

กระบวนการและงานทางวิทยาศาสตร์ข้อมูล

Data Science Process and Tasks



FACULTY OF ENGINEERING
AT SRI RACHA
DEPARTMENT OF COMPUTER ENGINEERING

03603351 วิทยาศาสตร์ข้อมูลเบื้องต้น

Introduction to Data Science

ภาคการศึกษาที่ 1 ปี 2562

ผศ.ดร. กุลวดี สมบูรณ์วิวัฒน์

kulwadee@eng.src.ku.ac.th

ภาควิชาวิศวกรรมคอมพิวเตอร์

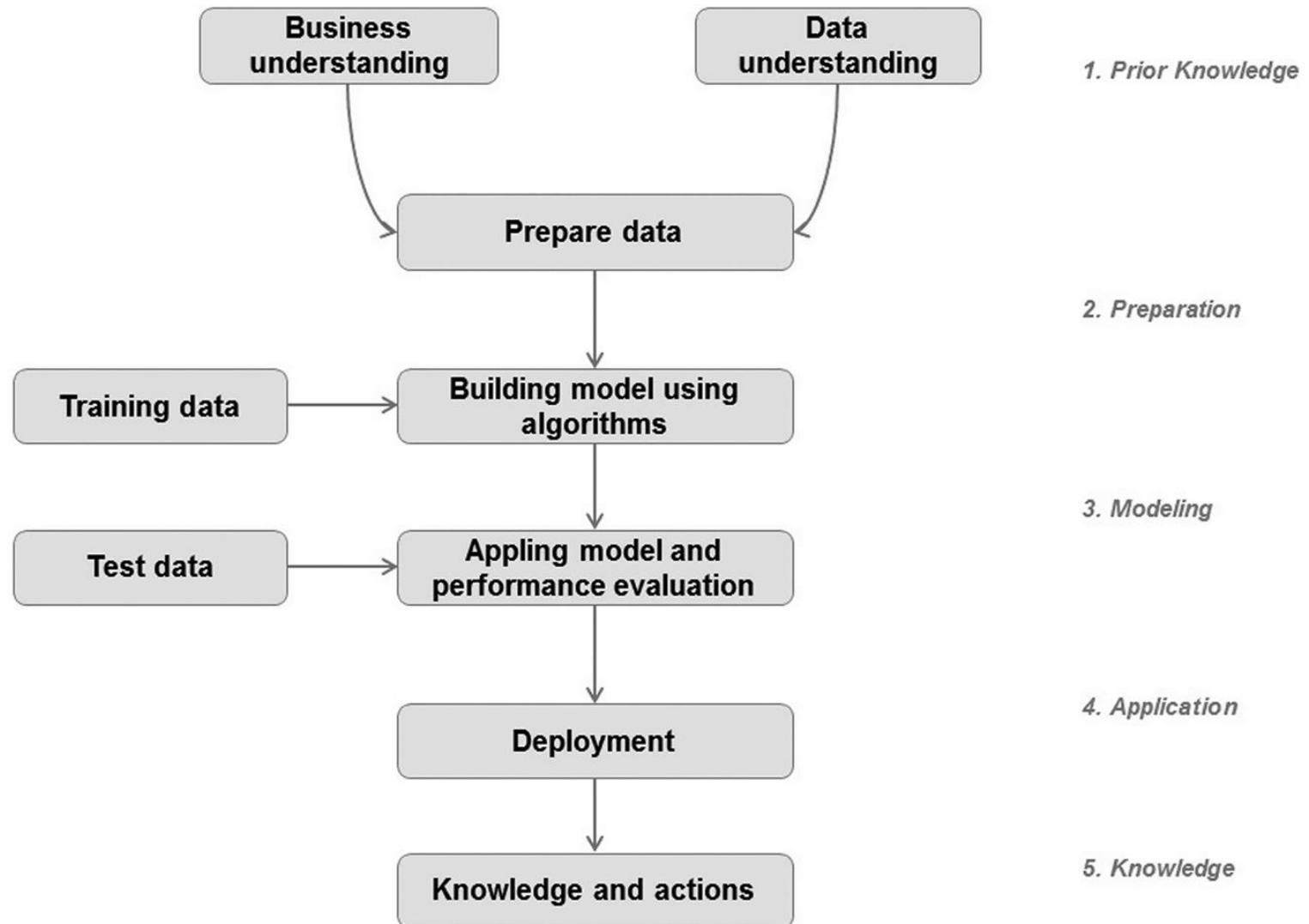
คณะวิศวกรรมศาสตร์ ศรีราชา

มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตศรีราชา

หัวข้อหลัก

- กระบวนการทางวิทยาศาสตร์ข้อมูล
 - นักวิทยาศาสตร์ข้อมูล ใช้กระบวนการทางวิทยาศาสตร์เป็นเครื่องมือในการค้นหารูปแบบແຜงในชุดข้อมูล
- งานหลักทางวิทยาศาสตร์ข้อมูลมี 5 ประเภท ได้แก่
 - การจำแนกประเภท
 - การวิเคราะห์การถดถอย
 - การจัดกลุ่ม
 - การวิเคราะห์ความสัมพันธ์
 - การตรวจจับความผิดปกติ

