

---

# Brokering multisource data with quality constraints

*Danilo Ardagna*  
Cinzia Cappiello  
Chiara Francalanci  
Annalisa Groppi



Politecnico di Milano, Italy

---

# Introduction

---

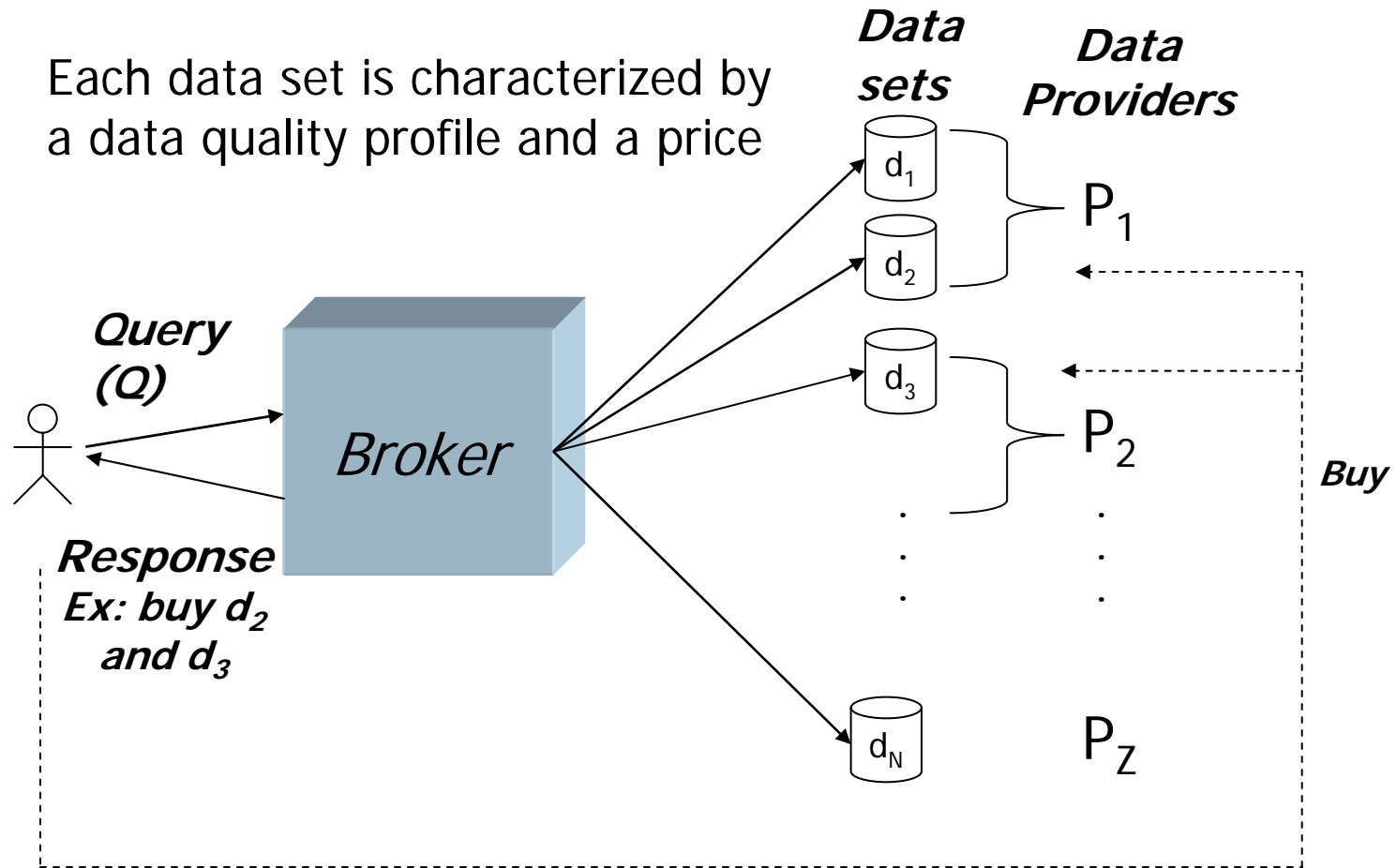
- Access to multisource heterogeneous data is a fundamental research issue
- Brokering approach to multisource data access provides greater flexibility with respect to the more traditional data integration
- The broker is submitted a query and has the responsibility to optimize the response
- Data quality perspective on data brokering focus on data accuracy
- Comparison between data visibility and data transparency approaches

