Mediation and Data Source Selection for Large Scale Virtual Organizations

Alexandra Pomares Quimbaya *

a.pomares267@uniandes.edu.co

Jury

Claudia Roncancio, Directeur de thèse
José Abásolo, Directeur de thèse
Rubby Casallas, Président
Marta Rukoz, Rapporteur
Sandra de Amo, Rapporteur
Marta Millán, Examinateur

Universidad de los Andes
COMIT Group
Bogotá, Colombia

Université de Grenoble
LIG Laboratory,
SIGMA Group
Grenoble, France

* Auspiciada por Colciencias y Pontificia Universidad Javeriana

Outline

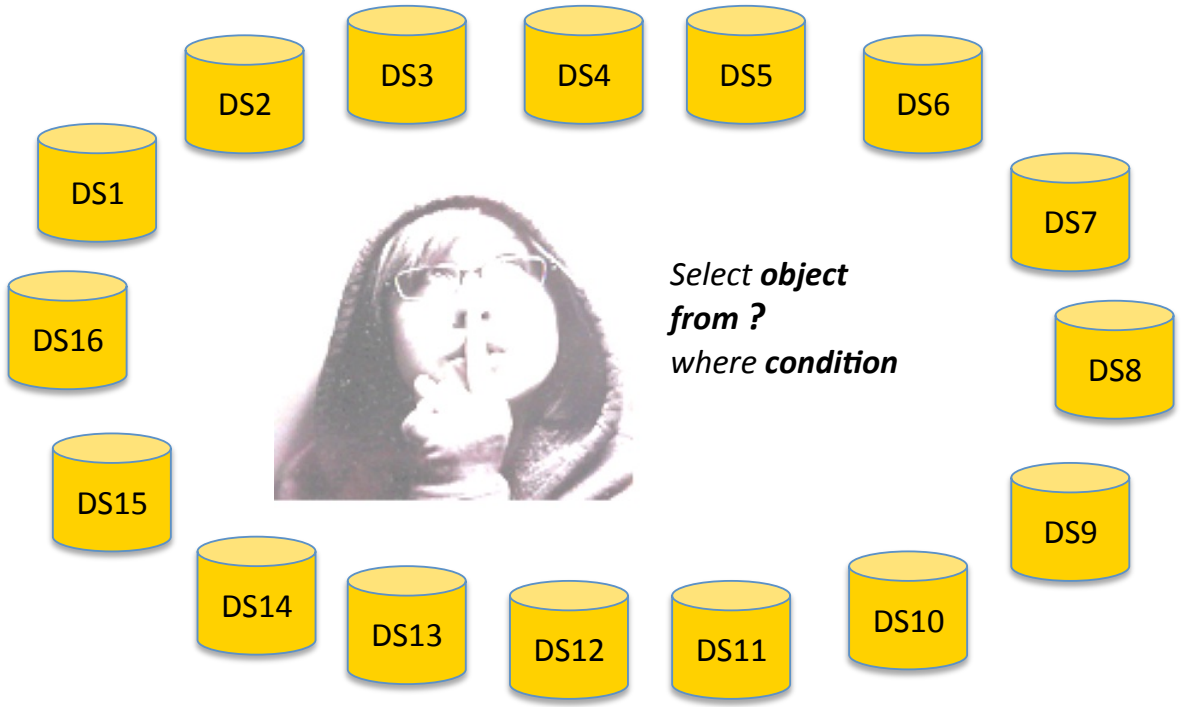
1. **Context and Motivation**: Data integration in virtual organizations
   - ❑ Data integration problem
   - ❑ Data integration solutions applied to virtual organizations
2. **Proposal: A Mediation System for Virtual Organizations**
   - ❑ OptiSource a source selection strategy
   - ❑ Organizational knowledge
   - ❑ Individual Contribution
   - ❑ Group Contribution
   - ❑ Evaluation process
3. **Validation of OptiSource**
   - ❑ Tests of precision and recall
   - ❑ Sensibility analysis
   - ❑ Comparison with IDrips
4. **Conclusions and Future Works**

Outline

1. **Context and Motivation**: Data integration in virtual organizations
   - ❑ Data integration problem
   - ❑ Data integration solutions applied to virtual organizations
2. **Proposal: A Mediation System for Virtual Organizations**
   - ❑ OptiSource a source selection strategy
   - ❑ Organizational knowledge
   - ❑ Individual Contribution
   - ❑ Group Contribution
   - ❑ Evaluation process
3. **Validation of OptiSource**
   - ❑ Tests of precision and recall
   - ❑ Sensibility analysis
   - ❑ Comparison with IDrips
4. **Conclusions and Future Works**

Data Integration Problem

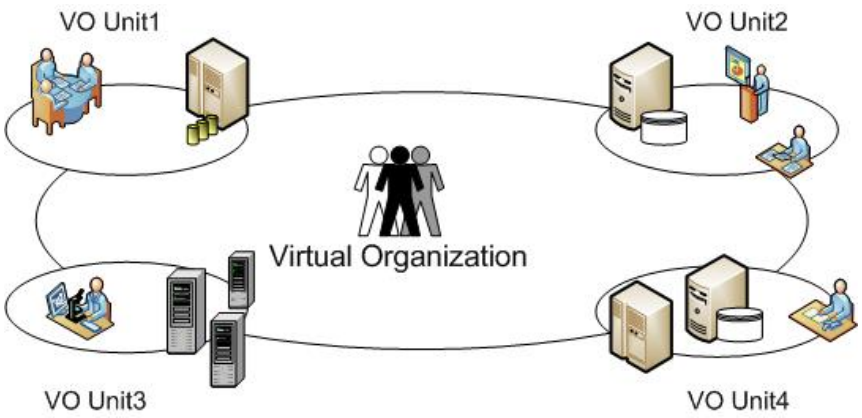
Data integration is the process of combining data residing at different data sources, and providing the user with a unified view of these data [1].



Heterogeneity ***Distribution*** ***Incompleteness*** ***Duplication***

Virtual Organizations

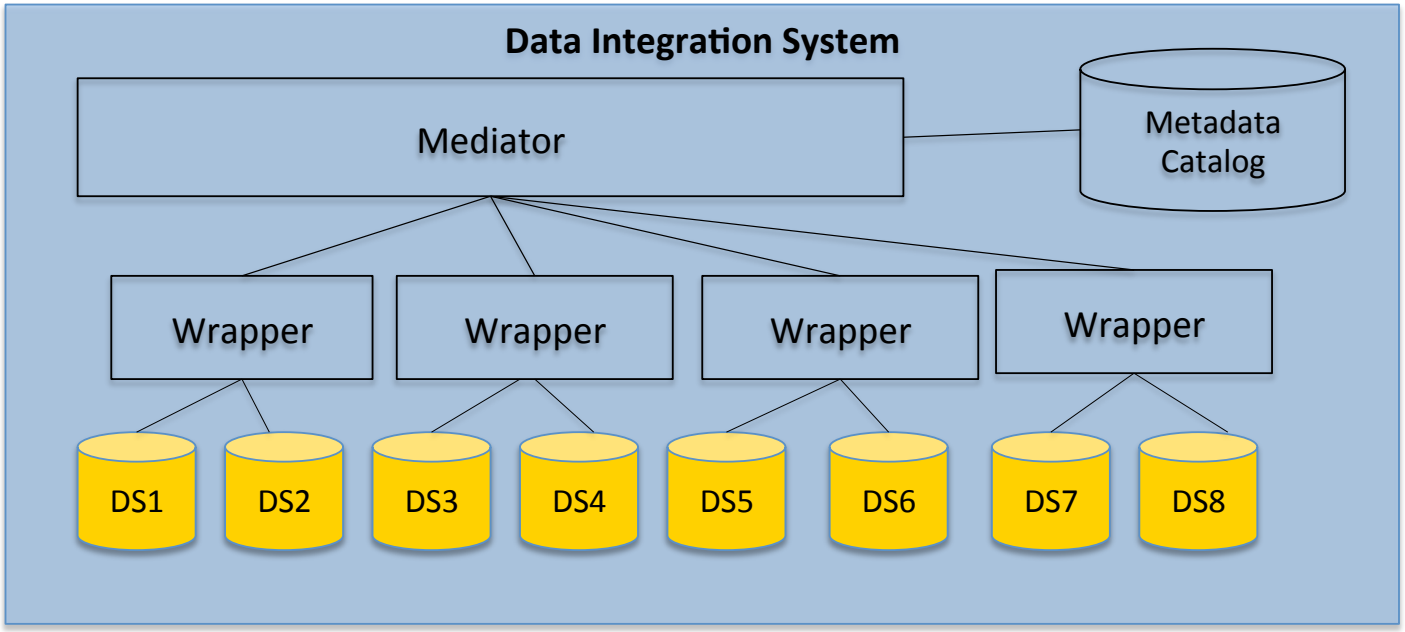
Set of autonomous collaborating organizations, called VO Units, working toward a common goal.



- Participants share resources and complementary competencies
- Participants are geographically dispersed
- Constructed over an alliance (with related sub-alliances)
- Based on information technology
- Based on mutual trust and shared processes
- Able to respond to environmental changes

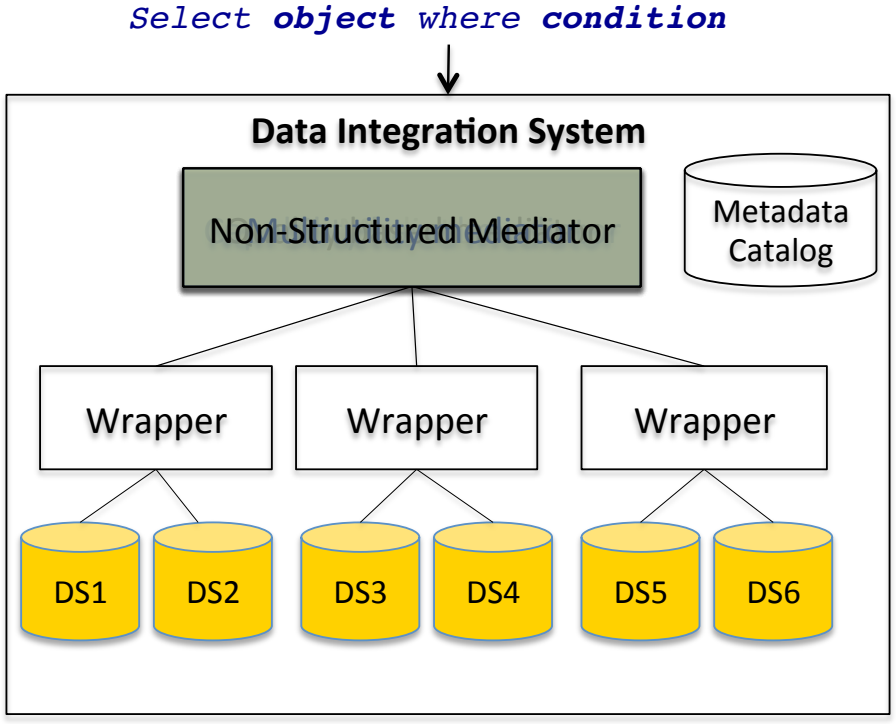
Data Integration Systems using a Mediator Architecture

Select ***object*** *where* ***condition***



Types of Mediators

According to their strategy of **source selection**.



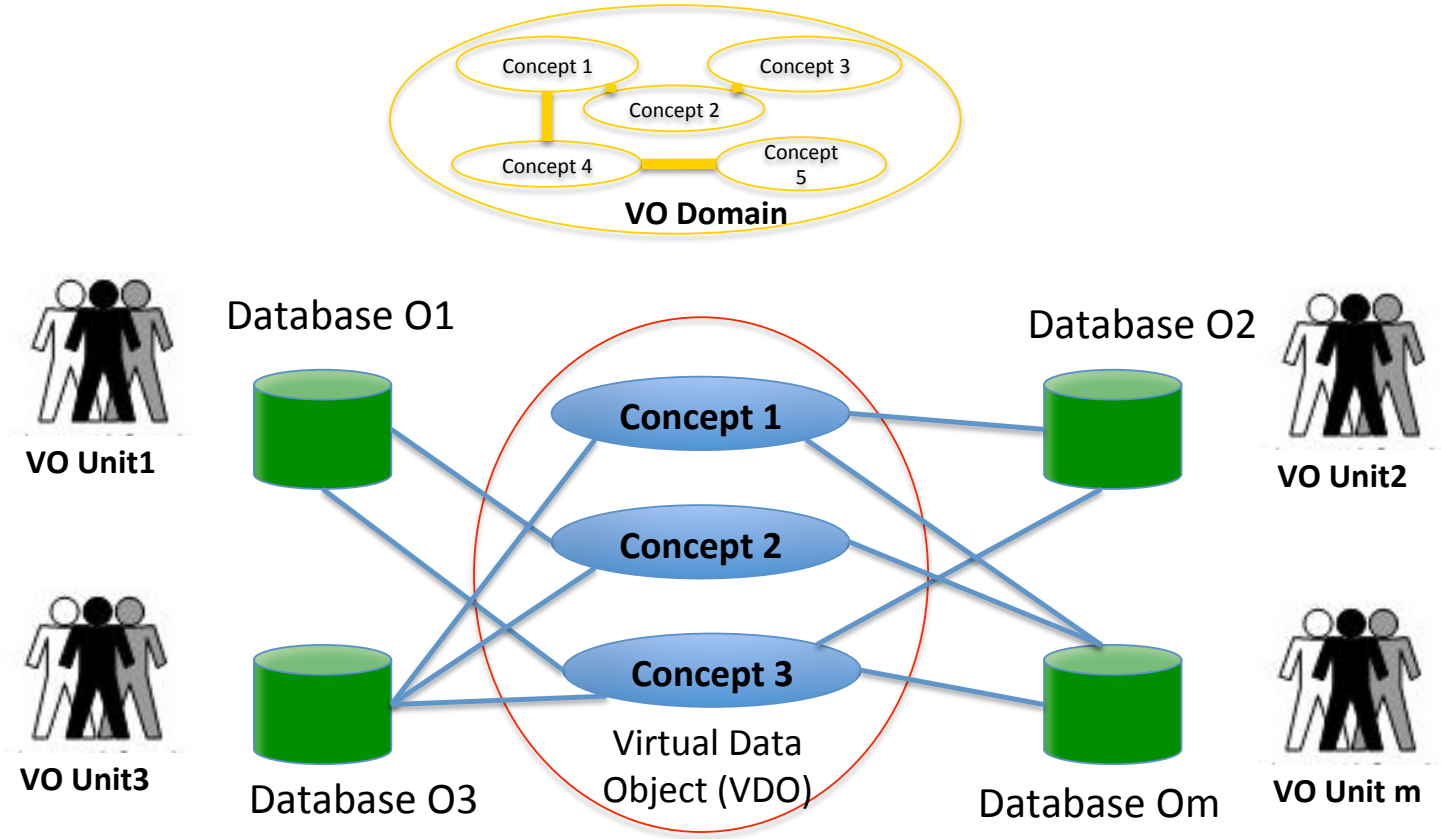
Capability Based
Ability to evaluate the structure of the query. e.g. Info. Manifold[2], TSIMMIS [3], PIER[4]

Quality Based
Internal and external quality.
e.g. QPIAD[5], Navigational Paths[6]

Multiutility Mediator
Prioritize plans using a generic utility function. e.g.IDrips[7], Streamer[8]

Non Structured Data Sources Mediator
Using a summary of their contents or sending probe queries to determine their utility. e.g. CORI[9], Qprober[10]

Virtual Organization Data Contexts



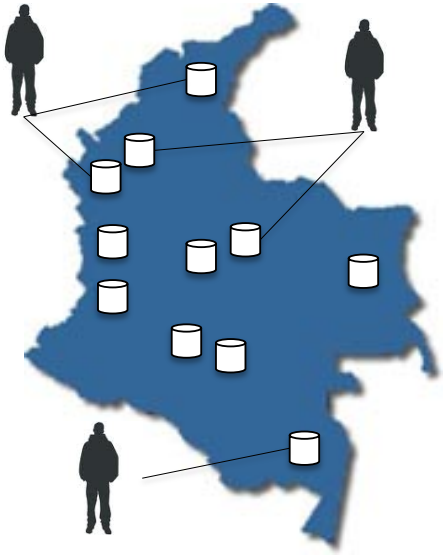
Large Scale Virtual Organizations Data Contexts

