Chapter 1

Chao Yei Ching Building

Introduction

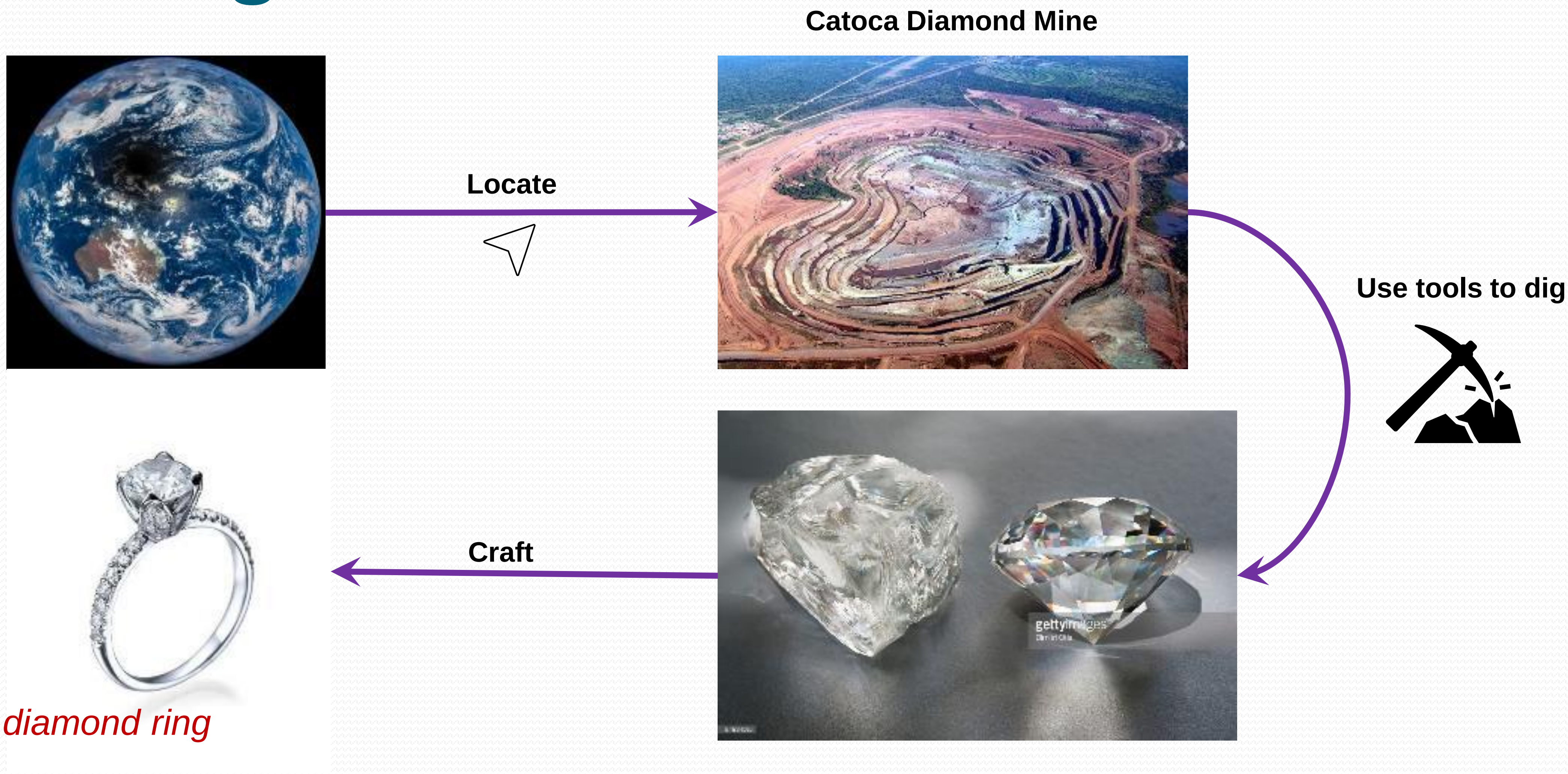
Overview

- *Data Mining* and *Knowledge Discovery*
- Data mining tasks
  - Classification
  - Association analysis
  - Cluster analysis
- The KDD process

What is Data Mining?

- Some terms we might have heard from database product vendors:
  - Data mining
  - Knowledge discovery
  - KDD
- Are these different things? Or are they just different terms of the same idea?

Data Mining and KDD



Catoca Diamond Mine

Locate

Use tools to dig

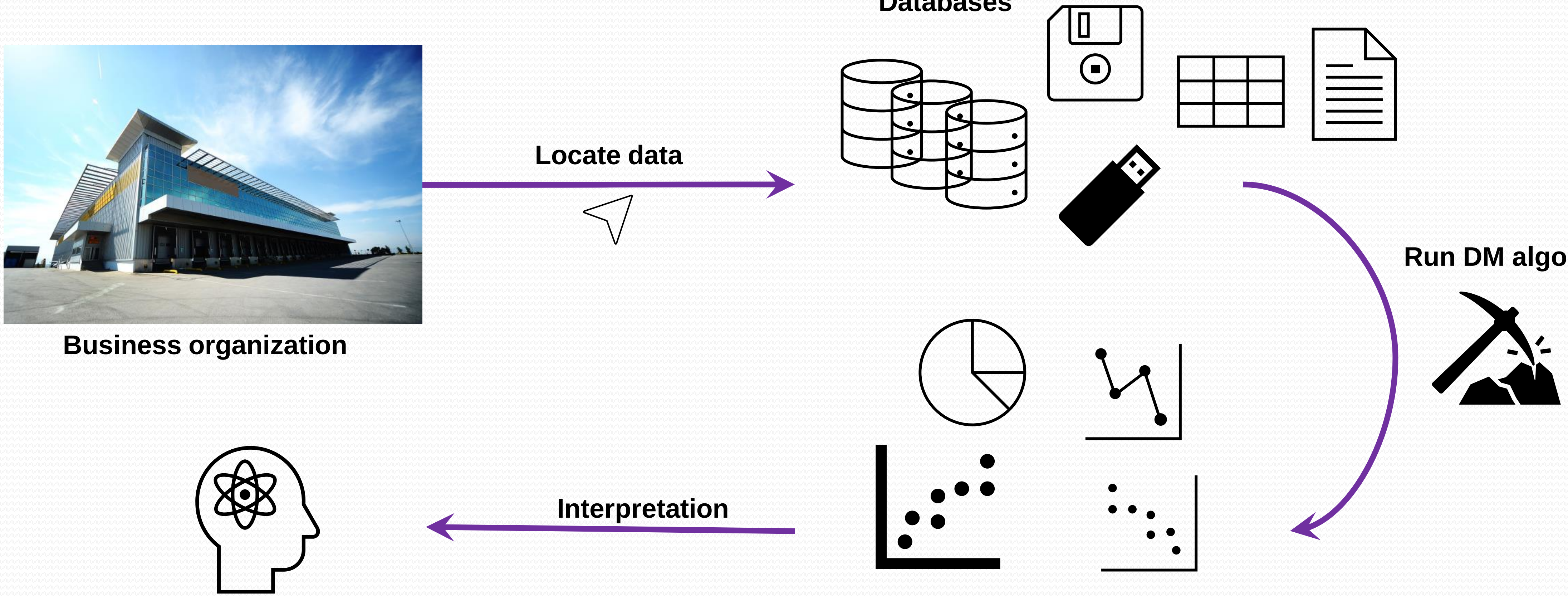
Craft

To craft a diamond ring

Data Mining and KDD

- *KDD* stands for *Knowledge Discovery in Databases*. It is a *process* of discovering useful knowledge from big collections of data
- *Data Mining* is a *step* within the KDD process in which interesting patterns are found. Some of these patterns are then interpreted and transformed into useful knowledge.

Data Mining and KDD



Databases

Locate data

Business organization

Run DM algos

Statistical patterns

Interpretation

Business intelligence and knowledge

Data and Knowledge

- Data: a collection of facts about certain group of objects
- Pattern: certain characteristics of data that are frequently observed
- Knowledge: some general rules about the objects

From data to knowledge

Data

Age	Height / m	Weight / lb	Sex
3	0.40	20	M
2	0.30	16	M
10	1.00	60	M
20	1.43	90	F
15	1.33	75	F
2	0.28	16	F
8	0.71	40	F
10	0.90	57	F
…	…	…	…

Knowledge

a rule: a person younger than 3 usually weighs < 40 lbs.

