

คพ341 (1-66) วิทยาการข้อมูล

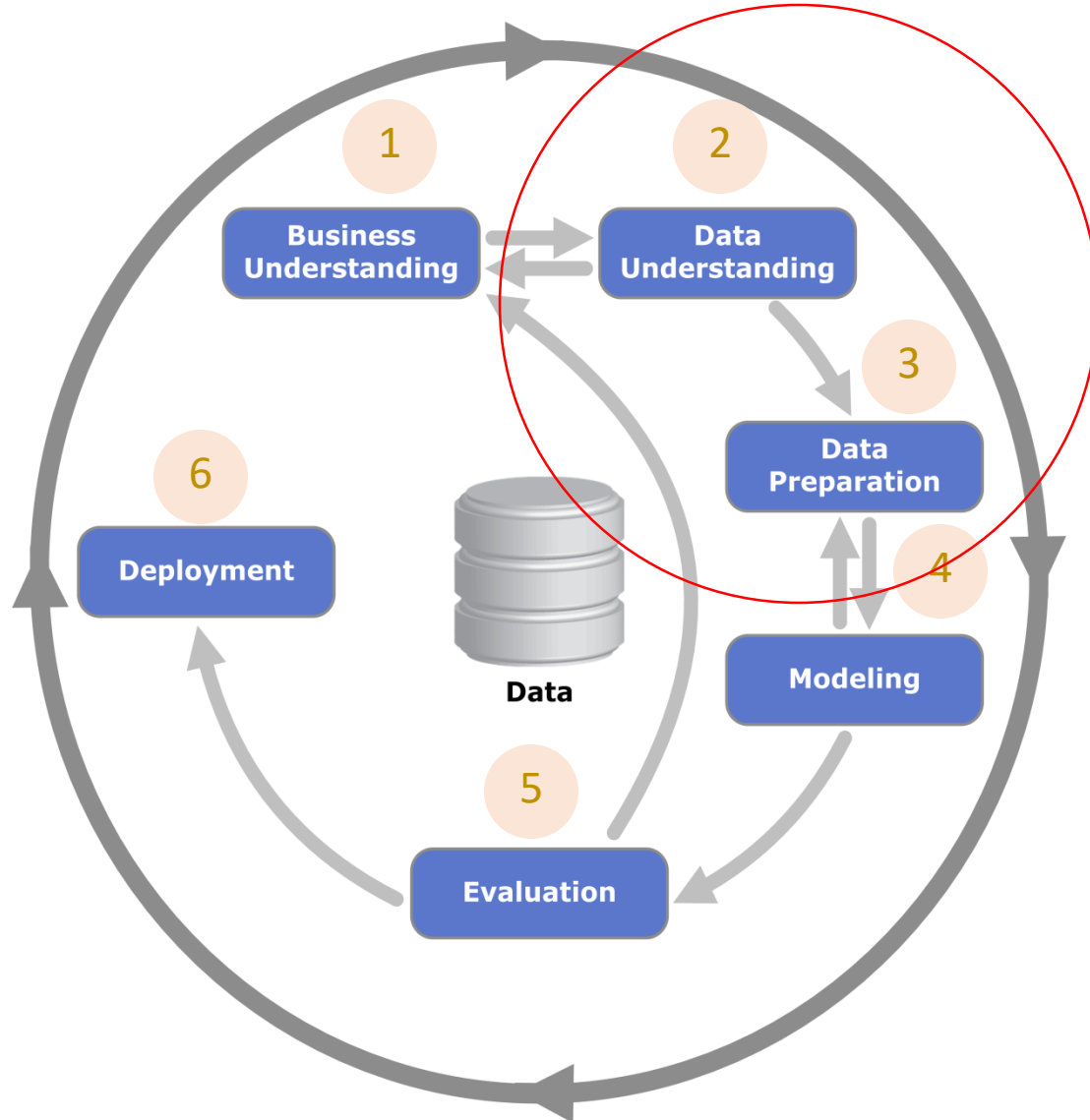
การทำความเข้าใจข้อมูล Data Understanding
การเตรียมข้อมูล Data Preparation

คาบ บรรยาย ที่ 3 วันที่ กรกฎาคม 2565

อ.ดร.พาสน์ ปราโมกษ์ชน

วิทยาการคอมพิวเตอร์ แม่โจ้

ขั้นตอนของ CRISP-DM



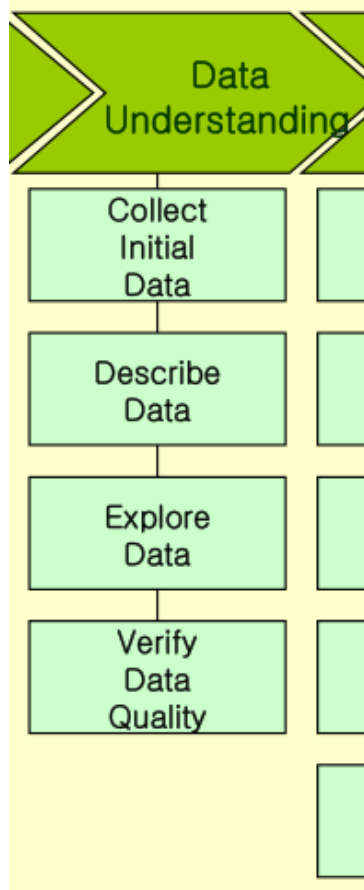
- ระเบียบวิธีการทำเหมืองข้อมูล (Data Mining Methodology)
- เป็นการเตรียมพิมพ์เขียวที่สมบูรณ์สำหรับทุกๆ งาน
- 6 phases Life cycle

CRISP-DM: 6 Phases

- Business/Research Understanding
 - ทำความเข้าใจในวัตถุประสงค์ของธุรกิจหรือปัญหางานวิจัยให้ชัดเจน
- Data Understanding
 - ทำความเข้าใจข้อมูลของธุรกิจหรือ ข้อมูลของปัญหาวิจัย รวมถึงระบุการเก็บข้อมูล
- Data Preparation
 - การจัดเตรียมข้อมูลเพื่อให้อยู่ในรูปที่เหมาะสมในการวิเคราะห์ ประมวลผล
- Modeling
 - การสร้างตัวแบบที่เหมาะสมกับวัตถุประสงค์ของธุรกิจ หรือ ปัญหางานวิจัย
- Evaluation
 - ประเมินตัวแบบเพื่อวัดประสิทธิภาพ
- Deployment
 - นำตัวแบบที่สร้างขึ้นไปใช้งานจริงเพื่อประเมินความสมบูรณ์โครงการ

มาตรฐาน CRISP-DM

2. Data Understanding



- การทำความเข้าใจข้อมูลทางธุรกิจ | การทำความเข้าใจข้อมูลงานวิจัย

2.1 เป็นจุดเริ่มต้นการเก็บรวบรวม (Collect) ข้อมูล

2.2 ทำความเข้าใจข้อมูลเพื่อสร้างความคุ้นเคยข้อมูล และสามารถอธิบาย (Describe) และ 2.3 ค้นหาความรู้ (Explore) ข้อมูลได้

- รวมถึงแง่กฎหมาย และ ทางเทคนิค

2.4 ประเมิน (Verify) คุณภาพ (Quality) ของข้อมูล

- มีปริมาณและมีรายละเอียดมากพอการวิเคราะห์หรือไม่?
- ข้อมูลมีความน่าเชื่อถือหรือไม่?

