Fast Data Anonymization with Low Information Loss

Gabriel Ghinita[1]

Panagiotis Karras[2]

Panos Kalnis[1]

Nikos Mamoulis[2]

[1] National University of Singapore
{ghinitag,kalnis}@comp.nus.edu.sg
[2] Hong Kong University
{pkarras,nikos}@cs.hku.hk

Privacy-Preserving Data Publishing

- Large amounts of public data
  - Research or statistical purposes
  - e.g. distribution of disease for age, city
- Data may contain sensitive information
  - Ensure data privacy

Privacy Violation Example

Age	ZipCode	Disease
42	52000	Ulcer
47	43000	Pneumonia
51	32000	Flu
55	27000	Gastritis
62	41000	Dyspepsia
67	55000	Dyspepsia

(a) Microdata

Name	Age	ZipCode	Disease
Andy	42	52000	Ulcer
Bill	47	43000	Pneumonia
Ken	51	32000	Flu
Nash	55	27000	Gastritis
Mike	62	41000	Dyspepsia
Sam	67	55000	Dyspepsia

(b) Voting Registration List (public)

k-anonymity[Sam01]

- QID generalization or suppression

Age	ZipCode	Disease
42-47	43000-52000	Ulcer
42-47	43000-52000	Pneumonia
51-55	27000-32000	Flu
51-55	27000-32000	Gastritis
62-67	41000-55000	Dyspepsia
62-67	41000-55000	Dyspepsia

(a) 2-anonymous microdata

Name	Age	ZipCode	Disease
Andy	42	52000	Ulcer or Pneumonia
Bill	47	43000	
Ken	51	32000	Flu or Gastritis
Nash	55	27000	
Mike	62	41000	Dyspepsia
Sam	67	55000	

(b) Voting Registration List (public)

Privacy Violation!

[Sam01] P. Samarati, "Protecting Respondent's Privacy in Microdata Release," in IEEE TKDE, vol. 13, n. 6, November/December 2001, pp. 1010-1027.

ℓ-diversity[MGKV06]

- At least ℓ sensitive attribute (SA) values "well-represented" in each group
  - e.g. freq. of an SA value in a group < 1/ℓ

[MGKV06] A. Machanavajjhala et al. ℓ-diversity: Privacy Beyond *k*-anonymity, Proceedings of the 22nd International Conference on Data Engineering (ICDE), 2006

Problem Statement

- Find *k*-anonymous/ℓ-diverse transformation
- Minimize information loss
- Incur reduced anonymization overhead

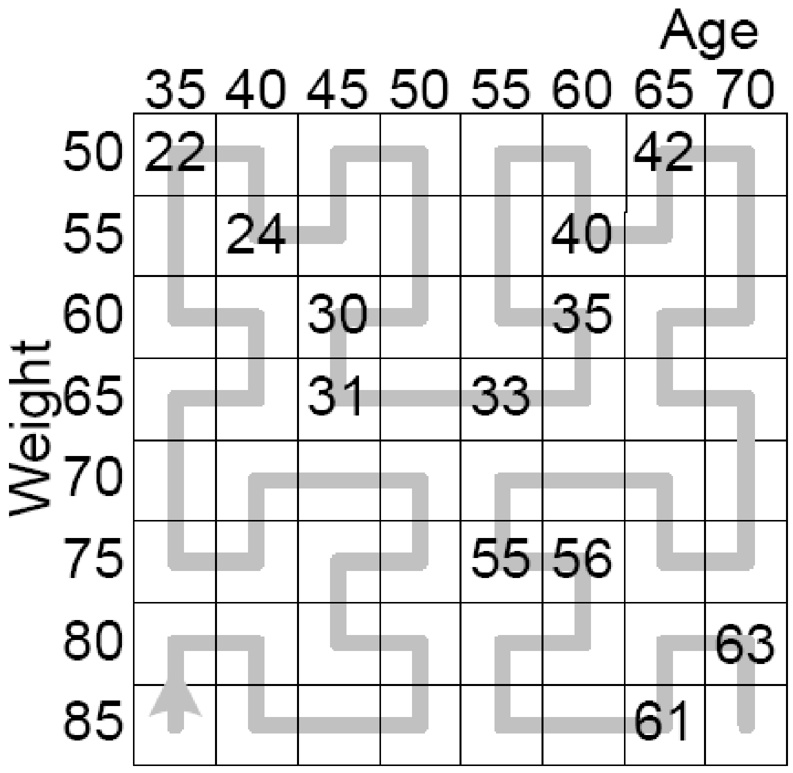
Contributions

- 1D QID
  - Linear, optimal *k*-anonymous partitioning
  - Polynomial, optimal ℓ-diverse partitioning
  - Linear heuristic for ℓ-diverse partitioning

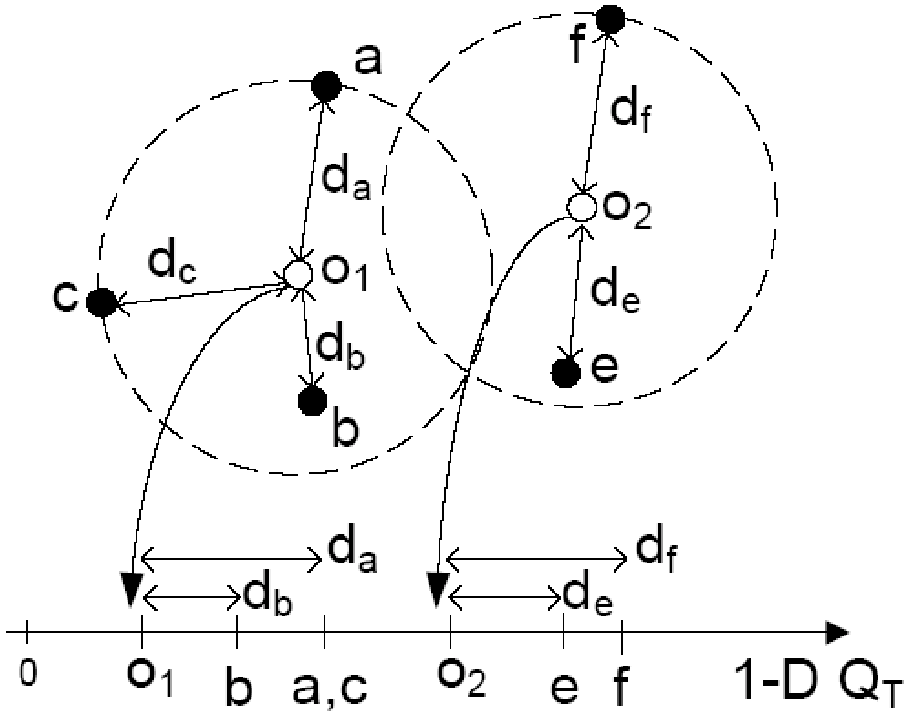
- Generalization to multi-dimensional QID
  - Multi-to-1D mapping
    - Hilbert Space-Filling Curve
    - i-Distance
  - Apply 1D algorithms

Multi-dimensional QID

- Dimensionality Mapping



(a) Hilbert curve

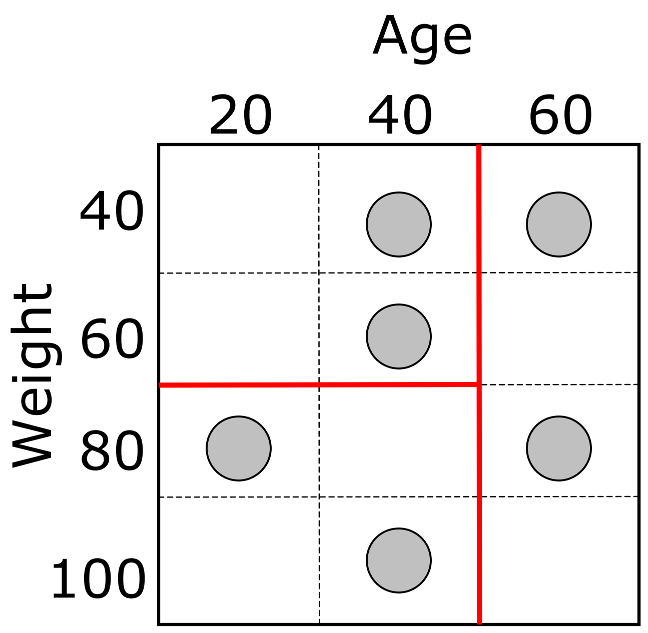


(b) iDistance

State-of-the-art: Mondrian[FWR06]

