Privacy and anonymity in data publishing and (mobility) data mining

Fosca Giannotti, Dino Pedreschi, Franco Turini

Pisa KDD Laboratory
Università di Pisa and ISTI-CNR, Pisa, Italy

Dottorato di Ricerca in Informatica, Scuola Galilei

Università di Pisa. Giugno-Luglio 2009

Privacy Preserving Data Mining: an overview

Fosca Giannotti & Dino Pedreschi

KDD Lab Pisa

PhD course, Department of Computer ScneceGiugno Luglio 2009

Plan of the Talk

- Motivations
  - re-identification examples
- Privacy-preserving data publishing & mining:
  - a condensed state-of-the-art:
    - Data publishing: k-anonymity
    - Data publishing: Data Perturbation and Obfuscation
    - Knowledge Publishing
    - Distributed Privacy Preserving Data Mining
    - Knowledge Hiding
    - Privacy Preserving Outsourcing
- Pointers to Resources

Traces

- Our everyday actions leave digital **traces** into the information systems of ICT service providers.
  - mobile phones and wireless communication,
  - web browsing and e-mailing,
  - credit cards and point-of-sale e-transactions,
  - e-banking,
  - electronic administrative transactions and health records,
  - shopping transactions with loyalty cards.

Traces: forget or remember?

- When no longer needed for service delivery, traces can be either forgotten or stored.
  - Storage is cheaper and cheaper.
- But why should we store traces?
  - From business-oriented information – sales, customers, billing-related records, …
  - To finer grained process-oriented information about how a complex organization works.
- Traces are worth being remembered because they may hide precious knowledge about the processes which govern the life of complex economical or social systems.

THE example: wireless networks

- Wireless phone networks gather highly informative traces about the human mobile activities in a territory
  - miniaturization
  - pervasiveness
    - 1.5 billions in 2005, still increasing at a high speed
    - Italy: # mobile phones ≈ # inhabitants
  - positioning accuracy
    - location technologies capable of providing increasingly better estimate of user location

Opportunities and threats

- Knowledge may be discovered from the traces left behind by mobile users in the information systems of wireless networks.
- Knowledge, in itself, is neither good nor bad.
- What knowledge to be searched from digital traces? For what purposes?
- Which **eyes** to look at these traces with?

The Spy and the Historian

- The malicious eyes of the **Spy** – or the detective – aimed at
  - discovering the individual knowledge about the behaviour of a single **person** (or a small group)
  - for **surveillance** purposes.
- The benevolent eyes of the **Historian** – or the archaeologist, or the human geographer – aimed at
  - discovering the collective knowledge about the behaviour of whole **communities**,
  - for the purpose of **analysis**, of understanding the dynamics of these communities, the way they live.

The privacy problem

- the donors of the mobility data are ourselves the citizens,
- making these data available, even for analytical purposes, would put at risk our own privacy, our right to keep secret
  - the places we visit,
  - the places we live or work at,
  - the people we meet
  - ...

The naive scientist's view (1)

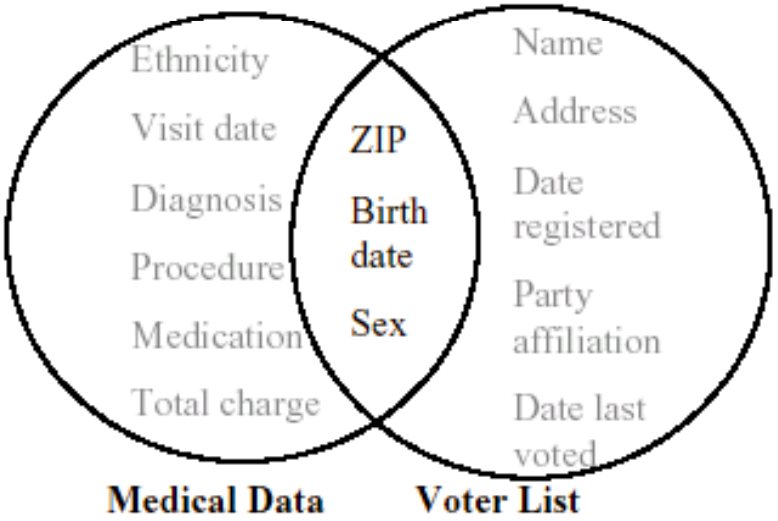
- Knowing the exact identity of individuals is not needed for analytical purposes
  - Anonymous trajectories are enough to reconstruct aggregate movement behaviour, pertaining to groups of people.
- Is this reasoning correct?
- Can we conclude that the analyst runs no risks, while working for the public interest, to inadvertently put in jeopardy the privacy of the individuals?

Unfortunately not!

- Hiding identities is not enough.
- In certain cases, it is possible to reconstruct the exact identities from the released data, even when identities have been removed and replaced by pseudonyms.
- A famous example of re-identification by L. Sweeney

Re-identifying "anonymous" data (Sweeney '01)

- She purchased the voter registration list for Cambridge Massachusetts
  - 54,805 people



- 69% unique on postal code and birth date
- 87% US-wide with all three (ZIP + birth date + Sex)

- Solution: *k*-anonymity
  - Any combination of values appears at least *k* times
- Developed systems that guarantee k-anonymity
  - Minimize distortion of results

Private Information in Publicly Available Data

Date of Birth	Zip Code	Allergy	History of Illness
03-24-79	07030	Penicillin	Pharyngitis
08-02-57	07028	No Allergy	Stroke
11-12-39	07030	No Allergy	Polio
08-02-57	07029	Sulfur	Diphtheria
08-01-40	07030	No Allergy	Colitis

Medical Research Database

Sensitive Information

Linkage attack: Link Private Information to Person

Quasi-identifiers

~~Date of Birth~~	Zip Code	Allergy	History of Illness
03-24-79	07030	Penicillin	Pharyngitis
08-02-57	07028	No Allergy	Stroke
11-12-39	07030	No Allergy	Polio
08-02-57	07029	Sulfur	Diphtheria
08-01-40	07030	No Allergy	Colitis

Victor is the only person born 08-02-57 in the area of 07028… Ha, he has a history of stroke!

Sweeney's experiment

- Consider the governor of Massachusetts:
  - only 6 persons had his birth date in the joined table (voter list),
  - only 3 of those were men,
  - and only … 1 had his own ZIP code!
- The medical records of the governor were uniquely identified from legally accessible sources!

The naive scientist's view (2)

- Why using quasi-identifiers, if they are dangerous?
- A brute force solution: replace identities or quasi-identifiers with totally unintelligible codes
- Aren't we safe now?
- No! Two examples:
  - The AOL August 2006 crisis
  - Movement data

A face is exposed for AOL searcher no. 4417749 [New York Times, August 9, 2006]

- No. 4417749 conducted hundreds of searches over a three months period on topics ranging from “numb fingers” to “60 single men” to “dogs that urinate on everything”.

- And search by search, click by click, the identity of AOL user no. 4417749 became easier to discern. There are queries for “landscapers in Lilburn, Ga”, several people with the last name Arnold and “homes sold in shadow lake subdivision gwinnet county georgia”.

A face is exposed for AOL searcher no. 4417749 [New York Times, August 9, 2006]

- It did not take much investigating to follow this **data trail** to Thelma Arnold, a 62-year-old widow of Lilburn, Ga, who loves her three dogs. “Those are my searches,” she said, after a reporter read part of the list to her.
- Ms. Arnold says she loves online research, but the disclosure of her searches has left her disillusioned. In response, she plans to drop her AOL subscription. “We all have a right to privacy,” she said, “Nobody should have found this all out.”
- http://data.aolsearchlogs.com

Mobility data example: spatio-temporal linkage [Jajodia et al. 2005]

- An anonymous trajectory occurring every working day from location A in the suburbs to location B downtown during the morning rush hours and in the reverse direction from B to A in the evening rush hours can be linked to
  - the persons who live in A and work in B;
- If locations A and B are known at a sufficiently fine granularity, it possible to identify specific persons and unveil their daily routes
  - Just join phone directories
- In mobility data, positioning in space and time is a powerful quasi identifier.

The naive scientist's view (3)

- In the end, it is not needed to disclose the data: the (trusted) analyst only may be given access to the data, in order to produce knowledge (mobility patterns, models, rules) that is then disclosed for the public utility.

- Only **aggregated information is published**, while **source data are kept secret**.

- Since aggregated information concerns **large** groups of individuals, we are tempted to conclude that its disclosure is safe.

Wrong, once again!

- Two reasons (at least):
- For **movement patterns**, which are sets of trajectories, the control on space granularity may allow us to re-identify a small number of people
  - Privacy (anonymity) **measures** are needed!
- From **rules** with high support (i.e., concerning many individuals) it is sometimes possible to deduce new rules with very limited support, capable of identifying precisely one or few individuals

An example of rule-based linkage [Atzori et al. 2005]

- **Age = 27 and**
  ZIP = 45254 and
  Diagnosis = HIV $\Rightarrow$ Native Country = USA
  [sup = 758, conf = 99.8%]
- Apparently a safe rule:
  - **99.8% of 27-year-old people from a given geographic area that have been diagnosed an HIV infection, are born in the US.**
- But we can derive that only the 0.2% of the rule population of 758 persons are 27-year-old, live in the given area, have contracted HIV and **are not born in the US**.
  - **1 person only! (without looking at the source data)**
- The triple Age, ZIP code and Native Country is a quasi-identifier, and it is possible that in the demographic list there is only one 27-year-old person in the given area who is not born in the US (as in the governor example!)

