CS341

Data Science:

คาบ บรรยาย ที่ 2

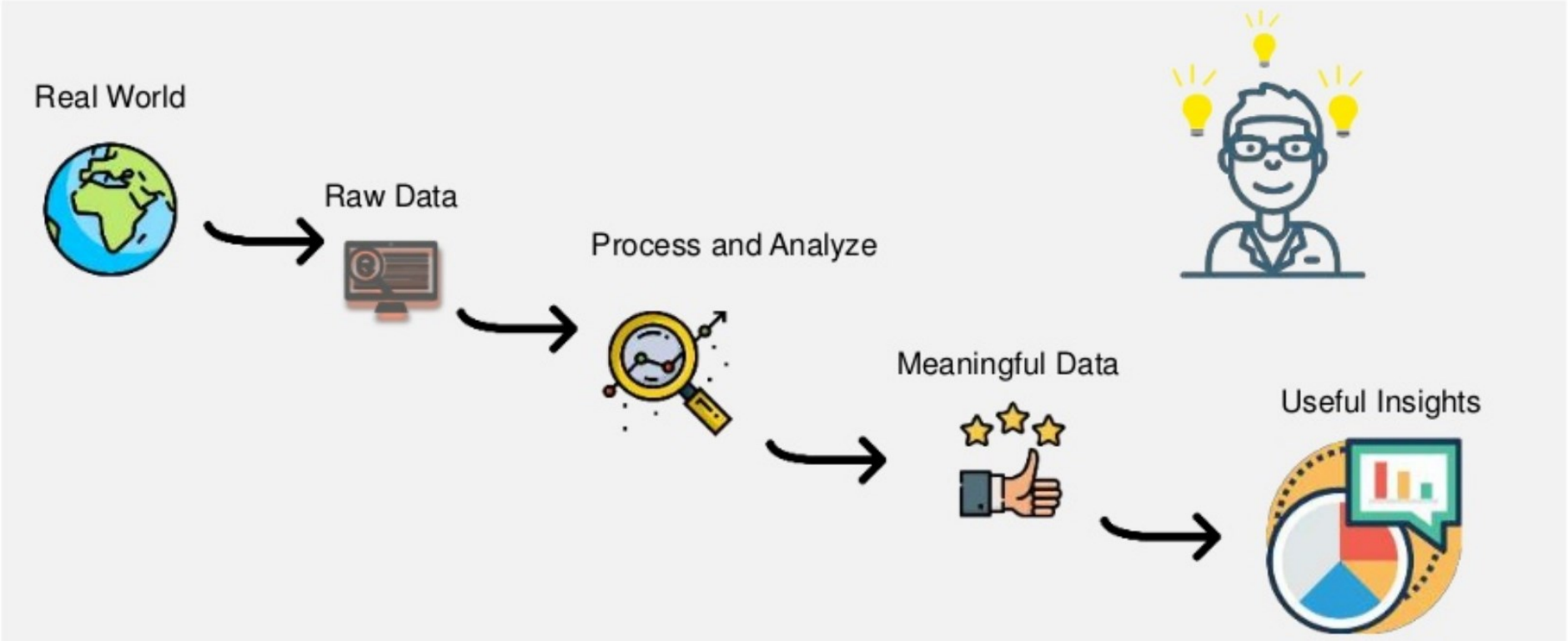
Data Science Process Cycle

วัฏจักรกระบวนการวิทยาการข้อมูล

จากบทที่ 1

What does a Data Scientist do?

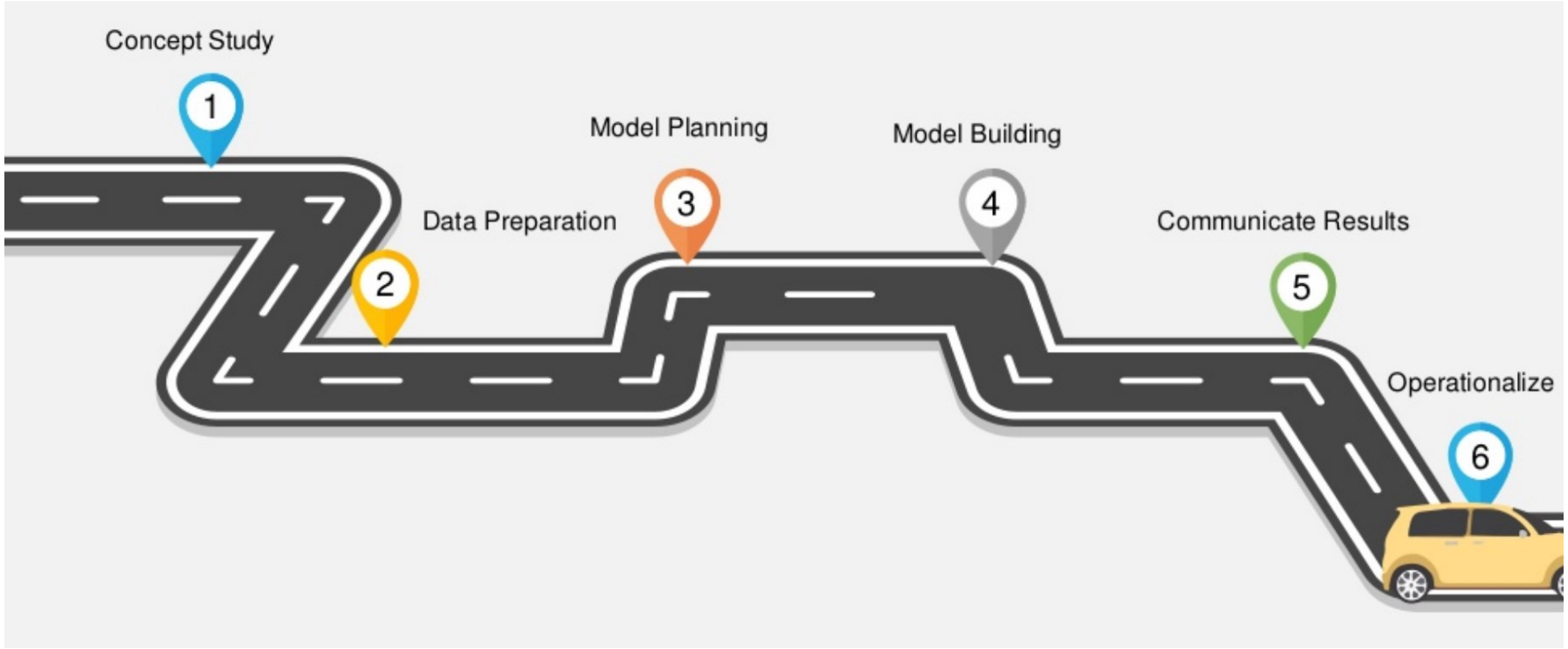
นักวิทยาการข้อมูลต้องทำกระบวนการใดบ้าง?



อย่างไรก็ตามขั้นตอนที่นักวิทยการทำข้อมูลได้ดำเนินการเป็นเพียงส่วนหนึ่งของโครงการพัฒนาระบบต่างๆ ที่ใช้เทคนิควิทยาการข้อมูล

ขั้นตอนกระบวนการทำงาน ในมุมมองโครงการด้านวิทยาการข้อมูล (Processes of Data Science Project)

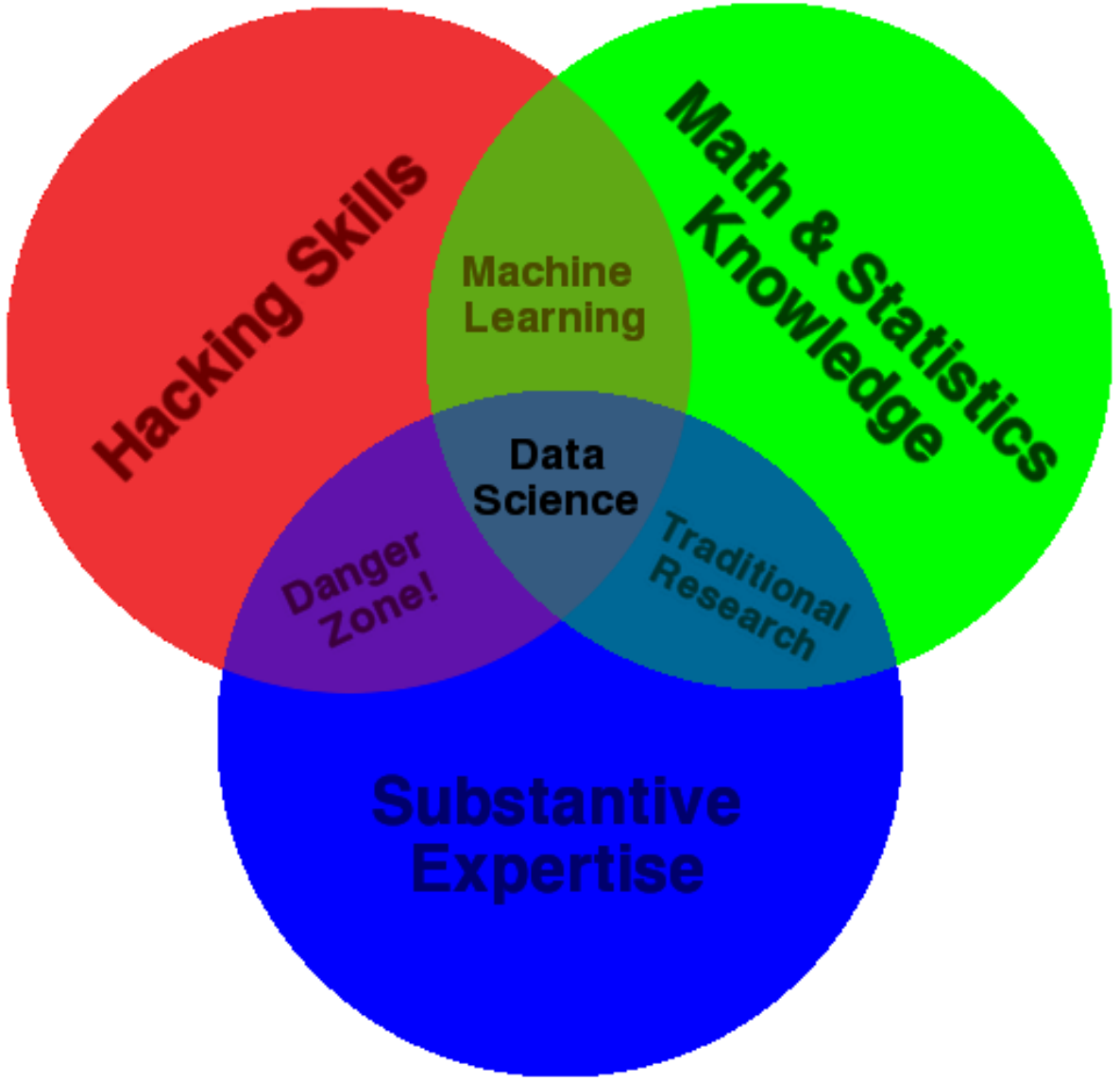
- ภาพรวมของการนำเทคนิควิทยาการข้อมูลไปใช้ประโยชน์ด้านต่าง ๆ
- เช่น ธุรกิจ การเกษตร การศึกษา ฯลฯ



Skills (ทักษะ) ที่จำเป็นของ “ทีม” นักวิทยาการข้อมูล

Data Science Team

Venn Diagram



กรอบการทำงานแบบ CRISP-DM

CRoss-industry Standard Process for Data Mining

- เป็นกระบวนการมาตรฐานอุตสาหกรรม สำหรับระบบการ ทำเหมืองข้อมูล (Data Mining)
  - เหมือนกับกระบวนการ ISO ในโรงงานอุตสาหกรรม
  - เหมือนกับกระบวนการ CMMI ในการพัฒนาซอร์ฟแวร์
- กำหนดมาตรฐานเพื่อ
  - เป็นขั้นตอนการทำงานมาตรฐาน (workflow) ที่ทำให้การพัฒนาเหมืองข้อมูลเหมาะสมต่อกลยุทธ์และความต้องการของธุรกิจหรือการวิจัย
  - ทำให้กระบวนการเหมืองข้อมูลนั้นเชื่อถือได้ (Reliable) และสามารถทำซ้ำได้ (repeatable) โดยผู้ที่มีพื้นฐานด้านเหมืองข้อมูลไม่มากนักได้

Data Mining หรือ ระบบการทำเหมืองข้อมูล เป็นการพัฒนาเทคนิค เพื่อค้นหารูปแบบที่มีประโยชน์ต่อธุรกิจที่แฝงอยู่ในข้อมูล ซึ่งต่อมากระบวนการเหมืองข้อมูลก็ถูกรวบรวมเข้ามาไว้ใน สาขาวิชา Data Science ซึ่งแนวคิดของมาตรฐาน CRISP-DM ก็มักถูกนำมาใช้ใน Data Science Process Management เช่นเดียวกัน

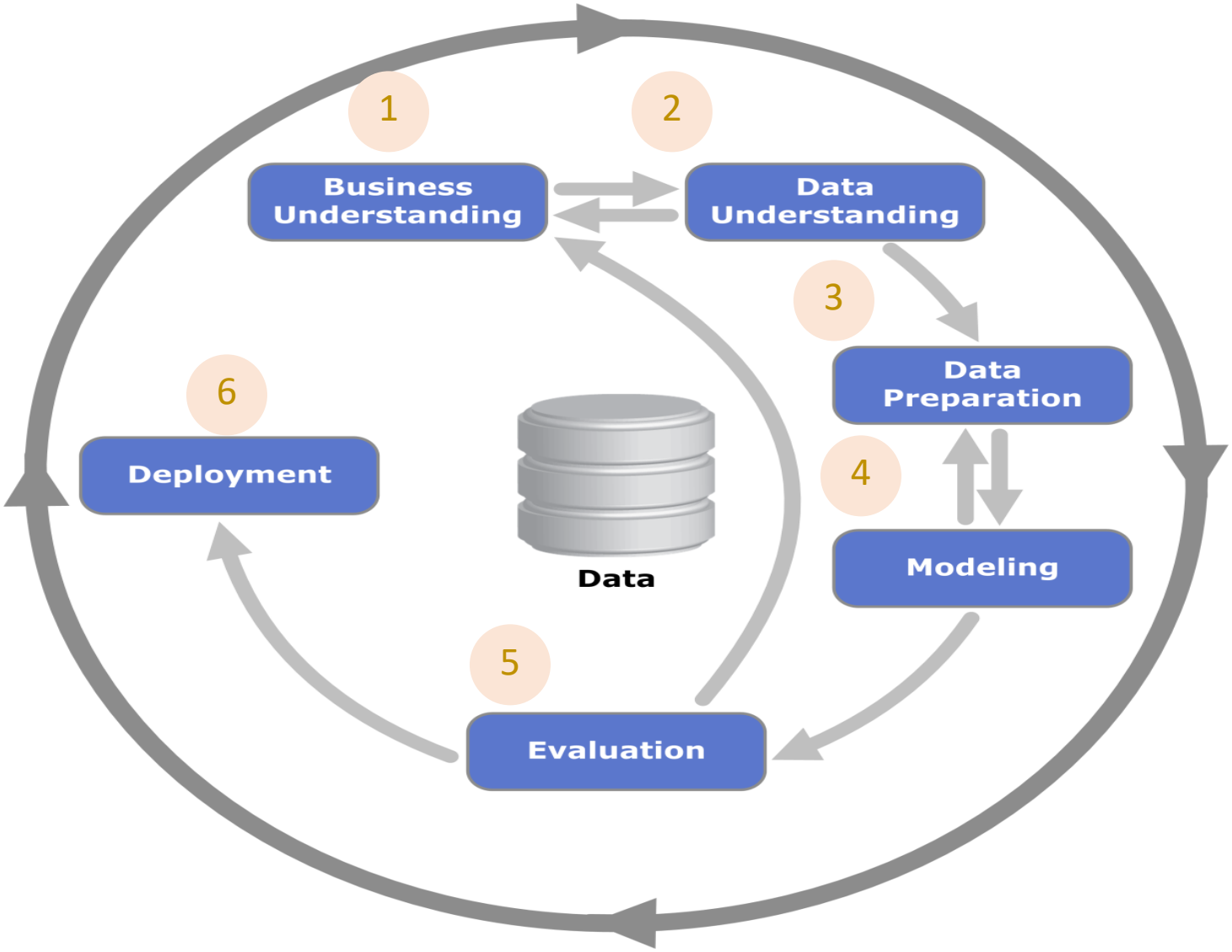
CRISP-DM → CRISP-DS (Data Science)

