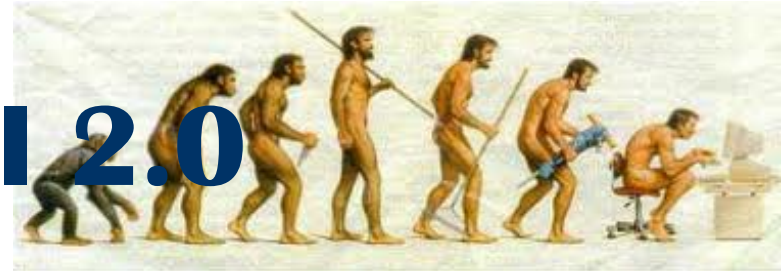
A Query Reformulation Framework for P2P OLAP (Extended Abstract)

Matteo Golfarelli
Wilma Penzo
Stefano Rizzi
Elisa Turricchia
(University of Bologna - Italy)
Federica Mandreoli
(University of Modena and Reggio Emilia - Italy)

Summary

- Collaborative BI
- Business Intelligence Networks (BINs)
- The query reformulation problem
- The BIN mapping language
- The BIN query reformulation framework
- Concluding remarks

From BI 1.0 to BI 2.0



- *Business intelligence* (BI) transformed the role of computer science in companies from a technology for storing data into a discipline for timely detecting key business factors and effectively solving strategic decisional problems
- In the current changeable and unpredictable market scenarios, the needs of decision makers are rapidly evolving
- To meet the new, more sophisticated user needs, a new generation of BI systems (BI 2.0) has been emerging

Collaborative BI

- One of the key features of BI 2.0 is **the ability to become collaborative** and extend the decision-making process beyond the boundaries of a single company
- **Collaboration** is seen today by companies as one of the major means for increasing flexibility and innovating so as to survive in today uncertain and changing market
- This is particularly relevant in inter-business collaborative contexts where companies organize and coordinate themselves to share opportunities, respecting their own autonomy and heterogeneity but pursuing a common goal

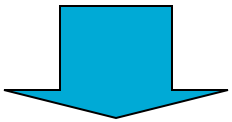
Benefits and requirements

- Main benefits to the corporate world:
  - enhance the decision making process and create new knowledge
  - build new inter-organizational relationships and coordination approaches
  - efficiently manage inter-company processes and safely share management information
- Main requirements of the corporate world:
  - Users need to access information anywhere it can be found, i.e. regardless of its location
  - Users need to transparently and uniformly access information scattered across several heterogeneous BI platforms
  - Information must be integrated on the fly and returned to users

But…

In a distributed business scenario **traditional BI systems are no longer sufficient** to maximize the effectiveness of decision making processes

- …most information systems were devised for individual companies and for operating on internal information, and they give **limited support to inter-company cooperation**
- …traditional BI applications are aimed at serving individual companies, and they **cannot operate over networks of companies** characterized by an organizational, lexical, and semantic heterogeneity

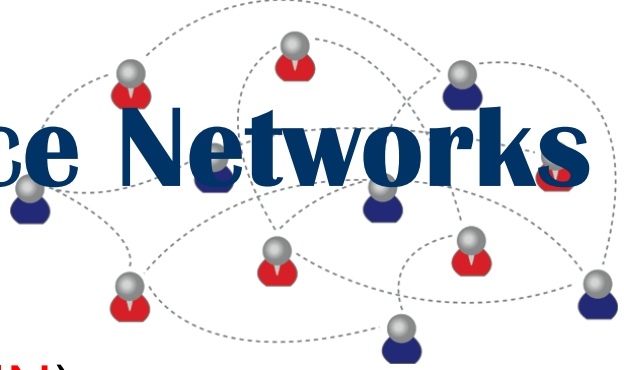


need for innovative approaches and architectures

Envisioned architecture

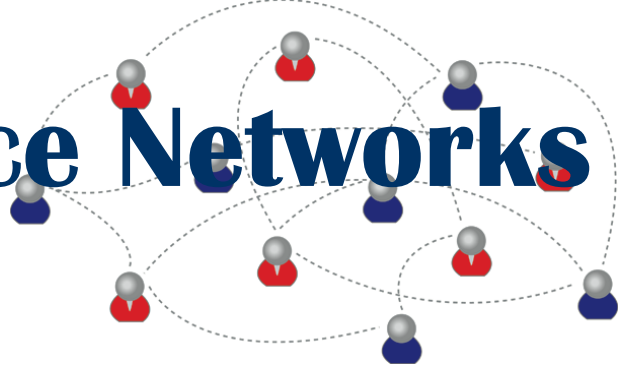
- Only a few works in the literature are focused on strategies for data warehouse integration and federation
- Problems related to data heterogeneity are usually solved by ETL processes that load data into a single multidimensional repository...
  - ...but a centralized architecture is hardly feasible in the context of collaborative BI
- **Peer Data Management Systems** (PDMSs) [Halevy et. Al., TKDE 2004]
  - architectures to support sharing of operational data across networks of peers while guaranteeing peers' autonomy
  - based on interlinked **semantic mappings** that mediate between the heterogeneous schemata exposed by peers

Business Intelligence Networks



- *Business Intelligence Network* (BIN): an architecture for sharing BI functionalities across a dynamic and collaborative network of heterogeneous and autonomous peers
- Each peer is equipped with an independent BI system, that relies on a local multidimensional schema to represent the peer's view of the business and exposes decision support functionalities aimed at sharing business information

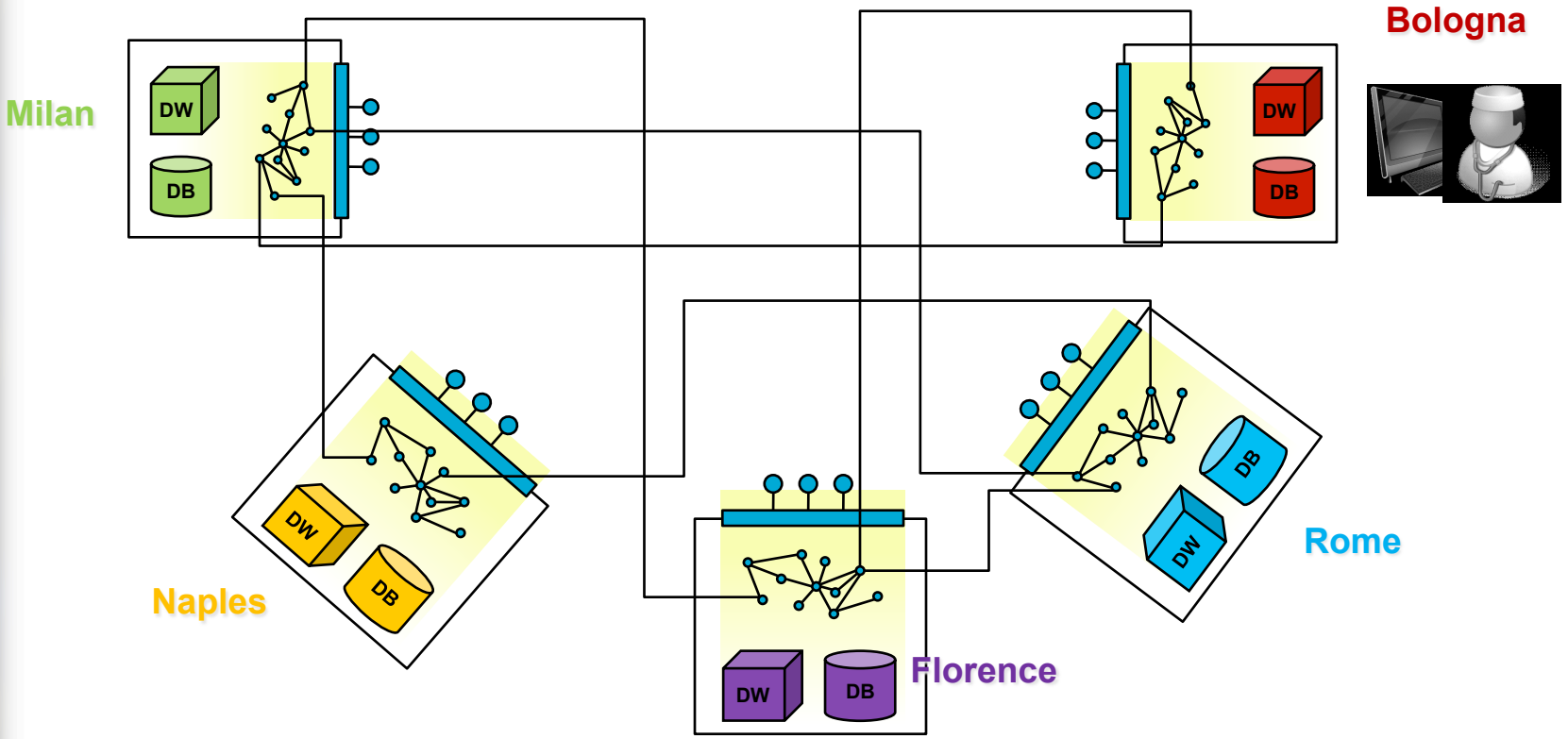
Business Intelligence Networks



- Features:
  1. Users transparently access business information distributed over the network in a pervasive and personalized fashion
  2. Access is secure, depending on the access control and privacy policies adopted by each peer
  3. Participants are collaborative, even if with different grades
  4. Inclination to collaboration does not reduce autonomy of participants, who are not subject to a shared schema
  5. A BIN is decentralized and scalable because the number of participants, the complexity of business models, and the workload can change