กระบวนการทางวิทยาศาสตร์ข้อมูล



กรณีศึกษา: ธุรกิจการให้สินเชื่อสำหรับลูกค้ารายย่อย

- วัตถุประสงค์ของการนำวิทยาศาสตร์ข้อมูลมาใช้
 - เพื่อทำนายอัตราดอกเบี้ยที่เหมาะสมสำหรับผู้ขอสินเชื่อรายใหม่
- บริบททางธุรกิจ
 - รายได้จากการเบี้ย
 - ความพึงพอใจของลูกค้า
 - ความเสี่ยงจากการให้สินเชื่อ (ผิดนัดชำระหนี้, หนี้เสีย)

ตารางที่ 2.1 ชุดข้อมูล (Dataset) ของผู้ขอสินเชื่อ

label, class label,
target variable

ลabeล

ค่าตัวแปรเป้าหมาย

input features, attributes

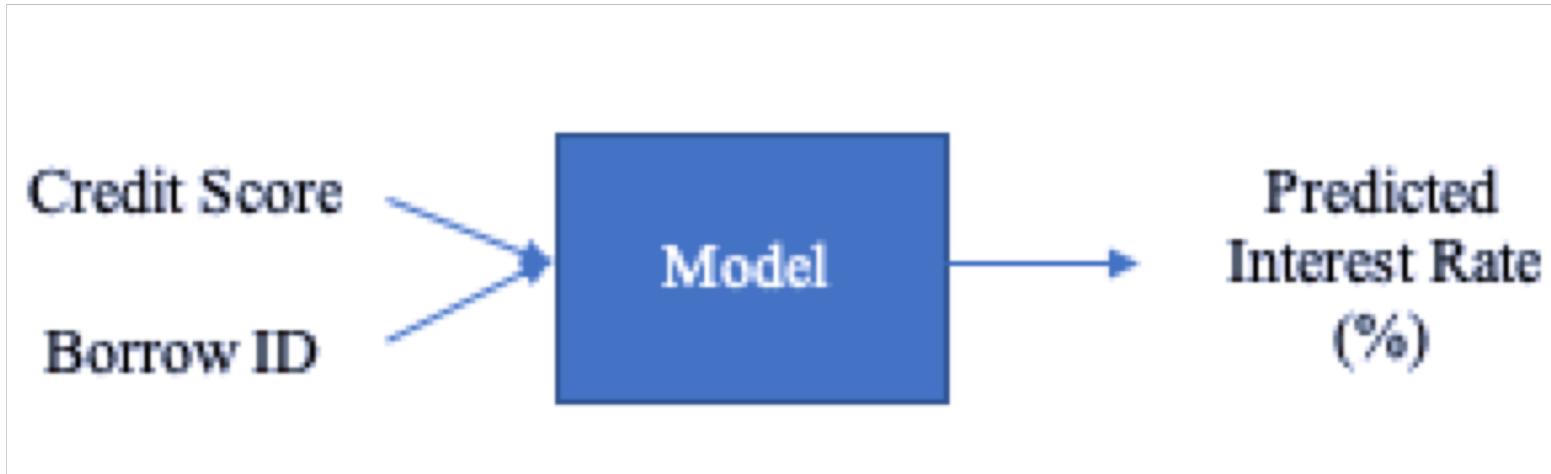
อินพุตฟีเจอร์

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

Data Understanding

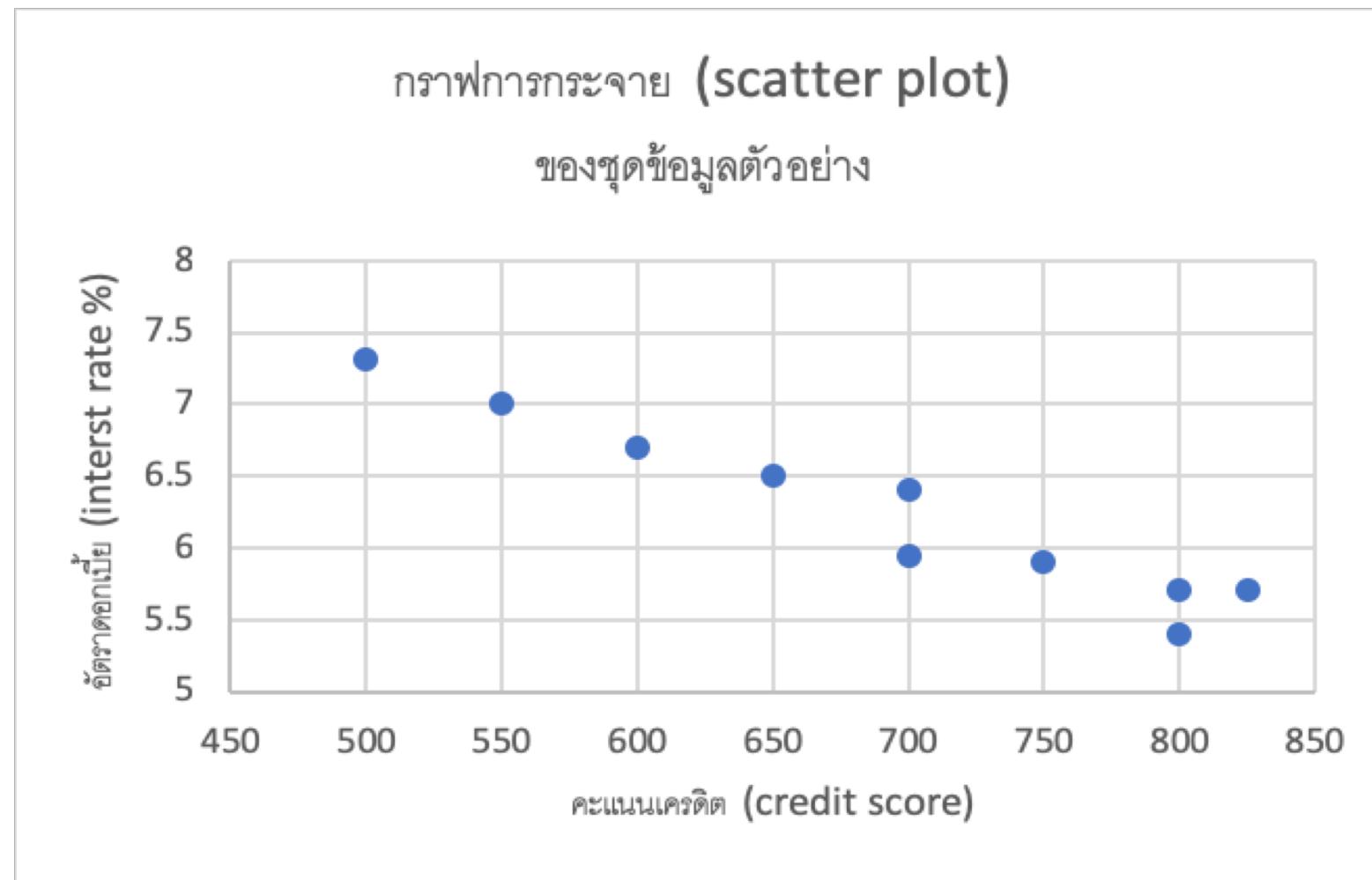
instance, sample,
data point
ตัวอย่าง

Data Understanding



ทำความเข้าใจข้อมูลเบื้องต้น (Data Exploration)

Data Preparation



การเตรียมข้อมูลสำหรับใช้สร้างโมเดล – data cleansing

Data Preparation

คุณภาพของข้อมูล มีผลต่อประสิทธิภาพของโมเดลที่ได้จากการเรียนรู้มาก หากเราป้อนข้อมูลที่มีคุณภาพต่ำ (เช่น มีข้อมูลซ้ำซ้อน ไม่ครบถ้วน) ให้กับอัลกอริทึมการเรียนรู้ ก็จะเป็นไปได้ยากมากที่โมเดลที่ได้จะมีประสิทธิภาพสูง

- การกำจัดเรkorด์ซ้ำ (elimination of duplicate records)
- การแยกค่าผิดปกติ (outliers)
- การทำค่าของแอทริบิวต์ให้อยู่ในรูปแบบ/ช่วงมาตรฐาน (standardization of attribute values)
- การแทนค่าที่ขาดหายไป (substitution of missing values)
