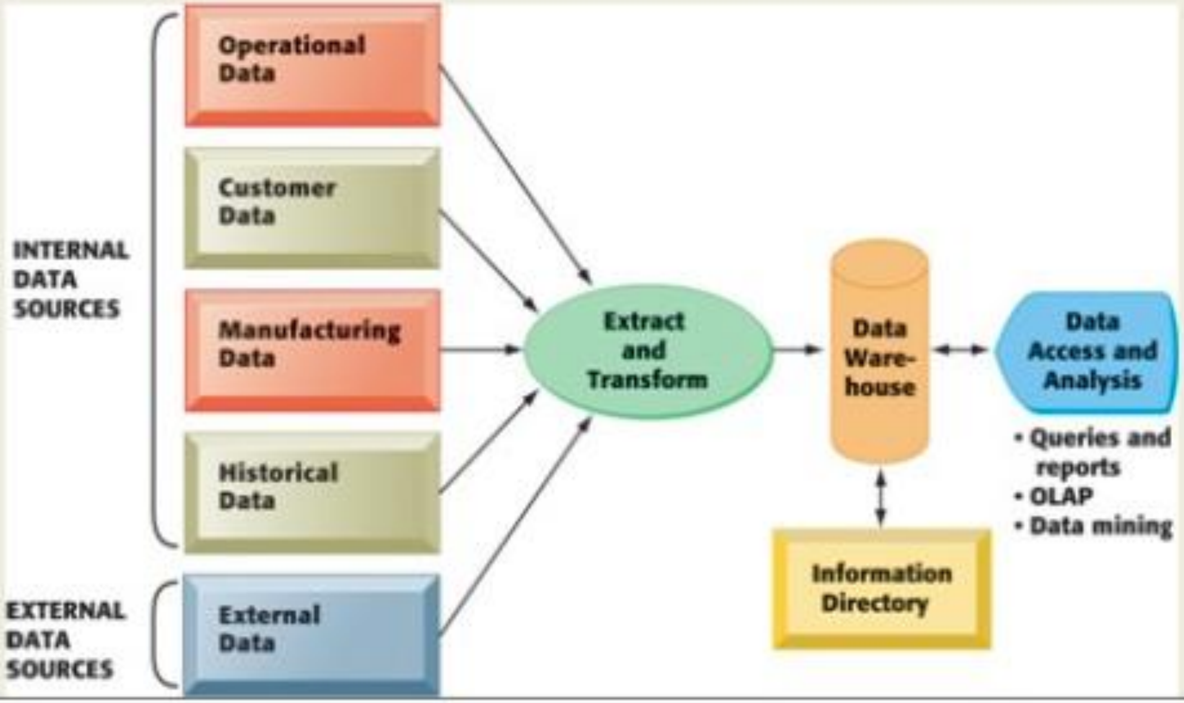
Introduction to Data Warehouse and Crisp-DM

DATABASE, DATA WAREHOUSE, DATA MART, DATA SET…?

- A database is an organized grouping of information within a specific structure. Most databases in use today are relational databases—they are designed using many tables which relate to one another in a logical fashion. Relational databases generally contain dozens or even hundreds of tables, depending upon the size of the organization.

Data Warehouse

- Many organizations need internal, external, current, and historical data
- Data Warehouse are designed to, typically, store and manage data from operational transaction systems, Web site transactions, etc.

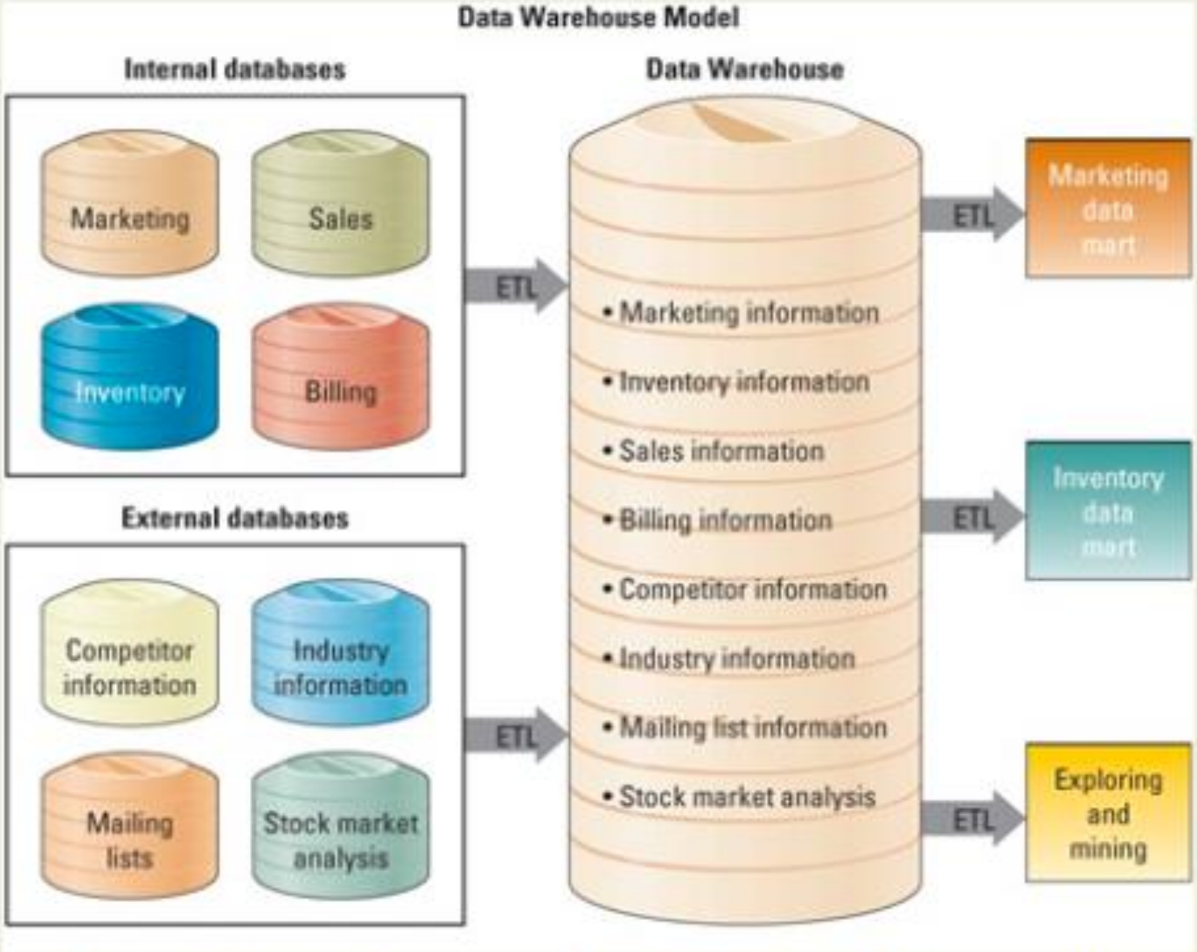


Data Warehouse Fundamentals

- ***Data warehouse*** - a logical collection of information – gathered from many different operational databases – that supports business analysis activities and decision-making tasks

- The primary purpose of a data warehouse is to aggregate information throughout an organization into a single repository for decision-making purposes

