AIML231/DATA302 — Week 04

Machine Learning Pipeline

Dr Bach Hoai Nguyen

School of Engineering and Computer Science

Victoria University of Wellington

Bach.Nguyen@ecs.vuw.ac.nz

Office Hour:1-2pm, Friday, Week 4-Week 7

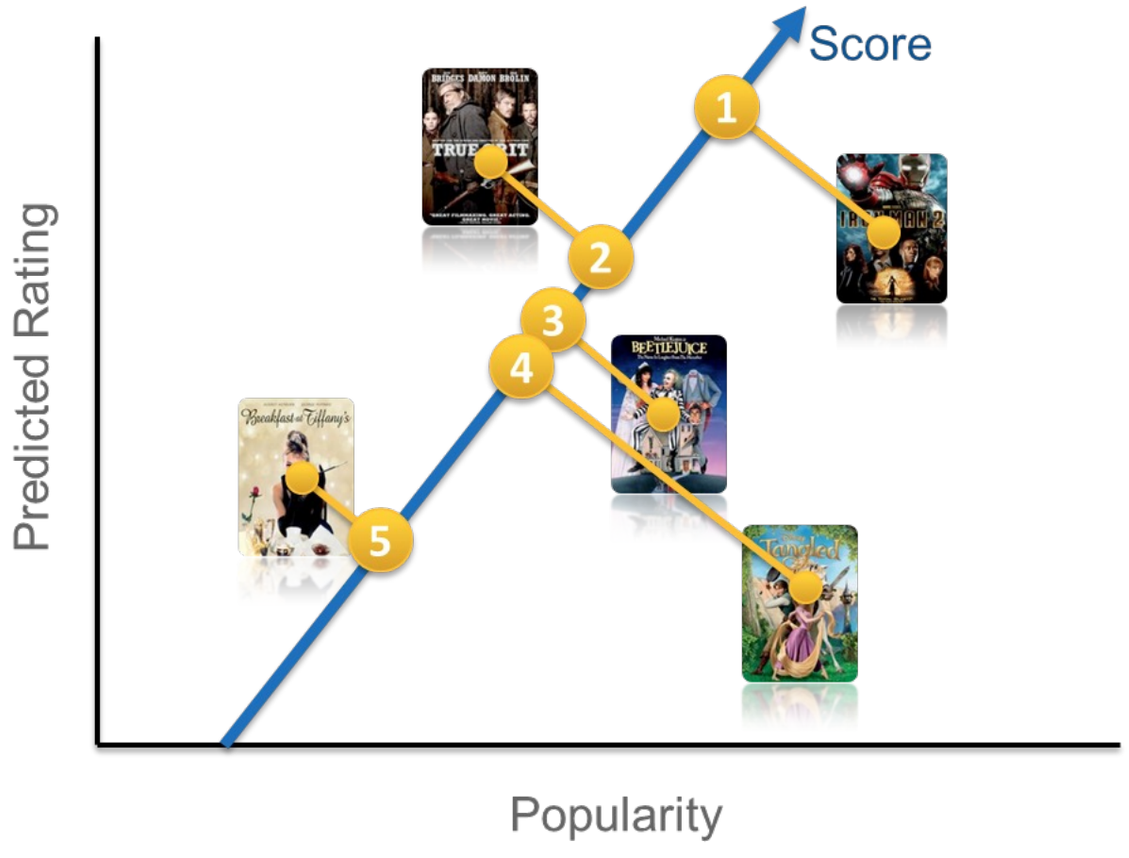
Room: CO364

Lecture Overview

- Machine Learning Applications
- Data Mining and Machine Learning
- The Six Phases in CRISP-DM

Machine Learning in Netflix

- Netflix: over 190 countries, over 260 million subscribers by 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



Machine Learning in Retail Industry

- Pricing strategy/optimisation
- Find groups of items that tend to occur together in transactions
- Fortune Business Insight's 2020: $12 billion in 2023 to $31.18 billion by 2028

