PORTIA *Privacy, Obligations, and Rights in Technologies of Information Assessment*

Privacy in a Networked World

Rebecca Wright, Rutgers University

Frontiers of Engineering Symposium
September 24, 2007

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

This work was supported by the National Science Foundation under grant number CNS-0331584

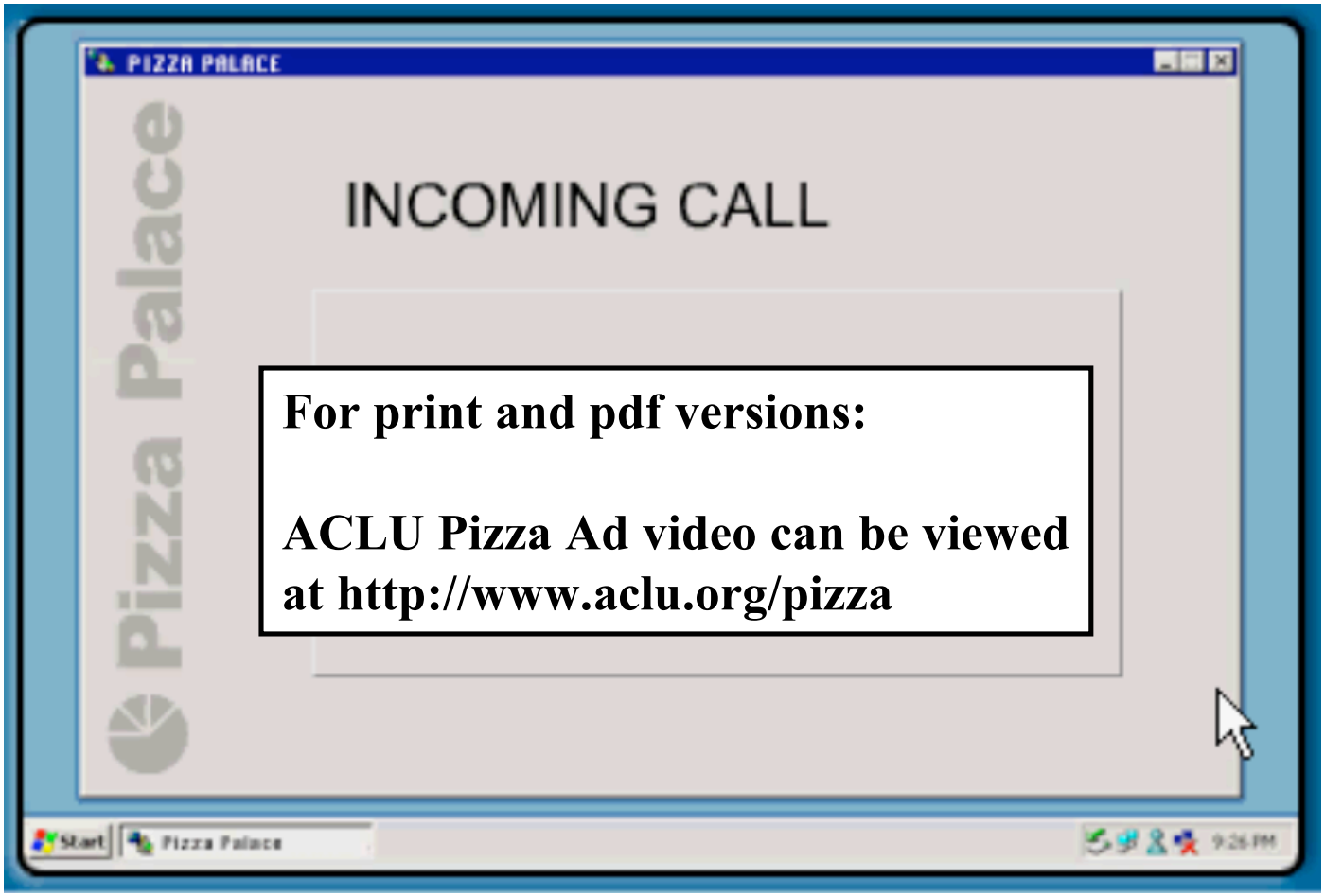
Erosion of Privacy

> "You have zero privacy. Get over it."
>
> – Scott McNealy, 1999

- Changes in technology are making privacy harder.
  - increased use of computers and networks
  - reduced cost for data storage
  - increased ability to process large amounts of data
- Becoming more critical as public awareness, potential misuse, and conflicting goals increase.

The Data Revolution

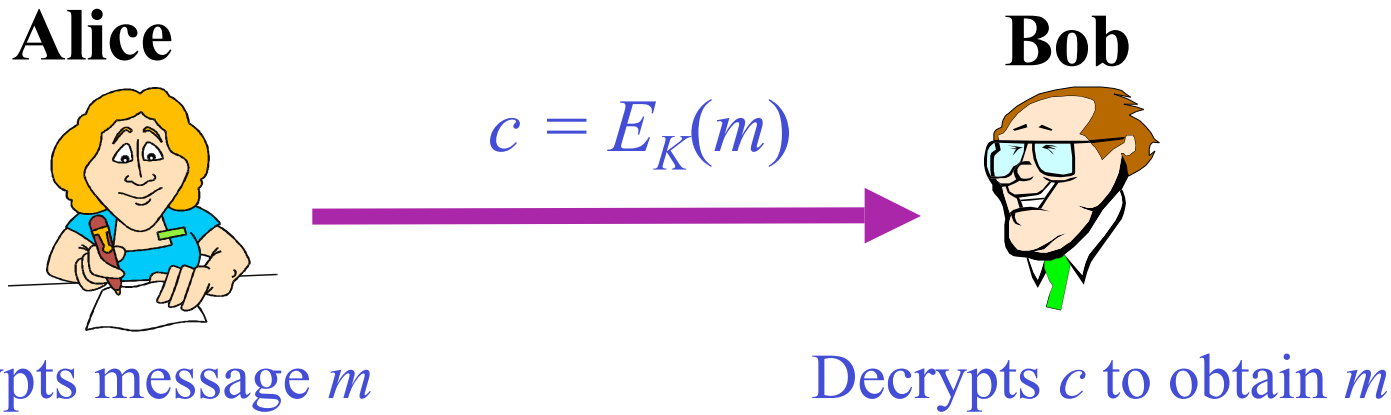
- We are in the midst of a data revolution fueled by the actual, perceived, and potential usefulness of data.
- Most electronic and physical activities leave some kind of data trail. These trails can provide useful information to various parties.
- However, there are also concerns about appropriate handling and use of sensitive information.
- Privacy-preserving methods of data handling seek to provide sufficient privacy as well as sufficient utility.



Courtesy of the ACLU.

Confidentiality vs. Privacy

Encryption works reasonably well to protect data confidentiality in transit and in storage.



Confidentiality vs. Privacy

Encryption works reasonably well to protect data confidentiality in transit and in storage.

"Reasonably" because issues remain with key management, end-host security, correct implementation.

Alice

$c = E_K$

Encrypts message m

Decrypts c to obtain m

Confidentiality vs. Privacy

Encryption works reasonably well to protect data confidentiality in transit and in storage.

"Reasonably" because issues remain with key management, end-host security, correct implementation.

Alice

$c = E_K$

Encrypts message m

Decrypts c to obtain m

In contrast, privacy is about what Bob can and will do with m.

Various Definitions of Privacy

Privacy means different things to different people, to different cultures, and in different situations.

- A reasonable working definition: The ability to control the use of one's personal information.
- Fair Information Practices [HEW73, OECD80]: A set of principles for providing control and/or notice to individuals regarding their personal information.
- Contextual integrity [Niss04]: a privacy breach is when information is used in a way that violates societal norms.
- Federal statistical agencies (e.g., Census bureau): No individual isolation or reidentification should be possible.
- Differential privacy [DN03, DN04, CDMSW05]: The difference in privacy for an individual if her data is in a given statistical database.

Various Definitions of Privacy

Privacy means different things to different people, to different cultures, and in different situations.

- A reasonable working definition: The ability to control the use of one's personal information.
- Fair Information Practice
  providing control and/or
  personal information.

