Non-Standard-Datenbanken und Data Mining

Prof. Dr. Ralf Möller

Universität zu Lübeck

Institut für Informationssysteme

Übersicht

- Semistrukturierte Datenbanken (JSON, XML) und Volltextsuche
- Information Retrieval
- Mehrdimensionale Indexstrukturen
- Cluster-Bildung
- Einbettungstechniken
- First-n-, Top-k-, und Skyline-Anfragen
- Probabilistische Datenbanken, Anfragebeantwortung, Top-k-Anfragen und Open-World-Annahme
- Probabilistische Modellierung, Bayes-Netze, Anfragebeantwortungsalgorithmen, Lernverfahren,
- Temporale Datenbanken und das relationale Modell, SQL:2011
- Probabilistische Temporale Datenbanken
- SQL: neue Entwicklungen (z.B. JSON-Strukturen und Arrays), Zeitreihen (z.B. TimeScaleDB)
- Stromdatenbanken, Prinzipien der Fenster-orientierten inkrementellen Verarbeitung
- Approximationstechniken für Stromdatenverarbeitung, Stream-Mining
- Probabilistische raum-zeitliche Datenbanken und Stromdatenverarbeitungsssysteme: Anfragen und Indexstrukturen, Raum-zeitliches Data Mining, Probabilistische Skylines
- Von NoSQL- zu NewSQL-Datenbanken, CAP-Theorem

Acknowledgements:

This presentation is based on the following two presentations

10 Years of Probabilistic Querying – What Next?
Martin Theobald
University of Antwerp

Temporal Alignment
Anton Dignös[1] Michael H. Böhlen[1] Johann Gamper[2]
[1]University of Zürich, Switzerland
[2]Free University of Bozen-Bolzano, Italy

Recap: Probabilistic Databases

A probabilistic database $\mathbf{D}^p$ (compactly) encodes a probability distribution over a finite set of deterministic database instances $\mathbf{D}_i$.

D_1: 0.42

WorksAt(Sub, Obj)	
Jeff	Stanford
Jeff	Princeton

D_2: 0.18

WorksAt(Sub, Obj)	
Jeff	Stanford

D_3: 0.28

WorksAt(Sub, Obj)	
Jeff	Princeton

D_4: 0.12

WorksAt(Sub, Obj)

- Special Cases:

(I) $\mathbf{D}^p$ tuple-independent

WorksAt(Sub, Obj)		p
Jeff	Stanford	0.6
Jeff	Princeton	0.7

(II) $\mathbf{D}^p$ block-independent

WorksAt(Sub, Obj)		p
Jeff	Stanford	0.6
	Princeton	0.4

Note: (I) and (II) are not equivalent!

- Query Semantics: ("Marginal Probabilities")
  - Run query Q against each instance $\mathbf{D}_i$; for each answer tuple t, sum up the probabilities of all instances $\mathbf{D}_i$ where t is a result.

Probabilistic & Temporal Databases

A temporal-probabilistic database $\mathbf{D}^{TP}$ (compactly) encodes a probability distribution over a finite set of deterministic database instances $\mathbf{D}_i$ *at each time point* of a finite time domain Ω^T.

BornIn(Sub,Obj)		T	p
DeNiro	Green-which	[1943, 1944)	0.9
DeNiro	Tribeca	[1998, 1999)	0.6

Wedding(Sub,Obj)		T	p
DeNiro	Abbott	[1936, 1940)	0.3
DeNiro	Abbott	[1976, 1977)	0.7

Divorce(Sub,Obj)		T	p
DeNiro	Abbott	[1988, 1989)	0.8

- Sequenced Semantics & Snapshot Reducibility: [Dignös, Gamper, Böhlen: SIGMOD'12]
  - Built-in semantics: **reduce temporal-relational operators** to their non-temporal counterparts at *each snapshot* (i.e., time point) of the database.
  - **Coalesce/split tuples** with consecutive time intervals based on their lineages.
- Non-Sequenced Semantics

$$\text{marriedTo}(x,y)[t_{b1},T_{max}) \Leftarrow \text{wedding}(x,y)[t_{b1},t_{e1}) \wedge \neg\text{divorce}(x,y)[t_{b2},t_{e2})$$

  - Queries can **freely manipulate timestamps** just like regular attributes. Single temporal operator $\leq^T$ supports all of Allen's 13 temporal relations.
  - **Deduplicate tuples** with overlapping time intervals based on their lineages.