Moral: protecting privacy when disclosing information is not trivial

- Anonymization and aggregation do not necessarily put ourselves on the safe side from attacks to privacy
- For the very same reason the problem is scientifically attractive – besides socially relevant.
- As often happens in science, the problem is to find an optimal trade-off between two conflicting goals:
  - obtain **precise, fine-grained** knowledge, useful for the analytic eyes of the Historian;
  - obtain **imprecise, coarse-grained** knowledge, useless for the sharp eyes of the Spy.

Privacy-preserving data publishing and mining

- Aim: guarantee anonymity by means of controlled transformation of data and/or patterns
  - little distortion that avoids the undesired side-effect on privacy while preserving the possibility of discovering useful knowledge.
- An exciting and productive research direction.

Privacy Preserving Data Publishing & Mining: Condensed State of the Art

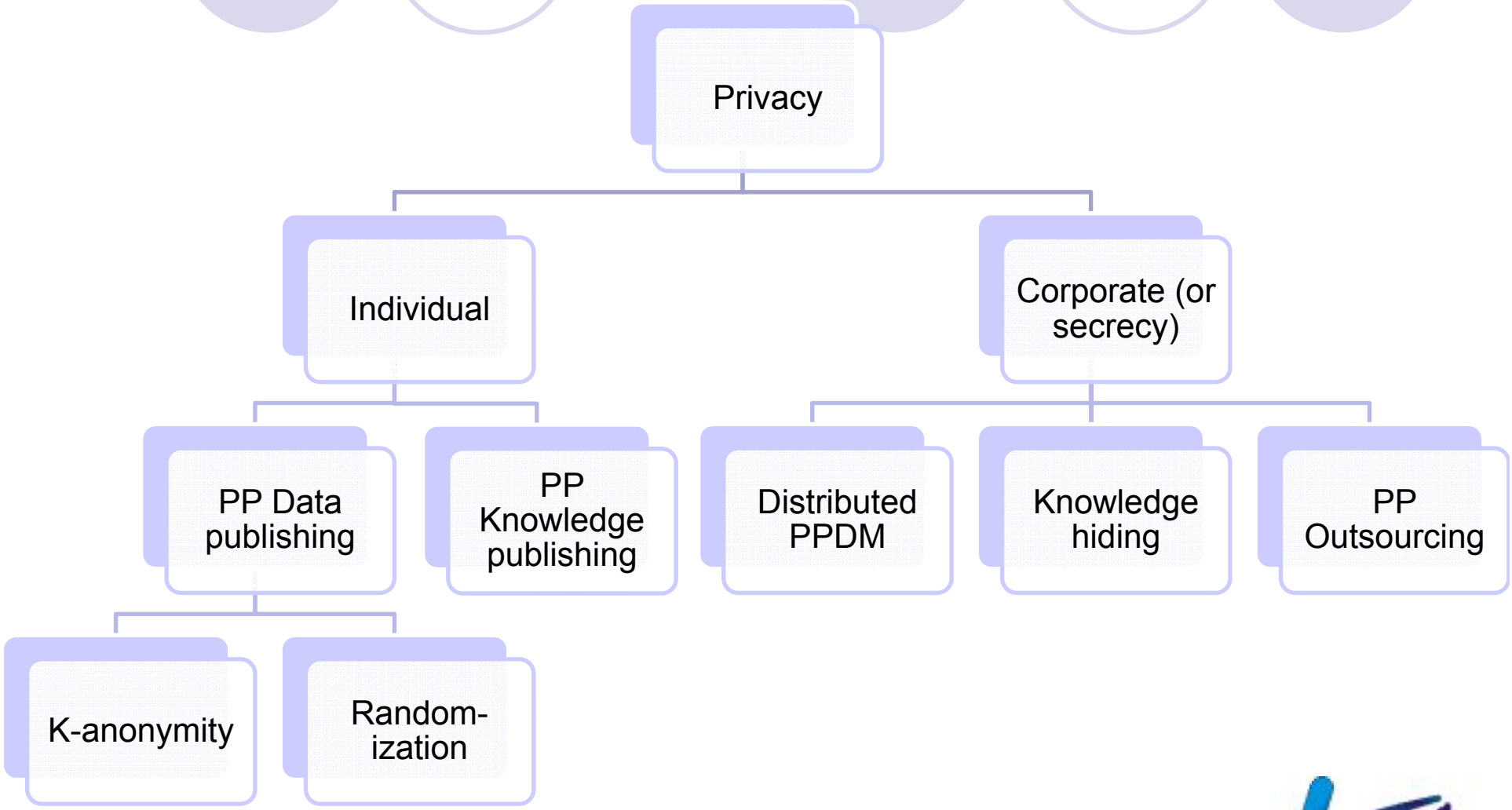
Privacy Preserving Data Mining

- We identify 6 main approaches, distinguished by the following questions:
  - *what is disclosed/published/shared?*
  - *what is hidden?*
  - *how is the data organized? (centralized or distributed)*

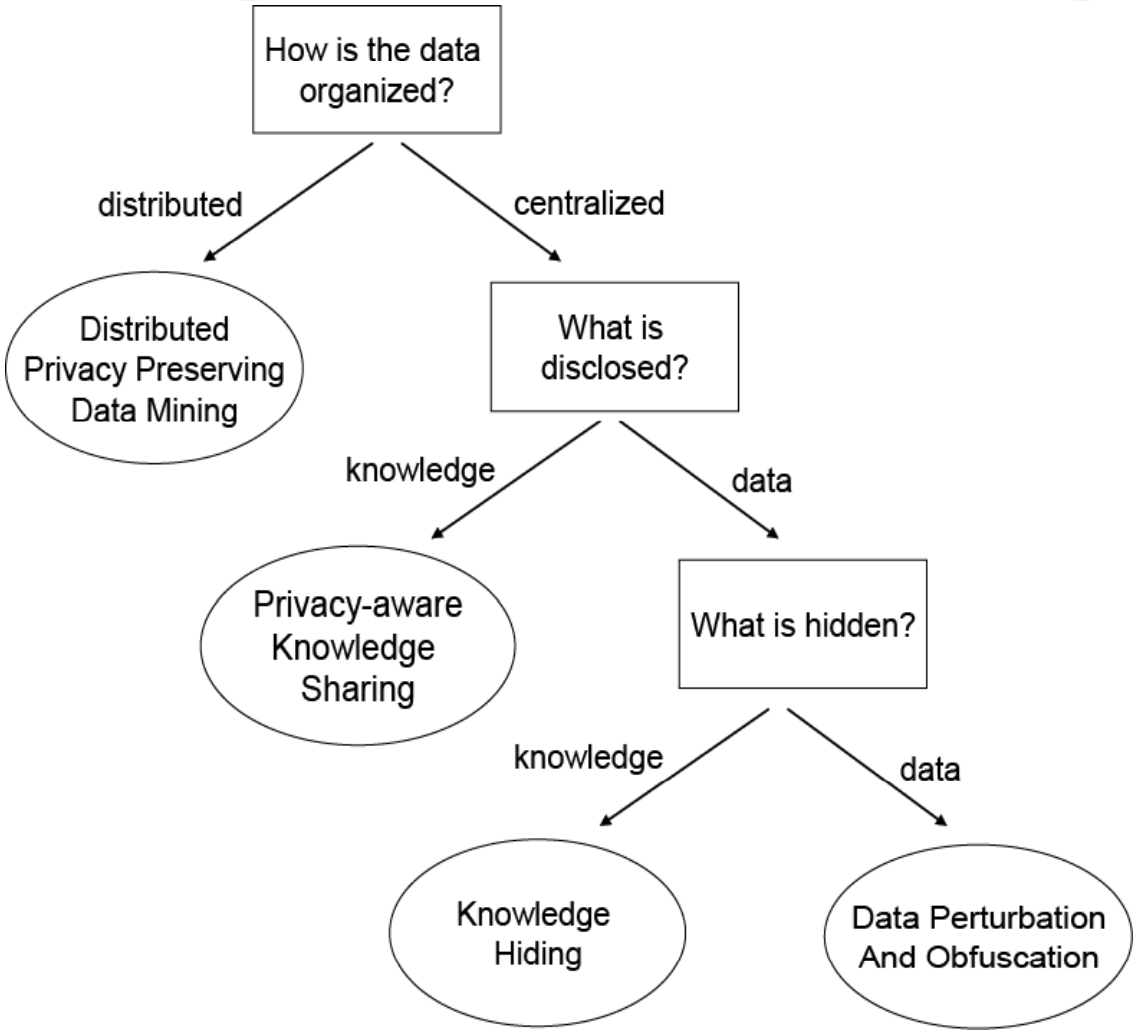
1. Data Publishing
   a) K-anonimity
   b) Data Perturbation and Obfuscation
2. Knowledge Publishing
3. Knowledge Hiding
4. Distributed Privacy Preserving Data Mining
5. Privacy Preserving Outsoursing

