Security and Privacy Topics in Big Data

CISC 6640 PRIVACY AND SECURITY IN BIG DATA

Instructor:
Md Zakirul Alam Bhuiyan
Assistant Professor
Department of Computer and Information Sciences
Fordham University

Quiz

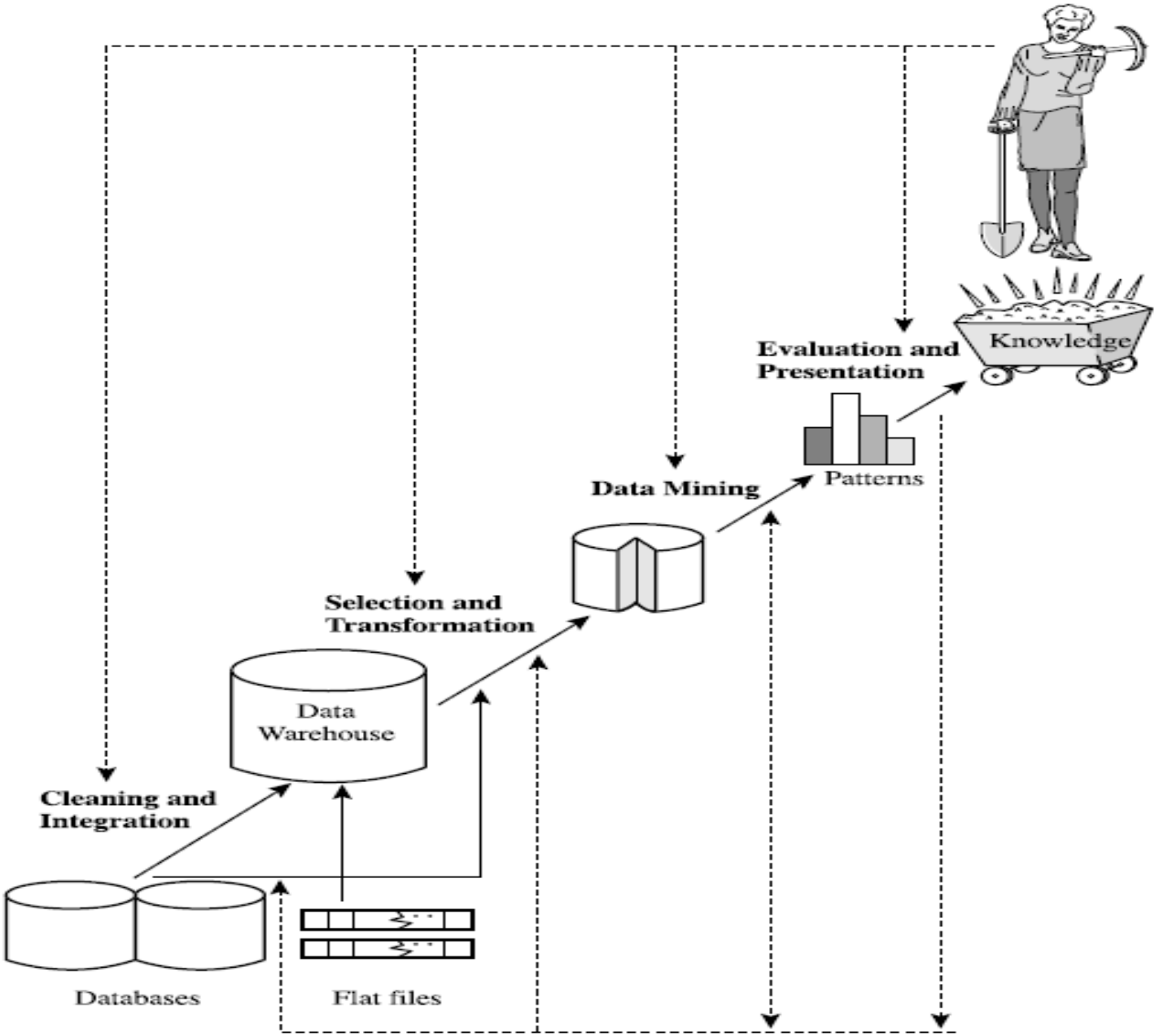
- Why is the direct use of a block cypher inadvisable?
- How are the block cyphers processed?
- What are the cryptographic modes of operation?
- What mode of operation maintains a link from previous block to the current block?
- Which mode of operation seems to be the best among all?
- How can the cryptographic attacks be made?
- Can you say some modes of cryptographic attacks?
- Can you say some chosen plain text attacks?

What We Are Going to Learn…

- Big Data Privacy and Data Mining
- Secure Multiparty Computation
- Security Proof Tools
- Privacy Preserving Data Mining Toolkit
- Trustworthy Decision-Making
- Privacy Preserving Public Auditing
- Data Centric Security

Big Data Privacy and Data Mining

- Data Mining
  - **The process of discovering interesting patterns and knowledge from large amounts of data**
  - **Applications**
    - Business intelligence, Web search, scientific discovery, digital libraries, etc.
  - **The term ``data mining'' is often treated as a synonym for another term ``knowledge discovery from data'' (KDD) which highlights the goal of the mining process.**


Evaluation and Presentation
Knowledge
Data Mining
Patterns
Selection and Transformation
Data Warehouse
Cleaning and Integration
Databases
Flat files

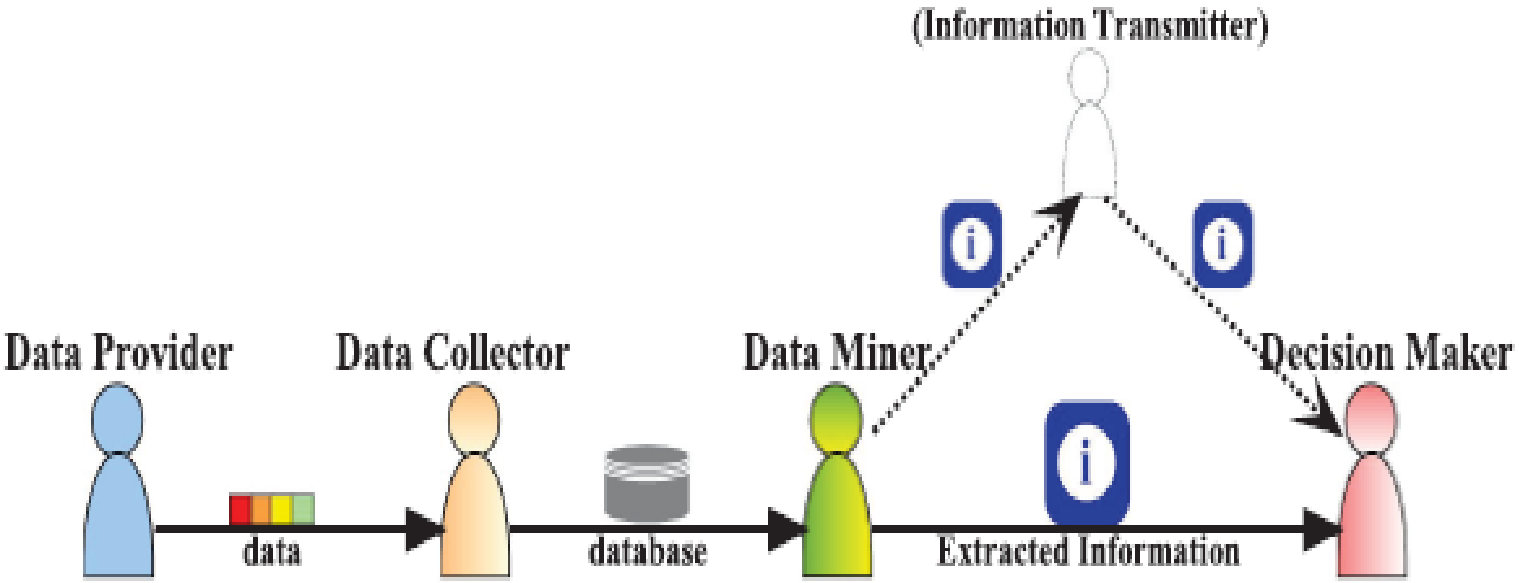
Big Data Privacy and Data Mining

- Individual's privacy may be violated due to the unauthorized access to personal data.
  - To deal with the privacy issues in data mining, a sub-field of data mining, referred to as privacy preserving data mining (PPDM) .
  - The aim of PPDM is to safeguard sensitive information from unsanctioned disclosure, and preserve the utility of the data.