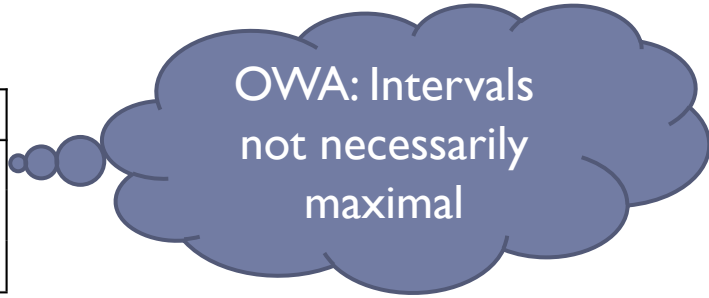
[Dylla, Miliaraki, Theobald: PVLDB'13]

Sequenced Semantics: Example

- **Input**: Employee N works for department D during time T.

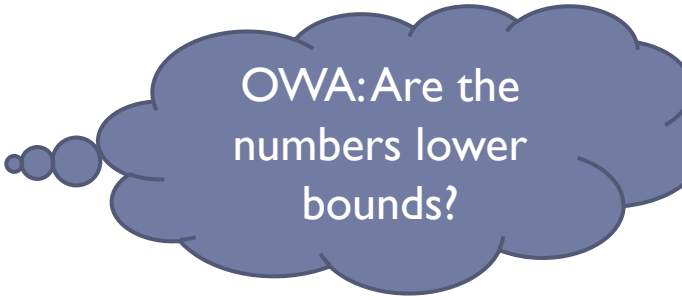
R

	N	D	T
r_1	Joe	DB	[Feb, Jul)
r_2	Ann	DB	[Feb, Sep)
r_3	Sam	AI	[May, Oct)



- **Query**: How did the average duration of contracts per department change?
- **Result**: Temporal Aggregation: ${}_{D}\vartheta^{T}_{AVG(DUR(T))}(\mathbf{R})$

	AVG	D	T
z_1	6	DB	[Feb, Jul)
z_2	7	DB	[Jul, Sep)
z_3	5	AI	[May, Oct)

