Carnegie Mellon
DATA PRIVACY LAB

Lecture 1: Overview of Privacy Technology

"a discussion about an emerging field within CS, how anonymity and policy specification fit in, some things we know how to do and some chances to save the world"

Latanya Sweeney

privacy.cs.cmu.edu

Carnegie Mellon
DATA PRIVACY LAB

Privacy Technology

technical integrations, add-ons and stand-alone technology with provable guarantees of privacy protection that enable many worthy uses.

In this talk, I will make 7 claims about privacy technology and provide examples.

Carnegie Mellon
DATA PRIVACY LAB

Privacy Technology

1. Privacy is here to stay.

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

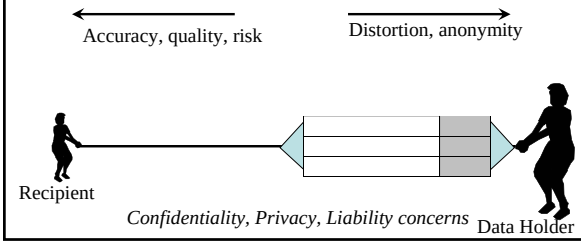
Example: VISA clearinghouse and Citibank

Carnegie Mellon
DATA PRIVACY LAB

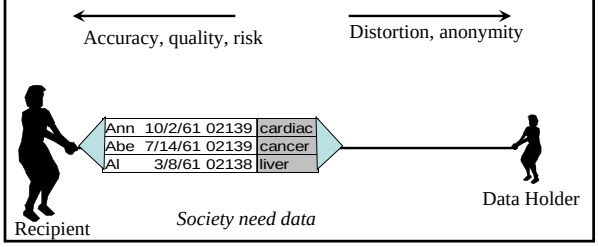
Privacy Technology

1. Privacy is here to stay.
2. Technology must have provable guarantees of privacy protection.

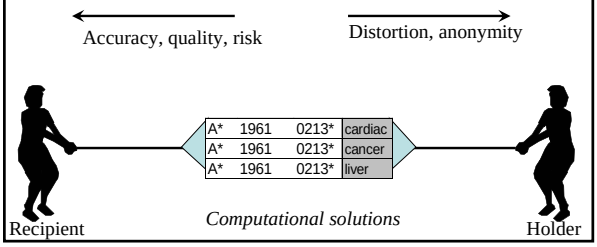
"Cannot Release Data"



"Privacy is Dead, Get over it!"



"Sharing data while providing guarantees of privacy"



Problem Addressed in this Talk

The emergence of many new technologies becomes increasingly hampered by privacy concerns because these technologies leave society vulnerable to privacy abuses.

Problem Addressed in this Talk

The emergence of many new technologies becomes increasingly hampered by privacy concerns because these technologies leave society vulnerable to privacy abuses.

Current situation.
Let society choose between benefiting from the technology and maintaining privacy protections.

Problem Addressed in this Talk

The emergence of many new technologies becomes increasingly hampered by privacy concerns because these technologies leave society vulnerable to privacy abuses.

Current situation.
Let society choose between benefiting from the technology and maintaining privacy protections.

Computer science solution → privacy technology.
Construct privacy technology with provable guarantees of privacy protection while allowing society to collect and share person-specific information for many worthy purposes .

Carnegie Mellon
DATA PRIVACY LAB

Privacy Technology

1. Privacy is here to stay.
2. Technology must have provable guarantees of privacy protection.
3. Privacy technology represents new pursuits in computer science.

Computer Scientists Must Save the World

Policy is limited by that which can be expressed in words.

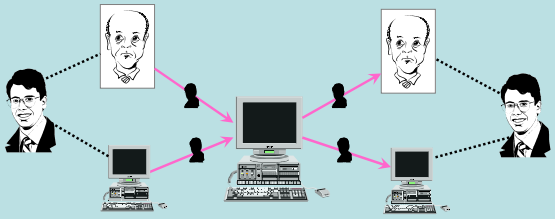
Traditional research based on **statistical and economic** approaches to policy problems tend to be retrospective and descriptive and assume technology is static.

