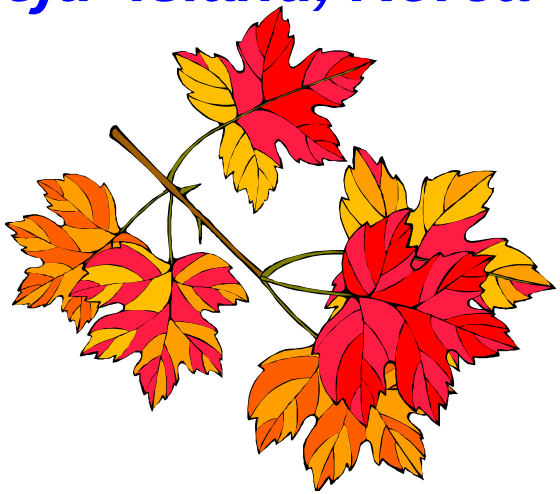
Data Semantics Revisited: Databases and the Semantic Web

John Mylopoulos
University of Toronto

DASFAA'04, March 17-19, 2004
Jeju Island, Korea



© 2004 John Mylopoulos

Next Generation Database Systems Won't Work Without Semantics!

Panel Discussion
ACM SIGMOD'98,
Seattle, June 4, 1998

...1998...

Data Semantics Can't Fail This Time!

Panel Discussion
CAiSE'98, Pisa
June 11, 1998

© 2004 John Mylopoulos

The Panelists

Panel:

Philip Bernstein (Microsoft)

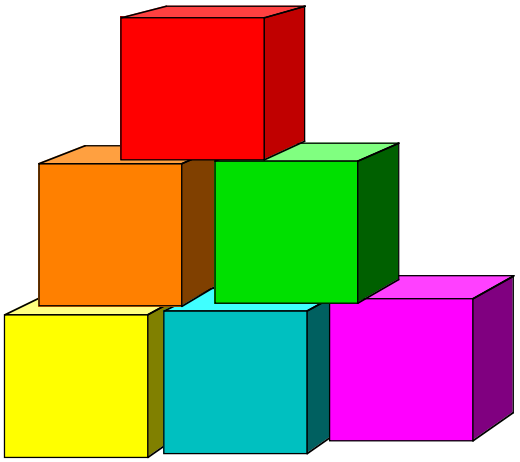
Umesh Dayal (HP Laboratories)

Sham Navathe (Georgia Tech)

Marek Rusinkiewicz (MCC)

Panel chair:

John Mylopoulos (Univ. of Toronto)



Panel:

Michael Brodie (GTE Labs)

Stefano Ceri (Politecnico di Milano)

Arne Solvberg (Univ. of Trondheim)

Panel chair:

John Mylopoulos (Univ. of Toronto)

© 2004 John Mylopoulos

"…The three most important problems in Databases used to be

Performance, Performance and Performance;

in the future, the three most important problems will be

Semantics, Semantics and Semantics…"

(paraphrase) Stefano Ceri

June 11, 1998

© 2004 John Mylopoulos

This Talk

- *Data Semantics: The problem and its history*
- *The Semantic Web: The vision and the challenges*
- *Towards a new theory of data semantics*

