Database Privacy

Integrated Session 3

“Sweet Analytics, ‘tis thou has ravished me!”

Marlow, *Faustus*, I, 34

David Sidi (dsidi@email.arizona.edu)

Administrative

- Privacy Week!
- Reminder: we have one more integrated session after this one

Database Privacy?

“What’s a database, what kinds of database are we thinking about here, and what’s the privacy problem?”

attributes

records

	Name	Age	Race	Marital Status
1	Jane Doe	37	White	Single
2	Joe Bloggs	40	Black	Married

statistic	group	count	median	mean
1A	total population	7	37	35
2A	female	4	39.5	41
2B	male	3	25	27
2C	black or African American	3	40	48
2D	white	4	26.5	25.25
3A	single adults	(D)	(D)	(D)
3B	married adults	4	38.5	45.25
4A	black or African American male	(D)	(D)	(D)
4B	black or African American female	(D)	(D)	(D)
4C	white male	(D)	(D)	(D)
4D	white female	(D)	(D)	(D)
5A	persons under 5 years	(D)	(D)	(D)
5B	persons under 18 years	(D)	(D)	(D)
5C	persons 64 years or over	(D)	(D)	(D)

attributes

records

	Name	Age	Race	Marital Status
1	Jane Doe	37	White	Single
2	Joe Bloggs	40	Black	Married

statistic	group	count	median	mean
1A	total population	7	37	35
2A	female	4	39.5	41
2B	male	3	25	27
2C	black or African American	3	40	48
2D	white	4	26.5	25.25
3A	single adults	(D)	(D)	(D)
3B	married adults	4	38.5	45.25
4A	black or African American male	(D)	(D)	(D)
4B	black or African American female	(D)	(D)	(D)
4C	white male	(D)	(D)	(D)
4D	white female	(D)	(D)	(D)
5A	persons under 5 years	(D)	(D)	(D)
5B	persons under 18 years	(D)	(D)	(D)
5C	persons 64 years or over	(D)	(D)	(D)

attributes

records

	Name	Age	Race	Marital Status
1	Jane Doe	37	White	Single
2	Joe Bloggs	40	Black	Married

MICRODATA TABLE

statistic	group	count	median	mean
1A	total population	7	37	35
2A	female	4	39.5	41
2B	male	3	25	27
2C	black or African American	3	40	48
2D	white	4	26.5	25.25
3A	single adults	(D)	(D)	(D)
3B	married adults	4	38.5	45.25
4A	black or African American male	(D)	(D)	(D)
4B	black or African American female	(D)	(D)	(D)
4C	white male	(D)	(D)	(D)
4D	white female	(D)	(D)	(D)
5A	persons under 5 years	(D)	(D)	(D)
5B	persons under 18 years	(D)	(D)	(D)
5C	persons 64 years or over	(D)	(D)	(D)

STATISTICAL TABLE

What is a constraint, informally?

Who would like to explain how we used statistics in the assignment to infer microdata? In what sense are the statistics “constraints?”

Who would like to explain how we used statistics in the assignment to infer microdata? In what sense are the statistics “constraints?”

Mean = 25

A1	A2
0	50

A1	A2
1	49

A1	A2
25	25

A1	A2
49	1

A1	A2
50	0

attributes

records

	Name	Age	Race	Marital Status
1	Jane Doe	37	White	Single
2	Joe Bloggs	40	Black	Married

MICRODATA TABLE

statistic	group	count	median	mean
1A	total population	7	37	35
2A	female	4	39.5	41
2B	male	3	25	27
2C	black or African American	3	40	48
2D	white	4	26.5	25.25
3A	single adults	(D)	(D)	(D)
3B	married adults	4	38.5	45.25
4A	black or African American male	(D)	(D)	(D)
4B	black or African American female	(D)	(D)	(D)
4C	white male	(D)	(D)	(D)
4D	white female	(D)	(D)	(D)
5A	persons under 5 years	(D)	(D)	(D)
5B	persons under 18 years	(D)	(D)	(D)
5C	persons 64 years or over	(D)	(D)	(D)

STATISTICAL TABLE

Why this?

“Who cares about databases with demographic attributes like age, sex, race, and marital status?”


NO JAPS
WANTED


Übersicht
mit
Hollerith
Lochkarten
DEUTSCHE HOLLERITH MASCHINEN GESELLSCHAFT M·B·H· BERLIN-LICHTERFELDE


EUROPEANS ONLY


UNITED STATES DEPARTMENT OF HOMELAND SECURITY
CENTER FOR DOMESTIC PREPAREDNESS
FEMA

Demographics microdata is private and can lead to serious harm, so it can’t be released to the public.

On the other hand, some of the benefits of that microdata can be had *only* if they or their products are released to the public.

