Information Preservation in Statistical Privacy and Bayesian Estimation of Unattributed Histograms

Bing-Rong Lin

Department of Computer Science & Engineering
Penn State University

August 14, 2013

(joint work with Daniel Kifer)
Outline

1. Introduction: Privacy, Utility, Usability
2. An Information Paradox
3. Properties for information preservation measures
4. A Consequence: Bayesian Decision Theory
5. Applications to Using Sanitized Data
**Scenarios**

- One-shot data publishing
  - Data owner publishes differentially private data
Scenarios

- One-shot data publishing
  - Data owner publishes differentially private data

Is it useful?
Scenarios

- One-shot data publishing
  - Data owner publishes differentially private data
  - Data publisher goes on vacation

Is it useful?

Query payment ? , ? , ?

Data user has limited privacy budget

Data owner sells privacy-preserving query answers for
Scenarios

- One-shot data publishing
  - Data owner publishes differentially private data
  - Data publisher goes on vacation
  - Data user must figure out how to use noisy data
Scenarios

- One-shot data publishing
  - Data owner publishes differentially private data
  - Data publisher goes on vacation
  - Data user must figure out how to use noisy data

- Query payment [?, ?, ?]
  - Data owner sells privacy-preserving query answers for
  - Data user has limited privacy budget
Utility Goals

- Scenarios incentivize the following behavior:
- First maximizes information content.
  - Data publishing: choose differentially private algorithm that best preserves statistical information.
  - Selling queries: choose $\varepsilon$-differentially private algorithm that maximizes information content for a fixed $\varepsilon$.
  - Choose algorithm $M$ maximizing some measure $\mu(M)$
Utility Goals

Scenarios incentivize the following behavior:

First maximizes information content.
- Data publishing: choose differentially private algorithm that best preserves statistical information.
- Selling queries: choose $\epsilon$-differentially private algorithm that maximizes information content for a fixed $\epsilon$.
- Choose algorithm $\mathcal{M}$ maximizing some measure $\mu(\mathcal{M})$

Then worry about usability
- Sanitized data may not be in the right form.
- Need to extract information
- Building statistical model
- Making business decisions
Utility Goals

- Scenarios incentivize the following behavior:
  - First maximizes information content.
    - Data publishing: choose differentially private algorithm that best preserves statistical information.
    - Selling queries: choose $\epsilon$-differentially private algorithm that maximizes information content for a fixed $\epsilon$.
    - Choose algorithm $\mathcal{M}$ maximizing some measure $\mu(\mathcal{M})$
  - Then worry about usability
    - Sanitized data may not be in the right form.
    - Need to extract information
    - Building statistical model
    - Making business decisions
- How to define $\mu$?
- How to use the sanitized data?
Outline

1. Introduction: Privacy, Utility, Usability
2. An Information Paradox
3. Properties for information preservation measures
4. A Consequence: Bayesian Decision Theory
5. Applications to Using Sanitized Data
Alice is conducting a survey.

Q: Do you like green eggs and ham?

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.75</td>
</tr>
<tr>
<td>No</td>
<td>0.5</td>
</tr>
</tbody>
</table>

**Algorithm** $M_1$

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.6</td>
</tr>
<tr>
<td>No</td>
<td>0.4</td>
</tr>
</tbody>
</table>

**Algorithm** $M_2$

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.5</td>
</tr>
<tr>
<td>No</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Alice must choose between sanitizers $M_1$ and $M_2$

$\mu$: worst-case error probability?

- Worst-case error probability of $M_1$: 0.5
- Worst-case error probability of $M_2$: 0.4
Defining \( \mu \) can be tricky

- Alice is conducting a survey.
  - Q: Do you like green eggs and ham?

<table>
<thead>
<tr>
<th>Input</th>
<th>Algorithm ( \mathcal{M}_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.75</td>
</tr>
<tr>
<td>No</td>
<td>0.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input</th>
<th>Algorithm ( \mathcal{M}_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.6</td>
</tr>
<tr>
<td>No</td>
<td>0.4</td>
</tr>
</tbody>
</table>

- Alice must choose between sanitizers \( \mathcal{M}_1 \) and \( \mathcal{M}_2 \)

- \( \mu \): worst-case error probability?
  - Worst-case error probability of \( \mathcal{M}_1 \): 0.5
  - Worst-case error probability of \( \mathcal{M}_2 \): 0.4
  - \( \mathcal{M}_2 \) is better?
Alice is conducting a survey.

- Q: Do you like green eggs and ham?