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.

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: login with password
Authorization: allowed to read/write data
Encryption: to avoid eavesdropping

BUT data can re-identify individual!

Relationships Between Related Areas

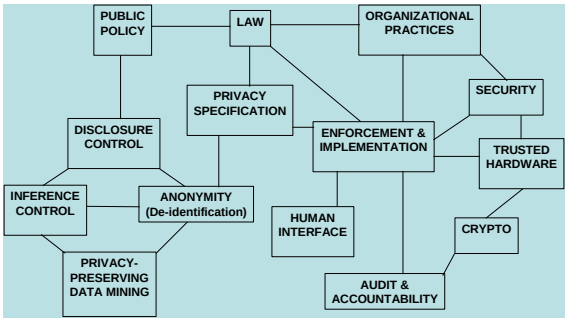


Diagram courtesy of Michael Shamos.

Data Anonymity (*new field* of CS)

The study of computational solutions for releasing data such that the data remain practically useful while the identities of the subjects of the data are not revealed.

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 consists of Anonymity and Privacy Rights, in single database and in ubiquitous settings

	anonymity	*rights mgt*	*database*	*ubiquitous*
AI	heavy	some	light	heavy
learning	some	light	light	heavy
theory	heavy	light	some	heavy
database	light	light	heavy	some
language	light	heavy	light	some
security		light	some	some
IS		light	light	light

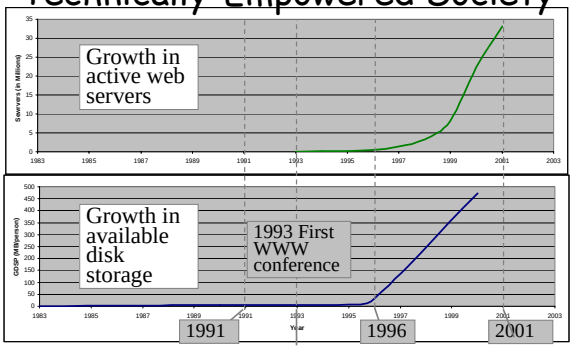
"AI" primarily concerns knowledge representation and semantics.
"Learning" focuses on data mining algorithms.
"Theory" includes zero-knowledge proofs and multi-party computations

Carnegie Mellon
DATA PRIVACY LAB

Privacy Technology

1. Privacy is here to stay.
2. Technology must have provable guarantees of privacy protection.
3. Privacy technology represents new pursuits in computer science.
4. Very hard to provide provable guarantees of privacy.

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

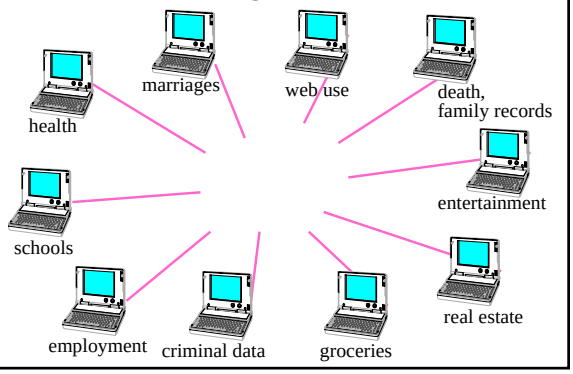
Typical Electronic Birth Certificate Fields in 1999 -*starting fields 196-210*

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



Carnegie Mellon DATA PRIVACY LAB

Privacy Technology

1. Privacy is here to stay.
2. Technology must have provable guarantees of privacy protection.
3. Privacy technology represents new pursuits in computer science.
4. Very hard to provide provable guarantees of privacy.
5. But we can learn interesting (sensitive) information from all that data!