- Generalization-based
  - data-space partitioning
  - similar to k-d-trees
    - split recursively as long as privacy condition holds

k = 2

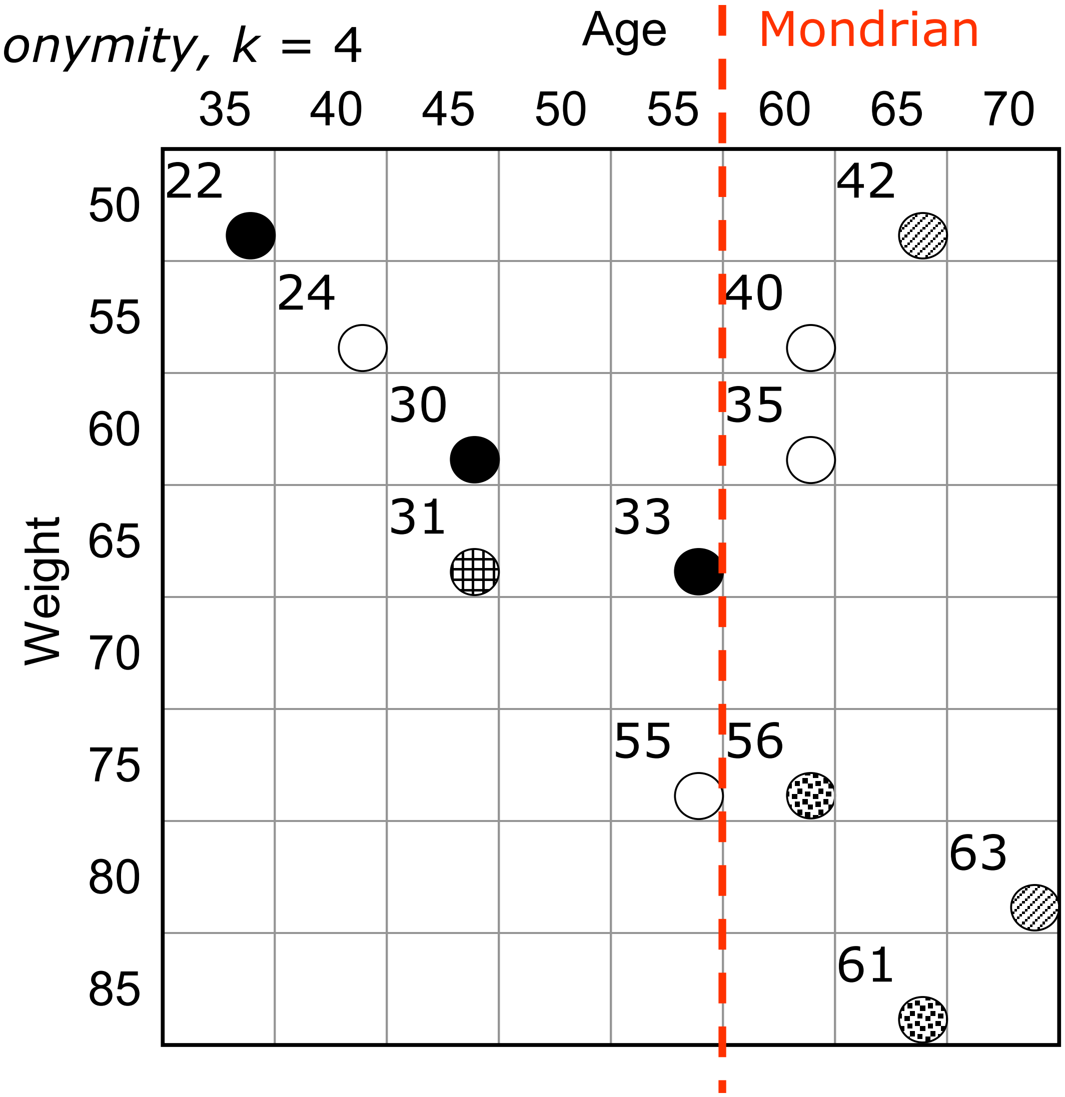


[FWR06] K. LeFevre et al. Mondrian Multidimensional *k*-anonymity, Proceedings of the 22nd International Conference on Data Engineering (ICDE), 2006

Motivating Example

k-anonymity, k = 4

Age	*Weight*	*Disease*	*1D*
35	50	Gastritis ●	22
40	55	Diabetes ○	24
45	60	Gastritis ●	30
45	65	Pneumonia ⊕	31
55	65	Gastritis ●	33
60	60	Diabetes ○	35
60	55	Diabetes ○	40
65	50	Alzheimer ▨	42
55	75	Diabetes ○	55
60	75	Flu ⊛	56
65	85	Flu ⊛	61
70	80	Alzheimer ▨	63

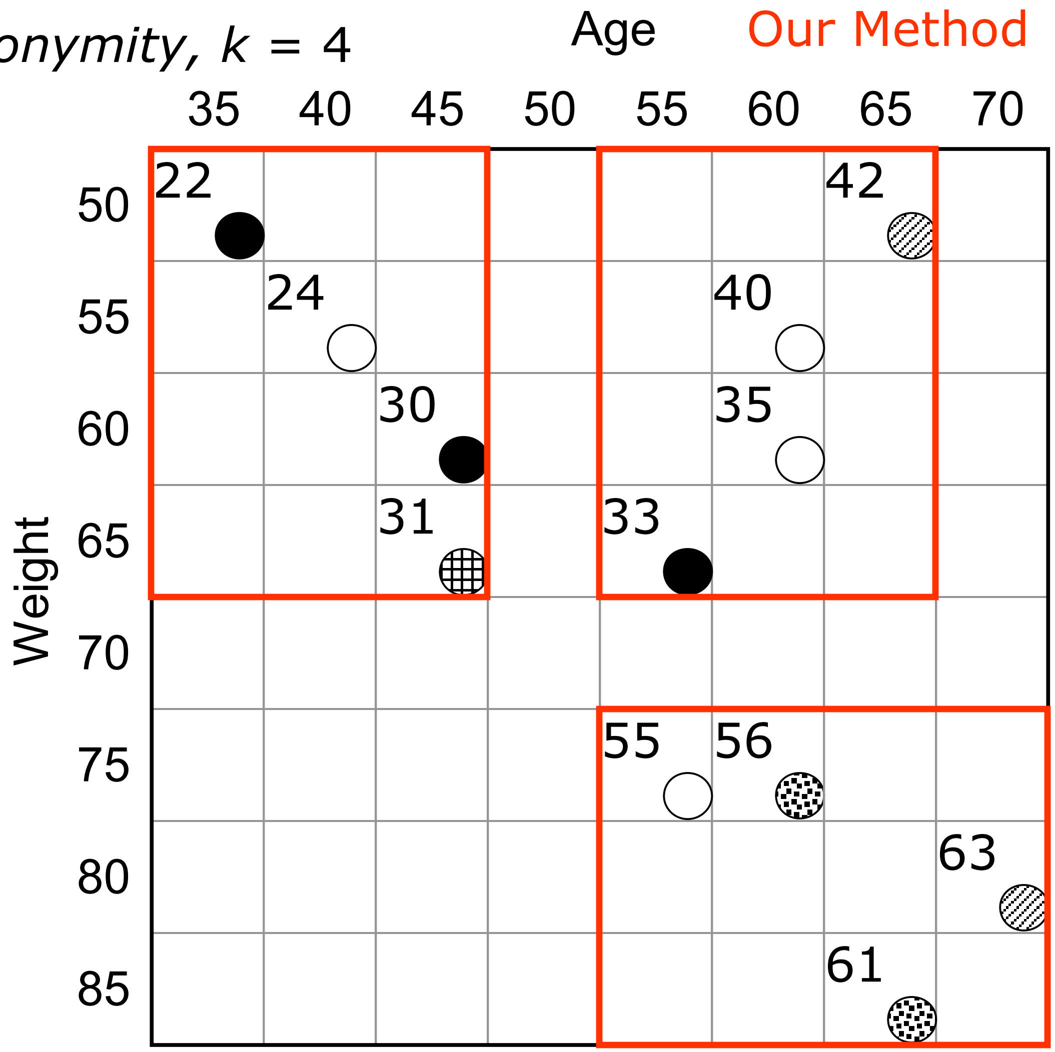


Motivating Example

k-anonymity, k = 4

Our Method

Age	*Weight*	*Disease*		*1D*
35	50	Gastritis	●	22
40	55	Diabetes	○	24
45	60	Gastritis	●	30
45	65	Pneumonia	⊞	31
55	65	Gastritis	●	33
60	60	Diabetes	○	35
60	55	Diabetes	○	40
65	50	Alzheimer	▨	42
55	75	Diabetes	○	55
60	75	Flu	▩	56
65	85	Flu	▩	61
70	80	Alzheimer	▨	63

