Introduction to Data Privacy

CMSC 23200/33250, Winter 2022, Lecture 21

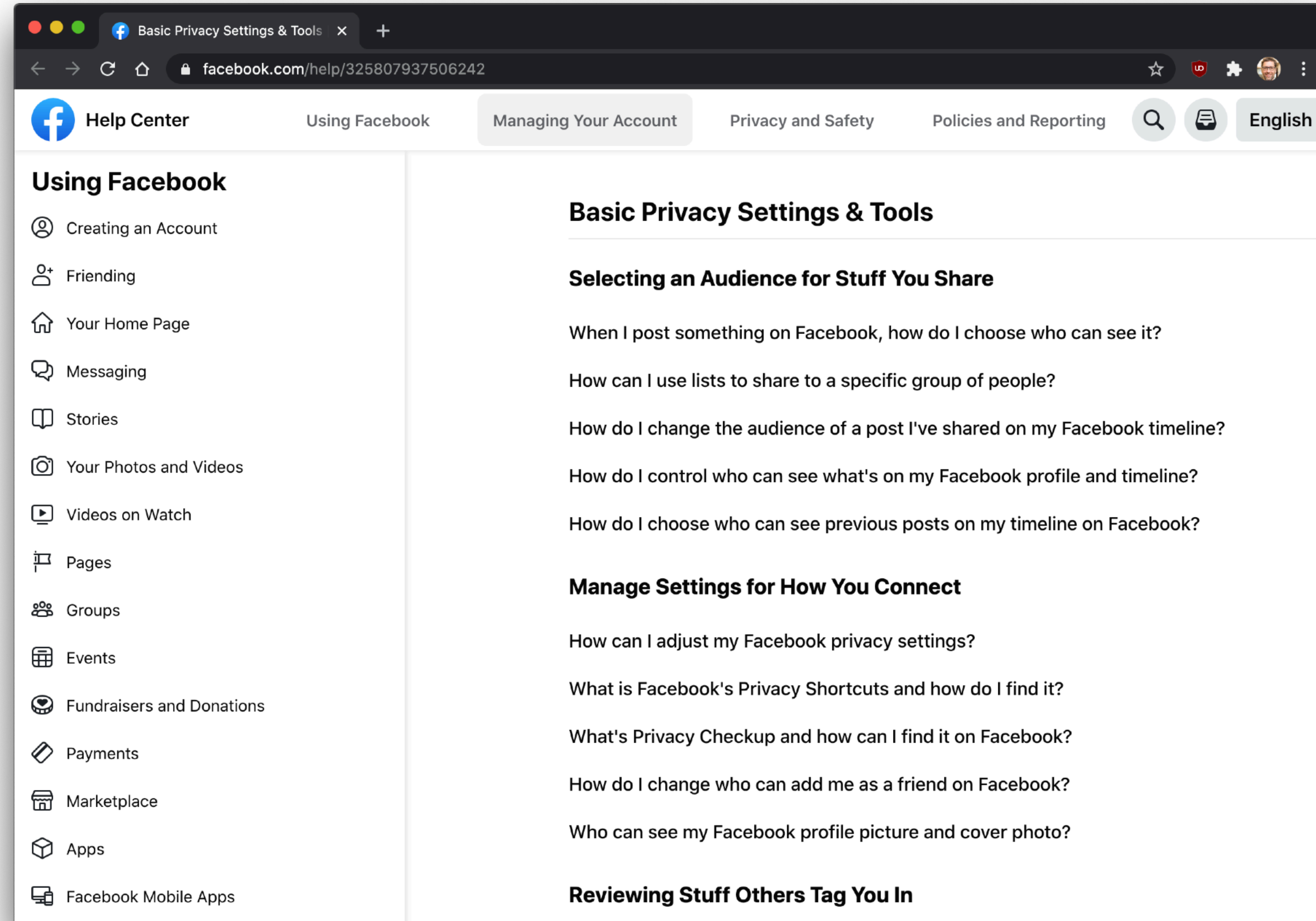
David Cash and Blase Ur

University of Chicago

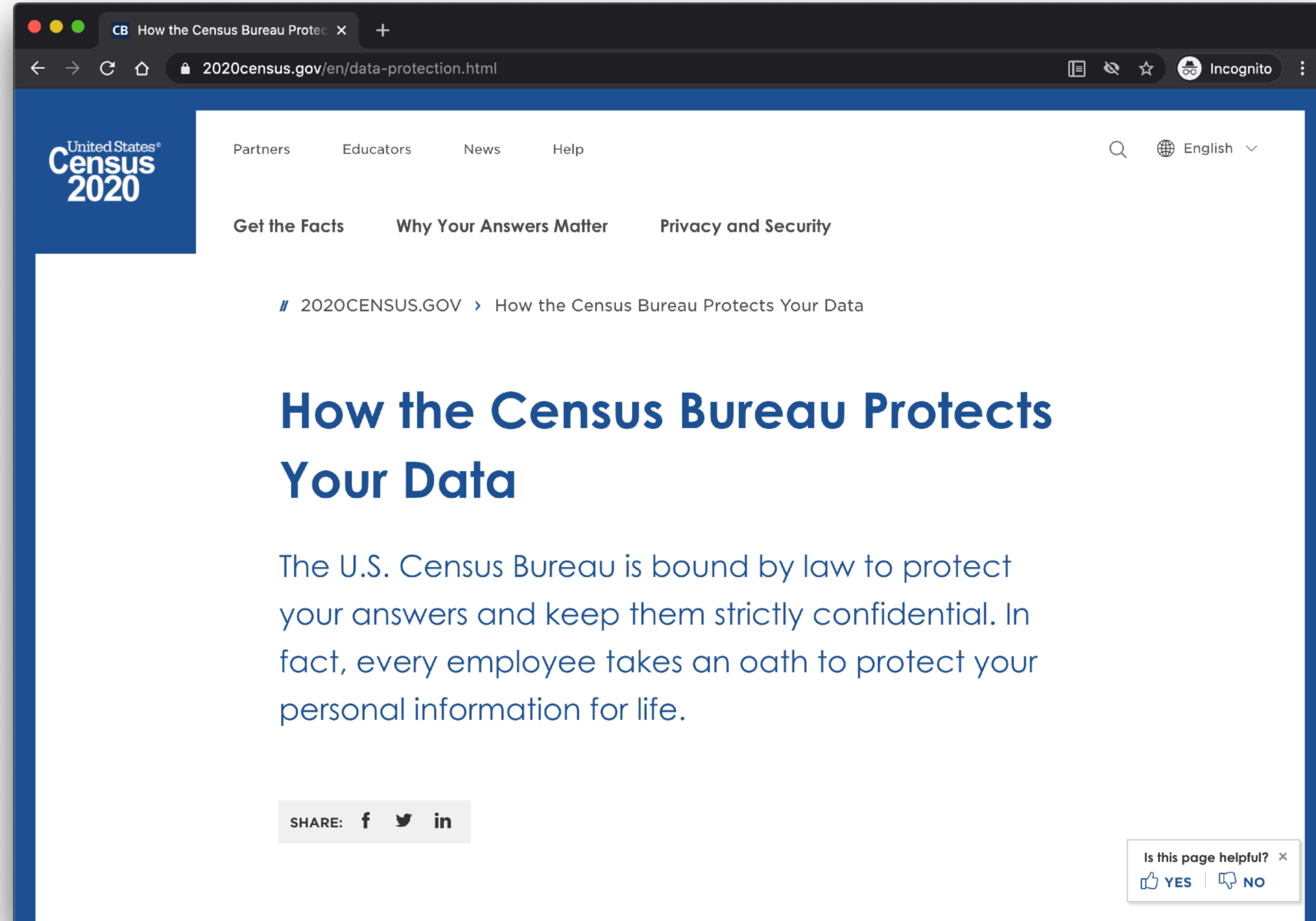
Outline

1. Problem setting for data privacy
2. Basic approaches to data privacy, and how to they fail
3. More advanced approaches, and how they also fail
4. A very interesting idea: Randomized Response

Privacy?


Basic Privacy Settings & Tools
facebook.com/help/325807937506242
Help Center
Using Facebook
Managing Your Account
Privacy and Safety
Policies and Reporting
English (
Using Facebook
Creating an Account
Friending
Your Home Page
Messaging
Stories
Your Photos and Videos
Videos on Watch
Pages
Groups
Events
Fundraisers and Donations
Payments
Marketplace
Apps
Facebook Mobile Apps
Basic Privacy Settings & Tools
Selecting an Audience for Stuff You Share
When I post something on Facebook, how do I choose who can see it?
How can I use lists to share to a specific group of people?
How do I change the audience of a post I've shared on my Facebook timeline?
How do I control who can see what's on my Facebook profile and timeline?
How do I choose who can see previous posts on my timeline on Facebook?
Manage Settings for How You Connect
How can I adjust my Facebook privacy settings?
What is Facebook's Privacy Shortcuts and how do I find it?
What's Privacy Checkup and how can I find it on Facebook?
How do I change who can add me as a friend on Facebook?
Who can see my Facebook profile picture and cover photo?
Reviewing Stuff Others Tag You In

