LMU LUDWIG-MAXIMILIANS-UNIVERSITY MUNICH | DEPARTMENT INSTITUTE FOR INFORMATICS | DATABASE SYSTEMS GROUP

36th International Conf. on Very Large Data Bases

VLDB SINGAPORE 2010

Similarity Search and Mining in Uncertain Databases

Matthias Renz*, Reynold Cheng**, Hans-Peter Kriegel*, Andreas Zuefle* and Thomas Bernecker*

*)
Ludwig-Maximilians-Universität München (LMU)
Munich, Germany
http://www.dbs.ifi.lmu.de
{renz, kriegel}@dbs.ifi.lmu.de

**)
University of Hong Kong (HKU)
Hong Kong
http://www.cs.hku.hk
ckcheng@cs.hku.hk

General Issues

1. Please feel free to ask questions at any time during the presentation

2. Main goal of the tutorial:
   – Foster understanding of different types of similarity search techniques efficiently supporting data retrieval and data analysis in the context of imprecise and inexact data
   – Learn core techniques for efficient similarity query processing on uncertain data

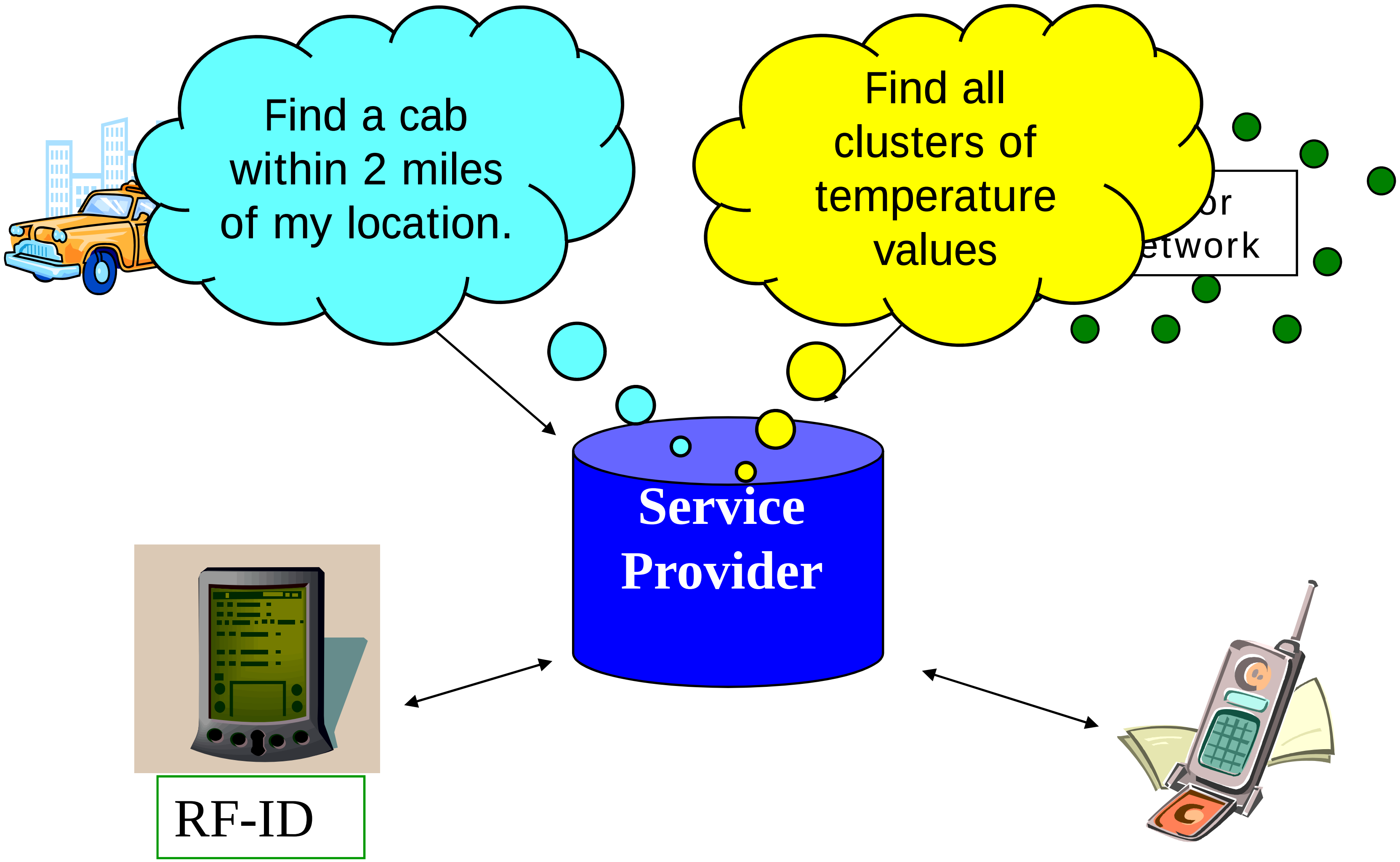
- The latest version of these slides will be made available within the next week:

http://www.dbs.ifi.lmu.de/~renz

Outline

- Introduction
  - Motivation
  - Uncertain Data Modelling
  - Challenges
- Similarity Search in Uncertain Data
  - Probabilistic Similarity Queries: Overview and Classification
  - Probabilistic ε-Range, NN, kNN and Ranking Queries
- Mining Uncertain Data
- Summary

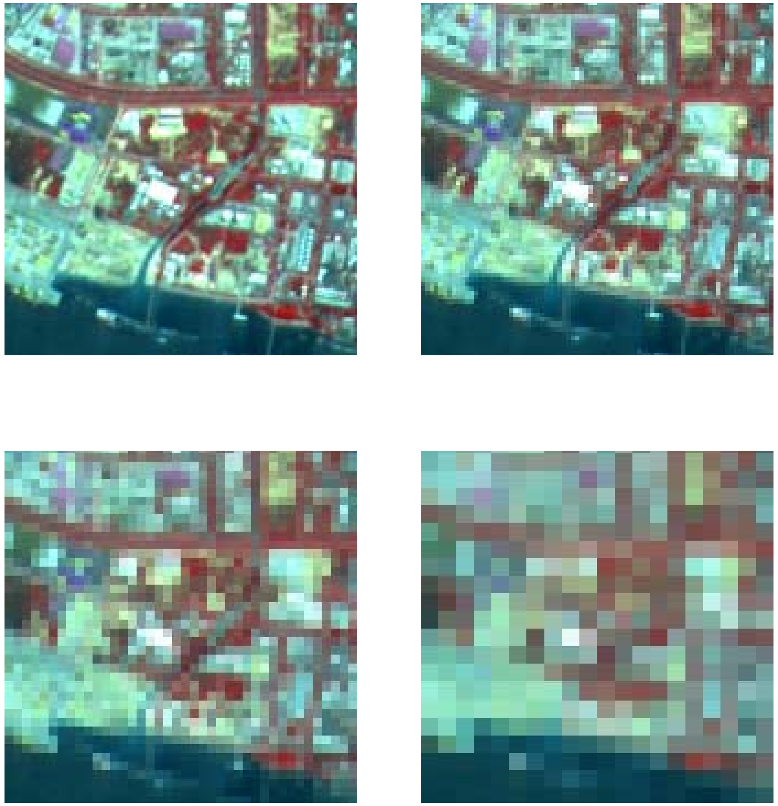
Location and Sensor Applications



Data Uncertainty in Location and Sensor Applications

- Due to limited network bandwidth and battery power, readings are just sampled
- The value of the entity being monitored (e.g., temperature, location) is changing
- The database stores old values only
- *Query/analysis results can be incorrect, resulting in poor service quality*

Location Uncertainty in Satellite Images

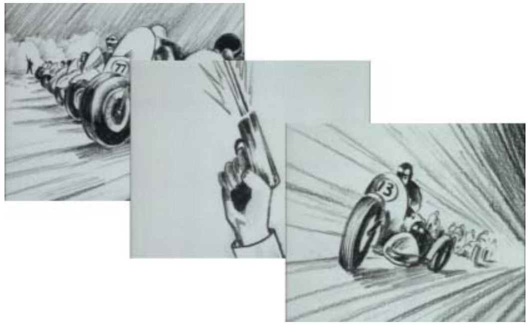


- Identities and locations of objects can be extracted from satellite images
- Due to the blurredness of the images, query, analysis, and pattern matching can be incorrect

Other Uncertain Data

- Temporal Data: uncertain time series

- Data Streams:
  - Uncertainty in audio/video data transmission



- Biological databases
  - Uncertainty in retina cell images

Data Uncertainty

- Data uncertainty exists in many emerging applications
  - Location-based services
  - Sensor data analysis
  - Biological image analysis
  - Economic decision making
  - Market surveillance

- How can we perform correct and efficient analysis on a large amount of imprecise and inexact data?

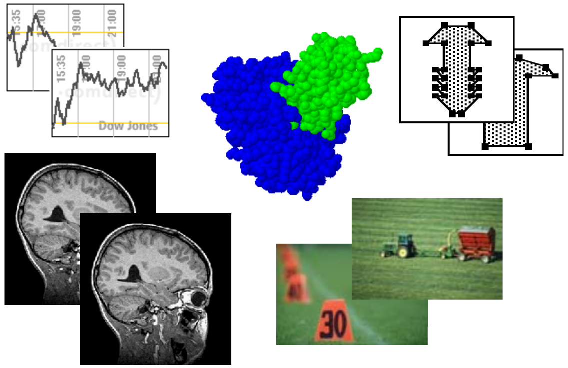
Similarity Search

- **Similarity Search:** return objects in a set of data collection which are similar or close to the *query object*
  - Data mining: identify groups of similar objects for clustering or classification
  - Pattern recognition: finds patterns that are highly similar to a given pattern

- Traditional similarity research focuses on:
  - Precise and accurate data
  - Definition of similarity metrics
  - Efficient and scalable similarity search algorithms on multi-dimensional space

Similarity Search: Basic Principles

- Basic Idea: Feature transformation



Feature transformation

histograms

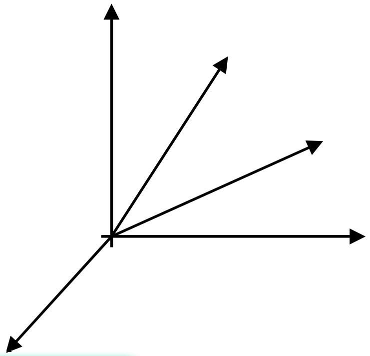
moment invariants

rectangle cover

binning

Fourier coefficients

etc.



Applica[tion (o…]

How can similarity be done on uncertain data?

[Featu]e space

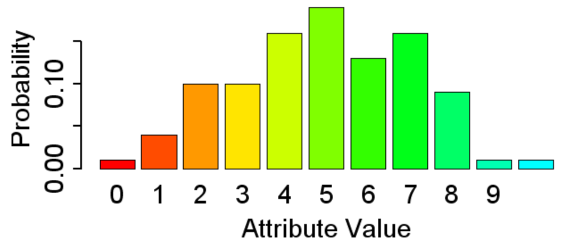
  - Extract a set of (usually numeric) features from the objects
  - Transform each object into a feature vector
  - Similarity of objects = vicinity of corresponding feature vectors
    - Similarity queries in the object space = neighbor queries in the feature space
    - Can be supported by spatial index structures

Uncertain Data Model

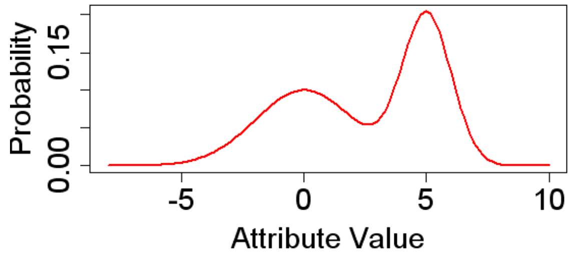
- Uncertain attribute

  An attribute *x* is uncertain if its value is given by a probabilistic density function (PDF), which describes all possible values *v* of *x*, associated with probability P(*x*=*v*).

  – Discrete PDF (e.g., temperature history data)



  – Continuous PDF (e.g., sensor measurement error)



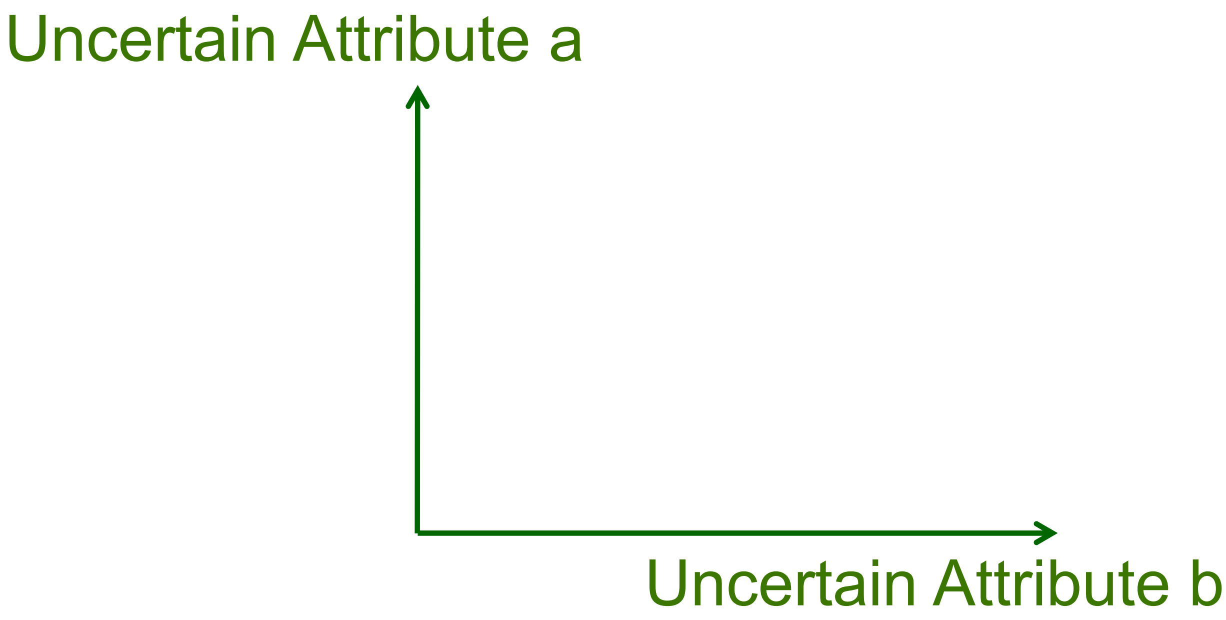
Uncertain Data Model

- Uncertain Object X
  - Has at least d ≥ 1 uncertain attributes.
  - Each uncertain attribute value of X is a random variable.
  - We say that X is a random variable, where the set of attribute values of X is described by a multi-dimensional PDF_X.

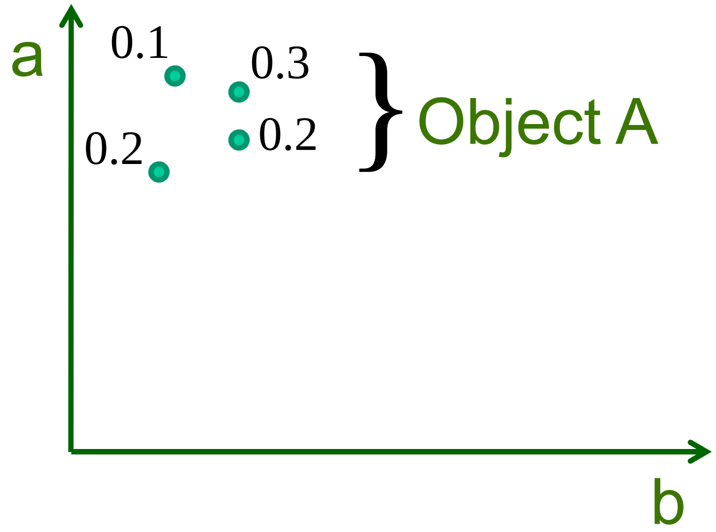
Uncertain Data Model

- Uncertain Object X
  - Has at least d ≥ 1 uncertain attributes.
  - Each uncertain attribute value of X is a random variable.
  - We say that X is a random variable, where the set of attribute values of X is described by a multi-dimensional PDF_X.
  - In the discrete case, X has a finite set of so-called instances $\{X_1, \ldots, X_m\}$ so that $P(X=t)>0$ if $t \in \{X_1, \ldots, X_m\}$, and $P(X=t)=0$ otherwise.
  - In the continuous case, X has a spatial region UR_X, the so-called Uncertain Region, so that $PDF_X(t)>0$ if $t \in UR_X$ and $PDF_X(t)=0$ otherwise.

