KNOWLEDGE MANAGEMENT

หลักสูตรทักษะการสร้าง Data Science Machine Learning เพื่อหาองค์ความรู้ใหม่และวิเคราะห์จากข้อมูลมหาศาล เพื่อสร้างมูลค่าต่อธุรกิจหรือองค์กร

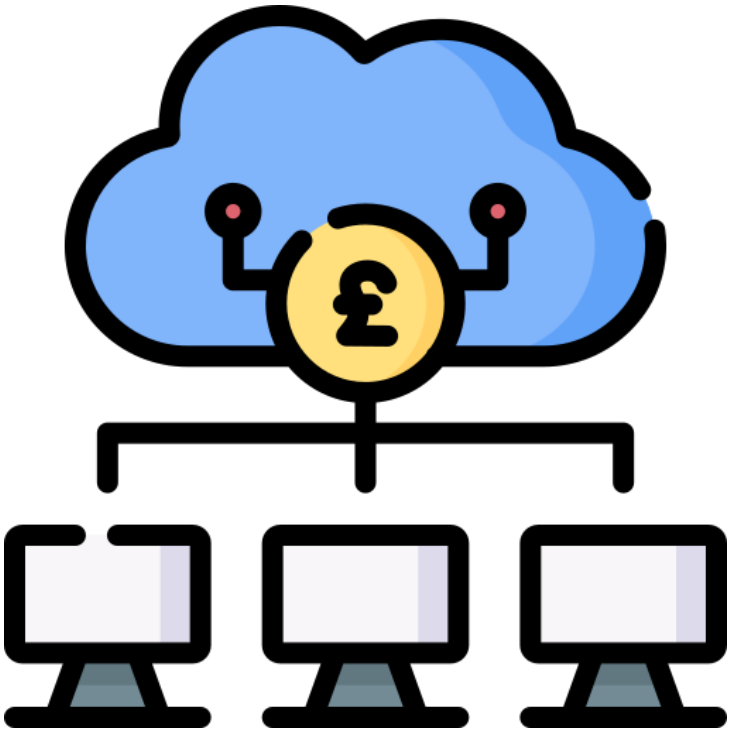
Developed Data Science Machine Learning for Business

ภายใต้แผนงานพัฒนาความสามารถทางเทคโนโลยีของบุคลากร
ภาคอุตสาหกรรม
โครงการสร้างกำลังคนและทักษะแห่งอนาคตในภูมิภาคเพื่อตอบโจทย์
การพัฒนานวัตกรรมของประเทศ ประจำปีงบประมาณ 2563

สารบัญ

บทที่ 1 : คำอธิบายโครงการและหลักสูตร

1.1 แนะนำโครงการ



BRAIN POWER SKILL UP

สร้างทักษะกำลังคนขั้นสูง เพื่อรับมือความเปลี่ยนแปลง วางแผนขับเคลื่อนธุรกิจสู่อนาคต

ภายใต้
แผนงานการพัฒนาความสามารถทางเทคโนโลยีของบุคลากรภาคอุตสาหกรรม
โครงการสร้างกำลังคนและทักษะแห่งอนาคตในภูมิภาคเพื่อตอบโจทย์การพัฒนานวัตกรรมของประเทศ

1 ภาพรวมแผนงาน "การพัฒนาความสามารถทางเทคโนโลยีของบุคลากรภาคอุตสาหกรรม (Brain Power Skill Up)"

รายละเอียด

หลักสูตรพัฒนาทักษะสำหรับภาคอุตสาหกรรม (upskill for future technology) จำนวน 20 หลักสูตร เพื่อรองรับการเปลี่ยนแปลง (transform) เทคโนโลยีของบริษัท เช่น AI, Data science, Big Data เป็นต้น

กลุ่มเป้าหมาย

บุคลากรในภาคอุตสาหกรรม

อุตสาหกรรมเป้าหมาย

วัตถุประสงค์

1. เพื่อสร้างกำลังคนและทักษะแห่งอนาคตในภูมิภาค ให้ตอบโจทย์การพัฒนานวัตกรรมของประเทศ
2. เพื่อสร้างระบบนิเวศเทคโนโลยีและนวัตกรรมในสถาบันการศึกษา

2 หลักการและความสำคัญ

- องค์กรรับมือวิกฤตหรือการเปลี่ยนแปลงได้ช้า
- บุคลากรใช้ทักษะเดิมที่ล้าสมัย ขาดองค์ความรู้ทางด้านมาตรฐาน
- โดนแชงหน้าด้วยคู่แข่งที่มีมาตรฐานสูง และล้ำสมัยกว่า

อุตสาหกรรมแบบเก่า

องค์กรขับเคลื่อนด้วยทักษะแห่งอนาคต

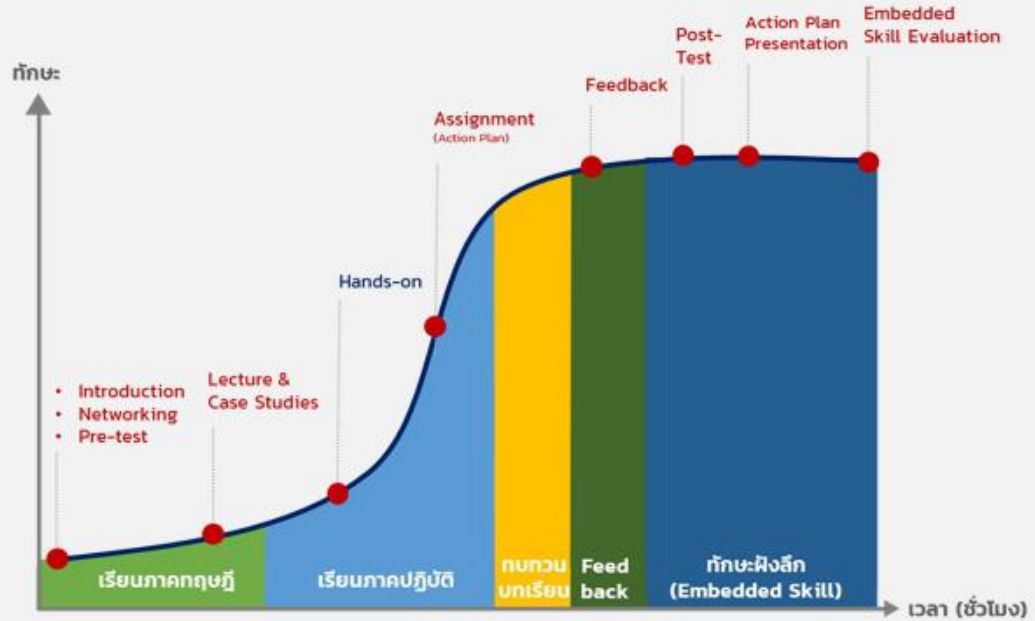
ก้าวทันความเปลี่ยนแปลงภายใต้มาตรฐาน

ติดอาวุธกำลังคน เพื่อสร้างมูลค่าและคุณค่าแก่องค์กร

อุตสาหกรรมแบบใหม่

3 กลไกการสร้างทักษะ (10 ขั้นตอน)

Josh Kaufman



1. **แนะนำโครงการ และ กิจกรรมสร้างเครือข่าย** (Networking Workshop)
2. **แบบทดสอบก่อนพัฒนาทักษะ** (Pre-Test)
   แบบประเมินทักษะก่อนเรียน (Pre-Embedded Skill Evaluation)
3. **เรียนภาคทฤษฎี** (Lecture)
4. **การอบรมเชิงปฏิบัติการ** (Case-Studies & Workshop)
5. **เรียนภาคปฏิบัติ** (Hands-On)
6. **การเขียนแผนงาน** ทุกวันหลังเลิกเรียน (Assignment – Action Plan)
7. **การให้คำปรึกษาแผนงานโดยผู้เชี่ยวชาญ** (Feedback) ทุกวันหลังเลิกเรียน
8. **บันทึกความก้าวหน้าการพัฒนาทักษะ Learning Curve Record** ทุกวันหลังเลิกเรียน
9. **แบบทดสอบหลังพัฒนาทักษะ** (Post-Test)
   แบบประเมินทักษะหลังเรียน (Post-Embedded Skill Evaluation)
10. **ผู้เรียนนำเสนอแผนงาน Action Plan**

