Privacy in D.A.T.A.

Latanya Sweeney, Ph.D.

Assistant Professor of Computer Science, Technology & Policy

Carnegie Mellon University

latanya@privacy.cs.cmu.edu

http://privacy.cs.cmu.edu/

The Question in this Talk

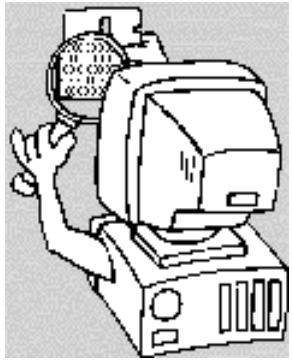
Can computer scientists provide both safety and privacy to society?

The Question in this Talk

Can computer scientists provide both safety and privacy to society?

Answer:

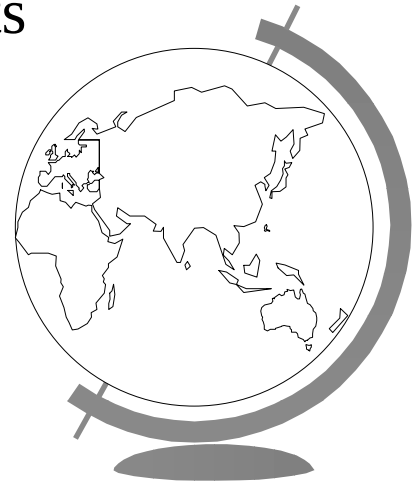
YES. Three goals: (1) understand the nature of real privacy threats; (2) design technical solutions to integrate with policy to avoid a setting in which society is forced to choose; and, (3) construct technical solutions that address these threats while keeping data useful.



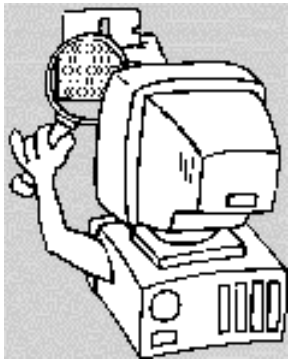
Data Privacy Laboratory at CMU

Ralph Gross
Yiheng Li
Bradley Malin
Elaine Newton
Michael Shamos
Latanya Sweeney
Ben Vernot
Aaron White

Joseph Barrett, JD
Sylvia Barrett, JD
Joseph Lombardo
Deanna Mool, JD
Julie Pavlin, MD
University of Pittsburgh Law Students



http://privacy.cs.cmu.edu/



Laboratory for International Data Privacy at CMU

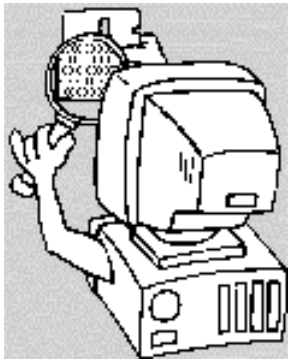
Work with real-world stakeholders:

- public health agencies
- government agencies
- private corporations

Kinds of projects currently underway:

- health data
- web data
- video surveillance data
- genetic data
- census surveys
- crime data
- grocery data, and so on…

http://privacy.cs.cmu.edu/



Laboratory for International Data Privacy at CMU

Data Linkage (“data detectives”):

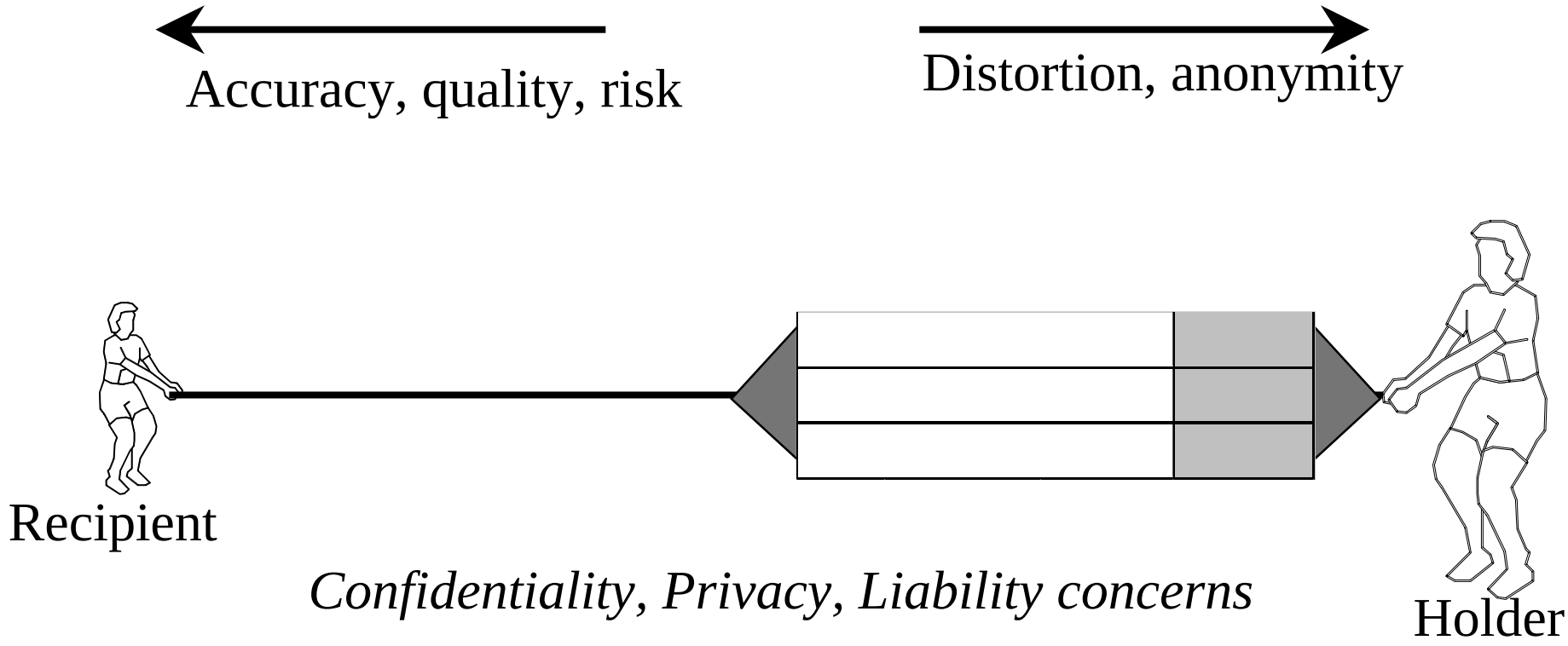
combining disparate pieces of entity-specific information to learn more about an entity

Privacy Protection (“data protectors”):

release information such that certain entity-specific properties (such as identity) cannot be inferred; restrict what can be learned

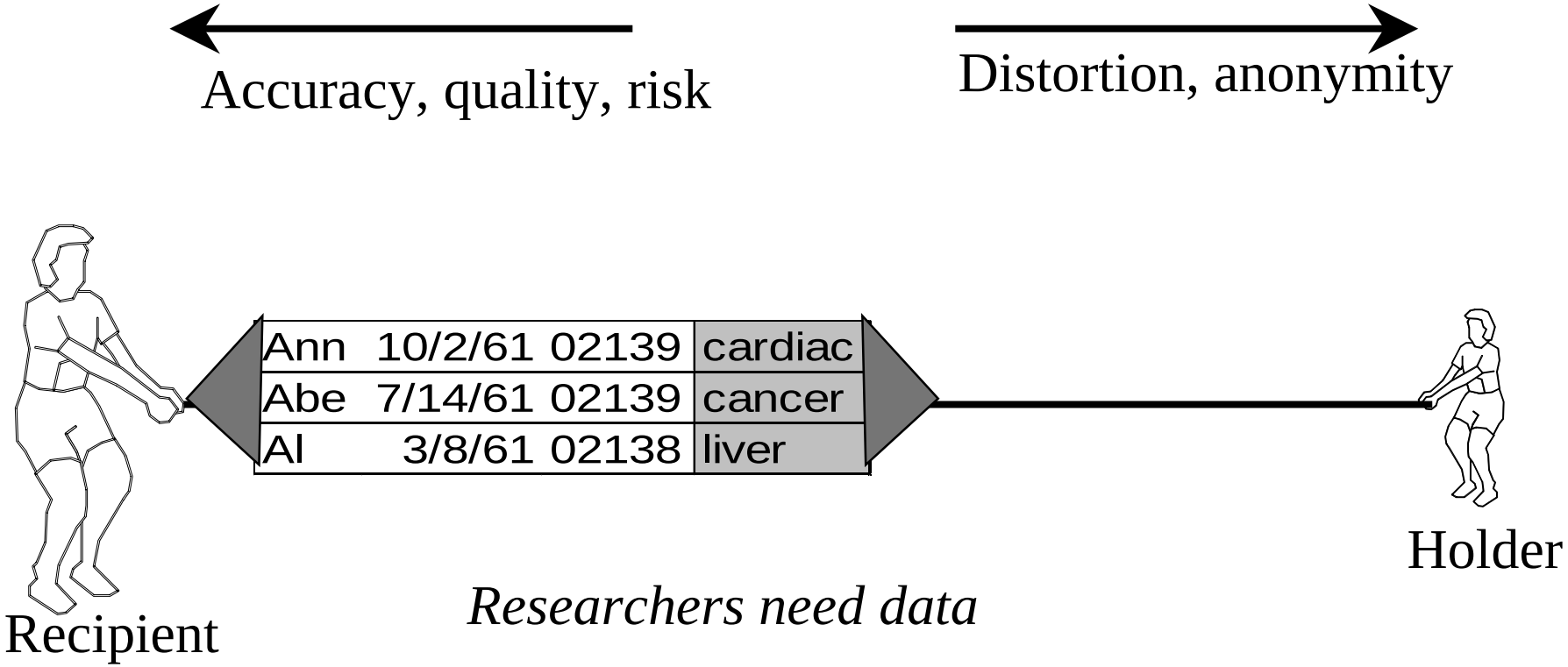
http://privacy.cs.cmu.edu/

“Can’t release data”



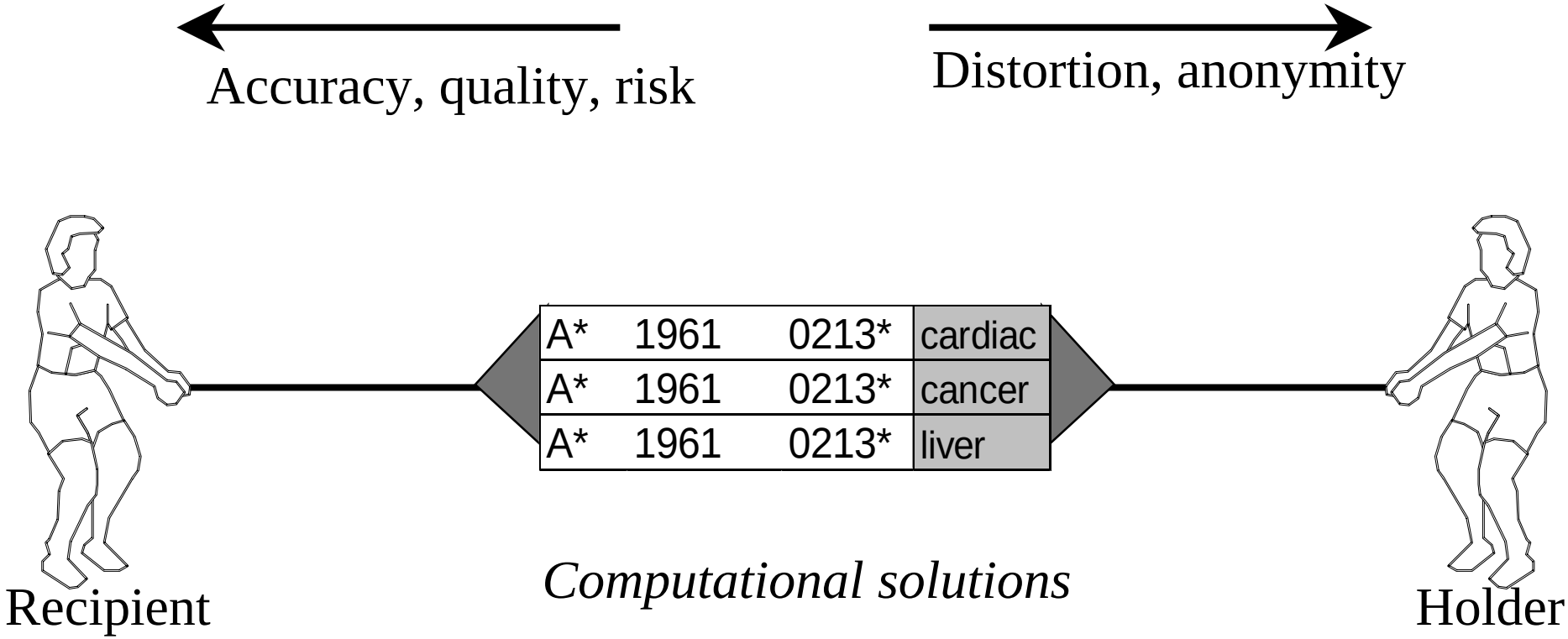
Confidentiality, Privacy, Liability concerns

“Privacy is dead, get over it”



Researchers need data

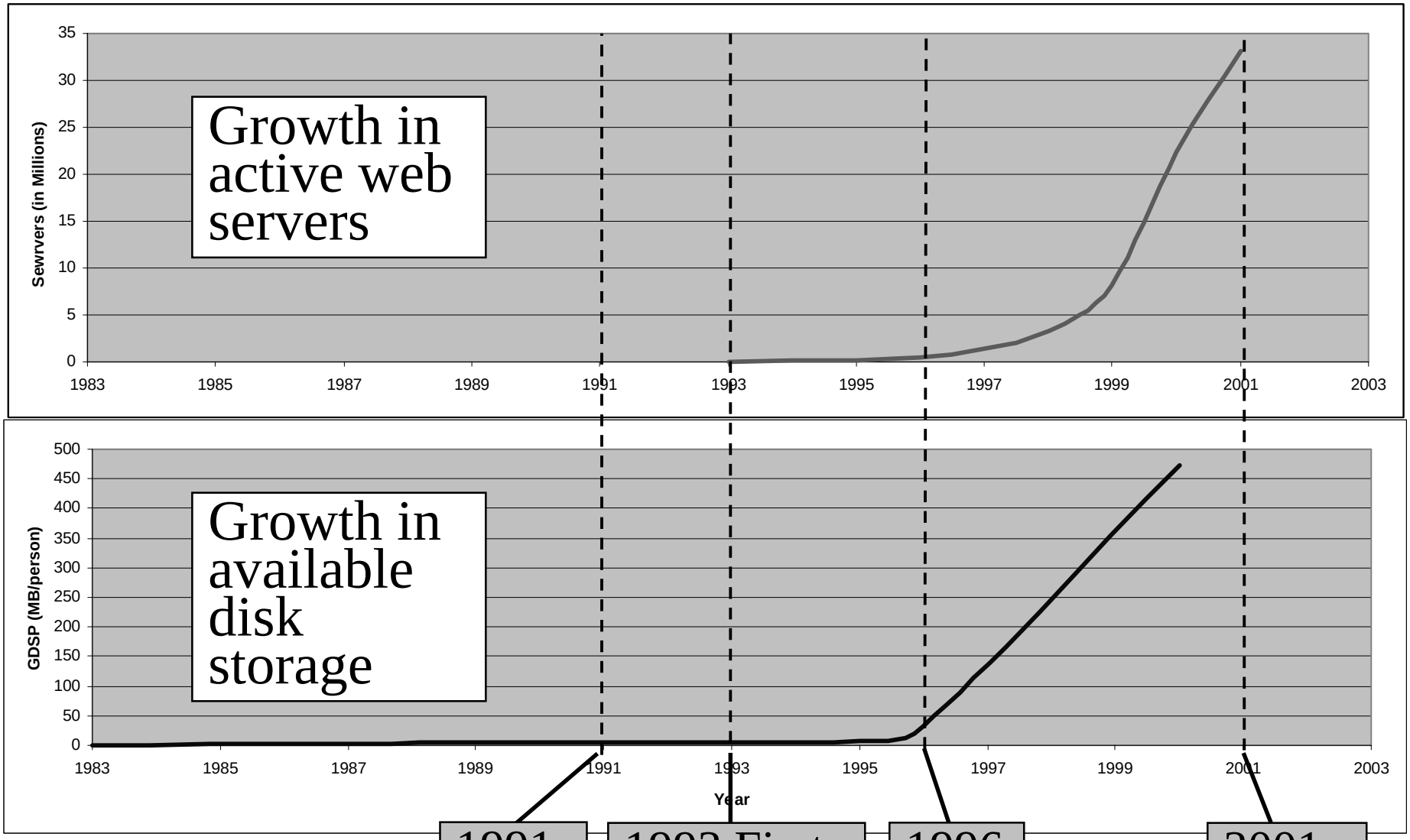
“Share data while guaranteeing anonymity”



This talk

- Data investigations
  - Lots of data out there
  - Use innocent looking data to learn sensitive information
- Data protection
- Surveillance

Technically-empowered Society



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 - *continued 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 - *continued 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 - *continued 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 - *continued 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 -*continued 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 -*continued 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 -*continued 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 -*continued 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 -*continued 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 -*continued fields 196-210*