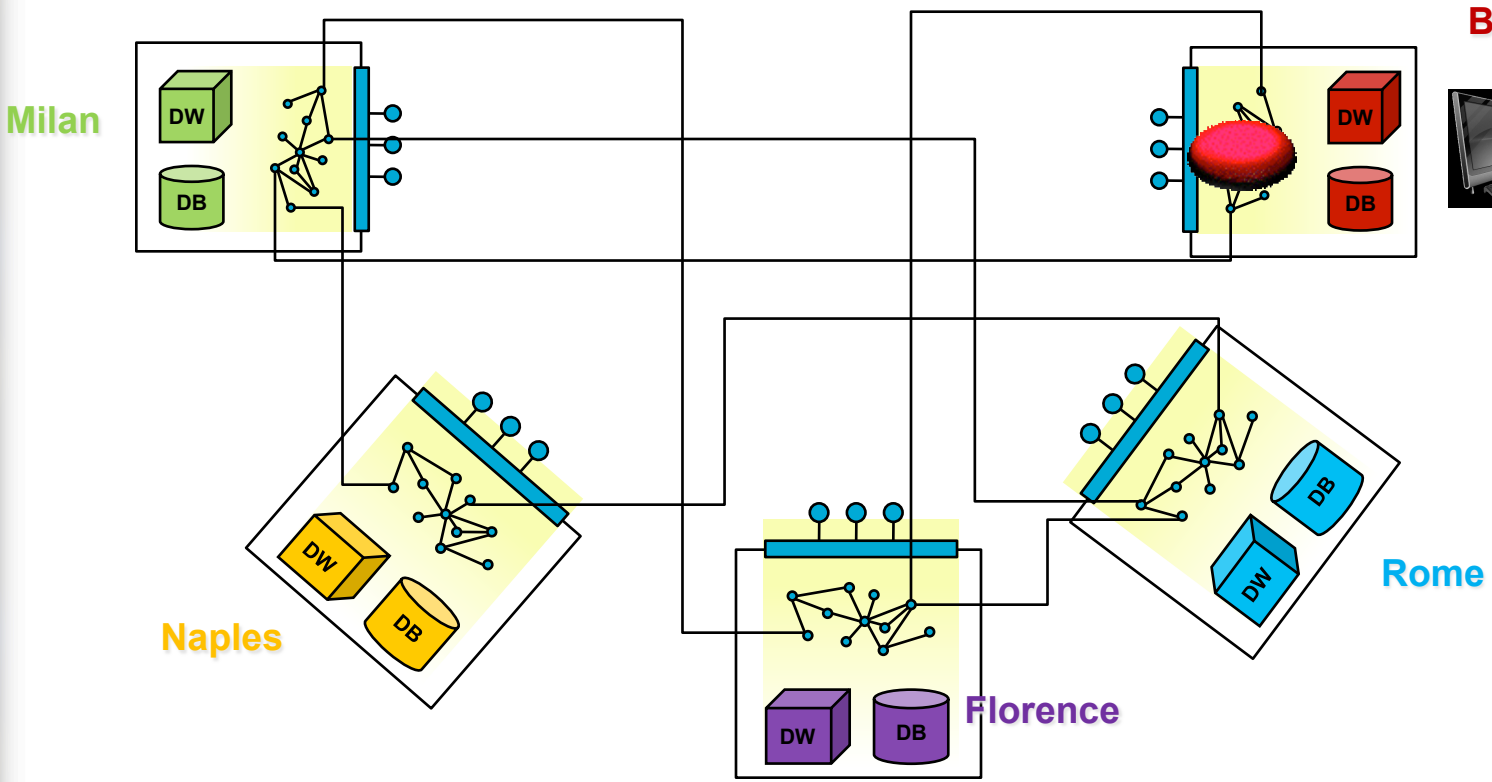
A typical user interaction

A user formulates an OLAP query *q* by accessing the local multidimensional schema of her peer, *p*



