


CIS9660: Data Mining for Business Analytics

1. From Business Problems to Data Mining Tasks

Foster Provost and Tom Fawcett. 2013. *Data Science for Business: What you need to know about data mining and data-analytic thinking. Chapter 2*



What is data
mining?

Definition: Imp All

Data mining is a process that uses statistical, mathematical, and artificial intelligence techniques to extract and identify useful patterns from large sets of data. These patterns can be in the form of business rules, affinities, correlations, trends, or prediction models. (Nemati and Barko 2001)

Why Data Mining

- Society produces huge amounts of data
 - Sources: business, science, medicine, economics, geography, environment, sports, ...
- This data is a potentially valuable resource
- Raw data is useless: need techniques to automatically extract information from it
 - Data: recorded facts
 - Information: patterns underlying the data



Supervised vs. Unsupervised Data Mining

Supervised:

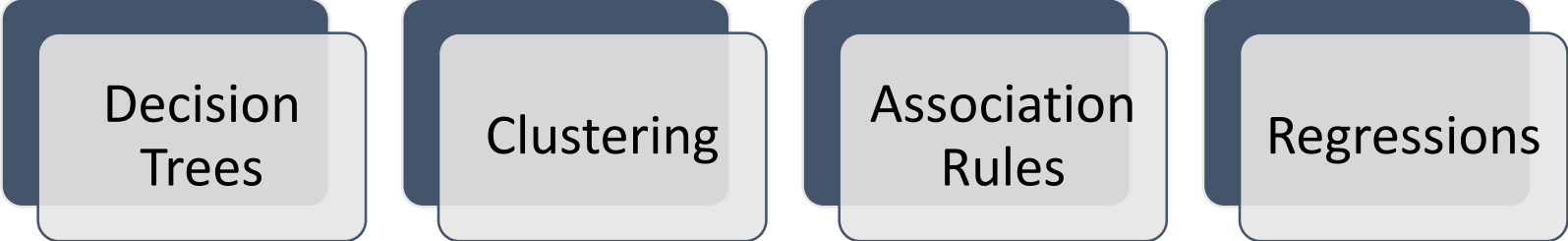
- This method operates under supervision by being provided with the actual outcome for each of the training examples.
- Classification and regression
- e.g. find groups of customers who are more likely to cancel their orders

Unsupervised:

- No actual outcome is provided, and algorithms are left to their own devices to discover and present interesting patterns in the data
- Clustering and association rule mining
- e.g. categorize customers into different groups based on similarity



Four Data Mining Techniques We Will Be Doing in this Class



Decision
Trees

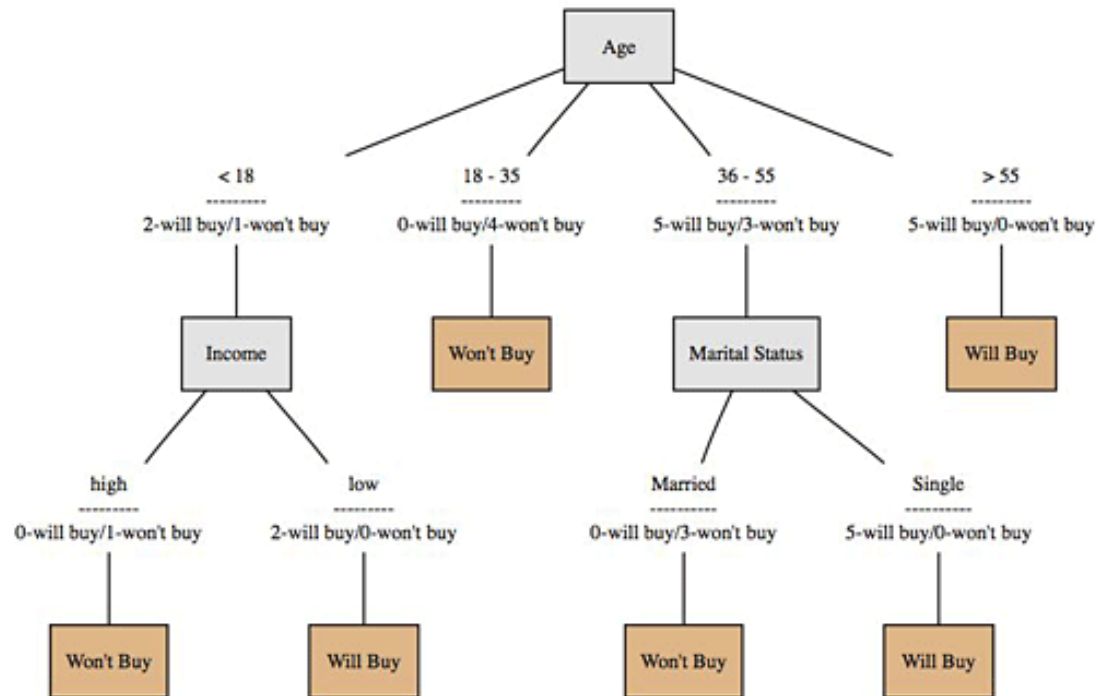
Clustering

Association
Rules

Regressions

Decision Tree

Decision tree is a **decision** support tool that uses a **tree-like** graph or model of **decisions** and their possible consequences, including chance event outcomes, resource costs, and utility.