กรอบการทำงานแบบ CRISP-DM

ประโยชน์ของการมีกรอบการทำงาน (framework) มาตรฐาน

- กรอบการทำงานช่วยในการบันทึกและเพิ่มประสบการณ์ (recording experience)
  - นำไปสู่การทำซ้ำแต่ละโปรเจคได้
- บูรณาการกับการทำการวางแผนและบริหารโครงการได้ (project planning and management)
- อำนวยความสะดวก(comfort) ให้กับผู้ที่รับไปใช้ (adopter)
  - สามารถแสดงพัฒนาการเติบโต (maturity) ของการทำเหมืองข้อมูลในโปรเจคได้
  - ลดการผูกขาดกับคนใดคนหนึ่งในคณะทำงานโครงการ (reduces dependency on stars)

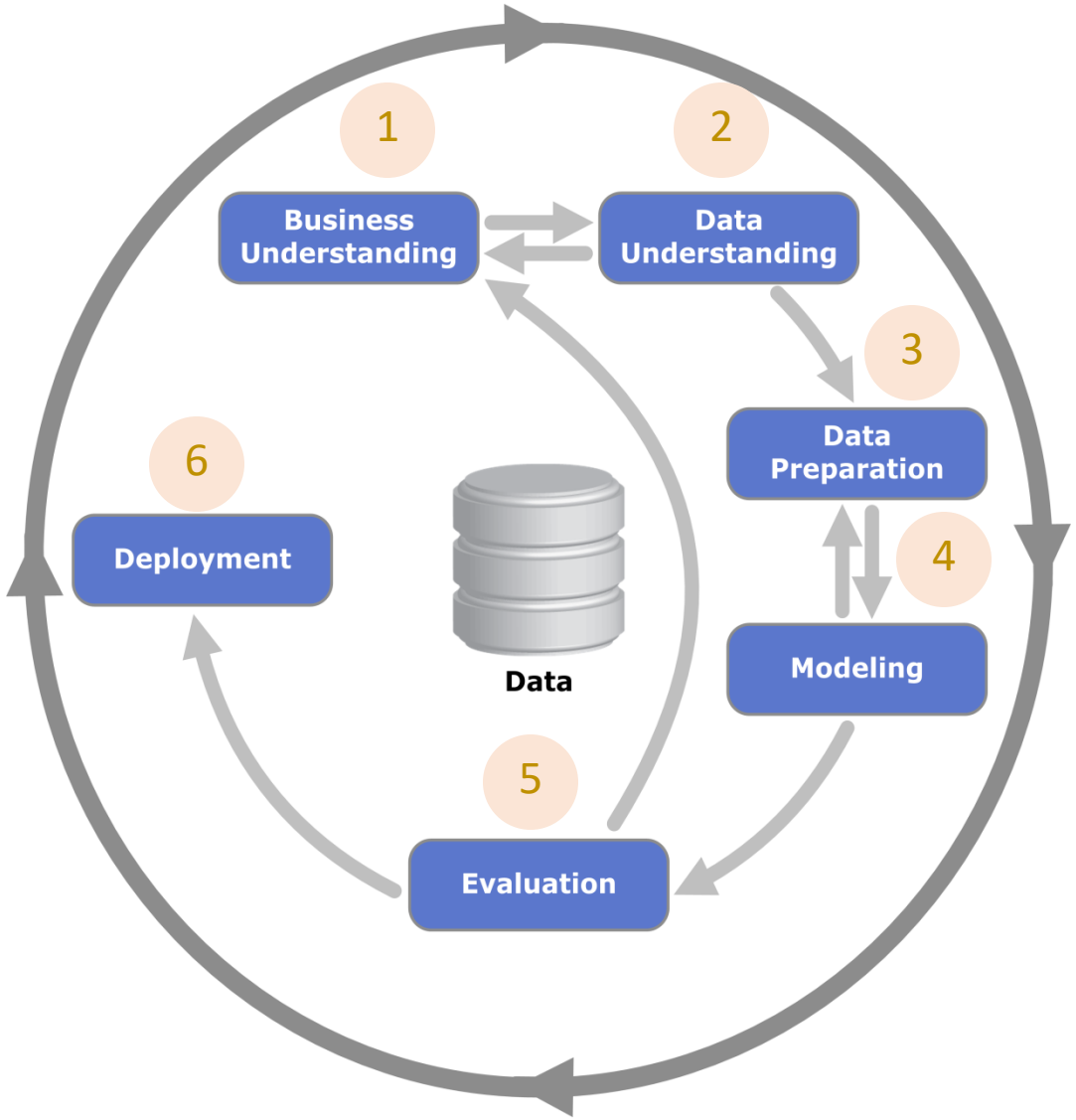
ขั้นตอนของ CRISP-DS



CRISP-DS: 6 Phases Process Cycle

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

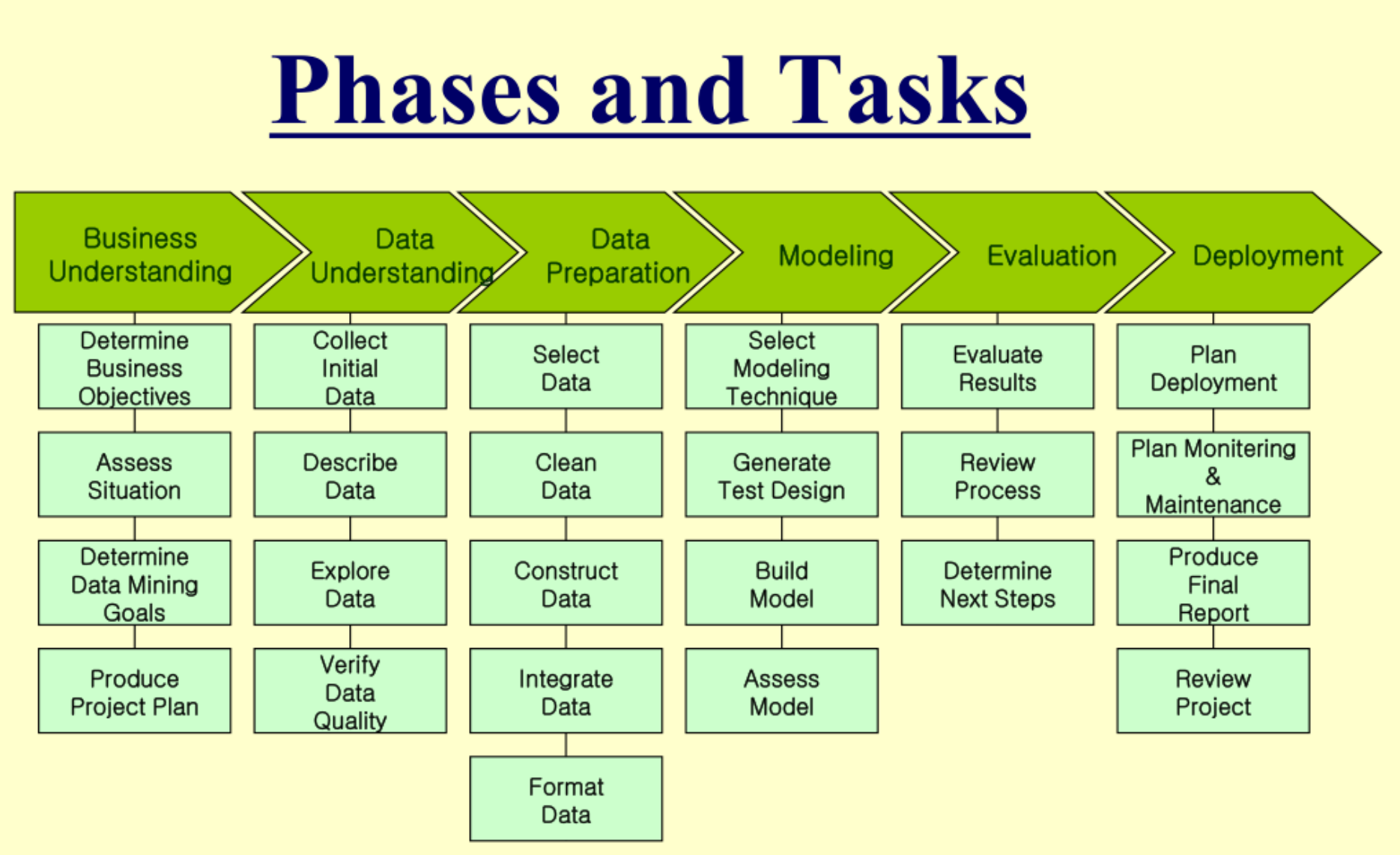
ขั้นตอนของ CRISP-DS



- Adaptive phases sequence
- Feedback process

โดยปกติ
ขั้นตอน 1-3 มักใช้เวลามากที่สุด

การดำเนินการในแต่ละ phases CRISP-DS



1. Business Understanding

- การทำความเข้าใจการทำธุรกิจ
- 1.1 Determine Business Objectives

  ระบุวัตถุประสงค์ (Objectives) และความจำเป็นของโครงการอย่างชัดเจน
- 1.2 Assess Situation

  ประเมินสถานการณ์ (Situation) ระบุข้อจำกัด หรือ กฎเกณฑ์
- 1.3 Determine Goals

  ระบุผลลัพธ์ หรือ เป้าหมาย (Goals) ที่ต้องการได้จากการทำเหมืองข้อมูล การวิเคราะห์ข้อมูล
- 1.4 Produce Project Plan

  จัดเตรียมวางแผนกลยุทธ์ (Plan) เบื้องต้นสำหรับบรรลุวัตถุประสงค์

ตัวอย่าง Goal

- ทำอย่างไรถึงเพิ่มยอดขายให้กับสินค้าชนิดต่างๆได้
- ทำอย่างไรให้ลูกค้ากลับมาซื้อสินค้าอีก
- อยากรู้ว่าลูกค้าคนใดบ้างที่มีโอกาสตั้งครรภ์
- อยากทำนายปริมาณน้ำฝนที่ตกในอีก 2 วันข้างหน้า
- อยากแบ่งกลุ่มนักศึกษาออกมาตามทักษะของแต่ละคน

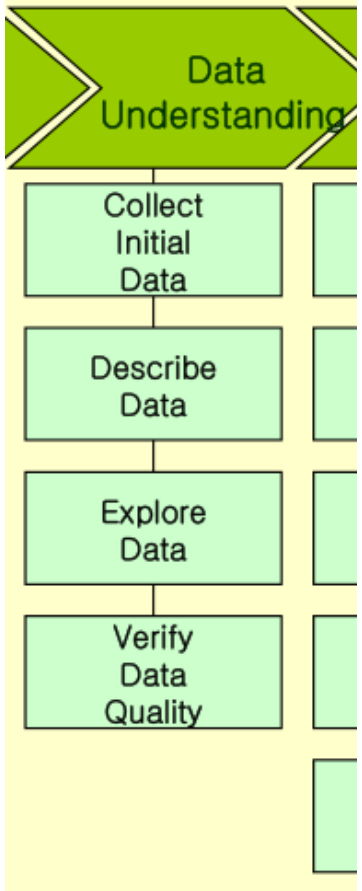
ตัวอย่าง CRIPS-DM

Business Understanding

- บริษัทผู้ผลิต smart phone แห่งหนึ่งกำลังจะวางตลาดสินค้าระดับ premium รุ่นใหม่
- ต้องการทำยอดขายสินค้านี้ในปริมาณมาก และ กระตุ้นตลาดอย่างต่อเนื่อง (Objective)
- บริษัทมีข้อมูลการซื้อ smart phone รุ่นก่อนหน้านี้ผ่านช่องทาง shopping online ของกลุ่มลูกค้าอยู่แล้ว (Situation)

- **Business Understanding**
  - เพื่อให้ยอดจำหน่ายเกิดขึ้นตลอดเวลาและสม่ำเสมอ
  - (Situation) บริษัททราบดีว่าโปรโมชั่นไม่จำเป็นต้องส่งให้กับลูกค้าทุกๆคน สามารถส่งเฉพาะบางคนในบางช่วงเวลา เพราะลูกค้ามีพฤติกรรมการสั่งซื้อหลังเริ่มวางขายผลิตภัณฑ์ที่ไม่เหมือนกัน (พวกที่ซื้อทันที พวกที่รอการรีวิว พวกที่รอสินค้าติดตลาด)
  - บริษัทจึงหาวิธีแบ่งกลุ่มลูกค้าออกเป็นกลุ่มๆ ได้ตามช่วงเวลาที่มีการสั่งซื้อหลังเริ่มวางขาย (Goal)
  - บริษัทส่งข้อมูลโปรโมชั่นที่เข้ากันกับลูกค้าให้ลูกค้าแต่ละกลุ่มที่ไม่เหมือนกัน ตามจังหวะเวลาที่เหมาะสม (Plan)

2. Data Understanding



- การทำความเข้าใจข้อมูลทางธุรกิจ | การทำความเข้าใจข้อมูลงานวิจัย
  - 2.1 Collect Data การเก็บรวบรวมข้อมูล (Data) ที่เกี่ยวข้อง
    - แหล่งข้อมูลมาจากไหน ภายใน|ภายนอก
  - 2.2 Describe Data ทำความเข้าใจข้อมูลเพื่อสร้างความคุ้นเคยข้อมูลและสามารถอธิบาย (describe) ข้อมูลได้
    - รวามถึงแง่กฎหมาย และ ทางเทคนิค
  - 2.3 Explore Data ค้นหา (explore) ทำความเข้าใจเชิงลึกเบื้องต้น (graph, attribute, missing value)
  - 2.4 Verify Quality ประเมินคุณภาพของข้อมูล
    - มีปริมาณและมีรายละเอียดมากพอการวิเคราะห์หรือไม่?
    - ข้อมูลมีความน่าเชื่อถือหรือไม่?
- เพื่อสร้างสมมติฐานสำหรับการวิเคราะห์สารสนเทศที่ซ่อนอยู่ (Insight)

