

Carnegie Mellon
DATA PRIVACY LAB

Data Privacy *Friend or Foe?*

Bradley Malin, malin@cs.cmu.edu
Data Privacy Laboratory, School of Computer Science
Carnegie Mellon University
October 26, 2004

Carnegie Mellon
DATA PRIVACY LAB

Privacy Policy

- Some Limitations
 - Need robust language
 - Need enforcement
 - Scope of world / interaction
 - Syntax, not semantics
- What is Data Privacy?
 - *WHERE* does data come from?
 - *WHAT* does data reveal?
 - *HOW* do we prove data does not reveal more than specified?

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

Privacy Policy

- Some Positives
 - *Procedure*
 - Specifies how data can (not) be used
 - *Logical Cognition*
 - Requires active involvement and thought regarding information
 - *Standardization*
 - equal opportunity
 - *Legal Enforcement*

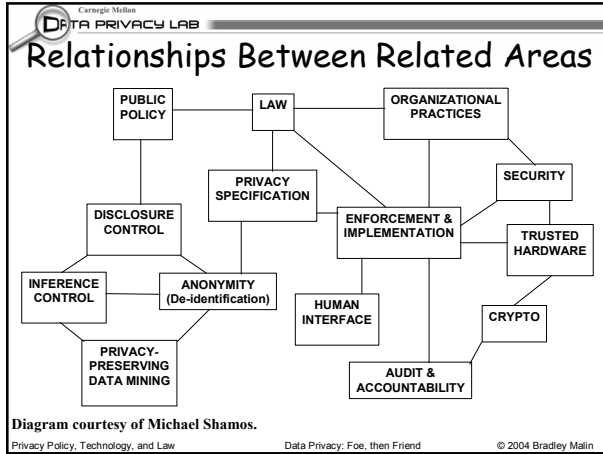
Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

What is Data Privacy?

- The study of computational solutions for releasing data such that the data remain practically useful while the aspects of the subjects of the data are not revealed.
- Privacy Protection (“data protectors”):
 - release information such that entity-specific properties (e.g. identity) are controlled
 - restrict what can be learned
- Data Linkage (“data detectives”)
 - combining disparate pieces of entity-specific information to learn more about an entity

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin



DATA PRIVACY LAB

Data. Data. Data.

- What kind of data? *All kinds!*
 - Field Structured Databases
 - Text Documents
 - Genomic
 - Image
 - Video
 - Network (Physical or Social)
 - Communications

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

DATA PRIVACY LAB

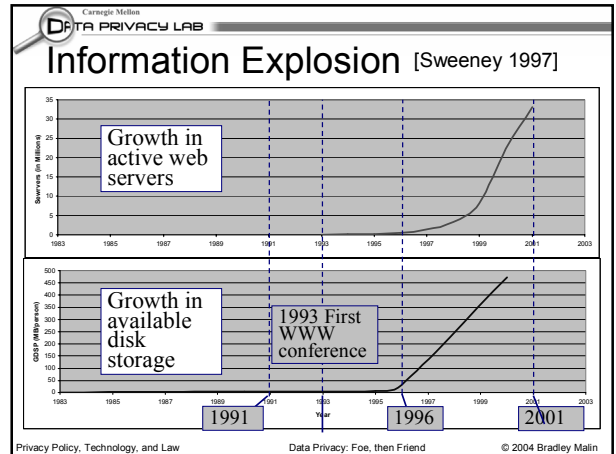
Data Privacy is Interdisciplinary

Table courtesy of Latanya Sweeney.

	<i>anonymity</i>	<i>rights mgt</i>	<i>database</i>	<i>ubiquitous</i>
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

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin



Carnegie Mellon
DATA PRIVACY LAB

Anonymity & De-identification

- **Anonymous:** Data can not be “manipulated” or linked to identify an individual
- **De-identified:** All explicit identifiers, such as name, address, & phone number are removed, generalized, or replaced with made up values
- Does Anonymous = De-identified?

