COLLEGE OF SOCIAL & BEHAVIORAL SCIENCES
School of Information

Database Privacy III

Information Privacy with Applications

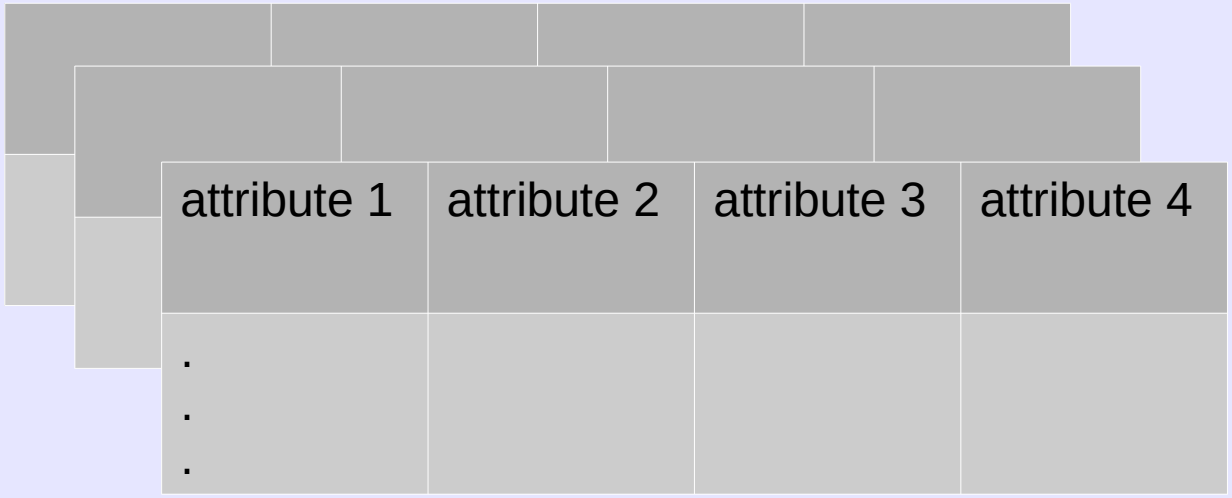
David Sidi (dsidi@email.arizona.edu)

Administrative

- Face detection assignment: how is it going?
- It’s Privacy Week! Share events you come across

Setting

- N databases d_1, d_2, ... , d_N
- Each database d_i has r_i records



Aim of data matching

- Identify and match individual records that describe the same entities

PatTbl

PatientID	Name	DOB	Age	Gender	StreetAddress	Suburb	Postcode
P1273489	John Smith	8/10/1960	51	M	8/42 Miller Street	Melbourne	3011
Q6549234-2	Mick Meyer	30/01/1948	63	M	10 Port Road	Ferny Grove	7004
P7693427-8	Joanna Smith	12/11/1984	27	F	76 George Crest	Sydeny	2020

AdmittedPatients

PID	Surname	GivenName	BirthDate	Sex	AID
25198	Smith	Jo Anna	19841112	1	A347
55642	Smith	John W.	19601008	0	A135
15907	Meier	Michael	19480101	0	A810
99801	Meyer	Mike	19790320	0	A135

Addresses

AID	Street	Location
A135	42 Miller St	3000 Melbourne
A347	16 George Crs	2000 Sydney
A810	PO Box 553	7000 Brisbane

Aim of data matching

- Identify and match individual records that describe the same entities

PatTbl

PatientID	Name	DOB	Age	Gender	StreetAddress	Suburb	Postcode
P1273489	John Smith	8/10/1960	51	M	8/42 Miller Street	Melbourne	3011
Q6549234-2	Mick Meyer	30/01/1948	63	M	10 Port Road	Ferny Grove	7004
P7693427-8	Joanna Smith	12/11/1984	27	F	76 George Crest	Sydeny	2020

AdmittedPatients

PID	Surname	GivenName	BirthDate	Sex	AID
25198	Smith	Jo Anna	19841112	1	A347
55642	Smith	John W.	19601008	0	A135
15907	Meier	Michael	19480101	0	A810
99801	Meyer	Mike	19790320	0	A135

Addresses

AID	Street	Location
A135	42 Miller St	3000 Melbourne
A347	16 George Crs	2000 Sydney
A810	PO Box 553	7000 Brisbane