Store Layout

Recommendation Engines

Targeted Marketing

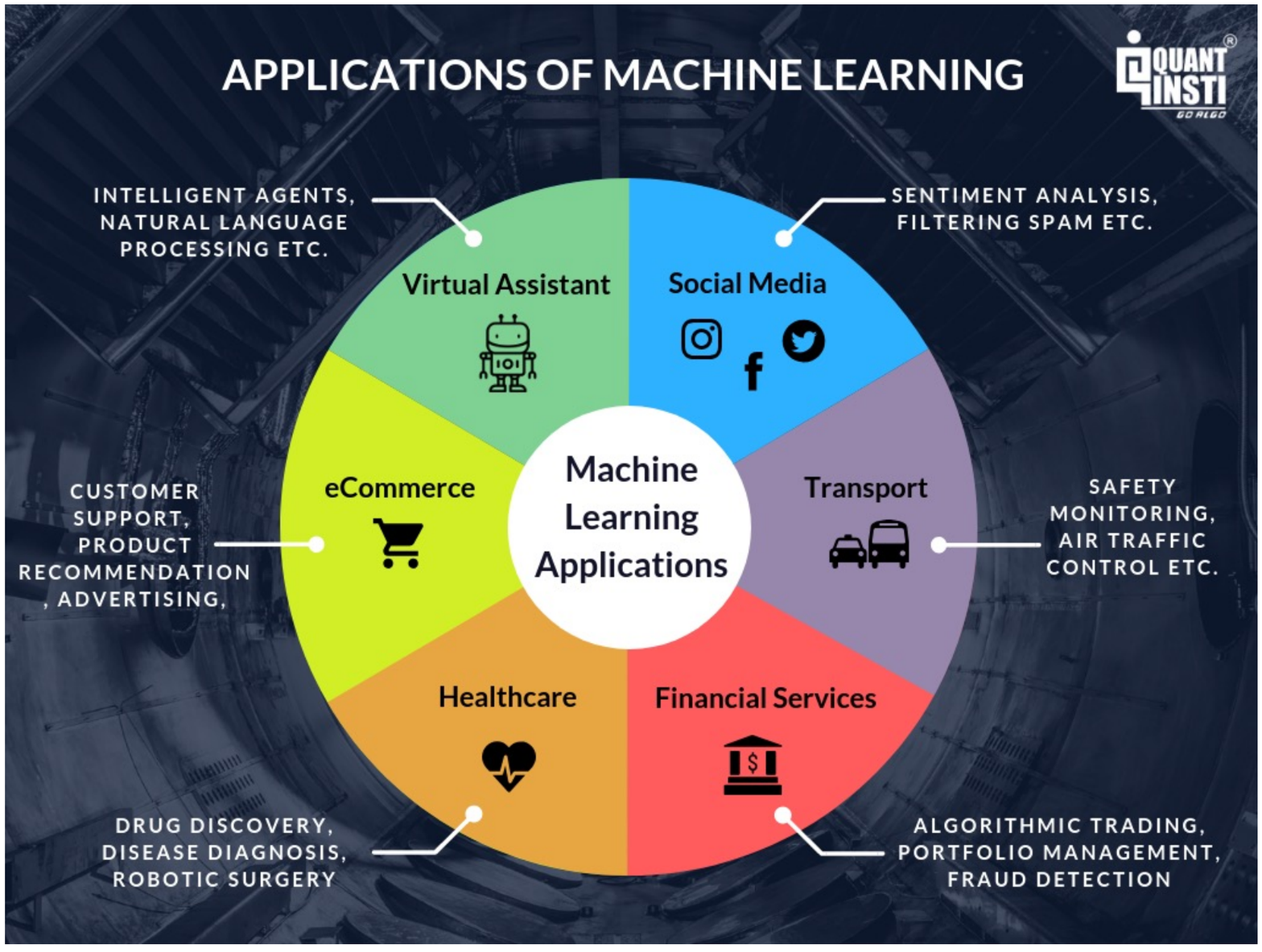
Up Sell & Cross Sell

Catalogue Design

Customer Experience

Typical Machine Learning Applications

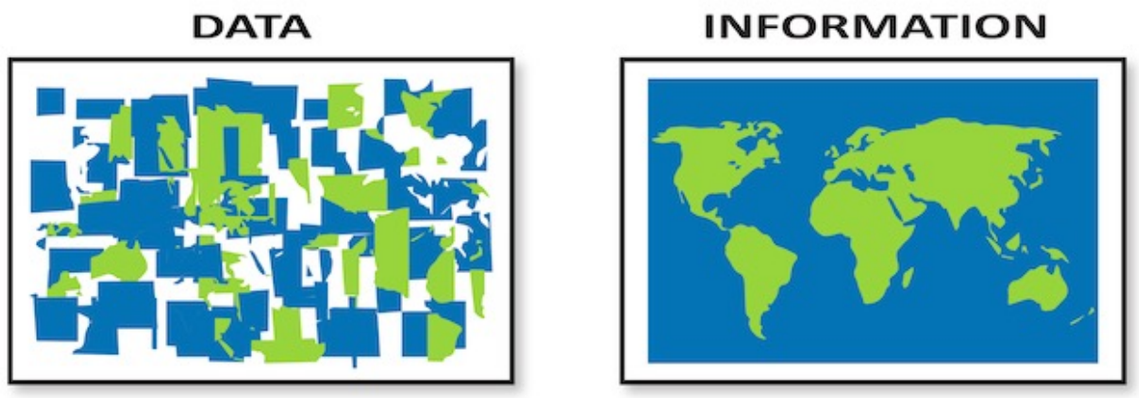
- Address many complex business problems/ opportunities
- Very successful and helpful in many areas



https://www.linkedin.com/pulse/machine-learning-its-applications-vidhi-kapoor/

From Data to Information

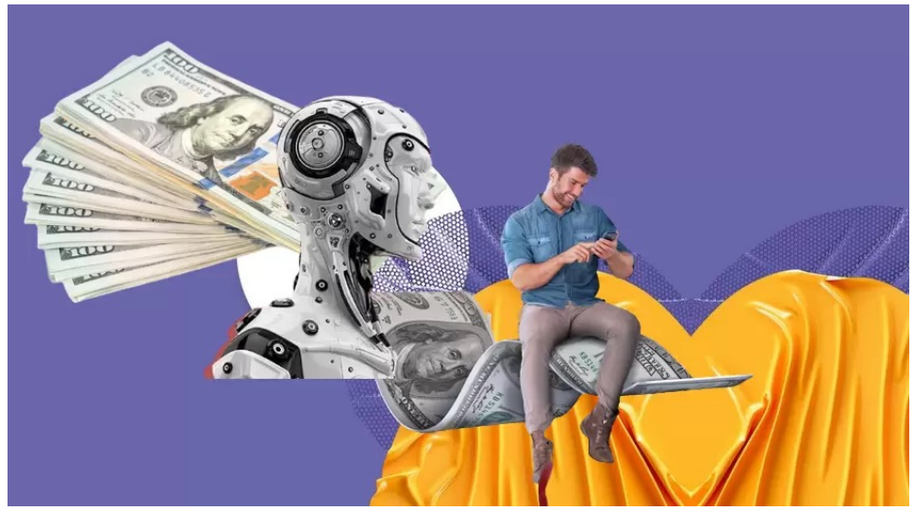
- Society produces huge amounts of data
  - Sources: business, science, medicine, economics, geography, environment, sports, …
- Data is potentially valuable resource but raw data is useless
- Need techniques to automatically extract information/pattern from data
  - Data: recorded facts
  - Information: patterns underlying the data



