BayesStore

Probabilistic Relational
Database System

# Managing Large, Uncertain Data Repositories with Probabilistic Graphical Models

Daisy Zhe Wang[+], Eirinaios Michelakis[+],
Minos Garofalakis[*+], Joseph M. Hellerstein[+]

University of California Berkeley[+], Yahoo! Research[*]

25th August 2008, VLDB

# Uncertainty in Real Systems

Sensor Networks

Data Extraction Systems

DBLife

Yahoo!/PSOX

IBM/Avatar/SystemT

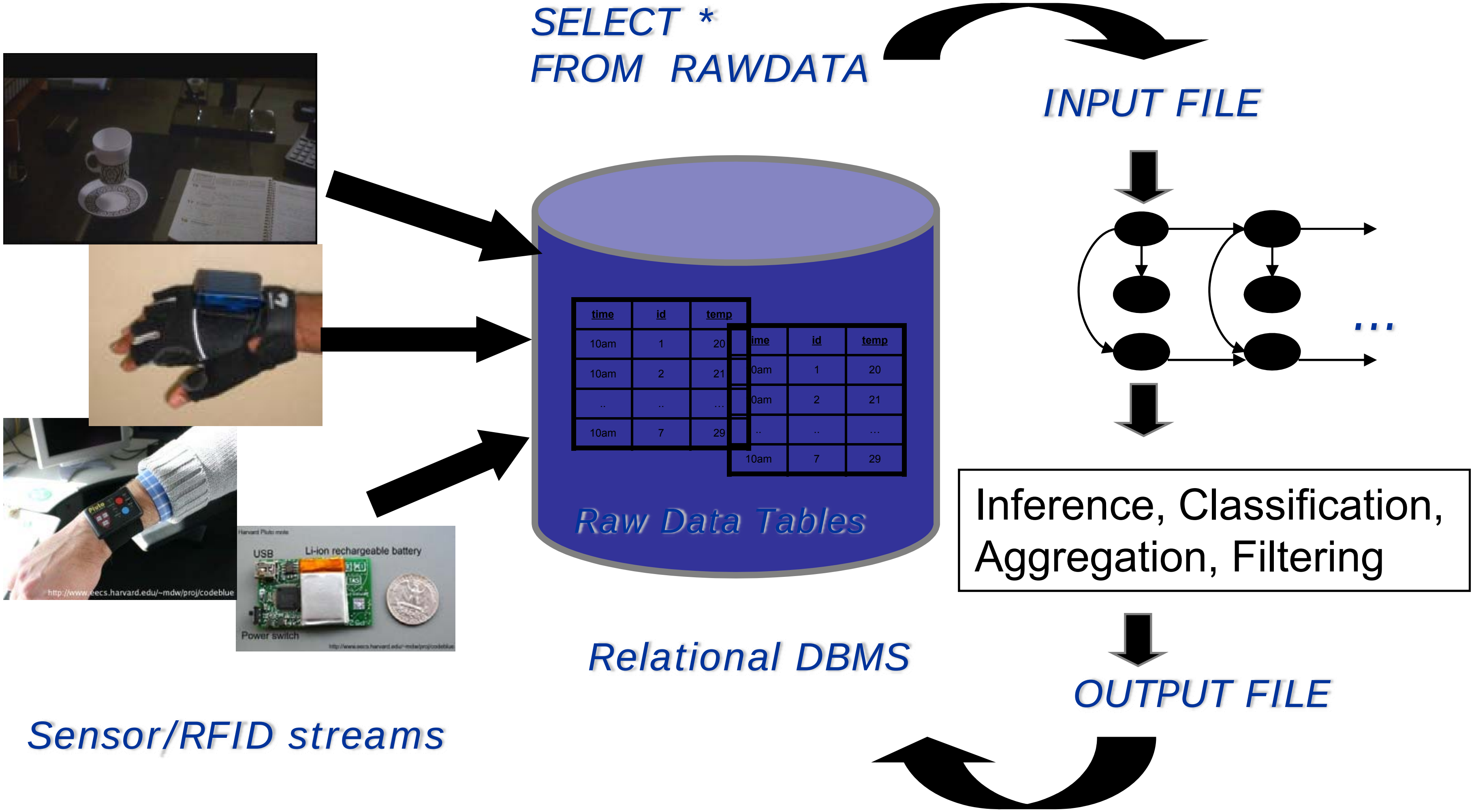
Social Networks

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


Harvard Pluto mote
USB
Li-ion rechargeable battery
Power switch
http://www.eecs.harvard.edu/~mdw/proj/codeblue
http://www.eecs.harvard.edu/~mdw/proj/codeblue
…

*SELECT **

*FROM RAWDATA*

*INPUT FILE*

| time | id | temp |
|---|---|---|
| 10am | 1 | 20 |
| 10am | 2 | 21 |
| .. | .. | … |
| 10am | 7 | 29 |

| ime | id | temp |
|---|---|---|
| 0am | 1 | 20 |
| 0am | 2 | 21 |
| .. | .. | … |
| 10am | 7 | 29 |

*Raw Data Tables*

Inference, Classification, Aggregation, Filtering

*Relational DBMS*

*OUTPUT FILE*

*Sensor/RFID streams*

# 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 -- $\mathbf{R^p}$
2. Distribution over Possible Worlds – **F**

**Sensor1(Time(T), Room(R), Sid, Temperature(Tp) $^p$, Light(L) $^p$)**

***Incomplete Relation of Sensor1$^p$***

| | T | R | Sid | $Tp^p$ | $L^p$ |
|---|---|---|---|---|---|
| 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 |