Data Warehouse Fundamentals



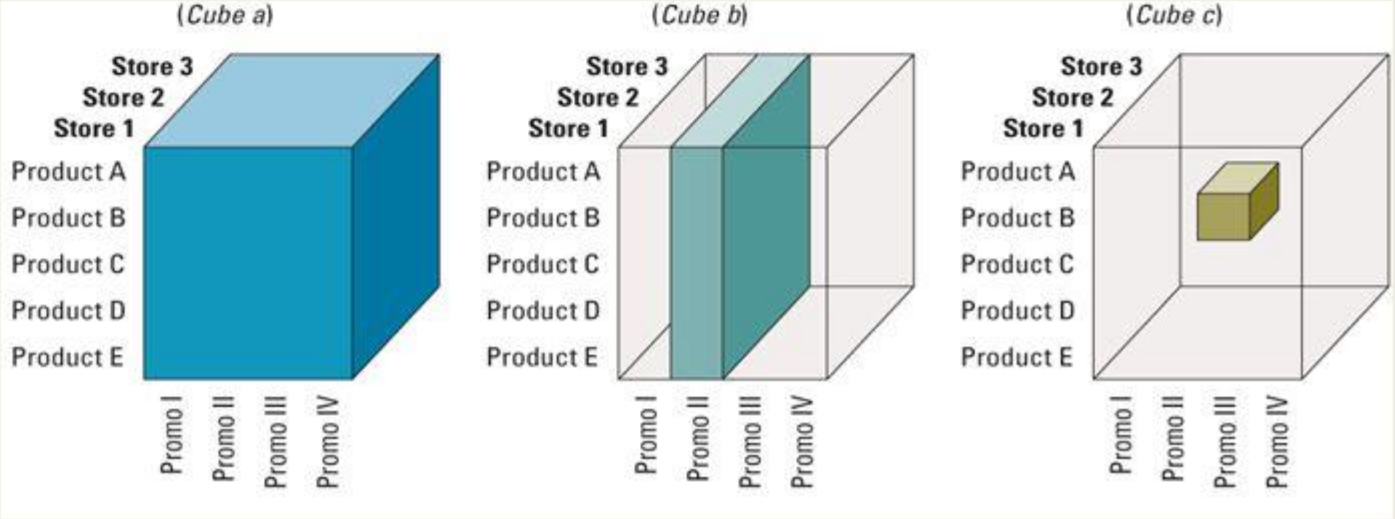
- **Extraction, transformation, and loading (ETL)** a process that extracts information from internal and external databases, transforms the information using a common set of enterprise definitions, and loads the information into a data warehouse.

Data Mart

- Subset of data warehouses that is highly focused and isolated for a specific population of users
- Example: Marketing data mart, Sales data mart, etc.

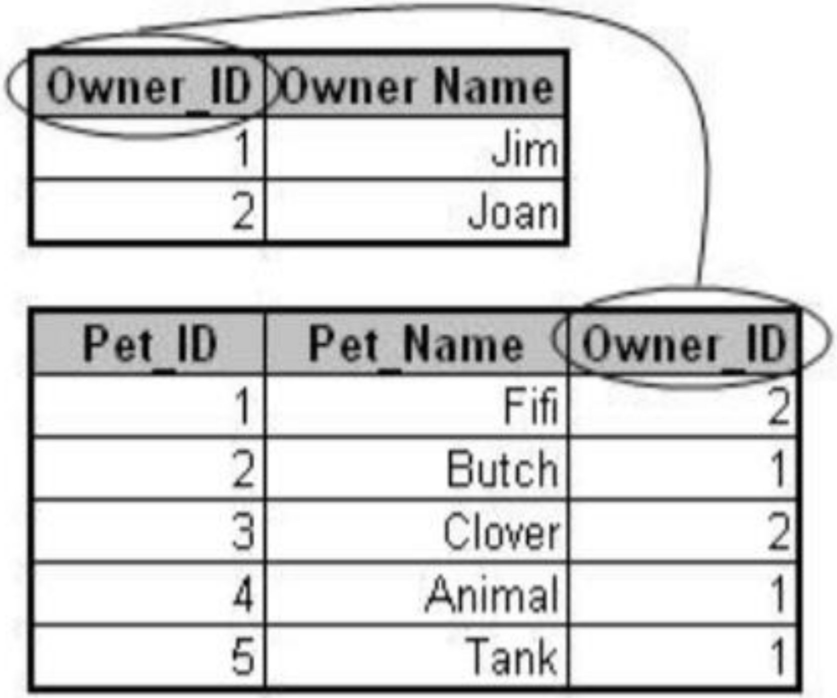
Database vs. Data Warehouse

- Databases contain information in a series of two-dimensional tables
- In a Data Warehouse and data mart, information is multidimensional, it contains layers of columns and rows



	Operational Database	Data Warehouse
Purpose	For data retrieval, updating and management	For data analysis and decision making
Systems/ Applications	OLTP (Online Transaction Processing System)	Analytical Software like Data Mining Tools, Reporting Tools and OLAP tools
Format	■ Normalised ■ Relational Database ■ Lowest level of granularity (e.g. individual transactions)	■ Denormalised and integrated ■ Multi-dimensional arrays or relational format ■ Subject-Oriented ■ Granularity level depends on subject
Time Frame	Current / Real-Time	Historical