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

How Machine Learning Works

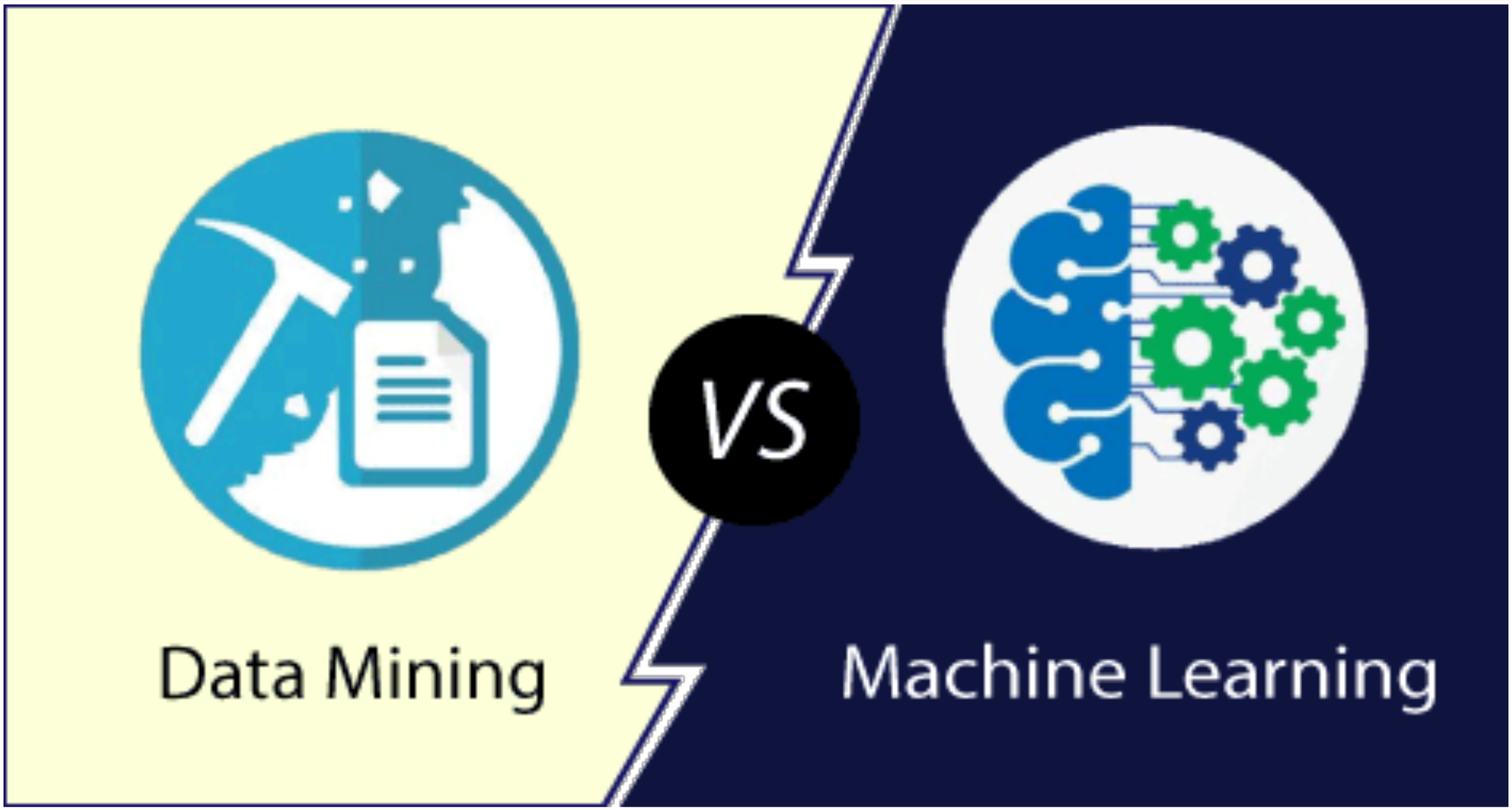
- Need a standardised process
  - to systematically conduct machine learning/data mining - project planning and management, encourage best practices and help to obtain better results.
  - as a framework for recording experience
  - "comfort factor" for new adopters - demonstrates maturity of data mining and reduces dependency on "stars"



- Several standardised processes have been developed

- Most popular one: Cross-industry standard process for Data Mining (CRISP-DM)

Thinking

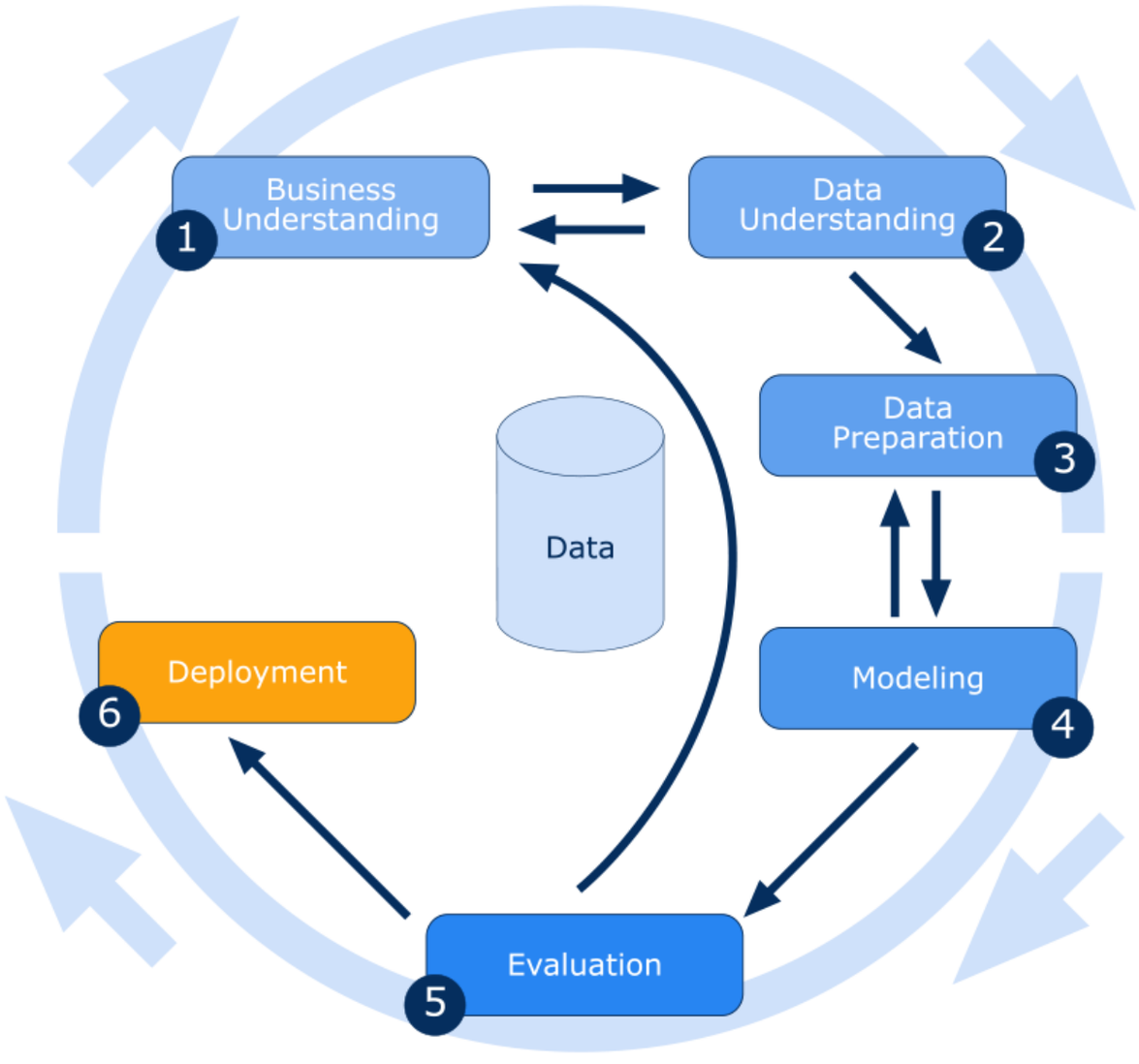


What differences and relationships exist between data mining and machine learning?

