Managing Uncertainty in Data Streams

Aleka Seliniotaki

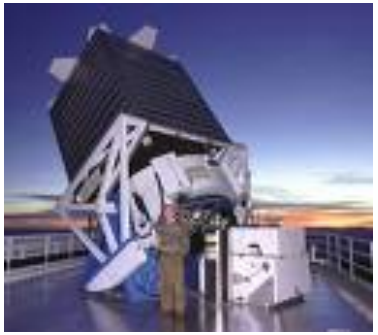
Project Presentation HY561

Heraklion, 22/05/2013

Introduction
Uncertain Data Streams

Data: *incomplete, imprecise, misleading*

Results: *unknown quality*



Introduction

- Uncertainty is the component of a reported value that characterizes the range of values where the true value should lie (e.g. the height of a person)
- Raw data streams don't give us the opportunity to process the data and correct the errors in real time (the data flows continuously)
- Capturing uncertainty from input data to query output becomes a key component of scientific data management systems.

Introduction
Biosensors

- Biosensors are analytical tools for monitoring programs and characterizing complex biological systems enabling manipulation of those systems
  - Biosensors for environmental applications:
    - Water treatment and desalination plants
    - Environmental pollution monitoring
    - Food Safety
  - Biosensors for medicine applications:
    - support continuous monitoring of patient conditions, providing a degree of self-diagnosis and enabling effective real-time decision making to reduce fatalities
- Query detecting increase in the fouling mass concentration

```
Select    sensorID,avg(S.Biofouling)
From      Biosensor S
Where     avg(S.Biofouling)>200
```

Quality of the returned alert?

1. Incomplete definition of observed quantities
2. Sampling effects, interferences
3. Approximations of the measurement process

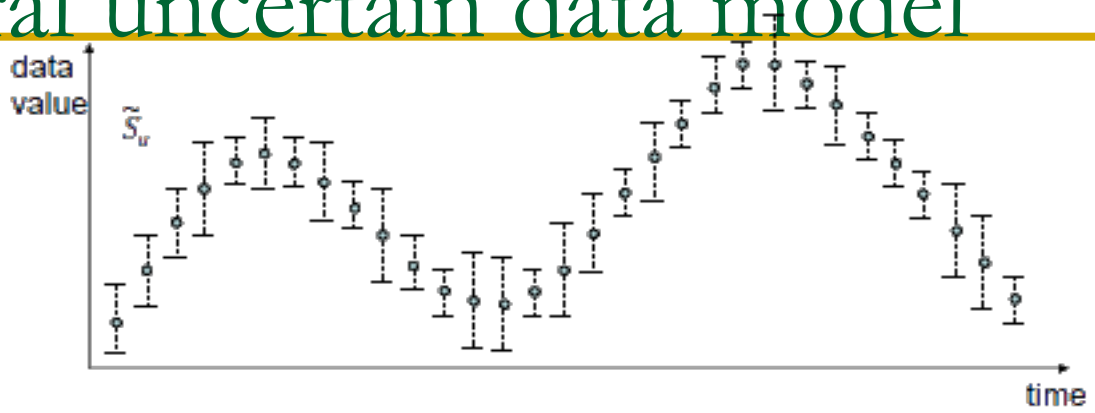
Introduction
RFID Tracking and Monitoring

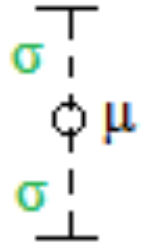
- RFID technology used for object tracking and monitoring
  - E.g. a warehouse, a retail store or a library
- Raw RFID readings are noisy and incomplete
- Inference yields stream with object locations

(time, tag_id, weight, (x,y), size)

1.Varying environmental conditions
2.Inherent inaccuracies of the equipment
3.Variations in repeated observations of the measurand under the same conditions

Introduction
General uncertain data model





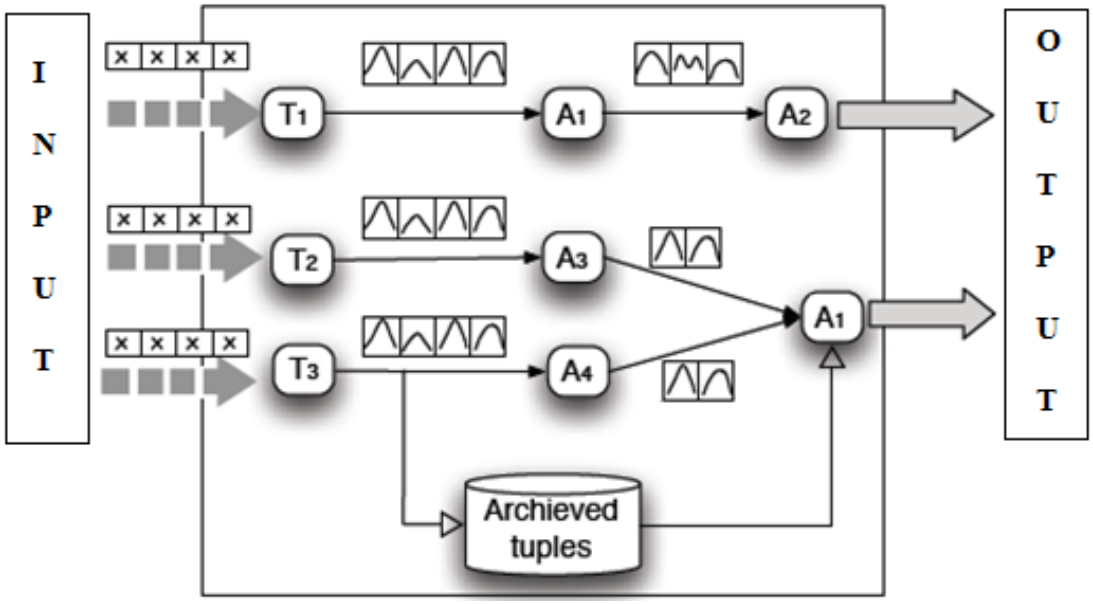
- The data value at each timestamp can be related with uncertainty
- The uncertainty at each time point can be modeled as a **continuous random variable** (the data can take infinitely values in the possible range of the random variable).
- Only the *mean* (μ) and the *deviation* (σ) of the random variable at each timestamp are available
- Each stream of data is considered as a sequence of random variables

Introduction
Main challenge

1.Characterizing the uncertainty in raw data before the insertion to the operators and in the query results

2. In the output, we should not have worse results regarding the uncertainty in measurements

Quantification of uncertainty in raw data streams



Quantification of uncertainty in the query results

PODS and CLARO systems:
support stream processing for uncertain data

Contents

- PODS system
  - Scope and contribution
  - Data Model
  - Aggregation of Uncertain tuples
  - Joins of Uncertain tuples
  - Conclusions
- CLARO system
  - Scope and contribution
  - Data Model
  - Selection
  - Aggregation with TEP≤1
  - Join
  - Query planning
  - Conclusions
- Conclusions

PODS system
Scope and contribution

- PODS system supports stream processing for uncertain data, using *continuous random variables*
- Characterizes the distribution of each tuple produced from uncertain data (before and after the operator) for the relational operators *aggregation* and *join*
- The architectural design of PODS is to extend the box-arrow example for stream processing such as:
  - The tuples carry distributions to describe uncertainty
  - Relational operators transform these distributions while processing tuples.

