do you spend each week	yes	yes		
Additional data for accepting checks				
Bank			yes	yes
Bank account number			yes	yes

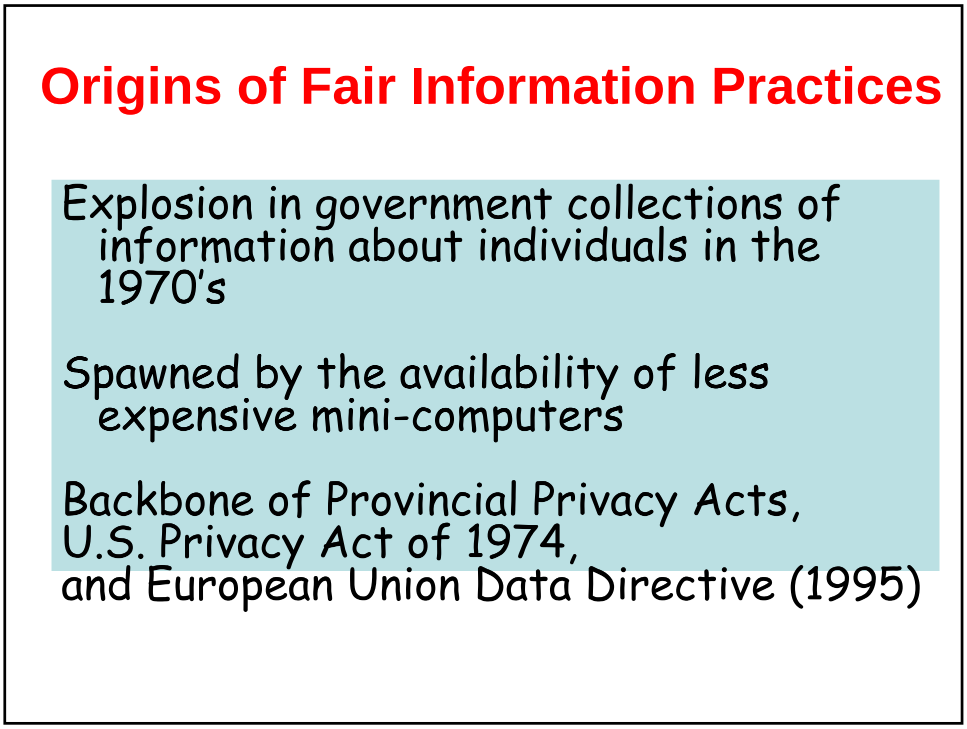


Trends in Data Collection Behaviors: starting in Late 1990's, to solve a problem

Collect more.
Expand an existing person-specific data collection.

Collect specifically.
Replace an existing aggregate data collection with a person-specific one.

Collect it if you can.
Given a question or problem to solve or merely provided the opportunity, gather information by starting a new person-specific data collection.

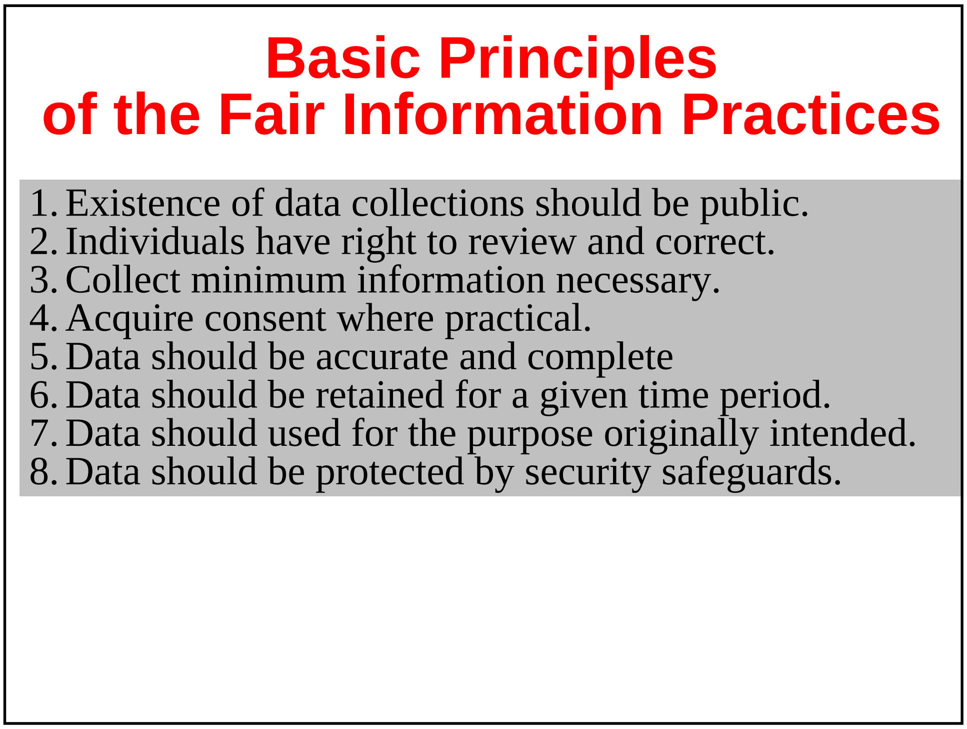


Origins of Fair Information Practices

Explosion in government collections of information about individuals in the 1970's

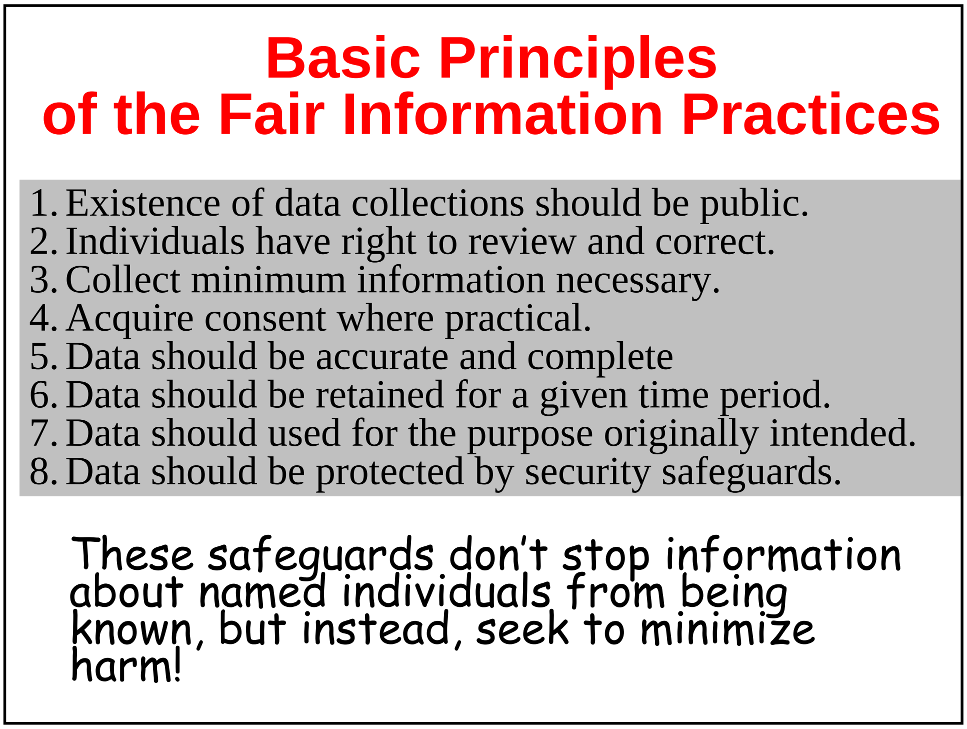
Spawned by the availability of less expensive mini-computers

Backbone of Provincial Privacy Acts,
U.S. Privacy Act of 1974,
and European Union Data Directive (1995)



Basic Principles of the Fair Information Practices

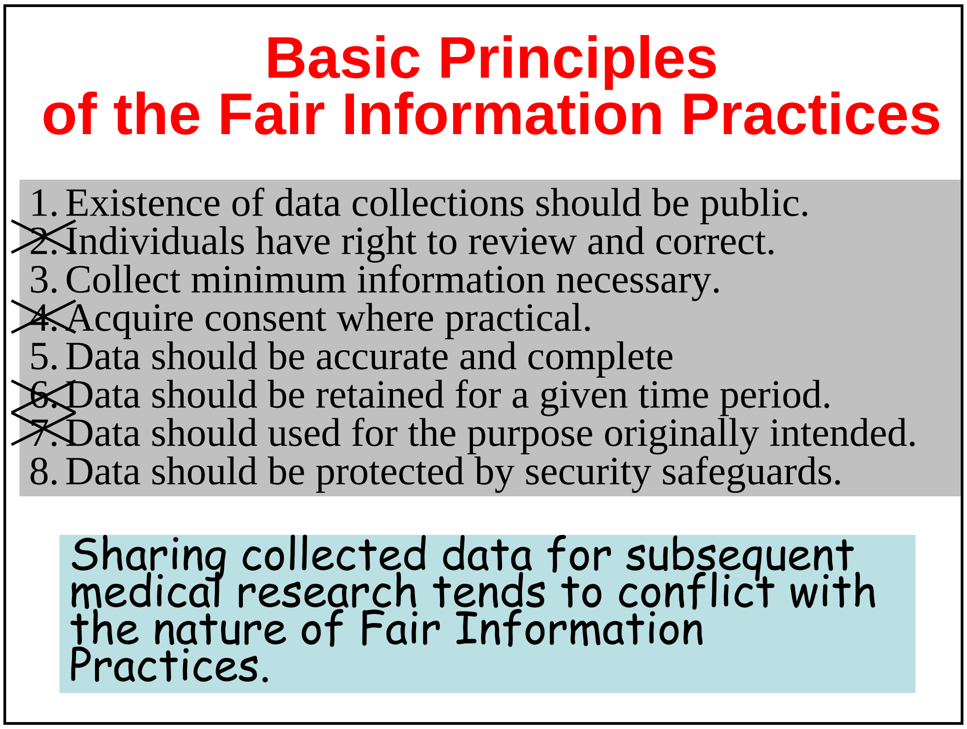
1. Existence of data collections should be public.
2. Individuals have right to review and correct.
3. Collect minimum information necessary.
4. Acquire consent where practical.
5. Data should be accurate and complete
6. Data should be retained for a given time period.
7. Data should used for the purpose originally intended.
8. Data should be protected by security safeguards.



Basic Principles of the Fair Information Practices

1. Existence of data collections should be public.
2. Individuals have right to review and correct.
3. Collect minimum information necessary.
4. Acquire consent where practical.
5. Data should be accurate and complete
6. Data should be retained for a given time period.
7. Data should used for the purpose originally intended.
8. Data should be protected by security safeguards.

These safeguards don't stop information about named individuals from being known, but instead, seek to minimize harm!



Basic Principles of the Fair Information Practices

1. Existence of data collections should be public.

~~2.~~ Individuals have right to review and correct.

3. Collect minimum information necessary.

~~4.~~ Acquire consent where practical.

5. Data should be accurate and complete

~~6.~~ Data should be retained for a given time period.

~~7.~~ Data should used for the purpose originally intended.

8. Data should be protected by security safeguards.

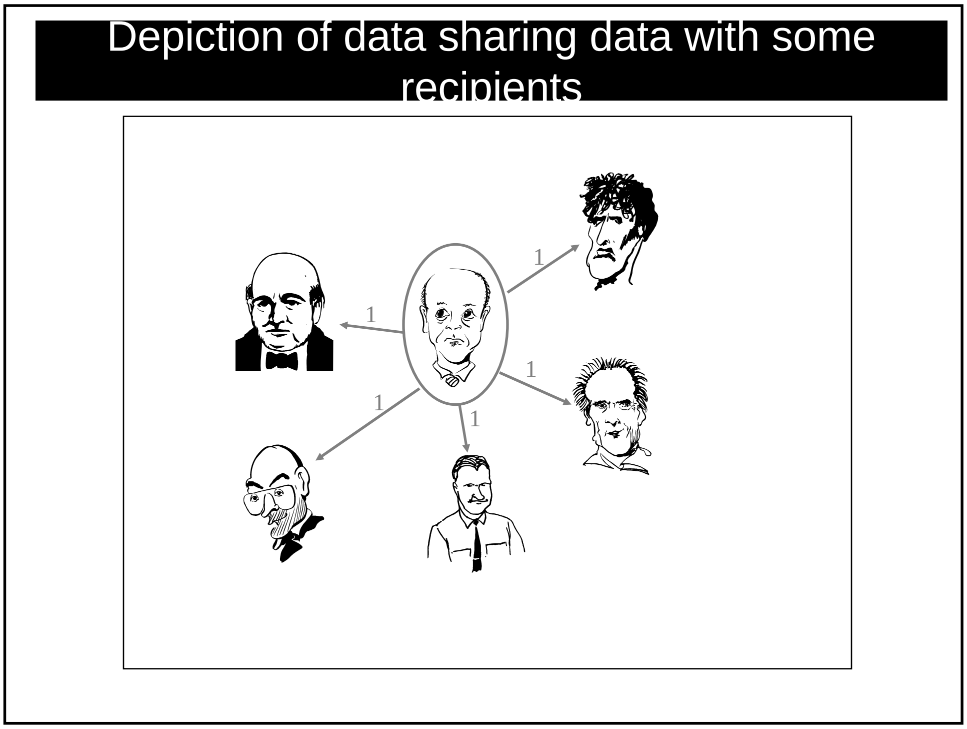
Sharing collected data for subsequent medical research tends to conflict with the nature of Fair Information Practices.

Not Fair Information Practices, But Data Anonymity

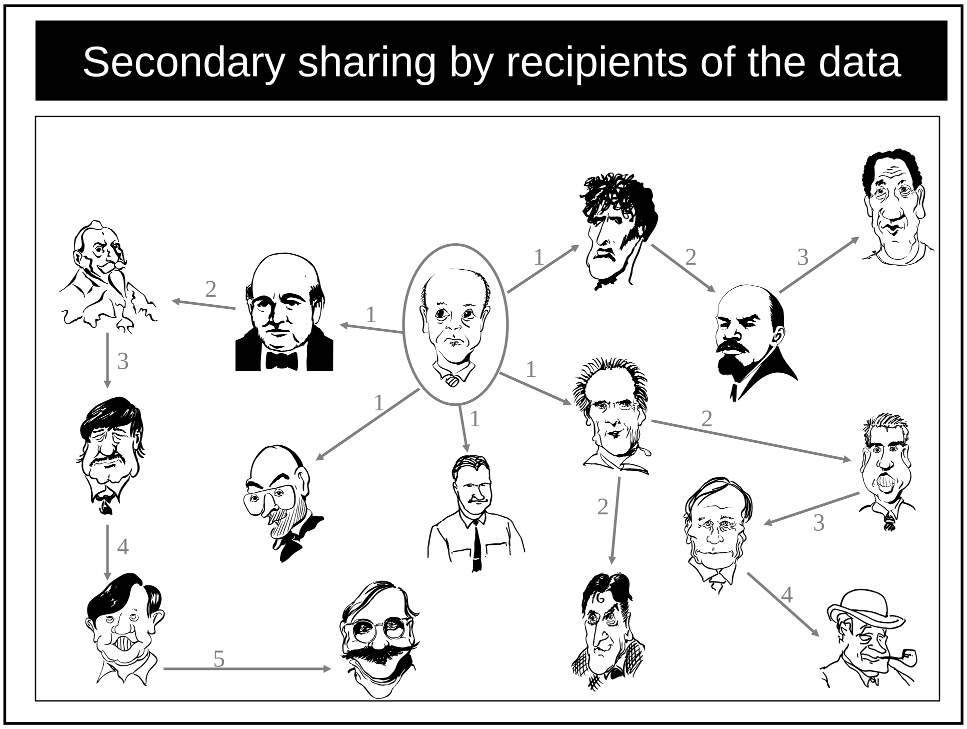
Provide a version of the data so that no one who is the subject of the data can be re-identified.

… can move beyond merely minimizing harm (as with Fair Information Practices), to actually providing privacy protection. Information can be known about a person without knowing who the person may be.

