Approximate Matching in XML

Sihem Amer-Yahia, Nick Koudas, Divesh Srivastava
AT&T Labs–Research
http://www.research.att.com/~{sihem, koudas, divesh}/

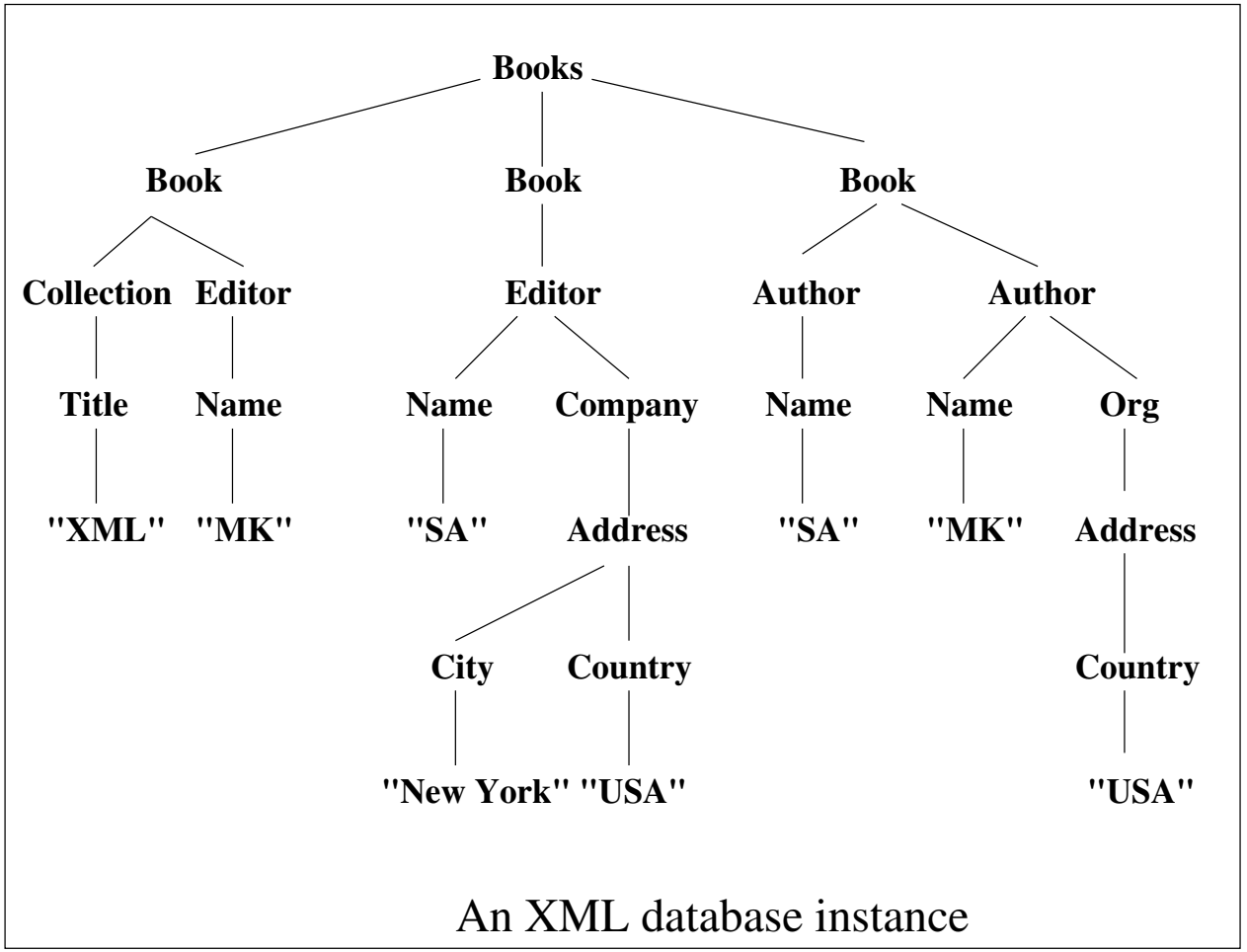
Outline of Seminar

- Motivation, language proposals: Divesh Srivastava
- Matching query twig to data tree: Sihem Amer-Yahia
- Matching data tree to data tree: Nick Koudas

What Makes XML Appealing?