- Pre-HIPAA
- “De-identified” hospital records
- Attributes included ZIP, DOB, gender
- 87% uniquely identified (61% in replication)

More on Golle's Replication

- Threat varies by age: Sharp reduction in risk for those most at risk of identification (the "Muggsys" of the world) at around age 20. **Why?**

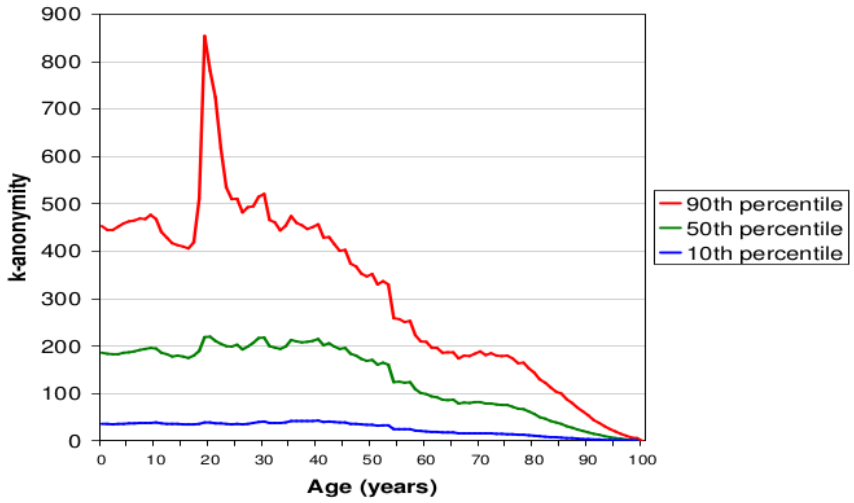


Figure 3: Anonymity of the U.S. population, by age, given {Gender, ZIP code, Year of birth}.

More on Golle's Replication

- Suppose you're friends with an elderly Muggsy, and want to help him. What might you do? (Group privacy)

