# Privacy and Data Mining: New Developments and Challenges

Stan Matwin

School of Information Technology and Engineering
Universit[é|y] [d' |of]Ottawa, Canada
stan@site.uottawa.ca

# Plan

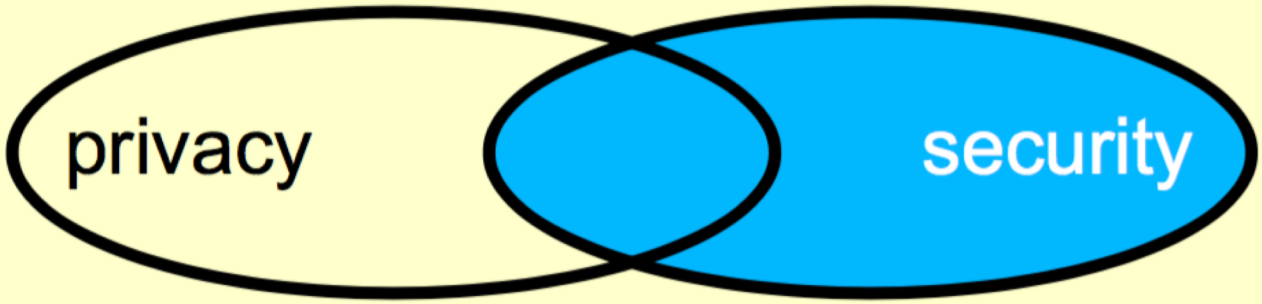
- Why privacy??
- Classification of Privacy-preserving Data Mining research (PPDM)
- Examples of current PPDM work
- Challenges

# Why privacy and data mining?…

- Like any technology can be used for « good » and « bad » purposes …
- It's Computer Science that has developed these tools, so…
- A moral obligation to develop solutions that will alleviate [potential] abuses and problems

# Privacy

- „fuzzy”, over-general concept
  - legal
  - economic
- Security?



# Instrumental vs. intrinsic value

## Instrumental

Privacy offers us protection against harm. For example, in some cases if person's medical condition were publicly known, then that person would risk discrimination

## Intrinsic

Rachaels suggests that privacy is valuable because it enables us to form varied relationships with other people. [Rachaels, 1975, p. 323]

Johnson proposes that we regard "privacy as an essential aspect of autonomy". [Johnson, 1994, p. 89]

# Instrumental vs. intrinsic value

https://9to5google.com/2018/04/11/zuckerberg-testimony-congress/

# Privacy

- Freedom from being watched ("*to be left alone*")
- …being able to control who knows what about us, and when [Moor]

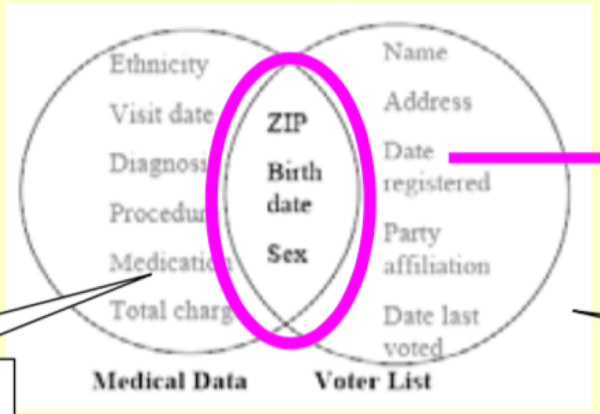
# Privacy

- A CS « perspective»
  - I am a database
  - Privacy is the ability to control the *views*
- Threats to privacy due to:
  - The Internet
  - Distributed databases
  - Data mining
- « greased » data

# …more precisely

- Privacy preservation: what does that mean?
- Given a table of instances (rows), we cannot associate any instance with a given person
- Naive anonymization…
- …is not sufficient, due to pseudo-identifiers

- L. Sweeney published this « attack » in 2001:
- **anoymized** (*de-linked*) health records of all 135,000 employees+families of the state of Massachussetts was placed on-line
- Electoral list of Cambridge, MA – bought for $20 (54 805 people)




Linking Attributes

Sensitive Data

Public Data

- 69% records are unique wrt birthdate, ZIP; 87% are unique wrt to bday, ZIP, sex…
- Governor's health records were identified
- …naive anonymization is not sufficient

# Other privacy fiascos

- AOL search engine queries published

  2006
- Netflix publicly released a data set containing movie ratings of 500,000 Netflix subscribers *between December* 1999 and December 2005.
- By matching no more than 8 movie ratings and approximate dates, 96% of subscribers can be uniquely identified.

# In statistics

- Statistical Disclosure Control
- A table is published, and the whole table has to be protected
- Risk/quality dilemma
- SDC ignores the use of the table
  - Classification
  - Associations
  - Distributed data

