MODULE 1:
PROBLEM FORMULATION
Module 1: What is Privacy?

• Privacy Problem Statement

• What privacy is \textit{not} …
  – Encryption
  – Anonymization
  – Restricted Query Answering

• What \textit{is} privacy?
Statistical Database Privacy

Function provided by the analyst

Output can disclose sensitive information about individuals

Module 1  Tutorial: Differential Privacy in the Wild
Statistical Database Privacy

Privacy for individuals (controlled by a parameter $\varepsilon$)

$f_{private}(DB, \varepsilon)$

Person 1 $r_1$
Person 2 $r_2$
Person 3 $r_3$
Person N $r_N$
Statistical Database Privacy

Utility for analyst

\( f_{\text{private}}(DB) \approx f(DB) \)

\( f_{\text{private}}(DB, \varepsilon) \)

Server

DB

Person 1

\( r_1 \)

Person 2

\( r_2 \)

Person 3

\( r_3 \)

\ldots

Person N

\( r_N \)
Statistical Database Privacy (untrusted collector)

Server wants to compute $f$

Individuals do not want server to infer their records

Module 1

Tutorial: Differential Privacy in the Wild
Statistical Database Privacy (untrusted collector)

Perturb records to ensure privacy for individuals and Utility for server

Module 1 Tutorial: Differential Privacy in the Wild
## Statistical Databases in real-world applications

<table>
<thead>
<tr>
<th>Application</th>
<th>Data Collector</th>
<th>Private Information</th>
<th>Analyst</th>
<th>Function (utility)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical</td>
<td>Hospital</td>
<td>Disease</td>
<td>Epidemiologist</td>
<td>Correlation between disease and geography</td>
</tr>
<tr>
<td>Genome analysis</td>
<td>Hospital</td>
<td>Genome</td>
<td>Statistician/Researcher</td>
<td>Correlation between genome and disease</td>
</tr>
<tr>
<td>Advertising</td>
<td>Google/FB/Y!</td>
<td>Clicks/Browsing</td>
<td>Advertiser</td>
<td>Number of clicks on an ad by age/region/gender ...</td>
</tr>
<tr>
<td>Social Recommendations</td>
<td>Facebook</td>
<td>Friend links / profile</td>
<td>Another user</td>
<td>Recommend other users or ads to users based on social network</td>
</tr>
</tbody>
</table>
Statistical Databases in real-world applications

• Settings where data collector may not be trusted

<table>
<thead>
<tr>
<th>Application</th>
<th>Data Collector</th>
<th>Private Information</th>
<th>Function (utility)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location Services</td>
<td>Verizon/AT&amp;T</td>
<td>Location</td>
<td>Traffic prediction</td>
</tr>
<tr>
<td>Recommendations</td>
<td>Amazon/Google</td>
<td>Purchase history</td>
<td>Recommendation model</td>
</tr>
<tr>
<td>Traffic Shaping</td>
<td>Internet Service Provider</td>
<td>Browsing history</td>
<td>Traffic pattern of groups of users</td>
</tr>
</tbody>
</table>
Privacy is *not* ...
Statistical Database Privacy is not …

• Encryption:
Statistical Database Privacy is not …

• Encryption:
  Alice sends a message to Bob such that Trudy (attacker) does not learn the message. Bob should get the correct message …

• Statistical Database Privacy:
  Bob (attacker) can access a database
  - Bob must learn aggregate statistics, but
  - Bob must not learn new information about individuals in database.
Statistical Database Privacy is not ...

- Computation on Encrypted Data:
Statistical Database Privacy is not …

- Computation on Encrypted Data:
  - Alice stores encrypted data on a server controlled by Bob (attacker).
  - Server returns correct query answers to Alice, without Bob learning *anything* about the data.

- Statistical Database Privacy:
  - Bob is allowed to learn aggregate properties of the database.
Statistical Database Privacy is not ...