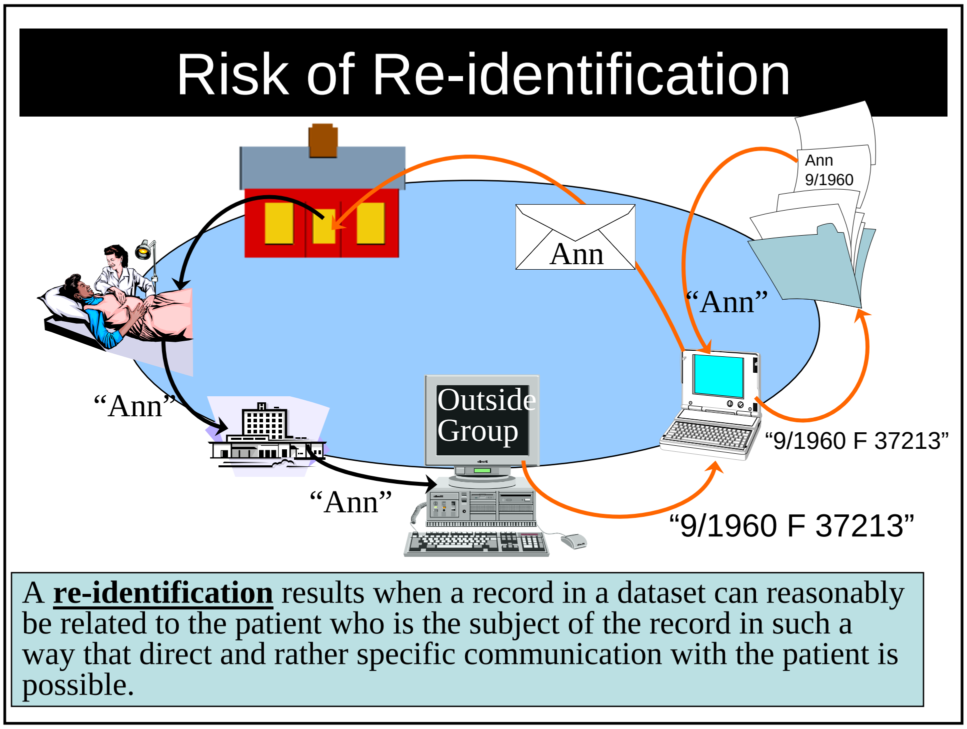
Depiction of no data sharing by the data collector



Depiction of data sharing data with some recipients

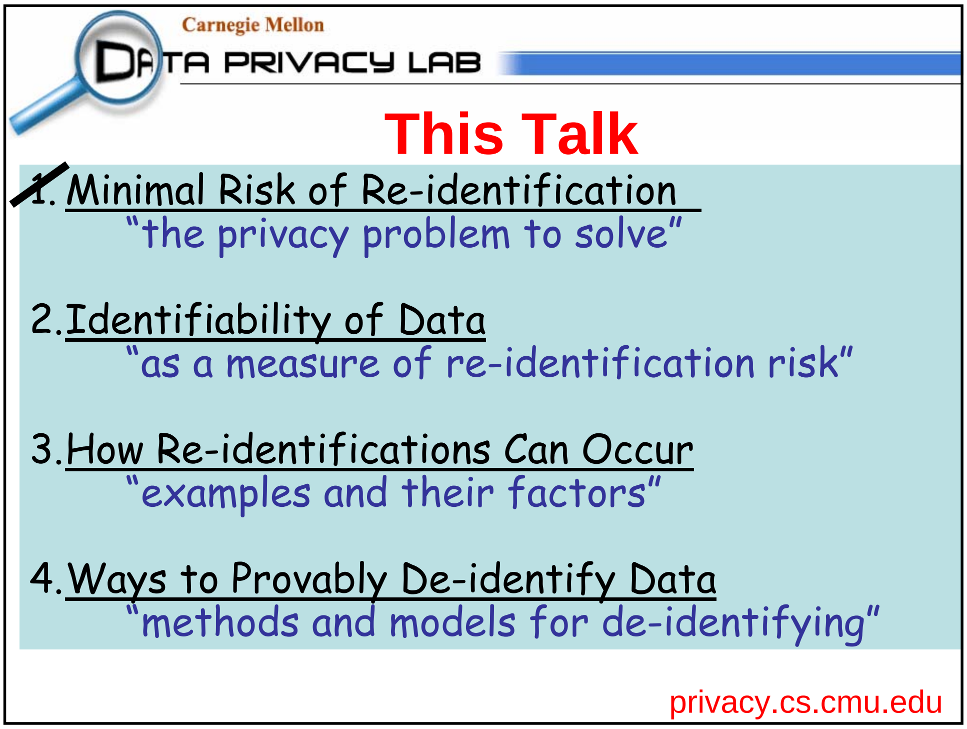


Secondary sharing by recipients of the data



Risk of Re-identification

A **re-identification** results when a record in a dataset can reasonably be related to the patient who is the subject of the record in such a way that direct and rather specific communication with the patient is possible.

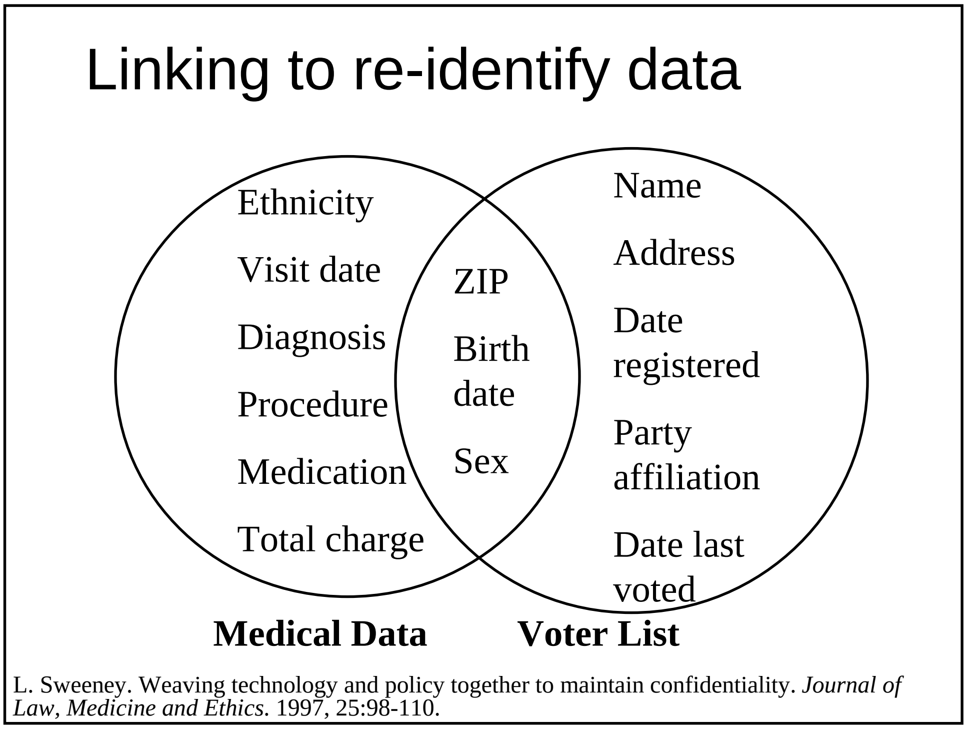


Carnegie Mellon
DATA PRIVACY LAB

This Talk

1. Minimal Risk of Re-identification
   "the privacy problem to solve"
2. Identifiability of Data
   "as a measure of re-identification risk"
3. How Re-identifications Can Occur
   "examples and their factors"
4. Ways to Provably De-identify Data
   "methods and models for de-identifying"

privacy.cs.cmu.edu



Linking to re-identify data

Ethnicity

Visit date

Diagnosis

Procedure

Medication

Total charge

ZIP

Birth date

Sex

Name

Address

Date registered

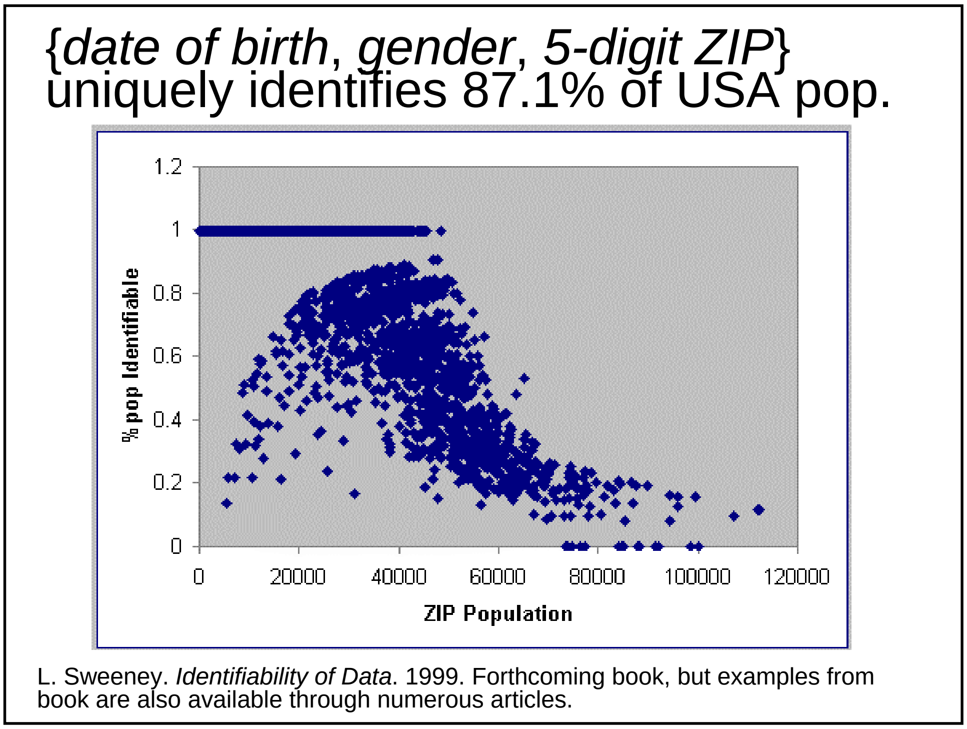
Party affiliation

Date last voted

Medical Data

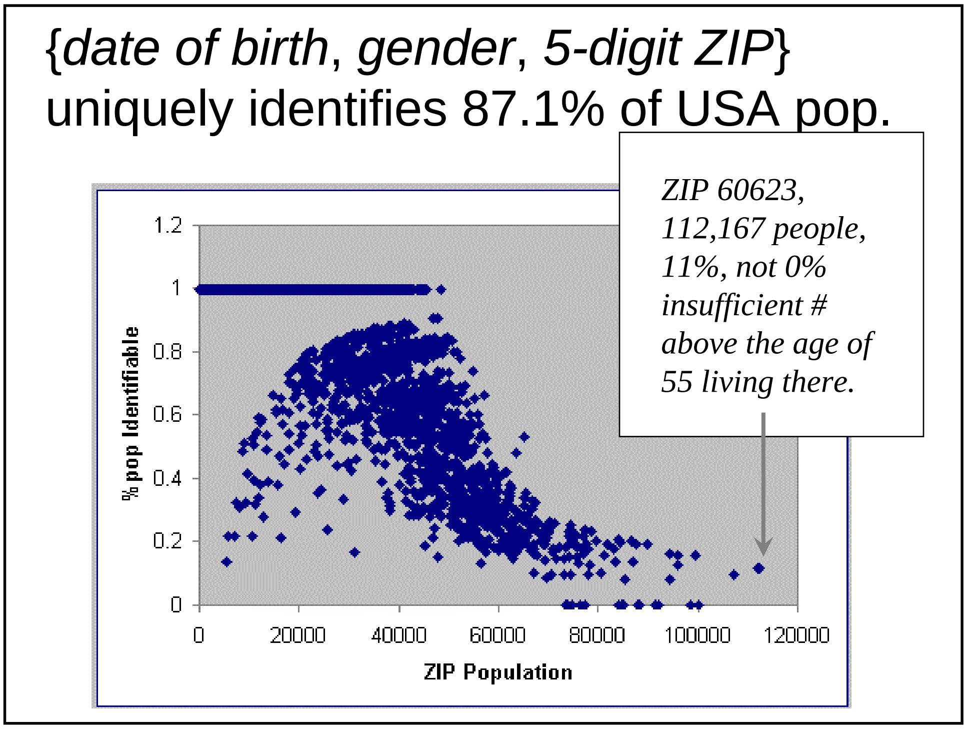
Voter List

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



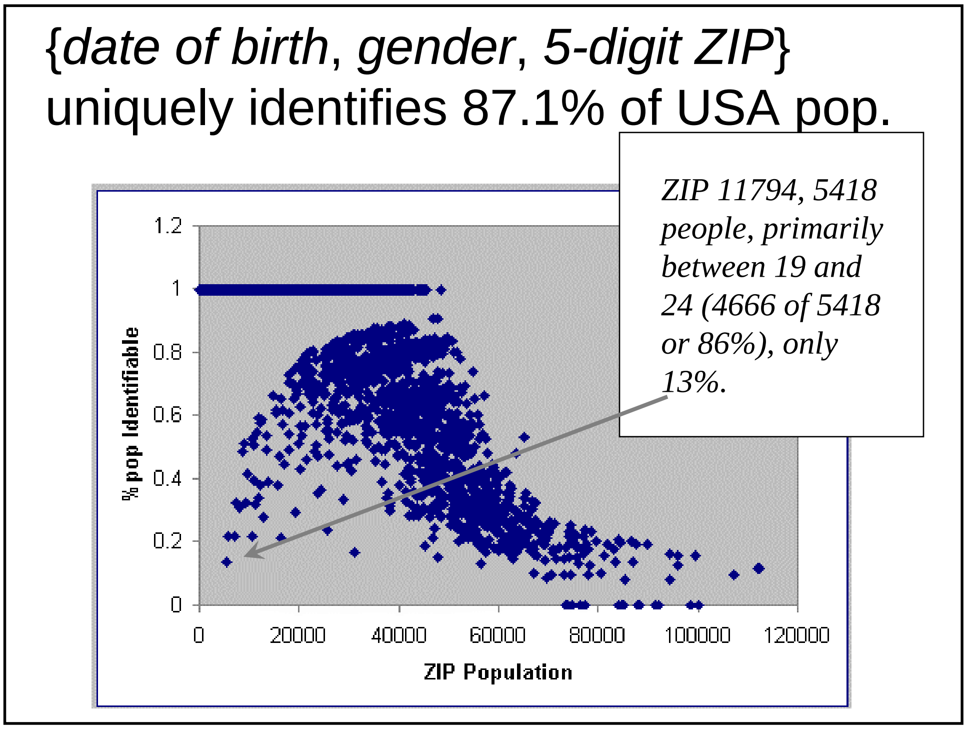
{*date of birth*, *gender*, *5-digit ZIP*} uniquely identifies 87.1% of USA pop.

L. Sweeney. *Identifiability of Data*. 1999. Forthcoming book, but examples from book are also available through numerous articles.



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

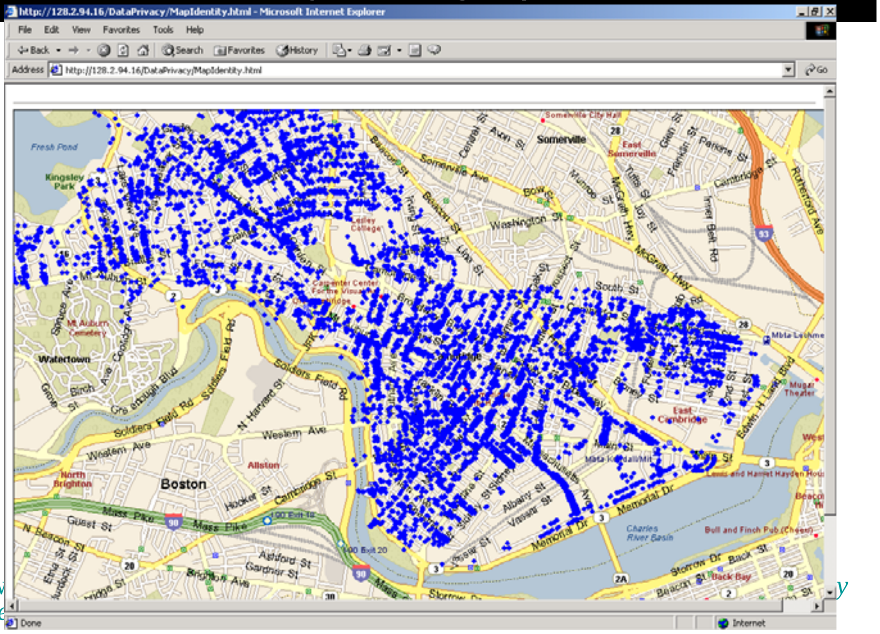
ZIP 60623, 112,167 people, 11%, not 0% insufficient # above the age of 55 living there.



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

ZIP 11794, 5418 people, primarily between 19 and 24 (4666 of 5418 or 86%), only 13%.

Identifiability, Geographical Plot

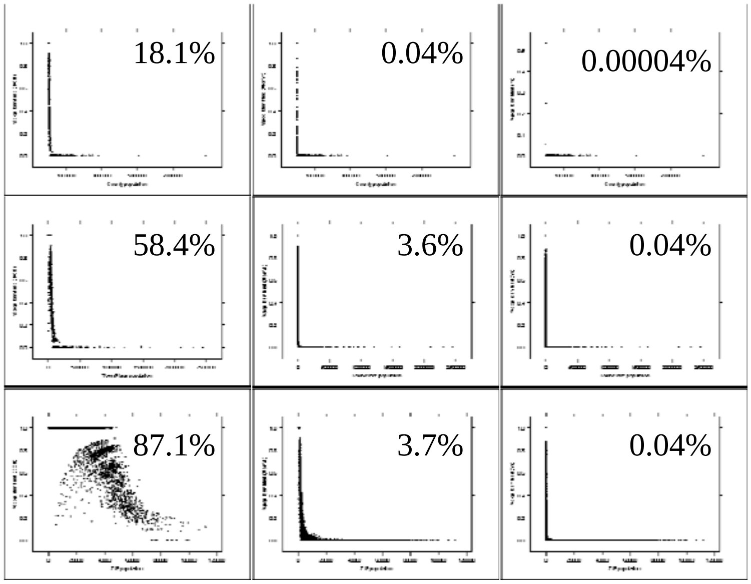

http://128.2.94.16/DataPrivacy/MapIdentity.html - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://128.2.94.16/DataPrivacy/MapIdentity.html
Go
Somerville
Watertown
Boston
Allston
East Cambridge
Charles River Basin
Done
Internet

;;-Sw

Prote

y

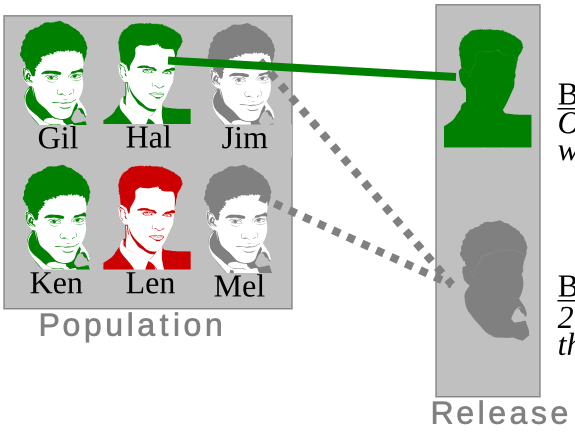
Re-identification Trade-offs



	Date of Birth	Mon/Yr Birth	Year of Birth
County	18.1%	0.04%	0.00004%
Town/ Place	58.4%	3.6%	0.04%
ZIP 5-digit	87.1%	3.7%	0.04%

Gender

Measuring Identifiability

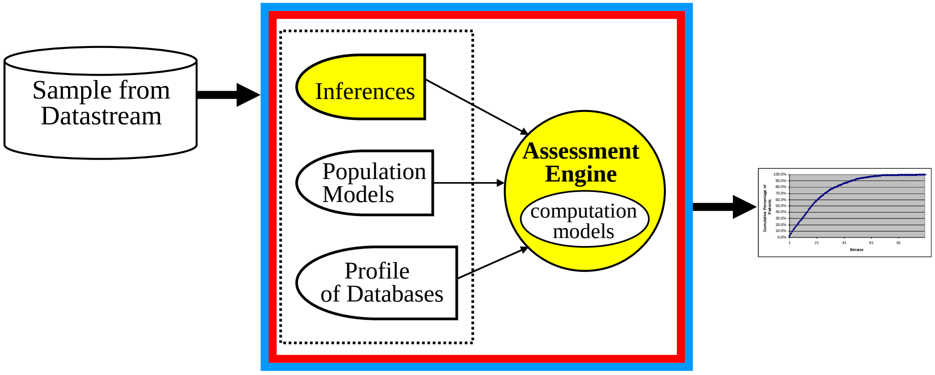


Binsize of 1
Only 1 person is green with that shape head.

