# Introduction to privacy-preserving data analysis

DS 102, Fall 2019

Moritz Hardt

# Part I

Many valuable applications of data science touch on sensitive personal data

Advertising

Health data

Census and
government data

Location data and
mobile phone activity

Finance

Smart meter data

How can we perform useful data analysis while protecting individual privacy?

# Today’s lecture

Failure of ad-hoc anonymization techniques in practice

The fundamental law of information recovery

Privacy attacks: Effective ways to breach privacy

Randomized response: An early randomization scheme

Next time: Differential privacy

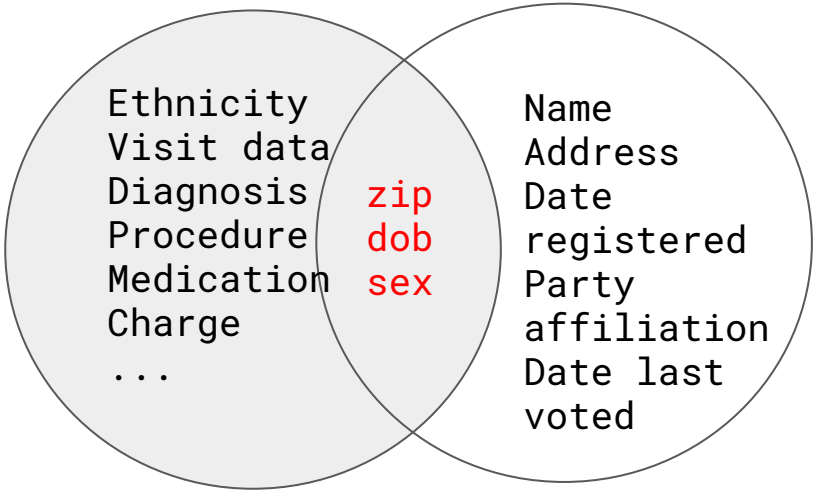
# Personally identifiable information (PII) and why it’s not enough to remove it

Common idea is to remove “sensitive attributes” from data to anonymize individuals

E.g. HIPAA safe harbor provision specifies such a rule for medical data

Name, location, phone, email, IP, SSN, medical record numbers, health plan numbers, device identifiers, account numbers, ...

# Sweeney's surprise for the Massachusetts governor (1997)




Medical data

Voter list

Latanya Sweeney

# This is called a *linkage attack*

Multiple data sources are combined to de-anonymize (or re-identify) records in a database

It’s one of endless attacks against ad-hoc anonymization schemes

# *k*-anonymity

Sweeney (1998)

