Mining Uncertain and Probabilistic Data

Problems, Challenges, Methods, and Applications

Jian Pei[1], Ming Hua[1], Yufei Tao[2], Xuemin Lin[3]

[1]Simon Fraser University

[2]The Chinese University of Hong Kong

[3]The University of New South Wales

Some figures in the slides are borrowed from some papers in the references

Outline

- Uncertainty and uncertain data, where and why?
- Models for uncertain and probabilistic data
- (coffee break)
- OLAP on uncertain and probabilistic data
- Mining uncertain and probabilistic data
- Tools: querying uncertain and probabilistic data
  - Indexing uncertain and probabilistic data
  - Ranking queries and spatial queries
- Summary and discussion

Uncertainty Is (Almost) Everywhere

- Uncertainty is often caused by our limited perception and understanding of reality
  - Limited observation equipment
  - Limited resource to collect, store, transform, analyze, and understand data
- Uncertainty can be inherent in nature
  - How much do you like/dislike McCain and Obama?

Data Collection Using Sensors

- Sensors are often used to collect data
  - Thermal, electromagnetic, mechanical, chemical, optical radiation, acoustic, …
  - Applications: environment surveillance, security, manufacture systems, …
- Ideal sensors
  - Ideal sensors are designed to be linear: the output signal of a sensor is linearly proportional to the value of the measured property
  - Sensitivity: the ratio between output signal and measured property

Measurement Errors – Certain

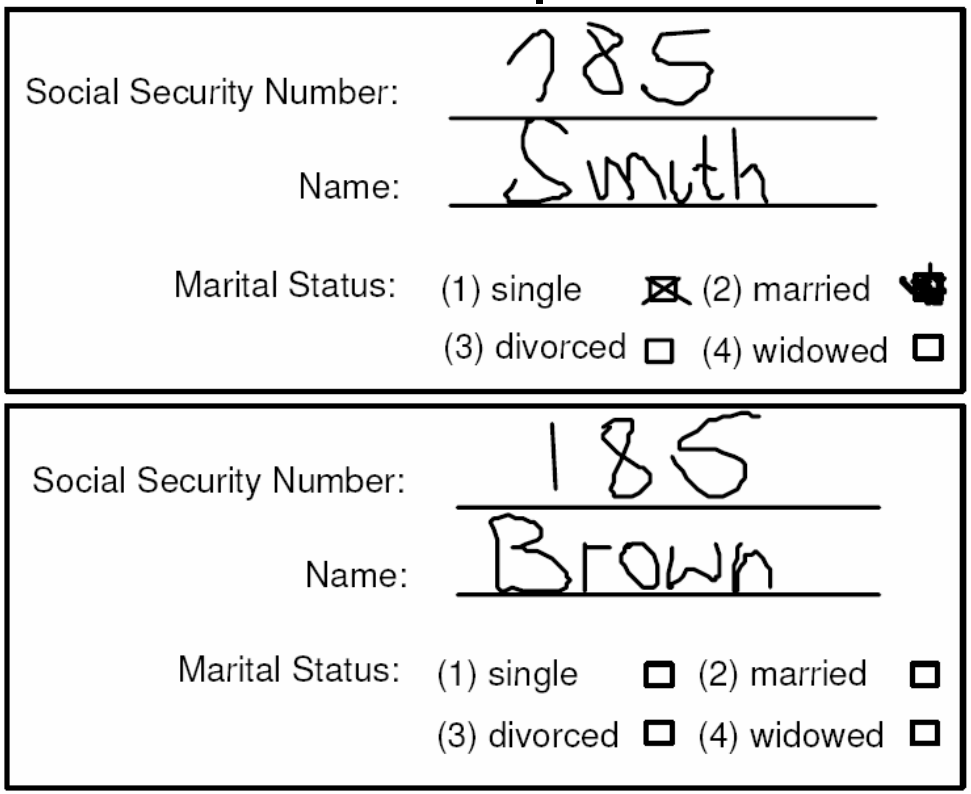
- Sensitivity error: the sensitivity differs from the value specified
- Offset (bias): the output of a sensor at zero input
- Nonlinearity: the sensitivity is not constant over the range of the sensor

Uncertain (Dynamic) Errors

- Dynamic error: deviation caused by a rapid change of the measured property over time
- Drift: the output signal changes slowly independent of the measured property
  - Long term drift: a slow degradation of sensor properties over a long period
- Noise: random deviation of the signal varying in time
- A sensor may to some extent be sensitive to properties (e.g., temperature) other than the one being measured
- Dynamic error due to sampling frequency of digital sensors

Uncertainty in Survey Data

- Social security number: 185 or 785
  - Exclusiveness: SSN should be unique
- Is Smith married?
  - Single or married, but not both



Antova et al. ICDE'07

Uncertainty due to Data Granularity

- Which state is p9 in?
- What is the total repair cost for F150's in the East?

	Auto	*Loc*	*Repair*	*Text*	*Brake*
p1	F-150	NY	\$200	...	$\langle 0.8, 0.2\rangle$
p2	F-150	MA	\$250	...	$\langle 0.9, 0.1\rangle$
p3	F-150	CA	\$150	...	$\langle 0.7, 0.3\rangle$
p4	Sierra	TX	\$300	...	$\langle 0.3, 0.7\rangle$
p5	Camry	TX	\$325	...	$\langle 0.7, 0.3\rangle$
p6	Camry	TX	\$175	...	$\langle 0.5, 0.5\rangle$
p7	Civic	TX	\$225	...	$\langle 0.3, 0.7\rangle$
p8	Civic	TX	\$120	...	$\langle 0.2, 0.8\rangle$
p9	F150	East	\$140	...	$\langle 0.5, 0.5\rangle$
p10	Truck	TX	\$500	...	$\langle 0.9, 0.1\rangle$

Burdick et al. VLDB'05

Uncertainty in Data Integration

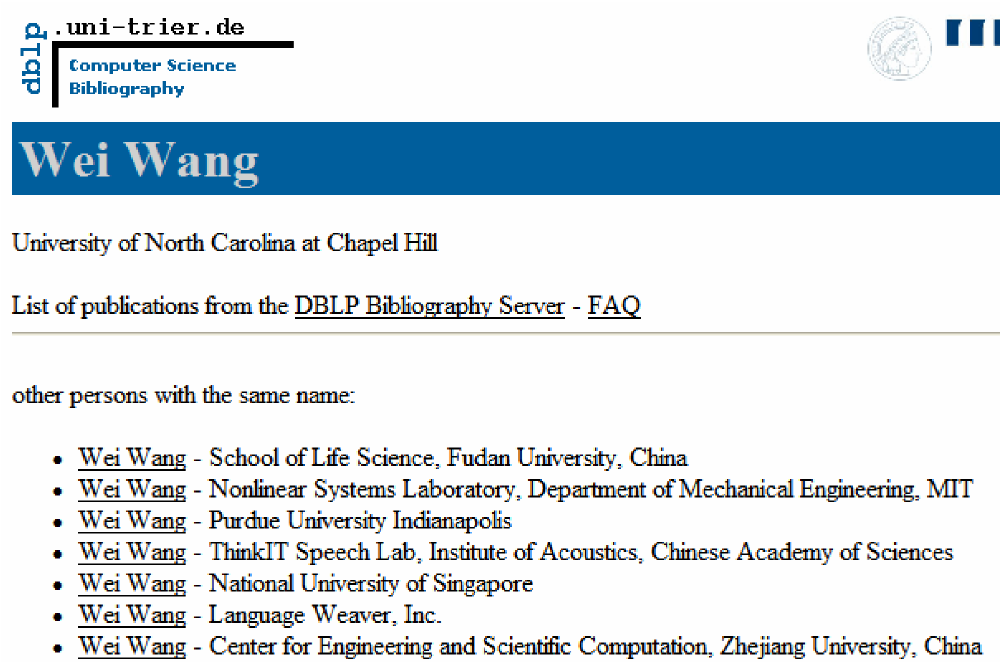
- Schema 1: (pname, email-addr, permanent-addr, current-addr)
- Schema 2: (name, email, mailing-addr, home-addr, office-addr)
- How to map the two schemas?

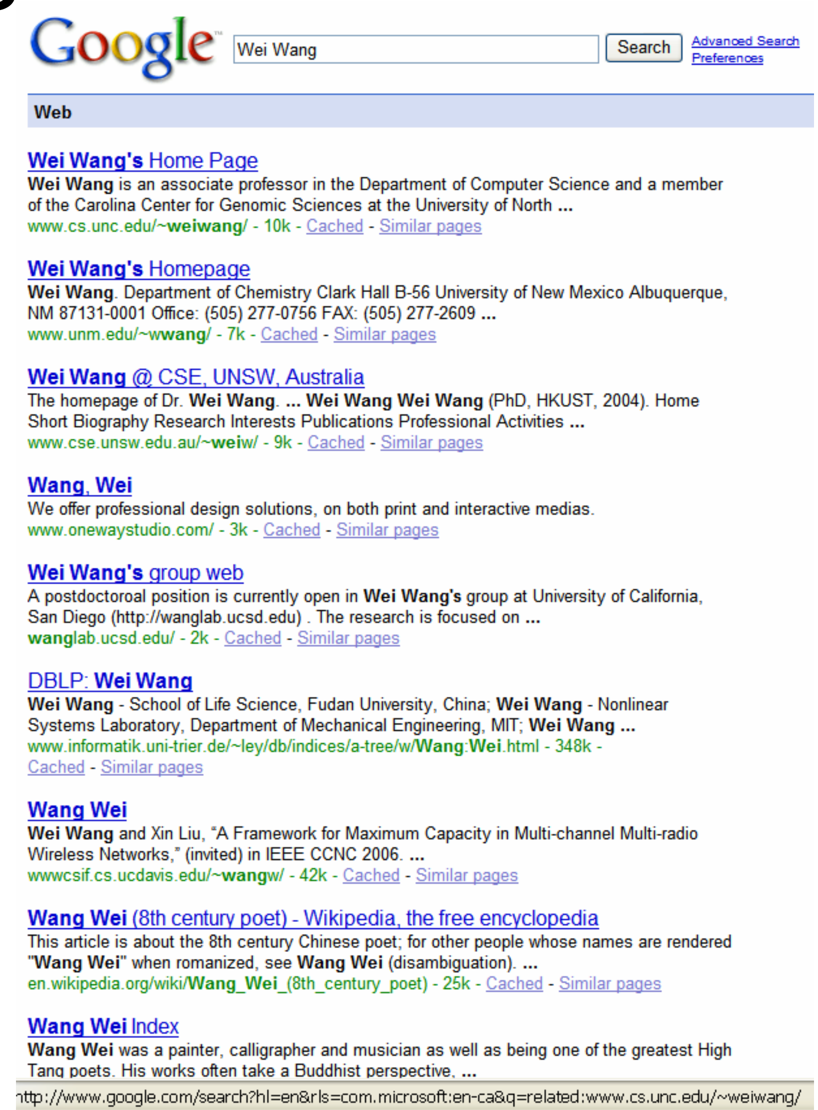
	Possible Mapping	Prob
$m_1 =$	{(pname, name), (email-addr, email), (current-addr, mailing-addr), (permanent-addr, home-address)}	0.5
$m_2 =$	{(pname, name), (email-addr, email), (permanent-addr, mailing-addr), (current-addr, home-address)}	0.4
$m_3 =$	{(pname, name), (email-addr, mailing-addr), (current-addr, home-addr)}	0.1

Dong et al. VLDB'07

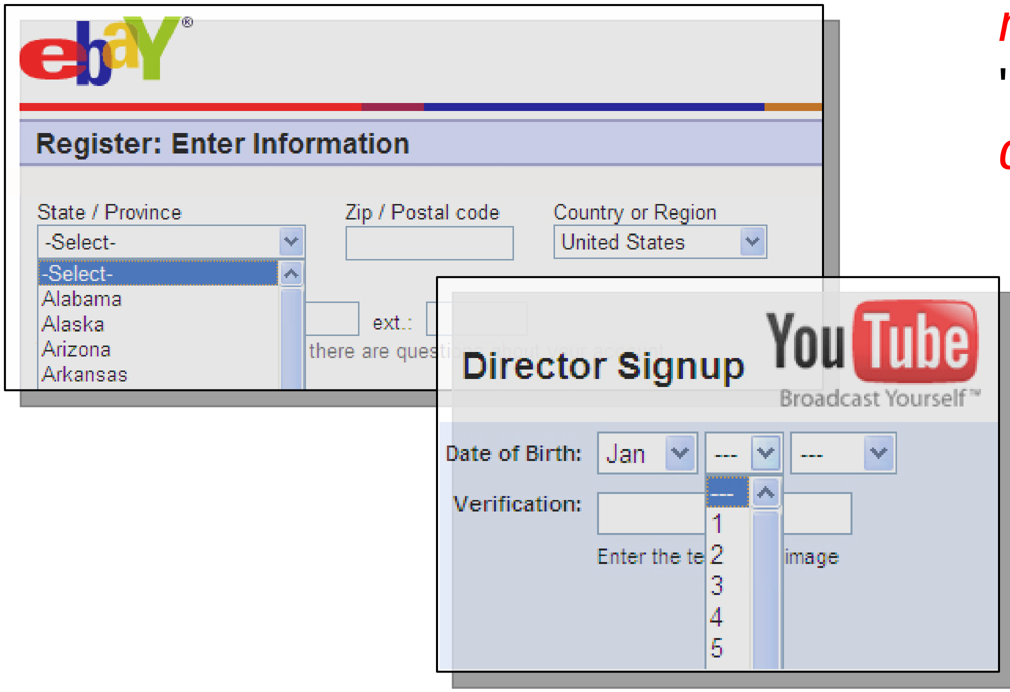
Ambiguous Entities

- Entity identification is a challenging task





Disguised Missing Data



Information about "State" is *missing*

"Alabama" is used as the *disguise*

Hua and Pei KDD'07

Disguised Missing Data

- Disguised missing data is the missing data entries that are not explicitly represented as such, but instead appear as potentially valid data values
- Disguised missing data also introduces uncertainty

Why Uncertain Data Is Still Useful?

- For a temperature sensor, suppose the difference between the real temperature and the sensed temperature follows normal distribution
- The real temperature can be modeled by a probability density function
- What is the real temperature? Uncertain
- What is the probability that the real temperature is over 50C? Certain!

Uncertainty and Confidence

- Uncertain data can provide probabilistic answers to aggregate questions
  - How can we estimate the percentage of married voters supporting Obama from survey data?
  - What is the total repair cost for F150's in the East?
- An answer derived from uncertain data may often be a function on probability or confidence

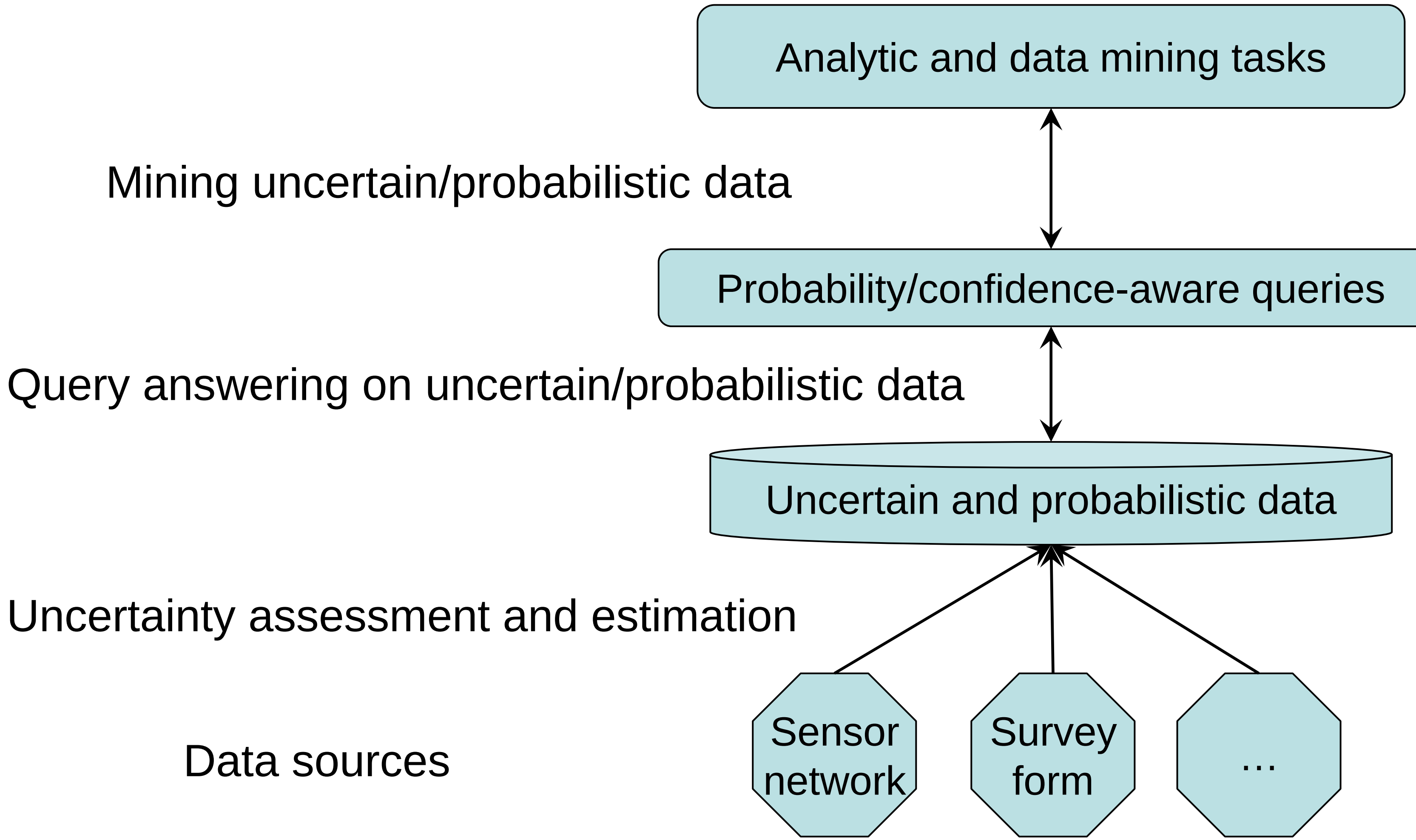
Reducing Uncertainty

- Removing uncertain entries
  - Removing uncertain attribute values
  - Removing uncertain records
  - Cons: reducing available data
- Generalization
  - Remove attribute city if some entries on the attribute is uncertain
  - Can accurately answer questions at level city or above
  - Still cannot answer questions at level city or below

Being Certain or Uncertain?