Binsize of 2
2 people are gray with that shape head.

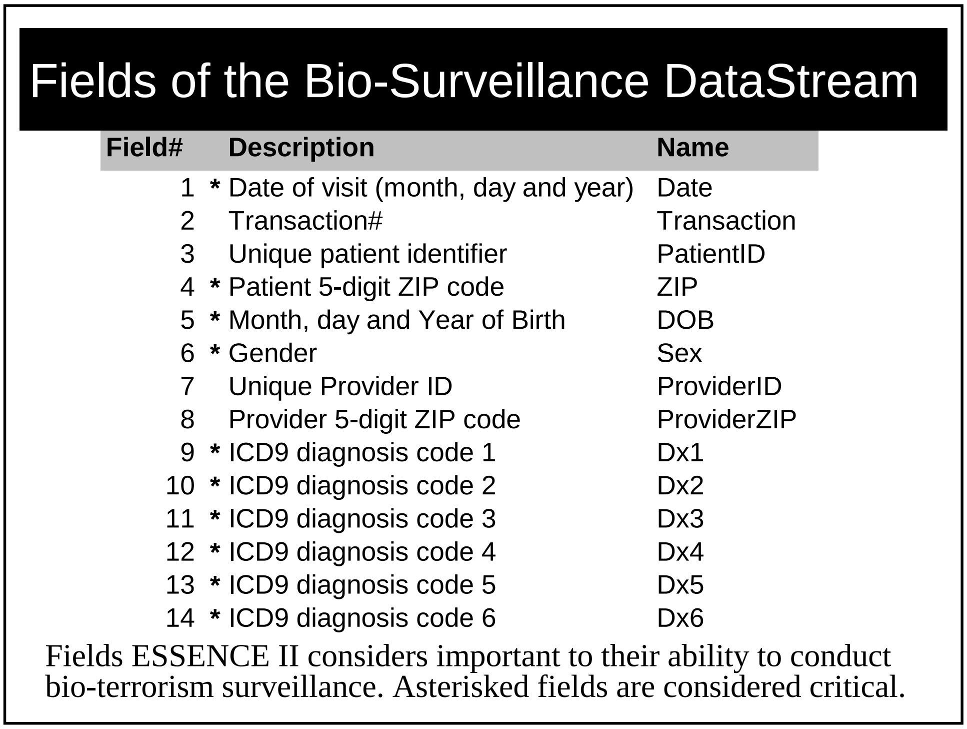
Identifiability estimates, in graduated sized groupings, the number of people to which a released record is apt to refer. These groupings are called **binsizes**.

Risk Assessment Server



The **Risk Assessment Server** identifies which fields and/or records in a datastream are vulnerable to known re-identification inference strategies. The output of the assessment server is a report on the identifiability of the datastream (not just a sample) with respect to those inference strategies.

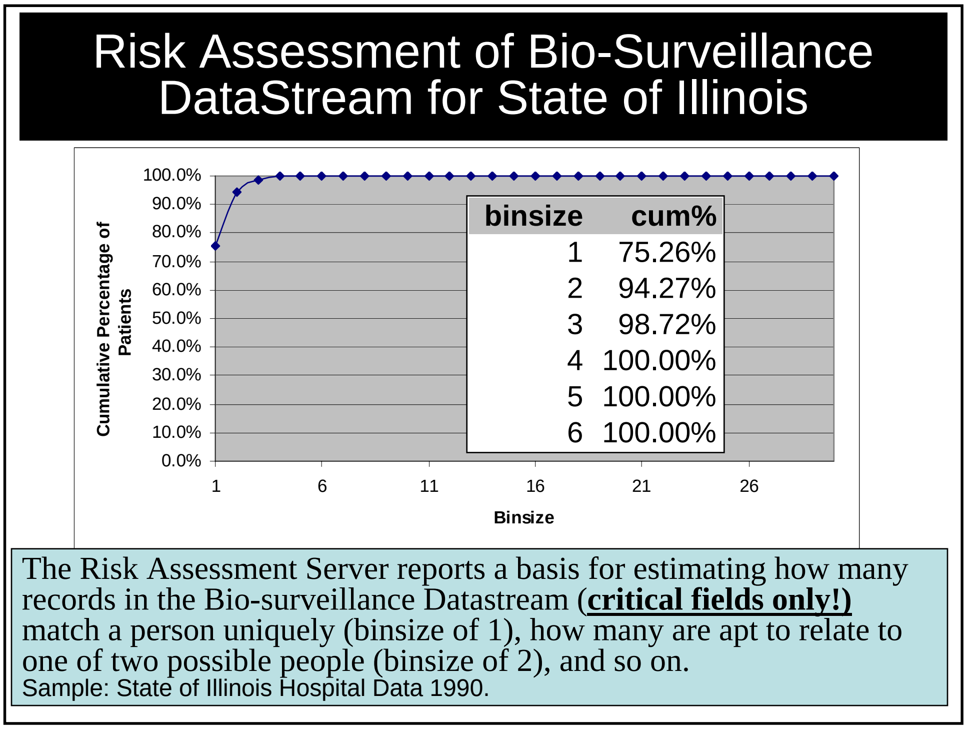
The Risk Assessment Server is licensed to Privacert, Inc. Diagram is courtesy of Privacert. All rights reserved. http://www.privacert.com



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

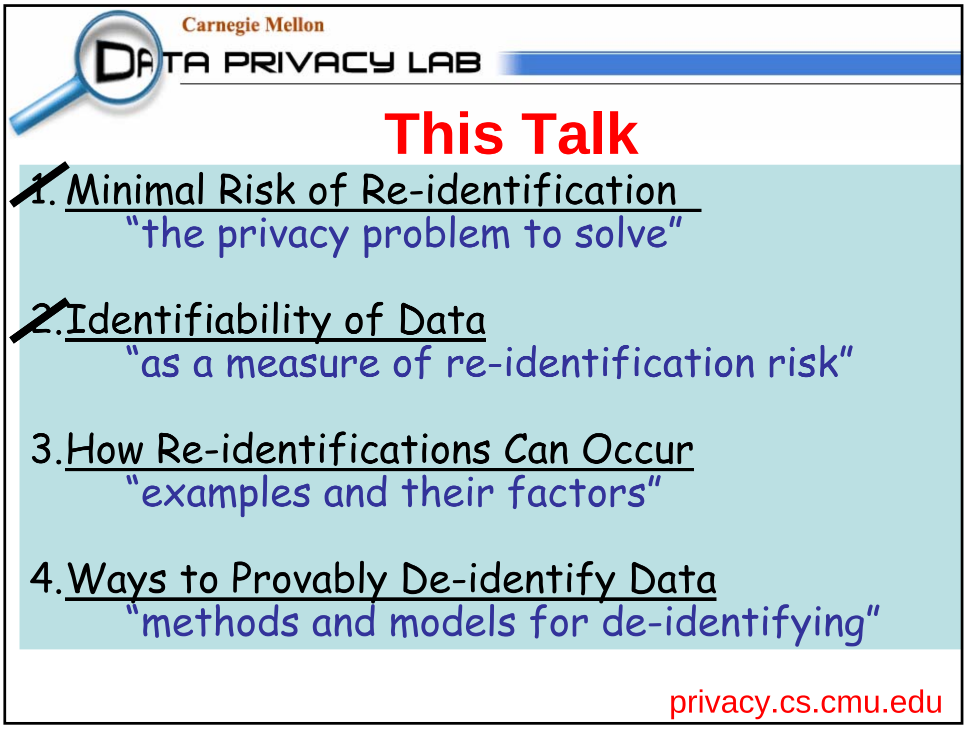
Fields ESSENCE II considers important to their ability to conduct bio-terrorism surveillance. Asterisked fields are considered critical.



Risk Assessment of Bio-Surveillance DataStream for State of Illinois

binsize	cum%
1	75.26%
2	94.27%
3	98.72%
4	100.00%
5	100.00%
6	100.00%

The Risk Assessment Server reports a basis for estimating how many records in the Bio-surveillance Datastream (**critical fields only!)** match a person uniquely (binsize of 1), how many are apt to relate to one of two possible people (binsize of 2), and so on.
Sample: State of Illinois Hospital Data 1990.

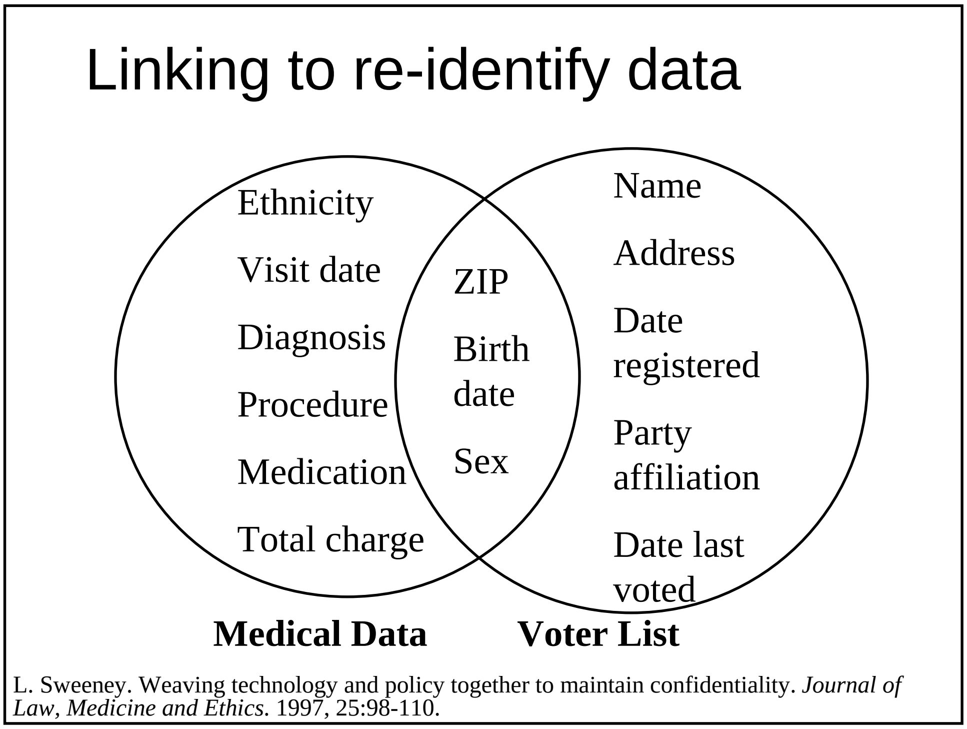


Carnegie Mellon DATA PRIVACY LAB

This Talk

1. Minimal Risk of Re-identification
   "the privacy problem to solve"
2. Identifiability of Data
   "as a measure of re-identification risk"
3. How Re-identifications Can Occur
   "examples and their factors"
4. Ways to Provably De-identify Data
   "methods and models for de-identifying"

privacy.cs.cmu.edu



Linking to re-identify data

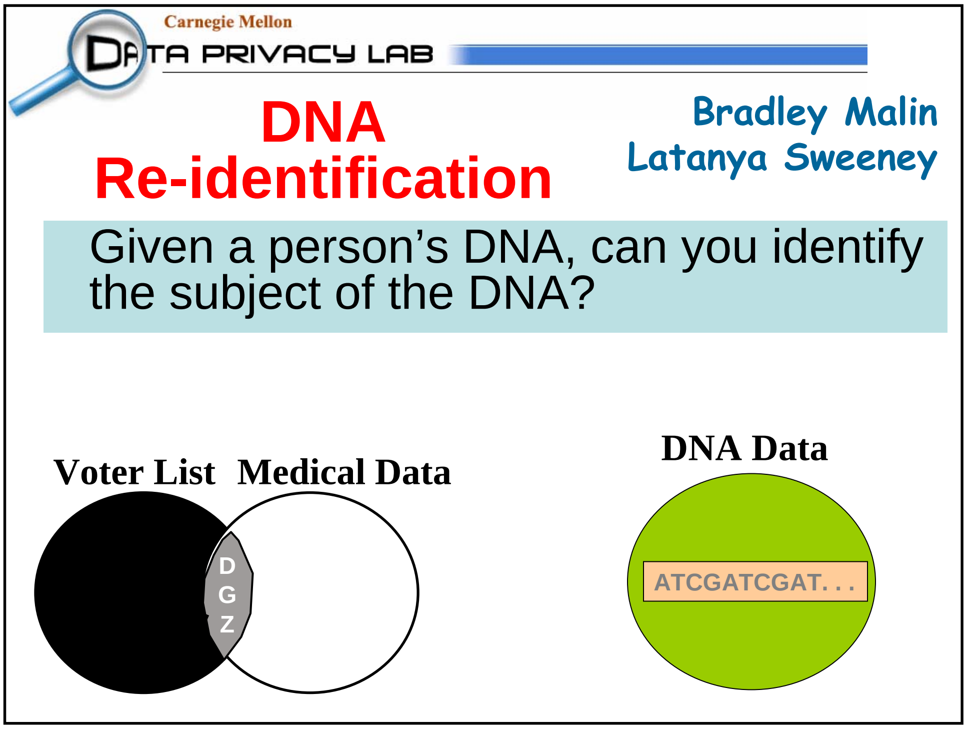
Ethnicity
Visit date
Diagnosis
Procedure
Medication
Total charge

ZIP
Birth date
Sex

Name
Address
Date registered
Party affiliation
Date last voted

Medical Data **Voter List**

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

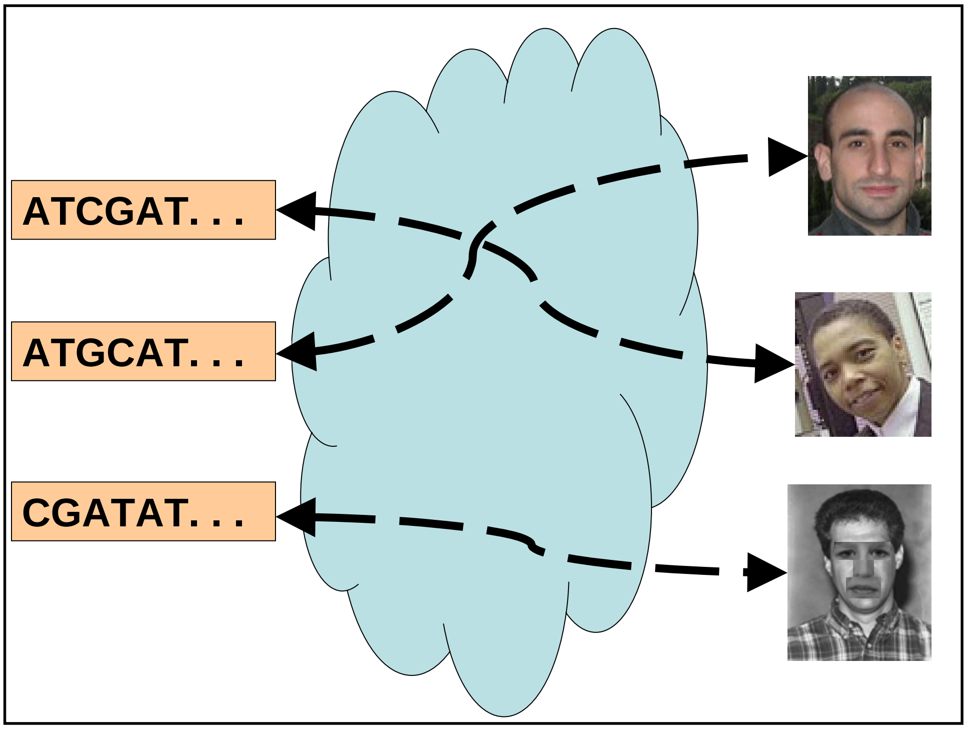


Carnegie Mellon DATA PRIVACY LAB

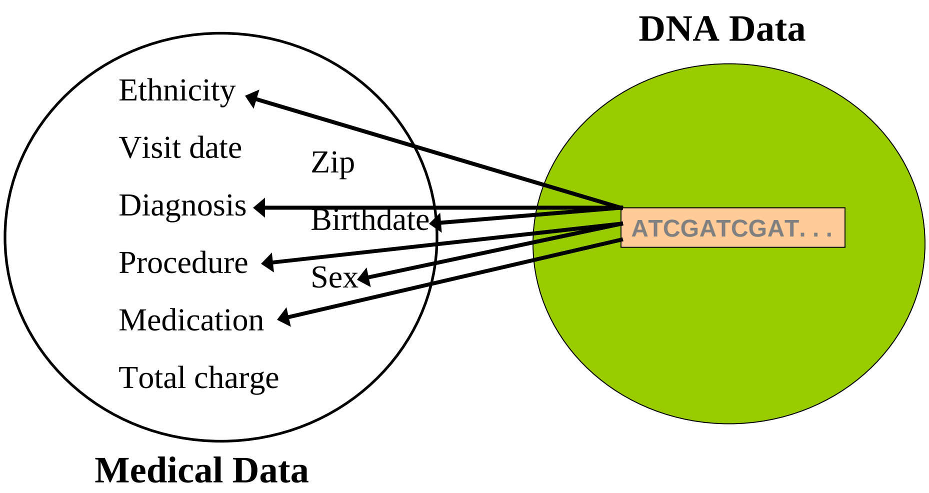
DNA Re-identification

Bradley Malin
Latanya Sweeney

Given a person’s DNA, can you identify the subject of the DNA?

