Dealing with Uncertain Data: Instances, Objects, Possible Worlds, and Probability Distributions

Jian Pei

Simon Fraser University

http://www.cs.sfu.ca/~jpei

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?"

Uncertainty in Data

Mobile object tracking

Sensor data processing

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

Social surveys

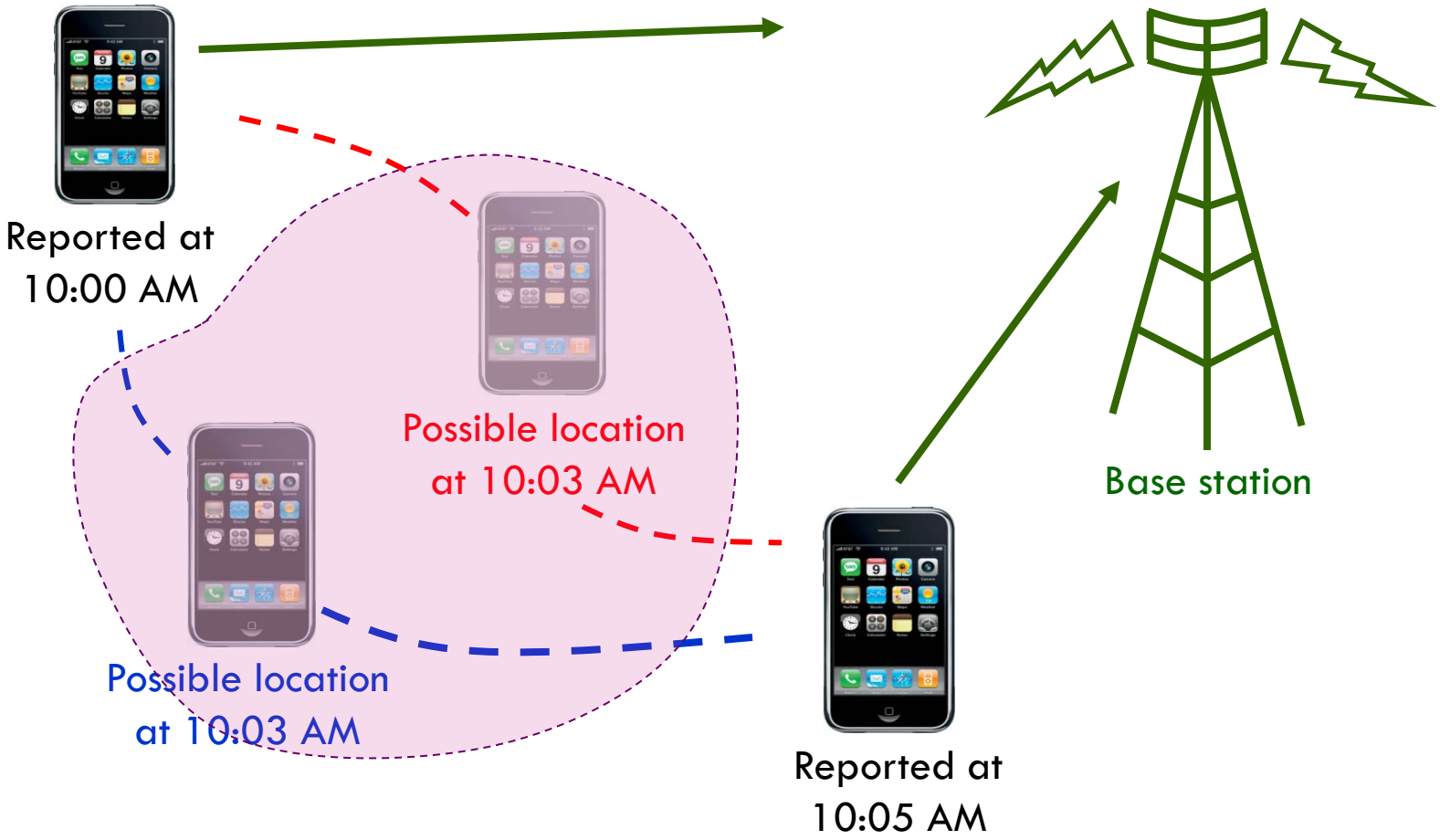
Applications

Uncertain Data

Simplifying uncertain data to certain data? It may not use the full potential of data

Probability: a new dimension
Probabilistic answers are interesting

Example


Reported at
10:00 AM
Possible location
at 10:03 AM
Possible location
at 10:03 AM
Base station
Reported at
10:05 AM

Related Work on Uncertain Data Analysis

Models

- Uncertain object model

  [Cheng *et al*. SIGMOD'03, Pei *et al*. VLDB'07, …]
- Probabilistic database model

  [Sarma *et al*. ICDE'06, Soliman *et al*. ICDE'06, …]
- Graphical models for uncertain data

  [Deshpande *et al*. Book Chapter 2009, ….]
- Possible worlds semantics

  [Abiteboul *et al*. SIGMOD'87,…]

Queries

- Skyline queries on uncertain data

  [Pei *et al*. VLDB'07, Atallah and Qi PODS'09, …]
- Indexing uncertain data

  [Kanagal *et al*. SIGMOD'09, Tao *et al*. VLDB'05, …]
- Ranking queries on uncertain data

  [Li *et al*. VLDB'09, Cormode *et al*.VLDB'09, …]
- …

Uncertain Data Analysis

Systems

- Trio [Stanford University]
- The Probabilistic Databases project

  [University of Maryland, College Park]
- The Probabilistic logics project

  [University of Maryland, College Park]
- HeisenData [University of California, Berkeley]
- Orion [Purdue University]

Other related study

- Typicality analysis in Psychology

  [Dubois et al. Journal of Intelligent Systems, 1991,…]
- Skyline analysis on deterministic data