No "one size fits all" definition.

⇒ No "one size fits all" solution.

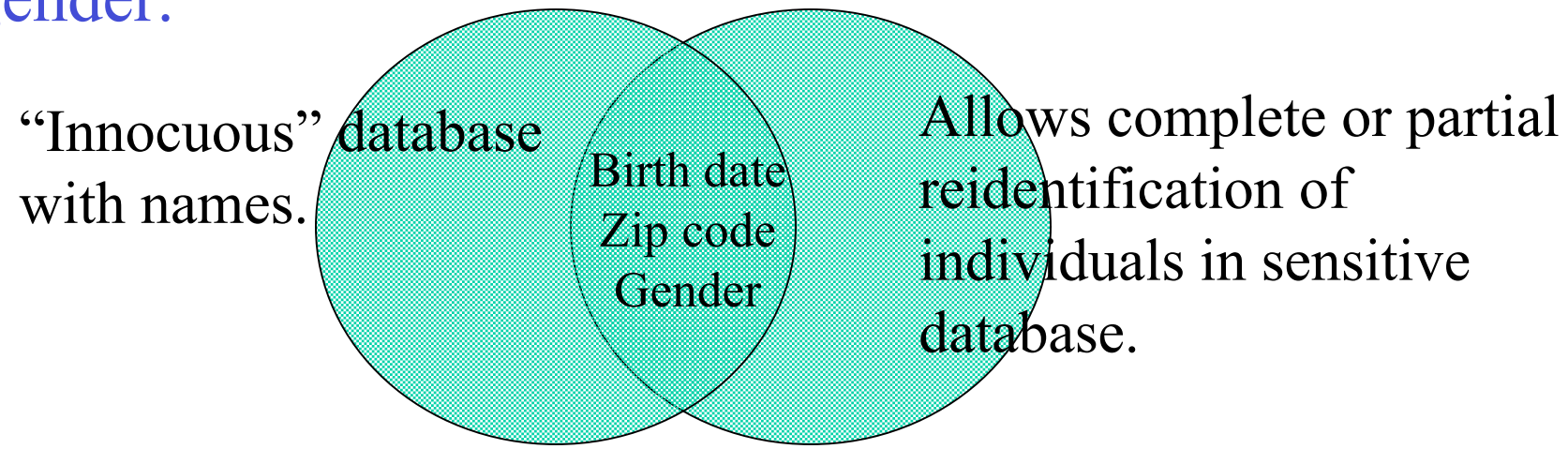
- Contextual integrity [Niss04]: a privacy breach is when information is used in a way that violates societal norms.
- Federal statistical agencies (e.g., Census bureau): No individual isolation or reidentification should be possible.
- Differential privacy [DN03, DN04, CDMSW05]: The difference in privacy for an individual if her data is in a given statistical database.

Challenges in Today's World

- Multiple data sources, widely available and extremely useful to individuals, enterprises, and society.
  - Examples: WWW coupled with search engines, social networks, blogs, data brokers
  - Level of detail and accessibility provide elaborate dossiers that go beyond most people's expectations
  - Even data sources intended to be confidential or restricted are often revealed. (E.g., Data breaches such as Choicepoint, many universities, VA, AOL search logs.)
- Concept of "personally identifiable information" (PII) isn't very robust in the face of such data.

Reidentification

- Sweeney: 87% of the US population can be uniquely identified by their date of birth, 5-digit zip code, and gender.



- AOL search logs released August 2006: user IDs and IP addresses removed, but replaced by unique random identifiers. Some queries provide information about who the querier is, others give insight into the querier's mind.

Abuses of Sensitive Data

- Embarrassment, loss of employment, cost of health coverage, effect on personal relationships
- Identity theft:
  - Current authentication mechanisms often rely on personal "private" information such as SSN, mother's maiden name, etc.
  - This information can often be easily learned and used to impersonate others.
  - The repeated use of a fixed, communicated authenticator has long been known to be insecure, but ease of use has made it widespread.
- Not just for individuals: Unfair business advantage, compromise of national security

Advantages of Privacy Protection

- protection of personal information: protects individuals and helps maintain their trust
- protection of proprietary or government-sensitive information
- enables collaboration between different data holders (since they may be more willing or able to collaborate if they need not reveal their information)
- compliance with legislative policies (e.g., HIPAA, EU privacy directives)

Privacy-Enhancing Technologies

Many tools, systems, and methods exist for enabling some activities or outcomes while protecting the privacy of some or all sensitive information. Some categories and examples:

- Disclosure and informed consent: P3P and related tools, automated generation and evaluation of privacy policies
- Computing without revealing: anonymous credentials [Chaum82] , secure multiparty computation [Yao82, Yao86, BGW88]
- Hiding real information among other information: mix networks [Chaum81] , Crowds [RR98], TrackMeNot [HN07]

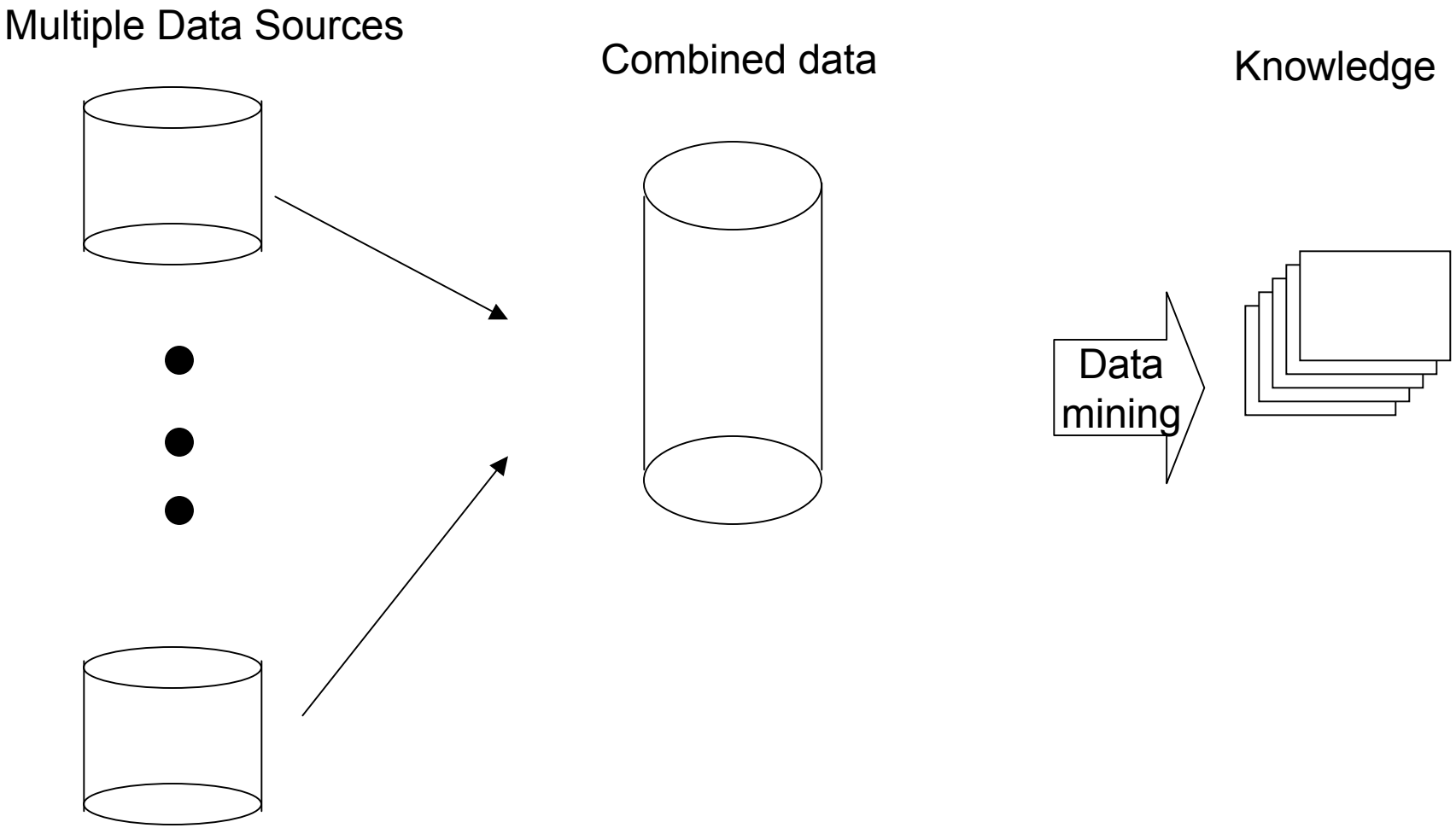
Data Collection and Mining

- Analyze large amounts of data from diverse sources, looking for patterns, anomalies, representations that can be used to classify other data.

