Data Quality Problems (DQPs) at the Instance Level

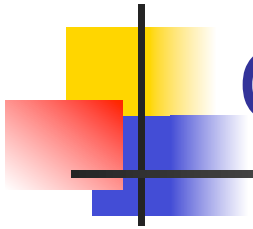
Paulo Oliveira

isep Instituto Superior de Engenharia do Porto

gecad
Grupo de Investigação em Engenharia do Conhecimento e Apoio à Decisão

Toward Efficient Portuguese and Brazilian Electricity Markets Workshop

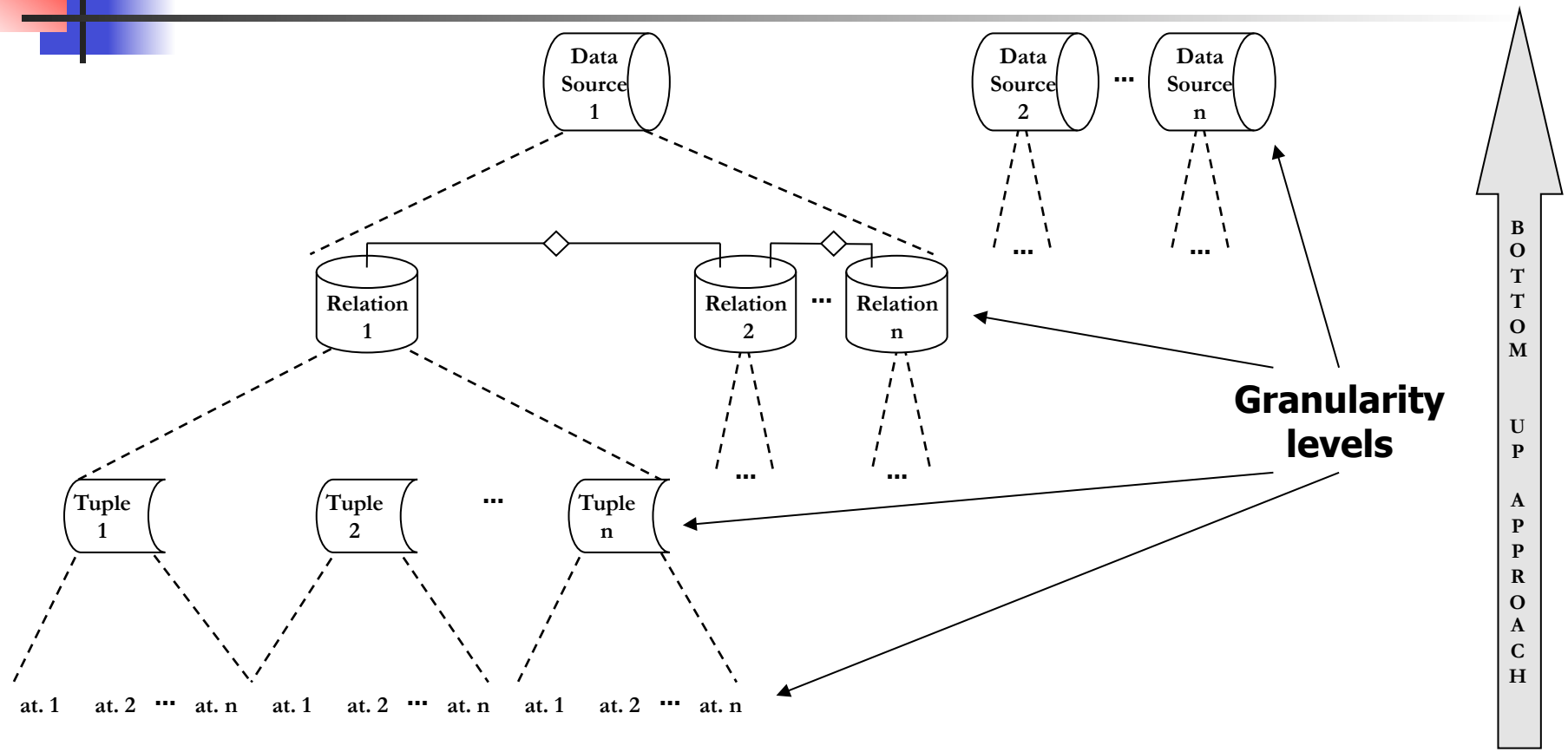
September 25, 2013



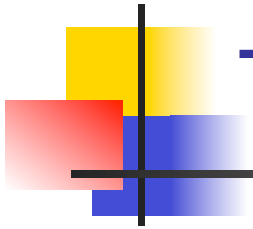
Concept of DQP

- DQP is restricted to problems related with the quality of the data values or instances
- Data values are affected by different kinds of quality problems (errors, anomalies, or *dirtiness*)
- DQPs arise in:
  - single data source
  - data migration
  - integration of multiple data sources
  - data-based projects
    - data warehouses
    - data mining
- Important due to the *Garbage In Garbage Out* principle

Approach to Identify DQPs



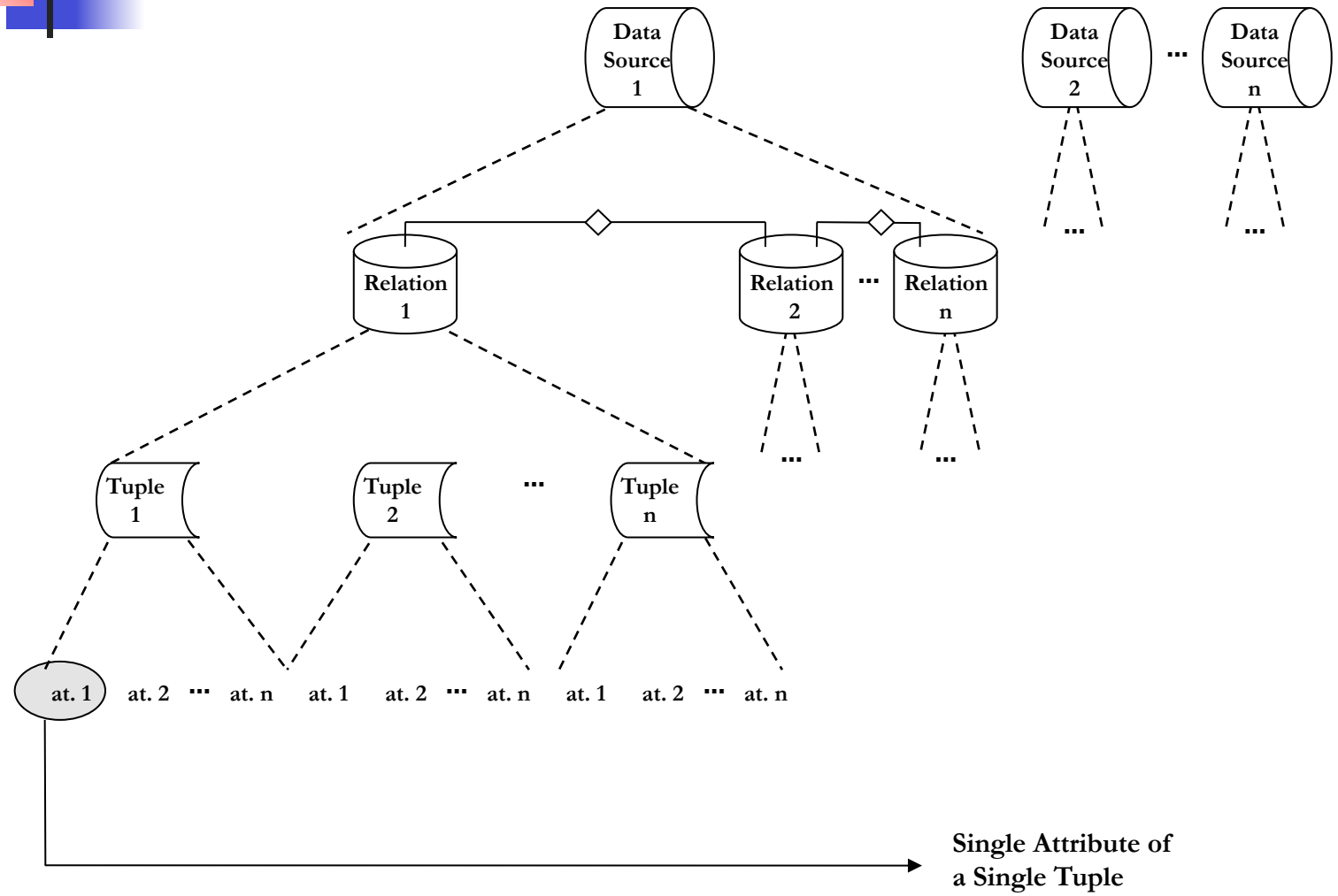
- Based on this model of data organization
- Identification of all the DQPs that can be found at each granularity level