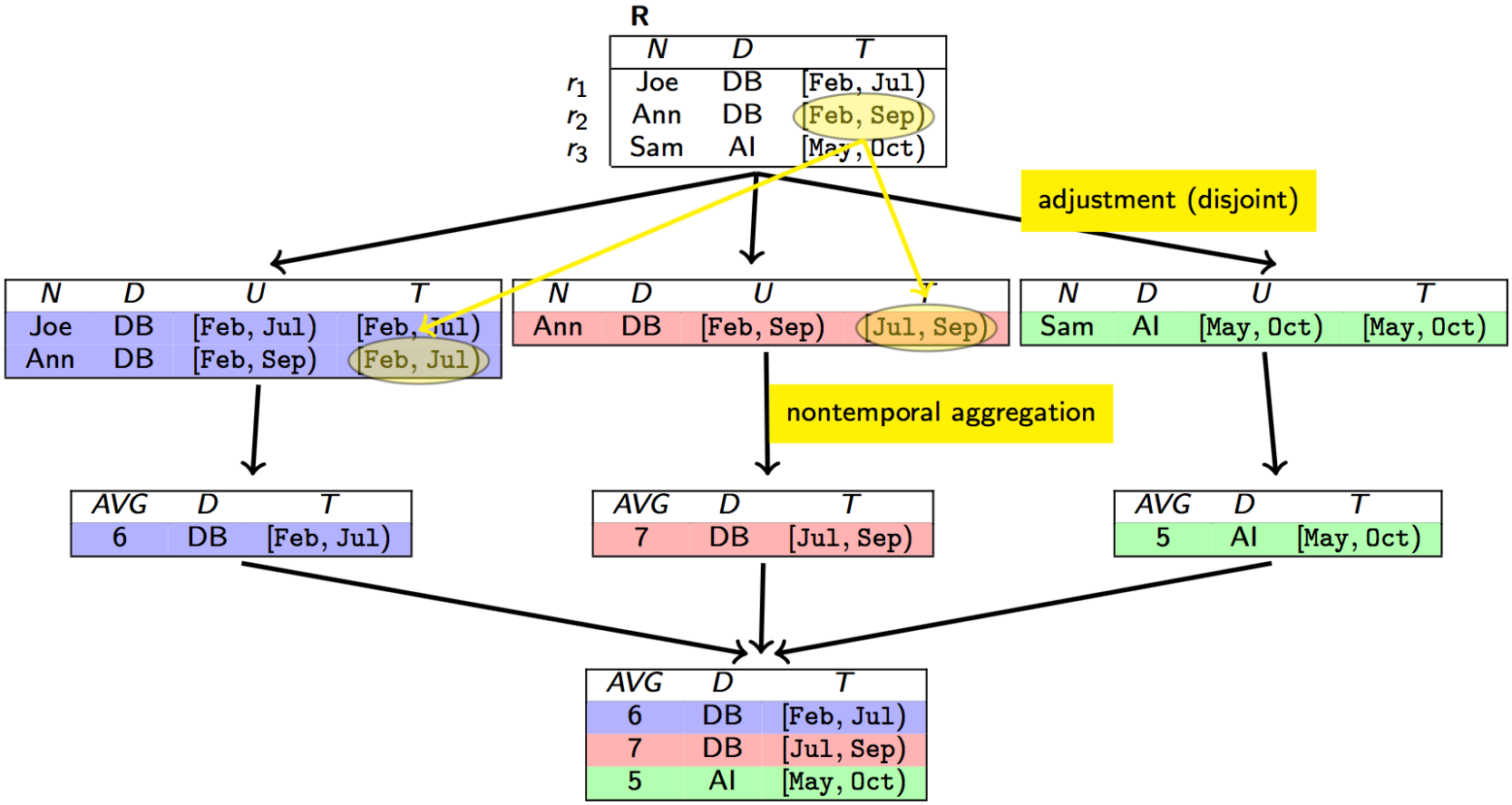


Timestamps must be adjusted for the result.

[Dignös, Gamper, Böhlen: SIGMOD'12]

Temporal Splitter / Snapshot Reduction