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

Direct Linkage

- Uses the combination of attributes to determine the uniqueness of an entity in a dataset
- Second dataset with identified subjects is used to make the re-identification by drawing inferences between the two datasets on the related attributes
- The attributes do not have to be equal, but there must exist some ability for inference of between attributes.

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

Medical Information

- Attributes Recommended by NAHDO (outside scope of HIPAA)

<input type="checkbox"/> Patient Zip Code	<input type="checkbox"/> Principle Diagnosis Codes (ICD-9)
<input type="checkbox"/> Patient Birth Date	<input type="checkbox"/> Procedure Codes
<input type="checkbox"/> Patient Gender	<input type="checkbox"/> Physician ID Number
<input type="checkbox"/> Patient Racial Background	<input type="checkbox"/> Physician Zip Code
<input type="checkbox"/> Patient Number	<input type="checkbox"/> Total Charges
<input type="checkbox"/> Visit Date	

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

Linking to re-identify data

Medical Data

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

Linking to re-identify data

A circular diagram representing a Voter List with the following fields: Name, Address, Zip, Date registered, Birthdate, Party affiliation, Sex, and Date last voted.

Voter List

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

{date of birth, gender, 5-digit ZIP} uniquely identifies 87.1% of USA [Sweeney 97, 98]

A scatter plot showing the percentage of the population identifiable (y-axis, 0 to 1.2) versus ZIP Population (x-axis, 0 to 120,000). A horizontal line is drawn at 1.0. A text box notes: "Few fields are needed to uniquely identify individuals."

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

Linking to re-identify data [Sweeney 97, 98]

A Venn diagram with two overlapping circles: Medical Data (left) and Voter List (right). The intersection is shaded black and contains the text: "87% of the United States is RE-IDENTIFIABLE".

Medical Data **Voter List**

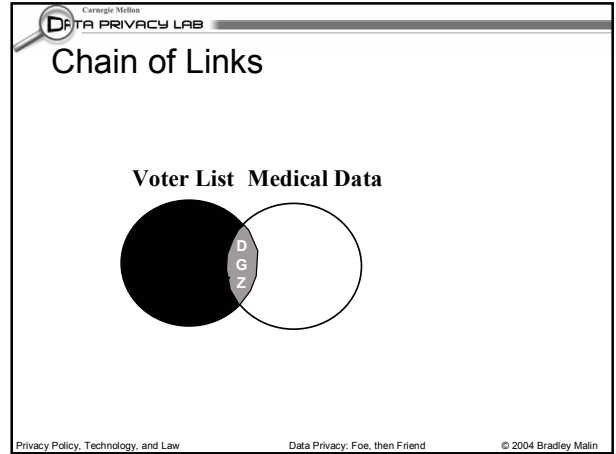
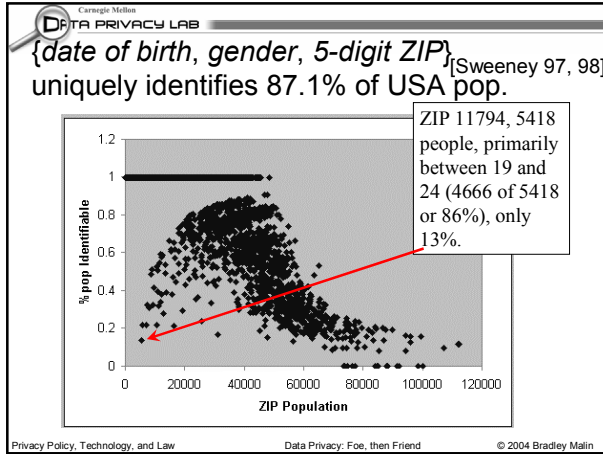
Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

{date of birth, gender, 5-digit ZIP} uniquely identifies 87.1% of USA [Sweeney 97, 98]

A scatter plot showing the percentage of the population identifiable (y-axis, 0 to 1.2) versus ZIP Population (x-axis, 0 to 120,000). A horizontal line is drawn at 1.0. A callout box points to a data point at approximately 112,167 people, stating: "ZIP 60623, 112,167 people, 11%, not 0% insufficient # above the age of 55 living there."