Application Determine whether an investment will pay off

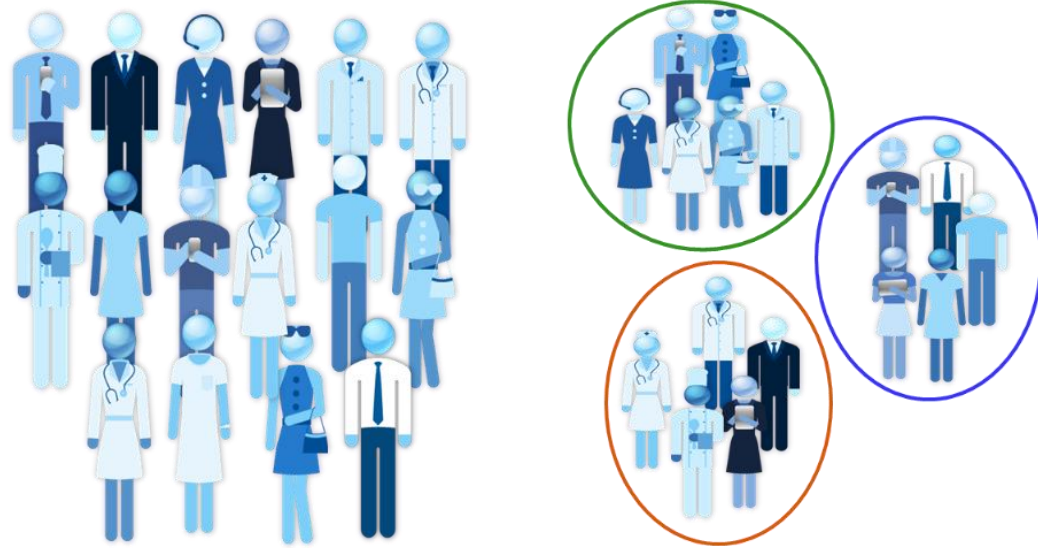
Predict whether a customer will default

Determine the species of an animal

Clustering

Used to group similar items together

Based on similarities between objects, or alternatively distance between objects



Application

Identify similar customers or similar products for recommendations


Optimize good delivery by finding the optimal number of launch locations

Detect insurance or credit card fraud

Association Rule Mining

Find out which events or items go together

Frequently bought together



Total price: **\$410.50**

Add both to Cart

Add both to List

i One of these items ships sooner than the other. [Show details](#)

- ✓ **This item:** Canon EOS Rebel T5 Digital SLR Camera Kit with EF-S 18-55mm IS II Lens **\$397.55**
- ✓ SanDisk 32GB Ultra Class 10 SDHC UHS-I Memory Card Up to 80MB, Grey/Black (SDSDUNC-032G-GN6IN) **\$12.95**

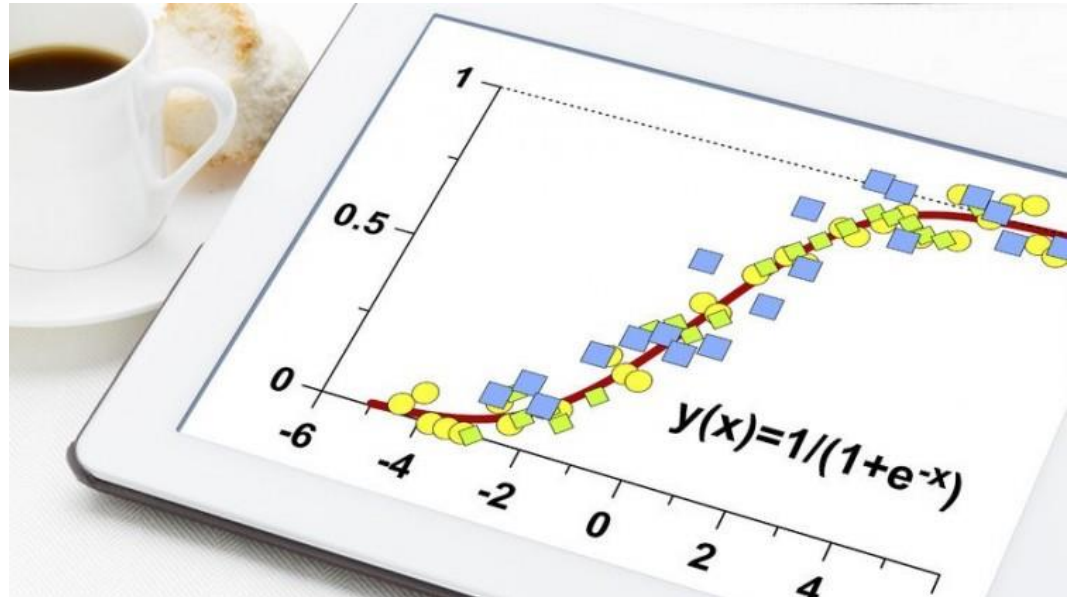
Application What products are bought together?