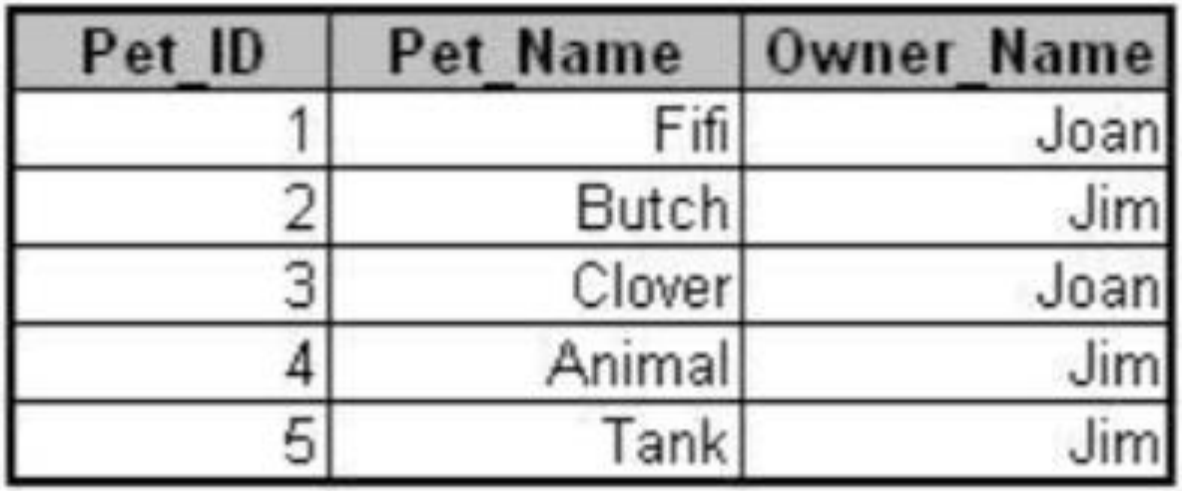
Relational Database



Owner_ID	Owner Name
1	Jim
2	Joan

Pet_ID	Pet_Name	Owner_ID
1	Fifi	2
2	Butch	1
3	Clover	2
4	Animal	1
5	Tank	1

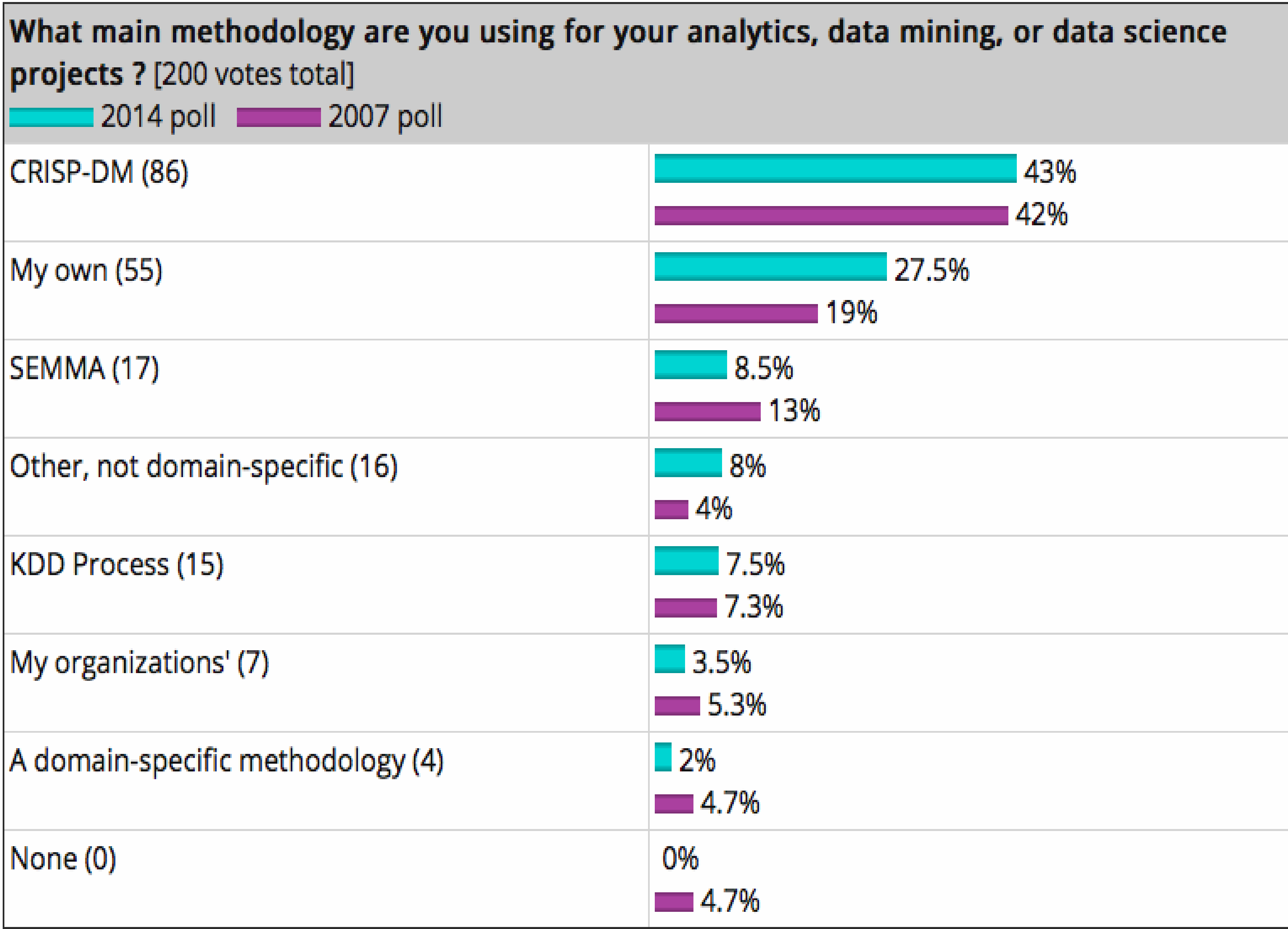
Denormalization



Pet_ID	Pet_Name	Owner_Name
1	Fifi	Joan
2	Butch	Jim
3	Clover	Joan
4	Animal	Jim
5	Tank	Jim

Dataset

- A data set is a subset of a database or a data warehouse. It is usually denormalized so that only one table is used. The creation of a data set may contain several steps, including appending or combining tables from source database tables, or simplifying some data expressions.
- Data sets may be made up of a representative sample of a larger set of data, or they may contain all observations relevant to a specific group.



What main methodology are you using for your analytics, data mining, or data science projects ? [200 votes total]

🟦 2014 poll 🟪 2007 poll

CRISP-DM (86)	43%	42%
My own (55)	27.5%	19%
SEMMA (17)	8.5%	13%
Other, not domain-specific (16)	8%	4%
KDD Process (15)	7.5%	7.3%
My organizations' (7)	3.5%	5.3%
A domain-specific methodology (4)	2%	4.7%
None (0)	0%	4.7%

CRISP-DM

CRoss-Industry Standard Process

for Data Mining

Why Should There be a Standard Process?

The data mining process must be **reliable and repeatable by people with little data mining background.**

Process Standardization

- **Initiative launched in late 1996 by three "veterans" of data mining market.**

  Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) , NCR

- **Developed and refined through series of workshops** (from 1997-1999)
- **Over 300 organization contributed to the process model**
- **Published CRISP-DM 1.0** (1999)
- **Over 200 members of the CRISP-DM SIG worldwide**

  - **DM Vendors** - SPSS, NCR, IBM, SAS, SGI, Data Distilleries, Syllogic, etc.

