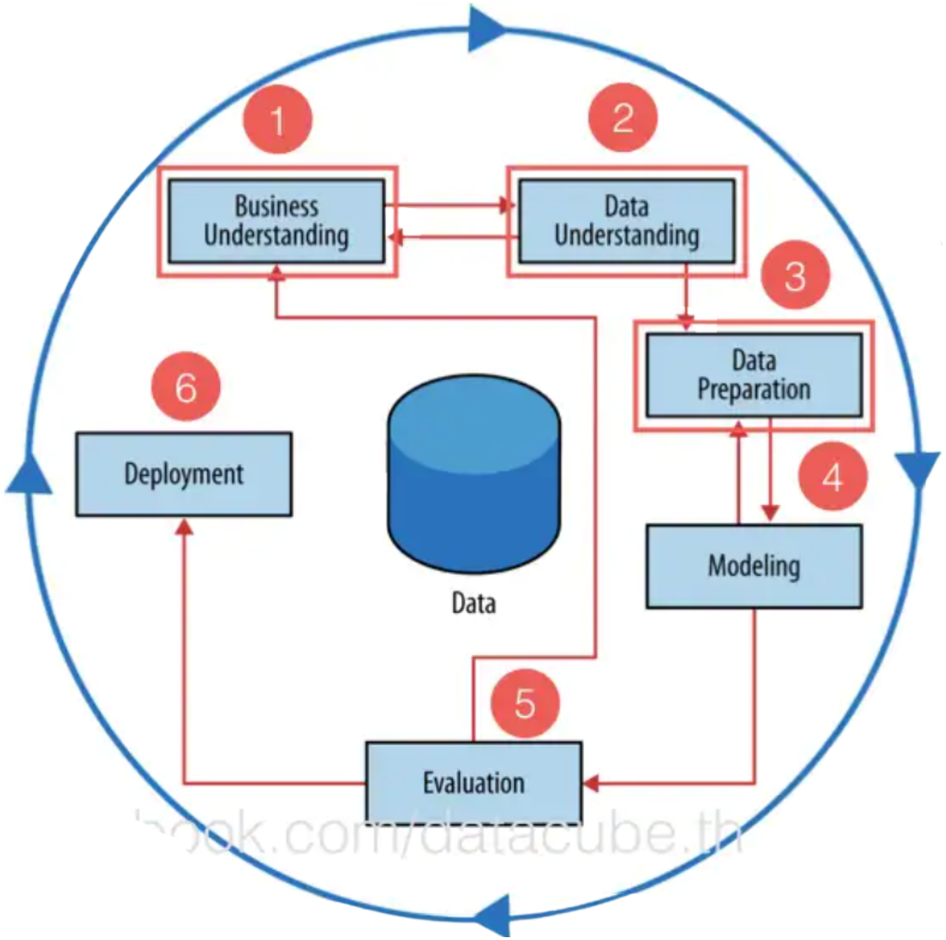
Chapter 2

Introduction to CRISP - DM

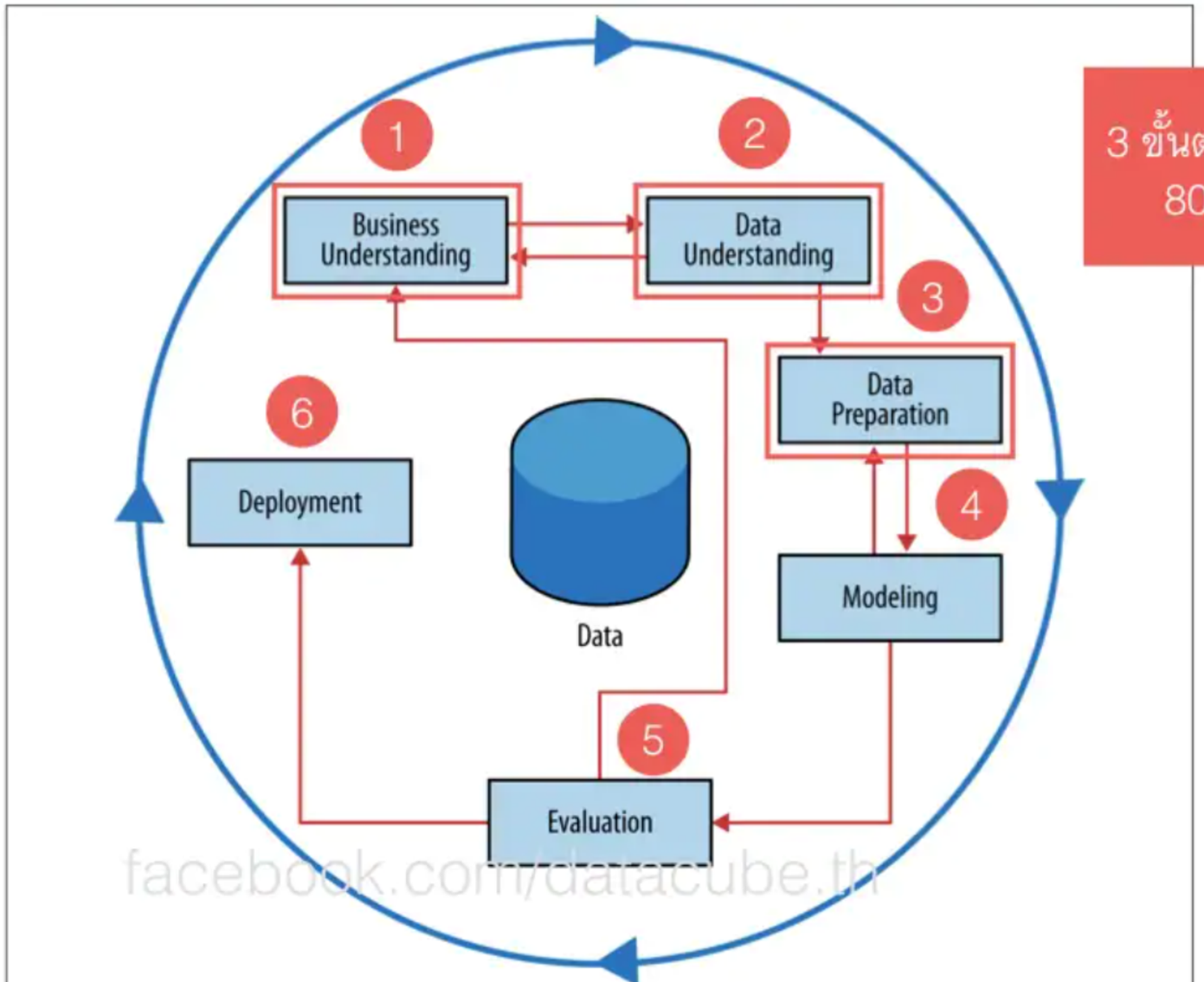
CRISP - DM

- CRoss-Industry Standard Process for Data Mining (CRISP-DM)
- พัฒนาขึ้นโดย 3 บริษัท
  - บริษัท SPSS
  - บริษัท DaimlerChrysler
  - บริษัท NCR
- เป็น Workflow มาตรฐานสำหรับการทำ data mining
- ประกอบด้วย 6 ขั้นตอน

CRISP - DM


1
Business Understanding
2
Data Understanding
3
Data Preparation
4
Modeling
5
Evaluation
6
Deployment
Data

CRISP - DM



3 ขั้นตอนแรกจะใช้เวลา
80% ของทั้งหมด

1. Business Understanding

- ขั้นตอนแรกของ CRISP-DM
  - ทำความเข้าใจกับปัญหา หรือ โอกาสเชิงธุรกิจ
  - ระบุ output หรือเป้าหมายที่ต้องการได้จากการวิเคราะห์ข้อมูลด้วย data mining
  - ตัวอย่างเช่น
    - ทำอย่างไรถึงเพิ่มยอดขายให้กับสินค้าชนิดต่างๆ ได้
    - ต้องการแบ่งกลุ่มนักศึกษาออกตามความสนใจ
    - ทำอย่างไรให้ลูกค้ากลับมาซื้อสินค้าได้อีก
    - อยากทำนายปริมาณน้ำฝนที่ตกใน 2 วันถัดไป
    - อยากรู้ว่าลูกค้าคนใดบ้างมีโอกาสป่วยเป็นโรคมะเร็ง

2. Data Understanding

- ในขั้นตอนนี้เป็นการ
