Context-aware Business Intelligence

Rafael Berlanga
Victoria Nebot

Temporal Knowledge Bases Group
Universitat Jaume I de Castellón

Overall Summary

- 1st PART
  - Definitions and Concepts
  - Review of context-aware BI
  - Future work
- 2nd PART
  - Use Case and Demos

Context-aware Business Intelligence

DEFINITIONS AND CONCEPTS

Business Intelligence

- BI is *"the process of **collecting** business data and turning it into information that is meaningful and **actionable** towards a strategic goal"*
- BI technology *"aims at gathering, transforming and **summarizing** available business data from available sources to generate **analytical information** suitable for decision-making tasks"*

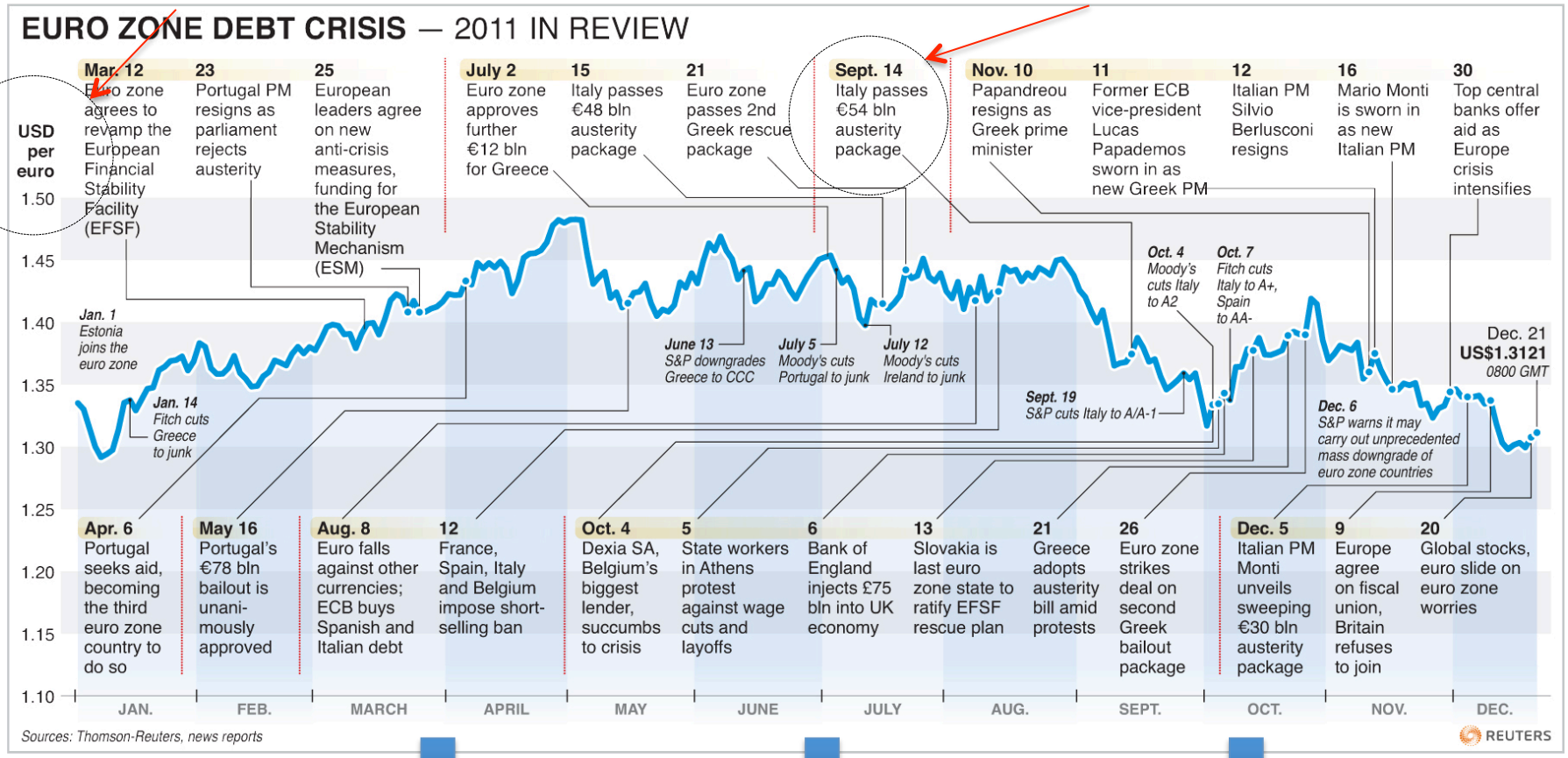
Context-awareness BI

- The context of a BI system comprises **all external events and facts** that could affect somehow the business indicators.
- Proper monitoring of the external context provide direct actionable information:
  - Early detection of potential **threats**: reputation attacks, fraud detection, unfavourable law changes, cyber-attacks, etc.
  - Identification of new business **opportunities**: favourable law changes, new marked trends, etc.

Context-aware BI

Business indicator

External event

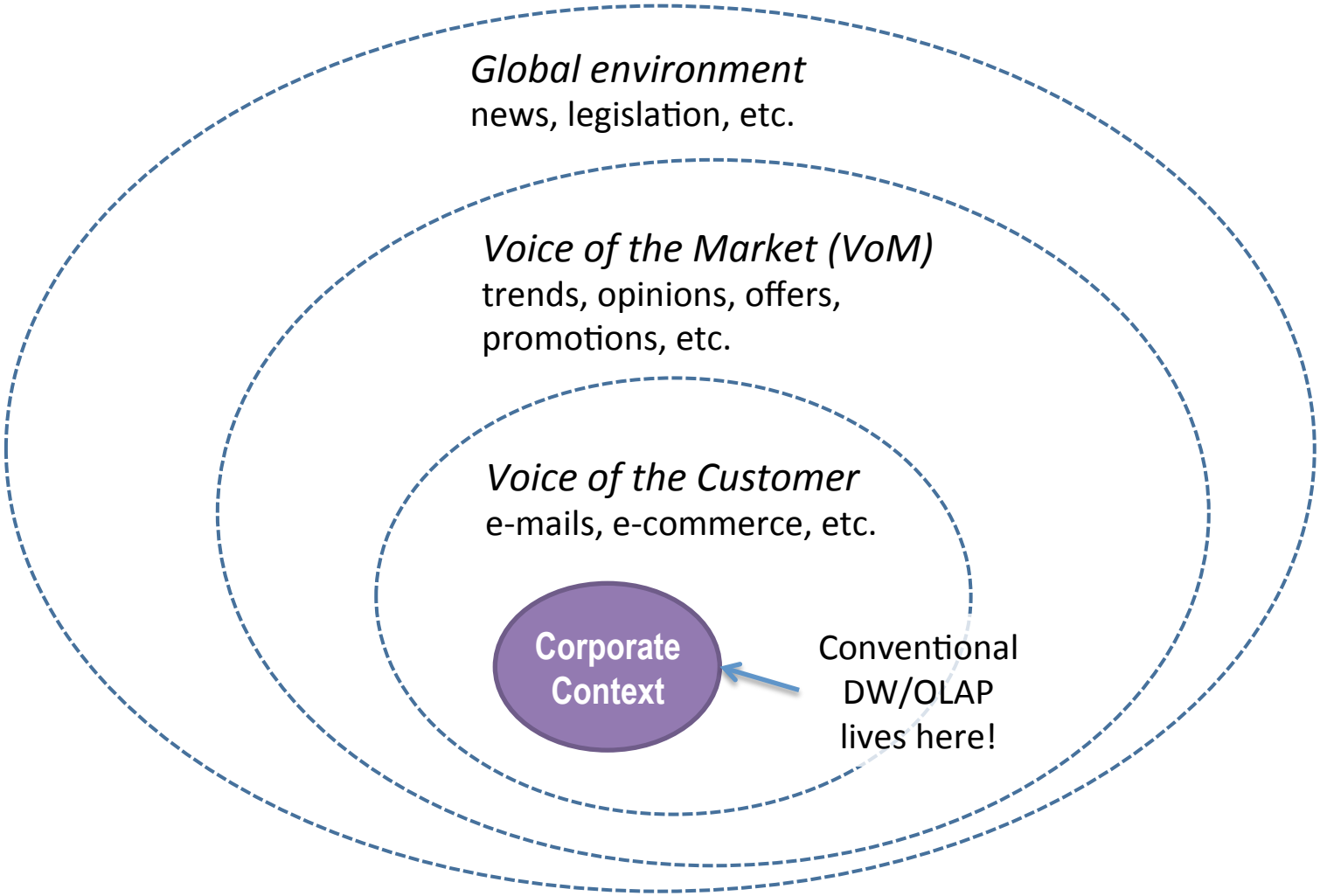


Monitoring

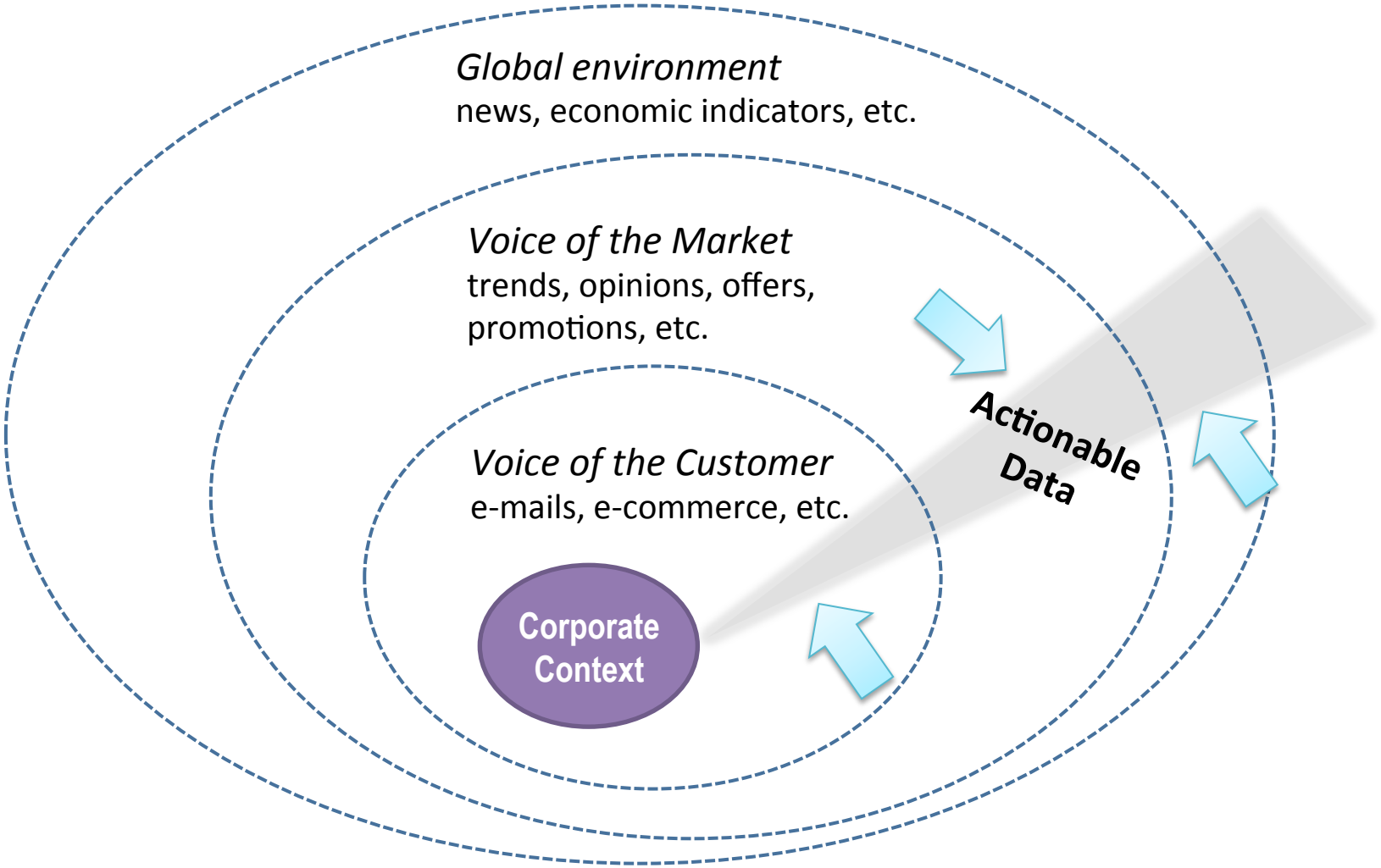
Exploring

Predicting

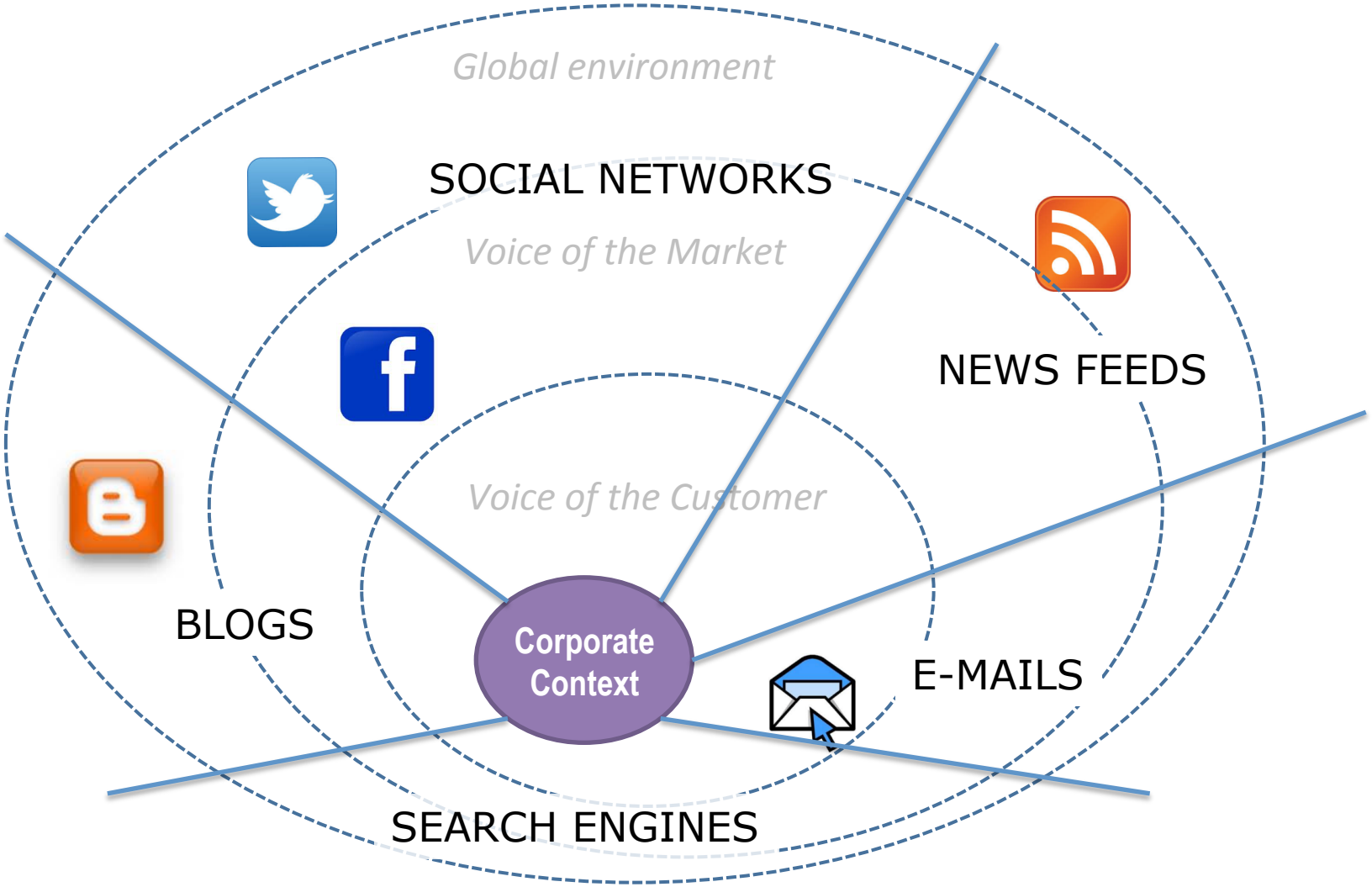
Context Layers for BI



Capturing actionable data



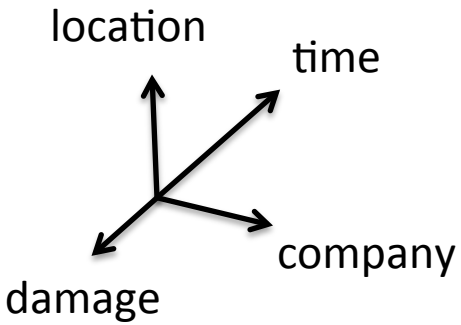
Context (Web) Sources



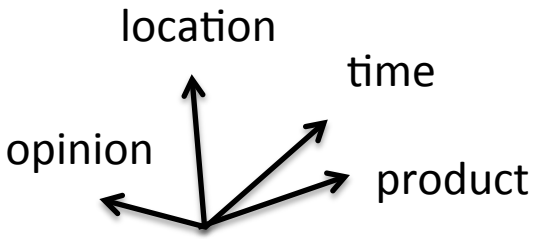
Context objects

- Context events/facts are:
  - expressed as **text-rich** data,
  - spatio-temporal located,
  - usually typified (e.g., topics),
  - multi-dimensional,
  - multi-granularity,
  - usually incomplete and imprecise.
- ... but not well-suited to conventional OLAP.

