COMP H3024

Data Mining, Level 7

Week 1
Introduction to Data Mining

Lecturer: Geraldine Gray

Contact Details

- Name: Geraldine Gray
- Office: E016
- Email: geraldine.gray@itb.ie

- Moodle - COMP H3024 Data Mining;
  - Enrolment key is ***mining***

Module Outline:

2 hour lecture & tutorial, 2 hour lab

- CA is work 50%
  - 15% lab work, assessed with **moodle quizes** at the end of each lab (5 quizes).
  - 35% assignment (due Dec 12th)
    - You will be given a dataset, and asked to analyse it following the CRISP-DM methodology covered in class.
    - More details on moodle in the coming weeks.
- Exam is worth 50%

Expectations & Communication

Expectation of students

- Attendance
- Assignments
  - Submitting lab worksheets & assignments on time (deduction made for late work)
  - Plagiarism & copying will results in 0% for that piece of work. Repeated plagiarism will result in 0% for the module.

Communication

- Emails via college email account
- Moodle – lecture notes + exercises, submission of work, grades

- Include ID and Name on all work submitted

Module Aims

Syllabus

- To cover in depth the steps involved in the discovery of information in data through the process of Data Mining
  - provide knowledge and understanding
- To give an understanding of how to prepare data for analysis
- To give an understanding of a variety of data mining algorithms

Module Objectives

Learning outcomes: upon completion of this module you will:

Knowledge

- have knowledge and understanding of the following processes and procedures in Data Mining:
  - exploratory data analysis to identify data quality problems and detect interesting subsets of data to form hypotheses for hidden information with use of visualisation
  - the data preparation phase and techniques used for data mining and the knowledge discovery process
  - the data modelling process, techniques and algorithms for classification and prediction
  - applications of data mining and its role in the market place including case studies where data mining was used successfully
  - data mining products and methodologies

Know-how & Skill

- have acquired the technical and knowledge skill range to implement and apply data mining for knowledge discovery on a dataset using CRISP-DM methodology
- be able to select and apply these skills for a data mining task using a data mining tool
- be competent at using and deriving algorithms to build Decision Tree and Back-Propagation Neural Network models for classification and prediction in a data mining task
- be able to apply and interpret data mining and association rules for market basket analysis
- be able to asses the performance and patterns produced by a data mining process

Competence

- be competent in the recognition, derivation and application of knowledge, processes, techniques and tools in a data mining task. The student will be expected to determine and achieve their personal outcomes and interact effectively as part of the learning group.

Recommended Reading

- "*Data Mining: Concepts and Techniques*", J. Han and M. Kamber, Morgan Kaufmann, 2001, ISBN: 1558604898
- "*Introduction to Data Mining*" Pang-Ning Tan, Michael Steinback, Vipin Kumar, Boston: Pearson Addison Wesley, 2005, ISBN: 0321321367
- "*Discovering Knowledge in Data: An Introduction to Data Mining*", Daniel T. Larose, Wiley, 2005, ISBN: 0471666572
- "*Data Mining: Practical Machine Learning Tools and Techniques*", Witten, I. H., Morgan Kaufmann, 2005, ISBN: 0120884090
- "*Data Mining: Building a Competitive Advantage*", Robert Groth, Prentice Hall PTR, c2000, ISBN: 0130862711

Lecture 1 outline

- Introduction to Data mining
  - Data mining applications
- Data mining tool
  - RapidMiner
- Methodology
  - CRISP- DM

Download V 5.3 only (free version).

What is Data Mining (DM)

- DM
  - concerned with extracting *useful information* (**knowledge**) from data sets
  - is the process of finding *trends* and *patterns* in data
  - to sort through large quantities of data and discover new (**interesting**) information
  - turning any newfound knowledge into **actionable** results

- Provides computational techniques to help analyse volumes/stores of data

What is Data Mining …

- Database programs can query (SQL) for specific information
  - e.g. how many patients are over age 70
  - E.g. query a web search engine for information about ‘Amazon’

- But there is potentially much more in the data than just specific information i.e. interesting new patterns
  - e.g. finding correlations in groups of people who are affected by a similar disease
  - e.g. group together similar documents returned by search engines according to their context i.e. ‘Amazon rainforest’ or ‘Amazon.com’

- KDD/DM searches through data for hidden relationships and patterns in the data

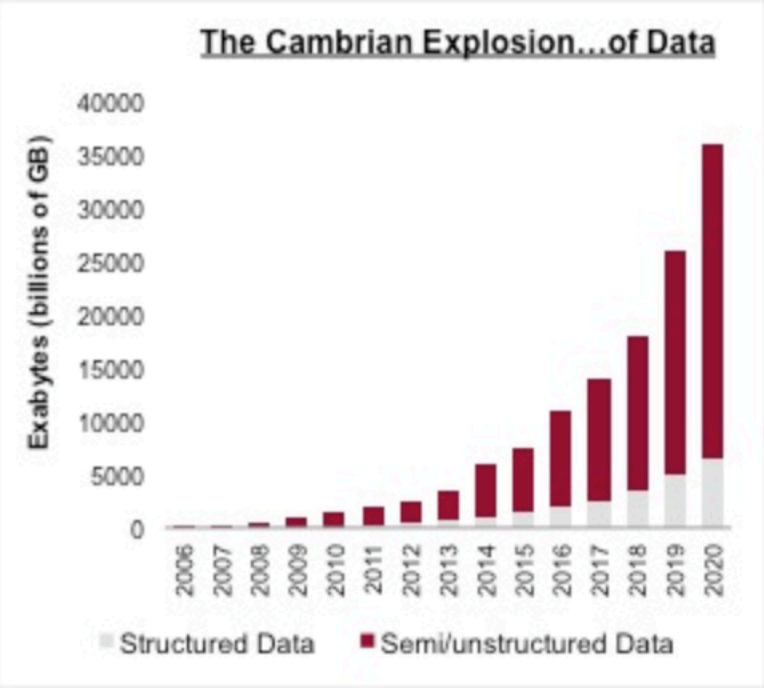
Why do we need Data Mining?

- Data being collected and accumulated in many forms at a dramatic pace (these include database systems, text files, web pages, social networking sites)
  - The volumn of Information in the world is estimated to double every **9 months** (kdnuggets.com)

- Need computational theories to assist in the extraction of knowledge from this rapidly growing volumn of data

Why do we need DM…

Data Explosion



- Automated data collection, ever maturing database technology – all leads to **huge stores of data in** databases, data warehouses and other data repositories
  - Data storage became easier (large amounts of computing power at low costs for processing and storage) but need some way of gaining knowledge from all this data
- DM process, technologies and practices provide a solution.

Examples of Data Mining Application Areas

- In Science
  - Astronomy…
  - Bioinformatics
- In Business
  - Advertising
  - Telecom
  - Targeted marketing
  - Customer Relationship Management (CRM) & Customer modelling…
  - E-commerce
  - Fraud detection…
  - Health care
  - Investments
  - Manufacturing
  - Sports/entertainment…
- Web
  - Search engines…
- Government
  - Anti-terrorism efforts
  - Law enforcement
  - Profiling tax cheats

Example Data Mining Application Areas…

- **Customer modelling** - important and widespread business application (predictive analytics) e.g. predicting attrition or churn
- **Astronomy** – performing image analyses, classification and cataloguing of sky objects from sky survey images (Fayyad, Djorgovski & Weir 1996). JPL and the Palomar Observatory
- **Credit risk** – identify the risk that a customer will not pay back a loan or credit card
- **Marketing** – database marketing systems which analyse customer databases to identify different customer groups and forecast their behaviour e.g. customer acquisition, cross-sell
- **Fraud detection and criminal activity** – monitoring credit card fraud, identifying financial transactions indicating money laundering
- **Web mining** – navigating through an information rich environment to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.
- **Sports** - IBM *Advanced Scout* analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat

Example Data Mining Application Areas…

E-commerce – **case study**

- A person buys a book (product) at Amazon.com

What is the task?