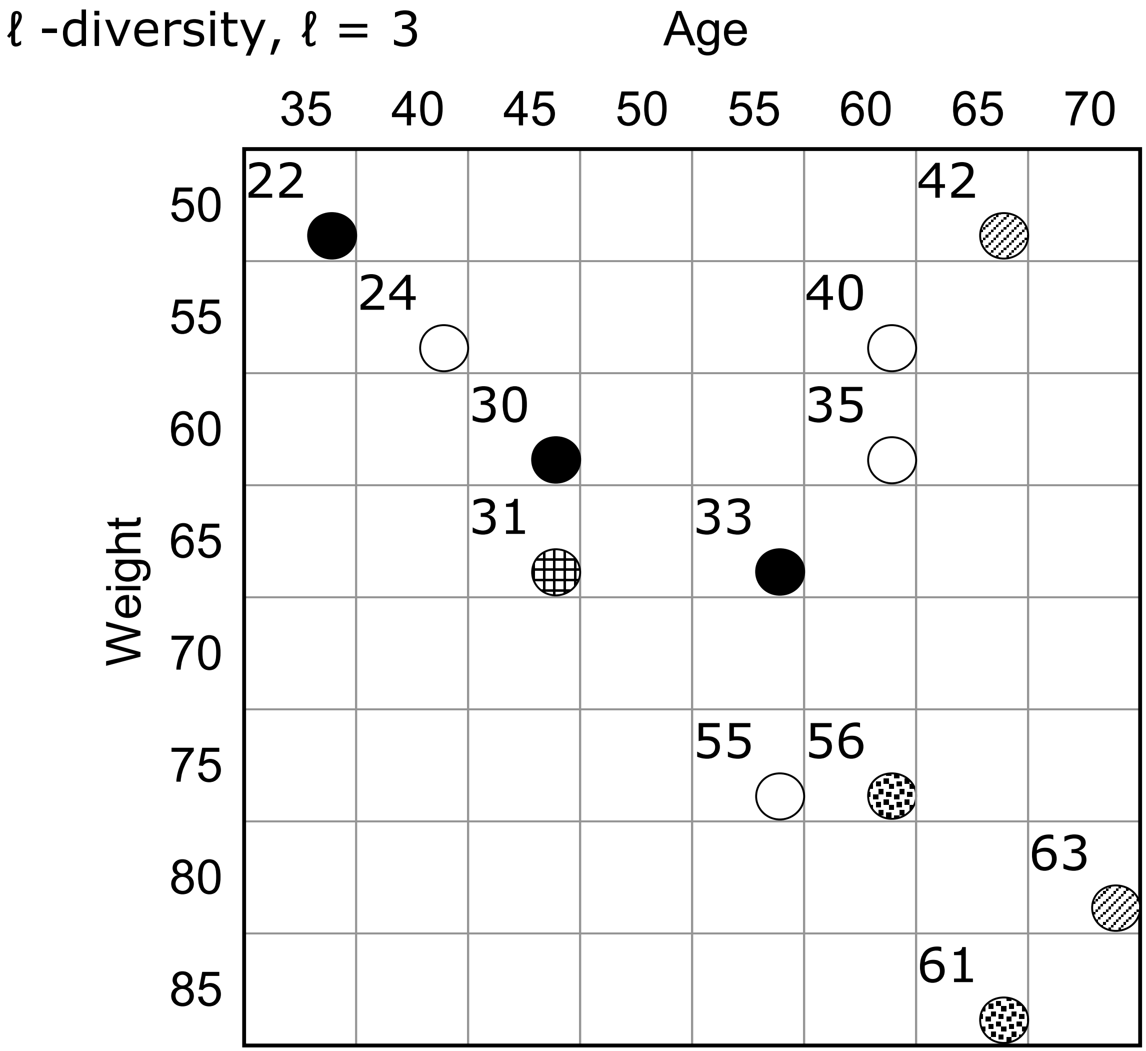


Motivating Example

Mondrian

Performs

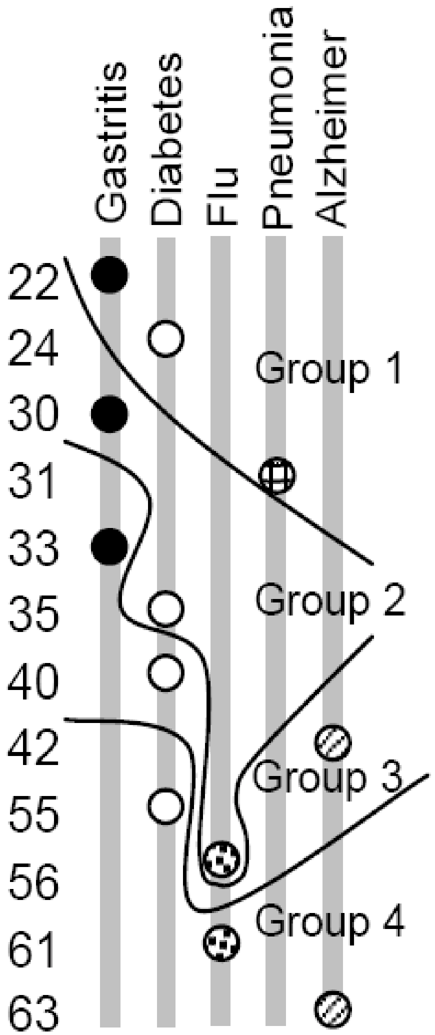
NO SPLIT!



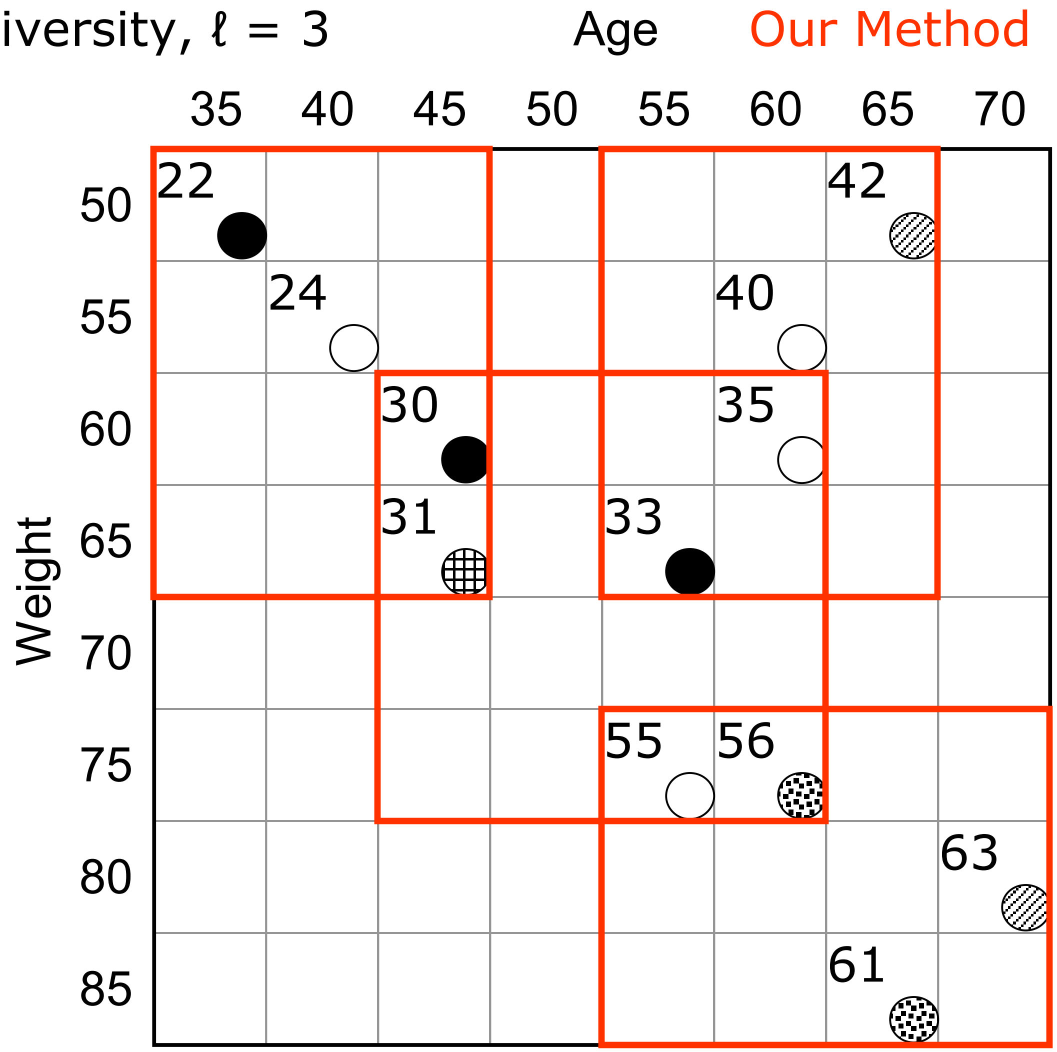
ℓ -diversity, ℓ = 3

Motivating Example

ℓ -diversity, ℓ = 3

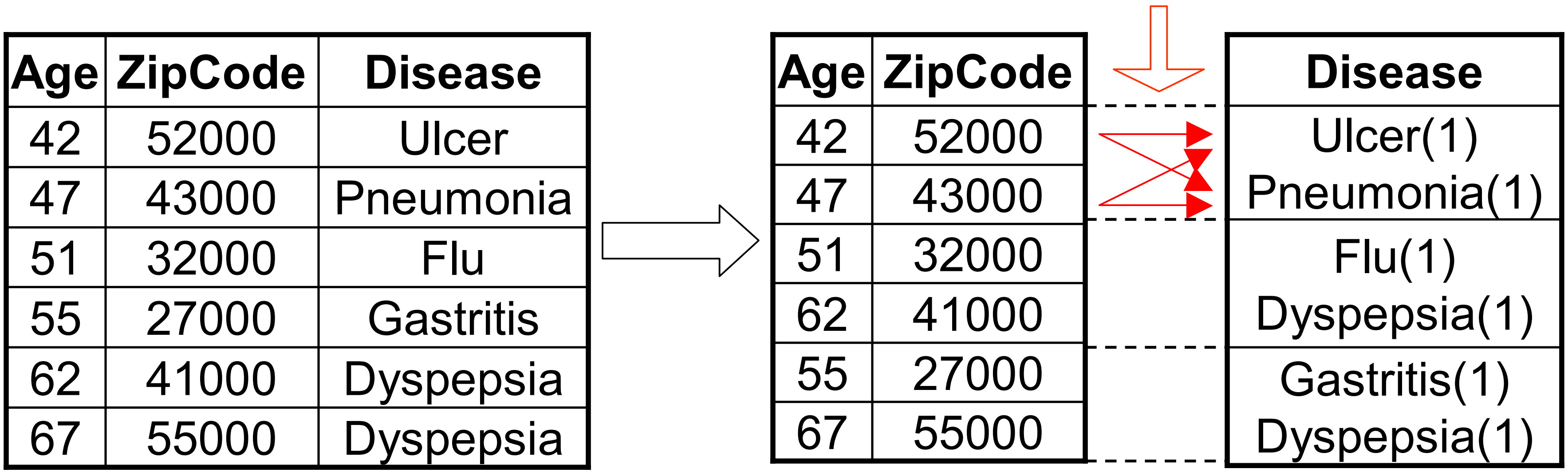


Our Method



State-of-the-art: Anatomy[XT06]

