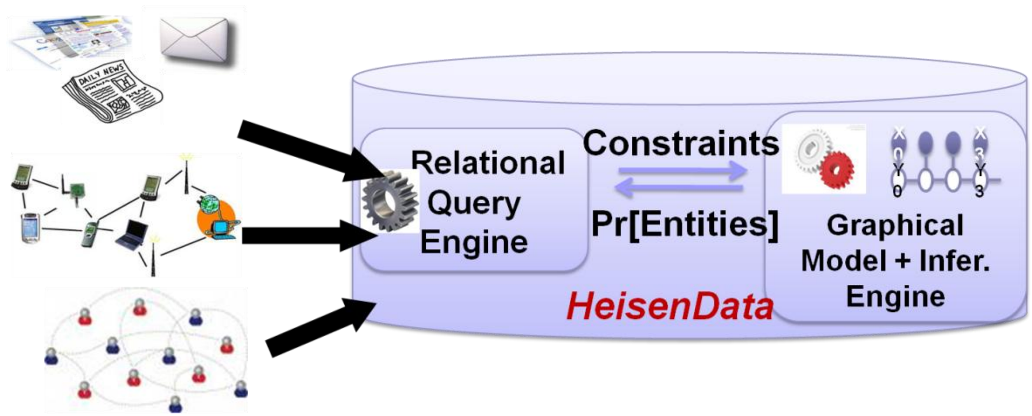
HeisenData : Towards Next-Generation Uncertain Database Systems



Minos Garofalakis**

Technical University of Crete, SoftNet Lab

minos@softnet.tuc.gr

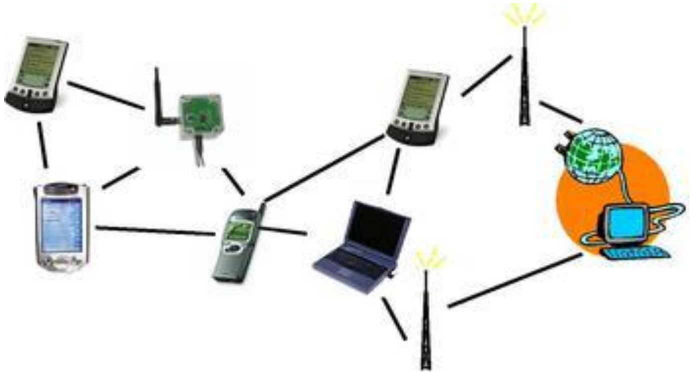
*** Partially supported by an FP7 Marie-Curie IRG Fellowship*

***Thanks to:* D. Wang, J. Hellerstein, M. Franklin, G. Cormode, A. Deligiannakis, E. Ioannou, E. Vazaios**

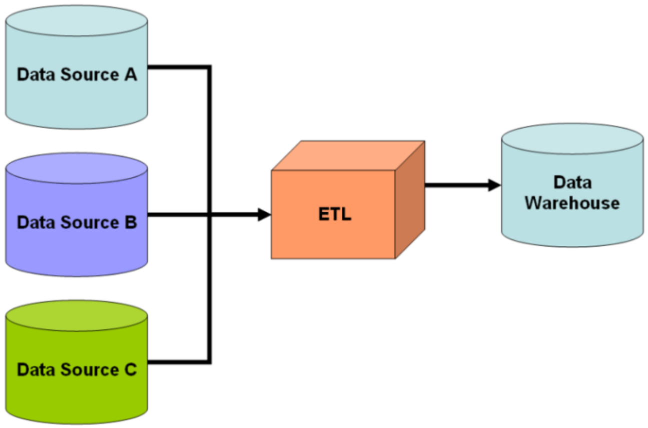
Probabilistic (Big) Data Analytics

Information Extraction Systems

Sensor Networks



Data Integration Systems



Social Networks

Information Extraction (IE)

Extracting structured entities and relationships from unstructured text

- “We are pleased that today's agreement guarantees our corporation will maintain a significant and long term presence in the Big Apple," McGraw-Hill president Harold McGraw III said in a statement.

--- *From New York Times April 24, 1997*

Information Extraction (IE)

- "We are pleased that today's agreement guarantees our corporation will maintain a significant and long term presence in the Big Apple," McGraw-Hill president Harold McGraw III said in a statement.

(prob=0.8)

--- *From New York Times April 24, 1997*

Labels:

Person Company Location Other

Information Extraction (IE)

- “We are pleased that today's agreement guarantees our corporation will maintain a significant and long term presence in the Big Apple," McGraw-Hill president Harold McGraw III said in a statement.

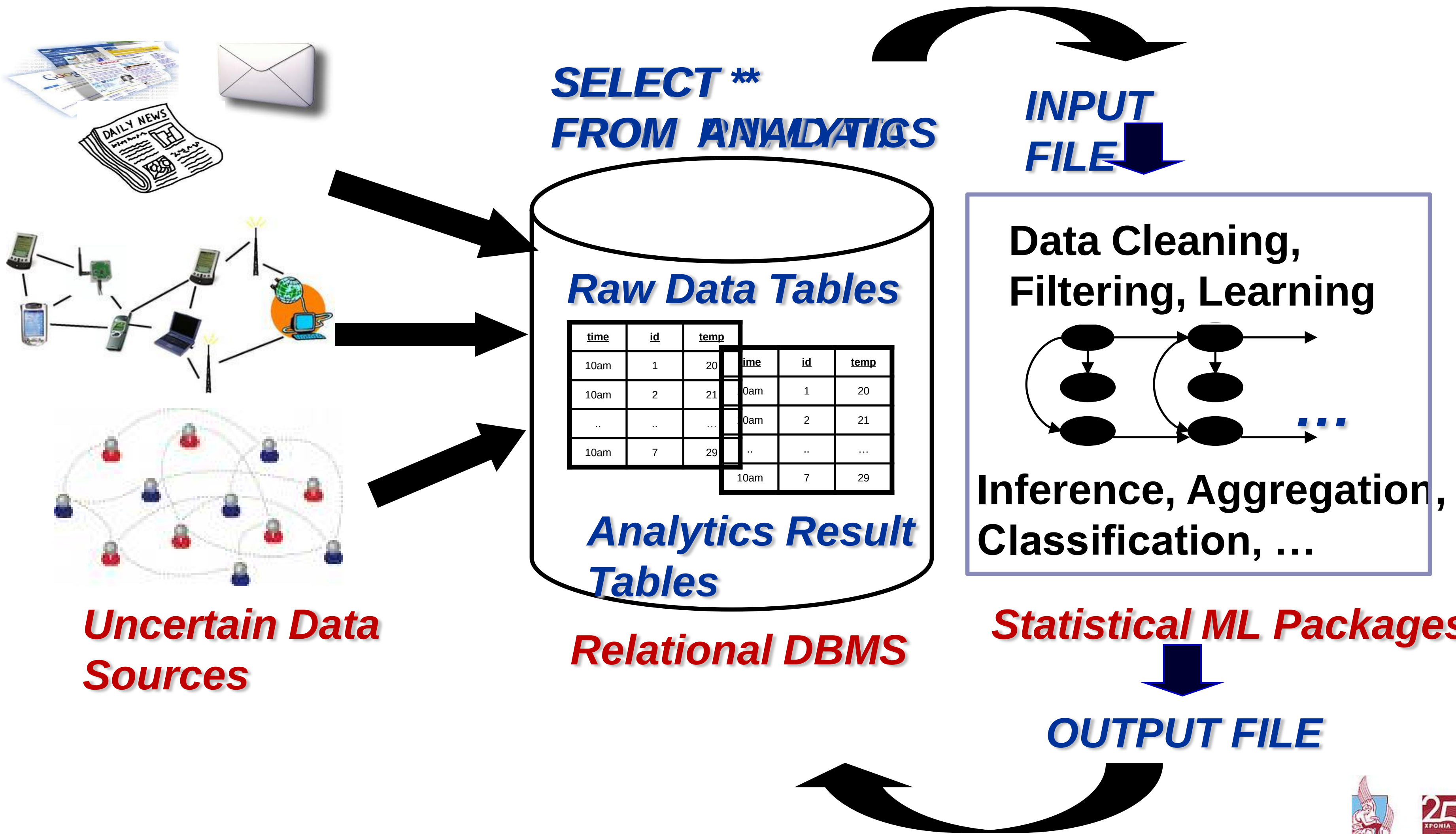
(prob=0.75)

--- *From New York Times April 24, 1997*

Labels:

Person Company Location Other

Standard Uncertain Data Analysis Loop



What's wrong with this picture…??