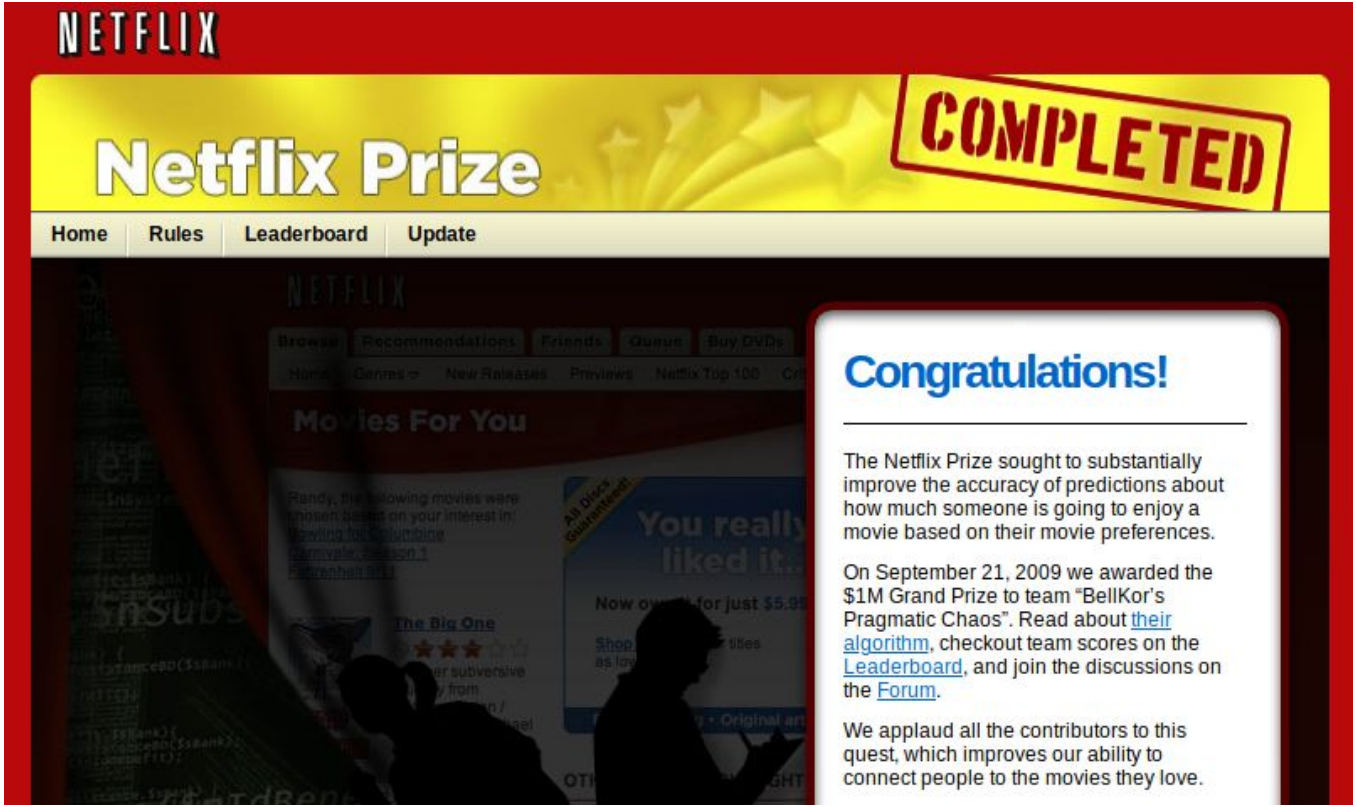
Divide data attributes into “quasi-identifiers” and “sensitive attributes”

Modify DB so that there are ≥ *k* rows for each combination of quasi-identifiers that is present

Many variants later on.

All broken.

# The Netflix Prize (2006-2009)




This is what web pages looked like in 2006?

# The Netflix data

18k movies

480k users

100M ratings {?, 0,1,...,5}

User names
Users replaced with random numbers

Official challenge goal: Predict missing entries

Kids, know what this is?

No, it's not a floppy disk.

It's a CD-ROM and the Netflix data fit on one of these. (650MB)

# The Netflix data

18k movies

480k users

100M ratings {?, 0,1,...,5}

User names
Users replaced with random numbers

Official challenge goal: Predict missing entries

As an side:

The Netflix challenge lead to lots of interesting technical work on *collaborative filtering* and *matrix completion*.

Idea: Fit a low rank approximation to the observed entries. Interpolate missing entries using low rank factors.

See e.g., Candes, Recht (2008); Recht (2009)

BRUCE SCHNEIER SECURITY 12.12.07 09:00 PM

# Why 'Anonymous' Data Sometimes Isn't

Another linkage attack!

**LAST YEAR, NETFLIX** published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using. The data was anonymized by removing personal details and replacing names with random numbers, to protect the privacy of the recommenders.

Arvind Narayanan and Vitaly Shmatikov, researchers at the University of Texas at Austin, de-anonymized some ofthe Netflix data by comparing rankings and timestamps with public information in the Internet Movie Database, or IMDb.

RYAN SINGEL SECURITY 03.12.10 02:48 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.

