



# Managing Large, Uncertain Data Repositories with Probabilistic Graphical Models

Daisy Zhe Wang<sup>+</sup>, Eirinaios Michelakis<sup>+</sup>,  
Minos Garofalakis<sup>\*+</sup>, Joseph M. Hellerstein<sup>+</sup>

University of California Berkeley<sup>+</sup>, Yahoo! Research<sup>\*</sup>  
25<sup>th</sup> August 2008, VLDB

# Uncertainty in Real Systems

Sensor Networks



Data Extraction Systems

**DBLife**

**Yahoo!/PSOX**

**IBM/Avatar/SystemT**



Social Networks



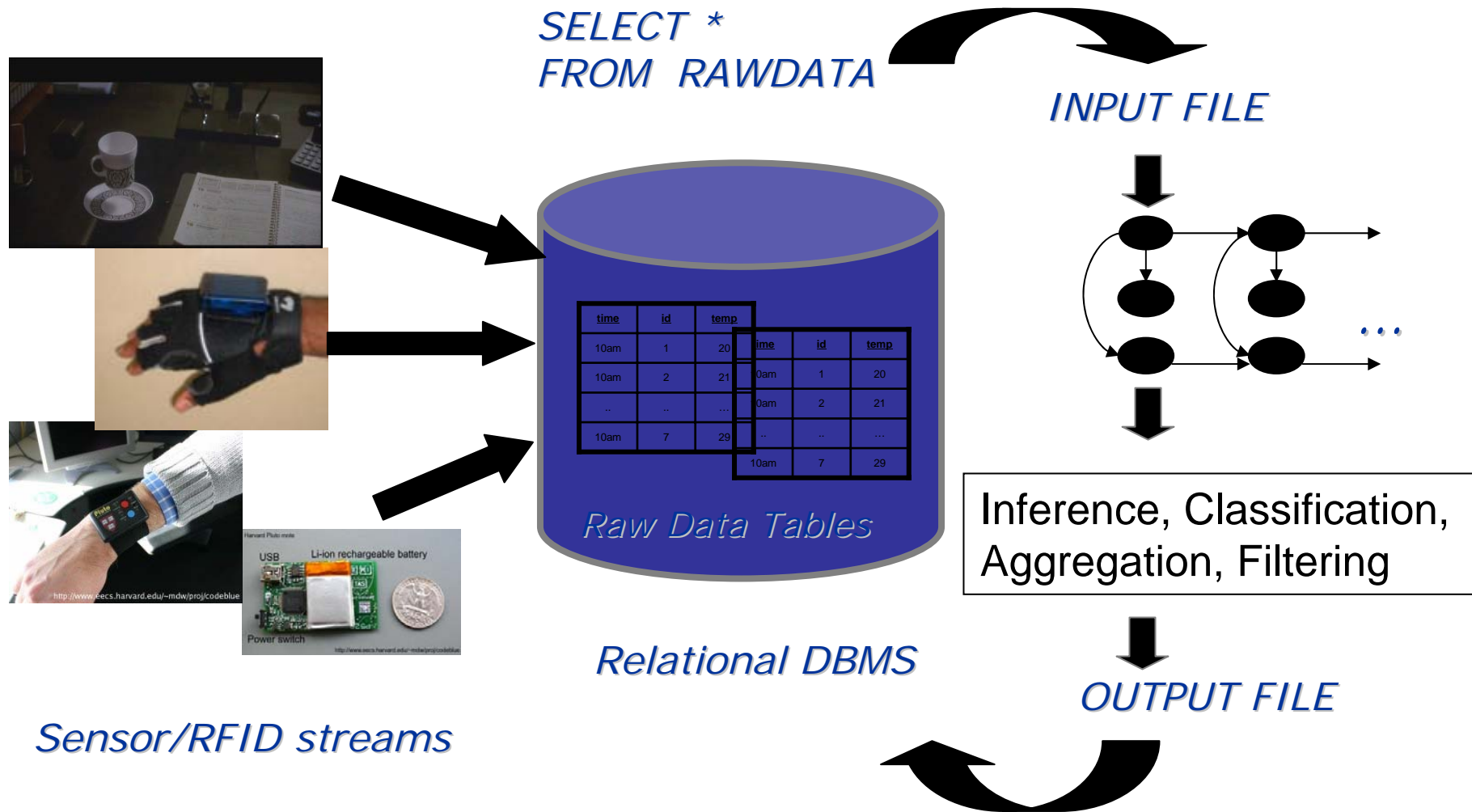
Data Integration Systems



# State of the Art – Probabilistic Data Management

- Machine Learning Research
  - Decision Tree, CRF Model
  - Bayesian Network
  - Probabilistic Relational Model