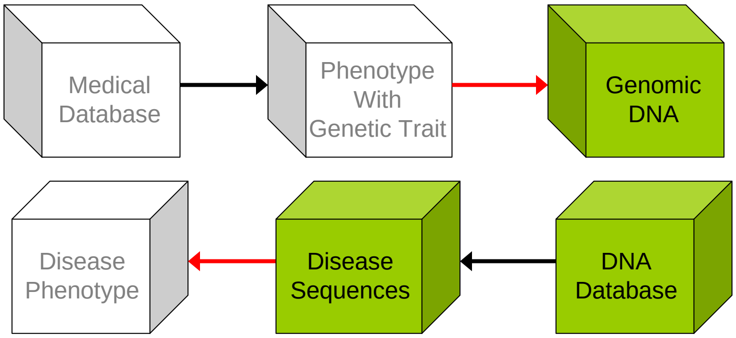


Build Up Relations

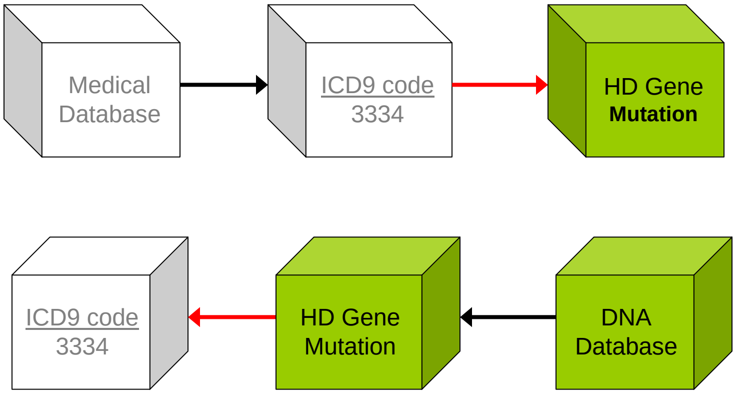


Genotype-Phenotype Relations

- Can infer genotype-phenotype relationships out of both DNA and medical databases



Example: Huntington's Disease



Medical Database

ICD9 code 3334

HD Gene **Mutation**

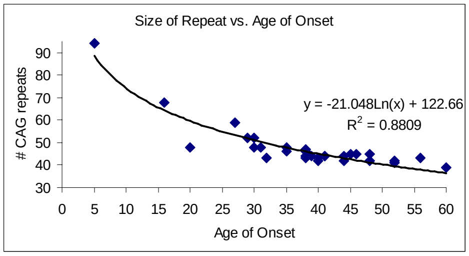
ICD9 code 3334

HD Gene Mutation

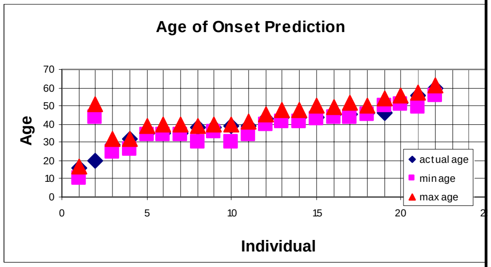
DNA Database

Experimental Results- DNA with Huntington's Disease

- Example: Huntington's disease
  - Exists strong correlation between age of onset and DNA mutation (# of CAG repeats)
  - Given longitudinal clinical info, accurately infer age of onset in 20 of 22 cases

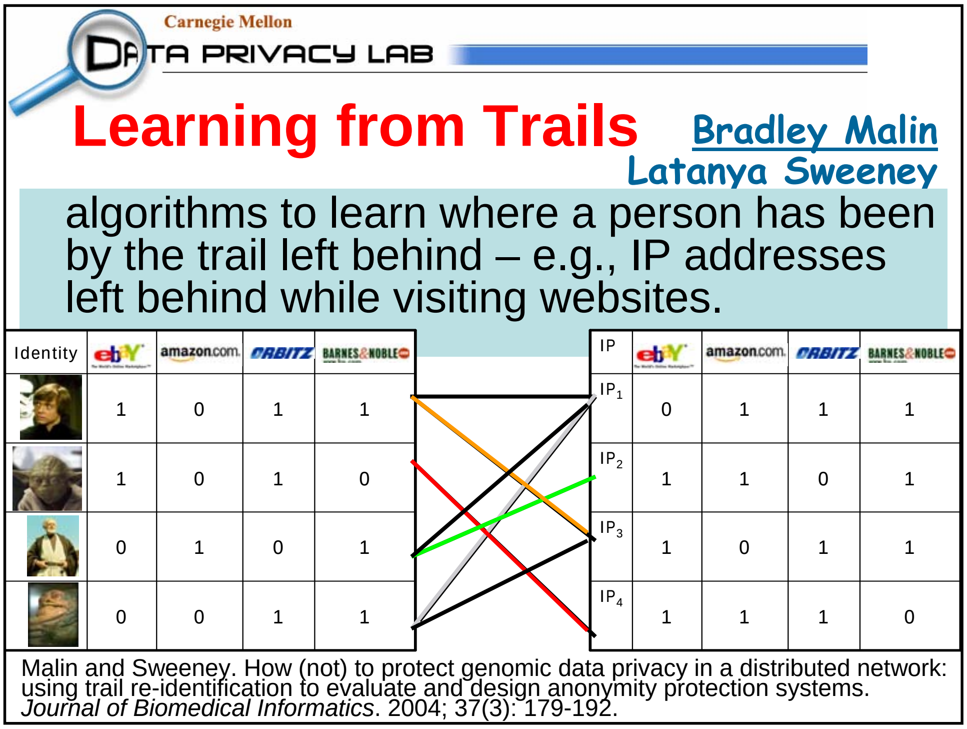


Size of Repeat vs. Age of Onset



Age of Onset Prediction

Malin B and Sweeney L. Inferring genotype from clinical phenotype through a knowledge-based algorithm. In *Pacific Symposium on Biocomputing*. pp. 41-52, Jan 2002.



Carnegie Mellon DATA PRIVACY LAB

Learning from Trails

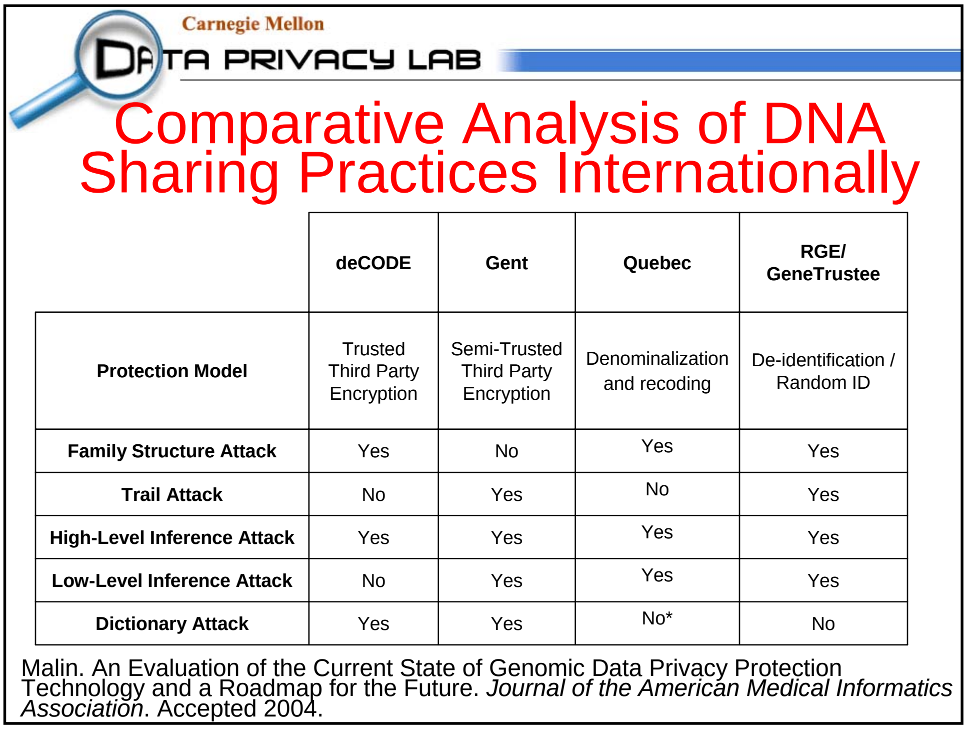
Bradley Malin
Latanya Sweeney

algorithms to learn where a person has been by the trail left behind – e.g., IP addresses left behind while visiting websites.

Identity	ebay	amazon.com	ORBITZ	BARNES&NOBLE
	1	0	1	1
	1	0	1	0
	0	1	0	1
	0	0	1	1

IP	ebay	amazon.com	ORBITZ	BARNES&NOBLE
IP_1	0	1	1	1
IP_2	1	1	0	1
IP_3	1	0	1	1
IP_4	1	1	1	0

Malin and Sweeney. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics*. 2004; 37(3): 179-192.

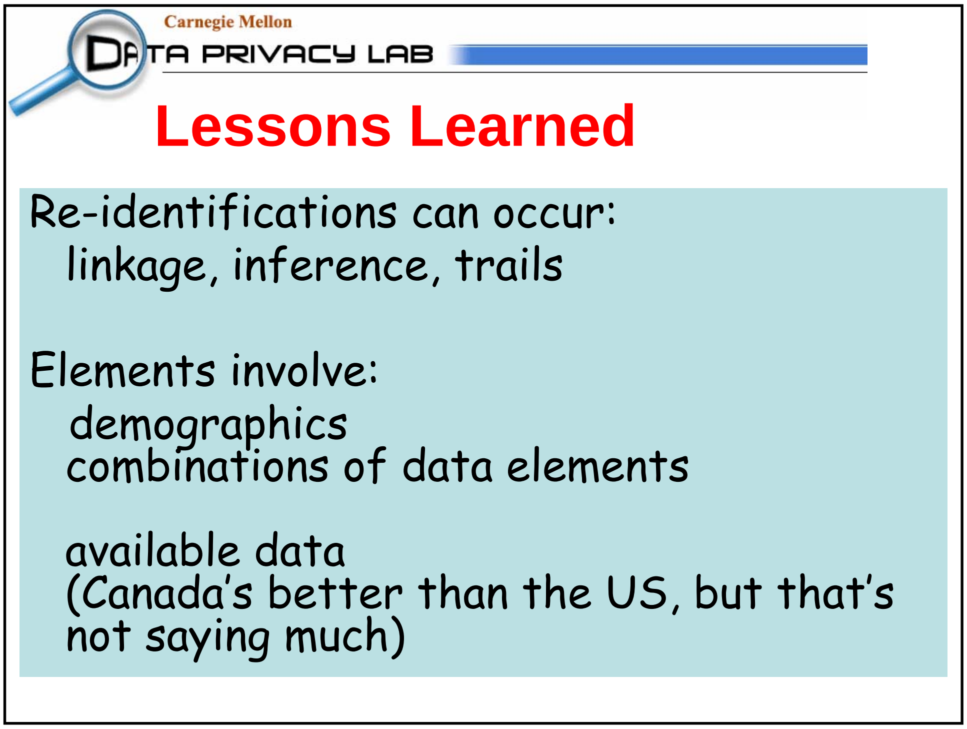


Carnegie Mellon DATA PRIVACY LAB

Comparative Analysis of DNA Sharing Practices Internationally

	deCODE	**Gent**	**Quebec**	**RGE/ GeneTrustee**
Protection Model	Trusted Third Party Encryption	Semi-Trusted Third Party Encryption	Denominalization and recoding	De-identification / Random ID
Family Structure Attack	Yes	No	Yes	Yes
Trail Attack	No	Yes	No	Yes
High-Level Inference Attack	Yes	Yes	Yes	Yes
Low-Level Inference Attack	No	Yes	Yes	Yes
Dictionary Attack	Yes	Yes	No*	No

Malin. An Evaluation of the Current State of Genomic Data Privacy Protection Technology and a Roadmap for the Future. *Journal of the American Medical Informatics Association*. Accepted 2004.



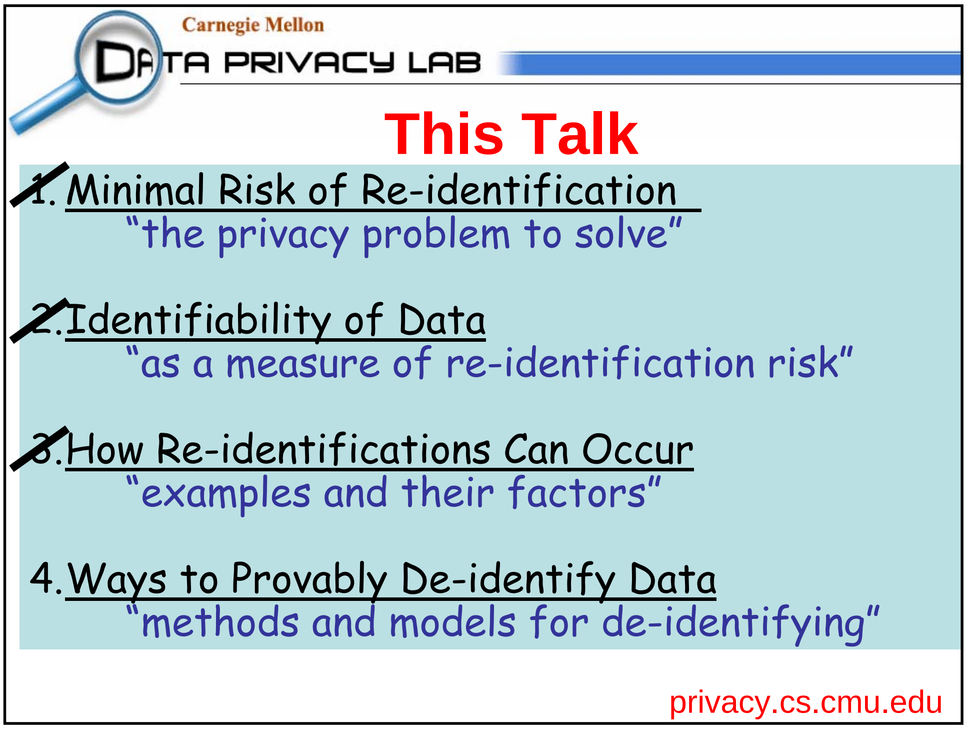
Carnegie Mellon
DATA PRIVACY LAB

Lessons Learned

Re-identifications can occur:
linkage, inference, trails

Elements involve:
demographics
combinations of data elements

available data
(Canada's better than the US, but that's not saying much)



Carnegie Mellon
DATA PRIVACY LAB

This Talk

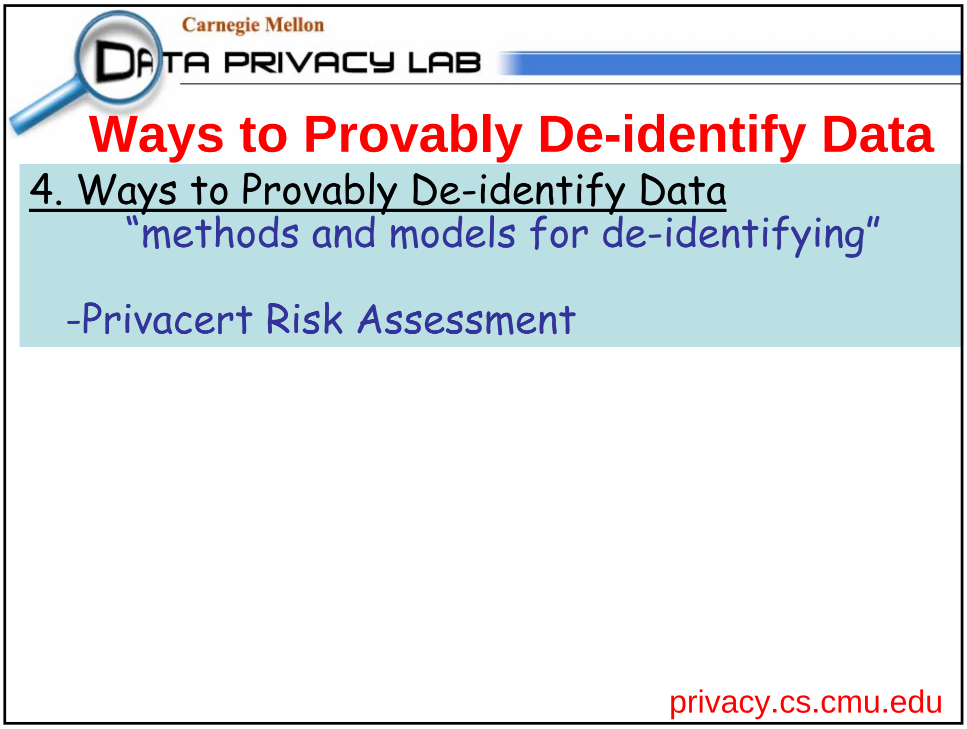
~~1~~. Minimal Risk of Re-identification
"the privacy problem to solve"

~~2~~.Identifiability of Data
"as a measure of re-identification risk"

~~3~~.How Re-identifications Can Occur
"examples and their factors"