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

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

- developed by big players in data analysis in 1996
- a nonproprietary standard methodology

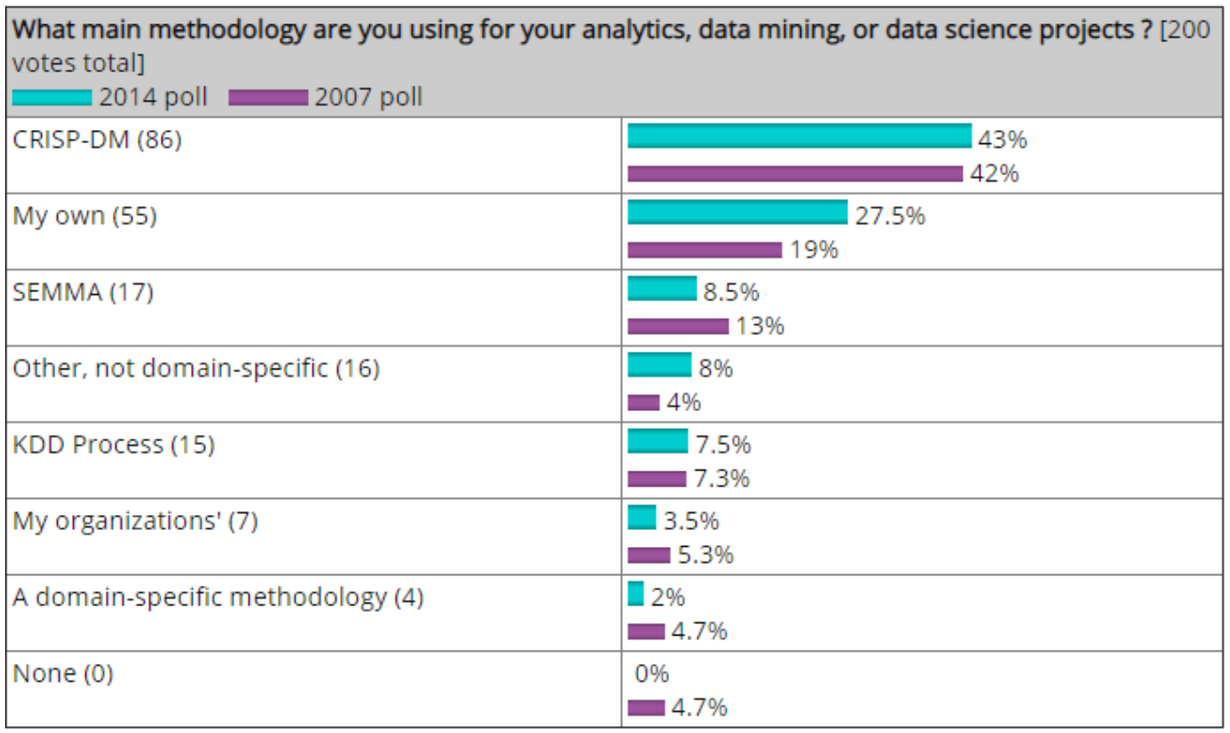


- consists of six phases with arrows indicating dependencies
- sequence of the phases is not strict
- flexible and can be customized easily
- can revisit phases
- CRISP-ML (Q) in 2021: very early stage

  Studer, Stefan, et al. "Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology." *Machine learning and knowledge extraction* 3.2 (2021): 392-413.

Why Successful?

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

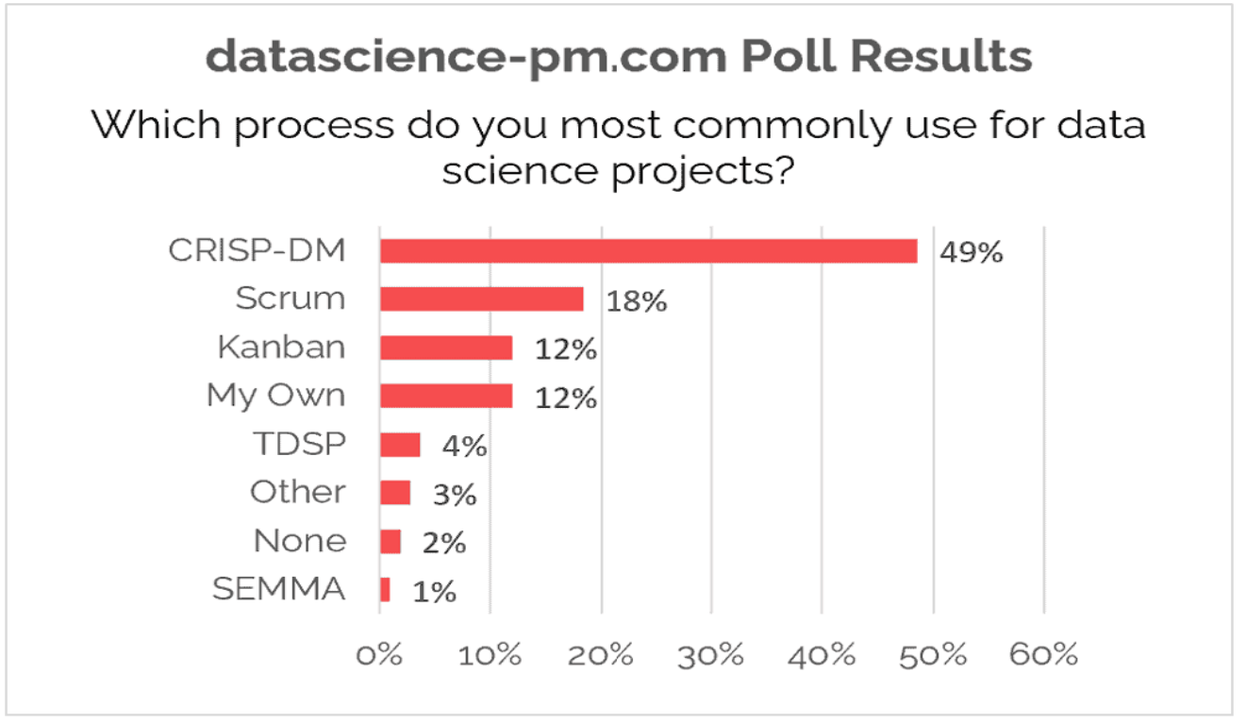


What main methodology are you using for your analytics, data mining, or data science projects ? [200 votes total]