- การคัดเลือกฟีเจอร์ (feature selection)

การเตรียมข้อมูลสำหรับใช้สร้างโมเดล – training/testing data

Data Preparation

เพื่อให้การประเมินประสิทธิภาพของโมเดลที่สร้างขึ้น มีความเที่ยงตรงแม่นยำ ข้อมูลที่ใช้สำหรับประเมินประสิทธิภาพของโมเดลจะต้องเป็นข้อมูลที่ไม่เคยถูกป้อนให้กับโมเดลมาก่อน ดังนั้น ก่อนเริ่มการฝึกฝน (model training) เราต้องแยกข้อมูลออกเป็นสองส่วนคือ ข้อมูลส่วนแรกใช้สำหรับการฝึกฝน และ ส่วนที่สองใช้สำหรับการทดสอบประเมินประสิทธิภาพ ซึ่งในทางปฏิบัติ มีหลักการแบ่งคือ ให้สุ่มเลือก 2 ใน 3 (หรือประมาณ 70%) ของข้อมูลทั้งหมด เป็นข้อมูลสำหรับฝึกฝน และข้อมูลที่เหลืออีก 1 ใน 3 เป็นข้อมูลสำหรับทดสอบประสิทธิภาพของโมเดล

การเตรียมข้อมูลสำหรับใช้สร้างโมเดล – training/testing data

Data Preparation

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

Training dataset

สุ่มเลือก ในอัตราส่วน 70/30

Testing 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

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

เลือกโมเดล

linear regression (การลดด้วยเชิงเส้น)

Model Building

$$y = wX + b$$

เมื่อ

y คือค่า เอาท์พุท หรือ ตัวแปรตาม (dependent variable)

X คือค่า อินพุท หรือ ตัวแปรต้น (independent variable)

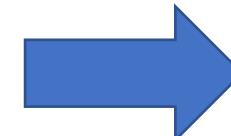
w คือค่าสัมประสิทธิ์ของสมการ (coefficients)

b คือจุดตัดแกน y (y-intercept)

