



DM,
Autumn, 2026



Data Mining

Faculty of DS & AI
Autumn semester, 2026

Trong-Nghia Nguyen



Content

- **About**
- What is data mining?
- Data organization
- CRISP-DM process

About

This course

Course site

(2 midterm: 1 paper/ 1 project)

Textbook

Data Mining for the Masses

Dr. Matthew North

Grading Component	Weight	Assessment Details
Attendance & Participation	10%	Random roll calls, active class participation (speaking up, doing board exercises).
Midterm Exam	40%	Midterm test / assignment / essay.
Final Exam	50%	Final examination / capstone project.
TOTAL	100%	

Rule Category	Detailed Policy Description
Base Attendance Score	Students with full attendance but no active participation will receive a maximum base score of 8.0/10 for this component.
Participation Bonus	Each instance of active participation (e.g., answering questions, doing exercises on the board) adds +0.25 points to the Attendance score.
Absence Penalty & Ban	Random roll calls will be conducted. If a student is absent for more than 4 sessions, their Attendance score will be 0 , and they will NOT be eligible to take the Final Exam .
Excess Point Transfer	The maximum recorded Attendance score is 10.0. If a student's total attendance points exceed 10.0, the excess points will be divided by 2 (50%) and added directly to their Midterm Exam score .
Calculation Example	Example: - A student has full attendance (8.0 pts) and participates 10 times ($10 \times 0.25 = 2.5$ pts). - Total raw attendance = 10.5 pts. - Final Attendance Score = 10.0 pts. - The excess 0.5 points are halved (0.25) and added to the Midterm score.

Content

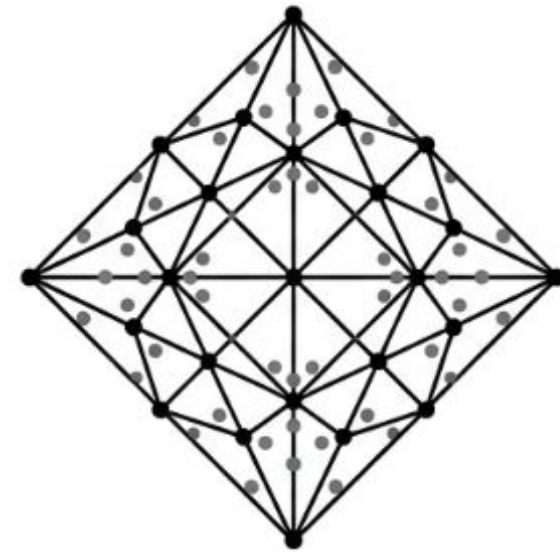
- About
- **What is data mining?**
- Data organization
- CRISP-DM process

What is data mining?

Definition



Data generation happens silently and transparently. Every loyalty card scan, credit card transaction, and web interaction creates massive datasets detailing user preferences, habits, and behaviors.



Definition: The systematic process of discovering hidden patterns, relationships, and actionable knowledge deep within large datasets. The ultimate goal is to support accurate and effective business decision-making.

