Structures, Semantics and Statistics

Alon Halevy
University of Washington, Seattle

VLDB, September 1, 2004

Abstractions 'R Us

- *Logical* vs. *Physical*; *What* vs. *How.*

Students:

SSN	Name	Category
123-45-6789	Charles	undergrad
234-56-7890	Dan	grad
	...	...

Takes:

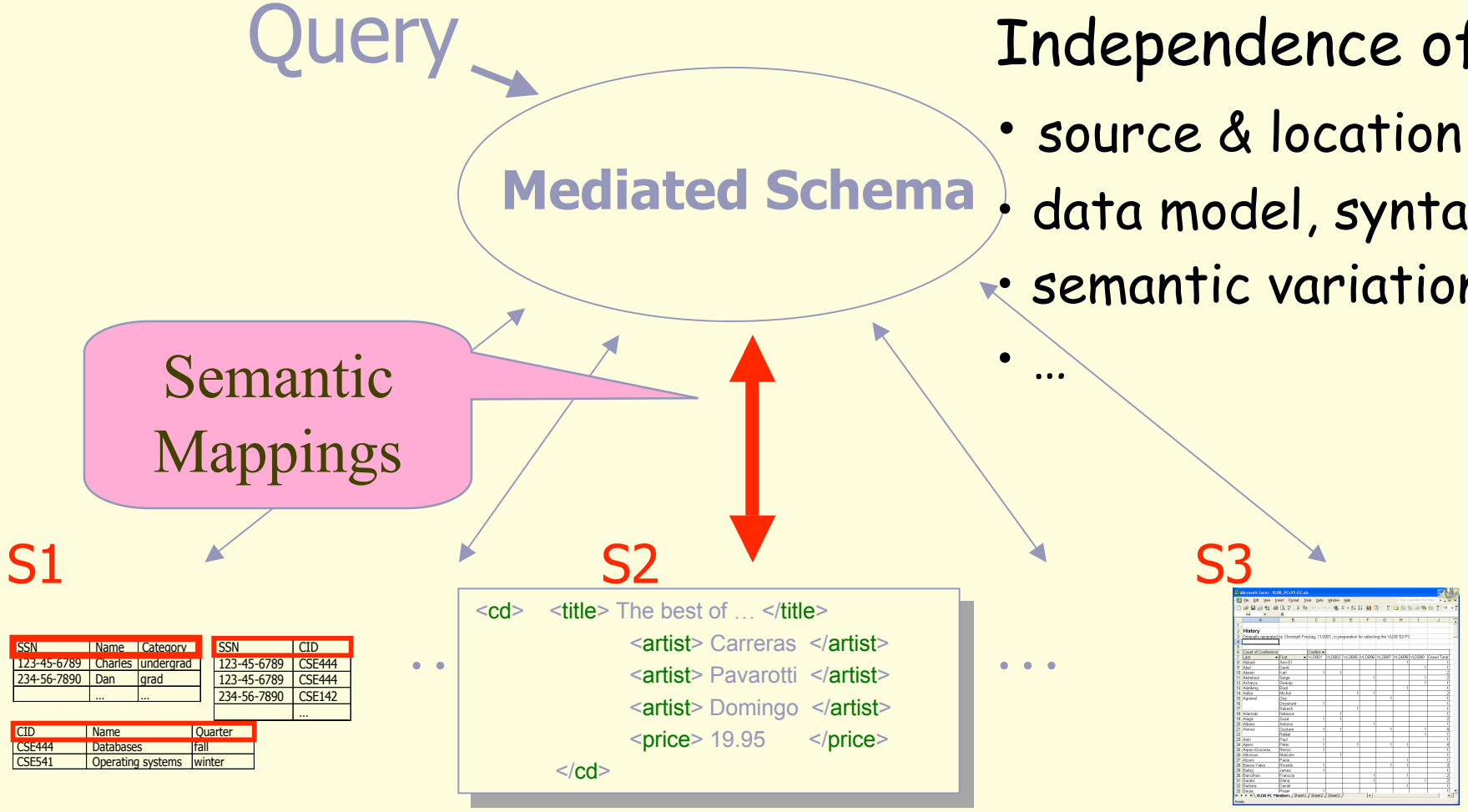
SSN	CID
123-45-6789	CSE444
123-45-6789	CSE444
234-56-7890	CSE142
	...

Courses:

CID	Name	Quarter
CSE444	Databases	fall
CSE541	Operating systems	winter

```
SELECT  C.name
FROM Students S, Takes T, Courses C
WHERE S.name="Mary" and
        S.ssn = T.ssn and T.cid = C.cid
```

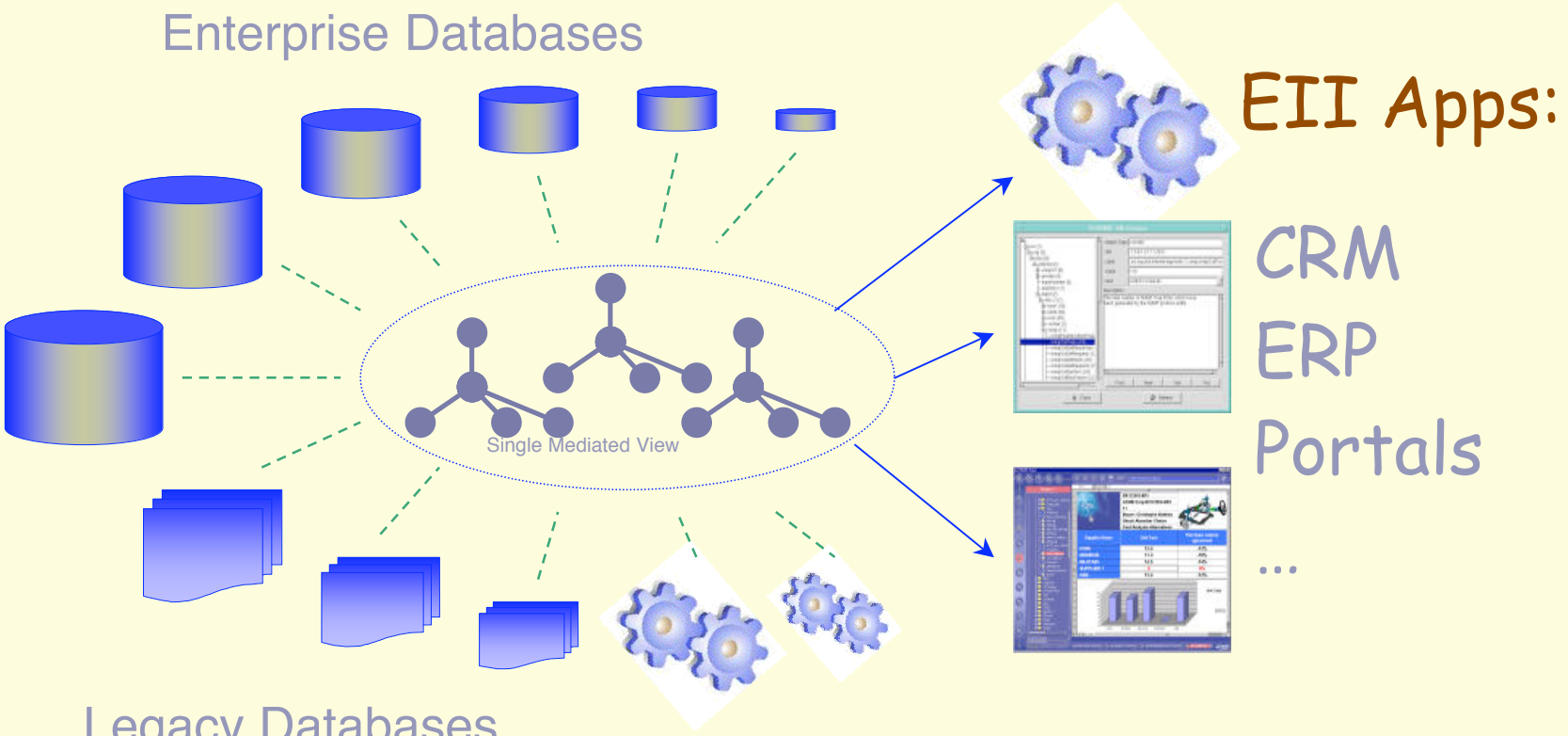
Data Integration: A Higher-level Abstraction



Independence of:

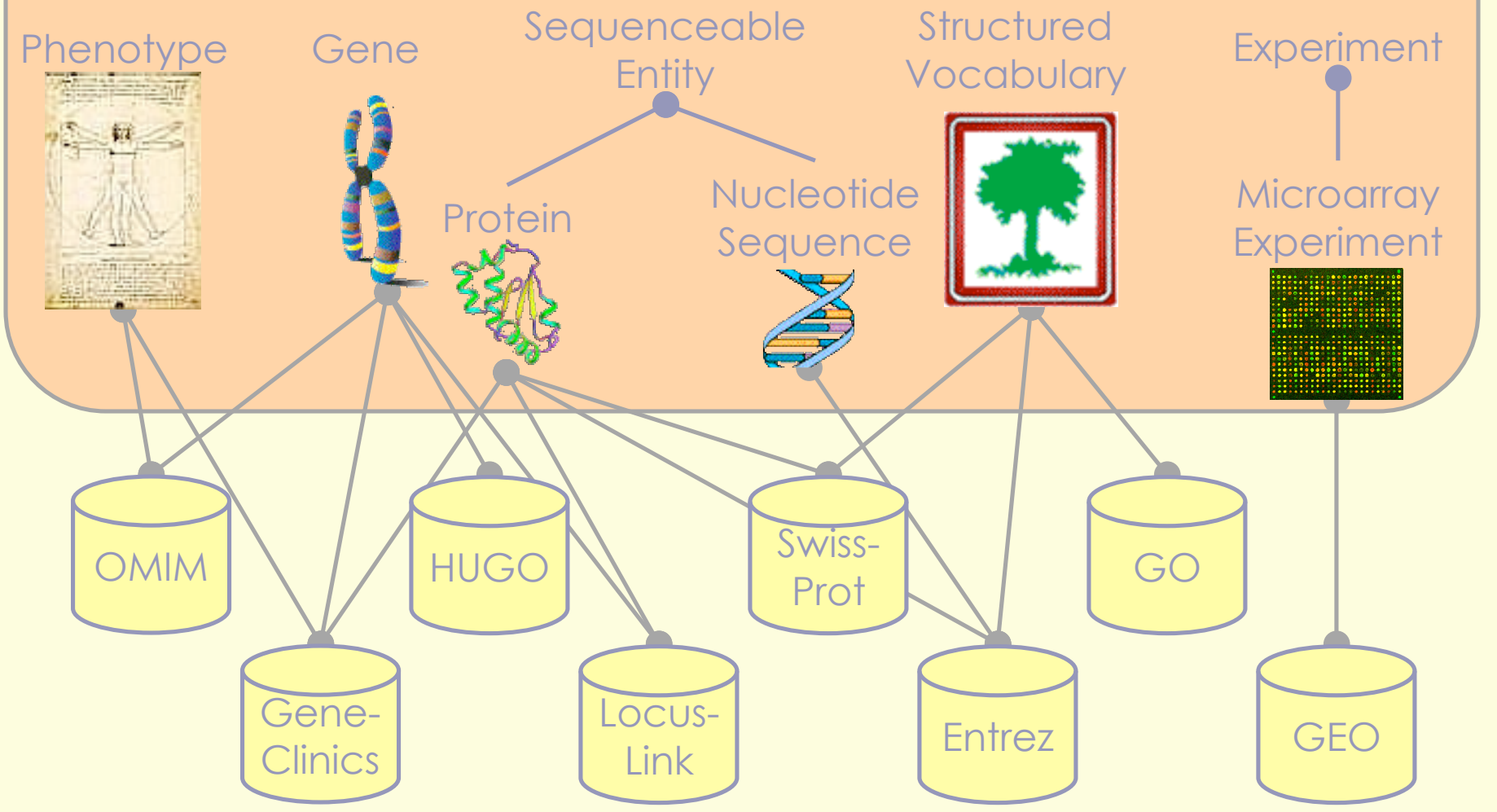
- source & location
- data model, syntax
- semantic variations
- …

Application Area 1: Business



50% of all IT $$$ spent here!

Application Area 2: Science



Hundreds of biomedical data sources available; growing rapidly!

Application Area 3: The Web

The Semantic Web

[Berners-Lee, Vision #2, Available in Beta]

- Knowledge sharing at *web* scale.
- Web resources are described by *ontologies:*
  - Rich domain models; allow reasoning.
  - RDF/OWL are the emerging standards
    - OWL-lite may actually be useful.
- Issues:
  - Too complex for users?
  - Killer apps?
  - Scalability of reasoning?
- Proposal: let's build the SW *bottom up*.

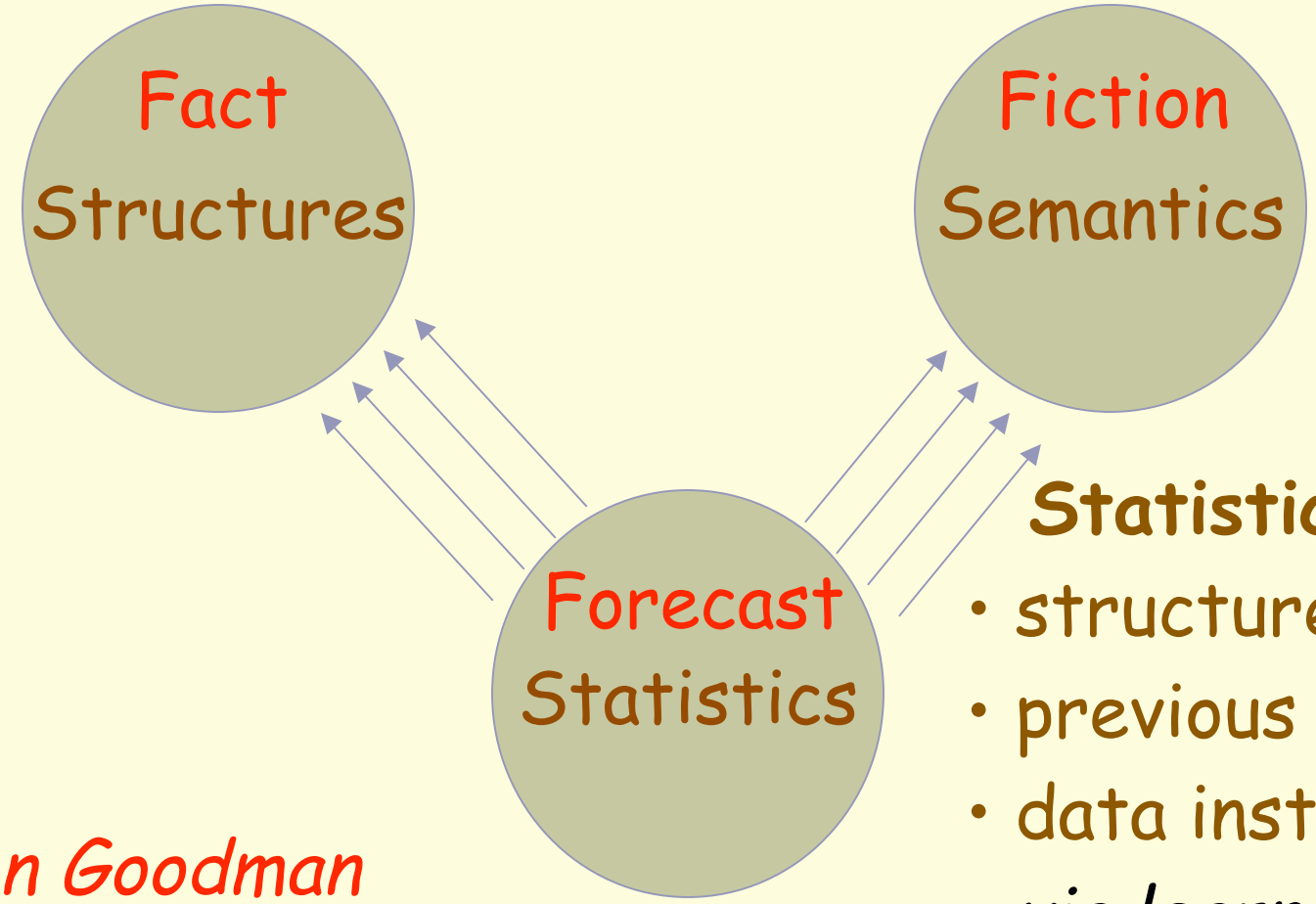
New Decade(s), *Old* Problem

- Data integration is listed on every 5-year DB introspective (the latest: Lowell, 2003).
- Data integration is a *necessity*:
  - Competitive advantage, B2B, science
- Significant new twists on the problem:
  - Number of sources
  - Types of data, queries, and answers
  - User skills
- It's plain hard!

Why is it Hard?