สร้างโมเดลโดยป้อนชุดข้อมูลฝึกฝนให้กับอัลกอริทึมการเรียนรู้

Model Building

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



$$y = wX + b$$

ทราบค่า X และค่า y ,
ต้องการหา w และ b

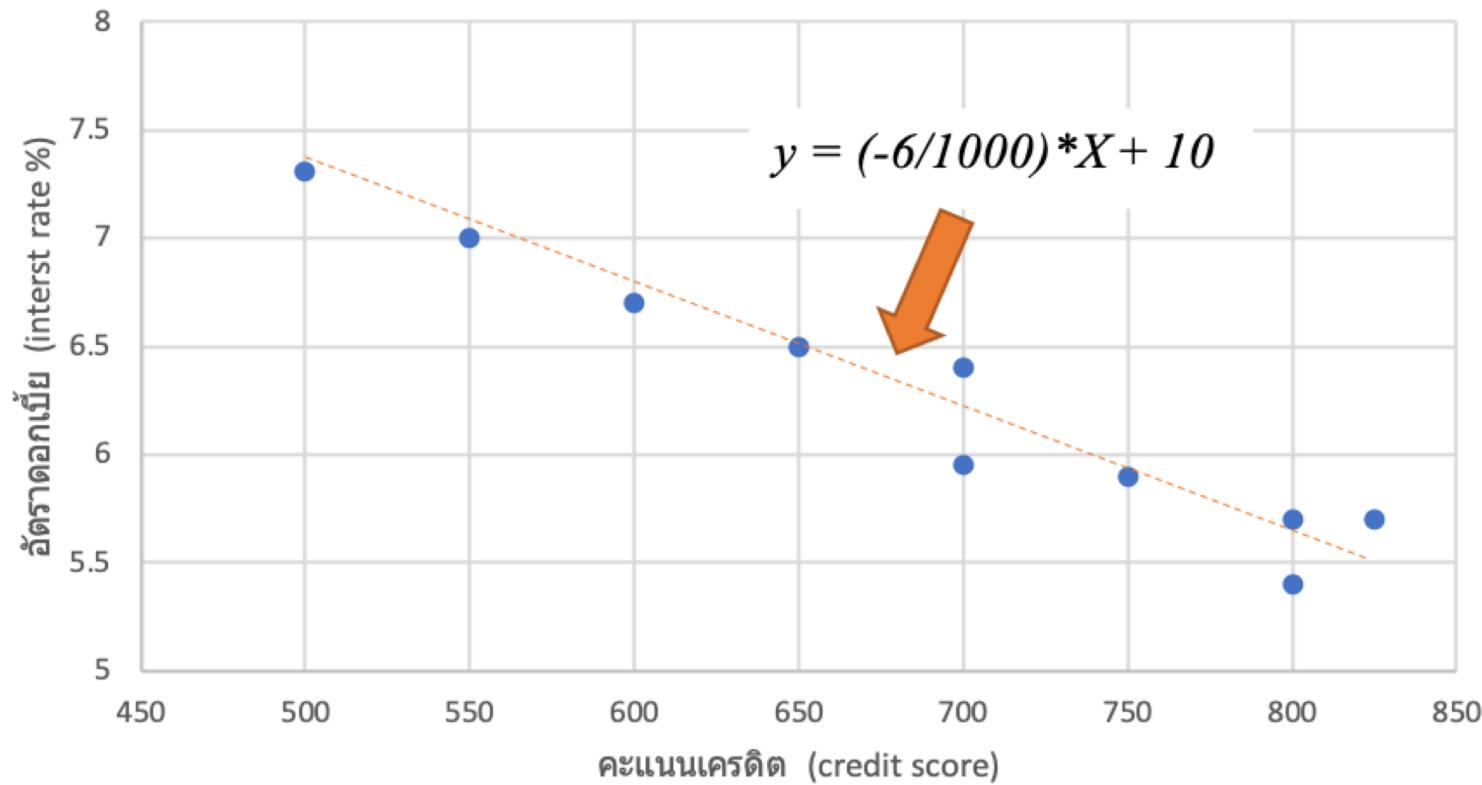


$$w, b$$

สร้างโมเดลโดยป้อนชุดข้อมูลฝึกฝนให้กับอัลกอริทึมการเรียนรู้

Model Building

กราฟการกระจาย (scatter plot)
ของชุดข้อมูลตัวอย่าง



$$W = -6/1000$$

$$b = 10$$

ประเมินประสิทธิภาพของโมเดลด้วยชุดข้อมูลทดสอบ

Borrow ID	Credit Score X	Interest Rate (%) y	คำทำนายอัตราดอกเบี้ยที่ได้ จากโมเดล $y = (-6/1000)X + 10$	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)

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

$$= 0.324$$

การนำโมเดลไปใช้งาน

Deployment

ผลลัพธ์ที่ได้จากการกระบวนการทางวิทยาศาสตร์ข้อมูล จะต้องถูกนำไปหலомรวมเข้ากับกระบวนการทางธุรกิจ (business process) ซึ่งส่วนมากจะอยู่ในรูปแบบแอปพลิเคชันซอฟต์แวร์ สิ่งที่ต้องคำนึงถึงในขั้นตอนนี้ ได้แก่ ความพร้อมในการนำไปใช้ในระบบจริง (production system) การผสานเข้ากับระบบงานอื่น (technical integration) เวลาตอบสนอง (response time) การอัพเดตโมเดล (model refresh) การส่งต่อผลลัพธ์ไปยังผู้ใช้งาน (assimilation)

