Privacy-Preserving Data Mining and E-commerce & E-government

Tu Bao Ho

School of Knowledge Science

Japan Advanced Institute of Science and Technology

and

IOIT, Vietnamese Academy of Science and Technology

Outline

Data mining and e-commerce, e-government

Privacy-preserving data mining

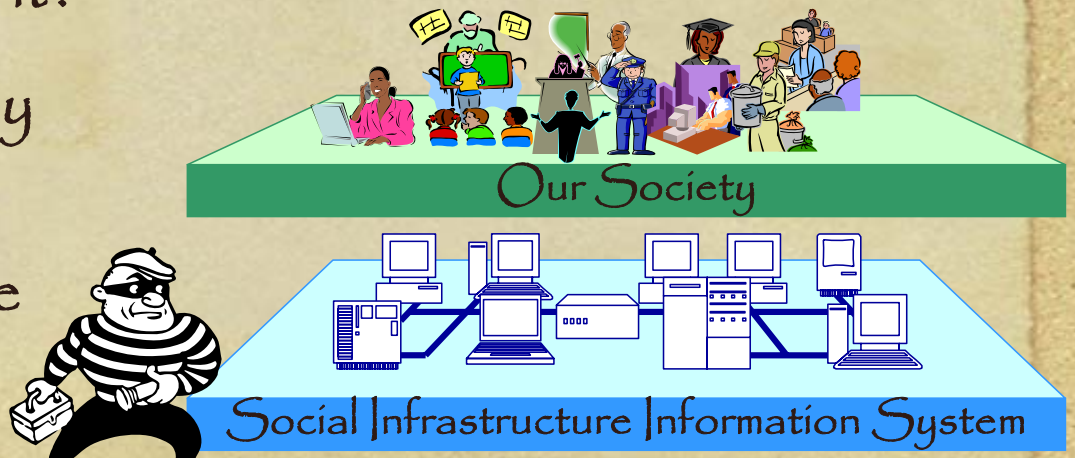
Requirements for trustworthy e-society

- E-government: networked and digitalized administration.
- E-commerce: Online business

- Can you trust e-society infrastructure information system and leave your life to it?
  - Is your private data illegally accessed or altered?
  - Is it possible for enterprise data to be stolen?

COE program on "Verifiable and Evolvable E-Society" (2004-2009)

1. Correctness
2. Accountability
3. Security
4. Fault Tolerance
5. Evolvability
6. Trustworthy infrastructures



Just the tip of the iceberg for consumers and for enterprises...

Data on individuals and enterprises are widely available from electronic databases

Business Intelligence
A Corp. ordering $35,000,000 of our product

Password Files

Financial Data
ABC corp. will be reporting a loss of $1.20 per share

Intellectual Property
Secret beverage recipe: Sugar, water, and a hint of CO_2

Data mining: An interdisciplinary field

- Data collection
- Data access
- Data analysis

Machine Learning

Build computer systems that learn as well as humans do (learning from data).

Discovery of new and useful knowledge (patterns/models) in databases

Statistics

Infer information from data (deduction and induction, mainly numeric data)

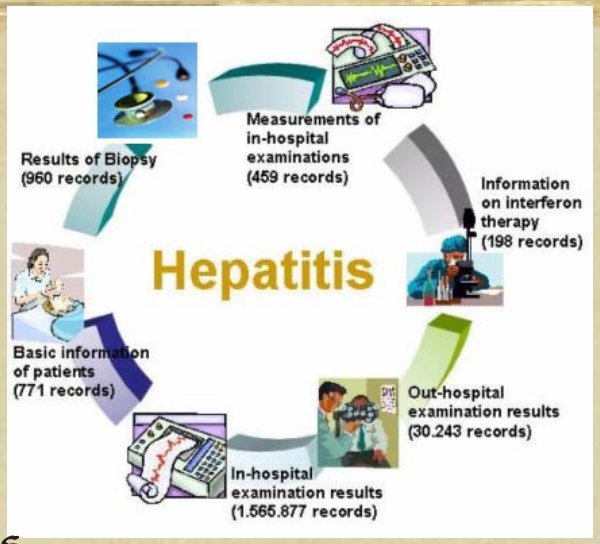
Databases

Store, access, search, update data (deduction)

