Contexts for Data Quality Assessment and Cleaning

Leopoldo Bertossi[⋆]

Carleton University
Ottawa, Canada

⋆: Faculty Fellow IBM CAS

The Issues, the Vision, and the Approach

Data about the temperatures of patients at a hospital

TempNoon

	Patient	Value	Time	Date
1	Tom Waits	38.5	11:45	Sep/5
2	Tom Waits	38.2	12:10	Sep/5
3	Tom Waits	38.1	11:50	Sep/6
4	Tom Waits	38.0	12:15	Sep/6
5	Tom Waits	37.9	12:15	Sep/7

Are these quality data?

If not, what is to be cleaned?

It depends ...

Actually the table is supposed to contain *temperature measurements for Tom taken at noon by a certified nurse with an oral thermometer*

Are these quality data? We still do not know ...

Maybe we can say something about the time: It may be good enough that the time is "around noon" (meaning?)

Questions about the quality of data like these make sense in a broader setting

The quality of the data depends on "the context"

A context allows us to make sense of the data, assess the data, etc.

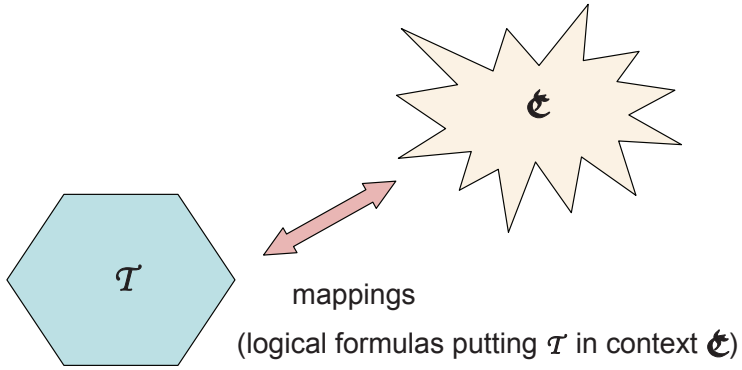
A precise, formalized, general, and usable notion of context is still missing

In particular for applications in data management, and data cleaning

Our vision for a general theory of context:

- A logical theory $\mathcal{T}$ is the one that has to be "put in context"

  For example, a relational database can be seen as a theory



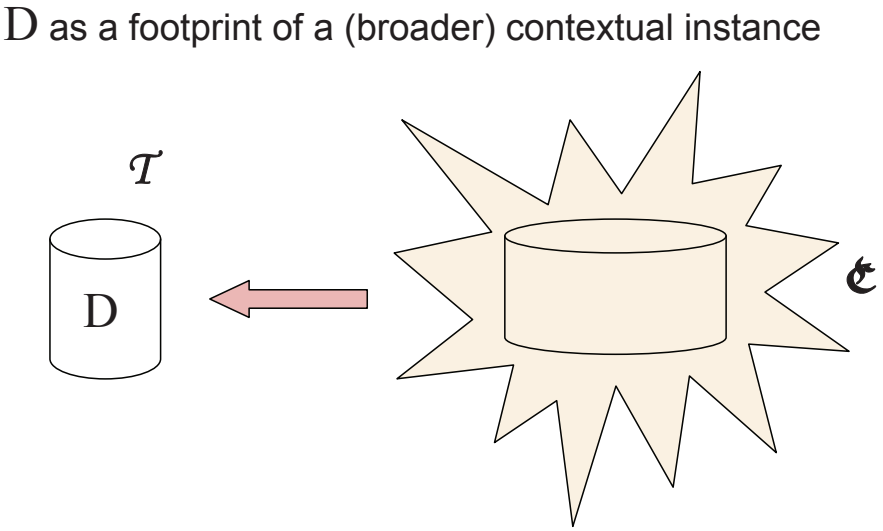
- The context is another logical theory $\mathfrak{C}$

  For example, an ontology, a virtual data integration system

- The connection between $\mathcal{T}$ and $\mathfrak{C}$ established through: connection predicates, possibly shared, and mappings