Amazon's recommendation engine

Medical diagnosis

Regression

Regression is a statistical method that allows you to examine the relationship between two or more variables of interest.

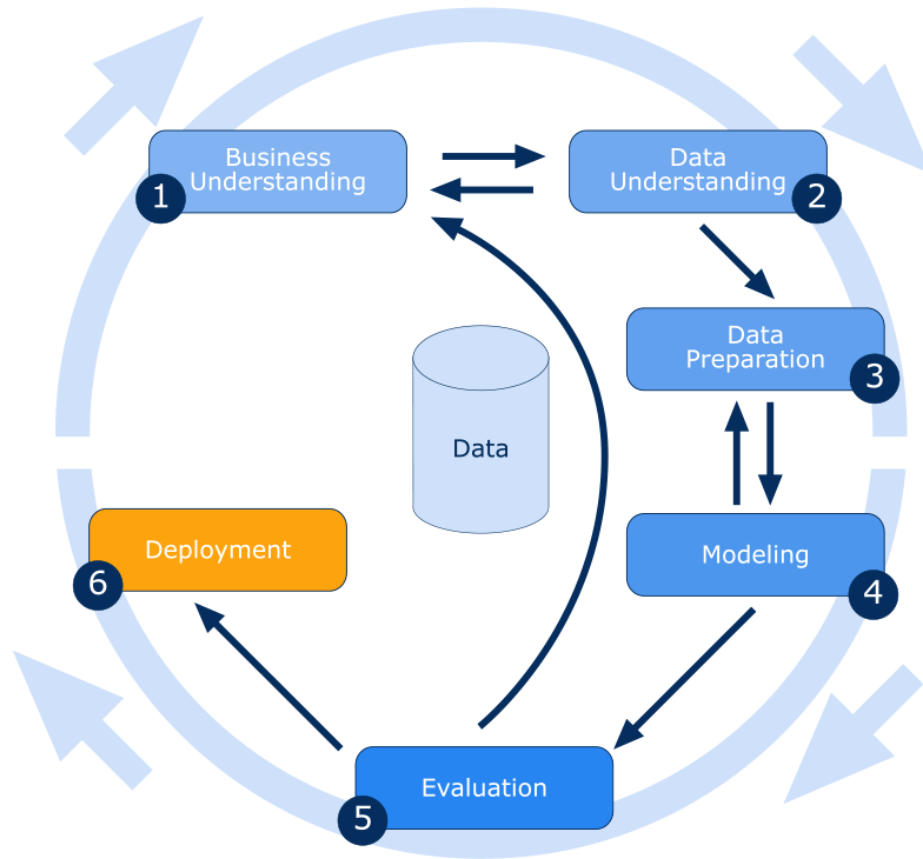


Application Test hypotheses

Estimate the correlation between events

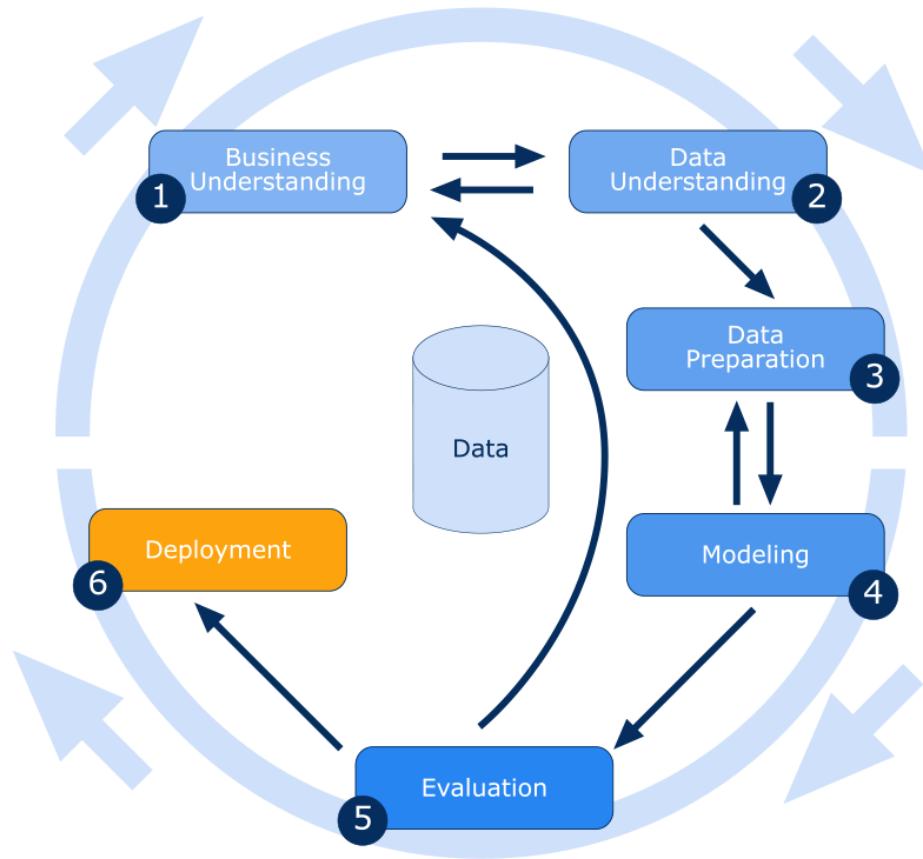
Predict future sales based on historical data

The Data Mining Process:



The CRIPS (Cross Industry Standard Process) Data Mining Process

The Data Mining Process:



Business Understanding:

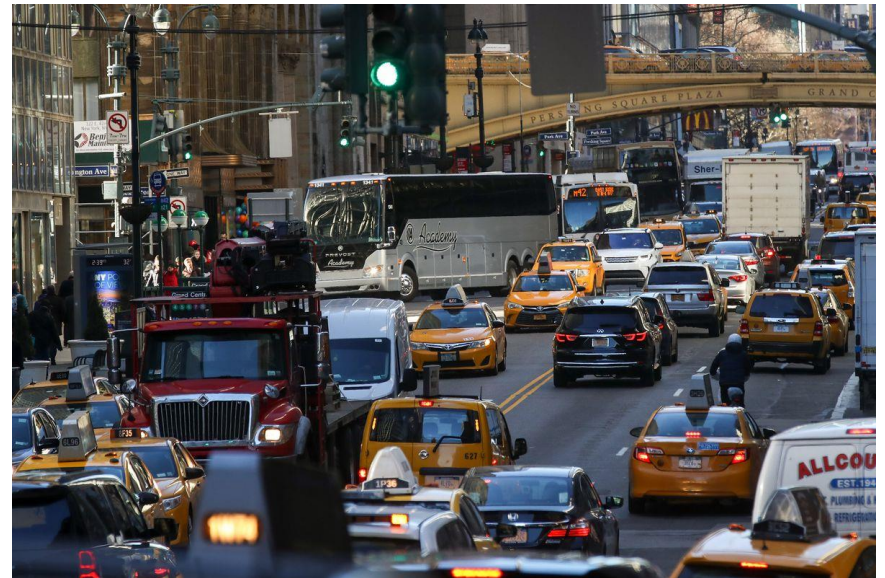
- Understand the business context
