Protecting Data Sources, Protecting Personal Data

Simson L. Garfinkel
Senior Scientist, Confidentiality and Data Access
U.S. Census Bureau

4th Kavli Symposium
Tuesday, February 20
Austin, TX

The views in this presentation are those of the author, and do not necessarily represent those of the U.S. Census Bureau.
Raise your hand if you use two-factor authentication to protect your email account.
Two factor authentication
Protecting Data Sources, Protecting Personal Data

Sources → Collection → Processing → Dissemination

- Communications Security
- Storage Security
- Publications Security
Outline

Communications security: How do you obtain confidential information from your sources?

Storage security: How do you maintain your secrets?

Publication security: How do you control the information released by your publication to prevent the inadvertent release of confidential information?
Bio: Simson L. Garfinkel

1987  Freelance science writer

2006  Associate Professor

2015  Computer Scientist

2017  Computer Scientist
I have spent 29 years trying to secure computers...
Today’s systems are frequently less secure than those of the 1970s.

Poor security is inherent in many information systems.

- Attack is easier and cheaper than defense.
- Cyber “defense in depth” does not work if a single vulnerability compromises.
- It’s easier to break things than to fix them.

Network Connectivity makes it easier to exploit vulnerable computers.

Fortunately, most journalists have modest security needs.
A methodology for thinking about your security needs

Identify your **critical assets** and **interactions** — what you are trying to protect.

Identify potential **threats** — what you are trying to protect against.

Identify potential **vulnerabilities** — how your threat could be harmed

Identify **risks** — the potential for harm

\[
\text{Asset} + \text{Threat} + \text{Vulnerability} = \text{Risk}
\]
There are many risk equations

Asset + Threat + Vulnerability = Risk

Risk = Threat * Vulnerability * Consequences

Risk = Impact * Probability

These equations *shouldn’t* be solved quantitatively.
Communications Security
Communications with sources: Securing data in flight

Primary risk: interception

Email  Phone  In-Person meeting

Asset: content & reputation
Which of these is has the most interception risk?

Answer depends on:

• **Threat** — who is attacking?
• **Vulnerability** — how are they attacking?
• **Consequence** — what is the impact of an interception?
In-person meetings are risky
On Amtrak, powerful people talk loudly and spill secrets.

“This is my conclusion based on five years’ field research commuting on Amtrak’s Acela between cities along the East Coast.”
Eavesdropping email or phone requires access.

There are two points of access:
1. The end-point devices.
2. The network.

Primary threat: spyware and malware

Primary threat: interception
Encryption doesn’t protect against malware

"https:" encryption protects data in flight against interception.

S/MIME and PGP (message encryption) also protects data at rest. See NIST SP800-188, “Trustworthy Email.”
Storage Security
Storage Security

Local Storage

Cloud Storage

Issues: Physical Access • Logical Access
Most of the data crimes in recent years have been unauthorized access to stored data.

Physical access:
- Attacker physically removed the data.

Logical access:
- Computer system allowed access
- Data were not encrypted to the attacker.
May 2013: Edward Snowden steals millions of documents from the US National Security Agency
March 2014:
IRS Employee Took Home Data on 20,000 Workers

A U.S. Internal Revenue Service employee took home a computer thumb drive containing unencrypted data on 20,000 fellow workers, the agency said in a statement today.

The tax agency’s systems that hold personal data on hundreds of millions of Americans weren’t breached, the statement said.

Photographer: Andrew Harrer/Bloomberg

March 2014:
Stolen F-35 secrets show up in China’s stealth Fighter
Apple Admits Celebrity Photos Were Stolen In Targeted Hack
Protecting Local Storage

Physical security.

Disk encryption.

Off-site backups.

MacOS FileVault

Oakland CA fires, 1989
Protecting Network Storage

Two-factor access

Account recovery

Google Drive  GMAIL  DropBox  OneDrive
Example: Google Authenticator’s 2-factor authentication protections against password stealing.
Universal Second Factor (U2F)
Here's what you need to do.

1. Visit Google's 2-Step Verification page and click on Get Started.
2. Enter your Google email and password (or just your password, if you're already logged in to Google).
3. Click the Start setup button.
4. Add a phone number that Google can send the six-digit verification code to.

More items...