- Systems reasons:
  - Managing different platforms
  - Query processing across multiple systems
- Social reasons:
  - Locating and capturing relevant data in the enterprise.
  - Convincing people to share (*data fiefdoms*)
    - Privacy and performance implications.
- Logic reasons:
  - Schema (and data) heterogeneity
  - *Challenge independent of integration architecture!*

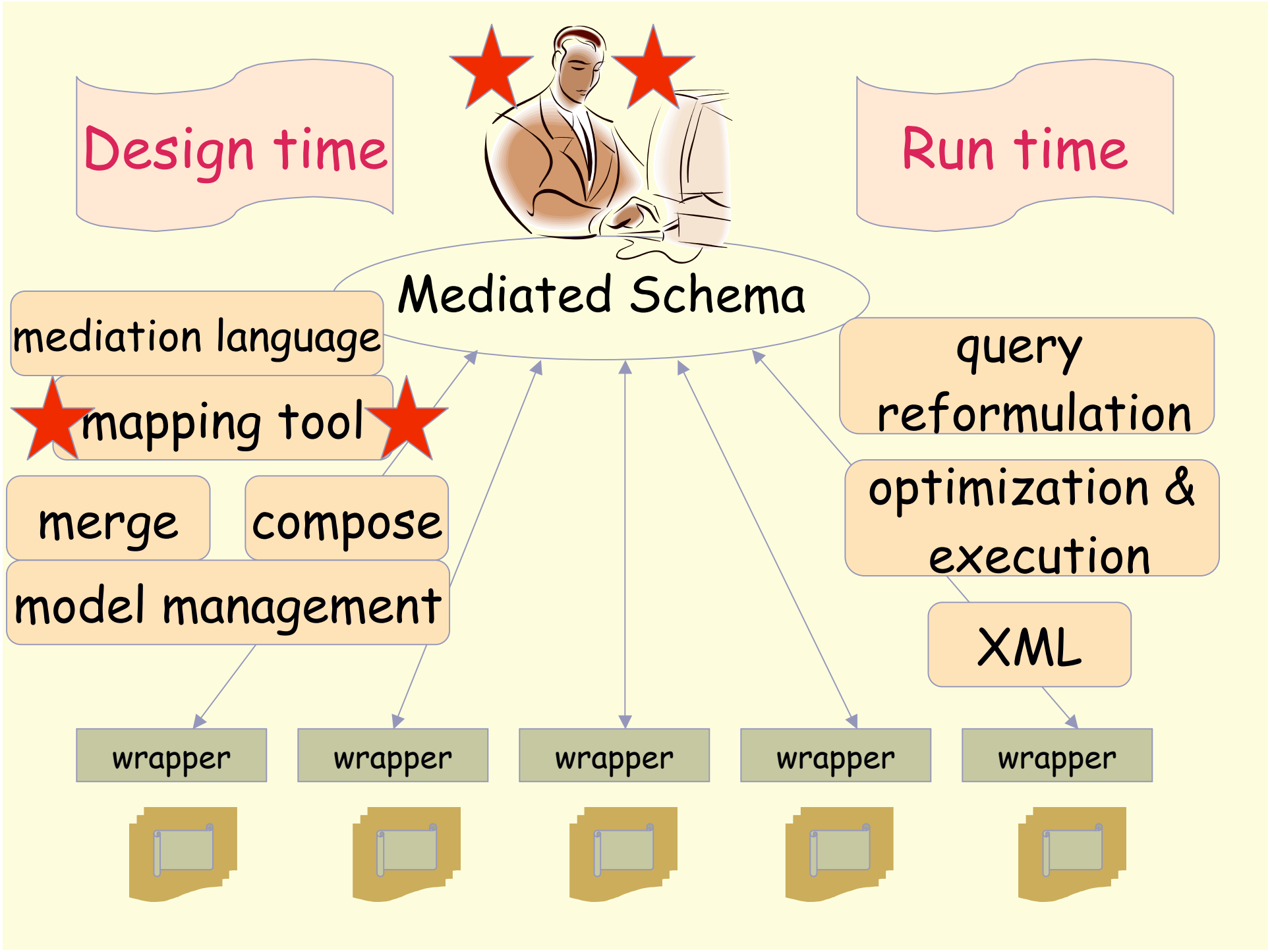
Explaining the Title

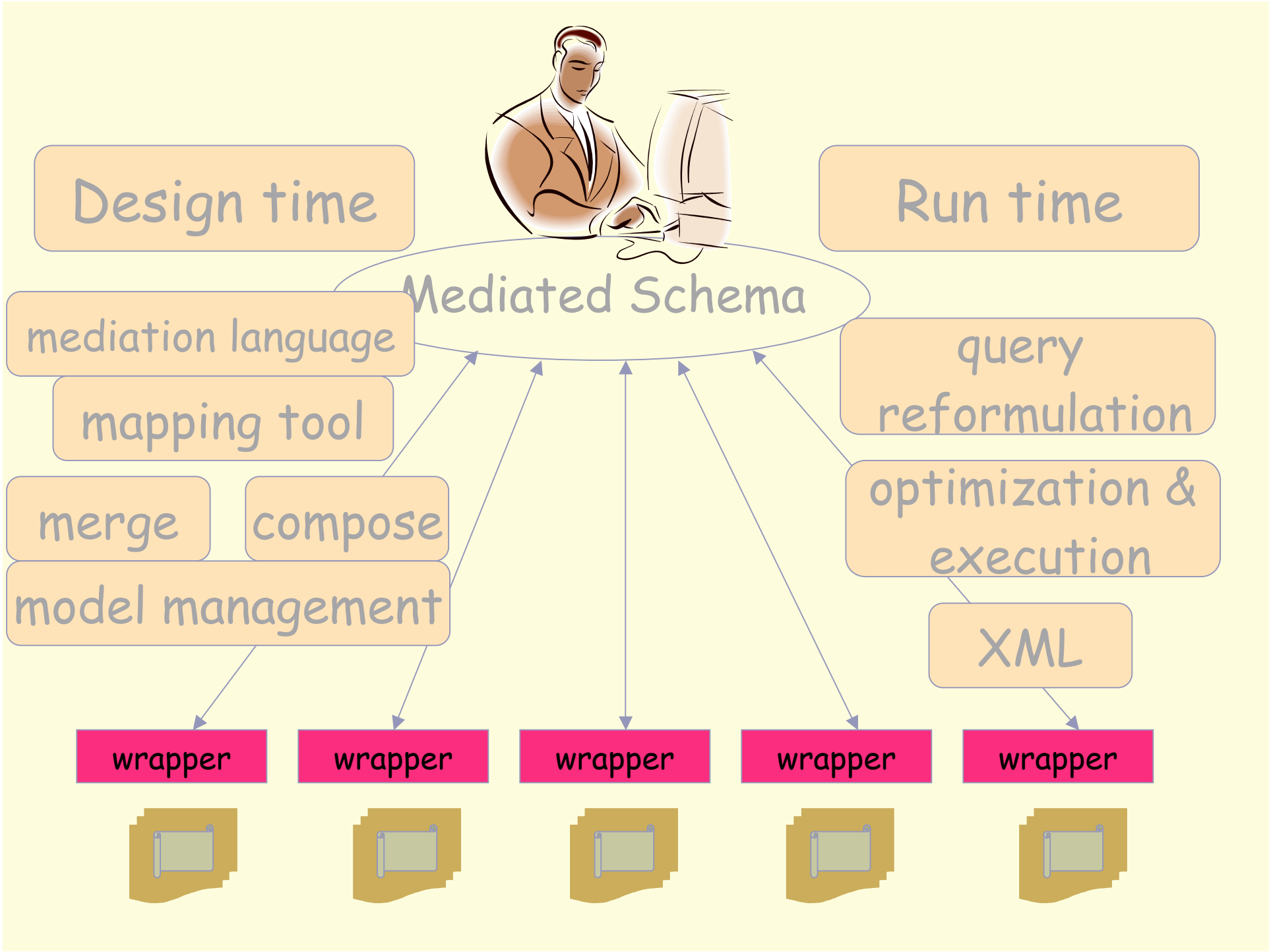


Statistics on:

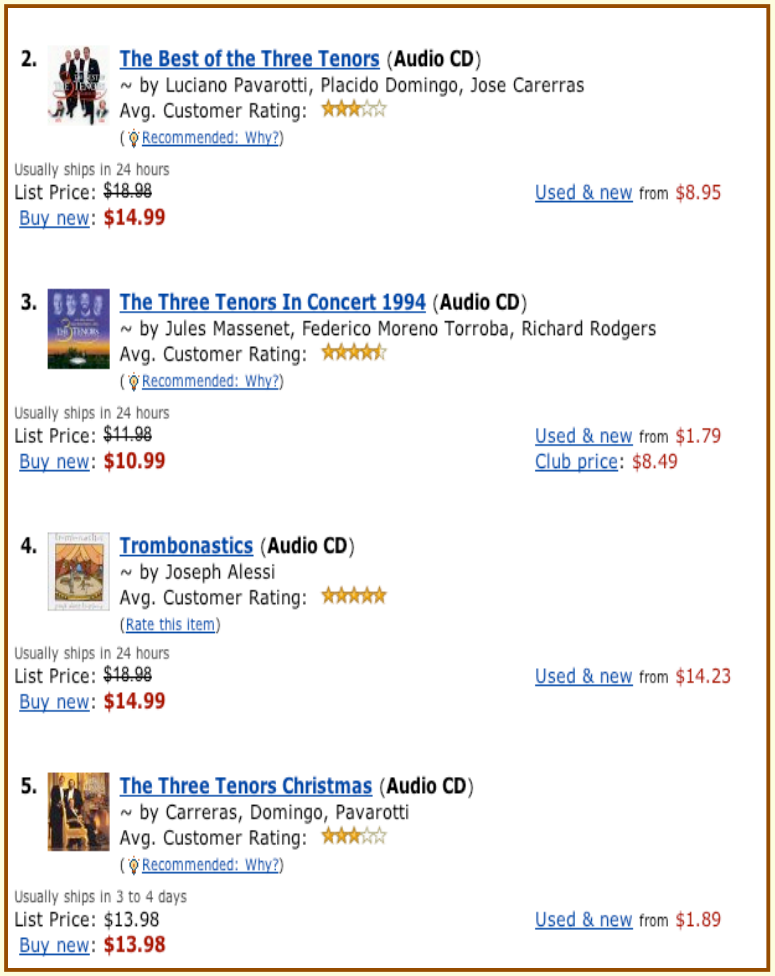
- structures
- previous actions
- data instances
- *via learning*

Nelson Goodman


Design time
Run time
Mediated Schema
mediation language
mapping tool
merge
compose
model management
query reformulation
optimization & execution
XML
wrapper
wrapper
wrapper
wrapper
wrapper


Design time
Run time
Mediated Schema
mediation language
mapping tool
merge
compose
model management
query reformulation
optimization & execution
XML
wrapper
wrapper
wrapper
wrapper
wrapper

Wrapper Construction



2. The Best of the Three Tenors (Audio CD)
~ by Luciano Pavarotti, Placido Domingo, Jose Carerras
Avg. Customer Rating:
(Recommended: Why?)
Usually ships in 24 hours
List Price: $18.98
Buy new: $14.99
Used & new from $8.95

3. The Three Tenors In Concert 1994 (Audio CD)
~ by Jules Massenet, Federico Moreno Torroba, Richard Rodgers
Avg. Customer Rating:
(Recommended: Why?)
Usually ships in 24 hours
List Price: $11.98
Buy new: $10.99
Used & new from $1.79
Club price: $8.49

4. Trombonastics (Audio CD)
~ by Joseph Alessi
Avg. Customer Rating:
(Rate this item)
Usually ships in 24 hours
List Price: $18.98
Buy new: $14.99
Used & new from $14.23

5. The Three Tenors Christmas (Audio CD)
~ by Carreras, Domingo, Pavarotti
Avg. Customer Rating:
(Recommended: Why?)
Usually ships in 3 to 4 days
List Price: $13.98
Buy new: $13.98
Used & new from $1.89

```
<cd>   <title> The best of … </title>
       <artist> Abiteboul </artist>
       <artist> Pavarotti  </artist>
       <artist> Domingo  </artist>
       <price> 19.95      </price>
  </cd>
  …
```

- Lixto [Vienna]
- Fetch [ISI]
- XQWrap [Georgia Tech]
- Wrapper induction [Dublin]

