FACULTY OF ENGINEERING
AT SRI RACHA
DEPARMENT OF COMPUTER ENGINEERING

03603351 วิทยาศาสตร์ข้อมูลเบื้องต้น

Introduction to Data Science

กระบวนการวิทยาศาสตร์ข้อมูล

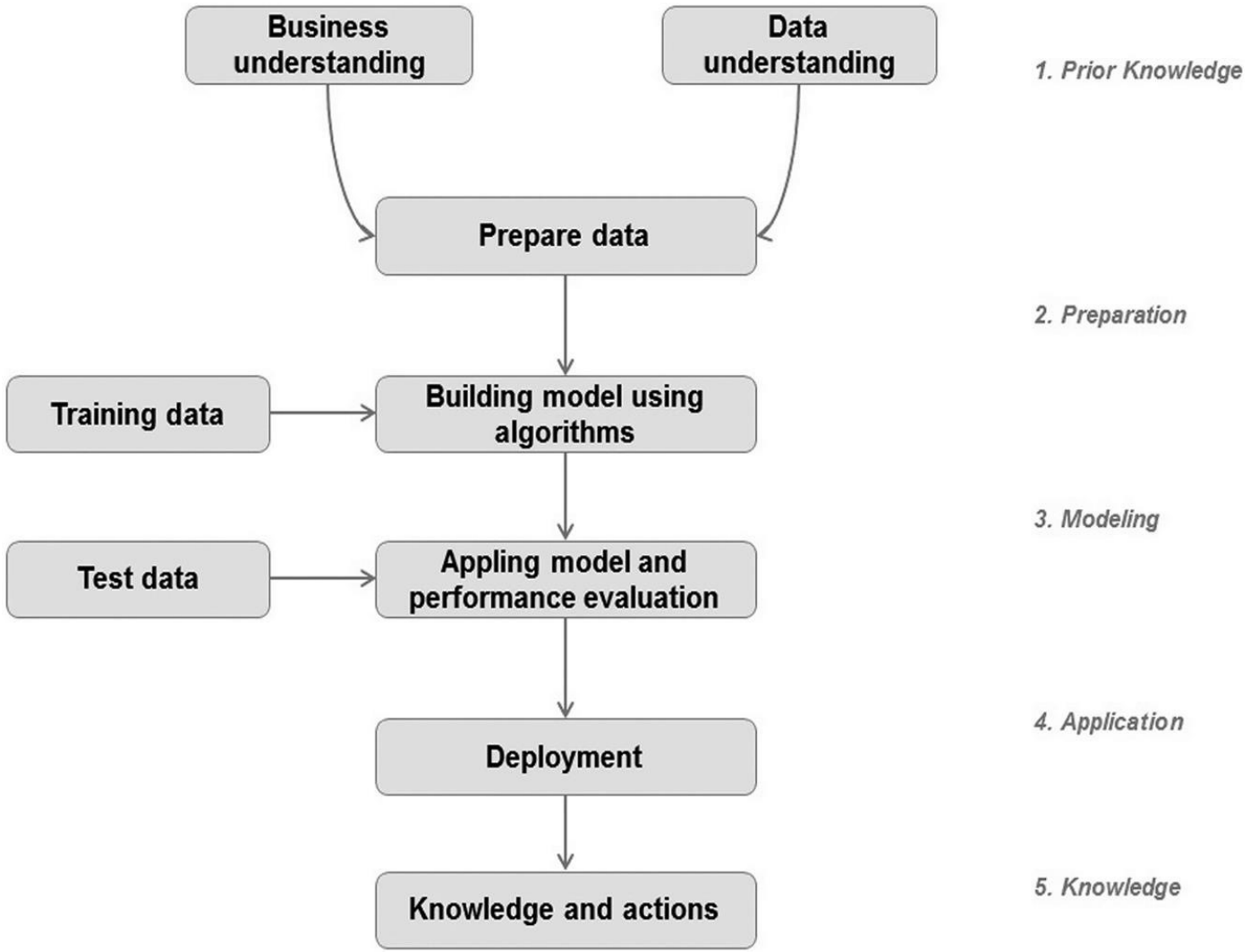
Data Science Process

ผศ.ดร. กุลวดี สมบูรณ์วิวัฒน์

kulwadee@eng.src.ku.ac.th

ภาคการศึกษาที่ 1 ปี 2565

กระบวนการทางวิทยาศาสตร์ข้อมูล (Data Science Process)



กรณีศึกษา: ธุรกิจการให้สินเชื่อสำหรับลูกค้ารายย่อย

=> ต้องการหาอัตราดอกเบี้ยการให้สินเชื่อที่เหมาะสมกับลูกค้าแต่ละราย

1. Prior Knowledge

1.1 กำหนด Business Problem

- ถือว่าเป็นขั้นตอนที่สำคัญที่สุดในกระบวนการวิทยาศาสตร์ข้อมูล และจำเป็นอย่างยิ่งที่จะต้องกำหนดปัญหาให้ชัดเจนถูกต้อง

ถ้าเรามีข้อมูลเกี่ยวกับอัตราดอกเบี้ยสินเชื่อและคะแนนเครดิตของผู้กู้ยืมในอดีต,
เราสามารถสร้างโมเดลสำหรับทำนาย
อัตราดอกเบี้ยที่เหมาะสมของผู้กู้ยืมรายใหม่จากคะแนนเครดิตได้หรือไม่?