- เพื่อสร้าง **สมมติฐานการวิเคราะห์สารสนเทศ** ที่ซ่อนอยู่

ทำความเข้าใจข้อมูลที่ใช้ในการวิเคราะห์

- ขั้นที่ 1 ของการวิเคราะห์ข้อมูล นักวิเคราะห์ต้องรู้จักและเข้าใจ **ประเภทของข้อมูล (type of data)** ทั้ง **ตัวแปรต้น** และ **ตัวแปรตาม**
- ขั้นที่ 2 คุณลักษณะ **รูปแบบ (pattern)** ของข้อมูลข้อมูลว่ามีการกระจายของข้อมูลการแจกแจงแบบใด เช่น **การกระจายแบบปกติ (Normal Distribution)** หรือไม่
- ขั้นที่ 3 ข้อมูลที่สนใจ(ตัวแปรต้น และ ตัวแปรตาม) มี **ค่าของข้อมูล** แค่แบบเดียว สองแบบ หรือมากกว่า
- ขั้นที่ 4 การเลือกใช้ **เทคนิคทางการวิเคราะห์** ที่สอดคล้องกับข้อมูล โดยพิจารณาจากขั้นตอน 1-3
- ขั้นที่ 5 การสรุปผลของเทคนิคที่ใช้กับข้อมูล ว่ามี **นัยยะสำคัญทางสถิติ** หรือไม่

ตัวอย่างขั้นตอน Data Understanding จากกรณีการแบ่งกลุ่มลูกค้าเพื่อส่ง Promotion

- จากขั้นตอน **Business understanding** ในคาบบรรยายที่ผ่านมา
- บริษัทคิดว่า **ลูกค้าน่าจะมีสี่กลุ่ม (customer group)** แบ่งตามช่วงเวลาที่ยกซื้อสินค้าหลังเริ่มวางจำหน่าย
 - *Innovator* คือ ลูกค้าที่เป็นผู้สนใจใช้สินค้า IT gadget และซื้อทันทีหลังสินค้าวางจำหน่ายในสัปดาห์แรก
 - *Early Adopter* คือ ลูกค้าที่สนใจสินค้า gadget แต่ละจะซื้อหลังสัปดาห์แรกไปจนถึงสัปดาห์ที่ 3
 - *Early Majority* คือ ลูกค้าที่ไม่ได้สนใจสินค้า gadget มากนัก ซึ่งจะซื้อหลังสัปดาห์ที่ 3 ไปจนถึงเดือนที่ 2
 - *Late Majority* คือ ลูกค้าที่ไม่ใช่ผู้สนใจสินค้า IT gadget ที่จะซื้อหลังจากเดือนที่ 2 ไปเป็นอย่งน้อย
- ซึ่งจากการพิจารณาบริษัทเห็นว่าสามารถใช้ **ข้อมูลการสั่งซื้อสินค้า online ผ่านเว็บ ที่ลูกค้าแต่ละคนเคยใช้บริการมาเป็นประโยชน์ได้**

ข้อมูลคืออะไร

- ในขั้นตอนการทำ Data understanding จำเป็นต้องเข้าใจนิยามคำว่า “ข้อมูล”
- คือการเก็บข้อมูลลักษณะ (attribute, แอตทริบิวต์) ของ สิ่งที่กำลังสนใจ (object)
- ตัวอย่างของ object
 - อาจเรียกว่า ระเบียน (record) (ในงาน database) เช่น ลูกค้า ใบสั่งซื้อ
 - จุด Point, กรณี case, ตัวอย่าง sample (ในงาน สถิติ) เช่น ผู้กรอกสัมภาษณ์
 - วัตถุ Instance (ในงาน Pattern Recognition หรือ Machine Learning) เช่น รถยนต์ ต้นพืช มนุษย์
- ตัวอย่างของ attribute
 - สีของนัยต์ตา, อุณหภูมิ, ความชื้น, คะแนนรายวิชา เป็นต้น
 - attribute อาจเรียกหลายชื่อ เช่น
 - Variable หรือ Predictor ในสถิติ
 - field ใน Database
 - Feature ใน Machine Learning

Attribute Values (ค่าของข้อมูลแอตทริบิวต์)

- Attribute values are **numbers or symbols** assigned to an attribute

ค่าที่เป็นได้ของข้อมูล attribute อาจจะเป็นตัวเลข หรือ อักขระ ก็ได้

- Distinction between attributes and attribute values

ความแตกต่างระหว่าง attribute กับ attribute value

- Same attribute can be mapped to different attribute values

Attribute ที่เหมือนกัน อาจมี attri value ที่ต่างกัน

- Example: height can be measured in feet or meters

Attribute ความสูง ฝั่งชาติตะวันตกมักใช้ ฟุต แต่เอเชียมักใช้เป็น เมตร ซึ่ง ฟุตกับเมตรคือ attribute value หน่วยข้อมูล

- Different attributes can be mapped to the same set of values

Example: attribute ที่ต่างกัน อาจใช้ att value ที่เหมือนกัน

- Attribute values for ID and age are integers บัตรประชาชน เลข 13 หลัก กับ อายุ เลข 3 หลัก

- But properties of attribute values can be different

- บัตรประชาชนเลขนั้นจะไม่มีจำกัด แต่ อายุมีค่ามากที่สุด และน้อยสุดแน่ 135 และ 0

- ID has no limit but age has a maximum and minimum value

Attribute Values (ค่าของแอตทริบิวต์)

ตัวอย่าง Attribute Value ของข้อมูลเพื่อการแบ่งลูกค้าของบริษัทจำหน่ายมือถือ

- นักศึกษาบอกได้หรือไม่ว่า
 - Value ของ Gender (เพศ) ประกอบด้วยค่าใดบ้าง,
 - Value ของ Website activity ประกอบด้วยค่าใดบ้าง
.....,

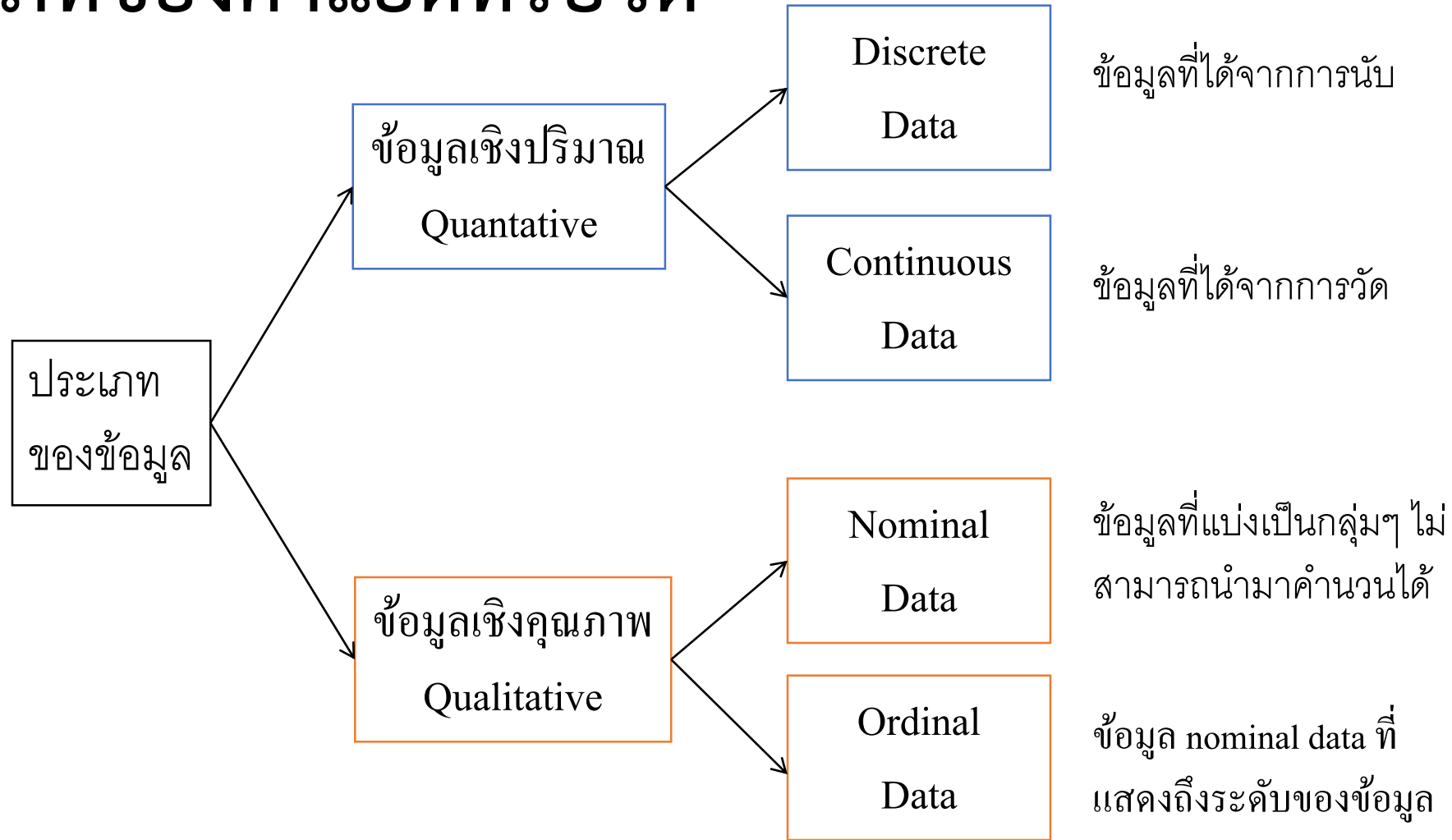
คำตอบผลลัพธ์

customer_ID	gender	age	website_activity	social_media_account	payment_method	customer_group
9123	M	58	rarly	no	bank transfer	Late Majority
4567	M	26	regular	no	bank transfer	Innovator
1254	F	30	rarly	yes	bank transfer	Early Adopter
3332	M	48	rarly	yes	website account	Early Adopter