- Answering questions on uncertain data in general can be more complicated
  - Probability is a new (and often difficult) dimension
- Simplifying uncertain data to certain data may not use the full potential of data
  - Many details may be lost
- Probabilistic answers on uncertain data are often interesting and useful

Uncertain Data Analysis Framework

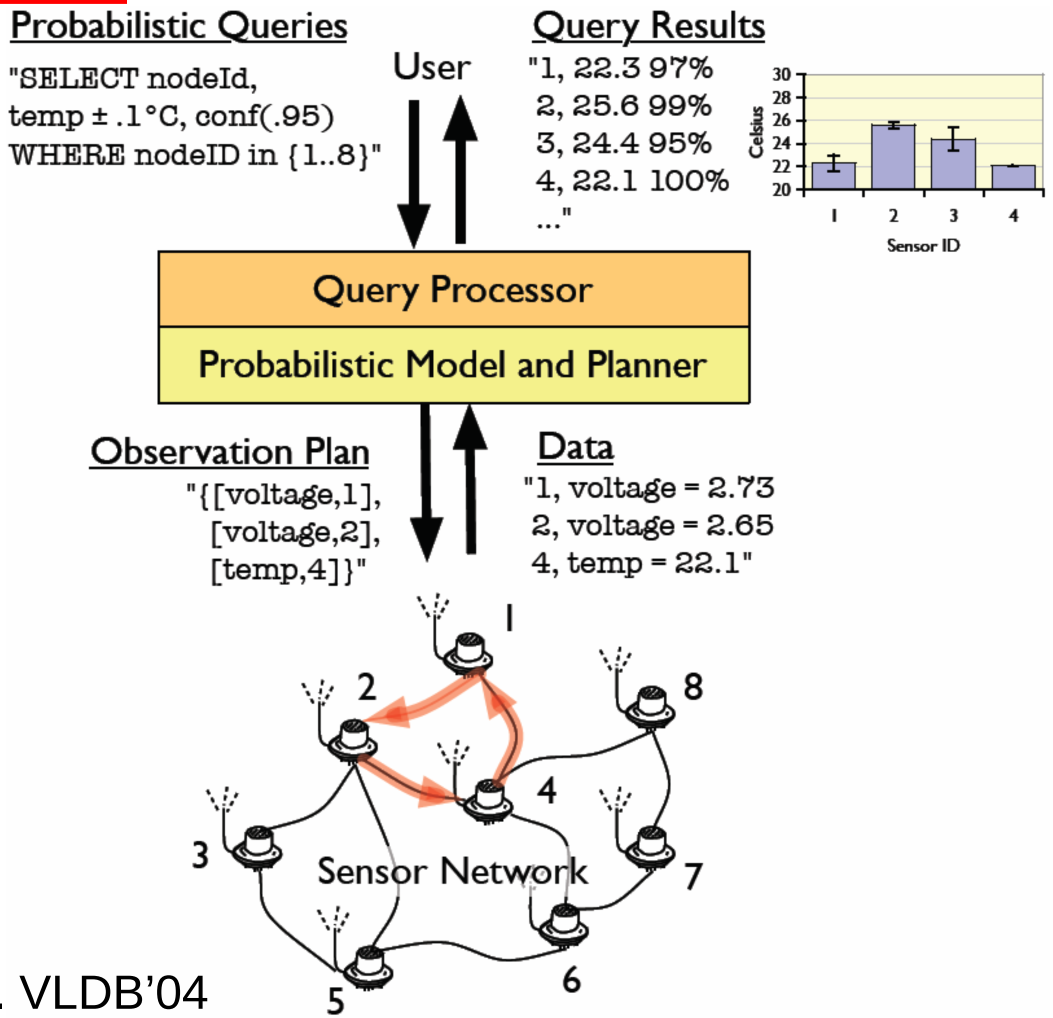


Uncertain Data Acquisition

- Statistics-based, model-driven approaches are often used
- Misrepresentations of data in sensor networks
  - Impossible to collect all relevant data – potentially infinite
  - Samples are non-uniform in time and space due to non-uniform placement of sensors in space, faulty sensors, high packet loss rates, …

A Model-driven Approach

- Treat each sensor as a variable
  - Hidden variables (e.g., whether a sensor faulty) can be added
- Learn a model (a multivariate probability density function)
  - A machine learning/data mining problem
- Given a query, compute a query plan optimal in communication cost to achieve the specified confidence



Deshpande et al. VLDB'04

Outline

- Uncertainty and uncertain data, where and why?
- Models for uncertain and probabilistic data
- (coffee break)
- OLAP on uncertain and probabilistic data
- Mining uncertain and probabilistic data
- Tools: querying uncertain and probabilistic data
  - Indexing uncertain and probabilistic data
  - Ranking queries and spatial queries
- Summary and discussion

Levels of Uncertainty

- Uncertainty can exist in object/tuple level and attribute level
- Object/tuple level uncertainty
  - An object/tuple takes a probability to appear (existing probability)
- Attribute level uncertainty
  - An attribute of an object/tuple takes a few possible values

Probabilistic Database Model

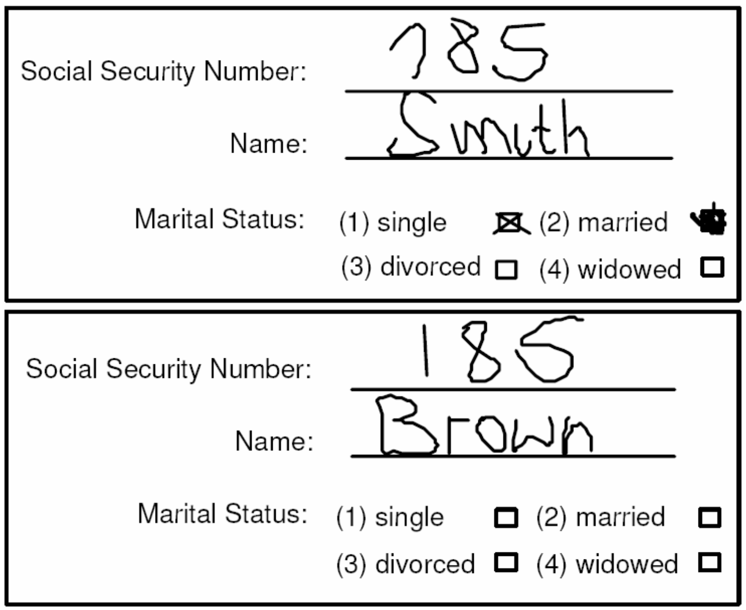
Speed of cars detected by radar

	Time	Radar Location	Car make	Plate No.	Speed	Confidence
t1	11:45	L1	Honda	X-123	130	0.4
t2	11:50	L2	Toyota	Y-245	120	0.7
t3	11:35	L3	Toyota	Y-245	80	0.3
t4	12:10	L4	Mazda	W-541	90	0.4
t5	12:25	L5	Mazda	W-541	110	0.6
t6	12:15	L6	Nissan	L-105	105	1.0

Generation rules: (t2⊕t3), (t4⊕t5)

- The values of each tuple are certain
- Each tuple carries an existing/membership probability
- Generation rules: constraints specifying exclusive tuples

Survey Data Example


Social Security Number: 185
Name: Smith
Marital Status: (1) single (2) married
(3) divorced (4) widowed

Social Security Number: 185
Name: Brown
Marital Status: (1) single (2) married
(3) divorced (4) widowed

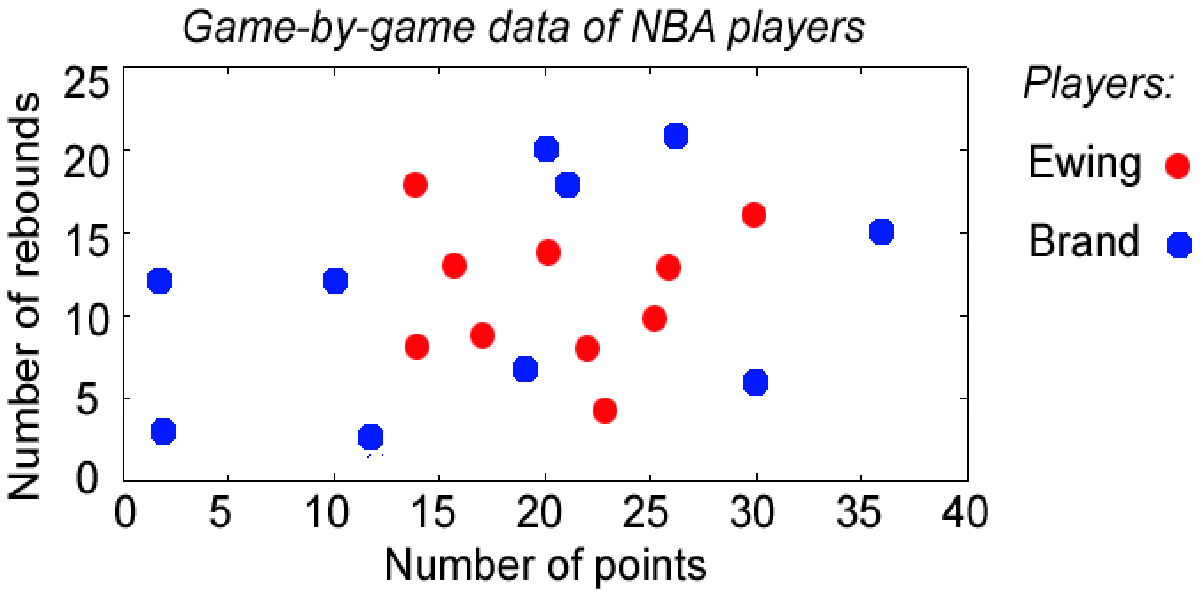
Antova et al. ICDE'07

TID	Name	SSN	Confidence
t1	Smith	185	40%
t2	Smith	785	60%
t3	Brown	185	50%
t4	Brown	186	50%

Generation rules:

$t1 \oplus t2$,
$t3 \oplus t4$,
$t1 \oplus t3$

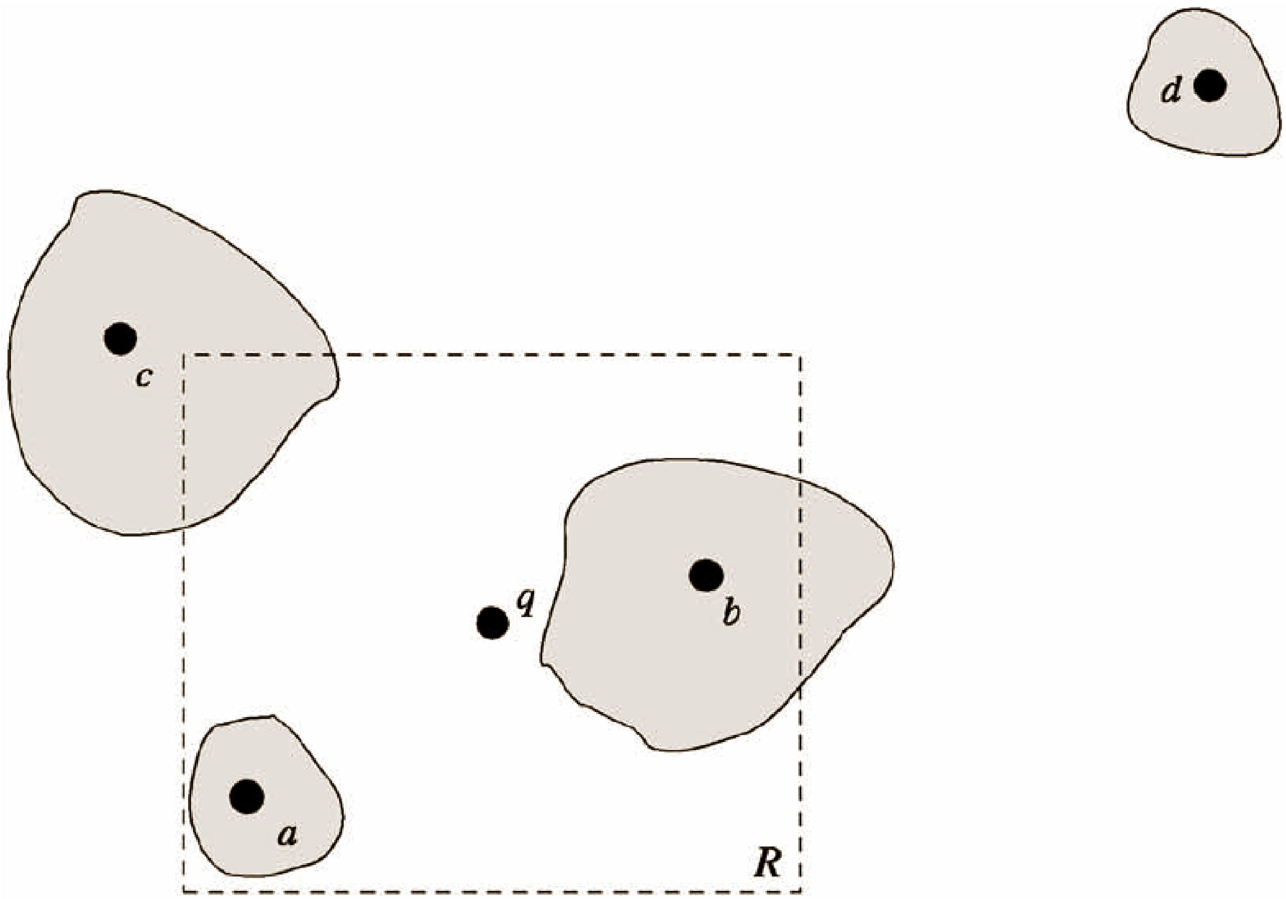
Uncertain Objects



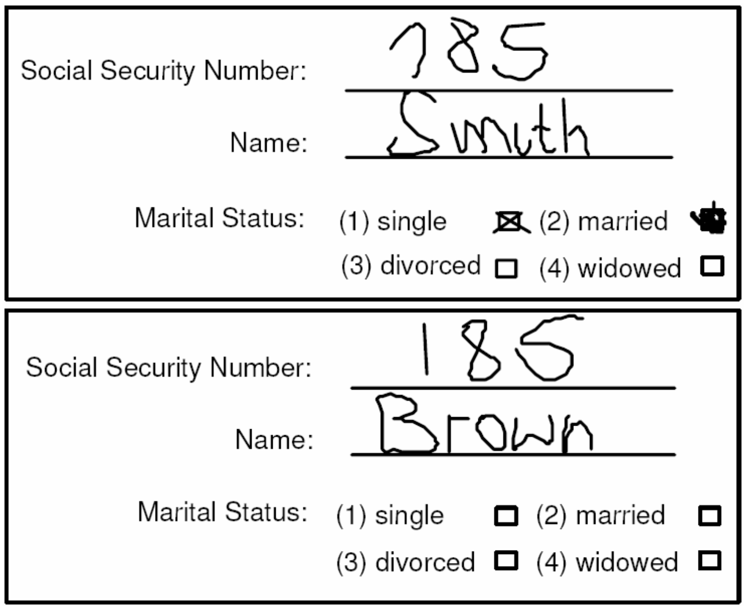
Uncertain objects: NBA players

- An object is uncertain in a few attributes
- Use a sample or a probability density function to capture the distribution on uncertain attributes

Uncertainty of Mobile Objects



Survey Data Example



Antova et al. ICDE'07

Object 1: Smith

Smith.SSN=785, 60%

Smith.SSN=185, 40%

Object 2: Brown

Brown.SSN=185, 50%

Brown.SSN=186, 50%

Constraints:

"Smith.SSN=185" $\oplus$ "Brown.SSN=185"

Prob Table vs. Uncertain Objects

- A probabilistic table can be represented as a set of uncertain objects
  - All tuples in a generation rule are modeled as an uncertain object
  - Use NULL instances to make the sum of membership probabilities in one object to 1
- Uncertain objects with discrete instances can be represented using a probabilistic table
  - One record per instance
  - All instances of an object are constrained by one generation rule
  - Uncertain objects with continuous probability density functions cannot be represented using a finite probabilistic table
- More complicated constraints may not be captured in the transformation

Prob Table vs. Uncertain Objects

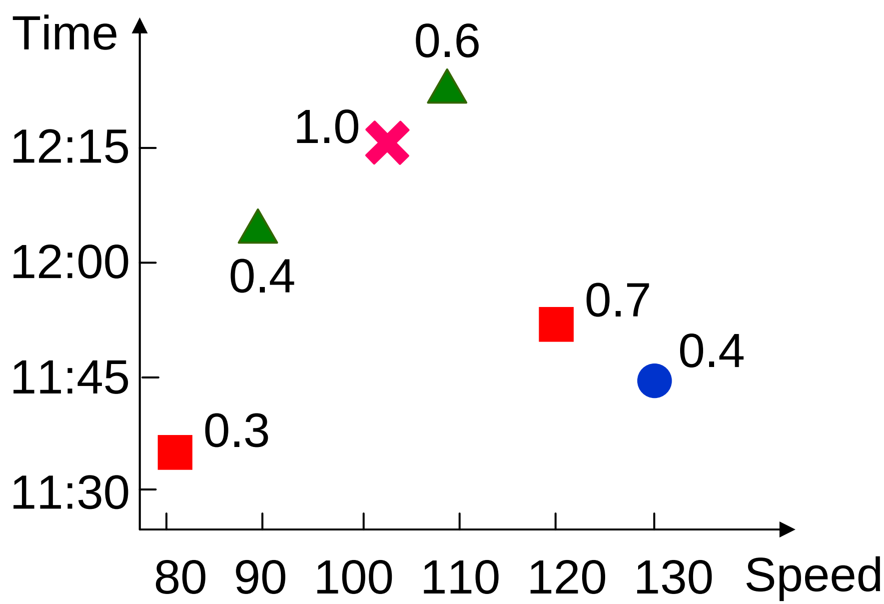
A probabilistic table | **A set of uncertain objects**

A tuple ⟷ An instance

A generation rule ⟷ An uncertain object

	Time	Radar Loc	Car Make	Plate No	Speed	Conf
t1	11:45	L1	Honda	X-123	130	0.4
t2	11:50	L2	Toyota	Y-245	120	0.7
t3	11:35	L3	Toyota	Y-245	80	0.3
t4	12:10	L4	Mazda	W-541	90	0.4
t5	12:25	L5	Mazda	W-541	110	0.6
t6	12:15	L6	Nissan	L-105	105	1.0

$Rules: (t2 \oplus t3), (t4 \oplus t5)$



Possible Worlds

- A possible world
  - a possible snapshot that may be observed
- Probabilistic database model
  - A possible world = a set of tuples
  - At most one tuple per generation rule in a possible world
- Uncertain object model
  - A possible world = a set of instances of uncertain objects
  - At most one instance per object in a possible world
- A possible world carries an existence probability

An Example of Possible Worlds

$0.112 = 0.4 \times 0.7 \times 0.4 \times 1.0$

$0.4 = 0.112 + 0.168 + 0.048 + 0.072$

	Time	Radar Loc	Car Make	Plate No	Speed	Conf
t1	11:45	L1	Honda	X-123	130	0.4
t2	11:50	L2	Toyota	Y-245	120	0.7
t3	11:35	L3	Toyota	Y-245	80	0.3
t4	12:10	L4	Mazda	W-541	90	0.4
t5	12:25	L5	Mazda	W-541	110	0.6
t6	12:15	L6	Nissan	L-105	105	1.0

$Rules: (t2 \oplus t3), (t4 \oplus t5)$

A probabilistic table

World	Prob.
$PW^1=\{t1,t2,t6,t4\}$	0.112
$PW^2=\{t1,t2,t5,t6\}$	0.168
$PW^3=\{t1,t6,t4,t3\}$	0.048
$PW^4=\{t1,t5,t6,t3\}$	0.072
$PW^5=\{t2,t6,t4\}$	0.168
$PW^6=\{t2,t5,t6\}$	0.252
$PW^7=\{t6,t4,t3\}$	0.072
$PW^8=\{t5,t6,t3\}$	0.108

Possible worlds

t2 and t3 never appear in the same possible world!

