*Pisa KDD Laboratory*

*http://www-kdd.isti.cnr.it/*

# Tutorial on Privacy Preserving Data Mining

## Handling Private Information in Data Mining


***Speaker:***

***Maurizio Atzori***

***PhD Student***

***KDD-Lab, ISTI-CNR***

***Maurizio.Atzori@isti.cnr.it***

# Summary

- Introduction on Privacy
  - EU and US laws
- Database Security
- Current Technology in PPDM
  - Randomization
  - Knowledge Hiding
  - Distributed DM
- Conclusions

# Introduction on Privacy

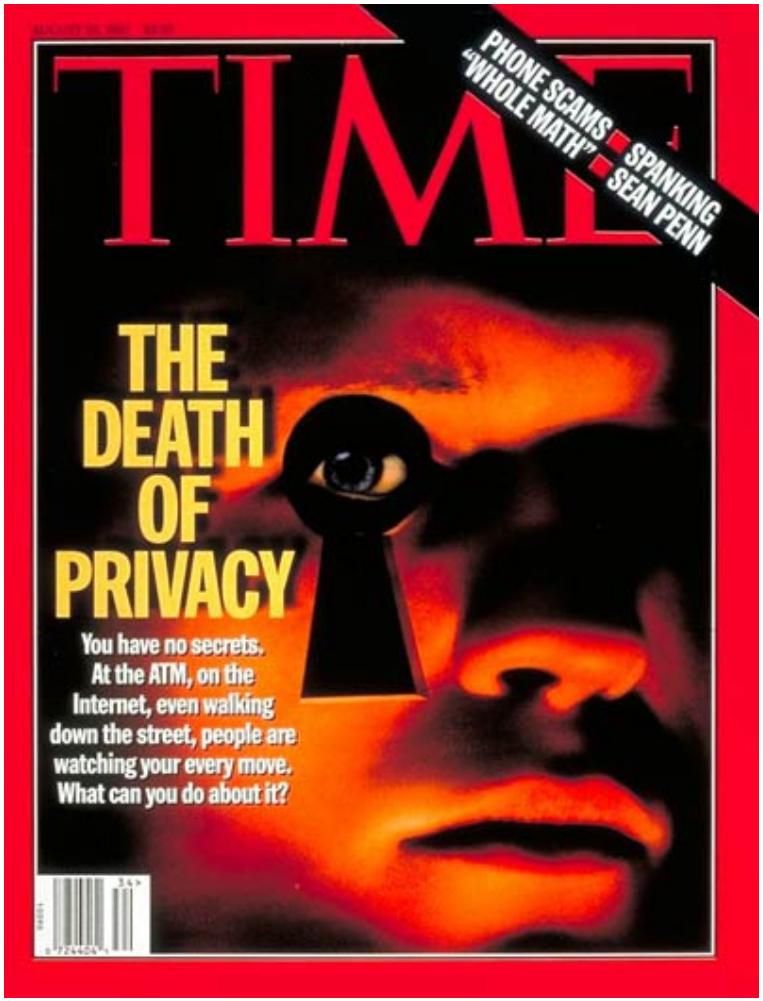
# Definition of privacy

What is privacy?

# Global Attention to Privacy

- Time (August 1997)
  - The Death of Privacy
- The Economist (May 1999)
  - The End of Privacy
- The European Union (October 1998)
  - Directive on Privacy Protection
- TIA-DARPA Project (2003)

# Time: The Death of Privacy



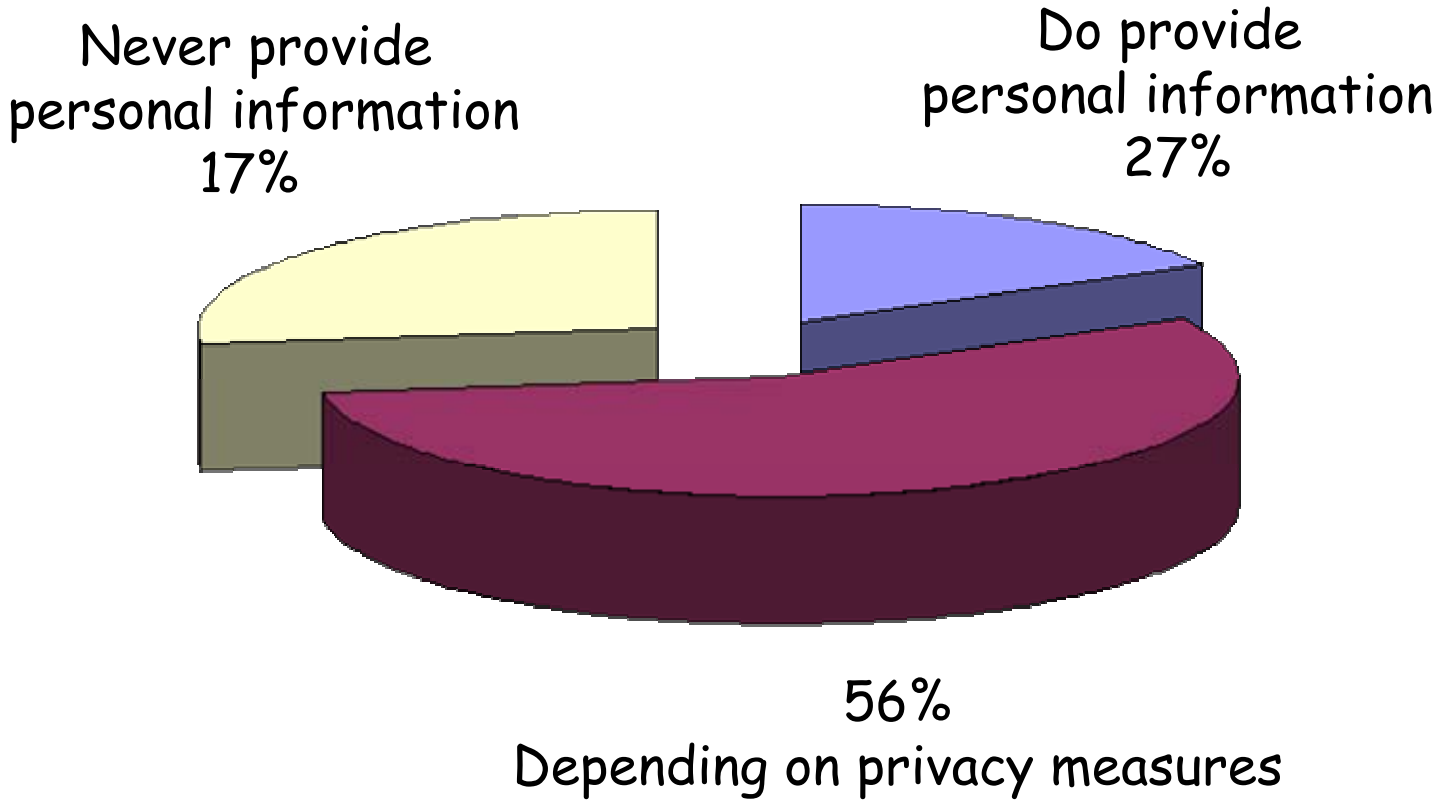

- Invasion of privacy
  - Our right to be left alone has disappeared, bit by bit, in little brotherly steps.
  - Still, we've got something in return, and it's not all bad

# The Economist

- **Remember, they are always watching you. Use cash when you can. Do not give your phone number, social-security number or address, unless you absolutely have to.**

Do not fill in questionnaires or respond to telemarketers. Demand that credit and data-marketing firms produce all information they have on you, correct errors and remove you from marketing lists.