- Represent structured, semi-structured, unstructured data
  - traditional databases: structure-rich
  - marked-up documents: text-rich
- Represent homogeneous, heterogeneous structure-rich data
  - repetition: chapter $\rightarrow$ section$+$, ...
  - optionality: book $\rightarrow$ cdrom$?$, ...
  - alternation: book $\rightarrow$ (editors $|$ authors), ...
  - recursion: section $\rightarrow$ section$*$, ...

XML Example: Data Trees



An XML database instance

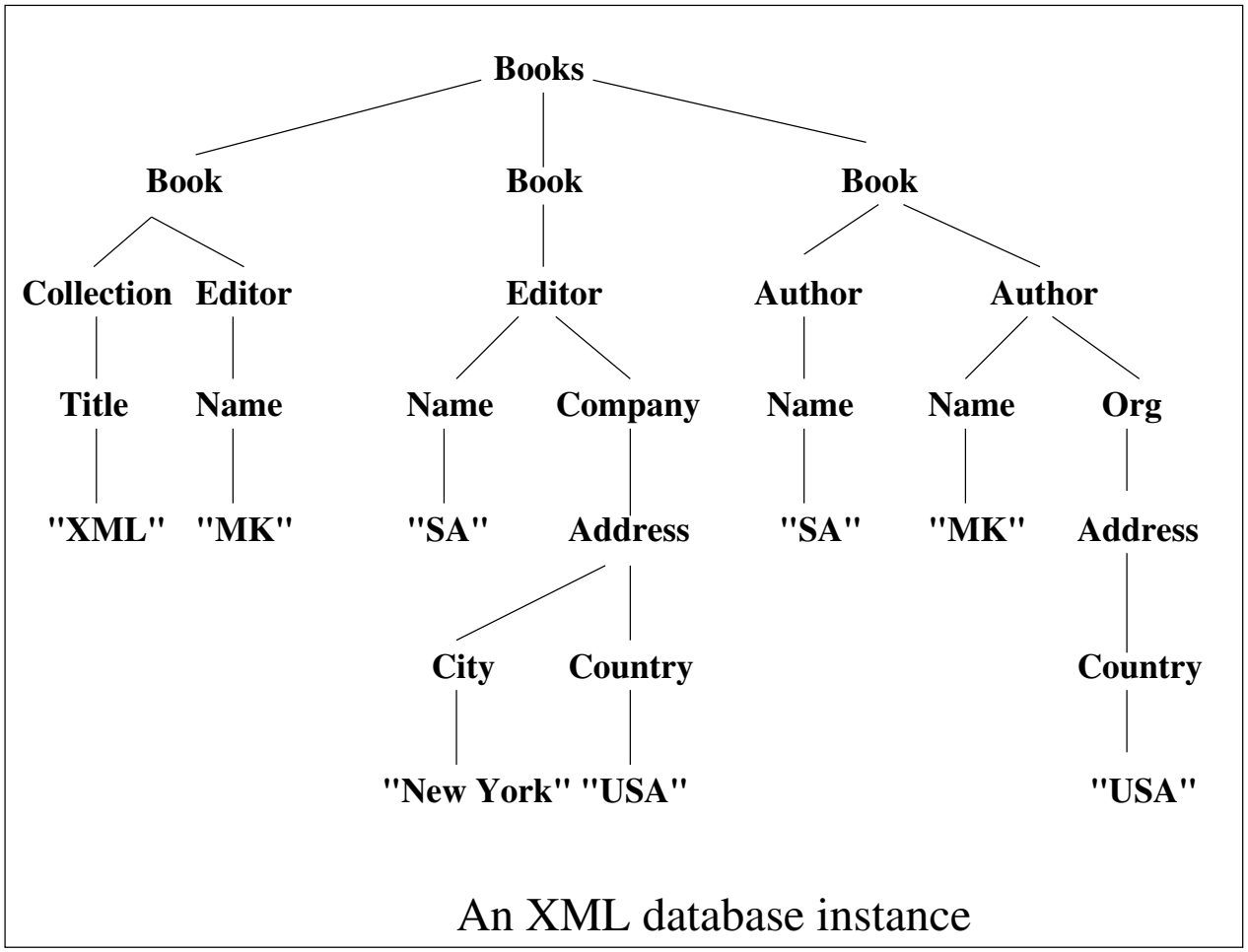
XML Queries

- Multiple XML query languages
  - semistructured: Lorel [AQM+97], YATL [SC98]
  - XML focus: XML-QL [DFF+98], XQL [RLS98], Quilt [CRF00]
  - W3C: XQuery [BCF+02]
- Basic features: pattern, filter, construction clauses [FSW99]

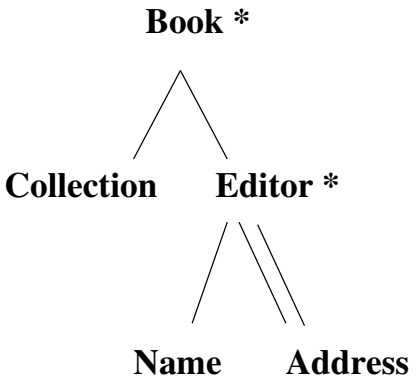
```
FOR $b IN document("books.xml")//book[./collection],
        $e IN $b/editor[./name AND .//address]
RETURN $e//country
```

- Heterogeneity: regular path expressions [CACS94,AQM+97]
  - book.(author)?.name, book.(editor|author).address, ...