Taxonomy of DQPs

- Covers
  - DQPs of tabular data
  - All data types, except multimedia
- Is complete
- Is limited to DQPs related with the attribute values

DQPs in a Single Attribute of a Single Tuple



Missing Value

	at_1	at_2	**name**	...	at_n
t_1	xxx	xxx	Carl Louis	...	xxx
t_2	xxx	xxx		...	xxx
. . .	. . .	. . .	. . .	. ... .	. . .
t_n	xxx	xxx	Ben Johnson	...	xxx

Syntax Violation

	at_1	at_2	**order_date**	...	at_n
t_1	xxx	xxx	24/10/2012	...	xxx
t_2	xxx	xxx	2012/10/26	...	xxx
. . .	. . .	. . .	. . .	. ... .	. . .
t_n	xxx	xxx	27/10/2012	...	xxx

Domain Violation

	at_1	at_2	**ordered_ quantity**	...	at_n
t_1	xxx	xxx	4	...	xxx
t_2	xxx	xxx	-1	...	xxx
. . .	. . .	. . .	. . .	. ... .	. . .
t_n	xxx	xxx	1	...	xxx

Misspelling Error

	at_1	at_2	**city**	...	at_n
t_1	xxx	xxx	New York	...	xxx
t_2	xxx	xxx	Bostom	...	xxx
.	.	.	.	.	.
.	.	.	.	...	.
.	.	.	.	.	.
t_n	xxx	xxx	Washington	...	xxx

Overloaded Attribute

	at_1	at_2	**name**	...	at_n
t_1	xxx	xxx	Bill Clinton	...	xxx
t_2	xxx	xxx	Dr. Barack Obama	...	xxx
. . .	. . .	. . .	. . .	. ... .	. . .
t_n	xxx	xxx	George Bush	...	xxx

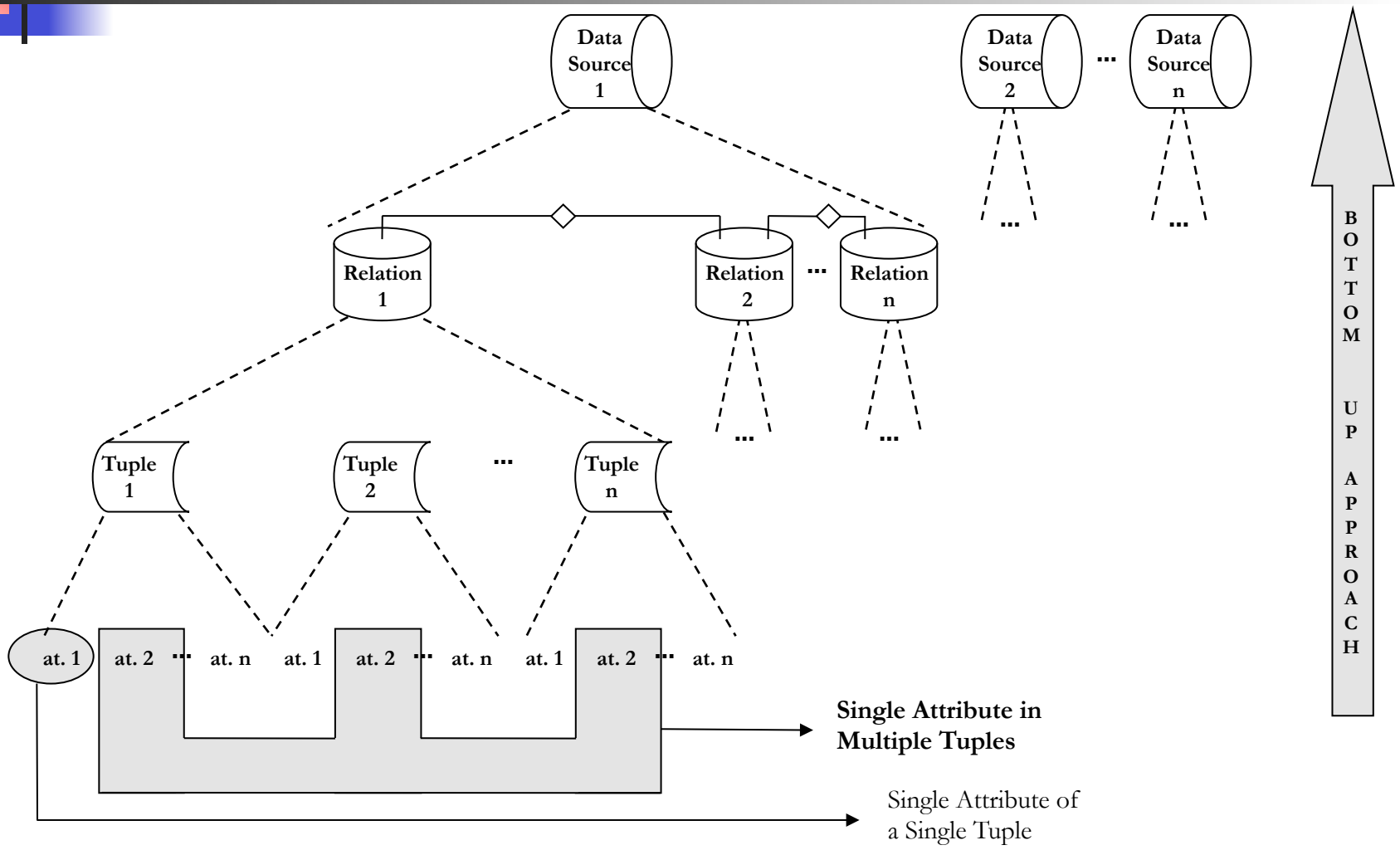
Incomplete Value

	at_1	at_2	**address**	...	at_n
t_1	xxx	xxx	Sun Street, 123	...	xxx
t_2	xxx	xxx	Flowers Street	...	xxx
. . .	. . .	. . .	. . .	. ... .	. . .
t_n	xxx	xxx	Baker Street, 321	...	xxx

Wrong Value

	at_1	at_2	**Marital Status**	...	at_n
t_1	xxx	xxx	Married	...	xxx
t_2	xxx	xxx	Single	...	xxx
.	.	.	.	.	.
.	.	.	.	...	.
.	.	.	.	.	.
t_n	xxx	xxx	Divorced	...	xxx

DQPs in a Single Attribute in Multiple Tuples



Uniqueness Violation

	name	at_2	**taxpayer_nr**	...	at_n
t_1	George Clooney	xxx	196 567 931	...	xxx
t_2	Harrison Ford	xxx	187 323 436	...	xxx
t_3	Brad Pitt	xxx	205 239 894	...	xxx
...	...	...	...	...	...
t_n	Judy Foster	xxx	187 323 436	...	xxx