Possible Worlds and Rules

- Possible worlds are governed by rules

	Time	Radar Loc	Car Make	Plate No	Speed	Conf
t1	11:45	L1	Honda	X-123	130	0.4
t2	11:50	L2	Toyota	Y-245	120	0.7
t3	11:35	L3	Toyota	Y-245	80	0.3
t4	12:10	L4	Mazda	W-541	90	0.4
t5	12:25	L5	Mazda	W-541	110	0.6
t6	12:15	L6	Nissan	L-105	105	1.0

World	Prob.
$PW^1=\{t1,t2,t6,t4\}$	0.16
$PW^2=\{t1,t2,t5,t6\}$	0.24
$PW^3=\{t2,t6,t4\}$	0.12
$PW^4=\{t2,t5,t6\}$	0.18
$PW^5=\{t6,t4,t3\}$	0.12
$PW^6=\{t5,t6,t3\}$	0.18

$$Rules : (t2 \oplus t3),\ (t4 \oplus t5),\ \boxed{(t1 \rightarrow t2)}$$

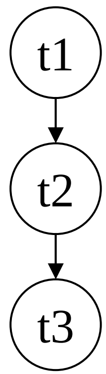
A new rule

Correlation and Dependencies

- An example of correlated tuples

TID	Confidence
t1	0.4
t2	0.42
t3	0.468

A probabilistic table



t1	f_1
0	0.6
1	0.4

t1	**t2**	f_{12}
0	0	0.9
0	1	0.1
1	0	0.1
1	1	0.9

t2	**t3**	f_{23}
0	0	0.7
0	1	0.3
1	0	0.3
1	1	0.7

Dependencies among tuples

- Factored representations

$$\Pr(t1 = x1, t2 = x2, t3 = x3)$$
$$= f_1(t1 = x1)\, f_{12}(t1 = x1, t2 = x2)\, f_{23}(t2 = x2, t3 = x3)$$

Possible Worlds

- Compute the joint probability of possible world assignments (Details in [Sen and Deshpande, ICDE'07])

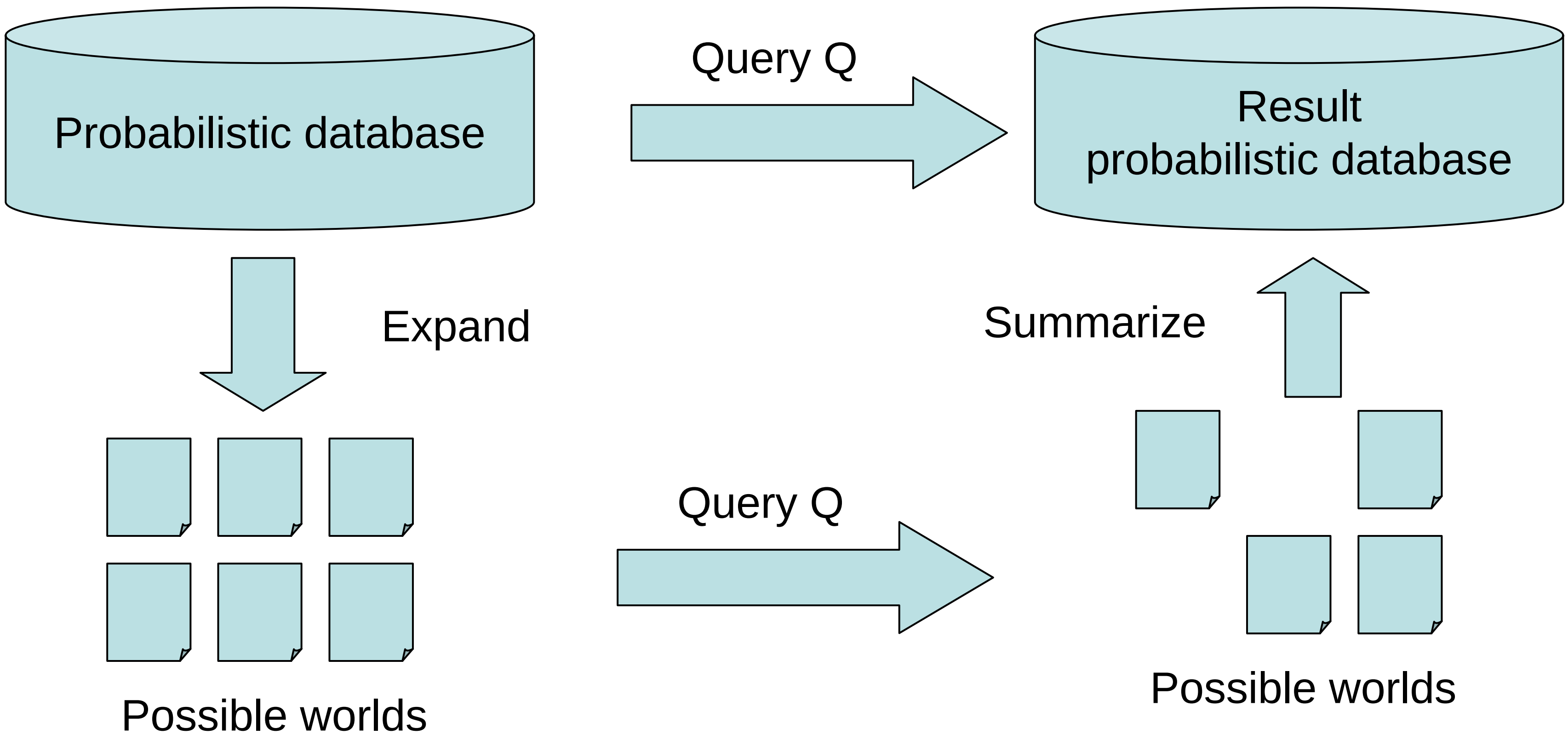
t1	t2	t3	Pr(t1,t2,t3)
0	0	0	0.378
0	0	1	0.162
0	1	0	0.018
0	1	1	0.042
1	0	0	0.028
1	0	1	0.012
1	1	0	0.108
1	1	1	0.252

Joint probability of (t1,t2,t3)

World	Probability
PW1=∅	0.378
PW2={t3}	0.162
PW3={t2}	0.018
PW4={t2,t3}	0.042
PW5={t1}	0.028
PW6={t1,t3}	0.012
PW7={t1,t2}	0.108
PW8={t1,t2,t3}	0.252

Possible worlds

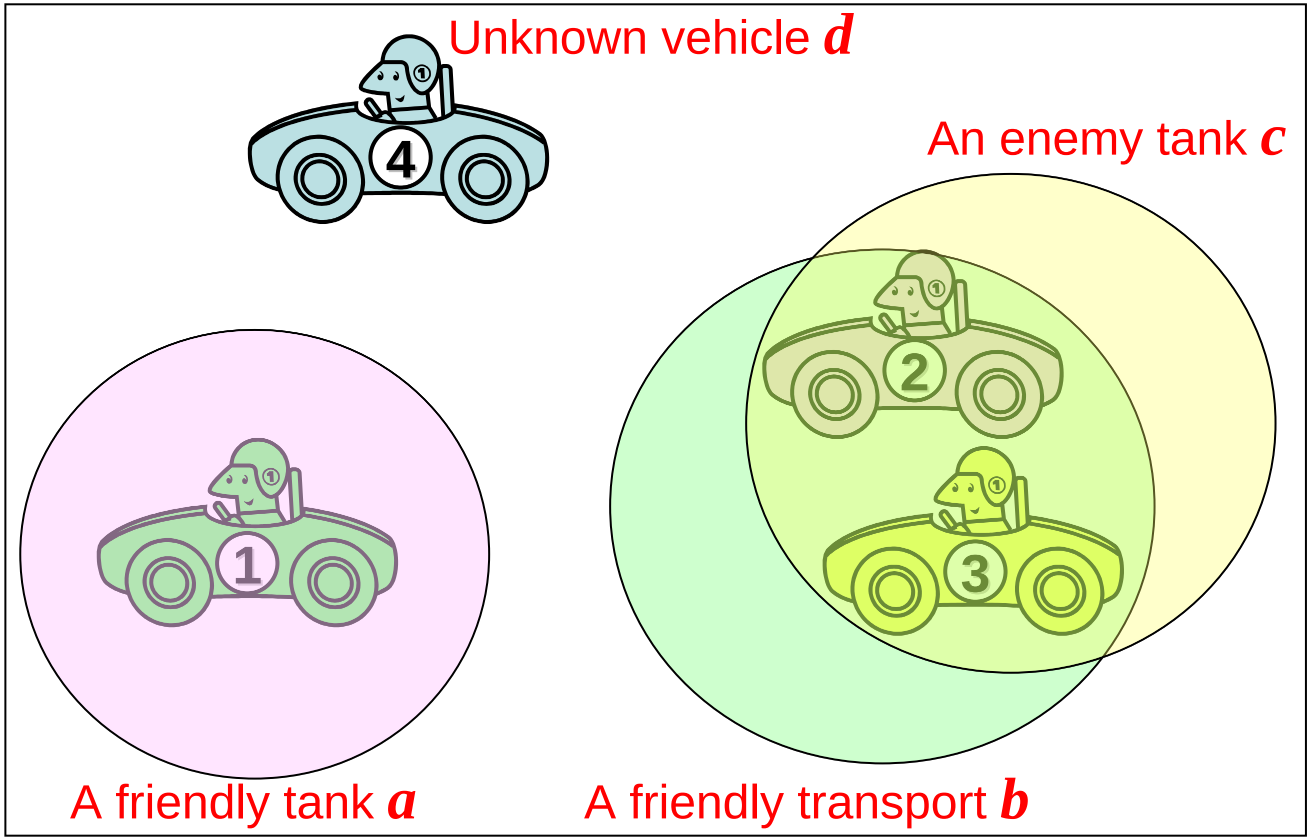
Conceptual Query Answering



Adapted from Singh et al. ICDE'08

Attribute Level Uncertainty

- An aerial photograph of a battlefield



Antova et al. ICDE'08

Attribute Level Uncertainty

- A relation R(ID, Type, Faction) with uncertain attributes
  - ID = { 1, 2, 3, 4 }
  - Type = { Tank, Transport }
  - Faction = { Friend, Enemy }
- Uncertainty in data
  - Vehicle 1 is a friendly tank *a*
  - Vehicle 2 and 3 are either
    - a friendly transport *b*, or
    - an enemy tank *c*
  - Vehicle 4 is unknown vehicle *d*

Vehicle	ID	Type	Faction
a	1	Tank	Friend
b	?	Transport	Friend
c	?	Tank	Enemy
d	4	?	?

Representing Uncertainty

- ID of vehicle b and c
  - "b's ID is 2 and c's ID is 3", or "b's ID is 3 and c's ID is 2"?
  - Random variable x={1,2}

- Type of Vehicle d
  - "Tank" or "Transport"?
  - Random variable y={1,2}

- Faction of Vehicle d
  - "Friend" or "Enemy"?
  - Random variable z={1,2}

Vehicle	**ID**	**Type**	**Faction**
a	1	Tank	Friend
b	**?**	Transport	Friend
c	**?**	Tank	Enemy
d	4	**?**	**?**

U-Relation

- Vertical Representation
  – Use a U-relation to represent each attribute of relation R

D	Vehicle	ID
	a	1
x=1	b	2
x=1	c	3
x=2	b	3
x=2	c	2
	d	4

U-relation for "ID"

D	Vehicle	Type
	a	Tank
	b	Transport
	c	Tank
y=1	d	Tank
y=2	d	Transport

U-relation for "Type"

D	Vehicle	Faction
	a	Friend
	b	Friend
	c	Enemy
z=1	d	Friend
z=2	d	Enemy

U-relation for "Faction"

Possible Worlds of U-Relations

x	y	z	World		
			b.ID & c.ID (x)	**d.Type (y)**	**d.Faction(z)**
1	1	1	b.ID=2, c.ID=3	Tank	Friend
1	1	2	b.ID=2, c.ID=3	Tank	Enemy
1	2	1	b.ID=2, c.ID=3	Transport	Friend
1	2	2	b.ID=2, c.ID=3	Transport	Enemy
2	1	1	b.ID=3, c.ID=2	Tank	Friend
2	1	2	b.ID=3, c.ID=2	Tank	Enemy
2	2	1	b.ID=3, c.ID=2	Transport	Friend
2	2	2	b.ID=3, c.ID=2	Transport	Enemy

Possible worlds

Transformation of U-Relation

- U-Relations can be transformed to a probabilistic table

Vehicle	ID	Type	Faction
a	1	Tank	Friend
b	?	Transport	Friend
c	?	Tank	Enemy
d	4	?	?

b.ID=2, c.ID=3 (30%)
b.ID=3, c.ID=2 (70%)
d.Type=Tank(50%),Transport(50%)
d.Faction=Friend (50%), Enemy(50%)

TID	Vehicle	ID	Type	Faction	Conf.
t1	a	1	Tank	Friend	1
t2	b	2	Transport	Friend	0.3
t3	c	3	Tank	Enemy	0.3
t4	b	3	Tank	Enemy	0.7
t5	c	2	Transport	Friend	0.7
t6	d	4	Tank	Friend	0.25
t7	d	4	Tank	Enemy	0.25
t8	d	4	Transport	Friend	0.25
t9	d	4	Transport	Enemy	0.25

Generation rules: $t2 \rightarrow t3$, $t4 \rightarrow t5$, $t2 \oplus t4$, $t3 \oplus t5$
$t6 \oplus t7 \oplus t8 \oplus t9$

Continuous Uncertain Model

- An attribute may take a continuous PDF as the value
- A table $T=(\Sigma_T, \Delta_T)$
  - Σ_T: a relational schema
  - Δ_T: dependency information including pdfs or joint pdfs
  - For each dependent group of uncertain attributes, store history Λ. When a new tuple is added, check whether the dependency remains

Car-id	Location
C1	Gaussian(mean 18, variance 6)
C2	Uniform(center (32, 26), radius 7)

More on Possible Worlds

- The possible world model can be enriched by various kinds of (arbitrarily complicated) constraints
  - Example: if instances A.a and B.b appear, instances C.c or D.d must appear
- Completeness and closure
  - A model M is closed under an operation Op if applying Op on any uncertain relation in M results in an uncertain relation that can be represented in M
  - M is complete if M is closed for all relational operations
  - Completeness → closure, but not the other way
- More details in [Sarma et al. ICDE'06]

Summary

- Object/tuple level and attribute level uncertainty
- Possible worlds model
- Expressiveness
  - Should be closed under the application operations
  - Completeness is even better
- Succinctness: representing a large number of worlds using fairly little space
- Evaluation efficiency: complexity in useful queries
  - Often a tradeoff between succinctness and efficiency
- Ease of use: can be put on top of an RDBMS
- [Antova et al. ICDE'08]

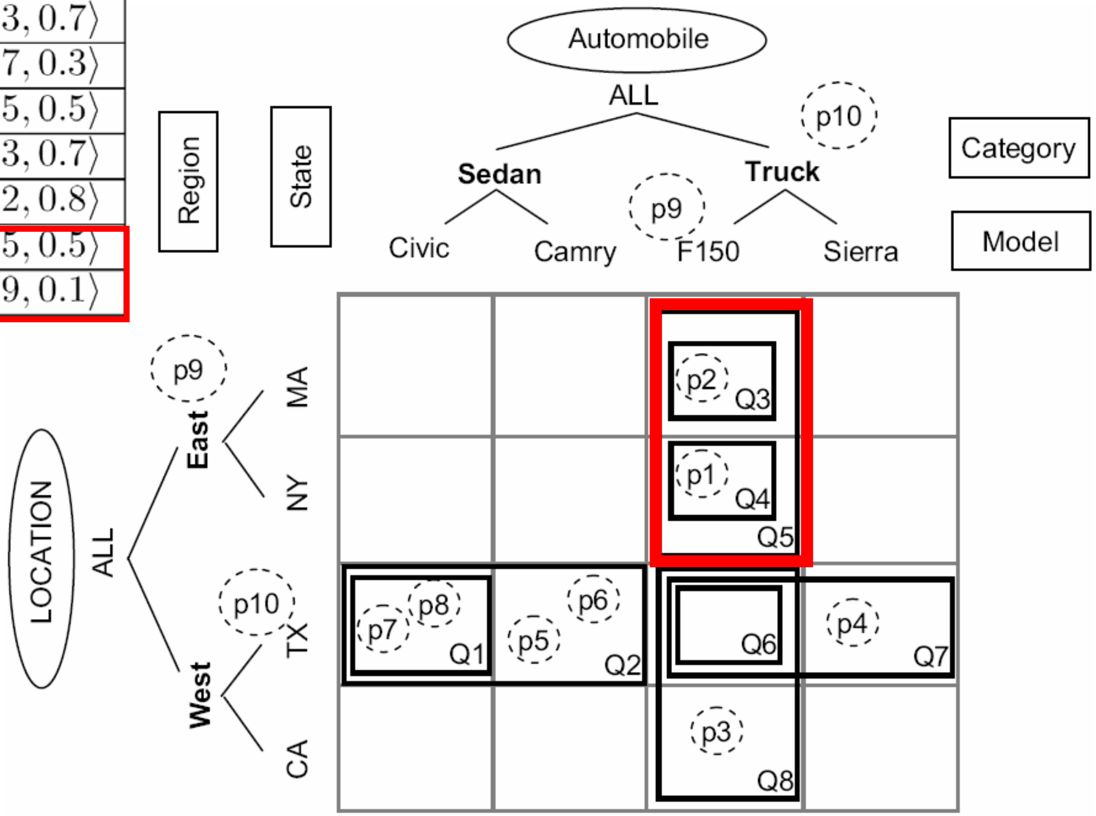
Outline

- Uncertainty and uncertain data, where and why?
- Models for uncertain and probabilistic data
- (coffee break)
- OLAP on uncertain and probabilistic data
- Mining uncertain and probabilistic data
- Tools: querying uncertain and probabilistic data
  - Indexing uncertain and probabilistic data
  - Ranking queries and spatial queries
- Summary and discussion

OLAP Query

	Auto	Loc	Repair	Text	Brake
p1	F-150	NY	\$200	...	$\langle 0.8, 0.2 \rangle$
p2	F-150	MA	\$250	...	$\langle 0.9, 0.1 \rangle$
p3	F-150	CA	\$150	...	$\langle 0.7, 0.3 \rangle$
p4	Sierra	TX	\$300	...	$\langle 0.3, 0.7 \rangle$
p5	Camry	TX	\$325	...	$\langle 0.7, 0.3 \rangle$
p6	Camry	TX	\$175	...	$\langle 0.5, 0.5 \rangle$
p7	Civic	TX	\$225	...	$\langle 0.3, 0.7 \rangle$
p8	Civic	TX	\$120	...	$\langle 0.2, 0.8 \rangle$
p9	F150	East	\$140	...	$\langle 0.5, 0.5 \rangle$
p10	Truck	TX	\$500	...	$\langle 0.9, 0.1 \rangle$

What are the total repair cost for F150's in the East?