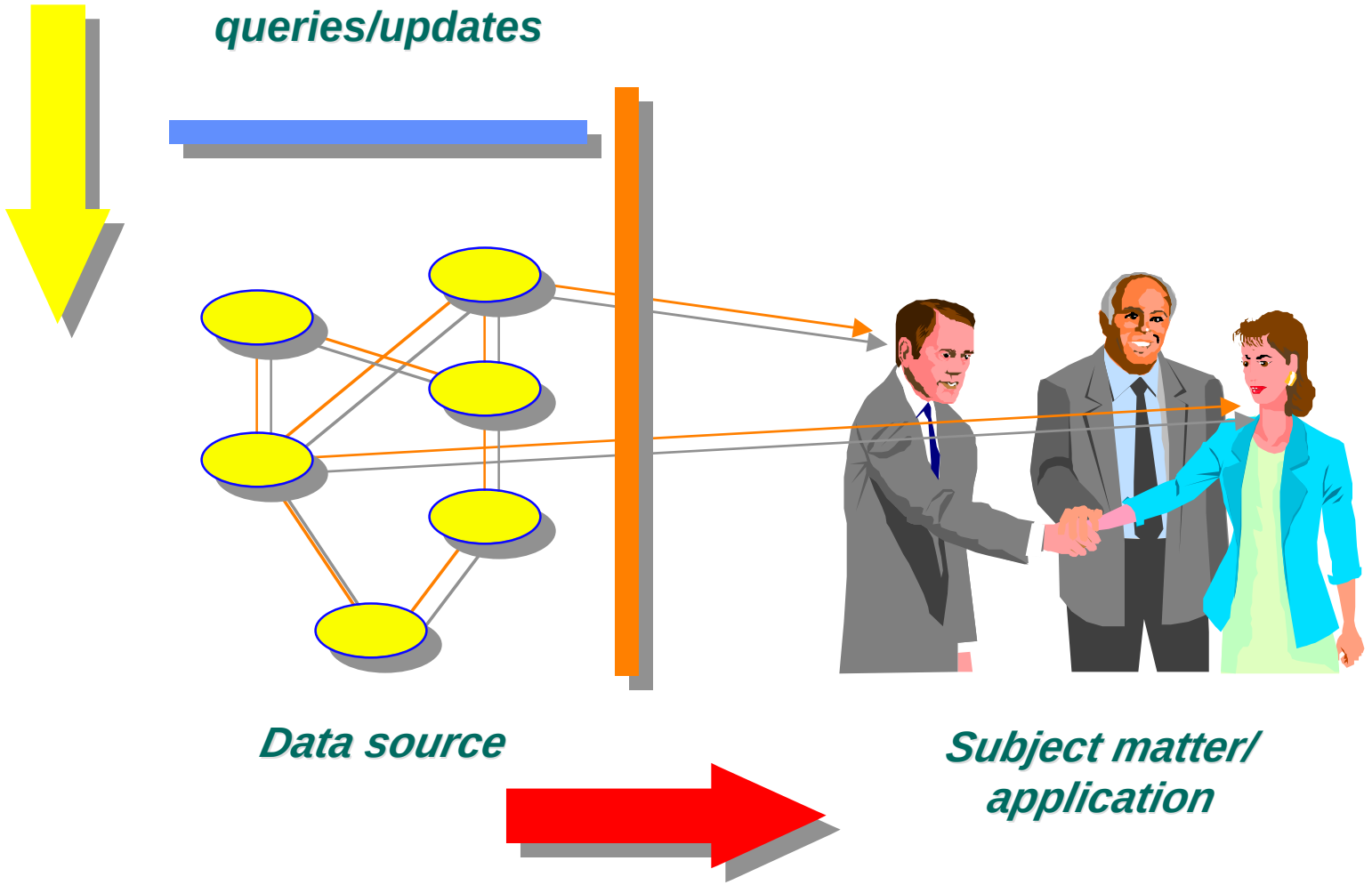
© 2004 John Mylopoulos

Data Semantics

- *Establish and maintain the correspondence between a data source, hereafter a **model**, and its intended subject matter.*
- *The model may be*
  - ✓ *A database storing data about employees in a company;*
  - ✓ *A database schema describing parts, projects and suppliers;*
  - ✓ *A website presenting information about a university;*
  - ✓ *A plain text file describing the battle of Waterloo.*

© 2004 John Mylopoulos

Machine State vs World Semantics



© 2004 John Mylopoulos

Semantic Data Models

- *Data models that attempt to capture "more world knowledge" [Codd79] than their logical counterparts.*
- *Make **ontological assumptions** about the subject matter, offer primitives accordingly.*
- *For example, the Entity-Relationship model "assumes that the world consists of entities and relationships" [Chen76].*
- *The Relational Model makes no ontological assumptions [Codd70].*

History

- *First semantic data models were proposed in 1974:*
  - ✓ *Jean-Robert Abrial;*
  - ✓ *Giampio Bracchi, Paolo Paolini, Giuseppe Pelagatti;*
  - ✓ *Jean-Luc Hainaut and Alain Pirotte;*
  - ✓ *Hans Schmid and Richard Swenson;*

Then in 1975

- ✓ *Peter Chen;*
- ✓ *Sham Navathe;*
- ✓ *Nick Roussopoulos and John Mylopoulos;*
- ✓ *….many others…*

Where do Semantic Data Models Fit?

- *Several possibilities, actually:*
  - ✓ *They are part of the technology --> Semantic DBMSs;*
  - ✓ *They are used during design;*
  - ✓ *They are part of the user interface to the database.*
- *Option 2 prevailed. Semantics were to be dealt with during **design-time**, rather than **run-time**, for performance reasons.*
- *But how does one use a database where semantics have been factored out?*

Rely on a stable environment of users and application programs to know the semantics!

© 2004 John Mylopoulos

...There is a down side to this data management practice:

Legacy data...

What did the Panel Experts See?

- *Factoring out the semantics of data won't work in dynamically changing, distributed, open environments, such as the web.*
- *In such settings, access of the data is not restricted to a small set of users and applications programs.*
- *This means that new users will have no clue what the data means; and the application programs that process the data may not have been designed specifically for these data.*

Hence the need to bundle together data and its semantics!

The Semantic Web

- *Unlike databases, hypertext data **are** designed for human consumption. However, these data are not machine processable.*
- *Hence the call for the Semantic Web [Berners-Lee01].*
- *Machine processable web data has come to mean "…having semantic metadata and ontologies for web content to enable information access, integration, interoperation and consistency…"*

Katia Sycara

ODBASE'03, November 7, 2003

The Layered Cake Architecture

- *From bottom to top…*
  - ✓ *Unicode, URI;*
  - ✓ *XML data, XML schema;*
  - ✓ *RDF data, RDF schema;*
  - ✓ *Ontology, vocabulary;*
  - ✓ *Logic;*
  - ✓ *Proof;*
  - ✓ *Trust.*

…but, who uses what, when??

© 2004 John Mylopoulos

Some Concerns

- *Hard to develop technologies for computationally demanding tasks, e.g., theorem provers, model checkers, deductive databases,…*
- *Scalability??*
- *Practitioners tend to not use logical specification languages, e.g., Z, Datalog,…*

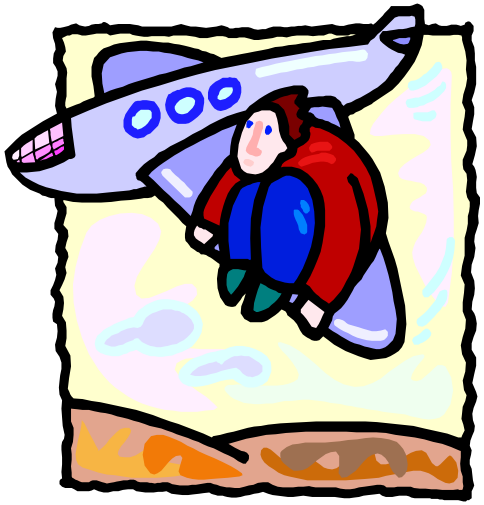
- ***From bottom to top…***
  - ✓ ***Unicode, URI;***
  - ✓ ***XML data, XML schema;***
  - ✓ ***RDF data, RDF schema;***
  - ✓ ***Ontology, vocabulary;***
  - ✓ ***Logic;***
  - ✓ ***Proof;***
  - ✓ ***Trust.***

We have to blend carefully technologies with methodologies

© 2004 John Mylopoulos

Towards a Novel Theory of Data Semantics

- *On-going (and very preliminary) work with Alex Borgida and Yuan An.*
- *Basic premise: If we are going to tackle the problem of data semantics for databases and/or web data -- again -- we better have a new angle at the problem!*