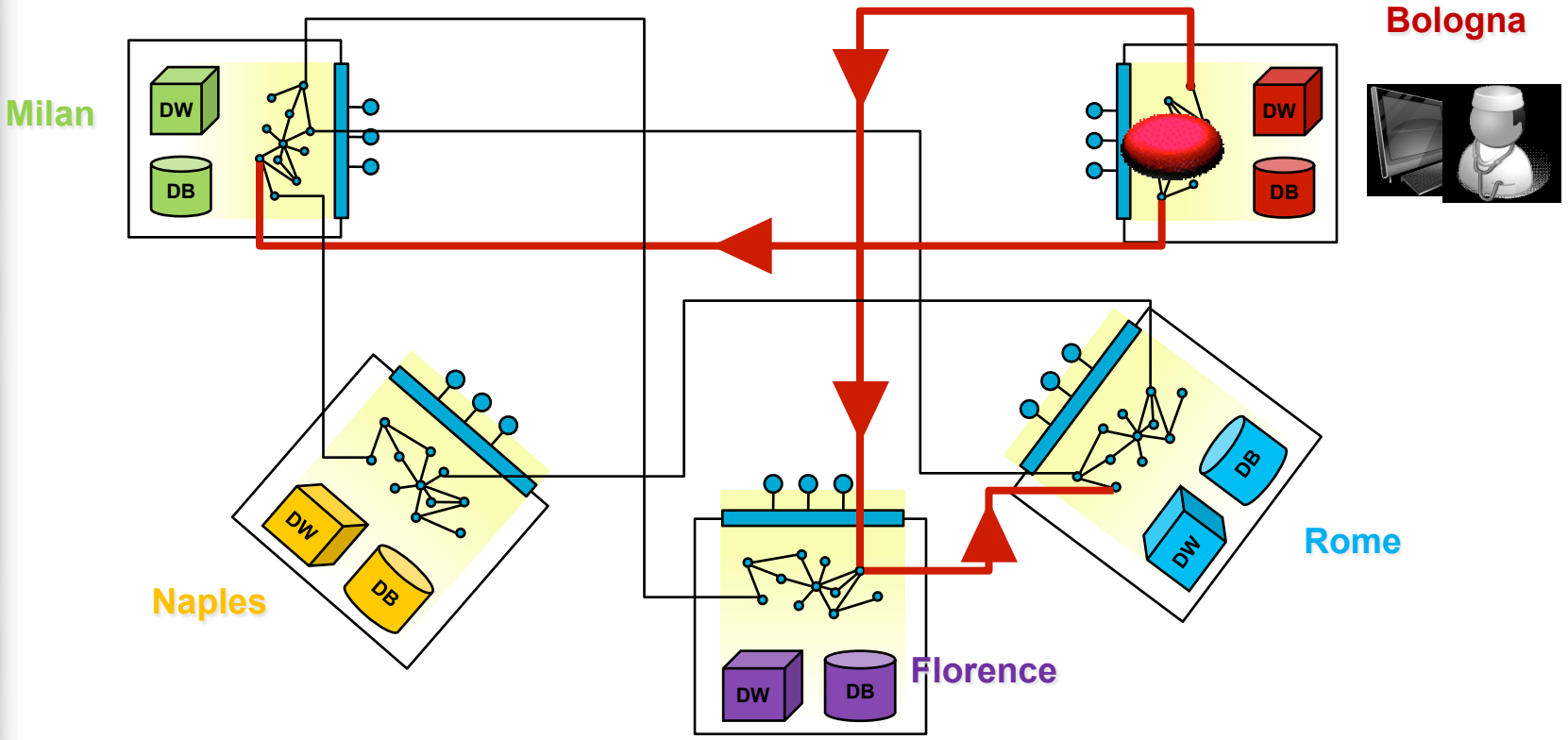
A typical user interaction

Query *q* is processed locally on the data warehouse of *p*



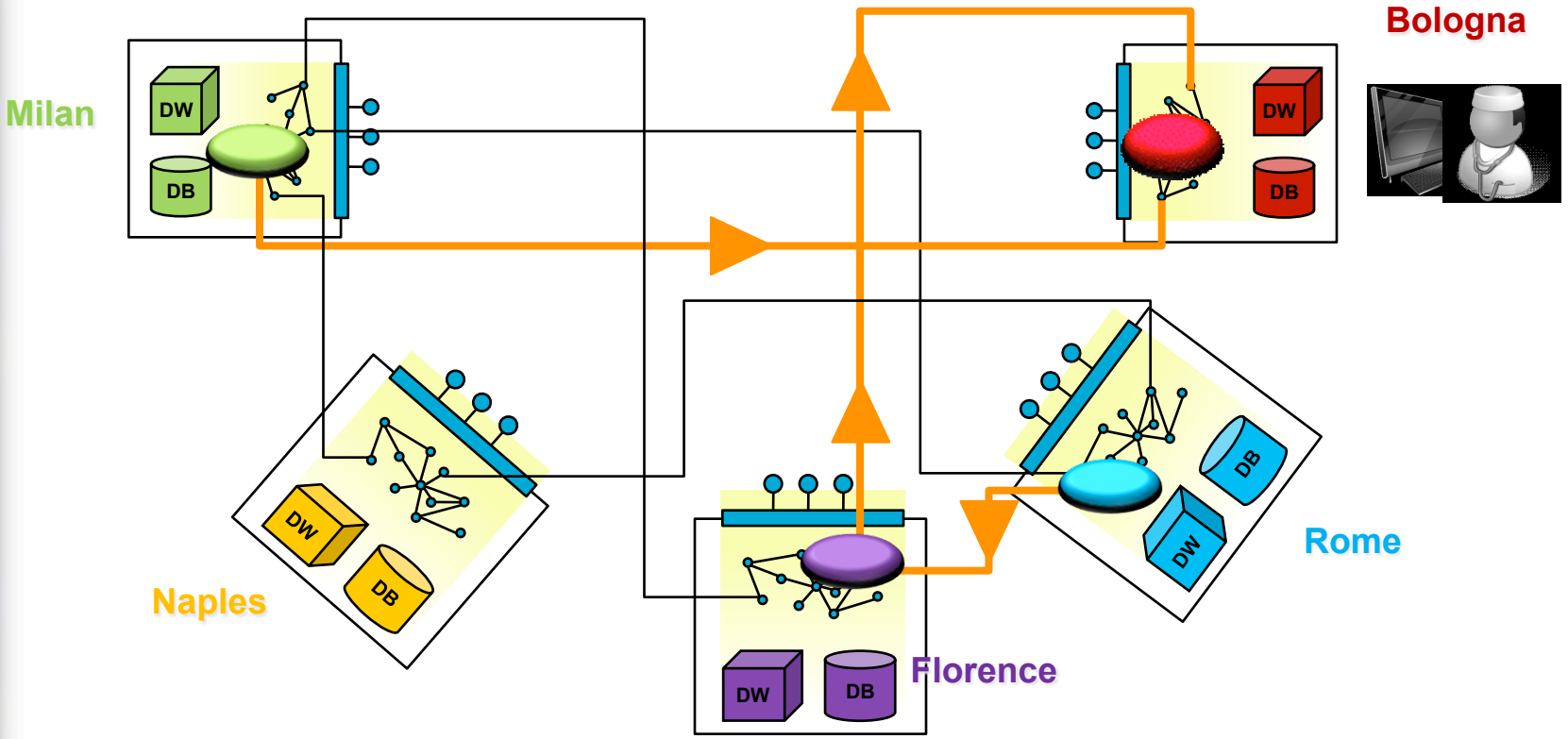
A typical user interaction

To enhance the decision making process, *q* is forwarded to the network



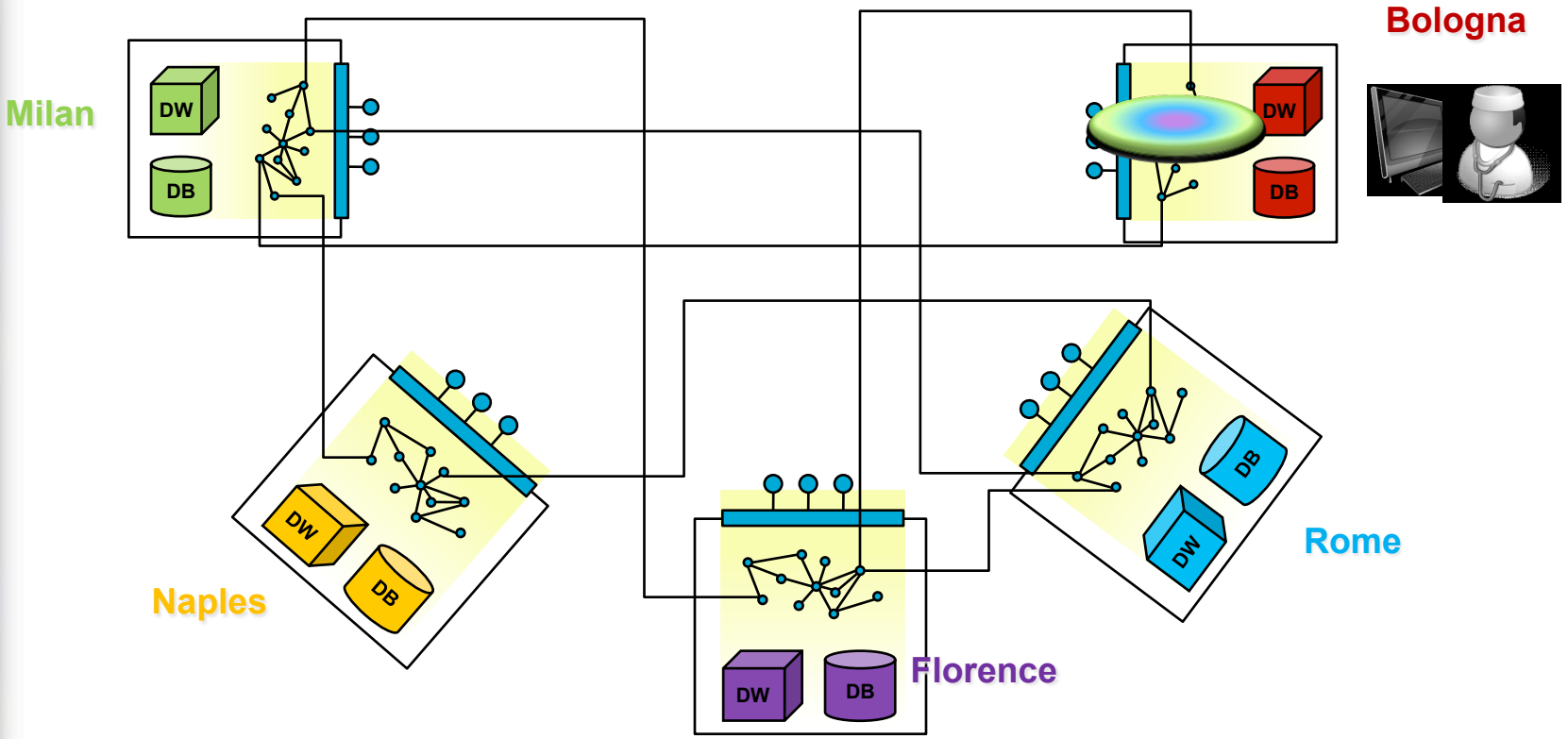
A typical user interaction

Each involved peer locally processes the query on its data warehouse and returns its (possibly partial or approximate) results to *p*



A typical user interaction

The results are integrated and returned to the user based on the lexicon used to formulate *q*

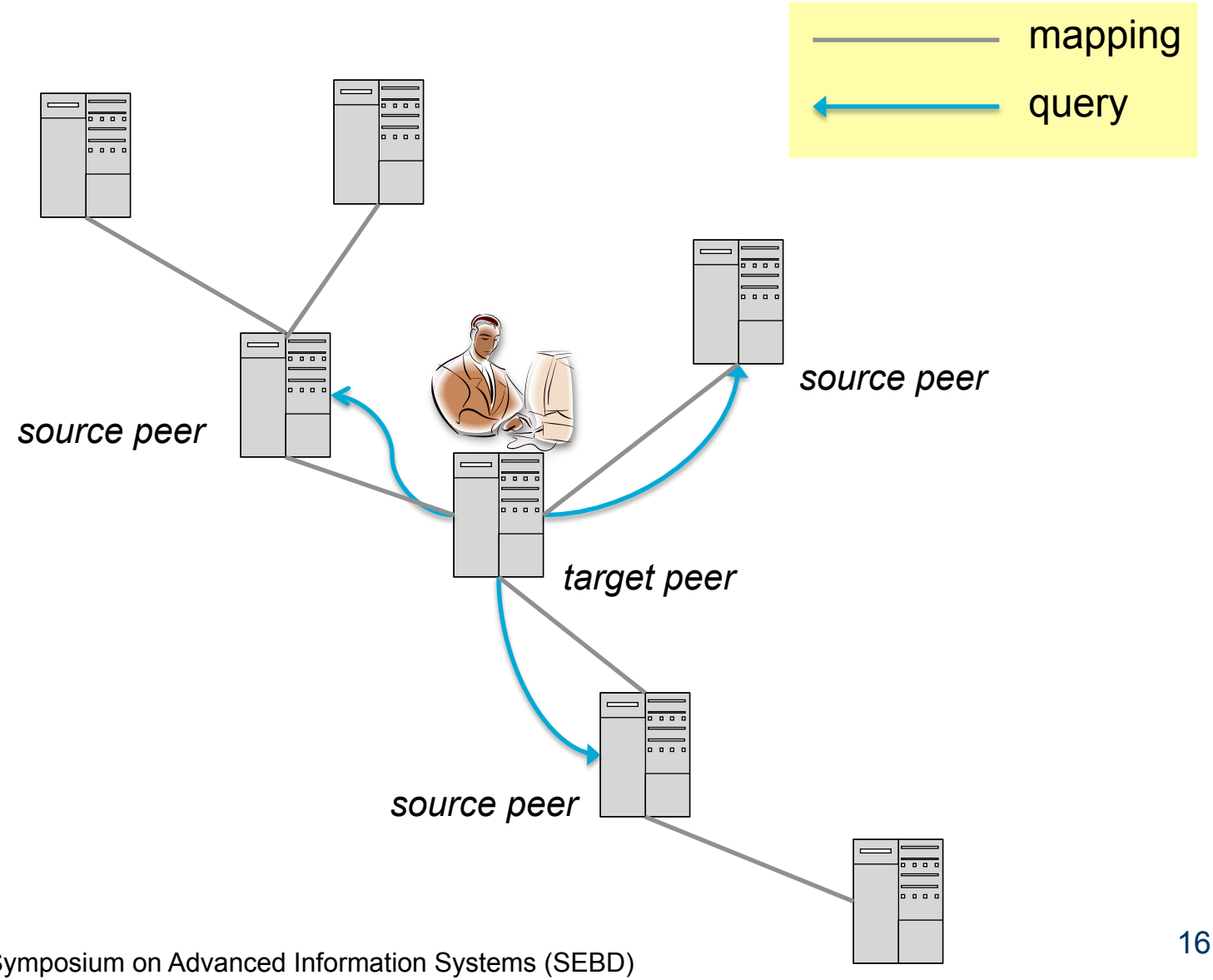


How to deal with heterogeneity?