# Machine Learning Approach



# State of the Art – Probabilistic Data Management

- Machine Learning Research
  - Bayesian Network, Markov Network
  - Probabilistic Relational Model
  - Markov Network Model
- Probabilistic/Uncertain Database Research
  - MystiQ System [Dalvi&Suciu04]
  - Trio System [Wid05, Das06]
  - MauveDB [D&M, 2006]
  - MayBMS [ICDE07]

# BayesStore Data Model

1. Incomplete Relation --  $R^P$
2. Distribution over Possible Worlds –  $F$

**Sensor1(Time(T), Room(R), Sid, Temperature(Tp)<sup>P</sup>, Light(L)<sup>P</sup>)**

*Incomplete Relation of Sensor1<sup>P</sup>*

	<b>T</b>	<b>R</b>	<b>Sid</b>	<b>TP<sup>P</sup></b>	<b>L<sup>P</sup></b>
t1	1	1	11	Hot	X1
t2	1	1	22	Cold	Drk
t3	1	1	33	X2	X3
t4	1	2	11	X4	Brk
t5	1	2	22	Hot	X5
t6	1	2	33	X6	X7

*Probabilistic Distribution of Sensor1<sup>P</sup>*

$$F = \text{Pr} [X_1, \dots, X_7]$$

N: number of missing values

|X|: size of the domain

$$|F| = \Theta (|X|^N)$$

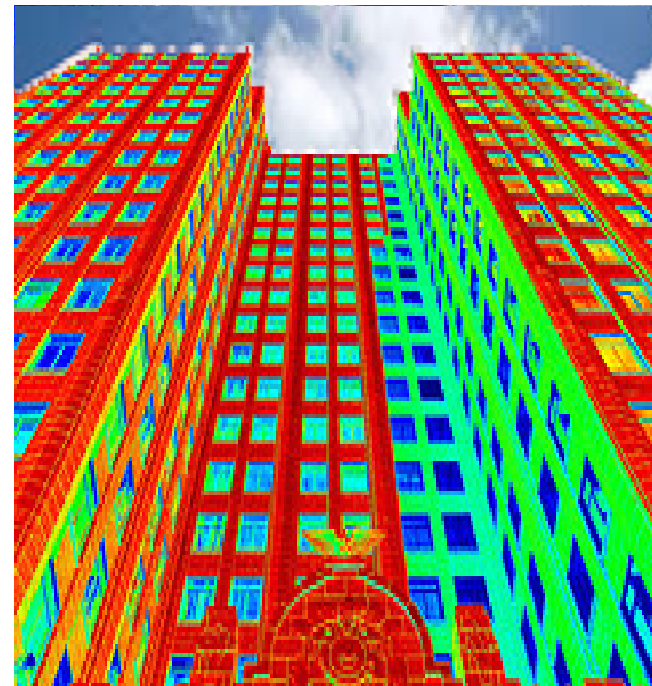
# The Skyscrapers Example

For all sensor in all rooms at all timestamp, Light and Temperature readings are correlated.

Light



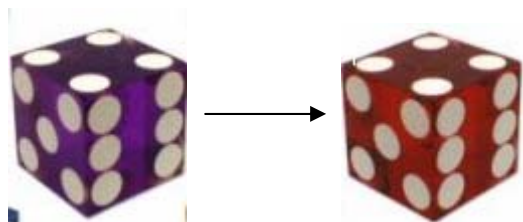
Temperature



# Definitions



**Stripe:** A family of random variables from the same probabilistic attribute.



**First-order Factor:** A family of local models, which share the same structure and conditional probability table(CPT).




**BayesStore Data Type:** The input and output abstract data type of queries in BayesStore, which consists of data and model.



**Possible Worlds**

# F as a First-order Bayesian Network (I)

**Sensor1<sup>p</sup>**

	T	R	Sid	Tp <sup>p</sup>	L <sup>p</sup>
t1	1	1	1	H 	X1
t2	1	1	2	Cold	Drk
t3	1	1	3	X2	X3
t4	1	2	1	X 	Brt
t5	1	2	2	Hot	X5
t6	1	2	3	X6	X7
t7	2	1	1	H 	X8
t8	2	1	2	Cold	Drk
t9	2	1	3	X9	X10
t10	2	2	1	X 	Brt
t11	2	2	2	Hot	X12
t12	2	2	3	X13	X14

**Stripe (FO Variable) Definitions**



*All Tp values in Sensor1<sup>p</sup> with Sid=1*

# F as a First-order Bayesian Network (I)

*Sensor1<sup>p</sup>*

	T	R	Sid	Tp <sup>p</sup>	L <sup>p</sup>
t1	1	1	1		X1
t2	1	1	2	C  d	Dr
t3	1	1	3	X	X3
t4	1	2	1	X	Br
t5	1	2	2	C	X5
t6	1	2	3	X	X7
t7	2	1	1	C	X8
t8	2	1	2	C  d	Dr
t9	2	1	3	X	X1
t10	2	2	1	X	Br
t11	2	2	2	C	X1
t12	2	2	3	X	X1