KDD is a *process*

- KDD is the process of identifying
  - valid
  - novel
  - potentially useful
  - understandable

  patterns in data and deriving knowledge from them.

Data Mining

- A *step* in the whole KDD process in which a specific algorithm is applied to the data for extracting certain specific (interesting) *patterns* in the data.
- Usually, a user has to decide what kinds of patterns he is interested in and to pick a particular data mining algorithm to extract such patterns.
- One more time:
  - KDD – the discovery of knowledge (much human involvement)
  - data mining – the discovery of patterns (mechanical, less human involvement)

An example

- A chain-store supermarket may collect millions of point-of-sale records every day. Each record consists of the items that are purchased together by a customer.
- What knowledge can we extract from such a database?

The KDD process (example)

- Step 1: *Goal setting*
  - If you were the manager of the supermarket, first ask yourself what kind of knowledge you would like to learn from your data.
  - "Hmmm … I am wondering … are there any special groups of items that people usually purchase together in a trip to the supermarket?"

The KDD process (example)

- Step 2: *Data collection*
    - In order to achieve the goal, the manager needs the point-of-sale records generated from the different stores be collected into a database.

The KDD process (example)

- Step 3: *Data transformation (or preparation)*
  - In order to apply an algorithm on the data to discover interesting patterns, we need to transform the data into certain format that is convenient for an algorithm to work on.

Example: How shall we identify the products?

We might want to transform the data so that a single id is used for the same kind of products, regardless of their brands and packaging.

The KDD process (example)

- Step 3: *Data transformation (or preparation)*
  - not all the data is useful, filter away those that are not needed.

    Examples: cash register id, payment method

The KDD process (example)

- Step 4: *Data mining*
  - Apply an algorithm on the data to determine which items are most likely purchased together

Example 1: Customers who buy diapers often buy beer as well.

Example 2: Customers who buy shampoo often buy conditioner as well.

The KDD process (example)

- Step 5: *Result interpretation and evaluation*
  - the manager sifts through the results obtained from the data mining step and selects those that are *actionable*, i.e., those that can be translated into knowledge that works to the advantage of the business.

Diapers & beer ☑

Shampoo & conditioner

The KDD Process

- is an iterative and interactive one
- Iterative – the result obtained from one run through an iteration of the KDD process may not get what you want. Very often, some steps of the process need to be refined, and then the whole process be repeated.
- Interactive – certain amount of human involvement is needed to monitor and to fine tune the steps.

Traditional Database Systems

- Don’t confuse data mining applications with traditional database systems.
- Again, *KDD* is for *knowledge* discovery; *data mining* is for interesting *patterns* discovery (with which knowledge may be derived).
- A traditional *DB* usually only provides the functions of *storing* and *retrieving facts*.

Traditional Database Systems

- A database system that supports simple aggregation functions, for example, should not be considered as providing data mining functionality.
- E.g., A software system that maintains a supermarket database may claim, “Users can *mine* the database to *discover* the total amount of milk purchased last month.”

Q: Do you consider the above software a data mining product?

Prediction and Description

- The knowledge resulting from data mining should carry certain degree of *predictive* ability or *descriptive* (explanatory) ability (or both).

Prediction

- Prediction involves using the database records that describe information about past behavior to automatically generate a model (or rule) that can *predict future behavior*.
- Example (predictive ability): the association between milk and bread can be used to predict that there is a high probability that the next customer buying milk would also buy bread.

Description

- Description involves deriving patterns that *summarize* the underlying *relationships in data* and to describe the characteristics of data.
- Example (descriptive ability): a set of documents could be partitioned into clusters such that documents in each cluster are highly similar. From each cluster, we can identify the most frequently occurring words, which can be interpreted as keywords for a given topic.

DSS & Data Warehouse

- A *decision-support system* (*DSS*) is a system that assists decision makers to make important decisions for an organization or business.
- KDD and data mining are important components in many DSS’s.
- To be effective, a data mining application must have access to organization-wide data. We can integrate departmental databases together into a *Data Warehouse*.

Data Warehousing

- A data warehouse is an integration of various departmental databases so that accesses to organization-wide data is possible.
- A data warehouse is a convenient place where KDD and data mining applications are performed.
- A data warehouse can also be used to support other DSS tools. For example, *On-Line Analytical Processing* (*OLAP*).