ตัวอย่าง CRIPS-DM

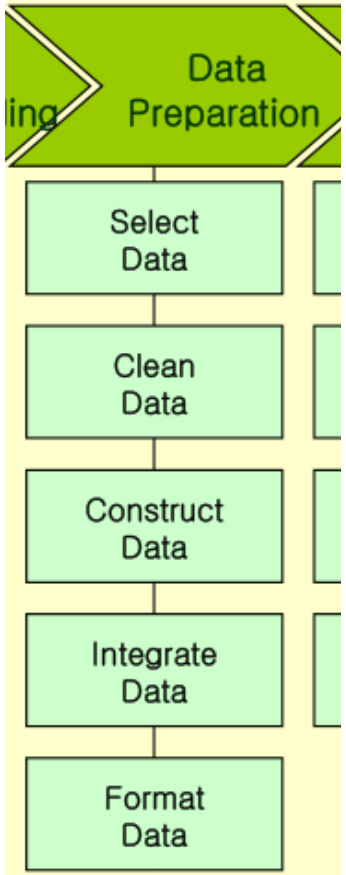
Data Understanding

- บริษัทเก็บข้อมูลการสั่งซื้อสินค้า online ผ่านเว็บในอดีต และสามารถนำมาใช้วิเคราะห์ได้ (Collect)
- บริษัทมีสมมติฐานสำหรับการวิเคราะห์สารสนเทศที่ซ่อนอยู่ และ อธิบาย (Describe) ว่า ลูกค้าสามารถแบ่งกลุ่มได้จากพฤติกรรมการซื้อสินค้าในแต่ละช่วงเวลาหลังสินค้าใหม่นั้นออกวางจำหน่าย
- บริษัทกำหนดว่า ลูกค้าน่าจะมีสี่กลุ่ม (customer group) แบ่งตามช่วงเวลาที่สั่งซื้อสินค้าหลังเริ่มวางจำหน่าย
  - Innovator คือ ลูกค้าที่ซื้อทันทีหลังสินค้าวางจำหน่ายในสัปดาห์แรก
  - Early Adaptor คือ ลูกค้าที่ซื้อหลังสัปดาห์แรกไปจนถึงสัปดาห์ที่ 3
  - Early Majority คือ ลูกค้าที่ซื้อหลังสัปดาห์ที่ 3 ไปจนถึงเดือนที่ 2
  - Late Majority คือ ลูกค้าที่ซื้อหลังจากเดือนที่ 2 ไปตลอด
- จำนวนข้อมูลน่าจะใช้ประมาณ 1000 คน
- ข้อมูลแบ่งเป็น ข้อมูลส่วนตัว (gender, age) และ พฤติกรรมการใช้งานเครือข่าย(Website_activity, Social_media_account, payment_method) และการทำธุรกรรม

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

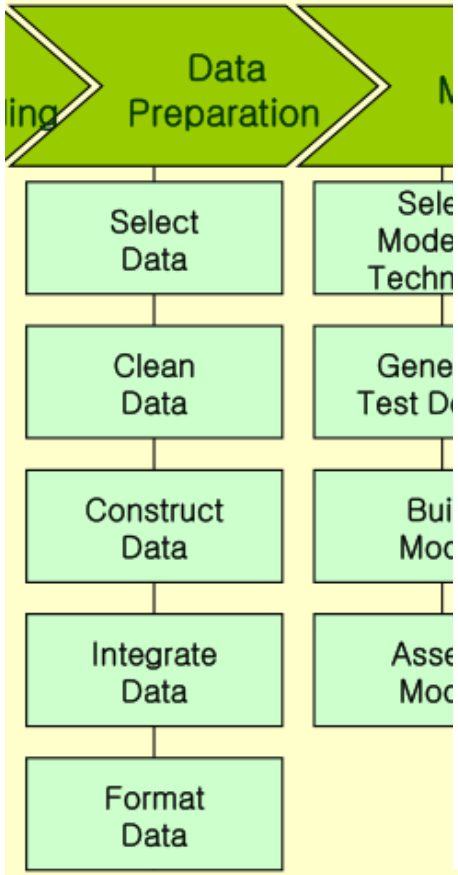
3. Data Preparation



- การเตรียมข้อมูล *ขั้นตอนที่ใช้เวลานานที่สุด
- เป็นขั้นตอนที่ทำการแปลงข้อมูลที่ได้รวบรวมมา (Raw data) ให้กลายเป็นข้อมูลที่สามารถนำไปวิเคราะห์ได้

- ตัวแบบ (Model) เพื่อการวิเคราะห์ข้อมูลที่ดี ขึ้นอยู่กับคุณภาพของข้อมูลที่ใช้วิเคราะห์ ซึ่งขึ้นกับ
  - 3.1 การคัดเลือกข้อมูล (data selection)
  - 3.2 การกลั่นกรองข้อมูล (data cleaning)
  - 3.3 การสร้างข้อมูล (data construction)
  - 3.4 การผสมผสานข้อมูล (data Integration)
  - 3.5 การทำข้อมูลให้อยู่ในรูปแบบ (format) ที่เหมาะสมต่อการวิเคราะห์ข้อมูล

3. Data Preparation



3.1 Data Selection การคัดเลือกข้อมูล

- เลือกเฉพาะข้อมูลที่เกี่ยวข้องกับสิ่งที่เราทำการวิเคราะห์
  - เลือกสุ่มข้อมูล (Sampling) แค่บางชุดออกมาจากข้อมูลทั้งหมด
  - เลือกเฉพาะบางลักษณะของข้อมูล (attribute)

แบบสำรวจความพึงพอใจ โครงการ การพัฒนาสื่อประสม เรื่องทรัพย์สินทางปัญญา
“การละเมิดลิขสิทธิ์ซอฟต์แวร์” ด้วยเทคนิค Animation 3D

คำชี้แจง แบบสอบถามชุดนี้ จัดทำขึ้นโดยนักศึกษา สาขางานคอมพิวเตอร์ธุรกิจ ระดับชั้น ปวส.2 วิทยาลัยอาชีวศึกษาสุพรรณบุรี มีวัตถุประสงค์เพื่อศึกษาและดำเนินการสร้างแอนิเมชั่น 3D เรื่องทรัพย์สินทางปัญญา “การละเมิดลิขสิทธิ์ซอฟต์แวร์” และเพื่อศึกษาความพึงพอใจต่อสื่อแอนิเมชั่น

ตอนที่ 1 ข้อมูลทั่วไป

1. เพศ O ชาย O หญิง
2. อายุ O 10-20 ปี O 21-30 ปี O 31-40 ปี O 40 ปีขึ้นไป
3. อาชีพ O ครู-อาจารย์ O นักเรียน-นักศึกษา O อื่นๆ ระบุ.....
4. วุฒิการศึกษา O ประถม O มัธยมต้น O มัธยมปลาย O ปวช. O ปวส. O ปริญญาตรีขึ้นไป

ตอนที่ 2 แบบสอบถามความพึงพอใจของผู้รับชมสื่อแอนิเมชั่น

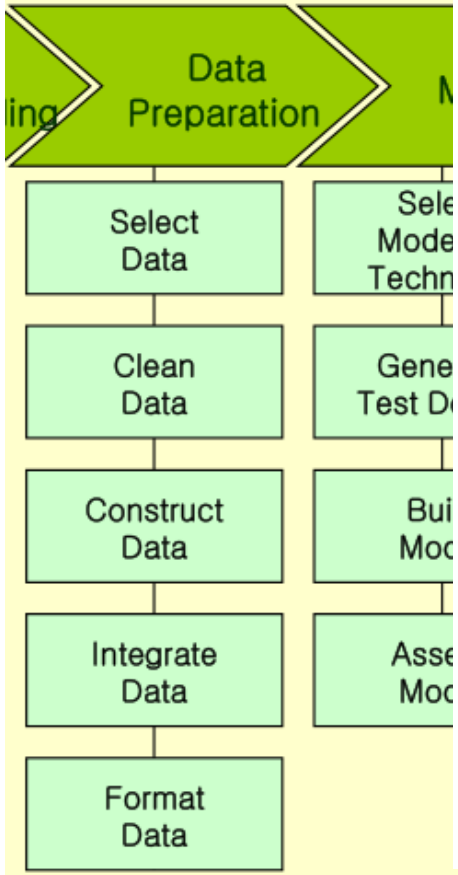
เกณฑ์การประเมิน ระดับ 1 หมายถึง น้อยที่สุด
ระดับ 2 หมายถึง น้อย
ระดับ 3 หมายถึง ปานกลาง
ระดับ 4 หมายถึง มาก
ระดับ 5 หมายถึง มากที่สุด

โปรดใส่เครื่องหมาย ✓ ลงในช่องที่ตรงกับความพึงพอใจของท่านมากที่สุด

รายการประเมิน	ระดับความพึงพอใจ				
	5	4	3	2	1
1. ด้านเนื้อหา					
1.1 ความเหมาะสมในการจัดลำดับการเล่าเรื่องมีความเข้าใจง่าย					
1.2 เนื้อหามีความถูกต้อง					
1.3 เนื้อหามีความเหมาะสม					
1.4 เนื้อหามีความกะทัดรัด เข้าใจง่าย					
1.5 Animation 3D มีภาพลักษณ์โดยรวมที่ดี					
1.6 การเล่าเรื่องสื่อถึงการละเมิดลิขสิทธิ์ซอฟต์แวร์					

แบบสอบถามใบที่	เพศ	อายุ	อาชีพ	วุฒิการศึกษา	ข้อที่ 1	ข้อที่ 2	ข้อที่ 3	ข้อที่ 4	ข้อที่ 5
1	ช	22	รับราชการ	ป ตรี	1	2	2	5	1
2	ญ	23	ข้าราชการ	ปริญญาตรี	1	5	1	3	3
3	ช	21ปี 9 เดือน	ส่วนตัว	ม 6	2	1	3	1	5
4	ชาย	24	freelance	มัธยมศึกษาตอนปลาย	5	2	3	3	3
5	หญิง	25.3	นักศึกษา	ไม่บอก	3	2	2	5	5

3. Data Preparation



3.2 Data Cleaning การทำความสะอาดข้อมูล

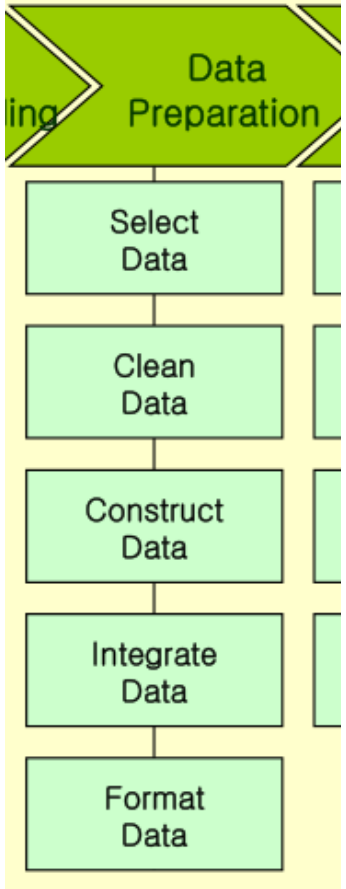
- ลบข้อมูลที่ซ้ำซ้อน (redundancy reduction)
- แก้ไขข้อมูลที่ผิดพลาด (missing values)
- ขจัดข้อมูลที่ผิดรูปแบบ (Outlier)

3. Data Preparation

ตัวอย่างข้อมูลที่จำเป็นต้องทำความสะอาด (Cleaning)

รหัส	เพศ	อายุ	ความสูง	น้ำหนัก
55001	ชาย	20	180	70
55002	ญ		120	45
55003	หญิง	21	160	250
55004	ช	19	168	89
	ผิดรูปแบบ	ขาดหาย	Outlier	

3. Data Preparation



3.3 Data Transformation การแปลงข้อมูล

3.3 การสร้างข้อมูลขึ้นมาใหม่ (Data Construction)

3.4 การรวมข้อมูลเข้าไว้ด้วยกัน (Data Integration)

3.5 การแปลงรูปแบบ (format) ข้อมูลใหม่ให้อยู่ในข่วงตัวเลข (Scaling) เดียวกัน

- ต้องไม่เปลี่ยนสารสนเทศที่อยู่ในข้อมูล

- เพื่อให้ข้อมูลอยู่ในรูปแบบที่พร้อมนำไปใช้ในการวิเคราะห์ตามเทคนิควิทยาการข้อมูลได้สะดวก

Data Transformation

ตัวอย่าง ใบเสร็จสินค้า

ข้อมูล Transaction ในฐานข้อมูลร้านค้า

ID	สินค้า	จำนวนที่ซื้อ
1	กะปิ	1
1	น้ำปลา	5
1	ซีอิ๊ว	3
2	พริก	1
2	กะปิ	1
3	พริก	2
3	กะปิ	1
3	น้ำปลา	1

ข้อมูลที่ถูก Transformation เพื่อกระบวนการต่อไป

ID	กะปิ	น้ำปลา	ซีอิ๊ว	พริก
1	TRUE	TRUE	TRUE	-
2	TRUE	-	-	TRUE
3	TRUE	TRUE	-	TRUE

Data Transformation

ตัวอย่าง การแปลงข้อมูลภาพให้อยู่ในรูปแบบที่เหมาะแก่การประมวลผล

จำนวน pixel สีแดง เขียว น้ำเงินที่ปรากฏในภาพ			
Pic ID	Red	Green	Blue
1	20	70	160
2			

ตัวอย่าง Data Preparation

- กรณีบริษัทจำหน่าย smart phone
  - ลบแอตทริบิวต์ (column) customer_ID จากหน้า13 ออกไป ได้เพราะเป็นแค่หมายเลขสมาชิกของลูกค้า ไม่มีประโยชน์ต่อการวิเคราะห์
  - แปลงอายุ (age) จากตัวเลขเป็นช่วงอายุ

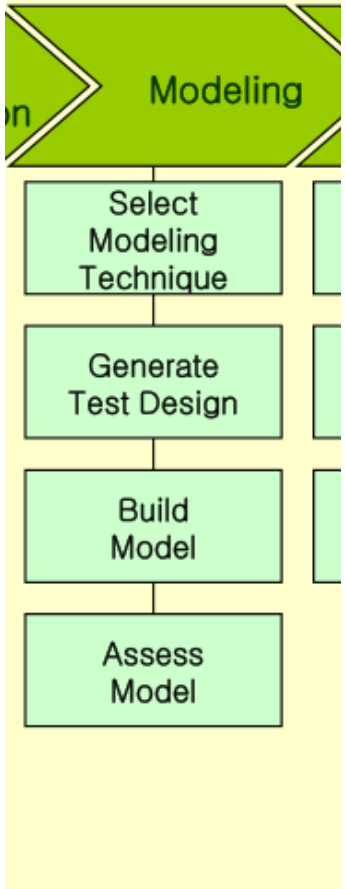
customer_ID
9123
4567
1254
3332

gender	age	website_activity	social_media_account	payment_method
M	High	rarly	no	bank transfer
M	Middle	regular	no	bank transfer
F	Middle	rarly	yes	bank transfer
M	High	rarly	yes	website account

คำตอบผลลัพธ์

customer_group
Late Majority
Innovator
Early Adopter
Early Adopter

4. Modeling



- เป็นขั้นตอนการทำงานวิเคราะห์ข้อมูลด้วยเทคนิค Statistical Modeling / Machine Learning

  4.1 เลือกและใช้เทคนิคเพื่อสร้างตัวแบบ (Model) ที่เหมาะสม บนพื้นฐานการวิเคราะห์วัตถุประสงค์และเป้าหมายจากขั้นตอนก่อนหน้านี้

  4.2 ออกแบบการทดสอบ (testing) ตัวแบบที่สร้างขึ้น

  4.3 สร้างตัวแบบ (Model)

  - รวมถึงการเลือกตัวแปร (parameter) เริ่มต้นของเทคนิคที่เหมาะสม
  - ศึกษาพฤติกรรมของตัวแบบ
    - วิเคราะห์ความไว (sensitivity analysis)

  4.4 ประเมิน (assess) ตัวแบบ

  - แปลผลการทำงานของตัวแบบ วินิจฉัยผลลัพธ์
  - สามารถใช้หหลายๆ เทคนิคเปรียบเทียบกันเพื่อให้ได้คำตอบที่ดีที่สุด

