Managing Temporal Clinical Data and Knowledge

A logic-based approach to represent temporal dependencies and trends with multiple granularities

Rosalba Rossato

Dipartimento di Informatica, Università degli Studi di Verona, Italy

AIME 2005 Doctoral Consortium

July 23, 2005 - Aberdeen, Scotland

Outline

- Introduction
- Background and Related Work
- The proposal
  - Temporal Functional Dependecies description
  - Trends description
- Conclusions and ongoing work

Introduction

- A database is a collection of data, typically organized to make common retrievals simple and efficient
- **Functional dependencies** (FDs) are widely used dependencies in the logical design of relational databases
  - **Static dependencies**: they are intended for snapshot databases which *do not manage any temporal dimension* of data
- **Temporal databases** allow one to describe the temporal evolution of information by associating one or more temporal dimensions with the stored data
  - **valid time**: the time when a *fact is true* in the modeled reality
  - **transaction time**: the time when a *fact is current* in the database
- **Temporal functional dependencies** (TFDs) have been used in temporal databases *to constrain the temporal relationships among data*

[Vianu'87,Jensen'96,Wijsen'95,Wang'97,Wijsen'99]

Introduction

- In the real world, when we describe a fact, we generally use a specific time unit
  - Follow-up visits are usually planned on working days
  - ICU events are described in terms of minutes and seconds
- A **time granularity** can be viewed as a partitioning of a temporal domain in groups of indivisible units called *granules*
- **Temporal functional dependencies with multiple granularities** describe *integrity constraints* which can include *different user-defined calendars*, i.e. finite sets of temporal granularities

Introduction

- In the context of clinical databases, the knowledge of the time during which the information is valid holds a crucial role
  - the information about the **past** and the **future** are important as the information about the **present**
- It is interesting to describe either temporal properties or temporal trends having specific time features
  - *the difference between two consecutive values of a patient's weight does not exceed a given value*
  - *there exists an increase of a parameter P during the administration of a particular drug, and only a finite number of exceptions are allowed in this time period*

Goals

- **The problem:** the existing proposals about FDs and TFDs capture **temporal features** which are **slightly different** one from each other
    - All the considered temporal features must be taken into account: the database has to verify *at each time point a FD* and it has to verify *constraints related to its updates*
- **Goal:** express in a **homogeneous way the FDs and the TFDs** proposed in the literature, and **express new temporal functional dependencies**

Goals

- **The problem:** at the best of our knowledge, few efforts have been devoted to providing a general framework allowing the specification of different trend features w.r.t. several time granularities
  - The existing proposal, related to *trend dependencies* (TDs), captures a class of data evolution constraints in the form:

    “The heigth of a patient should never decrease”

- **Goal:** define **temporal trends** involving different time granularities for expressing both **contraints** and **queries**

- Background and Related Work

Relations and FDs

- A **relation schema** is defined by a *relation name* and a *set of attributes*; the map dom associates a *domain of values* to each attribute
- A **tuple** over a set of attributes X is a total function from X into $\mathrm{dom}(X)$. A **relation** r over X is a set of tuples over X
- $t[a]$ represents the *value of attribute* a in the tuple t

Definition: A functional dependency (FD) is a statement $A \rightarrow B$, where A and B are sets of attributes. The functional dependency $A \rightarrow B$ is satisfied by a relation r over X, with $A, B \subseteq X$, if and only if for every $s, t \in r$:

$$s[A] = t[A] \rightarrow s[B] = t[B]$$

Example: If P_Id and BG are the attributes representing the *patient identifier* and the *blood group* of the patient, then the following FD must be true:

$$P\_Id \rightarrow BG$$

Representing time granularities

Definition. *Let* T *be the time domain and* I *the domain of the time granularity called index set. A granularity* $\mathcal{G}$ *is a mapping* $\mathcal{G} : \mathrm{I} \rightarrow 2^{\mathrm{T}}$ *such that:*

