ERIC
ENTREPÔTS, REPRÉSENTATION & INGÉNIERIE des CONNAISSANCES

EDA

EDA 2013

13-14 Juin 2013
Blois, France

Summarizability Issues in Multidimensional Models: A Survey*

AUTHORS:
MAROUANE HACHICHA
JÉRÔME DARMONT

*Problèmes d'additivité dus à la présence de hiérarchies complexes dans les modèles multidimensionnels : définitions, solutions et travaux futurs

Lyon 1
UNIVERSITÉ LUMIÈRE LYON 2 UNIVERSITÉ DE LYON
UNIVERSITÉ DE LYON

Introduction

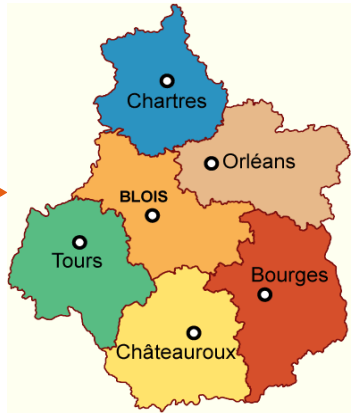
- Data warehousing and OLAP tools → decision making process
- The development of these systems is based on multidimensional modeling of real-world situations
- Particular nature of some real-world situations
- Summarizability problems

Particular real-world situations

Incomplete Hierarchies (Geographical dimension: *country-region-city*)

Country: France

Regions of France



Cities of *Région Centre*

Particular real-world situations

Incomplete Hierarchies (Geographical dimension: *country-region-city*)

Country: San Marino

Cities of San Marino

There is not any region in San Marino

Particular real-world situations

Incomplete Hierarchies (Geographical dimension: *country-region-city*)

City	Sales
Blois	2000 €
Tours	1500 €
Paris	1500 €
San Marino	2500 €
Acquaviva	2500 €
Total	**10000 €**

Particular real-world situations

Incomplete Hierarchies (Geographical dimension: *country-region-city*)

City	Sales
Blois	2000 €
Tours	1500 €
Paris	1500 €
San Marino	2500 €
Acquaviva	2500 €
Total	**10000 €**

Region	Sales
Centre	3500 €
Ile-de-France	1500 €
Total	**5000 €**

Particular real-world situations

Incomplete Hierarchies (Geographical dimension: *country-region-city*)

City	Sales
Blois	2000 €
Tours	1500 €
Paris	1500 €
San Marino	2500 €
Acquaviva	2500 €
Total	**10000 €**

Region	Sales
Centre	3500 €
Ile-de-France	1500 €
Total	**5000 €**

Country	Sales
France	5000 €
Total	**5000 €**

Particular real-world situations

Incomplete Hierarchies (Geographical dimension: *country-region-city*)

City	Sales
Blois	2000 €
Tours	1500 €
Paris	1500 €
San Marino	2500 €
Acquaviva	2500 €
Total	**10000 €**

Region	Sales
Centre	3500 €
Ile-de-France	1500 €
Total	**5000 €**

Country	Sales
France	5000 €
Total	**5000 €**

Country	Sales
France	5000 €
San Marino	5000 €
Total	**10000 €**

If we do not take into account the problem related to regions (Incomplete hierarchies) → Summarizability problems → incorrect analysis results → erroneous decisions.

Particular real-world situations

Non-strict Hierarchies (Date dimension: *week-month*)

Plan of the Month:
4 Weeks to More Muscle

Particular real-world situations

Non-strict Hierarchies (Date dimension: *week-month*)

Plan of the Month:
4 Weeks to More Muscle

1 week = 7 days
4 weeks = 28 days <= 1month

1 month = 4 weeks and some days of the 5th week

Particular real-world situations

Non-strict Hierarchies (Date dimension: *week-month*)

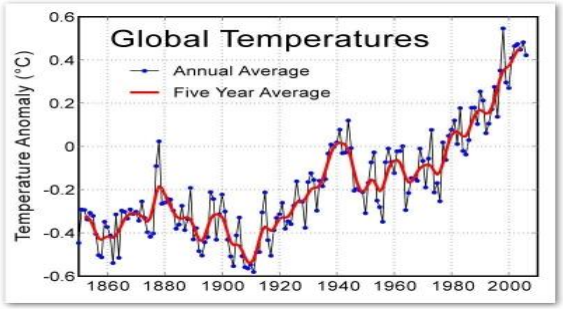
Week	Sales
Week#1 April 2013	10
W#2 April 2013	10
W#3 April 2013	10
W#4 April 2013	10
W#5 April / W#1 May 2013	10
W#2 May 2013	10
W#3 May 2013	10
W#4 May 2013	10
W#5 May 2013 / W#1 June 2013	10
Total	90

Month	Sales
April 2013	50
May 2013	50
Total	100

Double counting problem for sales due to non-strictness.