XML Example: Twig Query



An XML database instance

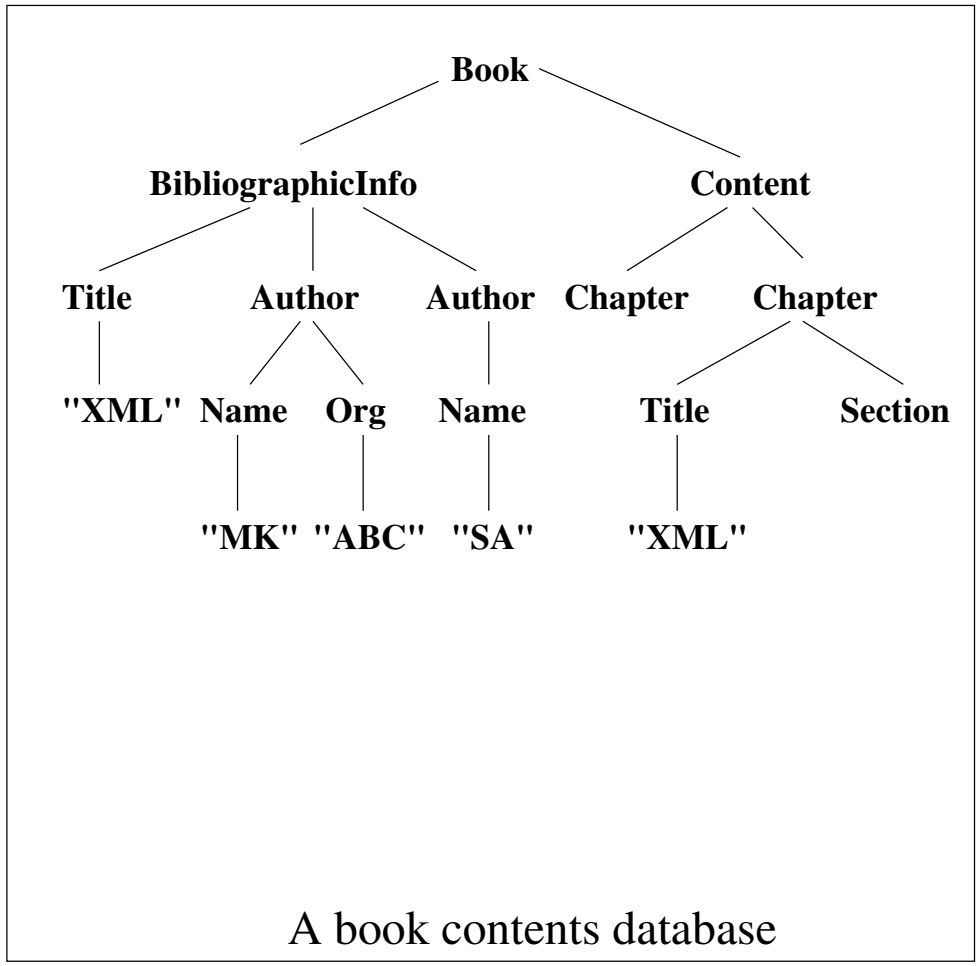


A tree pattern

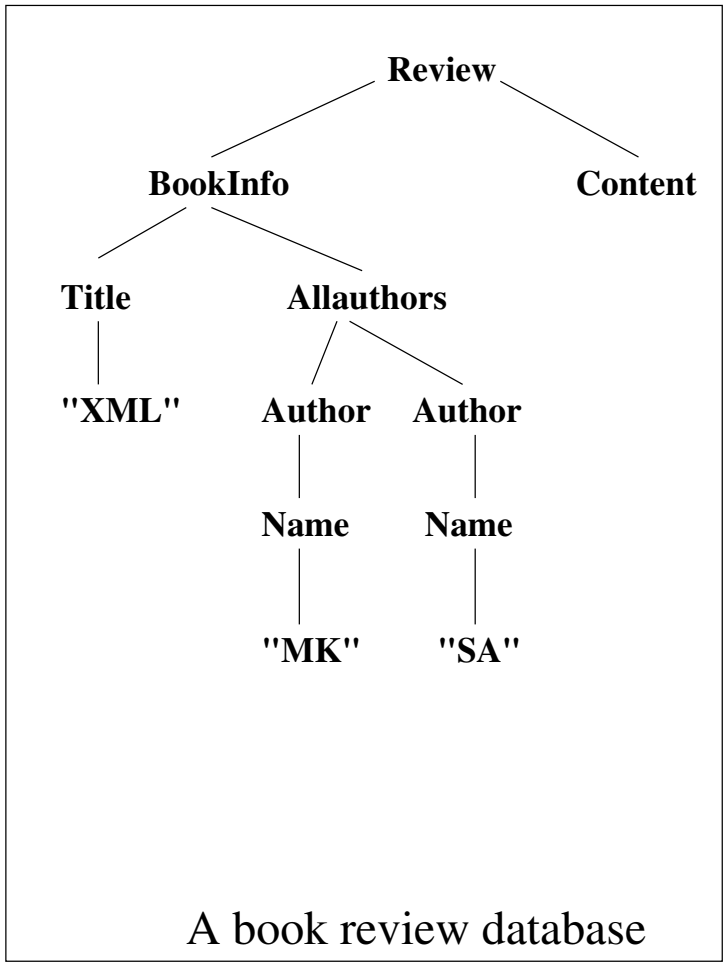
Motivation: Approximate Matching Applications

- Naive-user querying
  - goal: specify query twig, get ranked list of data tree matches
  - generalizes, focuses keyword-based search
- XML data correlation/integration
  - goal: specify matching elements, get ranked list of data tree correlates
  - generalizes string similarity-based correlations