Data Privacy



Privacy vs Security

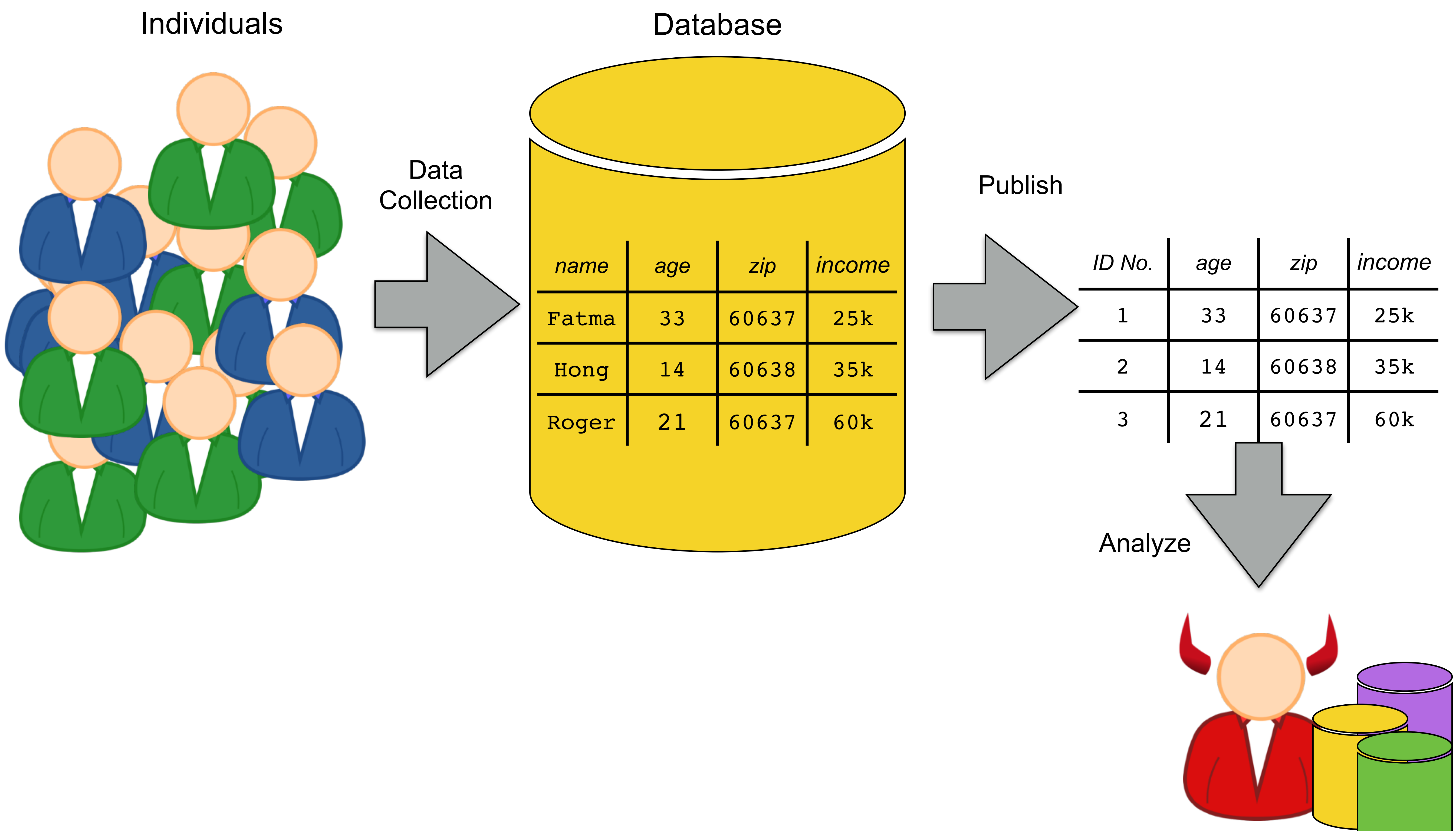
- *Privacy* is about individuals controlling how their personal data are collected, used, and published.

[Personal data is] any information relating to an identified or identifiable natural person.

- General Data Protection Regulation of the European Union

- *Security* is part of it. Confidentiality, authentication, authorization, and availability are ingredients.

Modern Data Privacy: Problem Setting in this Lecture



Examples

- Governments
- Medical research
- Financial/insurance companies
- Tech companies
- Advertisers
- Schools and Universities

Basic Data Privacy Mechanisms

- Simply enforce rules regulating data sharing and collection
- *De-identification:* Remove names, unique id numbers, addresses, etc
    - Health Insurance Portability and Accountability Act of 1996 (HIPAA)
    - Family Educational Rights and Privacy Act of 1974 (FERPA)
- *Segmentation:* Chop tables up vertically before publishing

name	*age*	*zip*	*income*
Fatma	33	60637	25k
Hong	14	60638	35k
Roger	21	60637	60k

Notable Privacy Failure #1: Mass. Grp Insurance (90s)

- Group Insurance Commission published info researchers (left circle)
- Sweeney purchased voter registration info from local government (right circle)
- "87% of the U.S. Population are uniquely identified by {date of birth, gender, ZIP}."