*Stripe (FO Variable) Definitions*



*All Tp values in Sensor1<sup>p</sup> with Sid=1*



*All Tp values in Sensor1<sup>p</sup> with Sid=2*



*All Tp values in Sensor1<sup>p</sup> with Sid !=2*



*All Tp values in Sensor1<sup>p</sup>*



*All L values in Sensor1<sup>p</sup>*

# F as a First-order Bayesian Model

## Mapping between Stripes

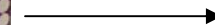
*All  $T_p$  values*



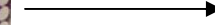
*All L values*



*All  $T_p$  values*



*All L values*



....

....



*All  $T_p$  values  
with Sid=1*



*All  $T_p$  values  
with Sid=2*



....

*All  $T_p$  values  
with Sid=1*

*All  $T_p$  values  
with Sid=2*

# F as a First-order Bayesian Model

## First-order Factor Definitions

*All Tp values*



*All L values*



*All Tp values  
with Sid=1*



*All Tp values  
with Sid=2*



*All Tp values  
with Sid !=2*

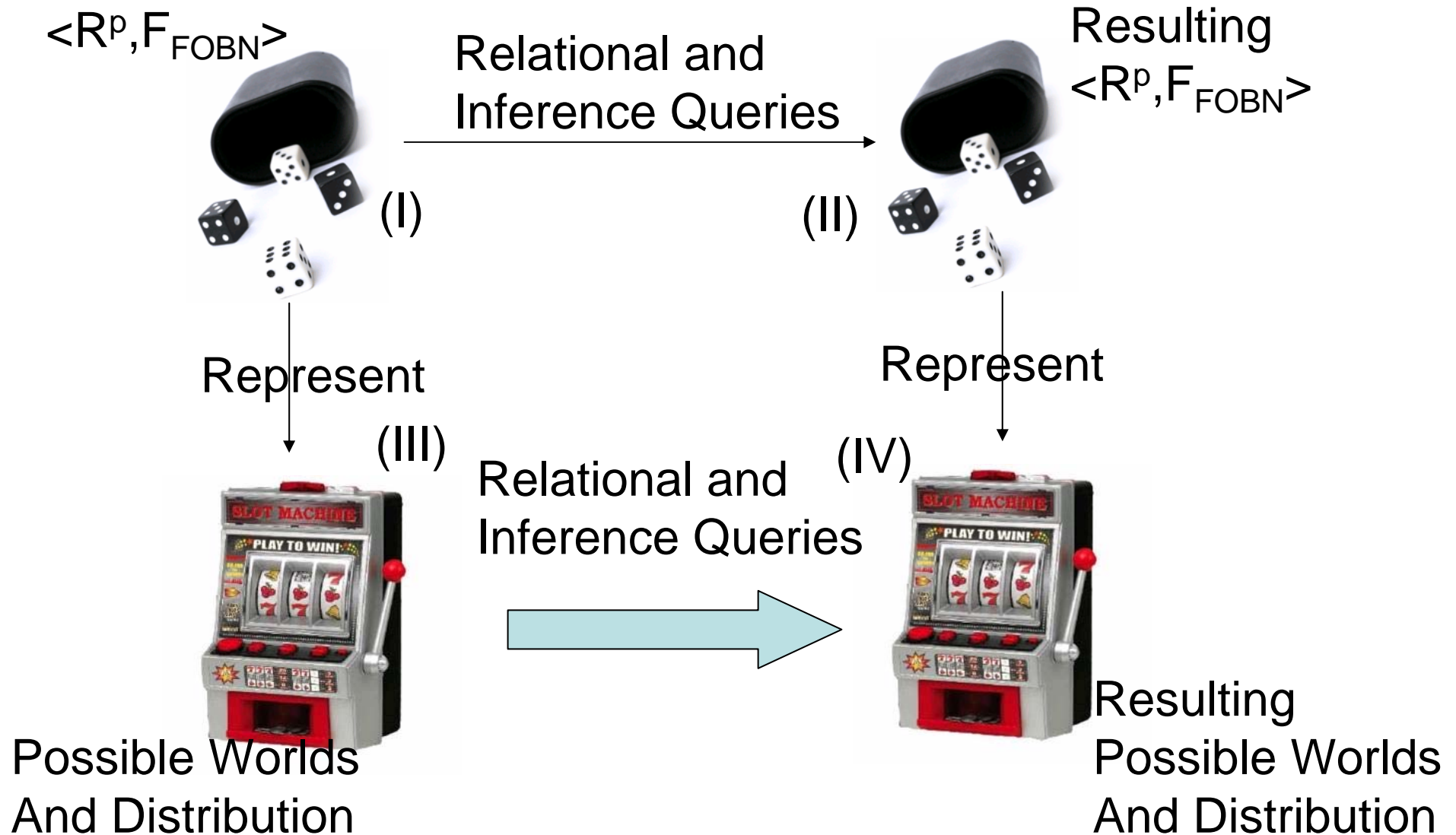


Tp	L	p
Cold	Brk	0.1
Hot	Brk	0.9
Hot	Drk	0.1
Cold	Drk	0.9

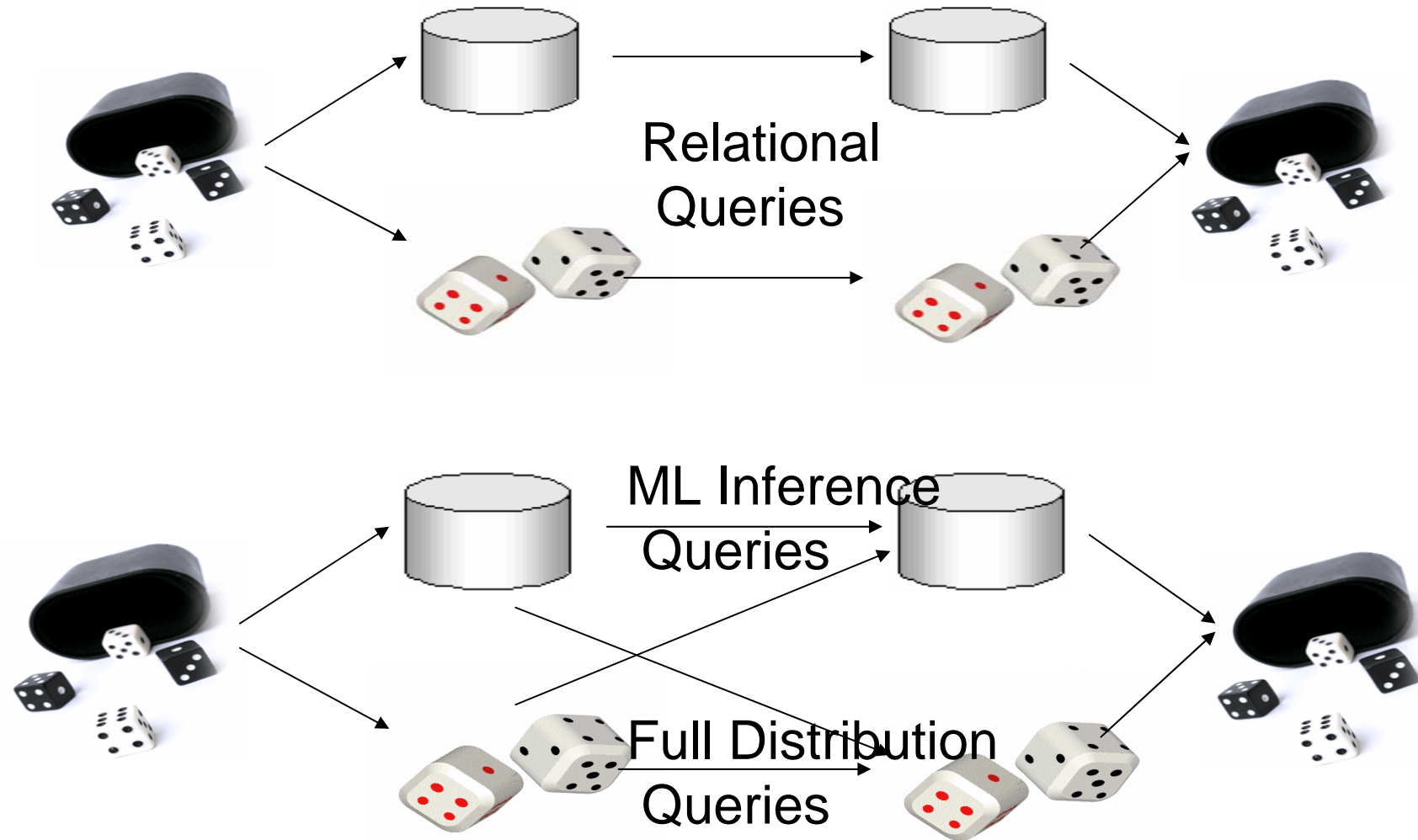
Tp1	Tp2	p
Cold	Cold	0.1
Cold	Hot	0.9
Hot	Hot	0.1
Hot	Cold	0.9