# Outline

---

- Data broker architecture
- Data Trasparency and Visibility Approaches
- Data brokering methodology
- Experimental results
- Conclusions and Future work

# Broker architecture



# The broker model

---

- Local-As-View (LAV) perspective, representing the schema of a source as a view of a global schema (GS)
- The GS is defined as a set of relations, called Global Relations, which are tied to each other by join attributes
- The Universal Relation (UR) is defined as the join of all global relations
- Users specify the minimum level of quality that they consider acceptable  $QD_r^*$  and rank the quality dimensions by specifying a set of normalized weights
- A constraint  $q^*$  on the overall value of quality  $q$  is also specified

# Data Transparency and Visibility Approaches

---

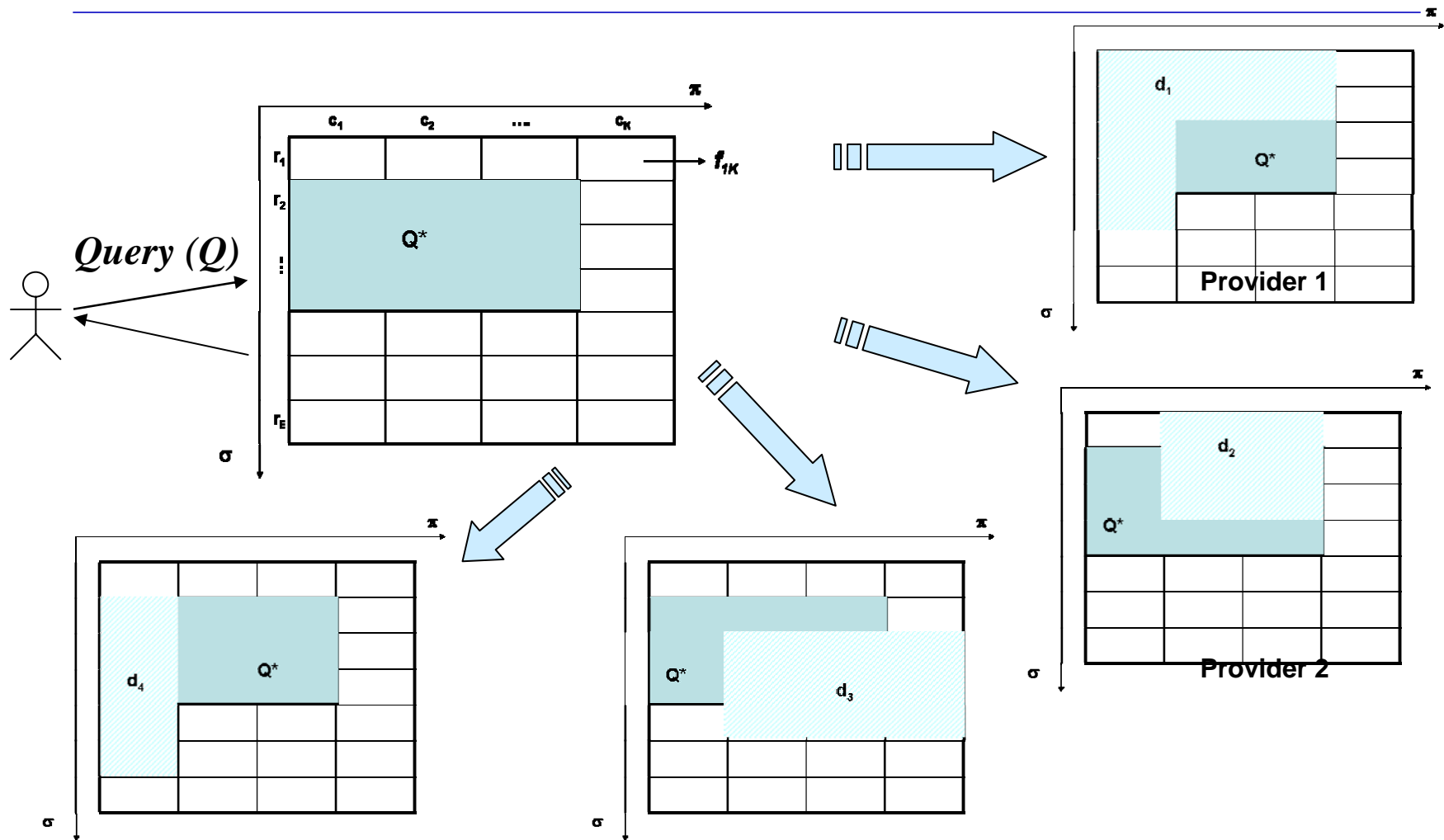
- DTA, the broker:
  - ▶ cannot view data values
  - ▶ it can view metadata
  - ▶ knows the quality of local data sets and the extent to which local data sets overlap with each other
- DVA, the broker:
  - ▶ can view the data values supplied by providers
  - ▶ quality values are associated with values of attributes included in the data sets
  - ▶ can perform data cleaning activities to improve the quality of local data and perform quality-based merging of local data