Design time

Run time

Mediated Schema

mediation language

mapping tool

merge

compose

model management

query reformulation

optimization & execution

XML

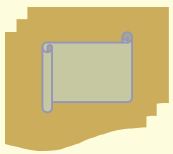
wrapper

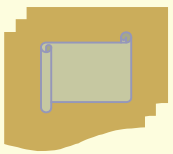
wrapper

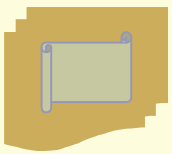
wrapper

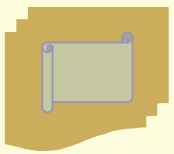
wrapper

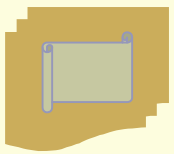
wrapper





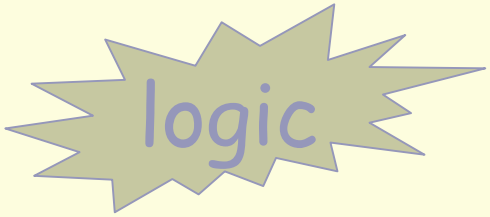




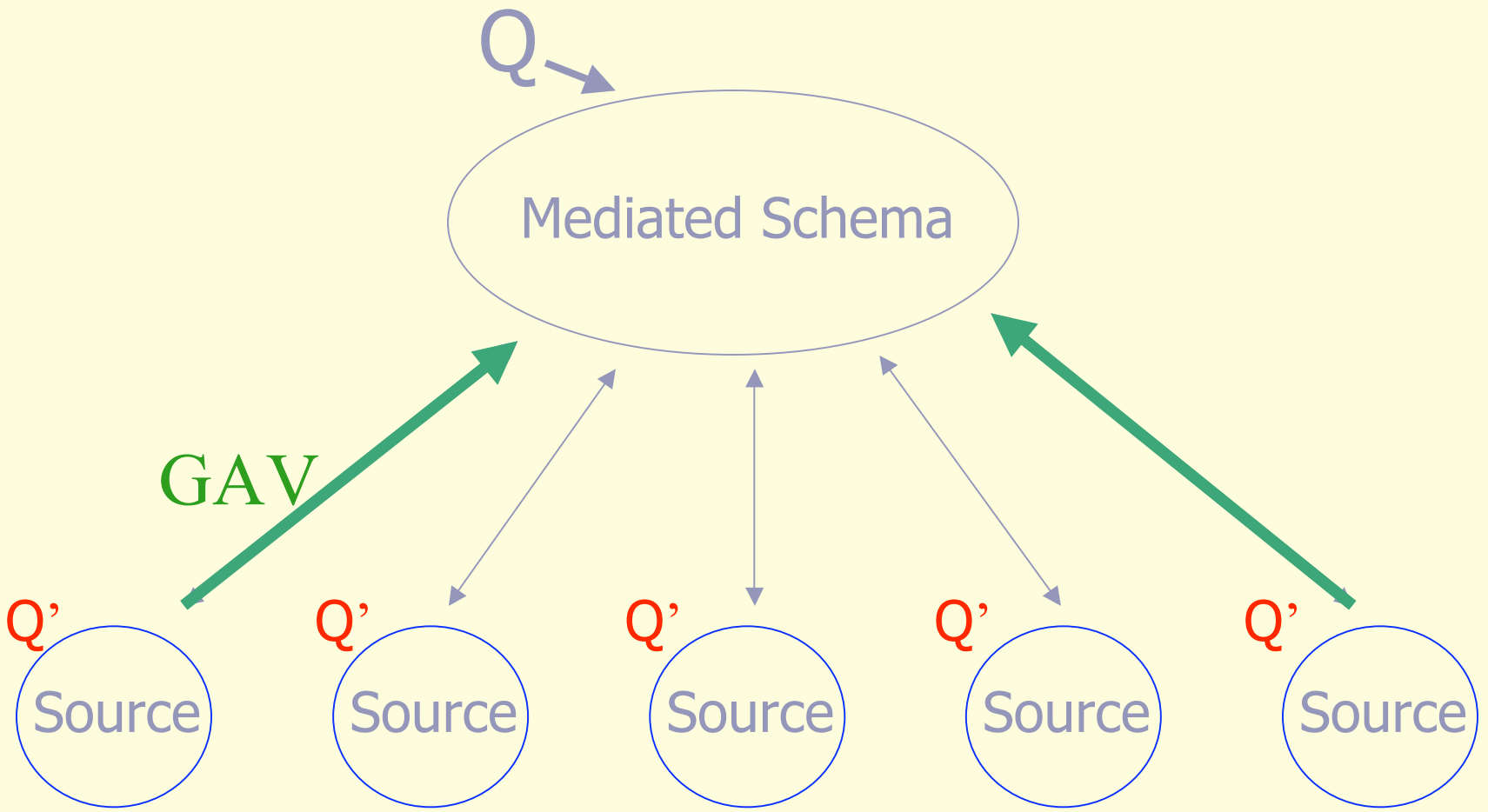


Mediation Languages

Mediated Schema
CD: ASIN, Title, Genre,…
Artist: ASIN, name, …



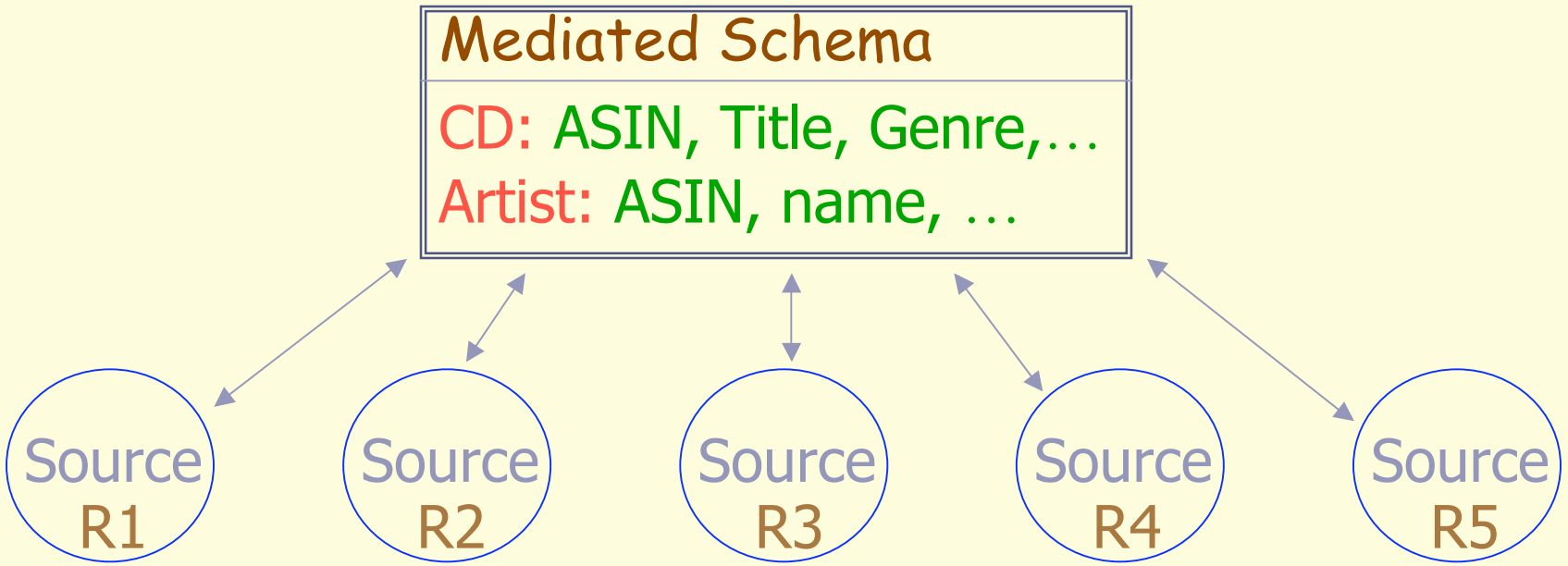
Mappings as Query Expressions



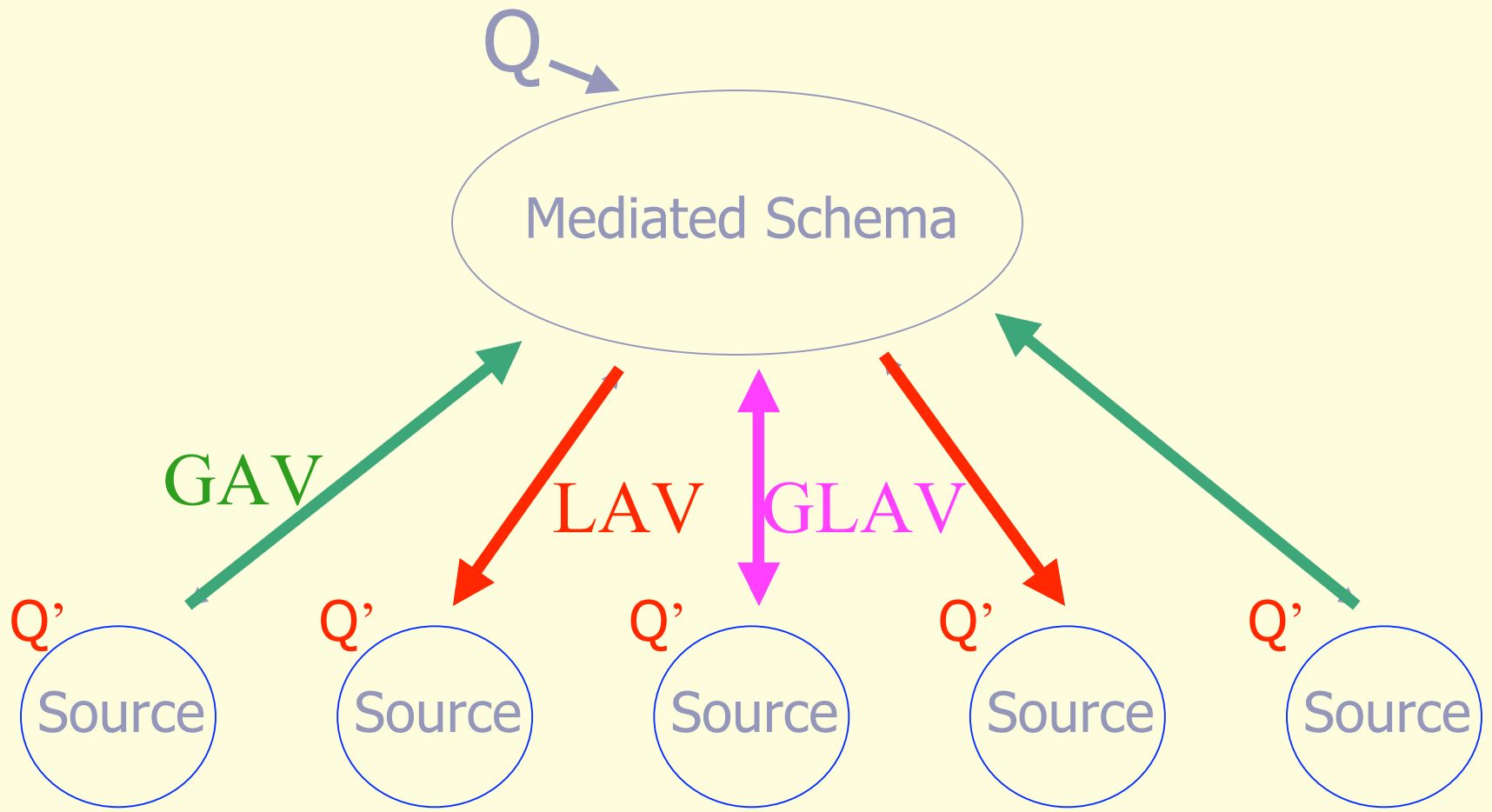
Global-as-View (GAV)

Mapping:

CD(x,y,z) :- R1(x,y,z)
CD(x,y,z) :- R2(x,z), R3(z,y)
…



Languages for Schema Mapping



Local-as-View (LAV,GLAV)

Mapping:

The details:

[Lenzerini, PODS 02]

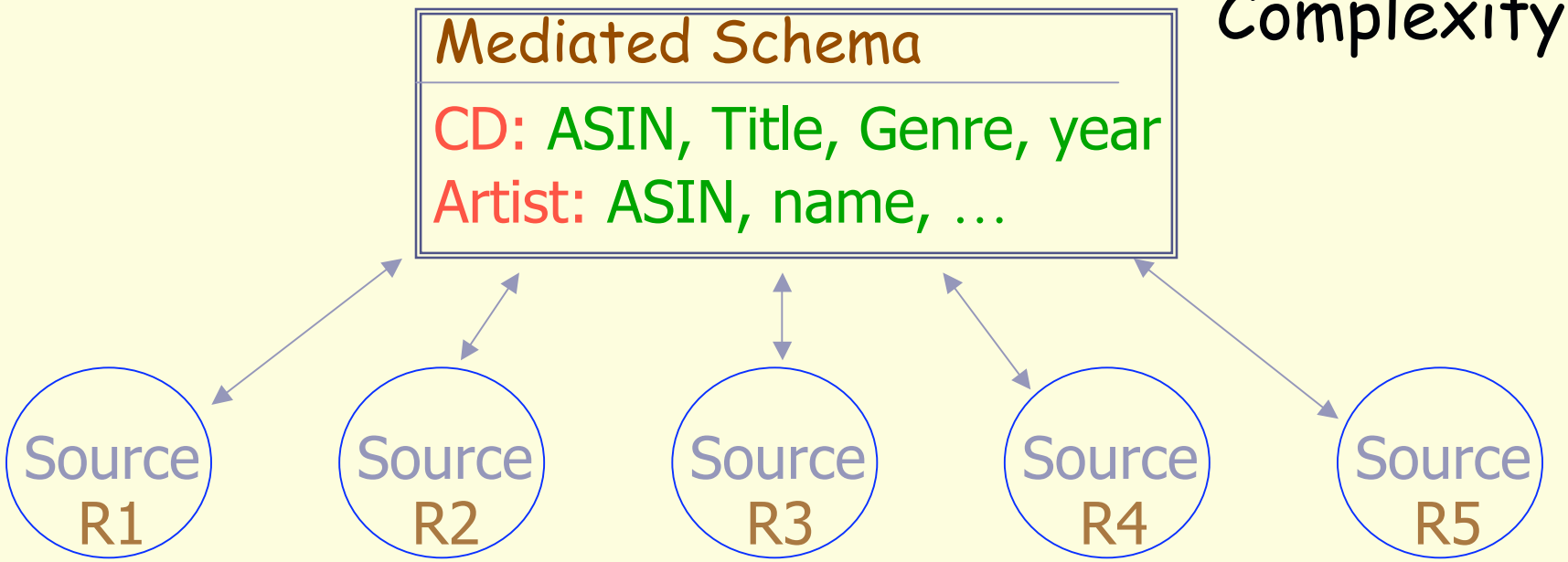
[Halevy, VLDBJ 01]

R1(x,y,t) :- CD(x,y,z), Artist(x,t), y< 1970

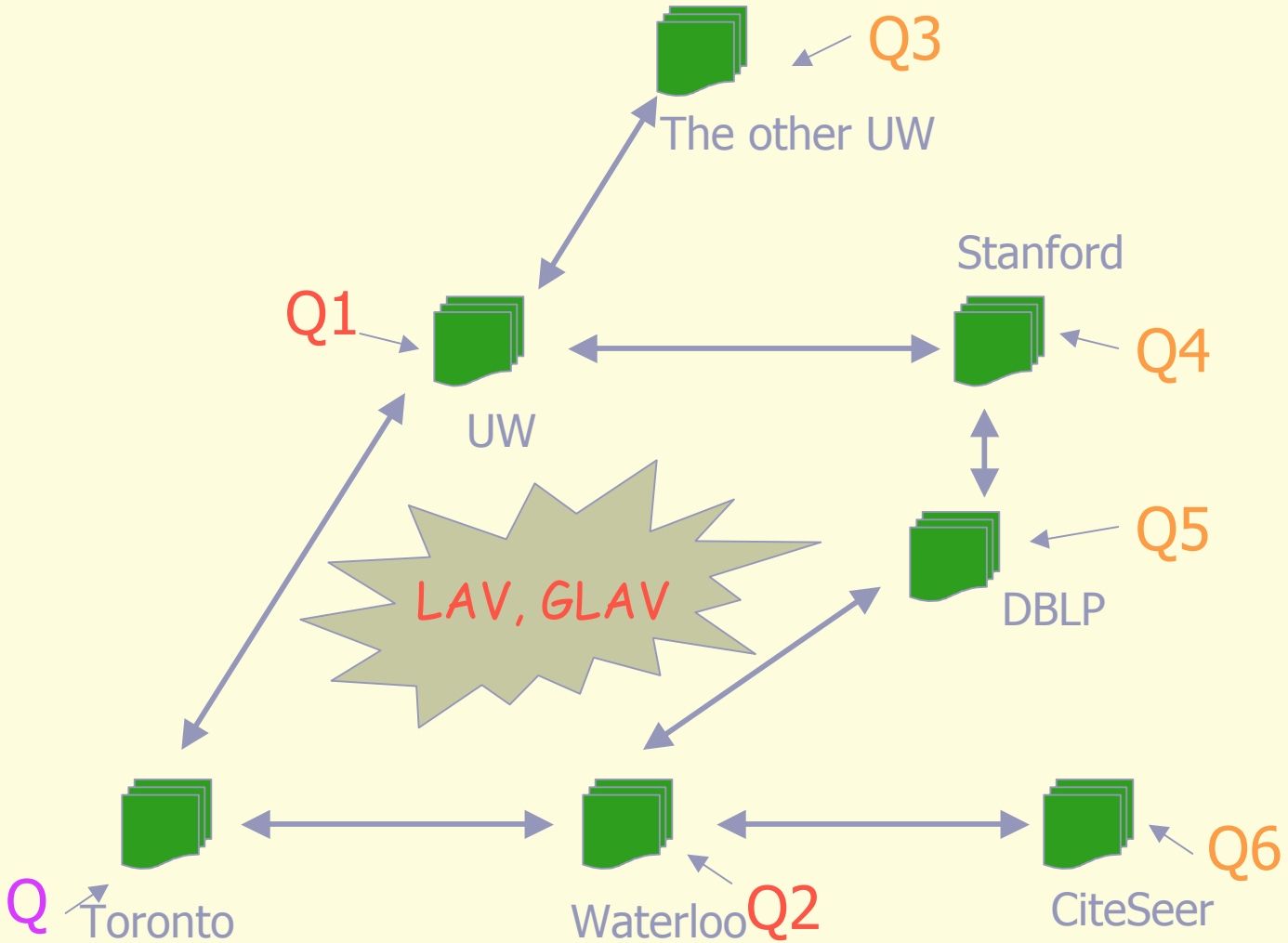
R2(x,y) :- CD(x,y,"French",z)

…

Complexity?



Peer Data Management Systems

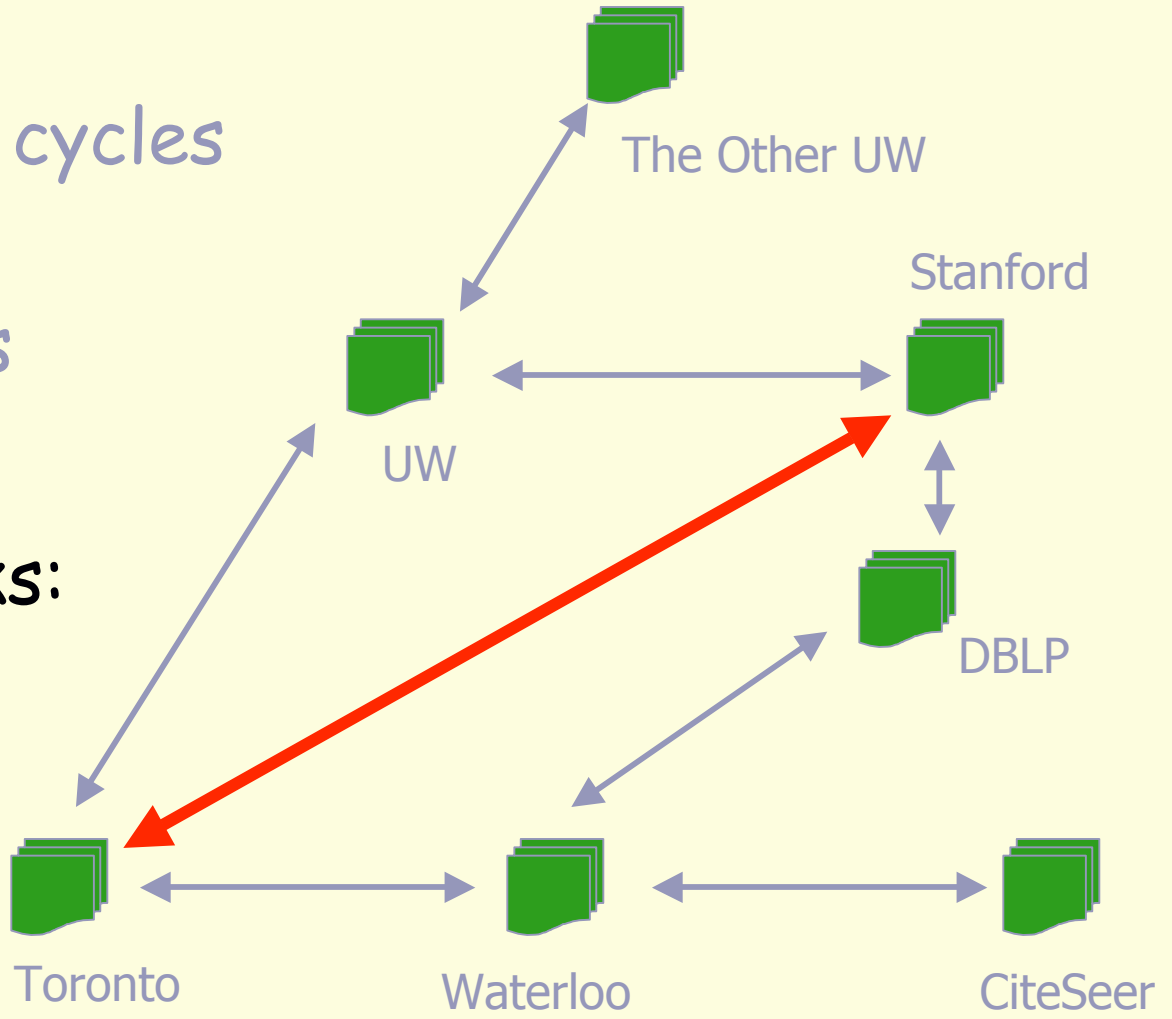


PDMS-Related Projects

- Piazza (Washington)
- Hyperion (Toronto)
- PeerDB (Singapore)
- Local relational models (Trento, Toronto)
- Active XML (INRIA)
- Edutella (Hannover, Germany)
- Semantic Gossiping (EPFL Lausanne)
- Raccoon (UC Irvine)
- Orchestra (U. Penn)