1.2 ศึกษาบริบทในเชิงธุรกิจ

- เข้าใจภาพกว้างของธุรกิจการให้สินเชื่อ
- รายละเอียดเกี่ยวกับกระบวนการและข้อมูลที่ใช้ในการสมัคร
- วิธีการกำหนดดอกเบี้ยที่เหมาะสม

ตารางที่ 2.1 ชุดข้อมูลล (Dataset) ของผู้ขอสินเชื่อ

input features, attributes
อินพุทฟีเจอร์

label, class label,
target variable
ลาเบล
ค่าตัวแปรเป้าหมาย

Borrow ID	Credit Score	Interest Rate (%)
01	500	7.31
02	600	6.70
03	700	5.95
04	700	6.40
05	800	5.40
06	800	5.70
07	750	5.90
08	550	7.00
09	650	6.50
10	825	5.70

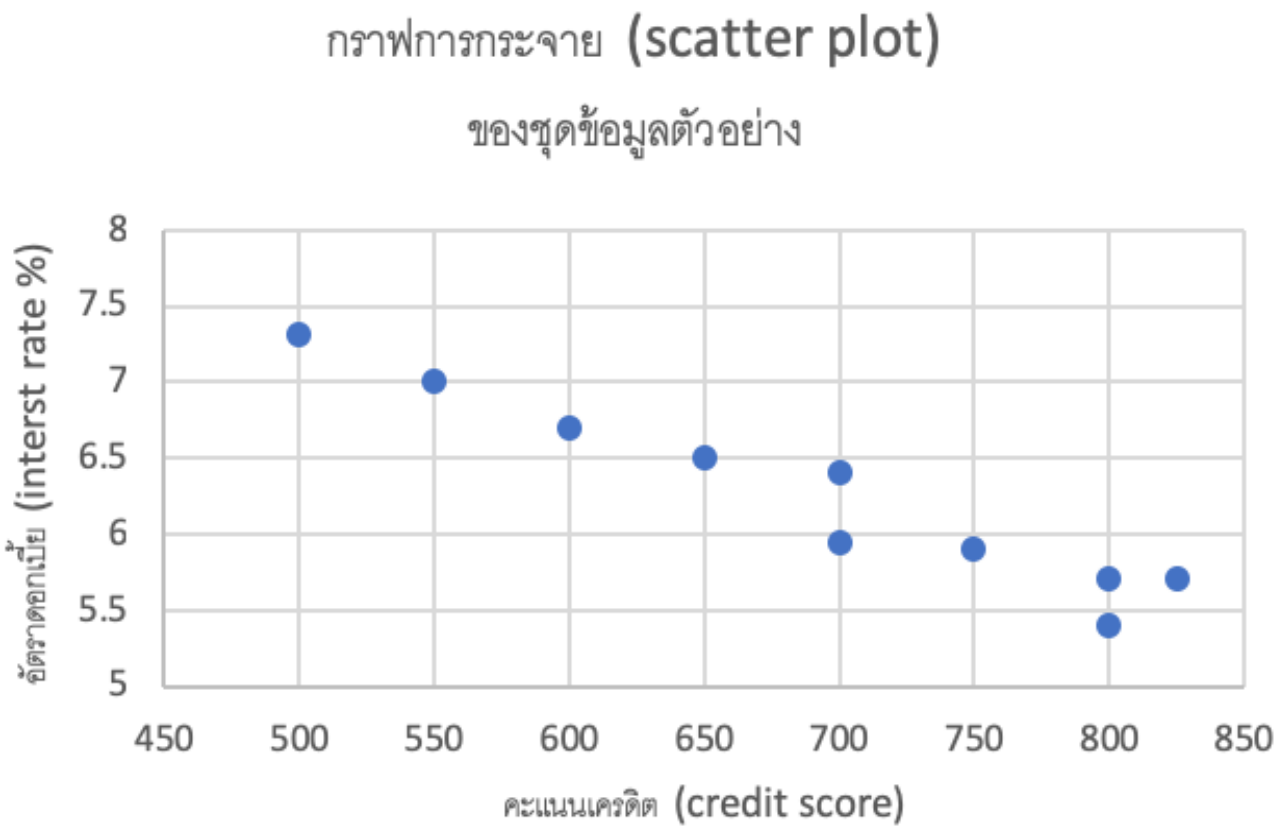
instance, sample,
data point
ตัวอย่าง

1.3 รวบรวมและทำความเข้าใจข้อมูล

2. Model Building

2.1 Data Exploratory Analysis

- การสำรวจข้อมูล (data exploration หรือ exploratory data analysis) คือการทำความเข้าใจเกี่ยวกับข้อมูลเบื้องต้น โดยการประยุกต์ใช้เครื่องมือพื้นฐานสำหรับการวิเคราะห์ข้อมูล เช่น สถิติพรรณนา (descriptive statistics) และการทำให้เห็นเป็นภาพ (data visualization) เป็นต้น



ความสัมพันธ์ระหว่างคะแนนเครดิต กับอัตราดอกเบี้ยมีลักษณะ**แปรผกผัน** กล่าวคือ ยิ่งคะแนนเครดิตสูง อัตราดอกเบี้ยที่เหมาะสมก็จะต่ำลง