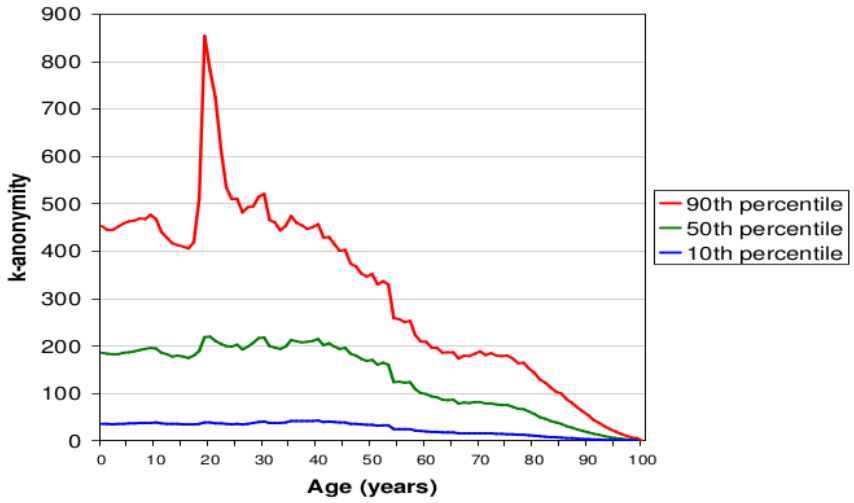


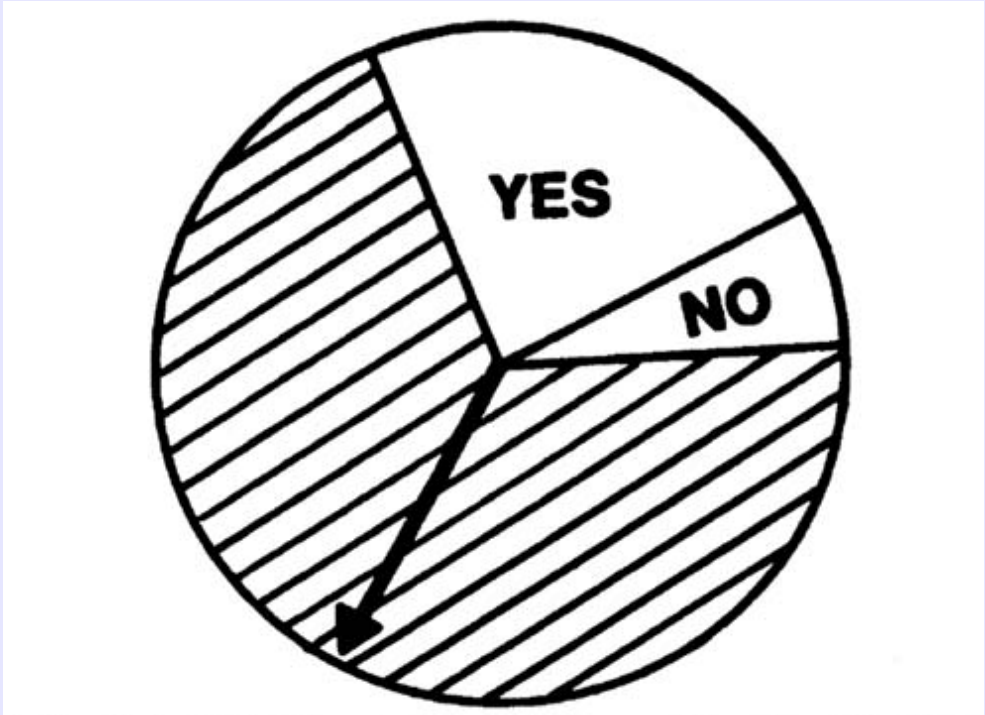
Figure 3: Anonymity of the U.S. population, by age, given {Gender, ZIP code, Year of birth}.

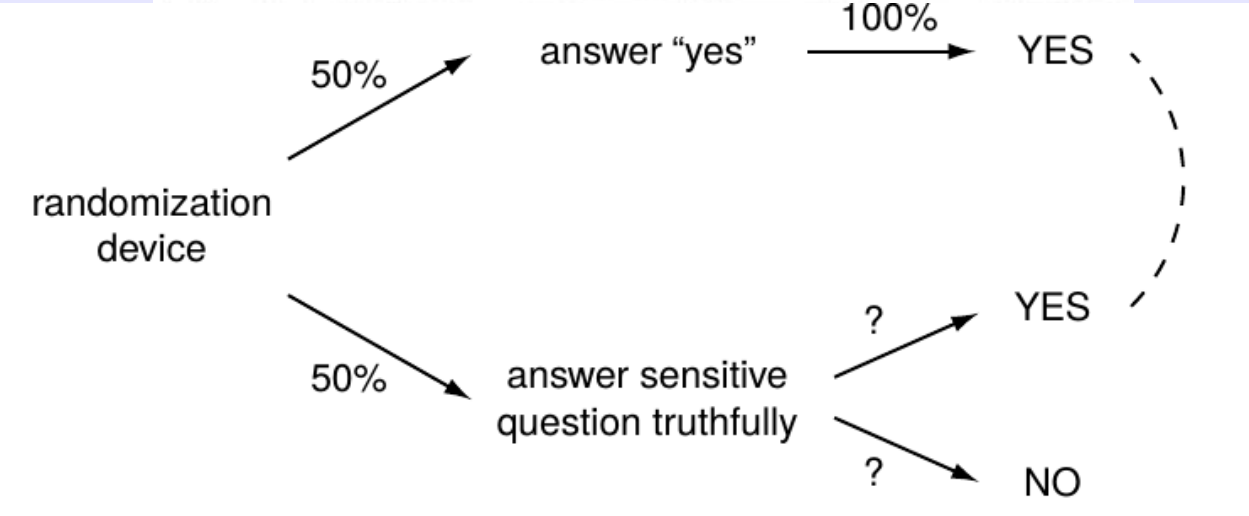
“those willing to sacrifice truthfulness for optimal anonymity should claim, when asked for their age and ZIP code, to be a 21-year-old male from Camp Pendleton, California (ZIP code 92054); or, if female, to be a 19-year-old from College Station, Texas (ZIP code 77840). They will share these characteristics with respectively 4,099 other males and 3,744 other females.”

What’s the name for this approach to privacy protection?