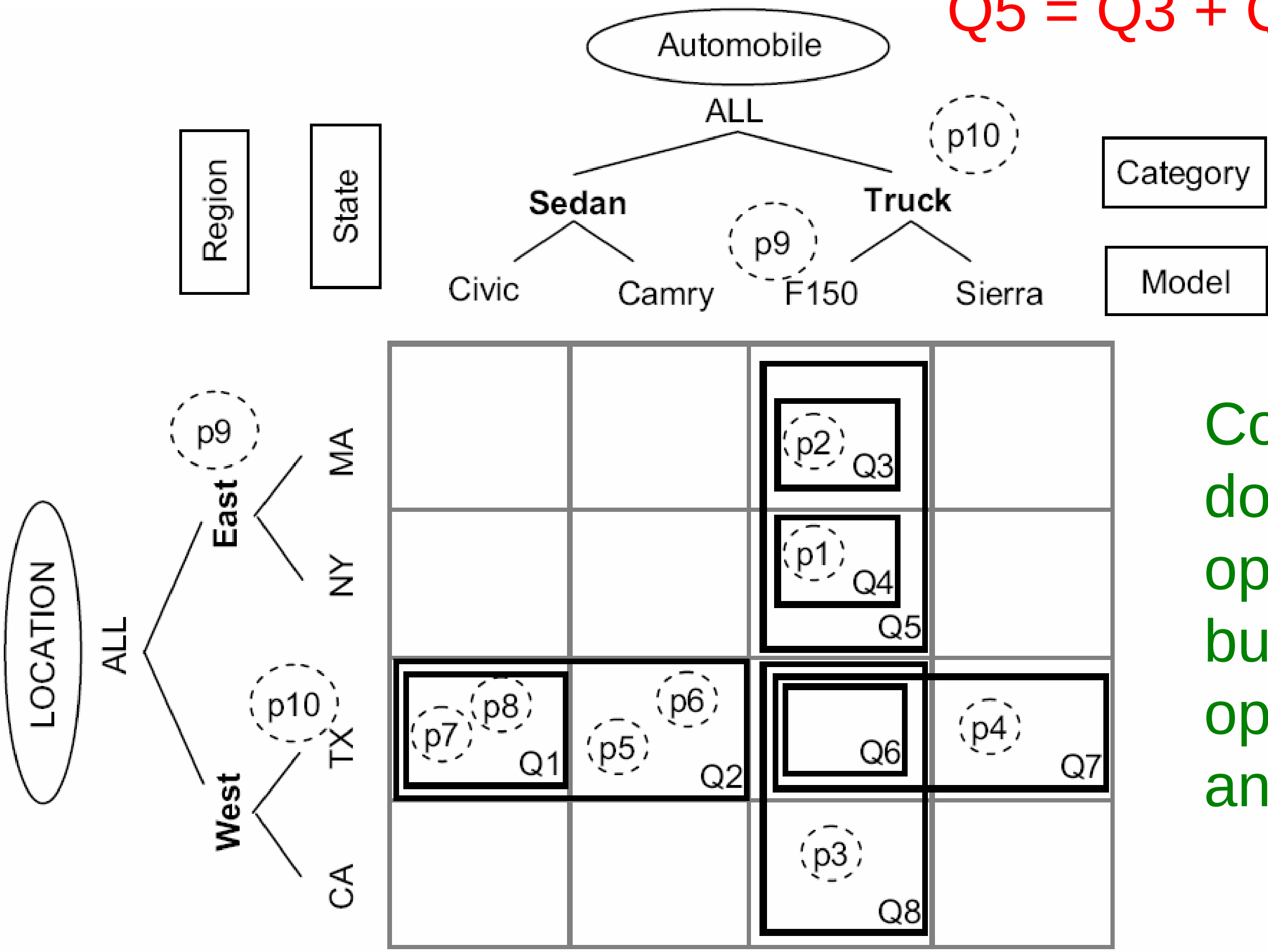


Three Options

- None: ignore all imprecise facts
  - Answer: p1, p2
- Contains: include only those contained in the query region
  - Answer: p1, p2, p9
- Overlaps: include all imprecise facts whose region overlaps the query region
  - Answer: p1, p2, p9, p10

Consistency among OLAP Queries

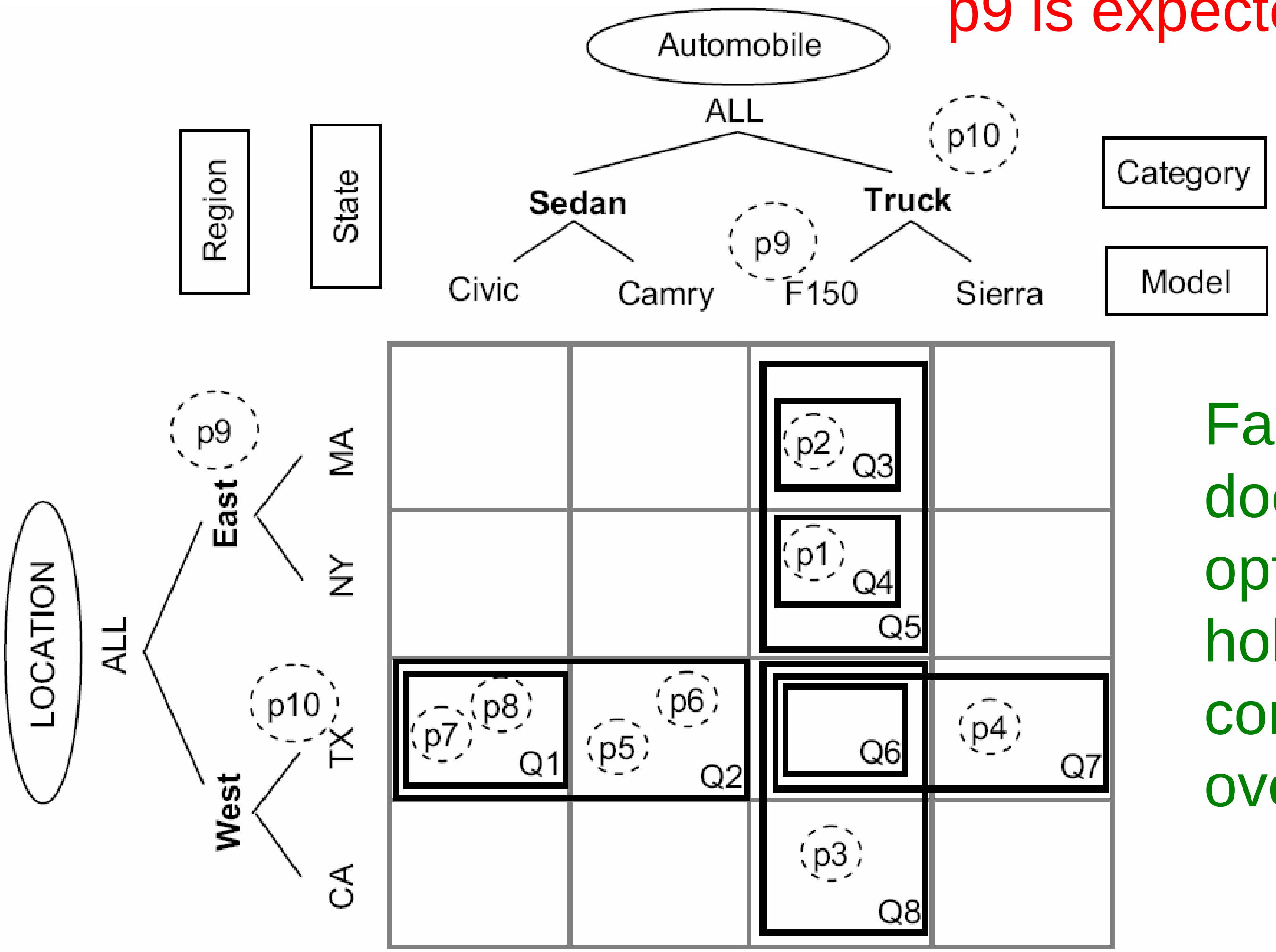
Q5 = Q3 + Q4 is expected!



Consistency does not hold for option contains, but holds for options none and overlaps!

Faithfulness of OLAP Queries

p9 is expected in Q5!

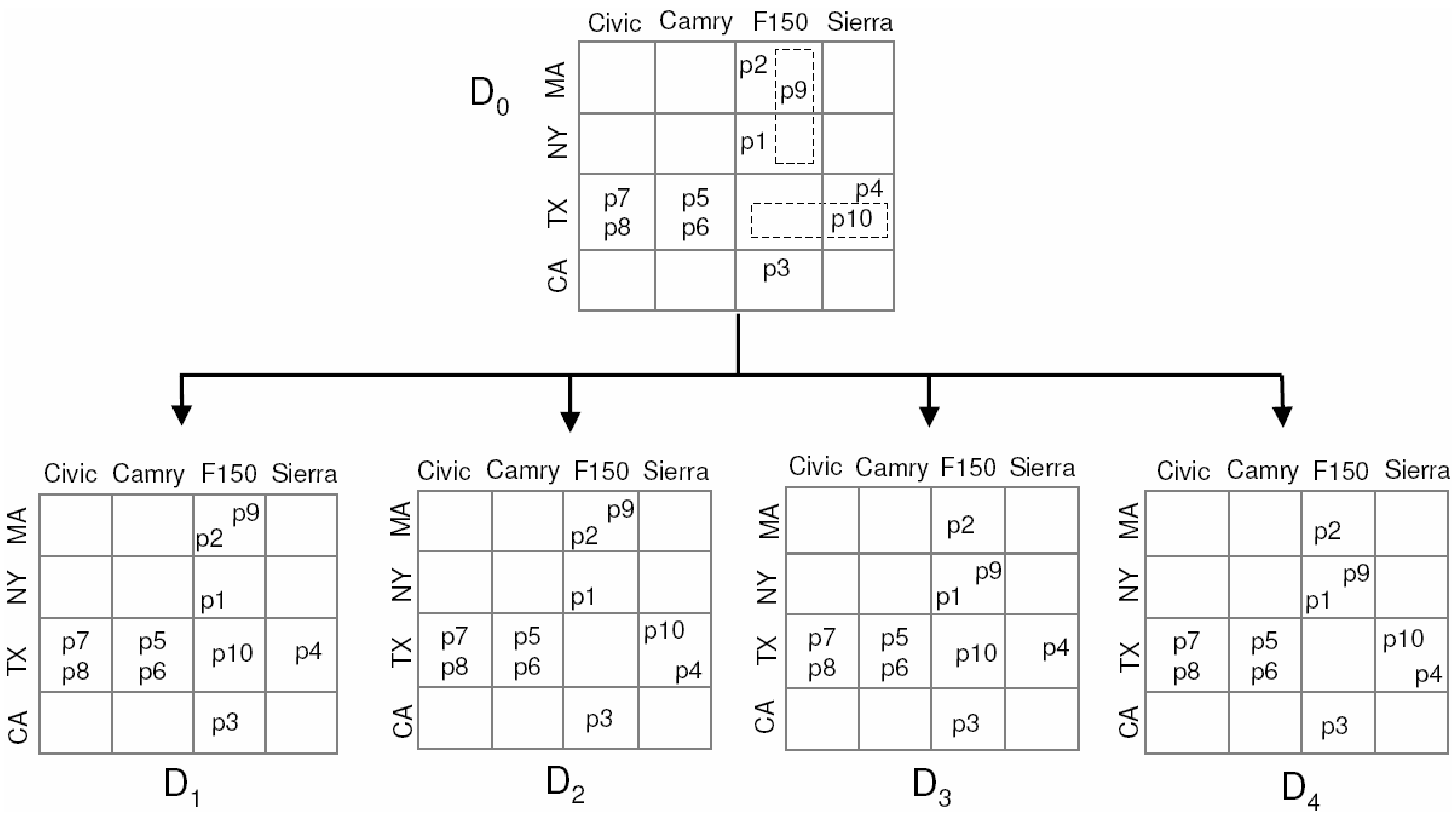


Faithfulness does not hold for option none, but holds for options contains and overlaps!

OLAP Requirements

- Consistency (summarizability): some natural relationships hold between answers to aggregation queries associated with different (connected) regions in a hierarchy
- Faithfulness: imprecise data should be considered properly in query answering

Possible Worlds



Allocation and Query Answering

- The allocation weights encode a set of possible worlds $D_1, \ldots, D_m$ with associated weights $w_1, \ldots, w_m$
- The answer to a query is a multiset $\{v_1, \ldots, v_m\}$
- Problem: how to summarize $\{v_1, \ldots, v_m\}$ properly?

Answer Variable

- Consider multiset $\{v_1, \ldots, v_m\}$ of possible answers to a query Q
- Define the answer variable Z associated with Q to be a random variable with probability density function

$$\Pr[Z=v_i]=\Sigma_{j \text{ s.t. } vi=vj}\, w_j,\ 1 \leq i, j \leq m$$

Answer Variable

- The answer to a query can be summarized as the first and the second moments (expected value and variance) of the answer variable Z
- Basic faithfulness is satisfied if answers to queries are computed using the expected value of the answer variable

Query Answering

- Identify the set of candidate facts and compute the corresponding allocations to Q
  - Identifying candidate facts: using a filter for the query region
  - Computing the corresponding allocations: identifying groups of facts that share the same identifier in the ID column, then summing up the allocations within each group
- Identify the information necessary to compute the summarization while circumventing the enumeration of possible worlds

Allocation Policies

- Dimension-independent allocation such as uniform allocation
- Measure-oblivious allocation such as count-based allocation
  - If Vancouver and Victoria have 100 and 50 F150's, respectively, and there are another 30 in BC as imprecise records, then allocate 20 and 10 to Vancouver and Victoria, respectively

Outline

- Uncertainty and uncertain data, where and why?
- Models for uncertain and probabilistic data
- (coffee break)
- OLAP on uncertain and probabilistic data
- Mining uncertain and probabilistic data
- Tools: querying uncertain and probabilistic data
  - Indexing uncertain and probabilistic data
  - Ranking queries and spatial queries
- Summary and discussion

Probabilistic Transactions

- A transaction t contains a number items where each item x is associated with a positive probability $P_t(x)$
  - Assuming items in a transaction are independent
  - Itemset xyz has probability $P_t(x)P_t(y)P_t(z)$ to happen in t
- In a probabilistic transaction database D of d transactions, an itemset X is frequent if its expected support is at least ρd, where ρ is a user-specified support threshold
  - [Chui et al., PAKDD'07]

Possible Worlds of Transactions

- Enumerating all possible worlds to compute the expected supports is computationally infeasible for large transaction databases

W_1

	A	B
t_1	✔	✔
t_2	✔	✔

W_2

	A	B
t_1	✔	✔
t_2	✔	✘

W_3

	A	B
t_1	✔	✔
t_2	✘	✔

W_4

	A	B
t_1	✔	✘
t_2	✔	✔

W_5

	A	B
t_1	✘	✔
t_2	✔	✔

W_6

	A	B
t_1	✔	✔
t_2	✘	✘

W_7

	A	B
t_1	✘	✘
t_2	✔	✔

W_8

	A	B
t_1	✔	✘
t_2	✔	✘

W_9

	A	B
t_1	✘	✔
t_2	✘	✔

W_{10}

	A	B
t_1	✘	✔
t_2	✔	✘

W_{11}

	A	B
t_1	✔	✘
t_2	✘	✔

W_{12}

	A	B
t_1	✘	✘
t_2	✔	✘

W_{13}

	A	B
t_1	✘	✘
t_2	✘	✔

W_{14}

	A	B
t_1	✘	✔
t_2	✘	✘

W_{15}

	A	B
t_1	✔	✘
t_2	✘	✘

W_{16}

	A	B
t_1	✘	✘
t_2	✘	✘

Chui et al., PAKDD'07

Independent Transactions

- If transactions are independent, expected support can be calculated efficiently transaction by transaction

$$S_e(X) = \sum_{j=1}^{d} \prod_{x \in X} P_{t_j}(x)$$

- Anti-monotonicity still holds: if X is infrequent, then every super set of X cannot be frequent
- U-Apriori: extending Apriori straightforwardly

Insignificant Support Contributions

- If a, b, c have existence probabilities 5%, 0.5%, and 0.1%, respectively in a transaction t, t contributes only 0.00000025 to the support of abc
  - In certain transactions, every transaction contributes 1 to the support of an itemset
- Counting many insignificant support contributions is costly

The Data Trimming Framework

- Obtain D^T by removing the items with existential probabilities smaller than a trimming threshold ρ_t
  - ρ_t can be either global to all items or local to each item
  - Estimate the error e(X) in support counting introduced by reducing D to D^T
- Mine D^T using U-Apriori
  - If X is frequent in D^T, X must be frequent in D
  - If X is infrequent in D^T, X may or may not be infrequent in D
- If $sup_{DT}(X) + e(X) < \rho d$, then X can be pruned
  - Check supports for only those itemsets that cannot be pruned

Decremental Pruning

- Estimate upper bounds of candidate itemsets' expected supports progressively when transactions are processed
- If a candidate's upper bound falls below the support threshold, the candidate can be pruned immediately
- For X' $\subset$ X, k $\geq$ 0, sup(X) $\leq$ s(X, X', k), where

$$S(X, X', k) = \sum_{i=1}^{k} \prod_{x \in X} P_{t_i}(x) + \sum_{i=k+1}^{d} \prod_{x \in X'} P_{t_i}(x)$$

- Using singleton itemsets or prefix-sharing itemsets to comput s(X, X', k) efficiently
  - Details in [Chui and Kao, PAKDD'08]

Is Expectation Good Enough?

- In D1, if the support threshold is 0.5, then a is frequent, however, a has only 50% chance to have support 0.5
- In D2, if the support threshold is 0.5, then a is infrequent. However, a has a probability of 0.9 to be frequent

D1

TID	Items
t1	(a: 0.5)
t2	(b:0.6)

D2

TID	Items
t1	(a: 0.9), (b: 0.1)
t2	(c:1)

Probabilistic Heavy Hitters

- An item is a (ρ, τ)-probabilistic heavy hitter if

$$\sum_{w \in W, \sup_w(x) \geq \rho d} \Pr(w) \geq \tau$$

  - τ is the probability/confidence threshold
- Dynamic programming using Poisson Binomial Recurrence

Zhang et al., SIGMOD'08

$$B^t[0,0] = 1$$

$$B^t[i,0] = 0 \quad (i \geq 1)$$

$$B^t[i,j] = \begin{cases} B^t[i, j-1] & \text{if } w_j \neq t; \\ B^t[i, j-1](1-p_j) + B^t[i-1, j-1]p_j & \text{if } w_j \neq t. \end{cases}$$

Classification on Uncertain Data

- Many studies exist in machine learning (particularly statistical learning)
  – Examples: [M. Mohri. Learning from Uncertain Data. COLT'03] and [S. Jain et al. Absolute Versus Probabilistic Classification in a Logical Setting. ALT'05]
- New problem: how does uncertain data affect classification?
  – How can we apply the existing classification with minor revision on uncertain data?

