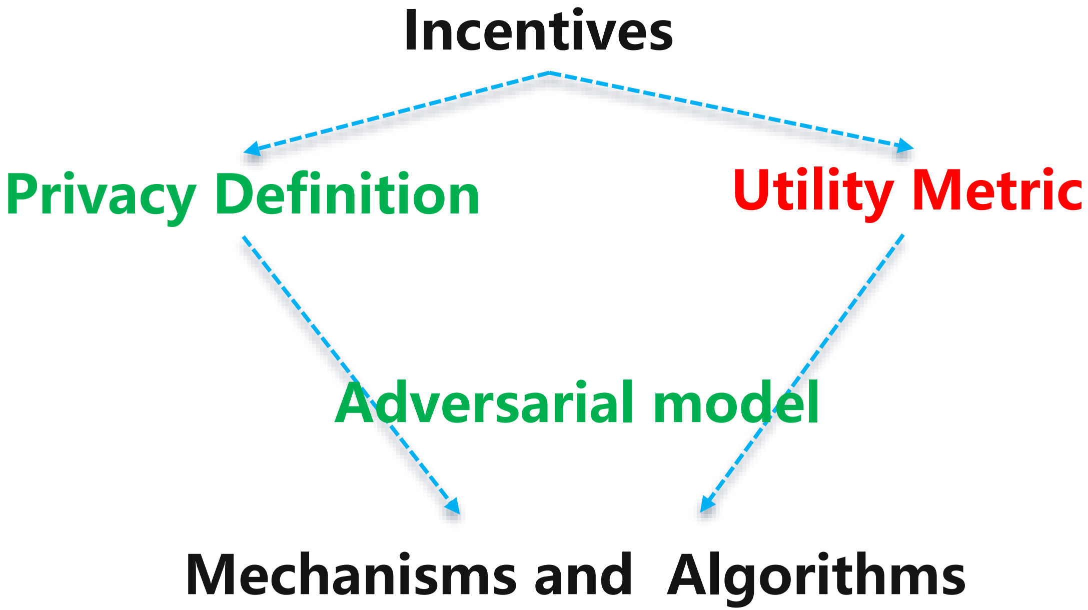
数据隐私方法伦理和实践
Methodology, Ethics and Practice of Data Privacy

3. 数据可用性
Data Utility

张兰
中国科学技术大学 计算机学院
2020春季

How Much



Incentives

Privacy Definition

Utility Metric

Adversarial model

Mechanisms and Algorithms

A data publisher seeks to release data that are not only safe, but also useful.

The temptation to form premature theories upon insufficient data is the bane of our profession.

-Sherlock Holmes (Sir Arthur Conan Doyle)

Measures of Utility

- **Quantitative measures of information loss**
  - Simple example: number of rows suppressed in a table
- **Quality of the results for queries in some fixed class**
  - Hope the class is representative, so other uses have low distortion
  - Costly: some methods enumerate all queries, or all anonymizations
- **Empirical Evaluation**
  - Perform experiments with a reasonable workload on the result
  - Compare to results on original data (e.g. Netflix prize problems)
- **Combinations of multiple methods**
  - Optimize for some surrogate, but also evaluate on real queries

A data publisher seeks to release data that are not only safe, but also useful.

1. Quantitative measures of information loss

[1] P. Samarati, “Protecting respondents’ identities in microdata release,” in Transactions on Knowledge and Data Engineering, pp. 1010–1027, 2001

[2] A. Meyerson and R. Williams, “On the complexity of optimal *k*-anonymity,” in *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, 2004.

[3] V. S. Iyengar, “Transforming data to satisfy privacy constraints,” in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.

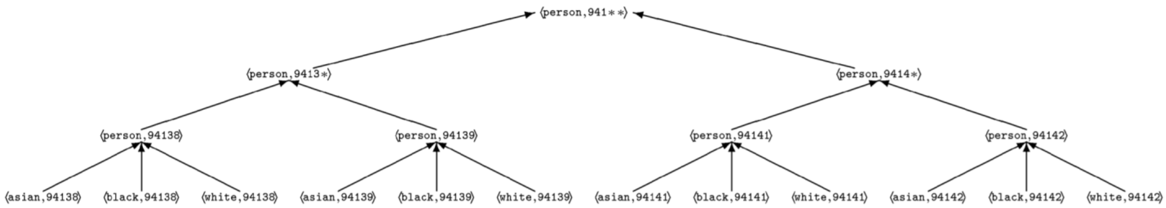
[4] R. Bayardo and R. Agrawal, “Data privacy through optimal k-anonymity,” in Proceedings of the 21st International Conference on Data Engineering (ICDE), 2005.

[5] M. E. Nergiz and C. Clifton, “Thoughts on k-anonymization,” Data & Knowledge Engineering, vol. 63, no. 3, pp. 622–645, 2007.

Generalization/Suppression Counting

- **Number of anonymization operations** performed on a data set.
  - E.g., changing “age =20” to “age ∈ [10 – 30]”
  - E.g., if generalization is the only operation being performed, then measure information loss by the number of generalization steps performed.
- **Generalization height**[1]:

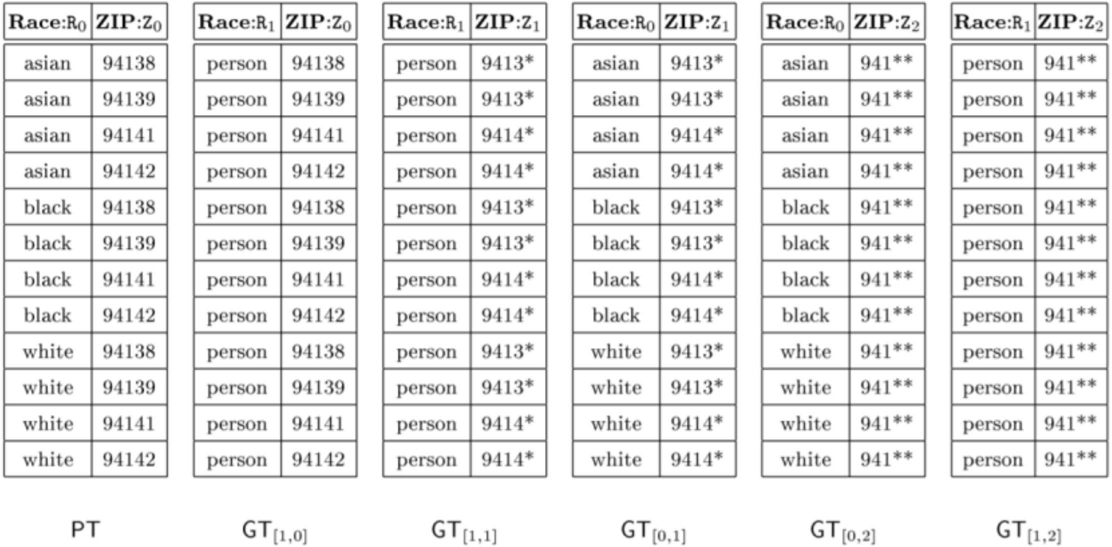
Generalization Strategy 1



- **Total number of attribute values that were suppressed**[2].
- Weighted version of these methods

Generalization/Suppression Counting

- Problem: not all operations affect utility in the same way



Race:R_0	**ZIP**:Z_0
asian	94138
asian	94139
asian	94141
asian	94142
black	94138
black	94139
black	94141
black	94142
white	94138
white	94139
white	94141
white	94142

PT

Race:R_1	**ZIP**:Z_0
person	94138
person	94139
person	94141
person	94142
person	94138
person	94139
person	94141
person	94142
person	94138
person	94139
person	94141
person	94142

$GT_{[1,0]}$

Race:R_1	**ZIP**:Z_1
person	9413*
person	9413*
person	9414*
person	9414*
person	9413*
person	9413*
person	9414*
person	9414*
person	9413*
person	9413*
person	9414*
person	9414*

$GT_{[1,1]}$

Race:R_0	**ZIP**:Z_1
asian	9413*
asian	9413*
asian	9414*
asian	9414*
black	9413*
black	9413*
black	9414*
black	9414*
white	9413*
white	9413*
white	9414*
white	9414*

$GT_{[0,1]}$

Race:R_0	**ZIP**:Z_2
asian	941**
asian	941**
asian	941**
asian	941**
black	941**
black	941**
black	941**
black	941**
white	941**
white	941**
white	941**
white	941**

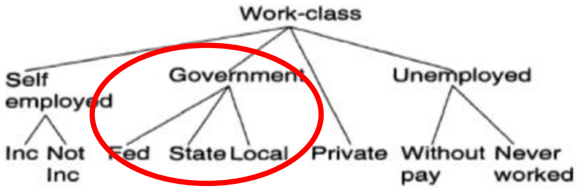
$GT_{[0,2]}$

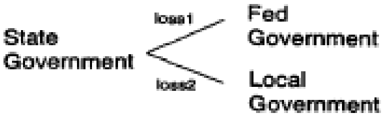
Race:R_1	**ZIP**:Z_2
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**

$GT_{[1,2]}$

Loss Metric (LM)

- LM[3] is defined in terms of a normalized loss for each attribute of every tuple.

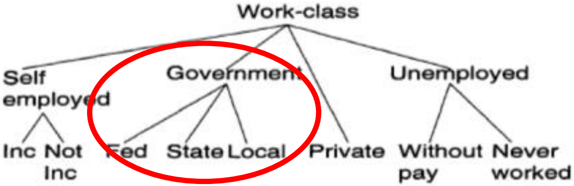


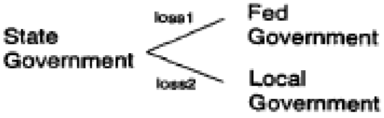


- Quantify the loss when a leaf node value cannot be disambiguated from another value due to the generalization.
- **Categorical attribute A**: For a tuple t, suppose the value of t[A] has been generalized to x. Letting |A| represent the total number of leaf nodes in the tree; letting M represent the number of leaf nodes in the subtree rooted at x, then the **loss for t[A] is (M – 1)/(|A| – 1).**
- What is the loss for “State”?
- The loss for attribute A is the average of the loss for all tuples t. The LM for the entire data set is the sum of the losses for each attribute.