The Correspondence Continuum

- *Consider:*
  - ✓ *A photo of a landscape is a model with the landscape (its subject matter);*
  - ✓ *A photocopy of the photo is a model of a model of the landscape;*
  - ✓ *A digitization of the photocopy is a model of the model of the model of the landscape….etc.*
- *Meaning is rarely a simple mapping from symbol to object; instead, it often involves a* ***continuum of (semantic) correspondences*** *from symbol to (symbol to)* object [Smith87]*

© 2004 John Mylopoulos

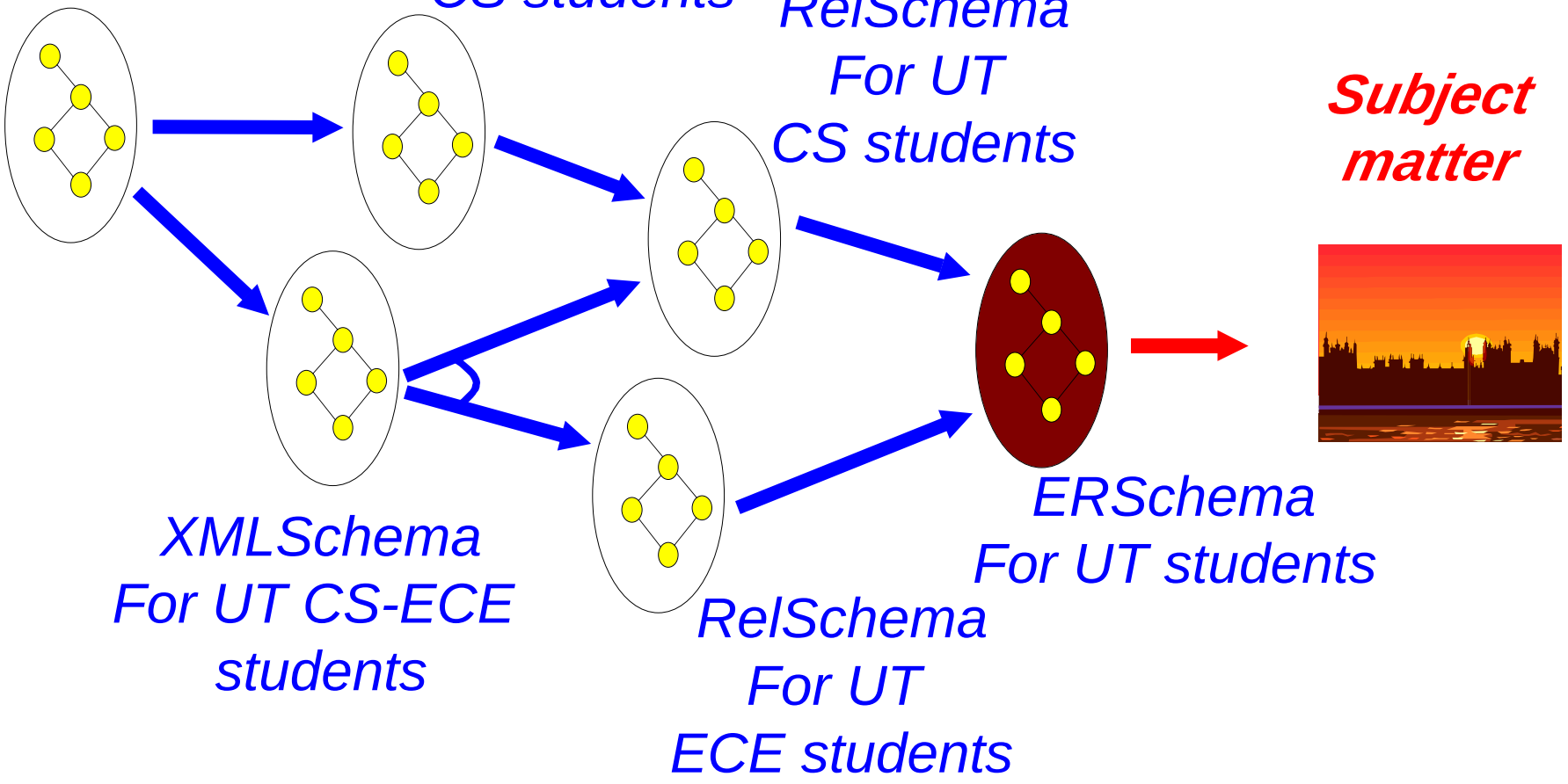
Correspondences and Mappings

- *Formally, **correspondences** constitute binary relations between symbols,*
  - ✓ *E.g., C = {(person,p), (pname,nm), (paddr,ad)}*
- ***Mappings** define translations from expressions defined with respect to one model to expressions defined with respect to another model,*
  - ✓ *E.g., person(x) ::= p(x)*

    pname(x,y) ::= p(x) and nm(x,y)

    paddr(x,y) ::= p(x) and ad(x,y)
- *The correspondence continuum gives rise to a mapping continuum.*

Example



© 2004 John Mylopoulos

Mapping Graphs

- *The graph associated with each mapping continuum is acyclic and has one or more anchors, its **semantic models ("ontologies"** in Semantic Web jargon).*
- *The semantic models constitute formally represented encyclopedias on a given subject matter; they are application-independent, specified in an expressive modeling language (e.g., OWL.)*
- *Every model has an associated (semantic) mapping to one or more other models.*