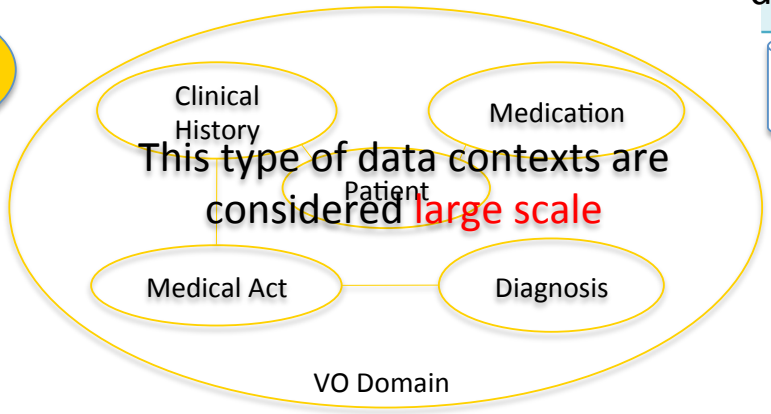
Intentional Overlapping

Extensional Overlapping

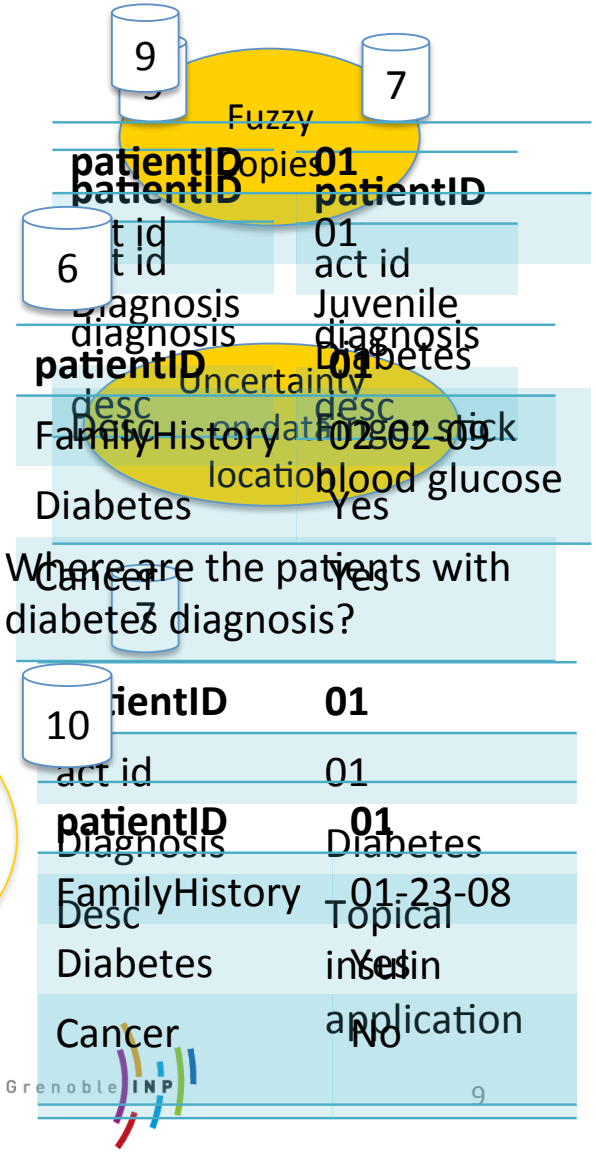
Large Number of Data Sources

300 - 1000





This type of data contexts are considered large scale



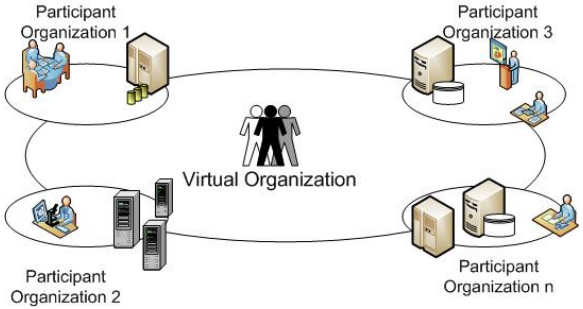
Current Strategies applied to Virtual Organizations

Mediator based on Capabilities

- Not useful to discard data sources when all of them have similar capabilities
- Designed for a hundred or less data sources

Multiutility Mediator

- Efficiency related to the used utility function
-Difficult to reflect extensional overlap



Mediator based on Quality

-Statistics not always available
-It does not obtain the better data sources if it is used alone when there are fuzzy copies and replication

Mediator for Non Structured Data Sources

- Limitations on the type of queries
- Difficulty to obtain the summary of data sources

Source Selection Strategies

The analysis of capabilities and quality of data sources are required during source selection, but...

- Intentional Overlapping
- Extensional Overlapping
- Fuzzy Copies
- Large Number of Data Sources
- Uncertainty on Data Location

need to be analized in order to scale up mediation systems and be efficient during the source selection process.

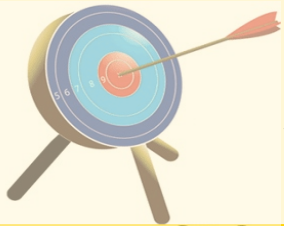
OptiSource

- Is a source selection strategy
- Aims to improve precision
- Uses organizational knowledge to describe the data context
- Created for large scale contexts

Outline

1. **Context and Motivation**: Data integration in virtual organizations
   - ❑ Data integration problem
   - ❑ Data integration solutions applied to virtual organizations
2. **Proposal: A Mediation System for Virtual Organizations**
   - ❑ OptiSource a source selection strategy
   - ❑ Organizational knowledge
   - ❑ Individual Contribution
   - ❑ Group Contribution
   - ❑ Evaluation process
3. **Validation of OptiSource**
   - ❑ Tests of precision and recall
   - ❑ Sensibility analysis
   - ❑ Comparison with IDrips
4. **Conclusions and Future Works**

Contribution of this thesis

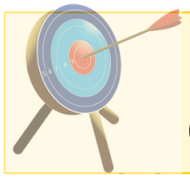


OptiSource
A strategy of data source selection for large scale virtual organizations

Objectives

- To increase scalability improving the precision
- To improve the selection of the more relevant data sources
- To reduce the level of redundancy in the final response

Contribution of this thesis: Approach

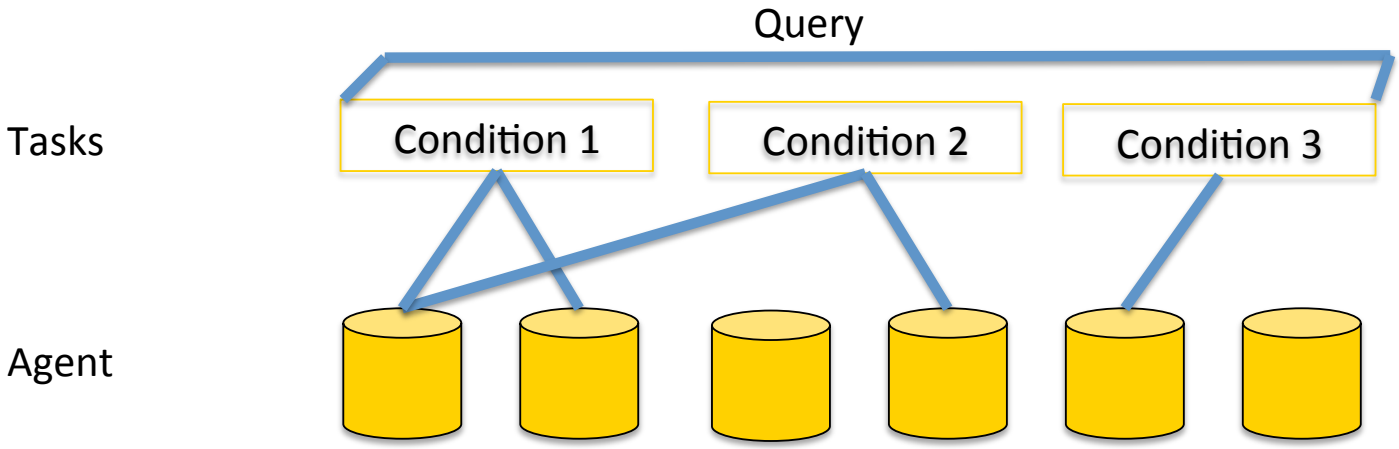


Approach of OptiSource

Optimize the assignment of conditions to data sources

- Obtains the better assignment of conditions to data sources using a combinatorial optimization model
- Estimates the benefit of a data source using organizational knowledge to identify the role it may play in a query
- Groups together data sources that tend to share instances using heuristic rules of the VO

OptiSource: Source Selection as an Assignment Problem



Objective Function

Maximize the benefit in term of instances

Minimize the number of queried data sources

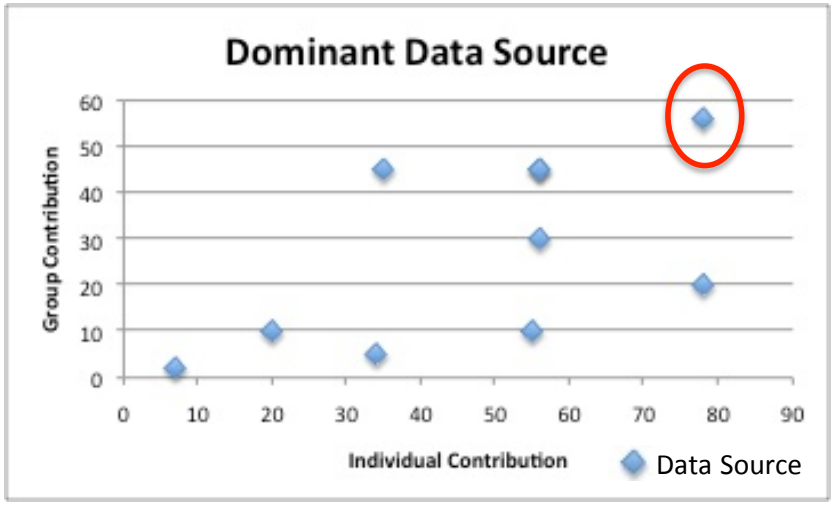
Query Assignment

A query condition must be assigned to the dominant data source for this condition.

Data Source Restrictions

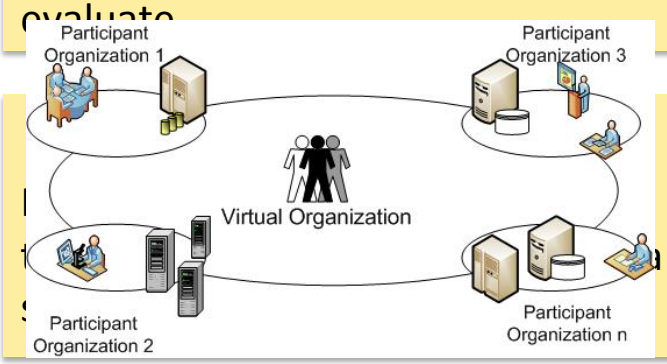
Assignments must respect data source restrictions like number of conditions it may evaluate.

OptiSource: Dominant Data Sources



Using organizational knowledge of the VO as metadata of the data context.

Individual contribution
Measures the ability of a data source to provide instances that match the query conditions it can evaluate



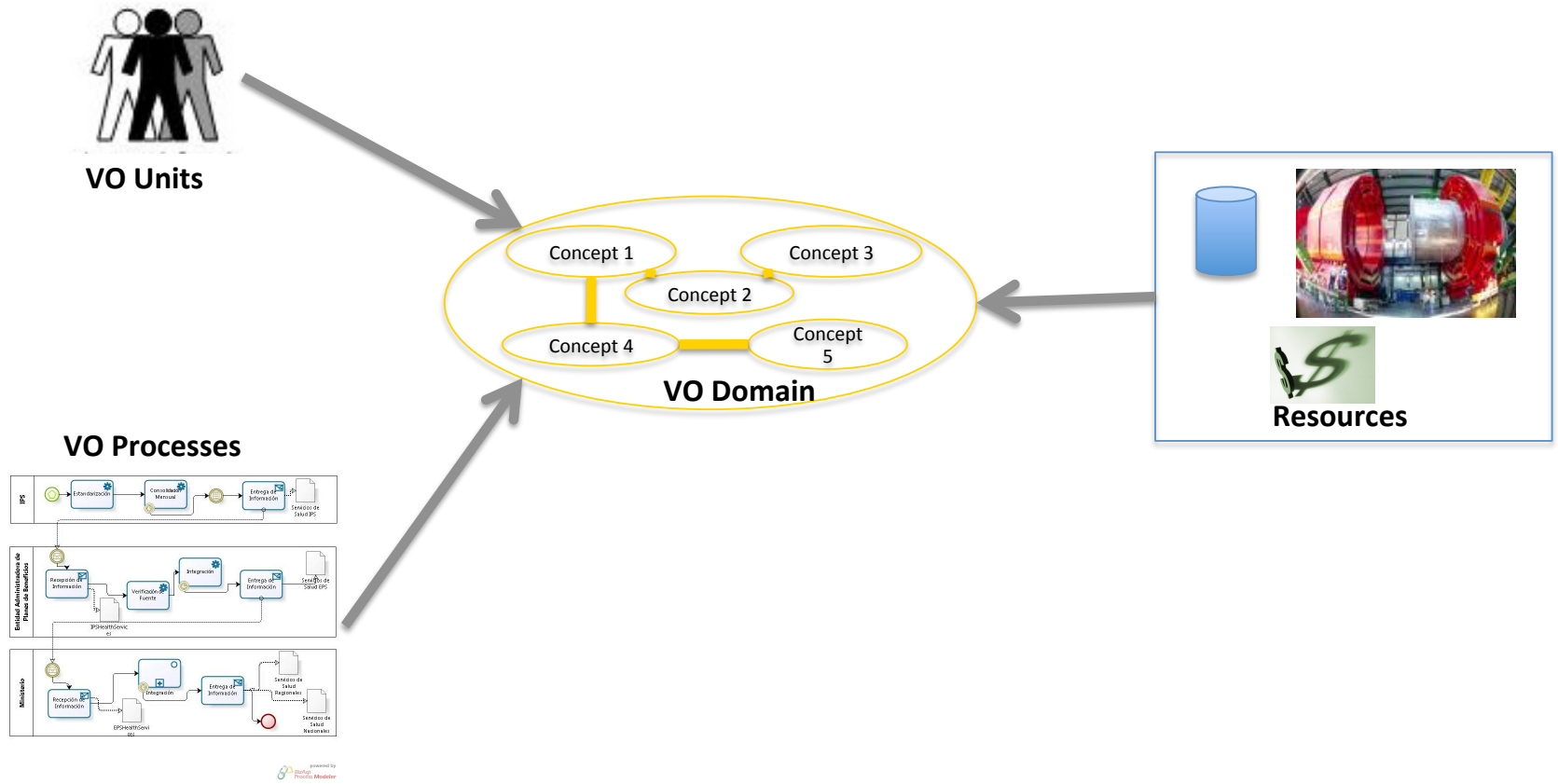
How to calculate the individual contribution of a data source?

How to determine the group contribution?

