Lecture 5: Privacy

CS 181S

Fall 2020

Confidentiality
Integrity
Availability

CIA

What is Privacy?

What is Privacy?

Cambridge Analytica

Privacy

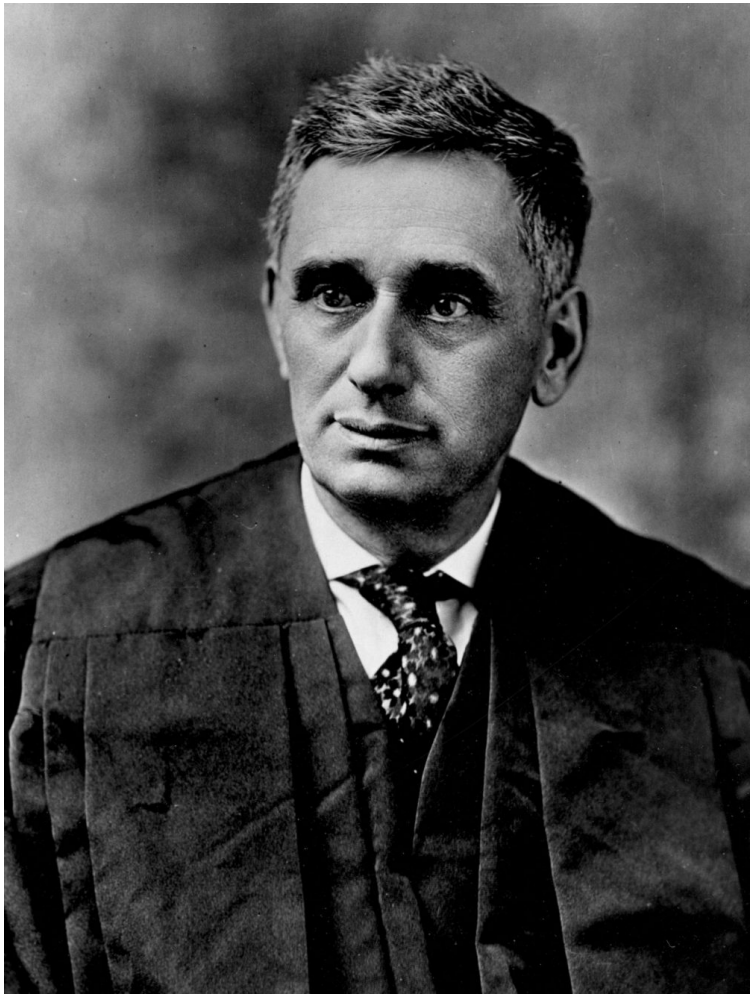
Privacy concerns information about individuals (people, organizations, etc.)

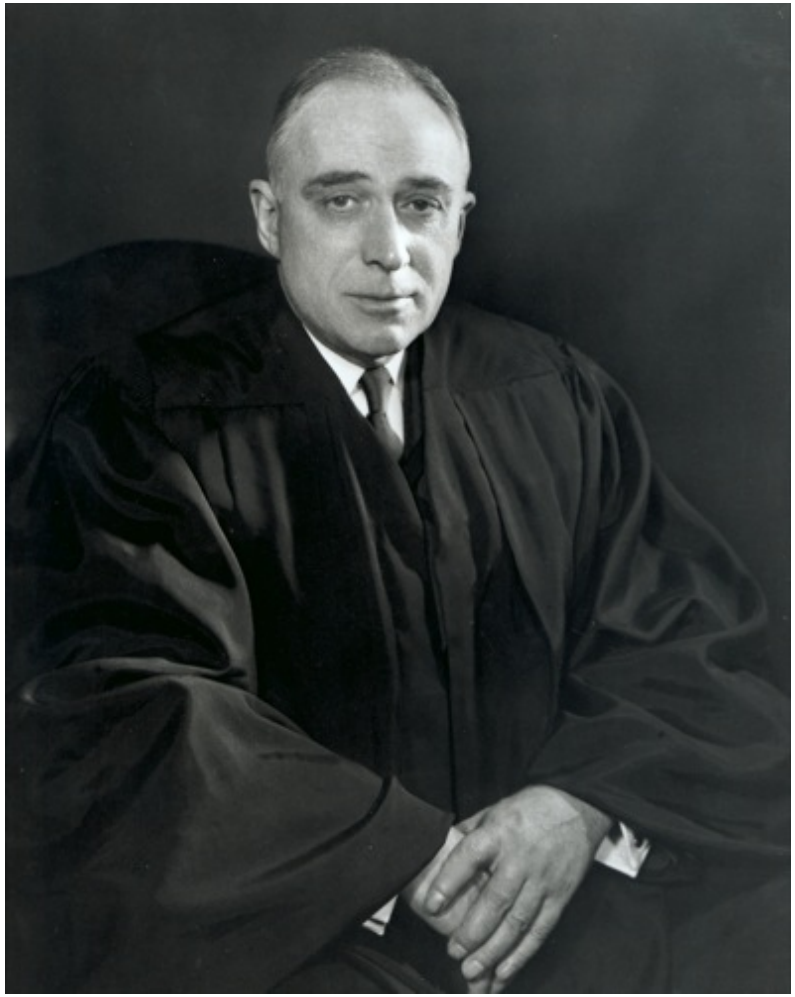
- Often construed as legal right
- *Privacy* is not a synonym for confidentiality or for secrecy

Exercise 1: What is a privacy violation?