PDMS Challenges

- Semantics:
  - careful about cycles
- Optimization:
  - Compose paths
  - Prune
- Manage networks:
  - Consistency
  - Quality
  - Caching



Design time

Run time

Mediated Schema

mediation language

mapping tool

merge

compose

model management

query reformulation

optimization & execution

XML

wrapper

wrapper

wrapper

wrapper

wrapper

Model Management

- Generic infrastructure to manage schemas
  - Manipulate models and mappings as bulk objects
  - Operators to create & compose mappings, merge & diff models,
  - Short operator scripts can solve schema integration, schema evolution, reverse engineering, etc.
- See [Bernstein, CIDR-03], [Melnik et al., SIGMOD 03], [Pottinger & B, VLDB-03, etc.]
- *Great opportunity for fundamental theory and systems work.*

Design time

Run time

Mediated Schema

mediation language

mapping tool

merge

compose

model management

query reformulation

optimization & execution

XML

wrapper

wrapper

wrapper

wrapper

wrapper

Query Processing

- Problems:
  - DQP: See [Ozsu & Valduriez, Kossmann survey]
  - Few statistics, if any.
  - Network behavior issues: latency, burstiness,…
- Solution: adaptive query processing.
  - Stonebraker saw it coming in Ingres.
  - Revivals by Graefe (1993) and DeWitt (1998).
  - Recent ideas: Query scrambling, eddies, adaptive data partitioning, inter-query adaptation.
- Challenge: reduce wasted work.
- *Great opportunity for some theory!*

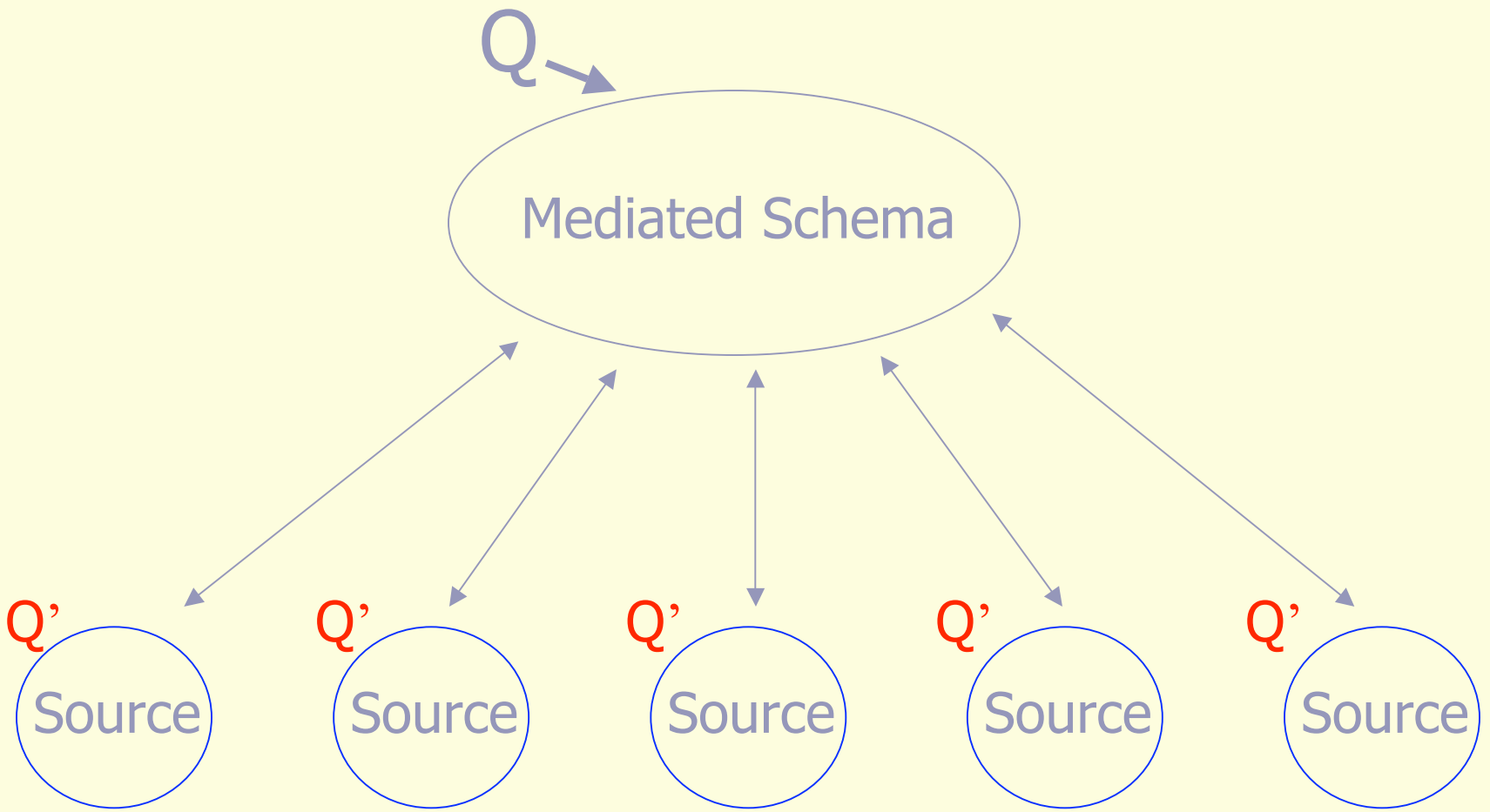
XML Query Processing

- XML = “data integration appetizer”.
- Industry went ahead of research:
  - Nimble, Enosys, XQRL
  - Inspiration from Tukwila, MIX, Strudel/Agora
- (some) Issues:
  - Designing the internal algebra
  - Dealing with evolving XQuery standard
- Our community has served an impressive smorgasbord of XML techniques.

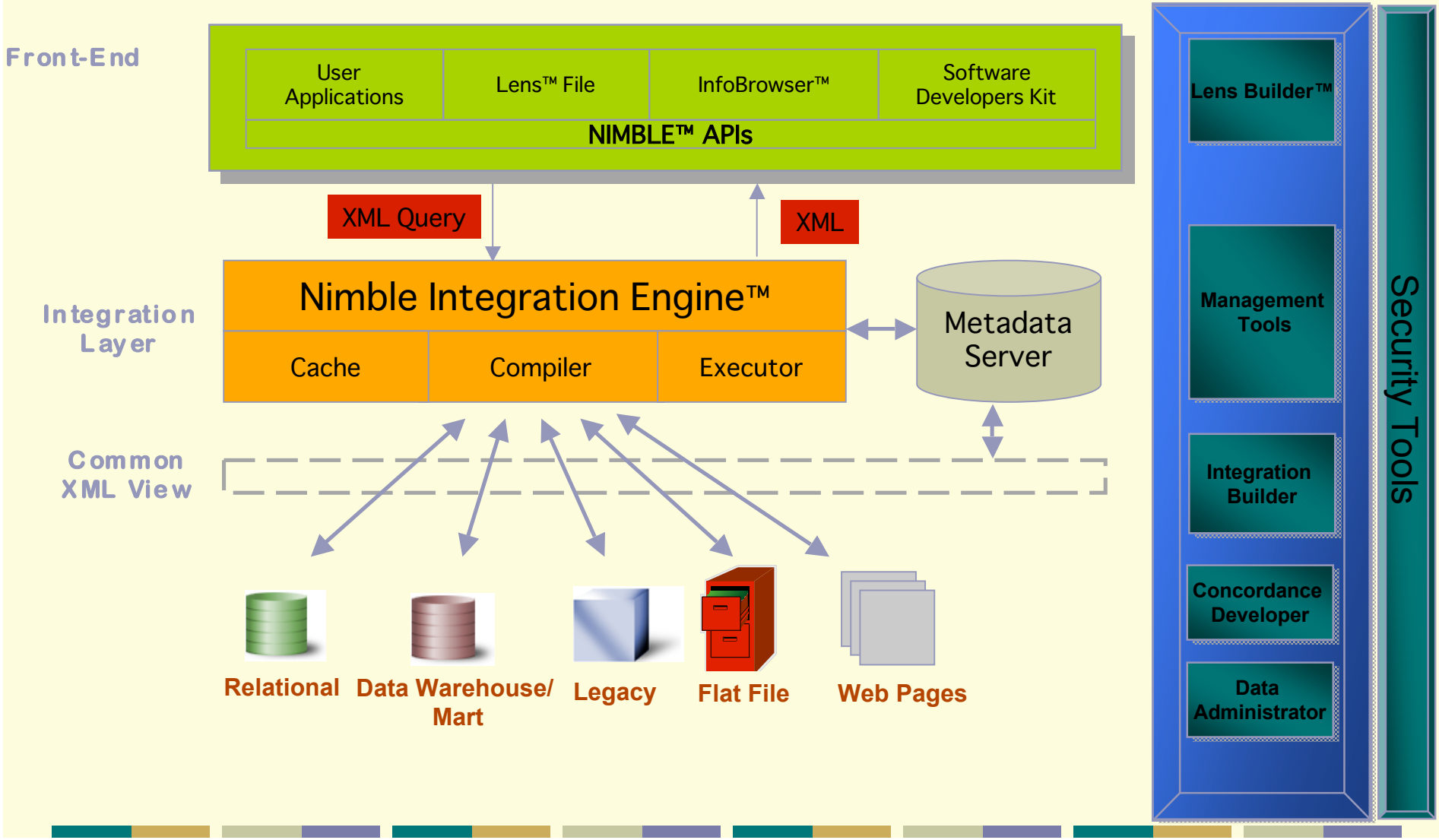
A Few Comments about Commerce

- Until 5 years ago:
  - Data integration = Data warehousing.
- Since then:
  - A wave of startups:
    - Nimble, Enosys, MetaMatrix, Calixa, Composite
  - Big guys made announcements (IBM, BEA).
  - [Delay] Big guys released products.
- Success: analysts have new buzzword – EII
  - New addition to acronym soup (with EAI).
- Lessons:
  - Performance was fine. Need management tools.

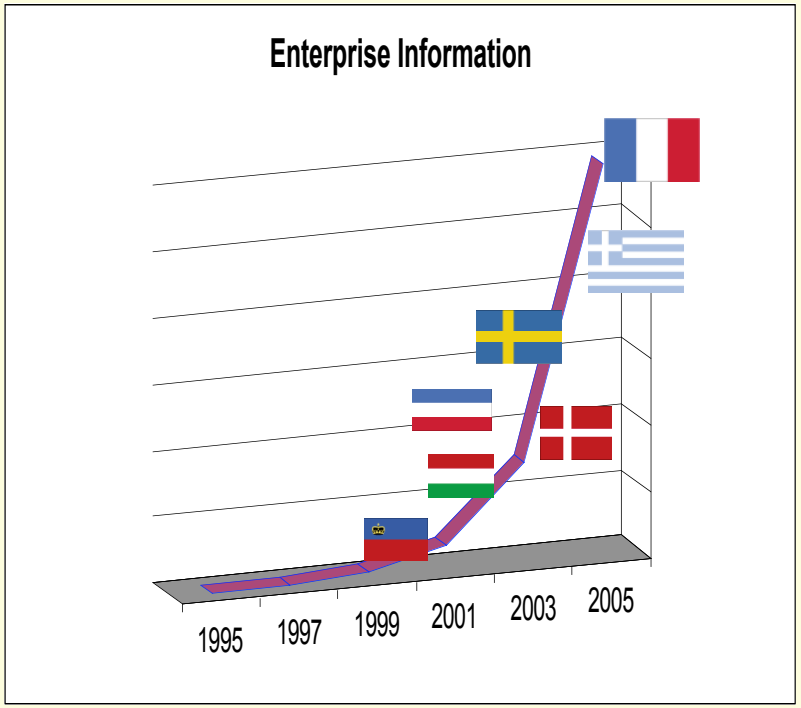
Data Integration: Before



Data Integration: After



Sound Business Models



Source: Gartner, 1999

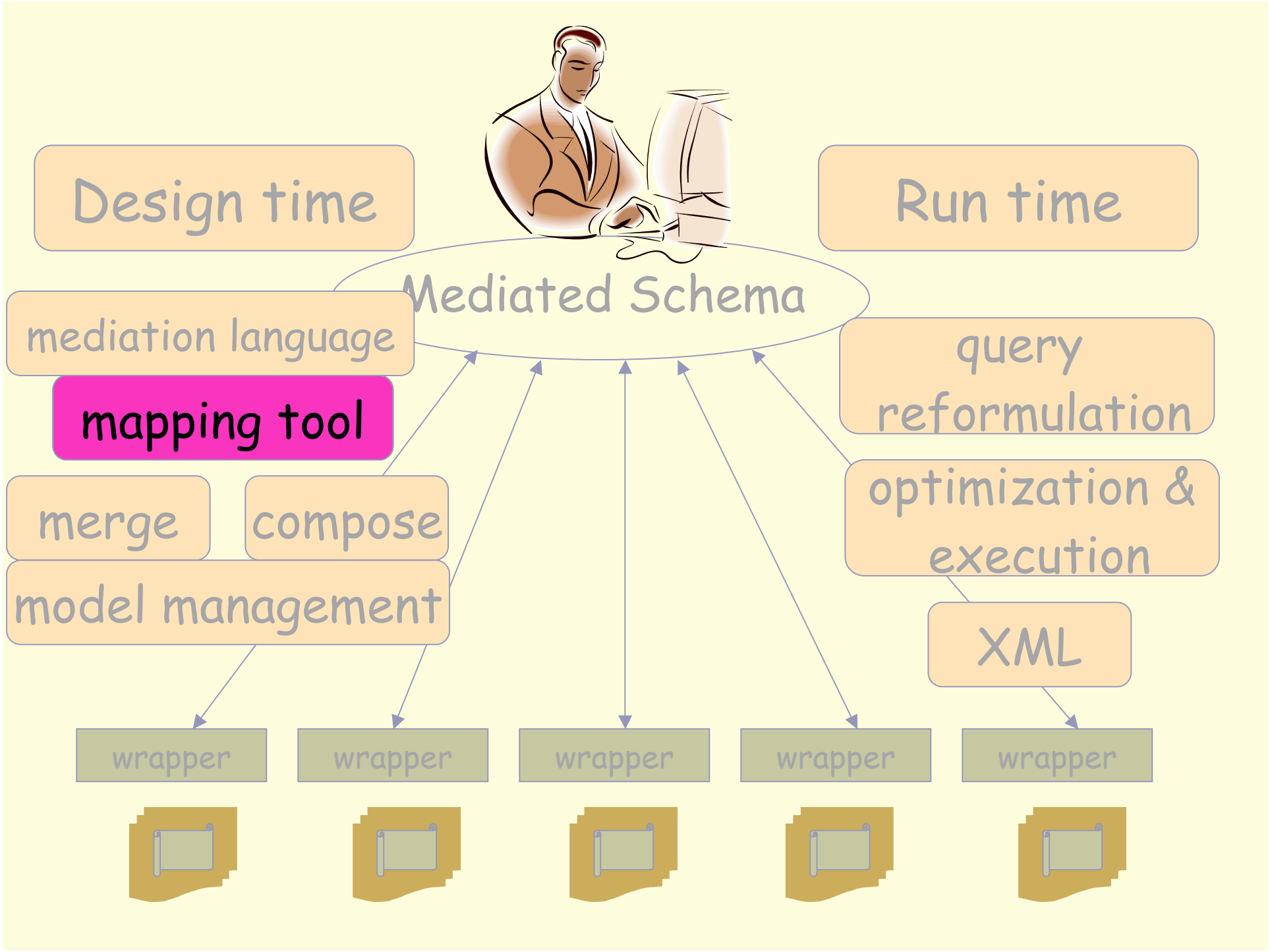
- Explosion of intranet and extranet information
- 80% of corporate information is unmanaged
- By 2004 30X more enterprise data than 1999
- The average company:
  - maintains 49 distinct enterprise applications
  - spends 50% of total IT budget on integration-related efforts