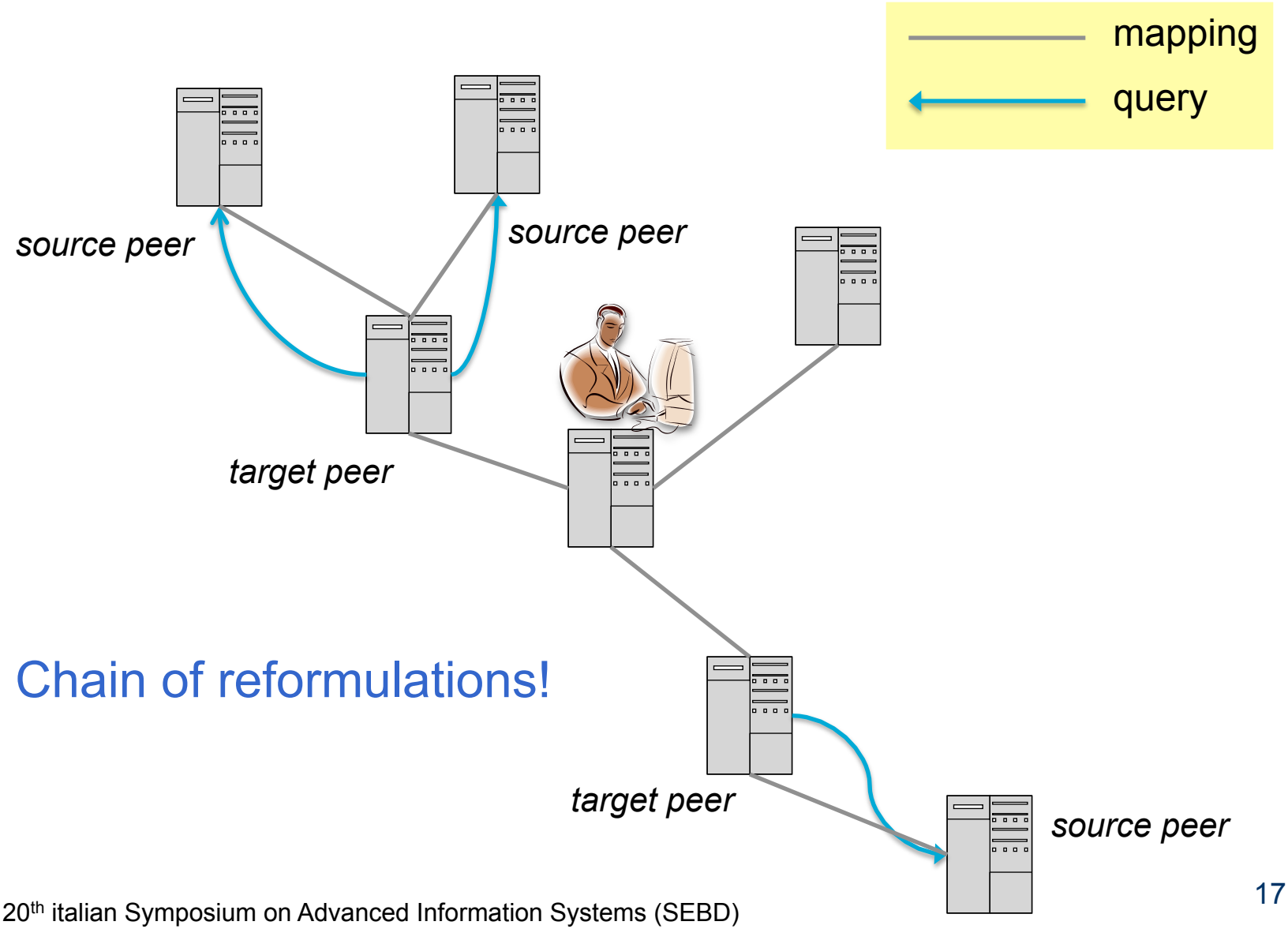
- Like in PDMSs, query reformulation in a BIN is based on *semantic mappings* that mediate between the different multidimensional schemata exposed by two peers

- Before a query issued on a peer can be forwarded to the network, it must be first reformulated according to the multidimensional schemata of the destination peers