$\mathcal{M}_2$ is preferred by worst-case error.
Alice is conducting a survey.

Q: Do you like green eggs and ham?

$M_2$ is preferred by worst-case error

But consider algorithm $A$

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.8</td>
</tr>
<tr>
<td>No</td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm $A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>data $\rightarrow$ $M_1$ $\rightarrow$ $A$ $\rightarrow$ output</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.75</td>
</tr>
<tr>
<td>No</td>
<td>0.25</td>
</tr>
</tbody>
</table>

| Algorithm $M_1$ |

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.6</td>
</tr>
<tr>
<td>No</td>
<td>0.4</td>
</tr>
</tbody>
</table>

| Algorithm $M_2$ |

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.6</td>
</tr>
<tr>
<td>No</td>
<td>0.4</td>
</tr>
</tbody>
</table>
Alice is conducting a survey.
  Q: Do you like green eggs and ham?

$M_2$ is preferred by worst-case error

But consider algorithm $A$

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.8</td>
</tr>
<tr>
<td>No</td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.75</td>
</tr>
<tr>
<td>No</td>
<td>0.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.6</td>
</tr>
<tr>
<td>No</td>
<td>0.4</td>
</tr>
</tbody>
</table>

$M_2$ can be simulated by running $M_1$ on data then $A$ on the result.

$A \circ M_1 = M_2$

So $M_1$ preserves more information.

Need $M_2$? Post-process output of $M_1$.

We want a $\mu$ that never prefers $M_2$ over $M_1$. 
Choose the set of outputs $\emptyset$

Choose loss function $\mathcal{L}(D, \omega)$

- What we “lose” if we output $\omega$ when input is $D$.

Worst-case error $\mu(\mathcal{M}) = \max_D \sum_{\omega \in \emptyset} \mathcal{L}(D, \omega) P(\mathcal{M}(D) = \omega)$

Average error $\mu(\mathcal{M}) = E_D \left[ \sum_{\omega \in \emptyset} \mathcal{L}(D, \omega) P(\mathcal{M}(D) = \omega) \right]$
Choose the set of outputs $\emptyset$

Choose loss function $\mathcal{L}(D, \omega)$
- What we “lose” if we output $\omega$ when input is $D$.

Worst-case error $\mu(\mathcal{M}) = \max_D \sum_{\omega \in \emptyset} \mathcal{L}(D, \omega) P(\mathcal{M}(D) = \omega)$

Average error $\mu(\mathcal{M}) = E_D \left[ \sum_{\omega \in \emptyset} \mathcal{L}(D, \omega) P(\mathcal{M}(D) = \omega) \right]$

Theorem: Counterexamples always exist unless $\mu$ is constant.
- $\mu(\mathcal{M}_2) > \mu(\mathcal{M}_1)$ but $\mathcal{M}_2 = A \circ \mathcal{M}_1$

Conclusion: They do not measure information the way we want.

Conclusion: Heuristically derived utility measure may have undesirable properties.

Conclusion: Need a testable approach to define $\mu$. 
• Axiomatic approach is testable
• \( \mu \) should be chosen based on desirable properties.
• Goal: list properties, then derive \( \mu \).
1. Introduction: Privacy, Utility, Usability
2. An Information Paradox
3. Properties for information preservation measures
4. A Consequence: Bayesian Decision Theory
5. Applications to Using Sanitized Data
Properties for information preservation measures

Sufficiency

- intuition: if $M_2$ can be simulated by $M_1$, then $\mu(M_2) \leq \mu(M_1)$.

Axiom (Sufficiency Axiom)

If $M_2 = A \circ M_1$ then $\mu(M_2) \leq \mu(M_1)$

- Caveat: can be violated in other applications
  - $M_1$ or $A$ are computationally expensive
  - $M_1$ or $A$ are difficult to implement
  - Aspects of usability
  - Violations of Sufficiency axiom imply information/usability tradeoff (future work).
Intuition: small changes to $M_1$ cause small changes to amount of information preserved.

**Axiom (Continuity)**

$\mu$ should be continuous in some metric.

$$d(M_1, M_2) = \sup_{i \in I} \| P(M_1(i) = \bullet) - P(M_2(i) = \bullet) \|_1$$
Branching

bullet intuition: the loss in utility ($\mu$) should only depend on the loss in information.
Properties for information preservation measures

Branching

- **Intuition**: The loss in utility ($\mu$) should only depend on the loss in information.