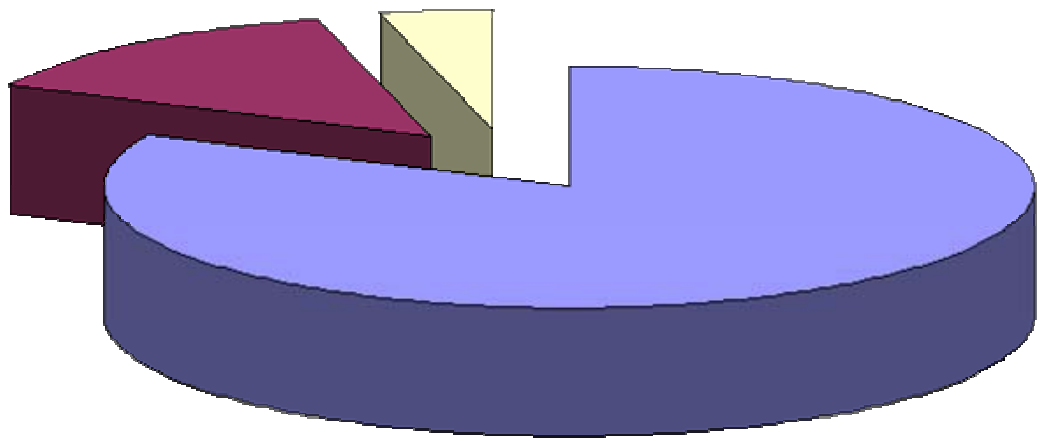
# Web Users: Attitudes




Source: *Special Issue on Internet Privacy.* Ed. L.F.Cranor (Feb 1999)

# Web Users: Privacy vs Benefits

86% of Web Users believe that participation in information-for-benefits programs is a matter of individual privacy choice




Source: *Freebies and privacy: What net users think.* A.F.Westin (July 1999)

# EU: Personal Data

- *Personal data* is defined as any information relating to an identity or identifiable natural person. An *identifiable person* is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.

http://www.legamedia.net/legapractice/reedsmith/2000/00-12/0012_olender_kurt_eu-privacy-directive.php

# EU: Processing of Personal Data

- The *processing of personal data* is defined as any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction.

# EU Privacy Directive

- The EU Privacy Directive provides:
  - That personal data must be processed fairly and lawfully
  - That personal data must be accurate
  - That data be collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes
  - That personal data is to be kept in the form which permits identification of the subject of the data for no longer than is necessary for the purposes for which the data was collected or for which it was further processed
  - That subject of the data must have given his unambiguous consent to the gathering and processing of the personal data
  - If consent was not obtained from the subject of the data, that personal data be processed for the performance of a contract to which the subject of the data is a party
  - That processing of personal data revealing racial or ethnical origin, political opinions, religious or philosophical beliefs, trade union membership, and the processing of data concerning health or sex life is prohibited

# 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
  - Can't reveal racial/ethnic origin, political/religious beliefs, trade union membership, health/sex life

# Safe Harbor (July 2000)

- The seven "safe harbor" principles are:
  - Notice
  - Choice
    - Opt-in in and opt-out
  - Onward Transfer
  - Security
  - Data Integrity
  - Access
  - Enforcement
- Note: voluntary compliance!
- Some patchwork of regulations (exceptions)

# Individually identifiable information

- Data that can't be traced to an individual not viewed as private
  - Remove identifiers (a list of 19)
- But can we ensure it can't be traced?
  - Candidate key in non-identifier information
  - Unique values for some individuals

*Data mining enables such tracing!*

# Individually identifiable information???

- Sweeney (2001) shows that "safe harbor" principles are not sufficient
  - From a set of 54805 people (voter list)
  - 69% unique on *postal code* and *birth date*
  - 87% US-wide with all 3 (*sex*)

- From Voter list to medical data!
- A solution is *k-anonymity*:
  - Any combination of values appears at least *k* times (distortion of results)

# Why Privacy Preserving Data Mining? An Example

- **Defense Advanced Research Projects Agency (DARPA)**
  - **Total Information Awareness (TIA)**
    - a set of technologies, including electronic searching tools to "mine" such records in the hopes of finding patterns indicating an imminent attack
    - TIA would violate individuals' privacy if it were used to inspect personal data, particularly financial transactions and phone records
- **The Solution:**
  - **Terrorism Information Awareness?!? …**
  - ***Privacy preserving data mining!!!***

http://www.govexec.com/dailyfed/0503/052003h2.htm

# Database Security

# Statistical Databases

- From works on Statistical Databases ('80)
  - Answer statistical queries while not disclosing actual values
- Query restriction
- Intrusion detection (sequential query analysis)
  - Query set overlap control
- Access control

- It is difficult to prove that some values are not released / can not be inferred

# Access Control

- Abstract reference architecture IETF
- Access control built into the database:
  - Hippocratic Databases (IBM)
- Access control outsourced
  - GUPster

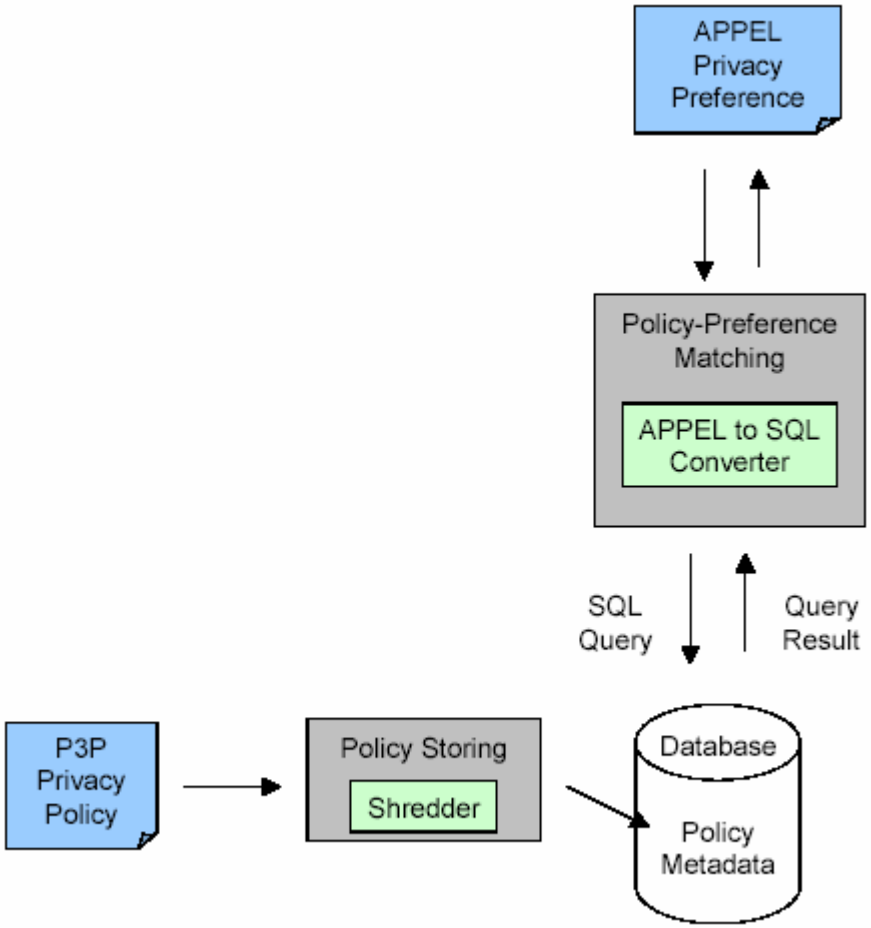
# Access Control Languages

- XACML (OASIS standard)
  - Used in GUPster prototype
- P3P/APPEL (W3C)
  - Used in Hippocratic DB prototype
  - P3P specifies Corporate data collection policy
  - APPEL specifies User Data collection policy
- GEOPRIV (IETF)

