University of Nebraska's

The Peter Kiewit Institute

College of Information Science & Technology

Privacy-preserving Data Mining of Medical Data using Data Separation Based Techniques

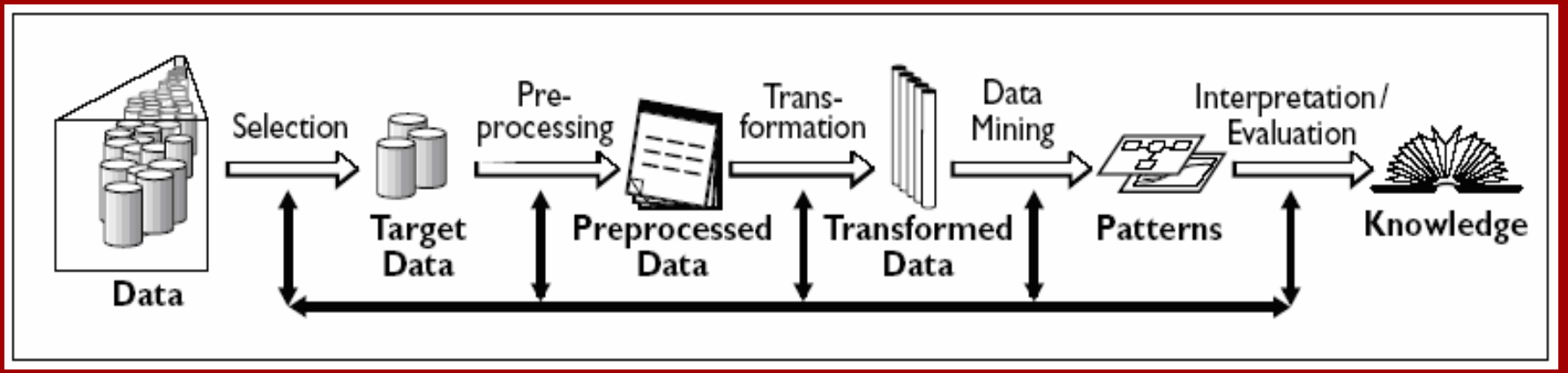
Gang Kou, Yi Peng, Yong Shi, and Zhengxin Chen

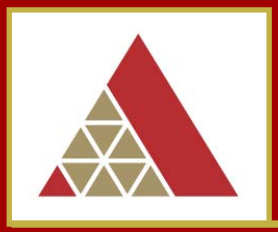
Data Mining and Knowledge Discovery

"Knowledge Discovery is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. "