Outline

1. **Context and Motivation**: Data integration in virtual organizations
   - ❑ Data integration problem
   - ❑ Data integration solutions applied to virtual organizations
2. **Proposal: A Mediation System for Virtual Organizations**
   - ❑ OptiSource a source selection strategy
   - ❑ Organizational knowledge
   - ❑ Individual Contribution
   - ❑ Group Contribution
   - ❑ Evaluation process
3. **Validation of OptiSource**
   - ❑ Tests of precision and recall
   - ❑ Sensibility analysis
   - ❑ Comparison with IDrips
4. **Conclusions and Future Works**

Organizational Knowledge



Organizational Knowledge: VO Units

Location
Physical and logical location

Commitments
It provides resources
It participates in one or more VO business process
It Plays a role in the VO

VO Units

Specialization
It works in an area of the VO Domain
e.g.: Research on Cancer Pathology

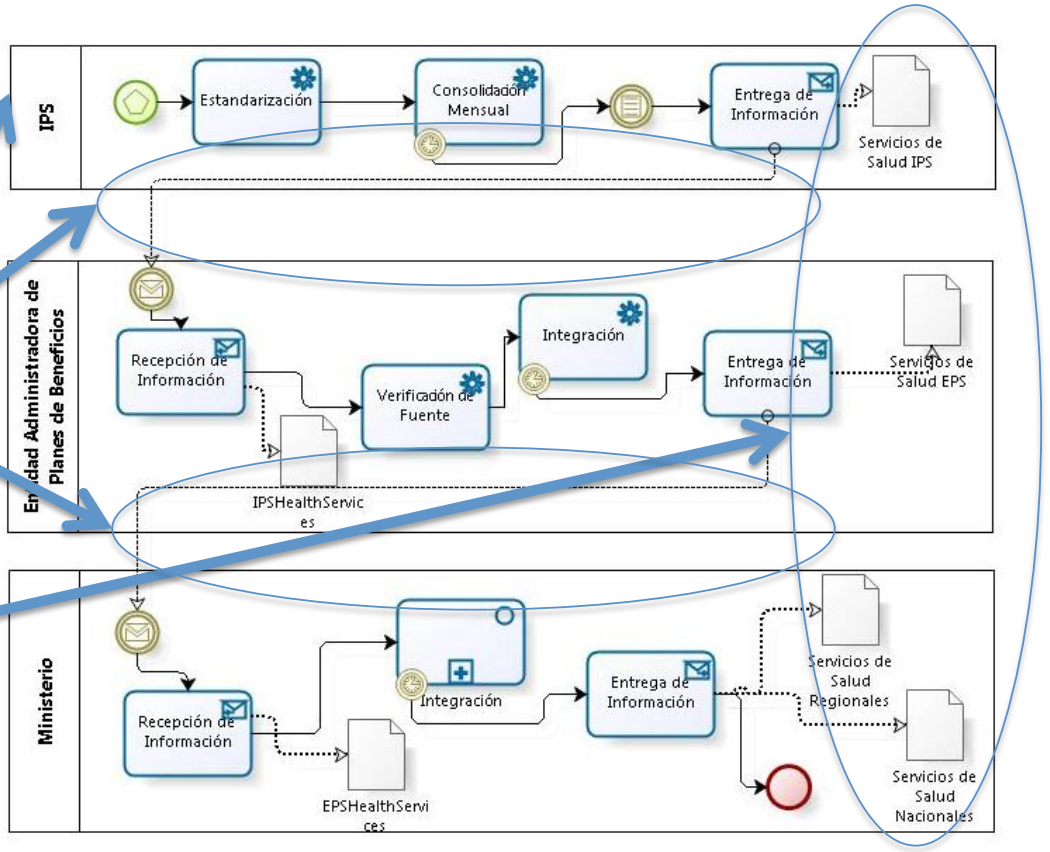
Alliances
Work together with other VOUnits

Organizational Knowledge: VO Business Processes

Data transfers between participants

Participants: VO Units

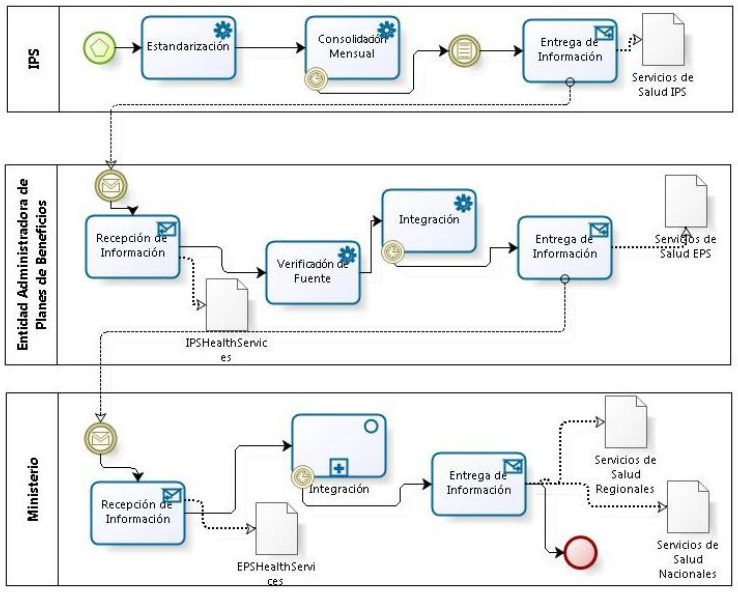
Involves VO Resources

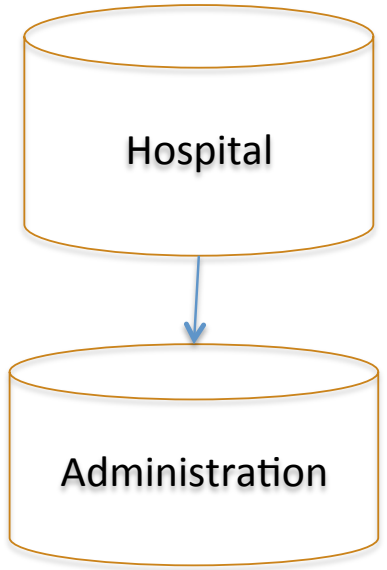


How is this knowledge used to describe the data context ?
VO Process

VO process

Data source relationships
Copies



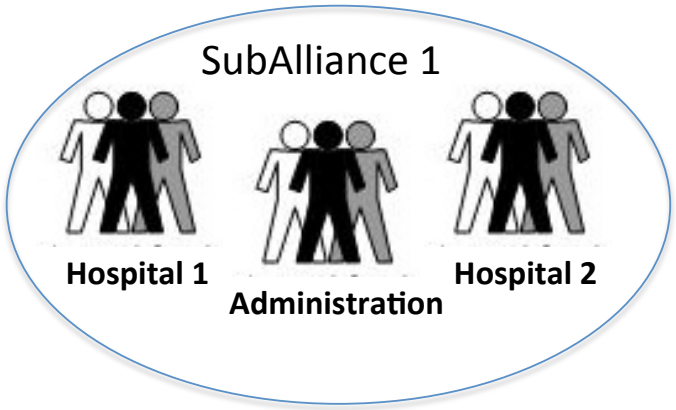


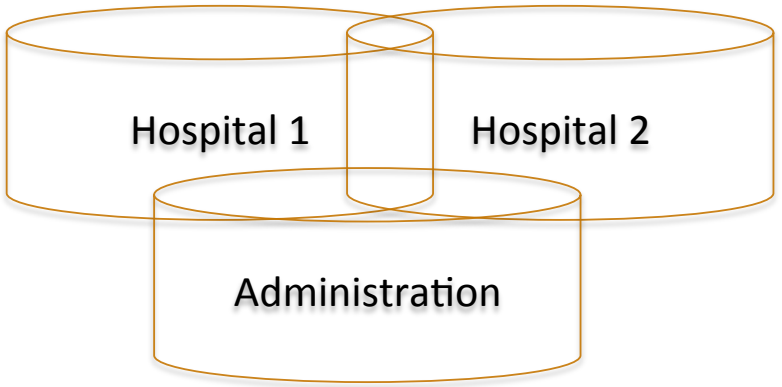
How is this knowledge used to describe the data context ?
VO Unit Characteristics

VO unit location

VO unit sub-alliances

Data source relationships
Extensional overlap





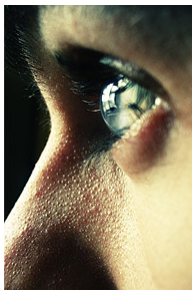
How is this knowledge used to describe the data context ?
VO Units Specializations

VO unit specialization

Data source role

Maternity Hospital

Specialist gender = F
Specialist diagnosis = pregnancy



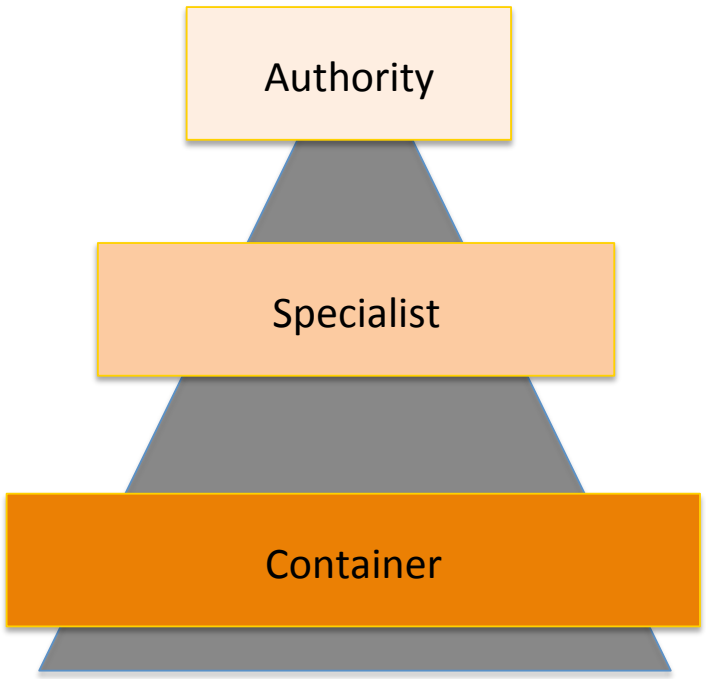
Ophthalmology Clinic

Specialist act = OperetiononTheEye
Container diagnosis = Glaucoma

Outline

1. **Context and Motivation**: Data integration in virtual organizations
   - ❑ Data integration problem
   - ❑ Data integration solutions applied to virtual organizations
2. **Proposal: A Mediation System for Virtual Organizations**
   - ❑ OptiSource a source selection strategy
   - ❑ Organizational knowledge
   - ❑ Individual Contribution
   - ❑ Group Contribution
   - ❑ Evaluation process
3. **Validation of OptiSource**
   - ❑ Tests of precision and recall
   - ❑ Sensibility analysis
   - ❑ Comparison with IDrips
4. **Conclusions and Future Works**

Individual Contribution: Roles of Data Sources



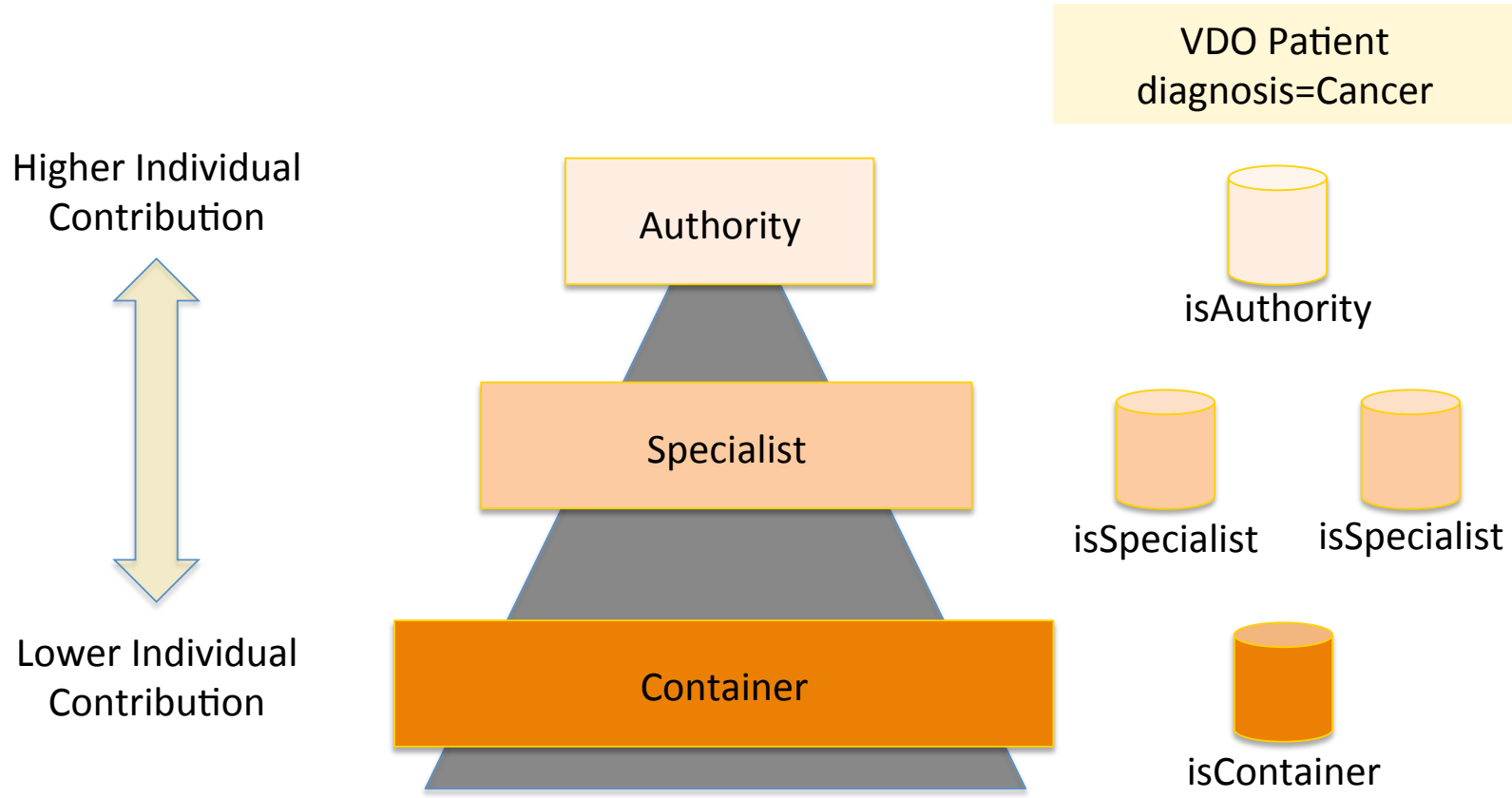
Data Source 1 is Authority of a VDO type because it contains all the instances of this type available in the VO

Data Source 3 is Specialist of a VDO type because it contains mostly instances of this type

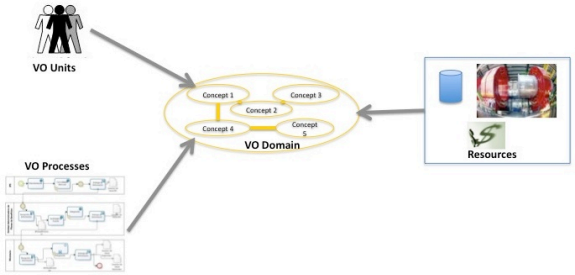
Data Source 2 is Container of a VDO type because it contains at least one instance of this type

Roles and Individual Contribution

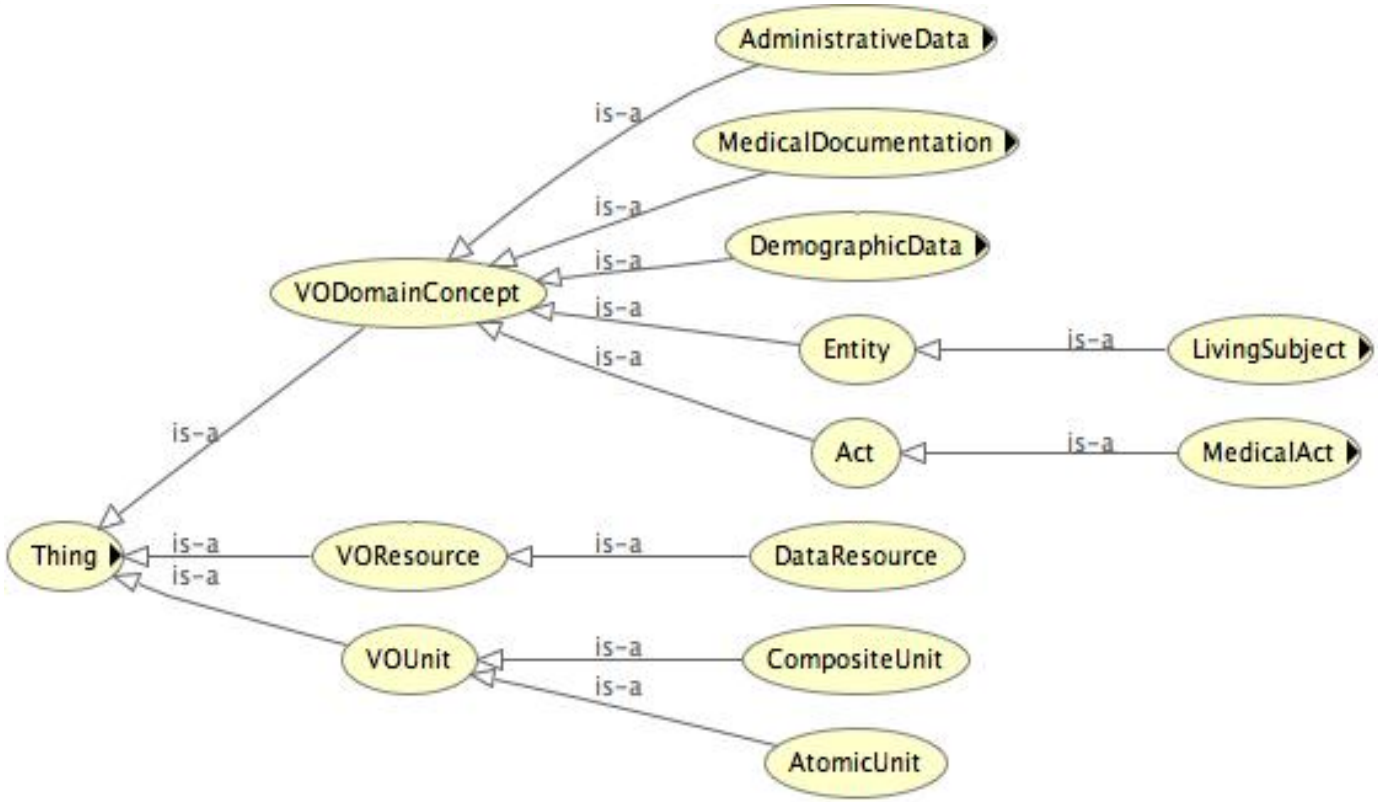
```
Select idPatient where diagnosis = Cancer
```