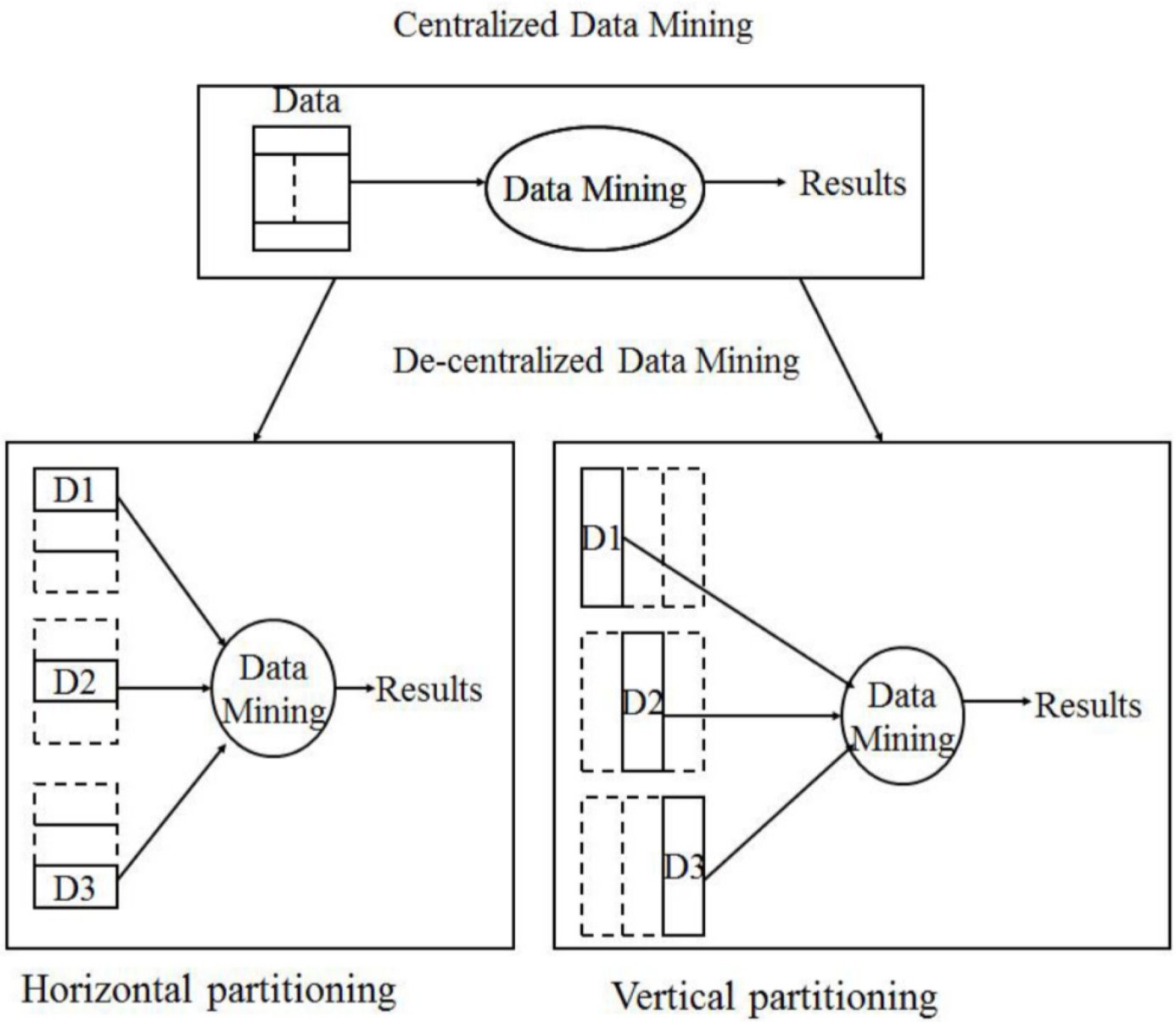
# Privacy-preserving Data Mining PPDM

- Data sharing
- Data publishing
- Cloud
- Two main dimensions:
  - What is being protected: data, results?
  - Data centralized or distributed?

# PPDM - dimensions

| | Data centralized | Data distributed |
|---|---|---|
| Protecting the data | •generalization/suppression [Sweeney]<br>•randomization [Du]/perturbation [Aggrawal] | •Horizontal/vertical: SMC-based [Clifton],<br>•Homomorphic encryption [Wright], [Zhang Matwin] |
| Protecting the results | *k*-anonymization of results :[Gianotti/Pedreschi] | [Jiang, Atziori], [Felty, Matwin] |

# Centralized vs. Distributed Data



# What is protected

- What is being protected:
  - the data: an attacker, given $T$,
    - will not be able to link any row in $T$ to a specific $i$ [**identity disclosure**]
    - will not be able to obtain a value $a_{ij}$ of a sensitive attribute $a_j$ of $i_i$ [**attribute disclosure**]
  - the inferred data mining result: an attacker, not knowing $T$ but given the results of the data mining operation, e.g. an association rule learned from $T$, will be able to identify some attributes of a specific $i_i$ [**model-based identity disclosure**]

# Privacy Goal: *k*-Anonymity

13

- Quasi-identifier (QID): The set of re-identification attributes.
- *k*-anonymity: Each record cannot be distinguished from at least *k-1* other records in the table wrt *QID*. [Sween98]

| Raw patient table | | | |
|---|---|---|---|
| **Job** | **Sex** | **Age** | **Disease** |
| Engineer | Male | 36 | Fever |
| Engineer | Male | 38 | Fever |
| Lawyer | Male | 38 | Hepatitis |
| Musician | Female | 30 | Flu |
| Musician | Female | 30 | Hepatitis |
| Dancer | Female | 30 | Hepatitis |
| Dancer | Female | 30 | Hepatitis |

| 3-anonymous patient table | | | |
|---|---|---|---|
| **Job** | **Sex** | **Age** | **Disease** |
| Professional | Male | [36-40] | Fever |
| Professional | Male | [36-40] | Fever |
| Professional | Male | [36-40] | Hepatitis |
| Artist | Female | [30-35] | Flu |
| Artist | Female | [30-35] | Hepatitis |
| Artist | Female | [30-35] | Hepatitis |
| Artist | Female | [30-35] | Hepatitis |

# Some Definitions

**Basic Notation.** Let $T = \{t_1, t_2, \ldots, t_n\}$ be a table with attributes $A_1, \ldots, A_m$. We assume that $T$ is a subset of some larger population $\Omega$ where each tuple represents an individual from the population. For example, if $T$ is a medical dataset then $\Omega$ could be the population of the United States. Let $\mathcal{A}$ denote the set of all attributes $\{A_1, A_2, \ldots, A_m\}$ and $t[A_i]$ denote the value of attribute $A_i$ for tuple $t$. If $\mathcal{C} = \{C_1, C_2, \ldots, C_p\} \subseteq \mathcal{A}$ then we use the notation $t[\mathcal{C}]$ to denote the tuple $(t[C_1], \ldots, t[C_p])$, which is the projection of $t$ onto the attributes in $\mathcal{C}$.

**Definition 2.1 (Quasi-identifier)** *A set of nonsensitive attributes $\{Q_1, \ldots, Q_w\}$ of a table is called a* quasi-identifier *if these attributes can be linked with external data to uniquely identify at least one individual in the general population $\Omega$.*

**Definition 2.2 ($k$-Anonymity)** *A table $T$ satisfies $k$-anonymity if for every tuple $t \in T$ there exist $k-1$ other tuples $t_{i_1}, t_{i_2}, \ldots, t_{i_{k-1}} \in T$ such that $t[\mathcal{C}] = t_{i_1}[\mathcal{C}] = t_{i_2}[\mathcal{C}] = \cdots = t_{i_{k-1}}[\mathcal{C}]$ for all $\mathcal{C} \in \mathcal{QI}$.*

l-Diversity: Privacy Beyond k-Anonymity by Machanavajjhala et al.

# Attacks on k-anonymity

## Homogeneity attack

| | Non-Sensitive | | | Sensitive |
|---|---|---|---|---|
| | Zip Code | Age | Nationality | Condition |
| 1 | 13053 | 28 | Russian | Heart Disease |
| 2 | 13068 | 29 | American | Heart Disease |
| 3 | 13068 | 21 | Japanese | Viral Infection |
| 4 | 13053 | 23 | American | Viral Infection |
| 5 | 14853 | 50 | Indian | Cancer |
| 6 | 14853 | 55 | Russian | Heart Disease |
| 7 | 14850 | 47 | American | Viral Infection |
| 8 | 14850 | 49 | American | Viral Infection |
| 9 | 13053 | 31 | American | Cancer |
| 10 | 13053 | 37 | Indian | Cancer |
| 11 | 13068 | 36 | Japanese | Cancer |
| 12 | 13068 | 35 | American | Cancer |

Fig. 1. Inpatient microdata.

| | Non-Sensitive | | | Sensitive |
|---|---|---|---|---|
| | Zip Code | Age | Nationality | Condition |
| 1 | 130** | $< 30$ | * | Heart Disease |
| 2 | 130** | $< 30$ | * | Heart Disease |
| 3 | 130** | $< 30$ | * | Viral Infection |
| 4 | 130** | $< 30$ | * | Viral Infection |
| 5 | 1485* | $\geq 40$ | * | Cancer |
| 6 | 1485* | $\geq 40$ | * | Heart Disease |
| 7 | 1485* | $\geq 40$ | * | Viral Infection |
| 8 | 1485* | $\geq 40$ | * | Viral Infection |
| 9 | 130** | 3* | * | Cancer |
| 10 | 130** | 3* | * | Cancer |
| 11 | 130** | 3* | * | Cancer |
| 12 | 130** | 3* | * | Cancer |

Fig. 2. 4-anonymous inpatient microdata.