Existence of Synonyms

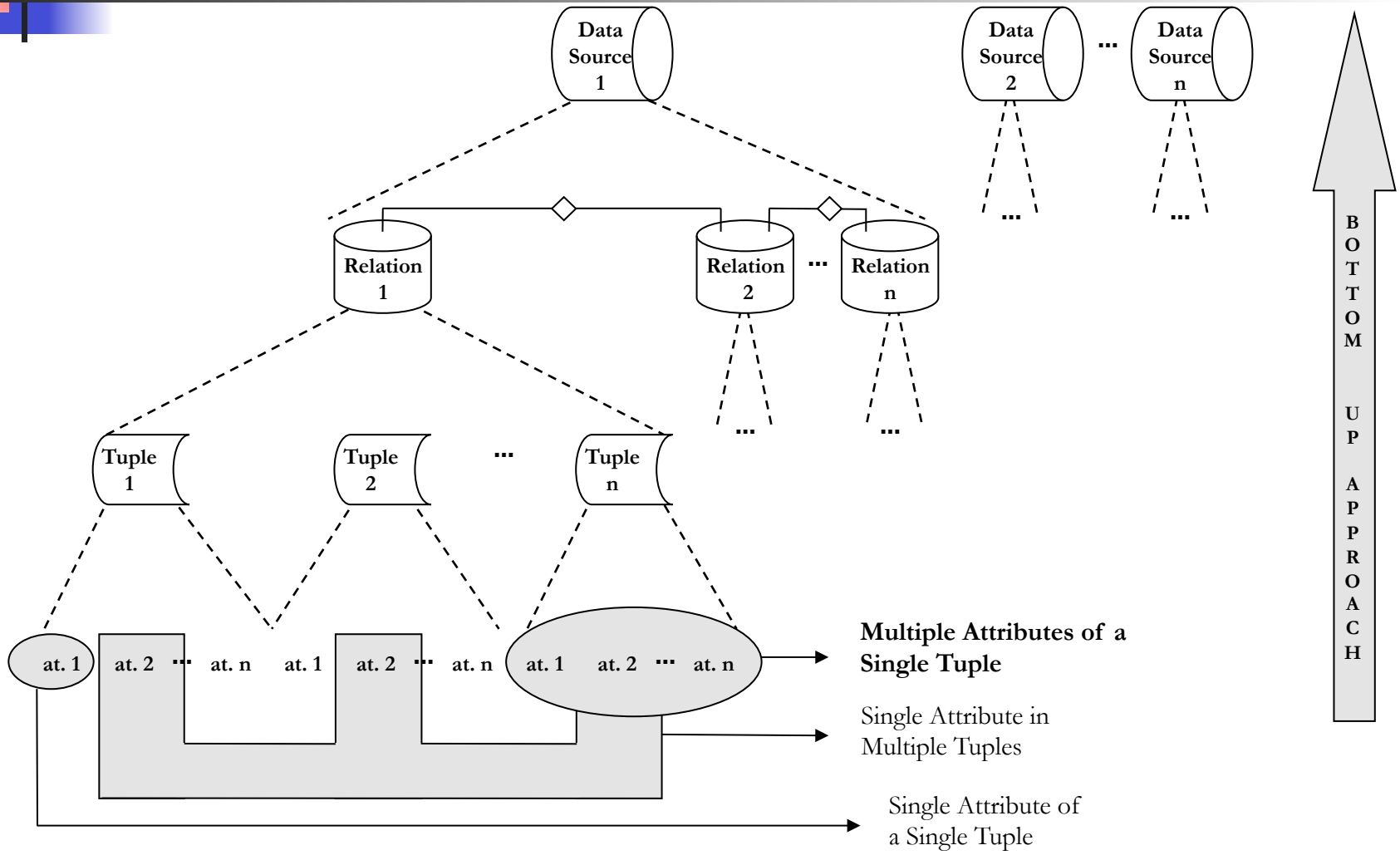
	at_1	at_2	**job**	...	at_n
t_1	xxx	xxx	Researcher	...	xxx
t_2	xxx	xxx	Schoolmaster	...	xxx
t_3	xxx	xxx	Electrician	...	xxx
. . .	. . .	. . .	. . .	. ... .	. . .
t_n	xxx	xxx	Teacher	...	xxx

Violation of Business Rule

	invoice_nr	at_2	**invoice_date**	...	at_n
t_1	20121100	xxx	25/10/2012	...	xxx
t_2	20121101	xxx	24/10/2012	...	xxx
t_3	20121102	xxx	25/10/2012	...	xxx
. . .	. . .	. . .	. . .	. ... .	. . .
t_n	20121178	xxx	05/11/2012	...	xxx

The values of *invoice_date* must appear by ascending order !

DQPs in Multiple Attributes of a Single Tuple

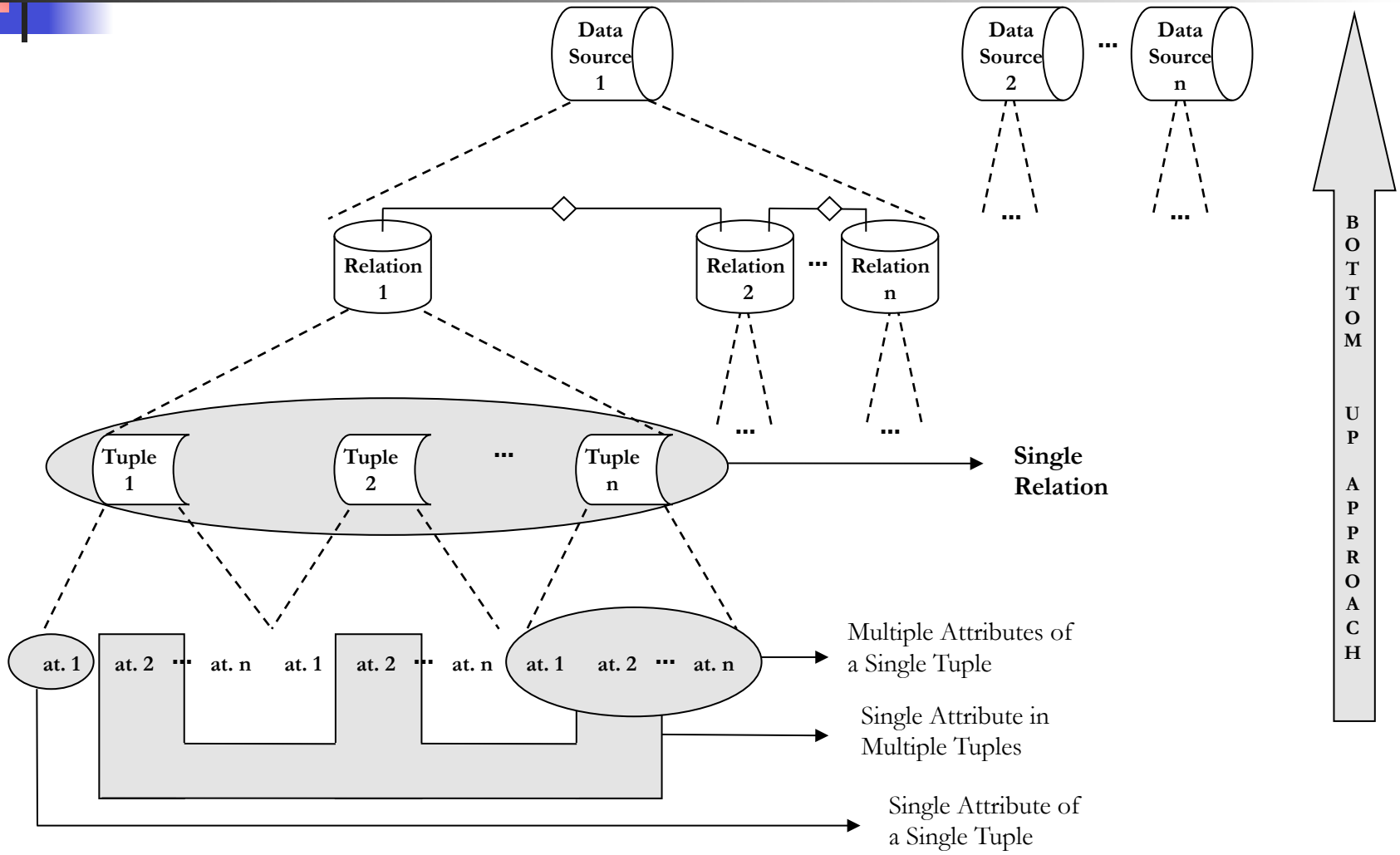


Violation of Business Rule

	at_1	**quantity**	**unit_price**	**total_prod**	...	at_n
t_1	xxx	2	3	6	...	xxx
t_2	xxx	2	5	5	...	xxx
t_3	xxx	3	4	12	...	xxx
. . .	. . .	. . .	. . .		. ... .	. . .
t_n	xxx	2	1	2	...	xxx

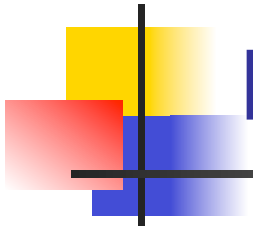
total_prod must be equal to *quantity* * *unit_price* !

