

# Privacy Preserving Data Publication

Yufei Tao / Tiancheng Li / Vitaly Smatikov / **Marianne Winslett**  
Chinese University of Hong Kong / Purdue University /  
University of Texas at Austin / **University of Illinois at Urbana-  
Champaign**

# How can we publish sensitive information without jeopardizing individuals' privacy?

No one should learn who had which disease.

Name	Age	Sex	Zipcode	Disease
Andy	5	M	12000	gastric ulcer
Bill	9	M	14000	dyspepsia
Ken	6	M	18000	pneumonia
Nash	8	M	19000	bronchitis
Joe	12	M	22000	pneumonia
Sam	19	M	24000	pneumonia
Linda	21	F	58000	flu
Jane	26	F	36000	gastritis
Sarah	28	F	37000	pneumonia
Mary	56	F	33000	flu

← **“Microdata”**

# What if we remove their names?

Name	Age	Sex	Zipcode	Disease
Andy	5	M	12000	gastric ulcer
Bill	9	M	14000	dyspepsia
Ken	6	M	18000	pneumonia
Nash	8	M	19000	bronchitis
Joe	12	M	22000	pneumonia
Sam	19	M	24000	pneumonia
Linda	21	F	58000	flu
Jane	26	F	36000	gastritis
Sarah	28	F	37000	pneumonia
Mary	56	F	33000	flu

publish  


Age	Sex	Zipcode	Disease
5	M	12000	gastric ulcer
9	M	14000	dyspepsia
6	M	18000	pneumonia
8	M	19000	bronchitis
12	M	22000	pneumonia
19	M	24000	pneumonia
21	F	58000	flu
26	F	36000	gastritis
28	F	37000	pneumonia
56	F	33000	flu

# We can reverse engineer the names by linking with external data.

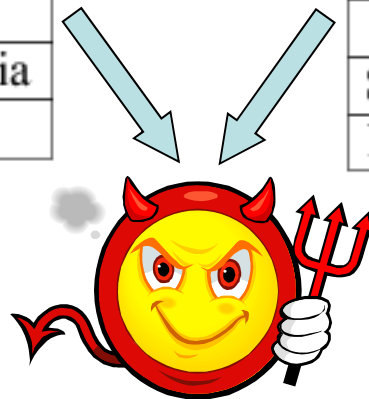
The published table

Age	Sex	Zipcode	Disease
5	M	12000	gastric ulcer
9	M	14000	dyspepsia
6	M	18000	pneumonia
8	M	19000	bronchitis
12	M	22000	pneumonia
19	M	24000	pneumonia
21	F	58000	flu
26	F	36000	gastritis
28	F	37000	pneumonia
56	F	33000	flu

Quasi-identifier (QI) attributes

A voter registration list

Name	Age	Sex	Zipcode
Andy	5	M	12000
Bill	9	M	14000
Ken	6	M	18000
Nash	8	M	19000
Mike	7	M	17000
Joe	12	M	22000
Sam	19	M	24000
Linda	21	F	58000
Jane	26	F	36000
Sarah	28	F	37000
Mary	56	F	33000



87% of Americans can be uniquely identified by {zip code, gender, date of birth}.

Latanya Sweeney [*International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002] used this approach to reverse engineer the medical record of an ex-governor of Massachusetts.



Real query logs, such as AOL's released log, can be very useful to CS researchers. But click history can uniquely identify a person.

*<AnonID, Query, QueryTime, ItemRank, URL clicked>*

What the New York Times did:

- Find all log entries for AOL user 4417749
- Multiple queries for businesses and services in Lilburn, GA (population 11K)
- Several queries for Jarrett Arnold
  - ✓ Lilburn has 14 people with the last name Arnold
- NYT contacts them, finds out AOL User 4417749 is Thelma Arnold



# What are the goals of privacy-preserving data publishing?

Publish a **distorted** version of the data set so that

Privacy: the privacy of all individuals is “adequately” protected;

Utility: the data set is useful for analyzing the characteristics of the microdata.

Paradox: Privacy protection ↑, utility ↓.

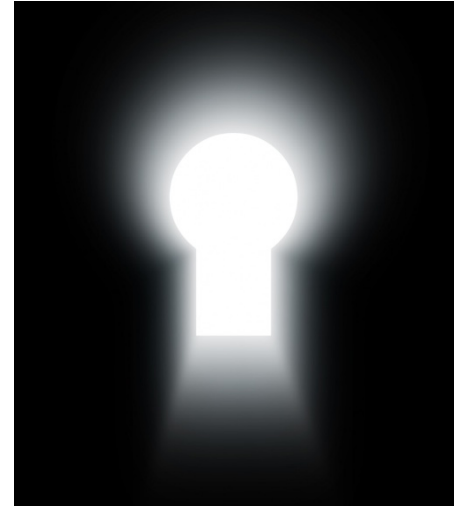
# Issues

## ➔ Privacy principle

What is adequate privacy protection?

## Distortion approach

How can we achieve the privacy principle, while maximizing the utility of the data?



# Different applications have different disclosure issues.

- ❑ **Membership disclosure:** Attacker cannot tell that a given person is in the data set (e.g., a set of AIDS patient records).
  - $\delta$ -presence [Nergiz et al., 2007].
- ❑ **Sensitive attribute disclosure:** Attacker cannot tell that a given person has a certain sensitive attribute
  - 1-diversity [Machanavajjhala et al., 2006].
  - t-closeness [Li et al., 2007].
- ❑ **Identity disclosure:** Attacker cannot tell which record corresponds to a given person
  - k-anonymity [Sweeney, 2002].

# Privacy principle 1: *k*-anonymity

[Sweeney, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002]

2-anonymous generalization:

Sensitive attribute

QI attributes

Name	Age	Sex	Zipcode
Andy	5	M	12000
Bill	9	M	14000
Ken	6	M	18000
Nash	8	M	19000
Mike	7	M	17000
Joe	12	M	22000
Sam	19	M	24000
Linda	21	F	58000
Jane	26	F	36000
Sarah	28	F	37000
Mary	56	F	33000

4 QI groups

Age	Sex	Zipcode	Disease
[1, 10]	M	[10001, 15000]	gastric ulcer
[1, 10]	M	[10001, 15000]	dyspepsia
[1, 10]	M	[15001, 20000]	pneumonia
[1, 10]	M	[15001, 20000]	bronchitis
[11, 20]	M	[20001, 25000]	pneumonia
[11, 20]	M	[20001, 25000]	pneumonia
[21, 60]	F	[30000, 60000]	flu
[21, 60]	F	[30000, 60000]	gastritis
[21, 60]	F	[30000, 60000]	pneumonia
[21, 60]	F	[30000, 60000]	flu

The **big** advantage of k-anonymity is that people can understand it.



*k*-anonymity does not provide privacy if the sensitive values in an equivalence class lack diversity and/or the attacker has background knowledge.

### Homogeneity Attack

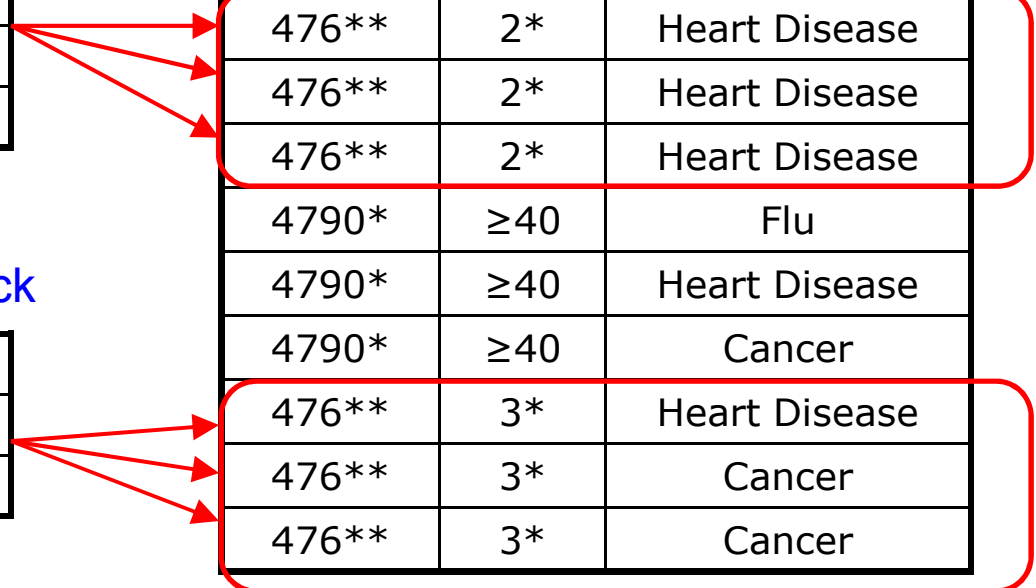
Bob	
<b>Zipcode</b>	<b>Age</b>
47678	27

### A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

### Background Knowledge Attack

Carl	
<b>Zipcode</b>	<b>Age</b>
47673	36



What is Joe's disease?