Big Data Privacy and Data Mining

- The 4 type of users in Data Mining process-
  - Data Provider
  - Data Collector
  - Data Miner
  - Decision Maker

Big Data Privacy and Data Mining



1. Data Provider: Security Concern

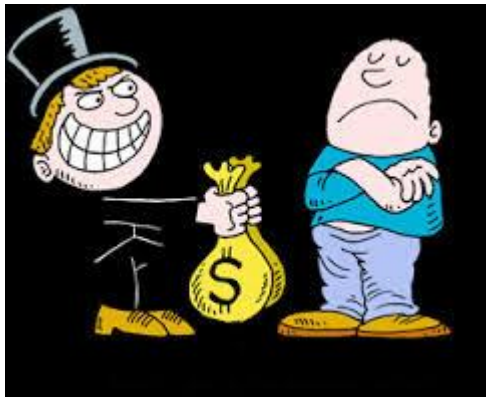
- The major concern of a data provider is whether he can control the sensitivity of the data he provides to others.
- On one hand,
  - The provider should be able to make his very private data, inaccessible to the data collector.
- On the other hand,
  - If the provider has to provide some data to the data collector, he wants to hide his sensitive information as much as possible and get enough compensations for the possible loss in privacy.

1. Data Provider: Security Concern

- **Approaches to privacy protection**
  - **Limit the access**
    - Security tools that are developed for internet environment to protect data
    - Anti-Tracking extensions.
      - Popular anti-tracking extensions include Disconnect , Do Not Track Me ,Ghostery etc
    - Advertisement and script blockers
      - Example tools include AdBlock Plus, NoScript, FlashBlock, etc.
    - Encryption tools-MailCloack and TorChat

1. Data Provider: Security Concern

- Approaches to privacy protection
  - **Trade privacy for benefit**
    - The data provider maybe willing to hand over some of his private data in exchange for certain benefit.
    - Such as better services or monetary rewards. The data provider needs to know how to negotiate with the data collector, so that he will get enough compensation for any possible loss in privacy



1. Data Provider: Security Concern

- Approaches to privacy protection
  - **Provide false data**
    - Using ``sockpuppets'' to hide one's true activities
    - Using a fake identity to create phony information
    - Using security tools to mask one's identity

"HEY! LOOK WHAT I CAN DO!!"

2. Data Collector: Security Concern

- The major concern of data collector is to guarantee that
  - the modified data contains no sensitive information but still preserve high utility.

2. Data Collector: Security Concern

- Approaches to privacy protection
  - 1. BASICS OF PPDP
    - PPDP mainly studies anonymization approaches for publishing useful data while preserving privacy.
    - Each record consists of the following 4 types of attributes:
      - Identifier (ID): Name, id, mobile number
      - Quasi-identifier (QID): Attributes that can be linked with external data to re-identify individual records, such as gender, age and zip code.
      - Sensitive Attribute (SA): Attributes that an individual wants to conceal, such as disease and salary.
      - Non-sensitive Attribute (NSA): Attributes other than ID,QID and SA

2. Data Collector: Security Concern

- Anonymization operations
  - 1. BASICS OF PPDP
    - **Generalization**. This operation replaces some values with a parent value in the taxonomy of an attribute.
    - **Suppression**. This operation replaces some values with a special value (e.g. a asterisk `*'), indicating that the replaced values are not disclosed.
    - **Anatomization**. This operation does de-associates the relationship between the two.
    - **Permutation**. This operation de-associates the relationship of attribute by partitioning a set of data records into groups and shuffling their sensitive values within each group.
    - **Perturbation**. This operation replaces the original data values with some synthetic data values.

2. Data Collector: Security Concern

- Anonymization operations
  - 1. BASICS OF PPDP

Age	Sex	Zipcode	Disease
5	Female	12000	HIV
9	Male	14000	dyspepsia
6	Male	18000	dyspepsia
8	Male	19000	bronchitis
12	Female	21000	HIV
15	Female	22000	cancer
17	Female	26000	pneumonia
19	Male	27000	gastritis
21	Female	33000	flu
24	Female	37000	pneumonia

Age	Sex	Zipcode	Disease
[1, 10]	People	1****	HIV
[1, 10]	People	1****	dyspepsia
[1, 10]	People	1****	dyspepsia
[1, 10]	People	1****	bronchitis
[11, 20]	People	2****	HIV
[11, 20]	People	2****	cancer
[11, 20]	People	2****	pneumonia
[11, 20]	People	2****	gastritis
[21, 60]	People	3****	flu
[21, 60]	People	3****	pneumonia

2. Data Collector: Security Concern

- Anonymization operations
  - 2.PRIVACY PRESERVING PUBLISHING OF SOCIAL NETWORK DATA
    - PPDP in the context of social networks mainly deals with anonymizing graph data
    - Which is much more challenging than anonymizing relational table data.

2. Data Collector: Security Concern

- Anonymization operations
  - 3. ATTACK MODEL

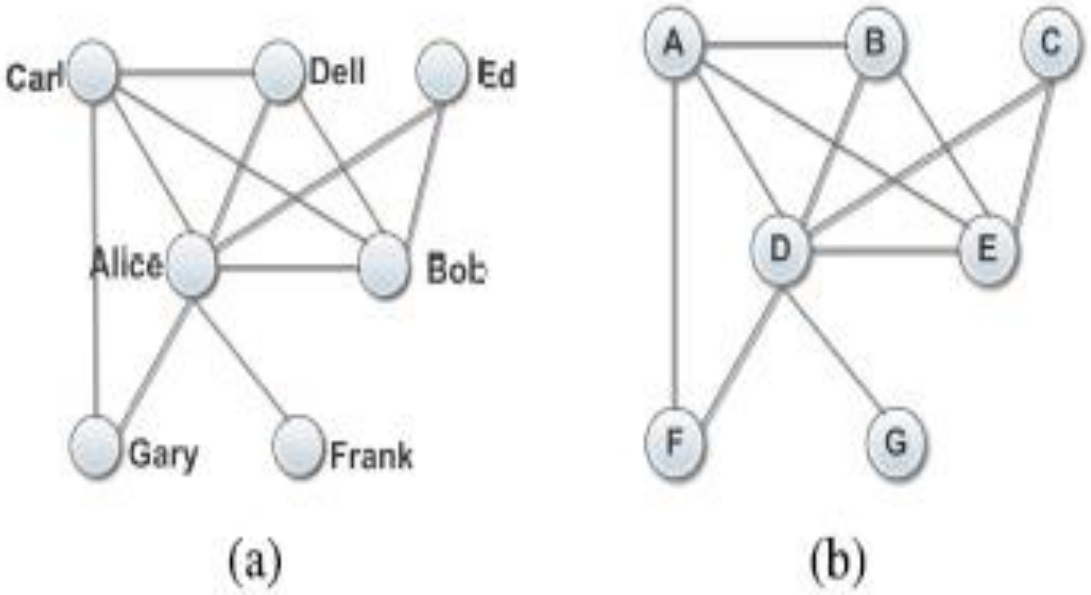
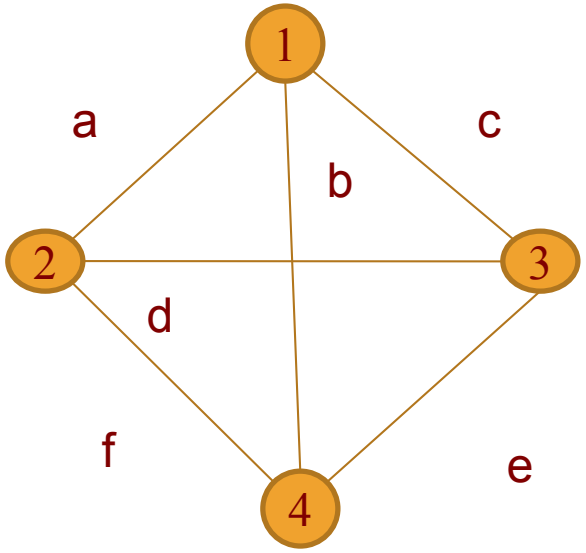


FIGURE 4. Example of mutual friend attack: (a) original network; (b) naïve anonymized network.

