Enhancing Traditional Databases to Support Broader Data Management Applications

Yi Chen
Computer Science & Engineering
Arizona State University

What Is a Database System?

- Of course, there are traditional relational database management systems (RDBMS)
- Was introduced in 1970 by Dr. E. F. Codd (of IBM)
- Commercial relational databases began to appear in the 1980s
- The focus of most work in the past 30 years

A Relational Database (RDBMS)

Table (Relation)

Column (Attribute)

Row
(Tuple, Record)

Refers to

Climber

Name	Skill	age
James	Beginner	21
Bob	Experienced	33

Climbs

Name	Route	Date	Duration
Bob	Last Tango	10/10/05	5
Bob	Last Tango	1/10/06	4.5

A predefined data structure (schema) is required.

Querying RDBMS: SQL

Climber

Name	Skill	age
James	Beginner	21
Bob	Experienced	33

selection: $\sigma_{Name = \text{"James"}}$

Climbs

Name	Route	Date	Duration
Bob	Last Tango	10/10/05	5
Bob	Last Tango	1/10/06	4.5

projection: $\prod_{Route = \text{"Last Tango"}}$

join: Climber $\bowtie_{Climber.name = climbs.name}$ Climbs

Name	Skill	Age	Route	Date	Duration
Bob	Experienced	33	Last Tango	10/10/05	5
Bob	Experienced	33	Last Tango	1/10/06	4.5

The Advantages of RDBMS

- Good data organization
- High efficiency for large datasets via indexing and query optimization
- Concurrency control and reliability

But, 80% of the World's Data is Not in RDBMS!

Examples:

- WWW, Emails
- Personal data, documents of various format
- Sensor data
- A lot of scientific data (experimental data, large images, documentation, etc)

- Why not?
- There are several assumptions in relational databases that do not fit for handling this data.

- My research addresses how to enhance RDBMS to manage them.

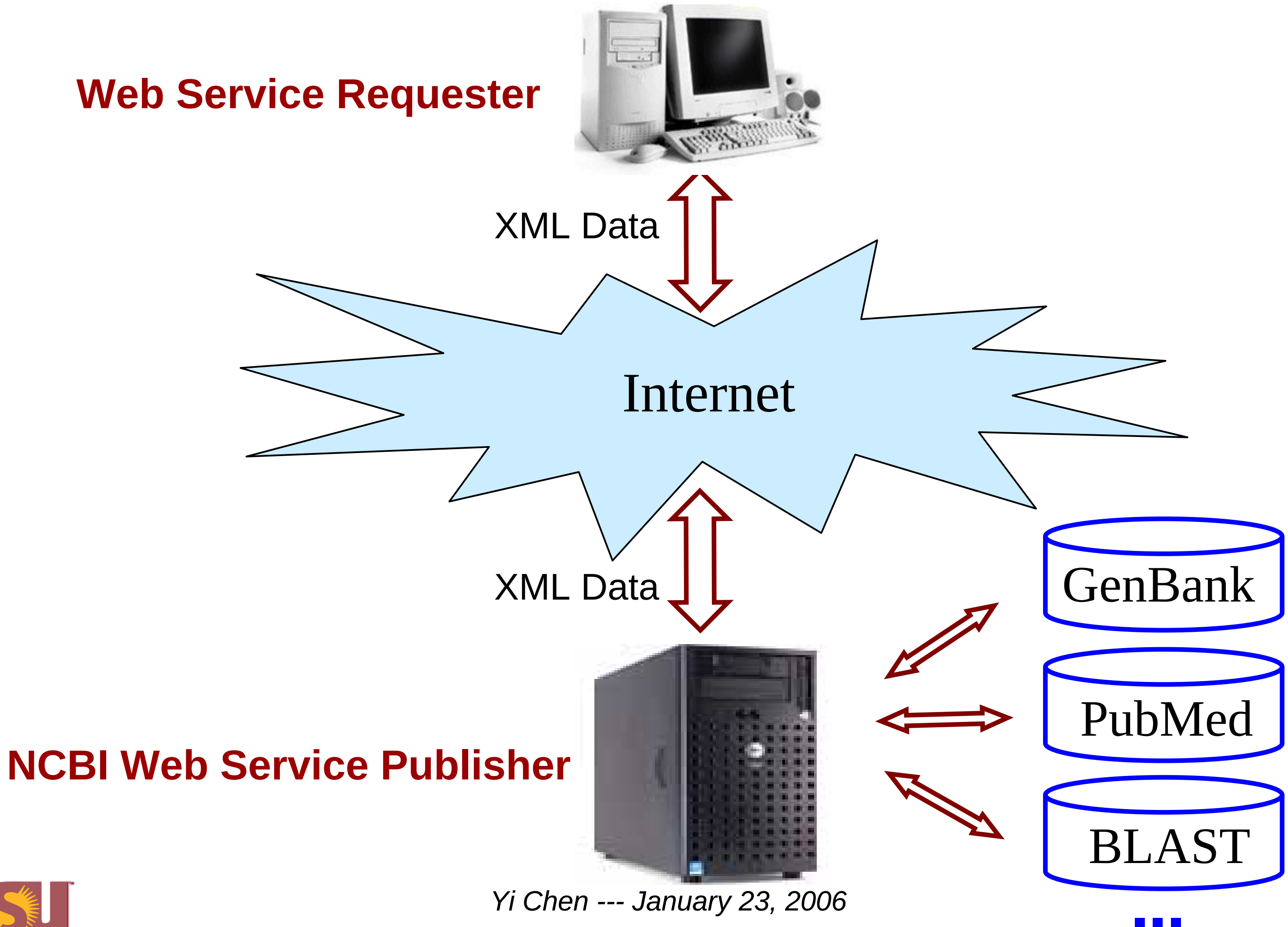
Challenges for RDBMS (I)

- RDBMS Assumption: data conforms to a predefined fixed schema, which is separated from the data itself
- Reality:
  - Data may be collected from different sources on the web, therefore has different schemas
  - Schema can change over time for a single source
- Requirements: We need to handle data of different schemas and have the schemas tightly associated with the data

XML as a Data Representation Format

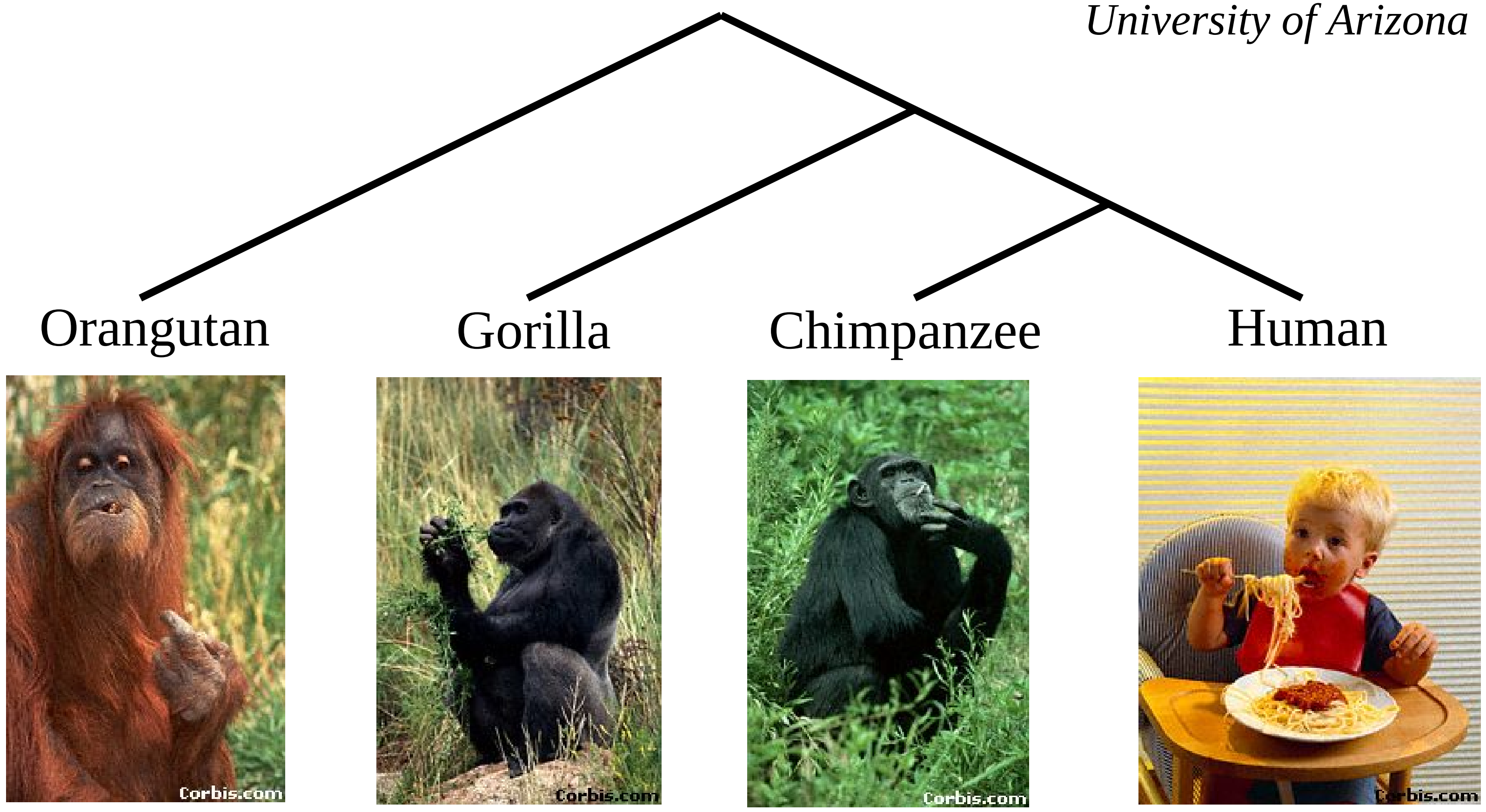
- XML has become a standard data format for various applications, because of:
  - Flexibility in schemas -- semi-structured data
  - Self - describing feature
  - Representing tree data model naturally

XML: the Standard for Web Data Representation



XML: Representing Phylogenetic Trees

From the Tree of the Life Website,
University of Arizona



Challenges for RDBMS (II)

- RDBMS Assumption: Data is clean and consistent.
- Reality: real world data is dirty
  - Data collected from different sources may have missing and conflicting information
  - Data that is obtained from data mining is often not error-prone
  - Experimental data often contains random errors
- Requirements: we need to measure data quality and handle imprecise and/or incomplete data

Roadmap of This Talk

- Managing XML by leveraging mature RDBMS [Chen et al 04]
  - Introduction to XML
  - A generic and efficient XML-to-RDBMS mapping
    - Data mapping from trees to tables
    - Query translation from tree navigation queries to SQL queries that are efficient
- Handling imprecise and incomplete data in DBMS [Chen et al 06]

Sample XML Data

<books>

<book>

<title>

The lord of the rings...

