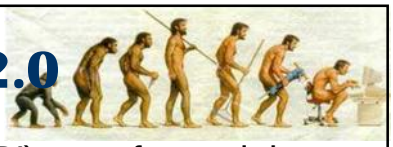
Towards OLAP Query Reformulation in P2P Data Warehousing

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

Summary

- Motivating scenario
- Envisioned architecture
- Mapping language
- A reformulation framework
- Summary and future work

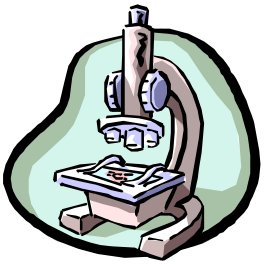
From BI 1.0 to BI 2.0



- **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

Issues in BI 2.0

- Pervasive BI
- On-demand BI
- Real-time BI
- Situational BI
- Collaborative BI ⇦
- BI as a service ⇦
- ....



Motivating scenario

- **Cooperation** 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
- Companies need strategic information about the outer world, for instance about trading partners and related business areas
- It is estimated that above 80% of waste in inter-company and supply-chain processes is due to a lack of communication between the companies involved

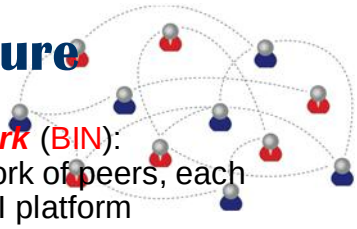
Motivating scenario

- A collaborative context where multiple partner companies organize and coordinate themselves to share opportunities, respecting their own autonomy but pursuing a common goal
  - Traditional stand-alone BI systems are not sufficient for decision making
  - Users need to *transparently* and *uniformly* access information scattered across several heterogeneous BI platforms
  - Information must be found through a semantic process and integrated on the fly

Envisioned architecture

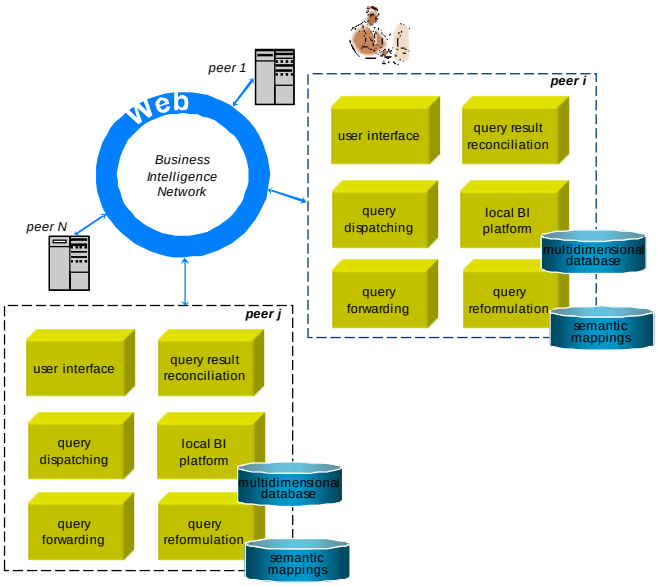
- Only a few works in the literature are focused on strategies for data warehouse integration and federation
- Problems related to data heterogeneity are usually solved by ETL processes that load data into a single multidimensional repository...
  - ...but a centralized architecture is hardly feasible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