*Homogeneity Attack.* Alice and Bob are antagonistic neighbors. One day Bob falls ill and is taken by ambulance to the hospital. Having seen the ambulance, Alice sets out to discover what disease Bob is suffering from. Alice discovers the 4-anonymous table of current inpatient records published by the hospital (Figure 2), and so she knows that one of the records in this table contains Bob's data. Since Alice is Bob's neighbor, she knows that Bob is a 31-year old American male who lives in the zip code 13053 (the quiet town of Dryden). Therefore, Alice knows that Bob's record number is 9, 10, 11, or 12. All of those patients have the same medical condition (cancer), and so Alice concludes that Bob has cancer.

# Attacks on k-anonymity

## Background knowledge attack

| | Non-Sensitive | | | Sensitive |
|---|---|---|---|---|
| | Zip Code | Age | Nationality | Condition |
| 1 | 13053 | 28 | Russian | Heart Disease |
| 2 | 13068 | 29 | American | Heart Disease |
| 3 | 13068 | 21 | Japanese | Viral Infection |
| 4 | 13053 | 23 | American | Viral Infection |
| 5 | 14853 | 50 | Indian | Cancer |
| 6 | 14853 | 55 | Russian | Heart Disease |
| 7 | 14850 | 47 | American | Viral Infection |
| 8 | 14850 | 49 | American | Viral Infection |
| 9 | 13053 | 31 | American | Cancer |
| 10 | 13053 | 37 | Indian | Cancer |
| 11 | 13068 | 36 | Japanese | Cancer |
| 12 | 13068 | 35 | American | Cancer |

Fig. 1. Inpatient microdata.

| | Non-Sensitive | | | Sensitive |
|---|---|---|---|---|
| | Zip Code | Age | Nationality | Condition |
| 1 | 130** | $< 30$ | * | Heart Disease |
| 2 | 130** | $< 30$ | * | Heart Disease |
| 3 | 130** | $< 30$ | * | Viral Infection |
| 4 | 130** | $< 30$ | * | Viral Infection |
| 5 | 1485* | $\geq 40$ | * | Cancer |
| 6 | 1485* | $\geq 40$ | * | Heart Disease |
| 7 | 1485* | $\geq 40$ | * | Viral Infection |
| 8 | 1485* | $\geq 40$ | * | Viral Infection |
| 9 | 130** | 3* | * | Cancer |
| 10 | 130** | 3* | * | Cancer |
| 11 | 130** | 3* | * | Cancer |
| 12 | 130** | 3* | * | Cancer |

Fig. 2. 4-anonymous inpatient microdata.

*Background Knowledge Attack.* Alice has a pen-friend named Umeko who is admitted to the same hospital as Bob and whose patient records also appear in the table shown in Figure 2. Alice knows that Umeko is a 21-year old Japanese female who currently lives in zip code 13068. Based on this information, Alice learns that Umeko's information is contained in record number 1,2,3, or 4. Without additional information, Alice is not sure whether Umeko caught a virus or has heart disease. However, it is well known that Japanese have an extremely low incidence of heart disease. Therefore Alice concludes with near certainty that Umeko has a viral infection.

# Homogeneity Attack on *k*-anonymity



- A data owner wants to release a table to a data mining firm for classification analysis on *Rating*

| Job | Country | Child | Bankruptcy | Rating | # Recs |
|---|---|---|---|---|---|
| Cook | US | No | Current | 0G/4B | 4 |
| Artist | France | No | Current | 1G/3B | 4 |
| Doctor | US | Yes | Never | 4G/2B | 6 |
| Trader | UK | No | Discharged | 4G/0B | 4 |
| Trader | UK | No | Never | 1G/0B | 1 |
| Trader | Canada | No | Never | 1G/0B | 1 |
| Clerk | Canada | No | Never | 3G/0B | 3 |
| Clerk | Canada | No | Discharged | 1G/0B | 1 |
| | | | | Total: | 24 |

- Inference: {Trader,UK} → fired
- Confidence = 4/5 = 80%
- An inference is sensitive if its confidence > threshold.

# p-Sensitive k-Anonymity

- for each equivalence class EC there is at least p distinct values for each sensitive attribute
- **Similarity attack** occurs when the values of sensitive attribute

| Age | Country | Zip Code | Health Condition |
|---|---|---|---|
| <30 | America | 142** | HIV |
| <30 | America | 142** | HIV |
| <30 | America | 142** | Cancer |
| <30 | America | 142** | Cancer |
| >40 | Asia | 130** | Hepatitis |
| >40 | Asia | 130** | Phthisis |
| >40 | Asia | 130** | Asthma |
| >40 | Asia | 130** | Heart Disease |
| 3* | America | 142** | Flu |
| 3* | America | 142** | Flu |
| 3* | America | 142** | Flu |
| 3* | America | 142** | Indigestion |

2-Sensitive 4-Anonymity

# l-Diversity

- every equivalence class in this table has at least $l$ *well represented* values for the sensitive attribute
- **Distinct *l*-diversity**: the number of distinct values for a sensitive attribute in each equivalence class to be at least $l$.
- $l$ -Diversity may be difficult and unnecessary to achieve and it may cause a huge information loss.

| | Non-Sensitive | | | Sensitive |
|---|---|---|---|---|
| | Zip Code | Age | Nationality | Condition |
| 1 | 1305* | ≤ 40 | * | Heart Disease |
| 4 | 1305* | ≤ 40 | * | Viral Infection |
| 9 | 1305* | ≤ 40 | * | Cancer |
| 10 | 1305* | ≤ 40 | * | Cancer |
| 5 | 1485* | > 40 | * | Cancer |
| 6 | 1485* | > 40 | * | Heart Disease |
| 7 | 1485* | > 40 | * | Viral Infection |
| 8 | 1485* | > 40 | * | Viral Infection |
| 2 | 1306* | ≤ 40 | * | Heart Disease |
| 3 | 1306* | ≤ 40 | * | Viral Infection |
| 11 | 1306* | ≤ 40 | * | Cancer |
| 12 | 1306* | ≤ 40 | * | Cancer |

3-diverse data [4]

# t-closeness

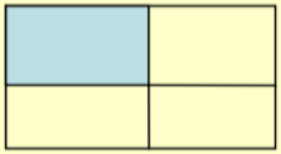
- An equivalence class EC is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold *t*. [5].

