Modelling and Querying Uncertain Spatio- Temporal Data

Tobias Emrich

joint work with
Andreas Züfle, Matthias Renz, Johannes Niedermayer, Hans-Peter Kriegel, Nikos Mamoulis, Lei Chen

Overview

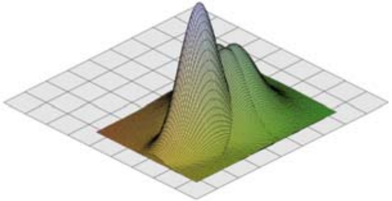
1. Uncertainty in Databases

2. Uncertain Spatio-Temporal Data
   1. Modelling UST Data
   2. Querying UST Data
   3. Follow Up Works

3. Future Directions

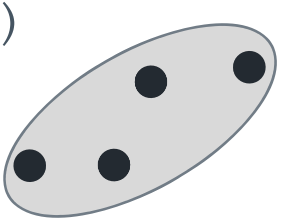
1. Uncertainty in Databases

Motivation [1]



› Uncertainty is inherent in many datasets:

- Automated Extraction of Information from HTML (i.e. John works at Google vs. John works at Microsoft)
- **Sensor Readings (i.e. RFID sensors tracking the position)**
- Human Readings (i.e. the seen Bird was either a Raven (75%) or a Crow (25%))
- Data Integration/Entity Resolution (i.e. do „John Doe“ and „J. Doe“ refer to the same person?)
- …

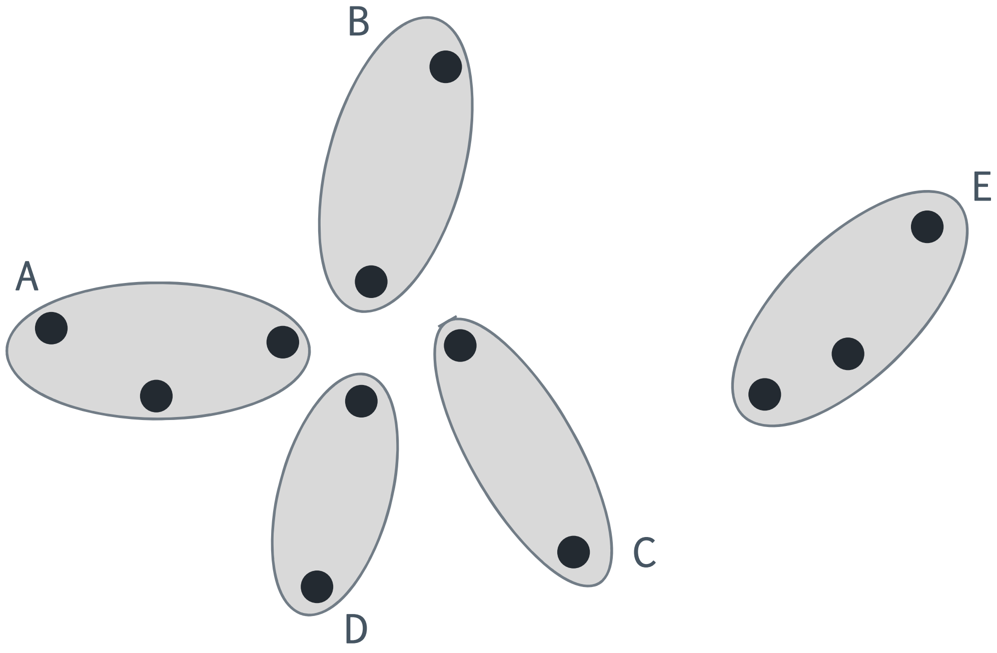


› Two approaches to solve this

- Cleaning (e.g. get rid of uncertainty)
- Management (e.g. handle the uncertainty)