- Special focus on:
  - Privacy in Spatio-Temporal and Mobility data

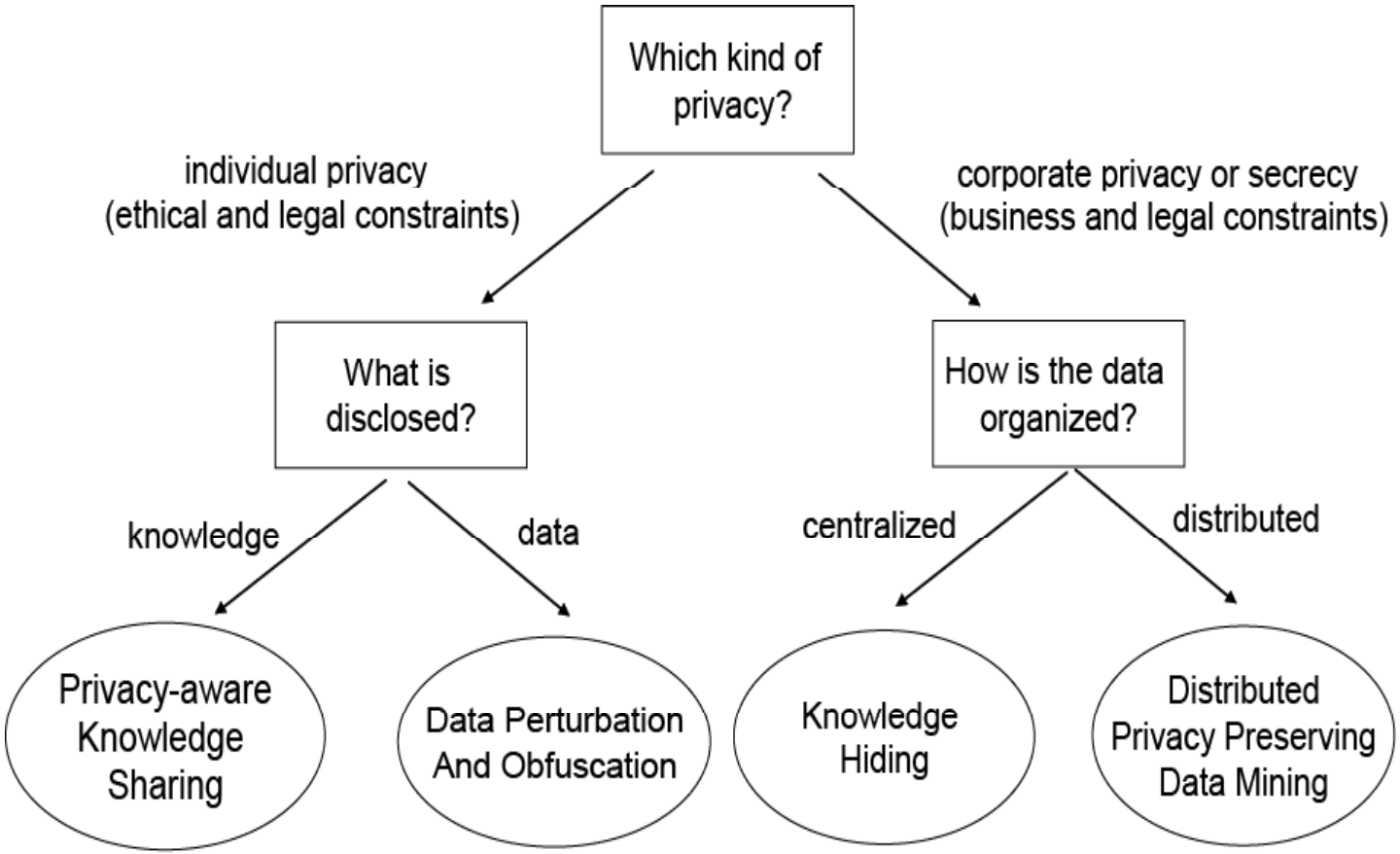
A taxonomy for PPDM



A taxonomy tree…



And another one…



Attack model and protection model

- In each problem setting, we must provide:
  - An **attack model**
    - What does the attacker know? Background knowledge
    - What does the attacker want to further know?
  - A **protection model**
    - Countermeasures: What is hidden? What is disclosed?
    - Privacy analysis: What is the probability that the attack succeeds?
    - Utility analysis: What is the analytical value of disclosed data/patterns?

Privacy-preserving data publishing: K-Anonymity

Data K-anonymity

- What is disclosed?
  - the data (modified somehow)
- What is hidden?
  - the real data
- How?
  - by transforming the data in such a way that it is not possible the re-identification of original database rows under a fixed anonymity threshold (individual privacy).

Motivation: Private Information in Publicly Available Data

Date of Birth	Zip Code	Allergy	History of Illness
03-24-79	07030	Penicillin	Pharyngitis
08-02-57	07028	No Allergy	Stroke
11-12-39	07030	No Allergy	Polio
08-02-57	07029	Sulfur	Diphtheria
08-01-40	07030	No Allergy	Colitis

Medical Research Database

Sensitive Information

Security Threat: May Link Private Information to Person

Quasi-identifiers

Date of Birth	Zip Code	Allergy	History of Illness
03-24-79	07030	Penicillin	Pharyngitis
08-02-57	07028	No Allergy	Stroke
11-12-39	07030	No Allergy	Polio
08-02-57	07029	Sulfur	Diphtheria
08-01-40	07030	No Allergy	Colitis

Victor is the only person born 08-02-57 in the area of 07028… Ha, he has a history of stroke!

k-Anonymity [SS98]: Eliminate Link to Person through Quasi-identifiers

Date of Birth	Zip Code	Allergy	History of Illness
*	07030	Penicillin	Pharyngitis
08-02-57	0702*	No Allergy	Stroke
*	07030	No Allergy	Polio
08-02-57	0702*	Sulfur	Diphtheria
*	07030	No Allergy	Colitis

k(=2 in this example)-anonymous table

Property of *k*-anonymous table

- Each value of quasi-identifier attributes appears ≥ k times in the table (or it does not appear at all)

⇒ Each row of the table is hidden in ≥ *k* rows

⇒ Each person involved is hidden in ≥ *k* peers

k-Anonymity Protects Privacy

Date of Birth	Zip Code	Allergy	History of Illness
08-02-57	0702*	No Allergy	Stroke
08-02-57	0702*	No Allergy	Stroke
*	07030	No Allergy	Polio
08-02-57	0702*	Sulfur	Diphtheria
	07030	No Allergy	Colitis

Which of them is Victor's record?
Confusing…

k-anonymity – Problem Definition

- Input: Database consisting of n rows, each with m attributes drawn from a finite alphabet.
- Assumption: the data owner knows/indicates which of the m attributes are *Quasi-Identifiers.*
- Goal: trasform the database in such a way that is K-anonymous w.r.t. a given *k,* and the QIs.
- How: By means of generalization and suppression.
- Objective: Minimize the distortion.
- Complexity: NP-Hard.
- A lot of papers on k-anonymity in 2004-2006 (SIGMOD, VLDB, ICDE, ICDM)

Privacy-preserving data publishing: Data Randomization, Perturbation and Obfuscation

Data Perturbation and Obfuscation

- What is disclosed?
  - the data (modified somehow)
- What is hidden?
  - the real data
- How?
  - by perturbating the data in such a way that it is not possible the identification of original database rows (individual privacy), but it is still possible to extract **valid** knowledge (models and patterns).
  - A.K.A. ***"distribution reconstruction"***

Data Perturbation and Obfuscation

- R. Agrawal and R. Srikant. Privacy-preserving data mining. In Proceedings of SIGMOD 2000.
- D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In Proceedings of PODS, 2001.
- W. Du and Z. Zhan. Using randomized response techniques for privacy-preserving data mining. In Proceedings of SIGKDD 2003.
- A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In Proceedings of PODS 2003.
- A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In Proceedings of SIGKDD 2002.
- K. Liu, H. Kargupta, and J. Ryan. Random Projection-based Multiplicative Perturbation for Privacy Preserving Distributed Data Mining. IEEE Transactions on Knowledge and Data Engineering (TKDE), VOL. 18, NO. 1.
- K. Liu, C. Giannella and H. Kargupta. An Attacker's View of Distance Preserving Maps for Privacy Preserving Data Mining. In Proceedings of PKDD'06