</title>

<section>

<title>

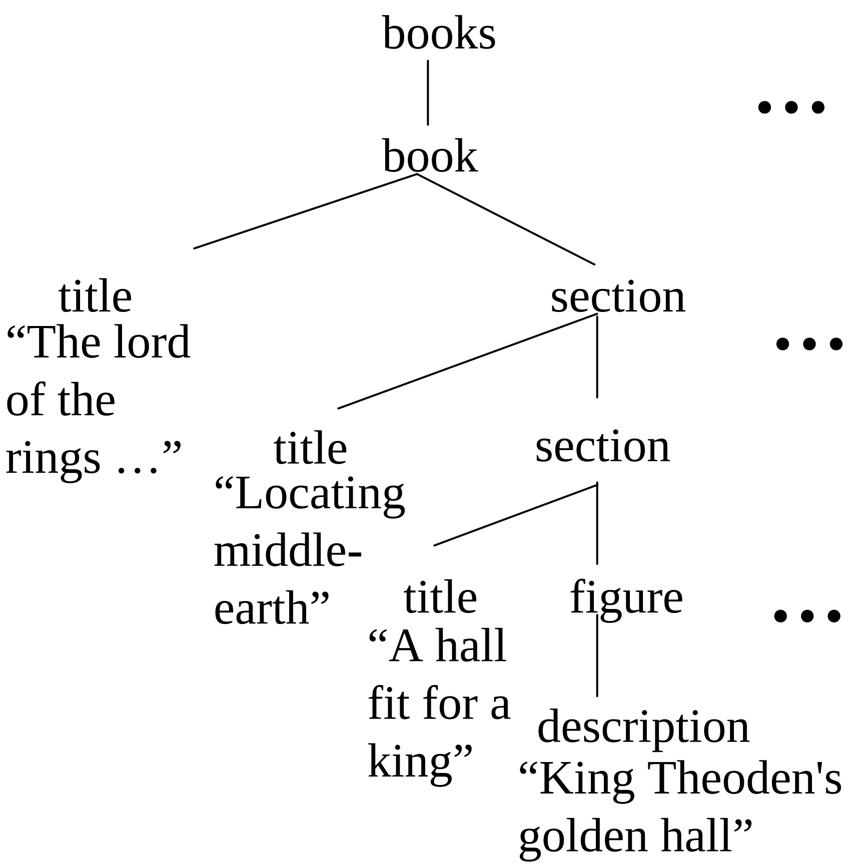
Locating middle-earth

</title> ...

</section> …

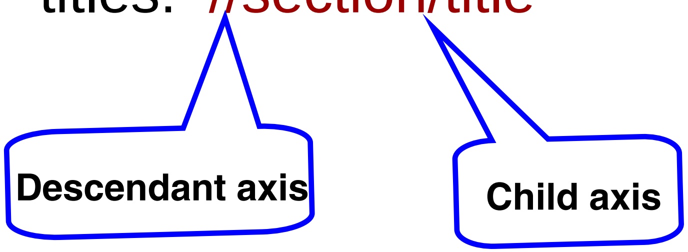
</book>

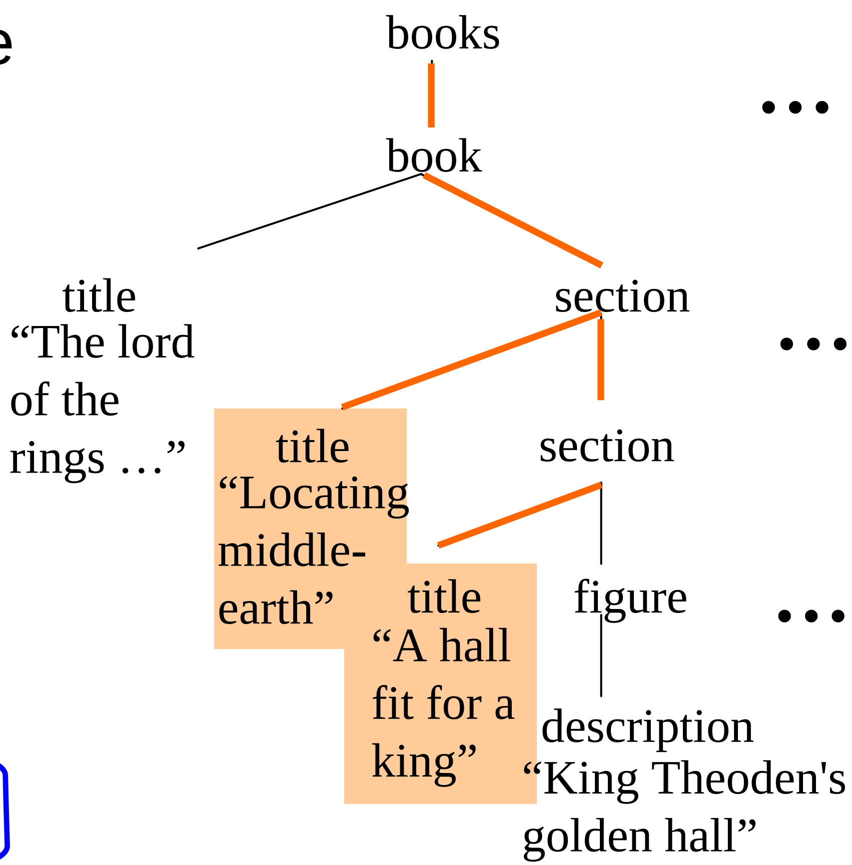
</books>



Sample XML Queries

- XML query languages are based on hierarchical structure navigation (e.g. XPath)
- Sample queries:
  - What are all the section titles: //section/title

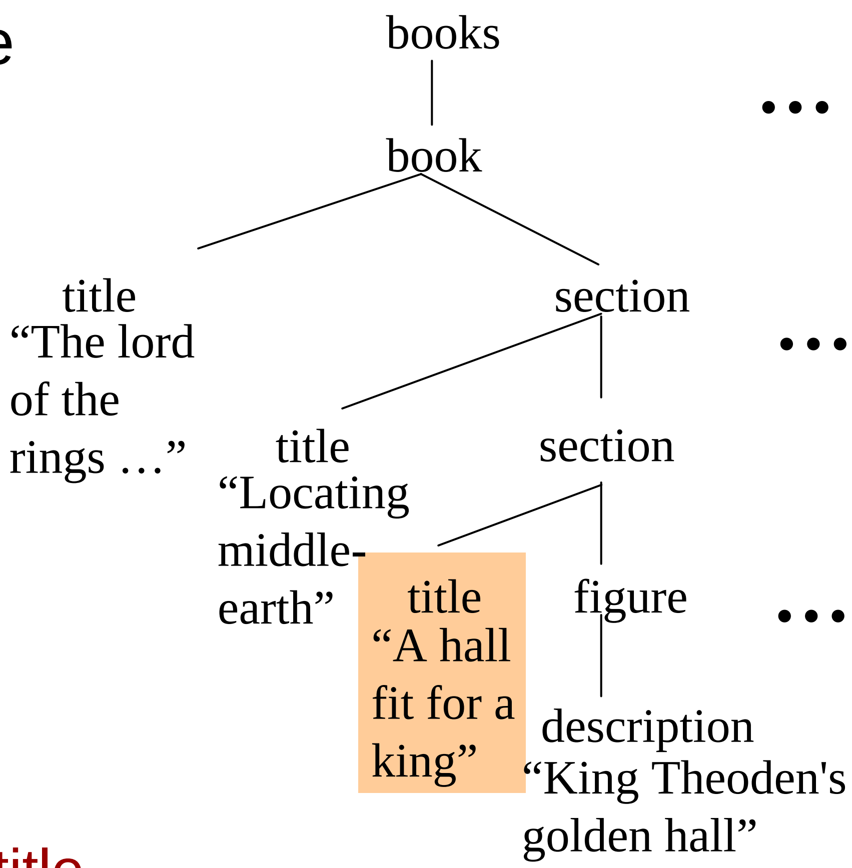




Sample XML Queries

- XML query languages are based on hierarchical structure navigation (e.g. XPath)
- Sample queries:
  - What are all the section titles: //section/title
  - What are the titles of sections that contain a figure: //section[/figure]/title

Predicates



How to Query XML Data efficiently?