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin



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.

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

DATA PRIVACY LAB

Chain of Links

Voter List Medical Data

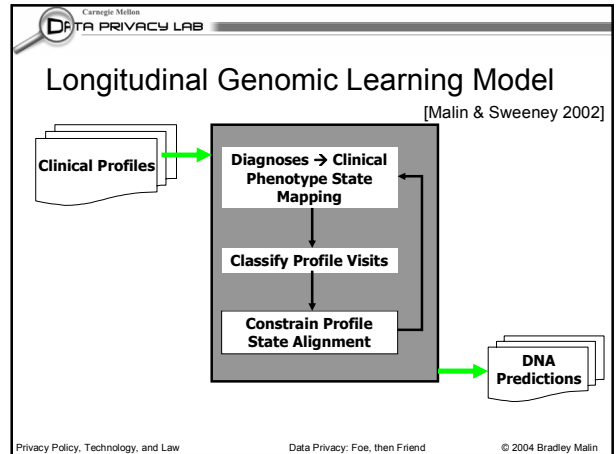
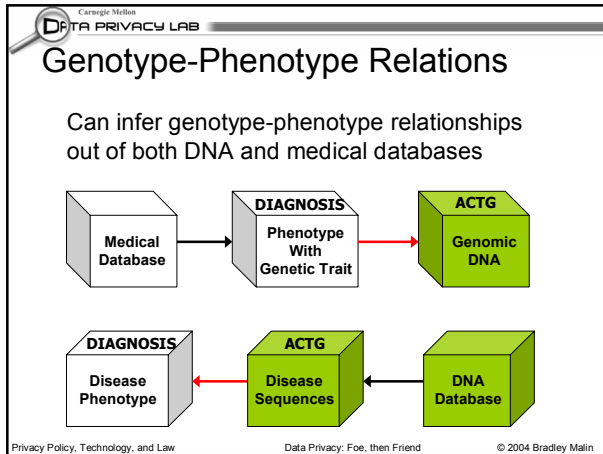
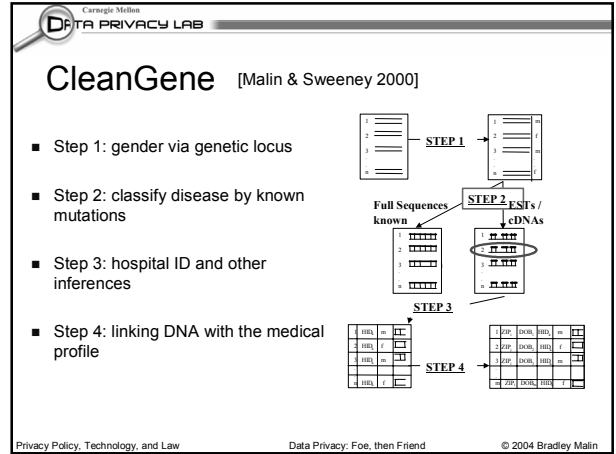
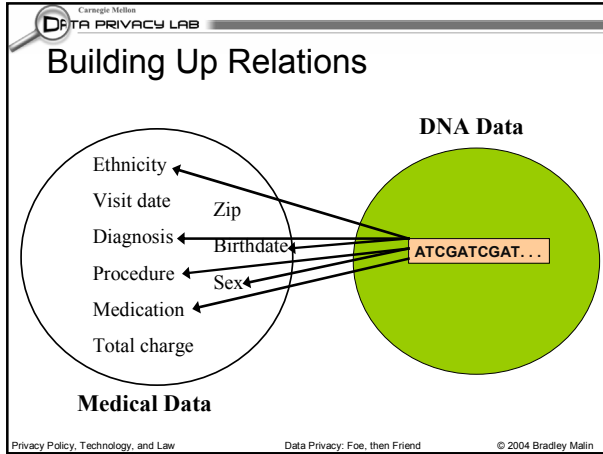
DNA Data

- Mutation Analysis
- Prediction and Risk
- Pharmaco-Genomic Relations
- Familial Relations

ATCGATCGAT...