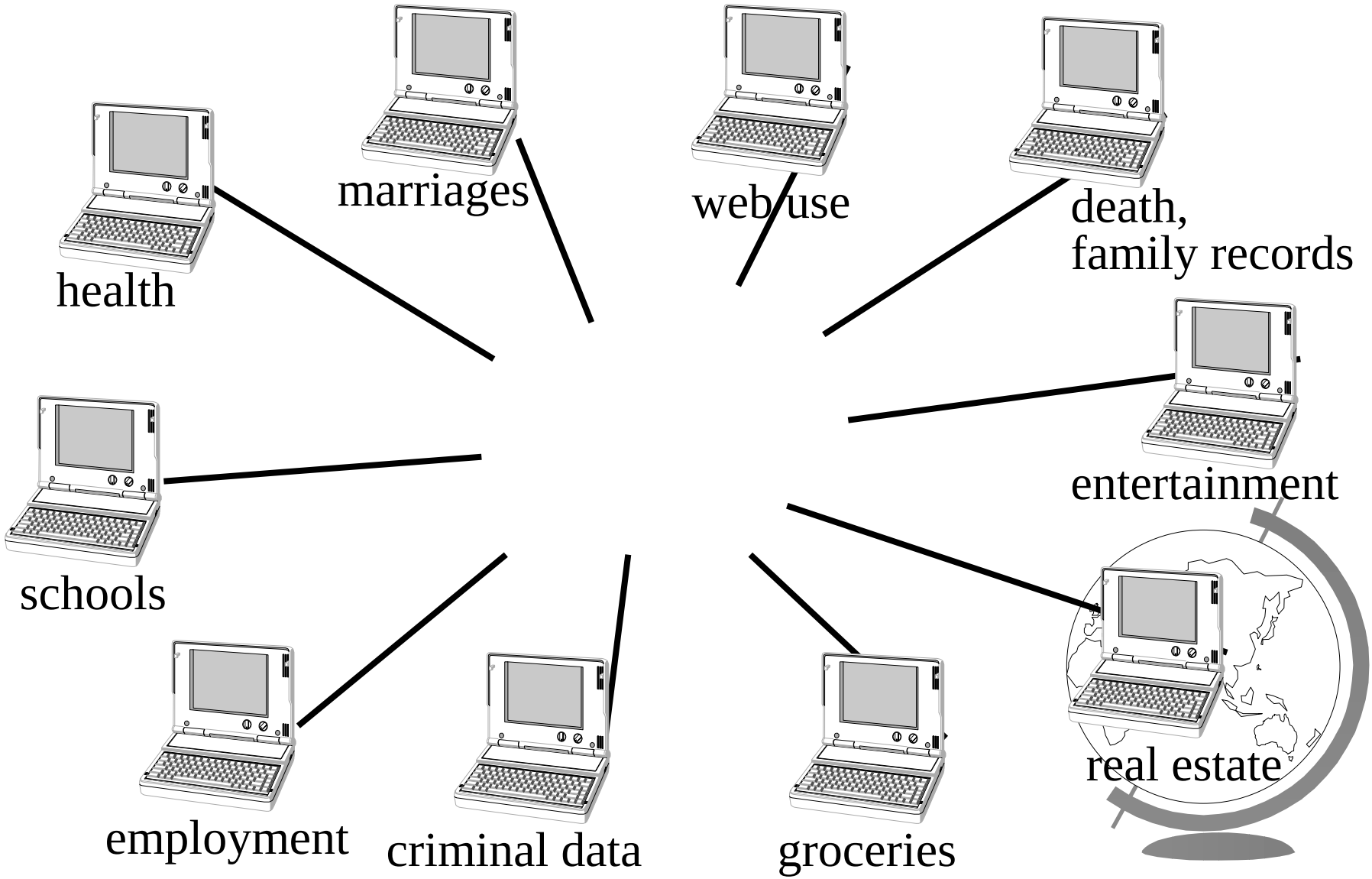
Field#	Size	Field name
196	40	Other Musculoskeletal/Integumental A
197	1	Down's Syndrome
198	40	Other Chromosomal Anomalies
199	1	No Congenital Anomalies
200	40	Other Congenital Anomalies
		CODE STRIP
201	1	Record Complete YN
202	1	Record Type
203	4	Facility ID
204	4	City of Birth
205	3	County of Birth
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 -*continued 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

On-line birth certificates (some California counties)

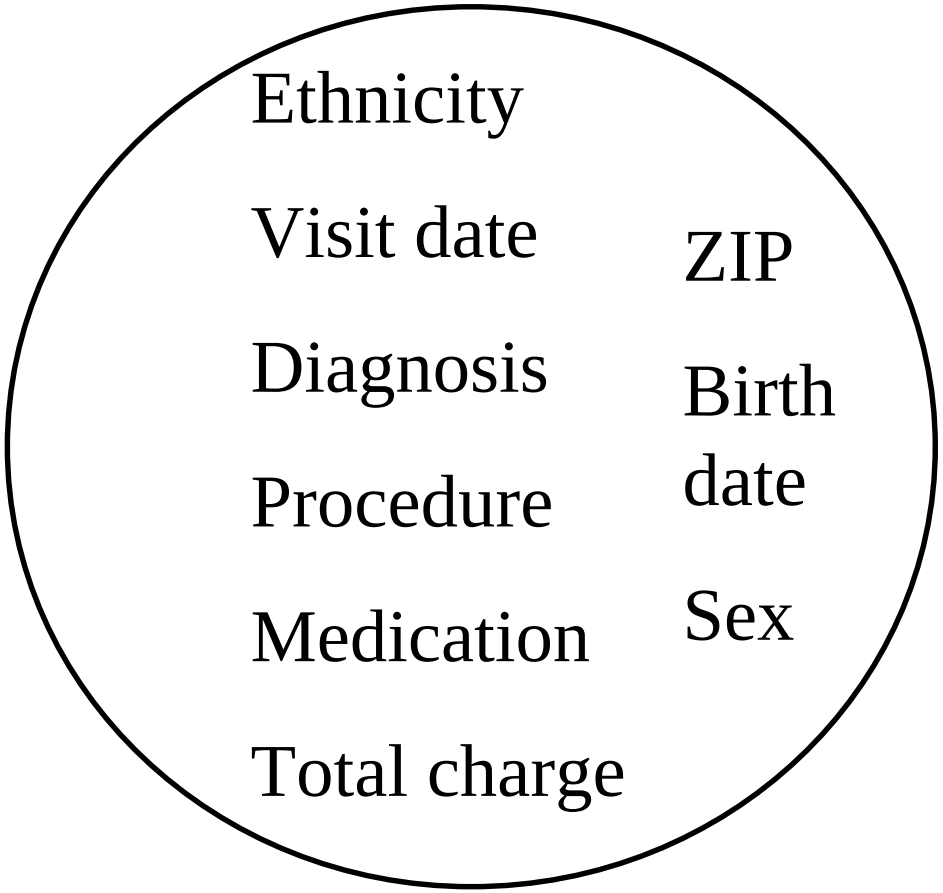
Numerous Efforts Underway to Fuse Available Data Together on Individuals



This talk

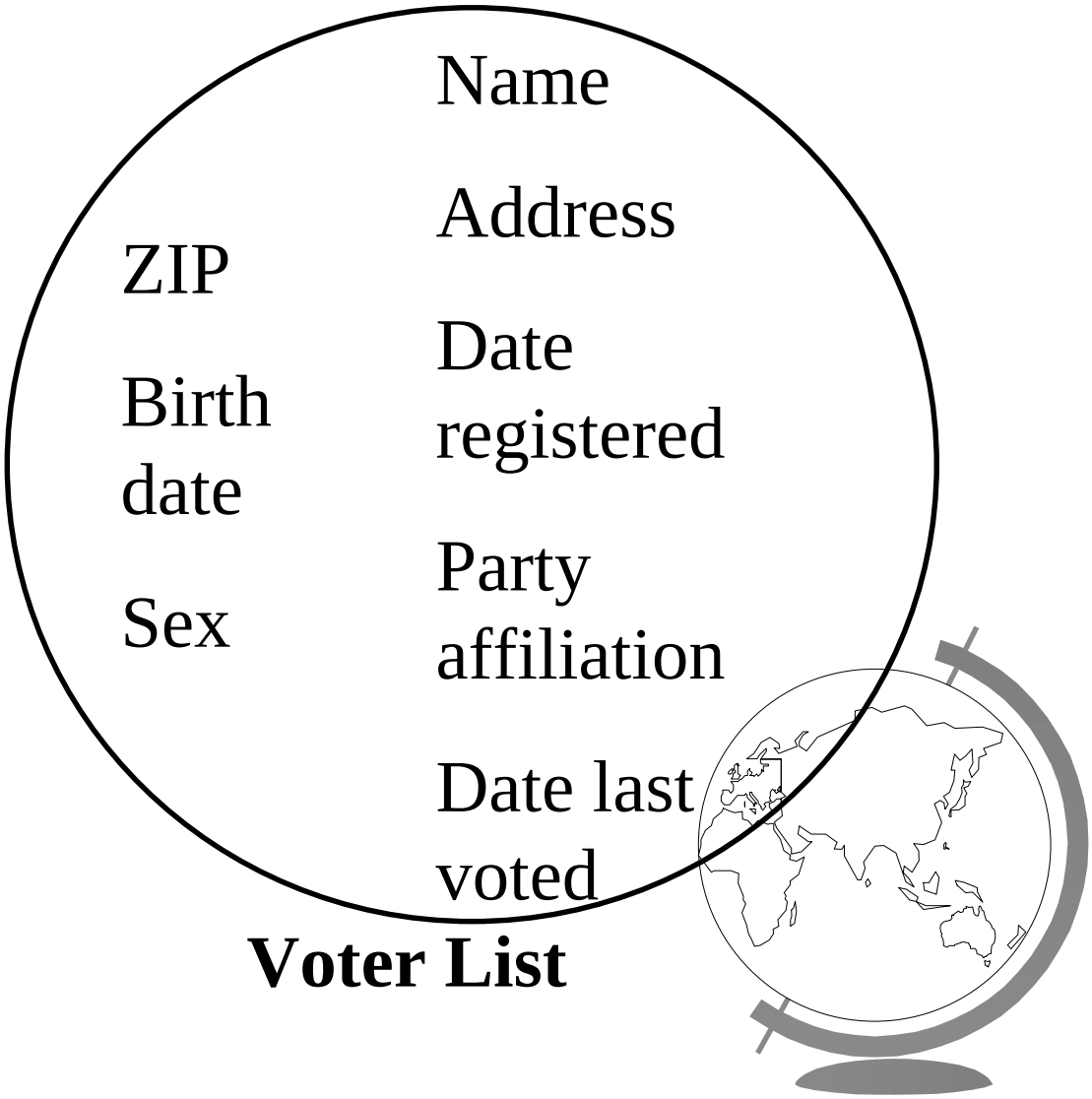
- Data investigations
  - ✓Lots of data out there
  - Use innocent looking data to learn sensitive information
- Data protection
- Surveillance

Health data (GIC example)

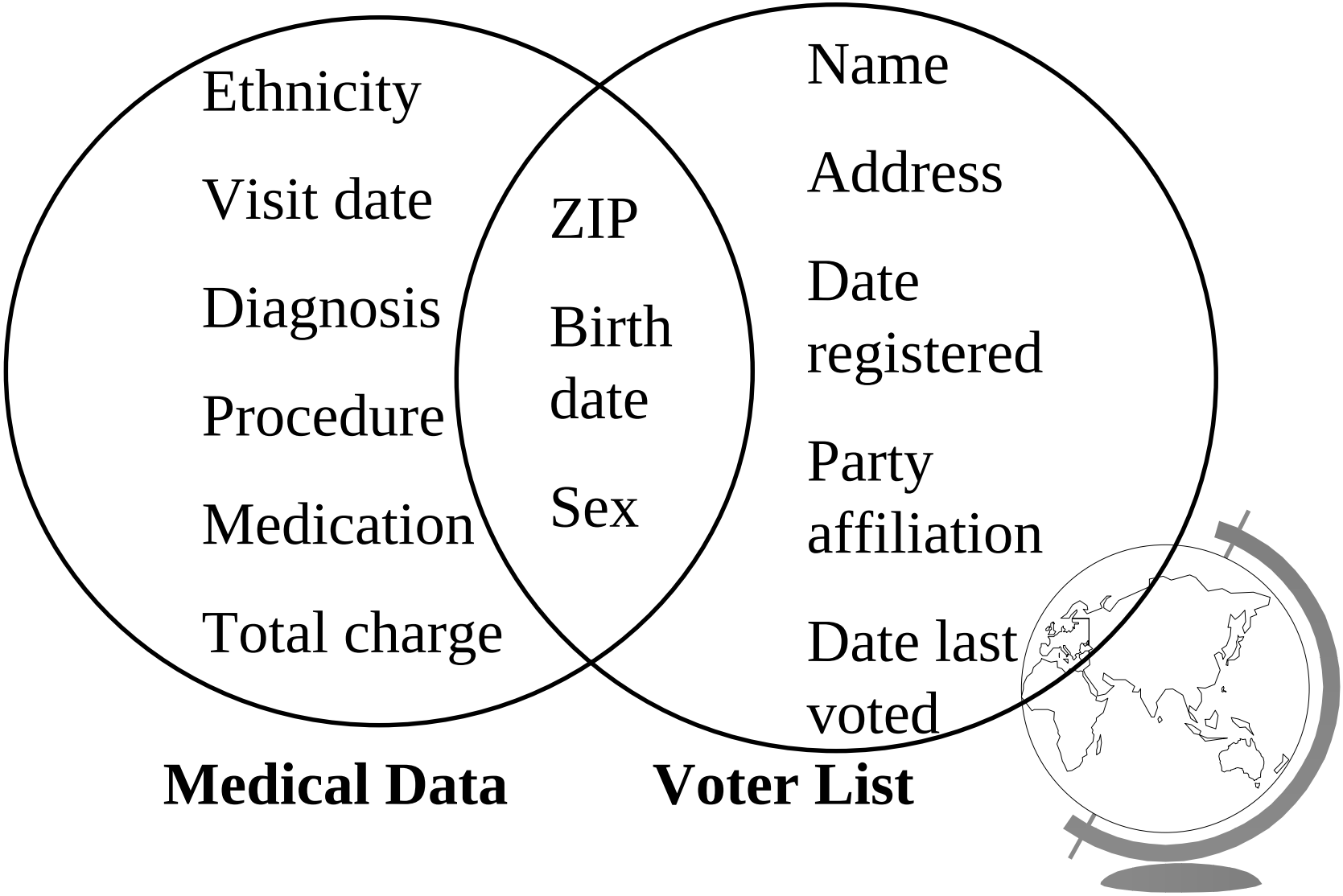


Medical Data

Population data (GIC example)



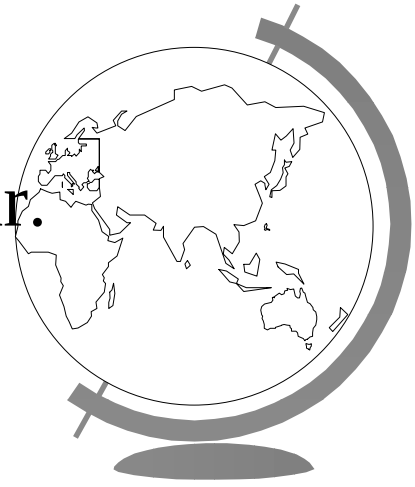
Linking to re-identify data



Uniqueness in Cambridge Voters

Birth date alone	12%
Birth date & gender	29%
Birth date & 5-digit ZIP	69%
Birth date & full postal code	97%

Birth date includes month, day and year.
Total 54,805 voters.