Models

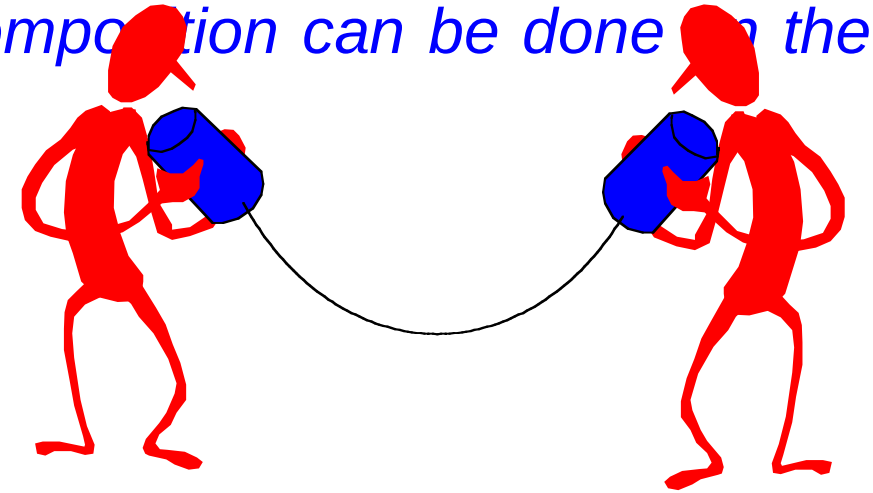
- *A model is intended to answer a specific set of questions about its subject matter (a model has a purpose!) [Ladkin97].*
- *For example, a model airplane can answer questions about the dimensions and aerodynamics of an aircraft; but not questions about its engine power, physical makeup, etc.*
- *For every model, we need a translation function (mapping) that will translate a query about the subject matter into one about the model (where applicable), and vice versa for the result of the query.*

Types of Models

- ***I-models*** *(intentional): Consist of a set of predicates with associated axioms; Database schemas (with integrity constraints), but also logical theories fit here.*
- ***E-models*** *(extensional): These have set-theoretic constructions, and query answering based on set-theoretic relationships; Tarskian and Kripke models, but also databases, fit here.*
- ***C-models*** (*computational*): *These are characterized by the fact that query answering is produced by running programs, e.g., a simulation program.*

Mappings

- *A mapping defines the semantics of a model with respect to its subject matter.*
- *This may be done in terms of GAV, (LAV?) or GLAV mappings, maybe others as well.*
- *Mappings have types too: denotations, representations, implementations, specifications,...*
- *Mapping composition can be done on the basis of their types.*



© 2004 John Mylopoulos

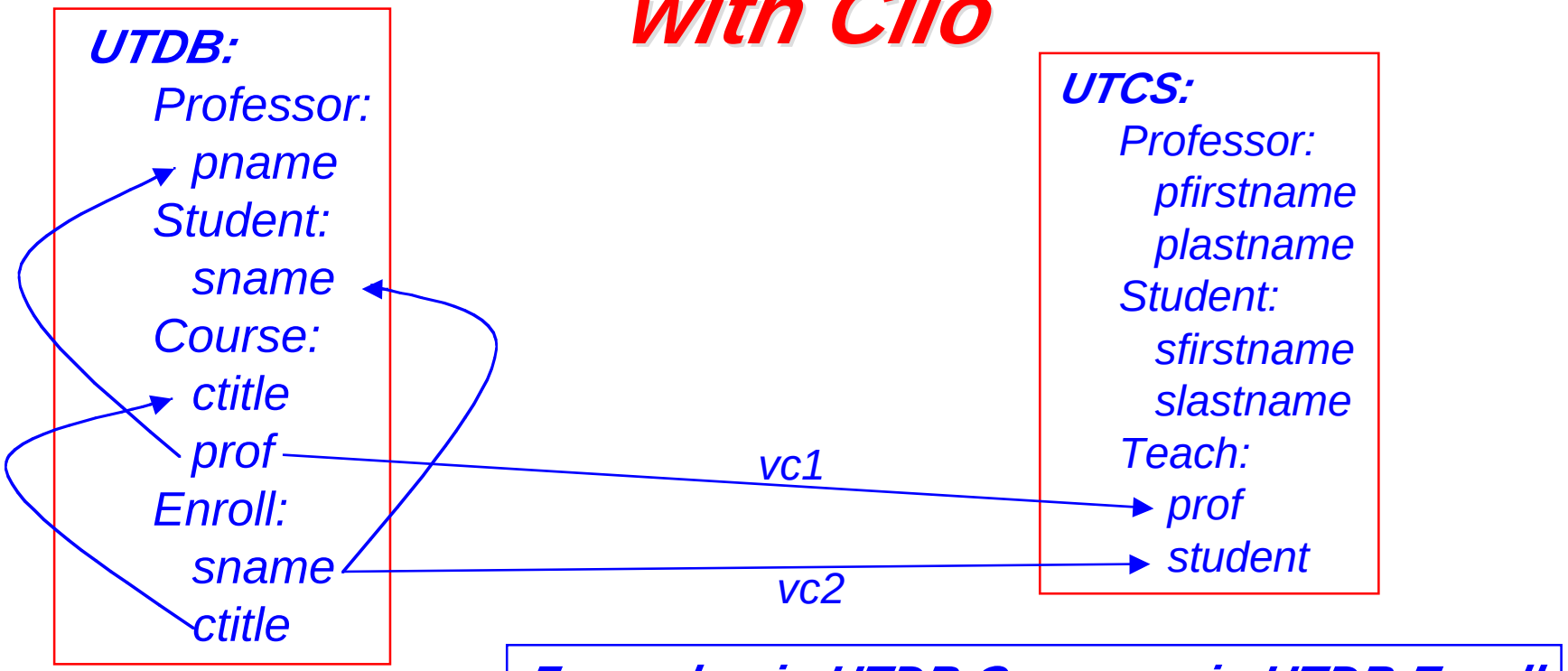
Compositions

- *The **semantics of a model** $\mathbf{m}$ consists of a composition of mappings $\mathbf{m_1}, \mathbf{m_2}, \ldots, \mathbf{m_n}$ that link it to a semantic model.*
- *The whole theory rests on the premise that we can come up with a rich enough class of mappings for which composition is meaningful and scalable.*