- Cast a business problem as one or more data science problems;
- The design team should think carefully about the problem to be solved and about the use scenario.

The CRIPS (Cross Industry Standard Process) Data Mining Process

Example 1:

How to transform this business problem into a data mining problem:

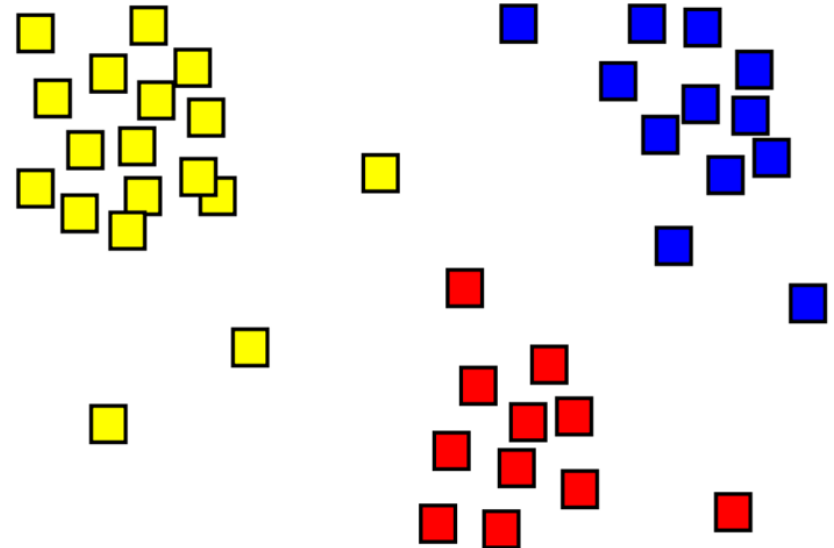
- Do ride-hailing companies like Uber and Lyft make a city's traffic worse or better?
 - Regression
 - Dependent variable: how to measure the performance of city traffic?
 - Explanatory variable
 - Correlation or causality?