Latanya Sweeney

Source: Wikipedia

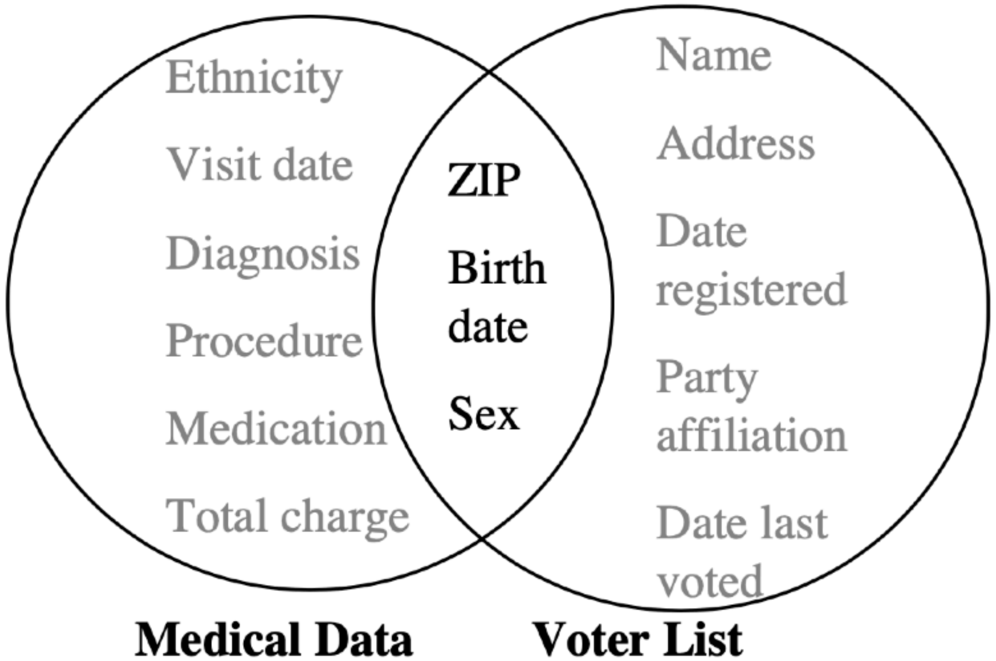
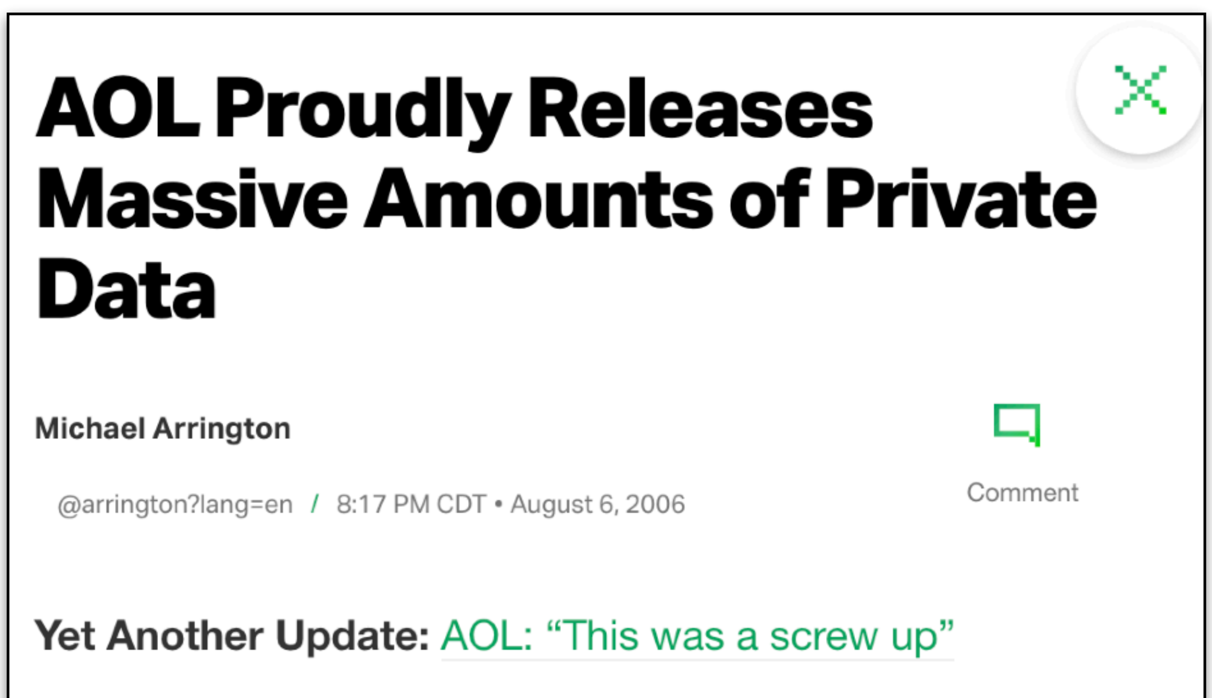


Figure 1 Linking to re-identify data

Source: L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

Notable Privacy Failure #2: AOL (2006)

- AOL publishes 20M search queries from 650k users.
- Names deleted, but query histories still associated with individuals



Source: xkcd

Notable Privacy Failure #2: AOL (2006)

A Face Is Exposed for AOL Searcher No. 4417749

By Michael Barbaro and Tom Zeller Jr.

Aug. 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

- Several individuals were identified. How would you guess?

Notable Privacy Failure #2: AOL (2006)

<u>User No. 4417749</u>

```
landscapers in Lilburn, Ga
John Arnold
numb fingers
Jenny Arnold
school supplies for Iraq children
60 single men
hand tremors
nicotine effects on the body
dog that urinates on everything
tea for good health
the best season to visit Italy
bipolar
safest place to live
…
```

Source: New York Times (previous slide)

Notable Privacy Failure #3: Netflix Prize (2006-2009)

- 2006: Netflix publishes movie rating data of 480K users
  - Meant to be used for recommendation system research
- Q from their FAQ: *"Is there any customer information in the dataset that should be kept private?"*
- *Netflix's answer:*

  *"**No, all customer identifying information has been removed; all that remains are ratings and dates.** This follows our privacy policy, which you can review here. Even if, for example, you knew all your own ratings and their dates you probably couldn't identify them reliably in the data because only a small sample was included (less than one-tenth of our complete dataset) and that data was subject to perturbation. **Of course, since you know all your own ratings that really isn't a privacy problem is it?"***

Notable Privacy Failure #3: Netflix Prize (2006-2009)

name	*Star Wars*	*Casablanca*	*Jurassic Park*	*<other movie>*
Fatma	★★★, 2/22/99	★★, 7/7/04	★, 8/17/03	★★★★, 8/22/00
Hong	★★, 5/6/02	★★★★★, 8/9/00	★★★, 6/16/03	★, 3/13/02
Roger	★★★★★, 4/29/98	★, 12/31/99	★★★★, 5/22/95	★, 4/29/00

- Idea: Cross-reference with IMDB
- Arvind+Vitaly: Knowing 8 ratings (w/dates) identifies 90% of users
- People rated movies on Netflix that they did not rate on IMDB.

Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

Source: Wikipedia

RYAN SINGEL SECURITY 03.12.2010 02:40 PM

NetFlix Cancels Recommendation Contest After Privacy Lawsuit

Netflix is canceling its second $1 million Netflix Prize to settle a legal challenge that it breached customer privacy as part of the first contest's race for a better movie-recommendation engine. Friday's announcement came five months after Netflix had announced a successor to its algorithm-improvement contest. The company at the time said it intended to [...]