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%

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/

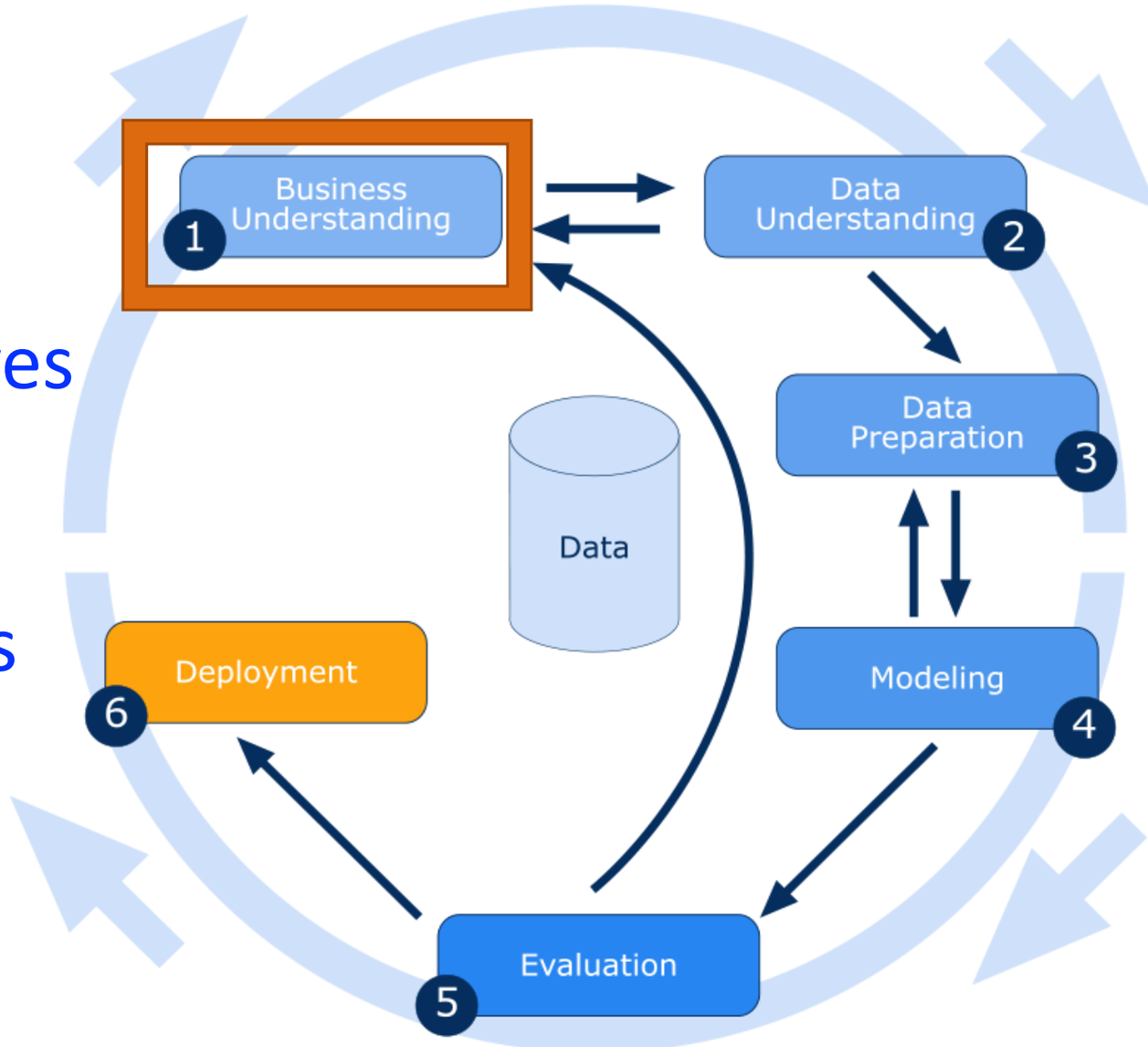
Business Understanding

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



Four Tasks in Business Understanding

- determine business objectives
  - to understand what to accomplish from a business perspective
  - often trade-off between several competing objectives with constraints
- assess situation
  - assess the resources availability, project requirements, assess risks and contingencies, and conduct a cost-benefit analysis
- determine data mining/machine learning goals
  - define successful outcomes in technical terms
- produce project plan
  - Select technologies and tools
  - define detailed plans for each project phase

Determine Business Objectives and Data Mining Goals

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

***Primary goal**: keep current borrowers by predicting when they are prone to move to a competitor - reducing borrowers churn*