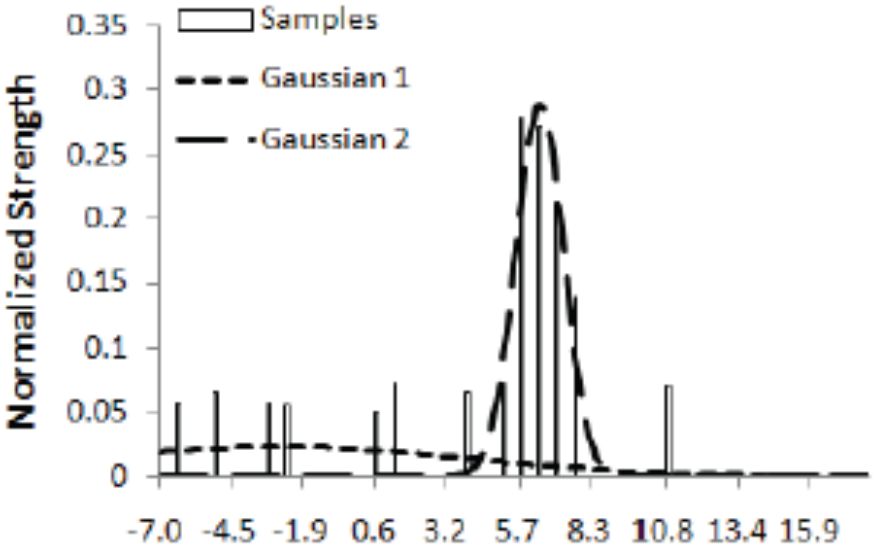
PODS system
Data Model

- The foundation of the PODS system is a data model based on Gaussian Mixture distributions that:
  - Captures a variety of uncertainties for continuous valued attributes
  - Gives us the opportunity to represent the unknown measurement values by making assumptions on the variance
- A Gaussian Mixture Model for a continuous random variable X is a mixture of m Gaussian variables $X_1, X_2, \ldots, X_m$. The probability density function (pdf) of X is

$$f_X(x) = \sum_{i=1}^{m} p_i f_{X_i}(x), \qquad f_{X_i}(x) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$

PODS system
Data Model-Why Gaussian distribution???

- This distribution is widely used in scientific applications, because
  - Can be completely specified by two parameters
    - Mean : defines the location of the random variable in space
    - Standard deviation: defines the scale of the random variable
  - Can be applied to situations in which the data is distributed very differently (central limit of the population the distribut approaches a normal distribution
- Gaussian properties and give flexibility in complex
- It is suitable for mod distributions



PODS system
Data Model

- Why not discrete random variables?

Tuple	Velocity	Prob
t1	10	0.7
	12	0.3
t2	7	0.4
	18	0.6
t3	16	0.2
	18	0.8
t4	22	0.5
	28	0.5

PWS

PW	Avg(Velocity)	Prob
1	13.75	0.028
2	14.75	0.012
...	...	...
16	19.00	0.072

(a) Discrete

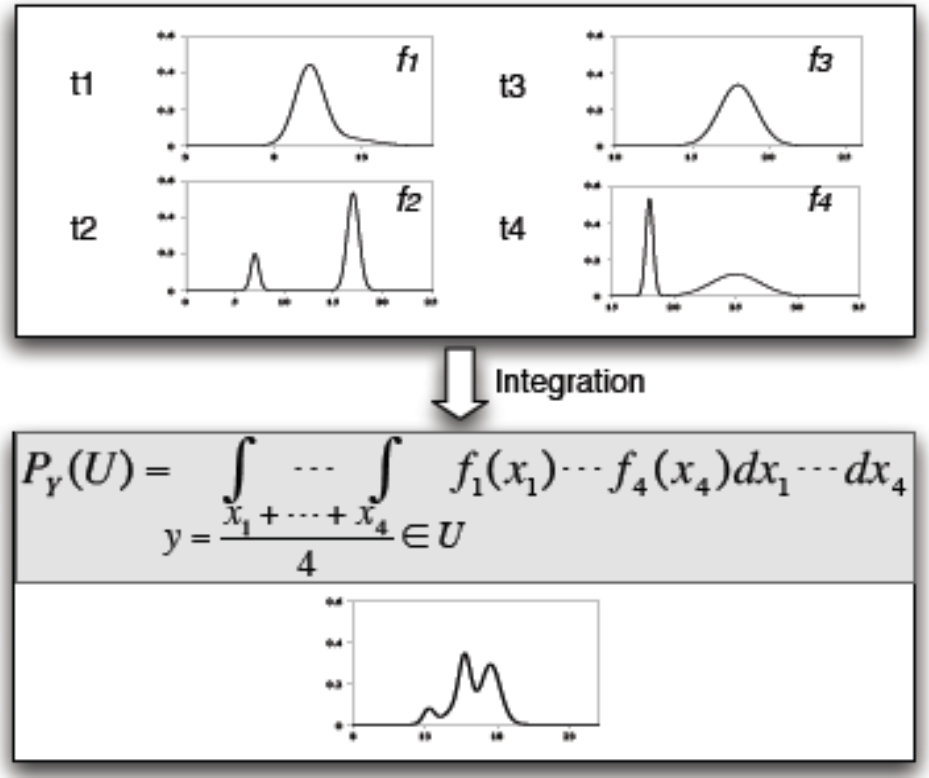
Discrete Random Variables model data uncertainty and the possible world semantics (PWS) model query processing.