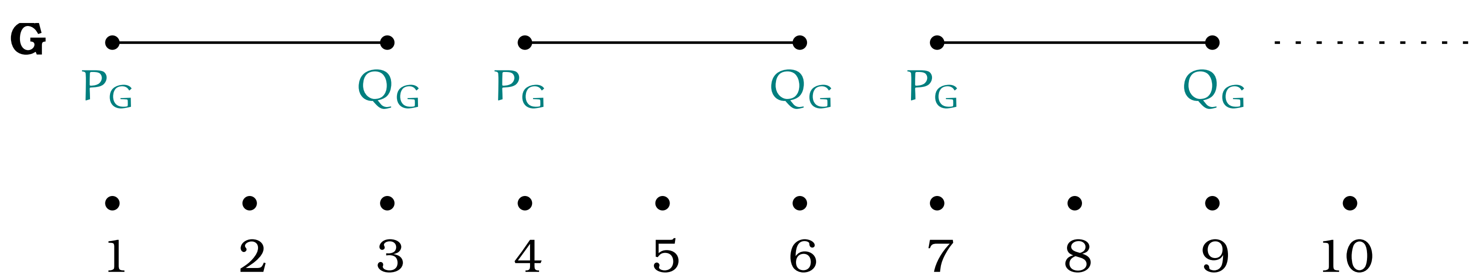
1. *for all* $i < j$, *for any* $n \in \mathcal{G}(i)$ *and* $m \in \mathcal{G}(j)$, $n < m$;
2. *for all* $i < j$, *if* $\mathcal{G}(j) \neq \emptyset$, *then* $\mathcal{G}(i) \neq \emptyset$.

- I is the domain of a granularity $\mathcal{G}$, called *index set*. The elements of the codomain of $\mathcal{G}$ are called *granules*
- The first condition states that **granules** in a granularity **do not overlap** and *their order is the same as their time domain order*
- The second condition states that the subset of the index set that maps to **nonempty granules forms an initial segment**

Labeled linear time structure for time granularities

- Temporal granularities can be represented by means of a **labeled linear time structure** [Combi,Franceschet,Peron ’04]
  - If $\mathcal{G}$ is a time granularity, then the symbols $P_{\mathcal{G}}$ and $Q_{\mathcal{G}}$ represent the **starting** and **ending point of each granule** of $\mathcal{G}$
  - A *time granularity* is described by means of a **logical PPLTL formula**

Example: Let G be the time granularity such that $G(0) = \{1, 2, 3\}$, $G(1) = \{4, 5, 6\}$, and $G(2) = \{7, 8, 9\}$. The labeled linear time structure representing the time granularity G is:



Temporal Functional Dependencies

- **TFDs without granularity**
  - *Dynamic constraints of Vianu* [Vianu87] describe constraints that the database has to respect during its evolution (sequence of updates)
  - *TFDs of Jensen et al.* [Jensen96] describe constraints which have to be satisfied in each time point
  - *Temporal and Dynamic dependencies of Wijsen* [Wijsen95] represents constraints which have to be satisfied for each time point and for each couple of consecutive time points, respectively
- **TFDs with multiple granularities**
  - *TFDs with granularities of Wang et al.* [Wang et al.'97] describe constraints related to tuples which are valid during the same granule of a given time granularity
  - *TFDs on Complex Objects of Wijsen* [Wijsen'99] extend the notion of TFDs with granularity to temporal databases that include the object-identity

Trends

- **Trend dependencies (TDs) of Wijsen** [Wijsen'01]:
  - Totally ordered infinite set of constants (**dom**, $\leq$) and a set of operators $\mathrm{OP} = \mathcal{P}(\{<, >, =\})$
  - Time is represented by the set $\mathbb{N}$ and *time accessibility relation* (TAR) is used to indicate which tuples have to be compared one with each other
  - *Directed Attribute Set* (DAS) over a set of attributes allows one to describe the attributes which have to be compared and the associated comparing operators
  - A TD is a logical implication $\phi \rightarrow_{\alpha} \psi$ between a couple of DASs (defined over the same set of attributes) with respect to a TAR α

"The heigth of a patient should never decrease"

$$(\mathrm{P\_Id}, =) \rightarrow_{\mathrm{Next}} (\mathrm{Heigth}, \leq)$$

- **Temporal-abstraction task of Shahar** *et al.* [Shahar'97]

Trends

- **Trend dependencies (TDs) of Wijsen** [Wijsen'01]
- **Temporal-abstraction task of Shahar** *et al.* [Shahar'97]: abstractions of data into higher-level concepts meaningful for specific domains
  - An abstraction has the form

    {⟨ parameter, value, context ⟩, interval}

    denoting the logical proposition:

    *"the **parameter** has a particular **value***
    *given a specific interpretation **context**"*

    is interpreted over a specific time ***interval***
  - The goal of the temporal abstraction task is, for example, *to evaluate and summarize* the state of a patient over a period, *to identify problems*, and *to support the management of the therapy plans*