— U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth

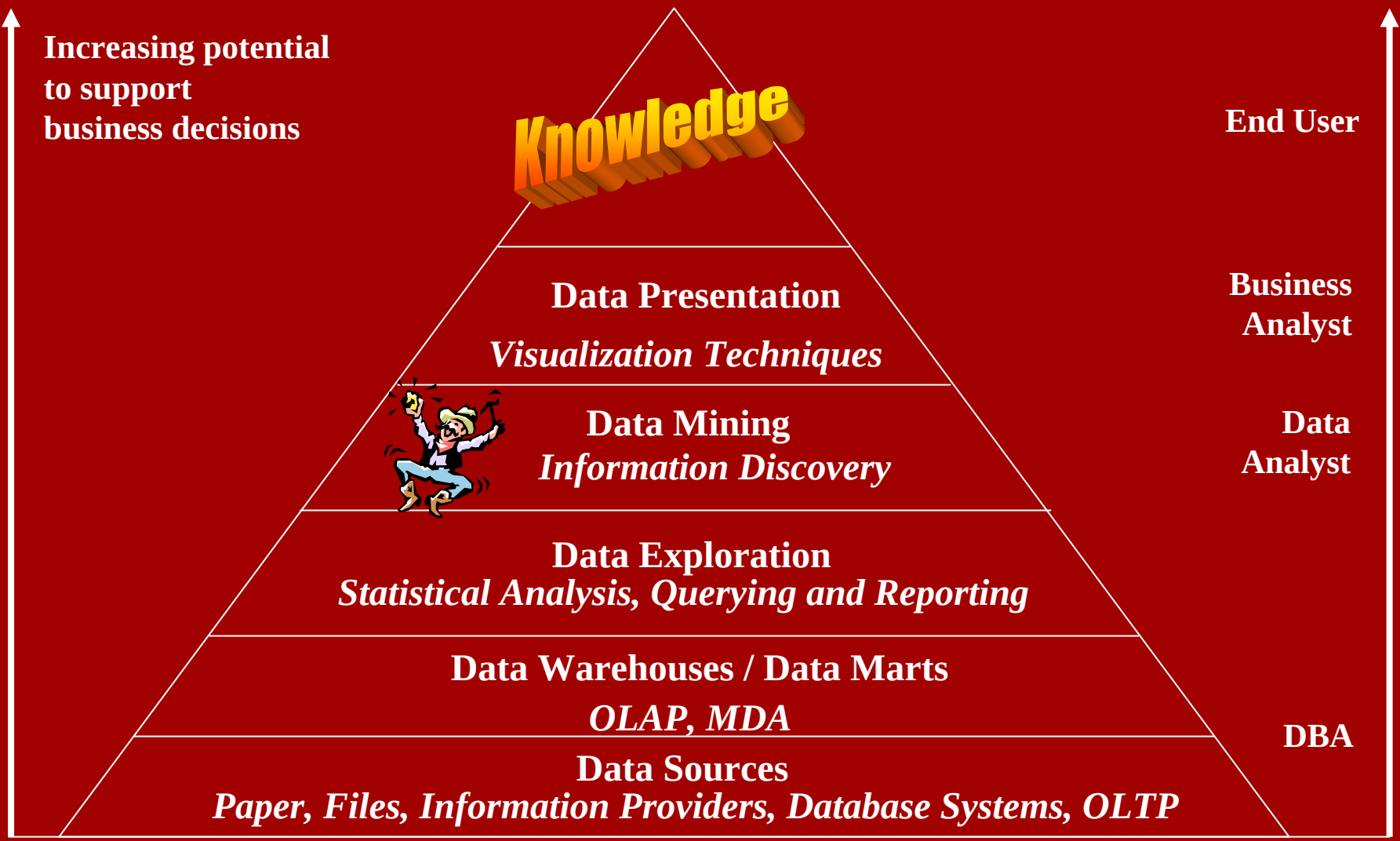
Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996





We are drowning in data, but starving for knowledge!

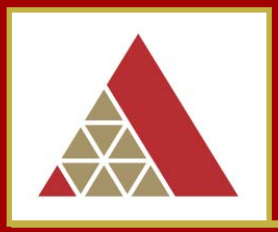
Adopted from Jiawei Han, UIUC



A Caveman's Wisdom

Source: The Futurist, December 1982, Tom Chalkley

Information is data that has been given context and is meaningful to a particular audience
Knowledge is information with a broader, general understanding of the subject matter
Wisdom is knowledge embedded with the reflection or application of that knowledge



Introduction

Data mining is concerned with the extraction of useful knowledge from various types of data.

Medical data mining has been a popular data mining topic of late. Compared with other data mining areas, medical data mining has some unique characteristics. Since medical files are related to human subjects, privacy concern is taken more seriously than other data mining tasks.

This paper applied data separation-based techniques to preserve privacy in classification of medical data.

We implement these two approaches using medical datasets from UCI KDD archive and report the experimental results.