Computing the distribution of the average of n discrete random variables may require the enumeration of an exponential number of possible worlds

PODS system

Aggregation –Existing approaches

- *Integral-based*: integrates two continuous random variables at a time, resulting in the use of *n* integrals to aggregate *n* variables.
  - not suitable for stream processing



Example: Illustration of an average of four continuous random variables , $Y=1/4(X_1+\ldots+X_4)$. The probability that Y is in the range of U is defined by the multivariate integral in the figure

PODS system

Aggregation –Existing approaches

- *Sampling-based*: generates a fixed number of samples from the distribution of each input tuple, computes aggregate values from these samples and constructs the output distribution using the aggregate values.
  - It is unknown how many samples are needed a priori
  - It does not provide knowledge of the true result distribution

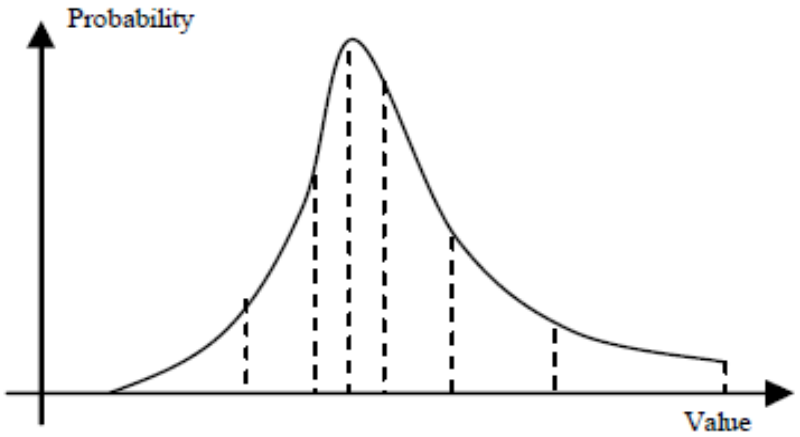


Figure: Illustrating discretization of a Gaussian distribution into intervals

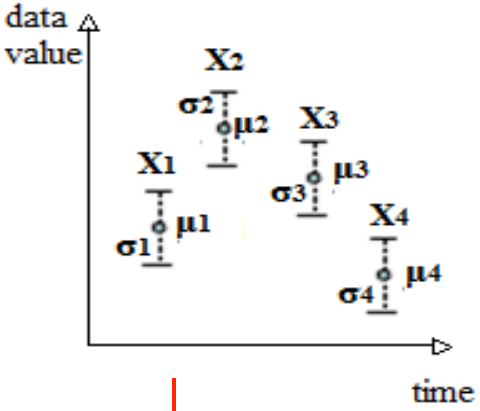
PODS system
Aggregation of Uncertain tuples

- *1st solution: Characteristic functions*
  - CFs are used to "characterize" distributions. The CF of a random variable U is defined as $\Phi_U(t) = E[e^{iUt}]$ (E denotes the expected value and i is the complex number sqrt(-1))and the pdf of U is

$$f_U(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-itx}\Phi_U(t)dt$$

  - Two steps to compute the result distribution:
    - Get the CF of each input tuple and take the product of these functions
    - For a given point, apply the inverse transformation of the CF

PODS system
Aggregation of Uncertain tuples



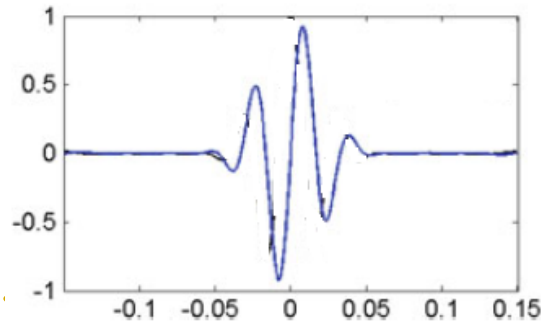
$sum = X_1 + X_2 + X_3 + X_4$

$$E[e^{iXjt}] = E[e^{i\,\mu\,jt-0.5\,\sigma\,j\hat{}2t\hat{}2}]$$

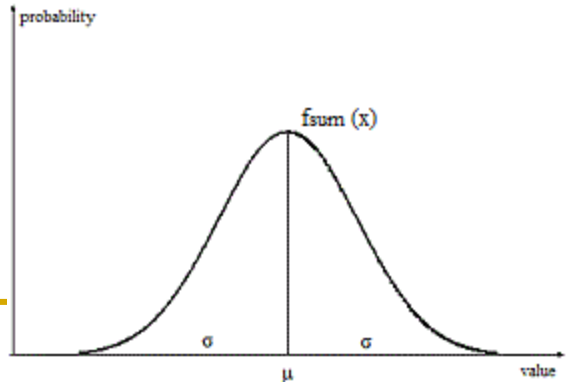
1. Produce characteristic function

$$\Phi_{sum}(t) = E[e^{i\,sum\,t}] = E[e^{i(X_1+X_2+X_3+X_4)t}] = E[e^{iX_1t}.e^{iX_2t}.e^{iX_3t}.\,e^{iX_4t}] =$$

$$E[e^{iX_1t}] \cdot E[e^{iX_2t}] \cdot E[e^{iX_3t}] \cdot E[e^{iX_4t}] = \Phi_{X1(t)}\,\Phi_{X2(t)}\,\Phi_{X3(t)}\,\Phi_{X4(t)}$$



2. For a given point apply the inverse transformation to yield $f_{sum}(x)$



PODS system
Aggregation of Uncertain tuples

- *Advantage*: gives a boost in performance compared to integral-based method which requires *n* parameterized integrals

- *Drawback*: The result distribution involves an unresolved parameterized integral. To get sufficient knowledge of the result distribution (e.g., calculating its mean and variance), one needs to repeat the inverse transformation for a large number of points