So we need a way to release statistical tables that resists attempts to infer the underlying microdata from the statistics.

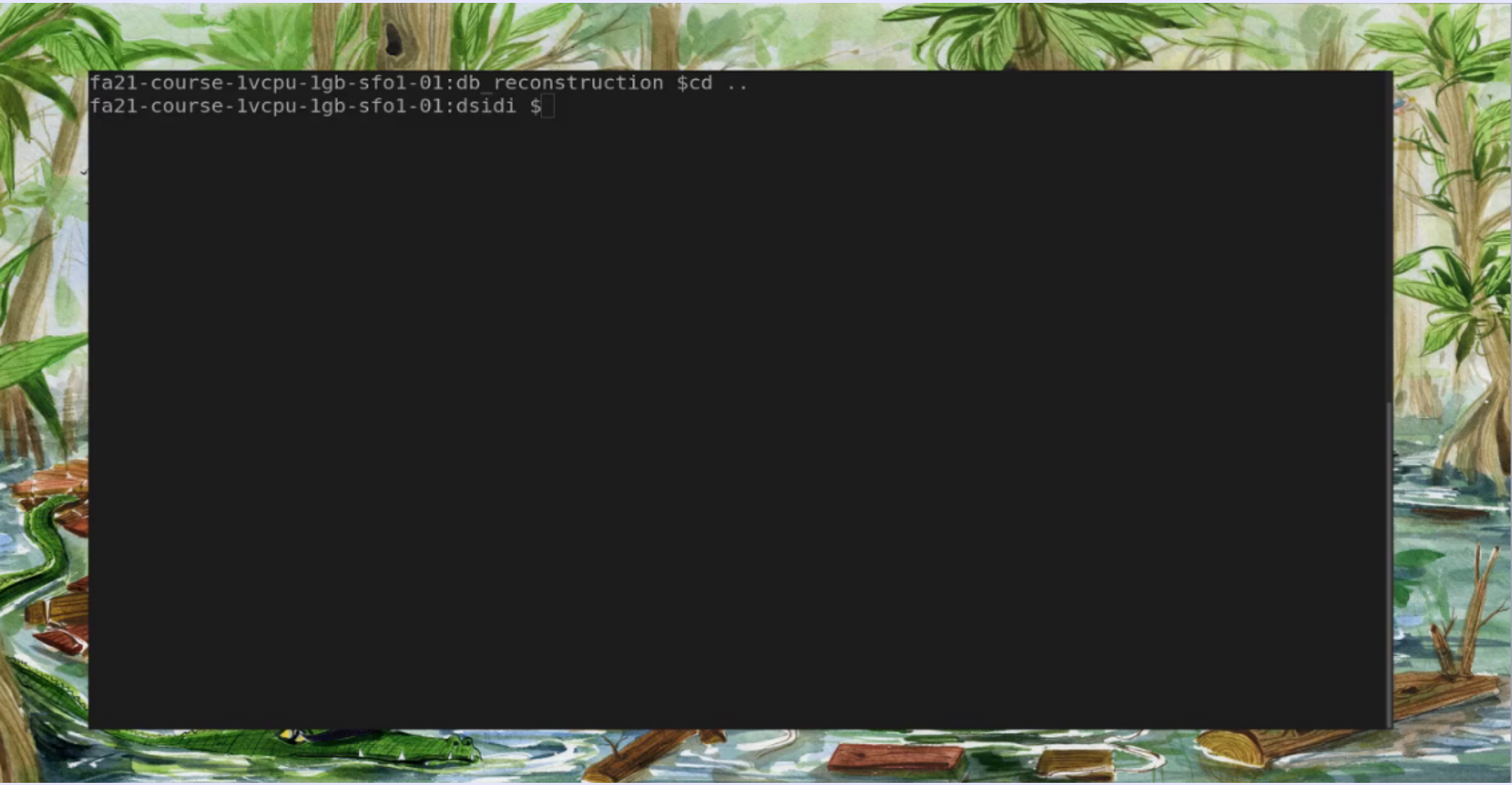
statistic	group	count	median	mean
1A	total population	7	37	35
2A	female	4	39.5	41
2B	male	3	25	27
2C	black or African American	3	40	48
2D	white	4	26.5	25.25
3A	single adults	1	42	42
3B	married adults	4	38.5	45.25
4A	black or African American male	2	32.5	32.5
4B	black or African American female	1	79	79
4C	white male	1	16	16
4D	white female	3	37	42.5
5A	persons under 5 years	0	0	0
5B	persons under 18 years	2	11	11
5C	persons 64 years or over	1	79	79