Carnegie Mellon DATA PRIVACY LAB

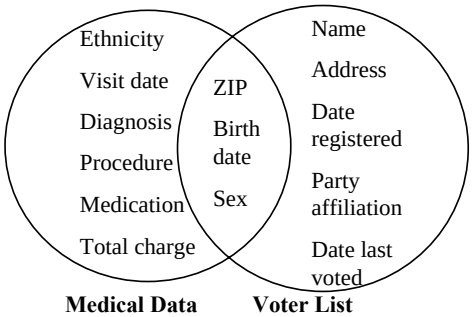
Semantic Learning

- Mining strategic information from text, and from video
- Automated profiles (putting disparate pieces together)
- Resolving ambiguous identities in data (e.g. *Michael Jordan*, the basketball player, vs. *Michael Jordan*, the computer scientist)

Edoardo Airoldi
William Gronim
Ralph Gross
Kishore Madhava
Bradley Malin

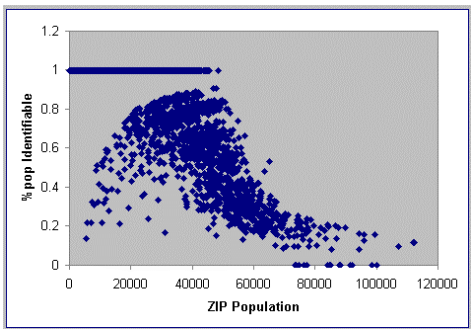
Algorithms for learning sensitive information from seemingly innocent information.

Linking to re-identify data

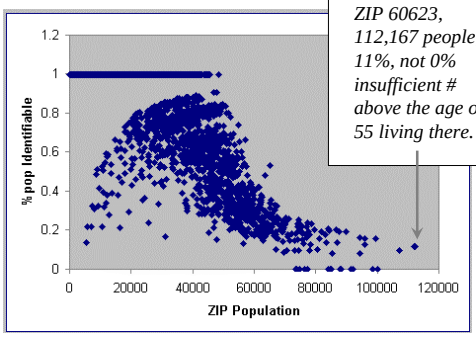


L. Sweeney. Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine and Ethics*. 1997, 25:98-110.

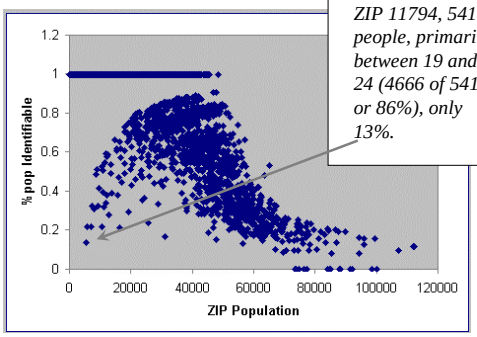
{*date of birth*, *gender*, *5-digit ZIP*} uniquely identifies 87.1% of USA pop.



{*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.

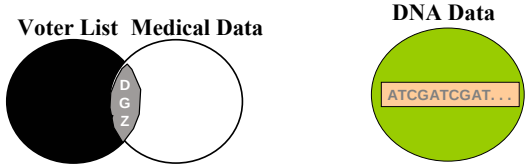


Carnegie Mellon
DATA PRIVACY LAB

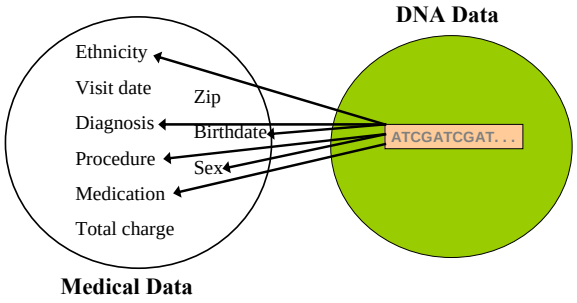
DNA Re-identification

Bradley Malin
Latanya Sweeney

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

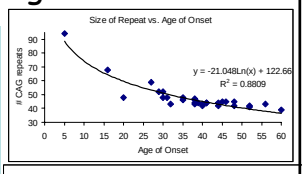


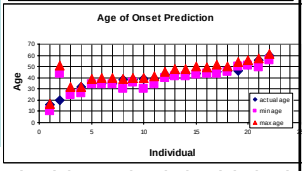
Build Up Relations



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





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
Elaine Newton
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