PODS system
Aggregation of Uncertain tuples

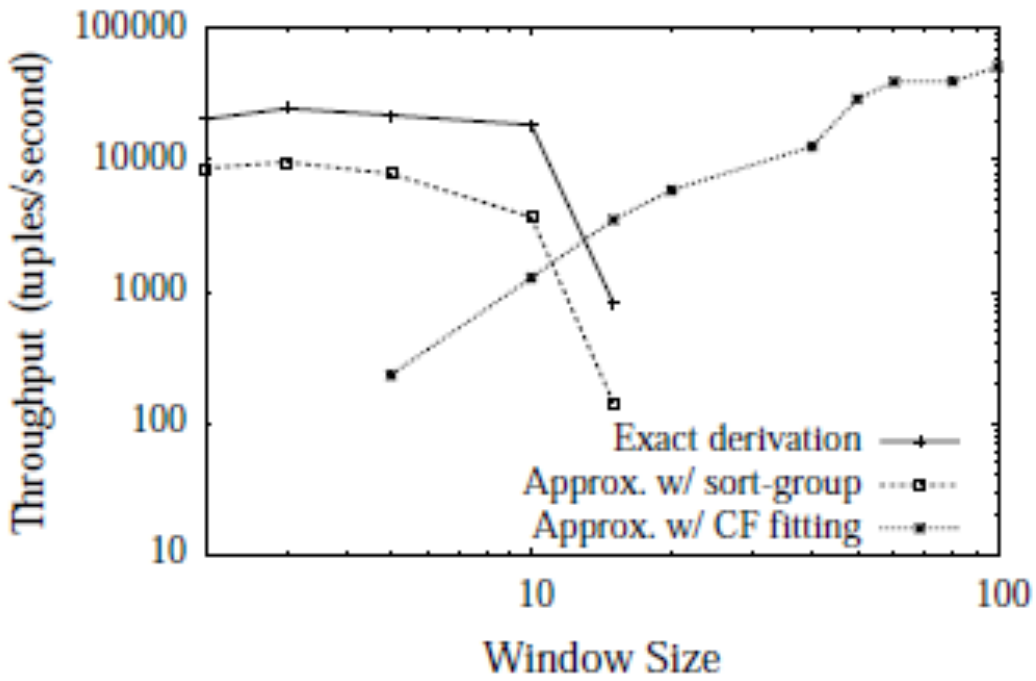
- *2nd solution: Exact derivation of result distributions.*
  - Let each X_i (i = 1..n) : mixture of m_i components identified by the parameters (μ_{ij} , σ_{ij}), ($j = 1..m_i$).
  - Result distribution of $U = \sum_{i=1}^{n} X_i$ Gaussian mixture of m_i components (corresponds to a unique combination that takes one component from each input Gaussian mixture),
  - Each result component is identified by (μ_k, σ_k):

$$\mu_k = \sum_{i=1}^{n} \mu_{i_j};\ \sigma_k = \sqrt{\sum_{i=1}^{n} \sigma_{i_j}^2}$$

  - *Advantage*: exact solution, so the accuracy is guaranteed
  - *Drawback* : Enumerate and compute all components of the result Gaussian mixture → exponentially growth in the number of tuples.

PODS system
Aggregation of Uncertain tuples

- *Sort-Gr* adjacent Gaussia roximates each gr uitable mean and var



- *CF fittin* nates the exact re ction fitting in the C Gaussian mixture distribution whose CF best fits this product function.

PODS system
Joins of Uncertain tuples

- Two types of joins of continuous random attributes:
  - The system employs *probabilistic views* to facilitate equi-joins and offers a closed-form solution in GMMs to represent join result distributions
  - The system combines tuples from two inputs for inequality and is modeled by a cross-product followed by a selection.

PODS system
Joins of Uncertain tuples

- What is a probabilistic view?
  - Let a stream S(**A,B,S'**), where:
    - **A** is a vector of attributes that can be deterministic/ probabilistic (e.g. a location stream consists of (X,Y) attributes)
    - **B** is another vector of deterministic/probabilistic attributes that are related to attributes in A (e.g a temperature stream consists of (X, Y, Temp) attributes)
    - **S'** is a vector for the rest of the attributes
- The *probabilistic view* of **B** as a function of **A** (denoted by $V_{B|A}(S)$), is a distribution of **B** for a given value of **A**, characterized by $p_B(b|A=a)$

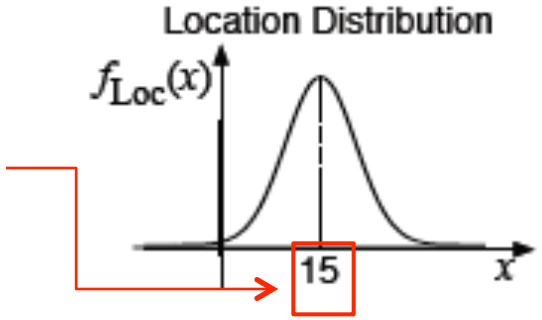
PODS system
Joins of Uncertain tuples

- Example of probabilistic view:

T1 Object Location

Tag id	Loc	Prob
0x333	10	0.5
	20	0.5

0.5*10 + 0.5*20 = 5+10= 15



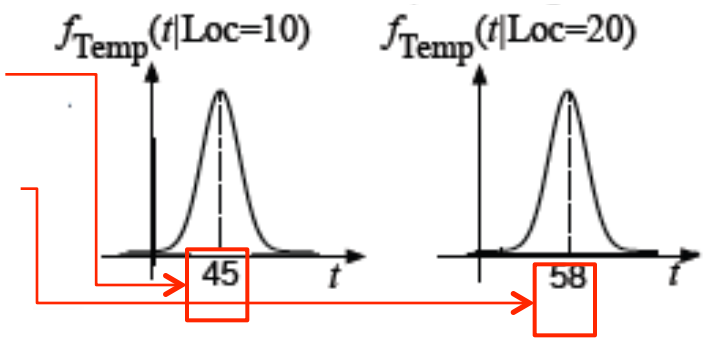
Probabilistic View of Temperature given Location:

T2 Temperature

Loc	Temp	Prob
10	30	0.2
	50	0.8
20	50	0.6
	70	0.4

0.2* 30 + 0.8*50 =6+40 ≈45

0.6* 50+ 0.4*70= 30+28=58



PODS system
Joins of Uncertain tuples

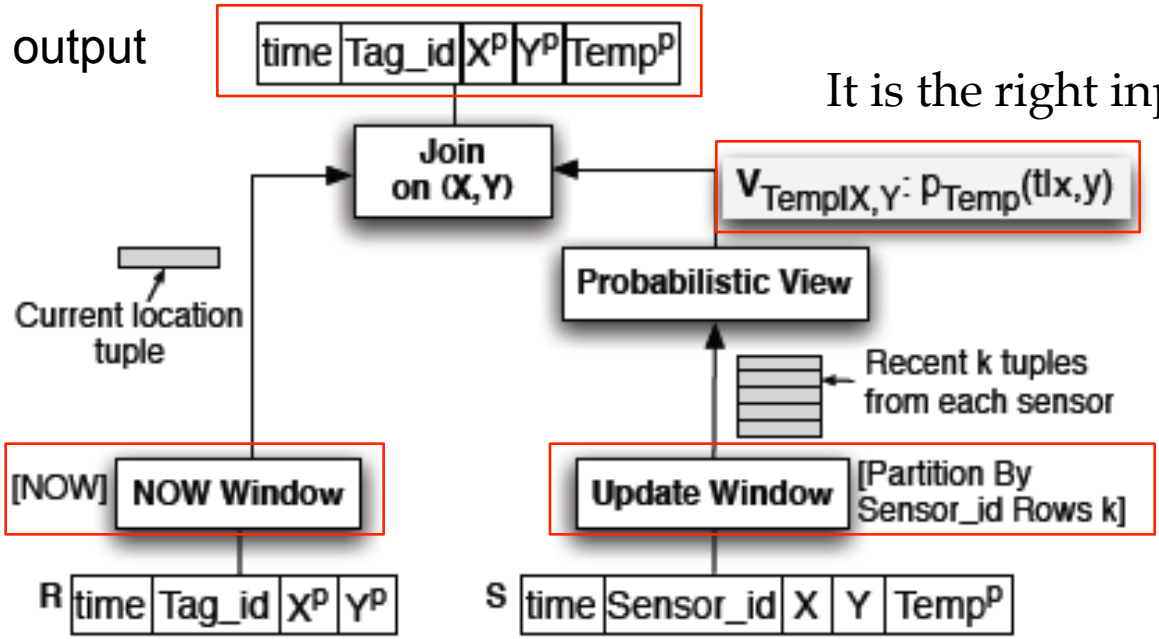
- Join using probabilistic views in GMMs:
  - Given two independent streams R(**A,R′**) , S(**A,B,S′**) an equi-join of R and S is a join of R and $V_{B|A}(S)$:
    - For any tuple *i* in *R*, denoted by ($\mathbf{A_i}$, $\mathbf{R'_i}$), the join combines the tuple with the view $V_{B|A}(S)$ and outputs a tuple ($\mathbf{A_i}$,$\mathbf{R'_i}$,**B**), with the joint distribution defined as:

$$f_{\mathbf{A_i},\mathbf{R_i},\mathbf{B}}(\mathbf{a},\bar{\mathbf{r}},\mathbf{b}) \equiv f_{\mathbf{A_i},\mathbf{R_i}}(\mathbf{a},\bar{\mathbf{r}}) \cdot p_{\mathbf{B}}(\mathbf{b}|\mathbf{S.A}=\mathbf{a})$$

  - If ($\mathbf{A_i}$, $\mathbf{R'_i}$) follows GMMs the output result ($\mathbf{A_i}$,$\mathbf{R'_i}$,**B**) also follows a GMM

PODS system
Joins of Uncertain tuples

- **Select** Rstream (R.Tag_id,R.x,R.y, T.Temp)
  from Object as R, Temperature as T
  where T.Temp>60 °C and R.x=T.x and R.y=T.y



PODS system
Conclusions

- The PODS system based on *GMMs* and *techniques for aggregates and joins* process the *uncertainty* in data streams
- The techniques further produce output distributions
- Extension of this work to:
  - Support a broader set of relational operators
  - Query optimization
  - Exploit inter-tuple correlations
  - Explore the combination of both discrete and continuous random variables

CLARO system

- Scope and contribution
- Data Model
- Selection
- Aggregation with TEP≤1
- Join
- Query planning
- Conclusions

CLARO system
Scope and contribution

- CLARO provides a system framework that supports stream processing for continuous uncertain data.
- Two types of uncertainty in data streams :
  - Attributes values (PODS system)
  - Tuple existence, indicating whether a tuple is present in a relation, is modeled by a discrete random variable (introduced from conditioning operations such as selections or group-bys)
- Support of a broader set of relational operators (selection, projection,aggreagation (min,max,count,sum, avg), group-by aggregation,join)

CLARO system
Tuple Existance Probability

- Figure shows an example of a database, Dp, with relations Sp (containing tuples s1 and s2 with probabilities 0.6 and 0.5 respectively) and Tp (containing tuple t1 with probability 0.4).

D^p

S^p

	A	B	*prob*
s_1	m	1	0.6
s_2	n	1	0.5

T^p

	C	D	*prob*
t_1	1	p	0.4

$pwd(D^p)$

instance	probability
$d_1 = \{s_1, s_2, t_1\}$	0.12
$d_2 = \{s_1, s_2\}$	0.18
$d_3 = \{s_1, t_1\}$	0.12
$d_4 = \{s_1\}$	0.18
$d_5 = \{s_2, t_1\}$	0.08
$d_6 = \{s_2\}$	0.12
$d_7 = \{t_1\}$	0.08
$d_8 = \emptyset$	0.12

CLARO system
Data Model

- An uncertain input stream is an infinite sequence of tuples that conform to the schema $\mathbf{A}^d \cup \mathbf{A}^p$
  - $\mathbf{A}^d$ are deterministic attributes (like those in traditional databases)
  - $\mathbf{A}^p$ are continuous valued uncertain attributes (such as location of an object). They are modeled by a vector of continuous random variables **X,** characterized by GMMs.
- The mixed-type distribution *g* of a tuple(with *m* continuous and *n* discrete uncertain attributes) is a pair *(p,f)* :
  - $p \in [0,1]$ is the tuple existence probability (TEP)
  - *f* is the joint density function for all uncertain attributes

$$f(A^x = x, A^y = y) = f_{A^x|A^y}(x \mid y) * P(A^y = y)$$

Claro System
Data Model Example

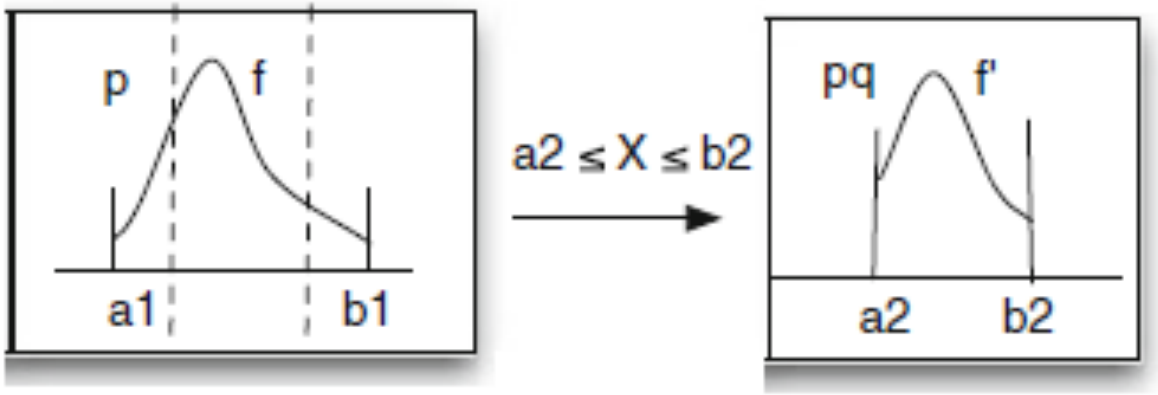
Location:(o_id, time, x^p, $luminosity^p$, $color^p$)

o_id	time	x^p	$luminosity^p$	$color^p$	p
1	08:15	10.1	201.25	R	0.9
2	10:24	9.5	98.63	G	0.8
3	15:30	21.3	312.6	B	0.75
4	17:42	32.7	135.8	Y	0.86