# 33 bits of entropy

33 bits are enough to index 8.5bln people

Rule of thumb: Given information source about individuals with > 33 bits of entropy, de-anonymization is possible and often easy

Example: Browsing history (even just, say, last 100 pages) is a unique identifier

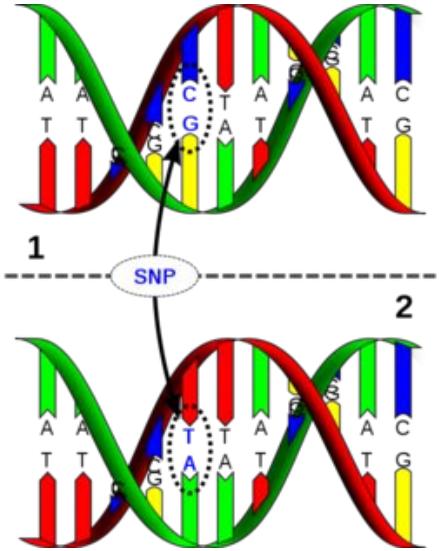
See 33bits.org (Blog by Narayanan on this topic)

# Genome Wide Association Studies (GWAS)

Typical Setup:

1. NIH takes DNA of 1000 test candidates with common disease
2. NIH releases minor allele frequencies (MAF) of test population at 100,000 positions (SNPs)

**Goal:** Find association between SNPs and disease



# Attack on GWAS data [Homer et al.]

Can **infer membership in test group** of an individual
with known DNA from published data!

| SNP | 1 | 2 | 3 | … | … | 100000 |
|---|---|---|---|---|---|---|
| MAF | 0.02 | 0.03 | 0.05 | … | | 0.02 |

Test population

| SNP | 1 | 2 | 3 | … | … | 100000 |
|---|---|---|---|---|---|---|
| MA | NO | NO | YES | … | | YES |

Moritz's DNA

| SNP | 1 | 2 | 3 | … | … | 100000 |
|---|---|---|---|---|---|---|
| MAF | 0.01 | 0.04 | 0.04 | … | | 0.01 |

Reference population (HapMap data, public)

# Attack on GWAS data [Homer et al.]

Can **infer membership in test group** of an individual
with known DNA from published data!

| SNP | 1 | 2 | 3 | … | … | 100000 |
|---|---|---|---|---|---|---|
| MAF | 0.02 | 0.03 | 0.05 | … | | 0.02 |

Test population

probably

| SNP | 1 | 2 | 3 | … | … | 100000 |
|---|---|---|---|---|---|---|
| MA | NO | NO | YES | … | | YES |

Moritz's DNA

| SNP | 1 | 2 | 3 | … | … | 100000 |
|---|---|---|---|---|---|---|
| MAF | 0.01 | 0.04 | 0.04 | … | | 0.01 |

Reference population (HapMap data, public)

# Interesting but typical characteristics

- Only innocuous looking data was released
  - Data was HIPAA compliant
- Data curator is trusted (NIH)
- Attack uses background knowledge (HapMap data set) available in public domain
- Attack uses unanticipated algorithm
- Curator pulled data sets (now hard to get)
- **Technical principle: Many weak signals combine into one strong signal**

# The fundamental law of information recovery

Dwork (ca 2014), Dwork and Roth (2014)

"Overly accurate information about too many queries to a data source allows for partial or full reconstruction of data (i.e., *blatant non-privacy*)."

Many formal incarnations: ***Reconstruction attacks***

# Boosting weak signals

# The signal boost lemma


Let $b$ in {-1, 1} be an unknown bit. (Think sensitive bit about one individual.)

**Query:** We can sample the distribution **B** = **Bernoulli**(½ + εb).

How many draws from **B** do we need to know $b$ with high confidence?

Answer: $\Theta(1/\varepsilon^2)$ samples are necessary and sufficient.

**Many reconstruction attacks reduce to some variant of signal boost lemma.**

Strategy: Identify source of mild correlation, boost into large correlation.

# The signal boost lemma

**Proof idea (sufficient):** Sample bits $b_1, \ldots, b_n$. If sum $S = (1/n)\sum_i b_i > 0$, *guess bit* $b'=1$, *else guess* $b'=-1$.