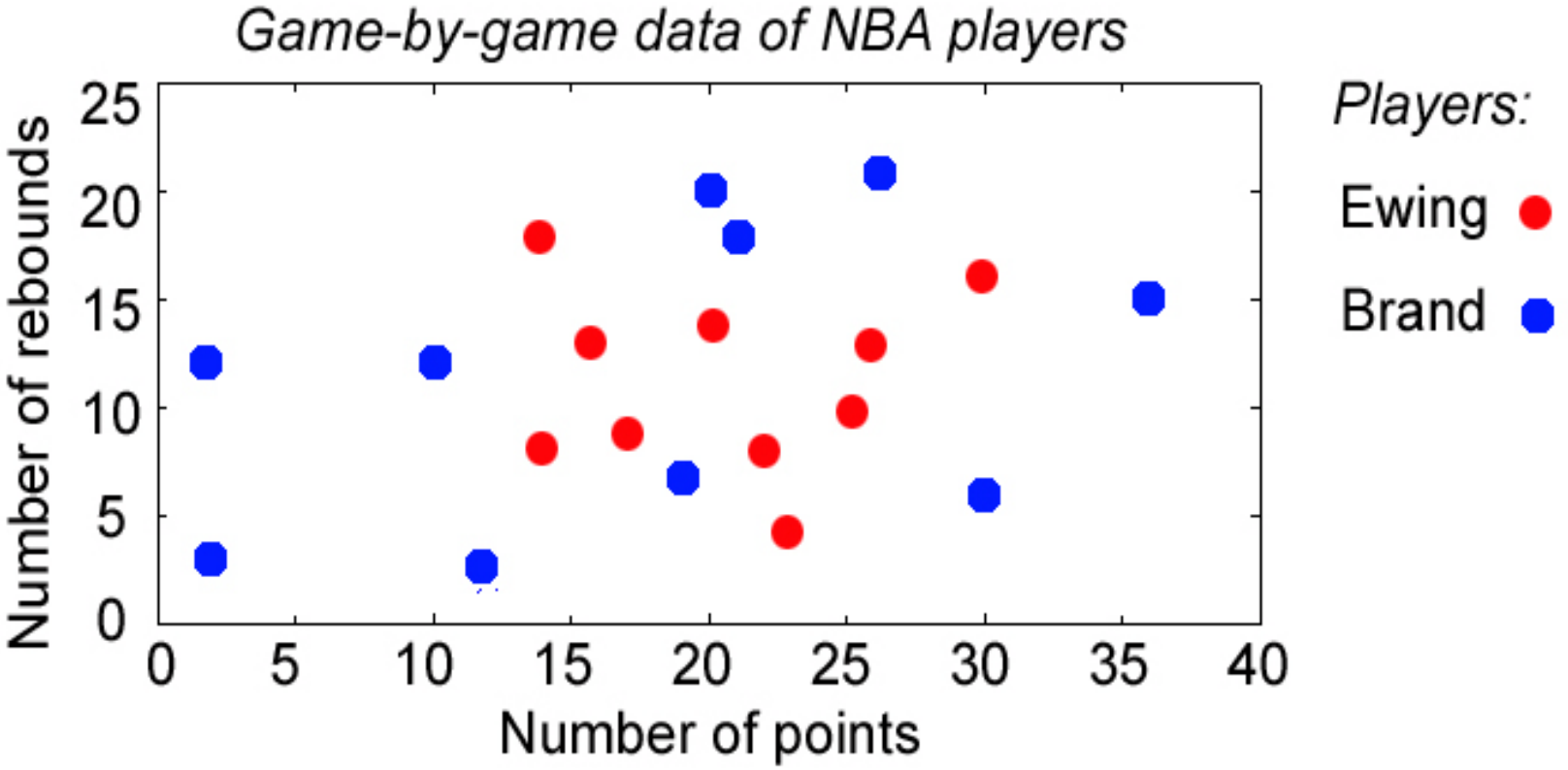
  [Borzsonyi *et al*. ICDE'01, Papadias *et al*. SIGMOD'03]
- Clustering analysis

  [Ester et al. KDD'96, Hartigan et al. Applied Statistics, 1979.]
- …

Models of Uncertain Data

- Uncertain Objects
  - An object is uncertain in a few dynamic attributes
  - Use a sample or a probability density function to capture the distribution
- Probabilistic database
  - The values of each tuple are certain
  - Each tuple carries an existing/membership probability
  - Generation rules: constraints specifying exclusive tuples

Uncertain Objects



Uncertain objects: NBA players

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

Possible Worlds [Abiteboul *et al*. SIGMOD'87]

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

Possible Worlds of Probabilistic Data

0.112=0.4×0.7×0.4×1.0

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

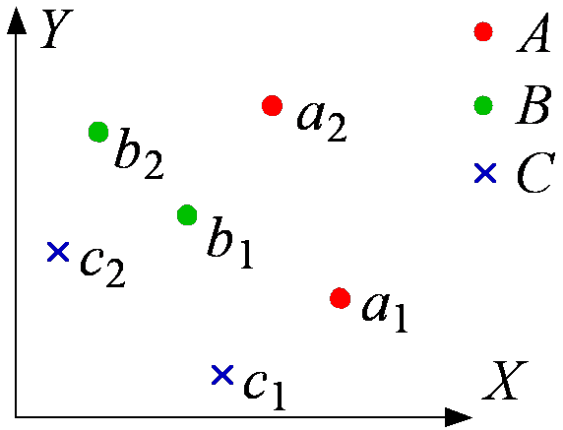
t2 and t3 never appear in the same possible world!

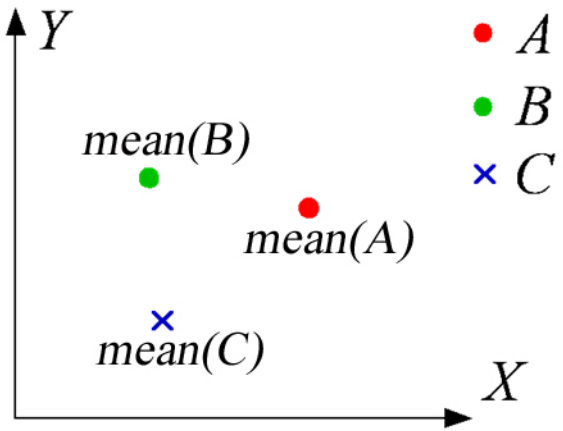
Yet Another Uncertain Data Paper

- Take a traditional database problem
  - Spatial database problems: nearest neighbor search, skyline, …
  - General database problems: ranking, join, …
- Apply on uncertain/probabilistic entries
  - A point → an uncertain object
  - A table → a probabilistic table
- Consider possible world semantics
  - Challenge: an exponential number of possible worlds
- Extend traditional queries to probability threshold queries
  - Objects with the highest probability, objects passing a probability threshold, …

Dominating Queries on Certain Data

- Conventional methods compute the skyline on
  - Individual game records
  - Aggregate: mean or median

