Carleton
UNIVERSITY

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

Is this 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*

Is this 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 this data's quality make sense in a broader setting

The quality of the data depends on "the context"

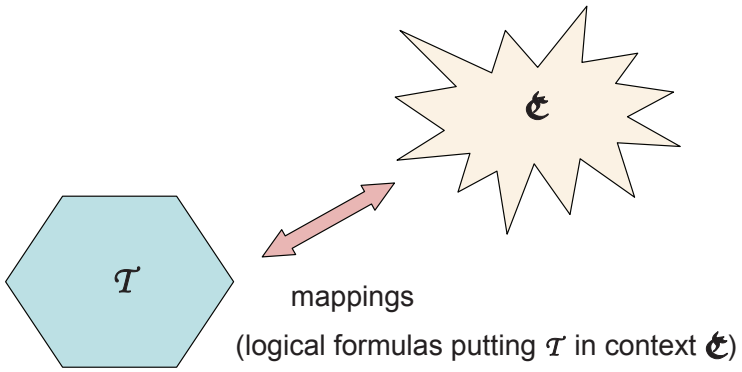
It allows us to make sense of the data, assess the data, etc.

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

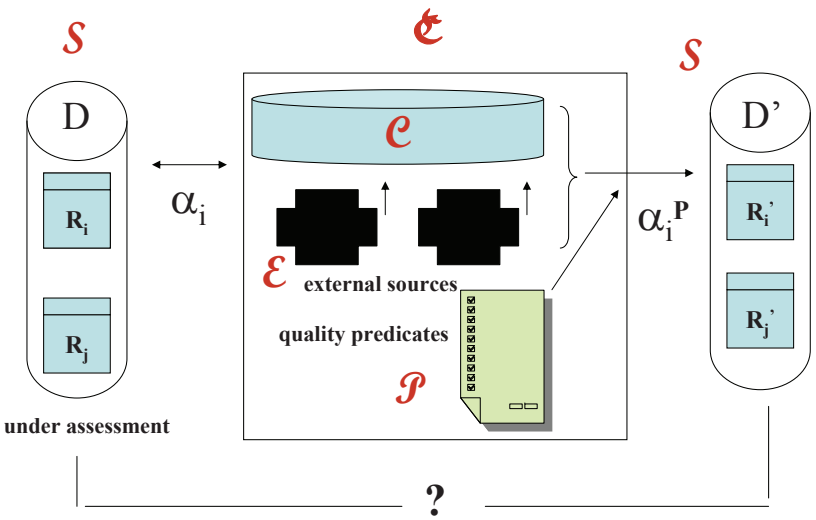
Our vision for a general theory of context, in particular for applications in data management, is as follows:

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



Contexts in Data Quality Assessment: (VLDB BIRTE WS 2010)



- Instance D is under assessment
- Context $\mathcal{C}$ (including $\mathcal{E}$, $\mathcal{P}$) on the RHS, as a virtual/(semi)materialized data integration system
- The α_i in between are the mappings, like in VDISs or data exchange
- The C_i are contextual predicates/relations

- Mappings to external sources E_i and quality predicates P_i

- D' (with schema as in $\mathcal{S}$) contains "ideal" contents for relations in D, as views

- Predicates in D' can be materialized through data in the R_i and additional massage via $\mathcal{C}$ (mapping composition at work)

- Quality-aware (QA) query answering about (or from) $\mathcal{S}$ can be done on top of D'

Techniques for query answering in VDISs can be applied (specially if D' is not materialized)

- Quality assessment of D can be done by comparing its contents with D' (there are some measures of distance between instances and collections thereof)

A particular case of QA query answering

Multidimensional Contexts (ongoing research)