(Data Sanitizer $\mathcal{M}$)

| Output $\omega_1$ | $P(\omega_1 | i_1)$ | $P(\omega_1 | i_2)$ | ... | $P(\omega_1 | i_n)$ |
|-------------------|---------------------|---------------------|-----|---------------------|
| Output $\omega_2$ | $P(\omega_2 | i_1)$ | $P(\omega_2 | i_2)$ | ... | $P(\omega_2 | i_n)$ |
| Output $\omega_3$ | $P(\omega_3 | i_1)$ | $P(\omega_3 | i_2)$ | ... | $P(\omega_3 | i_n)$ |
| ...               | ...                 | ...                 | ... | ...                 |

(Data Sanitizer $\mathcal{M}^*$ – merge first two rows of $\mathcal{M}$)

| Output $\omega^*$ | $P(\omega_1 \lor \omega_2 | i_1)$ | $P(\omega_1 \lor \omega_2 | i_2)$ | ... | $P(\omega_1 \lor \omega_2 | i_n)$ |
|-------------------|-----------------------------------|-----------------------------------|-----|-----------------------------------|
| Output $\omega_3$ | $P(\omega_3 | i_1)$ | $P(\omega_3 | i_2)$ | ... | $P(\omega_3 | i_n)$ |
| ...               | ...                 | ...                 | ... | ...                 |

**What information is lost in $\mathcal{M}^*$?**
- To distinguish between the case where $\mathcal{M}$ outputs $\omega_1$ and the case where $\mathcal{M}$ outputs $\omega_2$
Likelihood principle tells us
All of the information in a sample is contained in the likelihood function
- If output is $\omega_1$, statistical analysis should only depend on probabilities of generating $\omega_1$.
- If output is $\omega_2$, statistical analysis should only depend on probabilities of generating $\omega_2$.
- If output is $\omega_1 \lor \omega_2$ (we don’t know which one) then analysis should depend only on the probabilities of generating $\omega_1 \lor \omega_2$.

Axiom (Branching)

$$\mu(M) = \mu(M^*) + G(\begin{pmatrix} P(\omega_1 | i_1), & P(\omega_1 | i_2), & \ldots & P(\omega_1 | i_n) \\ P(\omega_2 | i_1), & P(\omega_2 | i_2), & \ldots & P(\omega_2 | i_n) \end{pmatrix})$$

This is a loss in utility (increase in error). Change only depends on
- Probabilities of generating $\omega_1$
- Probabilities of generating $\omega_2$
Outline

1. Introduction: Privacy, Utility, Usability
2. An Information Paradox
3. Properties for information preservation measures
4. A Consequence: Bayesian Decision Theory
5. Applications to Using Sanitized Data
What is Bayesian Decision Theory?

- Example: Data Analysis
- States: set of possible databases $\mathcal{I} = \{D_1, D_2, \ldots \}$
- Prior: $P(\text{data} = D_i)$
- Evidence: sanitized output $\omega = \mathcal{M}(\text{data})$
- Actions: $\mathcal{A}$
  - Set of possible model parameter values
  - Set of possible query answers
- Cost function $\mathcal{L} : \mathcal{A} \times \mathcal{I} \rightarrow \mathbb{R}$
- Choose action minimizing expected cost.

$$a^* = \arg\min_{a \in \mathcal{A}} \sum_{i=1}^{\infty} \mathcal{L}(a, D_i) P(\text{data} = D_i \mid \mathcal{M}(\text{data}) = \omega)$$

$$\text{cost for}(\mathcal{M}, \omega) = \min_{a \in \mathcal{A}} \sum_{i=1}^{\infty} \mathcal{L}(a, D_i) P(\text{data} = D_i \mid \mathcal{M}(\text{data}) = \omega)$$
A Consequence: Bayesian Decision Theory

Consequence of 3 Axioms

**Axiom (Sufficiency Axiom)**

If $\mathcal{M}_2 = A \circ \mathcal{M}_1$ then $\mu(\mathcal{M}_2) \leq \mu(\mathcal{M}_1)$

**Axiom (Continuity)**

$\mu$ should be continuous in some metric.

$$d(\mathcal{M}_1, \mathcal{M}_2) = \sup_{i \in I} ||P(\mathcal{M}_1(i) = \bullet) - P(\mathcal{M}_2(i) = \bullet)||_1$$

**Axiom (Branching)**

$$\mu(\mathcal{M}) = \mu(\mathcal{M}^*) + G \left( P(\omega_1 | i_1), P(\omega_1 | i_2), \ldots, P(\omega_1 | i_n) \right) + G \left( P(\omega_2 | i_1), P(\omega_2 | i_2), \ldots, P(\omega_2 | i_n) \right)$$
Theorem

\( \mu \) satisfies the three axioms if and only if \( \mu \) is (negative) average error of a Bayesian decision maker.