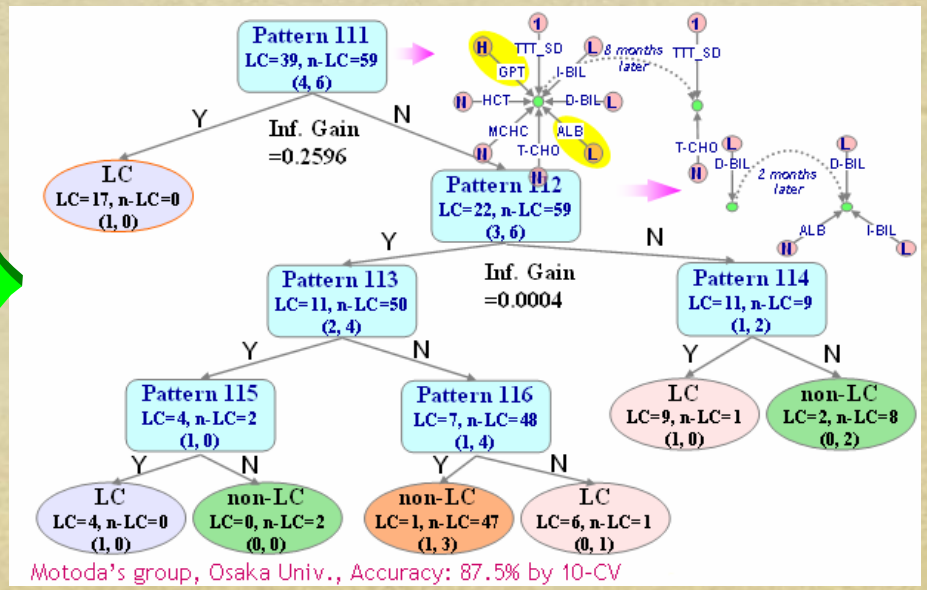
What does data mining do?

Tasks and methods

- Classification and Prediction
  - Decision trees
  - Neural network
  - Rule induction
  - Support vector machines
  - Hidden Markov Model
  - etc.
- Description
  - Association analysis
  - Clustering
  - Summarization
  - etc.



IF ALB = NormalToLow
AFTER
TP = high & peaks
THEN Liver cirrhosis (LC)



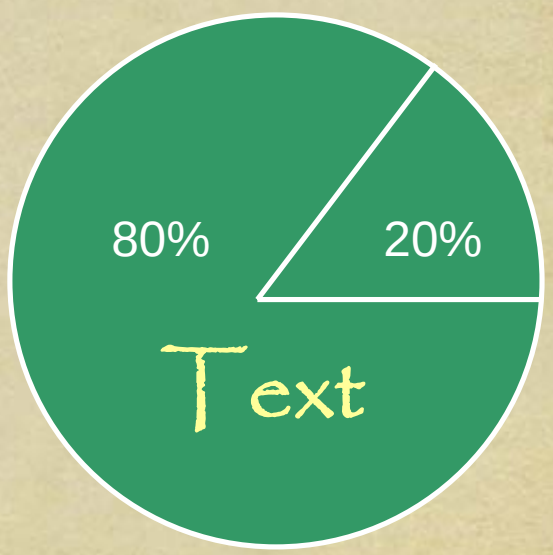
Motoda's group, Osaka Univ., Accuracy: 87.5% by 10-CV

Text mining: A typical example

- Extract pieces of evidence from article titles in the biomedical literature (Swanson and Smalheiser, 1997)
  - "stress is associated with migraines"
  - "stress can lead to loss of magnesium"
  - "calcium channel blockers prevent some migraines"
  - "magnesium is a natural calcium channel blocker"

Induce a new hypothesis not in the literature by combining culled text fragments with human medical expertise

- Magnesium deficiency may play a role in some kinds of migraine headache



Web mining

Typical data in a server access log

looney.cs.umn.edu han - [09/Aug/1996:09:53:52 -0500] "GET mobasher/courses/cs5106/cs5106l1.html HTTP/1.0" 200

mega.cs.umn.edu njain - [09/Aug/1996:09:53:52 -0500] "GET / HTTP/1.0" 200 3291

mega.cs.umn.edu njain - [09/Aug/1996:09:53:53 -0500] "GET /images/backgnds/paper.gif HTTP/1.0" 200 3014