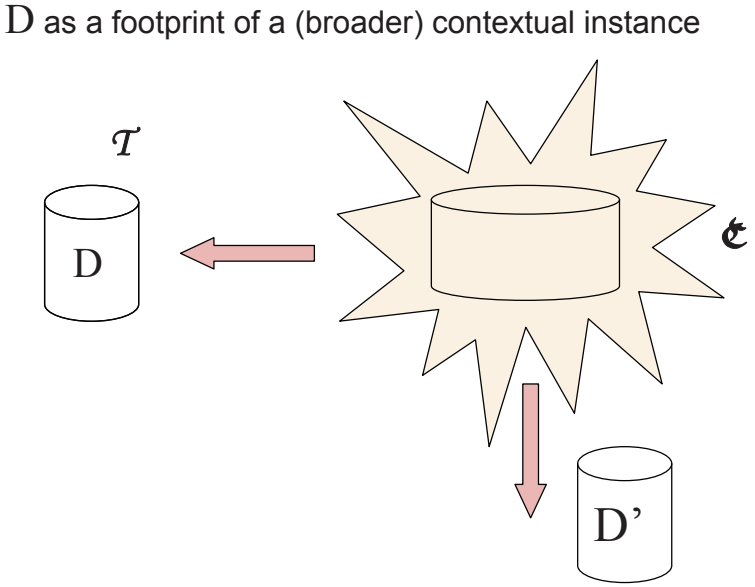
More concretely:

(A)



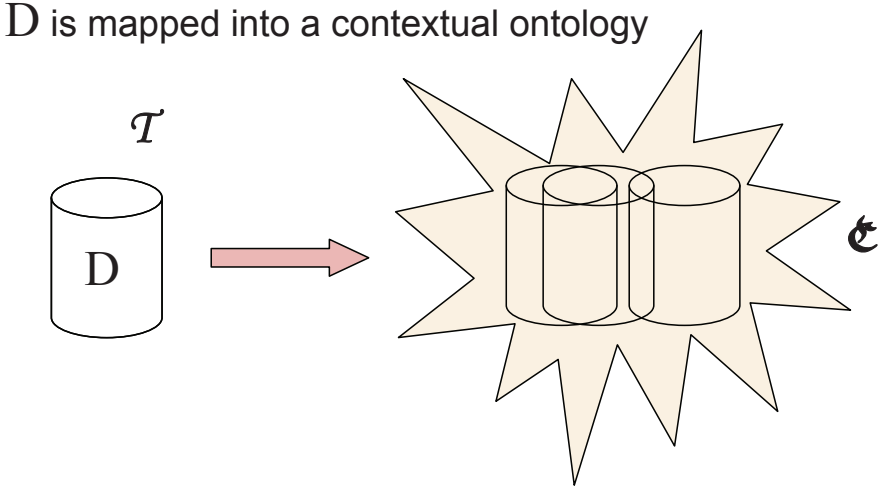
Data in $\mathfrak{C}$ (including D) is analyzed/cleaned

According to additional data available in or accessible from $\mathfrak{C}$; and quality criteria defined in $\mathfrak{C}$



A new version of D is obtained and can be compared with D for quality assessment

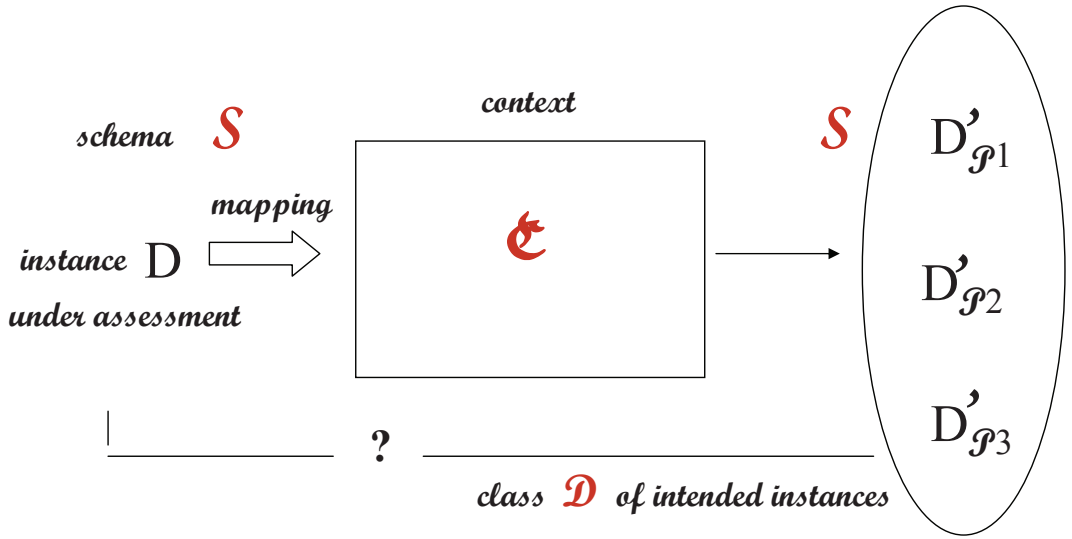
(B)



In principle several versions of D can be obtained at the contextual level

Depending on the mapping, assumptions about the sources of data (completeness?), availability of (partial) data at the context, etc.