*Note:* $\mathbf{E}[S] = 2\varepsilon b$, $V[S] = (1-4\varepsilon^2)/n \cong 1/n$

*Guess is good with probability, say, 9/10, if* $\varepsilon > C / n^{1/2}$.

**Proof idea (necessary):** Let **B** = **Bernoulli**(½ + ε), **B'** = **Bernoulli**(½ - ε). Let $\mathbf{B}^n$ denote *n* independent draws from **B**.

Show $TV(\boldsymbol{B}^n, \boldsymbol{B}'^n) = o(1)$ for $n = o(1/\varepsilon^2)$.

Bound Hellinger distance between **B**, **B'**, use product rule for Hellinger squared distance, relate Hellinger and TV. [Details in the notes.]

# Approximate inversion

# Linear reconstruction attacks

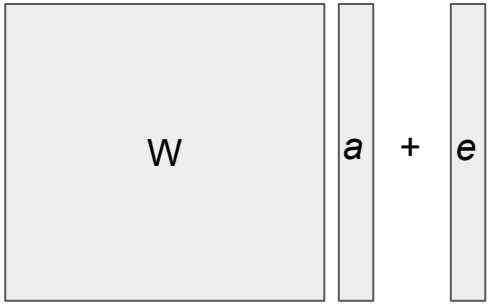
Binary vector *a*, corresponding to sensitive -1/1 bits of *n* individuals


**Query:** We can specify vector $w$ *in* $\{-1, 1\}^n$

***Query answer:*** Inner product $\langle a, w \rangle + e$, where $e$ is an **unknown** noise term.

Assuming some bound on the error term, how many queries do we need to approximately reconstruct $a$?

# Linear reconstruction attacks

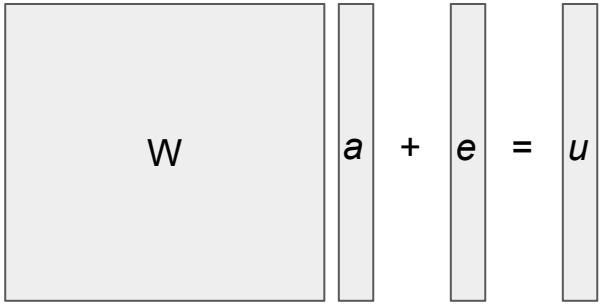



**Observation:** We can write multiple measurements as matrix W with -1/1 coefficients

Let's choose W to be n x n.

**Query answer:** *Wa* + *e*, where e is now a vector

# Main idea

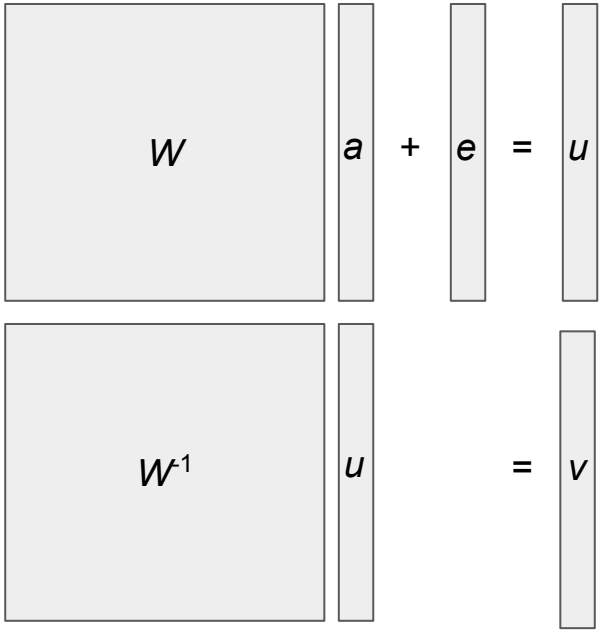



Suppose we have u = Wa + e, how do we get back a?

Assuming W is invertible, we can compute v = $W^{-1}u$

But when is this good?

