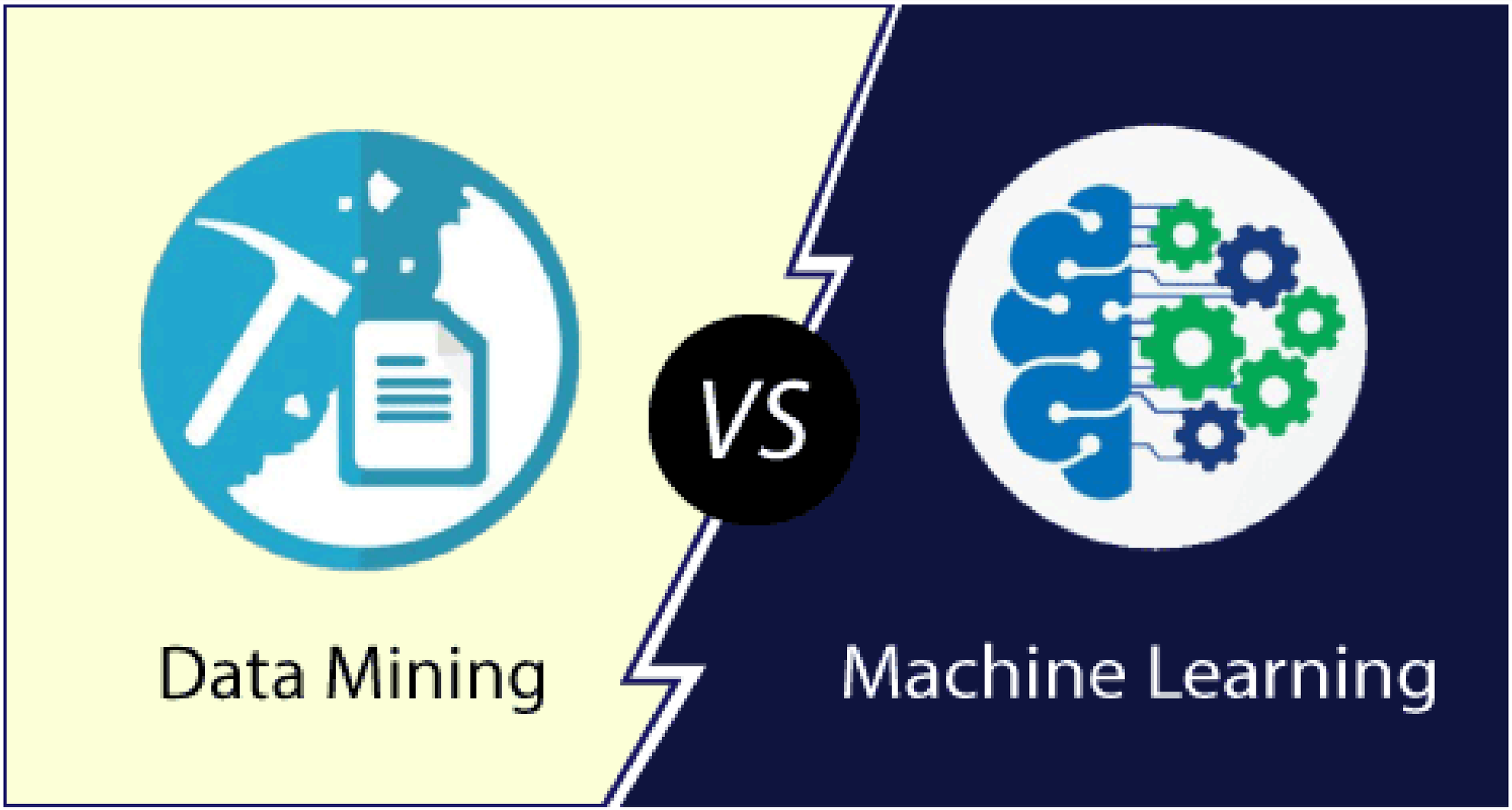
Data Mining

Data Mining Introduction

Overview

- What is Data Mining
- Data Mining and Machine Learning
- Data Mining: applications, patterns, tools
- Data Mining Process: CRISP-DM
- The Six Phases in CRISP-DM

Thinking

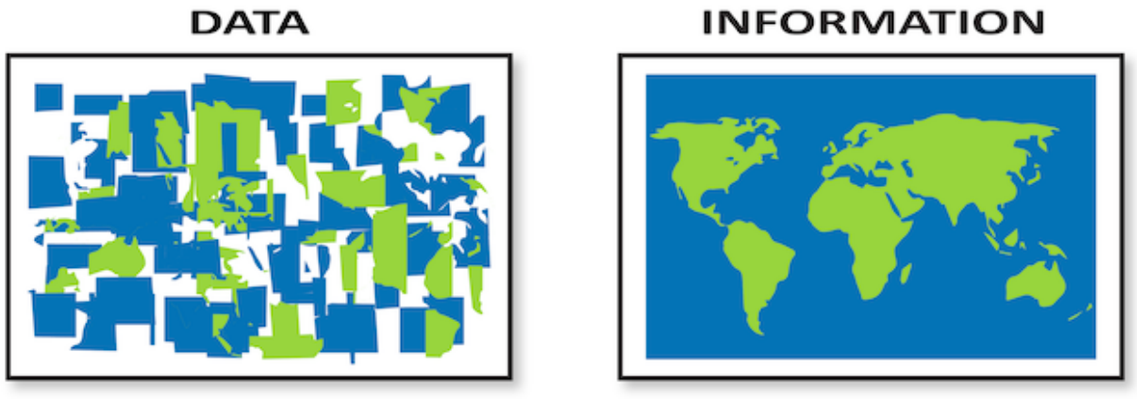


What differences and relationships between data mining and machine learning?

https://www.javatpoint.com/data-mining-vs-machine-learning

From Data to Information

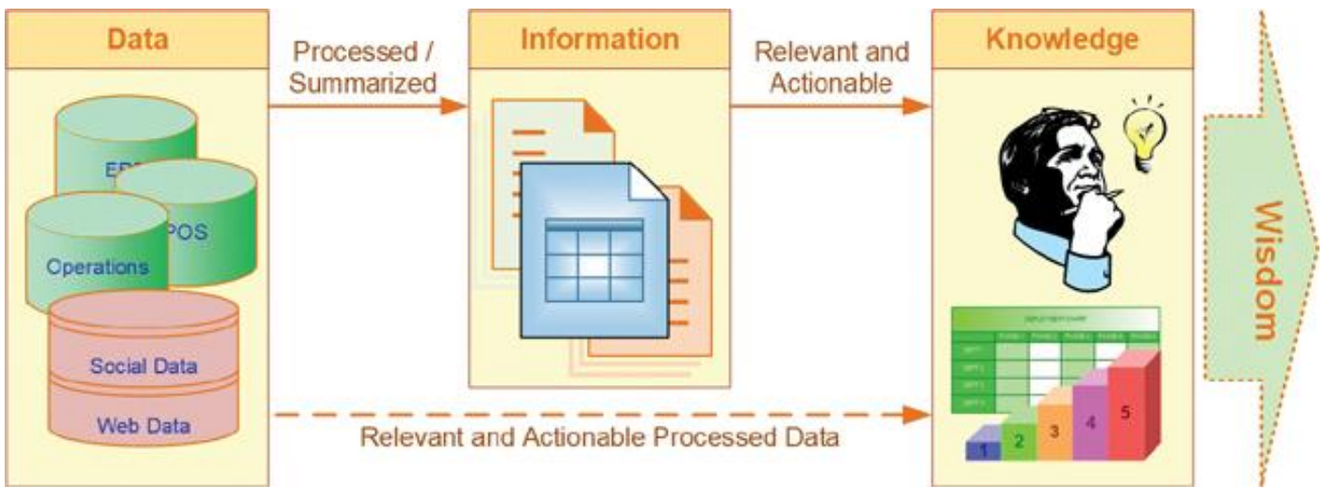
- Society produces huge amounts of data
  - Sources: E-commerce and Retail, Social media platforms, Transportation and Logistics, Healthcare and medicine, Economics, Geography, Environment, Sports, …
- Data is potentially valuable resource but raw data is useless
- Need techniques to automatically extract information from data
  - Data: recorded facts
  - Information: patterns underlying the data



https://effectualsystems.com/data-need-information/

Data Mining

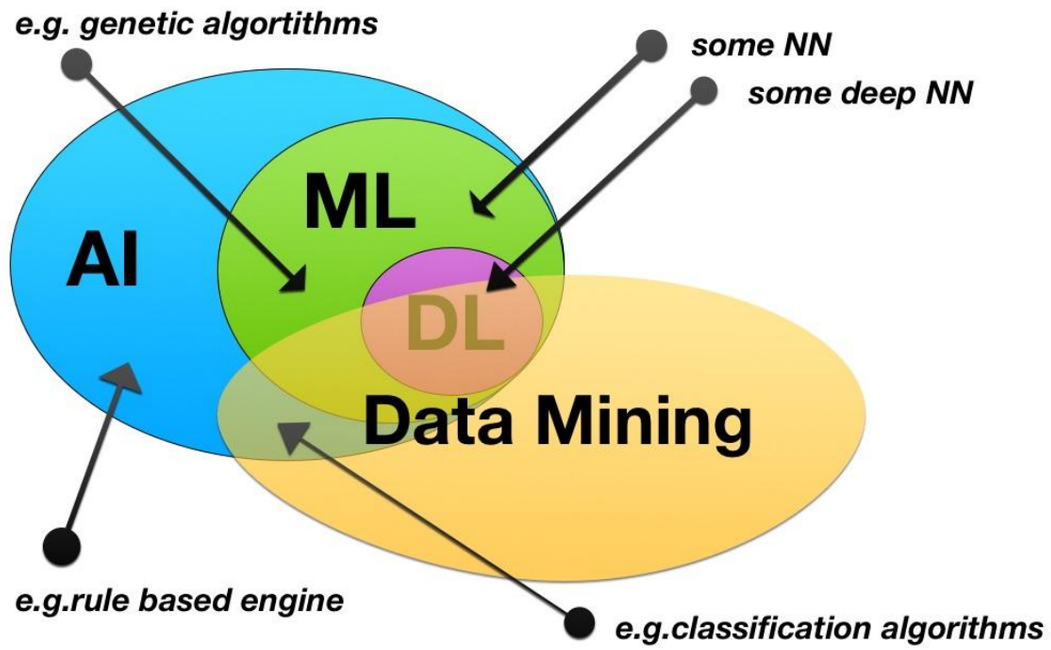
- Data Mining: AKA knowledge discovery in databases, "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases" – Fayyad et al.(1996)
    - process : data mining comprises many iterative steps
    - nontrivial: some experimentation-type search or inference is involved
    - valid: the discovered patterns should hold true on new data
    - novel: the patterns were not previously known to the user
    - potentially useful: lead to some benefit to the user or task
    - understandable: make business sense



Fayyad, Usama M., Gregory Piatetsky-Shapiro, and Padhraic Smyth. "Knowledge Discovery and Data Mining: Towards a Unifying Framework." In KDD, vol. 96, pp. 82-88. 1996.
Delen, Dursun. Predictive Analytics: Data Mining, Machine Learning and Data Science for Practitioners. FT Press, 2020.