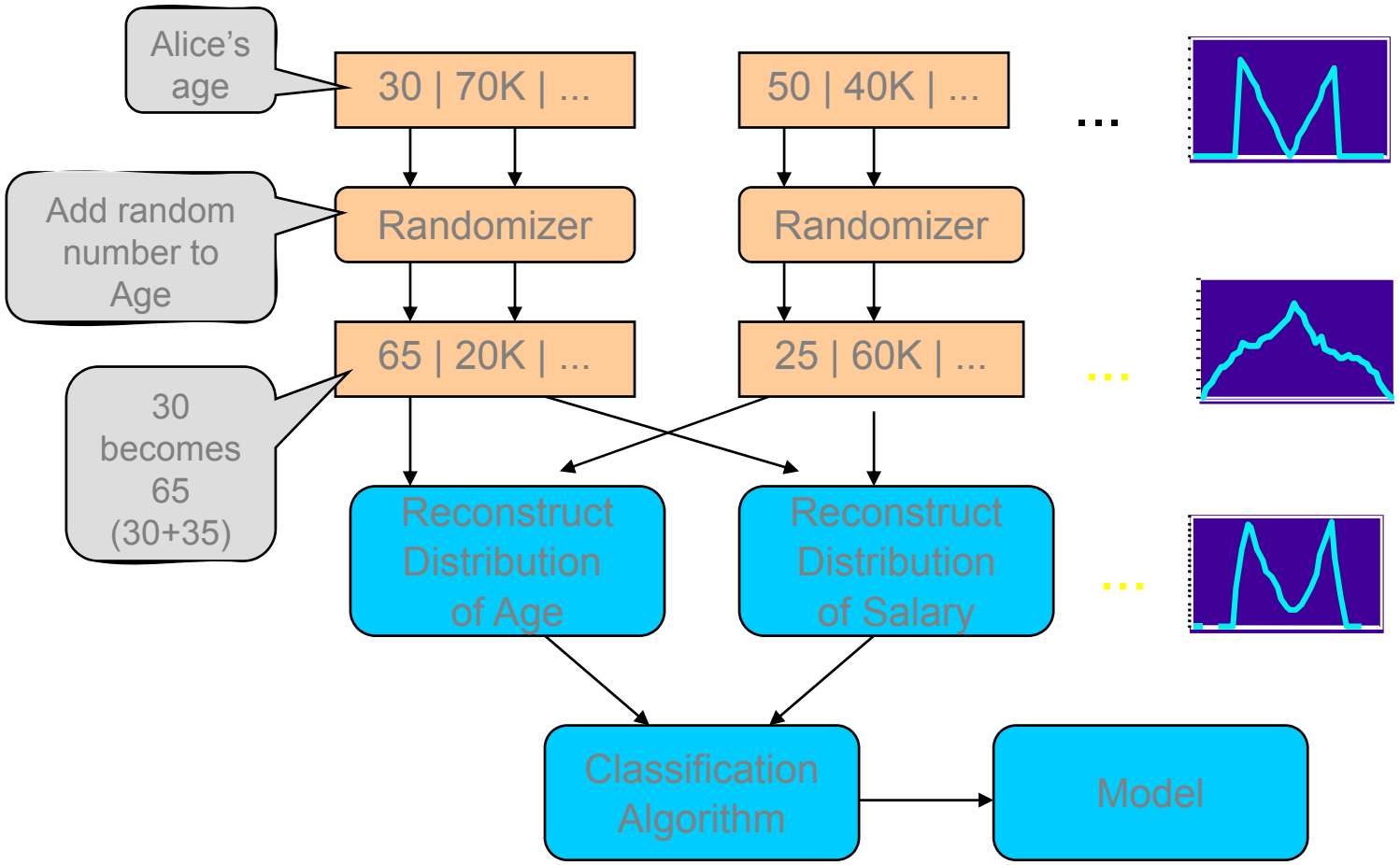
Data Perturbation and Obfuscation

- This approach can be instantiated to association rules as follows:
  - *D* source database;
  - *R* a set of association rules that can be mined from *D;*
  - Problem: define two algorithms *P* and M_P such that
    - *P(D) = D'* where *D'* is a database that do not disclose any information on singular rows of *D*;
    - $M_P(D') = R$

Decision Trees
Agrawal and Srikant '00

- Assume users are willing to
  - Give true values of certain fields
  - Give modified values of certain fields
- Practicality
  - 17% refuse to provide data at all
  - 56% are willing, as long as privacy is maintained
  - 27% are willing, with mild concern about privacy
- Perturb Data with Value Distortion
  - User provides $x_i + r$ instead of x_i
  - r is a random value
    - Uniform, uniform distribution between $[-\alpha, \alpha]$
    - Gaussian, normal distribution with $\mu = 0$, σ

Randomization Approach Overview

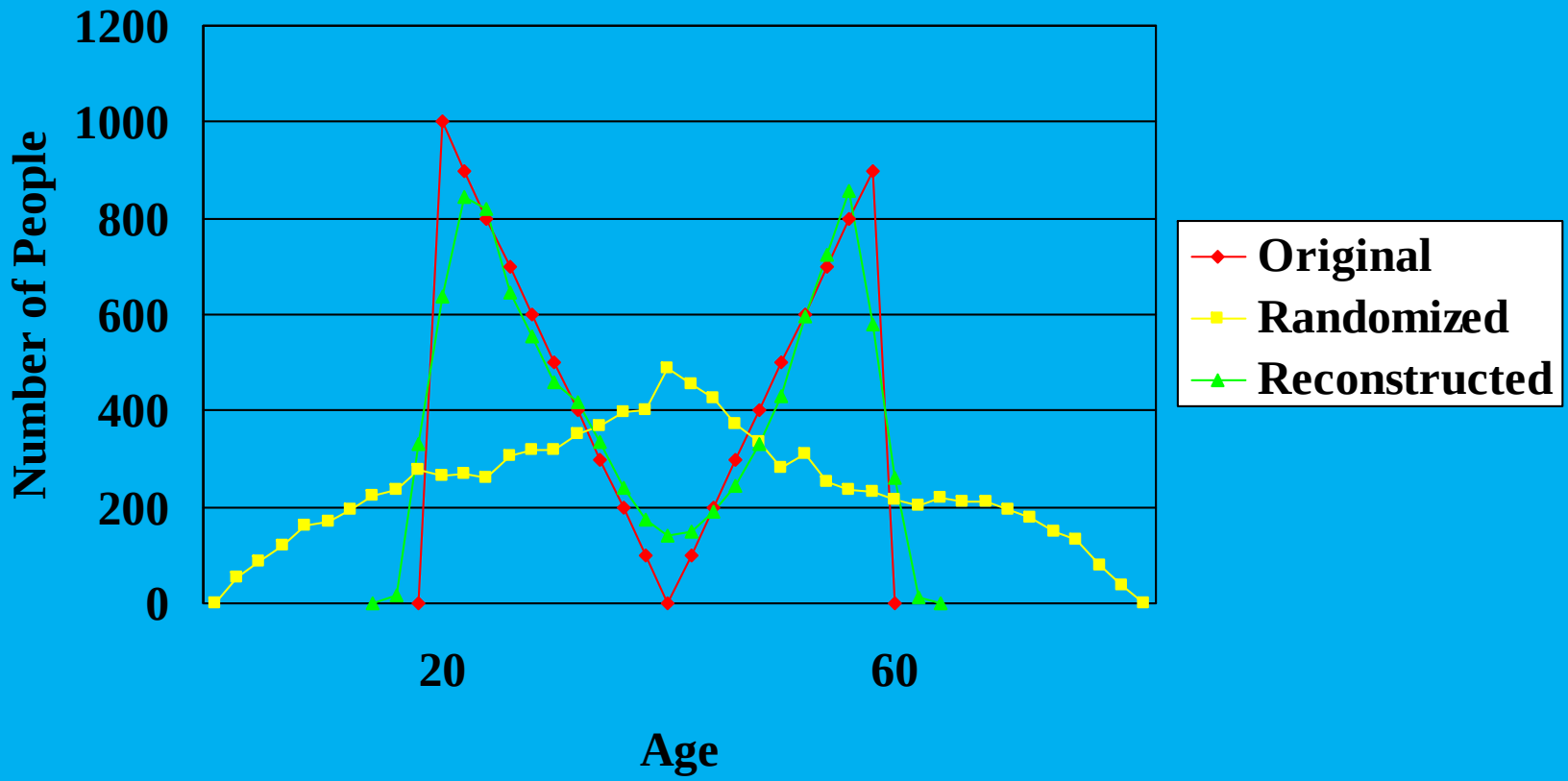


Reconstruction Problem

- Original values $x_1, x_2, \ldots, x_n$
  - from probability distribution X (unknown)
- To hide these values, we use $y_1, y_2, \ldots, y_n$
  - from probability distribution Y
- Given
  - $x_1+y_1, x_2+y_2, \ldots, x_n+y_n$
  - the probability distribution of Y

  Estimate the probability distribution of X.

Distribution reconstruction



Recap: Why is privacy preserved?

- Cannot reconstruct individual values accurately.
- Can only reconstruct distributions.

Privacy-aware Knowledge Publishing

Privacy-aware Knowledge Sharing

- What is disclosed?
  - the intentional knowledge (i.e. rules/patterns/models)
- What is hidden?
  - the source data

- The central question:

  "do the data mining results themselves violate privacy?"

- Focus on **individual privacy**: the individuals whose data are stored in the source database being mined.

Privacy-aware Knowledge Sharing

- M. Kantarcioglu, J. Jin, and C. Clifton. When do data mining results violate privacy? In Proceedings of the tenth ACM SIGKDD, 2004.
- S. R. M. Oliveira, O. R. Zaiane, and Y. Saygin. Secure association rule sharing. In Proc.of the 8th PAKDD, 2004.
- P. Fule and J. F. Roddick. Detecting privacy and ethical sensitivity in data mining results. In Proc. of the 27° conference on Australasian computer science, 2004.
- Atzori, Bonchi, Giannotti, Pedreschi. K-anonymous patterns. In PKDD and ICDM 2005, The VLDB Journal (accepted for publication).
- A. Friedman, A. Schuster and R. Wolff. *k*-Anonymous Decision Tree Induction. In Proc. of PKDD 2006.

Privacy-aware Knowledge Sharing

- Association Rules can be dangerous…

Example

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

$$sup(\{a_1, a_2, a_3\}) = \frac{sup(\{a_1, a_2, a_3, a_4\})}{conf} \approx \frac{80}{0.987} = 81.05$$

In other words, we know that there is just one individual for which the pattern $a_1 \wedge a_2 \wedge a_3 \wedge \neg a_4$ holds.

- How to solve this kind of problems?