- Value ของ Customer Group ประกอบด้วยค่าใดบ้าง
.....,,,

Types of Attribute value

ประเภทของค่าแอตทริบิวต์



Types of Attribute value

ประเภทของค่าแอตทริบิวต์

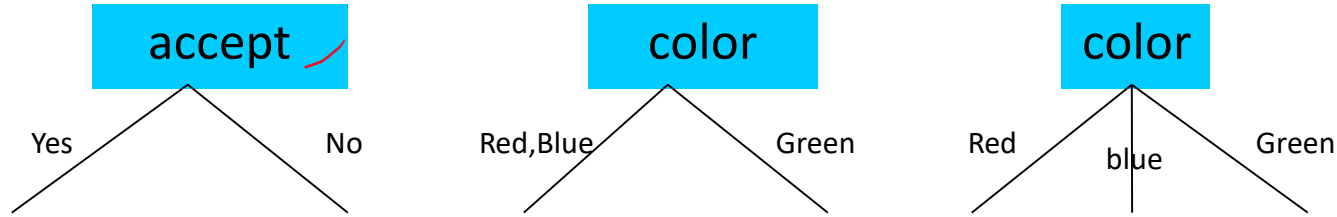
- There are different types of attributes
 - Discrete
 - Examples: number of customers, number of object in transaction, calendar date
 - Continuous
 - Examples: temperature, length, weight
 - Nominal
 - Examples: ID numbers, eye color, zip codes
 - Ordinal
 - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}

Types of Attributes value (เพิ่มเติม)

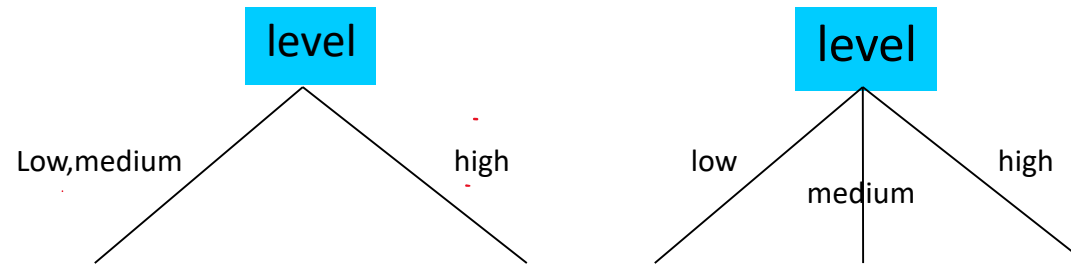
Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: { <i>male</i> , <i>female</i> }	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, { <i>good</i> , <i>better</i> , <i>best</i> }, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, <i>t</i> and <i>F</i> tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

ตัวอย่างของ ประเภทของแอตทริบิวต์ และค่าที่เป็นไปได้ของแอตทริบิวต์

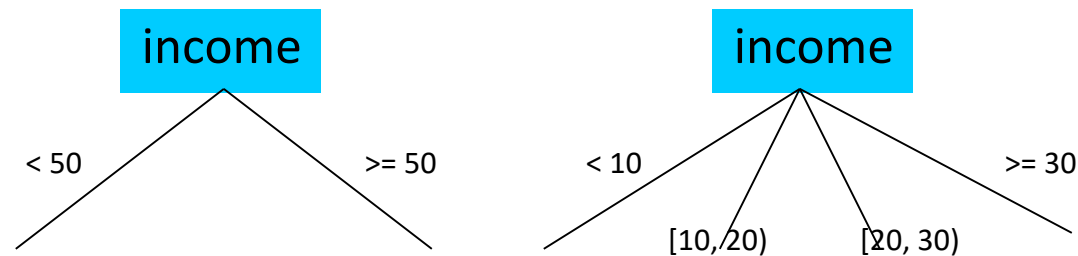
❖ **Nominal Attribute** ตัวแปรที่ไม่เป็นตัวเลข



❖ **Ordinal Attribute** ตัวแปรที่มีลักษณะเป็นระดับ



❖ **Continuous Attribute หรือ Interval Attribute** ตัวแปรที่มีลักษณะเป็นช่วง



ตัวอย่างการกำหนดประเภทของข้อมูลแอตทริบิวต์ในขั้นตอน Data Understanding

- กรณีแบ่งกลุ่มลูกค้าของบริษัท Smart Phone
- จะมีการเก็บกำหนด attribute ดังนี้

attribute	ความหมาย	ประเภทข้อมูล	หมายเหตุ
customer_id	รหัสลูกค้า	Numeric	เป็นหมายเลข
gender	เพศ	Nominal	
age	อายุ	Numeric	ไม่เกิน 100
website_activity	ความถี่การใช้งาน internet	Ordinal	มีระดับการใช้งาน
social_media_account	การเป็นสมาชิกเครือข่ายสังคมออนไลน์	Numeric	เป็นเลขจำนวนเต็ม
payment_method	การชำระเงิน	Nominal	
customer_group	ประเภทของลูกค้า	Nominal	คำตอบของการจัดกลุ่ม

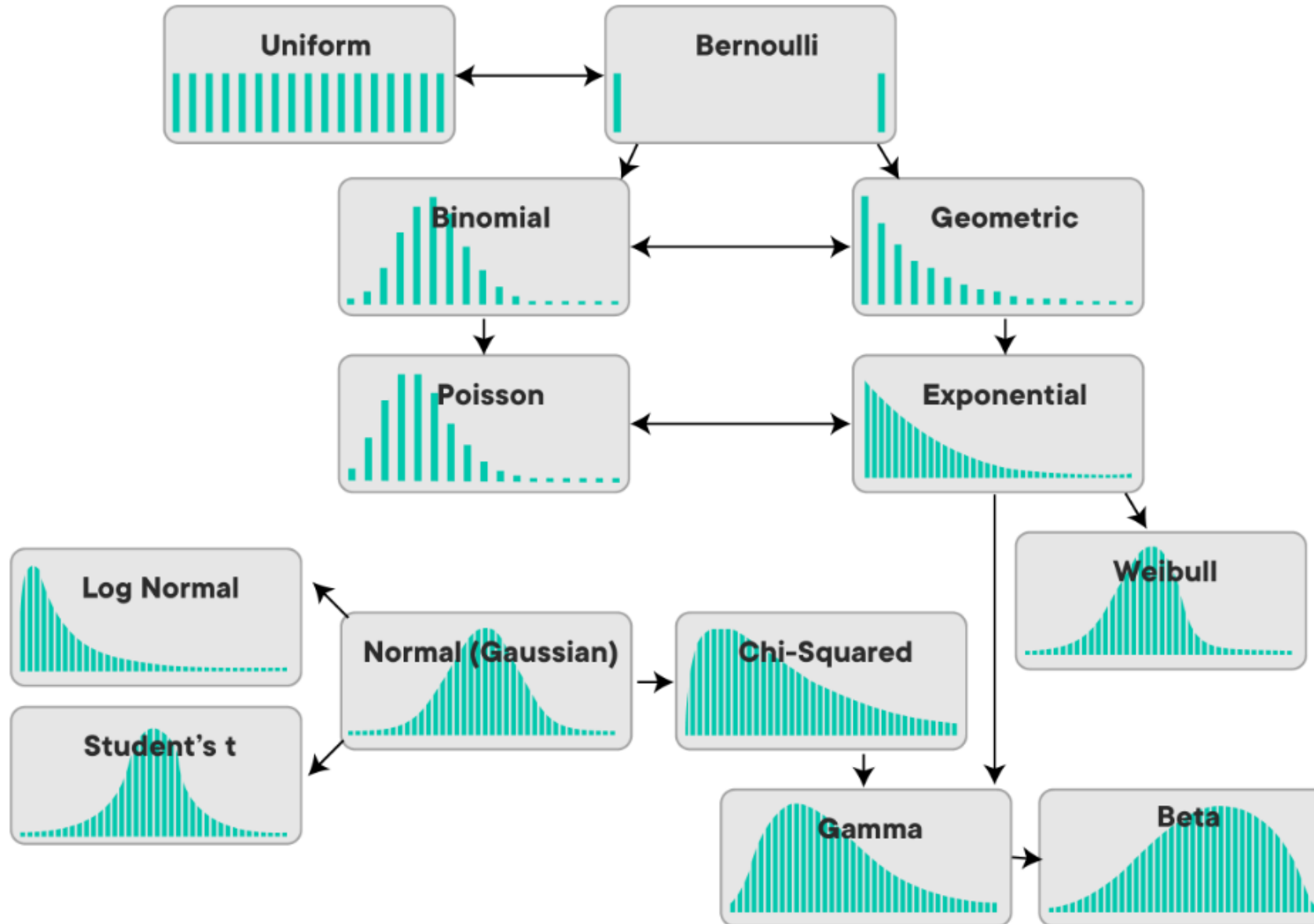
ตัวอย่างที่ 2 การกำหนดประเภทของข้อมูลแอตทริบิวต์ในขั้นตอน