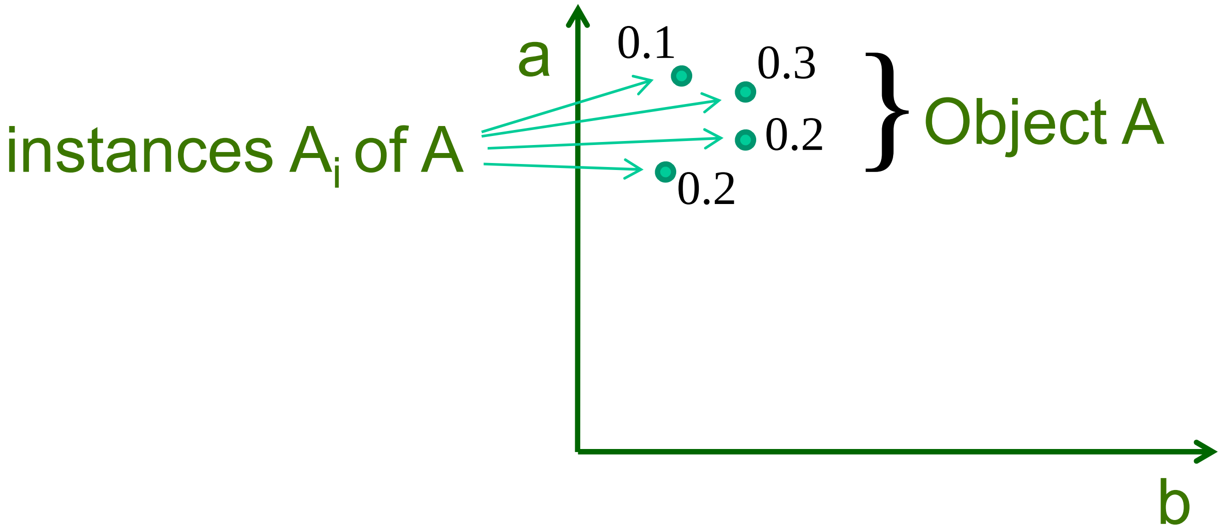
Discrete Uncertain Data



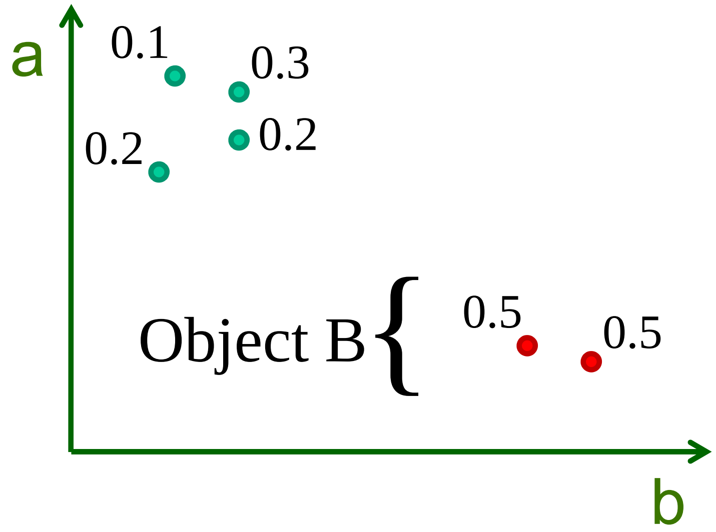
Discrete Uncertain Data



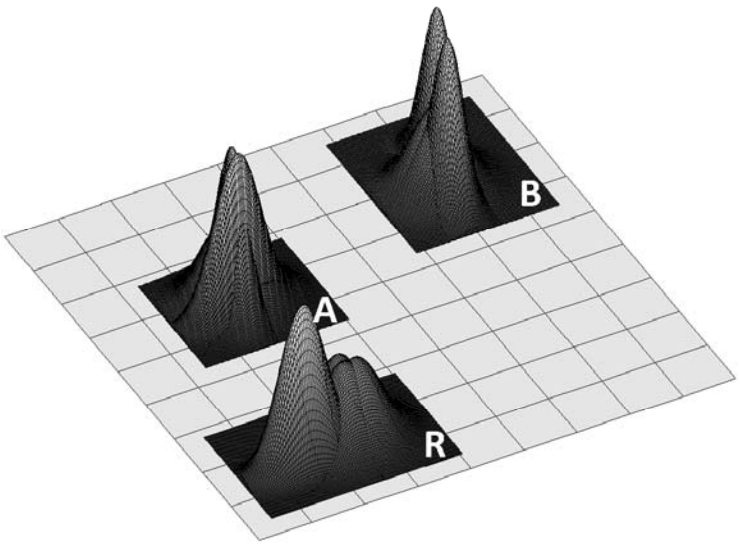
Discrete Uncertain Data



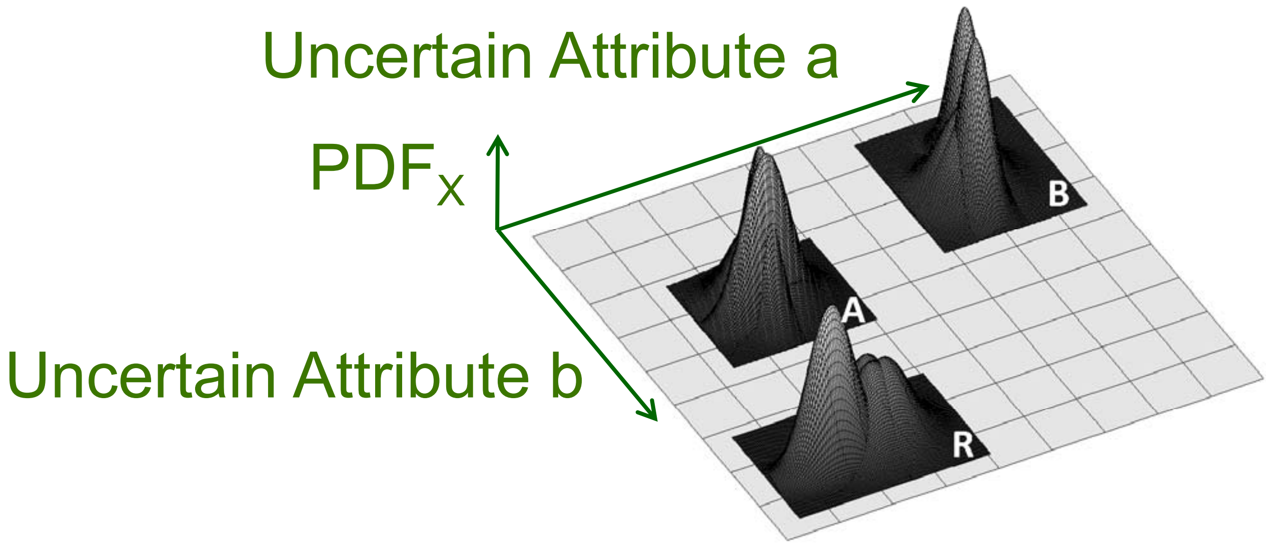
Discrete Uncertain Data



Continuous Uncertain Data



Continuous Uncertain Data



Uncertain Data Model

For each uncertain object *X*,

$$\sum_{v \in X_i} pmf_X(v) \leq 1$$ Discrete case

$$\int_{v \in UR_X} pdf_X(v)dv \leq 1$$ Continuous case

Attribute Uncertainty: Brief Survey

Uncertainty models

- Continuous uncertainty (**pdf + range**) [Sistla98,Pfoser99,Cheng03]
- Tuple uncertainty and continuous pdf attributes [Singh08]
- Sensor correlation models [Desphande04, Wang08]

Query Evaluation and Indexing

- Probabilistic query classification [Cheng03]
- Range queries [Sistla98,Kriegel06,Pfoser99, Cheng04b,Tao05,Tao07, Cheng07]
- Nearest-neighbor [Cheng04a,Kriegel07,Ljosa06,Cheng08a,Beskales08, Cheng09, Chen09, Cheng10a]
- MIN/MAX [Cheng03,Deshpande04]
- Skylines [Pei07]
- Reverse skylines [Lian08]
- Object Identification [Bohm06]

Summary on Other Uncertainty Models

Data Models

- Independent tuple/attribute uncertainty [Barbara92]
- **x-tuple** (ULDB) [Benjelloun06]
- Graphical model [Sen07]
- Categorical uncertain data [Singh07]
- World-set descriptor sets [Antova08]

Query Evaluation and Other Works

- Efficiency of query evaluation [Dalvi04]
- Top-k query evaluation [Soliman07,Re07,Yi08]
- Storing information extraction models [Sarawagi06]
- Continuous queries on data streams [Jin08]
- Handling schema matching uncertainty [Cheng10c]
- Uncertain data cleaning [Cheng08b, Cheng10b]

Challenge 1: Query Semantic

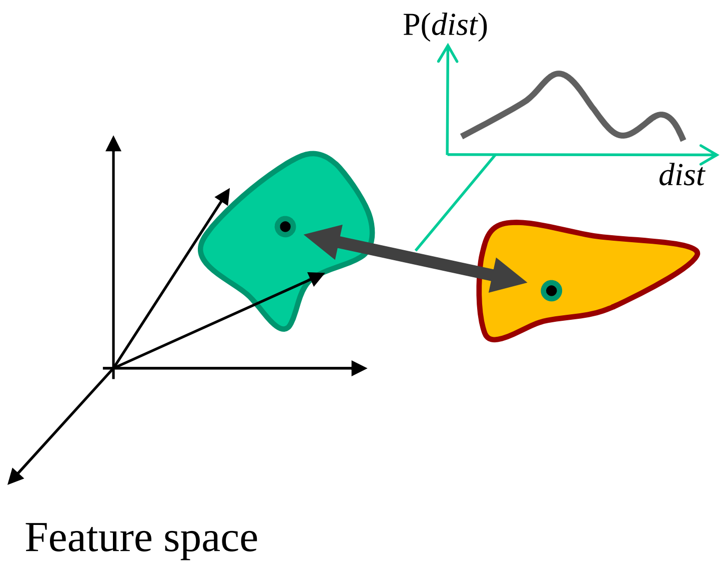
- How to define semantics of *similarity search* over uncertain data?

- Similarity (i.e., distances) between uncertain objects is uncertain
  - Validation of query predicate is uncertain
  - Similarity between two objects is probabilistic

- There exists a lot of different semantics/interpretations

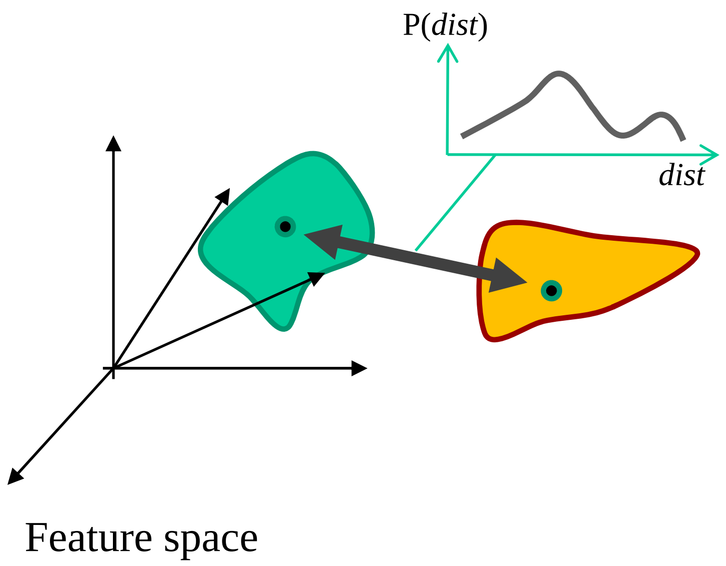
- How to understand and differentiate each of them?

Challenge 1: Query Semantic



- Uncertain objects yield uncertain location in feature space
- Similarity distance is ambiguous => neighbor queries in the feature space designed for uncertain distances are required

Challenge 2: Query Evaluation



- Traditional similarity query processing methods, which are designed for precise data, cannot be used anymore
- Probabilistic similarity is much more difficult to compute (see next)
- We need new evaluation tools (e.g., indexing and multi-step query processing)

Challenge 2: Query Evaluation

- Possible Worlds

  An instantiation of an uncertain databases is given by creating one instance of each uncertain object. If the probability of this instantiation is greater zero, it is called a *possible world.*

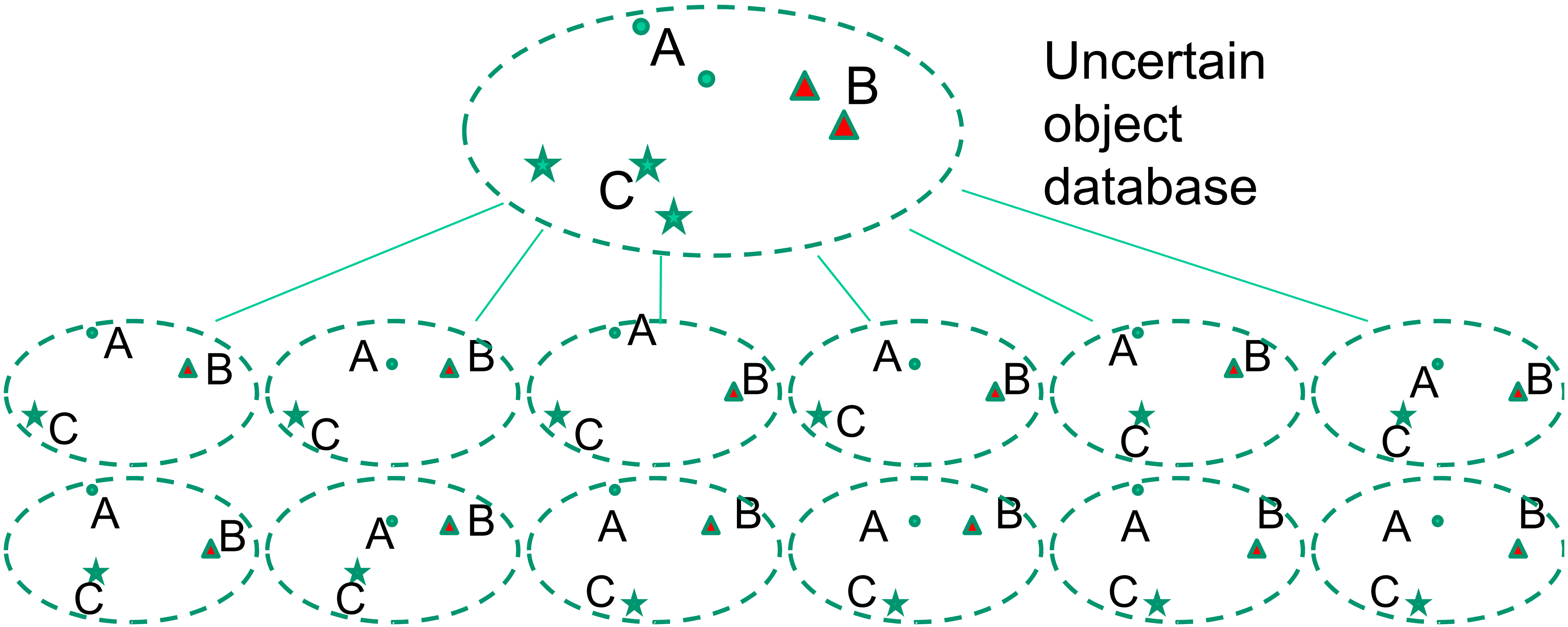
  In the discrete uncertainty model, a possible world is associated with a probability greater zero.

- The Problem:

  The number of possible worlds grows exponential in the number of uncertain objects.