Type compatibility problem

OLAP tool



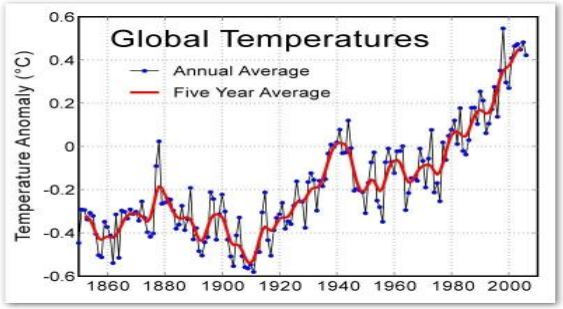
Q1: Average of temperatures

Q2: Sum of temperatures

User

Type compatibility problem

OLAP tool



Q1: Average of temperatures

Q2: Sum of temperatures

User

If we do not guide the user → erroneous decisions

Summarizability

- **Definition:** Correct computation of aggregate values with a coarser level of detail from aggregate values with a finer level of detail [Mazon *et al.* 09].

[Mazon *et al.* 09] **A survey on summarizability issues in multidimensional modeling,** In *Data & Knowledge Engineering*, 2009.

Summarizability

- **Definition:** Correct computation of aggregate values with a coarser level of detail from aggregate values with a finer level of detail [Mazon *et al.* 09].

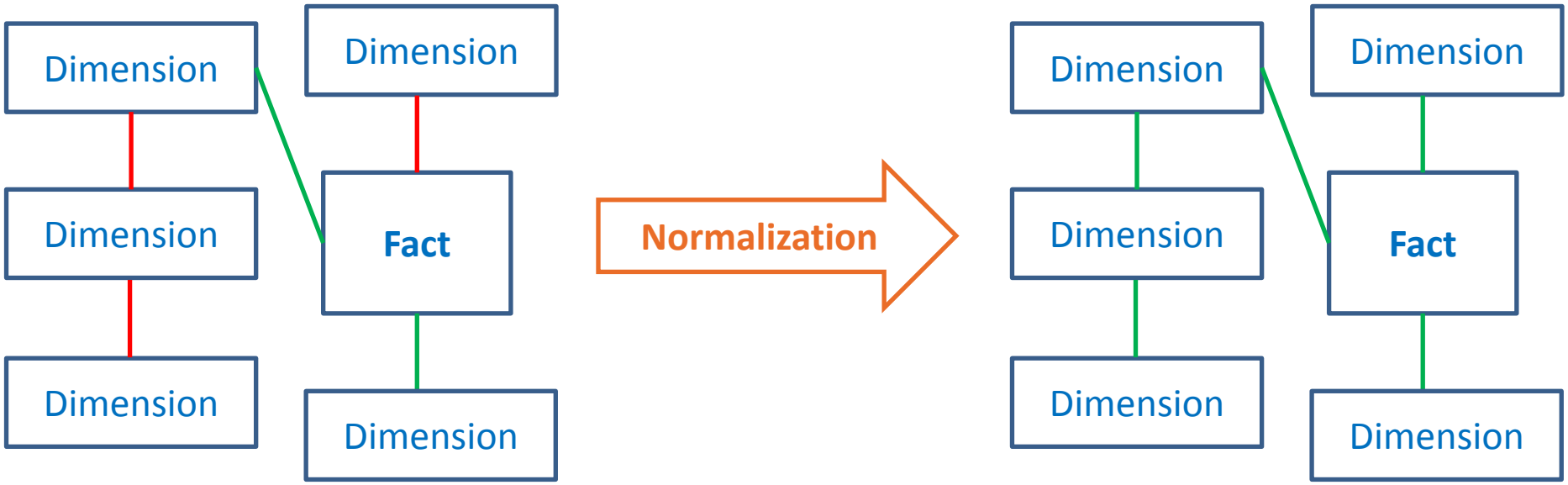
- **Constraints for Summarizability**
  - Let us consider the association between two levels of a dimension hierarchy (fact-dimension relationships);
  - Zero-to-many associations must be avoided (There must not be missing values). Otherwise: Incomplete hierarchy;
  - Many-to-many associations must be avoided. Otherwise: Non-strict hierarchy;
  - Incomplete hierarchy + Non-strict hierarchy = Complex hierarchy.
  - Type compatibility: ensures that the applied aggregate function is summarizable.