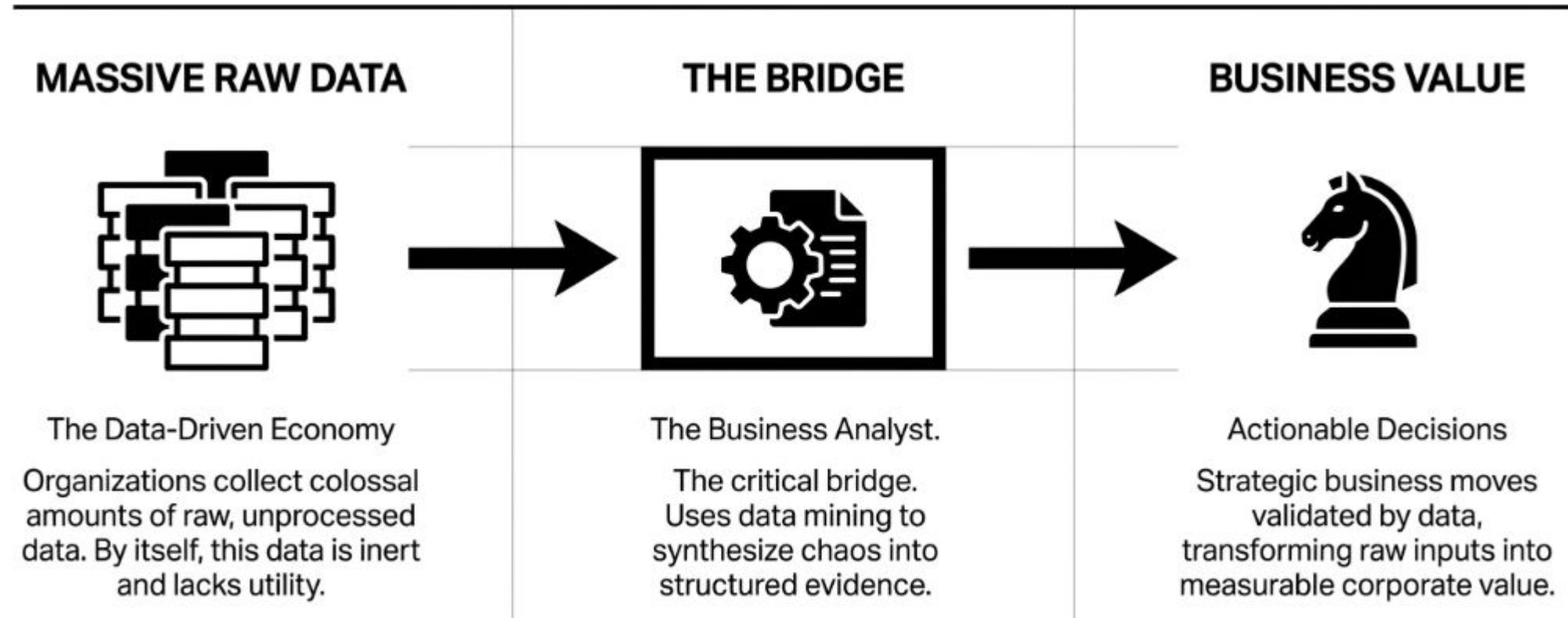
What is data mining?

Traditional Reporting vs. Data Mining

Traditional Reporting	Data Mining
Analytics Category Descriptive Analytics.	Predictive Category Predictive & Prescriptive Analytics.
Core Question What happened?	Core Question What is likely to happen? and Why?
Business Example What were total sales last quarter? How many customers churned?	Business Example Which specific customers will churn next month, and what factors are driving their decision?

What is data mining?

The Importance of Data Mining for Business Analytics



What is data mining?

Four major problem groups in data mining



DESCRIBE

Goal: Summarize and visualize data.

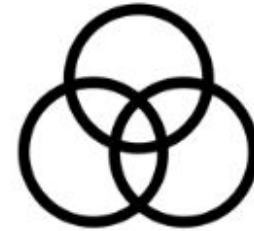
Key Question: What is the data saying?



RELATE

Goal: Identify correlations and associations.

Key Question: What usually goes together?



GROUP

Goal: Cluster objects with shared, natural traits.

Key Question: How many groups can these objects be divided into?



PREDICT

Goal: Utilize classification and regression.

Key Question: What will happen next?

What is data mining?

Four major problem groups in data mining

Industry/Company	Business Problem	Data Mining Solution	Core Problem Type
Retail & E-commerce (Shopee/Amazon)	Suggesting Products you may like or optimizing physical store locations via demographics (age, zip code)	Association Rules	Relate
Entertainment (Netflix/YouTube)	Recommending video content tailored to individual viewing habits	Clustering & Prediction	Group & Predict
Finance & Banking	Detecting and preventing credit card fraud in real-time	Classification Models	Predict
Corporate Management	Identifying at-risk clients to execute targeted retention strategies	Logistic Regression (Churn Rate)	Predict

What is data mining?