2. Data Collector: Security Concern

PRIVACY MODELS

- **If a network satisfies k-NMF anonymity then for each edge e, there will be at least k - 1 other edges with the same number of mutual friends as e. It can be guaranteed that the probability of an edge being identified is not greater than 1/k.**



a= 2 mutual friends
b=2 mutual friends
c=2 mutual friends
d=2 mutual friends
e=2 mutual friends
f=2 mutual friends

So 6-NMF

3. Data Miner: Security Concern

- The primary concern of data miner is how to prevent sensitive information from appearing in the mining results.

- To perform a privacy-preserving data mining, the data miner usually needs to modify the data he got from the data collector.

3. Data Miner: Security Concern

- Approaches

FIGURE 12. Data distribution. (a) centralized data. (b) horizontally partitioned data. (c) vertically partitioned data.

3. Data Miner: Security Concern

- Approaches
  - 1. PRIVACY PRESERVING ASSOCIATION RULE MINING
    - Various kinds of approaches have been proposed to perform association rule hiding .
      - Heuristic distortion approaches
      - Heuristic blocking approaches
      - Probabilistic distortion approaches
      - Exact database distortion approaches
      - Reconstruction-based approaches

3. Data Miner: Security Concern

- Approaches
  - 2. PRIVACY PRESERVING CLASSIFICATION
    - Classification is a form of data analysis that extracts models describing important data classes
    - To realize privacy-preserving decision tree mining,
      - Dowd et al. proposed a data perturbation technique based on random substitutions.
      - Brickell and Shmatikov present a cryptographically secure protocol for privacy-preserving construction of decision trees.

4. Decision maker: Security Concern

- The privacy concerns of the decision maker are following:
  - how to prevent unwanted disclosure of sensitive mining results
  - how to evaluate the credibility of the received mining results

4. Decision Maker: Security Concern

- Approaches
  - **Legal measures**
    - For example, making a contract with the data miner to forbid the miner from disclosing the mining results to a third party
    - The decision maker can utilize methodologies from data provenance, credibility analysis of web information, or other related research fields

Data Provenance

- The information that helps determine the derivation history of the data, starting from the original source
- Two kinds of information
  - the ancestral data from which current data evolved
  - the transformations applied to ancestral data that helped to produce current data.
- With such information, people can better understand the data and judge the credibility of the data.

Data Provenance

- Web information credibility
  - **5 ways Internet users to differentiate false information from the truth:**
    1. **Authority**: the real author of false information is usually unclear.
    2. **Accuracy**: false information does not contain accurate data
    3. **Objectivity**: false information is often prejudicial.
    4. **Currency**: for false information, the data about its source, time and place of its origin is incomplete, out of date, or missing.
    5. **Coverage**: false information usually contains no effective links to other information online.

Secure Multiparty Computation

Secure Multiparty Computation

- Applications for Privacy Preserving Distributed Data Mining

Distributed Data Mining

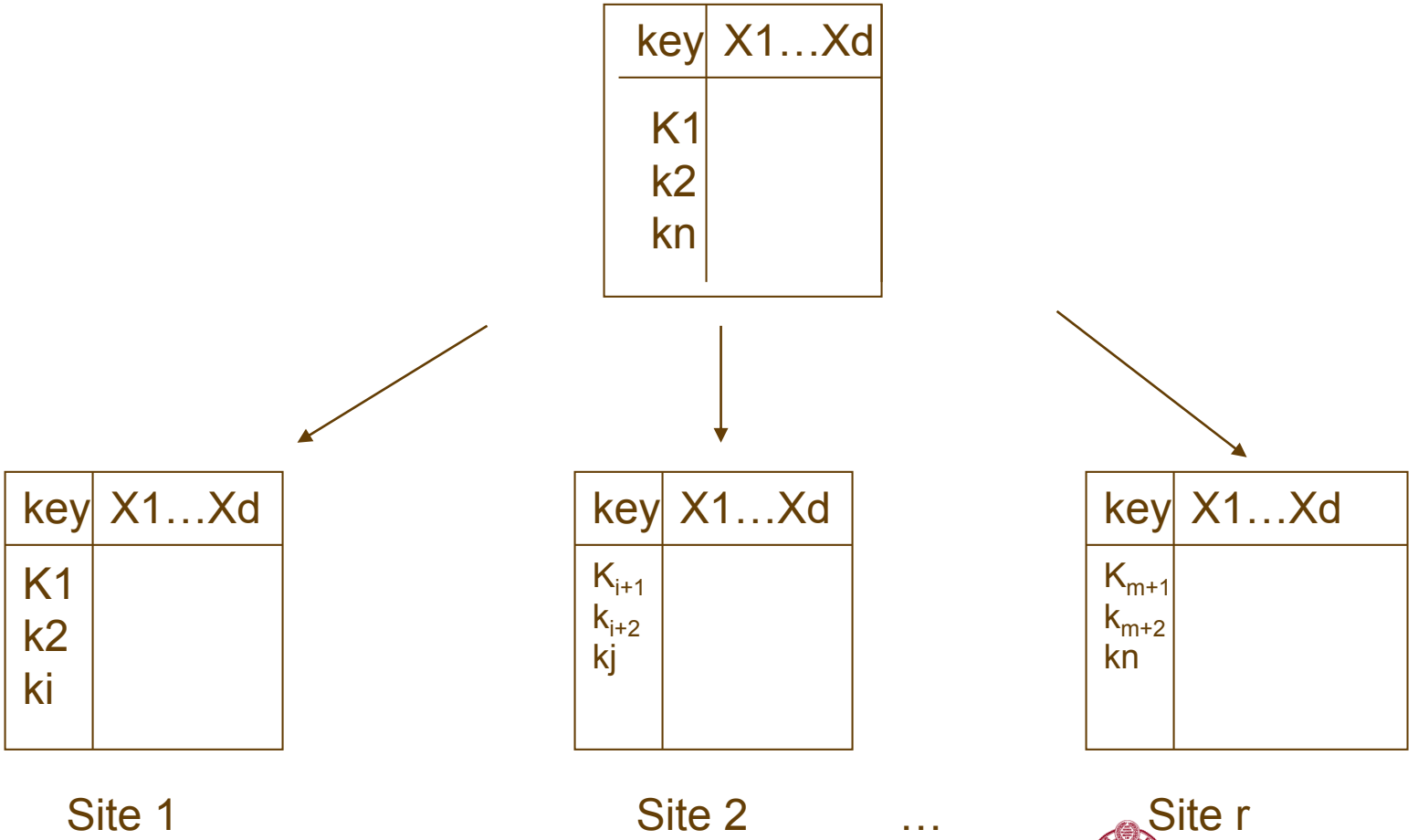
- **Government / public agencies. Example:**
  - The Centers for Disease Control want to identify disease outbreaks
  - Insurance companies have data on disease incidents, seriousness, patient background, etc.
  - But can/should they release this information?
- **Industry Collaborations/Trade Groups. Example:**
  - An industry trade group may want to identify best practices to help members
  - But some practices are trade secrets
  - How do we provide "commodity" results to all (Manufacturing using chemical supplies from supplier X have high failure rates), while still preserving secrets (manufacturing process Y gives low failure rates)?