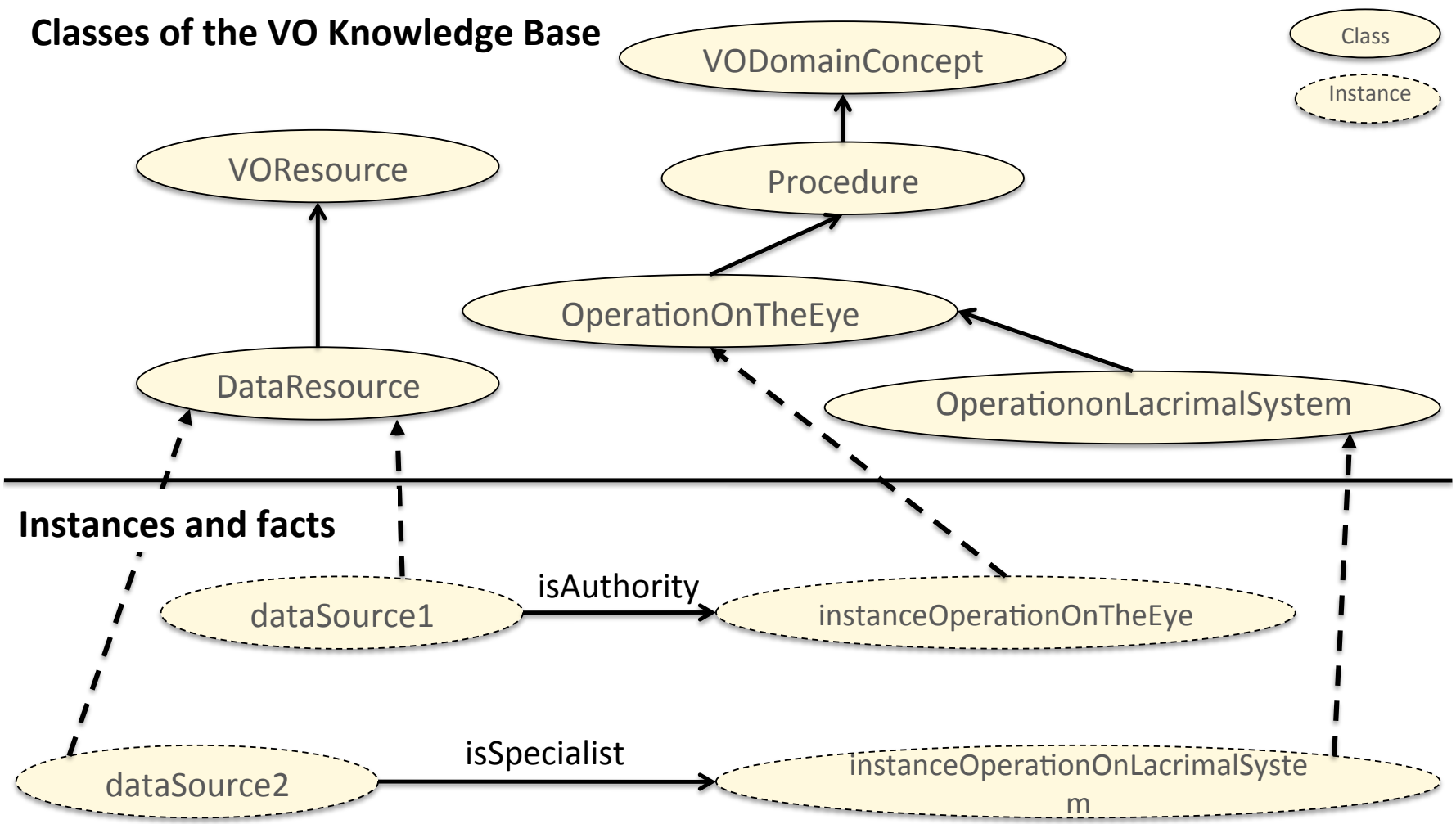
Individual Contribution in Practice



VO Knowledge Base

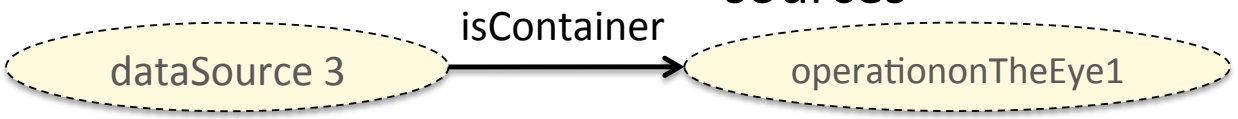


Individual Contribution in Practice (cont.)

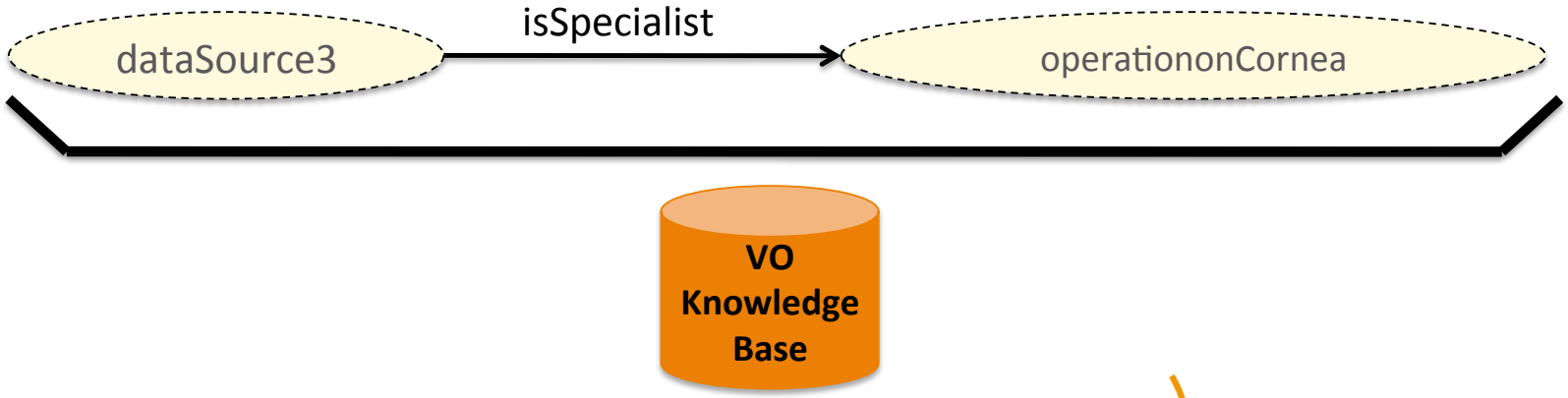


Individual Contribution in Practice (cont.)

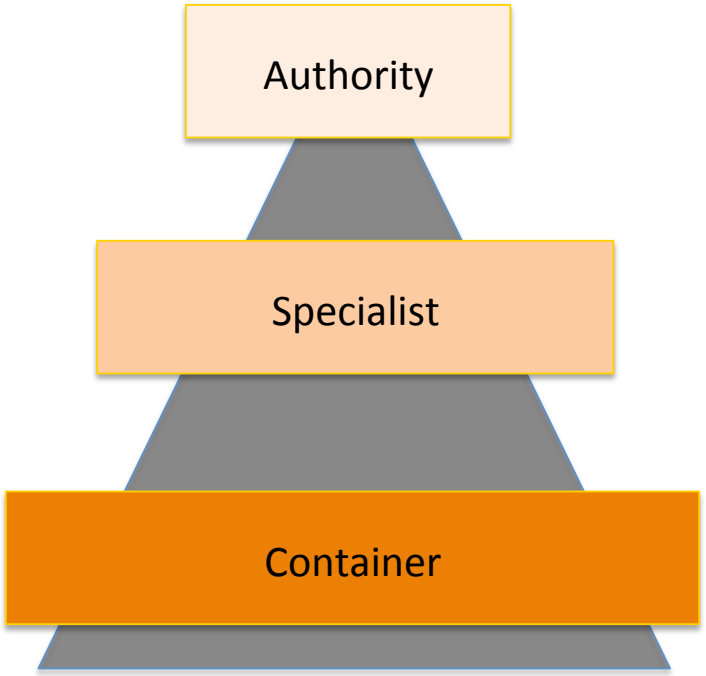
- VO expert knowledge
- Business process analysis
- Analysis of query log
- Data mining over data sources



Knowledge can be refined and enriched



Individual Contribution Measure



and

Number of Instances in the Data Source

$$IndividualCont(DS_i,VDO,c) = RoleFactor(DS_i,VDO,c) * \frac{card(ext(DS_i,VDO))}{\max(card(ext(DS_k,VDO)))\forall DS_k \in U}$$

Outline

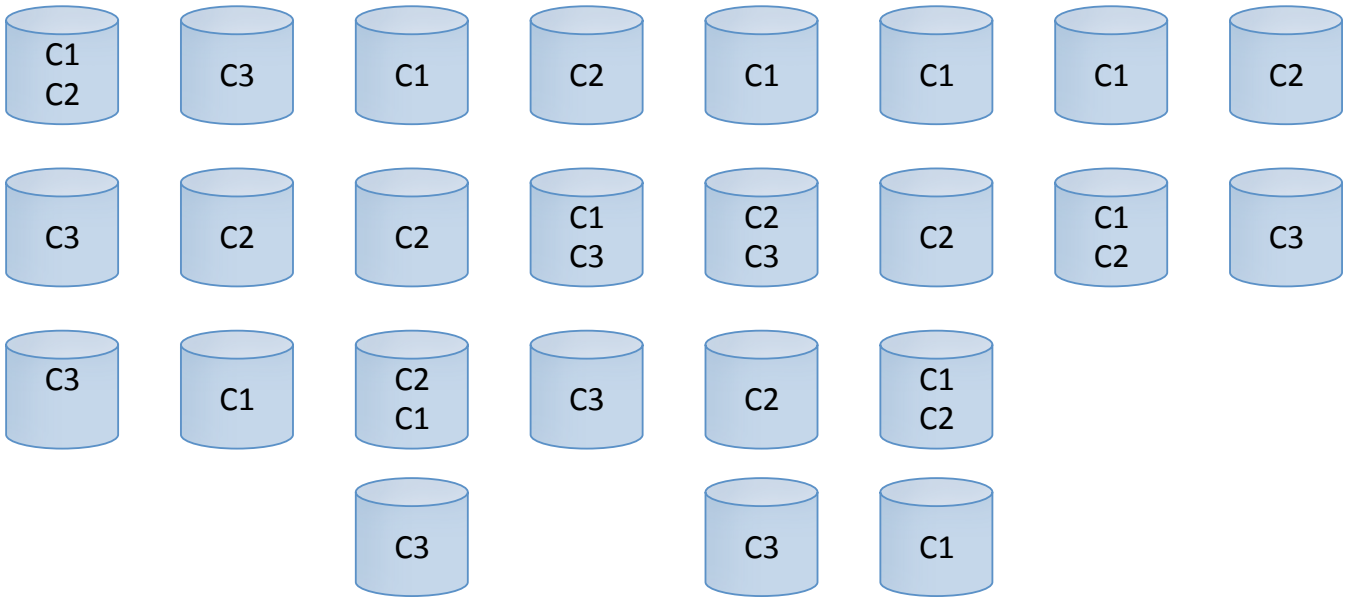
1. **Context and Motivation**: Data integration in virtual organizations
   - ❑ Data integration problem
   - ❑ Data integration solutions applied to virtual organizations
2. **Proposal: A Mediation System for Virtual Organizations**
   - ❑ OptiSource a source selection strategy
   - ❑ Organizational knowledge
   - ❑ Individual Contribution
   - ❑ Group Contribution
   - ❑ Evaluation process
3. **Validation of OptiSource**
   - ❑ Tests of precision and recall
   - ❑ Sensibility analysis
   - ❑ Comparison with IDrips
4. **Conclusions and Future Works**

Group Contribution

1. The answer may be incomplete.
2. Data sources may not share instances.

Query (condition1, condition 2, condition 3)

Integration Group
Set of data sources that share instances



Identifying Integration Groups

Heuristic rules to identify integration groups

Data sources provided by VO Units located in the same physical location.
e.g.: Bogotá

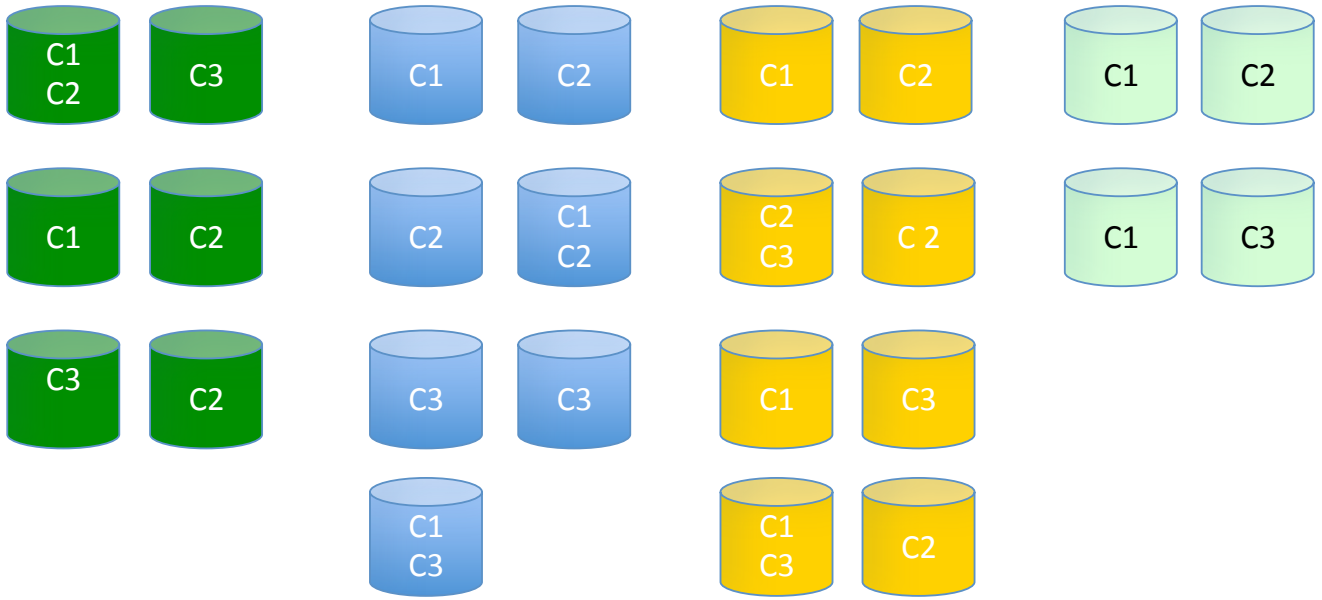
Data Sources provided by VO Units that belong to the same sub-alliance.
e.g.: Cancer research alliance

Data Sources provided by VO Units of the same type.
e.g.: Administration

Building Integration Groups

Can this assignment maximize the group contribution?