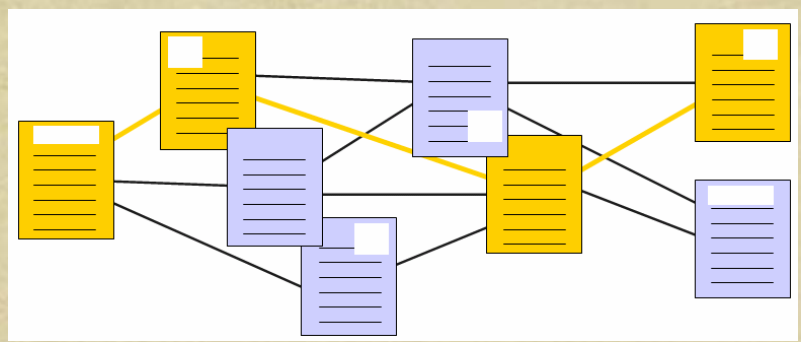
mega.cs.umn.edu njain - [09/Aug/1996:09:54:12 -0500] "GET /cgi-bin/Count.cgi?df=CS home.dat\&dd=C\&ft=1 HTTP

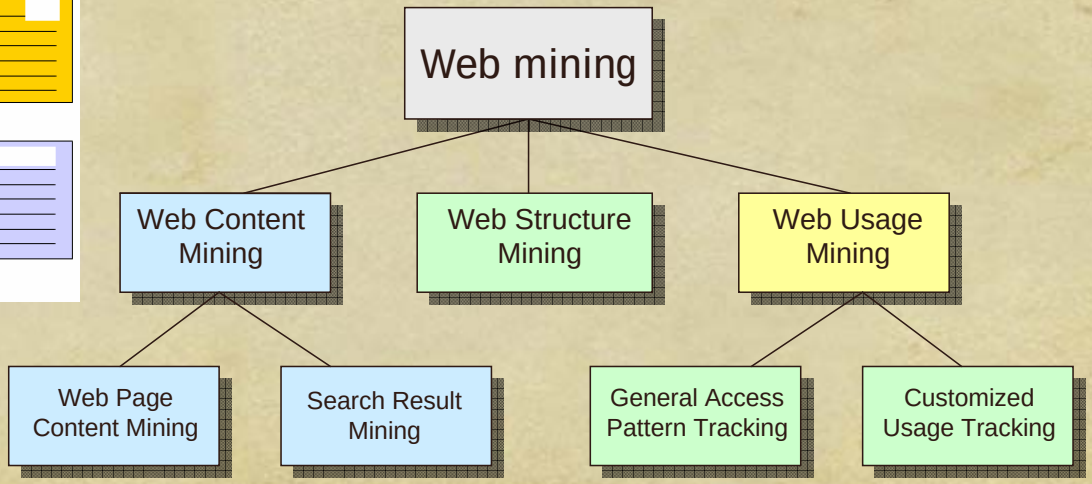
mega.cs.umn.edu njain - [09/Aug/1996:09:54:18 -0500] "GET advisor HTTP/1.0" 302

mega.cs.umn.edu njain - [09/Aug/1996:09:54:19 -0500] "GET advisor/ HTTP/1.0" 200 487

looney.cs.umn.edu han - [09/Aug/1996:09:54:28 -0500] "GET mobasher/courses/cs5106/cs5106l2.html HTTP/1.0" 200

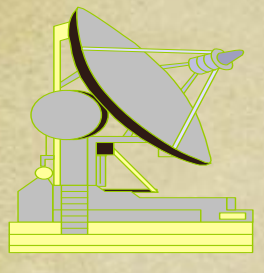
... ...





KDD: New and fast growing area

KDD'95, 96, 97, 98, ..., 04, 05 (ACM, America)
PAKDD'97, 98, 99, 00, ..., 04, 05 (Pacific & Asia)
http://www.jaist.ac.jp/PAKDD-05 (Hanoi)
PKDD'97, 98, 99, 00, ..., 04, 2005 (Europe)
ICDM'01, 02,..., 04, 05 (IEEE), SDM'01, ..., 04, 05 (SIAM)



Industrial Interest: IBM, Microsoft, Silicon Graphics, Sun, Boeing, NASA, SAS, SPSS, ...

Japan: FGCS Project focus on logic programming and reasoning; attention has been paid on knowledge acquisition and machine learning. Projects "Knowledge Science", "Discovery Science", and "Active Mining Project" (2001-2004)

E-commerce and Data Mining

- A person buys a book (product) at Amazon.com.
- Task: Recommend other books (products) this person is likely to buy
- Amazon does clustering based on books bought:
  - customers who bought "E-Commerce Security and Privacy", also bought "E-commerce & privacy: What net users want"
- Recommendation program is quite successful.

E-government and Data Mining

- Hospital data contains
  - Identifying information: name, id, address
  - General information: age, marital status
  - Medical information
  - Billing information
- Database access issues:
  - Your doctor should get every information that is required to take care of you
  - Emergency rooms should get all medical information that is required to take care of whoever comes there
  - Billing department should only get information relevant to billing
- Data usage issue: "who is doing what to which and how much critical information, when and from where"

Outline

Data mining and e-commerce, e-government

Privacy-preserving data mining

Data Mining and Privacy

- There is a growing concern among citizens in protecting their privacy
- Government and business have strong motivations for data mining
- Can we satisfy both the data mining goal and the privacy goal?