# P3P – Platform for Privacy Preferences

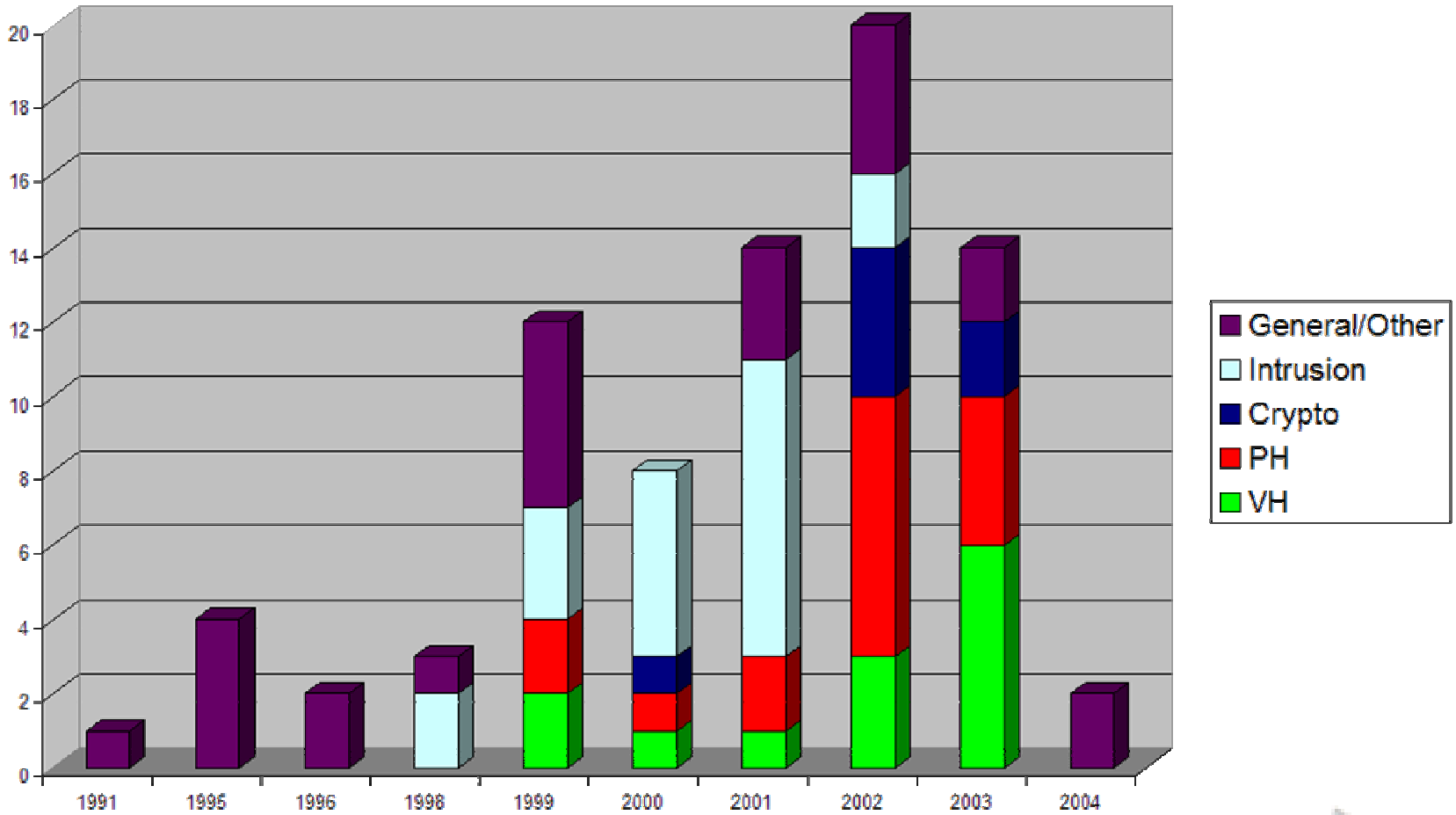
- PURPOSE: why data is collected
  - <current>: to complete current task
  - <contact>: to allow company to contact person
- RECIPIENT: who is to see the data
  - <ours>: ourselves
  - <same>: legal entities which follow our practices
  - <unrelated>: legal entities with unknown practices
- RETENTION: how long data is kept
- DATA-GROUP: lists of data items collected for stated purpose (i.e. Columns in the DB)
- CONSEQUENCE: human-readable description of usage of collected data

# Hippocratic DB simplified architecture



# Current Technology in PPDM

# PPDM Papers



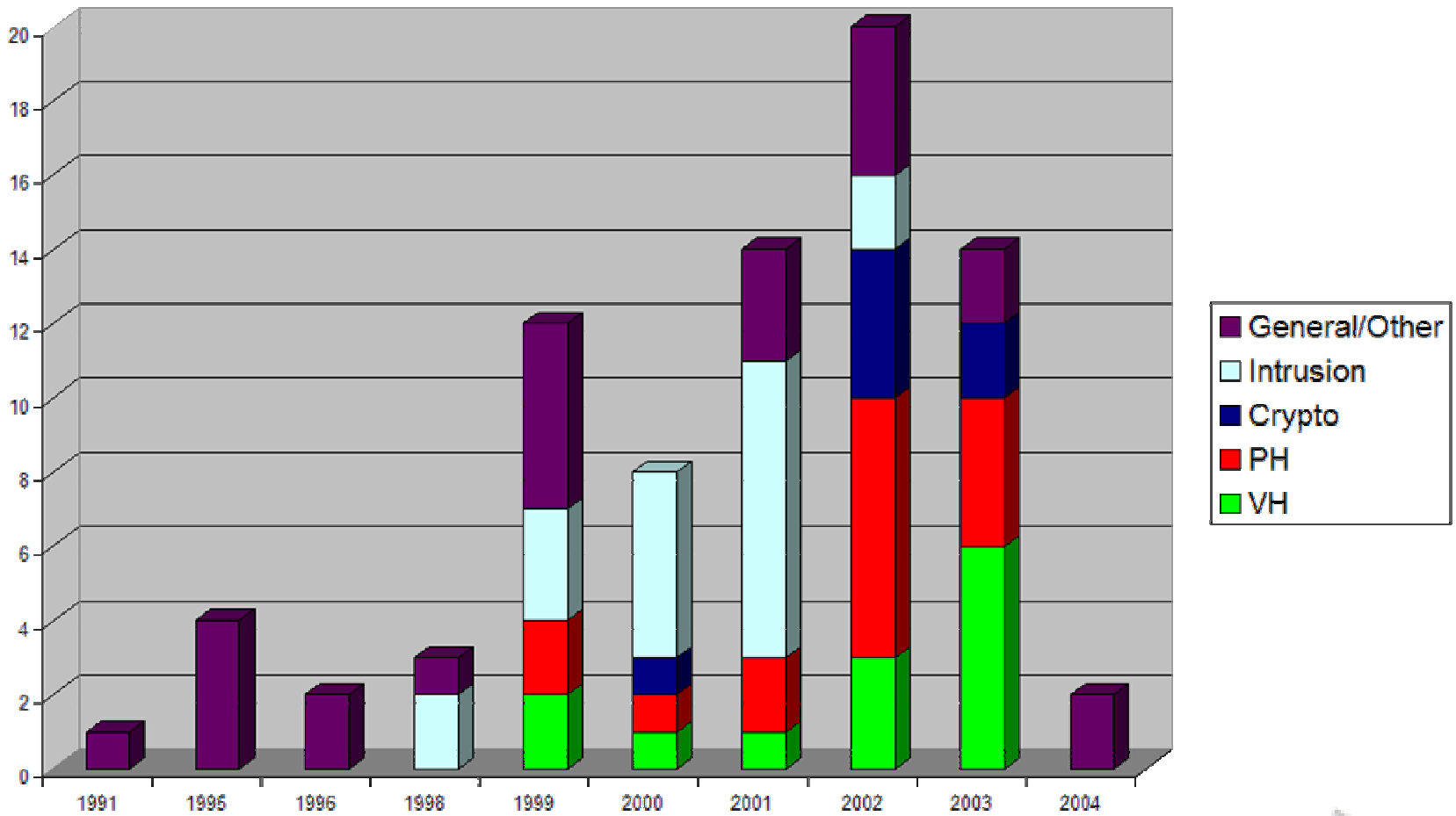

Source: *The Privacy, Security, and Data Mining Site.* Stanley Oliveira (Dec 2003)

http://www.cs.ualberta.ca/~oliveira/psdm/psdm_index.html

# Approaches

- **Centralized database**
  - **Value Hiding**
    - AKA: Data Perturbation, *Reconstruction Based*
  - **Pattern Hiding**
    - AKA: Data Sanitization, *Heuristic Based*
- **Distributed databases**
  - **Value Hiding during communications**
    - AKA: Secure Multiparty Computations (SMC), *Cryptographic Based*

# PPDM Papers




Source: *The Privacy, Security, and Data Mining Site.* Stanley Oliveira (Dec 2003)

`http://www.cs.ualberta.ca/~oliveira/psdm/psdm_index.html`

# Current Technology in PPDM

Randomization

# Value Hiding: the Idea

- Since the primary task in data mining is the development of models about aggregated data,
  - Can we develop accurate models without access to precise information in individual data records?