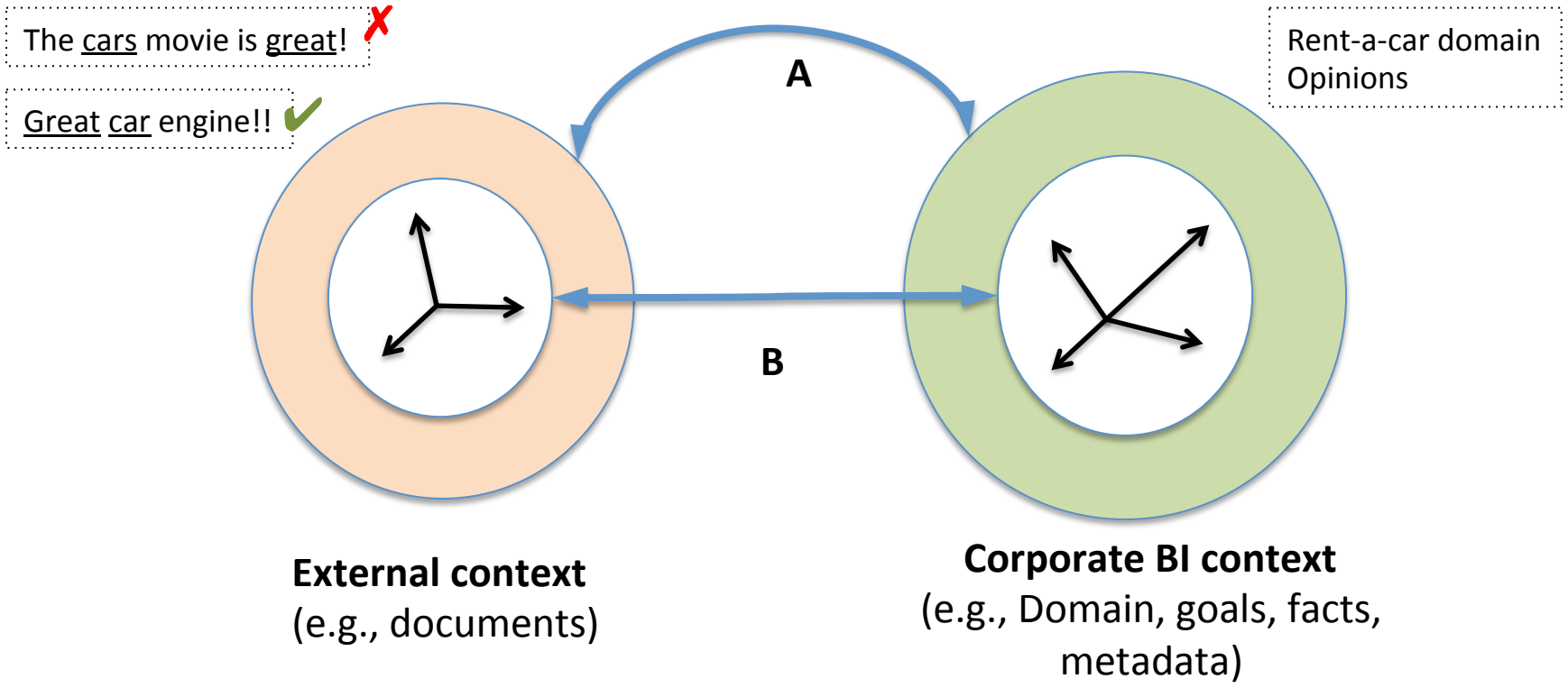
Flood disaster event



Product opinion fact



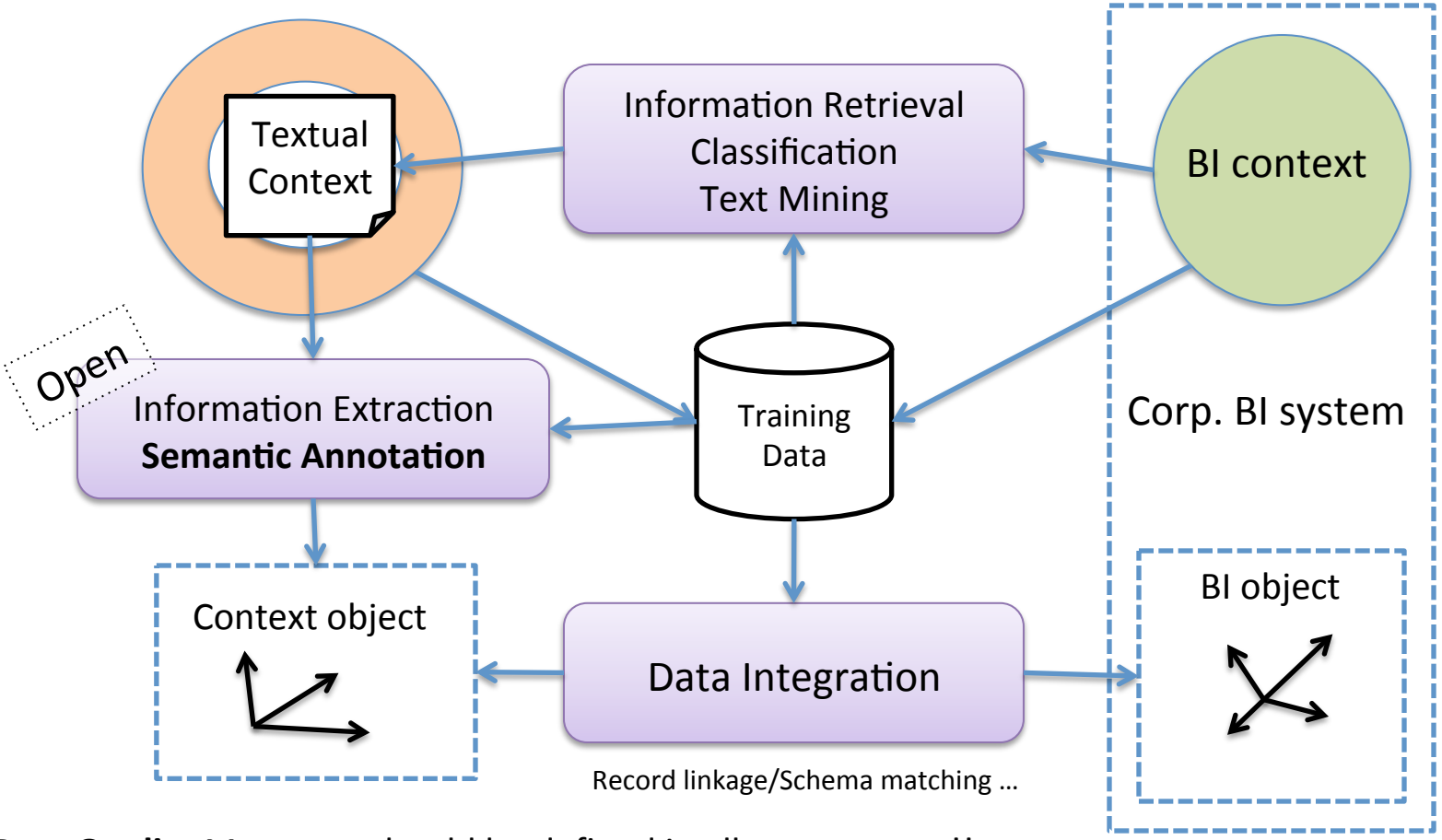
2-level context correlations



A: contexts correlation measures **relevance** of external w.r.t internal contexts.

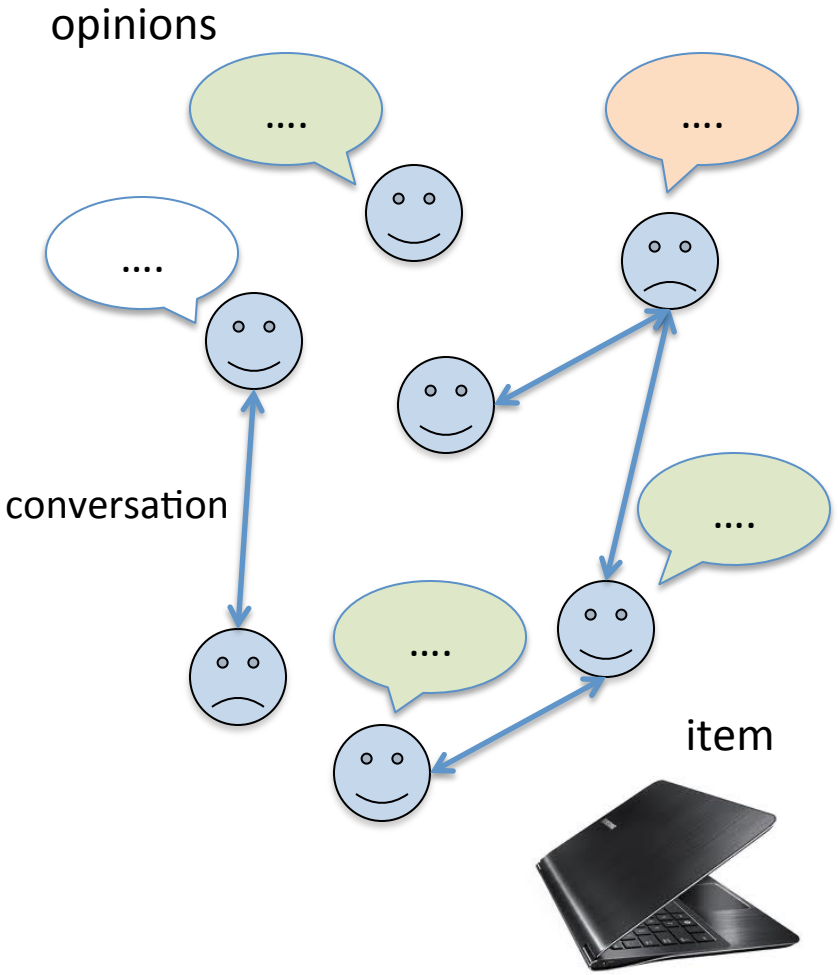
B: multidimensional object correlation ➔ classical data integration problem.

Techniques



Data **Quality** Measures should be defined in all components!!
Precision, Recall, F1-measure, etc.

Social Context

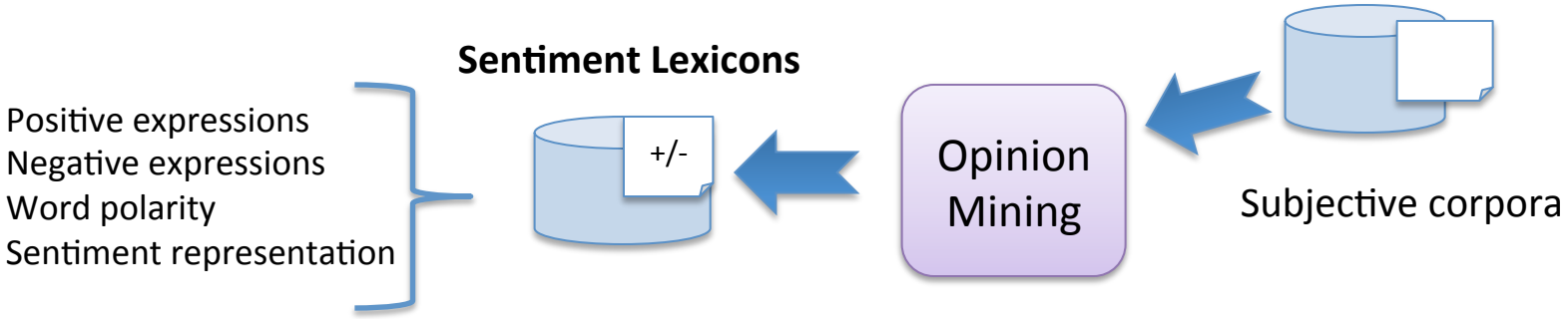


Social commerce

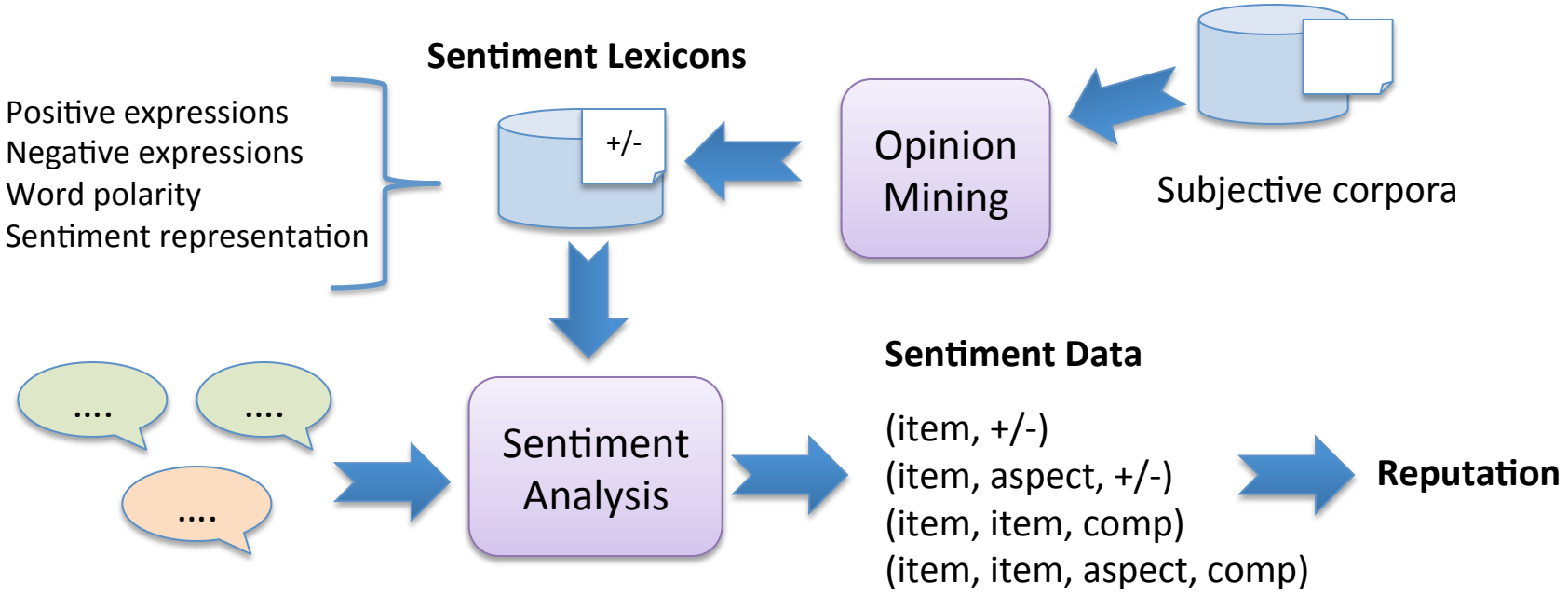
Listening to the VoM:

- global VoM opinion?
- best/worst scored aspects?
- customer needs?
- current trends?
- viral information?
- … and so on.