How to set up 2-step verification for Google and Gmail on your iPhone ...
Publication Security
Problems with attempts at anonymisation

- Black band over the eyes
- Pixelated face
- Cropped head

https://commons.wikimedia.org/wiki/Commons:Photographs_of_identifiable_people
SECRET OF HISTORY: The C.I.A. in Iran -- A special report.; How a Plot Convulsed Iran in '53 (and in '79)

By JAMES RISEN  APRIL 16, 2000

For nearly five decades, America's role in the military coup that ousted Iran's elected prime minister and returned the shah to power has been lost to history, the subject of fierce debate in Iran and stony silence in the United States. One by one, participants have retired or died without revealing key details, and the Central Intelligence Agency said a number of records of the operation -- its first successful overthrow of a foreign government -- had been destroyed.

The Times had attempted to redact the names Iranians who had assisted.

The Times “redacted” by putting black boxes over the PDF.

Cryptome.org removed the black boxes and re-published.

[http://cryptome.org/cia-iran.htm](http://cryptome.org/cia-iran.htm)
The digital means the NY Times used to black out names of persons it was advised might be put at risk by publication failed to do the job properly. All the deletions are readable. The unredacted report shall be published shortly on cryptome.org.

The unexpected consequences of digital security are worth pondering.
Dear Mr. Young, Thank you for informing us about the problem with this document. We are removing it from our site until we can delete the names in a more secure fashion. The names were obscured because of our concern for possible retribution against the families of the people named in this report, and we would strongly urge you to respect that judgment.

Sincerely, Rich Meislin
In online blunder, Dallas police revealed names of people reporting sexual assaults

Sarah Mervosh, Investigative Reporter
Data can be revealing, even without names.

In March 2014, the New York City Taxi & License Commission tweeted a “TAXI FACTS” infographic:

Chris Whong files a “Freedom of Information Law” request for all the data used to create the graphic.
NYC TLC provided Chris Whong with all of the data

175 million trips:

<table>
<thead>
<tr>
<th>medallion</th>
<th>hack_license</th>
<th>vendor_id</th>
<th>pickup_datetime</th>
<th>payment_type</th>
<th>fare_amount</th>
<th>surcharge</th>
<th>mta_tax</th>
<th>tip_amount</th>
<th>tolls_amount</th>
<th>total_amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>89D227B6555E5C2AECF13C3</td>
<td>BA96DE41E97116918944</td>
<td>1/1/13 15:11</td>
<td>CSH</td>
<td>6.5</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>0BD78F5B1A12B88E067BED</td>
<td>9DF869F08040DB5549F</td>
<td>1/6/13 0:18</td>
<td>CSH</td>
<td>5.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>0BD78F5B1A12B88E067BED</td>
<td>9DF869F08040DB5549F</td>
<td>1/5/13 18:49</td>
<td>CSH</td>
<td>5.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>DFD220EE0B7F8A8DC9A5780</td>
<td>51E68763205C985E843</td>
<td>1/7/13 23:54</td>
<td>CSH</td>
<td>5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>DFD220EE0B7F8A8DC9A5780</td>
<td>51E68763205C985E843</td>
<td>1/7/13 23:25</td>
<td>CSH</td>
<td>9.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10.5</td>
</tr>
<tr>
<td>7</td>
<td>20D9ECB8A0767CF7A01564I598BCE589C19B1658DDE</td>
<td>CMT</td>
<td>1/7/13 15:27</td>
<td>CSH</td>
<td>9.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>496644932DF9392605C22CE751319AD756FF14FE670</td>
<td>CMT</td>
<td>1/8/13 11:01</td>
<td>CSH</td>
<td>6</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.5</td>
</tr>
<tr>
<td>9</td>
<td>0B578B6363A2FEC3D83B1944</td>
<td>CDO43678417ED66349D</td>
<td>1/7/13 12:39</td>
<td>CSH</td>
<td>34</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4.8</td>
<td>39.3</td>
</tr>
<tr>
<td>10</td>
<td>2C0E91FF20A856C891383ED1DA2F65346A2B8ED934</td>
<td>CMT</td>
<td>1/7/13 18:15</td>
<td>CSH</td>
<td>5.5</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>

Every trip:
- Pickup date, time & GPS
- Drop-off date, time & GPS
- Fare & tip
- Encoded medallion number

With this data, you can make a map of NYC Taxi Service
Compare taxi prices and Uber prices:

http://qz.com/363759/data-proves-that-often-a-yellow-taxi-is-a-better-deal-than-an-uber/
Each taxi has a pseudonym, which allows taxi rides to be linked.
Oops. The taxi medallion numbers were not properly de-identified.