Privacy-aware Knowledge Sharing

- Association Rules can be dangerous…

Age = 27, Postcode = 45254, Christian ⇒ American
(support = 758, confidence = 99.8%)

Age = 27, Postcode = 45254 ⇒ American
(support = 1053, confidence = 99.9%)

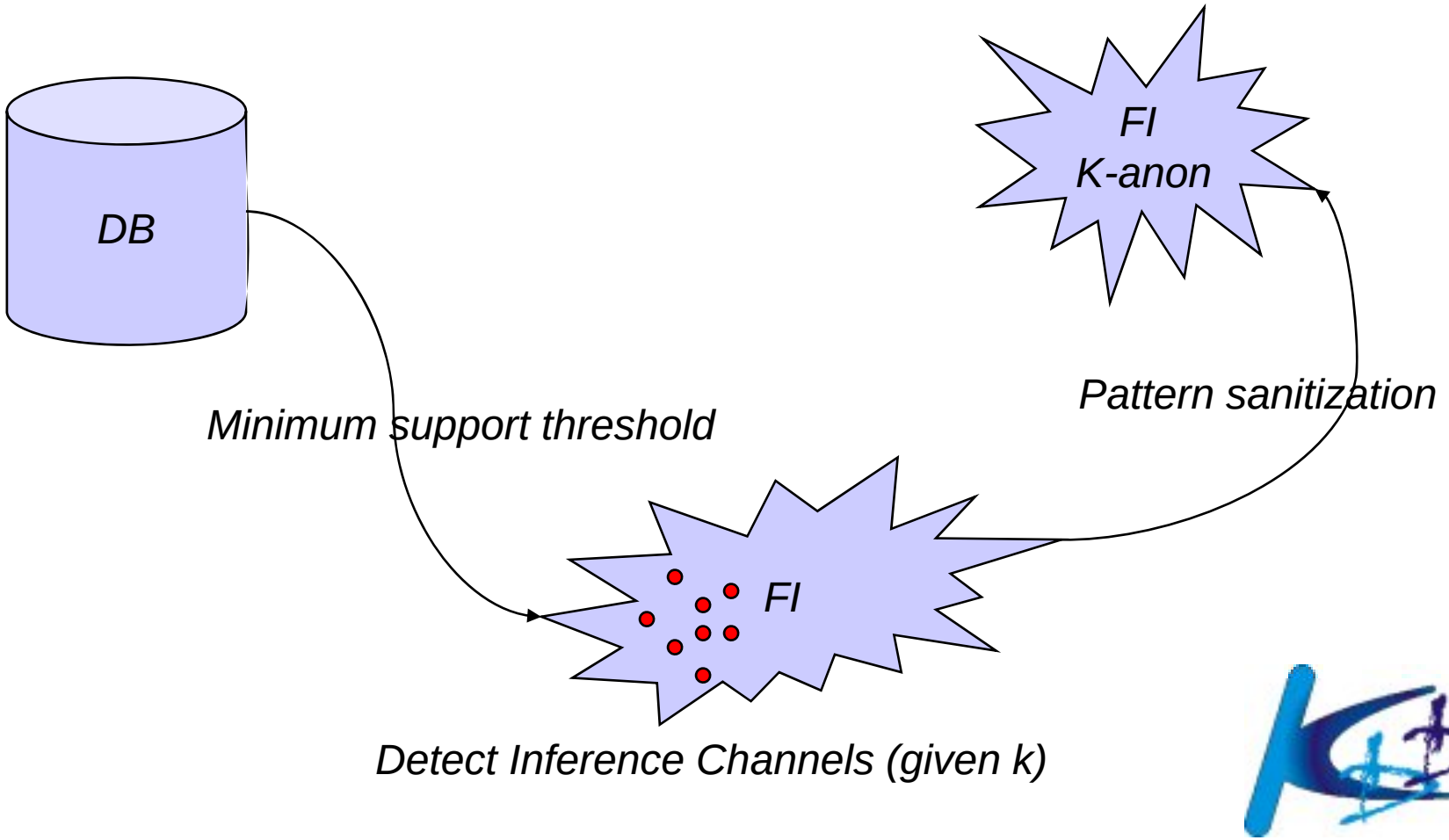
Since *sup(rule) / conf(rule) = sup(head)* we can derive:

Age = 27, Postcode = 45254, not American ⇒ Christian
(support = 1, confidence = 100.0%)

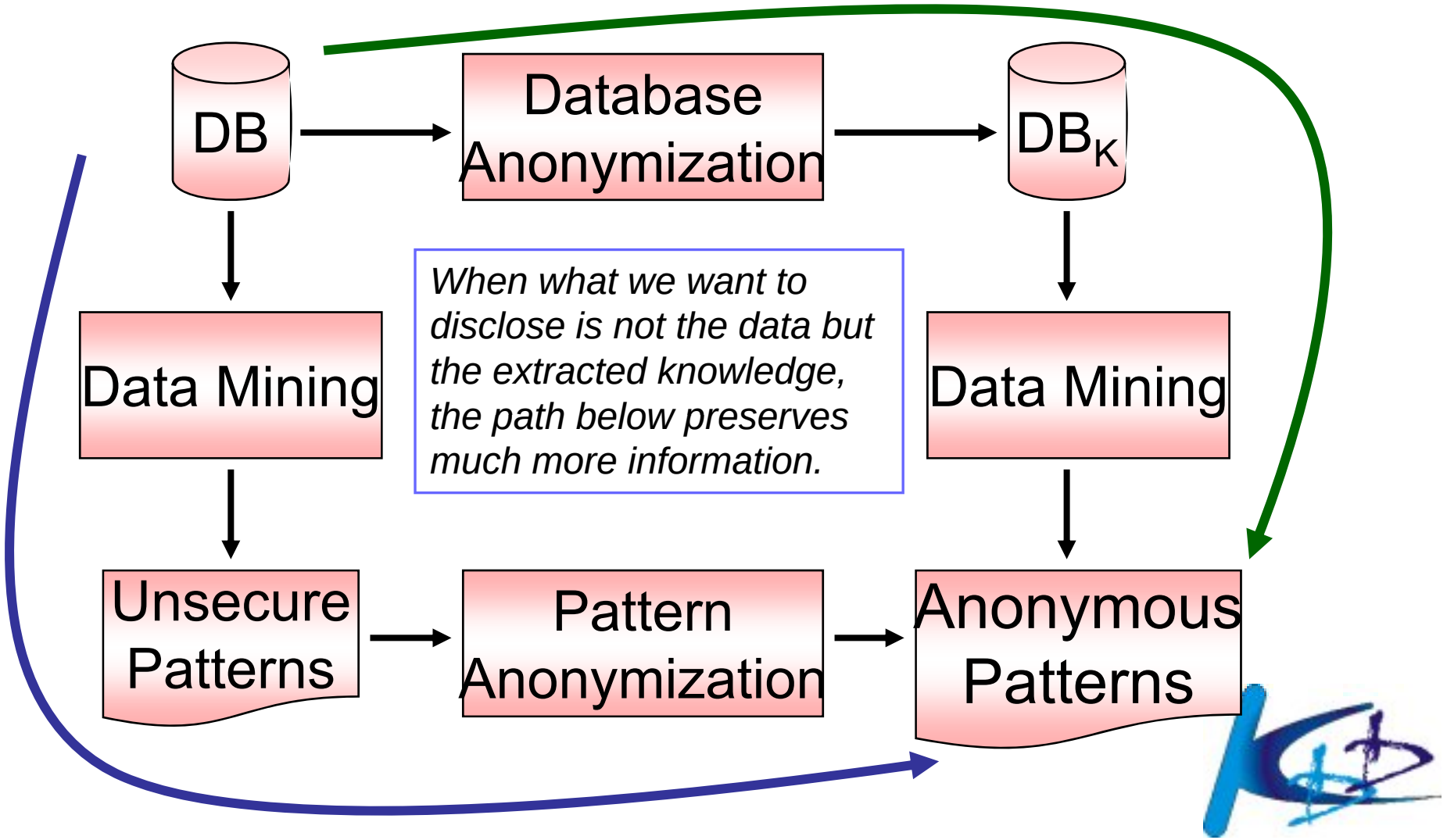
This information refers to my France neighbor…. he is Christian!
(and this information was clearly not intended to be released as it links public information regarding few people to sensitive data!)

- How to solve this kind of problems?