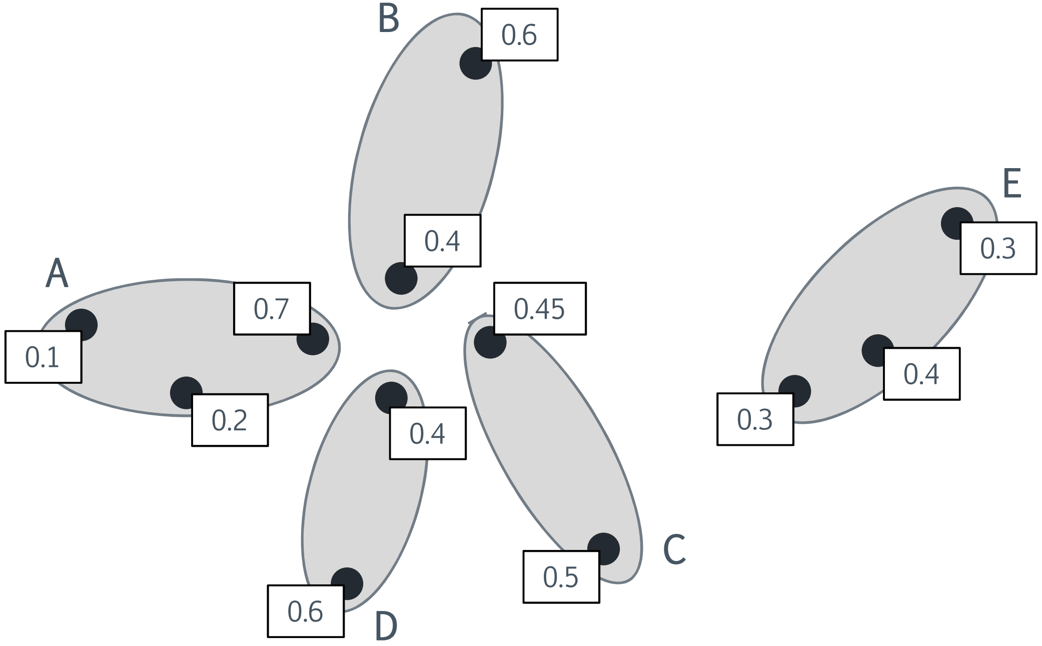
Example

› A spatial (discrete) uncertain Database may look like this



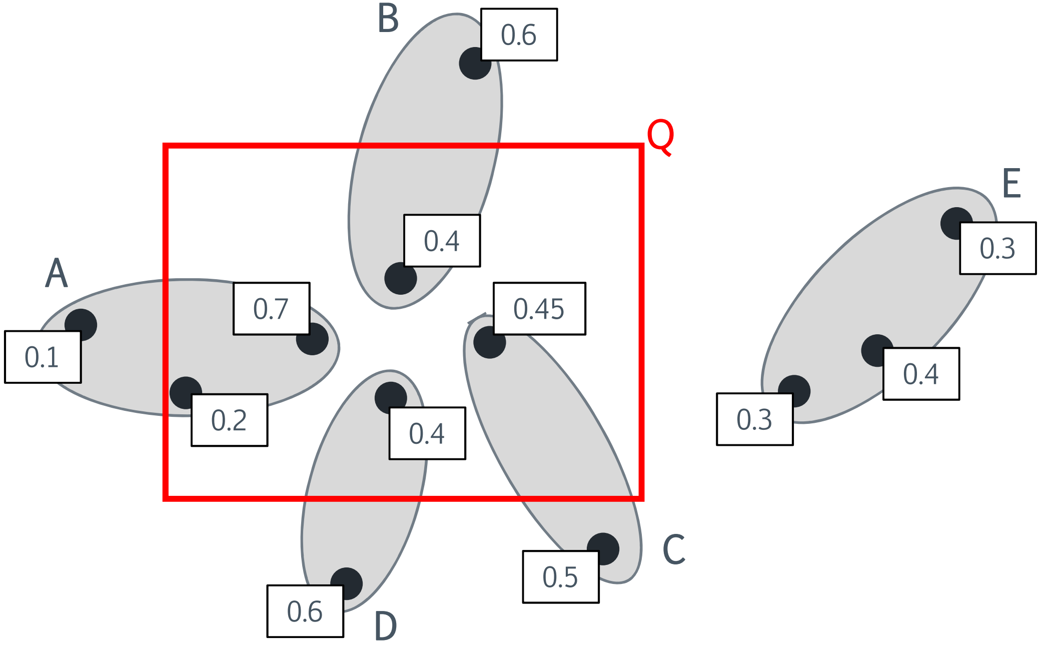
Example

› A spatial (discrete) uncertain Database may look like this



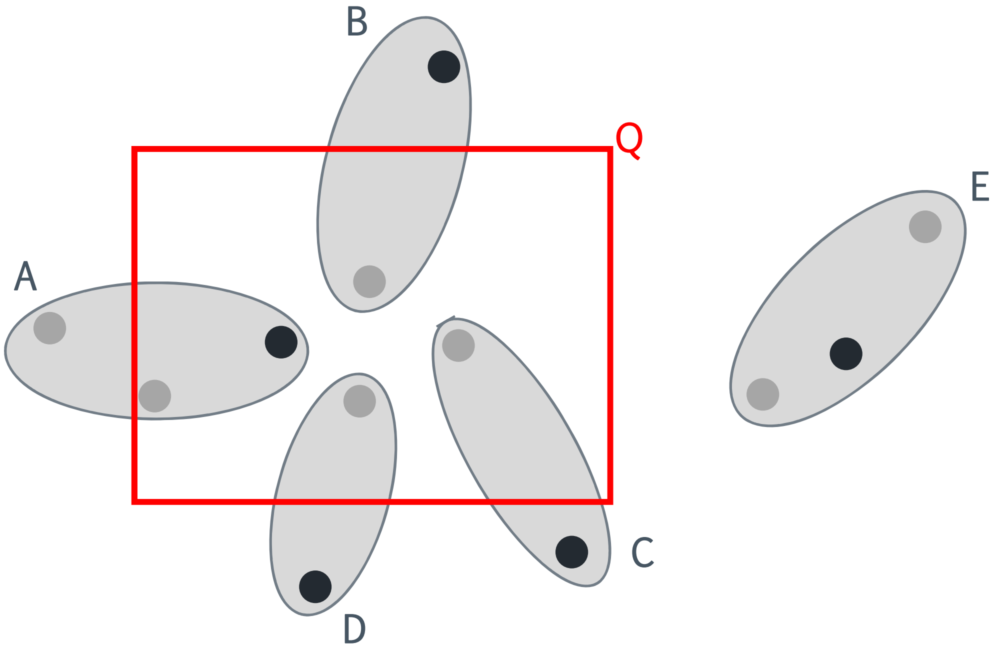
Example

- How many objects are in the query region?



Example

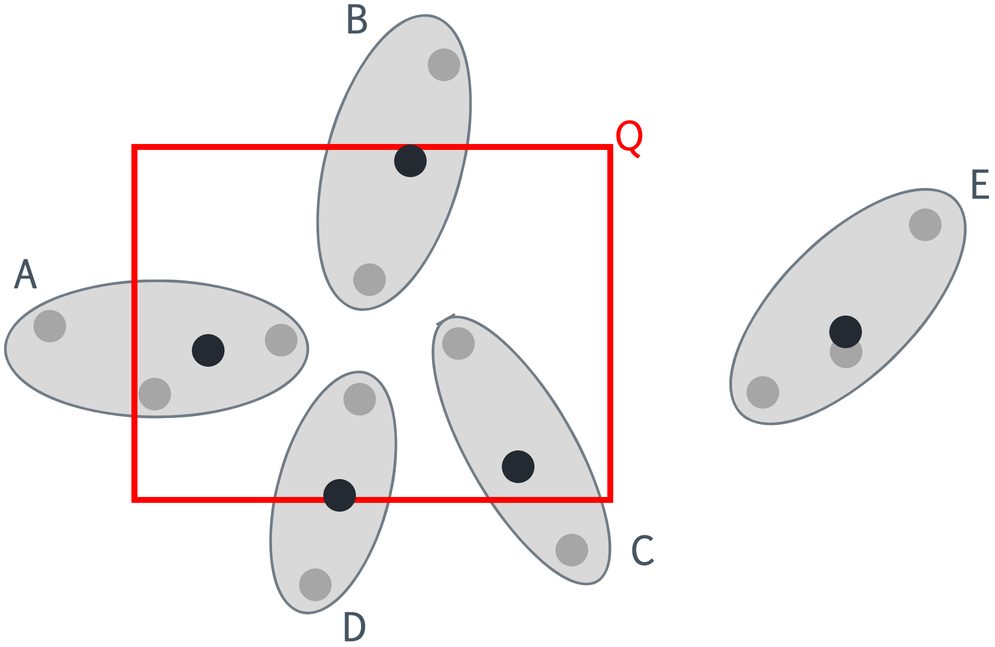
- **Cleaning** (take the most probable position)



Result = 1

Example

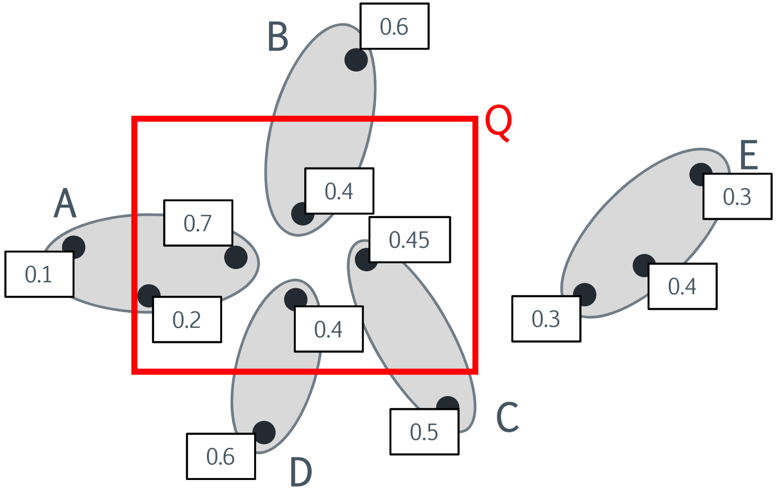
- **Cleaning** (take the expected position)

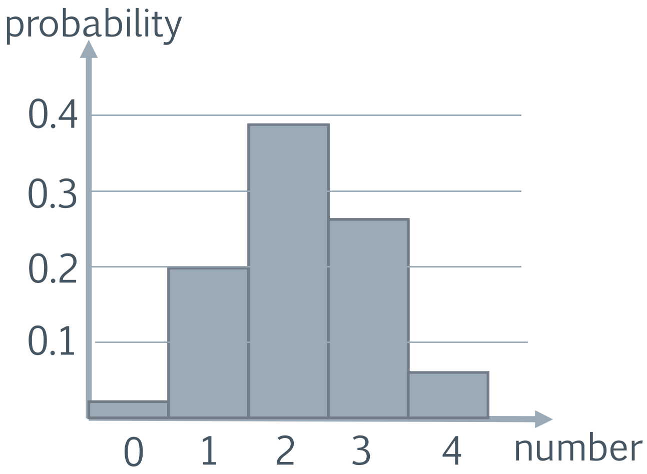


Result = 4

Example

- Managing considers all possible database instances (worlds)