1. Police read papers stored in your home
2. Police read papers you threw in the trash
3. Police read your medical records
4. Your parents read your medical records
5. Pomona uses your medical records in a research study
6. Police read your social media posts
7. Police read your emails
8. Google employee reads your emails
9. Google uses your emails to target personalized ads
10. Someone tracks your location for months (using phone)

Privacy in American Law





Contextual Integrity

- defines privacy relative to appropriate context
- considers information type, time, location, purpose, principals involved (subject, sender, receiver)
- dependent on social norms
- norms can change over time

General Guidelines

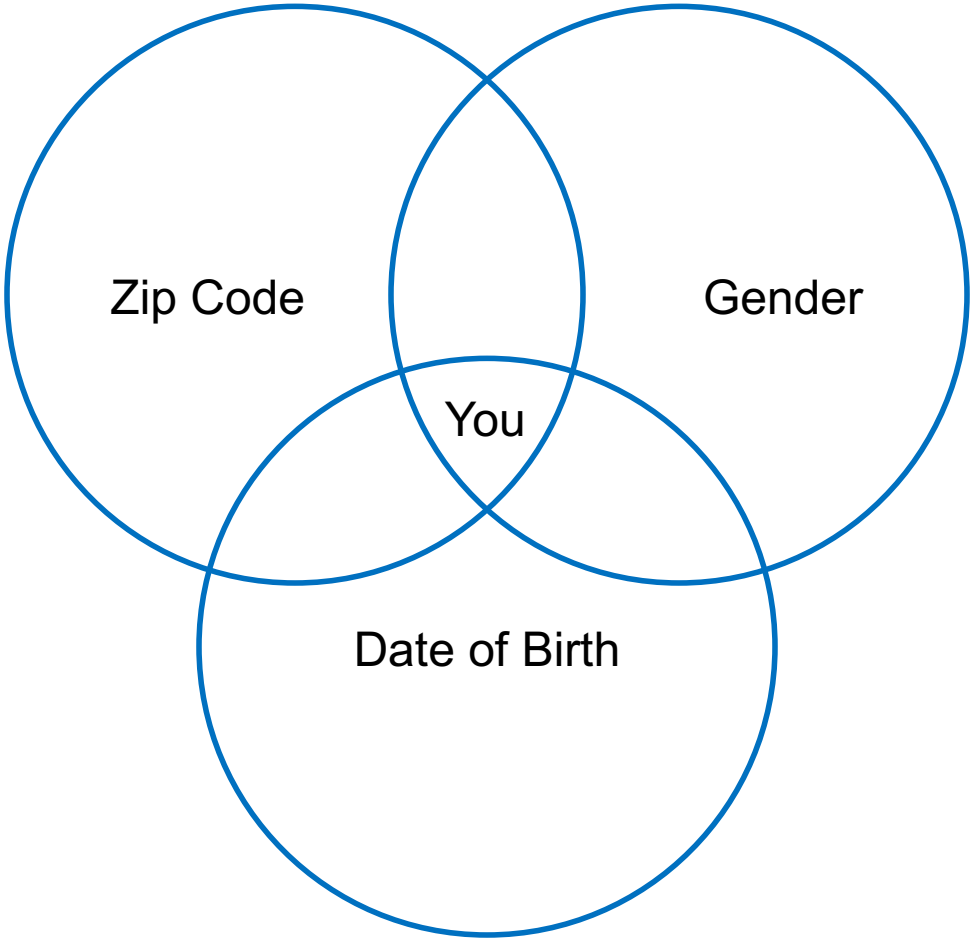
The FTC's Fair Information Practice Principals (FIPPs) are the most broadly recognized guidelines for handling private data in information systems

- Seek consent
- Minimize data use
- Limit storage
- Avoid linking

HIPAA

- Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information.
- A covered entity may determine that health information is not individually identifiable health information only if:
  - (1) An expert determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; or
  - (2) It removes names, geographic subdivisions smaller than a state, all elements of dates (except year), telephone and fax numbers, email addresses, SSNs, UIDs, URLs, IP addresses, biometric identifiers, full face photos,

Deanonymization



Deanonymization

k-Anonymity

Name	Pronouns	Year	Grade
Alice	she/her	2020	95
Bob	he/him	2020	80
Charlie	they/them	2020	95
David	he/him	2020	60
Edward	he/him	2021	80
Flora	she/her	2021	99
Georgia	she/her	2021	60

- **Quasi-identifiers (QIs)** are sets of attributes that can be exploited for linking
- A database is **k-anonymous** if each QI maps to at least k different individuals
- Techniques: suppression and generalization

Exercise 2: k-anonymity

- Modify this dataset to make it 2-anonymous with respect to Race/DOB/Sex

Race	DOB	Sex	Marital Status	Health Issues
asian	9/27/00	female	divorced	hypertension
asian	9/30/00	female	divorced	obesity
asian	4/18/00	male	married	chest pain
asian	4/15/00	male	married	obesity
black	3/13/99	male	married	hypertension
black	3/18/99	male	married	shortness of breath
black	9/13/00	female	married	shortness of breath
black	9/07/00	female	married	obesity
white	5/14/01	male	single	chest pain
white	4/08/01	male	single	obesity
white	9/15/01	female	married	shortness of breath

Exercise 2: k-anonymity

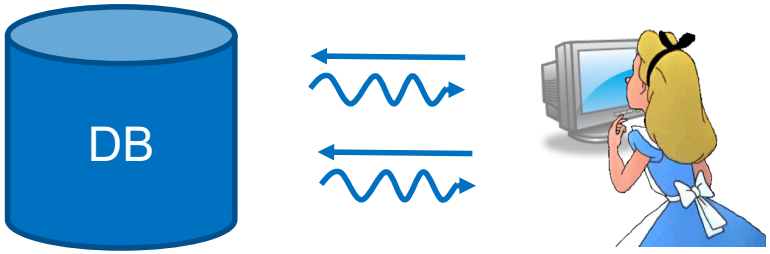
- Modify this dataset to make it 2-anonymous with respect to Race/DOB/Sex