Websites Share Weblogs

128.2.65.781	Yoda
134.6.8.91	Luke
34.1.687.21	Leah
82.912.32.1	Obi

ebay

81.2.1.541	Han
134.6.8.91	Luke
82.912.32.1	Obi

amazon.com

128.2.65.781	Yoda
134.6.8.91	Luke
322.46.7.93	C3PO
12.78.96.54	Jabba

ORBITZ

322.46.7.93	C3PO
82.912.32.1	Obi
34.1.687.21	Leah
51.3.5.677	Lando

BARNES&NOBLE

Websites Share Consumer Lists

128.2.65.781	Yoda
134.6.8.91	Luke
34.1.687.21	Leah
82.912.32.1	Obi

ebay

81.2.1.541	Han
134.6.8.91	Luke
82.912.32.1	Obi

amazon.com

128.2.65.781	Yoda
134.6.8.91	Luke
322.46.7.93	C3PO
12.78.96.54	Jabba

ORBITZ

322.46.7.93	C3PO
82.912.32.1	Obi
34.1.687.21	Leah
51.3.5.677	Lando

BARNES&NOBLE

Yoda
Luke
Leah
Obi
Han
Luke
Obi
Yoda
Luke
C3PO
Jabba
C3PO
Obi
Leah
Lando

Tracks and Trails

P — Identified

TRACK	ebay	amazon.com	ORBITZ	BARNES&NOBLE
Yoda	1	0	1	0
Luke	1	1	1	1
Leah	1	0	0	1
Obi	1	1	0	1
Han	0	1	0	1
C3PO	0	0	1	1
Jabba	0	0	1	0
Lando	0	0	0	1

TRAIL

trail(P, *p*)

N — De-identified

	ebay	amazon.com	ORBITZ	BARNES&NOBLE
128.2.65.781	1	0	1	0
134.6.8.91	1	1	1	0
34.1.687.21	1	0	0	1
82.912.32.1	1	1	0	1
322.46.7.93	0	1	0	0
322.46.7.93	0	0	1	1
12.78.96.54	0	0	1	0
513.5.677	0	0	0	1

Note: Size of track $T = |T|$ = # of distinct entities

REIDIT-I Example

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

Example of what was learned:

"Luke" (known by name) visited Amazon even though he never bought anything at Amazon.

Reidentified

Carnegie Mellon
DATA PRIVACY LAB

Privacy Technology

1. Privacy is here to stay.
2. Technology must have provable guarantees of privacy protection.
3. Privacy technology represents new pursuits in computer science.
4. Very hard to provide provable guarantees of privacy.
5. But we can learn interesting (sensitive) information from all that data!
6. And on the other side, we control what can be learned (protect privacy)!

Carnegie Mellon
DATA PRIVACY LAB

Privacy Technologies

- De-identify text
- De-identify video
- Distributed Surveillance
- Risk Assessment Server
- Formal protection models: k-map, null-map, wrong-map, k-anonymity

Edoardo Airoldi
Samuel Edoho-Eket
Ralph Gross
Kishore Madhava
Bradley Malin
Yiheng Li
Latanya Sweeney

Algorithms with provable guarantees of privacy protection that enable data sharing for many worthy purposes.

Sample letter

Wednesday, February 2, 1994

Marjorie Long, M.D.
St. John's Hospital
Huntington 18
Boston, MA 02151

RE: Virginia Townsend
CH#32-841-09787
DOB 05/26/86

Dear Dr. Lang:

I feel much better after seeing Virginia this time. Dot is a 7 and 6/12 year old female in follow up for insulin dependent diabetes mellitus diagnosed in June of 1993 by Dr. Frank at Brigham's. She is currently on Lily Human Insulin and is growing and gaining weight normally. She will start competing again with the U. S. Junior Gymnastics team. We will contact Mrs. Hodgkins at Marina Corp 473-1214 for a follow-up visit for her daughter.