Possible Worlds Semantics: Example

- Object Tracking Systems: Current locations of a set of moving objects are observed from multiple sensor devices



12 Possible Worlds

The number of possible worlds is exponential

Efficient Similarity Search (1)

- Query processing on uncertain data is in general very expensive
- A probabilistic similarity search is naturally more expensive than its non-probabilistic counterpart

- **High CPU cost,** due to the use of possible world semantics and numerical integration
  - Do not materialize all possible worlds
  - Use different efficient methods from statistics

- **High I/O cost**, since a lot of objects might be affected
  - Develop effective pruning strategies (spatial and probabilistic pruning)

Efficient Similarity Search (2)

- Modify query semantics and results, by:
- Imposing constraints on query semantics:
  - Instead of returning the most likely result, return objects which are most likely to be in the result.
- Restricting the query result, for example:
  - Return only the k-most probable results
  - Return only results for which the probability exceeds a given threshold
- Returning approximate results

Tutorial Goals

- Provide a classification of a variety of probabilistic similarity search queries
  - Compare the definitions and usage of different queries
- Understand the core evaluation techniques used in similarity search algorithms
- Hopefully, we can provide you key information knowledge for designing new similarity metrics and similarity search techniques, that are suitable for uncertain databases

General Issues

1. Please feel free to ask questions at any time during the presentation

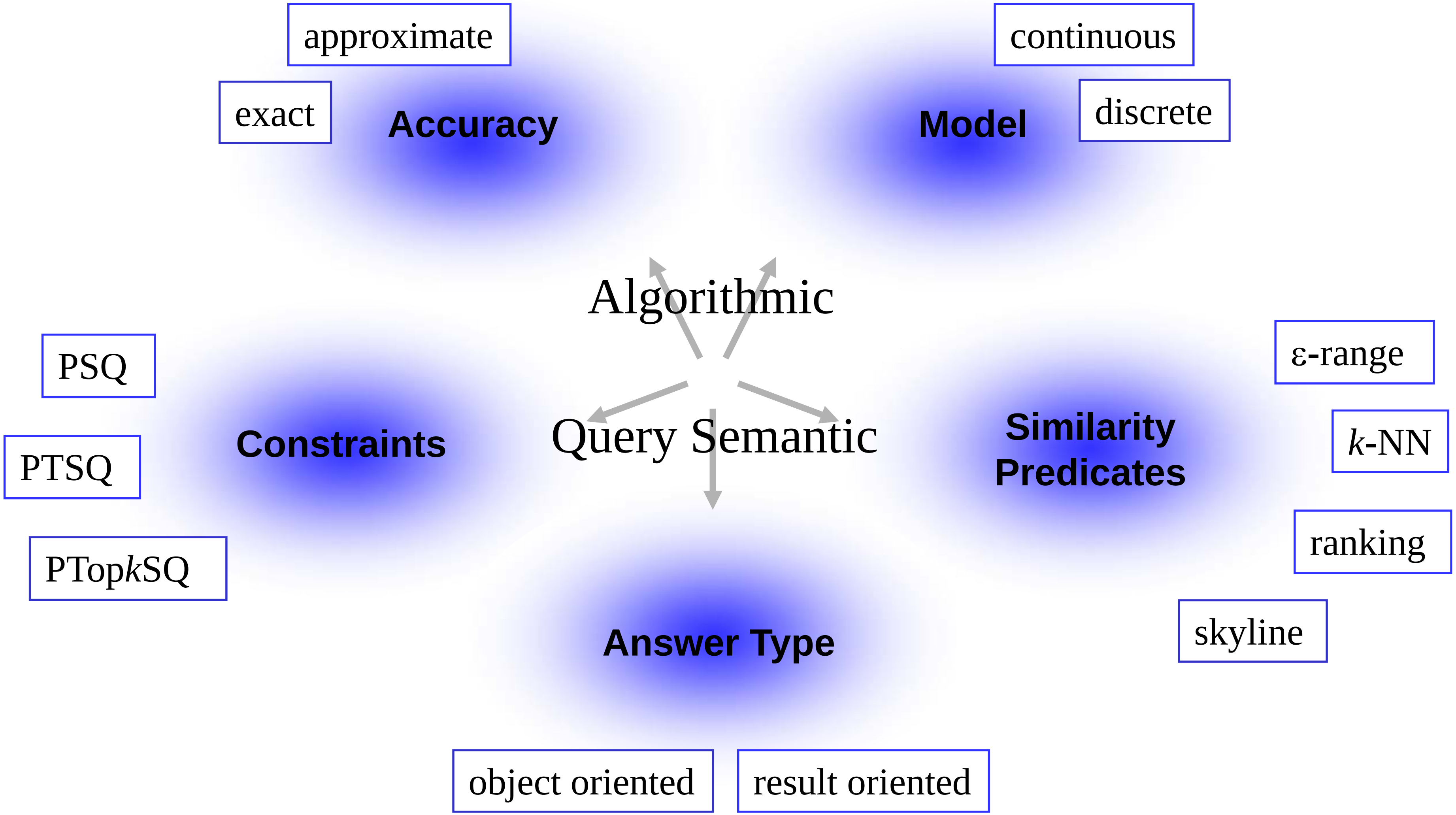
2. Main goal of the tutorial:
   - Foster understanding of different types of similarity search techniques efficiently supporting data retrieval and data analysis in the context of imprecise and inexact data
   - Learn core techniques for efficient similarity query processing on uncertain data

- The latest version of these slides will be made available within the next week:

http://www.dbs.ifi.lmu.de/~renz

Overview and Classification

- Classification of Probabilistic Similarity Queries (Overview)



Overview and Classification

- Similarity Query Predicates:

Similarity Predicates

ε-range

Report all objects within a given distance range of query q

k-NN

Report the k objects most closest to query q

ranking

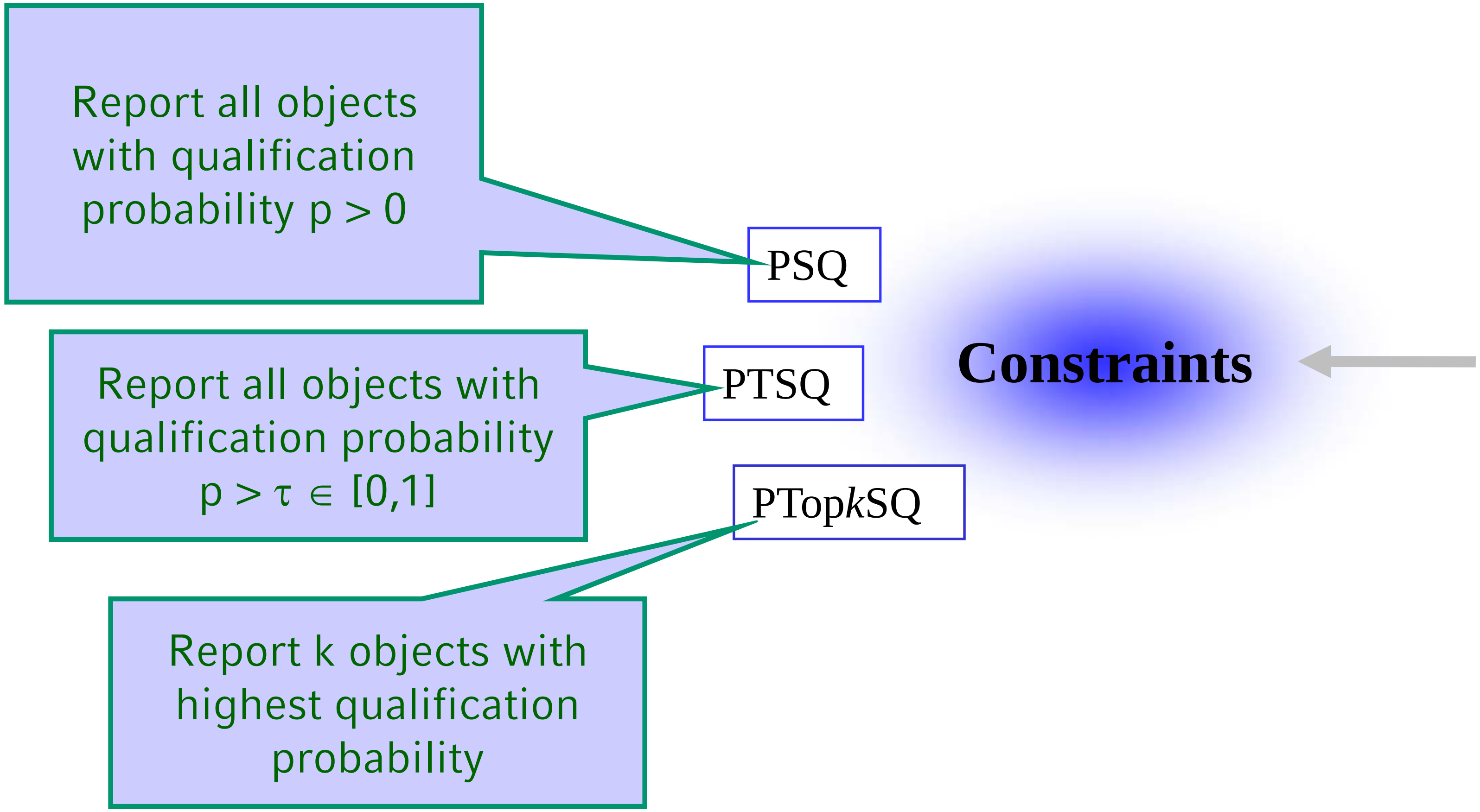
Report objects with ascending distance to query q

skyline

Report objects not dominated by other objects

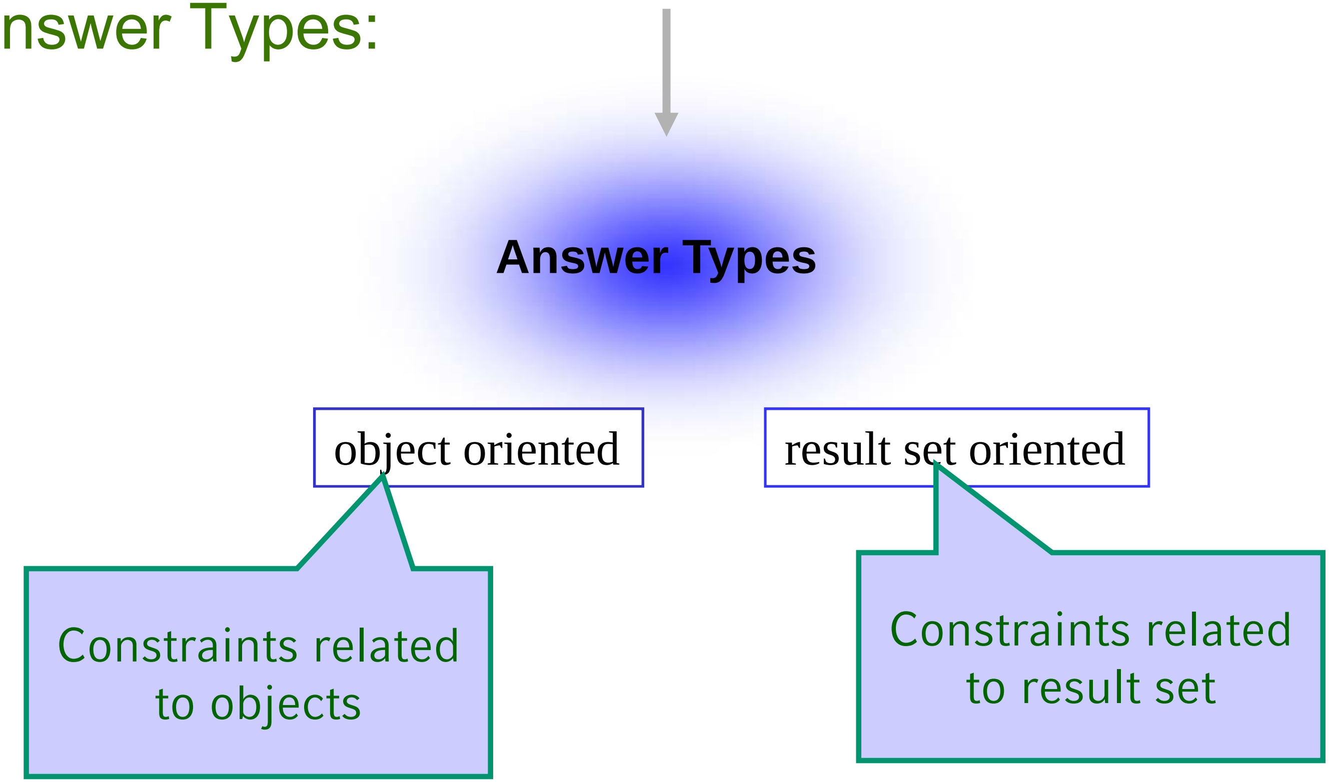
Overview and Classification

- Constraints on qualification probability for the result set:



Overview and Classification

- Answer Types:



Overview and Classification

- Answer Types:

OID	P(d(q,X)<ε)	P(d(q,X)>ε)
C	0.3	0.7
D	0.6	0.4
F	0.8	0.2
H	0.2	0.8

possible ε-range results:
{C}(0,0192), {D}(0,0672), {F}(0,1792), {H}(0,0112), {C,D}(0,0288), {C,F}(0,0768), {C,H}(0,0048), {D,F}(0,2688), {D,H}(0,0168), {F,H}(0,0448), {C,D,F}(0,1152), {C,D,H}(0,0072), {C,F,H}(0,0192), {D,F,H}(0,0672), {C,D,F,H}(0,0288)

Answer Types

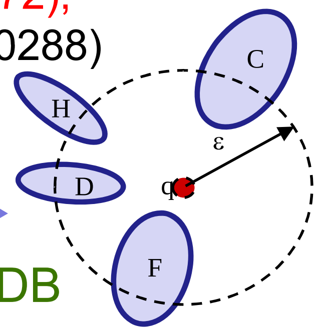
object oriented

query result (τ = 0.5):
{D(0.6),F(0.8)}

result set oriented

query result (τ = 0.05):
{{D}(0,0672), {F}(0,1792), {C,F}(0,0768), {D,F}(0,2688), {C,D,F}(0,1152), {D,F,H}(0,0672)}

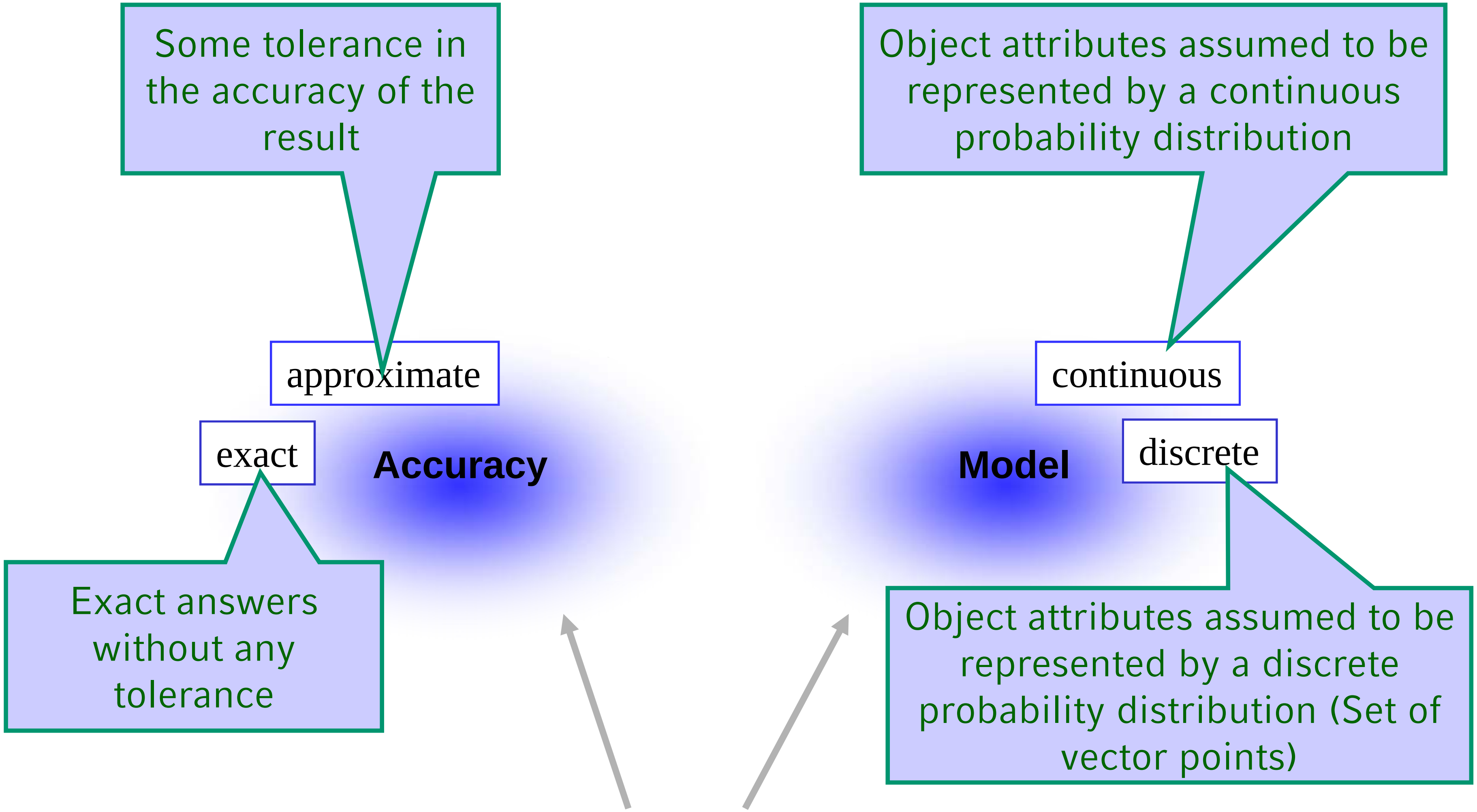
- Example:
  - Query type: *constraint* = PTSQ,
    similarity predicate = ε-range



query object and DB

Overview and Classification

- Algorithmic Issues:



Probabilistic Similarity Search

– In the following:

We will learn the most important concepts/tools for an efficient evaluation of probabilistic similarity queries following the above classification scheme, including:

- Indexing,
- Multi-step query processing,
- Spatial and probabilistic pruning
- Binomial recurrence / generating functions

Techniques introduced by a selection of sample queries grouped by the *similarity predicate* attribute, including

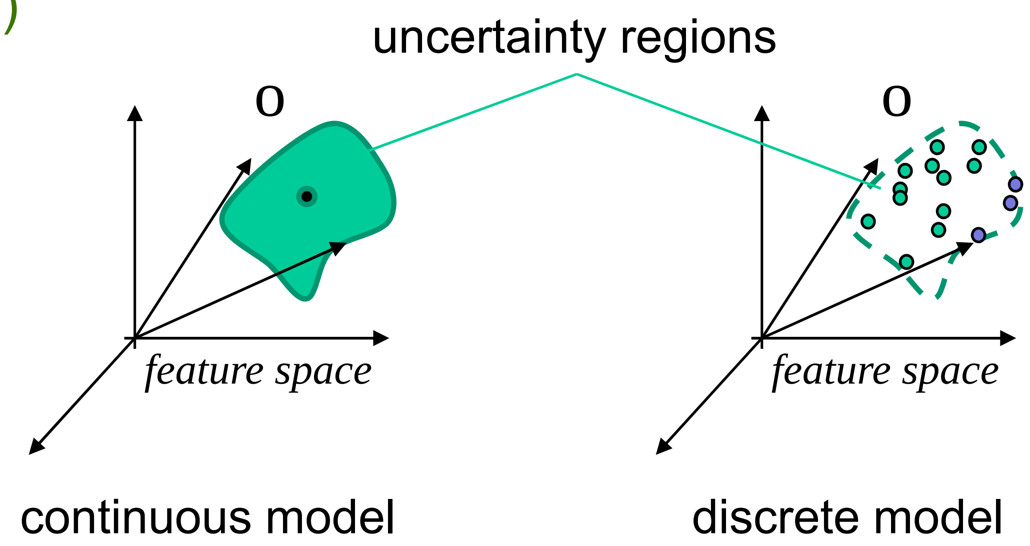
- Probabilistic ε-Range Query,
- Probabilistic Nearest Neighbor Query,
- Probabilistic k-Nearest Neighbor Query and
- Probabilistic Ranking Query

Uncertain Similarity Distance

- Preliminaries:
  - An uncertain object is represented by a region called **uncertainty region** covering all possible locations of the object
  - Possible locations of an object are specified by a

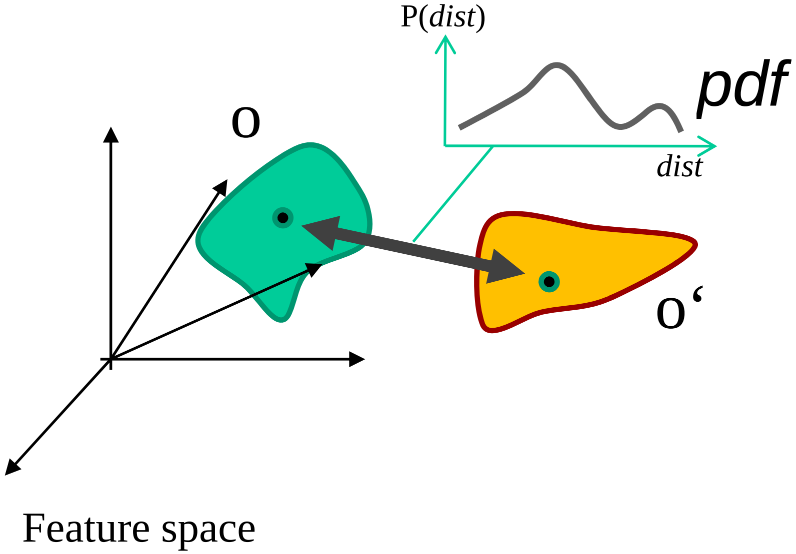
    probability density function (continuous model) defined within the uncertainty region or

    set of location instances assigned with a probability value (discrete model)



Uncertain Similarity Distance

- Preliminaries:
  - Similarity expressed by a distance function, e.g. Euclidean distance
  - The distance between uncertain objects is uncertain as well
  - The representation depends on the underlying uncertainty model
  - It can be expressed by a pdf (continuous model)

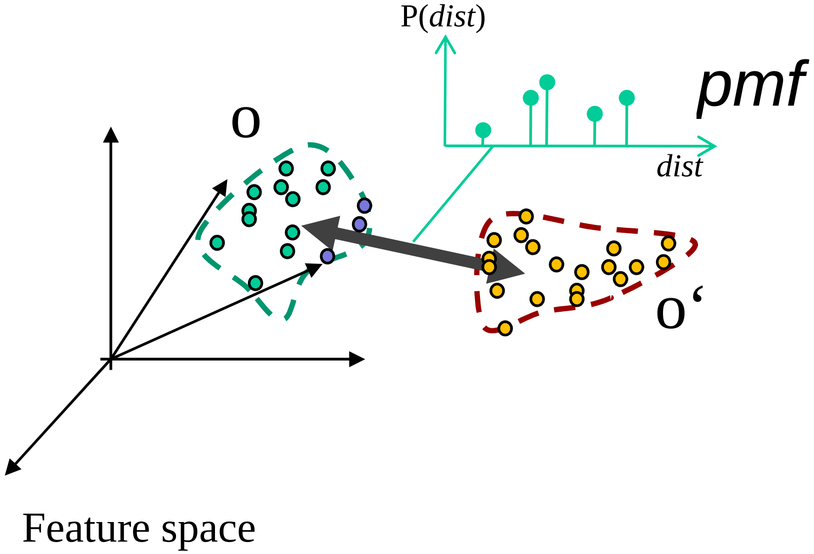


  - Formally:

$$P(a \leq dist(o,o') \leq b) = \int_a^b pdf_{dist(o,o')}(x)dx$$

Uncertain Similarity Distance

- Preliminaries:
  - Similarity expressed by a distance function, e.g. Euclidean distance
  - The distance between uncertain objects is uncertain as well
  - The representation depends on the underlying uncertainty model
  - It can be expressed by a pdf (continuous model) or by a pmf (discrete model)

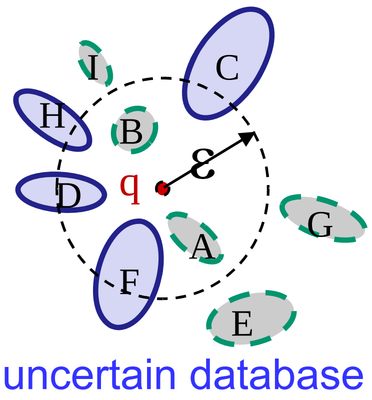


  - Formally:

distance $= \{(d, p) \mid x \in X, y \in Y, d = dist(x, y), p = P(x) \cdot P(y)\}$

Probabilistic Distance Range Query

- Probabilistic Distance Range Query
  - Query Semantic (Classification):
    - Similarity predicate: ε-range
    - Constraints: PTSQ (PSQ, PTopkSQ)
    - Answer type: object oriented



  - Given:
    - Query object *q, q*uery radius ε, probability threshold τ (or k for PTopkSQ)
  - Search:
    - All objects having probability > τ being within ε-range of query *q*, formally:

$$PTRQ(q,\varepsilon) = \{o \in DB \mid P(dist(q,o) \leq \varepsilon) > \tau\}$$

where

$$P(dist(q,o) \leq \varepsilon) = \int_0^{\varepsilon} pdf_{dist(q,o)}(x)dx \quad \text{(continuous model)}$$

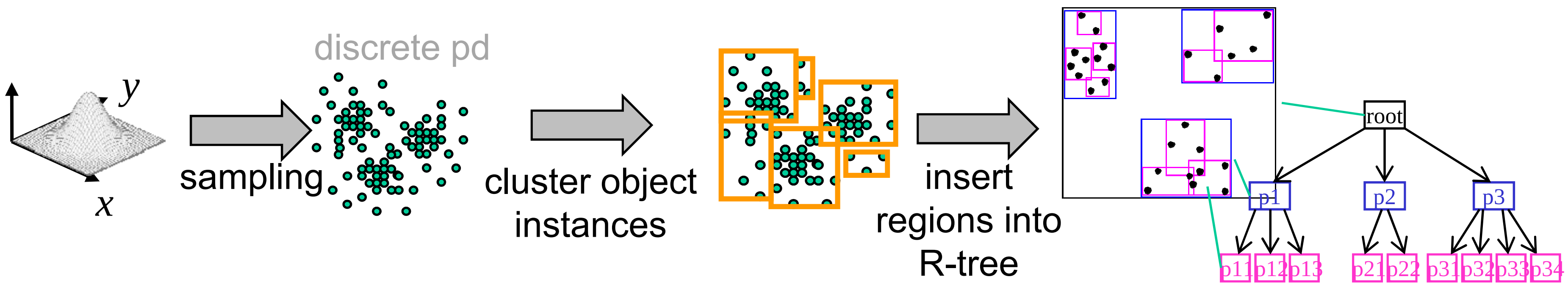
$$P(dist(q,o) \leq \varepsilon) = \sum_{x \leq \varepsilon} pmf_{dist(q,o)}(x) \quad \text{(discrete model)}$$

Probabilistic Distance Range Query

- Naive Approach:
  - For each object *o* compute the probability P(*dist*(*o*,*q*)≤ε)
  - Report objects with probability P(*dist*(*o*,*q*)≤ε) > τ
- Properties:
  - Too many objects accessed (large I/O-overhead)
  - Expensive integration for the evaluation of P(*dist*(*o*,*q*)≤ε) (high CPU cost)
- What we need to perform queries efficiently:
  - Appropriate index methods coping with uncertain data
    - primarily based on spatial keys (support spatial pruning)
    - consideration of probability distributions (support probabilistic pruning)
  - Efficient and effective pruning heuristics
  - Fast and accurate estimation of P(*dist*(*o*,*q*)≤ε)

Probabilistic Distance Range Query

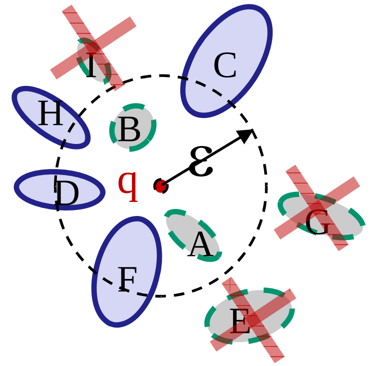
– Solution based on R-tree: [Kriegel06]
  - Algorithmic properties:
    - Accuracy: approximate
    - Model: discrete
    - Also applicable for exact solutions and continuous model
  - Object Management (Index):
    - Uncertain objects decomposed into a set of uncertainty regions
    - Each region represented by (mbr_i,p_i,oid_i)
    - Regions efficiently organized by an R-tree



R-tree used to organize the uncertainty regions

Probabilistic Distance Range Query

- Spatial Pruning:
  - Prune all objects which uncertainty region is outside of the ε-range of query q.

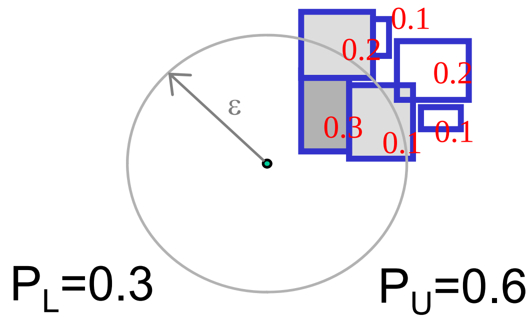


- Probabilistic Pruning:
  - Build lower/upper probability bounds P_L/P_U

$$P_L(o) = \sum_{ur \in o, MaxDist(ur,q) \leq \varepsilon} P_o(ur)$$

$$P_U(o) = \sum_{ur \in o, MinDist(ur,q) \leq \varepsilon} P_o(ur)$$

ur = uncertainty region partition
P_o(*ur*) = probability that object *o* in *ur*



Probabilistic Distance Range Query

- Query Processing (PTSQ):

  Filter:
  - Retrieve from the R-tree all objects and their uncertainty regions intersecting the query range (spatial pruning)
  - For each retrieved object o:
    - » Compute lower and upper bounds P_L and P_u
    - » If PL ≥ τ, then report o as true hit
    - » Else if PU < τ, then report o as true drop
    - » Else insert o into candidates

  Refinement:
  - For each o $\in$ candidates, compute $P(dist(o,q) \leq \varepsilon)$

- Pruning techniques also applicable for other types of queries, e.g. PTopkSQ [Kriegel06]

Probabilistic Distance Range Query

– Summary:
  - Approximate qualification probabilities (conservative and progressive) derived from decomposed uncertainty regions
  - Concept of multi-step query processing is applied to probability approximations used to prune objects (probabilistic pruning) and true hit detection (probabilistic filtering)
  - Filter performance depends on the granularity of the object decomposition: trade-off between redundancy and filter performance

  - Properties:

    ☺ Using an existing and well established index method (no specialized index necessary)

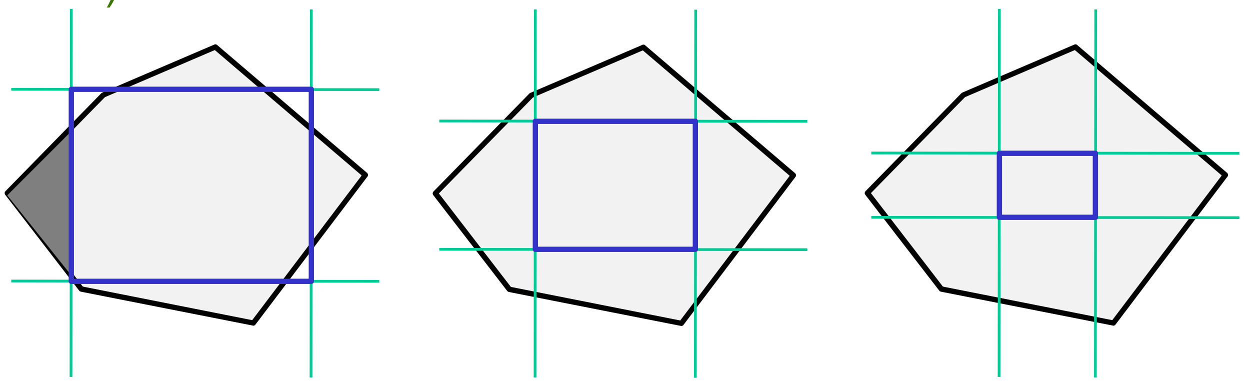
    ☺ Very easy to implement

    ☺ Object approximations adapts well to arbitrary uncertain region shapes

    ☹ high redundancy in terms of uncertainty-region decompositions

Probabilistic Distance Range Query

- Solution based on U-tree: [Tao05]
  - Algorithmic properties:
    - Accuracy: exact
    - Model: continuous
  - Object Management (Index):
    - Uncertain objects approximated by a set of probabilistic constrainted regions (PCRs)