4 ผลลัพธ์ทักษะ (Learning Output)

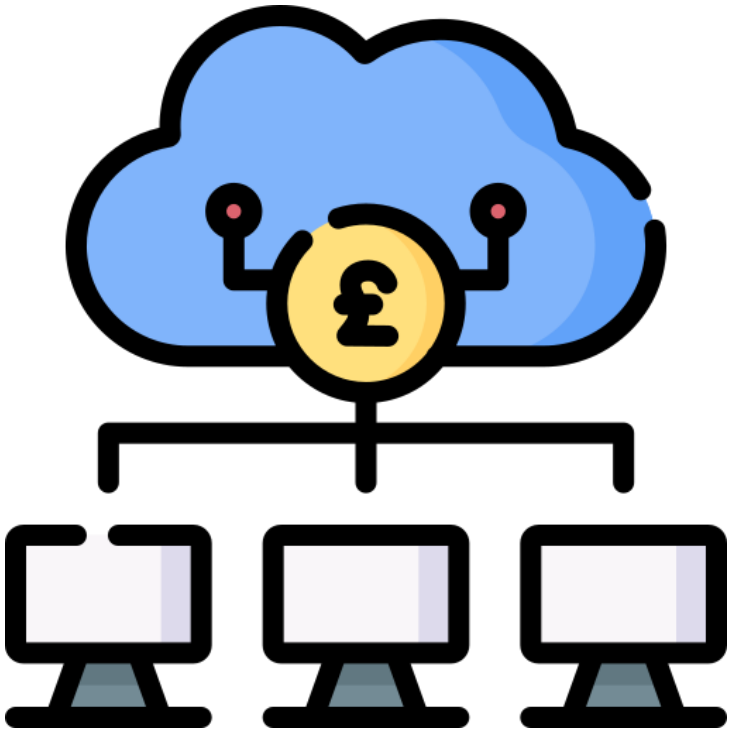
ผู้เรียนได้ทดสอบ
องค์ความรู้
ผ่าน **Pre-Test**
และ Post-Test

ผู้เรียนบันทึกและวางแผนการ
พัฒนาทักษะของตนเอง ผ่าน
Learning Curve Record

ผู้เรียนได้แผน
Action Plan
รายคน

ผู้เรียนได้รับการวัดผลสำเร็จ
การพัฒนาทักษะ ก่อนและหลัง
(Pre-Post Embedded Skill Evaluation)
โดยวัดผลลัพธ์ 5 ทักษะ
เพื่อนำมาทำ Radar Chart

1.2 คำอธิบายหลักสูตร



ประเภท 2 **การพัฒนาทักษะความสามารถทางเทคโนโลยีของบุคลากรขั้นสูง**

หลักสูตรที่ 3 **ทักษะการสร้าง Data Science Machine Learning เพื่อหาองค์ความรู้ใหม่และวิเคราะห์จากข้อมูลมหาศาล เพื่อสร้างมูลค่าต่อธุรกิจหรือองค์กร**

Developed Data Science Machine Learning for Business

คำอธิบาย : สร้างทักษะ การสร้าง Data Science Machine Learning เพื่อหาองค์ความรู้ใหม่และวิเคราะห์จากข้อมูลมหาศาล เพื่อสร้างมูลค่าต่อธุรกิจหรือองค์กร โดยต้องมีผู้เชี่ยวชาญที่สามารถจัดการนำข้อมูลที่มีมาวิเคราะห์อย่างมีประสิทธิภาพ จึงจะสามารถนำข้อมูลมาใช้งานได้อย่างมีประสิทธิภาพและเกิดประโยชน์สูงที่สุด

วัตถุประสงค์ : 1. สร้างความรู้ความเข้าใจและประโยชน์ในการนำ Data Science มาประยุกต์ใช้ในธุรกิจ
2. สร้างทักษะการสร้างและพัฒนา Data Science Machine Learning เพื่อมาประยุกต์ใช้ในธุรกิจ
3. สร้างทักษะการวิเคราะห์ข้อมูล เทคนิคการวิเคราะห์ที่จำเป็น

ผลลัพธ์ทักษะ : 1.ทักษะการวิเคราะห์ความต้องการทางด้านข้อมูล
2. ทักษะด้านการออกแบบกระบวนการการวิเคราะห์ข้อมูล
3. ทักษะในด้านการวิเคราะห์ข้อมูลเชิงสำรวจ (Exploratory Data Analysis)
4. ทักษะพื้นฐานการออกแบบการวิเคราะห์ข้อมูลด้วยเทคนิคภาษาธรรมชาติ
5. ทักษะพื้นฐานการเขียนโปรแกรมในด้านวิทยาการข้อมูล

อุตสาหกรรมเป้าหมาย : ทุกกลุ่มอุตสาหกรรม

กลุ่มเป้าหมาย : หัวหน้างาน พนักงาน

วิทยากร :

ลำดับ	ชื่อ-นามสกุล	ตำแหน่ง	หน่วยงาน	ความเชี่ยวชาญ	รูปภาพ
1	รศ.ดร. จักรพงศ์ นาทวิชัย	รองคณบดีฝ่ายยุทธศาสตร์และแผนงาน ภาควิชาวิศวกรรมคอมพิวเตอร์	ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเชียงใหม่	- Data Privacy - Information Systems - Database Systems	
2	ผศ.ดร. พร้อมพงศ์ สุกัณศีล	รองคณบดีฝ่ายวิชาการ	วิทยาลัยศิลปะ สื่อ และเทคโนโลยี มหาวิทยาลัยเชียงใหม่	- ปัญญาประดิษฐ์ - ขั้นตอนวิธีเชิงพันธุกรรม - โครงข่ายประสาทเทียม - การแก้ปัญหาด้วยวิธีการทางฮิวริสติกส์ การทำให้เหมาะสม	
3	ดร.ไพรัช พิบูลย์รุ่งโรจน์	อาจารย์	คณะเศรษฐศาสตร์ มหาวิทยาลัยเชียงใหม่	- Supply Chain Management, Supply Chain Economics.	

เนื้อหาที่เรียน :

ลำดับที่	เนื้อหาที่เรียน	ระยะเวลา (ชั่วโมง)
1	บทนำสู่วิทยาการข้อมูล (Introduction to Data Science)	0.75
2	วิทยาการข้อมูล (Data Science) • ความหมาย การครอบคลุม • ตัวอย่างของการนำองค์ความรู้ที่เกี่ยวกับวิทยาการข้อมูลไปประยุกต์ใช้	1.25
3	CRISP-DM	1.5
4	แนวปฏิบัติสำหรับกระบวนการทางวิทยาการข้อมูลโดยใช้ CRISP-DM • เหตุผล ความจำเป็น ตัวอย่าง Workshop : การฝึกปฏิบัติโดยใช้กรณีศึกษาจากองค์กรของผู้เข้าร่วมพัฒนาทักษะ	2.25
5	พื้นฐานการพัฒนาโปรแกรมทางวิทยาการข้อมูล	1.5
6	สภาพแวดล้อมพื้นฐานทางวิทยาการข้อมูล • สภาพแวดล้อมพื้นฐานสำหรับการพัฒนาโปรแกรมทางวิทยาการข้อมูล • พื้นฐานการพัฒนาโปรแกรม	1.25
7	การรวบรวมข้อมูล (Data Collecting) Workshop : การรวบรวมข้อมูล (Data Collecting) • กรณีศึกษา • กระบวนการในภาพรวม การรวบรวมข้อมูลจากหลายแหล่งซึ่งรวมไปถึงแหล่งข้อมูลที่ไม่เป็นทางการ (Non-conventional Data Source)	3.75
8	การมุ่งเน้นคุณภาพข้อมูล (Data Cleansing)	1.5
9	แนวคิดด้านคุณภาพข้อมูล • กระบวนการที่ทำให้เกิดคุณภาพข้อมูล	1.25
10	การจัดเตรียมและการวิเคราะห์ข้อมูล (Data Analytics) Workshop : การจัดเตรียมและการวิเคราะห์ข้อมูล (Data Analytics) • การจัดเตรียมข้อมูลให้เหมาะกับการวิเคราะห์ (Data Preparation) • การวิเคราะห์ข้อมูล (Data Analytics) ด้วยเครื่องมือสำเร็จรูปสำหรับผู้ใช้ (Business user tools)	3.75
11	การประมวลผลภาษาธรรมชาติ	1.5