  - รวบรวมข้อมูลที่เกี่ยวข้อง
  - ข้อมูลถูกต้องน่าเชื่อถือ
  - ข้อมูลที่ได้มีปริมาณมากพอหรือยัง
  - ข้อมูลที่ได้มีความเหมาะสม มีรายละเอียดเพียงพอต่อการนำไปใช้ในการวิเคราะห์
- ตัวอย่าง
  - ข้อมูลการซื้อสินค้าคนแต่ละบุคคล
  - ข้อมูลการลงทะเบียนและผลการศึกษาของนักศึกษา

3. Data Preparation

- ขั้นตอนการเตรียมข้อมูลเป็นขั้นตอนที่ใช้เวลานานที่สุด
- เนื่องจากโมเดลที่ได้จากการทำดาต้าไมน์นิ่งจะให้ผลลัพทธ์ที่ถูกต้องหรือไม่นั้น ขึ้นอยู่กับคุณภาพของข้อมูลที่ใช้ แบ่งออกได้เป็น 3 ขั้นตอนย่อยคือ
- 3.1 ทำการคัดเลือกข้อมูล (Data Selection)
  - กำหนดเป้าหมายก่อนว่าเราจะทำการวิเคราะห์อะไร
  - เลือกใช้เฉพาะข้อมูลที่เกี่ยวข้องกับสิ่งที่เราจะทำการวิเคราะห์

3. Data Preparation

- 3.2 ทำการกลั่นกรองข้อมูล (Data Cleaning)
  - ลบข้อมูลซ้ำซ้อน
  - แก้ไขข้อมูลที่ผิดพลาด
    - ข้อมูลผิดรูปแบบ
    - ข้อมูลที่หายไป
    - ข้อมูล outlier ที่แปลกแยกจากคนอื่น

ข้อมูลนักศึกษาชั้นปีที่ 1 ปีการศึกษา 2557

รหัส	เพศ	อายุ	ความสูง	น้ำหนัก
57001	ชาย	18	180	70
5702A	ญ		80	35
57123	หญิง	19	150	2500
58002	ช	17	175	90