# Value Hiding: the Problem

- **Given:**
  - **a database source *D*,**
  - **a subset $A_h$ of the attributes in *D***
- **We want:**
  - **a new database *D´* with the same attributes of *D* such that $\forall\ A \in A_h$ :**
    - For each record, we cannot know the original value of the attribute *A*
    - The distribution of *A* in *D´* is quite the same as the one in *D* (i.e. *D´* is good to be mined)

# Value Hiding: Brief History

- From works on Statistical Databases ('80)
  - Answer statistical queries while preserving individual "privacy"
  - Based on:
    - Query restriction
    - Noise addiction
      - Data Swapping
      - Value Discretization
      - Value Distortion

# Statistical DB: Data Swapping

- *k*-order statistics are those that employ exactly *k* attributes
- A database *D* is $\kappa$-transformable if there exists a database *D´* that has no records in common with *D*, but has the same *k*-order COUNTs for $k \in \{0, \ldots, \kappa\}$
  - Intractable problem
- Approximate Data Swapping
  - Replace the original *D* with randomly generated records, so that *D´* has similar *k*-order statistics as the original one

# Data Swapping in Classification

- The confidential attribute is the class attribute
- Build an induced decision tree
- Swap class values between records belonging to the same path
  - Now we have a new DB where the confidential attribute is "hidden"
  - Balancing privacy against precision:
    - Swap internal nodes (near the root) leads to more privacy
    - Swap only leaves leads to optimum precision, bad privacy

# Data Swapping in Classification

- Pro's
  - Each record is (in some ways) "privacy preserved"
  - You can induce a "good" classifier
  - Low cost
- Drawbacks
  - Algorithm depending (C4.5)
  - Unsuitability for on-line databases
  - Low precision if we want good privacy
  - <u>You can use the induced tree to perform privacy breaches!!!</u>

# The "Honest Data Miner" Assumption

an honest data miner

# On Line Noise Addiction

| Name | Age | Incomes |
|---|---|---|
| **Maurizio** | **26** | **15000** |

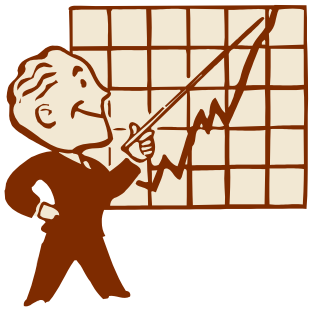
Perturbation (Client side) of:
Age, Incomes

| | | |
|---|---|---|
| **Maurizio** | **31** | **7234** |

Client side

Send to the server

Server side



# Value Discretization

- **Discretization is unuseful for privacy preserving data mining**
  - Many values: less privacy
  - Few classes: not very good privacy and no accuracy

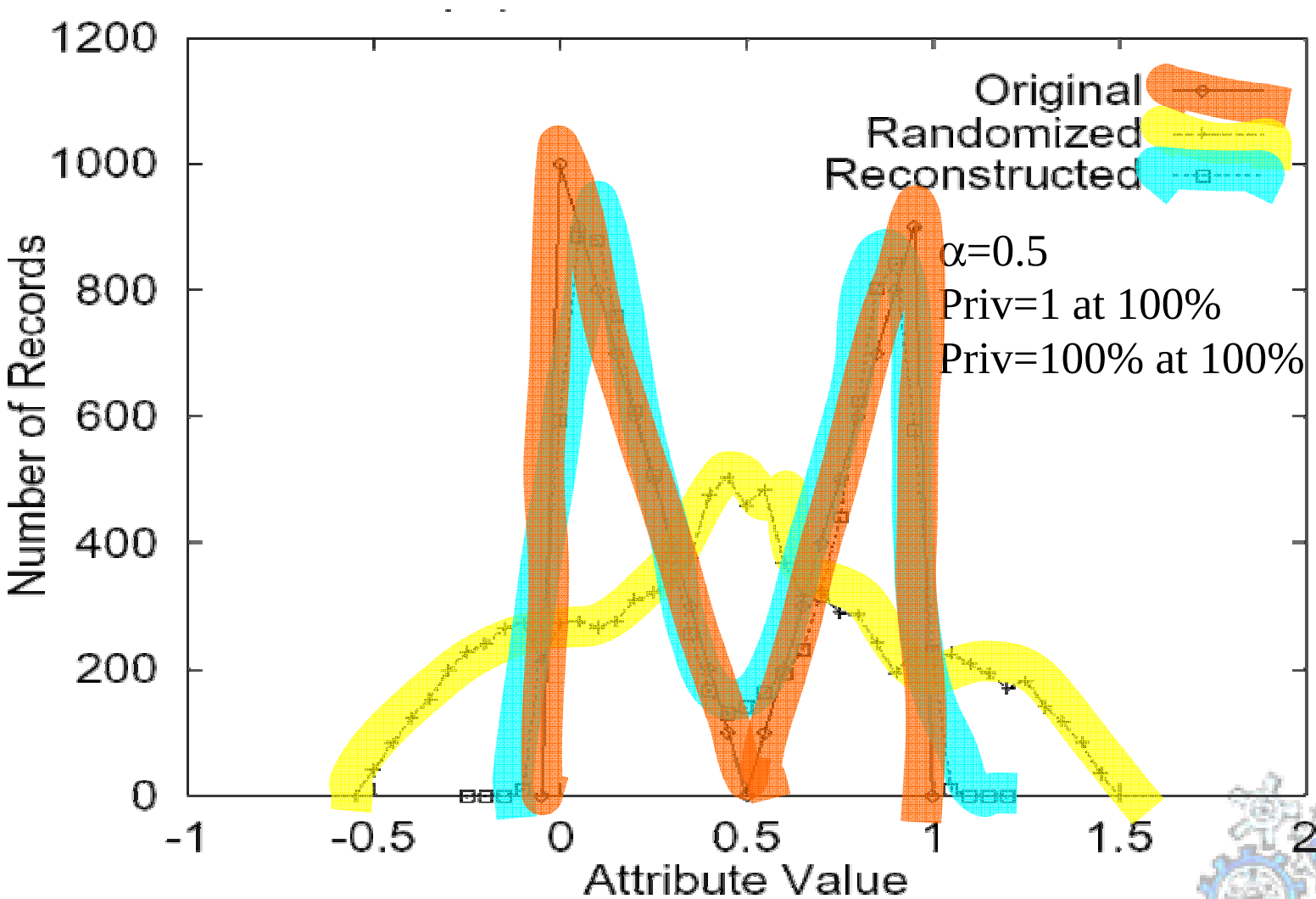
# Value Distortion

- **Basic idea:**
  - **The client return *x+r* instead of the actual value *x*, where:**
    - r is a random value from a known distribution
      - **Uniform**: random variable [-$\alpha$, +$\alpha$]
        - mean = 0
      - **Gaussian**: random variable
        - mean =0 , standard deviation = $\sigma$
  - **Note: The perturbation *r* of each entity should be fixed**
    - Repeated queries are unuseful for snoopers!