Protests over a National Registry

Privacy-Preserving Data Mining

- Allow multiple data holders to collaborate to compute important (e.g., security-related) information while protecting the privacy of other information.

- Particularly relevant now, with increasing focus on security even at the expense of some privacy.

Advantages of privacy protection

- Protection of personal information
- Protection of proprietary or sensitive information
- Fosters collaboration between different data owners (since they may be more willing to collaborate if they need not reveal their information)

10 challenging problems in data mining
(ICDM'05)

1. Developing
2. Scaling
3. Mining
4. Min
5. Dat
6. Distr
7. Data ms
8. D
9. [illegible]
10. Dealing with non-static, unbalanced and cost-sensitive data

The trade-off between sharing information for analysis and keeping it secret to corporate trade secrets and customer privacy is a growing challenge.

Three approaches to PPDM

- Distribute limited subset of data
  - E.g., Census bureau releases only some fields
  - Theory tells which subsets can be safely released
- Distribute purposely distorted data records
  - Nobody see the real data
  - Tell recipients the probabilistic distortion function
  - They can compute original data distribution, but not original data records
- Distribute the computation instead of data
  - Use cryptographic methods to assure privacy of intermediate computations

Distribute limited subset of data

- A naïve solution to the problem is de-identification —removing all identifying information from the data and then releasing it—but pinpointing exactly what constitutes identification information is difficult.
- Latanya Sweeney (2001)
  - Date of birth uniquely identifies 12% of the population of Cambridge, MA.
  - Date of birth + gender: 29%
  - Date of birth + gender + (9 digit) zip code: 95%
  - Sweeney was therefore able to get her medical information from an "annonymized" database

Distribute purposely distorted data records

- Goal: Hide the protected information
- Approaches: Data perturbation
  - Swap values between records: exchanging data values between records in ways that preserve certain statistics but destroy real values
  - Randomly modify data: adding noise to data to prevent discovery of the real values.
- Problems
  - Does it really protect the data?
  - Can we learn from the results?

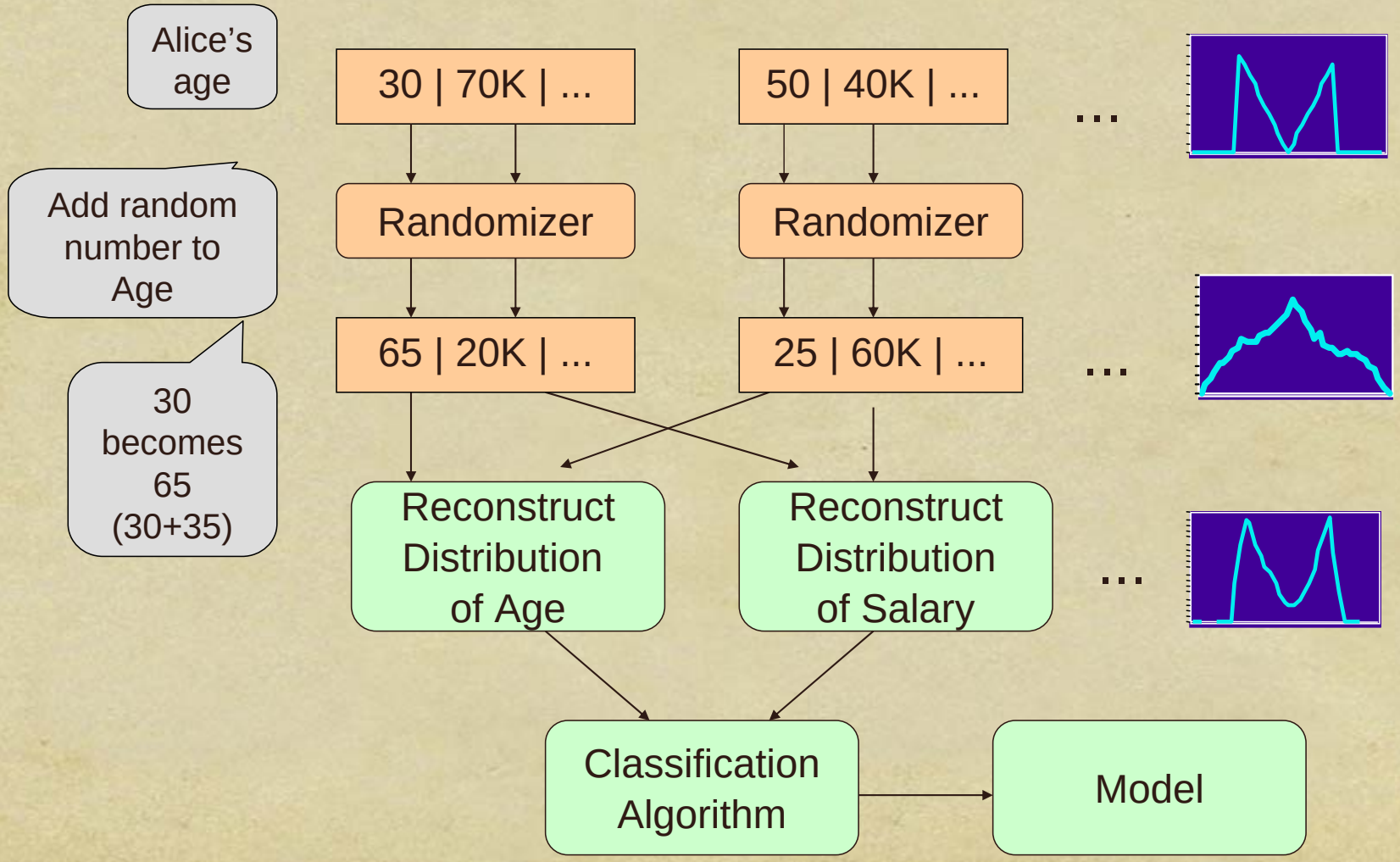
Distribute purposely distorted data records