Query (condition1, condition 2, condition 3)

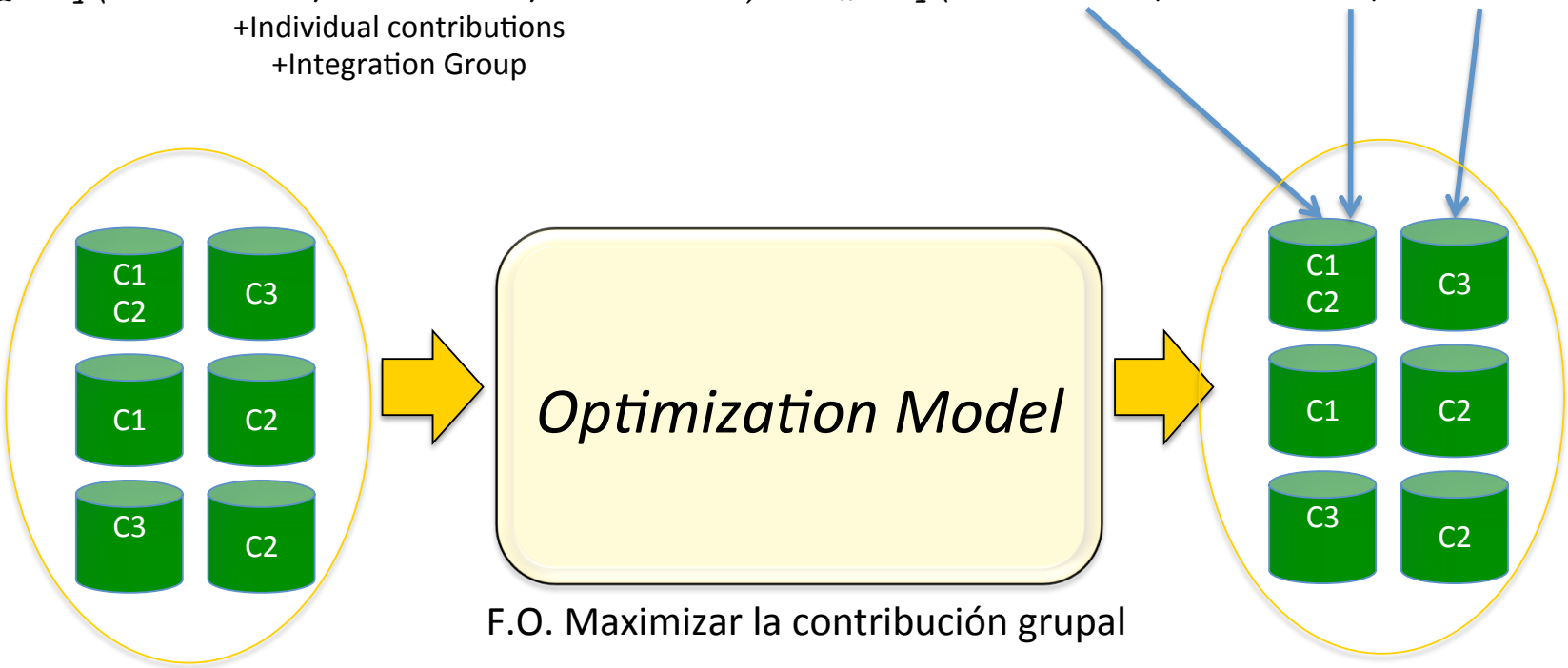


Query Answer = Result set 1 | Result set 2 | Result set 3 | Result set 4

Group Contribution in Practice

Input:
Query(condition1,condition2,condition3)
+Individual contributions
+Integration Group

Output:
Query(condition1,condition2,condition3)



F.O. Maximizar la contribución grupal

Optimization Model of OptiSource

$$\max \sum_{c=1}^{n} \sum_{i=1}^{m} Ben_{i,c} * (x_{i,c} + assig_{i,c}),$$

$$\sum_{i=1}^{m} x_{i,c} = 1, \forall c \in C$$

$$\sum_{i=1}^{m} y_i \leq k$$

$$\sum_{c=1}^{n} x_{i,c} \geq y_i \forall i \in I$$

$$\sum_{c=1}^{n} \mathrm{Re}\, s_{i,c} * assig_{i,c} \leq Max\,\mathrm{Re}\, s_i, \forall i \in I$$

$$\sum_{c=1}^{n} assig_{i,c} \leq MaxAssig_i, \forall i \in I$$

$$x_{i,c} \leq assig_{i,c}, \forall i \in I, \forall c \in C$$

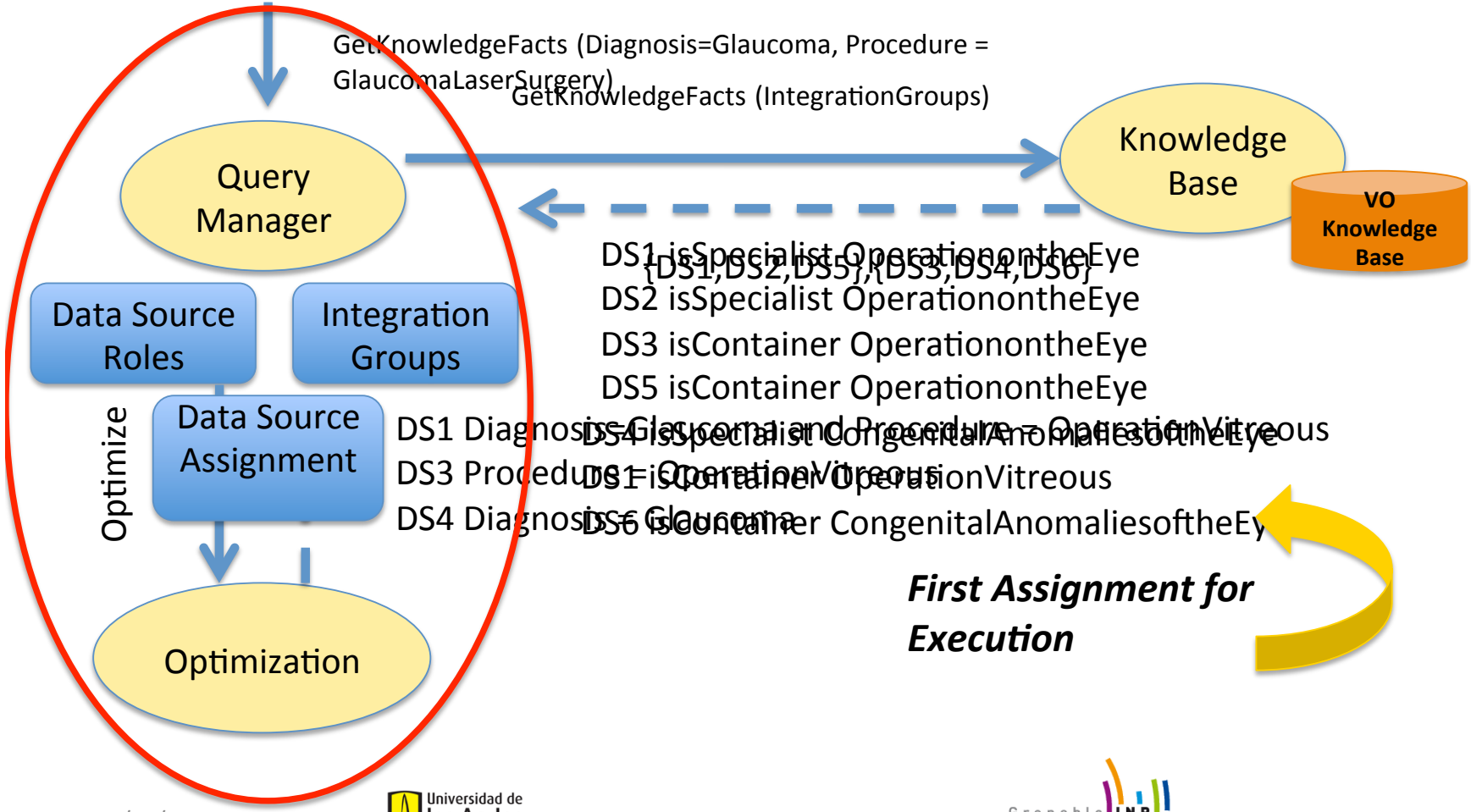
$$assig_{i,c} \leq y_i, \forall i \in I, \forall c \in C$$

Outline

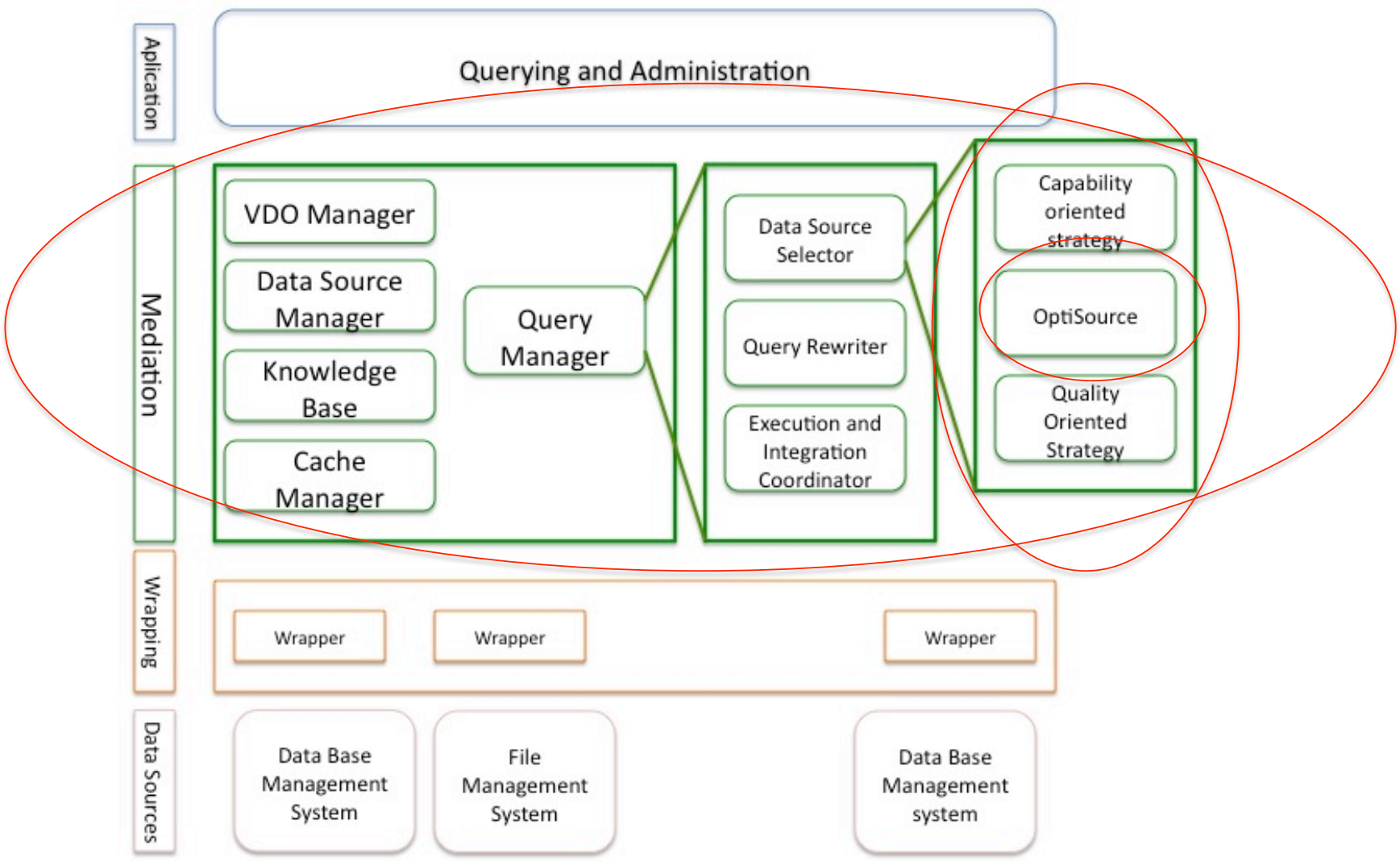
1. **Context and Motivation**: Data integration in virtual organizations
   - ❑ Data integration problem
   - ❑ Data integration solutions applied to virtual organizations
2. **Proposal: A Mediation System for Virtual Organizations**
   - ❑ OptiSource a source selection strategy
   - ❑ Organizational knowledge
   - ❑ Individual Contribution
   - ❑ Group Contribution
   - ❑ Evaluation process
3. **Validation of OptiSource**
   - ❑ Tests of precision and recall
   - ❑ Sensibility analysis
   - ❑ Comparison with IDrips
4. **Conclusions and Future Works**

OptiSource Process

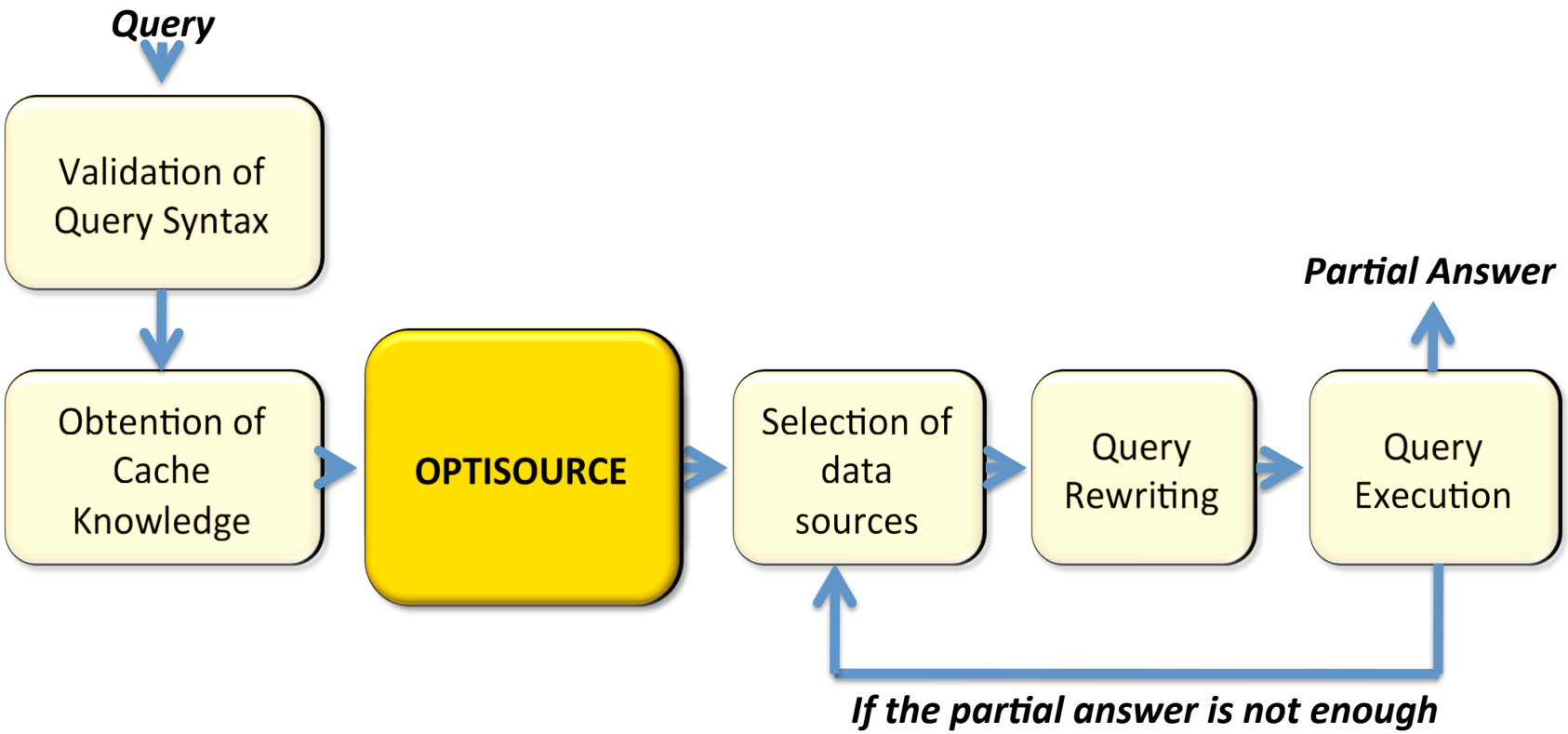
Diagnosis = Glaucoma
Procedure = OperationVitreous