***Probabilistic Distribution of Sensor1$^p$***

$$\mathbf{F = Pr\ [X_1, \ldots, X_7\,]}$$

N: number of missing values
|X|: size of the domain

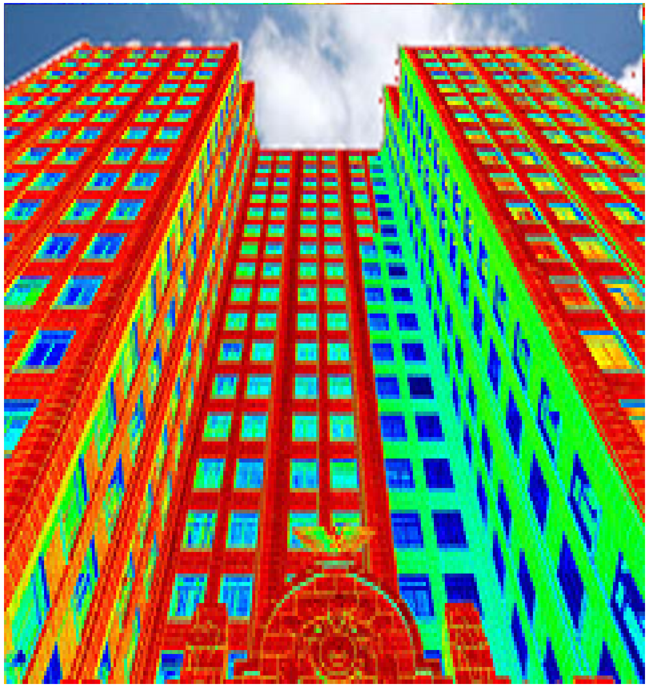
$$\mathbf{|F| = \Theta(|X|^N)}$$

# The Skyscrapers Example

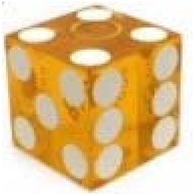
For all sensor in all rooms at all timestamp, Light and Temperature readings are correlated.

Light

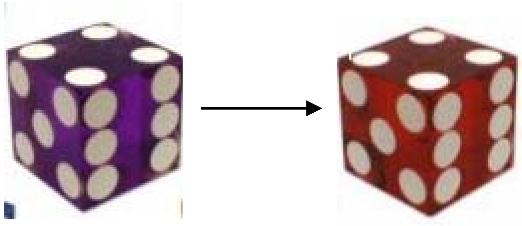
Temperature



# 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.*

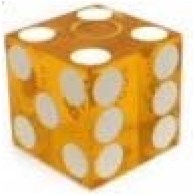
***Possible Worlds***

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

***Sensor1$^p$***

| | T | R | Sid | Tp$^p$ | L$^p$ |
|---|---|---|---|---|---|
| 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***



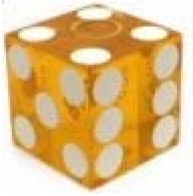
***All Tp values in Sensor1$^p$ with Sid=1***

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

***Sensor1$^p$***

|  | T | R | Sid | Tp$^p$ | L$^p$ |
|---|---|---|---|---|---|
| t1 | 1 | 1 | 1 | H | X1 |
| t2 | 1 | 1 | 2 | C d | Dr |
| t3 | 1 | 1 | 3 | X | X3 |
| t4 | 1 | 2 | 1 | X | Br |
| t5 | 1 | 2 | 2 | H | X5 |
| t6 | 1 | 2 | 3 | X | X7 |
| t7 | 2 | 1 | 1 | H | X8 |
| t8 | 2 | 1 | 2 | C d | Dr |
| t9 | 2 | 1 | 3 | X | X1 |
| t10 | 2 | 2 | 1 | X | Br |
| t11 | 2 | 2 | 2 | H | X1 |
| t12 | 2 | 2 | 3 | X 3 | X1 |