So what do you do?

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin



DATA PRIVACY LAB

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

$$y = -21.048L(x) + 122.66$$

$$R^2 = 0.8809$$

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.

DATA PRIVACY LAB

Websites Share Weblogs

128.2.65.781 Yoda
134.6.8.91 Luke
34.1.687.21 Leah
82.912.32.1 Obi

81.2.1.541 Han
134.6.8.91 Luke
82.912.32.1 Obi

128.2.65.781 Yoda
134.6.8.91 Luke
322.46.7.93 C3PO
12.78.96.54 Jabba

322.46.7.93 C3PO
82.912.32.1 Obi
34.1.687.21 Leah
51.3.5.677 Lando

Data Privacy: Friend

DATA PRIVACY LAB

Learning from Trails

[Malin & Sweeney '01, '04], [Malin '02]

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

Privacy Policy, Technology, and Law

DATA PRIVACY LAB

Websites Share Consumer Lists

128.2.65.781 Yoda
134.6.8.91 Luke
34.1.687.21 Leah
82.912.32.1 Obi

81.2.1.541 Han
134.6.8.91 Luke
82.912.32.1 Obi

128.2.65.781 Yoda
134.6.8.91 Luke
322.46.7.93 C3PO
12.78.96.54 Jabba

322.46.7.93 C3PO
82.912.32.1 Obi
34.1.687.21 Leah
51.3.5.677 Lando

Data Privacy: Friend

DATA PRIVACY LAB

REIDIT-I Example

Identity	eb*	amazon.com	BARBIZ	BARNEYS	NOBLE
	1	0	1	1	
	1	0	1	0	
	0	1	0	1	
	0	0	1	1	

IP	eb*	amazon.com	BARBIZ	BARNEYS	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

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

DATA PRIVACY LAB

Protection Models

	Jcd	cardiac
	Jwq	cancer
	Jxy	liver

Null-Map

	Al	3/8/61	02138	cardiac
	Ann	10/2/61	02139	cancer
	Abe	7/14/61	02139	liver

Wrong-Map

	A*	1961	0213*	cardiac
	A* <td>1961</td> <td>0213*</td> <td>cancer</td>	1961	0213*	cancer
	A* <td>1961</td> <td>0213*</td> <td>liver</td>	1961	0213*	liver

k-Anonymity

Private Information

	Ann	10/2/61	02139	cardiac
	Abe	7/14/61	02139	cancer
	Al	3/8/61	02138	liver

Universe

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

DATA PRIVACY LAB

Overview

- General Idea of Data Privacy
- Data Analysis in Personal Information Learning
 - Demographic Data
 - Genetic Data
- Data Protection
 - Formal Models
 - Video Data

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

DATA PRIVACY LAB

Idea of k -map and k -anonymity [Sweeney 97, 98]

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.

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

DATA PRIVACY LAB

Model Examples

k-map: For each tuple t in the release,
 t must indistinctly refer to at least k entities in the population

A*	1963	0213*	cardiac
A*	1961	0213*	cancer
A*	1964	0213*	liver

k-anonymity: "k in the release"

A*	1961	0213*	cardiac
A*	1961	0213*	cancer
A*	1961	0213*	liver

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

DATA PRIVACY LAB

Distributions of a Population Register

Register

There are three colors with frequencies: 1 red, 3 green and 2 blue.
 There are 2 types of figures, with 2 of one type and 4 of the other.
 The combination of color and figure labeled as Hal and Len are each unique.