Obfuscation doesn't have to be adversarial: Randomized Response Techniques

(These are from different sources; the numbers don't line up)





Let’s ask embarrassing questions

Simplest nontrivial matching task

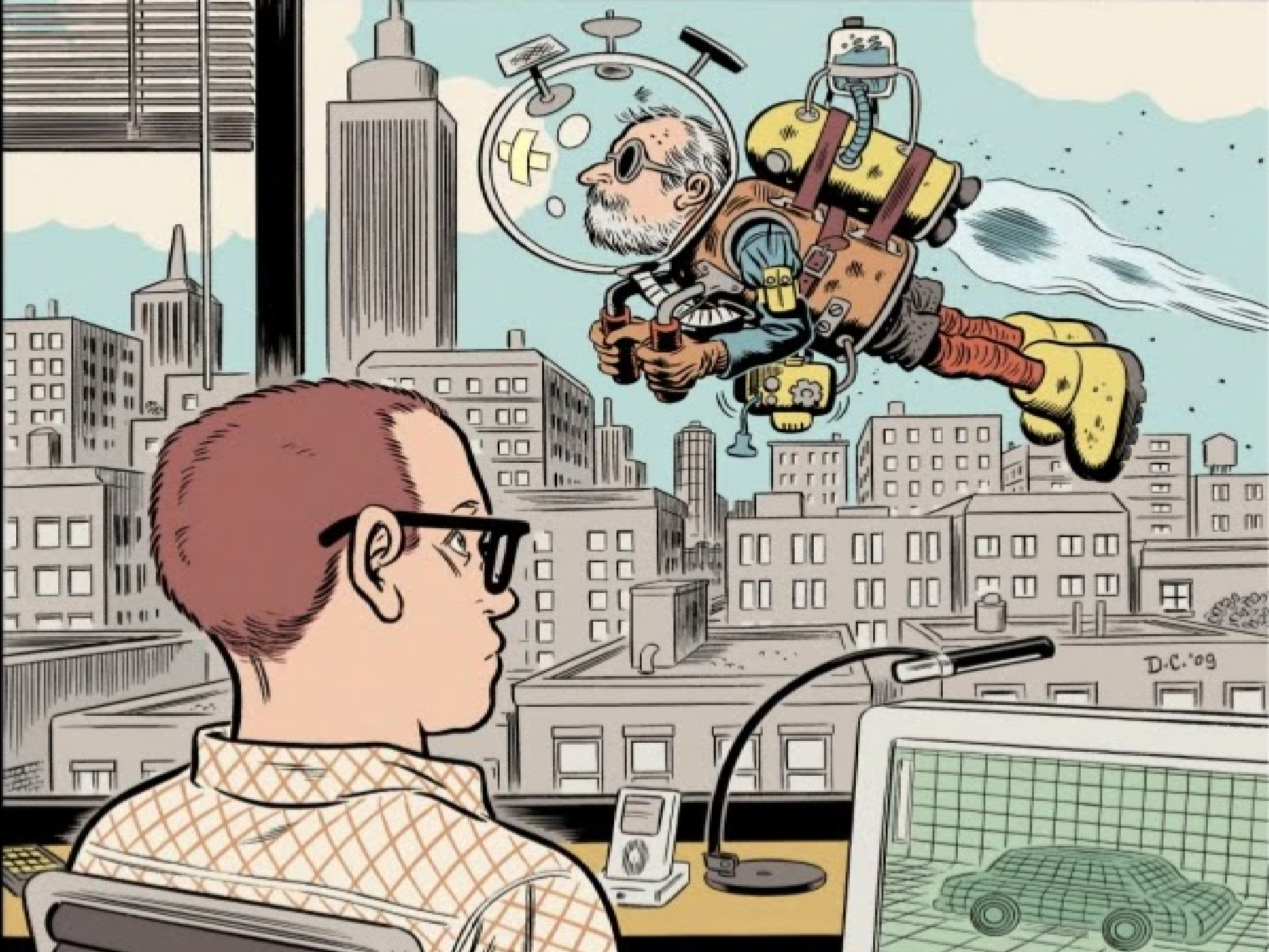
- Two databases are accurate, complete, unchanging, robust, and consistent over time
- The same unique entity identifiers are used in each of them
- Matching reduces to a database join

Simplest nontrivial matching task

- Two databases are accurate, complete, unchanging, robust, and consistent over time
- ~~The same unique entity identifiers are used in each of them~~
- Must use shared attributes

Simplest nontrivial matching task

- ~~Two databases are accurate, complete, unchanging, robust, and consistent over time~~
- ~~The same unique entity identifiers are used in each of them~~
- Must use shared attributes
- Must handle “dirty data”


D.C. '09