- Permutation-based method
  - discloses exact QID values
  - vulnerable to presence attacks



"Anatomized" table

|G|! permutations

Age	ZipCode	Disease
42	52000	Ulcer
47	43000	Pneumonia
51	32000	Flu
55	27000	Gastritis
62	41000	Dyspepsia
67	55000	Dyspepsia

Age	ZipCode
42	52000
47	43000
51	32000
62	41000
55	27000
67	55000

Disease

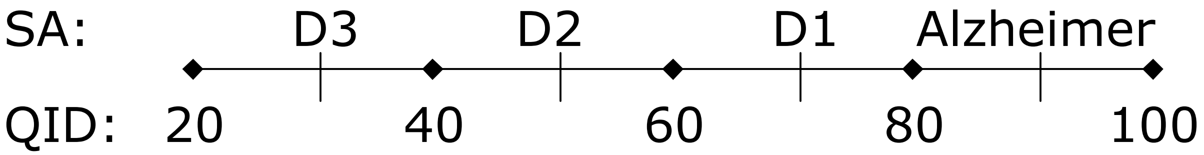
Ulcer(1)
Pneumonia(1)

Flu(1)
Dyspepsia(1)

Gastritis(1)
Dyspepsia(1)

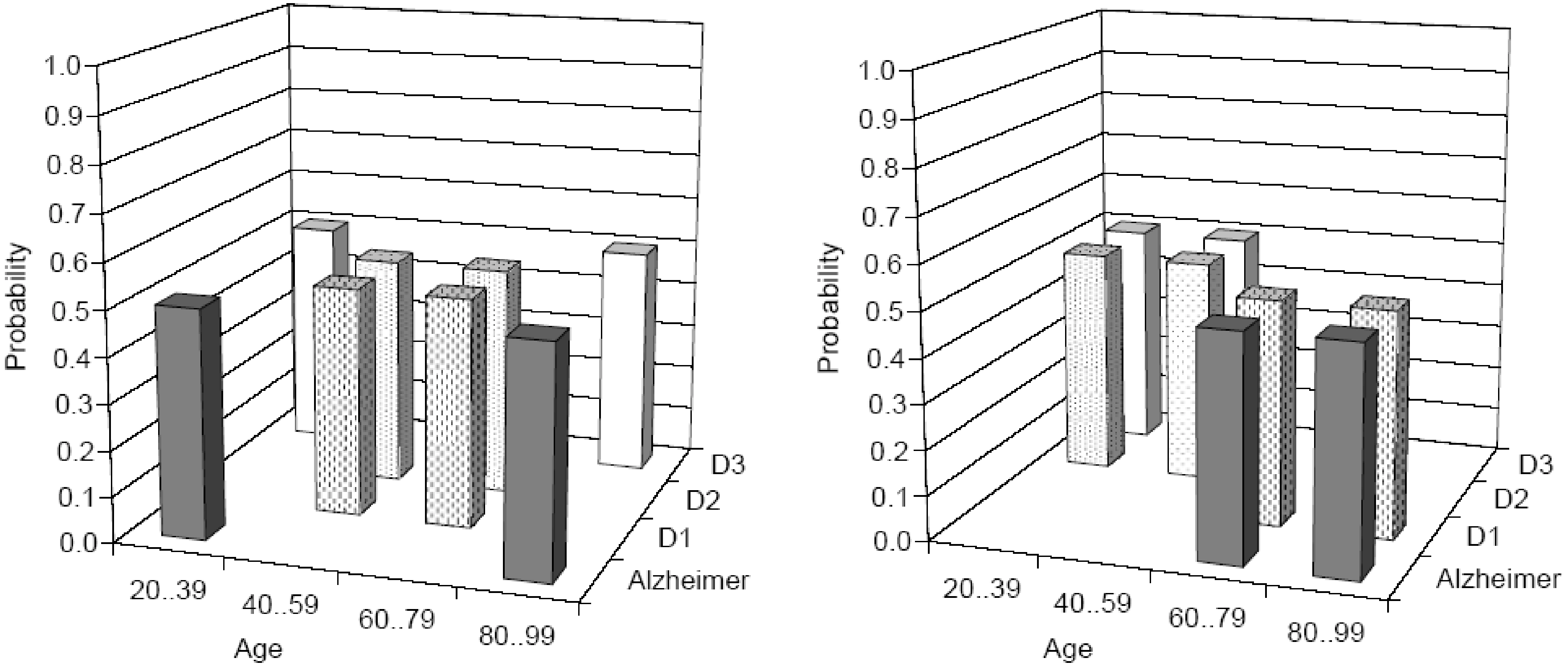
[XT06] X. Xiao and Y. Tao. Anatomy: simple and effective privacy preservation, Proceedings of the 32nd international conference on Very Large Data Bases (VLDB), 2006

Limitation of Anatomy



SA: D3 D2 D1 Alzheimer

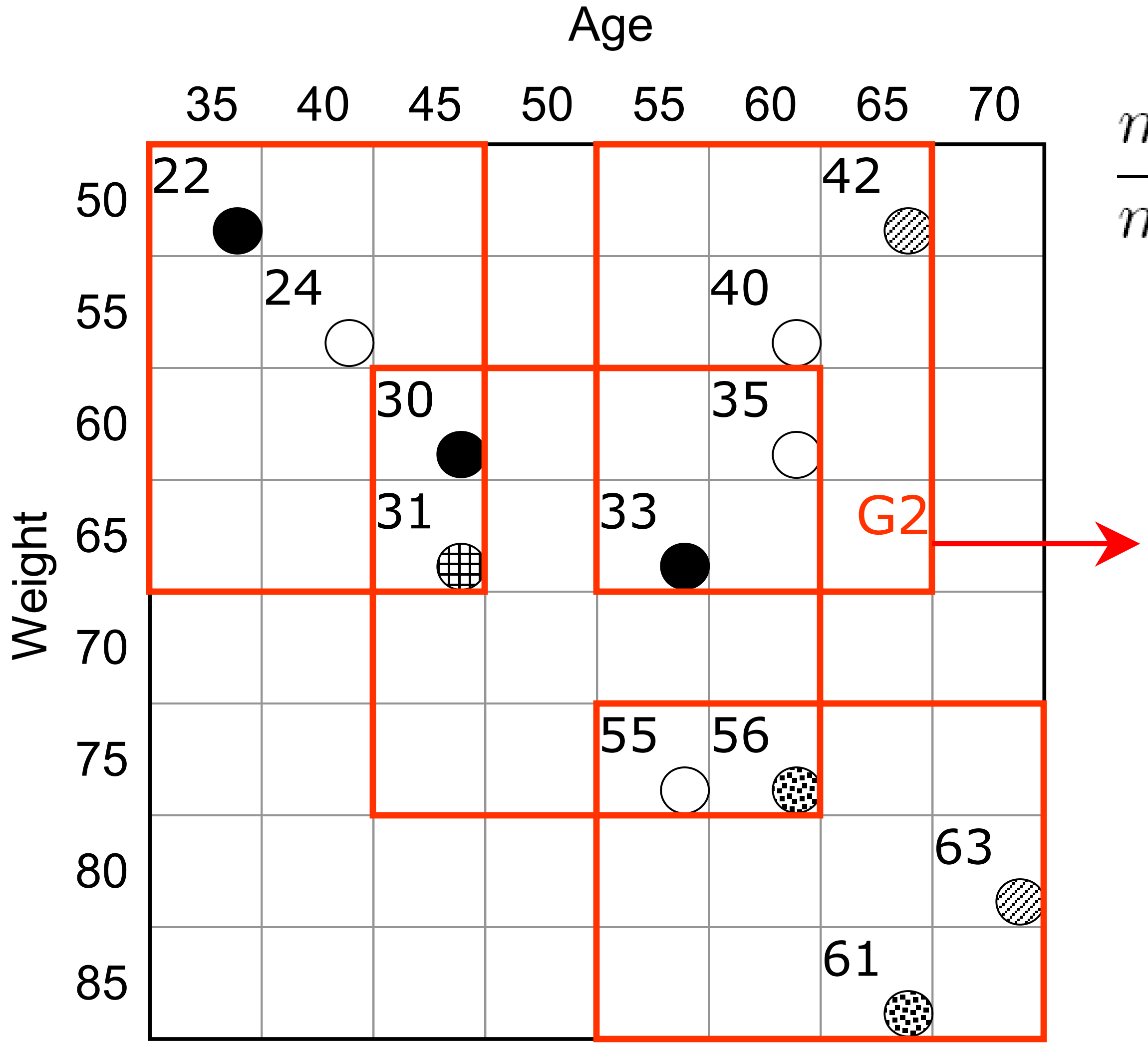
QID: 20 40 60 80 100



(a) Anatomy

(b) Our Approach

Information Loss (Numerical Data)

