Carleton
UNIVERSITY

Data Quality in Information Integration and Business Intelligence

Leopoldo Bertossi*

Carleton University
School of Computer Science
Ottawa, Canada

⋆: Faculty Fellow of the IBM Center for Advanced Studies

Data Quality and Business Intelligence

Business Intelligence (BI) requires modeling of different and complex operational aspects of business activities of an enterprise

This requires understanding those business activities and actors, and learning from experience and data

Business activities have become more complex: distributed tasks, interaction of different computational and human processes, increasing amount and complexity of data, ...

It becomes essential to be able to automate tasks that are related with data management

Many of them for automating and/or supporting decision making in organizations

- Integration of data from multiple and heterogeneous data sources

  Also data from "the web" and more complex and diverse "data spaces"

- Search in complex and non-traditional data spaces

  Data models? Query Languages? Data retrieval?

- Learning from data, e.g. machine learning, data mining

  Increasing amount of semi- or non-structured data

- Enterprise integration systems: local database systems (operational, DWHs), work-flows, external data sources, application programs, ERP systems, WWW, etc.

- Integration of data management with higher-level reasoning systems: intelligent information systems, knowledge bases, ontologies, semantic web, etc.

- Data quality assessment and data cleaning

  Making the right business decisions and inferences requires quality data

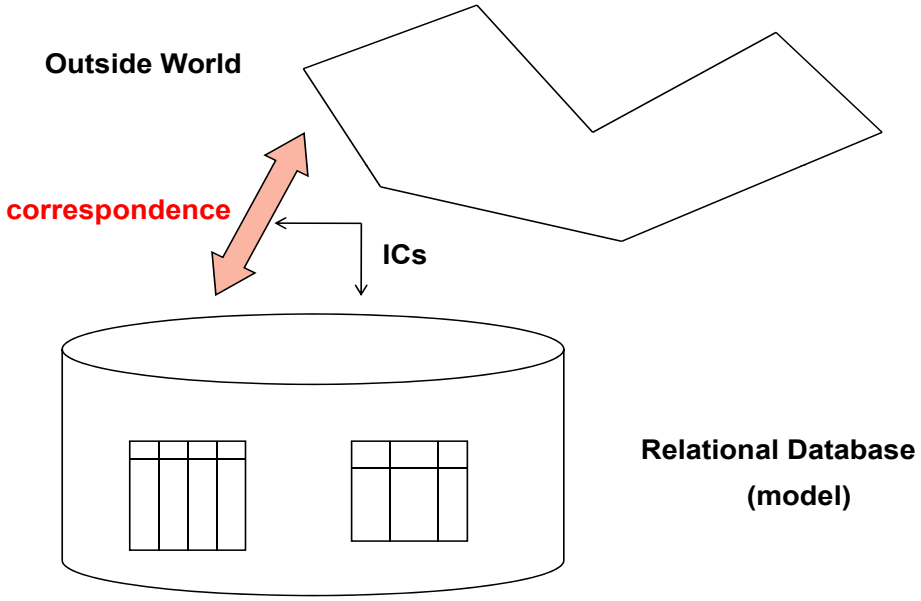
  It is not clear how to achieve it

  Actually, not even clear how to characterize it

  Existing solutions are specific and vertical

  Too many dimensions of data quality: accuracy, completeness, redundancy, freshness, consistency, etc.

Consistency of Databases



A database instance D is a model of an outside reality

An integrity constraint on D is a condition that D is expected to satisfy in order to capture the semantics of the application domain

A set IC of integrity constraints (ICs) helps maintain the correspondence between D and that reality

Example: Instance D violates the functional dependency FD: $Name \rightarrow Salary$

$Employee$	$Name$	$Salary$
	$Page$	5K
	$Page$	8K
	$Smith$	3K
	$Stowe$	7K

An inconsistent database, an undesirable situation ...

If D satisfies IC, we say that D is consistent

For several reasons a database may become inconsistent with respect to a given set of desirable ICs, e.g.