Few characteristics make a person unique

Birth includes month, day and year:

365 days x 100 years = 36,500 possibilities

Two genders and Five ZIP (5-digit) codes:

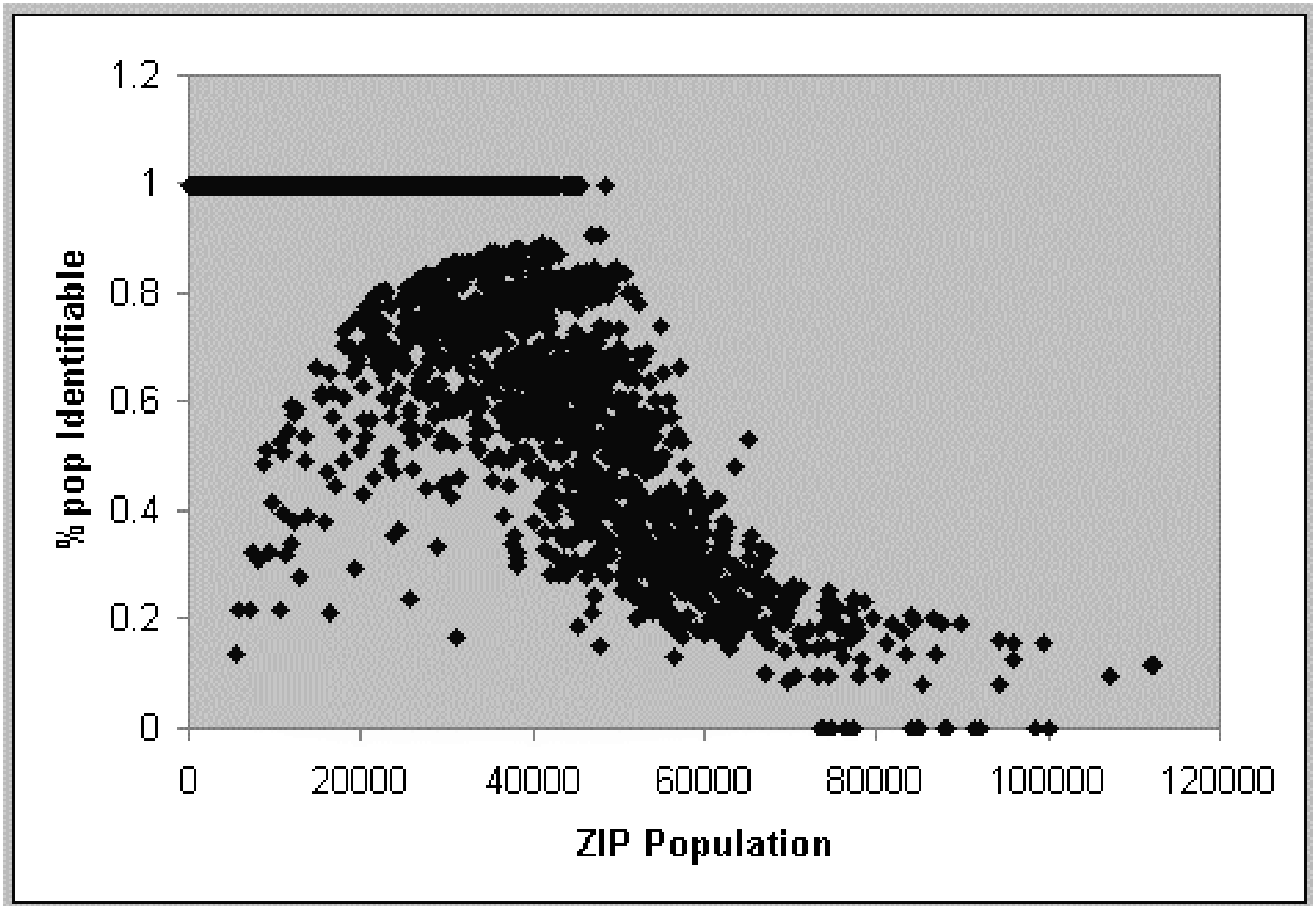
2 * 5 * 36,500 = **365,000 possibilities**

But the Cambridge Voter list had:

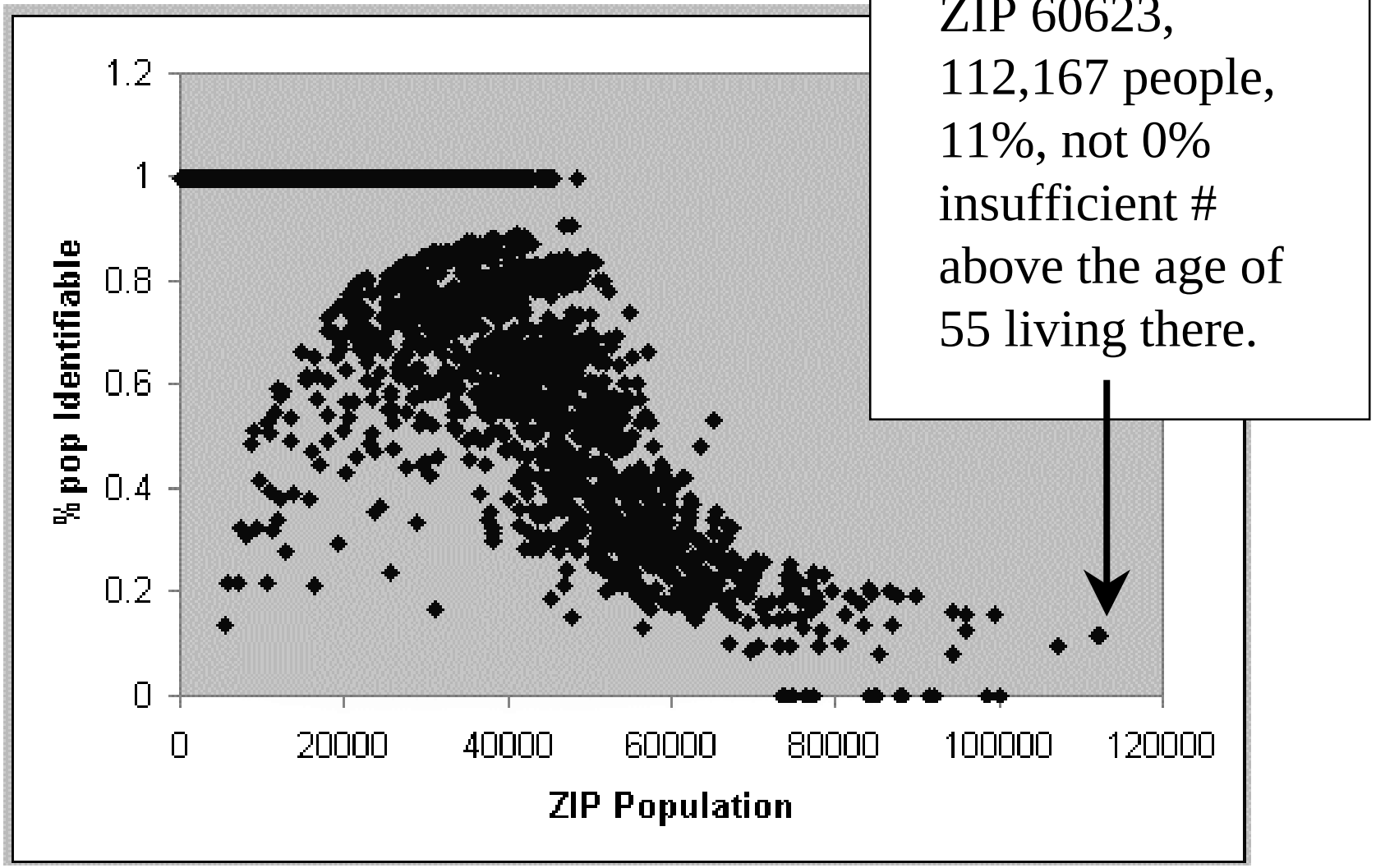
54,805 voters

So in general, using
(birth[mon,day,yr], gender, ZIP[5-digit])
provides a **unique quasi-identifier**.

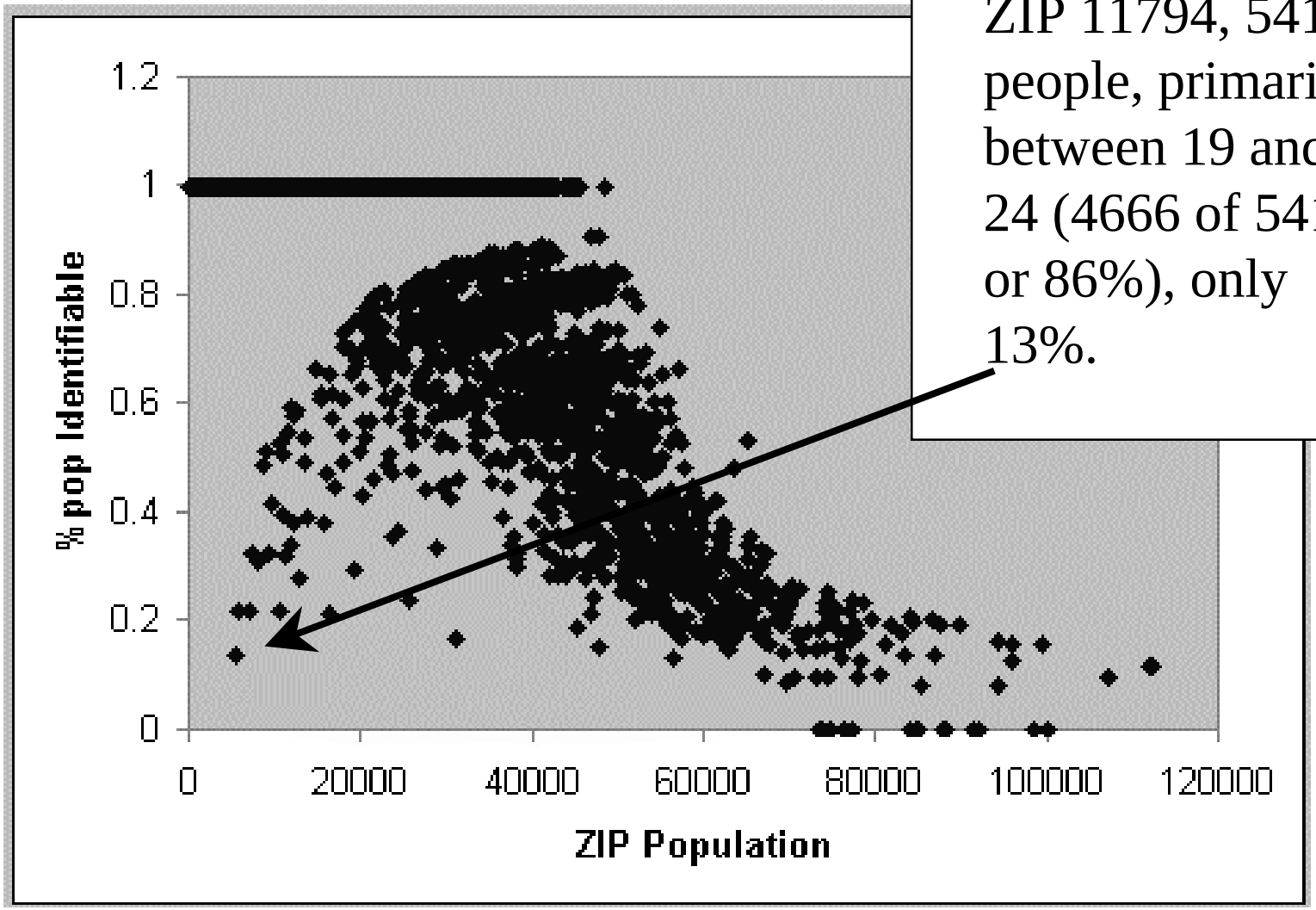
{*date of birth*, *gender*, *5-digit ZIP*}
uniquely identifies 87.1% of USA pop.



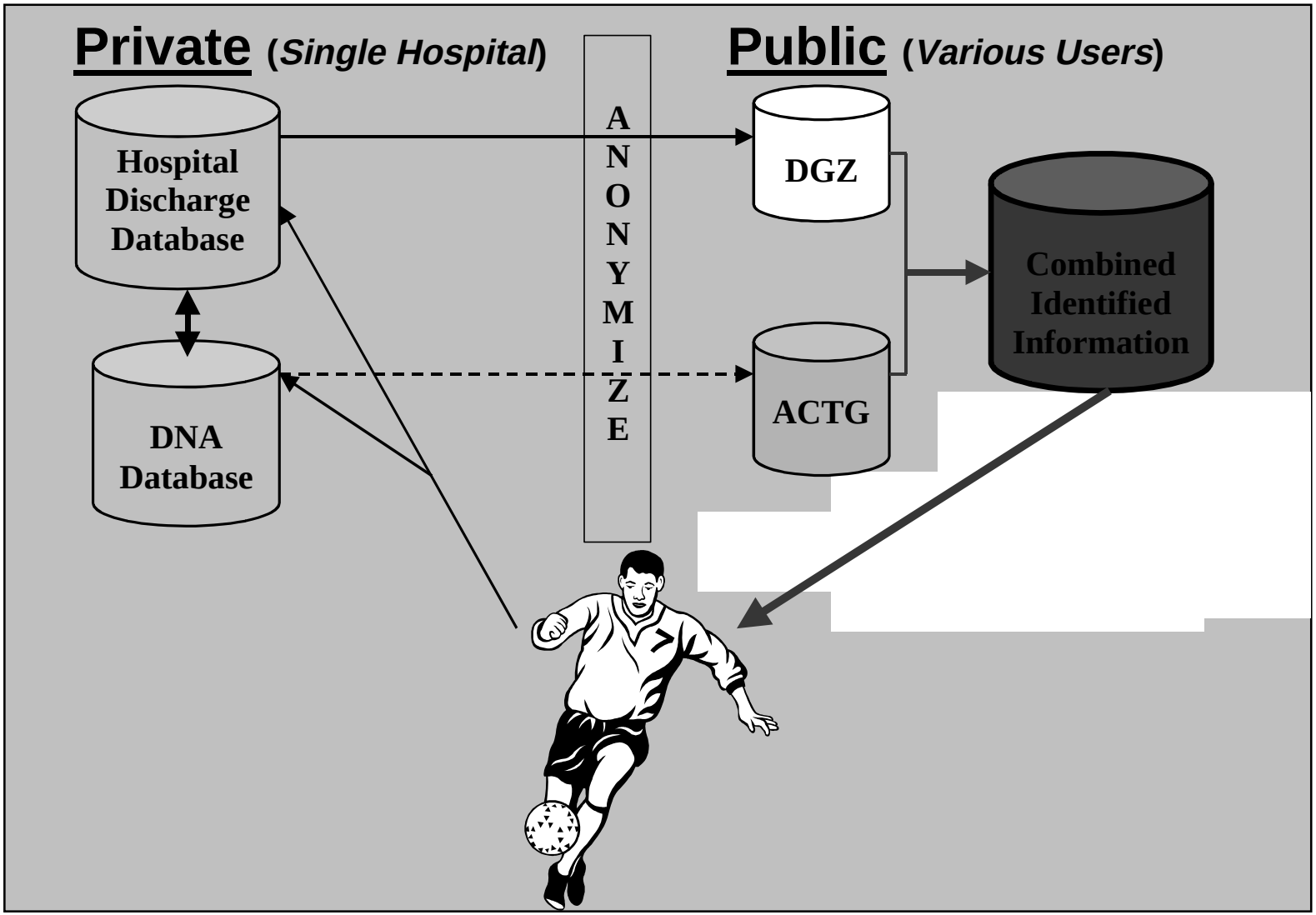
{*date of birth, gender, 5-digit ZIP*}
uniquely identifies 87.1% of USA pop.