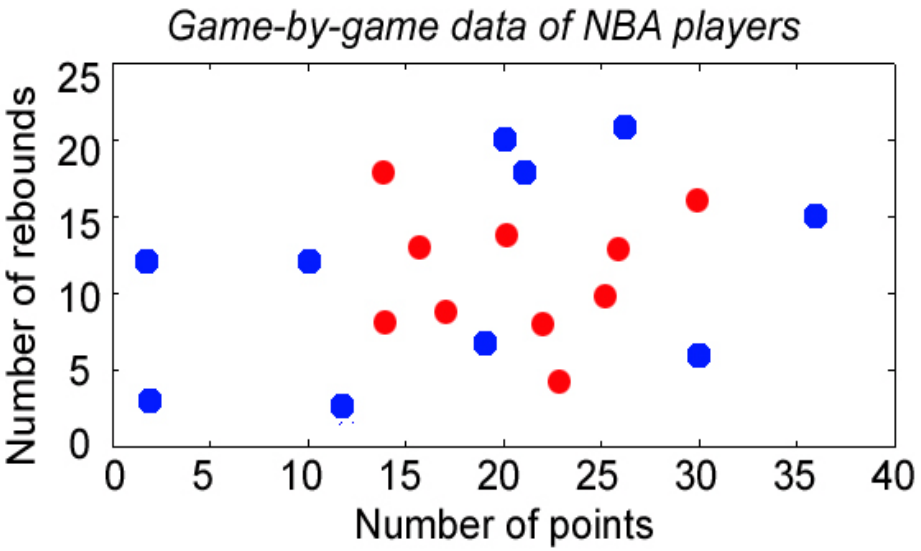


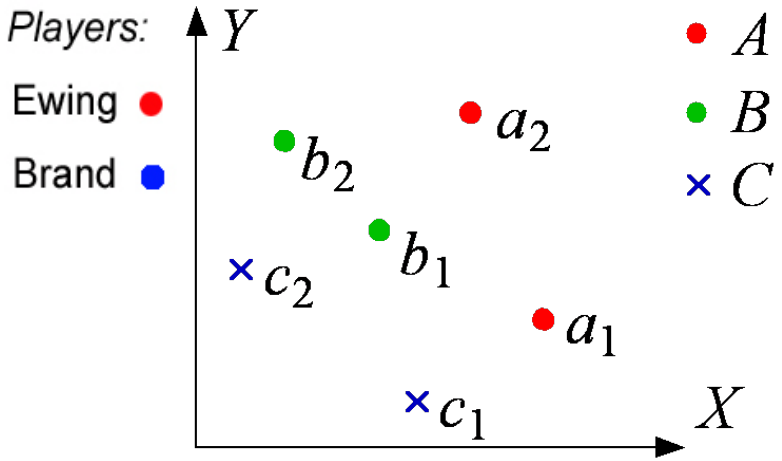


- Limitations
  - Aggregates may be misled by outliers
  - Data distribution is not captured

Probabilistic Skyline

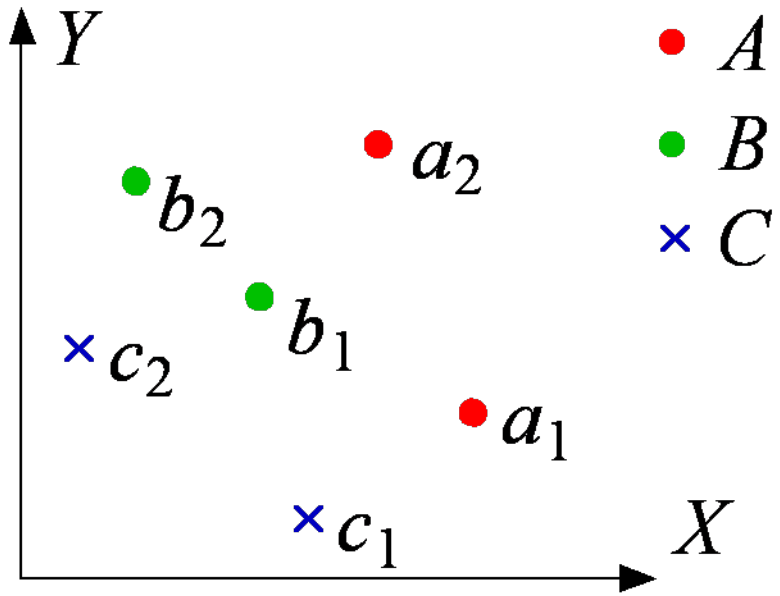
- Probabilistic skylines [Pei *et al*. VLDB'07]
  - An instance has a probability to represent the object
  - An object has a probability to be in the skyline





Skyline Probabilities

- Possible world: $W = \{a_i, b_j, c_k\}$ (i, j, k = 1 or 2)
  - $Pr(W) = 0.5 \times 0.5 \times 0.5 = 0.125$, $\sum_{W \in \Omega} Pr(W) = 1$
- $SKY(\{a_1, b_1, c_1\}) = \{a_1, b_1\}$
  - A and B are in $SKY(\{a_1, b_1, c_1\})$
- B is in the skyline of possible worlds $\{a_1, b_1, c_1\}$, $\{a_1, b_1, c_2\}$, $\{a_1, b_2, c_1\}$, and $\{a_1, b_2, c_2\}$
  - $Pr(B) = 4 \times 0.125 = 0.5$
- Pr(A) = 1, Pr(C) = 0

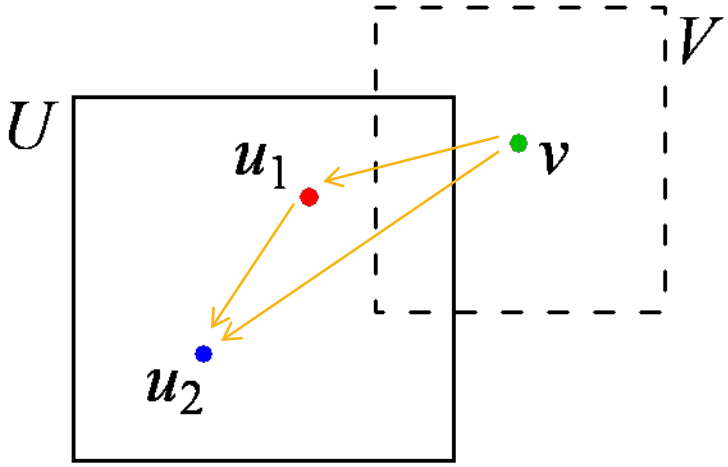


Probabilistic Skyline Computation

- p-skyline = {U | Pr(U) ≥ p} for a given threshold p
  - 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

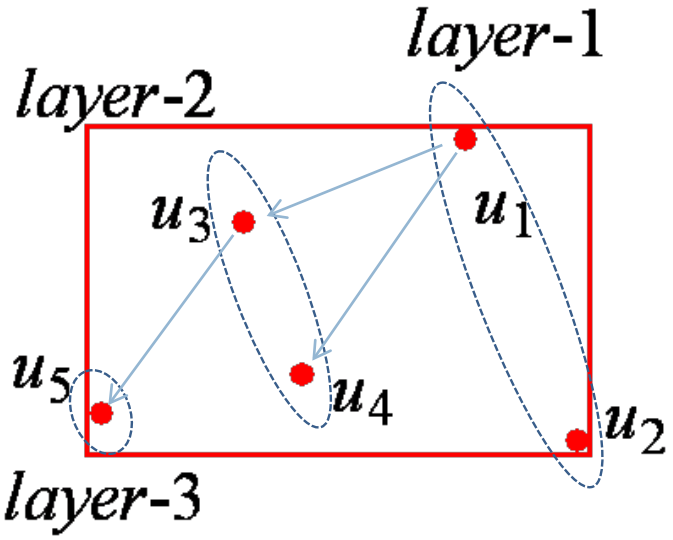
The Bottom-Up Method

- Key idea: sort the instances of an object according to the dominance relation such that their skyline probabilities are in descending order
- Two instances u_1 and $u_2 \in U$, if $u_1 \succ u_2$, then $Pr(u_1) \geq Pr(u_2)$



The Layer Structure

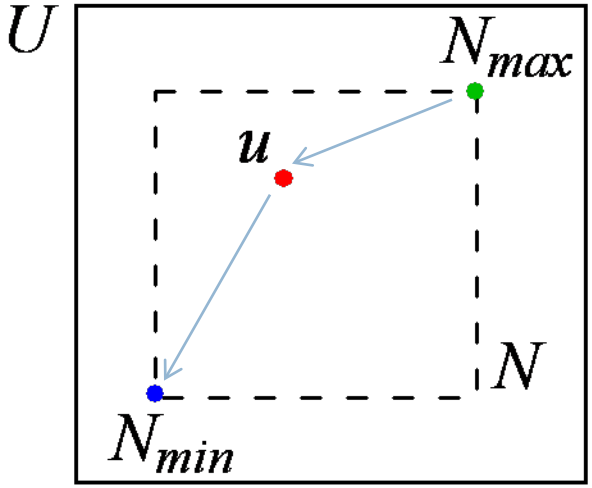
- layer-1: the skyline of all instances
- layer-k (k > 1): the skyline of instances except those at layer-1, …, layer-(k-1)