A voter registration list

Name	Age	Sex	Zipcode
Andy	5	M	12000
Bill	9	M	14000
Ken	6	M	18000
Nash	8	M	19000
Mike	7	M	17000
Joe	12	M	22000
Sam	19	M	24000
Linda	21	F	58000
Jane	26	F	36000
Sarah	28	F	37000
Mary	56	F	33000

No "diversity" in this QI group.

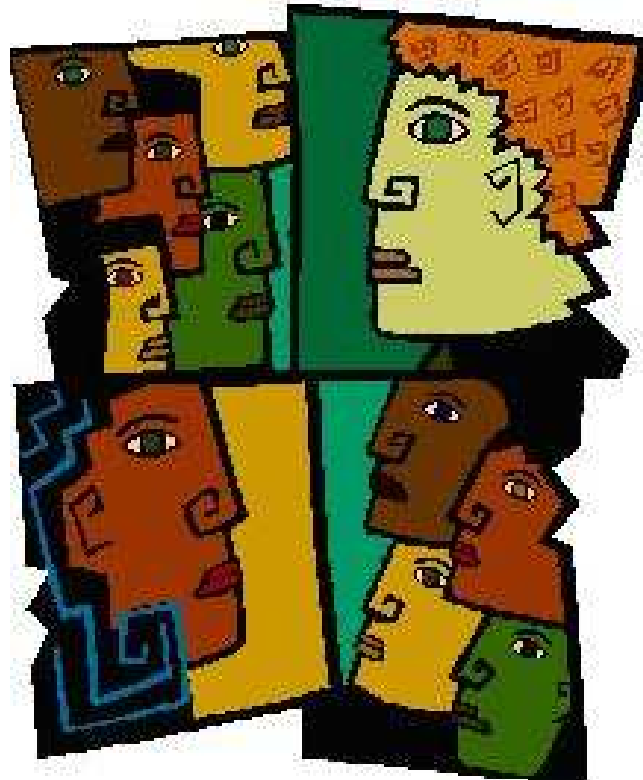
Age	Sex	Zipcode	Disease
[1, 10]	M	[10001, 15000]	gastric ulcer
[1, 10]	M	[10001, 15000]	dyspepsia
[1, 10]	M	[15001, 20000]	pneumonia
[1, 10]	M	[15001, 20000]	bronchitis
[11, 20]	M	[20001, 25000]	pneumonia
[11, 20]	M	[20001, 25000]	pneumonia
[21, 60]	F	[30000, 60000]	flu
[21, 60]	F	[30000, 60000]	gastritis
[21, 60]	F	[30000, 60000]	pneumonia
[21, 60]	F	[30000, 60000]	flu

Updates can also destroy k-anonymity.

# Principle 2: /-diversity

[Machanavajjhala et al., *ICDE*, 2006]

Each QI group should have at least / **“well-represented”** sensitive values.



*What does  
that mean?*

Maybe each QI-group has /  
different sensitive values?

A 2-diverse table

Age	Sex	Zipcode	Disease
[1, 5]	M	[10001, 15000]	gastric ulcer
[1, 5]	M	[10001, 15000]	dyspepsia
[6, 10]	M	[15001, 20000]	pneumonia
[6, 10]	M	[15001, 20000]	bronchitis
[11, 20]	F	[20001, 25000]	flu
[11, 20]	F	[20001, 25000]	pneumonia
[21, 60]	F	[30001, 60000]	gastritis
[21, 60]	F	[30001, 60000]	gastritis
[21, 60]	F	[30001, 60000]	flu
[21, 60]	F	[30001, 60000]	flu

# We can attack this probabilistically.

If we know Joe's QI group, what is the probability he has HIV?

A QI group with 100 tuples

...	Disease
	...
	HIV
	HIV
	...
	HIV
	pneumonia
	bronchitis
	...

98 tuples

**Implication: The most frequent sensitive value in a QI group cannot be too frequent.**

# Even then, we can still attack using background knowledge.

Joe has HIV.

Sally knows: Joe does not have pneumonia.

Sally can guess that Joe has HIV.

A QI group with 100 tuples

...	Disease
	...
	HIV
	...
	HIV
	pneumonia
	...
	pneumonia
	bronchitis
	...

50 tuples

49 tuples

# $l$ -diversity variants have been proposed to address these weaknesses.

## □ Probabilistic $l$ -diversity

- The frequency of the most frequent value in an equivalence class is bounded by  $1/l$ .

## □ Entropy $l$ -diversity

- The entropy of the distribution of sensitive values in each equivalence class is at least  $\log(l)$

## ➔ □ Recursive $(c, l)$ -diversity

- The most frequent value does not appear too frequently
- $r_1 < c(r_l + r_{l+1} + \dots + r_m)$ , where  $r_i$  is the frequency of the  $i$ -th most frequent value.

To address this weakness, we can control the 2nd most frequent value...

A QI group with 100 tuples

...	Disease
	...
	HIV
	...
	HIV
	pneumonia
	...
	pneumonia
	bronchitis
	...
	bronchitis
	...