Classification

- Data partition
  - Horizontally partitioned
  - Vertically partitioned
- Techniques
  - Data perturbation
  - Secure Multi-party Computation protocols
- Mining applications
  - Decision tree
  - Bayes classifier
  - …

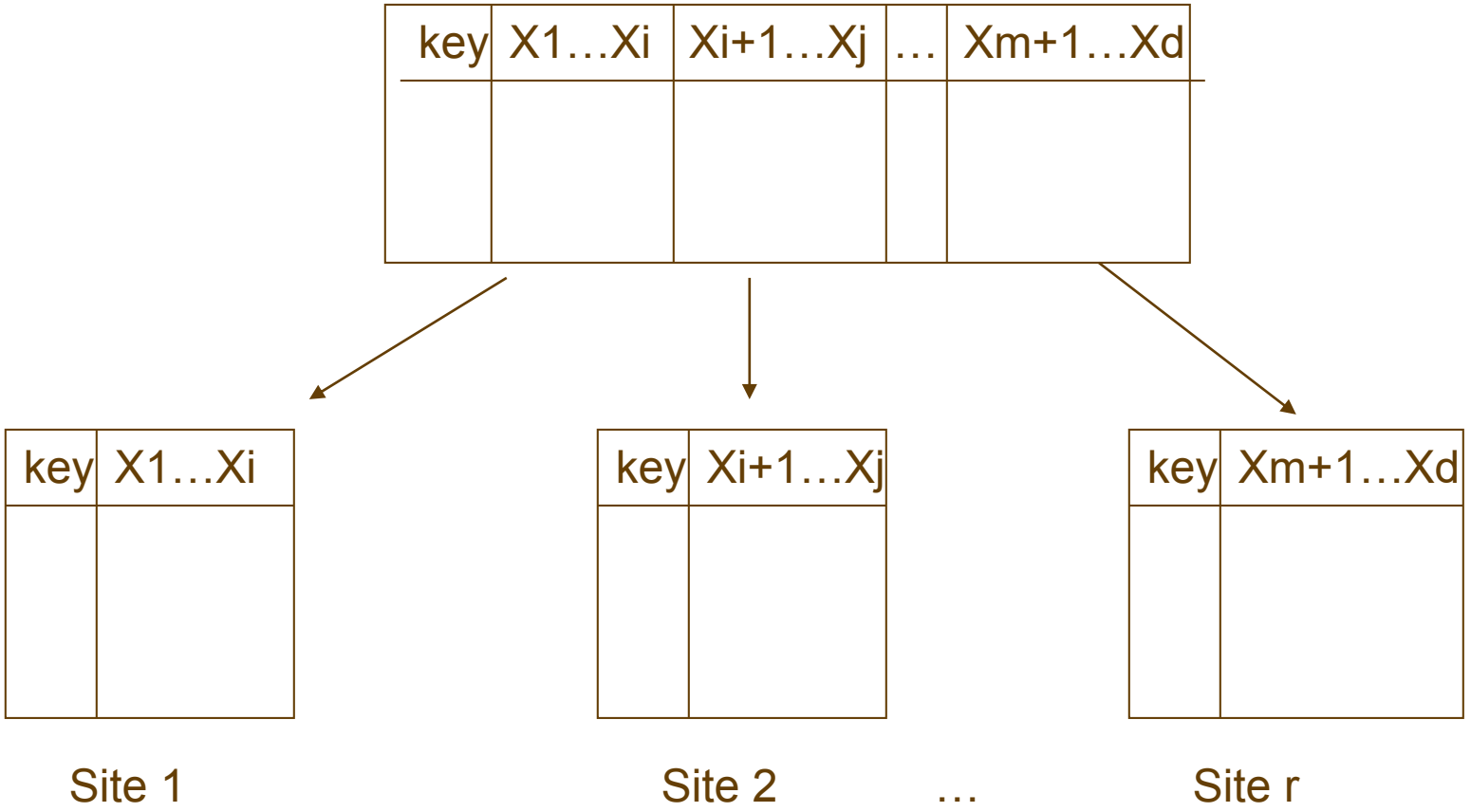
Horizontally Partitioned Data

- **Data can be unioned to create the complete set**



Vertically Partitioned Data

- Data can be joined to create the complete set



Secure Multiparty Computation

- **Goal**: Compute function when each party has some of the inputs
- Secure
  - Can be simulated by ideal model - nobody knows anything but their own input and the results
- Formally
  - Polynomial time S such that $\{S(x,f(x,y))\} \equiv \{View(x,y)\}$
- Semi-Honest model
  - Follow protocol, but remember intermediate exchanges
- Malicious: "cheat" to find something out

Secure Multiparty Computation

- Basic cryptographic tools
  - Oblivious transfer
  - Oblivious circuit evaluation
- Yao's Millionaire's problem (Yao '86)
  - Secure computation possible if function can be represented as a circuit
- Works for multiple parties as well (Goldreich, Micali, and Wigderson '87)

Secure Multiparty Computation

- Why aren't we done?
  - **Secure Multiparty Computation is possible**
    - But is it practical?
  - **Circuit evaluation: Build a circuit that represents the computation**
    - For all possible inputs
    - Impossibly large for typical data mining tasks
  - **The next step: Efficient techniques for specialized tasks and computations**

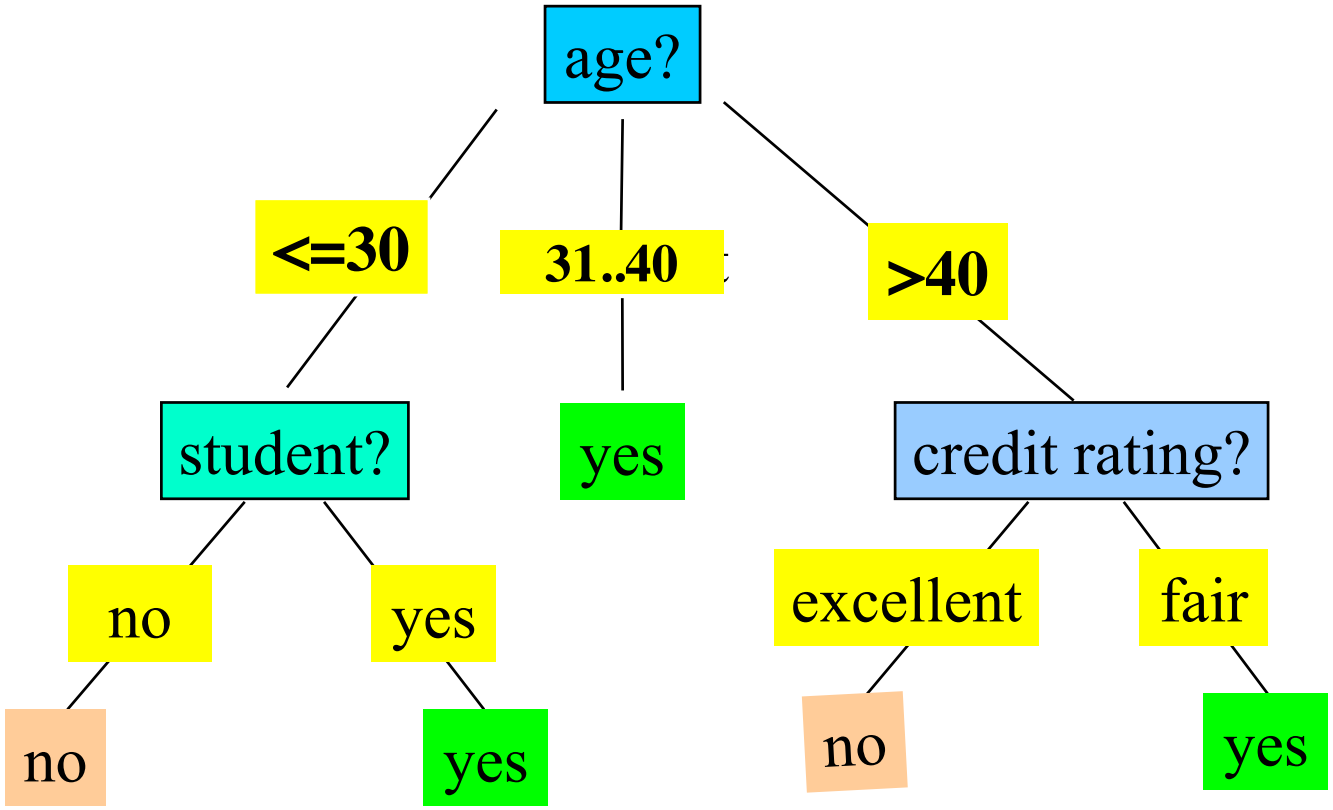
Secure Multiparty Computation

- Privacy preserving two-party decision tree mining (Lindell & Pinkas '00)
- Privacy preserving distributed data mining toolkit (Clifton '02)
  - Secure sum
  - Secure union
  - Association Rule Mining on Horizontally Partitioned Data (Kantarcioglu '04)