2.2 เตรียมข้อมูล

- เป็นขั้นตอนที่ใช้เวลานานที่สุดในกระบวนการวิทยาศาสตร์ข้อมูล เนื่องจาก โดยปกติชุดข้อมูลที่รวบรวมมาได้ จะอยู่ในรูปแบบที่ไม่เหมาะกับการประมวลผลของอัลกอริทึมทางวิทยาศาสตร์ข้อมูล ซึ่งส่วนใหญ่ต้องการอินพุทที่มีโครงสร้างแบบตาราง โดยแต่ละแถวคือหนึ่ง instance และแต่ละคอลัมน์คือ attribute
- กรรมวิธีที่ใช้ในการเตรียมข้อมูลมีหลายวิธี เช่น การเติมค่าที่หายไปด้วยค่าเฉลี่ย ค่าแปลงค่าให้อยู่ในช่วงมาตรฐาน การจัดการค่าผิดปกติ (outliers), การคัดเลือกฟีเจอร์ (feature selection), และการสุ่มตัวอย่างข้อมูล (data sampling)

2.3 สร้างชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ

Borrow ID	Credit Score	Interest Rate (%)
01	500	7.31
02	600	6.70
03	700	5.95
04	700	6.40
05	800	5.40
06	800	5.70
07	750	5.90
08	550	7.00
09	650	6.50
10	825	5.70

ตารางที่ 2.2 ชุดข้อมูลฝึกฝน (training dataset)

Borrow ID	Credit Score X	Interest Rate (%) y
01	500	7.31
02	600	6.70
03	700	5.95
05	800	5.40
06	800	5.70
08	550	7.00
09	650	6.50

ตารางที่ 2.3 ชุดข้อมูลทดสอบ (testing dataset)

Borrow ID	Credit Score X	Interest Rate (%) y
04	700	6.40
07	750	5.90
10	825	5.70

2.4 กำหนดโมเดลที่จะใช้ และสร้างโมเดลด้วยชุดข้อมูลฝึกฝน

Linear Regression Model: $y = wX + b$

2.4 กำหนดโมเดลที่จะใช้ และสร้างโมเดลด้วยชุดข้อมูลฝึกฝน

ตารางที่ 2.2 ชุดข้อมูลฝึกฝน (training dataset)

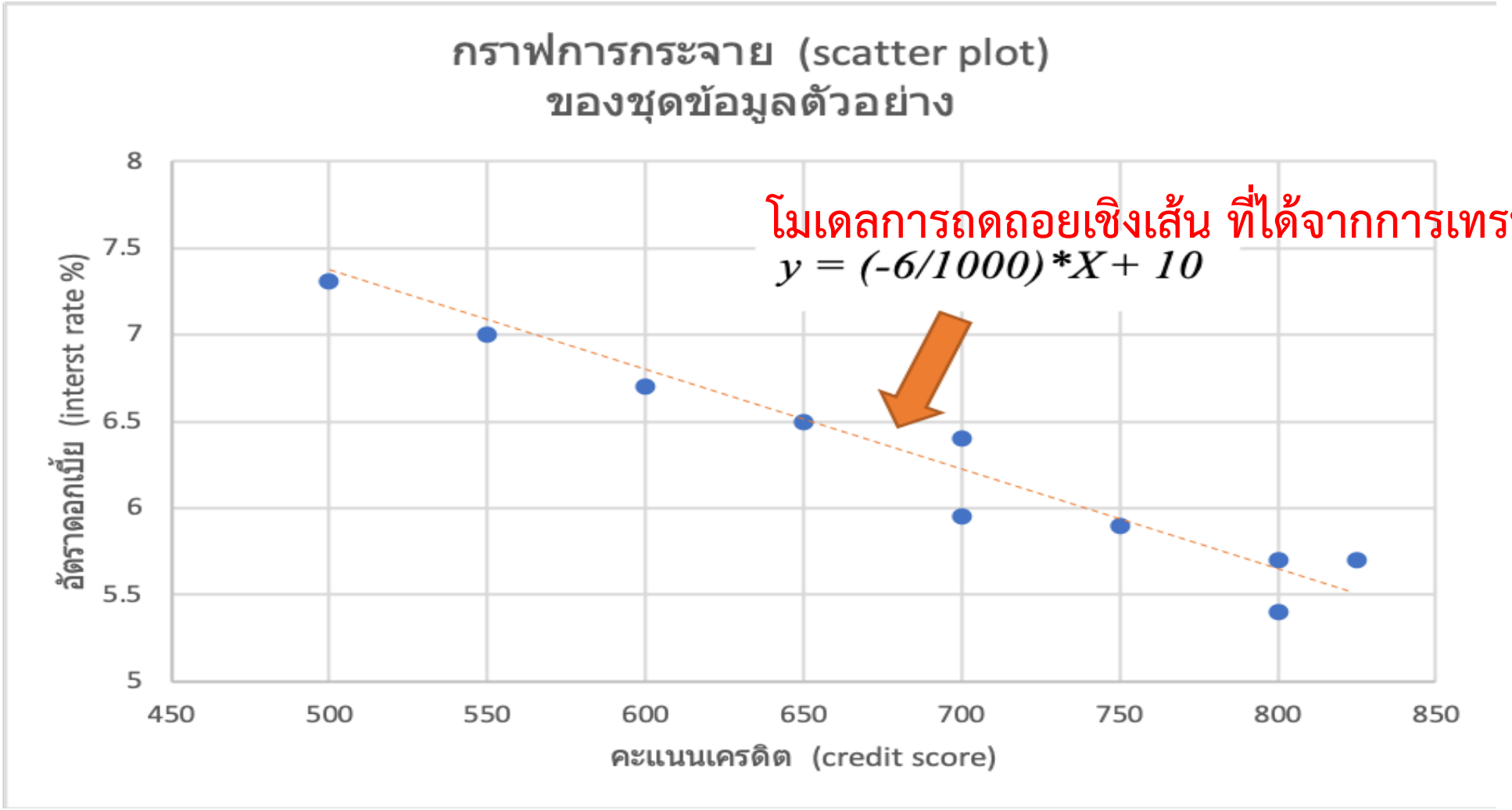
Borrow ID	Credit Score X	Interest Rate (%) y
01	500	7.31
02	600	6.70
03	700	5.95
05	800	5.40
06	800	5.70
08	550	7.00
09	650	6.50