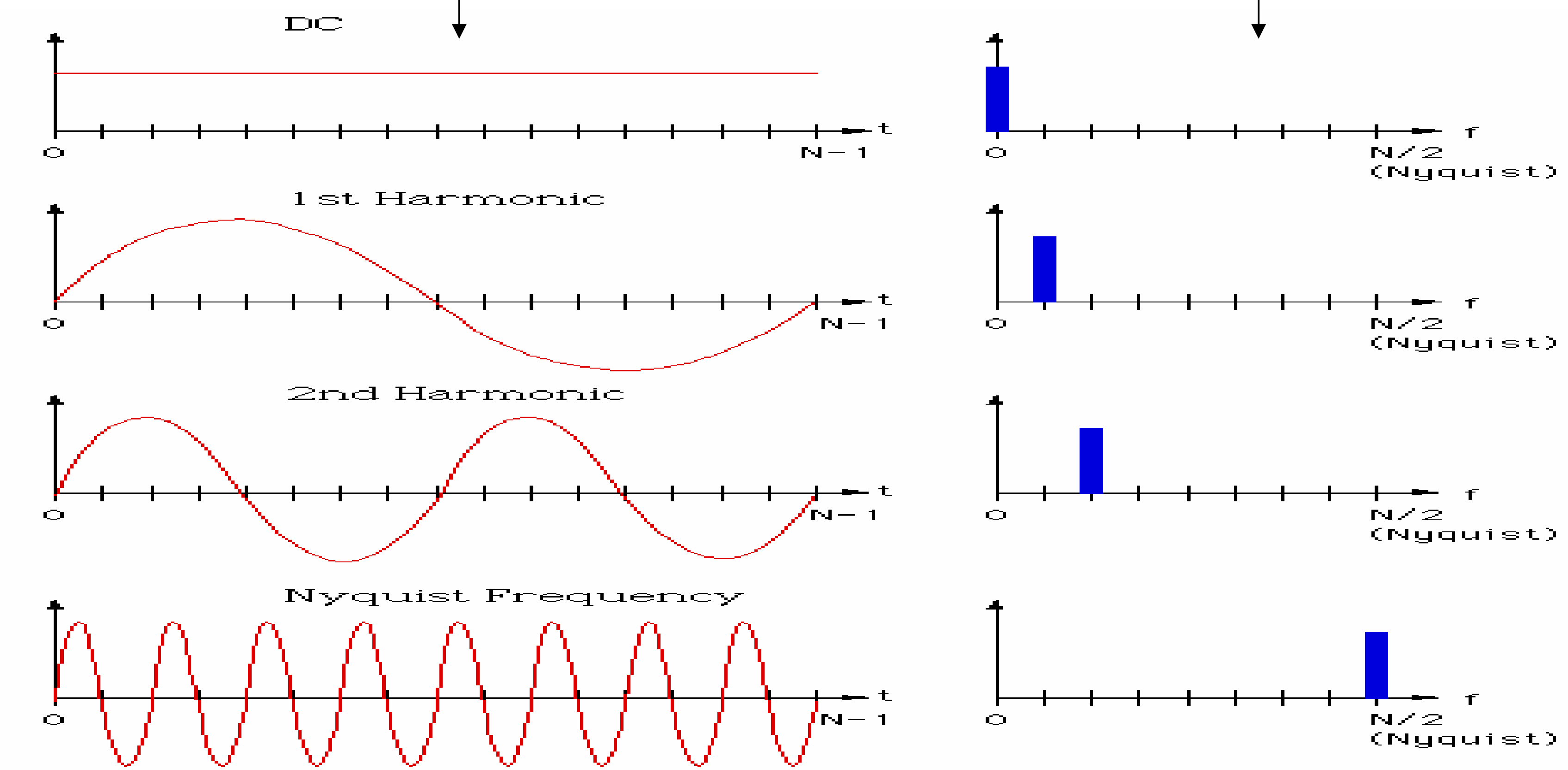
- RDBMS have achieved high performance in query evaluation.
- Can we leverage RDBMS by encoding XML to tables?

Analogy: Fourier Transforms

Complex | **Efficient**

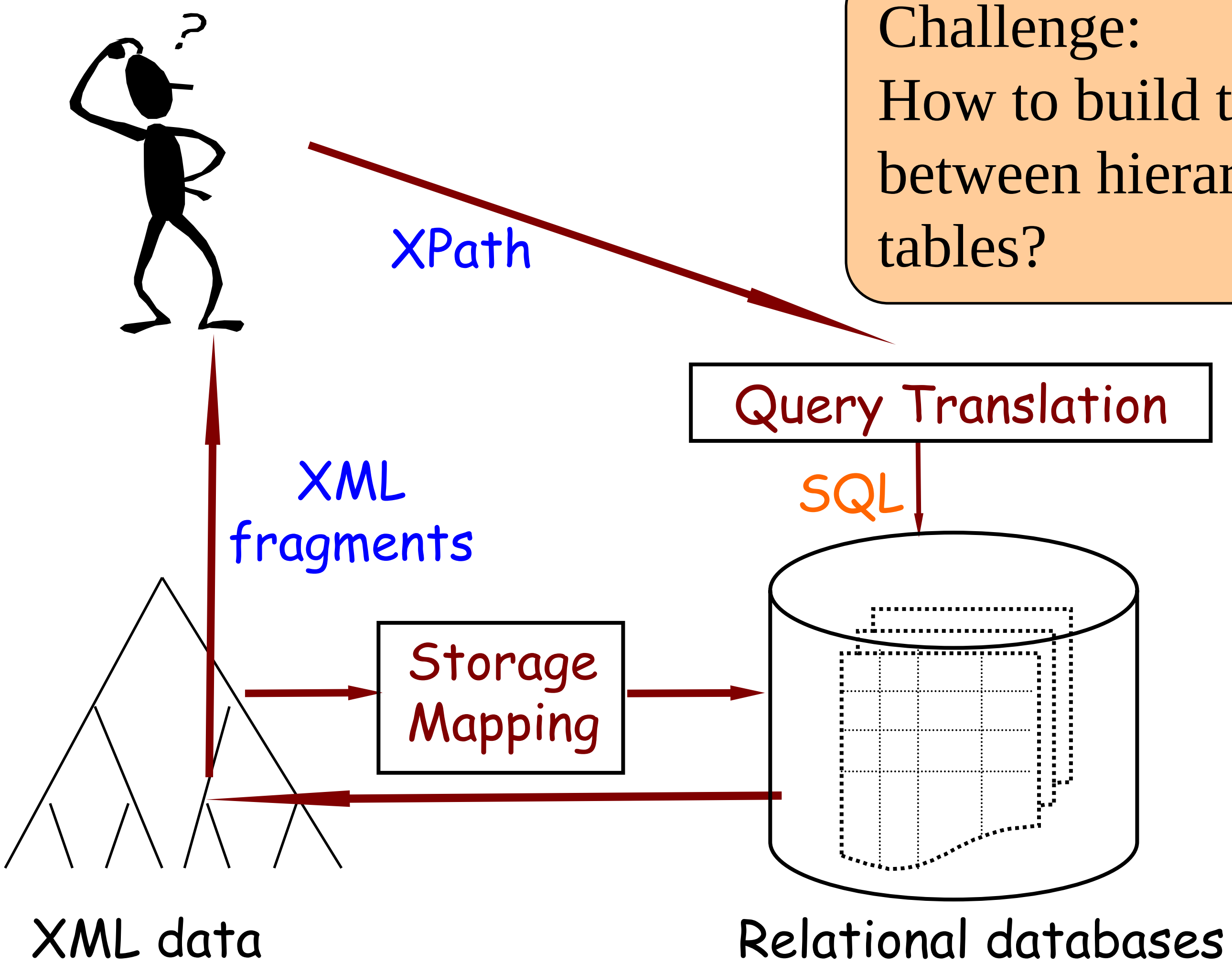
$$g * h = \iint_{-\infty}^{+\infty} g(u)h(u)du$$

$$G(f)H(f)$$

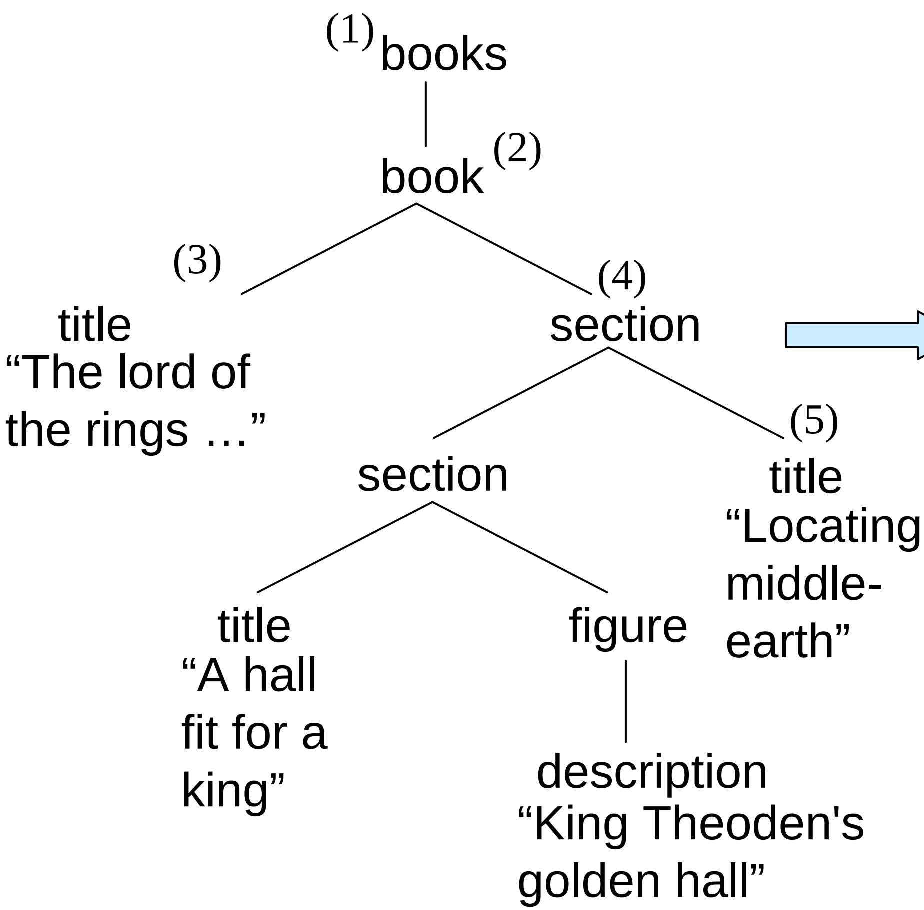


Mapping XML Data to RDBMS

Challenge:
How to build the bridge between hierarchies and tables?