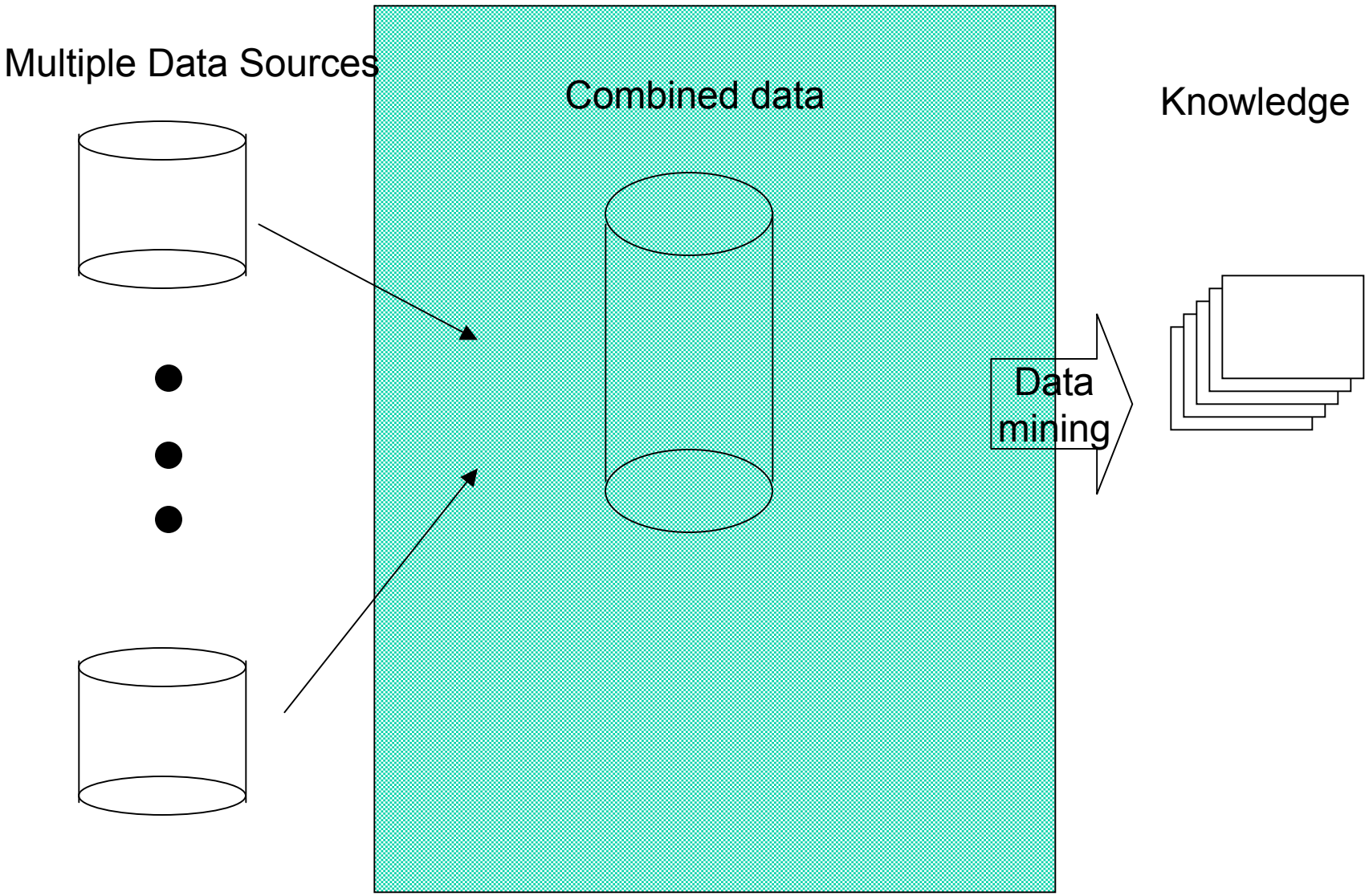
- Many applications:
  - Biomedical research
  - Marketing, personalized customer service
  - Law enforcement and homeland security
    - detect and thwart possible incidents before they occur
    - recognize that an incident is underway
    - identify and prosecute perpetrators after incidents occur

