



Machine Learning for Non-Coders (Anyone)

Eakasit Pacharawongsakda, Ph.D.

eakasit@datacubeth.ai

Co-Founder and Data Science Team Lead

Cube Analytics Consulting Co., Ltd.

About me

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

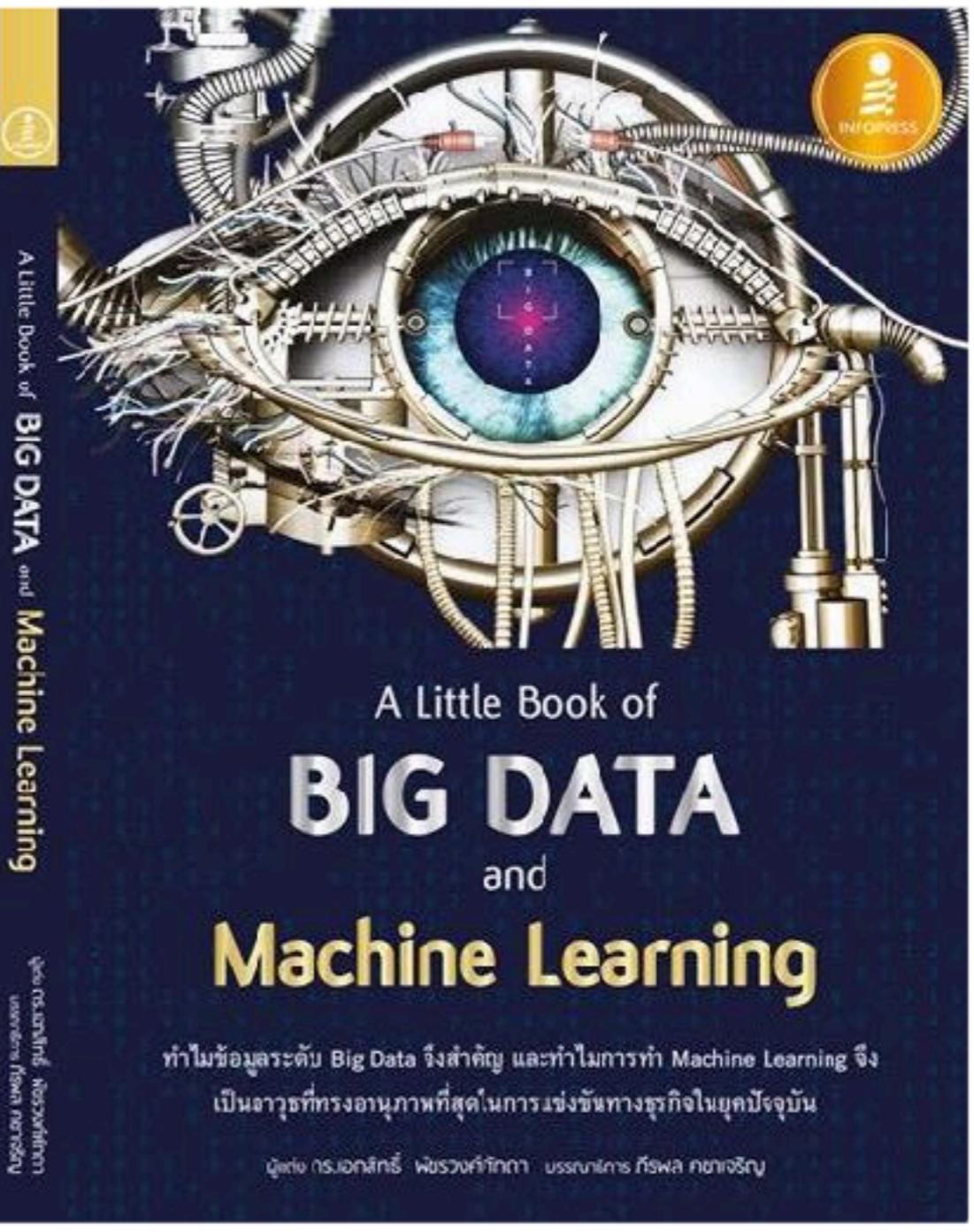
- ชื่อ: เอกสิทธิ์ พัชรวงศ์คักดา
- ประสบการณ์
 - Data Science Team Lead บริษัท คิวบ์ อนาโลเกติกส์ คอนซัลติ้ง จำกัด
 - อาจารย์ประจำหลักสูตร Big Data Engineering มหาวิทยาลัยธุรกิจบัลกิตร์
 - กี่ปรึกษาบริษัท ขายสินค้าทางด้านไอทีและการเงินแห่งหนึ่งในประเทศไทย
 - กี่ปรึกษาบริษัท พลังงานแห่งหนึ่งในประเทศไทย
 - กี่ปรึกษาบริษัท อสังหาริมทรัพย์แห่งหนึ่งในประเทศไทย
 - พัฒนาแบบจำลองในการแจ้งเตือนภัยพิบัติในจังหวัดต่างๆ ของประเทศไทย



About me

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- หนังสือ Big Data และ Machine Learning ฉบับภาษาไทย



About me

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- หนังสือ Data Mining ฉบับภาษาไทย



“Theory without practice is empty,
practice without theory is blind”

— Kant

Outlines

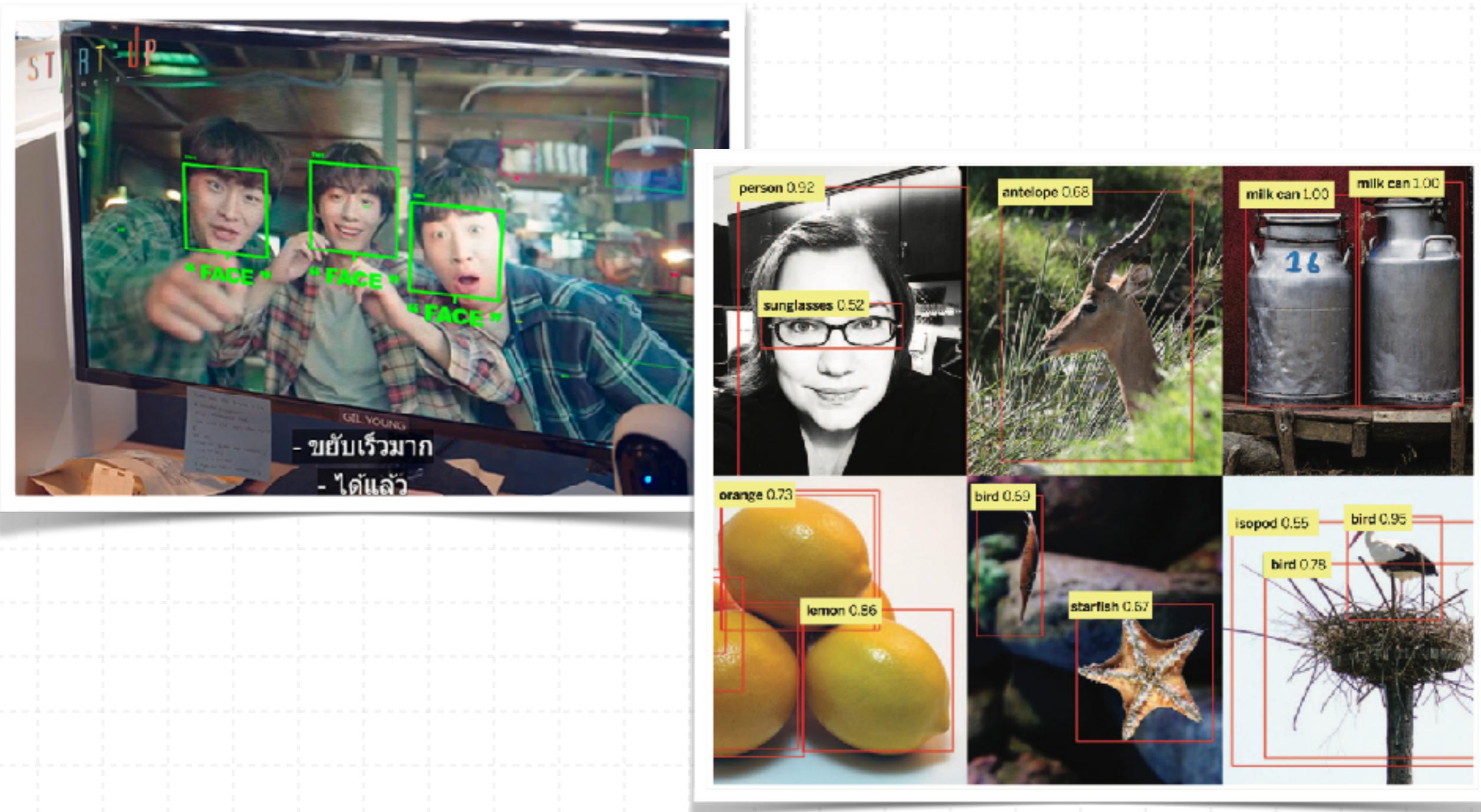
สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- **Machine Learning Applications**
- Data Management
- Data Roles
- Data Analytics
- Machine Learning
- Machine Learning Development Steps
- No-Code/Low-Code Machine Learning Platform

Machine Learning Application

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- ระบบตรวจจับวัตถุ (Image Recognition)



ที่มา: https://medium.com/@artisan_digital/สอนบันทึกเรื่อง-ai-เส้นทางพัฒนาธุรกิจชั้นนำจากเชียร์ส-start-up-f287ddaa843

<https://cs.uwaterloo.ca/~y328yu/mycourses/480-2018/readings/JordanMitchell.pdf>

Machine Learning Application

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- ระบบตรวจจับคนที่ไม่ใส่หน้ากากอนามัย

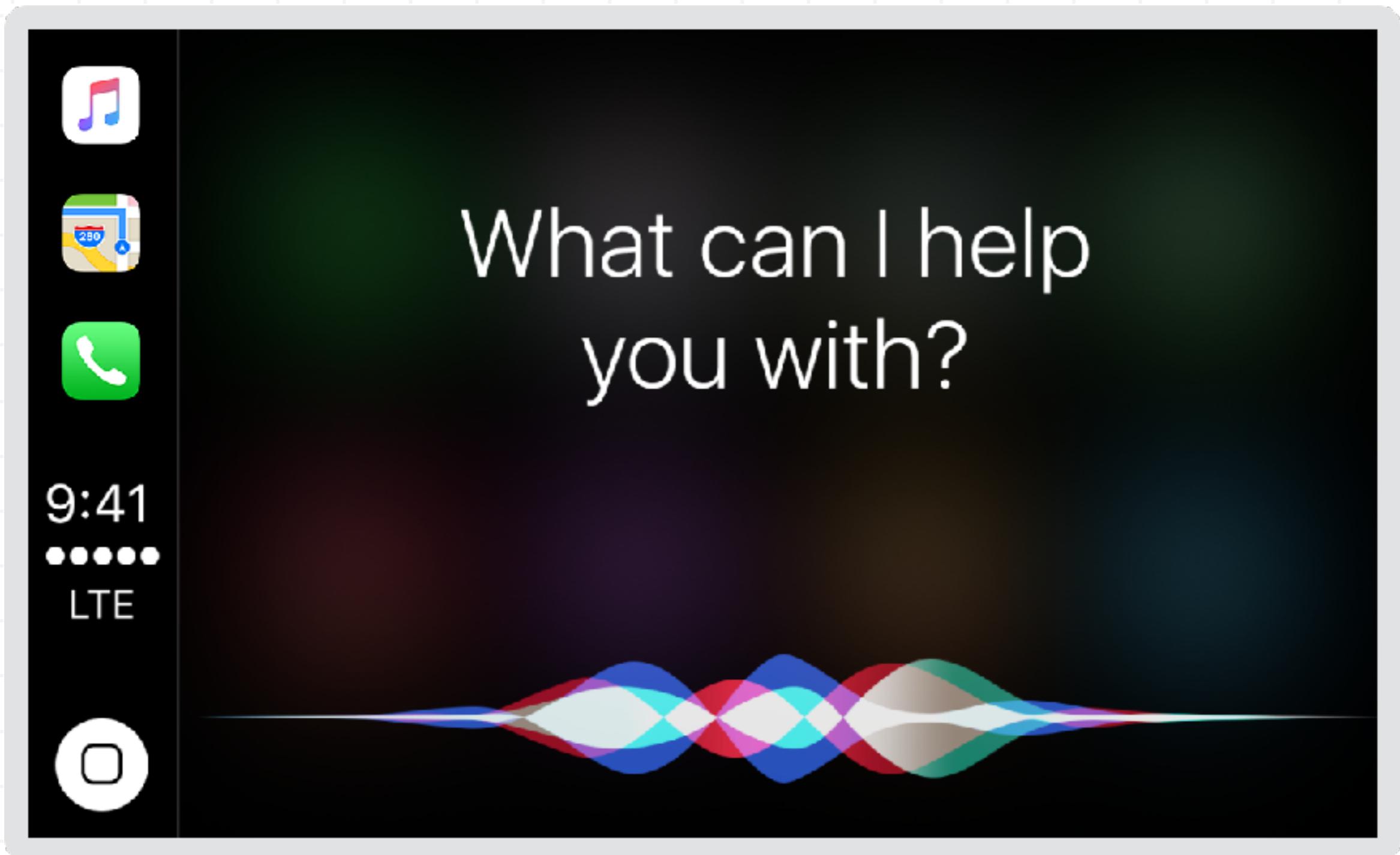


ที่มา: https://www.youtube.com/watch?v=S4TUM_p6H3E&feature=emb_logo

Machine Learning Application

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- ระบบรู้จำเสียง (Speech Recognition)



ที่มา: <https://developer.apple.com/design/human-interface-guidelines/carplay/interaction/voice/>

Machine Learning Application

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- จำลองอัจฉริยะ



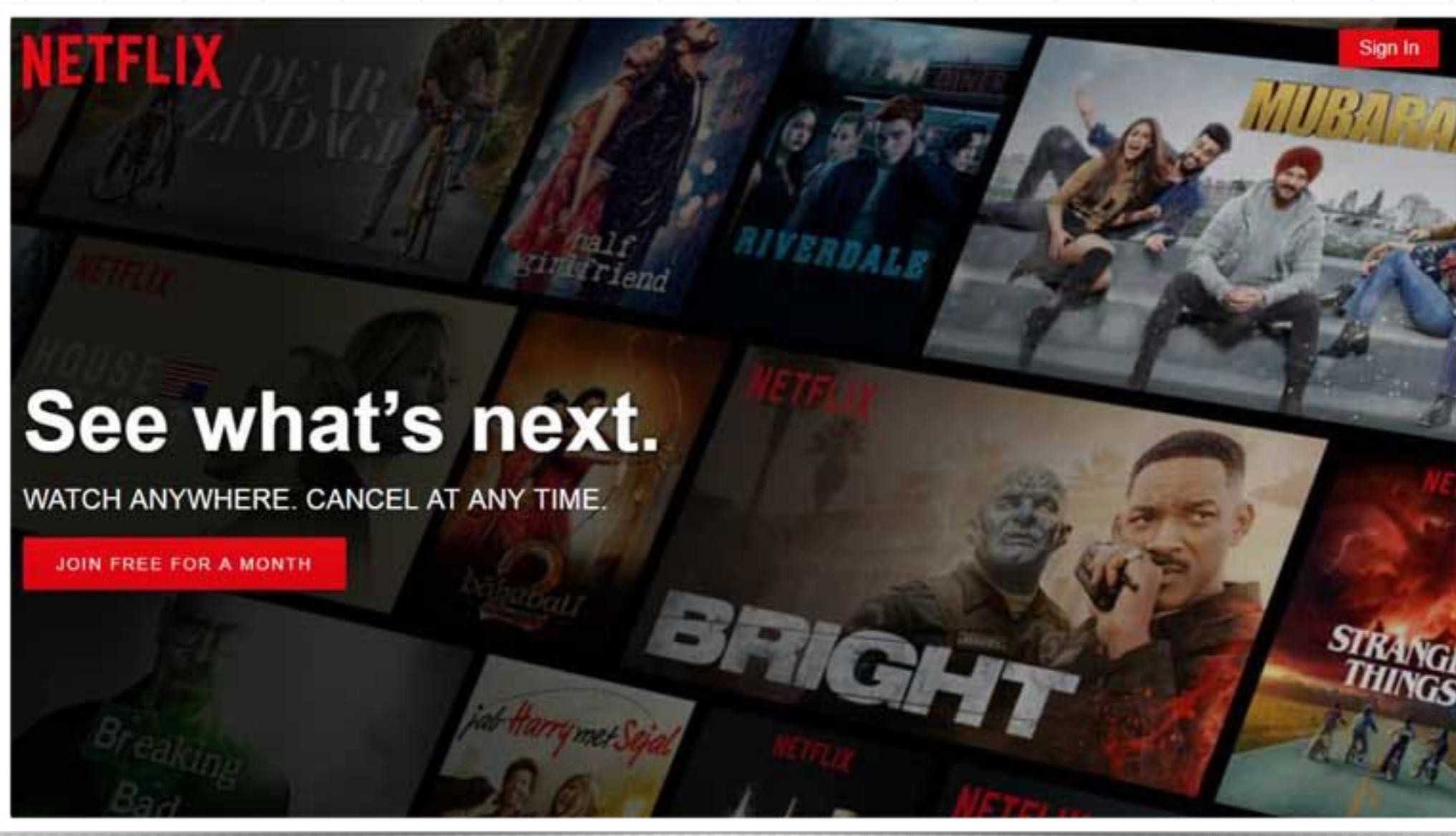
ที่มา: <https://www.zdnet.com/product/amazon-echo/>

<https://ninnygadget.com/review/mi-smart-speaker/>

Machine Learning Application

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- แนะนำสิ่งค้าที่เกี่ยวข้อง
 - Netflix แนะนำภาพยนตร์ที่คล้ายกับที่เคยดูมาก่อน

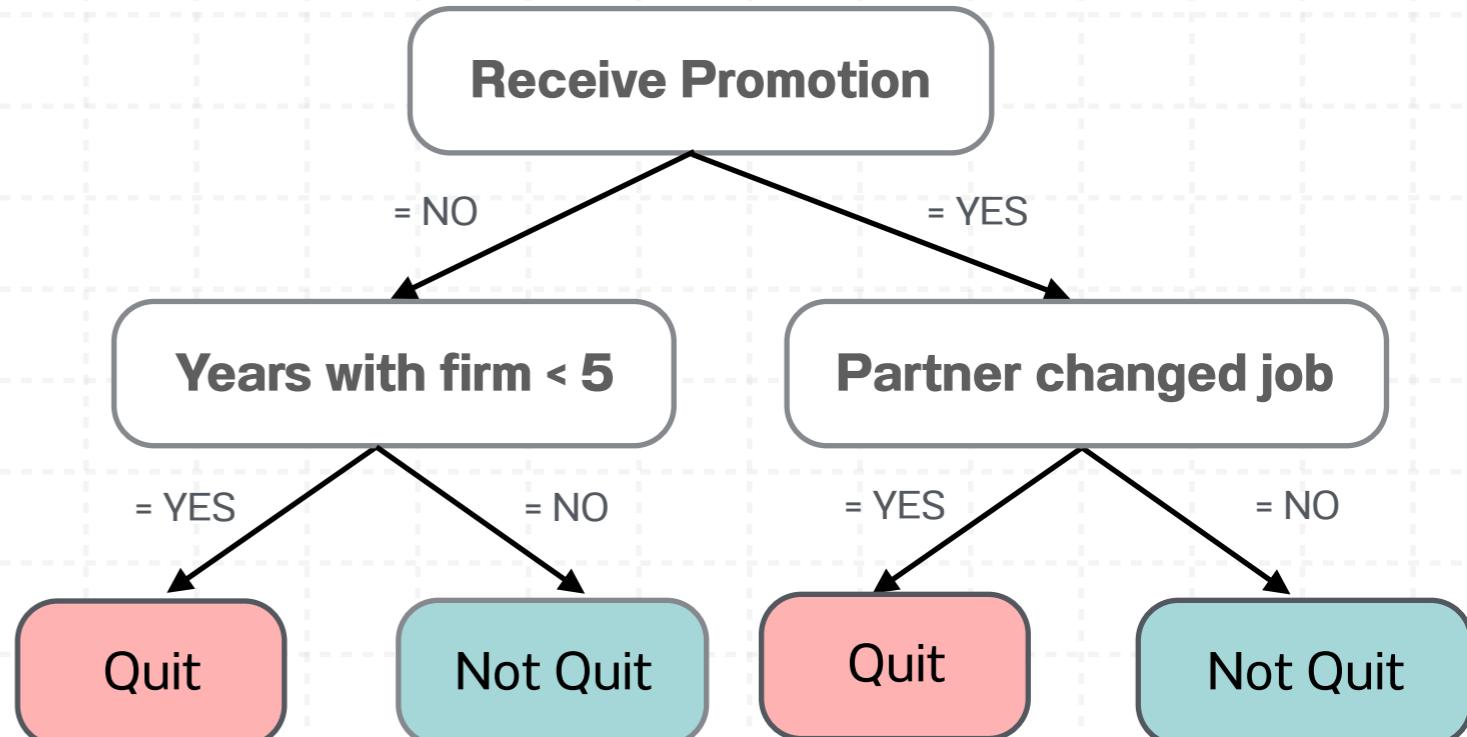


source: <https://indianexpress.com/article/technology/social/netflix-partnering-with-whatsapp-in-india-will-send-content-recommendations-5002335/>

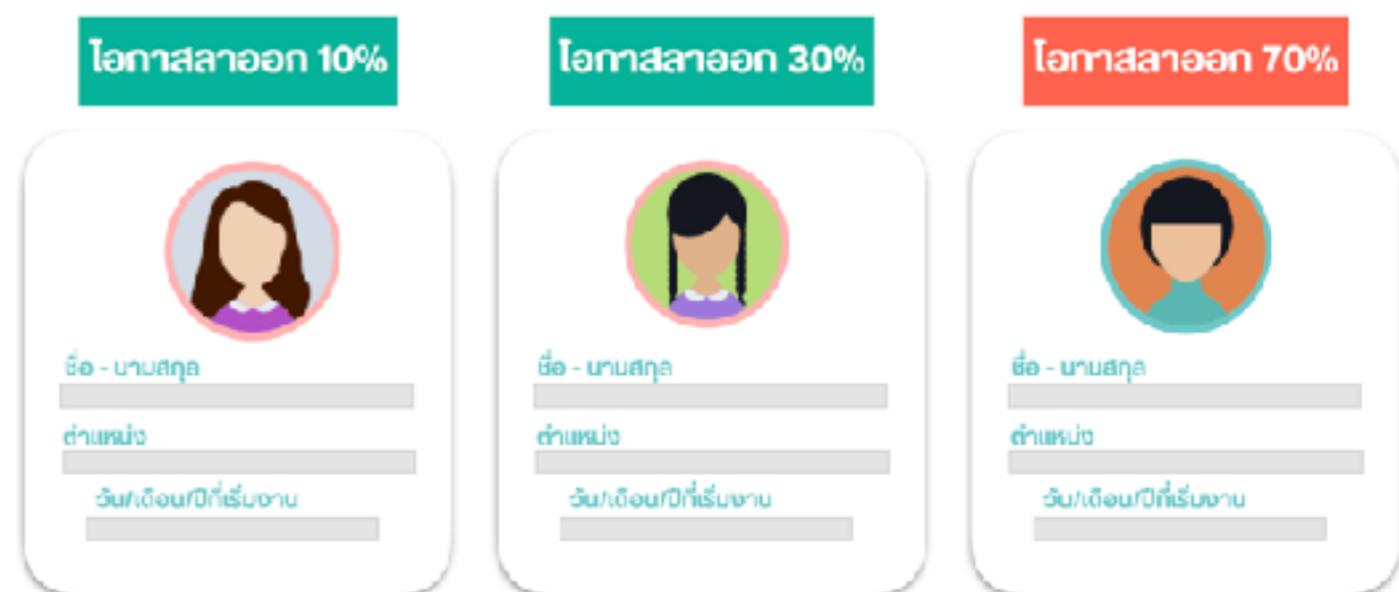
Machine Learning Application

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- คาดการณ์การลาออกจากพนักงาน
- ใช้ AI ดูพฤติกรรมของพนักงานที่ลาออกไปแล้ว และนำมาเทียบกับพนักงานที่ยังทำงานอยู่
- หาโอกาสการลาออกจากพนักงานแต่ละราย



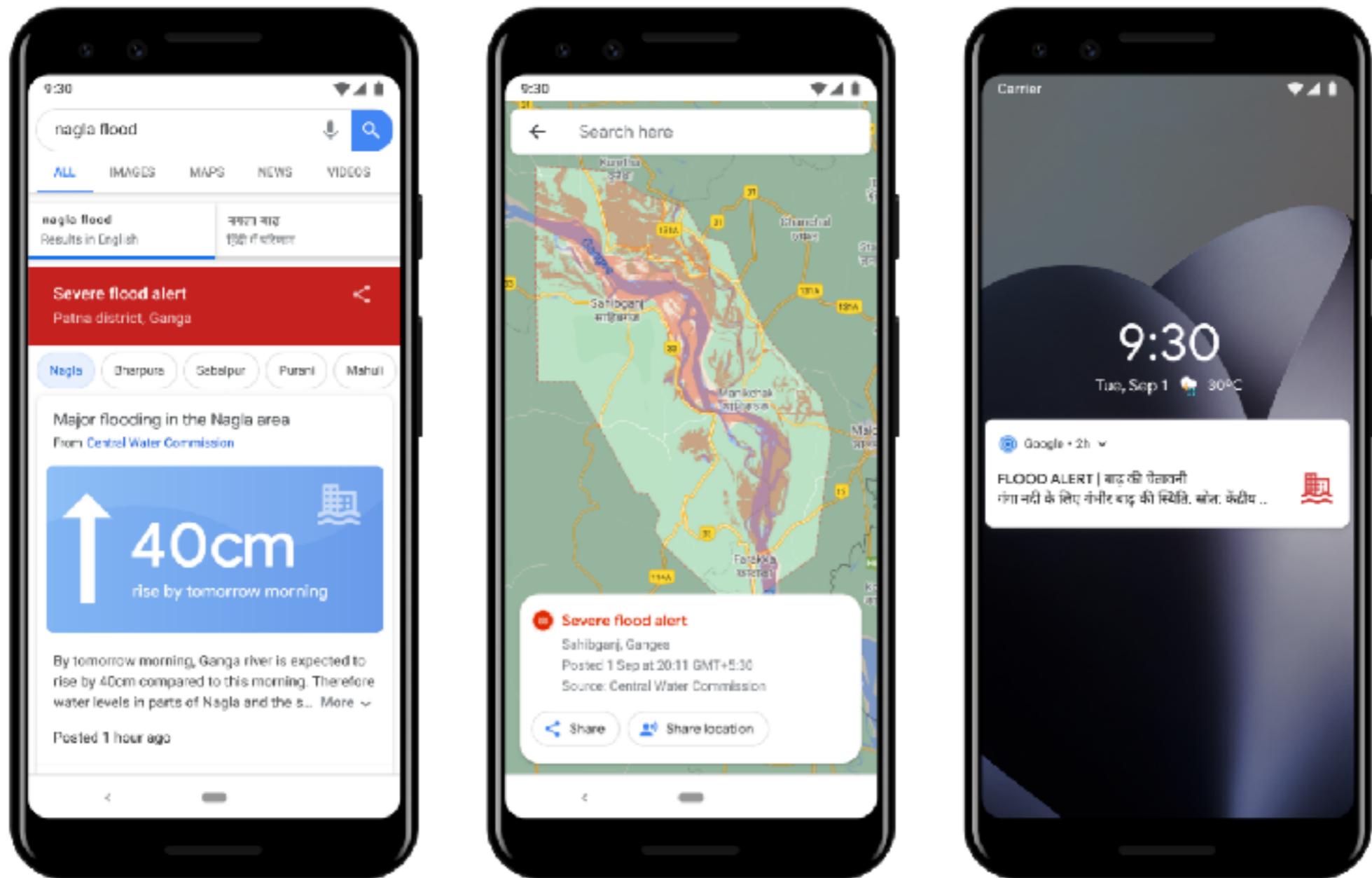
ตัวอย่างไม่เดลคาดการณ์การลาออกจากพนักงาน



Machine Learning Application

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- ระบบคาดการณ์และแจ้งเตือนภัยอุทกภัย

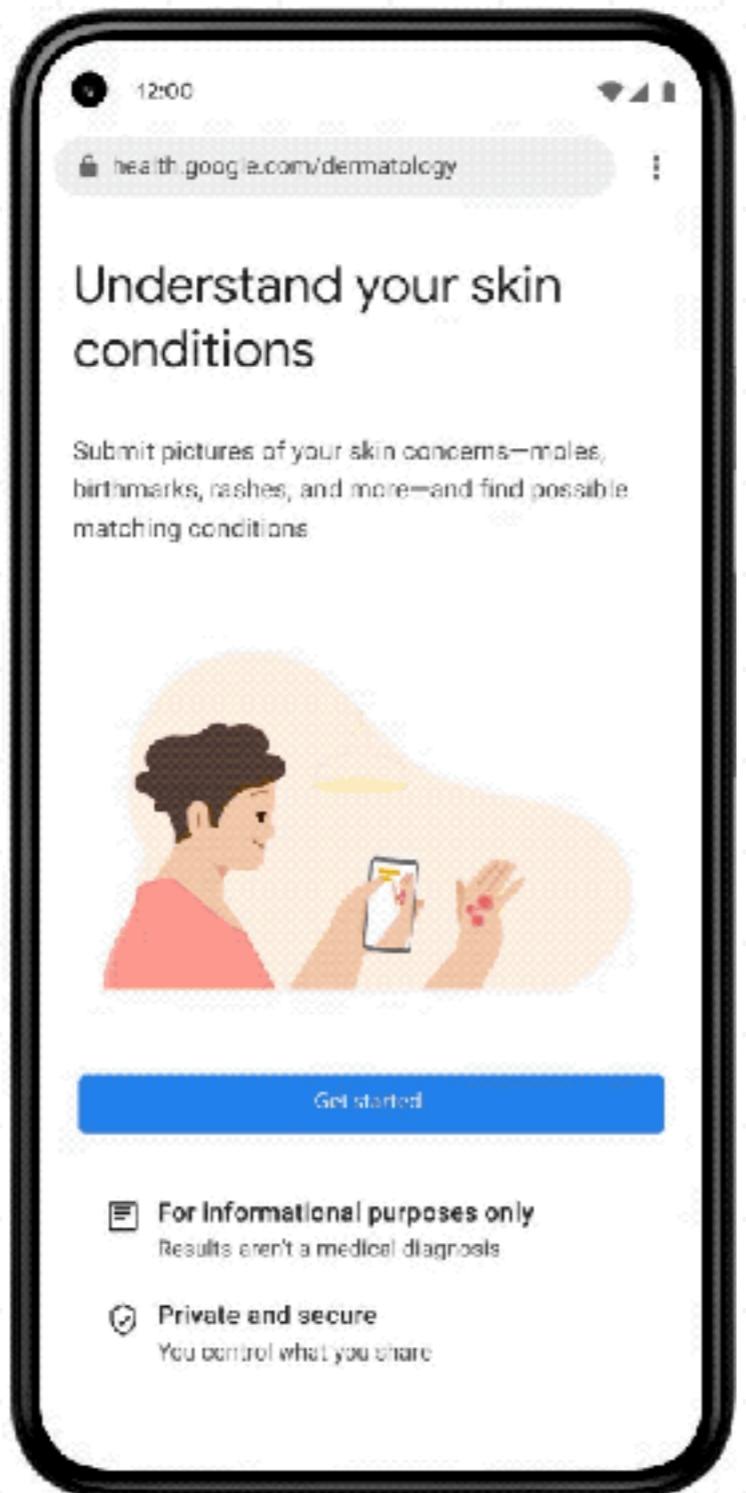


ที่มา: <https://blog.google/technology/ai/flood-forecasts-india-bangladesh/>

Machine Learning Application

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- ระบบคาดการณ์ความผิดปกติบนผิวหนัง



ที่มา: <https://blog.google/technology/health/ai-dermatology-preview-io-2021/>

Outlines

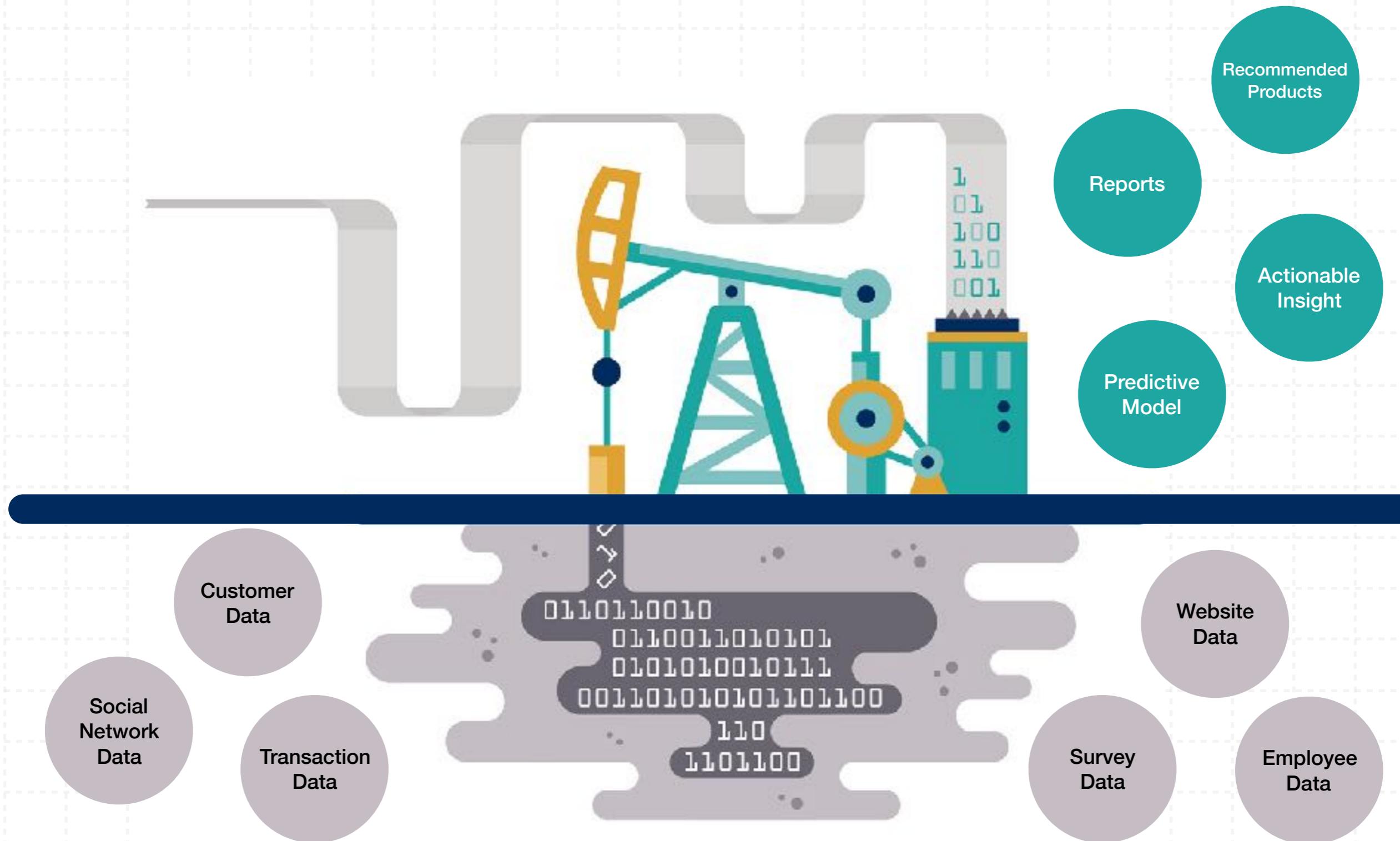
สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- Machine Learning Applications
- **Data Management**
- Data Roles
- Data Analytics
- Machine Learning
- Machine Learning Development Steps
- No-Code/Low-Code Machine Learning Platform

Data is a new oil

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

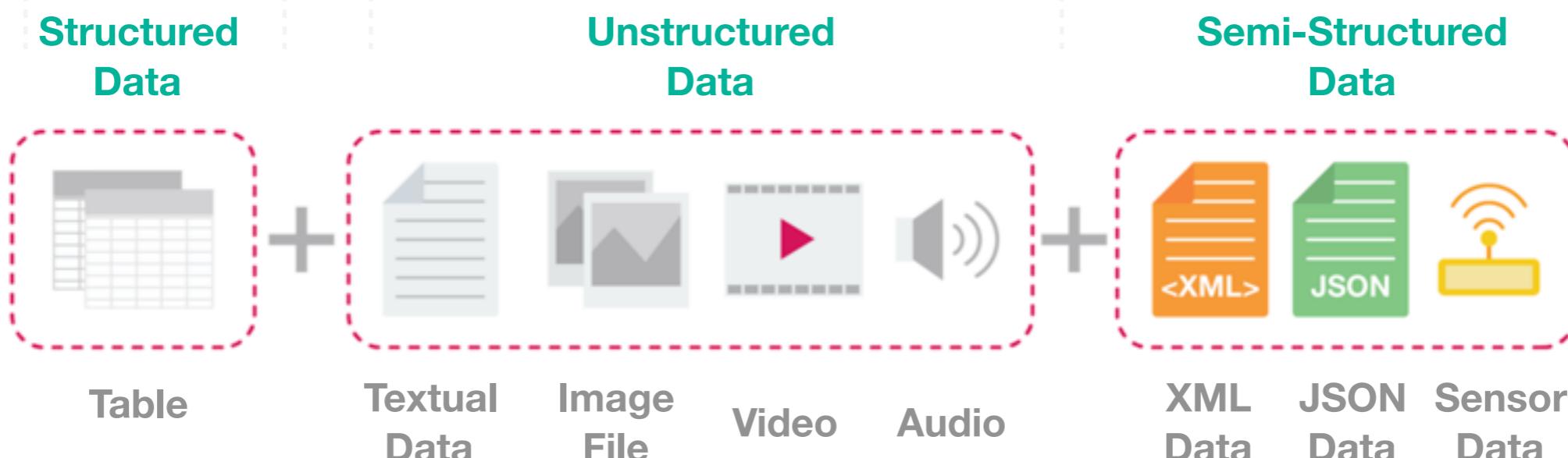
- “เมื่อข้อมูลมีค่าดั่งน้ำมัน”