The scenario



Privacy-aware Knowledge Sharing



Distributed Privacy Preserving Data Mining

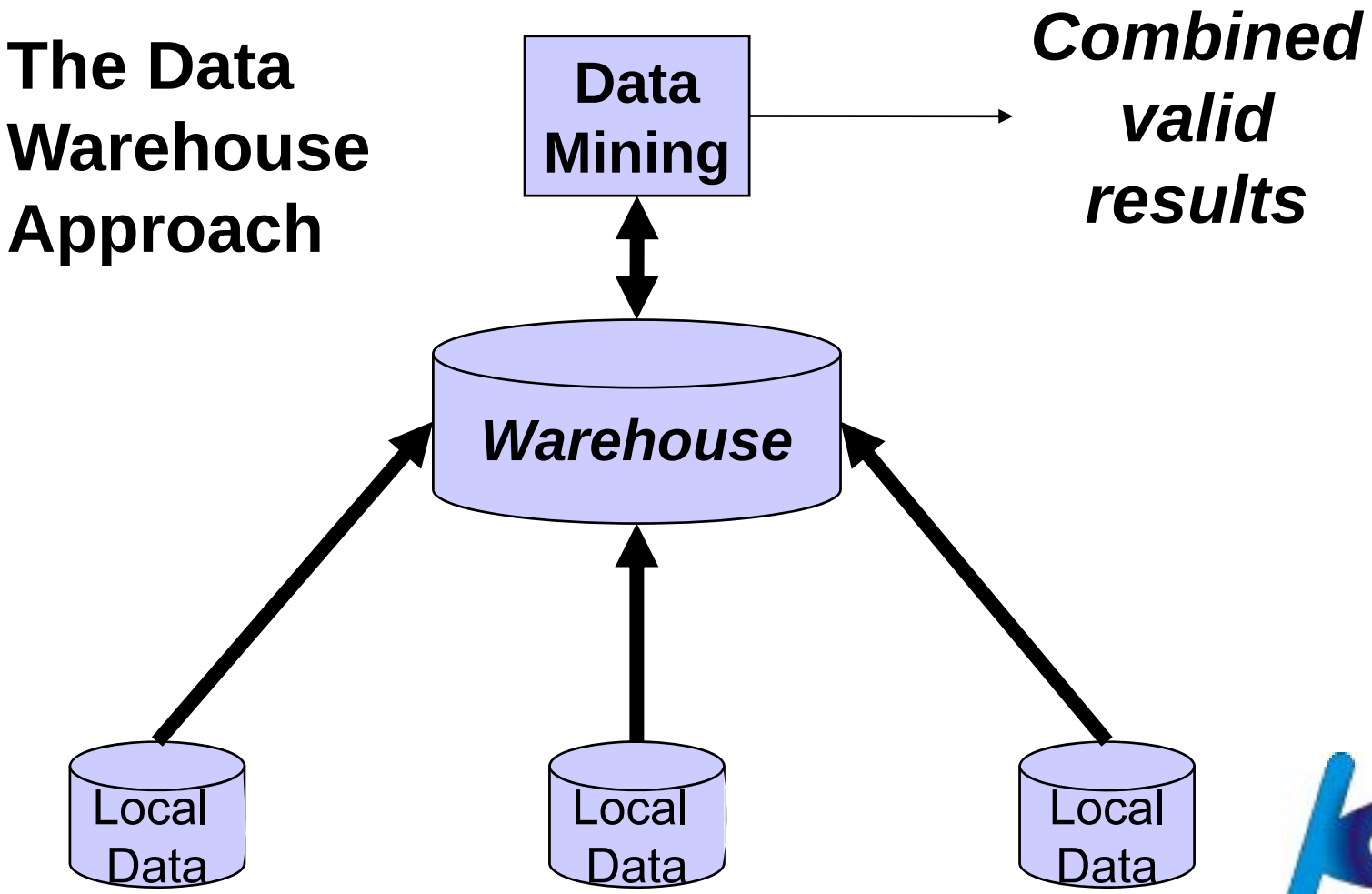
Distributed Privacy Preserving Data Mining

- Objective?
  - computing a valid mining model from several distributed datasets, where each party owing a dataset does not communicate its data to the other parties involved in the computation.
- How?
  - cryptographic techniques
- A.K.A. "*Secure Multiparty Computation*"

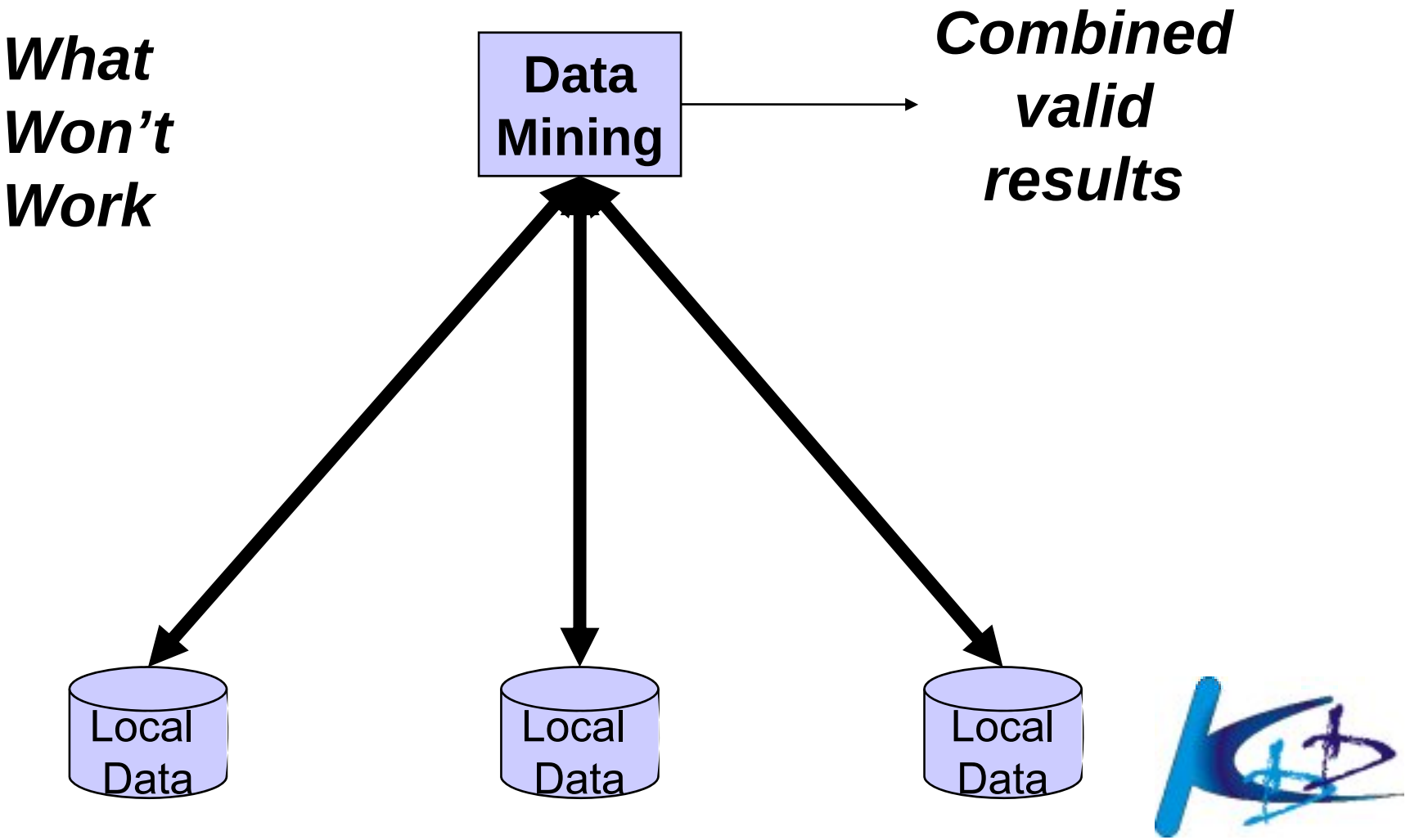
Distributed Privacy Preserving Data Mining

- C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y.Zhu. Tools for privacy preserving distributed data mining. SIGKDD Explor. Newsl., 4(2), 2002.
- M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. In SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'02), 2002.
- B. Pinkas. Cryptographic techniques for privacy-preserving data mining. SIGKDD Explor. Newsl., 4(2), 2002.
- J. Vaidya and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. In Proceedings of ACM SIGKDD 2002.

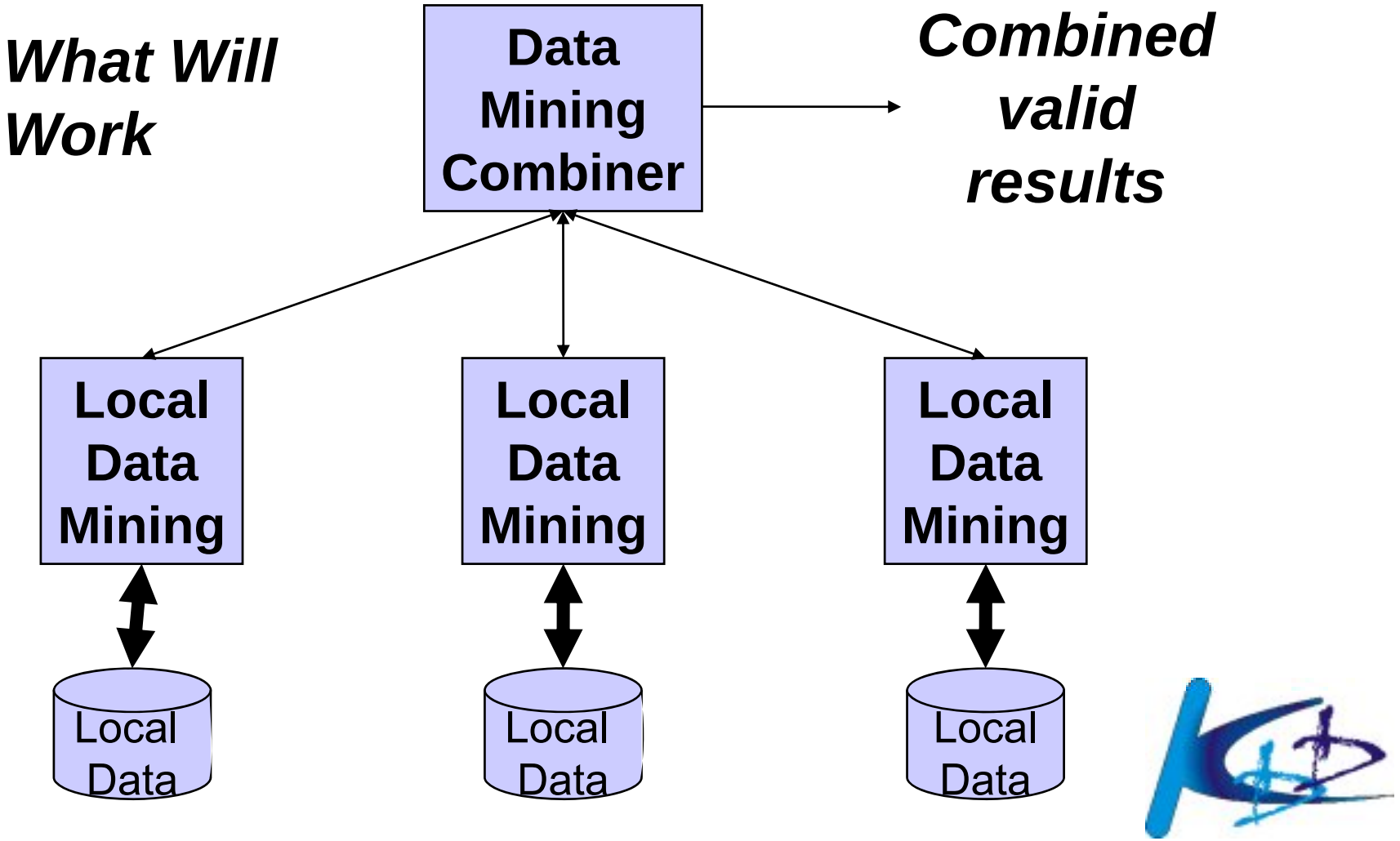
Distributed Data Mining: The “Standard” Method



Private Distributed Mining: What is it?