Sound Business Models

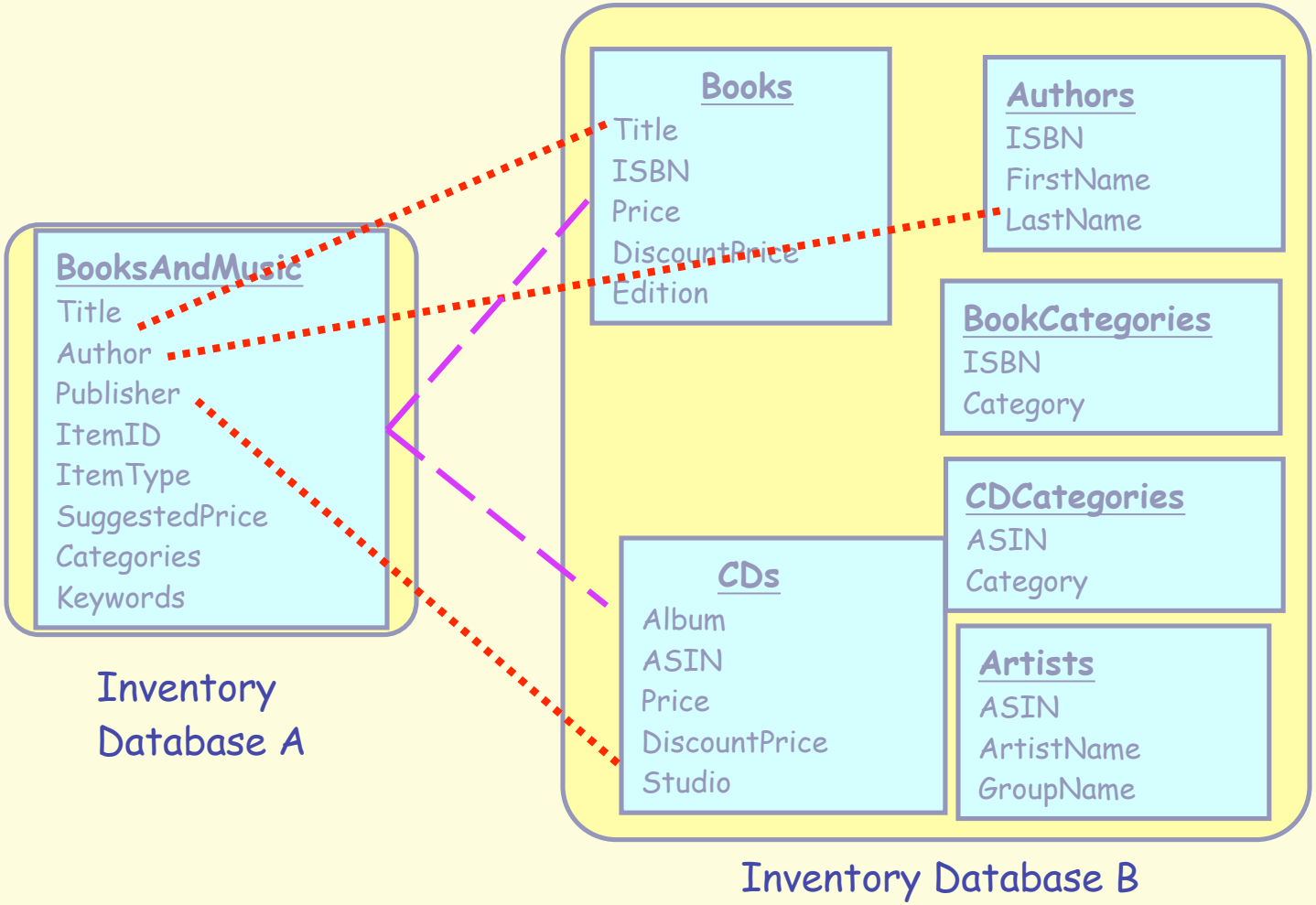
Source: Gartner, 1999

- Explosion of intranet and extranet information
- 80% of corporate information is unmanaged
- By 2004 30X more enterprise data than 1999
- The average company:
  - maintains 49 distinct enterprise applications
  - spends 50% of total IT budget on integration-related efforts

Hockey, eh?

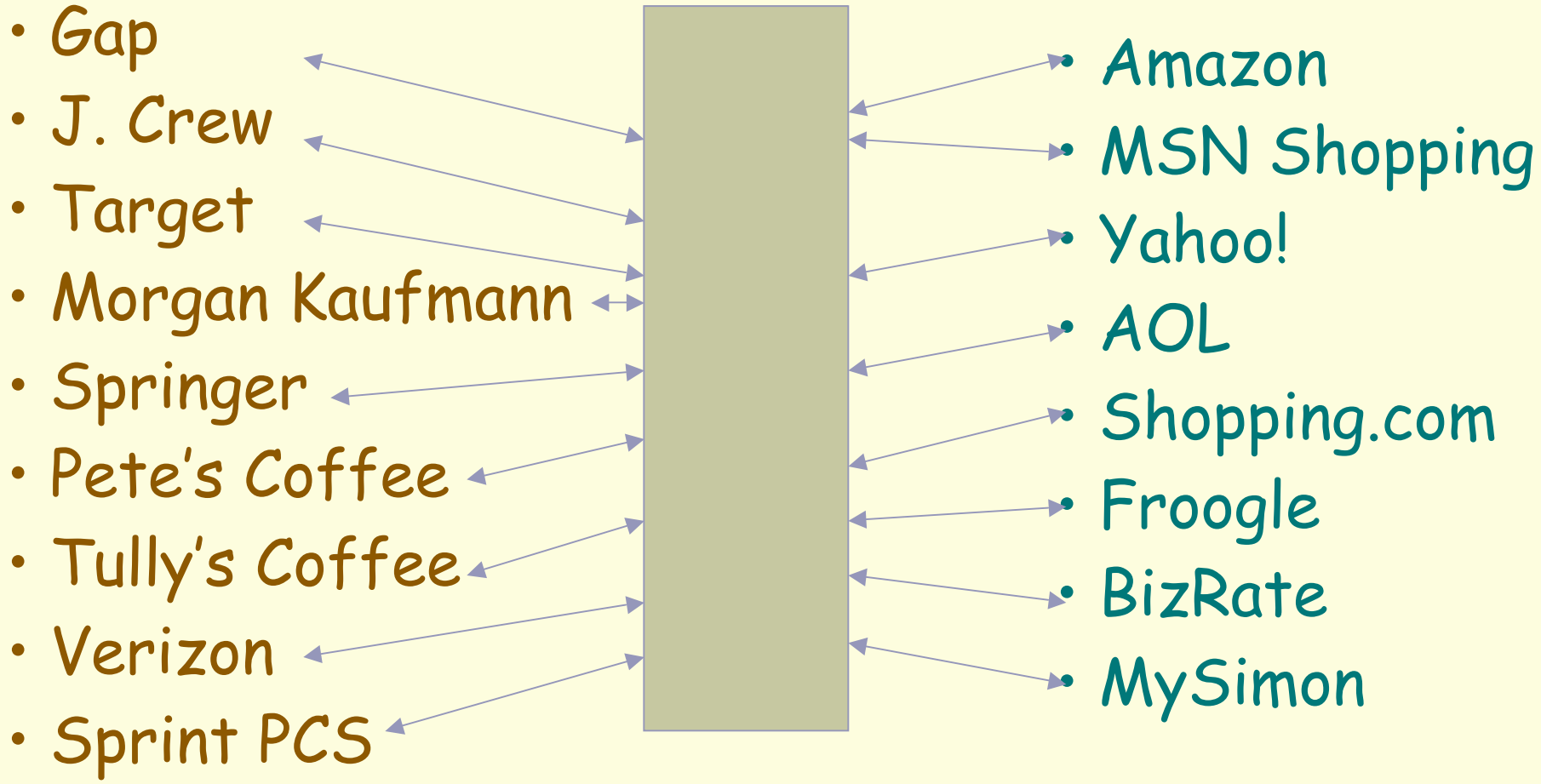

Design time
Run time
Mediated Schema
mediation language
mapping tool
merge
compose
model management
query reformulation
optimization & execution
XML
wrapper
wrapper
wrapper
wrapper
wrapper

Semantic Mappings



"Standards are great, but there are too many of them."

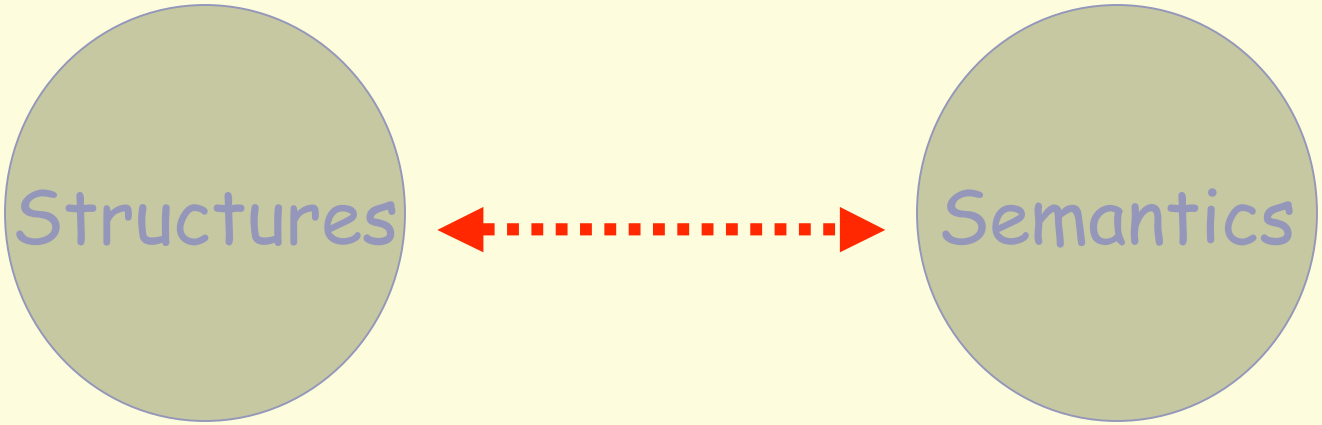
Web Services Intermediary



Mapping taxonomies and fields

Why is it so Hard?

- Schemas never fully capture their intended meaning:
  - They're just symbols and structures.



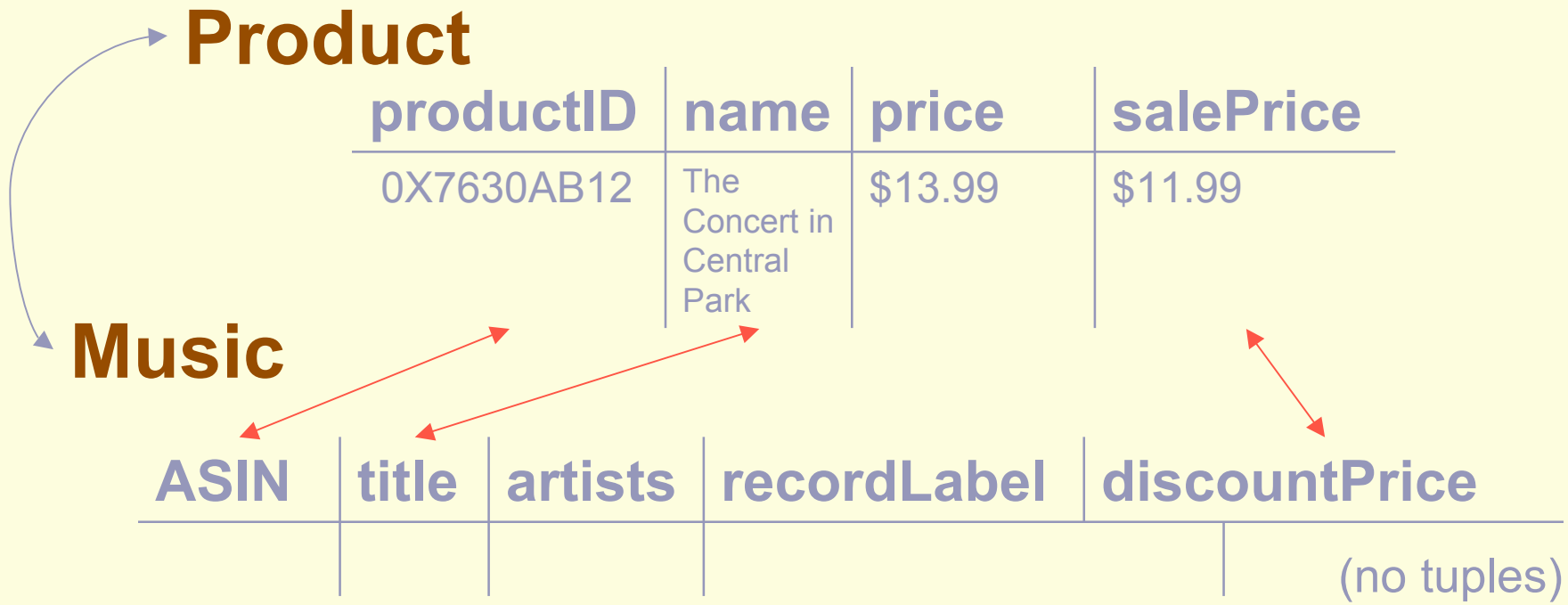
  - *Automatic schema matching is AI-Complete.*
  - Our goal: reduce the human effort.

Comparison-Based Matching

- Build a model for every element in the schema, and compare models.
- Models based on:
  - Names of elements
  - Data types
  - Data instances
  - Text descriptions
  - Integrity constraints

[Survey by Rahm and Bernstein, VLDBJ 2001]

Insufficient Information



Key Hypothesis: SSS

- Statistics offer clues about the semantics of the structures.
- Statistics can be gleaned from collections of schemas and mappings:
  - A collection provides *alternate* representations of domain concepts.
- Human experts do this unconsciously.

Obtaining More Evidence

Product CD

productID prodID	name albumName	price	salePrice
0X7630AB12	The Concert in Central Park	$13.99	$11.99

Corpus

MusicCD

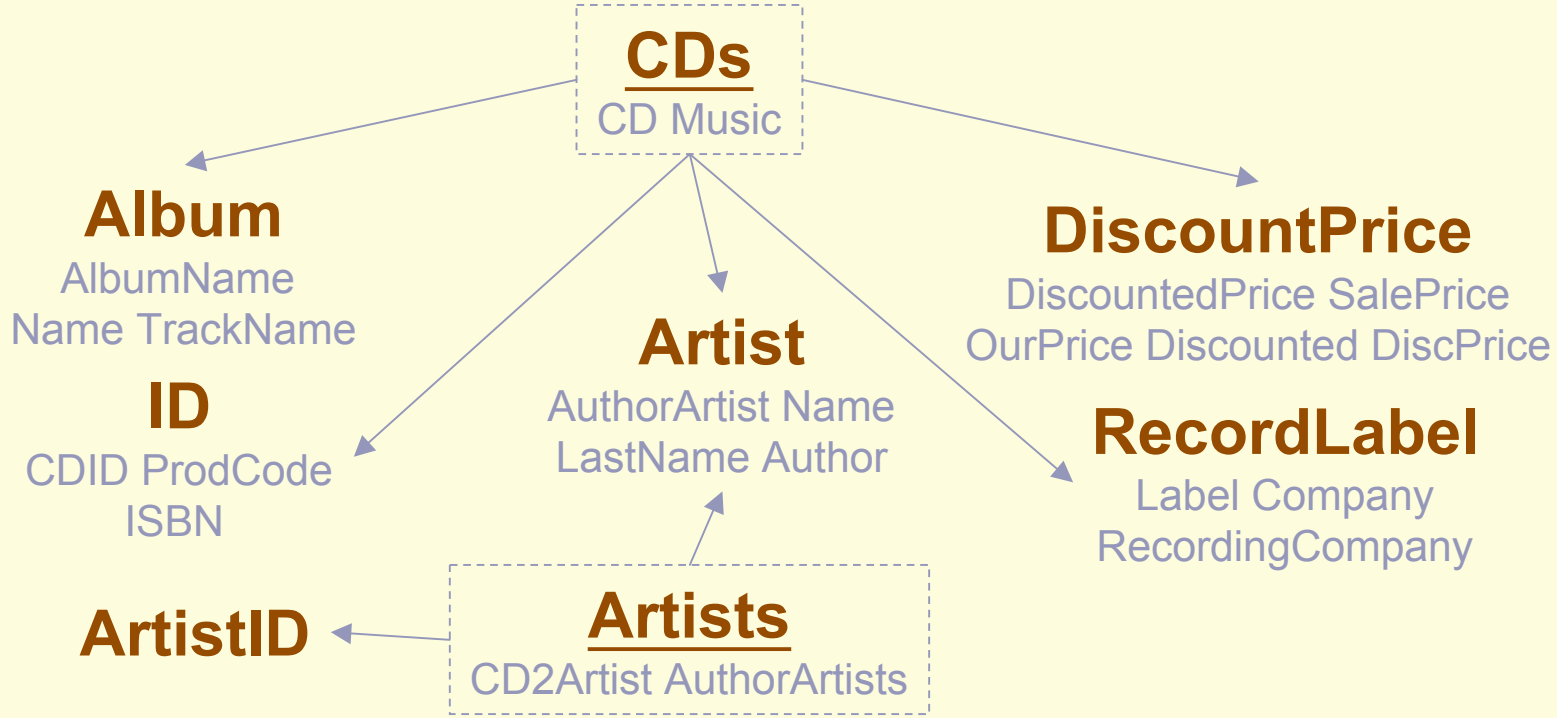
ASIN	album	artistName	price	discountPrice
4Y3026DF23	The Best of the Doors	The Doors	$16.99	$12.99