Race	DOB	Sex	Marital Status	Health Issues
asian	9/00	female	divorced	hypertension
asian	9/00	female	divorced	obesity
asian	4/00	male	married	chest pain
asian	4/00	male	married	obesity
black	3/99	male	married	hypertension
black	3/99	male	married	shortness of breath
black	9/00	female	married	shortness of breath
black	9/00	female	married	obesity
white	2001	*	single	chest pain
white	2001	*	single	obesity
white	2001	*	married	shortness of breath

Database Privacy

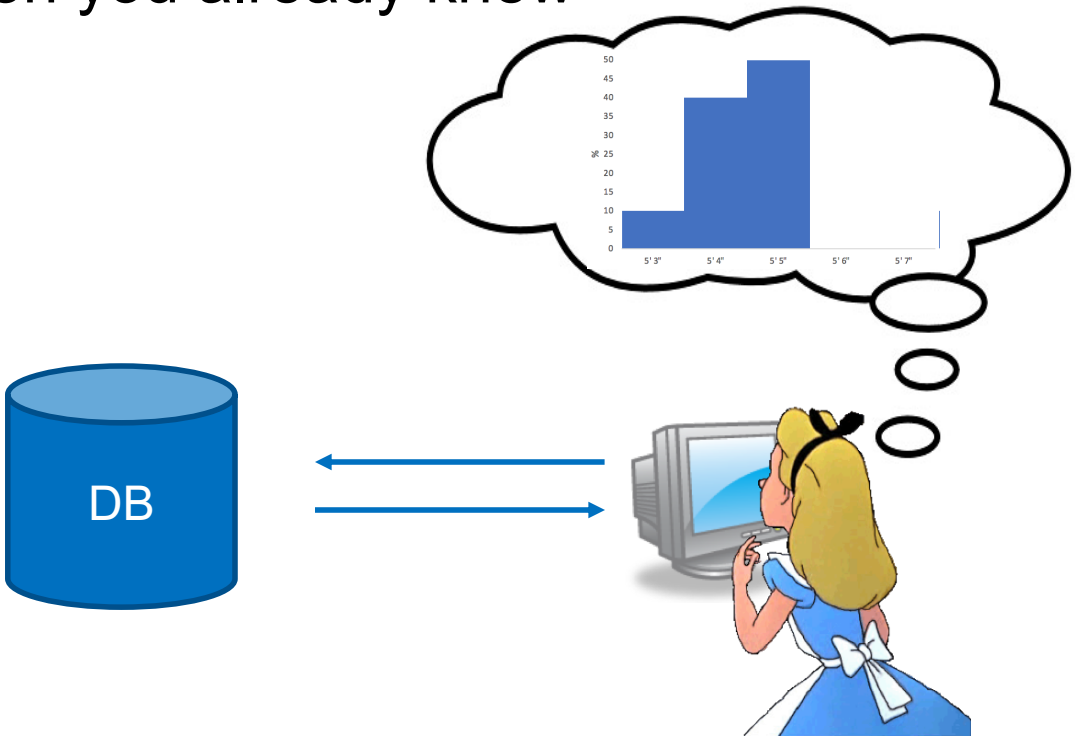
Offline Privacy

Online Privacy



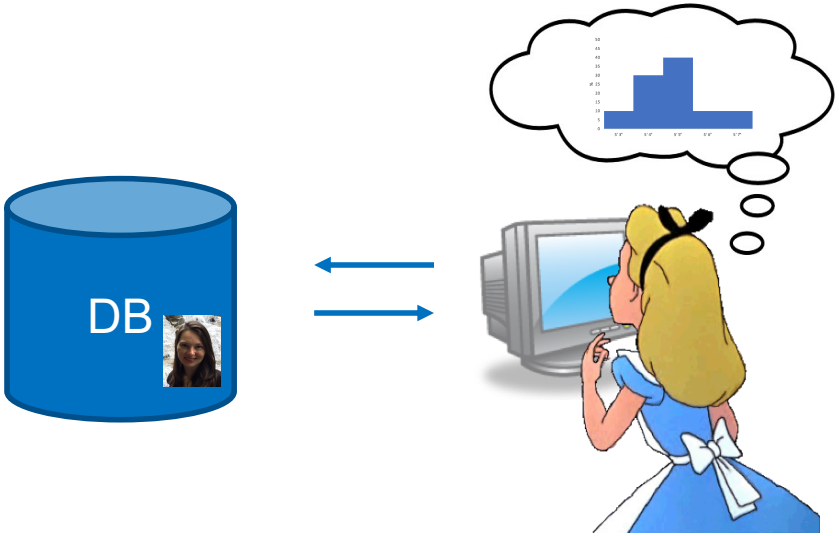
Defining Privacy: Try #1

- You don't know anything more after interacting with the database then you already knew

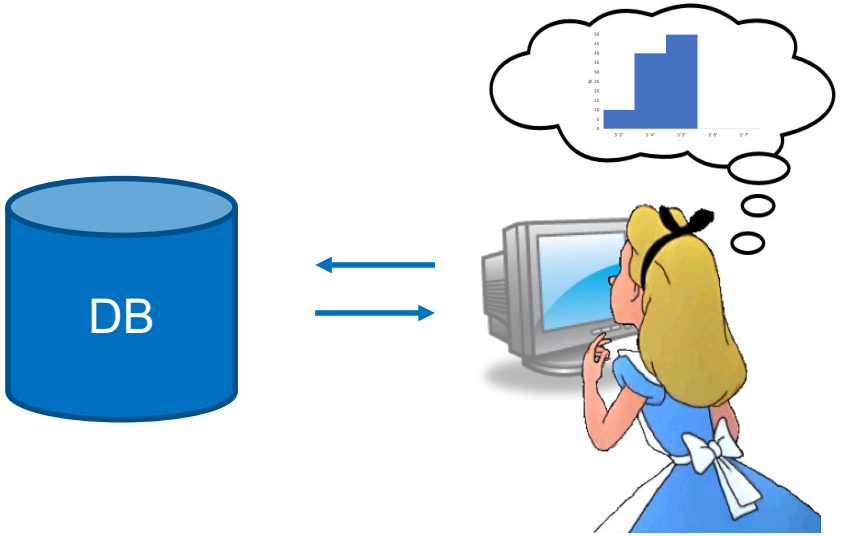


Differential Privacy

World 1



World 2



A query Q is ϵ-differentially private if $\forall D, r \in D$,
$$\Pr[Q(D) = x] \leq e^{\epsilon} \cdot \Pr[Q(D - r) = x]$$