***Stripe (FO Variable) Definitions***



***All Tp values in Sensor1$^p$ with Sid=1***

***All Tp values in Sensor1$^p$ with Sid=2***

***All Tp values in Sensor1$^p$ with Sid !=2***

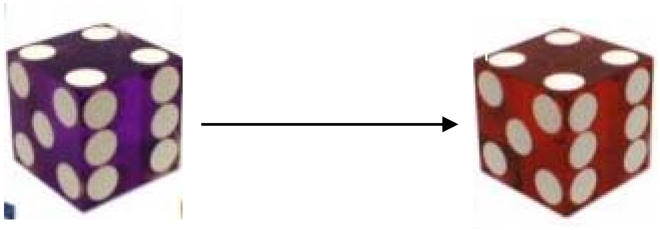
***All Tp values in Sensor1$^p$***

***All L values in Sensor1$^p$***

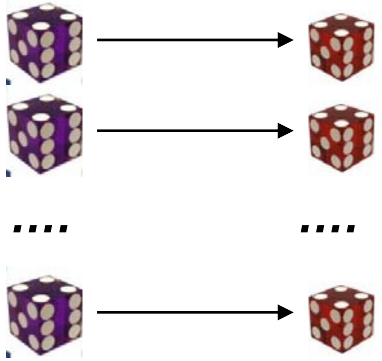
# F as a First-order Bayesian Model

***Mapping between Stripes***

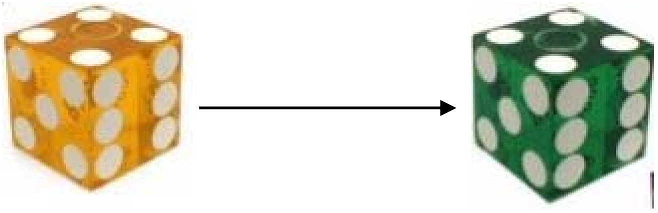
***All Tp values*** ***All L values***

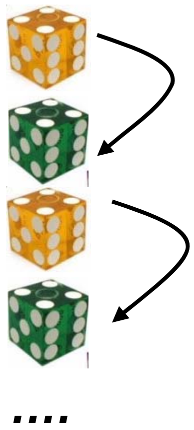


***All Tp values*** ***All L values***



***All Tp values with Sid=1*** ***All Tp values with Sid=2***





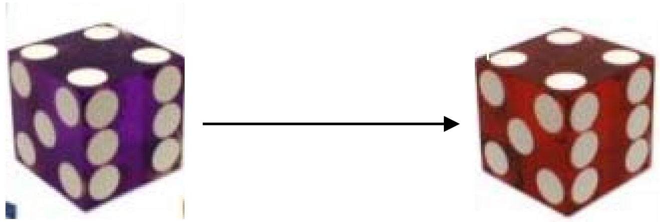
***All Tp values with Sid=1***

***All Tp values with Sid=2***

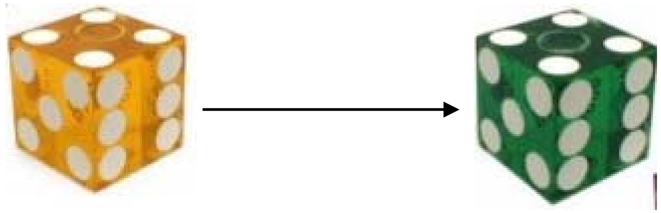
# F as a First-order Bayesian Model

***First-order Factor Definitions***

***All Tp values*** ***All L values***



***All Tp values with Sid=1*** ***All Tp values with Sid=2***



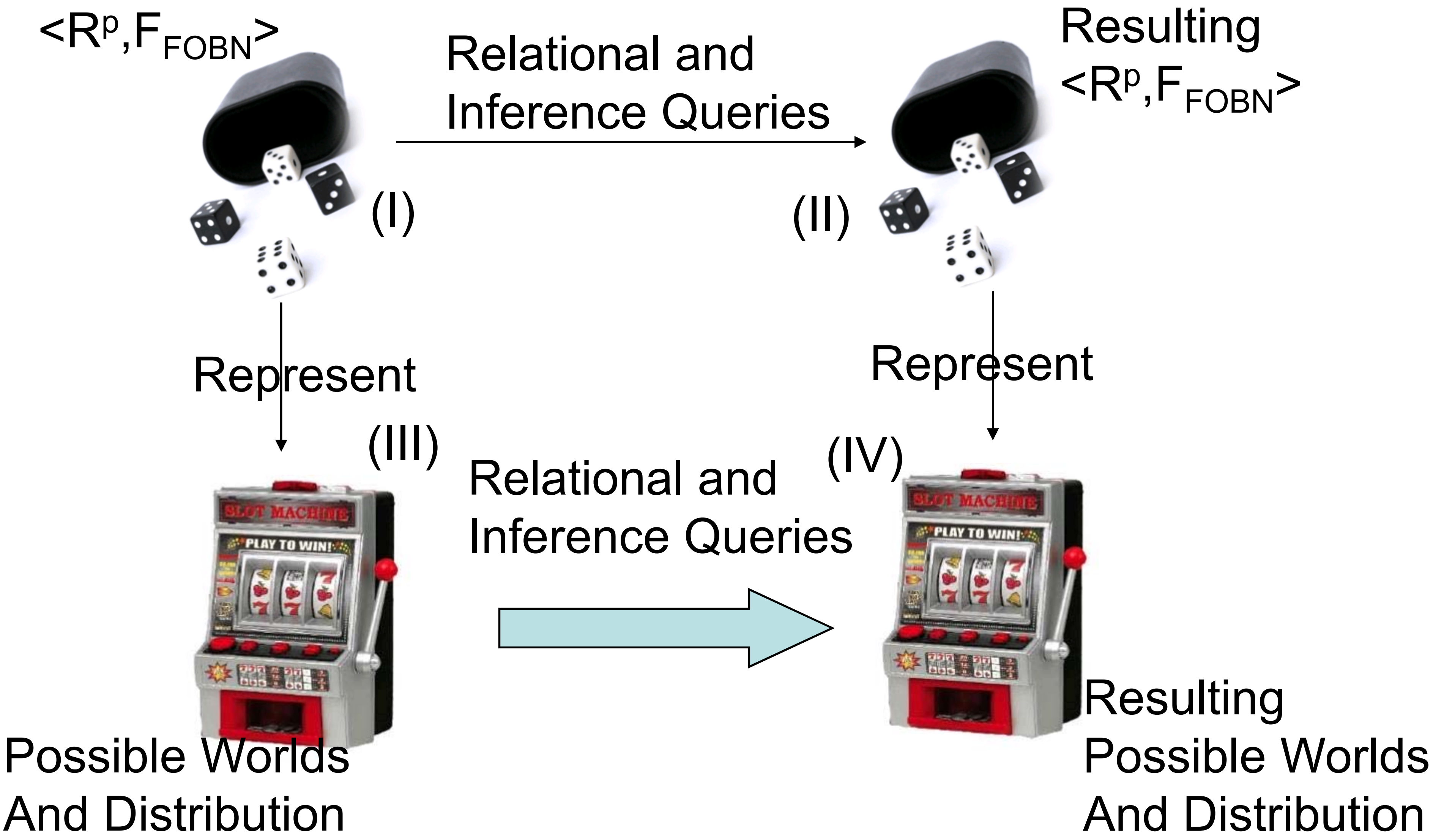
***All Tp values with Sid !=2***

| Tp | L | p |
|---|---|---|
| Cold | Brt | 0.1 |
| Hot | Brt | 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



$<R^p,F_{FOBN}>$

Relational and Inference Queries

Resulting $<R^p,F_{FOBN}>$

(I)

(II)

Represent

Represent

(III)