Decision Tree Construction

- Two-party horizontal partitioning
  - Each site has same schema
  - Attribute set known
  - Individual entities private
- Learn a decision tree classifier
  - ID3: Iterative Dichotomiser 3
- Essentially ID3 meeting Secure Multiparty Computation Definitions
  - Semi-honest model

A Decision Tree for "buys_computer"



- **Greedy algorithm - tree is constructed in a top-down recursive divide-and-conquer manner**
  - At start, all the training examples are at the root
  - A test attribute is selected that "best" separate the data into partitions - information gain
  - Samples are partitioned recursively based on selected attributes
- **Conditions for stopping partitioning**
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
  - There are no samples left

Privacy Preserving ID3

- **Input:**
  - R – the set of attributes, C – the class attribute
  - T – the set of transactions
  - Step 1: If R is empty, return a leaf-node with the class value with the most transactions in T
  - Set of attributes is public
  - Both know if R is empty
  - Use secure protocol for majority voting
    - Yao's protocol
    - Inputs ($|T_1(c_1)|,\ldots,|T_1(c_L)|$), ($|T_2(c_1)|,\ldots,|T_2(c_L)|$)
    - Output i where $|T_1(c_i)|+|T_2(c_i)|$ is largest

- **Step 2: If T consists of transactions which have all the same value c for the class attribute, return a leaf node with the value c**
  - Represent having more than one class (in the transaction set), by a fixed symbol different from ci,
  - Force the parties to input either this fixed symbol or ci
  - Check equality to decide if at leaf node for class ci
  - Various approaches for equality checking
    - Yao'86
    - Fagin, Naor '96
    - Naor, Pinkas '01

- **Step 3:(a) Determine the attribute that best classifies the transactions in T, let it be A**
  - Essentially done by securely computing x*(ln x) where x is the sum of values from the two parties
  - P1 and P2 , i.e., x1 and x2, respectively
  - Step 3: (b,c) Recursively call ID3 for the remaining attributes on the transaction sets T(a1),…,T(am) where a1,…, am are the values of the attribute A
    - Since the results of 3(a) and the attribute values are public, both parties can individually partition the database and prepare their inputs for the recursive calls

Security Proof Tools

- **Real/ideal model: the real model can be simulated in the ideal model**
  - **Key idea**
    - Show that whatever can be computed by a party participating in the protocol can be computed based on its input and output only
  - **$\exists$ polynomial time S such that {S(x,f(x,y))} ≡ {View(x,y)}**

- **Composition theorem**
  - **If a protocol is secure in the hybrid model where the protocol uses a trusted party that computes the (sub) functionalities, and we replace the calls to the trusted party by calls to secure protocols, then the resulting protocol is secure**
  - **Prove that component protocols are secure, then prove that the combined protocol is secure**

Privacy Preserving Data Mining Toolkit

- Many different data mining techniques often perform similar computations at various stages
  - e.g., computing sum, counting the number of items
- Toolkit
  - simple computations – sum, union, intersection …
  - assemble them to solve specific mining tasks – association rule mining, bayes classifier, …
- The protocols may not be truly secure but more efficient than traditional SMC methods

Privacy Preserving Data Mining Toolkit

- **Toolkit**
  - **Secure functions**
    - Secure sum
    - Secure union
    - …
  - **Applications**
    - Association rule mining for horizontally partitioned data
    - …

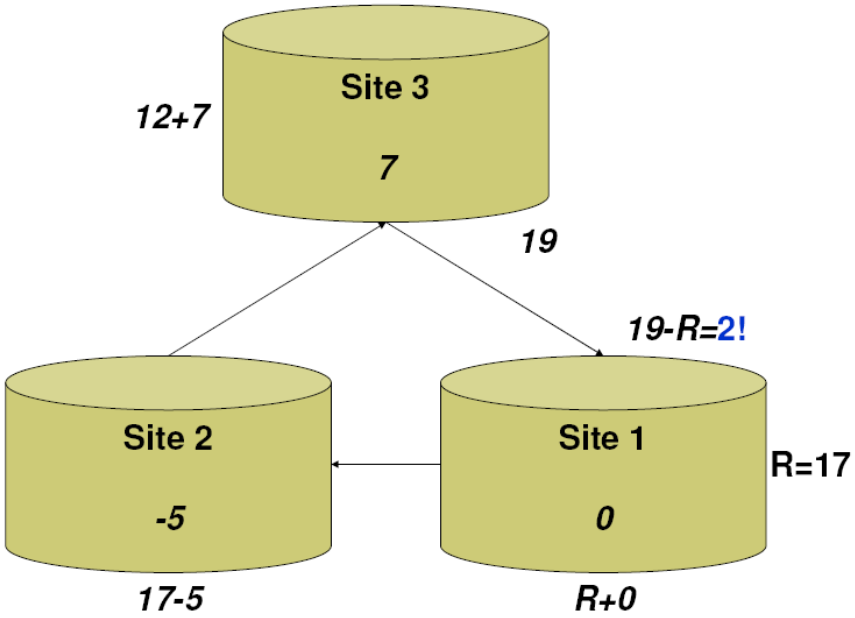
Privacy Preserving Data Mining Toolkit

- **Toolkit**
  - **Secure functions**
    - Secure sum

- Does not reveal the real number

Is it secure?

- Site can collude!
- Each site can divide the number into shares, and run the algorithm multiple times with permutated nodes



Privacy Preserving Data Mining Toolkit

Secure Union

- **Commutative encryption**
- **For any set of permutated keys and a message M**

$$E_{K_{i_1}}(\ldots E_{K_{i_n}}(M)\ldots) = E_{K_{j_1}}(\ldots E_{K_{j_n}}(M)\ldots)$$

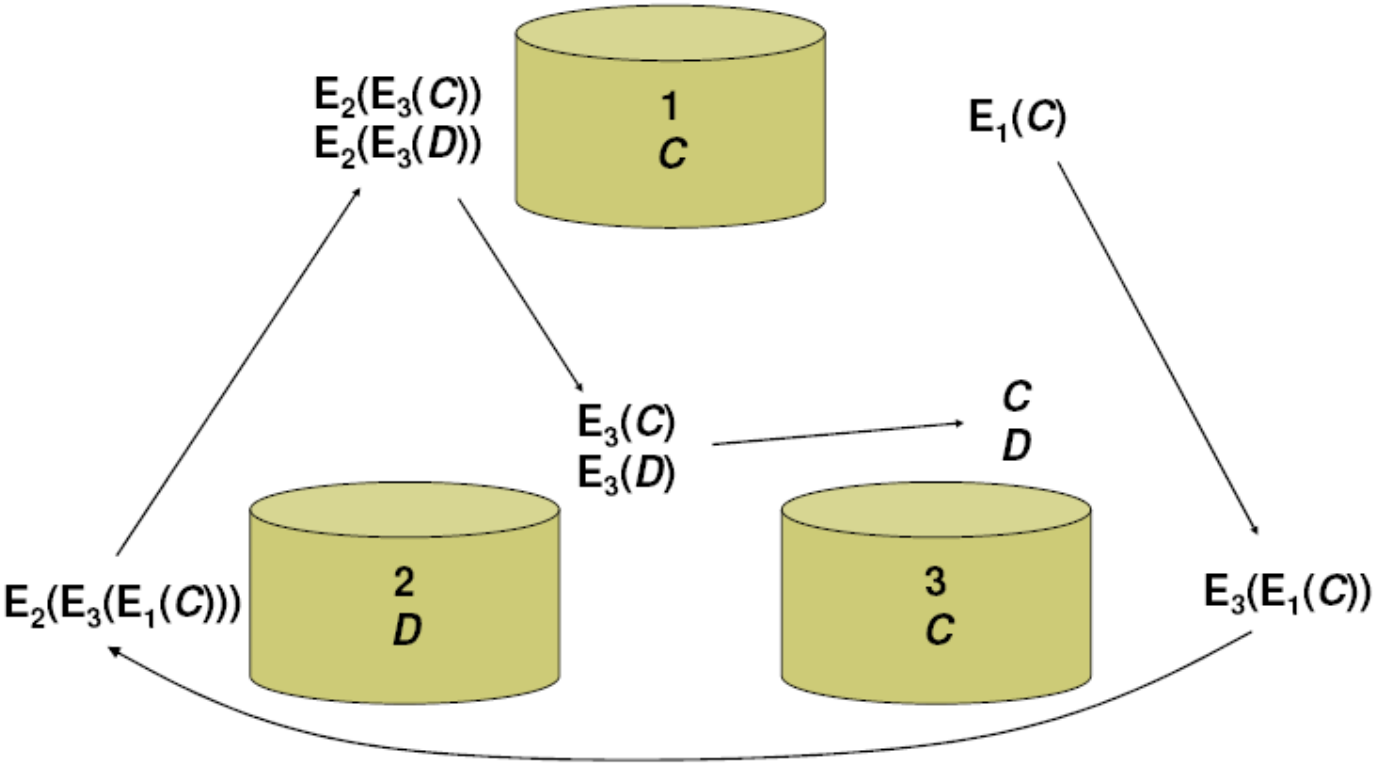
- **For any set of permutated keys and message M1 and M2**