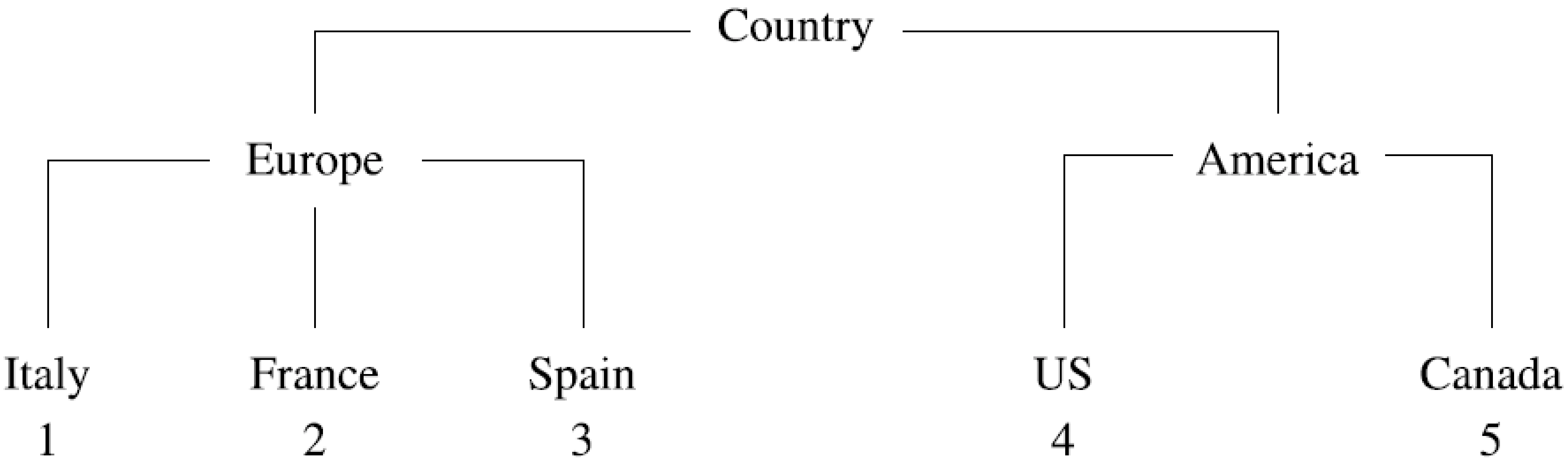


$$\frac{max^{G}_{A_{Num}} - min^{G}_{A_{Num}}}{max_{A_{Num}} - min_{A_{Num}}}$$

$$IL_{Age}(G_2) = \frac{65-55}{70-35}$$

$$IL_{Weight}(G_2) = \frac{65-50}{85-50}$$

Information Loss (Categorical Data)



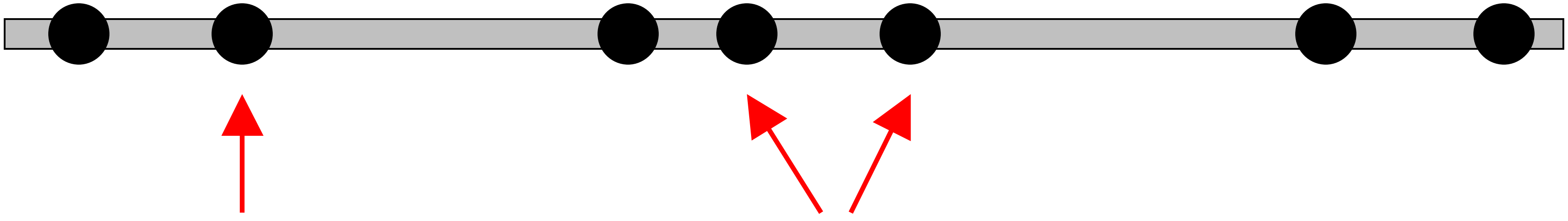
$$IL = \begin{cases} 0, & card(u) = 1 \\ card(u)/|A_{Cat}|, & otherwise \end{cases}$$

IL({Italy, Spain}) = 3/5

Optimal 1D *k*-anonymity

- Properties of optimal solution
  - Groups do not overlap in QID space
  - Group size bounded by 2*k*-1
- DP Formulation *O*(*kN*)

$$Opt(i) = \min_{i-2k<j\leq i-k} (Opt(j) + Opt_I([j+1, i]))$$

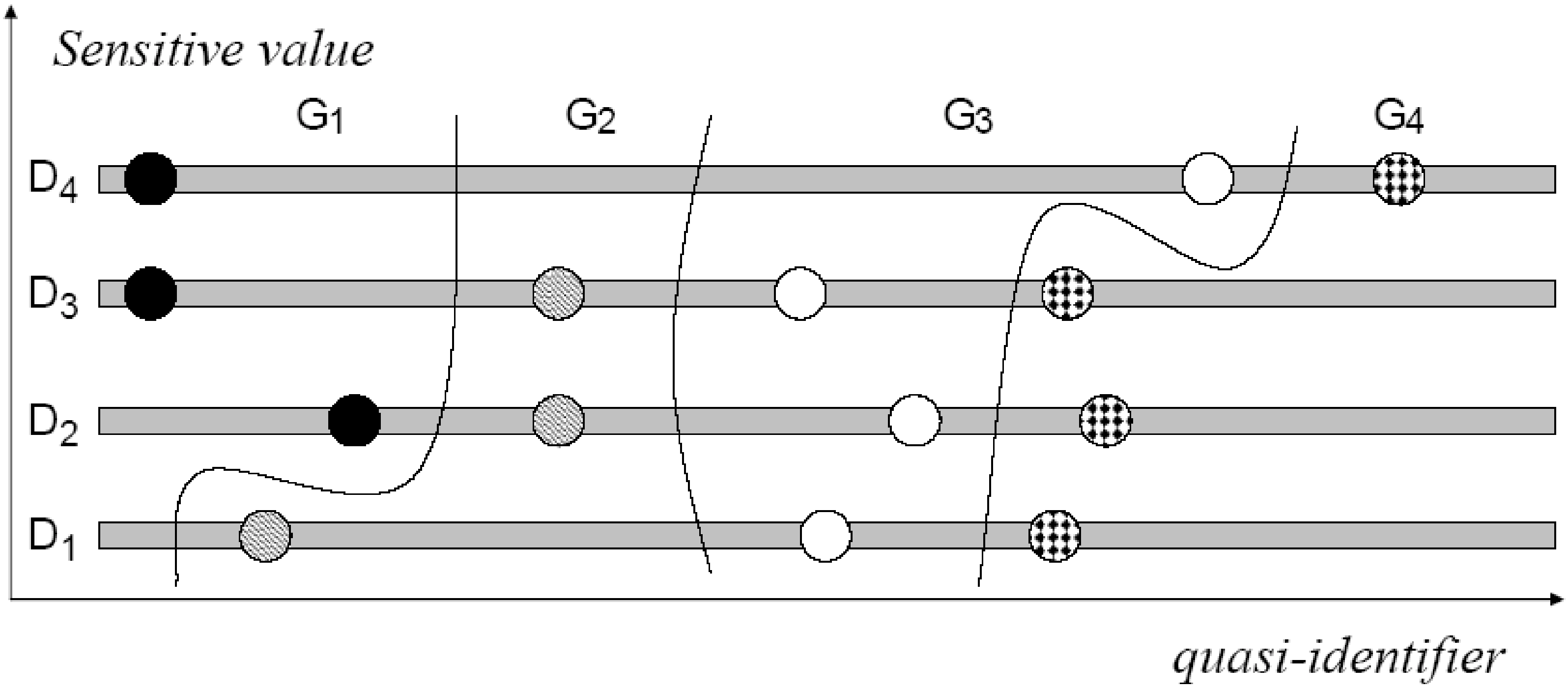


j: end record of previous group

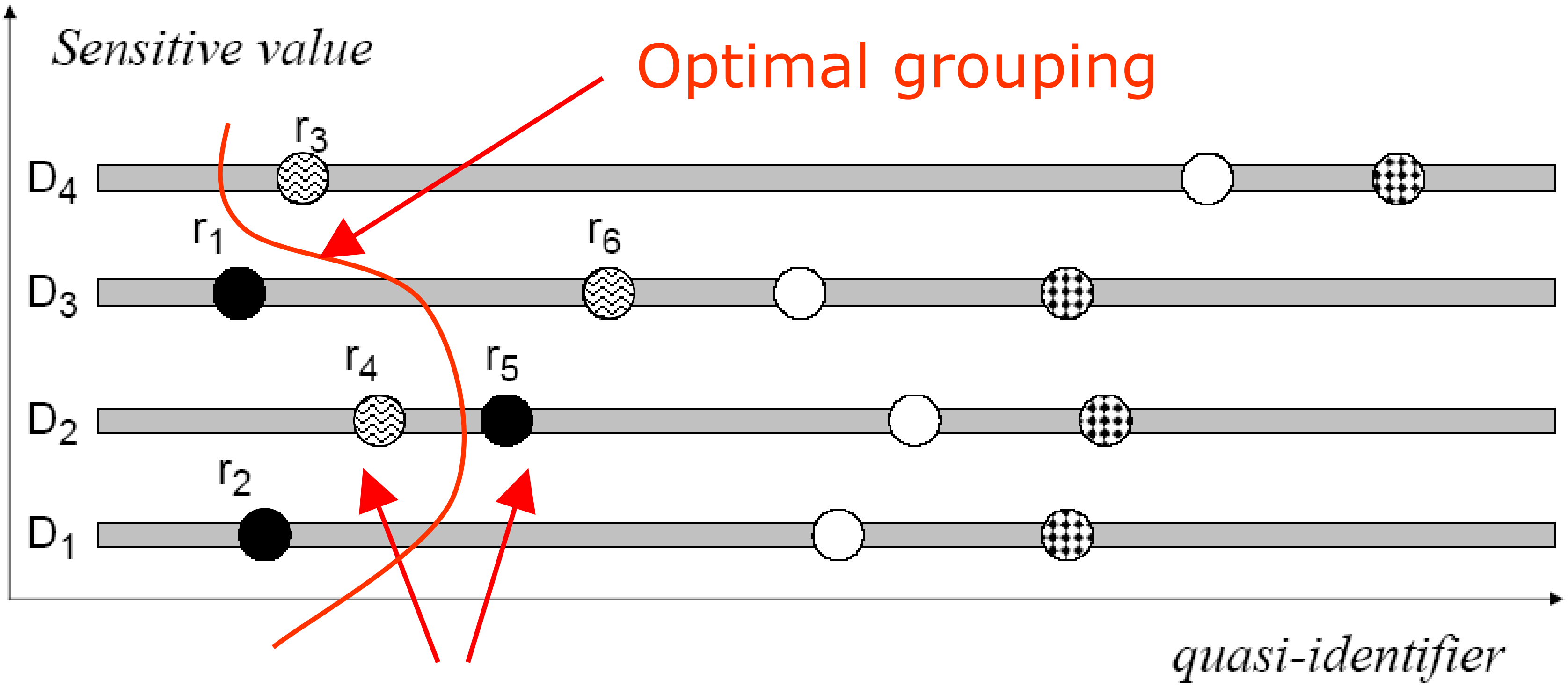
i: end record candidates for current group