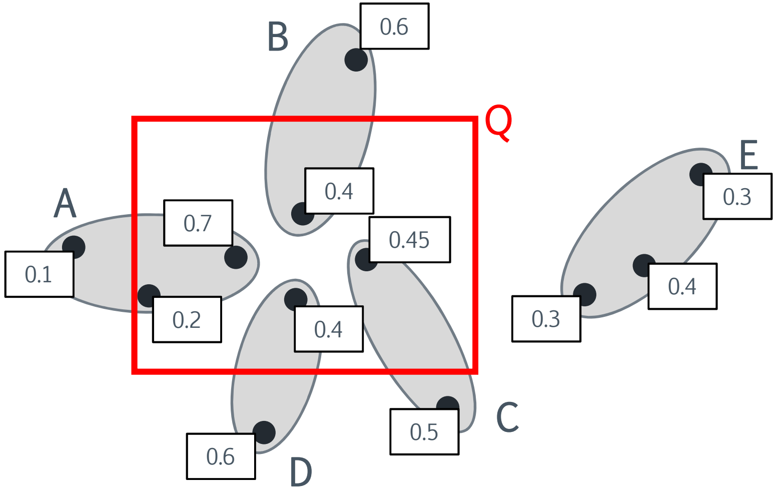


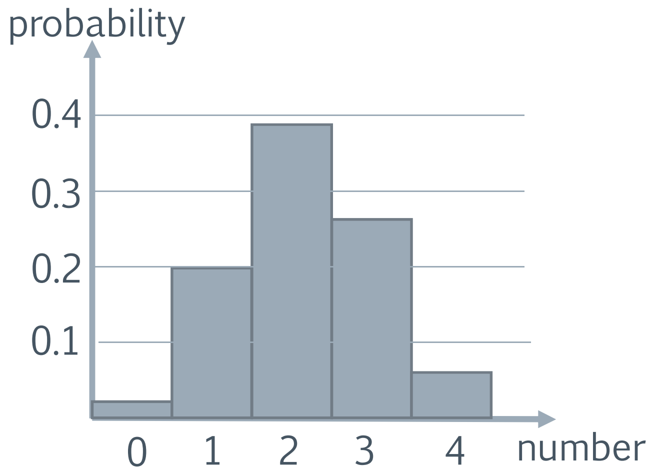


- We get all possible results together with a probability
- But there is an exponential number of possible worlds!

Example

- New efficient techniques have to be developed





- Generating Functions [2] solve this problem efficiently
- $$(0.9x \overset{A}{+} 0.1)(0.4x \overset{B}{+} 0.6)(0.45x \overset{C}{+} 0.5)(0.4x \overset{D}{+} 0.6)(0x \overset{E}{+} 1)= 0.0648\, x^4 +0.2736\, x^3 + 0.3914\, x^2 +0.2022\, x + 0.018$$

Goal of this research

- Jennifer Widom and others brought uncertainty in databases to the attention of the research community in ~2004

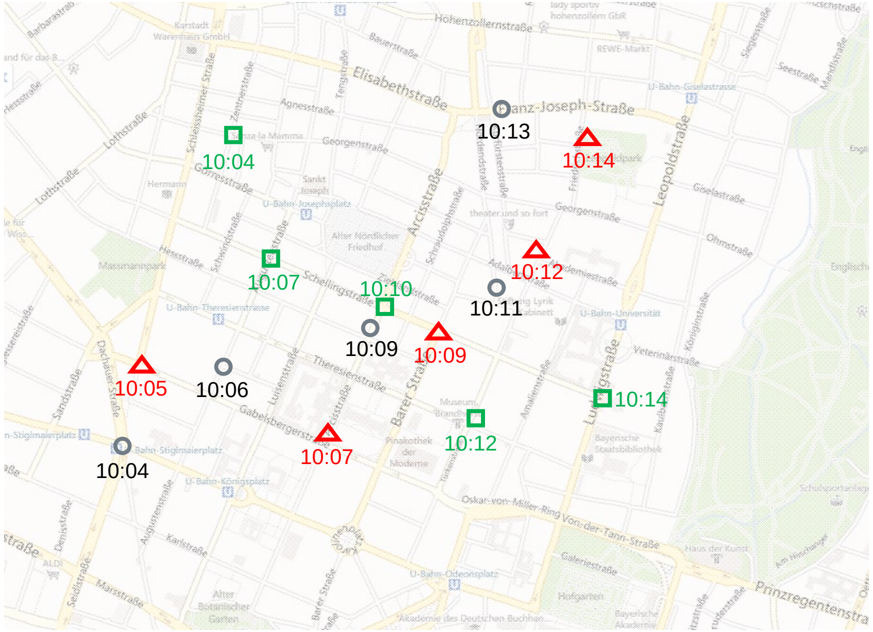
- Uncertainty has now been a hot topic for quite a while and many great ideas have been proposed to handle uncertainty **efficiently(!)**

- It's time to apply the lessons we learned to the area of spatio-temporal data where uncertainty was considered ~1998 by O. Wolfson, C. Jensen, D. Pfoser and others

2. Uncertain Spatio-Temporal Data

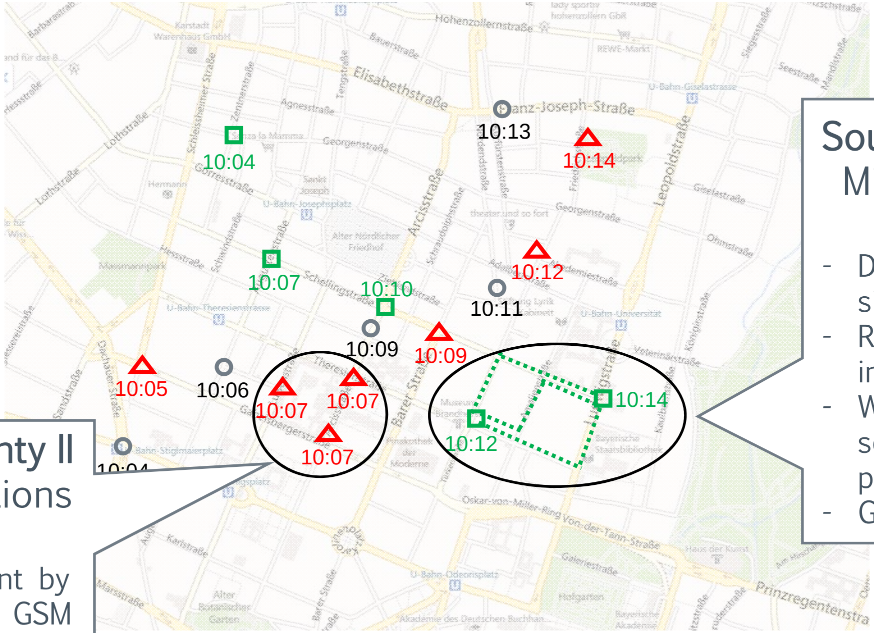
What is Uncertain in ST Data?

- Spatio-Temporal Data usually looks somehow like this



What is Uncertain in ST Data?

› But there are sources of uncertainty



Source of Uncertainty I
Missing Observations

- Delays between GPS signals
- RFID sensors located only in certain locations
- Wireless sensor nodes sending infrequently to preserve power
- Geo-application check-ins

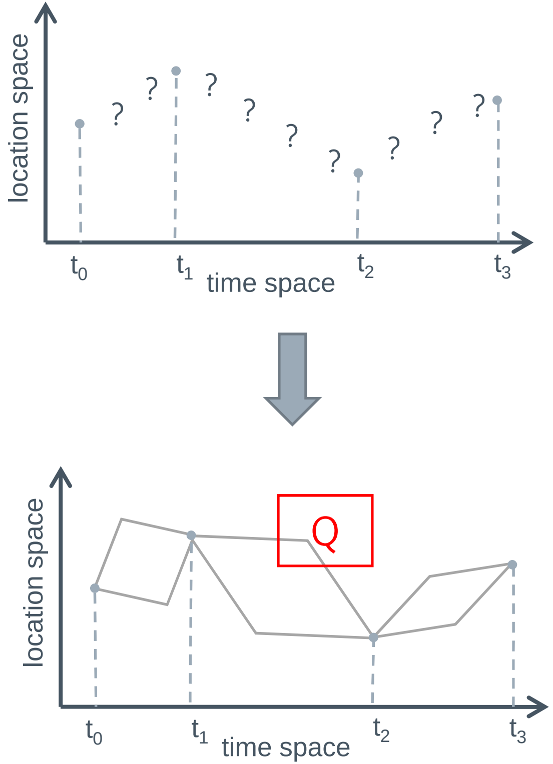
Source of Uncertainty II
Imprecise Observations

- Inexact Measurement by sensor devices e.g. GSM Triangulation
- Human errors e.g. Uncertain observations