[Mazon *et al.* 09] **A survey on summarizability issues in multidimensional modeling,** In *Data & Knowledge Engineering*, 2009.

Outline

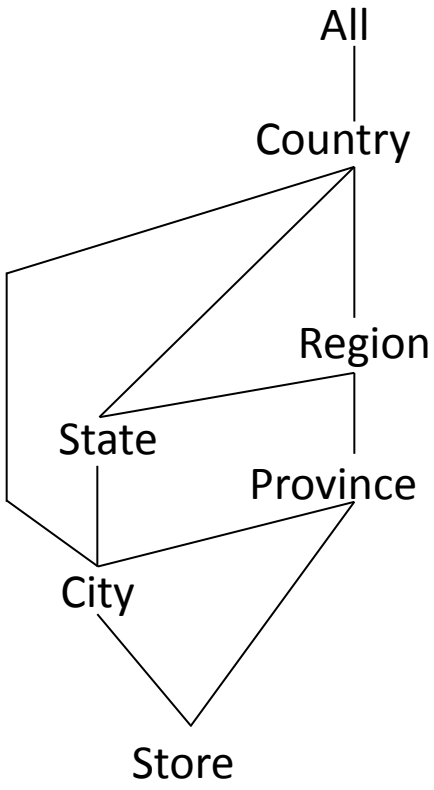
- **Solving Summarizability problems in complex hierarchies**
- Type compatibility
- Conclusion and Future Work

Normalization of multidimensional models



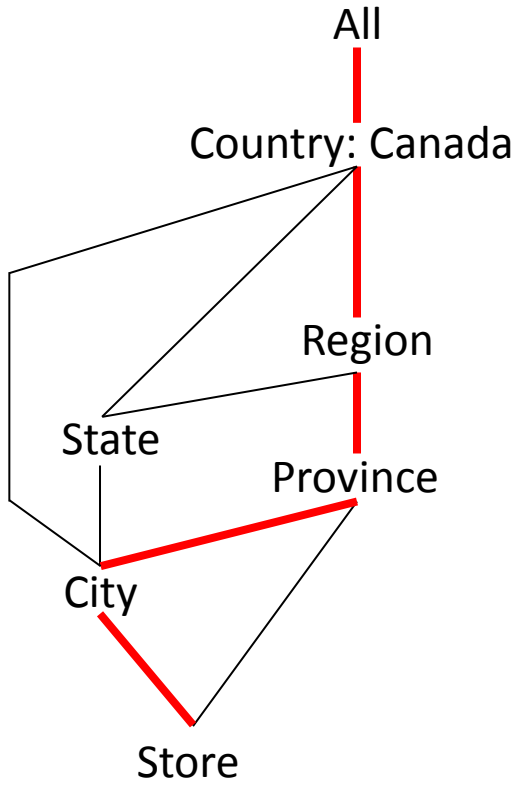
Transforming complex hierarchies to simple (*strict and complete*) hierarchies

Normalization of multidimensional models



(a) [Store, City]
(b) [City= "Paris"] ⇒ [Country= "France"]
(c) [City, ..., Country = "France" ∨ City, ..., Country = "USA"]
(d) [Province, ..., Country = "Canada"]

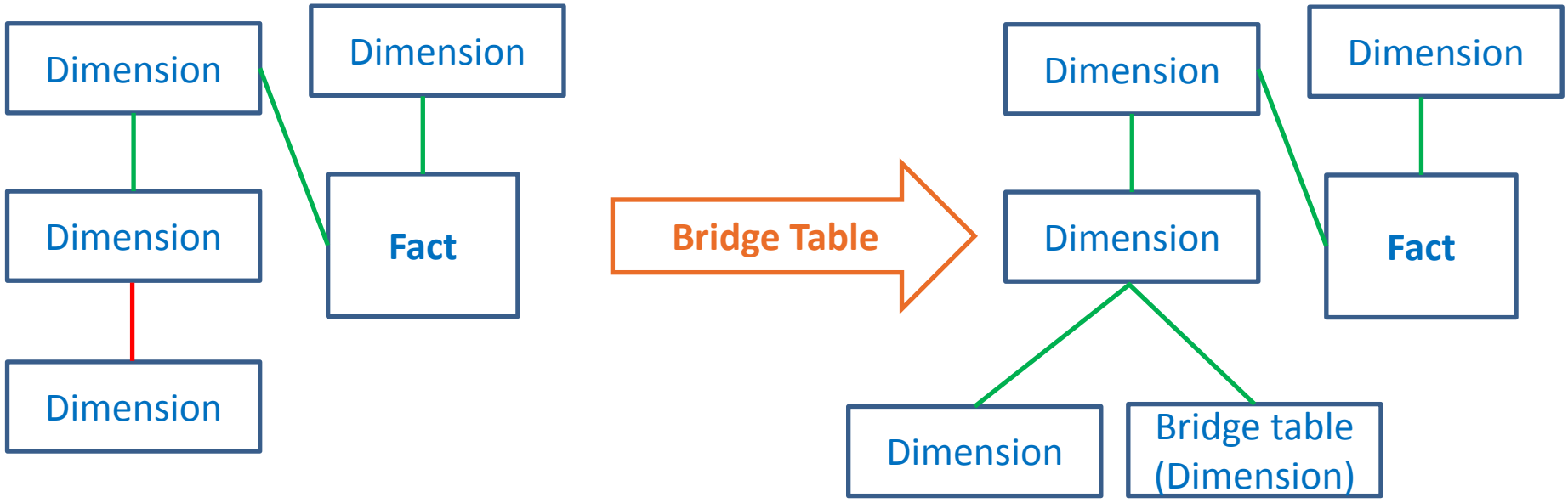
Integrity constraints for multidimensional model