Machine learning and Data Mining

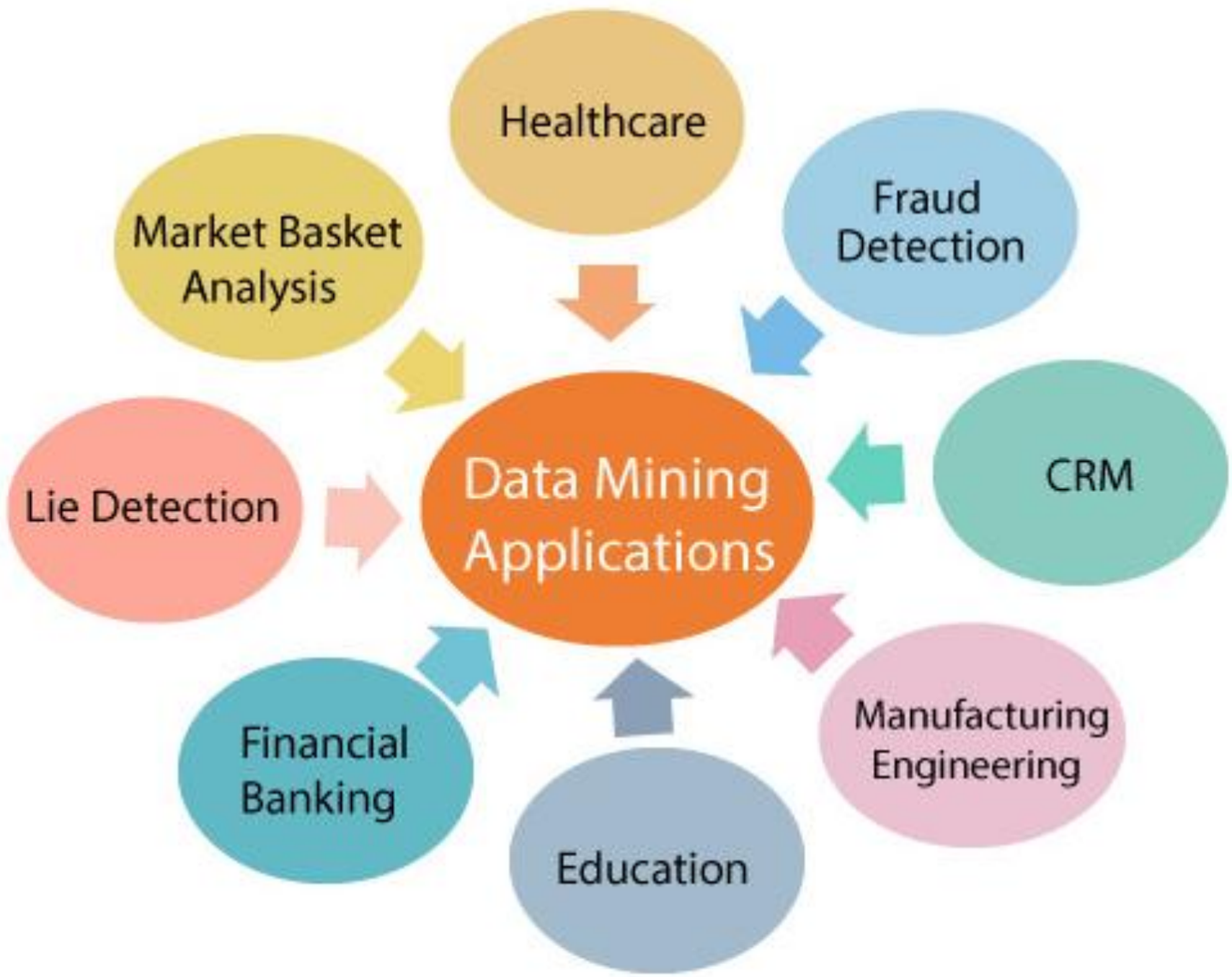
- Machine learning: new technology for mining knowledge from data
  - automatically finding patterns in data
  - finding structural descriptions of what is learned: explicit representation of the knowledge
  - helping to explain that data and make predictions



https://softwareengineering.stackexchange.com/questions/366996/distinction-between-ai-ml-neural-networks-deep-learning-and-data-mining

Typical DM Applications

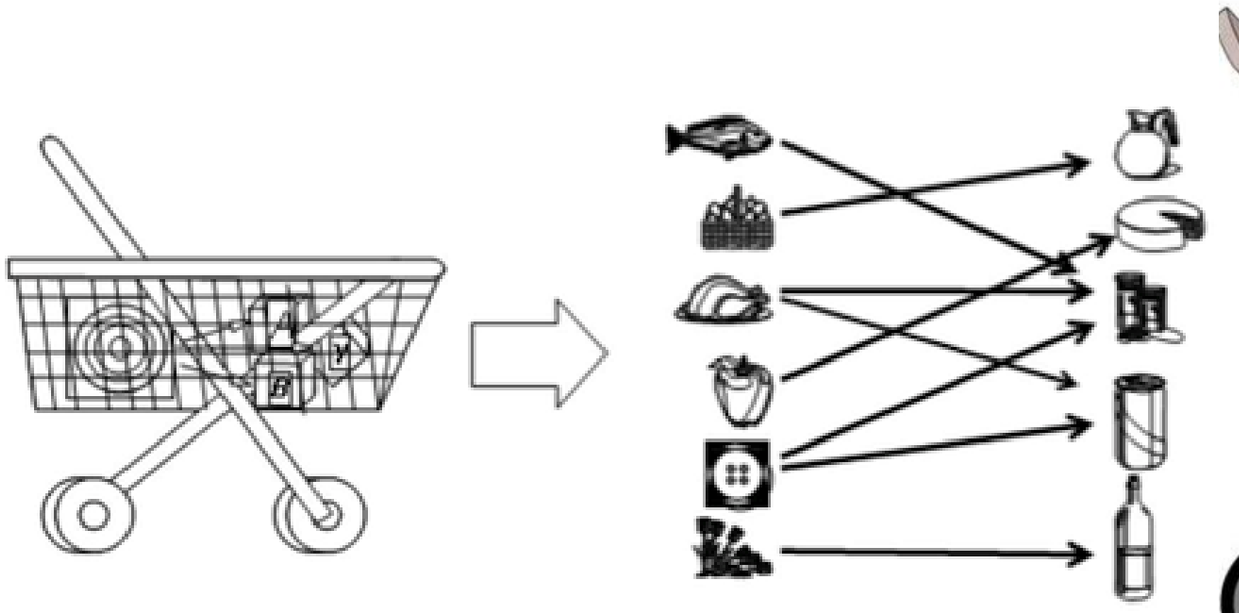
- Data mining: address many complex business problems/ opportunities
- being very successful and helpful in many areas



Marketing and Sales

Market Basket Analysis:

- find groups of items that tend to occur together in transactions
- Product association anlaysis, use association rule mining
- find association between different objects in a set
- marketers use these rules to strategize their recommendations



98% of people who purchased items A and B also purchased item C

https://www.analyticsvidhya.com/blog/2014/08/effective-cross-selling-market-basket-analysis/

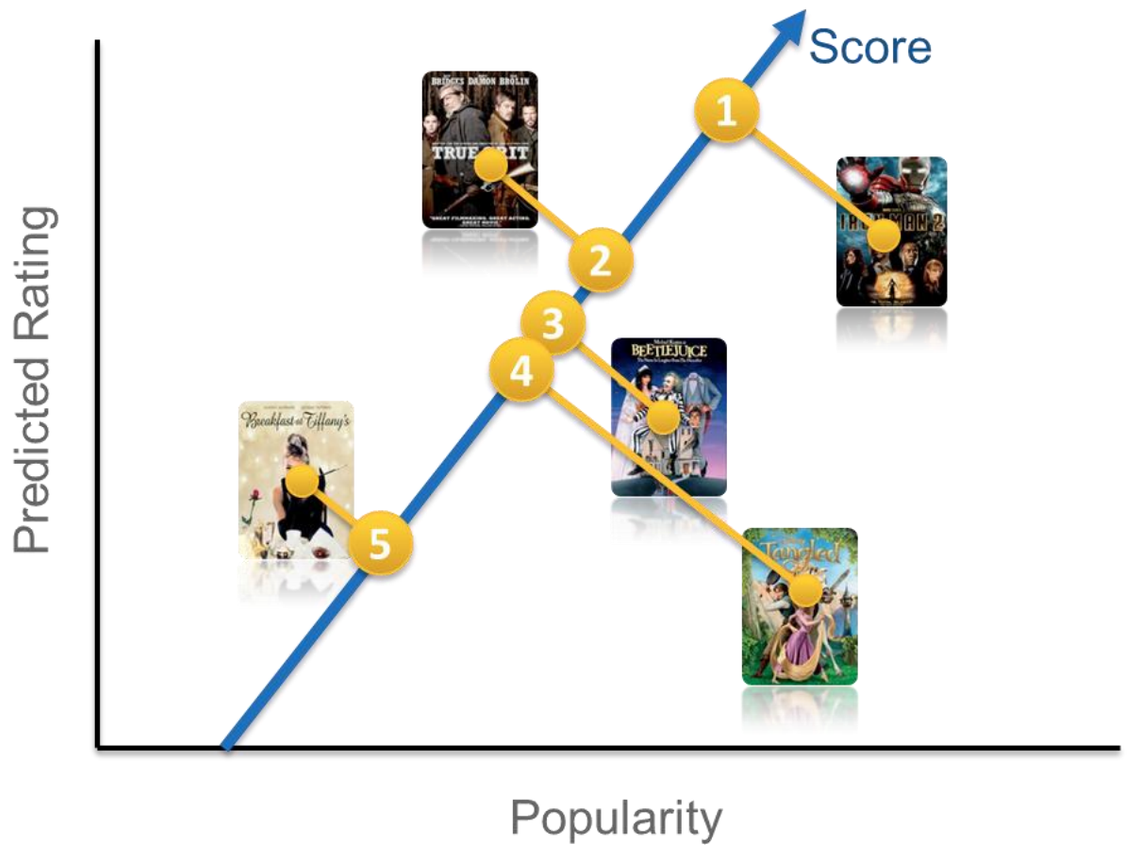
Healthcare

- Healthcare Management: identify chronic diseases, track high-risk regions reduce the spread of disease
- Effective Treatments: continuous comparison of symptoms, causes, and medicines