- All interesting data processing done *outside* the database!
- Lose *all key benefits of a database system (*30+ years of R&D)
  - Declarative querying, Persistence, Indexing, Caching, Parallelization, Automatic optimization, …
  - Poor performance, poor scalability
- No sharing of data/knowledge/abstractions, duplication of effort
- Information loss
  - Focus on top-few results, rather than *possible-world semantics*

Early Work on Probabilistic DBs (PDBs)

Simplistic uncertainty models that easily map to existing DB architectures

- Independent tuple-level confidences and attribute-value options (OR-tuples)

S^P

	A	B	
s1	'm'	1	**0.8**
s2	'n'	1	**0.5**

T^P

	C	D	
t1	1	'p'	**0.6**

MystiQ (UW) ***[VLDB04]***

(Witness, Car)	
(Amy,Honda):0.5 || (Amy,Toyota):0.3 || (Amy,Mazda):0.2	
(Betty,Acura):0.6	**?**

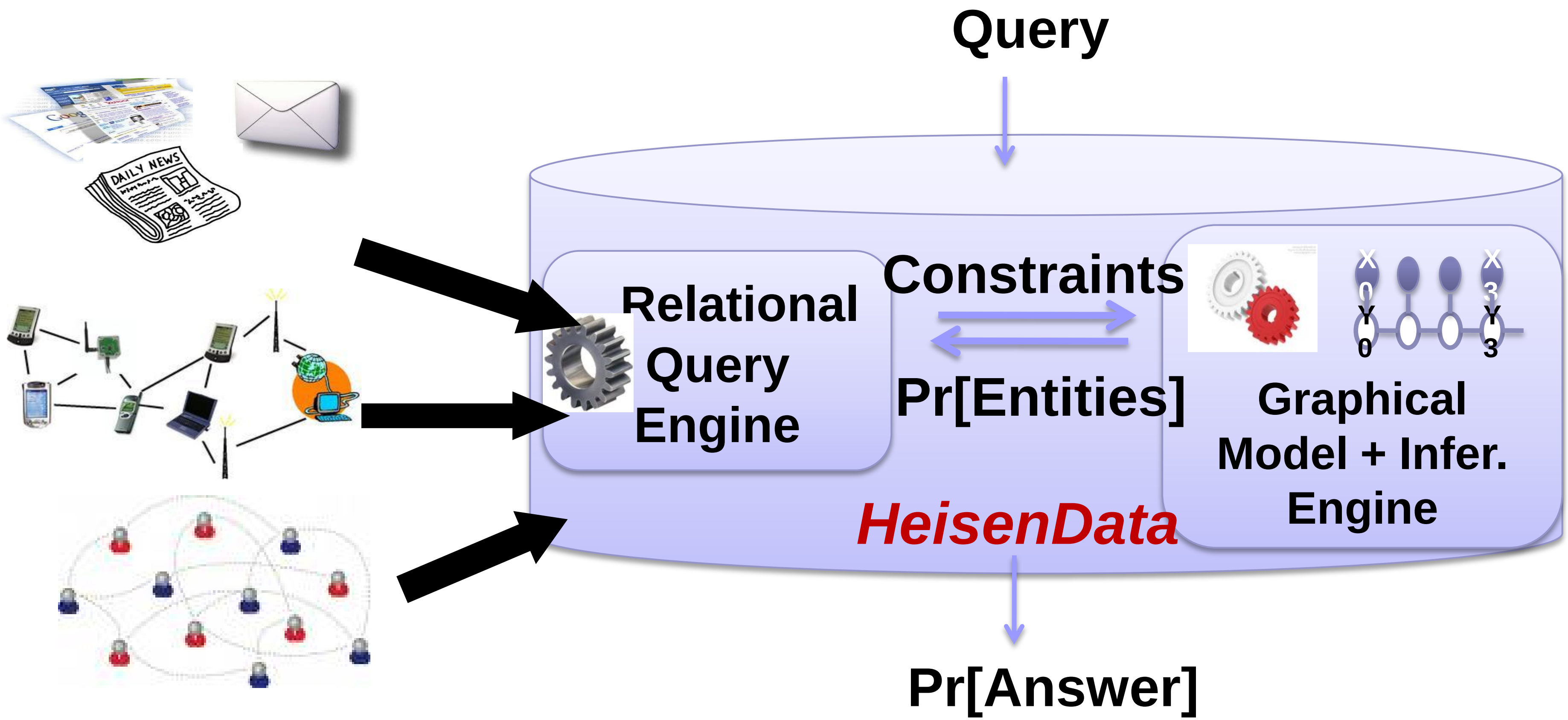
Trio (Stanford) ***[VLDB06]***

More Recent Work on PDBs…

- **MayBMS (Cornell):** Model correlations through factored relational table representations
- **PrDB (UMD):** Capture correlations using propositional/ grounded (per-tuple) Bayesian nets

- ***HeisenData*** **(Berkeley, now at TUC):** *Scalable, integrated data-management & probabilistic-reasoning platform*
  - *(First-Order (FO))* statistical models and reasoning as "first-class" citizens in the DBMS
  - Query processing = relational ops + statistical inference
    - *"Possible worlds"* semantics (data + stat model)
  - Application domains: Sensors, IE

HeisenData – Architecture



Uncertain Data Sources

Prototypes built on top of PostgreSQL 8.4

Key *HeisenData* Challenges

- What are the right language, physical/logical algebra, user interface?
  - Completeness, soundness
  - Expressiveness & ease of use
  - Extensibility (stat models, inference techniques, …)

- Query Processing & Optimization
  - Probabilistic queries with relational and inference operators!
  - Optimization & Approximation – Statistics for probabilistic data?
  - Inference is *expensive!*
    - Exploit massive parallelism (e.g., Hadoop) and/or approximation?
  - Physical DB design (indexes, access structs, views, …)?
  - Concrete Application Domains: Information Extraction

Talk Outline

- Introduction, Motivation, Challenges
- Example Data Model and Relational Query Processing [VLDB08]
- Managing Inference for Information Extraction [ICDE10,VLDB10,SIGMOD11]
- Statistics for Probabilistic Data [ICDE09,VLDB09]
- Conclusions & Future Work

HeisenData Example [VLDB08]

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}

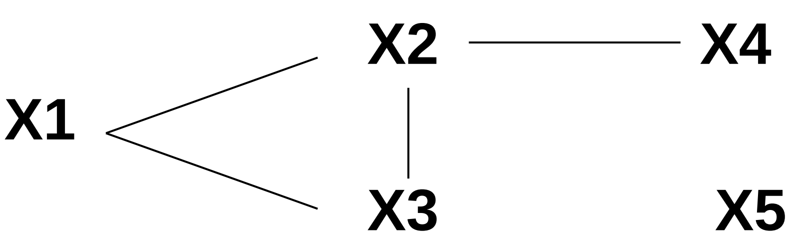
$$\mathbf{F = Pr\,[X_1, \ldots, X_7\,]}$$

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

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

Probabilistic Graphical Models (PGMs)

- PGM can compactly represent a joint PDF over large numbers of random variables (RVs) with complex correlation patterns
  - Take advantage of conditional independences
- Specified by: (1) Set of RVs, and (2) Set of factors over RVs
- Joint PDF = take product of all factors and normalize



Joint = (1/Z) f(X1,X2,X3) f(X2,X4) f(X5)

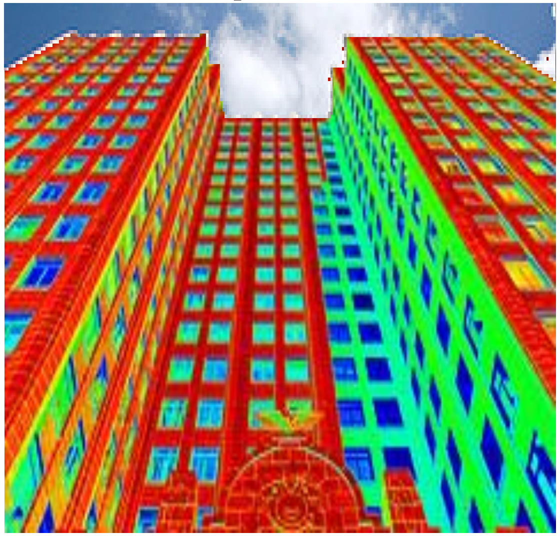
- Inference tasks
  - Find *mode* or *top-k* joint distribution points
  - Find *marginal PDF* on subset of RVs

First-Order (FO) PGMs

- Define factors/correlation patterns over FO families of RVs
  - RVs sharing the same correlation pattern
  - ***[VLDB08] RV stripes*** : defined using SQL queries over the incomplete relation schema
- Much more concise representation of joint PDF

Light

Temperature



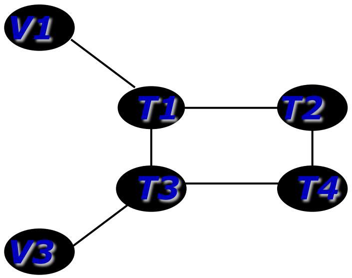
For all sensor in all rooms at all timestamps, Light and Temperature readings are correlated

HeisenData Data Model

time	id	temp	volt
10am	1	20	2.5
10am	2	21	XXX
..	..	…	
10am	7		2.8

Evidence Table(s)

+



FO Graphical Model
(factors stored as relational tables)