Data

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- แบ่งตามรูปแบบของข้อมูล
 - ข้อมูลแบบมีโครงสร้าง (Structured Data) เช่น ข้อมูลในตาราง
 - ข้อมูลแบบไม่มีโครงสร้าง (Unstructured Data) เช่น ข้อความ รูปภาพ
 - ข้อมูลแบบกึ่งโครงสร้าง (Semi-structured Data) เช่น ไฟล์ JSON, ไฟล์ XML



Data

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- แบ่งตาม **รูปแบบ** ของข้อมูล
- ข้อมูลแบบมีโครงสร้าง** (Structured Data) เช่น ข้อมูลที่เก็บในรูปแบบตาราง

customer_id	age	gender	region	income	married	children	car	response
ID12101	48	FEMALE	INNER_CITY	17546.0	NO	1	NO	NO
ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO
ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES
ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO

- ข้อมูลแบบไม่มีโครงสร้าง** (Unstructured Data) เช่น ข้อมูลที่เป็นข้อความ
ข้อมูลรูปภาพ



Data

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- แปลงตาม **รูปแบบ** ของข้อมูล
- **ข้อมูลแบบกึ่งมีโครงสร้าง** (semi-structure) เช่น ข้อมูล XML หรือ JSON

```
<?xml version="1.0" standalone="yes" ?>

<BankAccount>

    <Number>1234</Number>

    <Type>Checking</Type>

    <OpenDate>11/04/2020</OpenDate>

    <Balance>25999.00</Balance>

    <AccountHolder>

        <FirstName>Manee</FirstName>

        <LastName>Yindee</LastName>

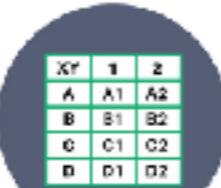
    <AccountHolder>

</BankAccount>
```

Data

Structured Data vs Unstructured Data

Can be displayed in rows, columns and relational databases



Numbers, dates and strings



Estimated 20% of enterprise data (Gartner)



Requires less storage



Easier to manage and protect with legacy solutions



vs

Unstructured Data

Cannot be displayed in rows, columns and relational databases



Images, audio, video, word processing files, e-mails, spreadsheets



Estimated 80% of enterprise data (Gartner)



Requires more storage



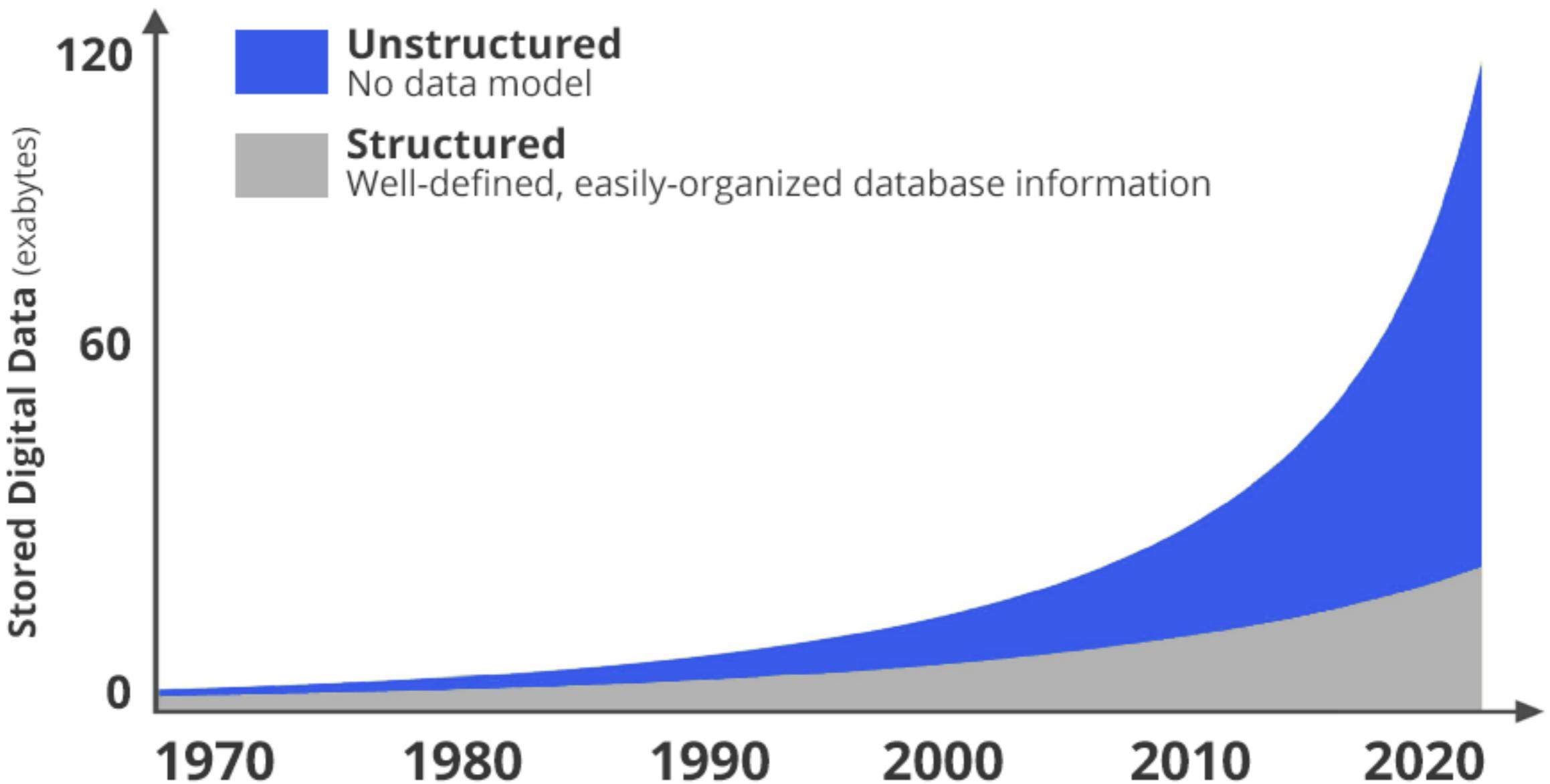
More difficult to manage and protect with legacy solutions



source: <https://www.igneous.io/blog/structured-data-vs-unstructured-data>

Data

สำหรับการบรรยายในงาน Dev Mountain Tech Festival



source: <https://www.m-files.com/blog/what-is-structured-data-vs-unstructured-data/>

Data Storage

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- การเก็บข้อมูลแบบไฟล์ (file based system)
 - เก็บข้อมูลไว้ในไฟล์ เช่น Excel หรือ Word หรือ Text File
 - แต่ละไฟล์ **มีรูปแบบการจัดเก็บที่ต่างกัน**
- ข้อจำกัดของการเก็บข้อมูลแบบไฟล์
 - ไม่มีรูปแบบมาตรฐาน** (no standard)
 - ข้อมูลมีการ**เก็บซ้ำซ้อนกัน** (duplicate) ระหว่างแต่ละคน
 - ข้อมูล **update ไม่ตรง** กัน
 - การดึงข้อมูลมาใช้งานจากหลาย ๆ ที่ **ทำได้ยาก**

Data Storage: Database

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- ฐานข้อมูล (database)
 - เป็นการเก็บข้อมูลไว้ในตารางซึ่งมีความสัมพันธ์กัน ทำให้สามารถ update, delete ได้อย่างถูกต้อง และช่วยให้การค้นหาข้อมูลง่ายขึ้น

ตาราง Customers

Customer_ID	Name	Surname	Gender	Birthdate
D2021-0001	ปีติ	ยันดี	ชาย	28/08/2545
D2021-0002	มาเน	มีความสุข	หญิง	22/06/2545
D2021-0003	บุญใจ	โปรดี	หญิง	12/05/2545

ตาราง Transactions

Customer_ID	Branch_ID	Product_ID	Qty	Date
D2021-0001	B101	00150	1	14/02/2021
D2021-0001	B102	00185	2	14/02/2021
D2021-0002	B102	182	1	12/05/2021
D2021-0002	B102	185	1	12/05/2021

Data Storage: Database

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

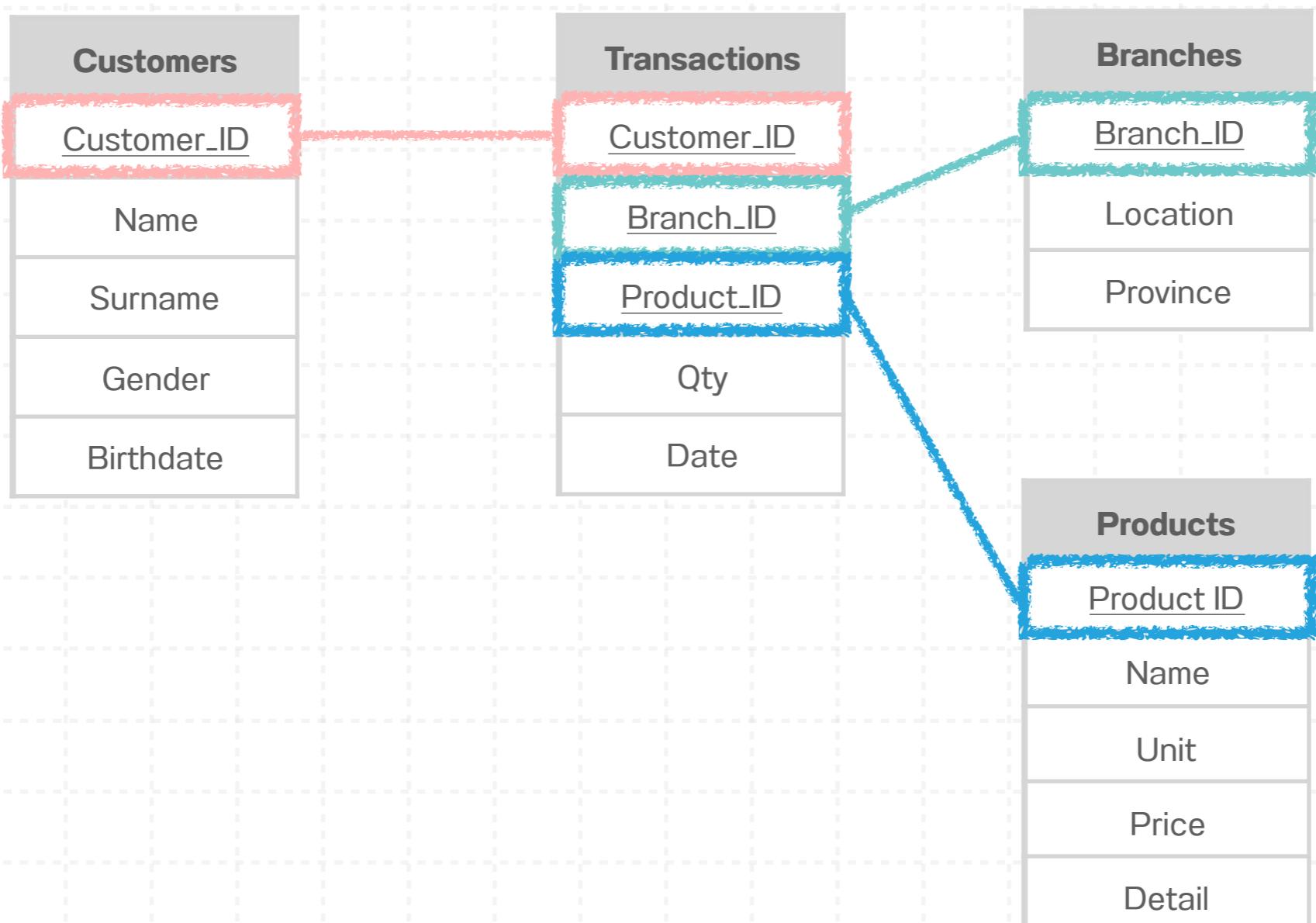
- ฐานข้อมูล (database) ประกอบด้วย
 - ตาราง (table หรือ relation) เก็บข้อมูลส่วนต่างๆ ไว้ และประกอบด้วย
 - ฟิลด์ (field) และคุณลักษณะ (attribute) ของข้อมูลแต่ละรายการ
 - เรคอร์ด (record) และข้อมูลข้อมูลแต่ละแถว (row) หรือตัวอย่าง (example)
 - ความสัมพันธ์ (relation) และความสัมพันธ์ระหว่างตาราง โดยใช้คีย์ (key) เป็นฟิลด์ที่เชื่อมโยงไปยังตารางอื่น



Data Storage: Database

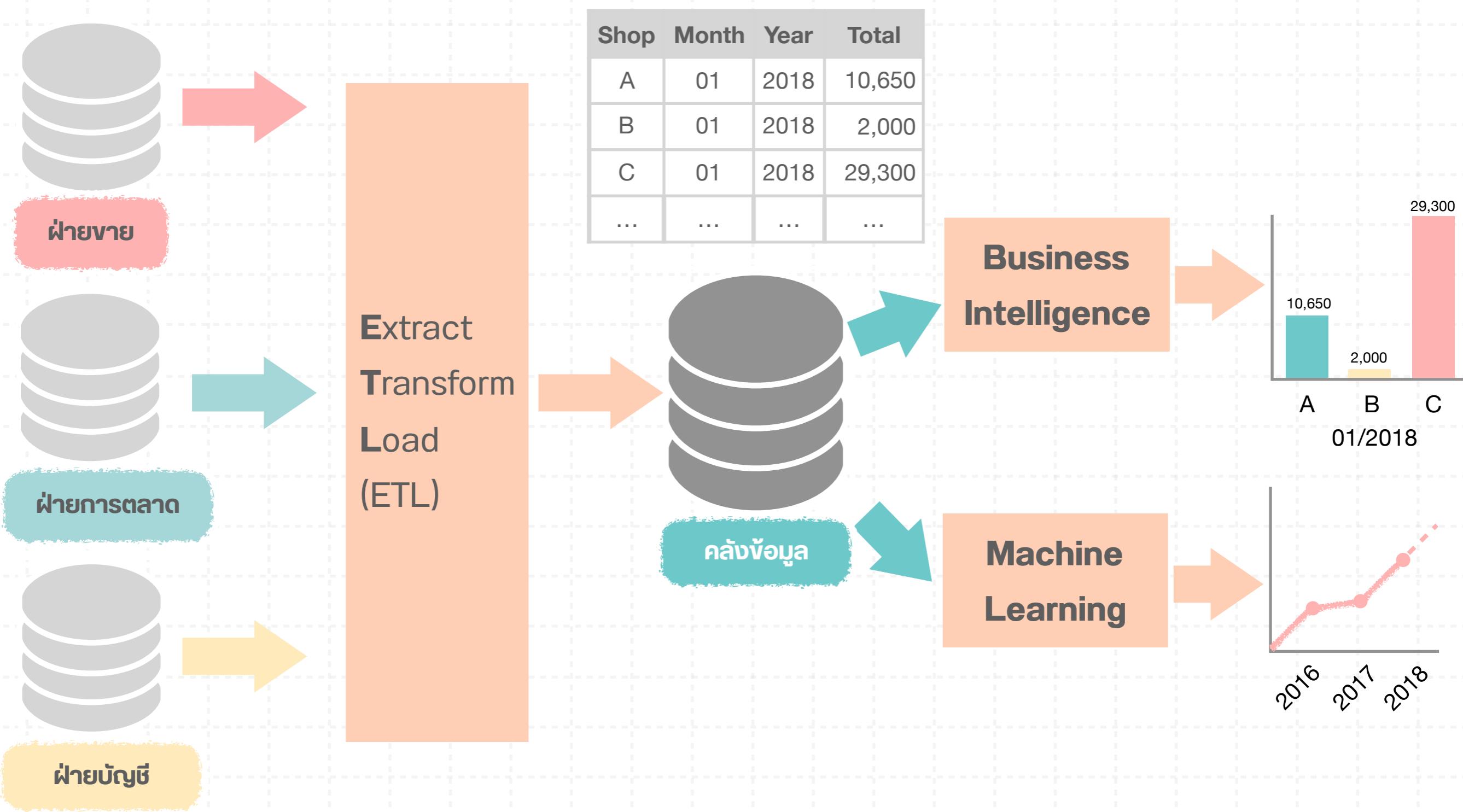
สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- ฐานข้อมูลจะแสดงในรูปแบบของ schema diagram ได้ เพื่อให้เห็นพาร์และความสัมพันธ์ระหว่างตาราง



Data Storage: Data Warehouse

สำหรับการบรรยายในงาน Dev Mountain Tech Festival



Data Storage: Data Warehouse

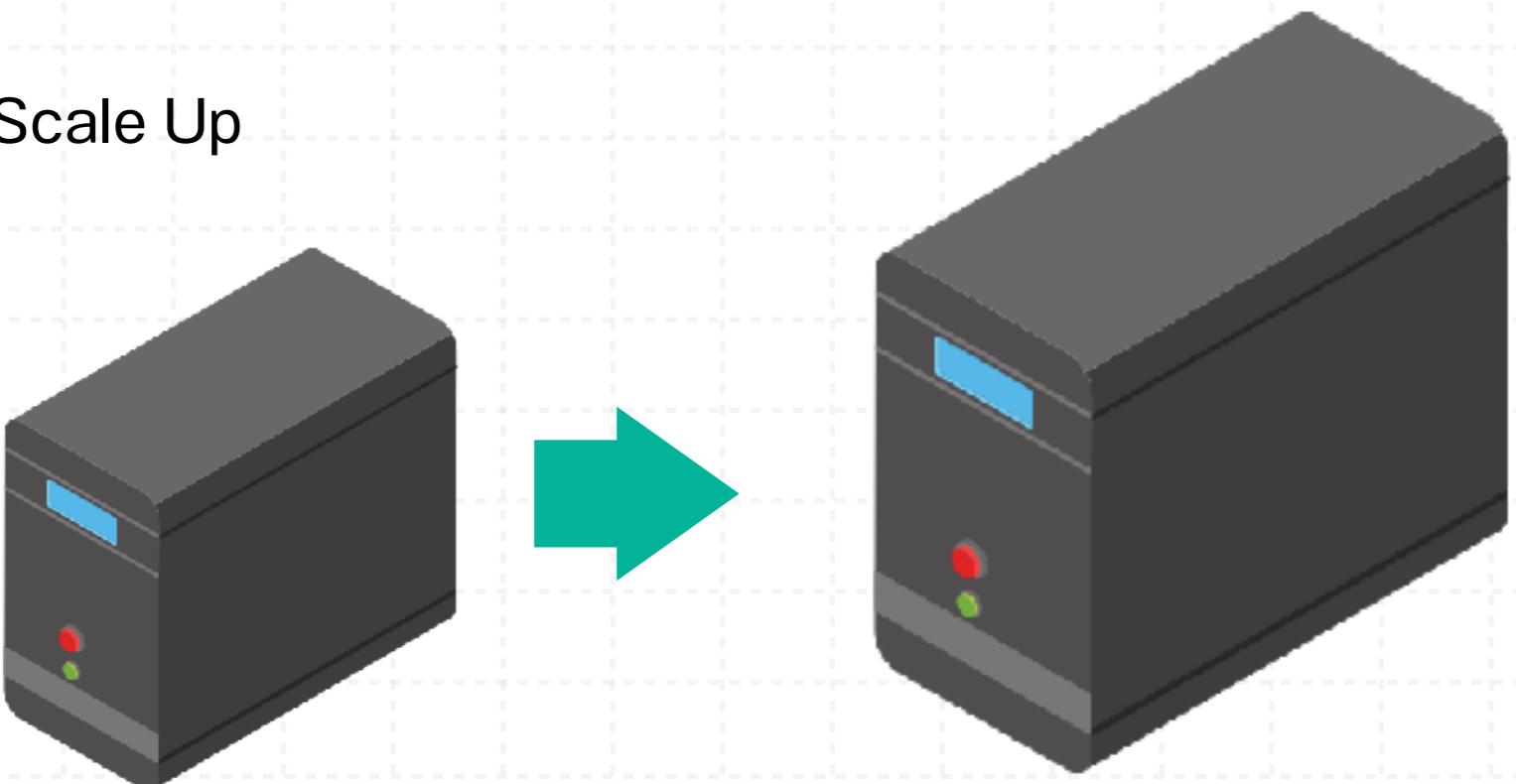
สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- **Database**
 - **ฐานข้อมูลใช้ในการจัดเก็บข้อมูล** ลดความซ้ำซ้อนของข้อมูล เน้นการจัดเก็บ เพิ่มแก้ไข และลบข้อมูล
- **Data Warehouse**
 - **คลังข้อมูลรวมข้อมูลจากหลายๆ ฐานข้อมูล** แปลงข้อมูลให้มีความเหมือนกัน หมายความว่า ให้มีรูปแบบเดียวกัน เช่น รหัสประจำตัวผู้คน เป็นต้น แล้วนำข้อมูลมา放在一起 ทำให้สามารถใช้งานได้สะดวก
- **Machine Learning**
 - เป็นการ**เรียนรู้จากข้อมูลในอดีตเพื่อนำมาสร้างแบบจำลอง** สำหรับใช้ในการพยากรณ์เหตุการณ์ในอนาคต

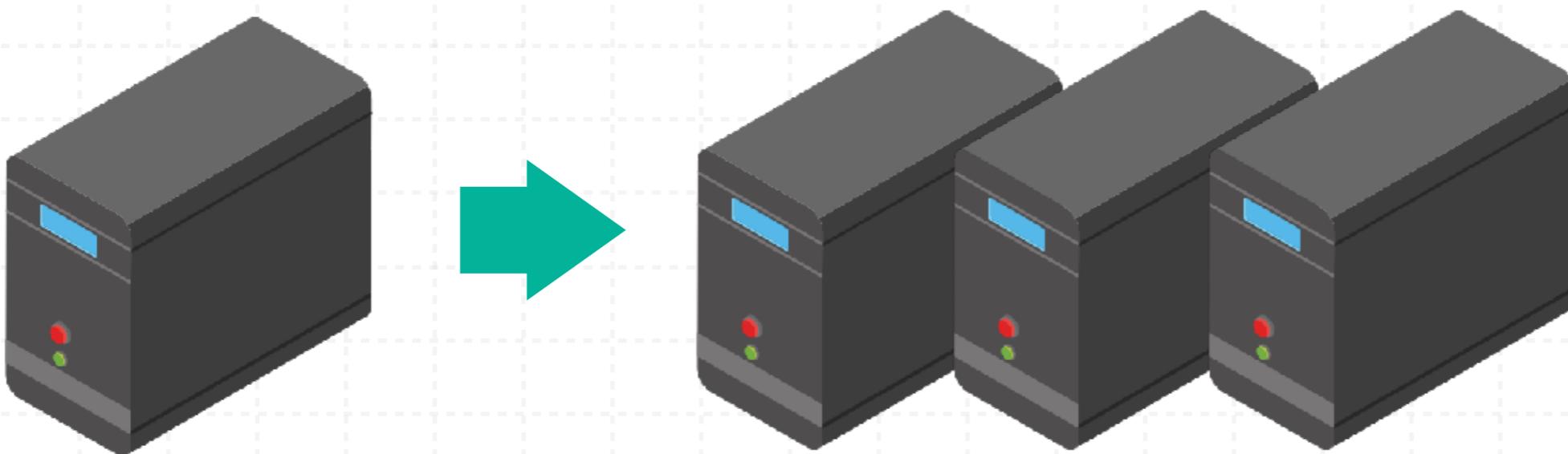
Scale Up v.s. Scale Out

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

Scale Up

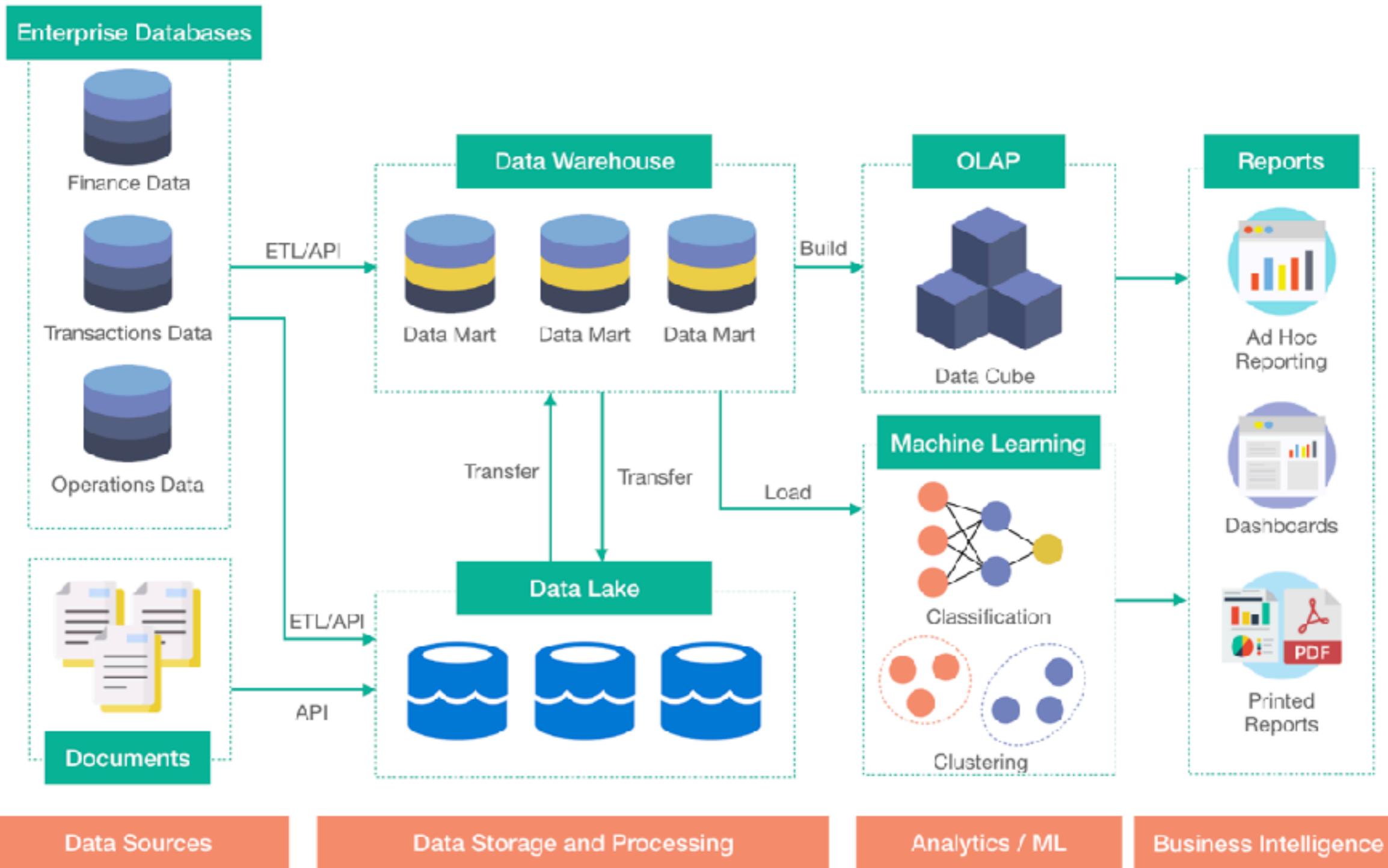


Scale Out



Big Data Framework

สำหรับการบรรยายในงาน Dev Mountain Tech Festival



Outlines

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- Machine Learning Applications
- Data Management
- **Data Roles**
- Data Analytics
- Machine Learning
- Machine Learning Development Steps
- No-Code/Low-Code Machine Learning Platform

Data Roles

สำหรับการบรรยายในงาน Dev Mountain Tech Festival



Data Analyst

(Business Analyst)

Skills

- Statistics
- Communication
- Business Knowledge

Tools

- Excel
- Tableau / Power BI
- SQL



Data Scientist

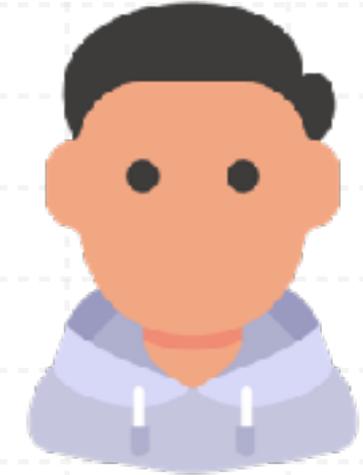
(Statisticians, Data Managers)

Skills

- Mathematics
- Programming
- Analytics

Tools

- SQL
- Python / R
- RapidMiner / SAS



Data Engineer

(Data Architects, Database Administrators)

Skills

- Programming
- Mathematics
- Big Data

Tools

- Hadoop / Spark
- NoSQL
- Python

Data Roles: Citizen Data Scientist

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

“a person who **creates or generates models** that leverage predictive or prescriptive analytics, but whose **primary job function is outside of the field** of statistics and analytics.”

– Gartner

Data Roles

สำหรับการบรรยายในงาน Dev Mountain Tech Festival



Data Scientist

จุดเด่น

- มีความรู้ทางสถิติและ
การวิเคราะห์ข้อมูล
- เขียน code ได้

จุดด้อย

- อาจจะไม่ค่อยเข้าใจธุรกิจ
- ใช้เวลาเรียนรู้ธุรกิจ



Citizen Data Scientist

จุดเด่น

- มีความรู้และความเชี่ยวชาญ
ในเชิงธุรกิจที่กำอยู่

จุดด้อย

- ไม่ค่อย懂บัดเดิน code

Outlines

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- Machine Learning Applications
- Data Management
- Data Roles
- **Data Analytics**
- Machine Learning
- Machine Learning Development Steps
- No-Code/Low-Code Machine Learning Platform

Data Analytics

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

“Data and analytics is the **management of data for all uses** (operational and analytical) and **the analysis of data to drive business processes** and **improve business outcomes through more effective decision making** and enhanced customer experiences.”

— Gartner

“เป็นการบริหารจัดการข้อมูลสำหรับการวิเคราะห์เพื่อช่วยขับเคลื่อนธุรกิจและสนับสนุนการตัดสินใจได้อย่างมีประสิทธิภาพ”

Data Analytics

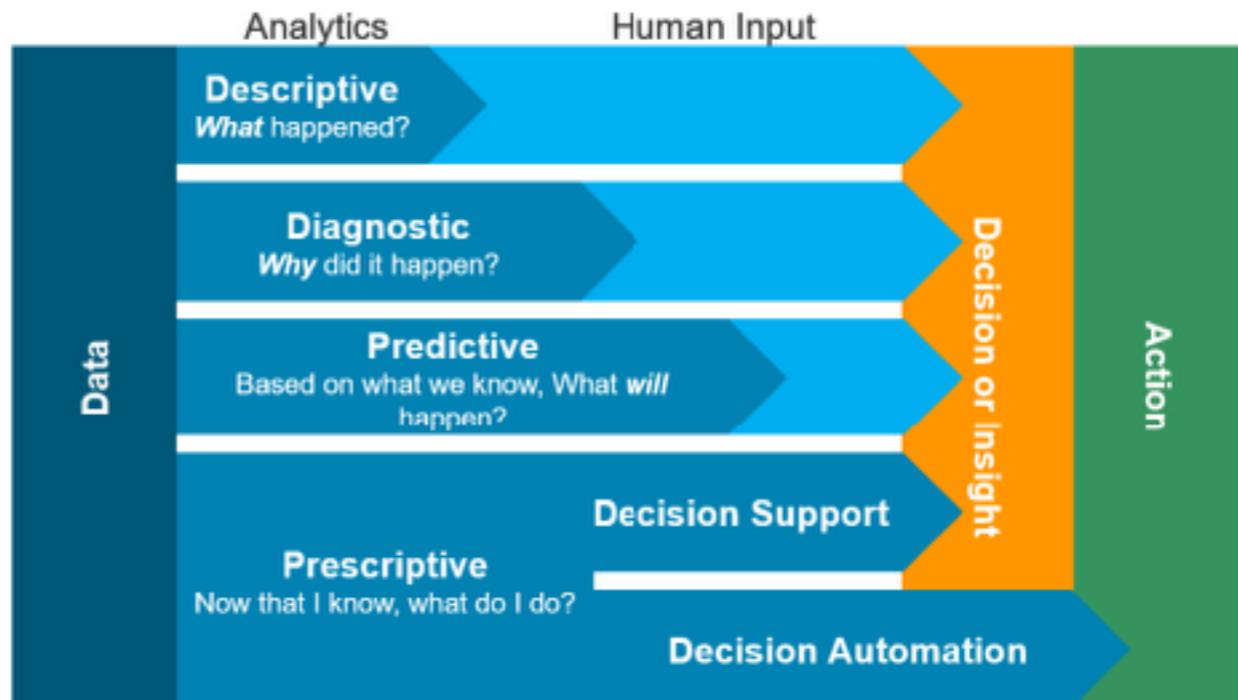
สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- วัตถุประสงค์หนึ่งของการทำ Data Analytics คือ การวิเคราะห์ข้อมูลเพื่อ **สนับสนุนการตัดสินใจ**
- Gartner ได้แบ่งระดับของการทำ Data Analytics ไว้ 4 ระดับ ได้แก่
 - การวิเคราะห์**เชิงพรรณนา** (Descriptive)
 - การวิเคราะห์**เชิงวินิจฉัย** (Diagnostic)
 - การวิเคราะห์**เชิงพยากรณ์** (Predictive)
 - การวิเคราะห์**เชิงให้คำแนะนำ** (Prescriptive)
- โดยจะเริ่มจากข้อมูล (data) เป็นหลักเพื่อช่วยสนับสนุนการตัดสินใจและถู่ว่าคบ (หรือมนุษย์) จะต้องเกี่ยวข้องกับการวิเคราะห์มากน้อยแค่ไหน

Data Analytics

สำหรับการบรรยายในงาน De-

- จากรูปจะแบ่งเป็น 5 ส่วน คือ
 - **Data** คือ ข้อมูลที่ใช้ในการวิเคราะห์
 - **Analytics** คือ การวิเคราะห์ข้อมูล รูปแบบต่างๆ
 - **Human Input** คือ สัดส่วนของการที่ คนเข้ามามีส่วนร่วมในการวิเคราะห์
 - **Decision or Insight** คือ การตัดสินใจและสิ่งสำคัญที่ค้นพบ (insight)
 - **Action** คือ การทำงาน



Adapted from Gartner Report July 2015

Data Analytics: Descriptive

- การวิเคราะห์เชิงพรรณนา (Descriptive)

พิจารณาข้อมูลในอดีตเพื่อถ่วงเวลาเมื่อไร

เกิดขึ้น (What happened?)