→ Linear Regression Training Algorithm →

$$y = \frac{-6}{1000}X + 10$$

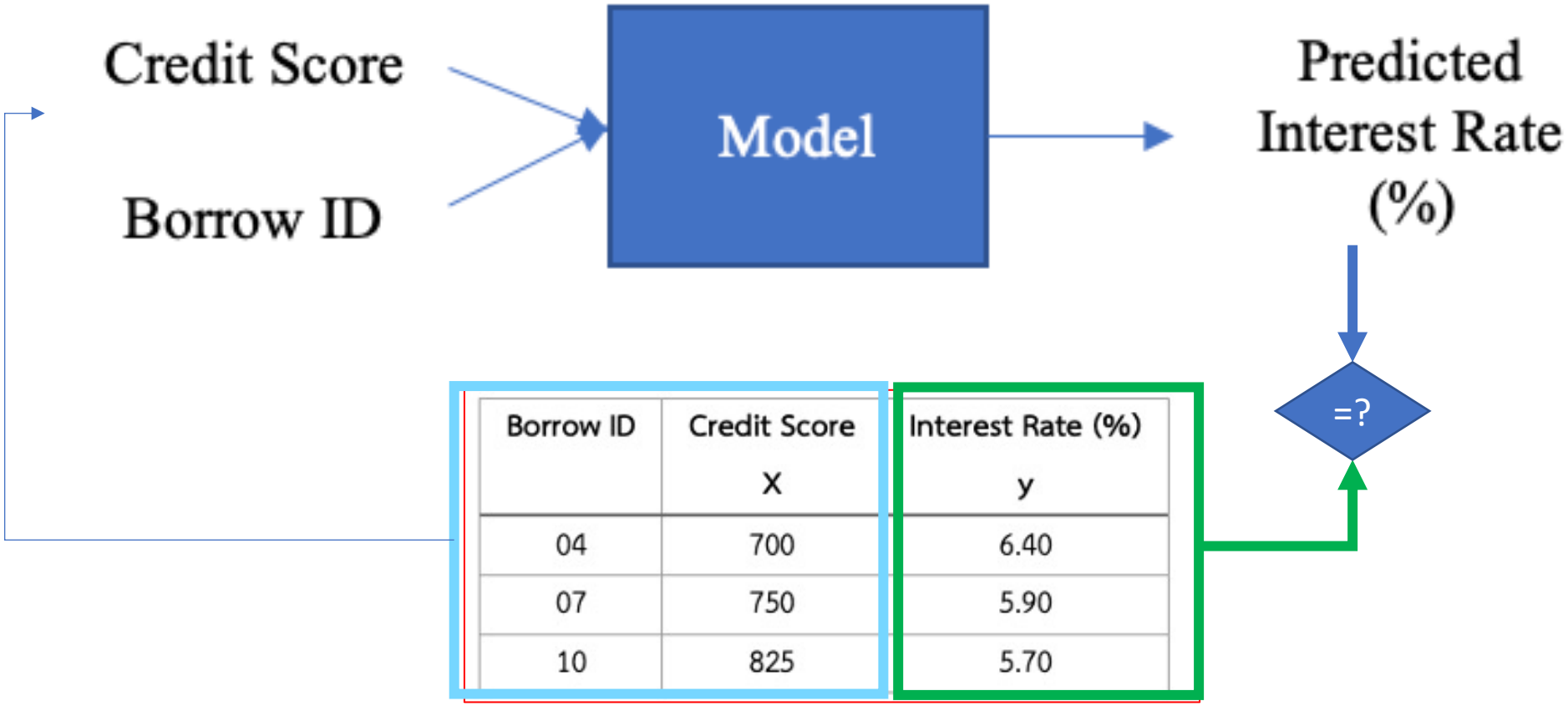
$$y = wX + b$$

2.4 กำหนดโมเดลที่จะใช้ และสร้างโมเดลด้วยชุดข้อมูลฝึกฝน



3. Model Testing

3.1 ประเมินประสิทธิภาพโมเดลด้วยชุดข้อมูลทดสอบ



Borrow ID	Credit Score X	Interest Rate (%) y
04	700	6.40
07	750	5.90
10	825	5.70

3.1 ประเมินประสิทธิภาพโมเดลด้วยชุดข้อมูลทดสอบ

ตารางที่ 2.4 การประเมินประสิทธิภาพของโมเดลบนชุดข้อมูลทดสอบ

Borrow ID	Credit Score X	Interest Rate (%) y	คำทำนายอัตราดอกเบี้ยที่ได้จากโมเดล y= (-6/1000)X+ 10	Errors	Squared Errors
04	700	6.40	5.8	-0.6	0.36
07	750	5.90	5.5	-0.4	0.16
10	825	5.70	5.05	-0.65	0.4225