DQPs at the Single Relation Level



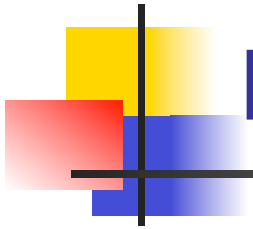
Violation of Functional Dependency

	at_1	**zip_code**	**city**	...	at_n
t_1	xxx	4000	Oporto	...	xxx
t_2	xxx	4000	Lisbon	...	xxx
t_3	xxx	1000	Lisbon	...	xxx
. . .	. . .	. . .	. . .	. ... .	. . .
t_n	xxx	4000	Oporto	...	xxx



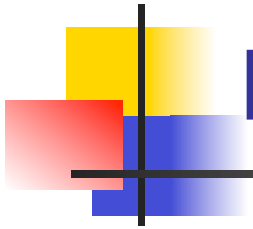
Duplicate Tuples (Equal)

	id	name	address	taxpayer_nr	...	at_n
t_1	xxx	xxx	xxx	xxx	...	xxx
t_2	10	**Cliff Barnes**	**Flowers Street, 123**	**205 239 894**	...	xxx
. . .	. . .	. . .	. . .		. ... .	. . .
t_n	72	**Cliff Barnes**	**Flowers Street, 123**	**205 239 894**	...	xxx



Duplicate Tuples (Approximate)

	id	name	address	taxpayer_nr	...	at_n
t_1	xxx	xxx	xxx	xxx	...	xxx
t_2	10	**Cliff Barnes**	**Flowers Street, 123**	**205 239 894**	...	xxx
. . .	. . .	. . .	. . .		. ... .	. . .
t_n	72	**C. Barnes**	**Flowers St., 123**	**205 239 894**	...	xxx



Duplicate Tuples (Inconsistent)

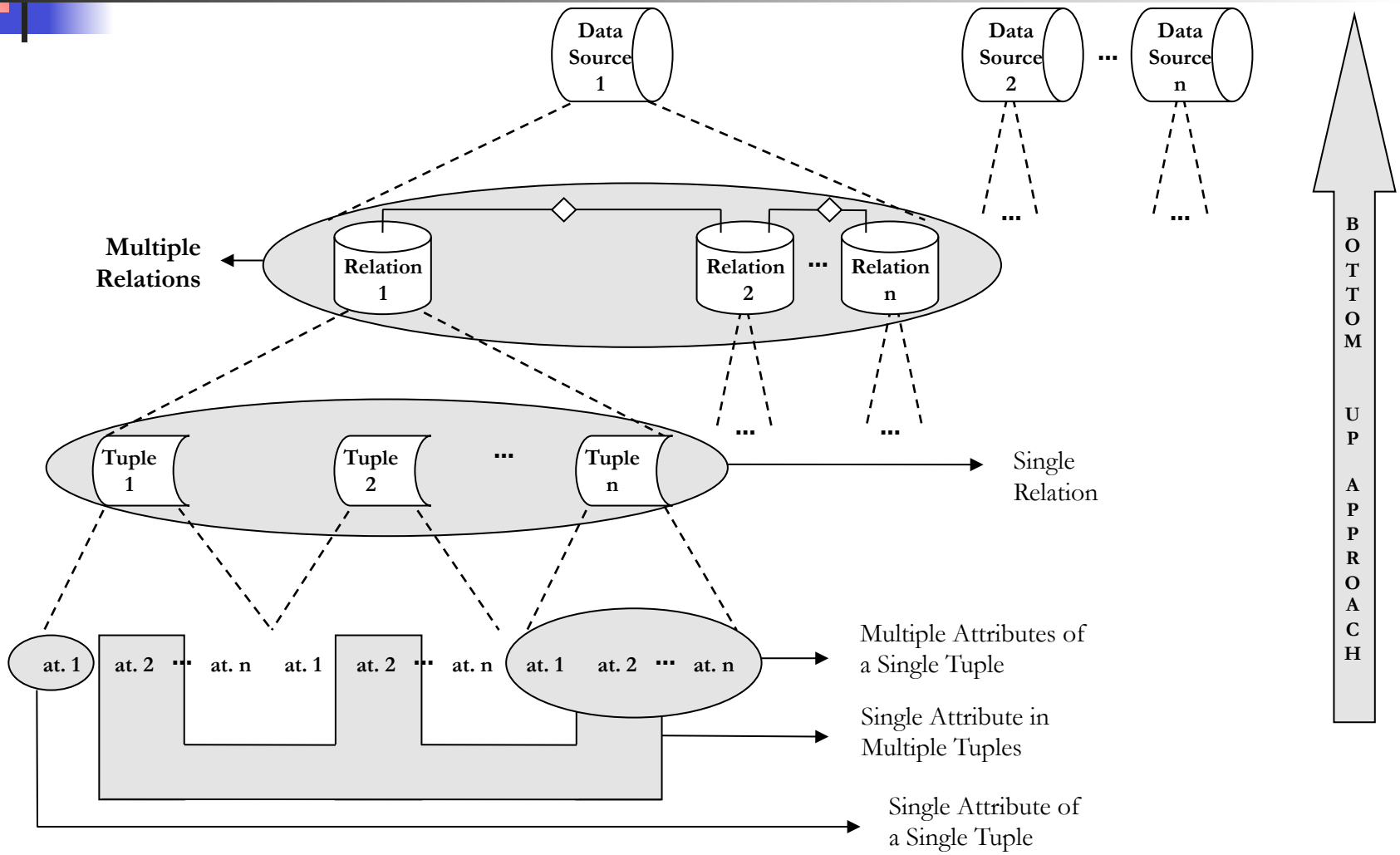
	id	name	address	taxpayer_nr	...	at_n
t_1	xxx	xxx	xxx	xxx	...	xxx
t_2	10	Cliff Barnes	**Flowers Street, 123**	205 239 894	...	xxx
. . .	. . .	. . .	. . .		. ... .	. . .
t_n	72	Cliff Barnes	**Sun Street, 321**	205 239 894	...	xxx