http://www.datamining-international.com/ehealth/

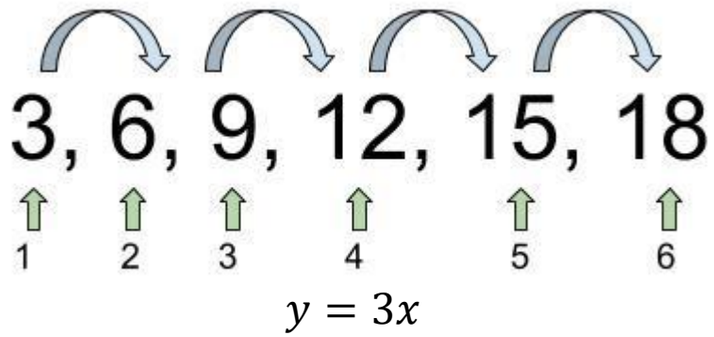
DM in Netflix

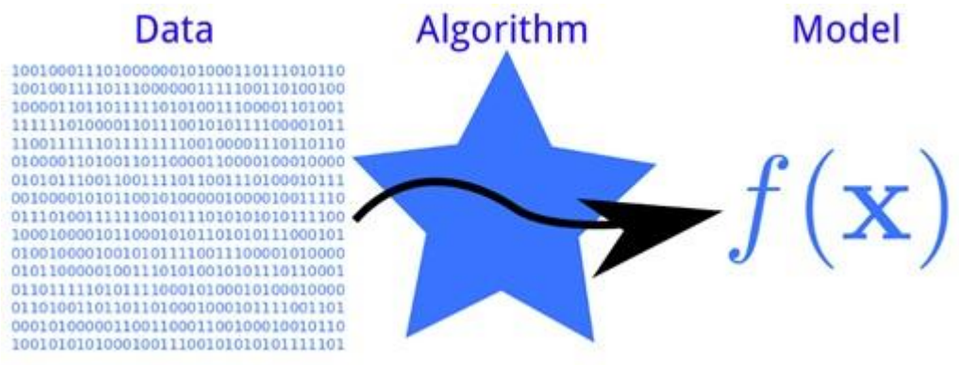
- Netflix: over 190 countries, over 238 million subscriber by the second quarter of 2023, several billion items
- Netflix's recommender system: achieve 80% of stream time
- present a number of attractive items for a person to choose from, to find a personalized ranking function
- produce rankings that balance popularity and predicted rating



Patterns Discovered

- Using the observed data, data mining builds models to identify patterns among the attributes /variables.
- models are usually the mathematical representations that identify the relationships among the attributes of the objects
- patterns can be explanatory - explain the interrelationships and affinities among the attributes, valuable for gaining insights into the data's behavior, identifying important factors
- Can also be predictive – predict future values of certain attributes, forecast outcomes, estimate values, or classify new data points into predefined categories.

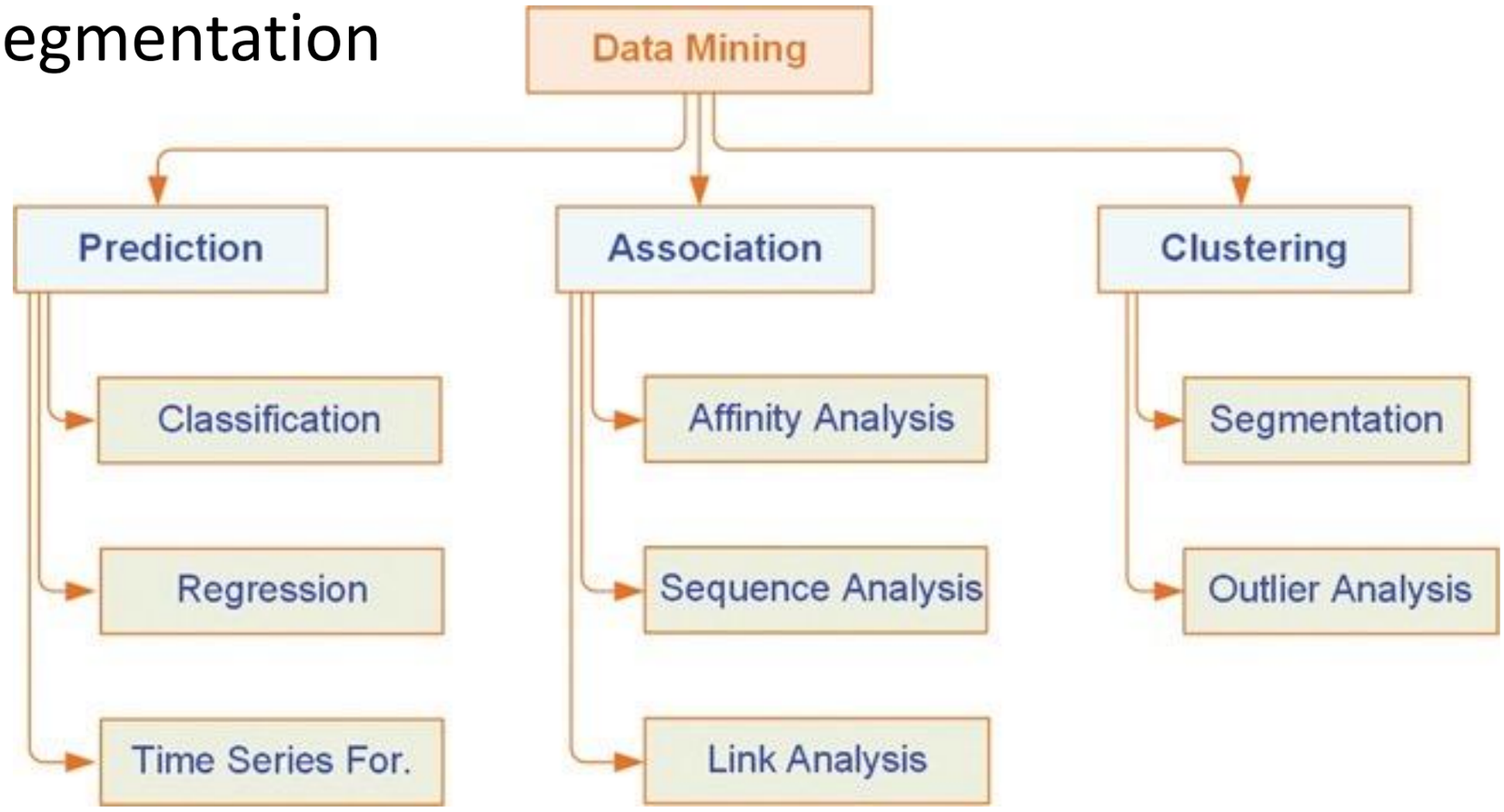




Three Major Types of Patterns

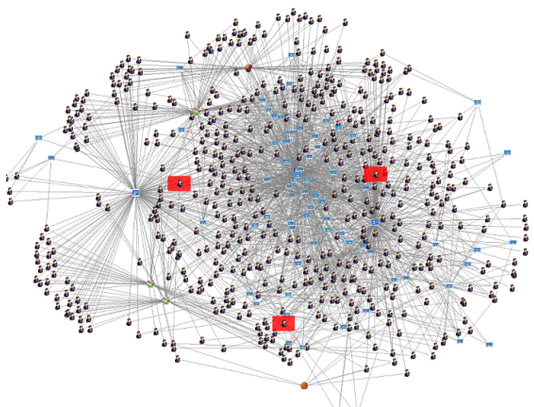
- Associations: find commonly co-occurring groupings of things, captures the sequences of things
- Prediction: tell the nature of future occurrences of certain events based on what has happened in the past
- Clustering: identify natural groupings of things based on their known characteristics: homogeneous Groups, anomaly detection,

Customer Segmentation



*A*ssociation Detection

- AKA association rule mining: discover relationships among variables in large databases
- link analysis: automatically discover the links among objects
    - Social Network Analysis: Studying connections between individuals in social networks to identify influencers, communities, or information flow.
    - Web Link Analysis: Analyzing hyperlinks on the web to determine the importance and relevance of web pages (e.g., Google's PageRank algorithm).
- sequence mining: find statistically relevant patterns between data examples that are delivered in a sequence
    - analyzing DNA/ protein sequences; determine buying patterns



Customer ID (CID)	TID	Itemset
1	1	$\{a, b, d\}$
1	3	$\{b, c, d\}$
1	6	$\{b, c, d\}$
2	2	$\{b\}$
2	4	$\{a, b, c\}$
3	5	$\{a, b\}$
3	7	$\{b, c, d\}$

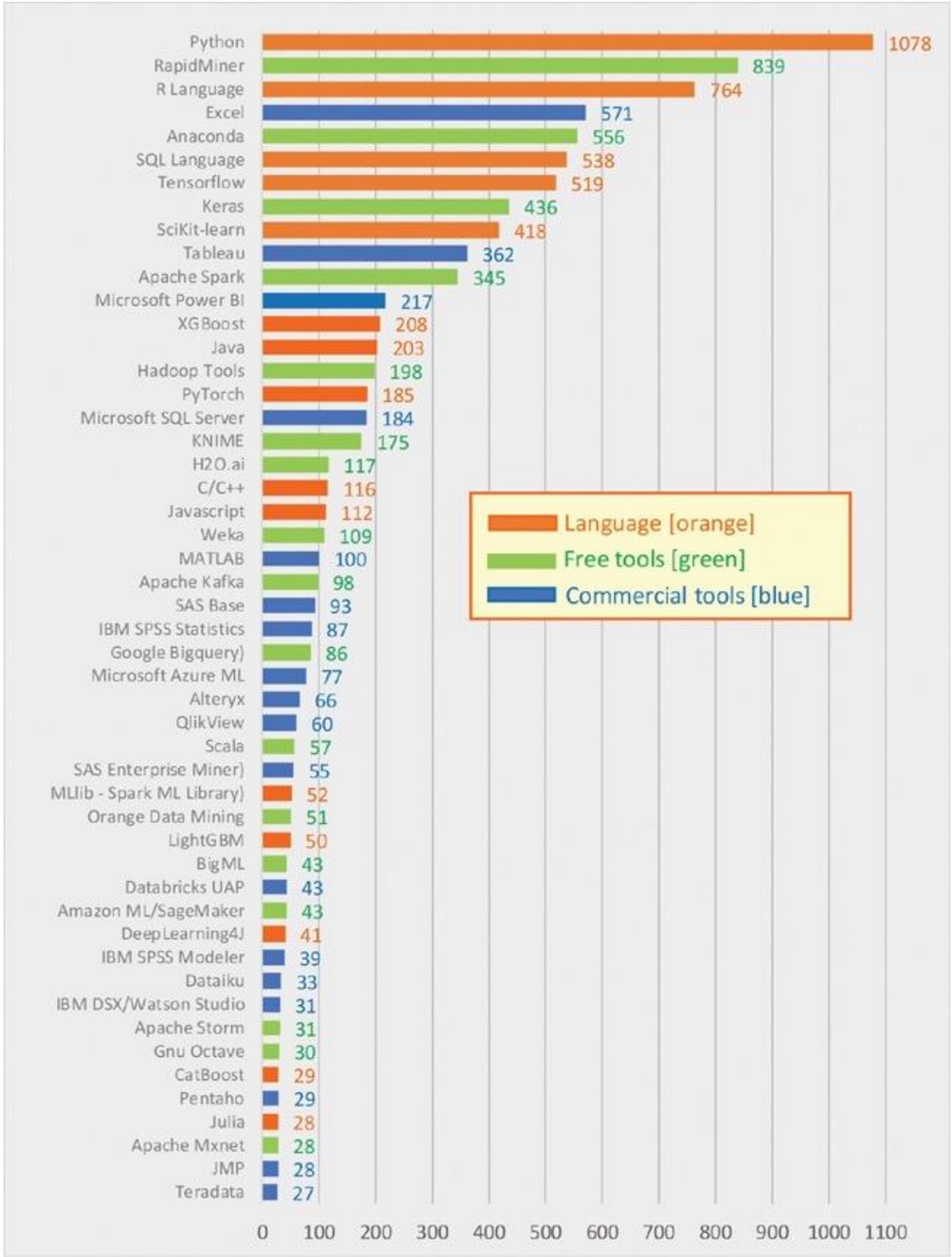
CID	Sequence
1	$(\{a, b, d\}, \{b, c, d\}, \{b, c, d\})$
2	$(\{b\}, \{a, b, c\})$
3	$(\{a, b\}, \{b, c, d\})$