CD

prodID	albumName	artists	recordCompany	price	salePrice
9R4374FG56	Saturday Night Fever	The Bee Gees	Columbia	$14.99	$9.99

Comparing with More Evidence

Corpus Statistics: Alternate Concept Representations



- Learn alternate names, data instances, names of related elements, data types, …

Statistics: Schema Design Patterns

- Relations between elements

Schema element dependency

CDs → price

discountPrice → price

numEmployees → manager

zipcode → Warehouses

fax → telephone

city → state

Frequently co-occurring concepts

(Warehouse, warehouseID, manager, telephone, fax)

(Availability, Books, CDs, Warehouses)

- Tables and likely columns

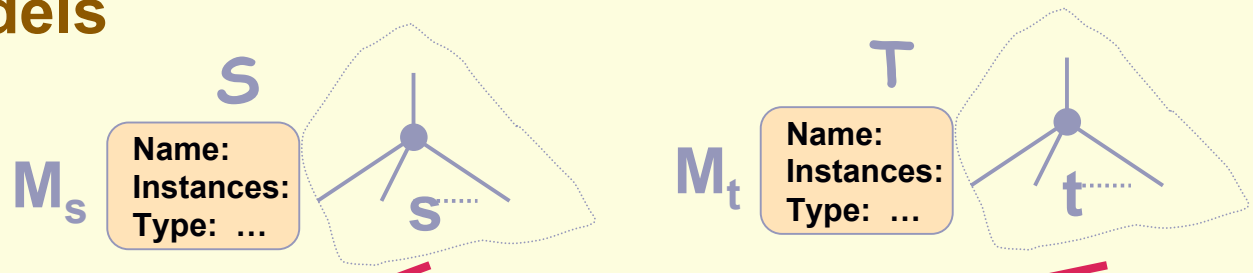
Table/column	Likely column/table	Other column/table
Warehouses	warehouseID, telephone, fax, manager, streetAddress, city	state, zip, numEmployees, capacity
title	Books	
isbn	Books, Availability	Keywords, Authors

The Corpus: Details

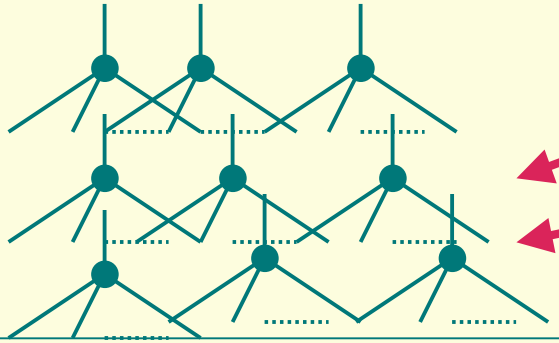
[Madhavan et al.]

- Learn model *ensemble* for each corpus element
  - Use names, data instances, types, structure, …
  - Model ensemble: combine a set of base learners.
- Training data:
  - Gleaned from schemas and mappings.
  - Positive examples:
    - Elements of the schema itself.
    - From previous mappings: *mapping reuse*.
- For learning dependencies:
  - Cluster elements in the corpus into concepts.

1. Build initial models



M_s M_t



Corpus of schemas and mappings

2. Find similar elements in corpus

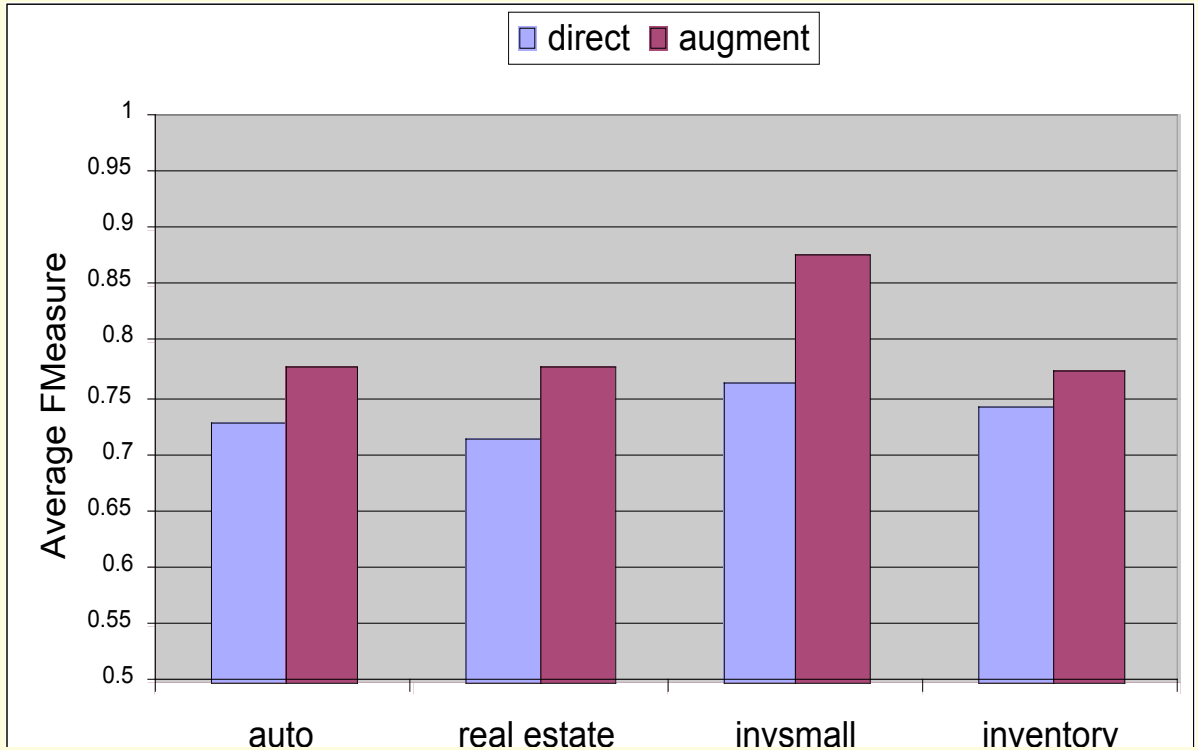
3. Build augmented models

M'_s M'_t

4. Match using augmented models

5. Use additional statistics (IC's) to refine match

Matching Performance



- 16-19 schemas and 6 mappings in the corpus
- 22-54 schema pairs being tested
- Results even better on *hard* matching tasks.

Examples of SSS Work

- [Doan et al.]: generalizing from manual mappings.
- [He and Chang]: creating a mediated schema for web form domains.
- [Kushmerick et al.]: classifying web forms.
- [Dong et al.]: Similarity search for web services in *Woogle* (This afternoon).
- Reformulating queries on unknown databases.
- Searching class libraries, deep web, enterprise data sources.

SSS Challenges

- Building corpora. In steps?
  - *Shell* of 1000
  - *Outer shell* of 10,000
  - The grand expanse of 450,000?
- Engineering a corpus:
  - Tuning: removing noise, selecting content.
  - Domain specific?

Design time

Run time

Mediated Schema

logic

mapping tool

merge

compose

model management

query reformulation

optimization & execution

wrapper

wrapper

"Corpus manager"
corpus reuse, cross domain?
Cross organization?

SSS Challenges

- Building corpora. In steps?
  - *Shell* of 1000
  - *Outer shell* of 10,000
  - The grand expanse of 300,000?
- Engineering a corpus:
  - Tuning: removing noise, selecting content.
  - Domain specific?
- Theory? What are we learning?
- Industry: www.transformic.com

Design time

Run time

Mediated Schema

mediation language

mapping tool

merge

compose

model management

query reformulation

optimization & execution

XML

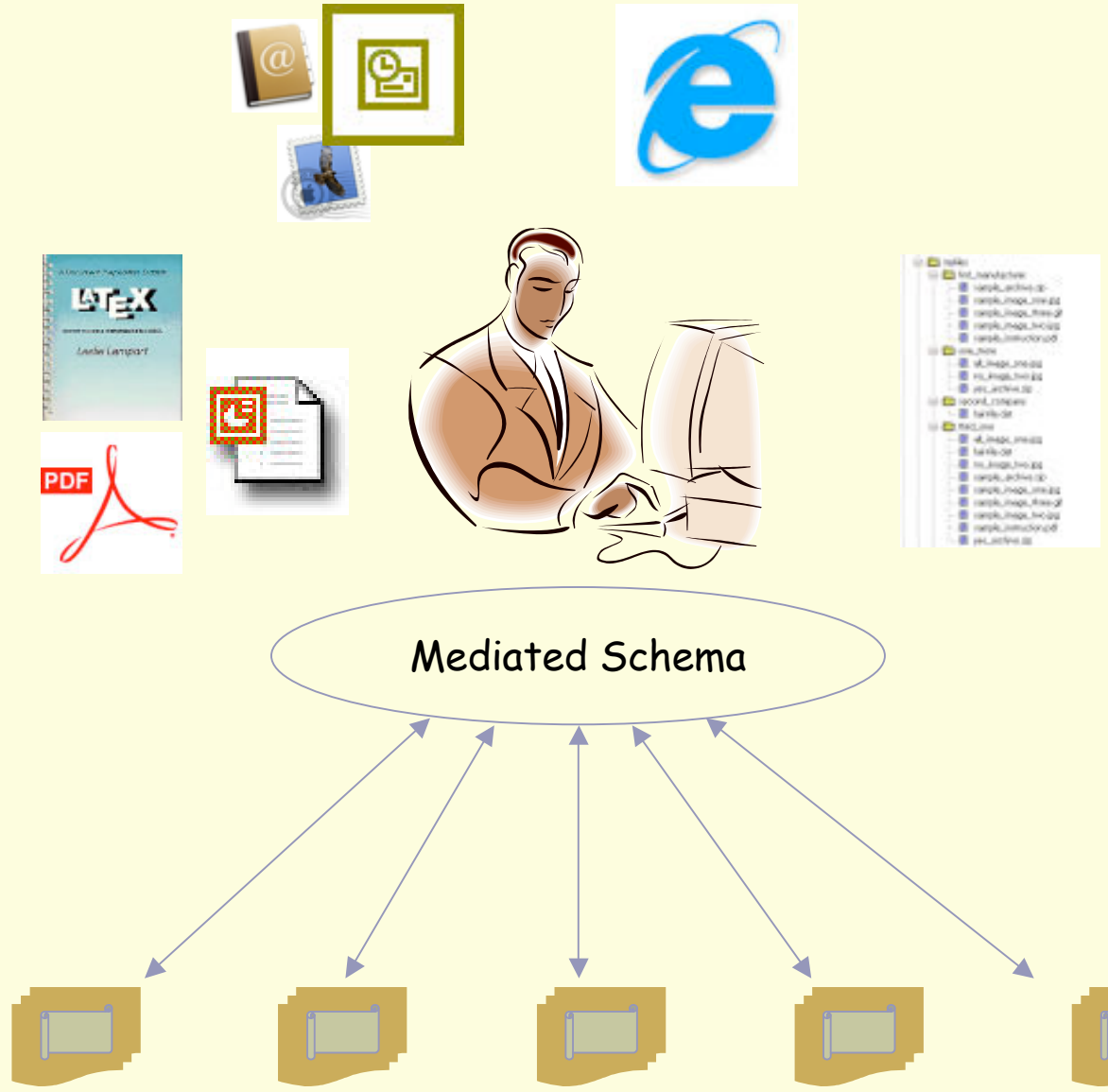
wrapper

wrapper

wrapper

wrapper

wrapper

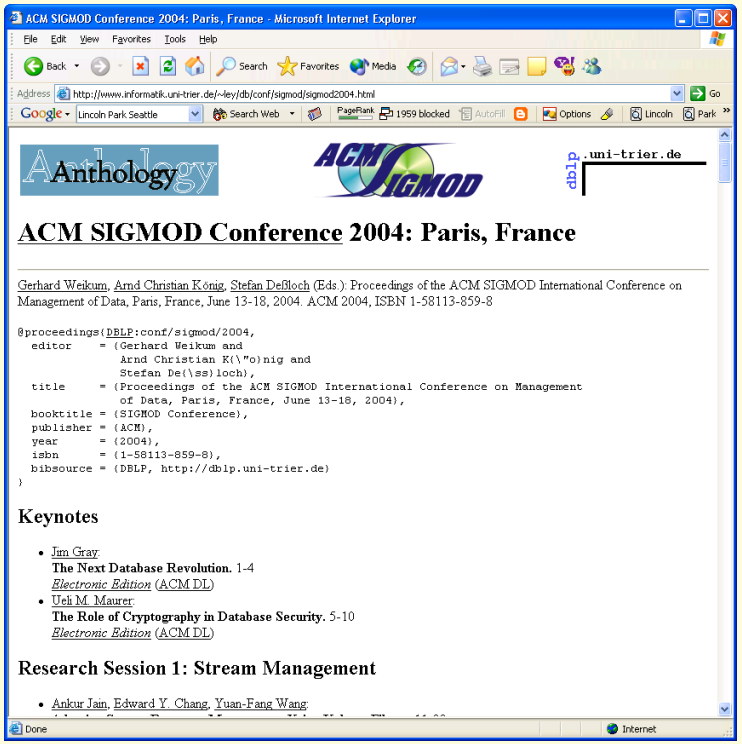

Mediated Schema

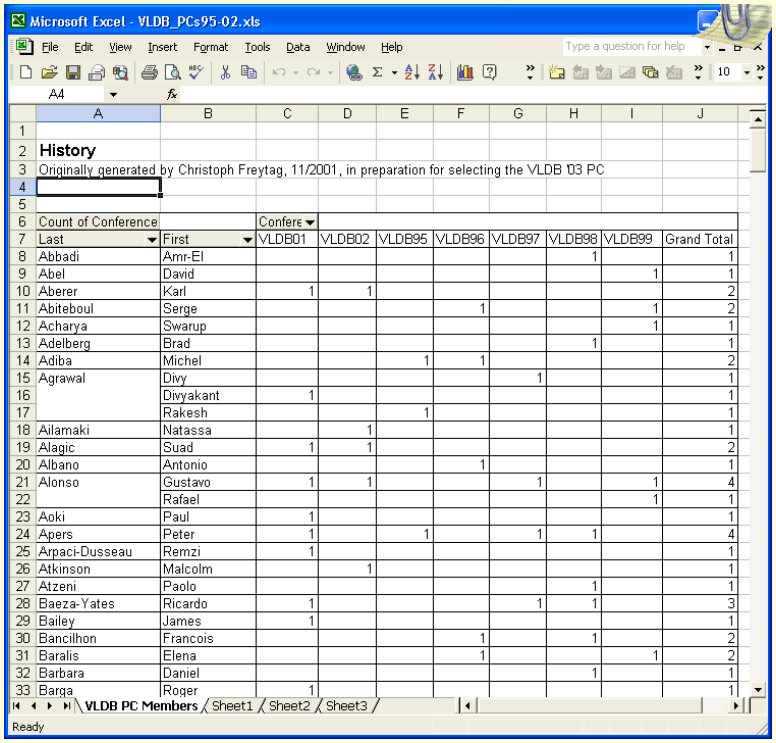
Hard to Find Information

- Find my VLDB04 paper; and the PowerPoint (maybe in an attachment).
- Find emails from my Californian friends.
- Which Ozsu paper did I cite in my VLDB04 paper?
- What quarter was Mary in my class and what grade did she get?
- Which experiment did I run with *NF1* and which emails discussed them?

On-the-Fly Data Integration

Who published at SIGMOD but was not recently on the PC?





The Deeper Issue

- We're not integrated in the user's habitat.