ผิดรูปแบบ ← ขาดหาย → outlier

3. Data Preparation

- 3.3 แปลงรูปแบบของข้อมูล (data transformation)
  - เป็นขั้นตอนการเตรียมข้อมูลให้อยู่ในรูปแบบที่พร้อมนำไปใช้ในการวิเคราะห์ตามอัลกอริทึมของ data mining ที่เลือกใช้

ID	สินค้า	จำนวนที่ซื้อ
1	ปากกา	1
1	ยางลบ	1
1	คลิป	10
2	สมุด	2
2	ปากกา	2
3	สมุด	1
3	ปากกา	3
3	ยางลบ	2

ID	สมุด	ปากกา	ยางลบ	คลิป
1	-	TRUE	TRUE	TRUE
2	TRUE	TRUE	-	-
3	TRUE	TRUE	TRUE	-

ข้อมูลสำหรับการหากฎความสัมพันธ์ (Association Rules)

3. Data Preparation

Business Understanding → Data Understanding → **Data Preparation** → Modeling → Evaluation → Deployment

- 3.3 **แปลงรูปแบบของข้อมูล** (data transformation)
  - เป็นขั้นตอนการเตรียมข้อมูลให้อยู่ในรูปแบบที่พร้อมนำไปใช้ในการวิเคราะห์ตามอัลกอริทึมของ data mining ที่เลือกใช้

เมื่อวันที่ 4 มกราคม 2557 เฟซบุ๊กเปิดตัวหน้าเพจใหม่ชื่อว่า Facebook A Look Back เมื่อผู้ใช้งานคลิกไปยังหน้านี้ก็จะแสดงคลิปวิดีโอที่บอกเล่าเรื่องราวของผู้ใช้งานคนนั้นๆ เช่น เริ่มเล่นเฟซบุ๊กครั้งแรกปีไหน, โพสต์แรกบนเฟซบุ๊ก, รูปภาพที่ถูกกดไลค์มากที่สุด, รูปภาพที่ถูกแชร์มากที่สุด และ 20 อันดับเรื่องราวต่าง ๆ ที่เกิดขึ้นในเฟซบุ๊กก็จะถูกแสดงและรวบรวมไว้ในคลิปวิดีโอนี้

เอกสารข่าว

ID	เฟซบุ๊ก	รูปภาพ	ไลค์	แชร์	คลิปวิดีโอ
1	4	2	1	1	2
2	…				

ตารางแสดงจำนวนความถี่ของแต่ละคำ

3. Data Preparation

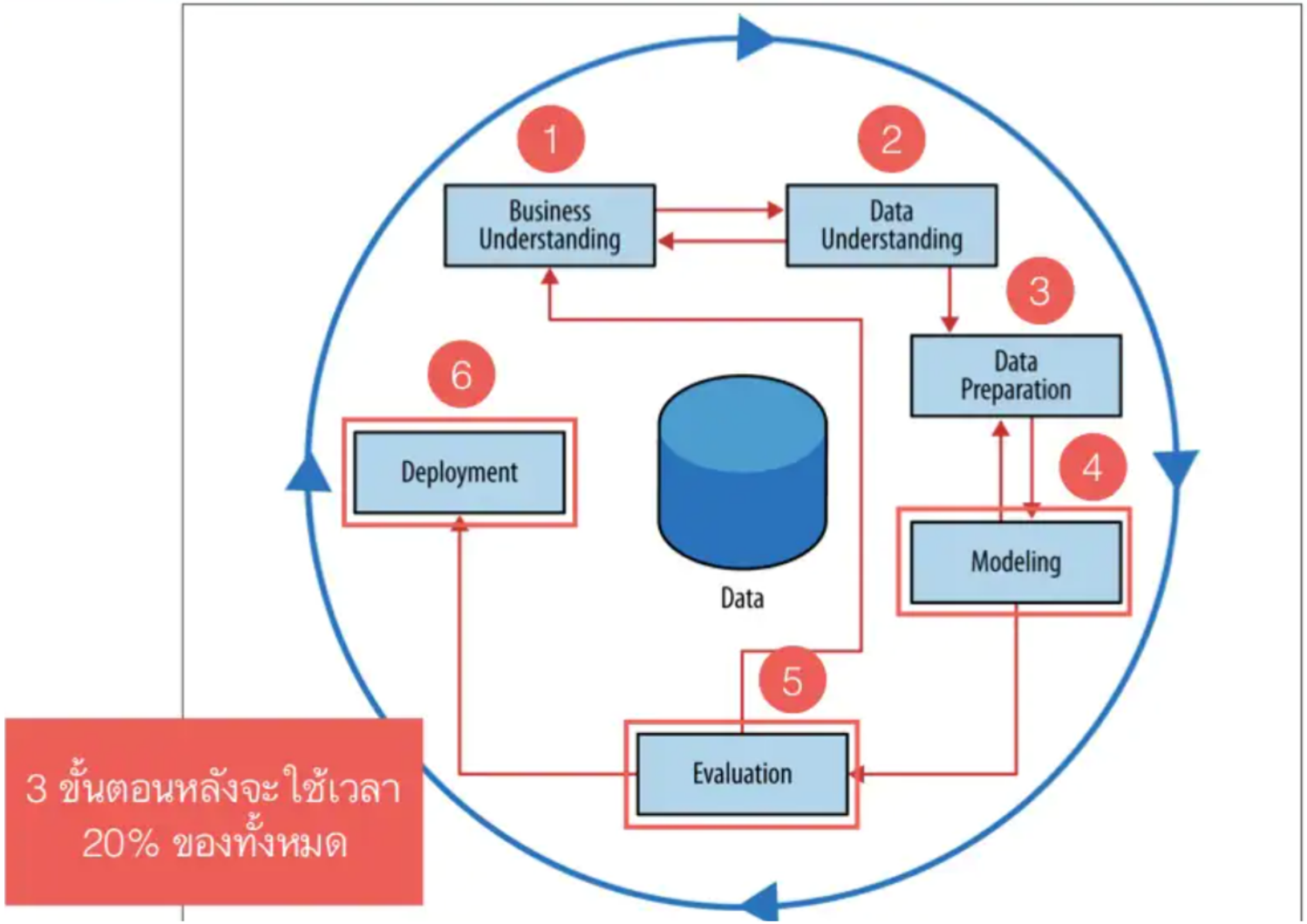
Business Understanding → Data Understanding → **Data Preparation** → Modeling → Evaluation → Deployment