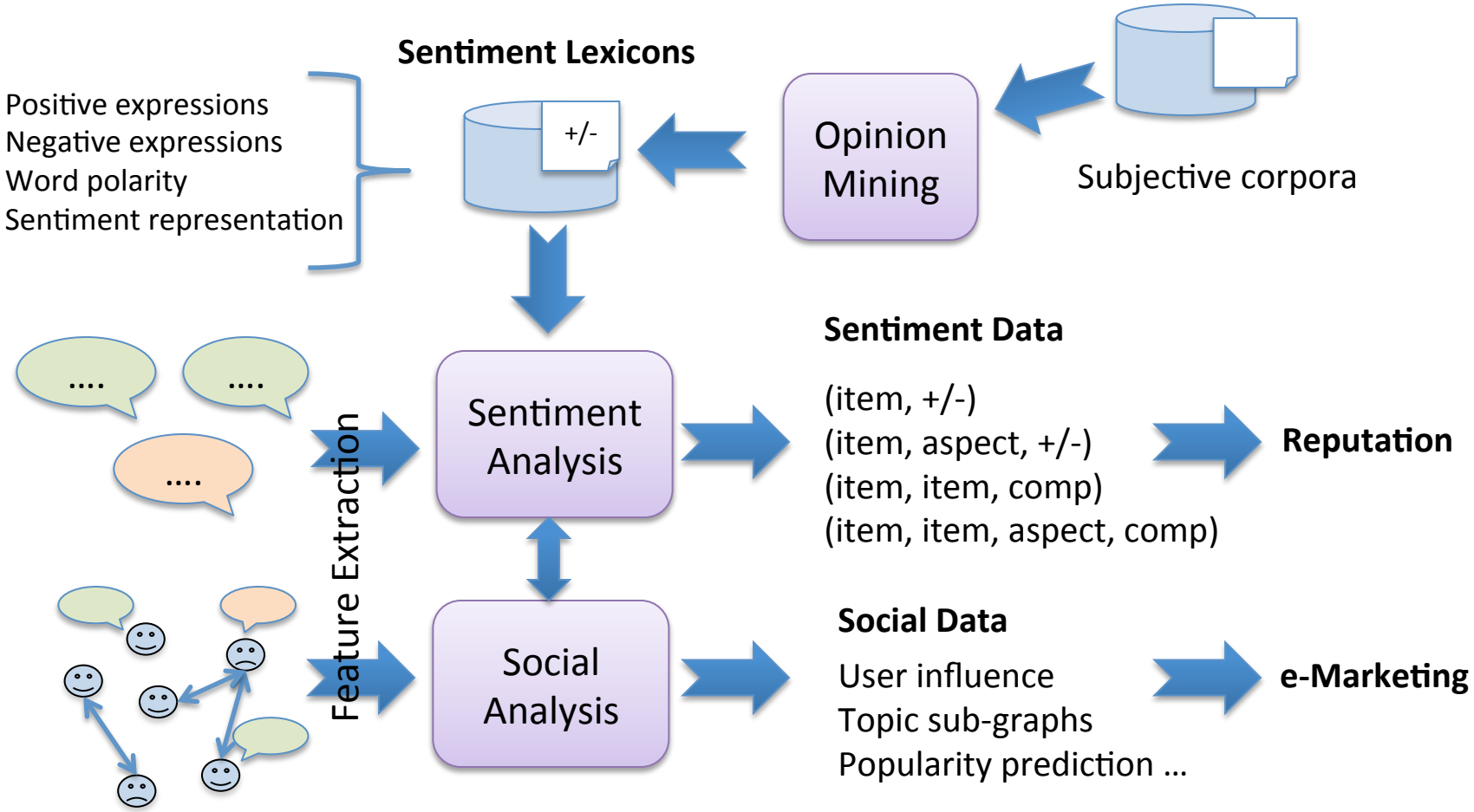
Social & Sentiment Analysis



Social & Sentiment Analysis



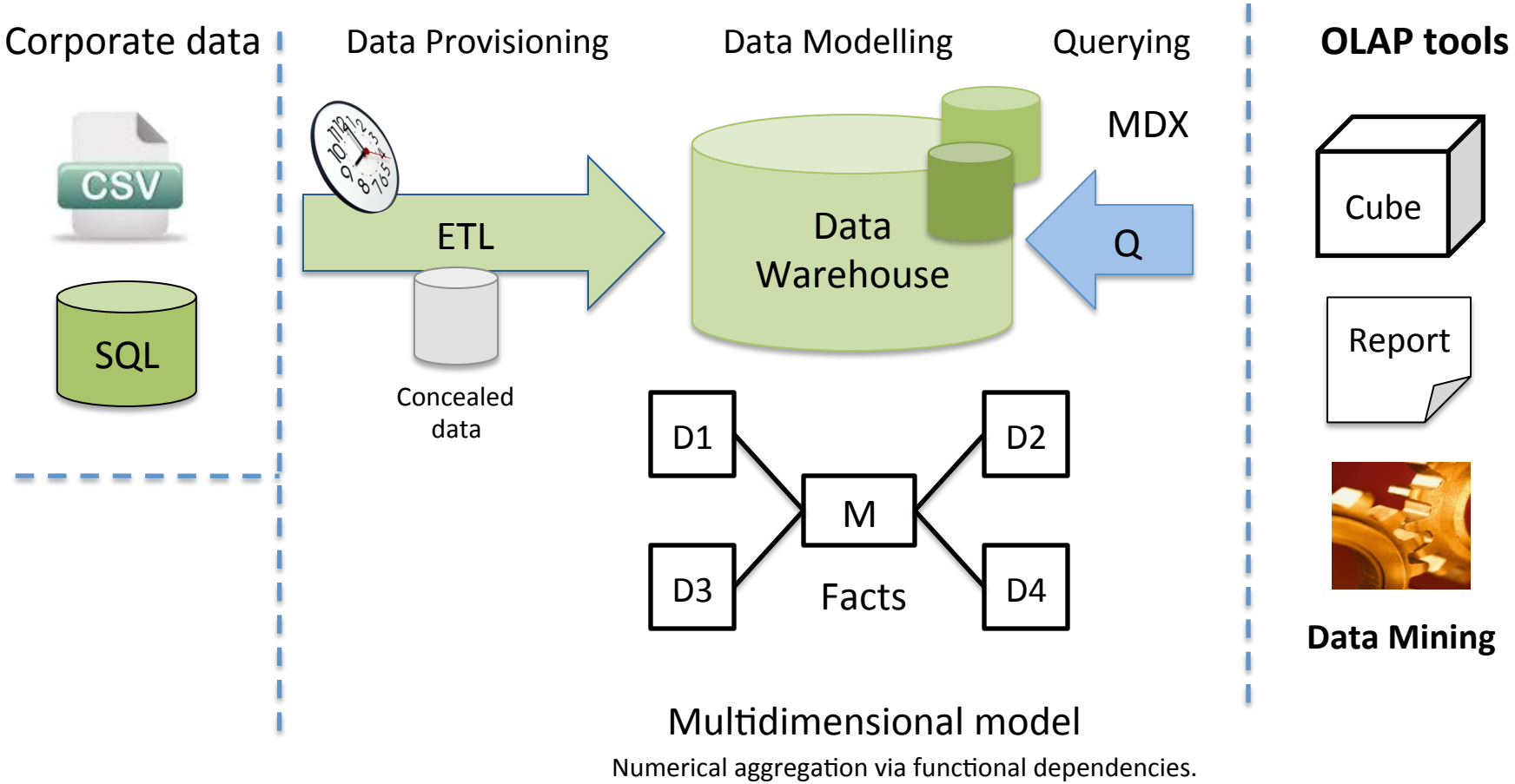
Social & Sentiment Analysis



Approaches to

CONTEXT-AWARE BI SYSTEMS

Conventional DW/OLAP

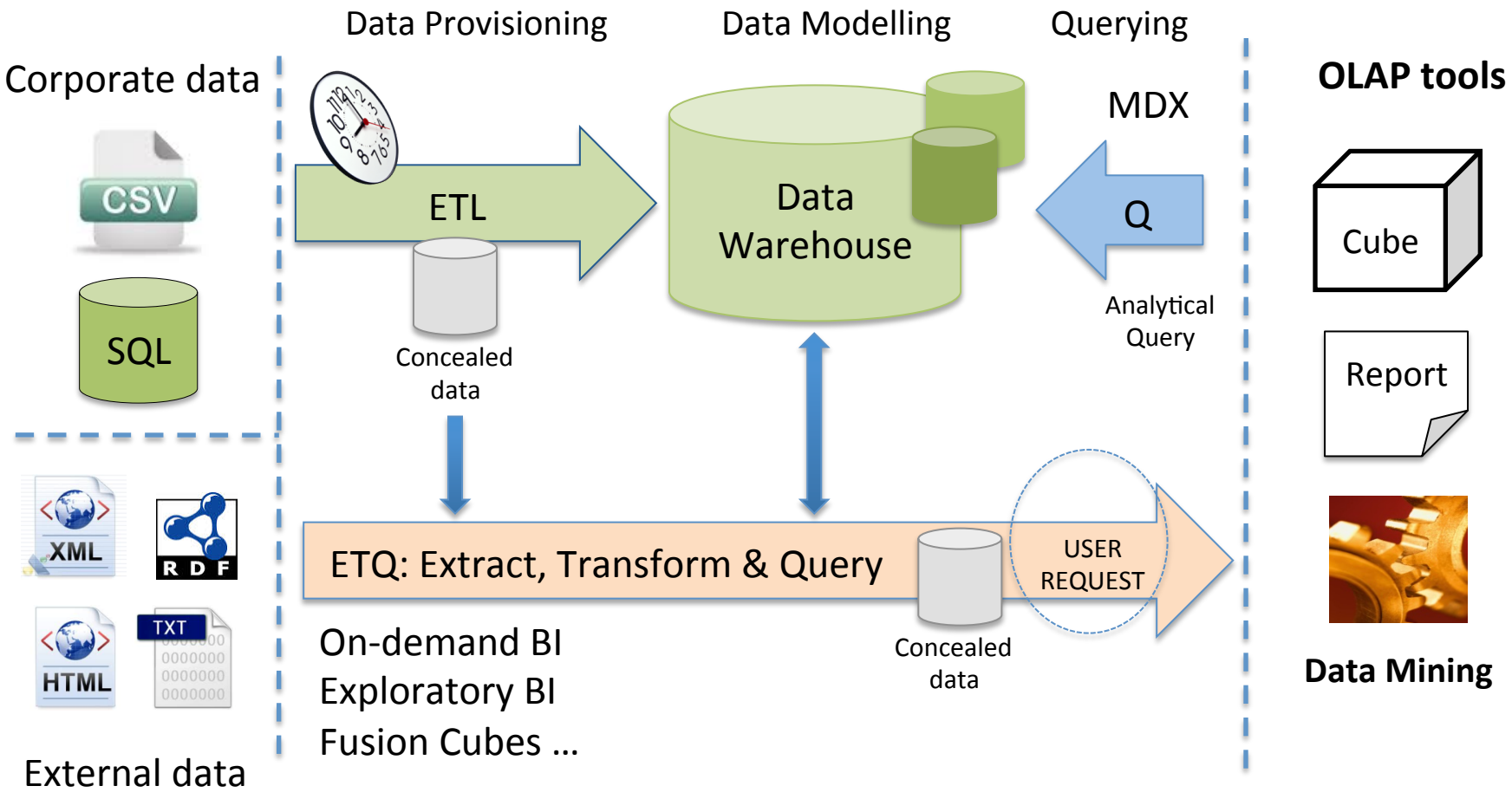


Internal Data vs. External Data

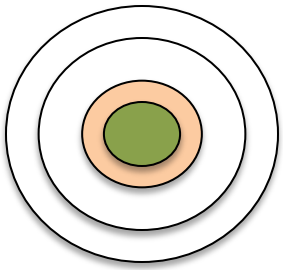
INTERNAL DATA	EXTERNAL DATA
• Slow changes	• Highly dynamic
• Mainly relational data	• Un- Semi-structured
• High quality data	• Low quality data
• Complete information	• Incomplete information
• Historical	• Up-to-date (fresh)

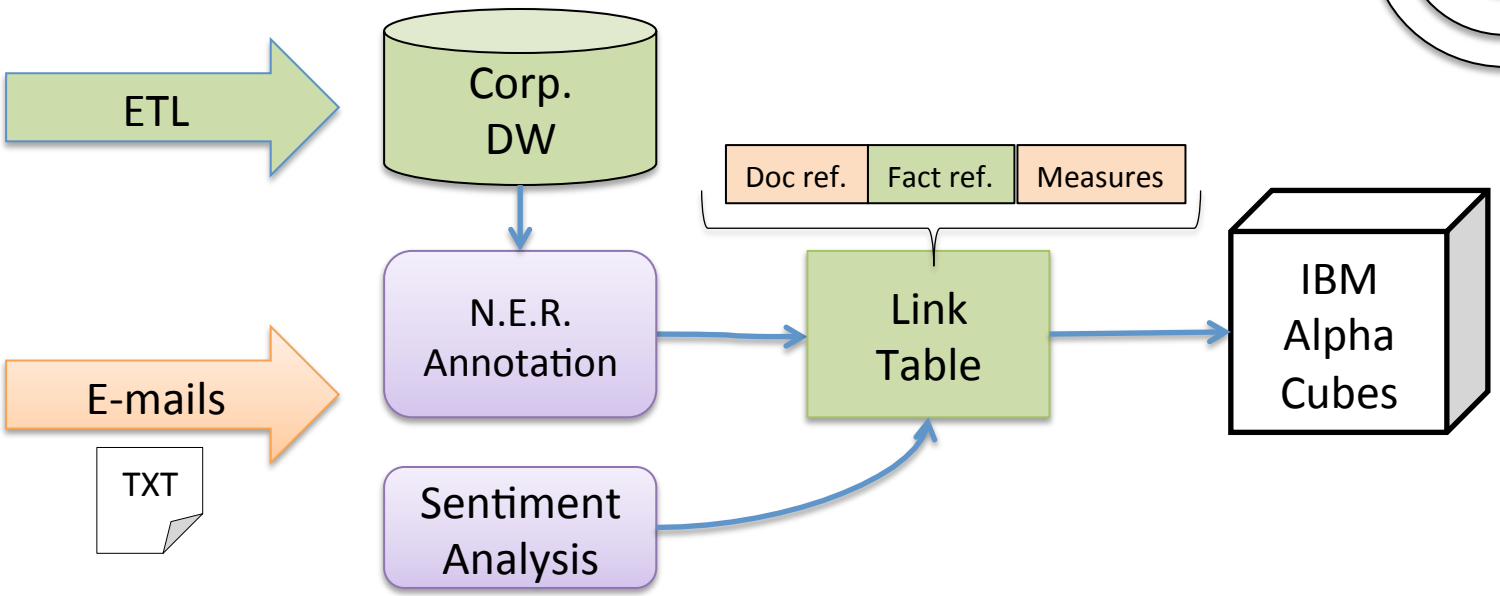
HARD CONFLUENCE!!!

On-demand BI Proposals



EROCS/LIPTUS 2006-2008

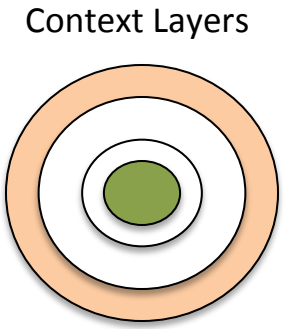




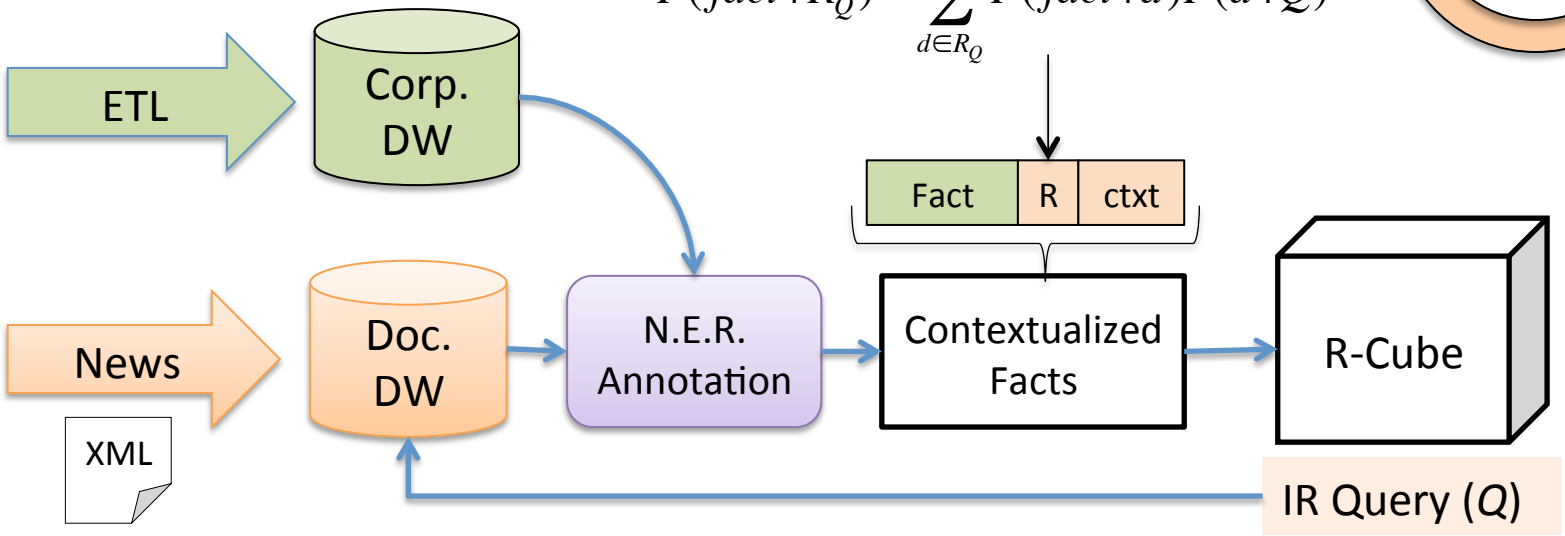
Goal: Find the DW fact that best matches the context (e-mails).

- Customer e-mails refer to complain/satisfaction opinions (business context).
- Document-fact relationship is considered 1:1.
- The extraction method uses N.E.R and TF*IDF for context relevance.
- To find the best annotation (fact) for a context is computational expensive.

R-cubes[2008]



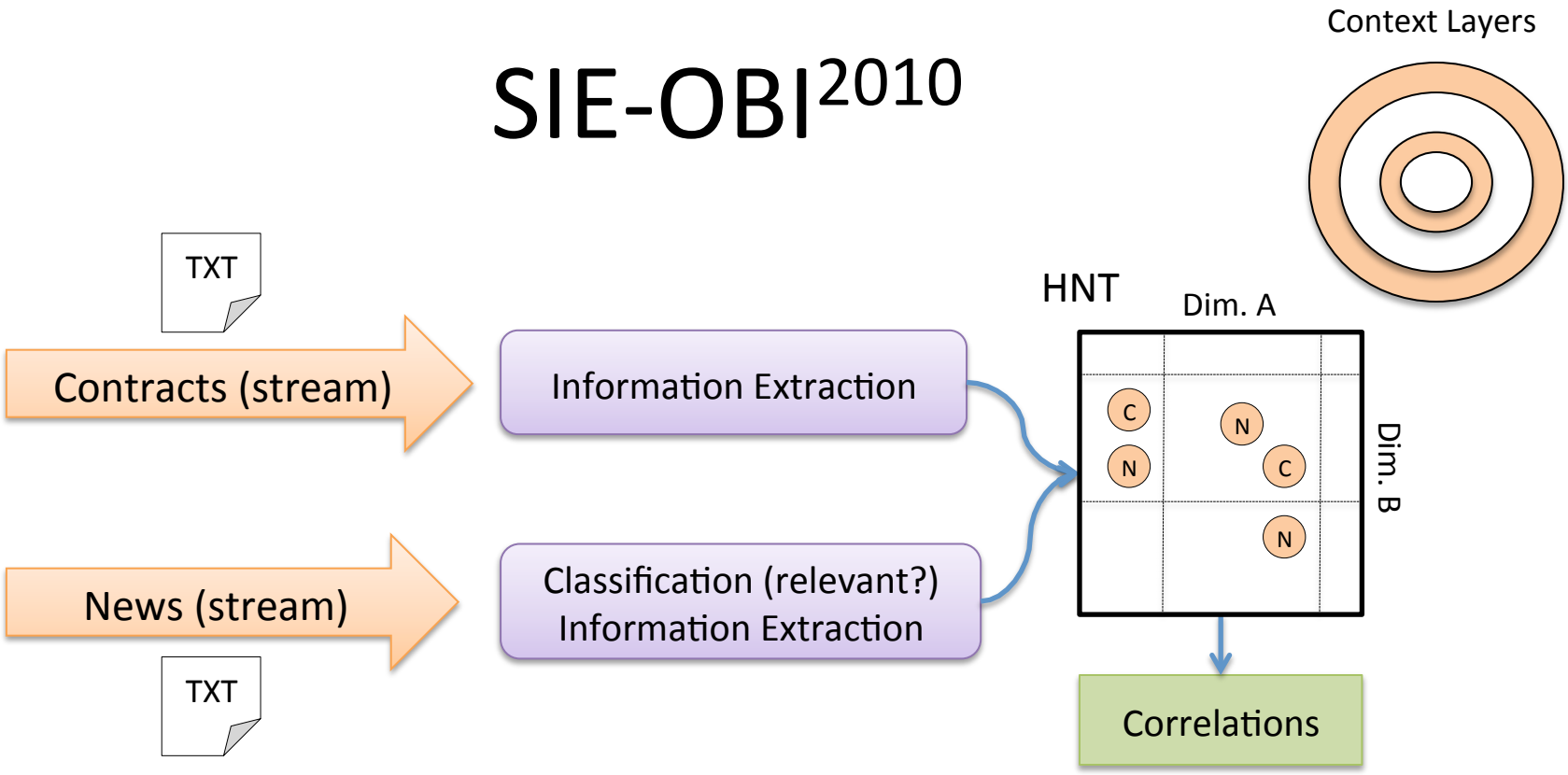
$$P(fact \mid R_Q) = \sum_{d \in R_Q} P(fact \mid d) P(d \mid Q)$$



Goal: Contextualize corporate facts with news.

- Many-to-many document-fact relationships.
- External contexts retrieved with IR Relevance Models (user guided).
- Data integration relies on N.E.R. and semantic annotation (dimension values)
- Sound OLAP support for external data: (R)elevance and context dimensions.

