PURDUE
UNIVERSITY

Privacy-Preserving Data Mining

Shared analysis without shared data

16 February 2011

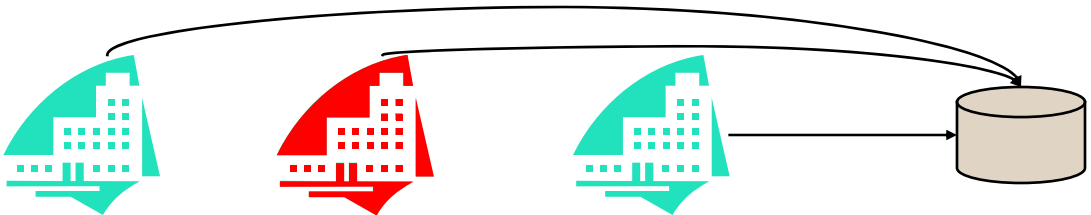
Chris Clifton

Idea: Collaborate to learn (only) desired results

- Data owners and FDA participate in protocol
  - Results same as if all data sent to FDA
  - Protocol ensures data not disclosed
- Solutions for many types of analysis
  - Theoretically possible for any (polynomially computable) function
- Protection beyond individual privacy
  - Even controls disclosure of which data owner responsible for which event

Association Rule Mining: Horizontal Partitioning

- Goal: Learn conditions unusually likely to lead to adverse outcome
  - Low creatinine clearance & >0.125mg digoxin ➔high ADE*
  - Identify *all such rules* – association rule mining
- Problem: rules occurring at only one participating site could be liability issue
  - Why is insurer *X* the only one with a particular rule?
  - Policies limiting coverage of (possibly more appropriate) medications?
- Solution: Reveal only rules, not source
  - But learn rules based on combined data from all sources