Data Mapping

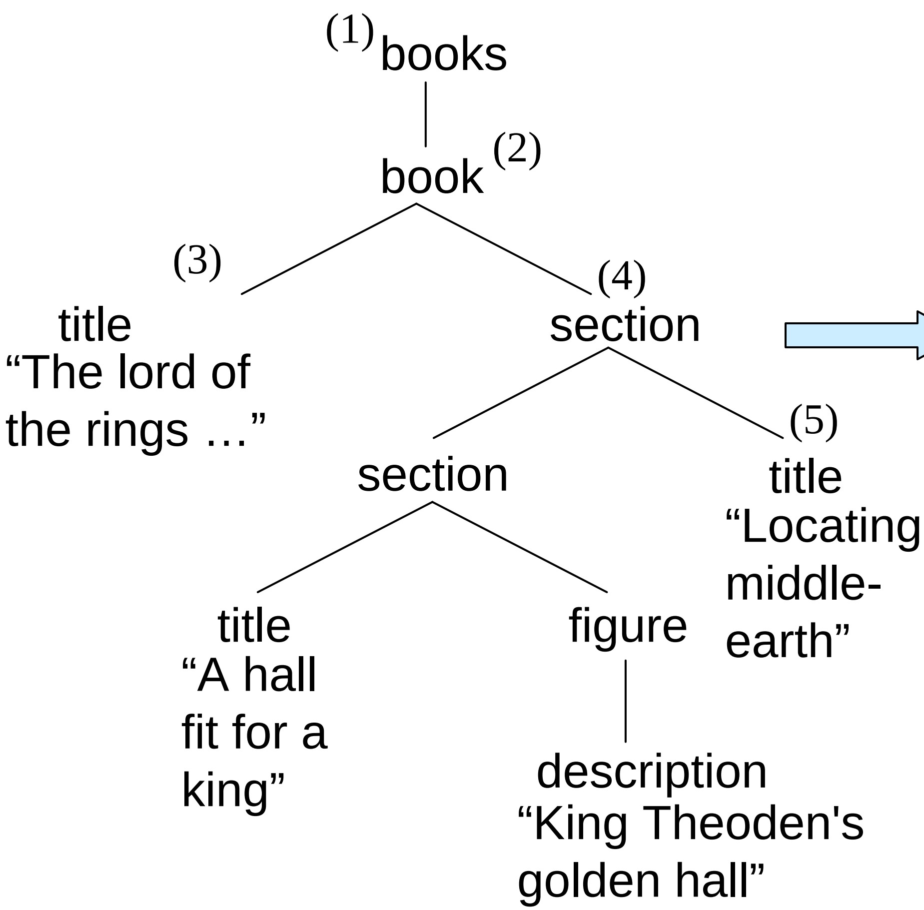


Parent ID
[Florescu & Kossmann 99]

T

ID	Tag	Value	Structural Information
1	books		
2	book		
3	title	The...	
4	section		
5	title	Locating…	
…	…	…	…

Data Mapping

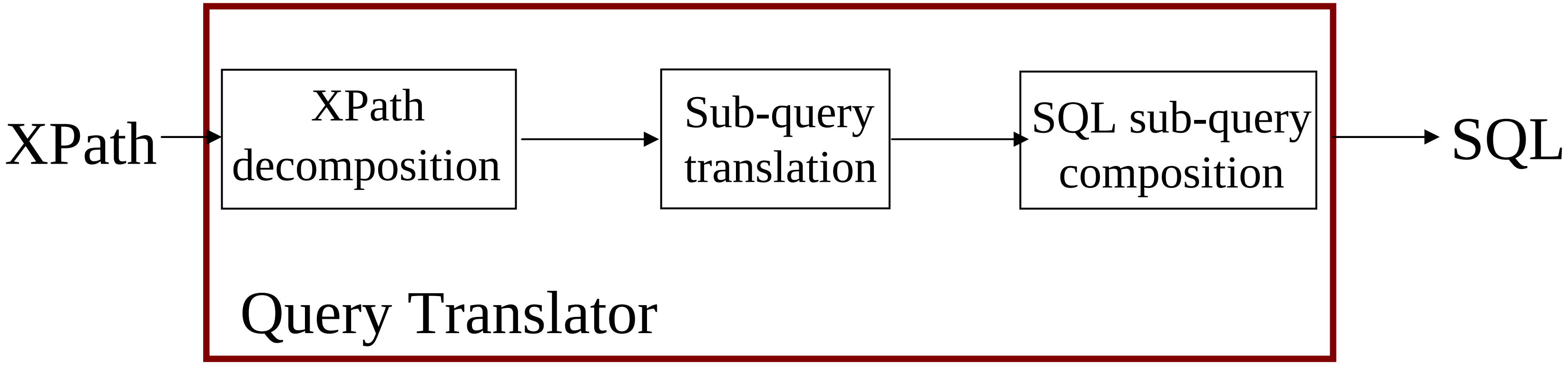


Design special labels to encode node relationships

T

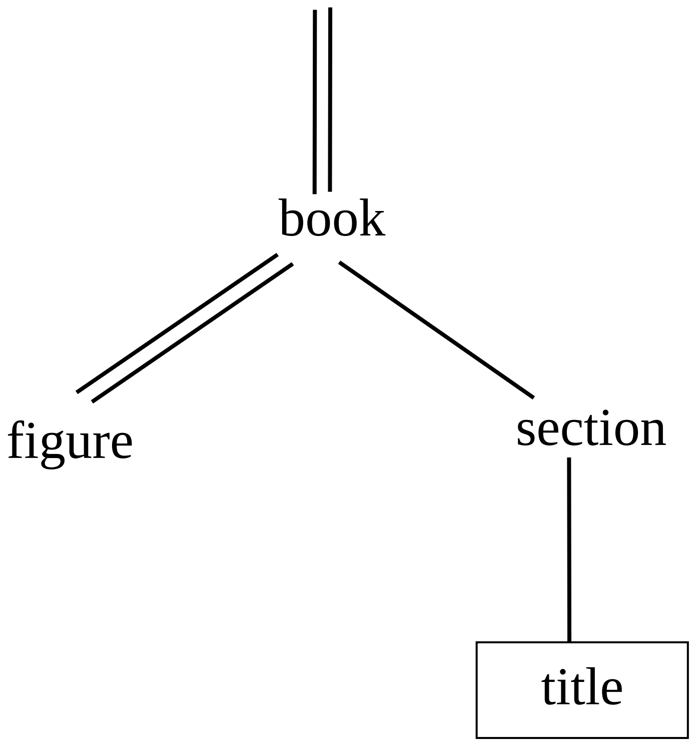
ID	Tag	Value	Structural Information
1	books		
2	book		
3	title	The...	
4	section		
5	title	Locating…	
…	…	…	…

Query Translator Architecture



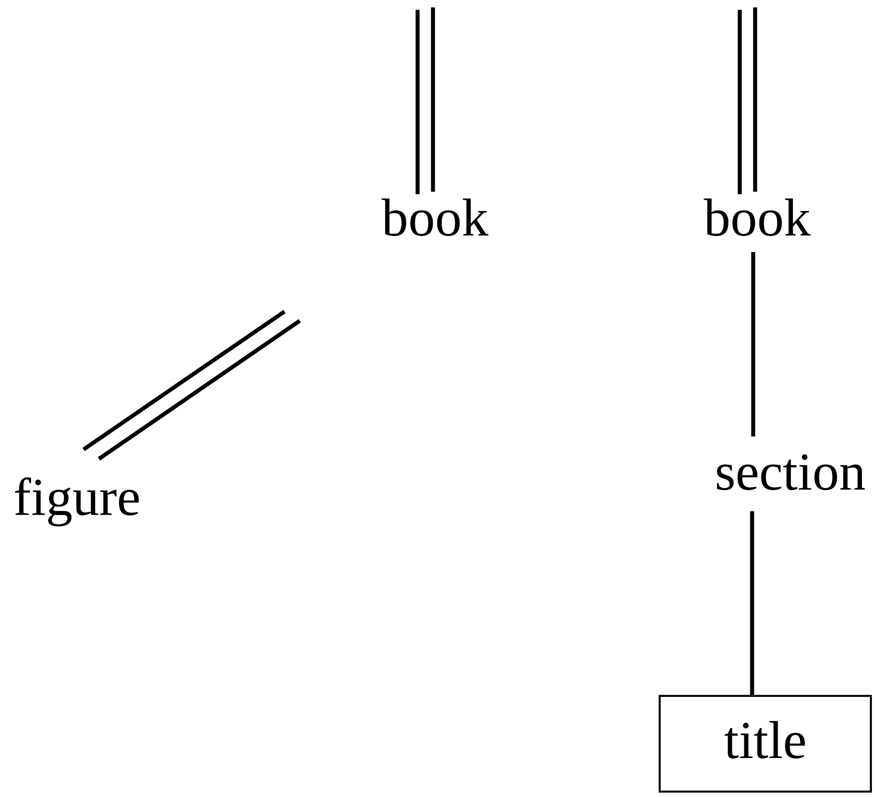
- How to choose XPath subqueries, such that:
  - they can be easily translated to SQL subqueries
  - the SQL subqueries can be efficiently evaluated
- How to combine SQL subqueries to a complete one?