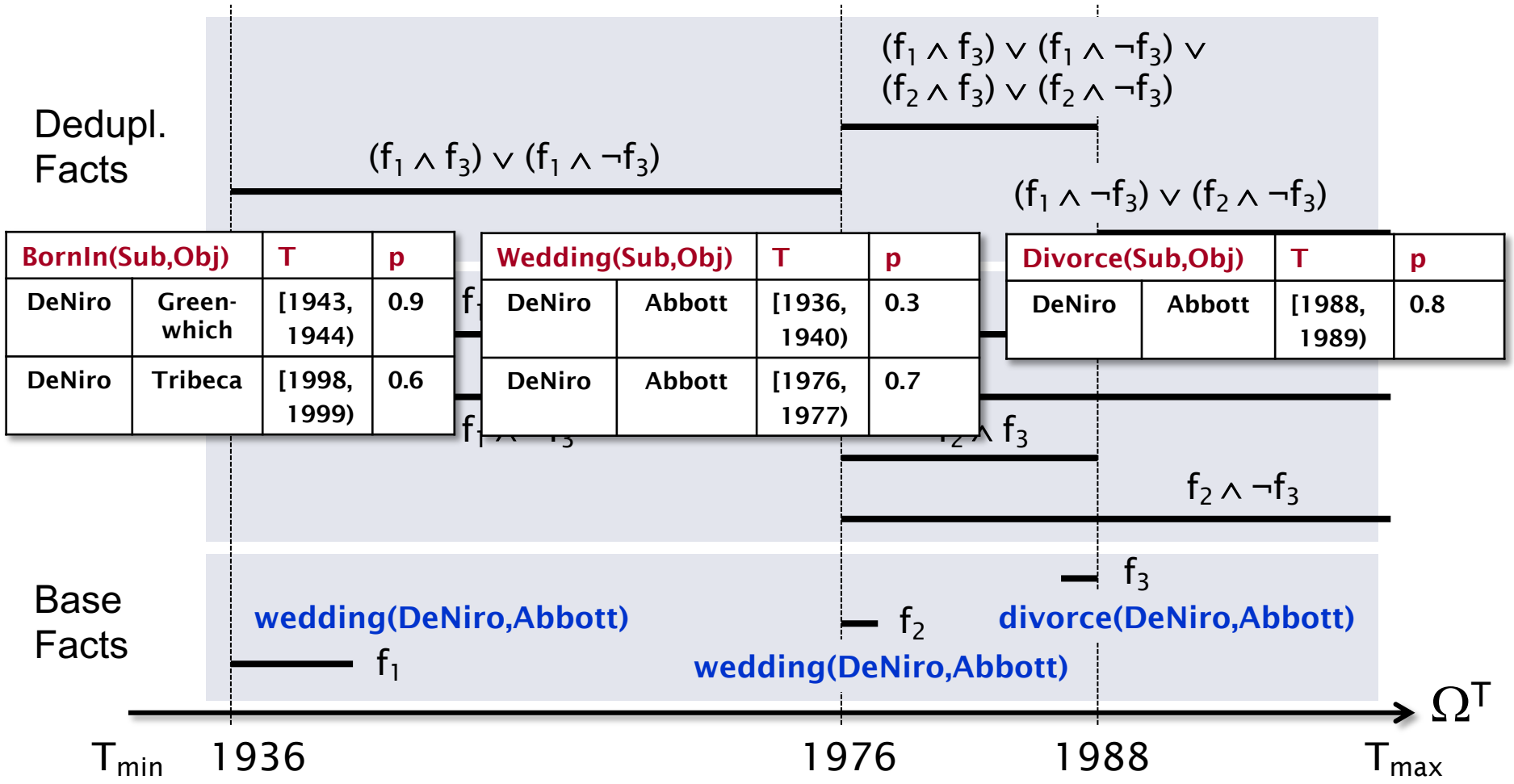
- Average duration of contracts per department: $_{D}\vartheta^{T}_{AVG(DUR(T))}(\mathbf{R})$