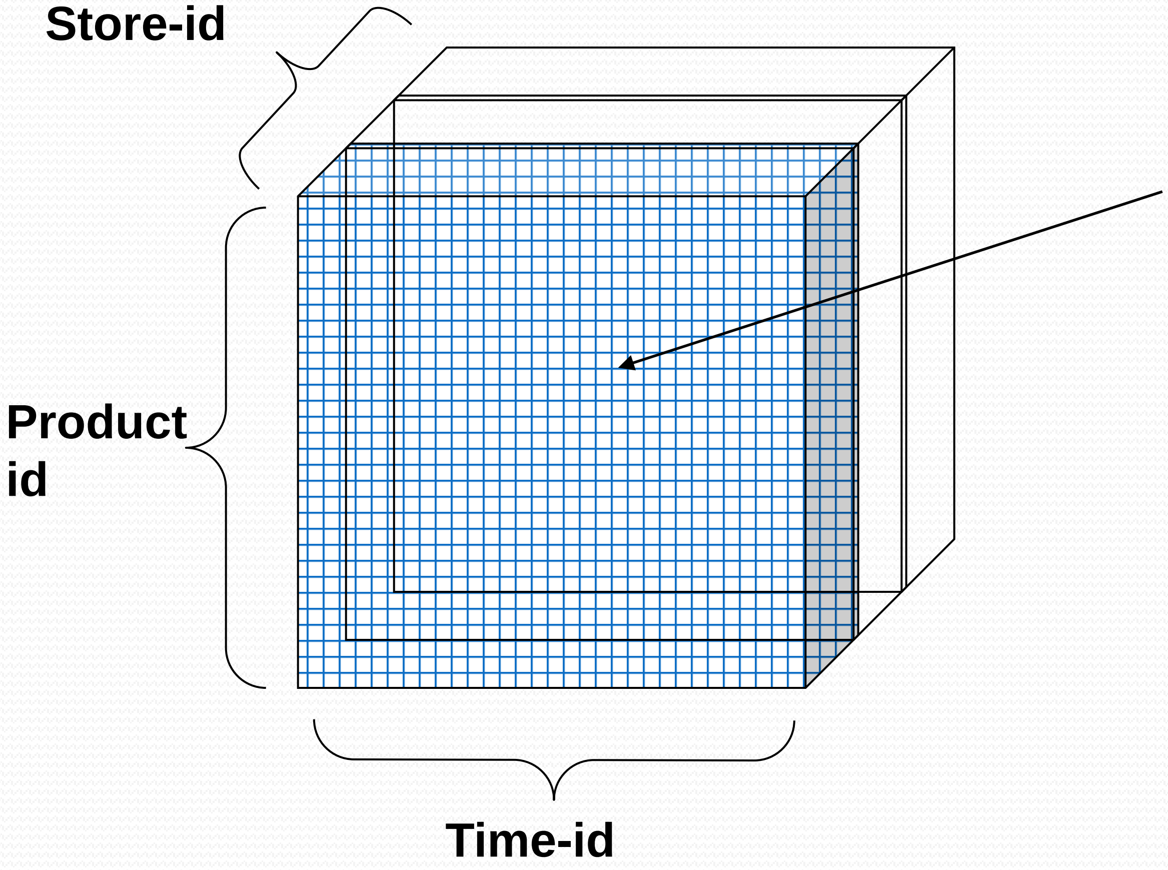
OLAP

A sale record

Toyota	**Corolla**	**Blue**	**$170K**	**Dealer A**	**1/1/2018**

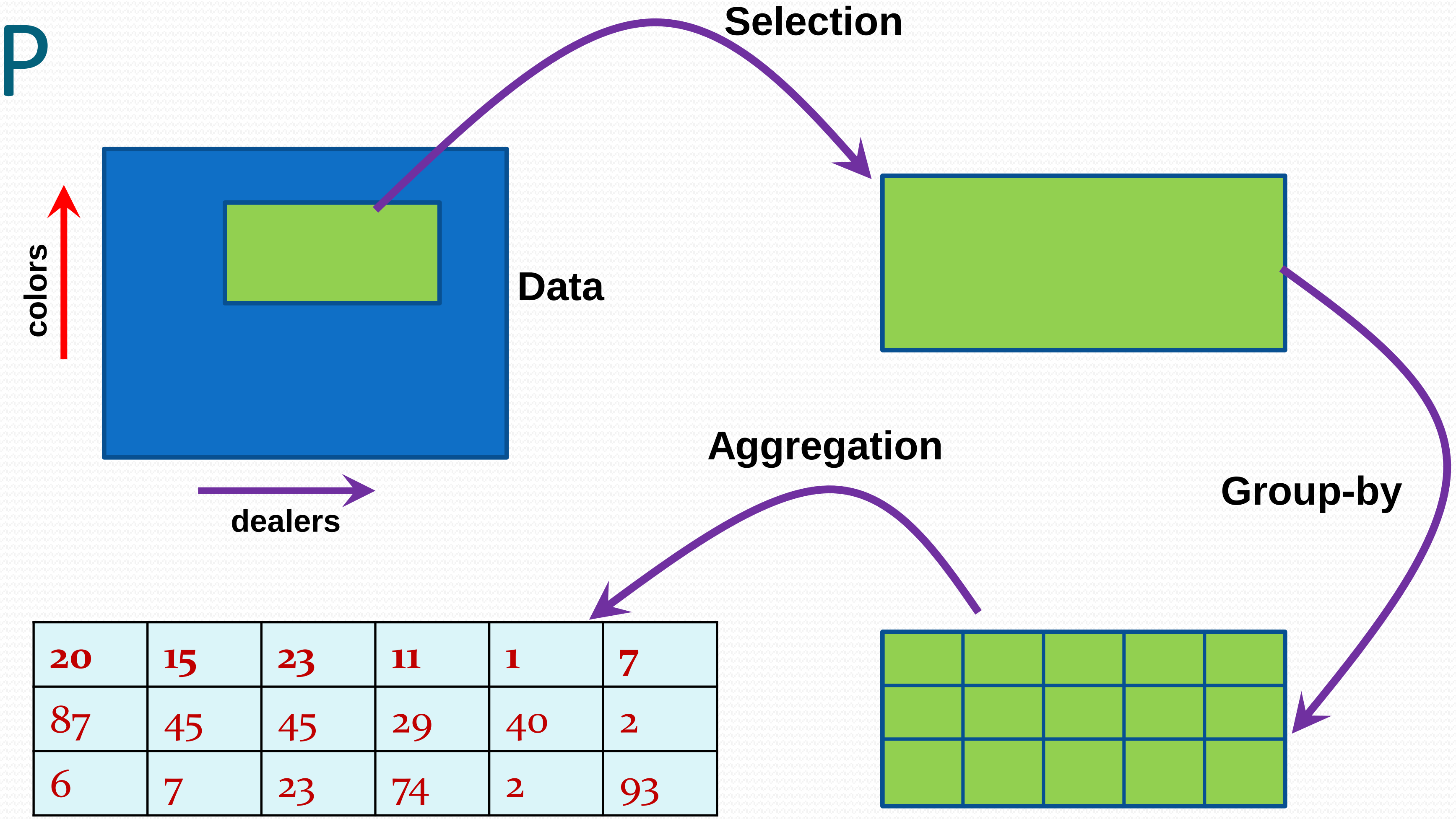
- The OLAP approach allows users to view data in a multi-dimensional model (a *Data Cube*), supporting fast aggregation and summarization operations.
- Example OLAP queries:
  - "What are the *total sale* figures of *blue-colored cars* for *each dealer* in California?"
  - "*How many* computers are sold in the *first quarter of 2000* in *each state of the U.S.*?"
- *Selection* → *Group-by* → *Summarization*

Multi-dimensional Array



each entry in the array stores the number of units sold for a particular <store-id, p-id, time-id> combination.

OLAP



20	**15**	**23**	**11**	**1**	**7**
87	45	45	29	40	2
6	7	23	74	2	93

Data Mining Tasks…

- Classification
- Clustering
- Association Analysis
- Regression Analysis

Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class* (or the *label, dependent variable*).
- Find a *model* that describes the class attribute as a function of the values of other attributes (*independent variables*).
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets. The training set used to build the model and the test set is used to evaluate the model.

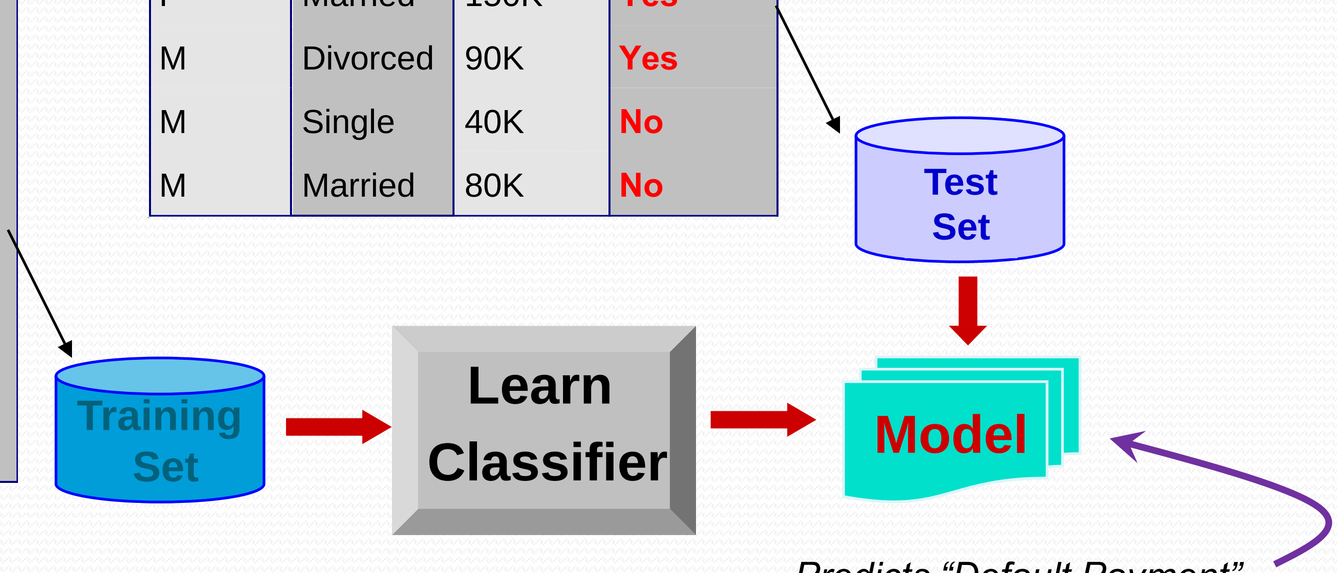
class attribute (label, dependent variable)

Independent attributes

Tid	Gender	Marital Status	Taxable Income	Default Payment
1	M	Single	125K	No
2	M	Married	100K	No
3	M	Single	70K	No
4	F	Married	120K	No
5	F	Divorced	95K	Yes
6	F	Married	60K	No
7	M	Divorced	220K	No
8	M	Single	85K	Yes
9	F	Married	75K	No
10	M	Single	90K	Yes

Credit holder data

Gender	Marital Status	Taxable Income	Default Payment
M	Single	75K	No
F	Married	50K	Yes
F	Married	150K	Yes
M	Divorced	90K	Yes
M	Single	40K	No
M	Married	80K	No