Solutions

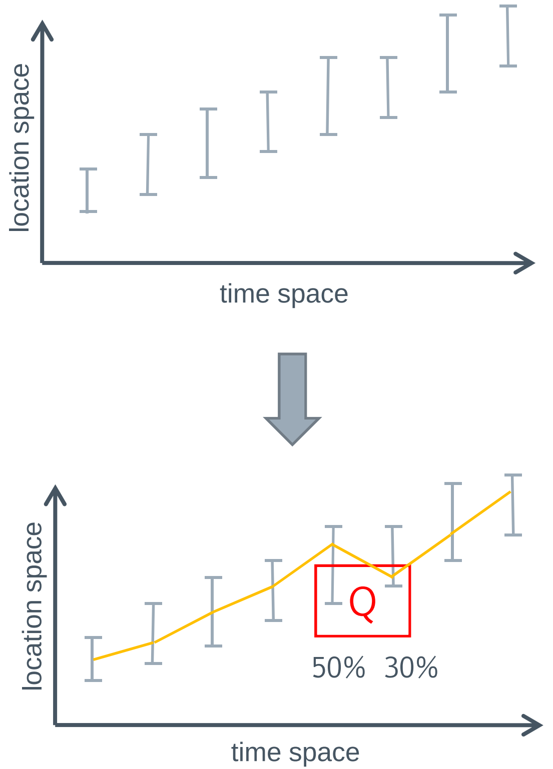
› Missing Observations

- Bound the set of possible (location,time) pairs of an object between **observations** by using spatio-temporal approximations (**diamonds**)
- e.g. by modeling knowledge about maximum speed
- Allows to make statements like „its possible that o intersects some query window Q“
- **But how likely is this event?** „*What is the probability of the object traveling through Q?*“



Solutions

- Imprecise Observations
  - Model position of the object at each point of time either with a discrete or a probabilistic probability density function (pdf)
  - Positions at each point of time are independent from the positions at previous points of time
  - This yields wrong results according to PWS
  - If e.g. an object can only move upwards (e.g. since it can go back on a highway) then the yellow path is not possible.
  - Probability to intersect Q
    - Independence: 1 – (1-0.5)*(1-0.3) = 0.65
    - Dependent location: = 0.5



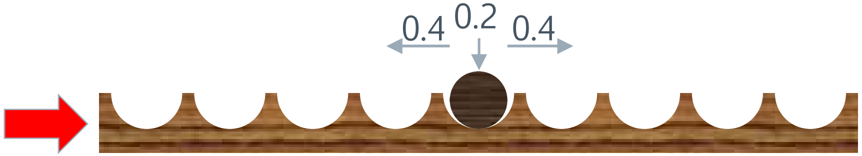
2.1. Modelling Uncertain Spatio-Temporal Data

Stochastic Processes for UST [QUeST11]

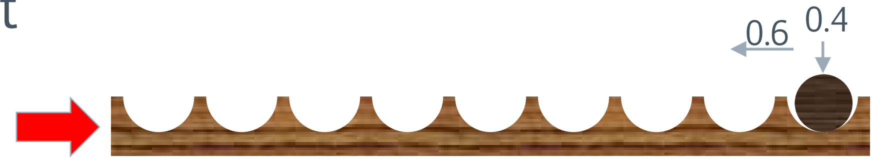
- Stochastic Processes are used to represent the evolution of some random value, or system, over time.
- A sound mathematical model which can be used to describe the uncertain location of an object over time.
- Many Stochastic Processes for different settings:
  - Markov Chain
  - Markov Process
  - Poisson Process
  - Wiener Process

A simple example

- Whenever the wooden board is hit, the ball stays or drops into one of the neighbour holes with certain probabilities.



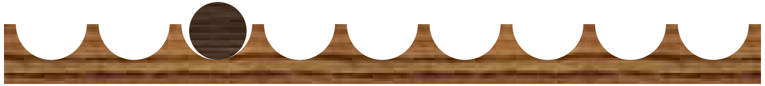
- At the border of the wood board these probabilities are different



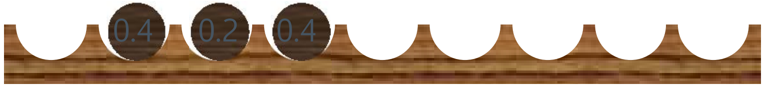
- This model is usually learned or given by experts

A simple example

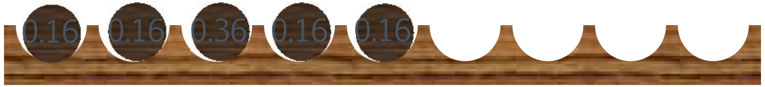
- Initial Position



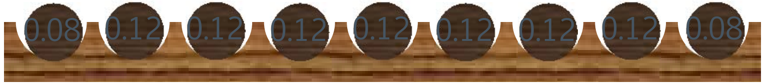
- After first hit



- After second hit



- After 40th hit



How can we model this?

› A Markov Chain is a “memoryless” Stochastic Process (the next state depends only on the current state)

› For our example we build the following transition Matrix M

$$\begin{array}{c|ccccccccc}
 & \multicolumn{9}{c}{\text{to bucket}} \\
\hline
\text{from bucket} & 0.4 & 0.6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 0.4 & 0.2 & 0.4 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 0 & 0.4 & 0.2 & 0.4 & 0 & 0 & 0 & 0 & 0 \\
 & 0 & 0 & 0.4 & 0.2 & 0.4 & 0 & 0 & 0 & 0 \\
 & 0 & 0 & 0 & 0.4 & 0.2 & 0.4 & 0 & 0 & 0 \\
 & 0 & 0 & 0 & 0 & 0.4 & 0.2 & 0.4 & 0 & 0 \\
 & 0 & 0 & 0 & 0 & 0 & 0.4 & 0.2 & 0.4 & 0 \\
 & 0 & 0 & 0 & 0 & 0 & 0 & 0.4 & 0.2 & 0.4 \\
 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.6 & 0.4
\end{array} = M$$

How can we model this?

- First hit

$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} * \begin{pmatrix} 0.4 & 0.6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.4 & 0.2 & 0.4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.4 & 0.2 & 0.4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.4 & 0.2 & 0.4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.4 & 0.2 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.4 & 0.2 & 0.4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.4 & 0.2 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.4 & 0.2 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.6 & 0.4 \end{pmatrix} = \begin{pmatrix} 0 & 0.2 & 0.4 & 0.2 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

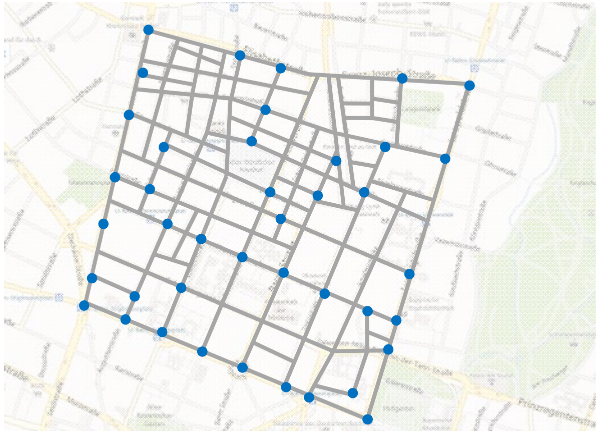
- Second hit

$$\begin{pmatrix} 0 & 0.4 & 0.2 & 0.4 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} * M = \begin{pmatrix} 0.16 & 0.16 & 0.36 & 0.16 & 0.16 & 0 & 0 & 0 & 0 \end{pmatrix}$$

- 40th hit

$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} * M^{40} = \begin{pmatrix} 0.8 & 0.12 & 0.12 & 0.12 & 0.12 & 0.12 & 0.12 & 0.12 & 0.08 \end{pmatrix}$$

Fusion of Model and Reality

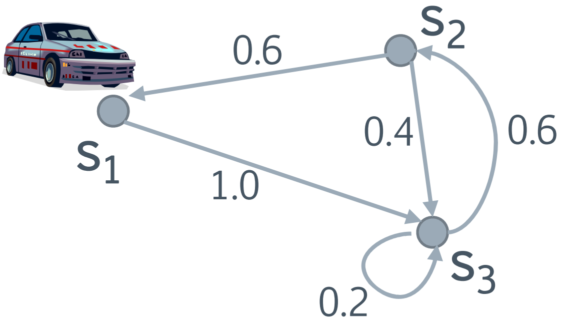


- Discretization of time and space
  - We usually treat intersections as states and add additional states on long streets
  - The time interval corresponding to a tick is 10 – 30 sec
- Estimation of model parameters
  - Transition probabilities from one state to another are learned from historical data (very sparse matrix!!)
  - Transition matrix can change over time and for different object groups

2.2. Querying Uncertain Spatio-Temporal

ST - Window Queries [ICDE12]

- Given the following state states and transition probabilities, what is the probability that the car is in s_1 or s_2 in the time interval T = [2,3]?