Query Translator



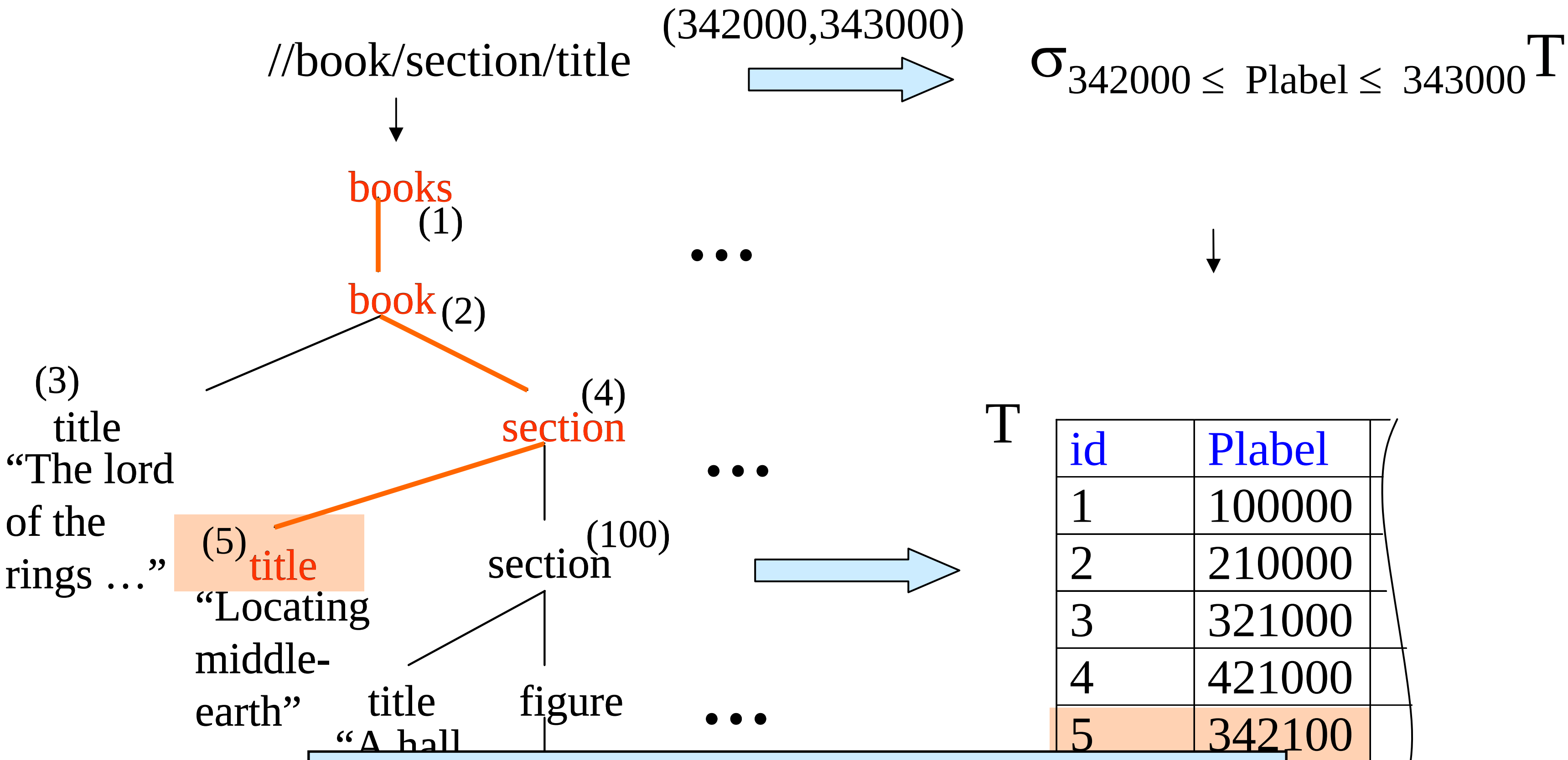
Q: //book[//figure]/section/title

Query Translator: (I) Decomposition to Suffix Paths



Q: //book[//figure]/section/title

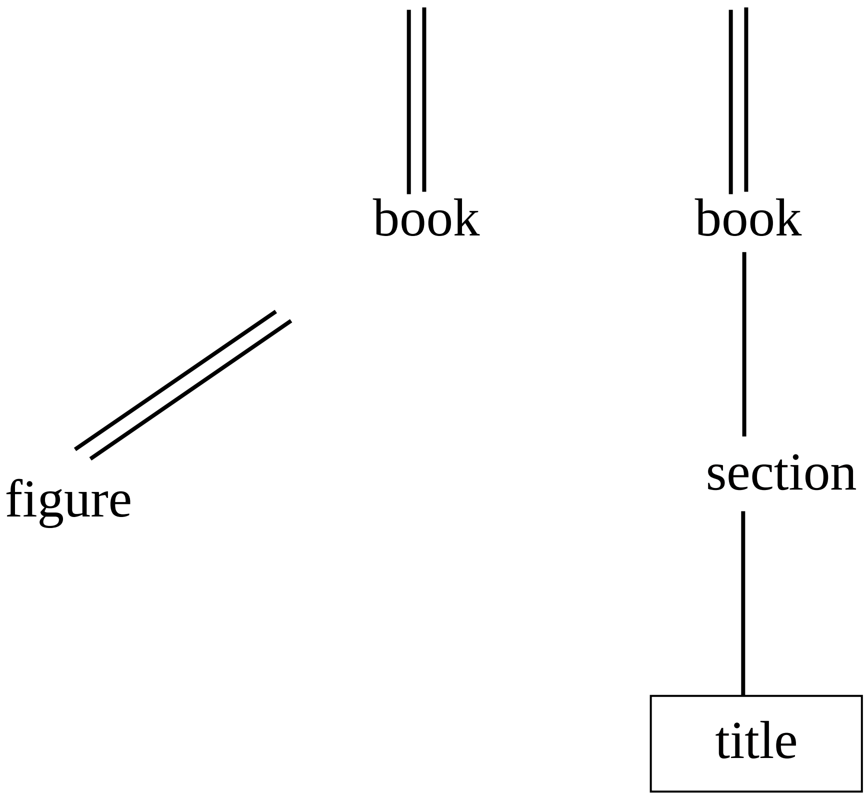
Encoding Suffix Paths Using P-labeling



id	Plabel
1	100000
2	210000
3	321000
4	421000
5	342100

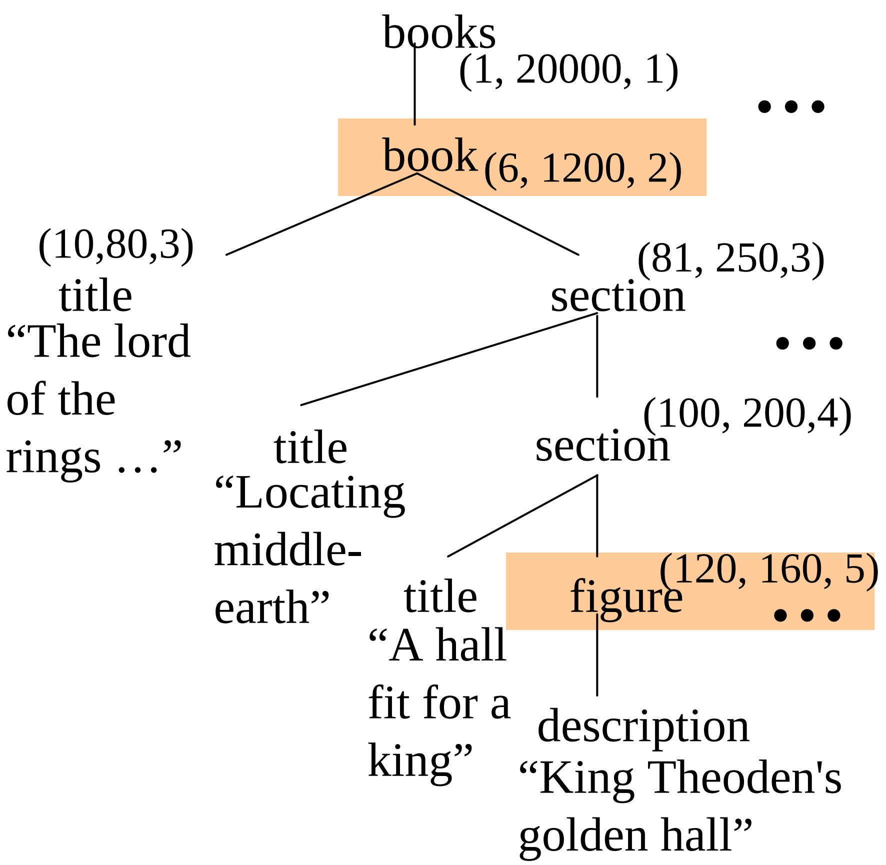
Evaluating suffix paths → SQL selections on P-labels

