Laboratoire d'Informatique et d'Automatique pour les Systèmes

Value-driven Approach Designing Extended Data Warehouses

Nabila BERKANI & Selma KHOURI

ESI
Algiers, Algeria
(n_berkani, s_khouri)@esi.dz

Ladjel BELLATRECHE

LIAS/ISAE-ENSMA
Poitiers, France
bellatreche@ensma.fr

Carlos ORDONEZ

University of Houston
USA
carlos@central.uh.edu

DOLAP'2019, Lisbon, Marsh 26, 2019

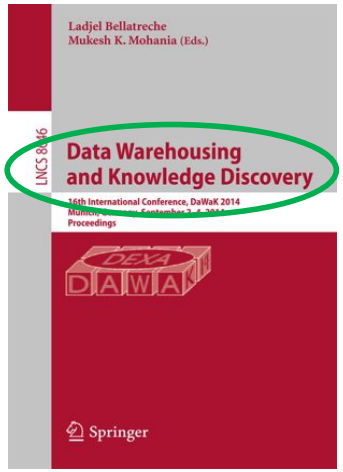
ENSi Poitiers – INGÉNIERIE pour la PROTECTION de L'ENVIRONNEMENT

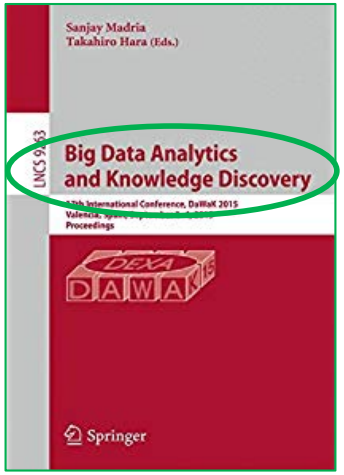
ISAE – Institut Supérieur de l'Aéronautique et de l'Espace – ENSMA