Prob=0.4

time	id	temp	volt
10am	1	20	2.5
10am	2	21	2.7
..	..	…	
10am	7	26	2.8

Prob=0.3

time	id	temp	volt
10am	1	20	2.5
10am	2	21	2.7
..	..	…	
10am	7	28	2.8

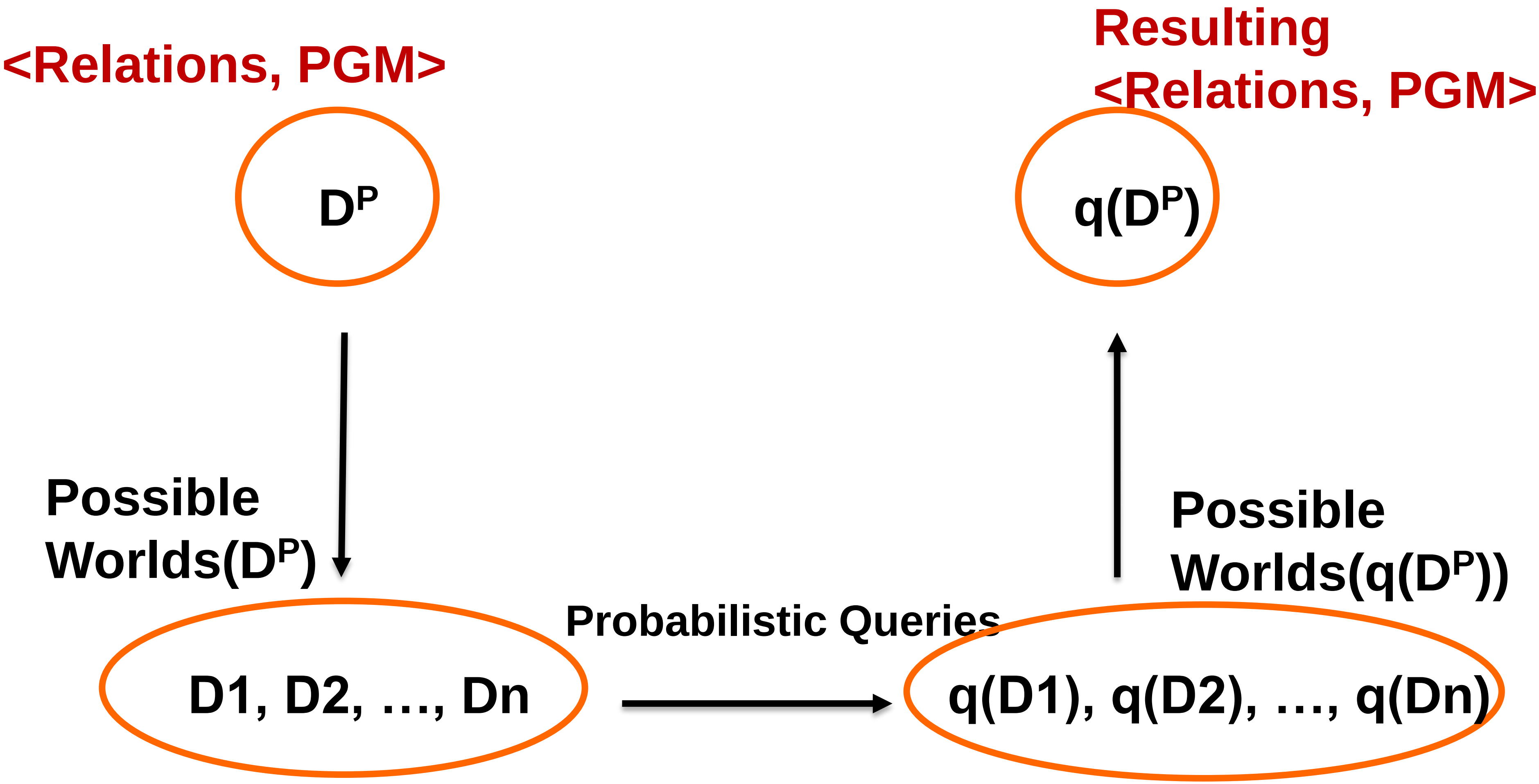
Prob=0.3

time	id	temp	volt
10am	1	20	2.5
10am	2	21	2.7
..	..	…	
10am	7	26	2.8

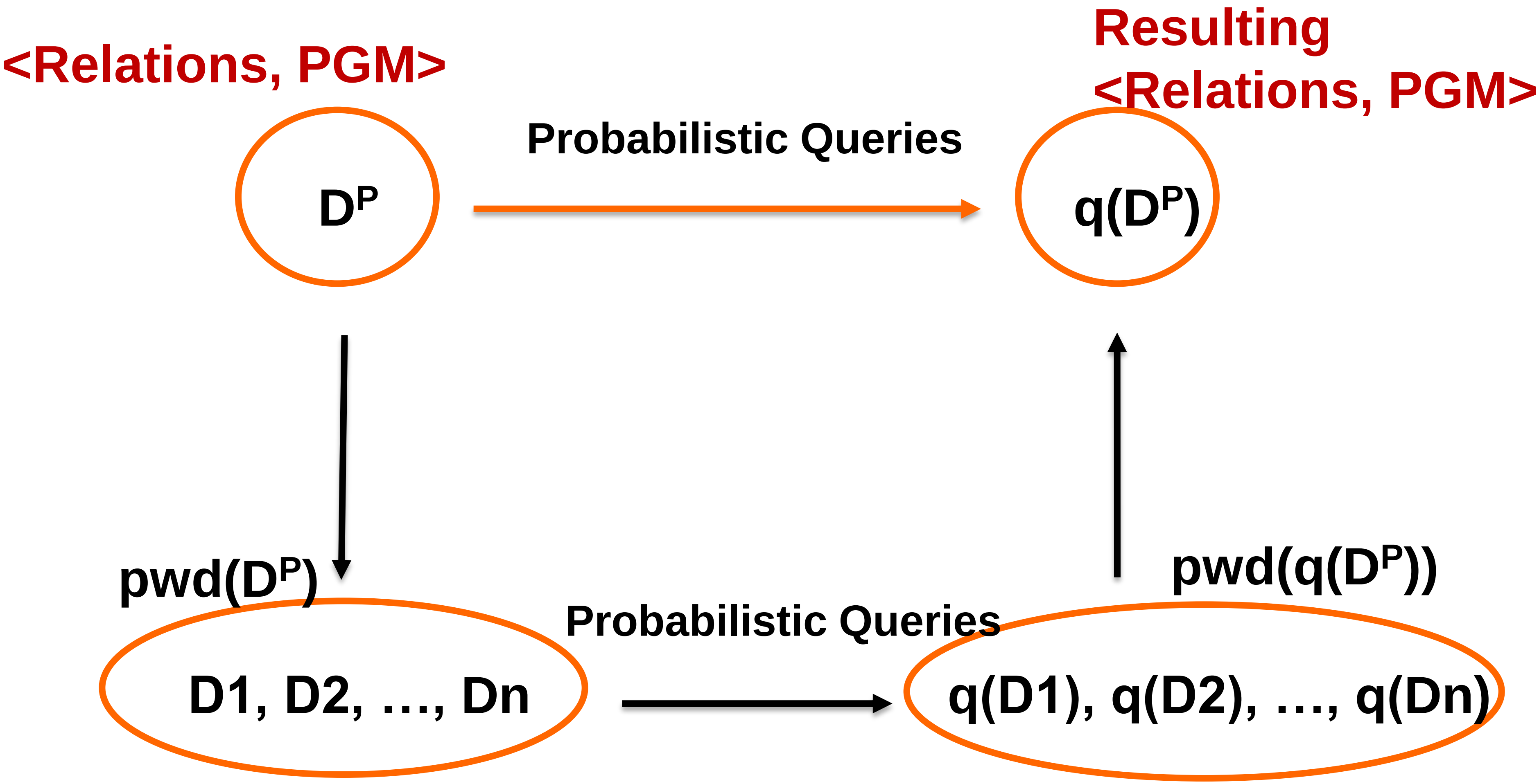
"Possible Worlds"

- (Evidence + Model) define a probability distribution over "possible worlds"
- *Complete* data model $Prob_{Model}(World \mid Evidence)$

Possible World Query Semantics



Possible World Query Semantics



HeisenData Relational Query Processing [VLDB08]

- Processing *general Select-Project-Join (SPJ) queries* directly over *<Incomplete Tables, FO Model>*

- Exploit SPJ query constraints to appropriately modify and/or shrink the model and uncertain data
  - Tools such as the Bayes Ball algorithm, Model-based filtering,…

- Did not really address probabilistic inference, other than simple optimizations
  - Exploit *FO Inference* in this setting…???

Talk Outline

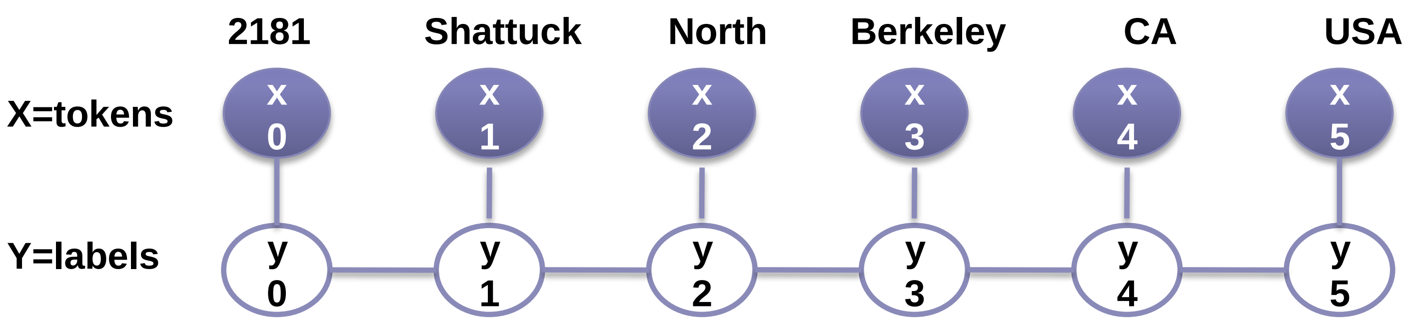
- Introduction, Motivation, Challenges
- Example Data Model and Relational Query Processing [VLDB08]
- Managing Inference for Information Extraction [ICDE10,VLDB10,SIGMOD11]
- Statistics for Probabilistic Data [ICDE09,VLDB09]
- Conclusions & Future Work

Conditional Random Fields (CRFs) for IE

Text (address string):

E.g., *"2181 Shattuck North Berkeley CA USA"*

CRF Model:



Possible Extraction Worlds:

x	2181	Shattuck	North	Berkeley	CA	USA	
y1	apt. num	street name	city	city	state	country	(0.6)
y2	apt. num	street name	street name	city	state	country	(0.1)
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

HeisenData Data Model for IE

2181 Shattuck North Berkeley CA USA

docID	pos	token	$Label^P$
1	0	2181	
1	1	Shattuck	
1	2	North	
1	3	Berkeley	
1	4	CA	
1	5	USA	

$TokenTable^P$

token	prevLabel	label	score
Shattuck	street num	street name	22
Shattuck	street num	street num	5
…	..	..	
Berkeley	street name	street name	10
Berkeley	street name	city	25
..	..	..	