# First Privacy Metric

- If it can be estimated with *c%* confidence that a value *x* lies in the interval *[x1, x2]* then the interval width *(x2-x1)* defines the amount of privacy at *c%* confidence level
- The privacy is alternatively expressed as a percentage: (interval width/attribute range of values)
- Example: Age=26, Uniform with $\alpha$=7
  - *r* = 5 $\Rightarrow$ Perturbed_Age = Age + r = 31
  - Privacy = 14 at 100% confidence level
    - If Age $\in$ [10..120], Privacy = 14/110 at 100% confidence level
  - Privacy = 7 at 50% confidence level

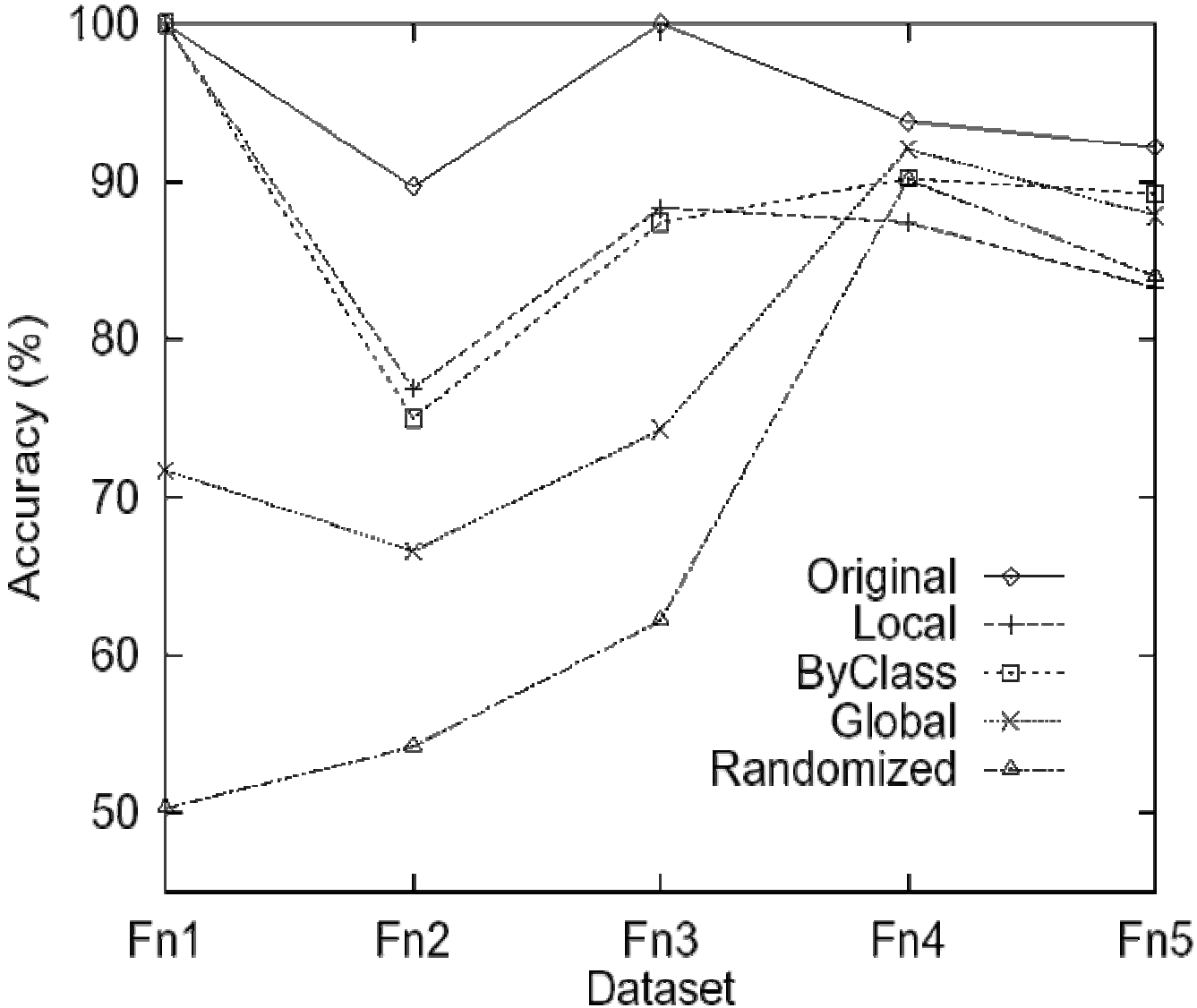
# The AS Algorithm



# DT-Classification Over Randomized Data

- 3 algorithms based on AS reconstruction:
  - Global
    - Reconstruct the distribution once at the beginning
  - ByClass
    - Once for each attribute, split the training data by class, then reconstruct the distributions separately for each class
  - Local
    - Like ByClass, but for each node instead of once

# AS Classification Performance



# AS Classification Results

- **Considerations:**
  - **Global is cheap but low accuracy**
  - **Local is expensive and accuracy is similar to ByClass**
    - ByClass is the best compromise!
- **Futhermore:**
  - **There is an accuracy/privacy tradeoff but:**
    - Original 90% accuracy
    - Reconstructed ByClass > 80% at 100% privacy, 70-80% accuracy at 200% privacy

# Second Privacy Metric

- Based on the concept of differential entropy of a random variable:

$$h(A) = -\int_{\Omega_A} f_A(a) \log_2 f_A(a) da$$

  - Where $\Omega_A$ is the domain of $A$ and $f_A$ is the density function of $A$

- The privacy of a random variable A is:

$$\Pi(A) = 2^{h(A)}$$

# Intuitions about $\Pi$

- A random variable *U* distributed uniformly between *0* and *a* has privacy:

$$\Pi(U) = 2^{h(U)} = 2^{\log_2(a)} = a$$

- Thus, if $\Pi(A)=2$ then *A* has as much privacy as a random variable distributed uniformly in an interval of length *2*

# Other Definitions

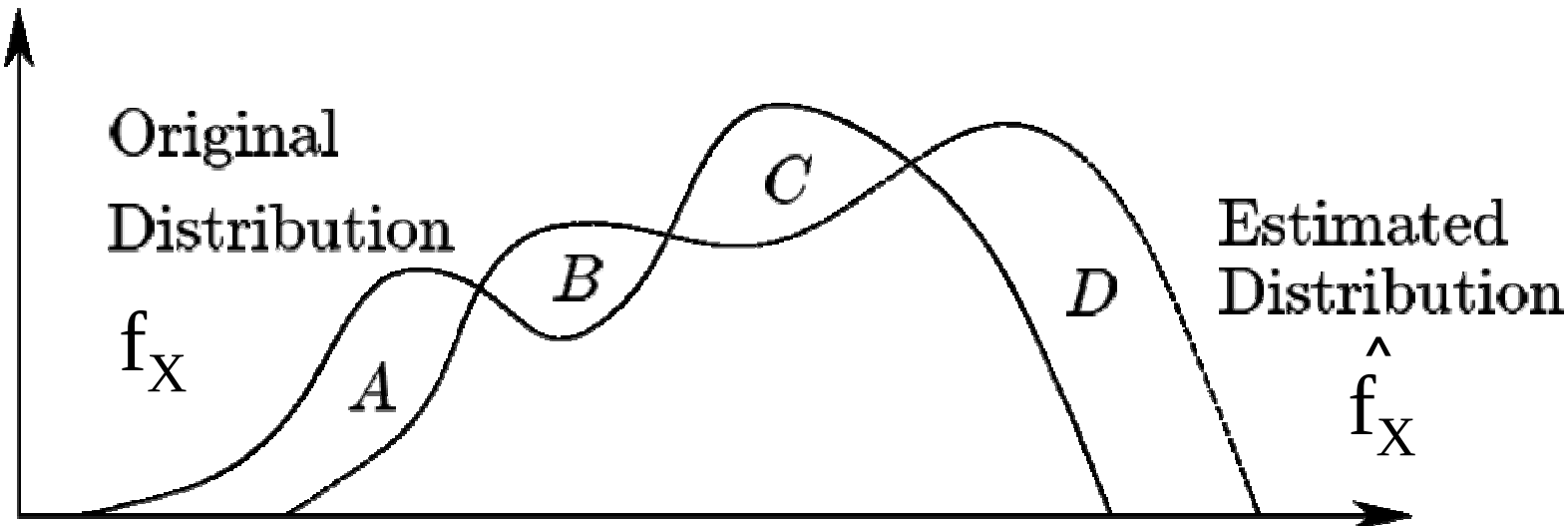
- Conditional privacy loss of *A* given *B*