Notable Privacy Failure #4: NYC Taxi Data (2014)

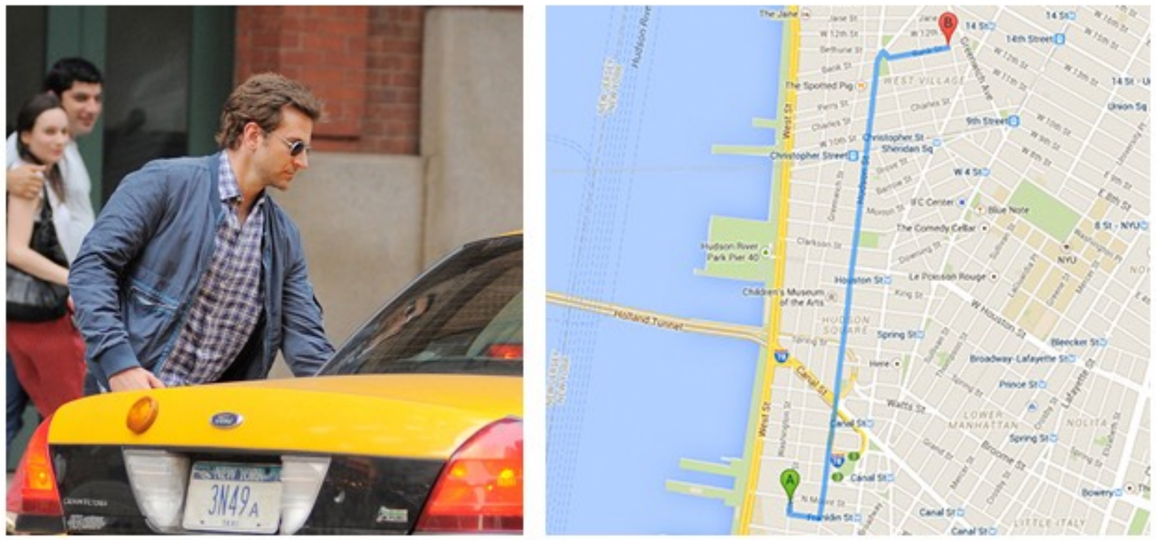
- NYC releases “anonymized” records of 173M taxi trips to researcher in response to Freedom of Information Act request
- Included start end location and time

10-02-14 | FAST FEED

NYC Taxi Data Blunder Reveals Which Celebs Don’t Tip–And Who Frequents Strip Clubs

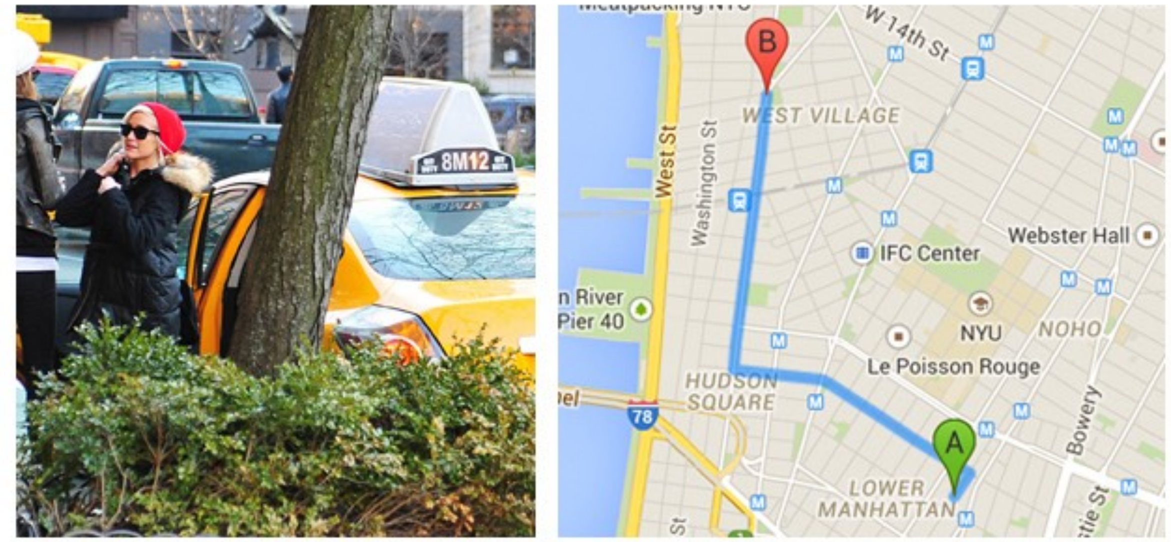
By cross-referencing de-anonymized trip data with paparazzi photos, a privacy research could tell how much Bradley Cooper paid his driver.

Notable Privacy Failure #4: NYC Taxi Data (2014)