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

DATA PRIVACY LAB

k-anonymity, enforce on release

- Quasi-identifier, profile {Birth, ZIP, Gender}
- Generalization 10/27/59 ↷ 1959
- Suppression 02139 ↷ ■■■■
- Encryption 3245123 ↷ 2168582

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

DATA PRIVACY LAB

Re-identification Example

Register

Release

There are 3 green figures and 2 figures having the same profile as the release.
 But only Hal is green and has the same figure type as the profile in the release. It is a unique match.

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

Re-identification Example

There are two matches for this profile, Jim and Mel. There is no unique match.

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

Overview

- General Idea of Data Privacy
- Data Analysis in Personal Information Learning
 - Demographic Data
 - Genetic Data
- Data Protection
 - Formal Models
 - Video Data

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

Formal Protection Example

To achieve k -map where $k=2$, agents for *Gil*, *Hal* and *Ken* agree to merge their information together. Information released about any of them results in the same merged image.

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

Video Data Privacy

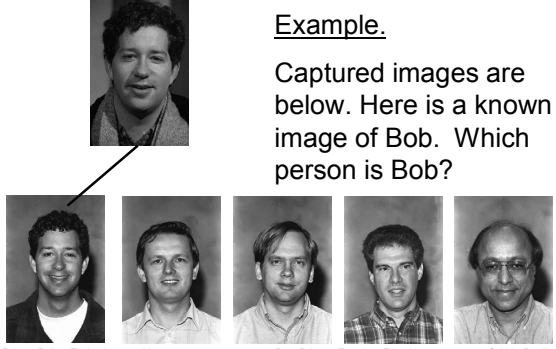
- Modify video images so that
 - Privacy: automated attempts to recognize faces fail
 - Utility: knowledge learned from data is useful
- Solution to problem
 - Enables sharing of data for specified purposes
 - Protects rights as specified in policy
 - e.g. your identity won't be revealed unless you have done something illegal

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

DATA PRIVACY LAB

De-identification of Faces

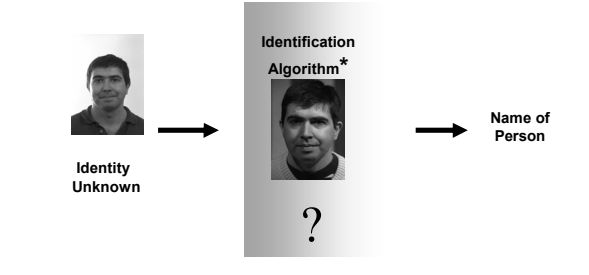
Example.
 Captured images are below. Here is a known image of Bob. Which person is Bob?



Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

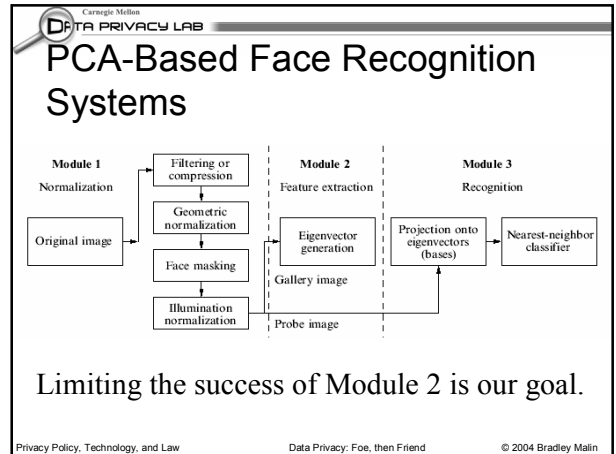
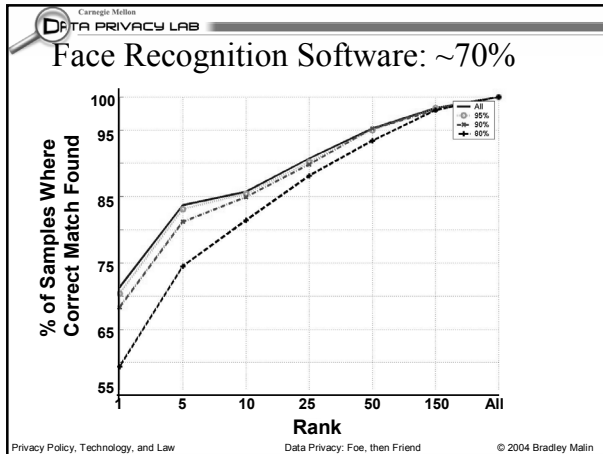
DATA PRIVACY LAB