Université de Poitiers – 1431

Impact of Big Data on DW

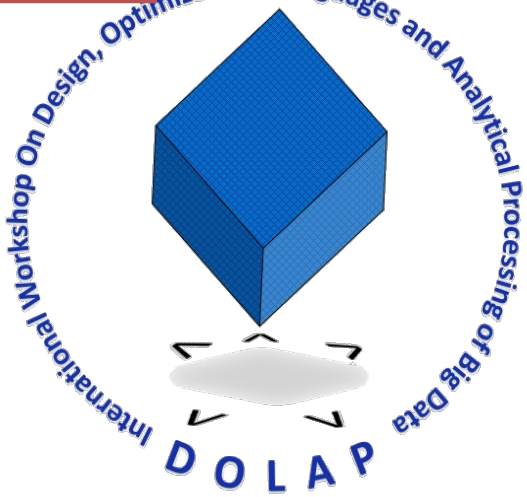
DaWak Conference





1999-2014 ⟶ 2015-Present

DOLAP Workshop

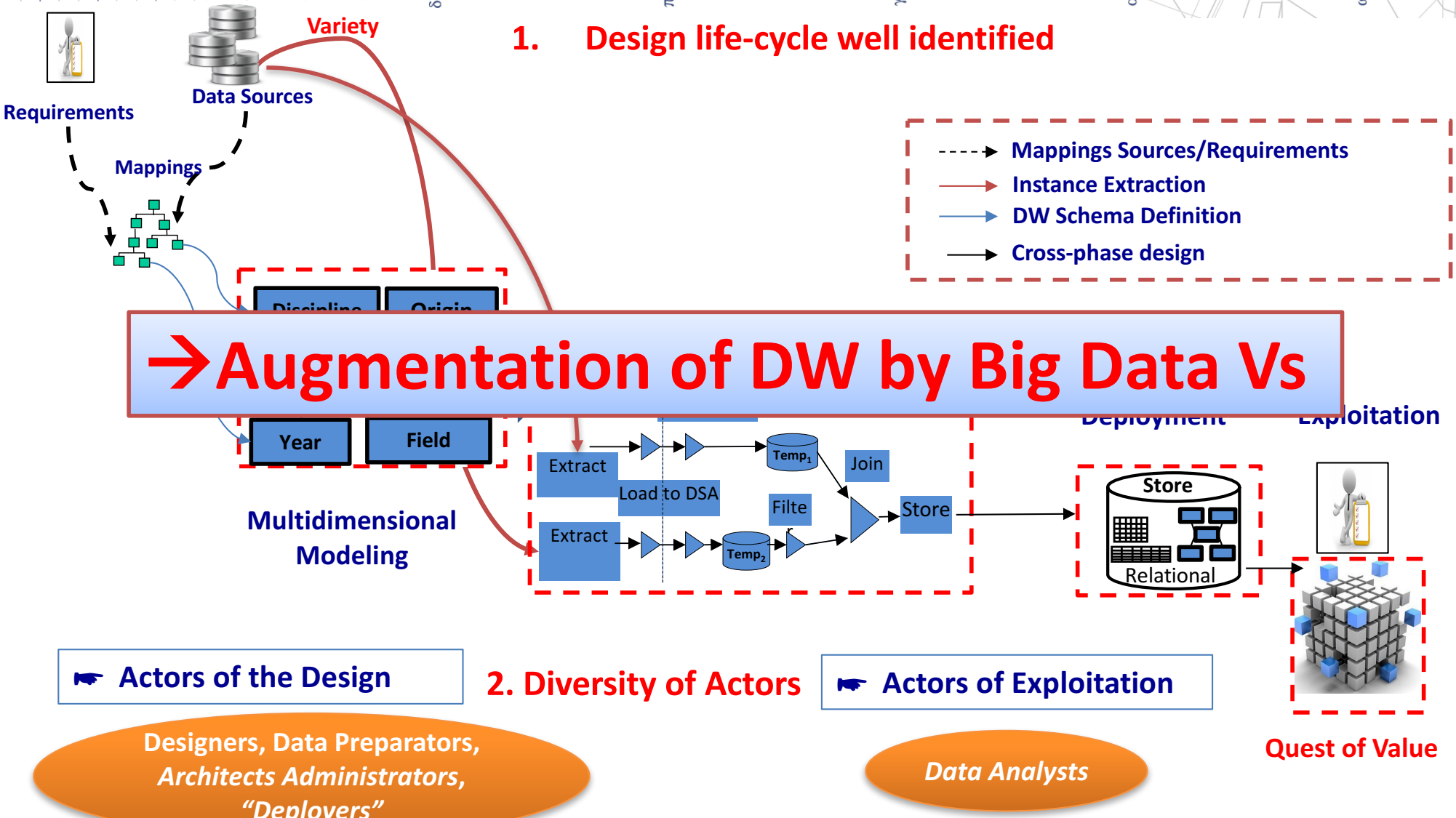


1998-2015 | 2016: Thinking | 2017-Present

New Tendency: Big Data Analytics

- **Teaching**
  - Master IT4BI → Big Data Management and Analytics (BDMA)
- **Storage Systems and Tools**
  - Hadoop, MongoDB, MapReduce, Spark, BigML, R.