# Data Transparency Approach

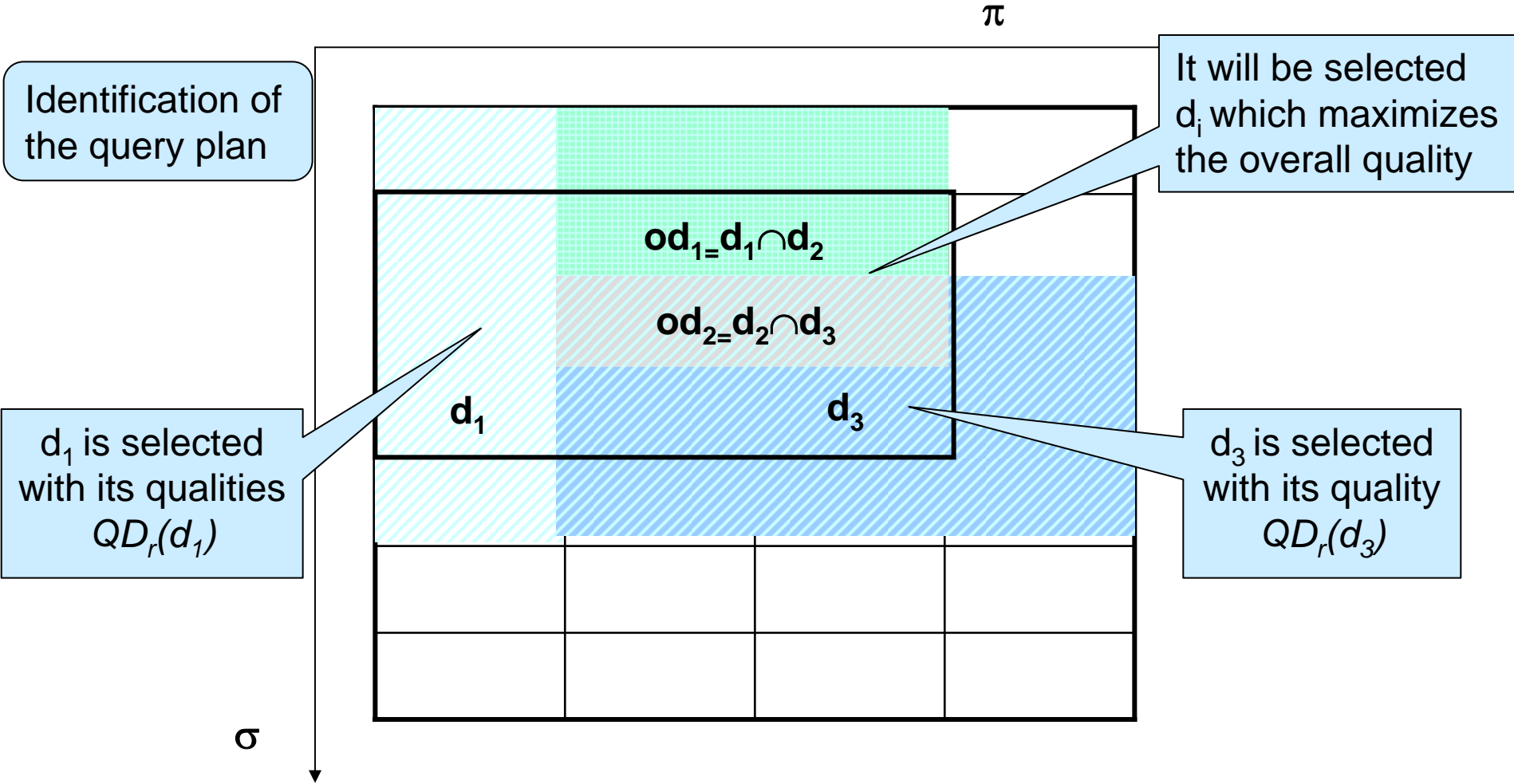
---

- Local data sets are divided into fragments and the broker knows the cardinality of all data fragments
- The response  $Q^*$  to query  $Q$  is a set of fragments
- Each data set  $d_i$  represents the smallest subset of data with which a provider can contribute to satisfying  $Q$
- Data quality is homogeneous within each data set, providers are responsible for the evaluation of the quality of their data
- The response to query  $Q$ , can be built by using multiple combinations of  $d_i$ , i.e query plans

# Identification of all the suitable provider



# Evaluation of the quality of data from multiple sources



# Data Visibility Approach and Data Cleaning

---

- Normalization of data format and/or representation
- Resolution of acronyms and abbreviations
- Elimination of duplicated records
- Reconciliation of contradicting records
- Control of external references
- Extraction of embedded values through data parsing

# Data Visibility Approach and Data Cleaning

---

- Normalization of data format and/or representation
- Resolution of acronyms and abbreviations
- Elimination of duplicated records
- Reconciliation of contradicting records
- Control of external references
- Extraction of embedded values through data parsing

# Data Cleaning

---

- Normalization of data format
  - ▶ Content accuracy
  - ▶ Format accuracy
  - ▶  $\Delta_{Acc}(t_i[a_k]) = [(\lambda \alpha) / n] Acc(t_i[a_k])$
- Resolution of acronyms and abbreviations:
  - ▶ Considered as format mismatch and evaluated as in the previous case

# Data Cleaning

---

- Deletion of duplicated and contradicting records:
  - ▶ both values are *Null*
  - ▶ one of the two values is *Null*
  - ▶ both values are not *Null* and are *similar*
  - ▶ both values are not *Null*, are not *similar* and the difference of their accuracy values is greater than a specified  $\Delta_{\min}$
  - ▶ both values are not null, the values are not similar and the difference of their accuracy values is lower than a specified  $\Delta_{\min}$