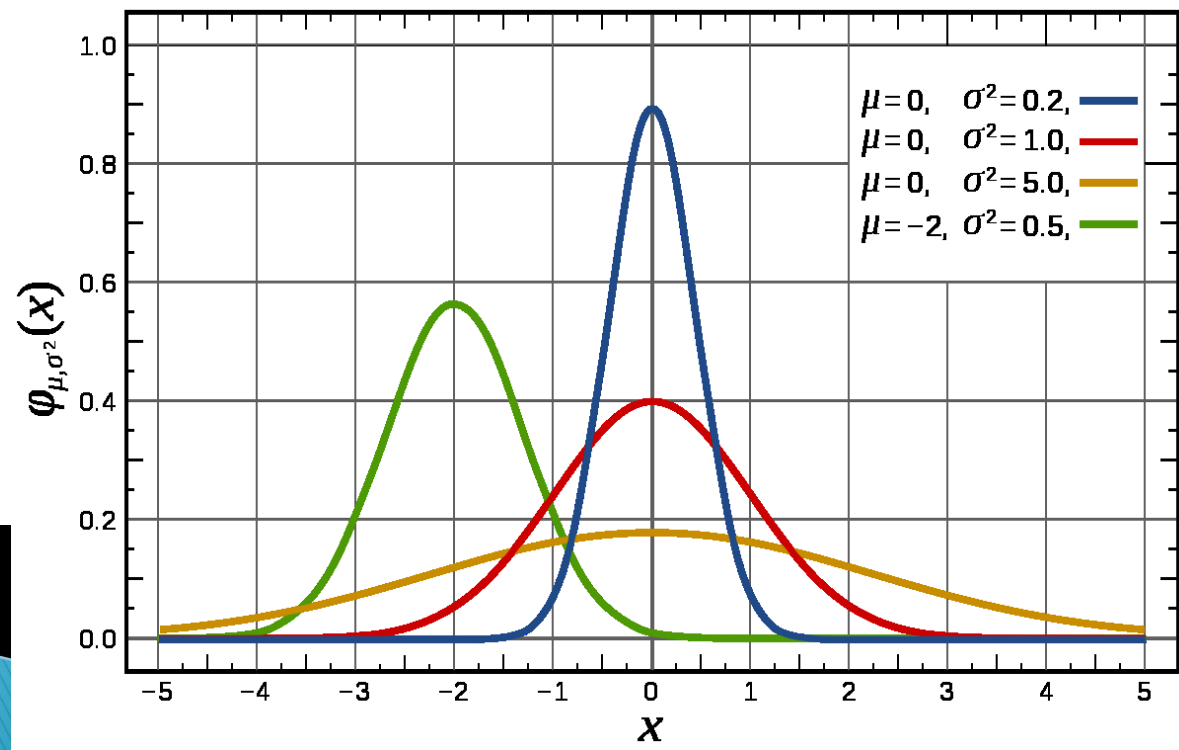
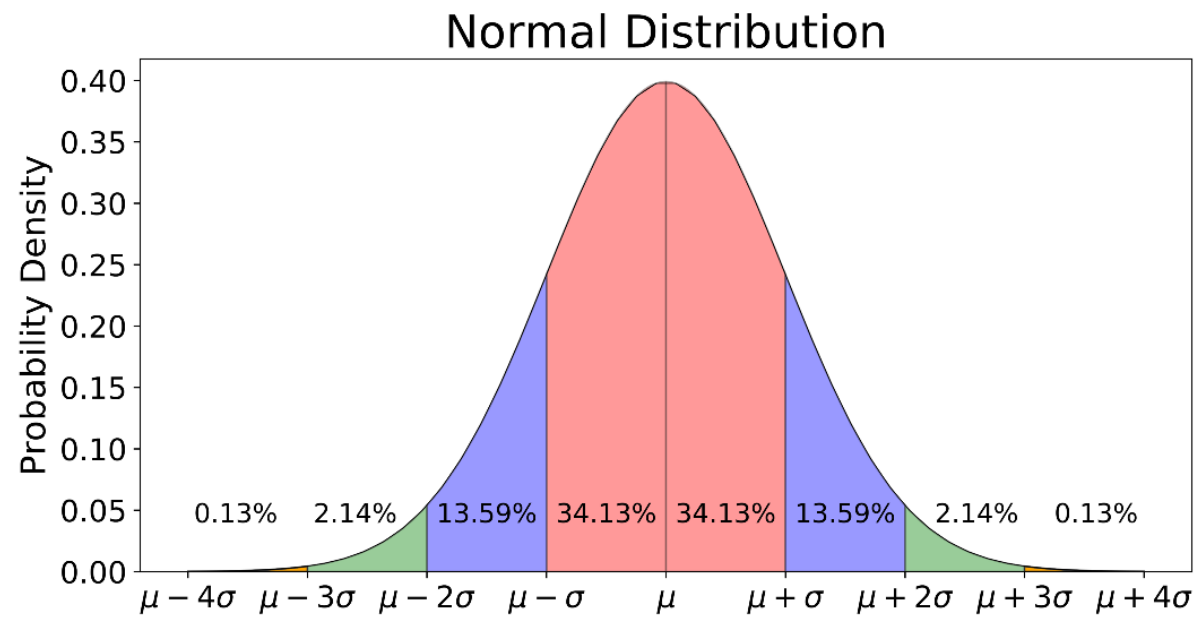
Data Understanding

- supermarket แห่งหนึ่งที่ทำ Business Understanding มีความเห็นว่า อยากออกโปรโมชั่นเพื่อเพิ่มยอดขาย โดยโปรโมชั่นที่วางแผนไว้จะเป็นการนำสินค้าที่ลูกค้าซื้อบ่อยๆ มาลดราคาร่วมกัน
- บริษัทมีแผนใช้ข้อมูลการซื้อสินค้าในใบเสร็จเพื่อเป็นข้อมูลการซื้อสินค้าร่วมกัน
- จากข้อมูลพื้นฐานบริษัทแบ่งประเภทสินค้าเป็นกลุ่มตามลักษณะสินค้า ซึ่งบริษัทมีความรู้เดิมที่คาดว่ามีสินค้าประเภทใดน่าจะสัมพันธ์กับสินค้าประเภทใด
- ดังนั้นในขั้นตอน Data Understanding จึงกำหนด attribute ดังนี้

attribute	ความหมาย	ประเภท	เพราะ
Reciept_id	หมายเลขใบเสร็จเงิน	numeric	หมายเลขเป็นตัวเลข
Member_id	หมายเลขสมาชิก	Norminal	หมายเลขเป็นตัวเลขแต่เป็น code ประจำตัวสมาชิก
Product_id	หมายเลขสินค้า	Norminal	หมายเลขสินค้าเป็นแค่ code เพื่อความสะดวกในการค้น
Product_amount	จำนวนสินค้าที่ซื้อ	numeric	เก็บจำนวนขึ้นที่ซื้อ
Discount	ส่วนลดราคา	numeric	เก็บจำนวนส่วนลดในการซื้อครั้งนั้น

Type of Probability Distributions





- ข้อมูลนักศึกษา

Attribute หรือ Feature

ไอดี (ID) แอตทริบิวต์แสดงหมายเลขของข้อมูล

แอตทริบิวต์ทั่วไป (attribute) ที่จะใช้ในการสร้าง
โมเดล หรือ เรียกว่าเป็น ฟีเจอร์ (feature) หรือ ตัว
แปรต้น (independent variable)