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

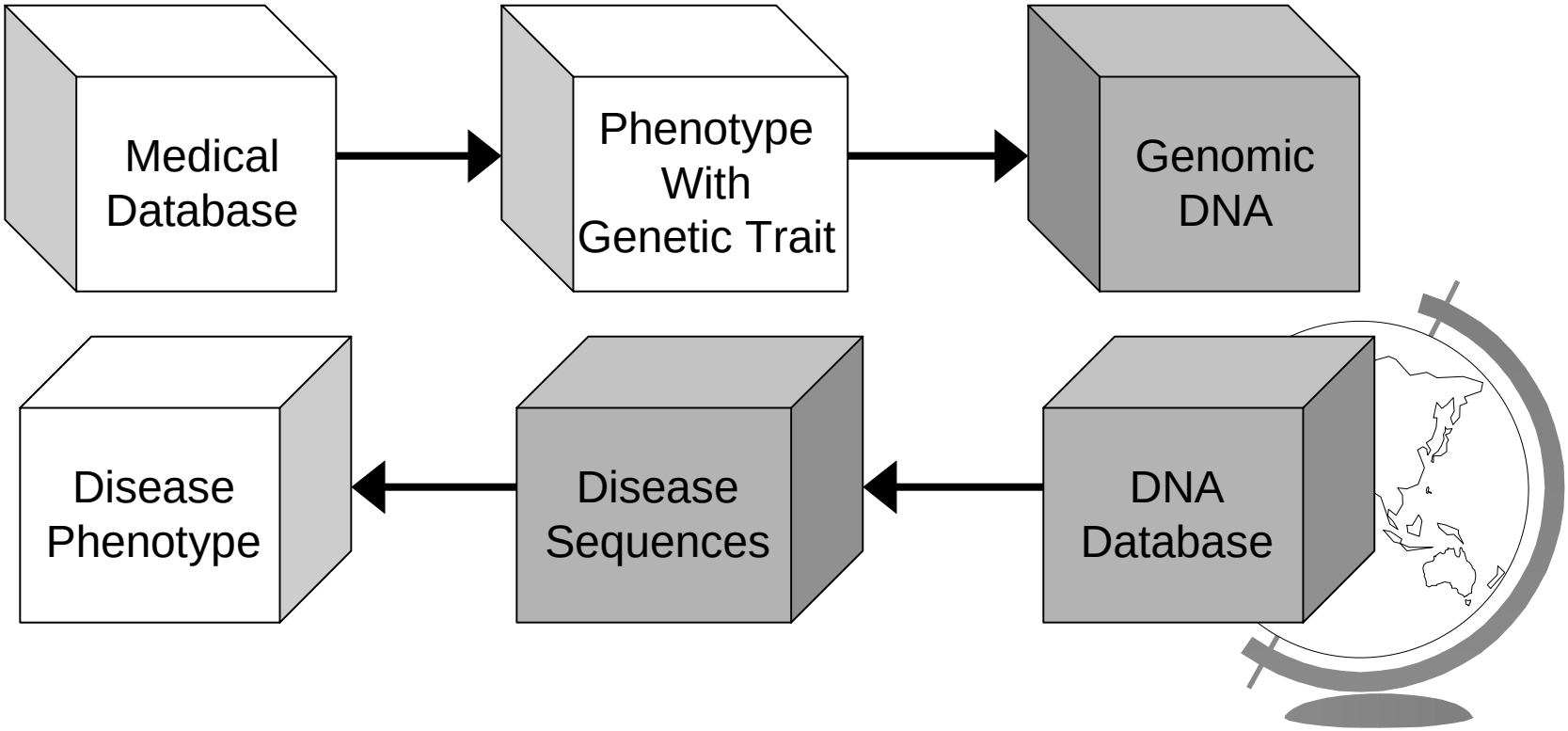


Disclosure Scenario

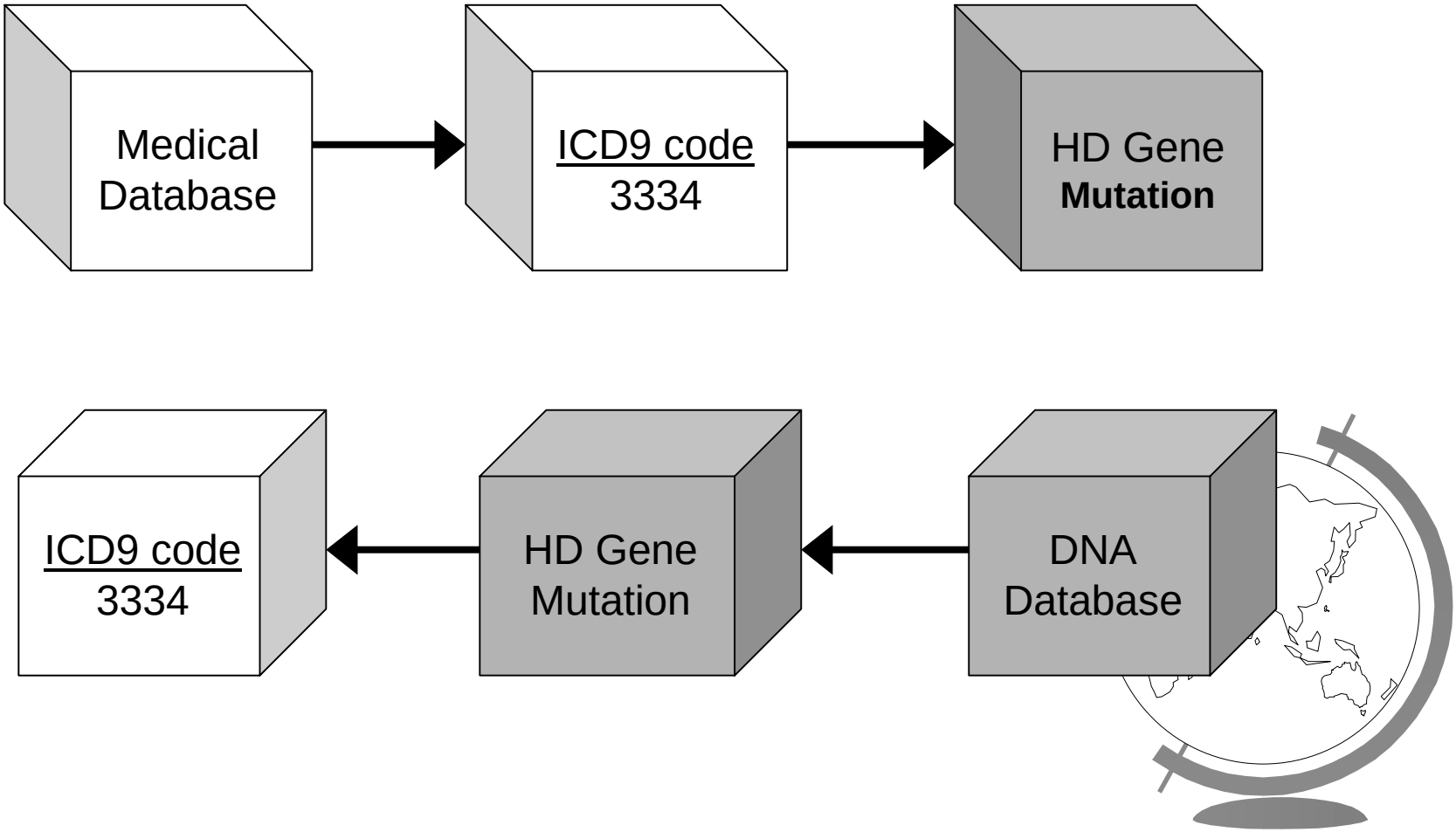


Genotype-Phenotype Relations

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



Example: Huntington's Disease



Uniqueness in Trails

logs

P1: 00000000000000000000000000000000000000000000000**1**0000000

P2: 00000000000000000000**1**0000000000000000000000000000000000

P3: 00000000000000000000**1**0000000000**1**0000000000000000000000

---

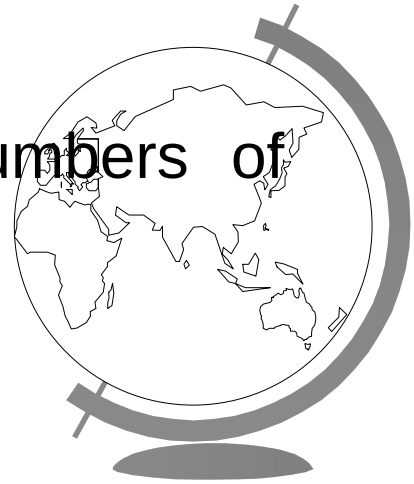
names

PA: 00000000000000000000000000000000000000000000000**1**0000000

PB: 00000000000000000000000000000**1**0000000000000000000000000

PC: 00000000000000000000**1**0000000000000000000000000000000000

Uniqueness of audit trails with large numbers of people and locations.



Uniqueness in Trails (Web logs)

logs

P1: 00000000000000000000000000000000000000000000000**1**0000000

P2: 00000000000000000000**1**0000000000000000000000000000000000

P3: 00000000000000000000**1**0000000000**1**0000000000000000000000

---

names

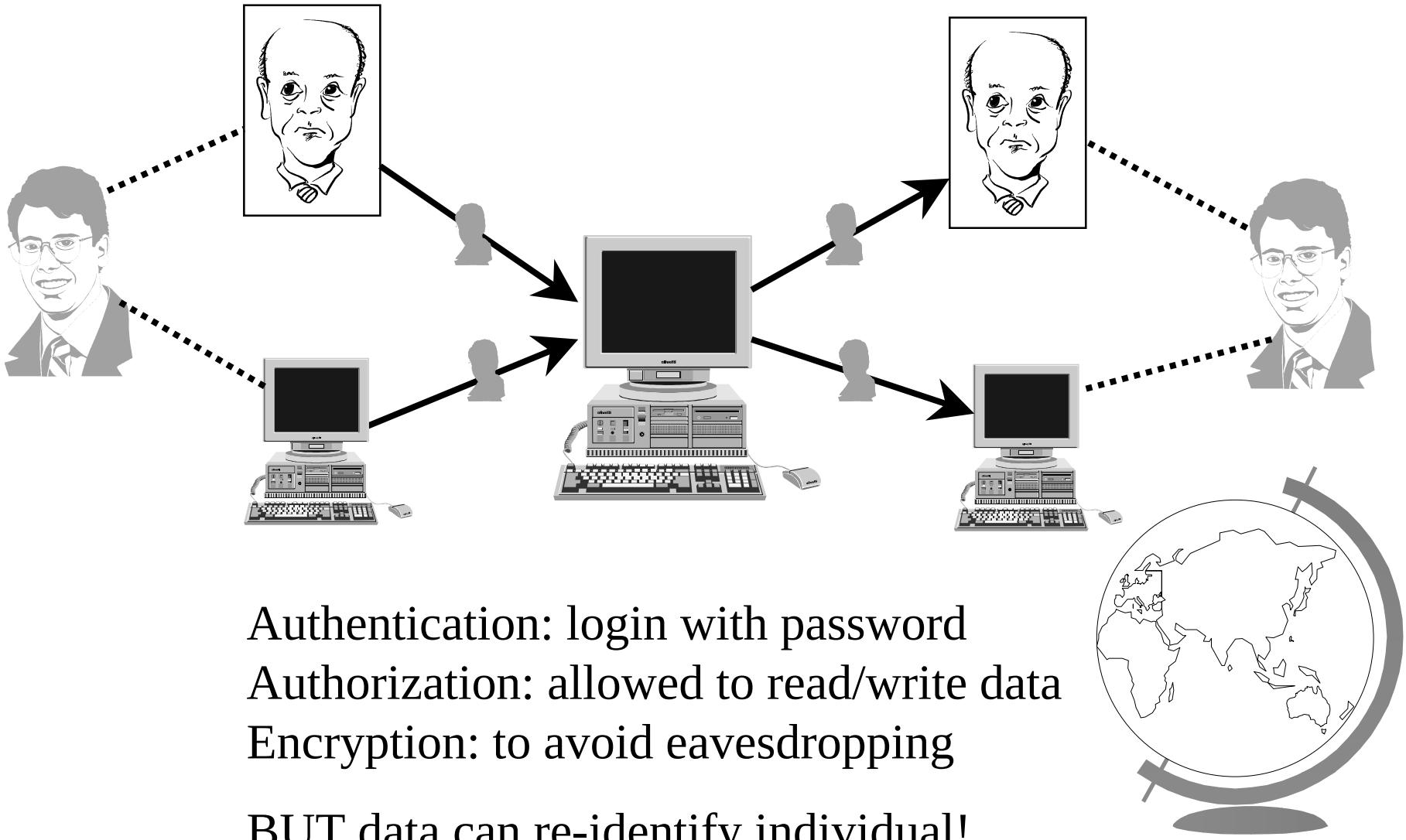
PA: 00000000000000000000000000000000000000000000000**1**0000000

PB: 00000000000000000000000000000**1**0000000000000000000000000

PC: 00000000000000000000**1**0000000000000000000000000000000000

Bradley Malin will talk about re-identifying people from the trails of data the leave behind.

Computer Security & Data Sharing



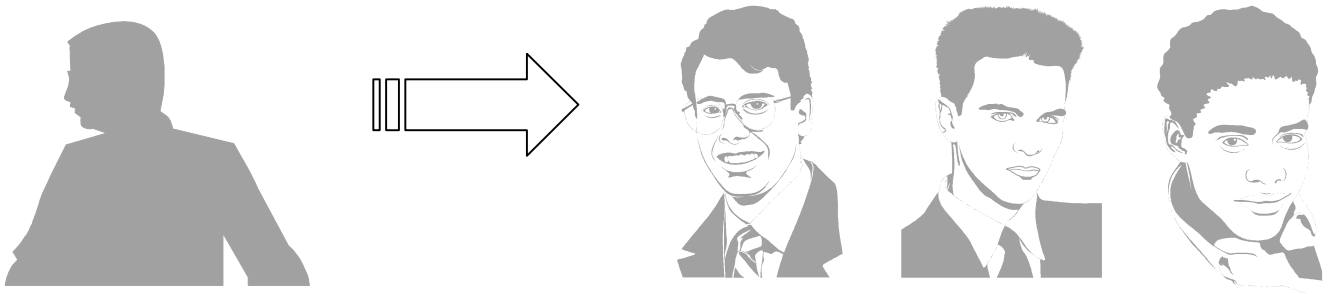
Authentication: login with password
Authorization: allowed to read/write data
Encryption: to avoid eavesdropping

BUT data can re-identify individual!

This talk

- ✓ Data investigations
- Data protection
  - Formal protection models
  - Effort-based models (evolving)
- Surveillance

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



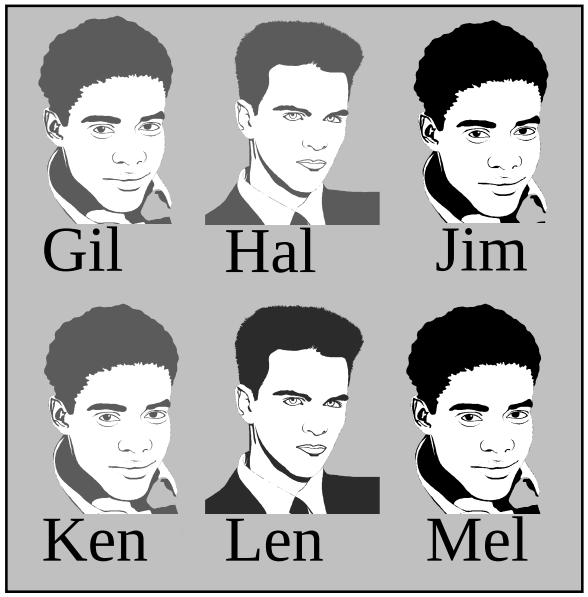
For every record released, there will be at least *k* individuals to whom the record indistinctly refers.

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

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

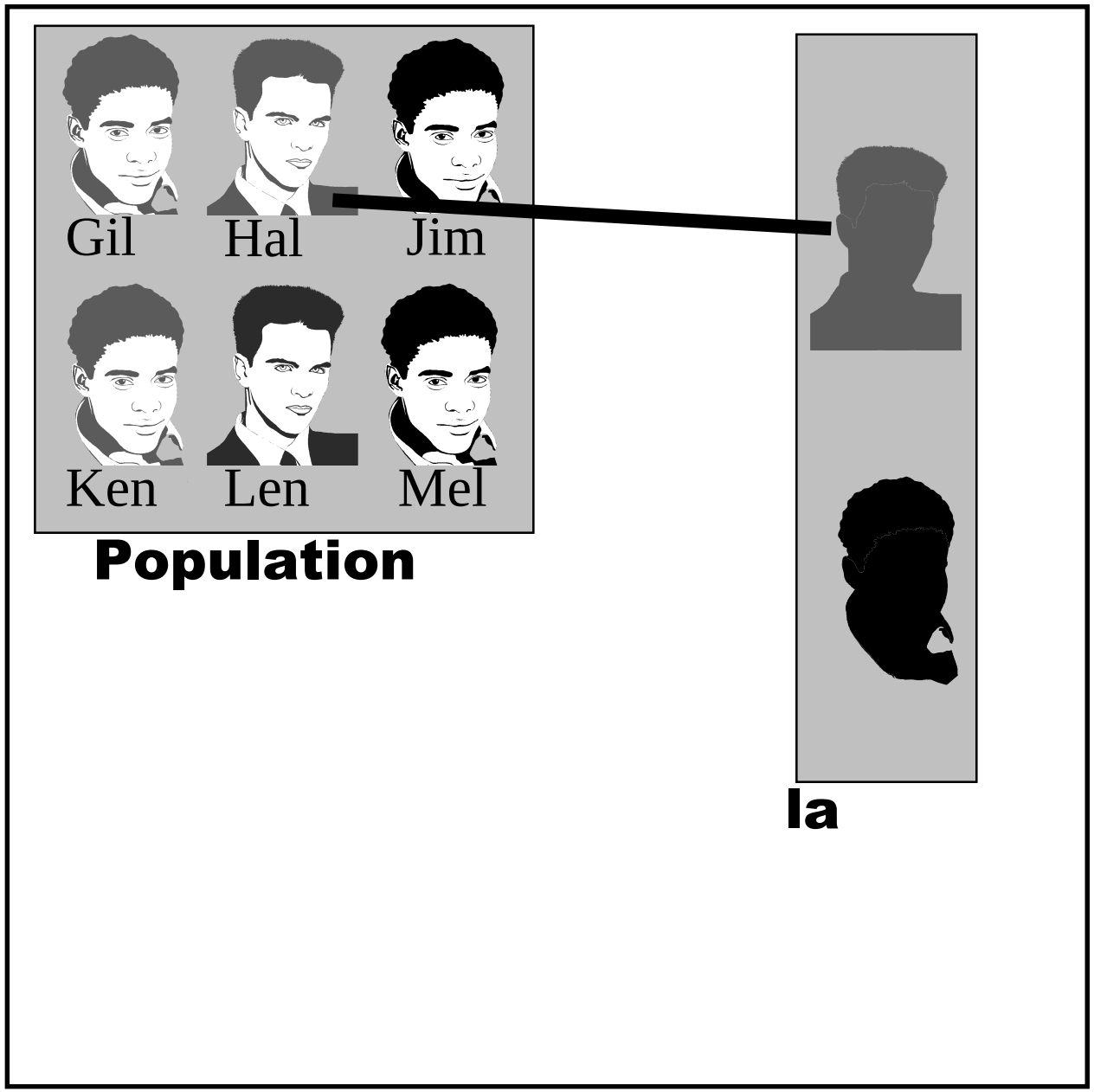
Sweeney 97 and 98

Sample population register of 6 people



Population

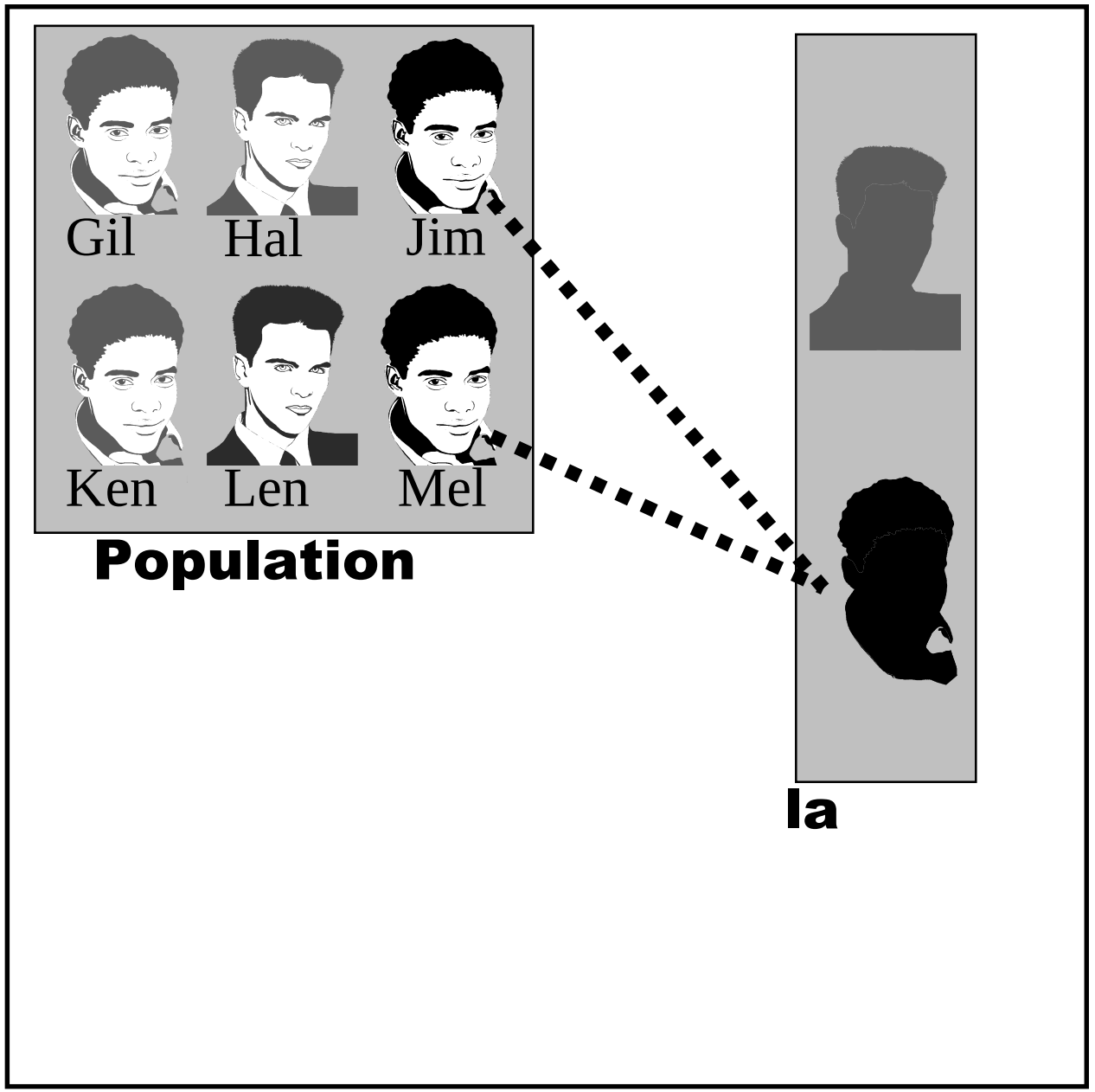
Re-identification Example



There are 3 green figures and 2 figures having the same profile as the release.