Code Example

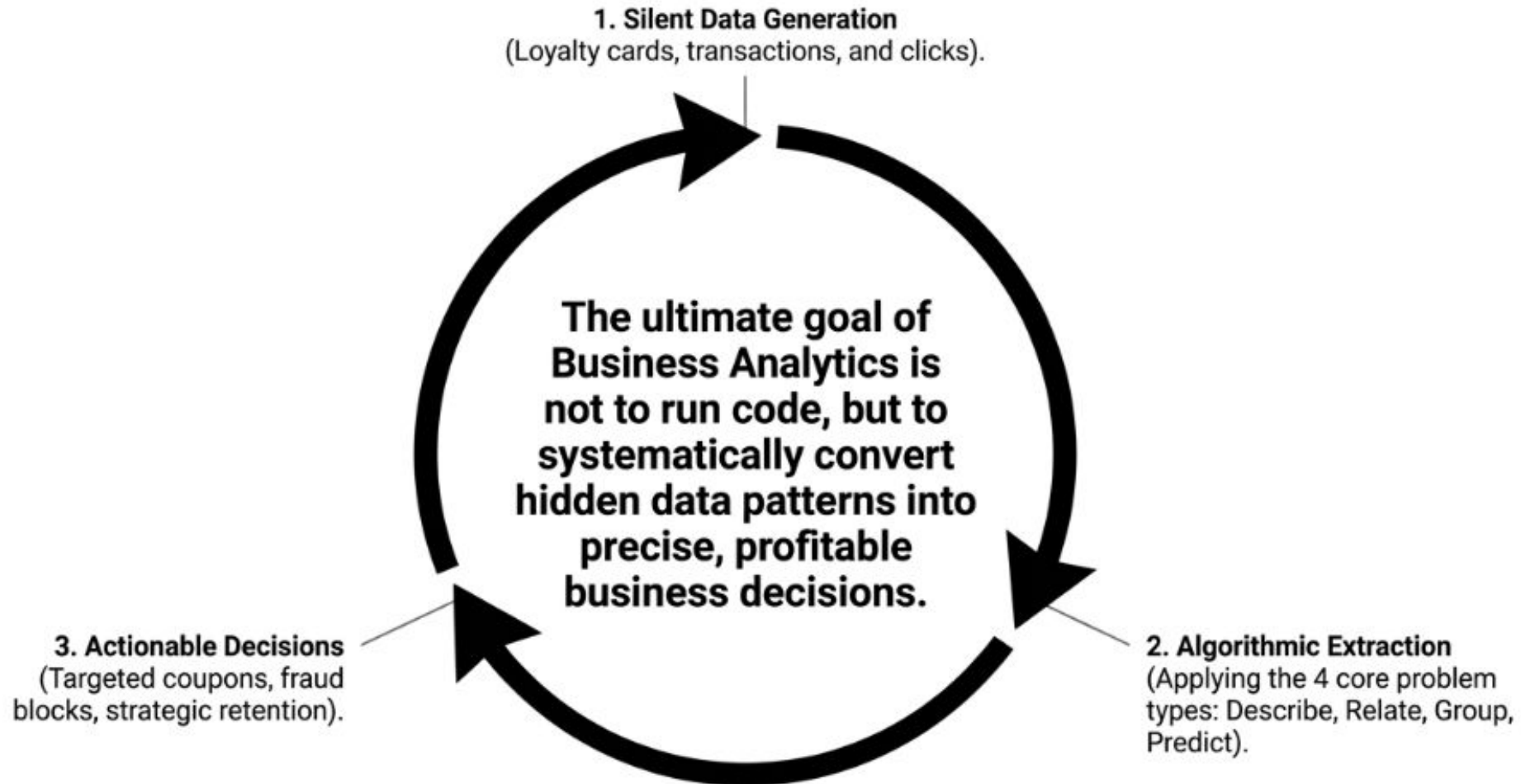
```
import pandas as pd ←  
from sklearn.cluster import KMeans  
  
# Load customer spending data  
data = pd.read_csv('customers.csv')  
  
# Initialize the grouping algorithm ←  
model = KMeans(n_clusters=3)  
  
# Extract patterns and assign groups  
data['Customer_Segment'] = model.fit_predict(data) ←
```

Tools like Python provide the mathematical infrastructure. The focus is on analytical thinking, not rote coding.

The Method: Clustering (Group). We ask the algorithm to find 3 natural segments based entirely on raw spending habits.

The Output: The algorithm independently discovers hidden rules, assigning each customer to a behavioral segment without human bias.