***Related questions**: will lower interest (mortgage) rates reduce the number of high-value customers who leave?*

- Covert business objectives to the definition of data mining problem and goals

***Business Goal**: increase the renewal rate (reduce churn) of borrowers whose contracts were expiring*

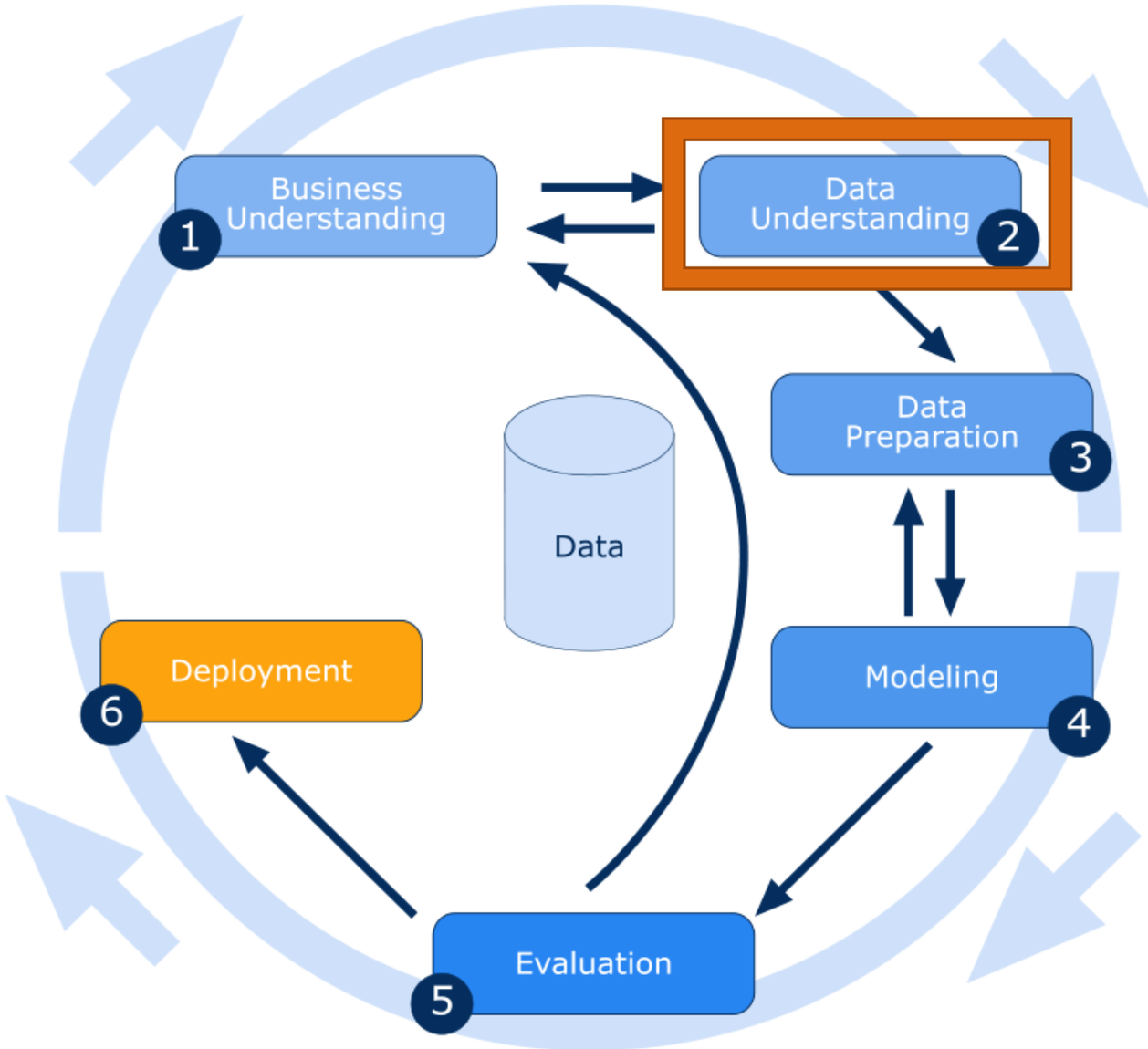
***Data Mining Goal**: Predict which customers were most likely to leave and/or how likely was a particular customer to accept an offer of a new plan*

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

- Collect initial data：acquire the data listed in the project resources, may need data loading, integrate multiple data sources
- Describe data: examine the “gross” or “surface” properties of the data (e.g. data format, quantity)
- Explore data: dig deeper into the data, query, visualize, and identify relationships among the data
- Verify data quality: Examine the quality of the data, addressing questions (e.g. is the data complete? Correct? Missing?)