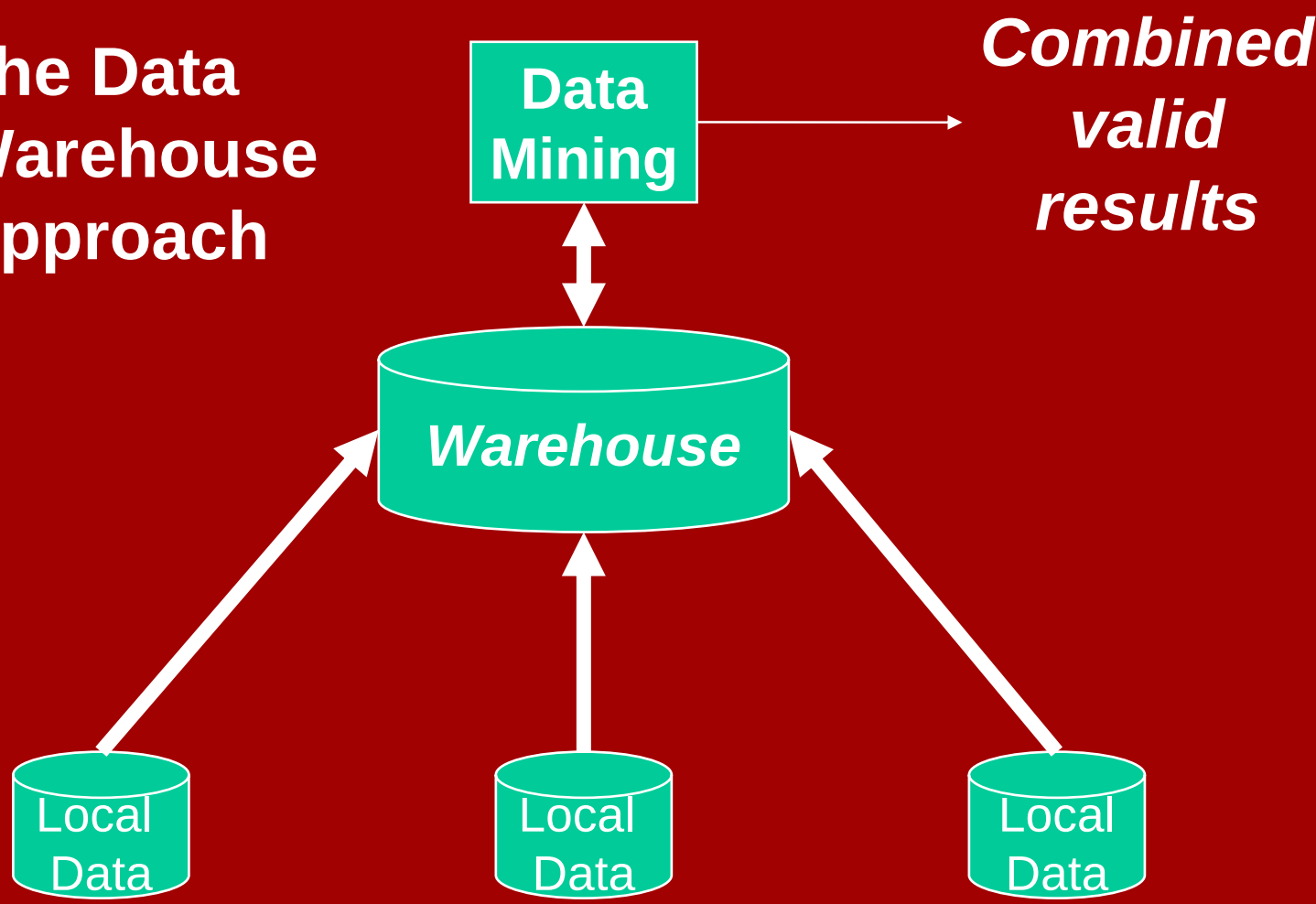
Medical Datasets in this research

Wisconsin prognostic breast cancer dataset has 699 records and 9 variables. These records belong to either benign or malignant class. We create nine different sub-datasets by removing one variable at a time.

The heart-disease dataset has 797 records and 13 variables. These records belong to either heart-disease or normal class. This data was collected from Cleveland Clinic Foundation, Hungarian Institute of Cardiology, University Hospital of Zurich, and Long Beach V.A. Medical Center.