Predicts “Default Payment” based on a record’s independent attribute values

Classification: Application

Occupation	Income	Age	…	Buy?
				Y
				N
…	…	…	…	…

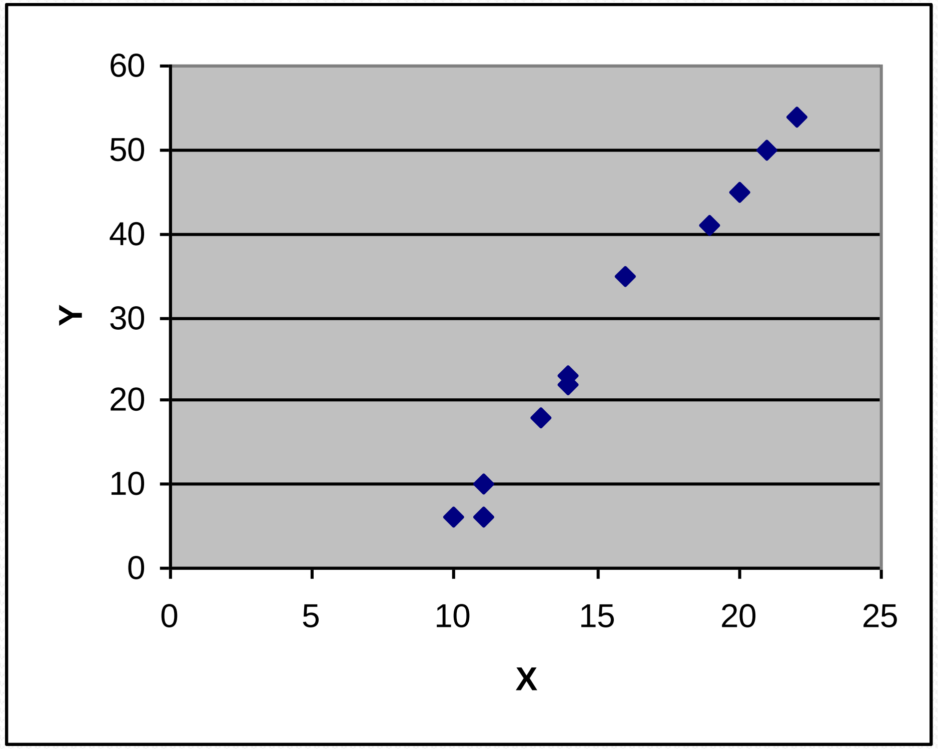
- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This {*buy, don't buy*} decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - occupation, where they live, how much they earn, gender, age, etc.
    - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

Regression

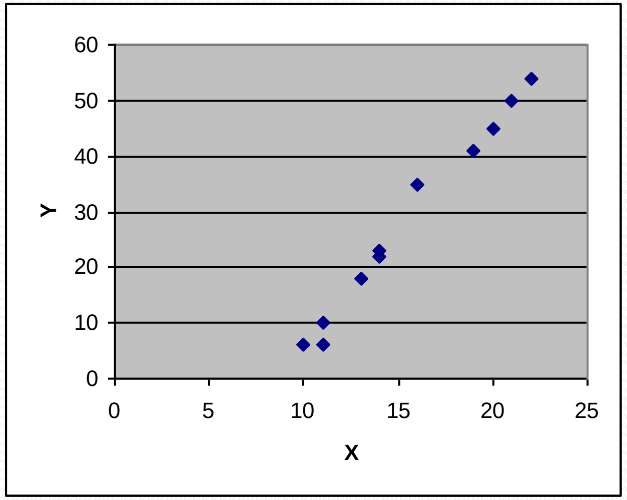
- Predict a value of a *numerical variable* based on the values of other variables.
- Extensively studied in statistics.
- Examples:
  - Predicting sales amounts of a new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.

Regression

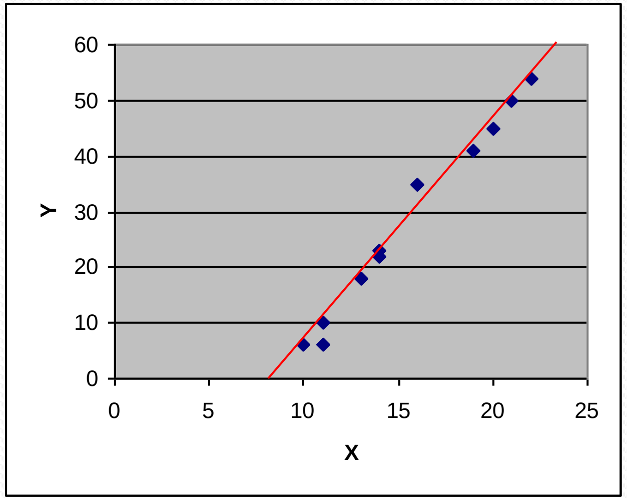


hmmm … looks like *X* and *Y* are linearly correlated

Regression



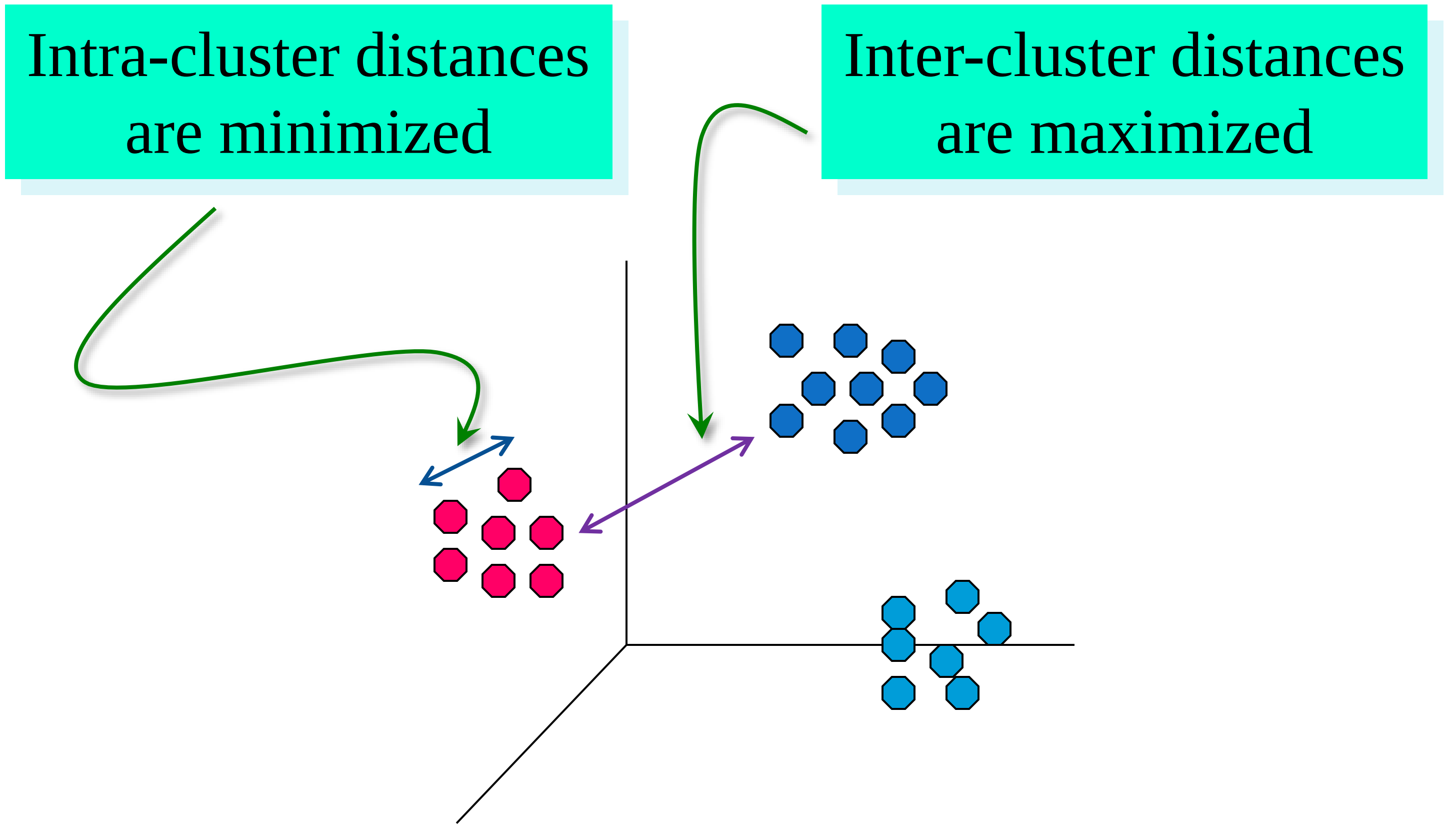
hmmm … looks like *X* and *Y* are linearly correlated



Clustering Definition

- Given a set of data objects, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Objects in one cluster are more similar to one another.
  - Objects in separate clusters are less similar to one another.
- Typically, cluster analysis requires a user to define a similarity measure between records. Clustering is then performed based on the principle of maximizing the intra-cluster similarity and minimizing the inter-cluster similarity (*distance-based clustering*).