- Miner doesn't see the real data
  - Some knowledge of how data obscured
  - Can't reconstruct real values
- Results still valid
  - Can reconstruct enough information to identify patterns
  - But not entities
- Example: Work of Agrawal & Srikant (2000)

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 $\mu = 0, \sigma$

Randomization approach overview



Reconstruction problem

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

Estimate the probability distribution of X.

Reconstructing the distribution

Combine estimates of where point came from for all the points:

Gives estimate of original distribution.

$$f_X = \frac{1}{n}\sum_{i=1}^{n} \frac{f_Y((x_i + y_i) - a)\, f_X^j(a)}{\int_{-\infty}^{\infty} f_Y((x_i + y_i) - a)\, f_X^j(a)}$$

Reconstruction: Bootstrapping

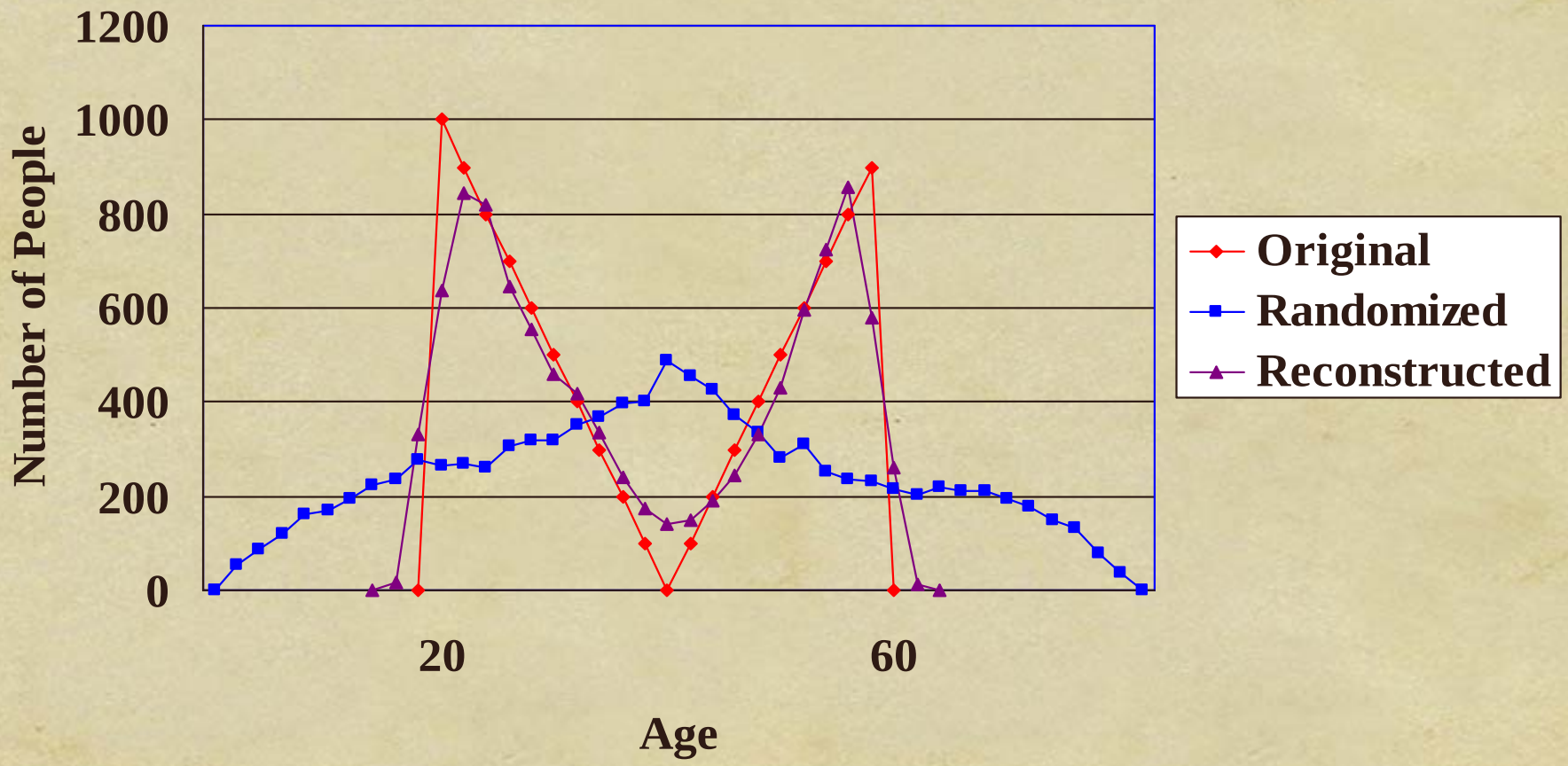
f_X^0 := Uniform distribution
j := 0 // Iteration number
repeat