- Task: Recommend other books (products) this person is likely to buy
- Amazon does clustering based on books bought:
  - customers who bought "**Advances in Knowledge Discovery and Data Mining**", also bought "**Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**"
- Recommendation program is quite successful

Example Data Mining Application Areas…

Assessing credit risk – case study

- Situation: Person applies for a loan
- Task: Should a bank approve the loan?
- Note: People who have the best credit don't need the loans, and people with worst credit are not likely to repay. Bank's best customers are in the middle

- Banks develop credit models using a variety of machine learning methods.
- Mortgage and credit card proliferation are the results of being able to successfully predict if a person is likely to default on a loan
- Widely deployed in many countries

Jobs in data mining

- Search any recruitment site with the key term ‘data analytics’ to see job opportunities

64 jobs on computerjobs.ie, 11/09/2014

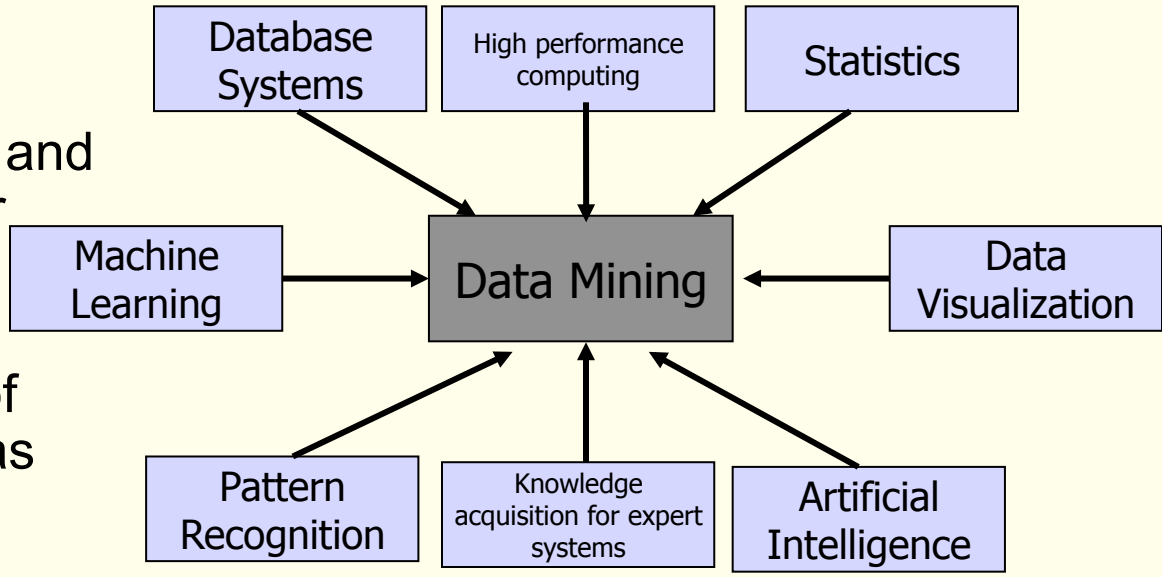
Some definitions

Data Mining / Knowledge discovery– is the **nontrivial process** of identifying **valid**, **novel**, **potentially useful** and ultimately **understandable patterns** in data (Fayyad et al)

- Process – knowledge discovery has many steps
- Nontrivial – some search or inference is involved (not a straightforward computation)
- Valid – discovered patterns should be valid on new data with some degree of certainty
- Novel – patterns to be novel to the user or system
- Potentially useful – lead to some benefit to the user or task
- Understandable – if not immediately then after some post-processing
- Pattern – high-level description from the data, summarising a trend that applies to a group of rows in the datset: E.g. 20% of people who buy nappies on a Friday evening also buy beer.

Interdisciplinary nature of DM

- DM important field – tools and applications developed for business, science, government & academia
- Involves the intersection of many different expert areas



These fields provide some of the methods and algorithms used for the data mining process which includes:

–How data is stored & accessed

–Methods & algorithms used to discover patterns & trends

–Scaling algorithms to capture massive data sets and still run efficiently

–How results are interpreted and visualised

Therefore DM very multi disciplinary encompassing techniques beyond the scope of any one discipline

CRISP-DM

CRoss-Industry Standard Process for Data Mining

Developed by

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

Why Should There be a Standard Process?

- **Framework for recording experience**
  - Allows projects to be replicated
- **Aid to project planning and management**
- **"Comfort factor" for new adopters**
  - Demonstrates maturity of Data Mining
  - Reduces dependency on "stars"

*CRISP - DM

Data Mining Algorithms

CRoss-Industry Standard Process for Data Mining

* Non-proprietary
* Application and Industry Neutral
* Tool neutral

www.crisp-dm.org

Why is there a standard process?

Framework for recording experience

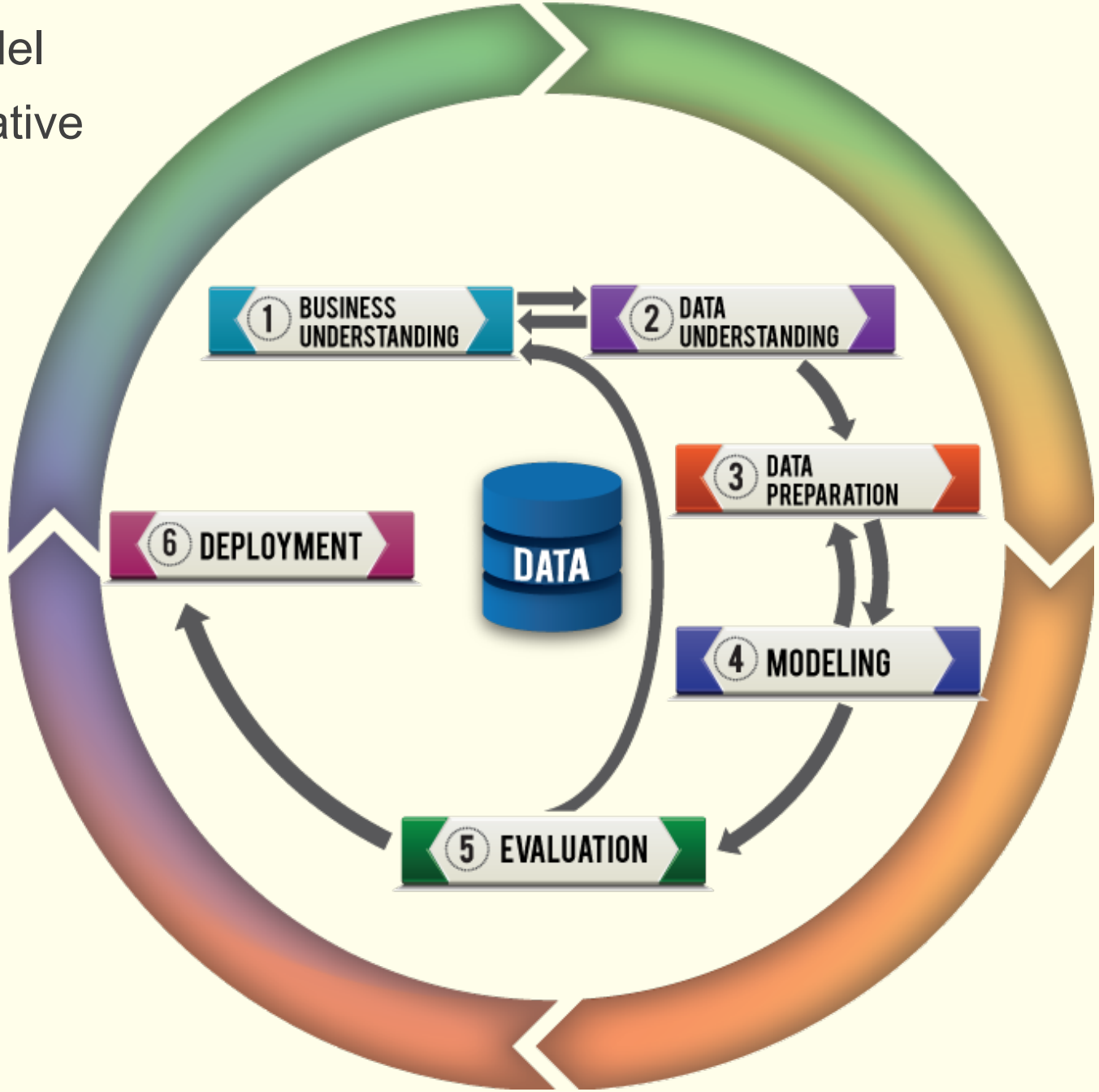
- Allows projects to be replicated

Aid project planning and management

- Demonstrates maturity of Data Mining

Adds a “Comfort factor” for new adopters

- Process Model
- 6-phase Iterative Lifecycle



*CRISP-DM: Phases

1. Business Understanding

   Project objectives and requirements understanding, Data mining problem definition