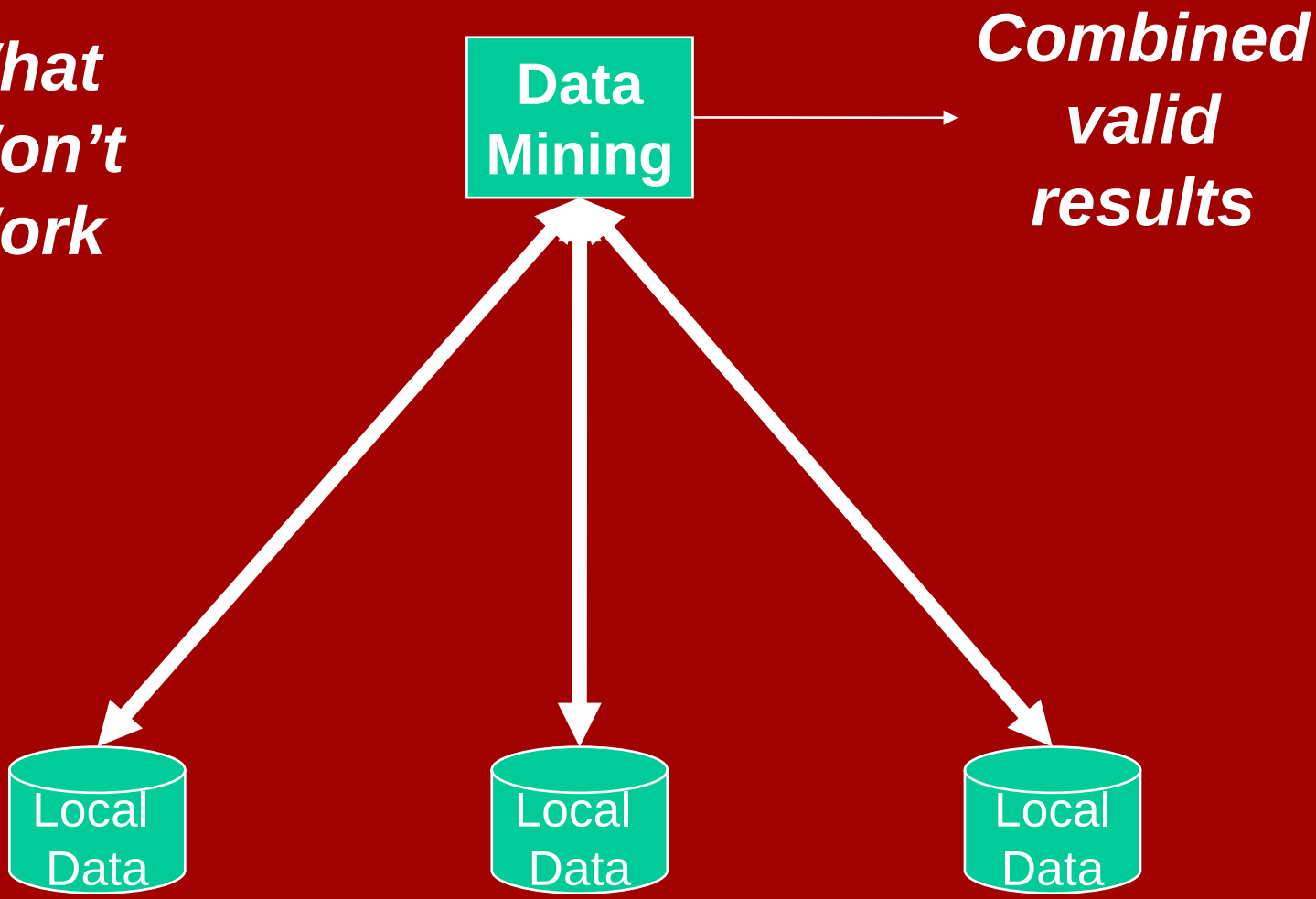
Distributed Data Mining: The "Standard" Method