1NN Classification on Certain Data

- Point x will be classified using point y since dist(x, y) < dist(x, z)

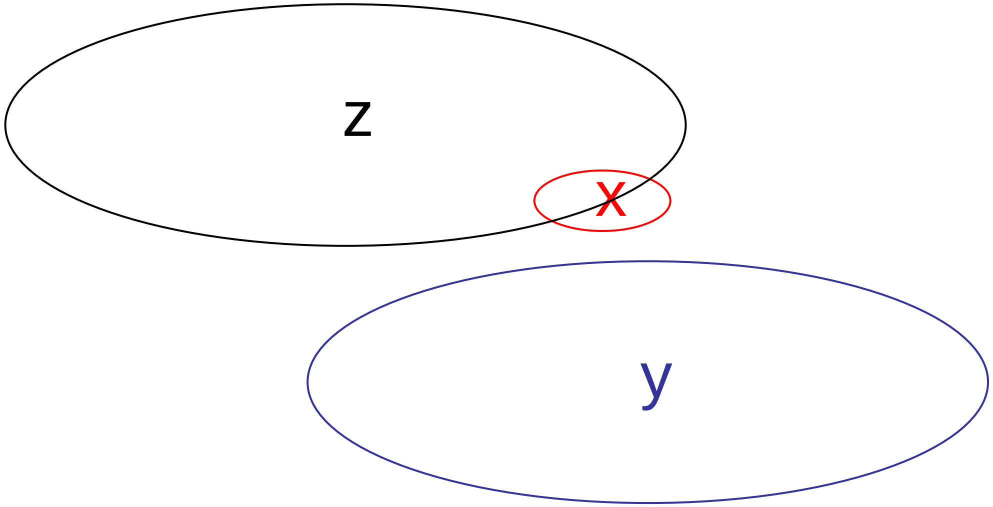
z

x

y

1NN on Uncertain Data

- Object x may have a good chance to be classified using z
  - Instances of x have a high probability to lie in the error boundary of z
- When classification on uncertain data, it is important to use the relative errors of different data points over the different dimensions in order to improve the accuracy



Aggarwal ICDE'07

Density Estimation with Errors

- Kernel estimation
  - General form $\overline{f}(x) = \frac{1}{N}\sum_{i=1}^{N} K'_h(x - \overline{X_i})$
  - Gaussian kernel with width h $\overline{f}(x) = \frac{1}{Nh\sqrt{2\pi}}\sum_{i=1}^{N} e^{-\frac{(x-\overline{X_i})^2}{2h^2}}$
- Error at point $\overline{X_i}$ can be modeled by function $\psi(\overline{X_i})$
- Error-based kernel

$$Q'_h(x - \overline{X_i}, \psi(\overline{X_i})) = \frac{h + \psi(\overline{X_i})}{\sqrt{2\pi}} e^{-\frac{(x-\overline{X_i})^2}{2(h^2 + \psi(\overline{X_i})^2)}}$$

$$\overline{f^Q}(x, \psi(\overline{X_i})) = \frac{1}{N}\sum_{i=1}^{N} Q'_h(x - \overline{X_i}, \psi(\overline{X_i}))$$

Error-Based Micro-Clustering

- Applying density estimation with errors on a large database may be costly
- Use micro-clusters to approximate
  – A BIRCH-like method [Zhang et al., SIGMOD'96]
  – Use the framework in [Aggarwal et al., VLDB'03], but maintain only q randomly chosen centroids
  – When assigning a point into a micro-cluster, use error-adjusted distance

$$dist(\overline{X}, c) = \sum_{i=1}^{d} \max\{0, (X_i - c_i)^2 - \psi_i(\overline{X})^2\}$$

- Micro-clusters can be used to generate classification rules

Fuzzy Clustering

- Each data point is certain
- Clusters are fuzzy (uncertain to some extent)
  - No sharp boundary between clusters, often perform better in some applications
  - Each point is assigned to a cluster with a probability (membership degree)
- Hoppner et al. Fuzzy cluster analysis. Wiley, 1999

Clustering Multi-represented Objects

- An object may have multiple representations
  - Molecules are characterized by an amino acid sequence, a secondary structure and a 3D representation
- Clustering multi-represented objects needs to consider all representations in question
  - Combine distance/neighborhoods in all representations into one global distance/neighborhood

Clustering Uncertain Objects

- Objects are fuzzy/uncertain, clusters can be certain or fuzzy
  - A fuzzy object can be represented by a probability density function or a set of instances
  - All instances of an object are in the same space, different objects may have a different number of instances
- In clustering, the distribution of the distance between two objects and the probability that an object is a cluster center should be considered

$$\Pr[a \le dist(o,o') \le b] = \int_a^b \Pr[dist(o,o') = x]dx$$

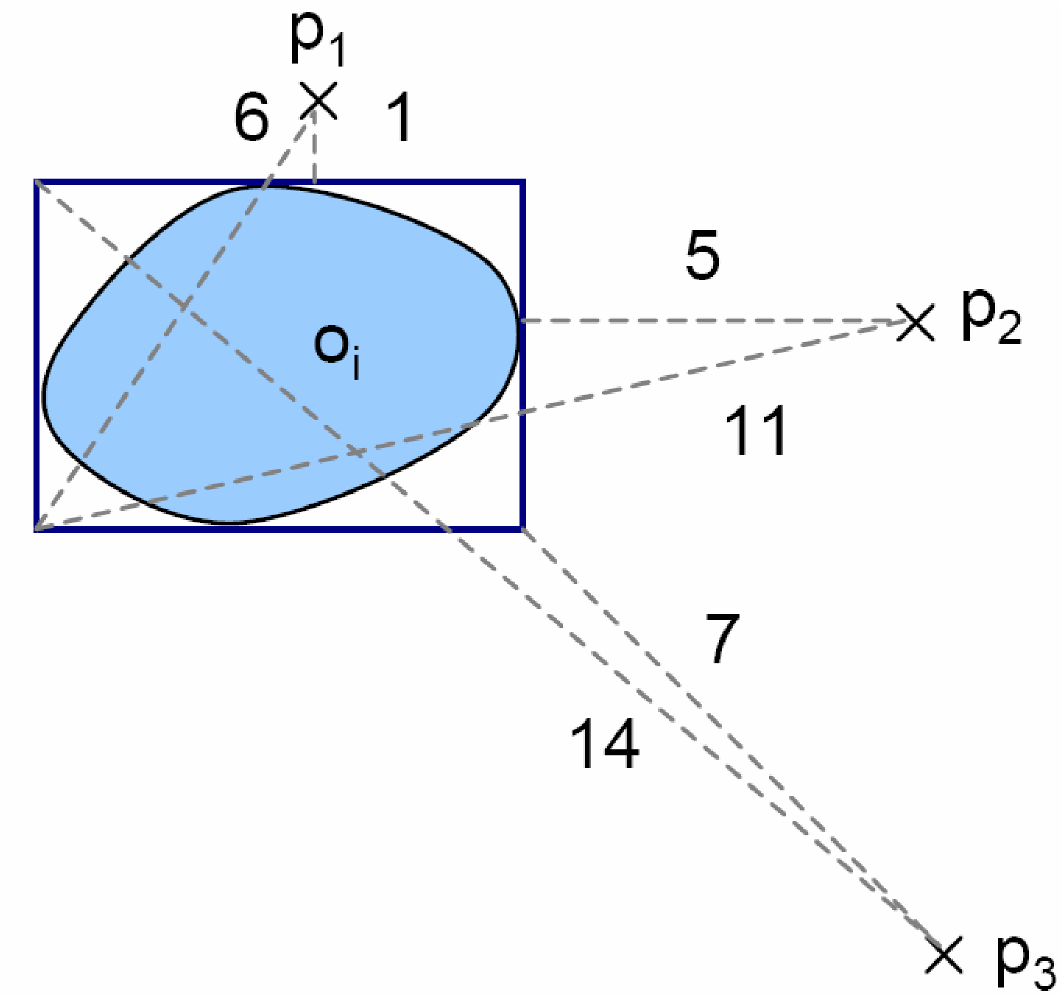
Kriegel and Pfeifle, KDD'05, ICDM'05

K-means on Uncertain Data

- Run k-means, use expectation of distance to assign objects/probabilistic points to clusters
- Computation can be sped up by using bounding rectangles or other polygon to bound PDF regions and approximate distance calculation

Example

- O_i cannot be assigned to p_3



Ngai et al., ICDM'06

(α, β)-bicriteria Approximation

- Optimal k-center, k-means, and k-median are NP-hard even for certain data
- A (α, β)-bicriteria approximation to k-clustering finds a clustering of size βk whose cost is at most α times the cost of the optimal k-clustering
- Assigned clustering: an object is assigned to a cluster
- Unassigned clustering: only cluster centers are computed – different instances of an object may be assigned to different clusters

Theoretical Results

Cormode and McGregor, PODS'08

Objective	Metric	Assignment	α	β
K-center Point probability	Any	Unassigned	$1 + \varepsilon$	$O(\varepsilon^{-1} \log^2 n)$
			$12 + \varepsilon$	2
K-center Discrete PDF	Any	Unassigned	$1.582 + \varepsilon$	$O(\varepsilon^{-1} \log^2 n)$
			$18.99 + \varepsilon$	2
K-means	Euclidean	Unassigned	$1 + \varepsilon$	1
		Assigned		
K-median	Any	Unassigned	$3 + \varepsilon$	1
	Euclidean		$1 + \varepsilon$	
	Any	Assigned	$7 + \varepsilon$	
	Euclidean		$3 + \varepsilon$	

K-center $(1 + \varepsilon)$ Approximation

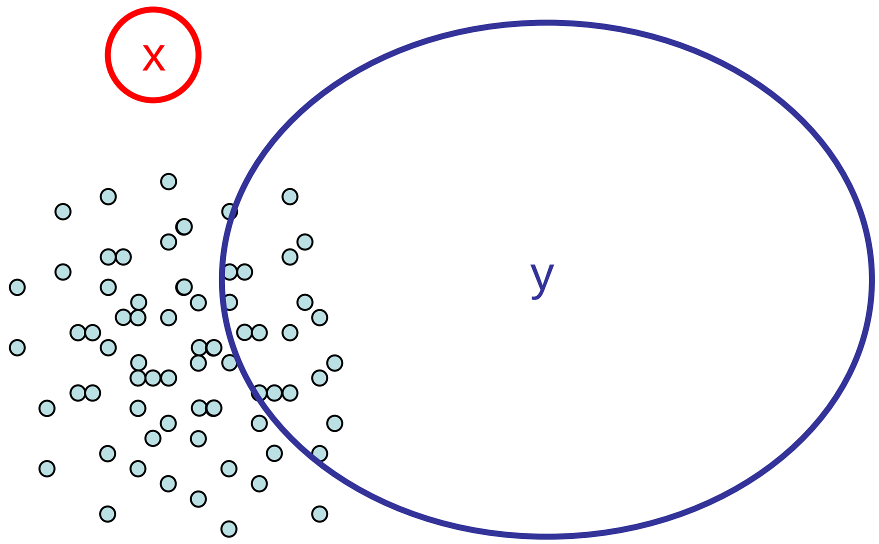
- For each point x, assign a weight $w_x = -\ln(1 - p_x)$
- Greedily select a set of centers
  - Suppose $c_1, \ldots, c_i$ are the current centers
  - A point x is assigned to a current cluster if it is within distance r to the center
  - Among the remaining points, find a new center c_{i+1} such that the total weight of points that can be assigned to c_{i+1} is maximized

Fuzzy Clustering of Uncertain Data

- Data points are probabilistic
- Clusters are fuzzy – each probabilistic point has a membership degree (between 0 and 1) to be assigned to a cluster
- Expectation maximization (EM) based on clustering of uncertain data [Dempster et al., J. of the Royal Stat. Society, 1977]

Outliers in Uncertain Data

- Which one is more an outlier, x or y?



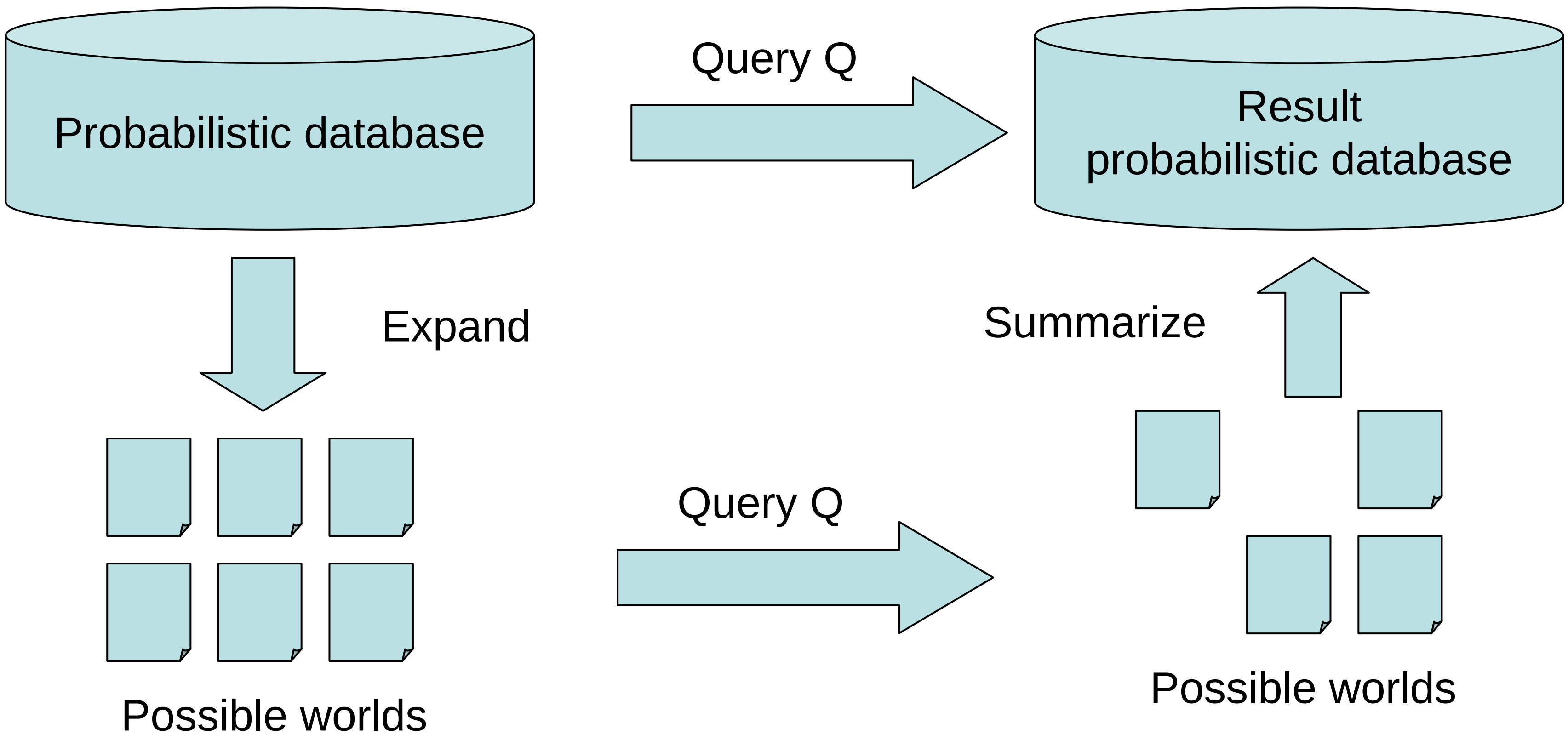
Outlier Detection on Uncertain Data

- The η-probability of a data point is the probability that it lies in a region with data density at least η
- (δ, η)-outlier: the η-probability of a point is some subspace is less than δ
- Enumerate all non-empty subspaces in a bottom-up breadth-first search, for each subspace, check whether there is any (δ, η)-outlier
  – Use sampling and micro-clusters to estimate density distribution
  – Details in [Aggarwal and Yu, SDM'08]

Outline

- Uncertainty and uncertain data, where and why?
- Models for uncertain and probabilistic data
- (coffee break)
- OLAP on uncertain and probabilistic data
- Mining uncertain and probabilistic data
- Tools: querying uncertain and probabilistic data
  - Indexing uncertain and probabilistic data
  - Ranking queries and spatial queries
- Summary and discussion

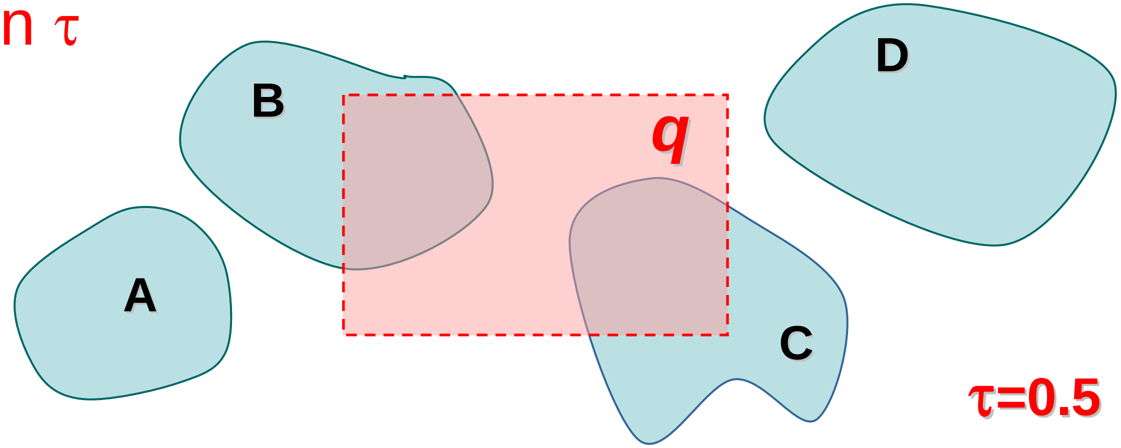
Conceptual Query Answering



Adapted from Singh et al. ICDE'08

U-Tree: Motivation

- Probabilistic range queries
  - Given query region *q* and probability threshold τ, return all the objects whose probability of being in q is higher than τ