What is data mining?

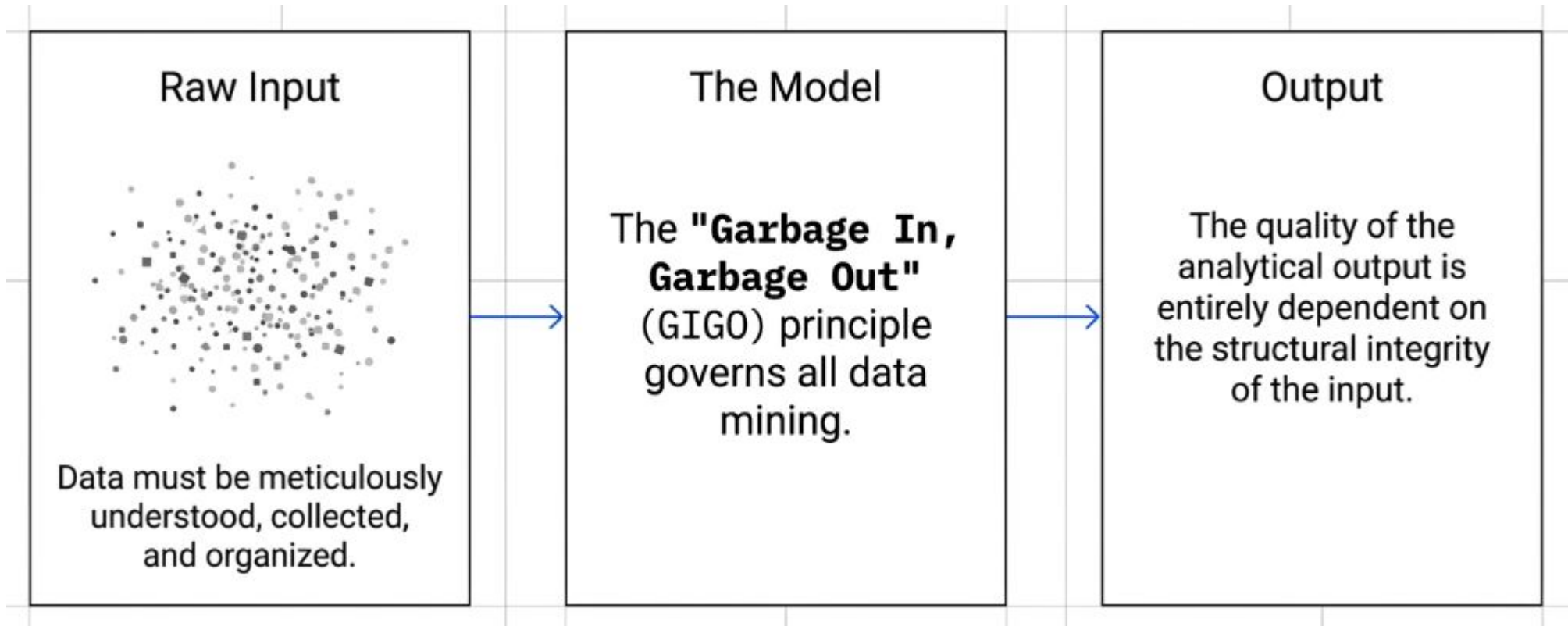


Content

- About
- What is data mining?
- **Data organization**
- CRISP-DM process

Data organization

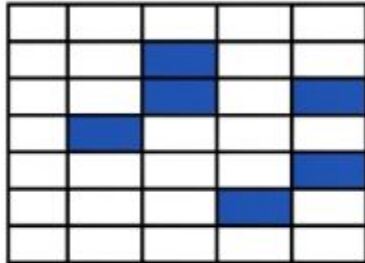
The prerequisite of data mining



Data organization

The prerequisite of data mining

Structured Data



- Neatly organized in 2D tables.
- Stored in Excel, Relational Databases, Data Warehouses, or Data Marts.
- Forms the foundation of core business analytics techniques.

Unstructured Data



- Free-form text, images, and audio.
- Requires complex preprocessing before modeling (e.g., Text Mining).

Data organization

Data table structure

Column
Attribute, feature, or variable.
A specific characteristic of the observation.