Illustrating Clustering



Intra-cluster distances are minimized

Inter-cluster distances are maximized

Clustering: Application

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms they contain.
  - Approach:
    - Identify frequently occurring terms in each document.
    - Form a similarity measure based on the frequencies of different terms.
    - Apply a clustering algorithm.
  - Applications:
    - Discover topics of interests
    - Recommend relevant documents (e.g., Google Now news feeds).

Association Rule Discovery: Definition

- Given a set of records each of which contains some items from a given collection;
  - Produce dependency rules which predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rule Discovery: Application

- Marketing and Sales Promotion:
  - Let the rule discovered be

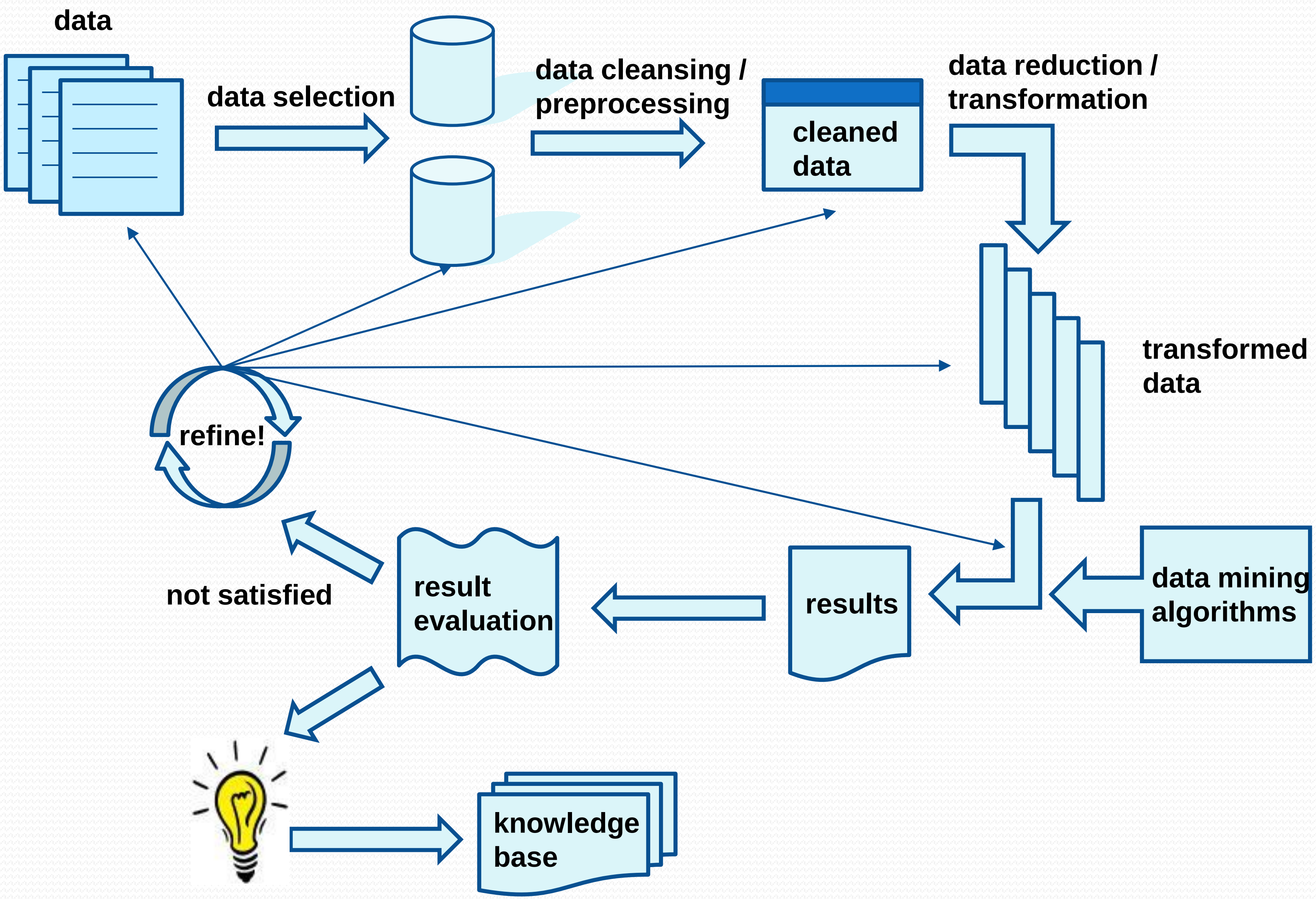
    $\{Bagels, Mayonnaise\} \rightarrow \{Potato\ Chips\}$
  - *Potato Chips as consequent* ⇒ Can be used to determine what should be done to boost its sales.
  - *Bagels in the antecedent* ⇒ Can be used to see which products would be affected if the store discontinues selling bagels.

Sequence Analysis

- A sequence database contains sequences of events. Sequence analysis is about finding interesting, frequently occurring (sub)sequences to predict future behavior.
- Example: renting movies, buying habits, web serving behavior, web log analysis.