"*The most profound technologies are those that disappear*". Mark Weiser

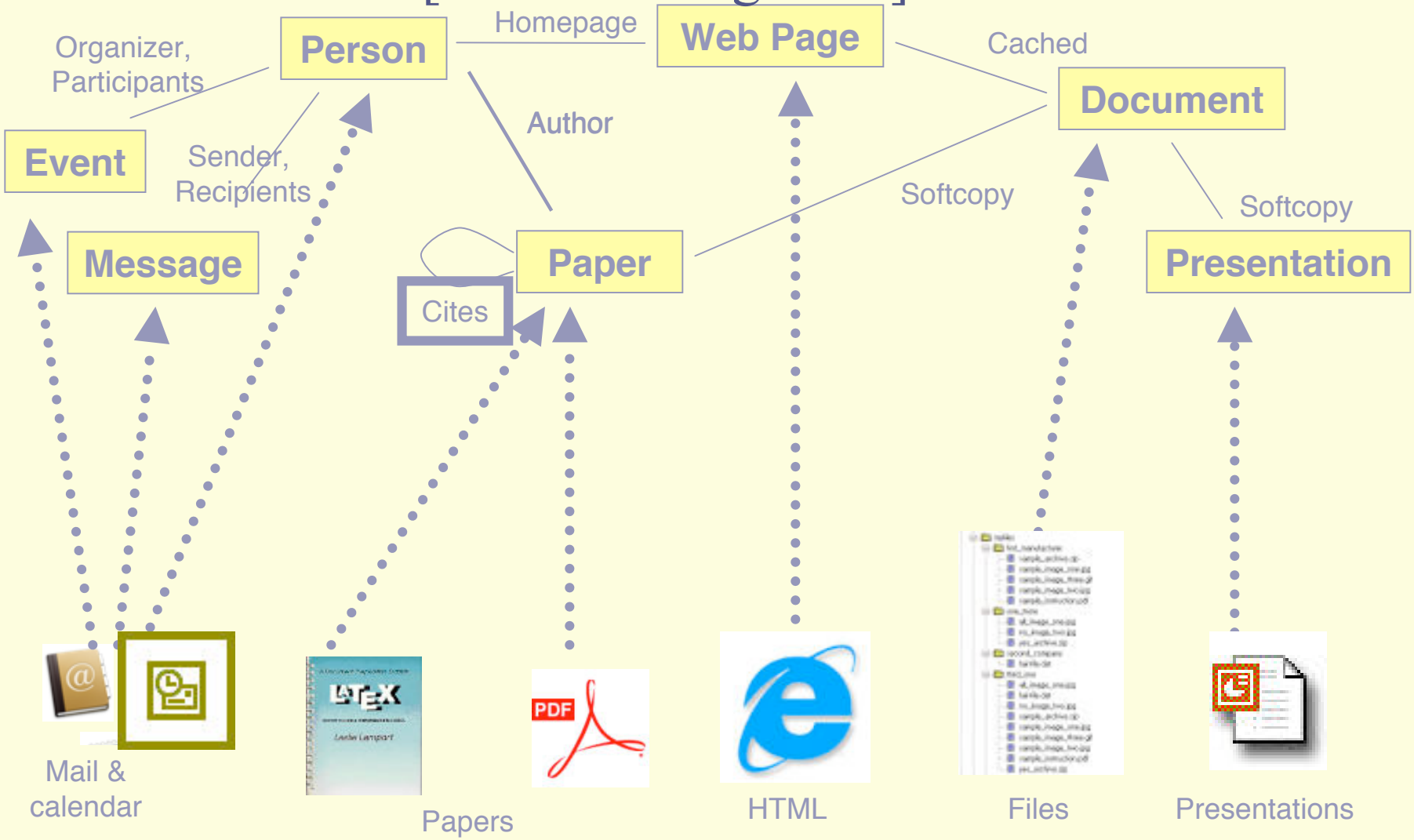
- But *we* are very visible:
  - Schema always comes first
  - Dichotomy between structured and unstructured data (the *structure chasm*)
  - Integration: only for high-volume needs.

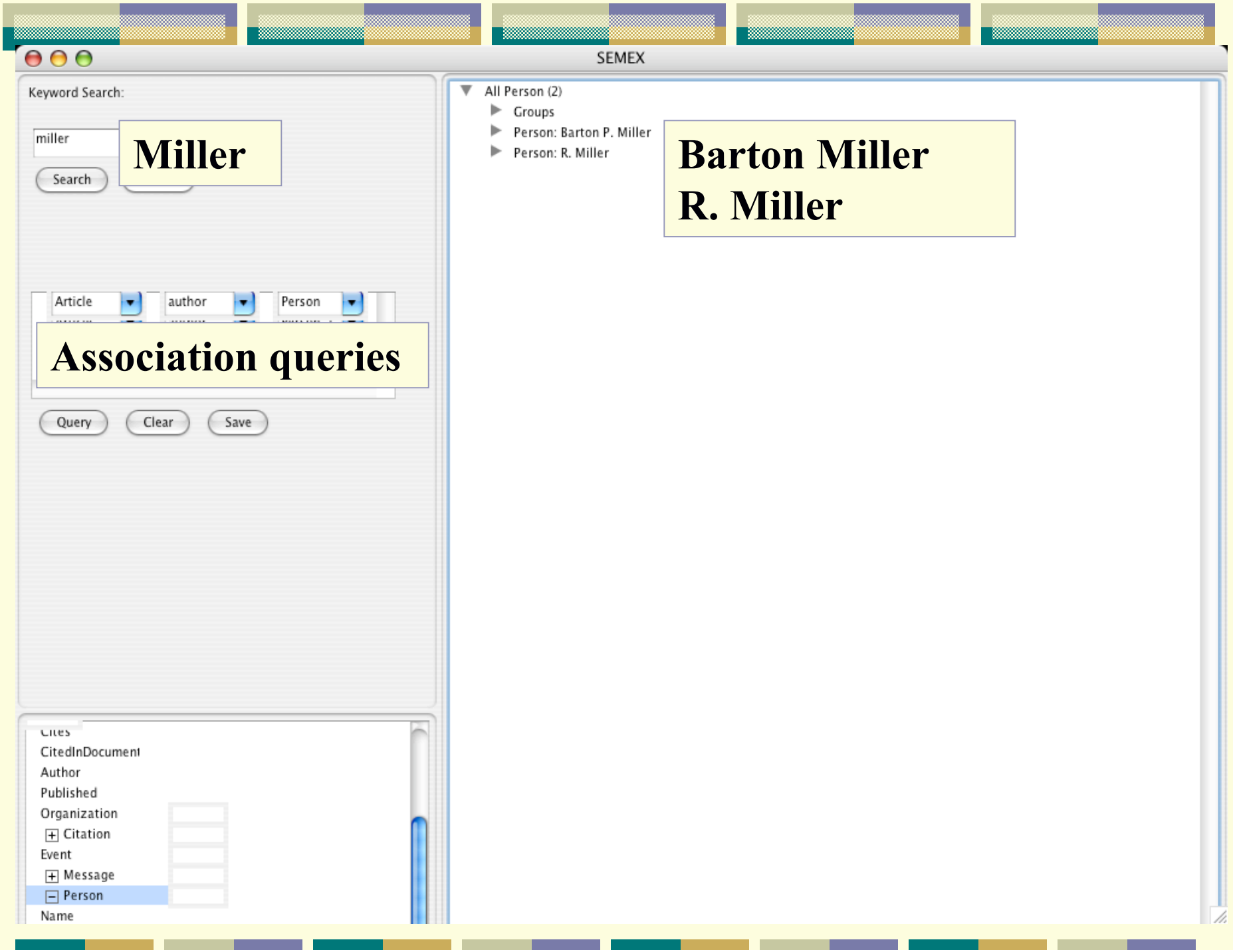
User-Centered Data Management

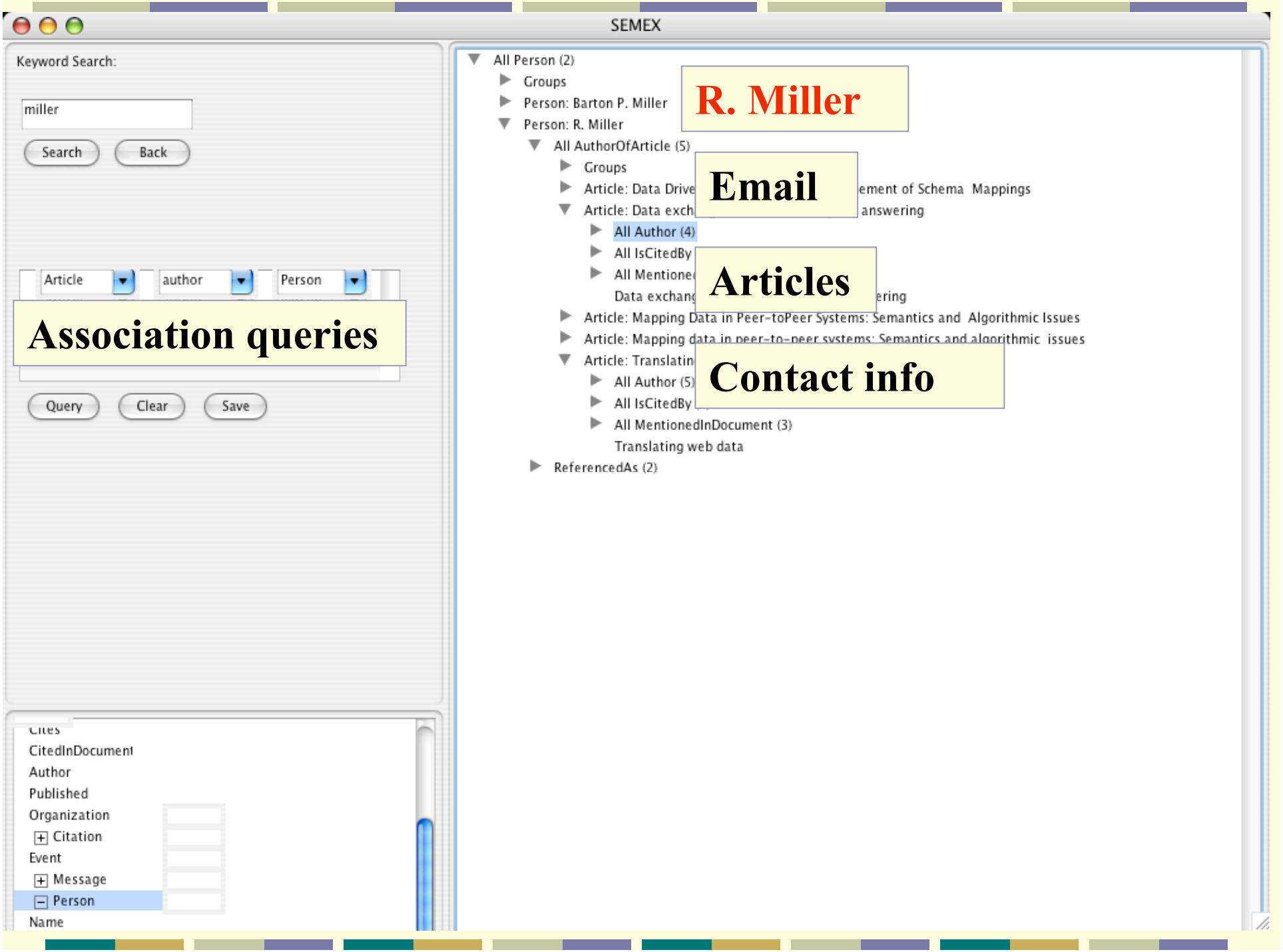
- Bring the benefits of data management to users, in their own habitat.
- Create *associations* between disparate objects and data sources.
- Create an ecology of cooperating personal Memex's.

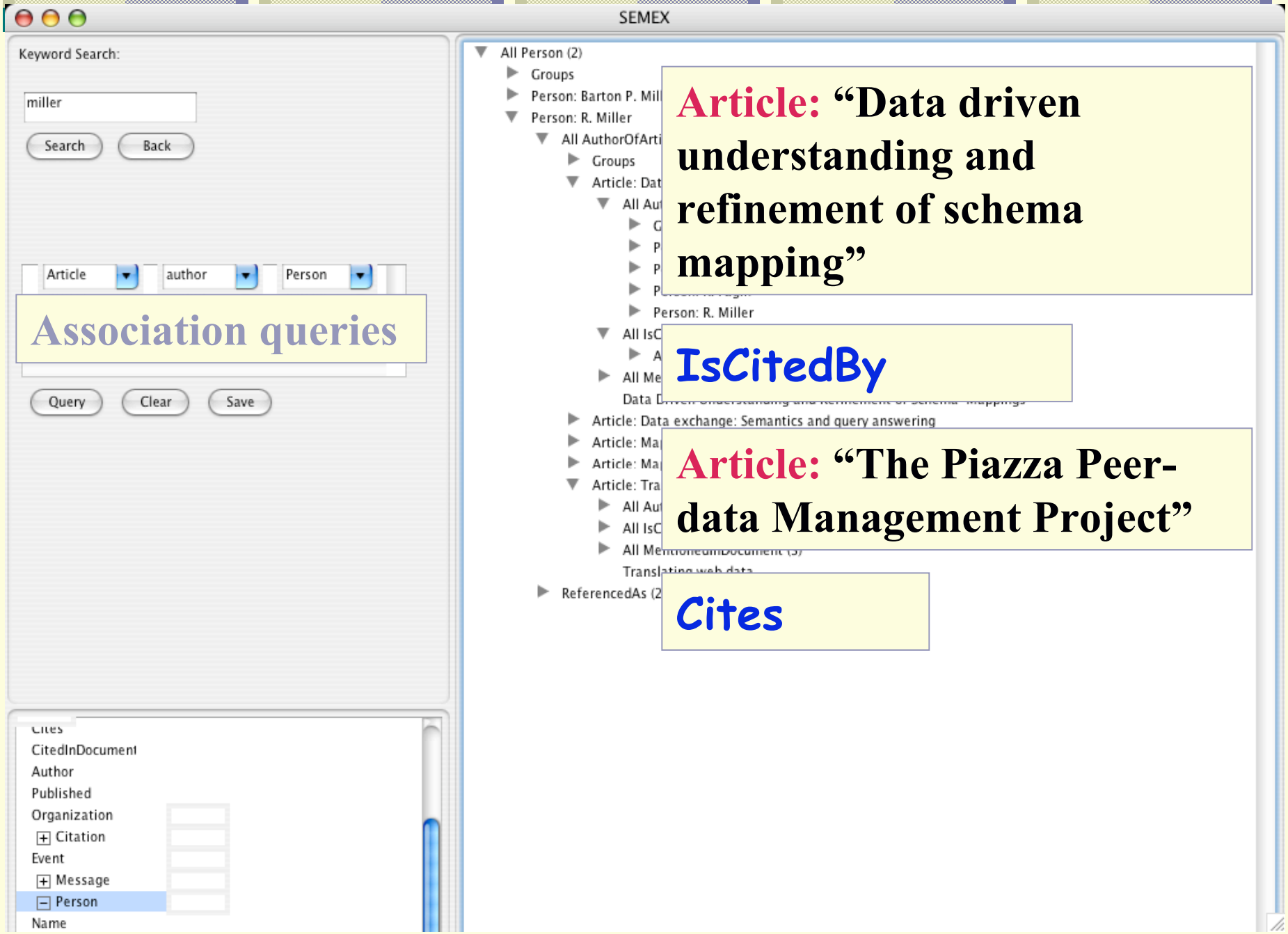
Semantic Explorer (Semex)

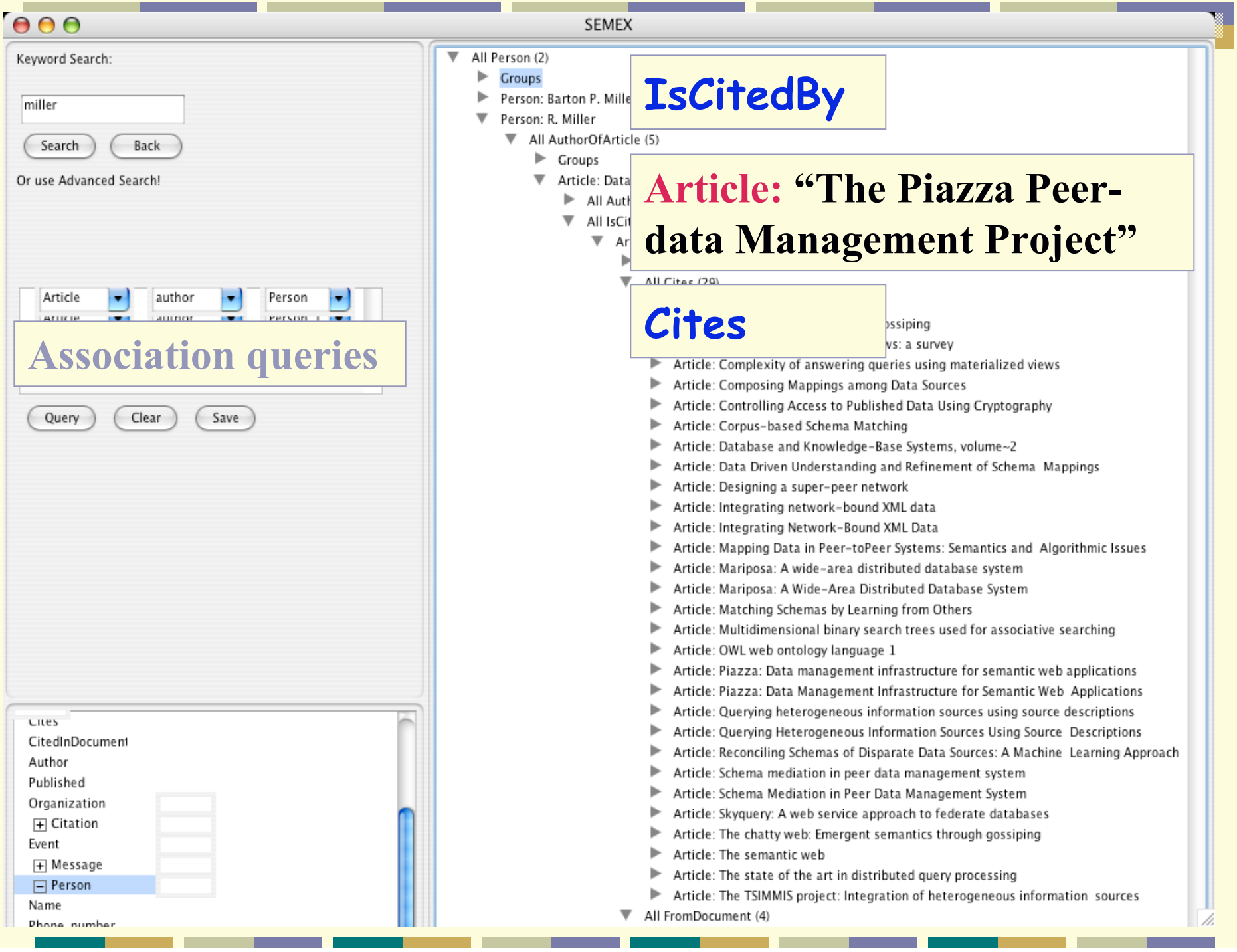
[Semex: Dong et al.]