Row
Observation, instance, case, record / tuple, or example (RapidMiner terminology). Represents one single entity.

PassengerId	Age	Gender	Survived
1	22	Male	0
2	38	Female	1
3	26	Female	1

Target / Label
The specific variable the data mining model seeks to predict or explain (e.g., passenger survival, or customer churn).

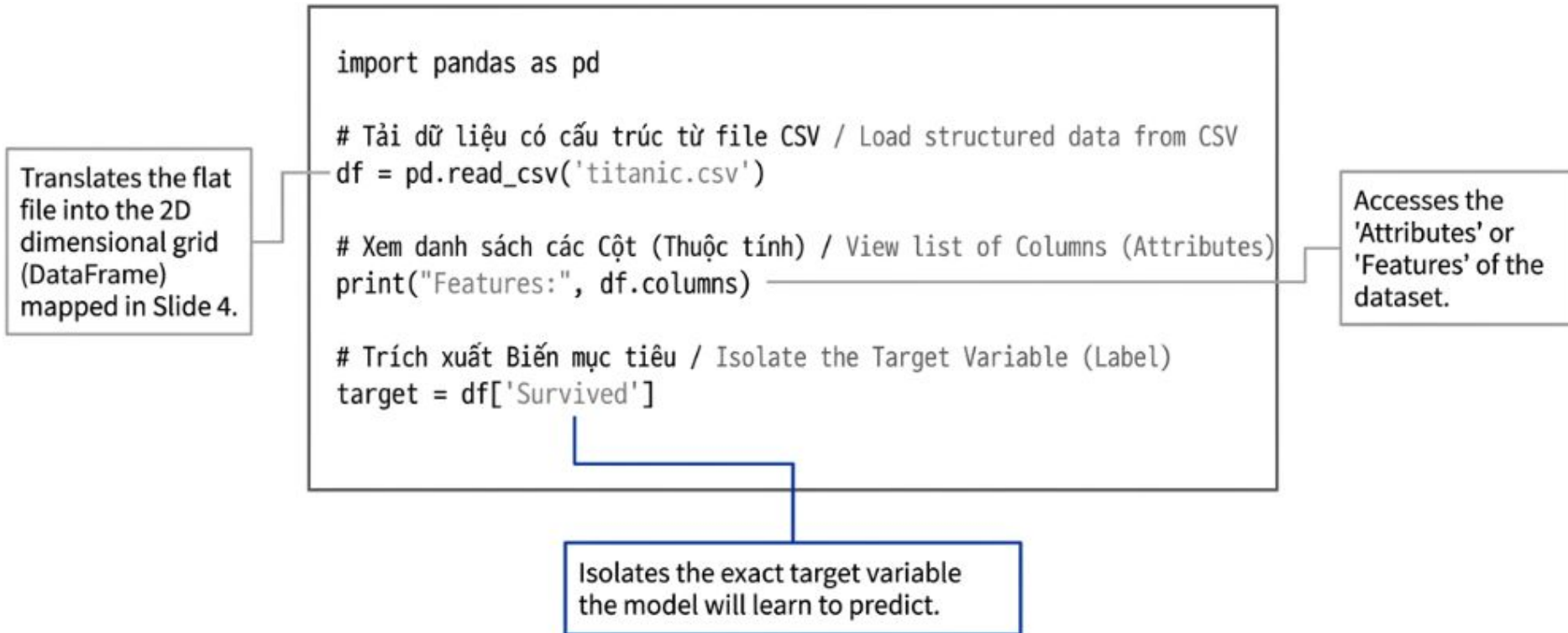
Data organization

Variable type

Variable Type	Statistical Nature	Real-World Examples	Machine Learning Implication
Numeric (Định lượng)	Contains numerical values capable of mathematical calculation.	Age, income, total order volume.	Algorithms scale and calculate distances between these numbers.
Categorical / Nominal (Phân loại)	Labels or group names with no inherent mathematical meaning or order.	Gender, city of residence, membership tier.	Must be encoded into numbers without implying one is "greater" than another.
Ordinal (Thứ bậc)	Categorical variables that possess a clear, inherent order or ranking.	Satisfaction level (Low, Medium, High).	Algorithms can leverage the sequential relationship between categories.
Datetime (Ngày tháng)	Stores specific timestamps or dates.	Purchase date, system login time.	Often parsed into smaller components (day of week, month) to find seasonal patterns.