Sensitivity

- The sensitivity Δ of a query Q is the maximum the answer to Q can possibly change between two databases that differ only by one person

- Q = number of people taller than 6 ft $\Delta = 1$
- Q = maximum height of a person $\Delta = 48$

Exercise 3: Sensitivity

- Assume you have a database containing the heights of 100 users specified in inches. You may assume that all heights are between 48 in and 96in.

- What is the sensitivity of the following queries?
  1. The number of people who are 5' 4"
  2. The median height in the dataset
  3. The mean height in the dataset

Exercise 3: Sensitivity

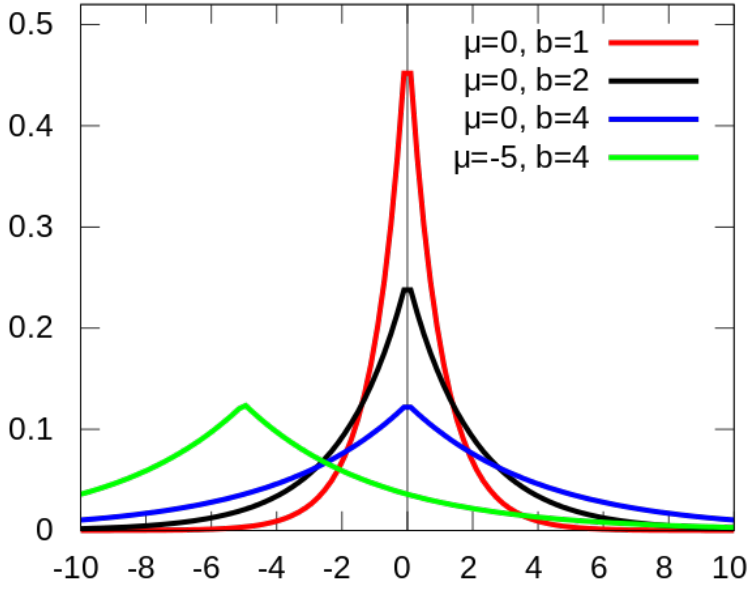
- Assume you have a database containing the heights of 100 users specified in inches. You may assume that all heights are between 48 in and 96in.

- What is the sensitivity of the following queries?
  1. The number of people who are 5' 4" $\Delta = 1$
  2. The median height in the dataset $\Delta = 48$
  3. The mean height in the dataset $\Delta = .48$

Laplacian Distribution

- Lap(b) is the probability distribution with the property that

$$\Pr[\, Lap(b) = x] = \frac{1}{2b} \cdot e^{-\frac{|x|}{b}}$$



Laplacian Mechanism

- Given a query Q on a database D that has sensitivity Δ, respond with Q(D)+Y where Y is drawn from the distribution $Lap(\frac{\Delta}{\epsilon})$
- Theorem: this mechanism satisfies ϵ-differential privacy

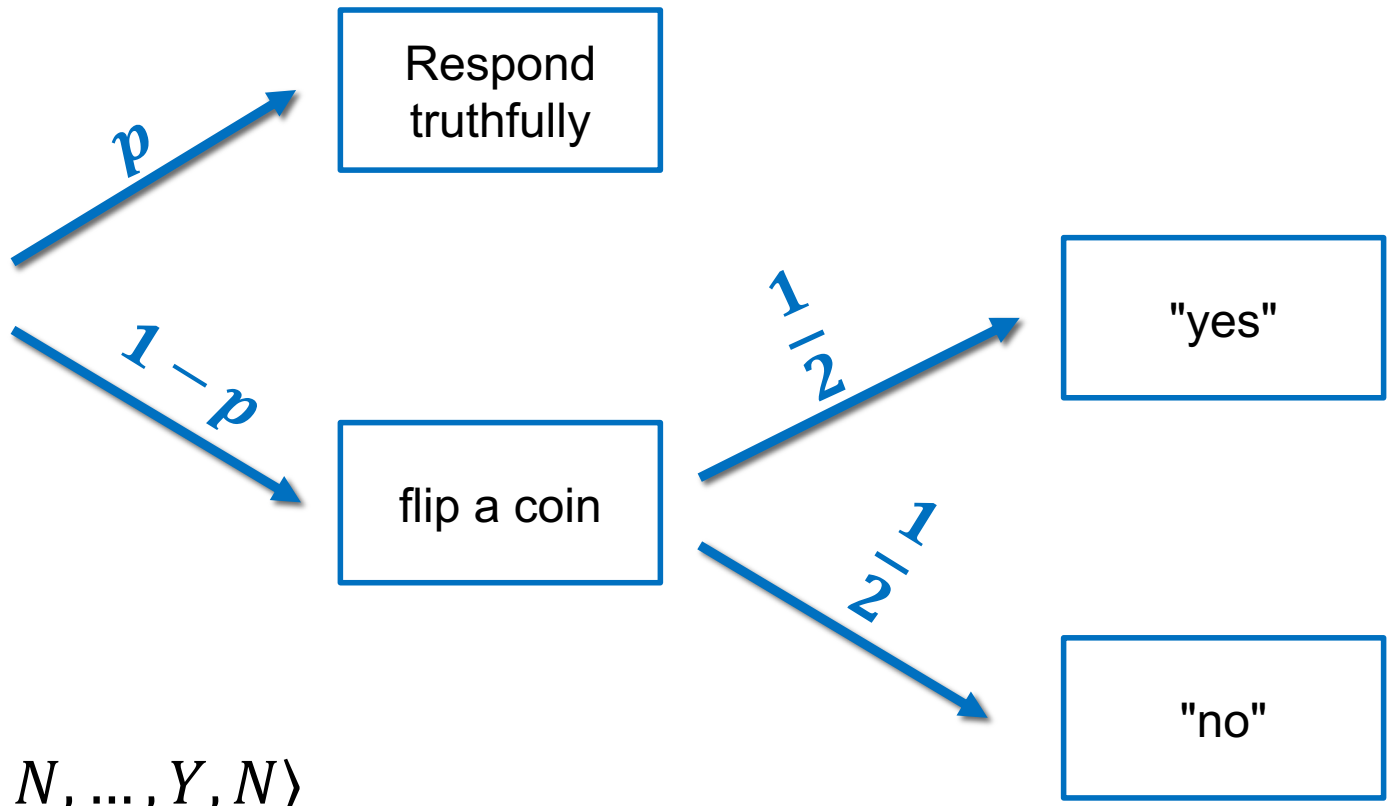
$$\frac{\Pr[\,Q(D)\,+\,Y\,=\,x]}{\Pr[Q(D-r)\,+\,Y\,=\,x]} = \frac{\Pr[\,Y\,=\,x\,-\,Q(D)]}{\Pr[Y\,=\,x\,-\,Q(D-r)]} = \frac{\frac{1}{2(\Delta/\epsilon)}\cdot e^{-\frac{|x\,-Q(D)|}{\Delta/\epsilon}}}{\frac{1}{2(\Delta/\epsilon)}\cdot e^{-\frac{|x\,-Q(D-r)|}{\Delta/\epsilon}}} = \frac{e^{-\frac{|x\,-Q(D)|}{\Delta/\epsilon}}}{e^{-\frac{|x\,-Q(D-r)|}{\Delta/\epsilon}}}$$