Query reformulation

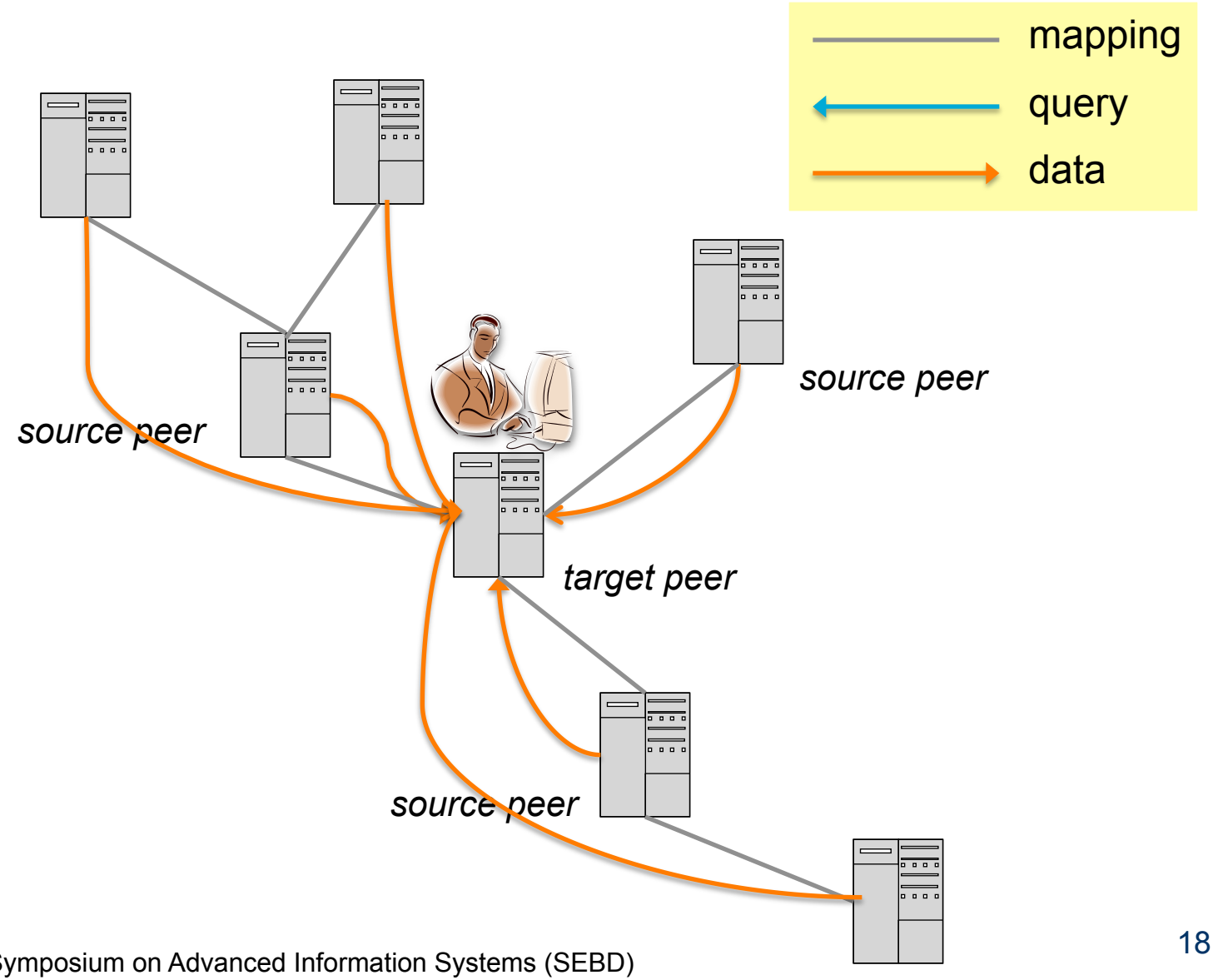


Query reformulation

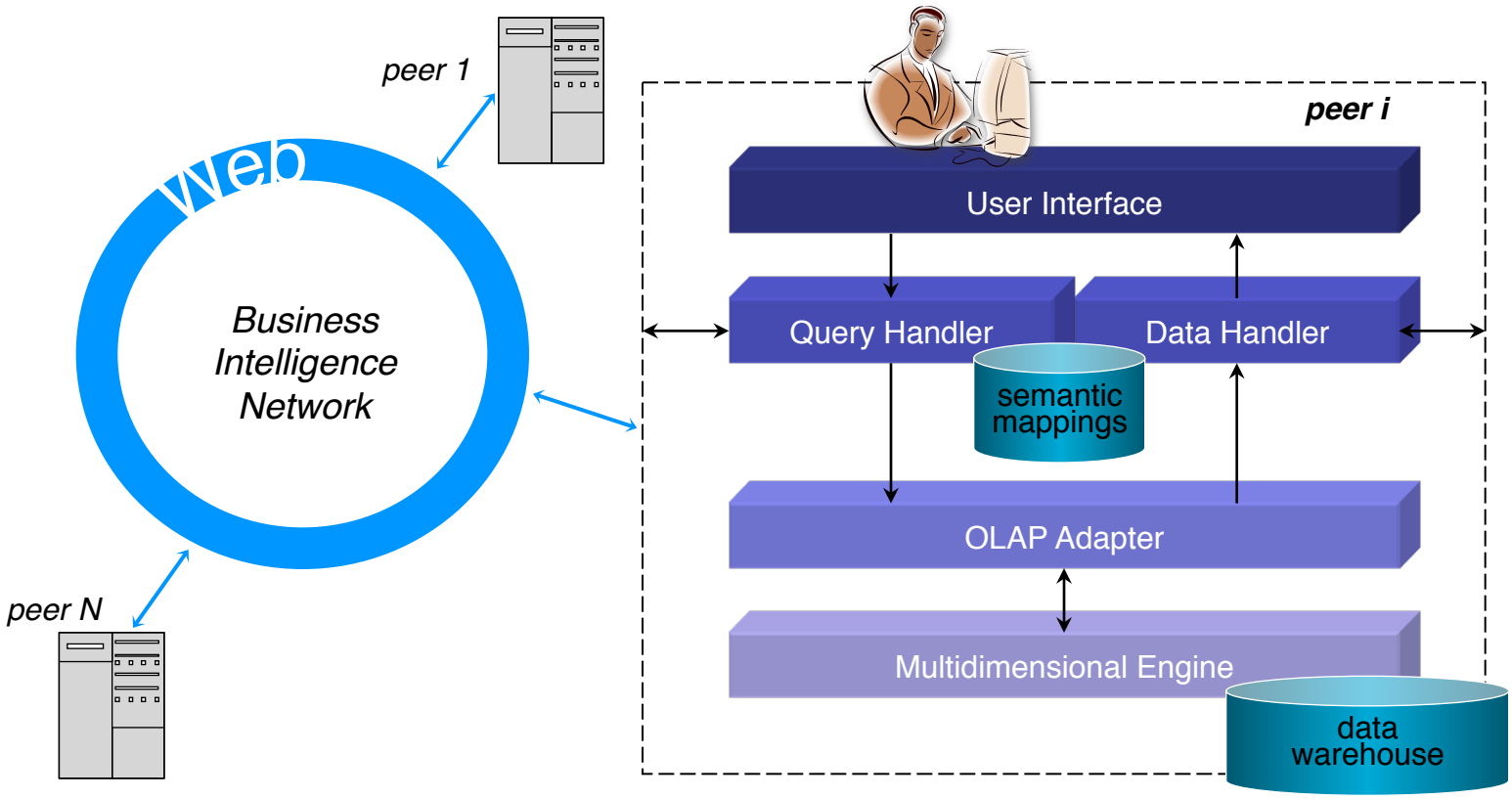


Chain of reformulations!