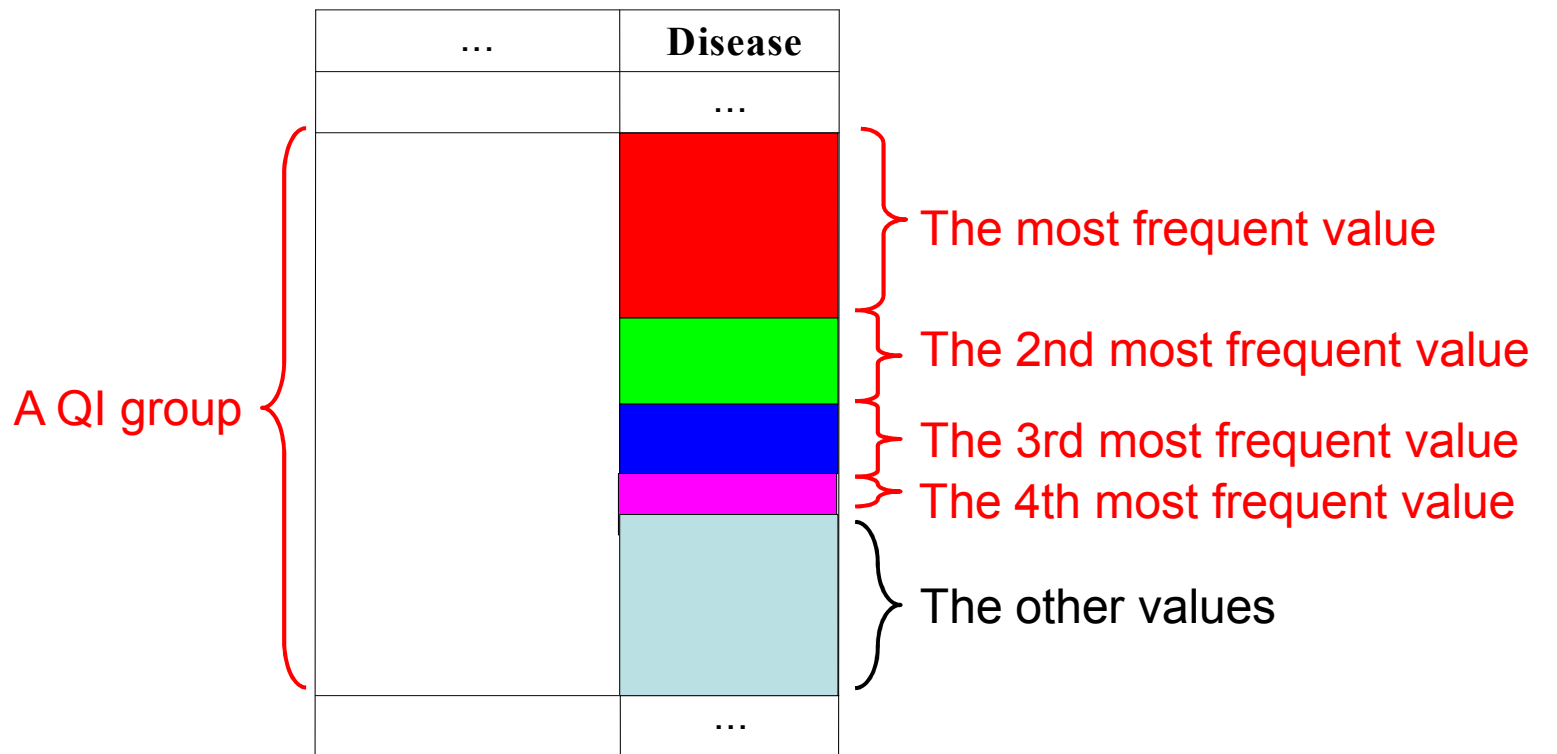
40 tuples

30 tuples

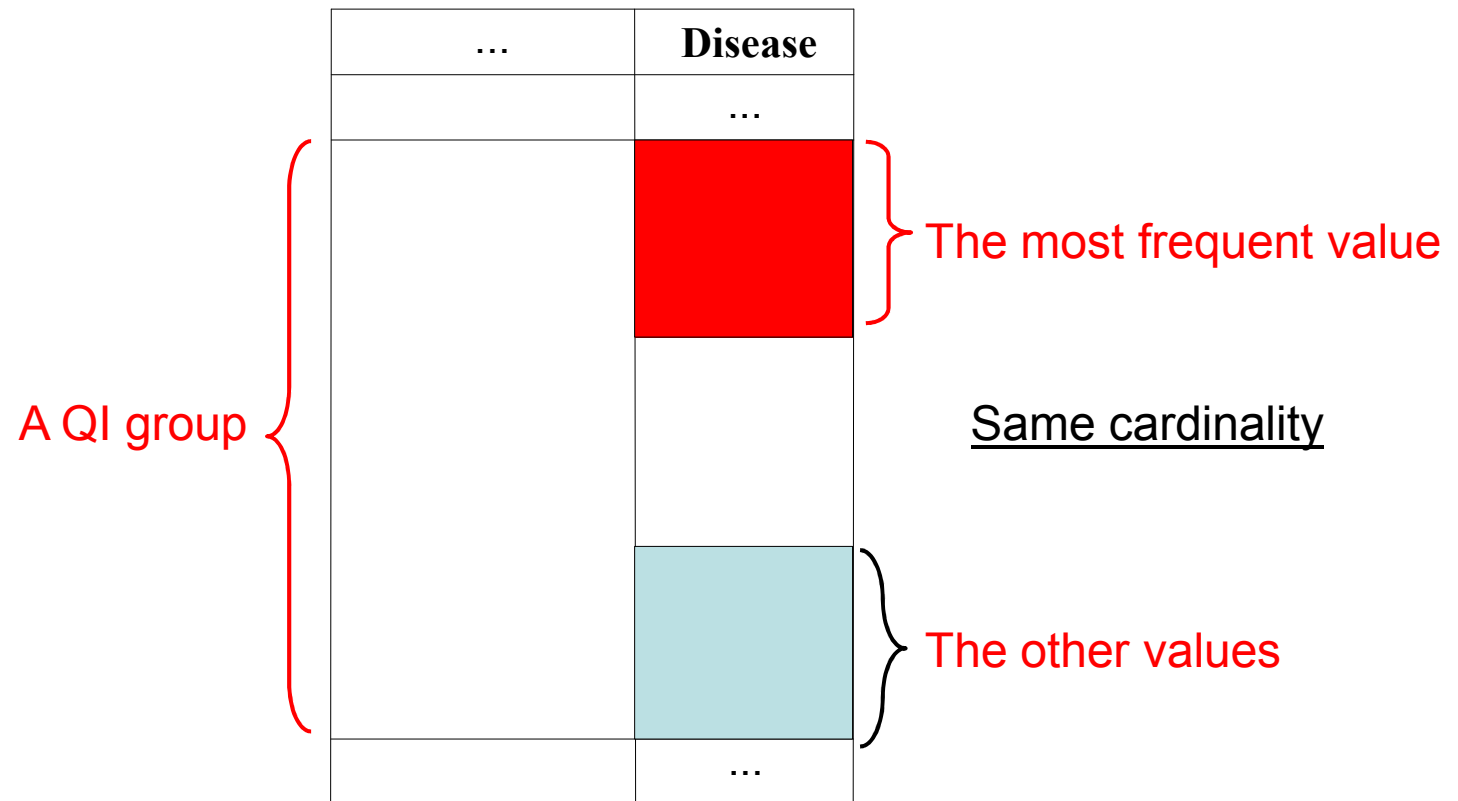
30 tuples

Now Sally knows that Joe has HIV with 40 / 70 probability.

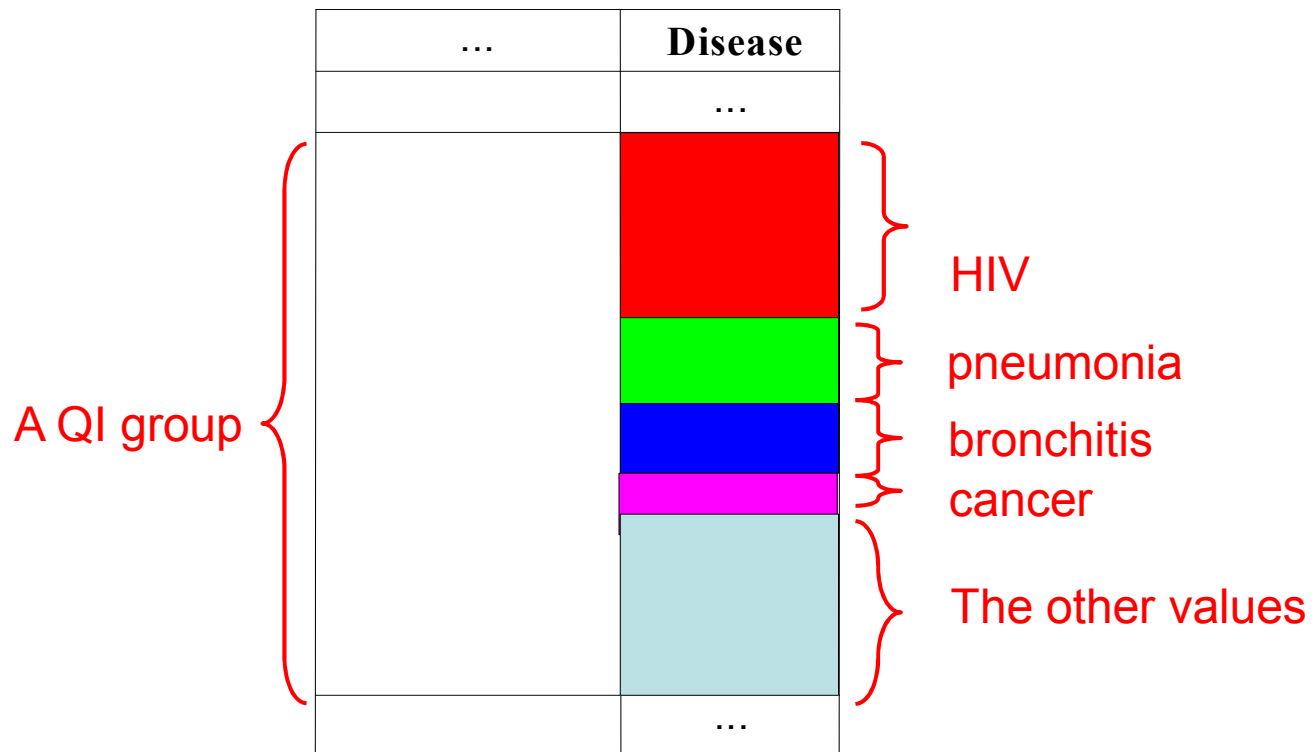
...