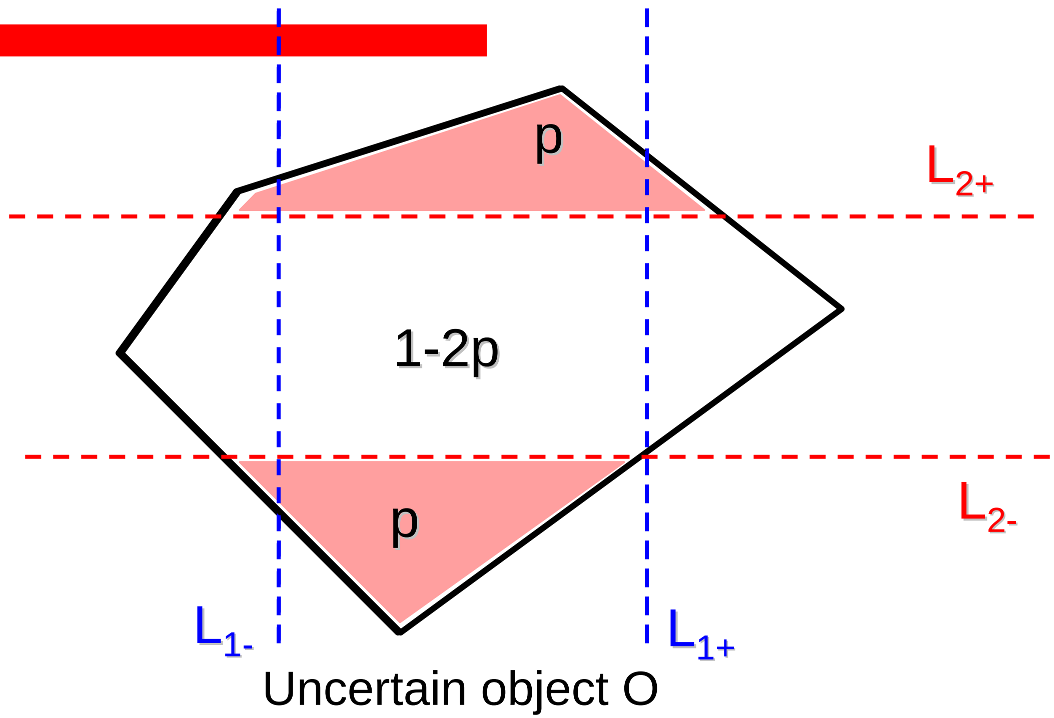


  - Appearance probability

$$\Pr(C \text{ is in } q) = \int_{q \cap C} f(x)dx = \frac{Area(q \cap C)}{Area(C)}$$

where f(x) is the pdf of C

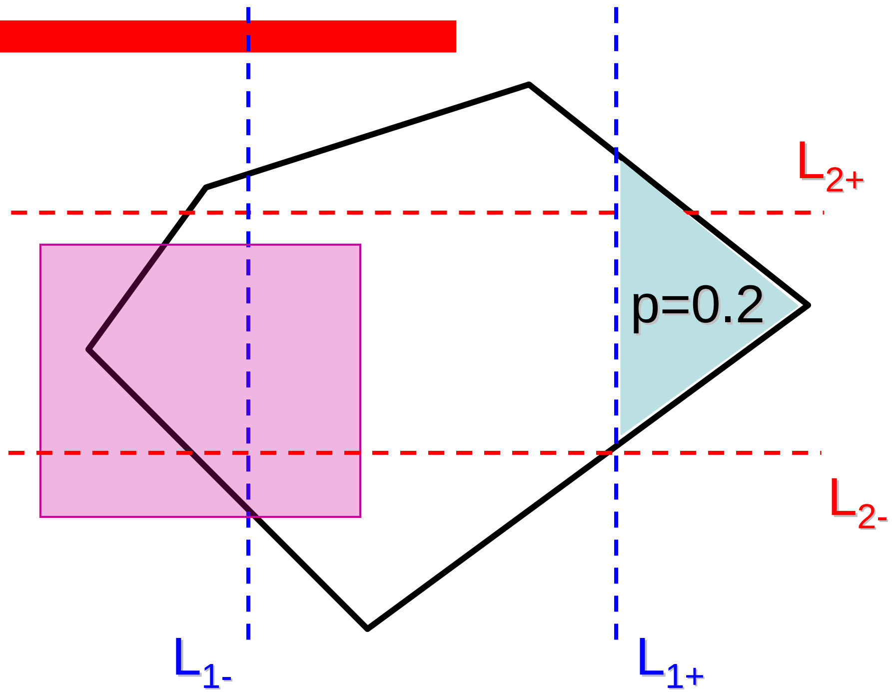
U-Tree: Idea



- Partition the object into three parts in one dimension (horizontally)
- Partition the object into three parts in the other dimension (vertically)

U-Tree: Pruning

Range query q,
Prob. threshold
τ=0.8

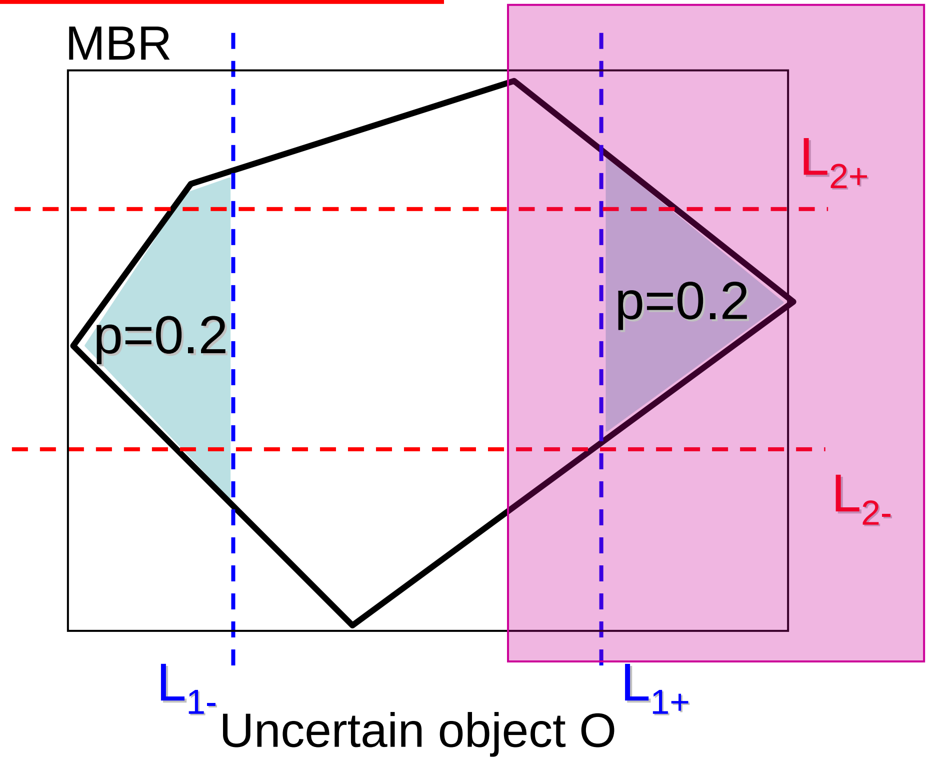


Uncertain object O

- Pr(O is in q)<τ, because *q* is disjoint with the right part of L_{1+}, whose probability is p=0.2
- Thus, O can be pruned

U-Tree: Validation

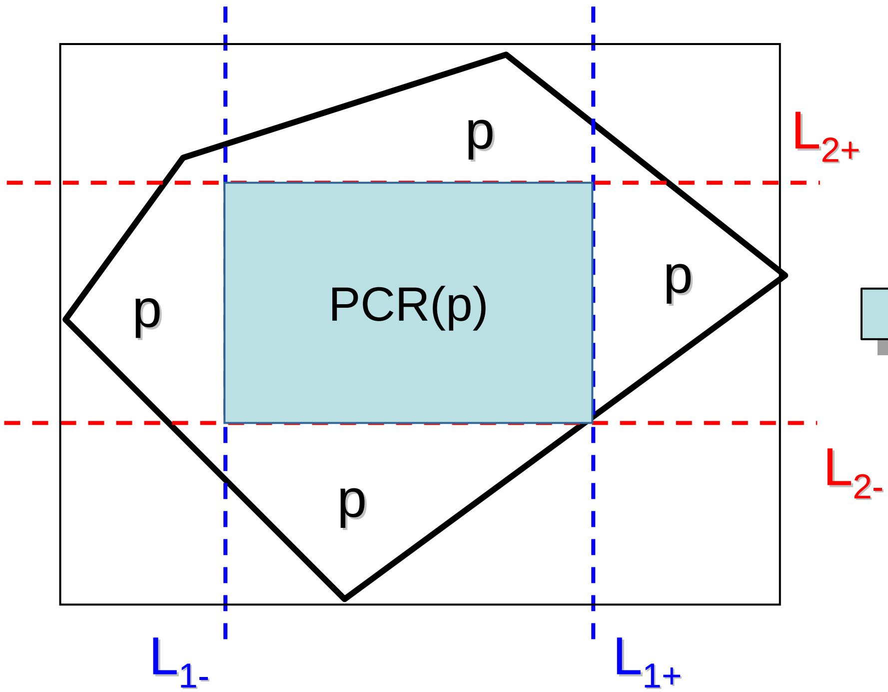
Range query q,
Prob. threshold
τ=0.2



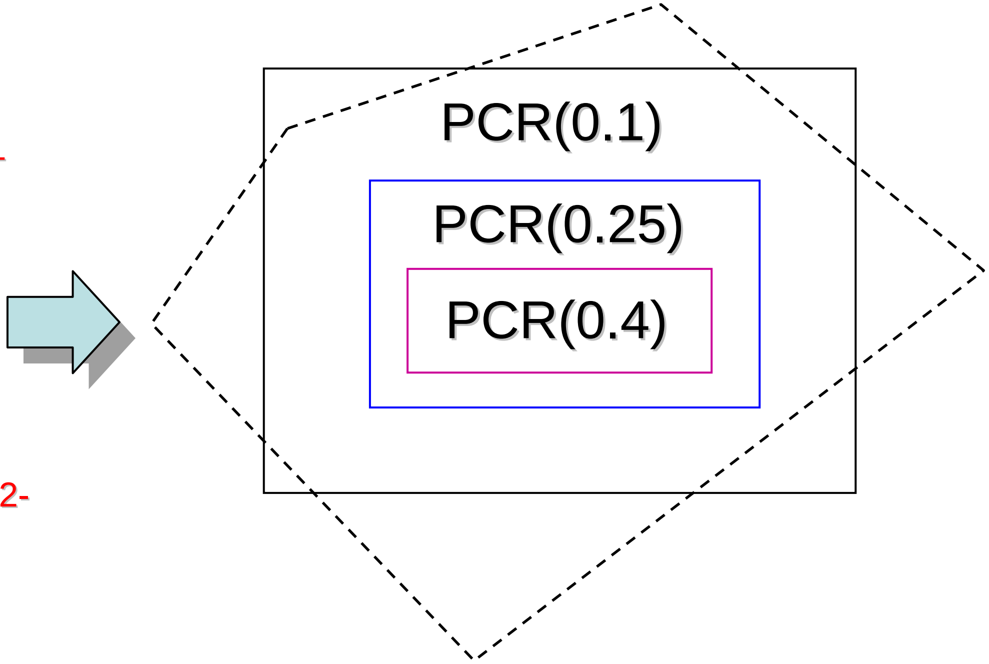
- Pr(O is in q)>τ, since *q* fully covers the part of O on the right side of L_{1+}, whose probability is p=0.2
- Thus, O can be validated

U-Tree: What to Store?

- Probabilistic constraint region (PCR)
- Select $0<p_1<\ldots<p_m<0.5$, and compute $PCR(p_1),\ldots,PCR(p_m)$



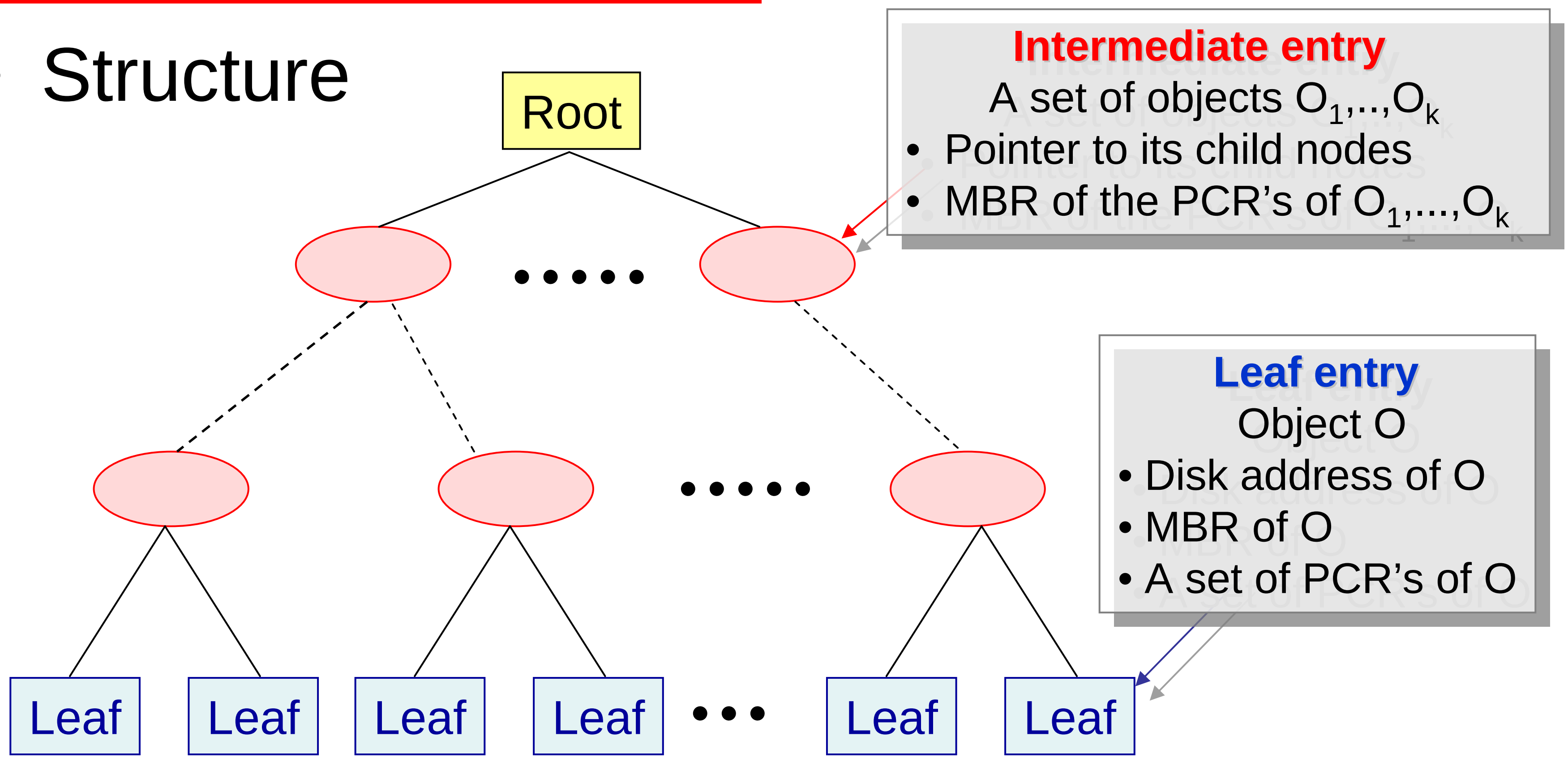
Uncertain object O



Information stored in a U-tree

U-Tree

- Structure



- Query evaluation

Probabilistic Categorical Data

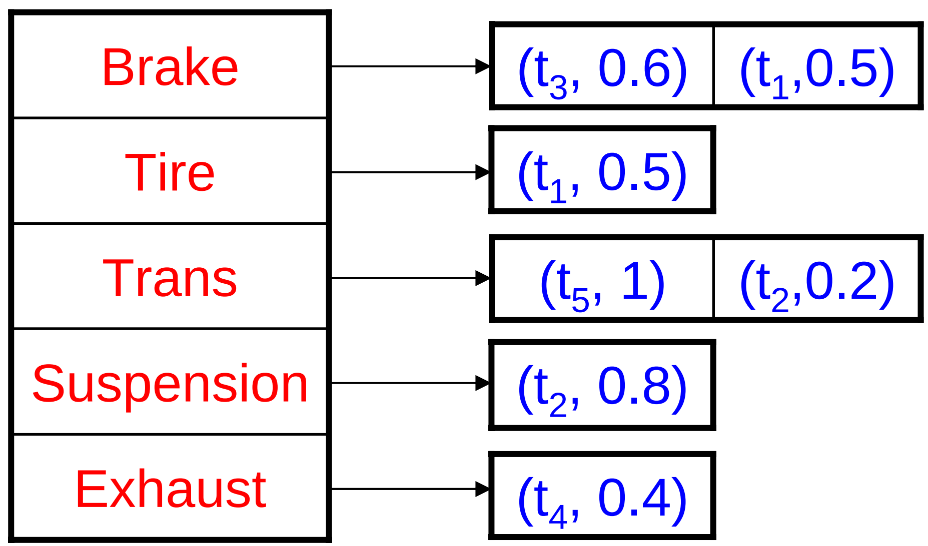
- Uncertain attribute "Problem" : derived from a text classifier
- Probabilistic threshold queries:
  - Find the tuples whose problem is "Brake" with probability 0.3 and "Tires" with probability 0.7
  - q={(Brake,0.3),(Tire,0.7)}
  - $Pr(t_1.Problem=q)=0.3\times0.5+0.5\times0.7=0.5$

	Make	Location	Date	Text	Problem
t_1	Explorer	WA	2/3/06	$\cdots$	{(Brake, 0.5), (Tires, 0.5)}
t_2	Camry	CA	3/5/05	$\cdots$	{(Trans, 0.2, (Suspension, 0.8)}
t_3	Civic	TX	10/2/06	$\cdots$	{(Exhaust, 0.4), (Brake, 0.6)}
t_4	Caravan	IN	7/2/06	$\cdots$	{(Trans, 1.0)}

Probabilistic Inverted Index

	Make	Location	Date	Text	Problem
t_1	Explorer	WA	2/3/06	···	{(Brake, 0.5), (Tires, 0.5)}
t_2	Camry	CA	3/5/05	···	{(Trans, 0.2, (Suspension, 0.8)}
t_3	Civic	TX	10/2/06	···	{(Exhaust, 0.4), (Brake, 0.6)}
t_4	Caravan	IN	7/2/06	···	{(Trans, 1.0)}

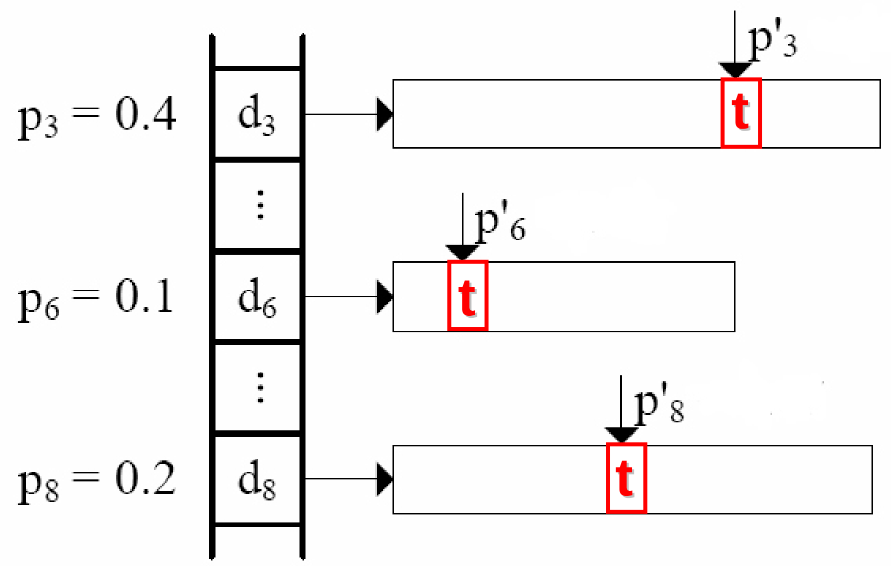
A list of
domain element



In probability descending order

Query Answering

- On Attribute A, a query $q=\{(d_3,0.4),(d_6,0.1),(d_8,0.2)\}$, $\tau=0.3$
  - $Pr(q=t.A) = p'_3 \times 0.4 + p'_6 \times 0.1 + p'_8 \times 0.2$
- Column pruning
  - If for each $d_i \in t.A$, $Pr(d_i) < \tau$, then t can be pruned