XML Example: XML Data Correlation



A book contents database



A book review database

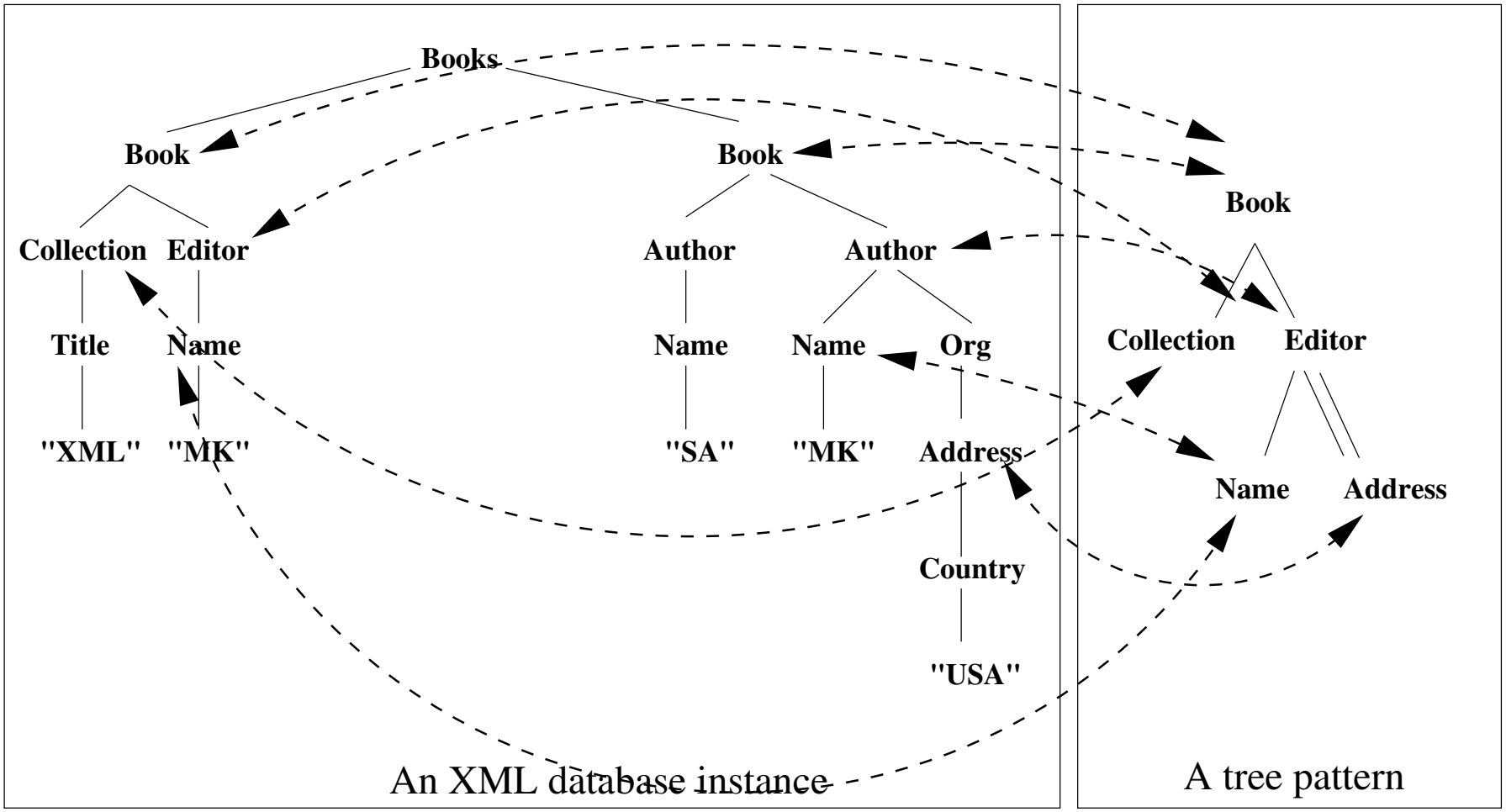
Approximate Matching: Rationale

- Traditional semantics: exact matching
  - embed query twig in data tree
  - isomorphism between data trees
  - useful for well-understood, homogeneous data
- Motivation for approximate matching: data heterogeneity
  - schema often allows heterogeneity, might not be known
  - schema complexity $\Rightarrow$ complex matching pattern
  - simple matching pattern: too few or no exact matches