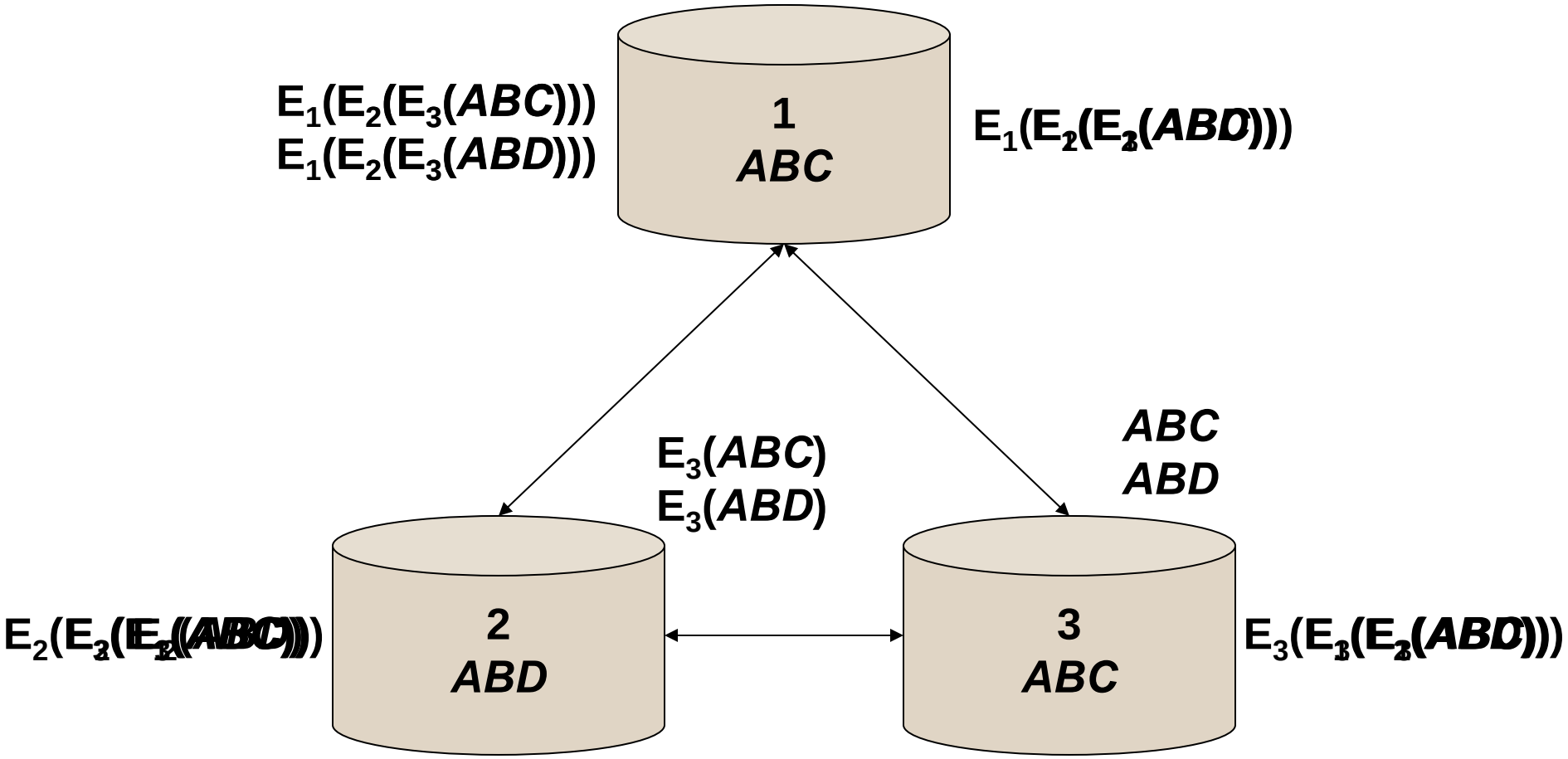


* doi:10.1345/aph.1E642

Overview of the Method
(Kantarcıoğlu and Clifton TKDE'04)

- Find the union of the locally large candidate itemsets securely
  - Any rule sufficiently strong at one site that it might be true globally
  - But don't reveal source (or even number of sites where the rule is significant)
- Compute statistics to determine global incidence of rules
  - E.g. A&B&C → Adverse Drug Event
  - But don't reveal any site-specific statistics

Computing Candidate Sets
Key – commutative encryption



Which Rules are *Globally* Frequent?

- Goal: Given a rule that is significant at (at least) one site, is it significant overall?

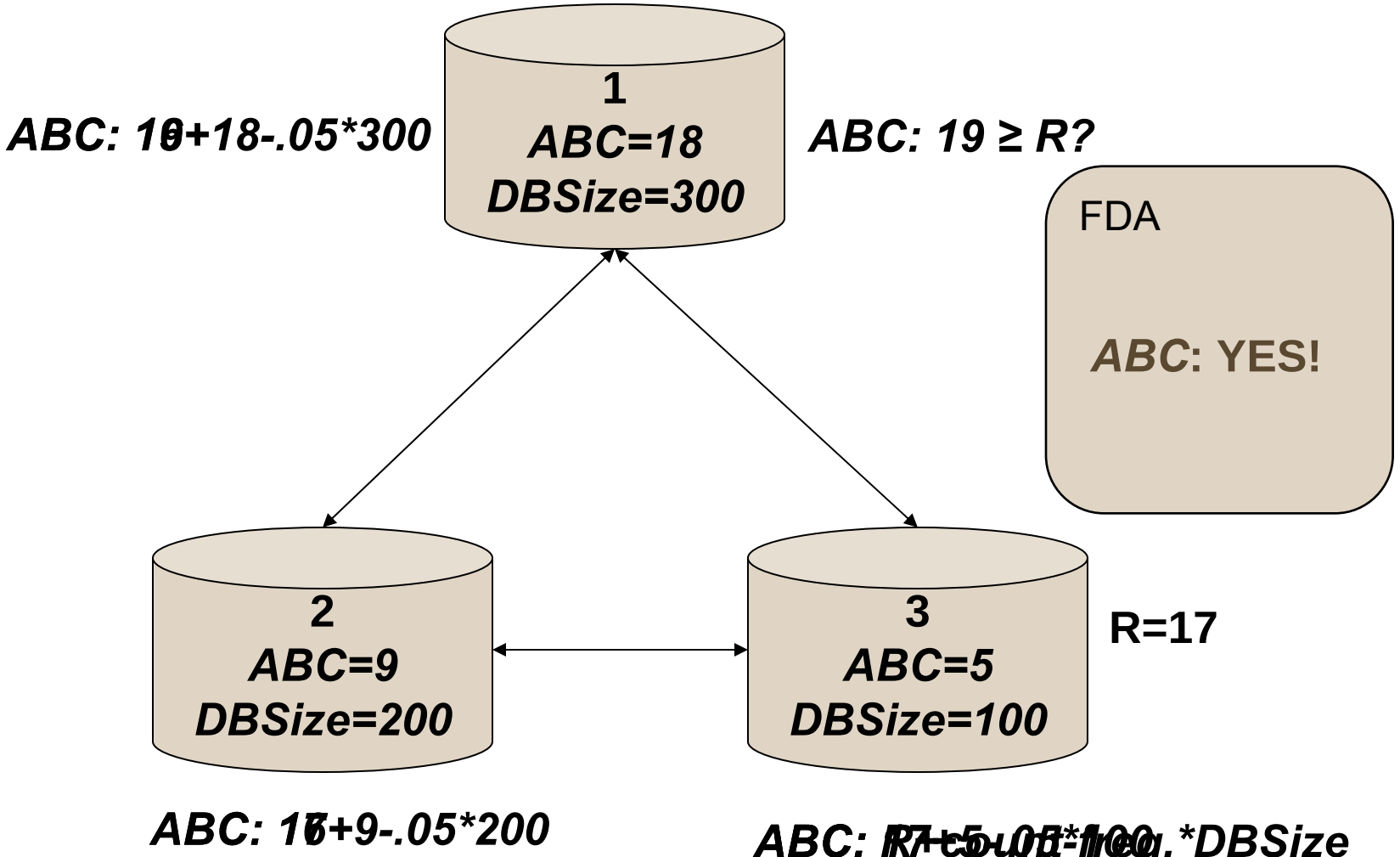
$$X.\text{sup} \geq s * \sum_{i=1}^{n} \left| DB_i \right| \quad (1)$$