- Can provide useful information, but also creates privacy concerns

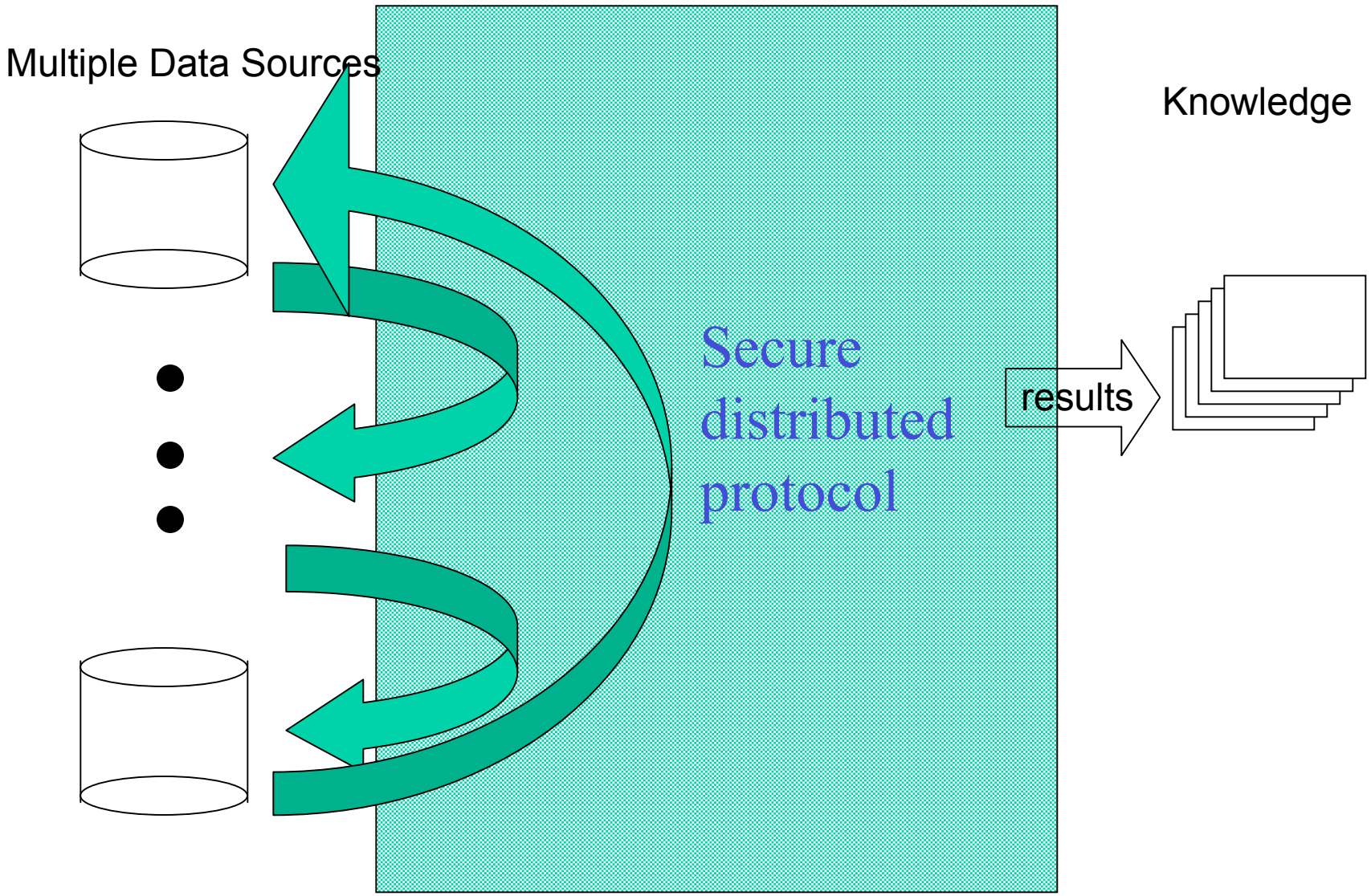
Data Mining


Multiple Data Sources
Combined data
Knowledge
Data
mining

Data Mining


Multiple Data Sources
Combined data
Knowledge
Data
mining

Privacy-Preserving Data Mining

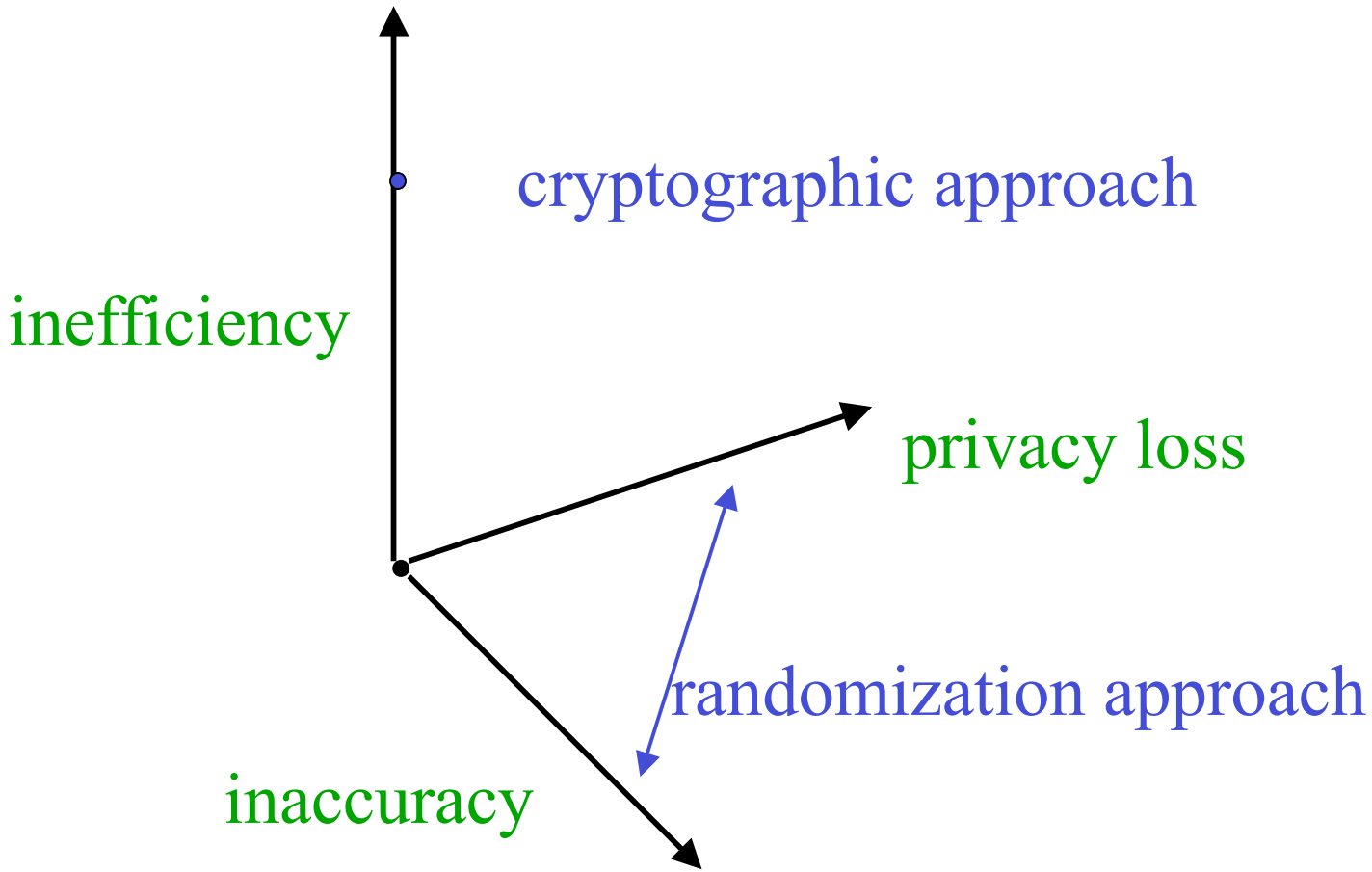


Privacy-Preserving Data Mining (PPDM)

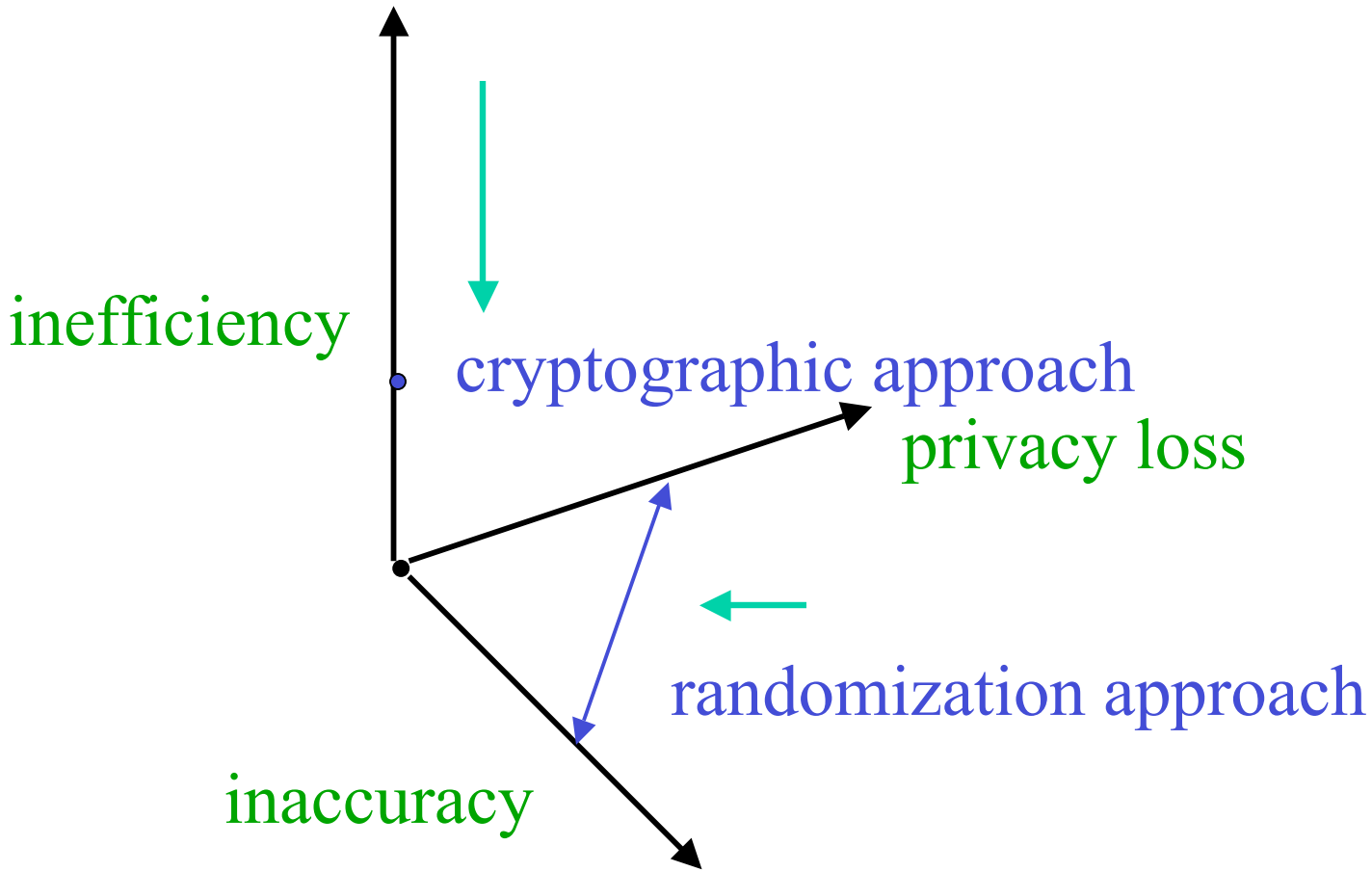
Enable mining for some kinds of results (**utility**), while protecting some kinds of information (**privacy**).