Loss Metric (LM)

- LM[3] is defined in terms of a normalized loss for each attribute of every tuple.





- Quantify the loss when a leaf node value cannot be disambiguated from another value due to the generalization.
- **Categorical attribute A**: For a tuple t, suppose the value of t[A] has been generalized to x. Letting |A| represent the total number of leaf nodes in the tree; letting M represent the number of leaf nodes in the subtree rooted at x, then the **loss for t[A] is (M – 1)/(|A| – 1).**
- What is the loss for “State”? **2/7**
- The loss for attribute A is the average of the loss for all tuples t. The LM for the entire data set is the sum of the losses for each attribute.

Loss Metric (LM)

- LM[3] is defined in terms of a normalized loss for each attribute of every tuple.
- **Numerical information**: For a tuple t, suppose the value of t[A] has been generalized to an interval i [Li, Ui]. Let the lower and upper bounds in the table for A be L and U. The normalized loss for this entry is given by **(Ui - Li)/(U - L)**.

Zipcode	Age	Salary	Disease
476**	20-30	20-40K	Gastric Ulcer
476**	20-30	20-40K	Gastritis
476**	20-30	20-40K	Stomach Cancer
4790*	30-40	40-60K	Gastritis
4790*	30-40	40-60K	Flu
4790*	30-40	40-60K	Bronchitis

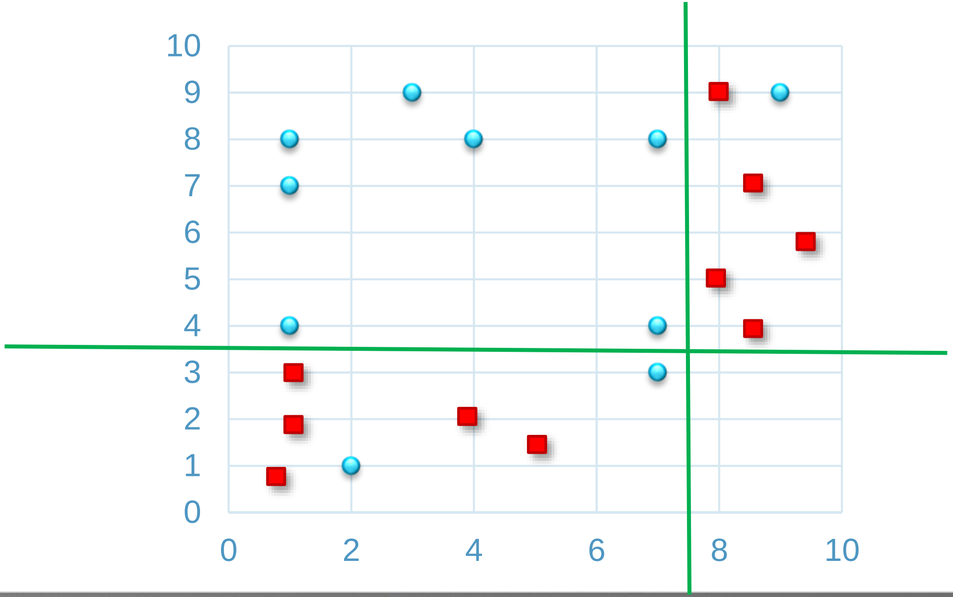
- The loss for age is? For salary is?

Classification Metric (CM)

- One possible use for the released table is to build predictive models for some attribute, e.g., recommendation.
- CM is designed to measure the **effect of the anonymization on a hypothetical classifier**. It penalizes impure groups that contain rows with different class labels.
- There is a distinguished class attribute, and tuples are placed into groups (by quasi-identifier value). Each tuple incurs a penalty of 1 if it is suppressed or if its class attribute is not the same as the majority class attribute in the group.
- The classification metric is defined as the average of the penalties of all tuples.

Classification Metric (CM)

- One possible use for the released table is to build predictive models for some attribute, e.g., recommendation.
- K=? CM=?



Discernibility Metric (DM)

- DM[4] assigns a penalty to each tuple based on how many other tuples in the database are indistinguishable from it.
- It works naturally in the k-anonymity framework.
- Impart a “penalty” on each tuple that reflects the information loss associated with its transformation or suppression.
  - For a database of size n, DM assigns a penalty of n for each suppressed tuple.
  - If a tuple is not suppressed, the penalty it receives is the total number of tuples in the database having the same quasi-identifier values.

$$C_{\mathrm{DM}}(g,k) = \sum_{\forall E \text{s.t.} |E| \geq k} |E|^2 + \sum_{\forall E \text{s.t.} |E| < k} |D||E|$$

Ambiguity Metric (AM)

- AM[5] is especially suitable for the *k*-anonymity framework.

- For each tuple $t*$ in the sanitized data, AM considers **the number of possible tuples in the *domain* of the data that could have been mapped (generalized) to $t*$.** This number is the ambiguity of $t*$.

- The AM for the sanitized data set is then the average ambiguity for all tuples in the sanitized data.

Discussion

- LM, CM, DM, AM are based on counting and oblivious to the distribution of actual attribute values in the data.
  - If the values of an attribute within a range is uniformed distributed and independent of all other attributes, then the generalization would have little effect on the analysis.
  - If the distribution is skewed, then different distributions could bias the analyst’s results.

Information Loss based on Distribution

- Measure the distance between the original probability distribution and the probability distribution reconstructed from the sanitized data.
- The larger the distance is, the greater the information loss.

- There are many ways of interpreting the sanitized data as a probability distribution.
  - If the sanitized version of the data is a set of **histograms**, then the histograms can be interpreted as constraints and the probability distribution p2 is the maximum entropy distribution consistent with those constraints.
  - Posit a **statistical model** such that the sanitized data form the sufficient statistics of the model.

Distance of Two Probabilistic Distributions

- Two distributions
  - $P = (p_1, p_2, \dots, p_m)$, $Q = (q_1, q_2, \dots, q_m)$
- Kullback-Leibler(KL) divergence: $D[P,Q] = \sum_{i-1}^{m} p_i log \frac{p_1}{q_i}$
- Entropy change: $D[P,Q] = \sum_{i-1}^{m} p_i log p_i$ -$\sum_{i-1}^{m} q_i log q_i$
- L_p norm: $D[P,Q] = \{\sum_{i=1}^{m} |p_i - q_i|^p\}^{1/p}$
- Variational distance: $D[P,Q] = \sum_{i=1}^{m} \frac{1}{2} |p_i - q_i|$
- Hellinger Distance: $D[P,Q] = \sqrt{\sum_m \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 /2}$
- Earth Mover's Distance

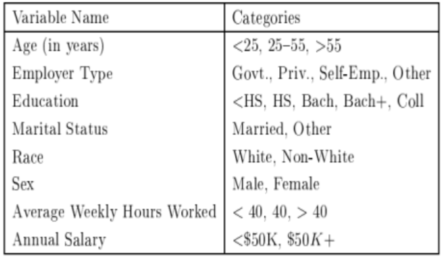
$$D[P,Q] = WORK(P,Q,F) = \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij}$$

KL-divergence & Entropy change

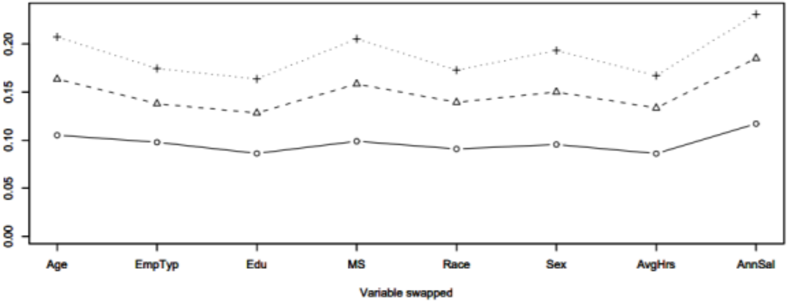
- $D[P,Q] = \sum_{i-1}^{m} p_i log \frac{p_1}{q_i}$
- Example: histograms form the sufficient statistics for a class of models known as log-linear models.
    - log-linear model takes the form of a function whose logarithm equals a linear combination of the parameters of the model, which makes it possible to apply (possibly multivariate) linear regression.
    - A model that overfits the original data has the maximum likelihood L_1 on this data set; A model using the sanitized data as sufficient statistics has lower likelihood L_1.
    - $log \frac{L_1}{L_2}$ is known as the log-likelihood ratio, measures the amount of likelihood that is not captured by the model built from sanitized data
    - The log-likelihood ratio is formally equivalent to KL-divergence.
- Entropy change: positive value indicate the generalization increases the uncertainty

Hellinger Distance

- $D[P,Q] = \sqrt{\sum_m \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 / 2}$
- $D[P,Q]$ is as the sine of the angle between the Hilbert vectors representing $\sqrt{p}$ and $\sqrt{q}$. Each square-root density can itself be interpreted as a point on the unit sphere in a real Hilbert space.
- 1%, 5%, 10% swap proportion

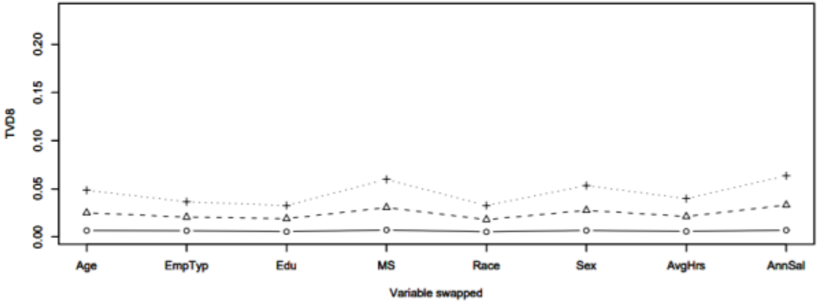


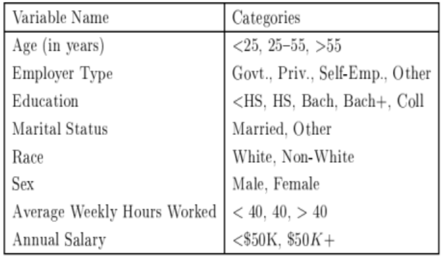
Variable Name	Categories
Age (in years)	<25, 25–55, >55
Employer Type	Govt., Priv., Self-Emp., Other
Education	<HS, HS, Bach, Bach+, Coll
Marital Status	Married, Other
Race	White, Non-White
Sex	Male, Female
Average Weekly Hours Worked	<40, 40, >40
Annual Salary	<\$50K, \$50K+