– $$\sum_{i=1}^{n} X.\text{sup}_i \geq \sum_{i=1}^{n} s^* \left| DB_i \right| \quad (2)$$

– $$\sum_{i=1}^{n} (X.\text{sup}_i - s^* \left| DB_i \right|) \geq 0 \quad (3)$$

- Checking (1) is equivalent to checking (3)

Computing Frequent: Is A&B&C→A.D.E ≥ 5%?



Privacy-Preserving Data Mining: Successes

Numerous machine learning tasks solved for horizontally and vertically partitioned data

- Decision tree learning and use
- K-Nearest Neighbor
- Clustering: K-Means, EM, General distance-based approaches
- Outlier / anomaly detection
- Collaborative Filtering
- Naïve Bayes, Bayes network structure
- And many more

PURDUE
UNIVERSITY

Privacy-Preserving Data Mining

Shared analysis without shared data

16 February 2011

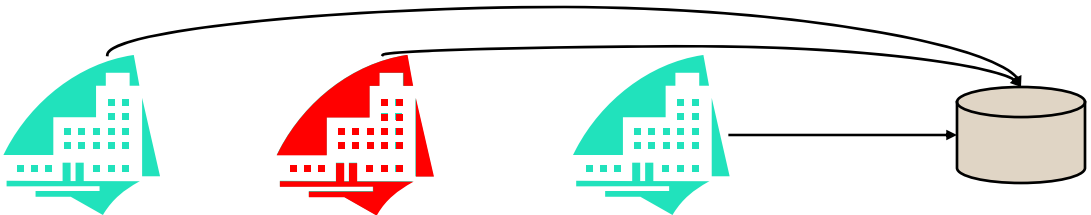
Chris Clifton

Idea: Collaborate to learn (only) desired results

- Data owners and FDA participate in protocol
  - Results same as if all data sent to FDA
  - Protocol ensures data not disclosed
- Solutions for many types of analysis
  - Theoretically possible for any (polynomially computable) function
- Protection beyond individual privacy
  - Even controls disclosure of which data owner responsible for which event

Association Rule Mining: Horizontal Partitioning

- Goal: Learn conditions unusually likely to lead to adverse outcome
  - Low creatinine clearance & >0.125mg digoxin ➔high ADE*
  - Identify *all such rules* – association rule mining
- Problem: rules occurring at only one participating site could be liability issue
  - Why is insurer *X* the only one with a particular rule?
  - Policies limiting coverage of (possibly more appropriate) medications?
- Solution: Reveal only rules, not source
  - But learn rules based on combined data from all sources