ส่วนใหญ่จะเป็นการพิจารณา

ตัวแปรเดียว เช่น ยอดขายของ

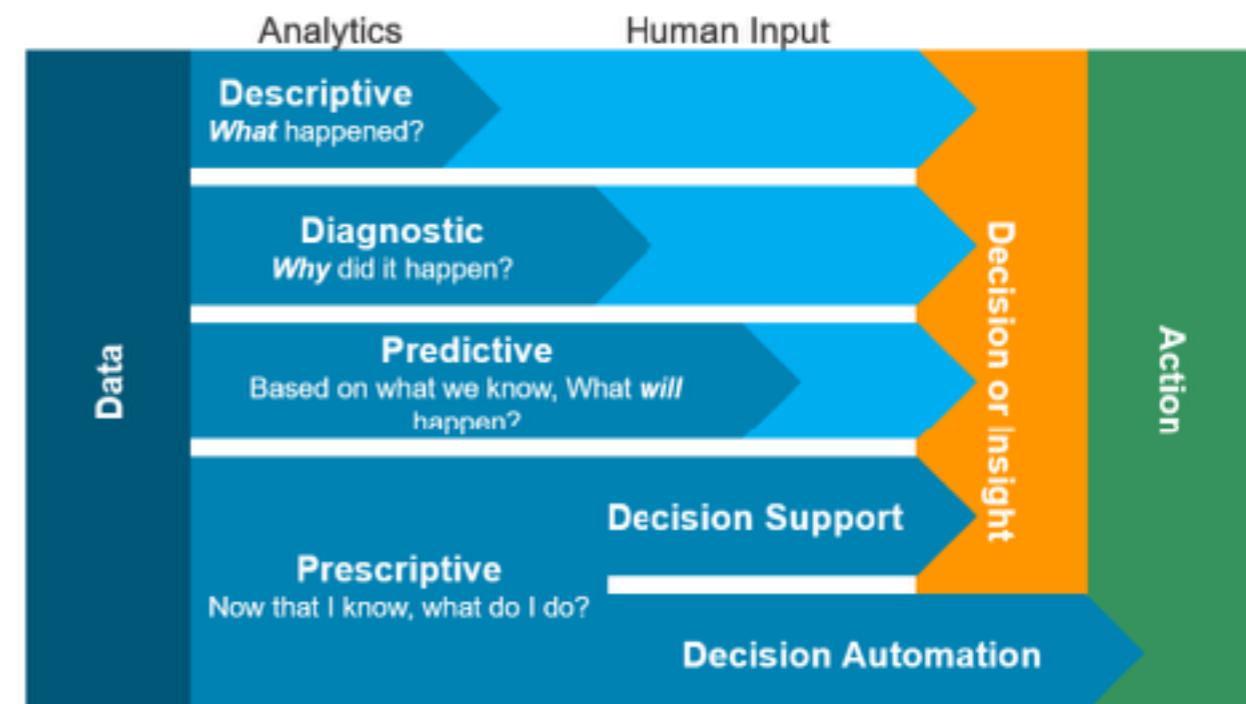
เครื่องดื่มชนิดหนึ่งของแต่ละเดือน

ในปีผ่านมา

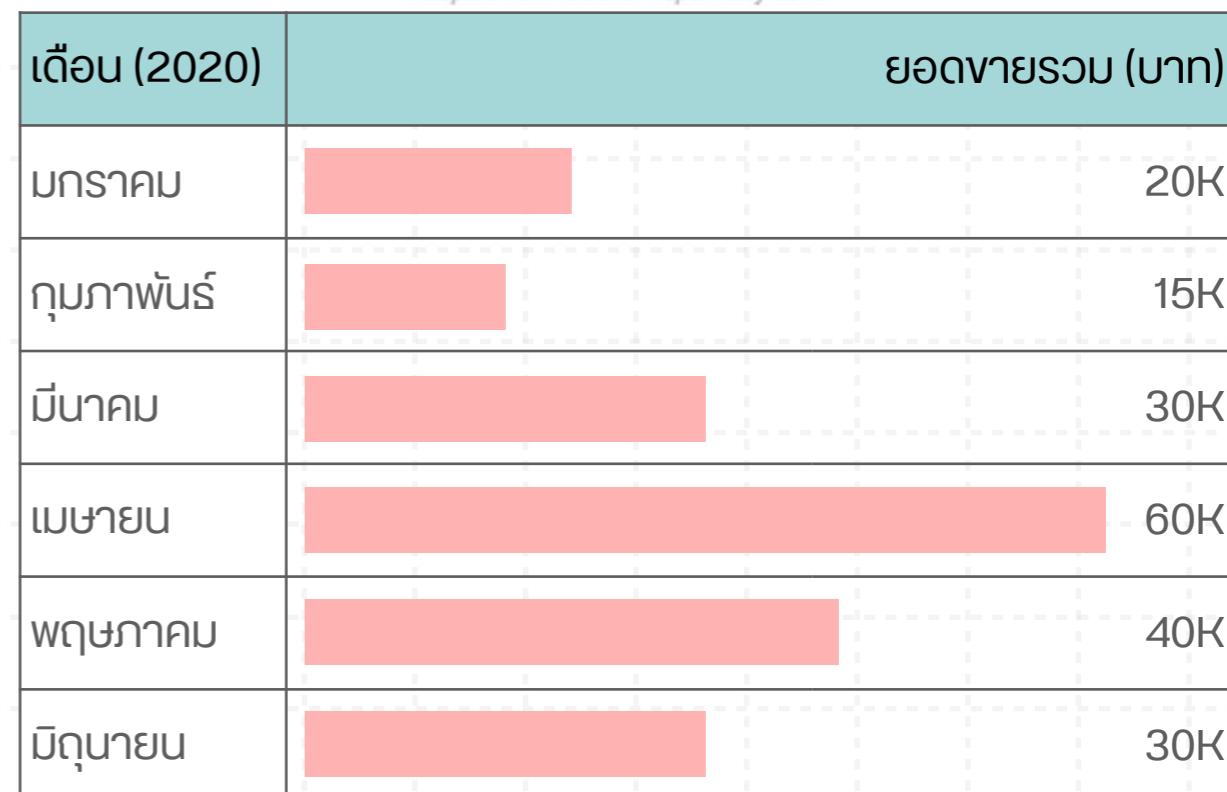
คนมีส่วนเกี่ยวข้องในการคิดวิเคราะห์

และตัดสินใจค่อนข้างมาก

(สังเกตุได้จากสีฟ้าในaccoverของรูป)



Adapted from Gartner Report July 2015



ยอดขายเครื่องดื่มใน 6 เดือนแรกของปี 2020

Data Analytics: Descriptive

- ## • การวิเคราะห์เชิงพรรณนา (Descriptive)

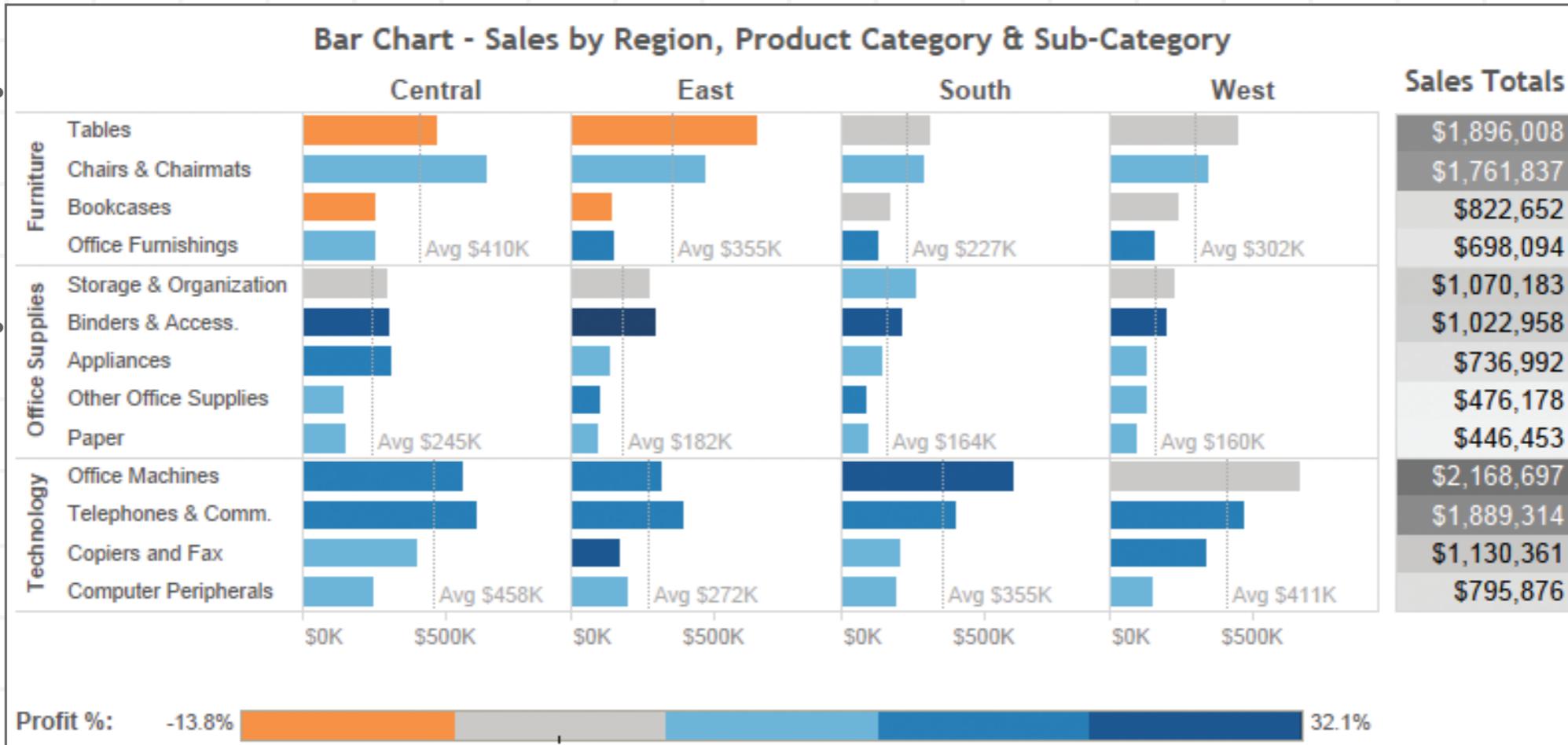


FIGURE 1-2 Sales Mix Analysis using a bar chart and heat map

- คบมีส่วนเกี่ยวข้องในการตัดวิเคราะห์และตัดสินใจค่อนข้างมาก
(สังเกตุได้จากสีฟ้าในแควแกรกของรูป)

image source: Tableau your data!

Data Analytics: Descriptive

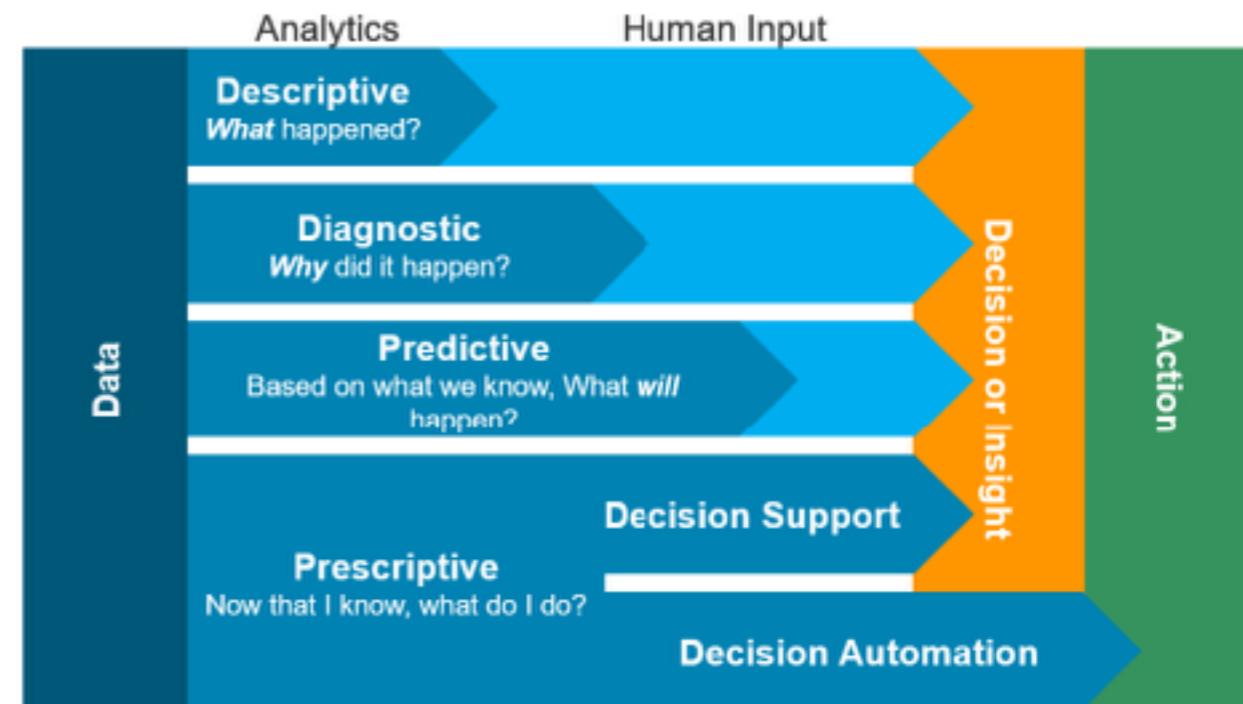
สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- จากรูปเป็นการแสดงยอดขายของสินค้าประเภทต่างๆ (แสดงในแบบแอด) ก็ง่ายได้ใน **ภูมิภาคต่างๆ** (แสดงในแนวคอกลั่นบี)
- ยอดขายแสดงด้วย Bar graph ยิ่ง **มีขนาดยาวมาก** แสดงว่าขายสินค้าได้เป็นปริมาณมาก โดยมีเส้นประ (dash line) เพื่อแสดงยอดขายเฉลี่ยของสินค้าประเภทนั้นๆ ในแต่ละภูมิภาค
- สีบ่งบอกถึงผลกำไร โดยที่
 - **สีส้ม** คือ ผลกำไรติดลบ (-)
 - **สีเขียวเข้ม** คือ กำไรเป็นบวก (+)
- จุดที่นำสังเกตุคือสินค้าบางประเภท เช่น Tables **เป็นสินค้าที่มียอดขายได้เยอะมาก** (มากกว่าค่าเฉลี่ย) ในภูมิภาค Central และ East แต่ว่ามีสีส้ม นั่นหมายถึงว่า **มีกำไรที่ติดลบ (-)**

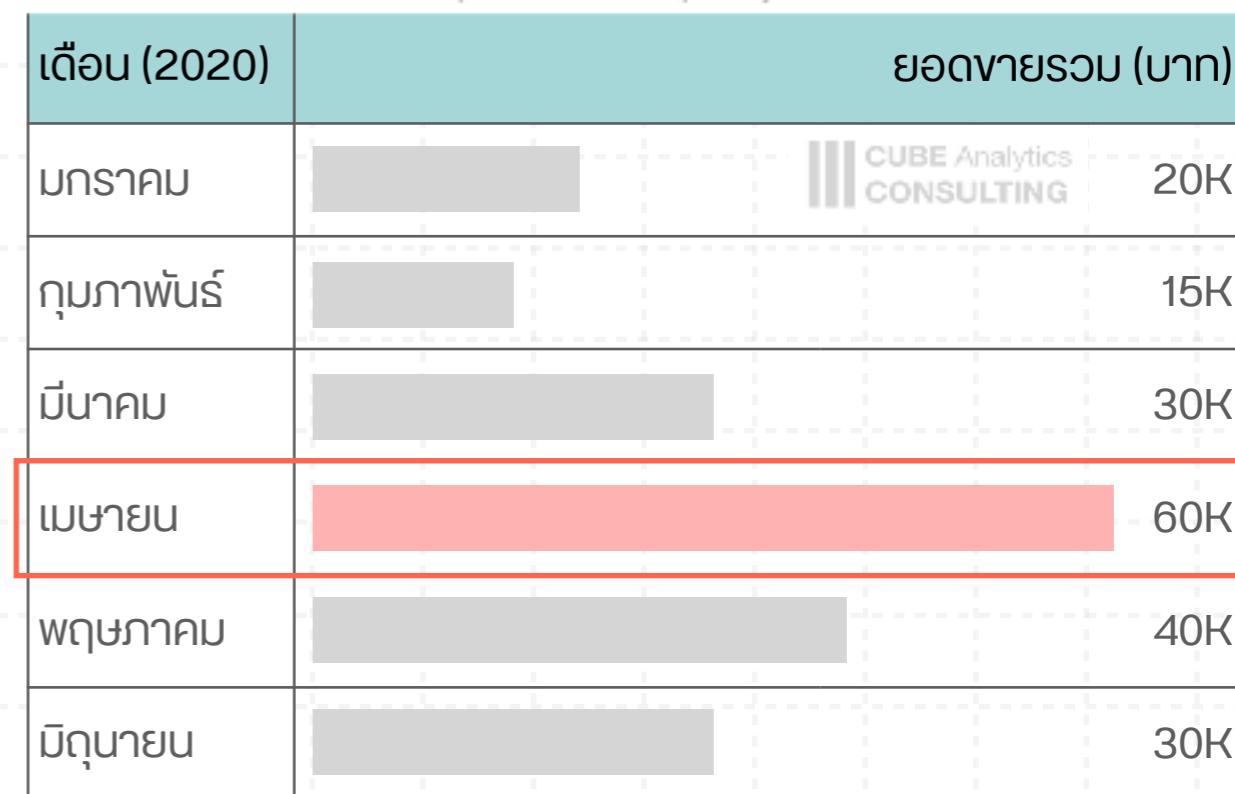
Data Analytics: Diagnostic

สำหรับการบรรยายในงาน Data

- การวิเคราะห์เชิงวินิจฉัย (Diagnostic)
 - พิจารณาข้อมูลในอดีตเพื่อ **หาเหตุผลว่าทำมาถึงเกิดเหตุการณ์แบบนี้ขึ้น** (Why did it happen?)
 - ส่วนใหญ่จะเป็นการ **พิจารณาหลายตัวแปรร่วมกัน เช่น**
 - ยอดขายที่หายได้ดีในเดือนเมษายนมีเหตุผลมาจากเป็นฤดูร้อนคนมีความกระหายเครื่องดื่มน้ำมากกว่าปกติ
 - คน **มีส่วนเกี่ยวข้องในการตัดวิเคราะห์และตัดสินใจยังมากอยู่** (สังเกตุได้จากสีฟ้าในดาวเทาของรูป)



Adapted from Gartner Report July 2015

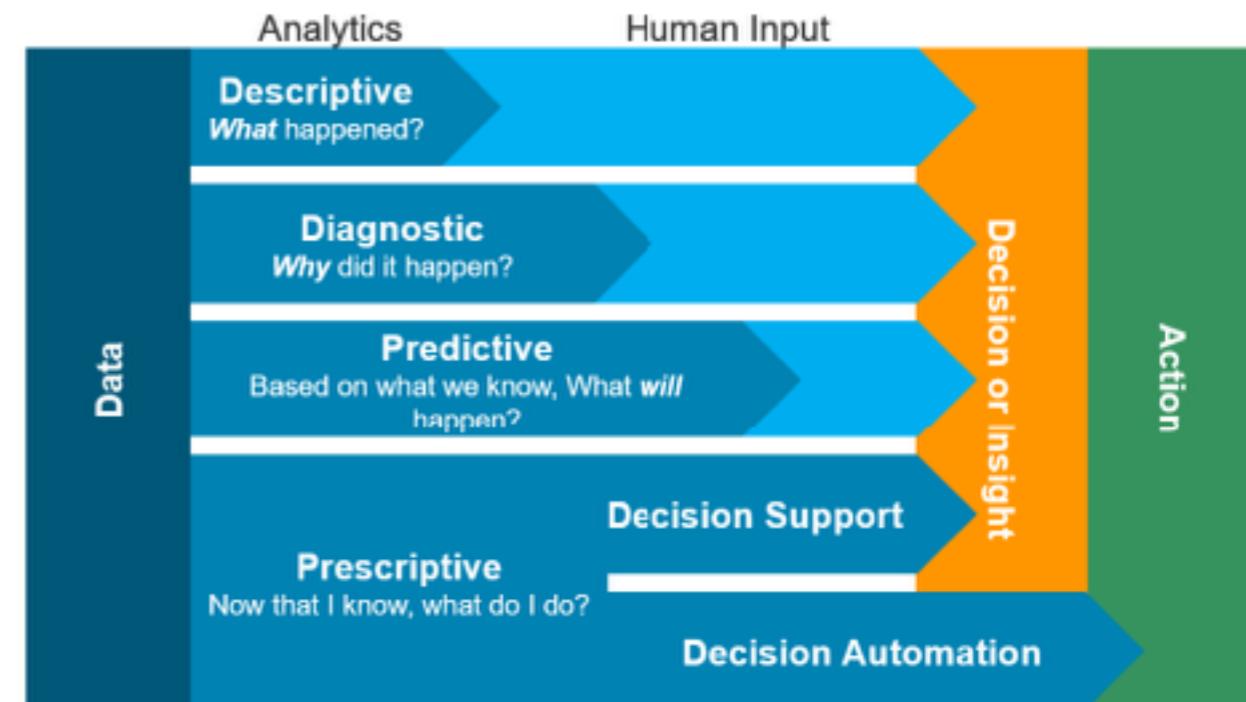


ยอดขายเครื่องดื่มใน 6 เดือนแรกของปี 2020

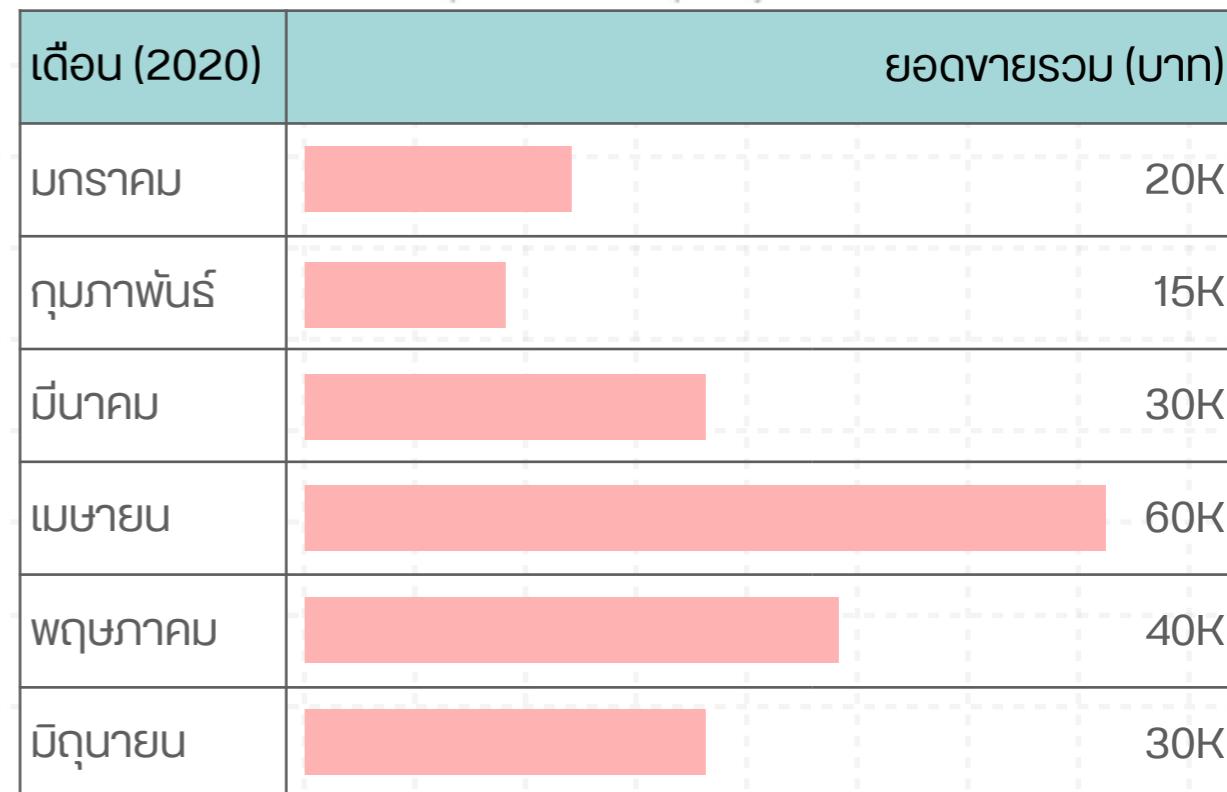
Data Analytics: Predictive

สำหรับการบรรยายในงาน De-

- การวิเคราะห์เชิงพยากรณ์ (Predictive)
 - ใช้ข้อมูลในอดีตเพื่อ **คาดการณ์สิ่งที่จะเกิดขึ้นในอนาคต** หรือใช้ตอบคำถามว่าจะเป็นไรเกิดขึ้นต่อไป (What will happen?)
 - คบมีส่วนเกี่ยวข้องในการคิดวิเคราะห์และตัดสินใจน้อยลง เนื่องจากระบบได้ทำการพยากรณ์สิ่งที่จะเกิดขึ้นมาให้ (สังเกตุได้จากสีฟ้าในแหวนแรกของรูป) เช่น
 - พยากรณ์จำนวนยอดขายที่จะขายได้ในเดือนถัดๆ ไปของปี 2021

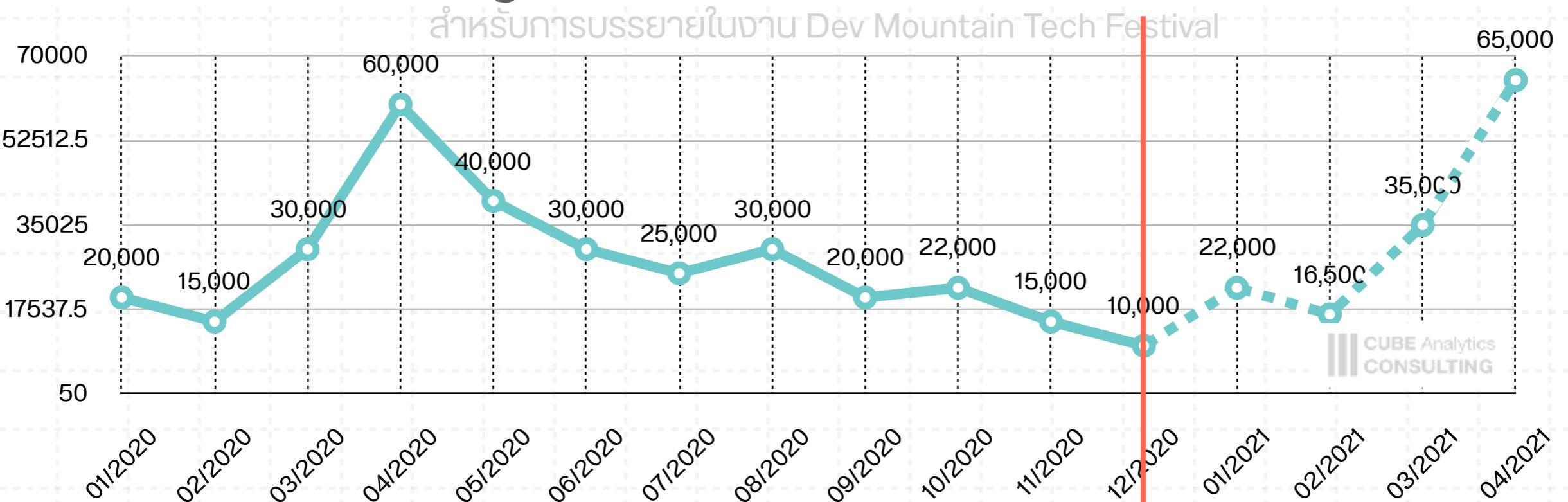


Adapted from Gartner Report July 2015



ยอดขายเครื่องดื่มใน 6 เดือนแรกของปี 2020

Data Analytics: Predictive



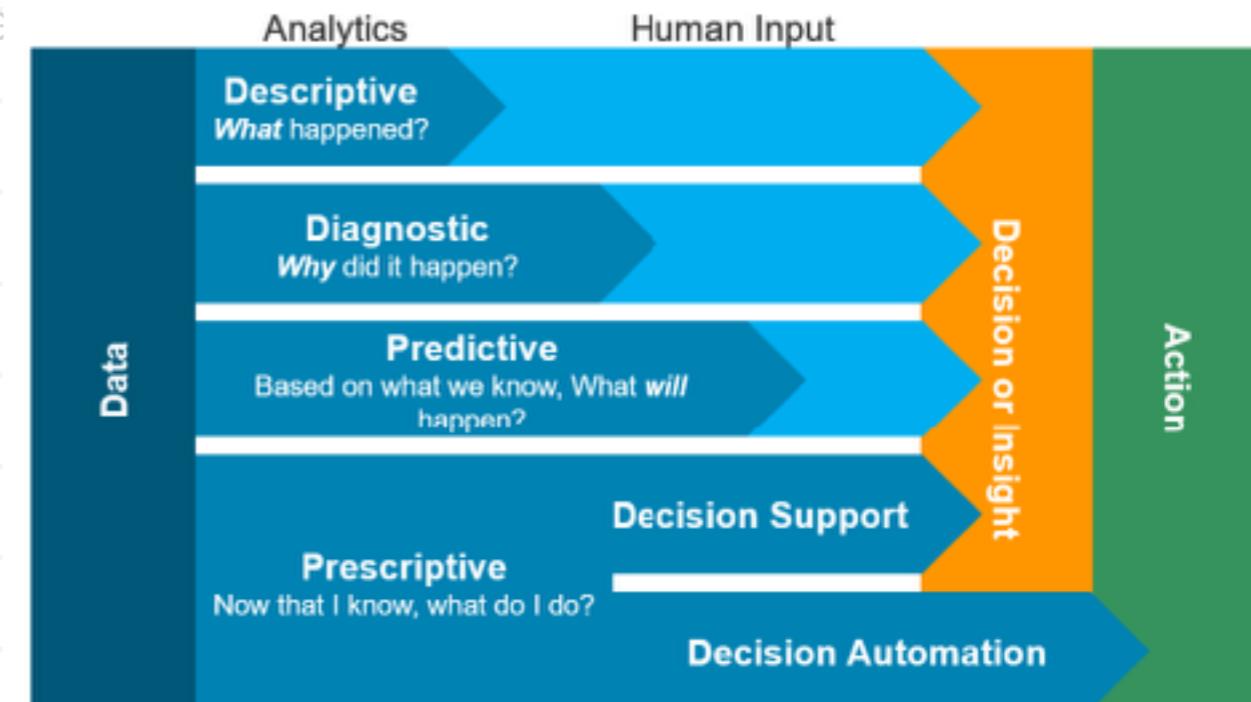
- ตัวอย่างแสดงยอดขายสินค้าของเครื่องดื่มชานชา

 - แสดงข้อมูลยอดขายในแต่ละเดือนของปี 2020
 - แสดงข้อมูลพยากรณ์ยอดขายในแต่ละเดือนของ 4 เดือนแรกปี 2021

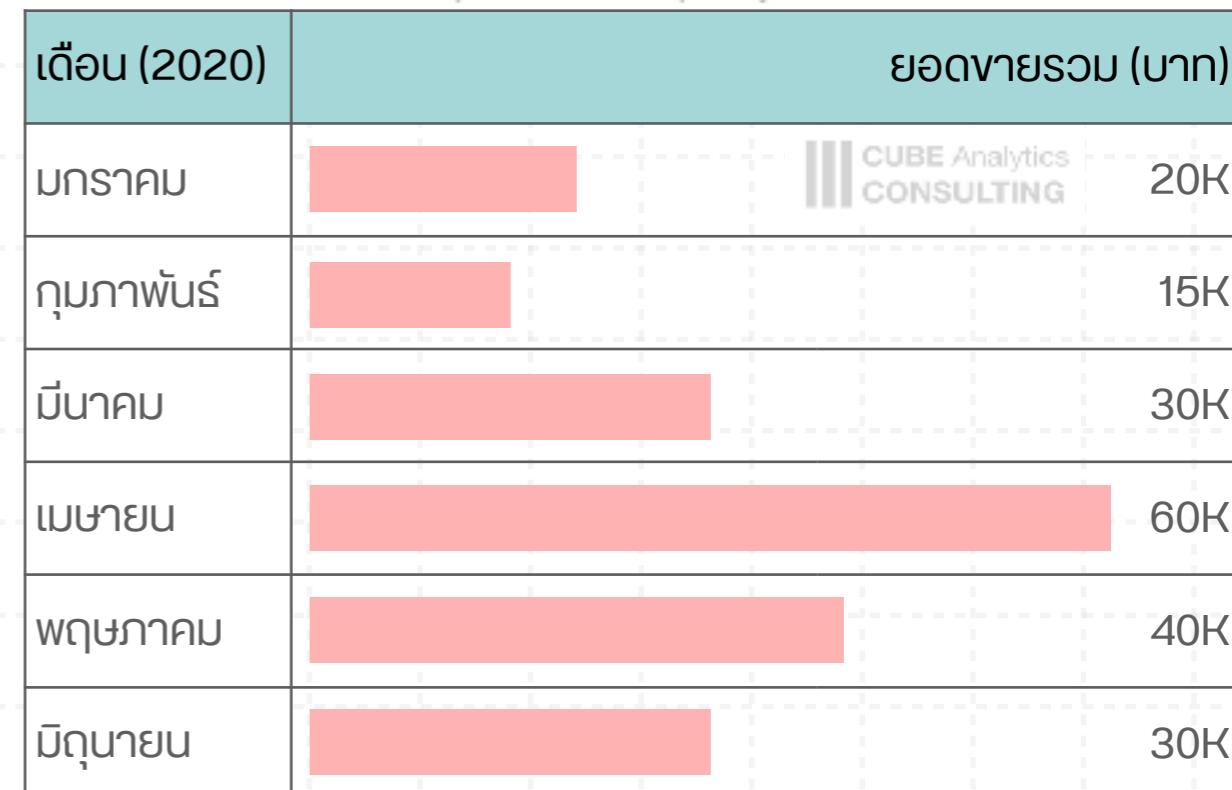
Data Analytics: Prescriptive

สำหรับการบรรยายในงาน Data

- การวิเคราะห์เชิงแบนนำ (Prescriptive)
 - ใช้ข้อมูลในอดีตเพื่อ **คาดการณ์สิ่งที่จะเกิดขึ้นในอนาคต** พร้อม **แนะนำแนวทางที่เหมาะสม** (What do I do?)
 - คน **ไม่มีส่วนเกี่ยวข้อง** ในการคิดวิเคราะห์และตัดสินใจเลย เป็นองค์ประกอบได้ **กำหนดพยากรณ์สิ่งที่จะเกิดขึ้นมาให้และแนะนำแนวทางที่เหมาะสมให้ด้วย** เช่น
 - พยากรณ์จำนวนยอดขายที่จะขายได้ในเดือนถัดๆ ไปของปี 2021 และแนะนำว่าจะต้องจัดจำหน่ายมาก่อนเท่านั้น



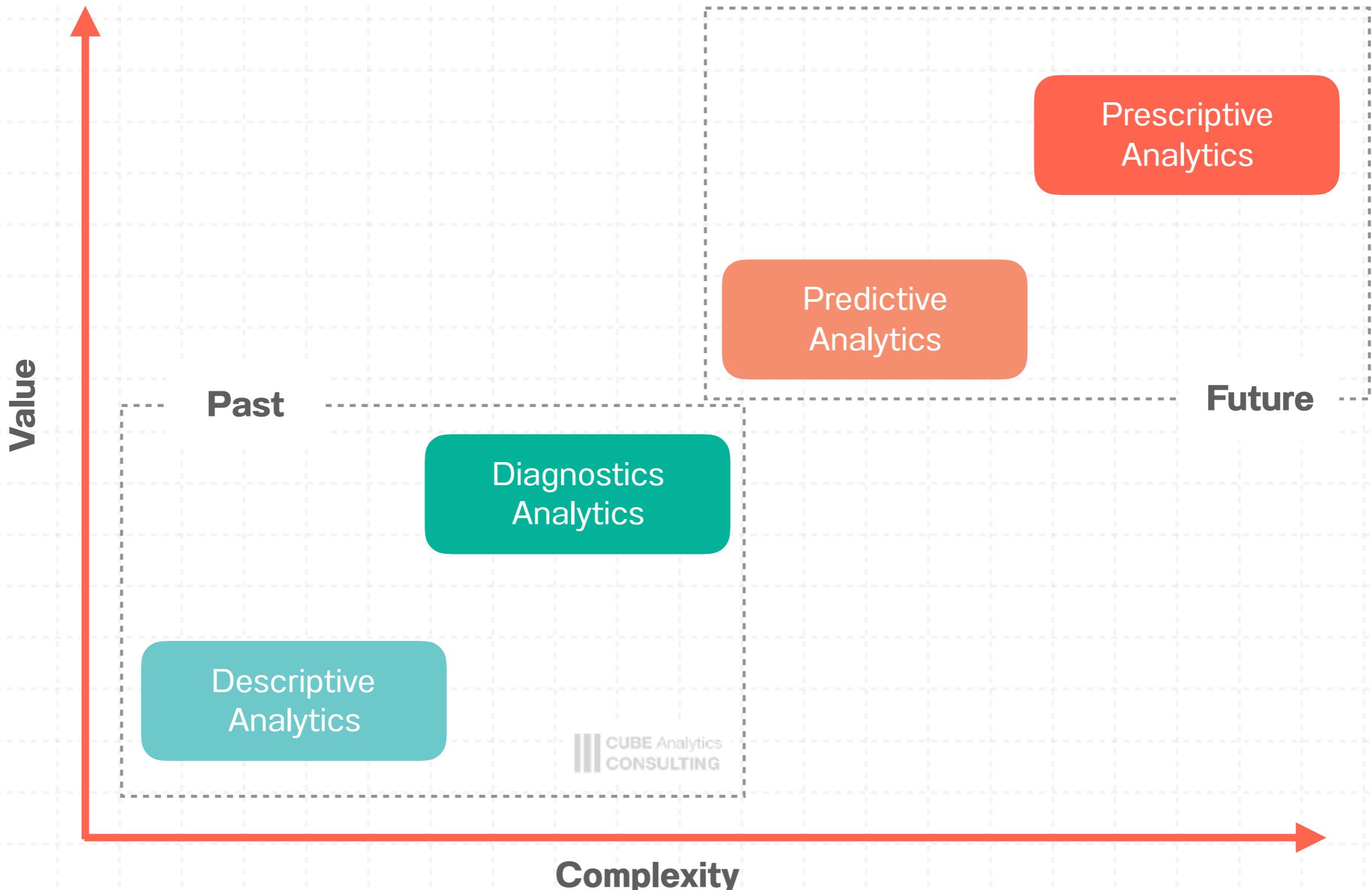
Adapted from Gartner Report July 2015



ยอดขายเครื่องดื่มใน 6 เดือนแรกของปี 2020

Data Analytics

สำหรับการบรรยายในงาน Dev Mountain Tech Festival



Outlines

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

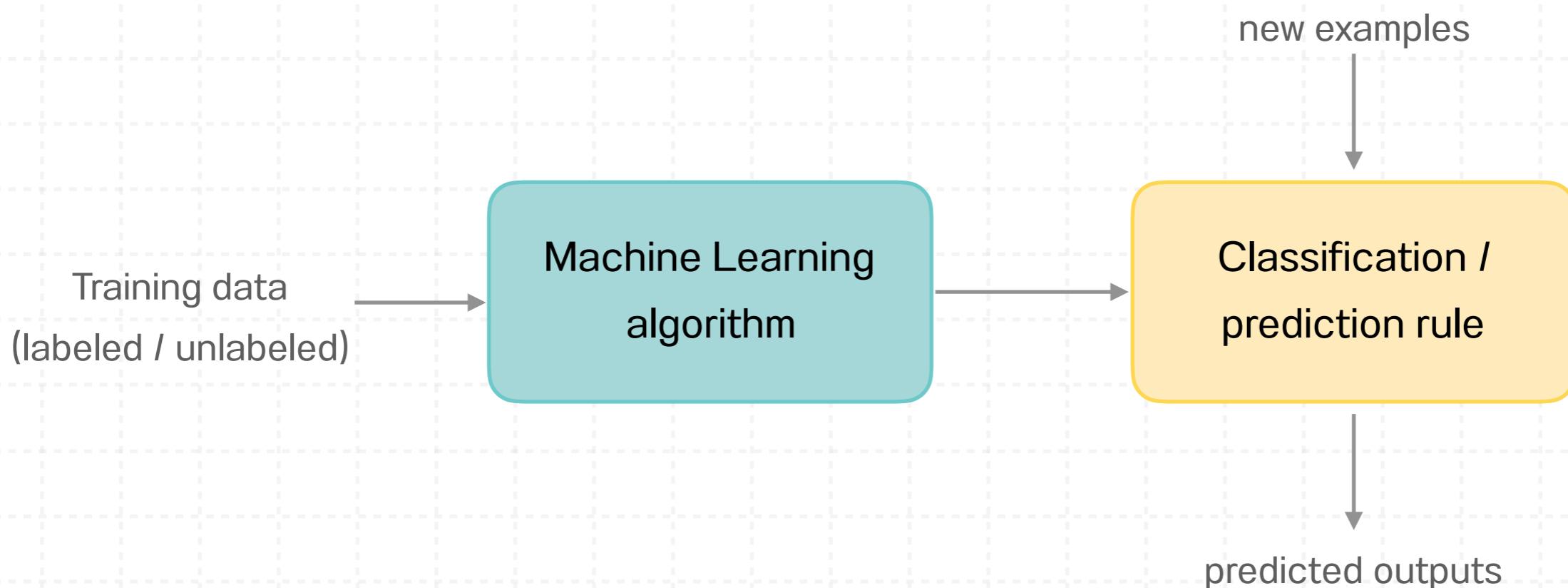
- Machine Learning Applications
- Data Management
- Data Roles
- Data Analytics
- **Machine Learning**
- Machine Learning Development Steps
- No-Code/Low-Code Machine Learning Platform