12	แนวคิดพื้นฐานของการประมวลผลภาษาธรรมชาติ (Natural Language Processing)	1.25
13	การประยุกต์ใช้วิทยาการข้อมูล (Supply Chain Management)	3.75
14	การประยุกต์ใช้วิทยาการข้อมูล (Tourism)	2.75
15	การเขียนแผนโครงการทางวิทยาการข้อมูลด้วย Data Canvas	3.75
รวม		**31.75**

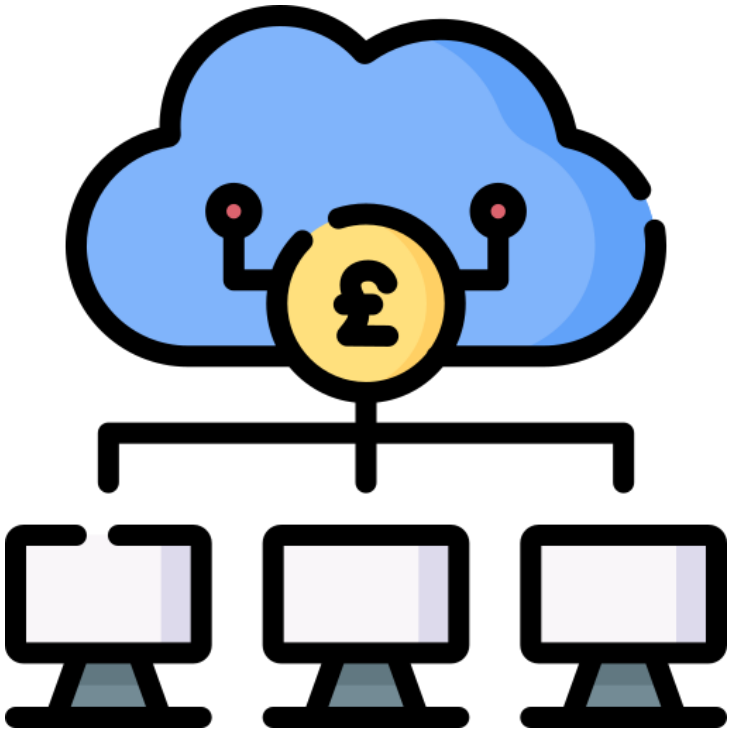
วิธีการเรียน :

1. แนะนำโครงการ Brain Power Skill Up
2. แบบทดสอบก่อนพัฒนาทักษะ (Pre-Test)/แบบประเมินทักษะก่อนเรียน (Pre-Embedded Skill Evaluation)
3. เรียนภาคทฤษฎี (Lecture)
4. การเขียนแผนงาน ทุกวันหลังเลิกเรียน (Assignment – Action Plan)
5. บันทึกความก้าวหน้าการพัฒนาทักษะ Learning Curve Record ทุกวันหลังเลิกเรียน
6. แบบทดสอบหลังพัฒนาทักษะ (Post-Test)/แบบประเมินทักษะหลังเรียน (Post-Embedded Skill Evaluation)

บทที่ 2 : แบบทดสอบและประเมินผลก่อนเรียน

2.1 แบบทดสอบก่อน

พัฒนาทักษะ (Pre-Test)



ส่วนที่ 1 ลงทะเบียน

1. กรุณากรอกชื่อ-นามสกุล.....

2. สถานประกอบการ.....

3. Email.....

4. เบอร์โทร.....

ส่วนที่ 2 แบบทดสอบก่อนพัฒนาทักษะ (Pre-Test)

คำชี้แจง 1. แบบทดสอบฉบับนี้เป็นแบบอัตนัย จำนวน 4 ข้อ 5 คะแนน

2. จงเลือกคำตอบที่ถูกต้องที่สุดเพียงข้อเดียว

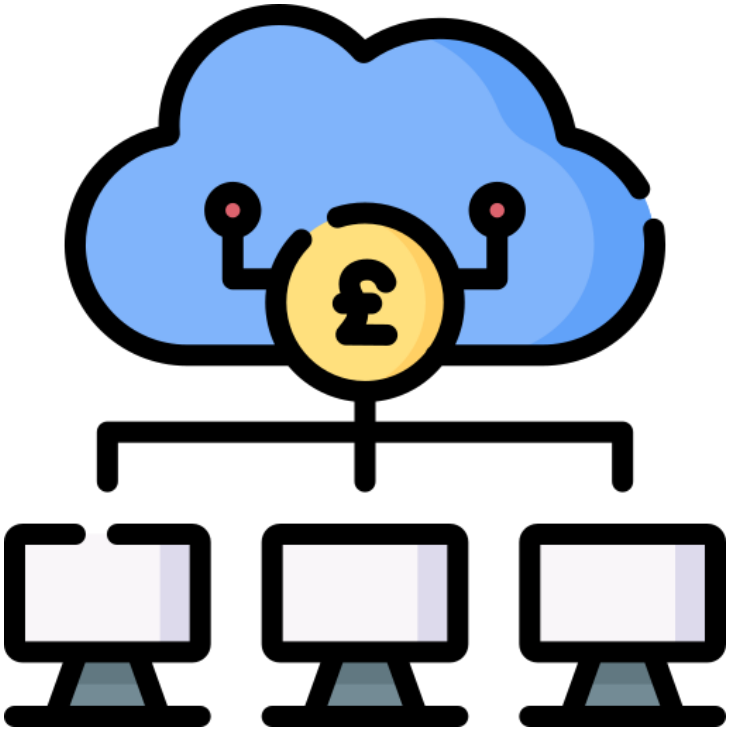
1. อะไรคือ CRISP-DM และทำไมต้องใช้ CRISP-DM จงอธิบายพร้อมยกตัวอย่างตามที่ท่านเข้าใจ

2. จงอธิบายความต่างระหว่างการเข้าใจบริบททางธุรกิจ (Business Understanding) และการเข้าใจข้อมูล (Data Understanding) จงอธิบายพร้อมยกตัวอย่างตามที่ท่านเข้าใจ

3. จงอธิบายปัญหาที่จะเกิดกับข้อมูล 1 ปัญหา พร้อม วิธีแก้ไข ตามที่ท่านเข้าใจ

4. จงอธิบายแนวทางการประยุกต์ใช้ วิทยาการข้อมูล กับ องค์กรของท่านตามแนวทาง CRISP-DM

2.2 แบบประเมินทักษะก่อนการพัฒนาทักษะ (Pre-Embedded Skill)



ส่วนที่ 1 สำหรับ ผู้เรียน

1.1 ข้อมูลทั่วไป

ชื่อ-นามสกุล.....

ชื่อสถานประกอบการ .....