FactorTable

Relational and Inference Queries

- Relational Operators
  - Select, Project, Join
  - Aggregation
- Inference Operators
  - Top-k Inference
  - Marginal Inference

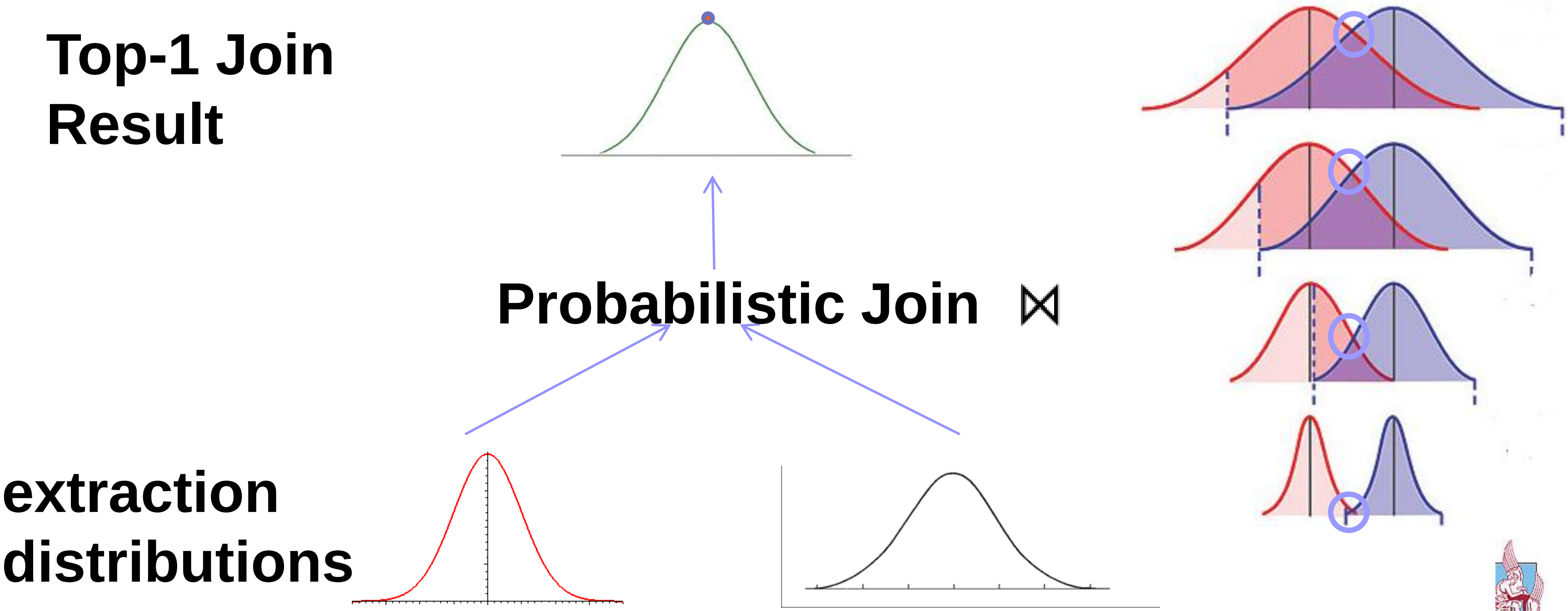
docID	pos	token	$Label^P$
1	0	2181	
1	1	Shattuck	
1	2	North	
1	3	Berkeley	
1	4	CA	
1	5	USA	

$TokenTable^P$

SELECT pos, token, top-k($Label^P$)
FROM *$TokenTable^P$*
WHERE docID <= 10

Top-k Probabilistic Join

SELECT	***Top-k join extractions***
FROM	***$Emails^P$ D1, $Emails^P$ D2***
WHERE	***D1.docID != D2.docID***
	and D1.$Label^P$ = D2.$Label^P$ = ‘company’
	and D1.token = D2.token and prob > T



Top-k Probabilistic Join

SELECT	***Top-k join extractions***
FROM	***Emails^P D1, Emails^P D2***
WHERE	***D1.docID != D2.docID***
	and D1.Label^P = D2.Label^P = 'company'
	and D1.token = D2.token and prob > T

Starting point: *Viterbi Dynamic Programing* (Top-1 Extraction)

- Incremental Viterbi → Ranked List of Extractions
- Probabilistic Rank-Join → Top-k Join Result

Viterbi DP for Max-Likelihood Extraction

Viterbi DP Algorithm:

$$V(i,y) = \begin{cases} \max_{y'} (V(i-1,y') + \sum_{k=1}^{K} \lambda_k \bullet f_k \bullet f(y,y',x_i)), & \text{if } i \geq 0 \\ 0, & \text{if } i = -1. \end{cases}$$

	pos	street num	street name	city	state	country
2181	0	5	1	0	1	1
Shattuck	1	2	15	7	8	7
North	2	12	24	21	18	17
Berkeley	3	21	32	32	30	26
CA	4	29	40	38	42	35
USA	5	39	47	46	46	50

Gave efficient in-database implementation (in SQL)! [ICDE10]

Incremental Viterbi [VLDB10]

- Novel variant of Viterbi-based CRF inference
- Input: States and Top-1 Extraction from Viterbi
- Algorithm: Incrementally computes the next highest probability extraction
  - Clever book-keeping and incremental evaluation
- Result: List of extractions ranked by probability

- Complexity: $O(T(|Y|+k)\log(|Y|+k)) < O(T|Y|^2)$

  when k is small, T (number of tokens),

  $|Y|$ (number of labels), k (extraction depth)

- [SIGMOD11] deals with alternative inference tools (e.g., MC sampling)

Talk Outline

- Introduction, Motivation, Challenges
- Example Data Model and Relational Query Processing [VLDB08]
- Managing Inference for Information Extraction [ICDE10,VLDB10,SIGMOD11]
- Statistics for Probabilistic Data [ICDE09,VLDB09]
- Conclusions & Future Work

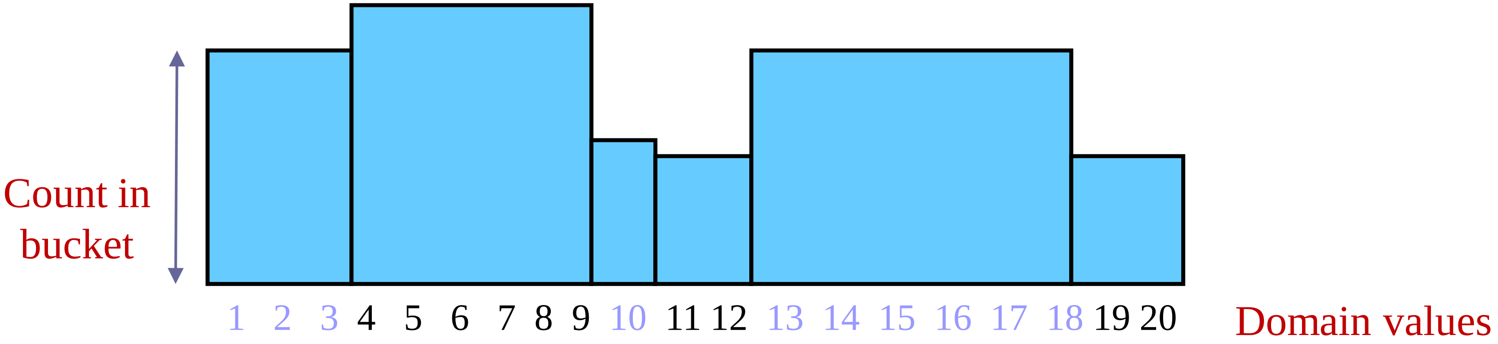
Probabilistic Data Reduction

- Probabilistic data can be difficult to work with
  - Even simple queries can be #P hard [Dalvi,Suciu'04]
    - joins and projections between (statistically) independent probabilistic relations
    - need to track the history of generated tuples
  - Want to avoid materializing all possible worlds

- *Our Goal:* Seek compact representations of probabilistic data
  - Data synopses which capture key properties and possible world semantics
  - Can perform expensive operations on compact summaries

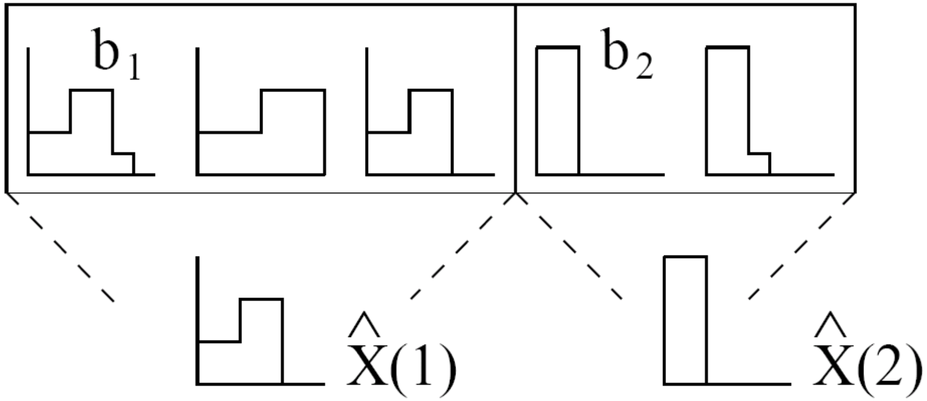
Traditional Histograms

- Compact, piecewise-constant approximations of large PDFs
  - Domain is split in B buckets
  - Each bucket approximated by a *single* value
    - Typically, the average probability mass / count in the bucket
  - Approximation using O(B) space