<table>
<thead>
<tr>
<th>Pseudonym</th>
<th>Taxi Medallion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0f76c35d4a069e0fe76b21d28f009639</td>
<td>5C27</td>
</tr>
<tr>
<td>be9f314926dd314b36496d926e42f4db</td>
<td>5C28</td>
</tr>
<tr>
<td>9ee993809f648d39d24f5ba8f862d7f1</td>
<td>5C29</td>
</tr>
<tr>
<td>23f7e8636fb9099822aa381054d215d4</td>
<td>5C30</td>
</tr>
</tbody>
</table>

The pseudonyms looked suspicious to Anthony Tockar, an intern at Neustar Research.

Tockar realized that the pseudonyms were MD5 hashes

MD5(“5C28”) = be9f314926dd314b36496d926e42f4db
Anthony Tockar identified the medallion number the records. He searched for photos in flickr that showed movie stars at taxis where he could read the medallion number.

MD5 can’t be reversed, but it’s possible to do a “brute force search” on all possible values.

A journalist at Gawker identified 9 other cab rides.
May 18, 1996: Massachusetts Governor William Weld Collapses at Bentley College Commencement

Massachusetts Governor Doing Well After Collapse

WALTHAM, Massachusetts (CNN, May 18) -- Gov. William Weld collapsed during a graduation ceremony at Bentley College, but doctors said he was doing well.

The governor was taken to Deaconess-Waltham Hospital, where he was undergoing a battery of tests, according to Bentley College spokeswoman Katherine Blake. Weld will remain in the hospital overnight for observation, she said.

Doctors said they performed an electrocardiogram, chest X-ray and blood tests, but found no immediate cause for concern.

"With all this testing we have done, nothing acute is showing," said Dr. Rifat Dweik.

"Right now, it looks like maybe the flu," said Pam Jonah, one of Weld's press aides.

Weld was receiving an honorary doctorate of law at 11 a.m. EDT when he was stricken, according to Blake.
In 1997, MIT Graduate Student Latanya Sweeney decided to search for William Weld’s medical records in the GIC data.

Sweeney obtains GIC dataset and looks for Weld's data.

- She knew that Weld lived in Cambridge, MA.
- Sweeney purchased Cambridge voter rolls for $20.
- Six people had the same birthday (July 31, 1945)
- Three were men
- One person had the same ZIP code.
"Linkage Attack"
Matching records using quasi-identifiers

- Weld’s records were uniquely identified.
- Sweeney estimated 87% of US population were uniquely identified by birthday, sex & ZIP

"Sensitive Data"
- Hospital admission info

"Quasi-Identifiers" or "Indirect-Identifiers"
- Birthday
- Sex
- ZIP Code

"Direct" or "Explicit" identifiers
- Name
- Address
- Phone
- SSN

United States Census Bureau
Sweeney invented K-Anonymity
A model for de-identifying structured data.

A dataset that you would like to release:

<table>
<thead>
<tr>
<th>Name</th>
<th>Race</th>
<th>Birthdate</th>
<th>Sex</th>
<th>Zip</th>
<th>Medication</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>Black</td>
<td>9/20/65</td>
<td>M</td>
<td>37203</td>
<td>M1</td>
<td>Gastric Ulcer</td>
</tr>
<tr>
<td>Bob</td>
<td>Black</td>
<td>2/14/65</td>
<td>M</td>
<td>37203</td>
<td>M1</td>
<td>Gastric Ulcer</td>
</tr>
<tr>
<td>Candice</td>
<td>Black</td>
<td>10/23/65</td>
<td>F</td>
<td>37215</td>
<td>M1</td>
<td>Gastritis</td>
</tr>
<tr>
<td>Dan</td>
<td>Black</td>
<td>8/24/65</td>
<td>F</td>
<td>37215</td>
<td>M2</td>
<td>Gastritis</td>
</tr>
<tr>
<td>Eliza</td>
<td>Black</td>
<td>11/7/64</td>
<td>F</td>
<td>37215</td>
<td>M2</td>
<td>Gastritis</td>
</tr>
<tr>
<td>Felix</td>
<td>Black</td>
<td>12/1/64</td>
<td>F</td>
<td>37215</td>
<td>M2</td>
<td>Stomach Cancer</td>
</tr>
<tr>
<td>Gazelle</td>
<td>White</td>
<td>10/23/64</td>
<td>M</td>
<td>37215</td>
<td>M3</td>
<td>Flu</td>
</tr>
<tr>
<td>Harry</td>
<td>White</td>
<td>3/15/64</td>
<td>F</td>
<td>37217</td>
<td>M3</td>
<td>Flu</td>
</tr>
<tr>
<td>Irene</td>
<td>White</td>
<td>8/13/64</td>
<td>M</td>
<td>37217</td>
<td>M3</td>
<td>Flu</td>
</tr>
<tr>
<td>Jack</td>
<td>White</td>
<td>5/5/64</td>
<td>M</td>
<td>37217</td>
<td>M4</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>Kelly</td>
<td>White</td>
<td>2/13/67</td>
<td>M</td>
<td>37215</td>
<td>M4</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>Lenny</td>
<td>White</td>
<td>3/21/67</td>
<td>M</td>
<td>37215</td>
<td>M4</td>
<td>Flu</td>
</tr>
</tbody>
</table>
First the identifiers are removed

<table>
<thead>
<tr>
<th>Name</th>
<th>Race</th>
<th>Birthdate</th>
<th>Sex</th>
<th>Zip</th>
<th>Quasi Identifiers</th>
<th>Medication</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>Black</td>
<td>9/20/65</td>
<td>M</td>
<td>37203</td>
<td></td>
<td>M1</td>
<td>Gastric Ulcer</td>
</tr>
<tr>
<td>Bob</td>
<td>Black</td>
<td>2/14/65</td>
<td>M</td>
<td>37203</td>
<td></td>
<td>M1</td>
<td>Gastric Ulcer</td>
</tr>
<tr>
<td>Candice</td>
<td>Black</td>
<td>10/23/65</td>
<td>F</td>
<td>37215</td>
<td></td>
<td>M1</td>
<td>Gastritis</td>
</tr>
<tr>
<td>Dan</td>
<td>Black</td>
<td>8/24/65</td>
<td>F</td>
<td>37215</td>
<td></td>
<td>M2</td>
<td>Gastritis</td>
</tr>
<tr>
<td>Eliza</td>
<td>Black</td>
<td>11/7/64</td>
<td>F</td>
<td>37215</td>
<td></td>
<td>M2</td>
<td>Gastritis</td>
</tr>
<tr>
<td>Feliz</td>
<td>Black</td>
<td>12/1/64</td>
<td>F</td>
<td>37215</td>
<td></td>
<td>M2</td>
<td>Stomach Cancer</td>
</tr>
<tr>
<td>Hazel</td>
<td>White</td>
<td>10/23/64</td>
<td>M</td>
<td>37215</td>
<td></td>
<td>M3</td>
<td>Flu</td>
</tr>
<tr>
<td>Harry</td>
<td>White</td>
<td>3/15/64</td>
<td>F</td>
<td>37217</td>
<td></td>
<td>M3</td>
<td>Flu</td>
</tr>
<tr>
<td>Irene</td>
<td>White</td>
<td>8/13/64</td>
<td>M</td>
<td>37217</td>
<td></td>
<td>M3</td>
<td>Flu</td>
</tr>
<tr>
<td>Jack</td>
<td>White</td>
<td>5/5/64</td>
<td>M</td>
<td>37217</td>
<td></td>
<td>M4</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>Kelly</td>
<td>White</td>
<td>2/13/67</td>
<td>M</td>
<td>37215</td>
<td></td>
<td>M4</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>Lenny</td>
<td>White</td>
<td>3/21/67</td>
<td>M</td>
<td>37215</td>
<td></td>
<td>M4</td>
<td>Flu</td>
</tr>
</tbody>
</table>
A dataset is “k-anonymous” if every record is in a set of at least k indistinguishable individuals