Frozen dimensions

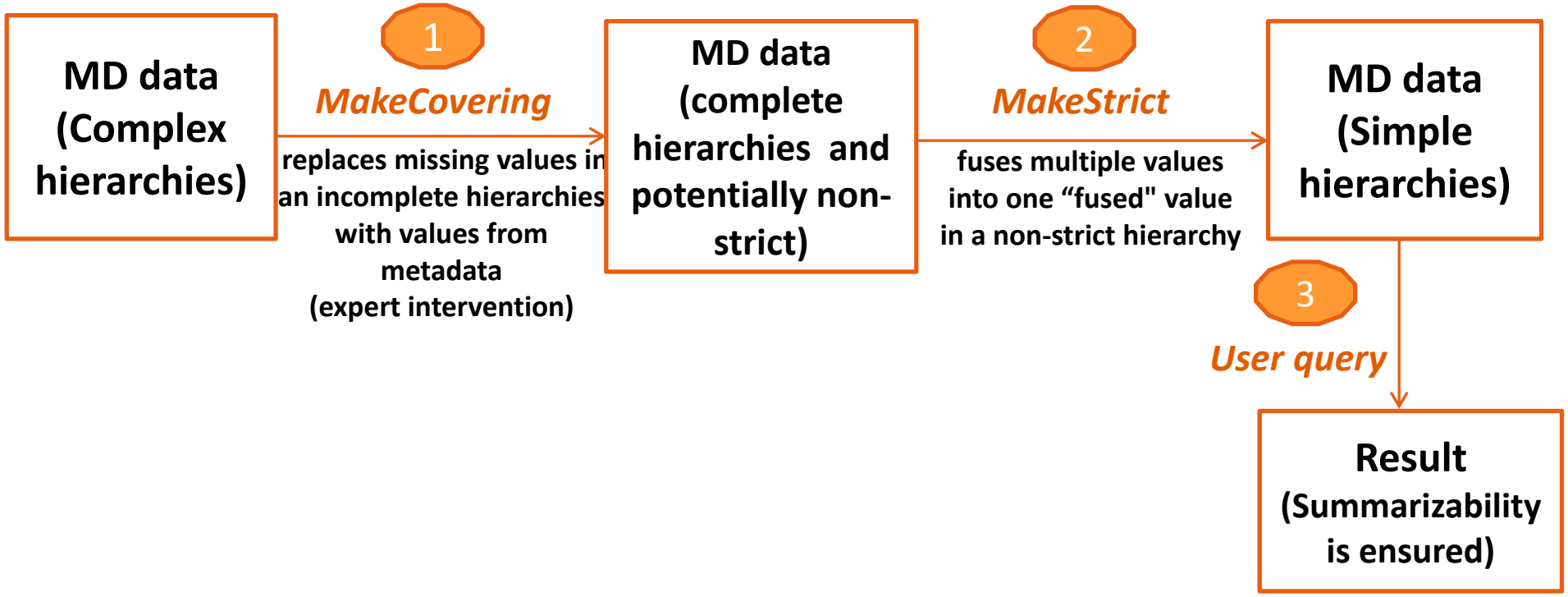
[Hurtado *et al.* 05] **Capturing Summarizability with Integrity Constraints in OLAP**, In *ACM Trans. Database Syst.*, 2005.