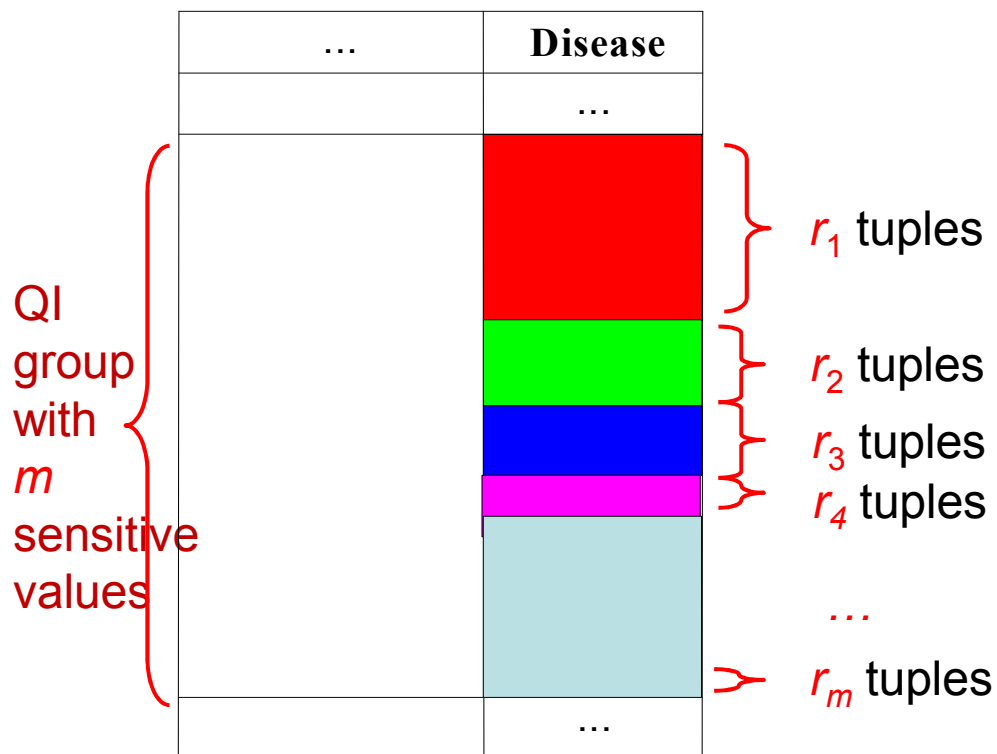
...



If Sally eliminates  $\leq 3$  diseases, she can guess Joe's disease with at most 50% probability.



$(c, l)$ -diversity: If Sally can eliminate only  $l - 1$  sensitive values, she can infer Joe's disease with probability at most  $1 / (c + 1)$ .



$$r_1 \leq c (r_1 + \dots + r_m)$$

Even (c,l)-diversity is not perfect. It does not consider personal preferences with respect to anonymization...

Andy does not want anyone to know that he had a stomach problem.  
Sarah does not mind at all if others find out that she had flu.

A 2-diverse table

Age	Sex	Zipcode	Disease
[1, 5]	M	[10001, 15000]	gastric ulcer
[1, 5]	M	[10001, 15000]	dyspepsia
[6, 10]	M	[15001, 20000]	pneumonia
[6, 10]	M	[15001, 20000]	bronchitis
[11, 20]	F	[20001, 25000]	flu
[11, 20]	F	[20001, 25000]	pneumonia
[21, 60]	F	[30001, 60000]	gastritis
[21, 60]	F	[30001, 60000]	gastritis
[21, 60]	F	[30001, 60000]	flu
[21, 60]	F	[30001, 60000]	flu

A voter registration list

Name	Age	Sex	Zipcode
Andy	4	M	12000
Bill	5	M	14000
Ken	6	M	18000
Nash	9	M	19000
Mike	7	M	17000
Alice	12	F	22000
Betty	19	F	24000
Linda	21	F	33000
Jane	25	F	34000
Sarah	28	F	37000
Mary	56	F	58000

# l-diversity can be overkill or underkill.

Original dataset

...	Cancer
...	Cancer
...	Cancer
...	Flu
..	Cancer
...	Cancer
...	Cancer
...	Cancer
...	Cancer
...	Cancer
...	Cancer
...	Flu
...	Flu

99% have cancer

Anonymization A

Q1	Flu
Q1	Flu
Q1	Cancer
Q1	Flu
Q1	Cancer
Q1	Cancer
Q2	Cancer
Q2	Cancer

Anonymization B

Q1	Flu
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q2	Cancer

99% cancer  $\Rightarrow$  quasi-identifier group is not "diverse"  
...yet anonymized database does not leak anything

50% cancer  $\Rightarrow$  quasi-identifier group is "diverse"  
**This leaks a ton of information**



/-diversity does not work if an individual can have multiple tuples in the microdata.

### Microdata

Name	Age	Sex	Zipcode	Disease
Andy	4	M	12000	gastric ulcer
Andy	4	M	12000	dyspepsia
Ken	6	M	18000	pneumonia
Nash	9	M	19000	bronchitis
Alice	12	F	22000	flu
Betty	19	F	24000	pneumonia
Linda	21	F	33000	gastritis
Jane	25	F	34000	gastritis
Sarah	28	F	37000	flu
Mary	56	F	58000	flu

A 2-diverse table

Age	Sex	Zipcode	Disease
4	M	12000	gastric ulcer
4	M	12000	dyspepsia
[6, 10]	M	[15001, 20000]	pneumonia
[6, 10]	M	[15001, 20000]	bronchitis
[11, 20]	F	[20001, 25000]	flu
[11, 20]	F	[20001, 25000]	pneumonia
[21, 60]	F	[30001, 60000]	gastritis
[21, 60]	F	[30001, 60000]	gastritis
[21, 60]	F	[30001, 60000]	flu
[21, 60]	F	[30001, 60000]	flu

A voter registration list

Name	Age	Sex	Zipcode
Andy	4	M	12000
Ken	6	M	18000
Nash	9	M	19000
Mike	7	M	17000
Alice	12	F	22000
Betty	19	F	24000
Linda	21	F	33000
Jane	25	F	34000
Sarah	28	F	37000
Mary	56	F	58000

# Principle 3: t-Closeness

[Li et al. ICDE '07]

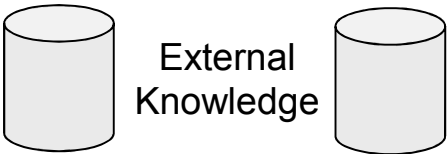
Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Distribution of sensitive attributes within each quasi-identifier group should be “close” to their distribution in the entire original database

Then we can bound the knowledge that the attacker gains by seeing a particular anonymization.

## Adversarial belief



Belief	Knowledge
$B_0$	 External Knowledge
$B_1$	Overall distribution $Q$ of sensitive values
$B_2$	Distribution $P_i$ of sensitive values in each equi-class

## A released table

Age	Zipcode	.....	Gender	Disease
2*	479**	.....	Male	Flu
2*	479**	.....	Male	Heart Disease
2*	479**	.....	Male	Cancer
.	.	.....	.	.
.	.	.....	.	.
.	.	.....	.	.
$\geq 50$	4766*	.....	*	Gastritis

# How can we measure the distance between $\mathbf{P}=(p_1,\dots,p_m)$ and $\mathbf{Q}=(q_1,\dots,q_m)$ ?

$\mathbf{Q}$ : {20K,30K,40K,50K,60K,70K,80K,90K,100K}

$\mathbf{P}_1$ : {20K,30K,40K}

$\mathbf{P}_2$ : {20K,60K,100K}

Intuitively,  $\mathbf{Q}$  is closer to  $\mathbf{P}_1$  than  $\mathbf{P}_2$

Distance $[\mathbf{P},\mathbf{Q}]$  should depend on the “ground distance” between sensitive values.

# Earth mover's distance: the amount of work needed to transform one distribution (histogram) into another

- $\mathbf{P}=(p_1,p_2,\dots,p_m)$ ,  $\mathbf{Q}=(q_1,q_2,\dots,q_m)$
- $d_{ij}$ : the ground distance between element  $i$  of  $\mathbf{P}$  and element  $j$  of  $\mathbf{Q}$ .
- Find a flow  $F=[f_{ij}]$  where  $f_{ij}$  is the flow of mass from element  $i$  of  $\mathbf{P}$  to element  $j$  of  $\mathbf{Q}$  that minimizes the overall work:

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

subject to the constraints:

$$f_{ij} \geq 0 \quad 1 \leq i \leq m, 1 \leq j \leq m \quad (c1)$$

$$p_i - \sum_{j=1}^m f_{ij} + \sum_{j=1}^m f_{ji} = q_i \quad 1 \leq i \leq m \quad (c2)$$

$$\sum_{i=1}^m \sum_{j=1}^m f_{ij} = \sum_{i=1}^m p_i = \sum_{i=1}^m q_i = 1 \quad (c3)$$

Simple formulas for EMD can be derived for several common ground distances.

# Earth Mover's Distance salary example

{3k,4k,5k} and {3k,4k,5k,6k,7k,8k,9k,10k,11k}

Move  $1/9$  probability for each of the following pairs:

✓ 3k→6k,3k→7k      cost:  $1/9*(3+4)/8$

✓ 4k→8k,4k→9k      cost:  $1/9*(4+5)/8$

✓ 5k→10k,5k→11k    cost:  $1/9*(5+6)/8$

➤ Total cost:  $1/9*27/8=0.375$

For {6k,8k,11k}, the cost is  $0.167 < 0.375$ .