ส่วนที่ 2 สำหรับ เจ้าของกิจการ หรือ หัวหน้างาน

2.1 การประเมินผู้เรียน

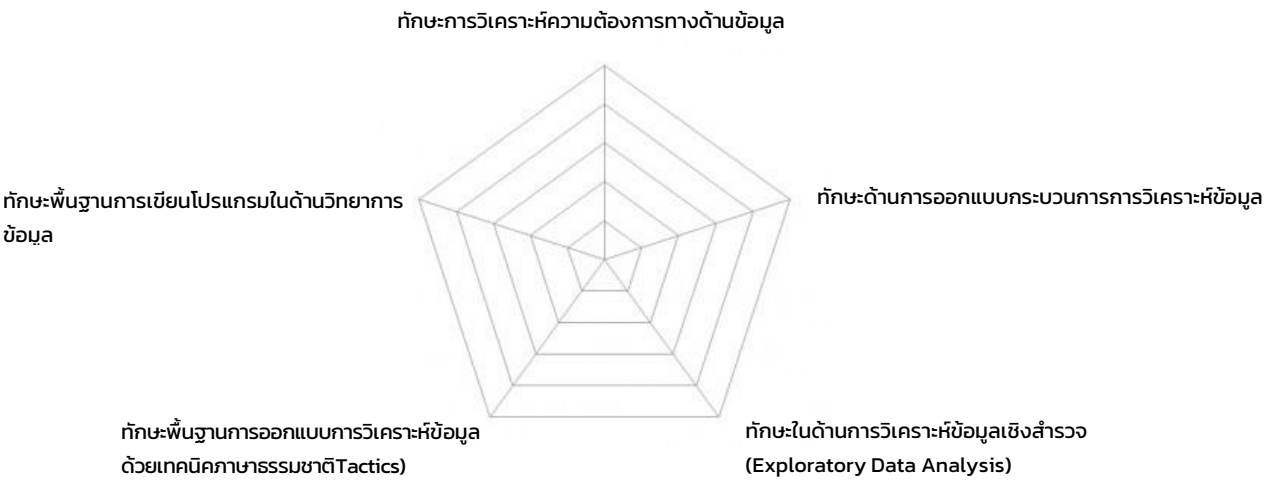
ความหมายระดับคะแนน

0 = Beginner	ไม่มีความรู้ ไม่มีทักษะ
1 = Learner	มีความเข้าใจในทฤษฎีเบื้องต้น
2 = Practitioner	มีความเข้าใจในทฤษฎีอย่างเต็มที่ มีความรู้ด้านปฏิบัติเล็กน้อย สามารถตอบคำถามหรือแก้ไขปัญหาที่ไม่ซับซ้อนได้
3 = Experienced	มีความเข้าใจในทฤษฎีและปฏิบัติอย่างเต็มที่ สามารถประยุกต์ใช้ความรู้เพื่อแก้ไขปัญหาซับซ้อนปานกลางได้
4 = Embedded	เกิดทักษะติดตัว สามารถเชื่อมโยงความรู้ในการแก้ไขปัญหาที่ซับซ้อนมากได้ และสามารถกำหนดแผนเพื่อปรับปรุงและพัฒนาประสิทธิภาพการทำงานในองค์กรได้และนำไปสู่การต่อยอดเพื่อลงมือทำจริง
5 = Broaden	เกิดทักษะอย่างท่องแท้ในระดับผู้เชี่ยวชาญ และสามารถถ่ายทอดทักษะให้แก่ผู้อื่นได้

กรุณา (√) ในช่องระดับคะแนน

ผลลัพธ์ทักษะ	ระดับคะแนน					
	0	1	2	3	4	5
1. ทักษะการวิเคราะห์ความต้องการทางด้านข้อมูล						
2. ทักษะด้านการออกแบบกระบวนการการวิเคราะห์ข้อมูล						
3. ทักษะในด้านการวิเคราะห์ข้อมูลเชิงสำรวจ (Exploratory Data Analysis)						
4. ทักษะพื้นฐานการออกแบบการวิเคราะห์ข้อมูลด้วยเทคนิคภาษาธรรมชาติ						
5. ทักษะพื้นฐานการเขียนโปรแกรมในด้านวิทยาการข้อมูล						

การวิเคราะห์ผลการพัฒนาทักษะด้วยกราฟเรดาร์ (Radar Chart)



บทที่ 3 : การพัฒนาทักษะ

3.1 บทที่ 1 :
Introduction to Data Science

กระบวนการทางข้อมูล

Data Process

ผู้ช่วยศาสตราจารย์ ดร. พรัอมพงศ์ สุกัณศีล

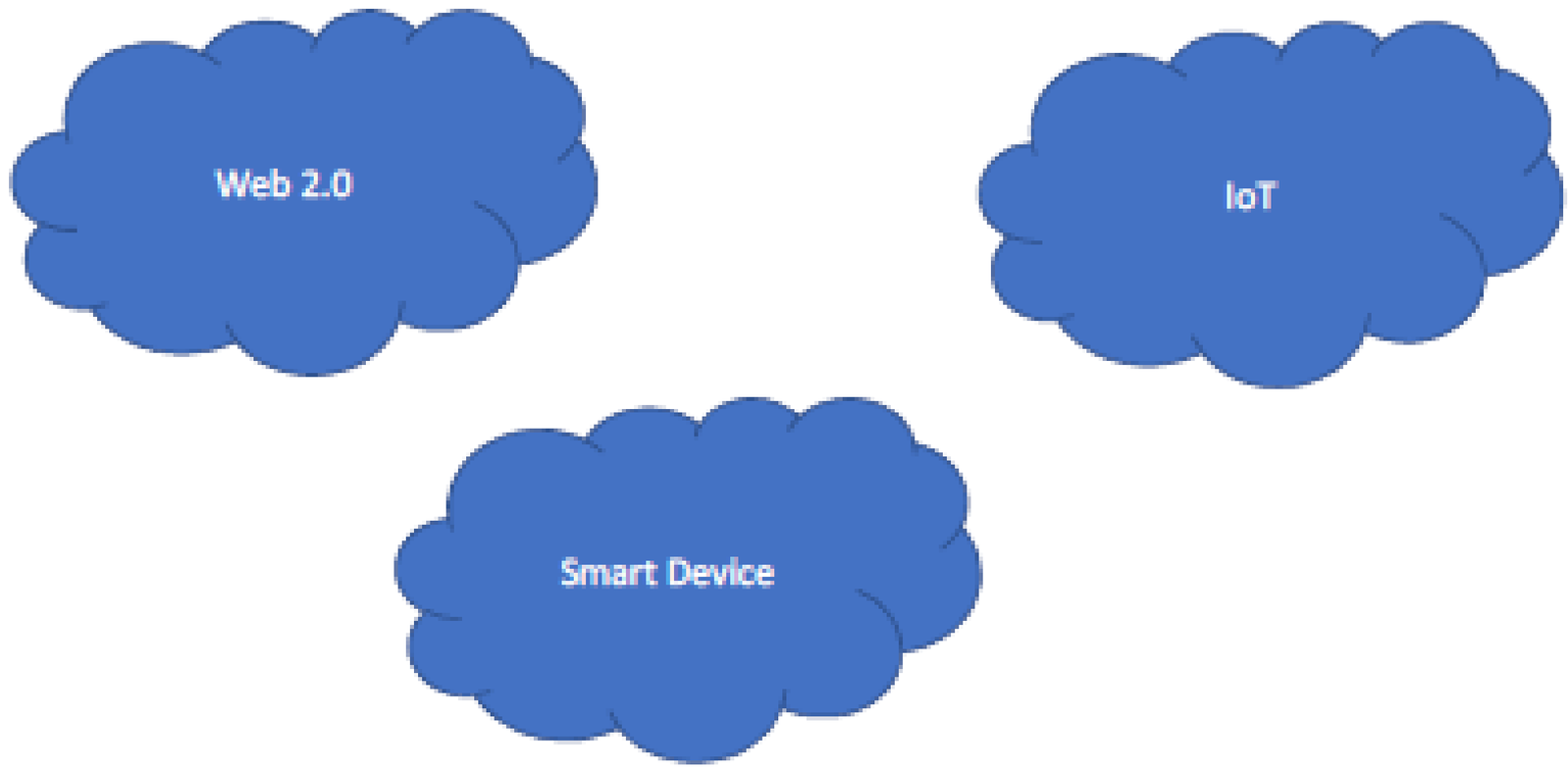
วิทยาการข้อมูลคืออะไร

Data science is an interdisciplinary field that uses **scientific methods**, **processes**, **algorithms** and **systems** to **extract** knowledge and insights from data.