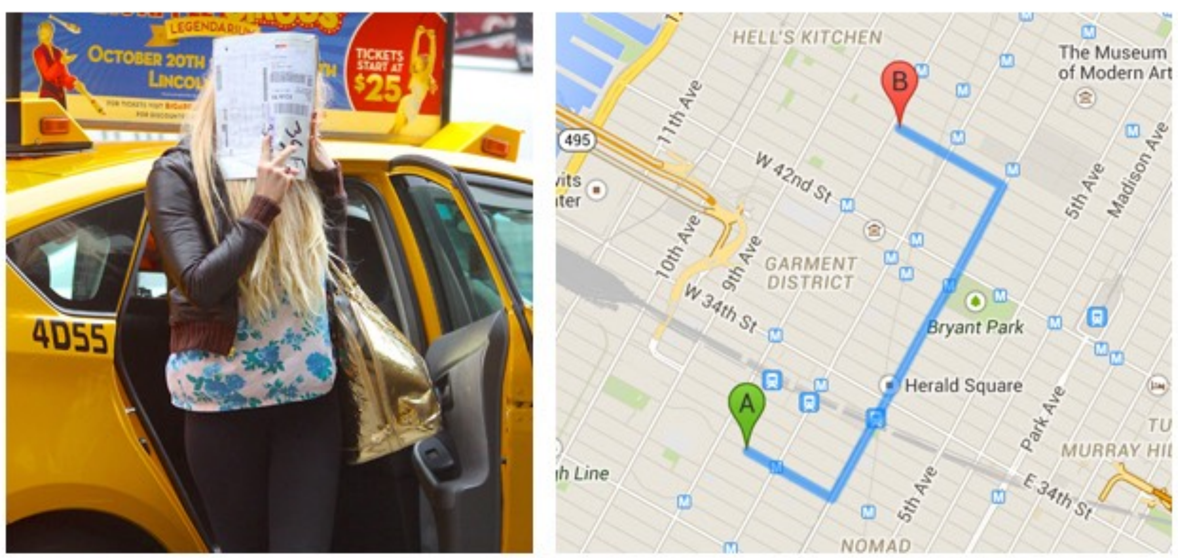
BRADLEY COOPER

JULY 8, 2013 • 7:34 PM - 7:44 PM
376 GREENWICH ST. TO 13 BANK ST.
$9.00 FARE • CASH; UNKNOWN TIP • ©SPLASH



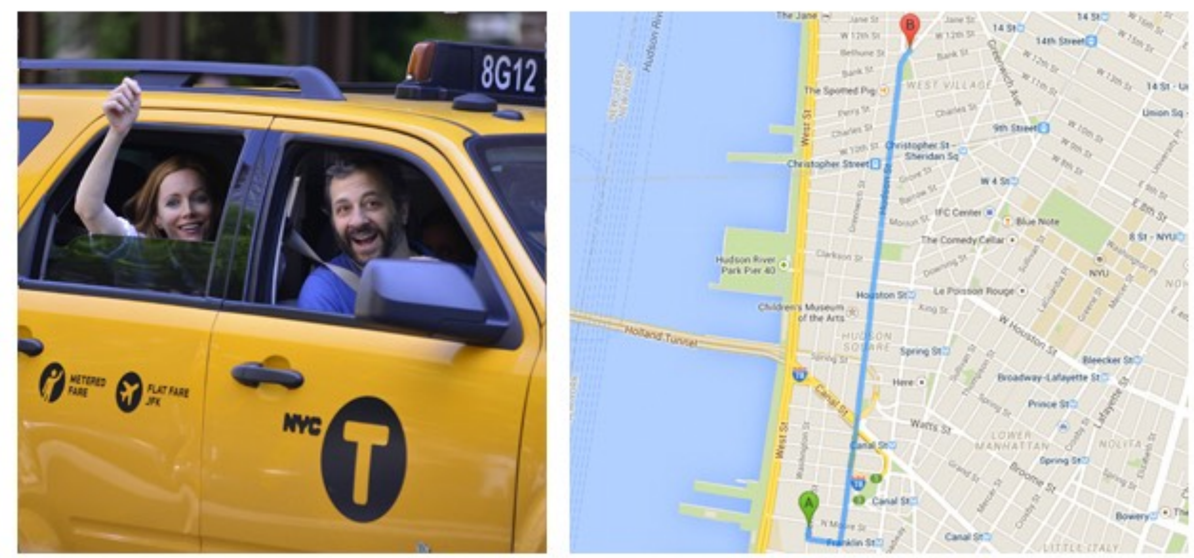
ASHLEE SIMPSON

JANUARY 6, 2013 • 3:29 PM - 3:38 PM
78 CROSBY ST. TO 580 HUDSON ST.
$7.50 FARE • $2 TIP • ©SPLASH



AMANDA BYNES

APRIL 11, 2013 • 5:43 PM - 6:02 PM
229 W 28TH ST. TO 271 W 47TH ST.
$13 FARE • CASH; UNKNOWN TIP • ©SPLASH



JUDD APATOW
LESLIE MANN

JUNE 21, 2013 • 11:28 AM - 11:35 AM
376 GREENWICH ST. TO 1 ABINGDON SQUARE
$7.00 FARE • $2.10 TIP • ©SPLASH

Source: https://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546

- Also: Dataset had taxi ID replaced with md5(taxiID)…

c.f. https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1

Privacy Failures: Why is this so hard?

- Hard to anticipate how individuals might be harmed
- Hard to anticipate what *side information* is available for linking
- Hard to anticipate what adversarial strategies might exist

Latanya Sweeney

Source: Wikipedia

- Sweeney: Take a principled approach!

1. Give precise *definition* of "sufficiently sanitized" data
2. Design sanitization methods that output data meeting definition.

Towards Modern Protection: k-Anonymity

<u>Definition</u>: A table is <u>*k-anonymous with respect to columns $C_1, \ldots C_n$*</u> if whenever a value ($v_1, \ldots, v_n$) appears for those columns in *some* row, it appears in at least k rows.

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Figure 2 Example of *k*-anonymity, where *k*=2 and QI={*Race*, *Birth*, *Gender*, *ZIP*}

Adapted from: L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

Processing Data/Queries for k-Anonymity

- Aggregate numerical columns. Generalize or redact others.

