

# Dealing with Semantic Heterogeneity by Generalization- Based Data Mining Techniques

Jiawei Han, Raymond T. Ng,  
Yongjian Fu, Son K. Dao

Thomas Chen

11/12/98

# Outline

- Problem Statement
- Solution
  - Construction of MLDB
  - Operations of MLDB
- Conclusion

# Problem Statement

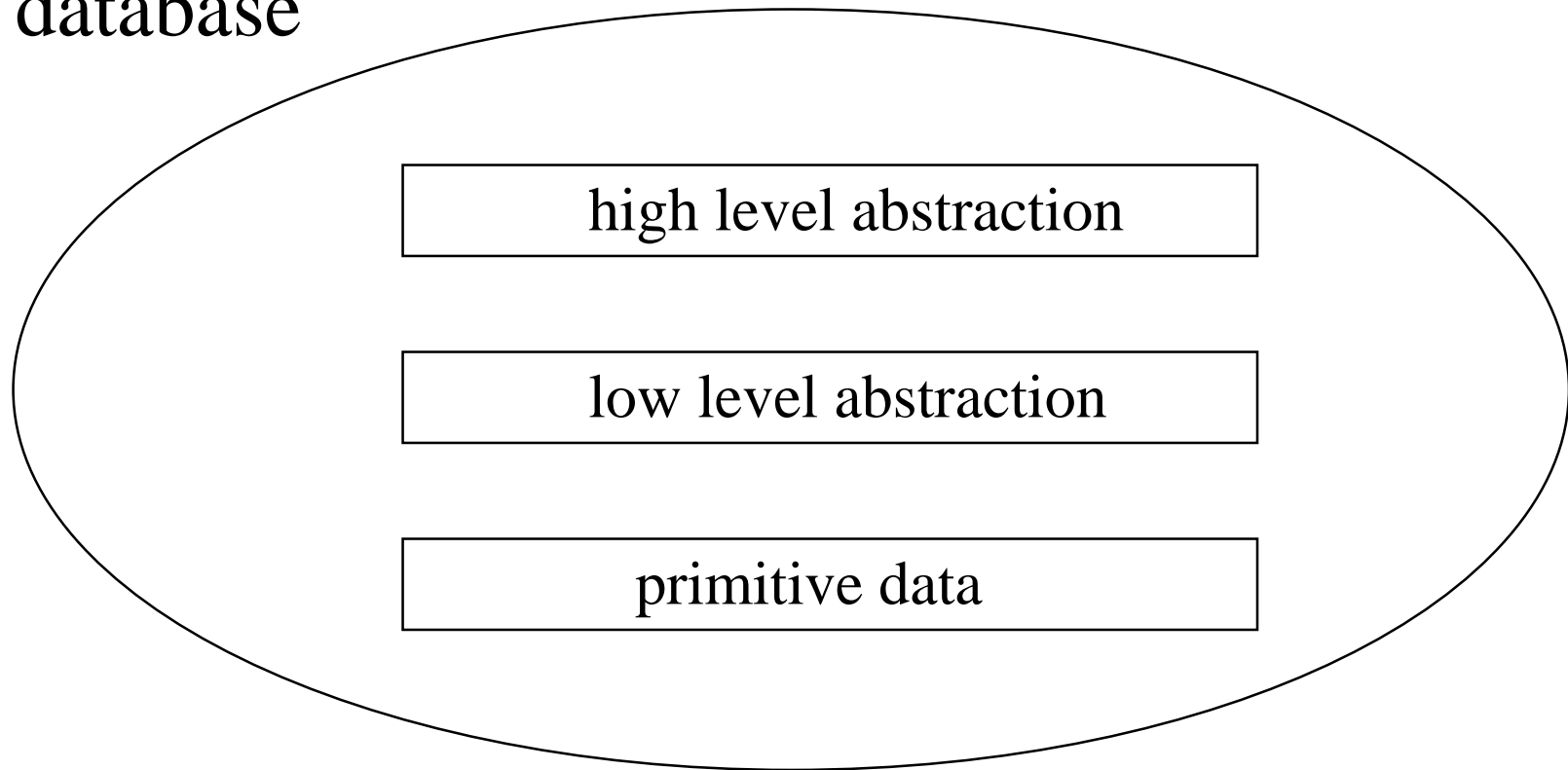
- How to deal with semantic heterogeneity presented by the multiple autonomous database?
  - schema-level analysis
  - data-level analysis