- $\forall$ u at layer-k, $\exists$ u' at layer-(k-1) such that u' $\succ$ u and Pr(u') $\geq$ Pr(u)
- max{Pr(u) | u is at layer-(k-1)} $\geq$ max{Pr(u) | u is at layer-k}
- Bounding
  - max{Pr(u1), Pr(u2)} $\geq$ max{Pr(u3), Pr(u4)} $\geq$ Pr(u5)

The Top-Down Method

- For instances u_1 and $u_2 \in U$, if $u_1 \succ u_2$, then $Pr(u_1) \geq Pr(u_2)$
  - N is a subset of instances of U, $\forall u \in N$, $Pr(N_{max}) \geq Pr(u) \geq Pr(N_{min})$



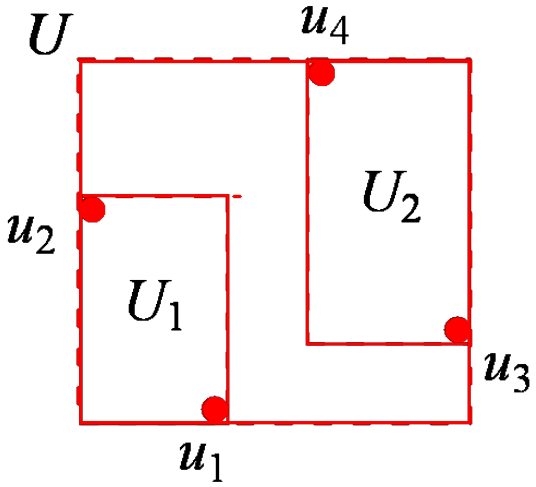
- Object U has k partitions N_1, …, N_k,

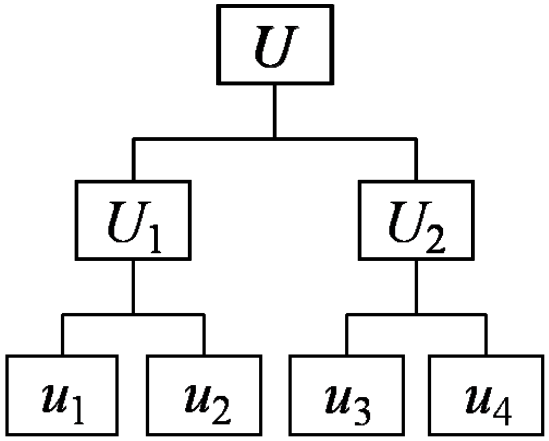
$$\frac{1}{|U|}\sum_{i=1}^{k}|N_i|\cdot Pr(N_{i,max}) \geq Pr(U) \geq \frac{1}{|U|}\sum_{i=1}^{k}|N_i|\cdot Pr(N_{i,min})$$

- Build a partition tree for each object to organize partitions

Partition Tree

- Binary tree





- Growing one level of the tree in each iteration
  - Choose one dimension in a round-robin fashion
  - Each leaf node is partitioned into two children nodes, each of which has half of instances
- Bound $Pr(N_{max})$ and $Pr(N_{min})$ of a partition N

Some General Techniques

- Enumerating possible world groups instead of possible worlds
  - All possible worlds having the same skyline instances form a group
  - An uncertain data search problem is in P if there are a polynomial number of groups needed to be considered
  - A good method considers as few groups as possible
- Examining groups in a good order
  - The more a group contributes to the answer, the earlier the group should be considered

So Simple?

- Subtlety in extending traditional problems to their probabilistic version
- Uncertainty at different levels

NN Search on Uncertain Data

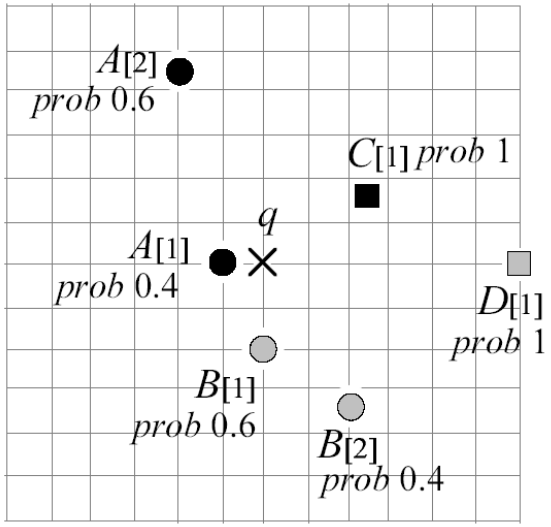
- In general, no object has absolute certainty to be the nearest neighbor of a query point
- Nearest Neighbor (NN) Probabilities
  - The NN probability of an uncertain object A is the probability that A is the nearest neighbor of query point q
- Anomalies
  - The largest NN probability can still trivial
  - Using a probability threshold may retrieve a large set of nearest neighbors
  - Wait a minute – how can I determine the probability threshold?

Superseding Nearest Neighbor Search

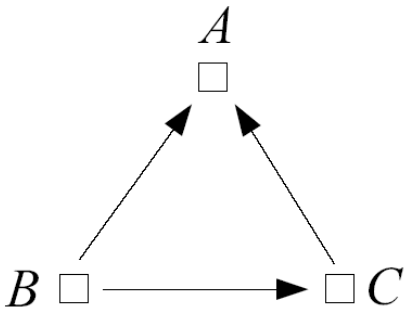
- Intuition: find a (very small) set of objects V such that every object in V is better than any object not in v in terms of being nearest neighbors
  - We do not need a probability threshold!
- Superseding Nearest Neighbors
  - Given two objects A and B, A supersedes B if the probability that q is nearer to A than to B exceeds 0.5

Superseding Nearest Neighbor Search

- B supersedes A
  - The probability that q is closer to B than to A is 0.6
- B is an all-game winner since B supersedes A and C