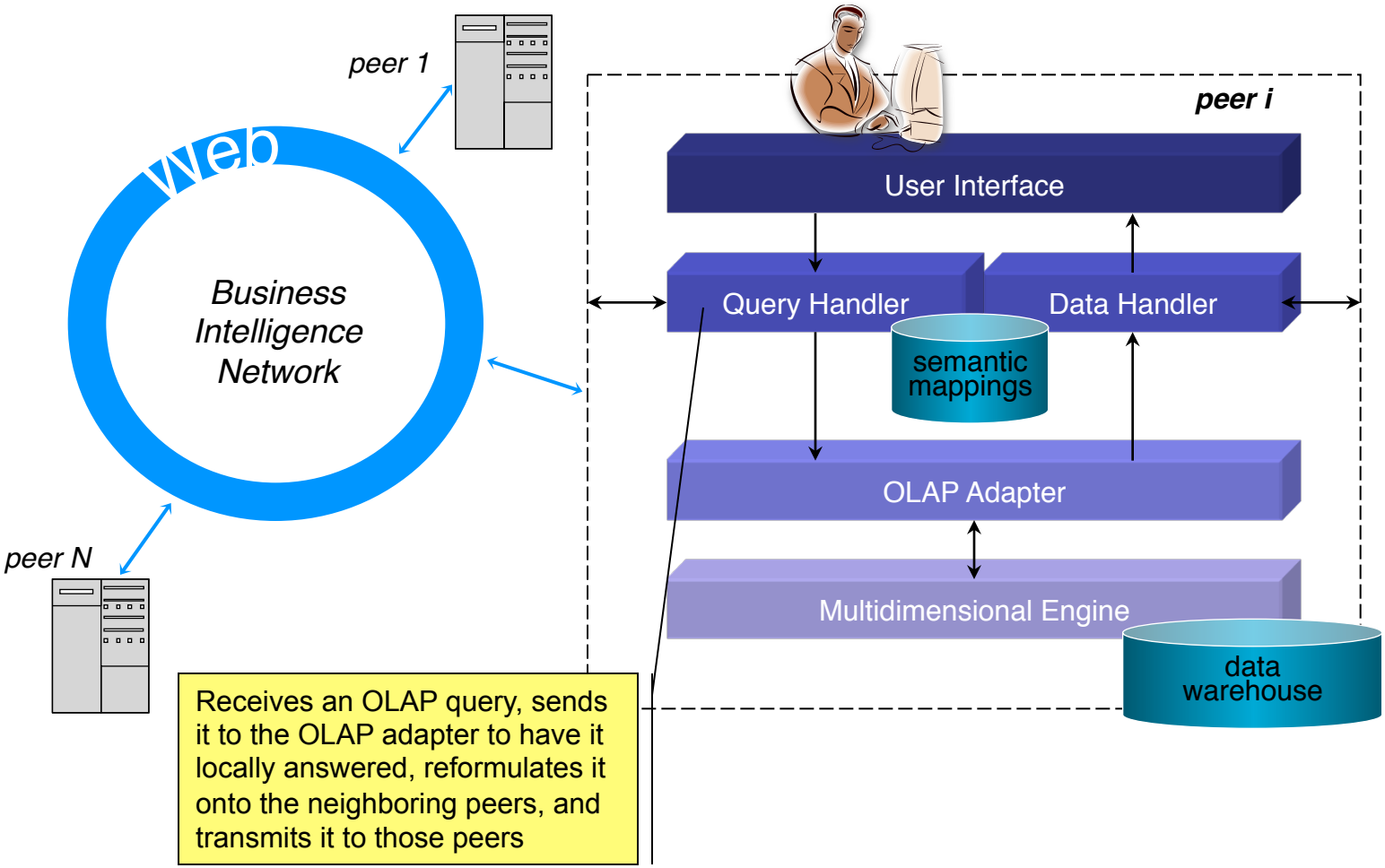
Query reformulation



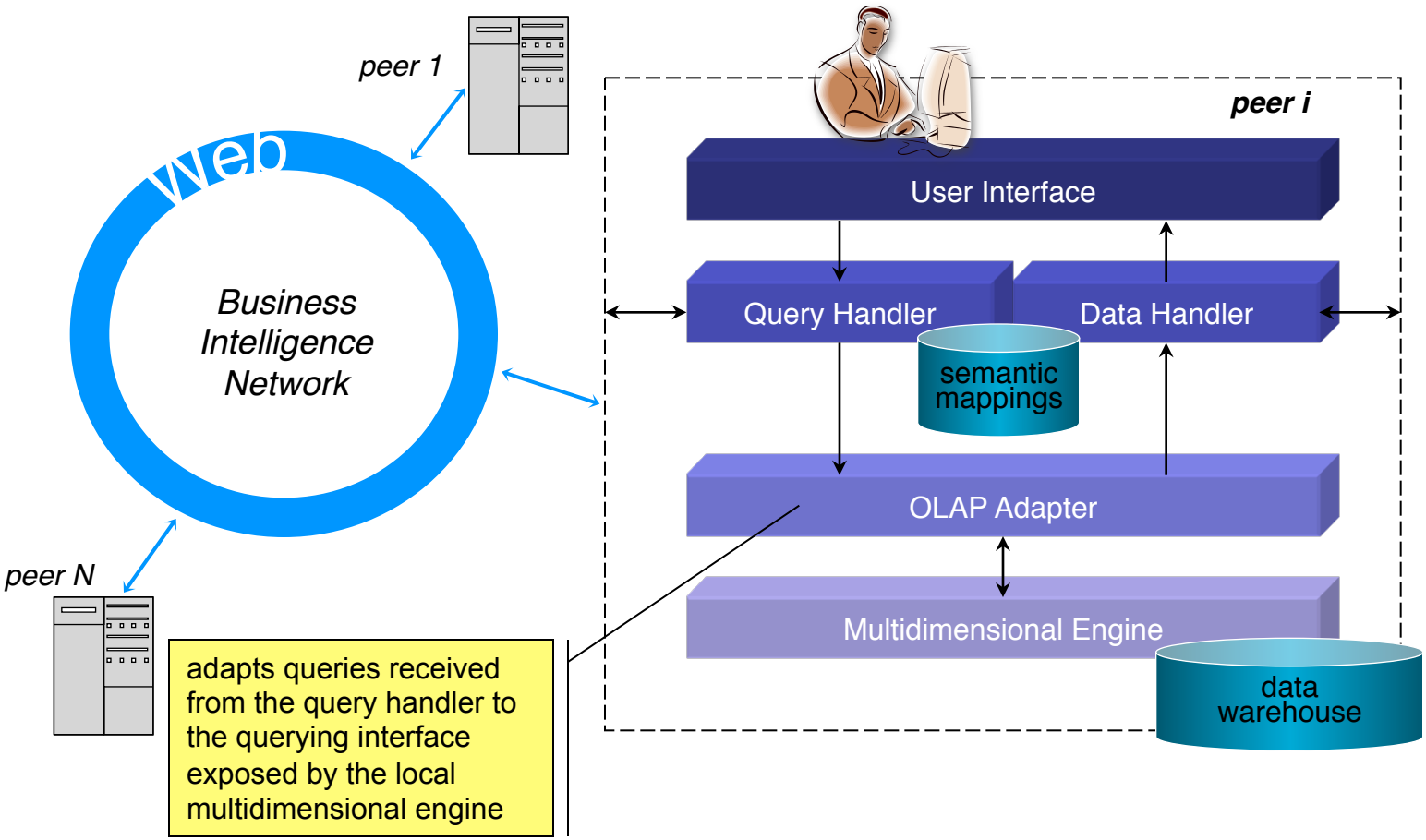
Envisioned architecture



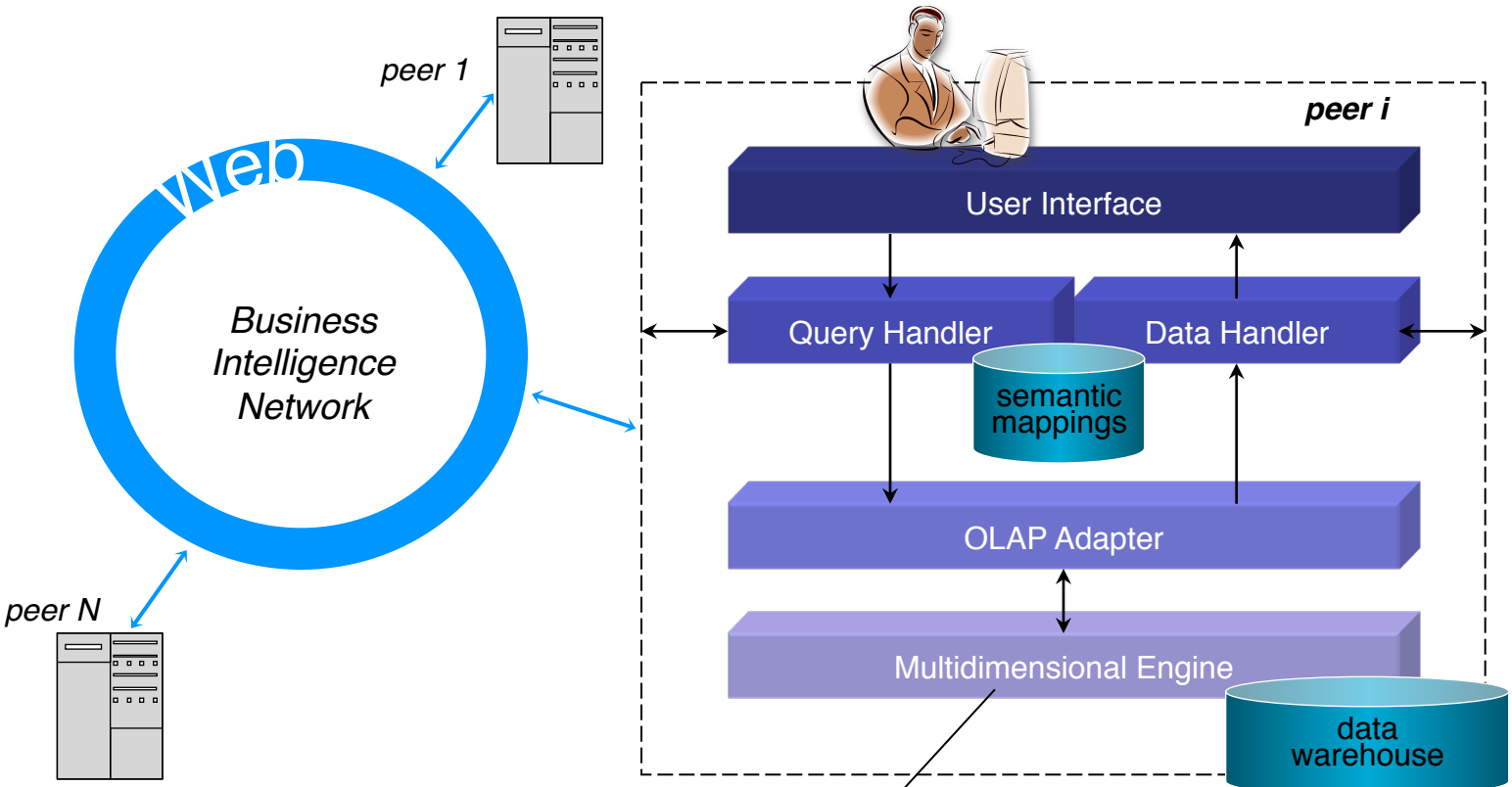
Envisioned architecture



Envisioned architecture



Envisioned architecture



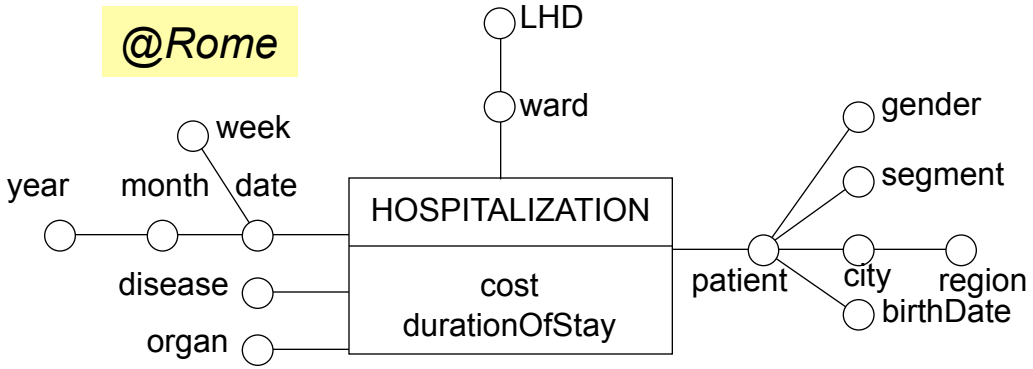
manages the local DW according to the multidimensional schema representing the peer's view of the business, and provides MDX-like query answering functionalities

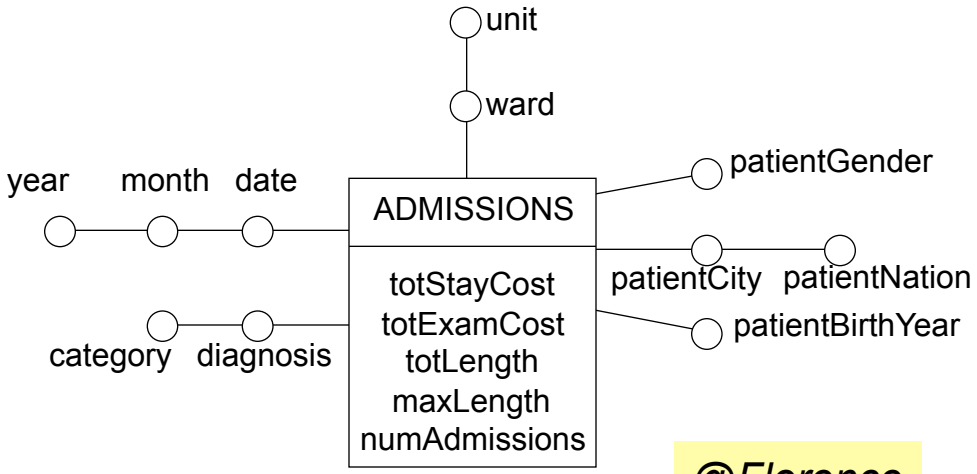
Semantic mappings

- They describe how the concepts in the multidimensional schema of one peer map onto those of another peer
- *Language requirements*:
  - Handling the asymmetry between dimensions and measures
  - Specifying the relationship between two attributes of different multidimensional schemata in terms of their granularity
  - Considering aggregation operators
  - Expressing also mappings at the instance level to transcode data