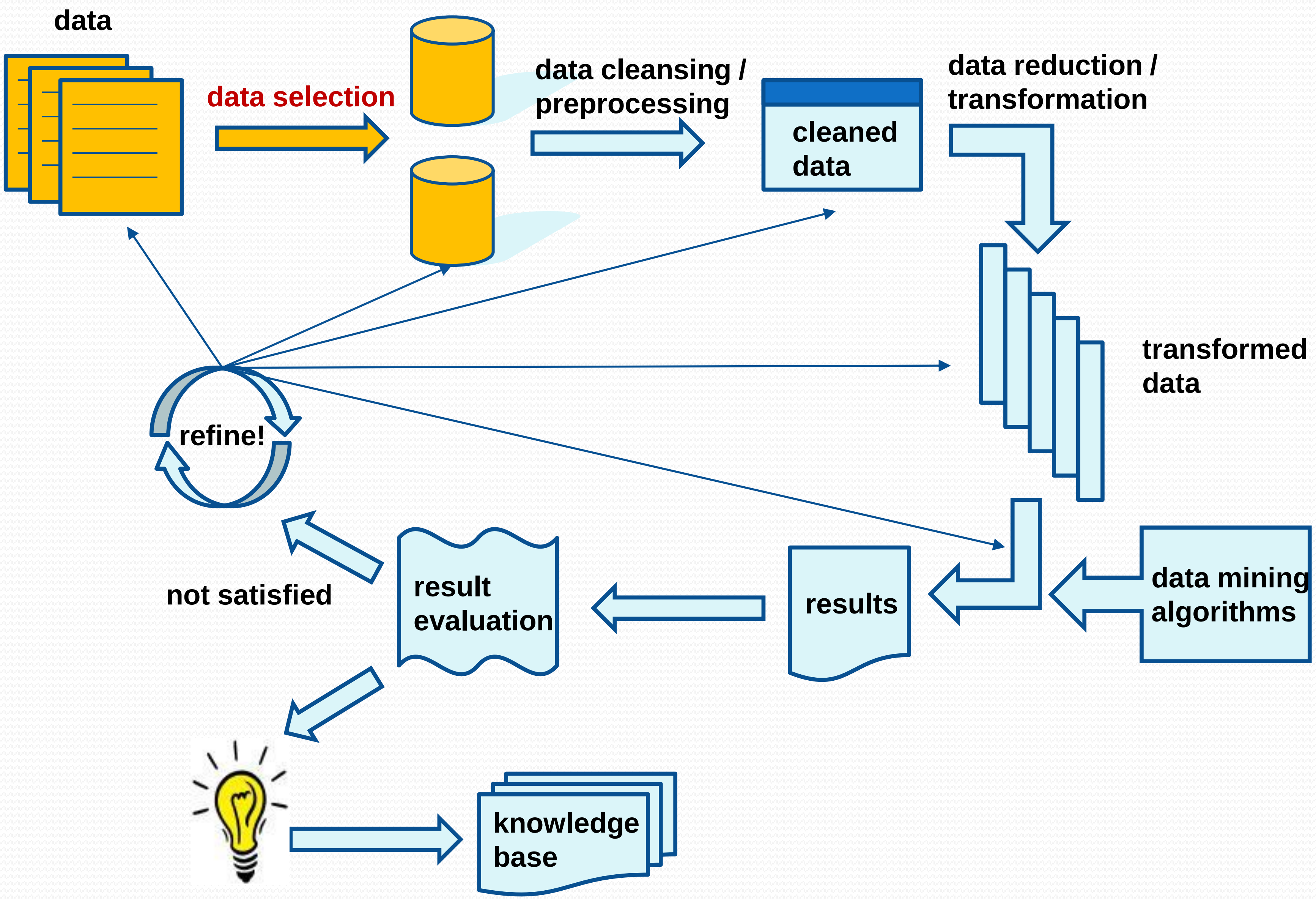
KDD Process

- The process of discovering knowledge from databases. It involves a number of steps.
- The process is typically interactive and iterative.


data
data selection
data cleansing / preprocessing
cleaned data
data reduction / transformation
transformed data
data mining algorithms
results
result evaluation
not satisfied
refine!
knowledge base

Step 1: Goal Identification

- Understand your application domain
- Obtain prior known knowledge to help the discovery process and to help set up the goal
- e.g., credit card company:
    - business: hard to retain customers
    - goal: to find out what characteristics defecting customers share

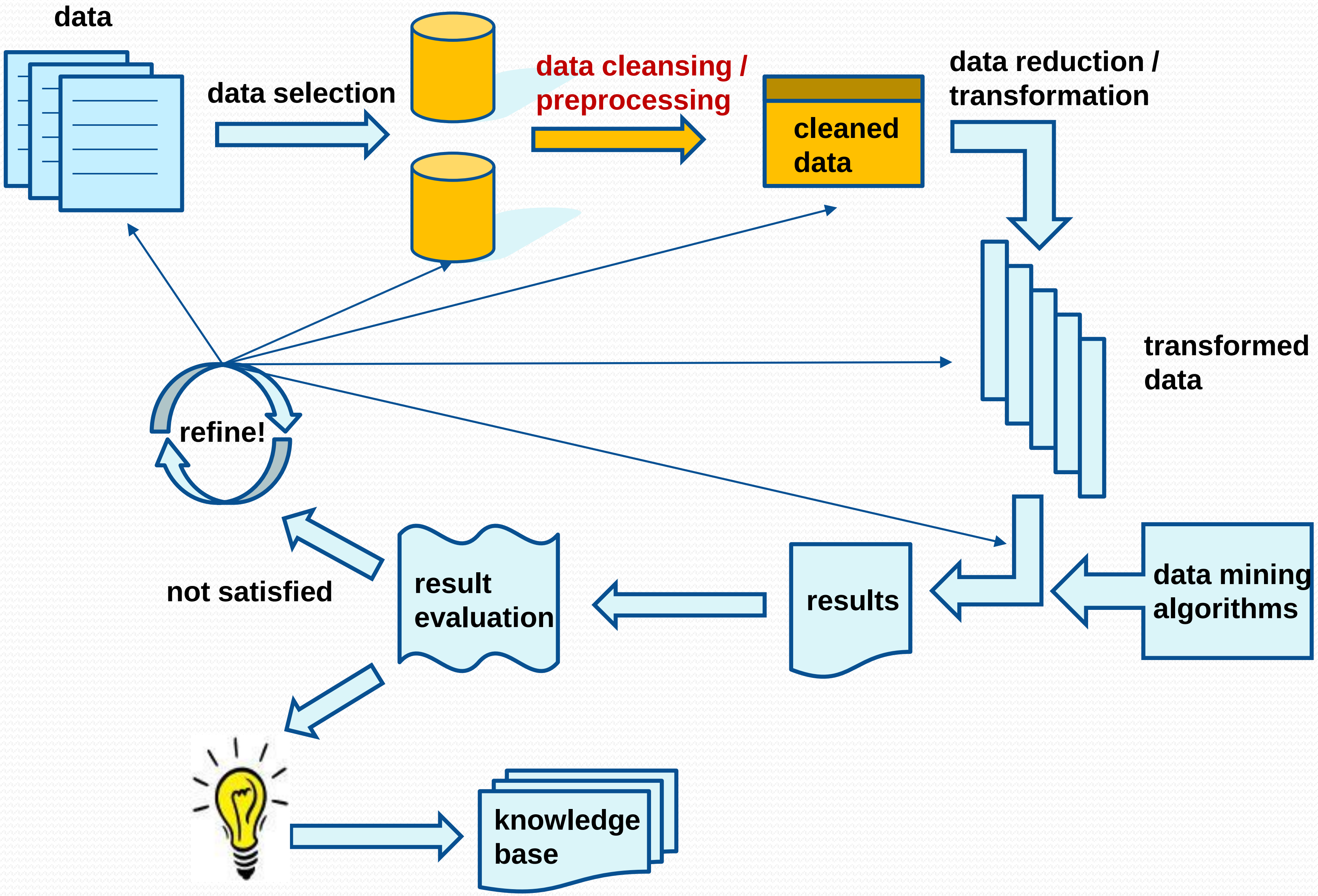

data
data selection
data cleansing /
preprocessing
cleaned
data
data reduction /
transformation
transformed
data
refine!
data mining
algorithms
results
result
evaluation
not satisfied
knowledge
base

Step 2: Data Collection and Selection

- What are some of the characteristics that would possibly affect how likely a customer would defect?
  - name? address? age? salary level? education?
- Where can I find all this data?
- If the data is scattered among various databases, shall I find a single place to store them (for later mining)? A data warehouse?