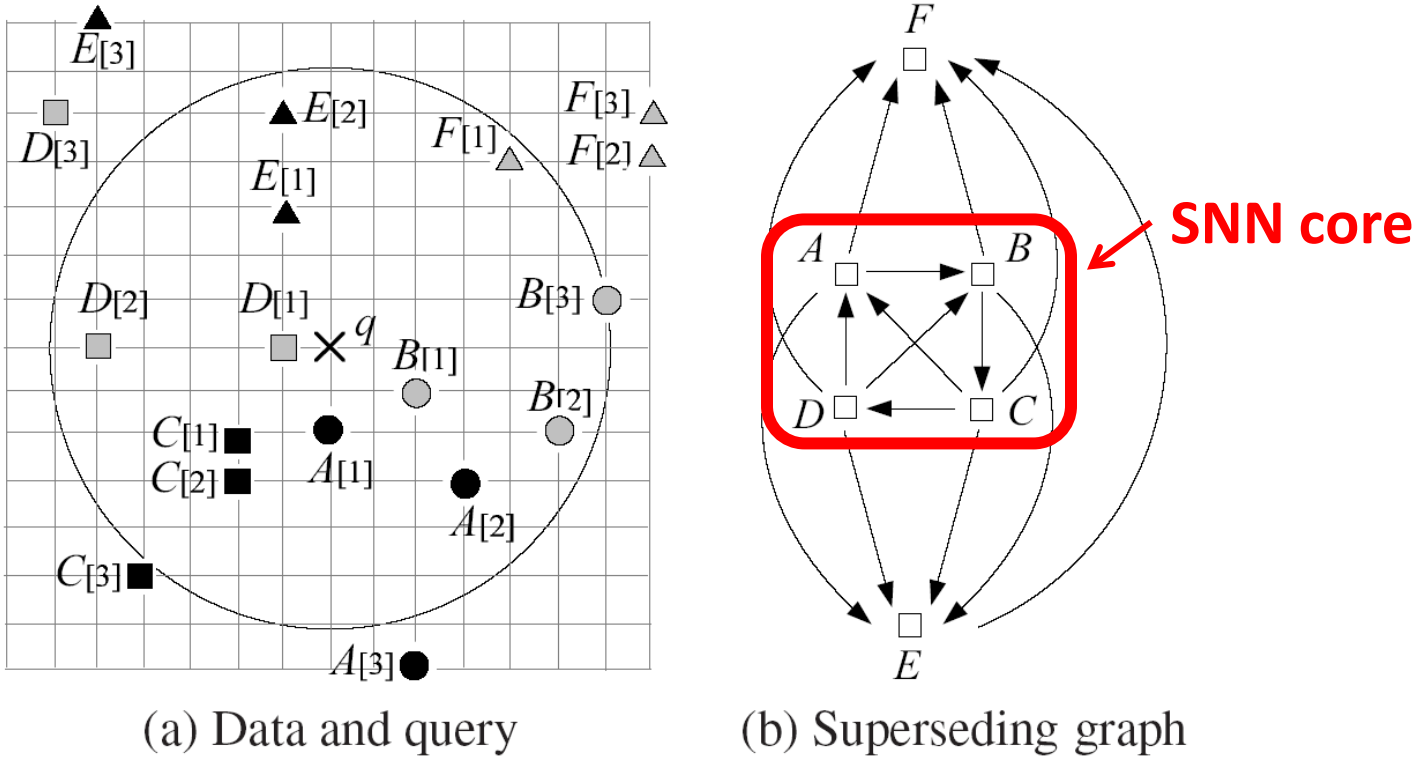
A set of uncertain objects



Superseding graph

Superseding Nearest Neighbor Search

- SNN-core
  - The smallest set of NN-candidates, each of which supersedes all the NN candidates outsider the core



(a) Data and query (b) Superseding graph

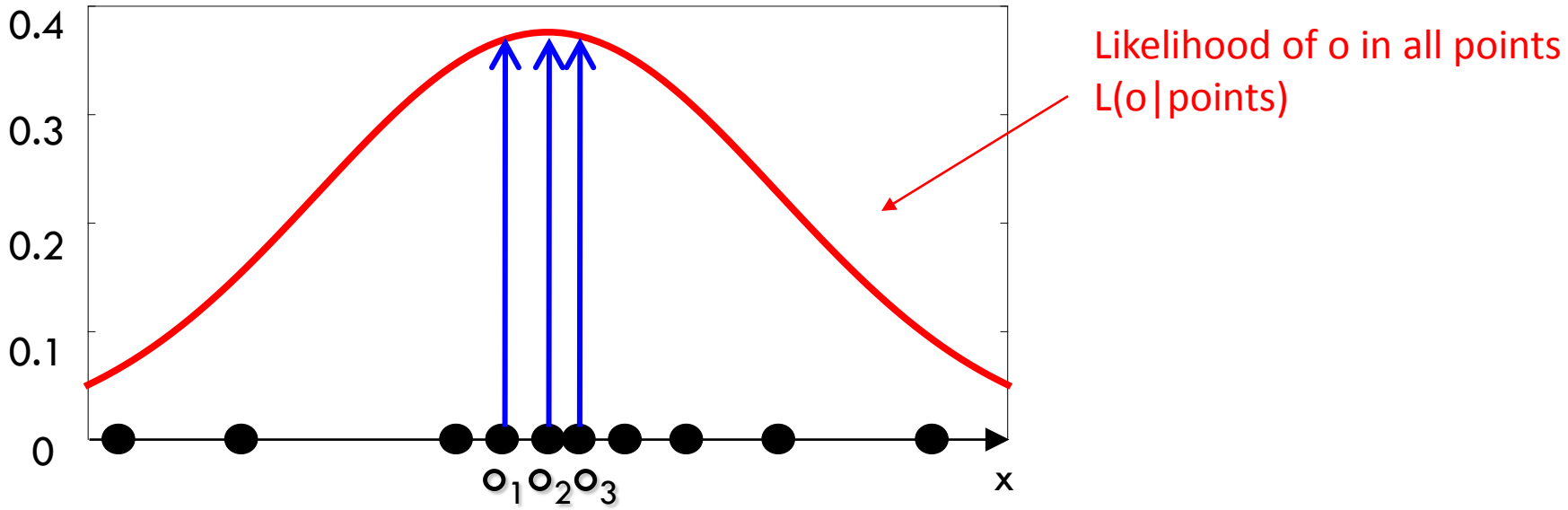
Understanding an Uncertain Object

- What is the most typical instance in uncertain object "mammal"? [Hua *et al*. VLDB'07, Hua *et al*. VLDBJ 2009]

Top-k Simple Typicality Queries

" What are the top-k most typical mammals? "

Top-k Simple Typicality Queries (on an uncertain object O)
Simple typicality score of an instance o in O = the likelihood of o in O.
A top-k simple typicality query returns the k instances with the highest simple typicality scores.