- Non-enforced ICs
- User constraints
- Data integration
- Legacy data without the currently required semantics

ICs can be expressed in languages of symbolic logic

The database can be seen as a mathematical structure

(In)consistency of data can be studied with mathematical models and techniques

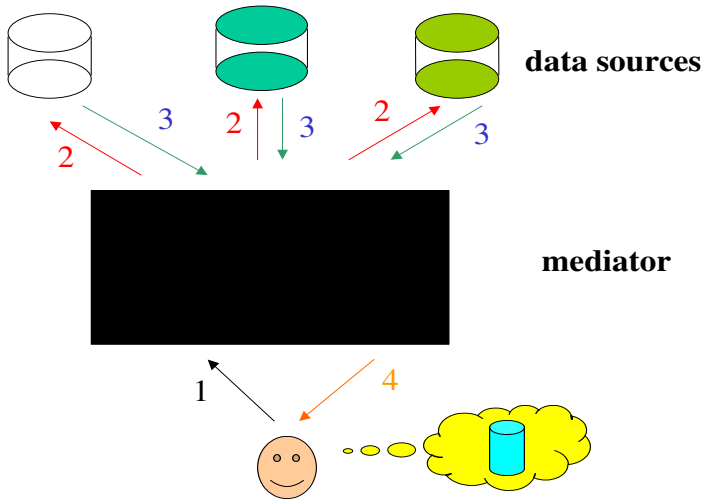
In particular, the problem of characterizing and obtaining semantically correct information from inconsistent database

A research program we started almost 10 years ago ...

Is it relevant?

An important scenario: when data is virtually integrated from several autonomous and heterogeneous data sources

Increasingly relevant given the diversity and number of data sources!



The user interacts with a mediator, having the feeling of being interacting with a single, real database, with a global schema

No way to maintain a set of global ICs satisfied!

What If the Database is Inconsistent?

Bringing it back to a consistent state may be difficult, impossible, nondeterministic, undesirable, unmaintainable, etc.

We may have to live with inconsistent data ...

The database (the model) is departing from the outside reality that is being modeled

However, the information is not all semantically incorrect

Most likely most of the data in the database is still "consistent"

Idea: (a) Keep the database as it is

(b) Obtain semantically meaningful information at query time; dealing with inconsistencies on-the-fly

Particularly appealing in virtual data integration ...

Characterizing Consistent Data

What is the consistent data in an inconsistent database?

What are the consistent answers to a query posed to an inconsistent database?

A mathematically precise definition is needed, that makes sense from the point of view of: (a) the intuitions behind the concept, and (b) its applications

In (Arenas,Bertossi,Chomicki; PODS99) such a characterization was provided

Intuitively, the consistent data in an inconsistent database D is invariant under all minimal ways of restoring D's consistency

That is, consistent data persists across all the minimally repaired versions of the original instance: the repairs of D

Example: For the instance D that violates FD: $Name \rightarrow Salary$

$Employee$	$Name$	$Salary$
	$Page$	5K
	$Page$	8K
	$Smith$	3K
	$Stowe$	7K

Two possible (minimal) repairs if only deletions/insertions of whole tuples are allowed: D_1, resp. D_2

$Employee$	$Name$	$Salary$
	$Page$	5K
	$Smith$	3K
	$Stowe$	7K

$Employee$	$Name$	$Salary$
	$Page$	8K
	$Smith$	3K
	$Stowe$	7K

$(Stowe, 7K)$ persists in all repairs: it is consistent information

$(Page, 8K)$ does not; actually it participates in the violation of FD

Consistent Query Answering

A consistent answer to a query $\mathcal{Q}$ from a database D that is inconsistent wrt IC is an answer that can be obtained as a usual answer to $\mathcal{Q}$ from every possible repair of D wrt IC