ความรู้และการกระทำ

Knowledge and Action

กระบวนการทางวิทยาศาสตร์ข้อมูลเริ่มต้นด้วย ความรู้ตั้งต้น (Prior Knowledge) และจบลงด้วย ความรู้เจ็บที่เพิ่มเติมขึ้น ซึ่งได้มาจากกระบวนการเรียนรู้จากข้อมูลแบบทำซ้ำ นักวิทยาศาสตร์ข้อมูล จะต้องคัดสรรความรู้ใหม่ที่มีนัยสำคัญ และนำไปใช้ในการตัดสินใจ หรือการกระทำอื่น ๆ ที่เป็นประโยชน์ในเชิงธุรกิจ

วิทยาศาสตร์ข้อมูลใช้ความรู้จากสาขาวิชาใดในข้อต่อไปนี้ ในการวิเคราะห์และค้นหารูปแบบที่มีนัยสำคัญในข้อมูล

- ก. สติติและคณิตศาสตร์
- ข. การบริหารจัดการข้อมูล
- ค. การเรียนรู้ของเครื่องจักร
- ง. ถูกทุกข์

ก่อนการสร้างโมเดลการเรียนรู้จากข้อมูล นักวิทยาศาสตร์ข้อมูลจะต้องทำสิ่งใด

- ก. ทำความเข้าใจปัญหา, business domain และทำความเข้าใจข้อมูลที่ต้องใช้ในการแก้ปัญหา
- ข. ทดสอบประสิทธิภาพของโมเดลที่ต้องการใช้
- ค. เตรียมข้อมูลให้พร้อม โดยการใช้เทคนิค data cleaning แบบต่าง ๆ และแบ่งชุดข้อมูลออกเป็น training dataset และ testing dataset
- ง. ข้อ (ก) และ (ค)

Data Science Tasks

- งานหลักทางวิทยาศาสตร์ข้อมูลมี 5 ประเภท ได้แก่
 - การจำแนกประเภท (Classification)
 - การวิเคราะห์การถดถอย (Regression)
 - การจัดกลุ่ม (Clustering)
 - การวิเคราะห์ความสัมพันธ์ (Association Analysis)
 - การตรวจจับความผิดปกติ (Outlier Detection)

Prediction Problems (Supervised Learning)