Envisioned architecture

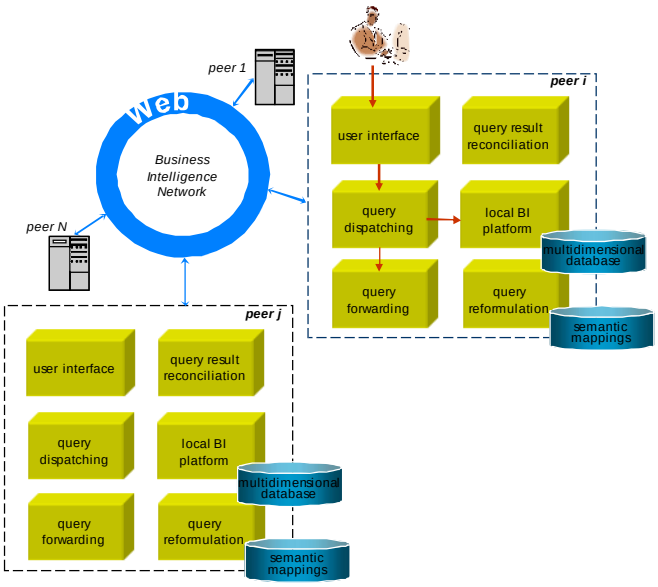


- ***Business Intelligence Network*** (BIN): a dynamic, collaborative network of peers, each hosting a local, autonomous BI platform
  1. Each peer relies on a local multidimensional schema that represents the peer's view of the business, and it offers monitoring and decision support functionalities to the other peers
  2. Users transparently access business information distributed over the network in a pervasive and personalized fashion
  3. Access is secure, depending on the access control and privacy policies adopted by each peer
  4. Participants are collaborative, even if with different grades
  5. Inclination to collaboration does not reduce autonomy of participants, who are not subject to a shared schema
  6. A BIN is decentralized and scalable because the number of participants, the complexity of business models, and the workload can change