Machine Learning methodology

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

“ Typically, ML methodologies **involves a learning process** with the objective **to learn from “experience”** (training data) to perform a task. ”

– Machine Learning in Agriculture: A Review



Types of Machine Learning

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- แบ่งได้เป็น 2 ประเภทใหญ่ๆ คือ
 - **Unsupervised Learning** เรียกว่า “**การเรียนรู้แบบไม่มีผู้สอน**”
 - เป็นการ**จัดกลุ่มข้อมูลที่มีลักษณะคล้ายคลึงกันเข้าไว้ด้วยกัน**
 - ตัวอย่างเทคโนโลย เช่น K-Means Clustering และ DBSCAN
 - **Supervised Learning** เรียกว่า “**การเรียนรู้แบบมีผู้สอน**”
 - เป็นการ**เรียนรู้จากข้อมูลในอดีตและนำมาสร้างโมเดล (model) เพื่อพยากรณ์** คำตอบให้กับข้อมูลใหม่
 - ตัวอย่างเทคโนโลย เช่น Decision Tree, Linear Regression และ Logistic Regression

Types of Machine Learning

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- **Unsupervised Learning (การเรียนรู้แบบไม่มีผู้สอน)**

- สร้างตัวแปรให้กับข้อมูลแต่ละตัวอย่าง
- แบ่งกลุ่มข้อมูลตามความคล้ายคลึงกันออกเป็นกลุ่มต่างๆ

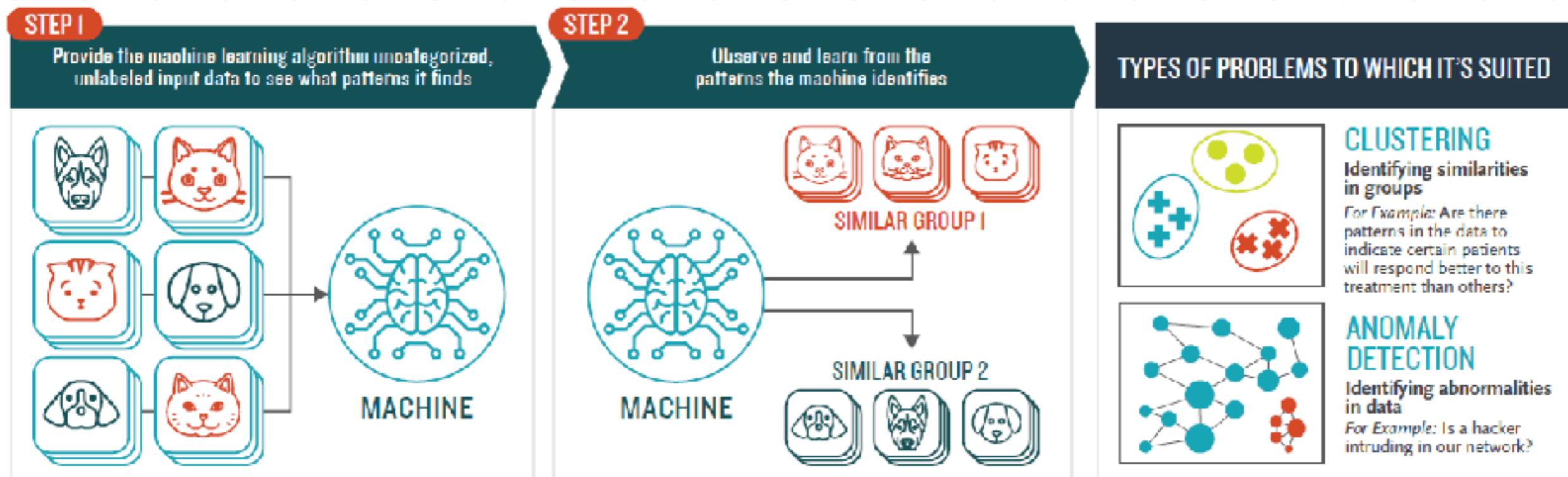


image source: <https://towardsdatascience.com/wtf-is-machine-learning-a-quick-guide-39457e49c65b>

Types of Machine Learning

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- **Supervised Learning (การเรียนรู้แบบมีผู้สอน)**

- สร้าง training data โดยการกำหนดคำต่อ (label) ให้กับข้อมูลแต่ละตัวอย่าง
- สร้างโมเดล (model) จากข้อมูล training data ที่เตรียมไว้

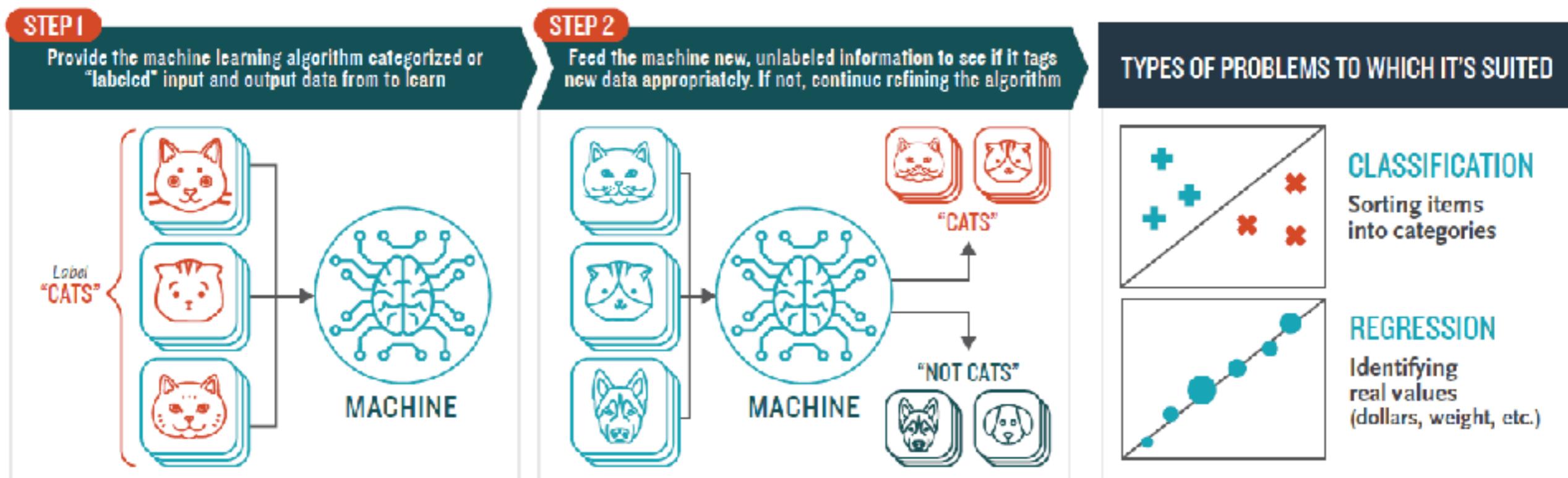


image source: <https://towardsdatascience.com/wtf-is-machine-learning-a-quick-guide-39457e49c65b>

Types of Machine Learning

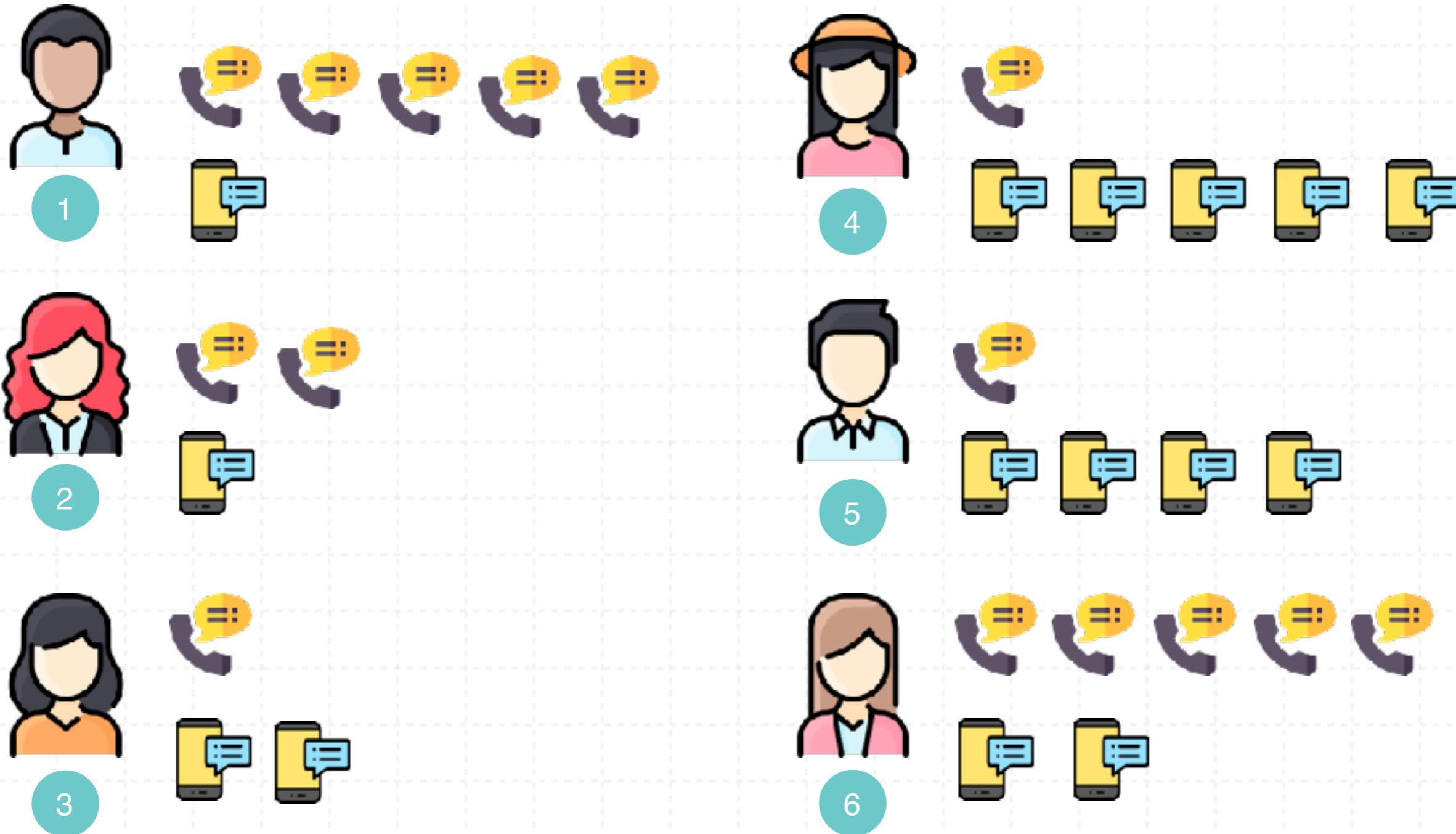
สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- แบ่งได้เป็น 2 ประเภทใหญ่ๆ คือ
 - **Unsupervised Learning** เรียกว่า “**การเรียนรู้แบบไม่มีผู้สอน**”
 - เป็นการจัดกลุ่มข้อมูลที่มีลักษณะคล้ายคลึงกันเข้าไว้ด้วยกัน
 - ตัวอย่างเทคโนโลยี เช่น K-Means Clustering และ DBSCAN
 - **Supervised Learning** เรียกว่า “**การเรียนรู้แบบมีผู้สอน**”
 - เป็นการเรียนรู้จากข้อมูลในอดีตและนำมาสร้างโมเดล (model) เพื่อพยากรณ์ค่าตอบแทนให้กับข้อมูลใหม่
 - ตัวอย่างเทคโนโลยี เช่น Decision Tree, Linear Regression และ Logistic Regression

Customer profile

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

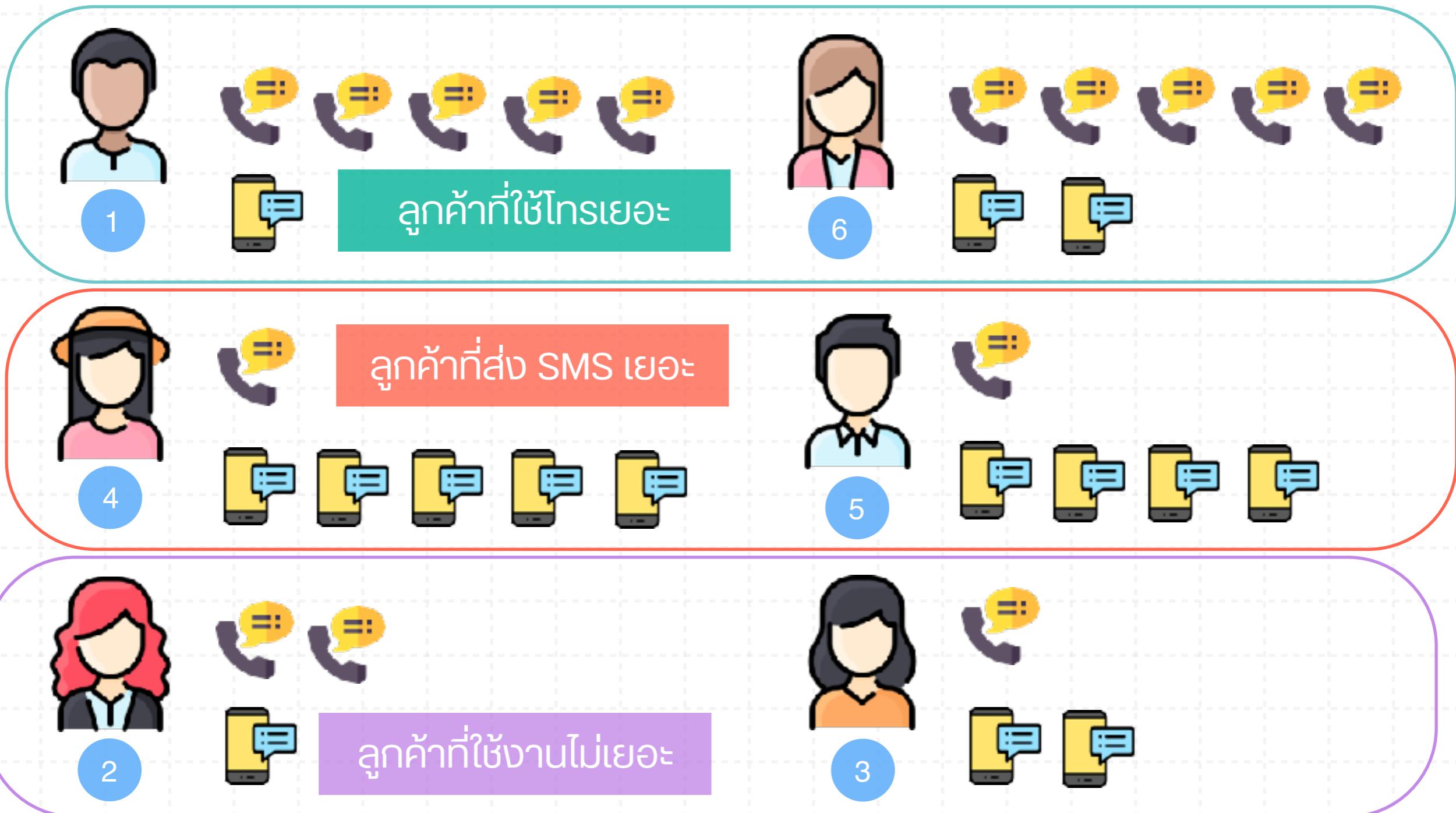
- การใช้งานโทรศัพท์มือถือของลูกค้า



Customer profile

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

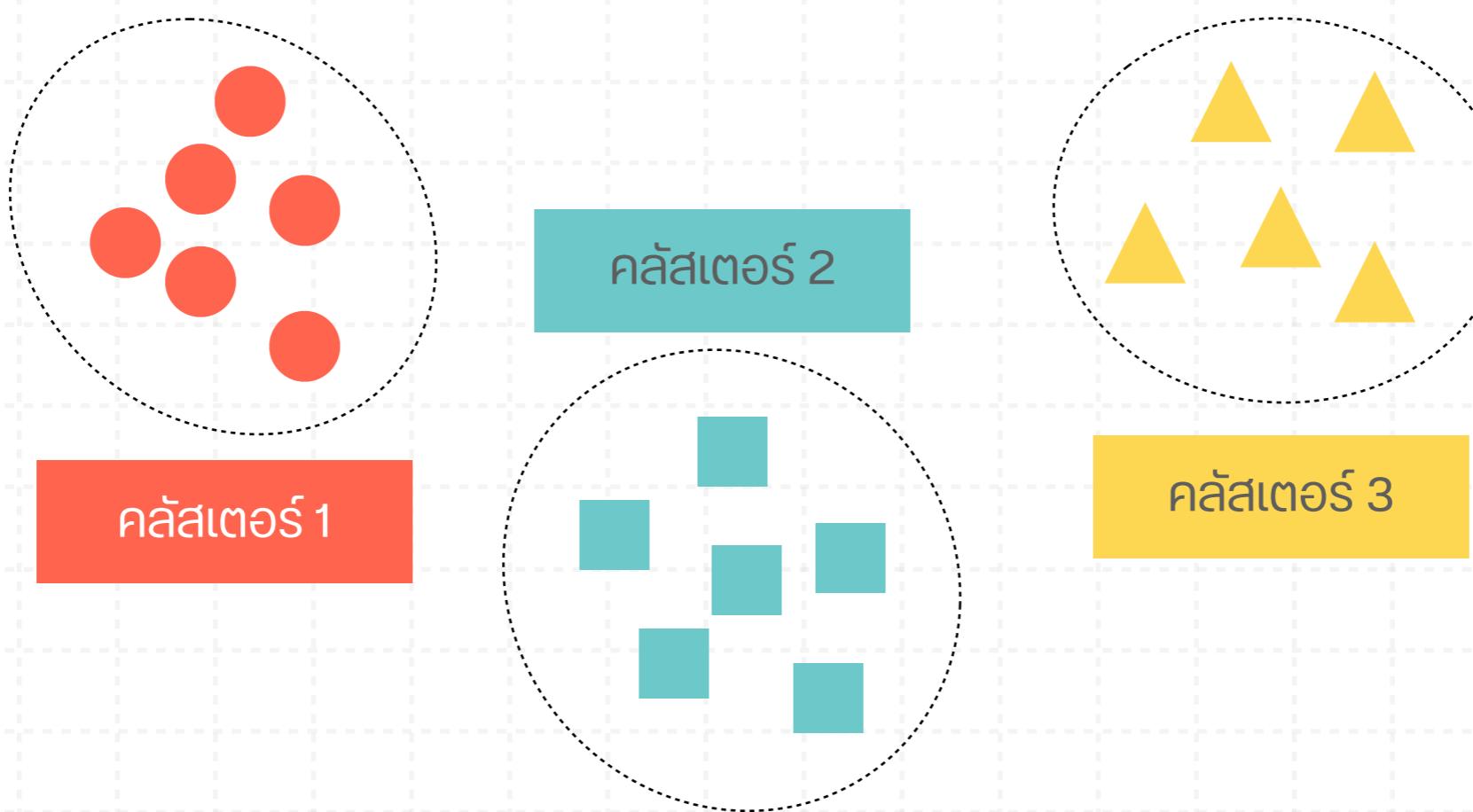
- หลังจากการแบ่งกลุ่มลูกค้าตามพฤติกรรมการใช้งาน



What is clustering?

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

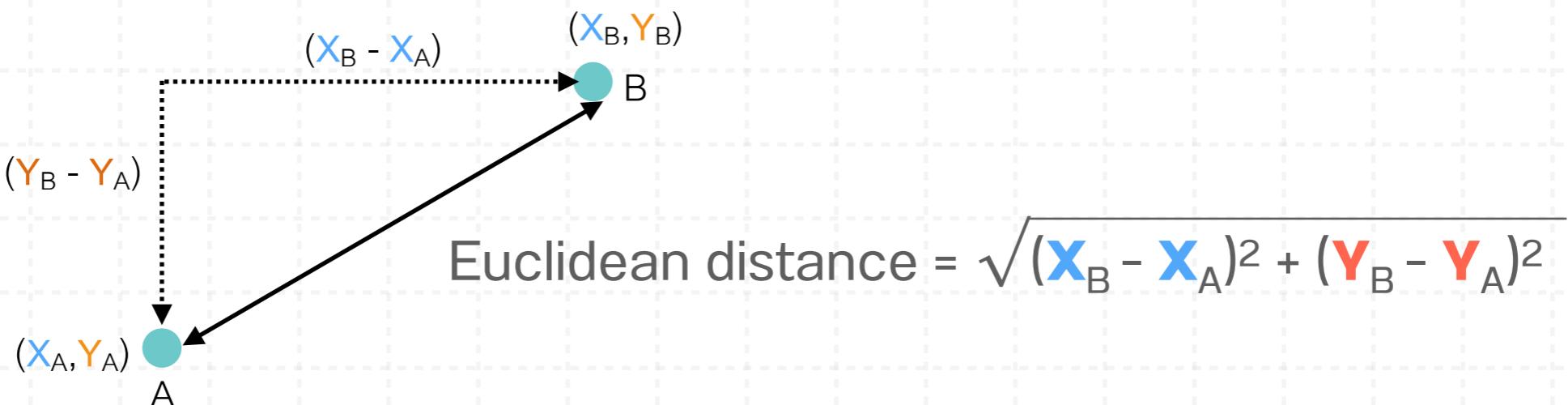
- Clustering คือ การแบ่งกลุ่มข้อมูล
 - ข้อมูลที่มีลักษณะคล้ายๆ กัน อยู่กลุ่มเดียวกัน
 - ข้อมูลที่อยู่คนละกลุ่มจะมีลักษณะที่แตกต่างกันมากๆ
- แต่ละกลุ่มจะเรียกว่า คลัสเตอร์ (cluster)



Distance function

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- การจัดข้อมูลให้อยู่ในกลุ่มต่างๆ จะต้องมีการวัดค่าความคล้ายคลึง (similarity) หรือ ค่าระยะห่าง (distance) ระหว่างข้อมูลแต่ละตัว
- ค่าระยะห่างที่นิยมใช้
 - ระยะห่างยูclidean (Euclidean distance)
 - ผลรวมของระยะห่างกำลังสอง



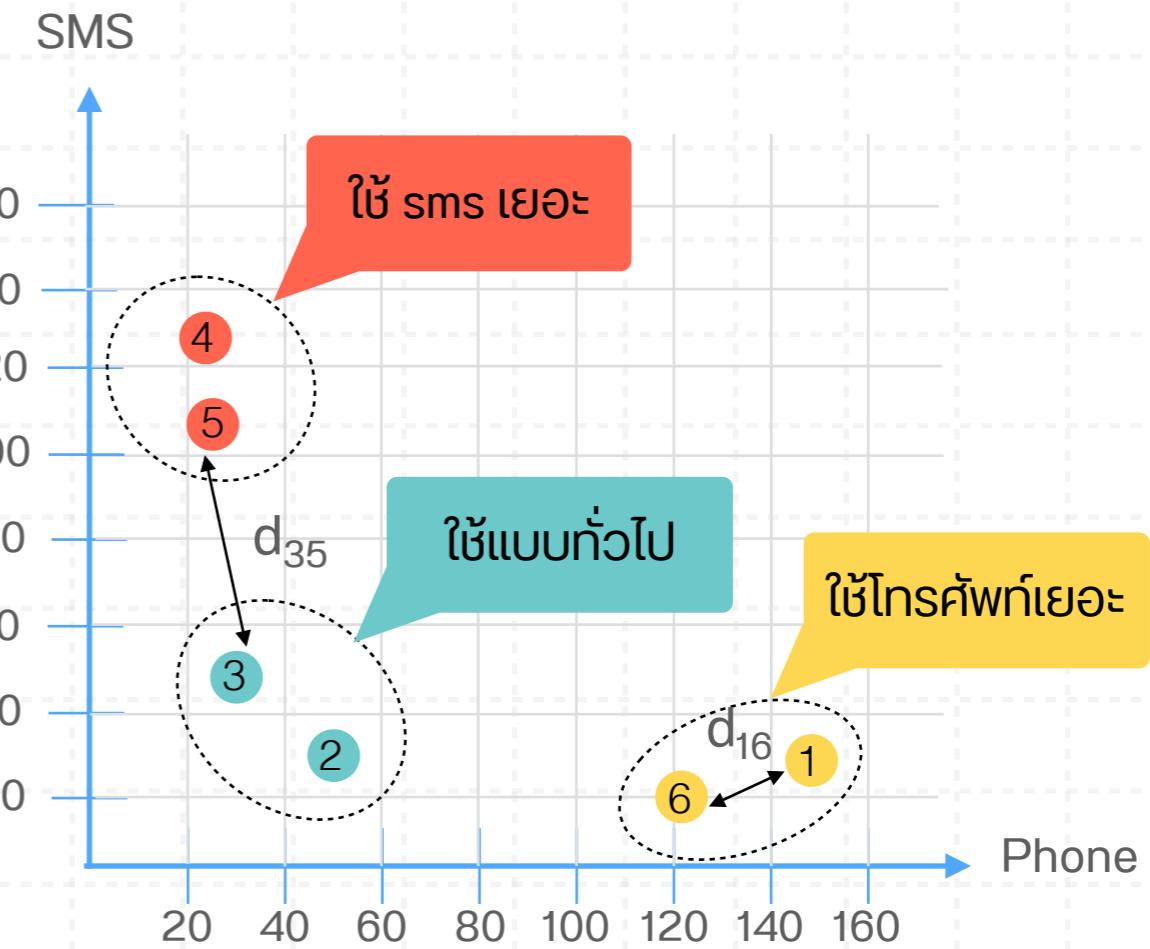
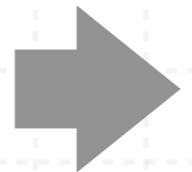
Clustering example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- ทำการ clustering ลูกค้าตามพฤติกรรมการใช้งาน SMS และ Phone

Phone

ID	Phone	SMS
1	144	27
2	44	32
3	30	41
4	21	125
5	23	105
6	121	20



Applications

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- แบ่งข้อมูลลูกค้าออกเป็นกลุ่มย่อยๆ เพื่อจะได้เข้าใจพฤติกรรมการบริโภคของลูกค้าได้ดีขึ้น
- ส่งโปรโมชันได้ตรงกับความต้องการของลูกค้าแต่ละกลุ่ม
- ช่วยในการจัดกลุ่มเอกสารตามความคล้ายคลึง

K-Means

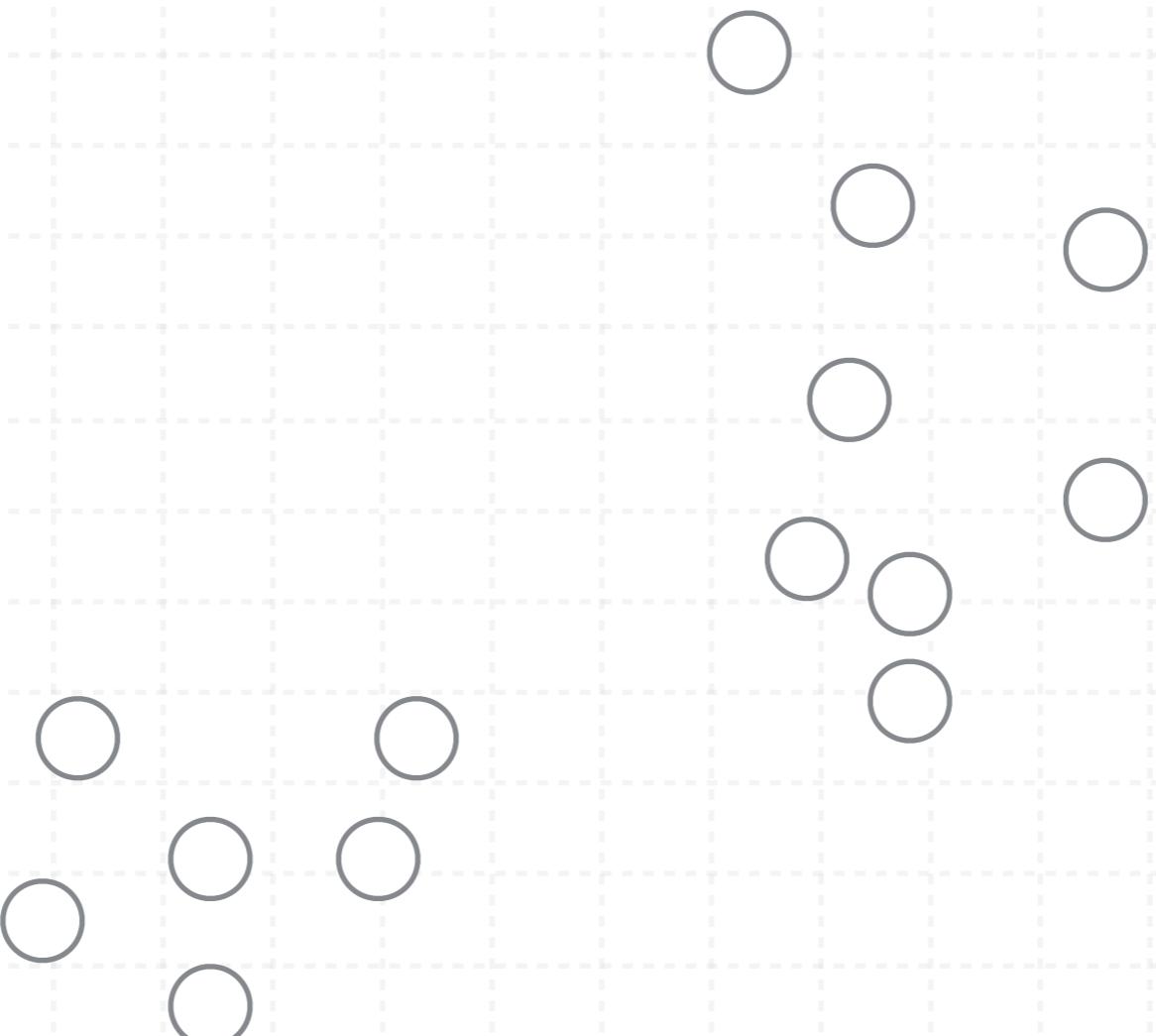
สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- นิยมใช้ในการทำ clustering
- เป็นเทคนิคแบบ partitional clustering
- ต้องกำหนดจำนวนกลุ่ม (หรือคลัสเตอร์) ที่ต้องการจะแบ่ง ในตัวแปร K
- วิธีการทำงาน
 - กำหนดจุดศูนย์กลางของแต่ละคลัสเตอร์
 - หาระยะห่างระหว่างข้อมูลแต่ละตัวกับจุดศูนย์กลาง (mean) ของแต่ละคลัสเตอร์
 - กำหนดให้ข้อมูลที่อยู่ในคลัสเตอร์ที่ใกล้สุด
 - คำนวณหาจุดศูนย์กลางของแต่ละคลัสเตอร์ใหม่
 - ทำการวนซ้ำจนข้อมูลอยู่ในคลัสเตอร์เดิมไม่มีการเปลี่ยนแปลง

K-Means

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

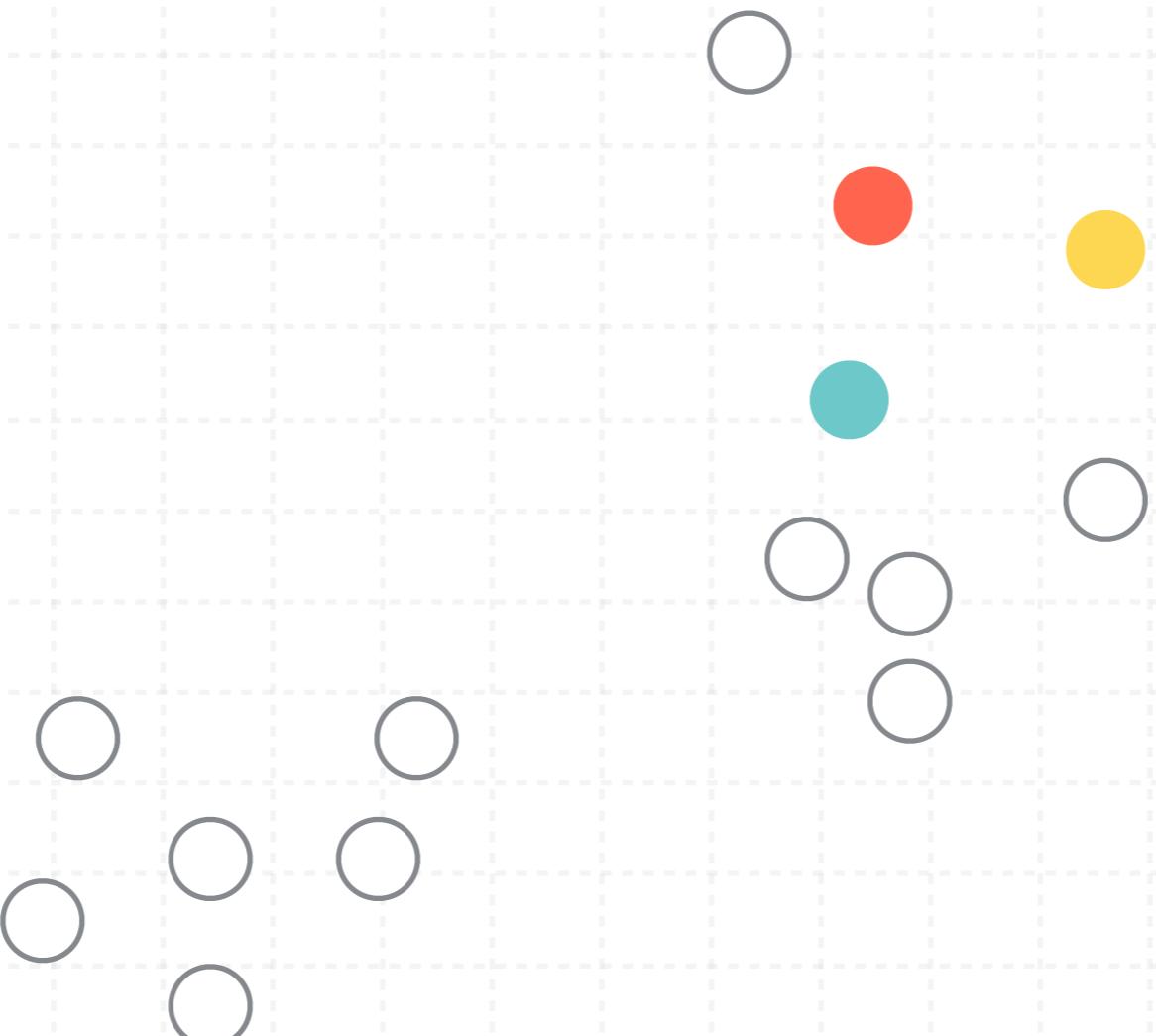
- แบ่งข้อมูลนี้ด้วยเทคนิค K-Means ออกเป็น 3 กลุ่ม