• The Millionaires Problem:
Statistical Database Privacy is not …

• Secure Multiparty Computation:
  - A set of agents each having a private input $x_i$ …
  - … Want to compute a function $f(x_1, x_2, …, x_k)$
  - Each agent can learn the true answer, but must learn no other information than what can be inferred from their private input and the answer.

• Statistical Database Privacy:
  - Function output must not disclose individual inputs.
Statistical Database Privacy is not …

• Access Control:
Statistical Database Privacy is not …

• Access Control:
  - A set of agents want to access a set of resources (could be files or records in a database)
  - Access control rules specify who is allowed to access (or not access) certain resources.
  - ‘Not access’ usually means no information must be disclosed

• Statistical Database:
  - A single database and a single agent
  - Want to release aggregate statistics about a set of records without allowing access to individual records
Privacy Problems

• In today’s cloud context, a number of privacy problems arise:
  – Encryption when communicating data across an unsecure channel
  – Secure Multiparty Computation when different parties want to compute on a function on their private data without using a centralized third party
  – Computing on encrypted data when one wants to use an unsecure cloud for computation
  – Access control when different users own different parts of the data

• Statistical Database Privacy:
  Quantifying (and bounding) the amount of information disclosed about individual records by the output of a valid computation.
What *is* privacy?
The Massachusetts Governor Privacy Breach [Sweeney IJUFKS 2002]

Medical Data Release

- Name
- SSN
- Visit Date
- Diagnosis
- Procedure
- Medication
- Total Charge
- Zip
- Birth date
- Sex
The Massachusetts Governor Privacy Breach [Sweeney IJUFKS 2002]

Medical Data Release

- Name
- SSN
- Visit Date
- Diagnosis
- Procedure
- Medication
- Total Charge

Voter List

- Name
- Address
- Date Registered
- Party affiliation
- Date last voted

Module 1
Tutorial: Differential Privacy in the Wild
Linkage Attack

Medical Data Release

- Name
- SSN
- Visit Date
- Diagnosis
- Procedure
- Medication
- Total Charge

Voter List

- Name
- Address
- Date Registered
- Party affiliation
- Date last voted
- Zip
- Birth date
- Sex

Governor of MA uniquely identified using ZipCode, Birth Date, and Sex.

Name linked to Diagnosis
Linkage Attack

- 87% of US population uniquely identified using ZipCode, Birth Date, and Sex.

Medical Data Release
- Name
- SSN
- Visit Date
- Diagnosis
- Procedure
- Medication
- Total Charge

Voter List
- Name
- Address
- Date Registered
- Party Affiliation
- Date last voted

Quasi Identifier
Privacy Breach: Informal Definition

A privacy mechanism $M(D)$ that allows an unauthorized party to learn sensitive information about any individual in $D$, which could not have been learnt without access to $M(D)$. 
Alice

Is this a privacy breach? NO
Privacy Breach: Revised Definition

A privacy mechanism $M(D)$ that allows an unauthorized party to learn sensitive information about any individual Alice in $D$, which could not have been learnt without access to $M(D)$ if Alice was not in the dataset.
K-Anonymity: Avoiding Linkage Attacks

• If every row corresponds to one individual …

… every row should look like k-1 other rows based on the quasi-identifier attributes
## K-Anonymity