Probabilistic Histograms [VLDB09]

- A powerful approximate representation of uncertain data
- Represent each bucket with a PDF
  - Capture prob. of each item appearing i times



- Complete representation
- Target several metrics
  - EMD, Kullback-Leibler divergence, Hellinger Distance
  - Max Error, Variation Distance (L1), Sum Squared Error etc

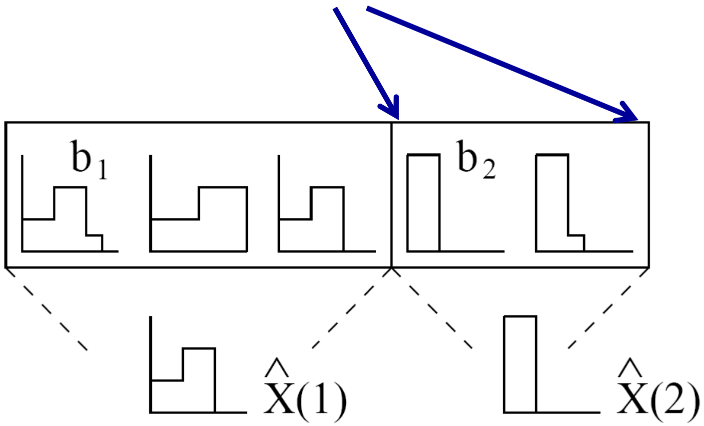
Probabilistic Data Model

- Ordered universe $\mathcal{U}$ of data items (i.e., {1, 2, …, N})
- Each item in $\mathcal{U}$ obtains values from a value domain $\mathcal{V}$
  - Each with different probability $\Rightarrow$ each item described by PDF
- Example:
  - PDF of item i describes prob. that i appears 0, 1, 2, … times
  - PDF of item i describes prob. that i measured value V_1, V_2 etc
- Can capture popular "independent tuple" models (Trio, Mystiq, …)
  - Handling correlations is an open problem…

Bucket Representation

- Goal: Partition universe $\mathcal{U}$ into buckets
- Within each bucket b = (s,e)
  - Approximate (e-s+1) pdfs with a piecewise constant PDF $\hat{X}(b)$

Start: s
End: e
of bucket



- Error of above approximation
  - Let d() denote a distance function of PDFs

$$Err(b) = \bigoplus_{i=s}^{e} d(\hat{X}(b), X_i)$$

Typically, summation or MAX

- Given a space bound (no. of piecewise constant terms), we need to determine
  - number of buckets
  - terms (i.e., pdf complexity) in each bucket

Target Error Metrics

Variation Distance (L1)	$d(X,Y) = \|X-Y\|_1 = \sum_{v\in\mathcal{V}} \lvert \Pr[X=v]-\Pr[Y=v] \rvert$
Sum Squared Error	$d(X,Y) = \|X-Y\|_2^2 = \sum_{v\in\mathcal{V}} (\Pr[X=v]-\Pr[Y=v])^2$
Max Error (L∞)	$d(X,Y) = \|X,Y\|_\infty = \max_{v\in\mathcal{V}} \lvert \Pr[X=v]-\Pr[Y=v] \rvert$
(Squared) Hellinger Distance	$d(X,Y) = H^2(X,Y) = \sum_{v\in\mathcal{V}} \frac{(\Pr[X=v]^{\frac{1}{2}} - \Pr[Y=v]^{\frac{1}{2}})^2}{2}$
Kullback-Leibler Divergence (relative entropy)	$d(X,Y) = KL(X,Y) = \sum_{v\in\mathcal{V}} \Pr[X=v]\log_2 \frac{\Pr[X=v]}{\Pr[Y=v]}$
Earth Mover's Distance (EMD)	Distance between probabilities at the value domain

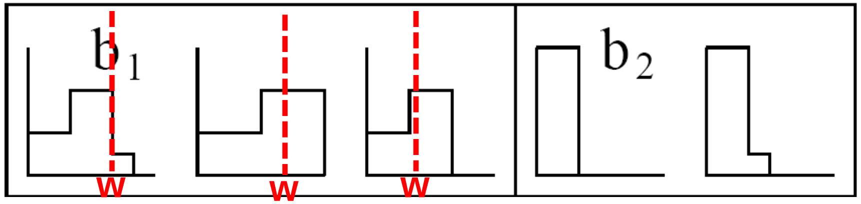
Common Prob. metrics

General DP Scheme: Inter-Bucket

- Let B-OPT^b[w,T] represent error of approximating up to $w \in \mathcal{V}$ first values of bucket b using T terms

Error approximating first w values of PDFS within bucket b

Using T terms for bucket b



- Let H-OPT[m, T] represent error of first m items in $\mathcal{U}$ when using T terms

$$\text{H-OPT}[m,T] = \min_{1\le k\le m-1, 1\le t\le T-1}\{\text{H-OPT}[k,T-t] + \text{B-OPT}^{(k+1,m)}[V+1,t]\}$$

Check all start positions of last bucket, terms to assign

Use T-t terms for the first k items

Where the last bucket starts

Approximate all V+1 frequency values using t terms

General DP Scheme: Intra-Bucket

- Compute efficiently per metric
- Utilize pre-computations

- Each bucket b=(s,e) summarizes PDFs of items s,…,e
  - Using from 1 to V=|$\mathcal{V}$| terms
- Let VALERR(b,u,v) denote minimum possible error of *1-term* approximating the frequency values in [u,v] of bucket b. Then:

$$B-OPT^{b}[w,T]=\min_{1\le u\le w-1}\{B-OPT^{b}[u,T-1]+VALERR(b,u+1,w)\}$$

Use T-1 terms for the first u frequency values of bucket

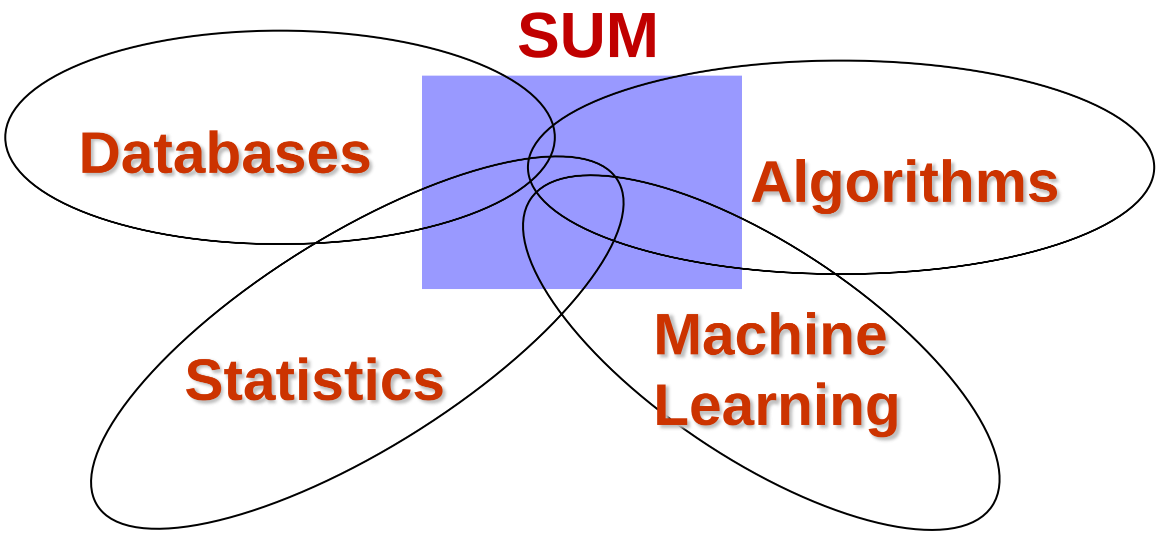
Where the last term starts

Efficient Probabilistic Histograms [VLDB09]

- Several optimizations for efficient DP computation under different error metrics
  - Efficient computation of VALERR() exploiting pre-computation
- Show how “possible worlds” queries can be handled using probabilistic histograms
- Have recently given more efficient approximate versions of the DP
  - Guaranteed *ε-approximate* probabilistic histograms

Conclusions

- *HeisenData* = Scalable data-management & probabilistic reasoning system
  - Integrate state-of-the-art DB and ML technology
  - Issues in data model, query processing, managing inference and IE, database statistics
  - Many-many more remain open…



<u>Very exciting field of research!!</u>

Future Work

- Many directions we currently pursue / want to pursue
  - Integrating the *Entity Resolution* step for IE
  - Exploiting modern cloud platforms and parallelism for inference
  - Managing and querying data lineage
  - Integrating FO inference techniques and ideas
  - Statistics in the presence of models and correlations
  - … and, designing physical algebras, costing operators, query optimization, ….