- It solves the attribute disclosure problems of l-diversity, i.e. skewness attack and similarity attack, [6]

| | ZIP Code | Age | Salary | Disease |
|---|---|---|---|---|
| 1 | 4767* | $\leq 40$ | 3K | gastric ulcer |
| 3 | 4767* | $\leq 40$ | 5K | stomach cancer |
| 8 | 4767* | $\leq 40$ | 9K | pneumonia |
| 4 | 4790* | $\geq 40$ | 6K | gastritis |
| 5 | 4790* | $\geq 40$ | 11K | flu |
| 6 | 4790* | $\geq 40$ | 8K | bronchitis |
| 2 | 4760* | $\leq 40$ | 4K | gastritis |
| 7 | 4760* | $\leq 40$ | 7K | bronchitis |
| 9 | 4760* | $\leq 40$ | 10K | stomach cancer |

0.167-closeness w.r.t. salary and

0.278-closeness w.r.t. Disease[5]



# Two basic approaches

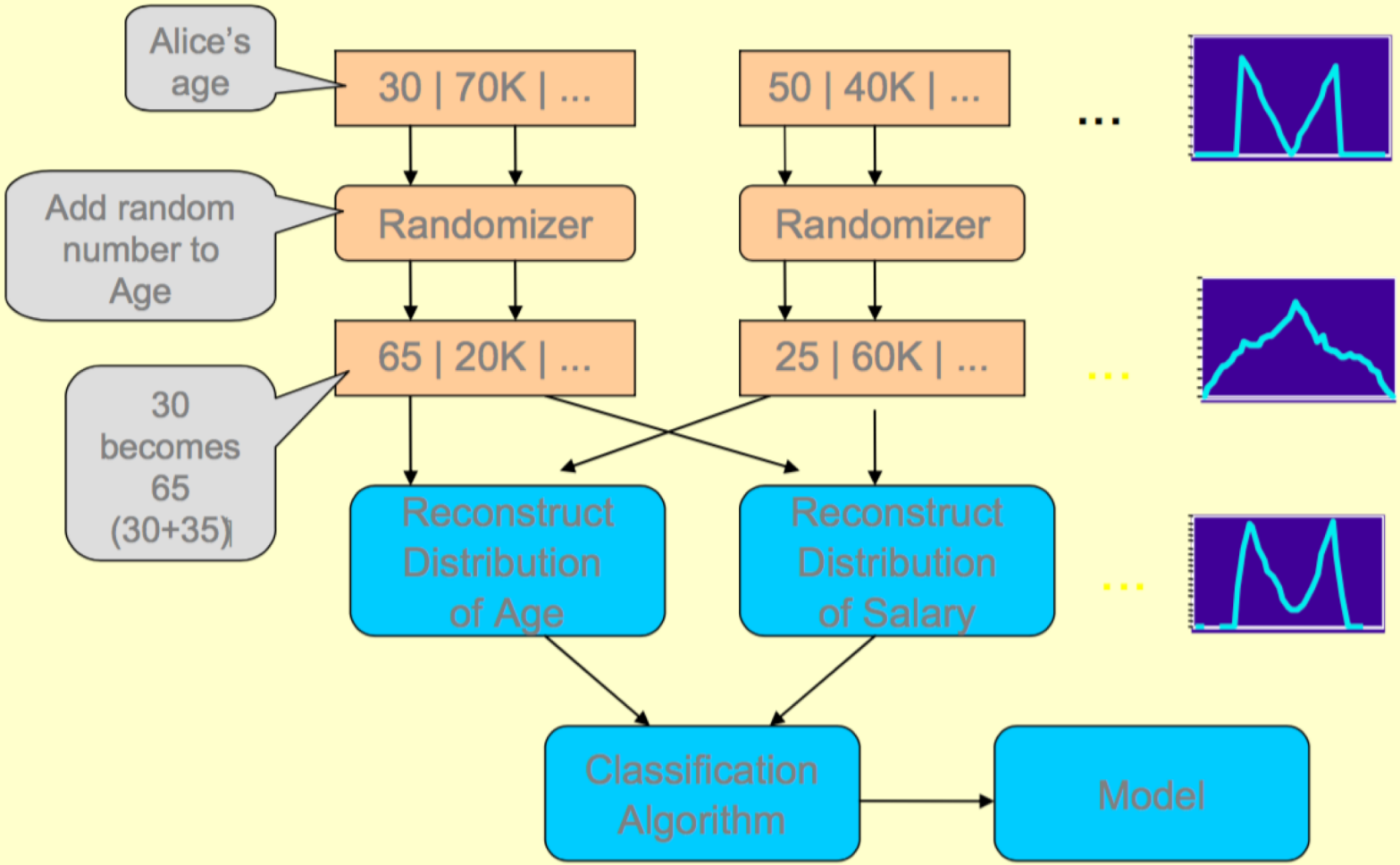
camouflage

Data modification/perturbation

hiding in the crowd

k-anonymization

# Randomization

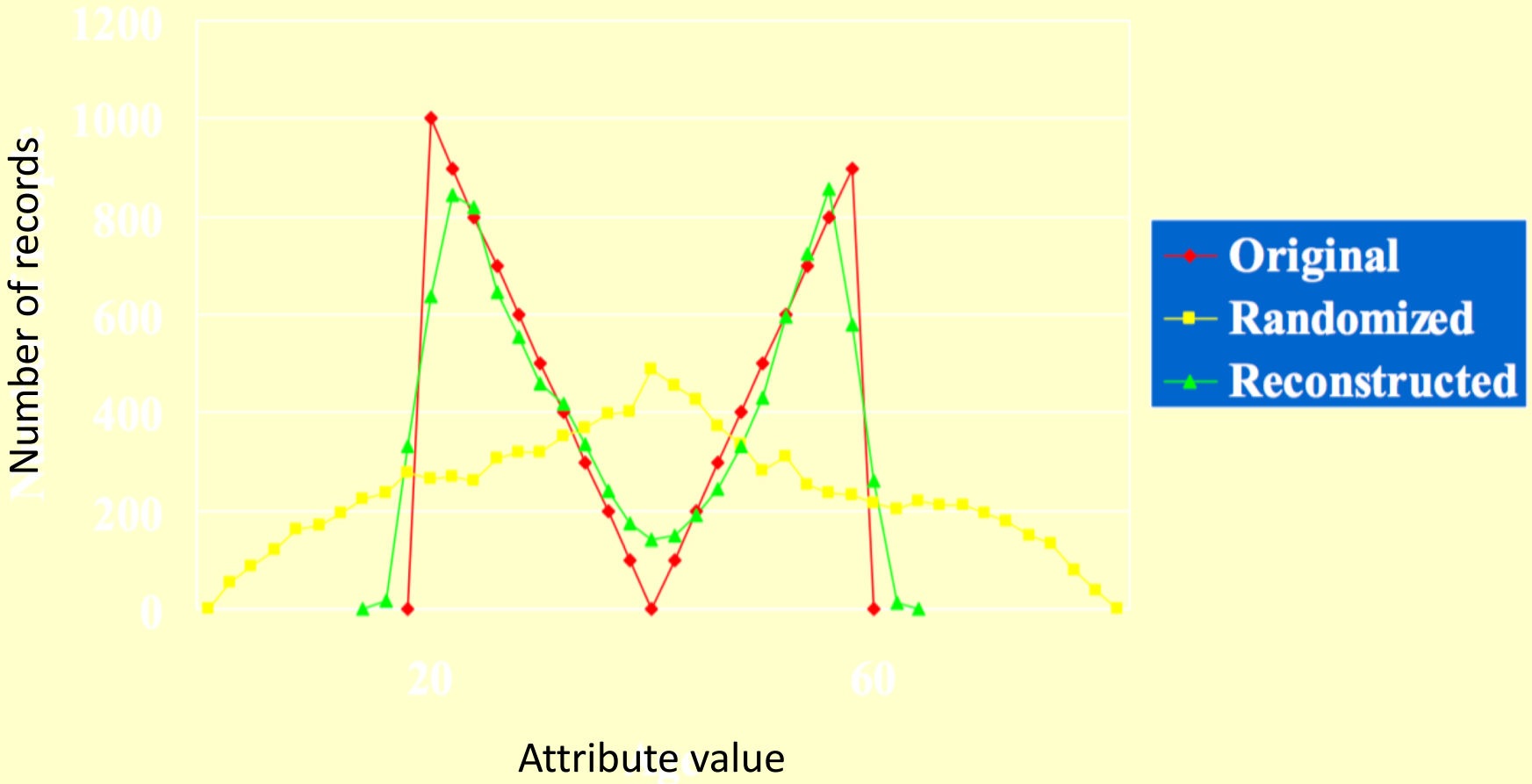


# Reconstruction (*linking*)

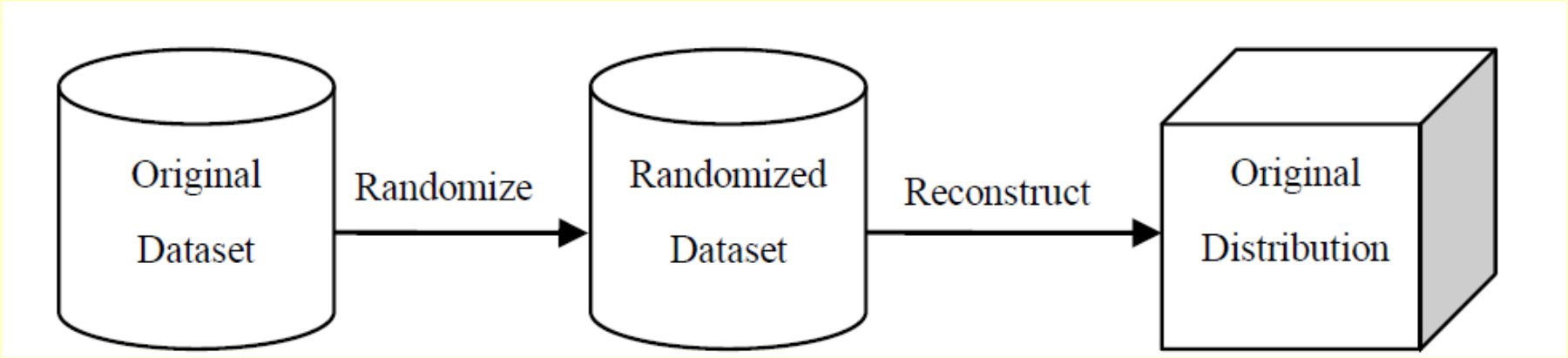
- initial (confidential) values $x_1, x_2, \ldots, x_n$ have an (uknown) distribution $X$
- For protection, we perturb them with values $y_1, y_2, \ldots, y_n$ with a *known* distribution $Y$