Optimal 1D ℓ-diversity

- Properties of optimal solution
  - Group size bounded by 2ℓ-1
  - But groups MAY overlap in QID space
- SA Domain Representation



Group Order Property

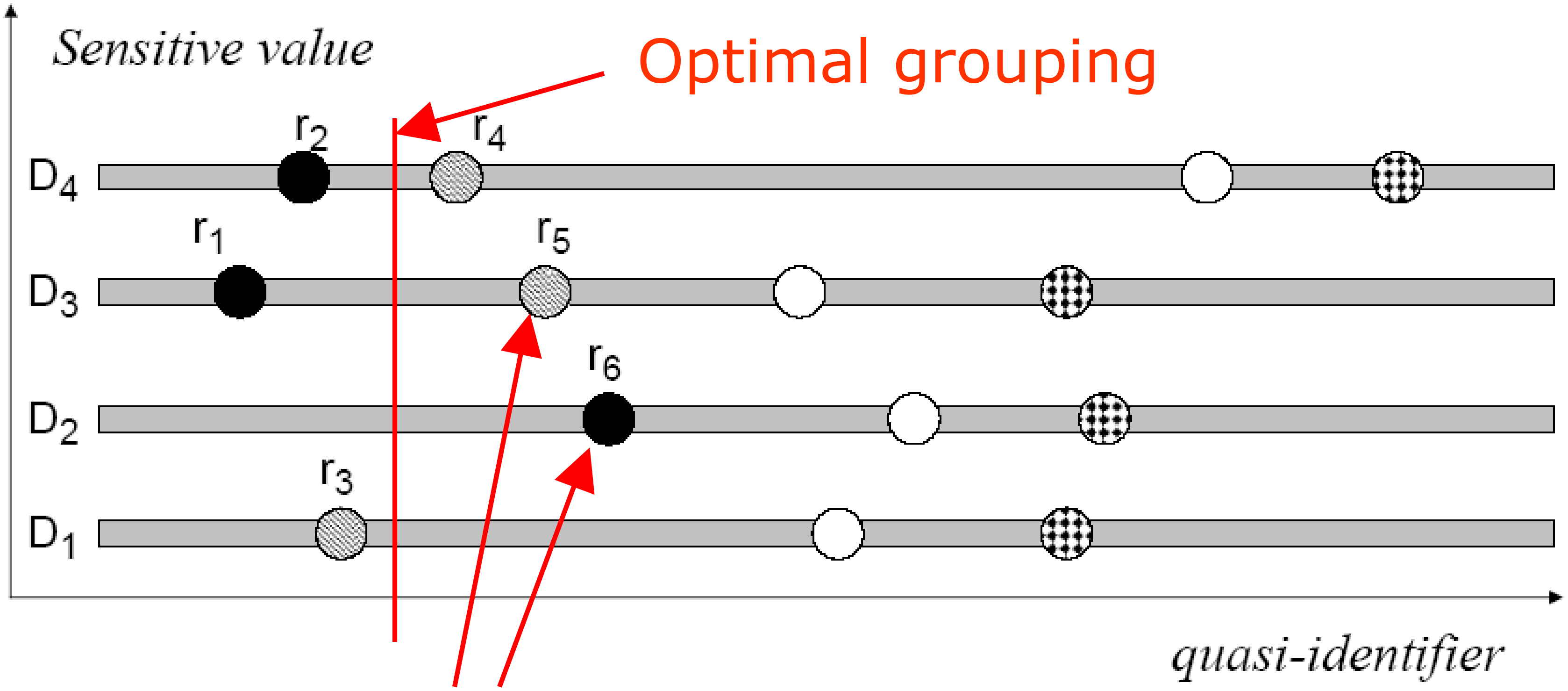


Optimal grouping

violation of group order

Order of groups in each domain is THE SAME

Border Order Property

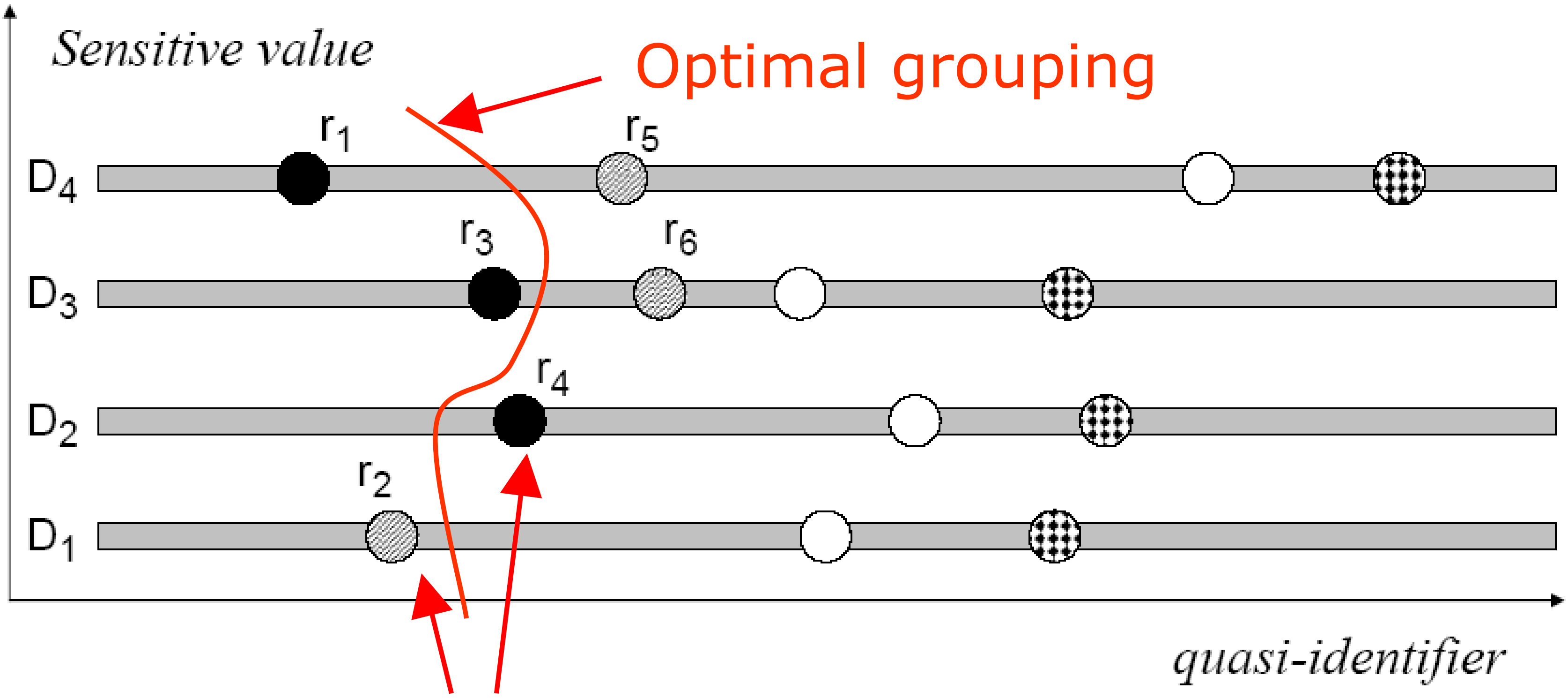


Optimal grouping

violation of border order

“begin” and “end” records in each group follow the same order

Cover Property



Optimal grouping

violation of cover order

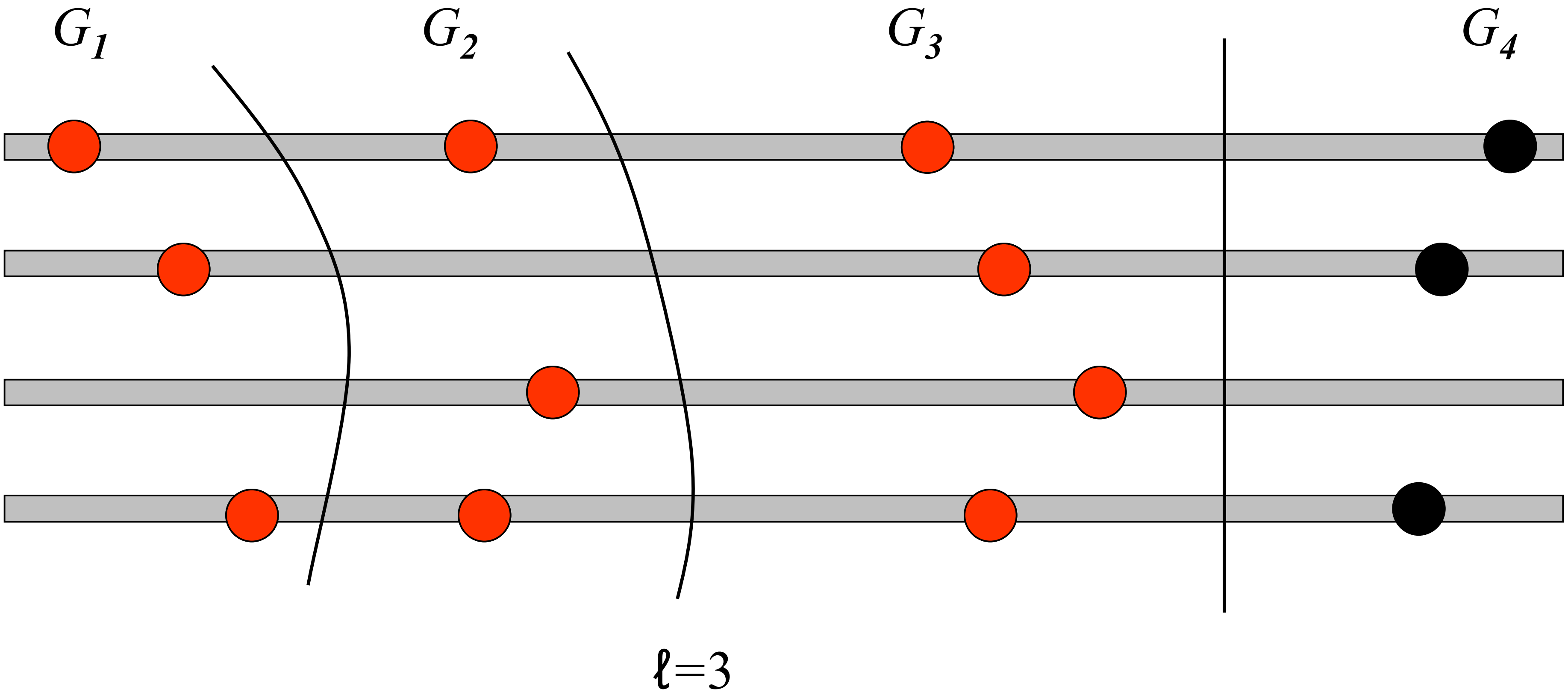
record *r* that can be added to two groups should belong to the “closest” group to *r*