# Data Cleaning

---

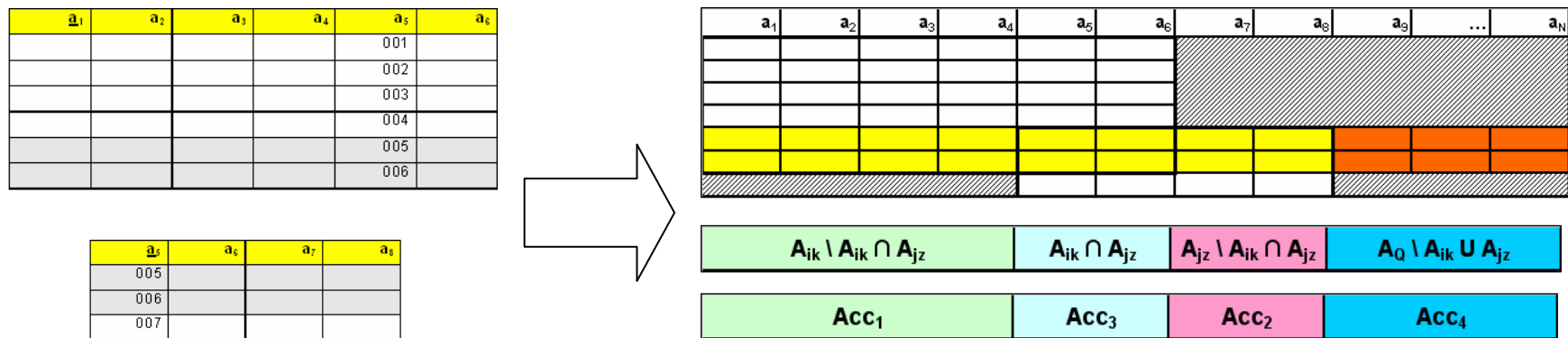
- Deletion of duplicated and contradicting records:
  - ▶ both values are *Null*: *Null* value is inserted in the final solution and the accuracy improvement is equal to 0
  - ▶ one of the two values is *Null*: the not *Null* value is considered for the final solution
  - ▶ both values are not *Null* and are *similar*: the value associated with the greater value of accuracy is considered
  - ▶ both values are not *Null*, are not *similar* and the difference of their accuracy values is greater than a specified  $\Delta_{\min}$ : the data value associated with highest accuracy is chosen
  - ▶ both values are not null, the values are not similar and the difference of their accuracy values is lower than a specified  $\Delta_{\min}$ : Business rules related to the specific attribute are applied

# Data Cleaning and Merge

---