Patrick Hayes, M.D. 34764

"Scrubbed" letter

February 1994

Erisa Cosborn, M.D.
Brighaul Hospital
Alberdam Way
Peabon, MA ***02100***

RE: *Kathel Wallams*
CH#*18-512-32871*
DOB ***05/86***

Dear Dr. ***Jandel***:

I feel much better after seeing *Kathel* this time. ***Cob*** is a 7 and 6/12 year old female in follow up for insulin dependent diabetes mellitus diagnosed in June of 1993 by Dr. ***Wandel*** at ***Naming's***. She is currently on Lily Human Insulin and is growing and gaining weight normally. She will start competing again with the ___ . We will contact Mrs. ***Learl*** at ***Garlaw*** Corp ***912-8205*** for a follow-up visit for her daughter.

Mank Brones, M.D. *21075*

Carnegie Mellon
DATA PRIVACY LAB

Scrub System (text de-identification)

Latanya Sweeney

Method	Letters
Straight search	37%
Search with cues	82%
Scrub System	98-100%

Percentage determined by:

$$\frac{\#\text{correct} - \#\text{false positives}}{\#\text{occurrences}}$$

Based on 3198 letters in pediatric data.

Sweeney L. Replacing Personally-Identifying Information in Medical Records, the Scrub System. In: Cimino, JJ, ed. *Proceedings, Journal of the American Medical Informatics Association.* Washington, DC: Hanley & Belfus, Inc., 1996.

Thwarting 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

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

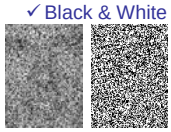
Ordinal

Threshold

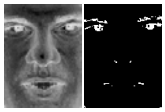
Ad Hoc Schemes that Don't Work II

Pixelation

Random
- ✓ Grayscale
- ✓ Black & White



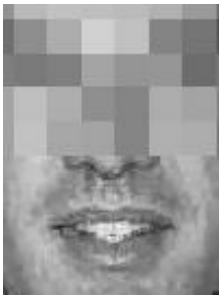
Negative
- ✓ Grayscale
- ✓ Black & White



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!



Carnegie Mellon DATA PRIVACY LAB

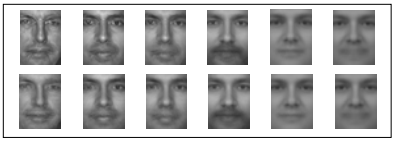
k-Same anonymizing faces

Elaine Newton
Latanya Sweeney
Bradley Malin

Thwarts face recognition while many facial details remain!

-Pixel

-Eigen



k = 2 3 5 10 50 100

Carnegie Mellon DATA PRIVACY LAB

Privacy Technology

1. Privacy is here to stay.
2. Technology must have provable guarantees of privacy protection.
3. Privacy technology represents new pursuits in computer science.
4. Very hard to provide provable guarantees of privacy.
5. But we can learn interesting (sensitive) information from all that data!
6. And on the other side, we control what can be learned (protect privacy)!
7. We can even *prove* privacy is protected!

Carnegie Mellon DATA PRIVACY LAB

k-Anonymity

Yiheng Li
Latanya Sweeney
Ryan Williams

- Proposed by Latanya Sweeney
- Given a table *T* of data, "suppress" or "generalize" entries of *T* so that for every row, *k*-1 other rows look identical

first	last	age	race
Harry	Stone	34	Afr-Am
John	Reyser	36	Cauc
Beatrice	Stone	47	Afr-Am
John	Ramos	22	Hisp

↓

first	last	age	race
*	Stone	30-50	Afr-Am
John	R*	20-40	*
*	Stone	30-50	Afr-Am
John	R*	20-40	*