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Fig. 1. Inpatient Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

Fig. 2. 4-Anonymous Inpatient Microdata

- NP-Hard (i.e. intractable) to do “optimally”

Source: A. Machanavajjhala et al. l-Diversity: Privacy Beyond k-Anonymity. TKDD 2007.

Problems with k-Anonymity: Homogeneity Attack

- If I know your Zip Code is 13053 and that you are in your 30s….

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

Fig. 2. 4-Anonymous Inpatient Microdata

Source: A. Machanavajjhala et al. l-Diversity: Privacy Beyond k-Anonymity. TKDD 2007.

Problems with k-Anonymity: Background Knowledge

- If I know your Zip Code is 13068, that you're 21 years old, and that you seem pretty healthy generally…

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

Fig. 2. 4-Anonymous Inpatient Microdata

Source: A. Machanavajjhala et al. l-Diversity: Privacy Beyond k-Anonymity. TKDD 2007.

Another attempt: L-Diversity

Definition: A table is *L-anonymous with respect to columns C_1, … C_n and sensitive column C^** if whenever a value (v_1, …, v_n) appears for columns C_1, … C_n, at least L different values appear in C^* in those rows.

(Note: actual definitions in paper cited below are more nuanced.)

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

Fig. 4. **3-Diverse** Inpatient Microdata

- Ensure that sensitive columns are “well represented” to defeat both attacks

Source: A. Machanavajjhala et al. l-Diversity: Privacy Beyond k-Anonymity. TKDD 2007.

Attacking L-Diversity

- Correlations still lead to violations even with diversity

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

Table 4. A 3-diverse version of Table 3

- Another patch suggested: t-Closeness, but conclusion is unclear

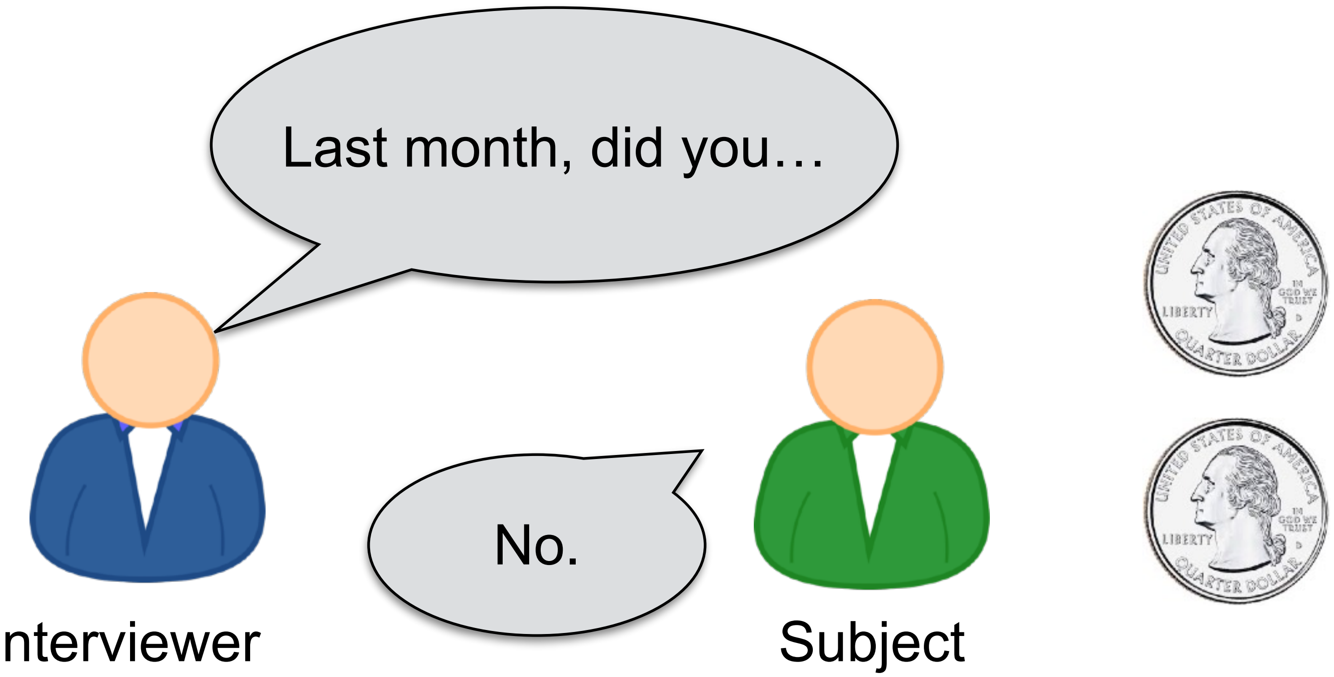
Source: N. Li et al. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. ICDE 2007.

Back to the 1960's (and then to the '00s next lecture)

- Want to survey a population about engaging in an embarrassing or illegal behavior X (e.g. X=drug use, X=cheating, …)
- Not interested in individuals. Only want to know fraction of the population.
- Discussion: what's wrong with just interviewing people and asking

"Did you engage in X in the last month?"