K-Means

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

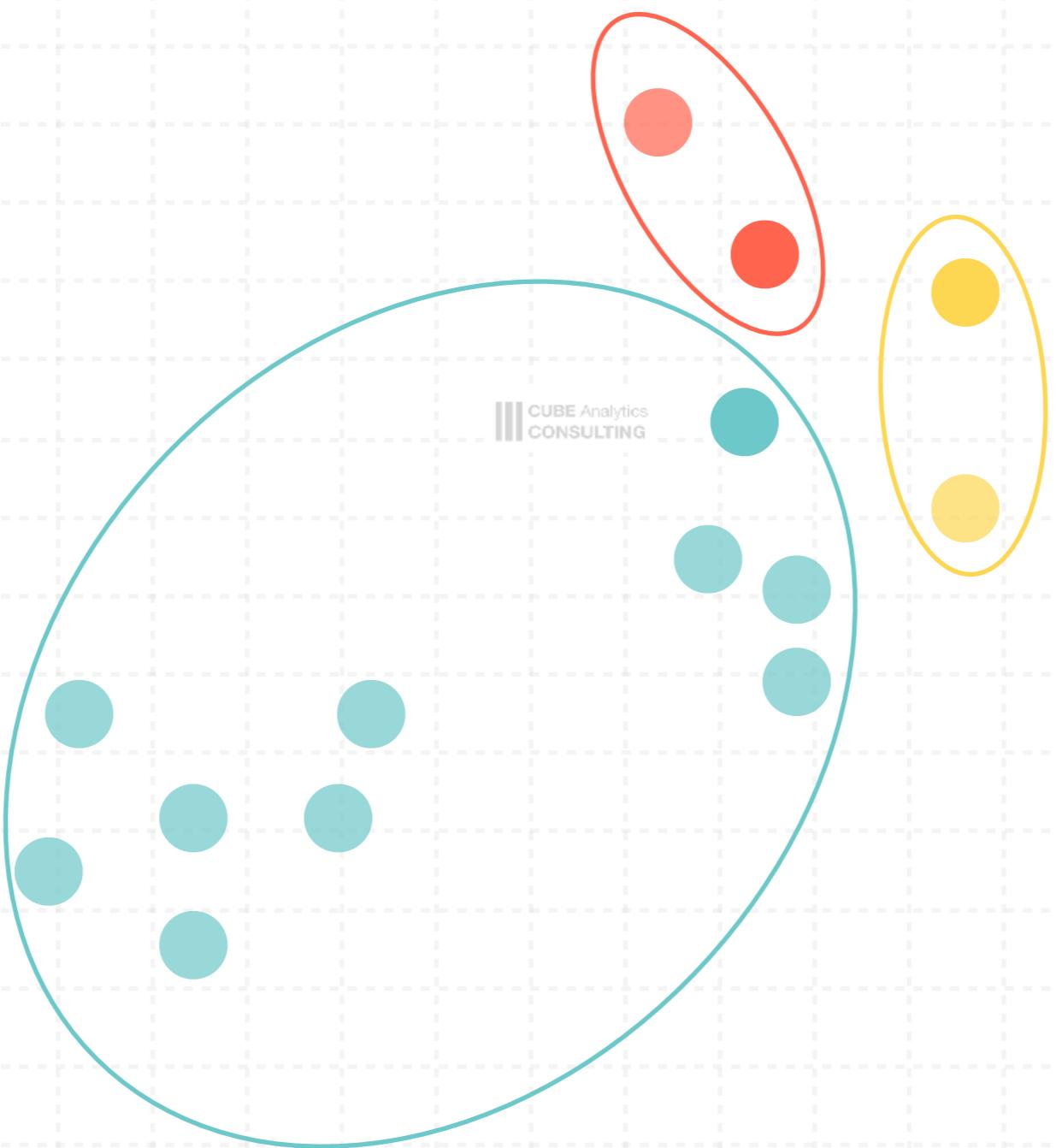
- ทำการสุ่มข้อมูล 3 ตัวเพื่อเป็นจุดศูนย์กลางของคลัสเตอร์



K-Means

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

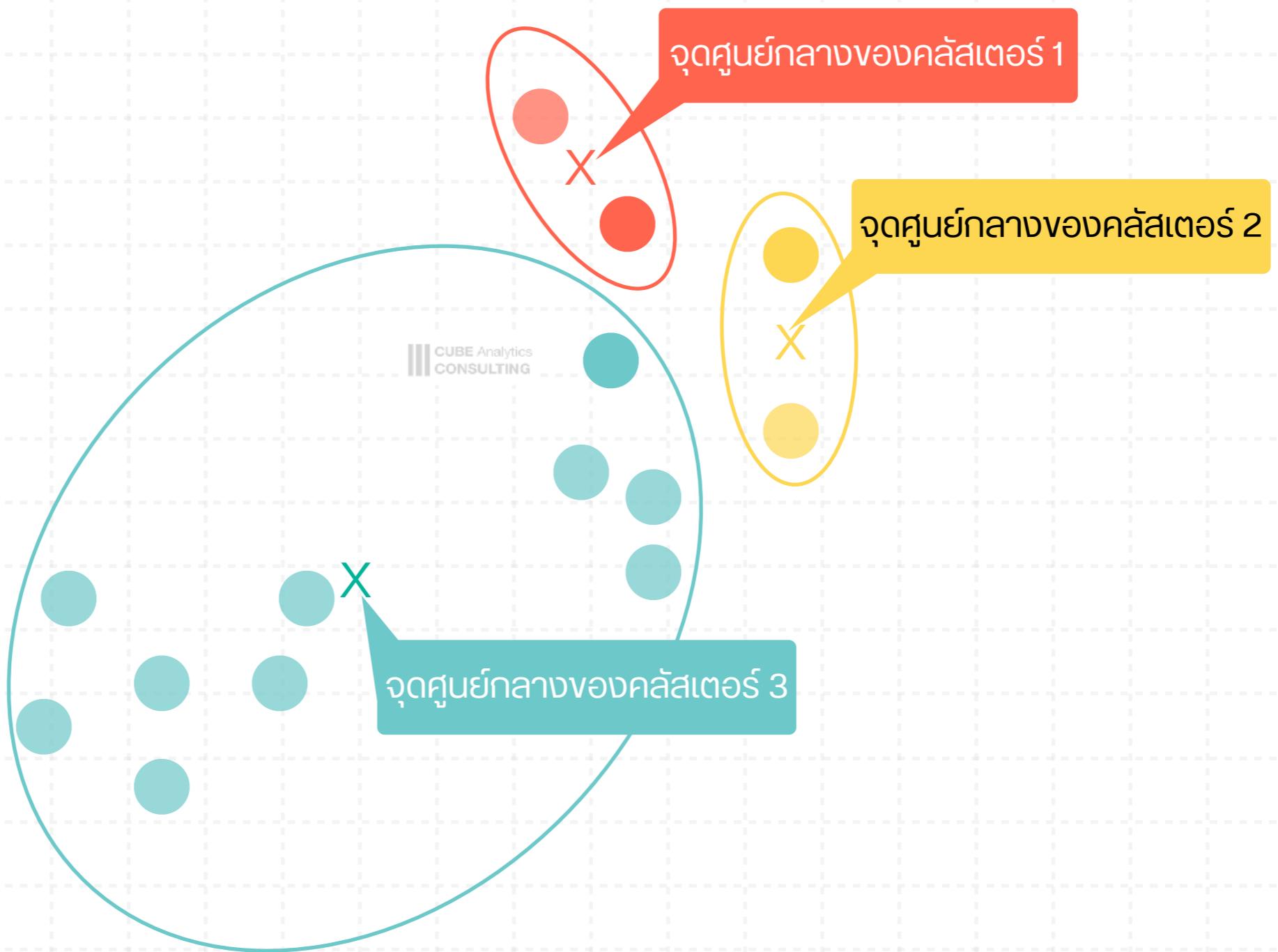
- จัดให้ข้อมูลแต่ละตัวอยู่ในกลุ่มที่มีระยะห่างจากจุดศูนย์กลางน้อยสุด



K-Means

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

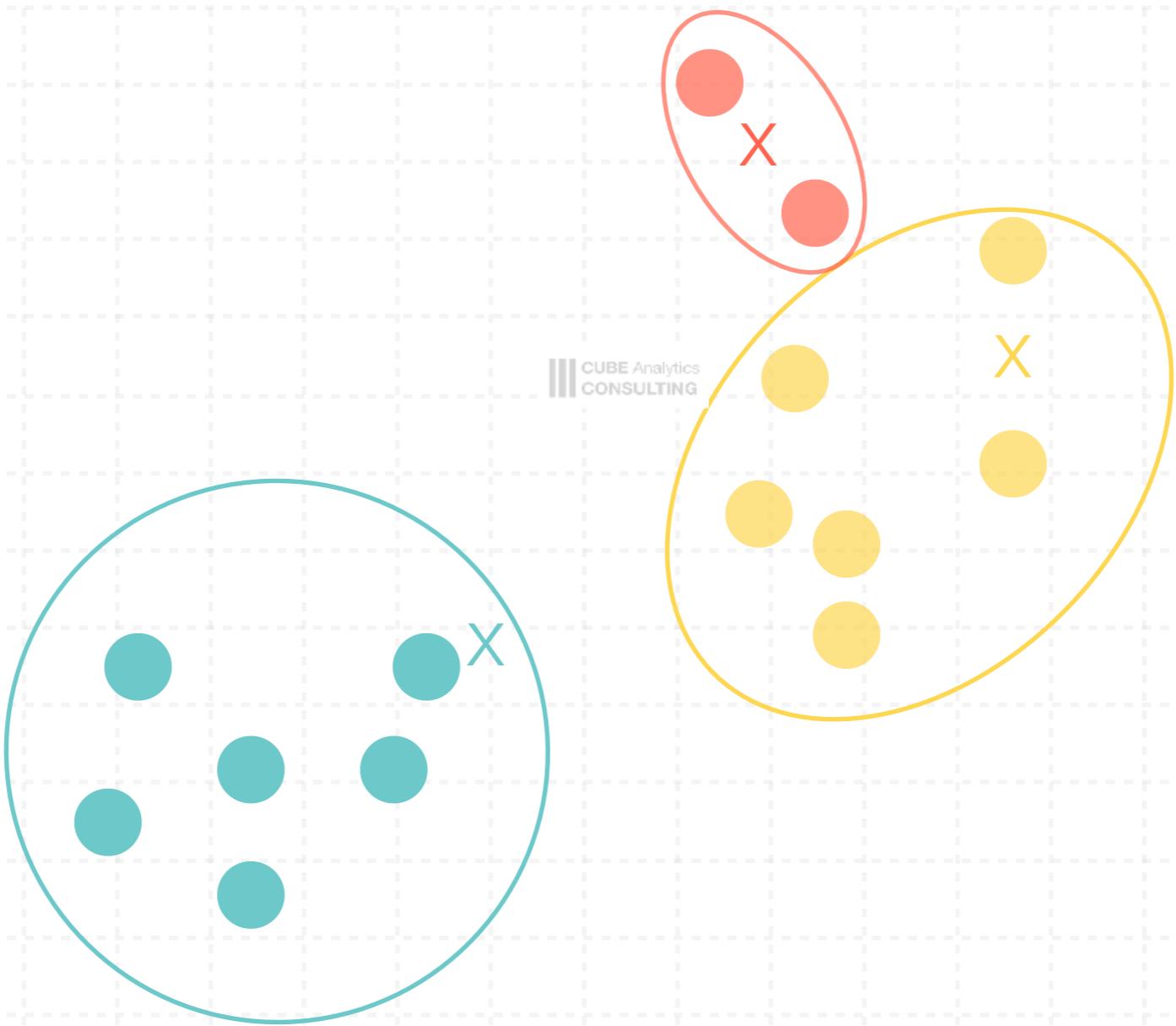
- หาจุดศูนย์กลางของแต่ละคลัสเตอร์



K-Means

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

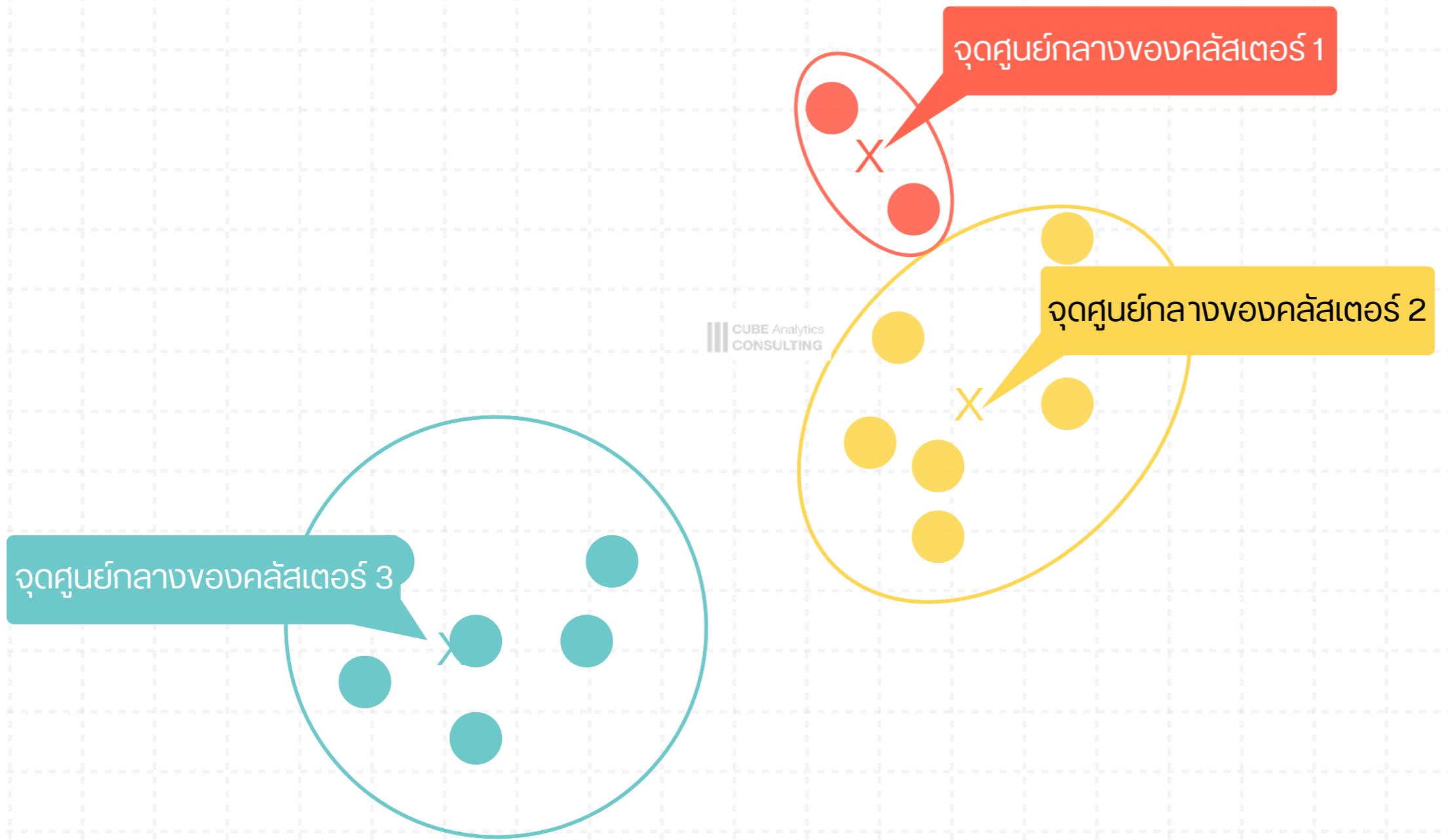
- วัดระยะห่างเทียบกับจุดศูนย์กลางใหม่อีกครั้ง



K-Means

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

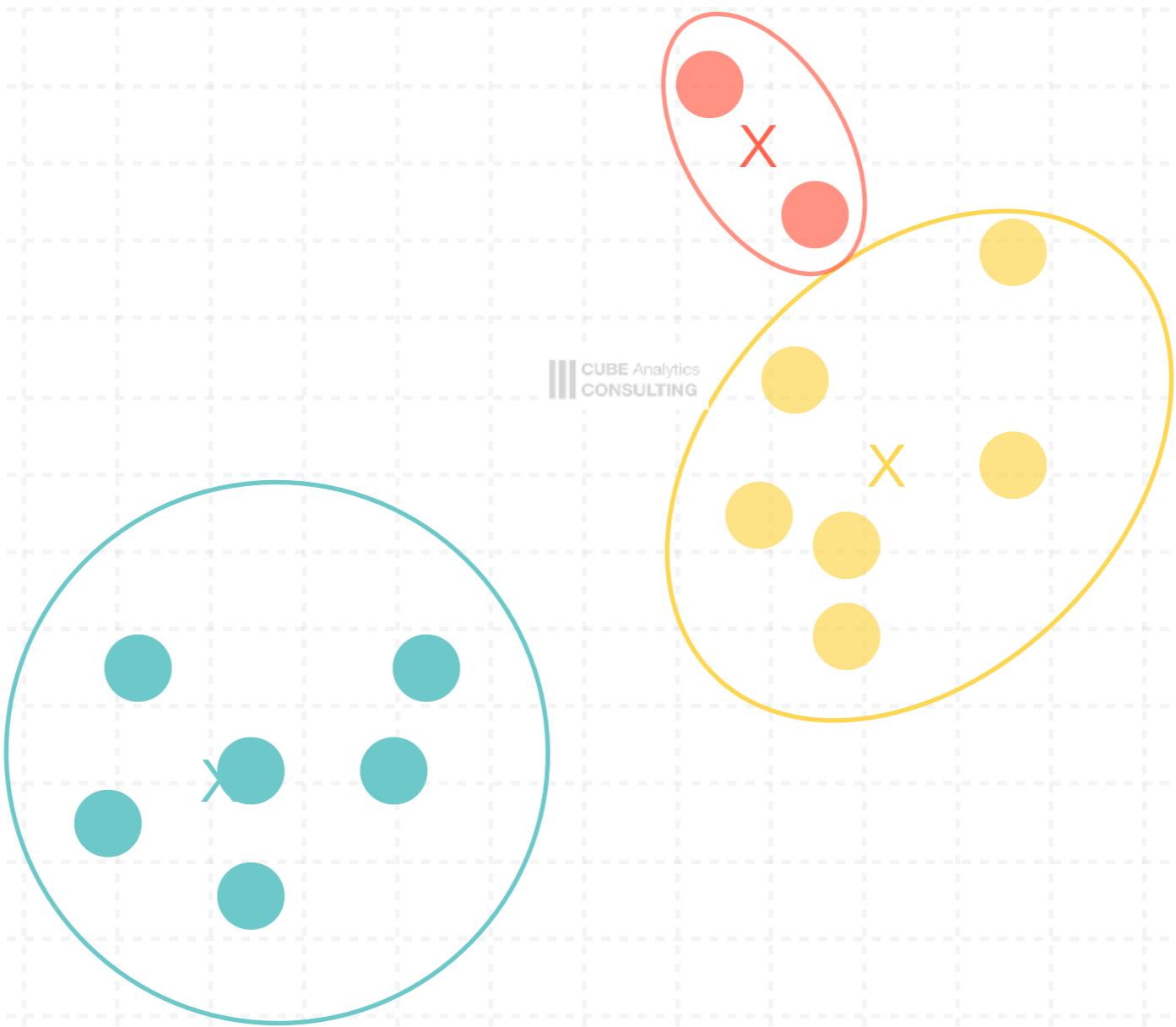
- หาจุดศูนย์กลางของแต่ละคลัสเตอร์ใหม่อีกครั้ง



K-Means

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

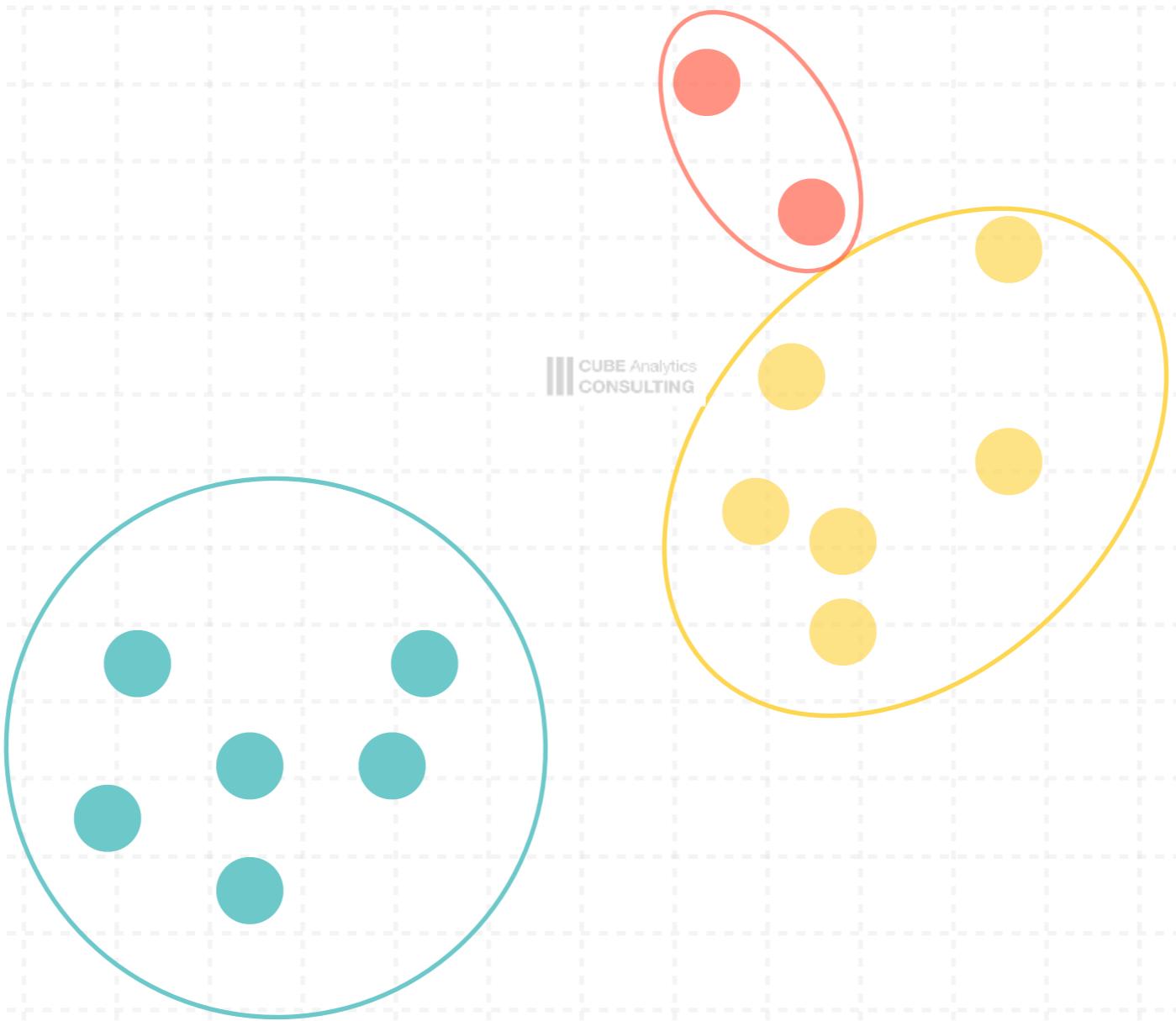
- วัดระยะห่างเทียบกับจุดศูนย์กลางใหม่อีกครั้ง



K-Means

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- พบว่าข้อมูลไม่มีการเปลี่ยนแปลงจึงหยุดการทำงาน

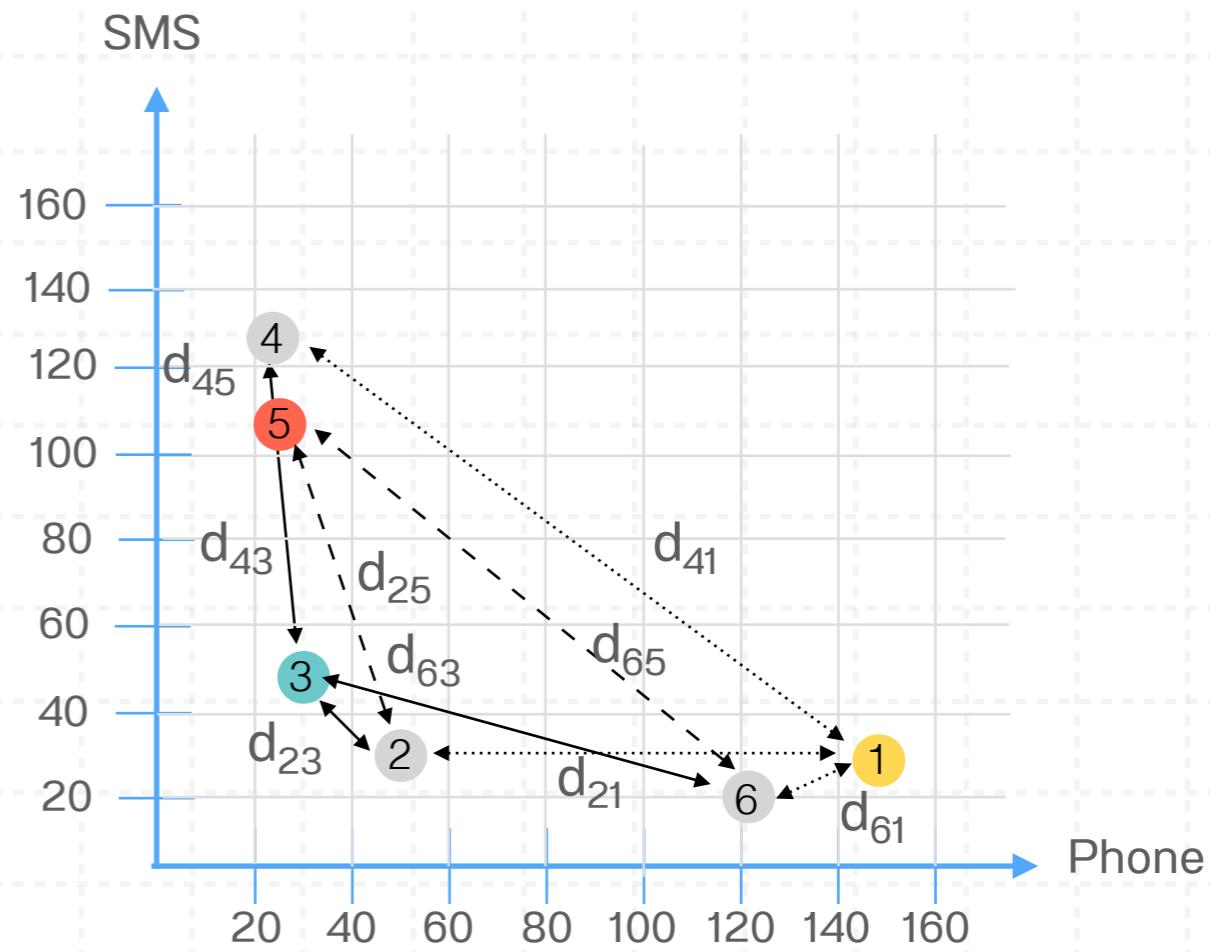


K-Means

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- กำหนด $K = 3$
- กำหนดให้ข้อมูลหมายเลข 1,3,5 เป็นจุดศูนย์กลางของแต่ละคลัสเตอร์
- คำนวณระยะห่างระหว่างจุดศูนย์กลางของคลัสเตอร์

ID	Phone	SMS	Center
1	144	27	X
2	44	32	
3	30	41	X
4	21	125	
5	23	105	X
6	121	20	



K-Means

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- คำนวณระยะห่างด้วย Euclidean distance

ID	distance to 1	distance to 3	distance to 5
2	$\sqrt{(32-27)^2 + (44-144)^2} = 100.12$	$\sqrt{(32-41)^2 + (44-30)^2} = 16.64$	$\sqrt{(32-105)^2 + (44-23)^2} = 75.96$
4	$\sqrt{(125-27)^2 + (21-144)^2} = 157.27$	$\sqrt{(125-41)^2 + (21-30)^2} = 84.48$	$\sqrt{(125-105)^2 + (21-23)^2} = 20.10$
6	$\sqrt{(20-27)^2 + (121-144)^2} = 24.04$	$\sqrt{(20-41)^2 + (121-30)^2} = 93.39$	$\sqrt{(20-105)^2 + (121-23)^2} = 129.73$

- ผลการแบ่งคลัสเตอร์รอบแรก
 - คลัสเตอร์ที่ 1 ประกอบด้วยข้อมูลหมายเลข 1,6
 - คลัสเตอร์ที่ 2 ประกอบด้วยข้อมูลหมายเลข 2,3
 - คลัสเตอร์ที่ 3 ประกอบด้วยข้อมูลหมายเลข 4,5

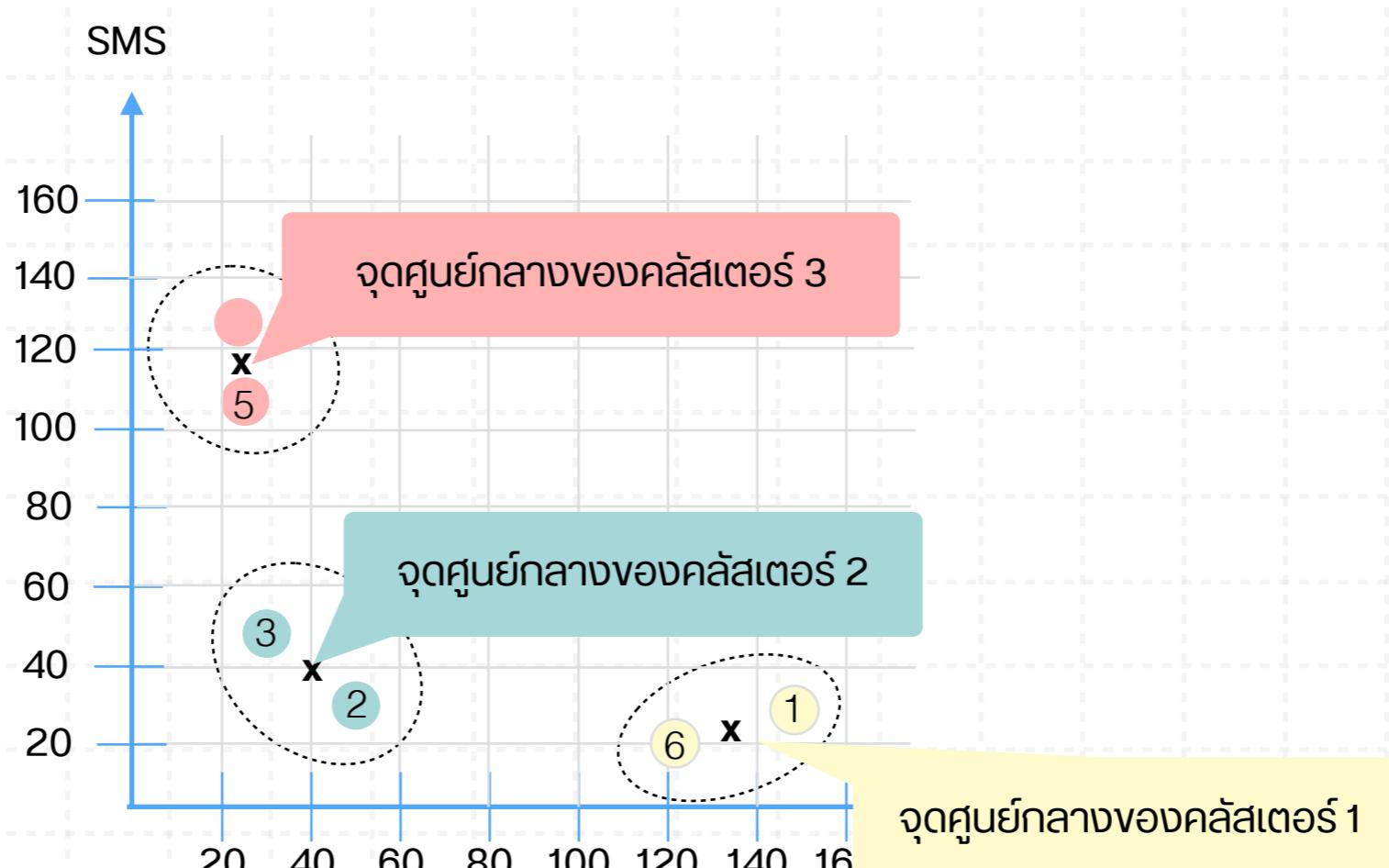
K-Means

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- หาจุดศูนย์กลางของแต่ละคลัสเตอร์

CUBE Analytics
CONSULTING

Attribute	จุดศูนย์กลางของคลัสเตอร์ 1	จุดศูนย์กลางของคลัสเตอร์ 2	จุดศูนย์กลางของคลัสเตอร์ 3
SMS	$(27 + 20)/2 = 23.5$	$(32 + 41)/2 = 36.5$	$(125 + 105)/2 = 115$
Phone	$(144 + 121)/2 = 132.5$	$(44 + 30)/2 = 37$	$(21 + 23)/2 = 22$



K-Means

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- คำนวณระยะห่างกับจุดศูนย์กลางของแต่ละคลัสเตอร์อีกครั้ง

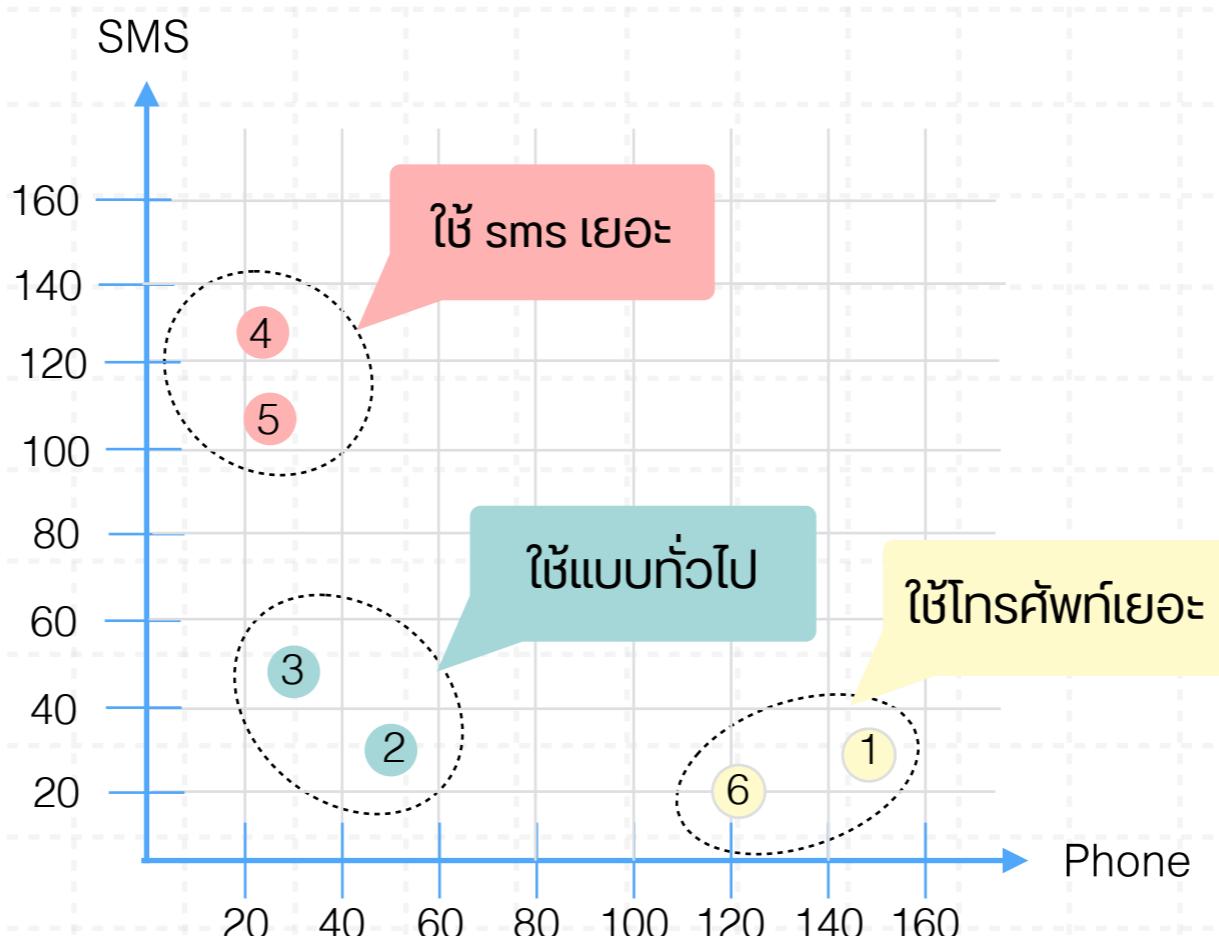
ID	ระยะห่างระหว่างจุดศูนย์กลาง คลัสเตอร์ 1	ระยะห่างระหว่างจุดศูนย์กลาง คลัสเตอร์ 2	ระยะห่างระหว่างจุดศูนย์กลาง คลัสเตอร์ 3
1	$\sqrt{(27-23.5)^2 + (144-132.5)^2} = 12.02$	$\sqrt{(27-36.5)^2 + (144-37)^2} = 107.42$	$\sqrt{(27-115)^2 + (144-22)^2} = 150.43$
2	$\sqrt{(32-23.5)^2 + (44-132.5)^2} = 88.91$	$\sqrt{(32-36.5)^2 + (44-37)^2} = 8.32$	$\sqrt{(32-115)^2 + (44-22)^2} = 85.87$
3	$\sqrt{(41-23.5)^2 + (30-132.5)^2} = 103.98$	$\sqrt{(41-36.5)^2 + (30-37)^2} = 8.32$	$\sqrt{(41-115)^2 + (30-22)^2} = 74.43$
4	$\sqrt{(125-23.5)^2 + (21-132.5)^2} = 150.78$	$\sqrt{(125-36.5)^2 + (21-37)^2} = 89.93$	$\sqrt{(125-115)^2 + (21-22)^2} = 10.05$
5	$\sqrt{(105-23.5)^2 + (23-132.5)^2} = 136.50$	$\sqrt{(105-36.5)^2 + (23-37)^2} = 69.92$	$\sqrt{(105-115)^2 + (23-22)^2} = 10.05$
6	$\sqrt{(20-23.5)^2 + (121-132.5)^2} = 12.02$	$\sqrt{(20-36.5)^2 + (121-37)^2} = 85.61$	$\sqrt{(20-115)^2 + (121-22)^2} = 137.21$

- ข้อมูลทุกตัวอยู่ในคลัสเตอร์เดิม จึงหยุดการทำงาน

K-Means

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- ผลการแบ่งกลั๊สเตอร์
 - กลั๊สเตอร์ที่ 1 ประกอบด้วยข้อมูลหมายเลข 1,6
 - กลั๊สเตอร์ที่ 2 ประกอบด้วยข้อมูลหมายเลข 2,3
 - กลั๊สเตอร์ที่ 3 ประกอบด้วยข้อมูลหมายเลข 4,5



Types of Machine Learning

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- แบ่งได้เป็น 2 ประเภทใหญ่ๆ คือ
 - Unsupervised Learning เรียกว่า “การเรียนรู้แบบไม่มีผู้สอน”
 - เป็นการจัดกลุ่มข้อมูลที่มีลักษณะคล้ายคลึงกันเข้าไว้ด้วยกัน
 - ตัวอย่างเทคโนโลยี เช่น K-Means Clustering และ DBSCAN
 - Supervised Learning** เรียกว่า “**การเรียนรู้แบบมีผู้สอน**”
 - เป็นการเรียนรู้จากข้อมูลในอดีตและนำมาสร้างโมเดล (model) เพื่อพยากรณ์ค่าตอบให้กับข้อมูลใหม่
 - ตัวอย่างเทคโนโลยี เช่น Decision Tree, Linear Regression และ Logistic Regression

Classification example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- ตัวอย่าง spam e-mail classification
 - จะบุ่าว่า e-mail ไหนบ้างที่เป็น spam e-mail

ID	Type	Text
1	spam	Please call our customer service representative on FREE PHONE 0808 145 4742 between 9am-11pm as you have WON a guaranteed £1000 cash
2	spam	You have won \$1,000 cash or a \$2,000 prize! To claim, call 09050000327
3	normal	I'm gonna be home soon and I don't want to talk about this stuff anymore tonight
4	normal	Is that seriously how you spell his name?
5	spam	Double mins and txts 4 6months FREE Bluetooth on Orange. Available on Sony, Nokia phone
6	spam	FREE RINGTONES text FIRST to 87131 for a poly or text GET to 87131 for a true tone!
7	normal	Sorry, I'll you call later in meeting.
8	spam	Congratulations - in this week's competition draw u have won the £1450 prize to claim just call 09050002311
9	normal	Thanks a lot for your wishes on my birthday. Thanks you for making my birthday truly memorable.
10	normal	Hello, What are you doing? Did you attend the training course today?