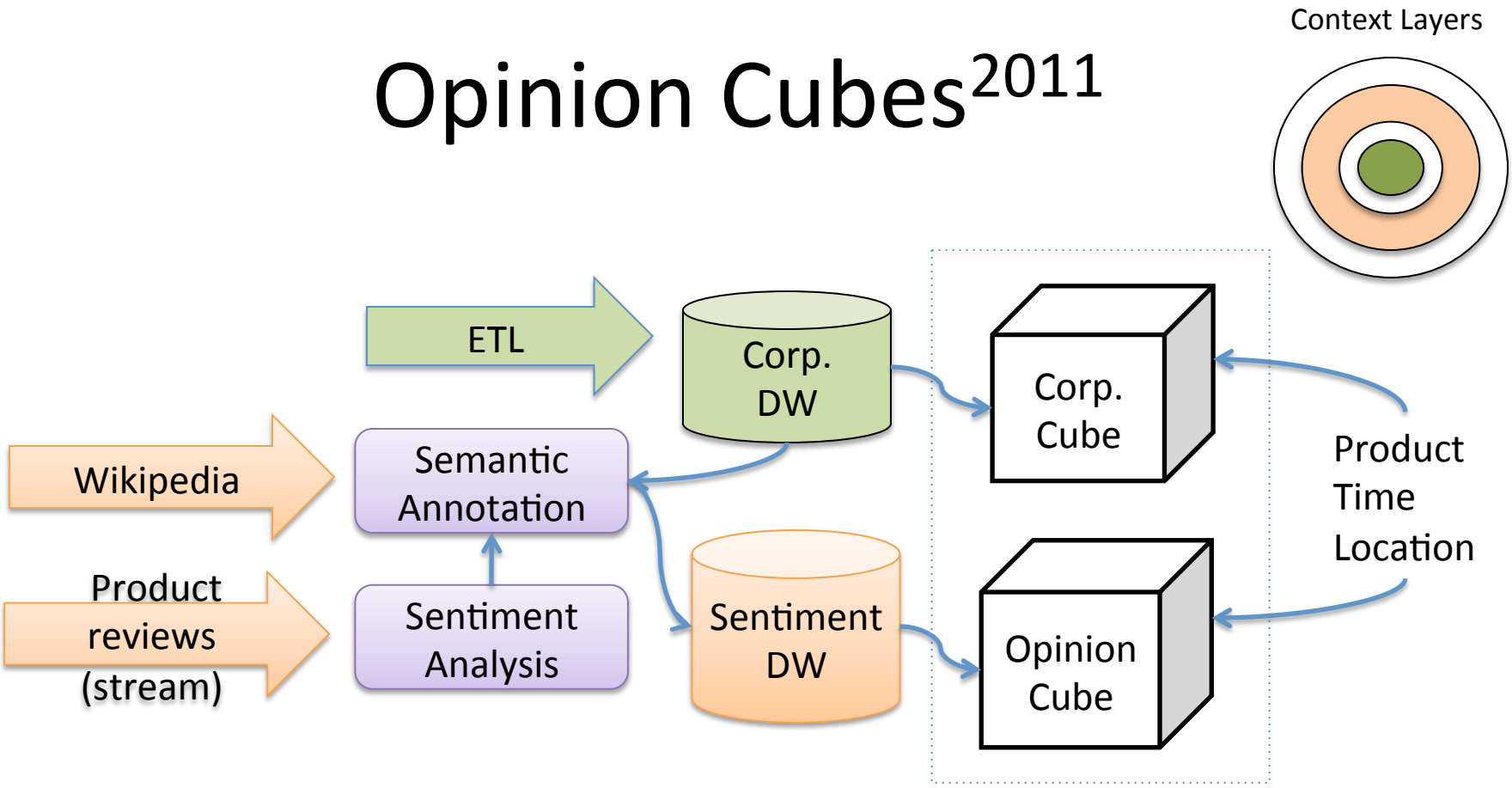
SIE-OBI^{2010}



Goal: Find multidimensional correlations between contracts and news.

- Many-to-many document-fact relationships.
- Internal DW corporate DW is not considered.
- Relevance of news context is performed with classifiers.
- Data integration relies on a neighbourhood distance over the shared MD space (HNT - Hierarchical Neighbourhood Trees).

Opinion Cubes[2011]



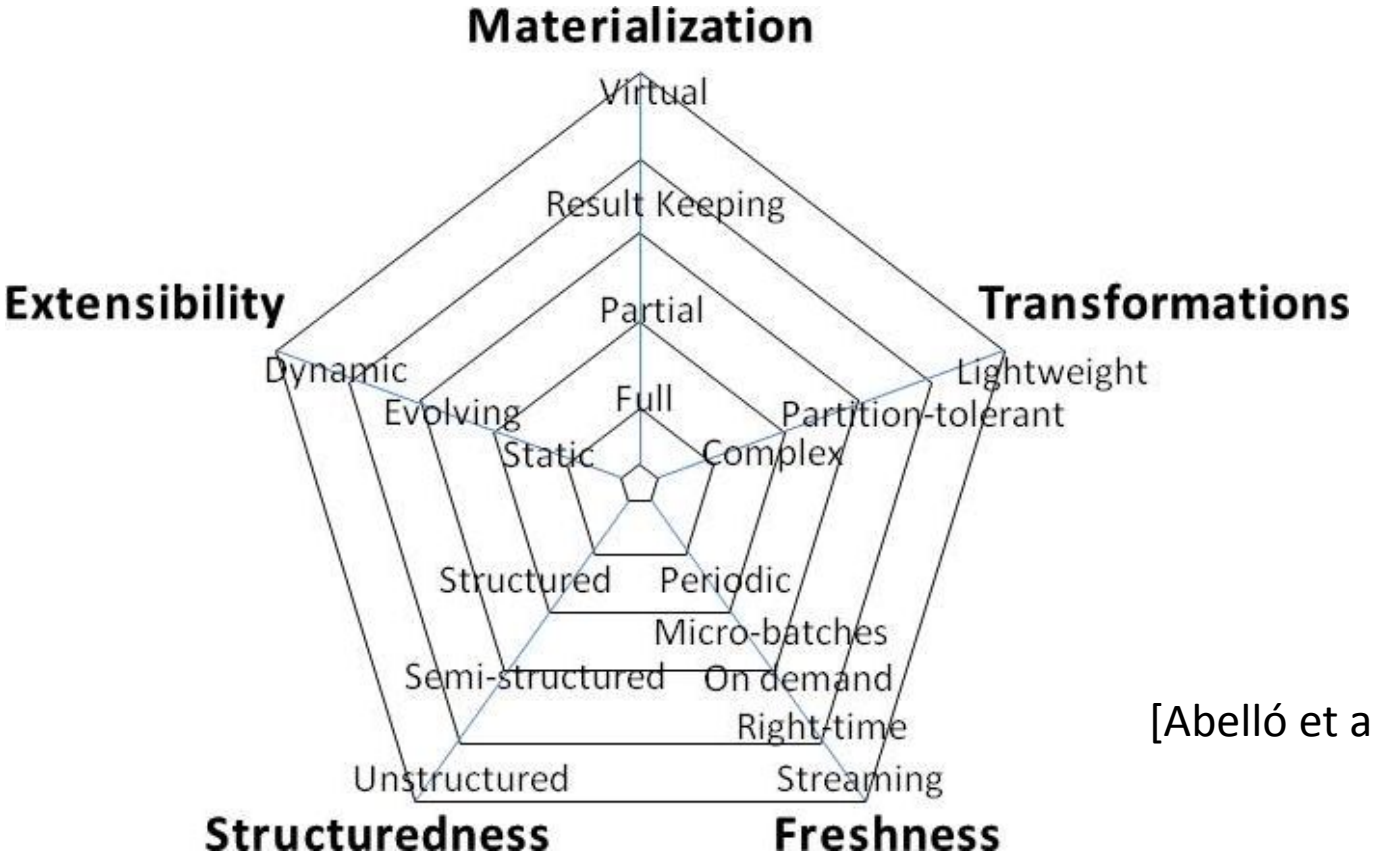
Goal: Correlate product opinions (VoM) with corporate analyses.

- *Tailored* semantic annotation (using DW metadata & Wikipedia categories).
- Context *relevance* determined by the sentiment analysis component.
- Correlations are analyzed by joining corporate and opinion cubes.

Limitations

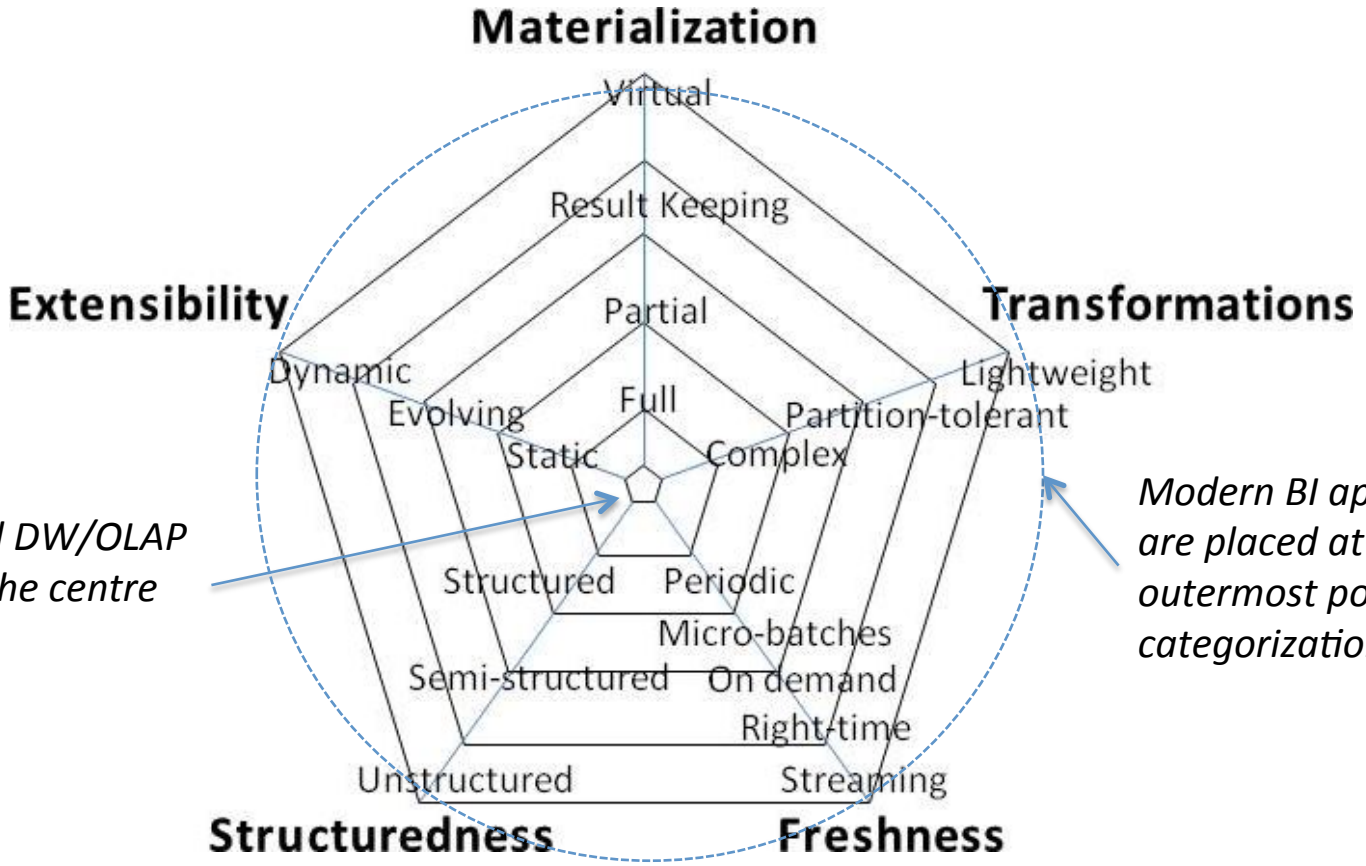
- Only one external source at time.
- No approach covers all context layers.
- Data provisioning implies a **high cost**.
- **No flexibility** for adapting to source changes.
- **Scalability** issues are not regarded.
- **Stream** data are processed in **batch** mode.
- It is assumed a **direct match** between external and business objects (not always true).
- Data **quality** issues poorly treated.

BI systems categorization



[Abelló et al. 2015]

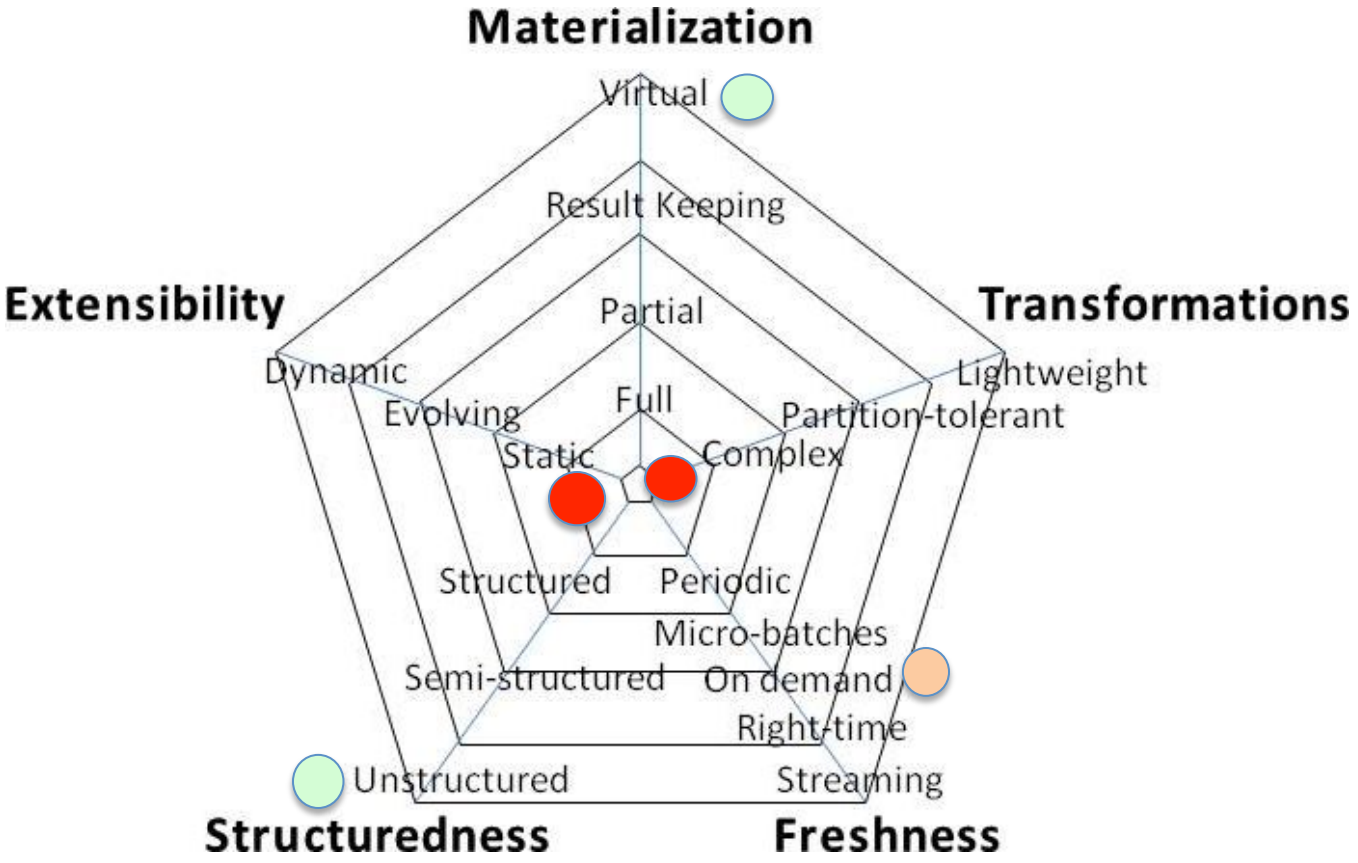
BI systems categorization



Conventional DW/OLAP is placed at the centre of the map.

Modern BI approaches are placed at the outermost points of the categorization map.