วิทยาการข้อมูล คือ สหวิทยาการที่ใช้ **ขั้นตอนทางวิทยาศาสตร์** **กระบวนการ** **ขั้นตอนวิธี** และ**ระบบ** เพื่อ**สกัด** องค์ความรู้และข้อมูลเชิงลึก จากข้อมูล

https://datascience.virginia.edu/pages/uva-plans-new-school-data-science

โอกาสของวิทยาการข้อมูล



จะเกิดอะไรขึ้นหาก ไม่สนใจข้อมูล....

- ขาดข้อมูลความเห็นของลูกค้า

Reviews of Lodge of Four Seasons Golf Resort, Marina & Spa from real guests

Provided by verified guests of agoda

AGODA REVIEWS (2) | BOOKING.COM REVIEWS (360)

5.2/10 Review score
From 2 reviews

Cleanliness	4.0	Facilities	4.0
Location	9.0	Service	4.0
Value for money	3.0		

Rating
- 9+ Exceptional (0)
- 8-9 Excellent (0)
- 7-8 Very Good (0)
- 6-7 Good (1)
- <6 Below Expectation (0)

Time of the year
- Mar - May (0)
- Jun - Aug (1)
- Sep - Nov (0)
- Dec - Feb (0)

9.5 Exceptional

What guests loved the most:

"1) Excellent and fast service anywhere in the resort.
2) Staff are very friendly and attendant.
3) Beautiful sight at the beach and main pool area.
4) Excellent kids club facility and activities.
5) Spa is amazing.
6) Airport shuttle was up to expectation and worth it.
7) Attention to details by resort management is clear and well appreciated."

Pickles

wait

I really do hope I can visit you for everything!"

uth Korea

to be if all you want is the a rooms are romantic and the great.
planned every day to keep ou finish your vacation book.

จะเกิดอะไรขึ้นหาก ไม่สนใจข้อมูล....

ธนาคารสีม่วงห่วยและช้ามากกกกกก

ธนาคาร | บัตรเครดิต | การเงิน | ช้อปทุกข์ | ธุรกรรมทางการเงิน

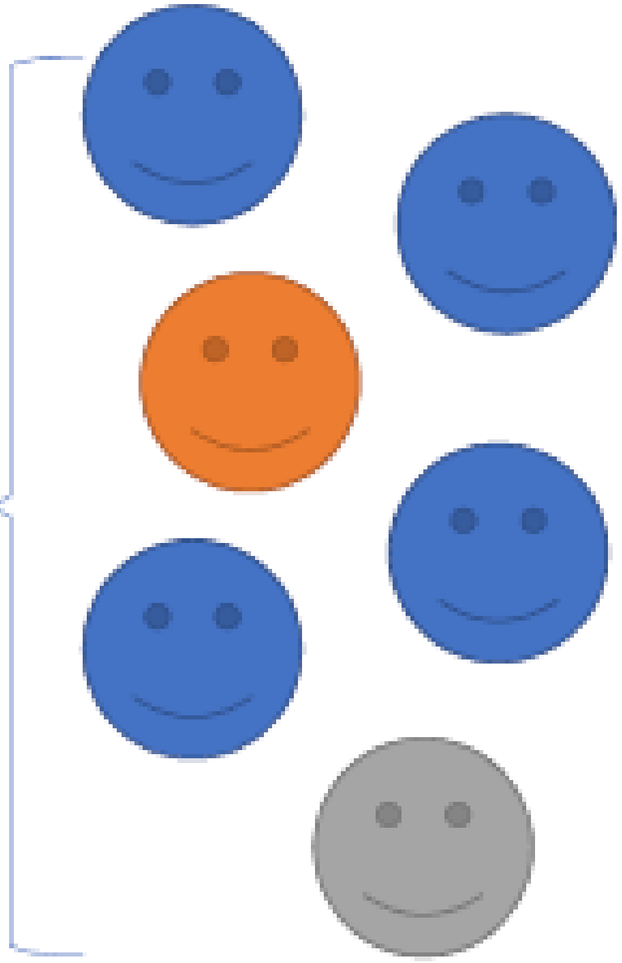
เรามีน้องชายคนนึง แล้วมีช่วงนึงเอาบัตรเครดิตธนาคารสีม่วงของน้องไปเติมน้ำมันและซื้อของ
คือบัตรน้องชายเราเป็นแบบเอาเงินค้ำไว้ 20,000 บาท แล้ววันที่ 25 สิงหาคม 2561
เราให้น้องชายเอาเงินไปปิดบัตรเครดิตที่ธนาคารสีม่วง
เพราะธนาคารบอกว่า อาทิตย์เดียวได้เงินคืน ก็เลยคิดว่าแค่ 7 วัน คงทันเอาไปจ่ายประกันให้พ่อ
ตอนนี้วันที่ 19 กันยายน 2561 ก็จะเดือนนึงแล้วค่ะ
บอกแค่ว่าเรื่องอยู่ส่วนกลางใหญ่ ถามอีกทีก็บอกว่าจะส่งเล่มกลับมาที่สาขา
แค่ส่งสมุดบัญชีคืนเนี่ย มันใช้เวลาเป็นอาทิตย์เลยเหรอคะ

https://pantip.com/topic/38073050

จะเกิดอะไรขึ้นหาก ไม่สนใจข้อมูล....

- ไม่สามารถตอบโจทย์ความต้องการของลูกค้าได้

จะเกิดอะไรขึ้นหาก ไม่สนใจข้อมูล....



Customer Profiling

- Personalization
- Empowerment
- Quick response
- Competitive value

Customer Profiling

- Demography
- Face recognition
- Online activities
- History
- Social network

จะเกิดอะไรขึ้นหาก ไม่สนใจข้อมูล....

61% of marketing decision makers said they struggled to access or integrate the data they needed last year.

https://www.thinkwithgoogle.com/marketing-resources/data-measurement/marketing-analytics-data-challenges-opportunities/

จะเกิดอะไรขึ้นหาก ไม่สนใจข้อมูล....

- ไม่ได้ใช้ประโยชน์จากอดีต

What if your biggest customer, customers was suddenly to decline by 20%.