$$\mathcal{P}(A|B) = 1 - \frac{\Pi(A|B)}{\Pi(A)} = 1 - 2^{-I(A;B)}$$

- Information loss

$$I(f_X, \hat{f}_X) = \frac{1}{2} E\left[\int_{\Omega_X} \left| f_X(x) - \hat{f}_X(x) \right| dx\right]$$

# Information Loss




$$I(f_X, \hat{f}_X) = \frac{1}{2} E\left[ \int_{\Omega_X} \left| f_X(x) - \hat{f}_X(x) \right| dx \right]$$

It is equal to 1-$\alpha$, where $\alpha$ is the area shared by both distributions

# The EM Algorithm

- Theorem: when there is a very large number of data observations, then the EM algorithm provides zero information loss
- For reasonably large perturbations:
  - 20000 points $\Rightarrow$ < 0.5% Information Loss

Note: in some sense, this result is related to *k-anonimity,* because if points are few then we get no information

# AR randomization

- Similar approaches to the problem of hiding items
  - each item changes its status (present or not present in the transaction) with probability *p*
    - *Items can be removed*
    - *New items can be inserted*
- Problems for itemsets (shown to be few privacy preserving)

# Random Data Perturbation breaches

- A paper asserts that using random matrices theory it is possible to predict structure in the spectral domain
  - A matrix-based spectral filtering technique has been shown to predict original data from observed data, not only the distribution
- Some (strong?) assumptions on data
  - E.g., SNR (signal-to-noise ratio) > 1
- Some other breaches in AR item hiding
  - Trying to classify and deeply understand privacy breaches

# PP Clustering by Data Transformation

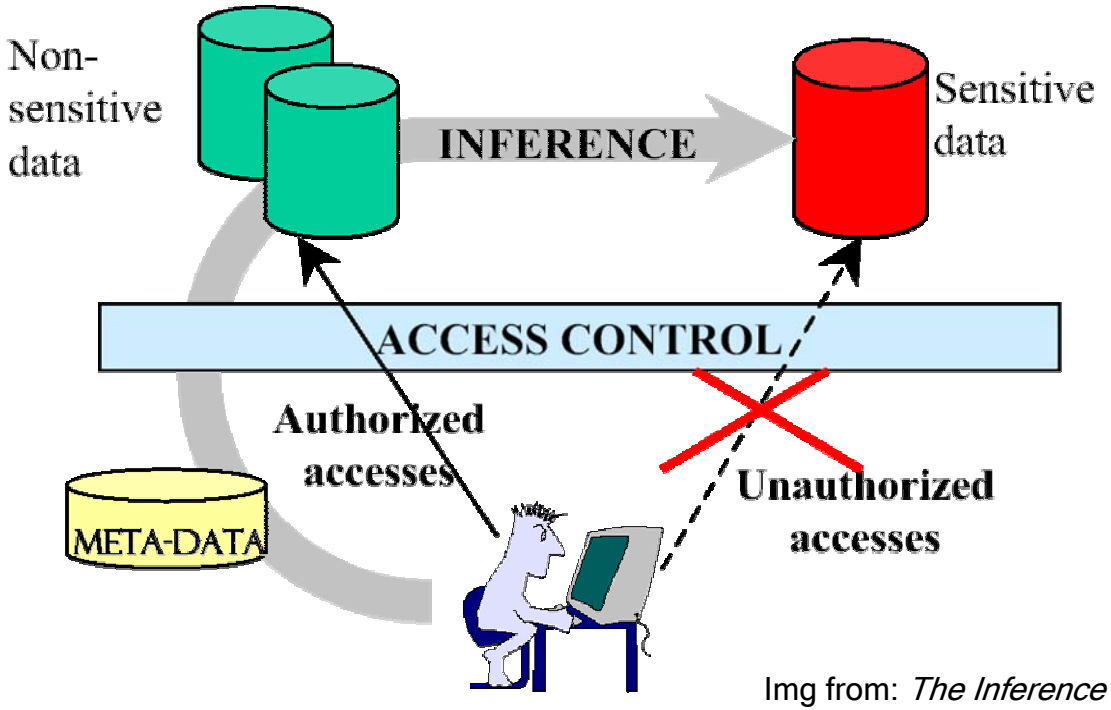
- The authors use GDTMs (geometric data transformation methods) to "randomly" modify the data, but preserving geometric structure
- The dataset (sensible data projection) can be viewed as a matrix
  - Translation
  - Rotation
  - Scaling

# Current Technology in PPDM

Knowledge Hiding

# Pattern Hiding: the Idea

- Clifton's Tutorial title: *When Do Data Mining Results Violate Privacy?*
  - Question: Do the results themselves violate privacy?
  - Very Related to the Inference Problem




Img from: *The Inference Problem: A Survey.* C.Farkas, S. Jajodia

# Pattern Hiding: the Problem

- **Given:**
  - **a database source $D$,**
  - **a subset $R_h$ of the set of significant patterns $R$ that can be mined from $D$**
- **We want:**
  - **a new (sanitized) database $D'$ with the same attributes of $D$ such that $\forall A \in P$ :**
    - $R_h$ cannot be mined from $D'$
    - $R / R_h$ can still be mined from $D'$

# Hiding AR using Confidence and Support

- **Conf(X ⇒ Y)= Supp(XY) /Supp(X)**
  - **E.g. A,C ⇒ B (conf=c,supp=s)**
- **3 strategies**
  - **Decreasing the Confidence**
    - Increasing support of the rule antecedent X, through transactions that partially support both X and Y
      - E.g. A ⇒ AC
    - Decreasing support of the rule consequent Y, in transactions that support both X and Y
      - E.g. ABC ⇒ AC
  - **Decreasing the Support**
    - Decreasing the support of either the rule antecedent X or the rule consequent Y
      - E.g. ABC ⇒ AB

# Using Unknowns

- The previous proposal can bring to misleading rules
  - This is not good if rules are used in diagnosis!
- Solution: as before but
  - replace "1" and "0" with "?"

# AR Hiding in general

- The problem is NP-hard:
  - Heuristics are used
  - Iterative process
  - No guarantees to converge in few passes
    - The final dataset can be very different from the original
    - The sanitization process can take too much time
- The sanitization process is "algorithm dependent"!!!
  - I.e, what if we mine Correlation Rules instead of AR rules?