# How anonymous is this k-anonymous, l-diverse, and t-close dataset?

Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
Asian/AfrAm	787XX	HIV+	Shingles
Caucas	787XX	HIV-	Acne
Caucas	787XX	HIV-	Shingles
Caucas	787XX	HIV-	Acne

# That depends on what the attacker knows.

*Bob is Caucasian and I heard he was admitted to the hospital with flu...*

This is against the rules, because "flu" is not a quasi-identifier.



Life does not always follow the rules.

Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
Asian/AfrAm	787XX	HIV+	Shingles
Caucas	787XX	HIV-	Acne
Caucas	787XX	HIV-	Shingles
Caucas	787XX	HIV-	Acne

# There are many, many other proposed privacy principles.

## ❑ *k*-gather

- [Aggarwal et al., *PODS* 2006]
- Suffers from the problems of *k*-anonymity.

## ❑ (*a*, *k*)-anonymity

- [Wong et al., *KDD* 2006]

## ❑ Personalized anonymity

- [Xiao and Tao, *SIGMOD* 2006]

## ❑ ***Your thesis here***

# Issues

- Privacy principle

- What is adequate privacy protection?

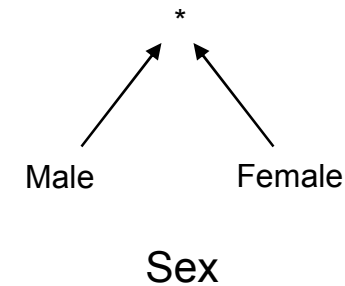
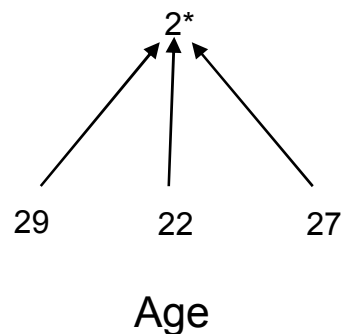
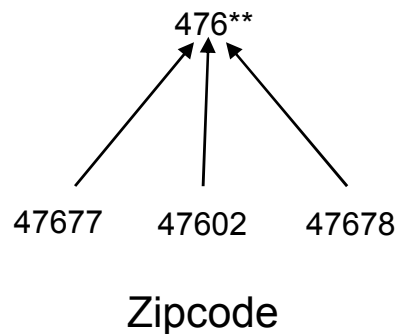
- ➔ □ Distortion approach

- How can we achieve the privacy principle?

# We can generalize or bucketize

## □ Generalization

- Replace with less-specific but semantically-consistent values

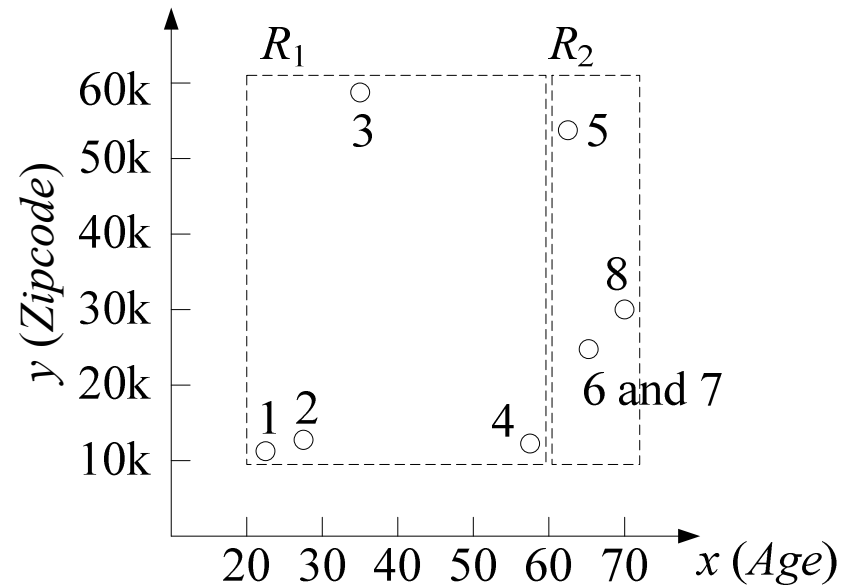


## □ Bucketization (also called “anatomy”)

Each of these approaches can be integrated with all the privacy principles discussed earlier.

# Generalization can be viewed geometrically.

tuple ID	Age	Sex	Zipcode	Disease
1	[21, 60]	M	[10001, 60000]	pneumonia
2	[21, 60]	M	[10001, 60000]	dyspepsia
3	[21, 60]	M	[10001, 60000]	dyspepsia
4	[21, 60]	M	[10001, 60000]	pneumonia
5	[61, 70]	F	[10001, 60000]	flu
6	[61, 70]	F	[10001, 60000]	gastritis
7	[61, 70]	F	[10001, 60000]	flu
8	[61, 70]	F	[10001, 60000]	bronchitis

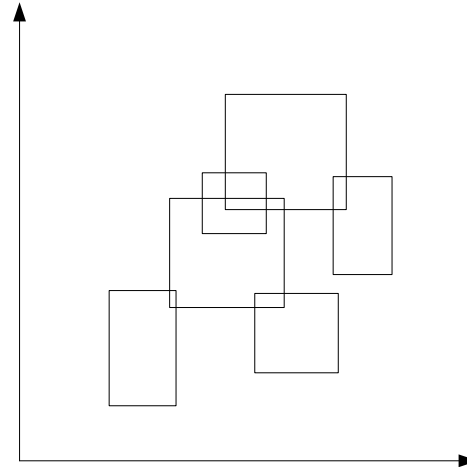


# Taxonomy of generalization

[LeFevre et al. *SIGMOD*, 2005]

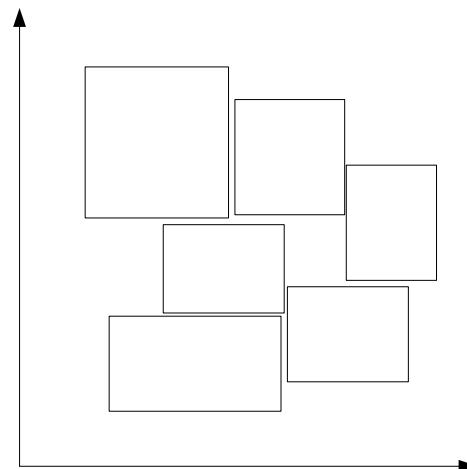
## ❑ Local recoding

- (Generalized) rectangles may overlap.
- Suppression is a special case of local recoding.



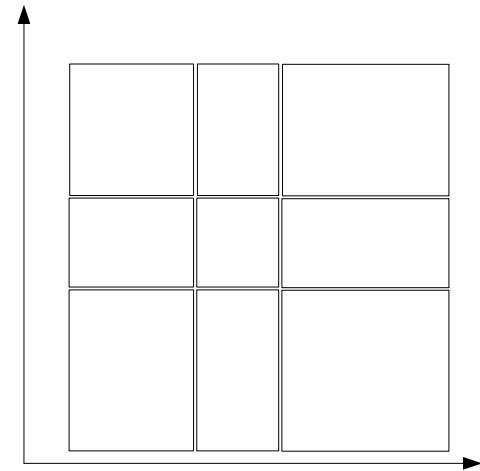
## ❑ Global recoding

- All rectangles are disjoint.



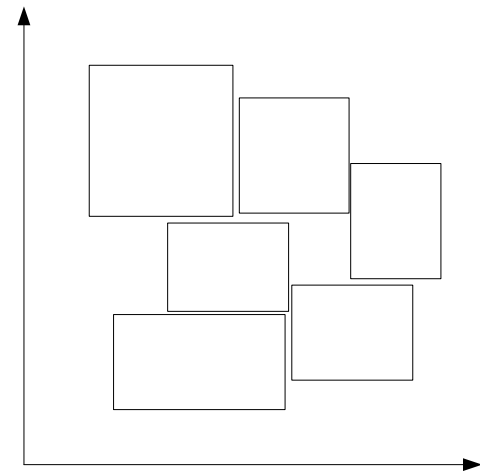
❑ Single-dimension global recoding

➤ Rectangles form a grid.