Tp	p
Cold	0.6
Hot	0.4

# Query Semantics



# Query Algebra

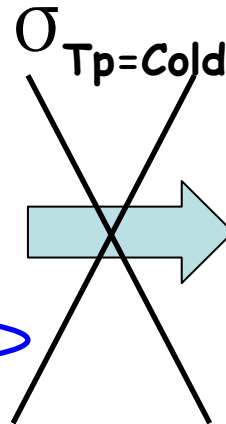


# Selection

- Selection over Incomplete Relation  $R^p$
- Selection over Model  $M_{\text{FOBN}}$

*Sensor1<sup>p</sup>*

	T	R	Sid	Tp <sup>p</sup>	L <sup>p</sup>
t1	1	1	1	Hot	X1
t2	1	1	2	Cold	Drk
t3	1	1	3	X2	X3
t4	1	2	1	X4	Brk
t5	1	2	2	Hot	X5
t6	1	2	3	X6	X7



*Sensor1<sup>p</sup>*

	T	R	Sid	Tp <sup>p</sup>	L <sup>p</sup>
t2	1	1	2	Cold	Drk
t3	1	1	3	X2	X3
t4	1	2	1	X4	Brk
t6	1	2	3	X6	X7

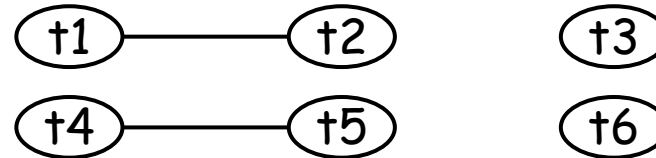
# Selection

- Selection over Incomplete Relation  $R^P$
- Selection over Model  $M_{FOBN}$

**Sensor1<sup>P</sup>**

	T	R	Sid	Tp <sup>P</sup>	LP
t1	1	1	1	Hot	X1
t2	1	1	2	Cold	Drk
t3	1	1	3	X2	X3
t4	1	2	1	X4	Brt
t5	1	2	2	Hot	X5
t6	1	2	3	X6	X7

**Tuple Correlation Graph (TCG)  
for  $F_{FOBN}$  (Sensor1)**



$\sigma_{Tp=Cold|Null}$

	T	R	Sid	Tp <sup>P</sup>	LP
t2	1	1	2	Cold	Drk
t3	1	1	3	X2	X3
t4	1	2	1	X4	Brt
t6	1	2	3	X6	X7

**Compute Transitive  
Closure over TCG**

	T	R	Sid	Tp <sup>P</sup>	LP
t1	1	1	1	Hot	X1
t2	1	1	2	Cold	Drk
t3	1	1	3	X2	X3
t4	1	2	1	X4	Brt
t5	1	2	2	Hot	X5
t6	1	2	3	X6	X7

# Selection

- Selection over Incomplete Relation  $R^p$
- **Selection over Model  $M_{FOBN}$**