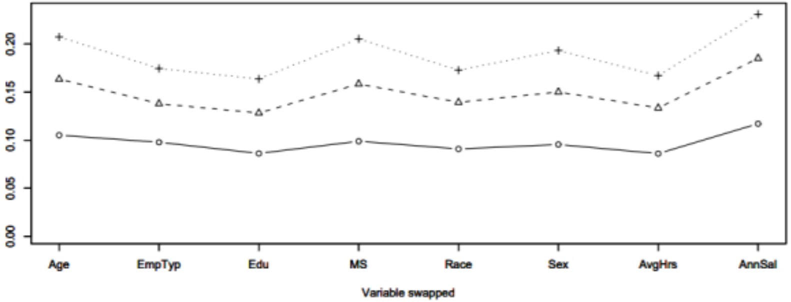
Hellinger Distance

- $D[P,Q] = \sqrt{\Sigma_m(}$
- $D[P,Q]$ is as th representing $\sqrt{i}$ interpreted as a
- 1%, 5%, 10% sw





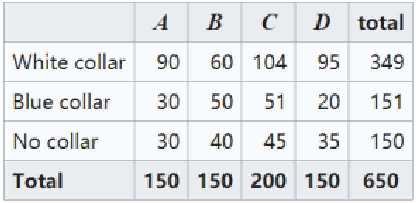
Variable Name	Categories
Age (in years)	<25, 25–55, >55
Employer Type	Govt., Priv., Self-Emp., Other
Education	<HS, HS, Bach, Bach+, Coll
Marital Status	Married, Other
Race	White, Non-White
Sex	Male, Female
Average Weekly Hours Worked	< 40, 40, > 40
Annual Salary	<\$50K, \$50K+



Bivariate Measures

- For a pair of attributes A and B, they compute the χ2 statistic in both the original data and the sanitized data.
- χ2 statistic *(test of independency)*

Contingency Table



	A	*B*	*C*	*D*	total
White collar	90	60	104	95	349
Blue collar	30	50	51	20	151
No collar	30	40	45	35	150
Total	**150**	**150**	**200**	**150**	**650**

1,000,000 residents with four neighborhoods. The null hypothesis is that each person's neighborhood of residence is independent of the person's occupational classification.

$$150 \times \frac{349}{650} \approx 80.54$$

$$\frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \frac{(90 - 80.54)^2}{80.54} \approx 1.11$$

- Cramer’s V: for a $m \times n$ contingency table, N samples
  - $V = \sqrt{\frac{\chi^2}{N \min(m-1, n-1)}}$
  - Varies from 0 (no association) to 1 (complete association)
  - Information loss: change of V
- Pearson’s contingency coefficient $C = \sqrt{\frac{\chi^2}{\chi^2 + \text{n}}}$

Workload-Aware Metrics

- The utility metric should depend on the intended uses of the sanitized data (in cases where the use is known beforehand).
  - Classification: weighted average of the entropy of the class attribute in each group. If there are multiple class attributes, then the total information loss is the sum of the information loss for each attribute.
  - Regression: the class attribute is continuous, so the information loss is measured as the weighted average of the variance of the class attribute in each group.
  - Count queries: information loss for each query is called *imprecision*-the number of points in all groups that overlap with the selection region of the query minus the true answer.

First- and Second-Order Statistics

- Use measures of information loss that are minimized when the original data and the sanitized data have the same first- and second-order statistics.
  - Assume there are p attributes (all numeric) and n tuples.
  - Mean variation: $\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{|x_{ij}-x'_{ij}|}{|x_{ij}|}$
  - Variation of the means: $\frac{1}{p}\sum_{i=1}^{p}\frac{|\mu_i-\mu'_i|}{|\mu_i|}.$
  - Variation of covariances: $\frac{1}{p(p+1)/2}\sum_{i=1}^{p}\sum_{1\leq j\leq i}\frac{|v_{ij}-v'_{ij}|}{|v_{ij}|}$
  - variation of variances: $\frac{1}{p}\sum_{i=1}^{p}\frac{|v_{ii}-v'_{ii}|}{|v_{ii}|}$
  - Clearly countless variations of these measures can be produced.

Analytical Validity

- The amount of information present in the sanitized data is known as *analytical validity*,
- It is evaluated by building models over both the original data and the sanitized data, and then comparing the learned parameters.
- Usually this is done by computing confidence intervals for the parameters learned from the original data and observing how many times the parameters from the sanitized model fall into the computed confidence intervals.
- Invariance： anonymization schemes should be devised so that they do not alter the output of pre-selected data mining algorithms.

Discussion

- Recent results indicate that utility optimization approach should be used with caution.
  - Poorly chosen measure of information loss could degrade the quality of the sanitized data. For example, recent work by Nergiz and Clifton has shown experimentally that if the goal is to build a good classifier from sanitized data, then optimizing for the LM, DM, CM, or AM metrics (discussed in the following sections) may provide little benefit.
  - In certain cases, the act of optimizing an information loss measure subject to privacy constraints can itself leak additional information.

“

> **Errors using inadequate data are much less than those using no data at all.**
>
> — Charles Babbage

How to make good use of data, given that they have been sanitized.

2. Quality of the results for queries

[1] N. N. Dalvi and D. Suciu, “Efficient query evaluation on probabilistic databases,” *VLDB Journal*, vol. 16, no. 4, pp. 523–544, 2007.
[2] D. Burdick, P. M. Deshpande, T. S. Jayram, R. Ramakrishnan, and S. Vaithyanathan, “Efficient allocation algorithms for OLAP over imprecise data,” in Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB), 2006.
[3] D. Burdick, P. M. Deshpande, T. S. Jayram, R. Ramakrishnan, and S. Vaithyanathan, “OLAP over uncertain and imprecise data,” VLDB Journal, vol. 16, no. 1, pp. 123–144, 2007.
[4] R. Agrawal, R. Srikant, and D. Thomas, “Privacy Preserving OLAP,” in Proceedings of the 23th ACM SIGMOD Conference on Management of Data, 2004.
[5] R. J. A. Little, “Statistical analysis of masked data,” Journal of Official Statistics, vol. 9, no. 2, pp. 407–426, 1993.

Queries over Sanitized Data

- The resulting sanitized data sets are usually **imprecise** (e.g., some attribute values have been generalized) or **probabilistic** (e.g., attribute values have been perturbed with random noise).

- Query processing for imprecise and uncertain data is an active and extensive research area in its own right

Example

Q1: the number of patients who have cancer and are less than 40 years old.

Q2: the number of patients who have cancer and are less than 50 years old.

	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	20-29	Any	130**	Asian	Heart disease
2	Bruce	20-29	Any	130**	Asian	Heart disease
3	Cary	20-29	Any	130**	Asian	Viral infection
4	Dick	20-29	Any	130**	Asian	Viral infection
5	Eshwar	40-59	Any	130**	Asian	Cancer
6	Fox	40-59	Any	14***	Asian	Flu
7	Gary	40-59	Any	14***	Asian	Heart disease
8	Helen	40-59	Any	14***	Asian	Flu
9	Igor	30-39	Any	1322*	American	Cancer
10	Jean	30-39	Any	1322*	American	Cancer
11	Ken	30-39	Any	1322*	American	Cancer
12	Lewis	30-39	Any	1322*	American	Cancer

Probabilistic Query Processing

- Represent the data set in terms of a probabilistic database.

Age	Gender	Zip Code	Nationality	BID
28	F	13053	Korean	**1**
29	M	13068	Chinese	**1**
21	F	13068	Japanese	**1**
23	M	13053	American	**1**
50	M	13053	Indian	**2**
55	M	14750	Japanese	**2**
47	M	14562	Chinese	**2**

BID	Condition
1	Heart disease
1	Heart disease
1	Viral infection
1	Viral infection
2	Cancer
2	Flu
2	Heart disease

Name	Age	Gender	Zip Code	Nationality	Condition	Probability
Ann	28	F	13053	Korean	Heart disease	**0.5**
Ann	28	F	13053	Korean	Viral infection	**0.5**
Bruce	29	M	13068	Chinese	Heart disease	**0.5**
Bruce	29	M	13068	Chinese	Heart disease	**0.5**
Cary	21	F	13068	Japanese	Viral infection	**0.5**
Cary	21	F	13068	Japanese	Viral infection	**0.5**
Dick	23	M	13053	American	Viral infection	**0.5**
Dick	23	M	13053	American	Viral infection	**0.5**

Probabilistic Query Processing

- Represent the data set in terms of a probabilistic database.
  - This table can be interpreted to mean that each row appears with probability 0.5.
  - Does not capture the fact that Ann has either heart disease or viral infection, but not both.