❑ Multi-dimension global recoding

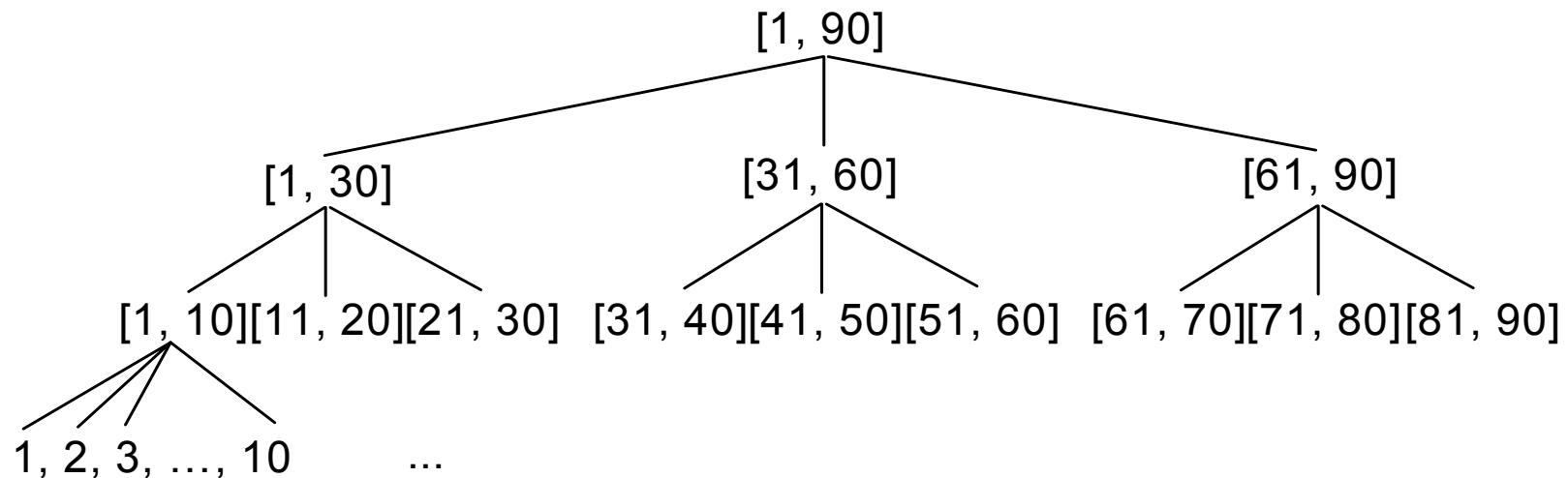
➤ They don't.



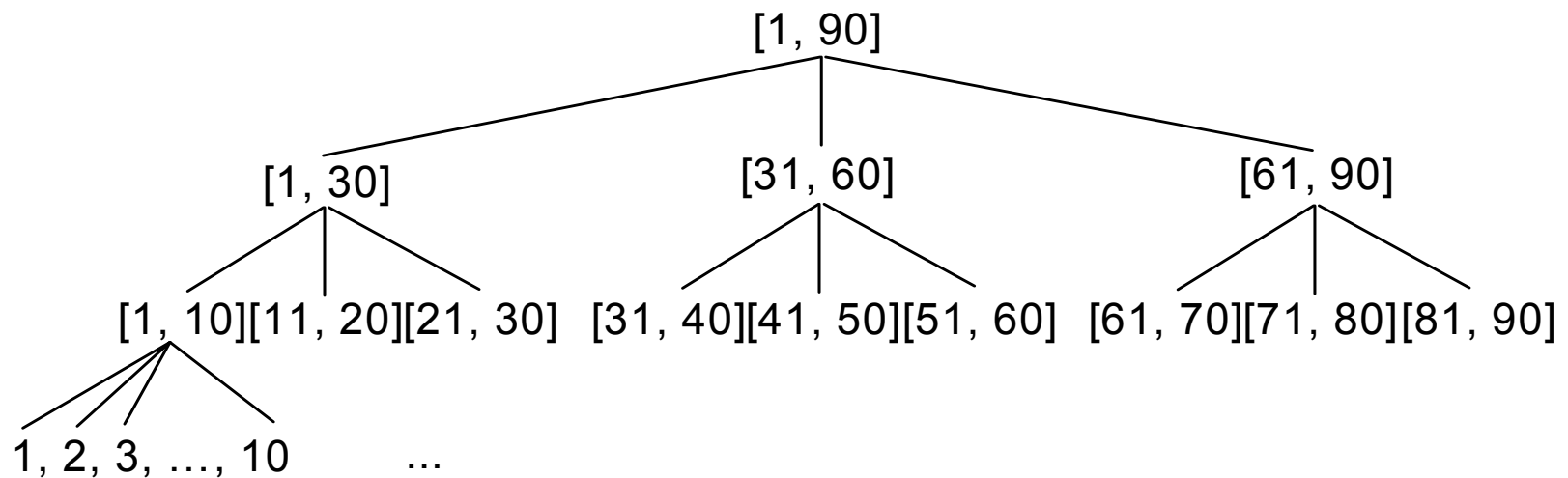
# Single-dimension recoding has two subtypes: full-domain and full-subtree.

Both assume a hierarchy on each QI attribute.

Example: hierarchy on *Age*



In **full-domain recoding**, all values are generalized to the same level of the hierarchy.



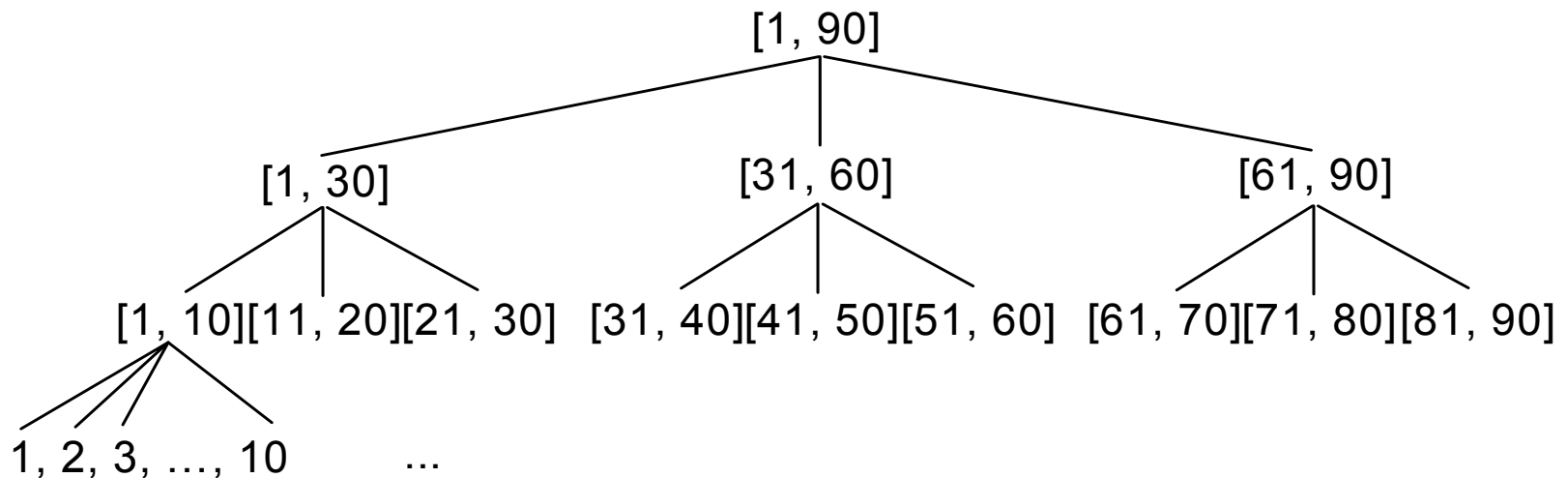
# In full-subtree recoding, the subtrees of all generalized values must be disjoint.

➤ Permissible generalization:

✓ [1, 30], [31, 40], [41, 50], [51, 60], [61, 90].

➤ Illegal generalization:

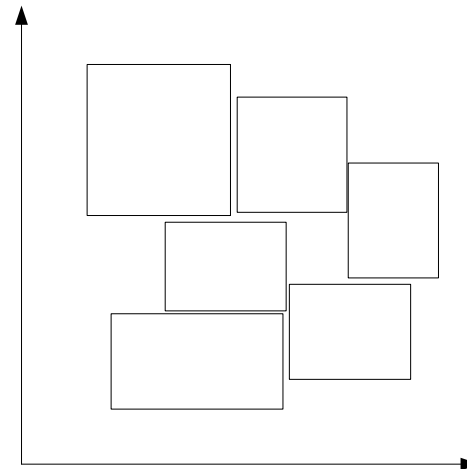
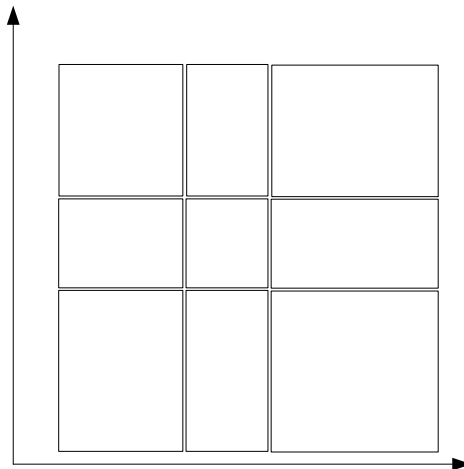
✓ [1, 10], [1, 30], [31, 60], [61, 90].