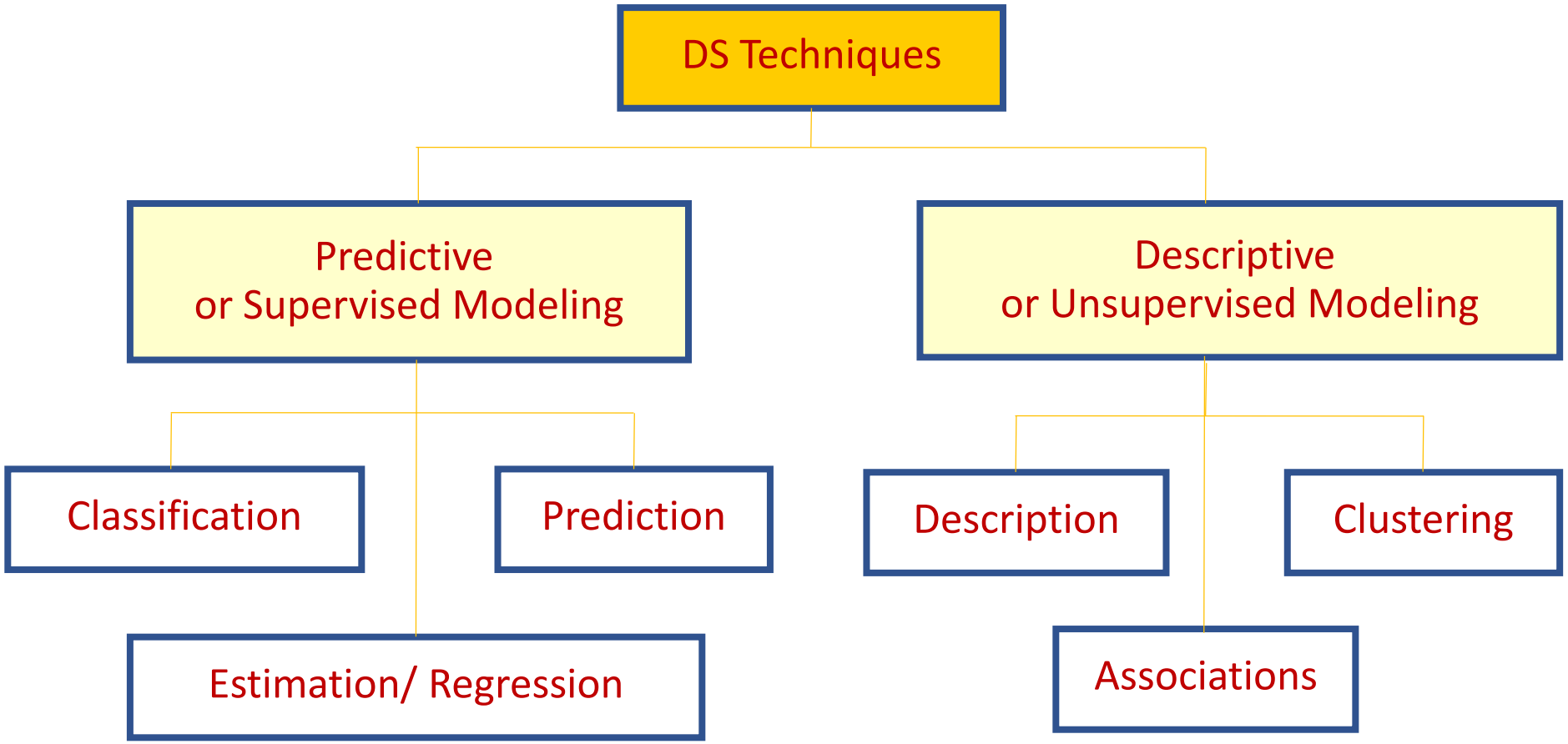
Modeling Techniques

“ไม่มีเทคนิคหรือเครื่องมือเพียงชนิดเดียวของกระบวนการวิทยาการข้อมูลที่เหมาะสมกับงานทุกชนิด งานในแต่ละชนิดก็จะมีเทคนิคที่เหมาะสมที่แตกต่างกันออกไป”

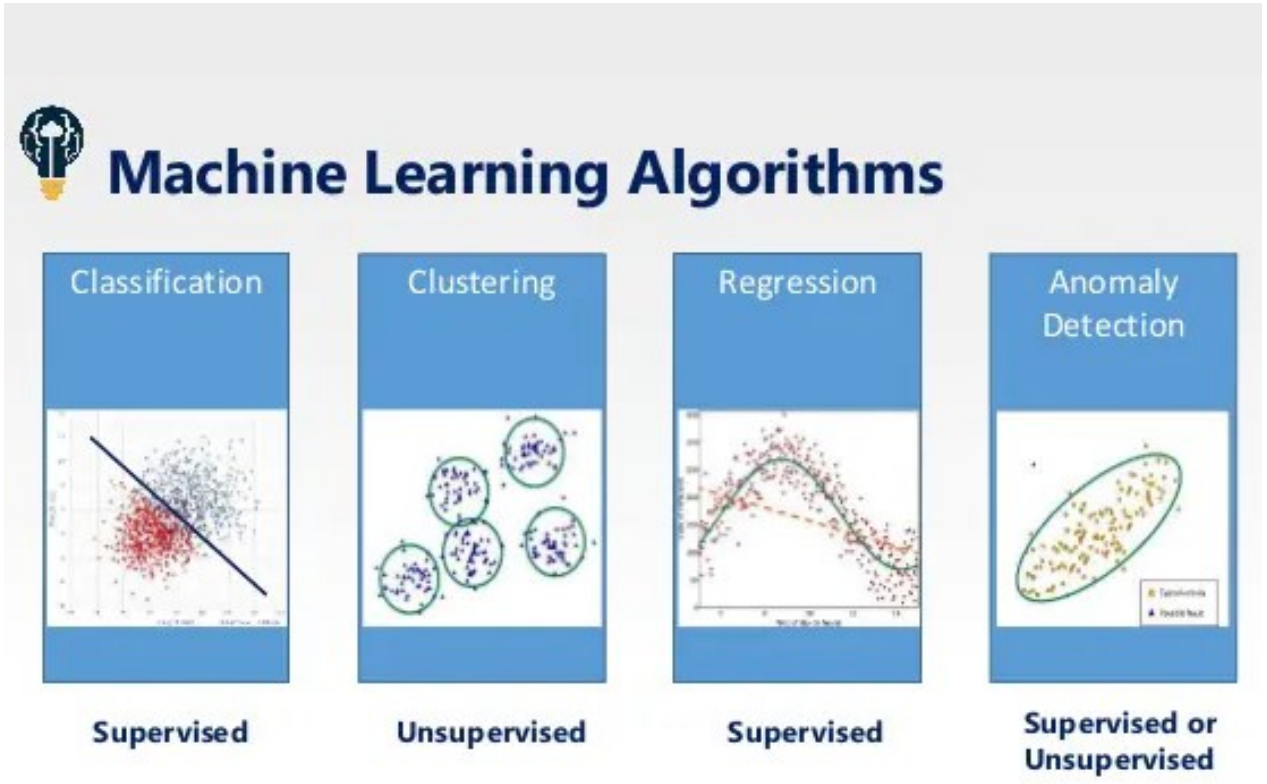
เทคนิคของการสร้างตัวแบบวิเคราะห์สำหรับวิทยาข้อมูล แบ่งเป็น 2 ประเภทหลัก

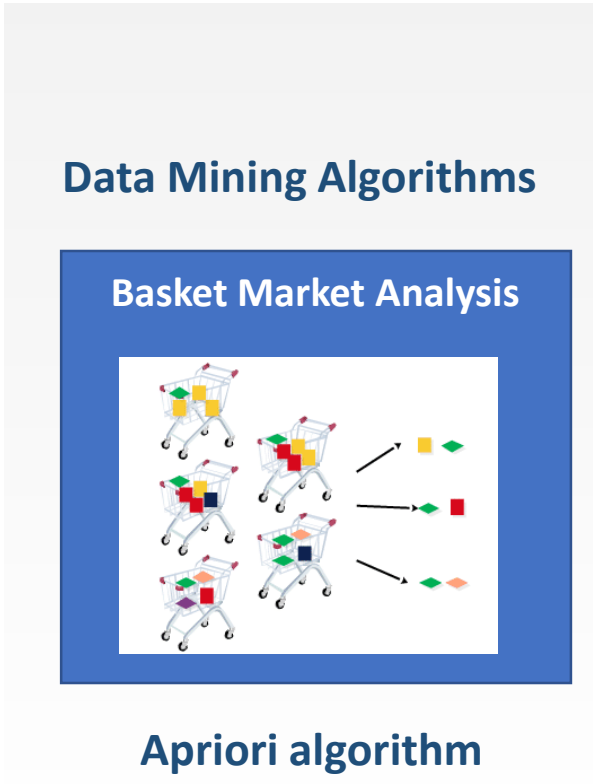
- แบบจำลองในการบรรยาย (Descriptive/ Unsupervised Modeling) ในที่นี้ อาจเป็นการหาความสัมพันธ์ต่างๆ (Association) หรือหาการจัดกลุ่มข้อมูล (Clustering) ซึ่งไม่ได้มีจุดมุ่งหมายเพื่อการทำนาย
    1. การพรรณนา (Description) (ในเชิงสถิติ)
    2. การหาความสัมพันธ์ (Association)
    3. การจัดกลุ่ม (Clustering)
- แบบจำลองในการทำนาย (Predictive/ Supervised Modeling) เป็นผลลัพธ์ที่สร้างจากการอนุมาน (Inference) ชุดข้อมูลปัจจุบัน เพื่อใช้ในการทำนายประเภทตัวอย่างในอนาคต
    1. การจัดหมวดหมู่ (Classification)
    2. การประเมินค่า (Estimation)
    3. การทำนายล่วงหน้า (Prediction)

ประเภทของเทคนิคในการทำตัวแบบวิเคราะห์ของวิทยาการข้อมูล



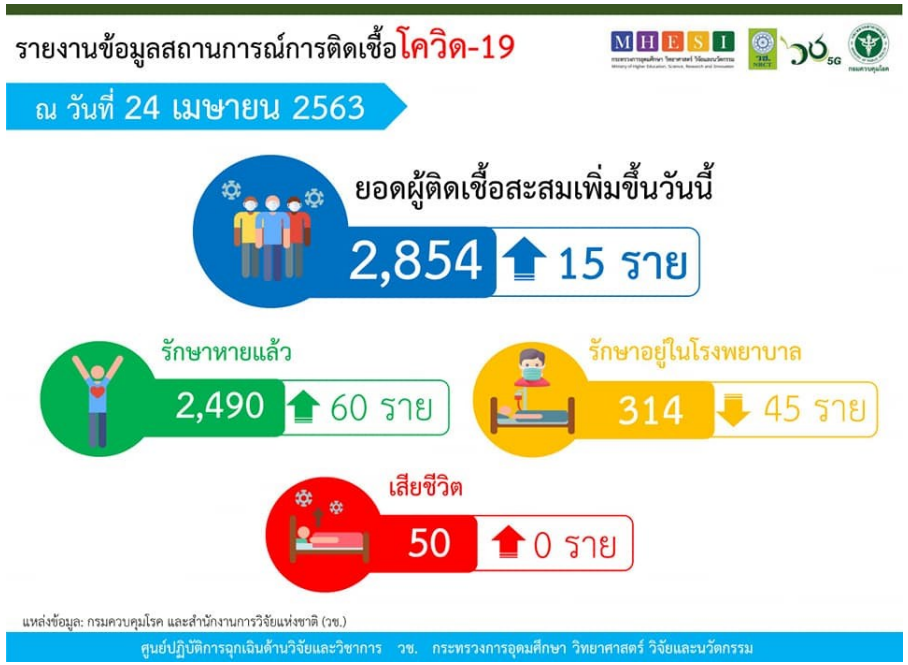
เทคนิคต่างๆ ที่อยู่ในกระบวนการวิทยาการข้อมูล

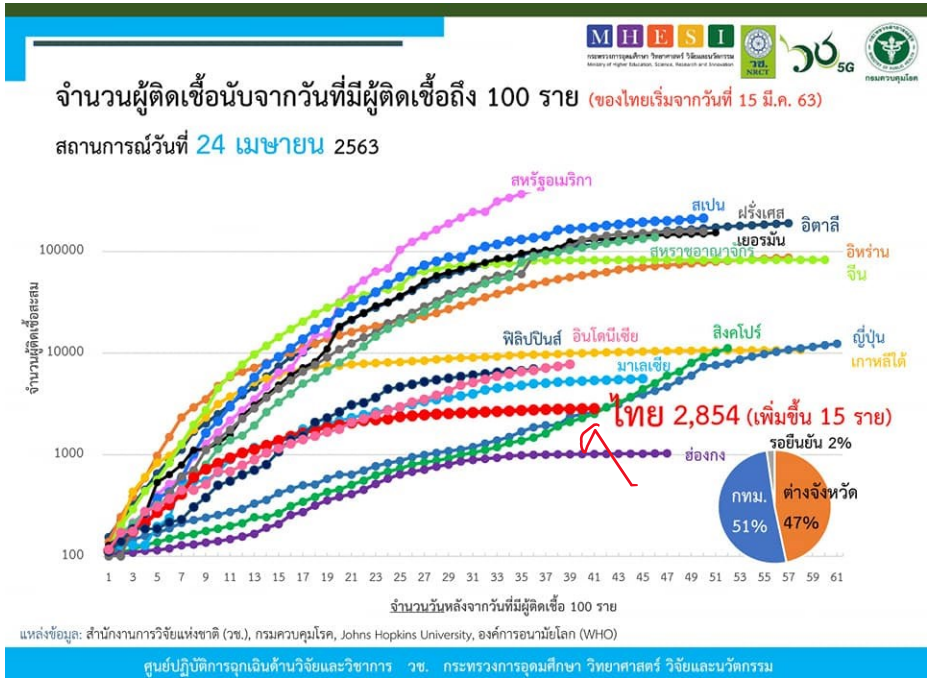




1. การพรรณนา (Description)

- วัตถุประสงค์ของการทำวิทยาการข้อมูล คือ ต้องการอธิบายรูปแบบและแนวโน้มของฐานข้อมูล
  - สามารถแปลความหมายและเข้าใจได้ง่าย
- เพื่อเพิ่มความเข้าใจในส่วนของประชากร ผลิตภัณฑ์ หรือ กระบวนการให้มากขึ้น
- การวิเคราะห์ข้อมูลโดยการสำรวจเชิงกราฟในการหารูปแบบและแนวโน้ม





2. การทำเหมืองกฎความสัมพันธ์ (Mining Association Rules)

- เป็น Descriptive หรือ Unsupervised Modeling
- การค้นหากฎความสัมพันธ์ มักเป็นงานทำเหมืองบนฐานข้อมูล Transactional เพื่อค้นหาสหสัมพันธ์ (correlation) ของสิ่งของ ส่วนใหญ่จะใช้ในการช่วยการวิเคราะห์ Market basket analysis
- การหากฎความสัมพันธ์แสดงอยู่ในรูปแบบ

$$X \rightarrow Y$$

- หมายถึง การเกิดขึ้นของไอเท็มเซต x เกิดขึ้นร่วมกันของไอเท็มเซต Y ด้วยค่าสนับสนุน (Support) และค่าความเชื่อมั่น (Confidence)
- ไอเท็มเซต เช่น เซตของสินค้าในร้าน หรือเซตของประเภทบริการ