Query Translator: (II) Selection on P-labels



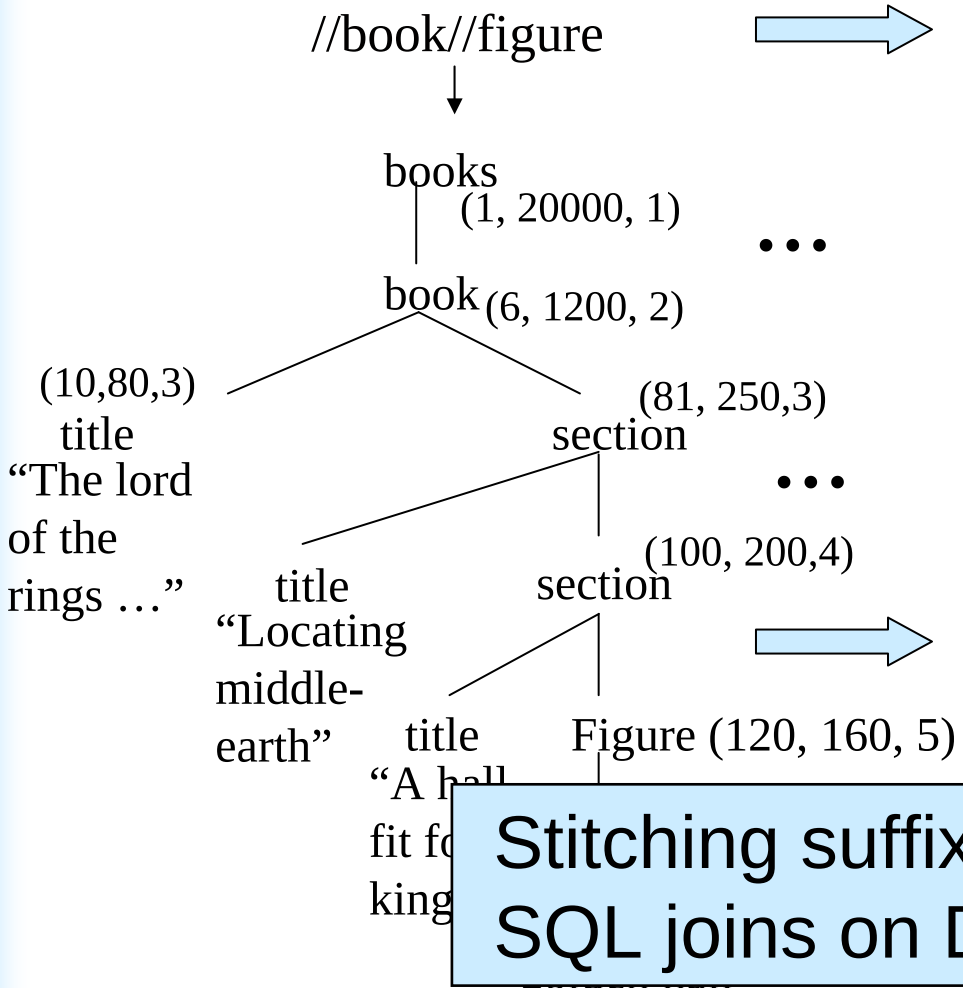
Q: //book[//figure]/section/title

D-labeling Scheme



- D-labeling is used to connect suffix paths.
- D-labels (start, end, depth) can be used to detect ancestor-descendant relationships between nodes in a tree.

A Bi-labeling Based Query Translation



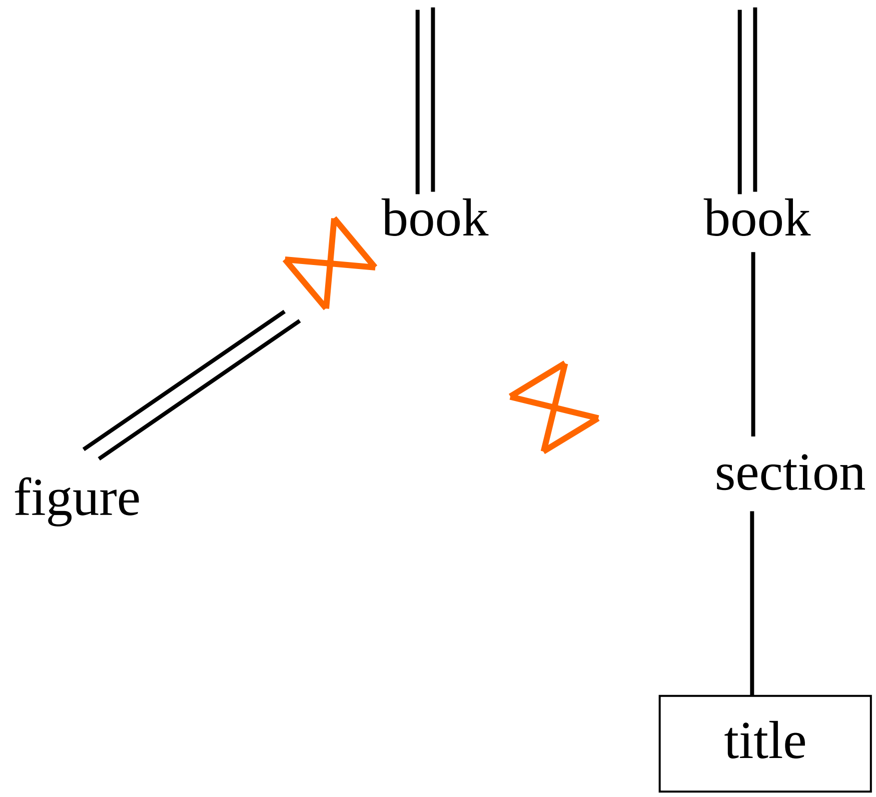
$$(\sigma_{\text{Plabel for “//book”}}(\rho_{T \to T'})) \bowtie_{T'.Start \le T.Start \wedge T'.End \ge T.End} (\sigma_{\text{Plabel for “//figure”}} T)$$

T

Plabel	Start	End	Depth	Val
100000	1	20000	1	…
210000	6	1200	2	…
321000	10	80	3	…
421000	81	250	3	…
442100	100	200	4	…
[illegible]	[illegible]	[illegible]	5	…
[illegible]	[illegible]	[illegible]	…	…

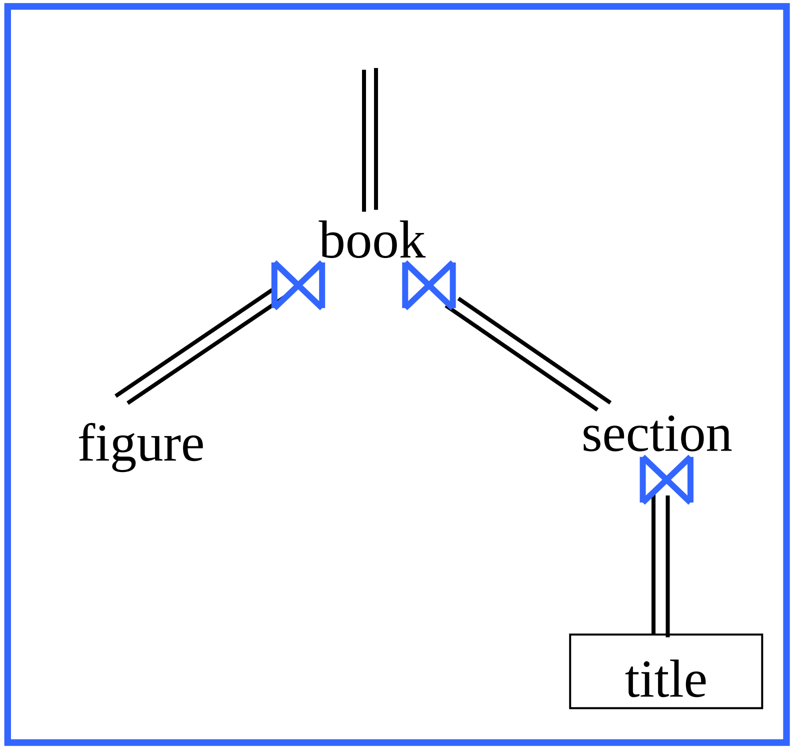
Stitching suffix paths → SQL joins on D-labels