- $\mathcal{Q}_1 : Employee(x, y)?$

  Consistent answers: $(Smith, 3K), (Stowe, 7K)$

- $\mathcal{Q}_2 : \exists y Employee(x, y)?$

  Consistent answers: $(Page), (Smith), (Stowe)$

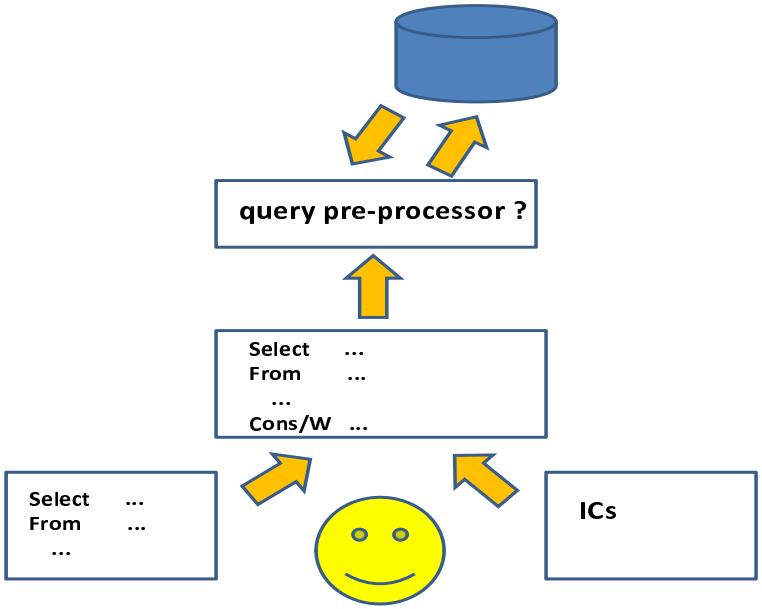
CQA may be different from classical data cleaning!

However, CQA is relevant for data quality; an increasing need in business intelligence

It also provides concepts and techniques for data cleaning

Next DBMSs should provide more flexible, powerful, and user friendlier mechanisms for dealing with semantic constraints

In particular, they should allow to be posed queries requesting for consistent data; and answer them



Why not an enhanced SQL?

SELECT	Name, Salary
FROM	Employee
CONS/W	FD: Name -> Salary;

(FD not maintained by the DBMS)

Paradigm shift: ICs are constraints on query answers, not on database states!

All Kinds of Interesting Problems!

- Logical Problems

Query answering in databases follows classical predicate logic (relational calculus)

What is the logic followed by CQA?

What about compositionality? How to compute consistent answers combining consistent answers to subqueries?

- Algorithmic Problems

How to compute consistent answers?

We have to avoid as much as possible computing and materializing all possible repairs

Try to use the only instance D at hand, the inconsistent one ...

Is it possible to rewrite a query $\mathcal{Q}$ that expects consistent answers into a new query $\mathcal{Q}'$, whose usual answers from D are the consistent answers to $\mathcal{Q}$?

In the example, the first query can be transformed into a standard SQL query to be posed to the original database

```
SELECT   Name, Salary
FROM     Employee;
```

Transformed into $\mapsto$

```
SELECT   Name, Salary
FROM     Employee
WHERE    NOT EXISTS (
           SELECT *
           FROM   Employee E
           WHERE  E.Name = Name AND
                  E.Salary <> Salary);
```

No repairs needed to obtain consistent answers!

Query on the RHS is easy ... Always possible?

- Mathematical Problems

What is the intrinsic complexity of CQA?

As a decision problem? As a data retrieval problem?

Expressive power of logical languages to rewrite queries for CQA?

Tractable vs. intractable cases

Approximation algorithms for intractable cases?

Use information theory to characterize degrees of consistency of databases as mathematical structures

- Computational Implementation

How to enhance DBMS for doing CQA?

How to couple a DBMS with a reasoning system for CQA?

• Applications

Many, where consistency of data is an issue ...

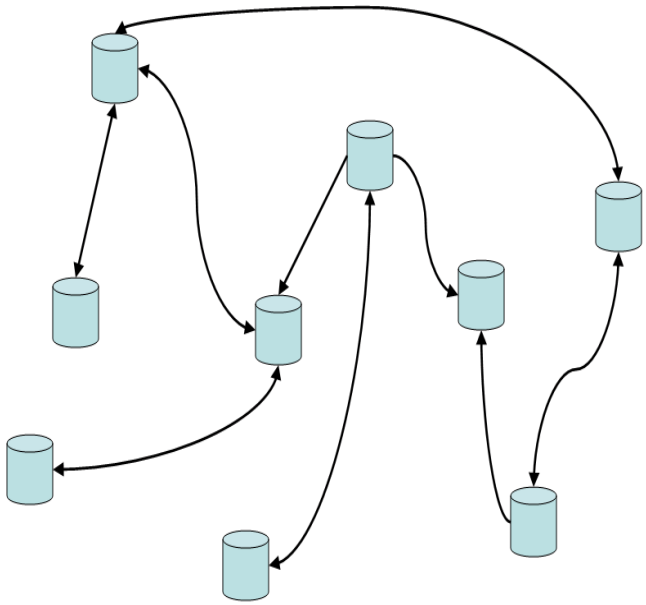
- Query answering in peer-to-peer data exchange systems