ตัวอย่าง เทคนิค Association Rule Discovery

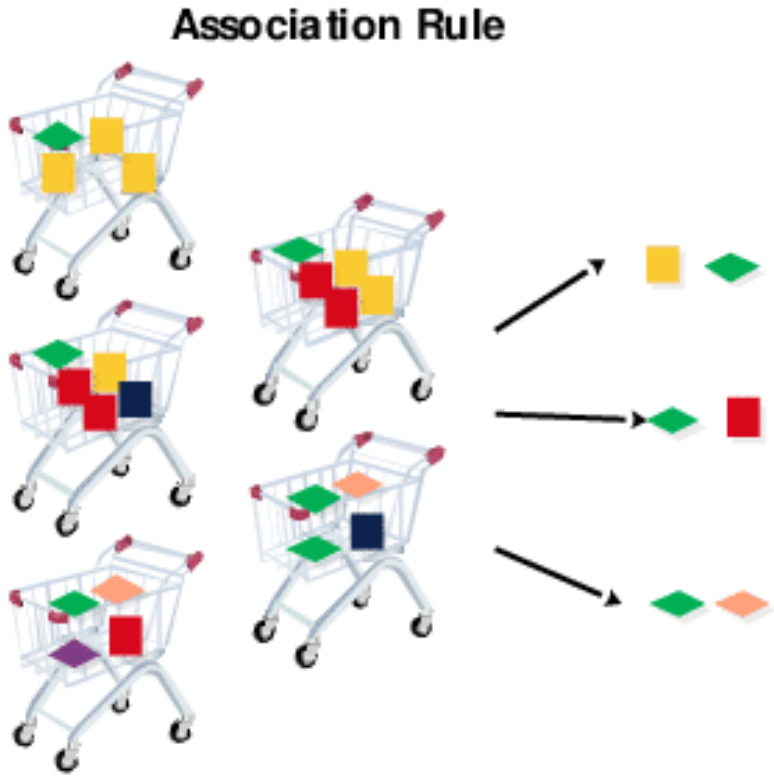
- ระบบแนะนำหนังสือให้กับลูกค้าแบบอัตโนมัติ ของ Amazon คือ
  - ลูกค้าที่ซื้อหนังสือเล่มหนึ่งๆ มักจะซื้อหนังสือเล่มใด พร้อมกันด้วยเสมอ เช่น

buys (x , database) -> buys (x , data mining) [80% , 60%]

  - หมายความว่า เมื่อซื้อหนังสือ database แล้วมีโอกาสที่จะ

ซื้อหนังสือ data mining ด้วย 60 % และมีการซื้อทั้งหนังสือ

database และหนังสือ data mining พร้อม ๆ กัน 80 %



- การซื้อสินค้าของลูกค้า 1 ครั้ง ต้องการทราบว่าสินค้า

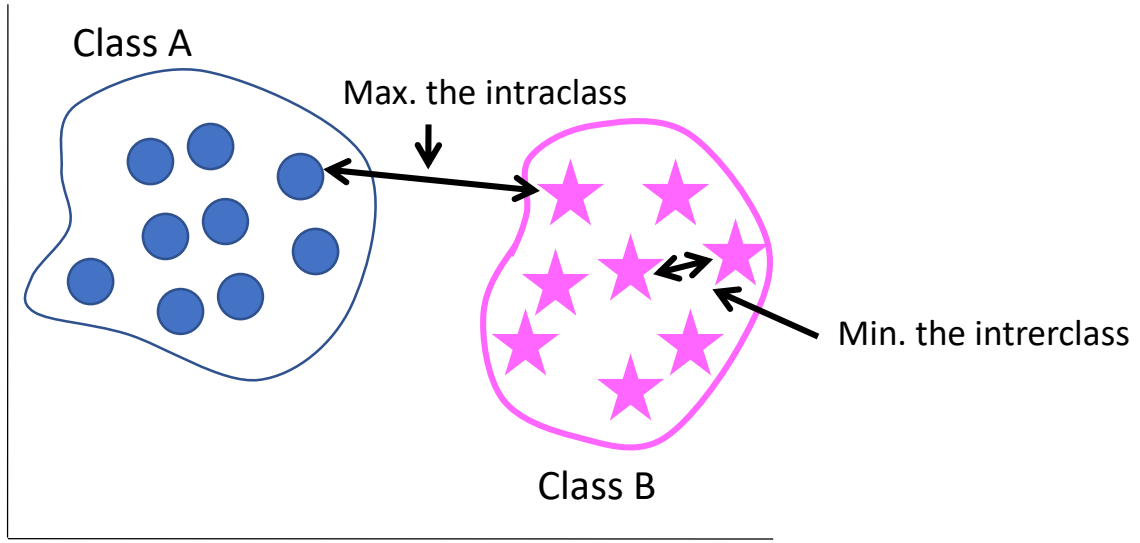
ใดบ้างที่ลูกค้ามักซื้อด้วยกัน เพื่อน าไปพิจารณาปรับปรุงการจัดวาง

สินค้าในร้าน เช่น

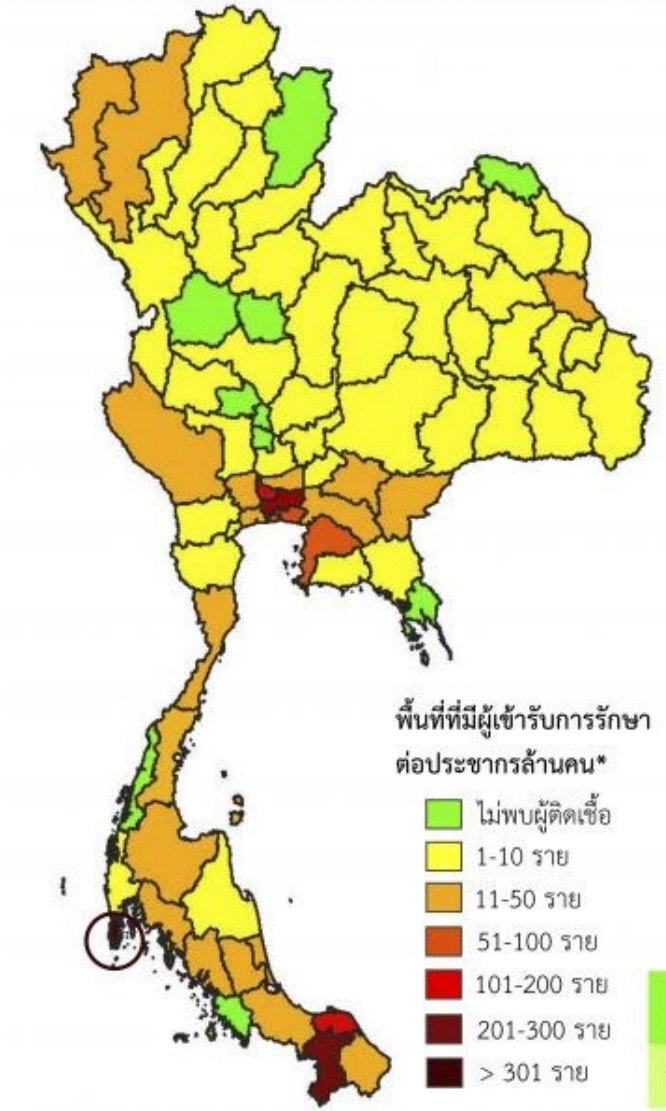
{ รองเท้า,ถุงเท้า } หรือ { ปากกา , หมึก }

3. การวิเคราะห์เพื่อการจัดกลุ่ม (Clustering)

- เป็น Descriptive หรือ Unsupervised Modeling
- เป็นการตรวจหากลุ่มตามธรรมชาติของข้อมูล โดยพิจารณาจากค่ามาตรวัดที่กำหนด ว่าวัตถุที่อยู่กลุ่มเดียวกันจะมีความคล้ายคลึงกันมากที่สุด และวัตถุต่างกลุ่มจะมีความคล้ายคลึงน้อยที่สุด



จำนวนผู้ติดเชื้อโรคโควิด-19 ต่อประชากร



ณ วันที่ 4 พ.ค. 2563

จังหวัด	ประชากร	จำนวนผู้ติดเชื้อสะสม*	State Quarantine	จำนวนผู้ติดเชื้อต่อประชากรล้านคน*
กรุงเทพมหานคร	5,666,264	1,526	14	269.14
ภูเก็ต	416,582	220		523.81
นนทบุรี	1,265,387	157		123.62
ยะลา	536,330	118	8	218.52
สมุทรปราการ	1,344,875	114		84.44
ชลบุรี	1,558,301	87	4	55.83
ปัตตานี	725,104	79	12	108.95
สงขลา	1,435,968	44	19 (+60)	30.56
เชียงใหม่	1,779,254	40		22.47
ปทุมธานี	1,163,604	39		33.62
นราธิวาส	808,020	28	6	34.65
นครปฐม	920,030	22		23.91
นครราชสีมา	2,648,927	19		7.17
สตูล	323,586	-	18	0

* จำแนกตามจังหวัดที่เข้ารับการรักษา

จังหวัดที่ยังไม่มีรายงานการรับรักษาผู้ป่วย (+1 จาก State Quarantine)

กำแพงเพชร, ชัยนาท, ตราด, น่าน, บึงกาฬ, พิจิตร, ระนอง, สิงห์บุรี, อ่างทอง, (+สตูล)

ศูนย์ปฏิบัติการด้านนวัตกรรมการแพทย์ และการวิจัยและพัฒนา วช. กระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัยและนวัตกรรม

แบบจำลองในการทำนาย (Predictive/ Supervised Modeling)

ได้แก่ 3. การจำแนกประเภทข้อมูล (Classification) 4. การประมาณค่า (Estimation) 5. การทำนาย (prediction)

- เป็นการค้นหาแบบจำลองหรือฟังก์ชันทางคณิตศาสตร์และสถิติจากข้อมูลที่มีอยู่

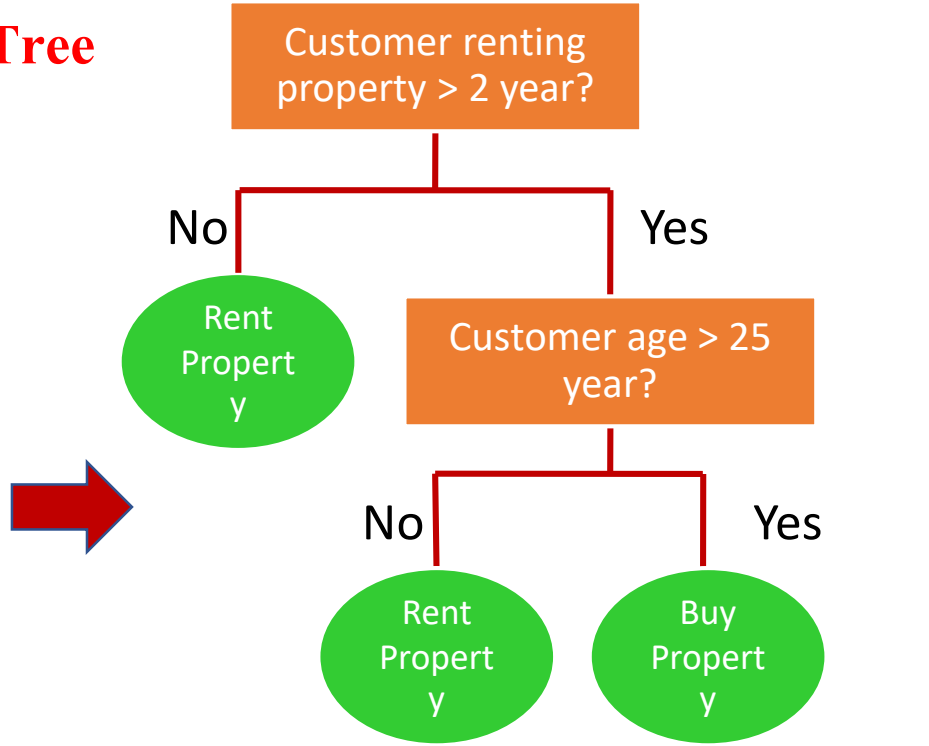
- ผลลัพธ์ที่ได้อาจจะอยู่ในแบบบตัวแบบ หรือ ฟังก์ชัน เช่น ต้นไม้ตัดสินใจ (Decision Tree) กฎการจำแนกประเภทข้อมูล หรือ เครือข่ายประสาทเทียม (Neural Network) เป็นต้น

ตัวอย่าง: เทคนิคการจำแนกข้อมูล (Classification)

- **Classification: Decision Tree**

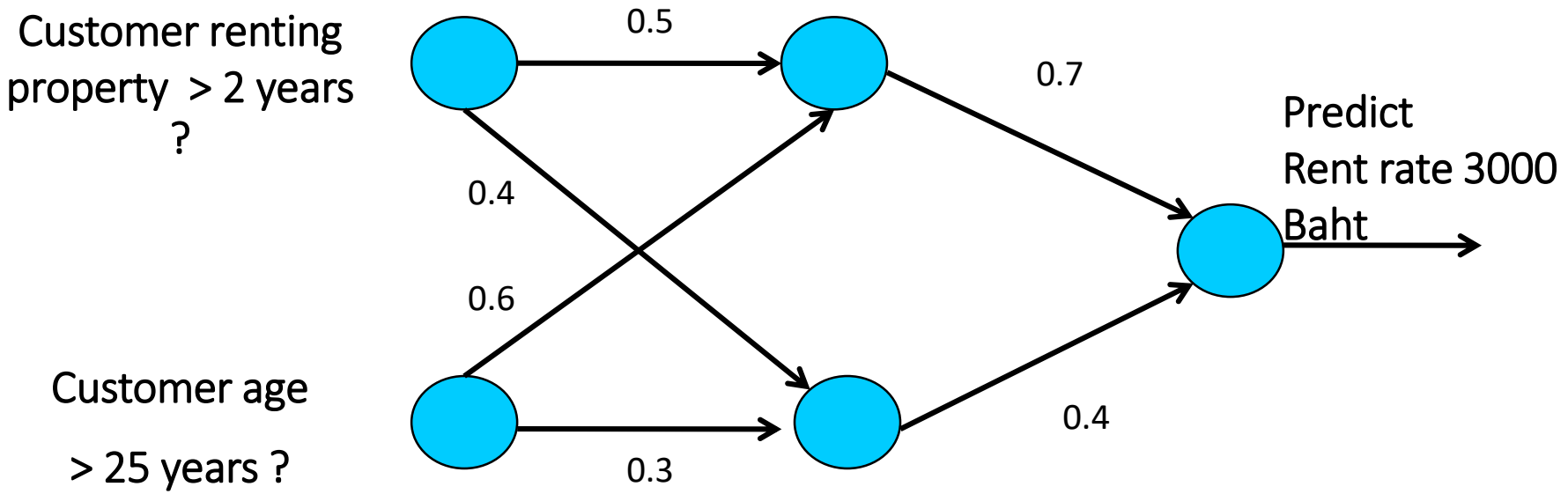
Business Info

Age	Rent Period	Buy
23	3	No
36	1.5	No
20	1.5	No
27	2	Yes
20	1	No
50	2.5	Yes
36	1	No
36	2	Yes
22	2.5	no



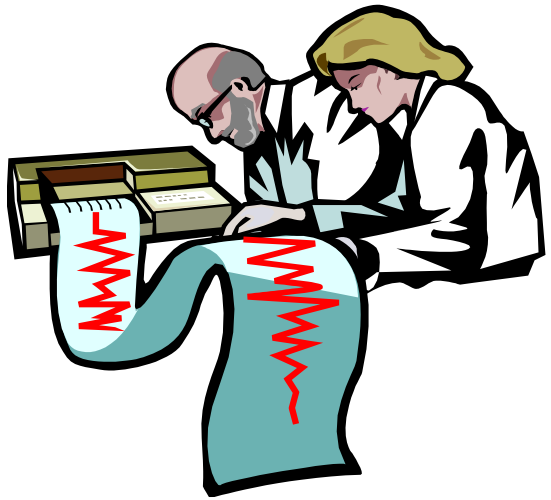
ตัวอย่าง: เทคนิคการประมาณค่า (Estimation)