Private Distributed Mining: What is it?



Distributed Privacy Preserving Data Mining

- This approach can be instantiated to association rules in two different ways corresponding to two different data partitions: vertically and horizontally partitioned data.

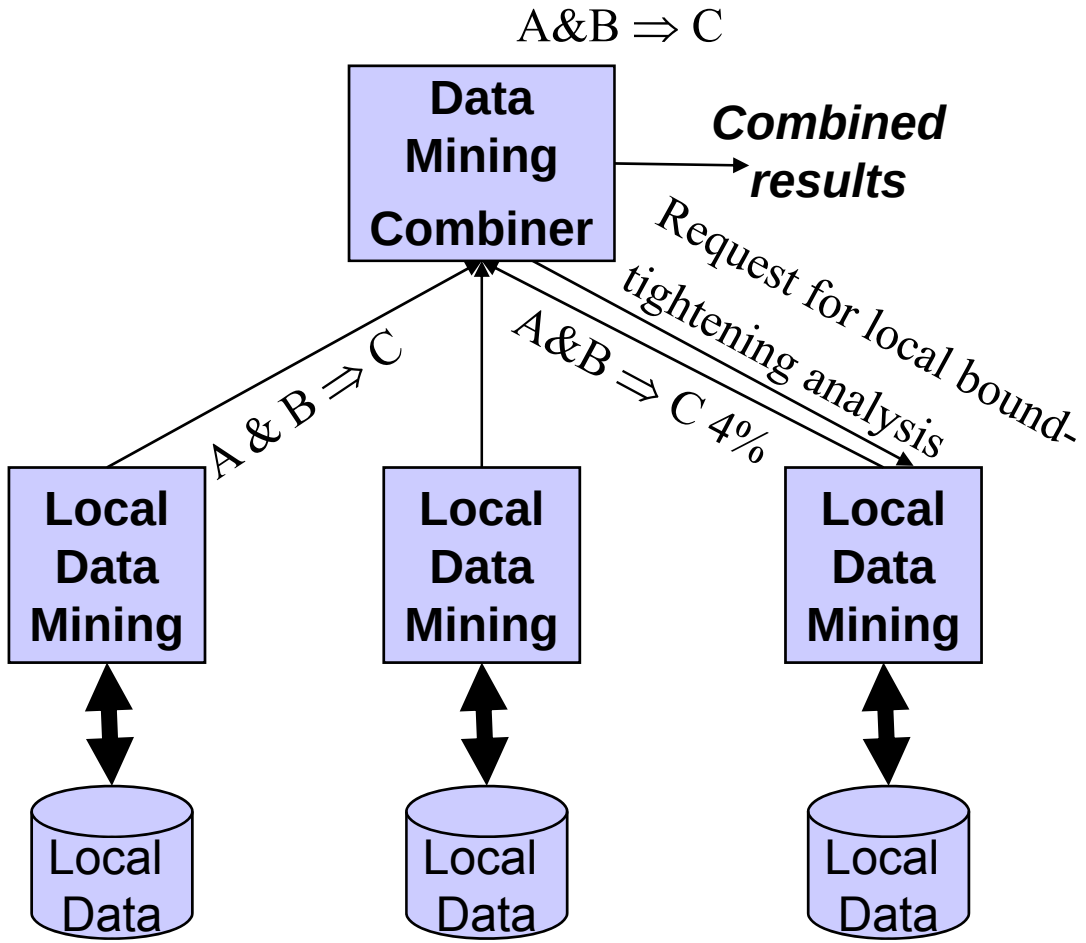
1. Each site *s* holds a portion *Is* of the whole vocabulary of items *I*, and thus each itemset is split between different sites. In such situation, the key element for computing the support of an itemset is the“secure” scalar product of vectors representing the subitemsets in the parties.

2. The transactions of *D* are partitioned in *n* databases *D1, . . . ,Dn*, each one owned by a different site involved in the computation. In such situation, the key elements for computing the support of itemsets are the “secure”union and “secure” sum operations.

Example: *Association Rules*

- Assume data is horizontally partitioned
  - Each site has complete information on a set of entities
  - Same attributes at each site
- If goal is to avoid disclosing entities, problem is easy
- Basic idea: Two-Phase Algorithm
  - First phase: Compute candidate rules
    - Frequent globally $\Rightarrow$ frequent at some site
  - Second phase: Compute frequency of candidates

Association Rules in Horizontally Partitioned Data



Knowledge Hiding

Knowledge Hiding

- What is disclosed?
  - the data (modified somehow)
- What is hidden?
  - some “sensitive” knowledge (i.e. secret rules/patterns)
- How?
  - usually by means of data **sanitization**
    - the data which we are going to disclose is modified in such a way that the sensitive knowledge can non longer be inferred,
    - while the original database is modified as less as possible.

Knowledge Hiding

- E. Dasseni, V. S. Verykios, A. K. Elmagarmid, and E. Bertino. *Hiding association rules by using confidence and support*. In Proceedings of the 4th International Workshop on Information Hiding, 2001.

- Y. Saygin, V. S. Verykios, and C. Clifton. *Using unknowns to prevent discovery of association rules*. SIGMOD Rec., 30(4), 2001.

- S. R. M. Oliveira and O. R. Zaiane. *Protecting sensitive knowledge by data sanitization*. In Third IEEE International Conference on Data Mining (ICDM'03), 2003.

Knowledge Hiding

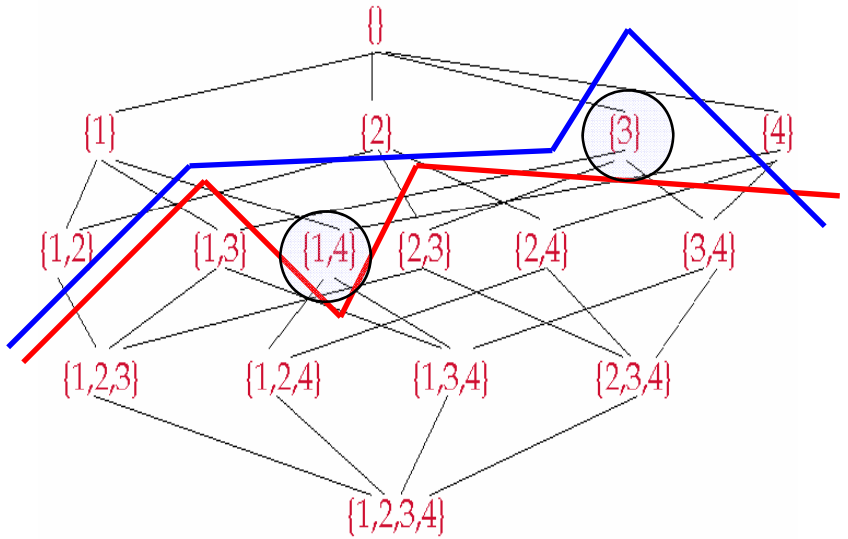
- This approach can be instantiated to association rules as follows:
  - *D* source database;
  - *R* a set of association rules that can be mined from *D;*
  - R_h a subset of *R* which must be hidden.
  - Problem: how to transform *D* into *D'* (the database we are going to disclose) in such a way that *R/* R_h can be mined from *D'.*

Knowledge Hiding

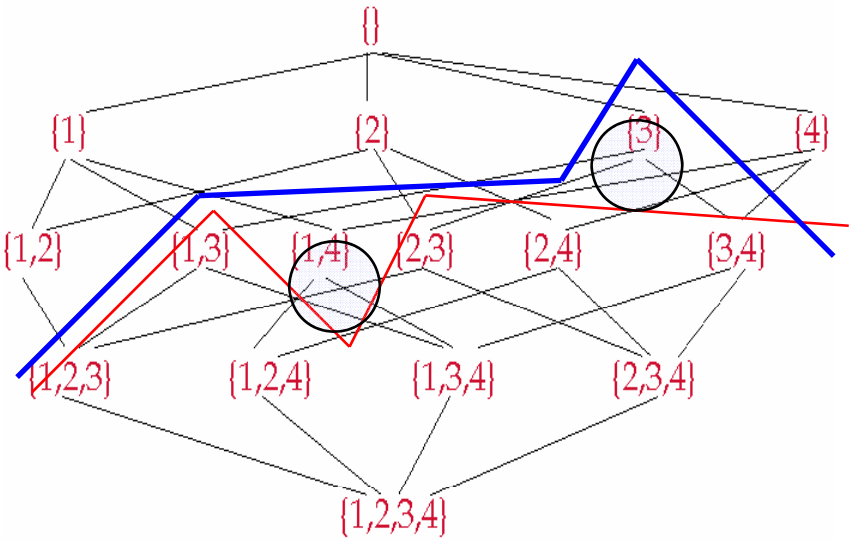
- *Mining frequent itemsets is the fundamental step for mining Association Rules*
- *Suppose min_sup = 2*

D	{1}	{2}	{3}	{4}
T1	1	1	0	0
T2	0	1	0	1
T3	1	0	1	1
T4	1	0	0	1
T5	1	1	0	0
T6	0	1	1	0
T7	0	0	1	0

itemset	support
{1}	4
{2}	4
{3}	3
{4}	3
{1,2}	2
{1,4}	2



D	{1}	{2}	{3}	{4}
T1	1	1	0	0
T2	0	1	0	1
T3	?	0	?	?
T4	?	0	0	?
T5	1	1	0	0
T6	0	1	?	0
T7	0	0	?	0



- [Intermediate table]: itemsets {3} and {1,4} have the '1's turned into '?'.
- Some of these '?' will later on be turned into zeros.
- Heuristics:
  - select which of the transactions {T3, T4, T6, T7} will be *sanitized*,
  - to which *extent* (meaning how many items will be affected),
  - and in which relative *order*.