1D ℓ-diversity Heuristic

- Optimal algorithm is polynomial
  - But may be costly in practice
- Linear heuristic algorithm
  - Considers single "frontier of search"
  - Frontier consists of first non-assigned record in each domain

1D ℓ-diversity Heuristic

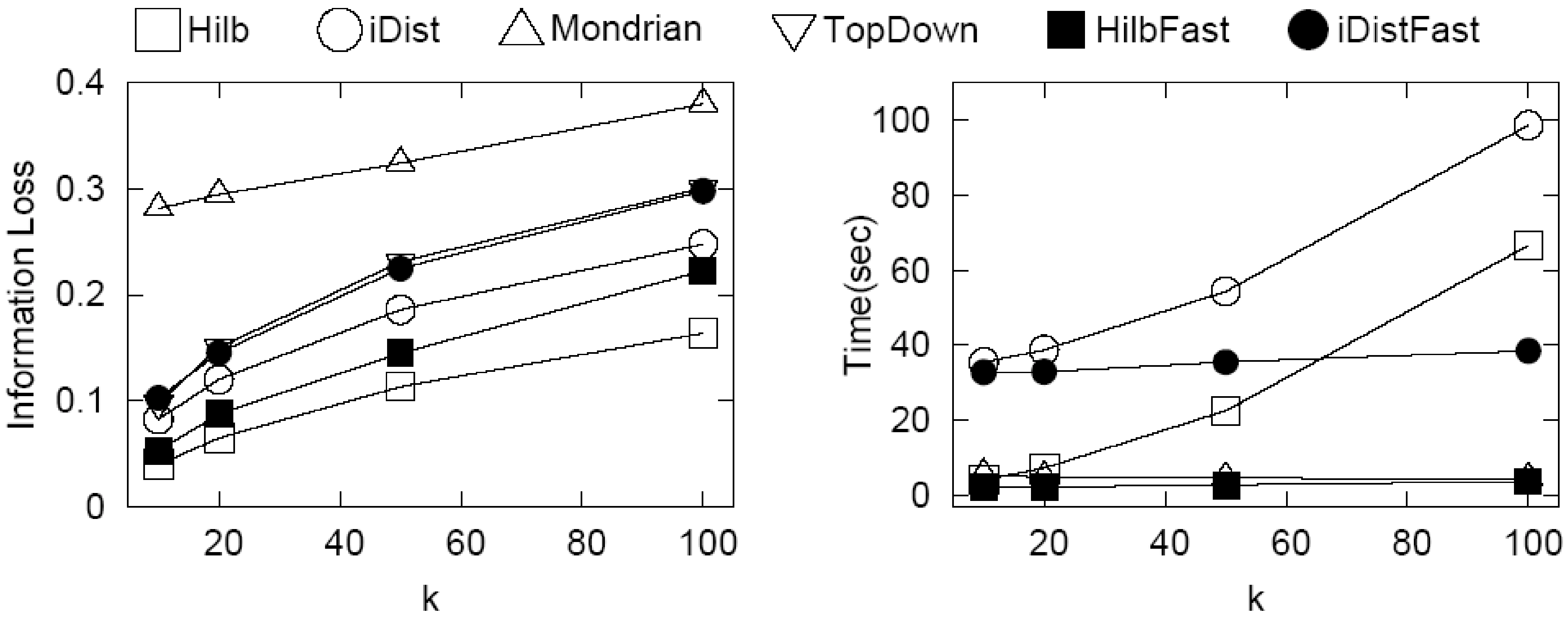
- □ use "frontier" of search
- □ check "eligibility condition" (for termination)



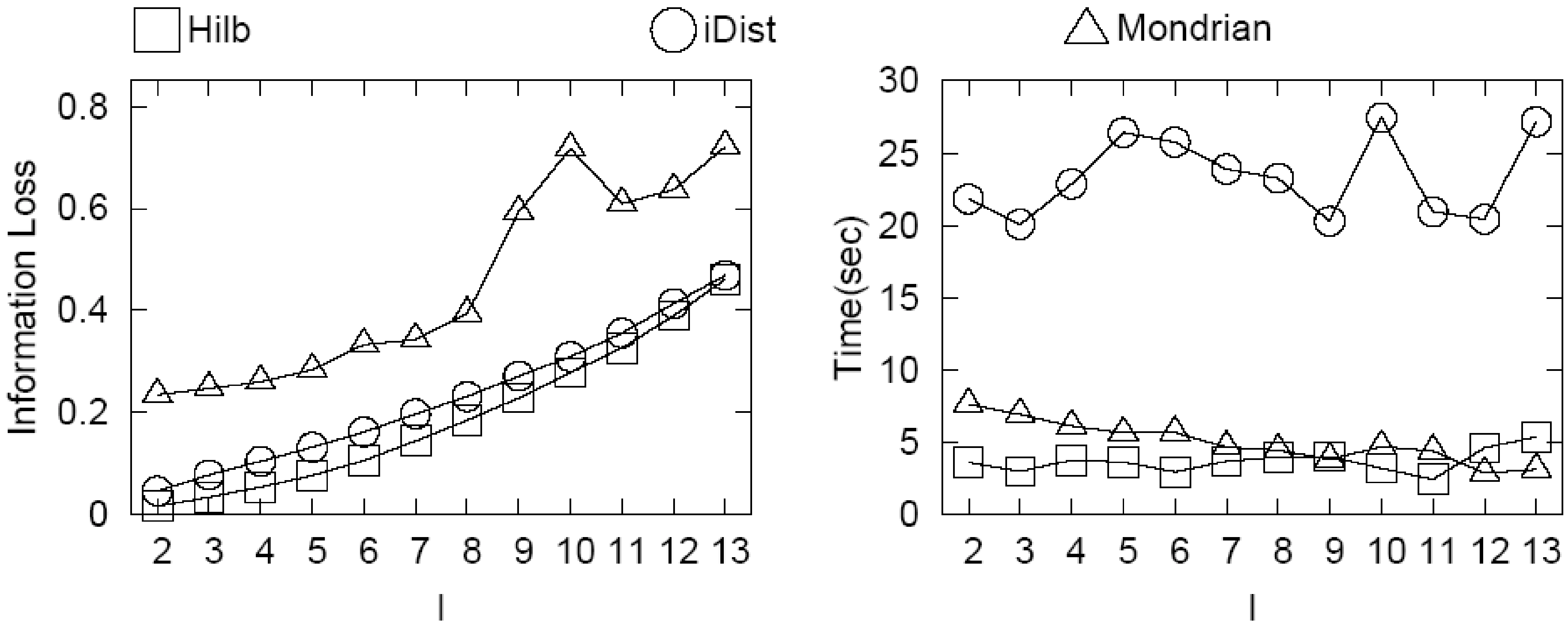
Experimental Setting

- Census dataset
  - Data about 500,000 individuals
- General purpose information loss metric
  - Based on group extent in QID space
- OLAP query accuracy
  - KL-divergence pdf distance