Semantic mappings

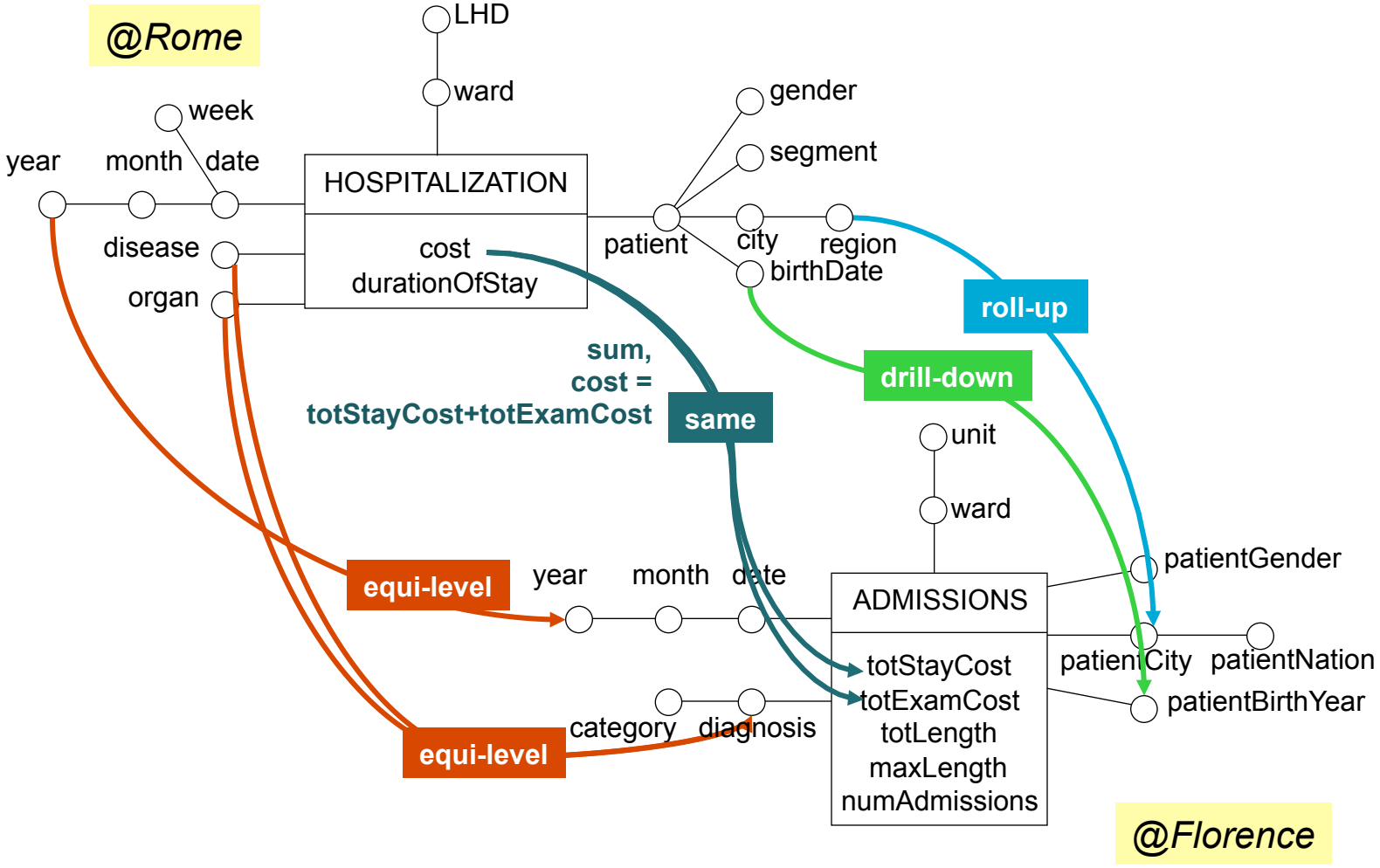
- Mapping language:





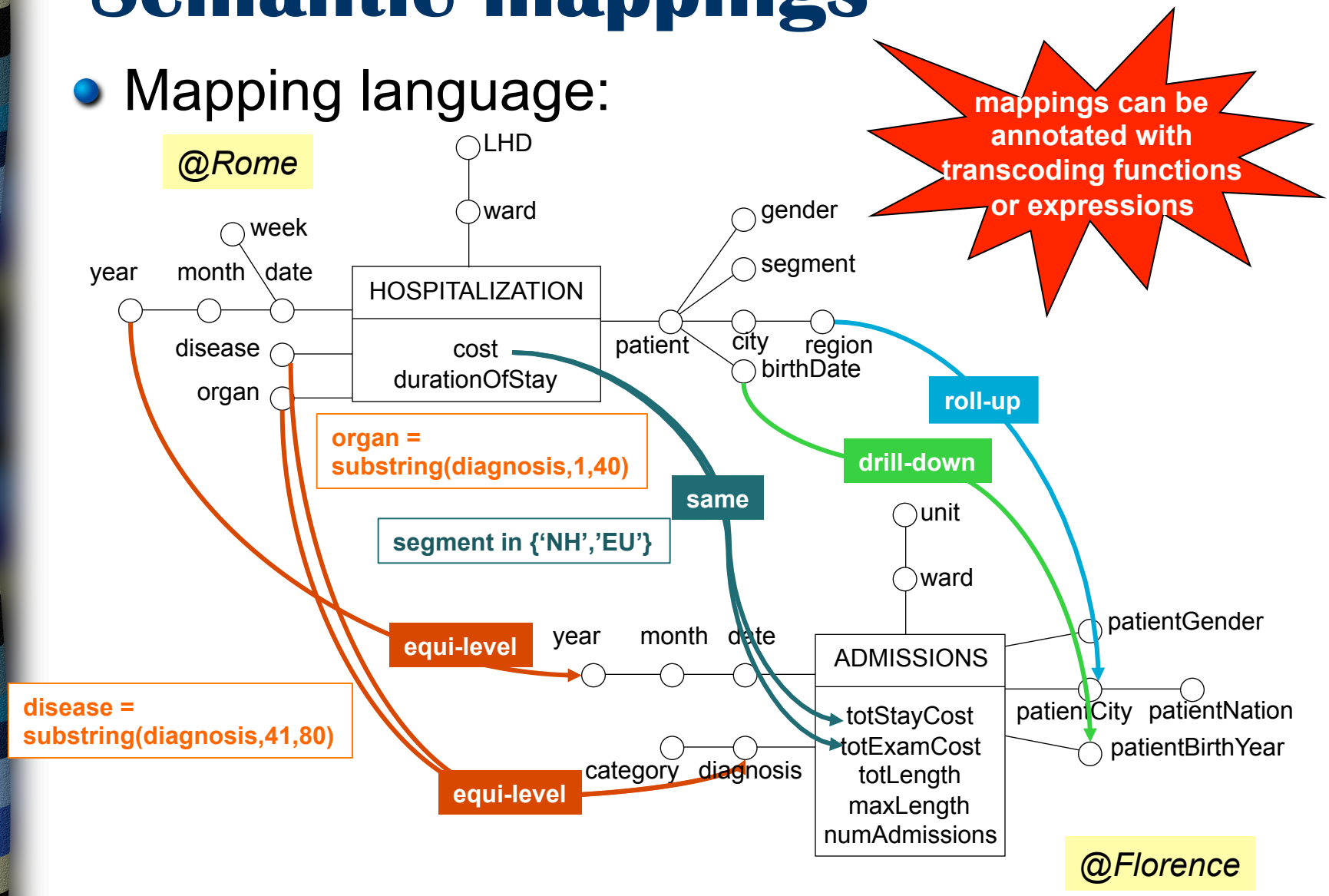
Semantic mappings

- Mapping language:



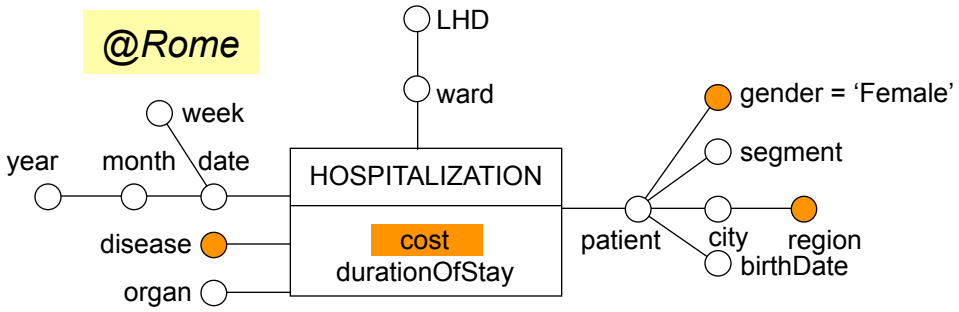
Semantic mappings

- Mapping language:



BIN Queries

"Total hospitalization costs of female patients for disease and region"

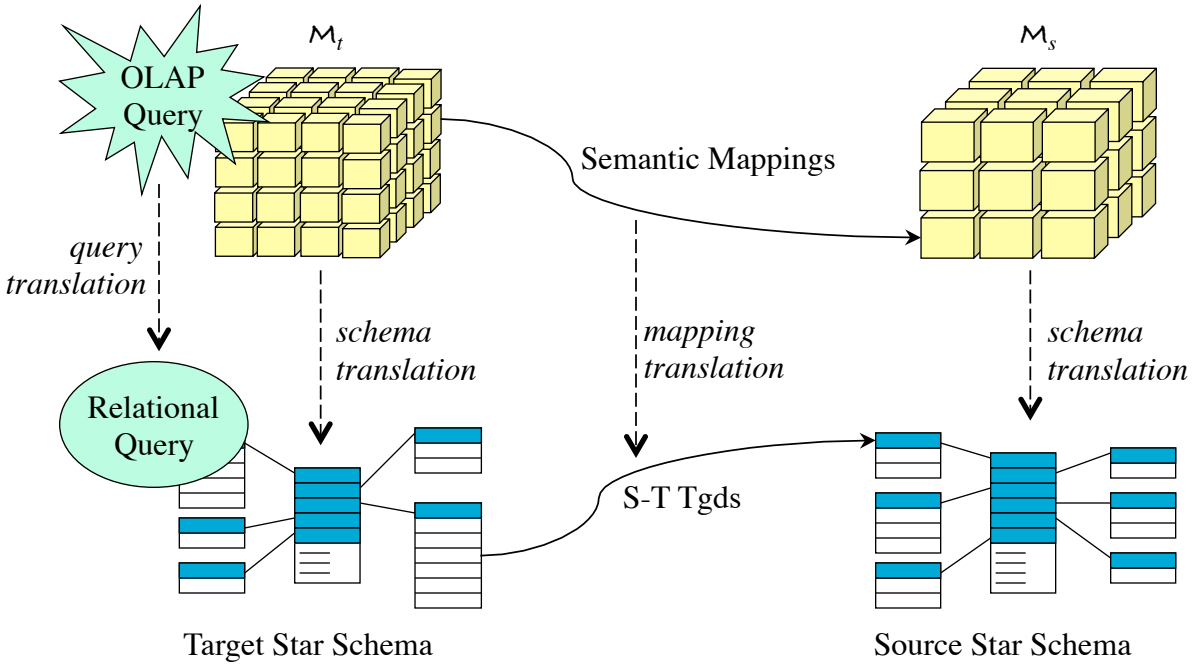


A GPSJ query

- a group-by clause
- an (optional) selection conjunctive predicate
- a numerical expression involving measures to be computed
- an aggregation operator to be used for each measure