Query Translator: (III) Join on D-labels

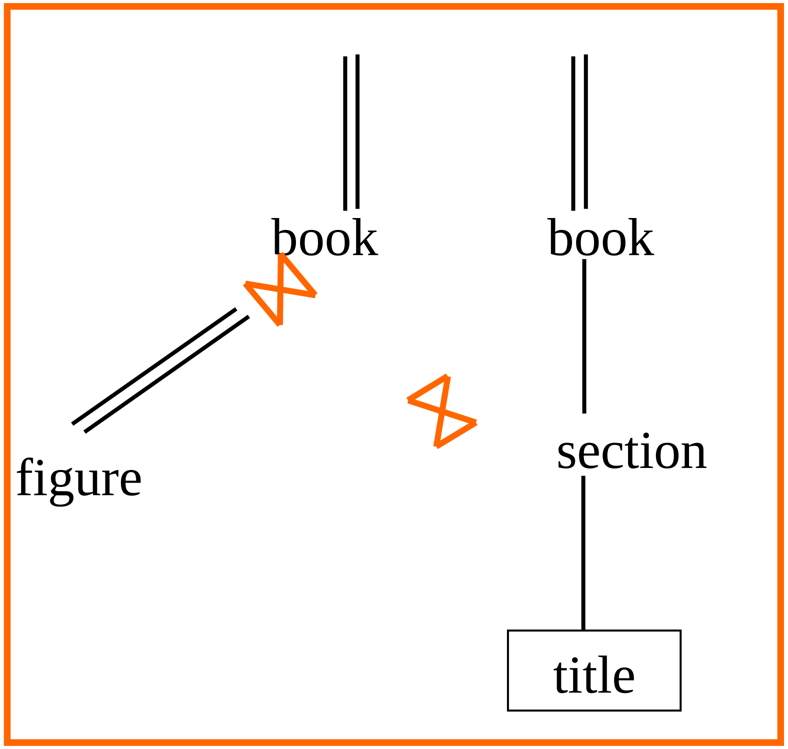


Q: //book[//figure]/section/title

Comparison with Previous Approach



Previous Approach
[Li & Moon 01, Zhang et al 01, Tatarinov et al 02, Grust 02, DeHaan et al 03, etc]



Ours:
fewer disk accesses, fewer joins

Experiment Setup

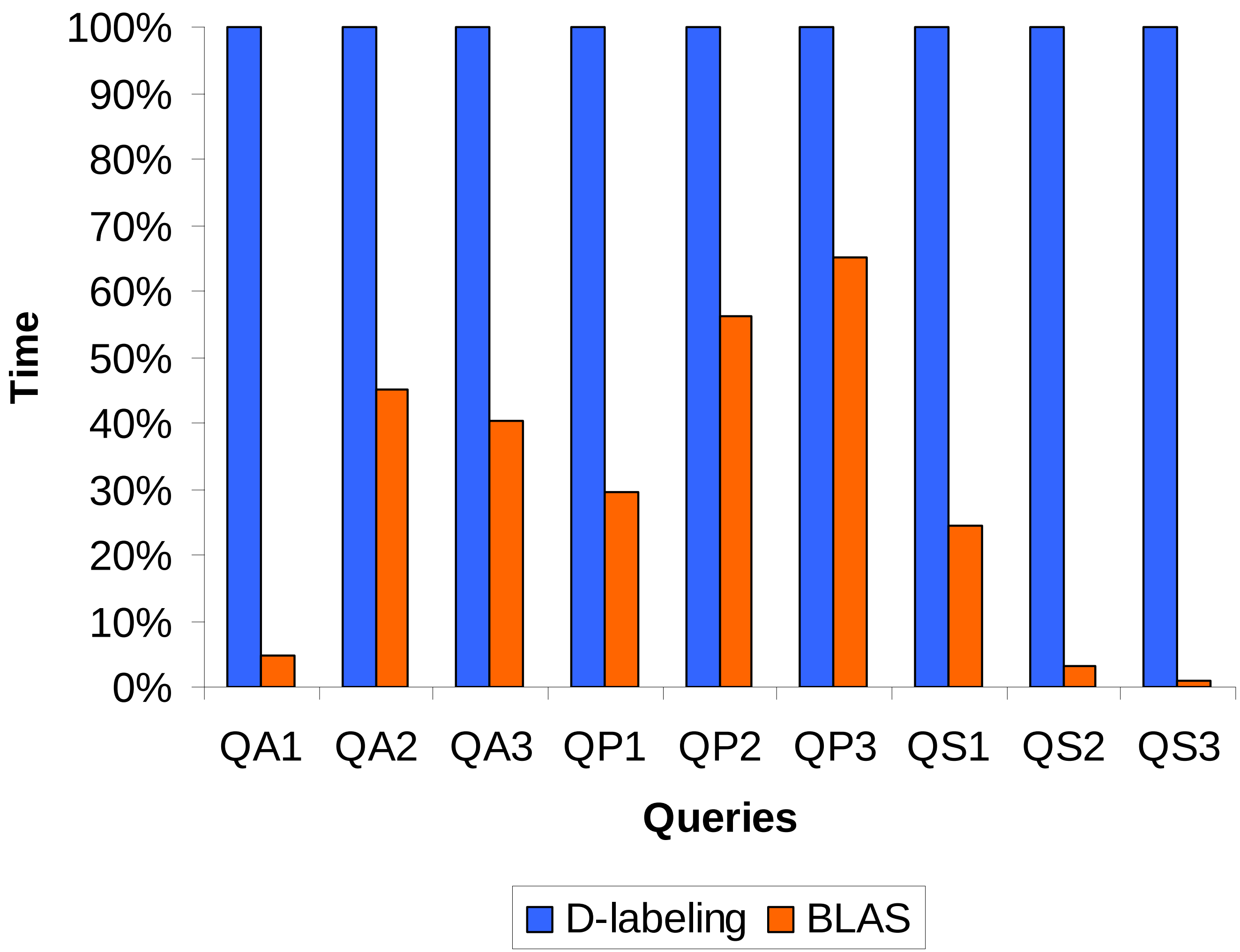
Compare our system (BLAS) with the previous approach using D-labeling scheme only

- Data sets

Data Set	Size(MB)	Nodes(K)	Tags	Depth	DTD
Protein	70	2277	66	7	Tree
Shakespeare	26	640	19	7	Acyclic graph
Auction	69	1238	77	12	Cyclic graph

- Query sets
  - Suffix path queries
  - Path queries
  - XPath queries
  - Benchmark queries
- Query Engines: DB2, TwigStack Join [Bruno et al 02]

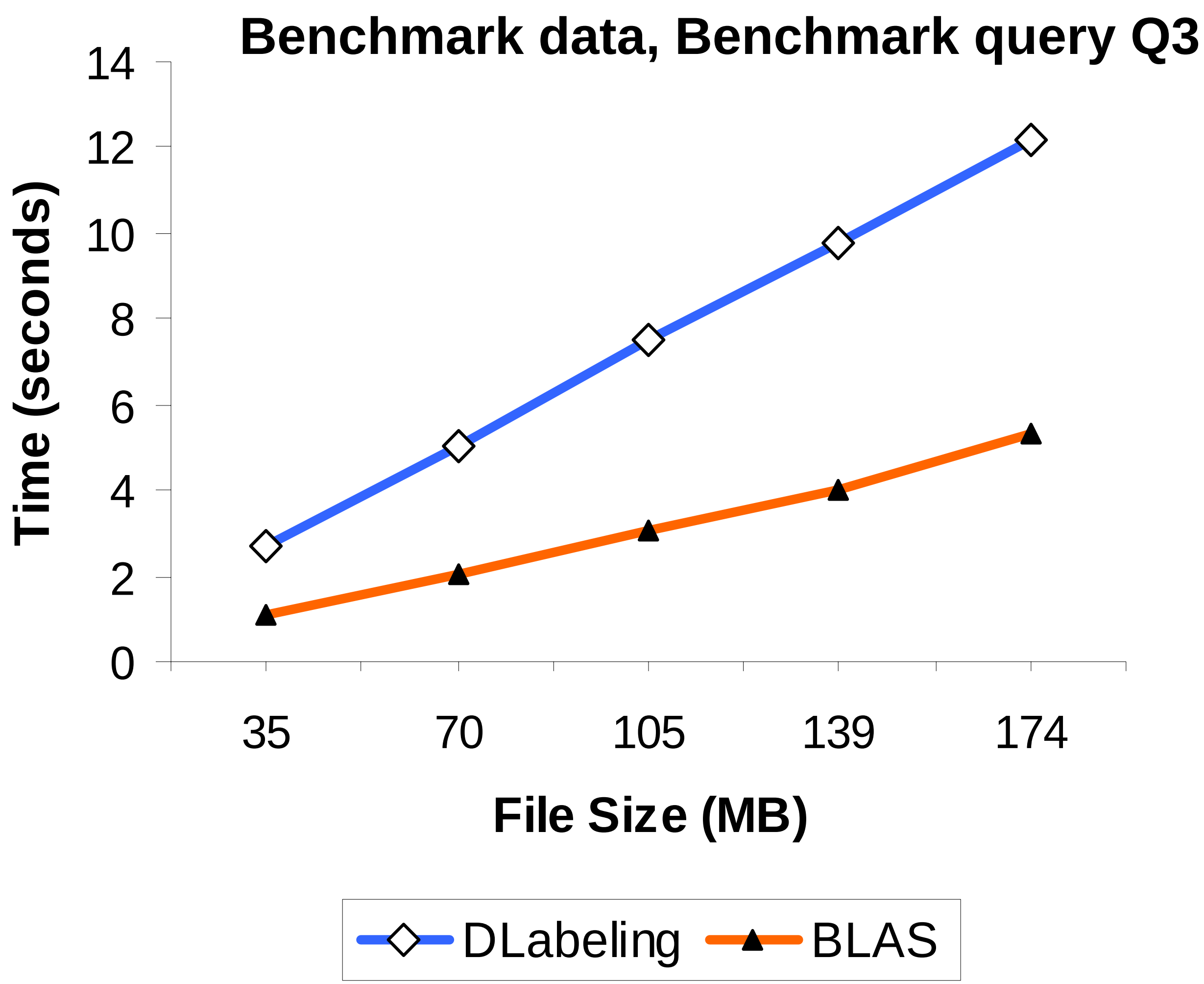
Query Execution Time



Query Name:
A:Auction
P: Protein
S: Shakespeare

1: suffix path query
2: path query
3: XPath query

Scalability



Summary of XML Data Management

- We proposed a generic XML-to-RDB mapping, based on a bi-labeling scheme.
- It is more efficient compared with previous approach, since it generates SQL queries that require:
    - fewer disk accesses
    - fewer joins
    - fewer intermediate results
- Experiments show the effectiveness