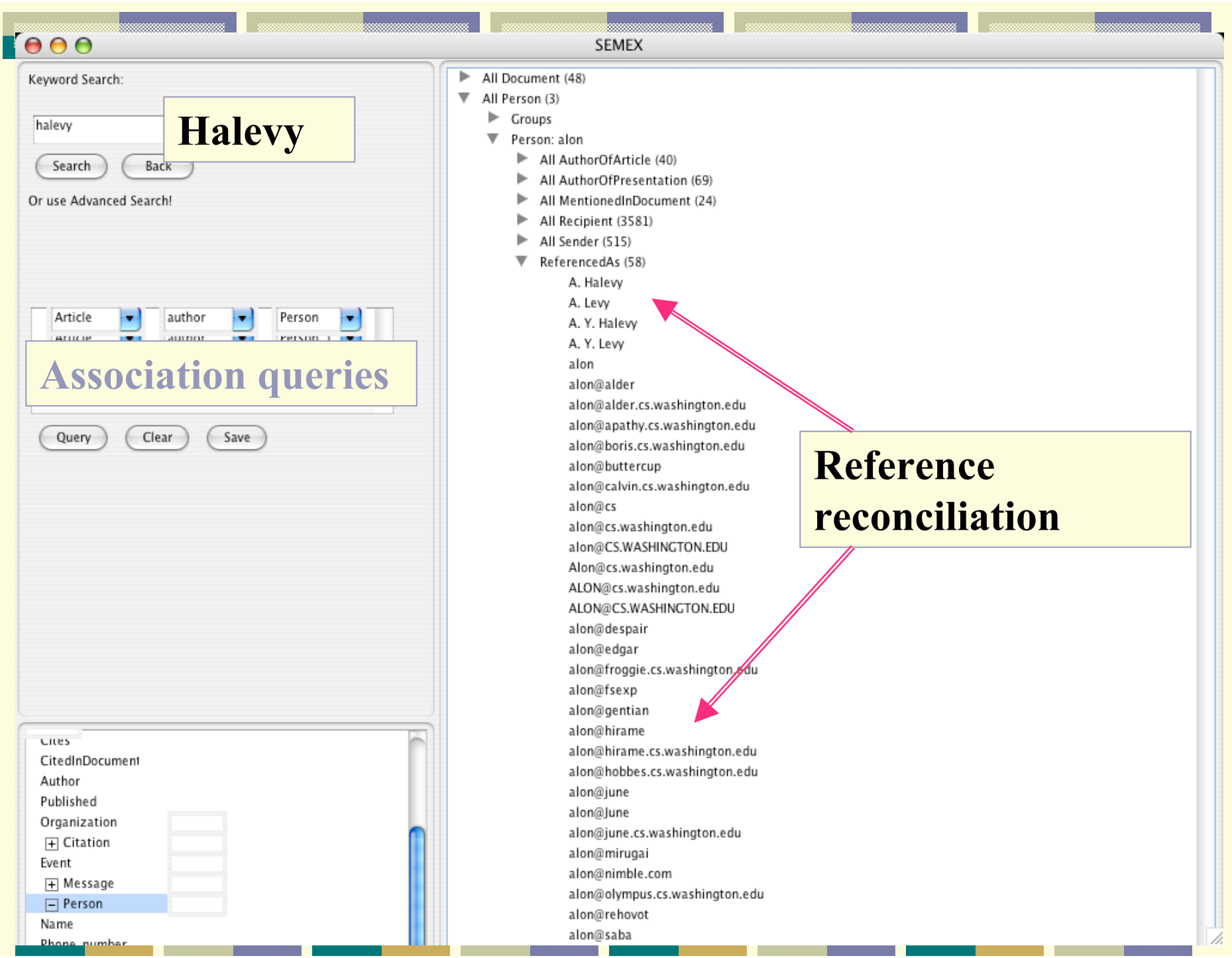



SEMEX
Keyword Search:
miller
Search
Miller
Article
author
Person
Association queries
Query
Clear
Save
CitedInDocument
Author
Published
Organization
Citation
Event
Message
Person
Name
All Person (2)
Groups
Person: Barton P. Miller
Person: R. Miller
Barton Miller
R. Miller

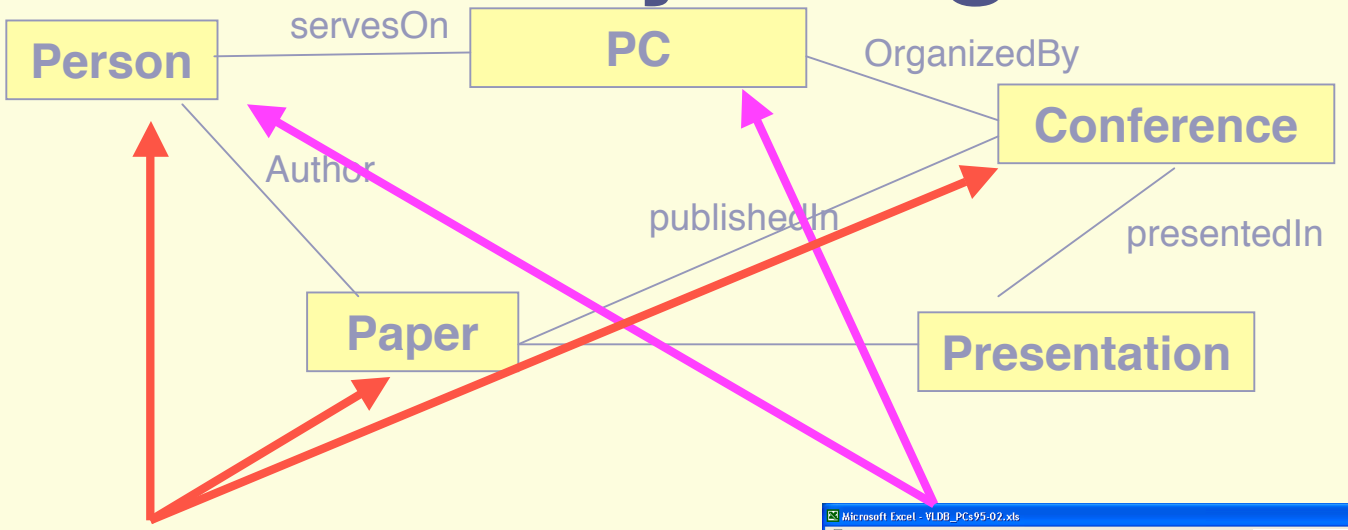

SEMEX
Keyword Search:
miller
Search
Back
Article
author
Person
Association queries
Query
Clear
Save
CitedInDocument
Author
Published
Organization
Citation
Event
Message
Person
Name
All Person (2)
Groups
Person: Barton P. Miller
Person: R. Miller
All AuthorOfArticle (5)
Groups
Article: Data Drive
ement of Schema Mappings
Article: Data exch
answering
All Author (4)
All IsCitedBy
All Mentione
Data exchang
ering
Article: Mapping Data in Peer-toPeer Systems: Semantics and Algorithmic Issues
Article: Mapping data in peer-to-peer systems: Semantics and algorithmic issues
Article: Translatin
All Author (5)
All IsCitedBy
All MentionedInDocument (3)
Translating web data
ReferencedAs (2)
R. Miller
Email
Articles
Contact info

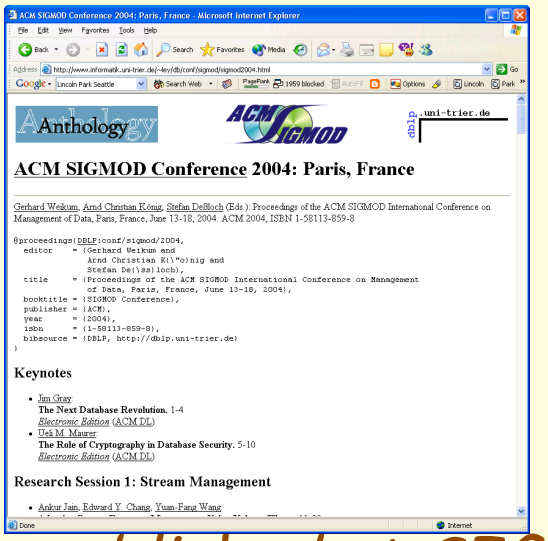

SEMEX
Keyword Search:
miller
Search
Back
Article
author
Person
Association queries
Query
Clear
Save
CitedInDocument
Author
Published
Organization
Citation
Event
Message
Person
Name
All Person (2)
Groups
Person: Barton P. Mil
Person: R. Miller
Groups
Person: R. Miller
Article: Data exchange: Semantics and query answering
ReferencedAs (2
Article: "Data driven understanding and refinement of schema mapping"
IsCitedBy
Article: "The Piazza Peer-data Management Project"
Cites

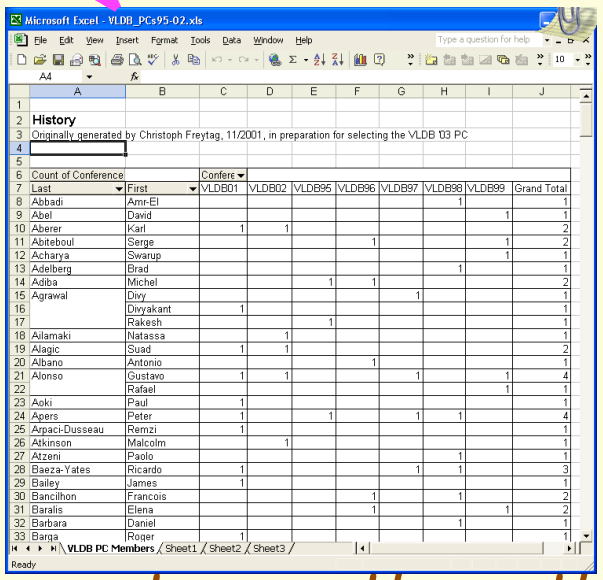

SEMEX
Keyword Search:
miller
Search
Back
Or use Advanced Search!
Article
author
Person
Association queries
Query
Clear
Save
Cites
CitedInDocument
Author
Published
Organization
Citation
Event
Message
Person
Name
All Person (2)
Groups
Person: Barton P. Mille
Person: R. Miller
All AuthorOfArticle (5)
Groups
Article: Data
All Auth
All IsCit
Ar
All Cites (29)
IsCitedBy
Article: "The Piazza Peer-data Management Project"
Cites
ossiping
vs: a survey
Article: Complexity of answering queries using materialized views
Article: Composing Mappings among Data Sources
Article: Controlling Access to Published Data Using Cryptography
Article: Corpus-based Schema Matching
Article: Database and Knowledge-Base Systems, volume~2
Article: Data Driven Understanding and Refinement of Schema Mappings
Article: Designing a super-peer network
Article: Integrating network-bound XML data
Article: Integrating Network-Bound XML Data
Article: Mapping Data in Peer-toPeer Systems: Semantics and Algorithmic Issues
Article: Mariposa: A wide-area distributed database system
Article: Mariposa: A Wide-Area Distributed Database System
Article: Matching Schemas by Learning from Others
Article: Multidimensional binary search trees used for associative searching
Article: OWL web ontology language 1
Article: Piazza: Data management infrastructure for semantic web applications
Article: Piazza: Data Management Infrastructure for Semantic Web Applications
Article: Querying heterogeneous information sources using source descriptions
Article: Querying Heterogeneous Information Sources Using Source Descriptions
Article: Reconciling Schemas of Disparate Data Sources: A Machine Learning Approach
Article: Schema mediation in peer data management system
Article: Schema Mediation in Peer Data Management System
Article: Skyquery: A web service approach to federate databases
Article: The chatty web: Emergent semantics through gossiping
Article: The semantic web
Article: The state of the art in distributed query processing
Article: The TSIMMIS project: Integration of heterogeneous information sources
All FromDocument (4)


SEMEX
Keyword Search:
halevy
Search
Back
Halevy
Or use Advanced Search!
Article
author
Person
Association queries
Query
Clear
Save
Cites
CitedInDocument
Author
Published
Organization
Citation
Event
Message
Person
Name
All Document (48)
All Person (3)
Groups
Person: alon
All AuthorOfArticle (40)
All AuthorOfPresentation (69)
All MentionedInDocument (24)
All Recipient (3581)
All Sender (515)
ReferencedAs (58)
A. Halevy
A. Levy
A. Y. Halevy
A. Y. Levy
alon
alon@alder
alon@alder.cs.washington.edu
alon@apathy.cs.washington.edu
alon@boris.cs.washington.edu
alon@buttercup
alon@calvin.cs.washington.edu
alon@cs
alon@cs.washington.edu
alon@CS.WASHINGTON.EDU
Alon@cs.washington.edu
ALON@cs.washington.edu
ALON@CS.WASHINGTON.EDU
alon@despair
alon@edgar
alon@froggie.cs.washington.edu
alon@fsexp
alon@gentian
alon@hirame
alon@hirame.cs.washington.edu
alon@hobbes.cs.washington.edu
alon@june
alon@June
alon@june.cs.washington.edu
alon@mirugai
alon@nimble.com
alon@olympus.cs.washington.edu
alon@rehovot
alon@saba
Reference reconciliation

On-The-Fly Integration







Who published at SIGMOD but was not recently on the PC?

Principles of User-Centered DM

- Create associations for the user:
  - Every data item should be associated with others.
- Adapt to and learn about the user:
  - Use statistics to leverage previous user activities!
- Manipulate *any* kind of data.
- Data and "schema" evolve over time:
  - life-long data management.
  - Modeling and querying may not be precise anymore.

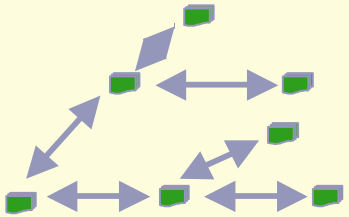
Summary

mapping
authoring
querying

Structures

Statistics

Semantics



model management

adaptive QP

mediation language

reformulation

XML

wrappers

Acknowledgements

- Phil Bernstein
- Anhai Doan (T)
- Pedro Domingos (F)
- Oren Etzioni (F)
- Mary Fernandez
- Zack Ives (T)
- Dan Suciu (F)
- Dan Weld (F)
- Luna Dong (J)
- Jayant Madhavan (J)
- Luke McDowell (J)
- Peter Mork (J)
- Rachel Pottinger (T)
- Divesh Srivastava
- Igor Tatarinov
- *NSF*

Some References

- www.cs.washington.edu/homes/alon
- Piazza: ICDE03, WWW03, VLDB-03, SIGMOD-04
- SSS: [Madhavan, forthcoming], VLDB-04.
- Semex: IIWeb-04
- Surveys on schema matching languages:
  - Halevy, VLDB Journal 01
  - Lenzerini, PODS 2002
- Teaching integration to undergraduates:
  - SIGMOD Record, September, 2003.