Normalization of multidimensional models



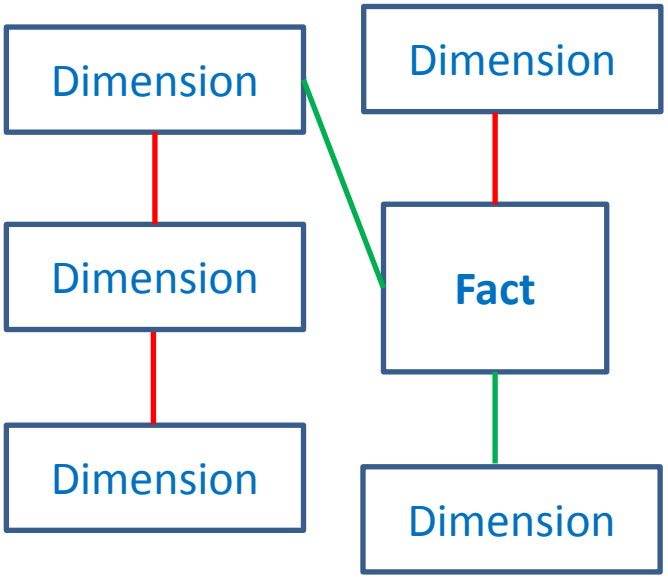
[Mazon *et al.* 08] **Solving Summarizability Problems in Fact-Dimension Relationships for Multidimensional Models**, In *DOLAP*, p 57-64, 2008.

Transformation of multidimensional data



[Pedersen *et al.* 99] **Extending Practical Pre-Aggregation in On-Line Analytical Processing,** In *VLDB*, p 663-674, 1999.

Query-time approaches



Multidimensional model with complex and simple hierarchies

User query

Detecting and solving summarizability problems

User

[Pedersen *et al*. 02] **A Powerful and SQL-Compatible Data Model and Query Language for OLAP,** In *ADC* 2002.

[Horner and Song 05] **A Taxonomy of Inaccurate Summaries and Their Management in OLAP Systems**, In *ER* 2005.

[Hachicha *et al*. 12] **A Novel Query-Based Approach for Addressing Summarizability Issues in XOLAP**, In *COMAD* 2012.

Comparison of solutions

	MD Schema	MD Data	Expert intervention	Solving Summarizability problemes
Normalization	Yes	No	Yes	Yes
Transformation	No	Yes	Yes	Yes
Query-time	No	No	No	Yes*

*[Hachicha *et al*. 12] **A Novel Query-Based Approach for Addressing Summarizability Issues in XOLAP.** In *COMAD*, Pune, India, pages 56-67, CSI, Maharashtra, 2012.

Outline

- Solving Summarizability problems in complex hierarchies
- **Type compatibility**
- Conclusion and Future Work

Type compatibility

- **Type compatibility:** ensures that the aggregate function applied to a measure is summarizable according to the type of the measure and the type of the related dimensions [Lenz and Shoshani 97]

- In a recent work, Prat *et al.* propose to associate **aggregation rules** to the multidimensional model [Prat *et al.* 11]

[Lenz and Shoshani 97] **Summarizability in OLAP and statistical data bases,** In *SSDBM* 1997.

[Prat *et al.* 11] **Combining objects with rules to represent aggregation knowledge in data warehouse and OLAP systems,** In *Data Knowl. Eng.* 70(8), 732–752.

Type compatibility

Aggregation rules	Definition	Example
Semantic rules	Semantics of dimensions, measures, aggregation functions	Ratios are not additive along any dimension
Syntactic rules	Making the sum of averages does not make sense	Sum of temperatures has no meaning
User preferences	When a given function should be used preferably to other aggregation functions	The aggregation function SUM should be used preferably to other aggregation functions
Aggregation execution rules	To deal with complex hierarchies	Sums along non-strict hierarchies are performed by consider null values as 0

[Prat *et al.* 11] **Combining objects with rules to represent aggregation knowledge in data warehouse and OLAP systems,** In *Data Knowl. Eng.* 70(8), 732–752.

Outline

- Solving Summarizability problems in complex hierarchies
- Type compatibility
- **Conclusion and Future Work**

Conclusion

- Importance of taking into account Summarizability issues in Multidimensional modeling and OLAP analysis
- Various solutions at various levels (MD schema – MD data – Query)
- Advantages/Disadvantages for each solution

Future Work

- DWs are more and more used in other novel areas (biology, multimedia…) ⇒ Ensuring Summarizability by considering specific semantics of each domain
- Including Type compatibility rules in query-time approaches
- Normalization / Transformation / Query-time

  ⇒ Summarizability should be considered in all DW and OLAP tools