- 3.3 **แปลงรูปแบบของข้อมูล** (data transformation)
  - เป็นขั้นตอนการเตรียมข้อมูลให้อยู่ในรูปแบบที่พร้อมนำไปใช้ในการวิเคราะห์ตามอัลกอริทึมของ data mining ที่เลือกใช้

ID	สีแดง	สีเขียว	สีน้ำเงิน
1	93	98	167
2	…		

จำนวน pixel สีแดง สีเขียว สีน้ำเงินที่ปรากฏ ในรูปภาพ

CRISP - DM

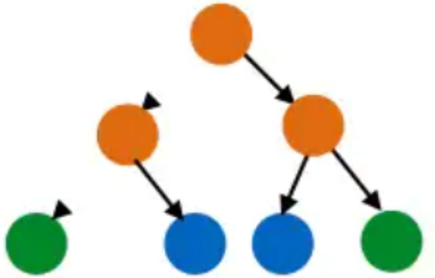


3 ขั้นตอนหลังจะใช้เวลา
20% ของทั้งหมด

4. Modeling

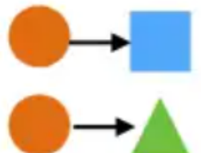
Business Understanding → Data Understanding → Data Preparation → **Modeling** → Evaluation → Deployment

- เป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคดาต้าไมน์นิ่ง
  - classification
    - สร้างโมเดลจากข้อมูลที่มีอยู่เพื่อทำนายอนาคต
    - เช่น ทำนายปริมาณน้ำฝนที่ตกในวันถัดไป
  - clustering
    - แบ่งข้อมูลหลายๆ กลุ่มตามความคล้ายคลึง
    - เช่น แบ่งกลุ่มนักศึกษาตามคะแนนที่ได้
  - association rules
    - หาความสัมพันธ์ของข้อมูลที่เกิดร่วมกัน
    - เช่น ค้นหาสินค้าที่มีการซื้อร่วมกันบ่อยๆ