- given
  - $x_1+y_1, x_2+y_2, \ldots, x_n+y_n$
  - distribution $Y$

  Find an estimation of the distribution $X$.

# Works well


1200
1000
800
600
400
200
0
Number of records
20
60
Attribute value
Original
Randomized
Reconstructed

# Randomized Response Model (RRM)




Randomized Response technique was first introduced by Warner [10] in 1965 as a technique to solve the following survey problem: to estimate the percentage of people in a population that has attribute $A$, queries are sent to a group of people.

1. I have the sensitive attribute $A$.
2. I do not have the sensitive attribute $A$.

$$
\begin{aligned}
P^*(A = yes) &= P(A = yes) \cdot \theta + P(A = no) \cdot (1 - \theta) \\
P^*(A = no) &= P(A = no) \cdot \theta + P(A = yes) \cdot (1 - \theta)
\end{aligned}
$$

# RRM for Decision Trees

## Condition: attributes are perturbed jointly

$$Gain(S, A) = Entropy(S) - \sum_{v \in A} \left(\frac{|S_v|}{|S|} Entropy(S_v)\right)$$

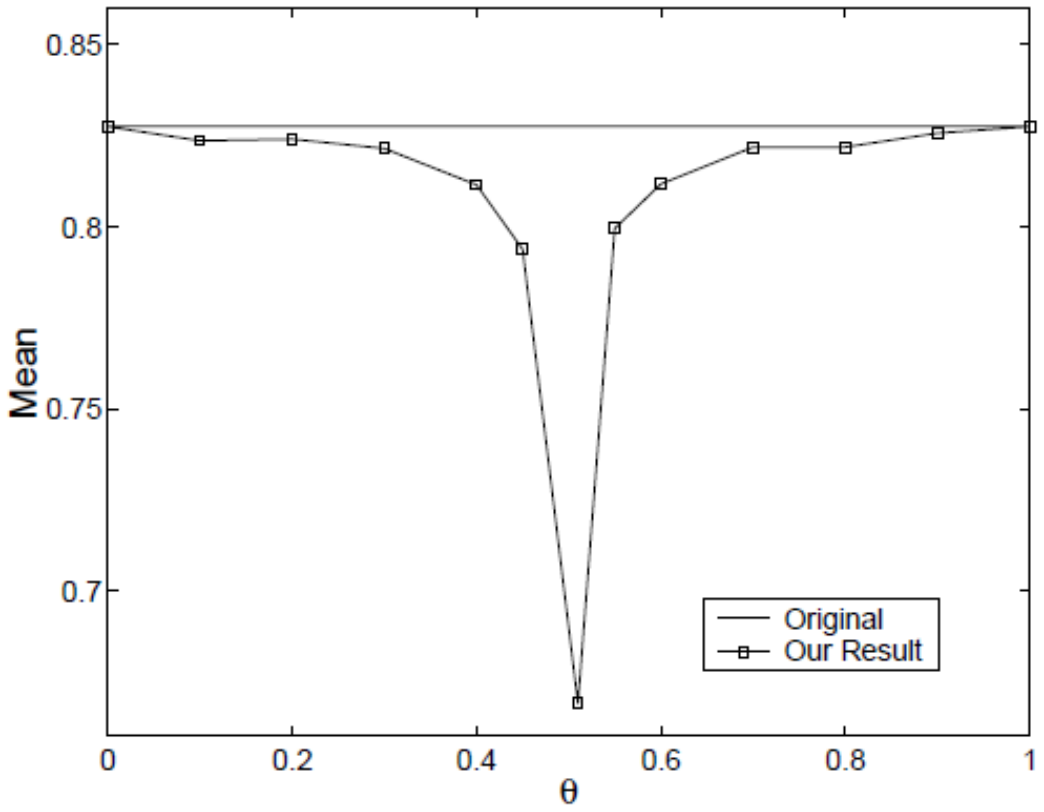
$$\begin{aligned} E &= (A_i = 1) \wedge (A_j = 0) \\ \overline{E} &= (A_i = 0) \wedge (A_j = 1) \end{aligned}$$