Name	Age	Gender	Zip Code	Nationality	Condition	Probability
Ann	28	F	13053	Korean	Heart disease	**0.5**
Ann	28	F	13053	Korean	Viral infection	**0.5**
Bruce	29	M	13068	Chinese	Heart disease	**0.5**
Bruce	29	M	13068	Chinese	Heart disease	**0.5**
Cary	21	F	13068	Japanese	Viral infection	**0.5**
Cary	21	F	13068	Japanese	Viral infection	**0.5**
Dick	23	M	13053	American	Viral infection	**0.5**
Dick	23	M	13053	American	Viral infection	**0.5**

OLAP

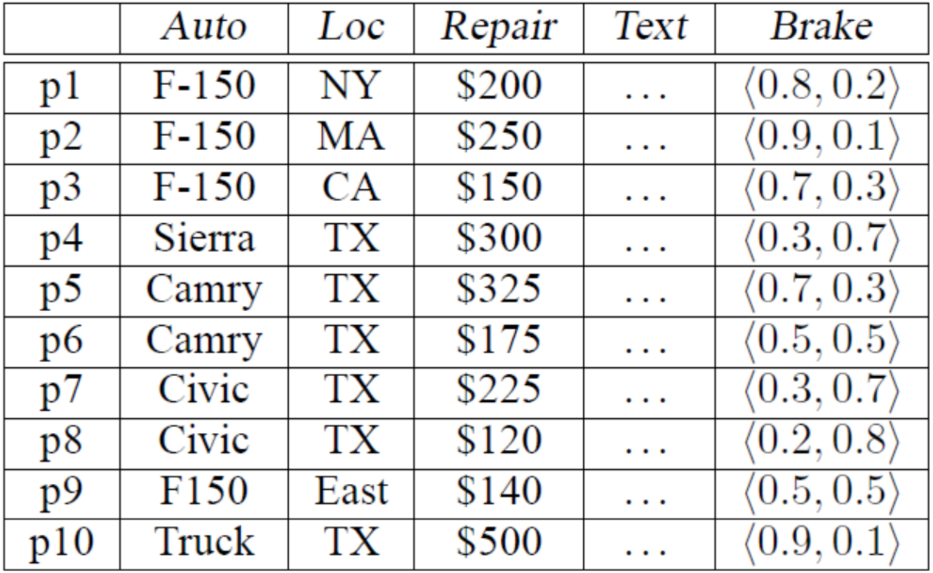
- OLAP (online analytic processing) : answer aggregate queries and perform OLAP analysis over ambiguous data.
  - **Imprecision domains**: an imprecise value is a non-empty set of possible values (a node in a generalization hierarchy) . Different records may be generalized to different granularities. If all attribute are leaf nodes, the observation is precise; otherwise, the observation is imprecise and describes a larger k-dimensional region.
  - **Uncertain domains**: the value of an attribute is a pdf.
  - Aggregate queries (e.g., SUM, COUNT, AVERAGE, etc.) over groups of any granularity.

OLAP

» Criteria that must be satisfied by any approach to handling data ambiguity in an OLAP setting:

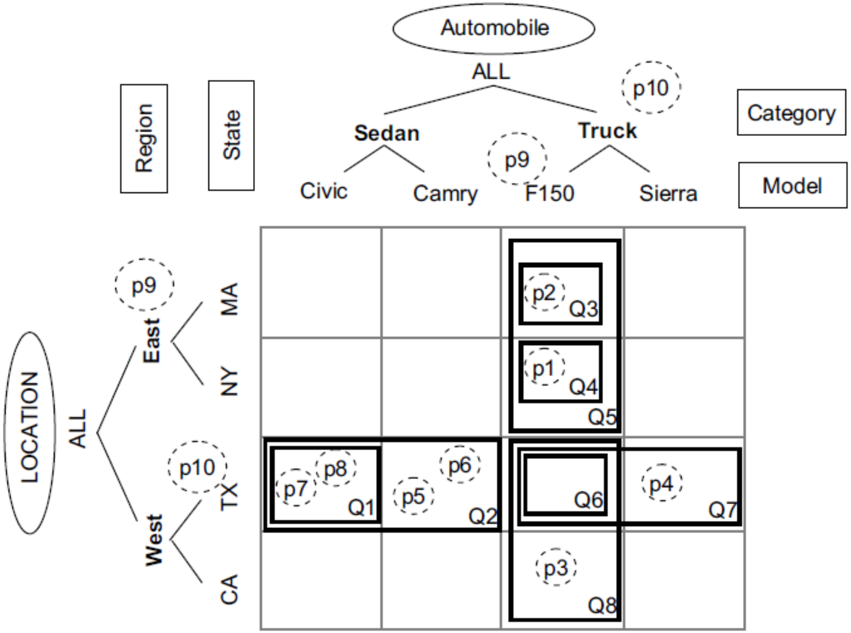
- **Consistency**: the relationship between similar queries issued at related nodes in a domain hierarchy in order to meet users' intuitive expectations as they navigate up and down the hierarchy.
- **Faithfulness**: more precise data should lead to better results.
- **Correlation-preservation:** the statistical properties of the data should not be affected by the allocation of ambiguous data records.

OLAP



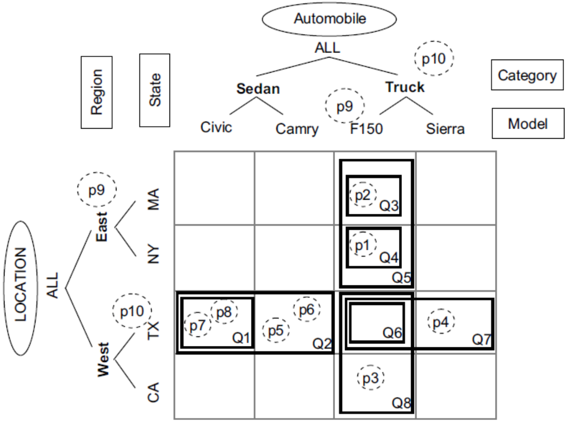
	Auto	*Loc*	*Repair*	*Text*	*Brake*
p1	F-150	NY	\$200	$\ldots$	$\langle 0.8, 0.2 \rangle$
p2	F-150	MA	\$250	$\ldots$	$\langle 0.9, 0.1 \rangle$
p3	F-150	CA	\$150	$\ldots$	$\langle 0.7, 0.3 \rangle$
p4	Sierra	TX	\$300	$\ldots$	$\langle 0.3, 0.7 \rangle$
p5	Camry	TX	\$325	$\ldots$	$\langle 0.7, 0.3 \rangle$
p6	Camry	TX	\$175	$\ldots$	$\langle 0.5, 0.5 \rangle$
p7	Civic	TX	\$225	$\ldots$	$\langle 0.3, 0.7 \rangle$
p8	Civic	TX	\$120	$\ldots$	$\langle 0.2, 0.8 \rangle$
p9	F150	East	\$140	$\ldots$	$\langle 0.5, 0.5 \rangle$
p10	Truck	TX	\$500	$\ldots$	$\langle 0.9, 0.1 \rangle$

OLAP


Automobile
ALL
Sedan
Truck
Civic
Camry
F150
Sierra
p9
p10
Category
Model
Region
State
LOCATION
ALL
East
West
MA
NY
TX
CA
p9
p10
p1
p2
p3
p4
p5
p6
p7
p8
Q1
Q2
Q3
Q4
Q5
Q6
Q7
Q8

OLAP

- Find-relevant (a1, ..., ak)
  - Precise fact within the region
  - **None**: ignore all imprecise facts
  - **Contain**: imprecise facts contained in the query region
  - **Overlap**: imprecise facts whose region overlaps the query region



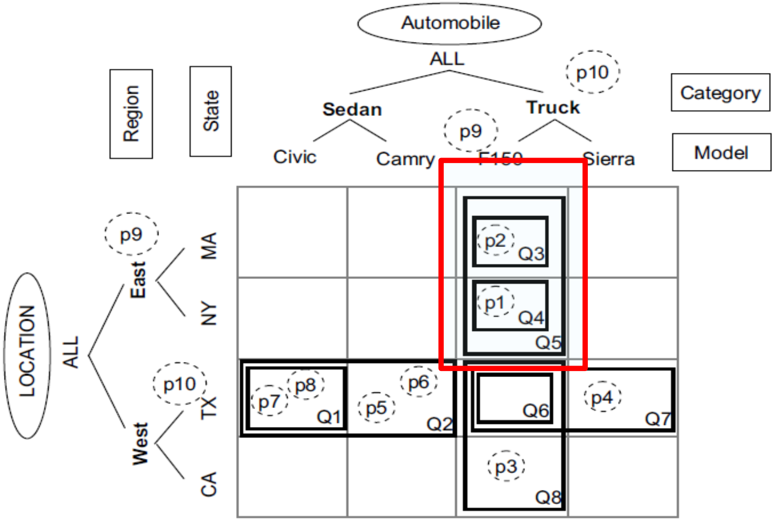
OLAP

- *Q3: What are the repair costs for F150's in MA?*
- *Q4: What are the repair costs for F150's in NY?*
- ***Q5: What are the repair costs for F150's in the East"?***

Q3=A(p2)

Q4=A(p1)

Q5=Q3+Q4?



OLAP

- *Q3: What are the repair costs for F150's in MA?*
- *Q4: What are the repair costs for F150's in NY?*
- ***Q5: What are the repair costs for F150's in the East"?***

For Q5=Q3+Q4,

We partially assign p9

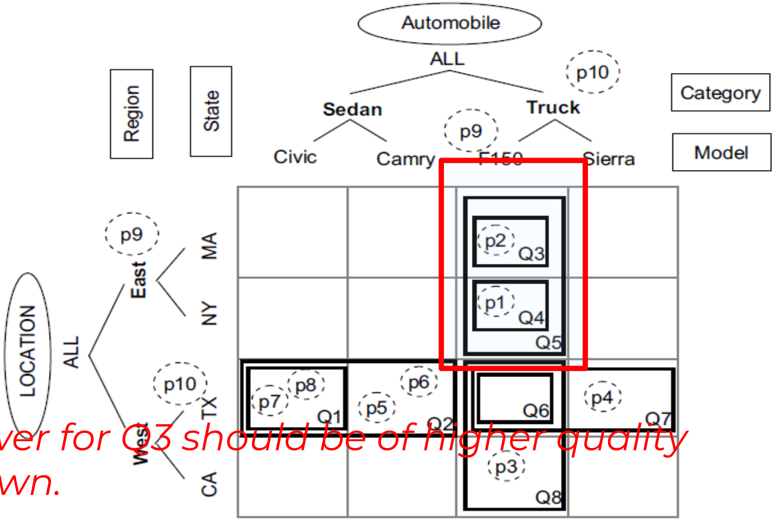
to both cells.