ตัวอย่าง
(example)
(instance)

ลำดับ	น้ำหนัก	ส่วนสูง	ขนาดเท้า (CM)	เพศ	ยี่ห้อสินค้าที่ชอบ
1	58	156	39	F	VANS
2	55	164	42	M	puma
3	47	162	40	F	converse
4	52	156	40	F	Nike
5	48	170	40	F	adidas

instance ที่ 1
instance ที่ 2
instance ที่ 3
instance ที่ 4
instance ที่ 5

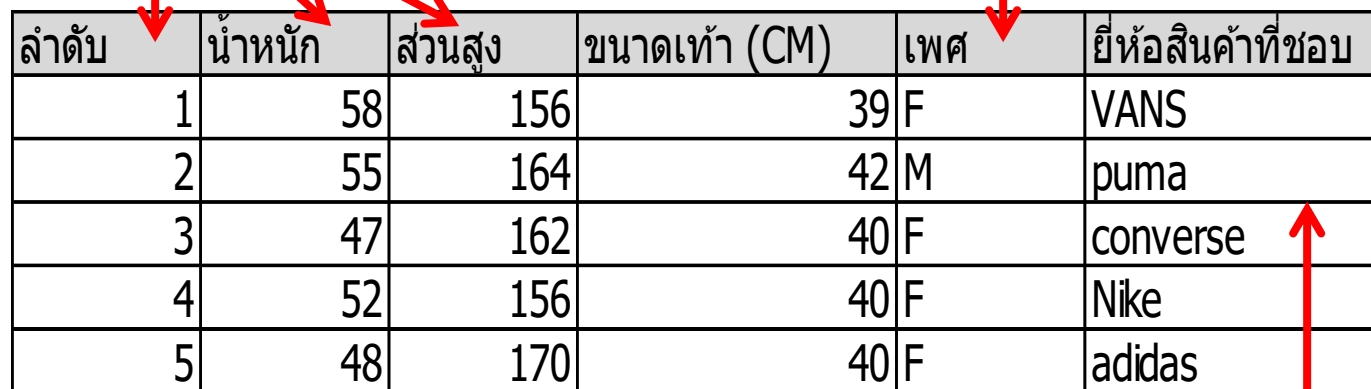
ระบบคาดเดายี่ห้อ
รองเท้ากีฬาจาก
นน ส่วนสูง ความยาวเท้า และ เพศ

ลาเบล (label) เป็นแอตทริบิวต์ชนิดพิเศษที่มักจะใช้
แสดงคำตอบของสิ่งที่เราต้องการจะสร้างโมเดลมา
ทำนาย หรือ เรียกว่า คลาส (class) หรือ ตัวแปรตาม
(dependent variable)

- ประเภทของข้อมูลในแต่ละแอตทริบิวต์
Attribute หรือ **Feature**

Binomial ข้อมูลประเภท category

(ข้อมูลที่ไม่ใช่ตัวเลข) มีค่าน้อยกว่า 2 กลุ่ม



ลำดับ	น้ำหนัก	ส่วนสูง	ขนาดเท้า (CM)	เพศ	ยี่ห้อสินค้าที่ชอบ
1	58	156	39	F	VANS
2	55	164	42	M	puma
3	47	162	40	F	converse
4	52	156	40	F	Nike
5	48	170	40	F	adidas

instance ที่ 1

instance ที่ 2

instance ที่ 3

instance ที่ 4

instance ที่ 5

อื่นๆ เช่น **Text** คือข้อมูลประเภทข้อความ

ข้อมูลประเภท **category**
(เป็นกลุ่มๆ ที่ไม่ใช่ตัวเลข) มีค่ามากกว่า 2 กลุ่ม
ขึ้นไป

CRISP-DM

3. Data Preparation

- ▶ เป็นขั้นตอนที่ทำการแปลงข้อมูลที่ได้รวบรวมมา (Raw data) ให้กลายเป็นข้อมูลที่สามารถนำไปวิเคราะห์ได้
- ▶ ปรับปรุงคุณภาพโดยรวมของข้อมูลก่อนการทำเหมือง



**Garbage in
Garbage out**
ถ้าป้อนขยะไร้ค่าเข้าไป
ก็ย่อมได้ขยะไร้ค่าออกมา

Data Preparation

ความจำเป็นที่ต้องมีการเตรียมข้อมูล

- ▶ เนื่องจากข้อมูลเบื้องต้น (Raw data, Primary Data) ในฐานข้อมูลในความเป็นจริงมีความสกปรก คือ
 - *ข้อมูลไม่สมบูรณ์ (Incomplete data)* เช่น ค่าของคุณลักษณะขาดหาย (Missing value) ขาดคุณลักษณะที่น่าสนใจ หรือ ทำ N/A ไว้ขาดรายละเอียดของข้อมูล
 - *ข้อมูลไม่สอดคล้อง (Inconsistent data)* เช่น ข้อมูลเดียวกัน แต่ตั้งชื่อต่างกัน หรือใช้ค่าแทนข้อมูลที่ต่างกัน **Human error**
 - *ข้อมูลฟุ่มเฟือยเกินไป (redundant data)* เช่น เก็บข้อมูล ความกว้าง ยาว และ ขนาดพื้นที่ตารางเมตร **Unnecessary data**
 - *ข้อมูลรบกวน (Noisy data)* เช่น ข้อมูลมีค่าผิดพลาด (Error) หรือมีค่าผิดปกติ (Outliers)
 - *ข้อมูลผิดปกติ (outlier data)* เช่น 1,2,3,4,5,6,7,8,9,10, 101 **ทำให้ data เกิดความแปรปรวนได้ง่าย**

ปัญหาบางประการของข้อมูล

นักศึกษาลองสังเกตและแยกความแตกต่างของปัญหาการเก็บข้อมูล

หมายเลขลูกค้า	รหัสไปรษณีย์	เพศ	รายได้	อายุ	สถานภาพสมรส	ค่าส่งสินค้า	ปีเกิด
1001	10048	ชาย	75000	M	M	5000	2520
1002	CM5361	หญิง	-40000	40	W	4000	2526
1003	90210	ชาย		45	S	7000	2521
1004	6269	ชายย	50000	0	S	1000	2566
10 05	55101	หญิง	99999	30	D	3000	2536

Noisy error

Inconsistent

Missing
value

Inconsistent

Inconsistent

อาจเป็น Single or Separated ก็ได้

Outlier

ขั้นตอนย่อยใน Data Preparation

- เพื่อให้ขั้นตอนการสร้างตัวแบบเหมืองข้อมูลได้ตัวแบบที่ดี ต้องขึ้นอยู่กับคุณภาพของข้อมูลที่ใช้วิเคราะห์ ซึ่งขึ้นกับ

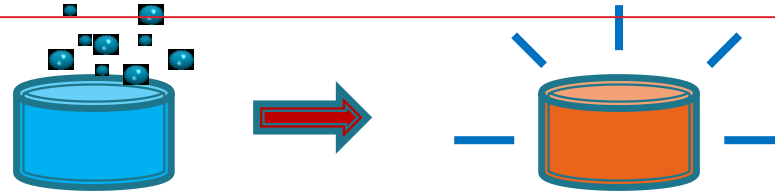
เนื้อหาในบทเรียนนี้เน้นที่

- Data Cleaning คือ การทำความสะอาดข้อมูล จัดการข้อมูลรบกวน
- Data Transformation คือ การแปลงข้อมูล
- Data Reduction คือ การลดรูปข้อมูล
- Data Integration คือ การผสานข้อมูล

เนื้อหาในบทเรียนนี้เน้นที่

เทคนิคการเตรียมข้อมูล

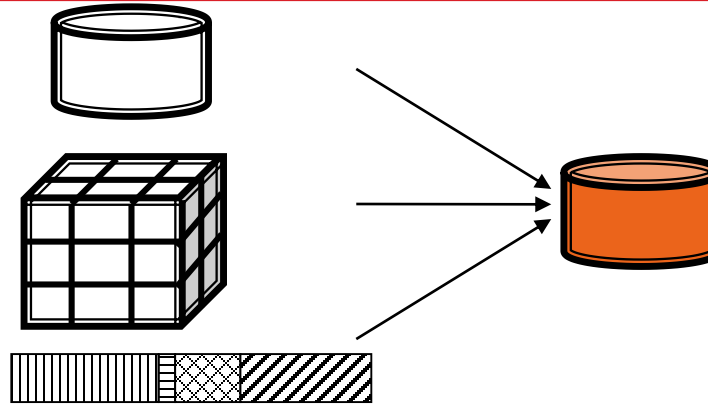
Data cleaning



Data transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data integration



Data reduction

transaction

attribute					
A1	A2	A3	A126	

transaction	attribute			
	A1	A2	A115
	T1			
	T2			
	...			
T2000				

Data Cleansing การทำความสะอาดข้อมูล

เนื้อหาในบทเรียนนี้เน้นที่

- แก้ไขข้อมูลที่ผิดพลาดหรือขาดหาย (missing values)
- ขจัดข้อมูลที่ผิดรูปแบบ (Outlier)
- ลบข้อมูลที่ซ้ำซ้อน (redundancy reduction)