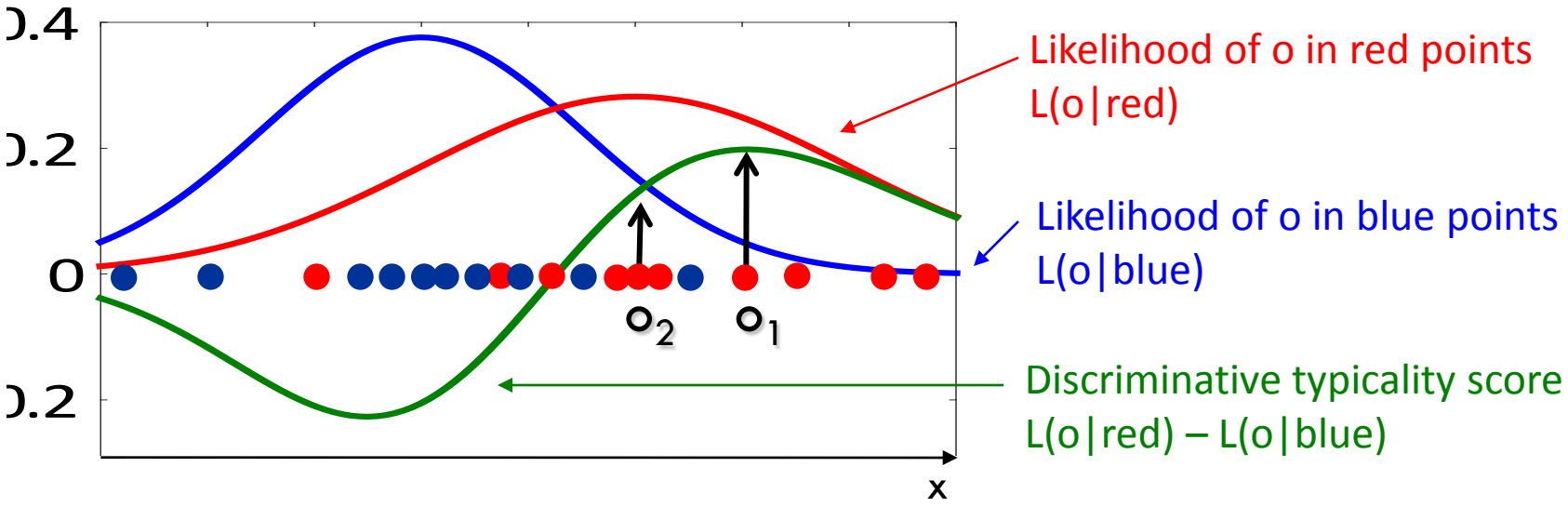


Top-k Discriminative Typicality Queries

" What are the top-k most typical carnivores distinguishing from herbivorous? "

Discriminative Typicality (given an target uncertain object O and an object S)

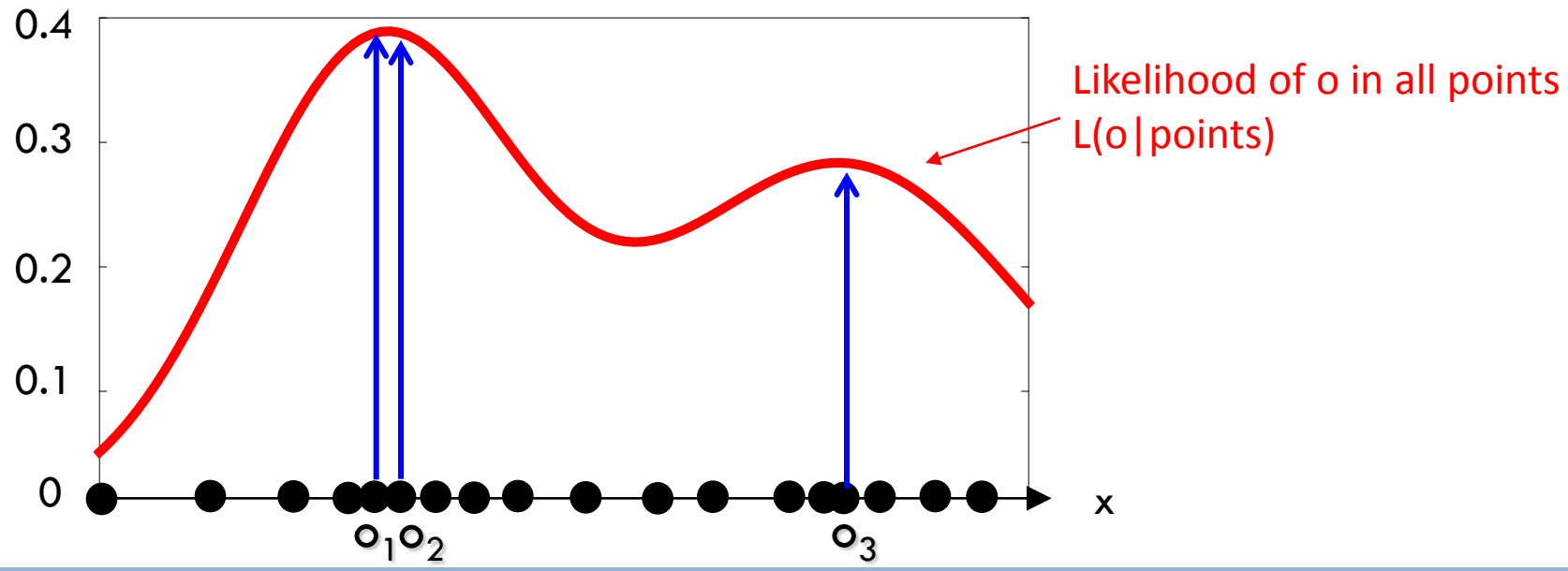
Discriminative typicality score of an instance o in O
= simple typicality score of o in O – simple typicality score of o in S



Top-k Representative Typicality Queries

" What are the top-k most typical mammals in whole representing different types of mammals? "

Group Typicality and Representative Typicality
Group $\{o_1, o_3\}$ represents the distribution of object O better than $\{o_1, o_2\}$.
A top-2 representative typicality returns $\{o_1, o_3\}$.