Data organization

Code Example



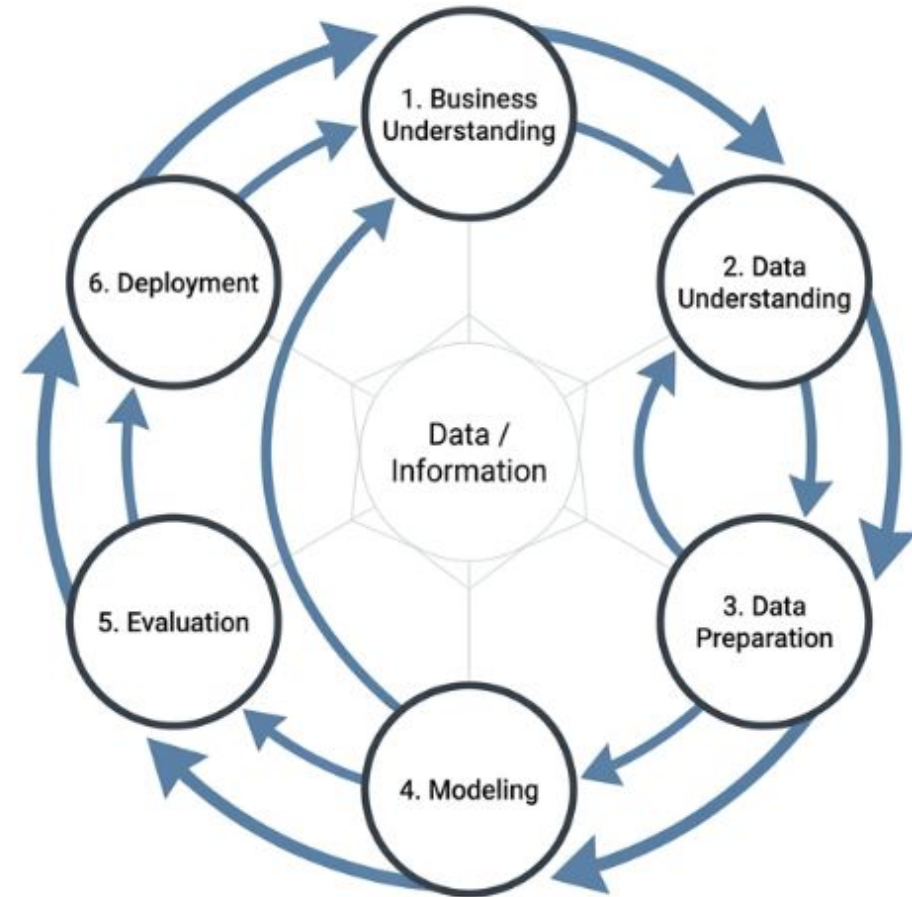
Content

- About
- What is data mining?
- Data organization
- **CRISP-DM process**

CRISP-DM process

Definition

- CRISP-DM = Cross-Industry Standard Process for Data Mining
- A standard, industry-independent process developed in the late 1990s and the most widely used in practice for organizing a data mining project.
- It consists of 6 stages and is cyclical-not a linear, one-way process.

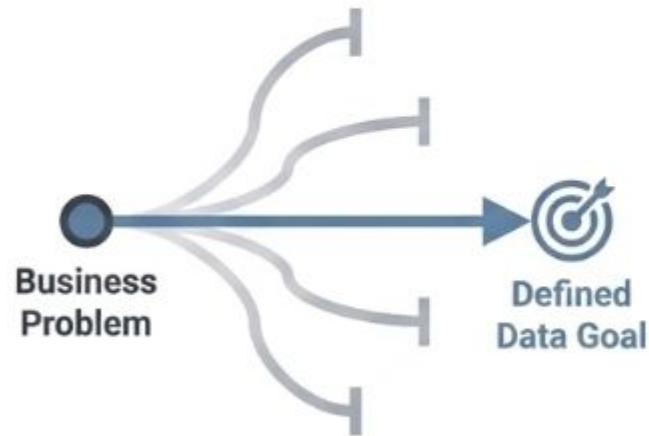


CRISP-DM process

Definition

1. Business Understanding

Translating business problems into data problems (e.g., "Reduce churn" becomes "Predict churn in 3 months").