- **Prediction: Neural Network**



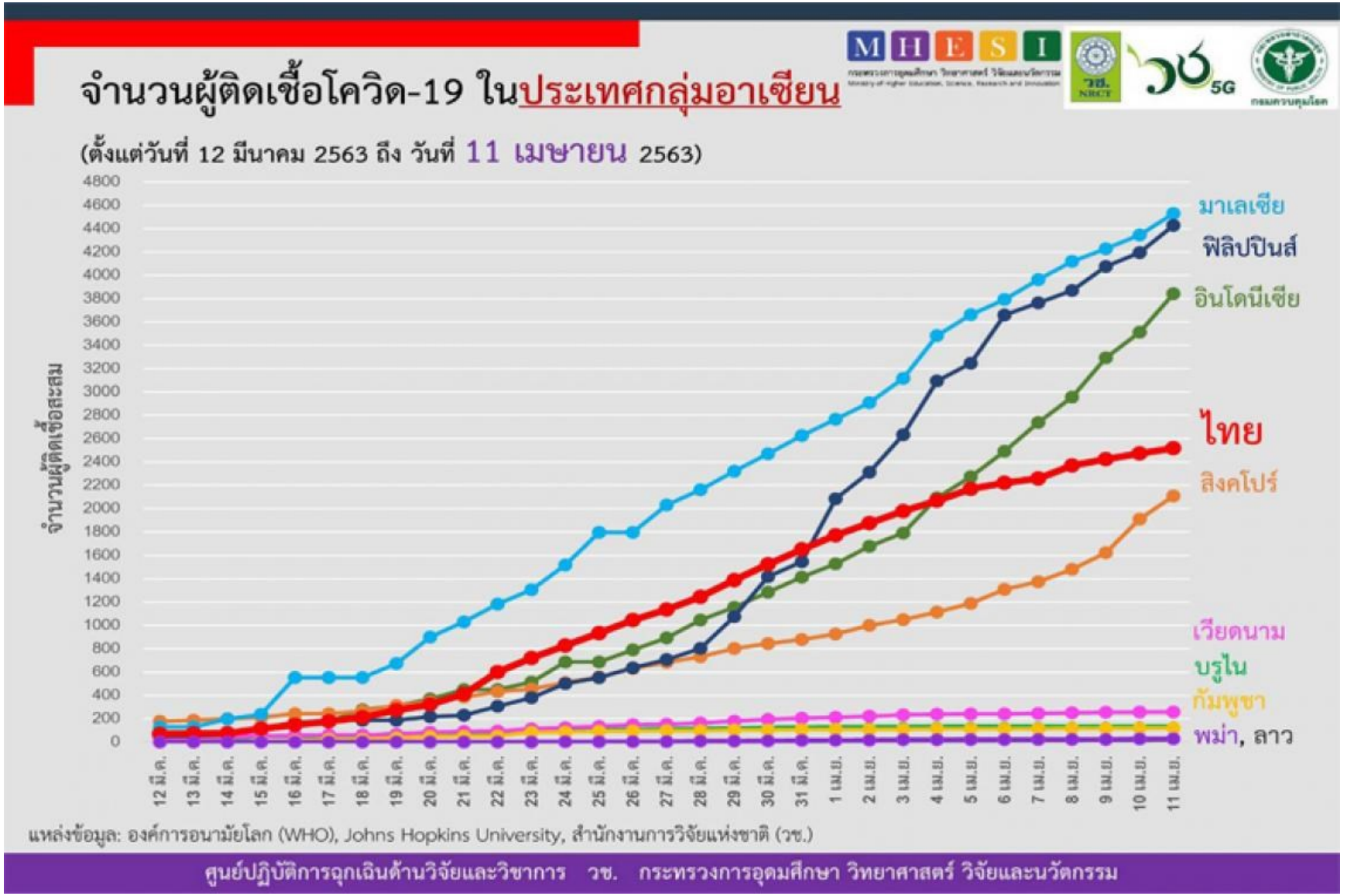
การวิเคราะห์แนวโน้มหรือวิวัฒนาการข้อมูล

- เป็นเทคนิควิทยการข้อมูลที่เกี่ยวกับเวลา เพื่อบรรยายและสร้างแบบจำลองของความสม่ำเสมอ หรือแนวโน้มของวัตถุซึ่งมีพฤติกรรมเปลี่ยนแปลงไปตามเวลา โดยช่วยทำนายแนวโน้มในอนาคต เช่น ราคาหุ้น
  - Regression

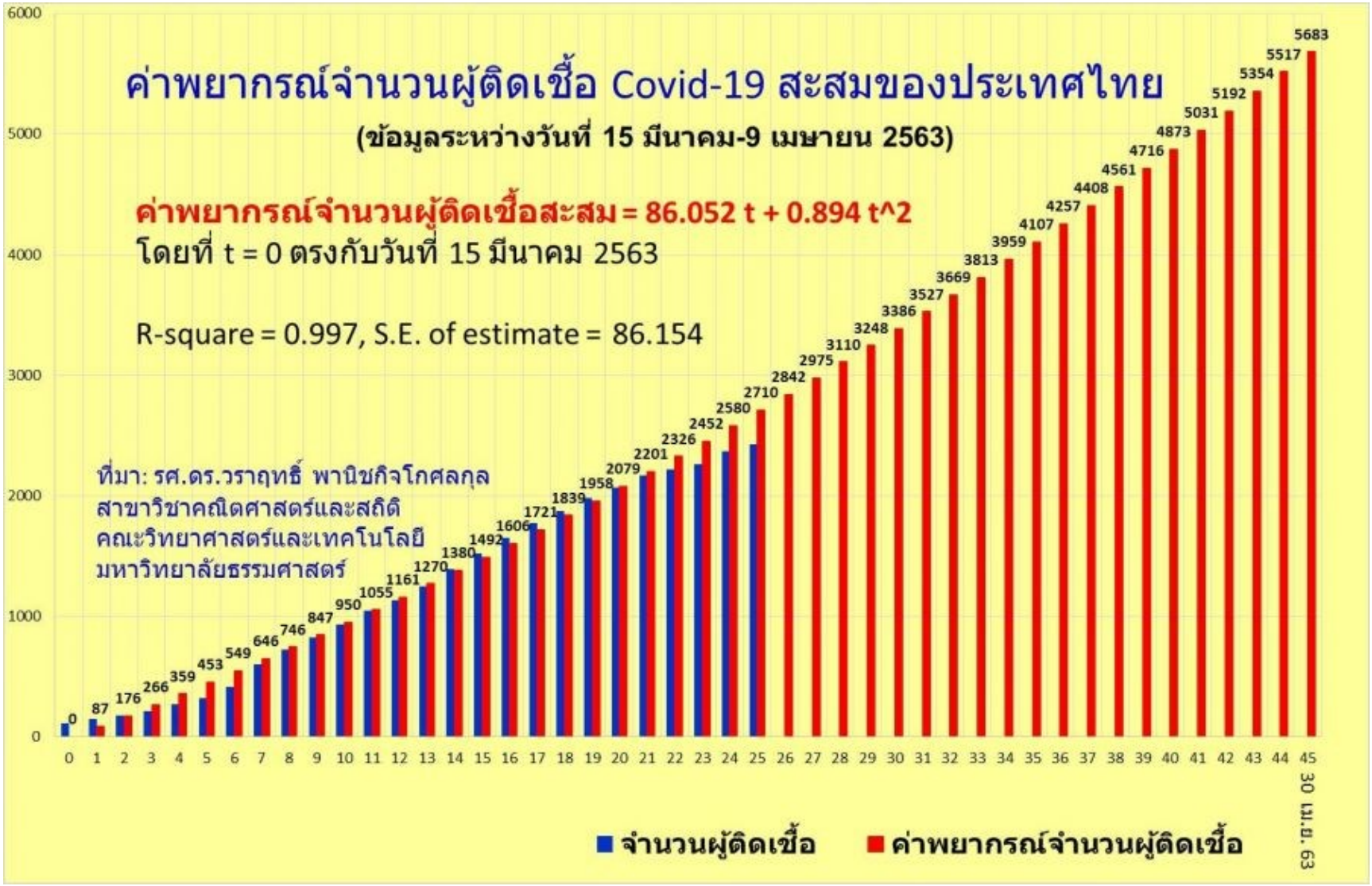


ตัวอย่าง: เทคนิคการวิเคราะห์แนวโน้ม (regression)

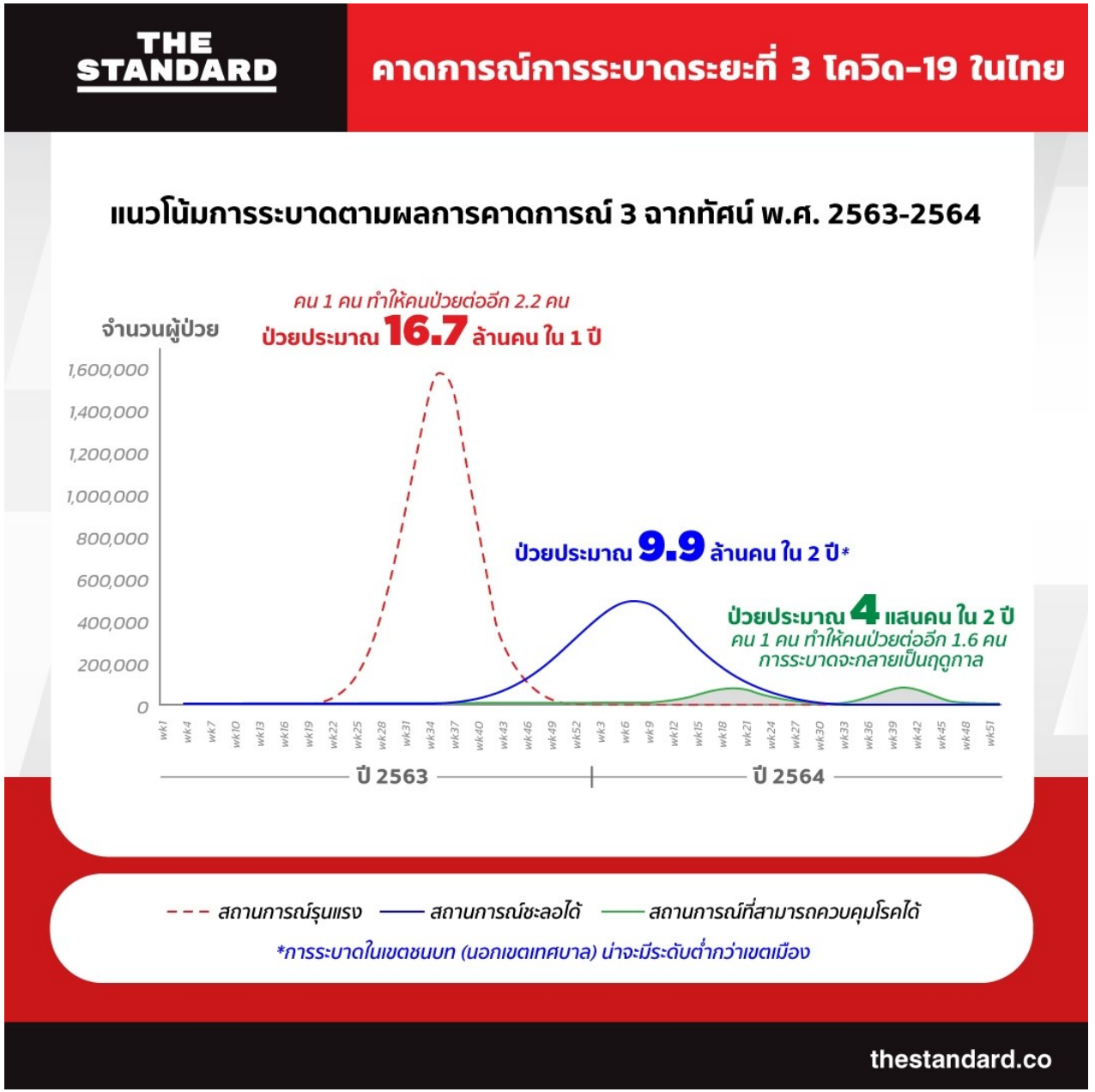
ช่วงโควิด นักศึกษาคุ้นกับแผนภูมินี้กันบ้างไหม



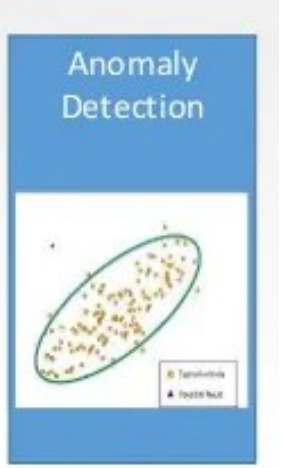
ตัวอย่าง: เทคนิคการพยากรณ์ (Prediction)



ตัวอย่าง: เทคนิคการพยากรณ์ (Prediction)



เทคนิควิทยาการข้อมูลที่ประยุกต์กับงานด้านอื่นๆ



การวิเคราะห์ข้อมูลผิดปกติ

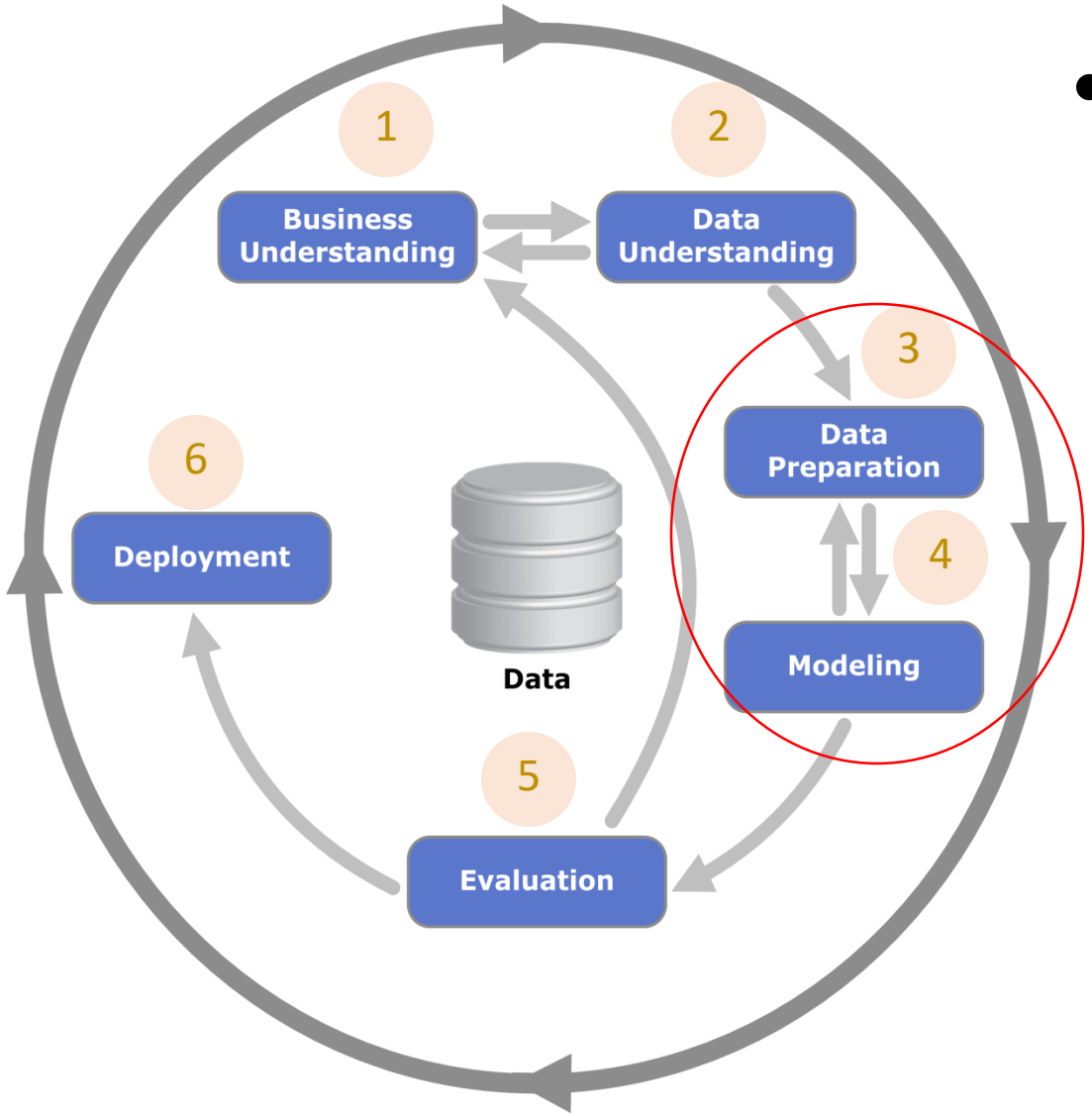
- ปกติข้อมูลที่มีค่าสูงหรือค่าต่ำกว่าผิดปกติ มักจะถูกเป็นข้อมูลรบกวน แต่บางกรณีมักจะมีประโยชน์ เช่น