# Current Technology in PPDM

Cryptography

# Distributed Data Mining

- Data is distributed among sites
  - Each site is allowed to see real data item
  - No site is allowed to see other's data
- No need to combine all data for mining
- Distribute computing
  - Each site participates to a protocol to get mining results
  - The protocol does not disclose private data to other sites

# Trusted Party Model

- In addition to the parties there is a trusted party who does not attempt to cheat
- All parties send their inputs to the trusted party, who computes the functions and sends back results to other parties
- A protocol is secure if anything that an adversary can learn in real world it can also learn in ideal world
- The protocol does not leak any unnecessary information

# Partial Leaks of Information

- It is possible to have partial leaks of information that are harmless
- It is hard to decide how much (which type) of leakage can be tolerated
- Cryptographic protocols aim to avoid any information disclosure, except for output

# Adversarial Behavior

- Semi-honest adversary
  - it is a party that follows the protocol specification, yet attempts to learn additional information by analyzing the messages received during the protocol execution
- Malicious adversary
  - it is a party that arbitrarily deviates from the protocol specification

# Protocol Design Approach

- First design a secure protocol for semi-honest case
- Then transform it into a protocol that is secure against malicious adversaries
  - for example, by means of zero-knowledge proofs
- However, semi-honest model is often a realistic one

# Protocol Building Blocks

- Oblivious Transfer
  - It was shown by Kilian that that given an implementation of oblivious transfer, and no other cryptographic primitive, one could construct any secure computation protocol
- Secure Multiparty Computation
  - Commutative Encryption
    - Secure Sum
    - Secure Set Union
    - Secure Set Intersection
    - Scalar Product

# Commutative Encryption

- Quasi-commutative hash functions h
  - given
  - the value
  - is the same for every permutation of $y_i$
  - if $x \neq x'$ then $z \neq z'$
- An example: public key encryption (RSA)
  - a function pair: $E_A, D_A$

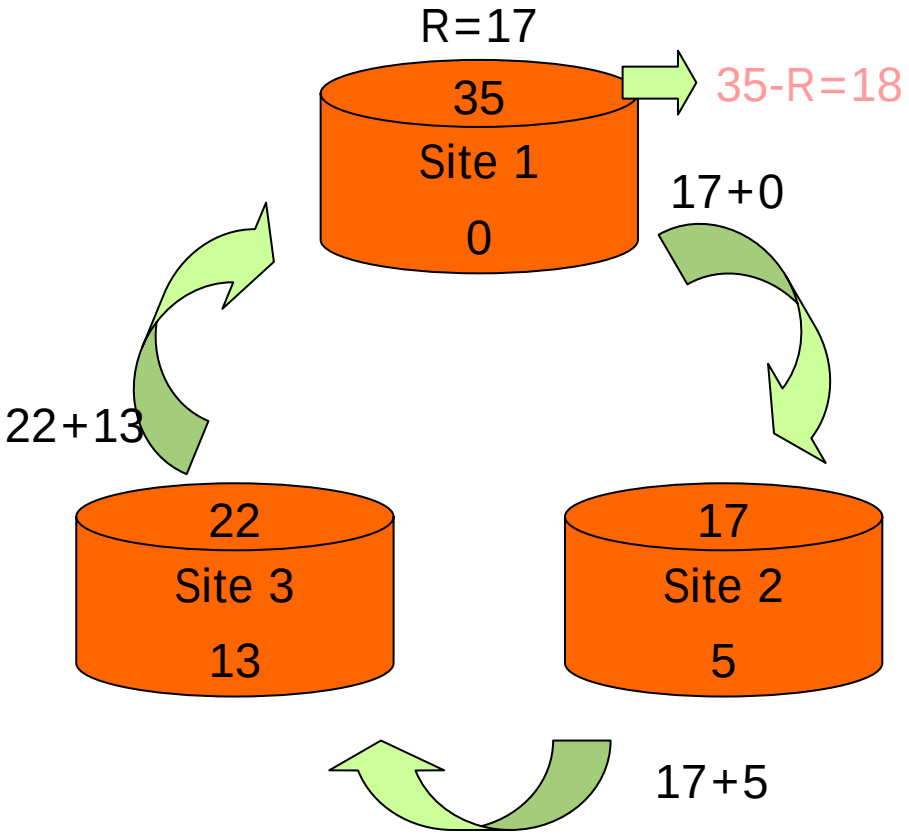
$$E_A(D_A(x)) = x$$

$$\Pr(E_B(x) = E_A(x)) \cong 0$$

$$E_A(E_B(x)) = E_B(E_A(x))$$

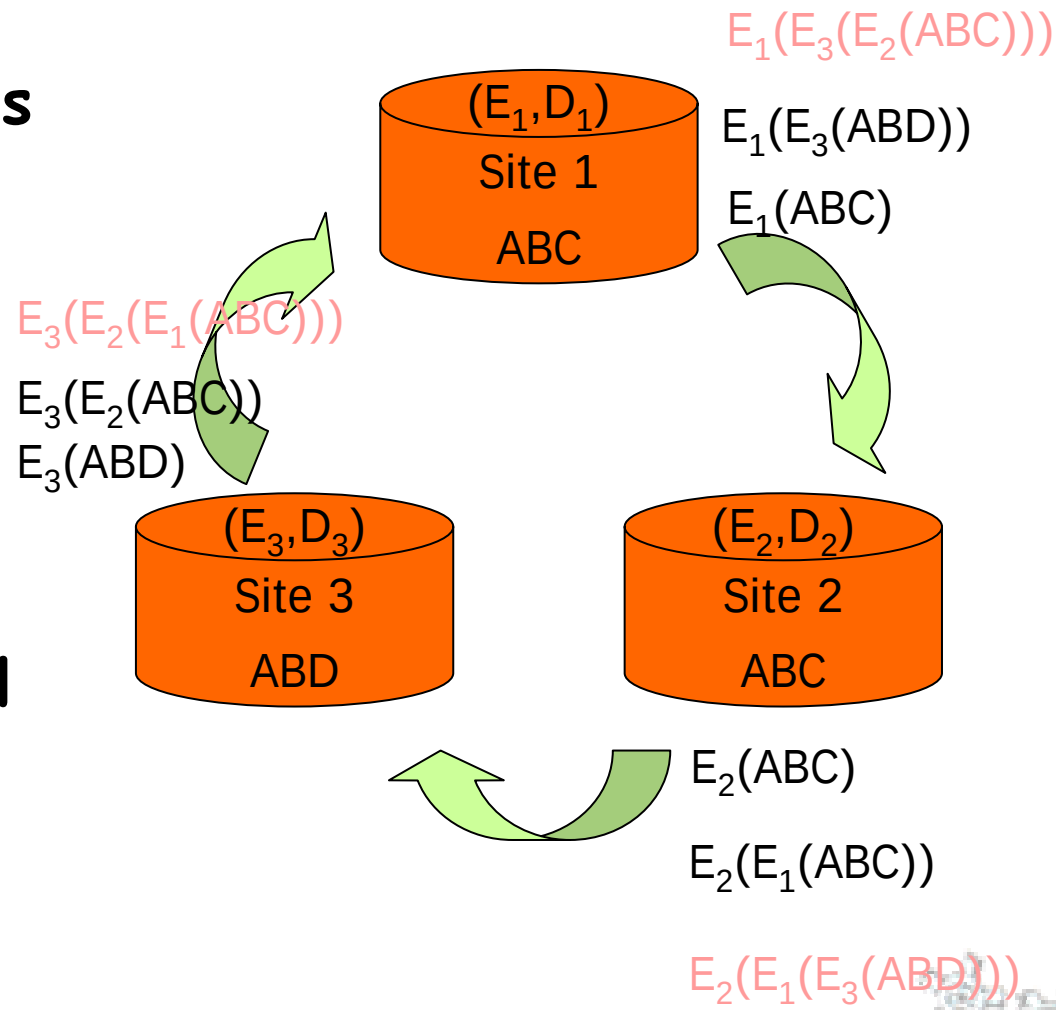
# Secure Sum

- One site designed as master
- Others are numbered from 2 to s
- Site 1 generates a random number R and compute $R+v_1$ mod n
- Site 2 learns nothing about $v_1$ and adds $v_2$ to value received
- For the remaining sites, protocol is analogous
- Site 1, knowing R, get actual result



# Secure Set Union/Intersection

- Each site i generates a key pair $(E_i, D_i)$
- Each site encrypts its items
- Each site encrypts items from other sites
- Duplicates in original values will be duplicates in encrypted values