Classification example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- ตัวอย่าง spam e-mail classification
 - หา keyword ที่ใช้ปัจบุกว่าเป็น spam e-mail

ID	Text	Type
1	Please call our customer service representative on FREE PHONE 0808 145 4742 between 9am-11pm as you have WON a guaranteed £1000 cash	spam
2	You have won \$1,000 cash or a \$2,000 prize! To claim, call 09050000327	spam
3	I'm gonna be home soon and I don't want to talk about this stuff anymore tonight	normal
4	Is that seriously how you spell his name?	normal
5	Double mins and txts 4 6months FREE Bluetooth on Orange. Available on Sony, Nokia Motorola phones.	spam
...

→

CUBE Analytics
CONSULTING

←

keywords

ID	Free	Won	Cash	Type
1	Y	Y	Y	spam
2	N	Y	Y	spam
3	N	N	N	normal
4	N	N	N	normal
5	Y	N	N	spam
6	Y	N	N	spam
7	N	N	N	normal
8	N	Y	N	spam
9	N	N	N	normal
10	N	N	N	normal

Classification example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- ตัวอย่าง spam e-mail classification
 - หา keyword ที่ใช้ปัจบุกว่าเป็น spam e-mail

ID	Text	Type
1	Please call our customer service representative on FREE PHONE 0808 145 4742 between 9am-11pm as you have WON a guaranteed £1000 cash	spam
2	You have won \$1,000 cash or a \$2,000 prize! To claim, call 09050000327	spam
3	I'm gonna be home soon and I don't want to talk about this stuff anymore tonight	normal
4	Is that seriously how you spell his name?	normal
5	Double mins and txts 4 6months FREE Bluetooth on Orange. Available on Sony, Nokia Motorola phones.	spam
...

→

CUBE Analytics
CONSULTING

←

keywords

ID	Free	Won	Cash	Type
1	Y	Y	Y	spam
2	N	Y	Y	spam
3	N	N	N	normal
4	N	N	N	normal
5	Y	N	N	spam
6	Y	N	N	spam
7	N	N	N	normal
8	N	Y	N	spam
9	N	N	N	normal
10	N	N	N	normal

Classification example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- ตัวอย่าง spam e-mail classification
 - หา keyword ที่ใช้ปัจบุกว่าเป็น spam e-mail

ID	Text	Type
1	Please call our customer service representative on FREE PHONE 0808 145 4742 between 9am-11pm as you have WON a guaranteed £1000 cash	spam
2	You have won \$1,000 cash or a \$2,000 prize! To claim, call 09050000327	spam
3	I'm gonna be home soon and I don't want to talk about this stuff anymore tonight	normal
4	Is that seriously how you spell his name?	normal
5	Double mins and txts 4 6months FREE Bluetooth on Orange. Available on Sony, Nokia Motorola phones.	spam
...

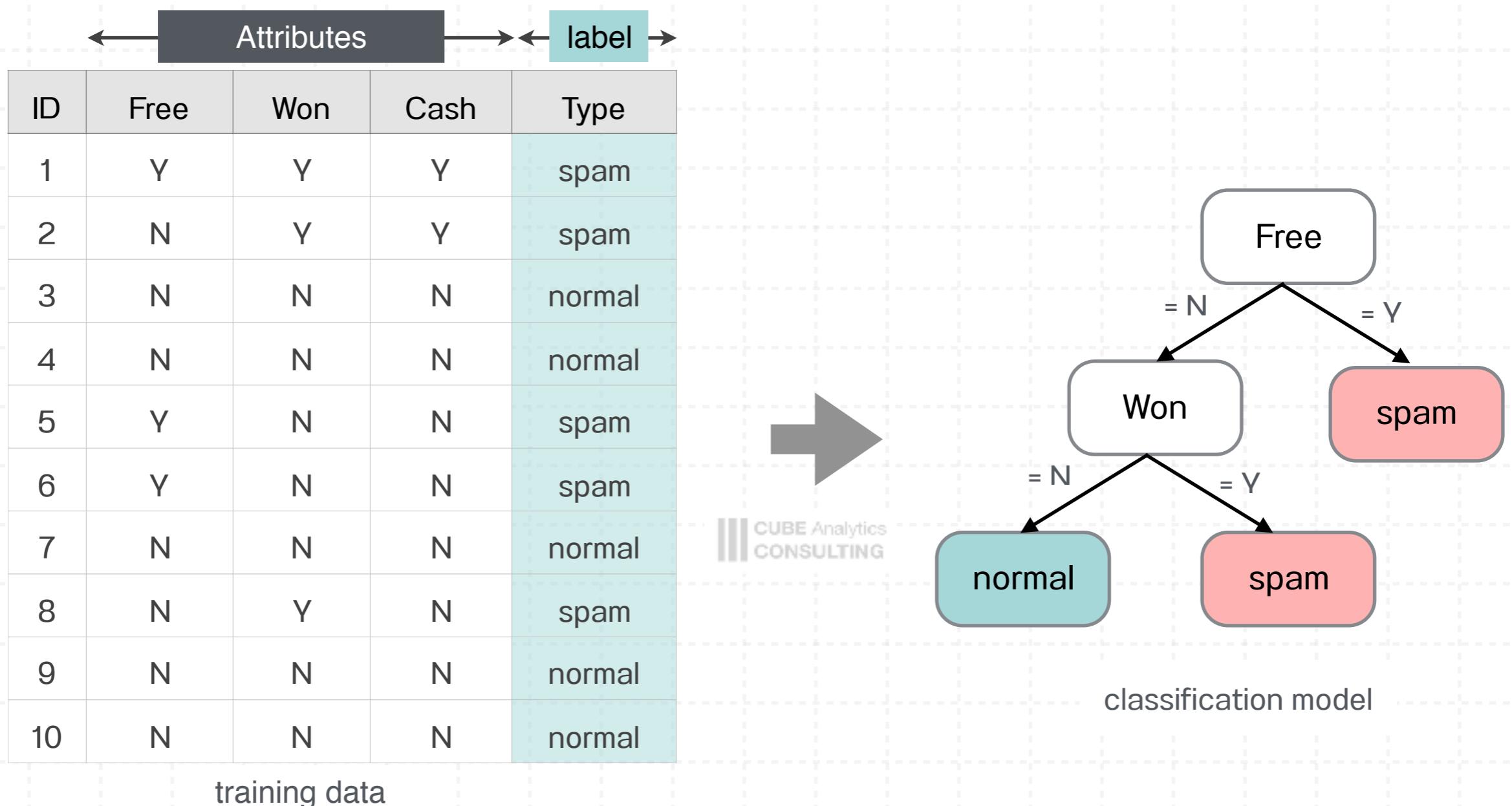
keywords

ID	Free	Won	Cash	Type
1	Y	Y	Y	spam
2	N	Y	Y	spam
3	N	N	N	normal
4	N	N	N	normal
5	Y	N	N	spam
6	Y	N	N	spam
7	N	N	N	normal
8	N	Y	N	spam
9	N	N	N	normal
10	N	N	N	normal

Classification example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

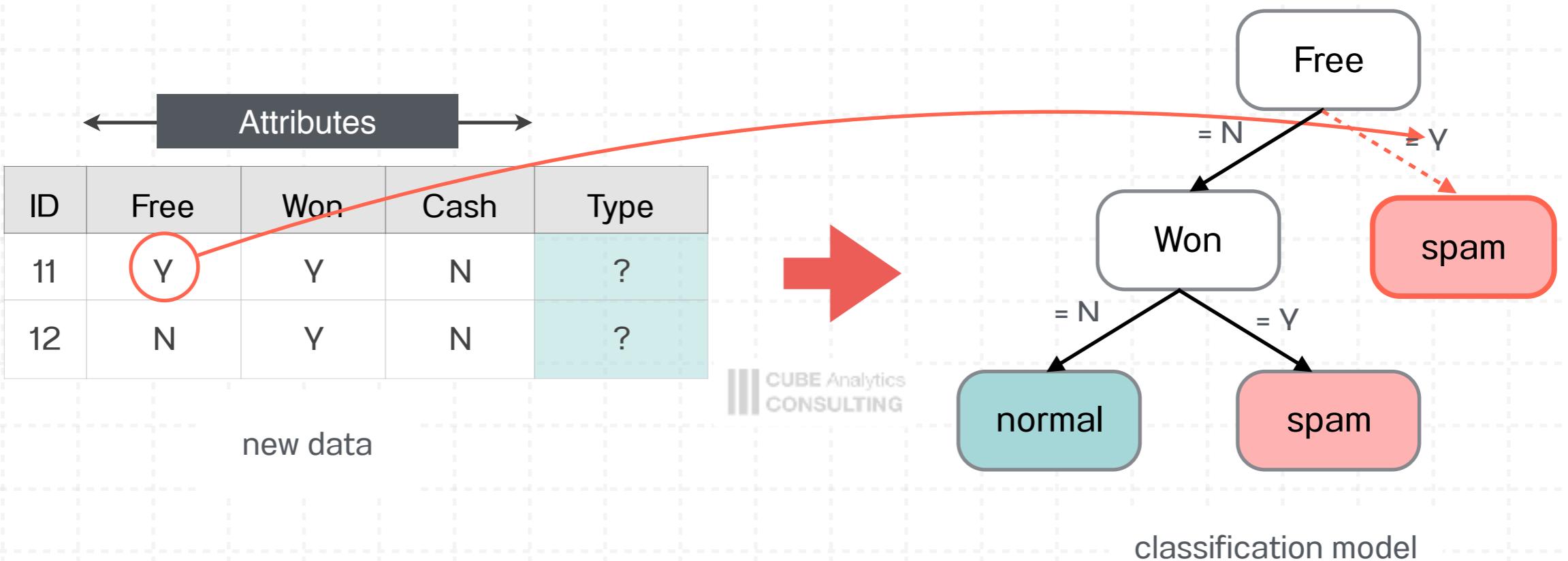
- ตัวอย่าง spam e-mail classification
 - สร้างโมเดล (classification model) จากข้อมูล training data ซึ่งมีลabea (label)



Classification example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

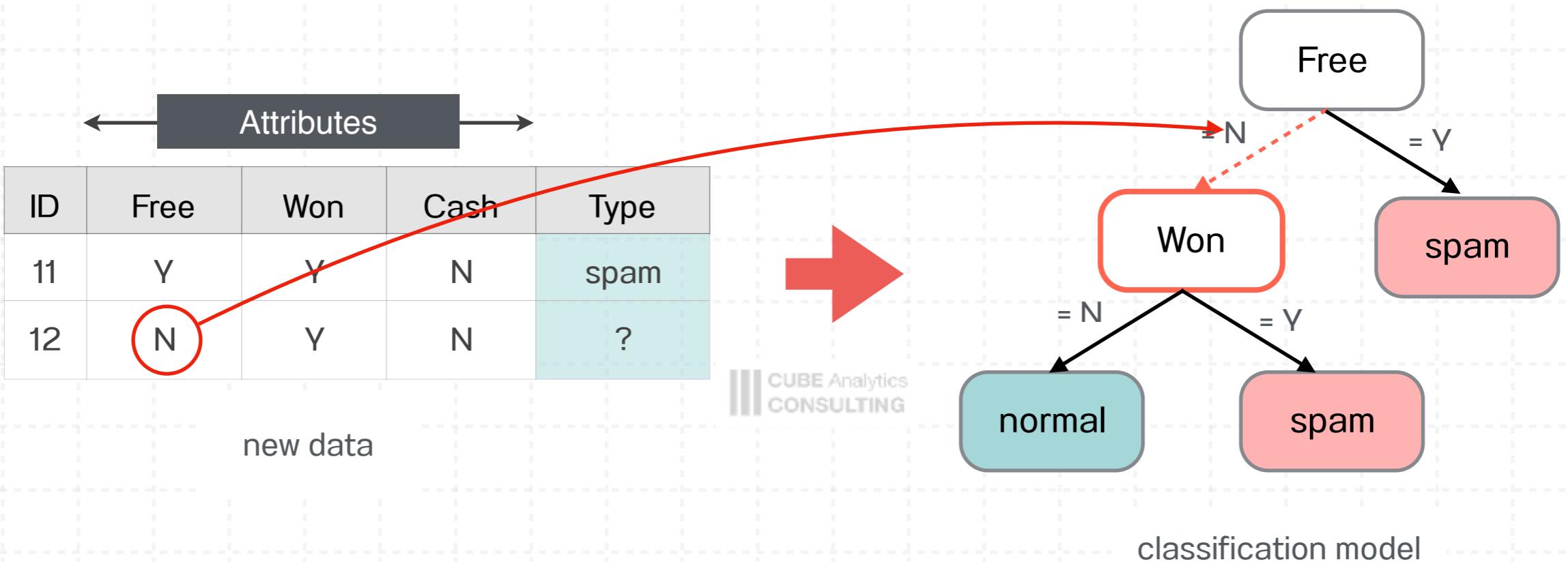
- ตัวอย่าง spam e-mail classification
 - นำข้อมูลใหม่ (unseen data) กำนายโดยใช้โมเดล



Classification example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

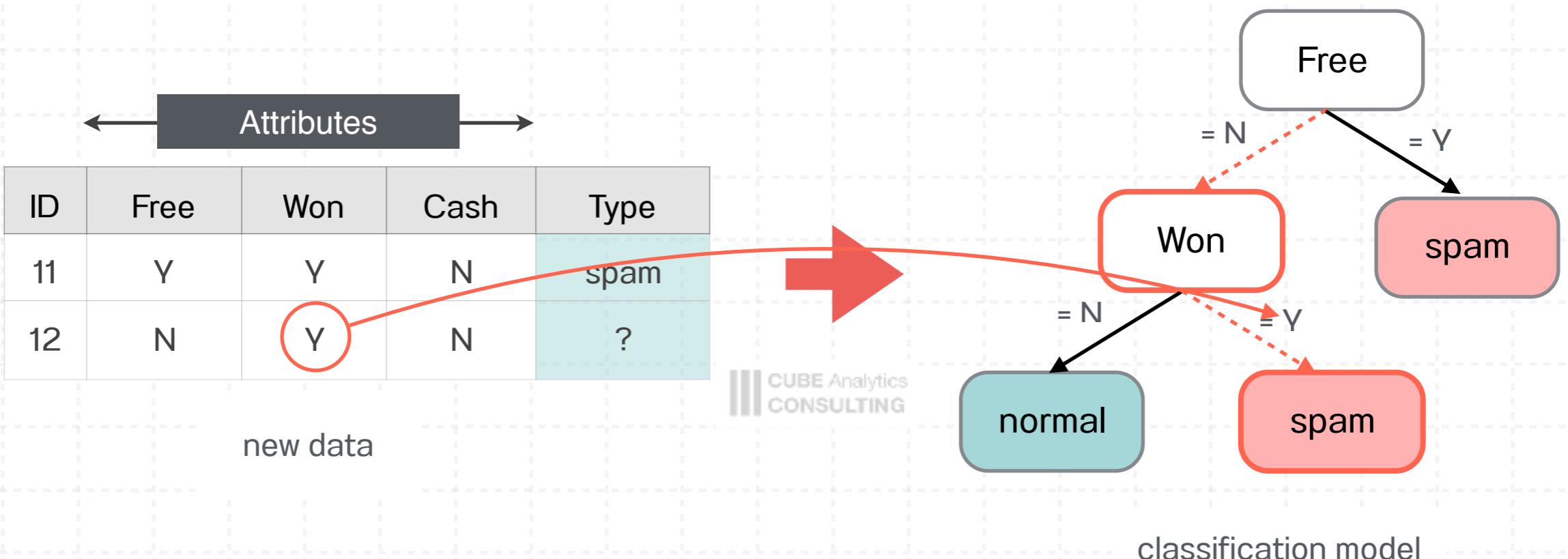
- ตัวอย่าง spam e-mail classification
 - นำข้อมูลใหม่ (unseen data) กำนายโดยใช้โมเดล



Classification example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

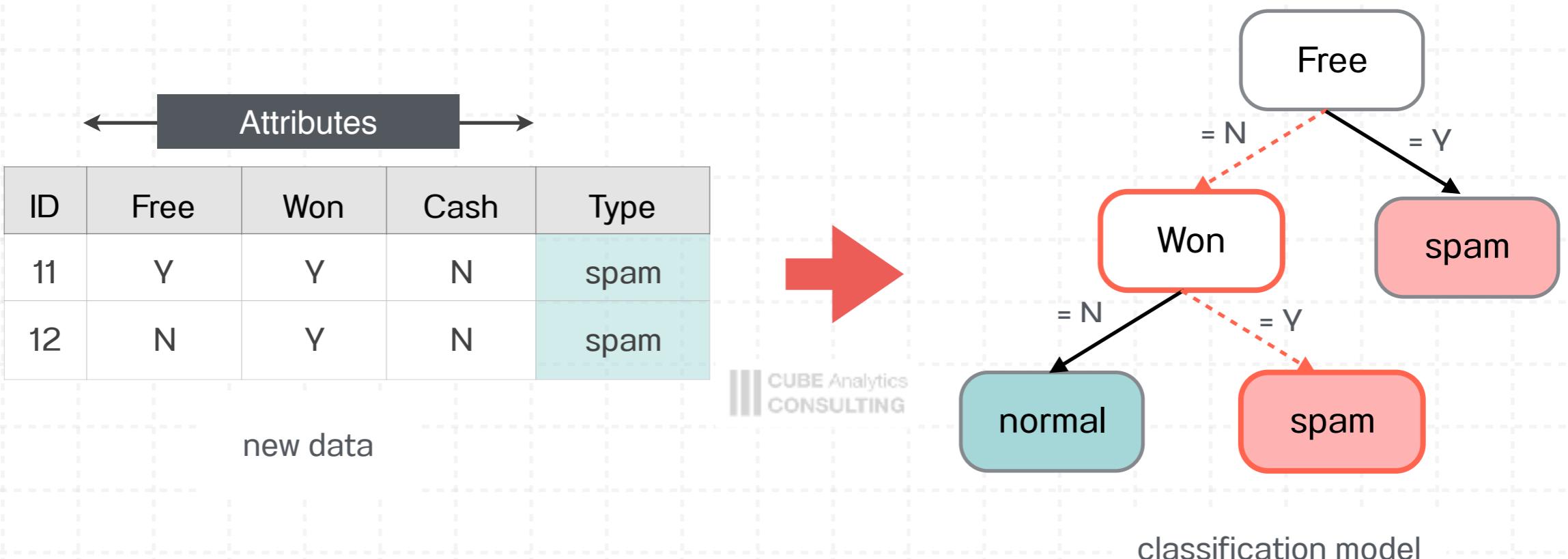
- ตัวอย่าง spam e-mail classification
 - นำข้อมูลใหม่ (unseen data) กำนายโดยใช้โมเดล



Classification example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

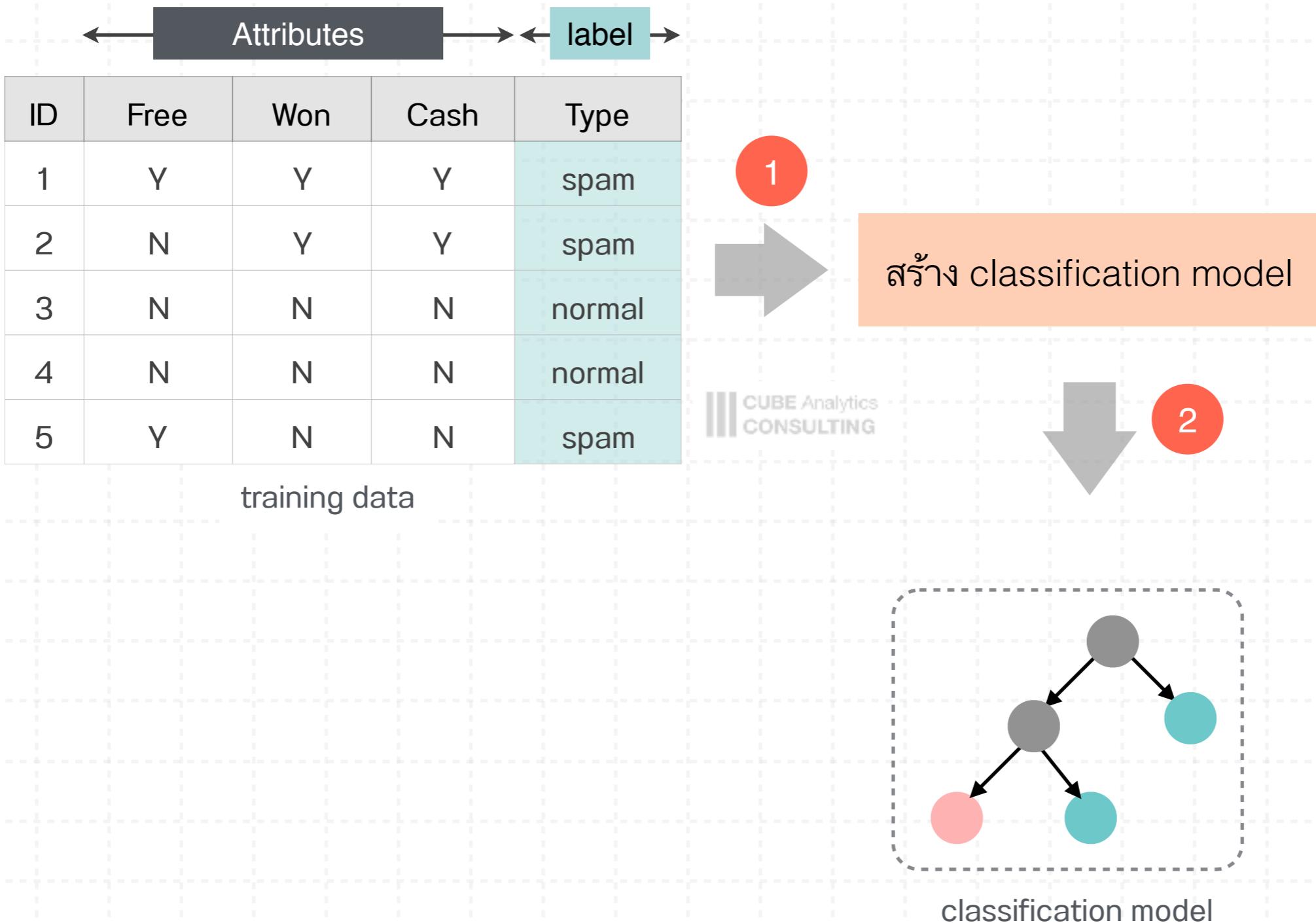
- ตัวอย่าง spam e-mail classification
 - นำข้อมูลใหม่ (unseen data) กำนายโดยใช้โมเดล



Classification example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- ตัวอย่าง spam e-mail classification



Classification example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- ตัวอย่าง spam e-mail classification

← attributes → ← label →

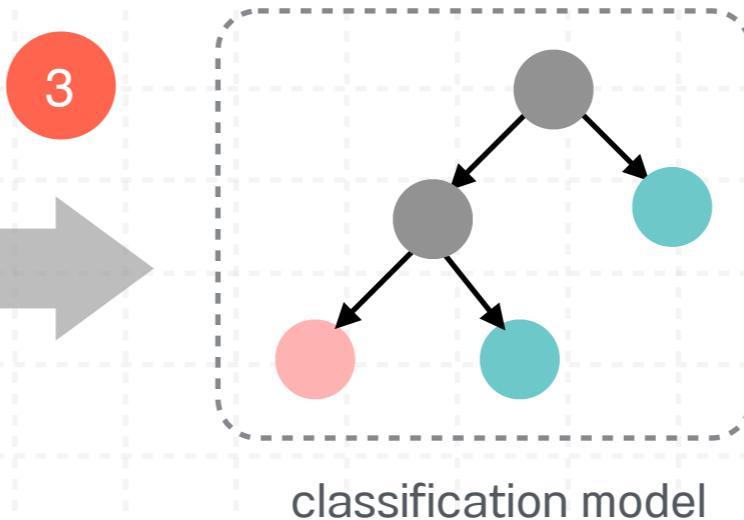
ID	Free	Won	Cash	Type
1	Y	Y	Y	spam
2	N	Y	Y	spam
3	N	N	N	normal
4	N	N	N	normal
5	Y	N	N	spam

training data



ID	Free	Won	Cash	Type
11	Y	Y	N	?
12	N	Y	N	?

new data



Classification example

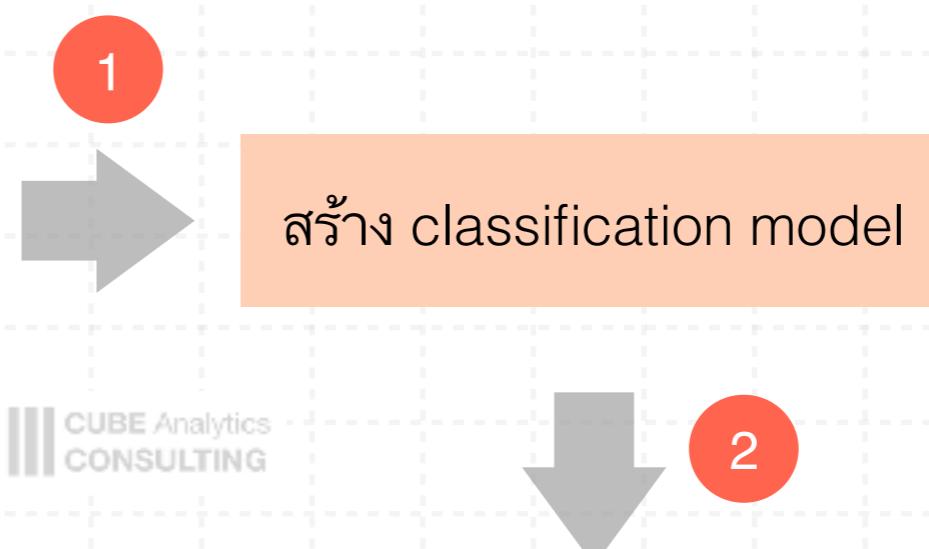
สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- ตัวอย่าง spam e-mail classification

← attributes → ← label →

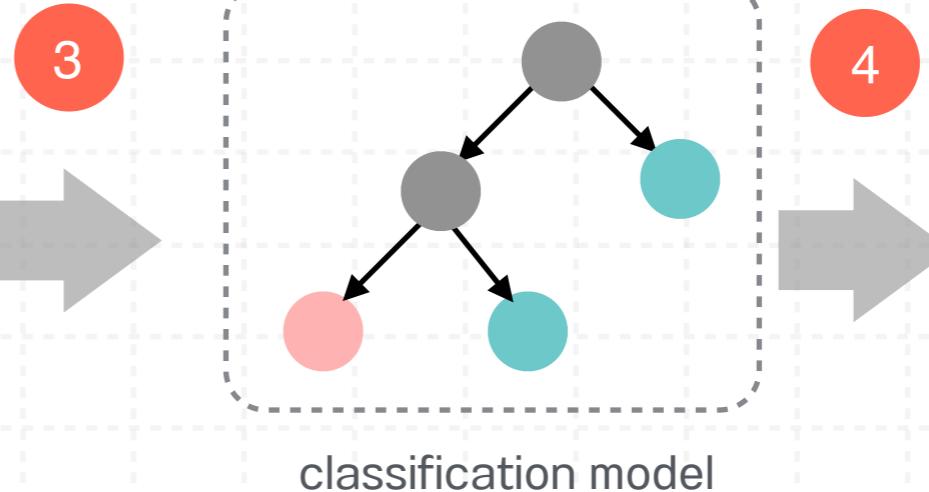
ID	Free	Won	Cash	Type
1	Y	Y	Y	spam
2	N	Y	Y	spam
3	N	N	N	normal
4	N	N	N	normal
5	Y	N	N	spam

training data



ID	Free	Won	Cash	Type
11	Y	Y	N	?
12	N	Y	N	?

new data



ID	Type
11	spam
12	spam

Classification & Regression

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

• การจำแนกประเภทข้อมูล (classification)

- นำข้อมูลเดิมที่มีคำต่อหนึ่ง หรือ คลาส (class) มาสร้างเป็นโมเดล (model) เพื่อหาคำต่อหนึ่งกับข้อมูลใหม่ (unseen data)
- คลาสคำต่อหนึ่งเป็น **ประเภท** (nominal)
 - ผบตอก หรือ ไม่ตอก
 - spam email หรือ normal email

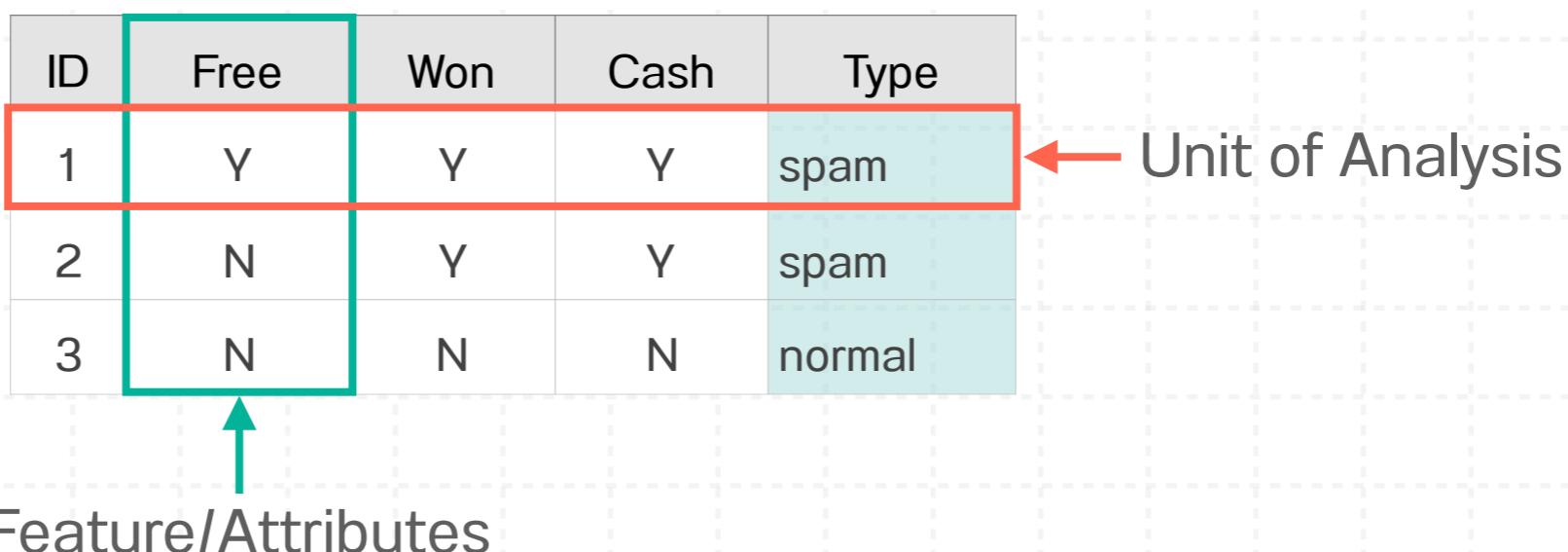
• การประมาณค่าข้อมูล (regression)

- มีลักษณะเหมือนกับ classification เพียงแต่คลาสคำต่อหนึ่งเป็น **ตัวเลข** (numeric)
 - อุณหภูมิในวันถัดไป
 - ยอดขายในไตรมาสถัดไป

Classification example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- **Analytical Base Table (ABT)** คือ **ตารางที่รวมข้อมูล** ไว้เพื่อใช้ในการสร้างโมเดล ซึ่งประกอบด้วย
 - **Unit of Analysis** คือ **ระดับของข้อมูลที่จะทำการวิเคราะห์แสดงด้วย** แถว (row) เช่น e-mail แต่ละฉบับ ลูกค้า (customer) หรือ แทรนแซกชัน (transaction)
 - **Feature/Attribute** คือ **ตัวแปรที่ใช้ในการวิเคราะห์ข้อมูลแสดงด้วย** คอลัมน์ (column) เช่น คำที่เกิดขึ้นใน e-mail ความถี่ในการซื้อสินค้า (frequency) หรือ อายุ (age)



ID	Free	Won	Cash	Type
1	Y	Y	Y	spam
2	N	Y	Y	spam
3	N	N	N	normal

↑
Feature/Attributes

← Unit of Analysis

Split Test

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- แบ่งข้อมูลออกเป็น 2 ชุด
 - training data (ประมาณ 80%) **สำหรับสร้างโมเดล** และ testing data (ประมาณ 20%) **สำหรับทดสอบ**



training data

ID	Free	Won	Cash	Type
1	Y	Y	Y	spam
2	N	Y	Y	spam



สร้าง classification model

CUBE Analytics
CONSULTING



testing data

ID	Free	Won	Cash	Type
3	N	N	N	normal



classification model



2

ID	Type	Pred.
3	normal	normal

4



Performance (classification)

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

ID	Type	Predicted
1	spam	spam
2	spam	spam
3	normal	normal
4	normal	spam
5	spam	spam
6	spam	spam
7	normal	spam
8	spam	normal
9	normal	normal
10	normal	normal
11	spam	spam
12	spam	spam
13	spam	normal
14	spam	normal
15	normal	normal

CUBE Analytics
CONSULTING

pred. \ true	true normal	true spam
pred. normal	TP	FP
pred. spam	FN	TN

Accuracy

- Accuracy

 - จำนวนข้อมูลที่ถูกนายดูออกทุกคลาส

 - True Positive + True Negative**

True Positive + True Negative + False Positive + False Negative

 - $10/15 \times 100 = 66.67\%$

Performance (regression)

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

ID	GPA (undergrad)	GPA (grad)	Predicted value	Error	Error ²
1	3.00	3.25	3.32	0.07	0.0049
2	2.00	2.50	2.54	0.04	0.0016
3	2.50	3.00	2.93	-0.07	0.0049
4	3.50	3.75	3.71	-0.04	0.0016
5	3.75	3.75	3.91	0.16	0.0256
6	2.75	3.25	3.12	-0.13	0.0169
7	2.50	2.75	2.93	0.18	0.0324
8	2.75	3.00	3.13	0.13	0.0169
9	3.00	3.50	3.32	-0.18	0.0324
10	3.10	3.50	3.32	-0.18	0.0324

- ตัววัดประสิทธิภาพของ Regression

- MAE (Mean Absolute Error)

$$\frac{\sum | \text{predicted value} - \text{actual value} |}{\text{number of examples}}$$

- MAE = $\frac{|0.07| + |0.04| + |-0.07| + \dots}{10}$
= 0.118

- RMSE (Root Mean Square Error)

$$\sqrt{\frac{\sum (\text{predicted value} - \text{actual value})^2}{\text{number of examples}}}$$