RMSE (Root Mean Squared Error)

$$= \frac{1}{3}\sqrt{(0.36 + 0.16 + 0.4225)}$$

$$= 0.324$$

4. Deployment

4. การนำโมเดลไปใช้งานจริง

- นำโมเดลที่ได้รวมเข้ากับแอปพลิเคชั่นขององค์กร เช่น ระบบงานอนุมัติสินเชื่อ
- ต้องมีการประสานงานกับส่วนงานอื่น ๆ ที่เกี่ยวข้องเช่น ผู้บริหาร ฝ่ายอนุมัติสินเชื่อ ฝ่ายไอที เป็นต้น

5. Knowledge and Actions

5. ความรู้และการกระทำ

- กระบวนการทางวิทยาศาสตร์ข้อมูลเริ่มต้นด้วย ความรู้ตั้งต้น (Prior Knowledge) และจบลงด้วยความรู้แจ้งที่เพิ่มเติมขึ้น ซึ่งได้มาจากกระบวนการเรียนรู้จากข้อมูลแบบทำซ้ำ นักวิทยาศาสตร์ข้อมูลจะต้องคัดสรรความรู้ใหม่ที่มีนัยสำคัญ และนำไปใช้ในการตัดสินใจ หรือการกระทำอื่น ๆ ที่เป็นประโยชน์ในเชิงธุรกิจ

การทำนายค่ามัธยฐานของราคาบ้านโดยใช้ California Housing Price dataset

- ไลบารี่ที่ใช้มีดังนี้คือ
  - numpy
  - pandas
  - matplotlib
  - sklearn

```
from sklearn.datasets import fetch_california_housing

housing = fetch_california_housing(as_frame=True)
```

```
from sklearn.model_selection import train_test_split

X = housing.data
y = housing.target
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

```
((14448, 8), (6192, 8), (14448,), (6192,))
```

```
from sklearn.linear_model import LinearRegression

lin_reg = LinearRegression()
lin_reg.fit(X_train, y_train)
```

```
from sklearn.metrics import mean_squared_error
import numpy as np # np.sqrt()