2. Data Understanding

   Initial data collection and familiarization, Data quality problems identification

3. Data Preparation

   Table, record and attribute selection, Data transformation and cleaning

*CRISP-DM: Phases

4. Modeling

Modeling techniques selection and application, Parameter calibration

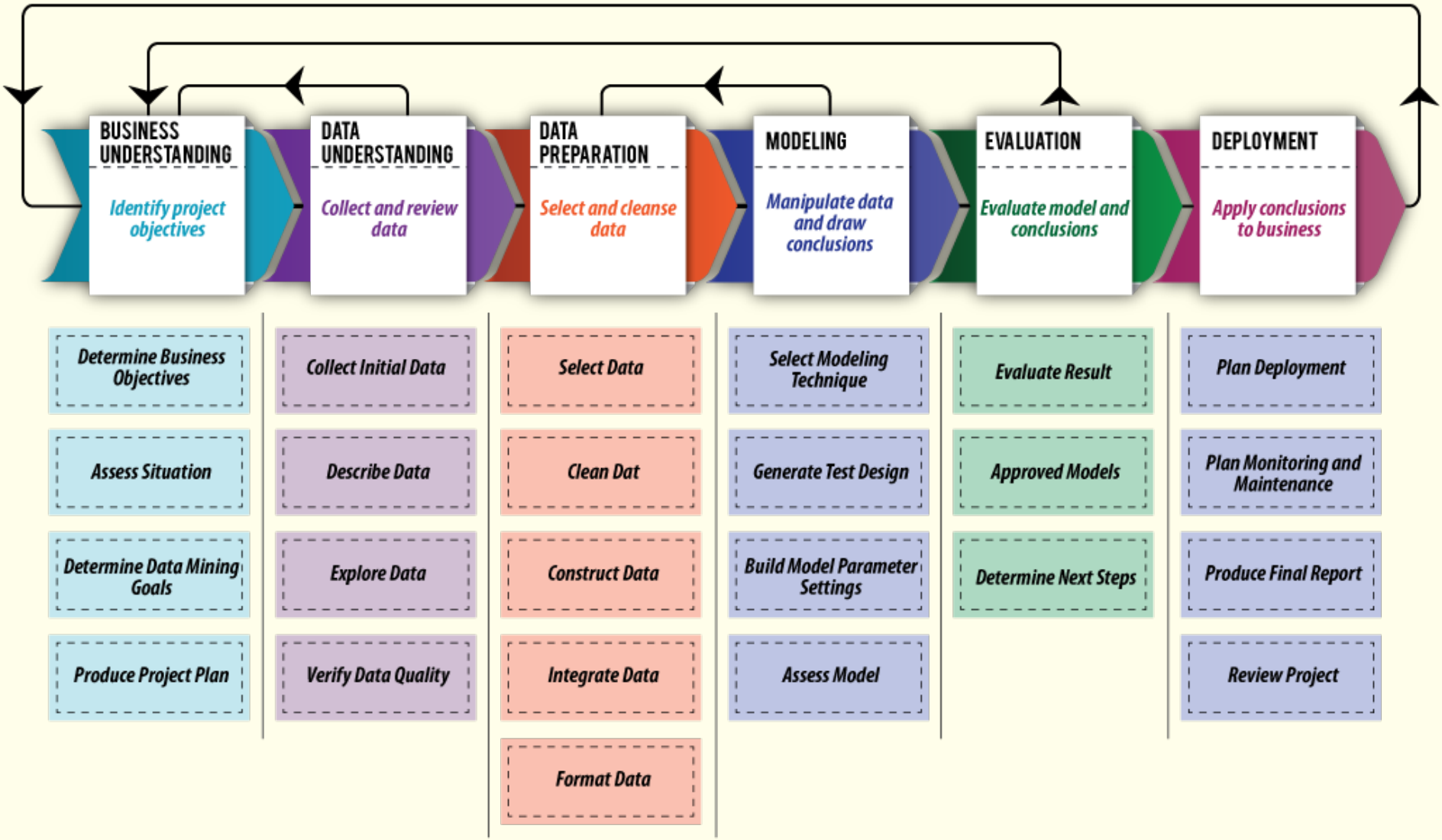
5. Evaluation

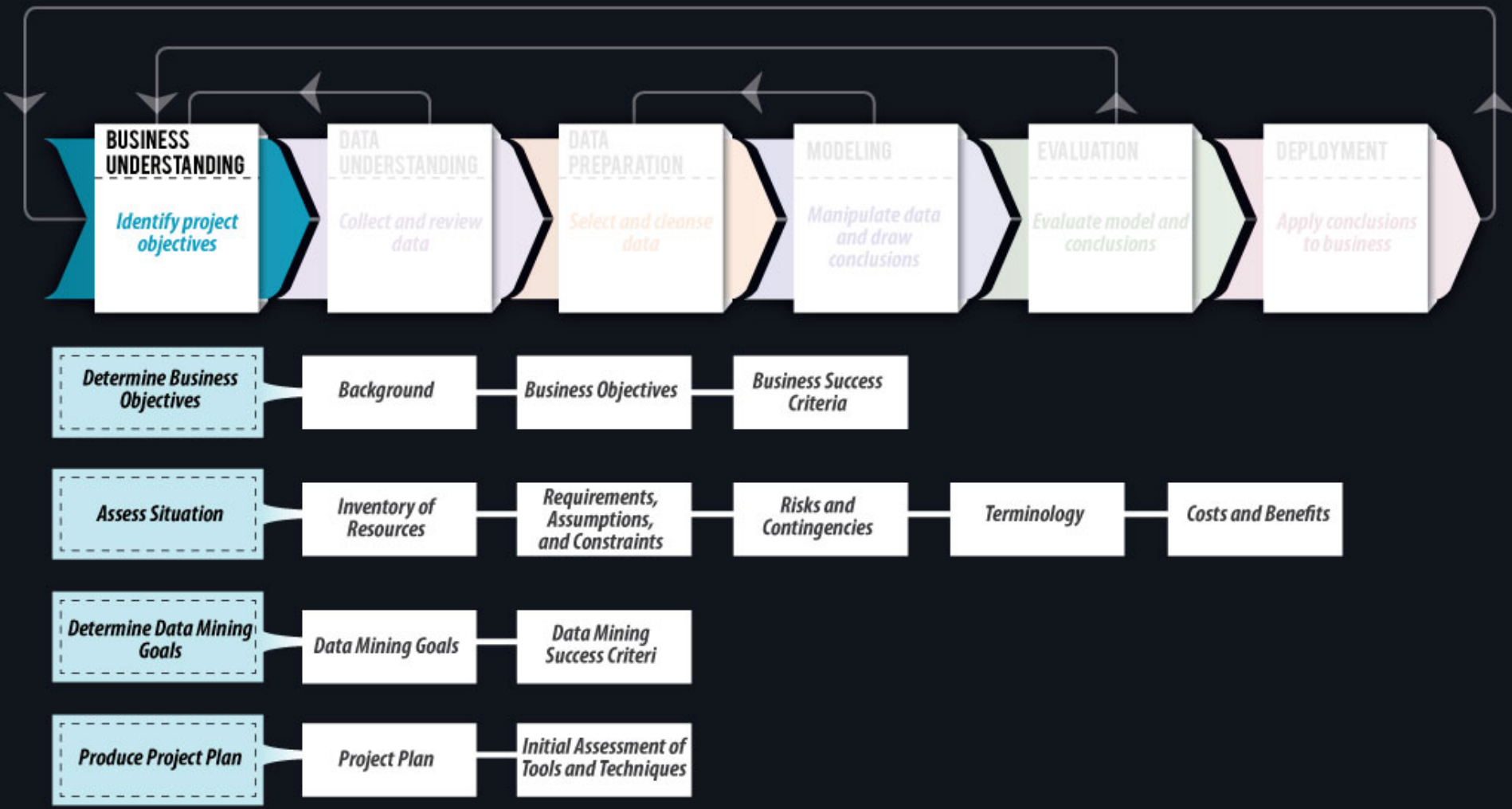
Business objectives & issues achievement evaluation