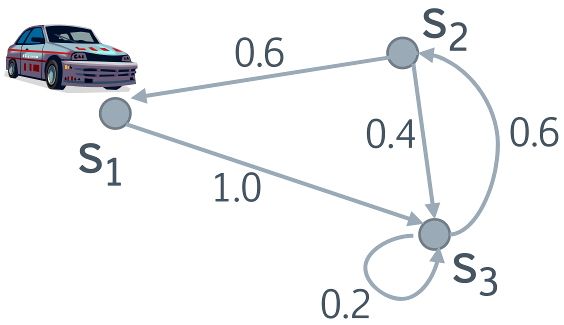


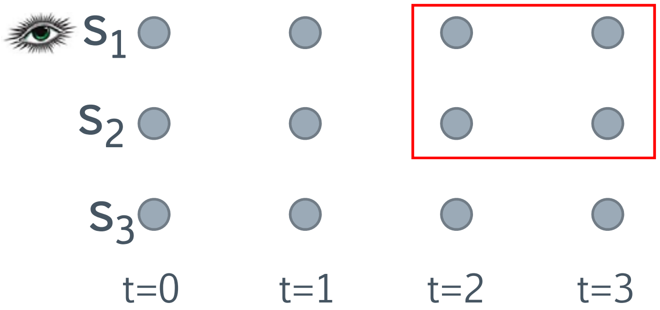
Note: Again we have an exponential number of possible paths the car might take!

$$M = \begin{pmatrix} 0 & 0 & 1 \\ 0.6 & 0 & 0.4 \\ 0 & 0.8 & 0.2 \end{pmatrix}$$

ST - Window Queries [ICDE12]

- Given the following state states and transition probabilities, what is the probability that the car is in s_1 or s_2 in the time interval T = [2,3]?



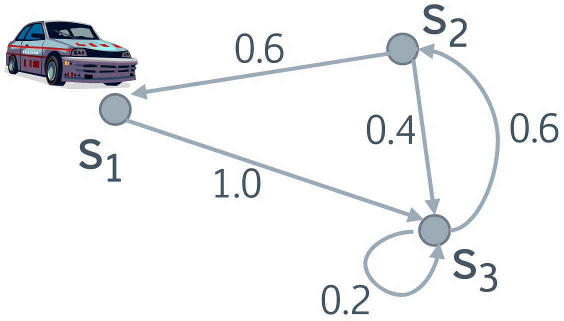


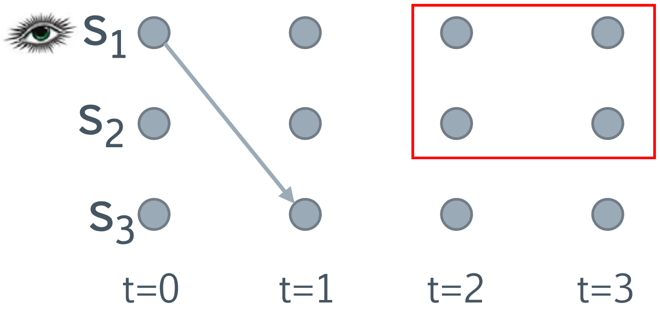
$$M = \begin{pmatrix} 0 & 0 & 1 \\ 0.6 & 0 & 0.4 \\ 0 & 0.8 & 0.2 \end{pmatrix}$$

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

ST - Window Queries [ICDE12]

› Given the following state states and transition probabilities, what is the probability that the car is in s_1 or s_2 in the time interval T = [2,3]?



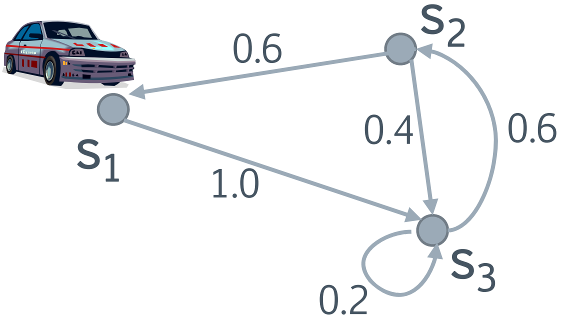


$$M = \begin{pmatrix} 0 & 0 & 1 \\ 0.6 & 0 & 0.4 \\ 0 & 0.8 & 0.2 \end{pmatrix}$$

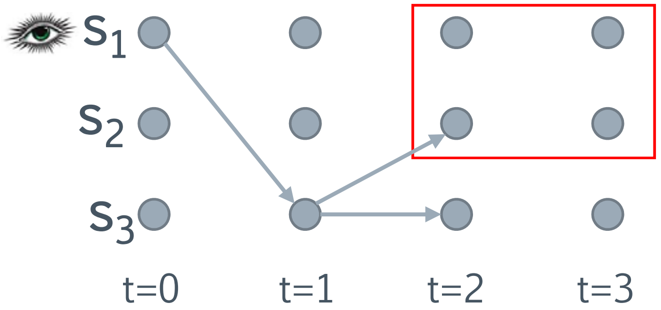
$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \xrightarrow{*M} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

ST - Window Queries [ICDE12]

- Given the following state states and transition probabilities, what is the probability that the car is in s_1 or s_2 in the time interval T = [2,3]?



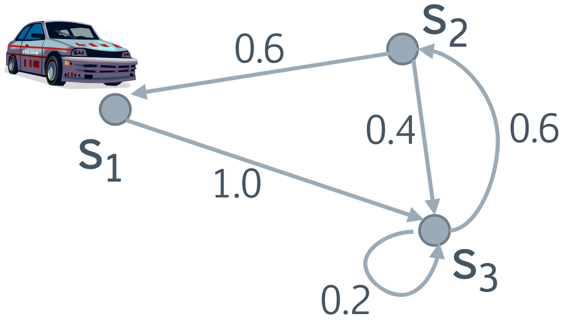
$$M = \begin{pmatrix} 0 & 0 & 1 \\ 0.6 & 0 & 0.4 \\ 0 & 0.8 & 0.2 \end{pmatrix}$$



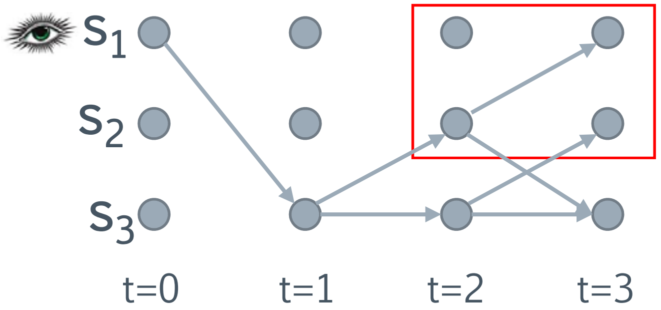
$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \xrightarrow{*M} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \xrightarrow{*M} \begin{pmatrix} 0 \\ 0.8 \\ 0.2 \end{pmatrix}$$

ST - Window Queries [ICDE12]

› Given the following state states and transition probabilities, what is the probability that the car is in s_1 or s_2 in the time interval T = [2,3]?



