Carleton
UNIVERSITY

Business Intelligence: Recent Experiences in Canada

Leopoldo Bertossi[⋆]

Carleton University
School of Computer Science
Ottawa, Canada

[⋆] Faculty Fellow of the IBM Center for Advanced Studies

Business Intelligence and Data Management

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 (DM)

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

Almost every area and problem of BI touches DM problems

The community of data management at large has been doing research in- and developing techniques and tools for BI

There are many open research problems in DM that are motivated, or made more relevant and urgent by BI needs

The database (data management) research community has been traditionally strong in Canada

There have been successful BI companies in Canada, even before the term was coined, e.g. Cognos, Business Objects, IBM (first IBM CAS created in Toronto), ...

The Canadian DM research community decided to put together a large BI research project

The BIN Initiative

The genesis of the The NSERC Strategic Network on "Data Management for Business Intelligence" (BIN)

In 2007 around 20 academic researchers across Canada started putting together a large research project in BI

- Potential industrial partners were approached and workshops with them were held, most prominently IBM, Cognos (later bought by IBM), Business Objects (later bought by SAP), etc.
- 15 principal investigators (PIs) were designated for a proposal submission

  Universities: Toronto, Ottawa, Carleton (Ottawa), British Columbia (Vancouver), Dalhousie (Halifax), Alberta (Edmonton), Univ. of Waterloo

- Proposal submitted to NSERC (Canada's main research funding agency) for an “NSERC Strategic Network”

- Eventually, among many participants, two networks were selected, one of them our “BIN Network”
- Started in 2009, for 5 years
- Approx. $1 million per year from NSERC

  $0.5 million per year from industrial partners, main contributors: IBM, SAP
- A Board of Directors was nominated, with external members from academia (USA, Germany, ...) and companies, in particular industrial partners
- An executive committee was created: network director, four theme leaders, and representatives from industrial partners (also acting as interface to the academic partners)

- Every year, PIs submit reports and proposals, including request for funding

  Funds come from the NSERC portion

  Industrial partners top-up if project is particularly relevant to industry

  There is room for both basic and applied research

  Funds are used mostly to pay graduate students, and post-docs; and exchanges among network members

The Network's Themes

Network director: Renee Miller (Univ. of Toronto)

1. "Strategy and Policy Management"
   Theme leader: John Mylopoulos (Univ. of Toronto)
2. "Capitalizing on Document Assets"
   Theme leader: Frank Tompa (Univ. Waterloo)
3. "Adaptive Data Cleaning"
   Theme leader: L. B. (Carleton Univ.)
4. "Business-Driven Data Integration"
   Theme leader: Iluju Kiringa (Univ. of Ottawa)

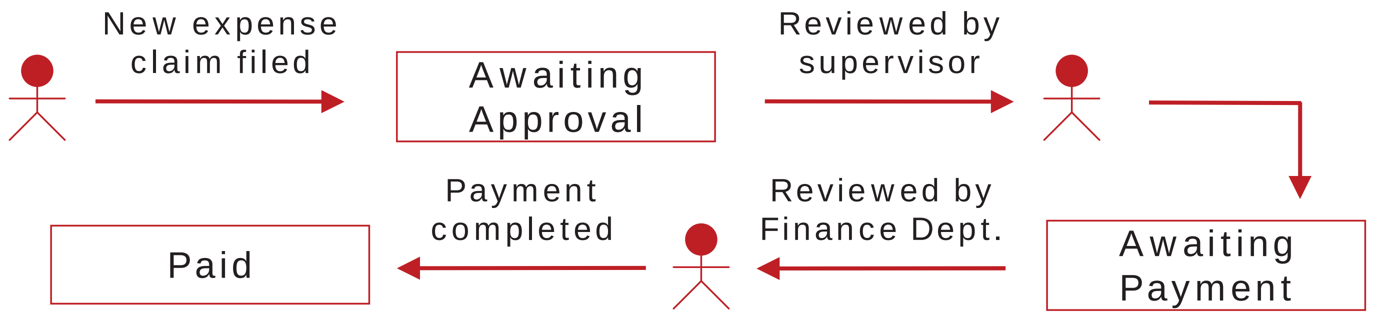
Other PIs, researchers, graduate students gather around a theme or several or them

1. Strategy and Policy Management:

- Modeling business processes
- Modeling and representing business policies (business rules)
- How to specify and integrate BPs and wokflows with data management
- How to automate the process from the specification
- How to run and monitor it
- Reasoning about the BP

  Does it guarantee that the desired business goals will be achieved?