... 2009		2010	
monthly	Payment (baht)	monthly	Payment (baht)
1	25,000.00	1	10,000.00
2	30,000.00	2	15,000.00
3	17,000.00	3	1,500,000.00
..	...		
12	23,500.00		

Outlier value can be detected

-Location
-Type of purchase
-Purchase frequency

ขั้นตอนของ CRISP-DM

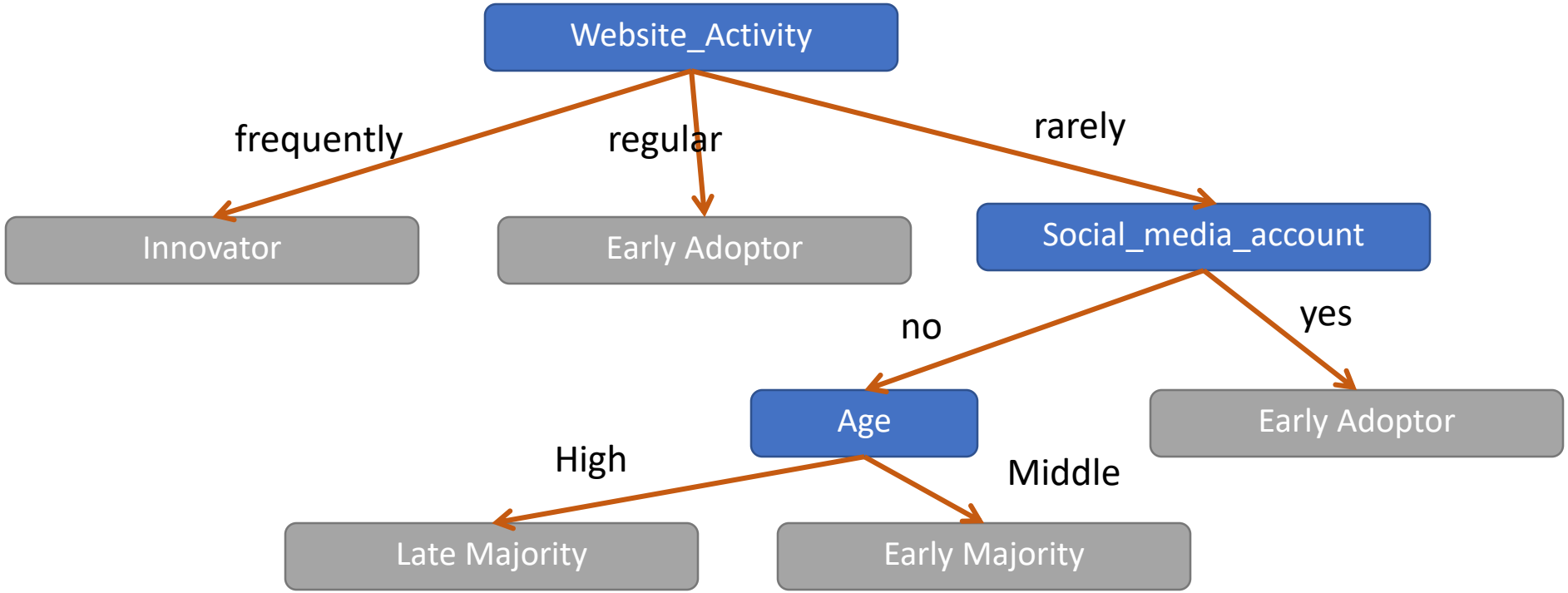


- Data preparation & Modeling

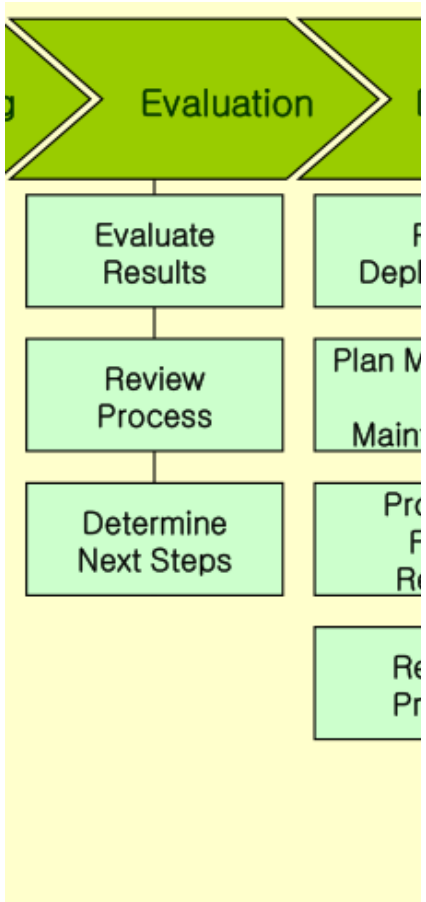
บางเทคนิคมีความเฉพาะเจาะจงกับรูปแบบของข้อมูล ดังนั้นสามารถย้อนกลับไปขั้นตอนการจัดเตรียมข้อมูลเพื่อแปลงข้อมูลบางส่วนให้เหมาะสมกับเทคนิค

ตัวอย่าง ขั้นตอนที่ 4 Modeling

- กรณีบริษัทจำหน่าย smart phone
- แบ่งข้อมูลเป็น 2 ส่วน ส่วนแรก 70 % นำมาใช้สร้างตัวแบบด้วยเทคนิคต้นไม้ช่วยการตัดสินใจ (decision tree) (จะได้เรียนในหัวข้อต่อๆไป)

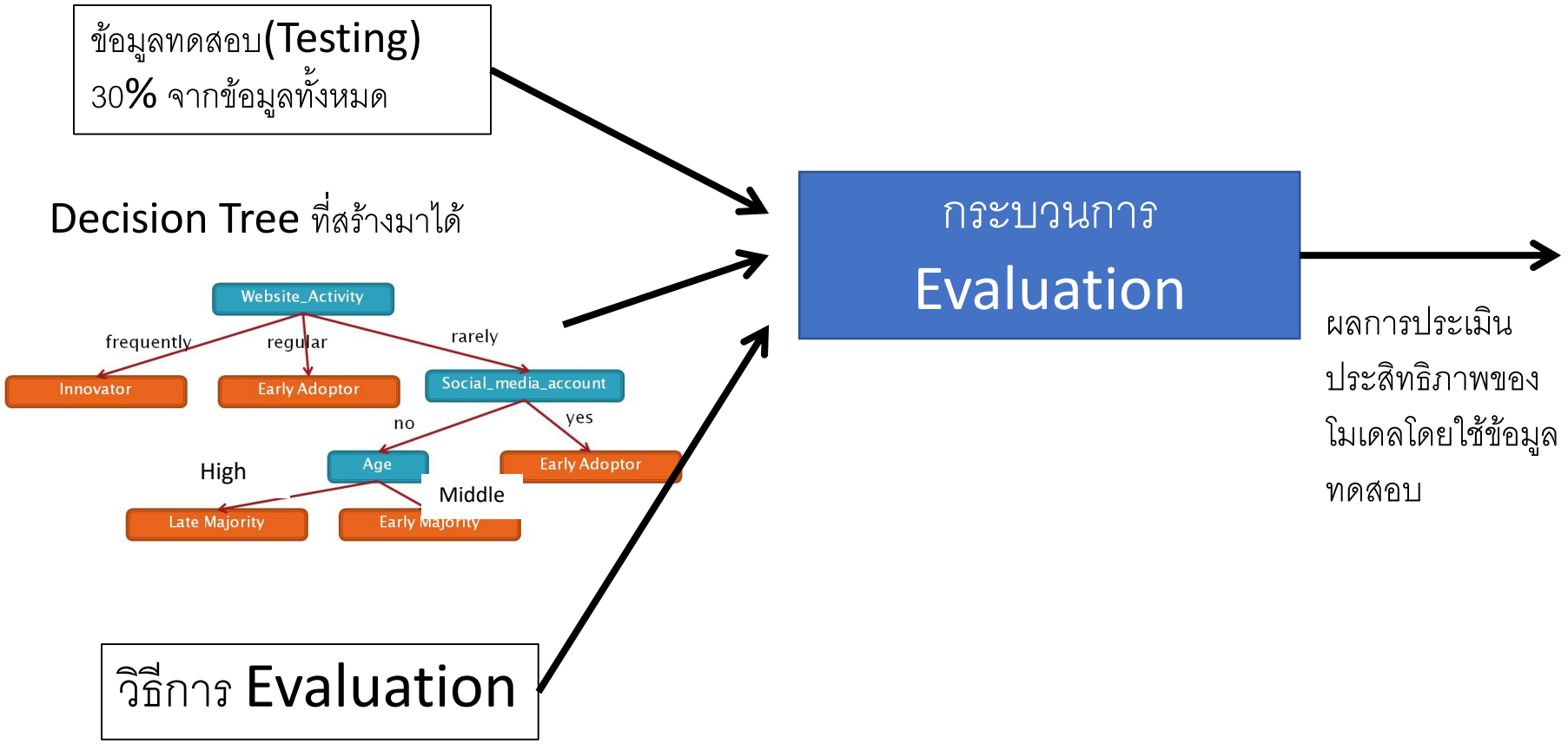


5. Evaluation

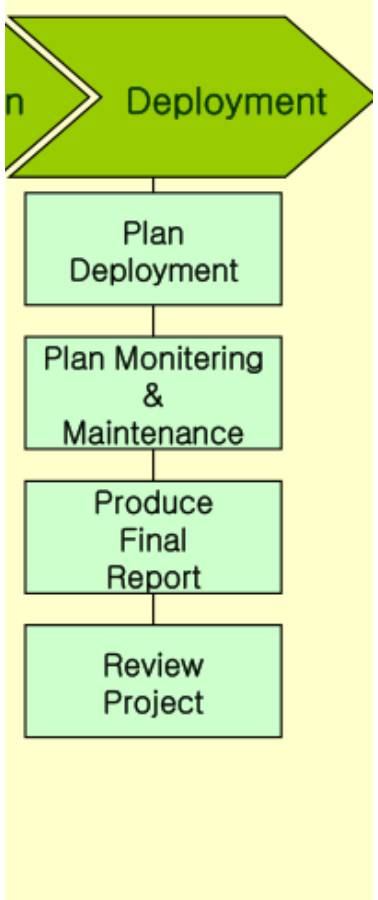


- 5.1 การตรวจสอบตัวแบบว่าทำงานได้ผลลัพธ์ (Result) เป็นอย่างไรด้วยข้อมูลทดสอบที่ใส่เข้าไป (Testing data)
- 5.2 วิธีทดสอบขึ้นกับประเภทของปัญหาและตัวแบบ
  - อาจเป็นการใช้ผู้เชียวชาญ (expert) เป็นผู้ประเมิน (Review)
  - การประเมินโดยใช้มุมมองทางธุรกิจ (Business Perspective)
    - ใช้การนิยามกลุ่มควบคุม (control groups)
    - ผลตอบแทนการลงทุน (expected return on investment)
- 5.3 เพื่อได้ข้อมูลที่นำไปสู่ขั้นถัดไป (Next Step)
  - ศักยภาพการนำไปใช้จริง
  - สถาปัตยกรรมที่ใช้เมื่อนำไปใช้จริง
  - ตัววัดความสำเร็จหลังจากนำไปใช้

ตัวอย่างการทำ Evaluation



6. Deployment



- การนำตัวแบบ (model) หรือ องค์ความรู้ที่ได้จากตัวแบบ (Knowledge) ไปใช้งาน (deploy) กับงานที่เฉพาะเจาะจงกับวัตถุประสงค์
  - 6.1 ต้องมีแผนกลยุทธ์ในการนำไปใช้ (strategy)
  - 6.2 การสอดส่องดูผลลัพธ์ (Monitoring) และ การบำรุงรักษา (Maintenance)
  - 6.3 เป็นรายงานความรู้ (Report)
    - เสนอในการประชุมของบริษัท
    - เสนอเพื่อออกโปรโมชั่นใหม่
    - บทความงานวิจัยใหม่
  - ต้องมีการสร้างรายงาน
  - มีการทำเอกสารประกอบการทำงาน (final report) ทุกขั้นตอน
  - มีแผนในการควบคุม (monitoring) และ ดูแล (maintenance)
    - 6.4 Review project

ตัวอย่างการทำ Deployment

งานวิจัยเรื่อง “การสร้างแบบจำลองรายการธุรกรรมผิดปกติของบัญชีเงินฝากออมทรัพย์ผู้สูงวัยโดยการทำเหมืองข้อมูล กรณีศึกษาธนาคารพาณิชย์แห่งหนึ่ง” ได้ *นำเสนอเทคนิคการคัดกรองข้อมูล และ เทคนิคการสร้างแบบจำลอง*

- ประยุกต์ใช้ในการคัดกรองข้อมูล เพื่อ *การวางแผนตรวจสอบรายการธุรกรรมบัญชีเงินฝากออมทรัพย์ของผู้สูงวัยที่ผิดปกติขององค์กร และทราบถึงลักษณะพฤติกรรมของกลุ่มธุรกรรมผิดปกติ*
- เป็นส่วนสำคัญในการ *ป้องกันความเสียหายที่จะเกิดขึ้นกับองค์กร* ทั้งในด้านที่เป็นตัวเงินและชื่อเสียงขององค์กร
- ด้วยการศึกษาจากปัจจัยต่างๆ ดังที่ได้กล่าวมา ทำให้ระบบช่วย *สร้างความเชื่อมั่นในการทำรายการธุรกรรม* ให้กับลูกค้าได้เป็นอย่างดี

อ้างอิง

บทความวิชาการเรื่อง “การสร้างแบบจำลองรายการธุรกรรมผิดปกติของบัญชีเงินฝากออมทรัพย์ผู้สูงวัยโดยการทำเหมืองข้อมูล กรณีศึกษาธนาคารพาณิชย์แห่งหนึ่ง” โดย ปรียานุช สมัครการ และ กมล เกียรติเรืองกมลา

ตัวอย่างการทำ Deployment

งานวิจัยเรื่อง “ระบบผู้เชี่ยวชาญสำหรับคัดแยกโรคพืชในไม้ผล : มะม่วงน้ำดอกไม้” ได้ *นำเสนอเทคนิคการสร้างแบบจำลอง และ การประยุกต์ใช้แบบจำลอง*

- ประยุกต์ใช้ในกาจำแนกข้อมูลอาการใบบนพืชอย่างมะม่วงน้ำดอกไม้สีทอง ด้วยเทคนิคต้นไม้การตัดสินใจ (Decision Tree) ทำให้สามารถระบุโรคพืชได้ เพื่อ *การวางแผนการรักษาโรคพืชและควบคุมการระบาดของโรคทำให้ลดการสูญเสียพืชและผลผลิต ช่วยควบคุมค่าใช้จ่ายการเพาะปลูกได้*
- เป็นส่วนสำคัญในการ *ป้องกันความเสียหายที่จะเกิดขึ้นกับแปลงปลูก และผลผลิตส่วนใหญ่ยังคงมีคุณภาพและขายได้ราคา*
- ด้วยการศึกษาจากปัจจัยอาการบนใบจากองค์ความรู้ทางการเกษตร ทำให้ระบบช่วย *สร้างความเชื่อมั่นในการเพาะ* ให้กับเกษตรกรได้เป็นอย่างดี