Violation of Business Rule

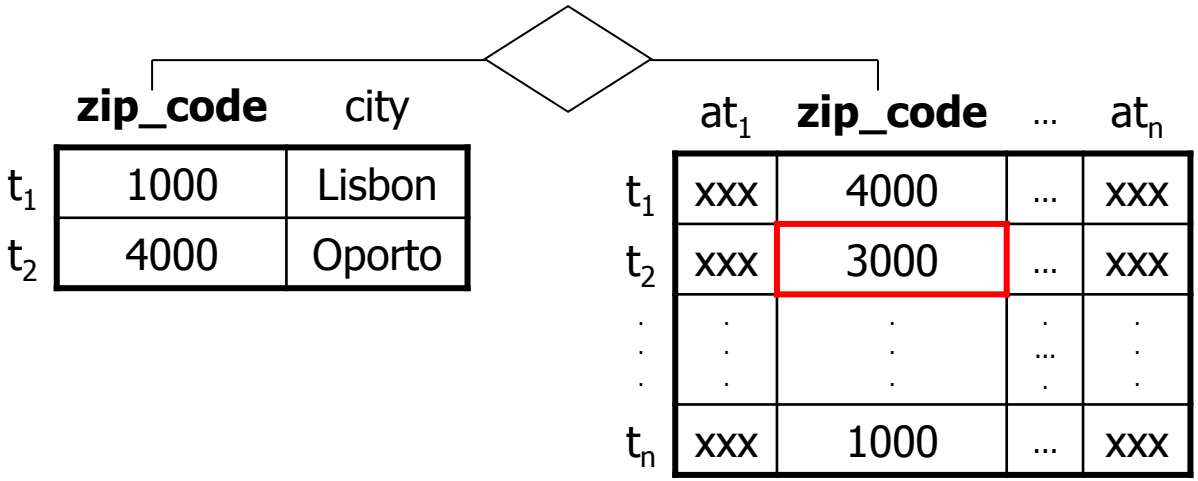
	at_1	at_2	...	at_n
t_1	xxx	xxx	...	xxx
t_2	xxx	xxx	...	xxx
.	.	.	.	.
.	.	.	...	.
.	.	.	.	.
t_{12}	xxx	xxx	...	xxx

The number of product families (tuples) must not be superior to 10!

DQPs at the Level of Multiple Relations



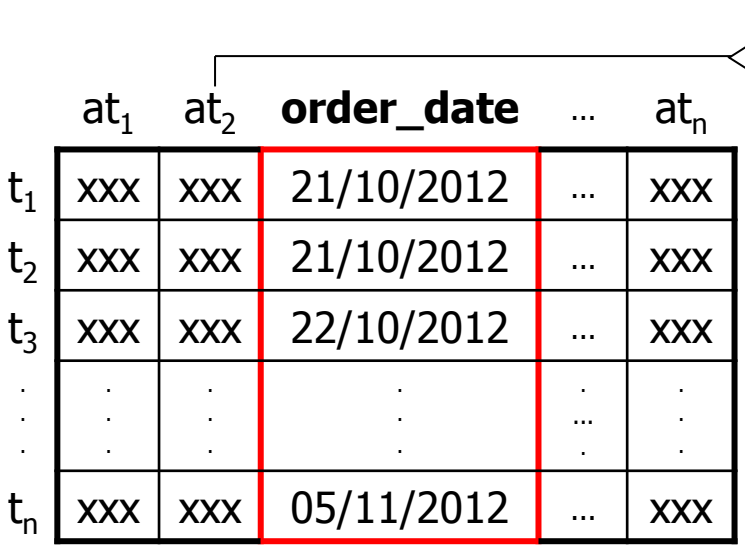
Referential Integrity Violation



	zip_code	city
t_1	1000	Lisbon
t_2	4000	Oporto

	at_1	**zip_code**	...	at_n
t_1	xxx	4000	...	xxx
t_2	xxx	3000	...	xxx
. . .	. . .	. . .	. ... .	. . .
t_n	xxx	1000	...	xxx

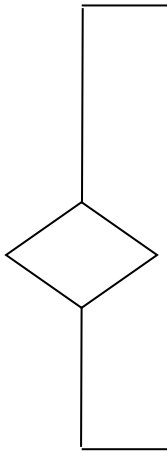
Heterogeneity of Syntaxes



	at_1	at_2	**order_date**	...	at_n
t_1	xxx	xxx	21/10/2012	...	xxx
t_2	xxx	xxx	21/10/2012	...	xxx
t_3	xxx	xxx	22/10/2012	...	xxx
...	...	...	...	...	...
t_n	xxx	xxx	05/11/2012	...	xxx

	at_1	at_2	**invoice_date**	...	at_n
t_1	xxx	xxx	2012/10/24	...	xxx
t_2	xxx	xxx	2012/10/27	...	xxx
...	...	...	...	...	...
t_n	xxx	xxx	2012/10/30	...	xxx

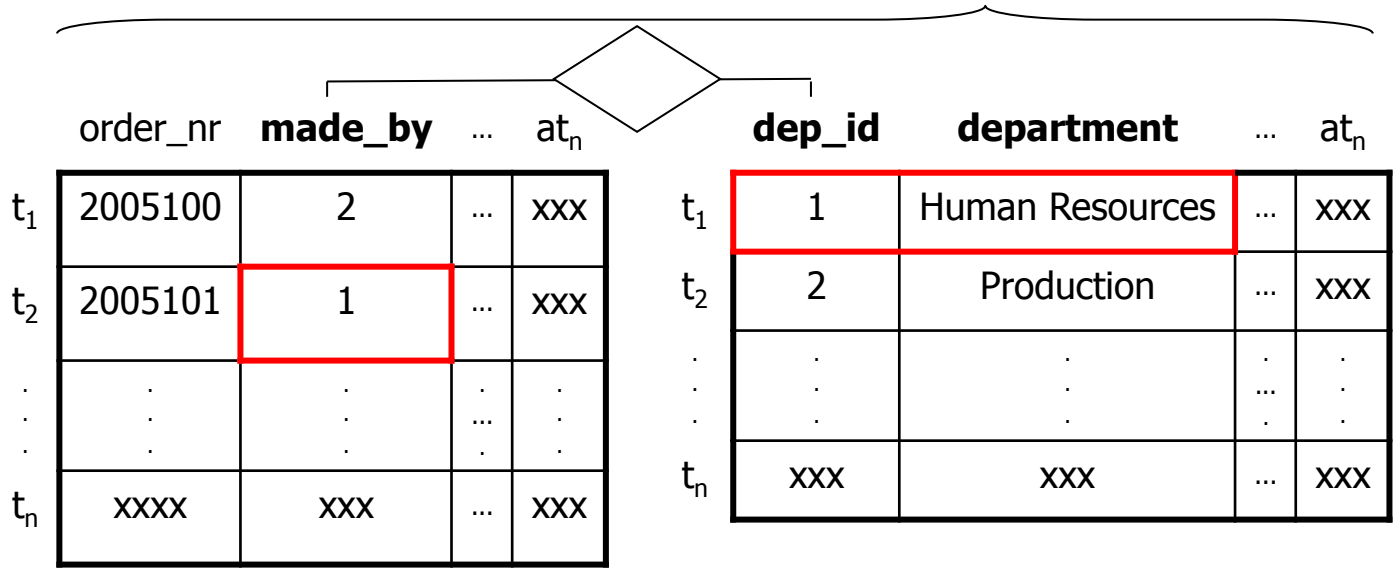
Circularity among Tuples in a Self-Relationship



	product_code	at_2	**sub_product_code**	...	at_n
t_1	X10	xxx	X40	...	xxx
t_2	X40	xxx	X10	...	xxx
. . .	. . .	. . .	. . .	. ... .	. . .
t_n	Y30	xxx	X10	...	xxx