$$Pr(E_{K_{i_1}}(\ldots E_{K_{i_n}}(M_1)\ldots) = E_{K_{j_1}}(\ldots E_{K_{j_n}}(M_2)\ldots)) < \epsilon$$

- **Secure union**
  - Each site encrypt its items and items from other site, remove duplicates, and decrypt

Privacy Preserving Data Mining Toolkit

- Secure Union



Privacy Preserving Data Mining Toolkit

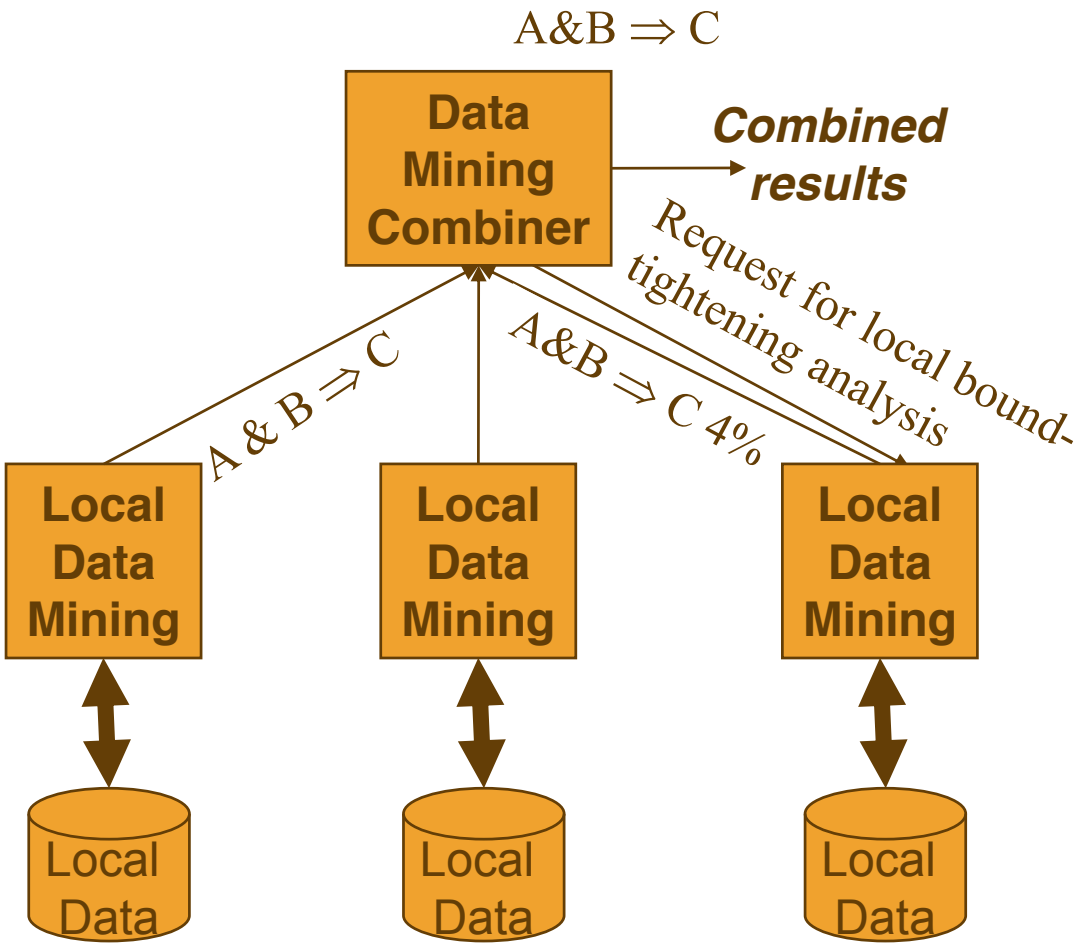
- Secure Union
  - **Does not reveal which item belongs to which site**
    - Is it secure under the definition of secure multi-party computation?
  - **It reveals the number of items that are common in the sites!**
    - Revealing innocuous information leakage allows a more efficient algorithm than a fully secure algorithm

Privacy-Preserving Association Rules Mining

- 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 ⇒ frequent at some site
  - Second phase: Compute frequency of candidates

Privacy-Preserving Association Rules Mining

- Association Rules in Horizontally Partitioned Data



Privacy-Preserving Association Rules Mining

- Association Rule Mining: Horizontal Partitioning
  - What if we do not want to reveal which rule is supported at which site, the support count of each rule, or database sizes?
  - Hospitals want to participate in a medical study
  - But rules only occurring at one hospital may be a result of bad practices

Privacy-Preserving Association Rules Mining

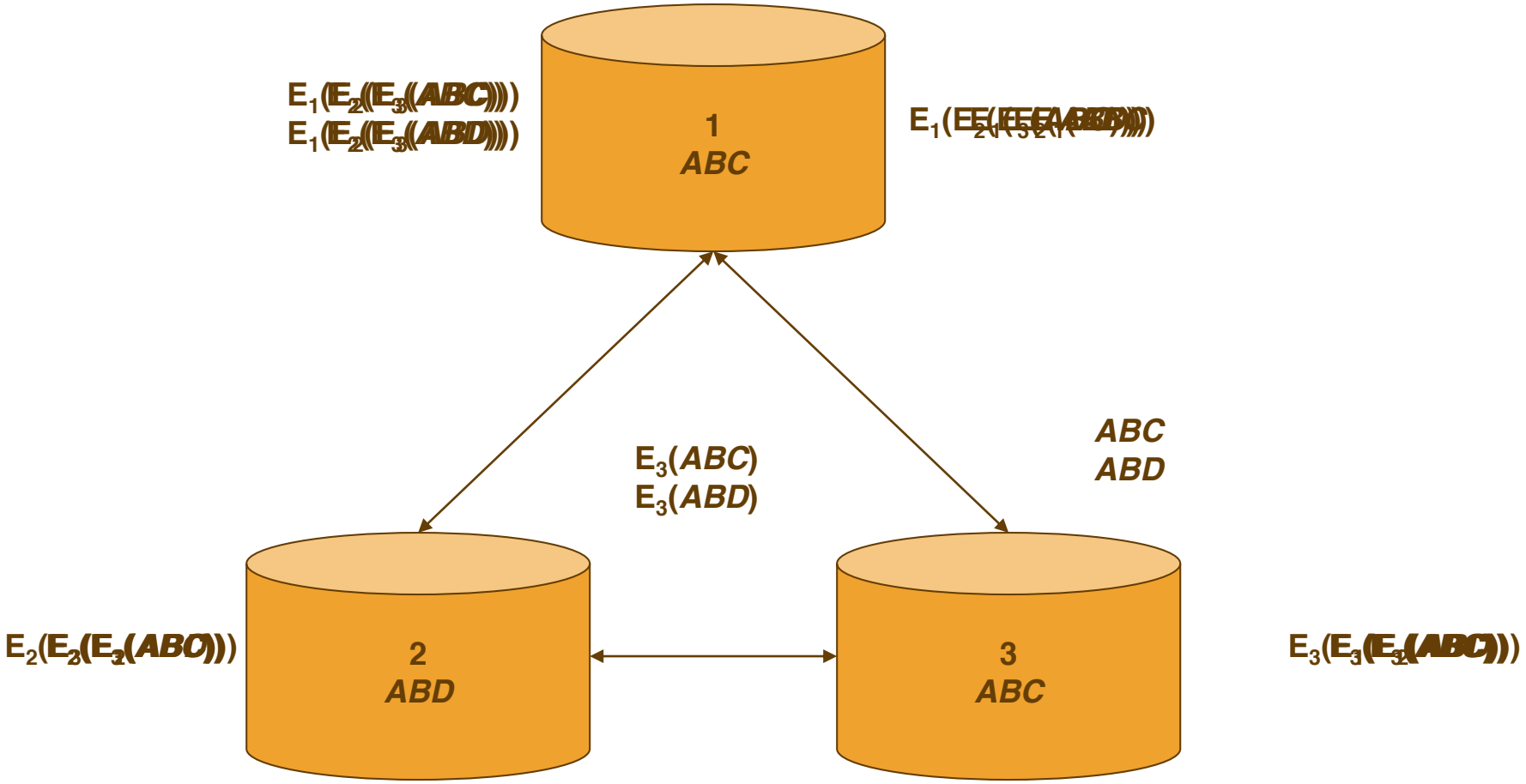
- Privacy-preserving Association rule mining for horizontally partitioned data (Kantarcioglu'04)
  - Find the union of the locally large candidate itemsets securely
  - After the local pruning, compute the globally supported large itemsets securely
  - At the end check the confidence of the potential rules securely