classification

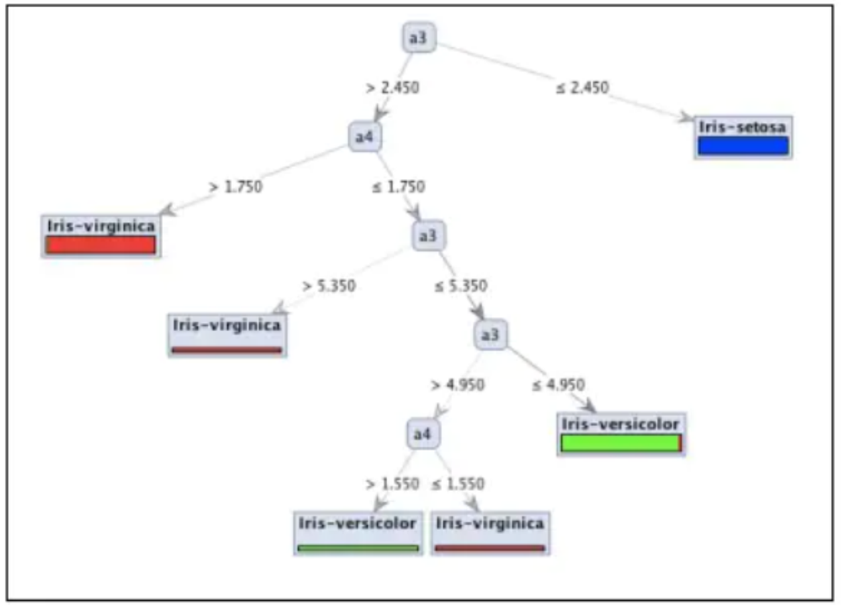
clustering



5. Evaluation

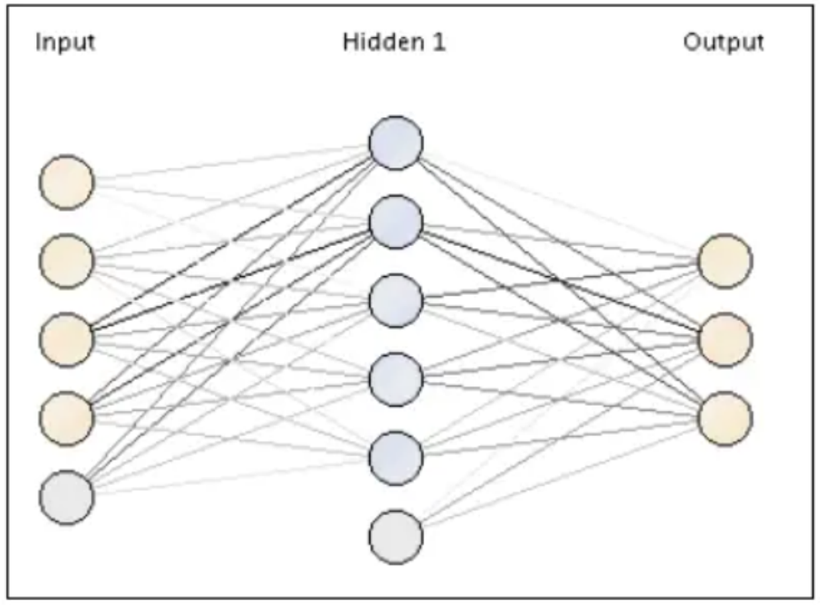
Business Understanding → Data Understanding → Data Preparation → Modeling → **Evaluation** → Deployment

- ประเมินหรือวัดประสิทธิภาพของโมเดลวิเคราะห์ข้อมูลในขั้นตอนก่อนหน้านั้น



โมเดล decision tree

VS



โมเดล neural network

6. Deployment

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → **Deployment**

- นำโมเดลที่ได้ หรือ ผลการวิเคราะห์ที่ได้ไปใช้งานจริง

CRISP - DM Example 1

- ตัวอย่าง CRISP-DM
  - อ้างอิงจากงานวิจัยเรื่อง **การใช้เทคนิคดาต้าไมน์นิงเพื่อพัฒนาคุณภาพการศึกษานิสิตคณะวิศวกรรมศาสตร์** *

1. Business Understanding

- นิสิตคณะวิศวกรรมศาสตร์ ม.เกษตรศาสตร์ จะเลือกภาควิชาเมื่อในชั้นปีที่ 2
- นิสิตเลือกภาควิชาไม่ตรงกับความสามารถของตนเอง
  - เลือกตามเพื่อน
  - เลือกตามที่ผู้ปกครองแนะนำ
- นิสิตบางคนได้ผลการเรียนตกต่ำและทำให้ต้องออกจากมหาวิทยาลัยกลางคัน

CRISP - DM Example 1

2. Data Understanding

- ข้อมูลนิสิตคณะวิศวกรรมศาสตร์ ม.เกษตรศาสตร์ช่วงปี พ.ศ. 2535 - 2542
  - นิสิตประมาณ 10,000 คน
  - ข้อมูลมีจำนวน 476,085 แถว
- ข้อมูลแบ่งเป็น 2 ส่วน
  - ข้อมูลประวัติส่วนตัวของนิสิต
    - เพศ, ที่อยู่, GPA ระดับมัธยมปลาย, GPA ชั้นปีที่ 1
  - ข้อมูลการลงทะเบียนของนิสิต
    - เกรดวิชาคณิตศาสตร์, เกรดวิชาฟิสิกส์, เกรดวิชาเคมี

CRISP - DM Example 1

3. Data Preparation

- คัดเลือกวิชาที่เกี่ยวข้องกับภาควิชาต่างๆ ในคณะวิศวกรรมศาสตร์
- แปลงข้อมูลให้เหมาะสมกับการวิเคราะห์

รหัสนิสิต	เพศ	ที่อยู่	GPA มัธยมปลาย	...	GPA
37058063	ชาย	กรุงเทพ	2.5	...	2.3
37058167	ชาย	สงขลา	3.4	...	3.3
...	...	...	...	...	...

ข้อมูลประวัติส่วนตัวของนิสิต

รหัสนิสิต	วิชา	ปีการศึกษา	เกรด
37058063	คอมพิวเตอร์	2537	C+
37058063	คณิตศาสตร์	2537	D
...	...	...	...

ข้อมูลการลงทะเบียนของนิสิต

รหัสนิสิต	เพศ	คอมพิวเตอร์	คณิตศาสตร์	...	GPA
37058063	ชาย	LOW	LOW	...	2.3
37058167	ชาย	HIGH	HIGH	...	3.3
...	...	...	...	...	...