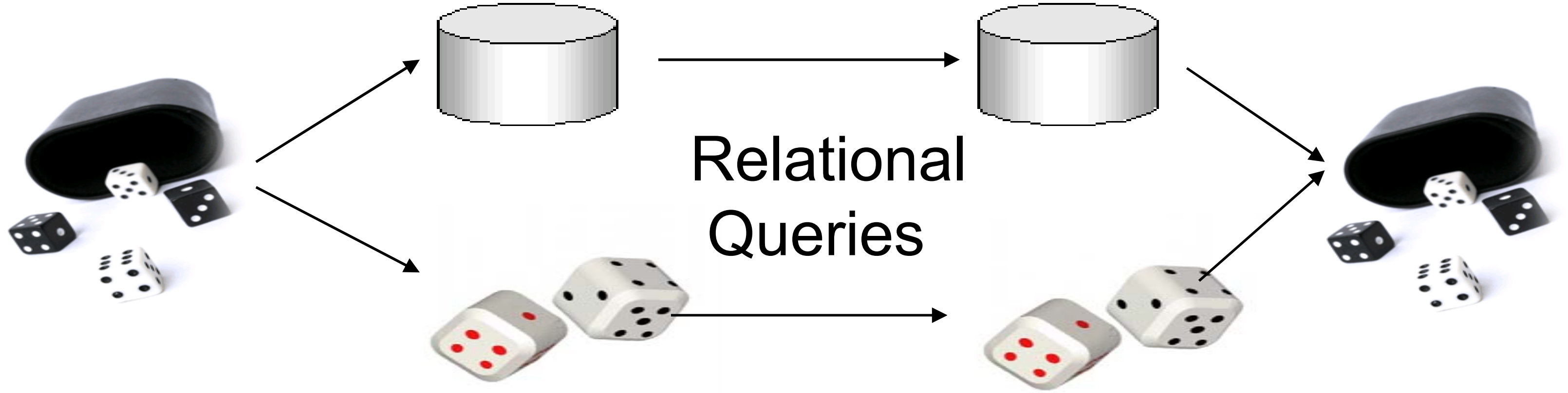
(IV)

Relational and Inference Queries

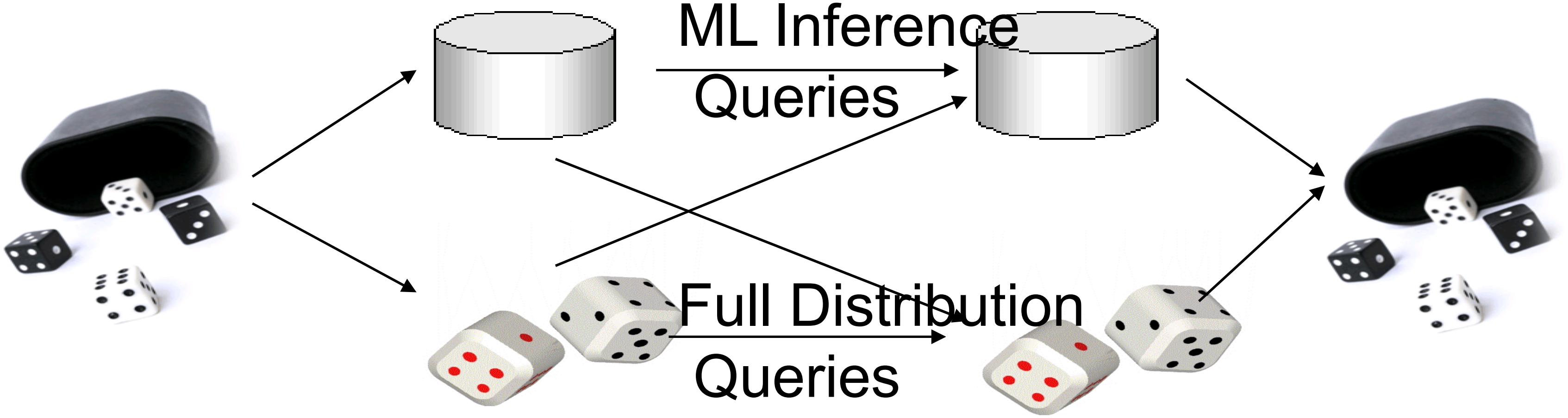
Possible Worlds And Distribution

Resulting Possible Worlds And Distribution

# Query Algebra



Relational Queries



ML Inference Queries

Full Distribution Queries

# Selection

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

***Sensor1$^p$***

|  | T | R | Sid | Tp$^p$ | L$^p$ |
|---|---|---|---|---|---|
| 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 |

$\sigma_{Tp=Cold}$

***Sensor1$^p$***

|  | T | R | Sid | Tp$^p$ | L$^p$ |
|---|---|---|---|---|---|
| t2 | 1 | 1 | 2 | Cold | Drk |
| t3 | 1 | 1 | 3 | X2 | X3 |
| t4 | 1 | 2 | 1 | X4 | Brt |
| t6 | 1 | 2 | 3 | X6 | X7 |

|  | T | R | Sid | Tp$^p$ | L$^p$ |
|---|---|---|---|---|---|
| t2 | 1 | 1 | 2 | Cold | X3 |

# Selection

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

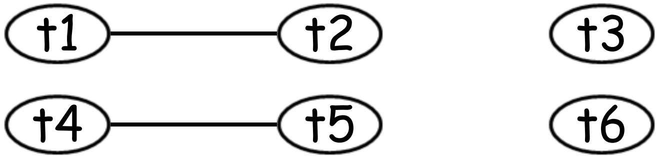
***Sensor1$^p$***

| | T | R | Sid | Tp$^p$ | L$^p$ |
|---|---|---|---|---|---|
| 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$^p$ | L$^p$ |
|---|---|---|---|---|---|
| t2 | 1 | 1 | 2 | Cold | Drk |
| t3 | 1 | 1 | 3 | X2 | X3 |
| t4 | 1 | 2 | 1 | X4 | Brt |
| t6 | 1 | 2 | 3 | X6 | X7 |