Data Mining Tools

- commercial tools:
  - IBM: SPSS Modeler
  - SAS: Enterprise Miner
  - Quest(StatSoft): STATISTICA Data Miner
- open source and/or free tools
  - Weka: from NZ (cs.waikato.ac.nz/ml/weka/)
  - RapidMiner (rapidminer.com)
  - Orange (orangedatamining.com)
  - KNIME Analytics Platform (knime.org)
  - computational efficiency, take longer for large datasets
- code-based tools: Python, R and JavaScript

Data Mining Tools

- KDnuggets 20th annual software poll on “what analytics, data science, machine learning software/tools have you used in the last three years (2017–2019) for a real project?”
- attracting more than 1,800 unique voters
- many users use more than one tool, the average number of tools used by a person or vendor was 6.1 in 2019 (compared to 3.7 in 2014).
- the most popular tool was Python (65% of the votes), as was the case in 2018; popularity of open-source tools and programming languages far exceeded commercial tools

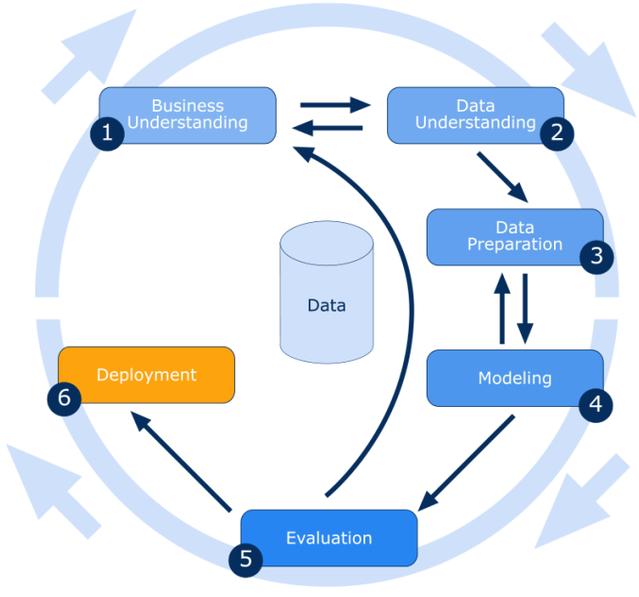


https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html

How Data Mining Works

- Need a standardized process
  - as a framework for recording experience
  - consistent: ensures that data mining projects are conducted in a consistent and repeatable manner
  - Efficient: reduce unnecessary rework
  - **transparency, effective communication, risk mitigation, aids project management, …**
- Several standardized processes have been developed: **SEMMA, KDD Process, TDSP, CRISP-DM**

- Most popular one: Cross-industry standard process for data mining (CRISP-DM)



The Dark Side of Data Mining

- Privacy issues: company share customer data with others without seeking the explicit consent of their customers
- Ethical issues, Discrimination and Bias : discriminating prediction models, catastrophic results of financial/political decisions
- Security issues: blackmail, damage reputation, identity theft
  - leak of data mine out from the cyberspace, e.g. fingerprint data
- Manipulation issues: the authenticity of the output of data mining
  - manipulative information makes data miners mine incorrect data
  - social media platforms propaganda false information/fake news
- Ethical and legal obligations: de-identify customer records prior to applying data mining applications

Data Mining

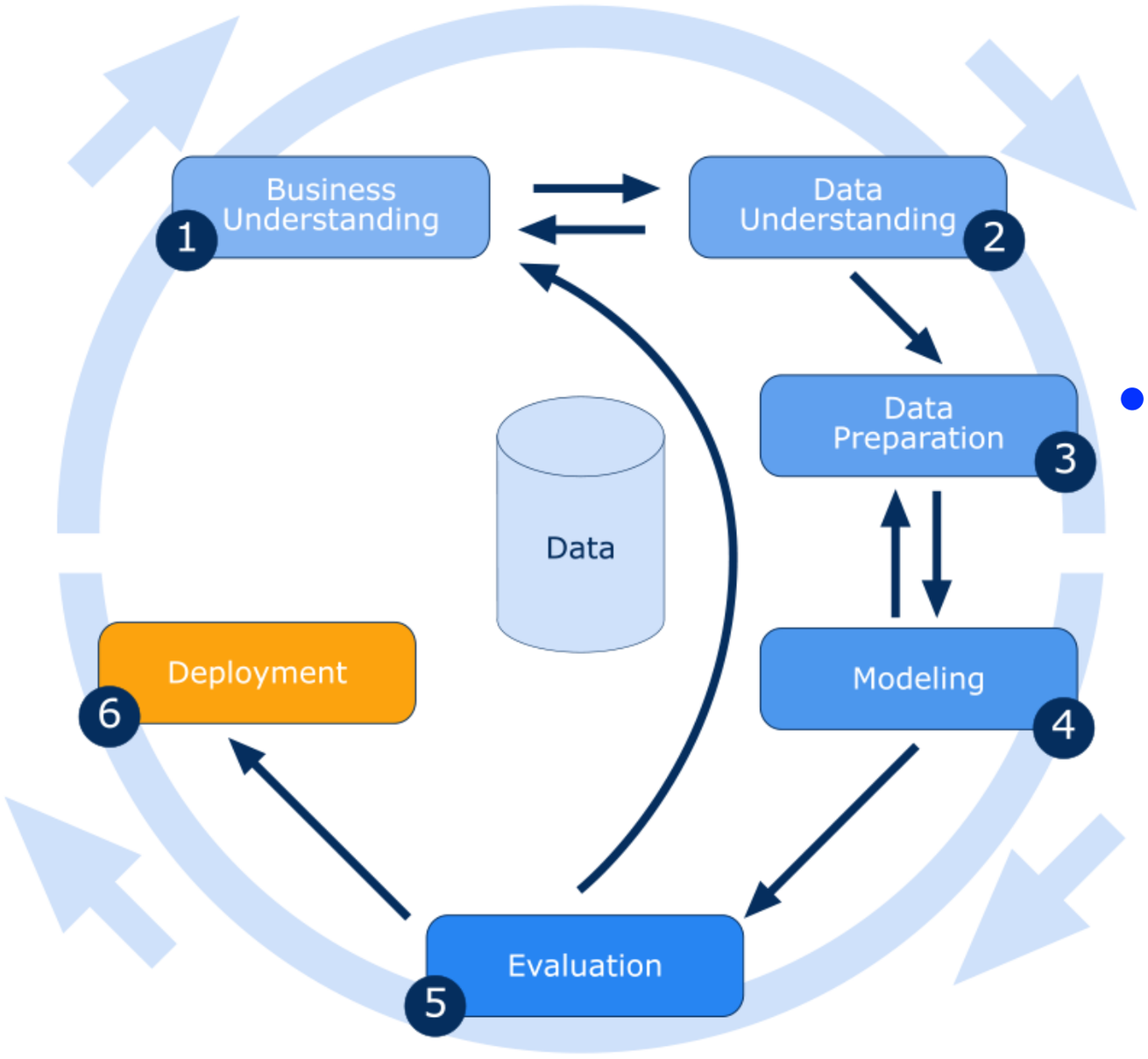
CRISP - Data Mining Process

Overview

- What is Data Mining
- Data Mining and Machine Learning
- Data Mining: applications, patterns, tools
- Data Mining Process: CRISP-DM
- The Six Phases in CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM)

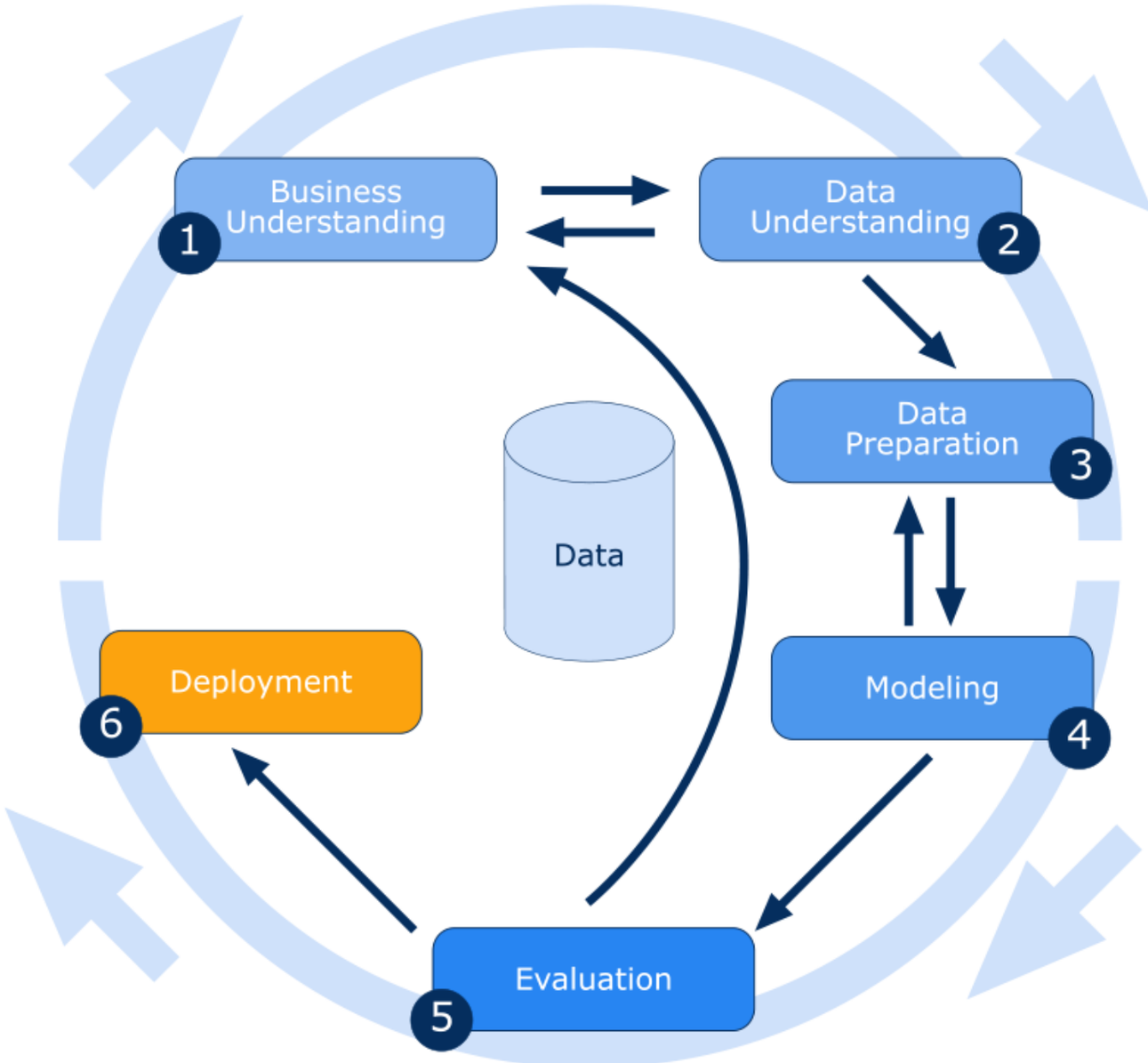
CRISP-DM is a widely used, non-proprietary, and industry agnostic methodology and procedures for best practices in data mining