- Row pruning
  - If t only contains d_6 and d_8 whose probability is smaller than τ, then the t can be pruned

Ranking Queries

- Find the top-2 sensors with highest temperature
  - Certain data: answer = {R1, R2}
  - Uncertain data
    - R1 and R2 may not co-exist in a possible world
    - In different possible worlds, the answers are different

RID	Loc.	Time	Sensor-id	Temperature	Conf.
$R1$	A	6/2/06 2:14	$S101$	25	0.3
$R2$	B	7/3/06 4:07	$S206$	21	0.4
$R3$	B	7/3/06 4:09	$S231$	13	0.5
$R4$	A	4/12/06 20:32	$S101$	12	1.0
$R5$	E	3/13/06 22:31	$S063$	17	0.8
$R6$	E	3/13/06 22:28	$S732$	11	0.2

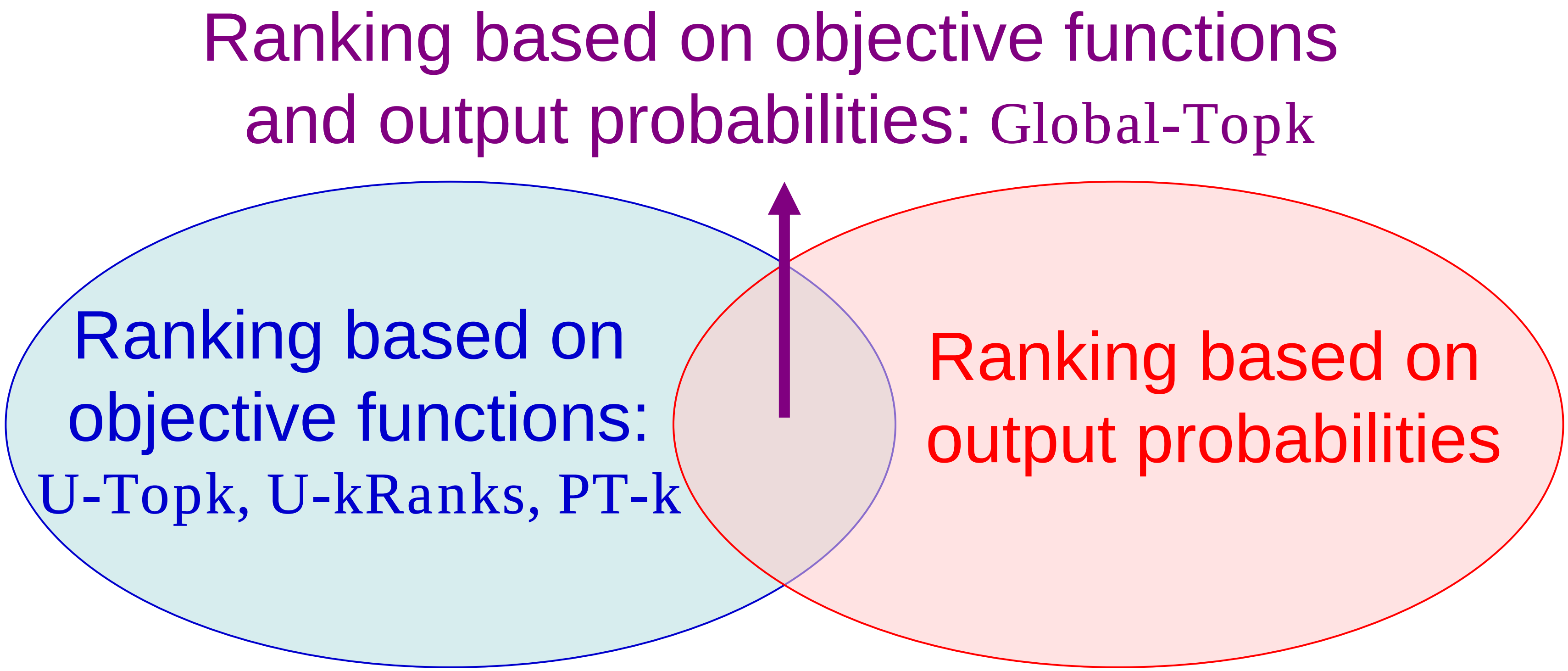
$R2 \oplus R3$ $R5 \oplus R6$

Challenges

- What does a probabilistic ranking query mean?
  - A ranking query on certain data returns the best k results in the ranking function
  - Ranking queries on uncertain data may be formulated differently to address different application interests
- How can a ranking query be answered efficiently?
  - Answering ranking queries on probabilistic databases can be very costly when the number of possible worlds is huge

Query Types

- How are tuples ranked?



Ranking Based on Objective Functions

- A scoring function is given
  - Rank the sensors in temperature descending order and select the top-2 results

$$R1 \prec R2 \prec R5 \prec R3 \prec R4 \prec R1$$

- How should the top-2 ranking results be captured?

RID	Loc.	Time	Sensor-id	Temperature	Conf.
$R1$	A	6/2/06 2:14	$S101$	25	0.3
$R2$	B	7/3/06 4:07	$S206$	21	0.4
$R3$	B	7/3/06 4:09	$S231$	13	0.5
$R4$	A	4/12/06 20:32	$S101$	12	1.0
$R5$	E	3/13/06 22:31	$S063$	17	0.8
$R6$	E	3/13/06 22:28	$S732$	11	0.2

$R2 \oplus R3 \quad R5 \oplus R6$

U-Topk Queries

- Find the most probable top-2 list in possible worlds
  - ⟨R1,R2⟩: p=0.12
  - ⟨R1,R5⟩: p=0.144
  - ⟨R1,R3⟩: p=0.03
  - ⟨R1,R4⟩: p=0.006
  - ⟨R2,R5⟩: p=0.224
  - ⟨R2,R4⟩: p=0.056
  - ⟨R5,R3⟩: p=0.28
  - ⟨R3,R4⟩: p=0.07
  - ⟨R5,R4⟩: p=0.056
  - ⟨R4,R6⟩: p=0.014
- Answer: ⟨R5,R3⟩

Possible world	Probability	Top-2 on Temperature
$W1 = \{R1, R2, R4, R5\}$	0.096	$R1, R2$
$W2 = \{R1, R2, R4, R6\}$	0.024	$R1, R2$
$W3 = \{R1, R3, R4, R5\}$	0.12	$R1, R5$
$W4 = \{R1, R3, R4, R6\}$	0.03	$R1, R3$
$W5 = \{R1, R4, R5\}$	0.024	$R1, R5$
$W6 = \{R1, R4, R6\}$	0.006	$R1, R4$
$W7 = \{R2, R4, R5\}$	0.224	$R2, R5$
$W8 = \{R2, R4, R6\}$	0.056	$R2, R4$
$W9 = \{R3, R4, R5\}$	0.28	$R5, R3$
$W10 = \{R3, R4, R6\}$	0.07	$R3, R4$
$W11 = \{R4, R5\}$	0.056	$R5, R4$
$W12 = \{R4, R6\}$	0.014	$R4, R6$

U-kRanks Queries

- Find the tuple of the highest probability at each ranking position
  - The 1st position
    - R1: p=0.3
    - R2: p=0.28
    - R5: p=0.336
    - R3: p=0.07
    - R4: p=0.014
  - The 2nd position
    - R5: p=0.368
- Answer: ⟨R5,R5⟩

Possible world	Probability	Top-2 on Temperature	
$W1 = \{R1, R2, R4, R5\}$	0.096	$R1$	$R2$
$W2 = \{R1, R2, R4, R6\}$	0.024	$R1$	$R2$
$W3 = \{R1, R3, R4, R5\}$	0.12	$R1$	$R5$
$W4 = \{R1, R3, R4, R6\}$	0.03	$R1$	$R3$
$W5 = \{R1, R4, R5\}$	0.024	$R1$	$R5$
$W6 = \{R1, R4, R6\}$	0.006	$R1$	$R4$
$W7 = \{R2, R4, R5\}$	0.224	$R2$	$R5$
$W8 = \{R2, R4, R6\}$	0.056	$R2$	$R4$
$W9 = \{R3, R4, R5\}$	0.28	$R5$	$R3$
$W10 = \{R3, R4, R6\}$	0.07	$R3$	$R4$
$W11 = \{R4, R5\}$	0.056	$R5$	$R4$
$W12 = \{R4, R6\}$	0.014	$R4$	$R6$

PT-k Queries

- Find the tuples whose probabilities to be in the top-2 list are at least p (p=0.35)
  - R1: p=0.3
  - R2: p=0.4
  - R3: p=0.38
  - R4: p=0.202
  - R5: p=0.704
  - R6: p=0.014
- Answer: {R2,R3,R5}

Possible world	Probability	Top-2 on Temperature
$W1 = \{R1, R2, R4, R5\}$	0.096	$R1, R2$
$W2 = \{R1, R2, R4, R6\}$	0.024	$R1, R2$
$W3 = \{R1, R3, R4, R5\}$	0.12	$R1, R5$
$W4 = \{R1, R3, R4, R6\}$	0.03	$R1, R3$
$W5 = \{R1, R4, R5\}$	0.024	$R1, R5$
$W6 = \{R1, R4, R6\}$	0.006	$R1, R4$
$W7 = \{R2, R4, R5\}$	0.224	$R2, R5$
$W8 = \{R2, R4, R6\}$	0.056	$R2, R4$
$W9 = \{R3, R4, R5\}$	0.28	$R5, R3$
$W10 = \{R3, R4, R6\}$	0.07	$R3, R4$
$W11 = \{R4, R5\}$	0.056	$R5, R4$
$W12 = \{R4, R6\}$	0.014	$R4, R6$

Global-Topk

- Find the top-2 tuples whose probabilities to be in the top-2 list are the highest
- Ranking based on objective functions and output probabilities
- Example
  - R1: p=0.3
  - R2: p=0.4
  - R3: p=0.38
  - R4: p=0.202
  - R5: p=0.704
  - R6: p=0.014
- Answer={R5,R2}

Possible world	Probability	Top-2 on Temperature
$W1 = \{R1, R2, R4, R5\}$	0.096	$R1, R2$
$W2 = \{R1, R2, R4, R6\}$	0.024	$R1, R2$
$W3 = \{R1, R3, R4, R5\}$	0.12	$R1, R5$
$W4 = \{R1, R3, R4, R6\}$	0.03	$R1, R3$
$W5 = \{R1, R4, R5\}$	0.024	$R1, R5$
$W6 = \{R1, R4, R6\}$	0.006	$R1, R4$
$W7 = \{R2, R4, R5\}$	0.224	$R2, R5$
$W8 = \{R2, R4, R6\}$	0.056	$R2, R4$
$W9 = \{R3, R4, R5\}$	0.28	$R5, R3$
$W10 = \{R3, R4, R6\}$	0.07	$R3, R4$
$W11 = \{R4, R5\}$	0.056	$R5, R4$
$W12 = \{R4, R6\}$	0.014	$R4, R6$

Query Answering Methods

- The dominant set property
  - For any tuple t, whether t is in the answer set only depends on the tuples ranked higher than t
  - The dominant set of t is the subset of tuples in T that are ranked higher than t
    - E.g. the dominant set of R3 is S_{R3}={R1,R2,R5}
- Framework of Query Answering Methods
  - Retrieve tuples in the ranking order
  - Evaluate each tuple based on its dominant set

Ranked tuples:

Temperature	25	21	17	13	12	11
RID	$R1$	$R2$	$R5$	$R3$	$R4$	$R6$

Answering PT-k Queries

- Position probability $Pr(t_i,j)$
  - The probability that t_i is ranked at the j-th position
  - E.g. $Pr(R3,2)=Pr(R3)\times Pr(S_{R3},1)$

Ranked tuples:

Temperature	25	21	17	13	12	11
RID	$R1$	$R2$	$R5$	$R3$	$R4$	$R6$

R3 is ranked 2^{nd}, if R3 appears, and 1 tuple in S_{R3} appears

- Generally: $$\Pr(t_i, j) = \Pr(t_i) \times \Pr(S_{t_i}, j-1)$$

Answering PT-k Queries

- Subset probability $\Pr(S_{ti},j)$
  - The probability that j tuples appear in S_{ti}
  - E.g. $S_{R3}=\{R5\} \cup S_{R5}$
  - $\Pr(S_{R3},2)=\Pr(R5) \times \Pr(S_{R5},1)+(1-\Pr(R5)) \times \Pr(S_{R5},2)$

Temperature	25	21	17	13	12	11
RID	$R1$	$R2$	$R5$	$R3$	$R4$	$R6$

2 tuples appear in S_{R3}, if $\begin{cases} \text{R5 appears, 1 tuple appears in } S_{R5} \\ \text{R5 does not appear, 2 tuples appear in } S_{R5} \end{cases}$

- Generally (Poisson Binomial Recurrence):

$$\Pr(S_{t_i}, j) = \Pr(t_i) \times \Pr(S_{t_{i-1}}, j-1) + (1-\Pr(t_i)) \times \Pr(S_{t_{i-1}}, j)$$

Summary of Query Answering Methods

- Optimal algorithms for U-Topk and U-kRanks queries in terms of the number of accessed tuples (Soliman *et al*. ICDE'07)
- Query answering algorithms for U-Topk and U-kRanks queries based on Poisson binomial recurrence (Yi *et al*. ICDE'08)
- Spatial and probabilistic pruning techniques for U-kRanks queries (Lian and Chen, EDBT'08)
- Efficient query answering algorithms and pruning techniques for PT-k queries (Hua *et al*. ICDE'08, SIGMOD'08)
- A sampling-based method (Silberstein *et al*. ICDE'06)

Ranking Based on Output Probabilities

- Query Q: find the average temperature of all sensors
- Ranking: find the top-2 results with the highest probabilities of being the answers to Q (output probabilities)
  - Answer: 14 (p=0.28), 16.67 (p=0.224)

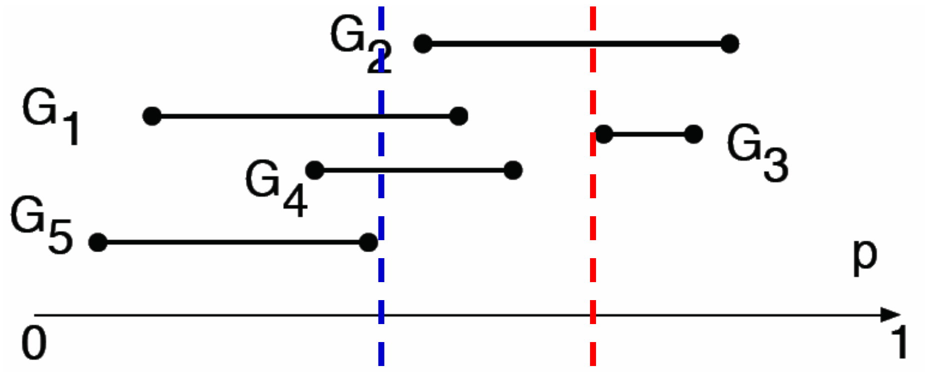
Possible world	Probability	Average temperature
$W1 = \{R1, R2, R4, R5\}$	0.096	18.75
$W2 = \{R1, R2, R4, R6\}$	0.024	17.25
$W3 = \{R1, R3, R4, R5\}$	0.12	16.75
$W4 = \{R1, R3, R4, R6\}$	0.03	15.25
$W5 = \{R1, R4, R5\}$	0.024	18
$W6 = \{R1, R4, R6\}$	0.006	16
$W7 = \{R2, R4, R5\}$	0.224	16.67
$W8 = \{R2, R4, R6\}$	0.056	14.67
$W9 = \{R3, R4, R5\}$	0.28	14
$W10 = \{R3, R4, R6\}$	0.07	12
$W11 = \{R4, R5\}$	0.056	14.5
$W12 = \{R4, R6\}$	0.014	11.5

Query Answering

- Monte Carlo Simulation (1 step)
  - Choose a possible world at random, and evaluate the query
  - Record the answer to the query and its frequency
- For example, if we run 100 steps of Monte Carlo simulation, and “14” is the answer in 30 steps
  - The output probability of “14” can be approximated by 30/100=0.3, with an error bound ε
  - The output probability of “14” lies in the probability interval [0.3-ε, 0.3+ε]
  - The more steps of Monte Carlo simulation we run, the smaller probability intervals we can get

Query Answering (cont.)