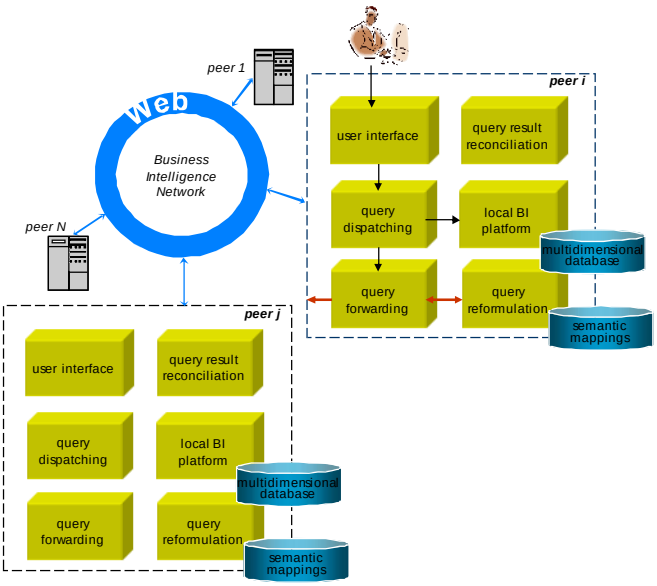
Envisioned architecture



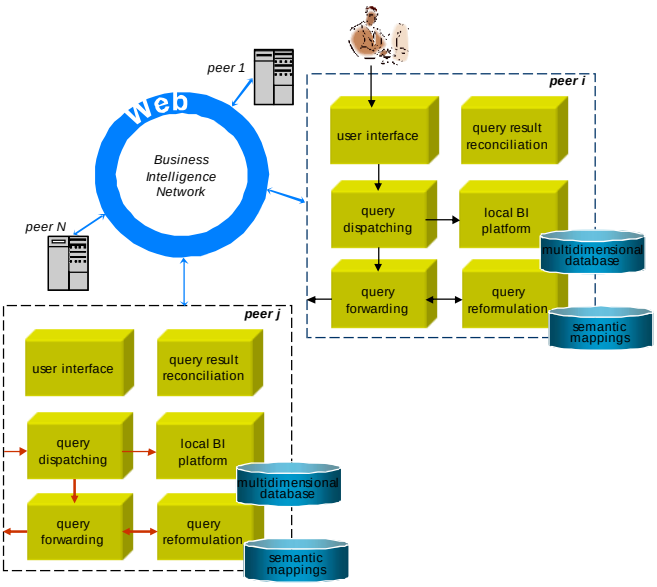
Envisioned architecture



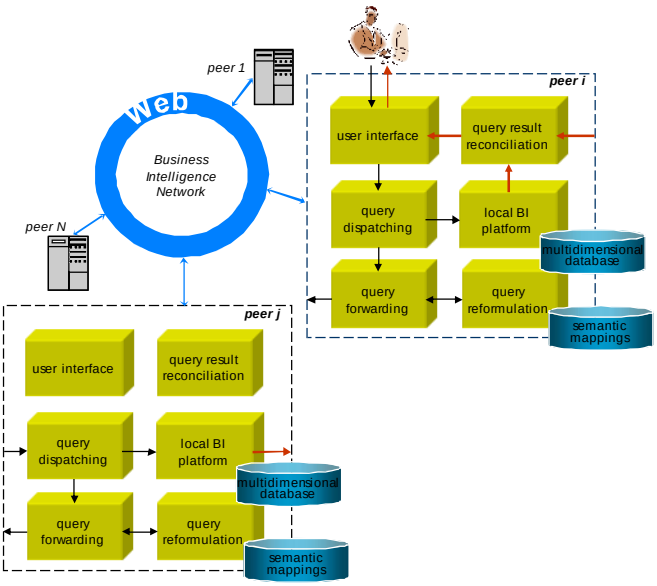
Envisioned architecture



Envisioned architecture



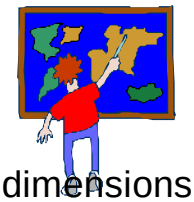
Envisioned architecture



Research issues

- Query reformulation on peers
- Query routing strategies to forward queries to the most promising peers only
- Advanced approaches for security
- Mechanisms for controlling data provenance and quality
- Data lineage to help users understand how data have been transformed
- Object fusion techniques to integrate returned data

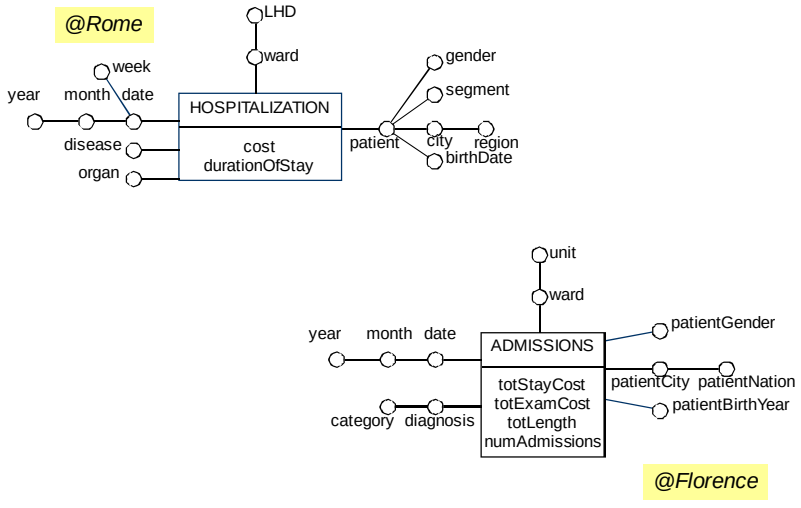
Mapping language



- 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 to avoid the risk of inconsistent query reformulations
- Expressing also mappings at the instance level to transcode data