$$= e^{\frac{|x\,-Q(D-r)|}{\Delta/\epsilon}-\frac{|x\,-Q(D)|}{\Delta/\epsilon}} = e^{\left(\frac{\epsilon}{\Delta}\right)\cdot(|x\,-Q(D-r)|\,-|x\,-Q(D)|)}$$

$$\leq e^{\left(\frac{\epsilon}{\Delta}\right)\cdot(|x\,-Q(D-r)\,-x+Q(D)|\,)} = e^{\left(\frac{\epsilon}{\Delta}\right)\cdot(|Q(D)-Q(D-r)\,|\,)}$$

$$\leq e^{\left(\frac{\epsilon}{\Delta}\right)\cdot\Delta} = \mathrm{e}^{\epsilon}$$

Randomized Response



$\langle Y, N, \boldsymbol{Y}, Y, Y, N, \ldots, Y, N \rangle$

Theorem: this mechanism satisfies ϵ-differential privacy

Randomized Response

- Theorem: this mechanism satisfies ϵ-differential privacy

$$\frac{\Pr[\langle Y, N, \boldsymbol{Y}, Y, Y, N, \ldots, Y, N\rangle \mid f(\mathrm{Bob}) = Y]}{\Pr[\langle Y, N, \boldsymbol{Y}, Y, Y, N, \ldots, Y, N\rangle \mid f(\mathrm{Bob}) = N]}$$

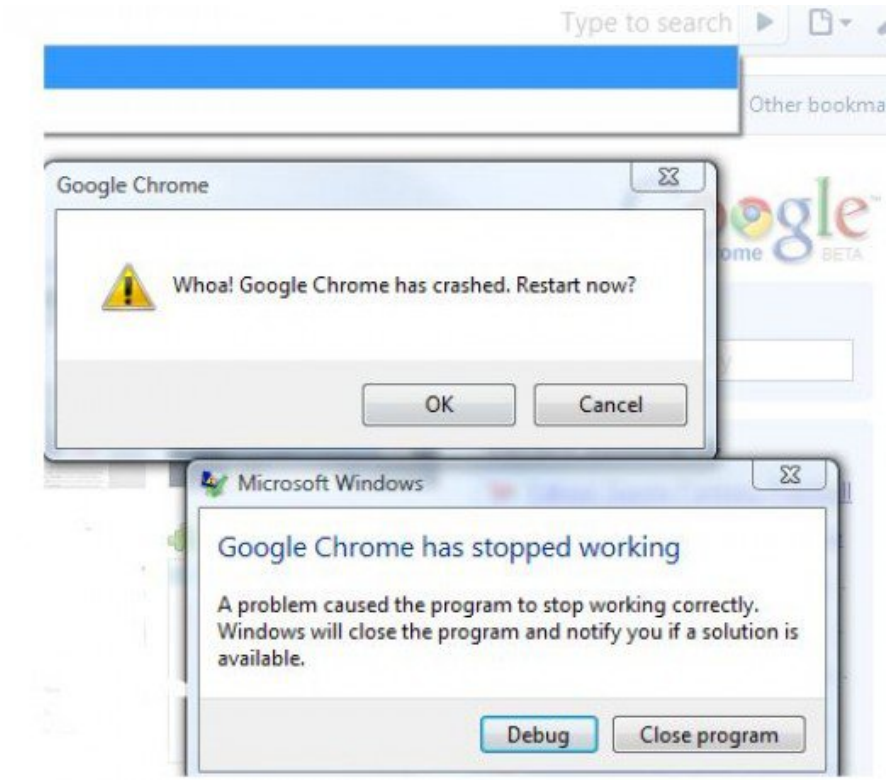
$$= \frac{\Pr[Y \mid f(P_0)] \cdot \Pr[\mathrm{N} \mid f(P_1)] \cdot \Pr[Y \mid f(Bob) = Y] \cdot \ldots \cdot \Pr[\mathrm{N} \mid f(P_{n-1})]}{\Pr[Y \mid f(P_0)] \cdot \Pr[\mathrm{N} \mid f(P_1)] \cdot \Pr[Y \mid f(Bob) = N] \cdot \ldots \cdot \Pr[\mathrm{N} \mid f(P_{n-1})]}$$