Profound Idea: Randomized Response



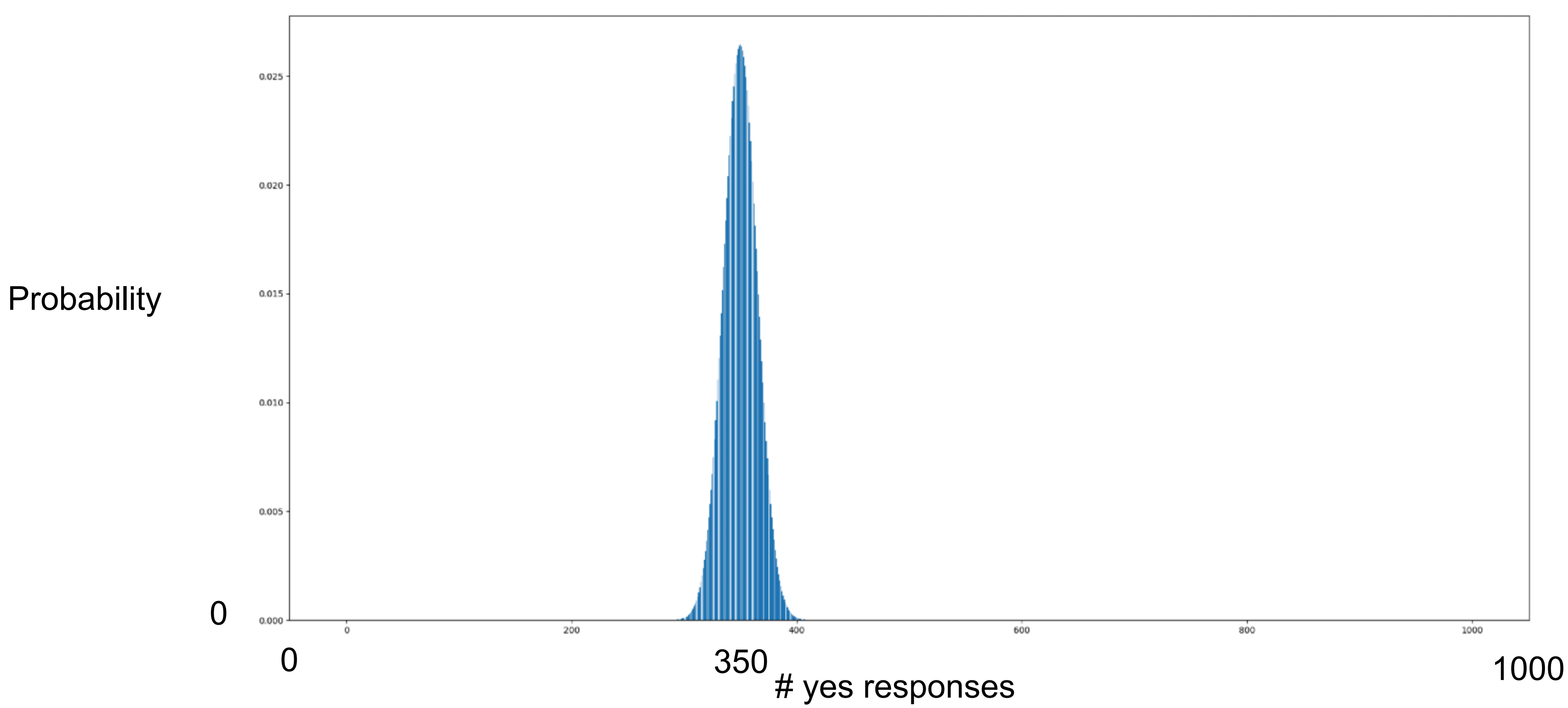
<u>Instructions for subject:</u>

1. Privately flip coins C_A and C_B
2. If C_A = Heads: Answer truthfully
3. Else: Answer randomly (use C_B)

Randomized Response: Example

- Suppose population is 1000.
- 200 engage in behavior and 800 do not.
- Expect to get 350 “yes” answers:

$$0.25 \cdot 800 + 0.50 \cdot 200 + 0.25 \cdot 200 = 350$$



Analyzing Randomized Response Data

<u>Claim</u>: If p-fraction of population engages in behavior ($0 \leq p \leq 1$), then expected proportion that say "Yes" is

$$y = 0.25(1 - p) + p(0.50 + 0.25)$$

- Measure y, then solve: $p = 2(y - 0.25)$

Randomized Response and Plausible Deniability

- High school students surveyed on drug use.
- Higher reported use on all drugs except hallucinogens (?)

Drug Use in Preceding Three Months: Means and Standard Errors

Drug category	Combined "6 + 1 item"		Randomized response procedure	
	μ	SE	μ	SE
All subjects				
Alcohol	10.63	3.697	18.79	13.019
Cannabis	3.68	0.779	3.04	1.329
Hallucinogens	0.35	0.174	0.26	0.134
Amphetamines ("speed")	0.11	0.048	0.43	0.200
Tranquilizers	0.26	0.097	0.81	0.232
Heroin	0.06	0.031	0.33	0.145
Excluding responses in excess of 100[a]				
Alcohol	5.19	0.420	10.98	3.393
Cannabis	3.01	0.618	3.51	1.244

[a] Hallucinogens, amphetamines ("speed"), tranquilizers and heroin were unaffected by this transformation.

Source: M. Goodstadt and V. Gruson. The Randomized Response Technique: A Test on Drug Use. J Amer Stat Assoc, 1975.

Changing Randomized Response

- How would you feel about using these instead?

```
Instructions for subject:
1. Privately roll a 6-sided die DA.
2. Privately toss a fair coin CB.
2. If DA = 1: Answer truthfully
3. Else: Answer randomly (use CB)
```

```
Instructions for subject:
1. Privately roll a 100-sided die DA.
2. Privately toss a fair coin CB.
2. If DA = 1: Answer truthfully
3. Else: Answer randomly (use CB)
```

The End