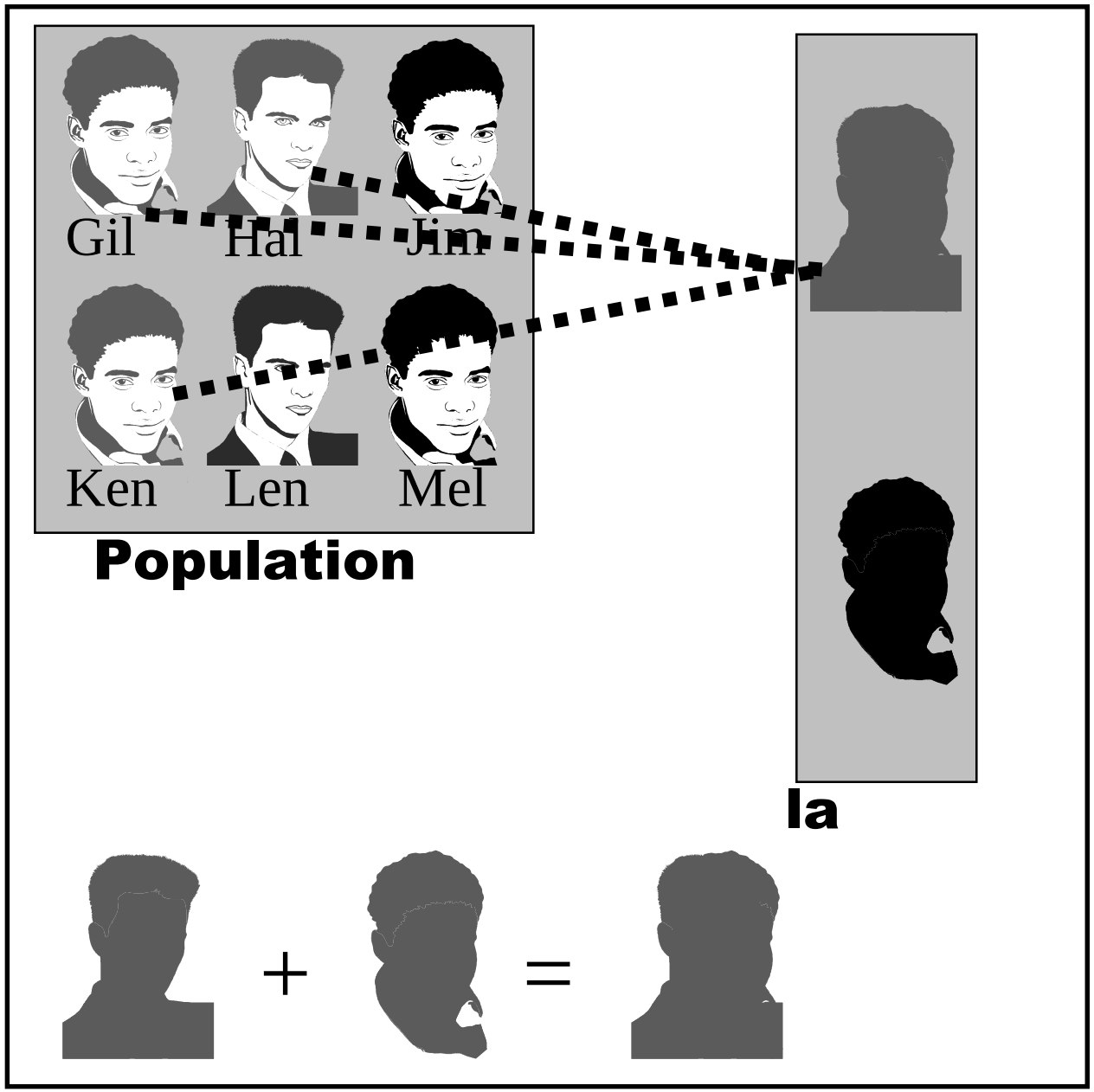
But only Hal is green and has the same figure type as the profile in the release. It is a unique match.

Re-identification Example



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

Re-identification Example

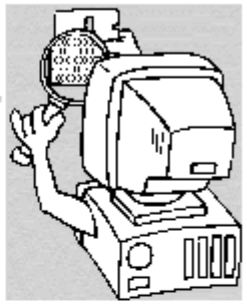


To achieve *k*-map where *k*=2, agents for *Gil*, *Hal* and *Ken* agree to merge their information together.

Information released about any of them results in the same merged image.

Datafly HIPAA Demonstration - Netscape

File Edit View Go Communicator Help



DATAFLY

HIPAA Demonstrations

Original data sample from a problem list

Race	DOB	Sex	ZIP	Problem
black	05/20/1965	M	02141	short of bre
black	08/31/1965	M	02141	chest pain
black	10/28/1965	F	02138	painful eye
black	09/30/1965	F	02138	wheezing
black	07/07/1964	F	02138	obesity
black	11/05/1964	F	02138	chest pain
white	11/28/1964	M	02138	short of bre
white	07/22/1965	F	02139	hypertension
white	08/24/1964	M	02139	obesity
white	05/30/1964	M	02139	fever
white	02/16/1967	M	02138	vomiting

Document: Done

ration - Netscape

unicator Help

Race	DOB	Sex	ZIP	Problem
black	1965	M	02141	short of bre
black	1965	M	02141	chest pain
black	1965	F	02138	painful eye
black	1965	F	02138	wheezing
black	1964	F	02138	obesity
black	1964	F	02138	chest pain
white	196*	*	021**	short of bre
white	196*	*	021**	hypertension
white	1964	M	02139	obesity
white	1964	M	02139	fever
white	1967	M	02138	vomiting

Datafly 2nd option preserving 5-digit ZIP

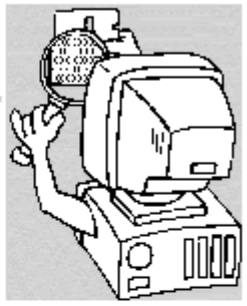
Race	DOB	Sex	ZIP	Problem
black	196*	M	02141	short of bre
black	196*	M	02141	chest pain
black	196*	F	02138	painful eye
black	196*	F	02138	wheezing
black	196*	F	02138	obesity
black	196*	F	02138	chest pain
white	196*	M	02138	short of bre
white	196*	*	02139	hypertension
white	196*	*	02139	obesity
white	196*	*	02139	fever
white	196*	M	02138	vomiting

Document: Done

privacy.cs.cmu.edu/datafly/

Datafly HIPAA Demonstration - Netscape

File Edit View Go Communicator Help



DATAFLY

HIPAA Demonstrations

Original data sample from a problem list

Race	DOB	Sex	ZIP	Problem
black	05/20/1965	M	02141	short of bre
black	08/31/1965	M	02141	chest pain
black	10/28/1965	F	02138	painful eye
black	09/30/1965	F	02138	wheezing
black	07/07/1964	F	02138	obesity
black	11/05/1964	F	02138	chest pain
white	11/28/1964	M	02138	short of bre
white	07/22/1965	F	02139	hypertension
white	08/24/1964	M	02139	obesity
white	05/30/1964	M	02139	fever

ration - Netscape

unicator Help

Race	DOB	Sex	ZIP	Problem
black	1965	M	02141	short of bre
black	1965	M	02141	chest pain
black	1965	F	02138	painful eye
black	1965	F	02138	wheezing
black	1964	F	02138	obesity
black	1964	F	02138	chest pain
white	196*	*	021**	short of bre
white	196*	*	021**	hypertension
white	1964	M	02139	obesity
white	1964	M	02139	fever
white	1967	M	02138	vomiting

Race	DOB	Sex	ZIP	Problem
		M	02141	short of bre
		M	02141	chest pain
		F	02138	painful eye
		F	02138	wheezing
		F	02138	obesity
		F	02138	chest pain
		M	02138	short of bre
		*	02139	hypertension
		*	02139	obesity
		*	02139	fever
		M	02138	vomiting

Document: Done

Ryan Williams will show that *k*-anonymity using generalization and suppression is NP hard in the general case.

privacy.cs.cmu.edu/datafly/

This talk

- ✓ Data investigations
- Data protection
  - ✓ Formal protection models
  - Effort-based models (evolving)
- Surveillance

Video Surveillance Cameras in Lower Manhattan

From http://www.appliedautonomy.com/isee

De-identification of Faces

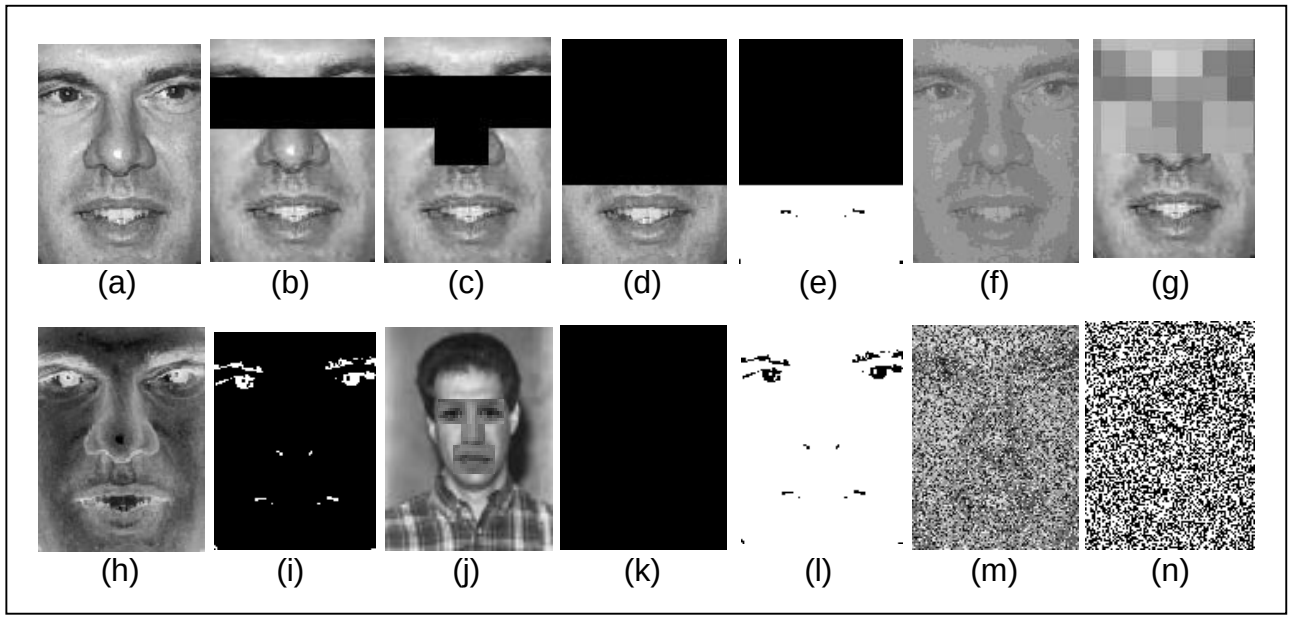
Example.

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

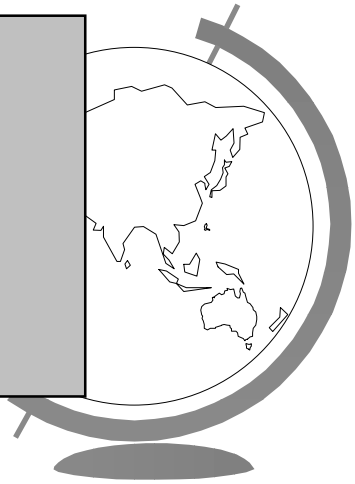
De-identification: T-mask

Example continued...

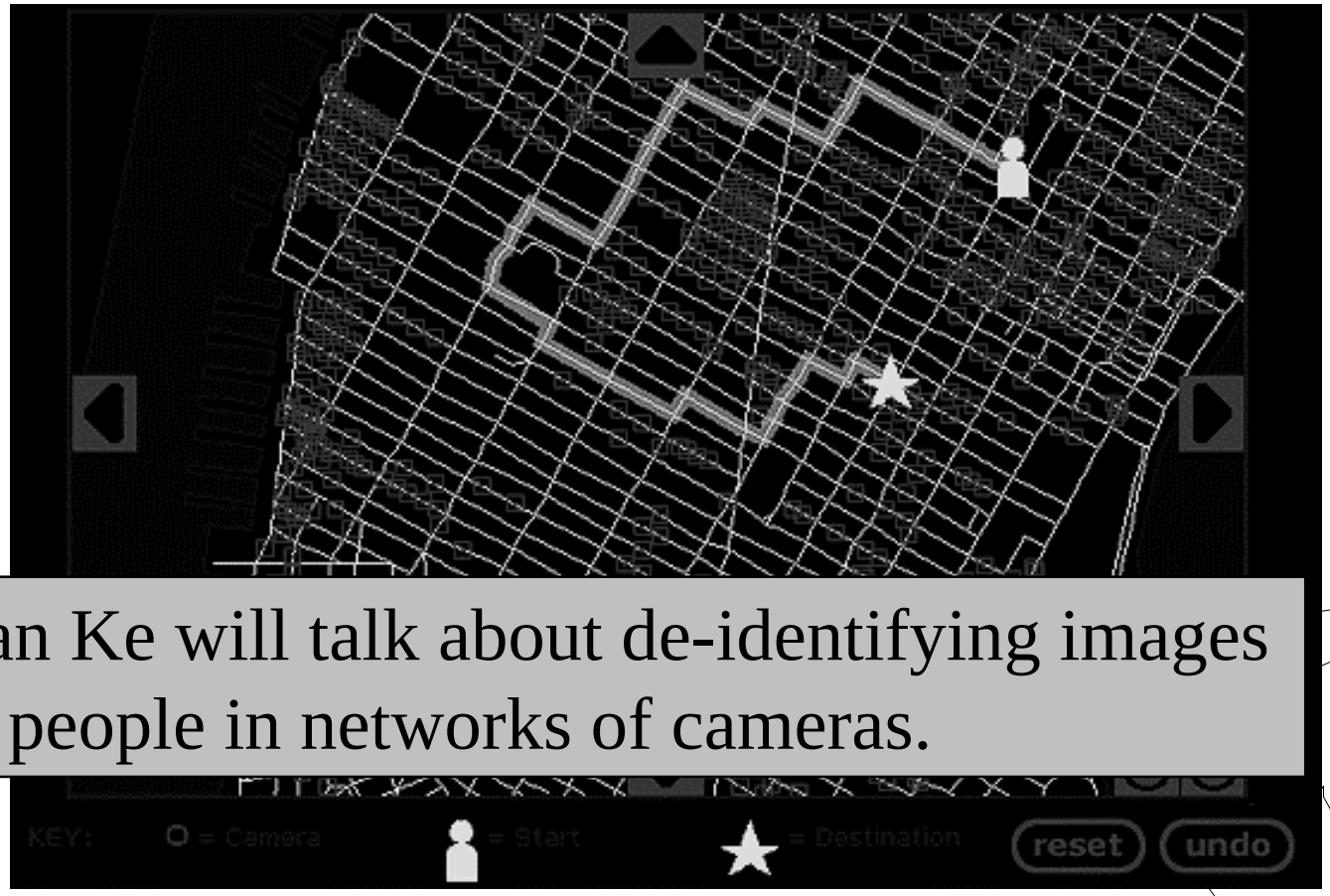
Captured images are de-identified below. Here is a known image of Bob. Which person is Bob?