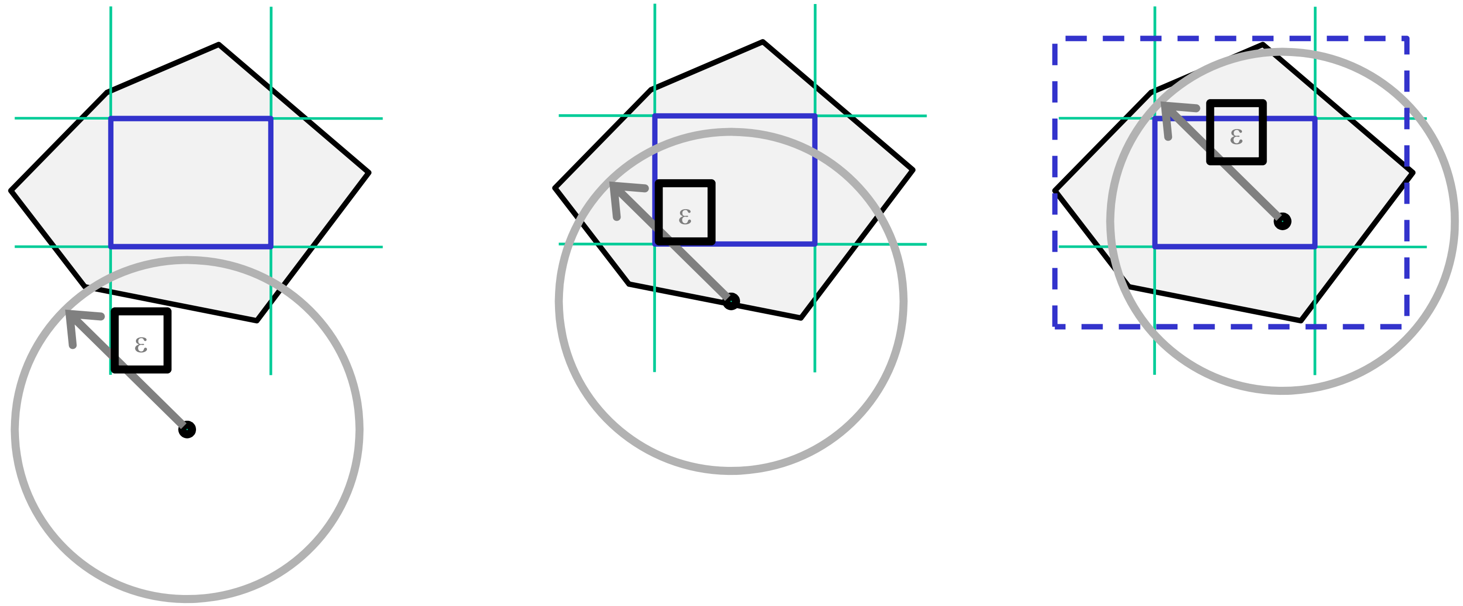


    - PCRs are indexed by an R-tree-like index called U-tree
    - Leave-Level: organizing a set of PCRs for each uncertain object
    - Non-Leave-Level: approximations of PCR-bounds (x-bounds) from leaf-level are provided to higher index levels

      → probabilistic pruning

Probabilistic Distance Range Query

- Probabilistic Pruning:



Upper bound qualification probability $P_U = 0.3$

Upper bound qualification probability $P_U = 0.7$

Lower bound qualification probability $P_L = 0.4$

- Mbr-aproximation provides spatial pruning as well
- Multi-step query processing strategies applicable for PTSQ and PTopkSQ similar to the R-tree based solution

– Summary:
  - Approximate qualification probabilities (conservative and progressive) derived from sets of PCRs
  - Concept of multi-step query processing can be applied for early pruning and true hit detection
  - Filter performance depends on the granularity of the PCRs

  - Properties:

    ☺ low redundancy

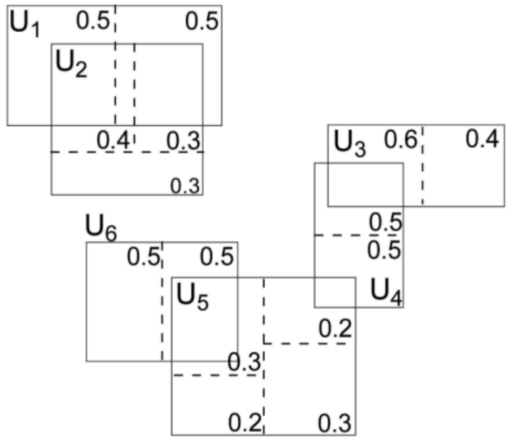
    ☺ provides probabilistic pruning on higher index levels

    ☹ specialized index structure

    ☹ approximation adapts not well to arbitrary uncertainty shapes

Probabilistic Distance Range Query

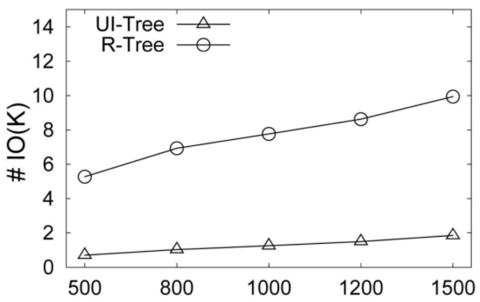
- Solution based on UI-tree [Zhang10]
  - Algorithmic properties:
    - Accuracy: approximative (exact)
    - Model: discrete (continuous)
  - Basic idea:
    - Decompose uncertainty regions into disjunctive partitions
    - **Merge part. with similar spatial key (reduce redundancy)**
    - Assign list of objects having a t least one part. within the merged part.
    - Index the merged part. with R-tree

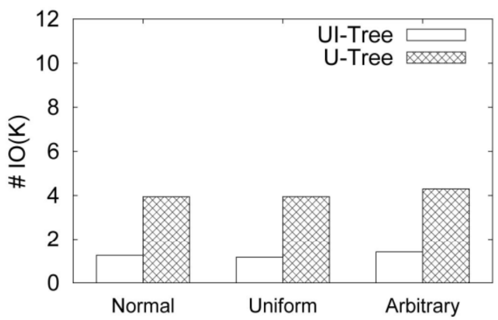


  - Uncertainty region approximation similar to R-tree based method => good adaption to arbitrary uncertainty region shapes
  - Reduced redundancy due to partition merging
  - Spatial and probabilistic pruning similar to R-tree based approach

Probabilistic Distance Range Query

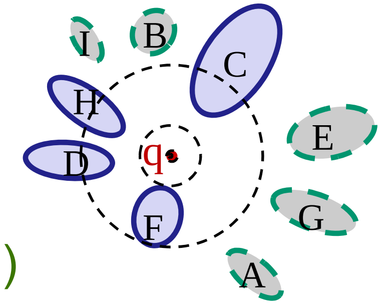
– Summary:
  - R-tree based solution:
    - Use existing index methods ☺
    - Good approximation quality ☺
    - High redundancy ☹
  - U-tree:
    - Compact object approximation ☺
    - Supports pruning on higher index levels ☺
    - Not well adaption to arbitrary pdf shapes ☹
  - UI-tree:
    - Good trade off between approximation quality and redundancy w.r.t. uncertainty regions





Probabilistic Nearest-Neighbor Query

- Probabilistic Nearest-Neighbor Queries
  - Query Semantic (Classification):
    - Similarity predicate: NN (1NN)
    - Constraints: PTSQ (PSQ, PTopkSQ)
    - Answer types: object oriented (result oriented)



Uncertain database

  - Given:
    - Query object *q*, query parameter *k*, probability threshold τ (or k for PTopkSQ)
  - Search:
    - All objects having probability > τ being NN of query *q*, formally:

$$PTNNQ(q,k) = \{o \in DB \mid P(\forall p \neq o : dist(q,o) \leq dist(q,p)) > \tau\}$$

where

(continuous model)

$$P(\forall p \neq o : dist(q,o) \leq dist(q,p)) = \int_0^\infty pdf_{dist(q,o)}(x) \cdot \prod_{p \neq o} (1 - \int_0^x pdf_{dist(q,p)}(y)dy)dx$$

(discrete model)

$$P(\forall p \neq o : dist(q,o) \leq dist(q,p)) = \sum_x pmf_{dist(q,o)}(x) \cdot \prod_{p \neq o} (1 - \sum_{y \leq x} pmf_{dist(q,p)}(y))$$

Probabilistic Nearest-Neighbor Query

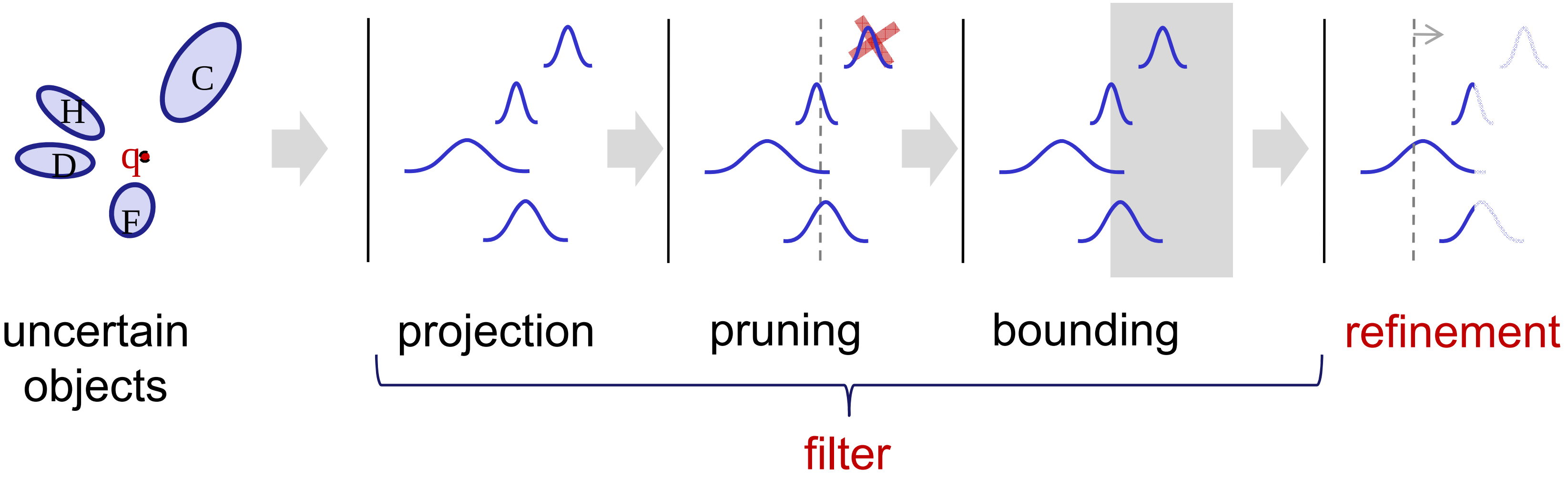
- Computation more complex than range queries:
  - Genrally: NN qualification of an object depends on the (uncertain) attributes of other objects
  - Impact on spatial and probabilistic pruning
  - We need new strategies supporting mutual pruning

- Concepts:
  - Spatial pruning: Projection-Pruning-Bounding [Cheng03], Sample based approach [Kriegel07]

  - Probabilistic pruning: Probabilistic Verfication [Cheng08]

Probabilistic Nearest-Neighbor Query

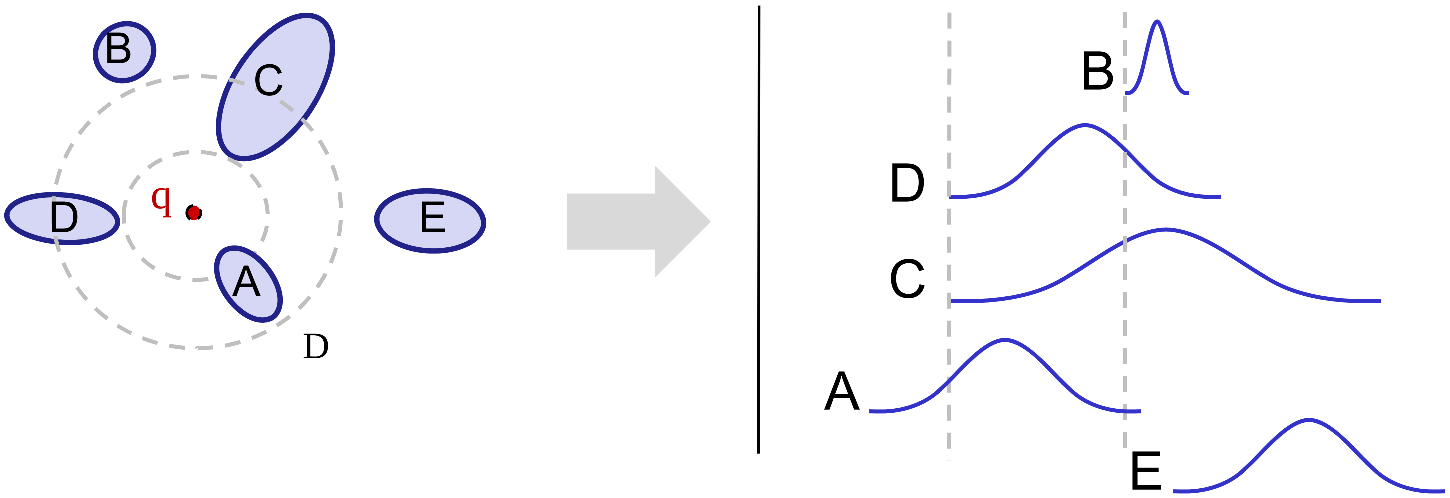
- Projection-Pruning-Bounding: [Cheng03]
  - Filter-refinement-pipeline:



  - We just concentrate on the filter to demonstrate the pruning concepts
  - Only spatial pruning supported

Probabilistic Nearest-Neighbor Query

1. step: Projection

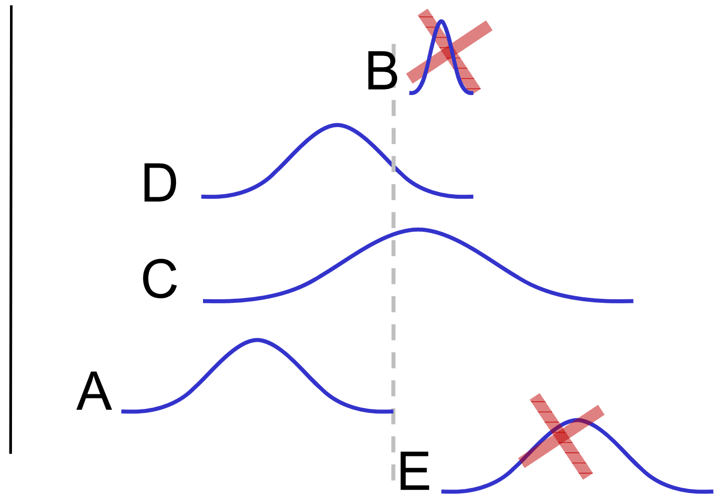


– Uncertain objects (object pdf) are projected to the distance space

– Distances between objects and query are represented by pdf and cdf repectively

Probabilistic Nearest-Neighbor Query

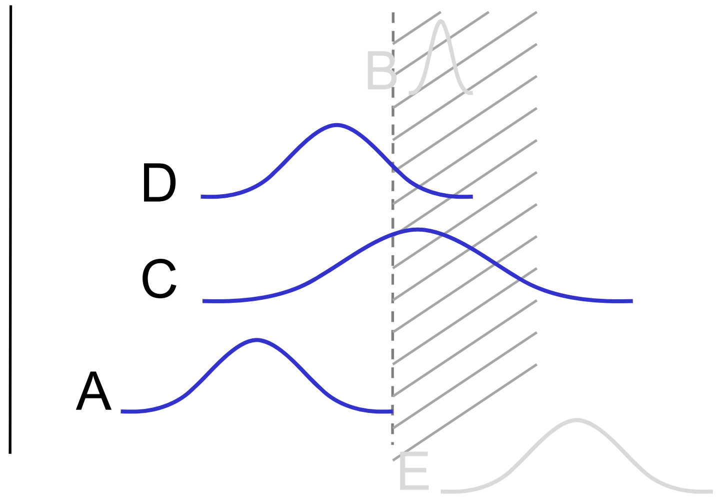
2. step: Pruning (spatial pruning)



- Compute lowest maximum distance D_{NN} (conservative estimation of the NN distance)
- Objects with distance dist(q,o) higher than D_{NN} can be safely pruned, e.g. objects B and E
- Reduced number of candidates to be considered

Probabilistic Nearest-Neighbor Query

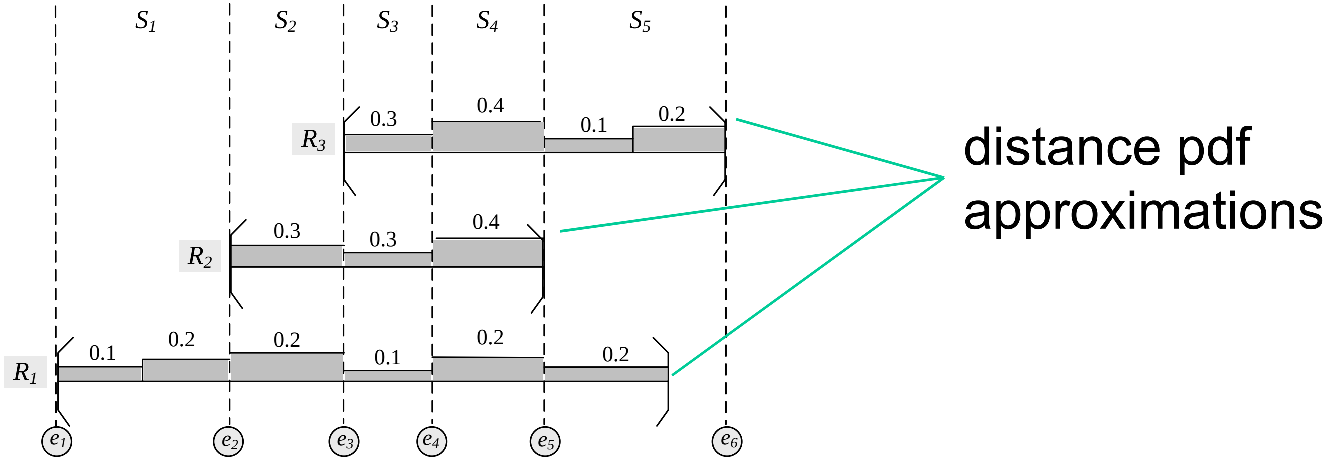
3. step: Bounding



– D_{NN} bounds the distance space where object pdfs have influence on the probabilistic NN result
– Distances above D_{NN} can be ignored
– Reduces integraion cost for the refinement step