- One input tuple contributes to at most one result tuple per month.

[Dignös, Gamper, Böhlen: SIGMOD'12]

Temporal Alignment & Deduplication Example



BornIn(Sub,Obj)		T	p
DeNiro	Green-which	[1943, 1944)	0.9
DeNiro	Tribeca	[1998, 1999)	0.6

Wedding(Sub,Obj)		T	p
DeNiro	Abbott	[1936, 1940)	0.3
DeNiro	Abbott	[1976, 1977)	0.7

Divorce(Sub,Obj)		T	p
DeNiro	Abbott	[1988, 1989)	0.8

Non-Sequenced Semantics:

$\text{marriedTo}(x,y)[t_{b1},T_{max}) \Leftarrow \text{wedding}(x,y)[t_{b1},t_{e1}) \wedge \neg\text{divorce}(x,y)[t_{b2},t_{e2})$

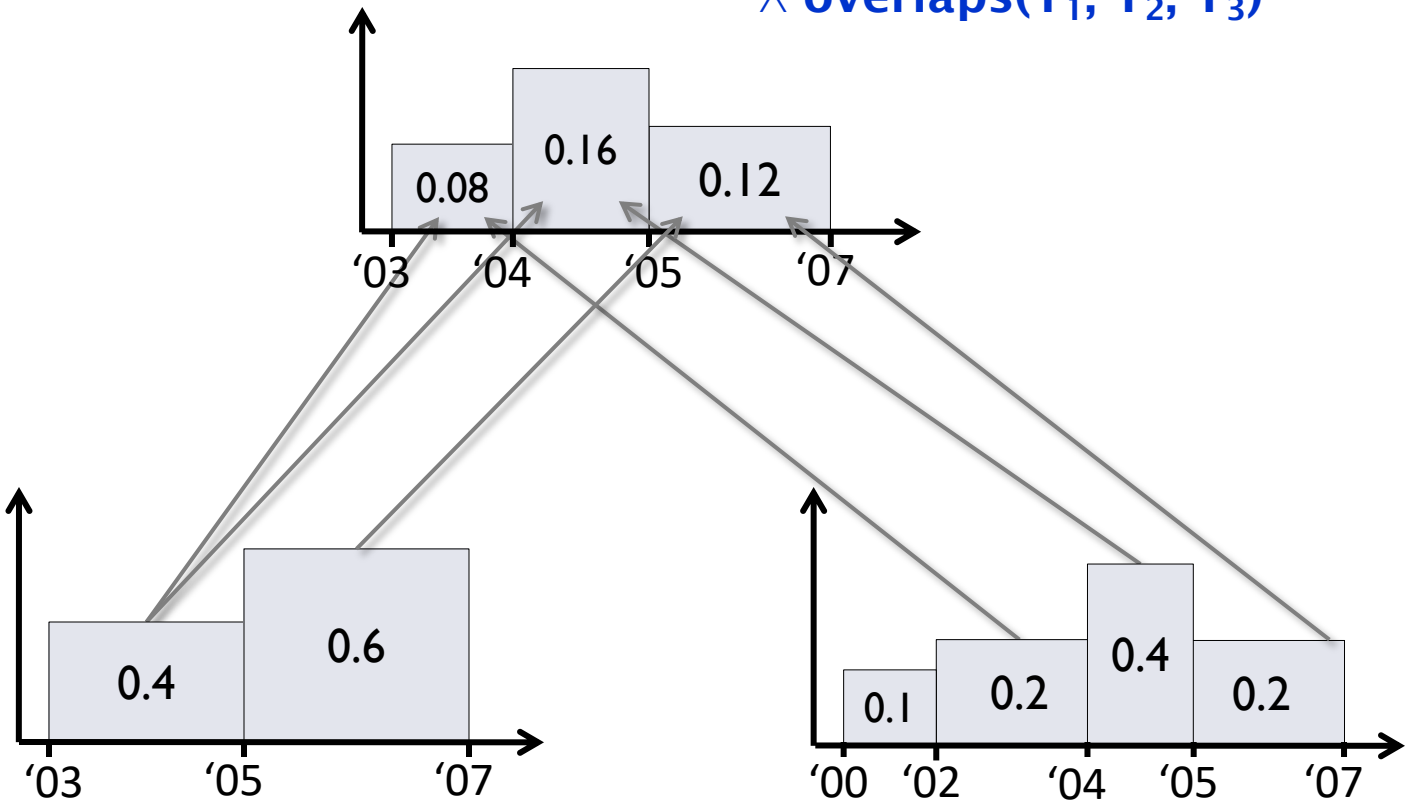
$\text{marriedTo}(x,y)[t_{b1},t_{e2}) \Leftarrow \text{wedding}(x,y)[t_{b1},t_{e1}) \wedge \text{divorce}(x,y)[t_{b2},t_{e2}) \wedge t_{e1} \leq^T t_{b2}$

Inference in Probabilistic-Temporal Databases

[Wang, Yahya, Theobald: MUD'10; Dylla, Miliaraki, Theobald: PVLDB'13]

Derived Facts

$$\text{teamMates}(\text{Beckham}, \text{Ronaldo}, T_3) \Leftarrow \text{playsFor}(\text{Beckham}, \text{Real}, T_1) \wedge \text{playsFor}(\text{Ronaldo}, \text{Real}, T_2) \wedge \text{overlaps}(T_1, T_2, T_3)$$