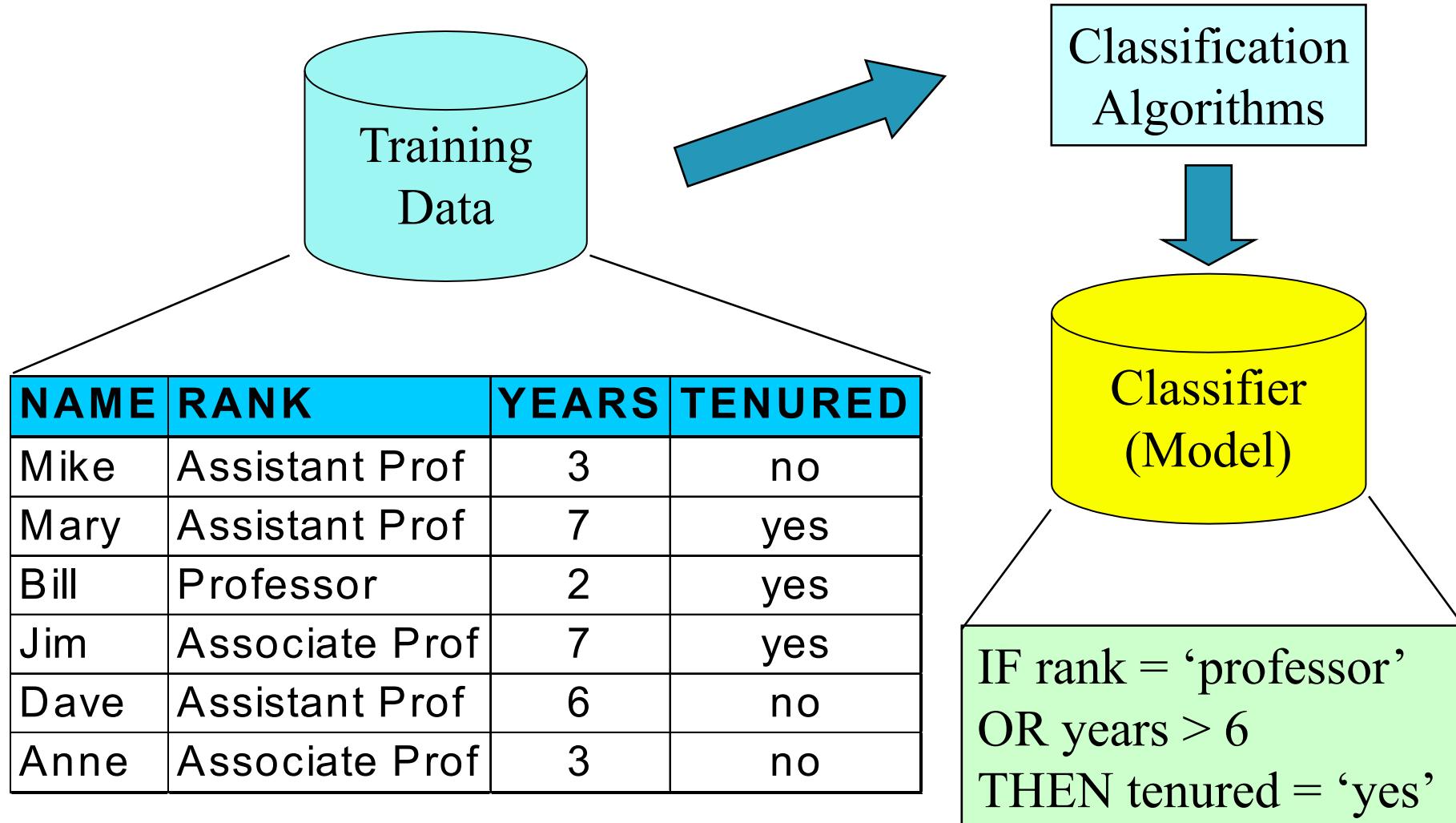
• Classification

- ทำนายประเภท หรือ คลาสลาเบล (class labels)
- สร้างโมเดลจำแนกประเภทด้วยชุดข้อมูลฝึกฝน (มีทั้งข้อมูลและ class labels) และนำโมเดลที่ได้ไปใช้ทำนายข้อมูลใหม่
- ตัวอย่าง: ทำนายประเภทของก้อนเนื้อว่าเป็นเนื้อร้ายหรือไม่เป็น, จำแนกประเภทของเว็บเพจ, อนุมัติเครดิต