Violation of Business Rule

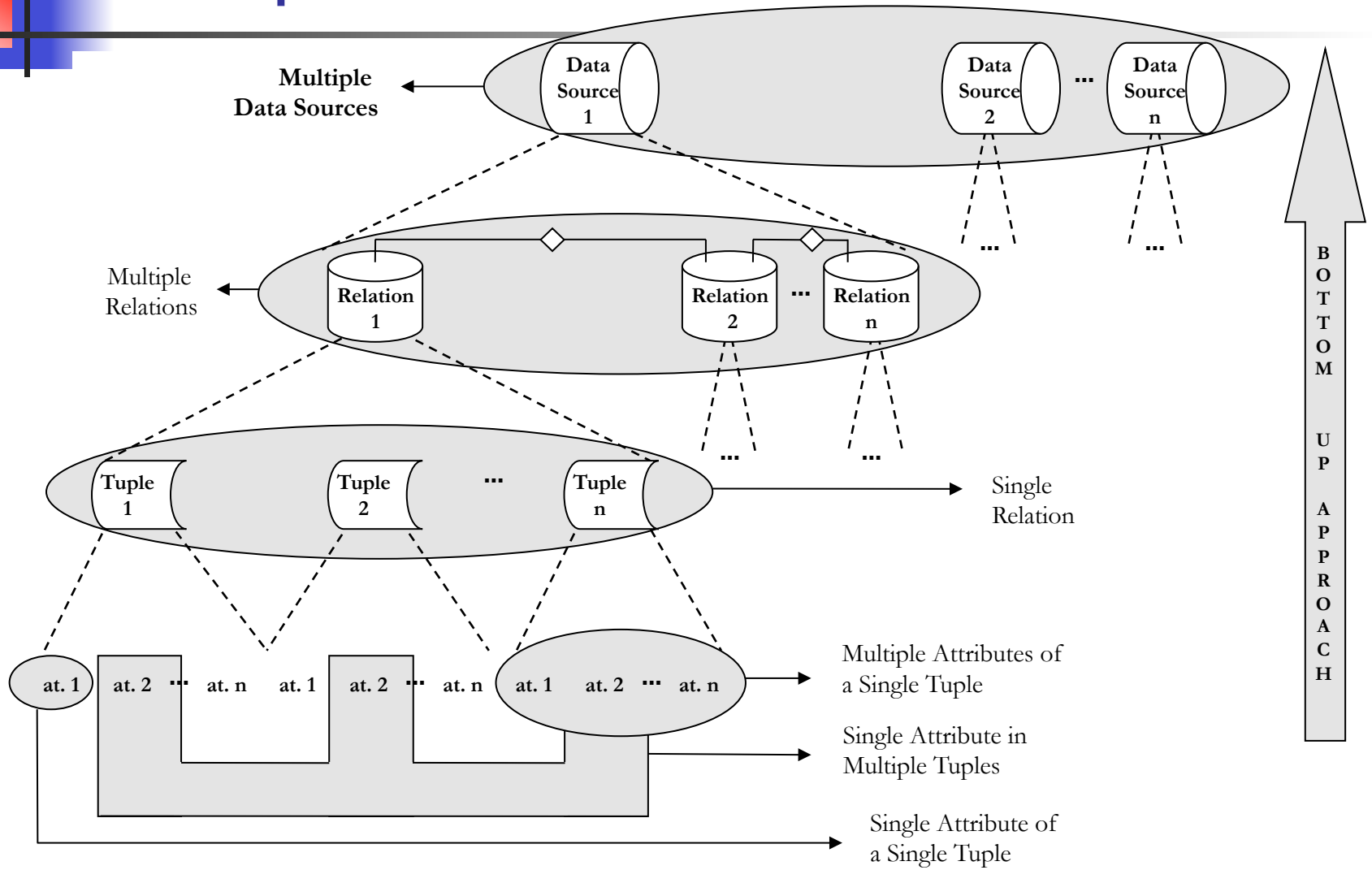
Orders can only be made by the provisions department or production department !



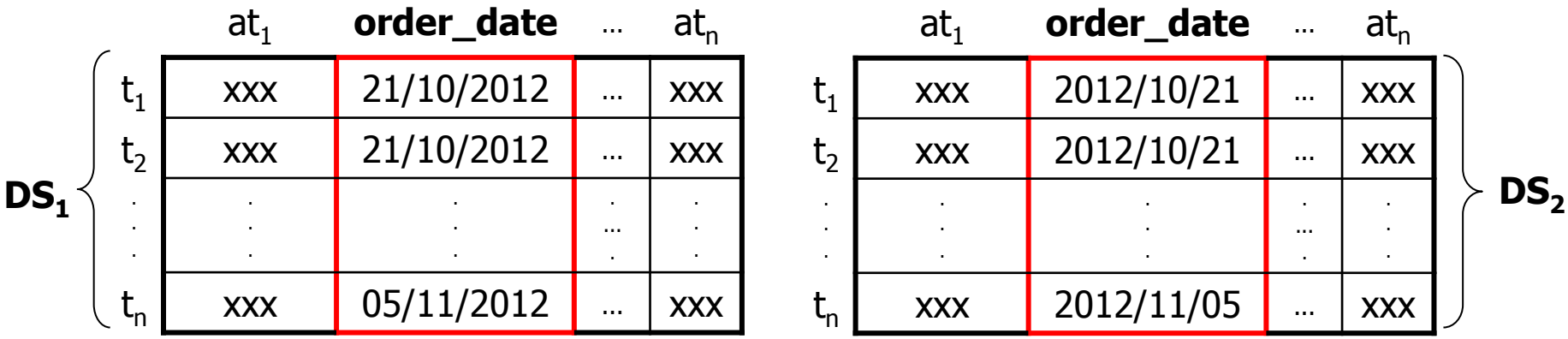
	order_nr	**made_by**	...	at_n
t_1	2005100	2	...	xxx
t_2	2005101	1	...	xxx
. . .	. . .	. . .	. ... .	. . .
t_n	xxxx	xxx	...	xxx

	dep_id	**department**	...	at_n
t_1	1	Human Resources	...	xxx
t_2	2	Production	...	xxx
. . .	. . .	. . .	. ... .	. . .
t_n	xxx	xxx	...	xxx

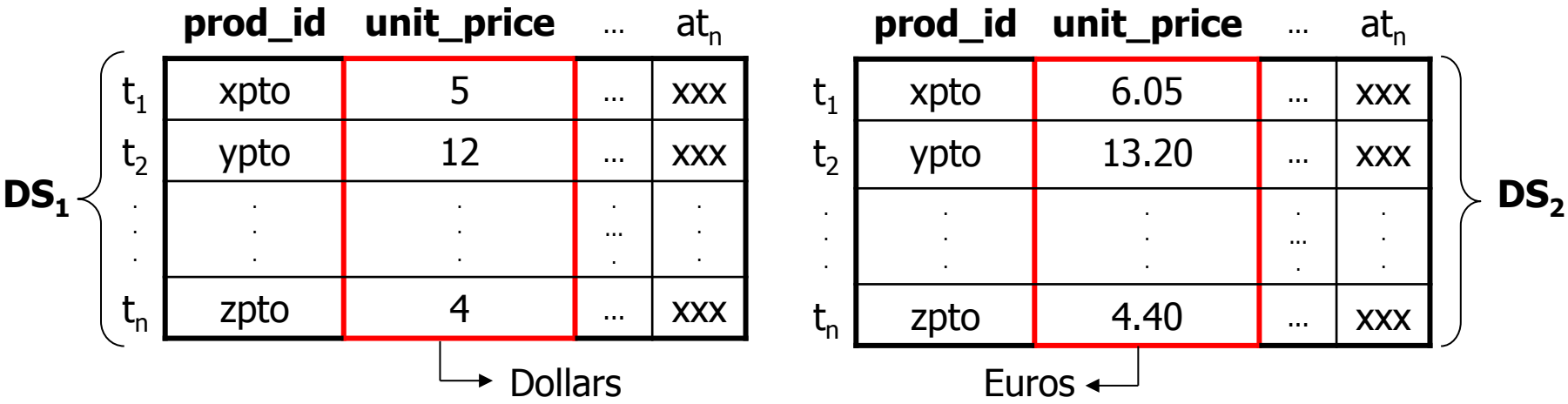
DQPs at the Level of Multiple Data Sources