- The proposal

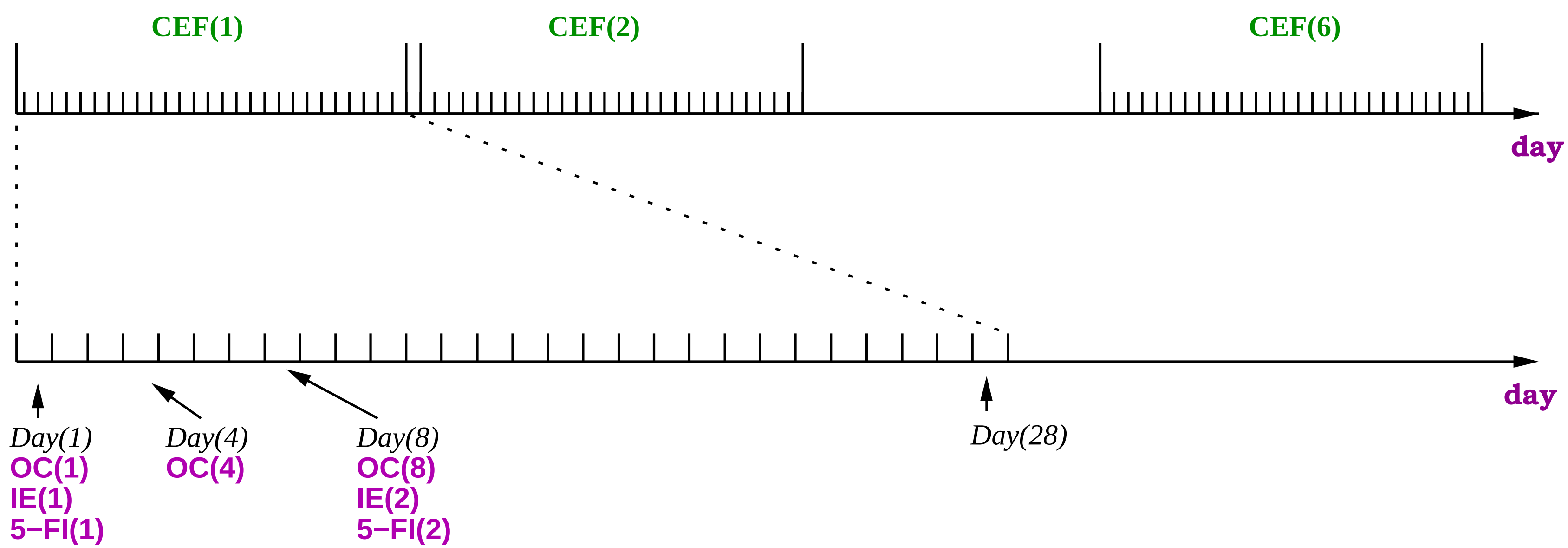
We propose a logical formalism for the definition of TFDs and temporal trends with multiple granularities and focus on its application to the management of clinical data and knowledge

The medical domain

“The recommended CEF regimen consists of 14 days of oral cycloshosphamide and intravenous injection of epirubicin and 5-fluorouracil on days 1 and 8. This is repeated every 28 days for 6 cycles.”

[Canadian Medical Association Journal, 2001]

- Let us assume that OC, IE and 5-FI are the *granularities corresponding to drugs* of the CEF regimen