30 years of existence: Maturity



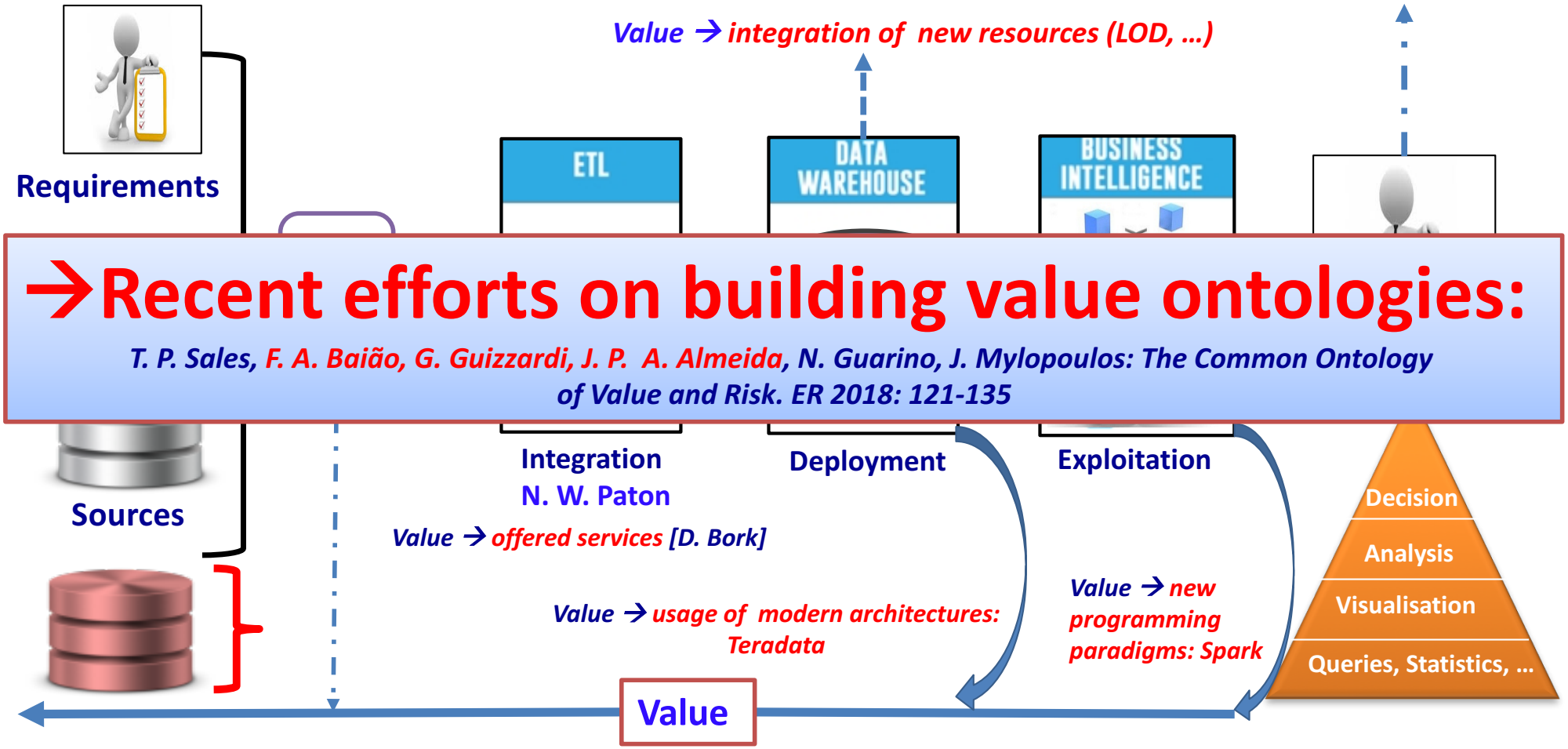
Agenda

- Value & Variety (2Vs)
- Augmenting DW by Linked Open Data
- 2Vs-driven Design Approach
- Case Study
- Summary

Value: # places

- *FR: Value → **money** [A. G. Sutcliffe'2018]*
- *NfR: Value → **satisfaction of qualities** (security, privacy, …)*

Value → user feedback:
N. Konstantinou



→Recent efforts on building value ontologies:

T. P. Sales, F. A. Baião, G. Guizzardi, J. P. A. Almeida, N. Guarino, J. Mylopoulos: The Common Ontology of Value and Risk. ER 2018: 121-135

→interdependencies between value (phases) & value (operational DW)

Value increases Variety

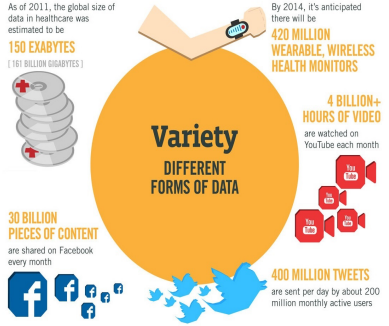
Person in Charge (PiC) of Value

VALEUR

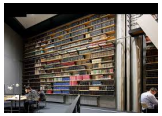
- Examples of requirements related to value[1] :
  - ***Media: Has the coverage of media changed over time?***
  - ***Politics: Speeches EU parliament that contain « human rights » by country***
  - ***Finance: Evolution of Debates related to Greece crisis by country***
- Measurement of the value depends on the studied domain