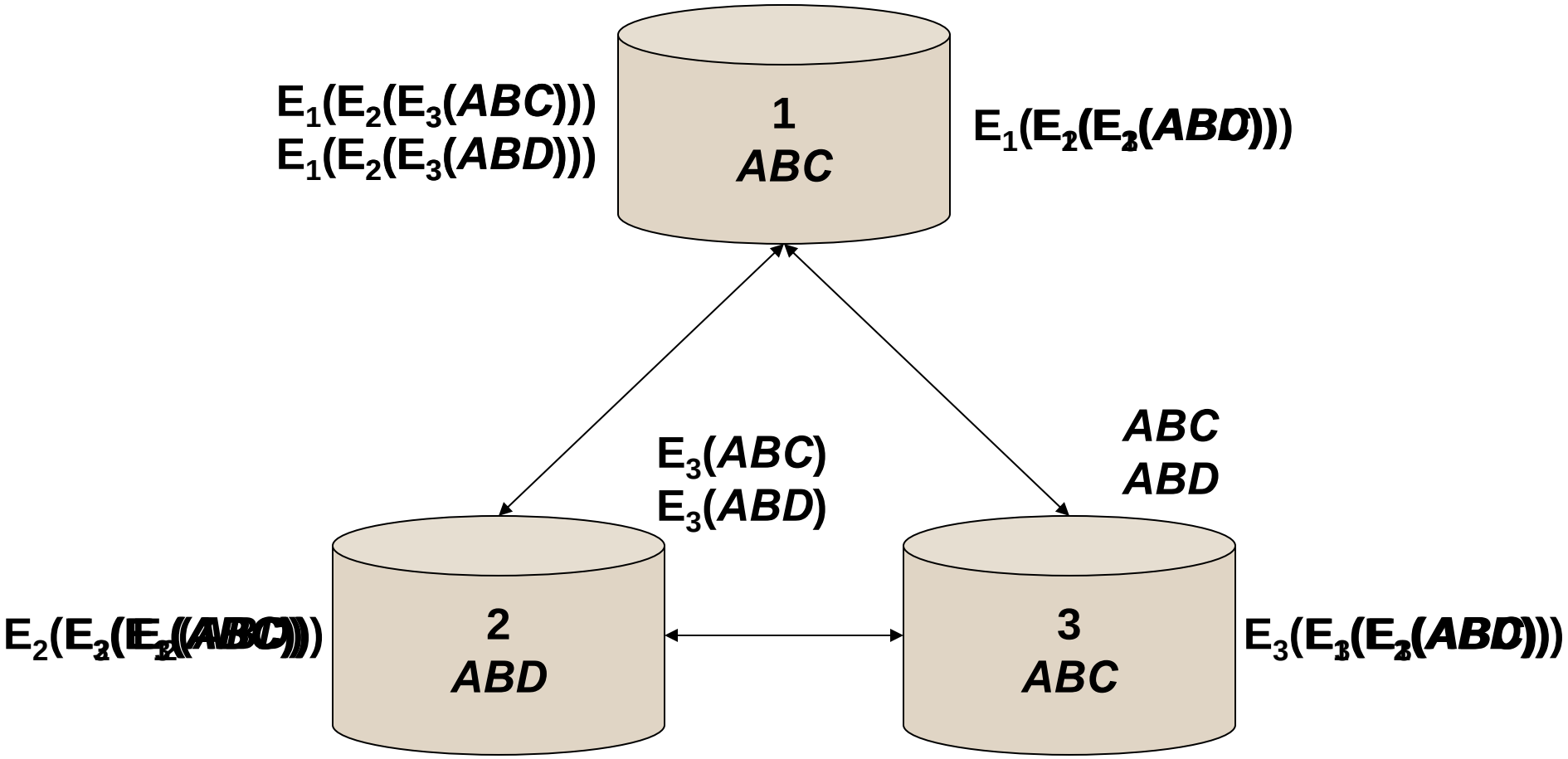


* doi:10.1345/aph.1E642

Overview of the Method
(Kantarcıoğlu and Clifton TKDE'04)

- Find the union of the locally large candidate itemsets securely
  - Any rule sufficiently strong at one site that it might be true globally
  - But don't reveal source (or even number of sites where the rule is significant)
- Compute statistics to determine global incidence of rules
  - E.g. A&B&C → Adverse Drug Event
  - But don't reveal any site-specific statistics

Computing Candidate Sets
Key – commutative encryption



Which Rules are *Globally* Frequent?

- Goal: Given a rule that is significant at (at least) one site, is it significant overall?