# Mining AR in Horizontally Partitioned Data

- Candidate Set Generation: intersect globally large (k-1)-itemsets with locally large (k-1)-itemsets to get $CG_{i(k)}$
- Local Pruning: for each X in $CG_{i(k)}$ scan $DB_i$ locally to compute local support $X.sup_i$. If X is locally large include it in $LL_{i(k)}$
- Itemset Exchange: securely compute the union of each $LL_{i(k)}$ to obtain $LL_{(k)}$ (using Secure Set Union)
- Support Count Exchange: securely compute support for each itemset in $LL_{(k)}$ (using Secur Sum)

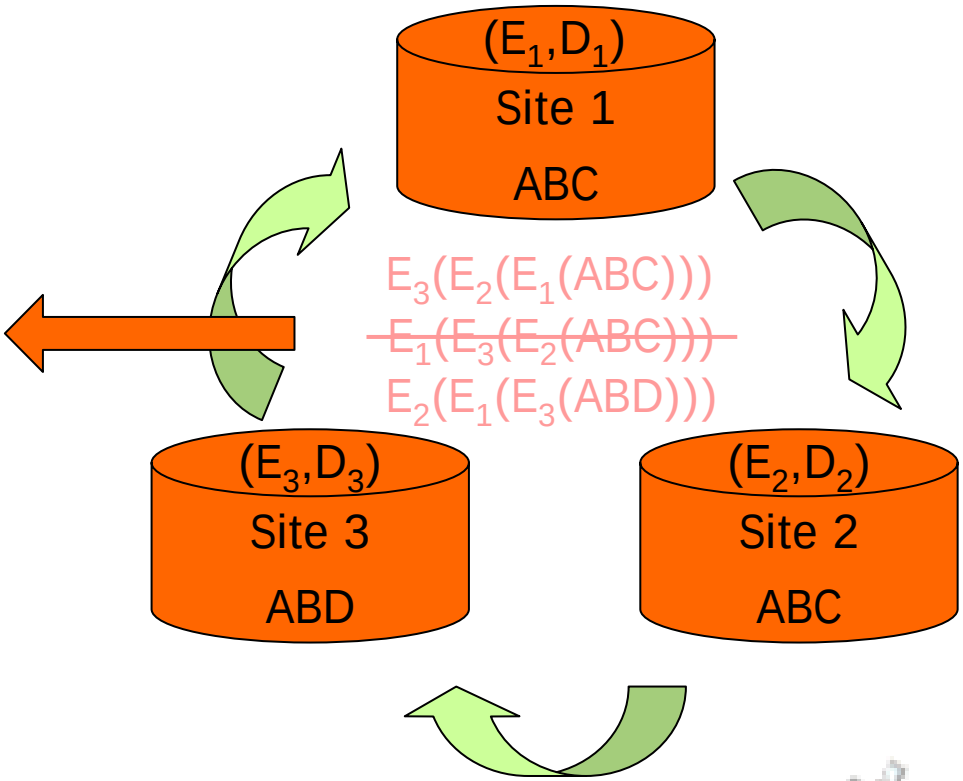
# Mining AR in Horizontally Partitioned Data (2)

## Finding secure union of large itemsets

$D_3(D_2(D_1(E_3(E_2(E_1(ABC)))))))$

$D_2(D_1(D_3(E_3(E_2(E_1(ABD)))))))$

{ABC,ABD}

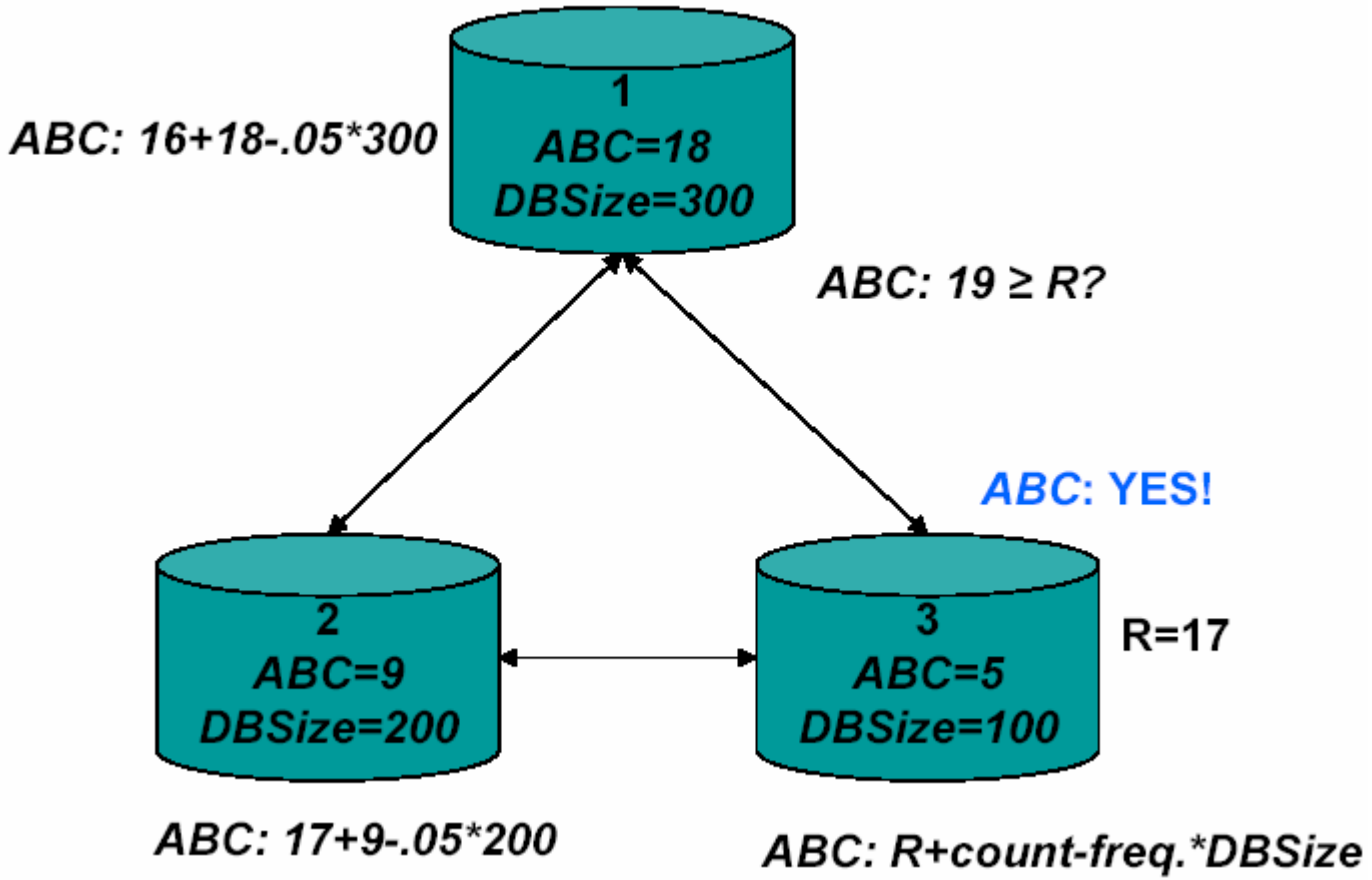


# Which Candidates Are Globally Supported?

- Now securely compute Sum ≥ 0:
  - Site 0 generates random R
    - Sends R+count0 – frequency*dbsize0 to site1
  - Site k
    - adds countk – frequency*dbsizek, sends to sitek+1
- Final result: Is sum at site n - R ≥ 0?
- Use Secure Two-Party Computation
- This protocol is secure in the semi-honest model

# Compunting frequency: ABC > 5% ?

⌘



# Conclusions

# Conclusions

- Still room for improvements
- A general accepted definition of privacy is still missing
- The main question still need an answer:
  - Do data mining results violate privacy?