การแก้ไขข้อมูลที่ขาดหาย (Handling Missing Value)



หมายเลขลูกค้า	รหัสไปรษณีย์	เพศ	รายได้	อายุ	สถานภาพสมรส	คำสั่งสินค้า	ปีเกิด
1001	10048	ชาย	75000	M	M	5000	2520
1002	CM5361	หญิง	40000	40	W	4000	2526
1003	90210	ชาย		45	S	7000	2521
1004	6269	ชาย	50000	0	S	1000	2566
10 05	55101	หญิง	99999	30	D	3000	2536

Missing
value

การแก้ไขข้อมูลที่ขาดหาย (Handling Missing Value)

1. Ignore the tuple

- ตัดทิ้งรายการที่มีข้อมูลสูญหาย
- นิยมใช้กับการจำแนกประเภท (Classification) ในกรณีที่ค่าคุณลักษณะขาดหายไปเป็นจำนวนมาก

2. Use a global constant to fill in the missing value

- เติมค่าคุณลักษณะของข้อมูลที่ขาดหายทุกค่า ด้วยค่าคงที่ค่าหนึ่ง เช่น ไม่รู้ค่า หรือ unknown หรือ N/A หรือ 0

3. Fill in the missing value manually

- เติมค่าที่ขาดหายด้วยมือ วิธีนี้ไม่เหมาะสมกรณีที่ชุดข้อมูลมีขนาดใหญ่ และมีข้อมูลขาดหายจำนวนมาก

4. Use the attribute mean to fill in the missing value

- ใช้ค่าเฉลี่ยของคุณลักษณะ เติมค่าข้อมูลที่ขาดหาย เช่น ถ้าทราบว่าลูกค้าที่รายได้เฉลี่ยเดือนละ 12000 บาท จะใช้ค่านี้แทนค่ารายได้ของลูกค้าที่ขาดหาย

5. Use the attribute mean for all samples belonging to the same class as the given tuple

- ใช้ค่าเฉลี่ยคุณลักษณะของตัวอย่างที่จัดอยู่ในประเภทเดียวกัน เพื่อเติมค่าข้อมูลที่ขาดหาย เช่น เติมค่ารายได้ของลูกค้าที่ขาดหาย ด้วยค่าเฉลี่ยของลูกค้าที่อยู่ในกลุ่มอาชีพเดียวกัน

6 Use the most propable value to fill in the missing value

- คล้ายกับวิธีที่ 4 และ 5 ใช้วิธีการประมาณค่าทางสถิติที่ซับซ้อนและมีความแม่นยำมากขึ้นเพื่อประมาณค่าที่สูญหายไป

ตัวอย่าง การแก้ไขข้อมูลที่ขาดหาย (Handling Missing Value)

1. Ignore the tuple

หมายเลขลูกค้า	รหัสไปรษณีย์	เพศ	รายได้	อายุ	สถานภาพสมรส	คำสั่งสินค้า	ปีเกิด
1001	10048	ชาย	75000	M	M	5000	2520
1002	CM5361	หญิง	-40000	40	W	4000	2526
1003	90210	ชาย		45	S	7000	2521
1004	6269	ชายย	50000	0	S	1000	2566
10 05	55101	หญิง	99999	30	D	3000	2536

- ตัดทิ้งรายการที่มีข้อมูลสูญหาย
- นิยมใช้กับการจำแนกประเภท (Classification) ในกรณีที่ค่าคุณลักษณะขาดหายไปเป็นจำนวนมาก



หมายเลขลูกค้า	รหัสไปรษณีย์	เพศ	รายได้	อายุ	สถานภาพสมรส	คำสั่งสินค้า	ปีเกิด
1001	10048	ชาย	75000	M	M	5000	2520
1002	CM5361	หญิง	-40000	40	W	4000	2526
1004	6269	ชายย	50000	0	S	1000	2566
10 05	55101	หญิง	99999	30	D	3000	2536

ตัวอย่าง การแก้ไขข้อมูลที่ขาดหาย (Handling Missing Value)

2. Use a global constant to fill in the missing value

หมายเลขลูกค้า	รหัสไปรษณีย์	เพศ	รายได้	อายุ	สถานภาพสมรส	ปริมาณธุรกิจ
1001	10048	ชาย	75000	M	M	5000
1002	CM5361	หญิง	-40000	40	W	4000
1003	90210	ชาย	NA	45	S	7000
1004	6269	ชาย	50000	0	S	1000
1005	55101	หญิง	99999	30	D	3000

3. Fill in the missing value manually

หมายเลขลูกค้า	รหัสไปรษณีย์	เพศ	รายได้	อายุ	สถานภาพสมรส	ปริมาณธุรกิจ
1001	10048	ชาย	75000	M	M	5000
1002	CM5361	หญิง	-40000	40	W	4000
1003	90210	ชาย	0	45	S	7000
1004	6269	ชาย	50000	0	S	1000
1005	55101	หญิง	99999	30	D	3000

ตัวอย่าง การแก้ไขข้อมูลที่ขาดหาย (Handling Missing Value)

4. Use the attribute mean to fill in the missing value

หมายเลขลูกค้า	รหัสไปรษณีย์	เพศ	รายได้	อายุ	สถานภาพสมรส	ปริมาณธุรกิจ
1001	10048	ชาย	75000	M	M	5000
1002	CM5361	หญิง	-40000	40	W	4000
1003	90210	ชาย	_____	45	S	7000
1004	6269	ชาย	50000	0	S	1000
1005	55101	หญิง	99999	30	D	3000

ทดลองแทนค่า
ค่าเฉลี่ย (mean) ของรายได้ของลูกค้า 4 คนที่เหลือ = —
= _____

การแก้ไขข้อมูลที่ขาดหาย (Handling Missing Value)

5. Use the attribute mean for all samples belonging to the same class as the given tuple

หมายเลขลูกค้า	รหัสไปรษณีย์	เพศ	รายได้	อายุ	สถานภาพสมรส	ปริมาณธุรกิจ
1001	10048	ชาย	75000	M	M	5000
1002	CM5361	หญิง	-40000	40	W	4000
1003	90210	ชาย	_____	45	S	7000
1004	6269	ชาย	50000	0	S	1000
1005	55101	หญิง	99999	30	D	3000

ทดลองแทนค่า โดยแบ่งกลุ่มจาก เพศ

ค่าเฉลี่ย (mean) ของรายได้ของลูกค้าเพศเดียวกันกับลูกค้า (1003) ที่ต้องการ
หาค่ารายได้ คือเพศชาย คือ ลูกค้าหมายเลข 1001 และ 1004 4 คนที่เหลือ