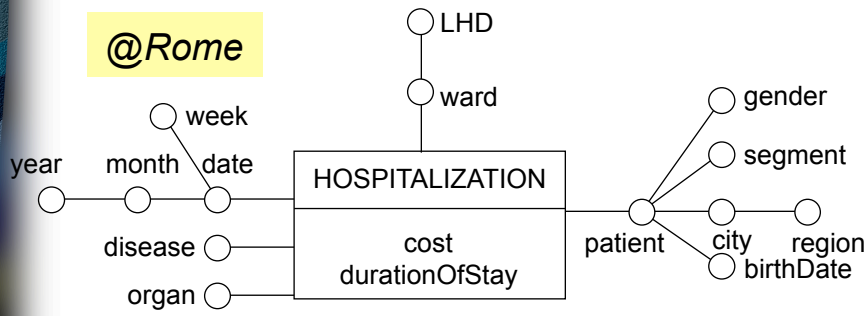
+ **transcodings** that can be applied to attributes and measures and can appear within the computed expression

Query reformulation framework

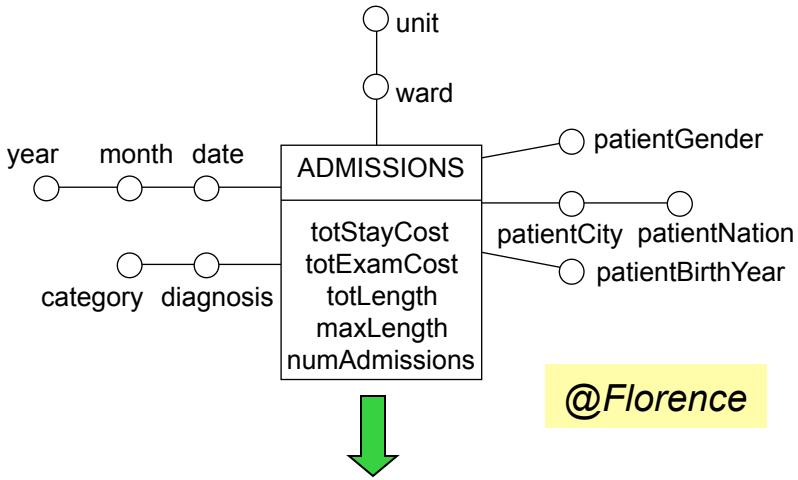


To translate semantic mappings we use a logical formalism called source-to-target tuple generating dependencies, asserting that if a pattern of facts appears in the source, then another pattern of facts must appear in the target

Example: Schema translation



HospFT(organ,disease,date,ward,patient,cost,durationOfStay)
OrganDT(organ)
DiseaseDT(disease)
DateDT(date,week,month,year)
WardDT(ward,LHD)
PatientDT(patient,birthDate,city,region,segment,gender)



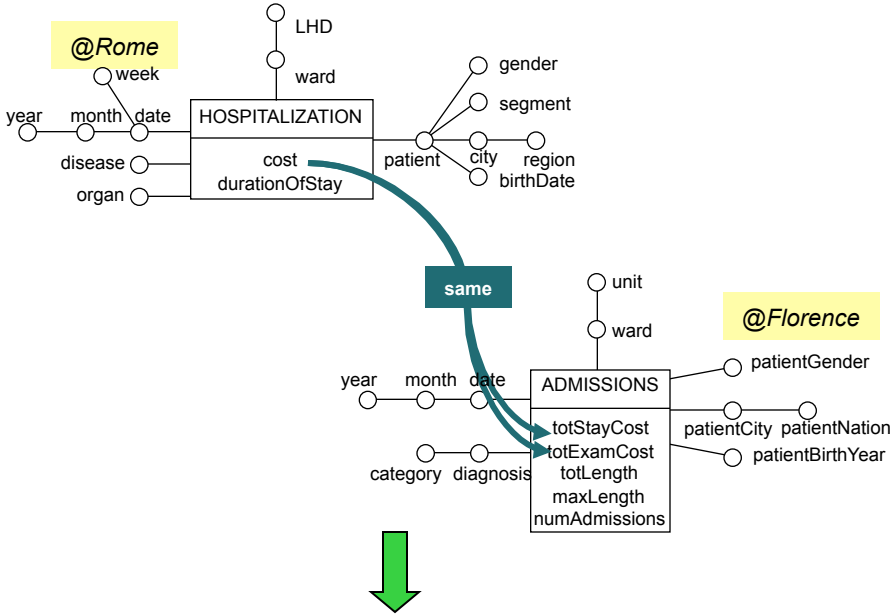
AdmFT(diagnosis,date,ward,patientCity,patientBirthYear,patientGender,
totStayCost,totExamCost,totLength,maxLength,numAdmissions)
DiagnosisDT(diagnosis,category)
DateDT(date,month,year)
WardDT(ward,unit)
PatientCityDT(patientCity,patientNation)
PatientBirthYearDT(patientBirthYear)
PatientGenderDT(patientGender)

Example: Query translation

"Total hospitalization costs of female patients for disease and region"

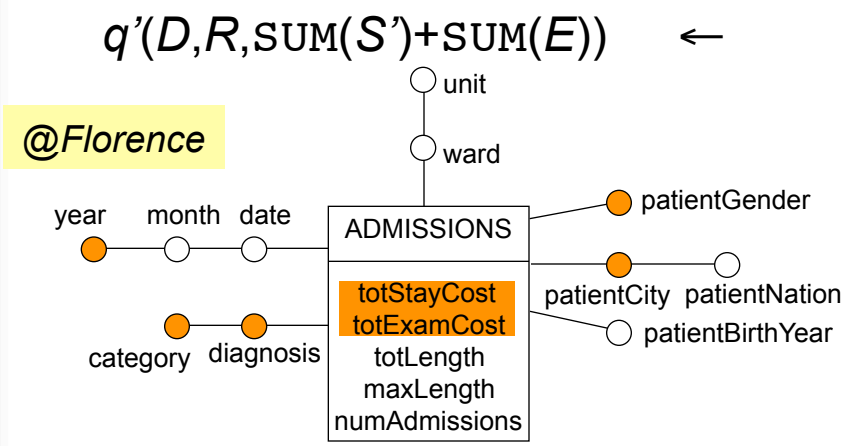
$q(R,D,\texttt{SUM}(C)) \leftarrow$ HospFT(_,D,_,_,*P*,*C*,_),
DiseaseDT(*D*),
PatientDT(*P*,_,_,*R*,_,G)),
G='Female'

Example: Mapping translation



$\forall S,E,C\ (\text{AdmFT}(\_,\ldots,S,E,\_,\_),\ C{=}S{+}E \rightarrow \text{HospFT}(\_,\ldots,C,\_))$

Example: Reformulation



AdmFT(D′,_,_,T,Y,P,S′,E, , ,),
DiagnosisDT(D′,C),
D = substring(D′, 1, 40),
C = categoryOf(substring(D′, 1, 40)),
PatientCityDT(T,_), R=regionOf(T),
PatientGenderDT(P),
completeGender(P) = 'Female'

- The group-by is reformulated using
  - the roll-up mapping from *region* to *patientCity*
  - the equi-level mapping from *disease* and *organ* to *diagnosis*
- The predicate is reformulated using the equi-level mapping from *gender* to *patientGender*
- Measure *cost* is derived using the same mapping

Inter-peer query reformulation

- The algorithm draws inspiration from the PIAZZA Q.R. Algorithm [Halevy et al., VLDBJ, 2005]
- It deals with approximate mappings, i.e. mappings that have no associated transcoding function
  - they induce no dependencies between the source and the target tuples
- Its output finds a corresponding query at the BIN level, so *the BIN query language is closed under reformulation*
- ✓ A necessary condition to implement chain of reformulations
- It is proved to be *sound and complete* with respect to the semantics of query answering, that in data sharing settings is usually given in terms of certain answers

Intra-peer query reformulation

- A BIN query cannot be directly executed on the peer local multidimensional engine
- How to bridge the language expressiveness gap between the BIN query language and the local multidimensional language?
  - Query rewriting using views
- The intra-peer reformulation algorithm
  - must deal with the presence of transcodings in the query group-by set
  - must properly manage non-distributive aggregation operators, like `avg`

Implementation issue

- How to share transcodings among peers?
  - *Public transcodings* are standard database functions that are shared by all peers
  - *Protected transcodings* are owned by a peer, that will make them available to its neighboring peers by attaching them to query messages
    - If protected transcodings are expressed as procedures, a shared programming language must be available in the BIN
    - Otherwise, transcodings can be expressed as look-up tables to be applied by a relational engine; in this case, an obvious drawback is the quantity of information to be transmitted over the network

Summary

- We have outlined a peer-to-peer architecture for supporting distributed and collaborative decision-making scenarios
- We have introduced the main query reformulation problems in the BIN context
- We have shown how an OLAP query formulated on one peer can be reformulated on a different peer, based on a set of inter-peer semantic mappings

More on BINs

- Matteo Golfarelli, Federica Mandreoli, Wilma Penzo, Stefano Rizzi, Elisa Turricchia.
  BIN: Business intelligence networks.
  In *Business Intelligence Applications and the Web: Models, Systems and Technologies*, IGI Global, 201.

- Matteo Golfarelli, Federica Mandreoli, Wilma Penzo, Stefano Rizzi, Elisa Turricchia.
  OLAP Query Reformulation in Peer-to-Peer Data Warehousing.
  Information Systems, 37(5): 393-411. 2012

Thank you for you attention

Questions?