OptiSource as part of ARIBEC, an Architecture of Mediation for VO



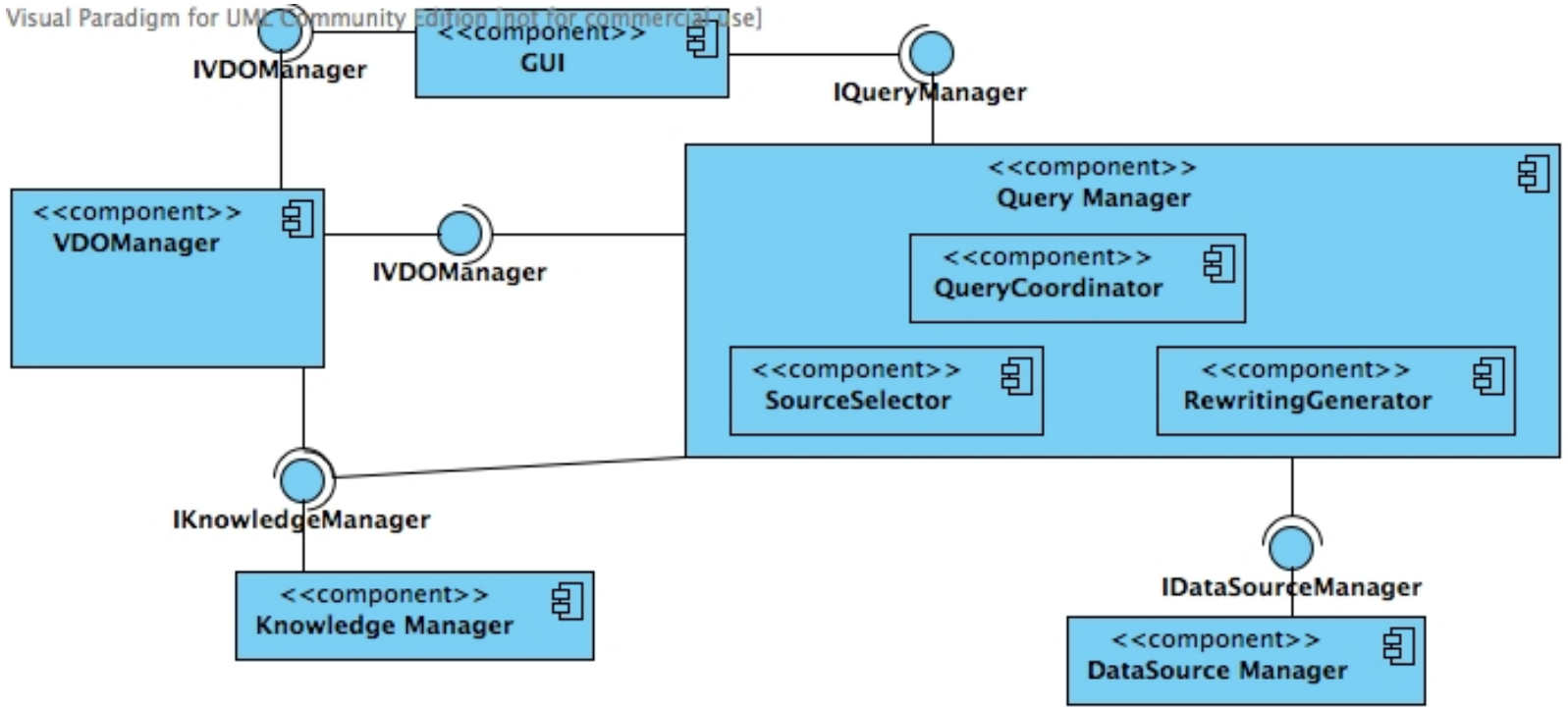
Execution of a Query using ARIBEC



Outline

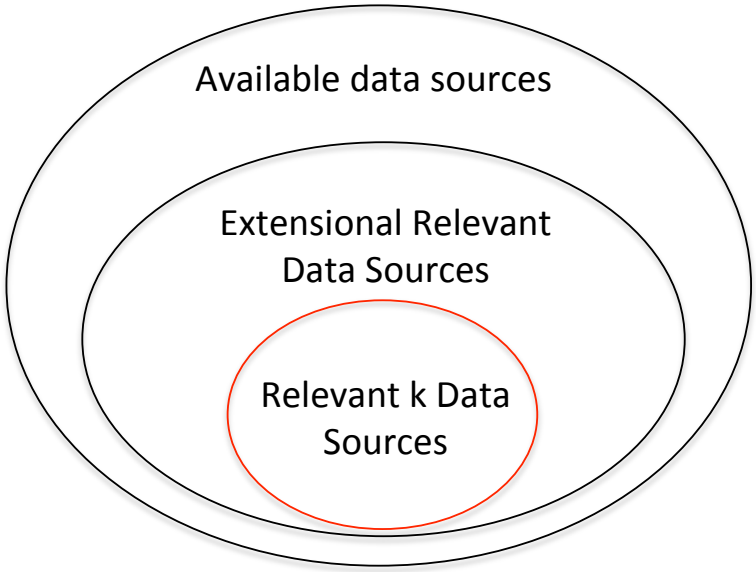
1. **Context and Motivation**: Data integration in virtual organizations
   - ❑ Data integration problem
   - ❑ Data integration solutions applied to virtual organizations
2. **Proposal: A Mediation System for Virtual Organizations**
   - ❑ OptiSource a source selection strategy
   - ❑ Organizational knowledge
   - ❑ Individual Contribution
   - ❑ Group Contribution
   - ❑ Evaluation process
3. **Validation of OptiSource**
   - ❑ Tests of precision and recall
   - ❑ Sensibility analysis
   - ❑ Comparison with IDrips
4. **Conclusions and Future Works**

OptiSource Validation: Prototype



VO Knowledge Base: OWL DL, Protégé as editor, Jena y Pellet as inference engine
Optimization model: Definition in GNU MathProg, execution using GNU Linear Programming Kit (GLPK), GLPK-Java, CPLEX 10.2

OptiSource Validation: Improvement on Source Selection Precision



Context Sets

Rext: Data sources relevant extensionally for Q
Rk: First k data sources relevant extensionally
Aext: Data sources relevant extensionally for Q selected by OptiSource
Ak: First k data sources selected by OptiSource

Precision

$$Extensional\ \Pr ecision = \frac{|A_{ext} \cap R_{ext}|}{A_{ext}}$$

$$\Pr ecisionK = \frac{|A_k \cap R_k|}{A_k}$$

Relevant data sources selected from the total set of data sources selected

Recall

$$\mathrm{Re}\, callK = \frac{|A_{ext} \cap R_k|}{|A_{ext} \cap R_{ext}|}$$

Relevant data sources selected from the total set of relevant data sources available

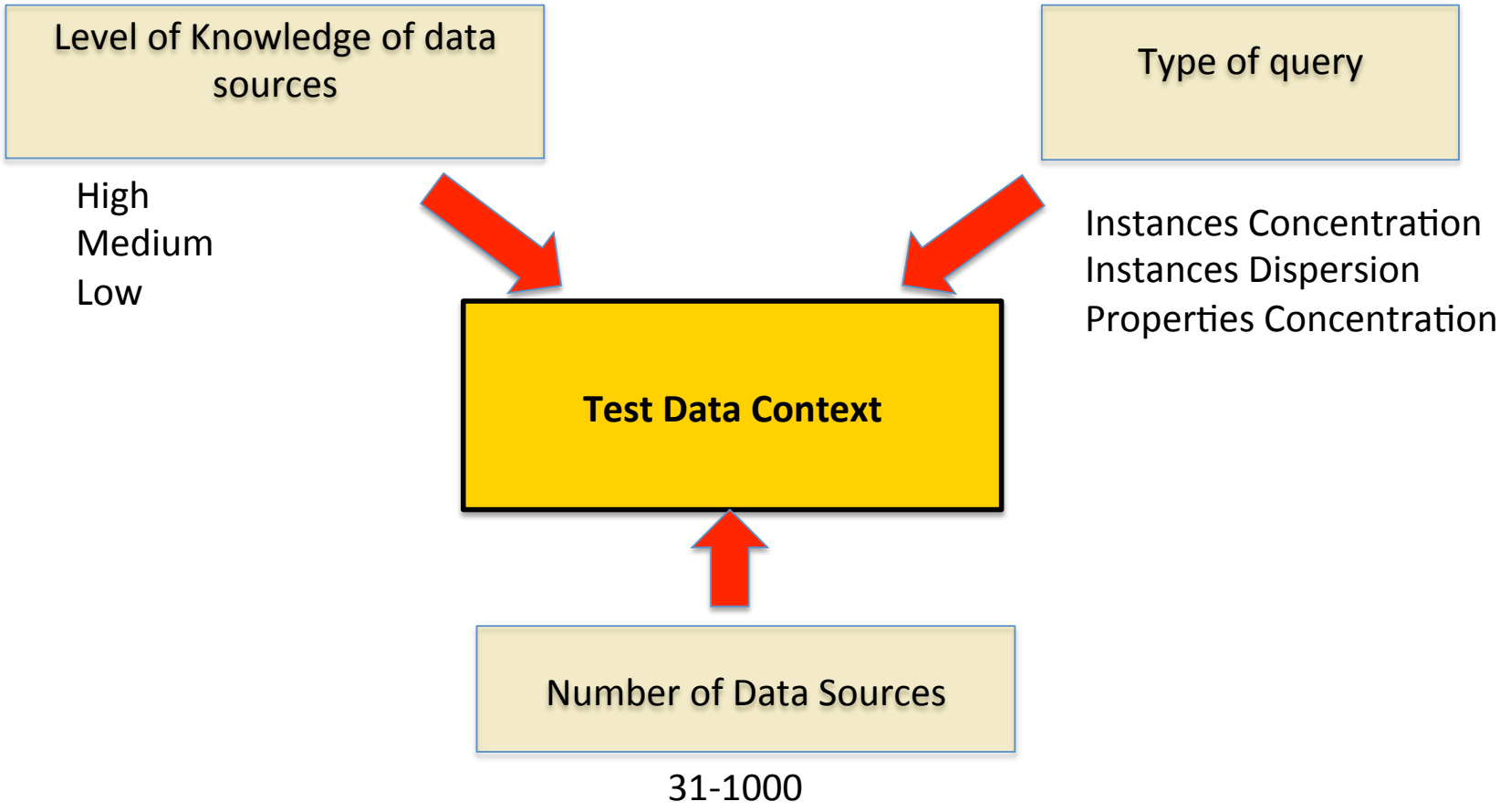
FallOut

$$FallOutExtensional = \frac{|A_{ext} \cap -R_{ext}|}{-R_{ext}}$$

$$FallOutK = \frac{|A_k \cap -R_k|}{-R_k}$$

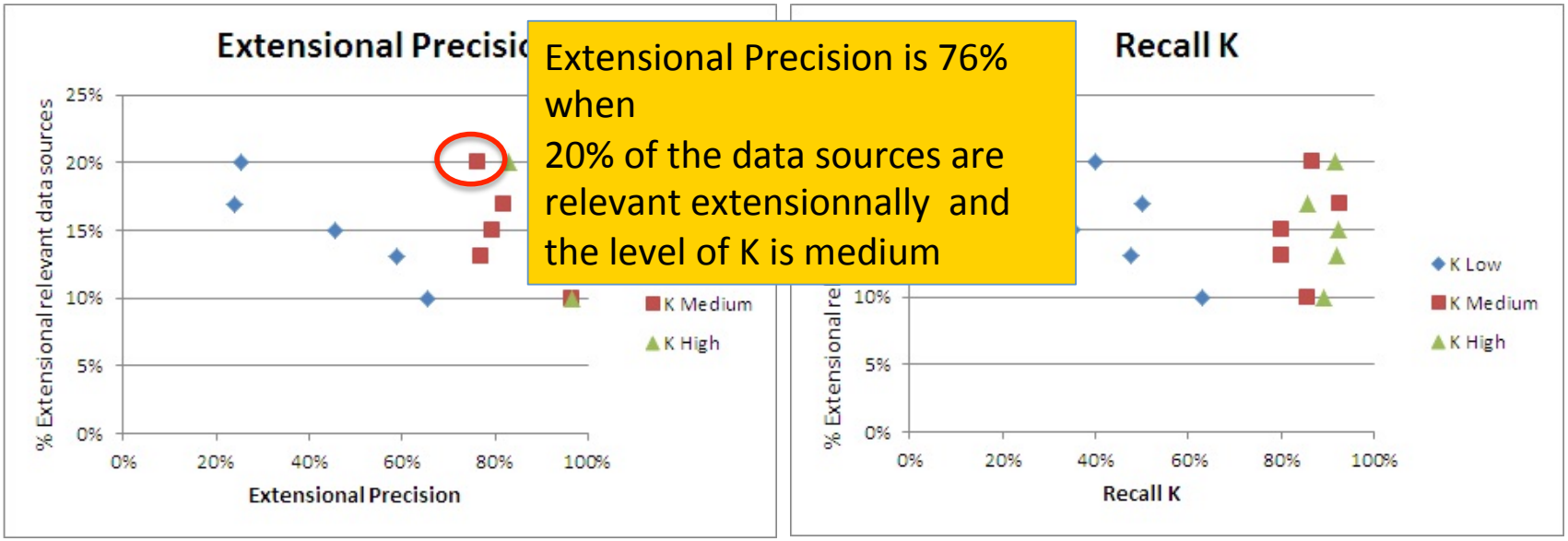
Not relevant data sources selected from the total set of not relevant data sources

OptiSource Validation: Methodology



Variation on the Level of Knowledge

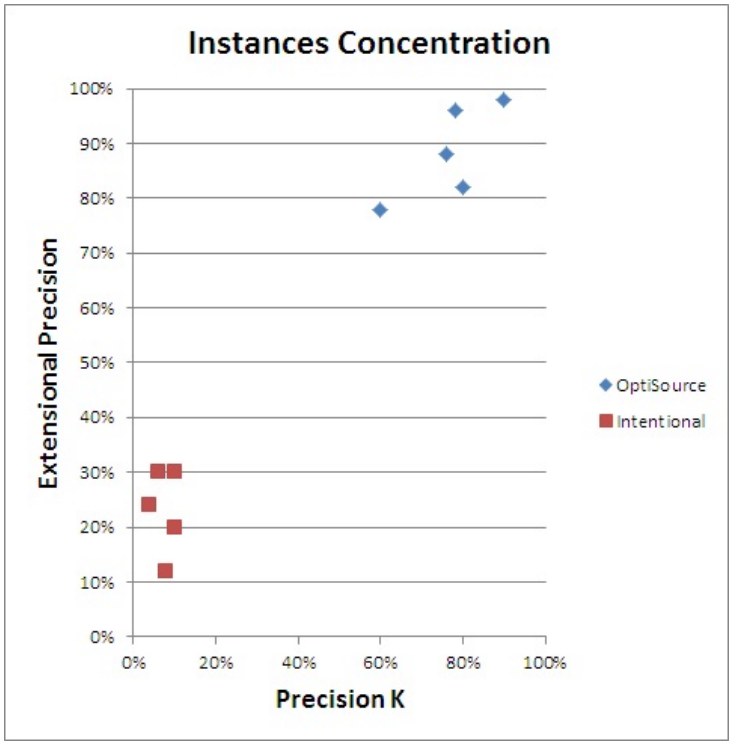
Objective: To evaluate the impact that changes in the level of knowledge have on the precision and recall of OptiSource



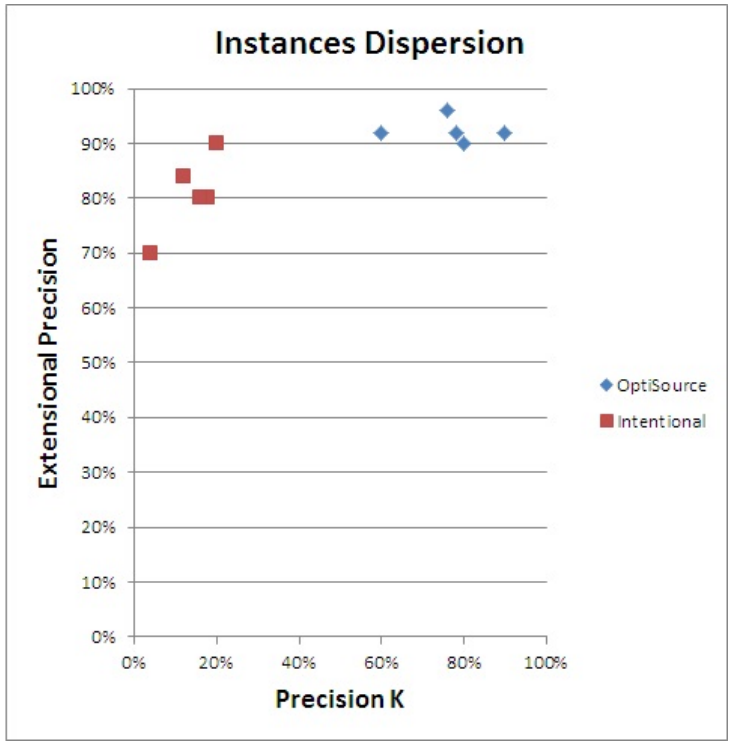
Query Procedure=OperationontheEye and Diagnosis=Glaucoma
Number of Data Sources 501-1000

Variation on the Type of Query-Context RelationShip

Objective: To evaluate the behavior of OptiSource when the relationship between the query and the data context varies and compare it with an intentional strategy.