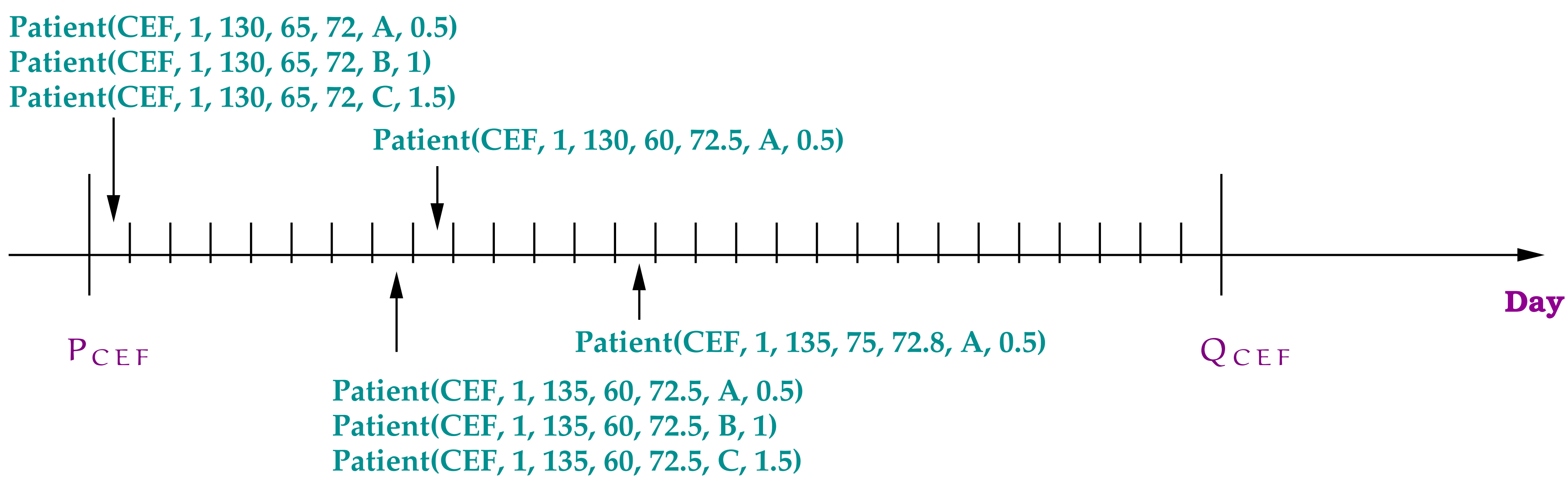


Representing temporal relations

- The temporal relation $\mathit{Patient}$ with atemporal schema

$$\mathit{Patient}(\mathit{Chemo}, \mathit{P\text{-}Id}, \mathit{SBP}, \mathit{DBP}, \mathit{Weight}, \mathit{Drug}, \mathit{Qty})$$

  describes, for each patient, the type of therapy (Chemo), the patient's identifier ($\mathit{P\text{-}Id}$), the value of systolic and diastolic blood pressures (SBP and DBP), the weight (Weight), and the assumed drug (Drug) and the related quantity (Qty)



- TFDs description

TFDs without granularity

- Let now consider the following dynamic dependency of Wijsen:

  "if the value of the attributes P-Id and Drug are preserved between times i and i+1, then the associated type of chemotherapy must be preserved as well".

- The dependency can be described as $\{P\_Id, Drug\}\mathbf{N}\{Chemo\}$
- It can be expressed with the following logical formula:

$$\mathbf{G}\quad \big(\ (\mathrm{Patient}(\mathrm{Ch}, \mathrm{Id}, \mathrm{Sbp}, \mathrm{Dbp}, \mathrm{W}, \mathrm{D}, \mathrm{Q})\ \wedge\ \mathbf{X}(\mathrm{Patient}(\mathrm{Ch}_1, \mathrm{Id}, \mathrm{Sbp}_1, \mathrm{Dbp}_1, \mathrm{W}_1, \mathrm{D}, \mathrm{Q}_1)\ \longrightarrow (\mathrm{Ch} = \mathrm{Ch}_1)\ \big)$$

- Two consecutive tuples having the same values for the attributes P_Id and $Drug$ must have the same value for the attribute $Chemo$

TFDs with granularity

- We now consider the following requirement:

  "a patient cannot assume different quantities of the same drug within a therapy cycle"

- The dependency can be identified as $\mathrm{P\text{-}Id}, \mathrm{Drug} \longrightarrow_{\mathrm{CEF}} \mathrm{Qty}$
- Assuming to have a formula ϕ_{CEF} describing the temporal granularity CEF, the logical formula which translates the dependency is the following:

$$\phi_{\mathrm{CEF}} \wedge \quad \mathbf{G}\Big(\big(P_{\mathrm{CEF}} \wedge \neg Q_{\mathrm{CEF}} \mathbf{U} \big(\mathrm{Patient}(\mathrm{Ch}, \mathrm{Id}, \mathrm{Sbp}, \mathrm{Dbp}, \mathrm{W}, \mathrm{D}, \mathrm{Q}) \wedge \\ (\neg Q_{\mathrm{CEF}} \mathbf{U} \mathrm{Patient}(\mathrm{Ch}_1, \mathrm{Id}, \mathrm{Sbp}_1, \mathrm{Dbp}_1, \mathrm{W}_1, \mathrm{D}, \mathrm{Q}_1))\big)\big) \rightarrow (\mathrm{Q} = \mathrm{Q}_1) \Big)$$