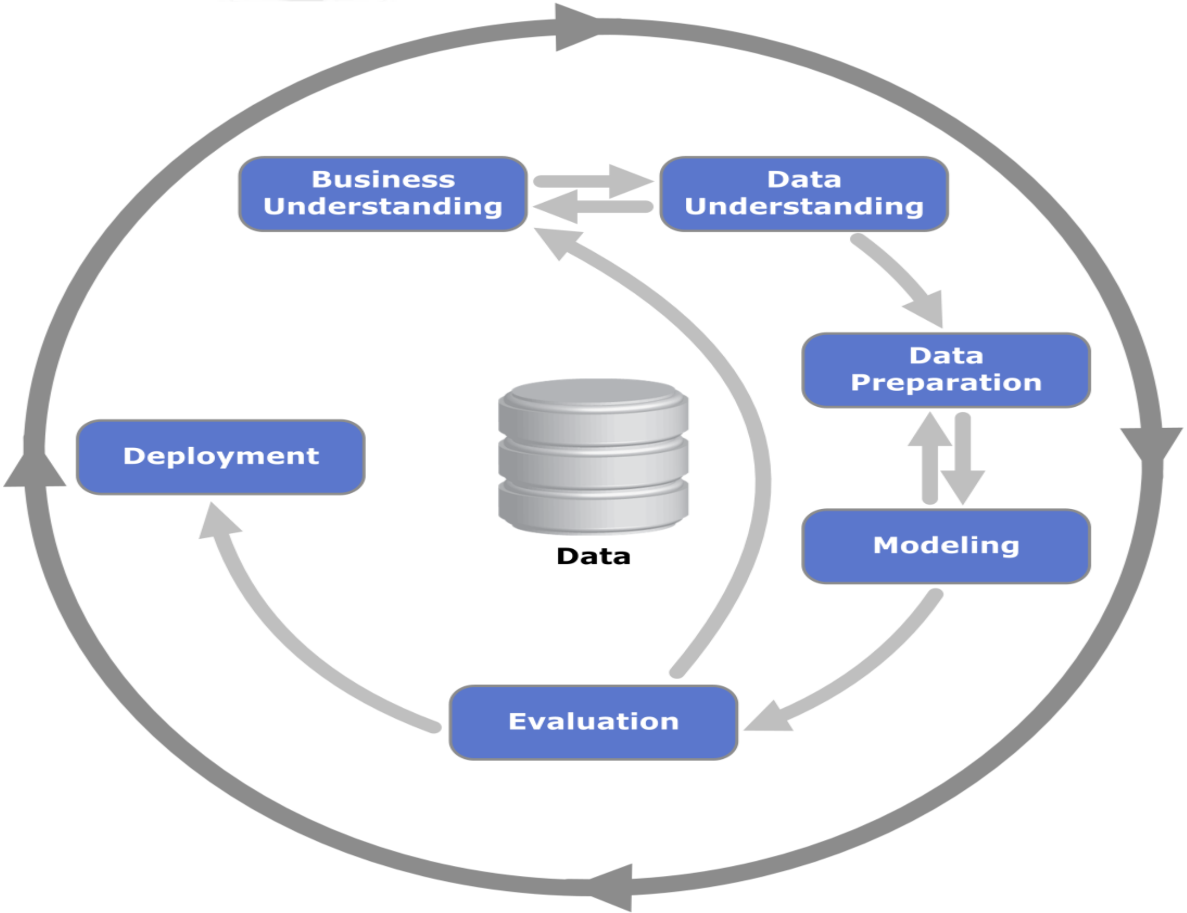
  - **System Suppliers / consultants** - Cap Gemini, ICL Retail, Deloitte & Touche, etc.

  - **End Users** - BT, ABB, Lloyds Bank, AirTouch, Experian, etc.

CRISP-DM

- Non-proprietary
- Application/Industry neutral
- Tool neutral
- Focus on business issues
  - As well as technical analysis
- Framework for guidance
- Experience base
  - Templates for Analysis

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Why CRISP-DM?

- The data mining process must be reliable and repeatable by people with little data mining skills

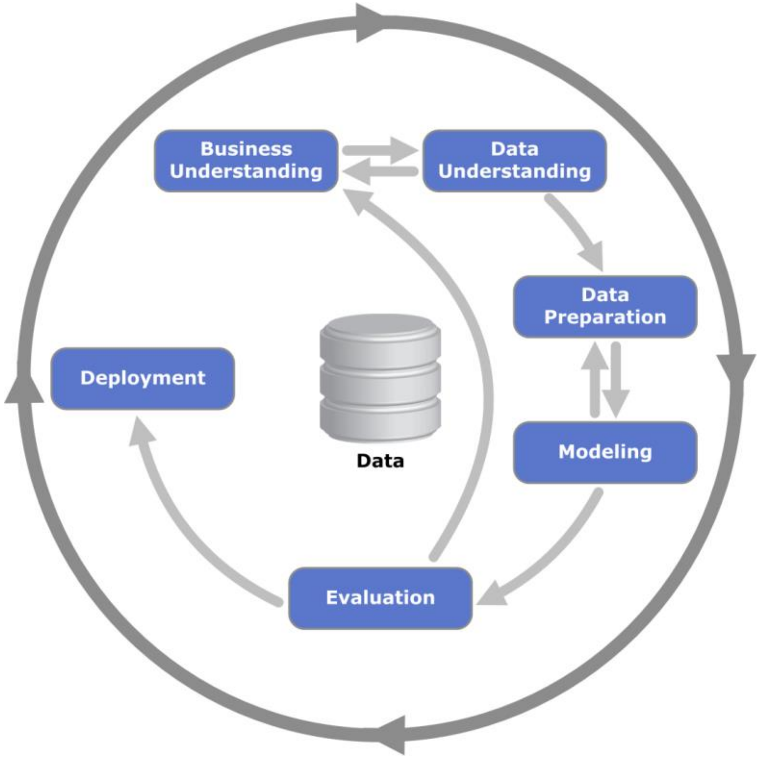
- CRISP-DM provides a uniform framework for
  - guidelines
  - experience documentation

- **Aid to project planning and management "Comfort factor" for new adopters**
  - Demonstrates maturity of Data Mining
  - Reduces dependency on "stars"

Source:- State University of New York & UC Berkerly School of Information

CRISP-DM

- Non-proprietary
- Application/Industry neutral
- Tool neutral
- Focus on business issues
  - As well as technical analysis
- Framework for guidance
- Experience base
  - Templates for Analysis



Phases and Tasks

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Situation Assessment
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goal
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Asessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Data Set
Data Set Description

Select Data
Rationale for Inclusion / Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Modeling

Select Modeling Technique
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Description