- There exists a set \( A \) of actions
- There exists a cost function \( \mathcal{L} \)
- There exists a prior \( P(\text{data} = D_i) \)
- Such that
  \[
  \mu(M) = -\sum_\omega \text{cost}_{\text{for}}(M, \omega)P(\omega)
  \]

Consequence: an optimization criteria in algorithm design.

Consequence: suggestion for how to analyze sanitized data.
  - Axioms do not mention priors, Bayes’ rule, etc.
  - Another (noncircular) justification for Bayesian methods.
  - Axiomatic approach is somewhat testable.
Outline

1. Introduction: Privacy, Utility, Usability
2. An Information Paradox
3. Properties for information preservation measures
4. A Consequence: Bayesian Decision Theory
5. Applications to Using Sanitized Data
Choosing a data sanitizer $M$ to maximize information measure $\mu$
  \[ \approx \text{choose } M \text{ to make sure Bayesian Decision Theory works well.} \]

Validation: Need to try Bayesian Decision Theory on many problems.

One problem: sorted histogram [?].

Add Laplace noise to each bucket for Differential Privacy

Goal: estimate the original sorted histogram
Choosing a data sanitizer $\mathcal{M}$ to maximize information measure $\mu$
  - $\approx$ choose $\mathcal{M}$ to make sure Bayesian Decision Theory works well.

Validation: Need to try Bayesian Decision Theory on many problems.

One problem: sorted histogram $[?]$. 

Add Laplace noise to each bucket for Differential Privacy

Goal: estimate the original sorted histogram

Nontrivial Problem
  - Prior methods not applicable (e.g., $[?]$).
Sorted Histogram Problem

- **What does Bayesian decision do?**
- **Components of Bayesian decision**
  - Prior: uniform over sorted histograms
  - Set of actions: sorted histograms
  - Loss function: squared loss between original and reconstructed histogram
  - Evidence: a sanitized output $\omega = M(data)$
- **Best action**: expected posterior sorted histogram $E_{D|\omega}[D]$.
- **But how?**
  - The problem can be modeled by hidden Markov Model (HMM) and the algorithm is based on forward-backward algorithm.
- **Comparisons**
  - Least-squares (LS) [?] (does not use probabilities of output).
  - Maximum likelihood (ML) (does not use loss function).
Number of bins (n): 11,342
Maximum value (m): 1,678
Running Time:
- LS: $O(n)$
- ML: $O(n \log(n))$
- HMM: $O(nm^2)$
## Error Results - Social Network

<table>
<thead>
<tr>
<th>Estimator</th>
<th>( \mu(\text{error}) )</th>
<th>( \sigma(\text{error}) )</th>
<th>Wins</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{S}_{HMM} )</td>
<td>577.2</td>
<td>49.1</td>
<td>100</td>
</tr>
<tr>
<td>( \hat{S}_{LS} )</td>
<td>1,111.2</td>
<td>66.2</td>
<td>0</td>
</tr>
<tr>
<td>( \hat{S}_{ML} )</td>
<td>889.4</td>
<td>52.0</td>
<td>0</td>
</tr>
</tbody>
</table>

\( \epsilon = 1 \)

<table>
<thead>
<tr>
<th>Estimator</th>
<th>( \mu(\text{error}) )</th>
<th>( \sigma(\text{error}) )</th>
<th>Wins</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{S}_{HMM} )</td>
<td>1,494.5</td>
<td>114.0</td>
<td>100</td>
</tr>
<tr>
<td>( \hat{S}_{LS} )</td>
<td>2,793.7</td>
<td>163.4</td>
<td>0</td>
</tr>
<tr>
<td>( \hat{S}_{ML} )</td>
<td>2,197.8</td>
<td>147.6</td>
<td>0</td>
</tr>
</tbody>
</table>

\( \epsilon = 0.5 \)

<table>
<thead>
<tr>
<th>Estimator</th>
<th>( \mu(\text{error}) )</th>
<th>( \sigma(\text{error}) )</th>
<th>Wins</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{S}_{HMM} )</td>
<td>7,376.5</td>
<td>1,360.3</td>
<td>100</td>
</tr>
<tr>
<td>( \hat{S}_{LS} )</td>
<td>21,983.1</td>
<td>2,324.9</td>
<td>0</td>
</tr>
<tr>
<td>( \hat{S}_{ML} )</td>
<td>16,430.9</td>
<td>1,720.4</td>
<td>0</td>
</tr>
</tbody>
</table>