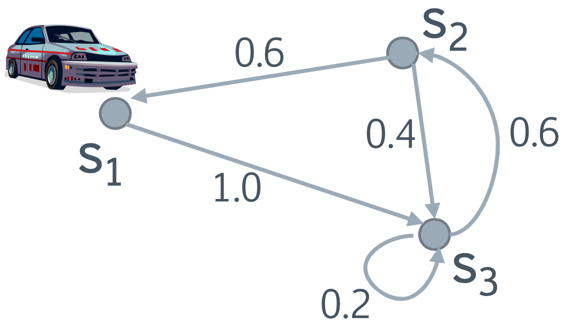
$$M = \begin{pmatrix} 0 & 0 & 1 \\ 0.6 & 0 & 0.4 \\ 0 & 0.8 & 0.2 \end{pmatrix}$$



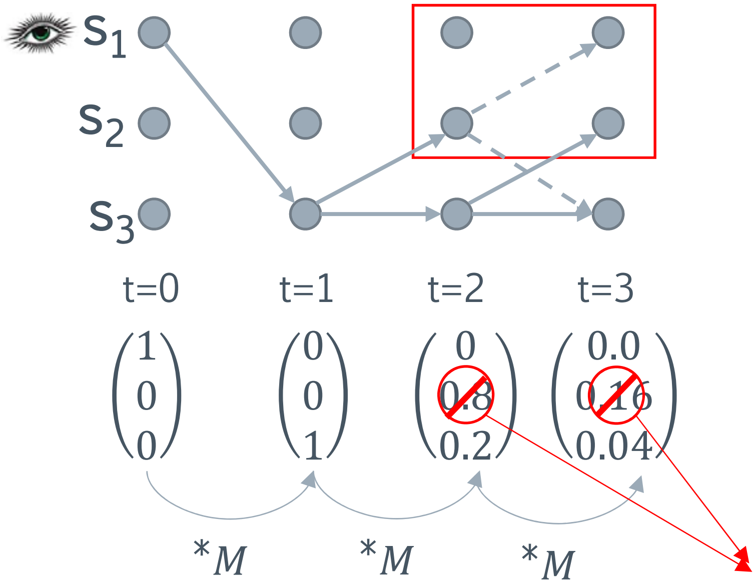
$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \xrightarrow{*M} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \xrightarrow{*M} \begin{pmatrix} 0 \\ 0.8 \\ 0.2 \end{pmatrix} \xrightarrow{*M} \begin{pmatrix} 0.48 \\ 0.16 \\ 0.36 \end{pmatrix}$$

ST - Window Queries [ICDE12]

› Given the following state states and transition probabilities, what is the probability that the car is in s_1 or s_2 in the time interval T = [2,3]?



$$M = \begin{pmatrix} 0 & 0 & 1 \\ 0.6 & 0 & 0.4 \\ 0 & 0.8 & 0.2 \end{pmatrix}$$



$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad \begin{pmatrix} 0 \\ \cancel{0.8} \\ 0.2 \end{pmatrix} \quad \begin{pmatrix} 0.0 \\ \cancel{0.16} \\ 0.04 \end{pmatrix}$$

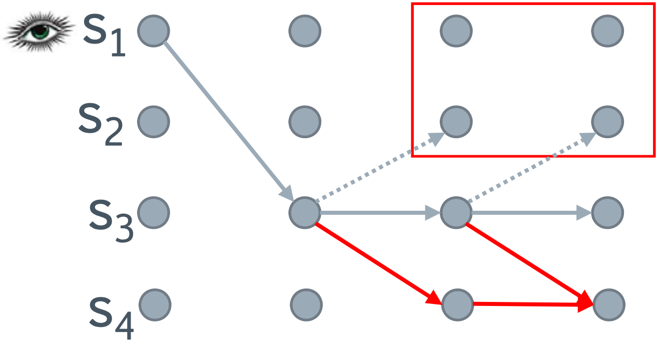
Result = 0.96

ST - Window Queries [ICDE12]

- Solution based on matrix multiplications introduces a new state for the winner trajectories and two matrices

$$M^- = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0.6 & 0 & 0.4 & 0 \\ 0 & 0.8 & 0.2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

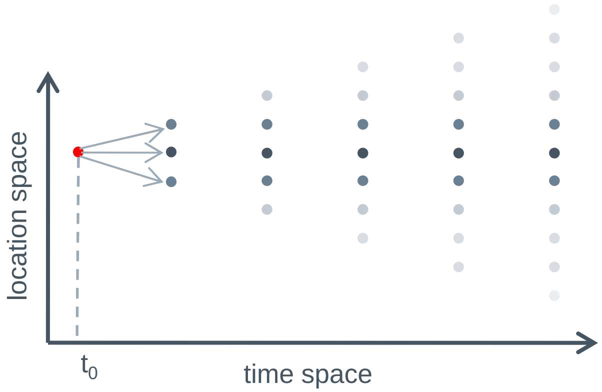
$$M^+ = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0.4 & 0.6 \\ 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$



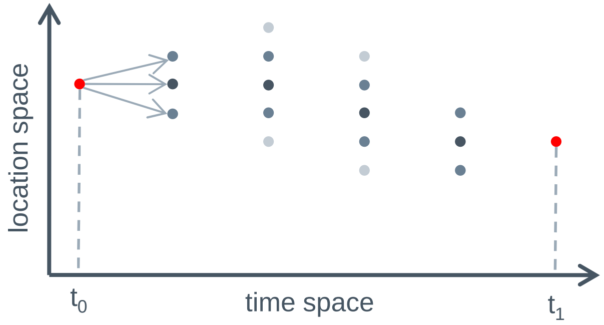
$$\begin{tikzcd} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \arrow[r, bend right, "*M^-"'] & \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \arrow[r, bend right, "*M^+"'] & \begin{pmatrix} 0 \\ 0 \\ 0.2 \\ 0.8 \end{pmatrix} \arrow[r, bend right, "*M^+"'] & \begin{pmatrix} 0 \\ 0 \\ 0.04 \\ \boxed{0.96} \end{pmatrix} \end{tikzcd}$$

Multiple Observations

- So far we had only one observation from which we could extrapolate

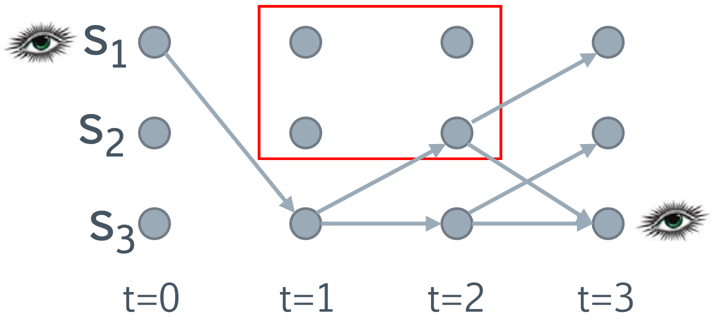


- This is not really of interest since cars do not move randomly



- With two observations we have to introduce more artificial states and adapt the techniques

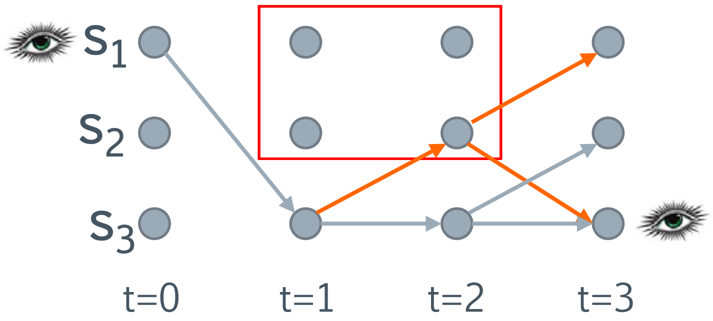
Multiple Observations



$$M = \begin{pmatrix} 0 & 0 & 1 \\ 0.6 & 0 & 0.4 \\ 0 & 0.8 & 0.2 \end{pmatrix}$$

- We need to track where true hit worlds are located
  - 2*|S| classes of equivalent worlds
  - One class S_i^- corresponding to worlds where o is located in state s_i, and o has not intersected the window
  - One class S_i^+ corresponding to worlds where o is located in state s_i, and o has not intersected the window