Carnegie Mellon DATA PRIVACY LAB

Nature of k-anonymity

Ryan Williams

- *Given an arbitrary table, what's the minimum number of entries that must be "suppressed" in order to achieve k-anonymity?*
- (And, what is that set of entries?)
- *NP-hard in general*
- However, there is a polynomial time algorithm suppressing no more than O(*k log k*) times the optimal number; uses set cover and other tricks

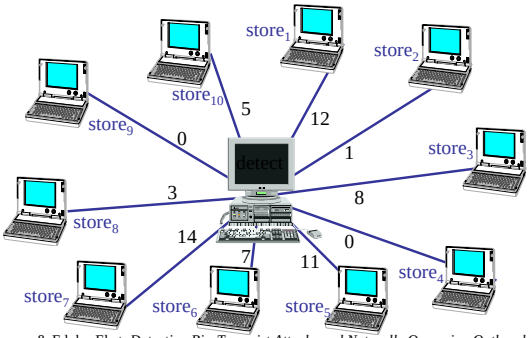
Carnegie Mellon
DATA PRIVACY LAB

Learning from k-anonymized data:

Yiheng Li

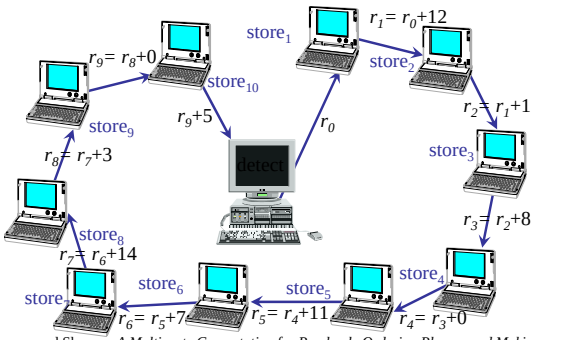
- GenTree is an algorithm that learns association rules from k-anonymized data
- *The approach led to more robust rule learning than was possible before in non-anonymized data (data mining improvement)*

"How many *x* occurred yesterday?"



Sweeney & Edoho-Eket. *Detecting Bio-Terrorist Attacks and Naturally Occurring Outbreaks Over a Distributed Network While Protecting Privacy and Confidentiality: the PrivaSum Protocol*. Carnegie Mellon, School of Computer Science, Tech Report, CMU-ISRI-04-111. 2004.

Total count: "How many *x* occurred?"



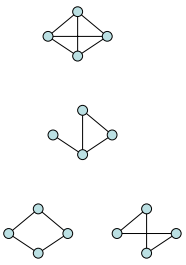
Sweeney and Shamos. *A Multiparty Computation for Randomly Ordering Players and Making Random Selections*. Carnegie Mellon University, School of Computer Science, Technical Report, CMU-ISRI-04-126. Pittsburgh: July 2004.

Carnegie Mellon
DATA PRIVACY LAB

Distributed Surveillance

Samuel Edoho-Eket
Michael Shamos
Latanya Sweeney

- Start with trivial secret sharing, multi-party
- Probabilistic multi-party solution (Improvement!)
- Randomized multi-party solution (new!)



Carnegie Mellon
DATA PRIVACY LAB

Privacy Technology

1. Privacy is here to stay.
2. Technology must have provable guarantees of privacy protection.
3. Privacy technology represents new pursuits in computer science.
4. Very hard to provide provable guarantees of privacy.
5. But we can learn interesting (sensitive) information from all that data!
6. And on the other side, we control what can be learned (protect privacy)!
7. We can even *prove* privacy is protected!

Carnegie Mellon
DATA PRIVACY LAB

CameraWatch

Ralph Gross
Kishore Madhava
Michael Shamos
Latanya Sweeney

- **Benefits: personal decisions about travel, weather, family members**
- **Privacy concerns: reactionary**
- **Testbed: CameraWatch database of 1000's of webcams**

privacy.cs.cmu.edu/dataprivacy/projects/camwatch