**Probabilistic Distribution  $F_{FOBN}$  of Sensor1<sup>p</sup>**

**All  $T_p$  values**



**All L values**



$T_p$	L	p
Cold	Brt	0.1
Hot	Brt	0.9
Hot	Drk	0.1
Cold	Drk	0.9

**All  $T_p$  values  
with  $Sid=1$**



**All  $T_p$  values  
with  $Sid=2$**

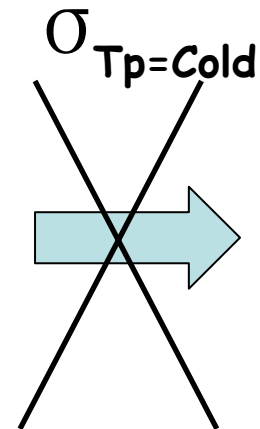


$T_{p1}$	$T_{p2}$	p
Cold	Cold	0.9
Cold	Hot	0.1
Hot	Hot	0.9
Hot	Cold	0.1

**All  $T_p$  values  
with  $Sid \neq 2$**



$T_p$	p
Cold	0.6
Hot	0.4

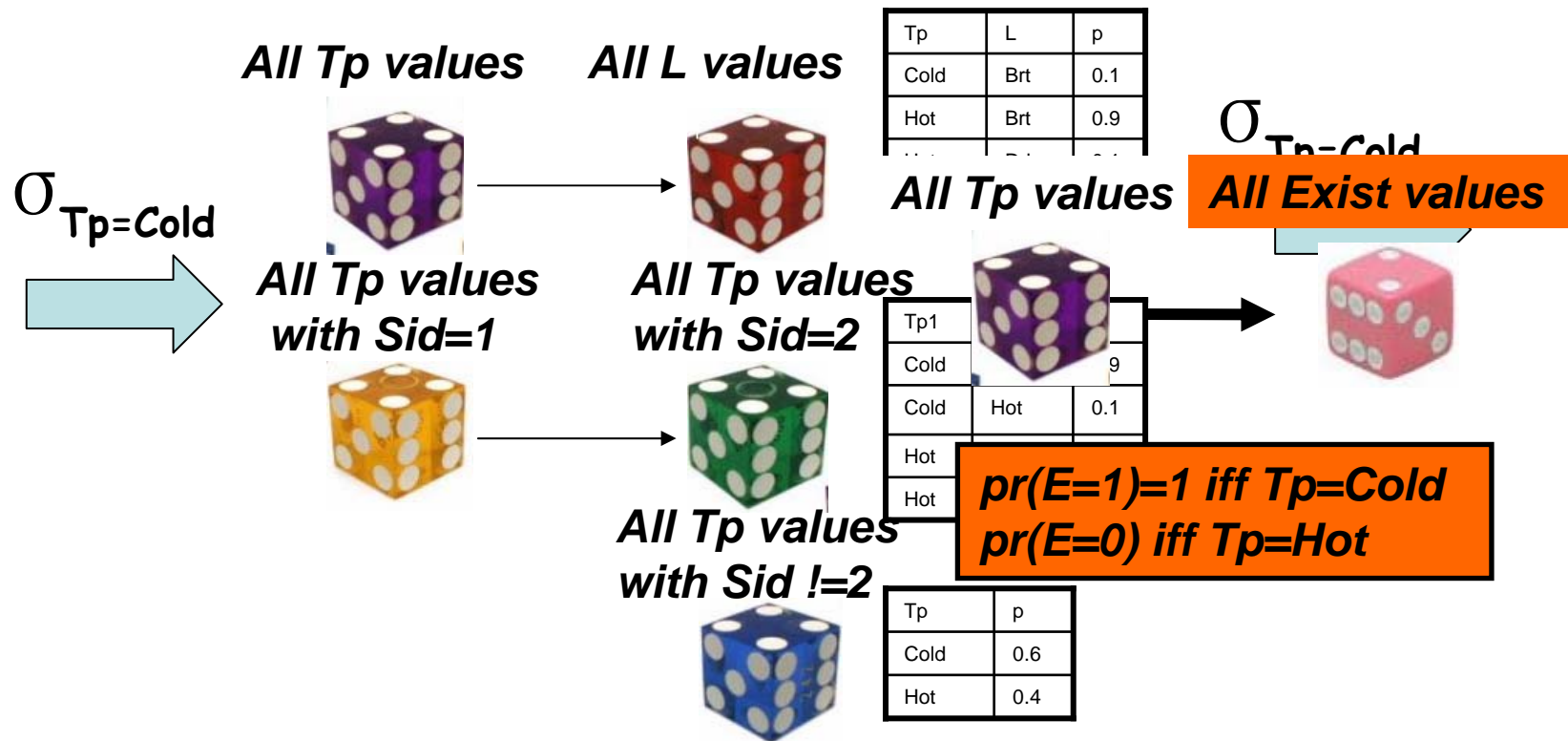


**$F_{FOBN} / T_p=Cold$**

# Selection

- Selection over Incomplete Relation  $R^p$
- **Selection over Model  $M_{FOBN}$**   $Sensor1(T, R, Sid, Tp^p, L^p, Exist(E)^p)$

$F_{FOBN}$  of  $Sensor1^p$



# Project & Join

- Project
  - Project over Incomplete Relation – projected attributes and correlated attributes
  - Project over Model – retrieve only part of the model relevant to the projected attributes
- Join
  - Join over Incomplete Relations with deterministic join condition (e.g.  $\text{Sensor1.Sid} = \text{Sensor2.Sid}$ )
  - Join over Models by merging the local models for  $\text{Exist}^p$  attribute
  - Probabilistic selection with probabilistic join condition (e.g.  $\text{Sensor1.Light}^p = \text{Sensor2.Light}^p$ )

# Optimizations (I)

- Selection over Incomplete Relation  $R^p$ 
  - **BayesBall Algorithm**
  - Model based Filtering

**Sensor1<sup>p</sup>**

	T	R	Sid	Tp <sup>p</sup>	L <sup>p</sup>
t1	1	1	1	Hot	X1
t2	1	1	2	Cold	Drk
t3	1	1	3	X2	X3
t4	1	2	1	X4	Brt
t5	1	2	2	Hot	X5
t6	1	2	3	X6	X7

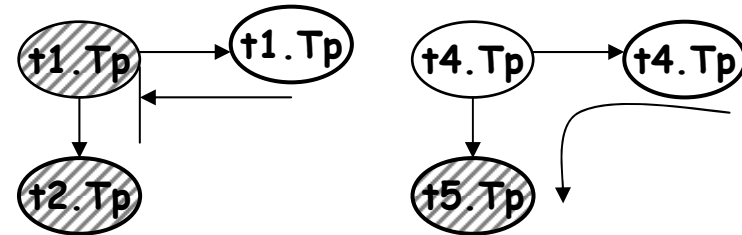
$\sigma_{Tp=Cold|Null}$

	T	R	Sid	Tp <sup>p</sup>	L <sup>p</sup>
t2	1	1	2	Cold	Drk
t3	1	1	3	X2	X3
t4	1	2	1	X4	Brt
t6	1	2	3	X6	X7