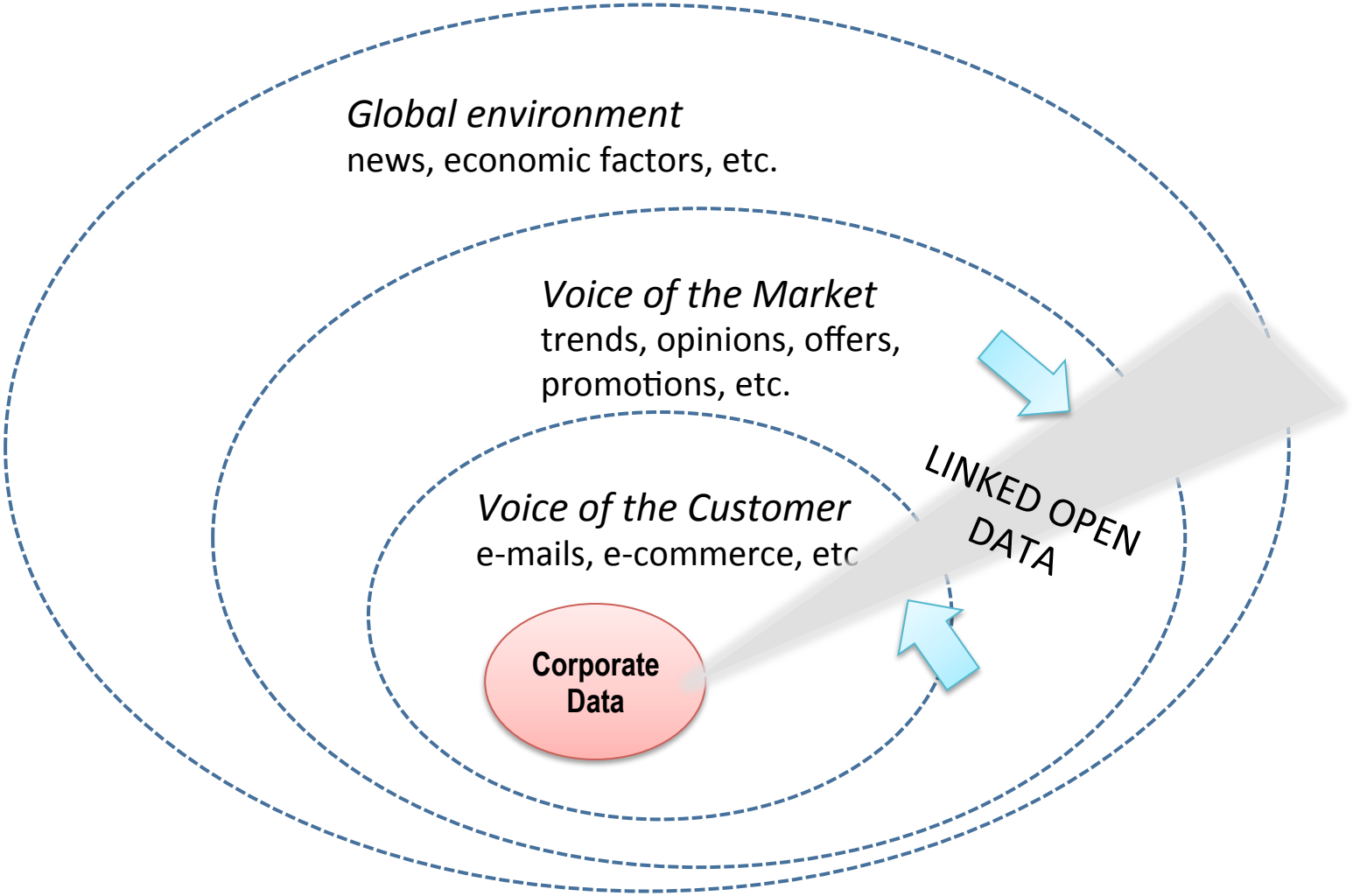
ETQ-like approaches



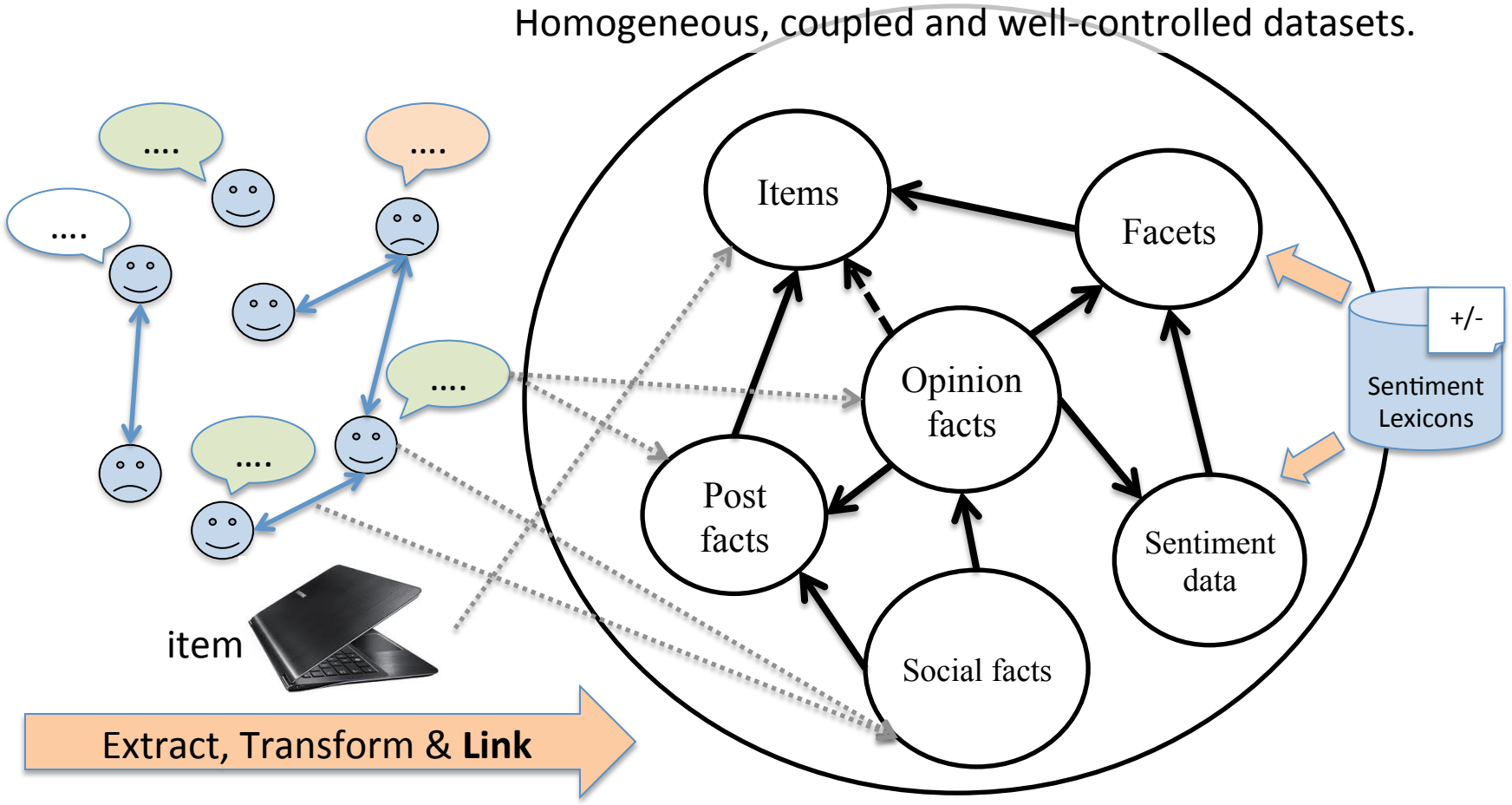
BI & Web of Data

- Linked Open Data (LOD) infrastructures
  - Publish and share datasets following web rules.
  - LOD offers standard data formats, as well as several tools to store, query and manage datasets.
- LOD provides widely-accepted vocabularies:
  - **Schema.org** (e-commerce micro-data)
  - (**qb**) RDF data cube (summarized data)
- LOD enables BI systems:
  - to **automatize** data provisioning and modelling
  - to **normalize** context objects
  - to **semantically enrich** context objects (exploratory)

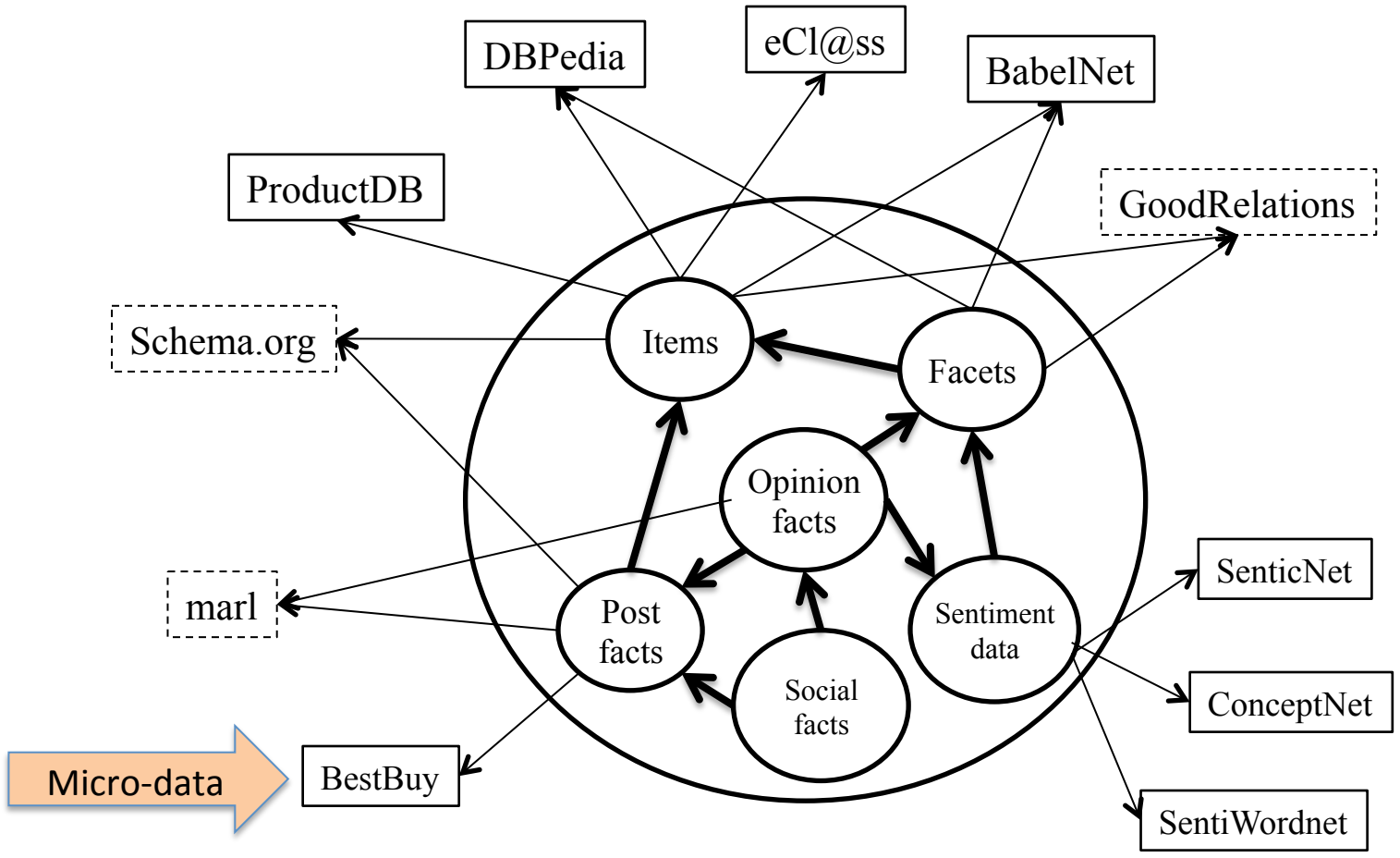
Web of Data & BI



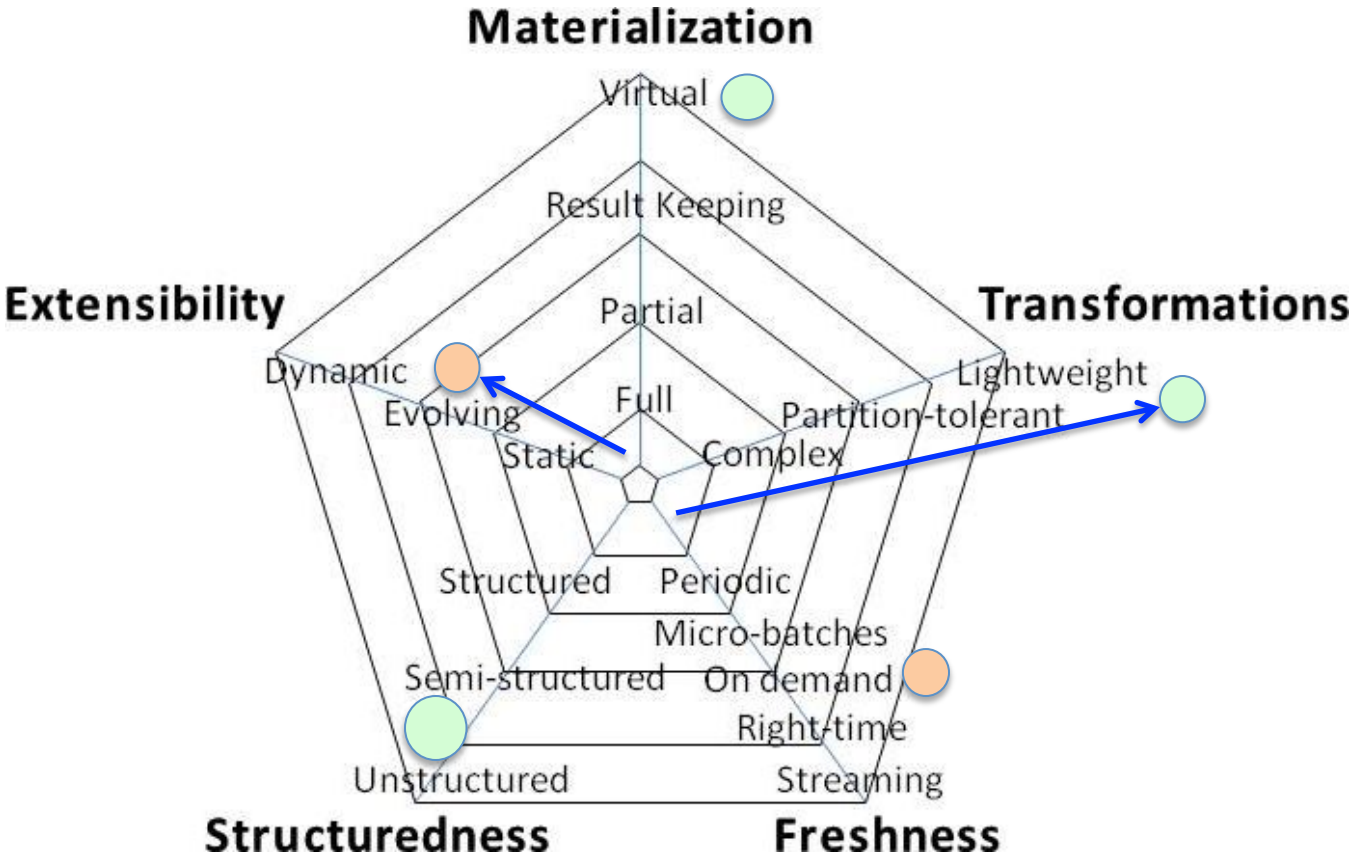
Revisiting opinion cubes: SLOD-BI[2015]



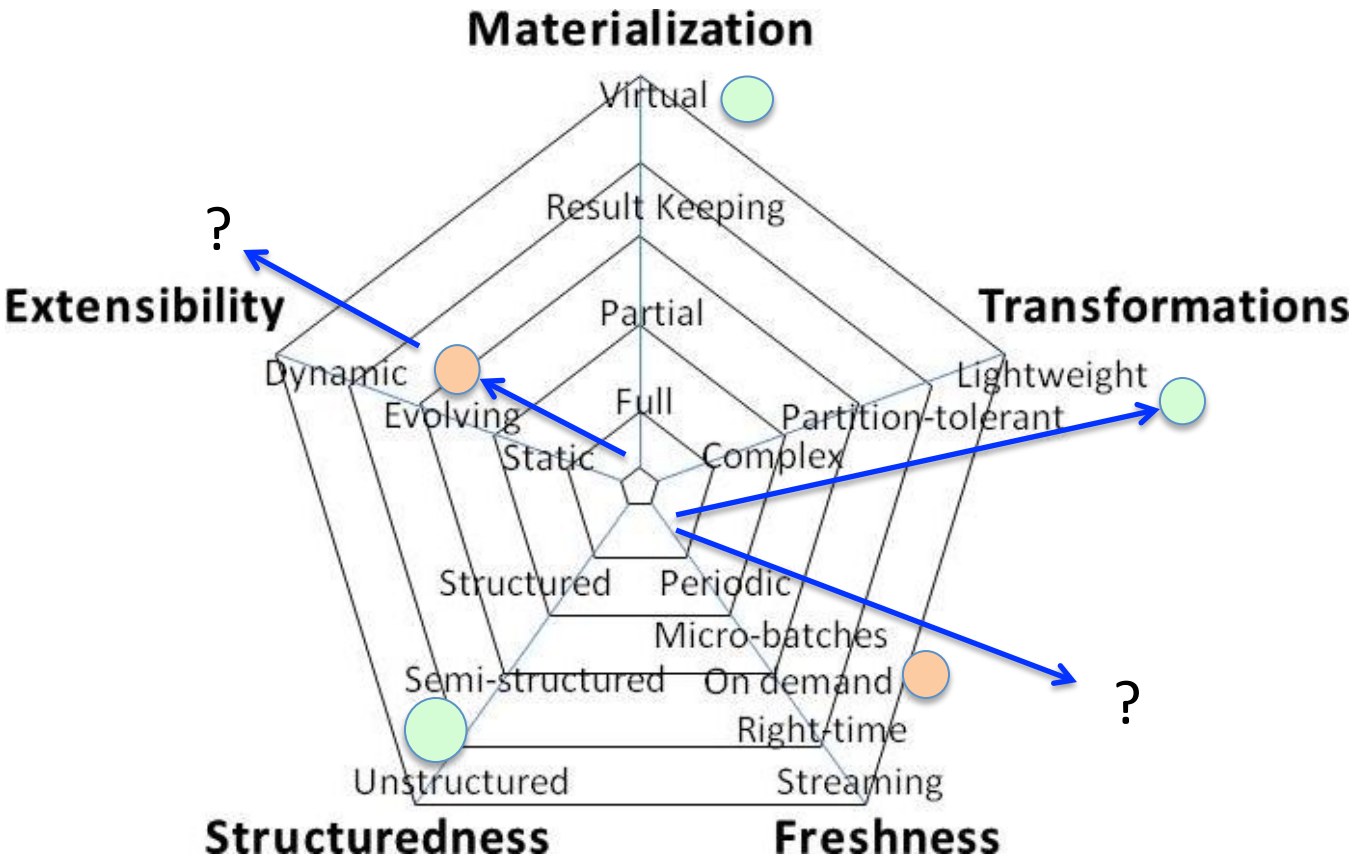
SLOD-BI as linked datasets



BI & LOD infrastructures



Next steps?


Materialization
Virtual
Result Keeping
Partial
Full
Extensibility
?
Transformations
Dynamic
Evolving
Static
Lightweight
Partition-tolerant
Complex
Structured
Periodic
Micro-batches
Semi-structured
On demand
Right-time
?
Unstructured
Streaming
Structuredness
Freshness

Future work

CHALLENGES AND ISSUES

Challenges and Issues

- Streaming technology:
  - Continuous monitoring of external data (big data).
  - Continuous training of ML components (concept drift)
  - Scalability issues always matter.
- Streaming + Linked Data?
  - Definition of **dynamic datasets**.
  - More flexible **data formats** like JSON-LD are needed.
  - **Scalable query processing** is required !!
  - Continuous adaptation of datasets to BI goals.
  - Continuous adaptation of semantic annotators.

Challenges and Issues

- Data Quality and Veracity (the 4th V of Big Data)
  - Reliability of analytical data should be studied
- Other short-term issues:
  - Automation of ETQ/ETLink processes.
  - New semantic-aware OLAP queries for social BI.
  - Better integration of corporate and social BI.
  - Summarizability issues (necessary in open scenarios?)
  - Adaptation/reformulation of MD indexing.
  - Fully integrated BI platforms in the cloud (e.g., IBM Watson Analytics).

THANKS FOR YOUR ATTENTION
Questions are welcome!