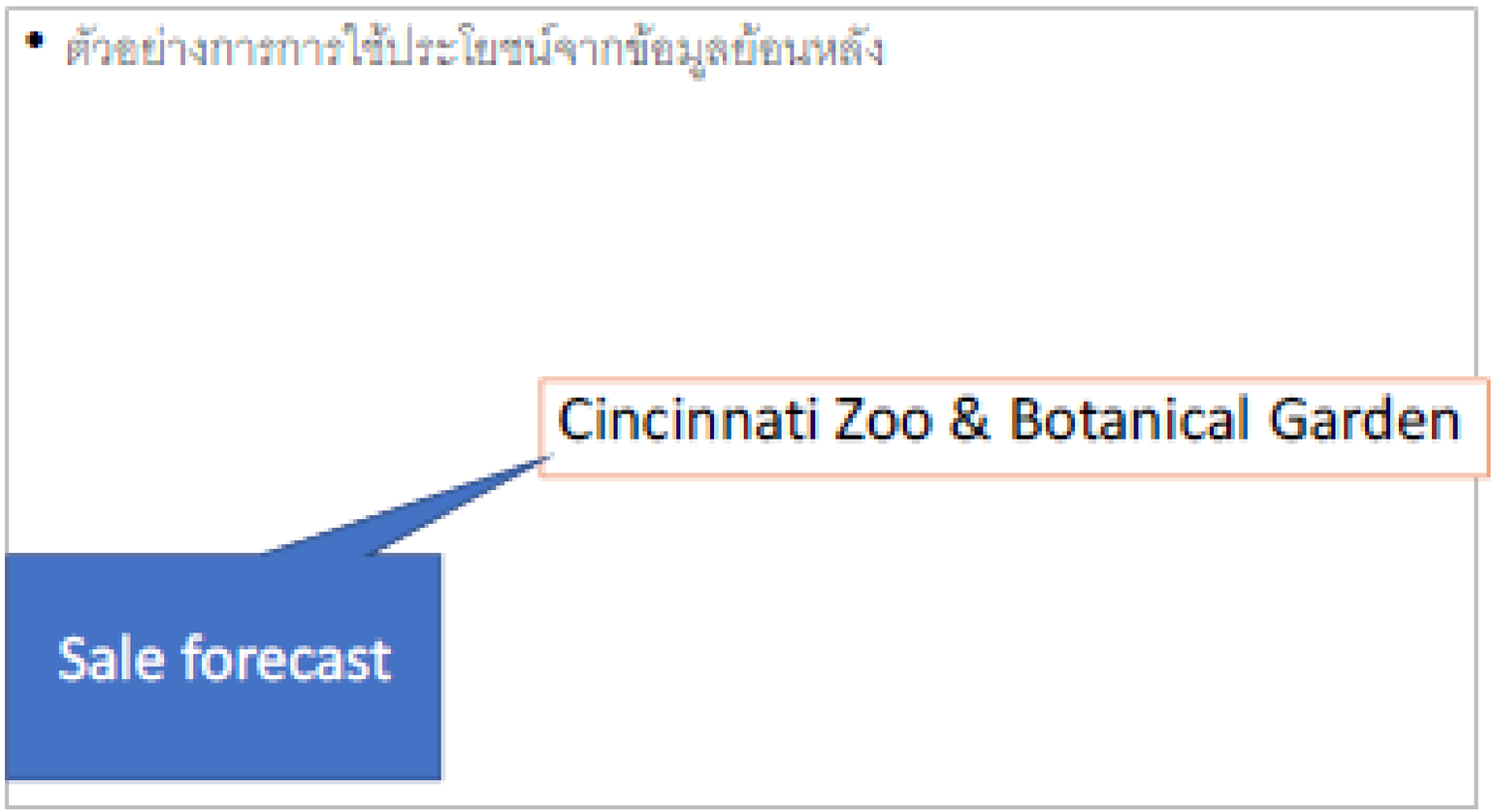
What if your store were to become ten times bigger.

What if a hurricane hit your city.

What if your opponent declined by 20%.

จะเกิดอะไรขึ้นหาก ไม่สนใจข้อมูล....

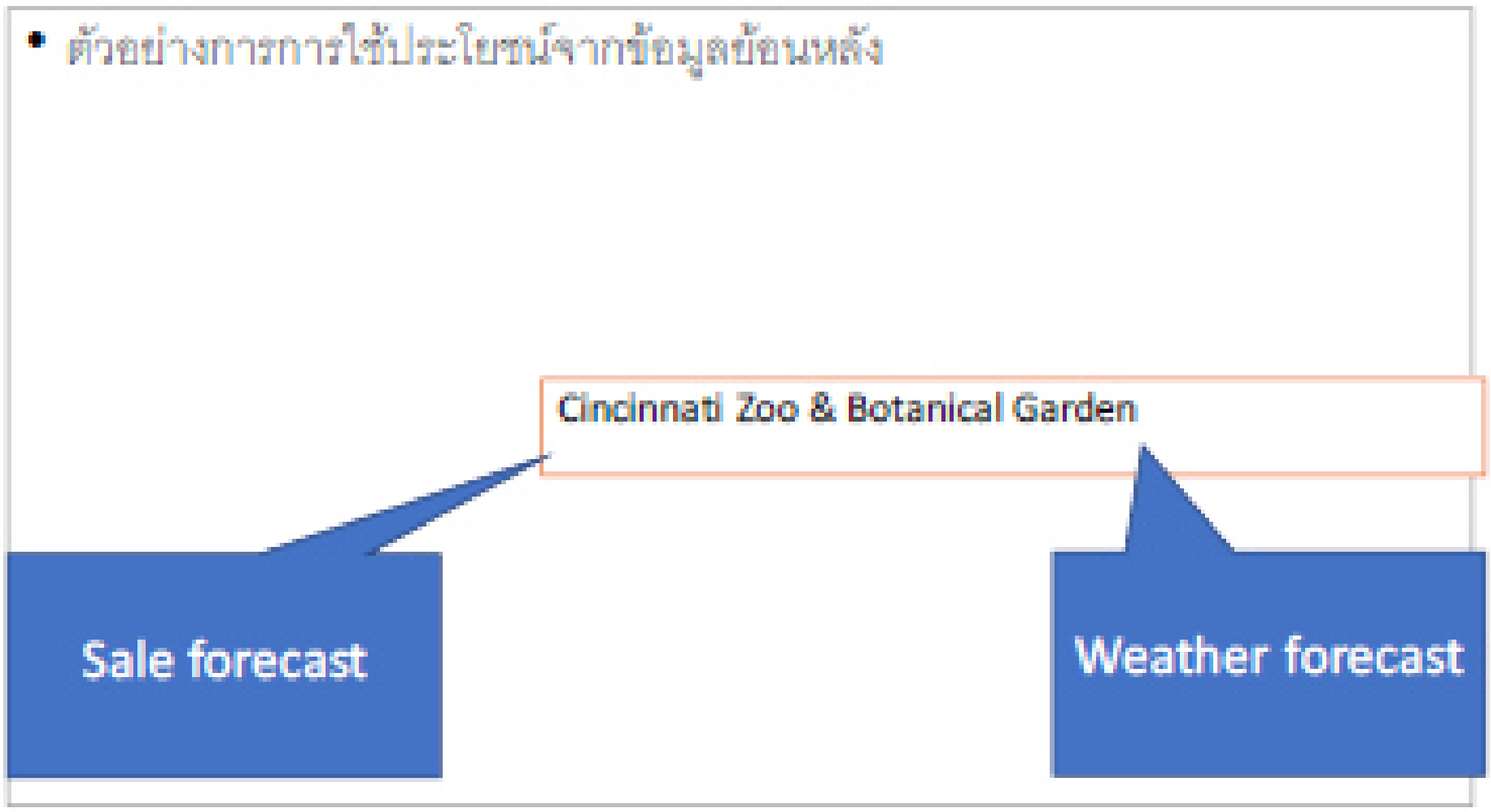
- ตัวอย่างการการใช้ประโยชน์จากข้อมูลย้อนหลัง



IBM Software Business Analtyics, "Cincinnati Zoo transforms customer experience and boosts profits," © IBM Corporation 2012.

จะเกิดอะไรขึ้นหาก ไม่สนใจข้อมูล....