statistic	group	count	median	mean
1A	total population	7	37	35
2A	female	4	39.5	41
2B	male	3	25	27
2C	black or African American	3	40	48
2D	white	4	26.5	25.25
3A	single adults	(D)	(D)	(D)
3B	married adults	4	38.5	45.25
4A	black or African American male	(D)	(D)	(D)
4B	black or African American female	(D)	(D)	(D)
4C	white male	(D)	(D)	(D)
4D	white female	3	37	42.5
5A	persons under 5 years	(D)	(D)	(D)
5B	persons under 18 years	(D)	(D)	(D)
5C	persons 64 years or over	(D)	(D)	(D)

statistic	group	count	median	mean
1A	total population	7	37	35
2A	female	4	39.5	41
2B	male	3	25	27
2C	black or African American	3	40	48
2D	white	4	26.5	25.25
3A	single adults	(D)	(D)	(D)
3B	married adults	4	38.5	45.25
4A	black or African American male	(D)	(D)	(D)
4B	black or African American female	(D)	(D)	(D)
4C	white male	(D)	(D)	(D)
4D	white female	(D)	(D)	(D)
5A	persons under 5 years	(D)	(D)	(D)
5B	persons under 18 years	(D)	(D)	(D)
5C	persons 64 years or over	(D)	(D)	(D)

```
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;; Statistic 1A: n=7, median=37, mean=35
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;

;; Median age 37
;; -------------

;; The ages are sorted, so A4 must be 37.
(= A4 37)

;; mean age: 35

;; ====================
;; Q2. YOUR ANSWER HERE
;; ====================
```

Let’s do it!


fa21-course-1vcpu-1gb-sfo1-01:db_reconstruction $cd ..
fa21-course-1vcpu-1gb-sfo1-01:dsidi $

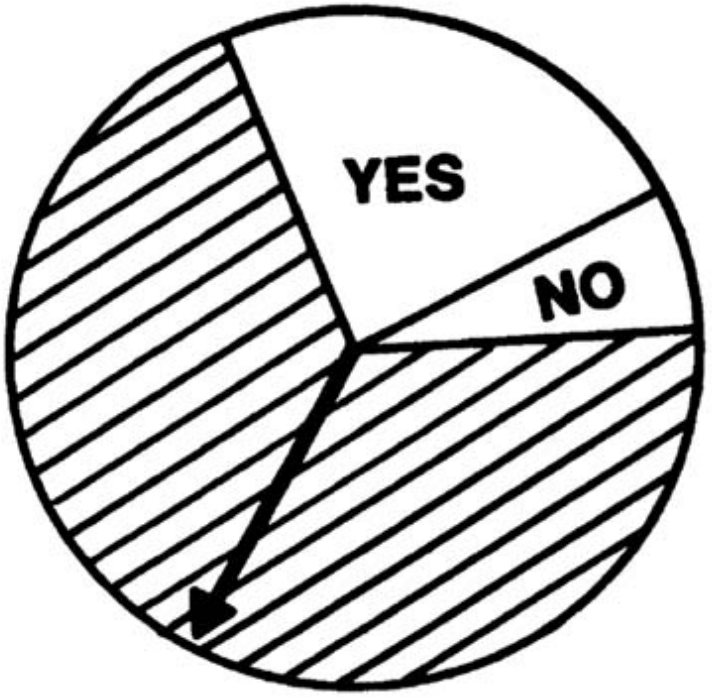
Name	Age	Gender	Zip Code	Smoker	Diagnosis
*	60–70	Male	191**	Y	Heart disease
*	60–70	Female	191**	N	Arthritis
*	60–70	Male	191**	Y	Lung cancer
*	60–70	Female	191**	N	Crohn's disease
*	60–70	Male	191**	Y	Lung cancer
*	***50–60***	***Female***	191**	N	HIV
*	50–60	Male	191**	Y	Lyme disease
*	50–60	Male	191**	Y	Seasonal allergies
*	***50–60***	***Female***	191**	N	Ulcerative colitis