# Main idea




$u = Wa + e$

$v = W^{-1}u$

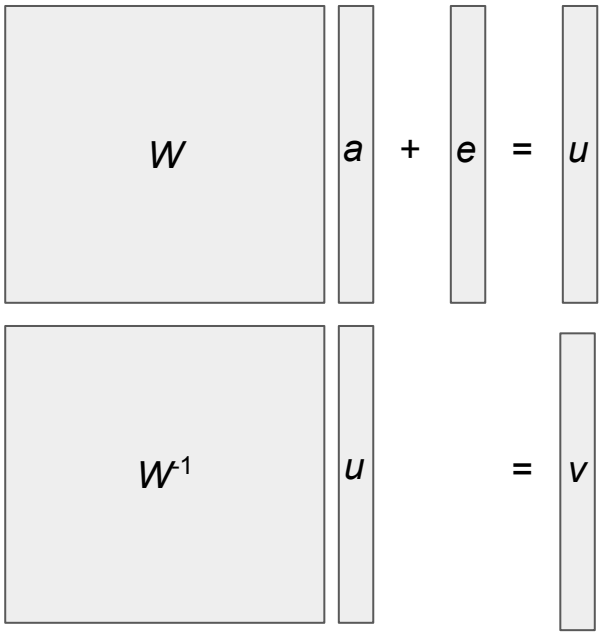
So:

$v = W^{-1}Wa + W^{-1}e = a + W^{-1}e$

Hence, we reconstruct a up to error term $W^{-1}e$

How can we make sure that $W^{-1}e$ has *small norm?*

# Main idea



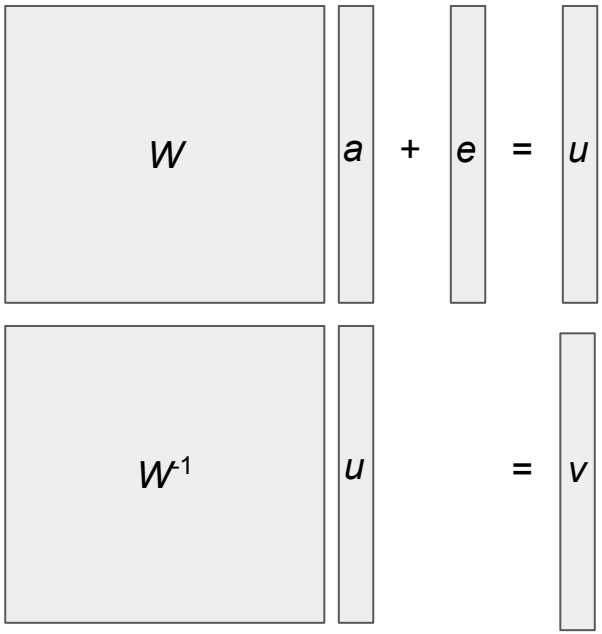

How can we make sure that $W^{-1}e$ has *small norm?*

*Note:* $||W^{-1}e|| \leq ||W^{-1}||\ \ ||e||$

Here $||W^{-1}||$ is the operator norm of $W^{-1}$.

It equals $1/\sigma_n(W)$, where $\sigma_n(W)$ is the smallest singular value of $W$

# Main idea



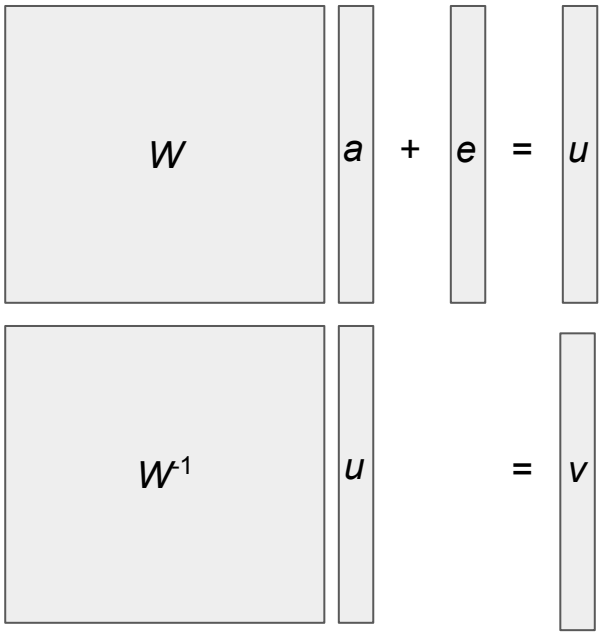

How do we maximize $\sigma_n(W)$ the smallest singular value of a binary -1/1 matrix $W$