- The simulation stops when the top-k output probabilities and their relative ranks are clear
  - E.g. There are 5 possible results G1, G2, G3, G4 and G5. After a few steps of Monte Carlo simulation, the output probability interval of each result is shown below
  - G3's output probability is in top-2. The other answer might be one of G1, G2, and G4 . But G5's output probability cannot be in top-2



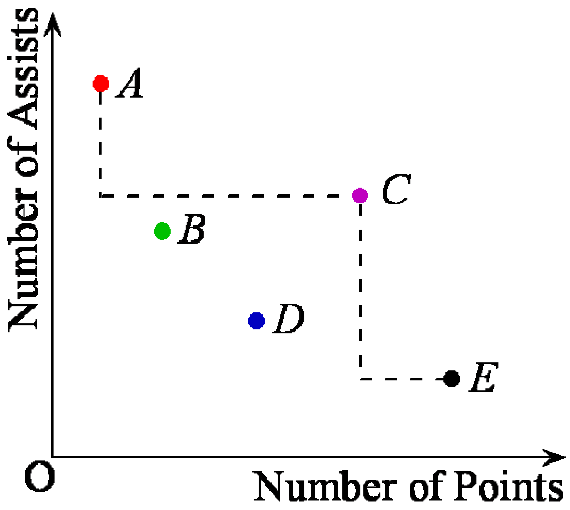
Re *et al*. ICDE'07

More on Monte Carlo Simulation

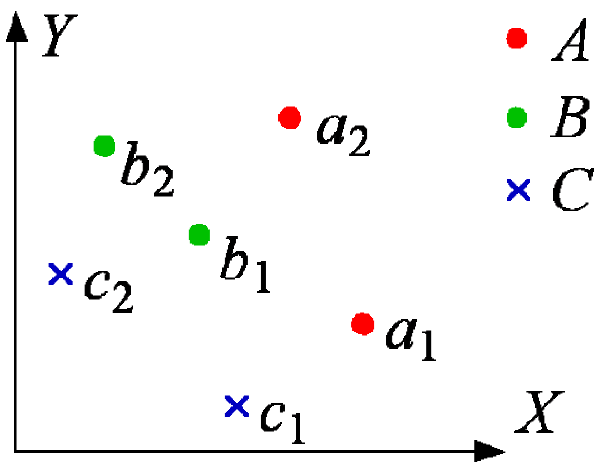
- Separate data schema and uncertain variables
  - Data schema is certain
  - Use random variables supported by variable generation (VG) functions to simulate uncertainty
- A naïve implementation: run Monte Carlo simulation until the result is stable
- Efficient Implementation
  - Run N times of Monte Carlo simulation once in batch
  - Delay random attribute materialization as long as possible
  - Reproduce values for random attributes when necessary
- Details in [Jampani et al., SIGMOD'08]

Probabilistic Skyline

- Probabilistic skylines
  - An instance has a probability to represent the object
  - An object has a probability to be in the skyline

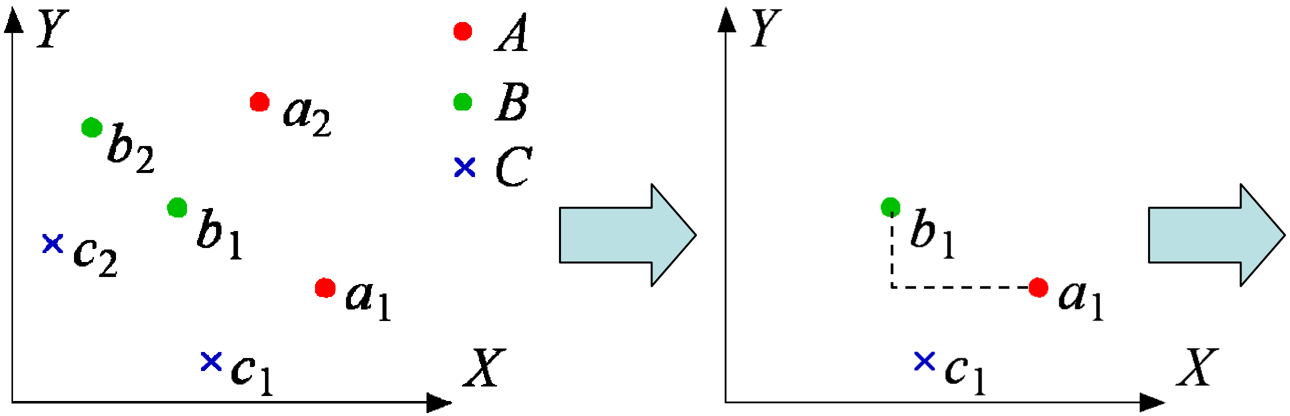


Skyline on certain objects



A set of uncertain objects

Skyline Probabilities



SKY($\{a_1,b_1,c_1\}$)=$\{a_1,b_1\}$

A set of uncertain objects

A possible world $\{a_1,b_1,c_1\}$

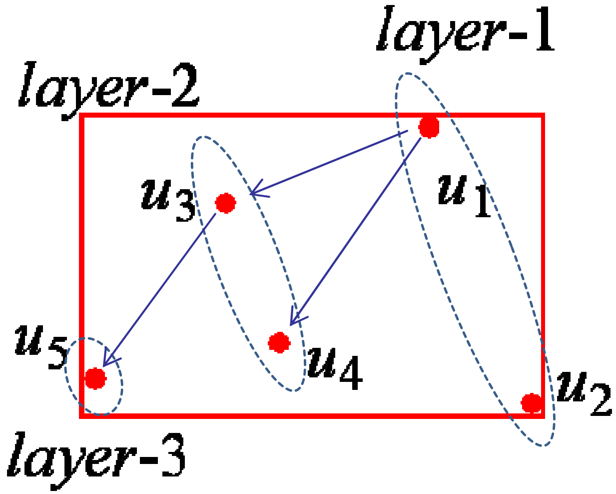
- Skyline probability
  - B is in the skyline of possible worlds $w_1=\{a_1,b_1,c_1\}$, $w_2=\{a_1,b_1,c_2\}$, $w_3=\{a_1,b_2,c_1\}$, and $w_4=\{a_1,b_2,c_2\}$
  - Thus, $Pr(B) = Pr(w_1)+Pr(w_2)+Pr(W_3)+Pr(W_4) = 4 \times 0.125 = 0.5$
- p-skyline = $\{ U \mid Pr(U) \geq p \}$ for a given threshold p

Probabilistic Skyline Computation

- Iteration: Bounding-Pruning-Refining
- Bounding
  - Bound $Pr(u)$: lower bound $Pr^{-}(u)$ and upper bound $Pr^{+}(u)$
  - Bound $Pr(U)$: $Pr(U) = \frac{1}{|U|}\sum_{u \in U} Pr(u)$
- Pruning
  - In p-skyline if lower bound $Pr^{-}(U) \geq p$
  - Not in p-skyline if upper bound $Pr^{+}(U) < p$
- Refining
  - Bottom-up method
  - Top-down method

Bottom-up Method

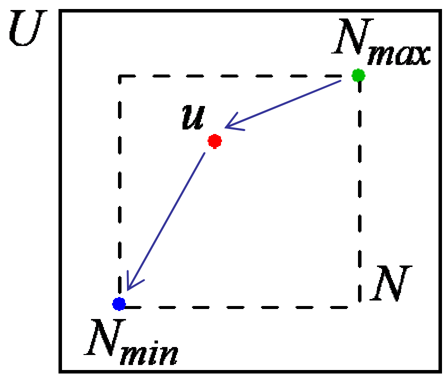
- Key Idea
  - Two instances u_1 and $u_2 \in U$, if u_1 dominants u_2, then $Pr(u_1) \geq Pr(u_2)$
- The layered structure
  - Sort the instances of an object according to the dominance relation



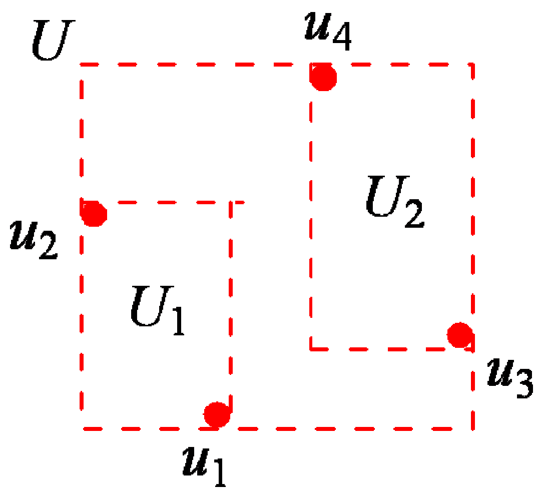
- Bounding
  - $\max\{Pr(u_1), Pr(u2)\} \geq \max\{Pr(u3), Pr(u4)\} \geq Pr(u5)$

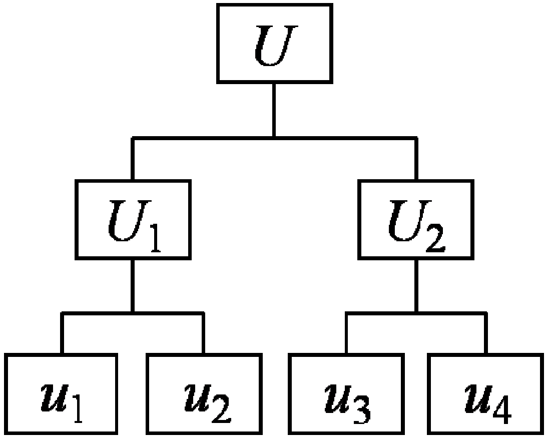
Top-down Method

- Bounding
  - Using the lower corner and upper corner to bound the skyline probability
  - $Pr(N_{min}) \leq Pr(u) \leq Pr(N_{max})$



- Iterative partitioning: binary tree





Outline

- Uncertainty and uncertain data, where and why?
- Models for uncertain and probabilistic data
- (coffee break)
- OLAP on uncertain and probabilistic data
- Mining uncertain and probabilistic data
- Tools: querying uncertain and probabilistic data
  - Indexing uncertain and probabilistic data
  - Ranking queries and spatial queries
- Summary and discussion

Summary

- Uncertain data becomes more and more important and prevalent
  - Critical applications: sensor networks, location-based services, web applications, user preferences, health-informatics, …
- Modeling uncertain data
  - Model uncertainty at various levels
  - Model correlation among data entries
- OLAP on uncertain data
- Mining uncertain data
- Tools: querying uncertain data
  - Simple queries, ranking queries, spatial queries
  - Using indexes to speed up query answering

Can Uncertainty Be Beneficial?

- In all the cases discussed so far, uncertainty leads to more complicated processing ☹
- Uncertainty and privacy preservation
  - Privacy preservation – preventing individuals from being re-identified, while keeping the aggregate data useful
  - Major approaches: perturbation and generalization – making data uncertain!
- [Aggarwal, ICDE'08]

Thank You

Future is uncertain because it will be what we make it.

– Immanuel Wallerstein

References(1)

- Charu C. Aggarwal. On density based transforms for uncertain data mining. In ICDE, 2007.
- Charu C. Aggarwal. On unifying privacy and uncertain data models. In ICDE, 2007.
- Charu C. Aggarwal and Philip. S. Yu. Outlier detection with uncertain data. In SDM, pages 483-493, 2008.
- Parag Agrawal, Omar Benjelloun, Anish Das Sarma, Chris Hayworth, Shubha U. Nabar, Tomoe Sugihara, and Jennifer Widom. Trio: A system for data, uncertainty, and lineage. In VLDB, pages 1151–1154, 2006.
- Parag Agrawal and Jennifer Widom. Confidence-aware joins in large uncertain databases. Technical report, Stanford University CA, USA.
- Lyublena Antova, Thomas Jansen, Christoph Koch, and Dan Olteanu. Fast and simple relational processing of uncertain data. In ICDE, pages 983-992, 2008.

References(2)

- Lyublena Antova, Christoph Koch, and Dan Olteanu. 10^10^6 worlds and beyond: Efficient representation and processing of incomplete information. In ICDE, pages 606–615, 2007.
- Daniel Barbará, Hector Garcia-Molina, and Daryl Porter. The management of probabilistic data. IEEE Trans. Knowl. Data Eng., 4(5):487–502, 1992.
- Omar Benjelloun, Anish Das Sarma, Chris Hayworth, and Jennifer Widom. An introduction to uldbs and the trio system. IEEE Data Eng. Bull., 29(1):5–16, 2006.
- Christian Böhm, Alexey Pryakhin, and Matthias Schubert. The gauss-tree: Efficient object identification in databases of probabilistic feature vectors. In ICDE, page 9, 2006.
- Douglas Burdick, Prasad Deshpande, T. S. Jayram, Raghu Ramakrishnan, and Shivakumar Vaithyanathan. Olap over uncertain and imprecise data. In VLDB, pages 970–981, 2005.

References(3)

- Douglas Burdick, Prasad M. Deshpande, T. S. Jayram, Raghu Ramakrishnan, and Shivakumar Vaithyanathan. Efficient allocation algorithms for olap over imprecise data. In VLDB, pages 391–402, 2006.
- Doug Burdick, AnHai Doan, Raghu Ramakrishnan, and Shivakumar Vaithyanathan. Olap over imprecise data with domain constraints. In VLDB, pages 39–50, 2007.
- Roger Cavallo and Michael Pittarelli. The theory of probabilistic databases. In VLDB, pages 71–81, 1987.
- Reynold Cheng, Dmitri V. Kalashnikov, and Sunil Prabhakar. Evaluating probabilistic queries over imprecise data. In SIGMOD Conference, pages 551–562, 2003.
- Reynold Cheng, Dmitri V. Kalashnikov, and Sunil Prabhakar. Querying imprecise data in moving object environments. IEEE Trans. Knowl. Data Eng., 16(9):1112–1127, 2004.
- Reynold Cheng, Jinchuan Chen, Mohamed Mokbel, and Chi-Yin Chow. Probabilistic verifiers: evaluating constrained nearest-neighbor queries over uncertain data. In ICDE, pages 973-982, 2008.

References(4)

- Reynold Cheng, Sarvjeet Singh, Sunil Prabhakar, Rahul Shah, Jeffrey Scott Vitter, and Yuni Xia. Efficient join processing over uncertain data. In CIKM, pages 738–747, 2006.
- Reynold Cheng, Yuni Xia, Sunil Prabhakar, Rahul Shah, and Jeffrey Scott Vitter. Efficient indexing methods for probabilistic threshold queries over uncertain data. In VLDB, pages 876–887, 2004.
- Chun-Kit Chui, Ben Kao, and Edward Hung. Mining frequent itemsets from uncertain data. In PAKDD, pages 47-58, 2007.
- Chun-Kit Chui and Ben Kao. A decremental approach for mining frequent itemsets from uncertain data. In PAKDD, pages 64-75, 2008.
- Graham Cormode and Andrew McGregor. Approximation algorithms for clustering uncertain data. In PODS, pages 191-199, 2008.
- Xiangyuan Dai, Man Lung Yiu, Nikos Mamoulis, Yufei Tao, and Michail Vaitis. Probabilistic spatial queries on existentially uncertain data. In SSTD, pages 400-417, 2005.
- Nilesh N. Dalvi and Dan Suciu. Efficient query evaluation on probabilistic databases. In VLDB, pages 864–875, 2004.

References(5)

- Nilesh N. Dalvi and Dan Suciu. Answering queries from statistics and probabilistic views. In VLDB, pages 805–816, 2005.
- Nilesh N. Dalvi and Dan Suciu. The dichotomy of conjunctive queries on probabilistic structures. In PODS, pages 293–302, 2007.
- A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39(1): 1-38, 1977.
- Lise Getoor. An introduction to probabilistic graphical models for relational data. IEEE Data Eng. Bull., 29(1):32–39, 2006.
- Todd J. Green and Val Tannen. Models for incomplete and probabilistic information. IEEE Data Eng. Bull., 29(1):17–24, 2006.
- Ming Hua, Jian Pei, Wenjie Zhang, and Xuemin Lin. Efficiently answering probabilistic threshold top-k queries on uncertain data (extended abstract). In ICDE, pages 1403–1405, 2008.
- Ming Hua, Jian Pei, Wenjie Zhang, and Xuemin Lin. Ranking queries on uncertain data: A probabilistic threshold approach. In SIGMOD, Vancouver, Canada, 2008.

References(6)

- Ravi Jampani, Fei Xu, Mingxi Wu, Luis Leopoldo Peres, Christopher Jermaine, and Peter Haas. MCDB: A Monte Carlo Approach to Managing Uncertain Data. In SIGMOD, Vancouver, Canada, 2008.
- Benny Kimelfeld and Yehoshua Sagiv. Maximally joining probabilistic data. In PODS, pages 303–312, 2007.
- Hans-Peter Kriegel, Peter Kunath, Martin Pfeifle, Matthias Renz. Probabilistic Similarity Join on Uncertain Data. In DASFAA, pages 295-309, 2006.
- Hans-Peter Kriegel and Martin Pfeifle. Density-based clustering of uncertain data. In KDD, 2005.
- Hans-Peter Kriegel and Martin Pfeifle. Hierarchical density-based clustering of uncertain data. In ICDM, 2005.
- Xiang Lian and Lei Chen. Monochromatic and bichromatic reverse skyline search over uncertain databases. In SIGMOD, Vancouver, Canada, 2008.
- Xiang Lian and Lei Chen. Probabilistic ranked queries in uncertain databases. In EDBT, pages 511-522, 2008.

References(7)

- Vebjorn Ljosa and Ambuj K. Singh. Apla: Indexing arbitrary probability distributions. In ICDE, pages 946–955, 2007.
- Michi Mutsuzaki, Martin Theobald, Ander de Keijzer, Jennifer Widom, Parag Agrawal, Omar Benjelloun, Anish Das Sarma, Raghotham Murthy, and Tomoe Sugihara. Trio-one: Layering uncertainty and lineage on a conventional dbms (demo). In CIDR, pages 269–274, 2007.
- Jian Pei, Bin Jiang, Xuemin Lin, and Yidong Yuan. Probabilistic skylines on uncertain data. In VLDB, pages 15–26, 2007.
- Christopher Re, Nilesh N. Dalvi, and Dan Suciu. Query evaluation on probabilistic databases. IEEE Data Eng. Bull., 29(1):25–31, 2006.
- Christopher Re, Nilesh N. Dalvi, and Dan Suciu. Efficient top-k query evaluation on probabilistic data. In ICDE, pages 886–895, 2007.
- Christopher Re and Dan Suciu. Materialized views in probabilistic databases for information exchange and query optimization. In VLDB, pages 51–62, 2007.

References(8)

- A. Das Sarma, O. Benjelloun, A. Halevy, S.U. Nabar, and J. Widom. Representing uncertain data: Models, properties, and algorithms. Technical report, Stanford University CA, USA.
- Anish Das Sarma, Martin Theobald, and Jennifer Widom. Exploiting lineage for confidence computation in uncertain and probabilistic databases. Technical report, Stanford University CA, USA.
- Prithviraj Sen and Amol Deshpande. Representing and querying correlated tuples in probabilistic databases. In ICDE, page 596-605, 2007.
- Adam Silberstein Silberstein, Rebecca Braynard, Carla Ellis, Kamesh Munagala, and Jun Yang. A sampling-based approach to optimizing top-k queries in sensor networks. In ICDE, page 68, 2006.
- Sarvjeet Singh, Chris Mayfield, Sunil Prabhakar, Rahul Shah, and Susanne E. Hambrusch. Indexing uncertain categorical data. In ICDE, pages 616–625, 2007.

References (9)

- Sarvjeet Singh, Chris Mayfield, Rahul Shah, Sunil Prabhakar, Susanne E. Hambrusch, Jennifer Neville, and Reynold Cheng. Database support for probabilistic attributes and tuples. In ICDE, pages 1053-1061, 2008.
- Mohamed A. Soliman, Ihab F. Ilyas, and Kevin Chen-Chuan Chang. Top-k query processing in uncertain databases. In ICDE, pages 896–905, 2007.
- Yufei Tao, Reynold Cheng, Xiaokui Xiao, Wang Kay Ngai, Ben Kao, and Sunil Prabhakar. Indexing multi-dimensional uncertain data with arbitrary probability density functions. In VLDB, pages 922–933, 2005.
- Yufei Tao, Xiaokui Xiao, and Reynold Cheng. Range search on multidimensional uncertain data. ACM Trans. Database Syst., 32(3):15, 2007.
- Jennifer Widom. Trio: A system for integrated management of data, accuracy, and lineage. In CIDR, pages 262–276, 2005.
- Ke Yi, Feifei Li, Divesh Srivastava, and George Kollios. Efficient processing of top-k queries in uncertain databases. In ICDE, pages 1406–1408, 2008.

References (10)

- Xi Zhang and Jan Chomicki. On the semantics and evaluation of top-k queries in probabilistic databases. In ICDE Workshops, pages 556–563, 2008.
- Zhengdao Xu and Hans-Arno Jacobsen. Evaluating proximity relations under uncertainty. In ICDE, pages 876-885, 2007.
- Qin Zhang, Feifei Li, and Ke Yi. Finding frequent items in probabilistic data. In SIGMOD, 2008.