Non-Standard-Datenbanken

Prof. Dr. Ralf Möller

Universität zu Lübeck

Institut für Informationssysteme

Felix Kuhr (Übungen)

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

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

  - 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

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

R

	N	D	T
r_1	Joe	DB	$[\texttt{Feb}, \texttt{Jul})$
r_2	Ann	DB	$[\texttt{Feb}, \texttt{Sep})$
r_3	Sam	AI	$[\texttt{May}, \texttt{Oct})$

- **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	$[\texttt{Feb}, \texttt{Jul})$
z_2	7	DB	$[\texttt{Jul}, \texttt{Sep})$
z_3	5	AI	$[\texttt{May}, \texttt{Oct})$

Timestamps must be adjusted for the result.

[Dignös, Gamper, Böhlen: SIGMOD'12]

Temporal Splitter / Snapshot Reduction

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

R

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

adjustment (disjoint)

N	D	U	T
Joe	DB	[Feb, Jul)	[Feb, Jul)
Ann	DB	[Feb, Sep)	[Feb, Jul)

N	D	U	T
Ann	DB	[Feb, Sep)	[Jul, Sep)

N	D	U	T
Sam	AI	[May, Oct)	[May, Oct)

nontemporal aggregation

AVG	D	T
6	DB	[Feb, Jul)

AVG	D	T
7	DB	[Jul, Sep)

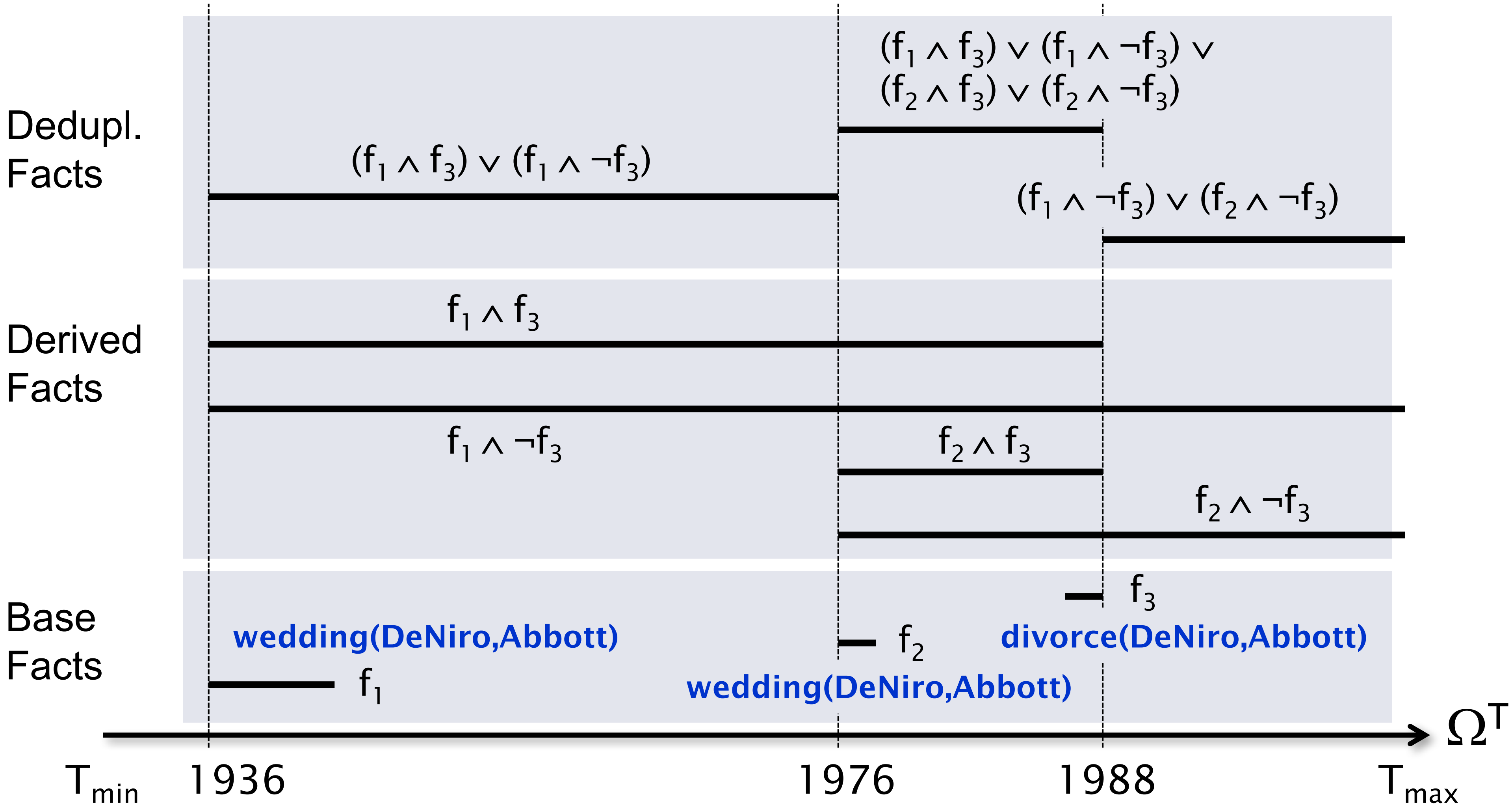
AVG	D	T
5	AI	[May, Oct)

