Data-Driven Multidimensional Design for OLAP

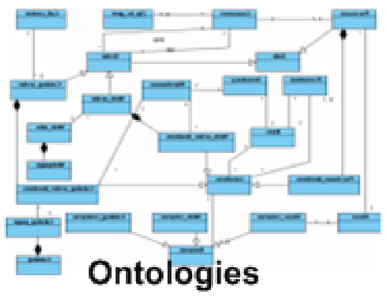
Oscar Romero **Alberto Abelló**

Universitat Politècnica de Catalunya, BarcelonaTech

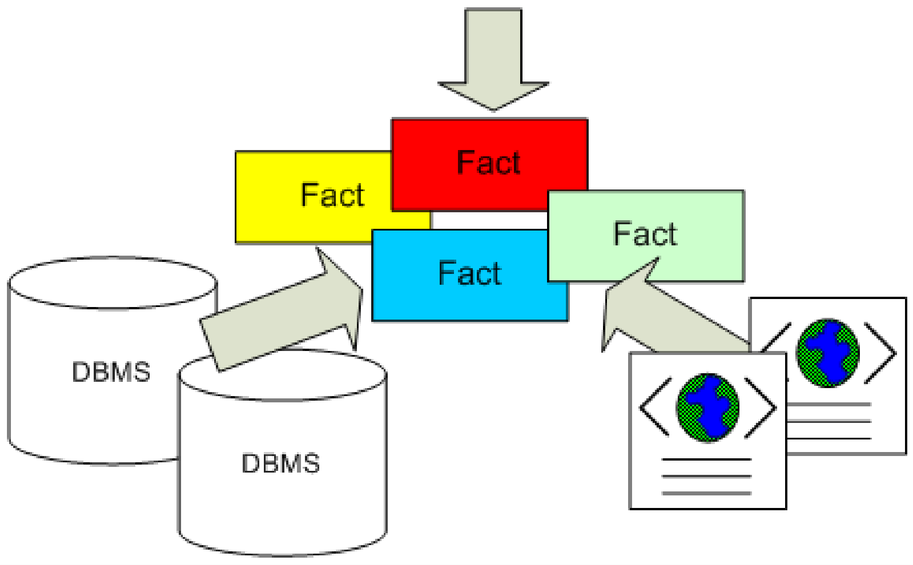
[oromero | aabello]@essi.upc.edu

We aim at <u>efficiently</u> identify relevant concepts for multidimensional (MD) modeling…

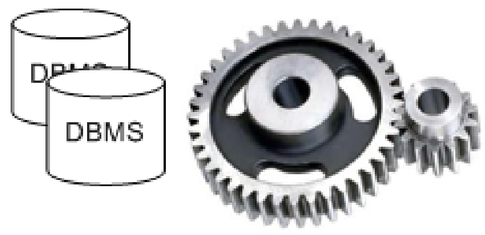
<u>1st Stage:</u> According to the available sources description / technologies, identify facts of interest