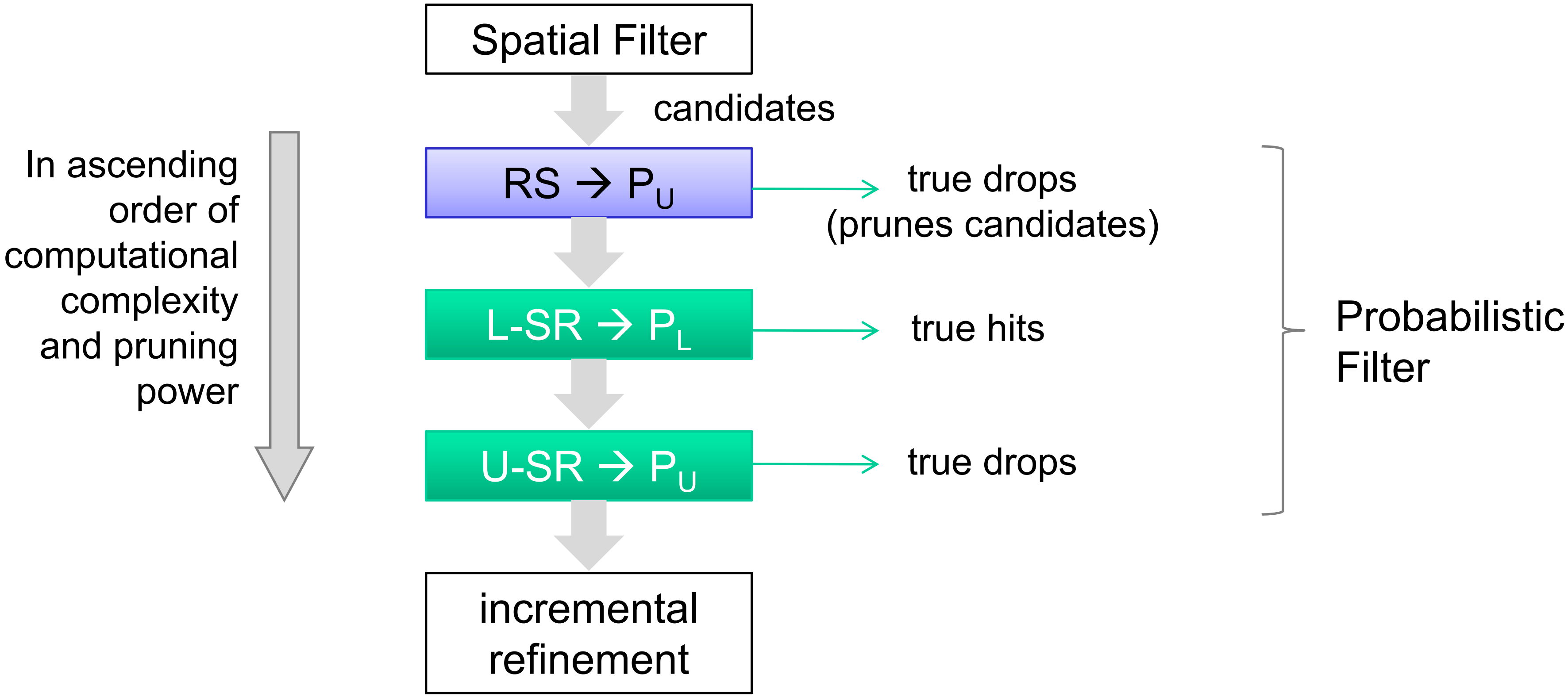
Probabilistic Nearest-Neighbor Query

- Probabilistic Verification: [Cheng08]
  - Supports probabilistic pruning
  - Mainly designed for approaximate queries → avoids expensive evaluation step
  - Basic idea:
    - After projection into distance space, decompose distance space into slots
    - Assume equal distribution within each slot (approximation)
    - Solve the problem slot-wise → avoids expensive integration
    - Apply cascade of spatial and probabilistic verification filters



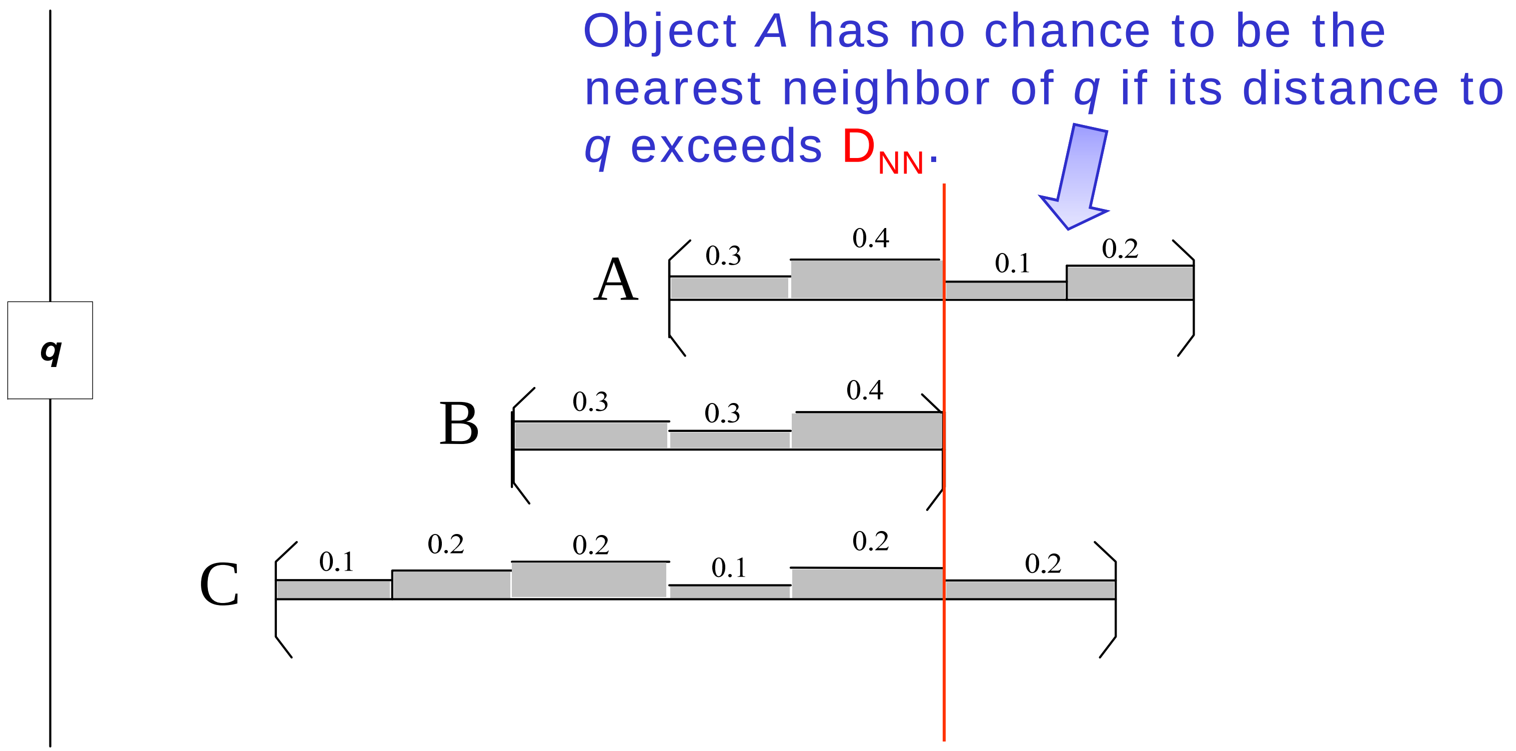
Probabilistic Nearest-Neighbor Query

- Filter-Refinement-Pipeline



Probabilistic Nearest-Neighbor Query

- Rightmost Subregion (RS) Verifier



– Probability that dist(q,A) > D_{NN} = 0.3 => P_U(A) = 1 - 0.3 = 0.7

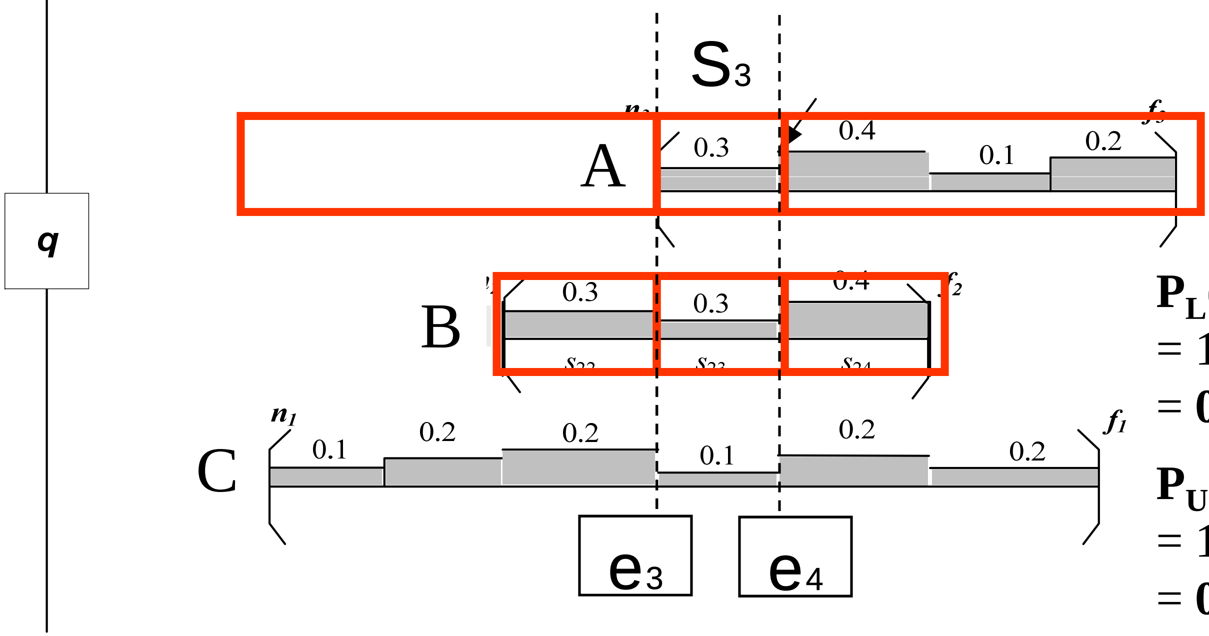
– Probability that dist(q,C) > D_{NN} = 0.2 => P_U(C) = 1 - 0.2 = 0.8

Probabilistic Nearest-Neighbor Query

- L-SR and U-SR Verifiers
  - Assume distance dist(q,X) is within slot S_j, then

$$P_L = \frac{1}{c_j} \prod_{x \neq o} (1 - cdf_{dist(q,x)}(e_j))$$

$$P_U = \frac{1}{2} \left(\prod_{x \neq o} (1 - cdf_{dist(q,x)}(e_j)) + \prod_{x \neq o} (1 - cdf_{dist(q,x)}(e_{j+1})) \right)$$

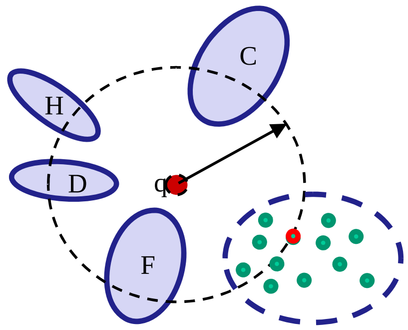


P_L(B | B in S_3)
= 1/3 (1-0)(1-0.5)
= **0.167**

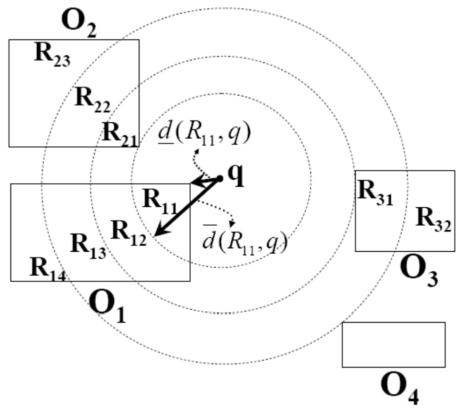
P_U(B | B in S_3)
= 1/2 ((1-0)(1-0.5)+(1-0.3)(1-0.6)
= **0.39**

Probabilistic Nearest-Neighbor Query

- Similar pruning techniques:
  - MC sample-based approach:
    [Kriegel07]
    - PTopkNNQ:
      - Similarity predicate: NN
      - Answer type: object oriented
      - Constraints: PSQ
      - Accuracy: approximate
      - Model: discrete

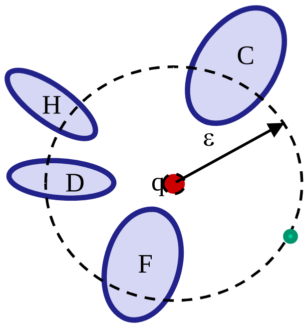


  - Uncertain region decomposition based approach:
    [Beskales08]
    - PTopkNNQ:
      - Similarity predicate: NN
      - Answer type: object oriented
      - Constraints: PTopkSQ
      - Accuracy: exact
      - Model: continuous



Probabilistic k-Nearest-Neighbor Query

- Probabilistic k-NN Queries
  - Pruning concepts used for PNN can be easily transferred to the P*k*NN (k>1) problem
    - Spatial pruning:
      - Prune all behind the k-th smallest maximum distance
    - Probabilistic pruning:
      - Pruning filter in consideration of the k-th smallest maximum distance
  - However, more complex computation of qualification probability than for PNN (P1NN)
    - Example: P(X=NN(q))=P(C,H,D and F outside of ε-range)



P(X=**3**NN(q))=

$$\left.\begin{array}{l}
\text{P(C,H outside of ε-range)·P(D,F inside of ε-range)+} \\
\text{P(C,D outside of ε-range)·P(H,F inside of ε-range)+} \\
\text{P(C,F outside of ε-range)·P(H,D inside of ε-range)+} \\
\text{P(H,D outside of ε-range)·P(C,F inside of ε-range)+} \\
\text{P(H,F outside of ε-range)·P(C,D inside of ε-range)+} \\
\text{P(D,F outside of ε-range)·P(C,H inside of ε-range)}
\end{array}\right\} \text{C^n_k possibilities to be considered}$$

Probabilistic k-Nearest-Neighbor Query

- Probabilistic kNN Query [Cheng09]
  - Query Semantic (Classification):
    - Similarity predicate: kNN (k≥1)
    - Constraints: PTSQ
    - Answer types: result oriented
  - Given:
    - Query object *q*, query parameter *k*, probability threshold τ
  - Search:
    - All *k*-sets (sets of card. k) having probability > τ containing the kNN set of query *q*
  - Pruning Techniques:
    - ***K-bound filtering*** (adaption of bounding-step (PNN)), effectively removes all objects that have no chance to be a query answer.
    - ***Probabilistic Candidate Selection***, efficiently detects unqualified k-sets (i.e. whose qualification probabilities are less than τ)
    - ***Verification***, computes lower and upper bounds of qualification probabilities

Probabilistic k-Nearest-Neighbor Query

- Probabilistic kNN Query [Cheng09]
  - Query Sema
    - Similarity p)
    - Constraint
    - Answer typ ente
  - Given:
    - Query object uery parameter *k*, pr
  - Search:
    - All *k*-sets (s s of card. k) having pr bability > τ of query *q*
  - Pruning Tech niques:
    - ***K-bound filtering*** (adaption of bounding all objects that have no chance to be ery answer.
    - ***Probabilistic Candidate S ection***, efficiently detects unqualified k-sets (i.e. whose qualification probabilities are less than τ)
    - ***Verification***, computes lower and upper bounds of qualification probabilities

Extending spatial pruning concept of PNN queries to PkNN

Extending RS pruning concept of PNN queries to PkNN

Extending L-SR/U-SR pruning concept of PNN queries to PkNN

- Challenge:
  - Coping with the exponential number of possible k-sets
- Basic Idea:
  - Only considering k-sets with qualification probability at least τ
  - A-priori-based k-set generation using monotonity criterion