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

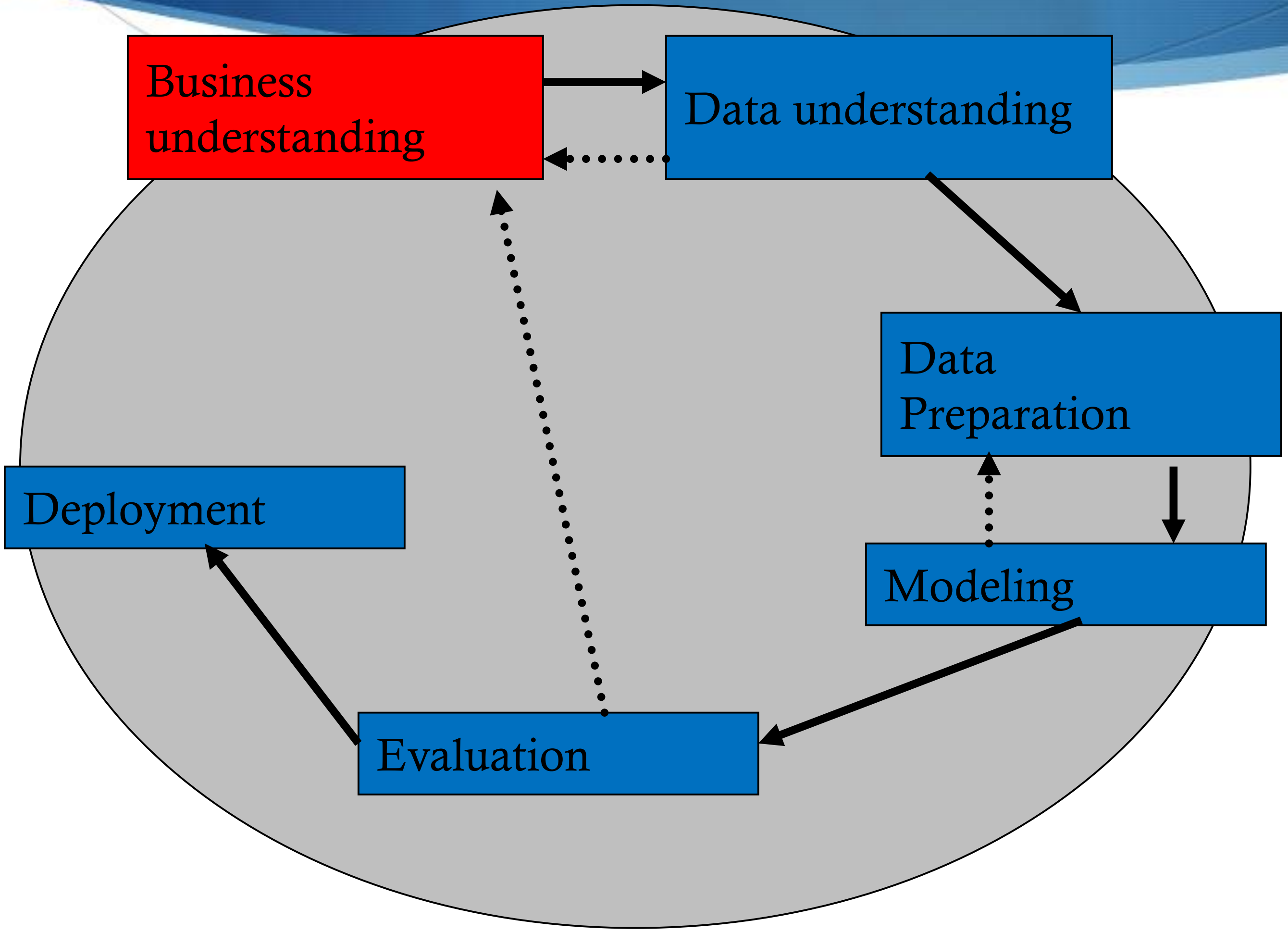
Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

Phase 1:- Business Understanding



TASKS

Business objective

Assess situation

Data mining goals

Project plan

Phase 1:- Business Understanding

- **Determine business objectives**

- **thoroughly understand, from a business perspective, what the client really wants** to accomplish

- **uncover important factors**, at the beginning, that can influence the outcome of the project

- neglecting this step is to expend a great deal of effort producing the right answers to the wrong questions

- **Assess situation**

- **more detailed fact-finding about all of the resources, constraints, assumptions and other factors** that should be considered

- flesh out **the details**

Phase 1:- Business Understanding

- **Determine data mining goals**

- a business goal states **objectives in business terminology**

- a data mining goal states **project objectives in technical terms**

ex:-

the business goal: "Increase catalog sales to existing customers."

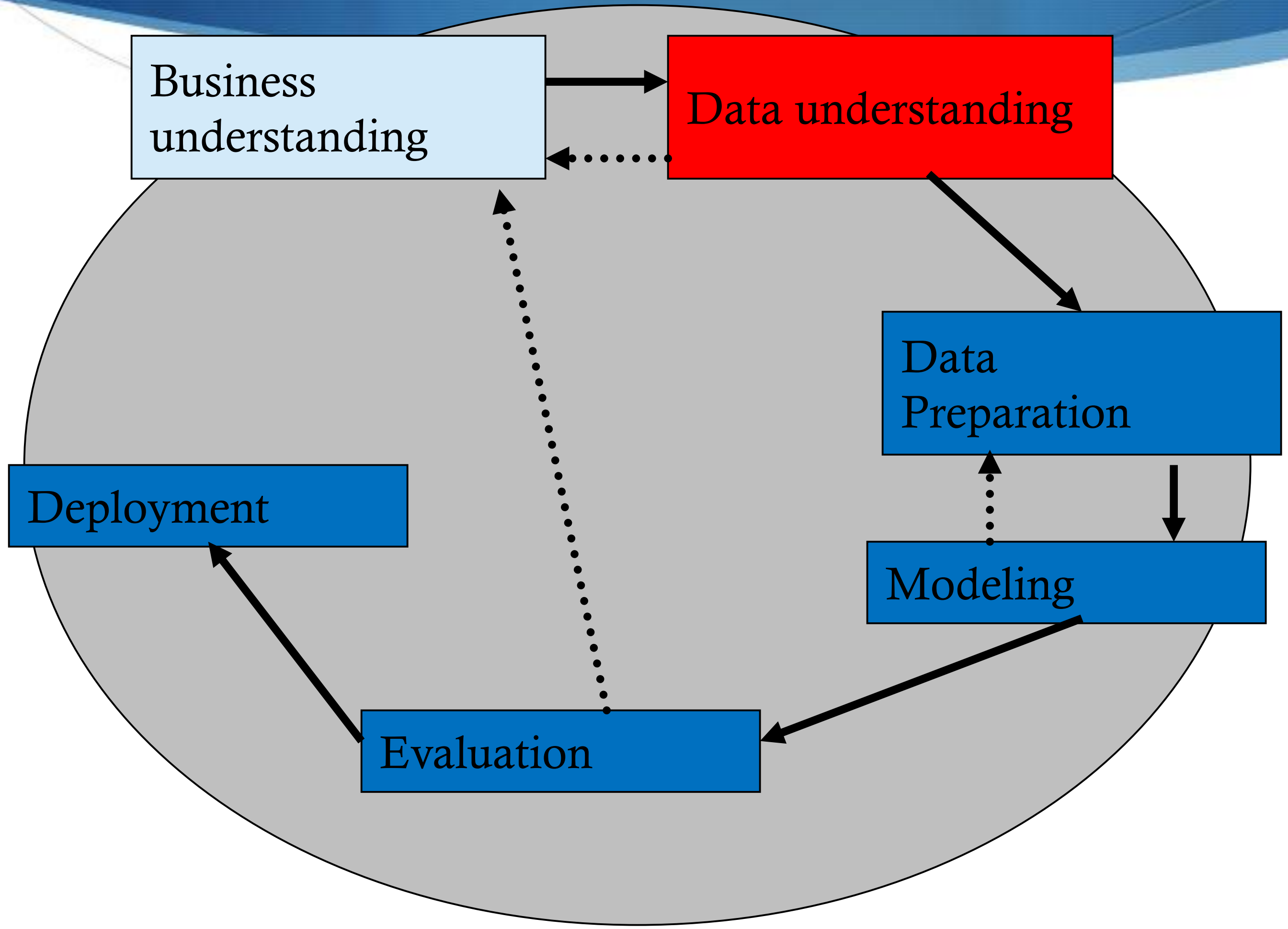
a data mining goal: "Predict how many widgets a customer will buy, given their purchases over the past three years, demographic information (age, salary, city) and the price of the item."