Compute BayesBall Algorithm over GBN

	T	R	Sid	Tp <sup>p</sup>	L <sup>p</sup>
t2	1	1	2	Cold	Drk
t3	1	1	3	X2	X3
t4	1	2	1	X4	Brt
t5	1	2	2	Hot	X5
t6	1	2	3	X6	X7

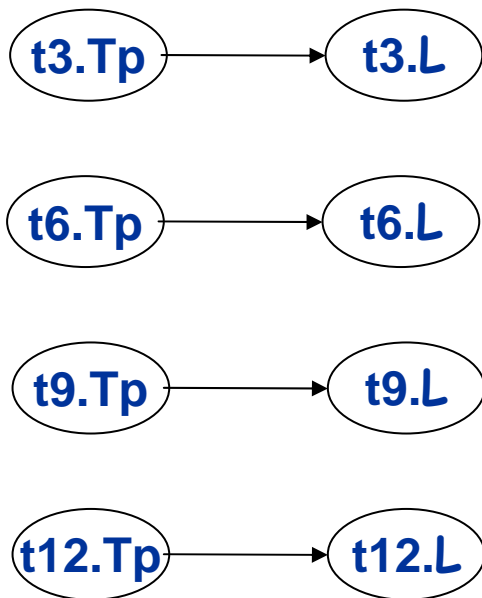
Grounded Bayesian Network (GBN) for  $F_{FOBN}$  (Sensor1)



# Optimizations (II)

*Sensor1<sup>p</sup>*

- Selection over Incomplete Relation  $R^p$ 
  - BayesBall Algorithm
  - Model based Filtering
- Simple First-order Inference Technique
  - **Sharing**



	T	R	Sid	Tp <sup>p</sup>	L <sup>p</sup>
t1	1	1	1	Hot	X1
t2	1	1	2	Cold	Drk
t3	1	1	3	X2	X3
t4	1	2	1	X4	Brk
t5	1	2	2	Hot	X5
t6	1	2	3	X6	X7
t7	2	1	1	Hot	X8
t8	2	1	2	Cold	Drk
t9	2	1	3	X9	X10
t10	2	2	1	X11	Brk
t11	2	2	2	Hot	X12
t12	2	2	3	X13	X14