1-subset	CP
$\{o_1\}$	1
$\{o_2\}$	1
$\{o_3\}$	1
$\{o_4\}$	0.5
$\{o_5\}$	0.2
$\{\mathbf{o_6}\}$	**0.1**

(a) Round 1

2-subset	CP
$\{o_1, o_2\}$	1
$\{o_1, o_3\}$	1
$\{o_1, o_4\}$	0.5
$\{o_1, o_5\}$	0.2
$\{o_2, o_3\}$	1
$\{o_2, o_4\}$	0.5
$\{o_2, o_5\}$	0.2
$\{o_3, o_4\}$	0.5
$\{o_3, o_5\}$	0.2
$\{\mathbf{o_4}, \mathbf{o_5}\}$	**0.1**

(b) Round 2

3-subset	CP
$\{o_1, o_2, o_3\}$	1
$\{o_1, o_2, o_4\}$	0.5
$\{o_1, o_2, o_5\}$	0.2
$\{o_1, o_3, o_4\}$	0.5
$\{o_1, o_3, o_5\}$	0.2
$\{\mathbf{o_1}, \mathbf{o_4}, \mathbf{o_5}\}$	**0.1**
$\{o_2, o_3, o_4\}$	0.5
$\{o_2, o_3, o_5\}$	0.2
$\{\mathbf{o_2}, \mathbf{o_4}, \mathbf{o_5}\}$	**0.1**
$\{\mathbf{o_3}, \mathbf{o_4}, \mathbf{o_5}\}$	**0.1**

(c) Round 3

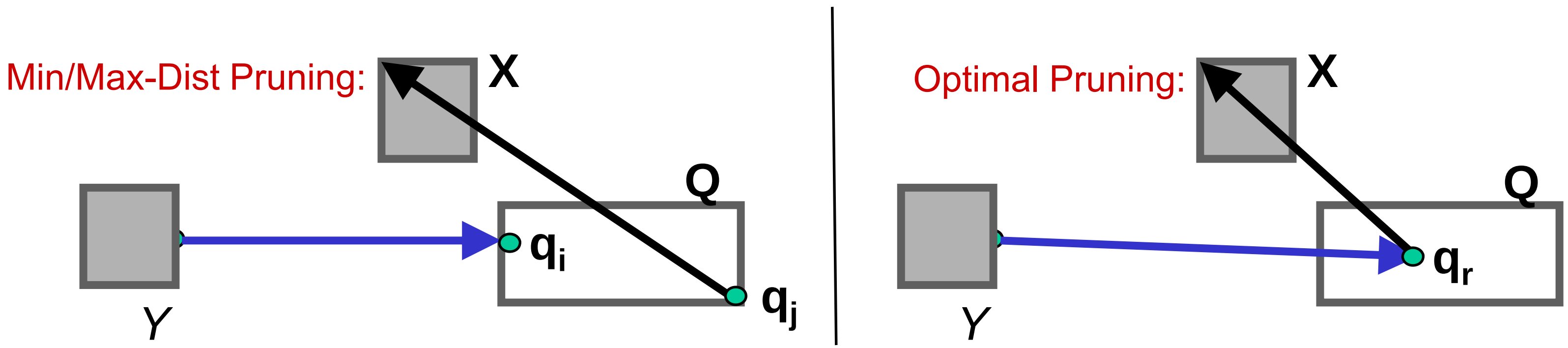
Step-by-step generating candidate subsets based on CP

Probabilistic k-Nearest-Neighbor Query

- Pruning with Uncertain Query Objects
  - Up to now, we only considered certain query object.
  - Most of the proposed concepts can be easily extended to cope with uncertain query objects, e.g.:
    - Filter: Adapting the spatial and probabilistic filter accordningly
    - Refinement: Integration over both query and database object
  - The approaches *Projection-Pruning-Bounding* and *Probabilistic Verification* already can cope with uncertain query objects since it is solved in the projection step (object space → distance space)
  - In the following:
    - a novel spatial pruning method for uncertain objects
    - more effective but not more expensive than traditional spatial pruning techniques
    - Universally applicable for many approaches using spatial pruning techniques

Probabilistic k-Nearest-Neighbor Query

- Enhancing Spatial Pruning: [Emrich10]
  - Up to now, most (spatial/probabilistic)-pruning approaches are based on min/max-distance comparisons
  - Min/max-distance pruning ignores dependency between distances

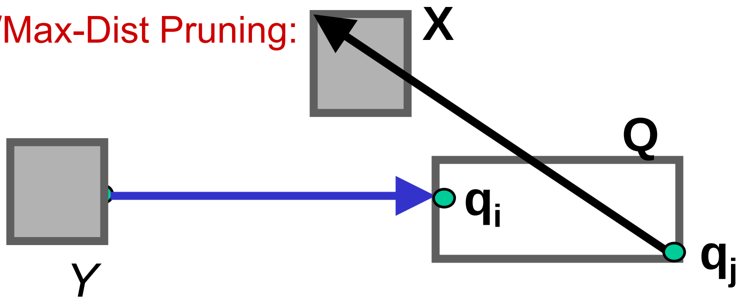


Min-dist(Y,Q) > Max-dist(X,Q)
$\Rightarrow$ P(X closer to Q than Y) = 1

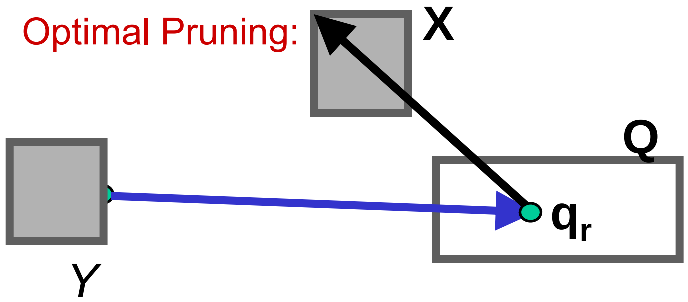
Probabilistic k-Nearest-Neighbor Query

- Enhancing Spatial Pruning: [Emrich10]
  - Up to now, most (spatial/probabilistic)-pruning approaches are based on min/max-distance comparisons
  - Min/max-distance pruning ignores dependency between distances
  - Generally, taking distance dependencies into account is very expensive (distance check for all possible locations of query object Q)

Min/Max-Dist Pruning:



Optimal Pruning:



Min-dist(Y,Q) > Max-dist(X,Q)
$\Rightarrow$ P(X closer to Q than Y) = 1

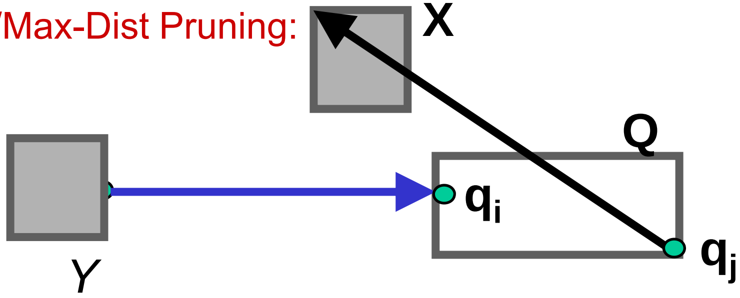
~~Min-dist(Y,Q) > Max-dist(X,Q)~~
~~$\Leftarrow$ P(X closer to Q than Y) = 1~~

$\forall q_r \in Q$: Min-dist(Y,q_r) > Max-dist(X,q_r)
$\Leftrightarrow$ P(X closer to Q than Y) = 1

Probabilistic k-Nearest-Neighbor Query

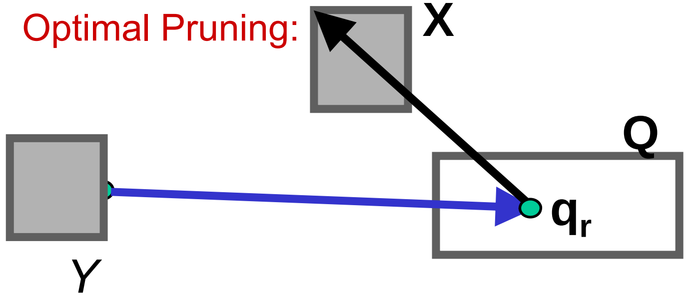
– Enhancing Spatial Pruning: [Emrich10]
  - Up to now, most (spatial/probabilistic)-pruning approaches are based on min/max-distance comparisons
  - Min/max-distance pruning ignores dependency between distances
  - Generally, taking distance dependencies into account is very expensive (distance check for all possible locations of query object Q)

Min/Max-Dist Pruning:



Min-dist(Y,Q) > Max-dist(X,Q)
$\Rightarrow$ P(X closer to Q than Y) = 1

Optimal Pruning:



$\forall q_r \in Q$: Min-dist(Y,q_r) > Max-dist(X,q_r)
$\Leftrightarrow$ P(X closer to Q than Y) = 1
$\Leftrightarrow$

$$\sum_{i=1}^{d} \max_{q_i \in \{Q_i^{\min}, Q_i^{\max}\}} \left(MaxDist(X_i, q_i)^2 - MinDist(Y_i, q_i)^2\right) < 0$$

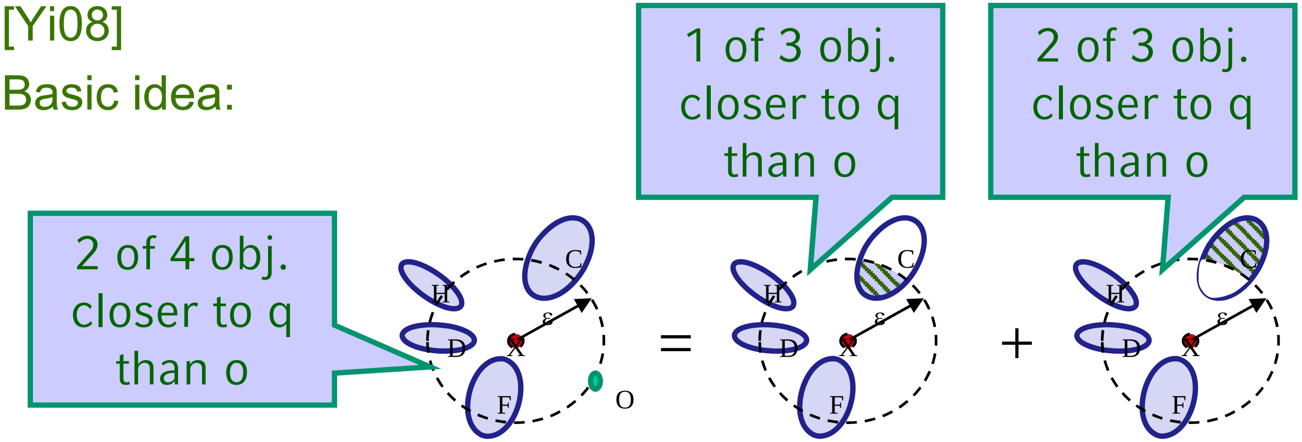
[Emrich10]

Probabilistic Ranking Query

- Probabilistic Ranking Query [Bernecker10]
  - Query Semantic (Classification):
    - Similarity predicate: ranking
    - Constraints: PSQ
    - Answer types: object oriented
  - Given:
    - Query object *q*
  - Search:
    - For all objects o and all ranking positions k report the probability that o is ranked at k w.r.t. distance to query *q*
    - General method which can be used to build prob. ranking results according to different query semantics, e.g. uk-ranks, expected rank [LI09]
  - Pruning Techniques:
    - K-bound-based spatial pruning
    - No probabilistic pruning since there is no constraint on the qualification probability

Probabilistic Ranking Query

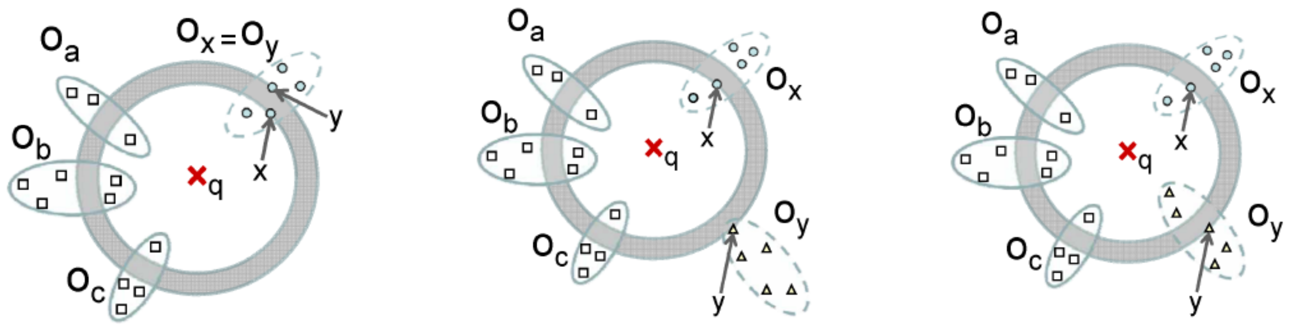
- Challenge:
  - For each object and each k, compute the probability that (k-1) objects are closer to q than o.
  - There are exponential number of possible (k-1)-sets to be considered
  - Very expensive computation
- 1. Solution:
  - Binomial recurrence technique firstly proposed for a similar problem in [Yi08]
  - Basic idea:



  - Enables to compute the probabilistic ranking for the first k ranking positions in $O(n^2)$ time

Probabilistic Ranking Query

– 2. Solution:
  - Iterative binomial recurrence technique as proposed in [Bernecker10] (equivalent to approach proposed in [LI09] based on generating functions)



(a) Case 1: Instances (x, p_x) and (y, p_y) belong to the same object.

(b) Case 2: Instance (y, p_y) is the first returned instance of object o_Y.

(c) Case 3: Instance (y, p_y) is not the first returned instance of object o_Y.

  - Enables to compute the probabilistic ranking for the first k ranking positions in O(n) time
  - Similar techniques proposed in [Zhang10] (prob. Pruning, interative refinement)

Other Probabilistic Similarity Search Approaches

- Approaches for advanced probabilistic similarity search queries based on similar spatial/probabilistic pruning concepts:
  – Probabilistic Reverse k-NN Queries:
    - Lian, Chen: *Efficient processing of probabilistic reverse nearest neighbor queries over uncertain data*, VLDB Journal 2009 (optimal pruning based on spherical object approximations)
    - Cheema, Lin, Wang: *Probabilistic Reverse Nearest Neighbor Queries on Uncertain Data*, TKDE 2010 (partial pruning concept)
  – Probabilistic Inverse Ranking:
    - [Lian09], [Lian10a]
  – Probabilistic (Reverse) Skyline Queries:
    - [Lian10b] [Lian08]
  – Probabilistic Top-k Dominating Queries:
    - [Zhang10], [Lian09]

Outline

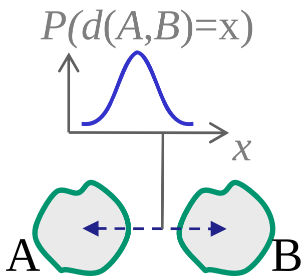
- Introduction
  - Motivation
  - Uncertain Data Modelling
  - Challenges
- Similarity Search in Uncertain Data
  - Probabilistic Similarity Queries: Overview and Classification
  - Probabilistic ε-Range, NN, kNN and Ranking Queries
- Mining Uncertain Data
- Summary

Mining Uncertain Data

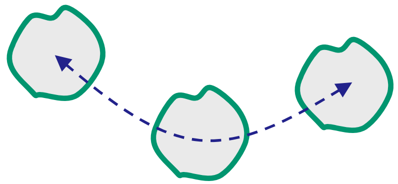
- Mining Uncertain Data
  - Recently, a number of mining applications for uncertain data have been proposed, including
    - Clustering
    - Frequent pattern mining
    - Classification
  - In the following: Selection of mining applications efficiently supported by similarity search techniques
  - Broad and detailed overviews are given in
    - [Aggarwal09]
    - [Tutorial: J. Pei, M. Hua, Y. Tao, and X. Lin, KDD'08]

Clustering Uncertain Data

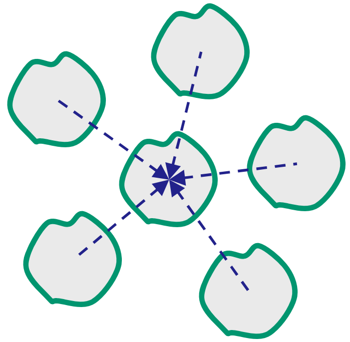
– Clustering Uncertain Data
  - Clustering methodologies are affected by uncertain distances between data points
  - Example 1: Density based clustering
    – FDBSCAN: probabilistic extension of DBSCAN [Kriegel05]



probabilistic distance



reachability probability

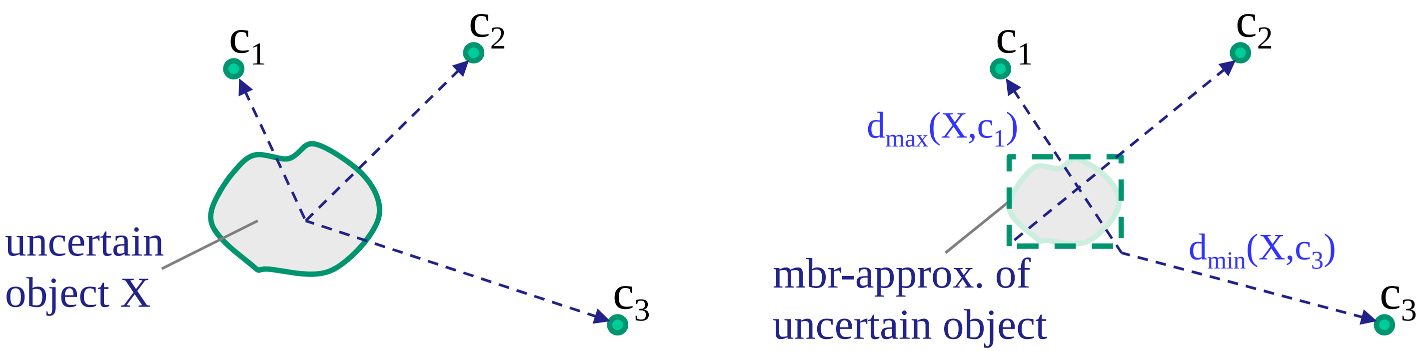


core object probability

    – $P^{reach}(p,o) = P^{core}(o) \cdot P(d(p,o) \leq \varepsilon)$
    – p assumed to be reachable from o, iff $P^{reach}(p,o) \geq 0.5$
  - Similarity search techniques can be applied here!