Quality criteria are imposed at the contextual level as before

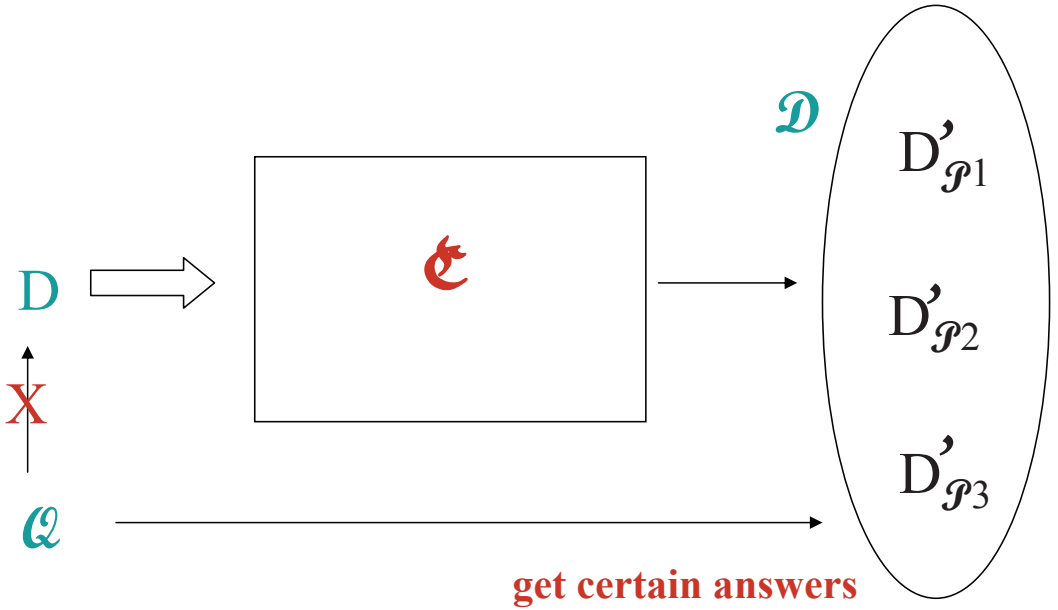


Quality of D can be measured through a distance to a class $\mathcal{D}$ of quality versions of it

This framework opens the ground for "quality query answering"

Given a query $\mathcal{Q}$ posed to original, dirty D

Quality answers from D to $\mathcal{Q}$ are certain wrt class $\mathcal{D}$



Issues:

- Data quality assessment vs. quality query answering
- Computation of quality answers
- Use of external sources at contextual level

Publications:

- Bertossi, L., Rizzolo, F. and Lei, J. "Data Quality is Context Dependent"
  Proc. WS on Enabling Real-Time Business Intelligence (BIRTE 2010). Collocated with VLDB 2010. Springer LNBIP 48, 2011, pp. 52-67.

- Bertossi, L. and Rizzolo, F. "Contexts and Data Quality Assessment"
  Submitted as invited paper to special issue on data quality of *Information Systems* , Feb. 2012

  `http://people.scs.carleton.ca/~bertossi/papers/journal10.pdf`

Multidimensional Contexts (ongoing research)

Temperature data at a hospital

Doctor requires temperatures taken with oral thermometer

TempNoon@Ward

Patient	Value	Time	Date	Ward
Tom Waits	38.5	11:45	Sep/5	1
Tom Waits	38.2	12:10	Sep/5	1
Tom Waits	38.1	11:50	Sep/6	1
Tom Waits	38.0	12:15	Sep/6	1
Tom Waits	37.9	12:15	Sep/7	1
Lou Reed	37.9	12:10	Sep/5	2
Lou Reed	37.6	12:05	Sep/6	2
Lou Reed	37.6	12:05	Sep/7	2