- RMSE = $\sqrt{(0.0049+0.0016+\dots)/10}$
= 0.13

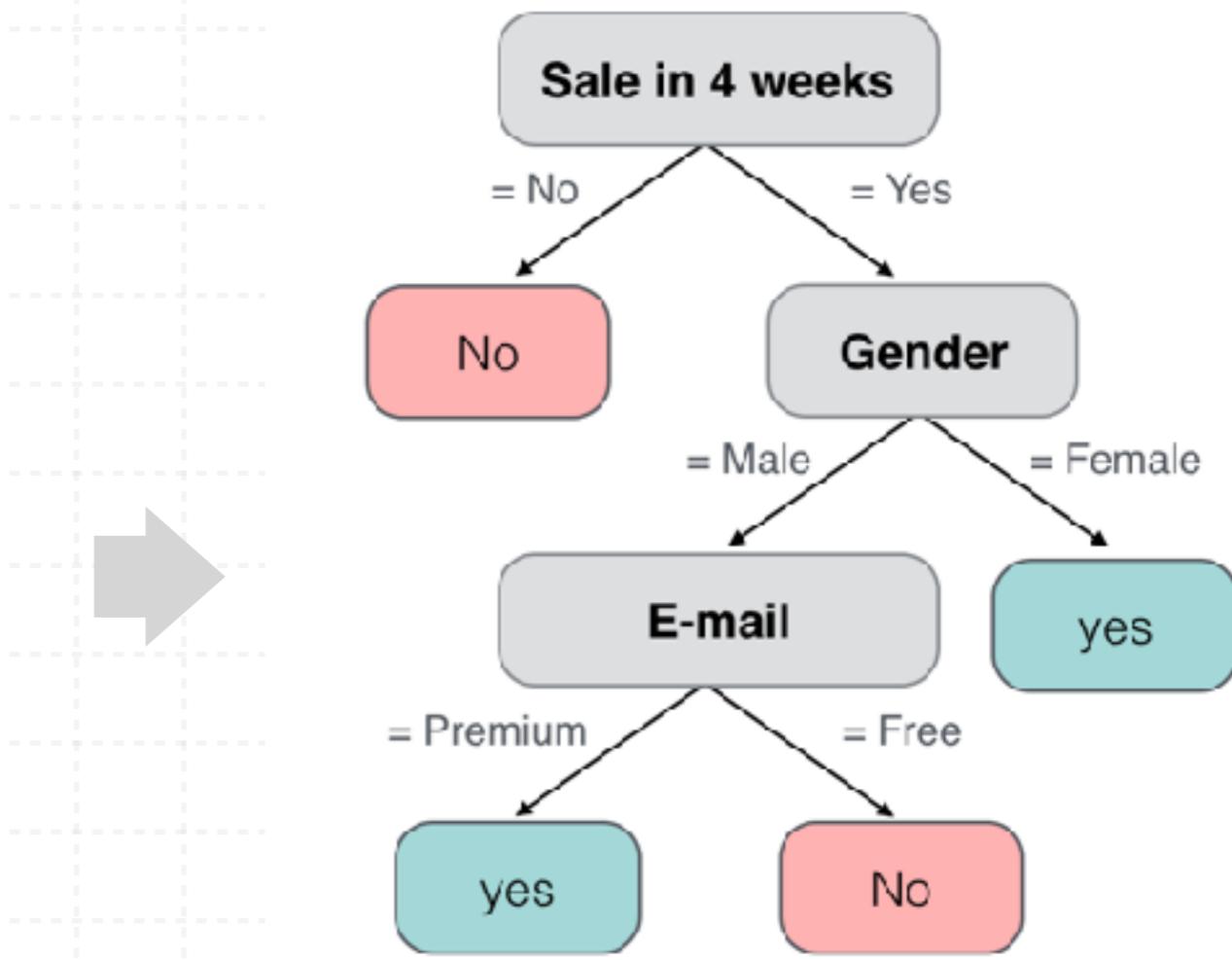
Decision Tree

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- ตัวอย่างการสร้างไมเต็ล Decision Tree

ID	Gender	E-mail	Sale in 4 Weeks	Response
1	Female	Premium	Yes	Yes
2	Male	Premium	Yes	Yes
3	Female	Free	No	No
4	Female	Premium	Yes	Yes
5	Female	Premium	Yes	Yes
6	Male	Premium	No	No
7	Female	Free	Yes	Yes
8	Male	Free	Yes	No
9	Female	Premium	No	No
10	Male	Free	Yes	No

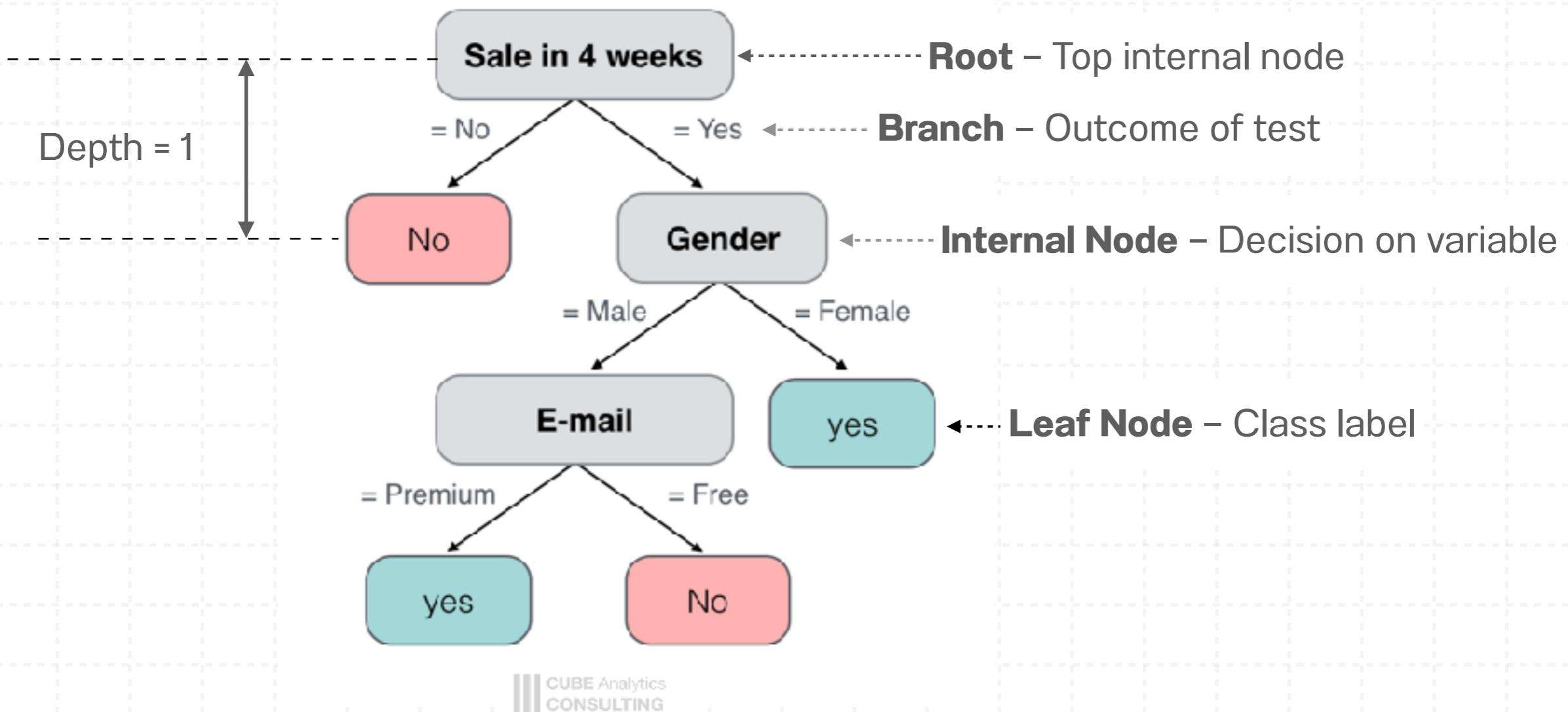
Training data



Decision Tree

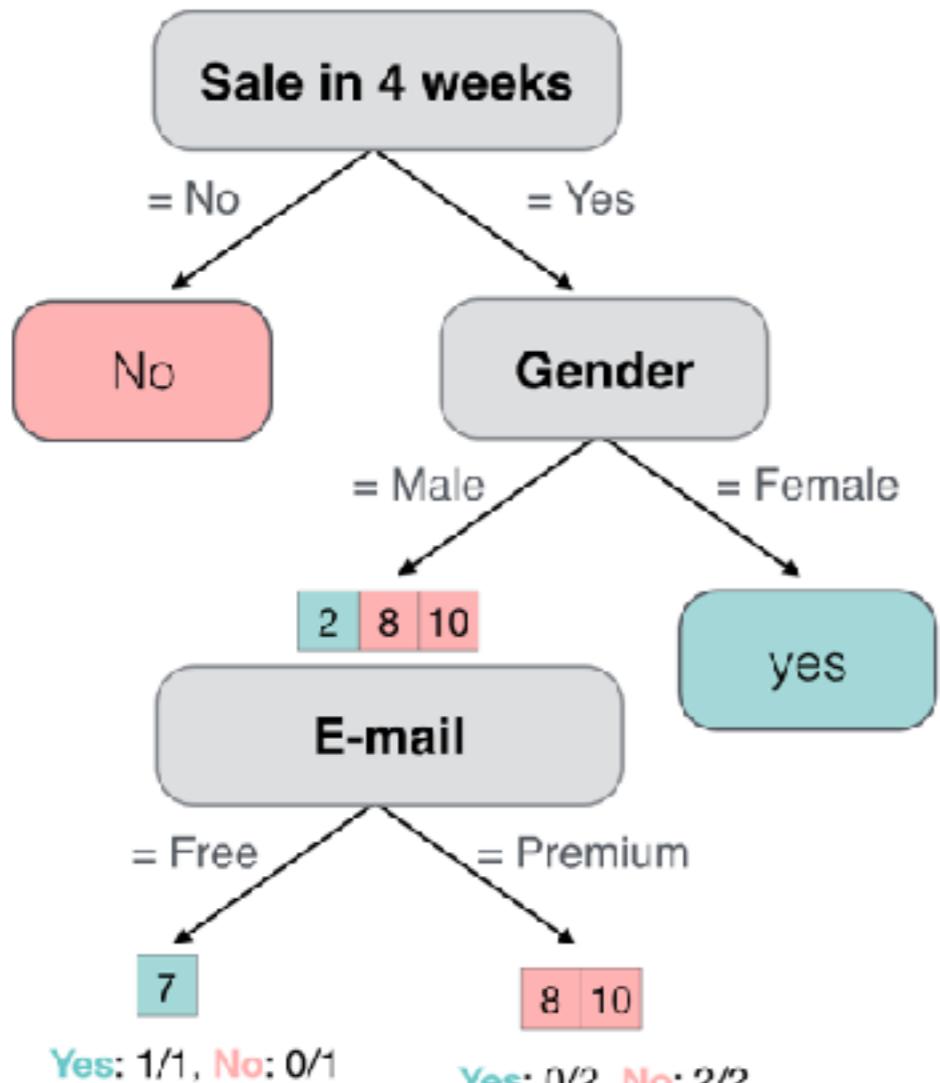
สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- องค์ประกอบของ Decision Tree



Decision Tree

สำหรับการบรรยายในงาน Dev Mountain Tech Festival



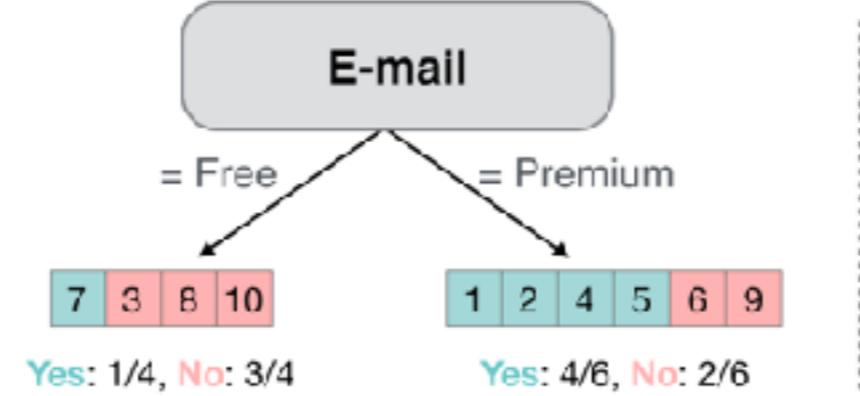
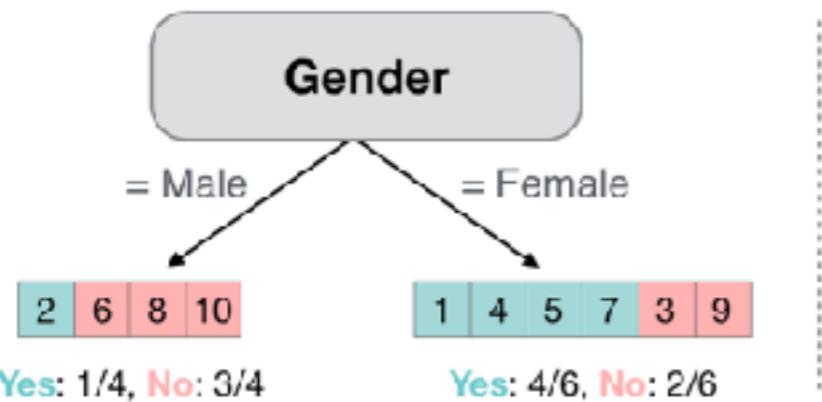
แนวคิดของการสร้างไมเดล Decision Tree

1. ไมเดล Decision Tree ต้องการจะ**แยกข้อมูลออกจากกัน**ให้ได้มากที่สุดเท่าที่จะทำได้
2. เลือกตัวแปร (variable) ที่**แบ่งแยกคลาสคำตอบได้ดีที่สุด**นิauważว่าเป็นไปบดແรက
3. หลังจากนั้นก็หาตัวแปรอื่นๆ มา**แบ่งข้อมูลในลำดับชั้นถัดไป**
4. กำช้ำจับข้อมูลแยกออกจากกัน

Decision Tree

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

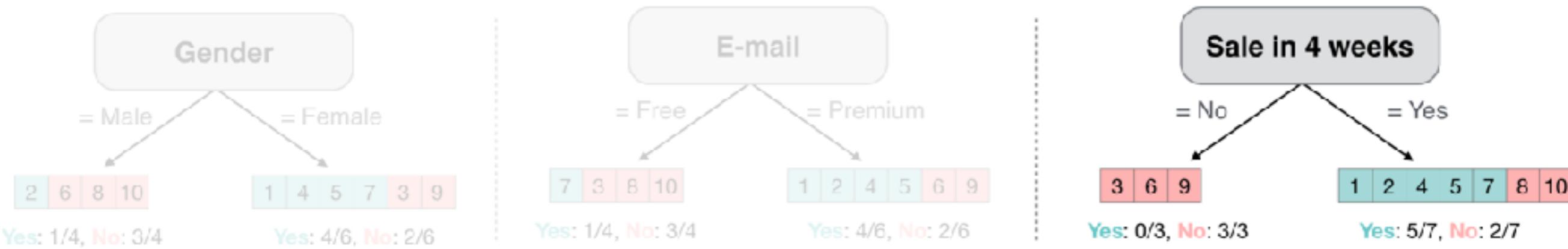
- เลือกแอดຕกรีบิวต์ที่สามารถแบ่งแยกคลาสคำตอบได้ดีที่สุดมาก่อน



Decision Tree

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

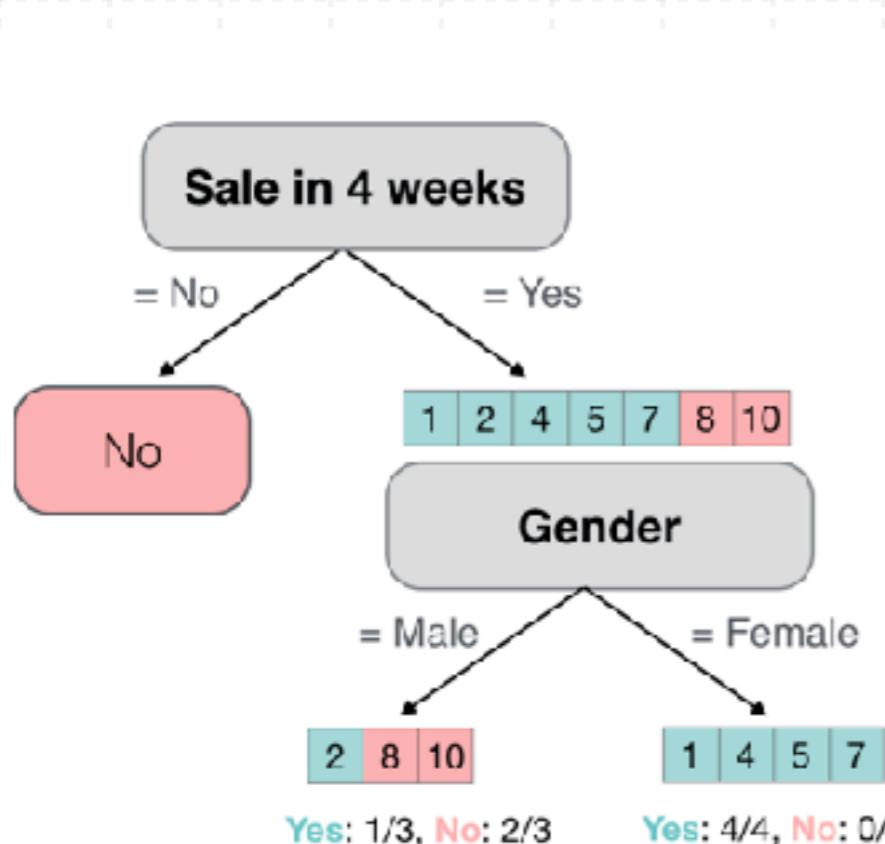
- เลือกແອຕກຣິບວ່າ Sale in 4 weeks ເນື່ອງຈາກດ້າມີຄໍາເປັນ No ຈະຕອບໄວ່ວ່າ responses = No



Decision Tree

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

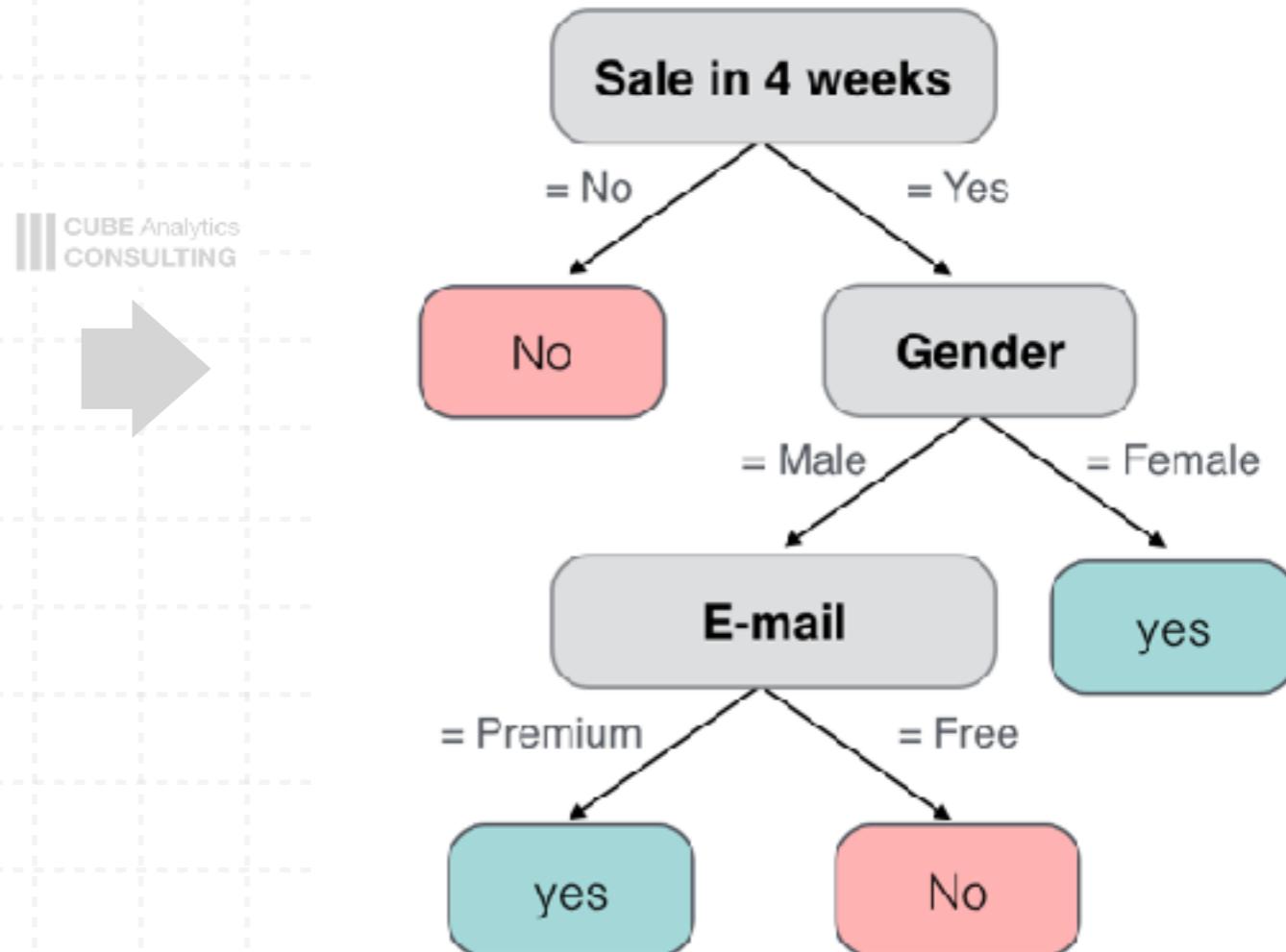
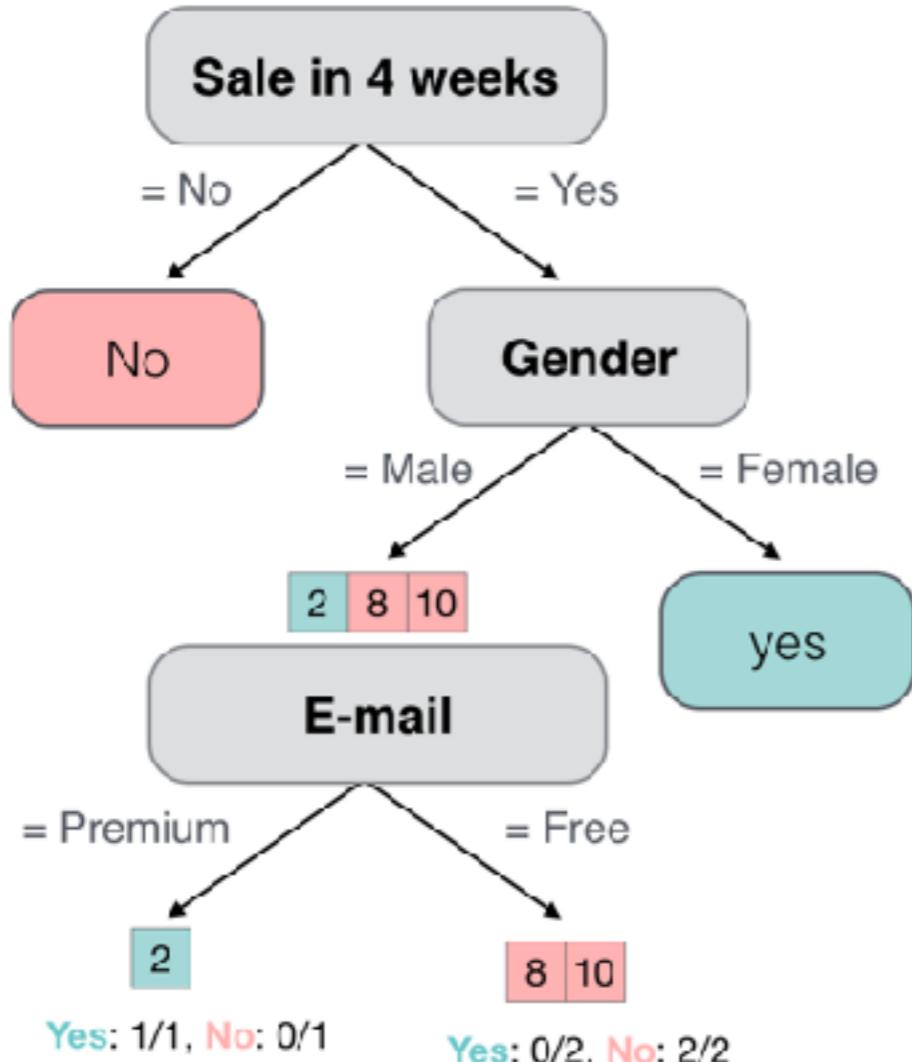
- หาແອຕກຣິບວ່າອື່ນໆ ມາແປ່ງຂ້ອມູລເນື່ອແອຕກຣິບວ່າ **Sale in 4 weeks** ໂປ່ງ Yes ຕ່ອໄປ



Decision Tree

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

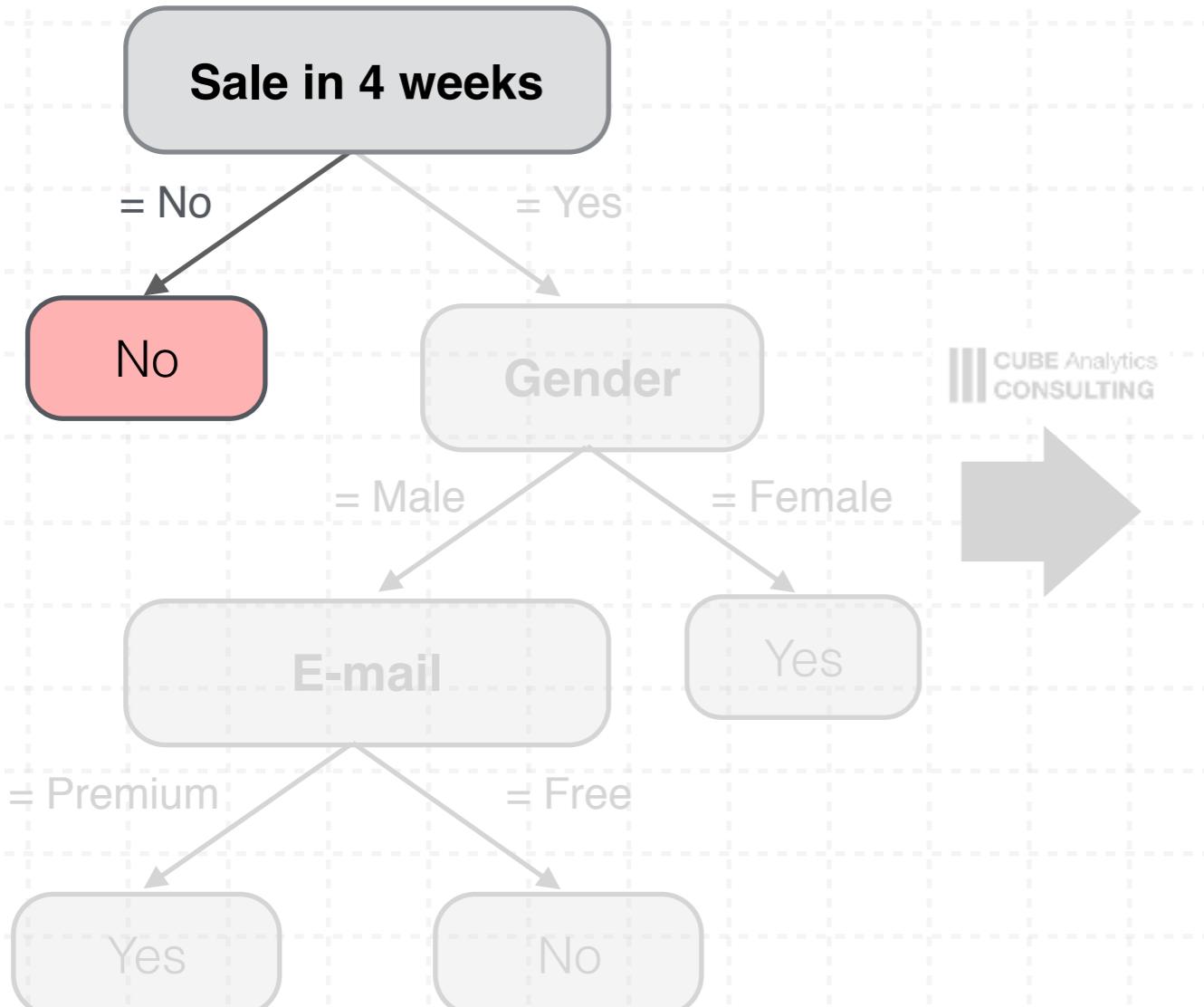
- หากแอ็ตกริบอตอื่นๆ มาแบ่งข้อมูลเมื่อแอ็ตกริบอต **Gender** เป็น Male ต่อไป



Decision Tree

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- แปลงโมเดล Decision Tree ให้กลายเป็นกฎ (rule)

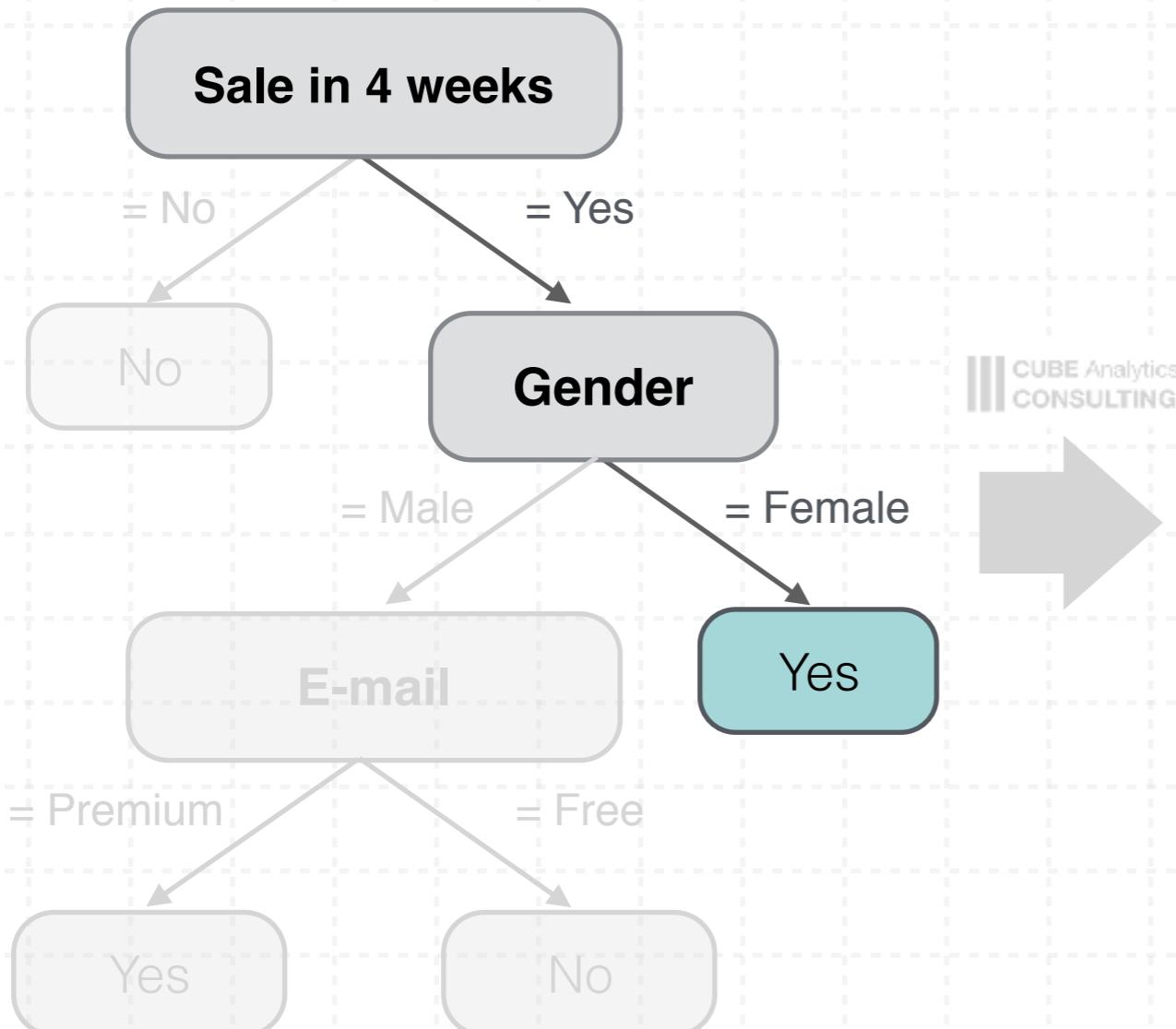


- IF Sale in 4 weeks = No THEN Response = No**

Decision Tree

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- แปลงโมเดล Decision Tree ให้กลายเป็นกฎ (rule)

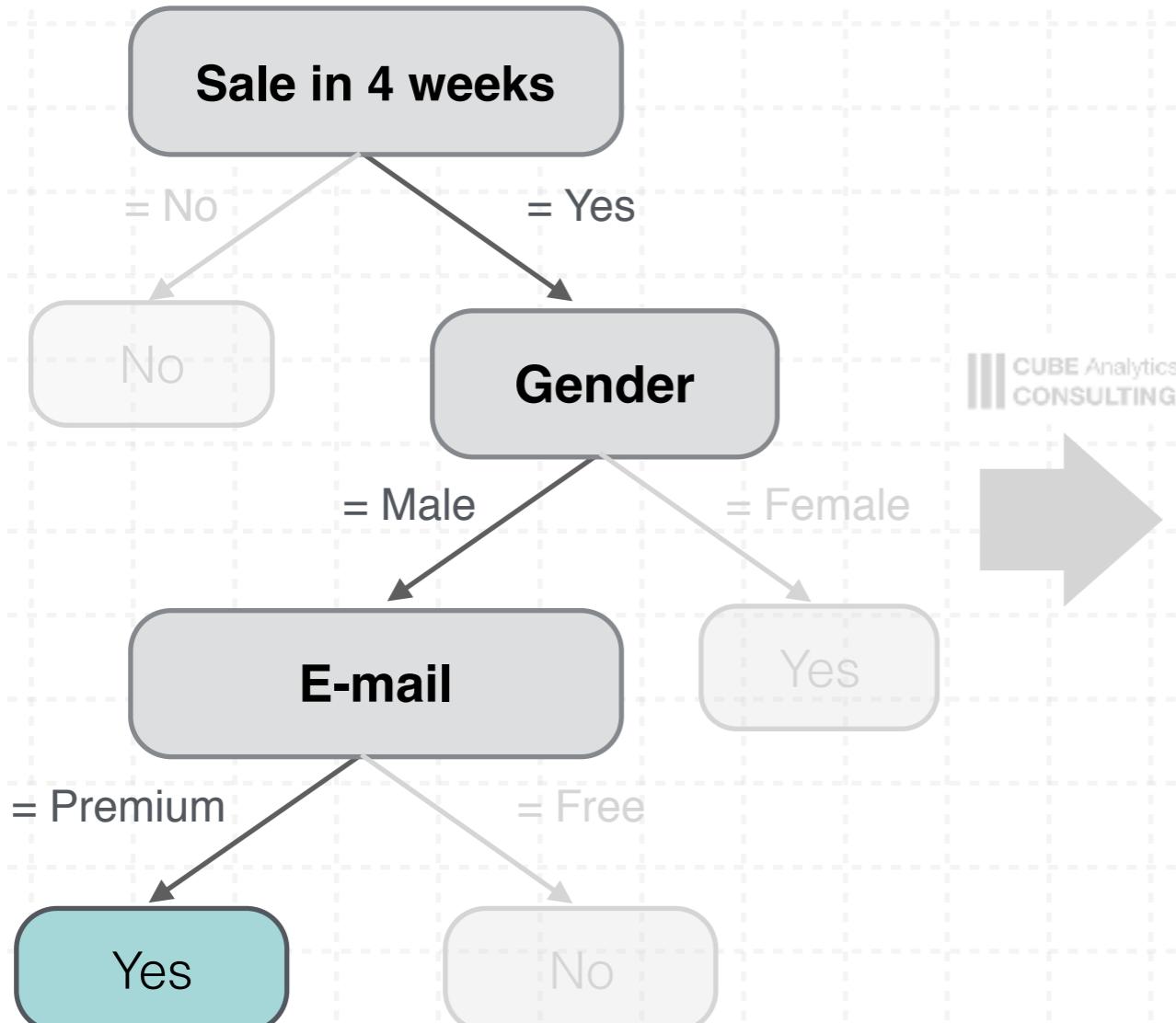


- IF Sale in 4 weeks = No THEN
Response = No
- IF Sale in 4 weeks = Yes
AND Gender = Female
THEN Response = Yes

Decision Tree

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- แปลงโมเดล Decision Tree ให้กลายเป็นกฎ (rule)