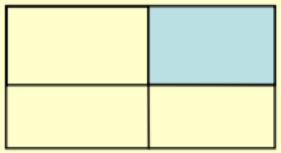
$$\begin{aligned} P^*(E) &= P(E) \cdot \theta + P(\overline{E}) \cdot (1 - \theta) \\ P^*(\overline{E}) &= P(\overline{E}) \cdot \theta + P(E) \cdot (1 - \theta) \end{aligned}$$

$\theta = 0$ $\rightarrow$ no privacy

$\theta = \frac{1}{2}$ $\rightarrow$ most privacy

DT accuracy on Census data





# Distributed data

- Vehicle/accident data
- To discover the causes of accidents we need to know the attributrs of different components from different manufacturers (brakes, tires)
- They will nolt disclose these values in the open
- **Vertical** partition

# Distributed data

- A medical study carried out in several hospitals
- Would like to *merge* the data for bigger impact of results (results on 20 000 patients instead of 5 000 each)
- For legal reasons, cannot just share then open data
- Horizontal partition

# *Association Rule Mining Algorithm* [Agrawal et al. 1993]

1. $L_1$ = large 1-itemsets
2. for $(k = 2; L_{k-1} \neq \phi; k++)$ do begin
3. $C_k = apriori - gen(L_{k-1})$
4. for all candidates $c \in C_k$ do begin
5. compute ***c.count***
6. end
7. $L_k = \{c \in C_k \mid c.count \geq \min\text{-}\sup\}$
8. end
9. Return $L = \bigcup_k L_k$

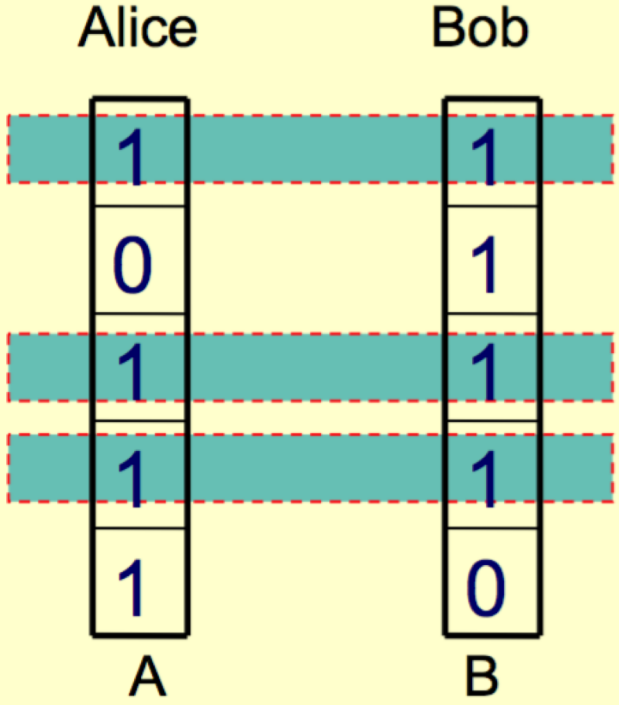
c.count is the frequency of an *itemset*.

to compute frequency, we need access to values of attributes belonging to different parties

# Example

- c.count is the scalar product.
- $A$ = Alice's attribute vector, $B$ = Bob'
- $AB$ is a candidate frequent itemset
- c.count = $A \bullet B = 3$.

- How to perform the scalar product preserving the privacy of Alice and Bob?