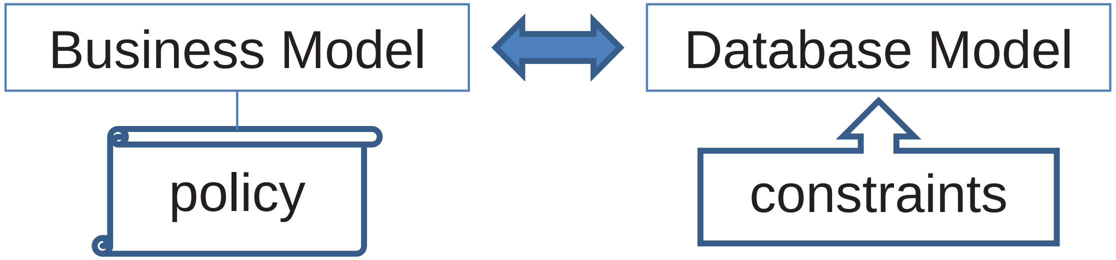
  Or undesirable behavior always avoided?



A business process (BP)

Different ways of modeling BPs, involving states, agents, policies, workflows, data management

Newer approach: Data-Centric Business Processes, in particular IBM's Business Artifacts ...



Stages of business objects in a BP are represented through database states, and their changes as database updates

Creation and elimination of business objects represented as database operations

The BP can be monitored and run via the underlying data management process, through database states, queries, integrity constraints, triggers, ...

The data layer becomes a footprint of the business layer

An idea to have into account when modeling both layers and mapping them ...

2. Capitalizing on Document Assets:

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

  Increasing amount of semi- or non-structured data

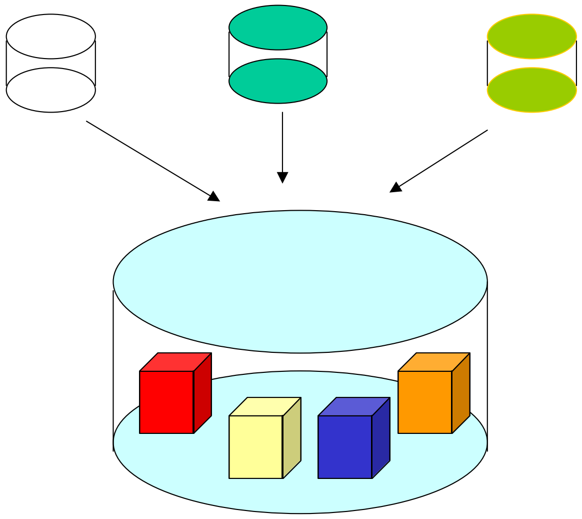
- Information retrieval

- Non-classical data mining:
  - web data, and web documents in general
  - text files
  - blogs
  - emails
  - social networks (learning about/from on-line communities)
  - sentiment analysis
  - recommender systems
  - ...