Related Research

- *Model management is concerned with the management of metadata (models) and the discovery of mappings between them [Bernstein00].*
- *Clio [Miller00] is a schema mapping tool that discovers mappings between two schemas, given a set of (value) correspondences between their elements (attributes).*
- *[Velegrakis03, Velegrakis04] extends Clio to support mapping evolution and schema and mapping management.*

Discovering Schema Mapping with Clio



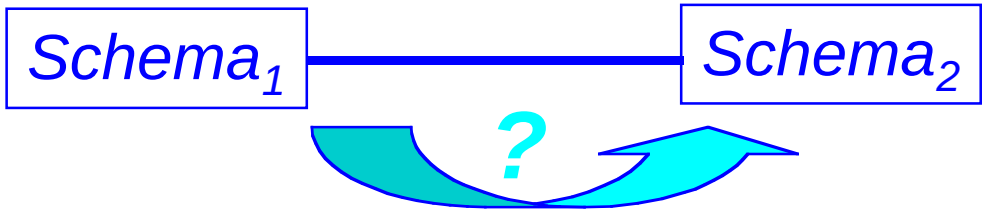
Foreach c in UTDB.Course, e in UTDB.Enroll
where c.Course.ctitle = e.Enroll.ctitle
exists t in UTCS.Teach
with c.Course.prof = t.teach.prof
and e.Enroll.sname = t.teach.student

© 2004 John Mylopoulos

Can we do Better?

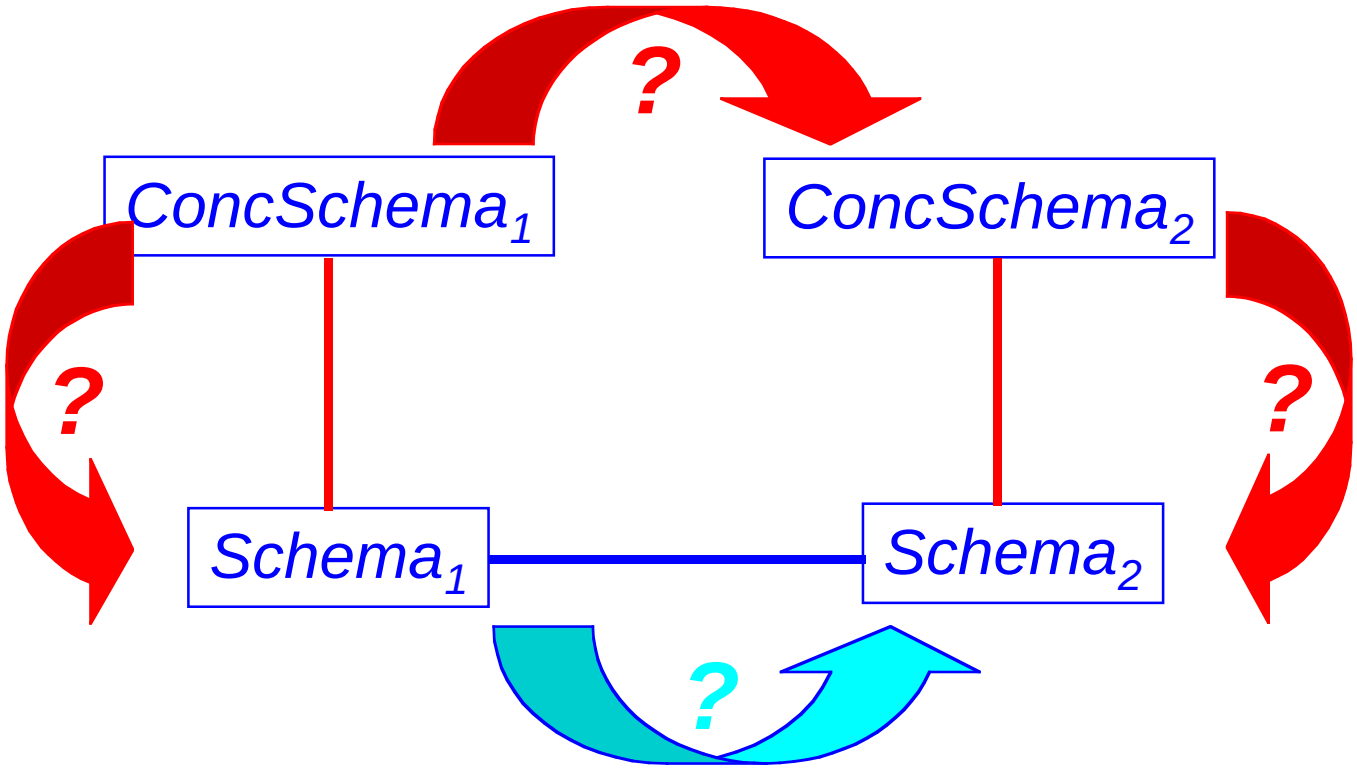
- *Clio relies on input value correspondences and foreign key constraints to discover "reasonable" mappings.*
- *Remains to be seen whether we can do better by also taking into account:*
  - ✓ *Disjointness and cardinality constraints;*
  - ✓ *Value restriction constraints;*
  - ✓ *Invariants;*
  - ✓ *…more…*
- *Such constraints may be expressed in terms of an ER schema, UML diagrams, or an OWL ontology.*