→ Interaction between designers and PiC of value: **multidisciplinary** in DW

Designer



DB



Libraries

News papers

Internal Sources + External Sources

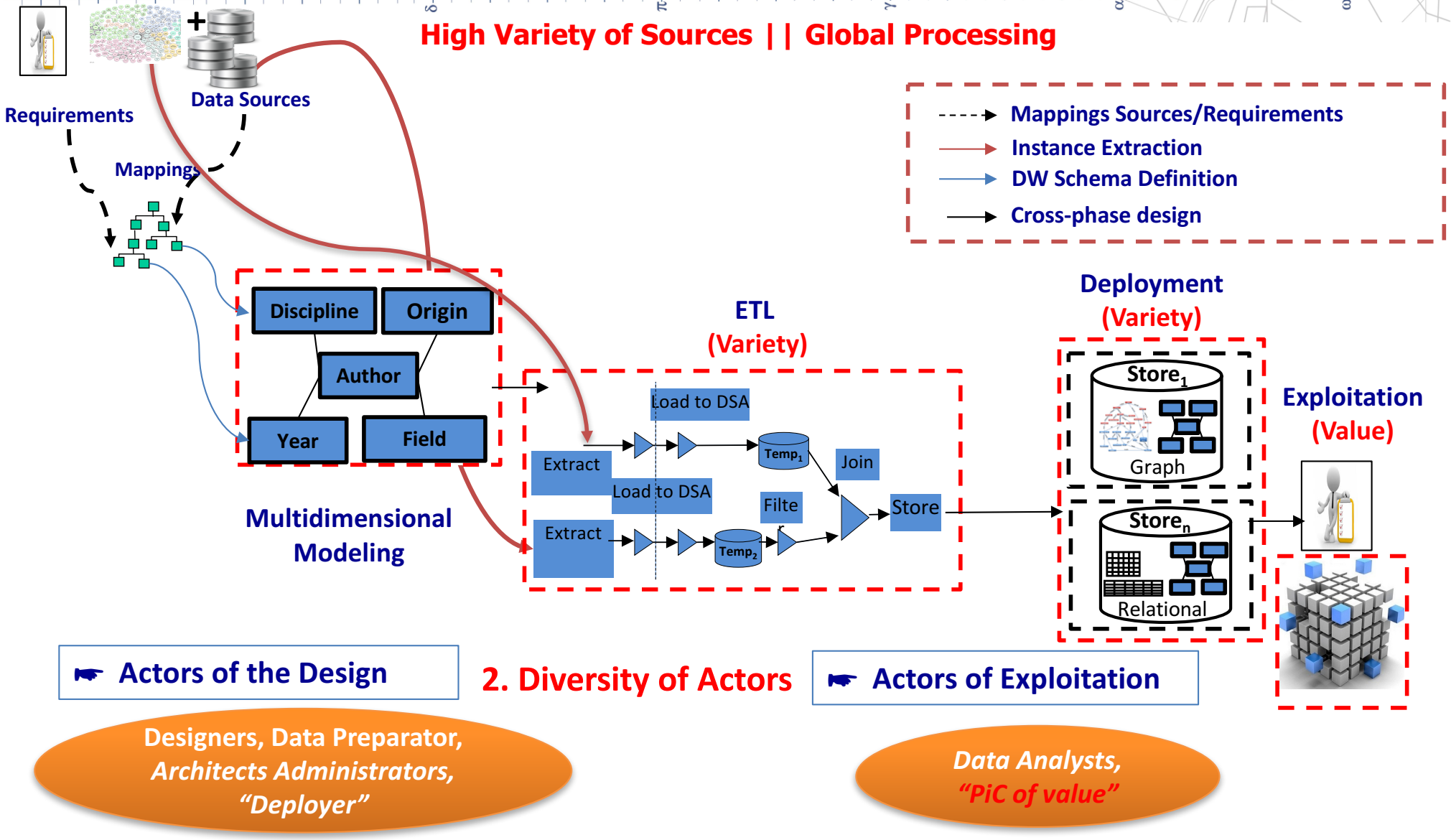
→ Usage of Linked Open Data :

- Traditional Management of Variety
  +
- Variety of Formalisms (graphs)

[1] http://www.talkofeurope.eu/data

Augmenting DW by 2Vs

- Variety & Value
- LOD & DW
- 2Vs Design Approach
- Case Study
- Summary



Formalisation

Inputs:

1. Set of internal sources: $S_{Int} = \{S_{I1}, S_{I2}, \ldots, S_{Im}\}$
2. Set of external resources: $S_{Ext} = = \{S_{E1}, S_{E2}, \ldots, S_{En}\}$
3. Each source (internal/external) S_i has:
   - Its own physical format (F_i)
   - Its conceptual model CM_i
   - Is related to a discipline D (medicine, engineering, etc.)
4. Set of requirements to be satisfied
5. [Optional]: An **operational** DW (**[Ravat et al. 2017]**), where:
   - Its conceptual model CM_{DW}
   - Its format(s) Format (S_{DW}) = $\{f_1, f_2, \ldots, f_k\}$ → **polystore storage**

Objective:

- Definition of all phases of DW augmenting its value

Challenges:

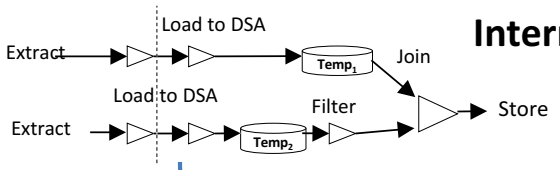
- Metrics of Value

  Value(DW)= **Operator**$_{(1 \leq i \leq n+m)}$ [Weight(S_i , D) * Value (S_i)]; $S_i \in S_{int} \cup S_{ext}$ [Ballou et al.]*

*Ballou, D. P., & Tayi, G. K. (1999). Enhancing data quality in data warehouse environments. Com *of the ACM*, *42*(1), 73-78

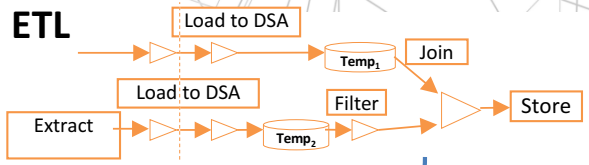
Scenarios

- Variety & Value
- LOD & DW
- 2Vs Design Approach
- Case Study
- Summary

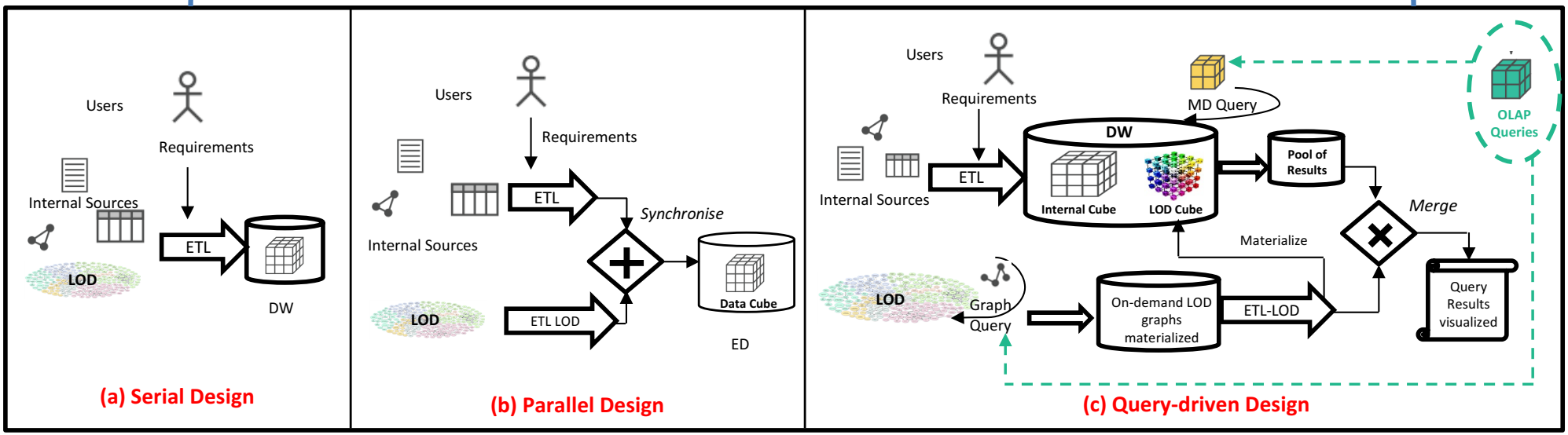


Internal ETL

External ETL



3 Scenarios



☛ LOD is seen as source

☛ Two Parallel ETL

☛ On-demand ETL: data extracted from the existing DW and LOD, then potentially loaded into DW → (requirement satisfaction)

Challenges?