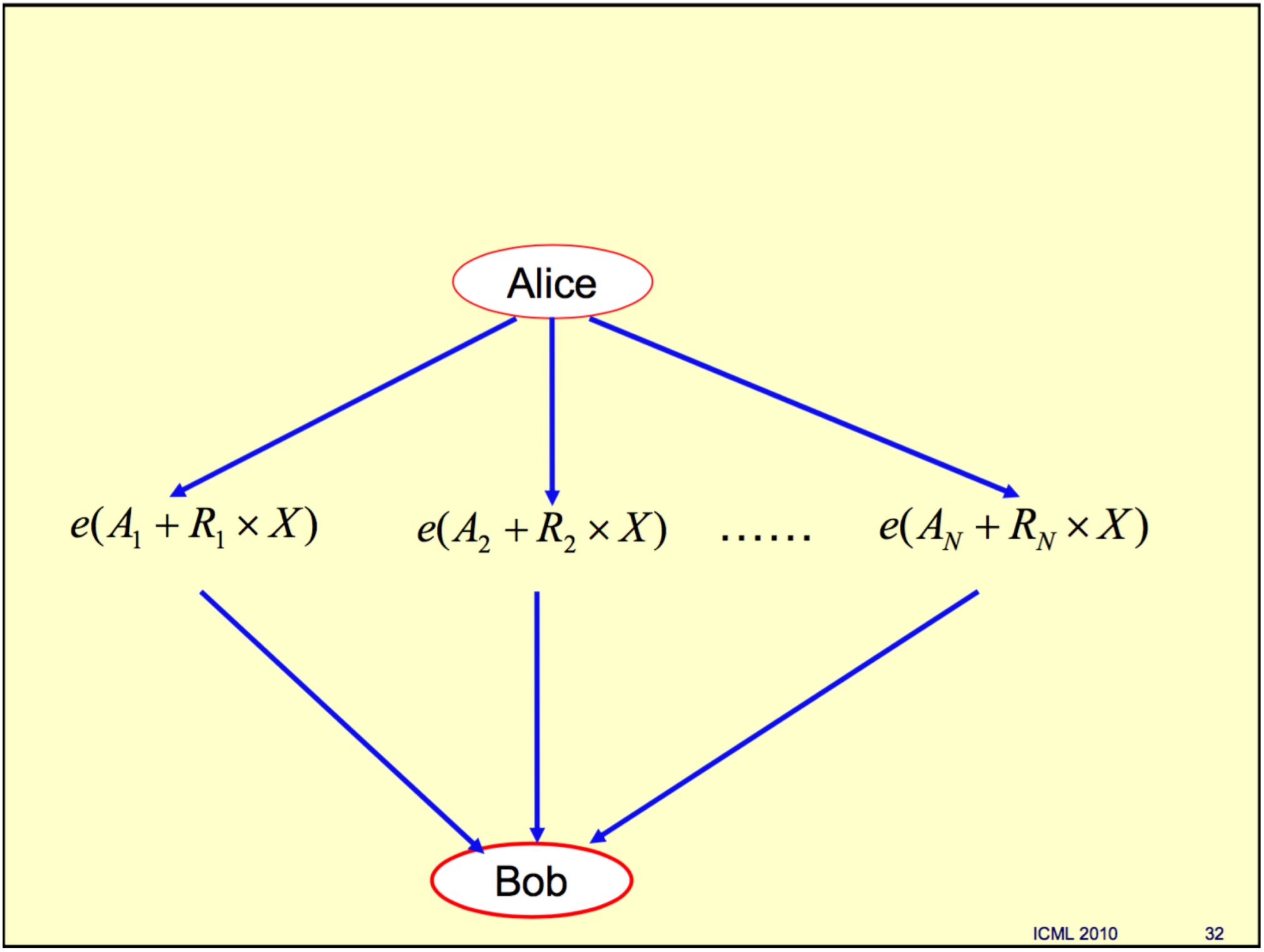
# *Homomorphic Encryption*

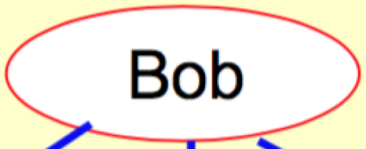
[Paillier 1999]

- Privacy-preserving protocol based on the concept of homomorphic encryption
- The homomorphic encryption property is

$$e(m_1) \times e(m_2) \times \cdots \times e(m_n) = e(m_1 + m_2 + \cdots + m_n)$$

- *e* is an encryption function $e(m_i) \neq 0$


Alice
$e(A_1 + R_1 \times X)$
$e(A_2 + R_2 \times X)$
……
$e(A_N + R_N \times X)$
Bob




$$W_1 = e(A_1 + R_1 \times X) \times B_1 \quad W_2 = e(A_2 + R_2 \times X) \times B_2 \quad \cdots W_N = e(A_N + R_N \times X) \times B_N$$

$$B_i = 0 \Rightarrow W_i = 0$$

$$B_i = 1 \Rightarrow W_i = e(A_i + R_i \times X) \times B_i = e(A_i + R_i \times X)$$

Bob computes $W' = [\prod_{j \neq 0} W_j] \bmod X = [\prod_{j \neq 0} e(A_j + R_j \times X)] \bmod X = [e(A_{j_1} + \ldots A_{j_m} +, (R_{j_1} + \ldots R_{j_m}) \times X] \bmod X$
encrypts , sends to Alice

# Last stage

- Alice decrypts $W'$ and computes modulo X.

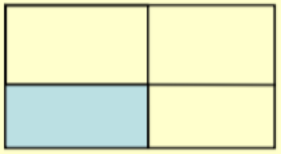
$$c.count$$

$$= d(e(A_1 + A_2 + \cdots + A_j + (R_1 + R_2 + \cdots + R_j + R') \times X)) \bmod X$$

$$(A_1 + A_2 + \cdots + A_j) \leq N < X$$

$$((R_1 + R_2 + \cdots + R_j + R') \times X) \bmod X = 0$$

- She obtains $A_1 + A_2 + \cdots + A_j$ for these $A_j$ whose corresponding $B_j$ are not 0, which is = c.count
- Privacy analysis



# Now looking at data mining results…

Can data mining results reveal personal information?

In some cases, yes: [Atzori et al. 05]:

An association rule :

$$a_1 \wedge a_2 \wedge a_3 \Rightarrow a_4 [\text{sup} = 80, conf = 98.7\%]$$

Means that $\text{sup}(\{a_1, a_2, a_3, a_4\}) = 80$

So $\text{sup}(\{a_1, a_2, a_3\}) = \frac{\text{sup}(\{a_1, a_2, a_3, a_4\})}{0.987} = \frac{0.8}{.0987} = 81.05$

And $a_1 \wedge a_2 \wedge a_3 \wedge \neg a_4$ has support =1, and identifies a person!!

# Protecting data mining results

- A *k-anonymous patterns* approach and an algorithm (*inference channels*) detect violations of *k*-anonymity of results