# Why all these generalization types?

## Reason 1:

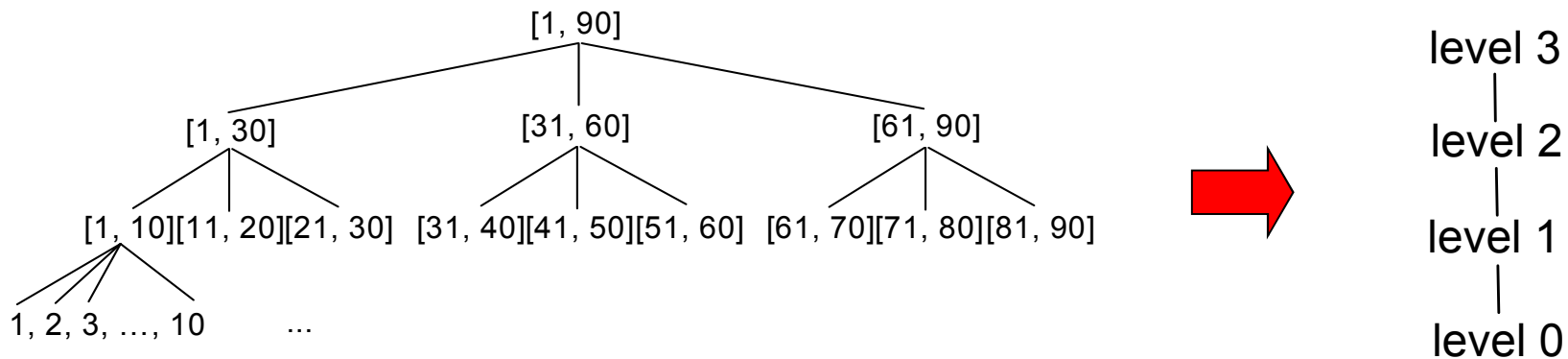
If a dataset is generalized in a **more restricted** manner, **less preprocessing** is required before it can be analyzed by a standard statistical tool (such as SAS).



# Why all these generalization types?

Reason 2:

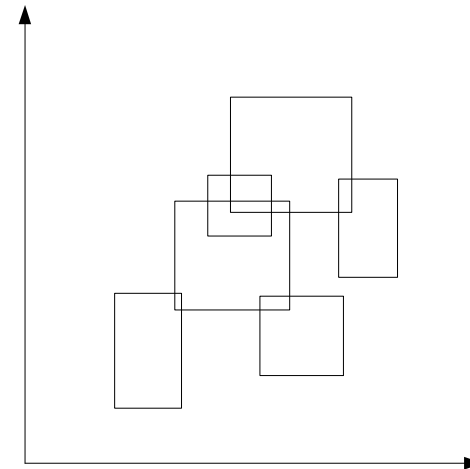
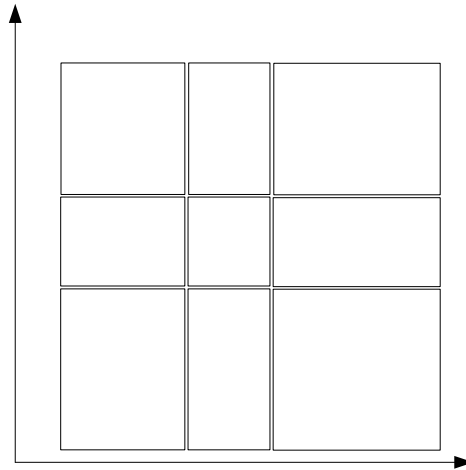
**More restrictive** generalizations are usually **faster to compute** and **easier to analyze**.



# Why all these generalization types?

Reason 3:

Less restrictive generalizations can provide **more accurate data analysis** (i.e., **higher utility**).



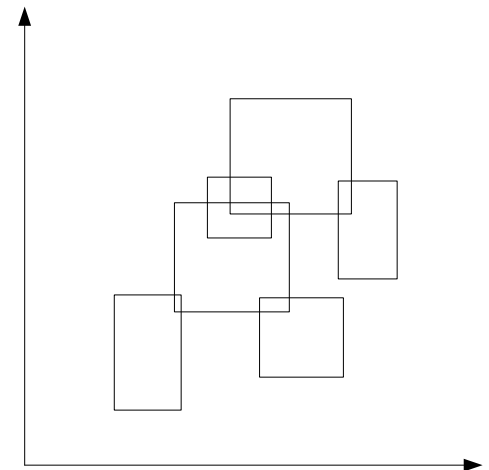
# Generalization algorithms employ a quality metric.

Examples:

- The generalization level (for full-domain recoding)
- Total rectangle size (for local recoding)
- ...

- ❑ Mostly heuristics-based.
- ❑ Finding the optimal generalization is often NP hard.

level 3  
|  
level 2  
|  
level 1  
|  
level 0



# Generalization can introduce needless inaccuracy.

Query A:           SELECT COUNT(\*) from Unknown-Microdata  
WHERE *Disease* = 'pneumonia' AND *Age* in [0, 30]  
AND *Zipcode* in [10001, 20000]

Age	Sex	Zipcode	Disease
[21, 60]	M	[10001, 60000]	pneumonia
[21, 60]	M	[10001, 60000]	dyspepsia
[21, 60]	M	[10001, 60000]	dyspepsia
[21, 60]	M	[10001, 60000]	pneumonia
[61, 70]	F	[10001, 60000]	flu
[61, 70]	F	[10001, 60000]	gastritis
[61, 70]	F	[10001, 60000]	flu
[61, 70]	F	[10001, 60000]	bronchitis

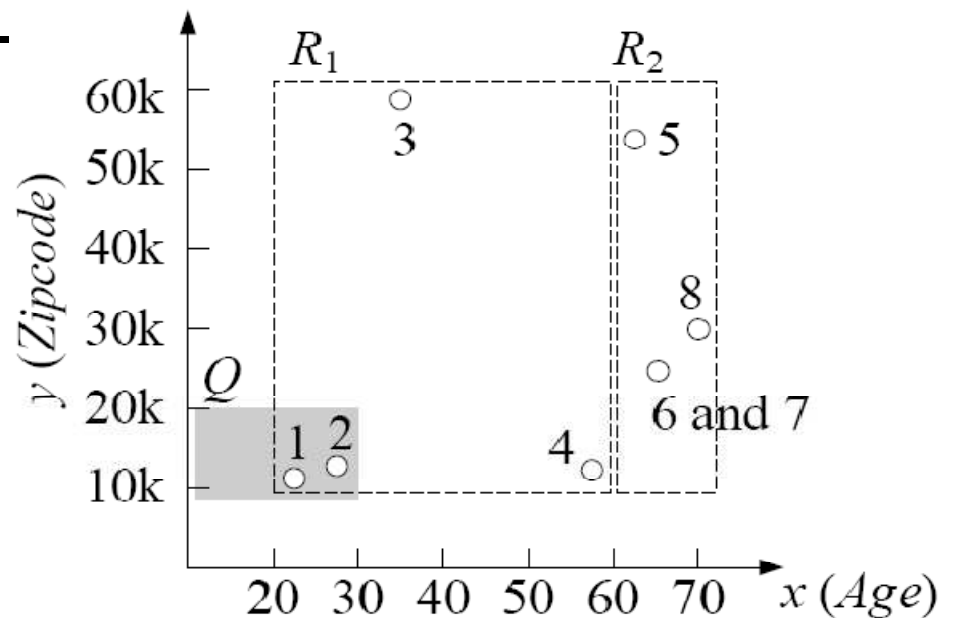
Estimated answer:  $2p$ , where  $p$  is the probability that each of the two tuples satisfies the query conditions on the *Age and Zipcode*.