1. **Pivot schema** : generic schema vs. LOD schema (graph)
2. **Redefinition** of operators (overloading)
3. **Synchronisation** of internal and external data: 3 scenarios

Value Metrics

Value(DW) = Operator$_{(1\leq i\leq n+m)}$ [weight(S_i, D) * Value(S_i)], where $S_i \in S_{int} \cup S_{Ext}$

❑ Three metrics related to:

1. Requirement satisfaction

$$Value(Req, S_i) = \frac{number\ of\ responses\ of\ requirement\ on\ S_i}{number\ of\ responses\ of\ all\ requirements}$$

2. Conceptual modelling (multidimensional concepts)

$$Value(Concepts, S_i) = \frac{number\ of\ concepts\ of\ DW\ schema\ by\ integrating\ S_i}{total\ number\ of\ the\ DW\ concepts}$$

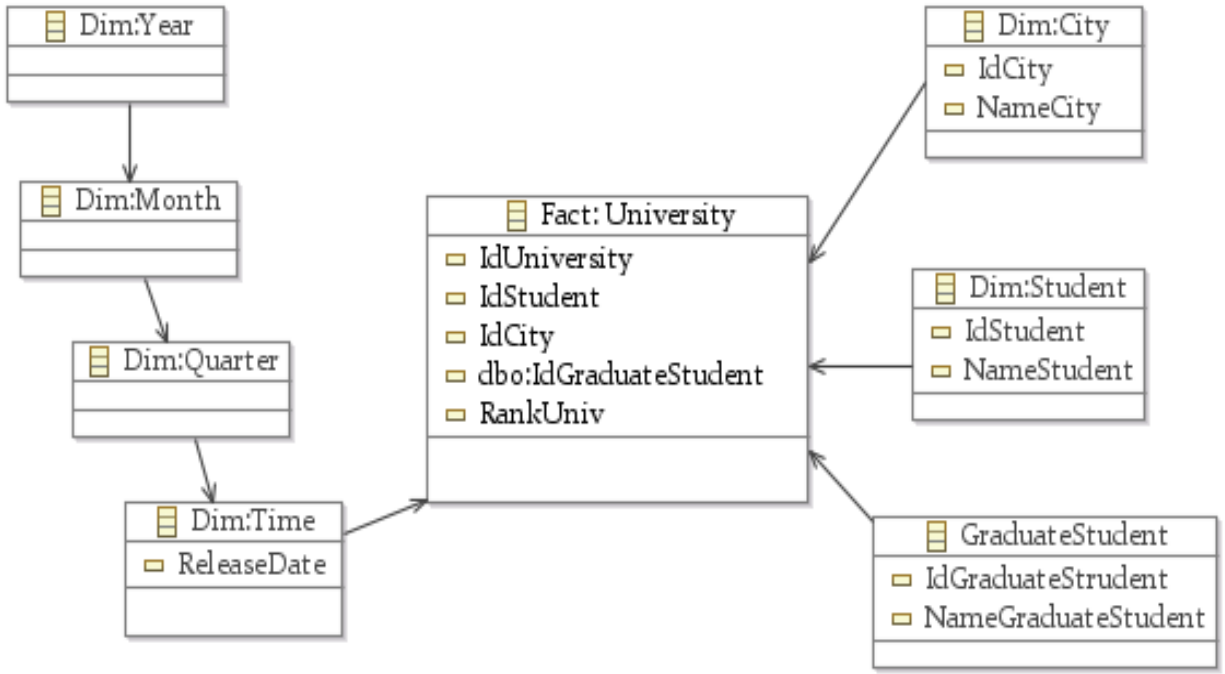
3. Target DW population

$$Value(Instances, Si) = \frac{number\ of\ instances\ of\ DW\ by\ integrating\ S_i}{total\ number\ of\ instances\ of\ the\ DW}$$

Case Study

☛ University Research Analysis

- 4 internal sources generated from LUBM benchmark
- 15 initial requirements



❑ **Analysis:**

- 6 requirements are not satisfied by internal sources (Oracle 12c release 1)