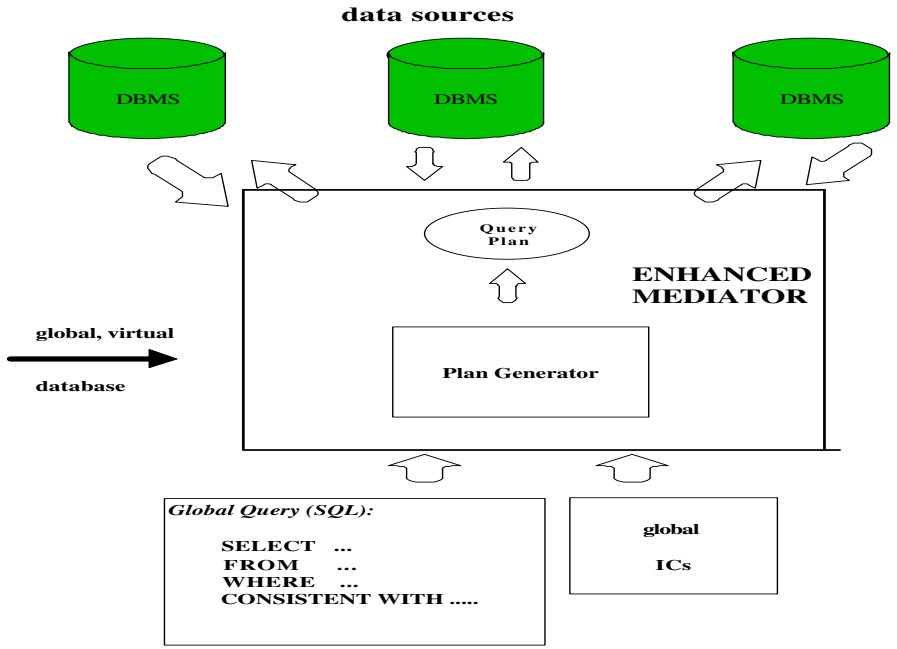
No centralized management

Peers exchange data at query answering time

Queries are posed to a peer who imports other peers' data or adjusts its own data

Trust relationships between peers may influence this process





- Virtual data integration

We can make the query plan generator to take into account the global ICs

The enhanced plan retrieves data from the sources, but returns only answers that are consistent with the global ICs

There are sufficiently expressive languages to express query plans

And can be evaluated using extended logic programming systems (answer set programming)

Query answers can be used to materialize an integrated and semantically correct global instance!

- Data quality and data cleaning

A vast territory still to be explored ...

Adaptive Data Cleaning

A research agenda: The NSERC Strategic Network on "Data Management for Business Intelligence" (BIN)

Around 20 academic researchers across Canada plus industrial participants; four themes:

1. "Strategy and Policy Management"
2. "Capitalizing on Document Assets"
3. "Adaptive Data Cleaning"

   Theme leader: L. Bertossi
4. "Supporting Top-down (Business-driven) Data Integration"

Main goals:

- Beyond Verticals

  So far: Specific and vertical solutions

  Not extendible or adaptable

  Start from scratch for every problem and application

  More generic approaches and techniques?

- Declarative Cleaning

  Specification of data quality assessment criteria

  Specification of data cleaning activities

  Generic and parameterizable

- New ways of understanding, modeling, improving information quality

  Quantify and assess data and metadata quality

More Specifically:

- What can be extracted, abstracted out from different domains, problems and techniques in DC?
- Identification of relevant parameters and dimensions of data quality assessment and cleaning
- Specification of quality data
- Identification and characterization of quality constraints
- Specification of quality constraints

  Particular case: integrity constraints

  Specification languages? What kind of primitives?
- Obtaining quality data: data cleaning vs. obtaining clean data at query or application time

  Characterization and retrieval of clean data in/from a dirty data source

- Data quality in data integration

  Materialized, virtual, ...

- Data quality assessment and data cleaning is context dependent

  Identification, characterization, specification and applications of contexts for data cleaning

- Specification and implementation of data cleaning activities

- Characterization of semantically clean data in data spaces with data accessible through search queries

Conclusions

Data management touches almost every aspect of BI

General and flexible solutions to old problems become crucial

$\Rightarrow$ Data Cleaning

Solutions to new problems becomes necessary

$\Rightarrow$ Consistency issues in virtual data integration

There are:

- Many stimulating scientific and technical challenges
- Important consequences for BI
- Much room and need for exciting interdisciplinary research: business vs. data management