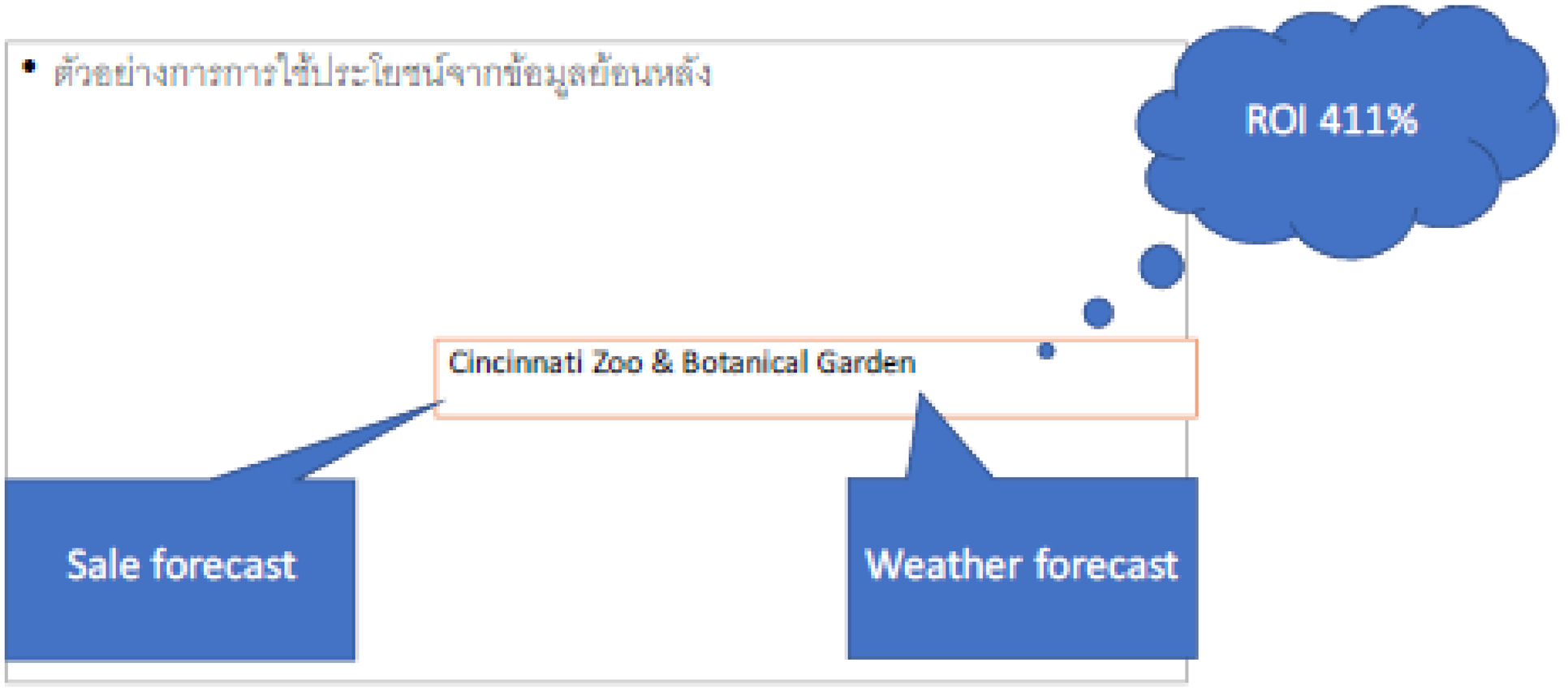
- ตัวอย่างการการใช้ประโยชน์จากข้อมูลข้อนหลัง



IBM Software Business Analtyics, "Cincinnati Zoo transforms customer experience and boosts profits." © IBM Corporation 2012.

จะเกิดอะไรขึ้นหาก ไม่สนใจข้อมูล....

- ตัวอย่างการการใช้ประโยชน์จากข้อมูลย้อนหลัง



IBM Software Business Analtyics, "Cincinnati Zoo transforms custome experience and boosts profits," © IBM Corporation 2012.

Customer targeting

- 4.2% rise in ticket sales

Optimize product mix

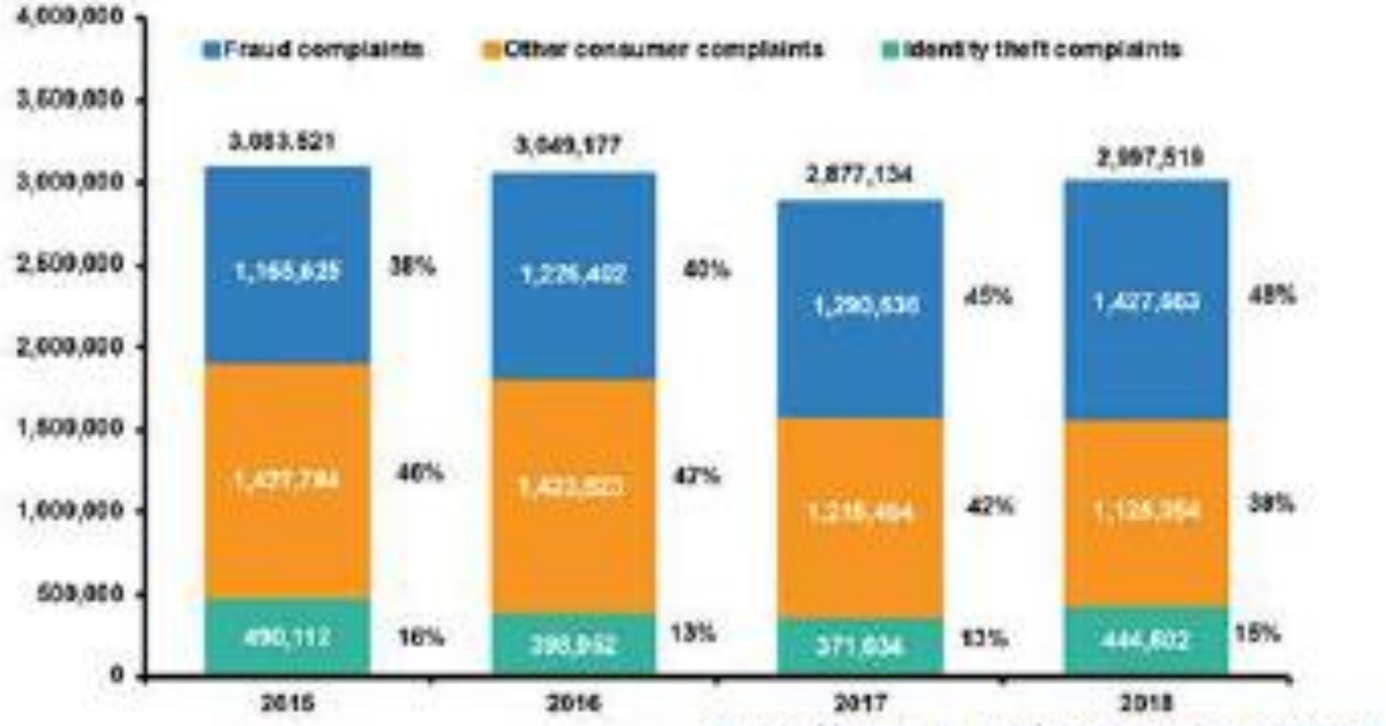
- 25% rise in food revenues

Eliminate ineffective campaigns

- Save 40,000$ in first year
- 43% reduce in advertising expenditure

จะเกิดอะไรขึ้นหาก ไม่สนใจข้อมูล....

- ไม่สามารถตรวจจับการข้อโกงได้



https://www.iii.org/graph-archive/96074

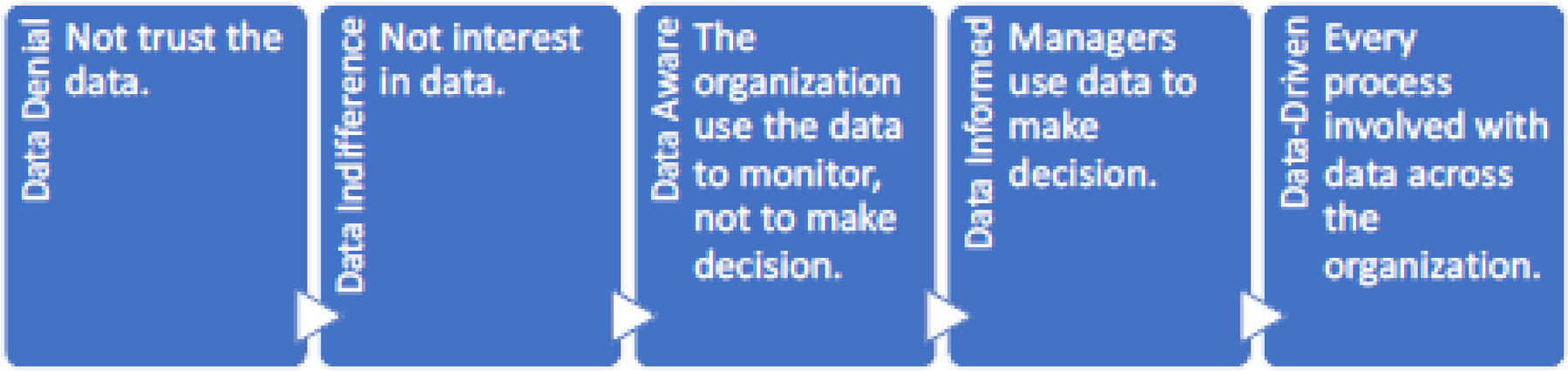
Source: Federal Trade Commission, Consumer Sentinel Network.