\( \epsilon = 0.1 \)
Applications to Using Sanitized Data

Error Results - Network

- **Number of bins (n):** 65,536
- **Maximum value (m):** 1,423
- **Running Time:**
  - LS: $O(n)$
  - ML: $O(n \log(n))$
  - HMM: $O(nm^2)$

Number of bins (n): 65,536
Maximum value (m): 1,423
Running Time:
- LS: $O(n)$
- ML: $O(n \log(n))$
- HMM: $O(nm^2)$
## Error Results - Network

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$\mu$(error)</th>
<th>$\sigma$(error)</th>
<th>Wins</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{S}_{HMM}$</td>
<td>166.3</td>
<td>29.4</td>
<td>100</td>
</tr>
<tr>
<td>$\hat{S}_{LS}$</td>
<td>264.9</td>
<td>33.9</td>
<td>0</td>
</tr>
<tr>
<td>$\hat{S}_{ML}$</td>
<td>227.3</td>
<td>31.7</td>
<td>0</td>
</tr>
<tr>
<td>$\epsilon = 0.5$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{S}_{HMM}$</td>
<td>438.8</td>
<td>73.2</td>
<td>100</td>
</tr>
<tr>
<td>$\hat{S}_{LS}$</td>
<td>751.8</td>
<td>98.8</td>
<td>0</td>
</tr>
<tr>
<td>$\hat{S}_{ML}$</td>
<td>630.6</td>
<td>87.6</td>
<td>0</td>
</tr>
<tr>
<td>$\epsilon = 0.1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{S}_{HMM}$</td>
<td>4,576.0</td>
<td>1,336.4</td>
<td>99</td>
</tr>
<tr>
<td>$\hat{S}_{LS}$</td>
<td>8,297.5</td>
<td>1,750.0</td>
<td>0</td>
</tr>
<tr>
<td>$\hat{S}_{ML}$</td>
<td>6,972.9</td>
<td>1,556.7</td>
<td>1</td>
</tr>
</tbody>
</table>
Number of bins (n): 32,768

Maximum value (m): 496

Running Time:
- LS: $O(n)$
- ML: $O(n \log(n))$
- HMM: $O(nm^2)$
## Error Results - Query Log

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$\mu$(error)</th>
<th>$\sigma$(error)</th>
<th>Wins</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{S}_{HMM}$</td>
<td>384.6</td>
<td>35.0</td>
<td>100</td>
</tr>
<tr>
<td>$\hat{S}_{LS}$</td>
<td>779.1</td>
<td>49.5</td>
<td>0</td>
</tr>
<tr>
<td>$\hat{S}_{ML}$</td>
<td>624.2</td>
<td>43.2</td>
<td>0</td>
</tr>
<tr>
<td>$\epsilon = 1$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$\mu$(error)</th>
<th>$\sigma$(error)</th>
<th>Wins</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{S}_{HMM}$</td>
<td>1,024.5</td>
<td>111.5</td>
<td>100</td>
</tr>
<tr>
<td>$\hat{S}_{LS}$</td>
<td>2,121.6</td>
<td>144.5</td>
<td>0</td>
</tr>
<tr>
<td>$\hat{S}_{ML}$</td>
<td>1,664.2</td>
<td>124.9</td>
<td>0</td>
</tr>
<tr>
<td>$\epsilon = 0.5$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$\mu$(error)</th>
<th>$\sigma$(error)</th>
<th>Wins</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{S}_{HMM}$</td>
<td>8,277.5</td>
<td>1,454.6</td>
<td>100</td>
</tr>
<tr>
<td>$\hat{S}_{LS}$</td>
<td>17,261.3</td>
<td>2,070.7</td>
<td>0</td>
</tr>
<tr>
<td>$\hat{S}_{ML}$</td>
<td>13,062.2</td>
<td>1,507.6</td>
<td>0</td>
</tr>
<tr>
<td>$\epsilon = 0.1$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Applications to Using Sanitized Data

Conclusions

- Formal justification on using expected error of Bayesian decision maker as a information preservation measure
- Bayesian decision theory should play a role in processing sanitized data
- Need efficient algorithms for
  - Bayesian decision for complex noise distribution
  - Designing privacy algorithm maximizing information retention
- Information preservation and usability tradeoff need to be studied more formally
Questions?