- CRISP-DM is a **methodology**
  - descriptions of typical phases
  - tasks involved with each phase
  - an explanation of the relationships between tasks
- CRISP-DM is a **process model**
  - provides an overview of the data mining life cycle

Cross Industry Standard Process for Data Mining (CRISP-DM)

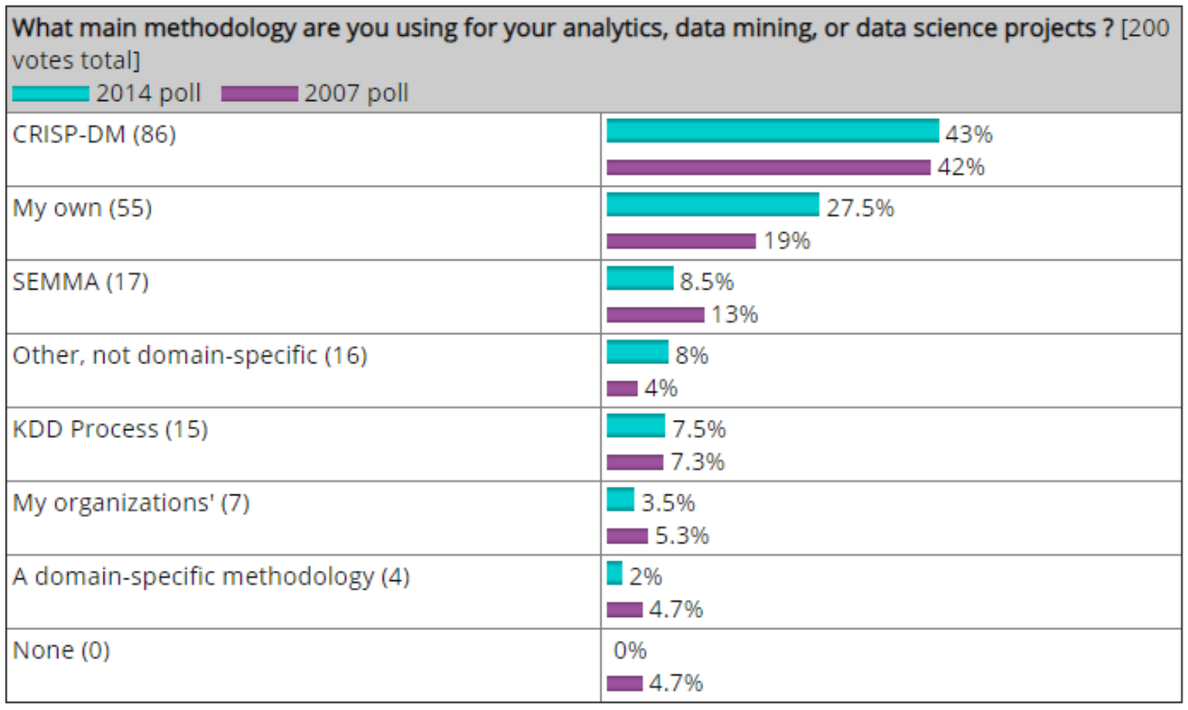
Developed by big players (CRISP-DM Special Interest Group) in data analysis in 1996



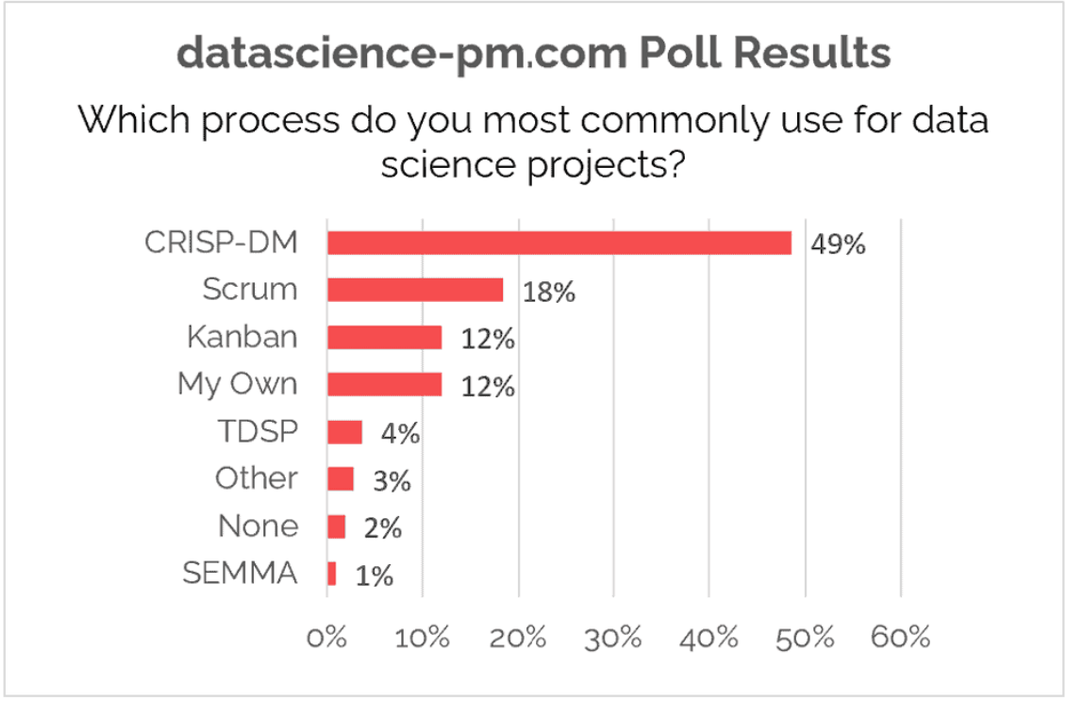
- consists of six phases with arrows indicating dependencies
- sequence of the phases is not strict
- flexible and can be customized easily
- can revisit phases

Why Successful?

- It's simple, only has six phases
- It's easy to implement
- domain-agnostic, works for industry and research communities



https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

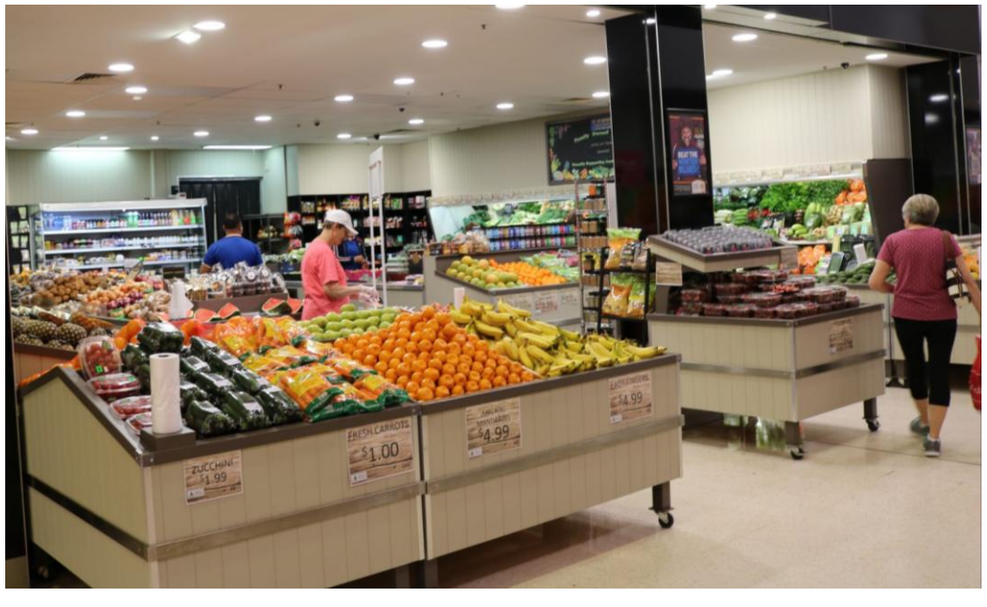


https://www.datascience-pm.com/crisp-dm-2/

Example-DM Task

"FruitMart" is a leading supermarket chain with stores all over the country. They have a vast selection of fruit products, including fresh fruits, canned fruits, and fruit-based snacks.

To stay competitive in the market, they decide to use data mining techniques to gain insights into their customers' preferences and behaviors, and subsequently, enhance their marketing efforts.