=

= _____

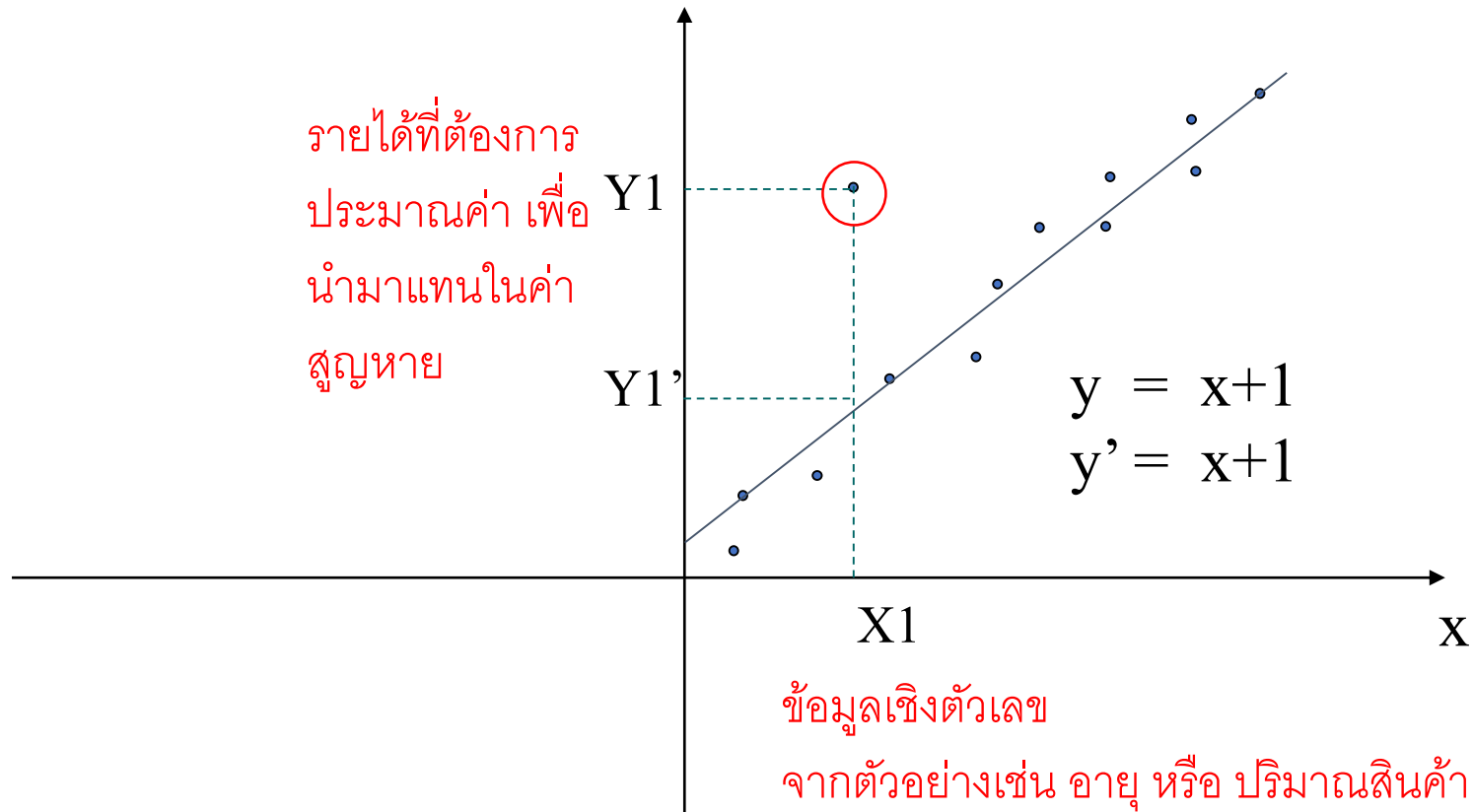
6 Use the most propable value to fill in the missing value

- ใช้ค่าที่เป็นไปได้มากที่สุด เติมแทนค่าข้อมูลที่ขาดหาย เช่น ค่าที่ได้จากสมการความถดถอย (Regression) ค่าที่ได้จากการอนุมาน โดยใช้สูตรของเบย์ (Bayesian formula) หรือ ค่าเฉลี่ยของจุดกึ่งกลางกลุ่ม (Mean of Centroid) หรือต้นไม้ตัดสินใจ (Decision tree) เช่น
ใช้ข้อมูลลูกค้า มาสร้างต้นไม้ตัดสินใจ เพื่อทำนายรายได้ของลูกค้า แล้วนำไปแทนค่าที่ขาดหาย
- นั่นคือการใช้การวิเคราะห์ข้อมูลสร้างปัญญาประดิษฐ์ (AI) เพื่อคาดเดาค่าของข้อมูลที่สูญหายและเติมกลับเข้าไปในชุดข้อมูล
- วิธีนี้นิยมกันแพร่หลาย เนื่องจากทำนายค่าข้อมูลที่ขาดหาย โดยพิจารณาจากค่าของข้อมูลชุดปัจจุบัน และความสัมพันธ์ระหว่างคุณลักษณะในชุดข้อมูล

6 Use the most propable value to fill in the missing value

ตัวอย่าง **Regression**

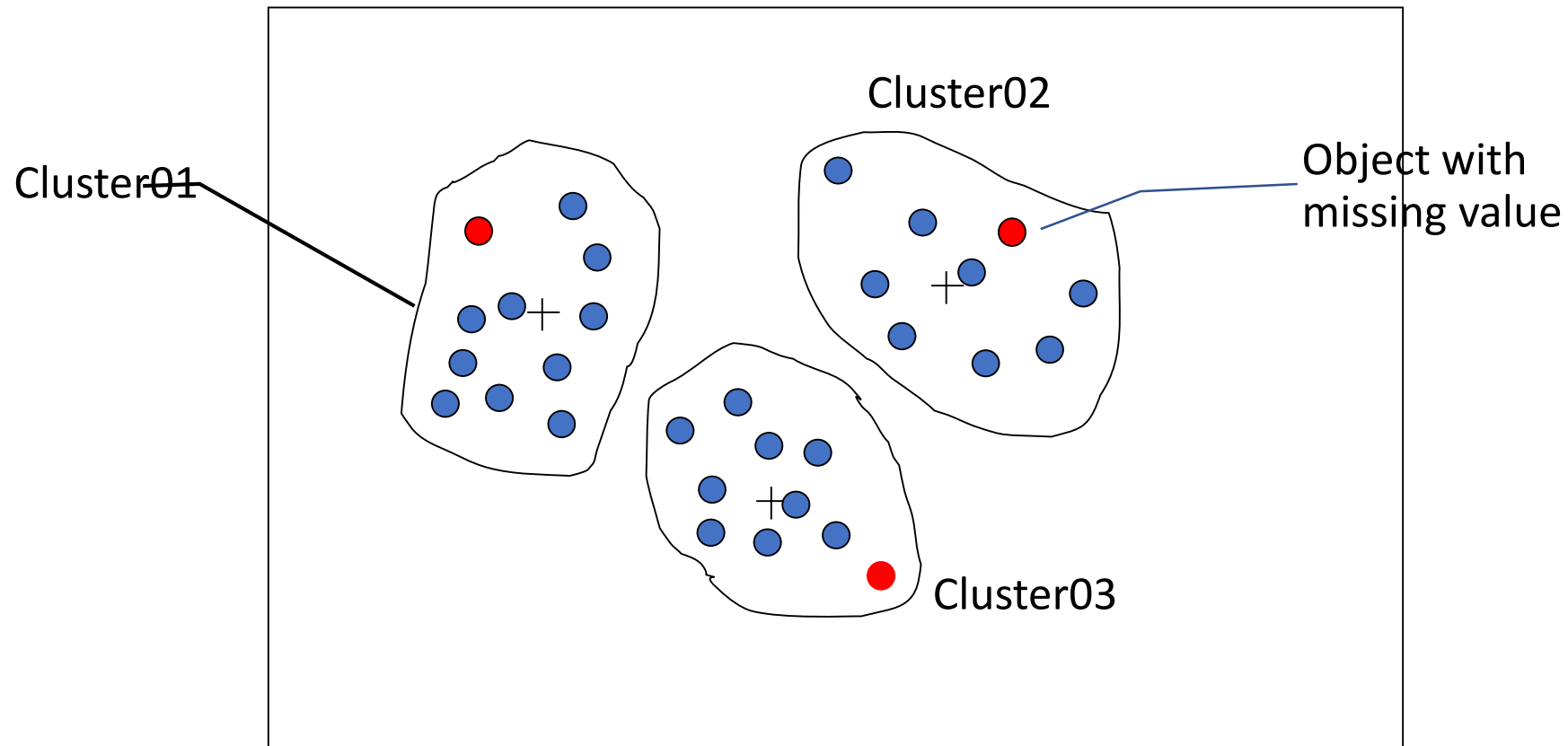
- วิธีความถดถอย ใช้การทำนายค่าของตัวแปรจากสมการความถดถอยที่หาได้ ด้วยวิธีความผิดพลาดน้อยที่สุด (Least-square error) จากชุดตัวอย่างตัวแปร



6 Use the most propable value to fill in the missing value

ตัวอย่าง Clustering

- คล้ายกันกับวิธีการที่ 5 แต่นักวิเคราะห์ใช้การวิเคราะห์การจัดกลุ่มแทนที่จะใช้กลุ่มจากแอตทริบิวต์เริ่มต้น จะช่วยตรวจหาค่าเฉลี่ยของแอตทริบิวต์ที่เป็นตัวแทนของสมาชิกในกลุ่มได้แม่นยำมากยิ่งขึ้น