References

- [EROCS] Bihde, M., Chakravarthy, V., Gupta, A., Gupta, H., Mohania, M., Puniyani, K., Roy, P., Roy, S., Sengaral, V. "Enhanced Business Intelligence using EROCS", ICDE 2008.
- [R-CUBES] Pérez-Martínez, J.M, Berlanga R., Aramburu M.J., Pedersen, T.B. "Contextualizing data warehouses with documents". Decision Support Systems 45(1): 77-94 (2008)
- [SIE-OBI] Castellanos, M., Gupta, C., Wang S., Dayal, U., Durazo, M. "A platform for situational awareness in operational BI". Decision Support Systems 52(4): 869-883 (2012)
- [Opinion Cubes] García-Moya, L., Kudama, S., Aramburu, M.J., Berlanga R. "Storing and analysing voice of the market data in the corporate data warehouse". Information Systems Frontiers 15(3): 331-349 (2013)
- [SLOD-BI] Berlanga, R., García-Moya, L., Nebot, V., Aramburu, M.J., Sanz, I., Llidó, L. "SLOD-BI: An Open Data Infrastructure for Enabling Social Business Intelligence". Inter. Journal of Data Warehousing and Mining, Vol. 11 (4): 1-28, 2015.
- Abelló, A., Romero, O. Pedersen, T. B., Berlanga, R., Simitsis, A., Nebot, V., Aramburu, M.J. "Using Semantic Web Technologies for Exploratory OLAP: A Survey", IEEE TKDE Vol 27(2), 2015.

CONTEXT-AWARE BI

USE CASE

CAR RENTAL DOMAIN

Rafael Berlanga Llavori
Victoria Nebot Romero

DLSI, Universitat Jaume I, Castellón, Spain

CAR RENTAL DOMAIN

Mission: cost effective and quality services

Typical management:

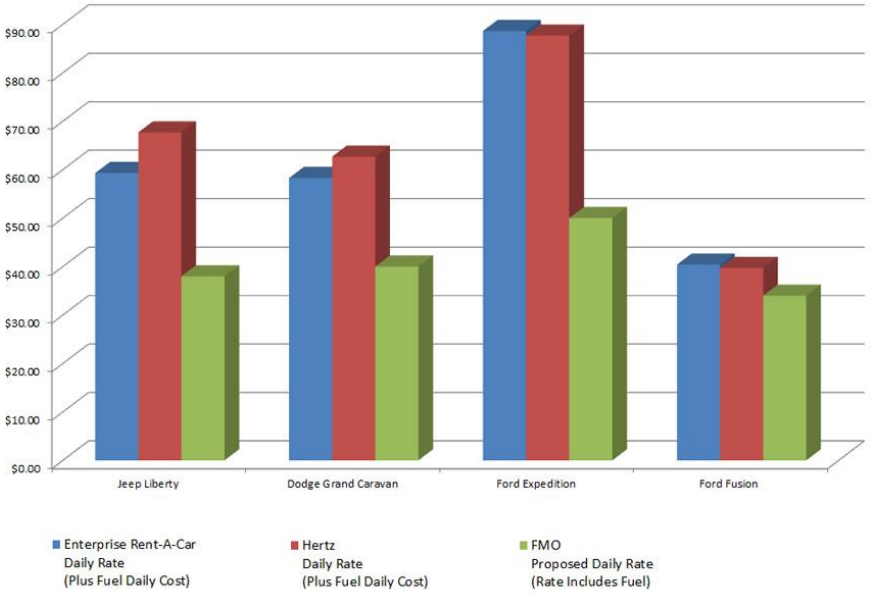
- Handling rental reservations (from individuals or companies)
- Handling car upgrades
- Offering the best promotional offer plans
- Changing the fleet
- Etc.

Corporate BI data:

- Rental’s cube
- Fleet’s cube
- Etc.

CAR RENTAL DOMAIN

- Q1: number of rental agreements per location, product and time
- Q2: most rented vehicles by location



NEW DEMANDS

- **Customer satisfaction:** how happy are customers with the new rental offices opened in the U.S. during the last year?

- **Opinion about company's products:** what are the features that customers value the most in the newly acquired Ford Focus fleet?

- **News related to the company's domain:** how's the merge of Fiat and General Motors going to affect our fleet?

- **News about competitors:** what do customers value the most about Sixt rental company?

CONTEXT-AWARE BI INFRASTRUCTURE

Car blogs

WHATCAR? Britain's biggest and best car buyer's guide

Auto Express THE CAR NEWS WEEKLY

CAR AND DRIVER

News feeds

The New York Times

theguardian

Twitter stream

Semantic middleware

Rent-a-car corporate data

Semantic middleware

Knowledge resources (KRs)

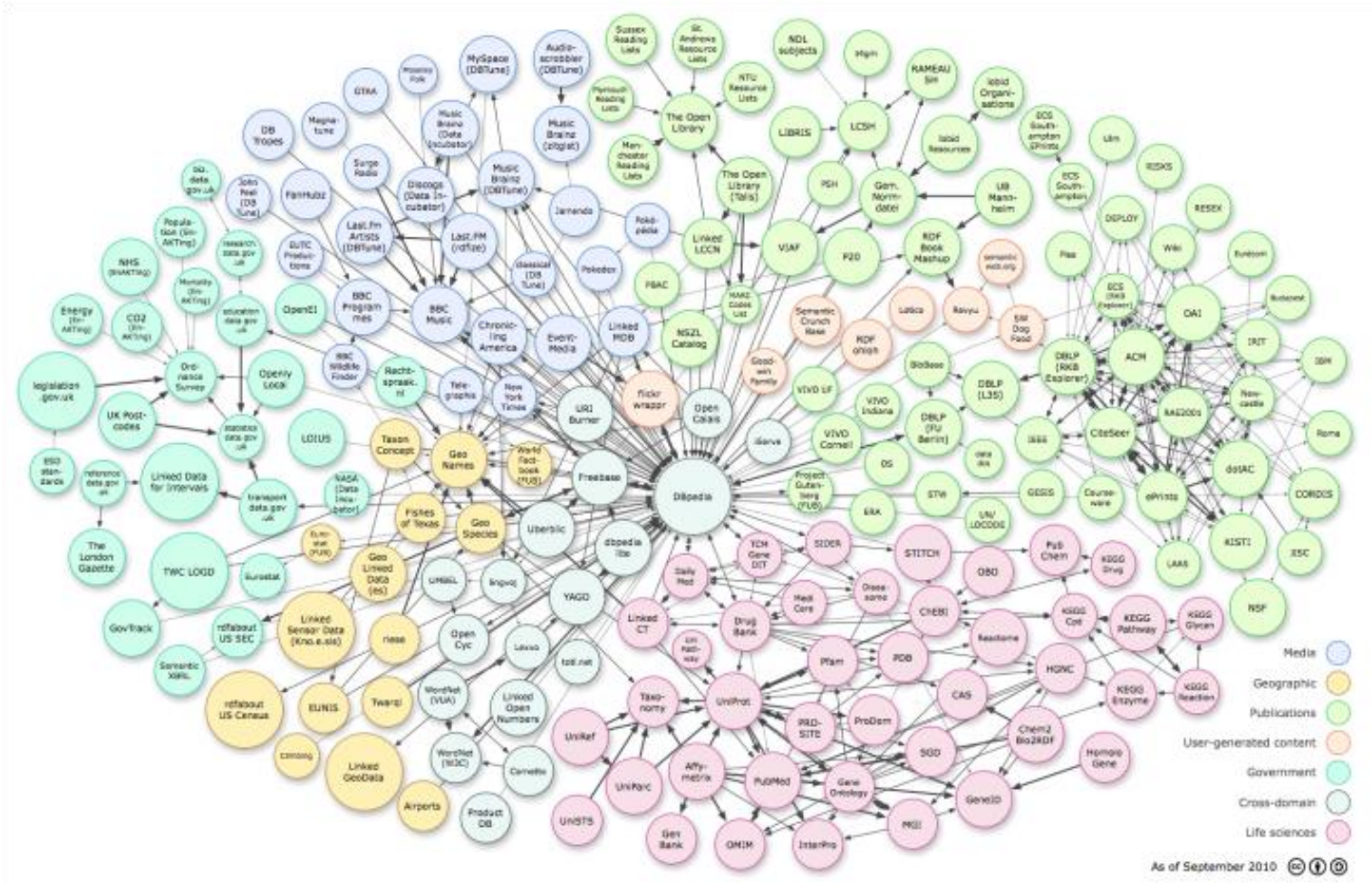
DBpedia

BabelNet

SEMANTIC MIDDLEWARE

LOD triples (s,p,o)

RDF



MAIN TASKS:

Extraction of sentiment data from texts

Express sentiment data as LOD (semantic middleware format)

Linking to corporate data

Semantic annotation

SEMANTIC ANNOTATION

Semantic annotation as a means to:

- Bring meaning to both internal and external data via public KRs
- Semantic integration of DW corporate data and external data
- Explicit link to external and shared KRs

http://babelnet.org/synset?word=bn:02438084n

"The Mazda 5 has a useless backseat for anyone with legs"

SEMANTIC ANNOTATION

Demos:

- **TagME**
- **Dbpedia spotlight**
- **Wikipedia miner**
- **BabelNet**

EXTERNAL DATA

Opinions

- "The Prius, in all its variations, is the last car on earth I would purchase."
- "no lithium-ion batteries? c'mon Toyota. time for a redesign of this and the Prius to incorporate that. you're falling behind in a segment that you once dominated."
- "The Mazda 5 has a useless backseat for anyone with legs. I sat in one before and the passengers in the 2nd row have to literally smack their knees on the backs of the 1st row occupant to give 3rd row peeps any kind of leg room at all."
- "BMW is garbage. I'd rather waste money on a Hyundai first."

News feeds

- It's Official: The Mazda CX-3 Is Most Economical Subcompact SUV
- Marchionne asks investors for help persuading GM to merge with Fiat Chrysler, report says
- Hyundai offers Google Android Auto system in cars

Car reviews

- The big points of the GT, however, are the chassis modification which creates a lower, more dynamic ride: lower = faster = cooler in car land; and the new series of mega-efficient, super-clean engines – the 1.6-litre BlueHDi 120 does an almost unbelievable 88.2 miles to the gallon with the merest whiff of CO_2, just 85g per km.

CONTEXT-AWARE BI INFRASTRUCTURE

Car blogs

WHATCAR? Britain's biggest and best car buyer's guide

Auto Express THE CAR NEWS WEEKLY

CAR AND DRIVER

News feeds

The New York Times

theguardian

Twitter stream

Semantic middleware

Rent-a-car corporate data

Semantic middleware

Knowledge resources

DBpedia

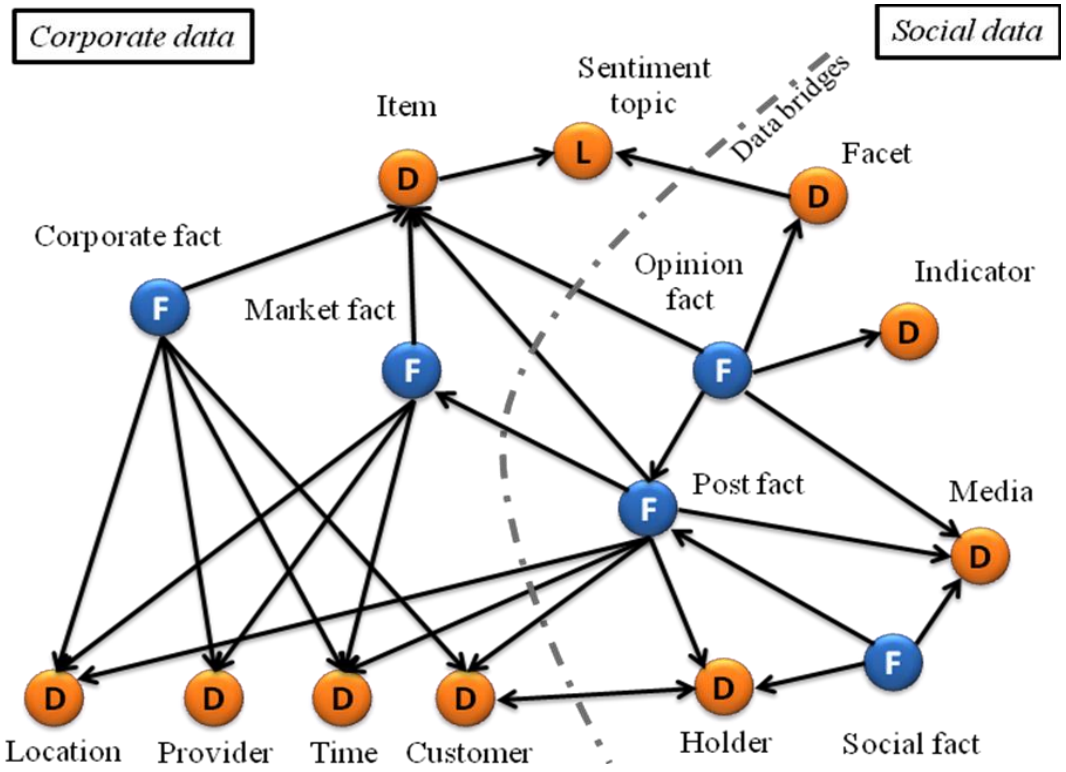
BabelNet

Semantic annotation

OPINIONS → SLOD-BI PLATFORM

Open semantic infrastructure for enabling social BI

- Transform social data to Semantic LOD BI patterns



SLOD-BI PATTERNS

***Opinion facts*: observations about sentiments concerning concrete facets about an item, along with their sentiment indicators.**

- E.g. “I don’t like the camera zoom”
  - Facet → “zoom”
  - Item → camera X
  - Sentiment indicator → “don’t like” (negative polarity).

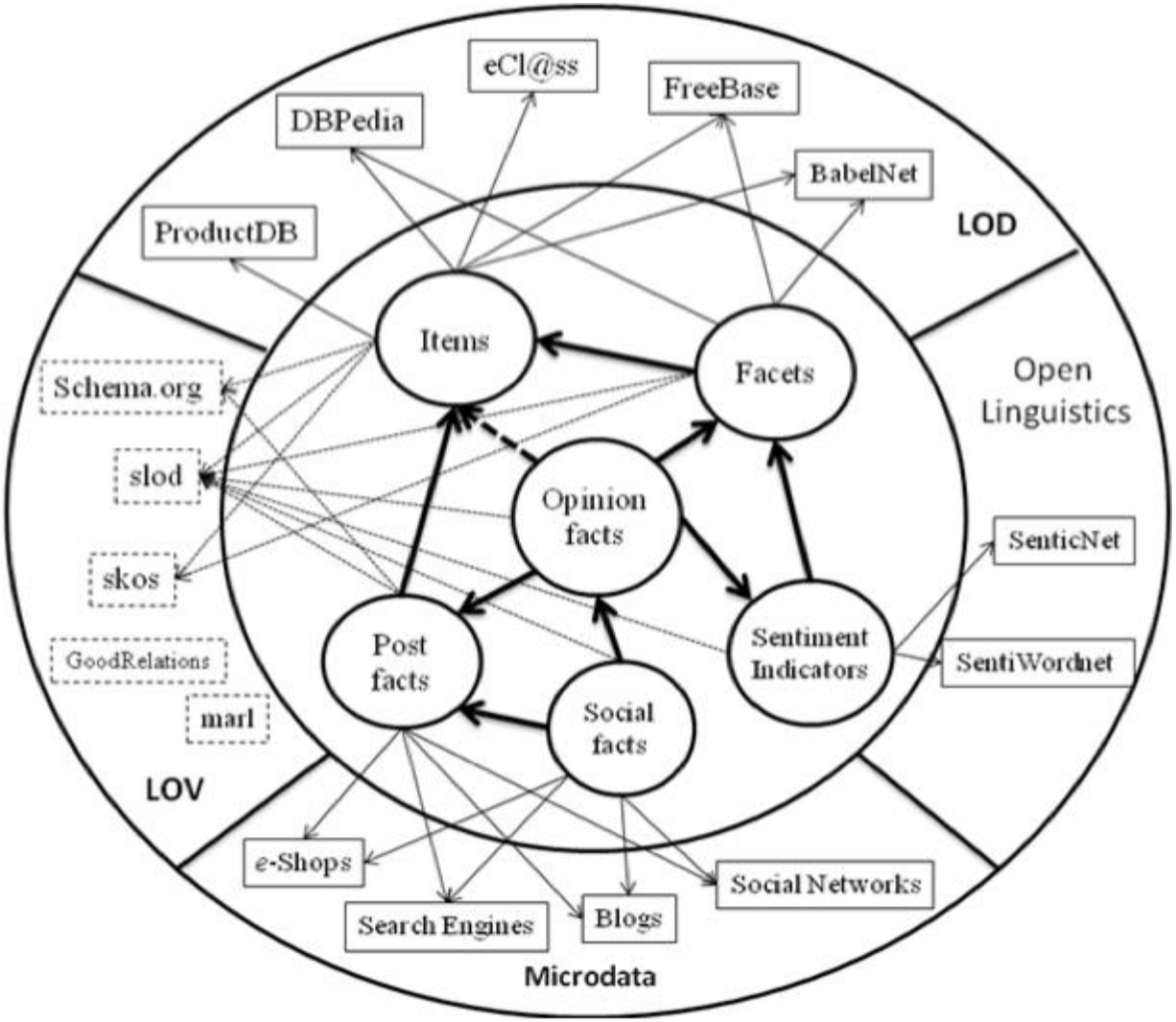
***Post facts*: observations of published information about some target item, which can include a series of opinion facts.**

- E.g. reviews, tweets, and comments published in a social network.