Temperature data at a hospital

Doctor requires temperatures taken with oral thermometer

Doctor expects this to be reflected in the table,

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

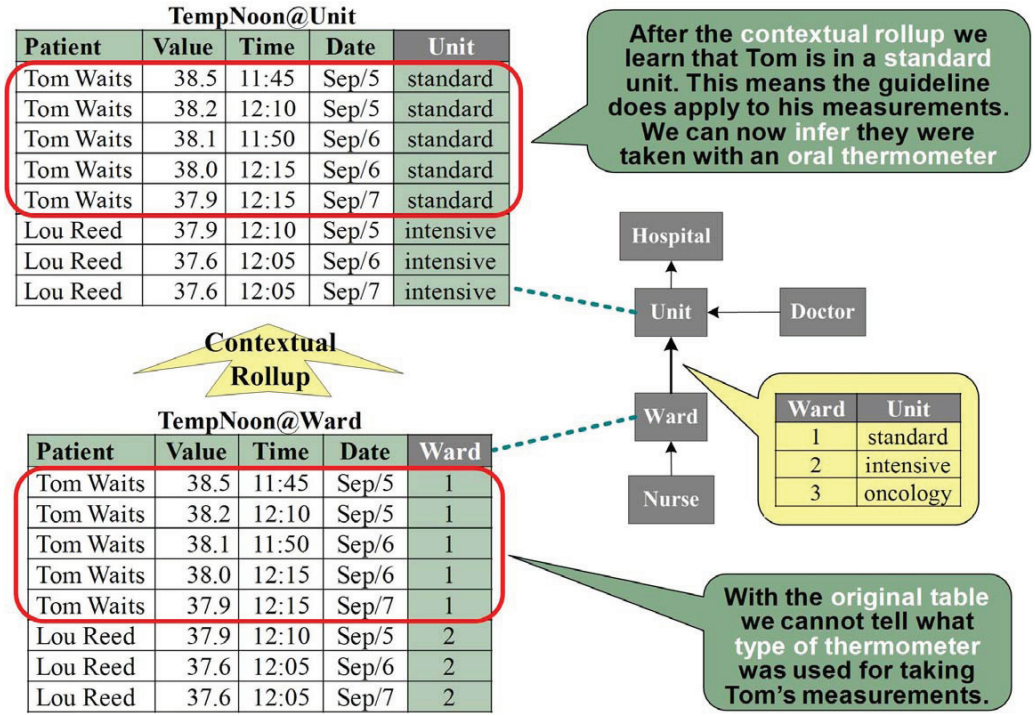
but the latter does not contain the information to make this assessment

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:

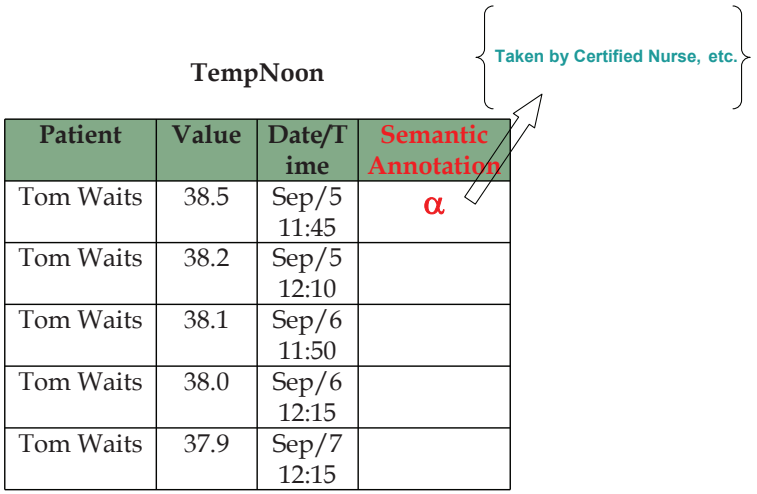
- Hospital guideline: *the temperature of patients in standard care units have to be taken with an oral thermometer*



- 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

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 quality predicates of the kind introduced in [Jiang, Borgida, Mylopoulos; ER'08]

Extra Slides

Contexts So Far

We find the term "context" in several places in computer science: databases, semantic web, KR, mobile applications, ...

Usually used for "*context aware* ... search, databases, applications, devices, ..."

Most of the time there is no explicit notion of context, but only some algorithms that take into account (or into computation) some contextual notions

Usually, time and geographic location, i.e. particular *dimensions*, and not much beyond

In our opinion, there is a lack of research in the area

A precise and formalized notion of context is still missing

There has been some research:

- Contexts in ontologies and semantic web

  Contexts are left implicit and logic programs are used to bridge them

- Contexts in KR

  They are denoted at the object level and a theory specifies their properties and dynamics

  It is possible to talk about things holding in certain (named) contexts

- Contexts in data management

  Usually in connection with specific dimensions of data, like time and place

A general theory of context has still to be developed

In particular for applications in data management

In the general formalization (page 4), the mappings and the way they are processed (reasoned about and from) have to be such that they enable what we expect from a context, i.e support for the following tasks

- Capturing and narrowing down semantics
  - By defining in $\mathcal{C}$ predicates that are used in $\mathcal{T}$ (e.g. "time close to noon")
  - Contributing in $\mathcal{C}$ with additional constraints for predicates used in $\mathcal{T}$, e.g. integrity constraints for table TempNoon)
- Term disambiguation (related to meaning)
- Dimensions for analysis and understanding of $\mathcal{T}$'s knowledge (generalizing multidimensional DBs, DWHS)
- Specifying and using notions of relevance

- Explanation, diagnosis, causality
- Capturing commonsense assumptions and practices
- Assessment, e.g. quality