- Two tuples which are valid inside the same granule of the granularity CEF, related to the same patient and having the same value for the attribute Drug, must have the same value for the attribute Qty ($\mathrm{Q} = \mathrm{Q}_1$)

- Trends description

Simple trends

- Simple trends common in the database [Wijsen'98] and AI area [Shahar'97]:
  - **Increase(Relation, Attribute, Rate)**, i.e. the value of the attribute Attribute of the relation Relation increases over the time and the difference between a value and the next value is greater than the value Rate
  - **Decrease(Relation, Attribute, Rate)**, i.e. the value of the attribute Attribute of the relation Relation decreases over the time and the difference between a value and the next value is smaller than the value Rate
  - **State(Relation, Attribute, Constant)**, i.e. the value of the attribute Attribute of the relation Relation is equal to Constant for a given time interval
  - **Stationary(Relation, Attribute, Threshold)**, i.e. the value of the attribute Attribute of the relation Relation over the time is inside a range such that the difference between a value and the next value (for the attribute Attribute) is not greater than a given threshold value Threshold

Examples of trends

- When we describe temporal trends involving specific temporal granularities, we can distinguish two kinds of trends
  - *intragranule trend*: temporal properties expressed by means of a trend have to be satisfied inside a given granule
  - *intergranule trend*: the properties involve different granules

Example. *We want to verify whether a patient, having an* Increase *trend of the parameter* *Weight* *during a cycle of chemotherapy with a given* Rate*, exists*

$$\exists \mathrm{Id}, \mathrm{Ch}, \mathrm{Ch}_1, \ldots \quad \mathbf{F}(\mathrm{P}_{\mathrm{CEF}} \wedge \neg \mathrm{Q}_{\mathrm{CEF}} \mathbf{U}(\mathrm{Patient}(\mathrm{Ch}, \mathrm{Id}, \mathrm{Sbp}, \mathrm{Dbp}, \mathrm{W}, \mathrm{D}, \mathrm{Q}) \wedge$$

$$\mathbf{X}(\mathrm{Patient}(\mathrm{Ch}_1, \mathrm{Id}, \mathrm{Sbp}_1, \mathrm{Dbp}_1, \mathrm{W}_1, \mathrm{D}_1, \mathrm{Q}_1)) \wedge \mathrm{W}_1 \geq \mathrm{W} + \mathrm{Rate}))$$

- The formula is satisfied when *at least* one increase trend between two consecutive values of the $Weight$ parameter, related to the same patient, exists

Examples of trends

- Both for intragranule and intergranule trends, we can discriminate two kinds of satisfiability:
  - *local satisfiability*: allows one to express that a trend must be valid over some time points of the considered granule(s)
  - *global satisfiability*: allows one to express that a trend must be valid over all the considered time points of the considered granule(s)

Example. *We want to verify whether a patient having, for a given treatment, an* Increase *trend of the parameter* Weight *during a whole granule of the granularity CEF, exists*

$$\exists \mathrm{Id}, \mathrm{Ch}.\ \mathbf{F} \quad (\mathrm{P}_{\mathrm{CEF}} \wedge \exists \mathrm{Sbp}, \mathrm{Dbp}, \mathrm{W}, \mathrm{D}, \mathrm{Q}.\ \mathrm{Patient}(\mathrm{Ch}, \mathrm{Id}, \mathrm{Sbp}, \mathrm{Dbp}, \mathrm{W}, \mathrm{D}, \mathrm{Q}) \wedge \mathbf{X}(\exists \mathrm{Sbp}_1, \mathrm{Dbp}_1, \mathrm{W}_1, \mathrm{D}_1, \mathrm{Q}_1.\ \mathrm{Patient}(\mathrm{Ch}, \mathrm{Id}, \mathrm{Sbp}_1, \mathrm{Dbp}_1, \mathrm{W}_1, \mathrm{D}_1, \mathrm{Q}_1)) \wedge (\mathrm{W}_1 > \mathrm{W})\mathbf{U}\mathrm{Q}_{\mathrm{CEF}})$$

- The formula is satisfied when an increase trend for the values of the parameter $Weight$ during a whole granule of the granularity CEF exists