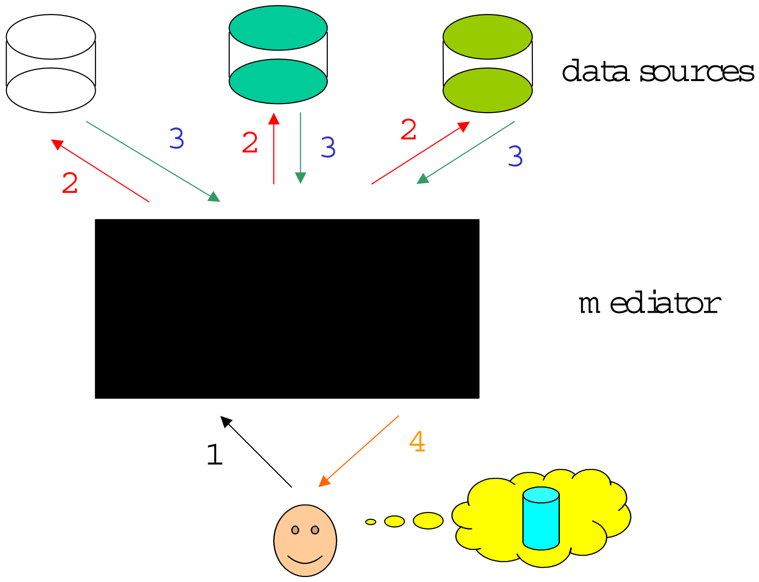
4. Business-Driven Data Integration:

- Integration of data from multiple and heterogeneous data sources, in different ways:

  Materialized integration, e.g. data warehouses

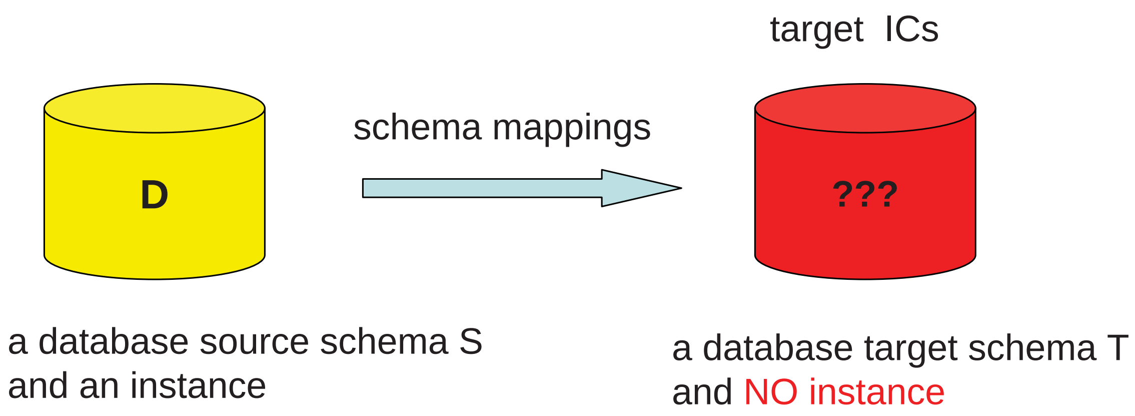


Virtual integration, e.g. mediator-based data integration systems



- Integration of data from WWW, and more complex and diverse “data spaces”
- Enterprise integration systems: local database systems (operational, DWHs), work-flows, external data sources, application programs, ERP systems, WWW, etc.

- Data exchange



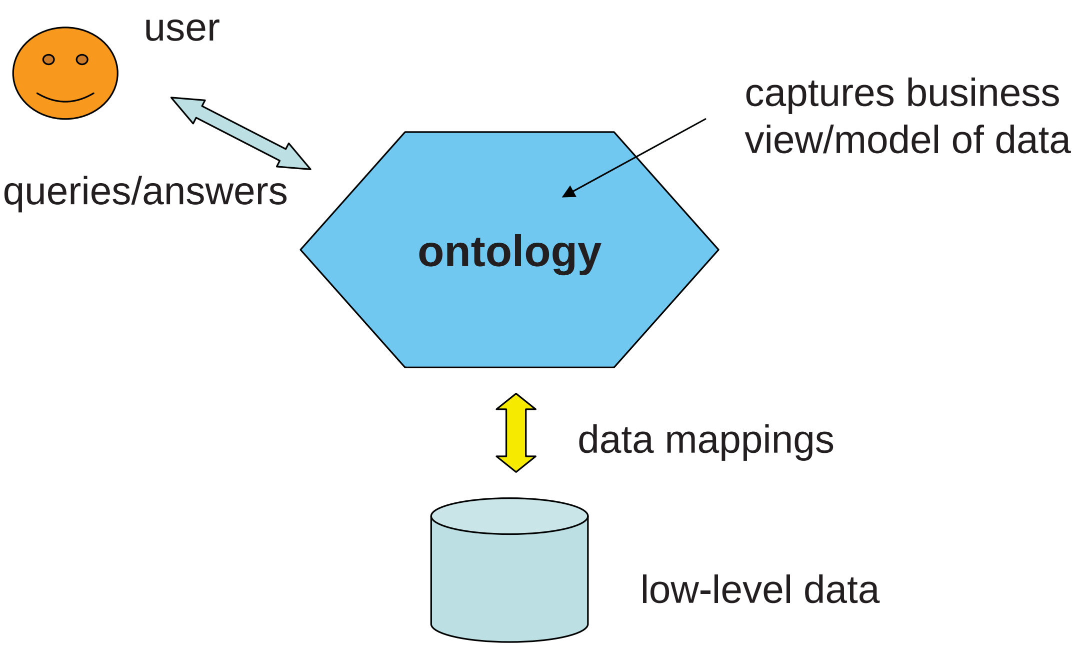
How to materialize an instance for the target schema