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



Compute Transitive Closure over TCG

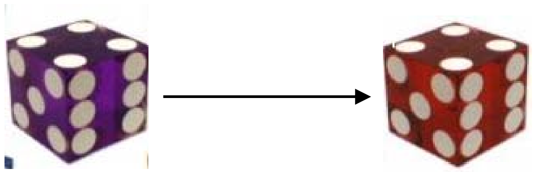
| T | R | Sid | Tp$^p$ | L$^p$ | |
|---|---|---|---|---|---|
| 1 | 1 | 1 | Hot | X1 | t1 |
| 1 | 1 | 2 | Cold | Drk | t2 |
| 1 | 1 | 3 | X2 | X3 | t3 |
| 1 | 2 | 1 | X4 | Brt | t4 |
| 1 | 2 | 2 | Hot | X5 | t5 |
| 1 | 2 | 3 | X6 | X7 | t6 |

# Selection

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

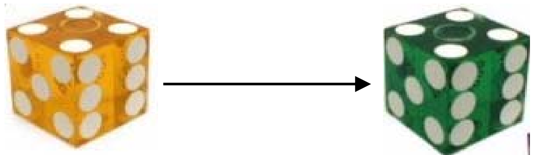
***Probabilistic Distribution $F_{FOBN}$ of Sensor1$^p$***

***All Tp values*** ***All L values***



| Tp | L | p |
|---|---|---|
| Cold | Brt | 0.1 |
| Hot | Brt | 0.9 |
| Hot | Drk | 0.1 |
| Cold | Drk | 0.9 |

***All Tp values with Sid=1*** ***All Tp values with Sid=2***

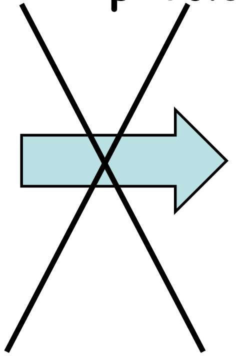


| Tp1 | Tp2 | p |
|---|---|---|
| Cold | Cold | 0.9 |
| Cold | Hot | 0.1 |
| Hot | Hot | 0.9 |
| Hot | Cold | 0.1 |