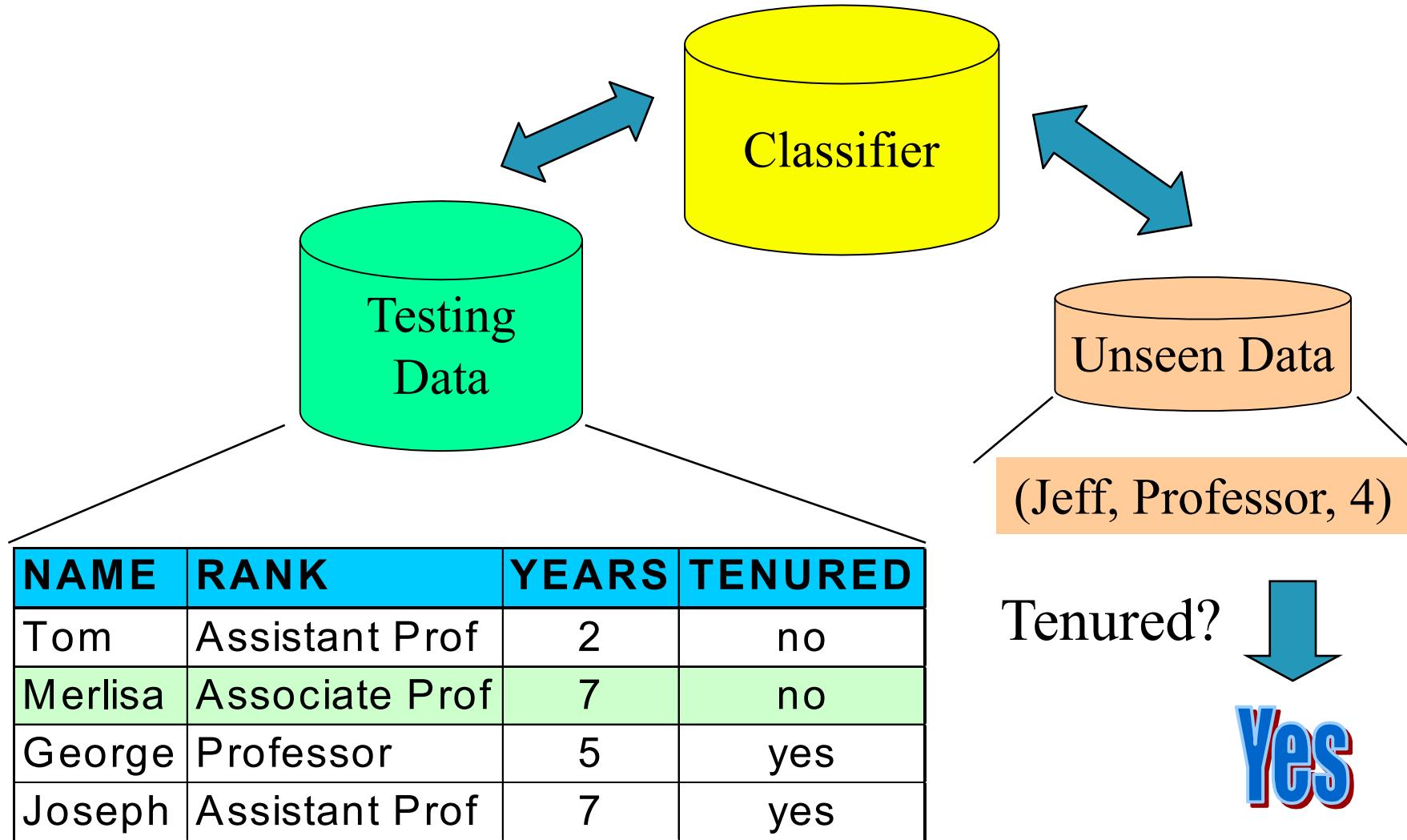
• Regression

- ทำนายค่าเป็นตัวเลข
- สร้างโมเดลทำนายค่าโดยใช้ชุดข้อมูลฝึกฝน (มีทั้งข้อมูลและค่าที่ต้องการทำนาย) และนำโมเดลที่ได้ไปใช้ทำนายค่าของข้อมูลใหม่
- ตัวอย่าง: ทำนายอัตราดอกเบี้ยการเช่าซื้อ, ทำนายราคาบ้าน, ทำนายอุณหภูมิของวันพรุ่งนี้

ขั้นตอนที่ 1 การสร้างโมเดล



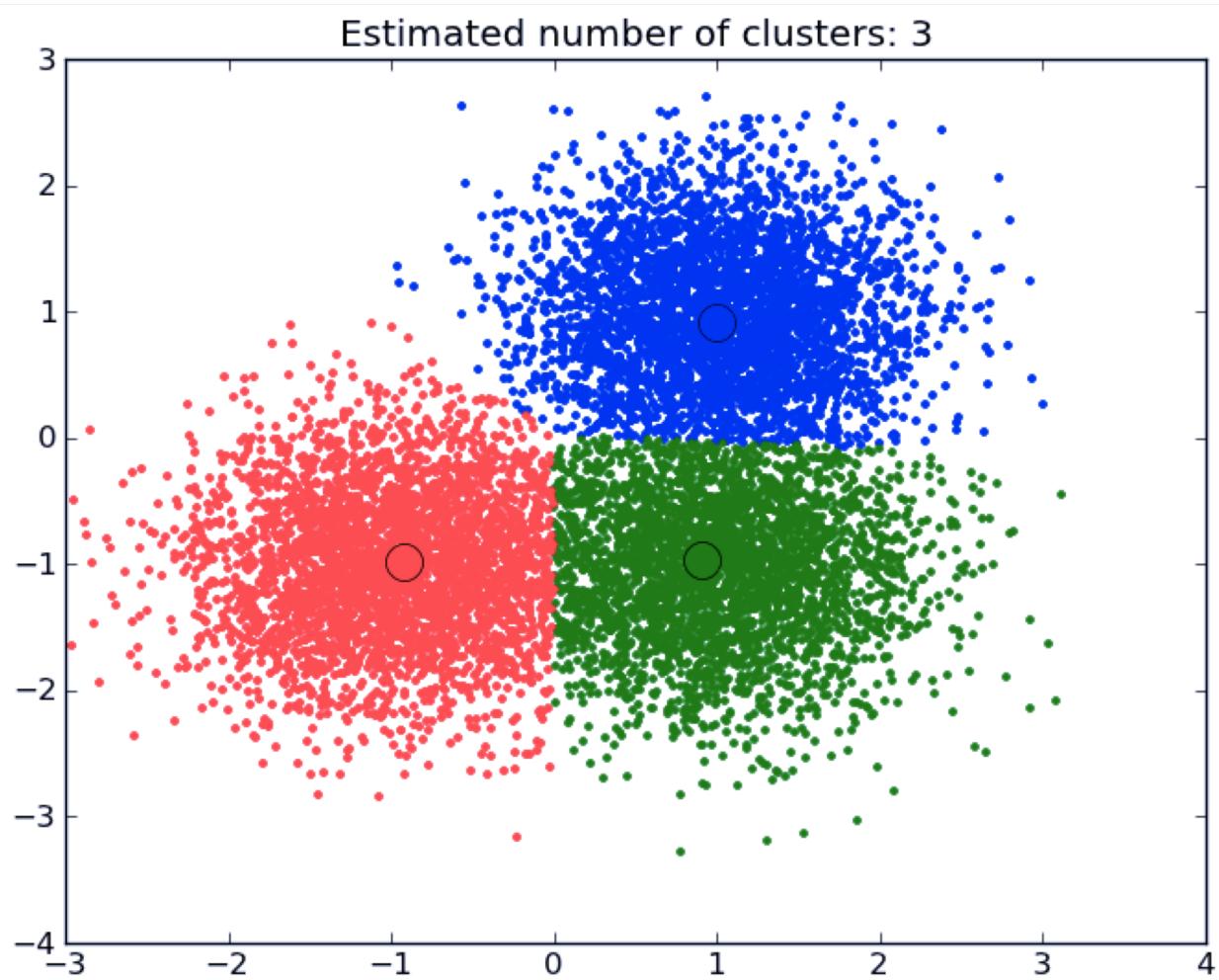
ขั้นตอนที่ 2 การนำโมเดลไปใช้งาน



โภมเดลการเรียนรู้ ที่ใช้กับ ปัญหาการจำแนกประเภท / การทำนายค่า

- Decision Tree (ต้นไม้มีการตัดสินใจ)
- Neural Network (เครือข่ายประสาทเทียม)
- Linear Regression (การวิเคราะห์การถดถอยเชิงเส้น)
- Logistic Regression (การวิเคราะห์การถดถอยแบบโลจิสติกส์)
-