Multiple Observations



$$\begin{tikzcd}
\begin{matrix} \neg\blacksquare \left\{ \begin{matrix} S_1^- \\ S_2^- \\ S_3^- \end{matrix} \right. \\ \blacksquare \left\{ \begin{matrix} S_1^+ \\ S_2^+ \\ S_3^+ \end{matrix} \right. \end{matrix}
& \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \arrow[r, bend right, "{*M^+}"']
& \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \arrow[r, bend right, "{*M^-}"']
& \begin{pmatrix} 0 \\ 0 \\ 0.2 \\ 0 \\ 0.8 \\ 0 \end{pmatrix} \arrow[r, bend right, "{*M^-}"']
& \begin{pmatrix} 0 \\ 0.16 \\ 0.04 \\ 0.48 \\ 0 \\ 0.32 \end{pmatrix}
\end{tikzcd}$$

$$M = \begin{pmatrix} 0 & 0 & 1 \\ 0.6 & 0 & 0.4 \\ 0 & 0.8 & 0.2 \end{pmatrix}$$

$$M^+ = \begin{pmatrix} M & 0 \\ 0 & M \end{pmatrix}$$

$$M^- = \begin{pmatrix} 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.4 & 0.6 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.2 & 0.0 & 0.8 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 0.0 & 0.6 & 0.0 & 0.4 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.8 & 0.2 \end{pmatrix}$$

Bayes' Theorem

- Now what is the probability that the trajectory passes the query window given the fact that the object was seen in s_3?

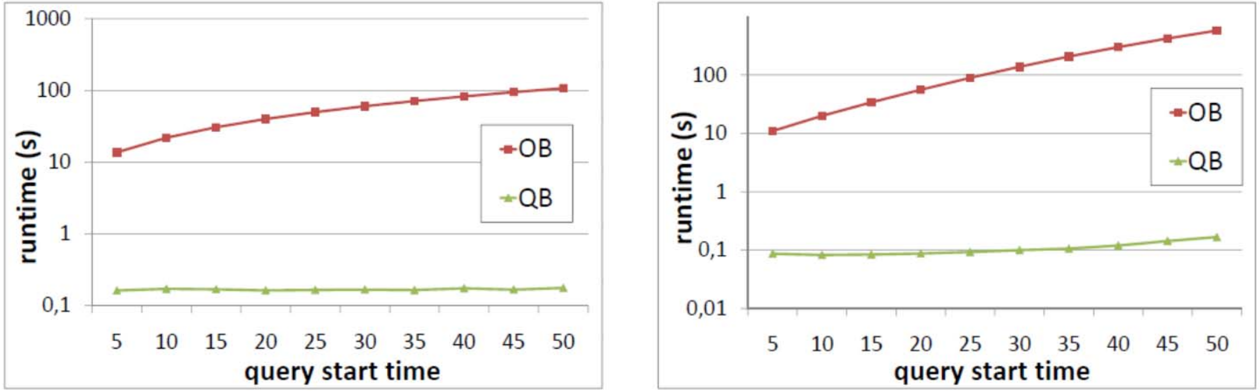
$$\begin{array}{l} S_1^- \\ S_2^- \\ S_3^- \\ S_1^+ \\ S_2^+ \\ S_3^+ \end{array} \begin{pmatrix} 0 \\ 0.16 \\ 0.04 \\ 0.48 \\ 0 \\ 0.32 \end{pmatrix} \begin{array}{l} \\ \\ \text{👁} \\ \\ \\ \text{👁} \end{array}$$

$$P(\blacksquare|\text{👁}) = \frac{P(\text{👁}|\blacksquare) * P(\blacksquare)}{P(\text{👁})} = \frac{P(\blacksquare \wedge \text{👁})}{P(\text{👁})}$$

$$= \frac{P(\blacksquare \wedge \text{👁})}{P(\text{👁} \wedge \blacksquare) + P(\text{👁} \wedge \neg\blacksquare)} = \frac{0.32}{0.32 + 0.04} = 0.89$$

Experimental Results

- For 10,000 objects and 100,000 states on a single machine



(a) Synthetic data (b) Munich dataset

- Can be distributed and parallelized!

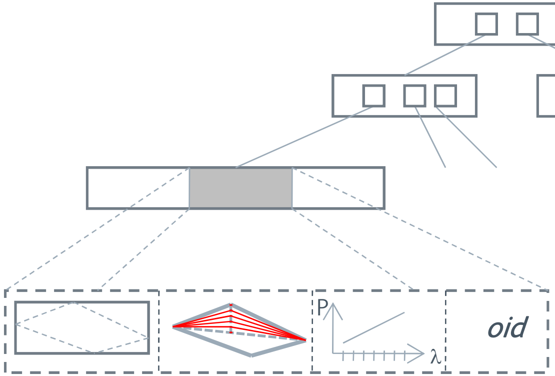
Summary

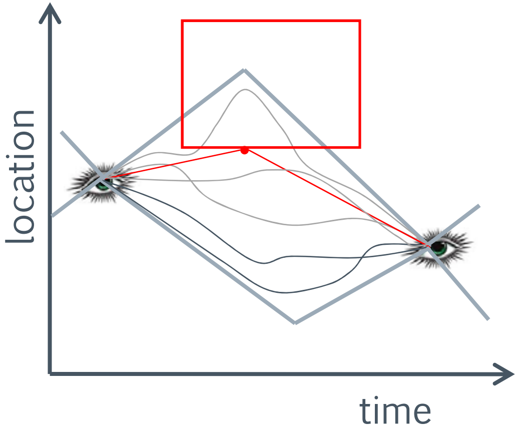
- Pros
  - Allows to answer queries according to possible worlds semantics
  - Considers location dependencies over time
  - Scales up very well since it is purely based on sparse matrix multiplications
  - Natively extendable for uncertain observations
  - Seems to work adequately on real-world data (more validation needed)
- Cons
  - Discrete time and space
  - Matching from time to tics might not be the perfect modelling

2.3. Follow-Up Works

Indexing UST Data [CIKM12]

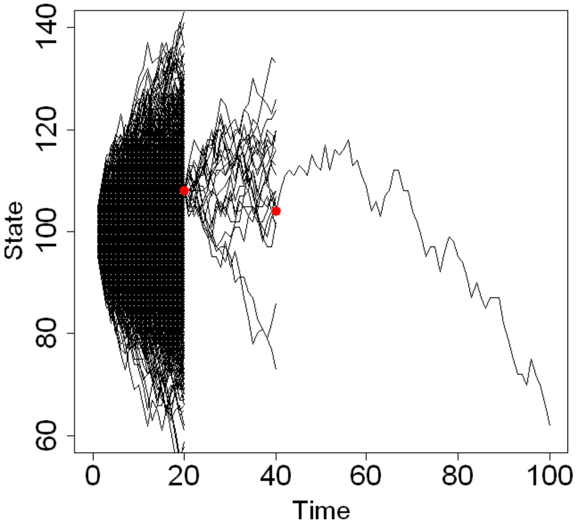
- With the current techniques we have to process each object in the database
- Index Structure based on R-Tree indexing the ST-Space
- The leafs contain the “intelligence” and enable probabilistic pruning (at max x% of the possible trajectories of o may intersect Q)





KNN queries + Sampling on UST Data [PVLDB13]

- Not all queries can be solved as elegant as window queries
- Popular in uncertain databases: Monte-Carlo-Sampling
  - Draw a sufficiently high number of samples
  - Approximate result probability = ratio of samples that satisfy the query and total number of drawn samples
- But how to draw samples efficiently such that they are conform with the observations?
- Solution: Adaption of transition matrices