The Clio Problem



© 2004 John Mylopoulos

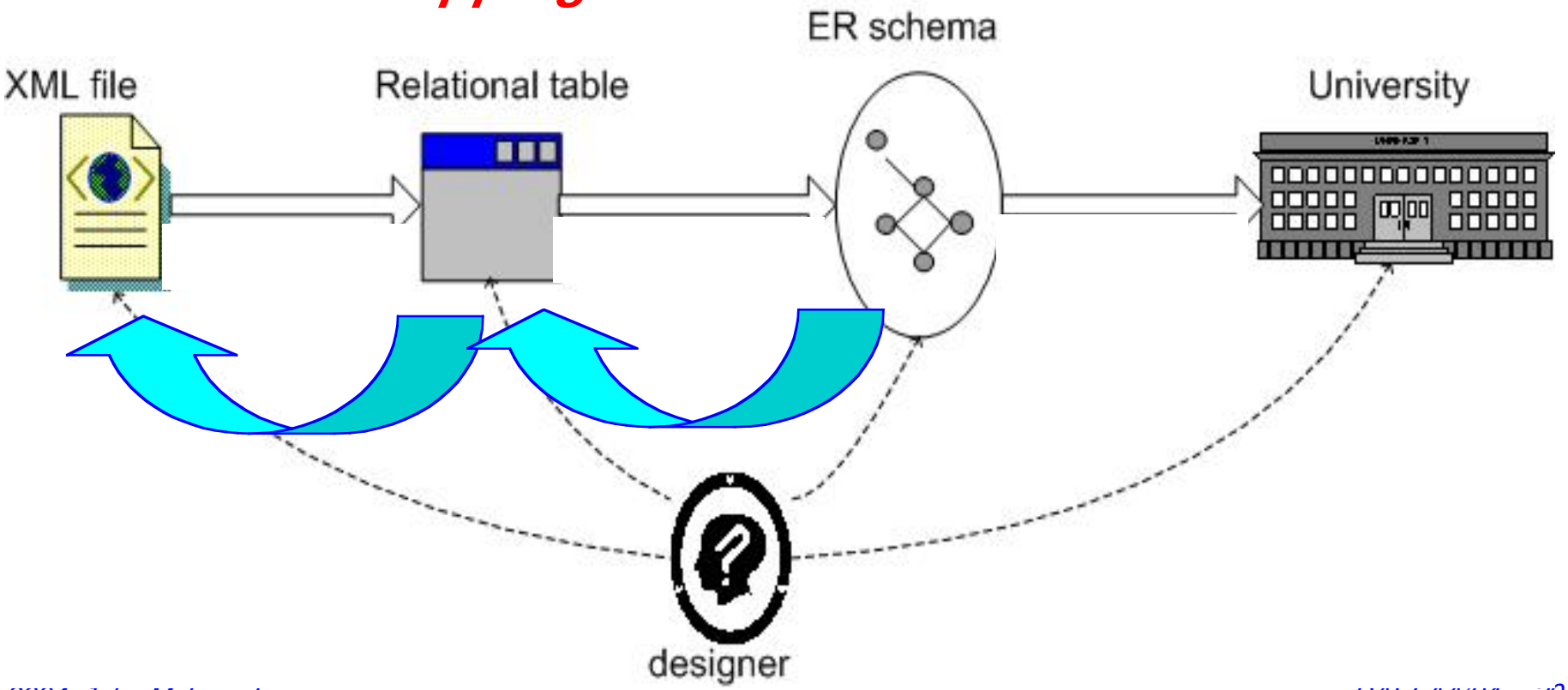
Extending the Clio Problem



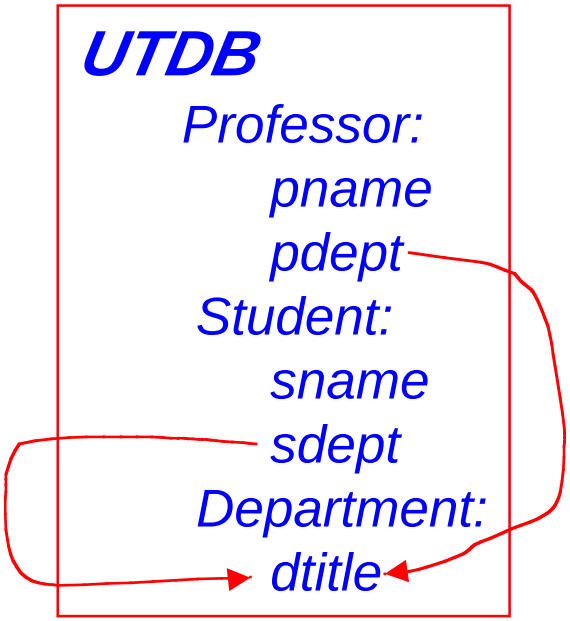
© 2004 John Mylopoulos

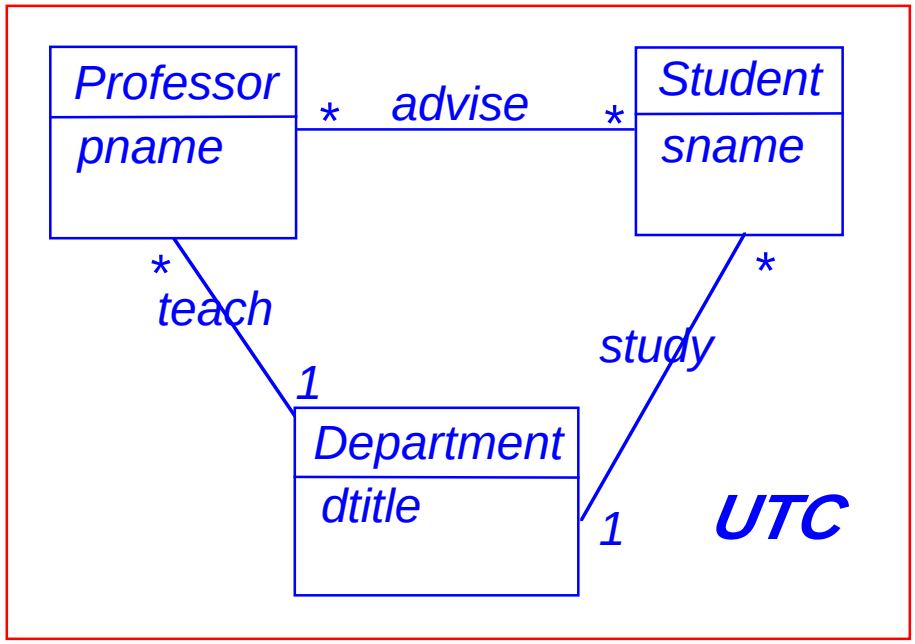
Some Mappings are Lineal

…and don't need to be discovered. These mappings are special in that they define the meaning of a schema, in the data semantics sense. Let's call them ***semantic mappings***