- The broker selects the data sets that build the most complete and accurate answer
- Each source has a unique ID and two tuples refer to the same object if they have the same ID
- Resolution function
- Merge operators:
  - ▶ Join-Merge Operator
  - ▶ Left (Right) Outerjoin-Merge-Operator
  - ▶ Full Outerjoin-Merge Operator
- We have defined these operators for the accuracy dimension

# Join-Merge Operator an example



- For every partition the final accuracy is evaluated by considering broker operations, e.g.:

$$Acc_1 = (AccDataCleaning(f_{ik}) \cdot T \cdot A_1) / (A \cdot N),$$

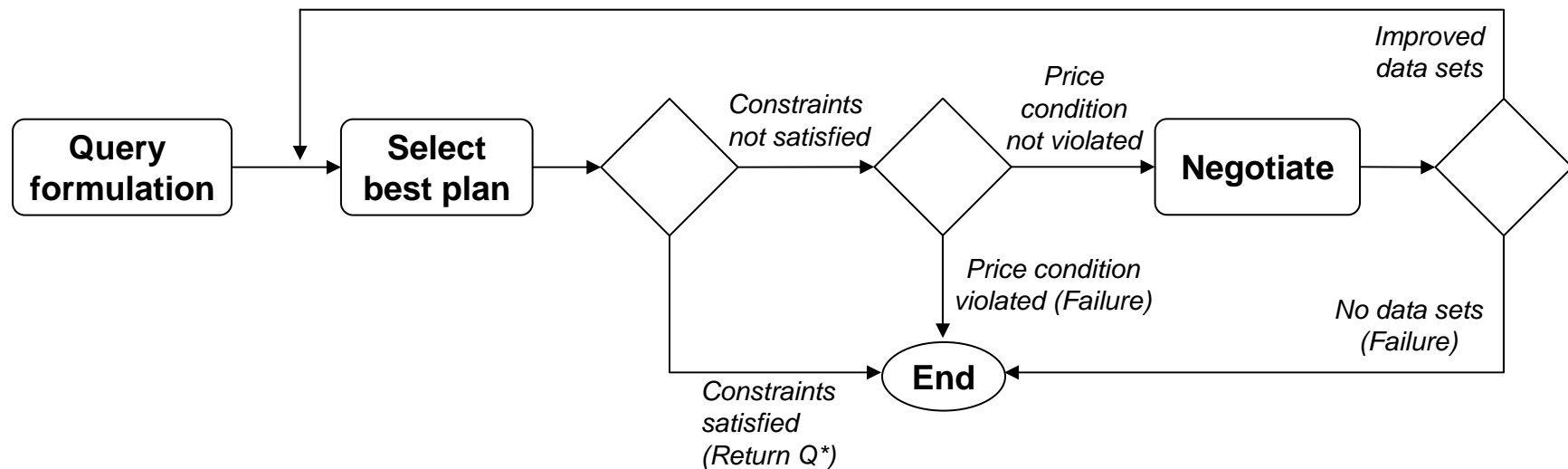
where  $AccDataCleaning(f_{ik}) = Acc_{initial}(f_{ik}) + \Delta_{AccDataCleaning}$ ,  $T$  and  $A_1$  are the number of records and attributes of the first partition while  $A$  and  $N$  are the total number of records and attributes of the UR