Face Recognition: The Big Idea



Identity Unknown → Identification Algorithm* → Name of Person

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin



DATA PRIVACY LAB

Basic Approach in Eigenfaces

1. Use a training set to identify a set of characteristic faces.
2. Given a gallery of known faces and a probe image of an unknown person, compare each face to the characteristic faces to get a distance measure for each.
3. The probe's identity is determined by the shortest distance to a gallery image.
4. There is one image per person in the gallery and one corresponding picture per person in the probe set.

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

DATA PRIVACY LAB

Eigenvectors

- The characteristic function:

$$|(A - \lambda I)| = 0$$


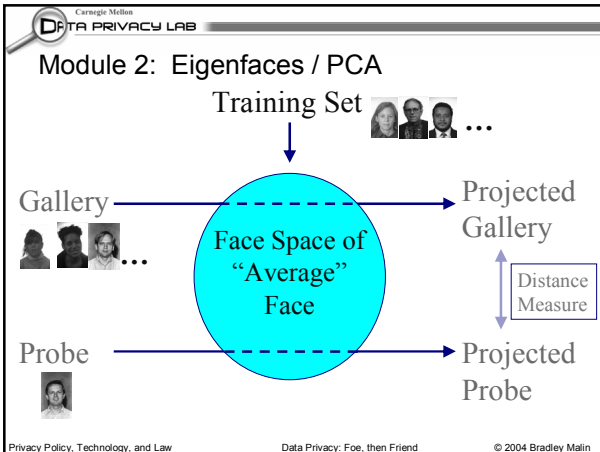


Figure 1. The first 16 eigenfaces of Ensemble 1 ($T = 1038$, $V = 3840$) and 8 of the rest.

where A is the covariance matrix C

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin




DATA PRIVACY LAB

De-identification: T-mask

Example continued...

Captured images are de-identified below. Here is a known image of Bob. Which person is Bob?




Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

De-identification: T-mask

Example continued...

Captured images are de-identified below. Here is a known image of Bob. Which person is Bob?




Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

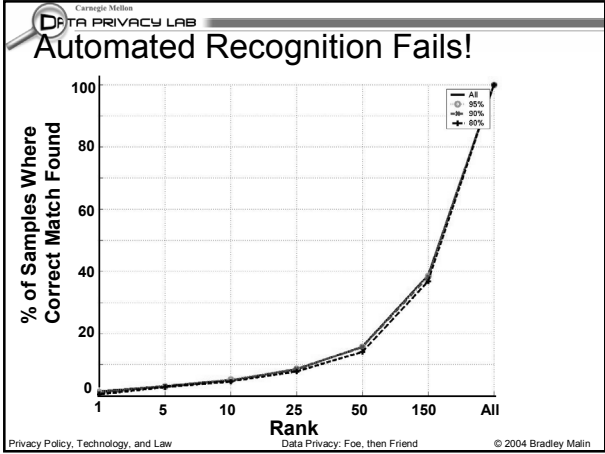
De-identification: pixel reduction

Example continued...

Captured images are de-identified below. Here is a known image of Bob. Which person is Bob?



Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin




Carnegie Mellon
DATA PRIVACY LAB

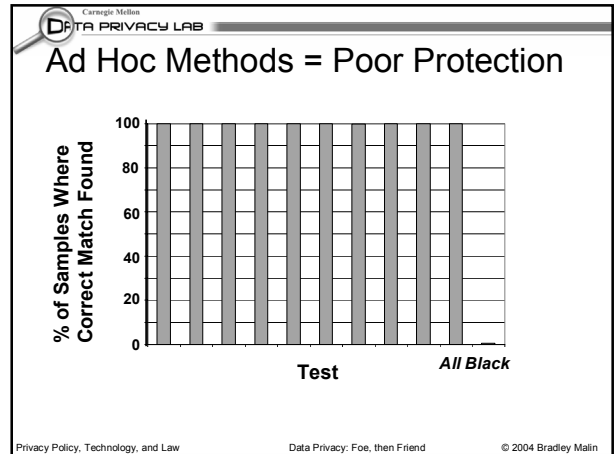
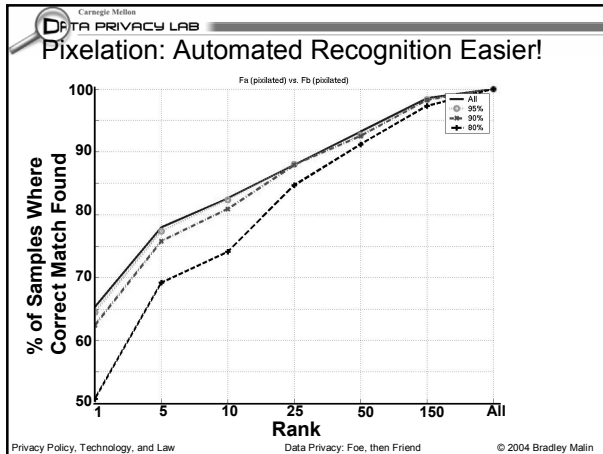
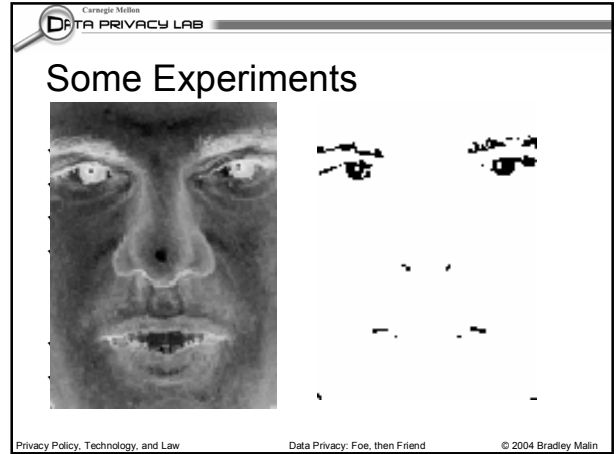
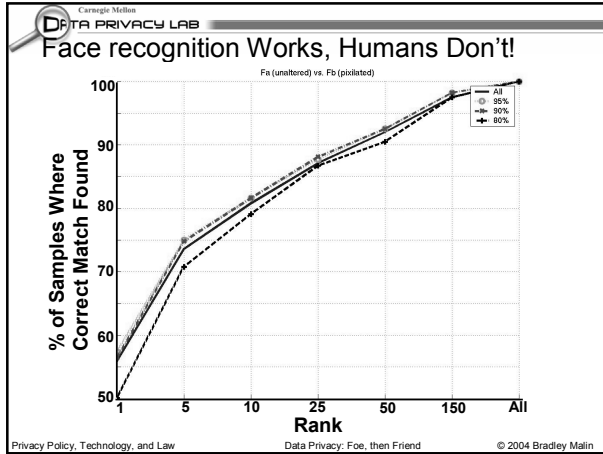
De-identification: pixel reduction

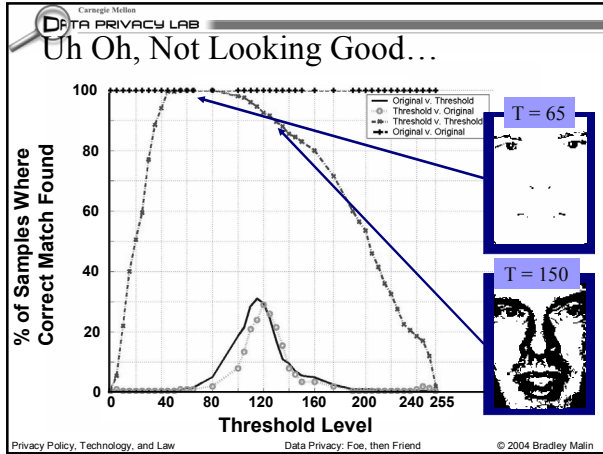
Example continued...