***All Tp values with Sid !=2***

| Tp | p |
|---|---|
| Cold | 0.6 |
| Hot | 0.4 |

$\sigma_{Tp=Cold}$



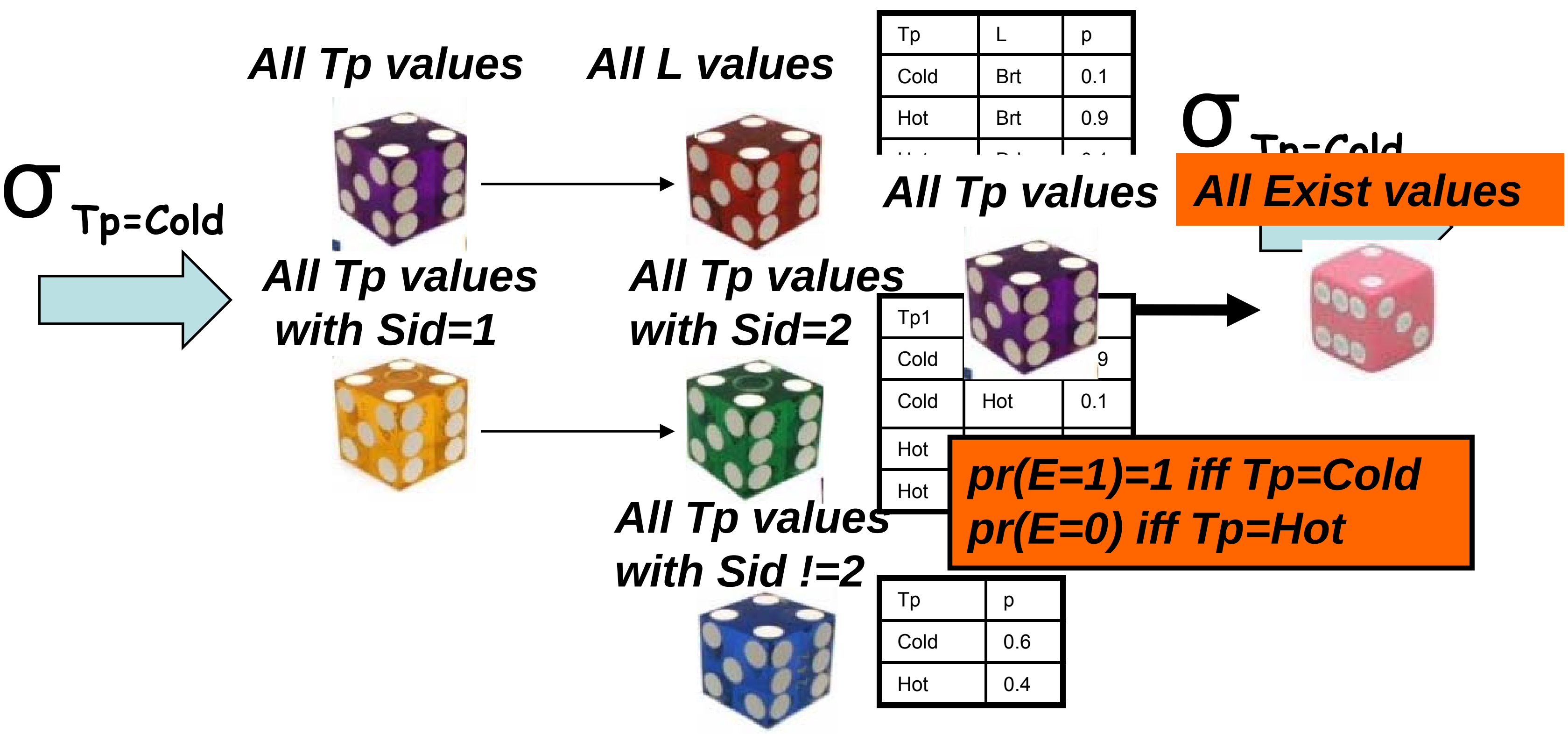
***$F_{FOBN}$ | Tp=Cold***

# Selection

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

***Sensor1(<u>T, R, Sid</u>, $Tp^p$, $L^p$,Exist(E)$^p$)***

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



$\sigma_{Tp=Cold}$

***All Tp values***

***All L values***

***All Tp values with Sid=1***

***All Tp values with Sid=2***

***All Tp values with Sid !=2***

***All Tp values***

$\sigma_{Tp=Cold}$

***All Exist values***

| Tp | L | p |
|---|---|---|
| Cold | Brt | 0.1 |
| Hot | Brt | 0.9 |
| [illegible] | [illegible] | [illegible] |

| Tp1 | | |
|---|---|---|
| Cold | | 9 |
| Cold | Hot | 0.1 |
| Hot | | |
| Hot | | |

***pr(E=1)=1 iff Tp=Cold***
***pr(E=0) iff Tp=Hot***

| Tp | p |
|---|---|
| Cold | 0.6 |
| Hot | 0.4 |

# 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. Sensor1.Sid = Sensor2.Sid)
  - Join over Models by merging the local models for $\text{Exist}^p$ attribute
  - Probabilistic selection with probabilistic join condition (e.g. $\text{Sensor1.Light}^p$ = $\text{Sensor2.Light}^p$)

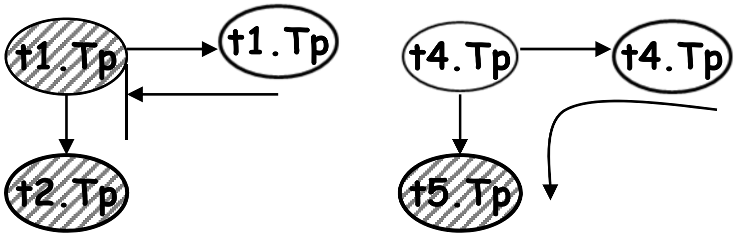
# Optimizations (I)

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

***Sensor1*** $^p$

| | T | R | Sid | $Tp^p$ | $L^p$ |
|---|---|---|---|---|---|
| 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 |

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



$\sigma_{Tp=Cold|Null}$

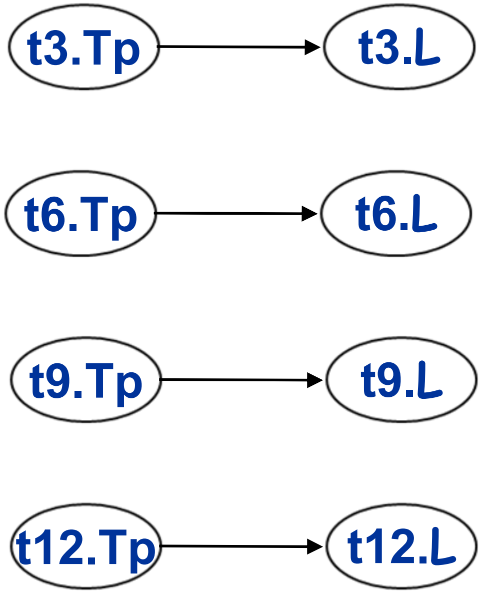
| | T | R | Sid | $Tp^p$ | $L^p$ |
|---|---|---|---|---|---|
| t2 | 1 | 1 | 2 | Cold | Drk |
| t3 | 1 | 1 | 3 | X2 | X3 |
| t4 | 1 | 2 | 1 | X4 | Brt |
| t6 | 1 | 2 | 3 | X6 | X7 |

Compute BayesBall Algorithm over GBN

| T | R | Sid | $Tp^p$ | $L^p$ | |
|---|---|---|---|---|---|
| 1 | 1 | 2 | Cold | Drk | t2 |
| 1 | 1 | 3 | X2 | X3 | t3 |
| 1 | 2 | 1 | X4 | Brt | t4 |
| 1 | 2 | 2 | Hot | X5 | t5 |
| 1 | 2 | 3 | X6 | X7 | t6 |