- **Produce project plan**

- **describe the intended plan** for achieving the data mining goals and the business goals

- the plan should **specify the anticipated set of steps to be performed** during the rest of the project including an initial selection of tools and techniques

Phase 2:- Data Understanding



<u>TASKS</u>

Collect data

Describe data

Explore data

Verify data quality

Phase 2:- Data Understanding

- Explore the data
- Verify the quality
- Find Outliers

Starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Phase 2. Data Understanding

- Collect initial data

- acquire within the project the data listed in the project resources

- includes data loading if necessary for data understanding

- possibly leads to initial data preparation steps

- if acquiring multiple data sources, integration is an additional issue, either here or in the later data preparation phase

- Describe data

- examine the "gross" or "surface" properties of the acquired data

- report on the results

Phase 2. Data Understanding

- Explore data

- tackles the data mining questions, which can be addressed using querying, visualization and reporting including:

distribution of key attributes, results of simple aggregations

relations between pairs or small numbers of attributes

properties of significant sub-populations, simple statistical analyses

- may address directly the data mining goals

- may contribute to or refine the data description and quality reports

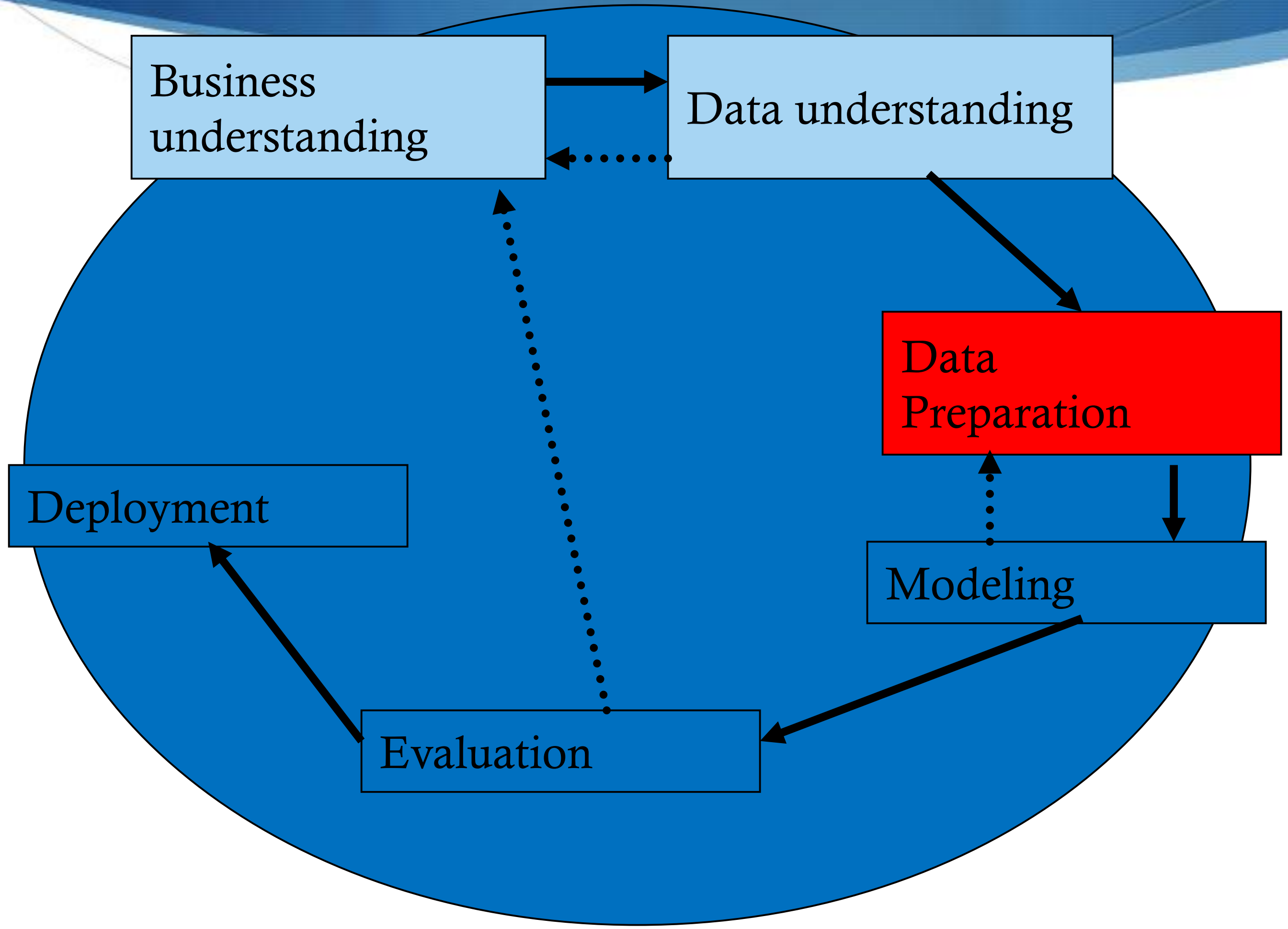
- may feed into the transformation and other data preparation needed

- Verify data quality

- examine the quality of the data, addressing questions such as:

"Is the data complete?", Are there missing values in the data?"

Phase 3:- Data Preparation



TASKS

Select data

Clean data

Construct data

Integrate data

Format data

Phase 3. Data Preparation

- Takes usually **over 90% of the time**

  - Collection
  - Assessment
  - Consolidation and Cleaning
  - Data selection
  - Transformations

Covers **all activities to construct the final dataset** from the initial raw data. Data preparation tasks are **likely to be performed multiple times** and **not in any prescribed order**. Tasks include **table, record and attribute selection** as well as **transformation and cleaning of data for modeling tools**.

Data Preparation

- **Select data**

- **decide on the data to be used** for analysis

- criteria include relevance to the **data mining goals, quality and technical constraints** such as limits on data volume or data types

- covers selection of attributes as well as selection of records in a table

Data Preparation

- **Clean data**

- **raise the data quality to the level required** by the selected analysis techniques

- may involve **selection of clean subsets of the data**, **the insertion of suitable defaults** or **more ambitious techniques such as the estimation of missing data** by modeling

Data Preparation

- **Construct data**

  - constructive data preparation operations such as the production of derived attributes, entire new records or transformed values for existing attributes