6. Deployment

Result model deployment, Repeatable data mining process implementation

Phases and Tasks




BUSINESS UNDERSTANDING
Identify project objectives
DATA UNDERSTANDING
Collect and review data
DATA PREPARATION
Select and cleanse data
MODELING
Manipulate data and draw conclusions
EVALUATION
Evaluate model and conclusions
DEPLOYMENT
Apply conclusions to business
Determine Business Objectives
Background
Business Objectives
Business Success Criteria
Assess Situation
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits
Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteri
Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

Phase: 1

Phase 1. Business Understanding

Focuses on understanding :

- The project objectives
- The requirements from the businesses stand point
  - » The Business Objective
- This information is then converted into a data mining problem definition
  - » The Data Mining objective
- A preliminary plan is designed to achieve the objectives

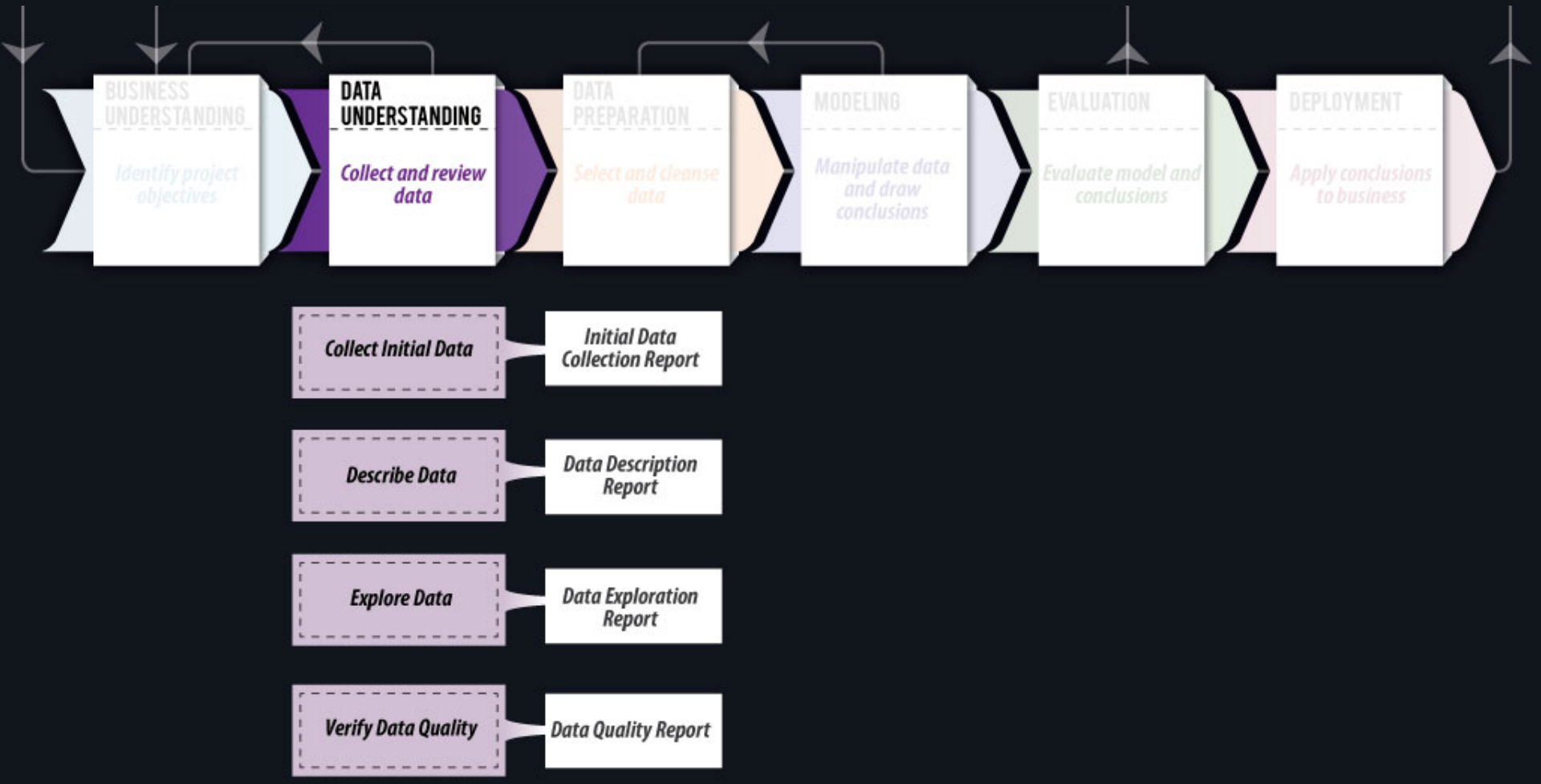
Phase 1. Business Understanding

Example of a Busines objective:

- **Increase catalog sales to existing customers**

Example of the corresponding data mining objective:

- **Build a classification model to predict how many items a customer will buy, given their purchases over the past three years, demographic information (age, salary, city) and the price of the item**