Other trend features

- Both for local and global satisfiability, it is possible to specify *existential* (**F**) or *universal* (**G**) temporal quantification with respect to the granules

  "Check whether an increase trend for SBP during a whole granule of CEF exists"

- The *duration* allows one to express the temporal length of the property

  "Check whether a patient, having two consecutive increases of SBP during a cycle of treatment, exists"

- The number of *exceptions* in a sequence of tuples allows one to express the maximum number of exceptions which might not satisfy the trend

  "Check whether a patient, having an increase trend for SBP during a whole cycle of treatment with at most two exceptions, exists"

- The *granule(s) frame* specifies that the required trend can be verified in a temporal window inside a granule or between granules

Conclusion and ongoing work

- This work aims to define a logical framework for expressing in a *uniform way* all the notions of temporal functional dependencies (with or without granularities) proposed in the database literature and moreover for the description of temporal trends over different granularities
- We have highlighted different orthogonal features related to the notion of temporal trends, *as local and global satisfiability*, *intragranule and intergranule trends*, *temporal quantification*, etc.
- Currently we aim to deal in a formal and detailed way with the problems of *model checking* and *temporal trend satisfaction*
- The final step will be the design and implementation of a system allowing one to query clinical databases according to specific related granularities (e.g., related to therapeutic cycles) and clinically relevant trends

References

Combi et al. 2004 C. Combi, M. Franceschet and A. Peron. *Representing and Reasoning about Temporal granularities*. *Journal of Logic and Computation*, 14(1):51-77, 2004.

Combi&Chittaro 1999 C. Combi and L. Chittaro. *Abstraction on clinical data sequences: an object-oriented data model and a query language based on the event calculus*. *Artificial Intelligence in Medicine*, 17(3): 271-301, 1999.

Combi&Montanari 2001 C. Combi and A. Montanari. *Data models with multiple temporal dimensions: Completing the picture*. *In Proc. of the 13th Conference on Advanced Information Systems Engineering*, vol. 2068 of LNCS, pp.187-202, 2001.

Chomicky 1995 J. Chomicki. *Efficient checking of temporal integrity constraints using bounded history encoding*. *ACM Transaction on Database Systems*, 20(2):149-186, 1995.

Chomicky&Niwinski 1995 J. Chomicki and D. Niwinski. *On the feasibility of checking temporal integrity constraints*. *Journal of Computer and System Sciences*, 51(3):523-535, 1995.

References

Jensen et al. 1996 C. Jensen, R. Snodgrass, and M. Soo. *Extending existing dependency theory to temporal databases*. *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 4, pp. 563–581, 1996.

Lipeck&Saake 1987 U.W. Lipeck and G. Saake. *Monitoring dynamic integrity constraints based on temporal logic*. *Information Systems*, 12(3):255-269, 1987.

Shahar 1997 Y. Shahar. *A Framework for Knowledge-Based Temporal Abstraction*. *Artificial Intelligence*, 90:79-133, 1997.

Ullman 1988 J.D. Ullman. *Principles of Database and Knowledge-Base Systems - Volume I.* Computer Science Press, 1988.

Vianu 1987 V. Vianu. *Dynamic functional dependency and database aging*. *Journal of the ACM*, vol. 34, no. 1, pp. 28–59, 1987.

Wang et al. 1997 X. S. Wang, C. Bettini, A. Brodsky, and S. Jajodia. *Logical design for temporal databases with multiple granularities*. *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 2, pp. 115–170, 1997.

References

Wang et al. 2000 X. Wang, C. Bettini and S. Jajodia. *Time granularities in Databases, Data Mining, and TemporalReasoning*. Springer, 2000.

Wang et al. 2002 X. S. Wang, P. Ning and S. Jajodia. *An algebraic representation of calendars*. *Annals of Mathematics and Artificial Intelligence*, vol. 36, pp. 5–38, 2002.

Wijsen 1995 J. Wijsen. *Design of temporal relational databases based on dynamic and temporal functional dependencies*. In *International Workshop on Temporal Databases*, ser. Recent Advances in Temporal Databases, J. Clifford and A. Tuzhilin, Eds. Springer, 1995, pp. 61–76.

Wijsen 1998 J. Wijsen. *Reasoning about qualitative trends in database*. *Information Systems*, 23(7):469-493, 1998.