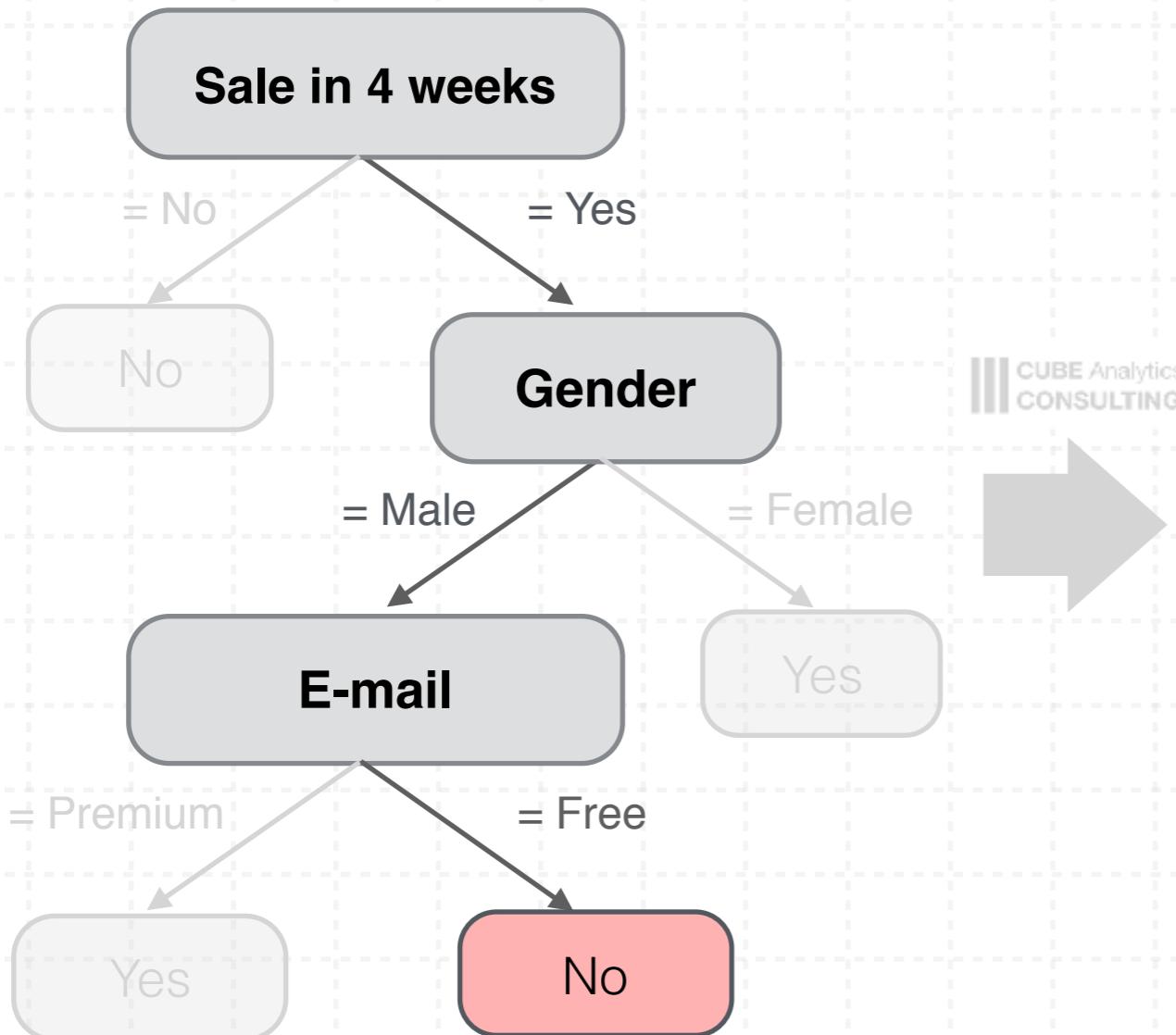


- IF Sale in 4 weeks = No THEN Response = No
- IF Sale in 4 weeks = Yes AND Gender = Female THEN Response = Yes
- IF Sale in 4 weeks = Yes AND Gender = Male AND E-mail = Premium THEN Response = Yes

Decision Tree

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- แปลงโมเดล Decision Tree ให้กลายเป็นกฎ (rule)

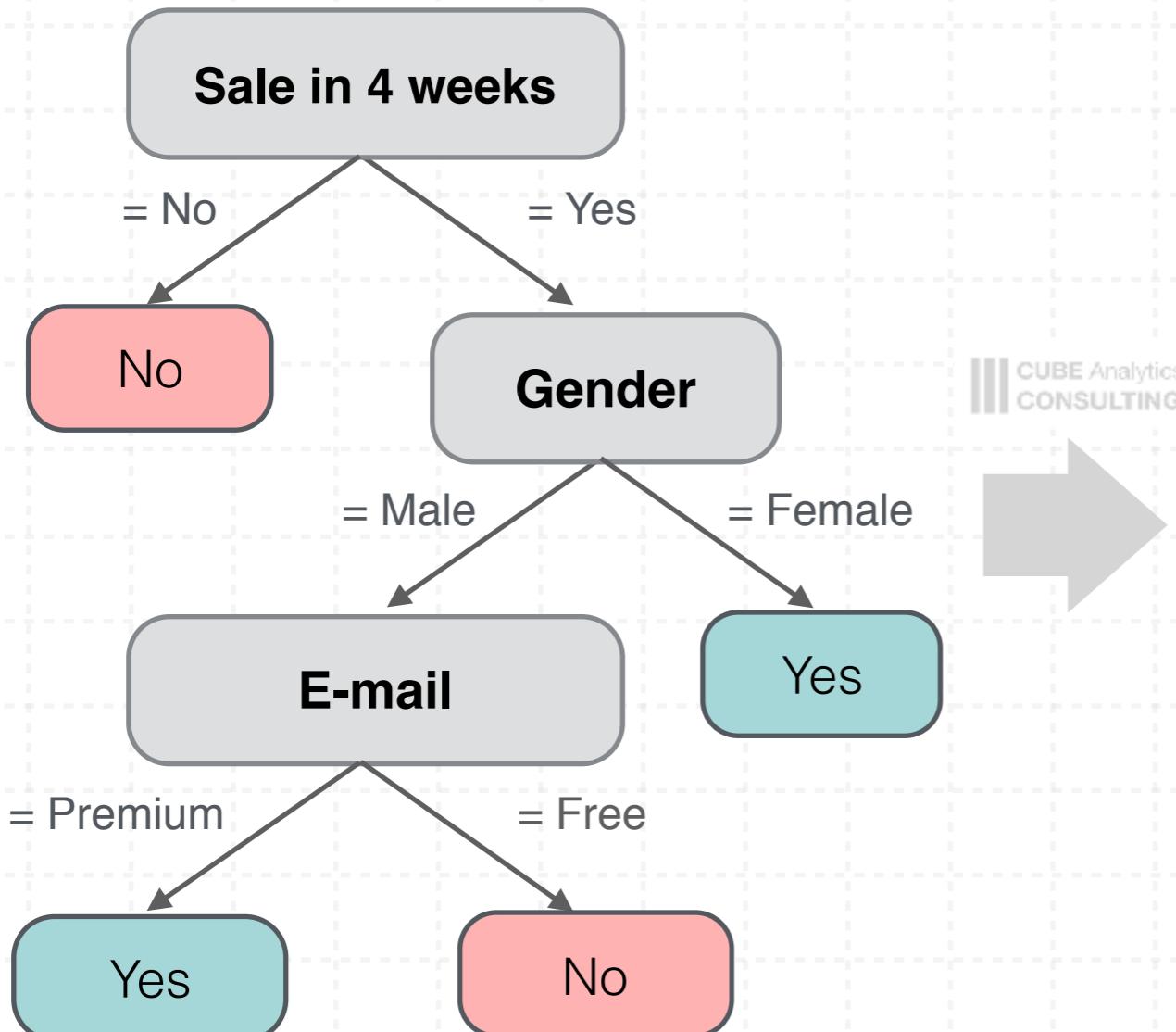


- IF Sale in 4 weeks = No THEN
Response = No
- IF Sale in 4 weeks = Yes
AND Gender = Female
THEN Response = Yes
- IF Sale in 4 weeks = Yes AND
Gender = Male AND E-mail = Premium
THEN Response = Yes
- IF Sale in 4 weeks = Yes AND
Gender = Male AND E-mail = Free
THEN Response = No

Decision Tree

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- แปลงโมเดล Decision Tree ให้กลายเป็นกฎ (rule)



- **IF Sale in 4 weeks = No THEN Response = No**
- **IF Sale in 4 weeks = Yes AND Gender = Female THEN Response = Yes**
- **IF Sale in 4 weeks = Yes AND Gender = Male AND E-mail = Premium THEN Response = Yes**
- **IF Sale in 4 weeks = Yes AND Gender = Male AND E-mail = Free THEN Response = No**

Decision Tree

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

• ข้อดี

- เป็นโมเดลที่ **แปลความได้ง่าย** โดยการอ่านໄล่จากบนลงล่าง
- ใช้ได้กับตัวแปรหรือแอตกริบิวต์ที่เป็นตัวเลข (numeric) และ category
- **เลือกตัวแปรหรือแอตกริบิวต์ที่สำคัญ** และตัดตัวแปรหรือแอตกริบิวต์ที่ไม่สำคัญ
- ปรับเปลี่ยนพารามิเตอร์ของ Decision Tree ได้ง่าย
- พยานกรณีคำตอบ (class) ที่เป็นประเภท category ได้มากกว่า 2 ค่า

• ข้อจำกัด

- ไม่เหมาะสมกับข้อมูลที่มีความซับซ้อน (complex) มาตรฐาน
- ข้อมูล[**แอตกริบิวต์หรือตัวแปรที่เป็นตัวเลขจะแบ่งได้เป็น 2 ช่วง**](#)หรือต้องแปลงให้เป็น category มาตรฐาน

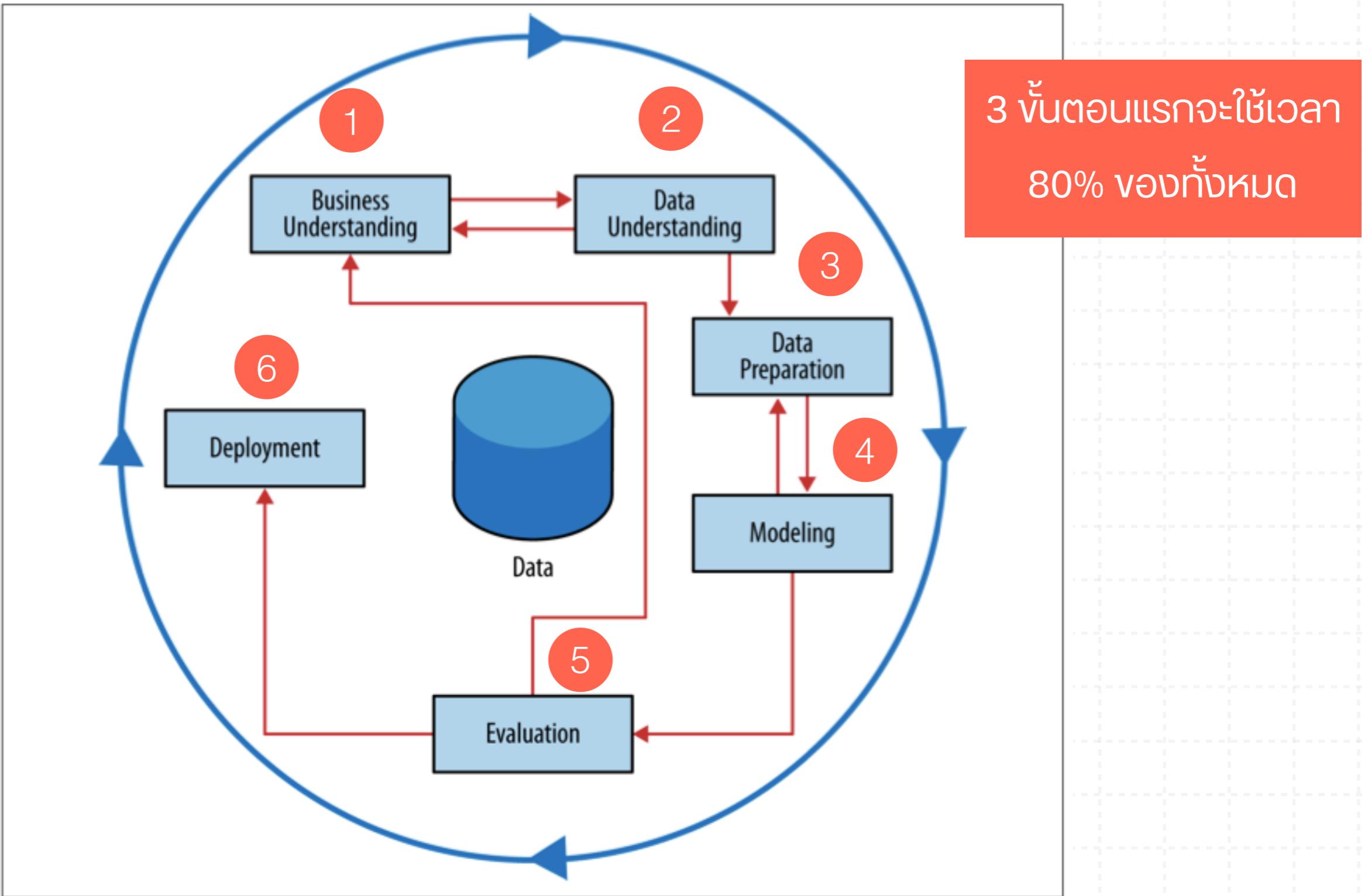
Outlines

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- Machine Learning Applications
- Data Management
- Data Roles
- Data Analytics
- Machine Learning
- **Machine Learning Development Steps**
- No-Code/Low-Code Machine Learning Platform

CRISP-DM

สำหรับการบรรยายในงาน Dev Mountain Tech Festival



CRISP-DM

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

STEP	DESCRIPTION
Business Understanding	Define the project.
Data Understanding	Examine the data; identify problems in the data.
Data Preparation	Fix problems in the data; create derived variables.
Modeling	Build predictive or descriptive models.
Evaluation	Assess models; report on the expected effects of
Deployment	Plan for use of models.

Machine Learning Use-cases

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

Use-case	คำอธิบาย
Customer Lifetime Value (CLV)	การคำนวณมูลค่าของลูกค้าแต่ละราย
Customer Segmentation	แบ่งกลุ่มลูกค้าตามพฤติกรรมต่างๆเพื่อจะได้เข้าใจลูกค้าได้มากขึ้น
Up- and Cross-Selling	แนะนำสินค้า/บริการเพื่อให้ลูกค้าซื้อมากขึ้นหรือแพงขึ้น
Next Best Action	นำเสนอสินค้า/บริการให้ตรงใจและเหมาะสมกับช่วงเวลาที่ต้องการ
Product Propensity	คาดการณ์ว่าลูกค้าคนใดบ้างที่มีโอกาสซื้อสินค้า/บริการ
Churn Prevention	คาดการณ์ว่าลูกค้าคนใดบ้างที่มีโอกาสจะยกเลิกการใช้บริการ
Fraud Detection	ค้นหาและคาดการณ์การใช้งานที่ผิดปกติ
Demand Forecasting	พยากรณ์ยอดขายหรือยอดผลิตสินค้าต่างๆ
Quality Assurance	ค้นหาปัญหาที่เกี่ยวข้องกับคุณภาพของสินค้าก่อนการจำหน่าย

Machine Learning Use-cases

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

Use-case	คำอธิบาย
Risk Management	การประเมินความเสี่ยงต่างๆ รวมถึงการให้สินเชื่อ
Price Optimization	ปรับเปลี่ยนราคาของสินค้า/บริการให้เหมาะสมกับความต้องการ
Predictive Maintenance	คาดการณ์ความผิดปกติของเครื่องจักรก่อนปัญหาจะเกิดขึ้น
Social Network Analysis	วิเคราะห์ข้อมูลจากเครือข่ายสังคมออนไลน์ (online)

Machine Learning Use-cases

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

Use-cases	Retail	Banking	Telecommunication	Insurance	Real Estate
Customer Lifetime Value (CLV)	●	●	●	●	●
Customer Segmentation	●	●	●	●	●
Up- and Cross-Selling	●	●	●	●	●
Next Best Action	●	●	●	●	●
Product Propensity	●	●	●	●	●
Churn Prevention	●	●	●	●	
Fraud Detection		●		●	
Demand Forecasting	●		●		●
Quality Assurance					●
Risk Management		●		●	●
Price Optimization	●				●
Predictive Maintenance					●

CRISP-DM Example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

1. Business Understanding

- บริษัทบัตรเครดิตแห่งหนึ่งให้บริการสินเชื่อซึ่งมีการเก็บข้อมูลว่าลูกค้าชำระเงินเท่ามาเมื่อวันที่เท่าไร
- บริษัทต้องการแบ่งกลุ่มลูกค้าตามพฤติกรรมการชำระเงิน เพื่อจะได้วางแผนได้ว่าลูกค้ากลุ่มไหนควรจะต้องติดตามอย่างไรบ้าง

2. Data Understanding

- ข้อมูลพฤติกรรมการชำระเงินแต่ละงวดของลูกค้าแต่ละคน

CRISP-DM Example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

2. Data Understanding

- ข้อมูลพื้นฐานของการชำระเงินแต่ละงวดของลูกค้าแต่ละคน

- Customer_ID** หมายเลขของลูกค้า

- Due_Date** วันที่กำหนดชำระเงิน

- Payment_Date** วันที่ลูกค้าชำระเงิน

- Amount** ยอดเงินที่ชำระ

Customer_ID ↑	Due_Date	Payment_Date	Amount
180001	Feb 28, 2018	Feb 26, 2018	1500
180001	Mar 28, 2018	Mar 27, 2018	1500
180001	Apr 28, 2018	Apr 28, 2018	1500
180001	May 28, 2018	May 25, 2018	1500
180001	Jun 28, 2018	Jun 25, 2018	1500
180001	Jul 28, 2018	Jul 23, 2018	1500
180001	Aug 28, 2018	Aug 29, 2018	1500
180001	Sep 28, 2018	Sep 28, 2018	1500
180001	Oct 28, 2018	Oct 28, 2018	1500
180001	Nov 28, 2018	Nov 23, 2018	1500
180002	Mar 31, 2018	Apr 1, 2018	3333.333
180002	Apr 30, 2018	Apr 29, 2018	3333.333
180002	May 31, 2018	May 26, 2018	3333.333
180002	Jun 30, 2018	Jul 1, 2018	3333.333
180002	Jul 31, 2018	Jul 27, 2018	3333.333
180002	Aug 31, 2018	Sep 1, 2018	3333.333

CRISP-DM Example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

3. Data Preparation

- สร้างตัวแปรใหม่จากข้อมูลประวัติการชำระเงิน
 - Percent การชำระเงินก่อนกำหนด, ตรงเวลา และ ล่าช้า
 - จำนวนวันน้อยสุด (min) ที่ชำระเงินก่อนกำหนด ตรงเวลา และล่าช้า
 - จำนวนวันมากสุด (max) ที่ชำระเงินก่อนกำหนด ตรงเวลา และล่าช้า
 - จำนวนวันเฉลี่ย (average) ที่ชำระเงินก่อนกำหนด ตรงเวลา และล่าช้า

4. Modeling

- Behavioral Segmentation
 - ใช้การแบ่งกลุ่มข้อมูล (clustering) ด้วยเทคนิค K-Means

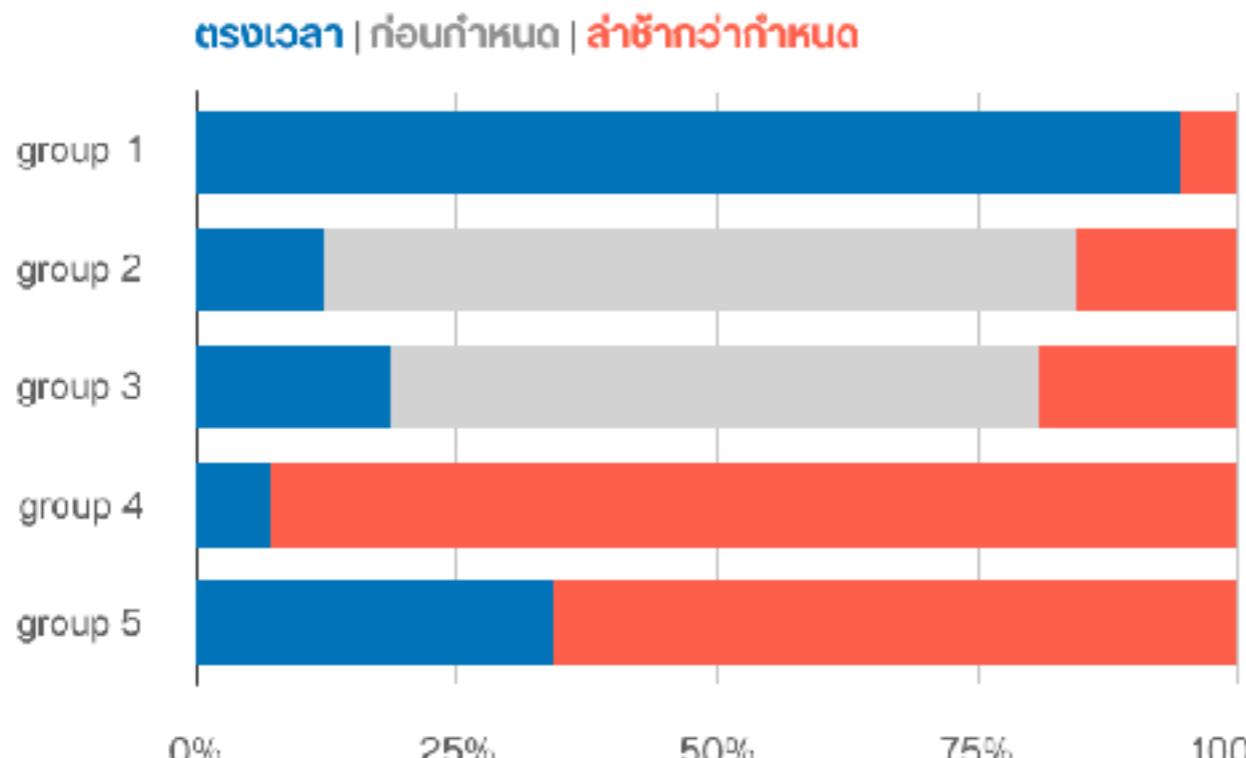
CRISP-DM Example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

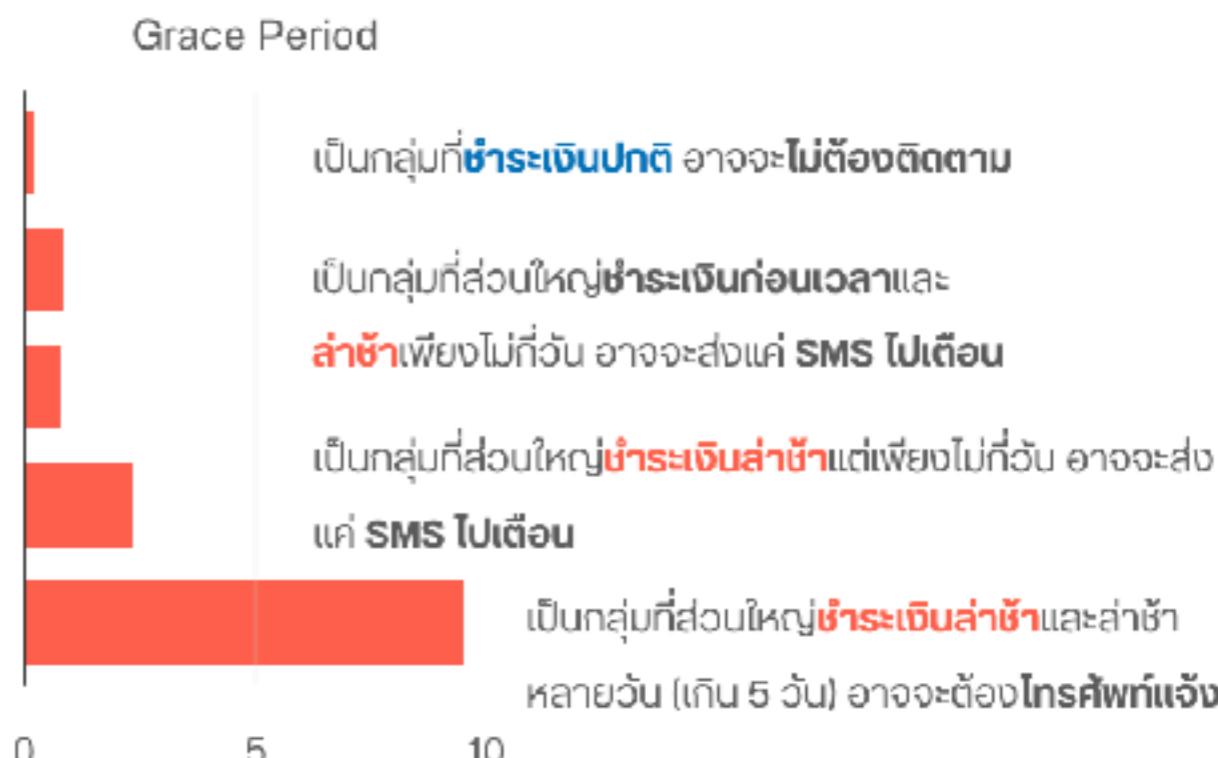
5. Evaluation

- แบ่งกลุ่มออกเป็น 5 กลุ่ม

สัดส่วนการชำระเงินลูกค้าแต่ละกลุ่ม



จำนวนวันที่ชำระเงินล่าช้ามากที่สุด



CRISP-DM Example

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

6. Deployment

จำนวนลูกค้ากั้งหนด

999

ยอดเงินกั้งหนด

21.7M

ยอดเงินเฉลี่ยต่อคน

22K

ยอดเงินน้อยสุดต่อคน

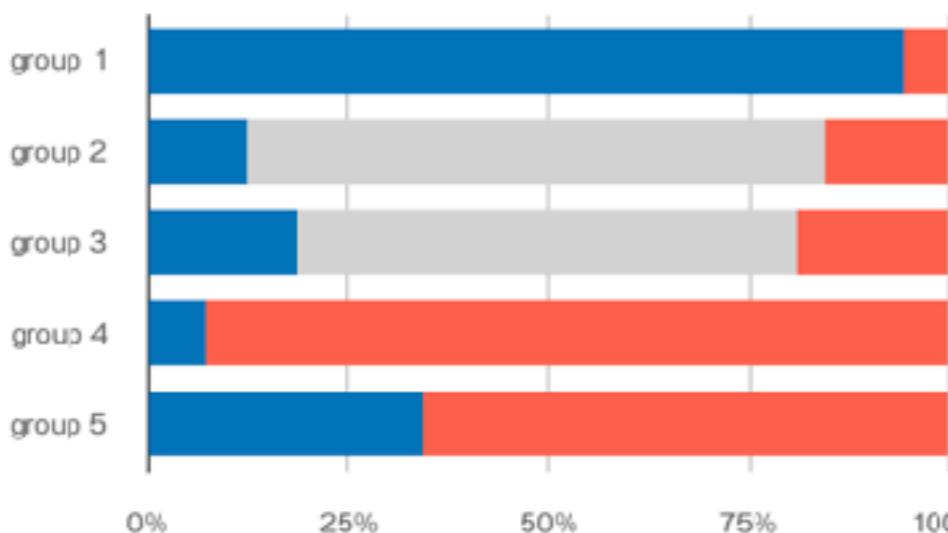
3K

ยอดเงินมากสุดต่อคน

100K

สัดส่วนการชำระเงินลูกค้าแต่ละกลุ่ม

ตรงเวลา | ก่อนกำหนด | **ล่าช้ากว่ากำหนด**



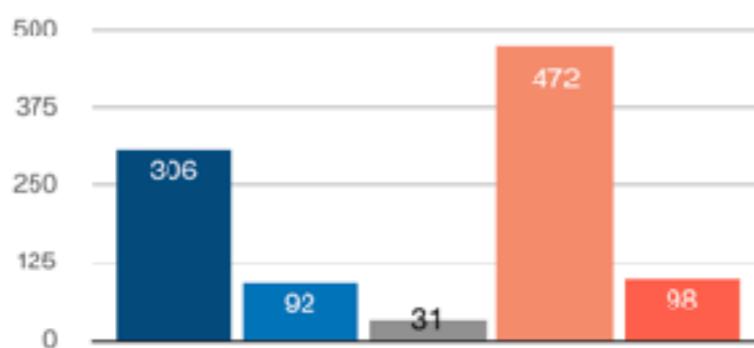
จำนวนวันที่ชำระเงินล่าช้ามากที่สุด

Grace Period

- เป็นกลุ่มที่ **ชำระเงินปกติ** อาจจะไม่ต้องติดตาม
- เป็นกลุ่มที่ส่วนใหญ่ **ชำระเงินก่อนเวลา** และ **ล่าช้าเพียงไปวัน** อาจจะส่งแค่ **SMS** ไปเตือน
- เป็นกลุ่มที่ส่วนใหญ่ **ชำระเงินล่าช้าแต่เพียงไม่ถ้วน** อาจจะส่ง **SMS** ไปเตือน
- เป็นกลุ่มที่ส่วนใหญ่ **ชำระเงินล่าช้า** และล่าช้า **หลายวัน** (เกิน 5 วัน) อาจจะต้อง **โทรศัพท์แจ้ง**

สัดส่วนของลูกค้าในแต่ละกลุ่ม

■ group 1 ■ group 2 ■ group 3 ■ group 4 ■ group 5



ยอดการชำระเงินล่าช้าของกลุ่มที่ 5 แยกตามจำนวนวันที่ชำระเงินล่าช้า



Outlines

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- Machine Learning Applications
- Data Management
- Data Roles
- Data Analytics
- Machine Learning
- Machine Learning Development Steps
- **No-Code/Low-Code Machine Learning Platform**

No-Code/Low-Code ML Platforms

สำหรับการสัมมนา Dev Mountain Tech Festival

- **No-Code ML Platforms**

- No-code machine learning platforms **empower businesses to utilize the power of machine learning through simple, drag-and-drop graphical user interfaces.**
- They allow users without any programming language or coding knowledge to create machine learning applications.

- **Low-Code ML Platforms**

- Low-code machine learning platforms are similar to their no-code counterpart but they **allow users to write a few lines of code or manipulate the same.**
- The percentage of editable code depends on the tool.
- Similar to no-code platforms, low-code machine learning tools are helpful for businesses lacking professionals with AI specialization.

source: <https://www.g2.com/articles/democratizing-ai-with-low-code-and-no-code-machine-learning-platforms>

No-Code/Low-Code ML Platforms

สำหรับการสัมมนา Dev Mountain Tech Festival

Advantages

- **Accessibility:**

- enable non-tech people or businesses to build AI systems from scratch
- making AI more accessible to a wider variety of companies.

- **Usability:**

- have an intuitive drag-and-drop interface where the complexity is reduced

- **Speed:**

- often have pre-built AI models, project templates, and ready-made datasets
- significantly accelerating the model development.

Disadvantages

- **Lack of customization:**

- mostly limited in functionality, because they're designed to cover a specific problem
- it's hard to come up with out-of-the-box, more complex solutions

- **Requires consultation or training:**

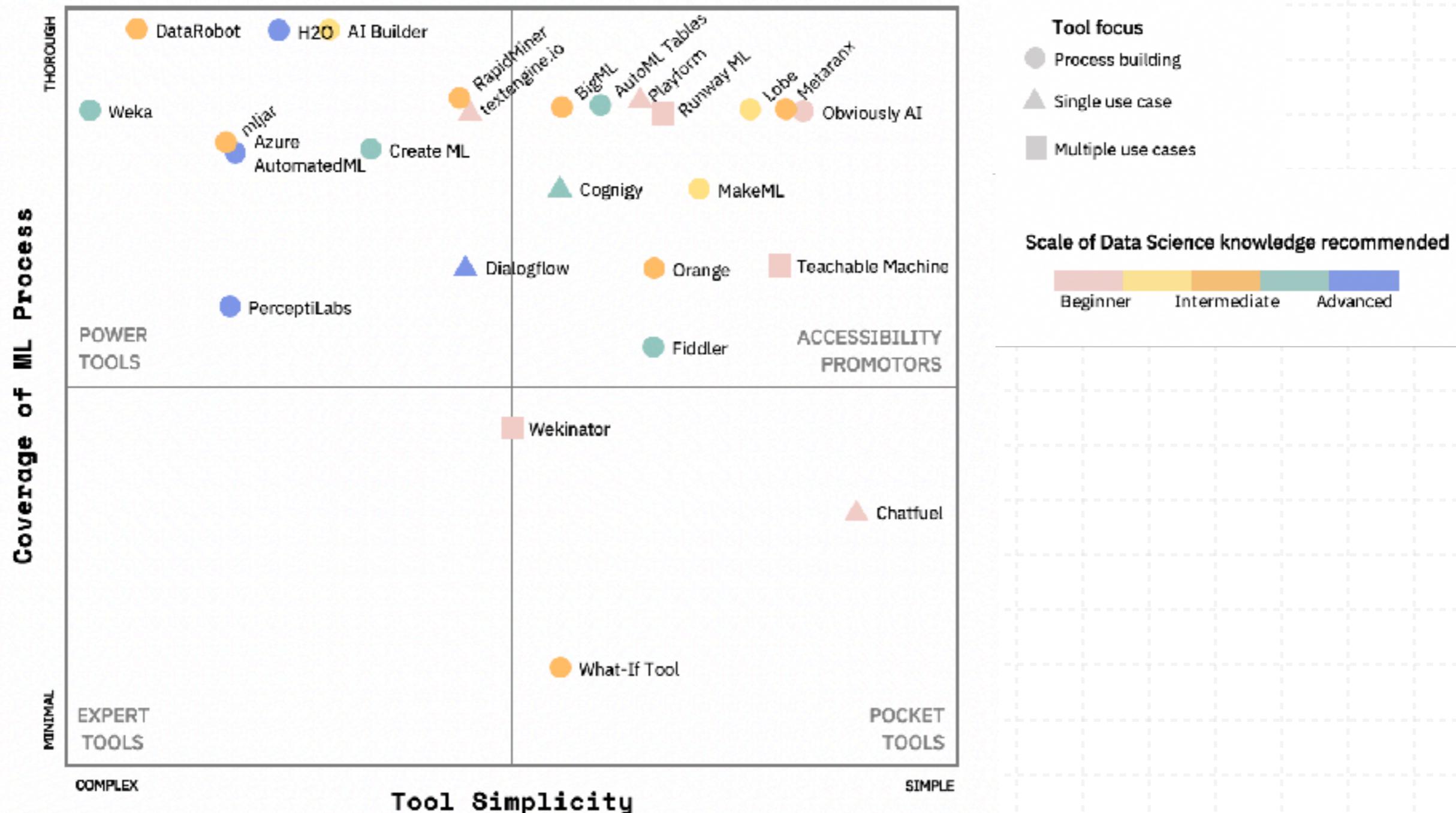
- Ideally, the ML engineer, human resources specialist, and marketing intern should be equally able to use the low-code/no-code platforms, but that's not always the case.

source: <https://blog.superannotate.com/guide-to-low-code-no-code-ai-platforms/>

No-Code/Low-Code ML Platforms

สำหรับการสัมมนา Dev Mountain Tech Festival

NO-CODE / LOW-CODE ML TOOLS ASSESSMENT

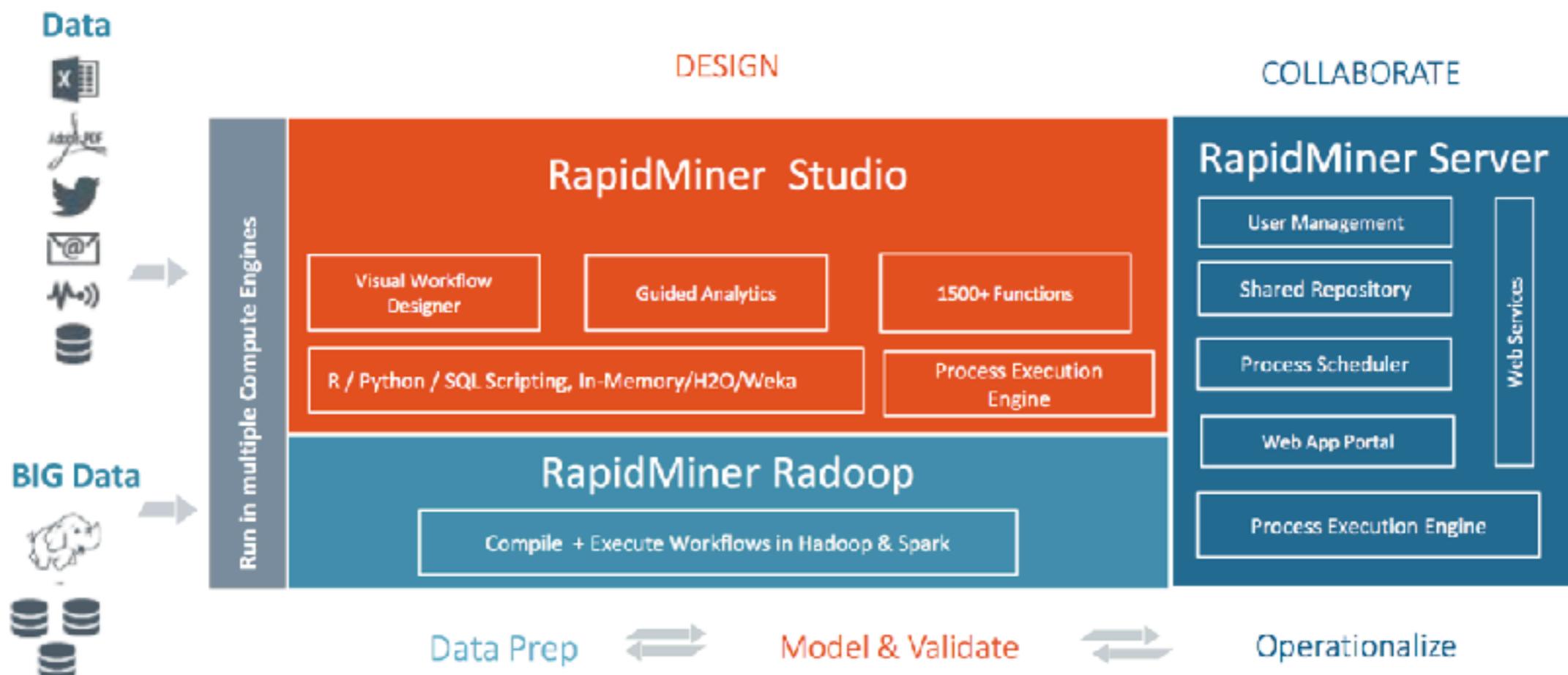


source: <https://medium.com/aixdesign/26-no-code-low-code-ml-tools-to-check-out-8b3414eaa489>

No-Code/Low-Code ML Platforms

สำหรับการสัมมนา Dev Mountain Tech Festival

Unified Data Science Platform



For more information

สำหรับการบรรยายในงาน Dev Mountain Tech Festival

- บริษัท คิวบ์ อนาคติกส์ คอนซัลติ้ง จำกัด



- website: <http://www.datacubeth.ai>
- facebook: <http://facebook.com/datacube.th> หรือ <http://facebook.com/sit.ake>
- email: eakasit@datacubeth.ai
- lineID: eakasitp