Example 2:

How to transform this business problem into a data mining problem:

- How can investors build a diversified portfolio?
 - Clustering
 - How to define diversified portfolio?



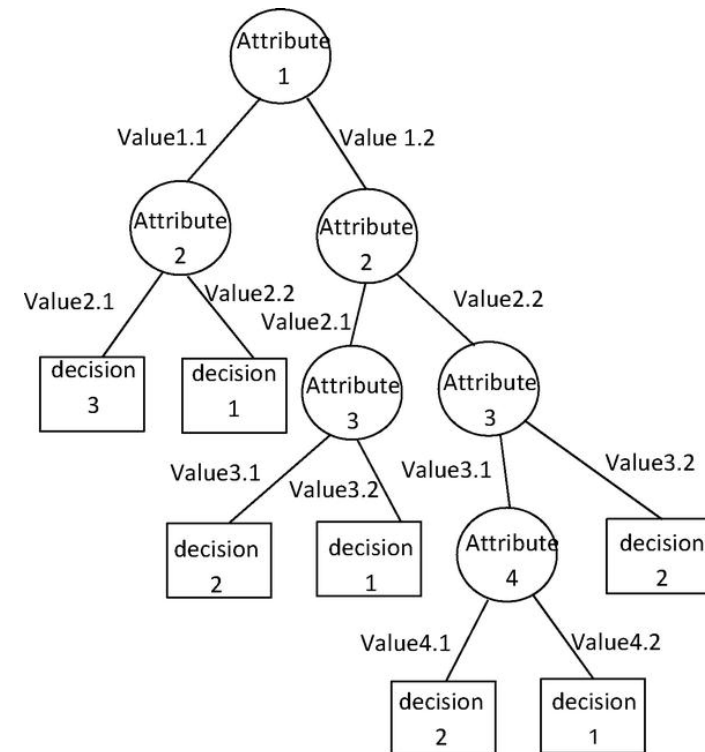
Example 3:

How to transform this business problem into a data mining problem:

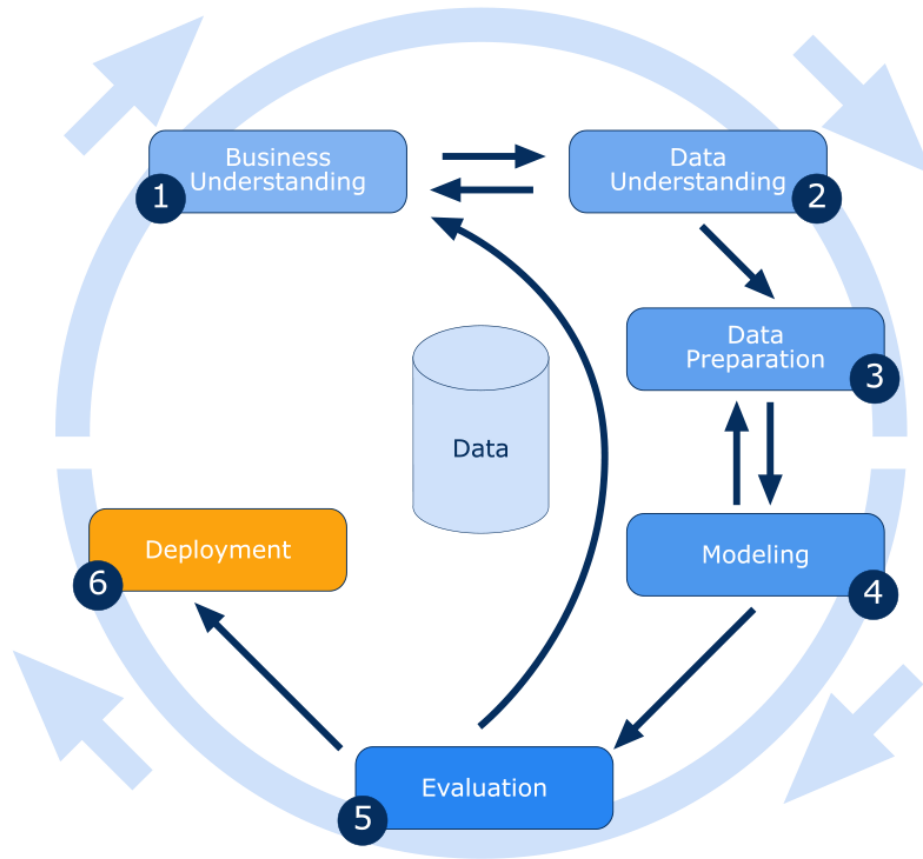
- How does an optician recommend lenses?
 - Decision tree
 - What factors would be considered when opticians recommend lenses?