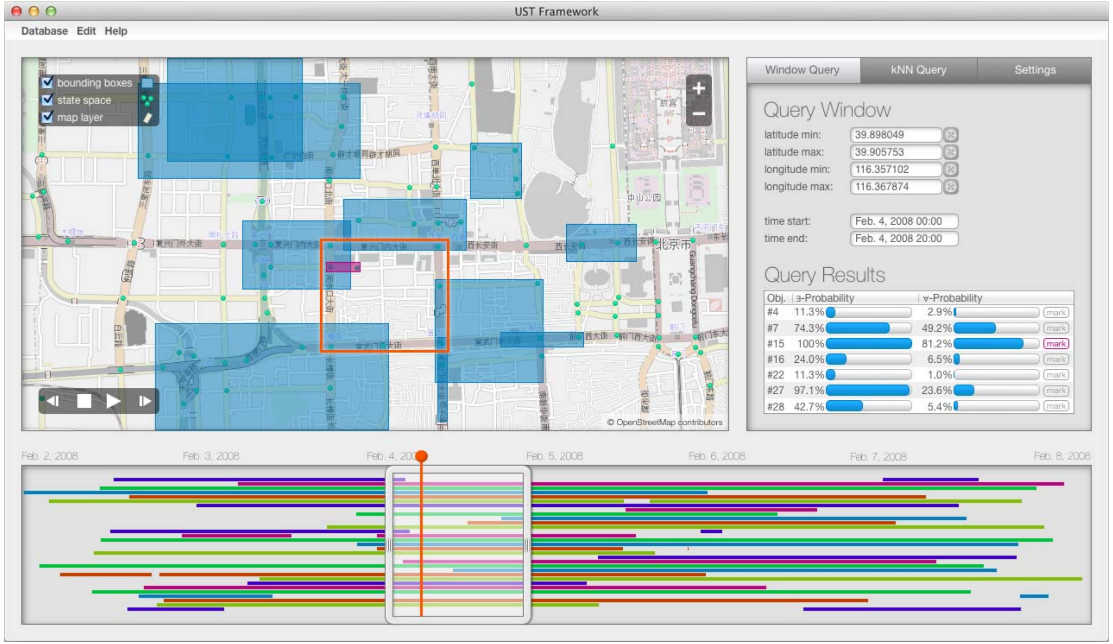
Other query predicates

- Similarity search on UST data [SISAP13]
  - How similar are two uncertain trajectories?
  - A probabilistic measure based on Longest Common Subsequence

- Reverse nearest neighbor queries [DASFAA14]
  - Not really intended, but to clarify an ICDE '13 paper that picked um the model
  - ...and Bali is nice ;-)

Demo [submitted to SIGMOD14]

- All code is available as C++ Code
- Together with a graphical user interface



Query Window

latitude min: 39.898049
latitude max: 39.905753
longitude min: 116.357102
longitude max: 116.367874

time start: Feb. 4, 2008 00:00
time end: Feb. 4, 2008 20:00

Query Results

Obj.	∃-Probability	∀-Probability	
#4	11.3%	2.9%	mark
#7	74.3%	49.2%	mark
#15	100%	81.2%	mark
#16	24.0%	6.5%	mark
#22	11.3%	1.0%	mark
#27	97.1%	23.6%	mark
#28	42.7%	5.4%	mark

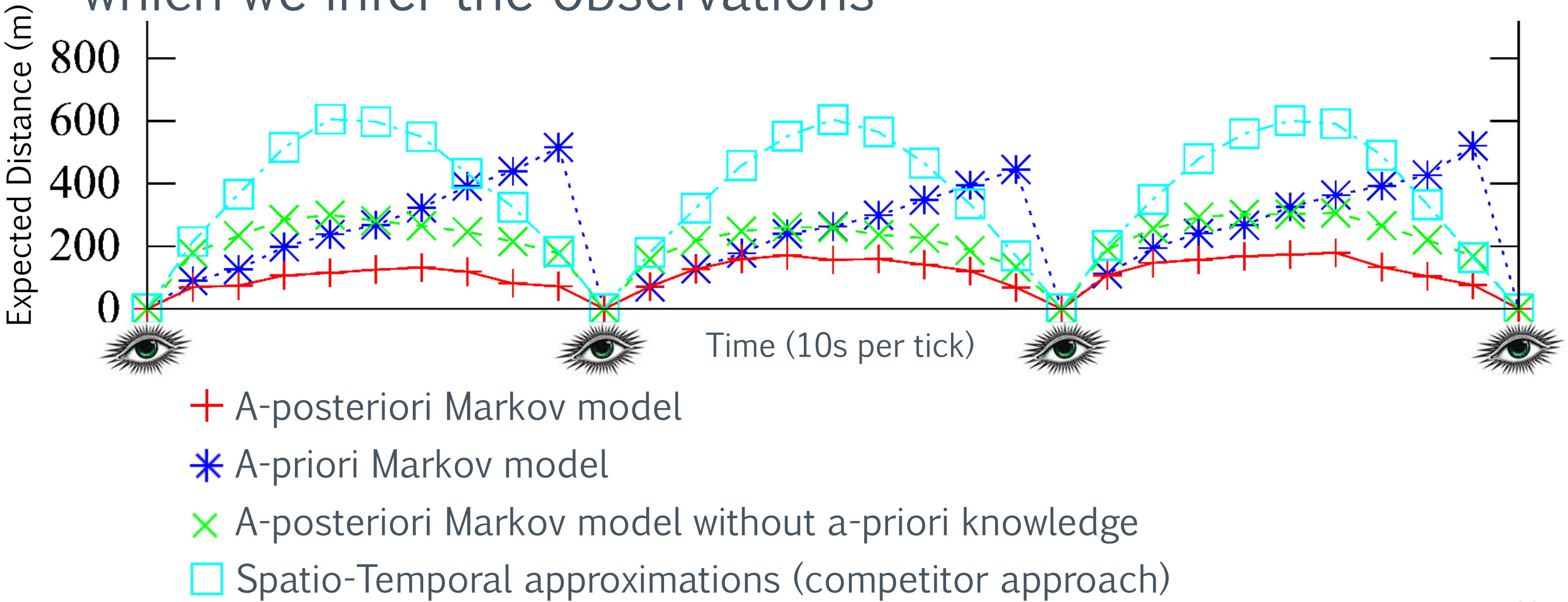
Future Directions for the UST Project

- Other probabilistic spatio-temporal queries
- Integration of other kinds of observations
- Analysis of other stochastic processes
  - Continuous space
  - Continuous time
  - Object dependence
- Learning of the parameters of stochastic processes
- Probabilistic Datamining on UST Data

Thanks for listening!

Does the Markov assumption hold in reality ?

- Of course single cars do not follow the Markov Chain (random walk)
- However the Markov Model is just the apriori Model in which we infer the observations



Related Work

- [QUeST11] T. Bernecker, L. Chen, T. Emrich, H.-P. Kriegel, N. Mamoulis, and A. Züfle. *Managing Uncertain Spatio-Temporal Data.* In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Querying and Mining Uncertain Spatio-Temporal Data (QUeST), Chicago, Illinois, 2011.
- [ICDE12] T. Emrich, H.-P. Kriegel, N. Mamoulis, M. Renz, and A. Züfle. *Querying uncertain spatio-temporal data.* In Proceedings of the 28th International Conference on Data Engineering (ICDE), Washington, DC, 2012.
- [CIKM12] T. Emrich, H.-P. Kriegel, N. Mamoulis, M. Renz, and A. Züfle. *Indexing uncertain spatio-temporal data.* In Proceedings of the 21th ACM International Conference on Information and Knowledge Management (CIKM), Maui, Hawaii, USA, 2012.
- [PVLDB13] Johannes Niedermayer, Andreas Züfle, Tobias Emrich, Matthias Renz, Nikos Mamoulis, Lei Chen, Hans-Peter Kriegel: *Probabilistic Nearest Neighbor Queries on Uncertain Moving Object Trajectories.* PVLDB 7(3): 205-216 (2013)

- Project Page: http://www.dbs.ifi.lmu.de/cms/Publications/UncertainSpatioTemporal

Related Work

- [1] http://infoblog.stanford.edu/2008/07/why-uncertainty-in-data-is-great-posted.html
- [2] Jian Li, Barna Saha, Amol Deshpande: A unified approach to ranking in probabilistic databases. VLDB J. 20(2): 249-275 (2011)