Privacy-Preserving Association Rules Mining

- Securely Computing Candidates
  - Compute local candidate set
  - Using secure union!

Privacy-Preserving Association Rules Mining

- Computing Candidate Sets



Privacy-Preserving Association Rules Mining

- Compute Which Candidates Are Globally Supported?
  - Goal: To check whether

$$X.\sup \geq s * \sum_{i=1}^{n} |DB_i| \qquad (1)$$

$$\sum_{i=1}^{n} X.\sup_i \geq \sum_{i=1}^{n} s^* | DB_i | \qquad (2)$$

$$\sum_{i=1}^{n} (X.\sup_i - s^* | DB_i |) \geq 0 \qquad (3)$$

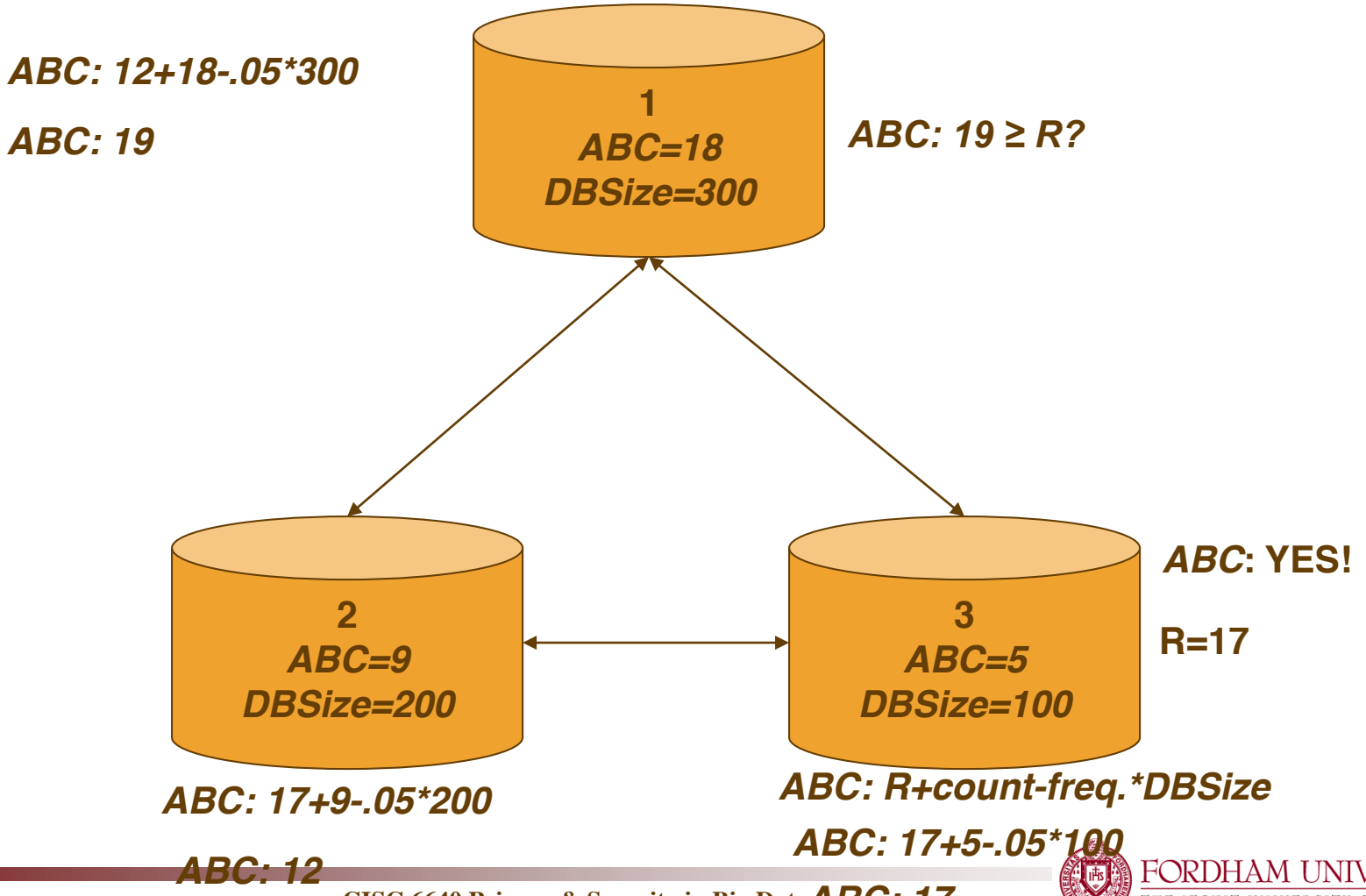
Note that checking inequality (1) is equivalent to checking inequality (3)

Privacy-Preserving Association Rules Mining

- **Compute Which Candidates Are Globally Supported?**
  - **Securely compute sum then check if sum ≥ o Is this a good approach?**
    - Sum is disclosed!
  - **Securely compute Sum - R**
  - **Securely compare sum ≥ R?**
    - Use oblivious transfer

Privacy-Preserving Association Rules Mining

- Computing Frequent: Is ABC ≥ 5%?