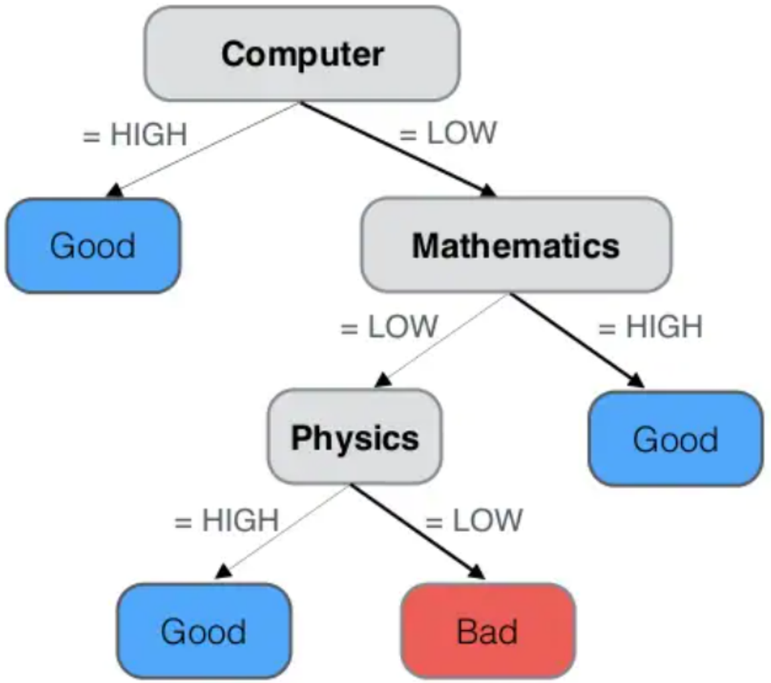
CRISP - DM Example 1

4. Modeling

- แบ่งข้อมูลออกเป็น 2 ส่วน คือ
  - 70% ของข้อมูลทั้งหมดใช้ในการสร้างโมเดล
  - 30% ของข้อมูลทั้งหมดใช้ในการทดสอบประสิทธิภาพของโมเดล
- สร้างโมเดลด้วยเทคนิค Decision Tree ซึ่งจะได้โมเดลที่สามารถเข้าใจได้ง่าย
- โมเดลแบ่งแยกตามภาควิชาต่างๆ เช่น ภาควิชาวิศวกรรมคอมพิวเตอร์ วิศกรรมไฟฟ้า
- คำตอบ (class) จะแบ่งเป็น 2 ประเภท คือ
  - GOOD หมายถึง นิสิตเรียนในภาควิชานี้แล้วจบมาได้ GPA อยู่ในช่วง 40% แรก (top 40%)
  - BAD หมายถึง นิสิตเรียนในภาควิชานี้แล้วจบมาได้ GPA อยู่ในช่วง 40% จากท้าย (bottom 40%)

CRISP - DM Example 1

4. Modeling



โมเดลของภาควิชาวิศวกรรมคอมพิวเตอร์

- **IF** Computer is HIGH **THEN** Graduate is Good
- **IF** Computer is LOW **AND** Mathematics is HIGH **THEN** Graduate is Good
- **IF** Computer is LOW **AND** Mathematics is LOW **AND** Physics is HIGH **THEN** Graduate is Good
- **IF** Computer is LOW **AND** Mathematics is LOW **AND** Physics is LOW **THEN** Graduate is Bad

เงื่อนไขที่สร้างได้จากโมเดล

CRISP - DM Example 1

5. Evaluation

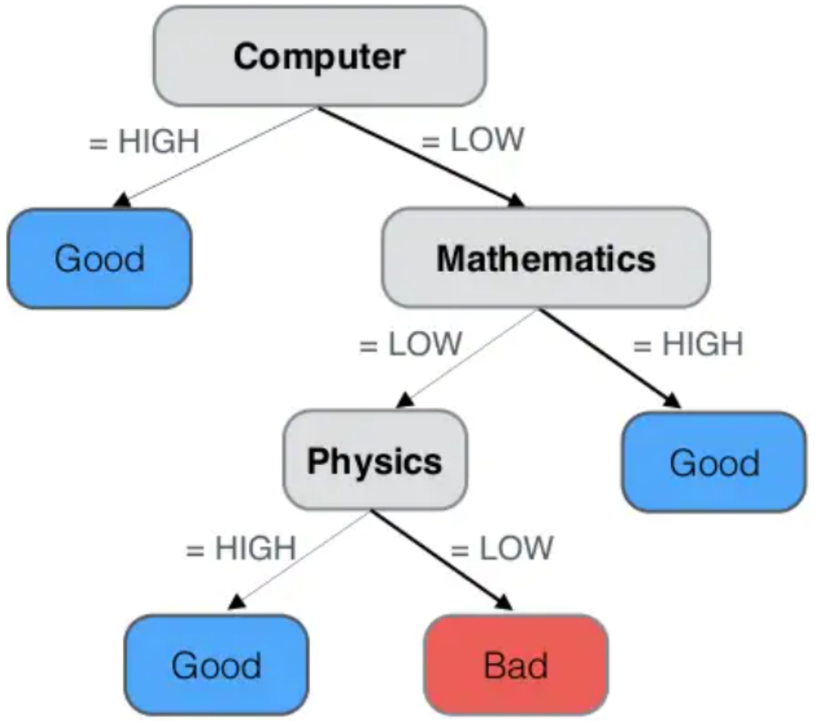
- ทดสอบด้วยข้อมูล 30% ที่แบ่งไว้
- คำนวณค่าความถูกต้อง

6. Deployment

- นำไปแนะนำนิสิตชั้นปีที่ 1 ที่กำลังจะเลือกภาควิชา
- พิจารณาจากเกรดตามโมเดลที่สร้างได้

รหัสนิสิต	เพศ	คอมพิวเตอร์	คณิตศาสตร์	...	Com Eng
5700123	ชาย	LOW	HIGH	...	??