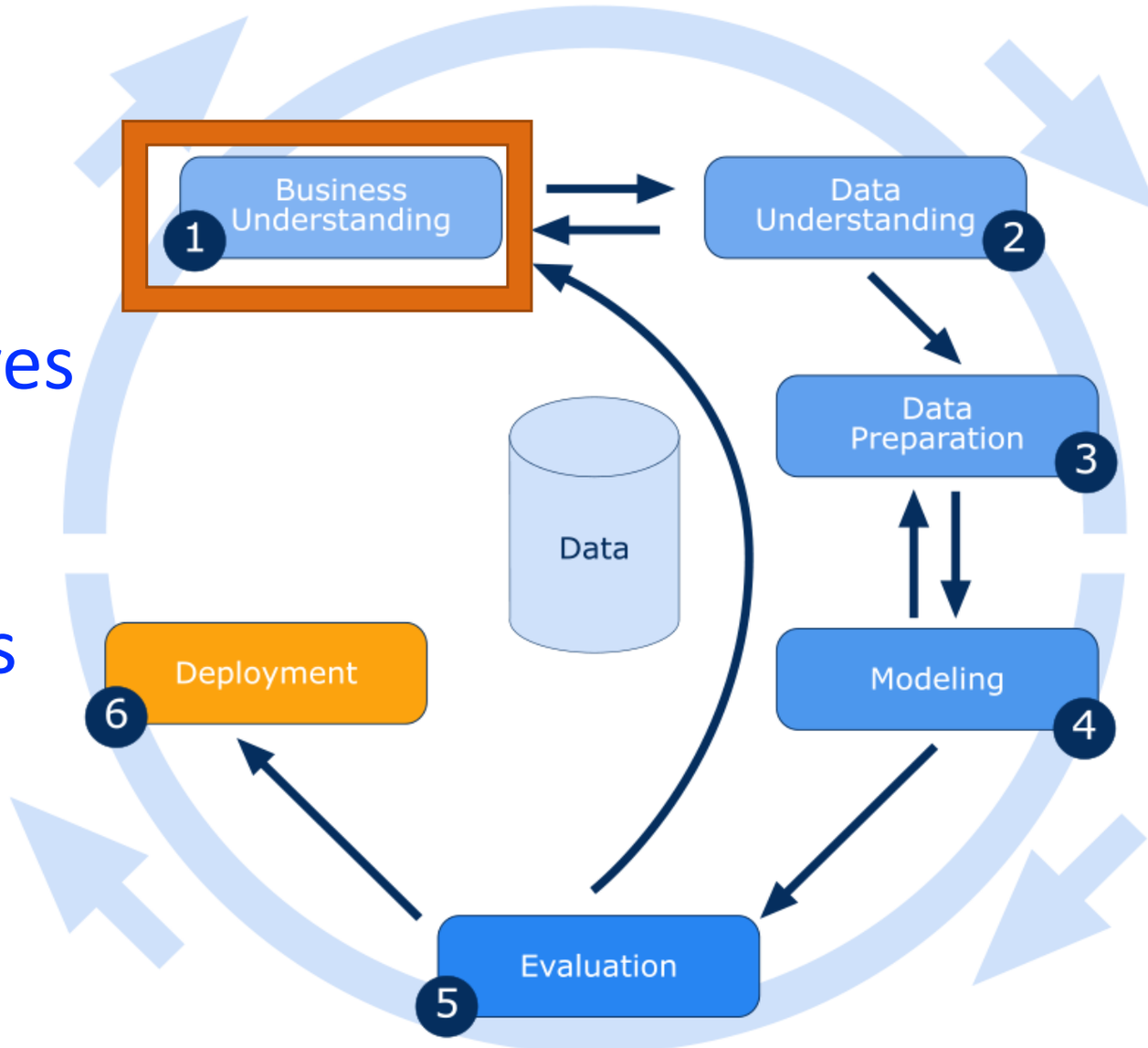
Business Understanding

Gain a true understanding of the business, and to identify the specific goals and problems a business wish to solve.

What does the business need?

This phase includes four tasks

- determine business objectives
- assess situation
- determine data mining goals
- produce project plan



Determine Business Objectives

- Business Objectives
  - describing primary objective from a business perspective
  - other related questions that would like to address

***Primary goal**: Inventory Optimization, to optimize the inventory management system, have the right amount of fresh fruits in stock to meet customer demand while minimizing wastage due to overstocking.*

***Related questions**:*

- *How customer preferences and purchasing patterns vary across different stores?*
- *Which fruits have the highest and lowest sales volumes in different time periods?*
- *How do promotions and discounts affect sales volumes and customer purchasing behaviour for different fruits?*

Determine Data Mining Goals

Covert business objectives to the definition of data mining problem and goals, state project objectives in technical terms

***Business Goal**: Inventory Optimization*

***Data Mining Goal**: Demand Forecasting, Market basket analysis to discover association rules among products, Customer Segmentation, Develop personalized recommendation models*

***Business Goal**: Sales Improvement*

***Data Mining Goal**: Customer Profiling, Outlier Detection*

- Define the data mining success criteria, might be subjective

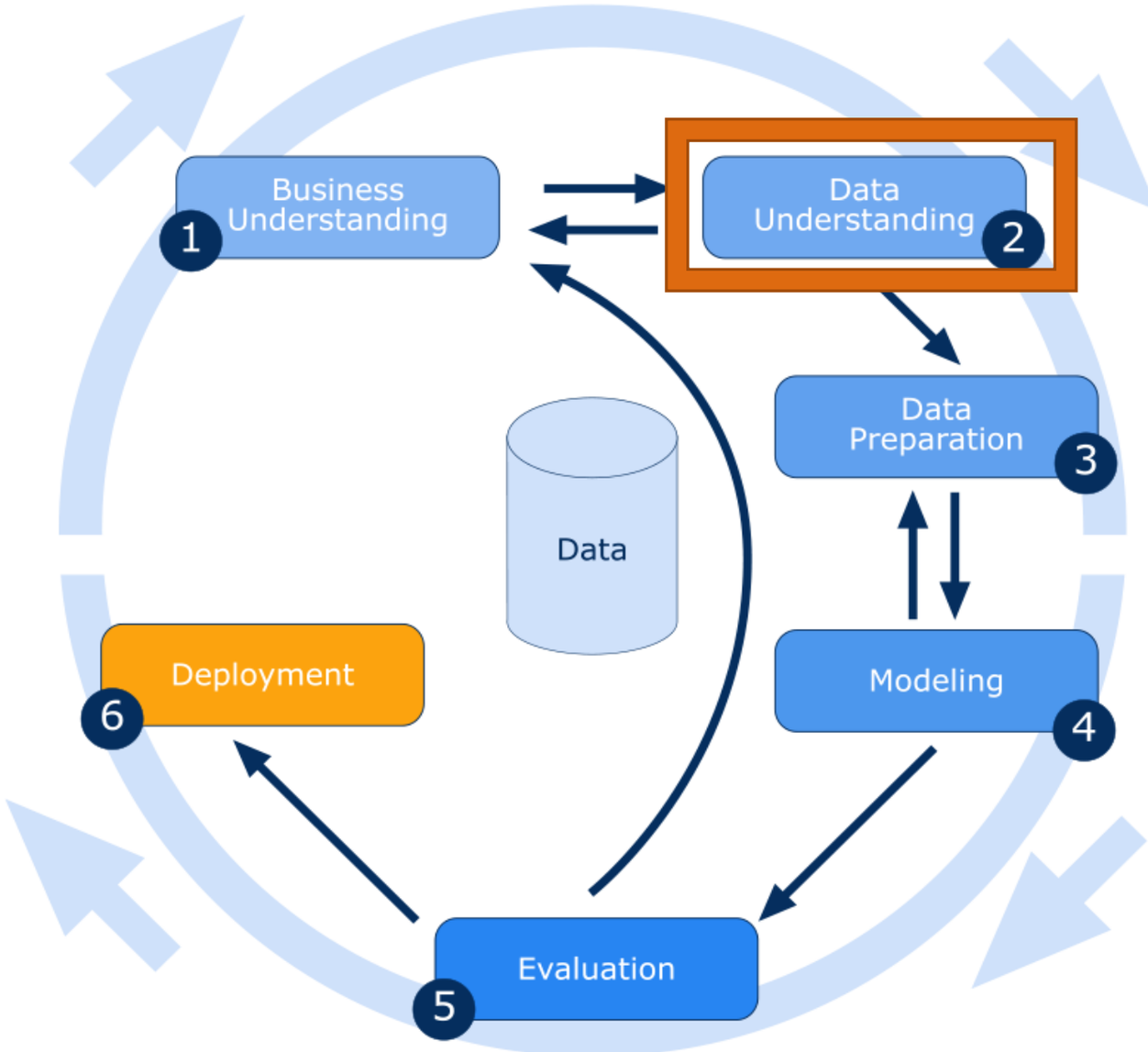
  (e.g. a certain level of forecasting accuracy)

Data Understanding

Take a closer look at the data, access and explore the data, match between the business problem and the data

This phase includes four tasks

- Collect initial data
- Describe data
- Explore data
- Verify data quality



Data Understanding Example

- Data Collection: gathers data from various sources, including:

  - Sales Data: Transaction records containing information on customer purchases
  - Customer Data: Information from loyalty program registrations
  - Product Data: Details about each fruit

- Data Description: examines the collected data to understand its characterise

  - Sales: verifies the integrity of the sales records, checks for consistency in data formats, and assesses the completeness of required fields.
  - Customer: examines the customer profiles for completeness and consistency
  - Product: reviews the product information for accuracy and ensures that essential attributes are present for each fruit.

- Data Exploration: explores the data to uncover insights and patterns

  - Sales trend: generates visualizations, to analyse sales trends over time and identify seasonal variations in fruit purchases.
  - Customer: uses bar charts and histograms to understand the distribution of customer age groups and locations.

- Data Quality Assessment: assesses the quality of the data and addresses any data quality issues: any missing values, outliers, incosistency

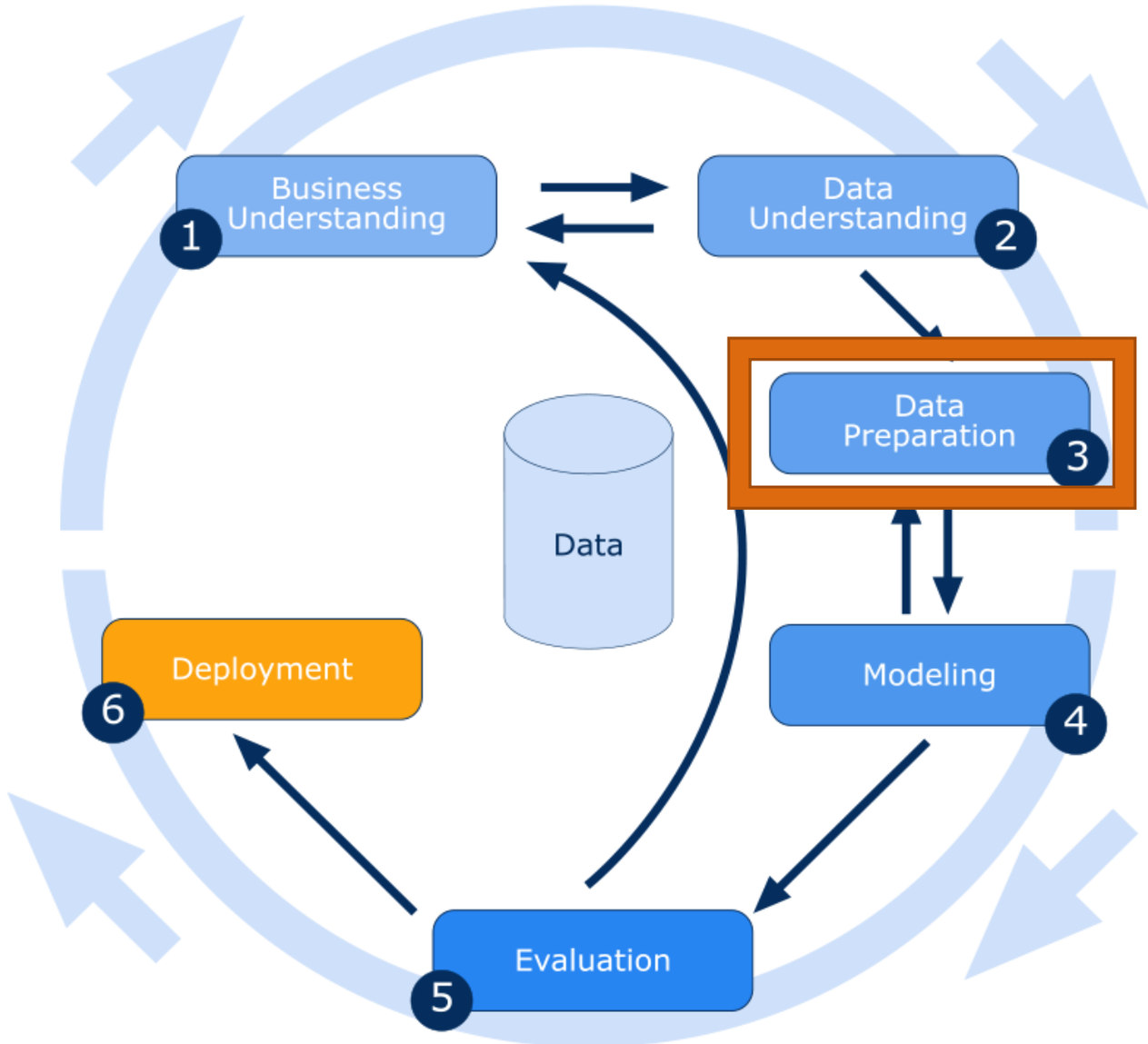
Data Preparation

"data munging": prepare the final data set(s) for modelling

- *a common rule of thumb – take 80% of the project time/effort*

This phase includes five tasks

- Data Selection
- Data Cleaning
- Data Construction
- Integrate data
- Format data