lin_predict = lin_reg.predict(X_test) # 6192 rows
lin_mse = mean_squared_error(y_test # actual value of housing price,
                             lin_predict)
lin_rmse = np.sqrt(lin_mse)
print("RMSE of Linear Regressor = ", lin_rmse)
```

AI (Artificial Intelligence)

Machine Learning

Data Mining

KDD (Knowledge Discovery in Databases)

Data Science

What is artificial intelligence? (AI)

Hard for Machines

General intelligence

Manipulation
Driving
Images
Faces
Language
Jeopardy!

'Deep' Learning

Prime factorization

Knowledge

Planning
Logic
Reasoning
Remembering
Math
Integration

'Classical' AI

Go
Chess

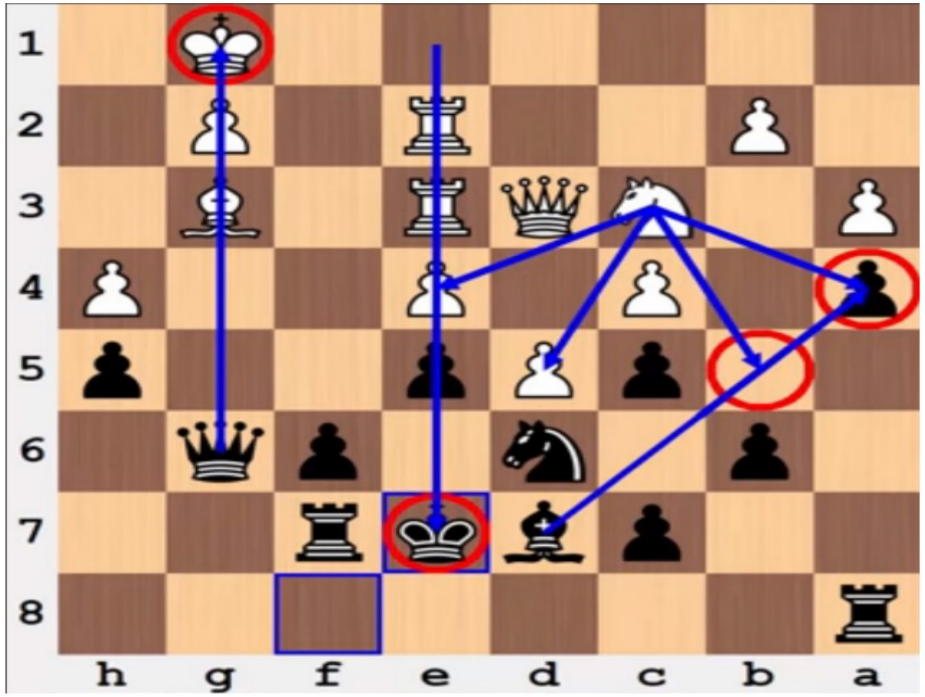
Easy for Machines

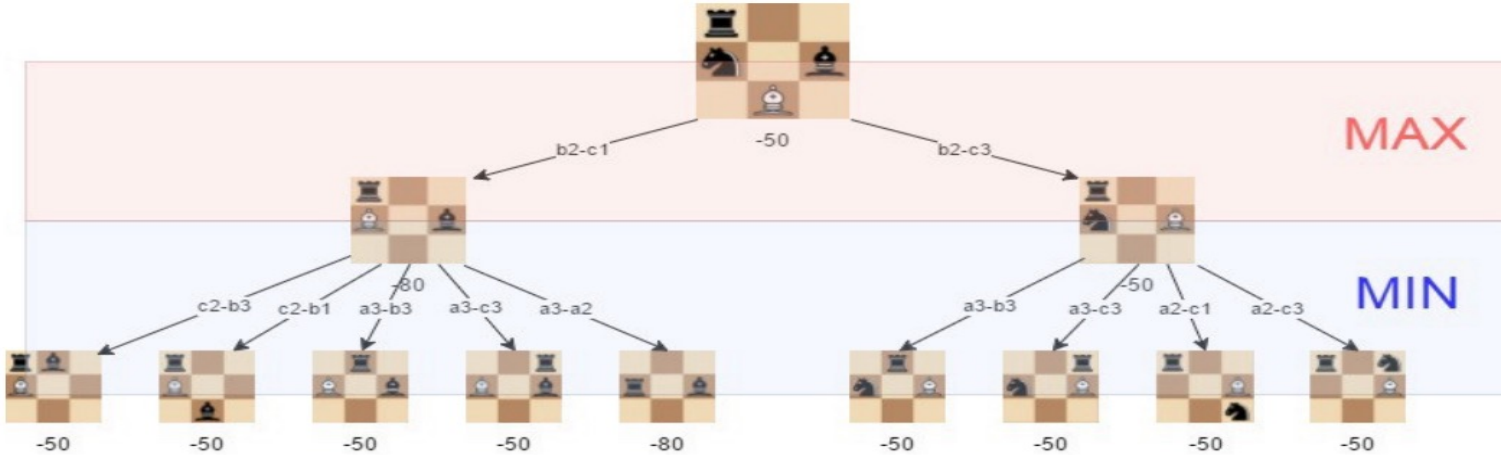
Easy for humans

Hard for humans

Source: https://mit6874.github.io/assets/sp2021/slides/l01.pdf

How do machines play chess?

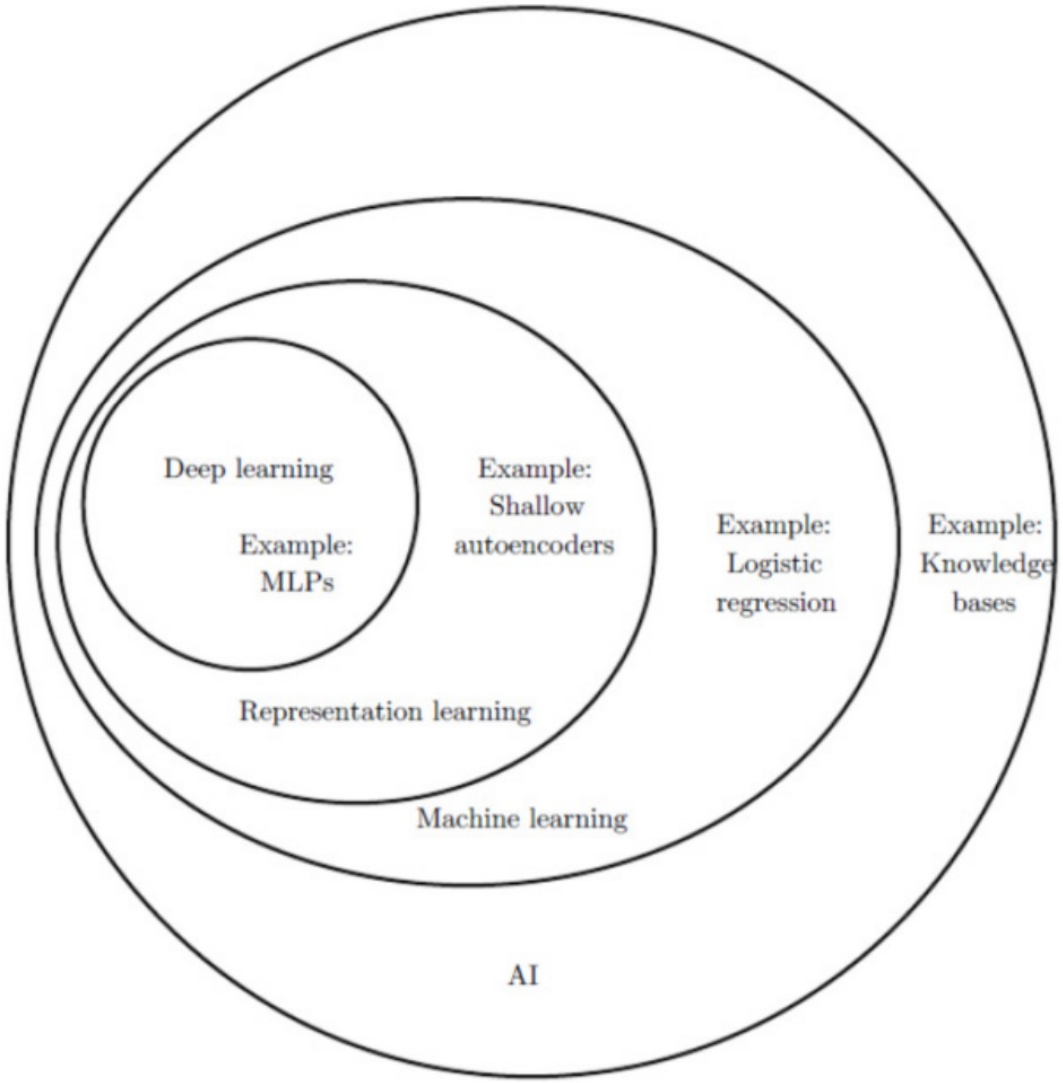


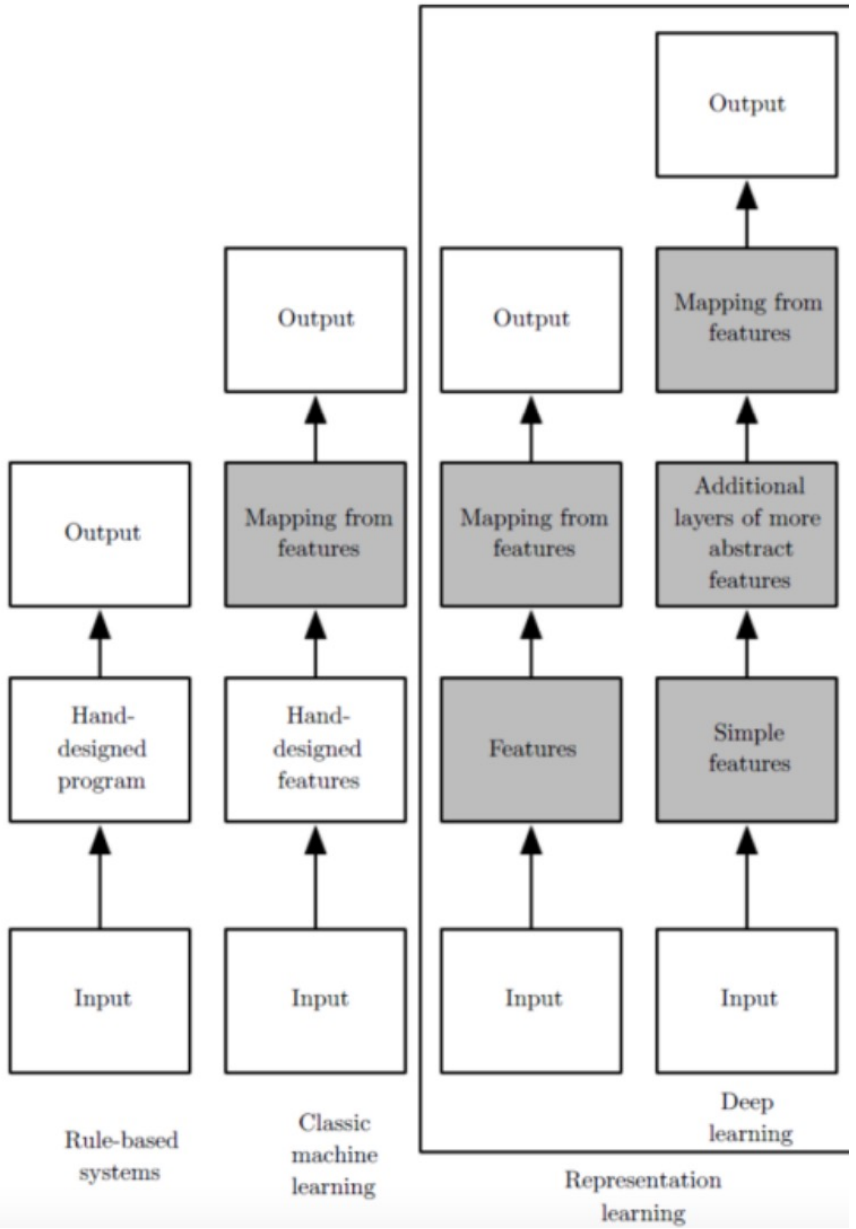


‘Classical’ AI approach (rule-based, tree search):

1. Human: Program in all the rules of chess
2. Human: Hand-craft a scoring function for each position
3. Search all moves that you can make (max score)
4. Search all moves that opponent can make (min score)