Doctor expects this to be reflected in the table

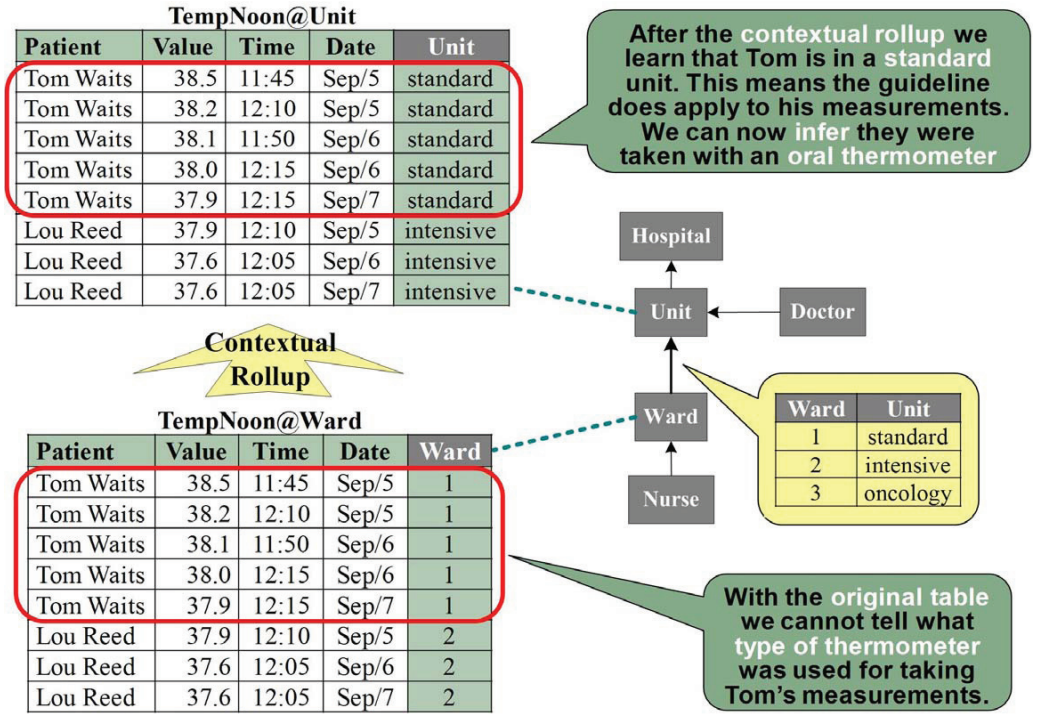
Table does not contain the information to make this assessment

An external context can provide that information, making it possible to assess the given data

The database under assessment is mapped into the context, for further analysis and cleaning

The information in the context is commonly of a multidimensional nature

- Hospital guideline (a rule): *temperatures of patients in standard care units have to be taken with an oral thermometer*



- A specification of the hierarchical and dimensional hospital structure

Other dimensions could be easily considered

Generating multidimensional (MD) contextual information

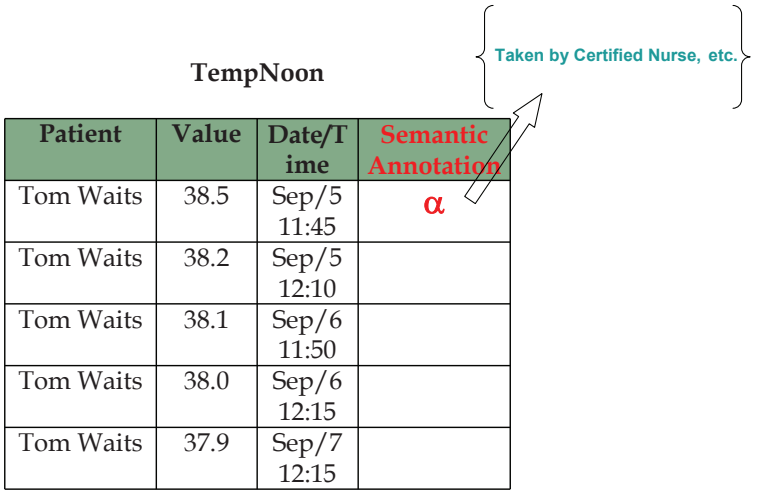
For additional and finer-granularity data quality assessment

We embed (an extension of) the Hurtado-Mendelzon model for MD DBs into our ontological context

Contextual roll-up can be used to access missing information at certain level, by lattice navigation

Mechanisms for querying database with taxonomies could be applied/embedded (Martinenghi & Torlone; ER10)

Making Sense of Data (ongoing research)



Use formal annotations to express sense or meaning of data

α is a symbolic, machine processable sentence

α expressed in terms that are described in the context by means of an ontology

These "sense predicates" can be used to define and apply higher-level quality predicates (cf. [Jiang, Borgida, Mylopoulos; ER'08])