*Q3=A(p2+w1*p9)*

*Q4=A(p1+w2*p9)*

Q5=A(p1,p2,p9)

Consistency !



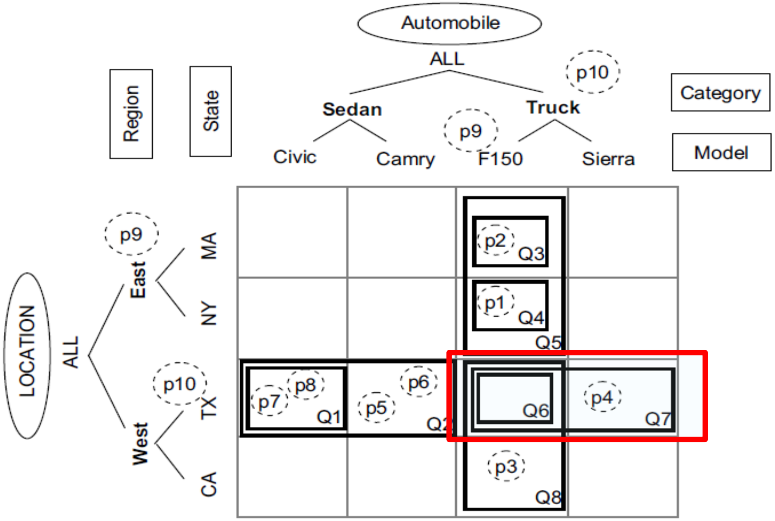
***Faithfulness** : the answer for Q3 should be of higher quality if p9 were precisely known.*

OLAP

Q6: What are the repair costs for F150's in TX"?

- If p10 is allocated to all cells in its region then Q6 can be answered. Otherwise, the answer to Q6 is undefined.

	Auto	*Loc*	*Repa*
p1	F-150	NY	$20
p2	F-150	MA	$25
p3	F-150	CA	$15
p4	Sierra	TX	$30
p5	Camry	TX	$32
p6	Camry	TX	$17
p7	Civic	TX	$22
p8	Civic	TX	$12
p9	F150	East	$14
p10	Truck	TX	$50

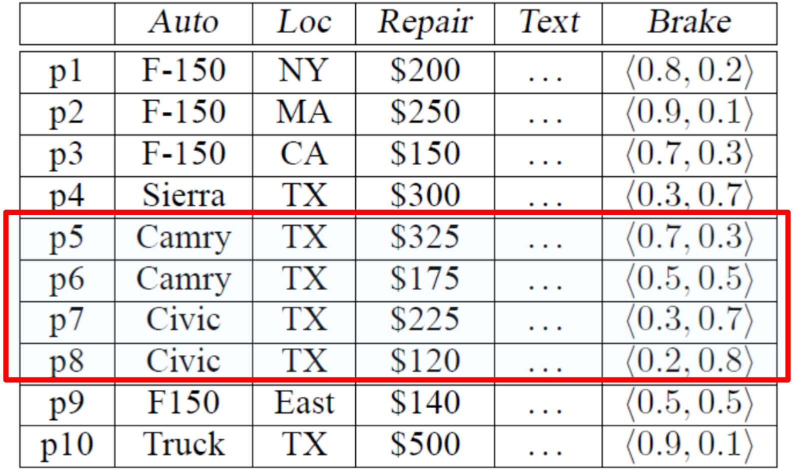


OLAP

- **_Q2: How likely are brake problems for sedans in TX?_**

Opinion pooling: provide a consensus opinion from a set of opinions.

Linear operation: produces a consensus pdf that is a weighted linear combination of the pdfs.



	Auto	*Loc*	*Repair*	*Text*	*Brake*
p1	F-150	NY	\$200	...	$\langle 0.8, 0.2 \rangle$
p2	F-150	MA	\$250	...	$\langle 0.9, 0.1 \rangle$
p3	F-150	CA	\$150	...	$\langle 0.7, 0.3 \rangle$
p4	Sierra	TX	\$300	...	$\langle 0.3, 0.7 \rangle$
p5	Camry	TX	\$325	...	$\langle 0.7, 0.3 \rangle$
p6	Camry	TX	\$175	...	$\langle 0.5, 0.5 \rangle$
p7	Civic	TX	\$225	...	$\langle 0.3, 0.7 \rangle$
p8	Civic	TX	\$120	...	$\langle 0.2, 0.8 \rangle$
p9	F150	East	\$140	...	$\langle 0.5, 0.5 \rangle$
p10	Truck	TX	\$500	...	$\langle 0.9, 0.1 \rangle$

Consistency

» User expects to see some natural relationships hold between the answers to aggregation queries associated with different (connected) regions in a hierarchy.

- **Sum-consistency**: SUM for a query region should equal the value obtained by adding the results of SUM for the query sub-regions that partition the region.

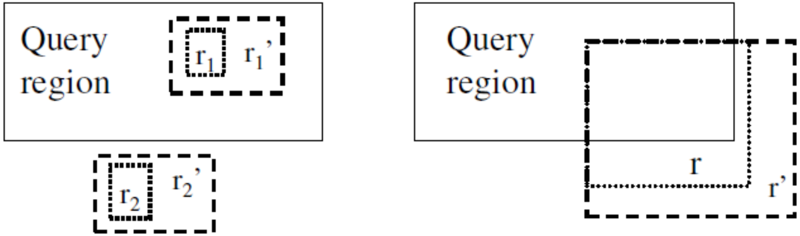
$$\hat{q} = \sum_i \hat{q}_i$$

- **Boundedness-consistency**: AVERAGE for a query region should be within the bounds of AVERAGE for the query sub-regions that partition the region. In the case of LinOp, the same property should hold element-wise for the associated pdfs.

$$\min_i\{\hat{q}_i\} \;\leq\; \hat{q} \;\leq\; \max_i\{\hat{q}_i\}$$

Faithfulness

- Increase imprecision in D by mapping facts in the database to larger regions. The answer to any query Q on this new database D’ will be different from the original answer.
- **Faithfulness** is intended to capture the intuitive property that this difference should be as small as possible.
  - Basic faithfulness: r1 and r1’ are completely contained in query region or completely disjoint from query region, i.e., identically precise.
  - Advanced: as the data in a query region becomes imprecise and grows outside the query region, SUM should be non-increasing.



a: No partial overlap

b: Partial order $\preccurlyeq_Q$

OLAP

» Criteria that must be satisfied by any approach to handling data ambiguity in an OLAP setting:

- **Consistency**: the relationship between similar queries issued at related nodes in a domain hierarchy in order to meet users’ intuitive expectations as they navigate up and down the hierarchy.
- **Faithfulness**: more precise data should lead to better results.
- **Correlation-preservation:** the statistical properties of the data should not be affected by the allocation of ambiguous data records.

- Only the Overlaps option for handling imprecision results in well-behaved queries in the context of OLAP.

Techniques specific to sanitization mechanisms

- A number of techniques have also been proposed to answer query based directly on the mechanisms used for sanitization.
  - Aggregate range queries v.s. to randomized response
  - Count queries v.s. subset true data with fake data inserted
  - Provide upper and lower bounds on answers to aggregate queries over bucketized data.

Aggregate Range Queries

» Multidimensional count aggregate from N clients

	Name	Age	Salary
1	Ann	28	25k
2	Bruce	29	20k
3	Cary	21	30k
4	Dick	23	28k
5	Eshwar	50	35k
6	Fox	55	42k
7	Gary	47	26k
8	Helen	49	33k
9	Igor	31	37k
10	Jean	37	45k
11	Ken	36	29k
12	Lewis	35	22k

Select count(*) from T
Where $R(A_1)$ and $R(A_2)$…

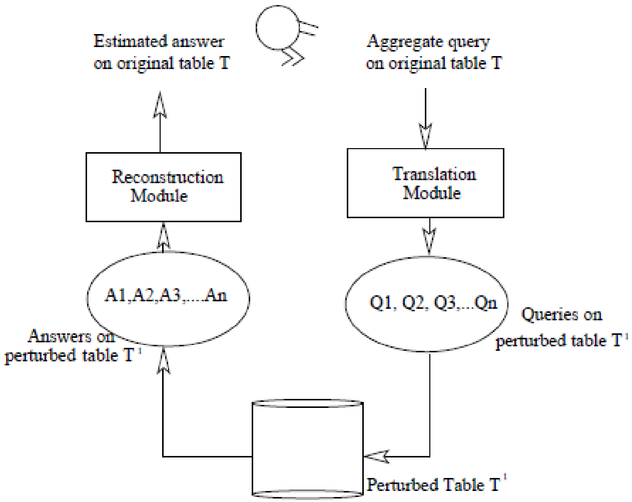
$R(A_1)$ is the range of A_1

Local randomization: each client perturb its row before sending it to the server. The perturbation algorithm is public.

Retention replacement scheme: an element is decided to be retained with probability **p** or replaced with an element selected from a **probability distribution function p.d.f**.

Aggregate Range Queries

- Multidimensional count aggregate from N clients



Estimated answer on original table T
Aggregate query on original table T
Reconstruction Module
Translation Module
A1,A2,A3,....An
Q1, Q2, Q3,...Qn
Answers on perturbed table T'
Queries on perturbed table T'
Perturbed Table T'

Approximate probabilistic reconstructability: the accuracy of the reconstruction algorithm.