4.Ways to Provably De-identify Data
"methods and models for de-identifying"

privacy.cs.cmu.edu



Carnegie Mellon DATA PRIVACY LAB

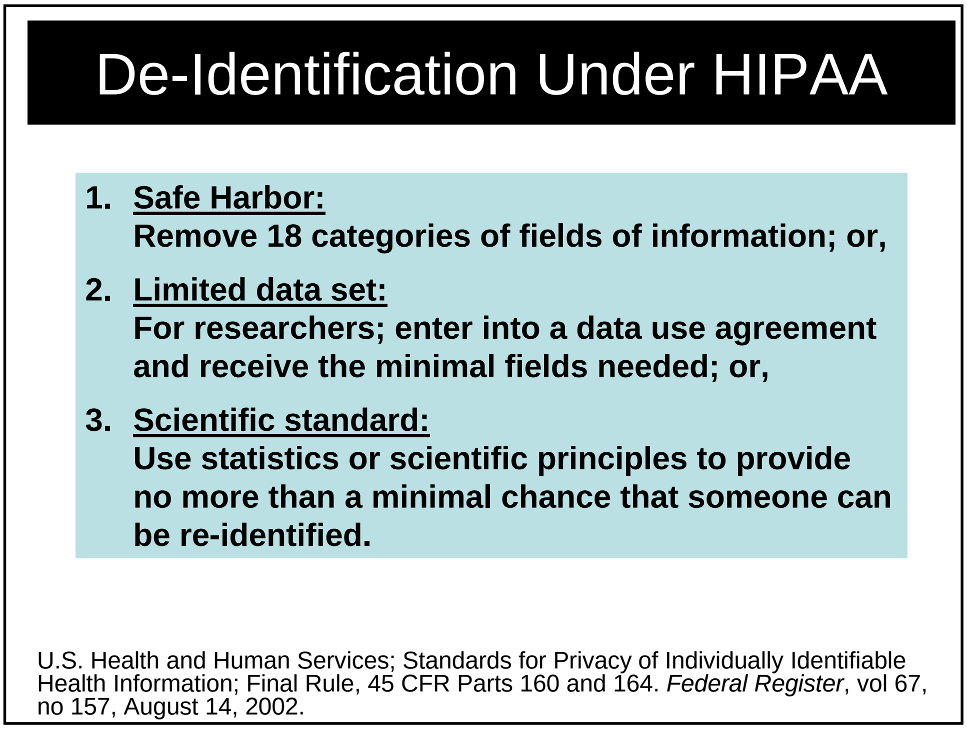
Ways to Provably De-identify Data

4. Ways to Provably De-identify Data

“methods and models for de-identifying”

-Privacert Risk Assessment

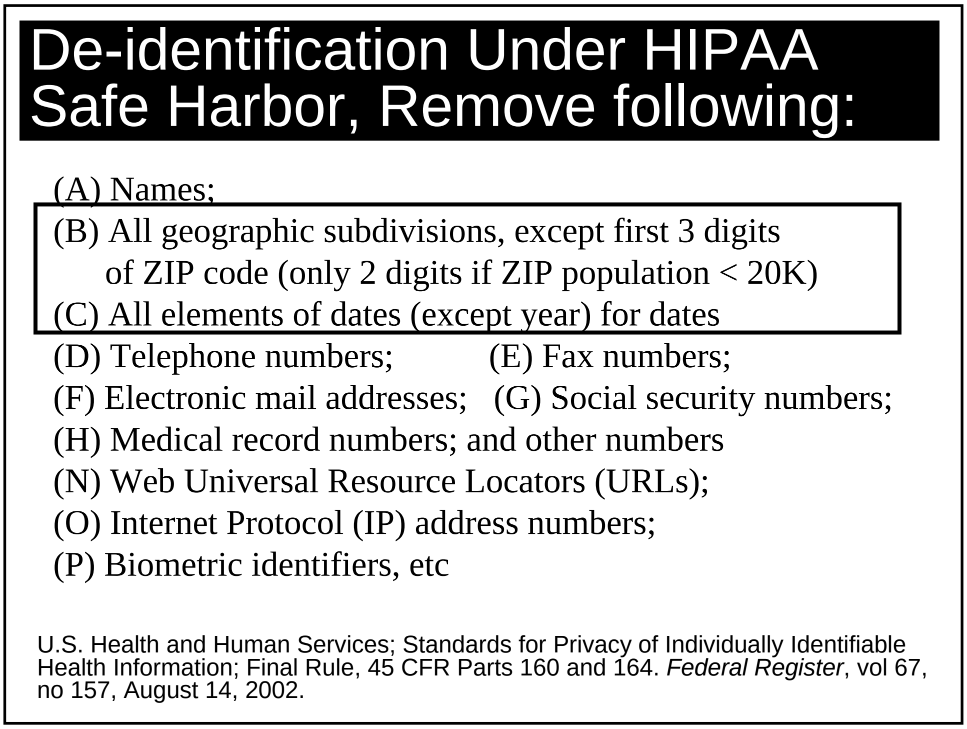
privacy.cs.cmu.edu



De-Identification Under HIPAA

1. **Safe Harbor:**
   Remove 18 categories of fields of information; or,
2. **Limited data set:**
   For researchers; enter into a data use agreement and receive the minimal fields needed; or,
3. **Scientific standard:**
   Use statistics or scientific principles to provide no more than a minimal chance that someone can be re-identified.

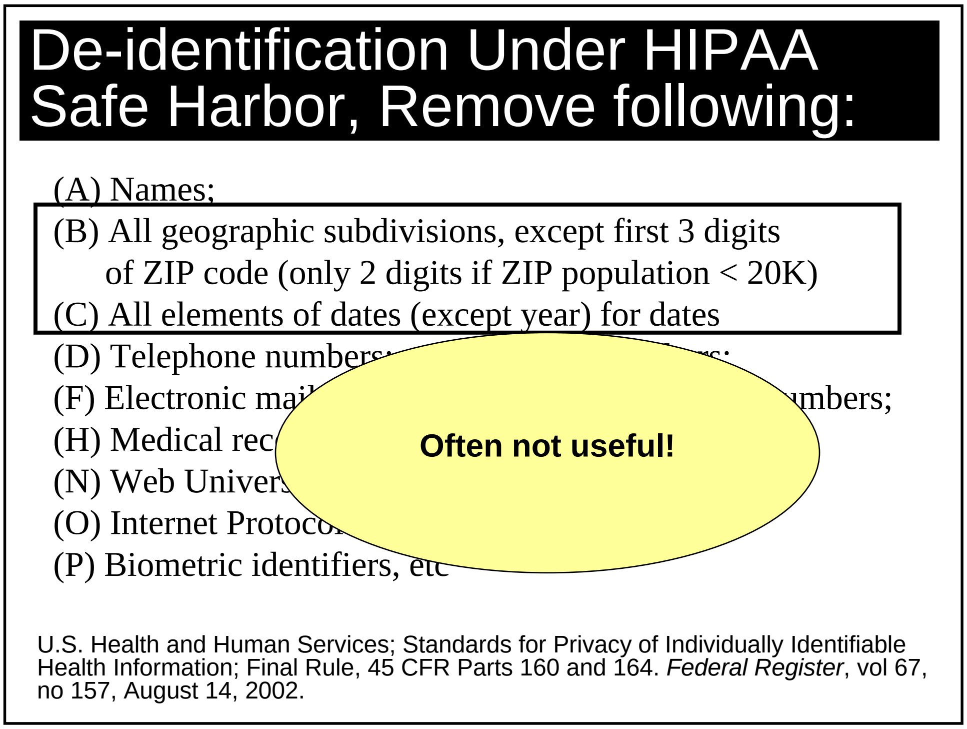
U.S. Health and Human Services; Standards for Privacy of Individually Identifiable Health Information; Final Rule, 45 CFR Parts 160 and 164. *Federal Register*, vol 67, no 157, August 14, 2002.



De-identification Under HIPAA 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 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; rs;
(F) Electronic mai umbers;
(H) Medical rec
(N) Web Univers
(O) Internet Protoco
(P) Biometric identifiers, etc

Often not useful!

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.

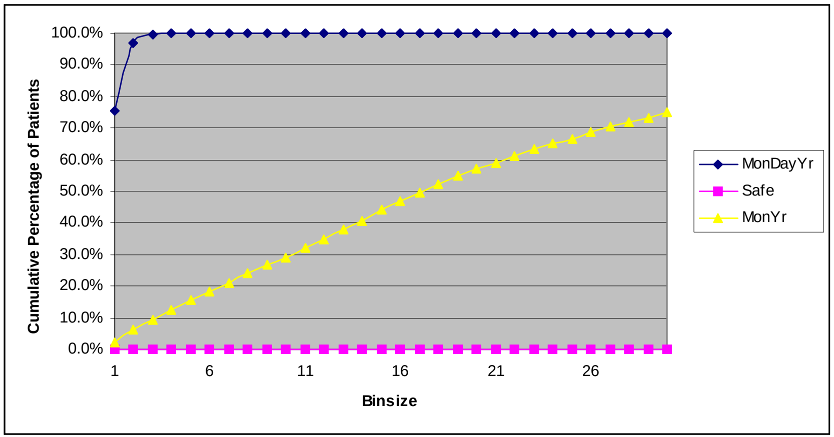
Fields of Bio-Surveillance Datastream Based on Usefulness For Detection -NY

	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	
	ICD9 diagnosis code 5	
	ICD9 diagnosis code 6	
Tier 2	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.

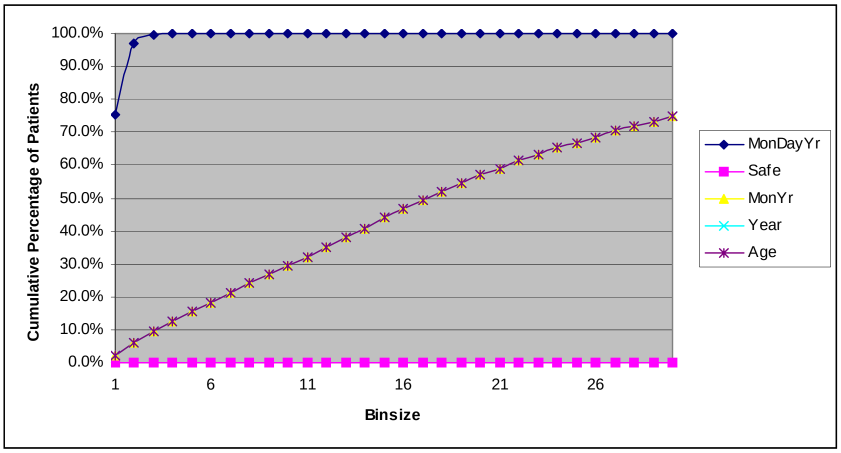
Fields of Bio-Surveillance Datastream Based on Usefulness For Detection -NY

	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	
	ICD9 diagnosis code 5	
	ICD9 diagnosis code 6	
Tier 2	Month, day and Year of Birth	DOB
	Gender	Sex

Generalize DOB more?

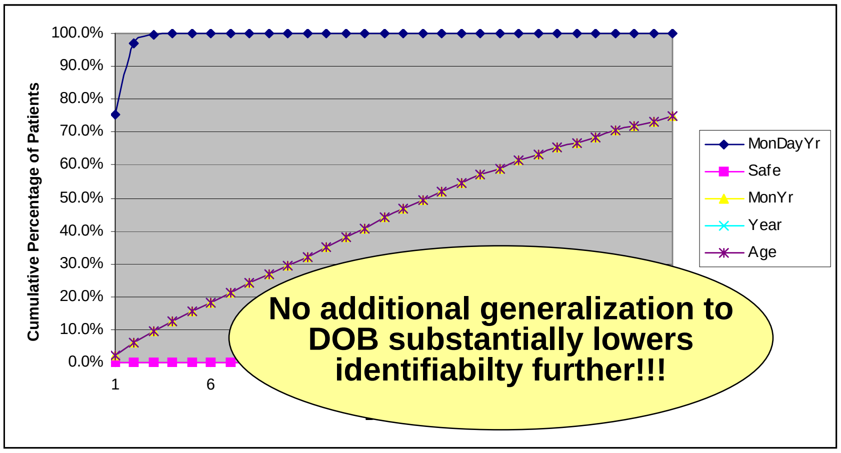
Given the improvement realized when date of birth was generalized to month and year of birth in in NY data, one might falsely believe generalizing DOB values further to year of birth, age or a 5-year range would provide further improvements -- not so!

Risk Assessment of Bio-Surveillance DataStream, Using Different Generalized Values for DOB



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

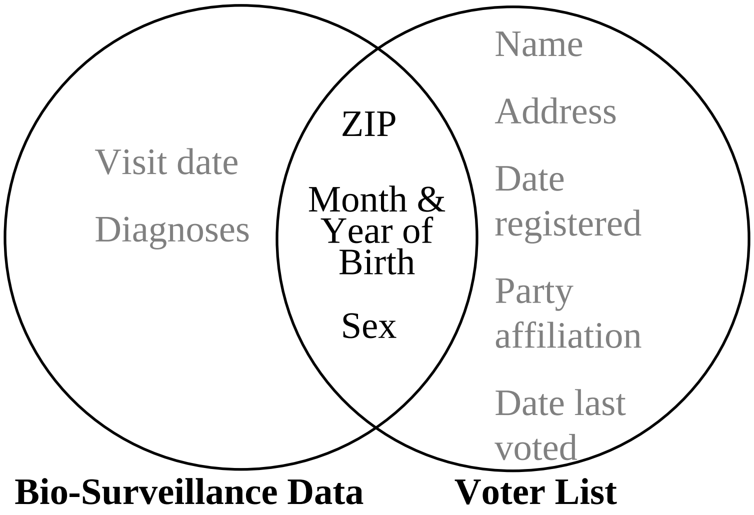
Risk Assessment of Bio-Surveillance DataStream, Using Different Generalized Values for DOB



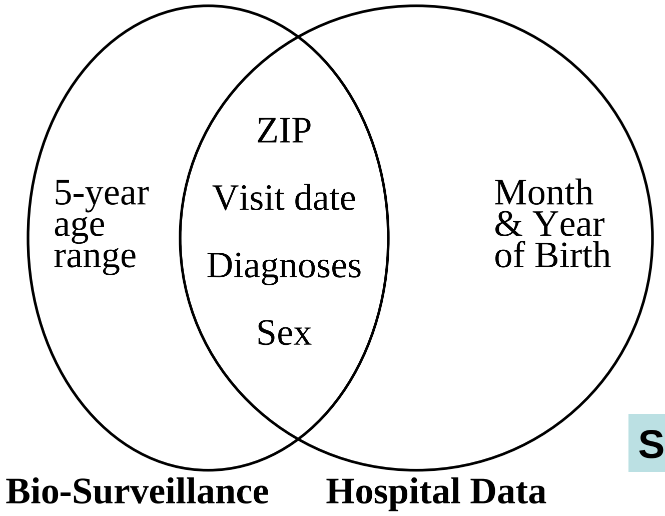
No additional generalization to DOB substantially lowers identifiabilty further!!!