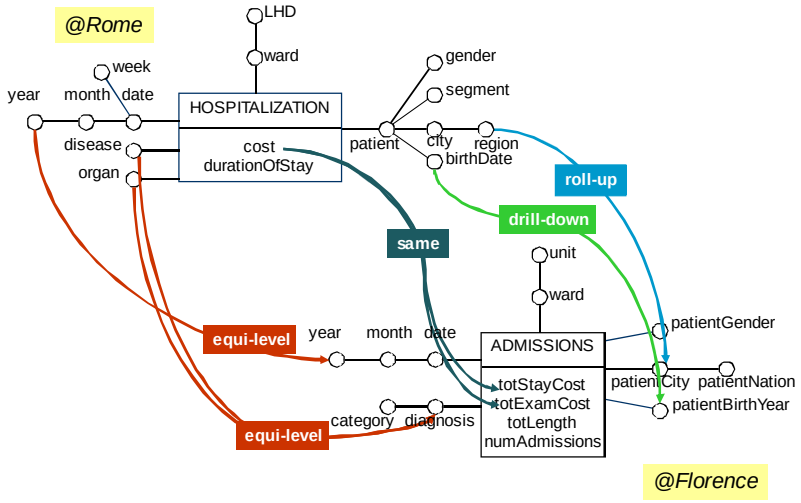
Mapping language

- Health-care example



Mapping language

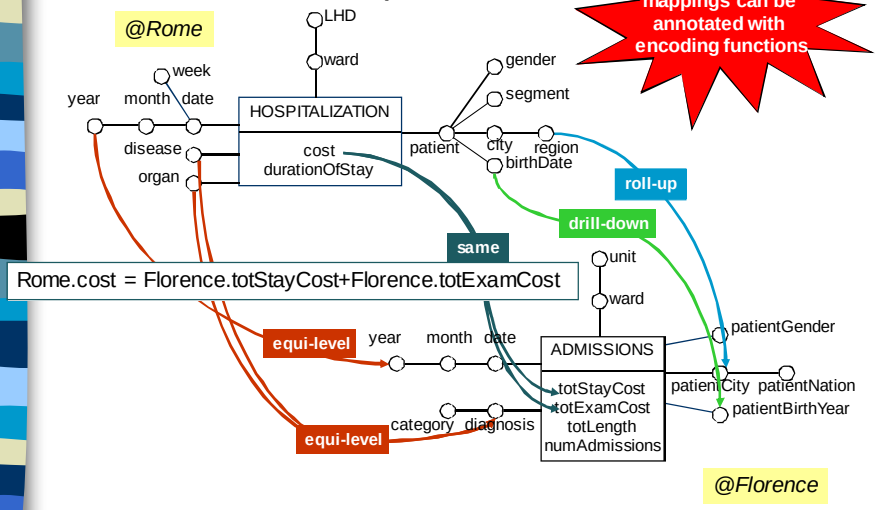
- Health-care example



Mapping language

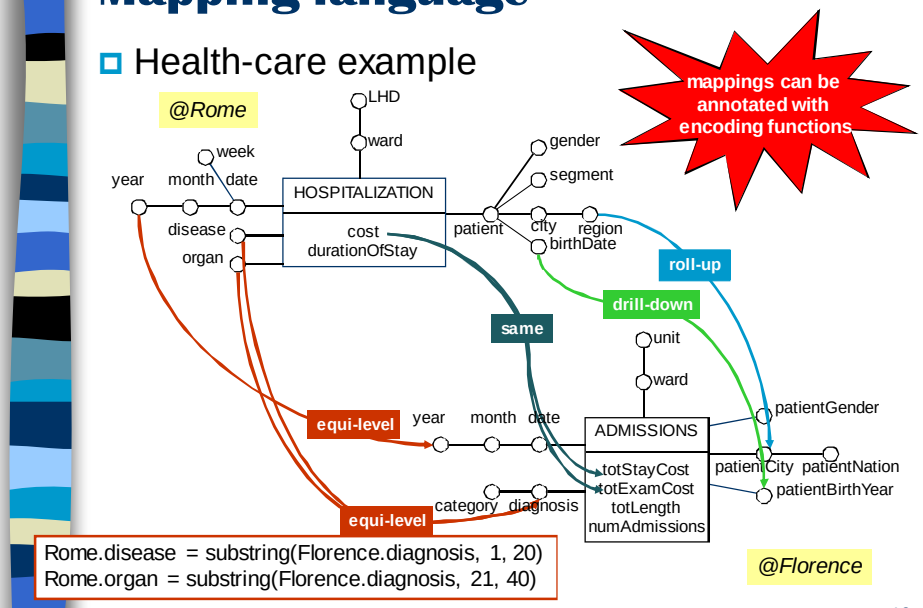
- Health-care example

mappings can be annotated with encoding functions



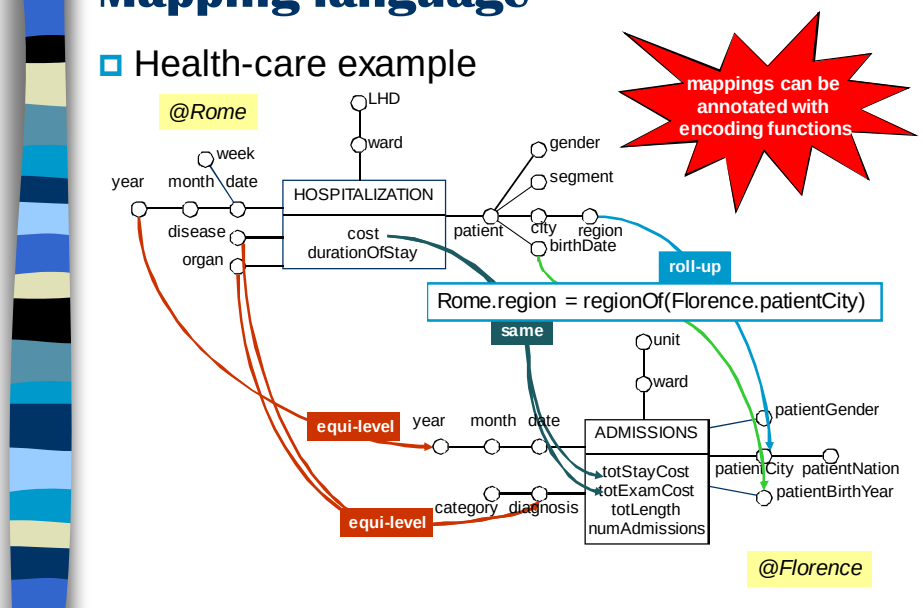
Mapping language

- Health-care example



Mapping language

- Health-care example

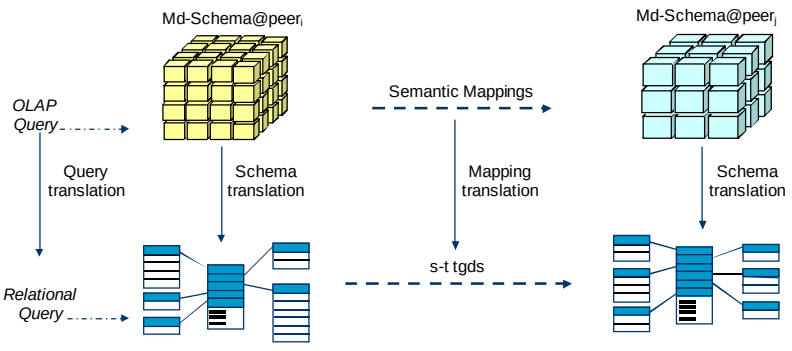


Query Reformulation Framework

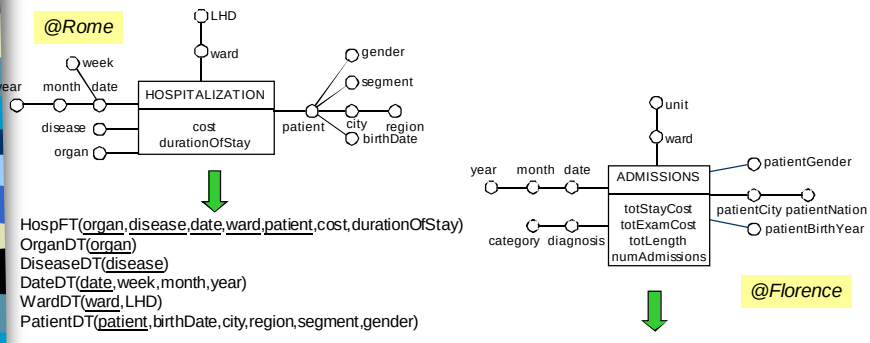
- Reformulation takes in input an OLAP query on a *target* schema *t* and the mappings between *t* and the schema of one of its neighbors, the *source* schema *s*, and it outputs an OLAP query that refers only to *s*
- In our approach, queries are reformulated by means of the underlying relational (star) schema

Query Reformulation Framework

- To translate semantic mappings we use a logical formalism called source-to-target tuple generating dependencies



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)