Mining Uncertain Data

- Example 2: Data partitioning based clustering
  - UK-means: probabilistic extension of k-means [Ngai06]



  - Distance probability distributions between object and cluster center
  - Object-cluster-assignment based on expected distances
  - Applying spatial pruning strategy as a filter step:
    - » object X cannot be assigned to cluster center c_3 because $d_{max}(X,c_1)<d_{min}(X,c_3)$

Mining Uncertain Data

- Frequent Pattern Mining of Uncertain Data
  - Given:
    - Set / Stream of uncertain transactions
    - An uncertain transaction consists of a set of uncertain items
    - Each uncertain item is associated with an existential probability value
  - Methods:
    - probabilistic extension of
      - » Apriori (U-Apriori) [Chui07]
      - » FP-growth (UFP-tree), H-Mine [Aggarwal09]

      Above approaches are based on expected support
    - Probabilistic FIM: based on probabilistic support

ID	Transaction
t_1	(A,0.8);(B,0.2);(D,0.5);(F,1.0)
t_2	(B,0.1);(C,0.7);(D,1.0);(E,1.0);(G,0.1)
t_3	(A,0.5);(D,0.2);(F,0.5);(G,1.0)
t_4	(D,0.8);(E,0.2);(G,0.9)
t_5	(C,1.0);(D,0.5);(F,0.8);(G,1.0)
t_6	(A,1.0);(B,0.2);(C,0.1)

Mining Uncertain Data

- Example 4: Probabilistic FIM [Bernecker09]

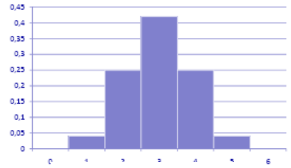
  – Query:
    » Itemsets having a high probability to be frequent, i.e. {itemset *is*: P(sup(*is*)≥sup_{min})} > τ

  – Basic idea:
    » Itemset generation: adaptation of Apriori
    » Efficient computation of P(sup(*is*)≥sup_{min}) (in linear time)
    » Prefer generation of most significant itemsets (best-first)

  – Properties:
    » Allows us to compute and report itemsets in an ite
    » Reports the itemsets in decreasing order of their significance

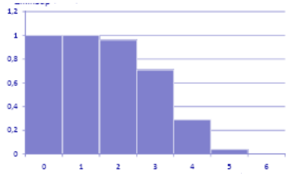
ID	Transaction
t_1	(A, 0.8) ; (B, 0.2) ; **(D, 0.5)** ;
t_2	(B, 0.1) ; (C, 0.7) ; **(D, 1.0)** ;
t_3	(A, 0.5) ; **(D, 0.2)** ; (F, 0.5) ;
t_4	**(D, 0.8)** ; (E, 0.2) ; (G, 0.9)
t_5	(C, 1.0) ; **(D, 0.5)** ; (F, 0.8) ;
t_6	(A, 1.0) ; (B, 0.2) ; (C, 0.1)

$P_i(D)$



support i

$P_{\geq minSup}(D)$



☺ minSup

Recent Development of FIM

- Explore the relationship between the support distribution of size-i itemset and size-(i+1) itemset [Sun10]
  - Develop top-down approach, where the maximal-size itemsets are discovered, before finding out smaller ones
  - Use Fourier Transform to speed up the process
  - Achieve an order of magnitude improvement (from $O(n^2)$ to $O(n \log n)$)
- Develop *model-based* approach for deriving support distribution with high accuracy [Wang10]
  - Support distribution can be modeled by a Poisson binomial distribution, which can be approximated with a Poisson distribution
- Efficient computation of *probabilistic association rules*, which are derived from FIM [Sun10]

Mining Uncertain Data

- Classification of uncertain data
- Main goal: Improve accuracy of traditional learning algorithms for handling uncertain data
- Method 1: Develop decision tree classifier [Tsang09]
  - Redefine split-points based on distributions of attributes
- Method 2: Develop Naïve Bayes classifiers [Ren09]
  - Extend the class conditional probability estimation in the Bayes model to handle pdf's.

Outline

- Introduction
  - Motivation
  - Uncertain Data Modelling
  - Challenges
- Similarity Search in Uncertain Data
  - Probabilistic Similarity Queries: Overview and Classification
  - Probabilistic ε-Range, NN, kNN and Ranking Queries
- Mining Uncertain Data
- Summary

Summary

- Traditional query and analysis tasks have to be enhanced or redeveloped, in order to handle uncertainty in data
- Probabilistic similarity search is a key component in many data mining and pattern recognition tasks for uncertain data
- We provide a classification of probabilistic similarity search queries, as well as their evaluation techniques

Future Research Directions

- Study the semantics and evaluation of different similarity measures for uncertain data
- Investigate similarity evaluation for more complex uncertain data models (e.g., joint distribution of attributes)
- Investigate mining of other types of data where uncertainty exists (e.g., trajectories)

References

- **[Yiu07]** M.L. Yiu and N. Mamoulis: *Efficient processing of top-k dominating queries on multi-dimensional data.* In Proc. VLDB'07
- **[Zhang10]** W. Zhang, X. Lin, Y. Zhang, J. Pei, W. Wang: *Threshold-based Probabilistic Top-k Dominating Queries*. VLDB Journal 2010
- **[Papadias03]** D. Papadias, Y. Tao, F. Greg, and B. Seeger: *Progressive skyline computation in database systems.* ACM TODS 2003
- **[Ljosa07]** VebjornLjosaand AmbujK. Singh. Apla: Indexing arbitrary probability distributions. In ICDE, pages 946–955, 2007.
- **[Tao05]** YufeiTao, ReynoldCheng, XiaokuiXiao, Wang Kay Ngai, Ben Kao, and Sunil Prabhakar. Indexing multi-dimensional uncertain data with arbitrary probability density functions. In VLDB, pages 922–933, 2005.
- **[Yi08]** Ke Yi, Feifei Li, George Kollios, Divesh Srivastava: Efficient Processing of Top-k Queries in Uncertain Databases. ICDE 2008
- **[Emrich10]** Tobias Emrich, Hans-Peter Kriegel, Peer Kröger, Matthias Renz, Andreas Züfle: Boosting spatial pruning: on optimal pruning of MBRs. SIGMOD 2010
- **[LI09]** Jian Li, Barna Saha, Amol Deshpande: A Unified Approach to Ranking in Probabilistic Databases. VLDB 2009
- **[Sistla98]** P. A. Sistla, O. Wolfson, S. Chamberlain, and S. Dao. Querying the uncertain position of moving objects. In Temporal Databases: Research and Practice. Springer Verlag, 1998.
- **[Pfoser99]** D. Pfoser and C. Jensen. Capturing the uncertainty of moving-objects representations. In SSDBM, 1999.
- **[Cheng03]** R. Cheng, D. Kalashnikov, and S. Prabhakar. Evaluating probabilistic queries over imprecise data. In Proc. ACM SIGMOD, 2003.

References (Attribute Uncertainty)

- **[Cheng04a]** R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J. S. Vitter. Efficient indexing methods for probabilistic threshold queries over uncertain data. In VLDB, 2004.
- **[Cheng04b]** R. Cheng, D. Kalashnikov and S. Prabhakar. *Querying Imprecise Data in Moving Object Environments. In IEEE Transactions on Knowledge and Data Engineering (IEEE TKDE), Vol. 16, No. 9, pp. 1112-1127, Sep 2004.*
- **[Desphande04]** A. Deshpande, C. Guestrin, S. Madden, J. Hellerstein, and W. Hong. Model-driven data acquisition in sensor networks. In VLDB, 2004.
- **[Cheng07]**
- **[Tao05]**Y. Tao, R. Cheng, X. Xiao, W. K. Ngai, B. Kao, and S. Prabhakar. Indexing multi-dimensional uncertain data with arbitrary probability density functions. In VLDB, 2005.
- **[Tao07]** Y. Tao, X. Xiao and R. Cheng. *Range Search on Multidimensional Uncertain Data. In ACM Transactions on Database Systems (TODS). 32(3):15, Aug 2007.*
- **[Pei07]** J. Pei, B. Jiang, X. Lin, and Y. Yuan. Probabilistic skylines on uncertain data. In VLDB, 2007.
- **[ICDE06]** A. Silberstein, R. Braynard, C. Ellis, K. Munagala, and J. Yang. A sampling-based approach to optimizing top-k queries in sensor networks. In ICDE, 2006.
- **[Kriegel06]** H. Kriegel, P. Kunath, M. Pfeifle and M. Renz. Probabilistic Similarity Join on Uncertain Data. In DASFAA, 2006
- **[Kriegel07]** H. Kriegel, P. Kunath, and M. Renz. Probabilistic nearest-neighbor query on uncertain objects. In DASFAA, 2007.
- **[Ljosa07]** V. Ljosa and A. K. Singh, “APLA: Indexing arbitrary probability distributions,” in *Proc. ICDE*, 2007.
- **[Cheng08a]** R. Cheng, J. Chen, M. Mokbel, and C. Chow. Probabilistic verifiers: Evaluating constrained nearest-neighbor queries over uncertain data. In ICDE, 2008.
- **[Singh08]** S. Singh et al. Database support for pdf attributes. In ICDE 2008.
- **[Lian08]** X. Lian and L. Chen. Monochromatic and bichromatic reverse skyline search over uncertain databases. In SIGMOD, 2008.

References (Attribute Uncertainty)

- **[Beskales08]** Efficient Search for the Top-k Probable Nearest Neighbors in Uncertain Databases. George Beskales, Mohamed A. Soliman, Ihab F. Ilyas. In VLDB 2008.
- **[Wang08]** BayesStore: Managing Large, Uncertain Data Repositories with Probabilistic Graphical Models. D. Wang, E. Michelakis, M. Garofalakis, J. Hellerstein. In VLDB, 2008.
- **[Chen09]** J. Chen, R. Cheng, M. Mokbel and C. Chow. Scalable Processing of Snapshot and Continuous Nearest- Neighbor Queries over One-Dimensional Uncertain Data. In Very Large Databases Journal (VLDBJ), Special Issue on Uncertain and Probabilistic Databases, Vol. 18, No. 5, pp. 1219-1240, 2009. (Awarded the Research Output Price in Department of Computer Science, Faculty of Engineering, HKU, 2010)
- **[Cheng10a]** R. Cheng, X. Xie, M. Y. Yiu, J. Chen and L. Sun. UV-diagram: A Voronoi Diagram for Uncertain Data. In the IEEE Intl. Conf. on Data Engineering (IEEE ICDE 2010), Long Beach, USA, Mar, 2010.
- **[Barbara92]** D. Barbara, H. Garcia-Molina, and D. Porter. The management of probabilistic data. Volume: 4, Issue: 5, page(s): 487-502, TKDE 1992.
- **[Dalvi04]** N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. In VLDB, 2004
- **[Agrawal06]** P. Agrawal, O. Benjelloun, A. D. Sarma, C. Hayworth, S. Nabar, T. Sugihara, and J. Widom. Trio: A system for data, uncertainty, and lineage. In VLDB, 2006.
- **[Benjelloun06]** O. Benjelloun, A. Sarma, A. Halevy, and J. Widom. ULDBs: Databases with uncertainty and lineage. In VLDB, 2006.
- **[Soliman07]** M. Soliman, I. Ilyas, and K. Chang. Top-k query processing in uncertain databases. In ICDE 2007.
- **[Re07]** C. Re, N. Dalvi, and D. Suciu. Efficient top-k query evaluation on probabilistic data. In ICDE, 2007.
- **[Sarawagi06]** S. Sarawagi. Creating Probabilistic databases with information extraction models. In VLDB 2006.
- **[Singh07]** S. Singh, C. Mayfield, S. Prabhakar, R. Shah and S. Hambrusch. Indexing uncertain categorical data. In ICDE 2007.

References (Attribute Uncertainty)

- **[Sen07]** P. Sen and A. Deshpande. "Representing and Querying Correlated Tuples in Probabilistic Databases". In Proc. ICDE, 2007.
- **[Antova08]** L. Antova, T. Jansen, C. Koch, and D. Olteanu. "Fast and Simple Relational Processing of Uncertain Data". In Proc. ICDE, 2008.
- **[Yi08]** K. Yi, F. Li, D. Srivastava and G. Kollios. Efficient processing of top-k queries in uncertain databases. In ICDE 2008.
- **[Jin08]** Sliding-Window Top-k Queries on Uncertain Streams. C. Jin, K. Yi, L. Chen, J. Yu, X. Lin.
- **[Cheng08b]** R. Cheng, J. Chen and X. Xie. Cleaning Uncertain Data with Quality Guarantees. In Very Large Databases Conf. (VLDB 2008), New Zealand, Aug 2008.
- **[Cheng10b**] R. Cheng, E. Lo, X. Yang, M. Luk, X. Li and X. Xie. Explore or Exploit? Effective Strategies for Disambiguating Large Databases. In Very Large Databases Conf. (VLDB 2010), Singapore, Sep, 2010.
- **[Cheng10c]** R. Cheng, J. Gong, and D. Cheung. Managing Uncertainty of XML Schema Matching. In the IEEE Intl. Conf. on Data Engineering (IEEE ICDE 2010), Long Beach, USA, Mar, 2010
- **[Sun10]** L. Sun, R. Cheng, D. W. Cheung, and J. Cheng. Mining Uncertain Data with Probabilistic Guarantees. In the 16th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (ACM SIGKDD 2010), Washington D.C., USA, Jul, 2010 (Full paper).
- **[Wang10]** L. Wang, D. Cheung, R. Cheng and S. Lee. Accelerating Probabilistic Frequent Itemset Mining: A Model-Based Approach. In the ACM 19thConf. on Information and Knowledge Management (ACM CIKM 2010), Toronto, Canada, Oct 2010.
- **[Tsang09]** S. Tsang, B. Kao, K. Y. Yip, W. Ho, and S. D. Lee. Decision Trees for Uncertain Data. In ICDE, 2009.

References (Attribute Uncertainty)

- **[Ren09]** J. Ren, S. D. Lee, X. Chen, B. Kao, R. Cheng, and D. Cheung. Naïve Bayes Classification of Uncertain Data. In the IEEE Intl. Conf. on Data Mining (IEEE ICDM 2009), Miami, USA, Dec, 2009.
- **[Lee07]** S. Lee, B. Kao and R. Cheng. Reducing UK-means to K-means. In the 1st Workshop on Data Mining of Uncertain Data (DUNE), co-located with the IEEE Conf. on Data Mining (IEEE ICDM 2007), USA, Oct, 2007.
- **[Ngai06]** J. Ngai, B. Kao, C. Chui, R. Cheng, M. Chau and K. Yip. Efficient Clustering of Uncertain Data. In the IEEE Intl. Conf. on Data Mining (IEEE ICDM 2006), Hong Kong, Dec, 2006.
- **[Chau06]** M. Chau, R. Cheng, B. Kao and J. Ng. Uncertain Data Mining: An Example in Clustering Location Data. In the Methodologies for Knowledge Discovery and Data Mining, Pacific-Asia Conference (PAKDD 2006)