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

First Inference Attack was a Direct Linkage Re-identification



Second Inference Attack was a 2-stage Linkage Re-identification



5-year age range

ZIP

Visit date

Diagnoses

Sex

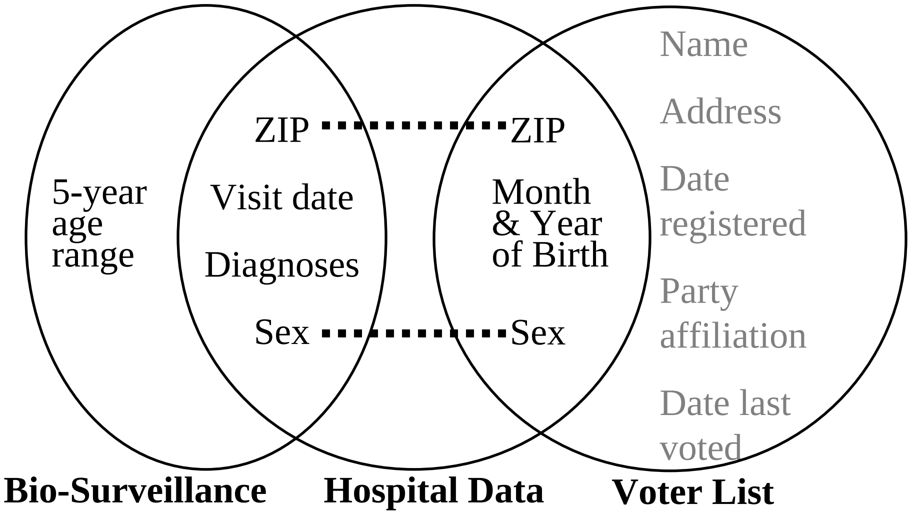
Month & Year of Birth

Bio-Surveillance

Hospital Data

Stage 1 of 2

Second Inference Attack was a 2-stage Linkage Re-identification



5-year age range

ZIP

Visit date

Diagnoses

Sex

ZIP

Month & Year of Birth

Sex

Name

Address

Date registered

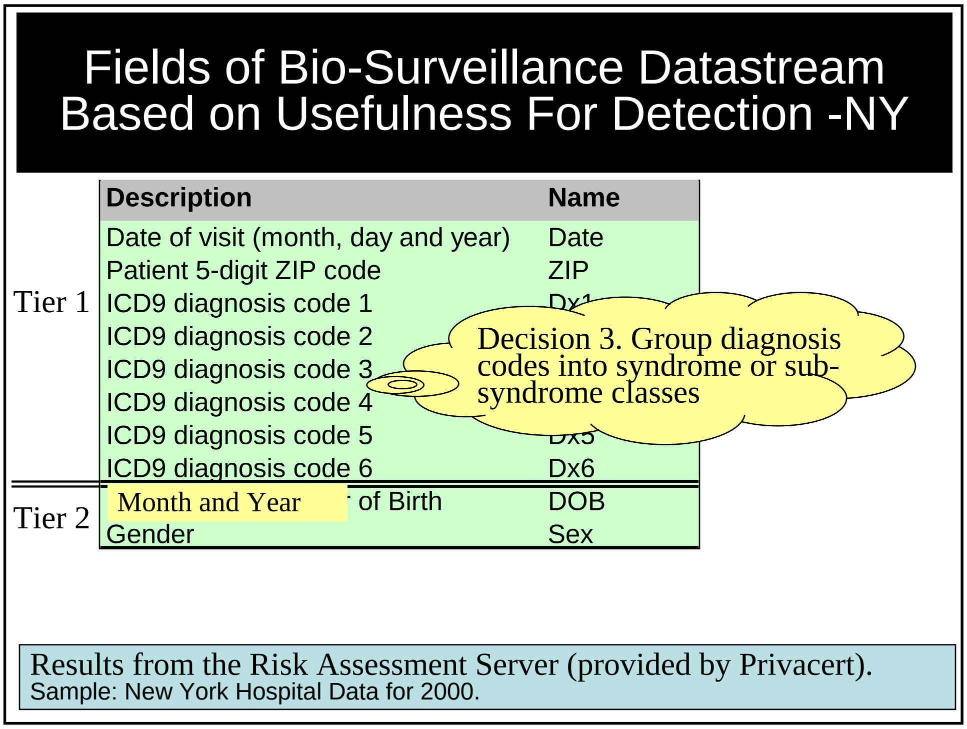
Party affiliation

Date last voted

Bio-Surveillance

Hospital Data

Voter List

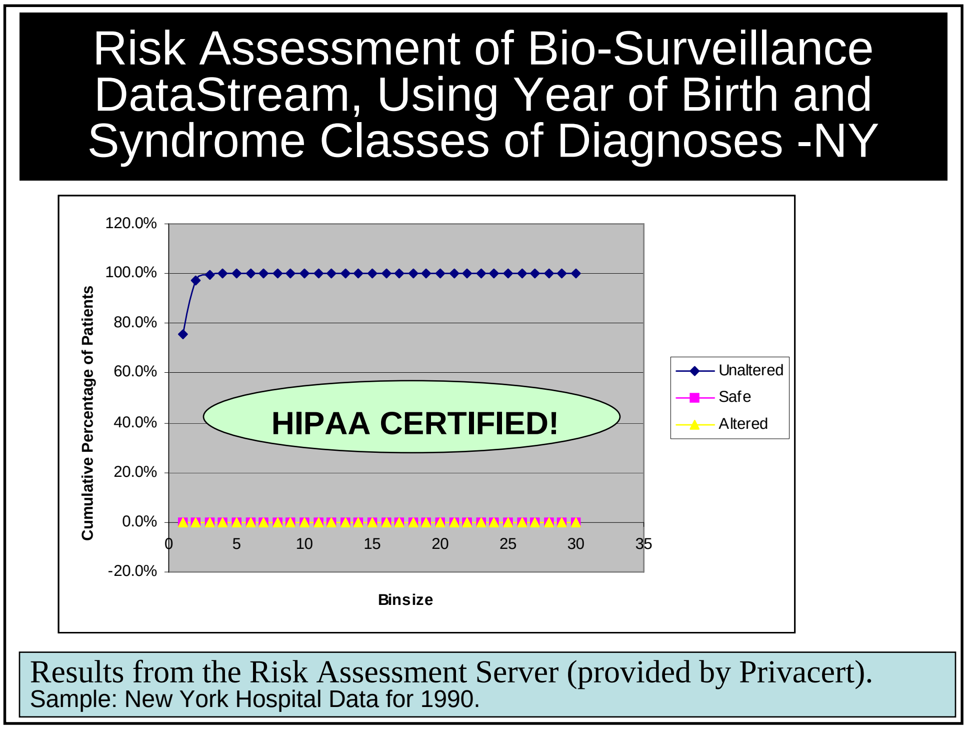


Fields of Bio-Surveillance Datastream Based on Usefulness For Detection -NY

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

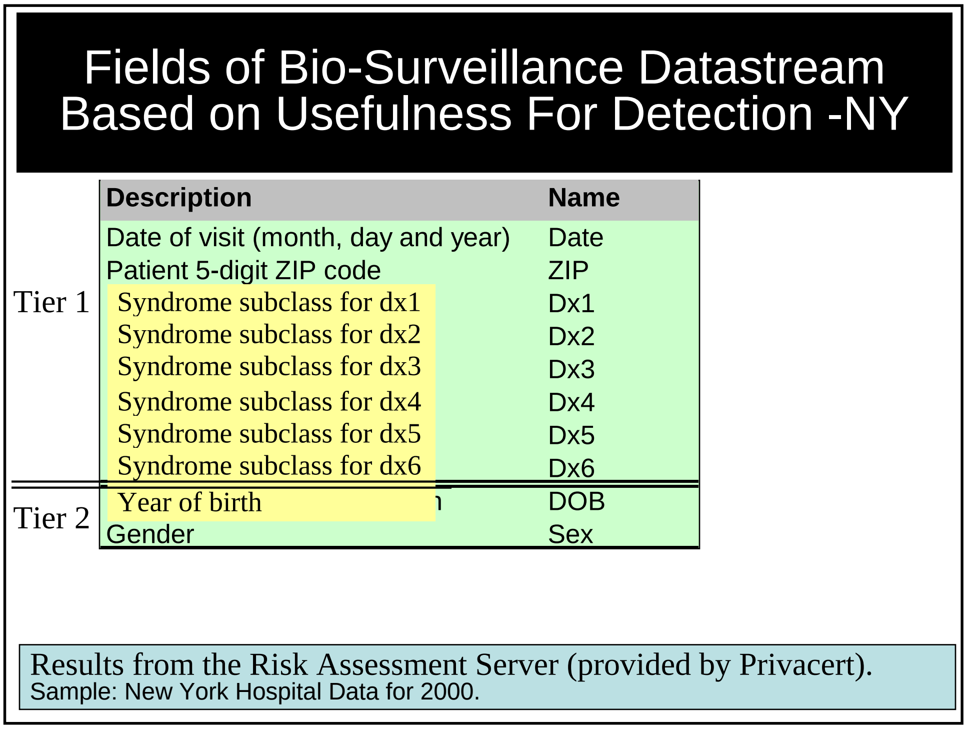
Decision 3. Group diagnosis codes into syndrome or sub-syndrome classes

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



Risk Assessment of Bio-Surveillance DataStream, Using Year of Birth and Syndrome Classes of Diagnoses -NY

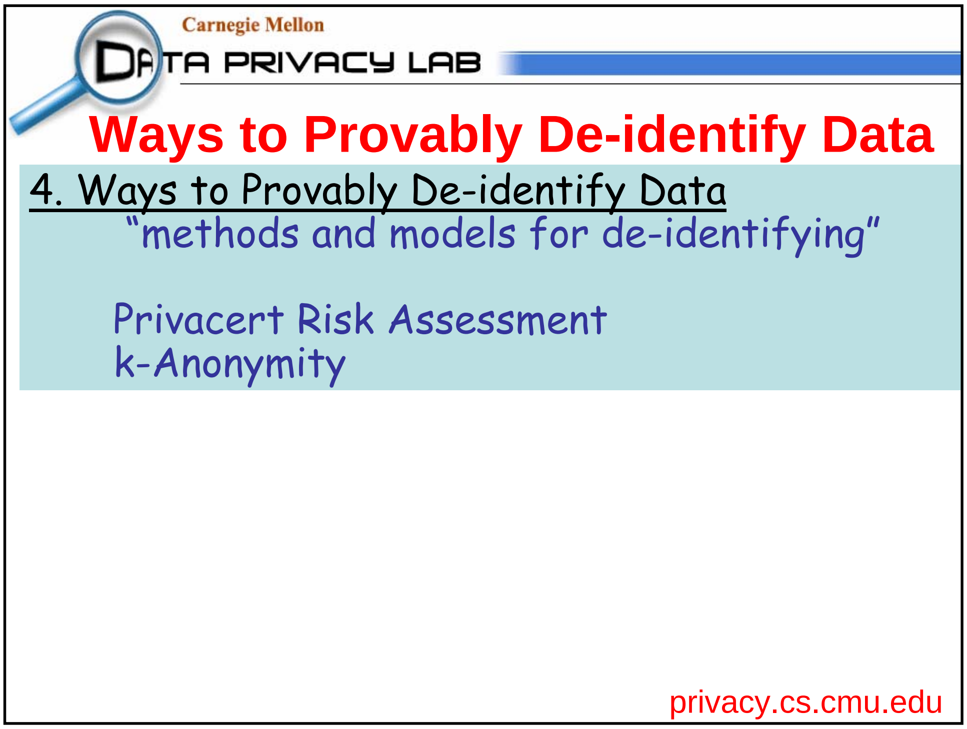
Results from the Risk Assessment Server (provided by Privacert).
Sample: New York Hospital Data for 1990.



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
	Syndrome subclass for dx1	Dx1
	Syndrome subclass for dx2	Dx2
	Syndrome subclass for dx3	Dx3
	Syndrome subclass for dx4	Dx4
	Syndrome subclass for dx5	Dx5
	Syndrome subclass for dx6	Dx6
Tier 2	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

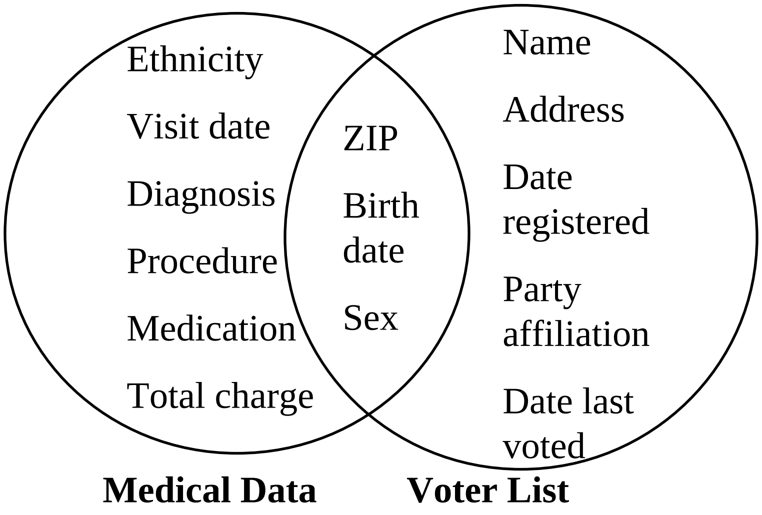
Ways to Provably De-identify Data

4. Ways to Provably De-identify Data
"methods and models for de-identifying"

Privacert Risk Assessment
k-Anonymity

privacy.cs.cmu.edu

Linking to re-identify data



Ethnicity

Visit date

Diagnosis

Procedure

Medication

Total charge

ZIP

Birth date

Sex

Name

Address

Date registered

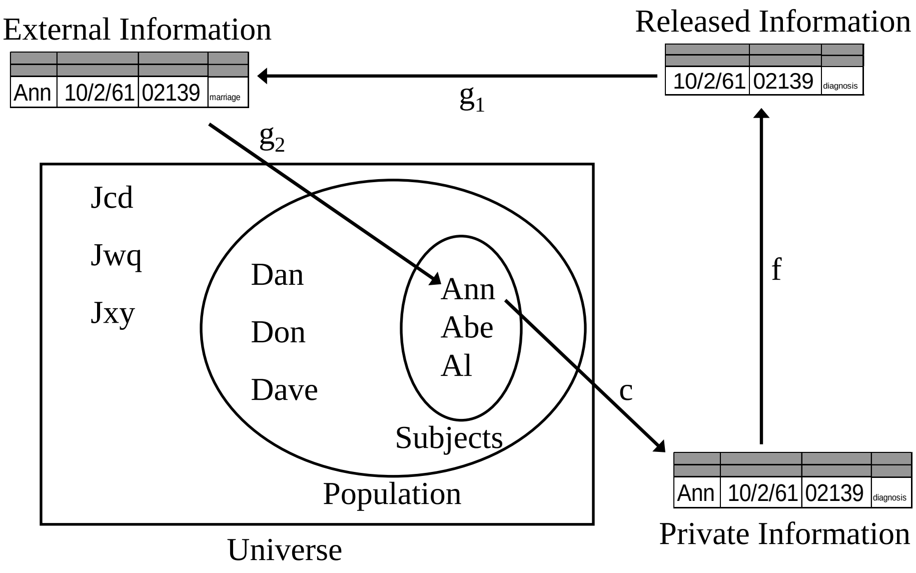
Party affiliation

Date last voted

Medical Data

Voter List

Disclosure overview



External Information

Ann	10/2/61	02139	marriage

Released Information

10/2/61	02139	diagnosis

g_1

g_2

f

c

Jcd

Jwq

Jxy

Dan

Don

Dave

Ann

Abe

Al

Subjects

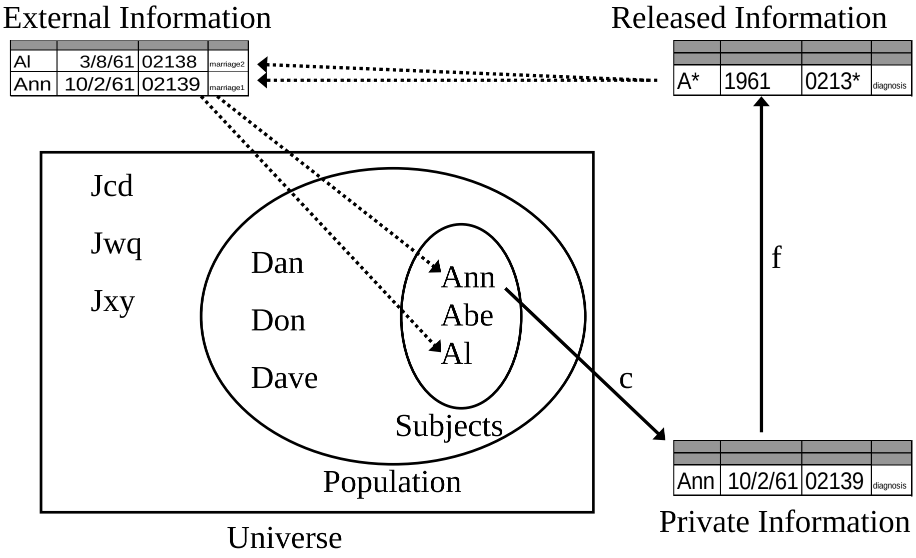
Population

Universe

Ann	10/2/61	02139	diagnosis

Private Information

Disclosure overview



External Information

Al	3/8/61	02138	marriage2
Ann	10/2/61	02139	marriage1

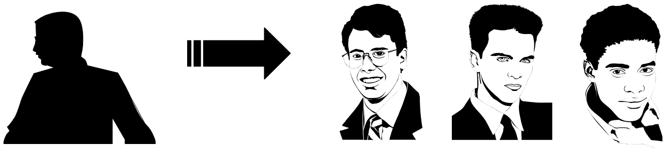
Released Information

A*	1961	0213*	diagnosis

Private Information

Ann	10/2/61	02139	diagnosis

Idea of *k*-map and *k*-anonymity

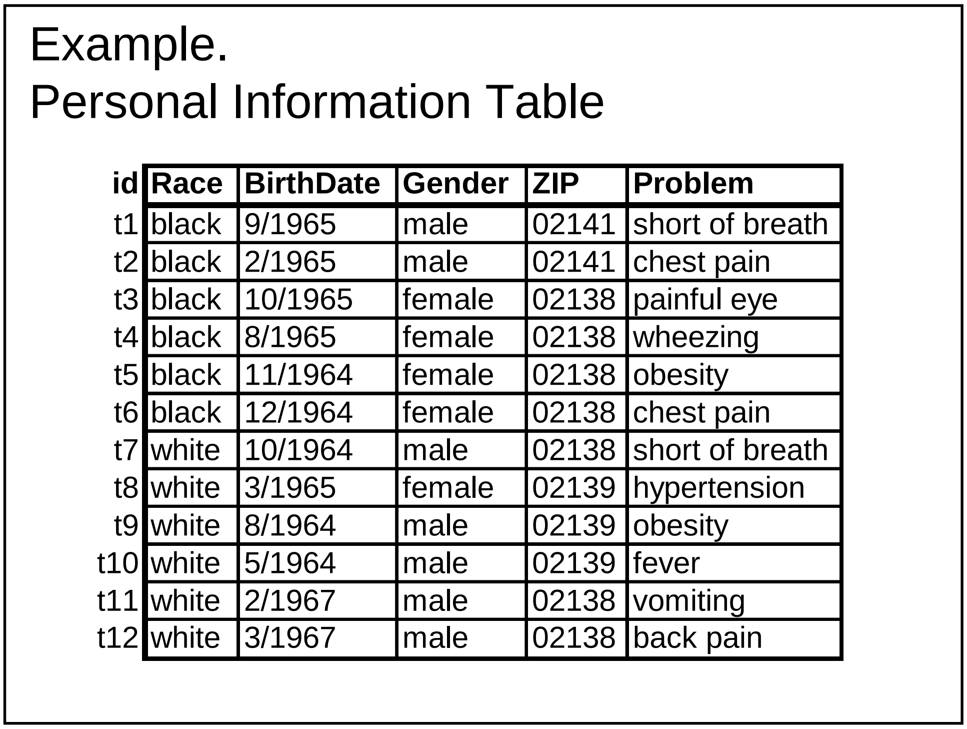


For every record released, there will be at least *k* individuals to whom the record indistinctly refers.

In *k*-map, the *k* individuals exist in the world.

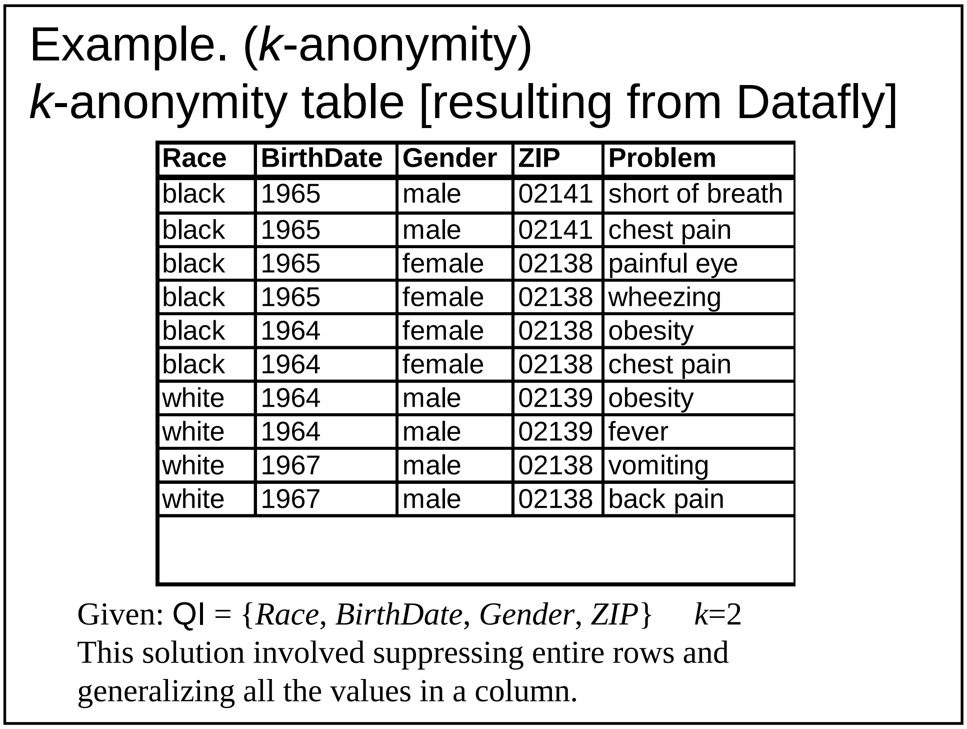
In *k*-anonymity, the *k* individuals appear in the release.