2. Data Understanding

Initial collection, preliminary exploration, and quality verification.

- Navigating organizational data silos
- Tracking data provenance
- Identifying missing data points

3. Data Preparation

Cleaning, formatting, and feature engineering (handling missing values, merging formats).

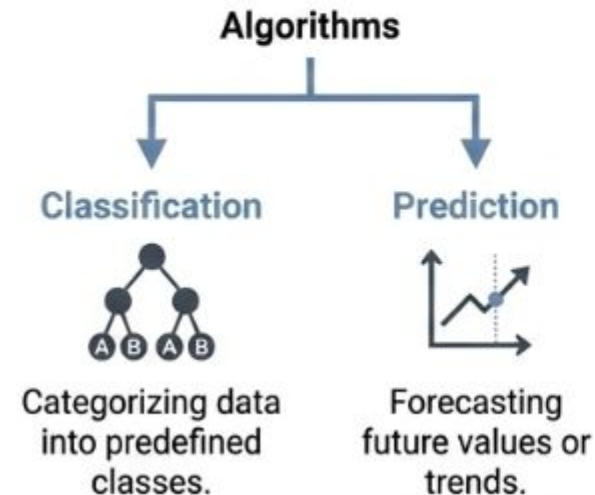
**70-80%
of Total
Project
Effort**

CRISP-DM process

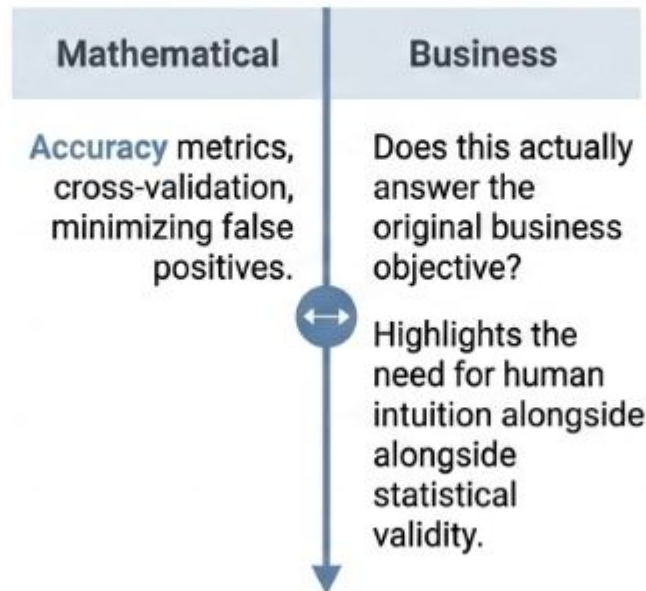
Definition

4. Modeling

Applying algorithms to extract patterns.



5. Evaluation



6. Deployment

System integration, automation, and drafting business recommendation memos.

The 1952 UNIVAC Anecdote

The computer **correctly predicted** the US Presidential election but was doubted by networks.

Overcoming user skepticism and **clear communication** are mandatory when deploying models.

CRISP-DM process

Definition

```
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score

# Constructing the pipeline
model_pipeline = Pipeline([
    ('scaler', StandardScaler()),
    ('classifier', DecisionTreeClassifier(max_depth=5))
])

# Execution and evaluation
model_pipeline.fit(X_train, y_train)
predictions = model_pipeline.predict(X_test)
score = accuracy_score(y_test, predictions)
```

Phase 3: *Data Preparation*
(Handling variable scales)

Phase 4: *Modeling*
(Applying the algorithm)

Phase 5: *Evaluation*
(Mathematical validation)

Content

- About
- What is data mining?
- Data organization
- CRISP-DM process
- Tools

Tools

Orange Data Mining



Cross-Platform & Free

Open-source software functioning seamlessly across all major operating systems.



No-Code Interface

Visual programming environment reliant entirely on intuitive drag-and-drop mechanics.