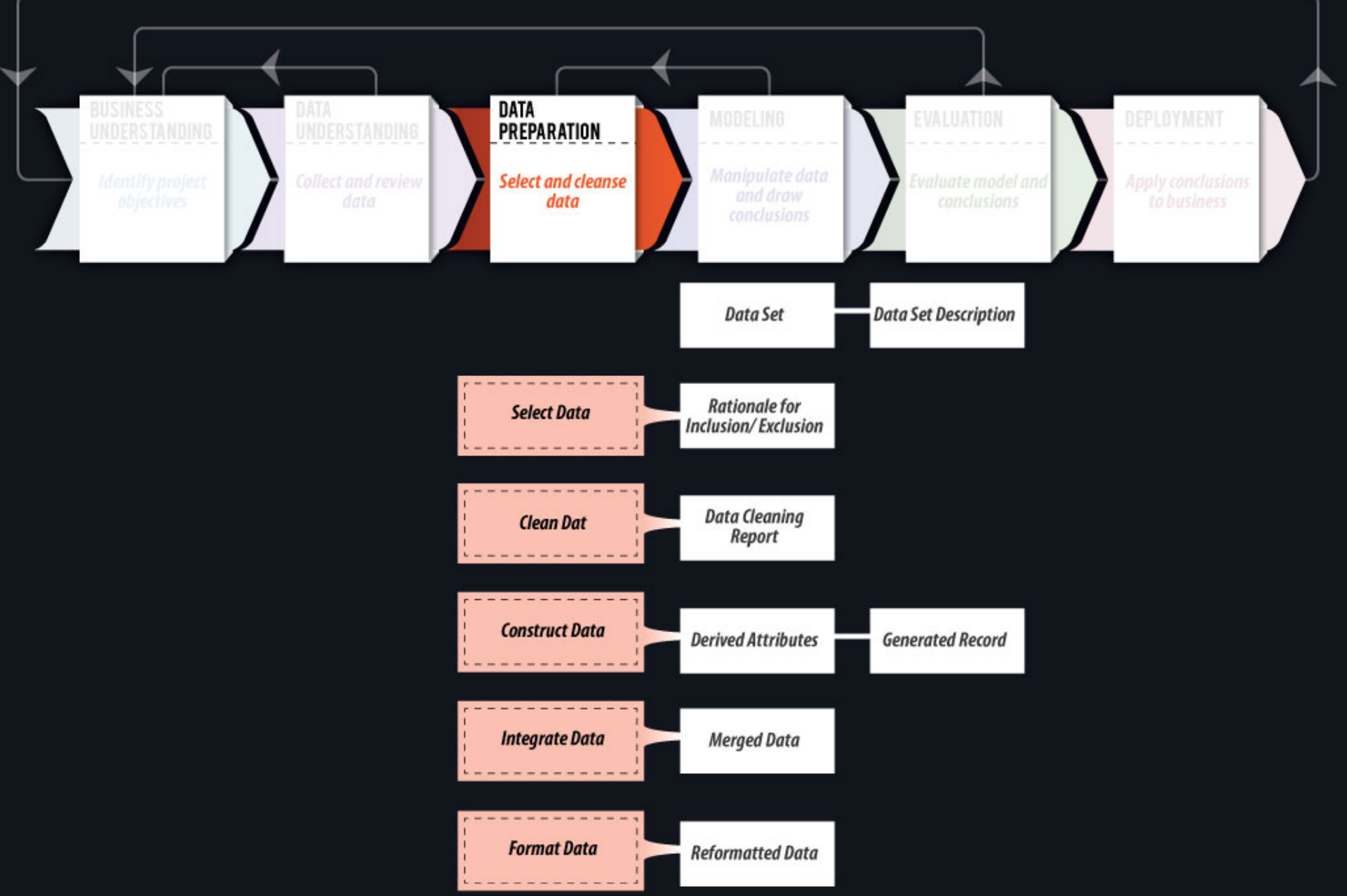

BUSINESS UNDERSTANDING
Identify project objectives
DATA UNDERSTANDING
Collect and review data
DATA PREPARATION
Select and cleanse data
MODELING
Manipulate data and draw conclusions
EVALUATION
Evaluate model and conclusions
DEPLOYMENT
Apply conclusions to business
Collect Initial Data
Initial Data Collection Report
Describe Data
Data Description Report
Explore Data
Data Exploration Report
Verify Data Quality
Data Quality Report

Phase: 2

Phase 2. Data Understanding

This phase involves:

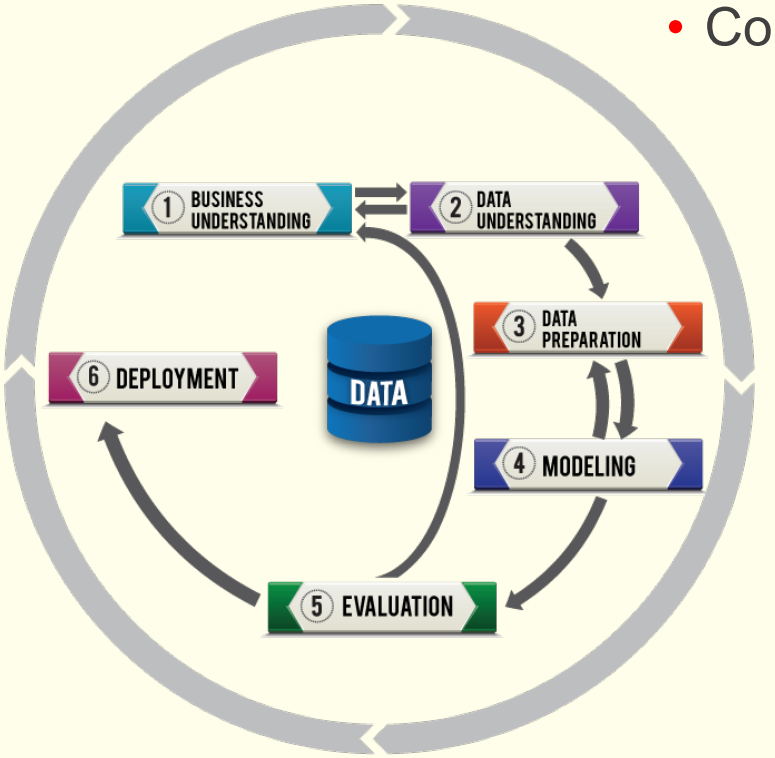
- Collecting initial data
- Describing the data
- Exploring the data
- Verifying the data quality


BUSINESS UNDERSTANDING
Identify project objectives
DATA UNDERSTANDING
Collect and review data
DATA PREPARATION
Select and cleanse data
MODELING
Manipulate data and draw conclusions
EVALUATION
Evaluate model and conclusions
DEPLOYMENT
Apply conclusions to business
Data Set
Data Set Description
Select Data
Rationale for Inclusion/ Exclusion
Clean Dat
Data Cleaning Report
Construct Data
Derived Attributes
Generated Record
Integrate Data
Merged Data
Format Data
Reformatted Data