Procedure=OperationontheEye and Diagnosis=Glaucoma
Level of Knowledge Medium

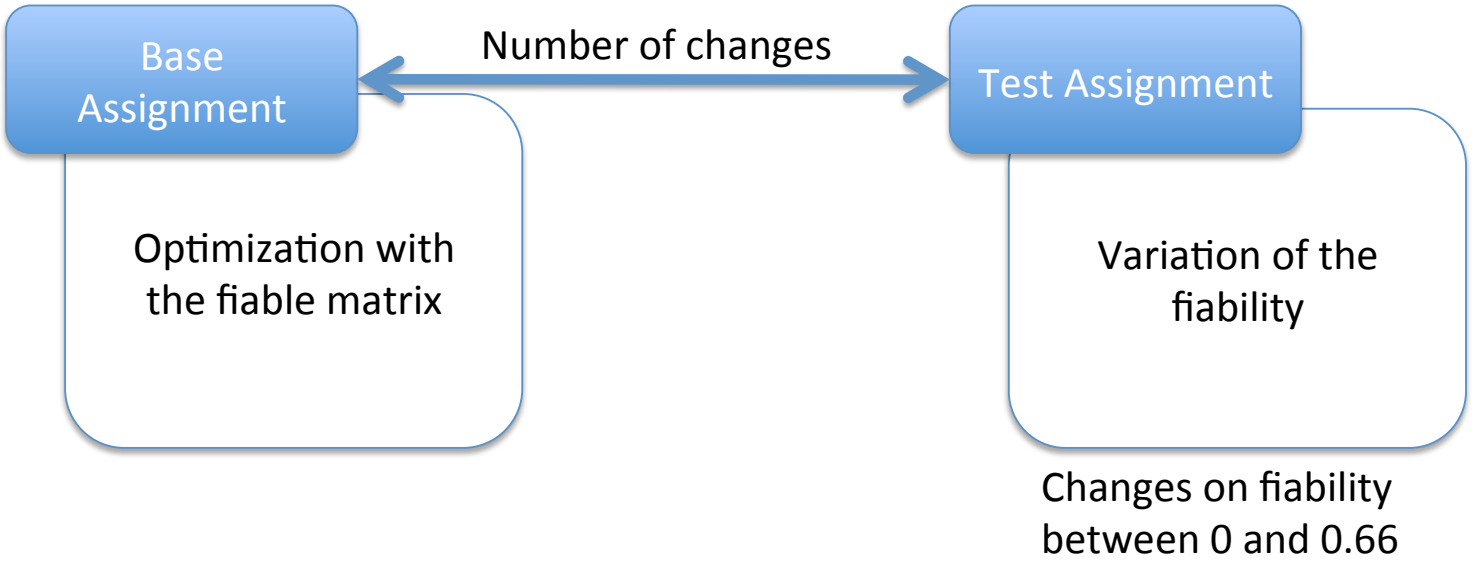


Procedure= Colonoscopy and Diagnosis=Diabetes
Number of Data Sources 501-1000

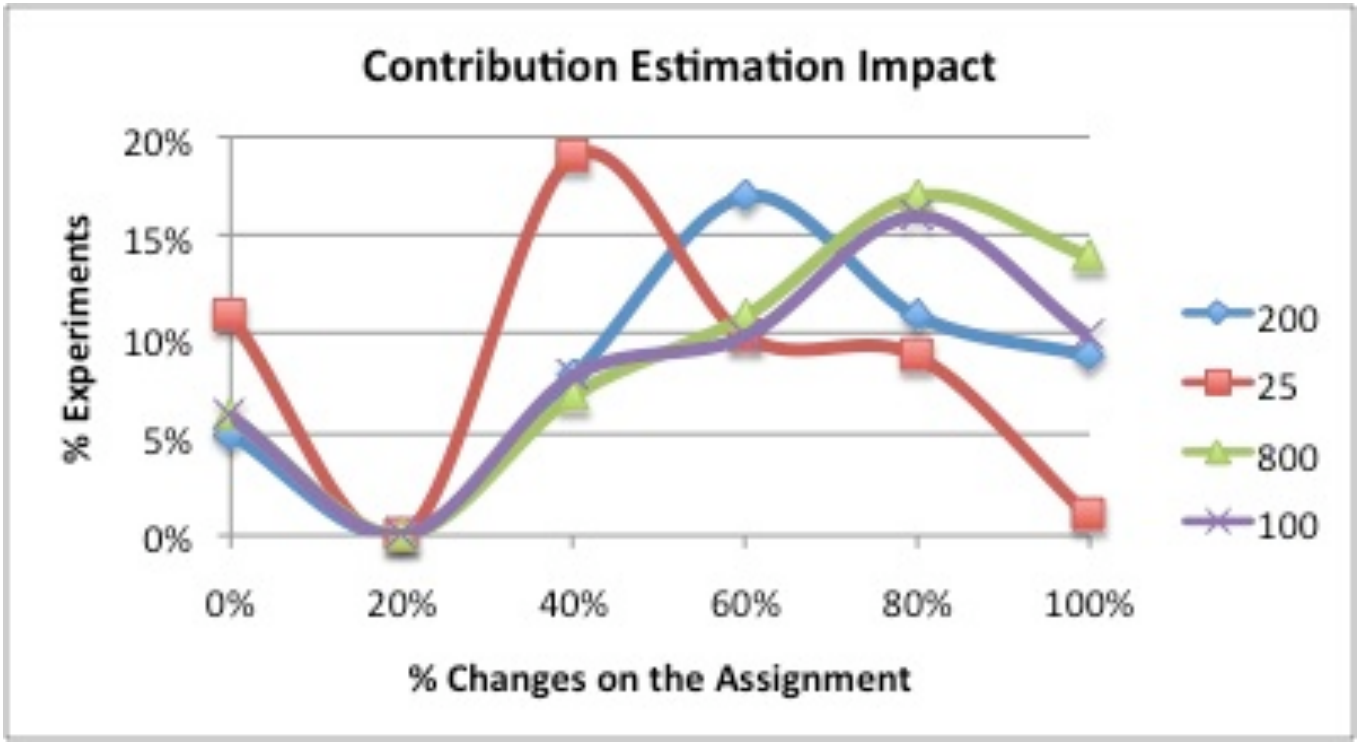
Sensibility Analysis of OptiSource: Methodology

Objective: To measure the impact that have the optimization model w.r.t. the fiability on the contribution estimation.

Number of conditions:5,10,20
Number of data sources: 25,50,100,200,400,800

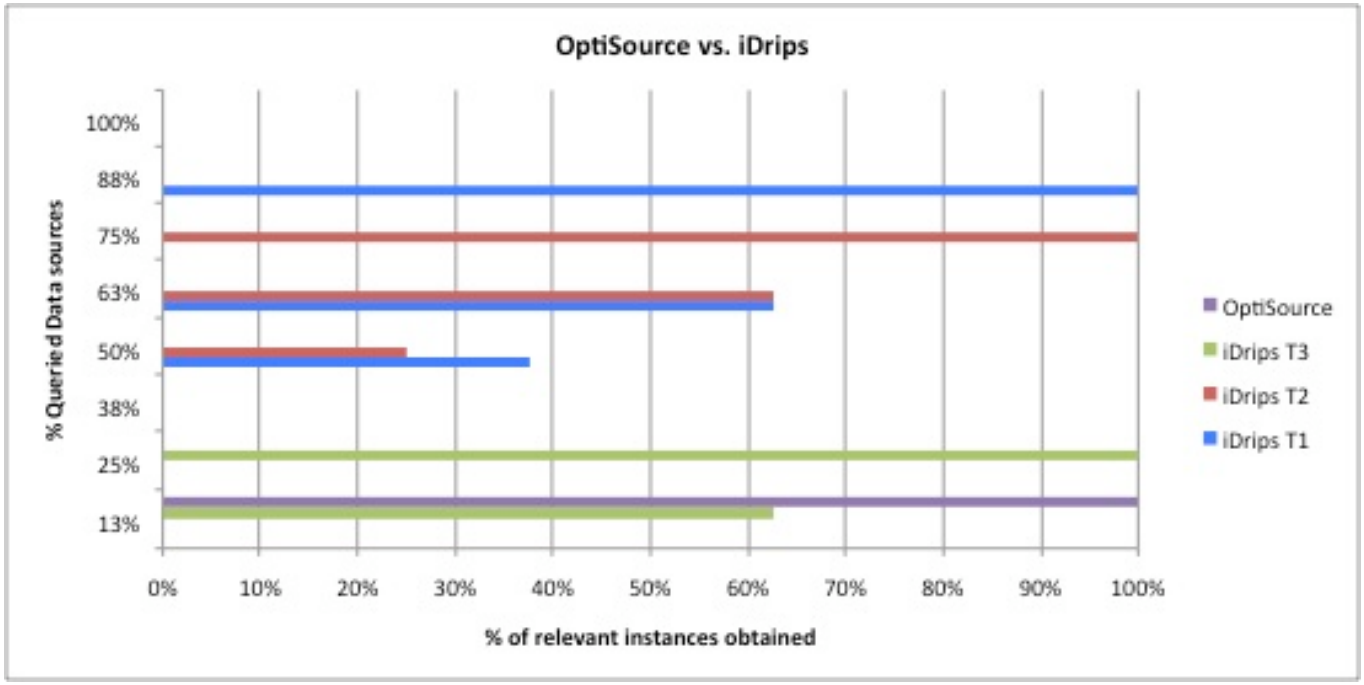


Analysis of Sensibility



Comparison with IDrips

Objective: Compare the behavior of OptiSource and Idrips using different utility measures for IDrips



Outline

1. **Context and Motivation**: Data integration in virtual organizations
   - ❑ Data integration problem
   - ❑ Data integration solutions applied to virtual organizations
2. **Proposal: A Mediation System for Virtual Organizations**
   - ❑ OptiSource a source selection strategy
   - ❑ Organizational knowledge
   - ❑ Individual Contribution
   - ❑ Group Contribution
   - ❑ Evaluation process
3. **Validation of OptiSource**
   - ❑ Tests of precision and recall
   - ❑ Sensibility analysis
   - ❑ Comparison with IDrips
4. **Conclusions and Future Works**

Contributions

- OptiSource a strategy of souce selection for large scale VO
  - Source selection as an optimization model
  - Proved to improve the precision w.r.t. current strategies
  - Proved to select the most relevant data sources
  - Proved to reduce the redundancy in the final response w.r.t. strategies based on general utility functions

- Knowledge Based Strategy of Metadata Definition
  - Allows to describe data contexts with flexibility
  - Uses organizational knowledge to describe data contexts
  - Allows to differentiate data sources using the concept of data source role

OptiSource

Contributions

- ARIBEC
  - An architecture that enriches the mediation level of the mediation architecture
  - Allows to select dynamically the source selection strategy
- Tool to evaluate the sensibility of models that solve the assignment problem

ARIBEC

Future Work

- Apply the principles of OptiSource to enrich other strategies of source selection.
- Integrate the proposal to commercial infrastructure of integration.
- Apply OptiSource in the deep web to improve the selection of data sources using the optimization model.
- Build tools and design other strategies to capture the role of data sources.

Thank you

Questions?

References

- [1] Maurizio Lenzerini. Data integration: a theoretical perspective. In PODS '02: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pages 233–246, New York, NY, USA, 2002. ACM.

- [2] Alon Y. Levy, Anand Ra jaraman, and Joann J. Ordille. Querying heterogeneous information sources using source descriptions. In VLDB '96: Proceedings of the 22th International Conference on Very Large Data Bases, pages 251–262, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc.
-
- [3], Hector Garcia-Molina, Yannis Papakonstantinou, Dallan Quass, Anand Ra jaraman, Yehoshua Sagiv, Jeffrey D. Ullman, Vasilis Vassalos, and Jennifer Widom. The tsimmis approach to media- tion: Data models and languages. Journal of Intel ligent Informa- tion Systems, 8(2):117–132, 1997.

- [4] Ryan Huebsch, Brent N. Chun, Joseph M. Hellerstein, Boon Thau Loo, Petros Maniatis, Timothy Roscoe, Scott Shenker, Ion Stoica, and Aydan R. Yumerefendi. The architecture of pier: an internet- scale query processor. In CIDR '05: Proceedings of the Second Biennial Conference on Innovative Data Systems Research, pages 28–43, 2005.

References

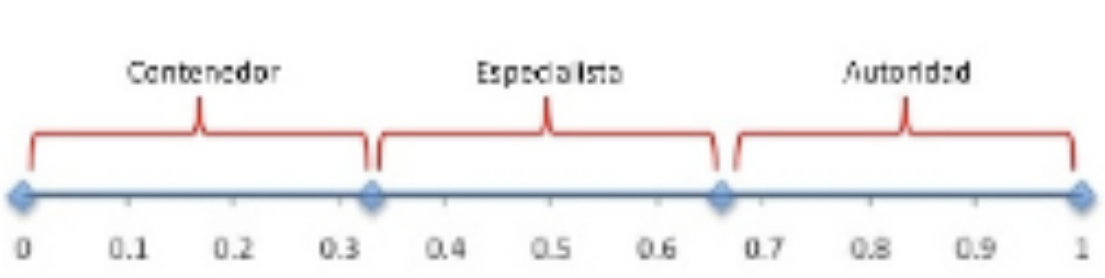
- [5] Hemal Khatri, Jianchun Fan, Yi Chen, and Subbarao Kambham- pati. Qpiad: Query processing over incomplete autonomous da- tabases. In ICDE, pages 1430–1432, 2007.

- [6] Jens Bleiholder, Samir Khuller, Felix Naumann, Louiqa Raschid, and Yao Wu. Query planning in the presence of overlapping sour- ces. In Proceedings of the International Conference on Extending Database Technology (EDBT), pages 811–828, 2006.

- [8] Jinxi Xu and Jamie Callan. Effective retrieval with distributed collections. In SIGIR '98: Proceedings of the 21st annual inter- national ACM SIGIR conference on Research and development in information retrieval, pages 112–120, New York, NY, USA, 1998. ACM.