SLOD-BI MEASURES

Measure	Example values	Fact type
Polarity	(-1,0,+1)	Opinion
Rating	(★,★★,★★★,…)	Post
Like	(👍, 👎)	Post
Popularity	(-10,..,+10)	Social
Credibility	(0,..,10)	Social

STRUCTURAL VIEW OF SLOD-BI



EXTERNAL VOCABULARIES

LOD

- ProductDB
- Dbpedia
- eClass
- Freebase
- BabelNet

LOV

- schema.org
- SKOS
- GoodRelations

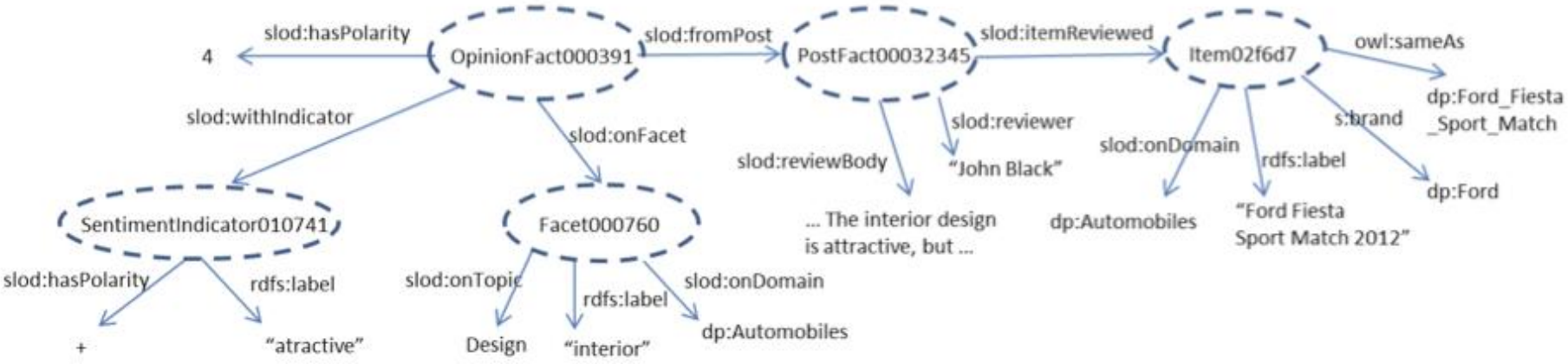
Open Linguistic

- SenticNet
- SentiWordNet

Microdata

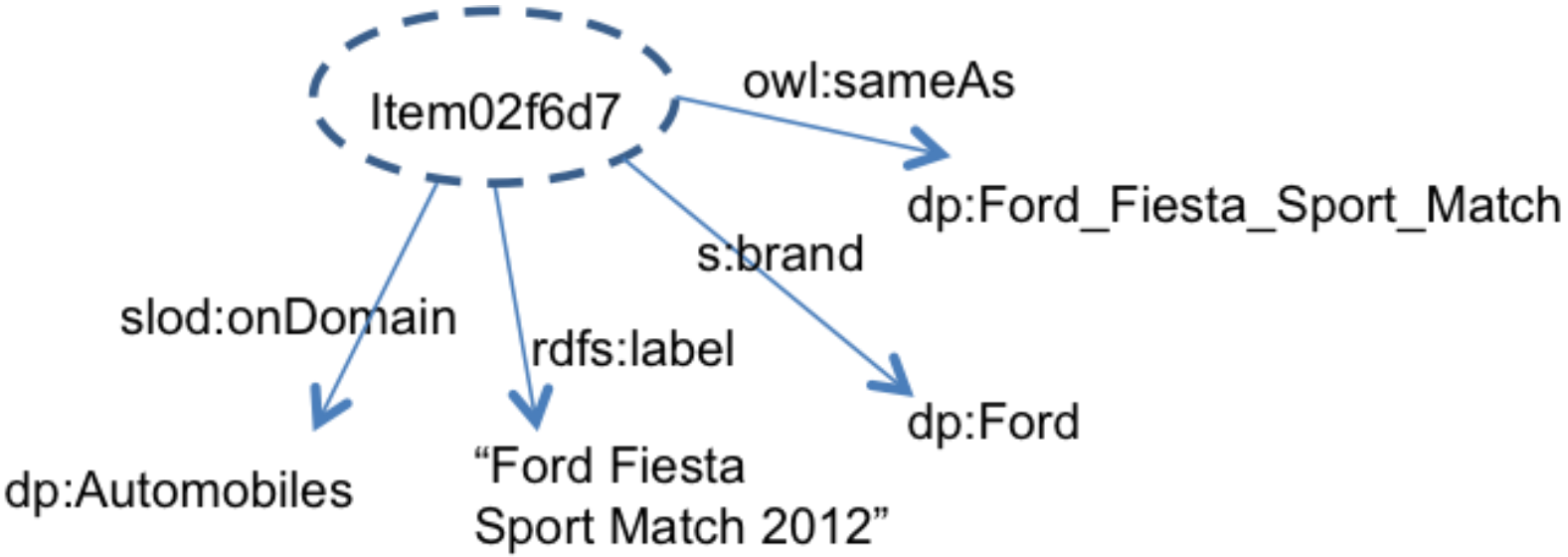
- eShops
- Search Engines
- Blogs
- Social Networks

EXAMPLE IN SLOD-BI

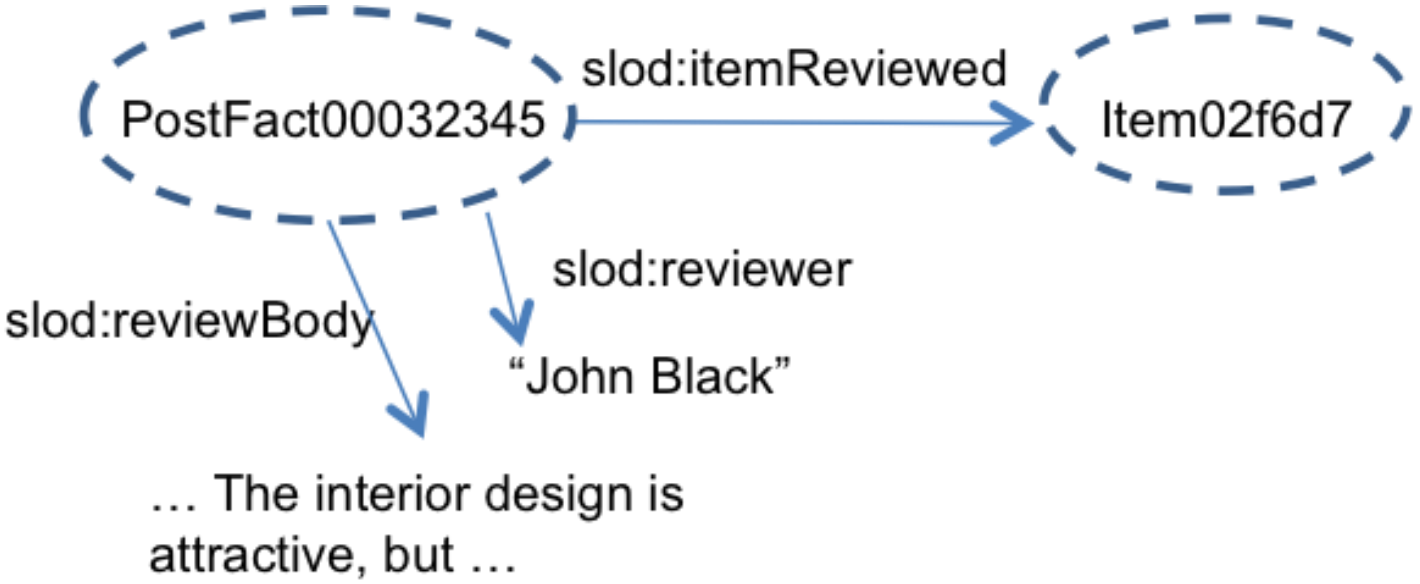


Namespaces
dp: → dbpedia.org
s: → schema.org
slod:, rdfs:, owl:

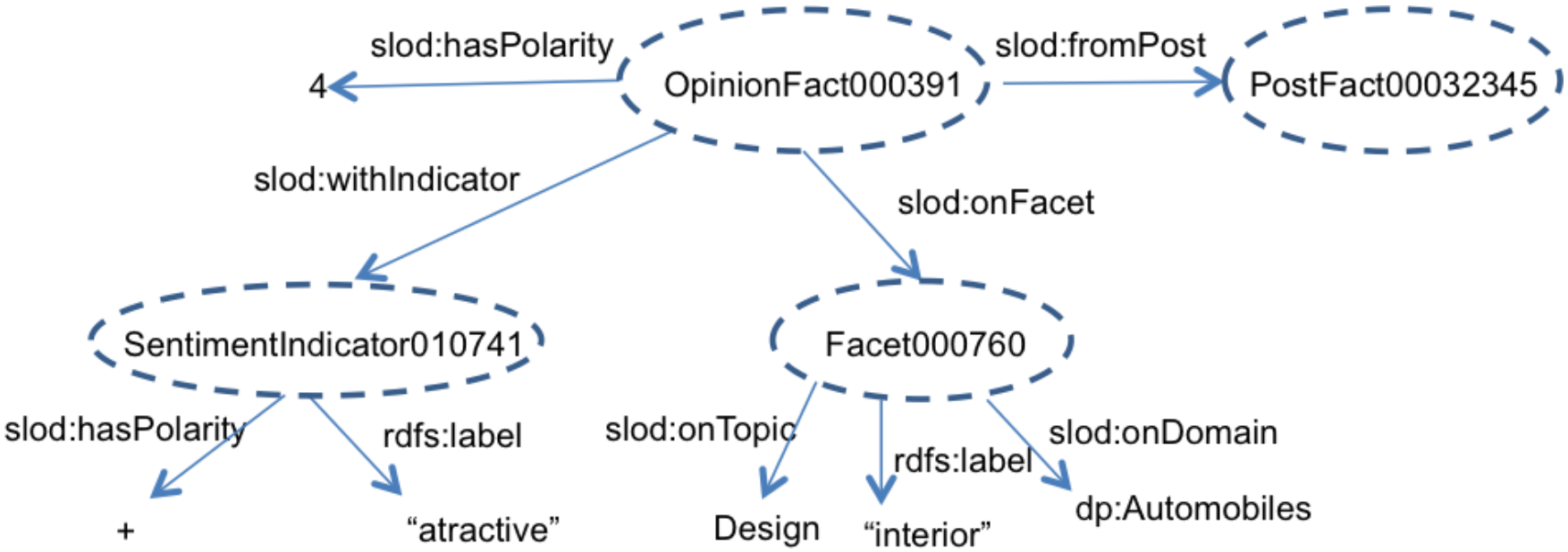
ITEMS


Item02f6d7
owl:sameAs
dp:Ford_Fiesta_Sport_Match
s:brand
slod:onDomain
rdfs:label
dp:Ford
dp:Automobiles
“Ford Fiesta
Sport Match 2012”

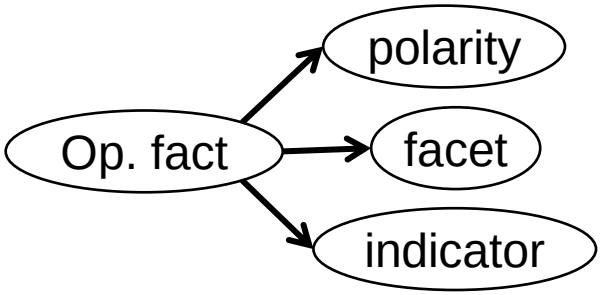
POST FACTS

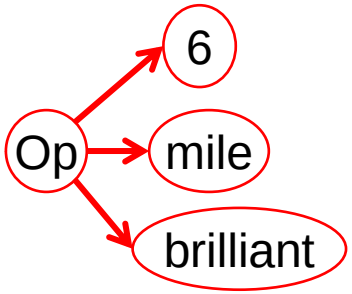

PostFact00032345
slod:itemReviewed
Item02f6d7
slod:reviewer
slod:reviewBody
"John Black"
… The interior design is attractive, but …

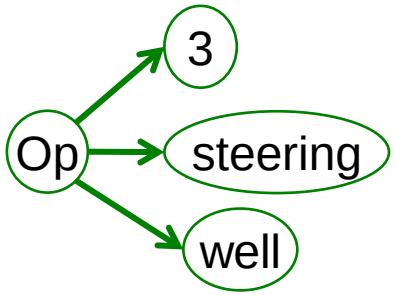
OPINION FACTS

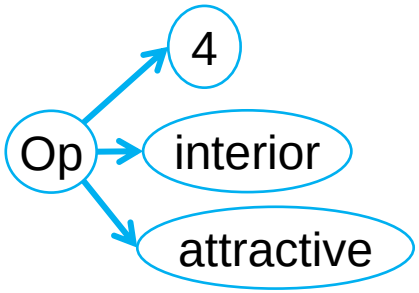


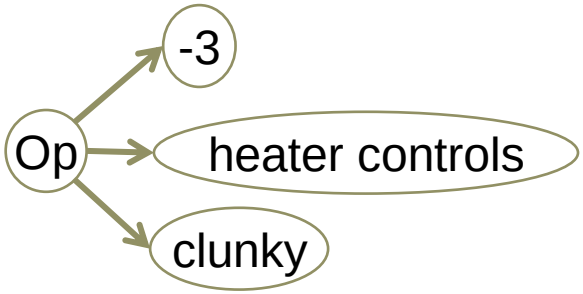
EXAMPLE OF REVIEW



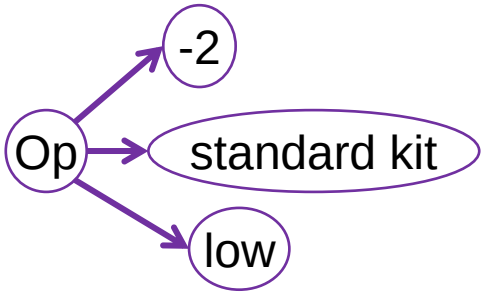


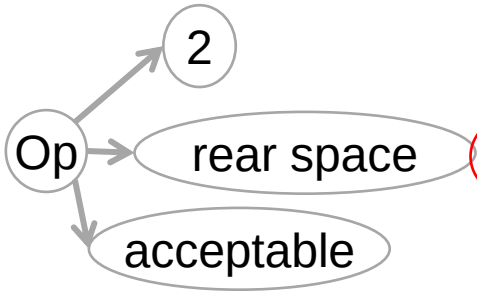


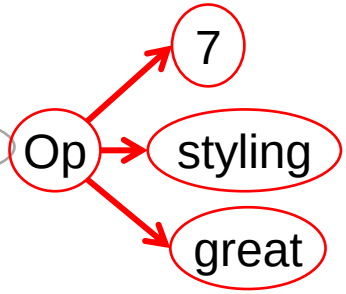


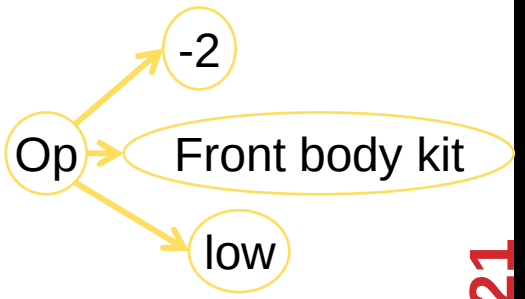


I got my car 2 months after passing my driving test, and based on the first 4 months and 9000 miles, I can conclude its brilliant … The steering is direct, full of feel, well weighted and communicative. The interior design is attractive, but clunky heater controls dissapoint. Standard kit is low, ... Rear space is acceptable for this type of car, … The styling is great, and has introduced me to loads of new people who spoke to me because they got attracted to such a swanky car, just beware, the front body kit is low.







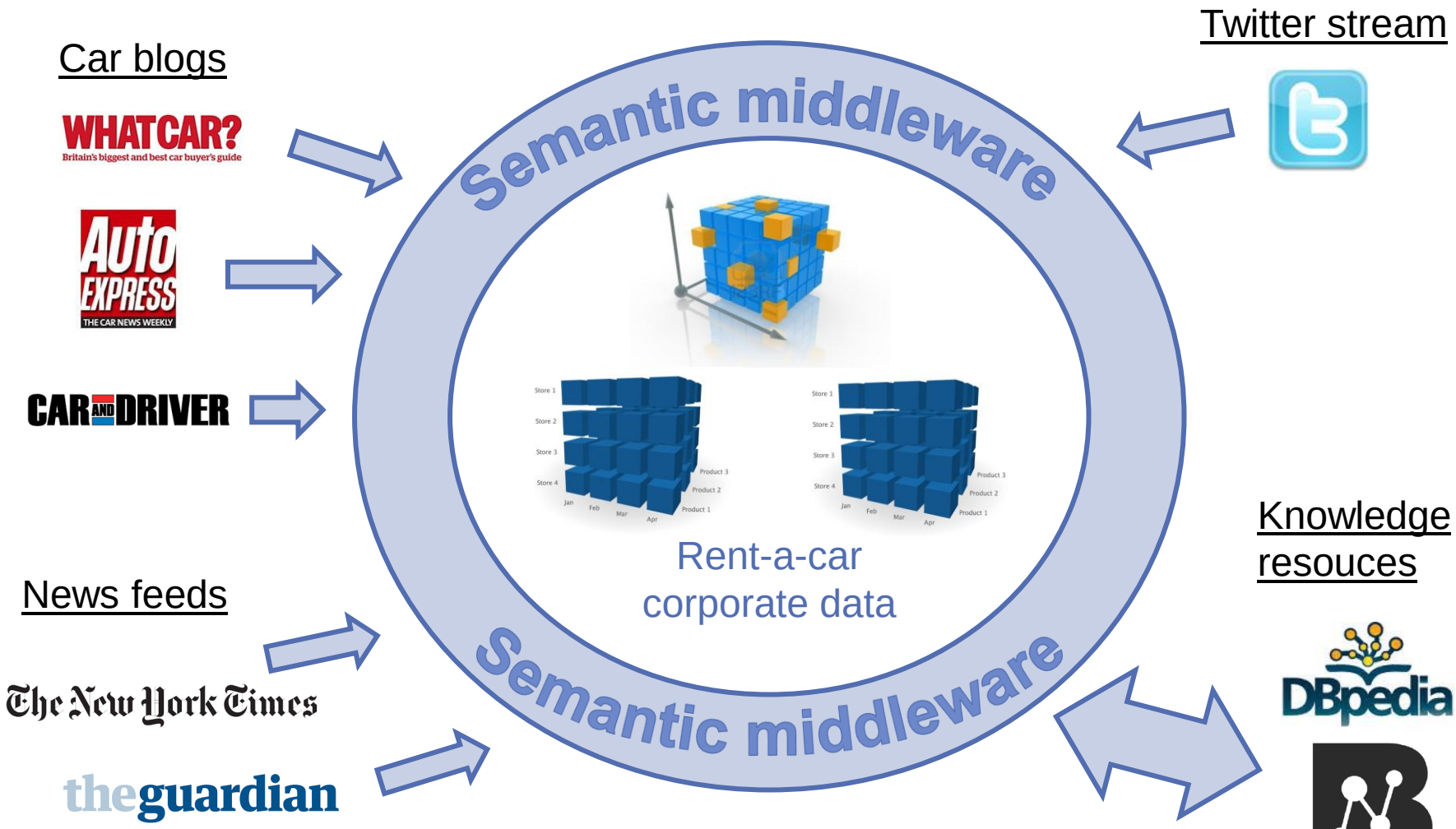


HOW IS TEXTUAL DATA TRANSFORMED TO SLOD-BI PATTERNS?