- First solutions were introduced in 2000, taking different approaches:
  - [LP00]: using cryptography, computes ID3 decision trees for data held by two parties, provably leaking nothing else.
  - [AS00]: using random data perturbation, computes reasonably accurate ID3 decision trees for data held by two parties, while obscuring original data.

Cryptography vs. Randomization



Cryptography vs. Randomization



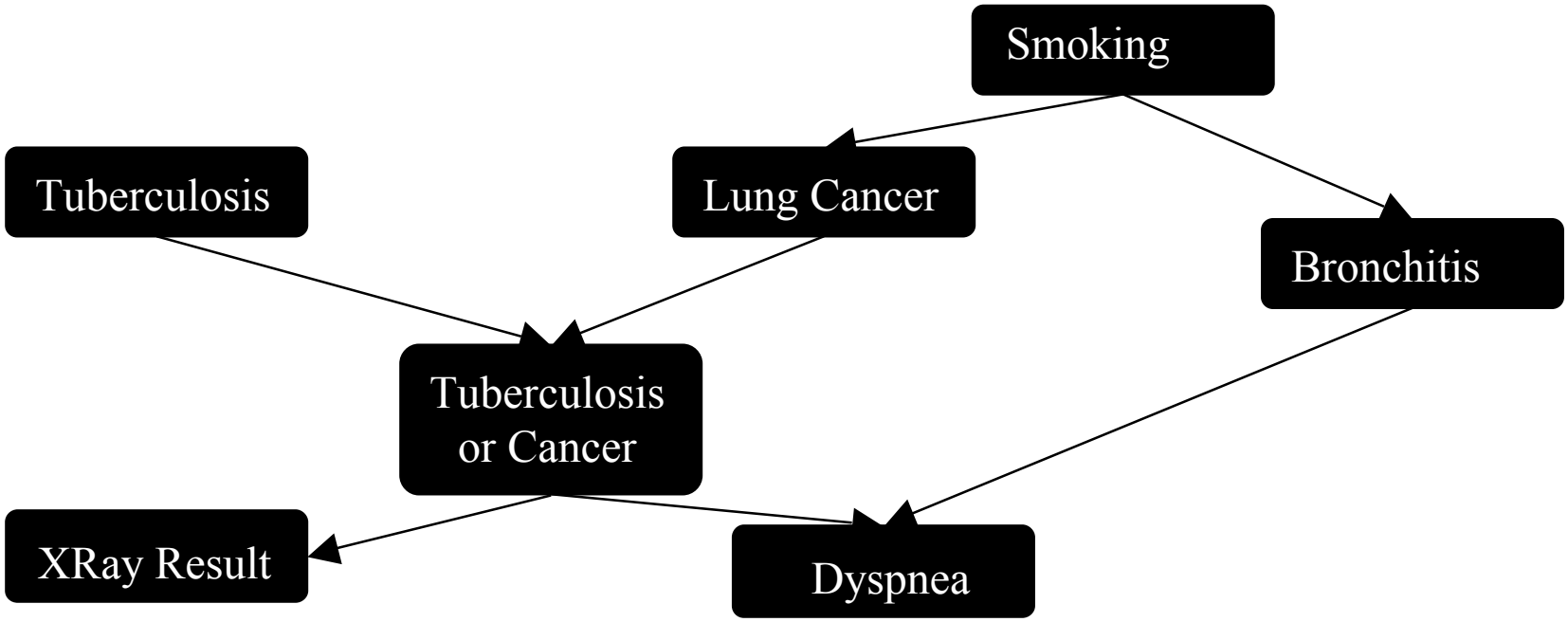
Some of Our PPDM Work

Our work takes the cryptographic approach.

- [WY04,YW05]: privacy-preserving construction of Bayesian networks from vertically partitioned data.
- [YZW05]: privacy-preserving frequency mining in the fully distributed model (enables naïve Bayes classification, decision trees, and association rule mining).
- [JW05, JPW06]: privacy-preserving clustering: *k*-means clustering for arbitrarily partitioned data and a divide-and-merge clustering algorithm for horizontally partitioned data.
- [ZYW05]: privacy-preserving solutions for a data publisher to learn a *k*-anonymized version of a fully distributed database.
- [RKWF05]: an experimental platform for PPDM.

Bayesian Networks

Bayesian networks: a graphical model that encodes probabilistic relations among variables.



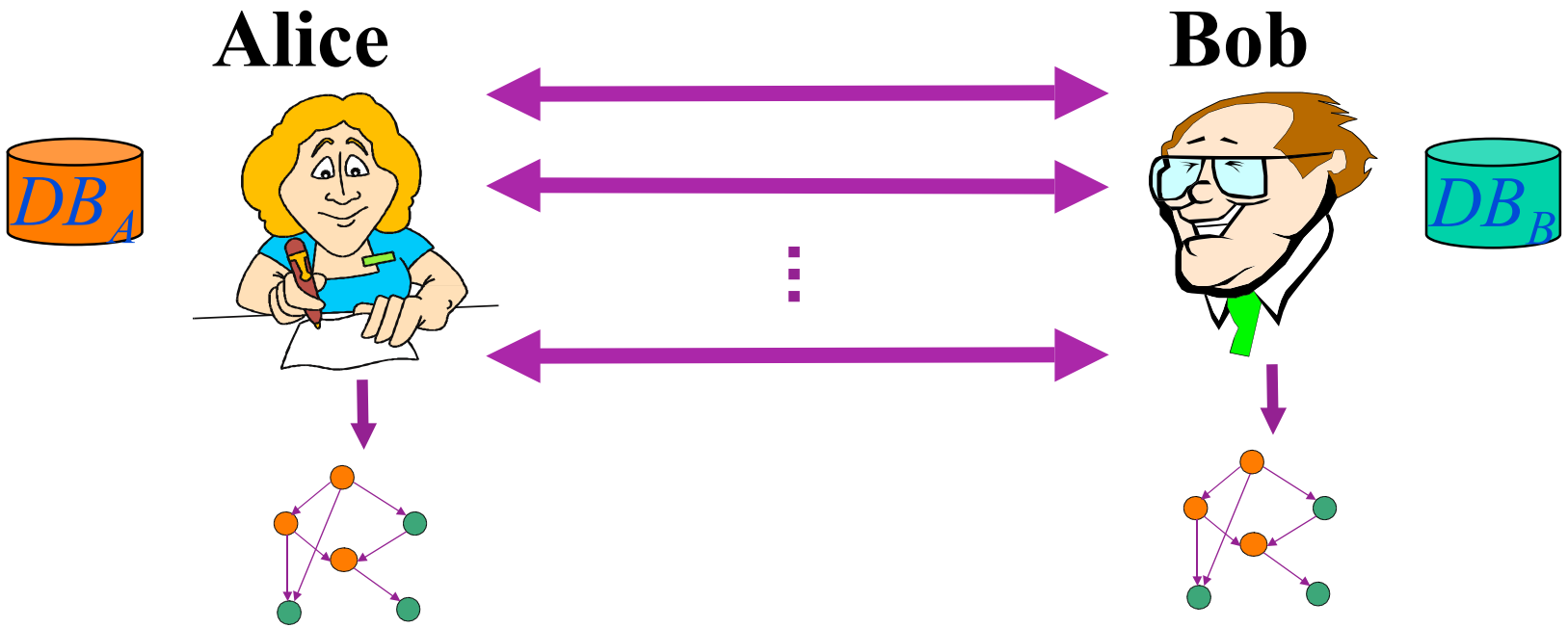
Once learned from some data, can be applied to predict unknown or uncertain variables in new data.