$$f_X^{j+1}(a) := \frac{1}{n}\sum_{i=1}^{n}\frac{f_Y((x_i+y_i)-a)\,f_X^j(a)}{\int_{-\infty}^{\infty} f_Y((x_i+y_i)-a)\,f_X^j(a)} \qquad \text{(Bayes' rule)}$$

j := j+1
until (stopping criterion met)

- Converges to maximum likelihood estimate.
  - D. Agrawal & C.C. Aggarwal, PODS 2001.

Works well

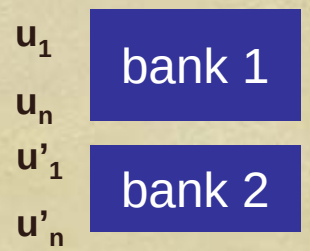

1200
1000
800
600
400
200
0
Number of People
20
60
Age
Original
Randomized
Reconstructed

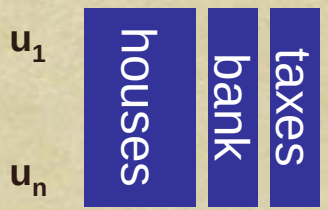
Distribute the computation instead of data

- Suppose
  - Multiple hospitals hold private patient data,
  - They wish to learn rules for SARS treatment effectiveness
  - But will not share details patient records
- Idea
  - Allow them to retain their data, and their individual privacy policies
  - Distribute computation
  - Use cryptographic techniques to maintain privacy of distributed computation

Scenario

- Multi database scenarios: Two or more parties with private data want to cooperate.
- Horizontally split: Each party has a large database. Databases have same attributes but are about different subjects. For example, the parties are banks which each have information about their customers.





- Vertically split: Each party has some information about the same set of subjects, e.g., the participating parties are government agencies; each with some data about every citizen.

Secure multiparty computation (SMC)