ข้อมูลของนักศึกษาปีที่ 1 ที่ต้องการได้รับคำแนะนำ



โมเดลของภาควิชาวิศวกรรมคอมพิวเตอร์

CRISP - DM Example 1

5. Evaluation

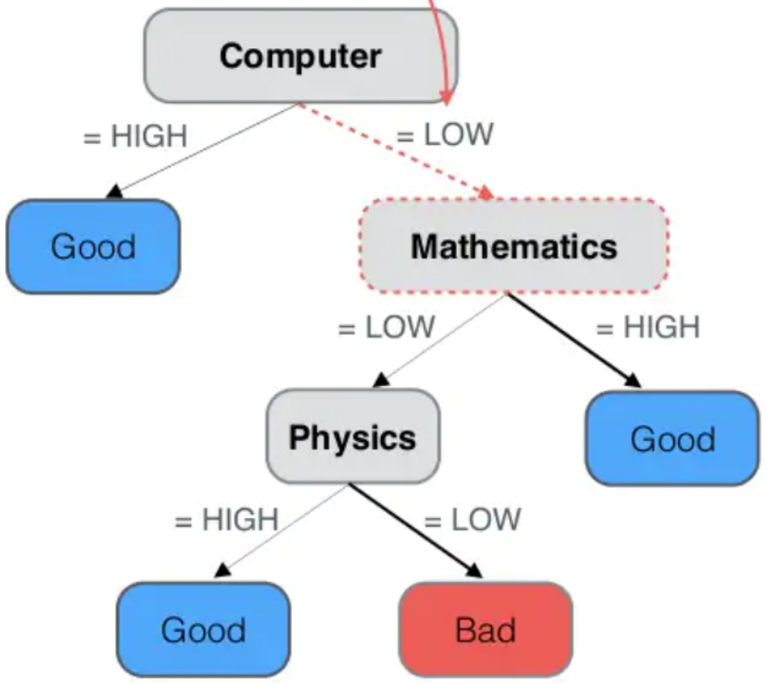
- ทดสอบด้วยข้อมูล 30% ที่แบ่งไว้
- คำนวณค่าความถูกต้อง

6. Deployment

- นำไปแนะนำนิสิตชั้นปีที่ 1 ที่กำลังจะเลือกภาควิชา
- พิจารณาจากเกรดตามโมเดลที่สร้างได้

รหัสนิสิต	เพศ	คอมพิวเตอร์	คณิตศาสตร์	...	Com Eng
5700123	ชาย	LOW	HIGH	...	??

ข้อมูลของนักศึกษาปีที่ 1 ที่ต้องการได้รับคำแนะนำ



โมเดลของภาควิชาวิศวกรรมคอมพิวเตอร์

CRISP - DM Example 1

5. Evaluation

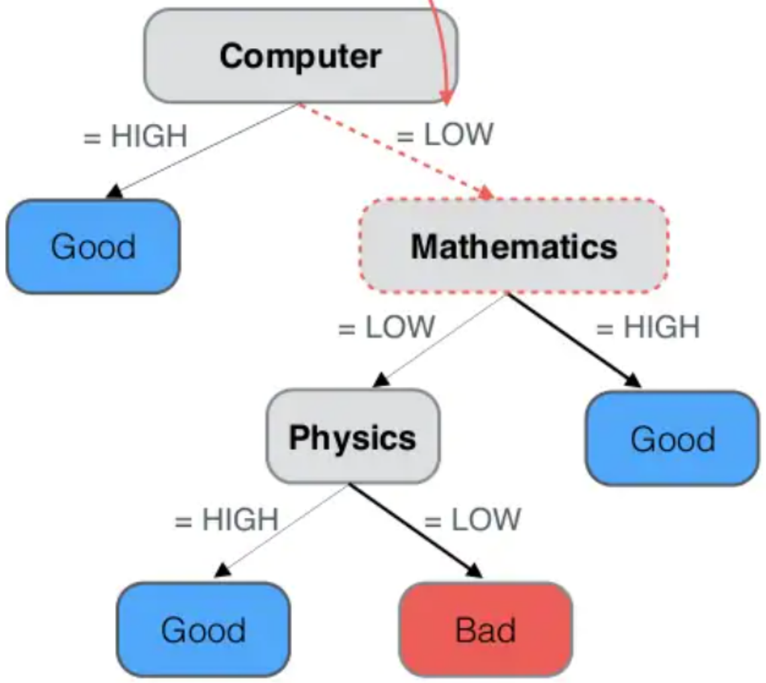
- ทดสอบด้วยข้อมูล 30% ที่แบ่งไว้
- คำนวณค่าความถูกต้อง

6. Deployment

- นำไปแนะนำนิสิตชั้นปีที่ 1 ที่กำลังจะเลือกภาควิชา
- พิจารณาจากเกรดตามโมเดลที่สร้างได้

รหัสนิสิต	เพศ	คอมพิวเตอร์	คณิตศาสตร์	...	Com Eng
5700123	ชาย	LOW	HIGH	...	??