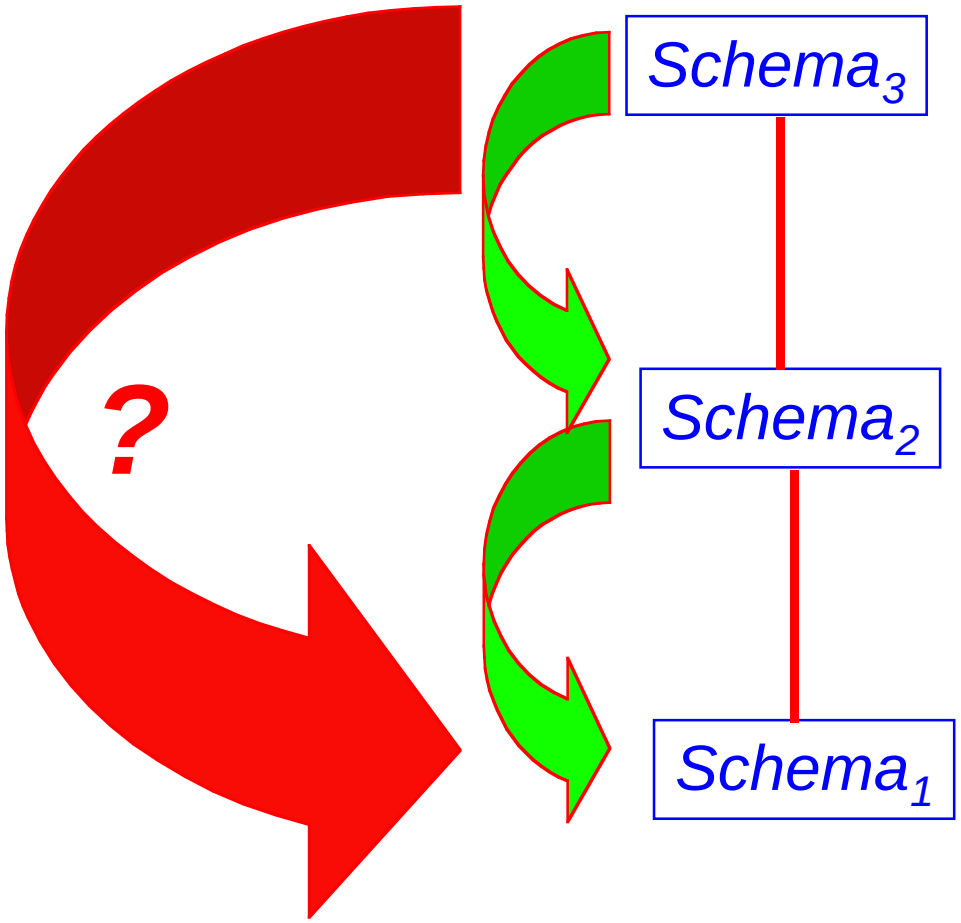
For Example…





Foreach path p:Professor----teach----d:Department in UTC
exists r in UTDB.Professors
where r.Professor.pname=p.Professor.pname
and r.Professor.pdept=d.Department.dtitle.