Roadmap of This Talk

- Managing XML by leveraging mature RDBMS [Chen et al 04]
  - Introduction to XML
  - A generic and efficient XML-to-RDBMS mapping
    - Data mapping from trees to tables
    - Query translation from tree navigation queries to SQL queries that are efficient
- Handling imprecise and incomplete data in DBMS [Chen et al 06]

Probabilistic Databases: Managing Imprecise Data

- How to measure the imprecision of the data?
- The simplest model: associate each tuple a probability

Climber

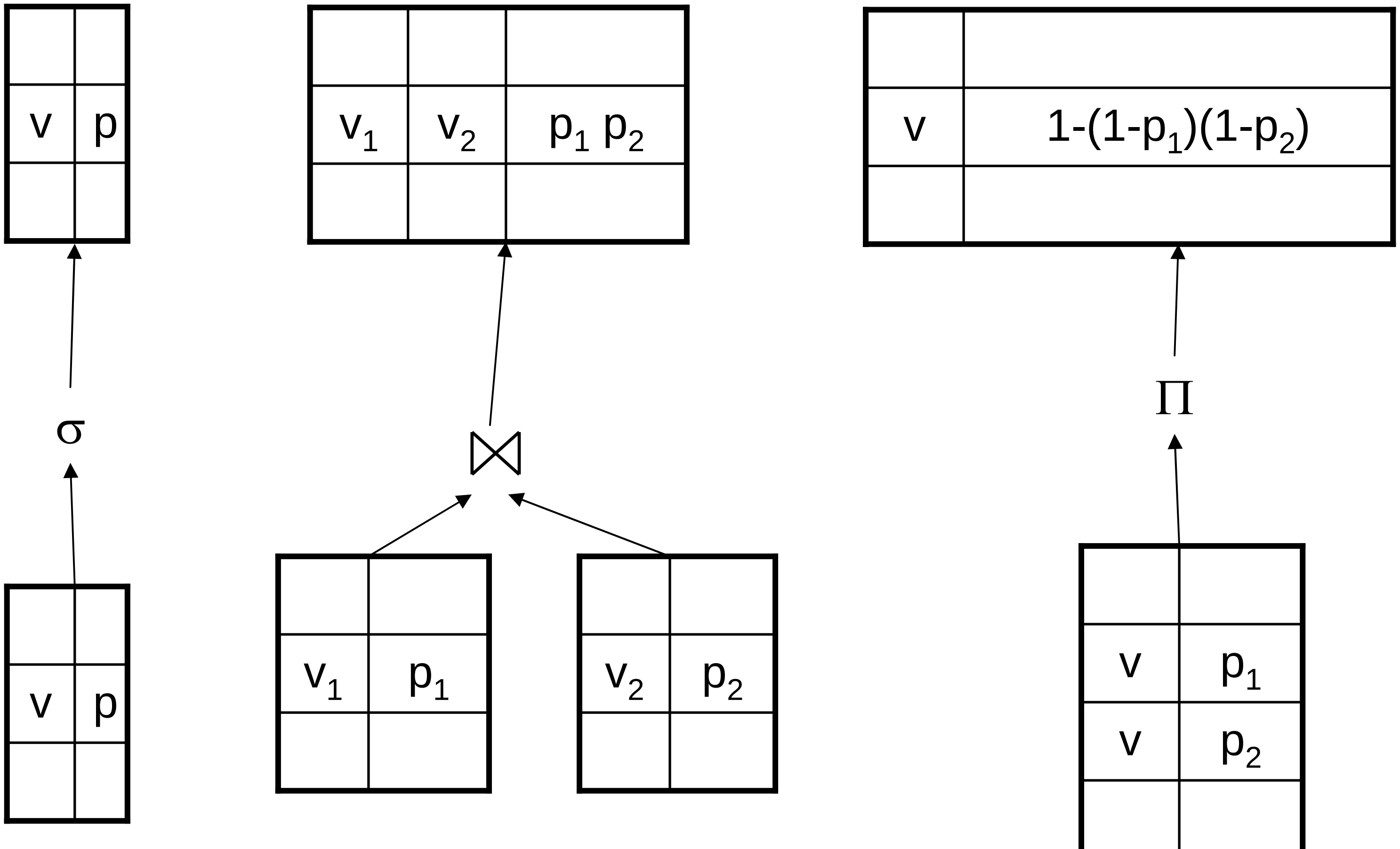
Name	Skill	age	Pr
James	Beginner	21	p1
Bob	Experienced	33	p2

Climbs

Name	Route	Date	Duration	Pr
Bob	Last Tango	10/10/05	5	q1
Bob	Last Tango	1/10/06	4.5	q2

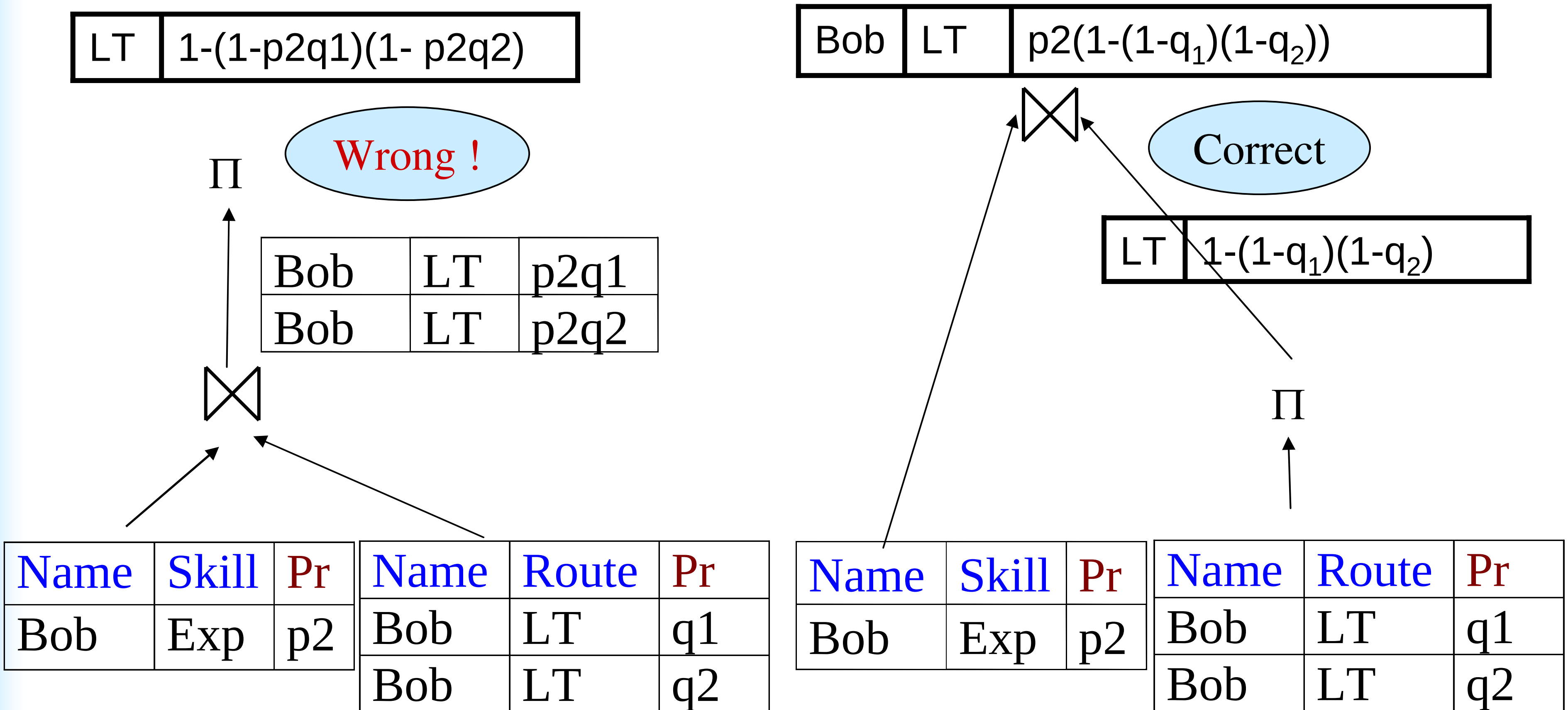
Assume that all the tuples are independent events

How Does This Affect Query Evaluation? [Fuhr&Roellke 97]



Challenges: Correctness Depends on Execution Order [Suciu et al 05]!

Find distinct climb routes that have been climbed by an experienced climber.



How to Handle Incomplete Data?

Find cars that are convertible and have price less than 10k.

Seller	Make	Model	Body Style	Year	Price
Mark	BMW	325Cic	convertible	2006	36000
Alice	Ford	Taurus		2001	8000

This can not be convertible !

Our techniques infer possible values with probability by mining data statistics

Querying Incomplete Databases

- Given missing value prediction

  Querying incomplete databases →

  Querying probabilistic databases

- What if we are not able to store the predicted values? --- e.g. data integration applications
- On-the-fly query rewriting

How should we handling imprecise and incomplete XML Data?

Conclusions

- Traditional RDBMS does not satisfy the requirements in ever growing scientific and web applications

- We have discussed two enhancement to RDBMS
  - Efficient XML data management
  - Handling imprecise and incomplete data

- Other enhancement to RDBMS that I am working on
  - Data stream processing
  - Scientific workflow modeling and query processing

Thank you!

Questions ?