TEP

$$f(A^x = (10.1,201.25), A^y = R)$$
$$= f_{A^x|A^y}(10.1,201.25|R) * P(A^y = R)$$
$$= P(A^x = (10.1,201.25)) * P(A^y = R)$$

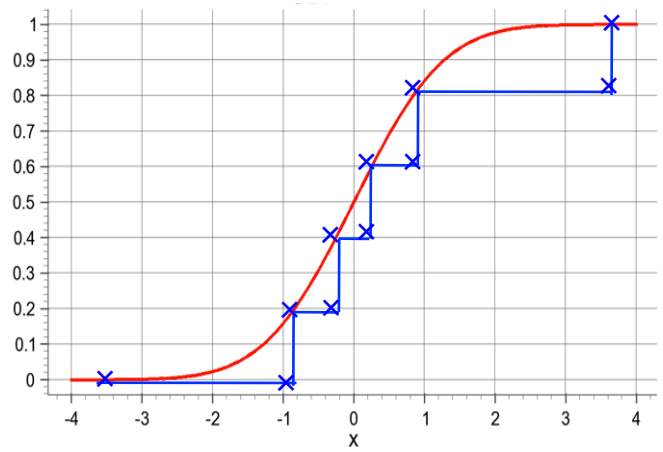
CLARO system
Selection

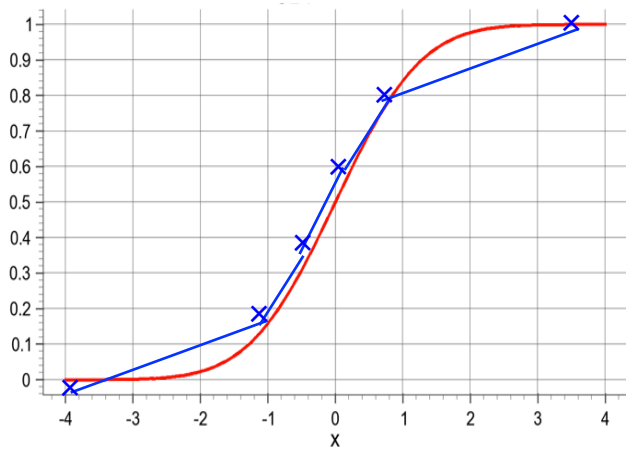


- Computation of the selection probability *q* (the probability mass of *f*= the joint density function in the selection range)
- Computation of the new tuple existence probability $p_{new} = p \cdot q$
- Truncation of joint distribution so that its support is restricted to the intersection of the original space and the selection range
- Normalization of the truncated distribution

CLARO system
Aggregation with TEP≤1

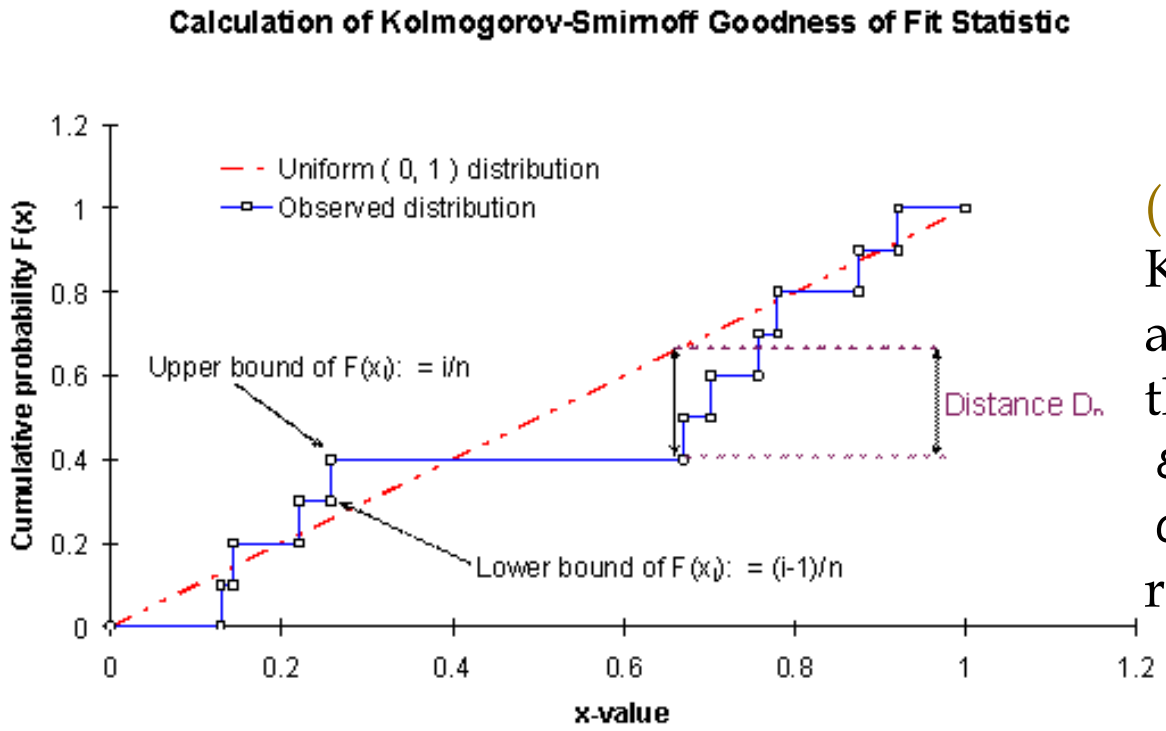
- Employ cumulative distribution functions (CDFs) (describes the probability that a real-valued random variable X with a given probability distribution will be found at a value less than or equal to x)
  - (1) It is non-decreasing function ranging from 0 to1
  - (2) It can be defined at any point in the real domain
- Two forms of CDFs: StepCDFs and LinCDFs