<table>
<thead>
<tr>
<th>Race</th>
<th>Birthdate</th>
<th>Sex</th>
<th>Zip</th>
<th>Medication</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>65</td>
<td>M</td>
<td>37203</td>
<td>M1</td>
<td>Gastric Ulcer</td>
</tr>
<tr>
<td>Black</td>
<td>65</td>
<td>M</td>
<td>37203</td>
<td>M1</td>
<td>Gastric Ulcer</td>
</tr>
<tr>
<td>Black</td>
<td>65</td>
<td>F</td>
<td>37215</td>
<td>M1</td>
<td>Gastritis</td>
</tr>
<tr>
<td>Black</td>
<td>65</td>
<td>F</td>
<td>37215</td>
<td>M2</td>
<td>Gastritis</td>
</tr>
<tr>
<td>Black</td>
<td>64</td>
<td>F</td>
<td>37215</td>
<td>M2</td>
<td>Gastritis</td>
</tr>
<tr>
<td>Black</td>
<td>64</td>
<td>F</td>
<td>37215</td>
<td>M2</td>
<td>Stomach Cancer</td>
</tr>
<tr>
<td>White</td>
<td>64</td>
<td>M</td>
<td>37217</td>
<td>M3</td>
<td>Flu</td>
</tr>
<tr>
<td>White</td>
<td>64</td>
<td>-</td>
<td>37217</td>
<td>M3</td>
<td>Flu</td>
</tr>
<tr>
<td>White</td>
<td>64</td>
<td>M</td>
<td>37217</td>
<td>M3</td>
<td>Flu</td>
</tr>
<tr>
<td>White</td>
<td>64</td>
<td>-</td>
<td>37217</td>
<td>M4</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>White</td>
<td>67</td>
<td>M</td>
<td>37215</td>
<td>M4</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>White</td>
<td>67</td>
<td>M</td>
<td>37215</td>
<td>M4 49</td>
<td>Flu</td>
</tr>
</tbody>
</table>
Attribute disclosure:
We know the Black / 65 / M had a Gastric Ulcer.

<table>
<thead>
<tr>
<th>Race</th>
<th>Age</th>
<th>Sex</th>
<th>Zip</th>
<th>Medication</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>65</td>
<td>M</td>
<td>37203</td>
<td>M1</td>
<td>Gastric Ulcer</td>
</tr>
<tr>
<td>Black</td>
<td>65</td>
<td>M</td>
<td>37203</td>
<td>M1</td>
<td>Gastric Ulcer</td>
</tr>
<tr>
<td>Black</td>
<td>65</td>
<td>F</td>
<td>37215</td>
<td>M1</td>
<td>Gastritis</td>
</tr>
<tr>
<td>Black</td>
<td>65</td>
<td>F</td>
<td>37215</td>
<td>M2</td>
<td>Gastritis</td>
</tr>
<tr>
<td>Black</td>
<td>64</td>
<td>F</td>
<td>37215</td>
<td>M2</td>
<td>Gastritis</td>
</tr>
<tr>
<td>Black</td>
<td>64</td>
<td>F</td>
<td>37215</td>
<td>M2</td>
<td>Stomach Cancer</td>
</tr>
<tr>
<td>White</td>
<td>64</td>
<td>M</td>
<td>3721</td>
<td>M3</td>
<td>Flu</td>
</tr>
<tr>
<td>White</td>
<td>64</td>
<td>-</td>
<td>37217</td>
<td>M3</td>
<td>Flu</td>
</tr>
<tr>
<td>White</td>
<td>64</td>
<td>M</td>
<td>3721</td>
<td>M3</td>
<td>Flu</td>
</tr>
<tr>
<td>White</td>
<td>64</td>
<td>-</td>
<td>37217</td>
<td>M4</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>White</td>
<td>67</td>
<td>M</td>
<td>37215</td>
<td>M4</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>White</td>
<td>67</td>
<td>M</td>
<td>37215</td>
<td>M4</td>
<td>Flu</td>
</tr>
</tbody>
</table>
De-identification caveats — what can go wrong

Mistakes happen:

- Metadata may contain identifiers.
- Direct identifiers can be missed.
- Hard to determine what's a quasi-identifier.