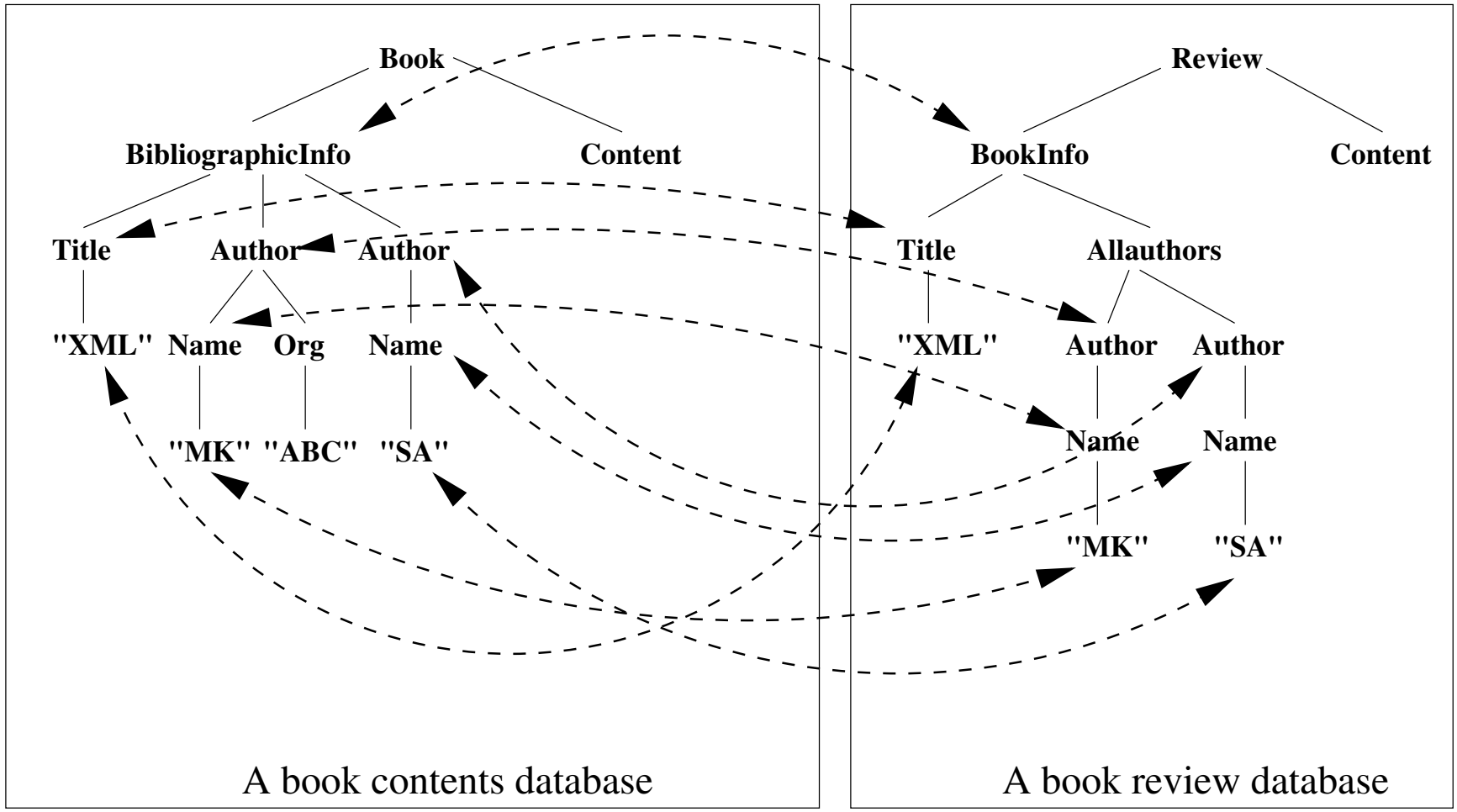
Approximate Matching: Key Notions

- Content matching
  - string edit distance, semantic distance
- Structure matching: variants of tree edit distance
  - node matching, cheapest edit script
  - edit operations: insert node, delete node, rename node, move subtree, ...
- Score of match, ranking, top-K
  - based on structure + content matching
  - edit distance, weighted edit distance (e.g., tf$*$idf)