- The accuracy for the final result is finally evaluated as:

$$Acc_1 + Acc_2 + Acc_3 + Acc_4$$

# A data quality brokering methodology

---



# Experimental results

---

- The effectiveness of our approach has been tested on a wide set of queries
- Case study: a distributed data citizen information database adopted by the Italian Public Administration
- Data fields accuracy values have been randomly generated according to a Gaussian distribution
- Analyses have been performed by varying the data field *Null* probability, the mean and variance of the Gaussian probability density function
- Queries with up to 16 data sets and 23 attributes have been considered
- The comparison of DVA and DTA has been performed by considering the same number of iterations
- On average, the DVA implies a 30% execution time overhead

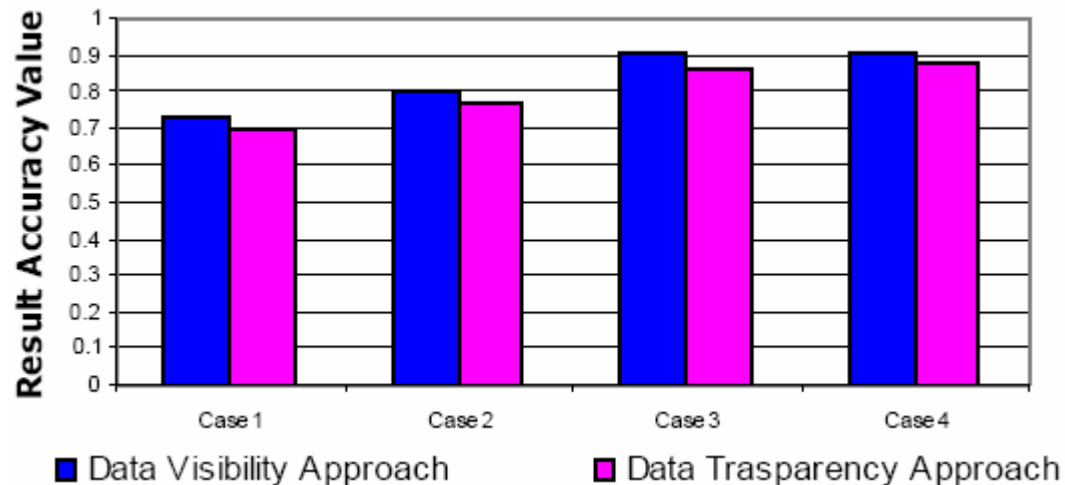
# Data Accuracy Heterogeneity vs. Data Cleaning Analysis

---

- Estimate how the improvement of accuracy changes with the heterogeneity of the local data sets
- Analyses consider four test cases:
  - ▶ Case 1: 16 data sets have an expected value of accuracy equal to 0.7;
  - ▶ Case 2: 8 data sets have an expected value of accuracy equal to 0.7, and 8 data sets have an expected value of accuracy equal to 0.8;
  - ▶ Case 3: 4 data sets have an expected value of accuracy equal to 0.7, 6 data sets have an expected value of accuracy equal to 0.8, and 6 data sets have an expected value of accuracy equal to 0.9;
  - ▶ Case 4: 4 data sets have an expected value of accuracy equal to 0.7, 4 data sets have an expected value of accuracy equal to 0.8, 4 data sets have an expected value of accuracy equal to 0.9 and 4 data sets have an expected value of accuracy equal to 0.99