Even worse:

- k-anonymity and l-diversity can significantly damage data quality.
- There is no mathematical proof that k-anonymity actually protects privacy.
All data are potentially identifying.
The Netflix Challenge (2008-2009)

Netflix published movie data for ~450,000 subscribers:

- Pseudonymized username
- Information on movies watched:
  
  _Movie Title_
  _Date watched_
  _Rating_

Challenge: Improve Netflix recommendation algorithm

Unintentional Challenge: Identify Netflix subscribers!
Re-identifying the Netflix Challenge Victims

“Sensitive Data”

Netflix Provided Data

Other Movies Watched & Movie Rankings

Movies Watched & Movie Rankings

IMDB username

“Direct” or “Explicit” identifiers
Figure 4. Adversary knows exact ratings and approximate dates.

Figure 8. Adversary knows exact ratings but does not know dates at all.

Figure 9. Effect of knowing less popular movies rated by victim. Adversary knows approximate ratings (±1) and dates (14-day error).
On Friday, Netflix announced on its corporate blog that it has settled a lawsuit related to its Netflix Prize, a $1 million contest that challenged machine learning experts to use Netflix’s data to produce better recommendations than the movie giant could serve up themselves.

The lawsuit called attention to academic research that suggests that Netflix indirectly exposed the movie preferences of its users by publishing anonymized customer data. In the suit, plaintiff Paul Navarro and others sought an injunction preventing Netflix from going through the so-called "Netflix Prize II," a follow-up challenge that Netflix promised would offer up even more personal data such as genders and zipcodes.

"Netflix is not going to pursue a sequel to the Netflix Prize," says spokesman Steve Swasey. "We looked at this, we heard some dissension and so we’ve settled it, resolved the issues and are moving on."
Differential Privacy: The Big Idea
Differential privacy is a new approach for assuring privacy in the release of statistical data.

Based on hope and assumptions.
1. Data are identify, quasi-identifying, or not-identifying
2. Future data sets will not be released that can be linked with previously released data
3. Adversaries have limited resources to pursue re-identification attacks

Based on math.
In traditional data publications, there are many ways that the contributions of an individual can leak out. It’s pretty easy to determine that the new kid is sad and has a 90.
Differential privacy’s core idea:
Create uncertainty regarding the presence any person in the dataset.

Noise is added to mask an individual’s contribution

### January

<table>
<thead>
<tr>
<th>Name</th>
<th>Affect</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex</td>
<td>Sad</td>
<td>30</td>
</tr>
<tr>
<td>Bobbie</td>
<td>Sad</td>
<td>50</td>
</tr>
<tr>
<td>Casey</td>
<td>Happy</td>
<td>80</td>
</tr>
<tr>
<td>Harper</td>
<td>Happy</td>
<td>100</td>
</tr>
</tbody>
</table>

- **Students**: 4
- **Percent Happy**: 45%
- **Average Grade**: 50

### February

<table>
<thead>
<tr>
<th>Name</th>
<th>Affect</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex</td>
<td>Sad</td>
<td>30</td>
</tr>
<tr>
<td>Bobbie</td>
<td>Sad</td>
<td>50</td>
</tr>
<tr>
<td>Casey</td>
<td>Happy</td>
<td>80</td>
</tr>
<tr>
<td>Emerson</td>
<td>Sad</td>
<td>90</td>
</tr>
<tr>
<td>Harper</td>
<td>Happy</td>
<td>100</td>
</tr>
</tbody>
</table>

- **Students**: 5
- **Percent Happy**: 60%
- **Average Grade**: 75
If we ran the statistics different times, we would get different results

<table>
<thead>
<tr>
<th>Name</th>
<th>Affect</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex</td>
<td>Sad</td>
<td>30</td>
</tr>
<tr>
<td>Bobbie</td>
<td>Sad</td>
<td>50</td>
</tr>
<tr>
<td>Casey</td>
<td>Happy</td>
<td>80</td>
</tr>
<tr>
<td>Harper</td>
<td>Happy</td>
<td>100</td>
</tr>
</tbody>
</table>

**Statistical Tabulation + noise**

- **January**
  - Students: 4
  - Percent Happy: 45%
  - Average Grade: 50

- **January**
  - Students: 4
  - Percent Happy: 55%
  - Average Grade: 75

- **January**
  - Students: 4
  - Percent Happy: 51%
  - Average Grade: 60

In this example, a *policy decision* requires that the number of students be accurately reported.
In this example, a *policy decision* requires that the exact number of students in the class be confidential.