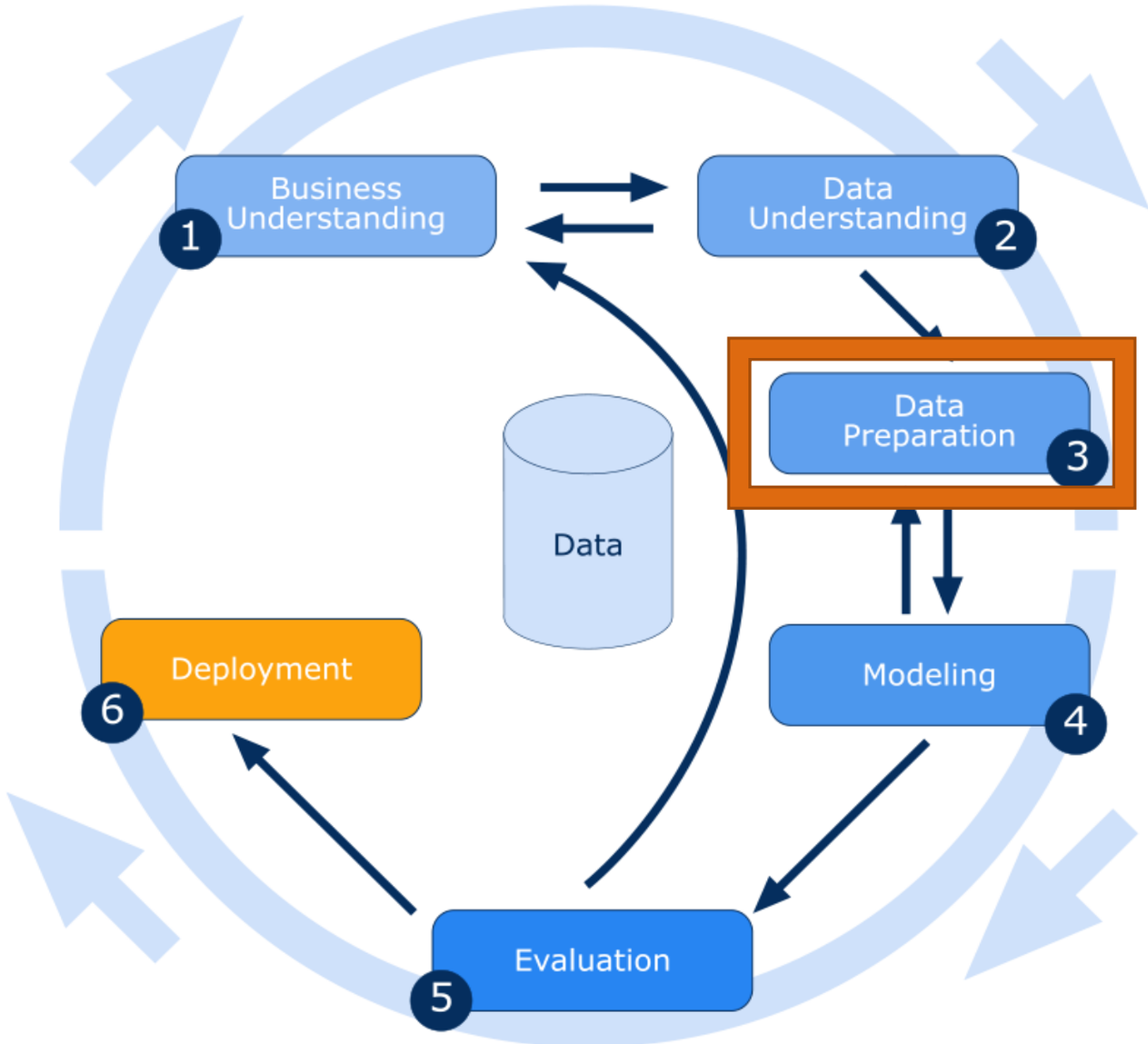
Data Preparation

"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



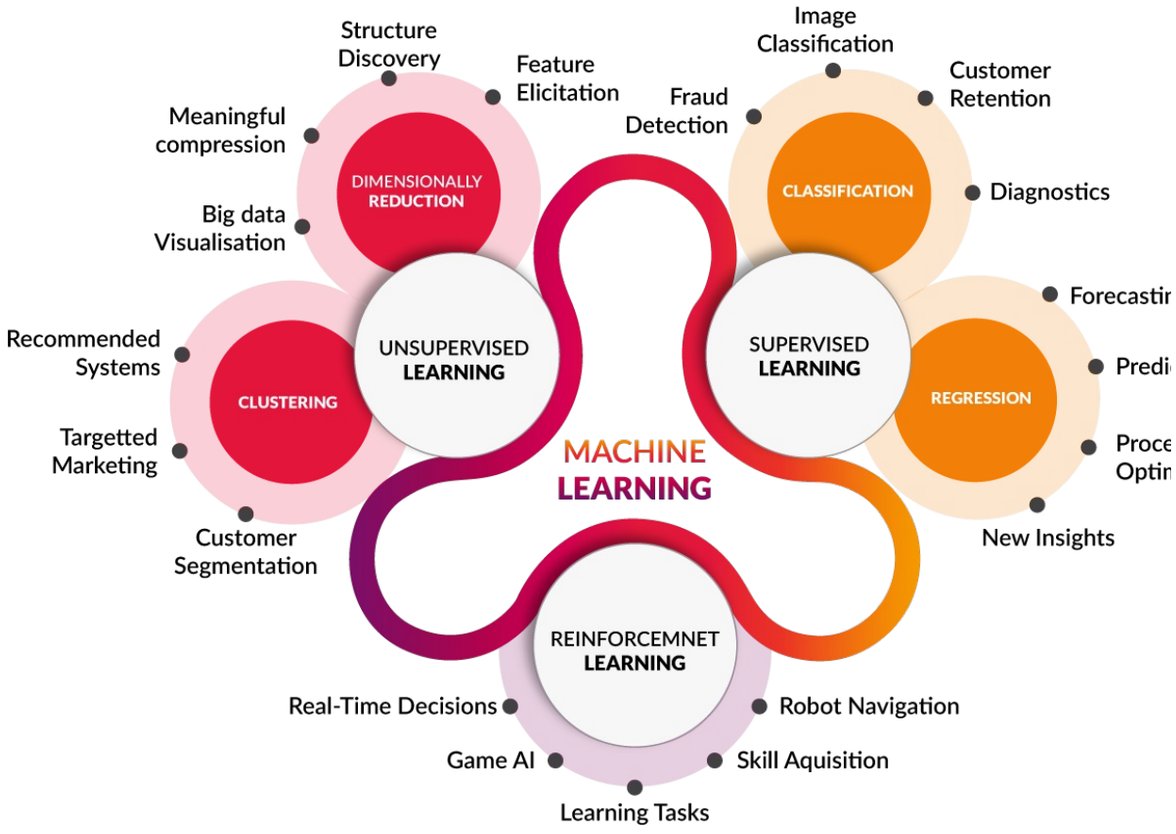
Data Preparation

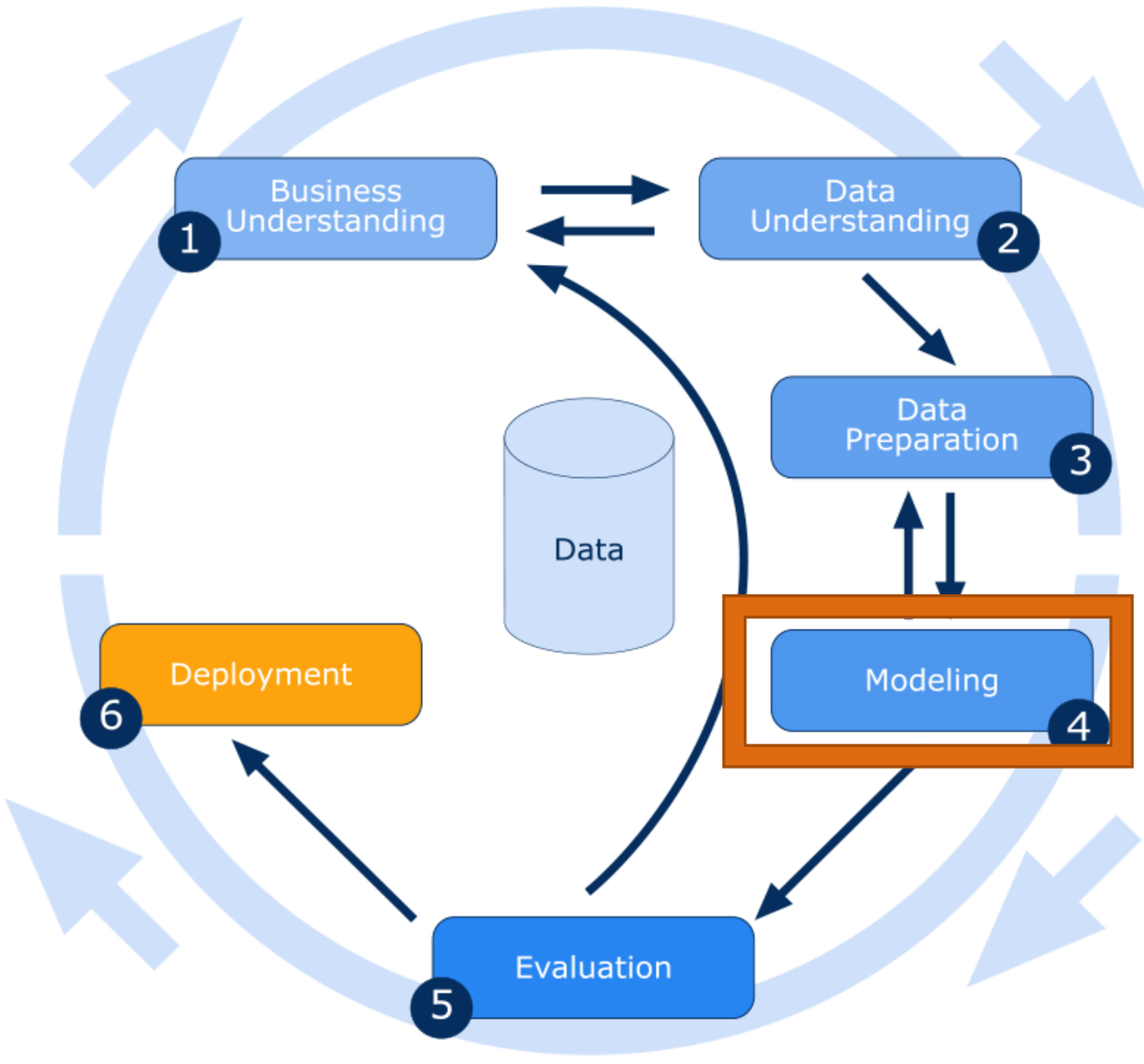
- Data Selection: determine data sets to be used, selection of features, selection of records/rows
- Data Cleaning: the lengthiest task, to correct, impute or remove erroneous values, missing values
- Data Construction: constructive data preparation
  - feature construction, instance generation, feature transformation
- Integrate data: create new records or values combining from multiple data sources
  - merge information from different sources, aggregations
- Format data: re-format data, convert to format convenient for modelling