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

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

[Dignös, Gamper, Böhlen: SIGMOD'12]

Temporal Alignment & Deduplication Example



<u>Non-Sequenced Semantics:</u>

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

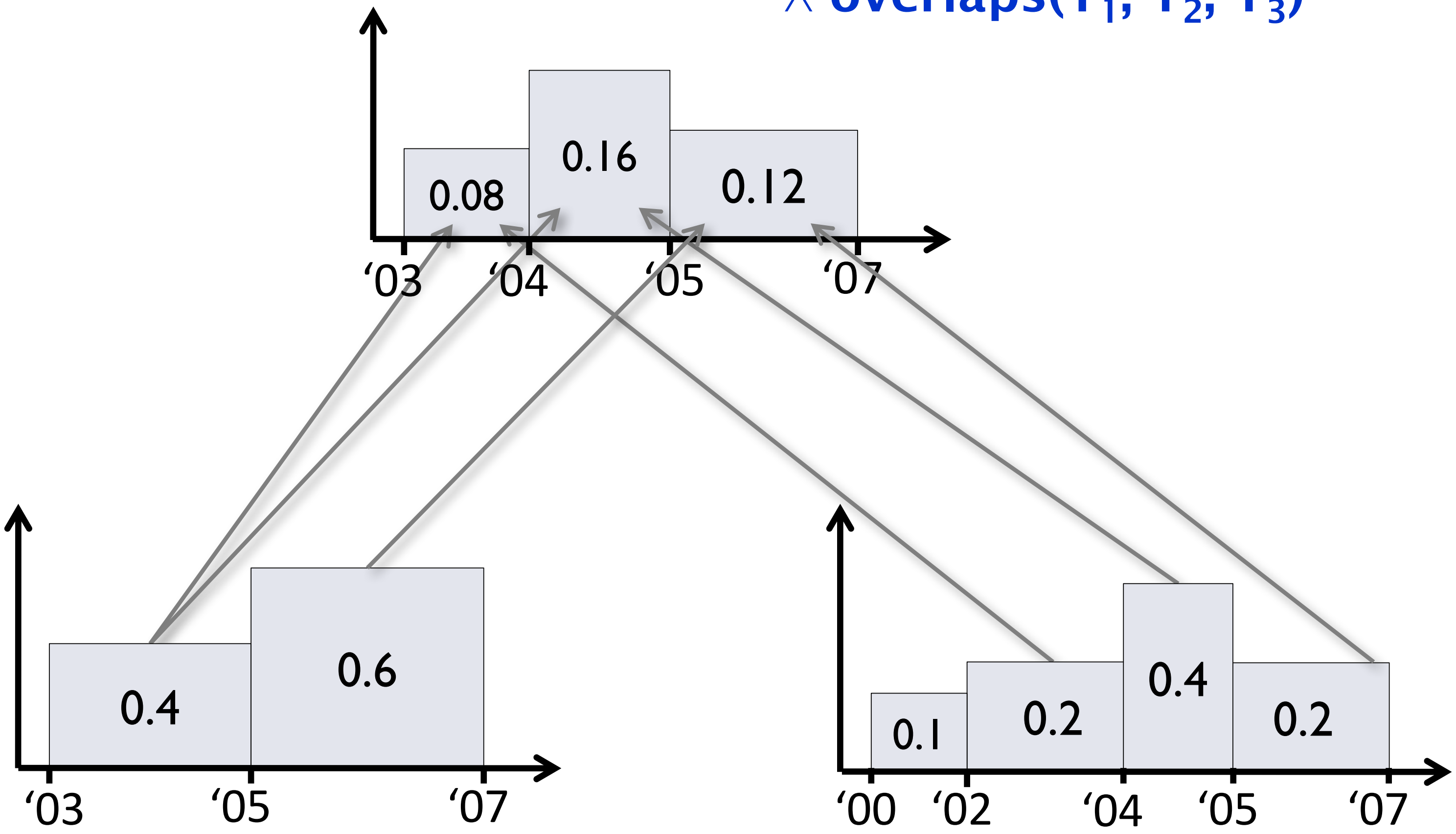
marriedTo(x,y)[t_{b1},t_{e2}) $\Leftarrow$ wedding(x,y)[t_{b1},t_{e1}) $\wedge$ divorce(x,y)[t_{b2},t_{e2}) $\wedge$ $t_{e1} \leq^T t_{b2}$