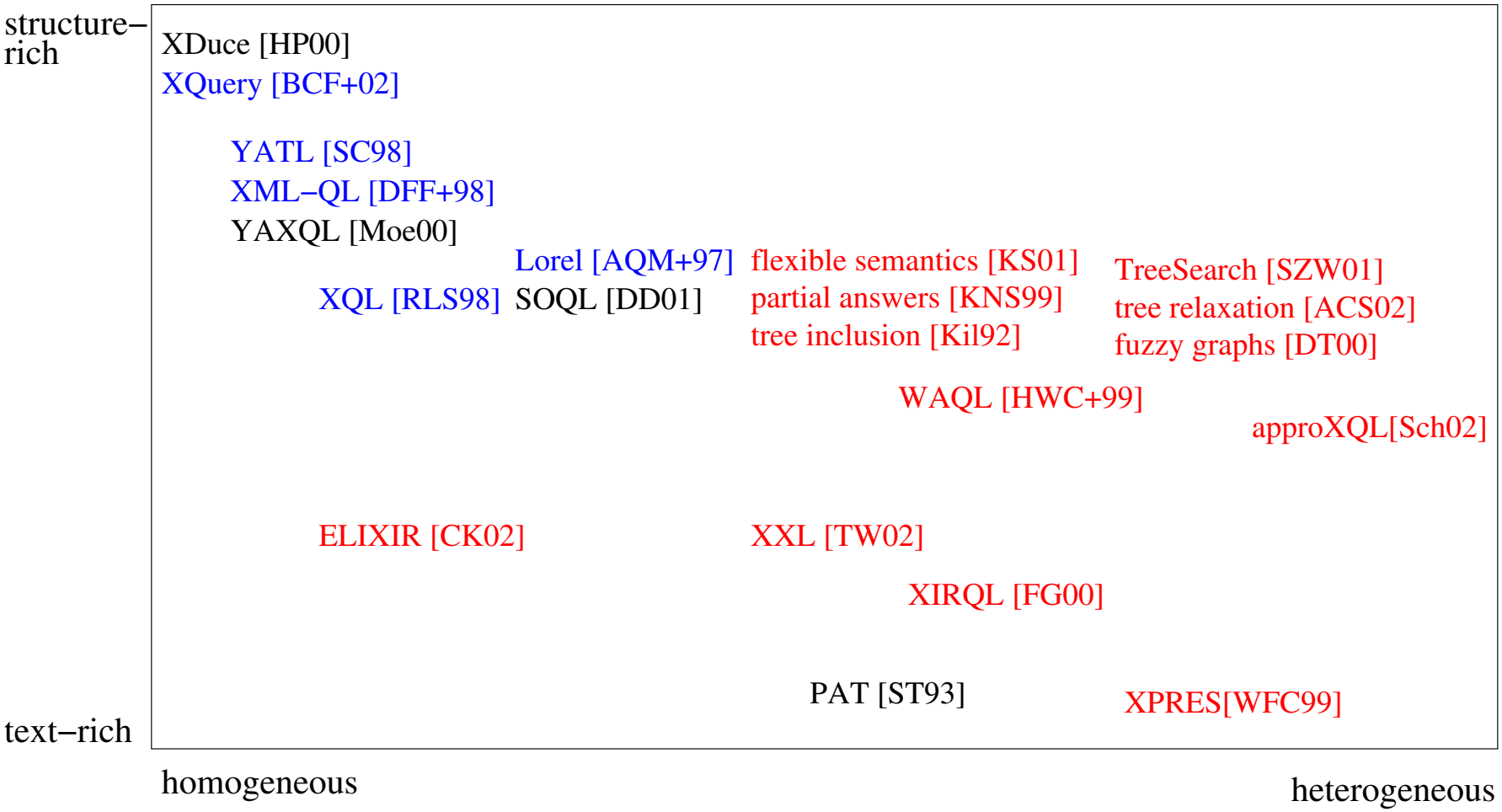
Structure Matching: Query-Data



Structure Matching: Data-Data



Taxonomy of XML Query Languages [Sch02]



XML Approximation Query Languages [Sch02]

query language or retrieval model	value matching	tag matching	delete nodes	insert nodes	permute nodes	scores ranking
approXQL [Sch02]	Yes	Yes	Yes	Yes	Yes	Yes
ELIXIR [CR02]	Yes					Yes
flexible semantics [KS01]					Yes	
fuzzy graphs [DT00]				Yes		Yes
partial answers [KNS99]			Yes			
tree inclusion [Kil92]				Yes		
tree relaxation [ACS02]		Yes	Yes	Yes	Yes	Yes
TreeSearch [SZW01]	Yes	Yes	Yes	Yes		Yes
WAQL [HWC+99]	Yes	Yes	Yes	Yes		
XIRQL [FG01]	Yes	Yes				Yes
XPRES [WFC99]	Yes					Yes
XXL [TW02]	Yes	Yes				Yes