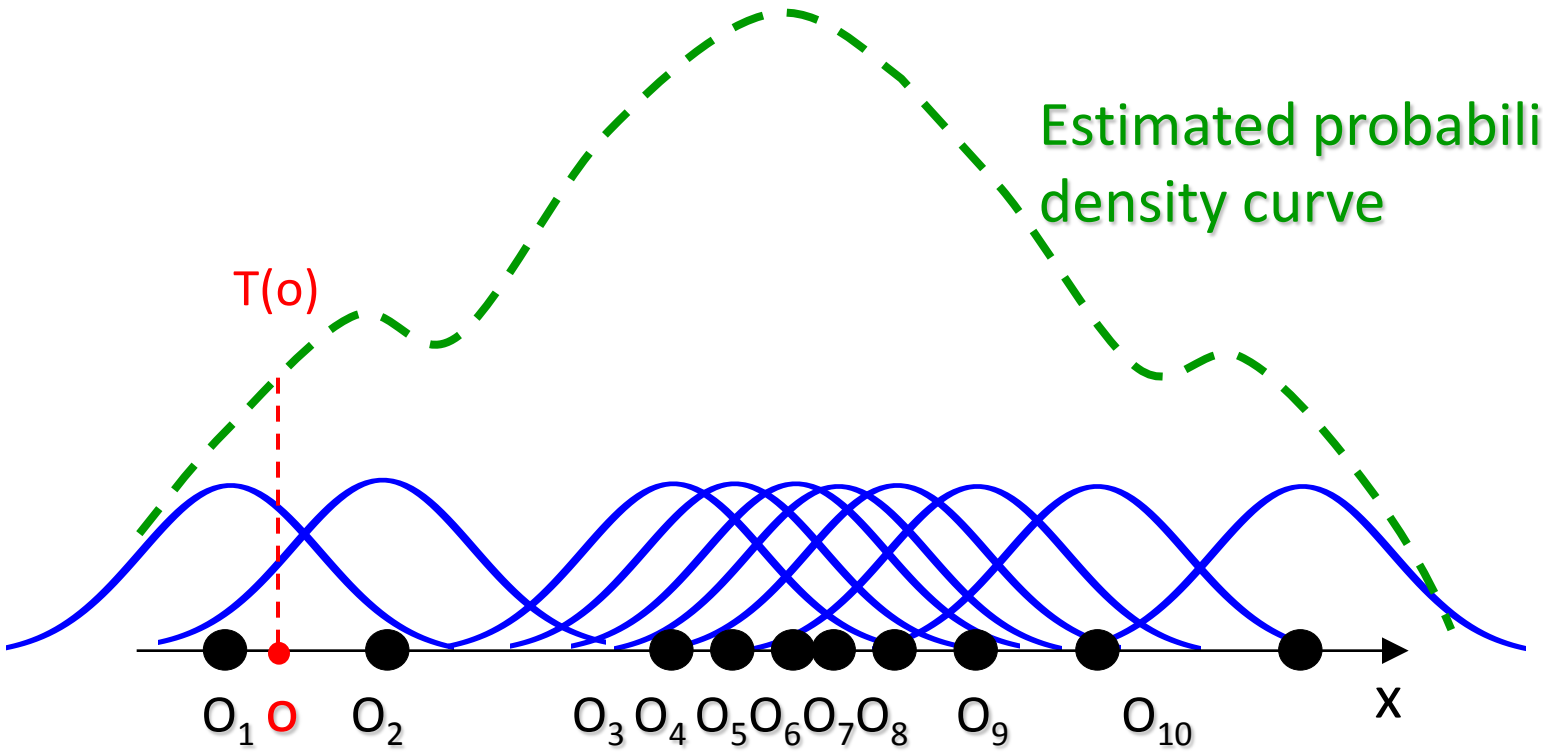
Probability Density Estimation

- Compute the likelihood using the Gaussian Kernel Density Estimation method

The likelihood of other points distributed around o_i
(Gaussian Kernel Function Curve)

The likelihood of o: $$G_h(o, o_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{dist(o,o_i)^2}{2h^2}}$$

Kernel Density Estimation



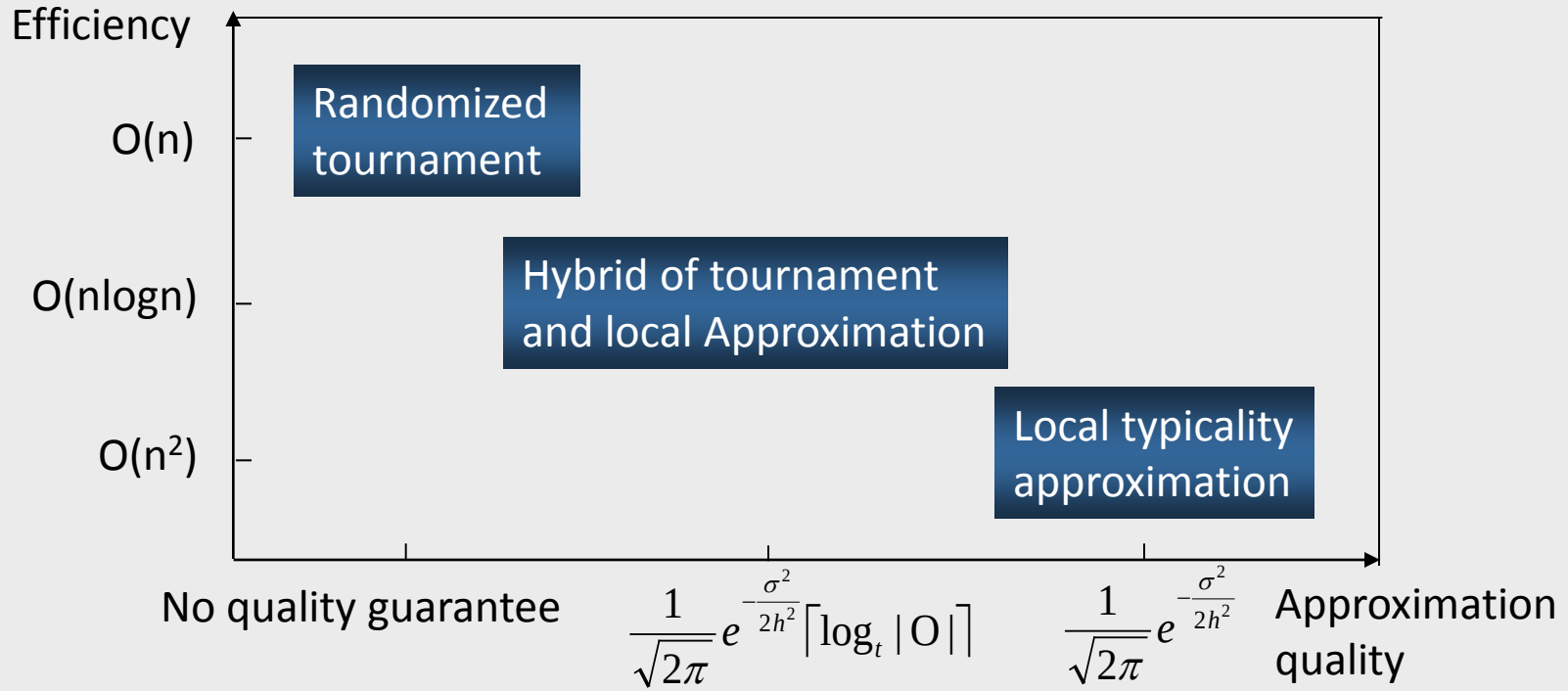
Typicality (Likelihood) of o: the average of all the likelihood contributed from other points

$$T(o) = \frac{1}{nh}\sum_{i=1}^{n} G_h(o, o_i) = \frac{1}{nh\sqrt{2\pi}}\sum_{i=1}^{n} e^{-\frac{dist(o,o_i)^2}{2h^2}}$$

Efficient and Scalable Query Answering

- Use kernel density estimation to compute the likelihood
- The exact query answering algorithm requires quadratic time

Three Approximation Methods



Some Results in Real Data Sets

The Zoo Database

Category	Most typical	Most atypical
Mammal	Leopard, Lion, …	Platypus
Bird	Sparrow, Wren, …	Penguin
Reptile	Slowworm	Seasnake

NBA 2005-2006 Season Statistics

Top-2 most typical guards

Name	3PT	Rebounds	Ast	Blk
Ronald Murray	2.4	2	2.6	0.1
Marko Jaric	2.3	3.1	3.9	0.3

Top-2 most atypical guards

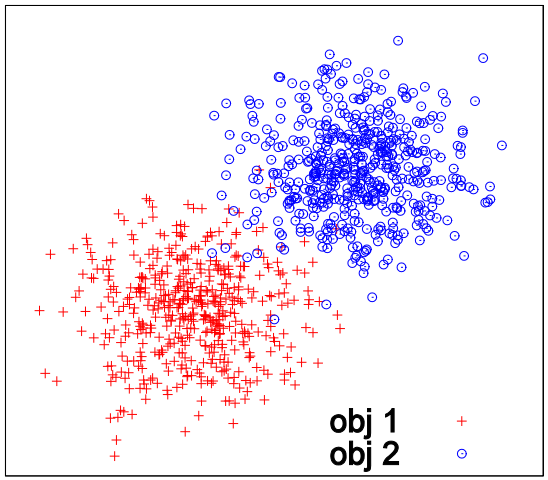
Name	3PT	Rebounds	Ast	Blk
Tracy McGrady	6.6	6.6	4.8	0.9
Corey Maggette	3.0	5.3	2.1	0.1