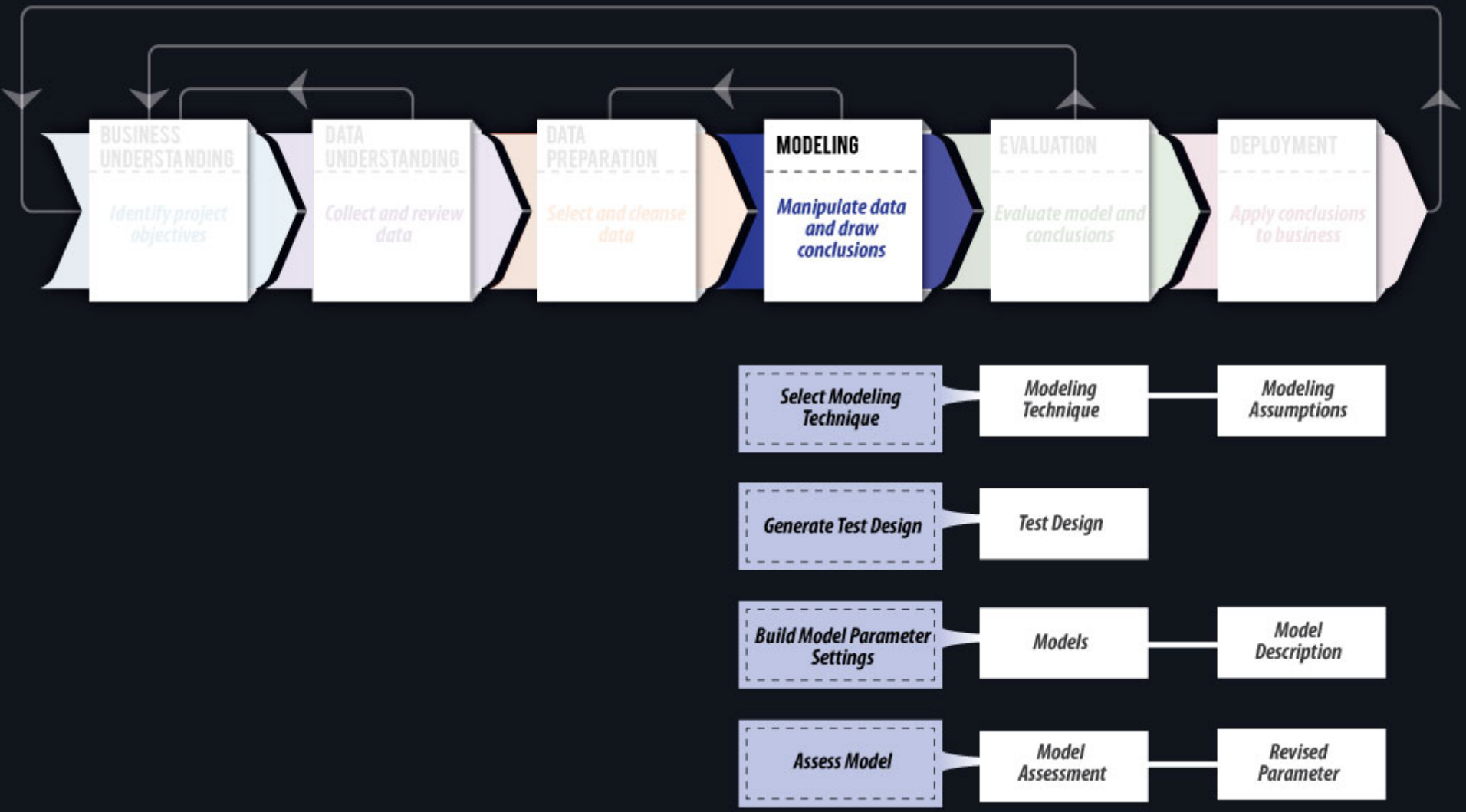
Phase: 3

Phase 3. Data Preparation

This phase usually takes over 80% of the time. It involves:



- Consolidating and Cleaning
- Data Selection
- Transformations
- Aggregations


BUSINESS UNDERSTANDING
Identify project objectives
DATA UNDERSTANDING
Collect and review data
DATA PREPARATION
Select and cleanse data
MODELING
Manipulate data and draw conclusions
EVALUATION
Evaluate model and conclusions
DEPLOYMENT
Apply conclusions to business
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

Phase: 4

Phase 4. Modelling

Select the modelling technique

- Based upon the data mining objective
- Eg. Decision tree, neutral network, regression

Generate test design

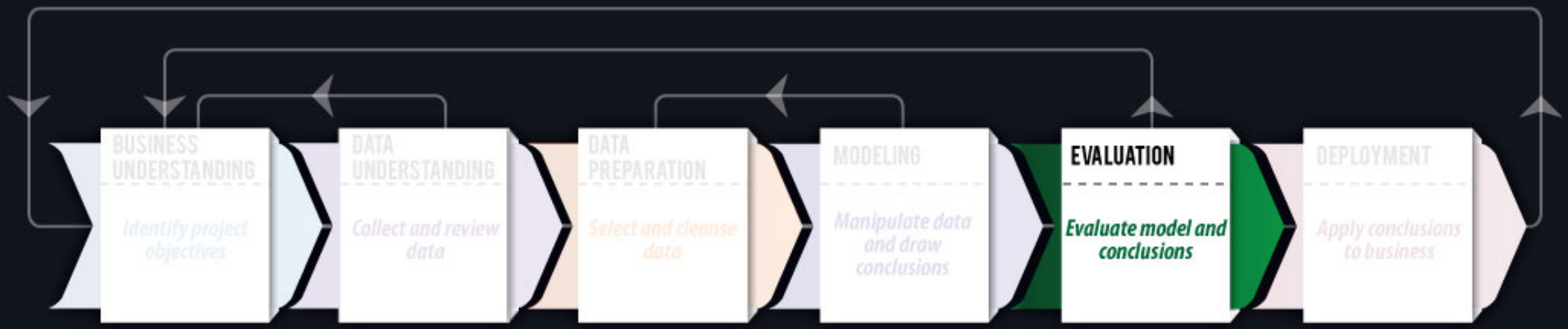
- Eg. In classification, it is common to use error rates as quality measures for data mining models

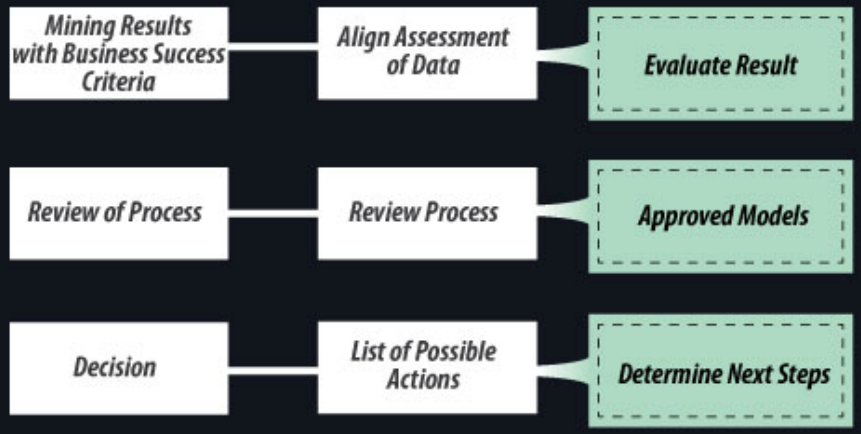
Build model

- Parameter settings

Assess model

- Rank the models


BUSINESS UNDERSTANDING
Identify project objectives
DATA UNDERSTANDING
Collect and review data
DATA PREPARATION
Select and cleanse data
MODELING
Manipulate data and draw conclusions
EVALUATION
Evaluate model and conclusions
DEPLOYMENT
Apply conclusions to business


Mining Results with Business Success Criteria
Align Assessment of Data
Evaluate Result
Review of Process
Review Process
Approved Models
Decision
List of Possible Actions
Determine Next Steps