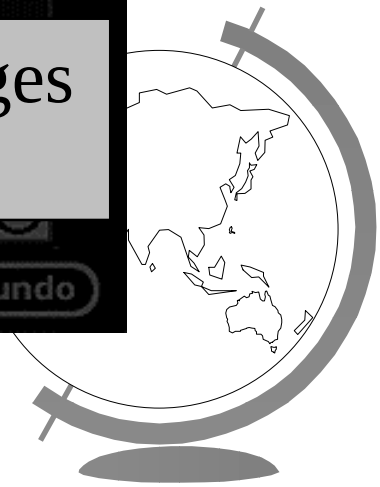
Ralph Gross (for Elaine Newton)
will show how faces can be de-identified to thwart any face recognition system yet preserving many details in the face.



Video Surveillance Cameras in Lower Manhattan



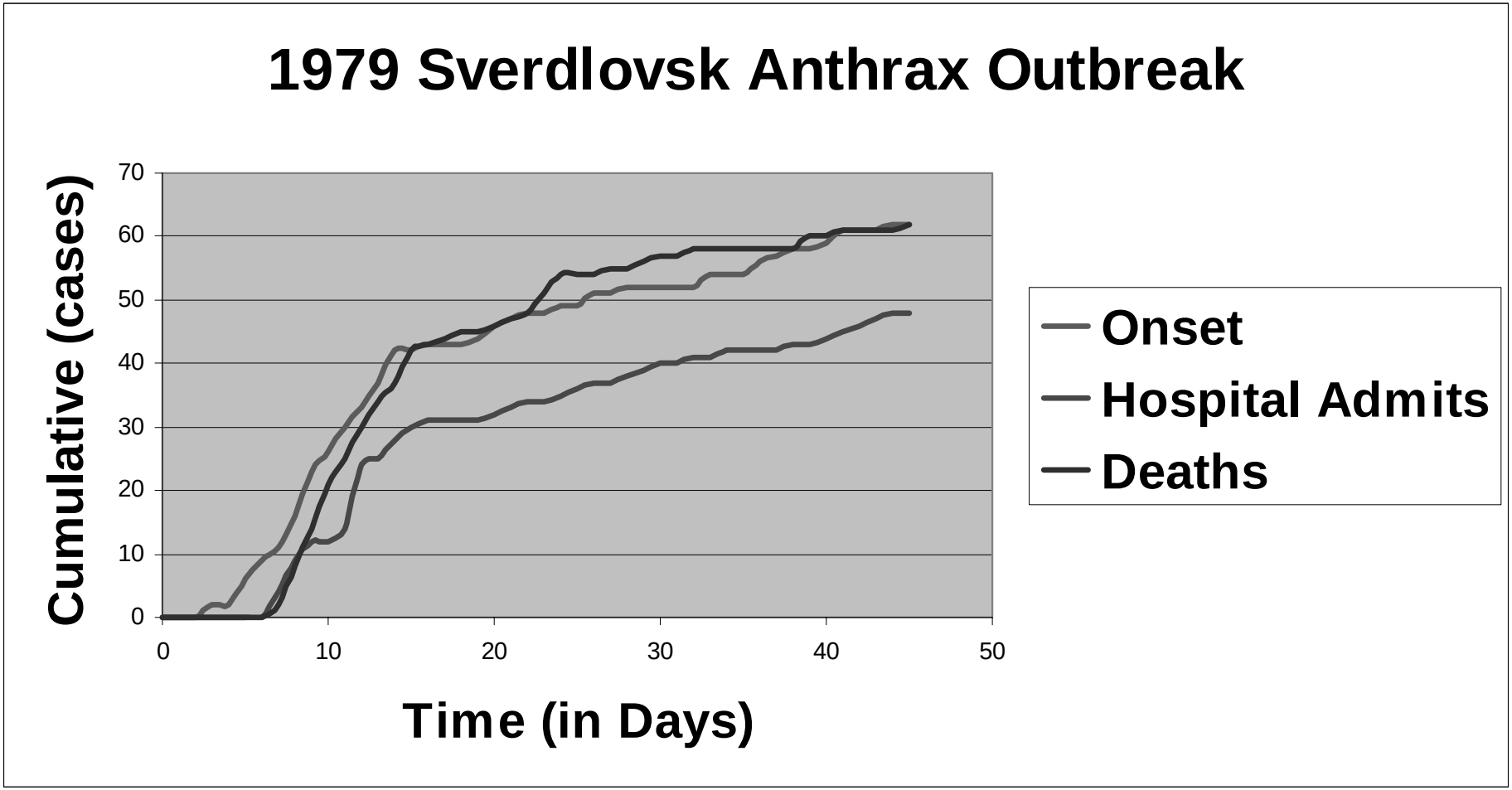
Yan Ke will talk about de-identifying images of people in networks of cameras.



This talk

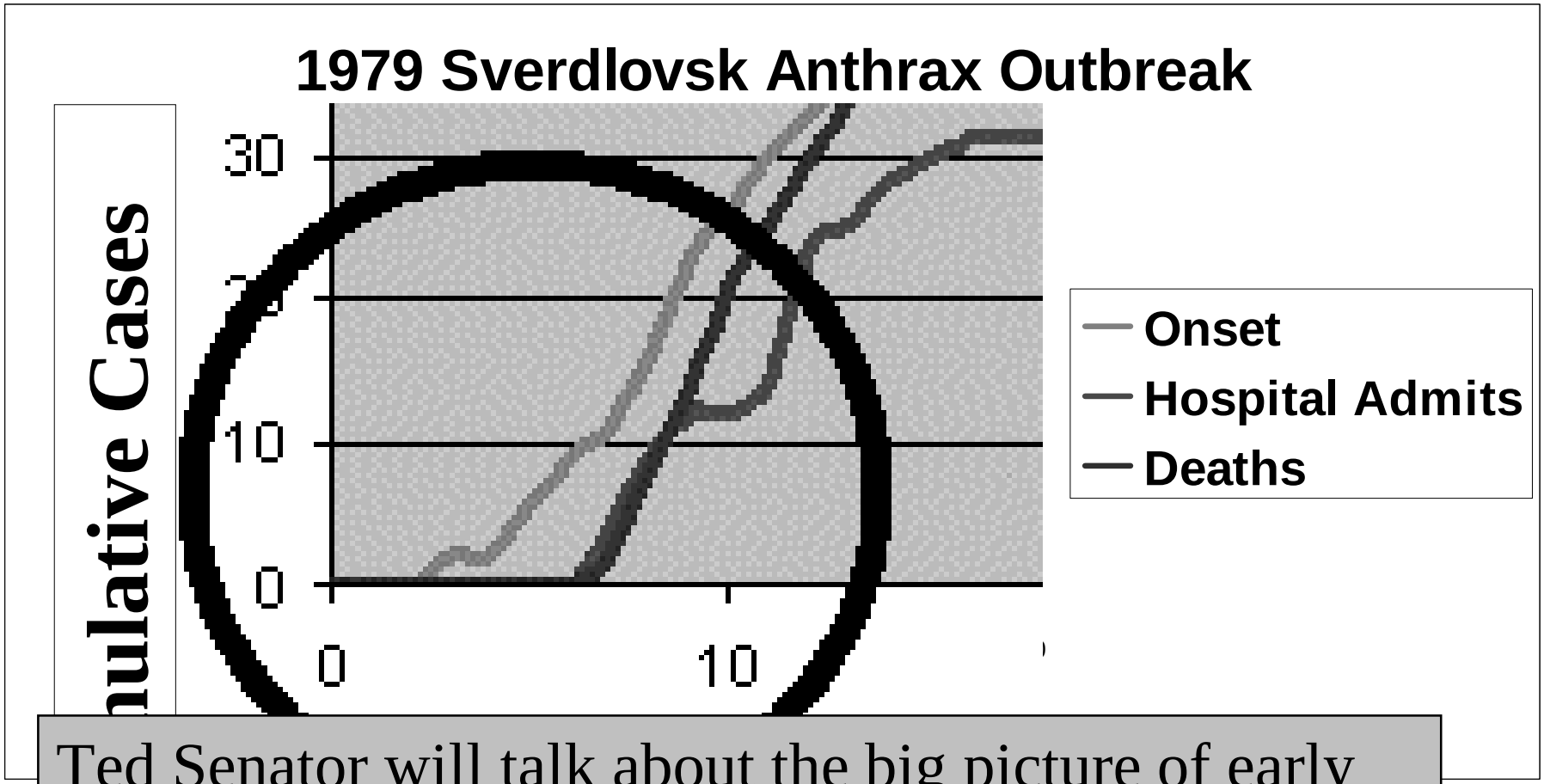
- ✓ Data investigations
- ✓ Data protection
- ● Surveillance

Detect Early using Onset, Coordinate Deaths & Hospital Admits



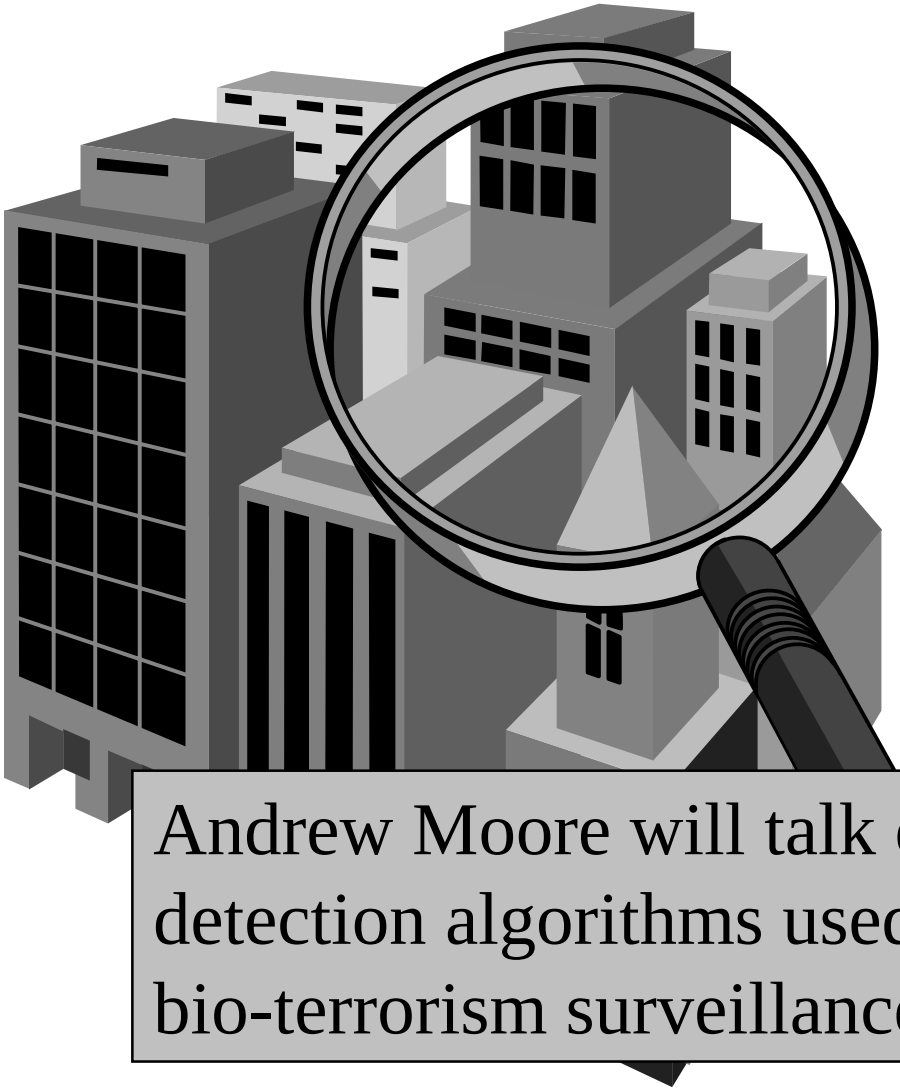
Based on results reported in Guillemin, 1999.

How can we detect onset?
How early on each can we predict?
How does coordination help?



Ted Senator will talk about the big picture of early detection bio-terrorism surveillance systems.

Continuously Observe Behaviors to Detect Onset of Symptoms



Prodromic surveillance:

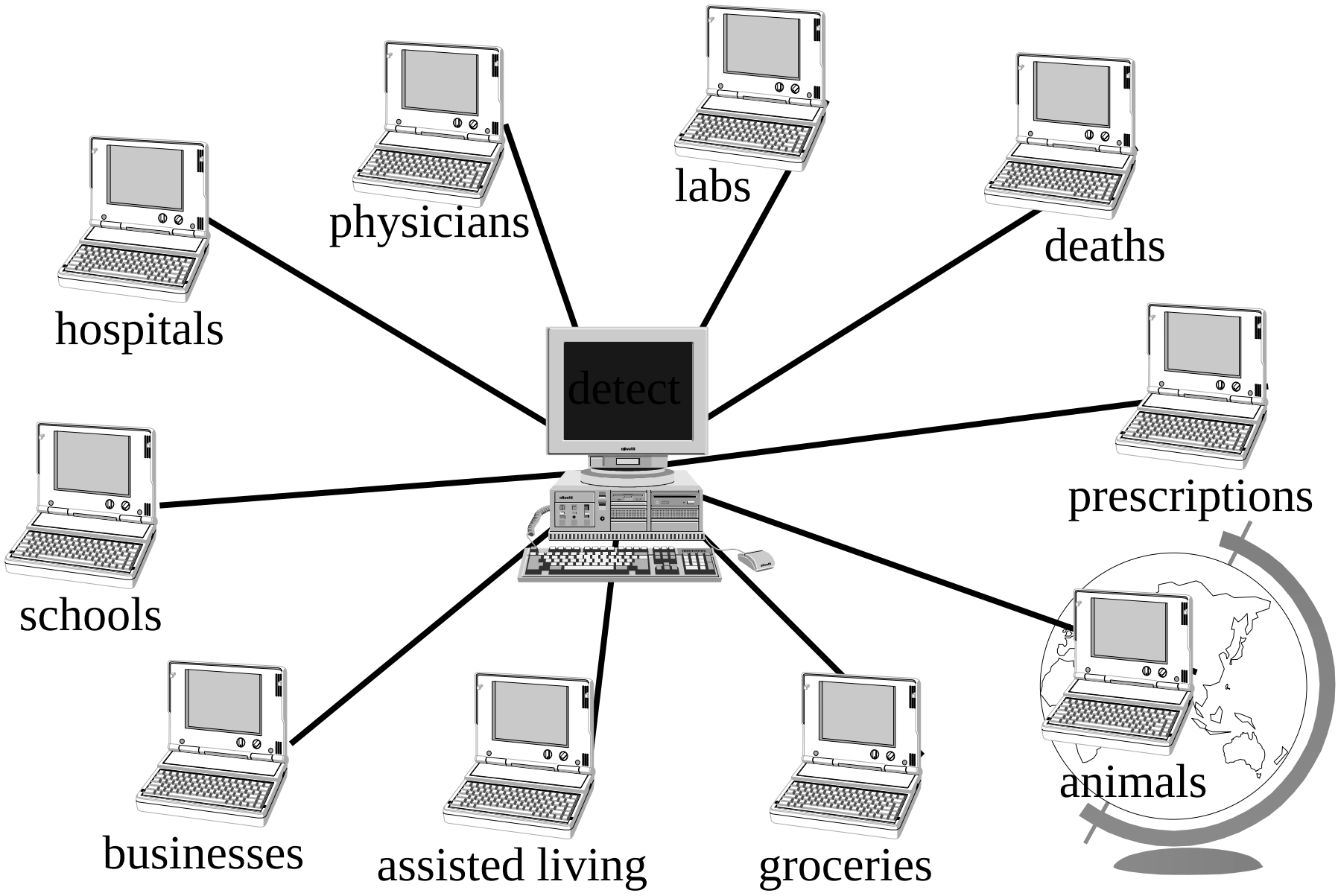
How many are acting ill?

Unusual behaviors→syndromes?

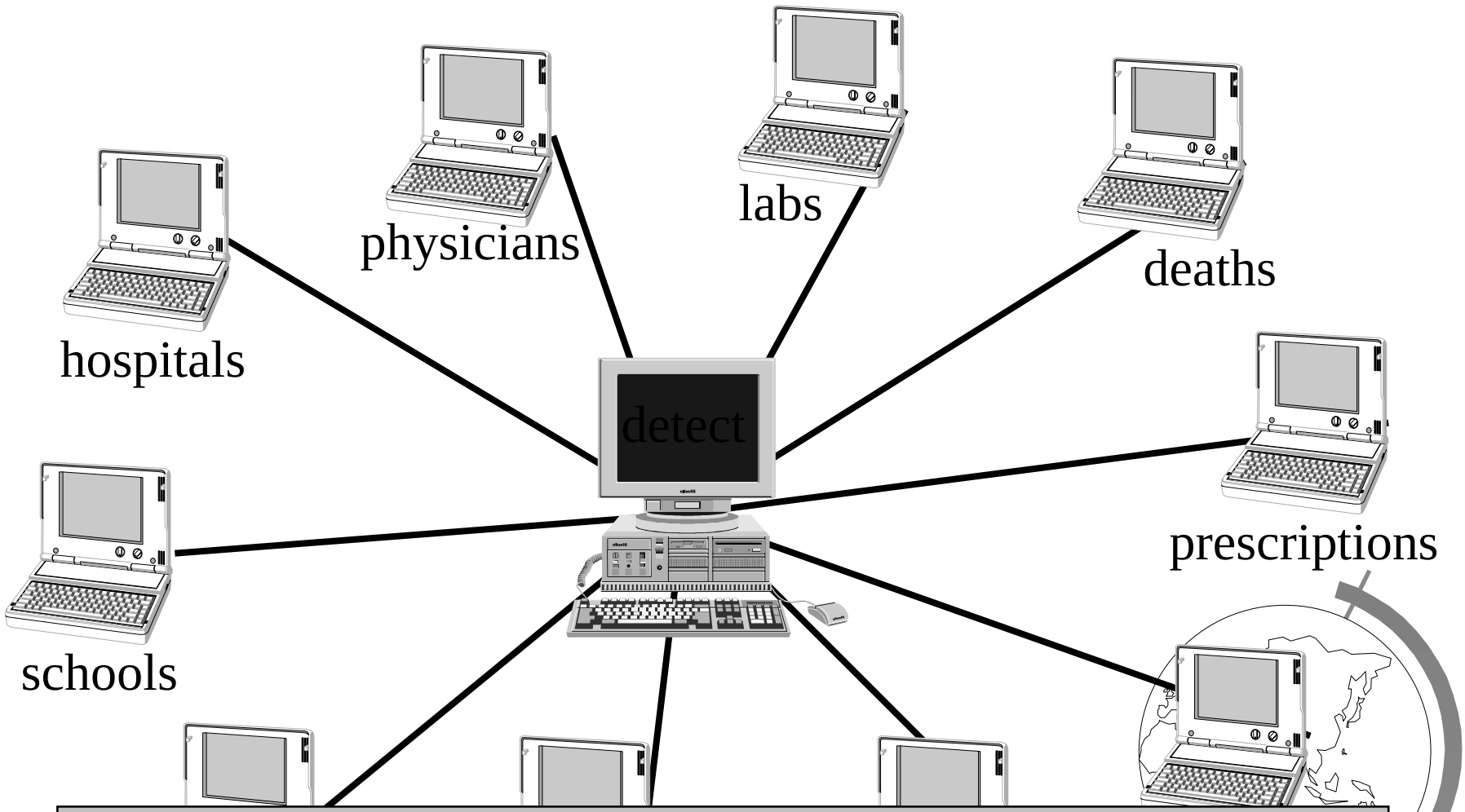
Not confirmed diagnoses!