What are the semantically correct data at the target?

- Querying data sources through ontologies

Data stay underneath, and may not represent the way the user sees the business

Ontology is metadata or an explicit, formal ER model, etc.



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

- How the business model drives the processes of data integration in any of the above forms?

3. Adaptive Data Cleaning:

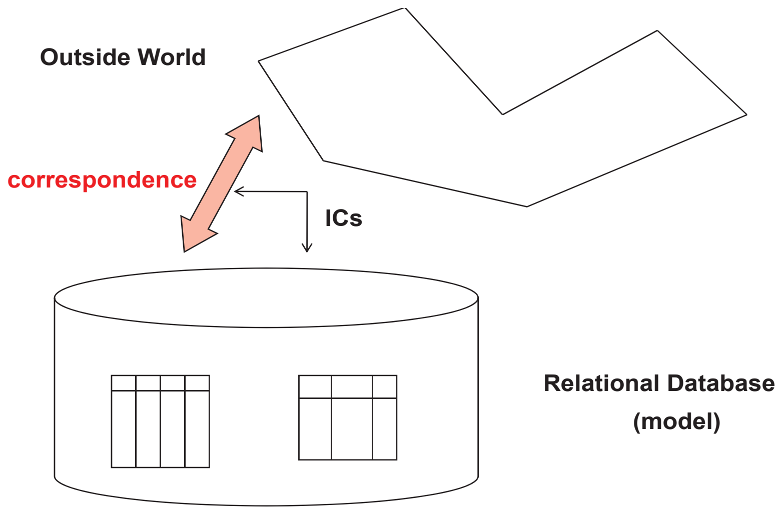
Making the right business inferences and decisions requires quality data

Not always clear how to assess the quality of data or clean data

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

Actually, not even clear how to characterize clean data within a data repository

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 specify/maintain the correspondence between D and that reality

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 ...
(no direct access to the data sources)

This requires:

(a) logically characterizing consistent data within an inconsistent database

(b) Developing algorithms for retrieving the consistent data

Why adaptive data cleaning?

We want to go beyond verticals

Most commonly data cleaning solutions are specific, domain dependent, and vertical

Not extendible or adaptable

Start from scratch for every problem and application

More generic approaches and techniques?

Declarative data cleaning

Specification of data quality assessment criteria

Specification of data quality properties/requirements

What properties should be satisfied for the data to be clean? (not necessarily explicitly how to clean)

Specification of data cleaning activities?

Possibly derived from data quality requirements

Generic and parameterizable solutions

New ways of understanding and modeling data quality

Quantify and assess data and metadata quality

Many problems, 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: ICs)

  More recently: entity-resolution, duplicate detection/resolution
- Specification languages?
- Definition of quality predicates

  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 are 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

New solutions to old problems becomes necessary

New problems have appeared that need solutions

In DC many stimulating scientific and technical challenges

Many fundamental problems have not been (properly) addressed

There is a lack of basic research

There is much room for both basic and applied research

Room for exciting interdisciplinary research: business and data management