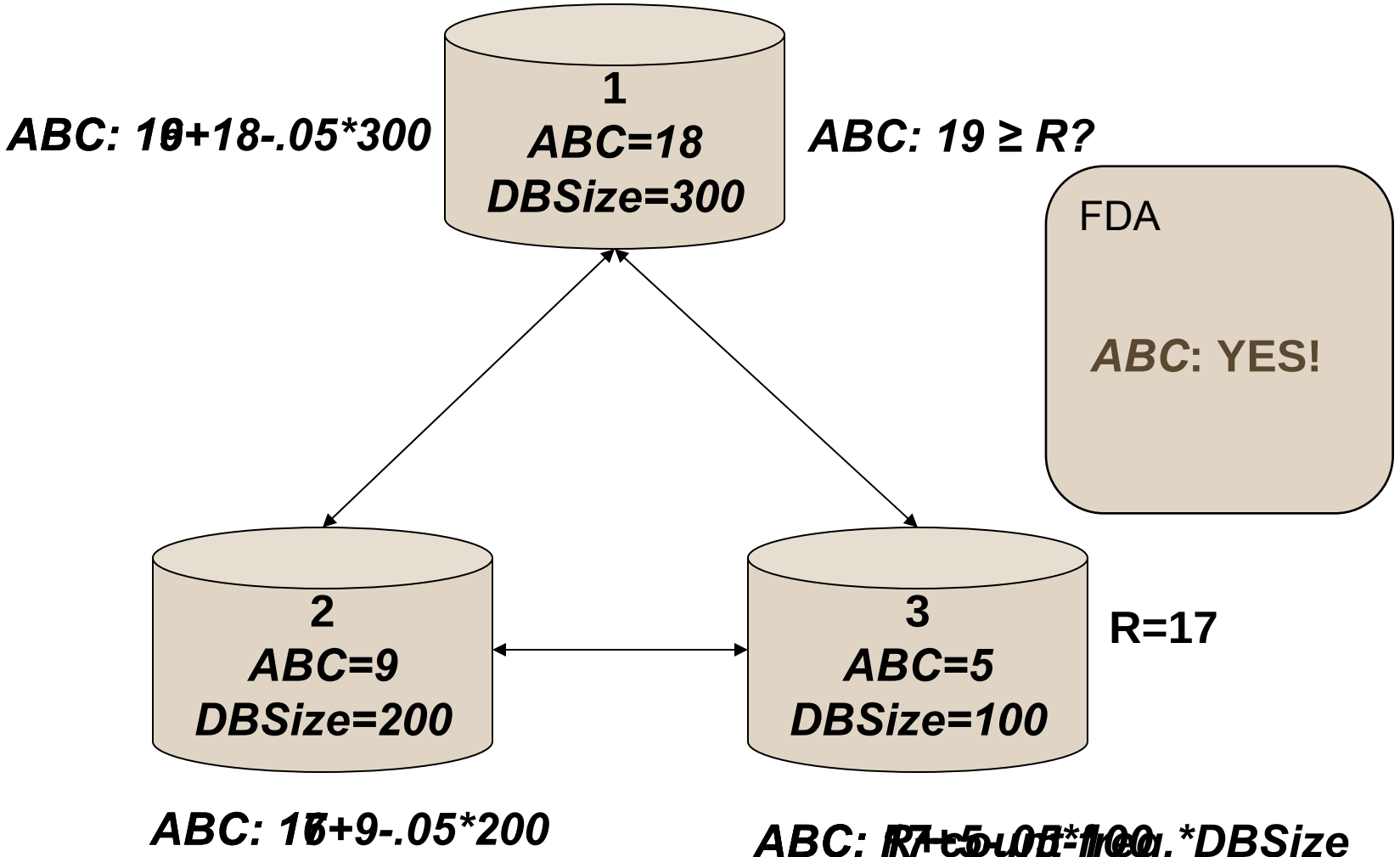
$$X.\text{sup} \geq s * \sum_{i=1}^{n} \left| DB_i \right| \quad (1)$$

– $$\sum_{i=1}^{n} X.\text{sup}_i \geq \sum_{i=1}^{n} s^* \left| DB_i \right| \quad (2)$$

– $$\sum_{i=1}^{n} (X.\text{sup}_i - s^* \left| DB_i \right|) \geq 0 \quad (3)$$

- Checking (1) is equivalent to checking (3)

Computing Frequent: Is A&B&C→A.D.E ≥ 5%?



Privacy-Preserving Data Mining: Successes

Numerous machine learning tasks solved for horizontally and vertically partitioned data

- Decision tree learning and use
- K-Nearest Neighbor
- Clustering: K-Means, EM, General distance-based approaches
- Outlier / anomaly detection
- Collaborative Filtering
- Naïve Bayes, Bayes network structure
- And many more