The Data Warehouse Approach

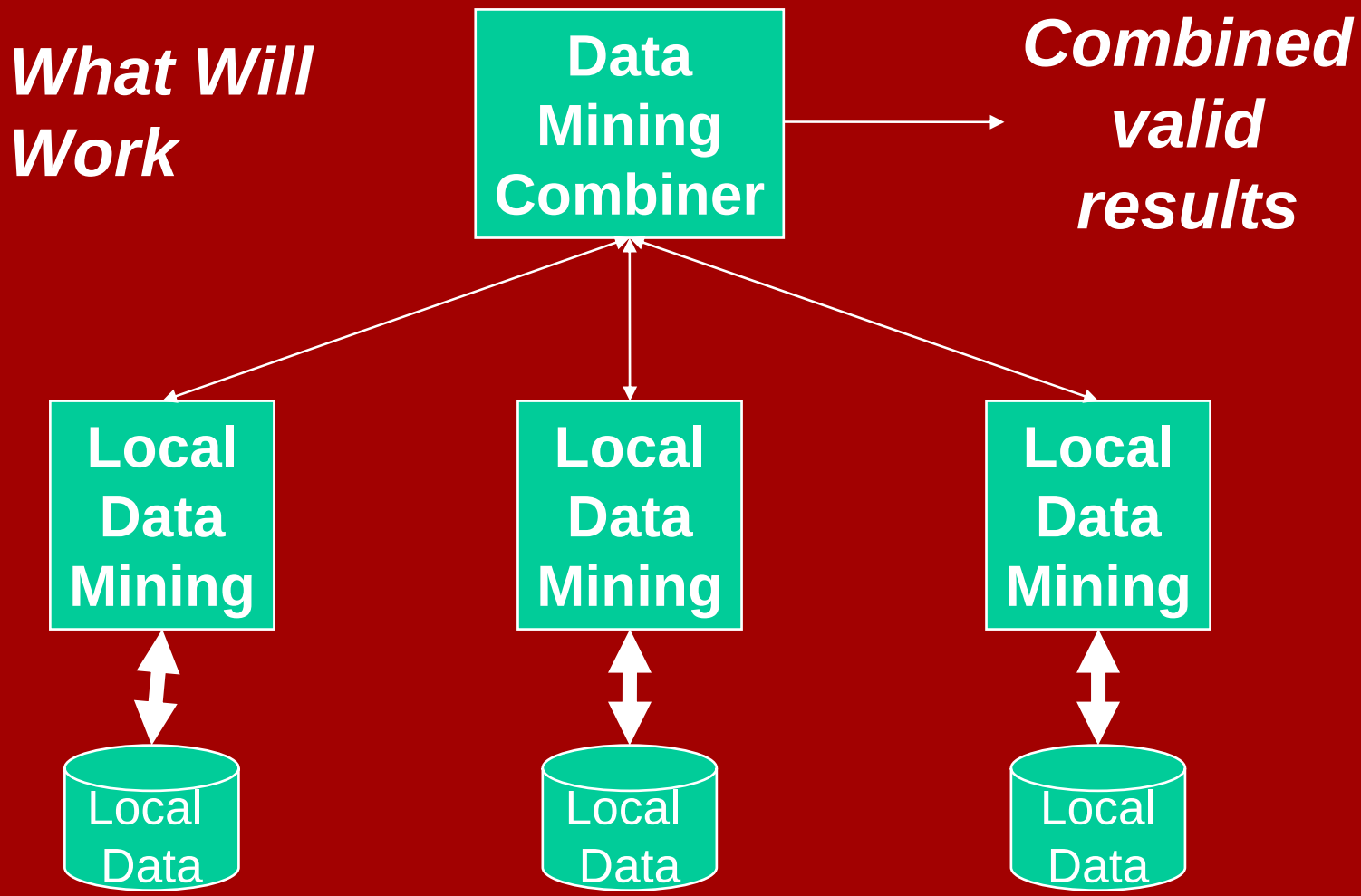


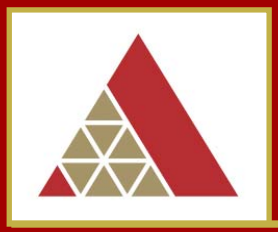
Private Distributed Mining: What is it?

What Won't Work



Private Distributed Mining: What is it?





Privacy-preserving Data Mining

– Data obfuscation

- Nobody sees the real data so it hide the protected information
- Approaches– Randomly modify data– Swap values between records– Controlled modification of data to hide secrets

– Summarization

- Only the needed facts are exposed
- Approaches– Overall collection statistics– Limited query functionality

– Data separation

- Data remains with trusted parties
- Approaches– Vertical separation– Horizontal separation

Privacy-preserving Classification (Vertical) Process

Input: The Medical dataset M , each of the medical records has m attributes

Output: Average classification accuracies for benign and malignant of the dataset in 10-fold cross-validation; scores for all records; decision trees ensemble.

Step 1 Generate m subsets with one different attribute removed from M at each time.

Step 2 Training each subset with See5 with adaptive boosting and 10-fold cross validation to get m decision trees.

Step 3 Ensemble the final decision function D, via majority vote of the m decision trees from step 2.

Step 4 Classify M by the final decision function.

END

Privacy-preserving Classification (Horizontal) Process

Input: The Medical dataset from r different sources, M1, M2, Mr, each of the medical records has m attributes

Output: Average classification accuracies for Normal and Heart-disease of the dataset in 10-fold cross-validation; scores for all records; decision trees ensemble.