Inference in Probabilistic-Temporal Databases

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

Derived Facts

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



Base Facts

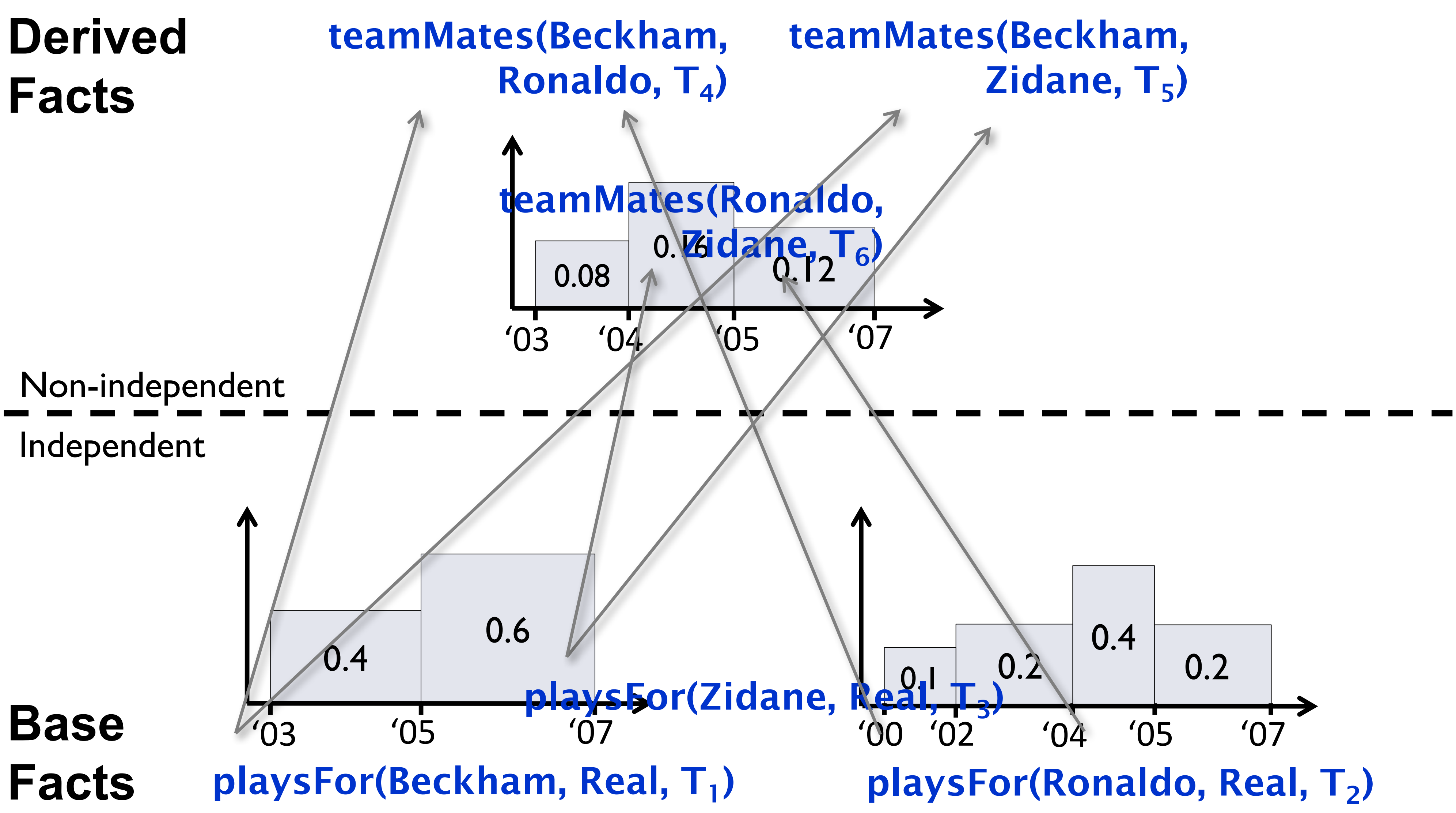
playsFor(Beckham, Real, T_1)

playsFor(Ronaldo, 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)

Need Lineage!