Phase: 5

Phase 5. Evaluation

Evaluation of model

- How well it performed on test data

Methods and criteria

- Depended on model type

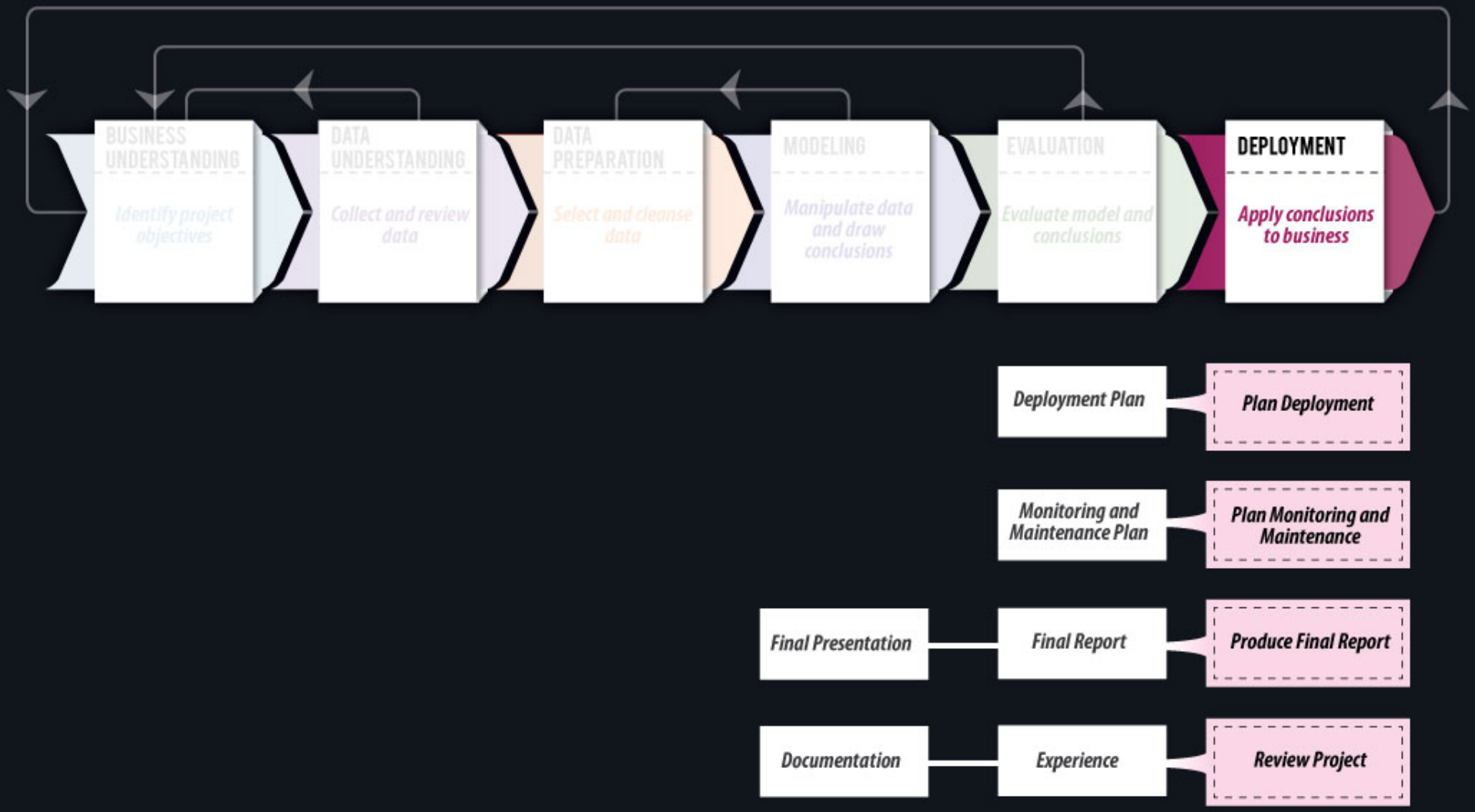
Interpretation of model

- How easy this is to do depends on the algorithm

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
- Also assesses other data mining results generated
- Review additional challenges arising, and/ or information for future directions


BUSINESS UNDERSTANDING
Identify project objectives
DATA UNDERSTANDING
Collect and review data
DATA PREPARATION
Select and cleanse data
MODELING
Manipulate data and draw conclusions
EVALUATION
Evaluate model and conclusions
DEPLOYMENT
Apply conclusions to business
Deployment Plan
Plan Deployment
Monitoring and Maintenance Plan
Plan Monitoring and Maintenance
Final Presentation
Final Report
Produce Final Report
Documentation
Experience
Review Project

Phase: 6

Phase 6. Deployment

Determine how the results need to be utilized

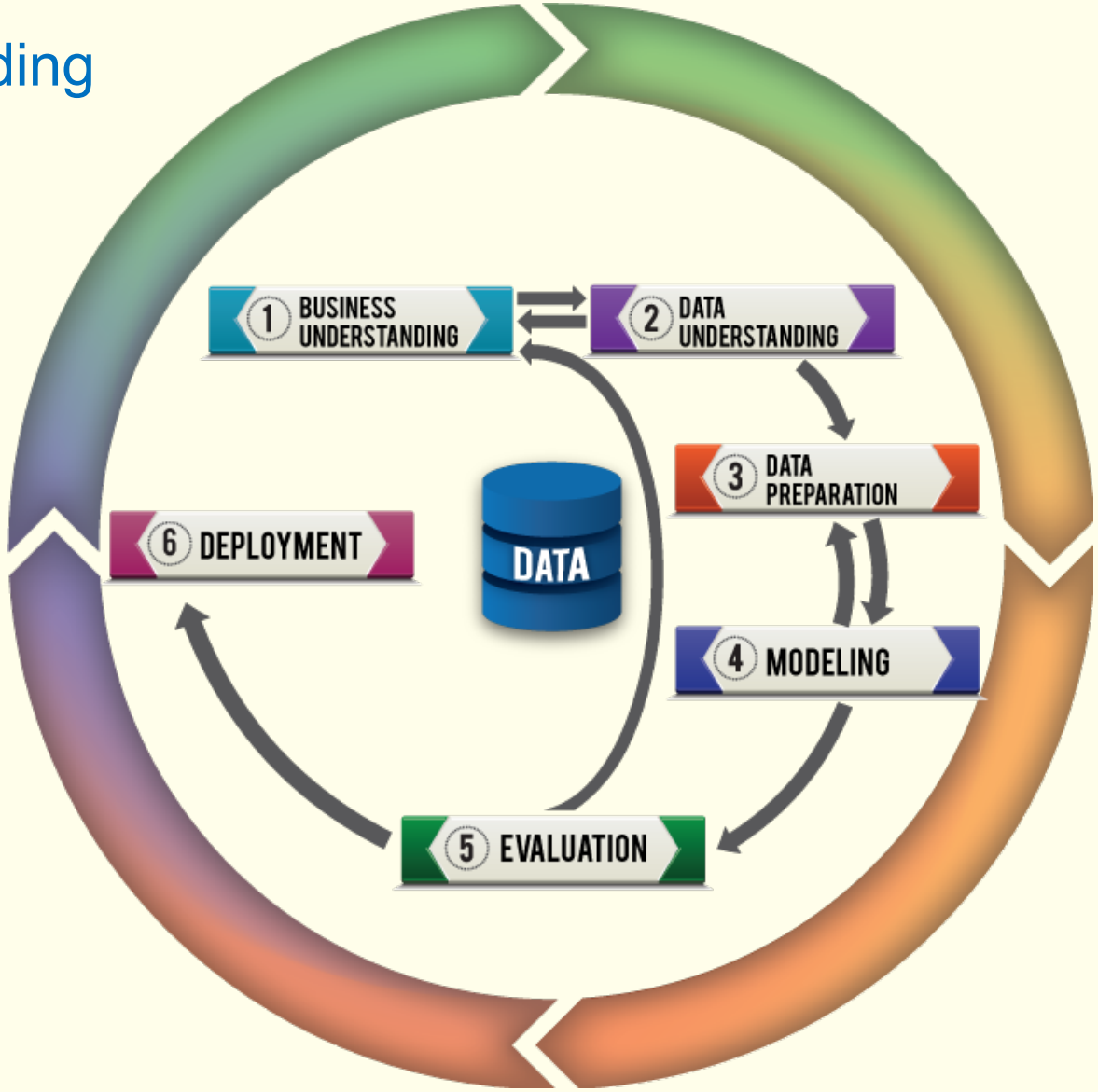
- Who needs to use them?
- How often do they need to be used?

Deploy Data Mining results by:

- Scoring a database (make predictions), utilizing results as business rules.

CRISP-DM Summary

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment



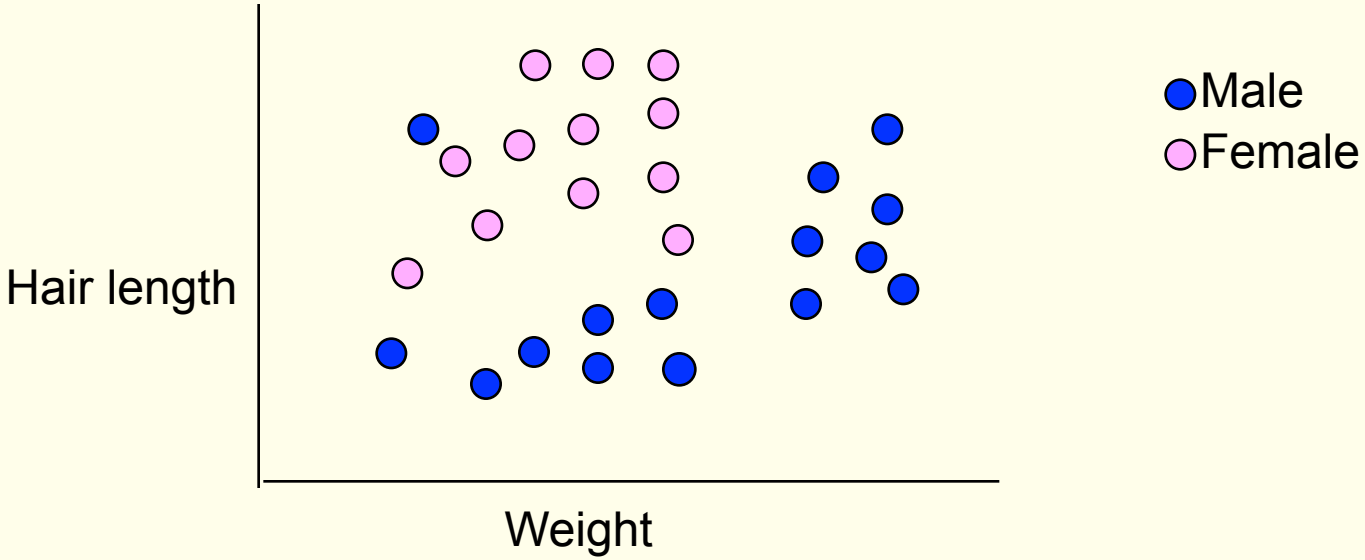
Data Mining Tasks

- **Classification**: predicting an item class
- **Clustering**: finding clusters in data
- **Associations**: e.g. A & B & C occur frequently
- **Visualization**: to facilitate human discovery
- **Summarisation**: describing a group
- **Deviation Detection**: finding changes
- **Link Analysis**: finding relationships