ETLink

- **Text pre-processing**
- **Sentiment analysis (ML and NLP)**
  - Vocabularies construction (García-Moya et al., 2013)
    - Facets (adapted to the domain)
      - Scarce LOD datasets of facets subject to opinions
      - Depend on the domain
    - Sentiment indicators (context-dependent)
      - On domain:
        - “unexpected” for movies (+), electronic devices (-)
      - On facet:
        - “long delay between shots” (-), “long battery life (+)
- **Automatic semantic annotation and linking of data**
  - Tailored semantic annotation (Berlanga et al., 2015)

CONTEXT-AWARE BI INFRASTRUCTURE



DEMO

http://krono.act.uji.es/EBISS/

- Data sources
- Open IE (triplify)
- Semantic annotation
- Sentiment analysis

SLOD-BI SENTIMENT TOPICS

Facets classified into 6 sentiment topics

- Comfortability: *“inside”, “reflection”, “room”, “climate”, “park”, “arm”, …*
- Safety: *“reliability”, “cruise”, “weather”, “snow”, “safety”, …*
- Driving perception: *“driving”, “visibility”, …*
- Design: *“interior”, “plastic”, “alloy”, “styling”, “dashboard”, “leather”, …*
- Mechanical issues: *“engine”, “breaks”, “fuel”, “wheel”, “mpg”, “gearbox”, “pedal”, …*
- Price: *“price”, “cost”, “money”, …*

DEMO

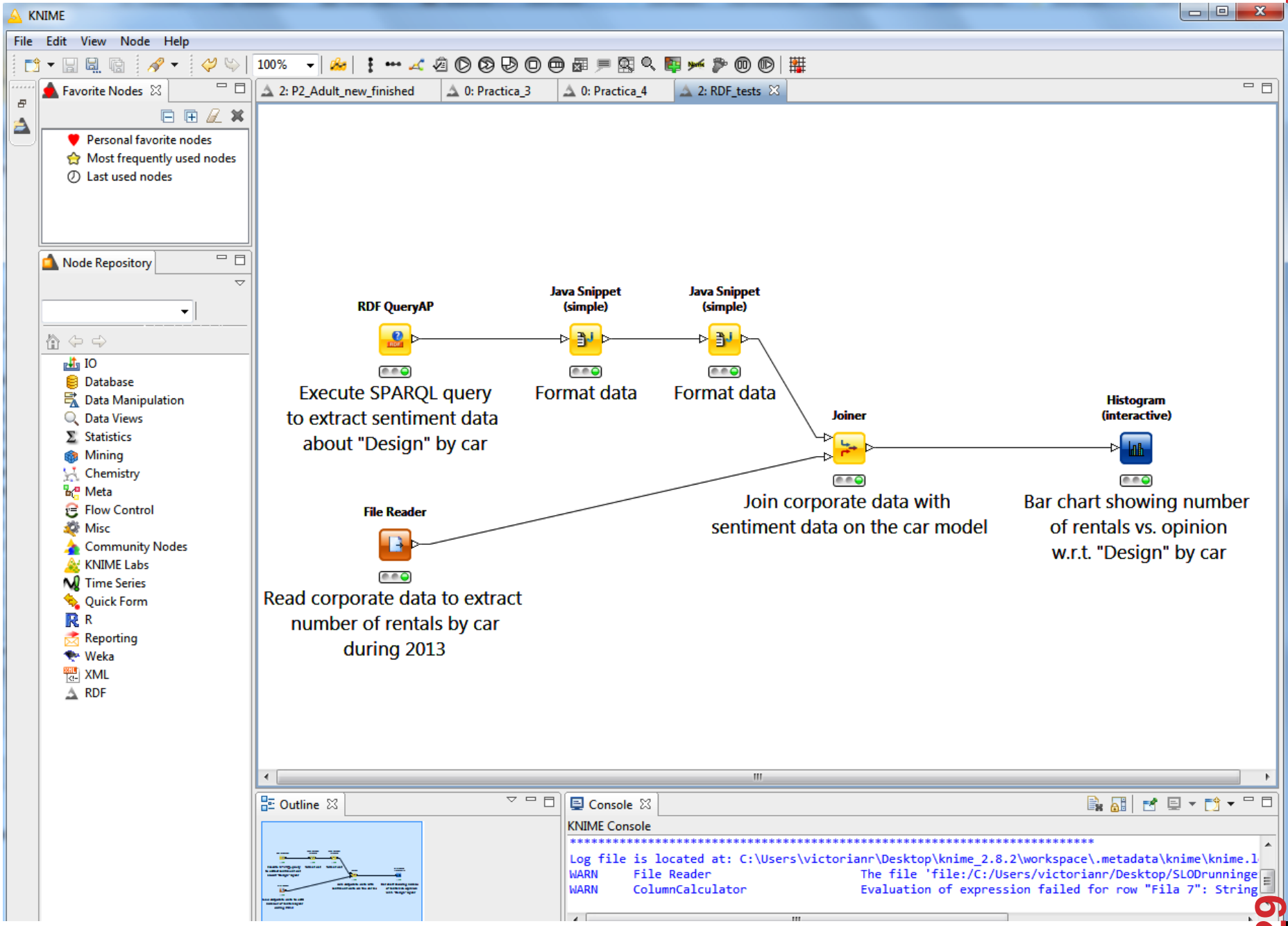
http://krono.act.uji.es/EBISS/

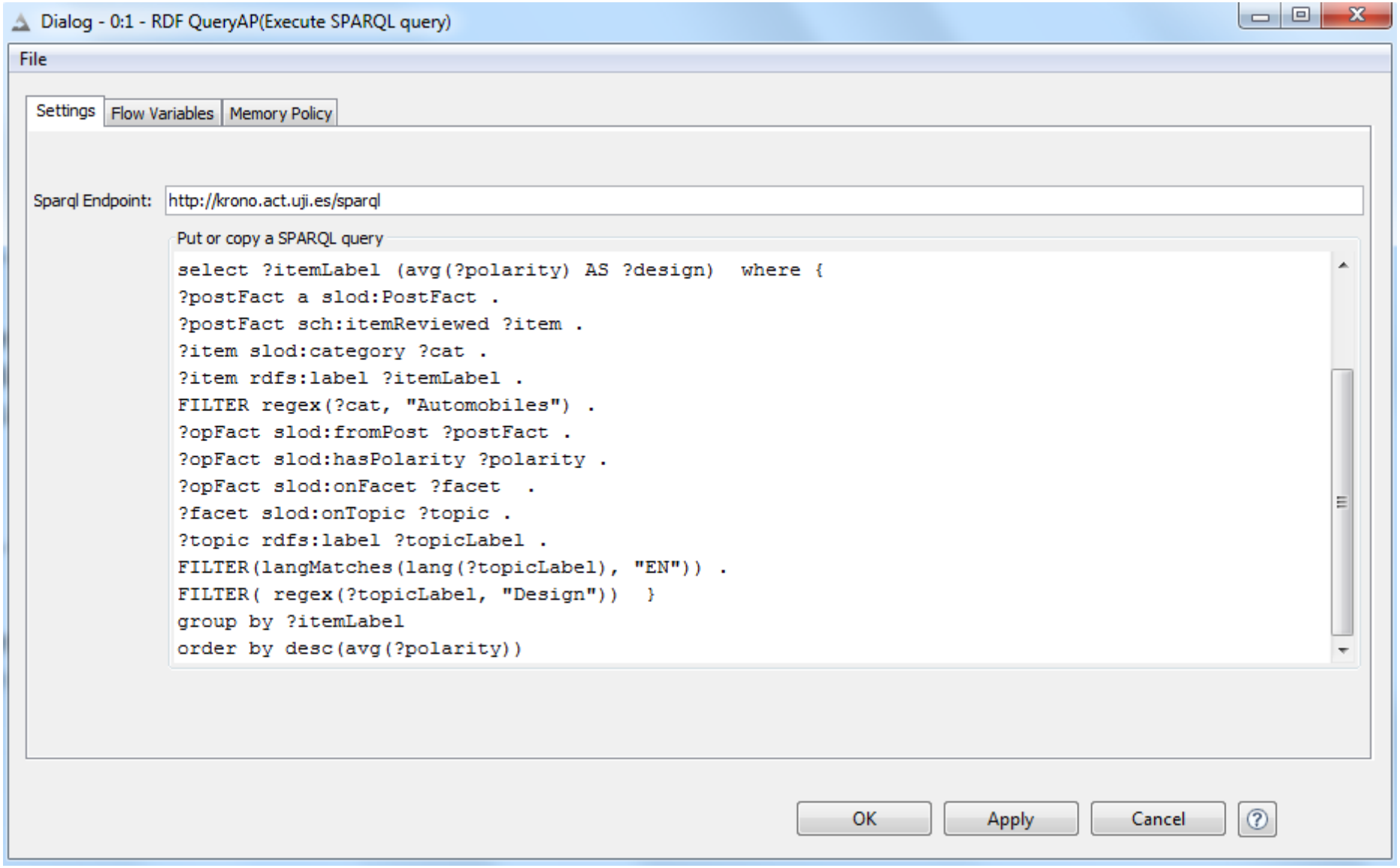
- SLOD-BI sparql endpoint
- SLOD-BI demo examples

SPARQL QUERY

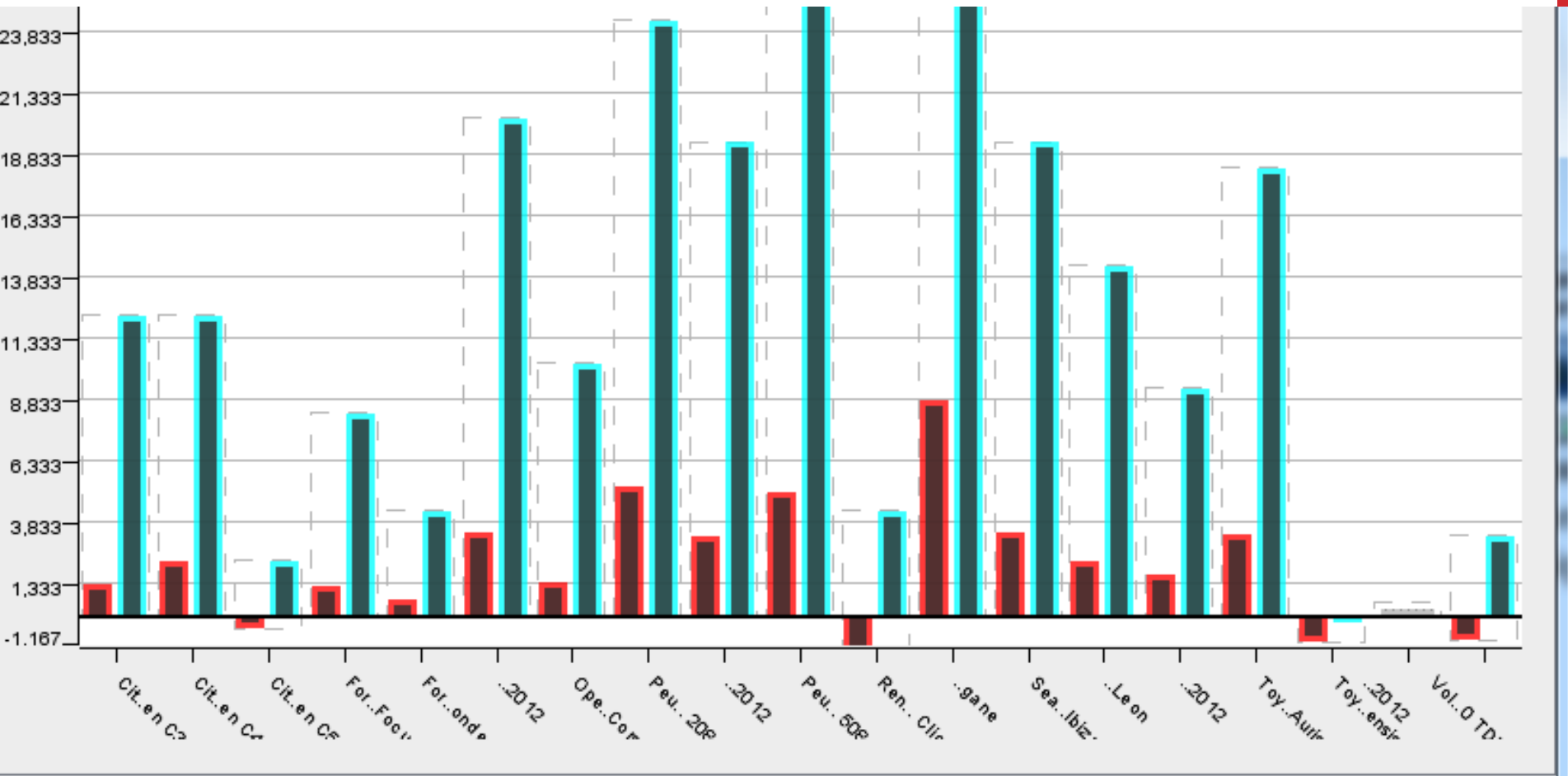
```
prefix slod: <http://krono.act.uji.es/datasets/ontologies/slodonto.rdf#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix rdfdata: <http://rdf.data-vocabulary.org/#>
prefix sch: <http://schema.org/>
select str(?itemLabel) str(?topicLabel)  avg(?polarity)  where {
   ?postFact a slod:PostFact .
   ?postFact sch:itemReviewed ?item .
   ?item rdfs:label ?itemLabel .
   ?item slod:category ?cat .
   FILTER regex(?cat, "Automobiles") .
   ?opFact slod:fromPost ?postFact .
   ?opFact slod:withIndicator ?sentIndicator .
   ?opFact slod:hasPolarity ?polarity .
   ?opFact slod:onFacet ?facet .
   ?facet slod:onTopic ?topic .
   ?topic rdfs:label ?topicLabel .
   FILTER(langMatches(lang(?topicLabel), "EN"))  }
group by ?itemLabel ?topicLabel
order by ?itemLabel
```

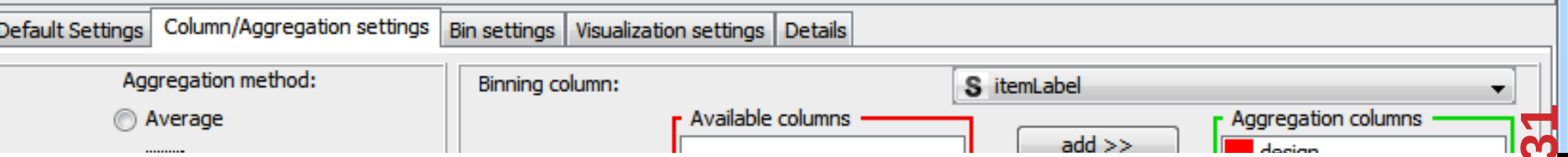
INTEGRATION OF CORPORATE DATA AND SOCIAL DATA USING KNIME


KNIME
File Edit View Node Help
100%
Favorite Nodes
Personal favorite nodes
Most frequently used nodes
Last used nodes
Node Repository
IO
Database
Data Manipulation
Data Views
Statistics
Mining
Chemistry
Meta
Flow Control
Misc
Community Nodes
KNIME Labs
Time Series
Quick Form
R
Reporting
Weka
XML
RDF
2: P2_Adult_new_finished
0: Practica_3
0: Practica_4
2: RDF_tests
RDF QueryAP
Execute SPARQL query to extract sentiment data about "Design" by car
Java Snippet (simple)
Format data
Java Snippet (simple)
Format data
Joiner
Join corporate data with sentiment data on the car model
Histogram (interactive)
Bar chart showing number of rentals vs. opinion w.r.t. "Design" by car
File Reader
Read corporate data to extract number of rentals by car during 2013
Outline
Console
KNIME Console
Log file is located at: C:\Users\victorianr\Desktop\knime_2.8.2\workspace\.metadata\knime\knime.l
WARN File Reader The file 'file:/C:/Users/victorianr/Desktop/SLODrunninge
WARN ColumnCalculator Evaluation of expression failed for row "Fila 7": String


Dialog - 0:1 - RDF QueryAP(Execute SPARQL query)
File
Settings
Flow Variables
Memory Policy
Sparql Endpoint:
http://krono.act.uji.es/sparql
Put or copy a SPARQL query
select ?itemLabel (avg(?polarity) AS ?design) where {
?postFact a slod:PostFact .
?postFact sch:itemReviewed ?item .
?item slod:category ?cat .
?item rdfs:label ?itemLabel .
FILTER regex(?cat, "Automobiles") .
?opFact slod:fromPost ?postFact .
?opFact slod:hasPolarity ?polarity .
?opFact slod:onFacet ?facet .
?facet slod:onTopic ?topic .
?topic rdfs:label ?topicLabel .
FILTER(langMatches(lang(?topicLabel), "EN")) .
FILTER(regex(?topicLabel, "Design")) }
group by ?itemLabel
order by desc(avg(?polarity))
OK
Apply
Cancel

NUMBER OF CAR RENTALS DURING 2013 (BLUE) VS. OPINION ABOUT DESIGN (RED)





CHALLENGES FACED

- **Static vs. dynamic infrastructure**
  - Continuous monitoring of external data
  - Continuous training of ML algorithms

- **SPARQL (endpoints)**
  - Limits usability (expressivity → complex queries)
    - Suggestion of MD queries →(Nebot et al., 2014)
  - Availability & scalability issues
    - Timeouts, limited results, …
  - Federated queries are not a reality yet

REFERENCES

- García-Moya, L., Anaya-Sánchez, H. and Berlanga, R. (2013). A Language Model Approach for Retrieving Product Features and Opinions from Customer Reviews. *IEEE Intelligent Systems*, 28(3), 19-27.
- Berlanga, R., Nebot, V. and Pérez, M. (2015) Tailored semantic annotation for semantic search. *Web Semantics: Science, Services and Agents on the World Wide Web*, 30, 69-81.
- Nebot, V. and Berlanga R. (2014) Towards MD analytical stars from Linked Data. *In Proc. of KDIR 2014*, 117-125.

THANKS FOR YOUR ATTENTION

QUESTIONS ARE WELCOME!

Rafael Berlanga: berlanga@uji.es
Victoria Nebot: victoria.nebot@uji.es