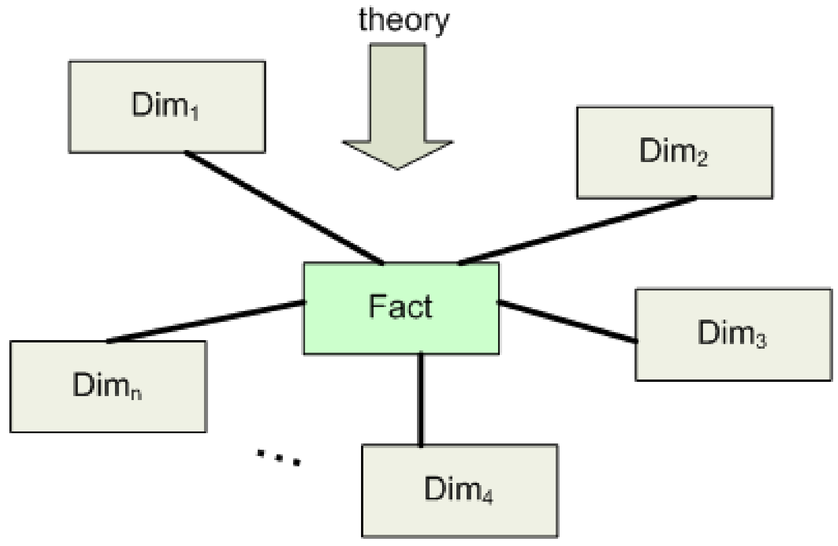
Ontologies



<u>2nd Stage:</u> For each fact, identify interesting dimensional concepts



Functional dependencies theory



<u>Current situation on MD design</u>

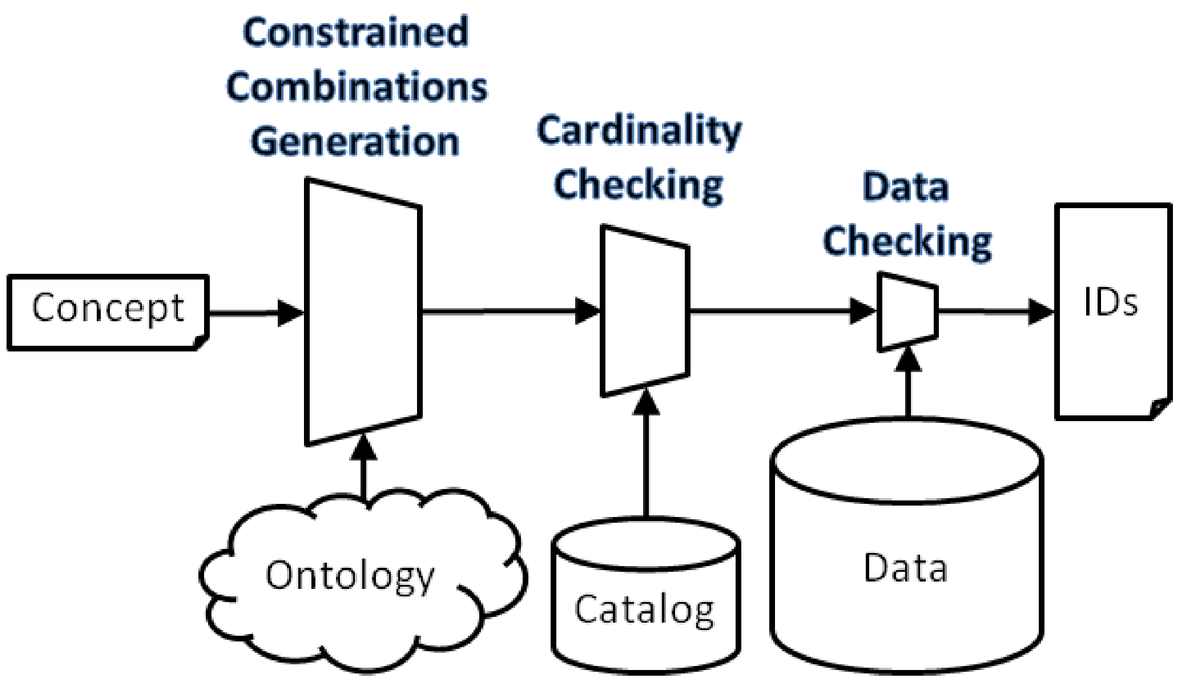
Although different approaches to identify facts are available, dimensional data is mainly identified by means of *functional dependencies (FDs)*. Thus:

• Methods working over RDBMS demand a certain degree of normalization to identify them through FKs-CKs at the logical level or apply expensive algorithms to discover FDs over the instances.

• Furthermore, these methods produce an unbearable amount of results that need to be *manually* filtered.

As consequence, dimensional data discovery is inefficient / too expensive in most real projects.

How Does our Approach Work?



- We use the available <u>ontological knowledge</u> to drive the FD search.

[objective: avoid a combinatorial explosion by also exploiting the domain conceptual layer]

-Metadata (e.g., DB catalog) is used for filtering purposes.

[objective: check necessary conditions that should be preserved before sampling data]

- A <u>statistical study</u> is performed to identify dimensions of interest.

- **ANOVA: A test on the variance to decide whether the difference in the means of several samplings are due to differences in the populations or can be reasonably attributed to chance fluctuations alone (we choose attributes of objective interest when *partitioning* the fact measures).**

[objective: avoid generating too many results according to objective evidences. FDs should NOT be considered as dimensional concepts by default!]