- Heuristics do not guarantee (in any way) the identification of the best possible solution: but they provide overall good solutions efficiently.
- A solution always exists! The easiest way to see that is by turning all '1's to '0's in all the 'sensitive' items of the transactions supporting the sensitive itemsets.

Privacy Preserving Outsourcing of Data Mining

Secure Outsourcing of Data Mining

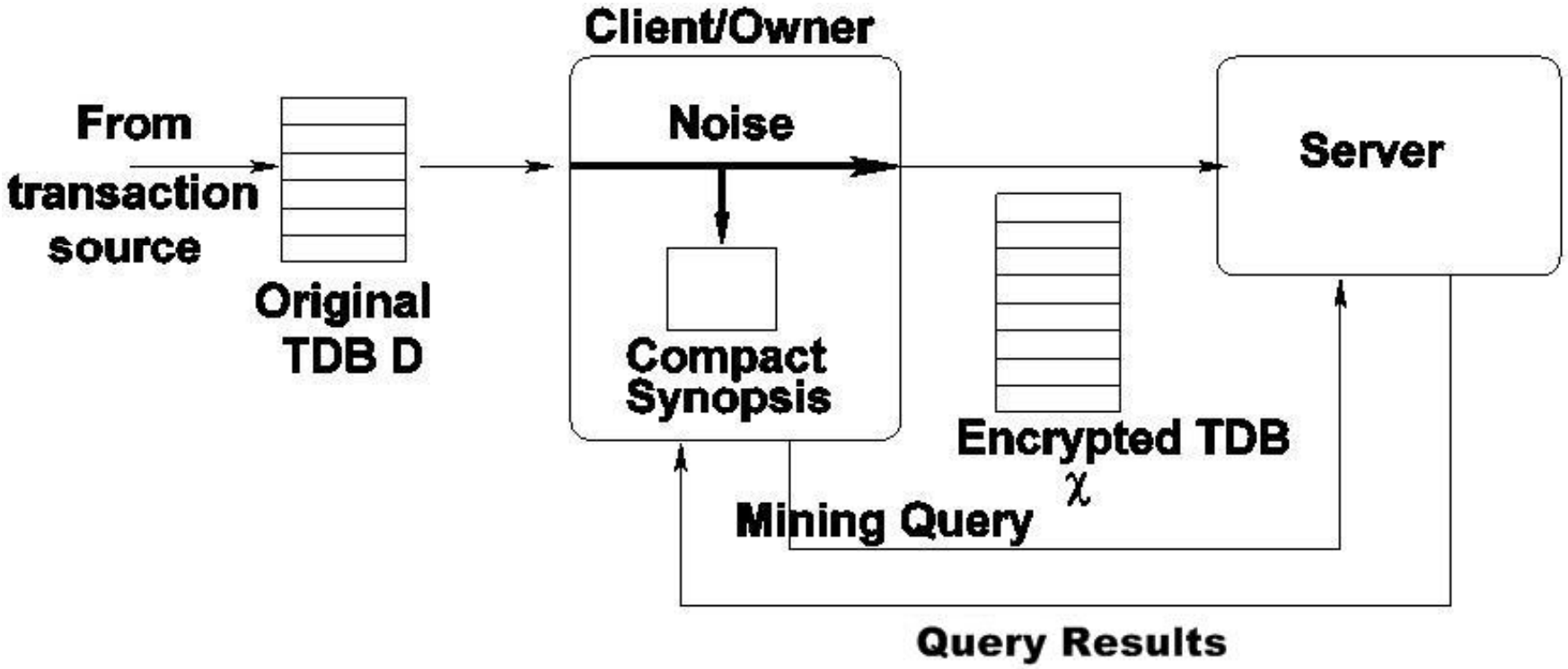
- Organizations could do not posses
  - **in-house expertise** for doing data mining
  - **computing infrastructure** adequate
- **Solution**: Outsourcing of data mining to a service provider
  - specific human resources
  - technological resources
- The server has access to data of the owner
- Data owner has the property of both
  - **Data** can contain personal information about individuals
  - **Knowledge** extracted from data can provide competitive advantages

The Problem

PROBLEM: Given a plain database D, construct an encrypted database D* such that:

- all encrypted transactions in D* and items contained in it are secure
- given any mining query the server can compute the encrypted result
- encrypted mining and analysis results are secure
- the owner can decrypt the results and so, reconstruct the exact result
- the space and time incurred by the owner in the process has to be minimum

Framework Architecture



Conclusions

PPDM research strives for a win-win situation

- Obtaining the advantages of collective mobility knowledge without disclosing inadvertently any individual mobility knowledge.

- This result, if achieved, may have an impact on
  - laws and jurisprudence,
  - the social acceptance of ubiquitous technologies.

- This research must be tackled in a multi-disciplinary way: the opportunities and risks must be shared by social analysts, jurists, policy makers, concerned citizens.

European Union Data Protection Directives

- Directive 95/46/EC
  - Passed European Parliament 24 October 1995
  - Goal is to ensure free flow of information
    - *Must preserve privacy needs of member states*
  - Effective October 1998
- Effect
  - Provides guidelines for member state legislation
    - Not directly enforceable
  - Forbids sharing data with states that don't protect privacy
    - Non-member state must provide adequate protection,
    - Sharing must be for "allowed use", or
    - Contracts ensure adequate protection

EU Privacy Directive

- Personal data is any information that can be traced directly or *indirectly* to a specific person
- Use allowed if:
  - Unambiguous consent given
  - Required to perform contract with subject
  - Legally required
  - Necessary to protect vital interests of subject
  - In the public interest, or
  - Necessary for legitimate interests of processor and doesn't violate privacy
- Some uses specifically proscribed (sensitive data)
  - Can't reveal racial/ethnic origin, political/religious beliefs, trade union membership, health/sex life

Anonymity according to 1995/46/EC

- The principles of protection must apply to any information concerning an identified or identifiable person;
- To determine whether a person is identifiable, account should be taken of *all the means likely reasonably to be used* either by the controller or by any other person to identify the said person;
- The principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable;

US Healthcare Information Portability and Accountability Act (HIPAA)

- Governs use of patient information
  - Goal is to protect the patient
  - Basic idea: Disclosure okay if anonymity preserved
- Regulations focus on outcome
  - A covered entity may not use or disclose protected health information, except as permitted or required…
    - To individual
    - For treatment (generally requires consent)
    - To public health / legal authorities
  - Use permitted where "there is no reasonable basis to believe that the information can be used to *identify an individual*"

The Safe Harbor “atlantic bridge”

- In order to bridge EU and US (different) privacy approaches and provide a streamlined means for U.S. organizations to comply with the European Directive, the U.S. Department of Commerce in consultation with the European Commission developed a "Safe Harbor" framework.
- Certifying to the Safe Harbor will assure that EU organizations know that US companies provides “adequate” privacy protection, as defined by the Directive.

The Safe Harbor “atlantic bridge”

- Data presumed not identifiable if 19 identifiers removed (§ 164.514(b)(2)), e.g.:
  - Name,
  - location smaller than 3 digit postal code,
  - dates finer than year,
  - identifying numbers
- Shown not to be sufficient (Sweeney)

Pointers to Resources

Web Links on Privacy Laws

English

- europa.eu.int/comm/justice_home/fsj/privacy/law/index_en.htm
- www.privacyinternational.org/
- www.export.gov/safeharbor/

Italian

- www.garanteprivacy.it
- www.interlex.it/
- www.iusreporter.it/
- www.privacy.it/

Web Resources on PPDM

- **Privacy Preserving Data Mining Bibliography (maintained by Kun Liu)**
  http://www.cs.umbc.edu/~kunliu1/research/privacy_review.html

- **Privacy Preserving Data Mining Blog**
  http://www.umbc.edu/ddm/wiki/index.php/PPDM_Blog

- **Privacy Preserving Data Mining Bibliography (maintained by Helger Lipmaa)**
  http://www.cs.ut.ee/~lipmaa/crypto/link/data_mining/

- **The Privacy Preserving Data Mining Site (maintained by Stanley Oliveira)**
  http://www.cs.ualberta.ca/%7Eoliveira/psdm/psdm_index.html [no longer updated]

- **IEEE International Workshop on Privacy Aspects of Data Mining (every year in conjunction with IEEE ICDM conference)**

  PADM'06 webpage: http://www-kdd.isti.cnr.it/padm06/