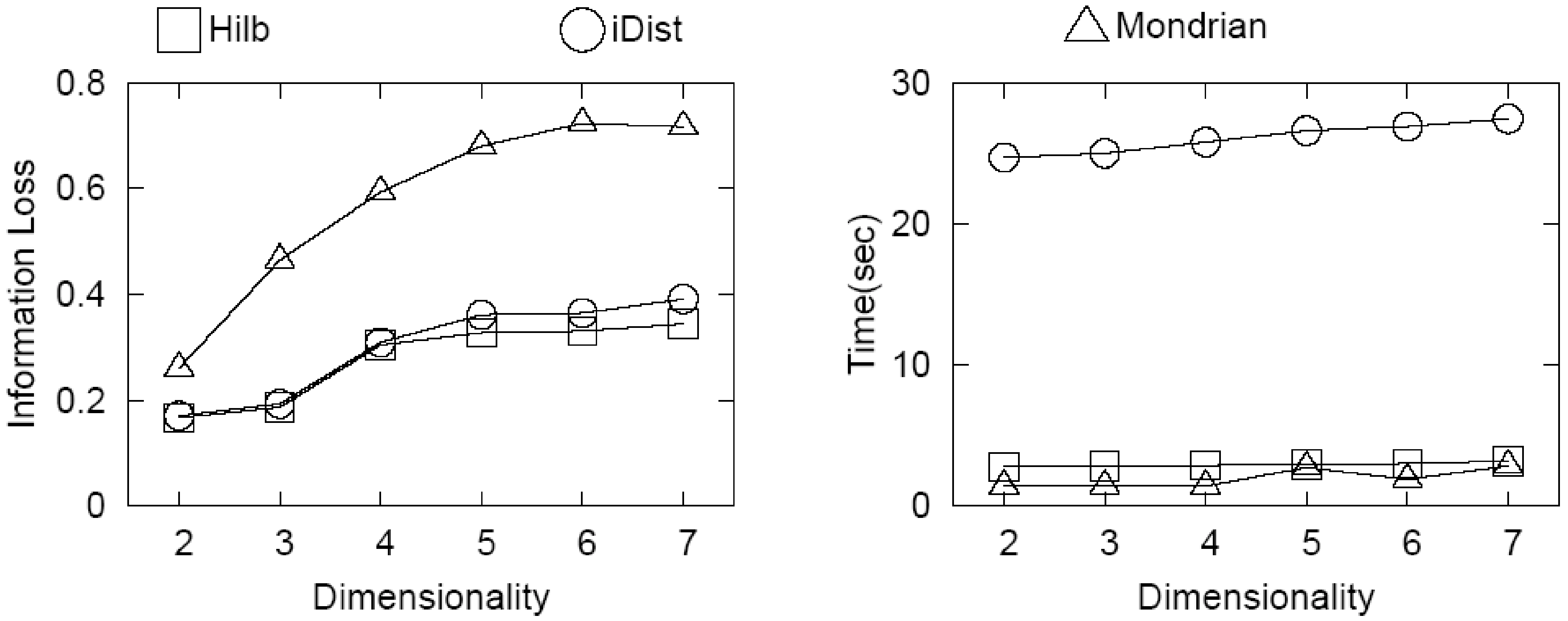
k-anonymity


□ Hilb
○ iDist
△ Mondrian
▽ TopDown
■ HilbFast
● iDistFast
Information Loss
0.4
0.3
0.2
0.1
0
20
40
60
80
100
k
Time(sec)
100
80
60
40
20
0
20
40
60
80
100
k

ℓ-diversity: General Info. Loss



ℓ-diversity: General Info. Loss

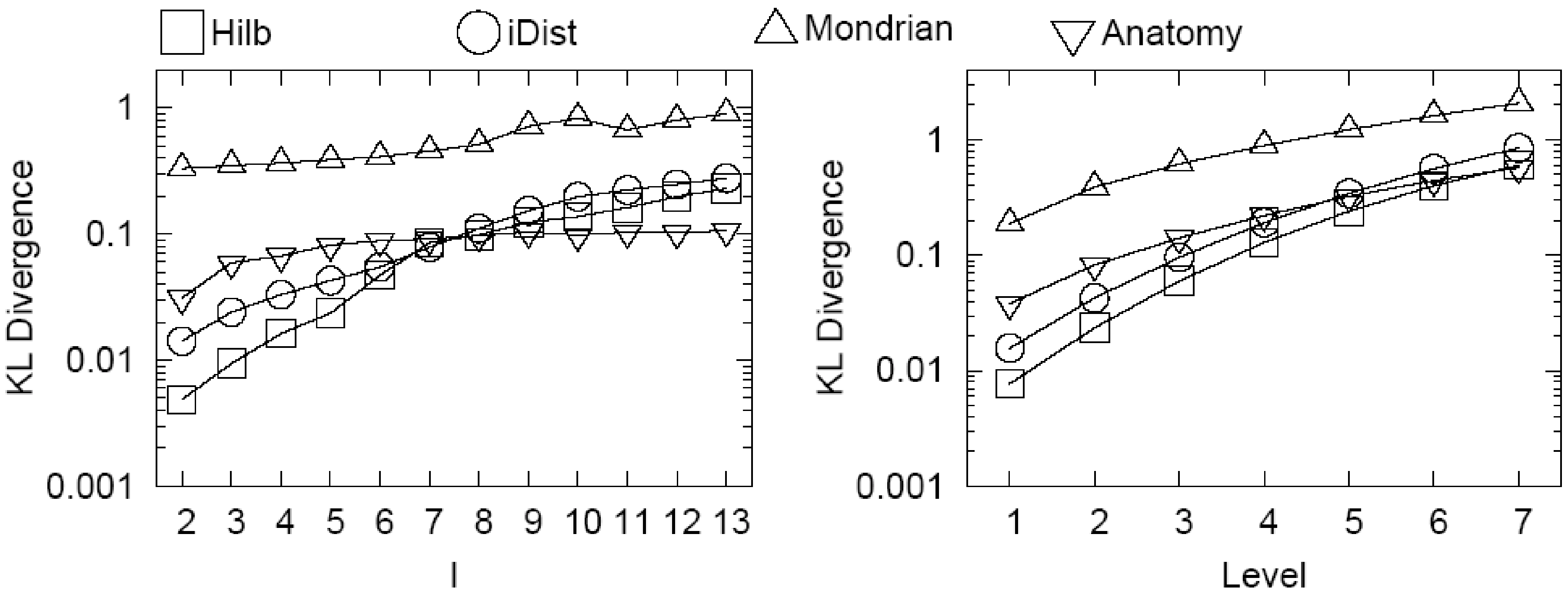


OLAP Queries

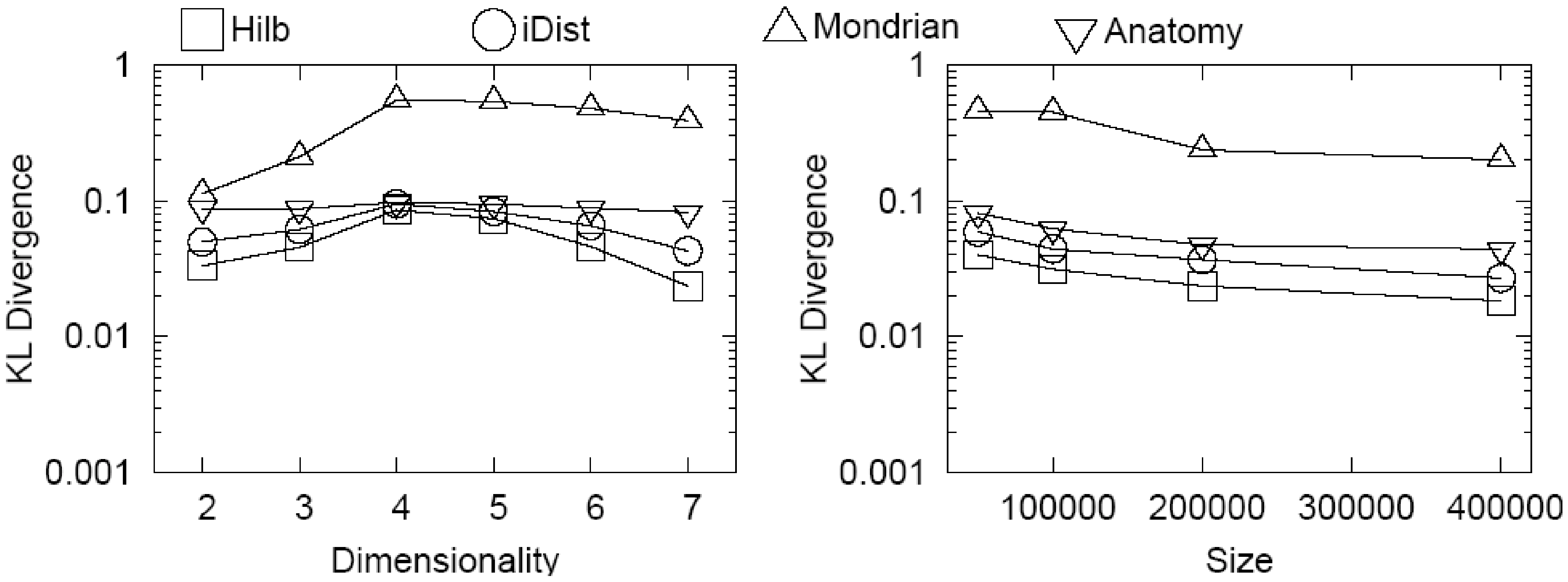
- Distance between actual and approximate OLAP cubes

```
SELECT QT1, QT2,..., QTi, COUNT(*)
FROM Data
WHERE SA = val
GROUP BY QT1, QT2,..., QTi
```

OLAP Query Accuracy



OLAP Query Accuracy



Conclusions

- Framework for *k*-anonymity and ℓ-diversity
  - Transform the multi-D QID problem to 1-D
  - Apply linear optimal/heuristic 1D algorithms

- Results
  - Clearly superior utility to Mondrian, with comparable execution time
  - Similar (or better) utility as Anatomy for aggregate queries, where Anatomy excels