# Example

- Schema-level Analysis
  - grading (name, id, university, semester, year, course name, grade)  
grading (Peter,1, UT, Fall, 94, CS304p, A)
- Data-level Analysis
  - express grades in terms of top-2%, top one third
  - classify classes according to difficulty  
grading (Peter,1, UT, Fall, 94, intro\_cs, top-2%)

# Multiple Layered Database

database

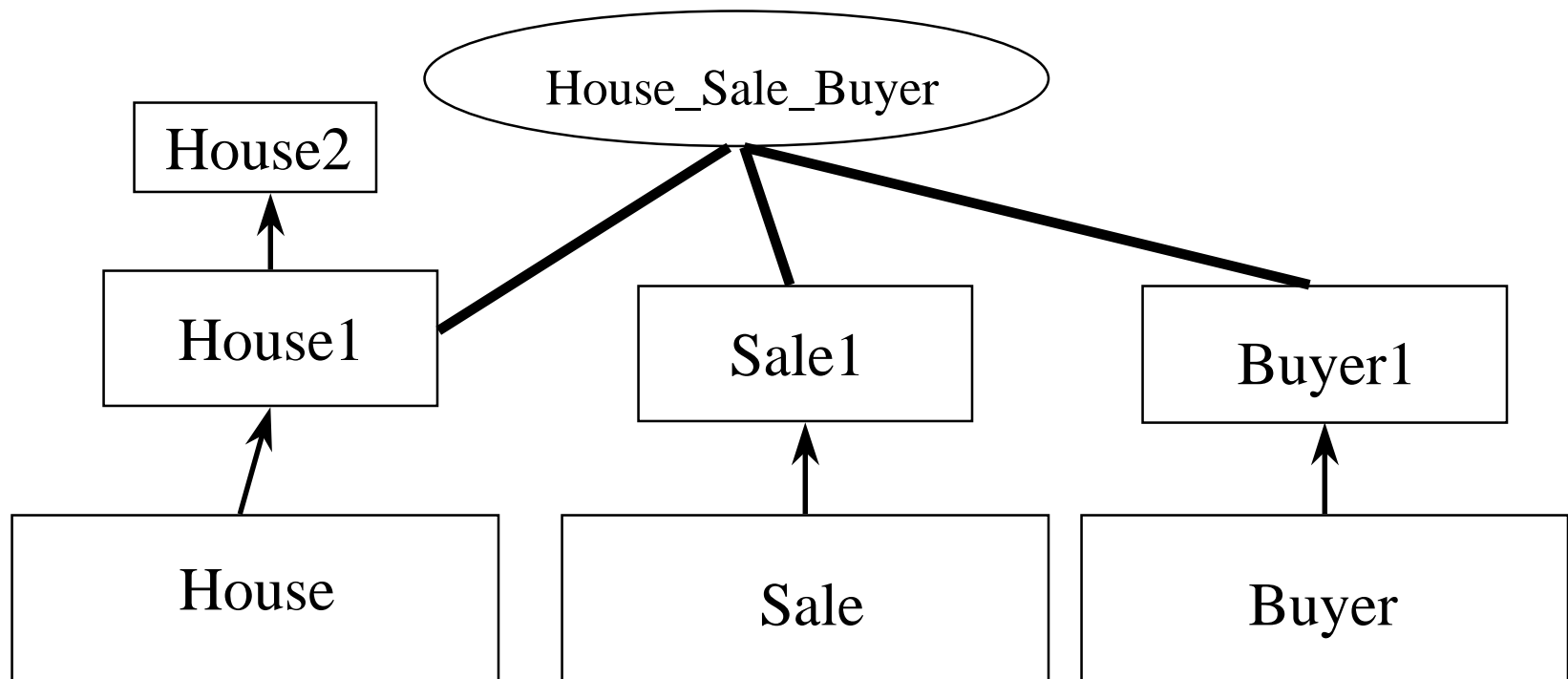


# MLDB

- Components
  - database schema
  - set of concept hierarchies
  - set of integrity constraints
  - set of database relations

# MLDB Example

- Database Schema



# MLDB Example

- Primitive Database Relations
  - house (id, address, construction\_date, constructor, owner, living\_room size, layout, price)
  - buyer (name, id, education, income, address, phone, spouse, children)
  - sales (house, buyer, agent, contract\_date, sell\_price, mortgage)
  - agent (...)