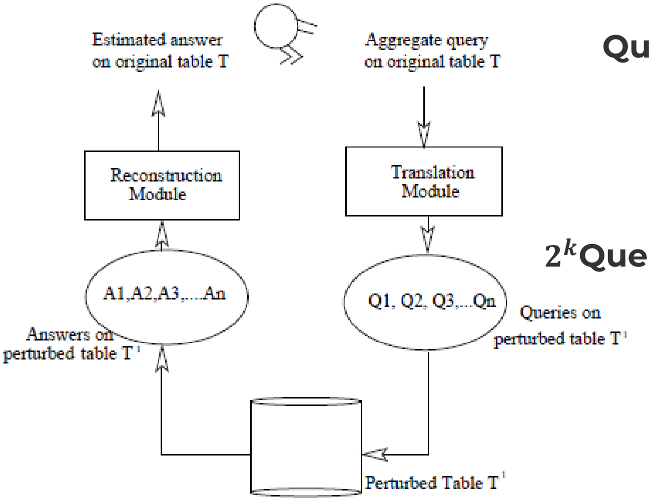
Reconstructability
f is a real-valued function computed over the original data, and f' is the estimator of f computed over the perturbed data, then f is (n, ε, δ) reconstructible if $|f - f'| < \max(\varepsilon, \varepsilon f)$ with probability at least $(1 - \delta)$ whenever the number of tuples in the original data is at least n.

Reconstructibility

- The approaches discussed so far measure the utility of the sanitized data that are actually produced.
- Measure utility in terms of the algorithm used to create the sanitized data, in this case, the result is usually a probabilistic utility guarantee.
- Eg., Agrawal et al. define the utility associated with a randomized anonymization algorithm in terms of the ability to reconstruct statistics from the sanitized data.

Aggregate Range Queries

- Multidimensional count aggregate from N clients


Estimated answer on original table T
Aggregate query on original table T
Reconstruction Module
Translation Module
A1,A2,A3,....An
Q1, Q2, Q3,...Qn
Answers on perturbed table T'
Queries on perturbed table T'
Perturbed Table T'

Query: $\text{count}\,(R(A_1) \cap R(A_2) \cap \cdots R(A_k))$

↓

2^kQueries: $\text{count}\,(R(A_1) \cap R(A_2) \cap \cdots R(A_k))$

$\text{count}\,(\neg R(A_1) \cap R(A_2) \cap \cdots R(A_k))$

....

$\text{count}\,(\neg R(A_1) \cap \neg R(A_2) \cap \cdots \neg R(A_k))$

↓

2^kQueries on T'

Aggregate Range Queries

» Reconstruct single column aggregates

- T’: uniform retention replacement perturbation with retention probability p
- n rows and a single column C, with domain [min, max]
- **Count (R(C))=count (C[low, high])**
- T’: $n_r = count(C[low, high]))$, estimated:

$$n_o = \frac{1}{p}(n_r - n(1-p)b), where\ b = \frac{high-low}{max-min}$$

1. The expected number of rows that get perturbed is n(1− p). For uniform perturbation, $n(1-p)b$ rows will be expected to lie within the [low, high] range.
2. n_r can be seen as the sum of those rows that were decided to be perturbed into [low, high] and those rows that were unperturbed in the original interval.
3. Subtracting the $n(1-p)b$ perturbed rows from n_r , we get an estimate for the number of unperturbed rows in [low, high] in T. This is scaled up by 1/p.

Aggregate Range Queries

» Reconstruct single column aggregates

- T': uniform retention replacement perturbation with retention probability p
- n rows and a single column C, with domain [min, max]
- **Count (R(C))=count (C[low, high])**
- T': $n_r = count(C[low, high])$), estimated:

$$\boldsymbol{n_o = \frac{1}{p}(n_r - n(1-p)b), where\ b = \frac{high-low}{max-min}}, \mathbf{a = 1 - b}$$

$$[n - \boldsymbol{n_o} \quad n]\begin{bmatrix}(1-p)a + p & (1-p)b \\ (1-p)a & (1-p)b + p\end{bmatrix} = [n - n_r \quad n_r]$$

count $(\neg R(C))$ count $(R(C))$ on T'

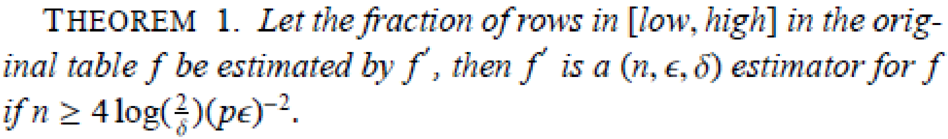
count $(\neg R(C))$ count $(R(C))$ on T

Aggregate Range Queries

- Reconstruct single column aggregates

$$f' = \frac{n_o}{n} = \frac{n_r}{pn} - \frac{(1-p)(high-low)}{p(max-min)}$$

1. It reconstructs an approximate answer with high probability



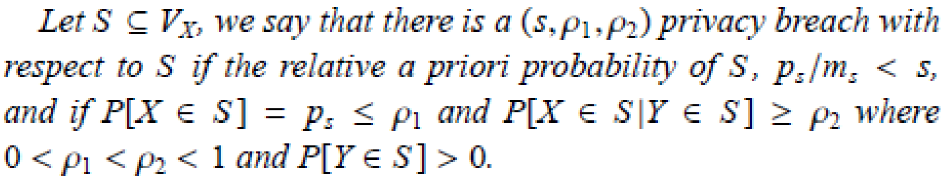
THEOREM 1. *Let the fraction of rows in* $[low, high]$ *in the original table* f *be estimated by* f', *then* f' *is a* (n, ϵ, δ) *estimator for* f *if* $n \geq 4\log(\frac{2}{\delta})(p\epsilon)^{-2}$.

How to reconstruct multiple column aggregates?

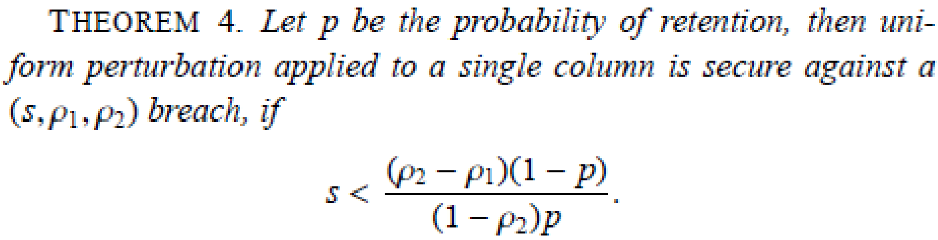
Aggregate Range Queries

- Reconstruct single column aggregates

2. The ratio $\frac{p_s}{m_s}$ (prior distribution P[X ∈ S] /replacing distribution P[Y ∈ S]) is called the relative a priori probability of the set S .



Let $S \subseteq V_X$, we say that there is a (s, ρ_1, ρ_2) privacy breach with respect to S if the relative a priori probability of S, $p_s/m_s < s$, and if $P[X \in S] = p_s \leq \rho_1$ and $P[X \in S|Y \in S] \geq \rho_2$ where $0 < \rho_1 < \rho_2 < 1$ and $P[Y \in S] > 0$.



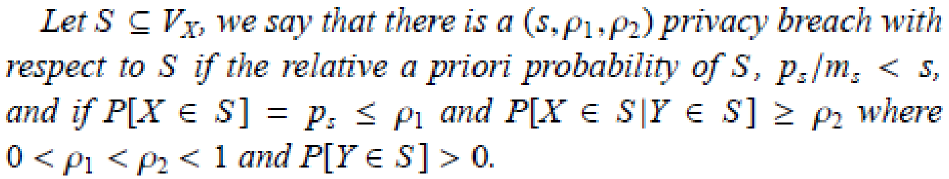
THEOREM 4. *Let p be the probability of retention, then uniform perturbation applied to a single column is secure against a (s, ρ_1, ρ_2) breach, if*

$$s < \frac{(\rho_2 - \rho_1)(1 - p)}{(1 - \rho_2)p}.$$

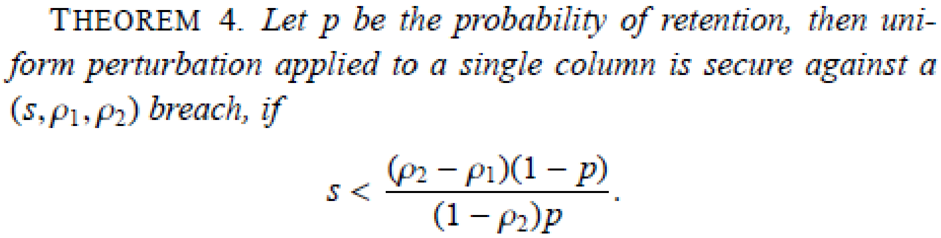
Aggregate Range Queries

- Reconstruct single column aggregates

2. The ratio $\frac{p_s}{m_s}$ (prior distribution P[X ∈ S] /replacing distribution P[Y ∈ S]) is called the relative a priori probability of the set S .



Let $S \subseteq V_X$, we say that there is a (s, ρ_1, ρ_2) privacy breach with respect to S if the relative a priori probability of S, $p_s/m_s < s$, and if $P[X \in S] = p_s \leq \rho_1$ and $P[X \in S|Y \in S] \geq \rho_2$ where $0 < \rho_1 < \rho_2 < 1$ and $P[Y \in S] > 0$.



THEOREM 4. *Let p be the probability of retention, then uniform perturbation applied to a single column is secure against a (s, ρ_1, ρ_2) breach, if*

$$s < \frac{(\rho_2 - \rho_1)(1-p)}{(1-\rho_2)p}.$$

Data Analysis over Sanitized Data

- Data analysis over sanitized data from the point of view of machine learning and data mining.

- The most direct measure of data utility is the **accuracy** (error rate or application-dependent metrics) of such models built on sanitized versions of data.
  - E.g., A/B test

Data Analysis over Sanitized Data

- Data analysis over sanitized data from the point of view of machine learning and data mining.

- The most direct measure of data utility is the **accuracy** (error rate or application-dependent metrics) of such models built on sanitized versions of data.

- Sanitization process often destroys some structure of the original data. Thus, learning or mining algorithms may need to be adapted in order to be applied to the sanitized data.

Evaluation Methodology

» To avoid obtaining overly optimistic accuracy estimates, **strict training/testing data separation** is recommended.