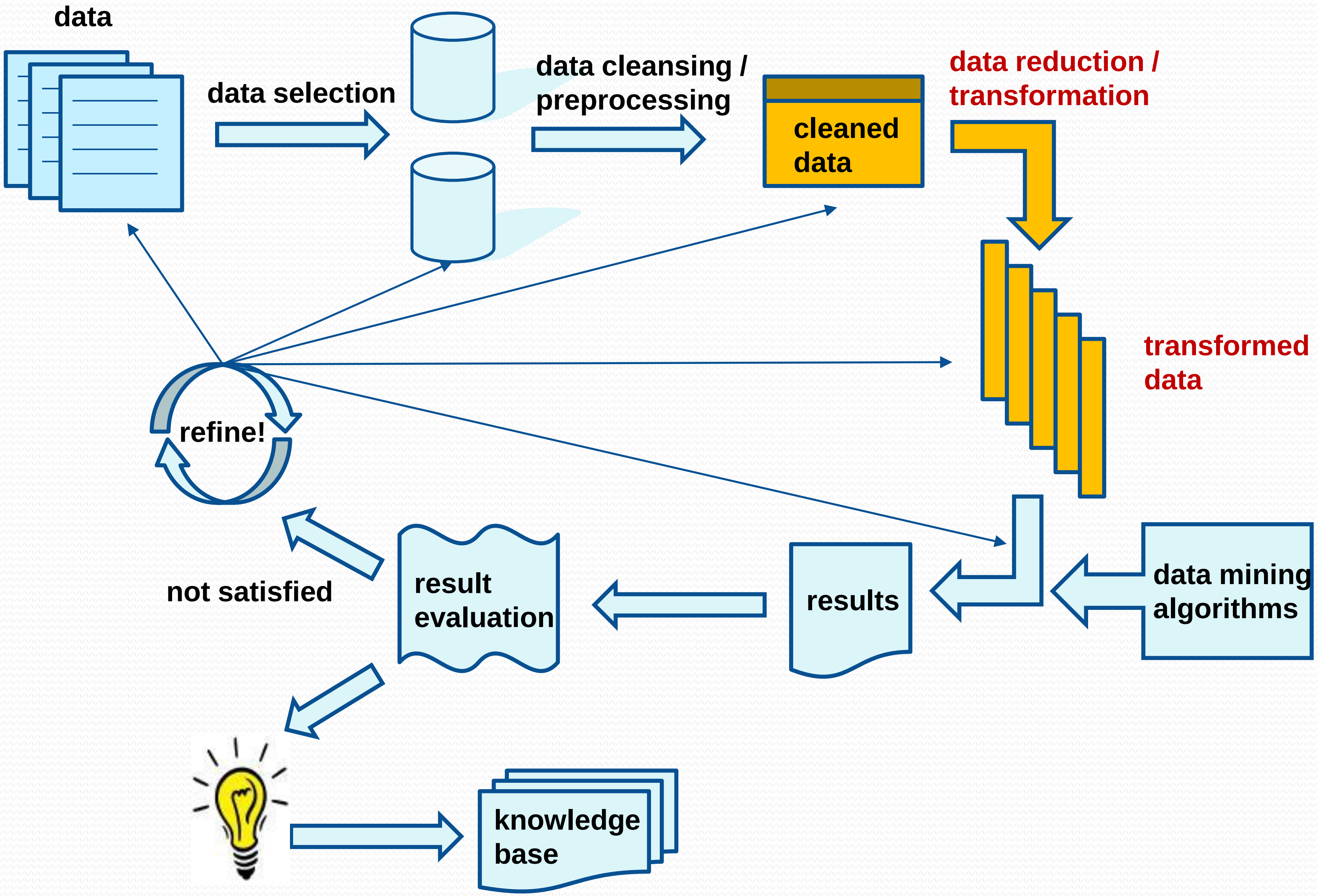
Step 2: Data Collection and Selection

- What are some of the characteristics that would possibly affect how likely a customer would defect?
  - name? address? age? salary level? education?
- Where can I find all this data?
- If the data is scattered among various databases, shall I find a single place to store them (for later mining)? A data warehouse?


data
data selection
data cleansing / preprocessing
cleaned data
data reduction / transformation
transformed data
data mining algorithms
results
result evaluation
not satisfied
refine!
knowledge base

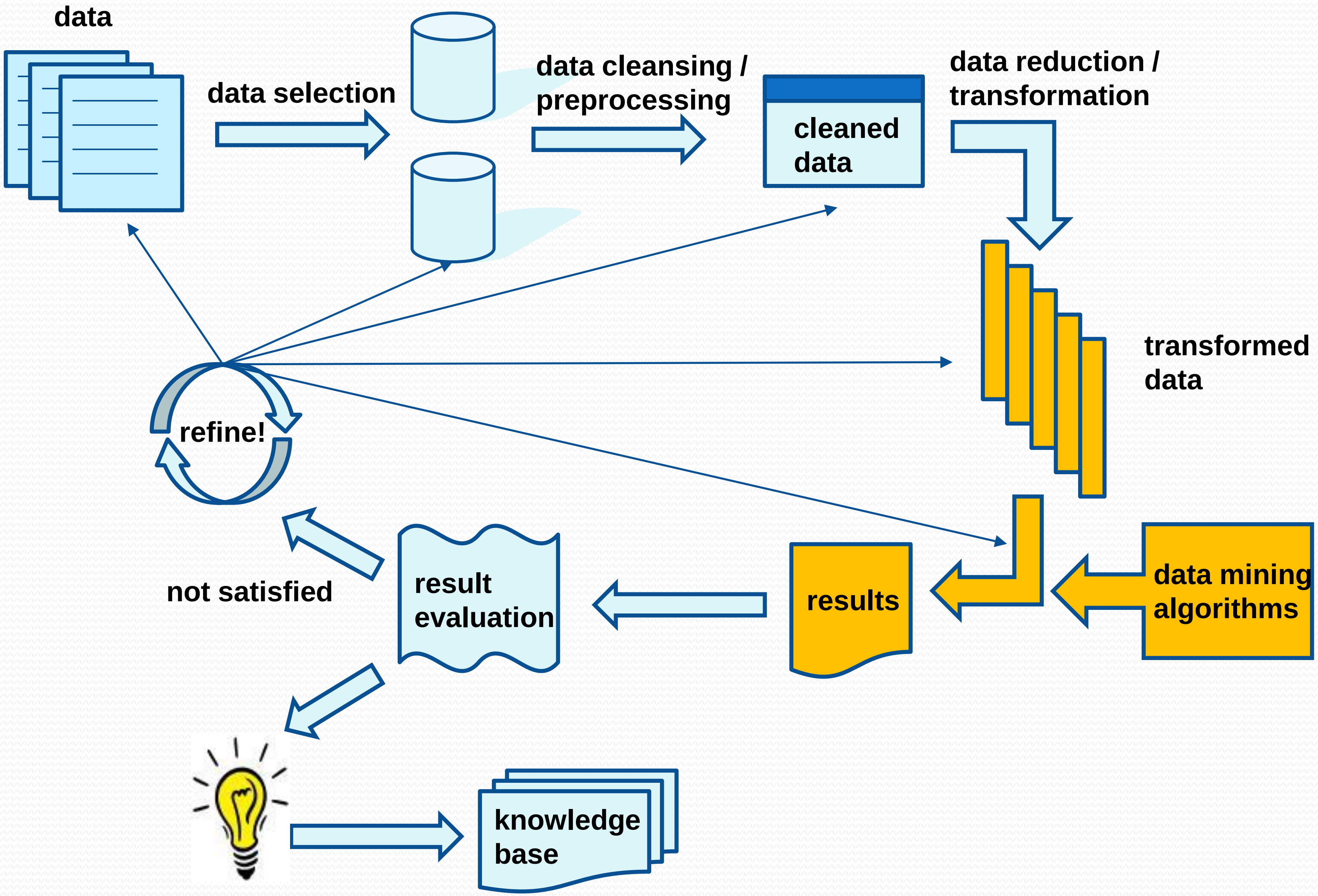
Step 3: Data Cleansing and Preprocessing

- Business data is usually incomplete with missing (and incorrect) values
    - e.g., operator errors, data entry errors, system measurement failures, database schema revision
- Missing data has to be either (1) thrown away, or (2) filled in
- To avoid incorrect data from affecting the performance of the KDD process, *noise* (or *outliers*) has to be detected


data
data selection
data cleansing / preprocessing
cleaned data
data reduction / transformation
transformed data
refine!
data mining algorithms
results
result evaluation
not satisfied
knowledge base