*One good choice: Random*

*Another good choice: Hadamard*

*Both have* $\sigma_n(W) \gtrsim n^{1/2}$

# Wrapping things up.



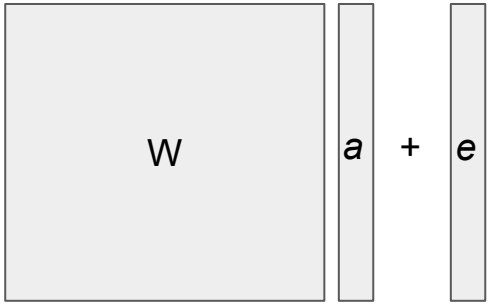

$v = a + W^{-1}e$

$||W^{-1}e|| \leq ||W^{-1}||\ \ ||e|| = ||e||/\sigma_n(W) \lesssim n^{-1/2}||e||$

$||a||^2 = n$, because $a$ is binary

Assume $||e||^2 = o(n^2)$. Then,
$||v - a||^2 = ||W^{-1}e||^2 \lesssim o(n)$.

# Linear reconstruction attacks




**Corollary:** Assume that each coordinate of the perturbation *e* has magnitude $o(n^{1/2})$.

Then, the linear reconstruction attack reconstructs *a* up to an average coordinate error of $o(1)$.

# A hint at how to ensure privacy

“Do you do drugs?”

Sensitive questions likely lead to *evasive answer bias*

# RANDOMIZED RESPONSE: A SURVEY TECHNIQUE FOR ELIMINATING EVASIVE ANSWER BIAS

STANLEY L. WARNER
*Claremont Graduate School*

Published
In 1965

For various reasons individuals in a sample survey may prefer not to confide to the interviewer the correct answers to certain questions. In such cases the individuals may elect not to reply at all or to reply with incorrect answers. The resulting evasive answer bias is ordinarily difficult to assess. In this paper it is argued that such bias is potentially removable through allowing the interviewee to maintain privacy through the device of randomizing his response. A randomized response method for estimating a population proportion is presented as an example. Unbiased maximum likelihood estimates are obtained and their mean square errors are compared with the mean square errors of conventional estimates under various assumptions about the underlying population.

# Basic idea

Suppose b in {-1, 1} is your private bit (answer to sensitive question).

Sample b’ from **Bernoulli**(½ + εb).

Report b’ instead of b.

Plausible deniability: Given that your reported value is, say, 1. You can plausibly claim that it was actually -1.

# Analysis idea

Suppose $n$ individuals report noisy bits $b_i' \sim$ **Bernoulli**$(½ + \varepsilon b_i)$.

We're interested in the average sensitive value mean$(b_1,\ldots,b_n)$.

**But note: E**[mean$(b_1',\ldots, b_n')$] = $½ + \varepsilon$ mean$(b_1,\ldots,b_n)$
and **V**[mean$(b_1',\ldots, b_n')$] = O(1/n).

So, for large enough n, we can reconstruct mean$(b_1,\ldots,b_n)$ from the noisy values.

# Historical note

Warner envisioned this approach for telephone surveys.

How would a respondent on the phone be able to create randomness?

# Some notes

The signal boost lemma shows that we can’t invoke randomized response too many times or else we compromise the private bit.

It’s not clear how to generalize the randomization scheme to multiple analysis in such a way that the privacy guarantee *composes* well.

We’ll see how to do this next time when we talk about *differential privacy*.

Apple and Google now use variants of randomized response at scale.