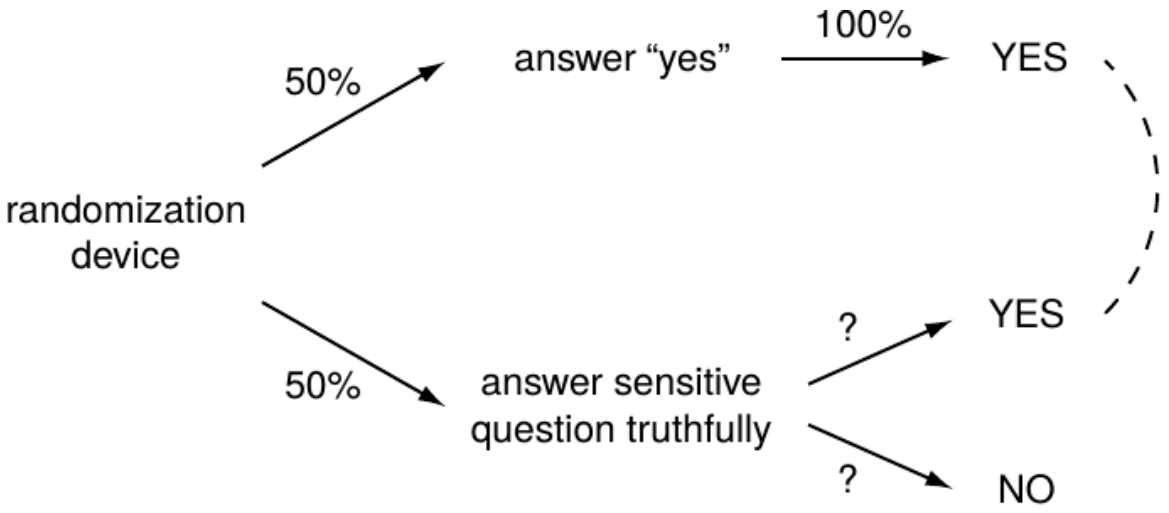
Name	Age	Gender	Zip Code	Diagnosis
*	***50–60***	***Female***	191**	HIV
*	***50–60***	***Female***	191**	Lupus
*	***50–60***	***Female***	191**	Hip fracture
*	60–70	Male	191**	Pancreatic cancer
*	60–70	Male	191**	Ulcerative colitis
*	60–70	Male	191**	Flu-like symptoms

from Kearns and Roth, *Ethical Algorithms*

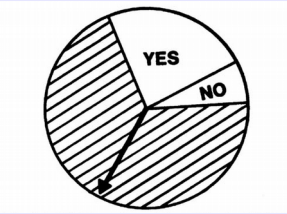
- Is a statistical database that resists microdata reconstruction today via suppression protected *forever*?
- What might protection depend upon over time?

“input perturbation”




randomization device
50%
answer "yes"
100%
YES
50%
answer sensitive question truthfully
?
YES
?
NO

- How might the following aspects of a Randomized Response Technique affect (1) reduction in privacy risk; and (2) utility of the data:
  - The size of the “tell the truth” region in the spinner (more generally, the probability that you’ll have to give a truthful answer in the protocol)



  - The mode of administration (online in a locked down browser, online in your own browser, on a computer in a lab setting, in person over video conference, in person IRL)
  - Whether the protocol for randomizing is a physical device, or a program you run locally on your computer, or on a remote server.

“subsampling”

- Occupation: Professional Basketball player (NBA)
- Height:
  - Muggsy Bogues: 5’3”
  - Manute Bol: 7’7”
- Occupation, year active, height is an indirect identifier for these two

	Date of birth	Height	…	Income
x	01/09/1965	63	…	1.5M

"Probability of Muggsy is low"

Unlikely that Muggsy matches anybody else

a match probably means the record came from Muggsy -- Muggsy is re-identified

	X_1 := Date of birth	X_2 := Height	…	X_d := Income	X_{d+1} := Diagnosis
$\mathbf{x}^{(1)}$					
$\mathbf{x}^{(2)}$					
⋮					
$\mathbf{x}^{(r)}$	01/09/1965	63	…	1.5M	Positive
⋮					
$\mathbf{x}^{(n\_D)}$					
⋮					
$\mathbf{x}^{(s)}$	01/09/1965	63	…	1.5M	Negative
⋮					
$\mathbf{x}^{(n)}$					

All the mitigation techniques we've seen so far transform the released data. What about handling this with just policy---leaving the collected microdata unchanged, but restricting access to it, and developing policies to govern its use?

What do you see as the strengths and weaknesses of this approach, as compared to techniques for protecting privacy by transforming released data?

Learn more

S. Garfinkel, J. M. Abowd, and C. Martindale, “Understanding database reconstruction attacks on public data,” *Commun. ACM*, vol. 62, no. 3, pp. 46–53, Feb. 2019, doi: 10.1145/3287287.

C. Dwork, “A firm foundation for private data analysis,” *Communications of the ACM*, vol. 54, no. 1, pp. 86–95, 2011.

L. Rocher, J. M. Hendrickx, and Y.-A. De Montjoye, “Estimating the success of re-identifications in incomplete datasets using generative models,” *Nature communications*, vol. 10, no. 1, pp. 1–9, 2019.

D. Sidi and J. Bambauer, “Plausible Deniability,” vol. 12276. Springer International Publishing, Cham, pp. 91–105, 2020. doi: 10.1007/978-3-030-57521-2_7.