# Data Accuracy Heterogeneity vs. Data Cleaning Analysis

---



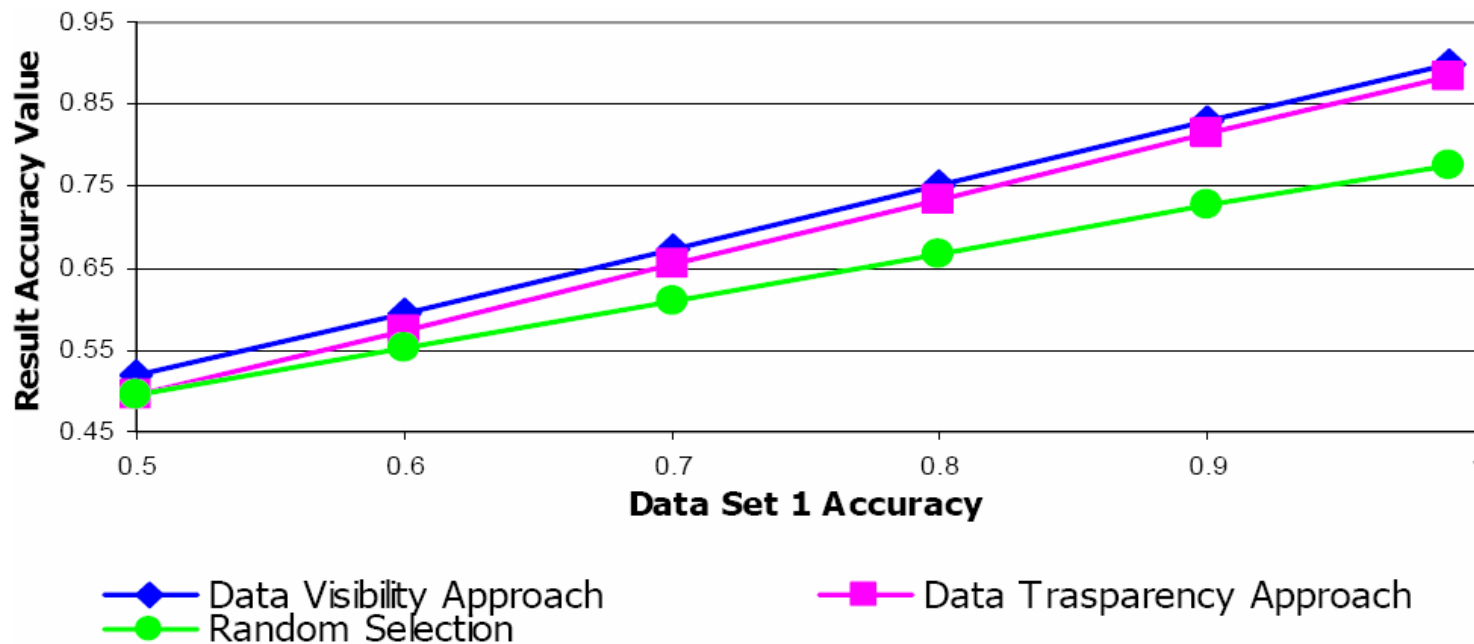
- as the accuracy of data sets increases, the improvement from data cleaning operations become less effective since there is a lower number of errors to correct
- as the number of data sets with a high value of accuracy increases, the merging procedure becomes less efficient since overlapping data often coincide with the data characterized by the highest average accuracy

# Data Accuracy Heterogeneity and Merging

---

- Determine how the quality heterogeneity of the local data sets influences the effectiveness of merging
- The greater the difference of accuracy for overlapping data, the greater the improvement obtained from merging
- Two data sets are considered: initially they have the same value of accuracy (0.5); then, the accuracy of the first data set is increased by 0.1 at each step
- The query includes the overlapping data fields that are not involved in data cleaning operations
- We consider also the random selection of tuples from the two data sets