อ้างอิง

บทความวิชาการเรื่อง “Expert System for Classification of Plant Disease In Fruit Plant: Barrracuda Mango” โดย ชุตินันท์ ตรงต่อกิจ และ พาสน์ ปราโมกข์ชน ICDAMT & NCON2018

ตัวอย่างการทำ Deployment

งานวิจัยเรื่อง “ระบบสนับสนุนการตัดสินใจเพื่อการลงทะเบียนของนักศึกษาสาขาวิทยาการคอมพิวเตอร์” ได้ *นำเสนอเทคนิคการหาความสัมพันธ์ และ การประยุกต์ใช้แบบจำลองความสัมพันธ์*

- ประยุกต์ใช้เทคนิคการหาความสัมพันธ์สร้างกฎความสัมพันธ์ระหว่างรายวิชาและผลการเรียน เพื่อ แนะนำนักศึกษาลงทะเบียนรายวิชาที่เหมาะสมกับทักษะความรู้ในอดีตทำให้นักศึกษาสามารถได้ผลการเรียนที่ดี
- เป็นส่วนสำคัญในการ ช่วยให้คำแนะนำที่ทำให้นักศึกษาเลือกเรียนในสิ่งที่ตัวเองถนัด และได้ผลการเรียนที่ดี
- ด้วยการศึกษาจากปัจจัยต่างๆ ดังที่ได้กล่าวมา ทำให้ระบบเหมือนเป็นที่ปรึกษาในการลงทะเบียนเรียนรายวิชาช่วย *สร้างความเชื่อมั่นในการทำเรียนในมหาวิทยาลัย* ให้กับนักศึกษาได้เป็นอย่างดี

อ้างอิง

บทความวิชาการเรื่อง “DECISION SUPPORT SYSTEM FOR SUBJECTS REGISTRATIONOF COMPUTER SCIENCE STUDENT” โดย ณรัฐศภรณ์ เหมรา และ พาสน์ ปราโมกข์ชน

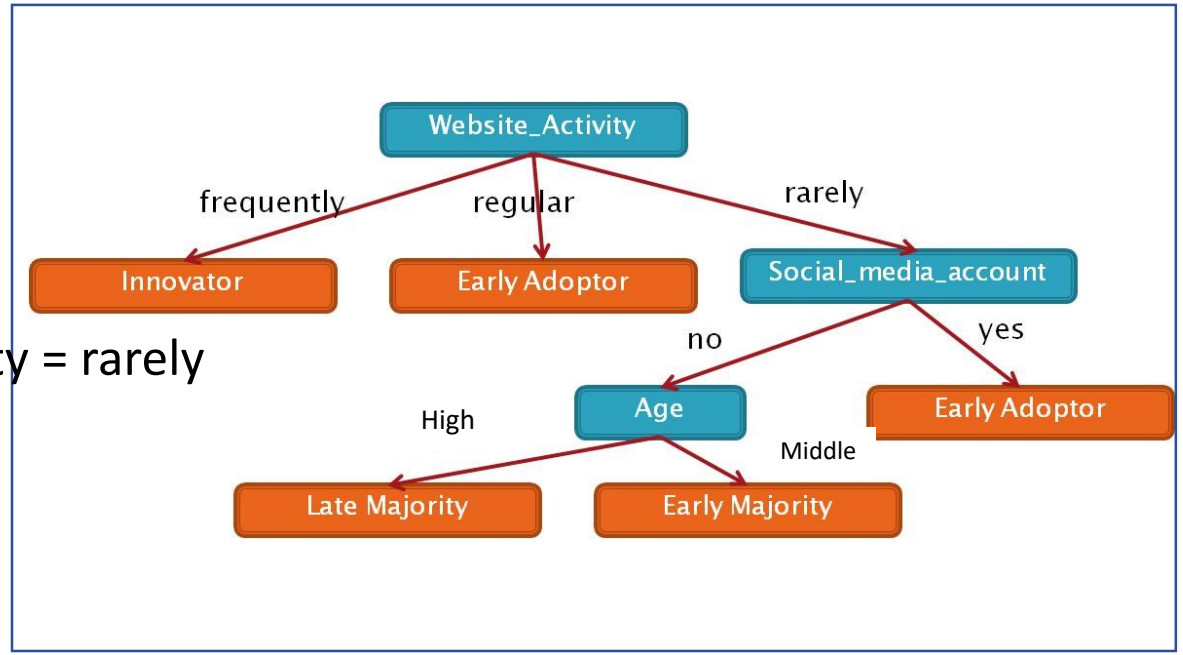
ตัวอย่าง Deployment

- นำข้อมูลของลูกค้าที่มีอยู่มาจำแนกกลุ่มว่าน่าจะเป็นลูกค้ากลุ่มใด
- แล้วส่งโปรโมชั่นที่เหมาะสมกับลูกค้าไปในช่วงเวลาที่เหมาะสม

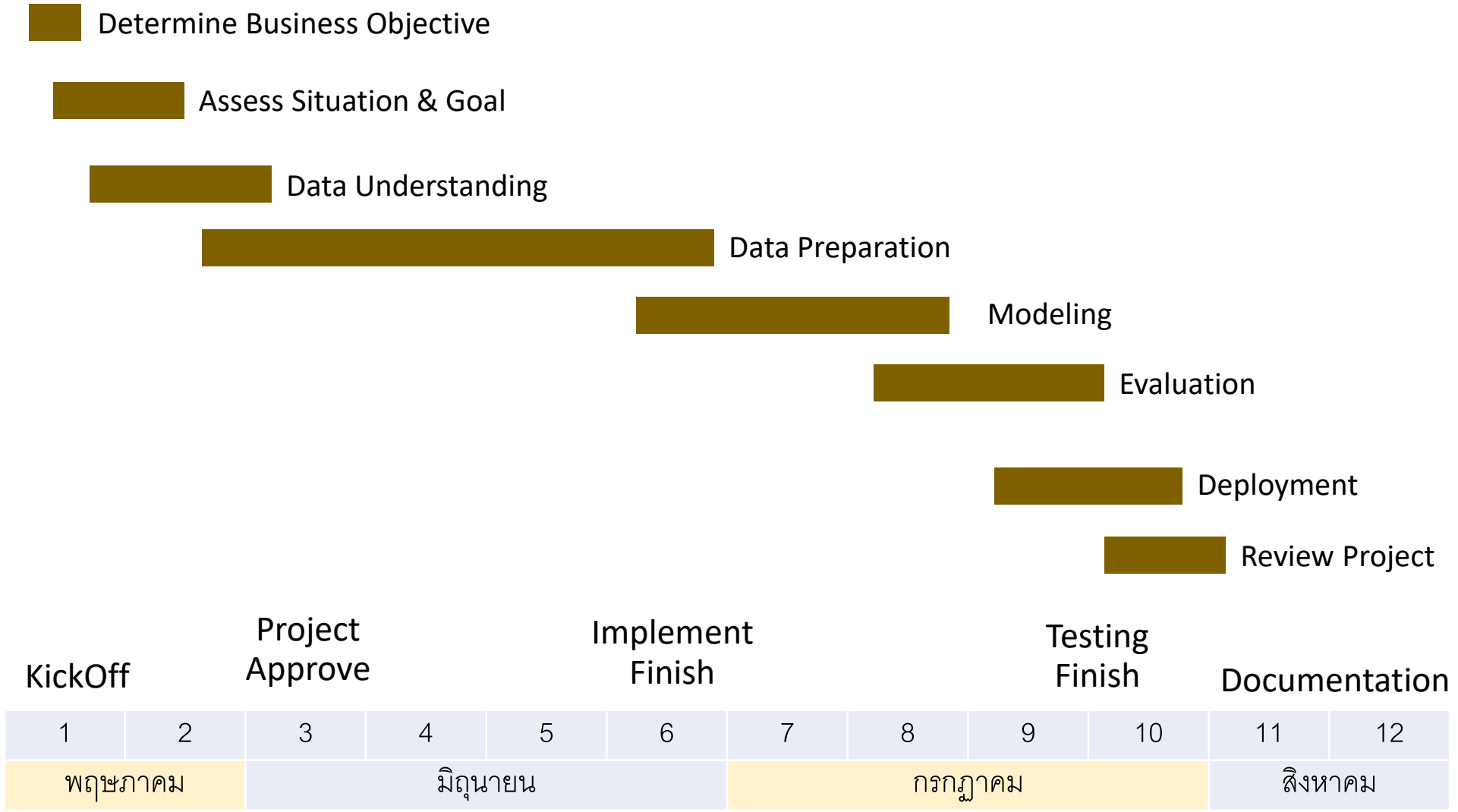
Decision Tree ที่สร้างมาได้

ข้อมูลลูกค้า
รายใหม่

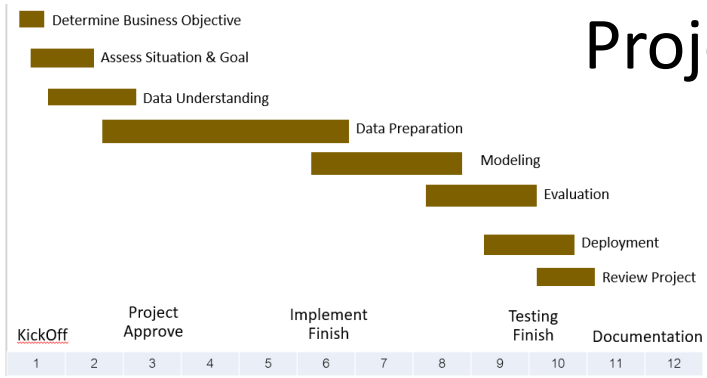
Website_Activity = rarely
Age > 25



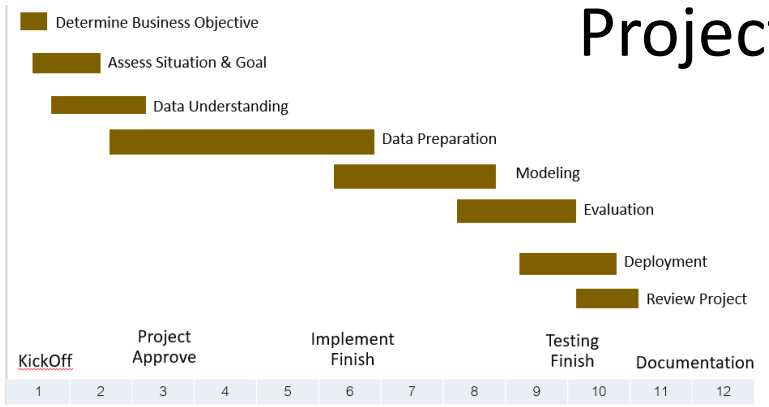
ผลการจำแนกลูกค้า
รายใหม่
เพื่อนำไปใช้ส่ง
โปรโมชั่นที่เหมาะสม


Determine Business Objective
Assess Situation & Goal
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Review Project
KickOff
Project Approve
Implement Finish
Testing Finish
Documentation
1
2
3
4
5
6
7
8
9
10
11
12
พฤษภาคม
มิถุนายน
กรกฎาคม
สิงหาคม

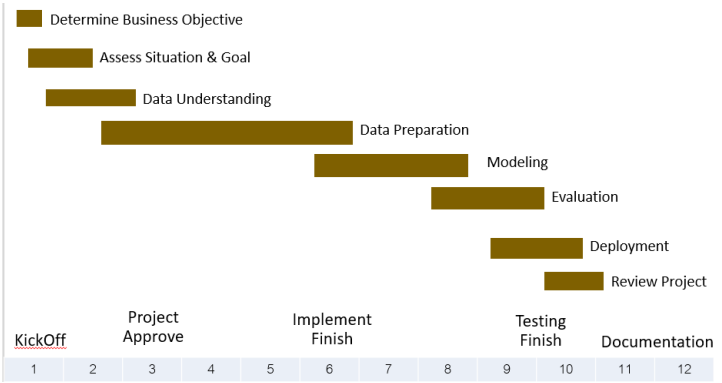
Project 1Phase1



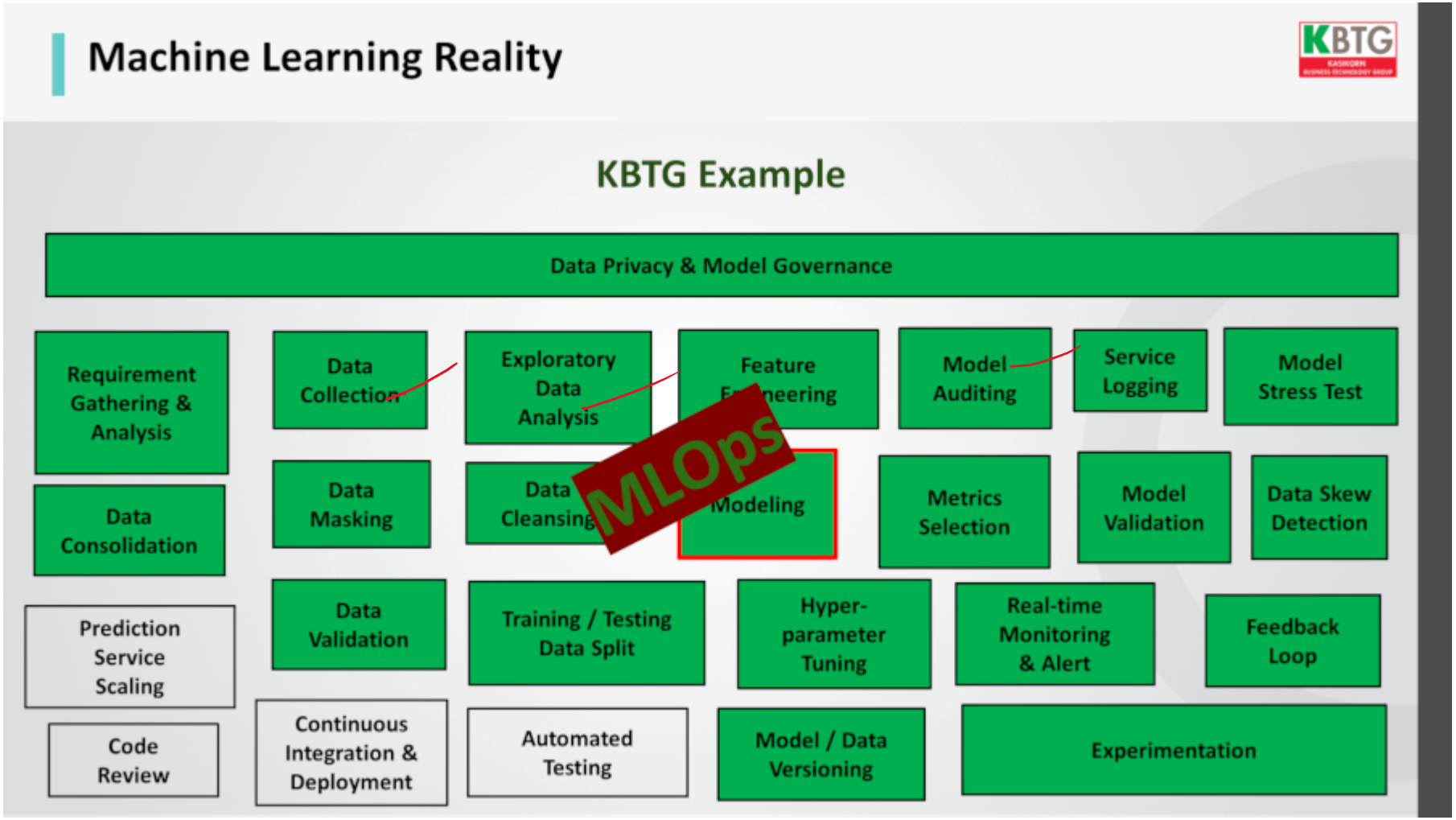
Project1 Phase2



Sprint to Project2 Phase1



CRISP-DS ไม่ได้เป็นแค่ Framework เดียวใน Data Science process



END

Q/A