Sweeney 97 and 98



Example.
Personal Information Table

id	Race	BirthDate	Gender	ZIP	Problem
t1	black	9/1965	male	02141	short of breath
t2	black	2/1965	male	02141	chest pain
t3	black	10/1965	female	02138	painful eye
t4	black	8/1965	female	02138	wheezing
t5	black	11/1964	female	02138	obesity
t6	black	12/1964	female	02138	chest pain
t7	white	10/1964	male	02138	short of breath
t8	white	3/1965	female	02139	hypertension
t9	white	8/1964	male	02139	obesity
t10	white	5/1964	male	02139	fever
t11	white	2/1967	male	02138	vomiting
t12	white	3/1967	male	02138	back pain



Example. (*k*-anonymity)
k-anonymity table [resulting from Datafly]

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	02139	obesity
white	1964	male	02139	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

Given: QI = {*Race*, *BirthDate*, *Gender*, *ZIP*} *k*=2
This solution involved suppressing entire rows and generalizing all the values in a column.

Example. (*k*-anonymity) [Table GT1]

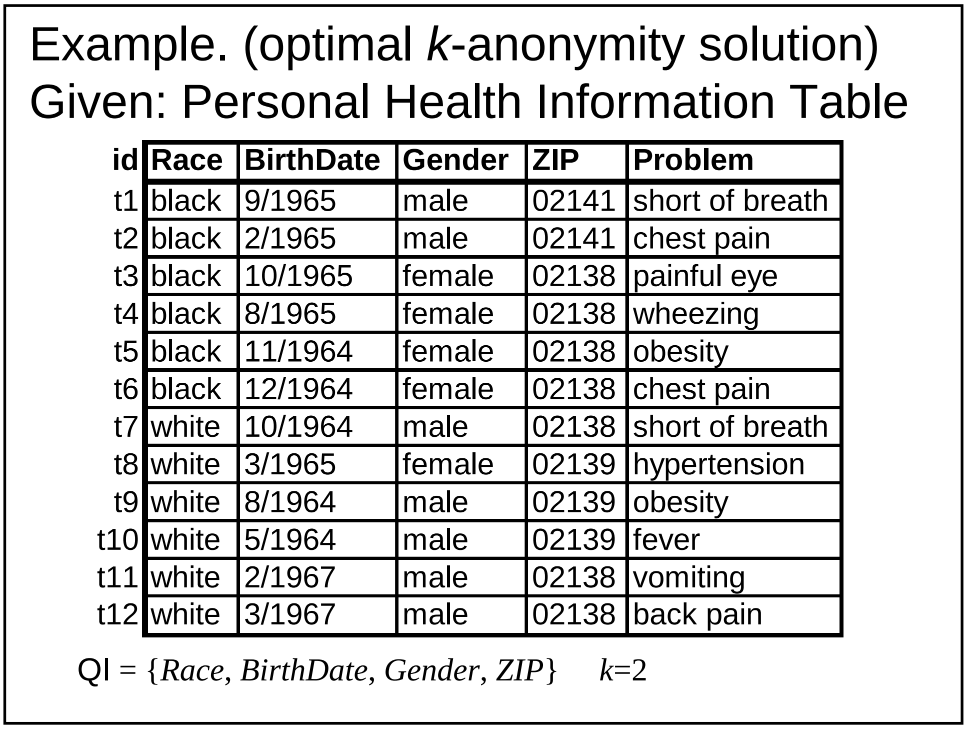
	Race	BirthDate	Gender	ZIP	Problem
t1	black	1965	male	02141	short of breath
t2	black	1965	male	02141	chest pain
t3	person	1965	female	0213*	painful eye
t4	person	1965	female	0213*	wheezing
t5	black	1964	female	02138	obesity
t6	black	1964	female	02138	chest pain
t7	white	1964	male	0213*	short of breath
t8	person	1965	female	0213*	hypertension
t9	white	1964	male	0213*	obesity
t10	white	1964	male	0213*	fever
t11	white	1967	male	02138	vomiting
t12	white	1967	male	02138	back pain

Given: QI = {*Race*, *BirthDate*, *Gender*, *ZIP*} *k*=2
cell-level generalization and suppression

Example. (*k*-anonymity) [Table GT3]

	Race	BirthDate	Gender	ZIP	Problem
t1	black	1965	male	02141	short of breath
t2	black	1965	male	02141	chest pain
t3	black	1965	female	02138	painful eye
t4	black	1965	female	02138	wheezing
t5	black	1964	female	02138	obesity
t6	black	1964	female	02138	chest pain
t7	white	1960-69	male	02138	short of breath
t8	white	1960-69	human	02139	hypertension
t9	white	1960-69	human	02139	obesity
t10	white	1960-69	human	02139	fever
t11	white	1960-69	male	02138	vomiting
t12	white	1960-69	male	02138	back pain

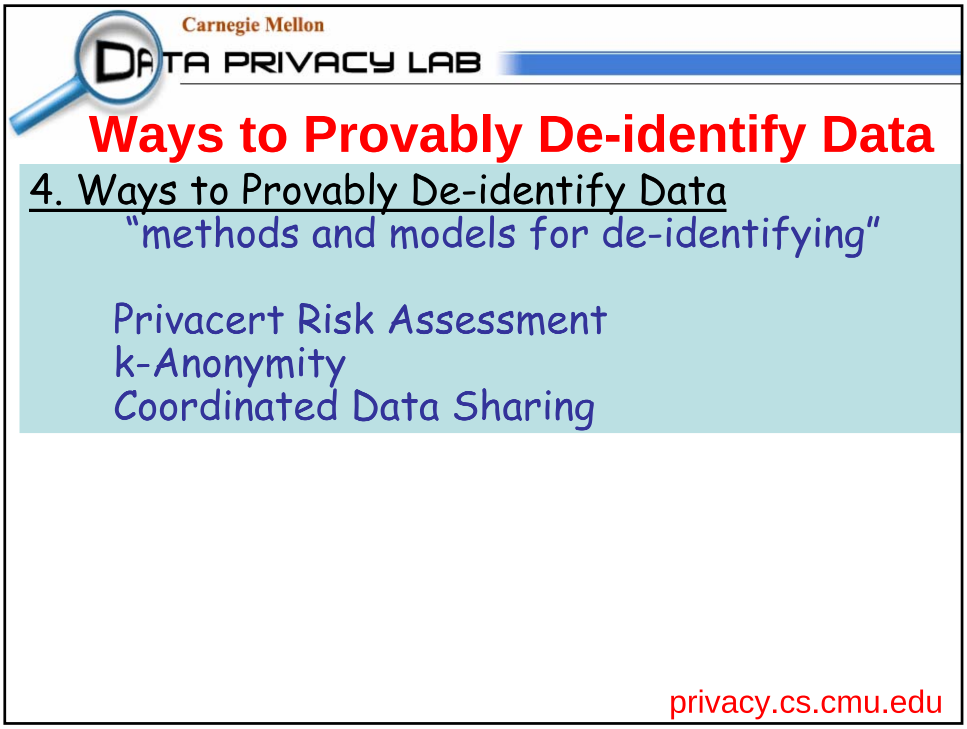
Given: QI = {*Race*, *BirthDate*, *Gender*, *ZIP*} *k*=2
cell-level generalization and suppression



Example. (optimal *k*-anonymity solution)
Given: Personal Health Information Table

id	Race	BirthDate	Gender	ZIP	Problem
t1	black	9/1965	male	02141	short of breath
t2	black	2/1965	male	02141	chest pain
t3	black	10/1965	female	02138	painful eye
t4	black	8/1965	female	02138	wheezing
t5	black	11/1964	female	02138	obesity
t6	black	12/1964	female	02138	chest pain
t7	white	10/1964	male	02138	short of breath
t8	white	3/1965	female	02139	hypertension
t9	white	8/1964	male	02139	obesity
t10	white	5/1964	male	02139	fever
t11	white	2/1967	male	02138	vomiting
t12	white	3/1967	male	02138	back pain

QI = {*Race*, *BirthDate*, *Gender*, *ZIP*}   *k*=2



Carnegie Mellon
DATA PRIVACY LAB

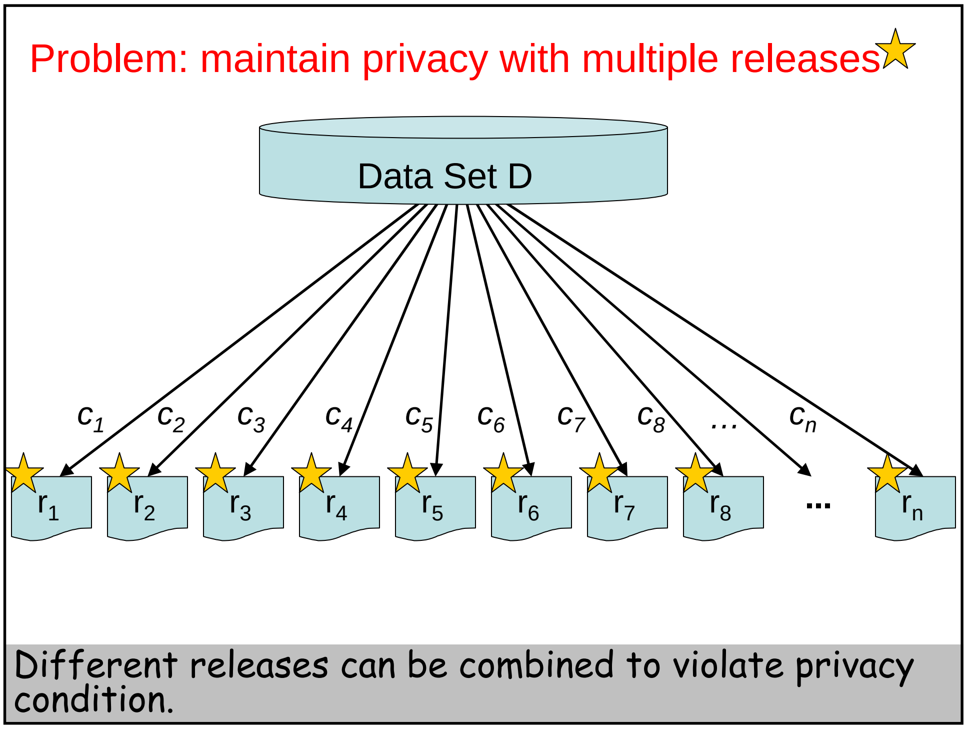
Ways to Provably De-identify Data

4. Ways to Provably De-identify Data

"methods and models for de-identifying"

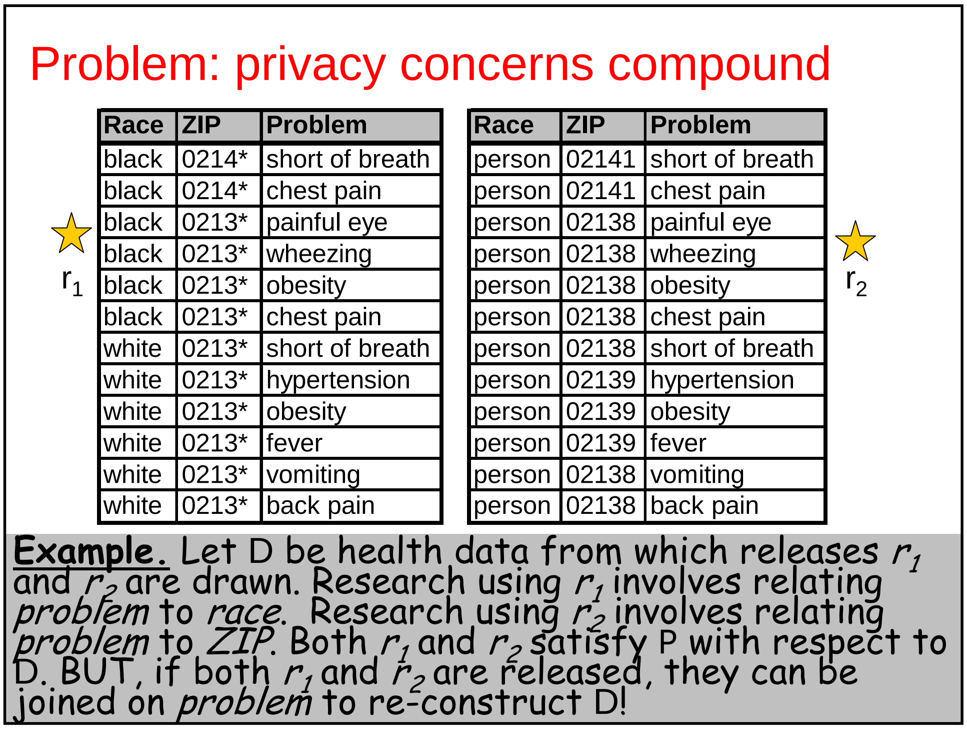
Privacert Risk Assessment
k-Anonymity
Coordinated Data Sharing

privacy.cs.cmu.edu



Problem: maintain privacy with multiple releases

Different releases can be combined to violate privacy condition.

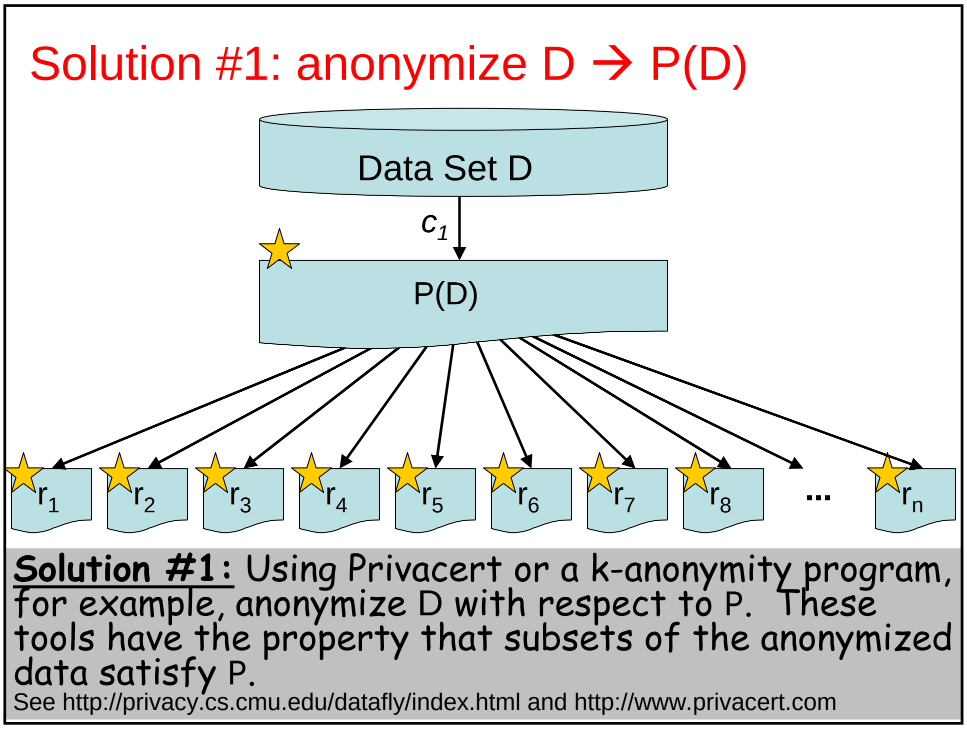


Problem: privacy concerns compound

Race	ZIP	Problem
black	0214*	short of breath
black	0214*	chest pain
black	0213*	painful eye
black	0213*	wheezing
black	0213*	obesity
black	0213*	chest pain
white	0213*	short of breath
white	0213*	hypertension
white	0213*	obesity
white	0213*	fever
white	0213*	vomiting
white	0213*	back pain

Race	ZIP	Problem
person	02141	short of breath
person	02141	chest pain
person	02138	painful eye
person	02138	wheezing
person	02138	obesity
person	02138	chest pain
person	02138	short of breath
person	02139	hypertension
person	02139	obesity
person	02139	fever
person	02138	vomiting
person	02138	back pain

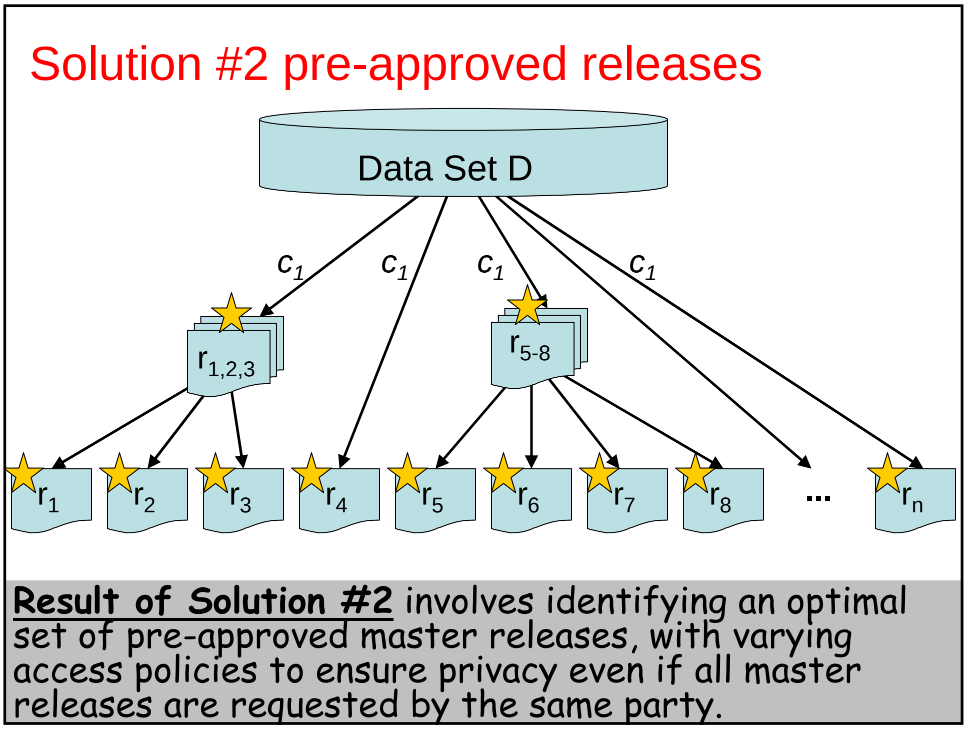
Example. Let D be health data from which releases r_1 and r_2 are drawn. Research using r_1 involves relating *problem* to *race*. Research using r_2 involves relating *problem* to *ZIP*. Both r_1 and r_2 satisfy P with respect to D. BUT, if both r_1 and r_2 are released, they can be joined on *problem* to re-construct D!