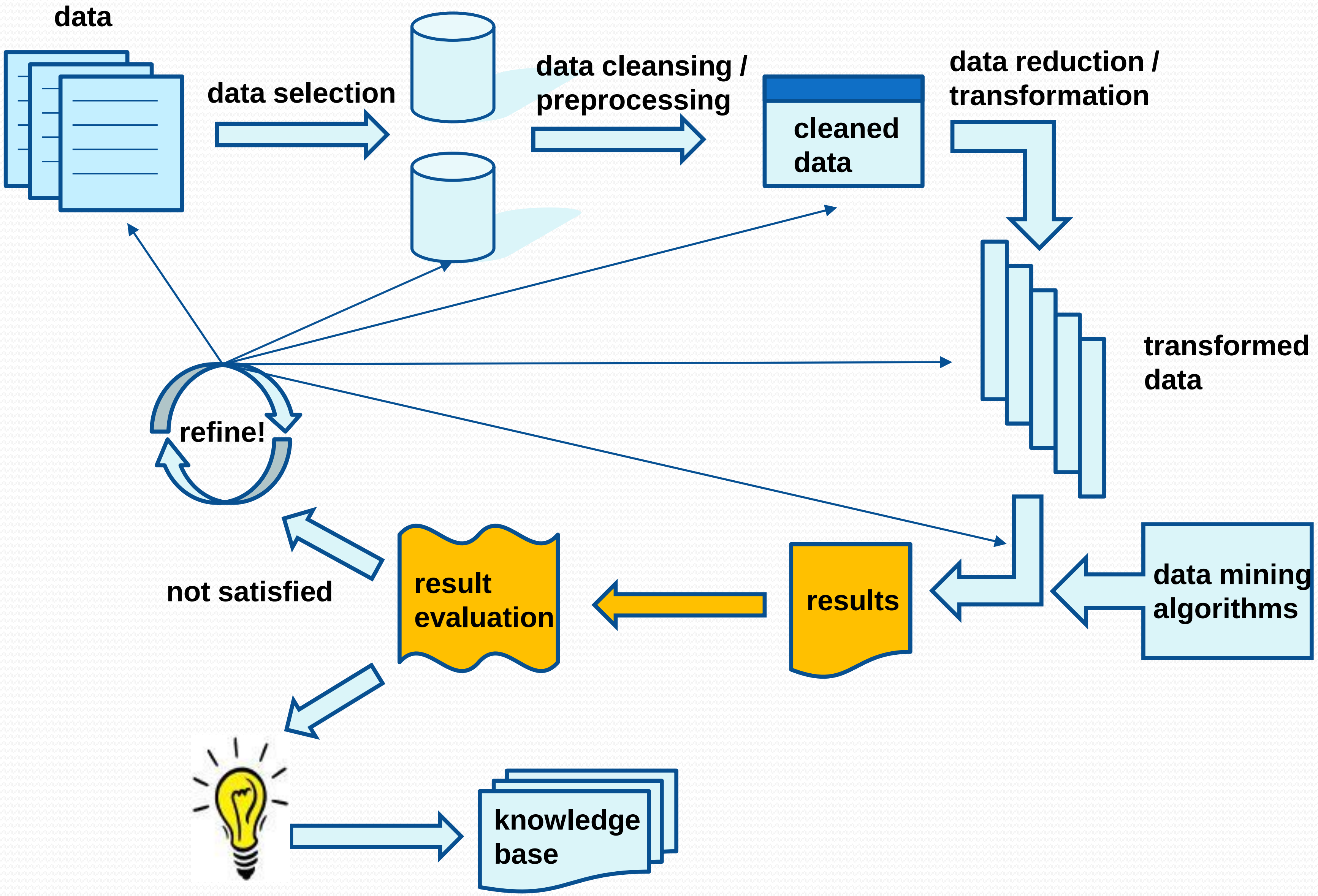
Step 4: Data Reduction and Transformation

- Compact data into a form that is less bulky than the original raw data
- Improve the efficient of the data mining algorithm
- e.g., data summarization, attribute subset selection


data
data selection
data cleansing / preprocessing
cleaned data
data reduction / transformation
transformed data
refine!
data mining algorithms
results
result evaluation
not satisfied
knowledge base

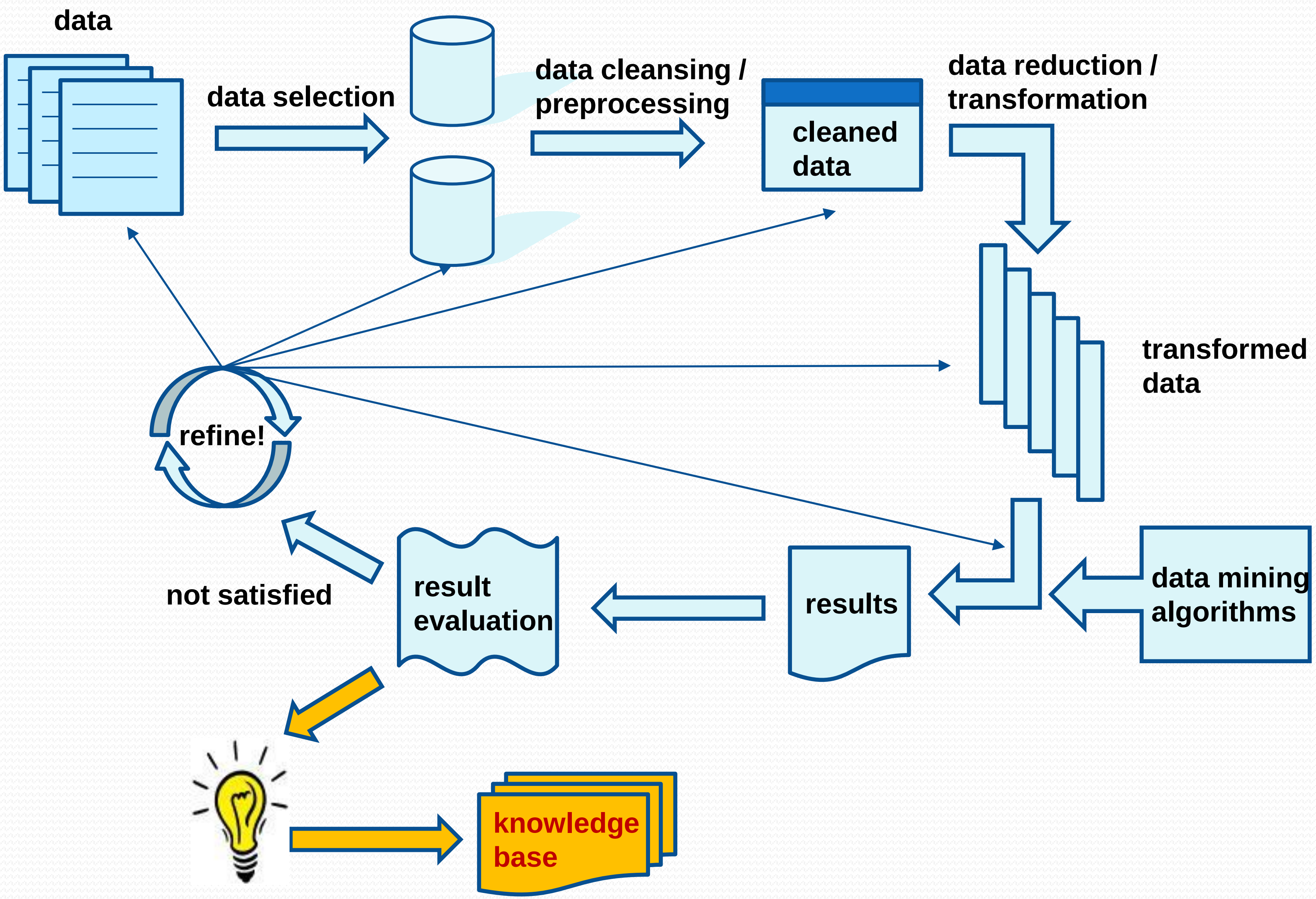
Steps 5-7: Data Mining

- Step 5: pick a data mining model
  - association rules? classification? sequence analysis?
- Step 6: pick a data mining algorithm
  - the Apriori algorithm, decision-tree classifier, neural-network classifier
- Step 7: apply the algorithm to the data


data
data selection
data cleansing / preprocessing
cleaned data
data reduction / transformation
transformed data
data mining algorithms
results
result evaluation
not satisfied
refine!
knowledge base

Step 8: Result Evaluation

- check if the results obtained satisfy your goals
- if not, refine some of the steps and re-run


data
data selection
data cleansing /
preprocessing
cleaned
data
data reduction /
transformation
transformed
data
refine!
data mining
algorithms
results
result
evaluation
not satisfied
knowledge
base

Step 9: Knowledge Consolidation

- document and report the useful knowledge to the users

Challenges of Data Mining

- Technical challenges
  - scalability
  - dimensionality
  - data stream
- Data challenges
  - complex and heterogeneous data
  - data quality
- Legal challenges
  - data ownership and distribution
  - privacy protection and regulations (e.g., GDPR)
  - Algorithmic biases
- Results
  - interpretation of patterns
  - Explanability

Privacy

- The “*fair information practice*”

“The primary purpose of the collection must be clearly understood by the consumer and identified at the time of the collection. Data mining, however, is a secondary, future use. As such it requires the explicit consent of the data subject or consumer.”

Summary

- Data Mining aims at finding non-trivial and useful patterns from large-scale data
- Data Mining is part of a larger process called “Knowledge Discovery in Databases (KDD)”
  - other parts include data preprocessing, integration, selection, cleaning, and also interpretation of mined patterns
- Several data mining tasks
  - classification, clustering, association analysis, etc.