Table 1.1 **The contact lens data.**

| Age | Spectacle prescription | Astigmatism | Tear production rate | Recommended lenses |
|----------------|------------------------|-------------|----------------------|--------------------|
| young | myope | no | reduced | none |
| young | myope | no | normal | soft |
| young | myope | yes | reduced | none |
| young | myope | yes | normal | hard |
| young | hypermetrope | no | reduced | none |
| young | hypermetrope | no | normal | soft |
| young | hypermetrope | yes | reduced | none |
| young | hypermetrope | yes | normal | hard |
| pre-presbyopic | myope | no | reduced | none |
| pre-presbyopic | myope | no | normal | soft |
| pre-presbyopic | myope | yes | reduced | none |
| pre-presbyopic | myope | yes | normal | hard |
| pre-presbyopic | hypermetrope | no | reduced | none |
| pre-presbyopic | hypermetrope | no | normal | soft |
| pre-presbyopic | hypermetrope | yes | reduced | none |
| pre-presbyopic | hypermetrope | yes | normal | none |
| presbyopic | myope | no | reduced | none |
| presbyopic | myope | no | normal | none |
| presbyopic | myope | yes | reduced | none |
| presbyopic | myope | yes | normal | hard |
| presbyopic | hypermetrope | no | reduced | none |
| presbyopic | hypermetrope | no | normal | soft |
| presbyopic | hypermetrope | yes | reduced | none |
| presbyopic | hypermetrope | yes | normal | none |



The Data Mining Process:

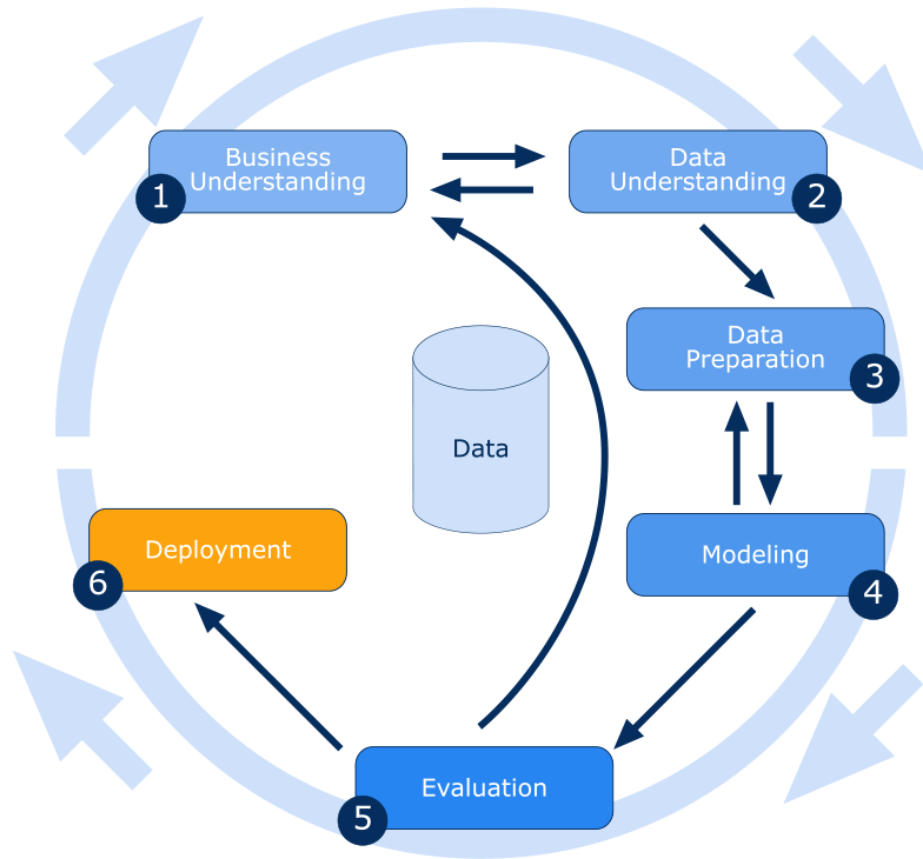


Data Understanding:

- Understand the strengths and limitations of the data;
- Estimate the costs and benefits of each data source and decide whether further investment is merited;

The CRIPS (Cross Industry Standard Process) Data Mining Process

The Data Mining Process:

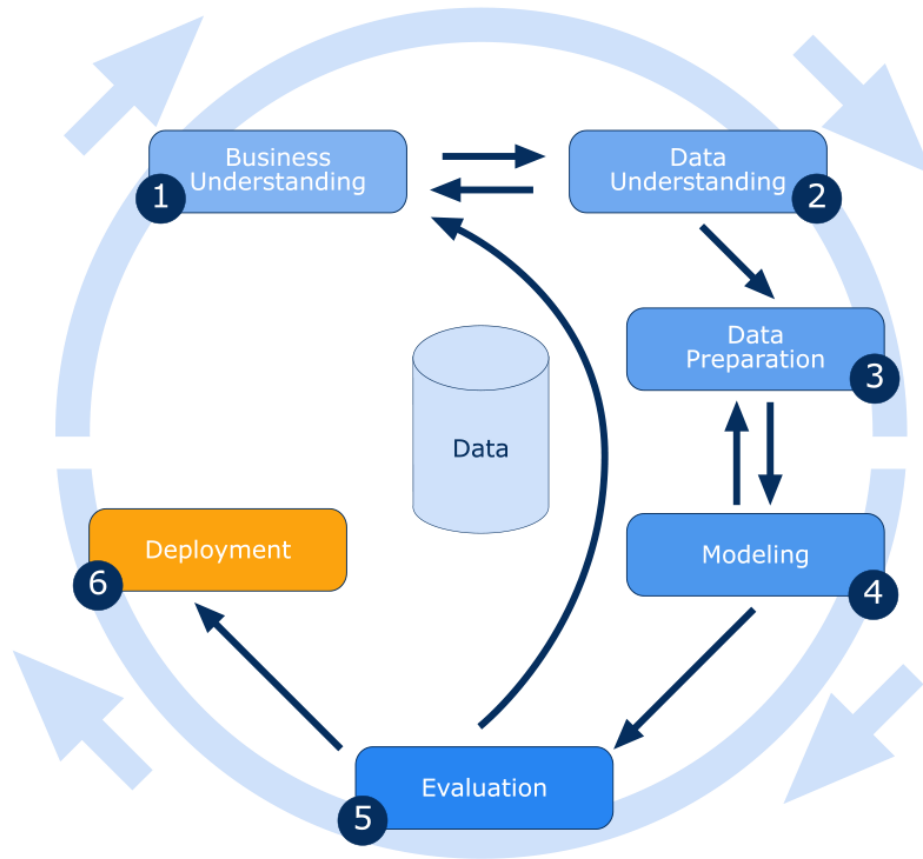


Data Preparation:

- Collect and clean data
- Manipulate and convert data into forms that yield better results

The CRIPS (Cross Industry Standard Process) Data Mining Process

The Data Mining Process:

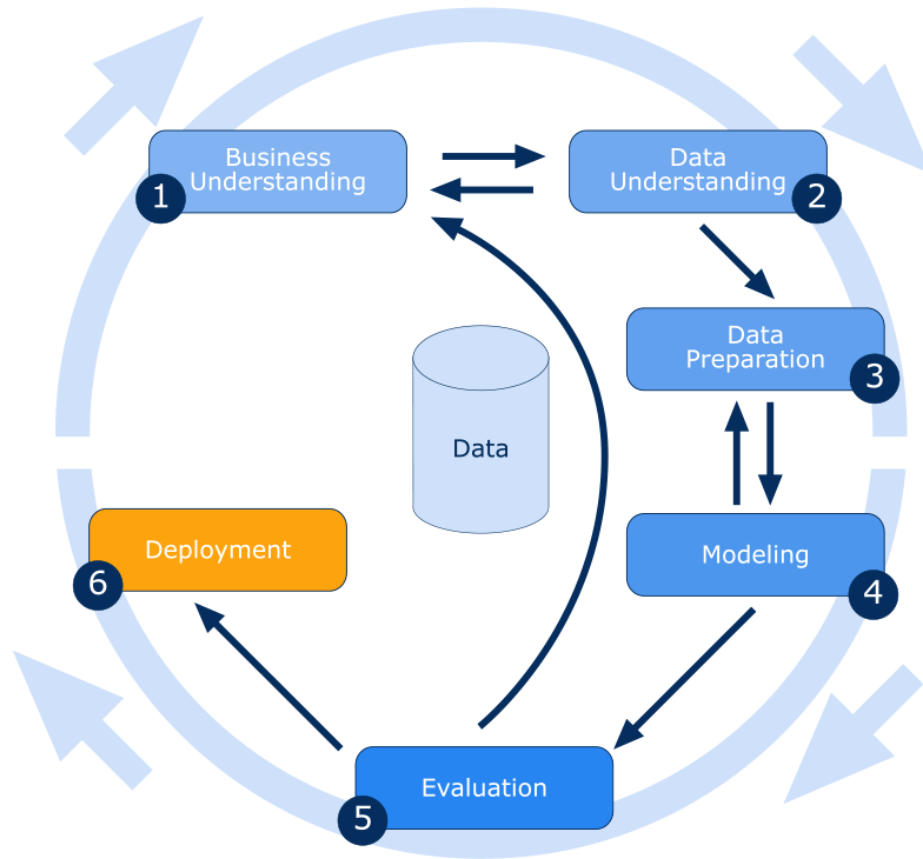


Modeling:

- Apply data mining techniques to the data
- No universally best methods or algorithms for data mining tasks

The CRIPS (Cross Industry Standard Process) Data Mining Process

The Data Mining Process:

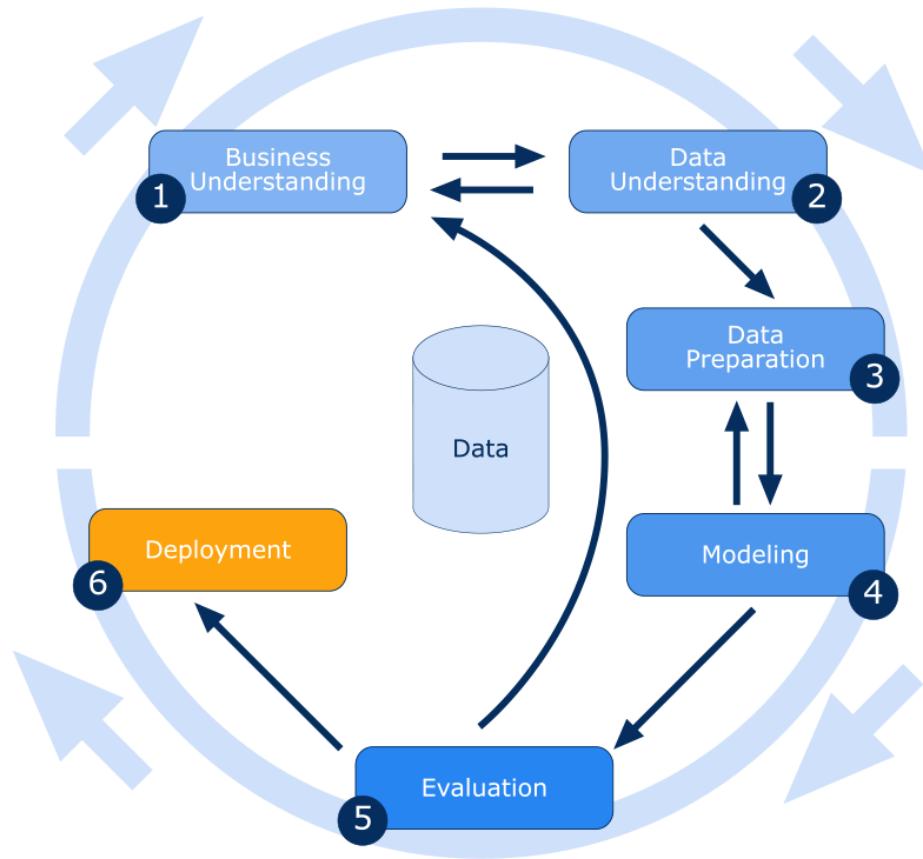


Evaluation:

- Assess the data mining results rigorously and evaluate if they are valid and reliable before moving on
- Easier, cheaper, quicker, and safer to test a model first in a controlled laboratory setting

The CRIPS (Cross Industry Standard Process) Data Mining Process

The Data Mining Process:



Deployment:

- The results of data mining—and increasingly the data mining techniques themselves—are put into real use in order to realize return on investment
- Often return to the Business Understanding phase for a second iteration, which can yield an improved solution;
- Results are not always worth deploying

The CRIPS (Cross Industry Standard Process) Data Mining Process

Problems of Data Mining

- Real data is imperfect
 - Some parts will be garbled, and some will be missing
- Many patterns will be banal and uninteresting. Others will be spurious, contingent on accidental coincidences in the particular dataset used.
- Anything discovered will be inexact
 - There will be exceptions to every rule and cases not covered by any rule
- Algorithms need to be robust enough to cope with imperfect data and to extract regularities that are inexact but useful.
- We also need to keep in mind that the data to which we will apply data science techniques are the product of some process that involved human decisions. We should not think that the data represent objective truth.

Practice:

Using data mining techniques to answer the following question: Is Grad School Worth the Cost?

- Transform the problem into a data mining problem
- Where can you get data for the analysis
- How can you evaluate your analysis?



Install Python and PyCharm

- <https://www.youtube.com/watch?v=kqtD5dpn9C8>

Install Packages in PyCharm

- <https://www.youtube.com/watch?v=zCO3KxV2zPI>