CLARO system
Aggregation with TEP≤1

- Approximation Framework: Metric
  - Kolmogorov-Smirnov (KS) distance:
    - Between two CDFs F, F' $KS(F, F') = \sup_x |F(x) - F'(x)|$



(ε , δ) approximation

KS distance between the approximate distribution and the exact distribution is at most ε , with probability (1- δ)

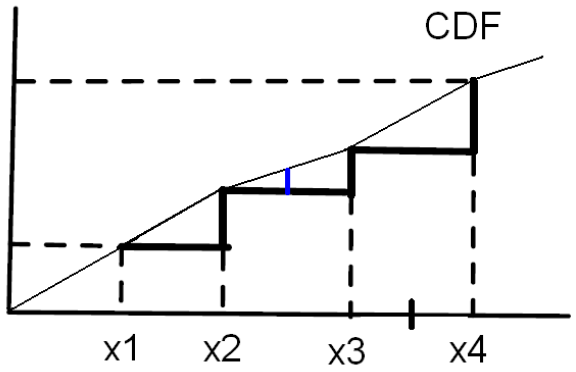
δ =0: deterministic, δ >0: randomized

CLARO system
Aggregation with TEP≤1

- Distributions of MIN and MAX

  $M_t = \max(Y_1, Y_2, \ldots, Y_t)$ (Y_i takes λ values from a finite universe)

  - Partition the universe into consecutive intervals dynamically
  - Maintain the estimates of the cumulative probabilities of its two ends
  - Since CDF is non-decreasing, if the estimates of the two ends are sufficiently close, none of these estimates is good for all the intermediate points



CLARO system
MAX: Analysis

1. Estimates of the two ends of an interval are bounded
   Estimates of any point in an interval are bounded

2. Number of intervals is bounded

$$|I| \leq 2 \log \epsilon^{-1} / \log(1 + \epsilon')$$

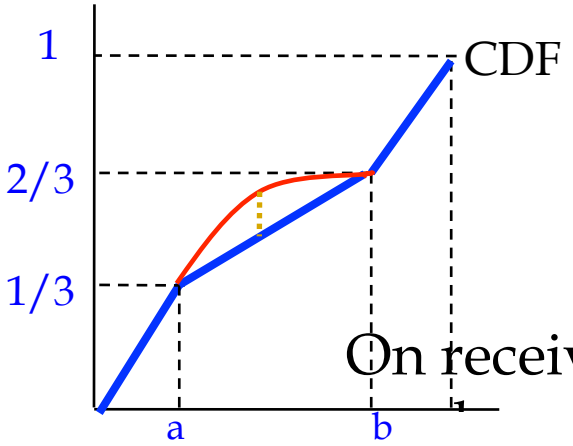
3. Number of times an interval is split is bounded, i.e., *logU*

(ε, 0) algorithm for max, update time is $O(\varepsilon^{-1} \log U \ln \varepsilon^{-1})$

Extend to continuous distributions:
A general approach is to consider a large universe of size 2^{64}. The complexity is then proportional to $\log 2^{64} = 64$.

CLARO system
Aggregation with TEP≤1

- Distributions of SUM: Approximate representation using quantiles (=the data values marking the boundaries between consecutive subsets)



$$P_{Q(\epsilon)}(F) = \{(x_1, \epsilon), (x_2, 2\epsilon), \ldots (x_k, 1)\}$$

$$\mathrm{KS}\left(F, \mathrm{LinCDF}_{P_{Q(\epsilon)}(F)}\right) \leq \epsilon$$

Assume each Y_i takes values from a finite set V_t of size at most λ

$$F_t^S(x) = \Sigma_{v \in V_t} F_{t-1}^S(x - v) P[Y_t = v]$$

On receiving each new tuple they can termediate approximation

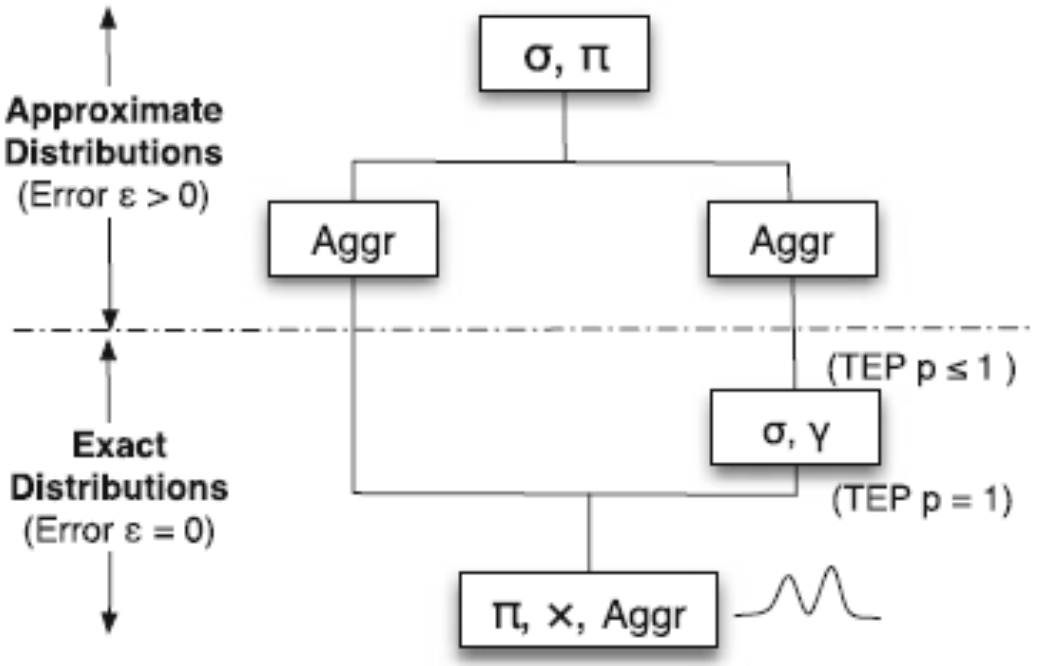
$$F(x) = \Sigma_{v \in V_t} LinCDF_{t-1}(x - v) P[Y_t = v]$$