Standalone Environment

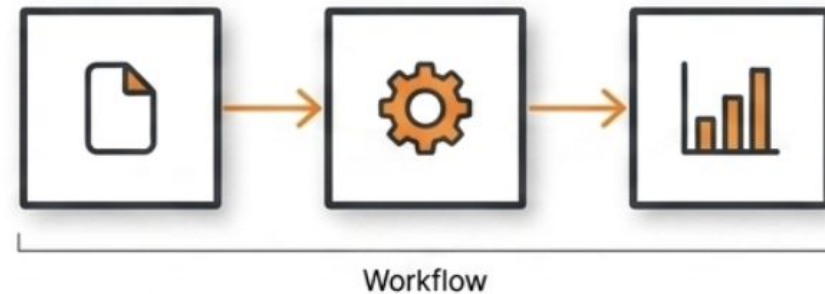
Installs as an independent ecosystem, requiring no complex environment configuration.

Tools

Orange Data Mining



Widget (Functional Block): The atomic unit of Orange. Each widget performs one specific data task (e.g., loading data, drawing a chart, or running an algorithm).



Workflow (The Pipeline): The continuous sequence created by connecting widgets. Data flows from left to right, building the complete analytical process.

Tools

Orange Data Mining

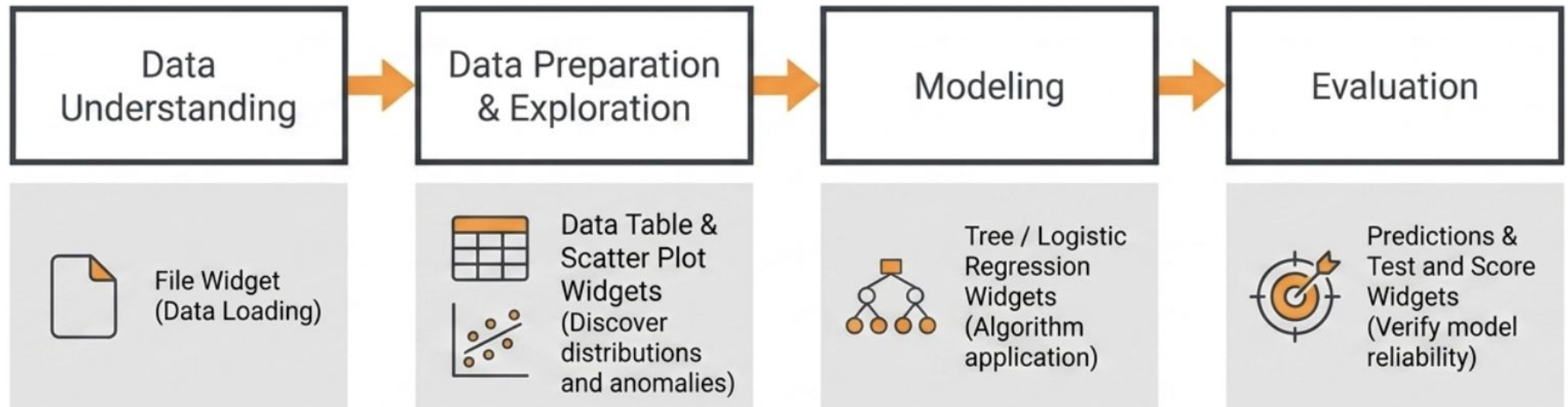
	The Coder's Focus	The Business Analyst's Focus (Orange)
Primary Challenge	Memorizing syntax and building algorithms from scratch.	Cultivating analytical thinking and solving business problems.
Learning Curve	Steep and technically demanding.	Gentle and approachable, preventing cognitive overload.

Power Under the Hood: While focused on business logic, Orange is not stripped down. It comes pre-packaged with a complete Python environment and industry-standard data science libraries, including **pandas** and **scikit-learn**.

Tools

Orange Data Mining

Mapping Orange Workflows to CRISP-DM



Orange Data Mining



The Role: Handles highly specific or custom data manipulation tasks that standard widgets do not natively support.

The Scope for Business Analysts: Python integration is utilized strictly at a read-and-understand level.

The Expectation: Students will tune parameters and refine data processing scripts. There is zero expectation to write complex algorithms from scratch.

Thank you!