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Nationality</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>13053</td>
<td>28</td>
<td>Russian</td>
<td>Heart</td>
</tr>
<tr>
<td>13068</td>
<td>29</td>
<td>American</td>
<td>Heart</td>
</tr>
<tr>
<td>13068</td>
<td>21</td>
<td>Japanese</td>
<td>Flu</td>
</tr>
<tr>
<td>13053</td>
<td>23</td>
<td>American</td>
<td>Flu</td>
</tr>
<tr>
<td>14853</td>
<td>50</td>
<td>Indian</td>
<td>Cancer</td>
</tr>
<tr>
<td>14853</td>
<td>55</td>
<td>Russian</td>
<td>Heart</td>
</tr>
<tr>
<td>14850</td>
<td>47</td>
<td>American</td>
<td>Flu</td>
</tr>
<tr>
<td>14850</td>
<td>59</td>
<td>American</td>
<td>Flu</td>
</tr>
<tr>
<td>13053</td>
<td>31</td>
<td>American</td>
<td>Cancer</td>
</tr>
<tr>
<td>13053</td>
<td>37</td>
<td>Indian</td>
<td>Cancer</td>
</tr>
<tr>
<td>13068</td>
<td>36</td>
<td>Japanese</td>
<td>Cancer</td>
</tr>
<tr>
<td>13068</td>
<td>32</td>
<td>American</td>
<td>Cancer</td>
</tr>
</tbody>
</table>

### Anonymized Table

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Nationality</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Flu</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Flu</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Flu</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Flu</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
</tbody>
</table>
Problem: Background knowledge

Adversary knows prior knowledge about Umeko

Adversary learns Umeko has Cancer

<table>
<thead>
<tr>
<th>Name</th>
<th>Zip</th>
<th>Age</th>
<th>Nat.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Umeko</td>
<td>13053</td>
<td>25</td>
<td>Japan</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Nationality</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Flu</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Flu</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
</tbody>
</table>
A privacy mechanism must be able to protect individuals’ privacy from attackers who may possess background knowledge.
Welcome to H-CUPnet

HCUPnet is a free, on-line query system based on data from the Healthcare Cost and Utilization Project (HCUP). It provides access to health statistics and information on hospital inpatient and emergency department utilization.

Begin your query here -

**Statistics on Hospital Stays**

- **National Statistics on All Stays**
  
  Create your own statistics for national and regional estimates on hospital use for all patients from the HCUP National (Nationwide) Inpatient Sample (NIS). Overview of the National (Nationwide) Inpatient Sample (NIS)

- **National Statistics on Mental Health Hospitalizations**
  
  Interested in acute care hospital stays for mental health and substance abuse? Create your own national statistics from the NIS.

- **State Statistics on All Stays**
  
  Create your own statistics on stays in hospitals for participating States from the HCUP State Inpatient Databases (SID). Overview of the State Inpatient Databases (SID)

**National Statistics on Children**

Create your own statistics for national estimates on use of hospitals by children (age 0-17 years) from the HCUP Kids’ Inpatient Database (KID). Overview of the Kids’ Inpatient Database (KID)

**National and State Statistics on Hospital Stays by Payer - Medicare, Medicaid, Private, Uninsured**

Interested in hospital stays billed to a specific payer? Create your own statistics for a payer, alone or compared to other payers from the NIS, KID, and SID.

**Quick National or State Statistics**

Ready-to-use tables on commonly requested information from the HCUP National (Nationwide) Inpatient Sample (NIS), the HCUP Kids’ Inpatient Database (KID), or the HCUP State Inpatient Databases (SID).
Hospital discharges in NJ of ovarian cancer patients, 2009

Counts less than k are suppressed achieving k-anonymity

<table>
<thead>
<tr>
<th>Age</th>
<th>#discharges</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian/Pcf HInd</th>
<th>Native American</th>
<th>Other</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#discharges</td>
<td>735</td>
<td>535</td>
<td>82</td>
<td>58</td>
<td>18</td>
<td>*</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>1-17</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>18-44</td>
<td>70</td>
<td>40</td>
<td>13</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>45-64</td>
<td>330</td>
<td>236</td>
<td>31</td>
<td>32</td>
<td>*</td>
<td>*</td>
<td>11</td>
<td>*</td>
</tr>
<tr>
<td>65-84</td>
<td>298</td>
<td>229</td>
<td>35</td>
<td>13</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>85+</td>
<td>34</td>
<td>29</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
# Hospital discharges in NJ of ovarian cancer patients, 2009