Thank you!

http://heisendata.softnet.tuc.gr/

http://www.softnet.tuc.gr/~minos/

minos@softnet.tuc.gr

Sum Squared Error & (Squared) Hellinger Distance

- Simpler cases (solved similarly). Assume bucket b=(s,e) and wanting to compute VALERR(b,v,w)
- (Squared) Hellinger Distance (SSE is similar)
  - Represent bucket [s,e]x[v,w] by single value p, where

$$p = \bar{p} = \left(\frac{\sum_{i=s}^{e} \sum_{j=v}^{w} \sqrt{\Pr[X_i = j]}}{(e-s+1)(w-v+1)} \right)^2$$

  - VALERR(b,v,w) = $\sum_{i=s}^{e} \sum_{j=v}^{w} \Pr[X_i = j] - (e-s+1)(w-v+1)\bar{p}$

Computed by 4 B[] entries

Computed by 4 A[] entries

  - VALERR computed in constant time using O(UV) pre-computed values, given

$$A[e,w] = \sum_{i=1}^{e} \sum_{j=1}^{w} \sqrt{\Pr[X_i = j]} \qquad B[e,w] = \sum_{i=1}^{e} \sum_{j=1}^{w} \Pr[X_i = j]$$

Variation Distance

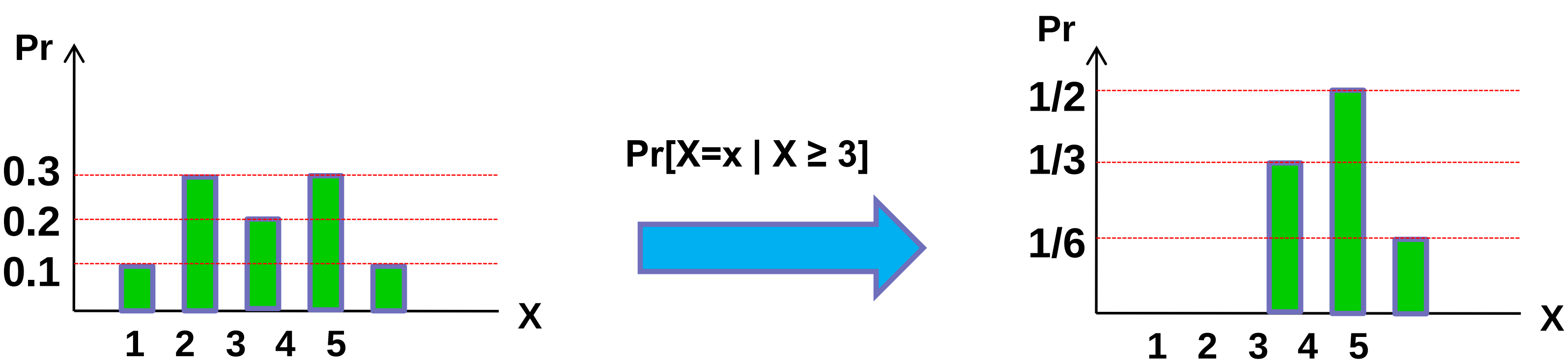
- Interesting case, several variations
- Best representative within a bucket = median P value
- $$\text{VALERR}(b,v,w) = \sum_{i=s}^{e}\sum_{j=v}^{w} \Pr[X_i = j] - 2I(i,j)\Pr[X_i = j]$$
- , $I(i,j)$ is 1 if $\Pr[X_i = j] \leq p_{med}$, and 0 otherwise
- Need to calculate sum of values below median $\Rightarrow$ two-dimensional range-sum median problem
- Optimal PDF generated is NOT normalized
- Normalized PDF produced by scaling = factor of 2 from optimal
- Extensions for ε-error (normalized) approximation

Other Distance Metrics

- Max-Error can be minimized efficiently using sophisticated pre-computations
  - No Intra-Bucket DP needed
  - Complexity lower than all other metrics: $O(TVN^2)$
- EMD case is more difficult (and costly) to handle
- Details in the paper…

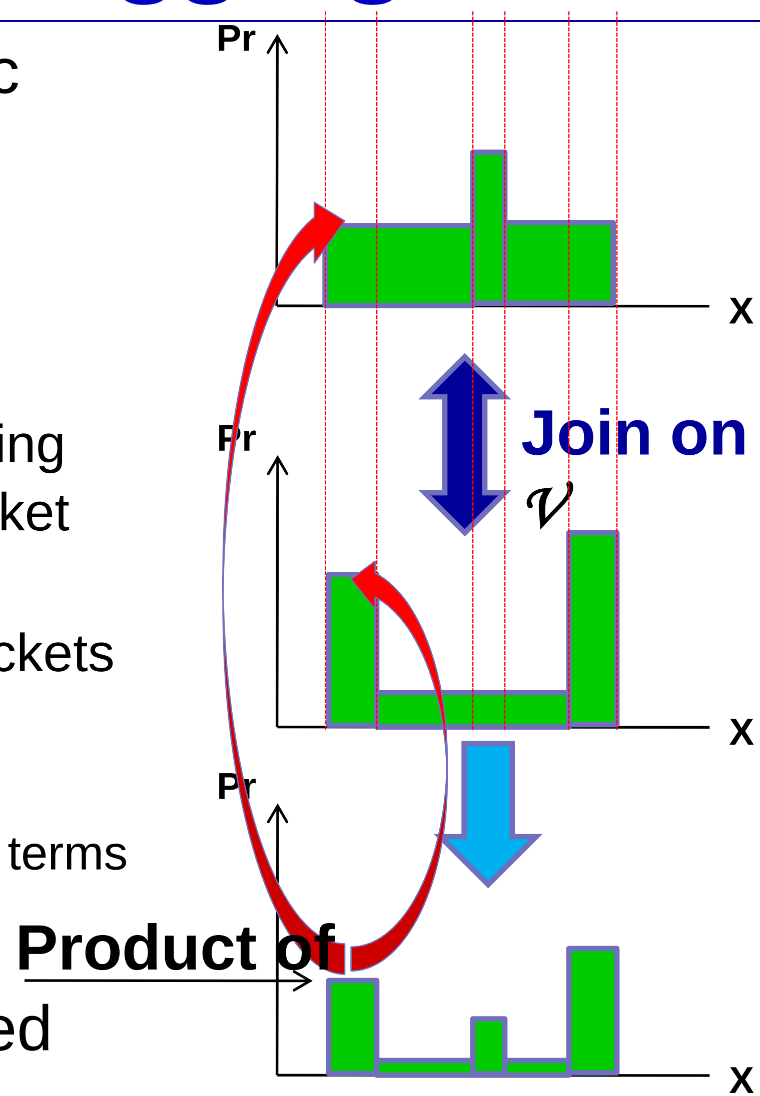
Handling Selections and Joins

- Simple statistics such as expectation are simple
- Selections on item domain are straightforward
  - Discard irrelevant buckets - Result is itself a prob. histogram
- Selections on the value domain are more challenging
  - Correspond to extracting the distribution conditioned on selection criteria
- Range predicates are clean: result is a probabilistic histogram of approximately same size



Handling Joins and Aggregates

- Result of joining two probabilistic relations can be represented by joining their histograms
  - Assume pdfs of each relation are independent
  - Ex: equijoin on $\mathcal{V}$: Form join by taking product of pdfs for each pair of bucket intersections
  - If input histograms have B1, B2 buckets respectively, the result has at most B1+B2-1 buckets
    - Each bucket has at most: T1+T2-1 terms



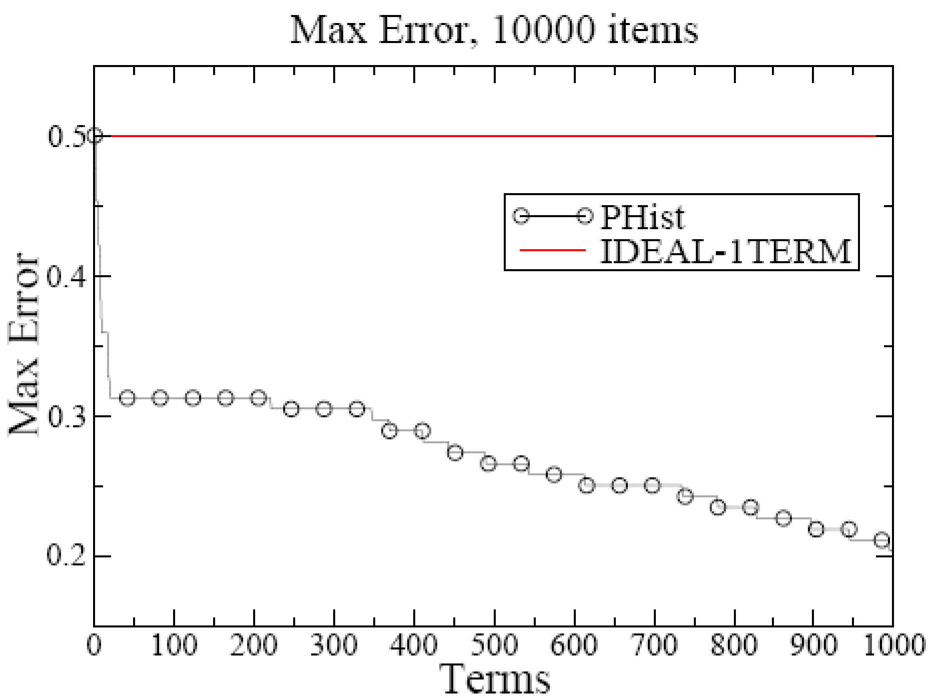
- Aggregate queries also supported
  - I.e., count(#tuples) in result
  - Details in the paper…