Step 1 Training r datasets with See5 with adaptive boosting and 10-fold cross validation to get r decision trees .

Step 2 Ensemble the final decision function D, via majority vote of the r decision trees from step 1.

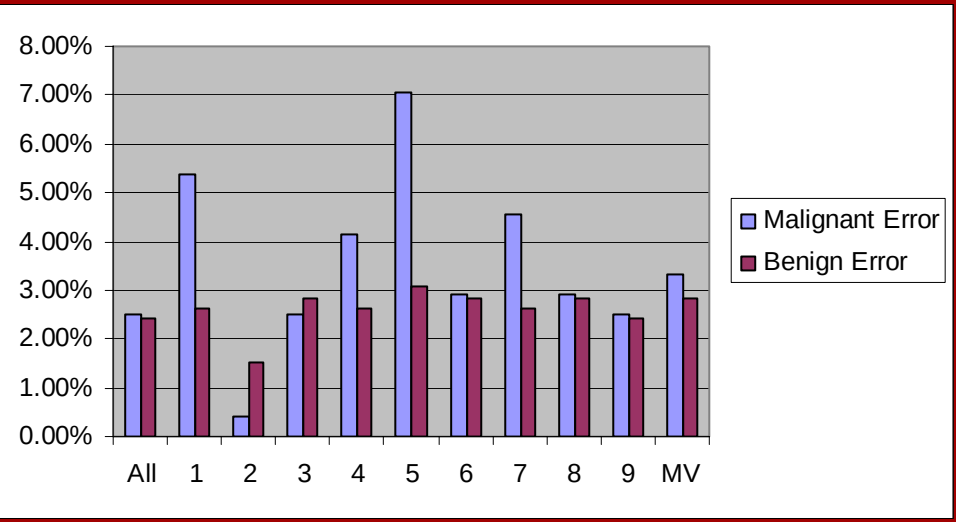
Step 3 Classify all r datasets by the final decision function.

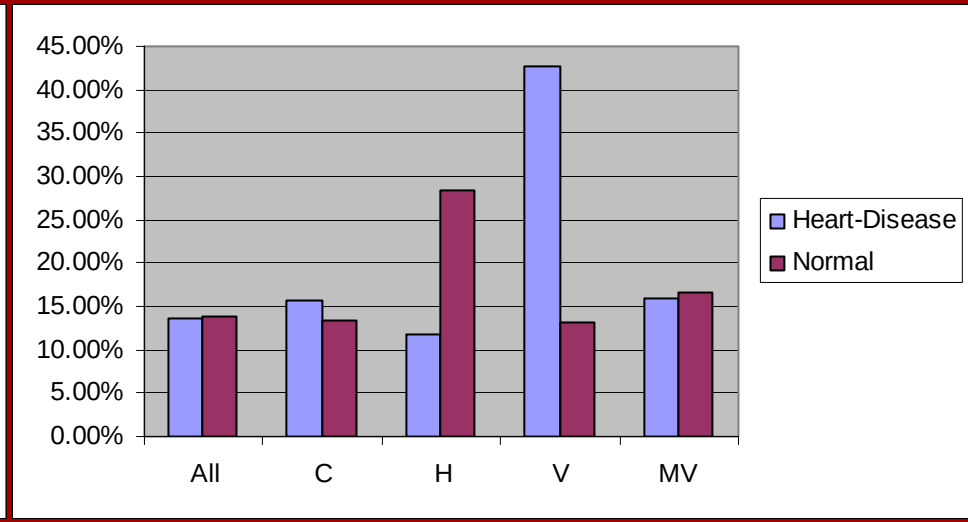
END

Privacy-preserving Data Mining

Vertical partition - each site uses a portion of the attributes to compute its results and the distributed results are assembled at a central trusted party using majority-vote ensemble method.

Horizontal partition - data are distributed among several sites. Each site computes its own data and a central party is responsible to integrate these results.





Conclusion

Privacy-preserving is an important issue in medical data mining. This paper investigates data separation techniques in medical data classification. The experiments demonstrate that data separation techniques can not only protect data privacy, but also increase classification accuracy sometimes (e.g., horizontally partitioned data).

University of Nebraska Omaha

University of Nebraska's

The Peter Kiewit Institute

Comment & Question

谢谢大家！ Thank You!