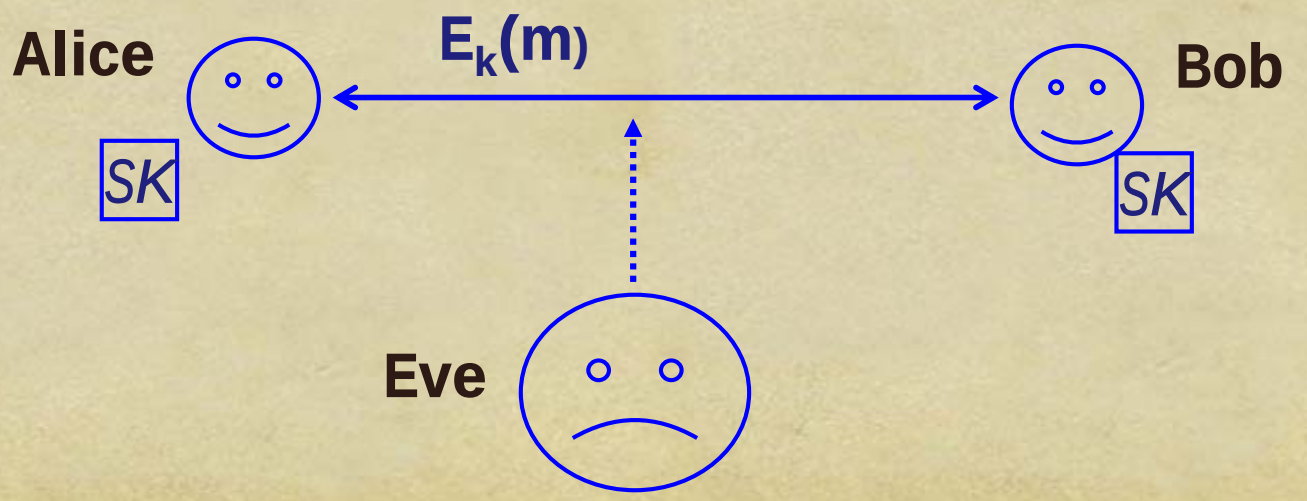
- A specialized form of privacy-preserving, distributed data mining.
- Parties that each know some of the private data participate in a protocol that generates the data mining results, yet that can be proven not to reveal data items to parties that don't already know them.
- The basic idea is that parties hold their own data, but cooperate to get the final result.

The methodology

- Because all interaction occurs through the messages sent and received, we simulate the views of all the parties by simulating the corresponding messages.
- If we can simulate these messages, then we can easily simulate the entire protocol just by running it.
- Instead, we use a notion from cryptography—the same message can be encrypted with different keys to look different, even though they represent the same message.

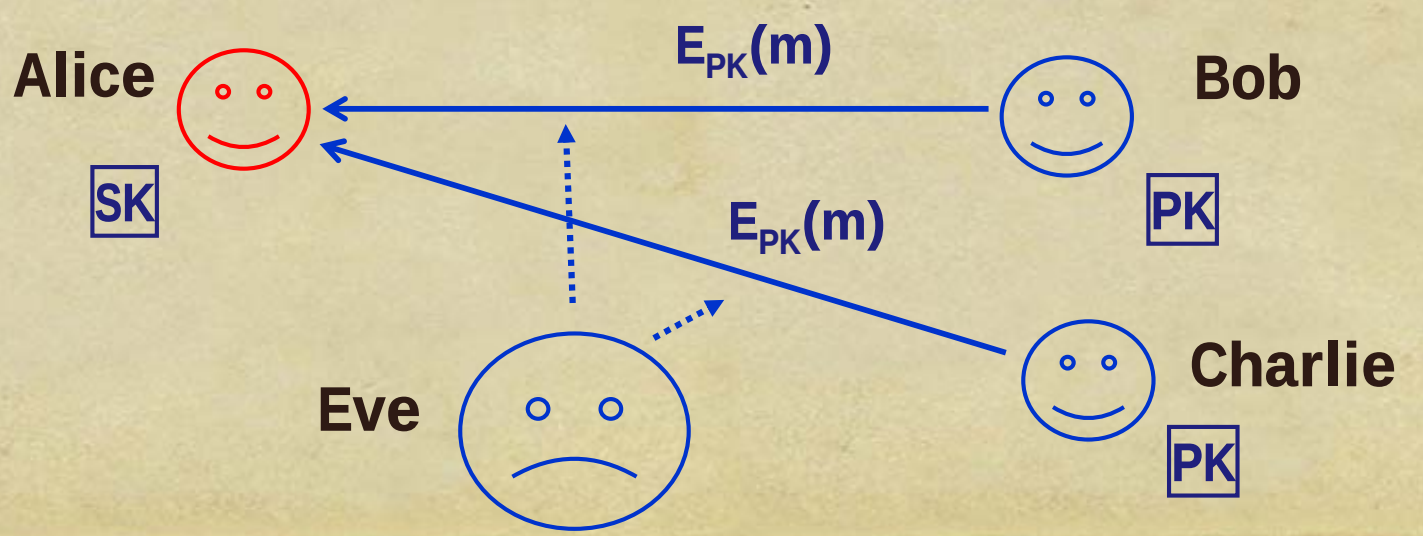
Encryption

- Alice wants to send a message $m \in \{0,1\}^n$ to Bob
  - Set-up phase is secret
  - Symmetric encryption: Alice and Bob share a secret key SK
- They want to prevent Eve from learning anything about the message.