CLARO system
Join on mixed-type tuples

- Two independent streams R(**A,R'**) , S(**A,B,S'**) :
  - TEP in R ≤ 1 : Each tuple in R →(p, f) then the join result tuple → (p', f'), where p=p' and f'=GMM
  - Uncertain S attributes :
    - When the view attributes B follow GMM the join result is also GMM.
    - When the join attributes A in B are uncertain →create views with smoothing techniques
  - TEP in S ≤1 : the sampling may return the empty set, indicating that a tuple does not exist → sampling more data points

CLARO system
Query planning

- How to handle errors due to mix of different operations in the context of a complete query



CLARO system
Query planning

- *Proposition on Selection:*
  - Selection on an attribute with (ε , δ) –approx, using a range condition is (2ε , δ)
  - If the selection uses a union of ranges, the approximation error is twice the sum, i.e., $2\varepsilon_i$
- Top-down approach to provision error bounds
  - If the error is ε , we should provision $\varepsilon/2$ for the approximation of sum

CLARO system
Conclusions

- CLARO system process data streams with two types of uncertainty:
  - Tuple existence uncertainty
  - Attributes uncertainty
- Query planning to compute the error that produce the approximation methods of operators
- Extension of this work to
  - Include query optimization
  - Exploit the correlations across tuples
  - Support user-defined functions

CLARO system
Conclusions

- Main drawbacks :
  - Hard to compute distributions for both continuous and discrete random variables (introduced by conditioning operations)
  - Reduced accuracy in the final results due to the approximate methods

Comparing two methods

PODS system

- One type of Uncertainty:
  - Attributes uncertainty
- Gaussian Mixture Model
- Operators:
  - Aggregation
  - Join

CLARO system

- Two types of Uncertainty:
  - Tuples existence uncertainty
  - Attributes Uncertainty
- Mixed-type Model (p,f)
- Operators:
  - Selection
  - Aggregation
  - Join
- Query planning for handling errors

Conclusions

- The intelligence of the methodologies lies in
  - The representation of uncertainties with continuous random variables following Gaussian distribution
  - The way of combining both uncertainty types
- We can decide how valid or not is an alert
- These systems have better performance for sparse data streams

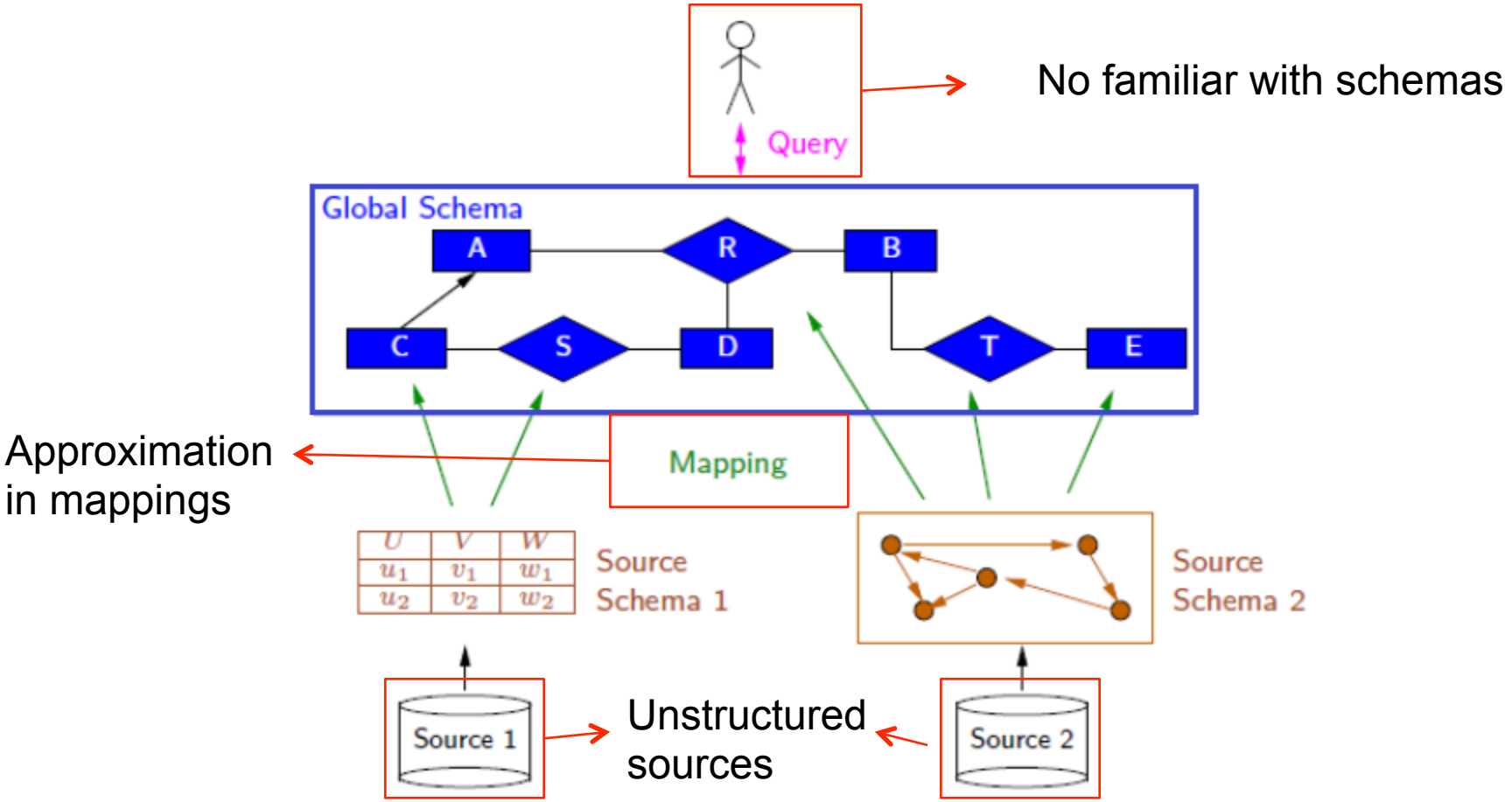
References

- "*PODS: A new data model and processing algorithms for uncertain data streams",* Thanh Tran, Liping Peng, Yanlei Diao, Andrew McGregor, Anna Liu
- "CLARO : *Modeling and processing uncertain data streams"*

  Thanh Tran, Liping Peng, Yanlei Diao, Andrew McGregor, Anna Liu
- *PROUD: A probabilistic approach to processing similarity queries over uncertain data streams,*

  Mi-Yen Yeh, Kun-Lung Wu, Philip yu, Min-Syan Chen

Sources of uncertainty in data integration



Thank you!!

Questions?