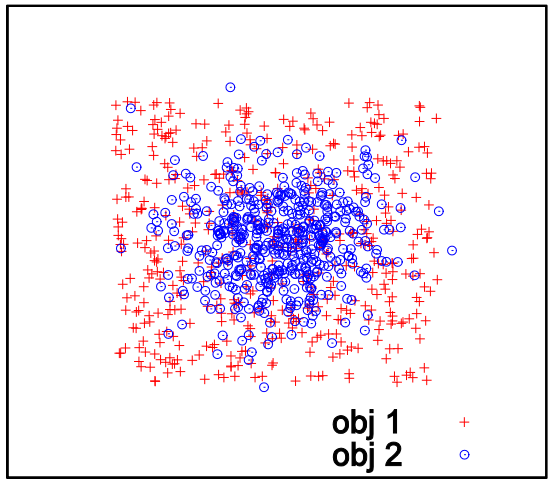
Possible Worlds and Distributions

- Possible world based approaches advocate instance level analysis and object level aggregation
- How can we conduct object level analysis of uncertain data?
  - In statistics, probability distribution has been used successfully to model uncertainty

Clustering Uncertain Objects

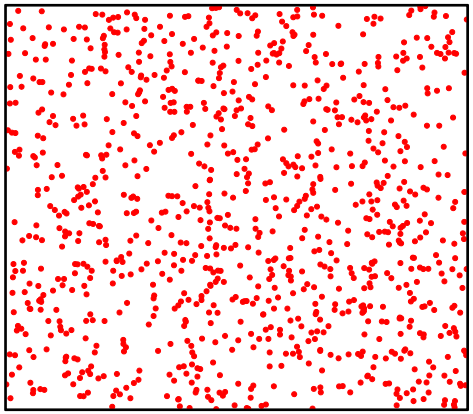
- Traditional methods only explore spatial properties
  - Expected distance-based, such as UK-means, …
  - Density-based, such as DBSCAN, …
  - Possible world model-based
- Spatial properties may not be enough to distinguish objects
  - Example: objects heavily overlapped



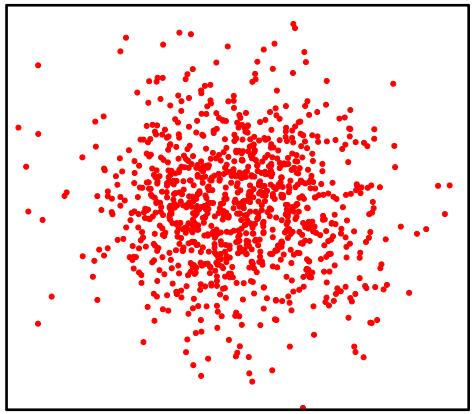


Using Distribution Information

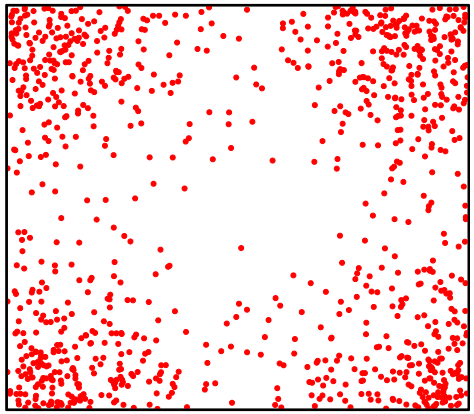
- Differences among the distributions (i.e., probability density functions) of objects



Uniform



Gaussian



Inverse-Gaussian

- PDFs can also capture spatial differences
- Modeling the similarity between PDFs by various statistical measurements, such as KL-divergence …

The K-Medoids Based Method

- Randomly select k representative objects
- Repeat
  - Assign each remaining object to the cluster with the nearest representative object
  - Randomly select a non-representative object
  - Compute the total cost S of swapping the old representative with the new representative
  - If S < 0 then swap and form a new set of k representative objects
- Until no change

Challenges

- Computing KL-divergence is costly
  - KL-divergence has to be derived using density estimation from a set of samples
- Idea: use fast Gauss transform to speed up

Conclusions

- Uncertainty is inherent and ubiquitous in many applications
- Analyzing large uncertain data sets
  - Handling uncertainty at different levels: instance level and object level
  - Instance based or object based analysis? – possible world model versus probability distribution similarity

What Is Next?

- How to obtain meaningful uncertain data?
  - Large data sets
  - More importantly, application needs and logics
- What are the problems inherent to uncertain data?
  - Do we have some problems that cannot exist without uncertainty?
- How can uncertain data analysis results be presented and understood well?

Our Recent Work (1)

- W. Zhang, X. Lin, Y. Zhang, J. Pei, and W. Wang. "Threshold-based Probabilistic Top-k Dominating Queries", To appear in the VLDB Journal, Springer Berlin / Heidelberg.
- M. Hua, J. Pei, A. W.-C. Fu, X. Lin, and H-F Leung. "Top-k Typicality Queries and Efficient Query Answering Methods on Large Databases". The VLDB Journal, Volume 18, Number 3, pages 809-835, June 2009Springer Berlin / Heidelberg.
- S. Yuen, Y. Tao, X. Xiao, J. Pei, D. Zhang. "Superseding Nearest Neighbor Search on Uncertain Spatial Databases". To appear in IEEE Transactions on Knowledge and Data Engineering, IEEE Computer Society.
- M. A. Cheema, X. Lin, W. Wang, W. Zhang, and J. Pei. "Probabilistic Reverse Nearest Neighbor Queries on Uncertain Data". To appear in IEEE Transactions on Knowledge and Data Engineering, IEEE Computer Society.
- M. Hua and J. Pei. "Continuously Monitoring Top-K Uncertain Data Streams: A Probabilistic Threshold Method". Distributed and Parallel Databases: An International Journal, Volume 26, Number 1, (special issue on ranking in databases), pages 29-65, August, 2009, Springer-Verlag.
- J. Pei, M. Hua, Y. Tao, and X. Lin. "Mining Uncertain and Probabilistic Data: Problems, Challenges, Methods and Applications''. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08), August 24-27, 2008, Las Vegas, NV, USA.

Our Recent Work (2)

- W. Zhang, X. Lin, J. Pei, and Y. Zhang. "Managing Uncertain Data: A Probabilistic Approach" (invited paper). In Proceedings of the 9th International Conference on Web-Age Information Management (WAIM'08), July 20-22, 2008, Zhangjiajie, China.
- J. Pei, M. Hua, Y. Tao, and X. Lin. "Query Answering Techniques on Uncertain and Probabilistic Data". In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD'08), June 11-14, 2008, Vancouver, Canada.
- M. Hua, J. Pei, W. Zhang, and X. Lin. "Ranking Queries on Uncertain Data: A Probabilistic Threshold Approach". In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD'08), June 11-14, 2008, Vancouver, Canada. (Implementation, the real data sets, the synthetic data generator and the data sets).
- M. Hua, J. Pei, W. Zhang, and X. Lin. "Efficiently Answering Probabilistic Threshold Top-k Queries on Uncertain Data". In Proceedings of the 24th International Conference on Data Engineering (ICDE'08), Cancún, México, April 7-12, 2008.
- J. Pei, B. Jiang, X. Lin, and Y. Yuan. "Probabilistic Skylines on Data". In Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB'07), Vienna, Austria, September 23-28 2007.
- M. Hua, J. Pei, A. W.-C. Fu, X. Lin, and H-F Leung. "Efficiently Answering Top-k Typicality Queries on Large Databases". In Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB'07), Vienna, Austria, September 23-28 2007.