Data Mining Tasks

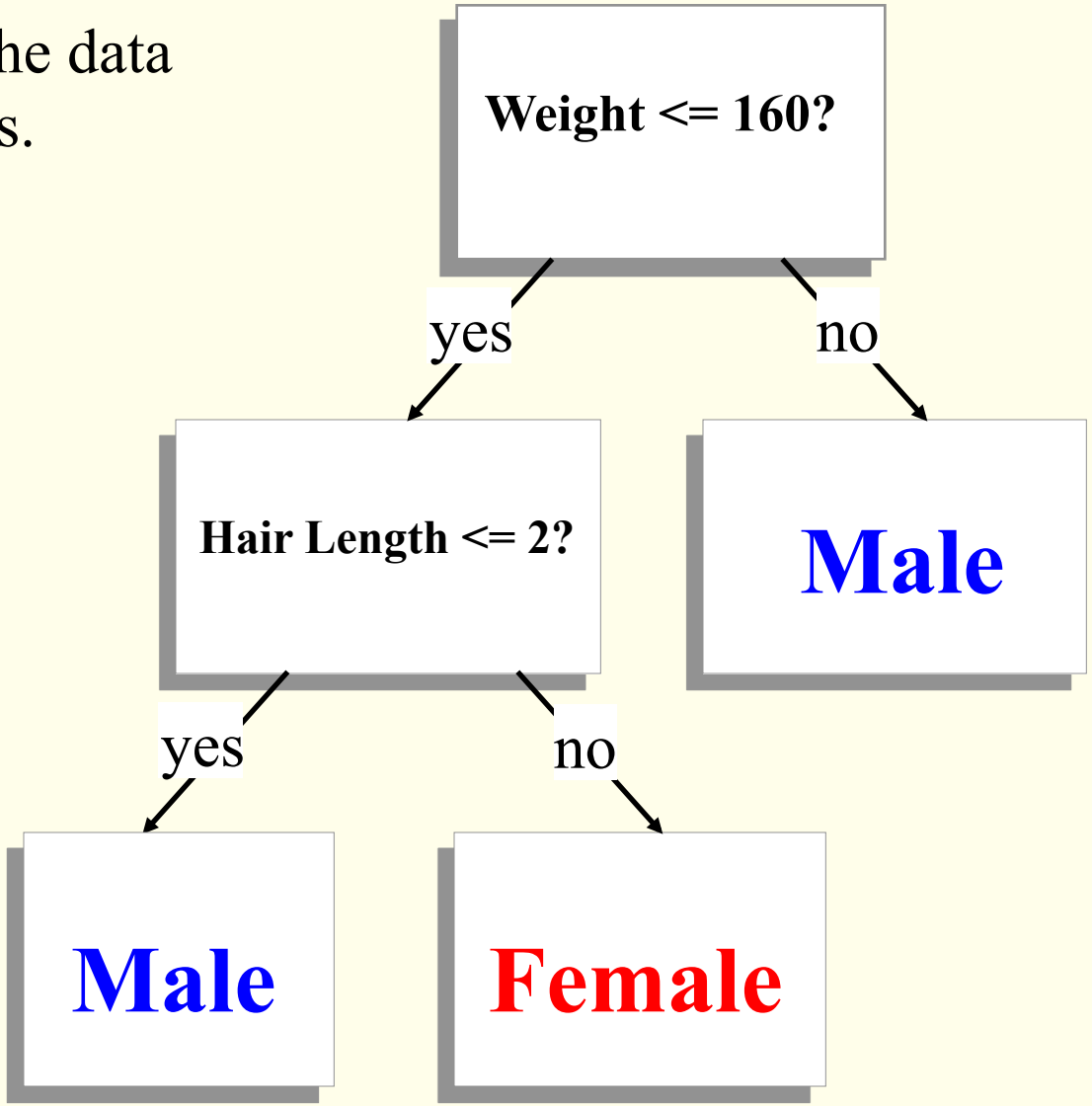
- **Classification and Prediction**
  - Construct models (functions) that describe and distinguish classes or concepts for future prediction
    - e.g., classify countries based on climate, or classify cars based on gas mileage
  - Presentation: decision-tree, classification rule, neural network
  - Predict some unknown or missing numerical values

  Learn a method for predicting the instance class from pre-labeled (classified) instances



We need don't need to keep the data around, just the test conditions.

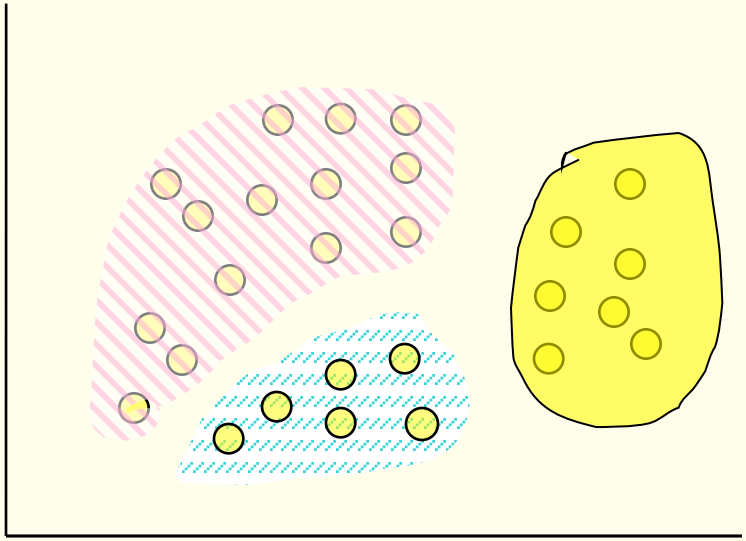
How would these people be classified?



Data Mining Tasks

- **Cluster analysis**
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity

Find "natural" grouping of instances given un-labeled data



Data Mining Tasks

- **Association**
  - Milk and Butter ➔ Tea

- **Outlier analysis**
  - Outlier: a data object that does not comply with the general behavior of the data

Are All the "Discovered" Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
  - Key objective:

FIND PATTERNS THAT ARE MEANINGFUL AND USEFUL

Module topics

Module topics follow CRISP-DM

- Business Understanding (done today)
- Data Understanding
- Data Mining
  - Classification
  - Clustering
  - Association analysis (time permitting)
- Model Evaluation
- Data Preparation Techniques
  - (usually done before mining, but topic will make more sense once you understand how mining algorithms work)

Lecture Summary

- Data Mining & Knowledge discovery: discovering interesting patterns from large amounts of data
- A natural evolution of database technology, in great demand, with wide applications
- The process includes Problem understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment
- Mining can be performed in a variety of information repositories
- Data mining functions: classification, clustering, association, outlier analysis, etc.

Reading & Links

- *"From Data Mining to Knowledge Discovery in Databases"*, U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, AAAI.
- Books in library

- http:// www.kdnuggets.com
- http:// www.crisp-dm.org
- www.spss.com
- www.sas.com
- Data Mining and Knowledge Discovery – journal http://www.springerlink.com/content/1573-756X/