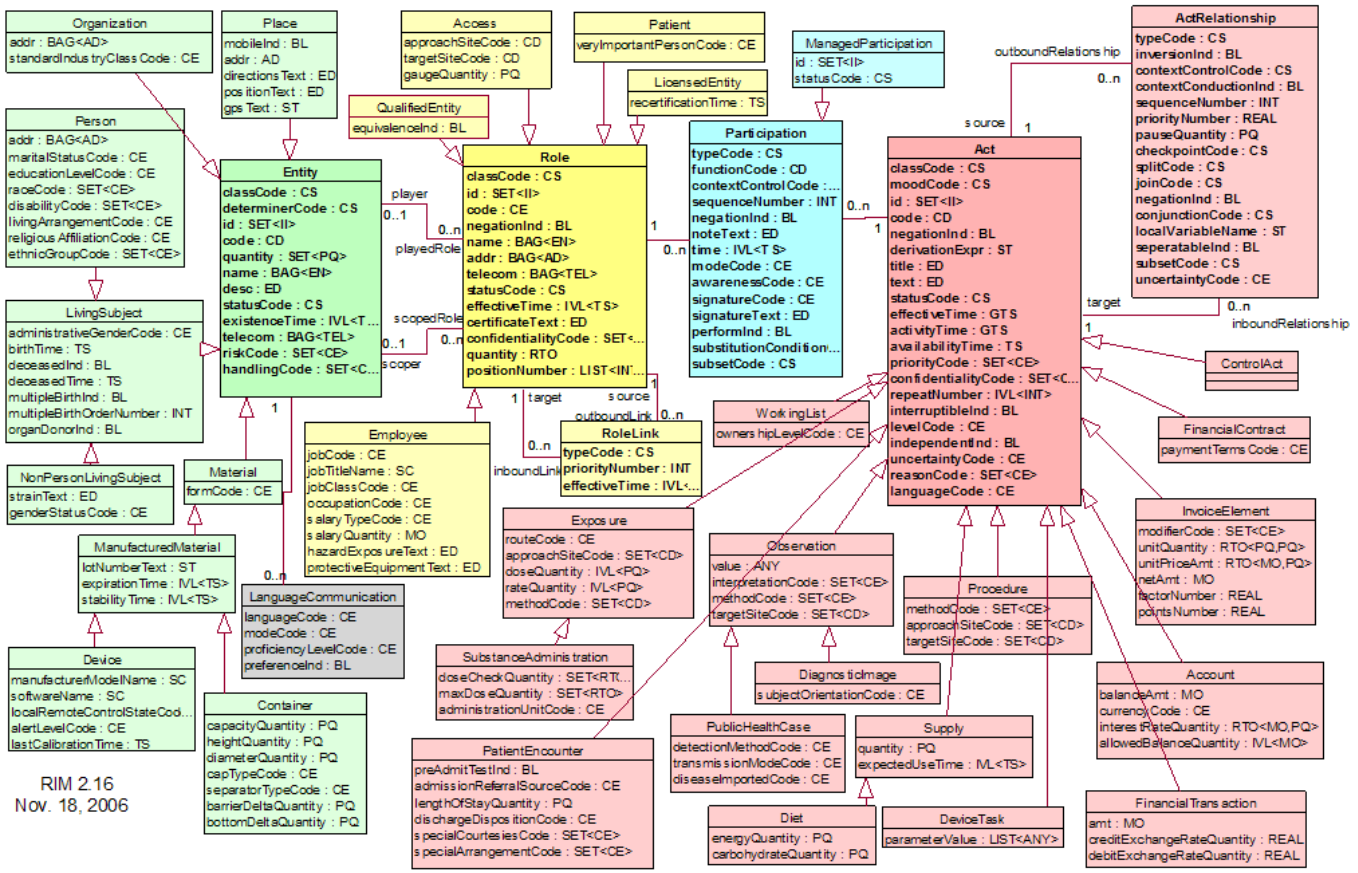
- [9] Panagiotis G. Ipeirotis and Luis Gravano. Distributed search over the hidden web: hierarchical database sampling and selection. In VLDB '02: Proceedings of the 28th international conference on Very Large Data Bases, pages 394–405, 2002.

Role Factor

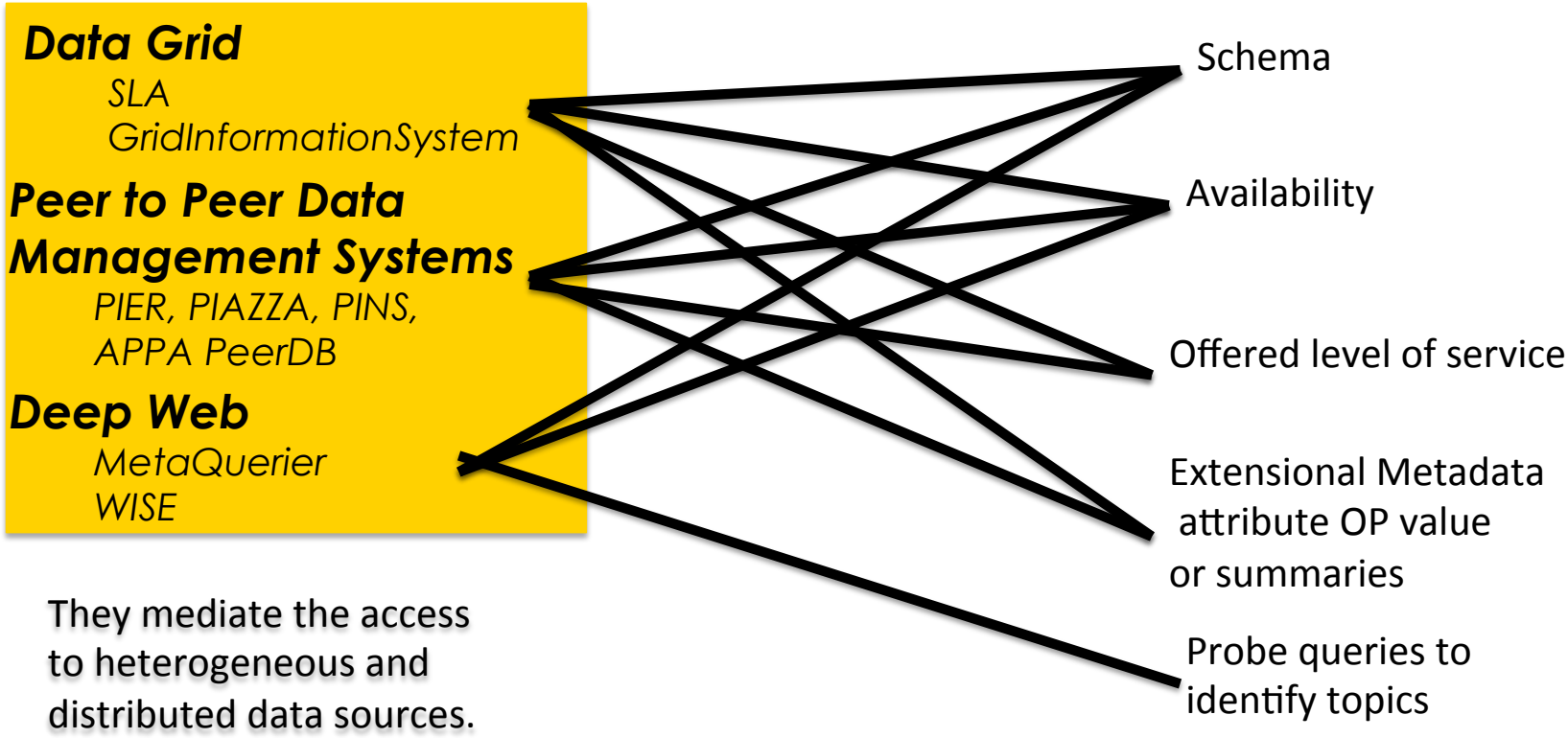


$$FactordeRol(DSi, VDOj, c) = \max((esContenedor(DSi, VDOj, c) * factorContenedor()),$$
$$(esEspecialista(DSi, VDOj, c) * factorEspecialista()),$$
$$(esAutoridad(DSi, VDOj, c) * factorAutoridad())) -$$
$$(diferenciaRol * factorNivelConocimiento(c, maxRol, DSi)$$
$$(6.1)$$

Virtual Organization Domain



Infrastructures



They mediate the access to heterogeneous and distributed data sources.

Complexity Analysis

$$\max \sum_{c=1}^{n} \sum_{i=1}^{m} Ben_{i,c} * (x_{i,c} + assig_{i,c}),$$

Restriction of the **Knapsack problem**

$$\sum_{i=1}^{m} x_{i,c} = 1, \forall c \in C$$

$$\sum_{i=1}^{m} y_i \leq k$$

$$\sum_{c=1}^{n} x_{i,c} \geq y_i \forall i \in I$$

$$\sum_{c=1}^{n} \mathrm{Re}s_{i,c} * assig_{i,c} \leq Max\,\mathrm{Re}s_i, \forall i \in I$$

$$\sum_{c=1}^{n} assig_{i,c} \leq MaxAssig_i, \forall i \in I$$

$$x_{i,c} \leq assig_{i,c}, \forall i \in I, \forall c \in C$$

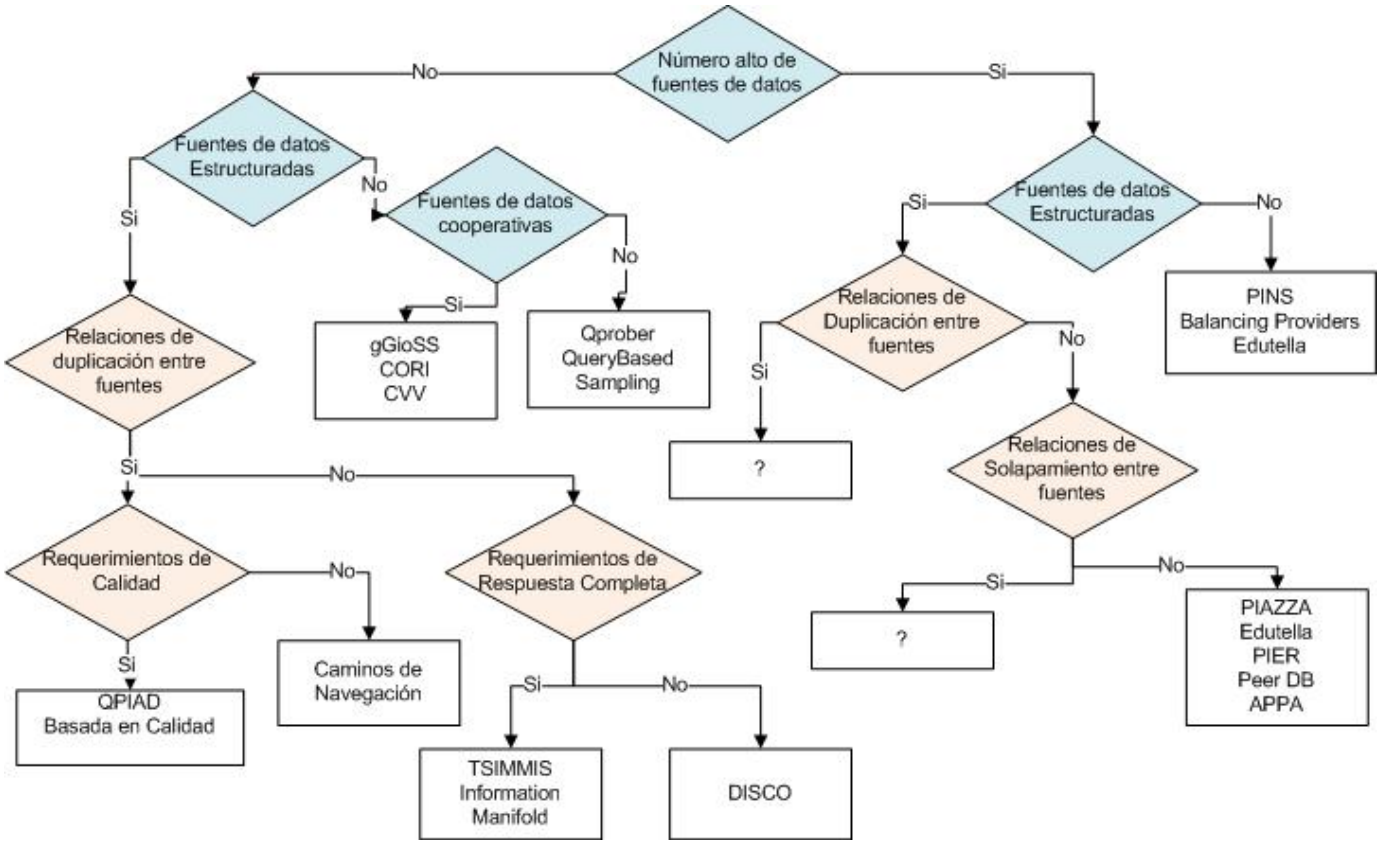
$$assig_{i,c} \leq y_i, \forall i \in I, \forall c \in C$$

Although it is NP Complete, Sensibility analysis proved that it is feasible to find its solution in practice

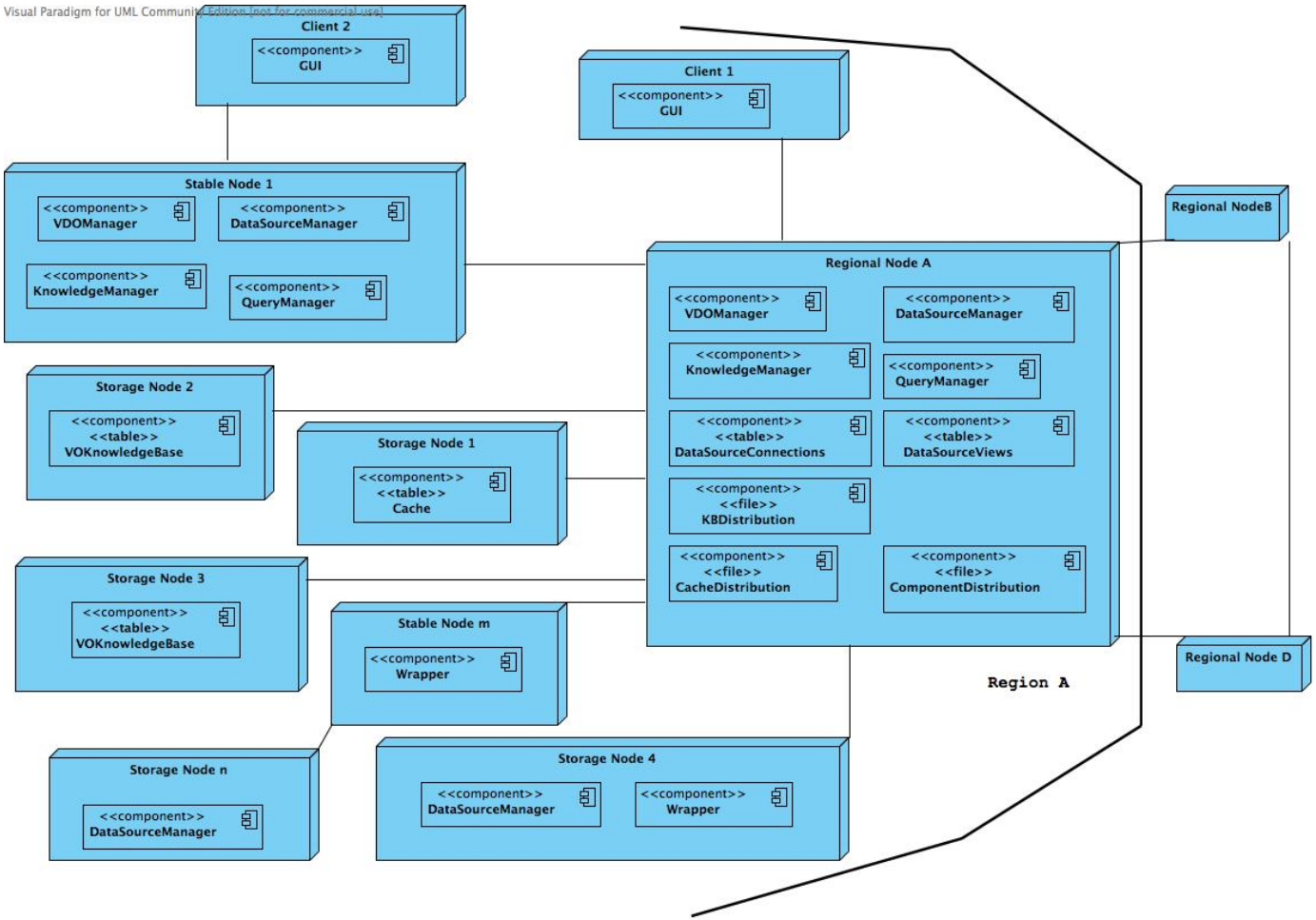
Scalability Problems of SISPRO

- Data Centralization
- ETL when there are updates in the data sources

GAP


Número alto de fuentes de datos
No
Si
Fuentes de datos Estructuradas
No
Fuentes de datos cooperativas
Si
No
Si
Relaciones de duplicación entre fuentes
gGioSS
CORI
CVV
Qprober
QueryBased
Sampling
Si
No
Requerimientos de Calidad
No
Requerimientos de Respuesta Completa
Si
QPIAD
Basada en Calidad
Caminos de Navegación
Si
No
TSIMMIS
Information
Manifold
DISCO
Si
Fuentes de datos Estructuradas
No
PINS
Balancing Providers
Edutella
Si
Relaciones de Duplicación entre fuentes
No
Si
?
Relaciones de Solapamiento entre fuentes
Si
No
?
PIAZZA
Edutella
PIER
Peer DB
APPA

Physical Architecture



Future Validation

- Changing different elements
- Grid Infrastructure