Wijsen 1999 J. Wijsen. *Temporal FDs on complex objects*. *ACM Transactions on Database Systems*, vol. 24, no. 1, pp. 127–176, 1999.

Wijsen 2001 J. Wijsen. *Trends in databases: Reasoning and mining*. *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 3, pp. 426–438, 2001.

The proposed logic

- The *set of objects* which can be *associated to the time point* i is:

$$L(i) = \begin{cases} P_{\mathcal{G}}, Q_{\mathcal{G}} & \text{where } \mathcal{G} \text{ is a defined granularity} \\ r(v_1, \ldots, v_n) & \text{where } r \text{ is the name of a relation,} \\ & v_i \text{ is the value associated to the i-th attribute} \end{cases}$$

- The *labeled linear time structure* $\mathcal{M}$ has the form $(\mathbb{N}, <, \mathsf{L})$ where $\mathsf{L} : \mathbb{N} \rightarrow 2^{\mathcal{L}}$ is the *labeling function* and $\mathcal{L}$ is the *objects universe*
- **Quantifiers over the variables**: $\forall x$ and $\exists x$ for each variables $x \in \mathcal{V}$
- **Temporal operators**: **X** (*next*), **U** (*until*), $\mathbf{X}^{-1}$ (*prec*), **S** (*since*), **F** (*sometime*), and **G** (*always*)
- **Valuation**: if x is a variable associated to the attribute a_j, then the valuation $\varphi(x)$ returns a value in the D_j domain

Syntax and Semantics

Definition. *The set of formulae is inductively defined as follows:*

$$F := p \mid r(x, y, \ldots, w, z) \mid x\theta y \mid x\theta\ y\ op\ z \mid x\theta\ y\ op\ c \mid F \wedge F \mid \neg F \mid \mathbf{X}F \mid F\mathbf{U}F \mid \mathbf{X}^{-1}F \mid F\mathbf{S}F \mid \forall x.F$$

where $x, y, \ldots, w, z \in \mathcal{V}$, $p \in \mathcal{P}_{\mathcal{G}}$, *and* $\theta \in \Theta$ *is a comparison operator, and* $op \in \{+, -\}$.

Definition. *Let* $\mathcal{M} = (\mathbb{N}, <, L)$ *be a* $\mathcal{L}$*-labeled linear time structure,* $i \in \mathbb{N}$*, and* $\varphi : \mathcal{V} \to \mathcal{D}$ *a valuation. We define a relation* $\mathcal{M}, i, \varphi \models \phi$ *by induction on the structure of* ϕ*:*

$\mathcal{M}, i, \varphi \models p$	*iff*	$p \in L(i)$ *for* $p \in \mathcal{P}_{\mathcal{G}}$*;*
$\mathcal{M}, i, \varphi \models r(x, \ldots, z)$	*iff*	$r(\varphi(x), \ldots, \varphi(z)) \in L(i)$*;*
$\mathcal{M}, i, \varphi \models x\ \theta\ y$	*iff*	$\varphi(x)\theta\varphi(y)$
$\mathcal{M}, i, \varphi \models x\ \theta\ y\ op\ z$	*iff*	$\varphi(x)\theta\ \varphi(y)\ op\ \varphi(z)$
$\ldots$	$\ldots$	$\ldots$
$\mathcal{M}, i, \varphi \models \mathbf{X}\psi$	*iff*	$\mathcal{M}, i+1, \varphi \models \psi$*;*
$\mathcal{M}, i, \varphi \models \phi\mathbf{U}\psi$	*iff*	$\mathcal{M}, j, \varphi \models \psi$ *for some* $j \geq i$ *and*
		$\mathcal{M}, k, \varphi \models \phi$ *for each* $i \leq k < j$*;*
$\mathcal{M}, i, \varphi \models \mathbf{X}^{-1}\psi$	*iff*	$i > 0$ *and* $\mathcal{M}, i-1, \varphi \models \psi$*;*
$\mathcal{M}, i, \varphi \models \phi\mathbf{S}\psi$	*iff*	$\mathcal{M}, j, \varphi \models \psi$ *for some* $j \leq i$ *and*
		$\mathcal{M}, k, \varphi \models \phi$ *for each* $j < k \leq i$*;*
$\mathcal{M}, i, \varphi \models \forall x.\psi$	*iff*	*for each* $a \in \mathcal{D}$*,* $\mathcal{M}, i, \varphi[x \mapsto a] \models \psi$.