Solution #1: anonymize D → P(D)

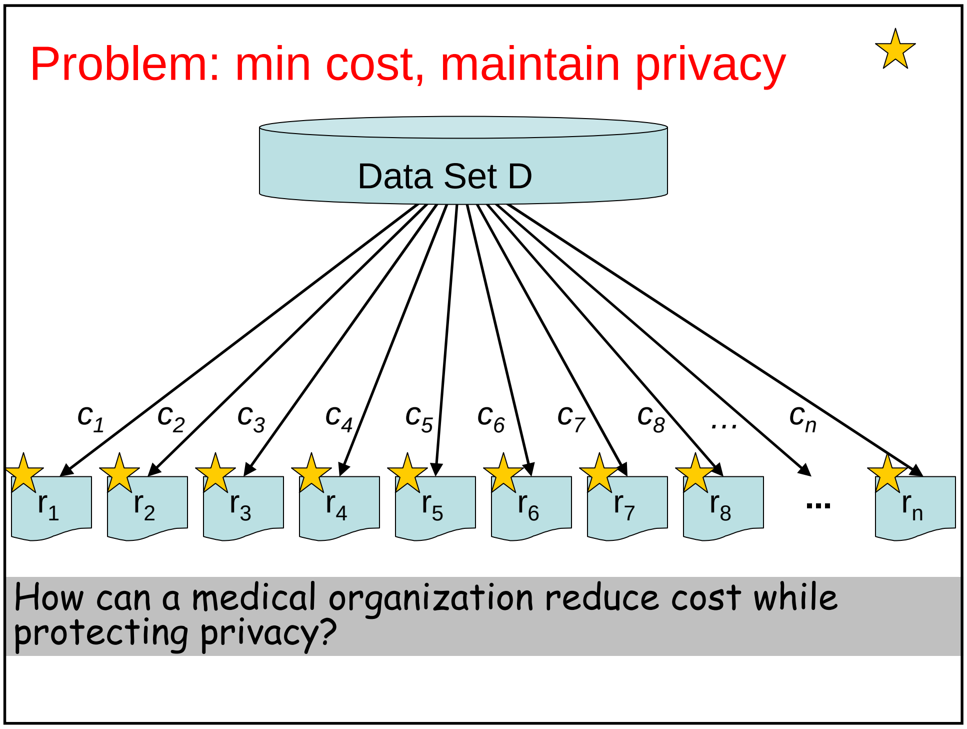
Solution #1: Using Privacert or a k-anonymity program, for example, anonymize D with respect to P. These tools have the property that subsets of the anonymized data satisfy P.

See http://privacy.cs.cmu.edu/datafly/index.html and http://www.privacert.com



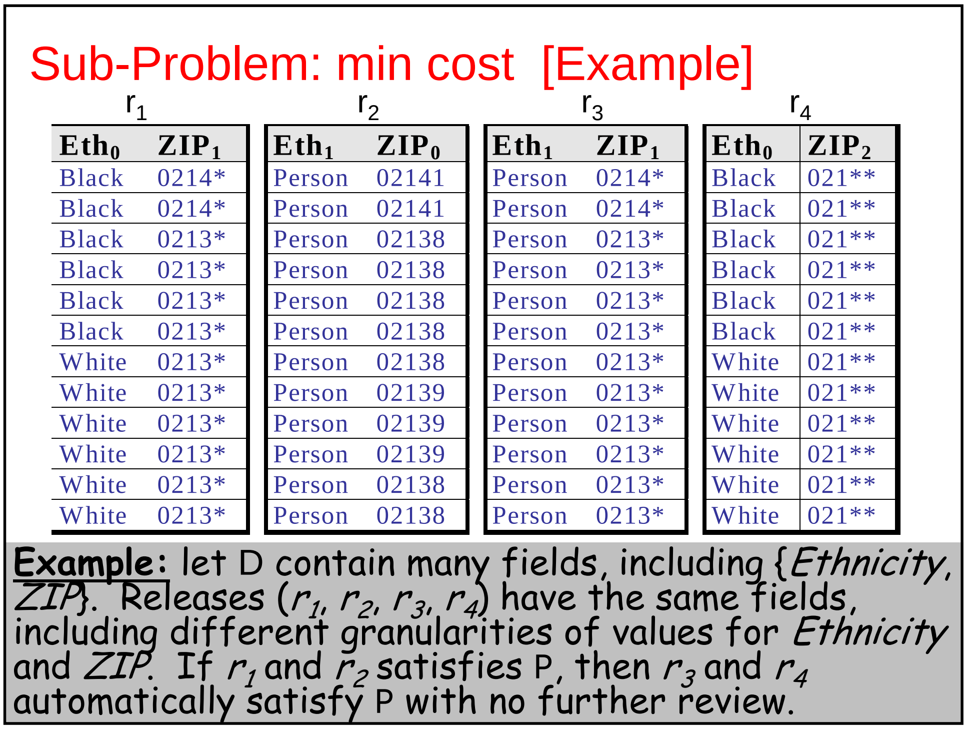
Solution #2 pre-approved releases

Result of Solution #2 involves identifying an optimal set of pre-approved master releases, with varying access policies to ensure privacy even if all master releases are requested by the same party.



Problem: min cost, maintain privacy

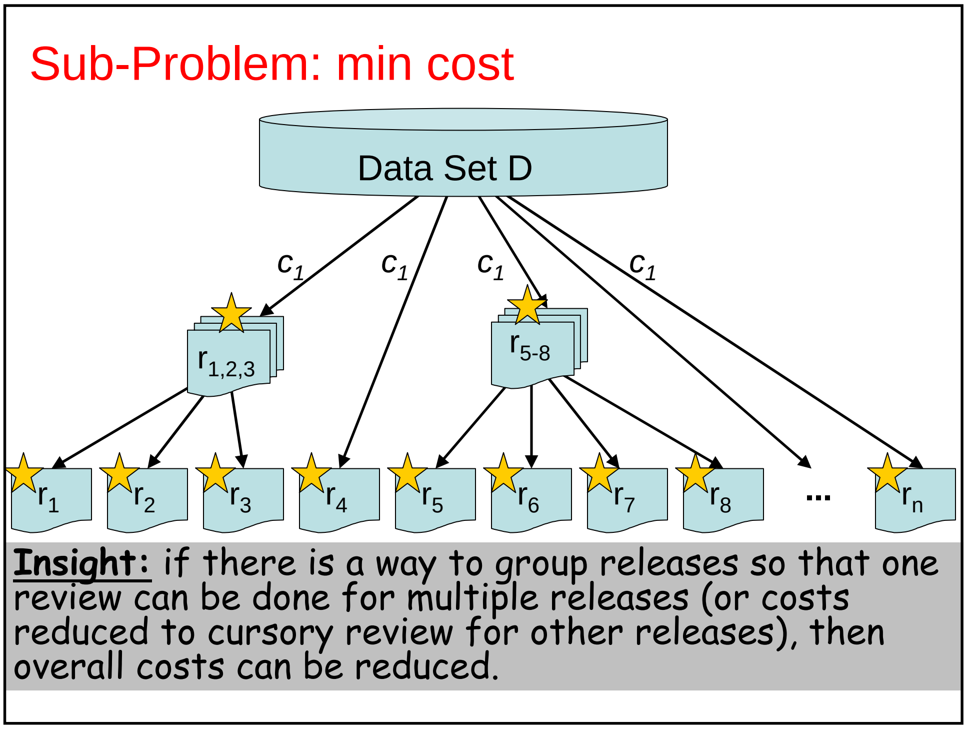
How can a medical organization reduce cost while protecting privacy?



Sub-Problem: min cost [Example]

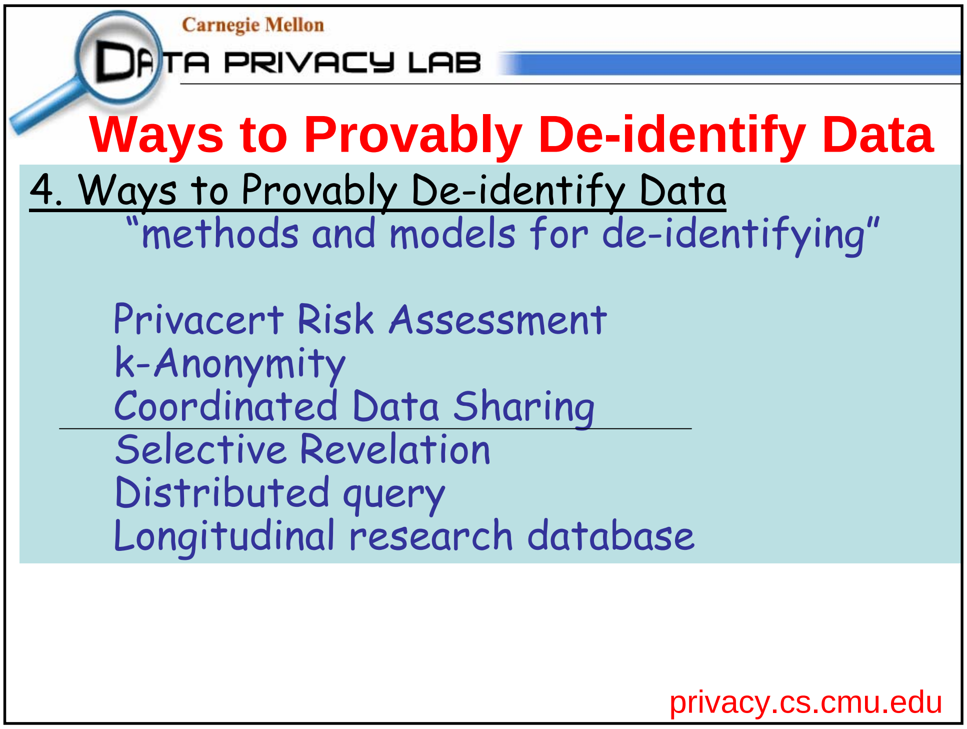
r_1		r_2		r_3		r_4	
Eth_0	**ZIP_1**	**Eth_1**	**ZIP_0**	**Eth_1**	**ZIP_1**	**Eth_0**	**ZIP_2**
Black	0214*	Person	02141	Person	0214*	Black	021**
Black	0214*	Person	02141	Person	0214*	Black	021**
Black	0213*	Person	02138	Person	0213*	Black	021**
Black	0213*	Person	02138	Person	0213*	Black	021**
Black	0213*	Person	02138	Person	0213*	Black	021**
Black	0213*	Person	02138	Person	0213*	Black	021**
White	0213*	Person	02138	Person	0213*	White	021**
White	0213*	Person	02139	Person	0213*	White	021**
White	0213*	Person	02139	Person	0213*	White	021**
White	0213*	Person	02139	Person	0213*	White	021**
White	0213*	Person	02138	Person	0213*	White	021**
White	0213*	Person	02138	Person	0213*	White	021**

Example: let D contain many fields, including {*Ethnicity, ZIP*}. Releases (r_1, r_2, r_3, r_4) have the same fields, including different granularities of values for *Ethnicity* and *ZIP*. If r_1 and r_2 satisfies P, then r_3 and r_4 automatically satisfy P with no further review.



Sub-Problem: min cost

Insight: if there is a way to group releases so that one review can be done for multiple releases (or costs reduced to cursory review for other releases), then overall costs can be reduced.



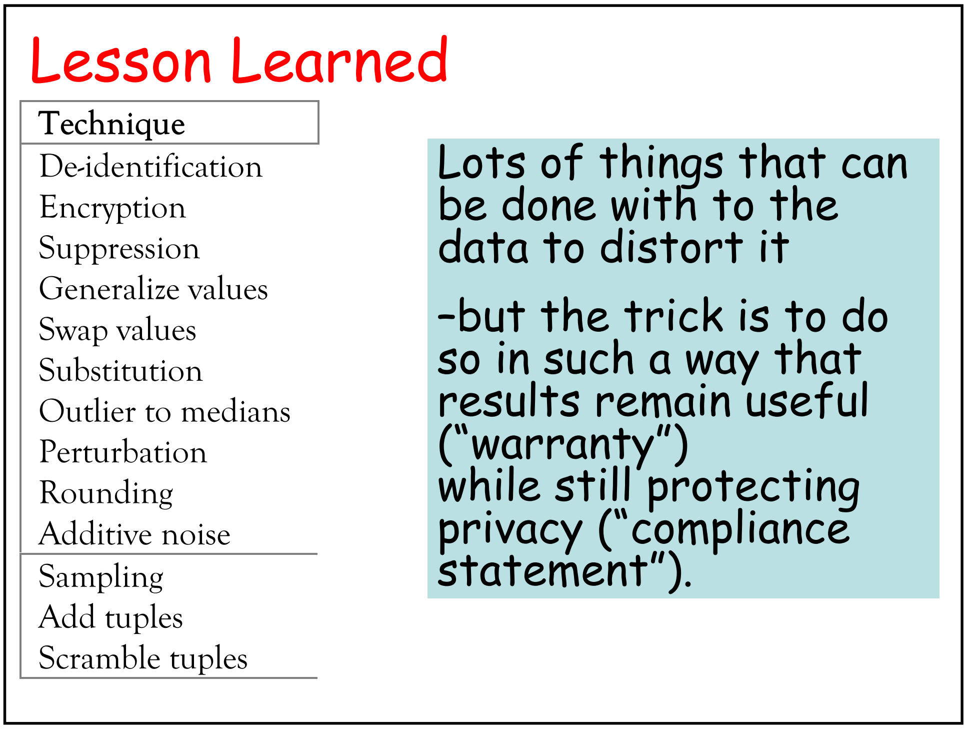
Carnegie Mellon
DATA PRIVACY LAB

Ways to Provably De-identify Data

4. Ways to Provably De-identify Data
"methods and models for de-identifying"

Privacert Risk Assessment
k-Anonymity
Coordinated Data Sharing
Selective Revelation
Distributed query
Longitudinal research database

privacy.cs.cmu.edu



Lesson Learned

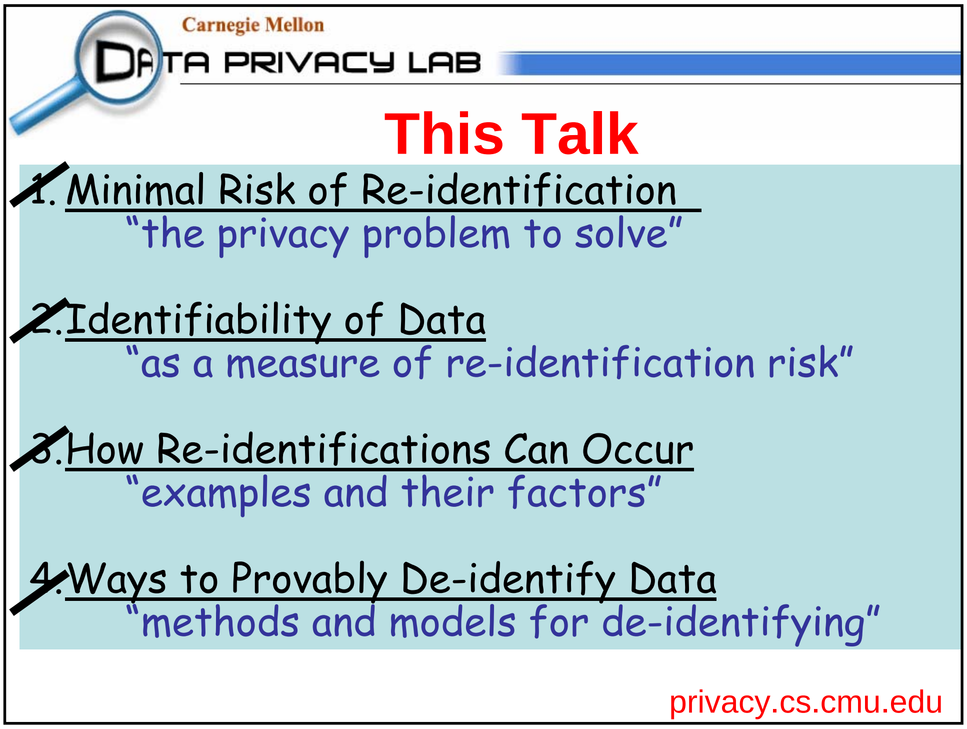
Technique
De-identification
Encryption
Suppression
Generalize values
Swap values
Substitution
Outlier to medians
Perturbation
Rounding
Additive noise
Sampling
Add tuples
Scramble tuples

Lots of things that can be done with to the data to distort it

-but the trick is to do so in such a way that results remain useful (“warranty”)
while still protecting privacy (“compliance statement”).

Carnegie Mellon
DATA PRIVACY LAB

Don’t use Ad Hoc Solutions



Carnegie Mellon

DATA PRIVACY LAB

This Talk

~~1~~. Minimal Risk of Re-identification
"the privacy problem to solve"

~~2~~. Identifiability of Data
"as a measure of re-identification risk"

~~3~~. How Re-identifications Can Occur
"examples and their factors"

~~4~~. Ways to Provably De-identify Data
"methods and models for de-identifying"

privacy.cs.cmu.edu