5. Repeat for many iterations
6. Choose move that gives best score

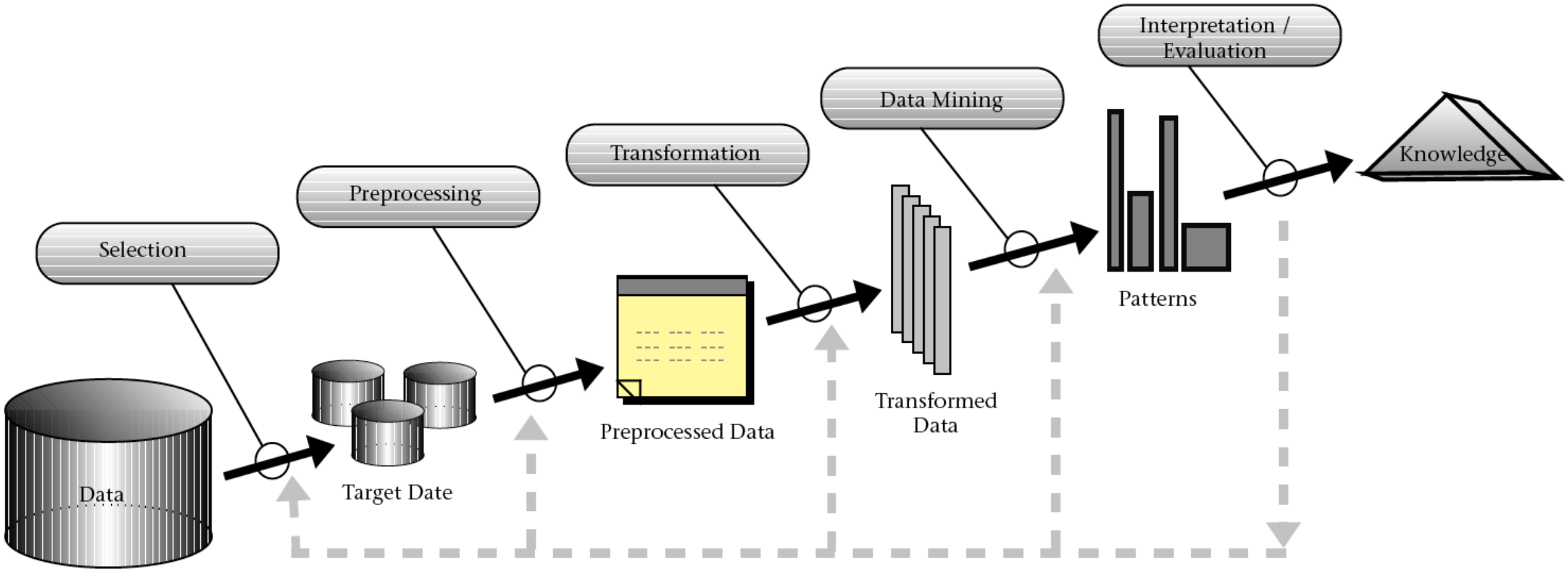
Source: https://mit6874.github.io/assets/sp2021/slides/l01.pdf

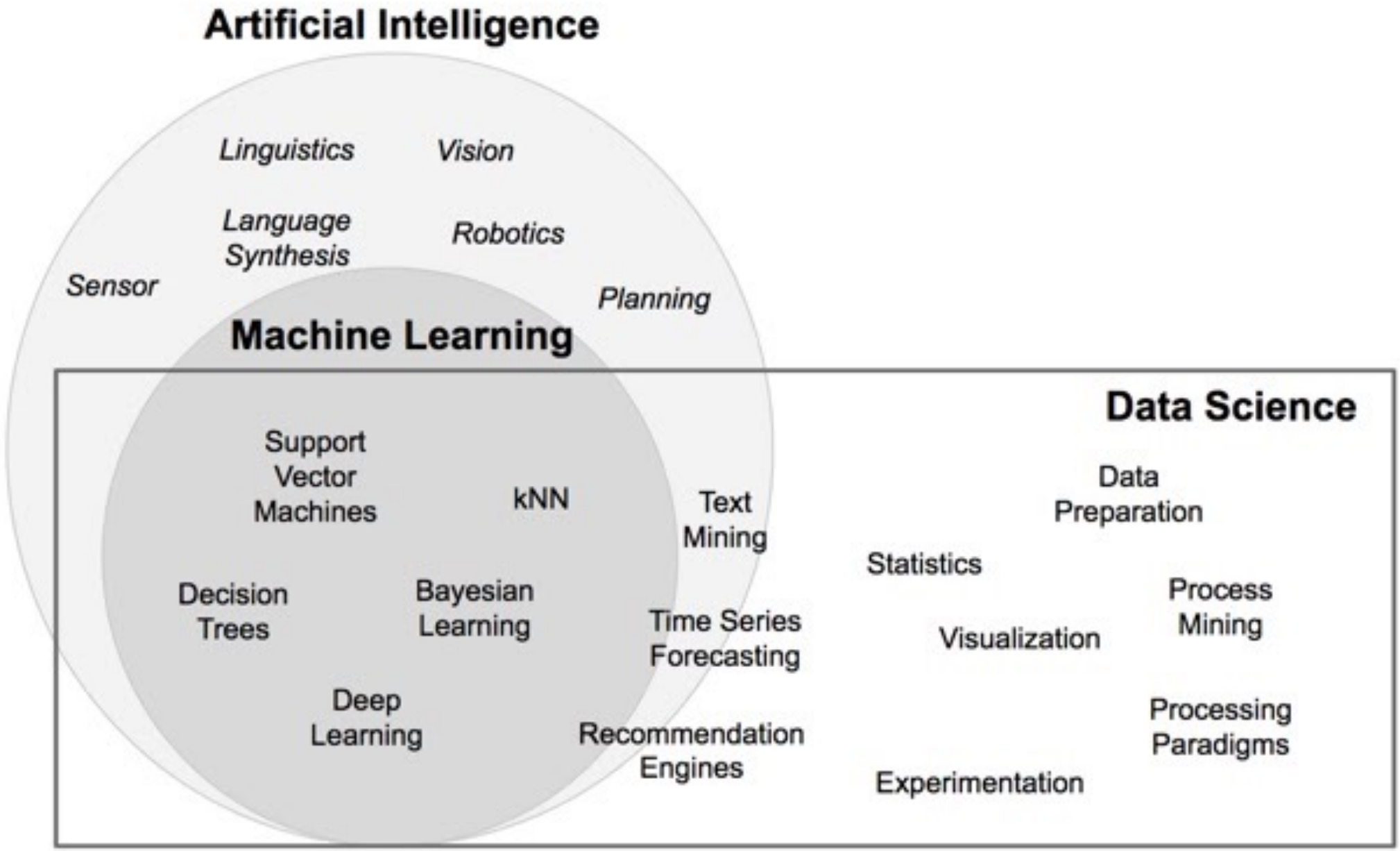




oodfellow 2016

KDD Process (Fayyad et al., 1996)




Artificial Intelligence
Linguistics
Vision
Language Synthesis
Robotics
Sensor
Planning
Machine Learning
Data Science
Support Vector Machines
kNN
Text Mining
Data Preparation
Statistics
Decision Trees
Bayesian Learning
Process Mining
Time Series Forecasting
Visualization
Deep Learning
Processing Paradigms
Recommendation Engines
Experimentation