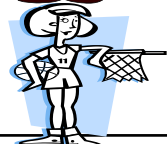



Data Transformation (การแปลงข้อมูล)

- การแปลงข้อมูลที่พบบ่อยในการทำเหมืองข้อมูลคือ การทำนอร์มอลไลซ์ (Normalization) โดยแปลงค่าข้อมูลให้อยู่ในช่วงสั้นๆ ที่อัลกอริทึมการทำเหมืองข้อมูลสามารถนำไปใช้ประมวลผลได้
- วิธีการทำนอร์มอลไลซ์ข้อมูลได้แก่
 - normalization by decimal scaling
 - min-max normalization
 - z-score normalization

ตัวอย่างปัญหาการใช้งานเหมืองข้อมูลที่ต้องมีการแปลงข้อมูลก่อนสร้างตัวแบบ

ให้นักศึกษาสังเกต และ อภิปรายว่าถ้าต้องนำข้อมูลนี้ไปวิเคราะห์ด้วย model ทางคณิตศาสตร์ attribute ใด น่าจะมีอิทธิพลต่อมา model มากกว่า attribute อื่นๆ

Customer	Age	Income	No. credit cards
John 	35	35,000	3
Rachel 	22	50,000	2
Hannah 	63	200,000	1
Tom 	59	170,000	1
Nellie 	25	40,000	4
David 	37	50,000	2

Data Transformation แบบ Decimal scaling

- เป็นการแปลงค่าข้อมูลเดิมให้เป็นเลขทศนิยม ตำแหน่งทศนิยมกำหนดโดยค่าสัมบูรณ์ที่มีค่ามากที่สุด

$$v' = \frac{v}{10^j}$$

- ตัวอย่างเช่น

ค่าที่เป็นไปได้ของคุณลักษณะ A อยู่ในช่วงระหว่าง -986 ถึง 917 จะได้ว่าค่าสัมบูรณ์ที่มากที่สุดคือ $|-986| = 986$ ดังนั้นเราจะหารข้อมูลแต่ละค่าด้วย 1000 โดยที่ $j=3$ ผลลัพธ์คือค่า -986 จะถูกแปลงเป็น -0.986

$$\frac{-986}{10^3} = -0.986$$

Min-Max Normalization

คะแนนปกติมาตรฐานน้อยที่สุด-มากที่สุด

- การแปลงข้อมูลให้อยู่ในช่วงใหม่ $[new_min_A - new_max_A]$

$$v' = \frac{v - min_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A$$

ตัวอย่าง เช่น

ข้อมูลคอลัมน์รายได้ (income) มีค่าต่ำสุด 12,000 บาท (min) และมีค่ามากที่สุด 98,000 บาท (max) ซึ่งข้อมูลที่ต้องการแปลงคือ 73,600 บาท ต้องการแปลงให้ข้อมูลนี้อยู่ในช่วงใหม่ คือ $[0 - 1]$ เพราะฉะนั้น 73,600 บาท จะมีค่าใหม่เป็นเท่าใด

$$\frac{73,600 - 12,000}{98,000 - 12,000} (1.0 - 0) + 0 = 0.716$$

Customer	Age	Income	No. credit cards
John	35	35,000	3
Rachel	22	50,000	2
Hannah	63	200,000	1
Tom	59	170,000	1
Nellie	25	40,000	4
David	37	50,000	2

Z-Score คะแนนมาตรฐาน

- เป็นการปรับการกระจายของข้อมูลให้มีค่าเท่ากับ 0 และค่าเบี่ยงเบนมาตรฐานเท่ากับ 1

- หาได้จากสูตร

$$v' = \frac{v - mean_A}{stand_dev_A}$$

ตัวอย่างเช่น

ข้อมูลคอลัมน์รายได้ (income) มีค่าเฉลี่ย 54,000 บาท (mean) และมีค่าเบี่ยงเบนมาตรฐาน 16,000 บาท (stand_dev) ต้องการแปลงค่ารายได้ 73,600 บาท เป็นค่าใหม่ตามวิธี Z-Score ผลลัพธ์จะเป็นทั้ง + และ -

$$\frac{73,600 - 54,000}{16,000} = 1.225$$

Customer	Age	Income	No. credit cards
John	35	35,000	3
Rachel	22	50,000	2
Hannah	63	200,000	1
Tom	59	170,000	1
Nellie	25	40,000	4
David	37	50,000	2

การบ้านมี 3 ข้อ Homework #1

- 1) จากตัวอย่างหน้า 22 ถ้ากรณีมี missing value ที่หมายเลขลูกค้า 1003 ที่ attribute ปริมาณธุรกิจ จงคำนวณค่าที่นำไปเติม พร้อมอธิบาย

หมายเลขลูกค้า	รหัสไปรษณีย์	เพศ	รายได้	อายุ	สถานภาพสมรส	ปริมาณธุรกิจ
1001	10048	ชาย	75000	M	M	5500
1002	CM5361	หญิง	-40000	40	W	4800
1003	90210	ชาย	46249.75	45	S	???
1004	6269	ชาย	50000	0	S	1300
10 05	55101	หญิง	99999	30	D	17000

- ค่า missing value ด้วยวิธี Use the attribute mean to fill in the missing value
- ค่า missing value ด้วยวิธี Use the attribute mean for all samples belonging to the same class as the given tuple ใช้ เพศ เป็น class

Homework #2

- 2) โรงพยาบาลแห่งหนึ่งทำการเก็บข้อมูล อายุและ ปริมาณไขมันในร่างกายของคนไข้โดยสุ่มข้อมูลมา 9 คนดังนี้

อายุ	23	23	27	27	39	41	47	49	50
ไขมัน	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

- 2.1 จงแปลงข้อมูล อายุ ด้วยวิธี Min-Max normalization ต้องการแปลงให้ข้อมูลนี้อยู่ในช่วงใหม่ คือ อายุ อยู่ในช่วง $[1,5]$ และเขียนอธิบายขั้นตอนการคำนวณ
- 2.2 จงแปลงข้อมูล ไขมัน ด้วยวิธี Z-score และเขียนอธิบายขั้นตอนการคำนวณ

ตัวอย่างการคำนวณการบ้านข้อที่ #2

- ย้ําว่ํา ข้อมูลตัวอย่างและช่วงตัวเลขนี้สมมติขึ้น
- 2.1 ถ้าต้องการแปลงอายุของคนใช้ 60 ให้อยู่ในช่วง $[0, 9]$
- แสดงว่า
$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$
 - ค่าต่ำสุดใหม่ $\text{New_min} = 0$
 - ค่าสูงสุดใหม่ $\text{New_max} = 9$
 - สมมติว่า ค่าต่ำสุดเดิม $\min = 10$ ค่าสูงสุดเดิม $\max = 40$
 - ค่าอายุของคนใช้คนที่ 1 จากเลขอายุ 23 จะแปลงไปเป็น
ตัวเลข 3.9 ซึ่งอยู่ในช่วง $[0, 9]$
$$\begin{aligned} &= \frac{30 - 10}{40 - 10} * (9 - 0) + 0 \\ &= 6 \end{aligned}$$

ตัวอย่างการคำนวณการบ้านข้อที่ 2

- ย้่าว่า ข้อมูลตัวอย่างนี้สมมติขึ้น

- 2.2 ถ้าต้องการแปลงปริมาณไขมันของคนไข้ 9.5 โดยวิธี z score

- สมมติว่า

$$v' = \frac{v - mean_A}{stand_dev_A}$$







- ค่าเฉลี่ย mean = 20 ค่าความแปรปรวน stand_dev = 5

- ค่าไขมันของคนไข้คนที่ 1 จากเลข 9.5 จะแปลงไปด้วยวิธี z score ได้เท่ากับ -2.1

$$\begin{aligned} &= \frac{9.5 - 20}{5} \\ &= -2.1 \end{aligned}$$

Homework#3

ให้นักศึกษาแปลงข้อมูล อายุ (age) ของ ... ให้อยู่ในช่วงตัวเลข [0 - 9]
และแปลงข้อมูล อายุ (age) ของ ด้วยวิธี Z-score

Customer	Age	Income	No. credit cards
John 	35	35,000	3
Rachel 	22	50,000	2
Hannah 	63	200,000	1
Tom 	59	170,000	1
Nellie 	25	40,000	4
David 	37	50,000	2

END