Privacy-Preserving Association Rules Mining

Computing Confidence

- **Checking confidence can be done by the previous protocol. Note that checking confidence for X ⇒ Y**

$$\frac{\{X \cup Y\}.\sup}{X.\sup} \geq c \Rightarrow \frac{\sum_{i=1}^{n} XY.\sup_i}{\sum_{i=1}^{n} X.\sup_i} \geq c$$

$$\Rightarrow \sum_{i=1}^{n} (XY.\sup_i - c * X.\sup_i) \geq 0$$

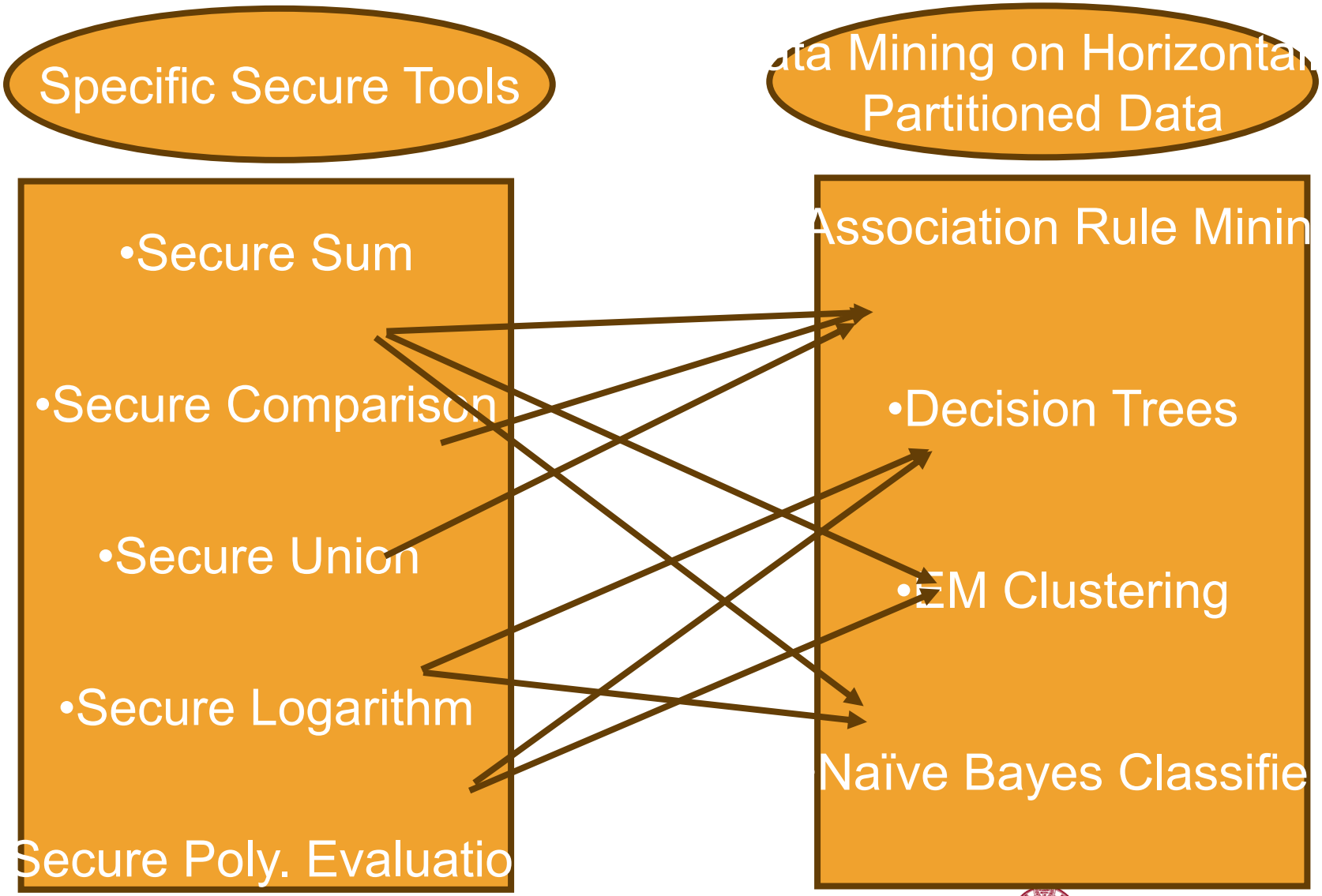
Privacy-Preserving Association Rules Mining

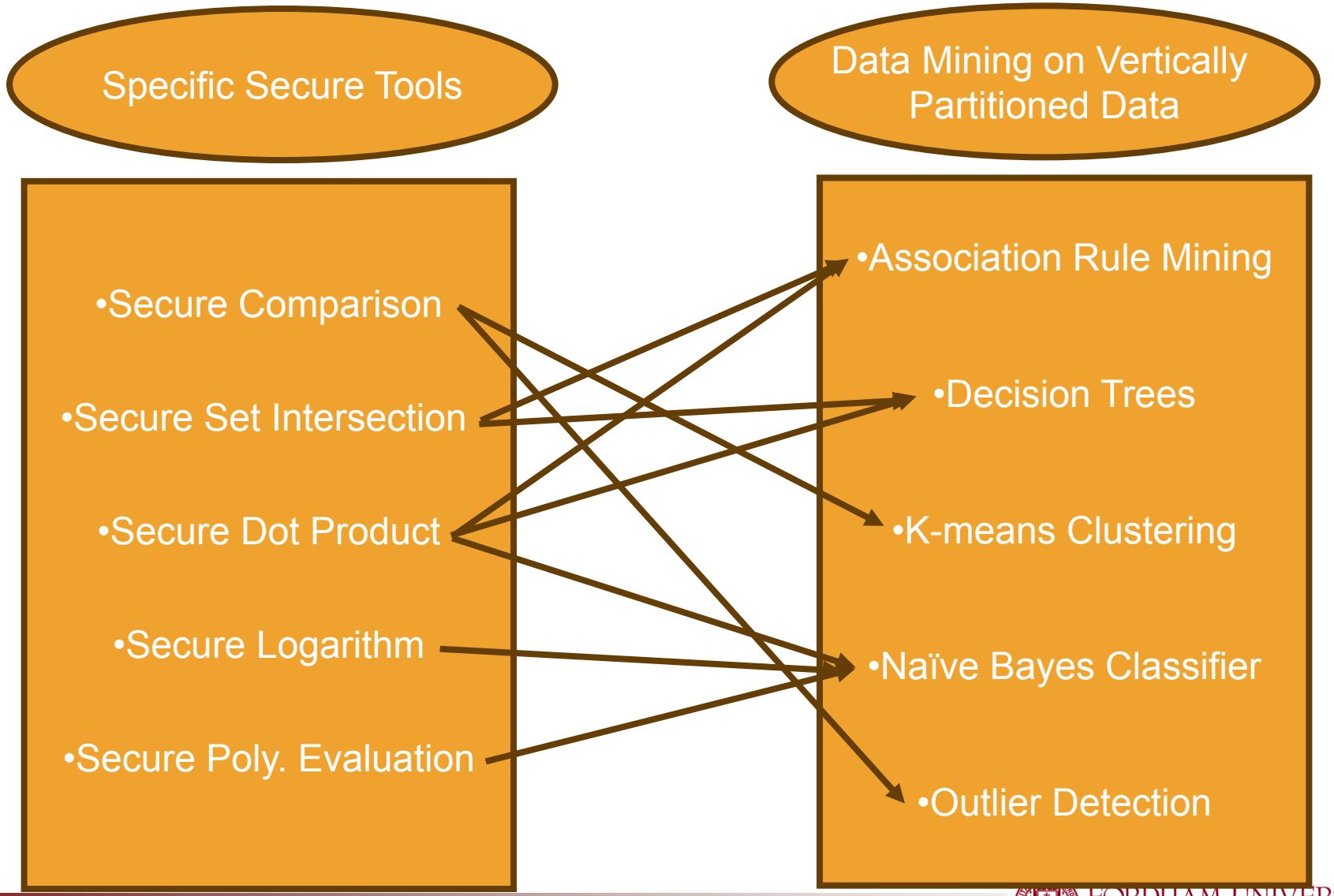
- **Secure Functionalities**
  - **Secure Comparison**
    - Comparing two integers without revealing the integer values.
  - **Secure Polynomial Evaluation**
    - Party A has polynomial P(x) and Part B has a value b, the goal is to calculate P(b) without revealing P(x) or b
  - **Secure Set Intersection**
    - Party A has set S_A and Party B has set S_B, the goal is to calculate $S_A \cap S_B$ without revealing anything else.

Privacy-Preserving Association Rules Mining

- **Secure Functionalities**
  - **Secure Set Union**
    - Party A has set SA and Party B has set SB , the goal is to calculate without revealing anything else.
  - **Secure Dot Product**
    - Party A has a vector X and Party B has a vector Y. The goal is to calculate X.Y without revealing anything else.

Privacy-Preserving Association Rules Mining




Specific Secure Tools
Data Mining on Vertically Partitioned Data
•Secure Comparison
•Secure Set Intersection
•Secure Dot Product
•Secure Logarithm
•Secure Poly. Evaluation
•Association Rule Mining
•Decision Trees
•K-means Clustering
•Naïve Bayes Classifier
•Outlier Detection