Bayes Network Applications

- Industrial
  - Processor Fault Diagnosis - by Intel
  - Auxiliary Turbine Diagnosis - GEMS by GE
  - Diagnosis of space shuttle propulsion systems - VISTA by NASA/Rockwell
- Military
  - Automatic Target Recognition - MITRE
  - Autonomous control of unmanned underwater vehicle - Lockheed Martin
  - Assessment of Intent
- Medical Diagnosis
  - Internal Medicine
  - Pathology diagnosis - Intellipath by Chapman & Hall
  - Breast Cancer Manager with Intellipath
- Commercial
  - Financial Market Analysis
  - Information Retrieval
  - Software troubleshooting and advice - Windows 95 & Office 97
  - Spam detection

Privacy-Preserving Bayes Networks [WY04,YW05]

Goal: Two parties cooperatively learn Bayesian network on a vertically partitioned database, without either party learning anything except the Bayesian network itself.



K2 Algorithm for BN Learning

- Determining the best Bayesian network structure for a given data set is NP-hard (i.e., probably computationally infeasible), so heuristics are used in practice.
- The K2 algorithm [CH92] is a widely used Bayesian network structure-learning algorithm. We use this as our starting point.
- Considers nodes in sequence. Adds new parent that most increases a score function *f*, up to a maximum number of parents per node.

$$f(i,\pi(i)) = \prod \frac{\alpha_0!\alpha_1!}{(\alpha_0+\alpha_1+1)!}$$

K2 Algorithm for BN Learning

- Determining the best Bayesian network structure for a given data set is NP-hard (i.e., probably computationally infeasible), so heuristics are used in practice.
- The K2 algorithm [CH92] is a widely used Bayesian network structure-learning algorithm. We use this as our starting point.
- Considers nodes in se[illegible] parent that most increases a score function *f*, up to a [illegible] number of parents per node.

α-parameters

$$f(i,\pi(i)) = \prod \frac{\alpha_0!\alpha_1!}{(\alpha_0+\alpha_1+1)!}$$

Our Solution: Approximate Score

Modified score function: approximates the same relative ordering, and lends itself well to private computation.

- Apply natural log to *f* and use Stirling's approximation
- Drop constant factor and bounded term. Result is:

$$g(i,\pi(i)) = \sum \left(\tfrac{1}{2}(\ln\alpha_0 + \ln\alpha_1 - \ln t) + \left(\alpha_0 \ln\alpha_0 + \alpha_1 \ln\alpha_1 - t \ln t\right)\right)$$

where $t = \alpha_0 + \alpha_1 + 1$

Cryptographic Primitives

- Homomorphic encryption:
  - Uses mathematical properties of underlying encryption scheme to allow specific computations on encrypted values without knowledge of decryption key and without revealing anything about the cleartext.
  - E.g., given $E(x)$ and $E(y)$, Bob can compute $E(x + y)$.
- Secret sharing:
  - Divides a value s into two random "shares" a and b. One alone yields no information about s, but a and b together reveal s.
  - Exclusive-or is an example: $s = a \oplus b$ for random a.
- Secure multiparty computation:
  - [Yao86]: allows two parties to privately compute any function of their inputs, with polynomial overhead in the size of their inputs and complexity of the function.

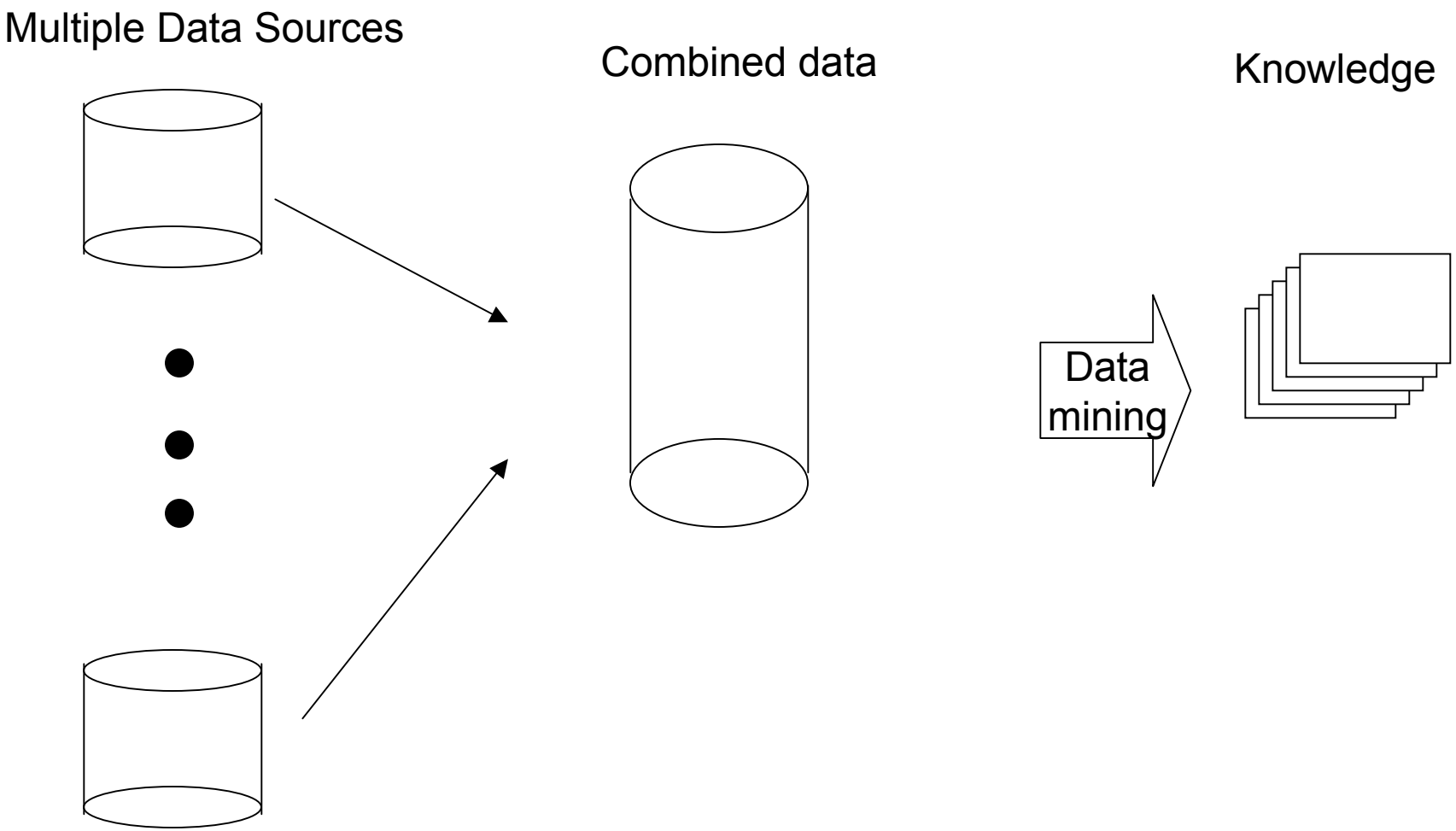
Our Solution: Components

Sub-protocols used:

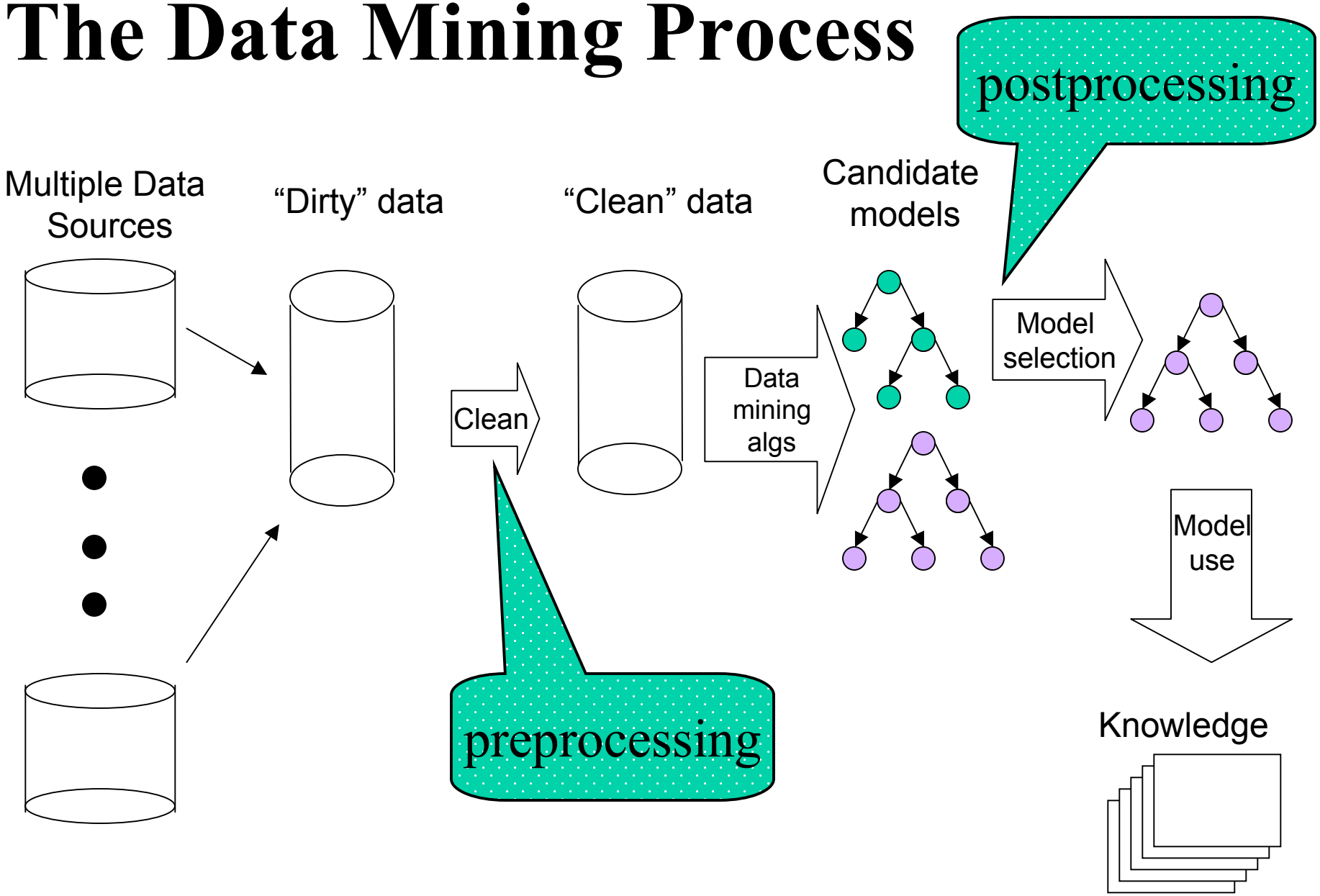
- Privacy-preserving scalar product protocol: based on homomorphic encryption
- Privacy-preserving computation of α-parameters: uses scalar product
- Privacy-preserving score computation: uses α-parameters, [LP00] protocols for $\ln x$ and $x \ln x$
- Privacy-preserving score comparison: uses [Yao86]

All intermediate values are protected using secret sharing.