Andrew Moore will talk describe anomaly detection algorithms used in real-world bio-terrorism surveillance systems.

Centralized Surveillance of Secondary Data



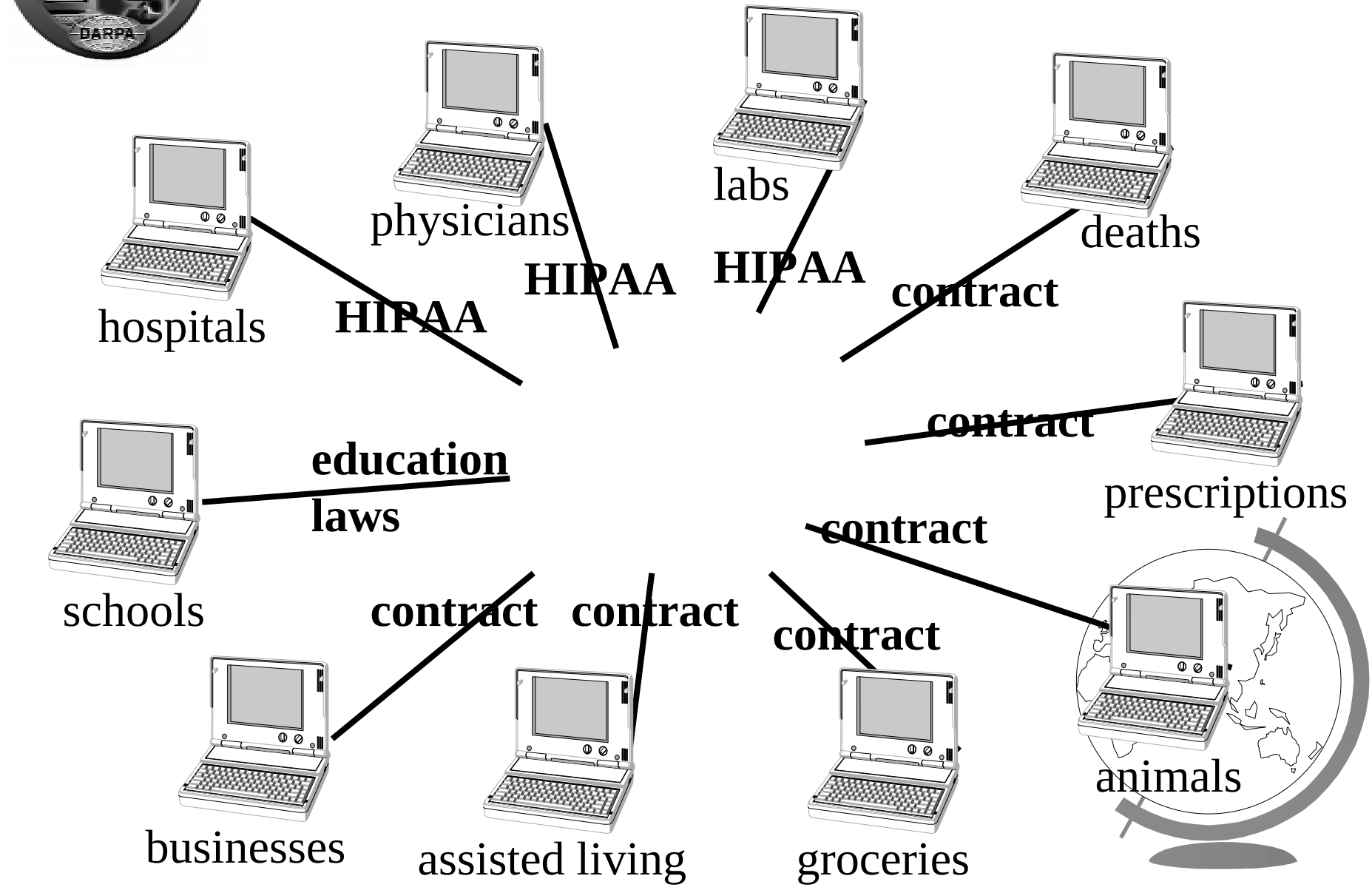
Centralized Surveillance of Secondary Data



Doug Dyer will talk about the surveillance for detecting terrorists.

Access Instruments

*Not including public health law

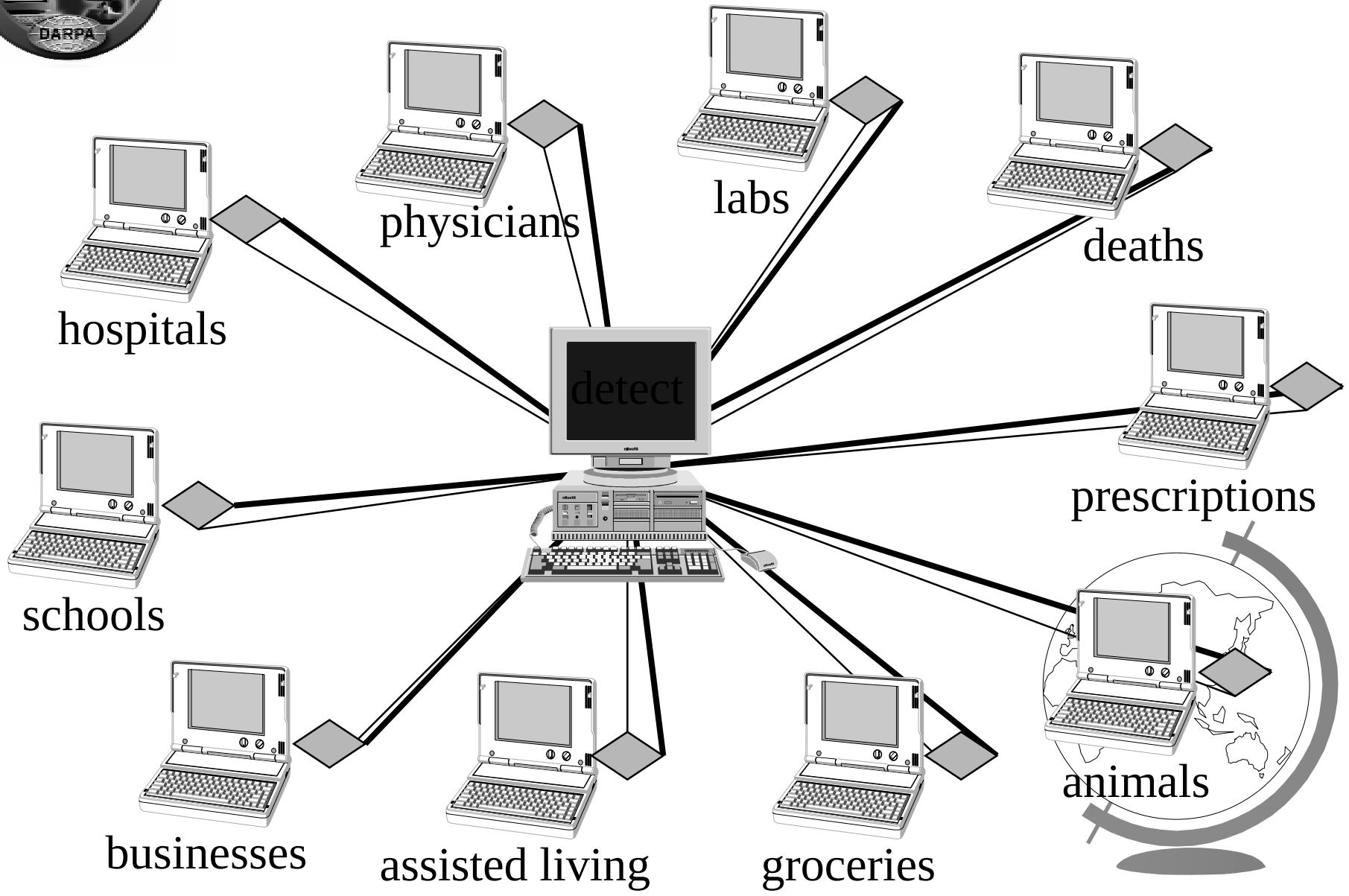


Policy Matters…

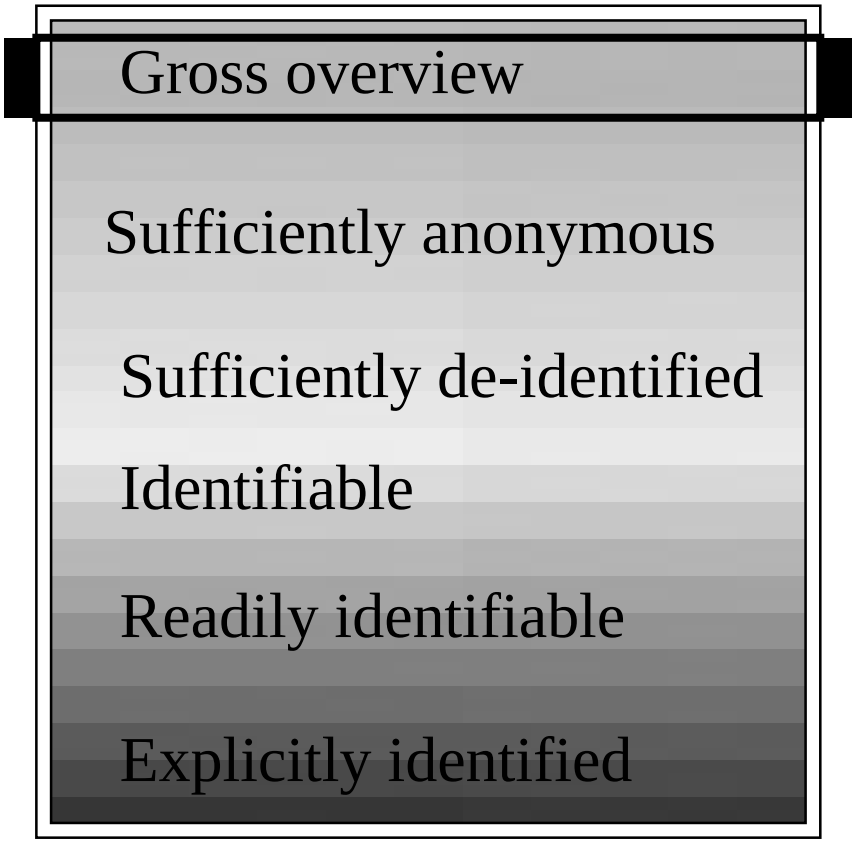
- **FOIA versus Privacy**
- **Law enforcement**
- **Intellectual property**
- **Medical privacy legislation**
- **Internet privacy**
- **Bio-terrorism surve**

Mike Shamos will describe how these laws, regulations and policies frame the mathematics behind solutions.