Data Preparation Example

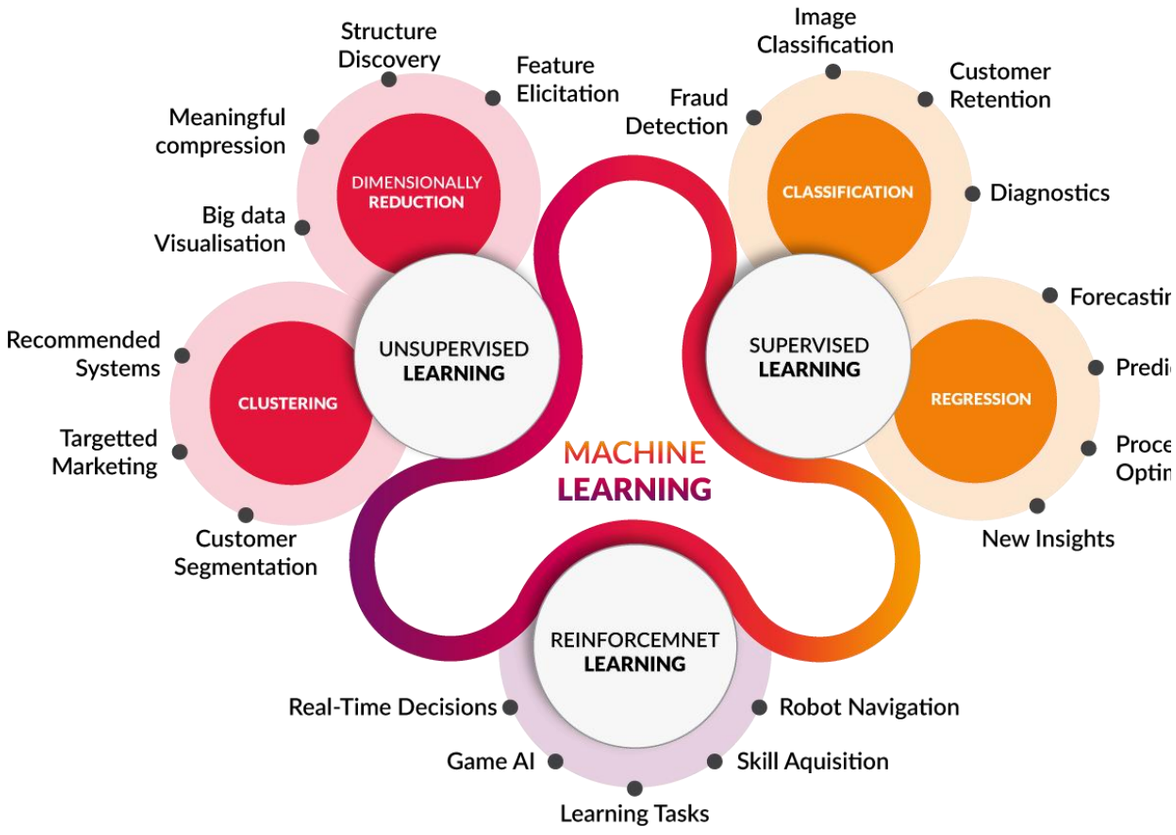
Based on the data exploration and quality assessment, prepares the data for subsequent data mining tasks:

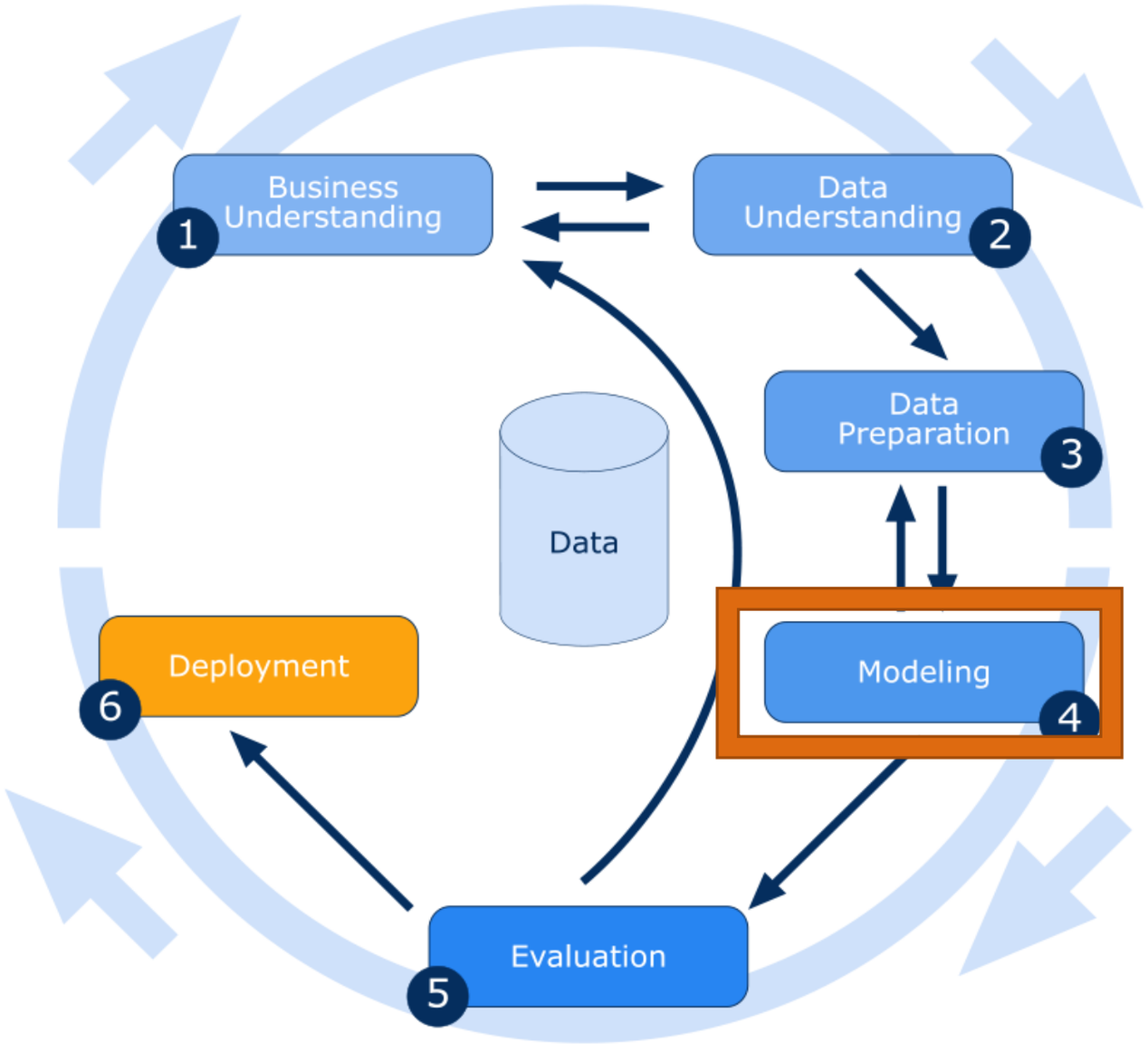
- Data Cleaning: removes duplicate records and handles missing values in customer profiles and product information.
- Data Integration: combines sales data, customer data, and product data using common identifiers to create a unified dataset for analysis.
- Data Transformation: Feature Engineering to creates new features (variables) that may provide more meaningful insights. If necessary, normalizes or scales numerical features to bring them to a similar range.
- Data Spliting: Dividing the Data into Training and Testing Sets

Model Building

Build and assess various models based on several different modeling techniques

- *widely regarded as data science's most exciting work but often the shortest in the process*





Model Building

- Select modelling technique: select the specific modelling technique, and record assumptions
- Generate test design: generate a procedure or mechanism to test the model's quality and validity (e.g. separate data into training and test)
- Build model: run the modelling tool on the prepared dataset to create one or more models
  - choose parameter settings, describe the resulting models
- Assess model: Interpret the models

  according to domain knowledge, data mining success criteria and desired test design

Model Building Example

- Demand Forecasting Model

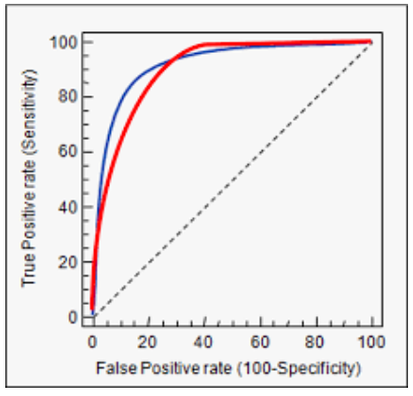
  - Time Series Forecasting: uses time series forecasting techniques, such as ARIMA (AutoRegressive Integrated Moving Average) or exponential smoothing methods, to predict future demand for specific fruits based on historical sales data

- Customer Segmentation Model

  - Clustering Algorithms: uses clustering algorithms, like k-means or hierarchical clustering, to segment customers based on their purchasing behaviour and preferences.

Access Model

- Evaluation of model: how well it performed on test data?