$$= \frac{\Pr[Y \mid f(Bob) = Y]}{\Pr[Y \mid f(Bob) = N]}$$

$$= \frac{p \cdot 1 + (1-p) \cdot \frac{1}{2}}{p \cdot 0 + (1-p) \cdot \frac{1}{2}} = \frac{(1+p) \cdot \frac{1}{2}}{(1-p) \cdot \frac{1}{2}} = \frac{(1+p)}{(1-p)}$$

$$= \mathrm{e}^{\ln(\frac{1+p}{1-p})}$$

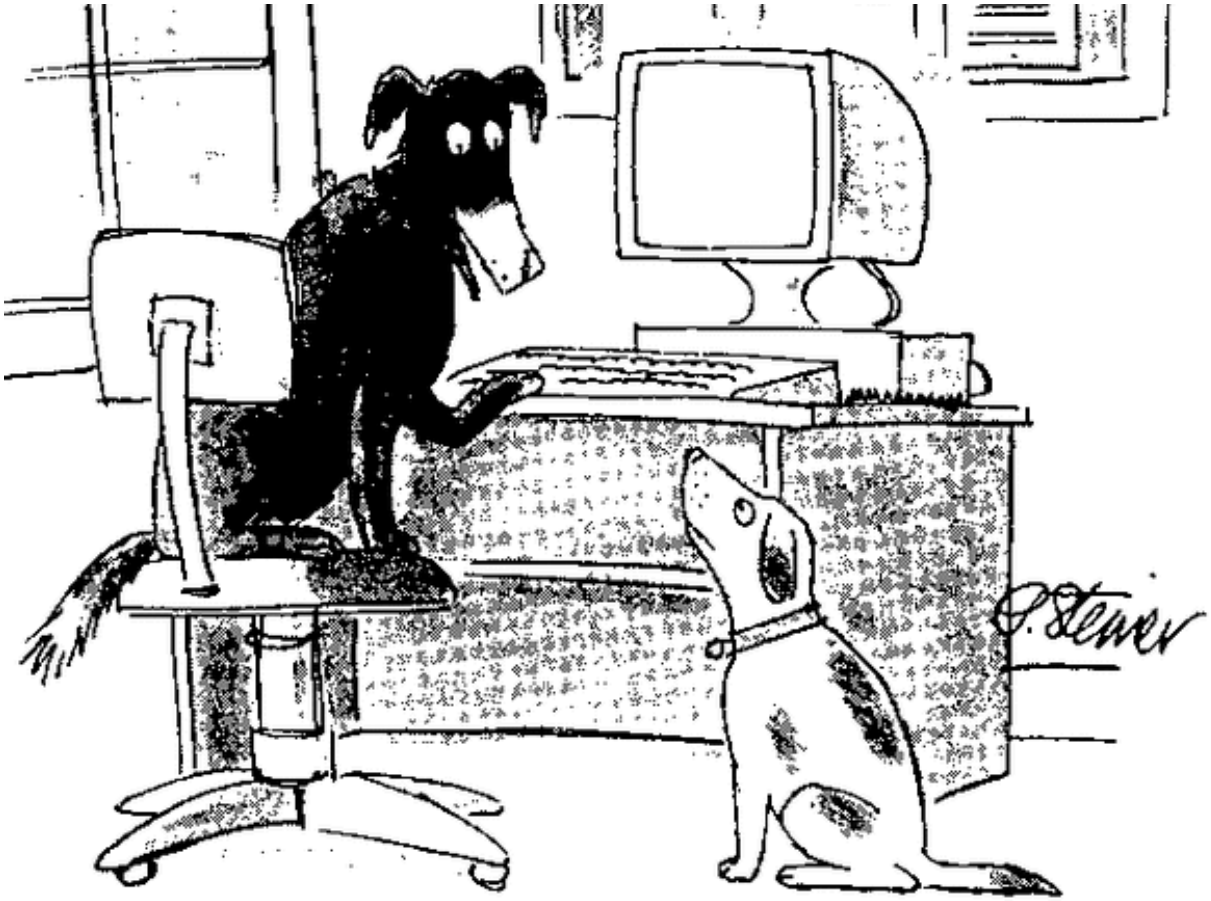
DP in action…



LinkedIn

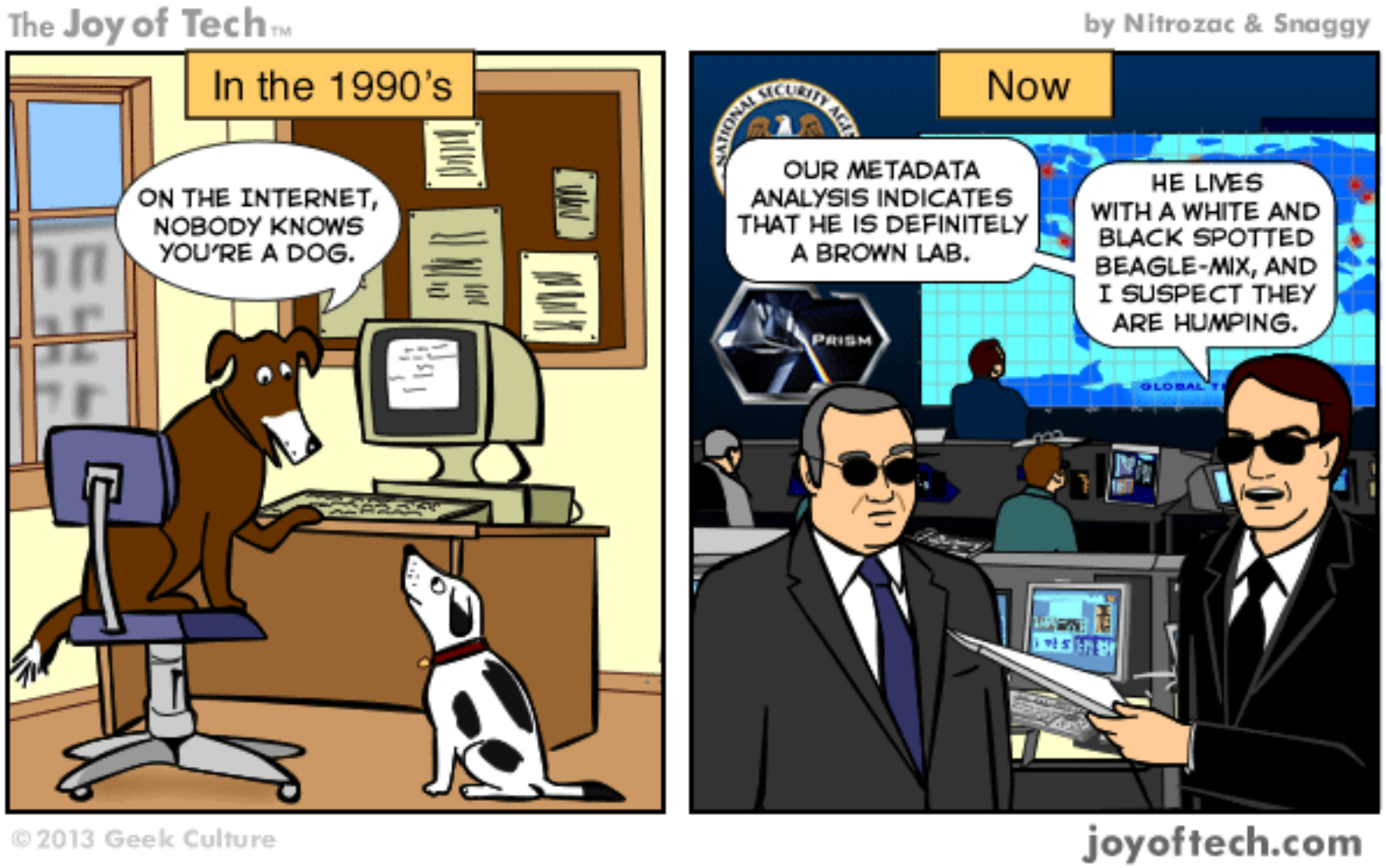
United States Census Bureau

Internet Privacy

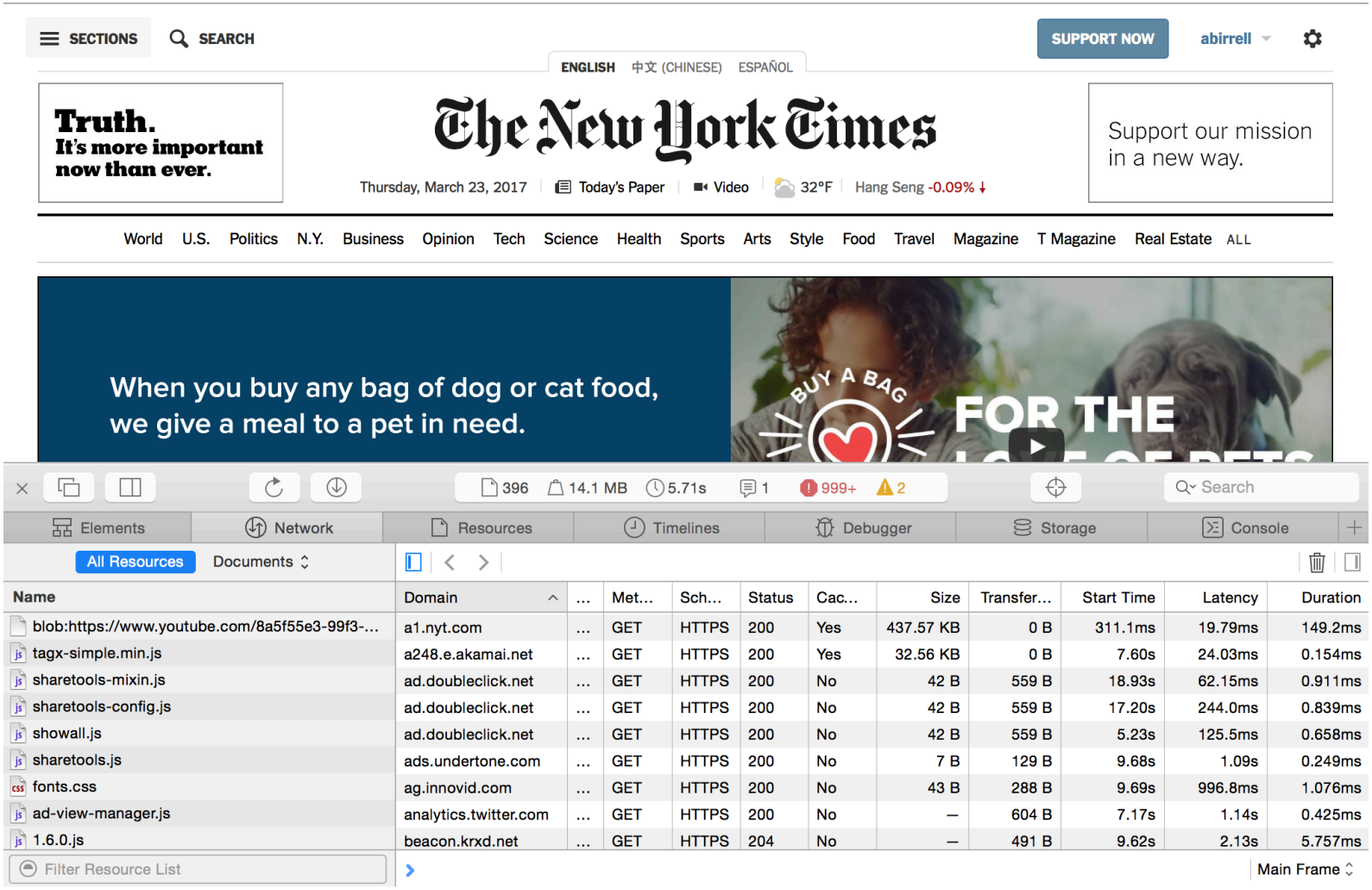


"On the Internet, nobody knows you're a dog."

Internet Privacy



Browser Tracking



The result…

Google Ad Settings

- 25-34 years old
- Female
- Sweetwater
- Action & Adventure Films
- Air Travel
- Android OS
- Antivirus & Malware