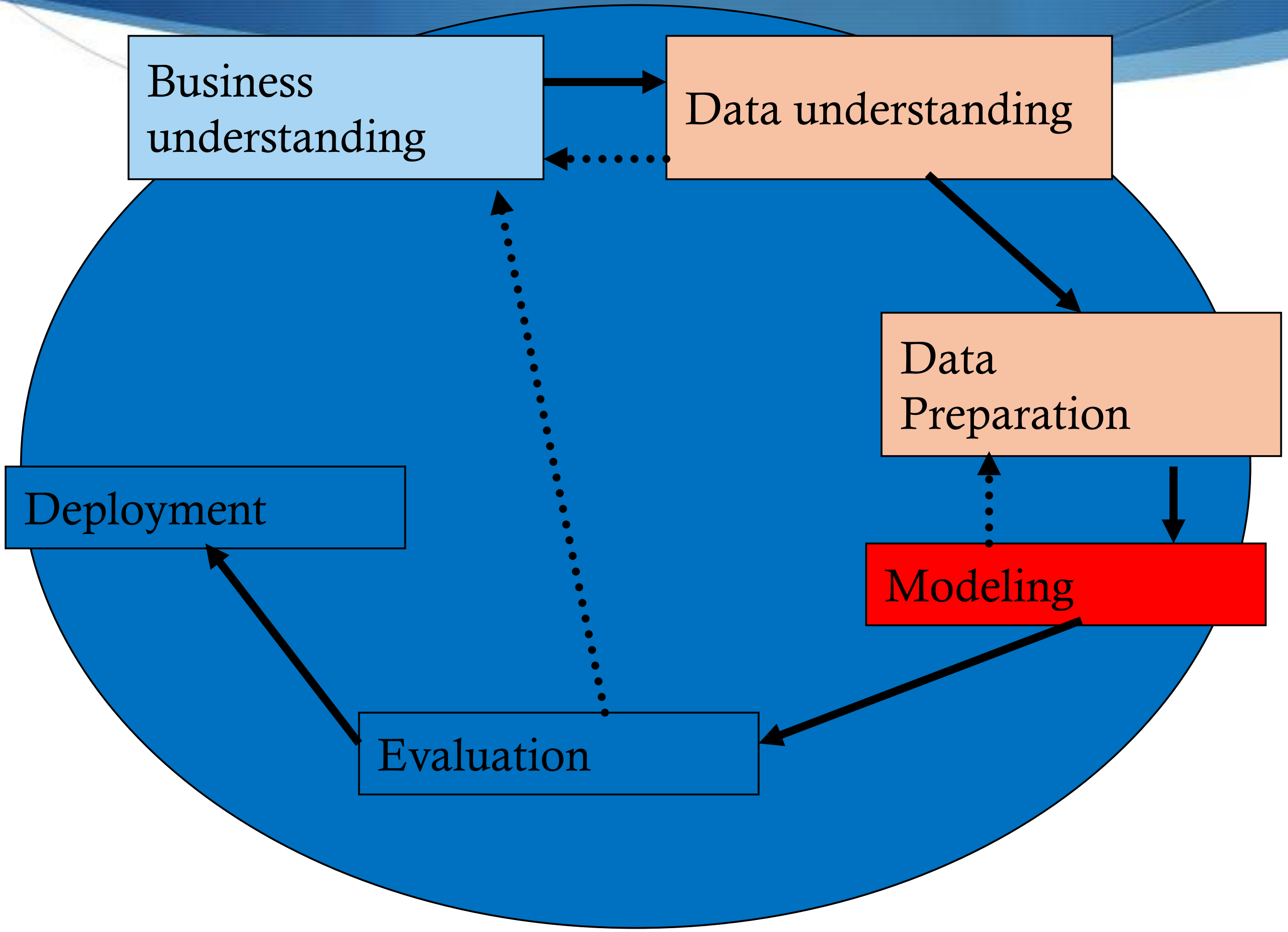
- **Integrate data**

  - methods whereby information is combined from multiple tables or records to create new records or values

- **Format data**

  - formatting transformations refer to primarily syntactic modifications made to the data that do not change its meaning, but might be required by the modeling tool

Phase 4:- Modelling



TASKS

Select modeling techniques

Design the test

Build model

Assess model

Phase 4. Modeling

- **Select the modeling technique**

  (based upon the data mining objective)

- **Build model**

  (Parameter settings)

- **Assess model** (rank the models)

  Various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Some techniques have specific requirements on the form of data. Therefore, **stepping back to the data preparation phase is often necessary.**

Phase 4. Modeling

- **Select modeling technique**

  - **select the actual modeling technique that is to be used**

  ex) decision tree, neural network

  - if multiple techniques are applied, perform this task for each techniques separately

- **Generate test design**

  - **before actually building a model, generate a procedure or mechanism to test the model's quality and validity**

  ex) In classification, it is common to use error rates as quality measures for data mining models. Therefore, typically separate the dataset into train and test set, **build the model on the train set and estimate its quality on the separate test set**

Phase 4. Modeling

- **Build model**

- **run the modeling tool on the prepared dataset to create one or more models**

- **Assess model**

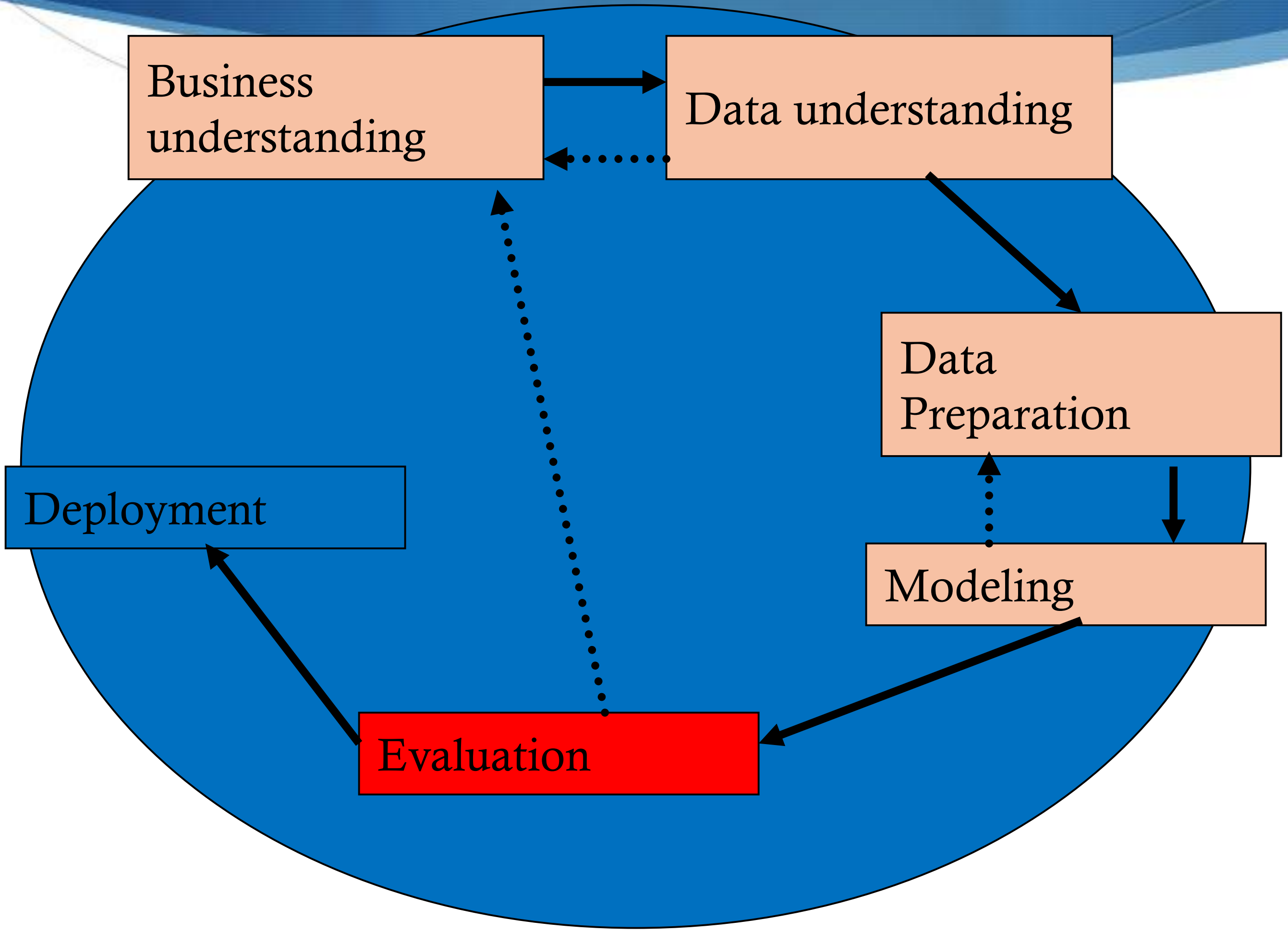
- **interprets the models** according to his domain knowledge, the data mining success criteria and the desired test design