# Optimizations (II)

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

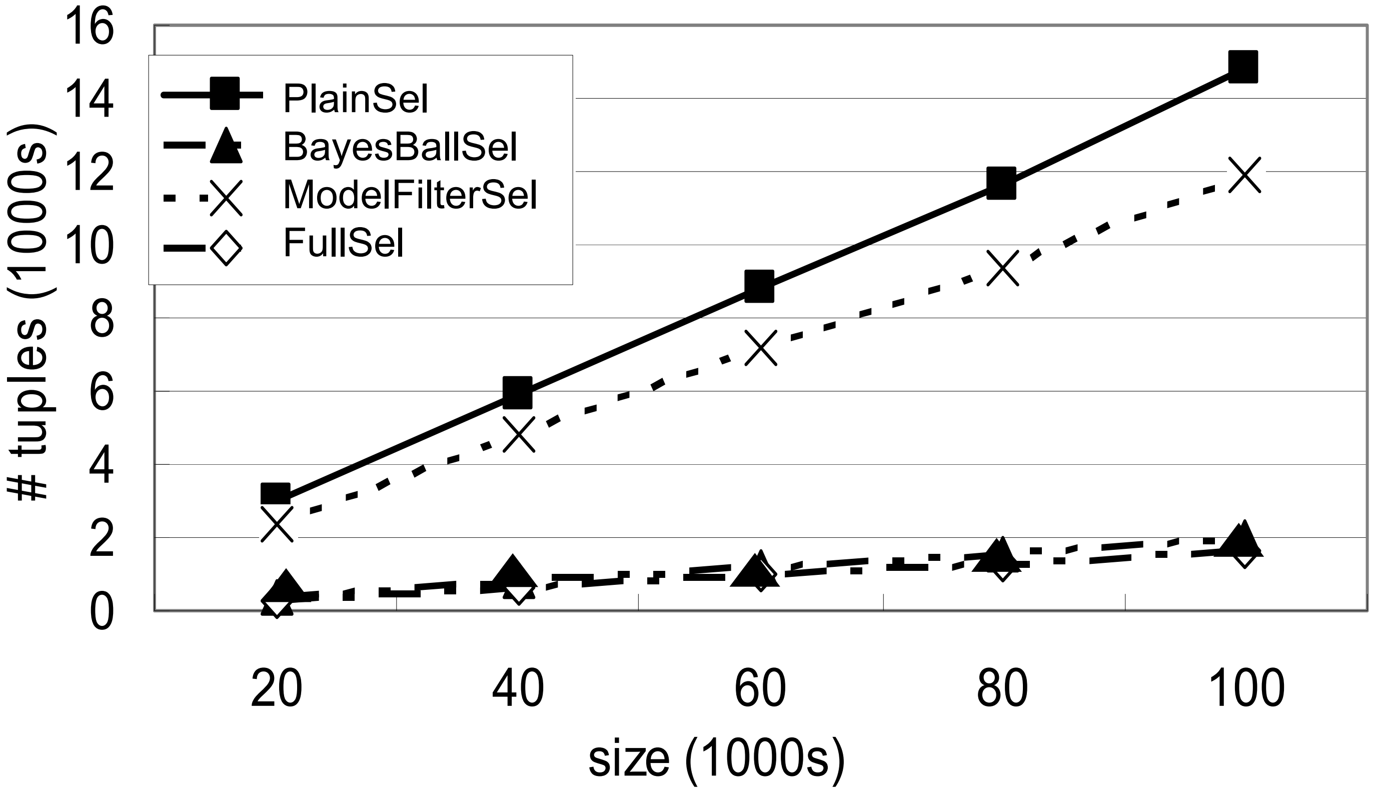


***Sensor1$^p$***

| | T | R | Sid | Tp$^p$ | L$^p$ |
|---|---|---|---|---|---|
| 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 |
| t7 | 2 | 1 | 1 | Hot | X8 |
| t8 | 2 | 1 | 2 | Cold | Drk |
| t9 | 2 | 1 | 3 | X9 | X10 |
| t10 | 2 | 2 | 1 | X11 | Brt |
| t11 | 2 | 2 | 2 | Hot | X12 |
| t12 | 2 | 2 | 3 | X13 | X14 |

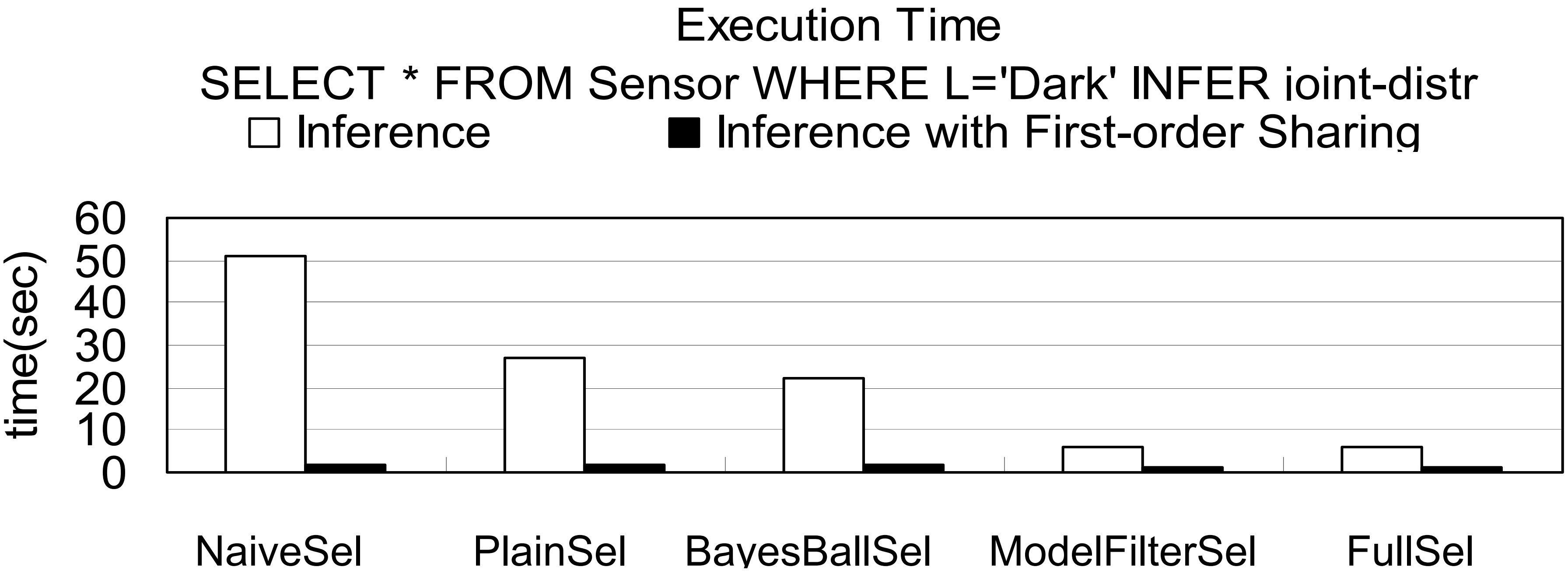
# Evaluation – Selection Algorithms

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.