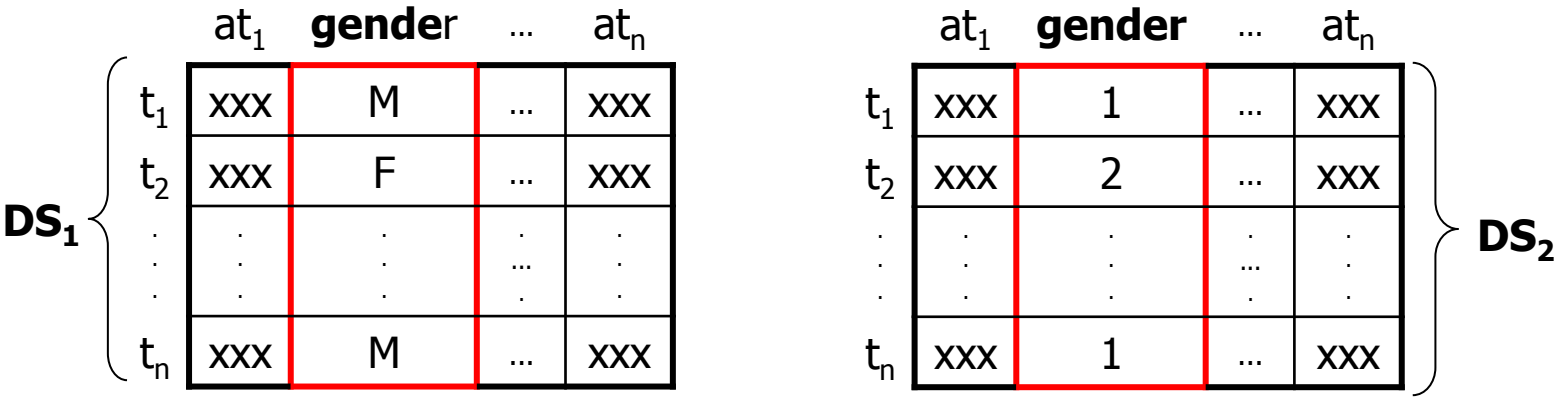
Heterogeneity of Syntaxes



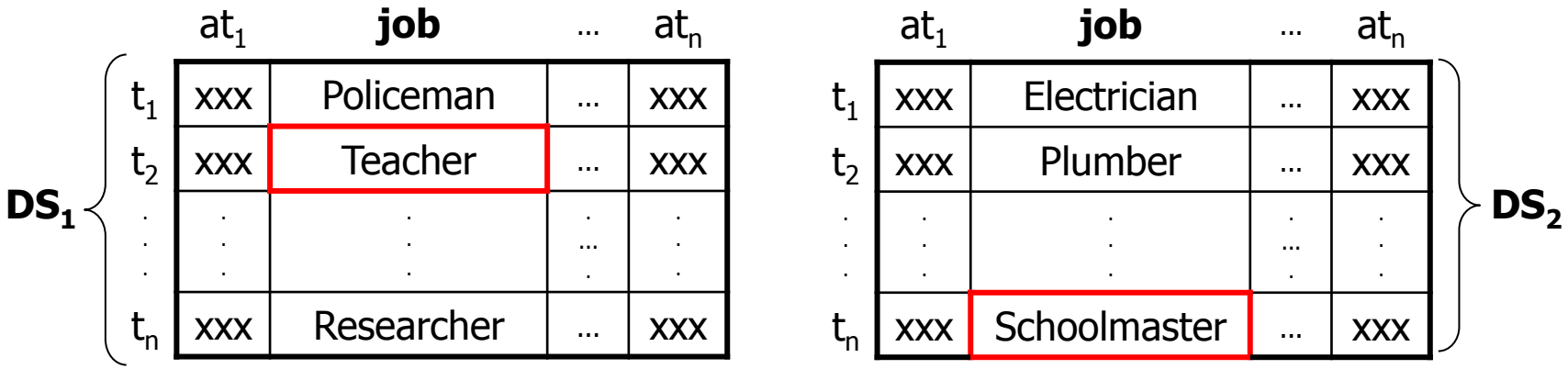
Heterogeneity of Measure Units



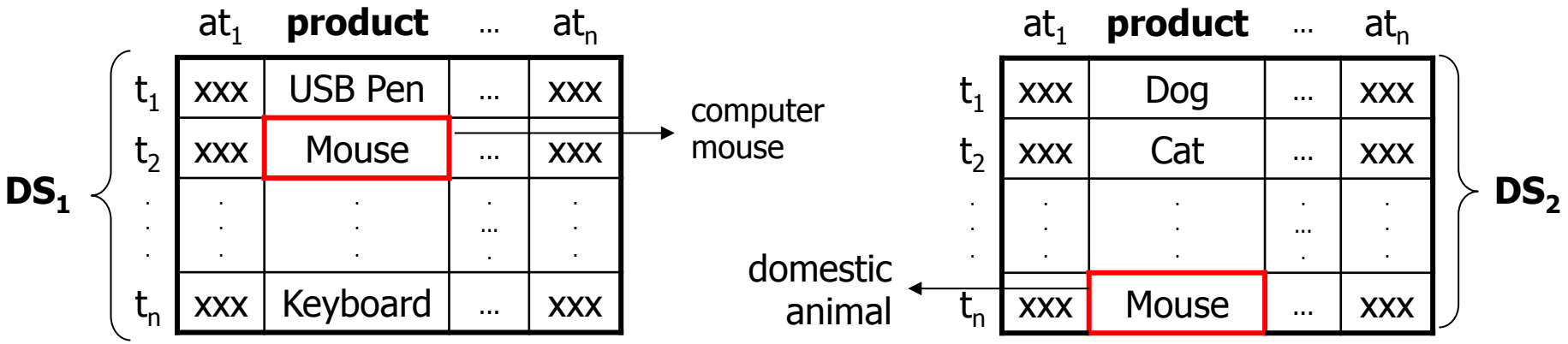
Heterogeneity of Domains



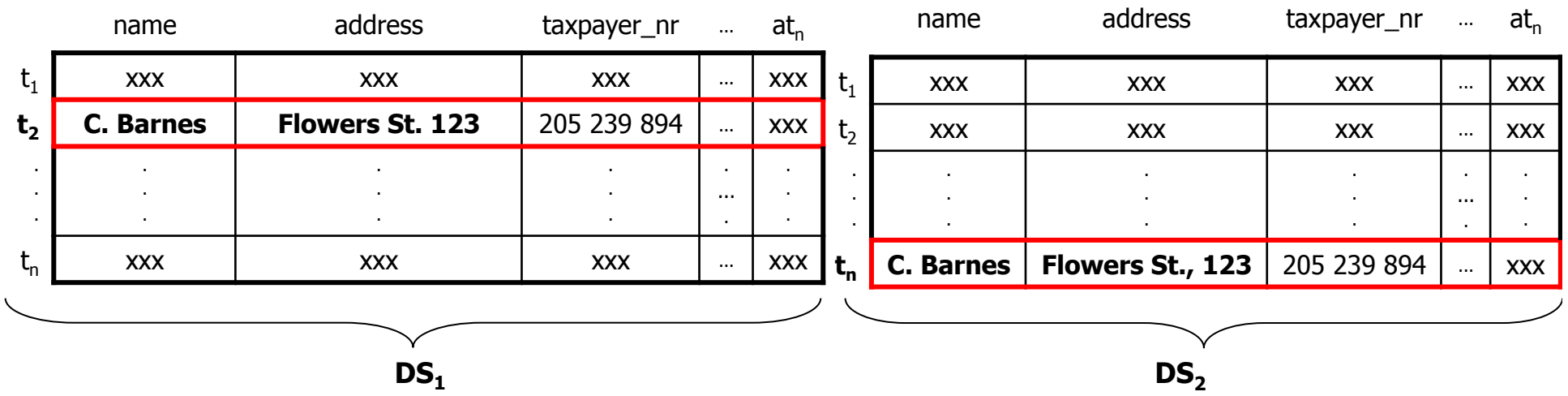
Existence of Synonyms



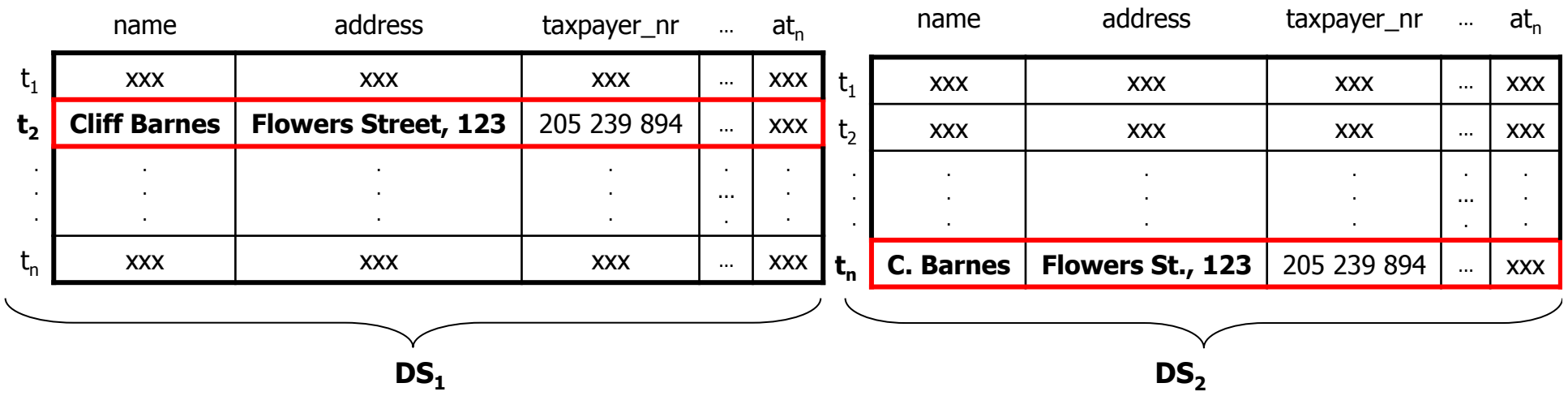
Existence of Homonyms



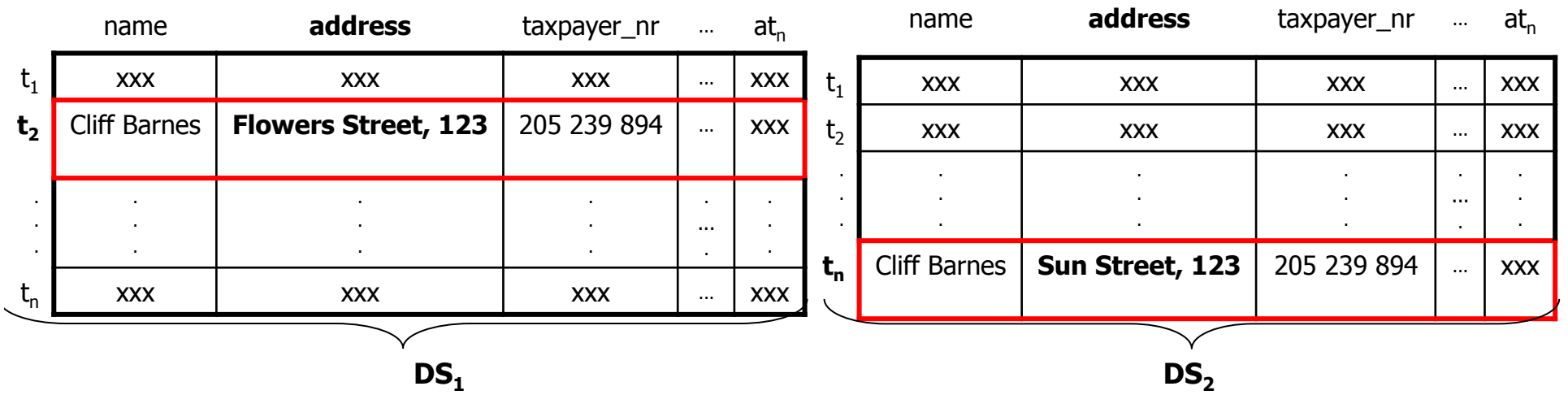
Duplicate Tuples (Equal)



Duplicate Tuples (Approximate)

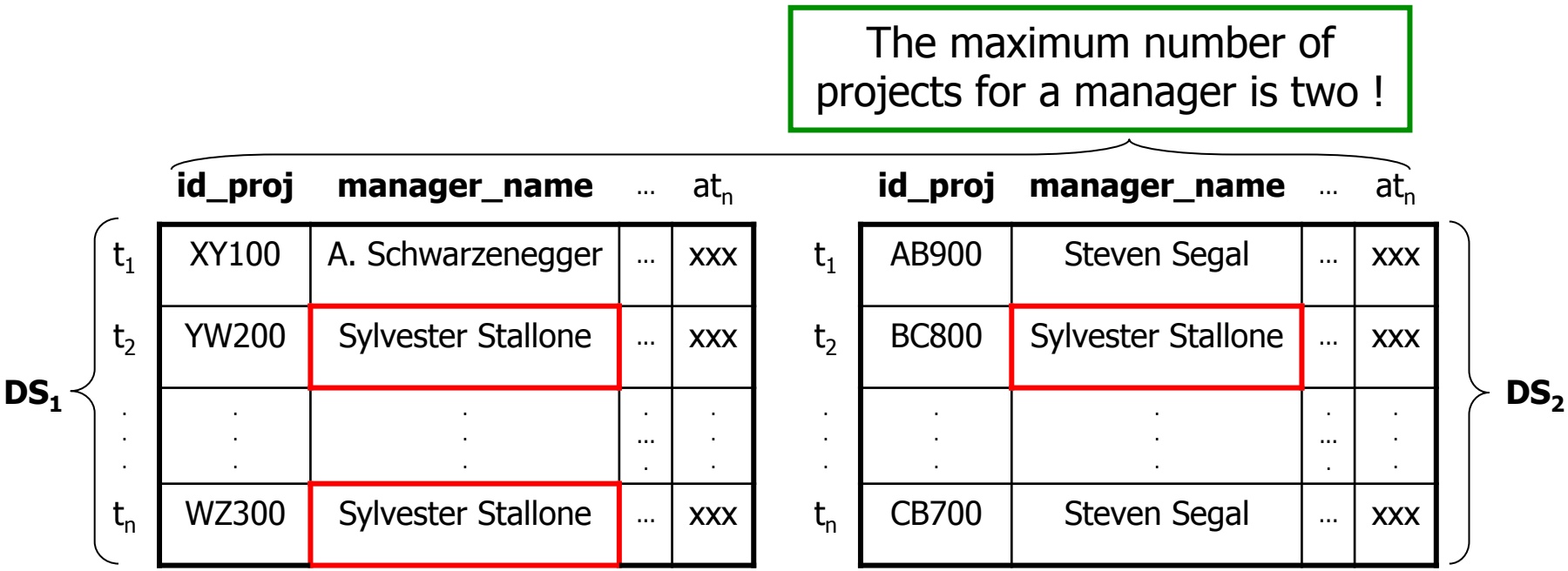


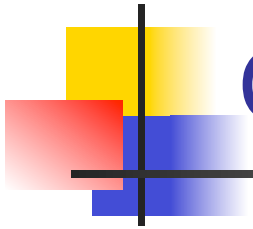
Duplicate Tuples (Inconsistent)



Violation of Business Rule

The maximum number of projects for a manager is two !





Conclusion

- Taxonomy of DQPs is useful to:
  - Alert for the existence of many DQPs
  - Assess the coverage of existing DQ tools

Data Quality Problems (DQPs) at the Instance Level

Paulo Oliveira

isep Instituto Superior de Engenharia do Porto

gecad
Grupo de Investigação em Engenharia do Conhecimento e Apoio à Decisão

Toward Efficient Portuguese and Brazilian Electricity Markets Workshop

September 25, 2013

Data Quality Problems (DQPs) at the Instance Level

Paulo Oliveira, Fátima Rodrigues and Pedro Henriques – "A Formal Definition of Data Quality Problems". In *Proceedings of the 10th International Conference on Information Quality*, MIT, Boston, EUA, November of 2005. p. 13-26.