Non-independent

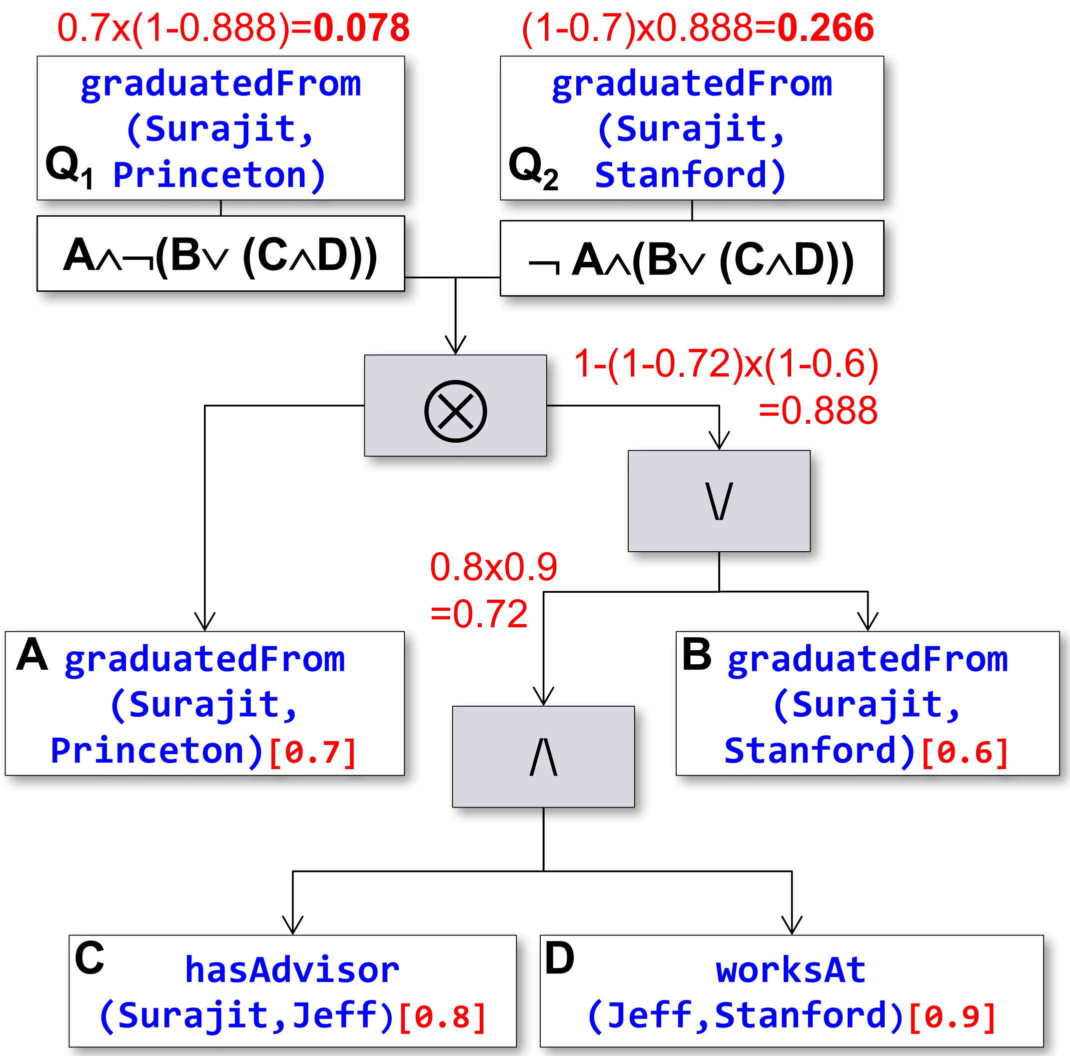
Independent

- **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

Lineage & Possible Worlds

[Das Sarma, Theobald, Widom: ICDE'08
Dylla, Miliaraki, Theobald: ICDE'13]

Query
graduatedFrom(Surajit, y)



1) Deductive Grounding
- Dependency graph of the query
- Trace lineage of individual query answers

2) Lineage DAG (not in CNF), consisting of
- Grounded soft & hard views
- *Probabilistic* base facts

3) Probabilistic Inference
→ Compute marginals:

P(Q): sum up the probabilities of all possible worlds that entail the query answers' lineage

P(Q|H): drop "impossible worlds"