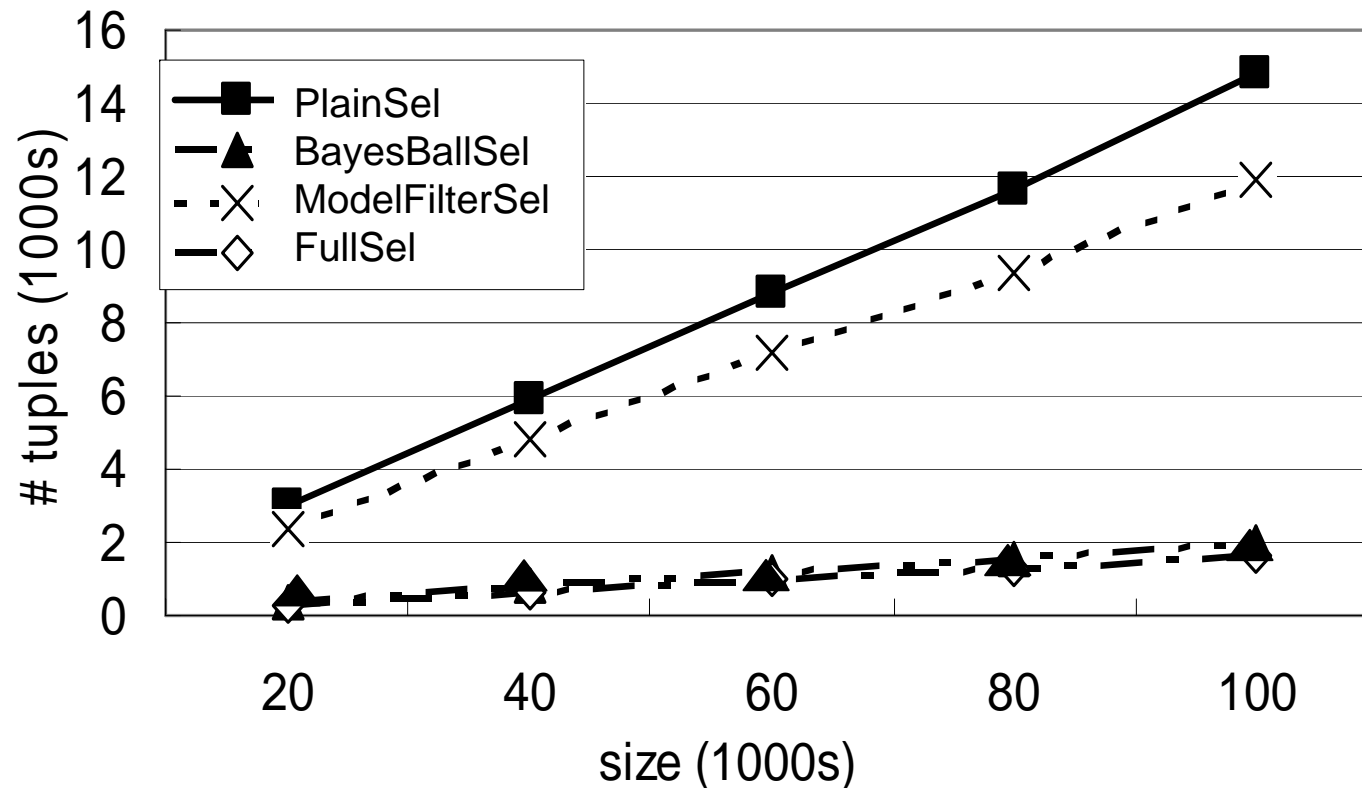
# Evaluation – Selection Algorithms

PlainSel: Selection over Incomplete Relation

BayesBallSel: Stop Transitive Closure using Bayes Ball Algorithm

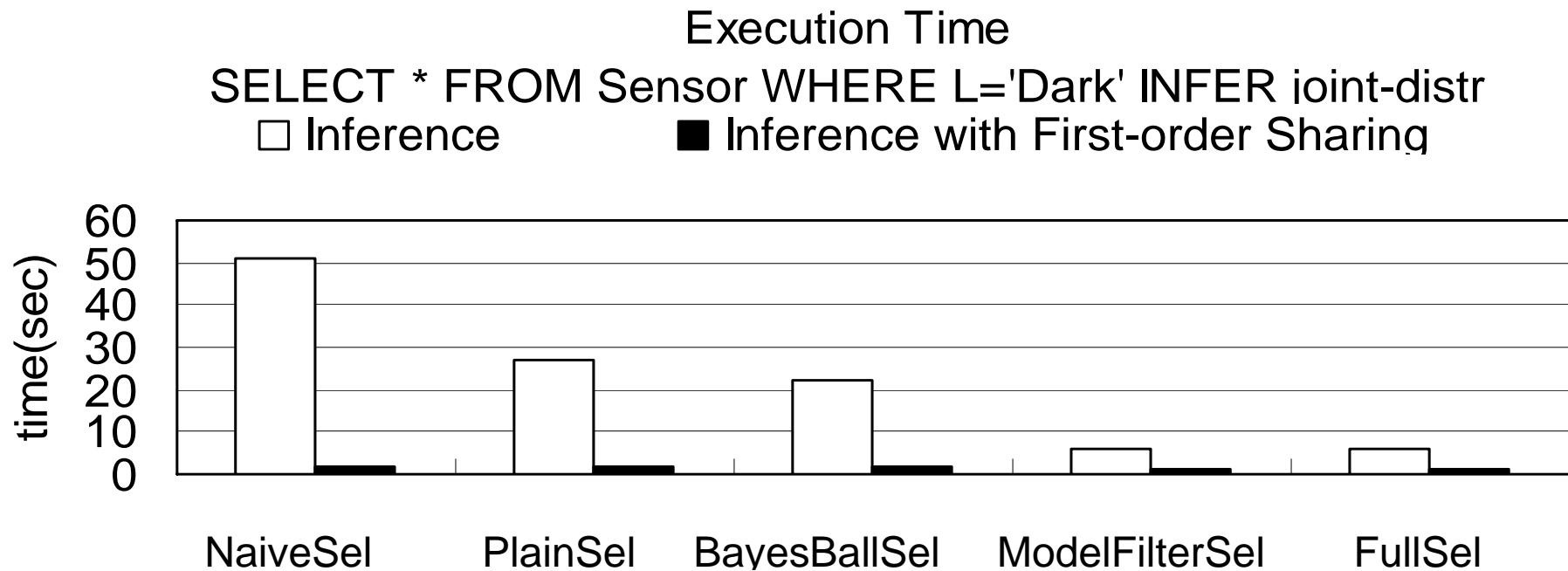
ModelFilterSel: Filter tuples with zero satisfying probability using Model

FullSel: Both BayesBall and ModelFilter Optimizations are used



# Evaluation – Inference Algorithms

First-order model enables the first-order inference optimizations.



# Current and Future Work

- First-order Inference & Model Learning
- Full System Implementation
- Aggregation Operators
- Query Optimizations
- Lineage Compression
- API Design

Questions?