Experimental Study

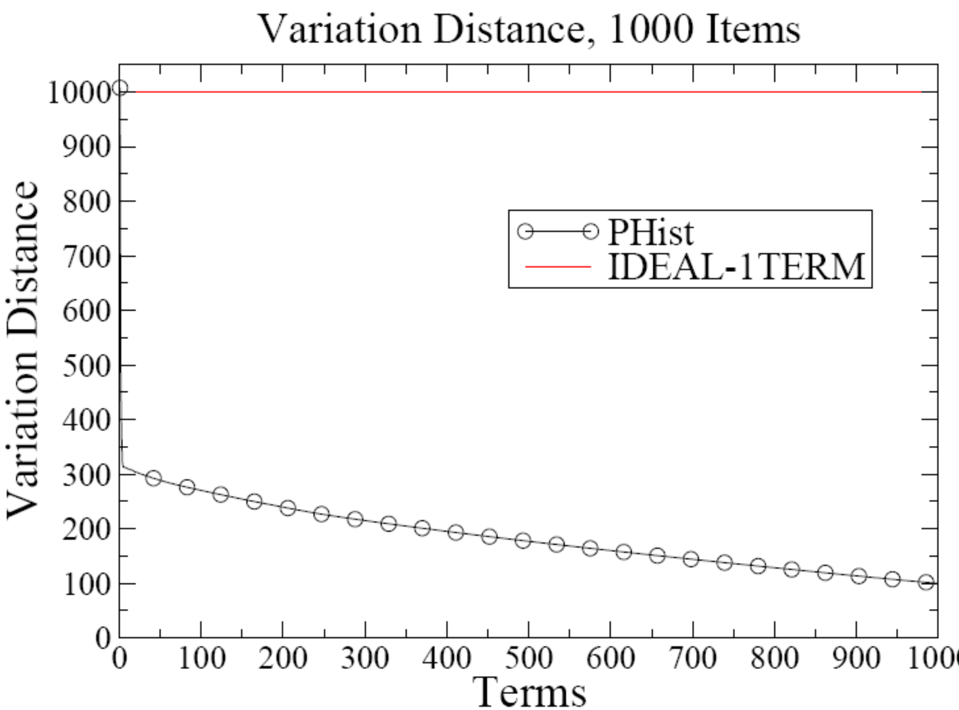
- Evaluated on two probabilistic data sets
  - Real data from Mystiq Project (127k tuples, 27,700 items)
  - Synthetic data from MayBMS generator (30K items)
- Competitive technique considered: IDEAL-1TERM
  - One bucket per EACH item (i.e., no space bound)
  - A single term per bucket

- Investigated:
  - Scalability of PHist for each metric
  - Error compared to IDEAL-1TERM

Quality of Probabilistic Histograms



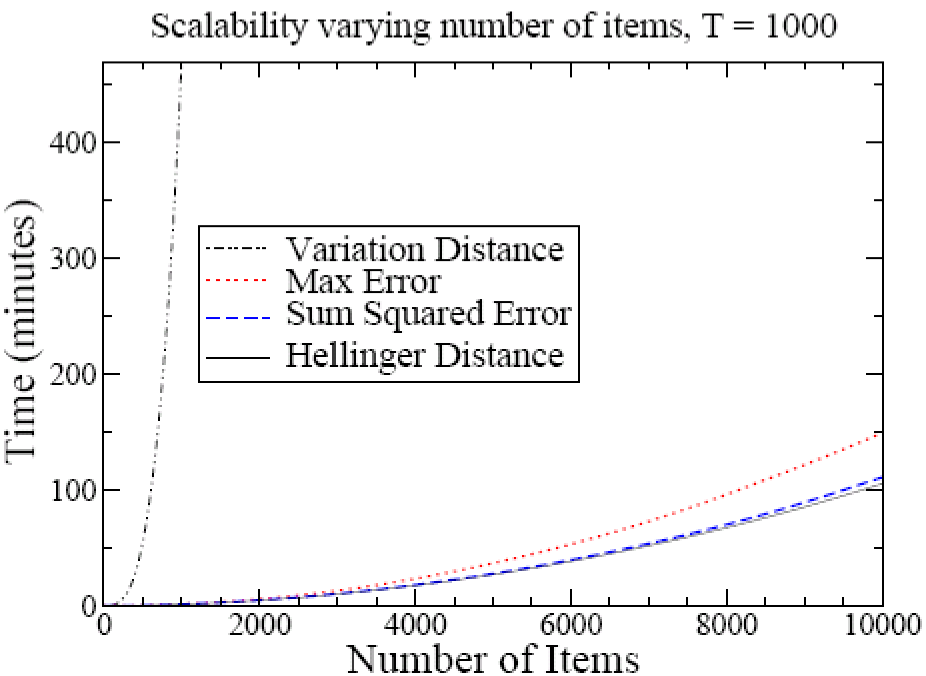
(b) Max-Error statistic



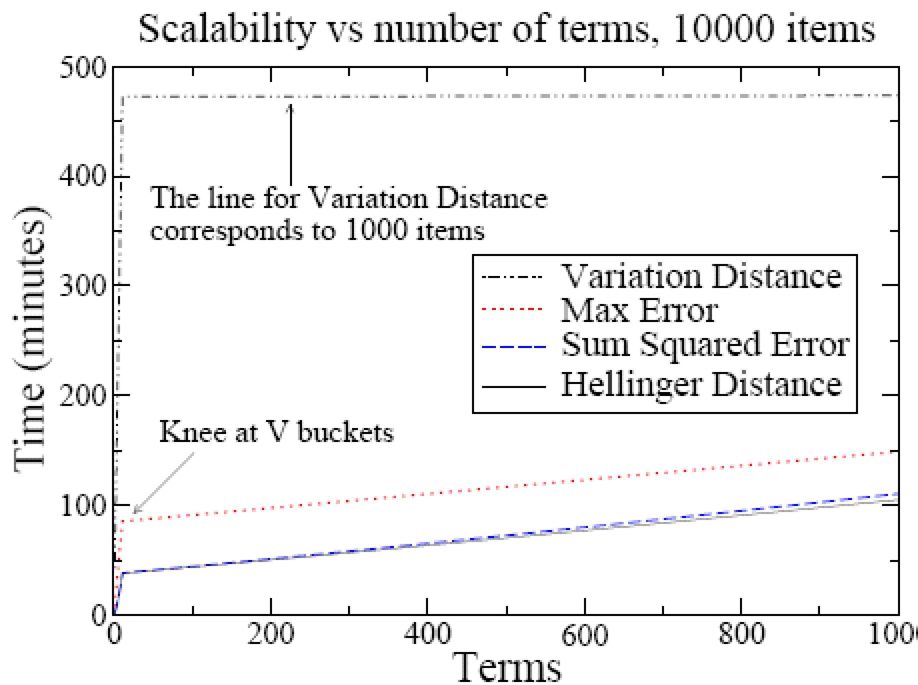
(d) Sum Variation Distance

- Clear benefit when compared to IDEAL-1TERM
  - PHist able to approximate full distribution

Scalability



(a) Time as the number of items N varies



(b) Time as T varies

- Time cost is linear in T, quadratic in N
  - Variation Distance (almost cubic complexity in N) scales poorly
- Observe “knee” in right figure. Cost of buckets with > V terms is same as with EXACTLY V terms => INNER DP uses already computed costs

Concluding Remarks

- Presented techniques for building probabilistic histograms over probabilistic data
  - Capture full distribution of data items, not just expectations
  - Support several minimization metrics
  - Resulting histograms can handle selection, join, aggregation queries
- Future Work
  - Current model assumes independence of items. How to deal with item correlations…?
  - Running time improvements
    - (1+ε)-approximate solutions [Guha, Koudas, Shim: ACM TODS 2006]
    - Prune search space (i.e., very large buckets) using lower bounds for bucket costs

Probabilistic Data Analysis

Information Extraction Systems

Sensor Networks

Extracted entities (e.g. names, locations) are probabilistic

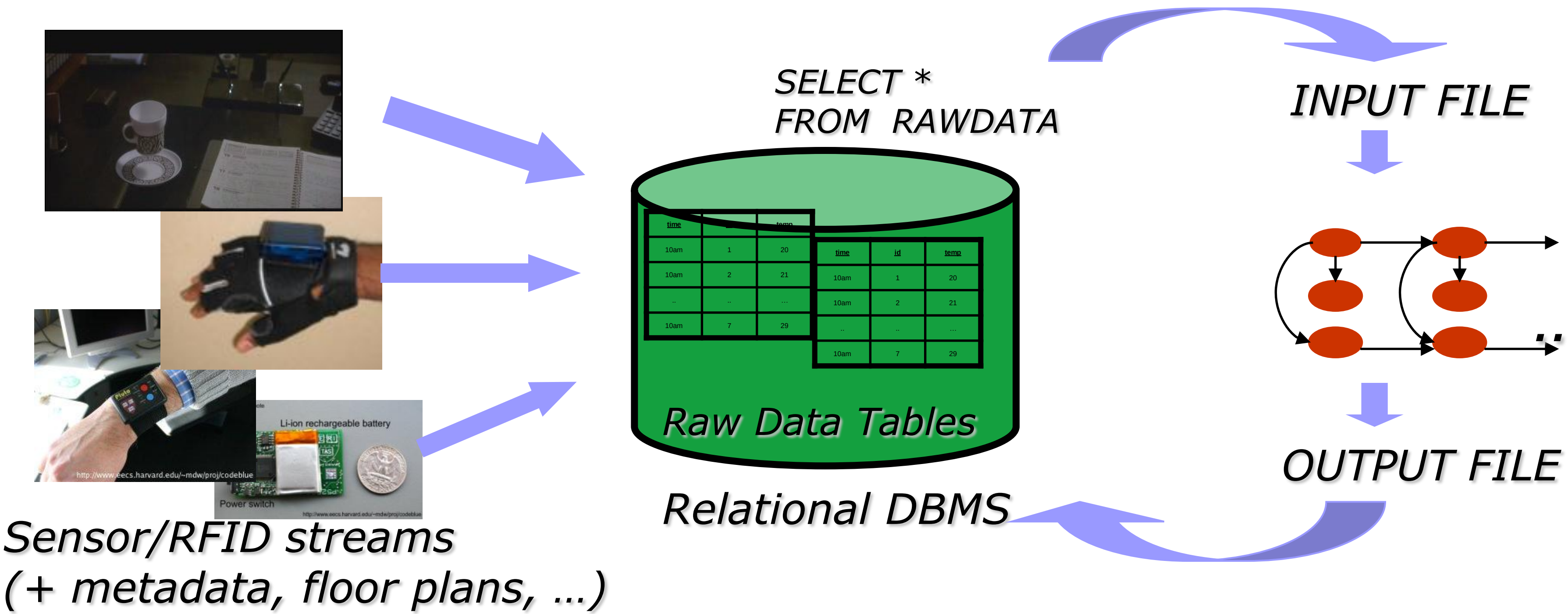
Sensor readings (e.g. light, temperature) are probabilistic

Which NYTimes articles mention ‘Apple’ as a company with top-k highest probability?