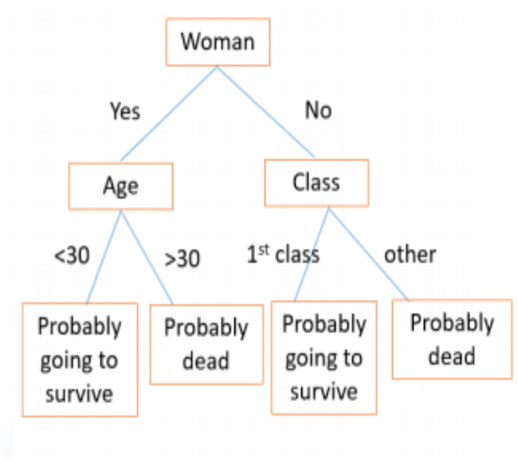


- Methods and criteria depend on model type:
  - e.g. confusion matrix with classification models, mean error rate with regression models

	Predicted: NO	Predicted: YES
Actual: NO	TN = ??	FP = ??
Actual: YES	FN = ??	TP = ??

$$\mathrm{M} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

- Interpretation of model: important or not, easy or hard depends on algorithm

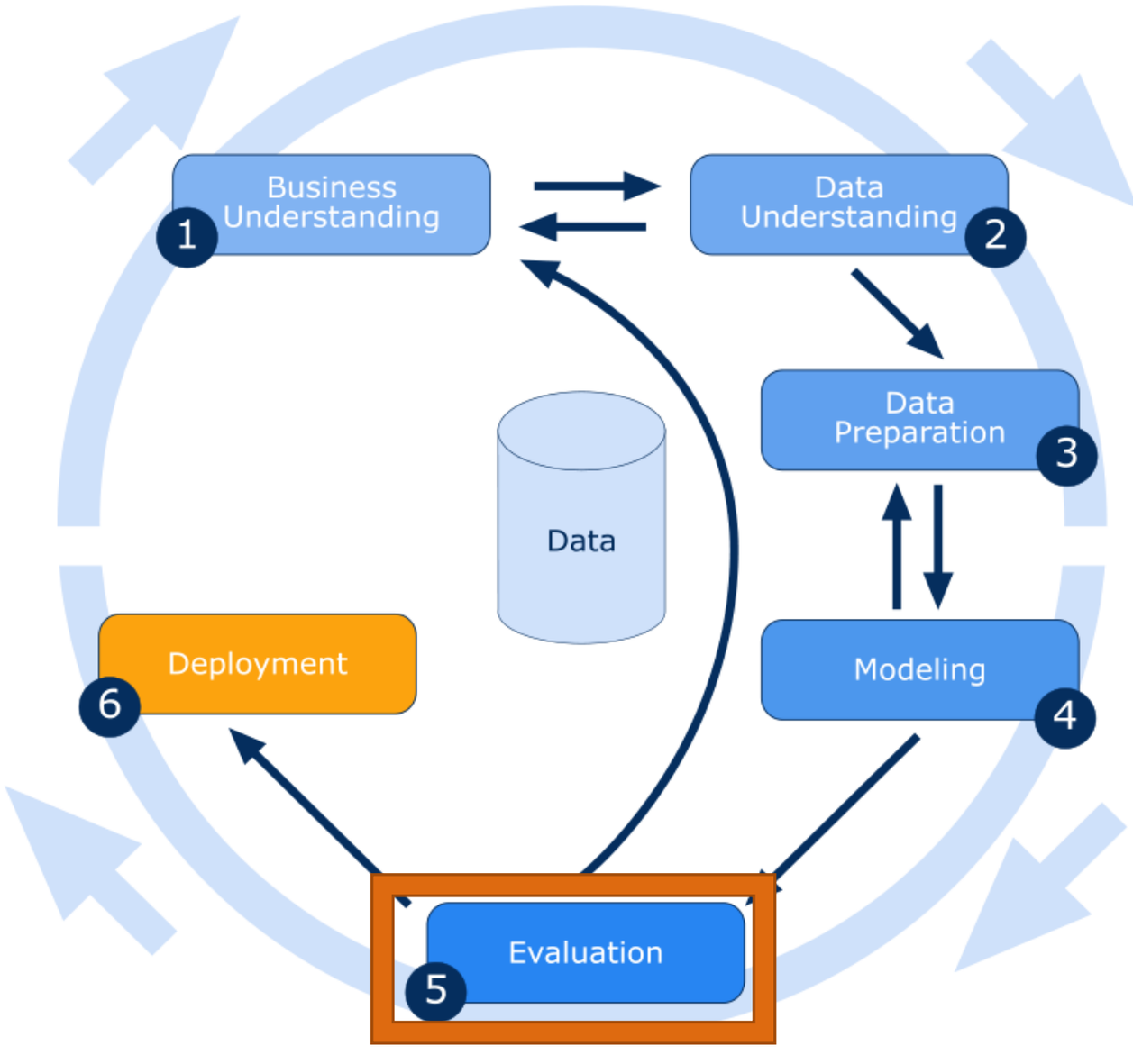


Model Evaluation

Evaluate and determine which model best meets the business and what to do next

This phase has three tasks:

- Evaluate results
- Review process
- Determine next steps

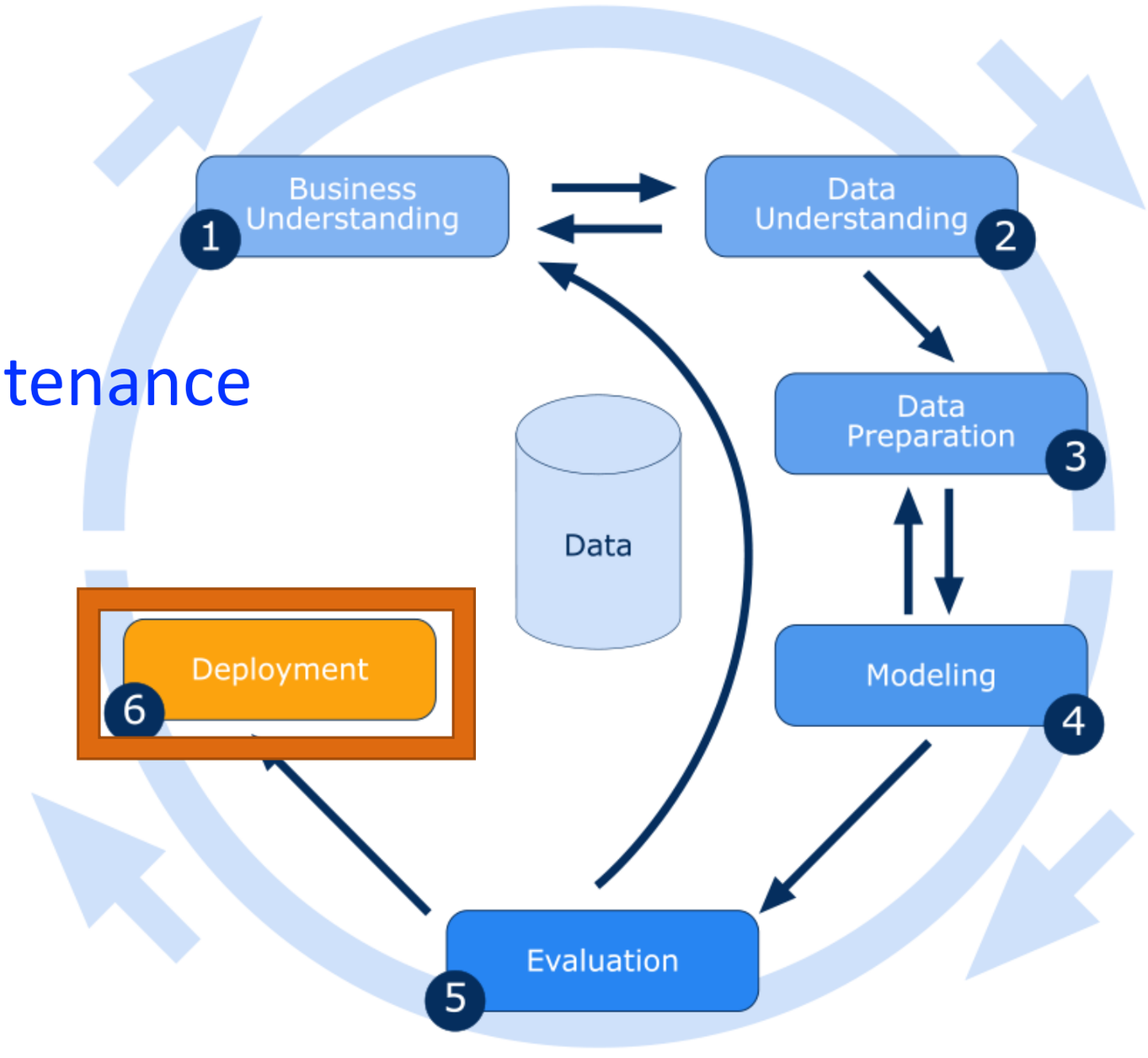


Deployment

The process of using new insights to make improvements, a formal integration of model, use the insights gained from data mining to make change

This phase has four tasks:

- Plan deployment
- Plan monitoring and maintenance
- Produce final report
- Review project



This table summarizes the main inputs to the deliverables. This does not mean that only the inputs listed should be considered—for example, the business objectives should be considered to all deliverables. However, the deliverables should address specific issues raised by their inputs.

Phase	Deliverable	Refers To	Closely Related To
Business Understanding	Background		
	Business Objectives	Background	Terminology
	Business Success Criteria	Business Objectives	
	Inventory of Resources		
	Requirements, Assumptions & Constraints	Business Objectives	
	Risks & Contingencies	Business Objectives; Business Success Criteria	
	Terminology	Background	Business Objectives
	Costs & Benefits	Business Objectives	Project Plan
	Data Mining Goals	Business Objectives; Requirements, Assumptions & Constraints	
	Data Mining Success Criteria	Business Success Criteria; Requirements; Assumptions & Constraints; Data Mining Goals	
	Project Plan	Business Objectives; Inventory of Resources; Requirements; Assumptions & Constraints; Risks & Contingencies	Costs & Benefits
Data Understanding	Initial Data Collection Report	Business Goals; Inventory of Resources; Data Mining Goals	
	Data Description Report	Business Goals; Initial Data Collection Report	Data Quality Report
	Data Quality Report	Business Goals; Initial Data Collection Report	Data Description Report
	Exploratory Analysis Report	Business Goals; Initial Data Collection Report	
Data Preparation	Dataset & Dataset Description	Business Goals; Data Mining Goals & Data Description Report; Data Quality Report; Exploratory Analysis Report	
Modeling	Test Design	Data Mining Goals; Data Mining Success Criteria	
	Models	Data Mining Goals	Parameter Settings
	Parameter Settings	Data Mining Goals	Models
	Model Description	Models; Parameter Settings; Test Design	
	Assessment	Data Mining Success Criteria; Test Design; Models	
Evaluation	Assessment w.r.t. Business Success Criteria	Business Success Criteria; Terminology	
	Review of Process	Business Goals; Assessment w.r.t. Business Success Criteria	
	Next Steps	Project Plan; Assessment w.r.t. Business Success Criteria	
Deployment	Deployment Plan	Business Goals; Requirements, Assumptions & Constraints	Maintenance Plan
	Maintenance Plan	Business Goals; Requirements, Assumptions & Constraints	Deployment Plan
	Final Report & Presentation	Business Goals; Terminology; Assessment w.r.t. Business Success Criteria	
	Experience Documentation	Project Plan; Review of Process	