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

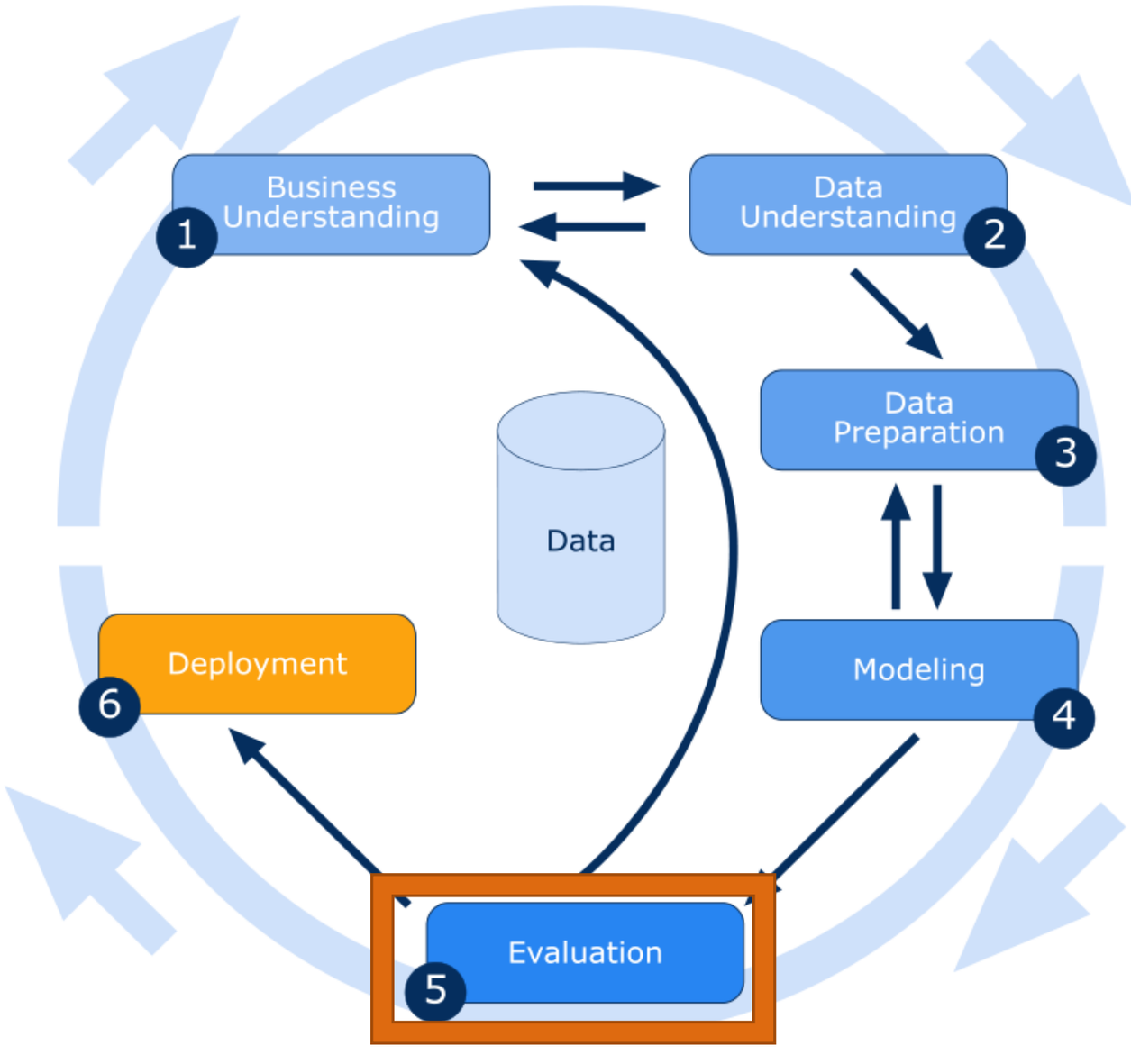
- Select modelling technique: select the specific modelling technique and record assumptions
- Generate test design: generate a procedure or mechanism to test the model's quality and validity (e.g. separate data into training and test)
- Build model: run the modelling tool on the prepared dataset to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

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



Model Evaluation

- Evaluate results: assesses model meets business objectives
- Review process: do a more thorough review of the data mining engagement , also cover quality assurance issues
- Determine next steps: decide how to proceed depending on the results of the assessment and the process review

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