What’s the Gaussian distribution of average temperature of the area?

Other Ongoing/Future Work: Probabilistic Data Management

- Managing *uncertain data*



- All interesting data processing done *outside* the database!
- Lose *all key benefits of a DBMS* (declarative querying, persistence, optimization, …)
- No sharing of data/knowledge/abstractions, duplication of effort

Probabilistic Data Management

- *Existing Probabilistic DBs:* Simplistic uncertainty models that easily map to existing DB architectures
  - Independent tuple-level confidences and attribute-value options (OR-tuples)

Year	Value	Confidence
1952	55^o F	0.7
1954	-22^o F	0.9
. . .	. . .	. . .

Owns (owner,car)
(Jimmy, Toyota) || (Jimmy, Mazda)
(Billy, Honda) || (Frank, Honda)
(Hank, Honda)

- *The **HeisenData** Project (originally UC Berkeley, now at TUC)*
  - *Scalable, integrated data-management & probabilistic-reasoning platform*
  - Statistical models and reasoning as "first-class" citizens in the DBMS
  - Query processing = relational ops + statistical inference
    - ***"Possible worlds"*** semantics (data + stat model)

HeisenData Challenges

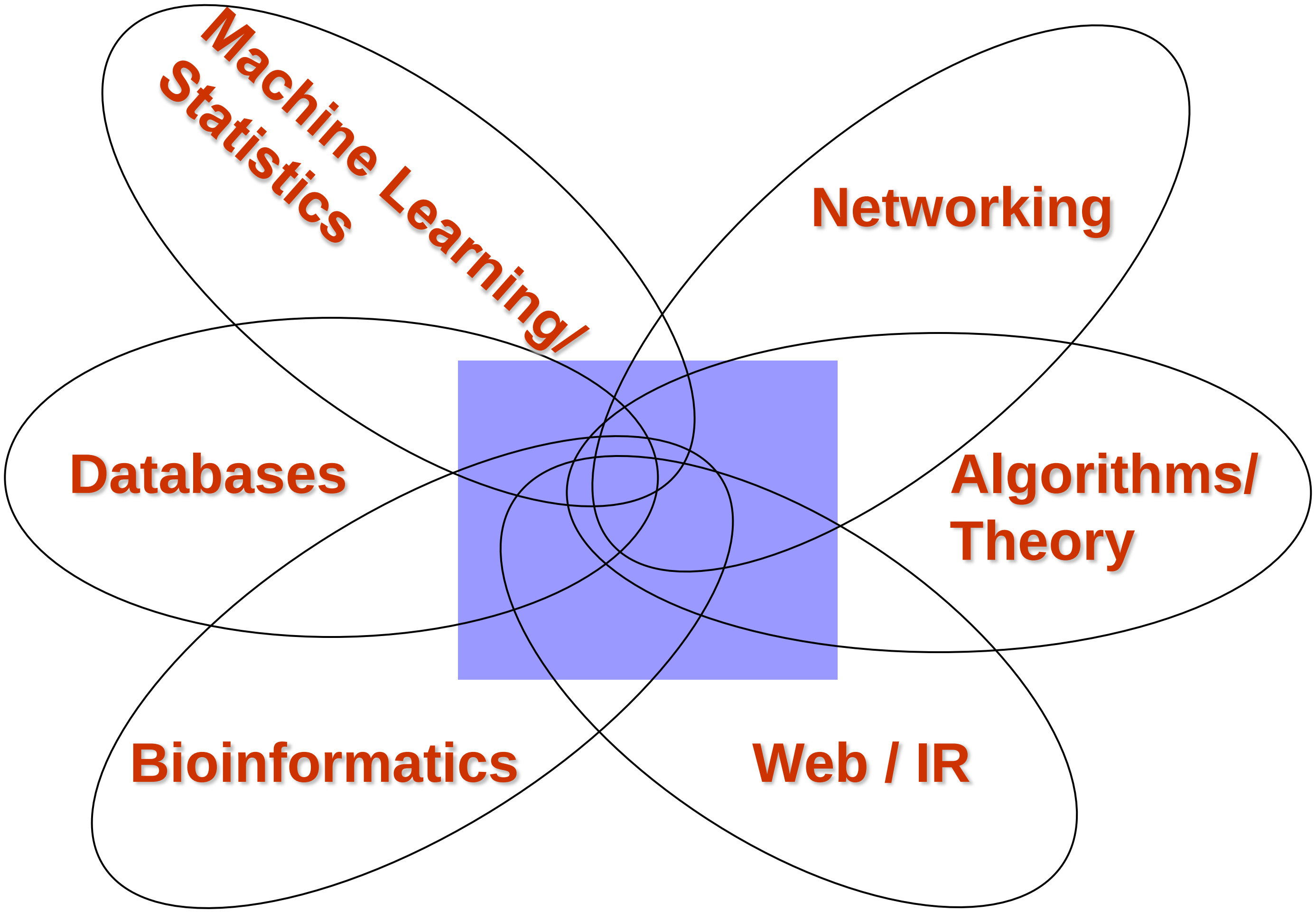
- What is the right language/algebra/interface?
  - Completeness, soundness
  - Expressiveness & ease of use

- Query Processing & Optimization
  - Probabilistic queries with relational and inference operators! *[MG+, VLDB'08]*
  - Inference is *expensive!*
    - Exploit massive parallelism (e.g., Hadoop) and/or approximation?
  - Statistics for probabilistic data? *[Cormode,MG,SIGMOD'07]*
  - Physical DB design (indexes, access structs, views, …)?
  - Extensibility (stat models, inference techniques, …)

App: Managing Information Extraction

- IE = Extracting structured entities from unstructured text
  - Based on sophisticated ML models and tools (e.g., CRFs)
  - *Lots of data:* many data sources, background/domain knowledge, extracted data (inferences), …
  - Results riddled with uncertainty
- Difficult challenges for Probabilistic DBMS
  - *Declarative IE:* Extraction as PDB query processing!
    - IE op algebra, optimizing IE query plans, statistics for IE, …
  - *Managing IE state*
    - Probabilistic query answering over extracted data
    - Maintaining/querying provenance of inferences (“explain”)
    - *Continuous* extraction (i.e., monitoring)
- Some initial steps in *[MG+, ICDE’10, Unpub’10]*

My View of Modern Data Management



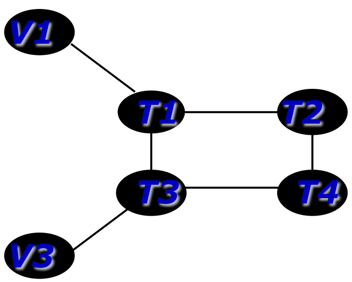
Really exciting times for Data-Management Research!!

BayesStore Model

time	id	temp	volt
10am	1	20	2.5
10am	2	21	XXX
..	..	…	
10am	7		2.8

Evidence Table(s)

+



Hierarchical FO Graphical Model

Prob=0.4

time	id	temp	volt
10am	1	20	2.5
10am	2	21	2.7
..	..	…	
10am	7	26	2.8

Prob=0.3

time	id	temp	volt
10am	1	20	2.5
10am	2	21	2.7
..	..	…	
10am	7	28	2.8

Prob=0.3

time	id	temp	volt
10am	1	20	2.5
10am	2	21	2.7
..	..	…	
10am	7	26	2.8

"Possible Worlds"

- (Evidence + Model) define a probability distribution over "possible worlds"
- *Complete* data model $Prob_{Model}(World \mid Evidence)$

BayesStore [MG+,VLDB'08]

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

$$F = Pr\,[X_1, \ldots, X_7\,]$$

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

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

The Skyscrapers Example

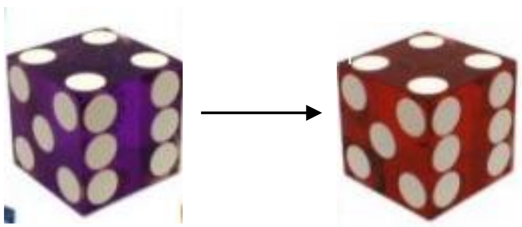
For all sensor in all rooms at all timestamp, Light and Temperature readings are correlated.

Light

Temperature

Definitions

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



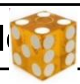
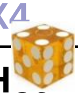
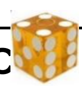
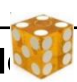
First-order Factor: *A family of local models, which share the same structure and conditional probability table (CPT).*

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

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	X4	Brt
t5	1	2	2	H 	X5
t6	1	2	3	X6	X7
t7	2	1	1	Hot	X8
t8	2	1	2	C 	Drk
t9	2	1	3	X9	X10
t10	2	2	1	X11	Brt
t11	2	2	2	H 	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

Sensor1^p

	T	R	Sid	Tp^p	L^p
t1	1	1	1	H	X1
t2	1	1	2	C l	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 l	Dr
t9	2	1	3	X	X1
t10	2	2	1	X	Br
t11	2	2	2	H	X1
t12	2	2	3	X	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

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

$\langle R^p, F_{FOBN} \rangle$

(I)

Relational and Inference Queries

Resulting $\langle R^p, F_{FOBN} \rangle$

(II)

Represent

Represent

(III)

Possible Worlds And Distribution

Relational and Inference Queries

(IV)

Resulting Possible Worlds And Distribution