- Split the dataset into two parts: the **training set D1** and the **test set D2**
- Apply a sanitization method to the training set **without any access to the test** (any statistics computed from the entire original data D should not be used in data sanitization)
- Build a model using only the sanitized training set D1*
- Evaluate on the test set D2
- **n-fold cross-validation**

Evaluation Methodology

» N-fold cross-validation

- It generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.
- The general procedure is as follows:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:

- Take the group as a hold out or test data set
- Take the remaining n-1 groups as a training data set
- Fit a model on the training set and evaluate it on the test set
- Retain the evaluation score and discard the model

4. Summarize the skill of the model using evaluation scores

Learning from Sanitized Data

- Learning from a generalized table

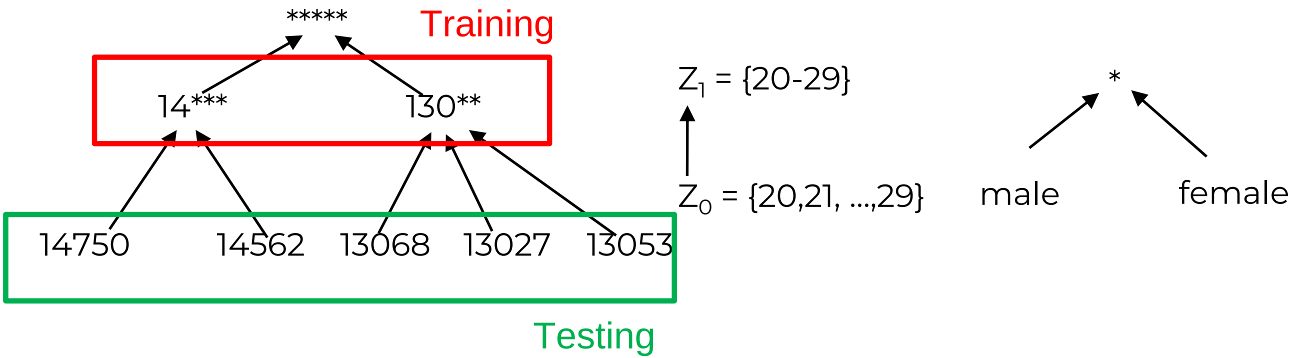
Name	Age	Gender	Zip Code	Nationality	Condition
Ann	20-29	Any	130**	Asian	Heart disease
Bruce	20-29	Any	130**	Asian	Heart disease
Cary	20-29	Any	130**	Asian	Viral infection
Dick	20-29	Any	130**	Asian	Viral infection
Eshwar	40-59	Any	130**	Asian	Cancer
Fox	40-59	Any	14***	Asian	Flu

↓ Predict

Name	Age	Gender	Zip Code	Nationality	Condition
Helen	49	F	14821	Korean	?
Igor	31	M	13222	American	?
Jean	37	F	13227	American	?
Ken	36	M	13228	American	?
Lewis	35	M	13221	American	?

Learning from Sanitized Data

- Learning from a generalized table



Learning from Sanitized Data

- Learning from a generalized table
  - One simple approach to handle this mismatch is to sample leaf-level values for records in D1* that have non-leaf values.
  - LeFevre et al. studied a **range encoding** method: replaces each generalized value with two separate features representing the upper and lower bounds of the range.
  - Decision tree learning, Naive Bayes learning, and rule learning for hierarchical attribute values.

Learning from Sanitized Data

- Learning from noisy data
  - Usually the noise distribution is assumed to be known.
  - Linear regression[1]
  - Decision tree induction [2]
  - Bayesian regression [252]
  - Support vector machine classifier [4]
  - Nearest neighbor classifier[5]

[1] G. A. F. Seber and A. J. Lee, *Linear Regression Analysis*. John Wiley and Son, 2003.

[2] R. Agrawal and R. Srikant, “Privacy-preserving data mining,” in Proceedings of the ACM SIGMOD International Conference on Management of Data, 2000.

[3] J.-A. Ting, A. D’Souza, and S. Schaal, “Bayesian regression with input noise for high dimensional data,” in Proceedings of the 23rd International Conference on Machine Learning,pp. 937–944, 2006.

[4] J. Bi and T. Zhang, “Support vector classification with input data uncertainty,” in Advances in Neural Information Processing Systems (NIPS), 2004.

[5] C. C. Aggarwal, J. Pei, and B. Zhang, “On privacy preservation against adversarial data mining,” in Proceedings of the 12th ACM SIGKDD, 2006.

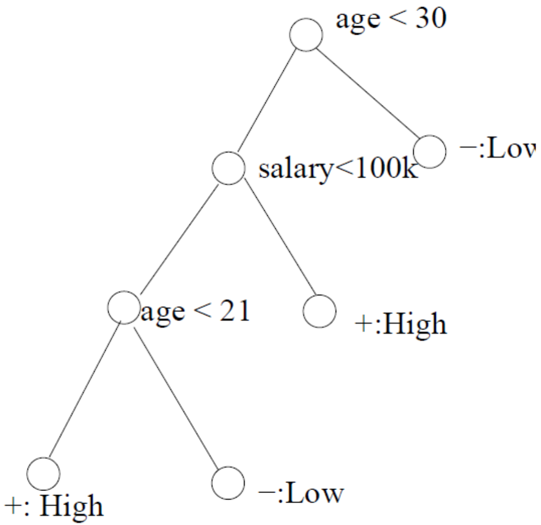
Learning from Sanitized Data

- Decision tree based on aggregating queries on multiple columns of randomized table T’.

Schema (age, salary, house-rent, class-variable) to predict the column class variable (high and low credit-risk).

The private columns among age, salary, house-rent and class-variable, are each independently perturbed by a retention replacement perturbation.

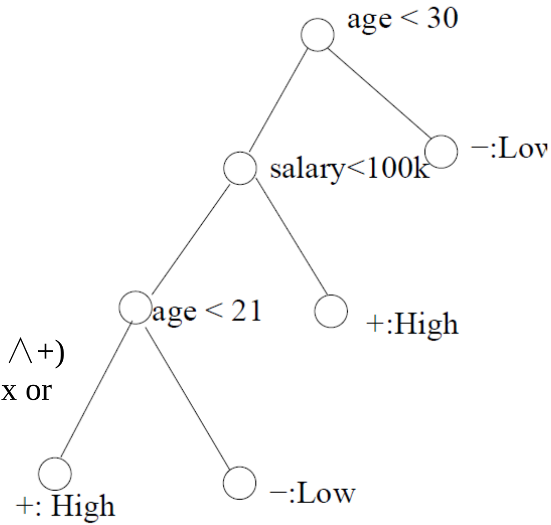
Q denotes the predicate (‘+’) while ¬Q denote the predicate (‘-’).



Learning from Sanitized Data

» Decision tree based on aggregating queries on multiple columns of randomized table T’.

1. on (age < 30), the Gini index is calculated using the estimated answers of the four queries: count(age[0-30] ∧-), count(¬age[0,30] ∧+), count(age[0-30]∧+) and count(¬age[0,30]∧-) on T.
2. With multi-column reconstruction the queries count(age[0-30] ∧ salary[25k-100k] ∧-), count(age[0,30] ∧salary[100k-200k] ∧-), count(age[0-30] ∧ salary[25k-100k] ∧ +) and count(age[0,30] ∧ salary[100k-200k] ∧+) are reconstructed for T, to calculate the Gini index or another split criterion at this level.



Learning from Sanitized Data

- Learning from group statistics
  - Sanitized data set is generated by first grouping or bucketizing records of the original data set and then releasing summary statistics for each group.
  - Build a model using only these statistics.
  - This setting is also related to the multiple instance learning problem.
- Learning from multiple views
  - the sanitized data set D1* consists of multiple aggregate views of the original data set D1, each of which contains a subset of the attributes of records in D1 and summary statistics (e.g., COUNT, AVERAGE, etc.)
  - For count views, which are commonly known as marginals (or marginal contingency tables), **iterative proportional fitting** is the classic method for estimating information about the original data table D1.

Statistical Analysis

- Data mining and machine learning literature provide methods to answer queries over sanitized data and build models from the data, these techniques usually **do not address whether a finding from the data is statistically significant or not.**
- For example, one may use a query processing technique to compute the sample mean; but one also needs to estimate the variance and/or confidence intervals.

Statistical Analysis

- General model for missing and coarsened data

Name	Age	Salary
Ann	28	25k
Bruce	29	20k
Cary	21	30k
Dick	23	28k
Eshwar	50	35k
Fox	55	42k
Gary	47	26k
Helen	49	33k
Igor	31	37k
Jean	37	45k
Ken	36	29k
Lewis	35	22k

- *Input: $n \times p$ data matri $\boldsymbol{X}$*
- *masking indicator matrix : M*
  - *The value of m_{ij} is 1 if the value of x_{ij} has been sanitized and 0 otherwise.*
- *masking treatment matrix : Z*
  - *The value of z_{ij} is the masked value of x_{ij} if m_{ij} = 1 and is unobserved otherwise.*
- *M and Z may not be known to the analyst and they would need to be treated as random variables.*

Statistical Analysis

- General model for missing and coarsened data

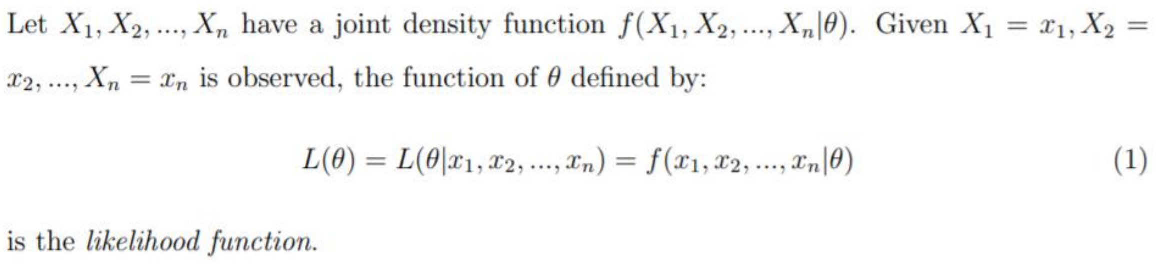
Name	Age	Salary
Ann	28	25k
Bruce	29	20k
Cary	21	30k
Dick	23	28k
Eshwar	50	35k
Fox	55	42k
Gary	47	26k
Helen	49	33k
Igor	31	37k
Jean	37	45k
Ken	36	29k
Lewis	35	22k

- *Partition X into two parts*
  - X_{obs} *is the subset of X that is observed in the output*
  - X_{mis} *is the subset of X for which the values are missing.*
- *Partition Z into two parts*
  - Z_{obs} *corresponds to values observed in the output*
  - Z_{mis} *corresponds to missing values Z*

Statistical Analysis

» Likelihood function

- A general functional relation between the unknown parameter(s) and the observed data.
- The goal of a statistical analysis is to estimate the unknown parameter(s) in the proposed model.
- It measures the support provided by the data for each possible value of the parameter.
- It allows us to make predictions.