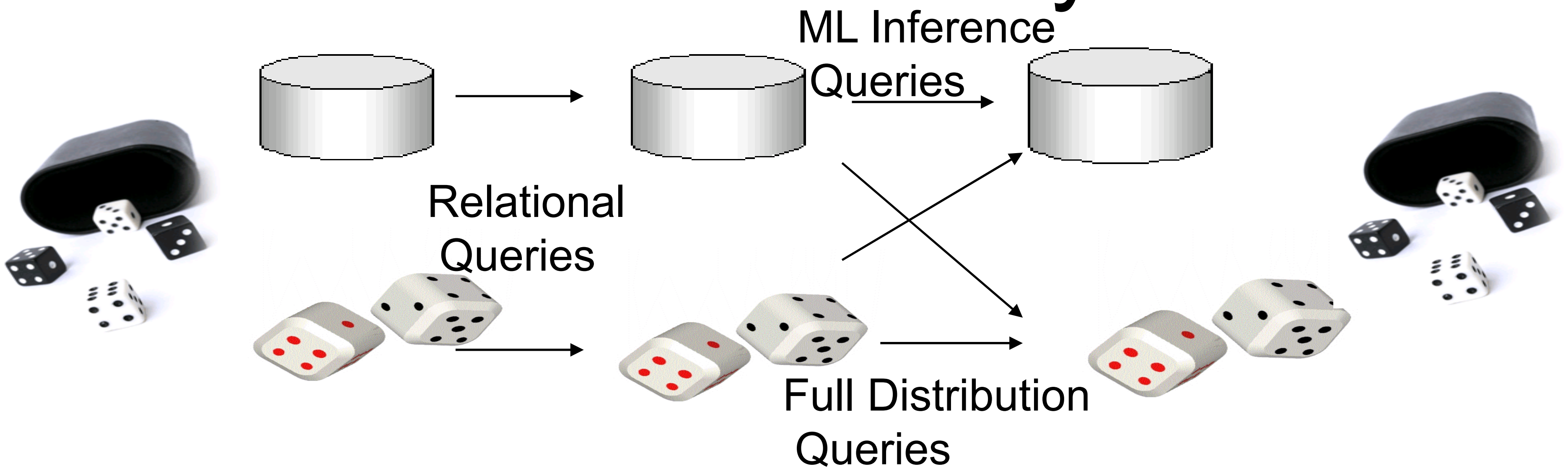
# Current and Future Work

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

# Questions?

# Backup Slides

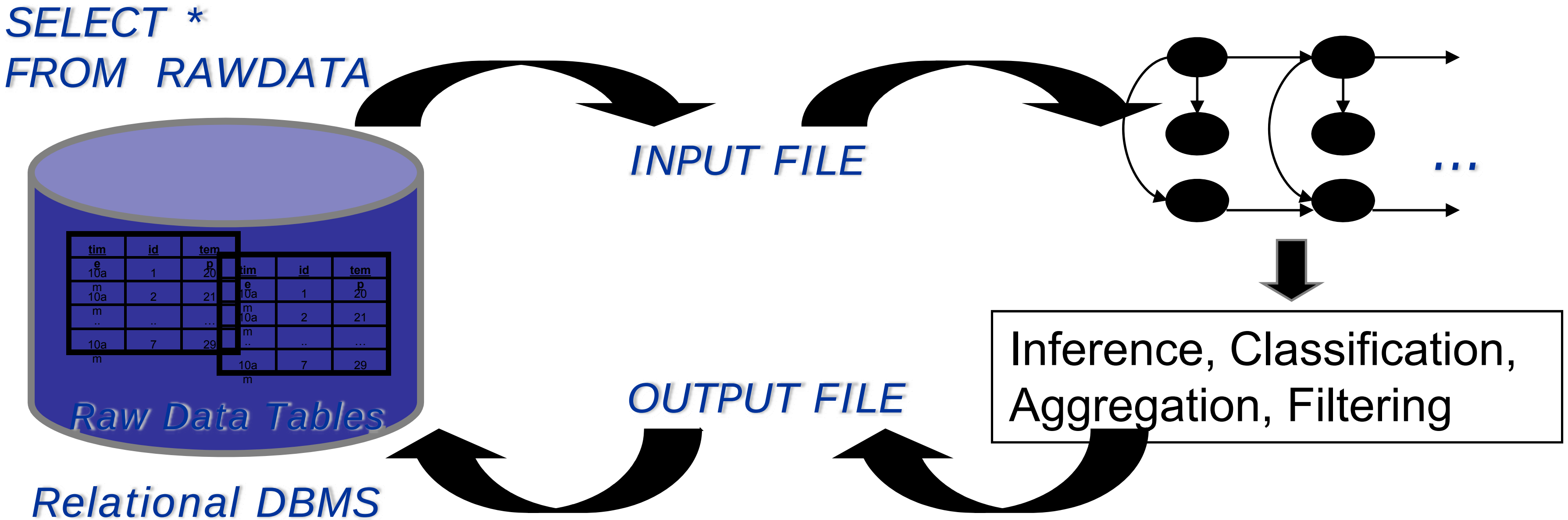
# Life of a Query



ML Inference Queries

Relational Queries

Full Distribution Queries



*SELECT **
*FROM RAWDATA*

*INPUT FILE*

…

| time | id | temp |
|---|---|---|
| 10am | 1 | 20 |
| 10am | 2 | 21 |
| .. | .. | … |
| 10am | 7 | 29 |

| time | id | temp |
|---|---|---|
| 10am | 1 | 20 |
| 10am | 2 | 21 |
| .. | .. | … |
| 10am | 7 | 29 |

*Raw Data Tables*

*OUTPUT FILE*

Inference, Classification, Aggregation, Filtering

*Relational DBMS*