- Apparel
- Apple iOS
- Audio Equipment
- Audio File Formats & Codecs
- Bars, Clubs & Nightlife
- Books & Literature
- Business & Industrial
- Business & Productivity Software
- Business News
- Business Services
- C & C++

Exercise 4: Ad Settings

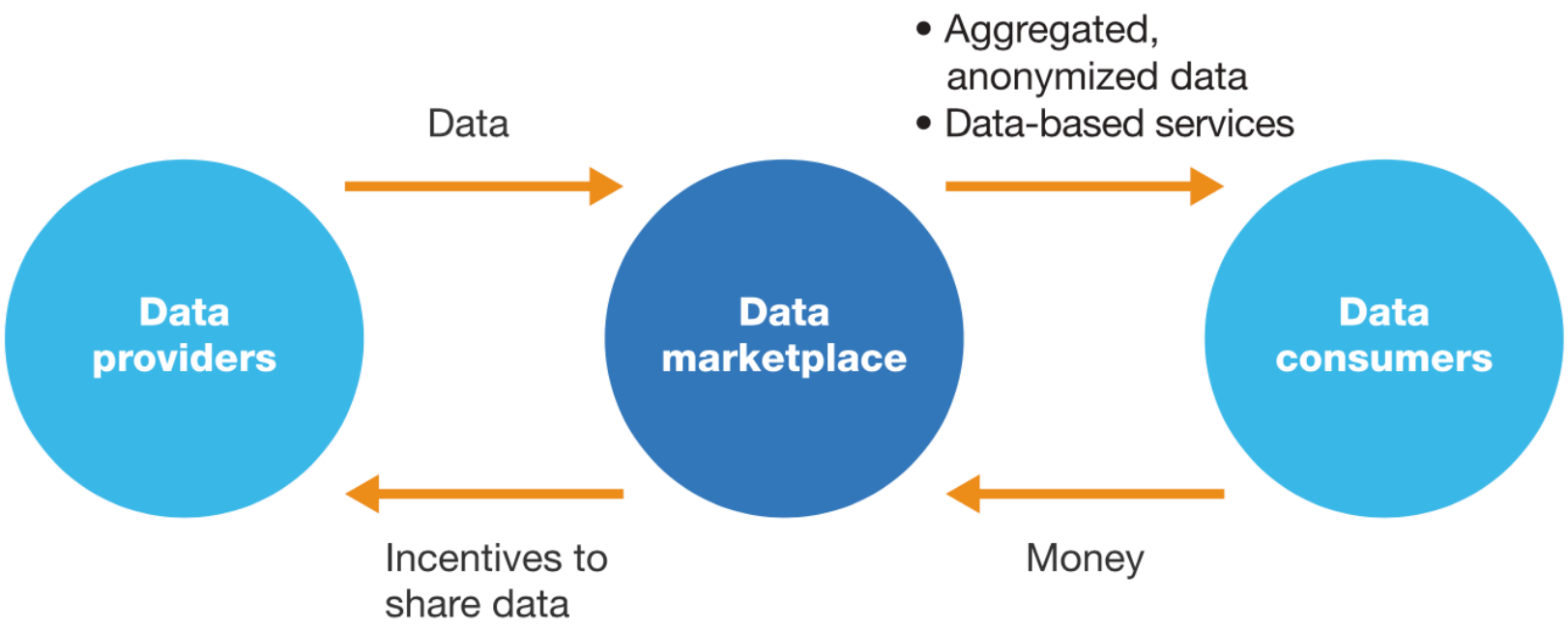
- If you have a google account, check your ad settings. If personalized ads are enabled, take a look at the categories associated with you: https://adssettings.google.com/
- How accure are these categories? Are there any that bother you?

Personalized Content

Targeted Advertising

© marketoonist.com

Data Marketplace



New Legal Regulations

- General Data Protection Regulation (GDPR), 2018
- California Consumer Privacy Act (CCPA), 2020
- Chile Privacy Bill Initiative, 2018
- New Zealand Privacy Bill, 2019
- Brazilian General Protection Law, 2020
- India Personal Data Protection Bill, 2020

Exercise 5: Feedback

1. Rate how well you think this recorded lecture worked
   1. Better than an in-person class
   2. About as well as an in-person class
   3. Less well than an in-person class, but you still learned something
   4. Total waste of time, you didn't learn anything

2. How much time did you spend on this video lecture (including time spent on exercises)?

3. Do you have any comments or feedback?

Data Privacy…

"Remember when, on the Internet, nobody knew who you were?"