# MLDB Example

- Layer-1 Database Relation
  - house1 (id, address, years\_old, total floor area, num\_room, price)
- Layer-2 Database Relation
  - house2 (area, year\_range, floor\_plan, price\_range, count)

# MLDB Example

- Conceptual Hierarchy

generalization rules that transform data from lower abstraction to higher abstraction

- speedway => central Austin
- \$164,000 => [150K, 200k]

- Types of Data

- non-numerical, numerical, structured, multimedia.

# Generalization Techniques

- non-numerical data
  - domain expert
- numerical data
  - clustering
  - k-means
- structured data
  - summarization

# Generalization Techniques

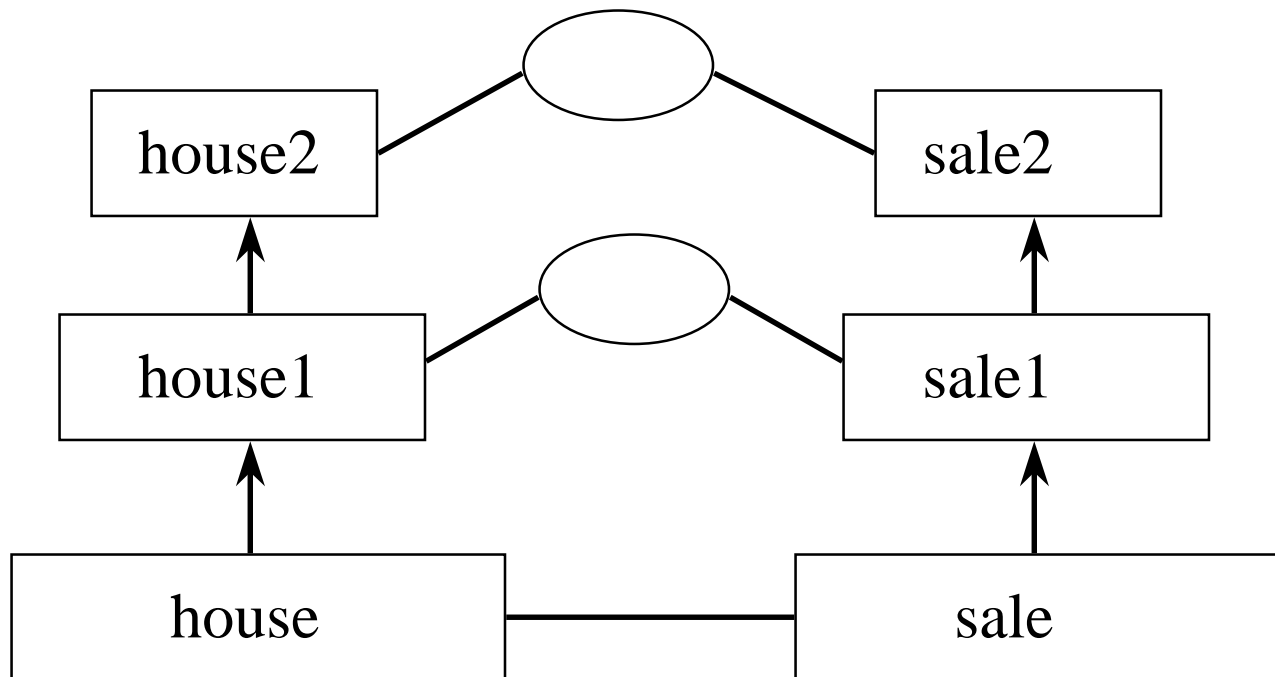
- Key-Preserving/Key-Altering
  - key-preserving: all keys (including foreign key) are preserved
  - key-altering: some keys are generalized, and cannot be used as join attributes.
    - different values maybe generalized to identical values in a higher layer so it might produce false tuples that is not present in the original data set.

# Query Answering

- Direct Query Answering
  - follow the query strictly without providing extra associative information
- primitive/higher layer
  - attribute expands several conceptual levels
  - transformation
  - find the highest joinable layer

# Direct Query Answering

- find the houses locate in north Austin, 3-bedroom and Sold in summer of 1995



# Query Answering

- Cooperative Query Answering
  - query provides associative information
    - What kind of house can be bought with \$300K in Vancouver area?
  - relax query condition: around \$300K
  - generalize answers with summarized data
    - 20% 20-30 years-old, medium-size 3-bedroom
  - comparison with neighborhood answers
    - 10% 20-30 yrs \$250-350K, 30% \$350-500K
  - progressive information focusing

# Query Answering