Let $X_1, X_2, ..., X_n$ have a joint density function $f(X_1, X_2, ..., X_n|\theta)$. Given $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$ is observed, the function of θ defined by:

$$L(\theta) = L(\theta|x_1, x_2, ..., x_n) = f(x_1, x_2, ..., x_n|\theta) \tag{1}$$

is the *likelihood function*.

Statistical Analysis

- The primary statistical interest is in estimation and hypothesis testing for a **parameter θ** which corresponds to a hypothetical **model that generated the data**. Thus the data distribution is $f_{X(X|\theta)}$.
  - Z depends on the data matrix through the distribution $f_{z(Z|X)}$
  - M depends on the rest through the distribution $f_{M(M|Z_{obs},X)}$
  - Inference can be based on the likelihood function for **θ**

$$\mathcal{L}(\theta\,|\,\mathbf{M},\mathbf{X}_{obs},\mathbf{Z}_{obs}) = \int f_M(\mathbf{M}\,|\,\mathbf{Z}_{obs},\mathbf{X})\; f_Z(\mathbf{Z}_{obs}\,|\,\mathbf{X}) \times f_X(\mathbf{X}\,|\,\theta)\; d\mathbf{X}_{mis}$$

  - In case the masking indicator matrix M is unknown, one would also have to integrate out the uncertainty in Z_{obs} and M.

Statistical Analysis

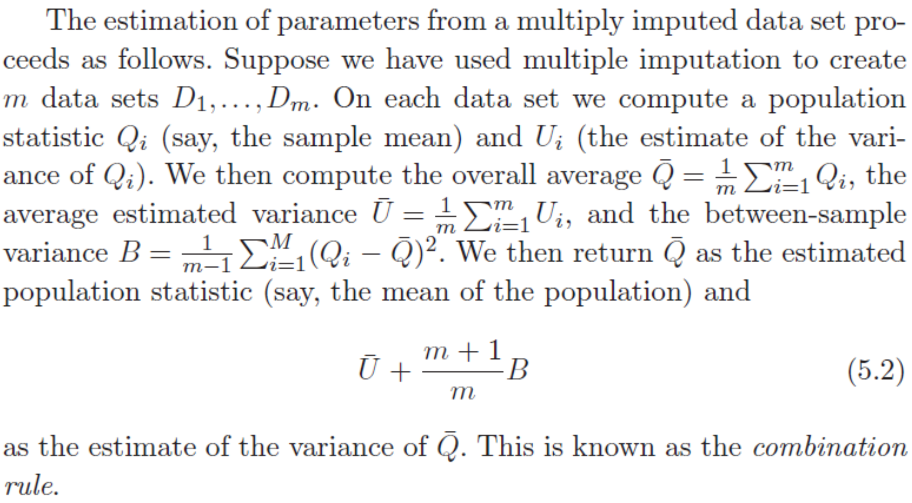
» Multiply Imputed Synthetic Data

- Multiple imputation (MI) is a tool for handling data sets that contain missing values.
- The basic idea is to create multiple complete data sets by filling in missing values by sampling them from a model built from the rest of the data.
- An alternative approach is to fill in each missing value with a single value (for example, by using a maximum likelihood method).
- Drawbacks: causing statistical software to underestimate the variance in the data because each filled-in value is treated in the same way as a value that is actually present in the data.

Statistical Analysis

» Multiply Imputed Synthetic Data

- Multiple imputation comes with a simple estimation procedure that helps avoid this problem.



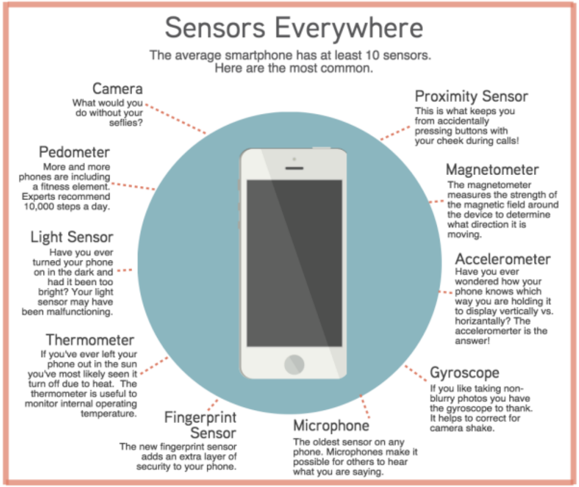
The estimation of parameters from a multiply imputed data set proceeds as follows. Suppose we have used multiple imputation to create m data sets $D_1, \ldots, D_m$. On each data set we compute a population statistic Q_i (say, the sample mean) and U_i (the estimate of the variance of Q_i). We then compute the overall average $\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} Q_i$, the average estimated variance $\bar{U} = \frac{1}{m}\sum_{i=1}^{m} U_i$, and the between-sample variance $B = \frac{1}{m-1}\sum_{i=1}^{M}(Q_i - \bar{Q})^2$. We then return $\bar{Q}$ as the estimated population statistic (say, the mean of the population) and

$$\bar{U} + \frac{m+1}{m}B \tag{5.2}$$

as the estimate of the variance of $\bar{Q}$. This is known as the *combination rule*.

Data Analysis over Sanitized Data

- Example:



Sensors Everywhere

The average smartphone has at least 10 sensors.
Here are the most common.

Camera
What would you do without your seflies?

Proximity Sensor
This is what keeps you from accidentally pressing buttons with your cheek during calls!

Pedometer
More and more phones are including a fitness element. Experts recommend 10,000 steps a day.

Magnetometer
The magnetometer measures the strength of the magnetic field around the device to determine what direction it is moving.

Light Sensor
Have you ever turned your phone on in the dark and had it been too bright? Your light sensor may have been malfunctioning.

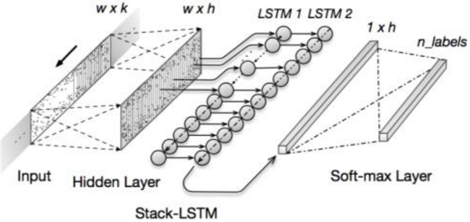
Accelerometer
Have you ever wondered how your phone knows which way you are holding it to display vertically vs. horizantally? The accelerometer is the answer!

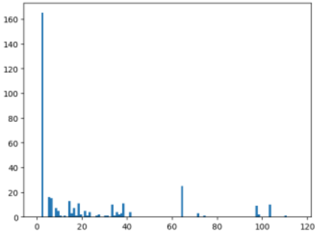
Thermometer
If you've ever left your phone out in the sun you've most likely seen it turn off due to heat. The thermometer is useful to monitor internal operating temperature.

Gyroscope
If you like taking non-blurry photos you have the gyroscope to thank. It helps to correct for camera shake.

Fingerprint Sensor
The new fingerprint sensor adds an extra layer of security to your phone.

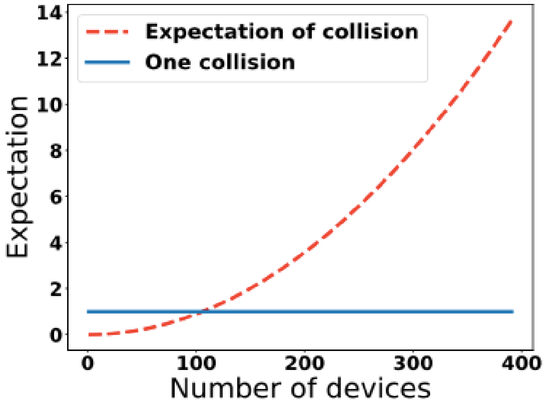
Microphone
The oldest sensor on any phone. Microphones make it possible for others to hear what you are saying.





Data Analysis over Sanitized Data

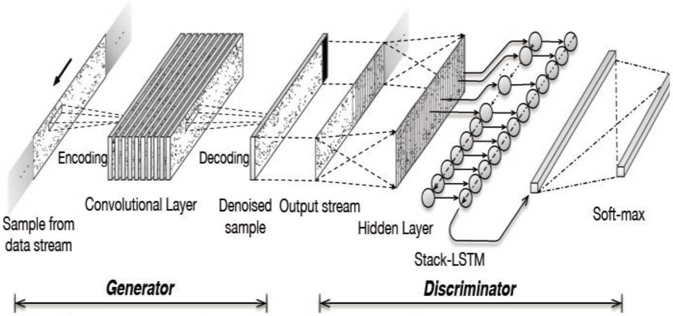
- Example:



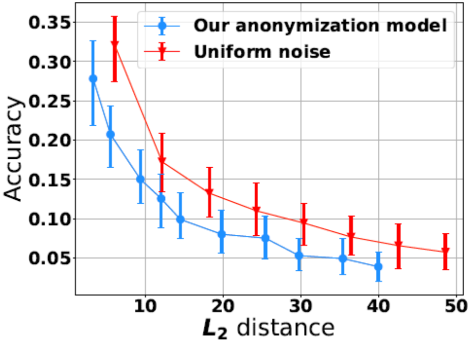
- 1秒钟的任意行为数据，即可达到93%的识别准确率。
- 10秒钟任意行为数据，可达到98.8%的识别准确率。

Data Analysis over Sanitized Data

- Example:



- 在不破坏数据语义的情况下，可将识别精度降至5%。



THANKS!

Any questions?

You can find me at:

- zhanglan@ustc.edu.cn