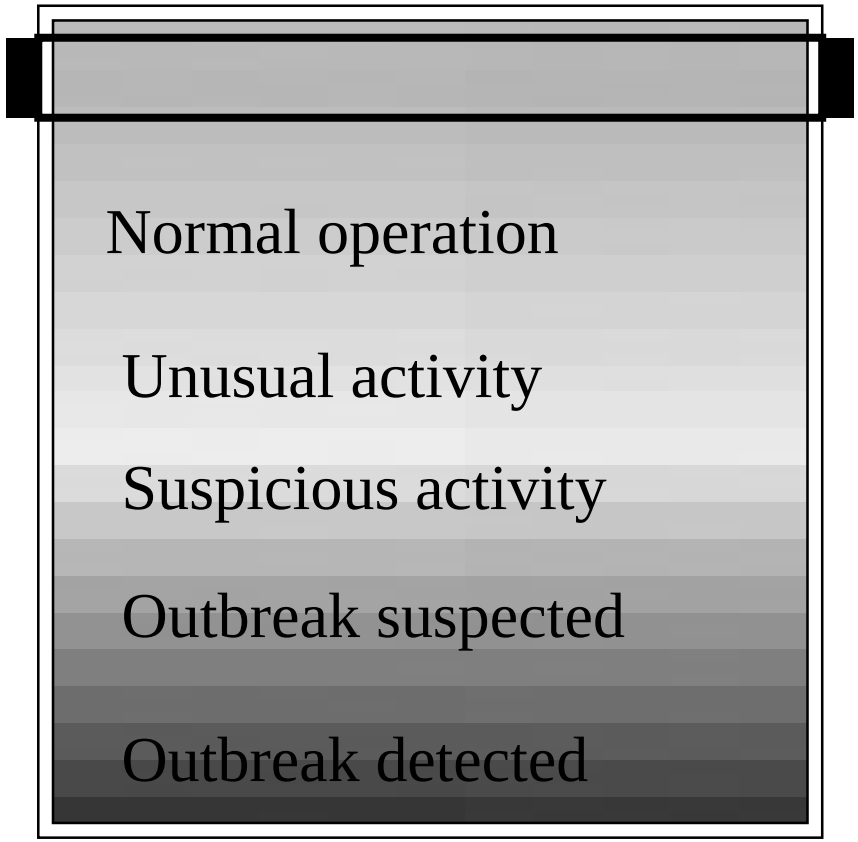
Automated Privacy Module



Mechanical distortion decisions typically renders data useless



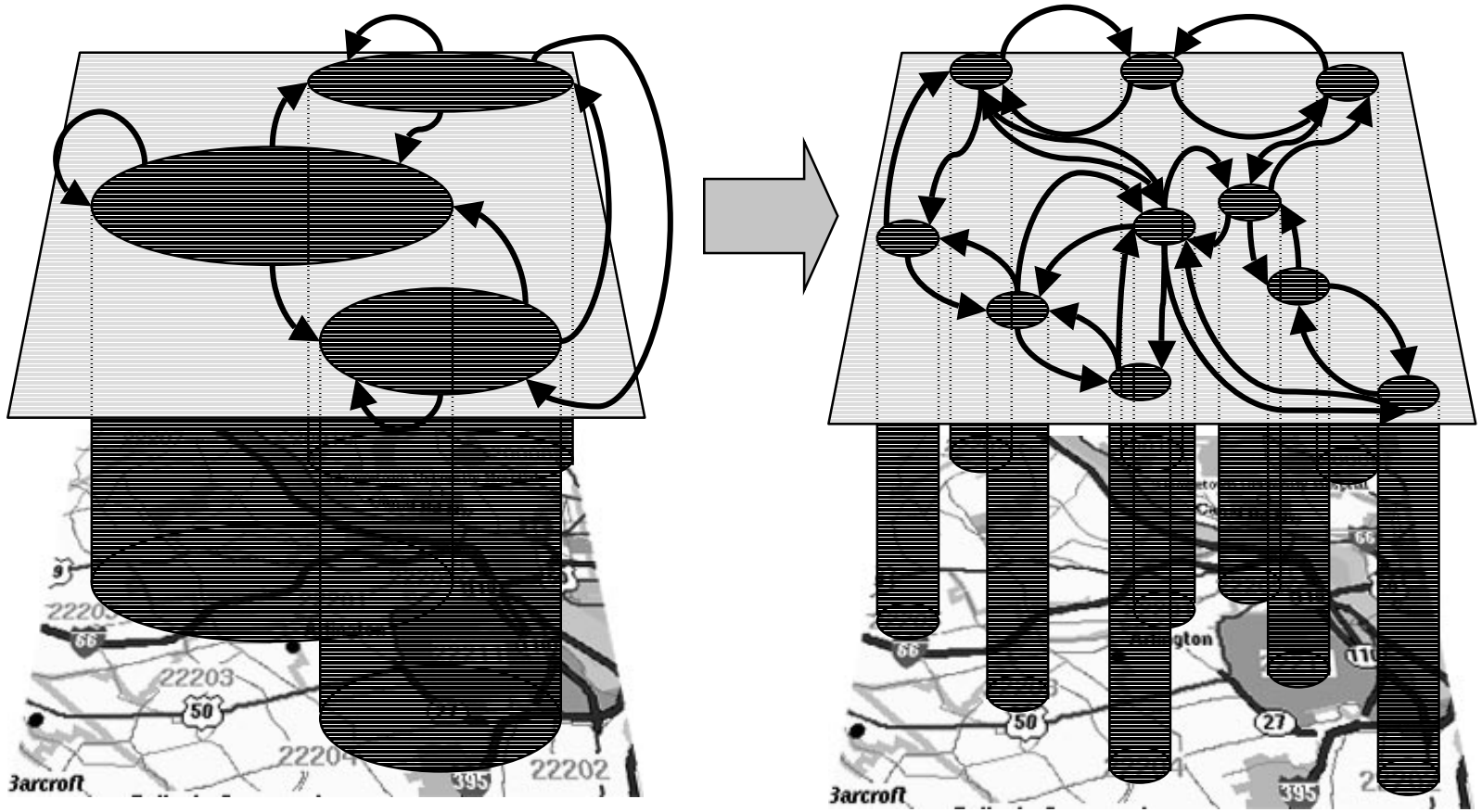
Datafly Idenifiability 0..1



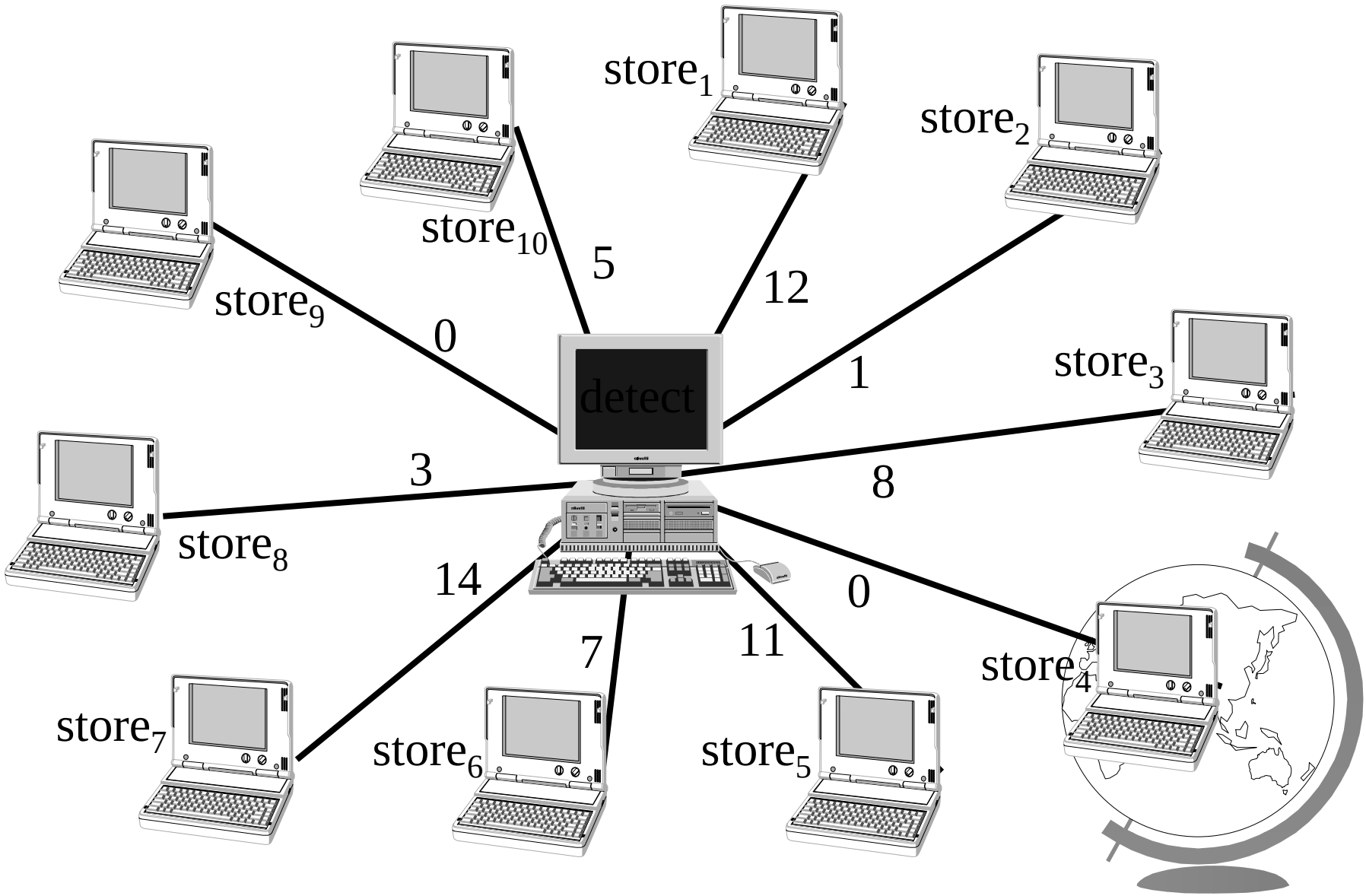
Detection Status 0..1

Dynamically Augment the Model When Surveillance Detects Possible Attack

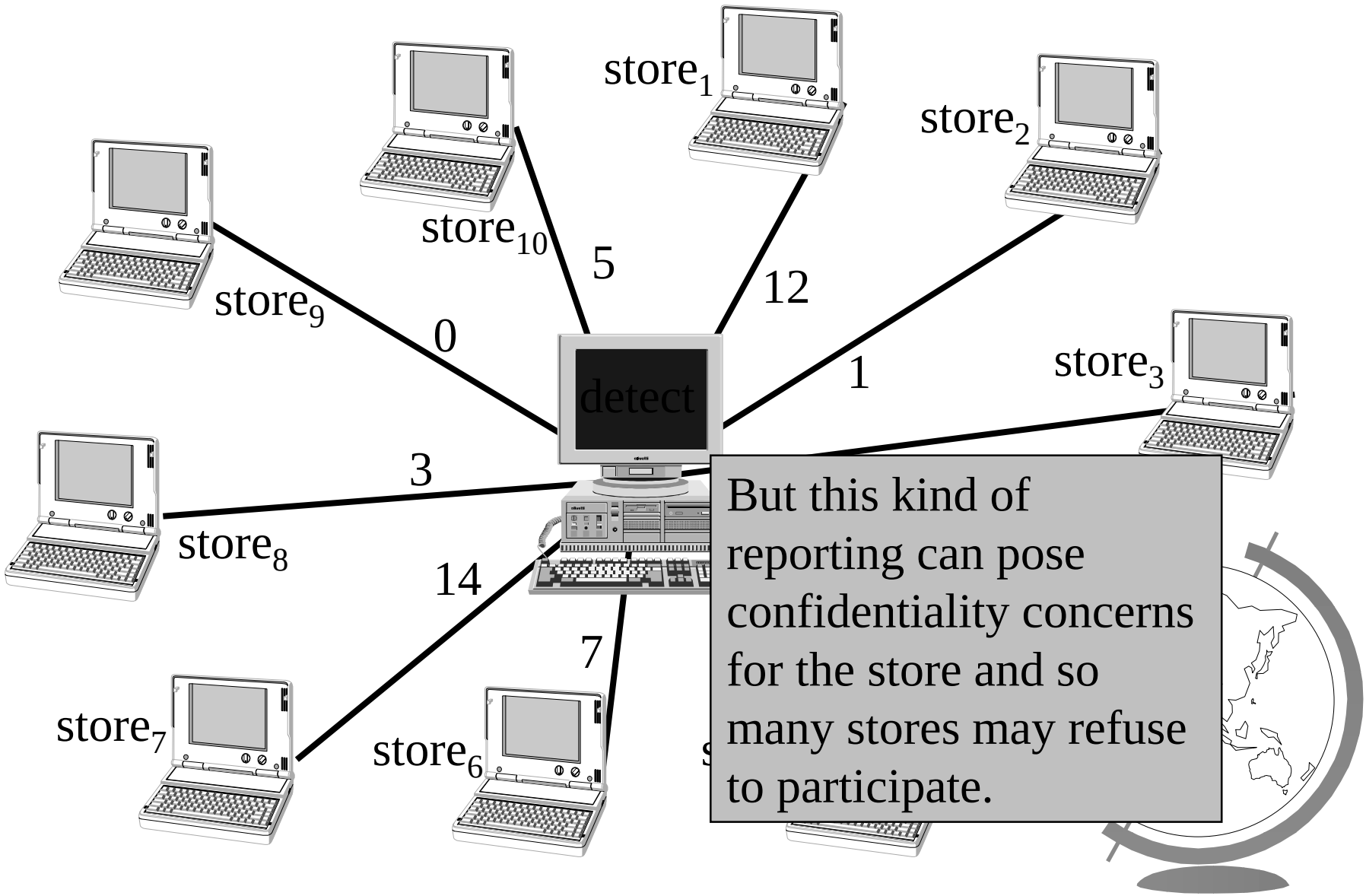
- Lower the privacy threshold when potential attack detected
  - Take advantage of disease-specific processing
  - Need to flush out early suspicions by looking at more detailed data



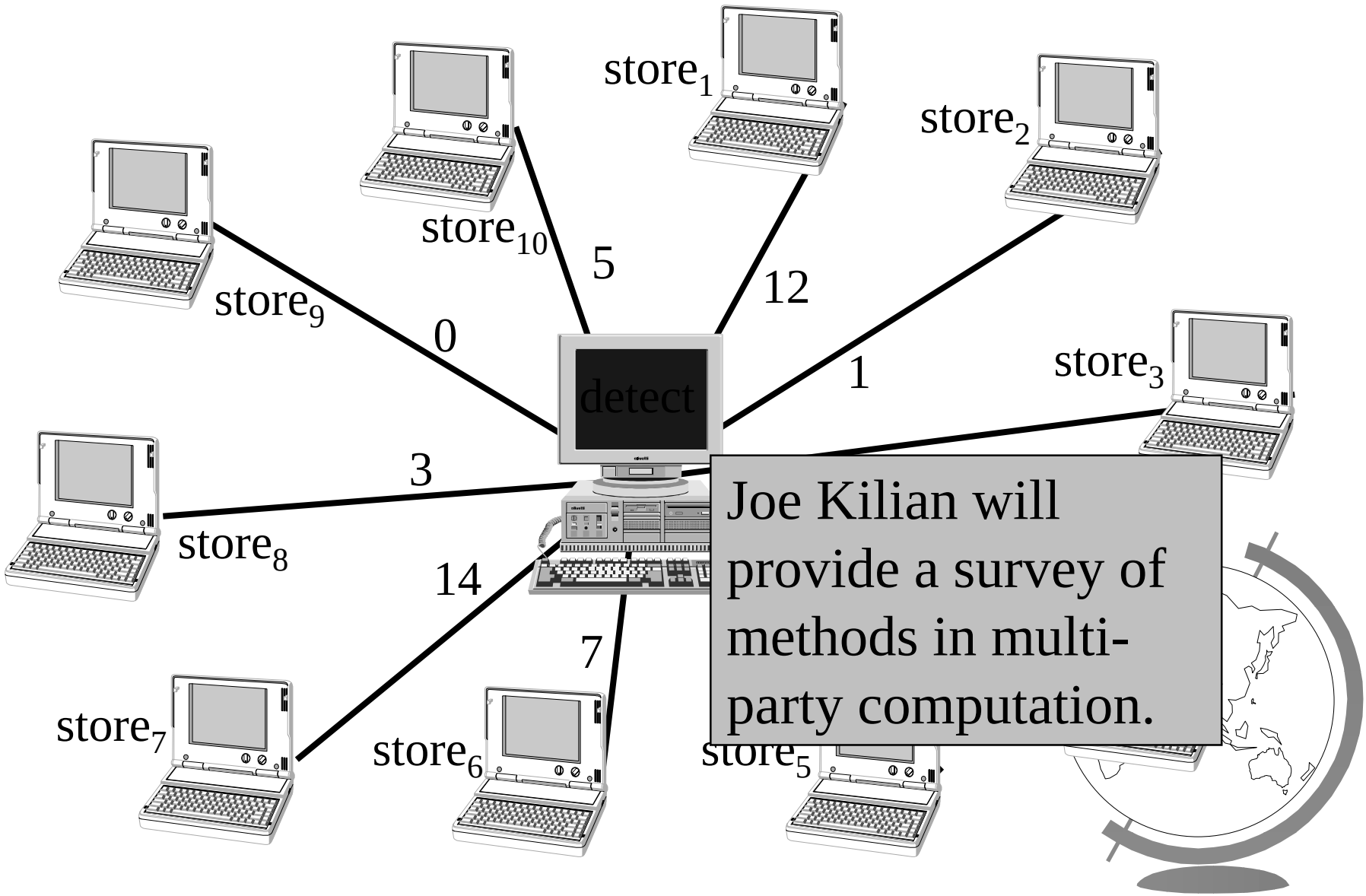
“How many *x* occurred yesterday?”



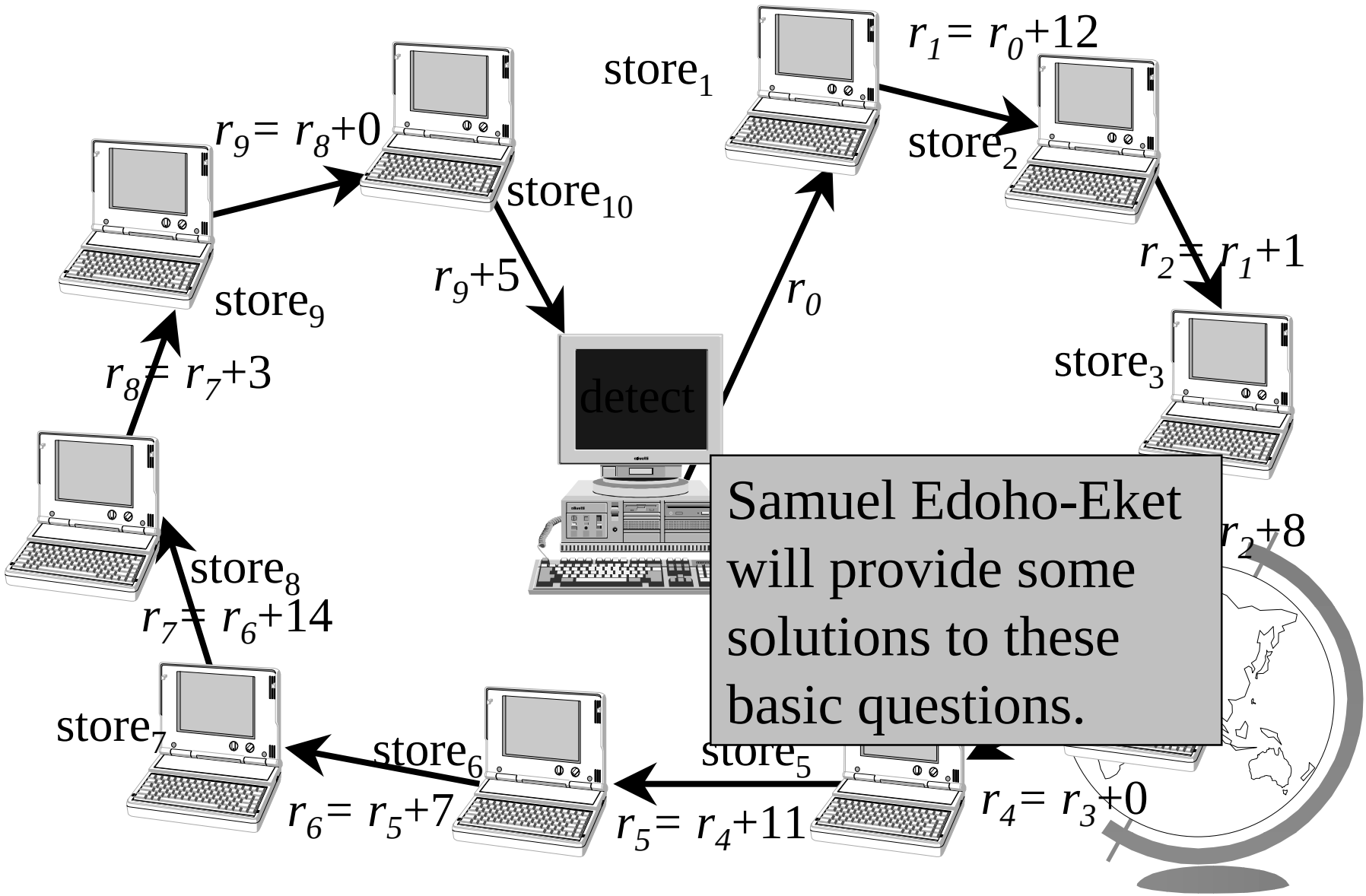
“How many *x* occurred yesterday?”



“How many *x* occurred yesterday?”



Total count: “How many *x* occurred?”



Other presentations

Privacy-preserving data mining:

- Rafail Ostrovsky
- Benny Pinkas
- Johannes Gehrke

Query restriction problem:

- Susmit Sarkar

Statistical approaches:

- Steve Fienberg
- Rebecca Wright

The Question in this Talk

Can computer scientists provide both safety and privacy to society?

Answer:

YES. Three goals: (1) understand the nature of real privacy threats; (2) design technical solutions to integrate with policy to avoid a setting in which society is forced to choose; and, (3) construct technical solutions that address these threats while keeping data useful.

This talk

- ✓ Data investigations
- ✓ Data protection
- ✓ Surveillance

Latanya Sweeney, PhD.
Assistant Professor of Computer Science,
Technology and Policy
latanya@privacy.cs.cmu.edu
http://privacy.cs.cmu.edu/