→External source: Dbpedia

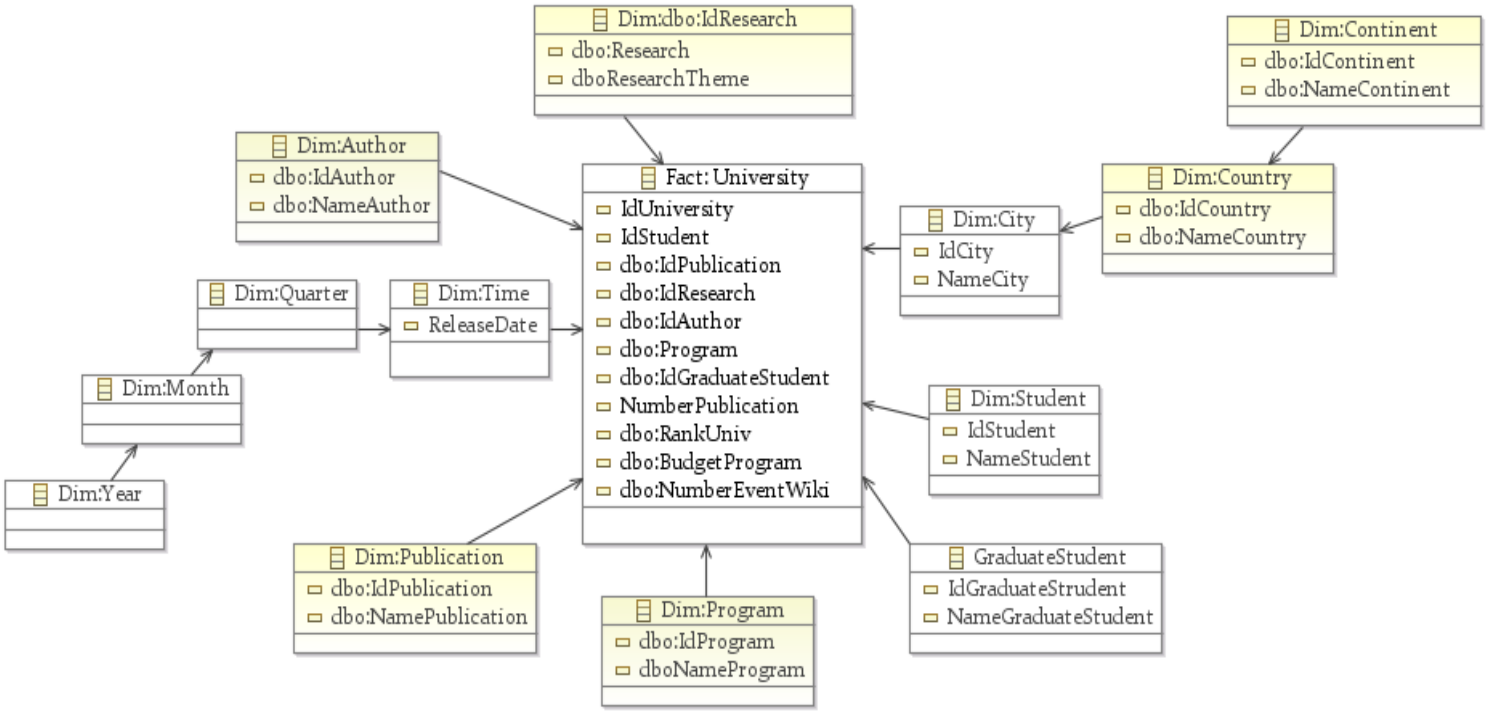
Experiments

- Variety & Value
- LOD & DW
- 2Vs Design Approach
- Case Study
- Summary

Metrics / Sources	Dimensions/ Measures	Value $(S^*)_{MD}$	Value$(S^*)_{Req.}$	Value $(S^*)_{Instances}$	Instances	Response time
Internal Sources	6/1	31%	6%	10%	550K	1.1
Serial Design	10/7	71%	80%	94%	$7{,}7x10^6$	3.2
Parallel Design	11/8	73%	84%	85%	$3{,}1x10^6$	2.6
Query-driven design	12/8	74%	96%	84%	$2{,}9x10^6$	1.7

*All sources have the same weight

*Operator: Avg



Summary

✓2Vs for the DW renaissance

✓Value = pool of multidisciplinary expertise

✓DW life cycle design revisited (new formalization)

✓3 augmented scenarios

☛Veracity & 2V

☛More automation (query rewriting)

☛Value Query Language (Thank Patrick)

Special issue on: Business Intelligence and Analytics for Value Creation in the Era of Big Data and Linked Open Data: International Journal of Information Management, Elsevier (Q1; IF=4.810)