ข้อมูลของนักศึกษาปีที่ 1 ที่ต้องการได้รับคำแนะนำ



โมเดลของภาควิชาวิศวกรรมคอมพิวเตอร์

CRISP - DM Example 1

5. Evaluation

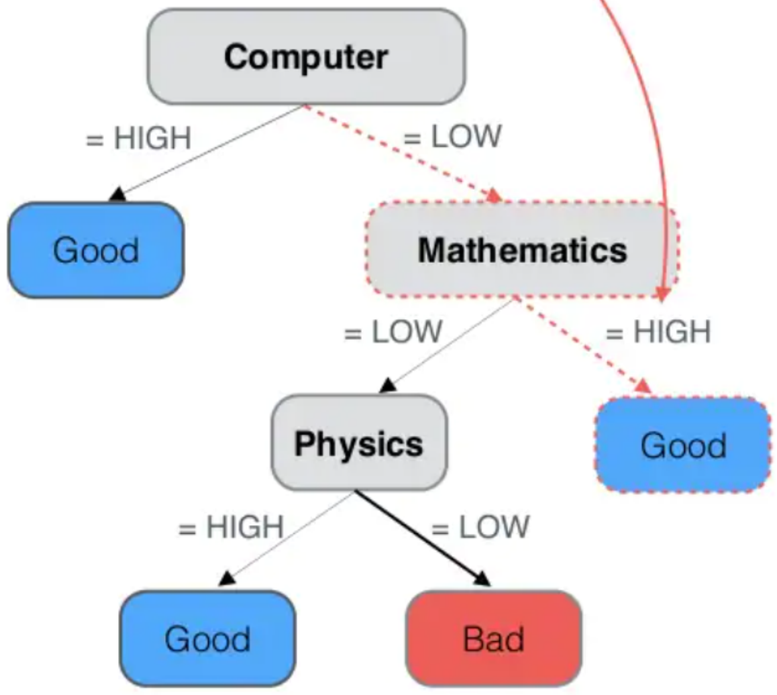
- ทดสอบด้วยข้อมูล 30% ที่แบ่งไว้
- คำนวณค่าความถูกต้อง

6. Deployment

- นำไปแนะนำนิสิตชั้นปีที่ 1 ที่กำลังจะเลือกภาควิชา
- พิจารณาจากเกรดตามโมเดลที่สร้างได้

รหัสนิสิต	เพศ	คอมพิวเตอร์	คณิตศาสตร์	...	Com Eng
5700123	ชาย	LOW	HIGH	...	??

ข้อมูลของนักศึกษาปีที่ 1 ที่ต้องการได้รับคำแนะนำ



โมเดลของภาควิชาวิศวกรรมคอมพิวเตอร์

CRISP - DM Example 1

5. Evaluation

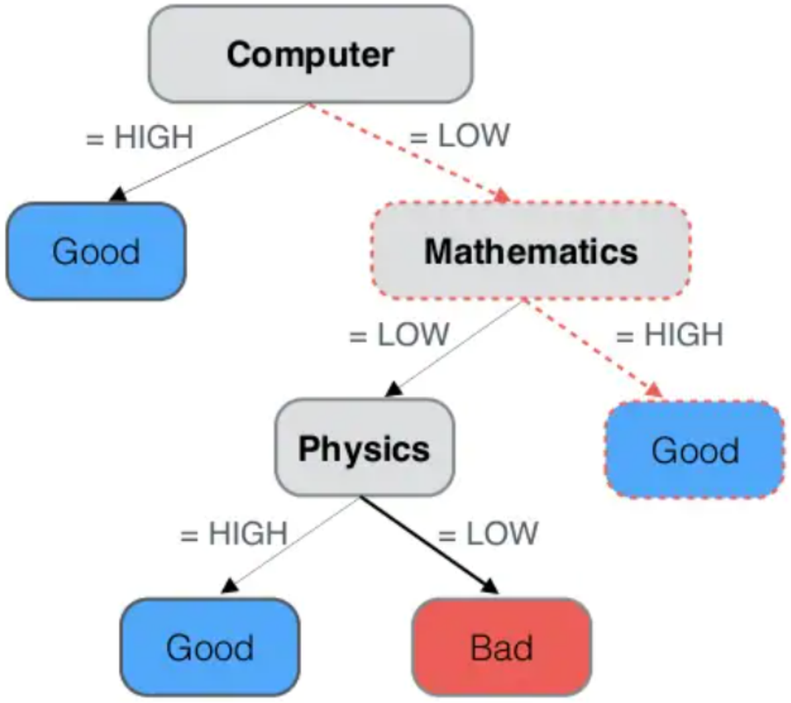
- ทดสอบด้วยข้อมูล 30% ที่แบ่งไว้
- คำนวณค่าความถูกต้อง

6. Deployment

- นำไปแนะนำนิสิตชั้นปีที่ 1 ที่กำลังจะเลือกภาควิชา
- พิจารณาจากเกรดตามโมเดลที่สร้างได้

รหัสนิสิต	เพศ	คอมพิวเตอร์	คณิตศาสตร์	...	Com Eng
5700123	ชาย	LOW	HIGH	...	Good

ข้อมูลของนักศึกษาปีที่ 1 ที่ต้องการได้รับคำแนะนำ



โมเดลของภาควิชาวิศวกรรมคอมพิวเตอร์