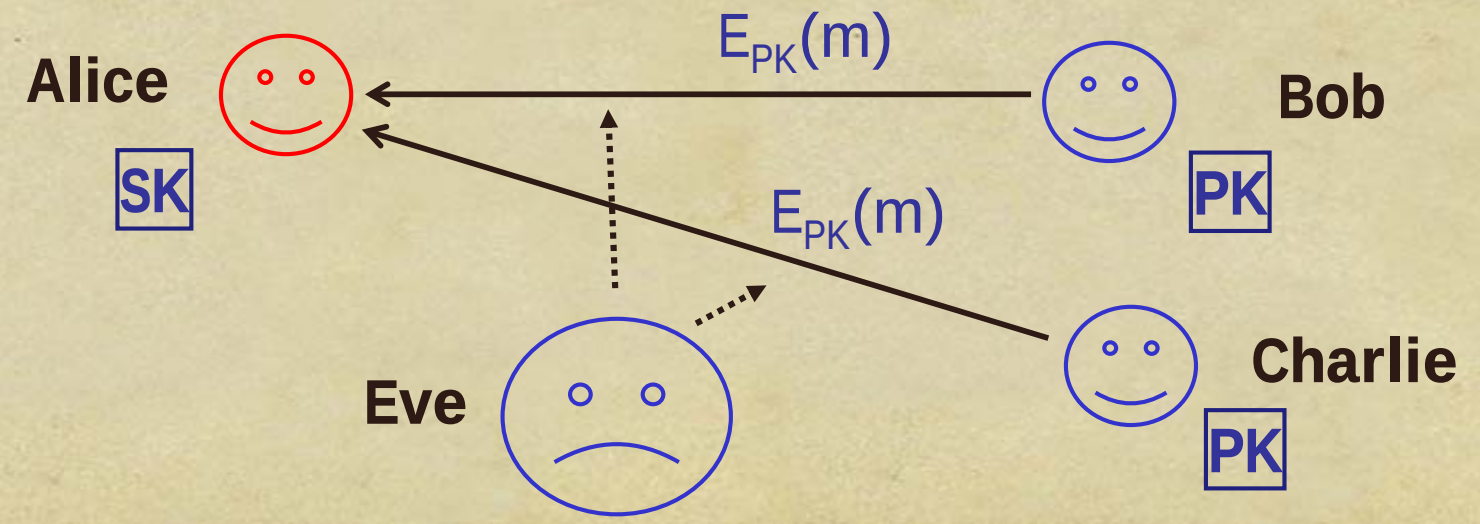
Public key encryption

- Alice generates a private/public key pair (SK,PK)
- Only Alice knows the secret key SK
- Everyone (even Eve) knows the public key PK, and can encrypt messages to Alice.
- Only Alice can decrypt (using SK)



Secure multiparty computation (SMC)

- A distribution of numbers is said to be computationally indistinguishable from another distribution of numbers if no polynomial time program can distinguish between the two distributions.
- As long as the sequence of numbers revealed during a protocol is computationally indistinguishable from numbers drawn from a random distribution, the protocol is assumed to be secure.



Progress has been made

- Secure two party-computation (Yao, 1986)
- Secure multiparty-computation (secure distributed computation) (Goldreich et al., 1987)
- SMC techniques for data mining: ID3 (Lindell and Pinkas, 2002), association rule mining (Clifton, 2003, 2004), etc.

Conclusion

- Privacy-preserving data mining is of growing importance, and some important progress has been made.
- Three ideas here:
  - Distribute only subset of data features
  - Distribute perturbed data records
  - Distribute computation rather than data
- Technical solutions can help (more work needed), but technology, policy, and education must work together.

References

- R. Agrawal, R. Srikant, "Privacy-Preserving Data Mining", ACM SIGKDD, 2000.
- C. Clifon, "Privacy-Preserving Distributed Data Mining", Tutorial ACM SIGKDD, 2003.
- T.B. Ho, Lecture on Knowledge Discovery and Data Mining, JAIST, 2005.
- Y. Lindell, B. Pinkas, "Privacy-Preserving Data Mining", J. Cryptology, Vol. 15, No. 3, 2002.
- T. Mitchell, "Privacy-Preserving Data Mining", CALD Summer School, June 2003.
- R. Srikant, "Privacy Preserving Data Mining: Challenges & Opportunities", Invited talk, PAKDD 2002.
- J. Vaidya, C. Clifon, "Privacy-Preserving Data Mining: Why, How and When", IEEE Security & Privacy, 2004.