Data Mining


Multiple Data Sources
Combined data
Knowledge
Data
mining

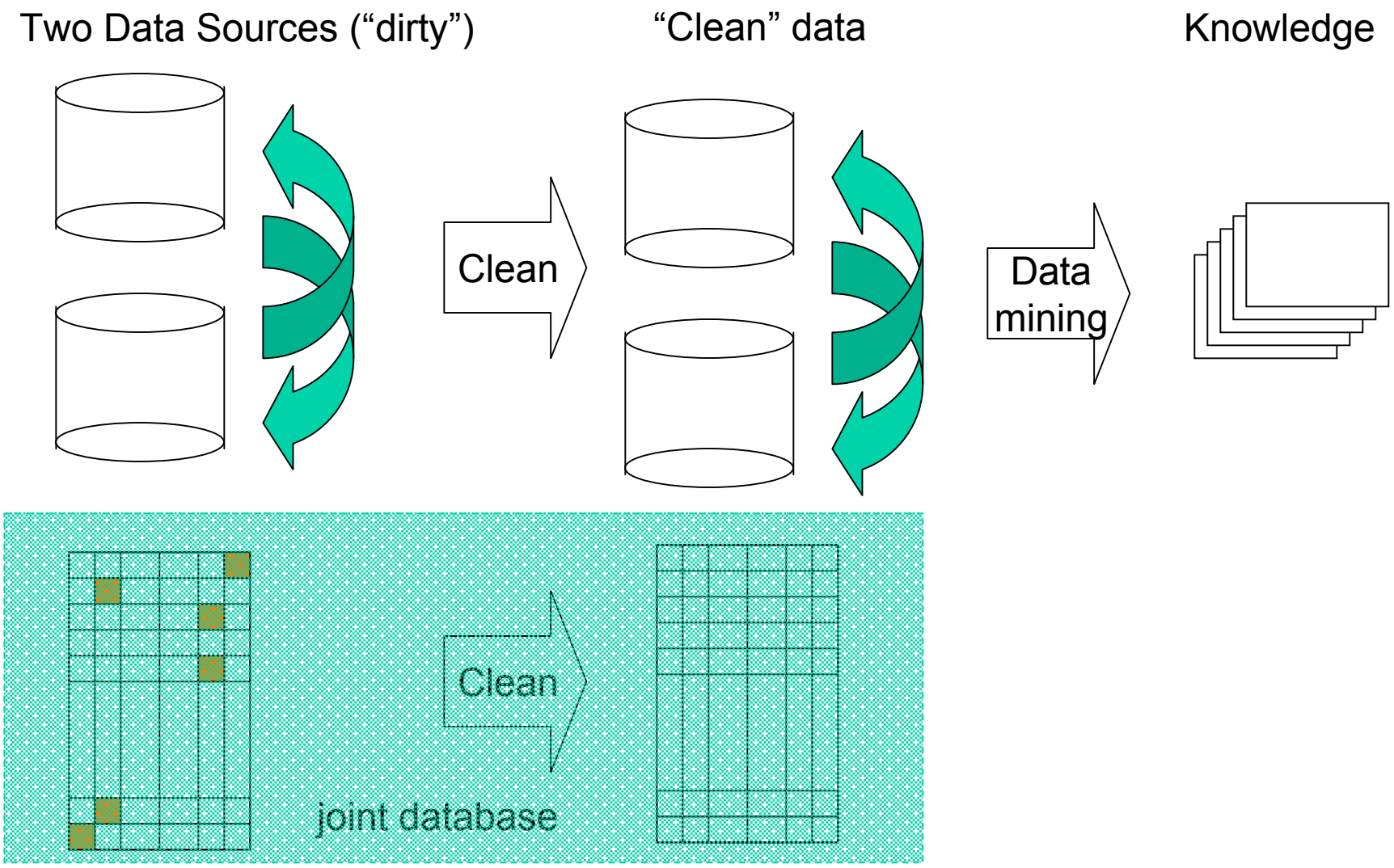
The Data Mining Process



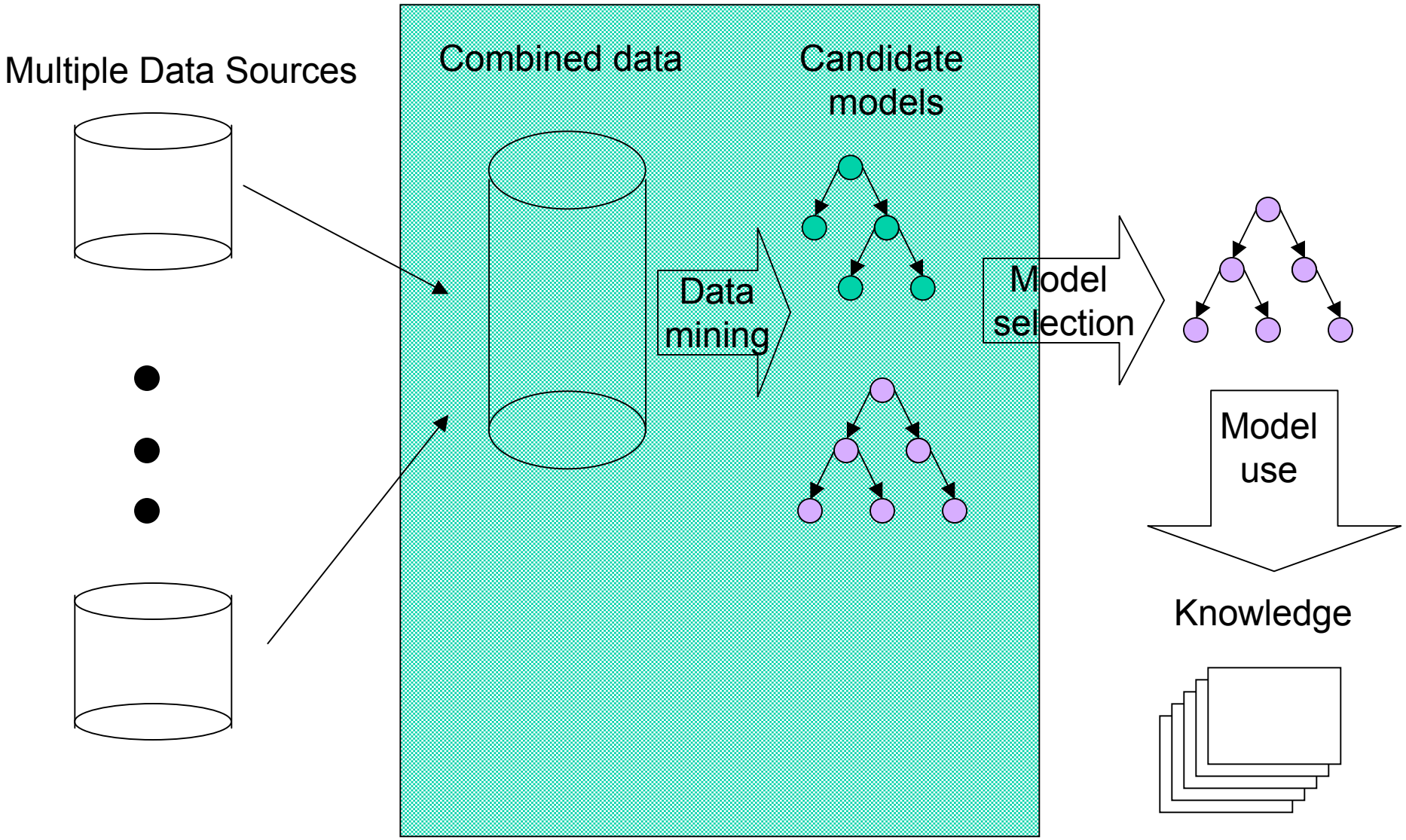
Extending the PPDM Boundary

- In order for PPDM to be useful, preprocessing and postprocessing must also incorporate privacy. Otherwise, either:
  - If performed, privacy is lost, because data must be revealed in order to carry out preprocessing and postprocessing steps.
  - If not performed, utility is lost, because quality of results is too low.

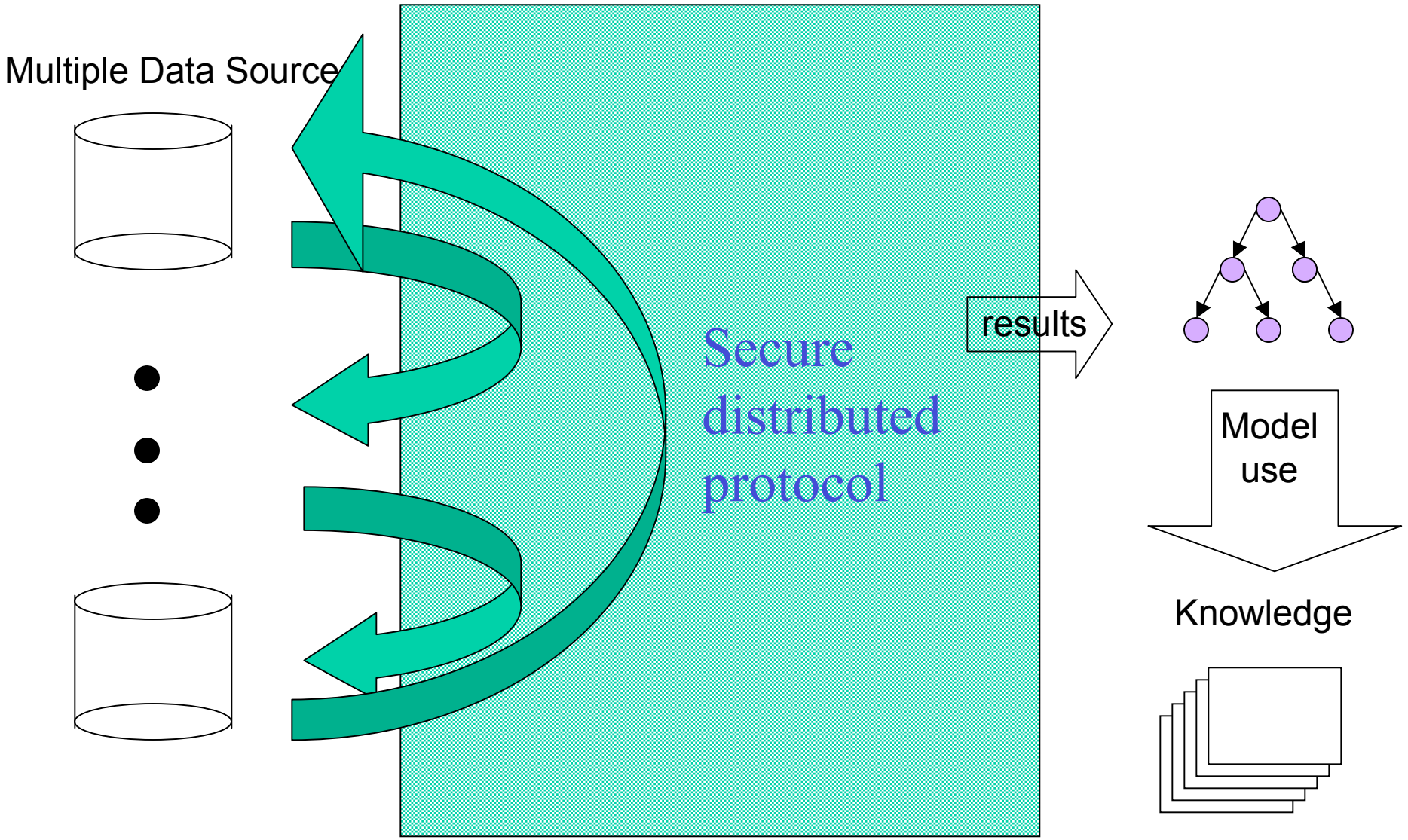
Privacy-Preserving Data Cleaning [JW06]



Privacy-Preserving Model Selection [YZW07]



Privacy-Preserving Model Selection [YZW07]



Challenges for the Future

- Improving existing solutions in various ways.
  - For example, policies for privacy-preserving data mining: languages, generation, reconciliation, and enforcement.
- Rigorous understanding of inherent tradeoffs in utility, privacy, efficiency, generality.
- Moving beyond the multitude of "point" solutions towards a comprehensive solution.

Challenges for the Future

- Improving existing solutions in various ways.
  - For exa
    mining
    enforce
- Rigorous
  utility, pr
- Moving t
  towards a

- Technology can enable new public policy decisions by instantiating solutions with new properties.
- Technology, policy, and education must work together in order to have a significant impact.