Captured images are de-identified below. Here is a known image of Bob. Which person is Bob?



Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin



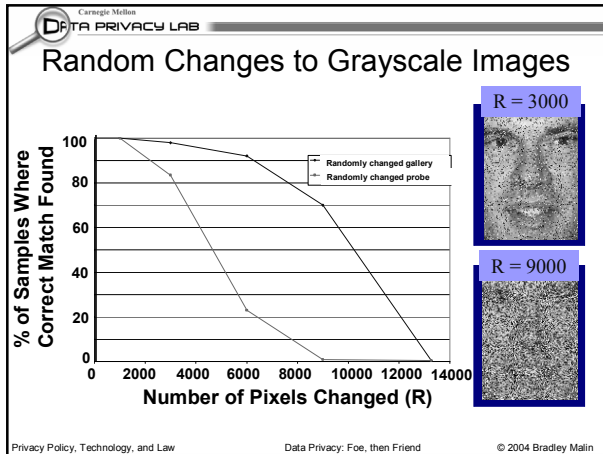


DATA PRIVACY LAB

Don't be Naïve

- Again, de-identified \neq anonymous
- Masks can be removed and trained against
- Some cases naïve de-identification even harms privacy!
 - pixelation and "blur" improves performance
- Time to get logical...

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin



DATA PRIVACY LAB

Back to k-Protection Models

[Newton, Sweeney, Malin 04,05]

- k-Anonymity: For every record, there are at least k individuals to whom it refers (realized upon release).

- k-Same: For every face, there are at least k people to whom that face refers. No face actually refers to a single real person.

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

Ranking of Faces

Example.

How does everyone rank against each other? Who is closest? Who is farthest?

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

Examples of k-Same...

-Pixel

-Eigen

k = 2 3 5 10 50 100

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

Example of k-Same for k=2

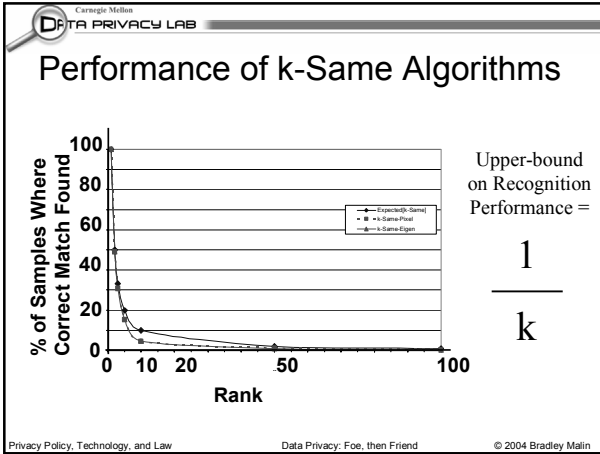
Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

Demonstration Time!

- K-Same Demo

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin



Carnegie Mellon
DATA PRIVACY LAB

Thanks!

- Some slides adapted from:
 - Elaine Newton
 - Michael Shamos
 - Latanya Sweeney
- More information:
 - <http://privacy.cs.cmu.edu>
 - <http://www.cs.cmu.edu/~malin>

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin

Carnegie Mellon
DATA PRIVACY LAB

Overview

- General Idea of Data Privacy
- Data Analysis in Personal Information Learning
 - Demographic Data
 - Genetic Data
- Data Protection
 - Formal Models
 - Video Data

Privacy Policy, Technology, and Law Data Privacy: Foe, then Friend © 2004 Bradley Malin