# Discrimination and data mining

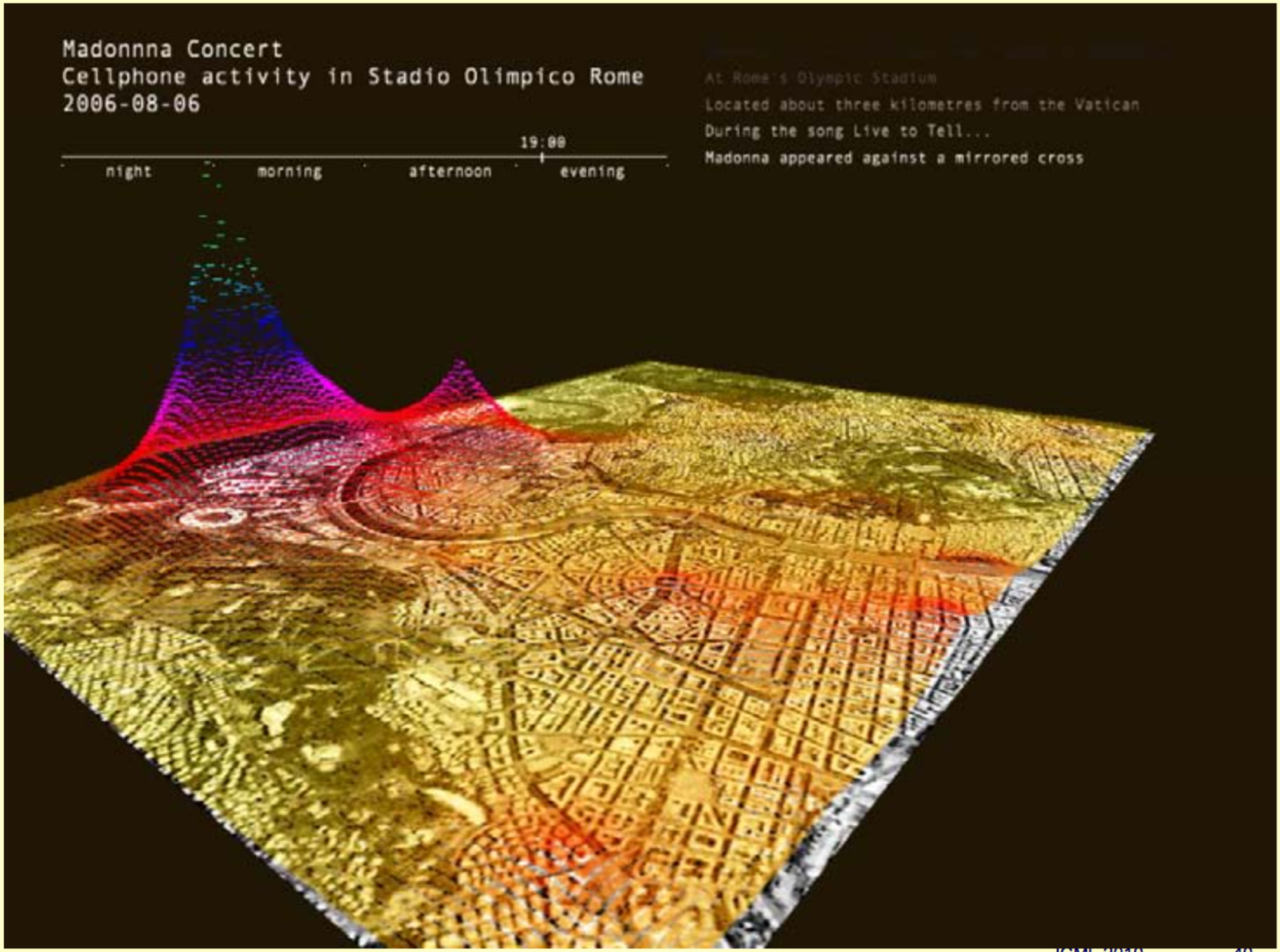
- [Pedreschi et al 07] shows how DM results can lead to discriminatory rules
- In fact, DM's goal is discrimination (between different sub-groups of data)
- They propose a measure of potential discrimination with lift : to what extent a sensitive is more assigned by a rule to a sensitive group than to an average group

# Other challenges

- Privacy and social networks
- Privacy definition – where to look for inspiration (economics?)
- Text data – perturbation/anonymization methods don't work
- Medical data: trails [Malin], privacy of longitudinal data
- Mobile data -

# GeoPKDD

- European project on Geographic Privacy-aware Knowledge Discovery and Delivery
- Data from GSM/UMTS and GPS


Madonnna Concert
Cellphone activity in Stadio Olimpico Rome
2006-08-06
19:00
night
morning
afternoon
evening
At Rome's Olympic Stadium
Located about three kilometres from the Vatican
During the song Live to Tell...
Madonna appeared against a mirrored cross

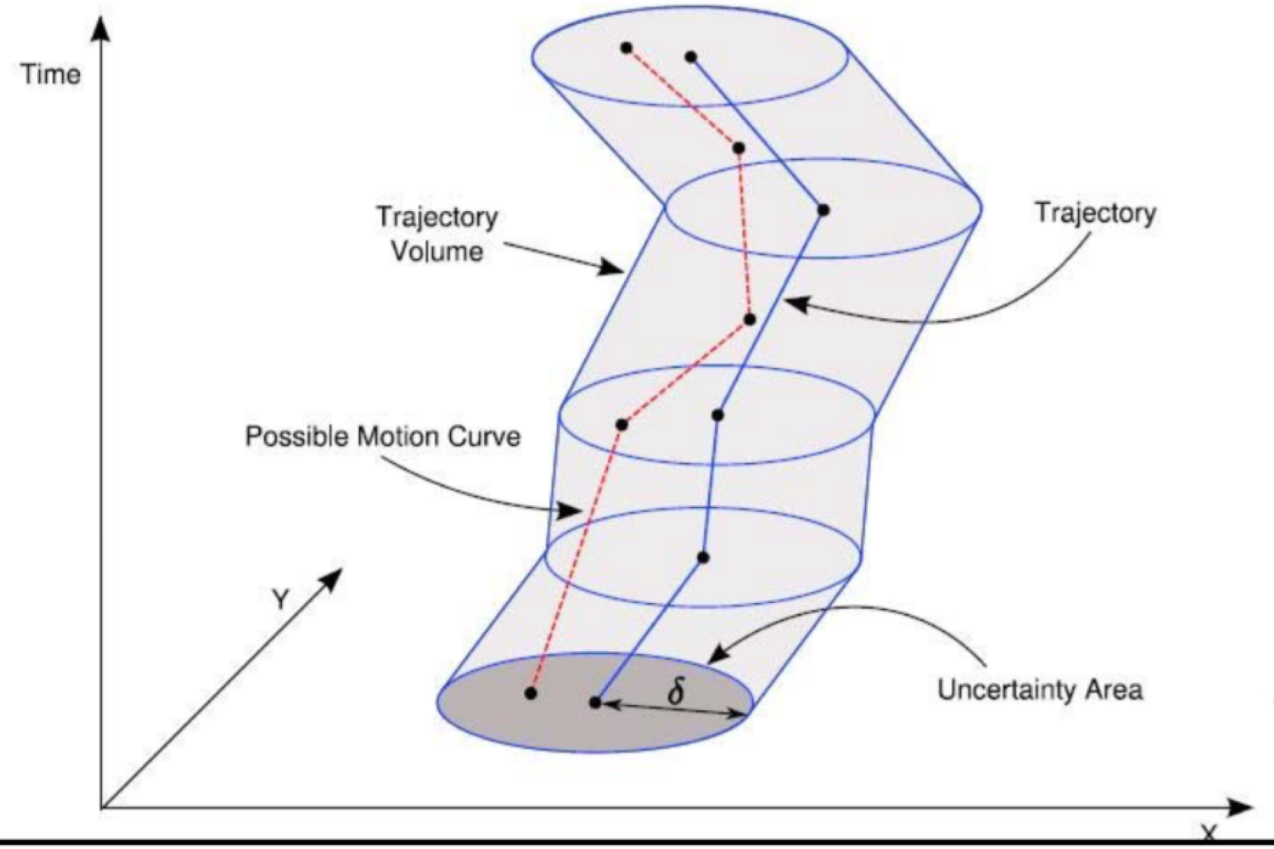
# First obtaining spatio-temporal trajectories, then patterns

Trajectory = sequence of points visiteddans in a temporal order

pattern= set of frequent trajectories with similar transition times

# Privacy of spatio-temporal data

- Modify the data in such a way each trajectory be indistinguishable from k other trajectories
- … by minimizing distorsion introduced into the data



# Conclusion

- A major challenge for database/data mining research
- Lots of interesting  contributions/papers, but lack of a systematic framework
- …?