<table>
<thead>
<tr>
<th>Age</th>
<th>#discharges</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian/Pcf</th>
<th>Hispanic/Pcf</th>
<th>Native American</th>
<th>Other</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-17</td>
<td>3</td>
<td>1</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>1</td>
<td>*</td>
</tr>
<tr>
<td>18-44</td>
<td>70</td>
<td>40</td>
<td>13</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>1</td>
<td>*</td>
</tr>
<tr>
<td>45-64</td>
<td>330</td>
<td>236</td>
<td>31</td>
<td>32</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>65-84</td>
<td>298</td>
<td>229</td>
<td>35</td>
<td>13</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>85+</td>
<td>34</td>
<td>29</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\text{Total} & = 535 - (40 + 236 + 229 + 29) \\
\end{align*}
\]
#Hospital discharges in NJ of ovarian cancer patients, 2009

<table>
<thead>
<tr>
<th>Age</th>
<th>#discharges</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian/Pcf Hlnder</th>
<th>Native American</th>
<th>Other</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-17</td>
<td>3</td>
<td>1</td>
<td>[0-2]</td>
<td>[0-2]</td>
<td>[0-2]</td>
<td>[0-2]</td>
<td>[0-2]</td>
<td>[0-2]</td>
</tr>
<tr>
<td>18-44</td>
<td>70</td>
<td>40</td>
<td>13</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>45-64</td>
<td>330</td>
<td>236</td>
<td>31</td>
<td>32</td>
<td>*</td>
<td>*</td>
<td>11</td>
<td>*</td>
</tr>
<tr>
<td>65-84</td>
<td>298</td>
<td>229</td>
<td>35</td>
<td>13</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>85+</td>
<td>34</td>
<td>29</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
#Hospital discharges in NJ of ovarian cancer patients, 2009

<table>
<thead>
<tr>
<th>Age</th>
<th>#discharges</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian/Pcf HLnder</th>
<th>Native American</th>
<th>Other</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-17</td>
<td>3</td>
<td>1</td>
<td>[0-2]</td>
<td>[0-2]</td>
<td>[0-2]</td>
<td>[0-2]</td>
<td>[0-2]</td>
<td>[0-2]</td>
</tr>
<tr>
<td>18-44</td>
<td>70</td>
<td>40</td>
<td>13</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>45-64</td>
<td>330</td>
<td>236</td>
<td>31</td>
<td>32</td>
<td>*</td>
<td>*</td>
<td>11</td>
<td>*</td>
</tr>
<tr>
<td>65-84</td>
<td>298</td>
<td>229</td>
<td>35</td>
<td>13</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>85+</td>
<td>34</td>
<td>29</td>
<td>[1-3]</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
Can reconstruct tight bounds on rest of data

In fact, when linked with queries giving other statistics, we can figure out that exactly 1 Native American woman diagnosed with ovarian cancer went to a privately owned, not for profit, teaching hospital in New Jersey with more than 435 beds in 2009. Furthermore, the woman did not pay by private insurance, had a routine discharge, with a stay in the hospital of 33.5 days, with her home residence being in a county with 1 million plus residents (large fringe metro, suburbs), and her age was exactly 75 years.

[Vaidya et al AMIA 2013]
Multiple Release problem

• Privacy preserving access to data must necessarily release some information about individual records (to ensure utility)

• However, k-anonymous algorithms can reveal individual level information even with two releases.
A privacy mechanism must satisfy composition …

… or allow a graceful degradation of privacy with multiple invocations on the same data.
Postprocessing the output of a privacy mechanism must not change the privacy guarantee.
Privacy must not be achieved through obscurity.

Attacker must be assumed to know the algorithm used as well as all parameters.
Summary

• Statistical database privacy is the problem of releasing aggregates while not disclosing individual records.

• The problem is distinct from encryption, secure computation and access control.

• Defining privacy is non-trivial
  – Desiderata include resilience to background knowledge and composition and closure under postprocessing.