<table>
<thead>
<tr>
<th>Name</th>
<th>Affect</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex</td>
<td>Sad</td>
<td>30</td>
</tr>
<tr>
<td>Bobbie</td>
<td>Sad</td>
<td>50</td>
</tr>
<tr>
<td>Casey</td>
<td>Happy</td>
<td>80</td>
</tr>
<tr>
<td>Harper</td>
<td>Happy</td>
<td>100</td>
</tr>
</tbody>
</table>

 statistical tabulation + noise

Students: 3
Percent Happy: 40%
Average Grade: 50

<table>
<thead>
<tr>
<th>Name</th>
<th>Affect</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex</td>
<td>Sad</td>
<td>30</td>
</tr>
<tr>
<td>Bobbie</td>
<td>Sad</td>
<td>50</td>
</tr>
<tr>
<td>Casey</td>
<td>Happy</td>
<td>80</td>
</tr>
<tr>
<td>Harper</td>
<td>Happy</td>
<td>100</td>
</tr>
</tbody>
</table>

 statistical tabulation + noise

Students: 6
Percent Happy: 45%
Average Grade: 45

<table>
<thead>
<tr>
<th>Name</th>
<th>Affect</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex</td>
<td>Sad</td>
<td>30</td>
</tr>
<tr>
<td>Bobbie</td>
<td>Sad</td>
<td>50</td>
</tr>
<tr>
<td>Casey</td>
<td>Happy</td>
<td>80</td>
</tr>
<tr>
<td>Harper</td>
<td>Happy</td>
<td>100</td>
</tr>
</tbody>
</table>

 statistical tabulation + noise

Students: 5
Percent Happy: 51%
Average Grade: 60

Data users understand that noise has been added.
Differential privacy uses the parameter $\varepsilon$ (epsilon) to describe the privacy/accuracy tradeoff.

- $\varepsilon = 0$ — No accuracy, full privacy
- $\varepsilon = \infty$ — No privacy, full accuracy

**How much noise do we add?**
That is a policy decision
Noise can be added in two places:
1) When data are collected.  2) When statistics are produced.

Input noise infusion:

<table>
<thead>
<tr>
<th>Name</th>
<th>Affect</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex</td>
<td>Sad + NOISE</td>
<td>30 + NOISE</td>
</tr>
<tr>
<td>Bobbie</td>
<td>Sad + NOISE</td>
<td>50 + NOISE</td>
</tr>
<tr>
<td>Casey</td>
<td>Happy + NOISE</td>
<td>80 + NOISE</td>
</tr>
<tr>
<td>Harper</td>
<td>Happy + NOISE</td>
<td>100 + NOISE</td>
</tr>
</tbody>
</table>

Advantages:
- Tabulator need not be trusted.
- More statistics do not pose additional privacy threats.

Output noise infusion:

<table>
<thead>
<tr>
<th>Name</th>
<th>Affect</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex</td>
<td>Sad</td>
<td>30</td>
</tr>
<tr>
<td>Bobbie</td>
<td>Sad</td>
<td>50</td>
</tr>
<tr>
<td>Casey</td>
<td>Happy</td>
<td>80</td>
</tr>
<tr>
<td>Harper</td>
<td>Happy</td>
<td>100</td>
</tr>
</tbody>
</table>

Advantages:
- More accurate for the same level of privacy
- Allows uses of confidential data that do not involve publication.
Other choices for policy makers

Where should the accuracy be spent?

What values should be reported exactly (with no privacy)

What are the possible bounds (sensitivity) of a person’s data?
  e.g. If reporting average student age, can students be 5..18 or 5..115?

How do we convey privacy guarantees to public?
Differential privacy was invented in 2006 by Dwork, McSherry, Nissim and Smith

Differential privacy is just 12 years old.

Today’s public key cryptography was invented in 1976-1978

Remember public key cryptography in 1990?

- No standardized implementations. No SSL/TLS. No S/MIME or PGP.
- Very few people knew how to build systems that used crypto.
In Summary

**Communications security:** Be careful when you get data from your sources.

**Storage security:** Be careful where you store data; use two-factor security.

**Publication security:** Be careful when you publish. Remember that data can be reverse-engineered if you do not take appropriate measures.

Questions?

Email: simson.l.garfinkel@census.gov