# Data Accuracy Heterogeneity and Merging



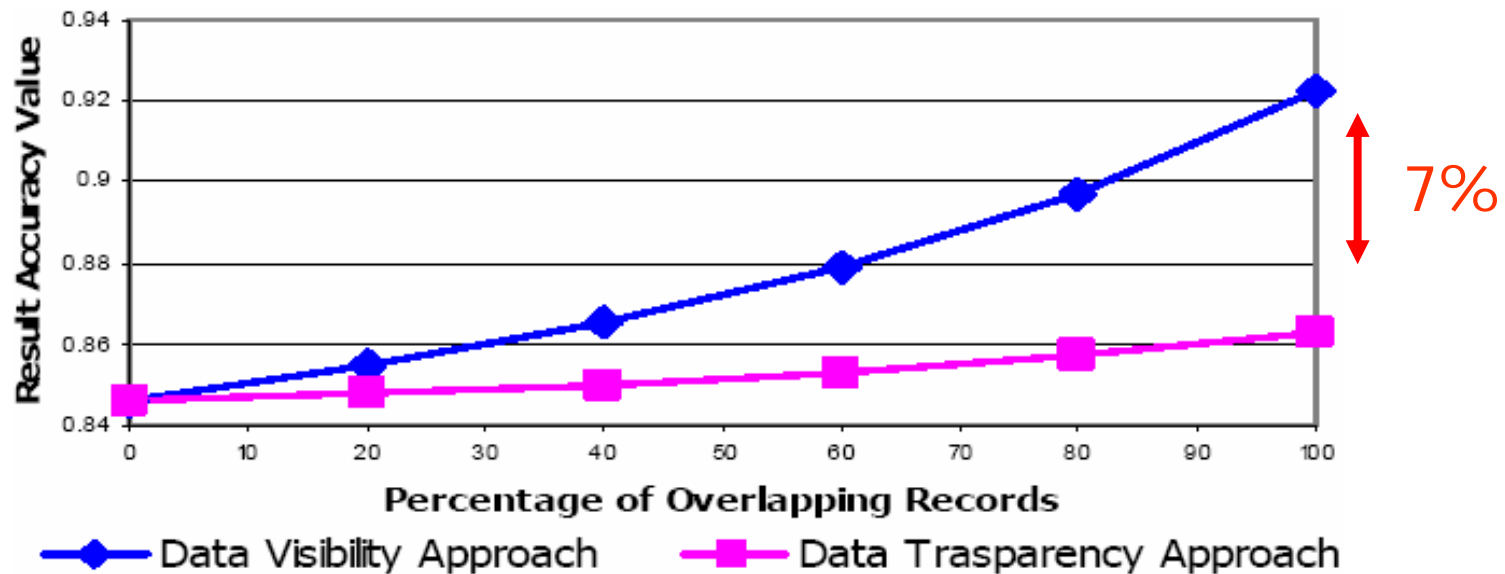
- The improvement with respect to the random selection increases, since if one of the two data sets becomes more accurate, the selection of the best data performed by the merging algorithm becomes more effective
- The improvement with respect to the DTA decreases, since if one of the two fragments becomes more accurate, the DVA and DTA select tuples from the same data set

# Fraction of Records and Merging

---

- Evaluate the dependency of the accuracy on the fraction of overlapping records
- Two data sets are considered. The query includes the overlapping data fields that are not involved in data cleaning operations
- The fraction of overlapping records has been varied in the range [0, 100%] with step 20%

# Fraction of Records and Merging



- As the percentage of overlapping records increases, the accuracy of the final query plan grows for both the DTA and DVA
- The improvement is more significant for the DVA since the merging algorithm always selects the best of the overlapping data

# Conclusions and Future work

---

- Quantitative comparison between DVA and DTA to data brokering
- Results show that the accuracy improvement obtained by using the DVA justifies the additional computational complexity
- Best results when data sets have low accuracy and data cleaning techniques are adopted
- Extend the optimization algorithm and implement column generation techniques in order to identify the global optimum
- Negotiation phase will be also considered with the introduction of the evaluation of price constraints in the optimization problem
- Consider other data quality dimensions

---

Thanks! Any questions?