Extending the Clio Problem Further

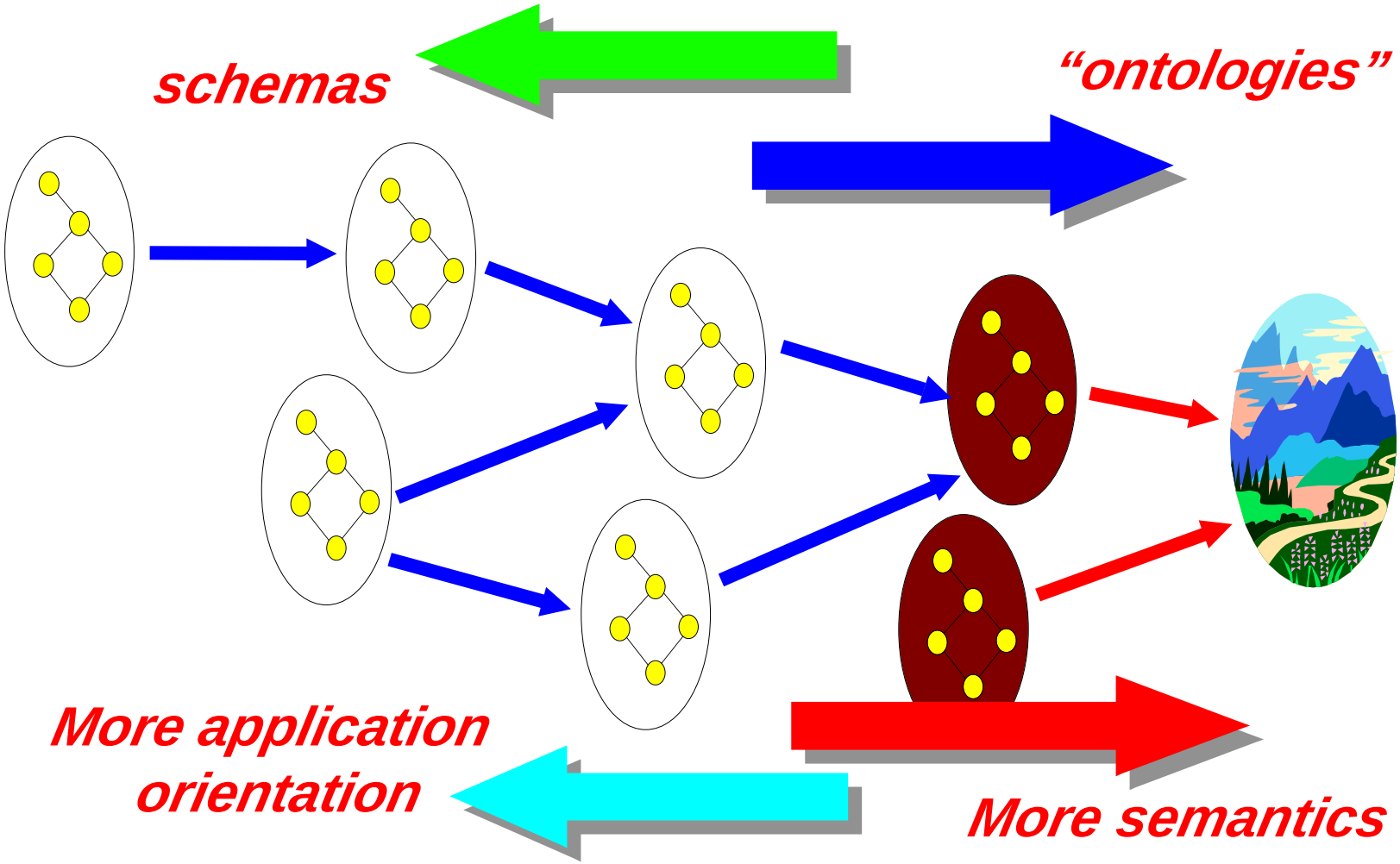


© 2004 John Mylopoulos

A New Dimension for Mapping Discovery

- *This (semantic) dimension is defined by expressive data models, such as the (Extended) ER Model, UML, or Description Logics.*
- *Note that the dimension is not strictly monotonic wrt expressiveness: There are constraints built into the ER Model that cannot be expressed in any Description Logic.*
- *For example, you can't state in any Description Logic for an n-ary relationship R, $n \geq 3$ "If two instances of R $r(e_1,\ldots e_n)$, $r'(e_1',\ldots e_n')$ are such that $e_1 = e_1',\ldots,e_n = e_n'$, then r = r'." [Calvanese].*

So, What Does All This Mean?



© 2004 John Mylopoulos

Remarks

- *Most computations use "simple" models, schemas, databases and XML data.*
- *Some others need access to the semantics of these models as well; mappings and mapping compositions are required here.*
- *Most contributors to the Semantic Web vision get to use well-known database technologies (schemas, queries, views,…).*
- *Semantic web applications resort to expensive semantic computations only when they need to.*

Claim: Mappings are easier to formalize/discover than concepts

© 2004 John Mylopoulos

This is a version of the Semantic Web with an emphasis on mappings and mapping compositions, rather than rich semantic models

Semantic Encapsulation

- *The proposed framework is founded on the premise that*

Every model comes with a semantic mapping to another model

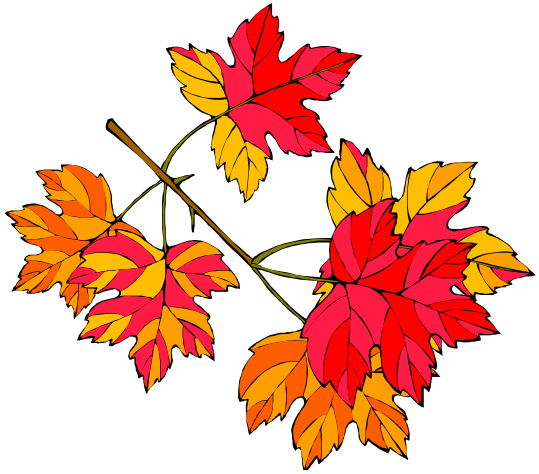
- *We have accepted behavioural encapsulation, why not semantic one?*
- *Of course, there is a price to be paid in requiring that every model (e.g., schema) comes with its semantics…*

© 2004 John Mylopoulos

...But there is a far greater price to pay with legacy data

Acknowledgements

This presentation is based on on-going work with Alex Borgida and Yuan An



© 2004 John Mylopoulos

VLDB'04 in Toronto:
August 30 - September 3
www.vldb04.org

References

- *[Berners-Lee01] Berners-Lee, T., Hendler, J., Lassila, O., "The Semantic Web: A New Form of Web Content that is Meaningful to Computers will unleash a Revolution of New Possibilities", Scientific American, May 2001.*
- *[Bernstein00] Bernstein, P., Halevy, A., Pottinger, R., "A Vision for Management of Complex Models, ACM SIGMOD Record 29(4), 2000*
- *[Chen76] Chen, P., "The Entity-Relationship Model: Towards a Unified View of Data," ACM Transactions on Database Systems 1(1), 1976.*
- *[Codd70] E. Codd. "A Relational Model for Large Shared Data Banks," Communications of the ACM 13(6), 377-387, 1970.*
- *[Codd79] E. Codd. "Extending the Database Relational Model to Capture More Meaning." ACM Transactions on Database Systems 4, 1979.*

© 2004 John Mylopoulos

More References

- *[King02] King, R., "The Story of Civilization and the Rubicon of Smart Data", Proceedings 28th International Conference on Very Large Databases (VLDB'02), Hong Kong, August 2002.*
- *[Ladkin97] Ladkin, P., "Abstraction and Modeling" Research Report RVS-Occ-97-04, University of Bielefeld, 1997; http://www.rvs.uni-bielefeld.de/publications/abstracts.html#AbsMod.*
- *[Miller00] Miller, R., Haas, L., Hernadez, M., "Schema Mapping as Query Discovery", Proceedings International Conference on Very Large Databases (VLDB'00), Cairo, September 2000.*
- *[Velegrakis03] Velegrakis, Y., Miller, R., Popa, L., "Mapping Adaptation Under Evolving Schemas", Proceedings International Conference on Very Large Databases (VLDB'03), Berlin, September 2003.*
- *[Velegrakis04] Velegrakis, Y., Miller, R., Mylopoulos, J., "Representing and Querying Data Transformations", (to appear.)*

© 2004 John Mylopoulos