Base Facts

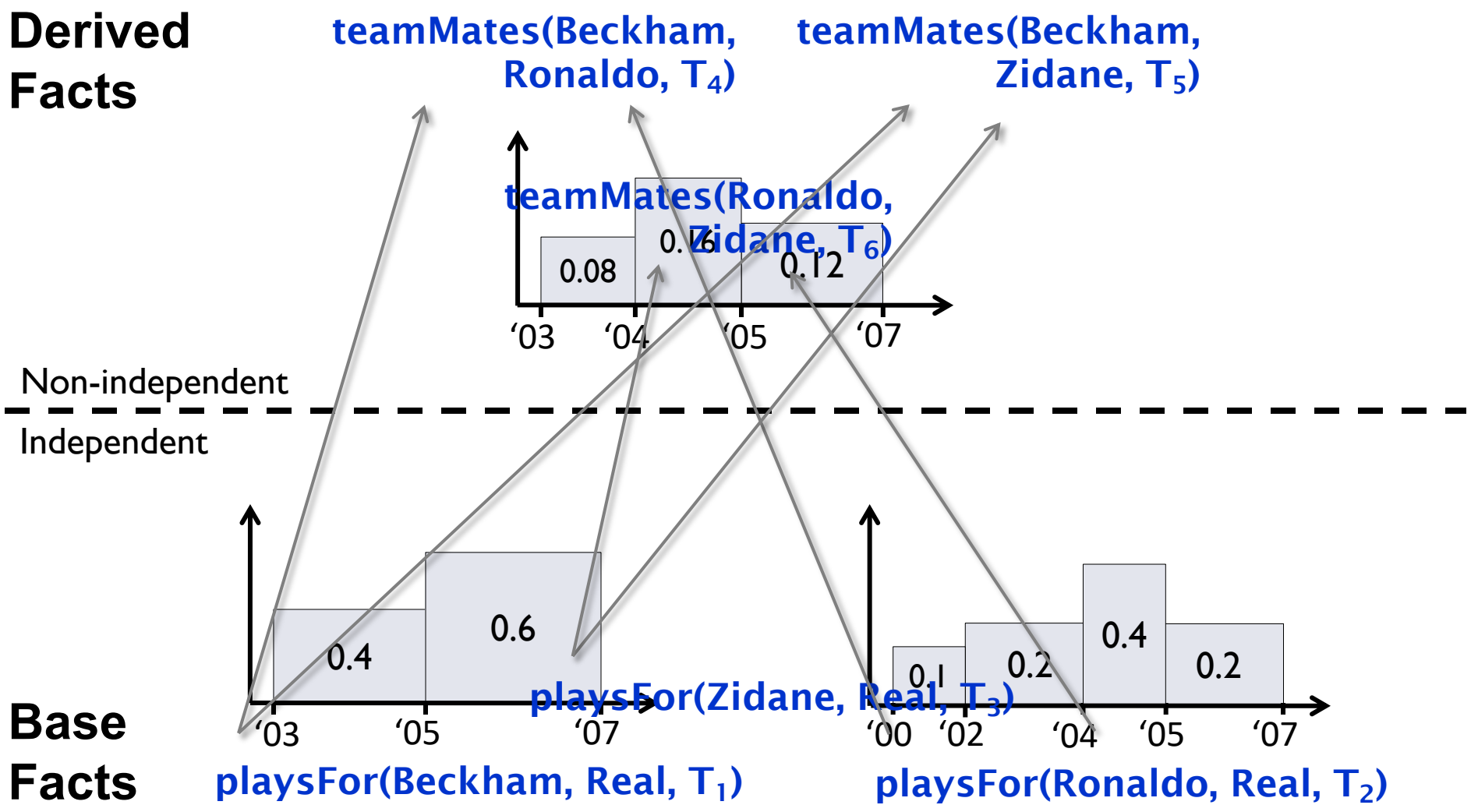
$\text{playsFor}(\text{Beckham}, \text{Real}, T_1)$

$\text{playsFor}(\text{Ronaldo}, \text{Real}, T_2)$

Example using the Allen predicate *overlaps*

Inference in Probabilistic-Temporal Databases

[Wang, Yahya, Theobald: MUD'10; Dylla, Miliaraki, Theobald: PVLDB'13]



Inference in Probabilistic-Temporal Databases

[Wang, Yahya, Theobald: MUD'10; Dylla, Miliaraki, Theobald: PVLDB'13]

Derived facts stored in views

teamMates(Beckham, Ronaldo, T_4)

teamMates(Beckham, Zidane, T_5)

teamMates(Ronaldo, Zidane, T_6)

Non-independent

Independent

- **Closed** and **complete** representation model (incl. lineage)
- **Temporal alignment** is **polyn.** in the number of input intervals
- **Confidence computation** per interval remains **#P-hard**
- In general requires Monte Carlo approximations (Luby-Karp for DNF, MCMC-style sampling), decompositions, or top-*k* pruning

Literature

[Das Sarma,Theobald,Widom: ICDE'08]
Das Sarma, Anish and Theobald, Martin and Widom, Jennifer, Exploiting Lineage for Confidence Computation in Uncertain and Probabilistic Databases. In: 24th International Conference on Data Engineering (ICDE 2008), IEEE Computer Society Press, pp. 1023-1032. **2008**

[Wang,Yahya,Theobald: MUD'10]
Wang, Yafang and Yahya, Mohamed and Theobald, Martin, Time-aware Reasoning in Uncertain Knowledge Bases. In: 4th International VLDB Workshop on Management of Uncertain Data, Vol. WP 10-, pp. 51-65, **2010**

[Dignös, Gamper, Böhlen: SIGMOD'12]
A. Dignös, M.H. Böhlen, J. Gamper. Temporal alignment. In Proc. of the SIGMOD-12, pages 433-444, Scottsdale, AZ, USA, May 20-24, **2012**

[Dylla,Miliaraki,Theobald: PVLDB'13]
Maximilian Dylla, Iris Miliaraki, Martin Theobald, A Temporal-Probabilistic Database Model for Information Extraction, Proceedings of the VLDB Endowment, Volume 6, Issue 14, **2013**