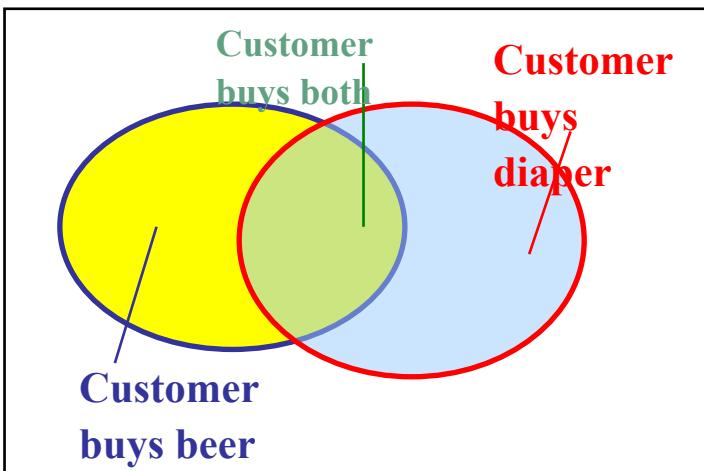
Clustering (Unsupervised Learning)



- ไม่ทราบ class labels
- ต้องการจัดกลุ่มจุดข้อมูลที่เหมือนหรือคล้ายกันไว้ในกลุ่มเดียวกัน
- ตัวอย่าง : customer segmentation, web community
- ไม่เดลการเรียนรู้: k-means, DBSCAN

Association Analysis

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



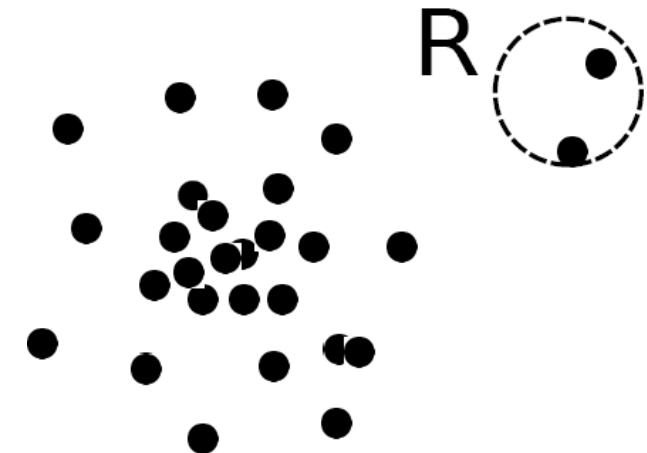
- หาสินค้าที่ถูกซื้อด้วยกันบ่อย ๆ (market basket analysis)

โดยใช้ข้อมูล sales transactions

- $Beer \rightarrow Diaper$ (60%, 100%)
- $Diaper \rightarrow Beer$ (60%, 75%)

Outlier Detection

- Outlier คือ จุดข้อมูลที่เบี่ยงเบนไปจากจุดข้อมูลปกติอย่างมีนัยสำคัญ
 - เช่น การใช้บัตรเครดิตที่ผิดปกติ, แต้มที่ไม่เคลื่อน จอร์แดน ทำได้ในการแข่งขัน NBA
- Outlier \neq Noise
 - Noise คือความผิดพลาดแบบสุ่มหรือความแปรปรวนของตัวแปรที่เราวัดค่า
 - Noise จะต้องถูกนำออกไปจากชุดข้อมูลก่อนที่เราจะทำการตรวจจับ Outlier
- ตัวอย่างการประยุกต์ใช้
 - การแบ่งกลุ่มลูกค้า
 - การวินิจฉัยโรค
 - การตรวจจับการฉ้อโกงบัตรเครดิต



อัลกอริทึมการเรียนรู้ใดต่อไปนี้ ไม่ใช่การเรียนรู้แบบมีผู้สอน (supervised learning)

- ก. Decision trees
- ข. k-Means
- ค. Support vector machines
- ง. Neural networks

อัลกอริทึมการเรียนรู้ใดต่อไปนี้ เป็นการเรียนรู้แบบไม่มีผู้สอน (unsupervised learning)

- ก. Hierarchical clustering
- ข. Random forest
- ค. Support vector machines
- ง. Linear regression

สรุป สิ่งที่ได้เรียนรู้ในวันนี้

- กระบวนการทางวิทยาศาสตร์ข้อมูล ได้แก่
 - ทำความเข้าใจธุรกิจ และข้อมูล, เตรียมข้อมูล, สร้างโมเดล, นำโมเดลไปใช้งาน, ความรู้และการกระทำ
- งานหลักทางวิทยาศาสตร์ข้อมูลมี 5 ประเภท ได้แก่
 - การจำแนกประเภท, การวิเคราะห์การถดถอย, การจัดกลุ่ม, การวิเคราะห์ความสัมพันธ์, การตรวจจับความผิดปกติ

หัวข้อถัดไป

- การสำรวจข้อมูล (Data Exploratory) ... การใช้สถิติพรรณนา (descriptive statistics) และ การทำให้เห็นภาพ (data visualization) เพื่อทำความเข้าใจข้อมูลเบื้องต้น