Query A: SELECT COUNT(\*) from Unknown-Microdata  
 WHERE *Disease* = 'pneumonia' AND *Age* in [0, 30]  
 AND *Zipcode* in [10001, 20000]

Age	Sex	Zipcode	Disease
[21, 60]	M	[10001, 60000]	pneumonia
[21, 60]	M	[10001, 60000]	pneumonia

$$p = \text{Area}(R_1 \cap Q) / \text{Area}(R_1) = 0.05$$

Estimated answer for Query A:  $2p = 0.1$



Query A:           SELECT COUNT(\*) from Unknown-Microdata  
WHERE *Disease* = 'pneumonia' AND *Age* in [0, 30]  
AND *Zipcode* in [10001, 20000]

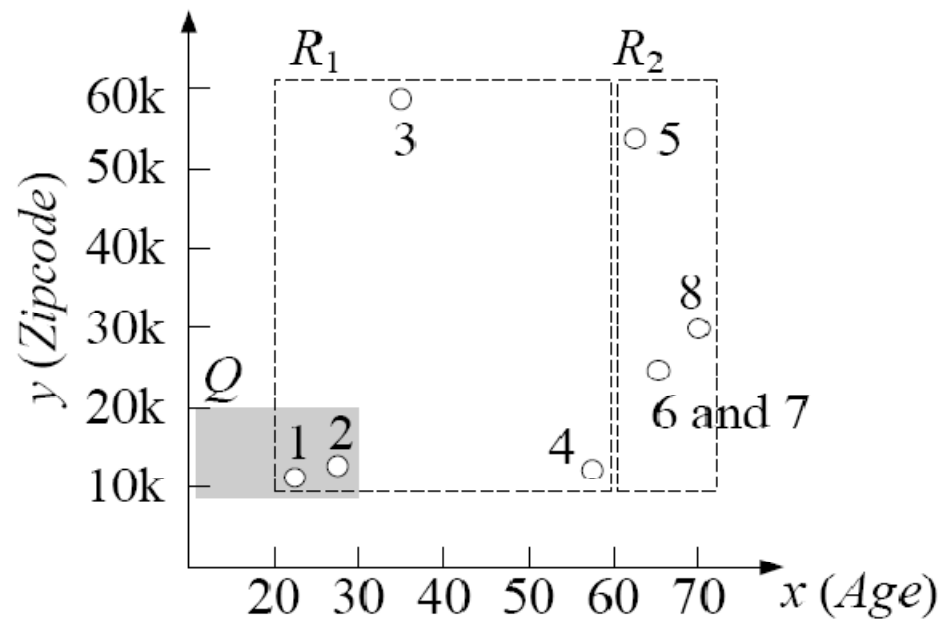
Estimated answer = 0.1

Exact answer = 1

Name	Age	Sex	Zipcode	Disease
Bob	23	M	11000	pneumonia
Ken	27	M	13000	dyspepsia
Peter	35	M	59000	dyspepsia
Sam	59	M	12000	pneumonia
Jane	61	F	54000	flu
Linda	65	F	25000	gastritis
Alice	65	F	25000	flu
Mandy	70	F	30000	bronchitis

# Cause of inaccuracy: QI distribution inside each QI group is lost!

Age	Sex	Zipcode	Disease
[21, 60]	M	[10001, 60000]	pneumonia
[21, 60]	M	[10001, 60000]	pneumonia



# Bucketization releases a quasi-identifier table (QIT) and a sensitive table (ST).

## Sensitive table (ST)

Group-ID	Disease	Count
1	dyspepsia	2
1	pneumonia	2
2	bronchitis	1
2	flu	2
2	gastritis	1

## Quasi-identifier table (QIT)

Age	Sex	Zipcode	Group-ID
23	M	11000	1
27	M	13000	1
35	M	59000	1
59	M	12000	1
61	F	54000	2
65	F	25000	2
65	F	25000	2
70	F	30000	2

## Microdata

Age	Sex	Zipcode	Disease
23	M	11000	pneumonia
27	M	13000	dyspepsia
35	M	59000	dyspepsia
59	M	12000	pneumonia
61	F	54000	flu
65	F	25000	gastritis
65	F	25000	flu
70	F	30000	bronchitis

# Bucketization step 1: choose an $l$ -diverse microdata partition.

	Age	Sex	Zipcode	Disease
QI group 1	23	M	11000	pneumonia
	27	M	13000	dyspepsia
	35	M	59000	dyspepsia
	59	M	12000	pneumonia
QI group 2	61	F	54000	flu
	65	F	25000	gastritis
	65	F	25000	flu
	70	F	30000	bronchitis

A 2-diverse partition

## Bucketization step 2: generate QIT and ST, based on the partition.

group 1

Age	Sex	Zipcode
23	M	11000
27	M	13000
35	M	59000
59	M	12000

group 2

61	F	54000
65	F	25000
65	F	25000
70	F	30000

quasi-identifier table (QIT)

Disease
pneumonia
dyspepsia
dyspepsia
pneumonia
flu
gastritis
flu
bronchitis

sensitive table (ST)

Age	Sex	Zipcode	Group-ID
-----	-----	---------	----------

23	M	11000	1
27	M	13000	1
35	M	59000	1
59	M	12000	1

61	F	54000	2
65	F	25000	2
65	F	25000	2
70	F	30000	2

quasi-identifier table (QIT)

Group-ID	Disease
----------	---------

1	pneumonia
1	dyspepsia
1	dyspepsia
1	pneumonia

2	flu
2	gastritis
2	flu
2	bronchitis

sensitive table (ST)

With bucketization, an adversary can infer the sensitive value of each individual with confidence at most  $1 / l$ .

Name	Age	Sex	Zipcode
Bob	23	M	11000

Age	Sex	Zipcode	Group-ID
23	M	11000	1
27	M	13000	1
35	M	59000	1
59	M	12000	1
61	F	54000	2
65	F	25000	2
65	F	25000	2
70	F	30000	2

Group-ID	Disease	Count
1	dyspepsia	2
1	pneumonia	2
2	bronchitis	1
2	flu	2
2	gastritis	1

sensitive table (ST)

quasi-identifier table (QIT)

# Bucketization can improve the published data's utility (accuracy of analysis).

Query A:                   SELECT COUNT(\*) from Unknown-Microdata  
WHERE *Disease* = 'pneumonia' AND *Age* in [0, 30]  
AND *Zipcode* in [10001, 20000]

Age	Sex	Zipcode	Group-ID
23	M	11000	1
27	M	13000	1
35	M	59000	1
59	M	12000	1
61	F	54000	2
65	F	25000	2
65	F	25000	2
70	F	30000	2

Quasi-identifier table (QIT)

Group-ID	Disease	Count
1	dyspepsia	2
1	pneumonia	2
2	bronchitis	1
2	flu	2
2	gastritis	1

Sensitive table (ST)

Query A:           SELECT COUNT(\*) from Unknown-Microdata  
WHERE *Disease* = 'pneumonia' AND *Age* in [0, 30]  
AND *Zipcode* in [10001, 20000]

	Age	Sex	Zipcode	Group-ID
$t_1$	23	M	11000	1
$t_2$	27	M	13000	1
$t_3$	35	M	59000	1
$t_4$	59	M	12000	1

2 patients contracted pneumonia

2 out of 4 patients satisfy the query conditions on *Age* and *Zipcode*

Estimated answer =  $2 * 2 / 4 = 1$ .

# Bucketization can leak the presence of an individual in the microdata.

Age	Sex	Zipcode	Group-ID
23	M	11000	1
27	M	13000	1
35	M	59000	1
59	M	12000	1
61	F	54000	2
65	F	25000	2
65	F	25000	2
70	F	30000	2

Group-ID	Disease	Count
1	dyspepsia	2
1	pneumonia	2
2	bronchitis	1
2	flu	2
2	gastritis	1

Age	Sex	Zipcode	Disease
[21, 60]	M	[10001, 60000]	pneumonia
[21, 60]	M	[10001, 60000]	dyspepsia
[21, 60]	M	[10001, 60000]	dyspepsia
[21, 60]	M	[10001, 60000]	pneumonia
[61, 70]	F	[10001, 60000]	flu
[61, 70]	F	[10001, 60000]	gastritis
[61, 70]	F	[10001, 60000]	flu
[61, 70]	F	[10001, 60000]	bronchitis

# This field is still very active.

- ❑ Tackle stronger background knowledge
  - [Martin et al., *ICDE* 2007]
- ❑ Improve utility
  - [Kifer and Gehrke, *SIGMOD* 2006]
- ❑ Apply to specific (non-trivial) applications
  - Location privacy
    - ✓ [Mokbel et al., *VLDB* 2006]
- ❑ Data re-publication with updated values (under submission)
- ❑ ***Your thesis here***