- **judges the success of the application of modeling and discovery techniques** more technically

- contacts business analysts and domain experts later in order to **discuss the data mining results in the business context**

- **only consider models** whereas the evaluation phase also takes into account all other results that were produced in the course of the project

Phase 5:- Evaluation



TASKS

Evaluate results

Review process

Determine next steps

Phase 5. Evaluation

- **Evaluation of model**

  - how well it performed on test data

- **Methods and criteria**

  - depend on model type

- **Interpretation of model**

  - important or not, easy or hard depends on algorithm

Thoroughly **evaluate the model** and **review the steps executed to construct the model** to be certain it **properly achieves the business objectives**. A key objective is **to determine if there is some important business issue that has not been sufficiently considered**. At the end of this phase, **a decision on the use of the data mining results should be reached**

Phase 5. Evaluation

- **Evaluate results**

- assesses the degree to which the model meets the business objectives

- seeks to determine if there is some business reason why this model is deficient

- test the model(s) on test applications in the real application if time and budget constraints permit

- also assesses other data mining results generated

- unveil additional challenges, information or hints for future directions

Phase 5. Evaluation

- Review process

- do a more thorough review of the data mining engagement in order to determine if there is any important factor or task that has somehow been overlooked

- review the quality assurance issues

ex) “Did we correctly build the model?”

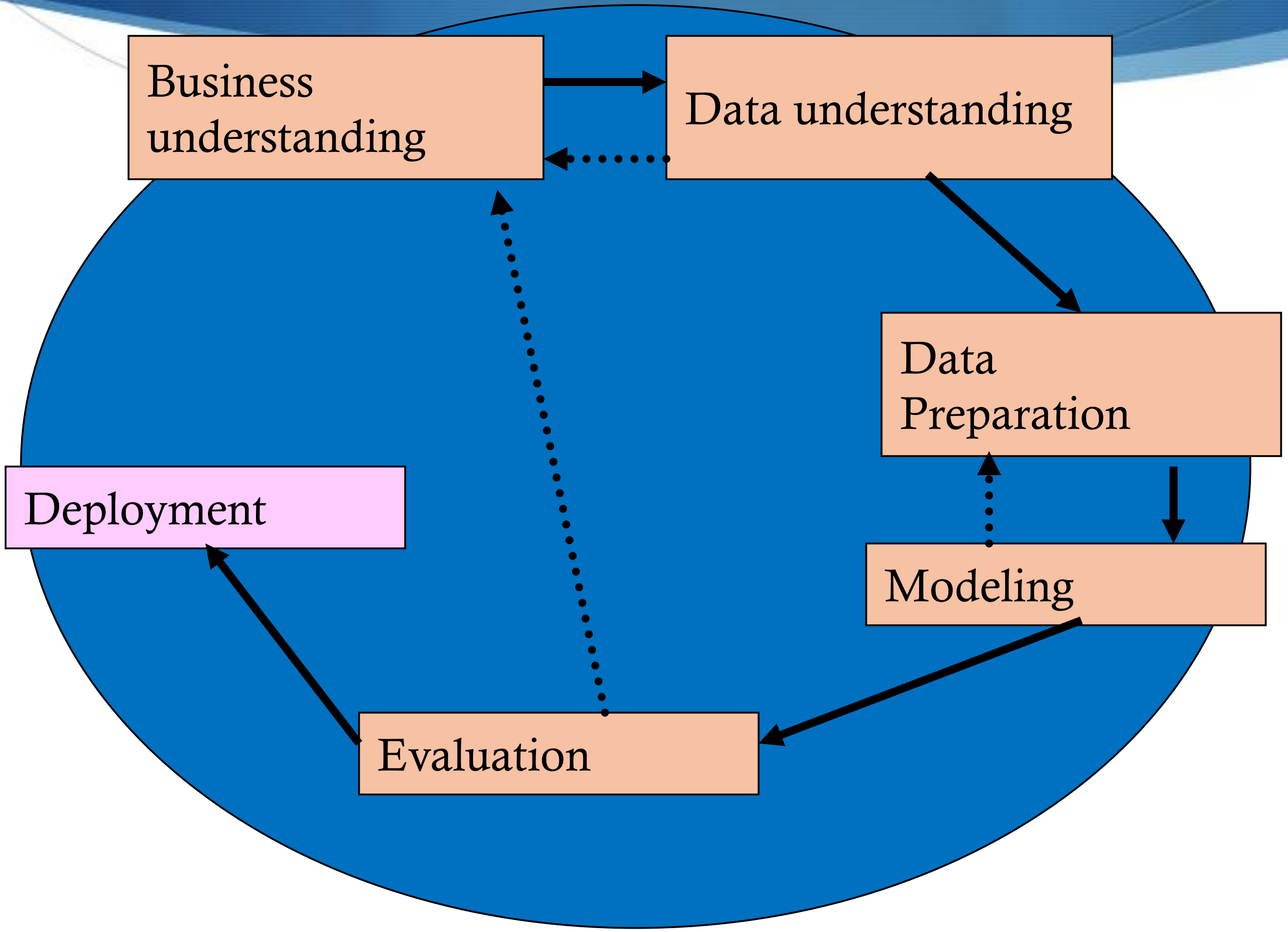
- Determine next steps

- decides how to proceed at this stage

- decides whether to finish the project and move on to deployment if appropriate or whether to initiate further iterations or set up new data mining projects

- include analyses of remaining resources and budget that influences the decisions

Phase 6:- Deployment



TASKS

Plan deployment

Plan monitoring and maintenance

Final report

Review project

Phase 6. Deployment

- Determine **how** the results need to be utilized
- **Who** needs to use them?
- **How often** do they need to be used

- Deploy Data Mining results by

Scoring a database, utilizing results as business rules,

interactive scoring on-line

The knowledge gained will need to **be organized and presented in a way that the customer can use it**. However, **depending on the requirements**, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

Phase 6. Deployment

- **Plan deployment**

- in order to deploy the data mining result(s) into the business, **takes the evaluation results and concludes a strategy for deployment**

- **document the procedure** for later deployment

- **Plan monitoring and maintenance**

- important if the data mining results become part of the day-to-day business and it environment

- **helps to avoid unnecessarily long periods of incorrect usage of data mining results**

- needs a detailed on monitoring process

- takes into account the specific type of deployment

Phase 6. Deployment

- **Produce final report**

- the project leader and his team **write up a final report**

- may be only a summary of the project and its experiences

- may be a final and comprehensive presentation of the data mining result(s)

- **Review project**

- **assess what went right and what went wrong, what was done well and what needs to be improved**