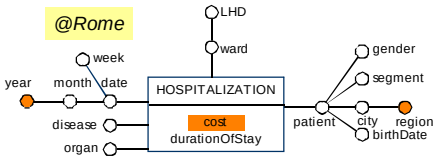
AdmFT(diagnosis,date,ward,patientCity,patientBirthYear,patientGender,
totStayCost,totExamCost,totLength,numAdmissions)
DiagnosisDT(diagnosis,category)
DateDT(date,month,year)
WardDT(ward,unit)
PatientCityDT(patientCity,patientNation)
PatientBirthYearDT(patientBirthYear)
PatientGenderDT(patientGender)

Example: Query translation

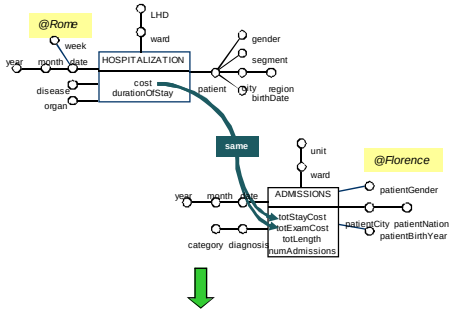
- Total hospitalization costs for region and year

$\pi_{region,year,\mathtt{SUM}(cost)}$ (HospFT ⋈ DateDT ⋈ PatientDT)

$q(R,Y,\mathtt{SUM}(C)) \leftarrow$ HospFT(_,_,D,_,P,C,_),
DateDT(D,_,_,Y),
PatientDT(P,_,_,R,_,_))



Example: Mapping translation

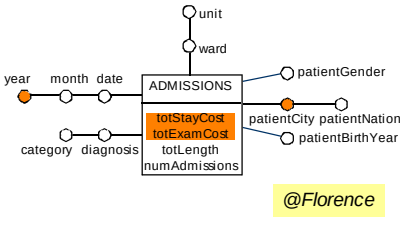


$\forall S,E,C$ (AdmFT(_,...,S,E,_,_), C=S+E→HospFT(_,...,C,_)

Example: Reformulation

- The group-by is reformulated using the roll-up mapping from *region* to *patientCity*, while measure *cost* is derived using the same mapping

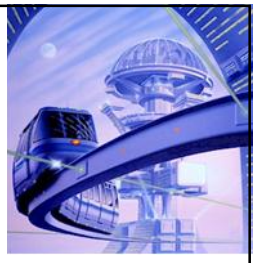
$\pi_{\text{year,regionOf(patientCity),SUM(totStayCost+totExamCost)}}$
(AdmFT ⋈ DateDT ⋈ PatientCityDT)



Summary

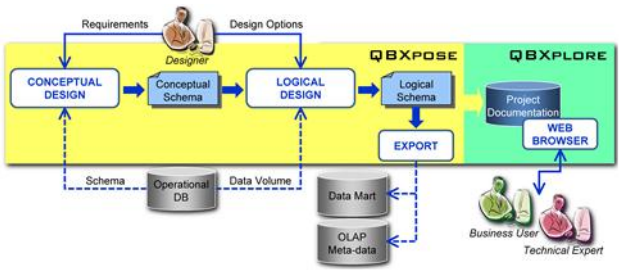
- We have outlined a peer-to-peer architecture for supporting distributed and collaborative decision-making scenarios
- We have shown how an OLAP query formulated on one peer can be reformulated on a different peer, based on a set of inter-peer semantic mappings

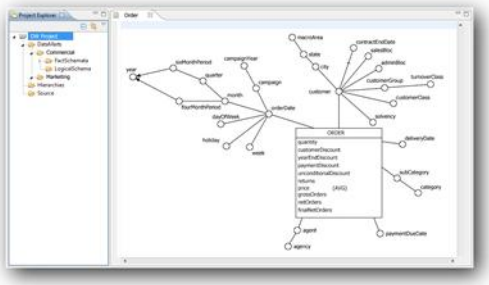
Future work



- Adopting MDX as the language for expressing multidimensional queries and closing the reformulation query language
- Devising multidimensional-aware object fusion techniques for integrating data returned by different peers
- Finding smart algorithms for routing queries to the most ñpromisingòpeers in the BI network

Some advertising...





Thank you for you attention

Questions?