การตัดสินใจโดยใช้ข้อมูล

Practice of basing decisions on the analysis of data,
rather than purely on intuition.

การตัดสินใจโดยใช้ข้อมูล

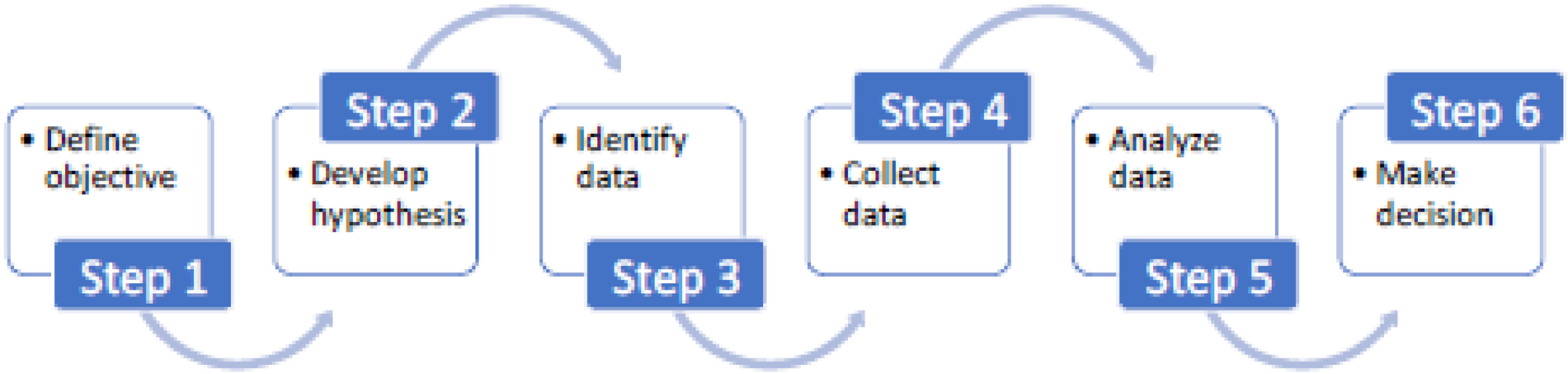
5 ระดับของการตัดสินใจโดยใช้ข้อมูล



https://www.smartsheet.com/data-driven-decision-making-manageme

การตัดสินใจโดยใช้ข้อมูล

7 ขั้นตอนของการตัดสินใจโดยใช้ข้อมูล



ข้อมูล และ การประมวลผล

- ขนาดข้อมูลเติบโตขึ้นอย่างรวดเร็ว
- เทคโนโลยี
  - พื้นที่เก็บข้อมูลถูกลงเรื่อยๆ
  - การประมวลผลกำลังสูงสามารถเข้าถึงได้ง่าย
  - มีหลายวิธีที่จะเก็บข้อมูล

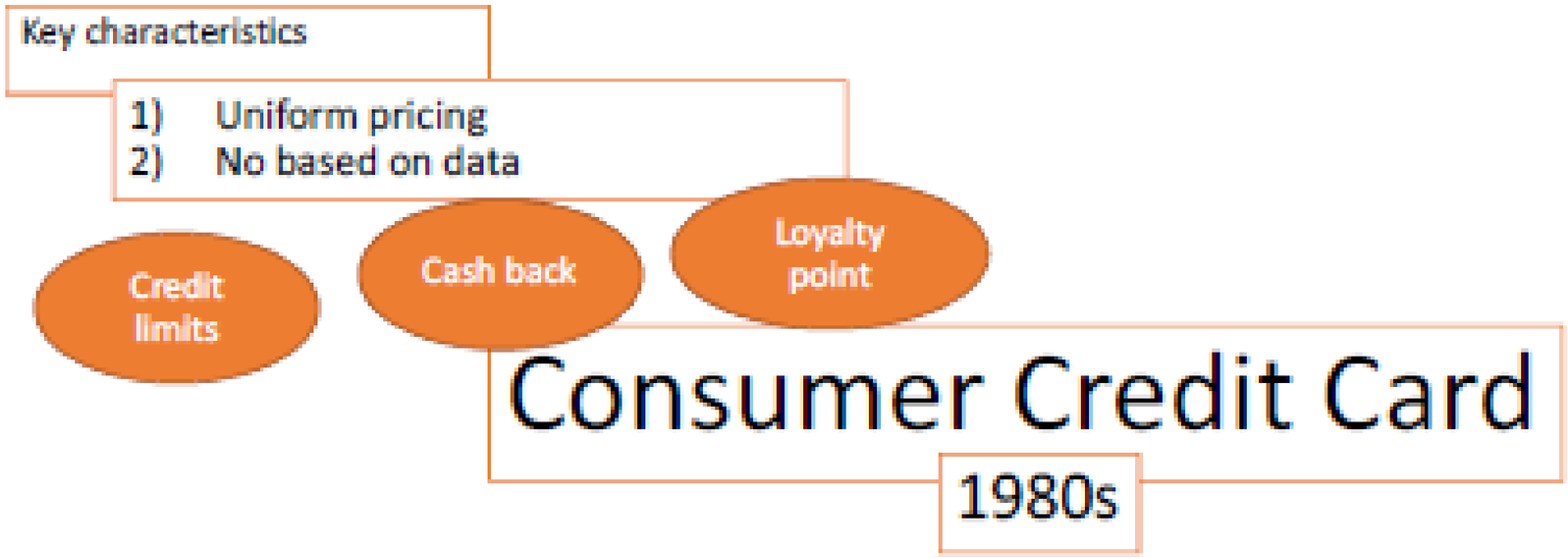
วิวัฒนาการของ Big Data

Year	Version	Key Developments
1994-2004	1.0	• E-commerce • Web mining (usage mining, structure mining, content mining)
2005-2014	2.0	• Social media mining • Sentimental analysis
2015 ++	3.0	• Streaming platform • Smart device and IOT

Lee, I. (2017). Big data: Dimensions, evolution, impacts and challenges. Business Horizons, 60(3), 293–303.doi:10.1016/j.bushor.2017.01.004

ตัวอย่างการวิเคราะห์ข้อมูลเชิงกลยุทธ์

ตัวอย่างการวิเคราะห์ข้อมูลเชิงกลยุทธ์



ตัวอย่างการวิเคราะห์ข้อมูลเชิงกลยุทธ์

Key characteristics

1) Uniform pricing
2) No based on data

Consumer Credit Card

1980s

Early stage problem

1) No data .....

Solution → Gather it !!!!

ตัวอย่างการวิเคราะห์ข้อมูลเชิงกลยุทธ์

- การเก็บข้อมูลคือทุกอย่าง !!!!
- มีการทำการทดลองที่หลากหลาย
  - อาจจะใช้เวลานาน (หลายปี)
- หลังจากได้โมเดลที่ดี ผลประกอบการของธนาคารก็เพิ่มสูงขึ้น

วิทยาการข้อมูลคืออะไร

- ประวัติศาสตร์ของวิทยาการข้อมูล (การเก็บข้อมูล)

การนับวันโดยการทำสัญลักษณ์บนเสา หรือ ไม้

↓

การจดบันทึกโดยใช้ตัวหนังสือ

↓

การเก็บข้อมูลโดยใช้ระบบอิเล็กทรอนิกส์

ตัวอักขระคูนิฟอร์ม (Cuneiform) ถูกคิดค้น
3200 BC ในพื้นที่เมโสโปเตเมีย

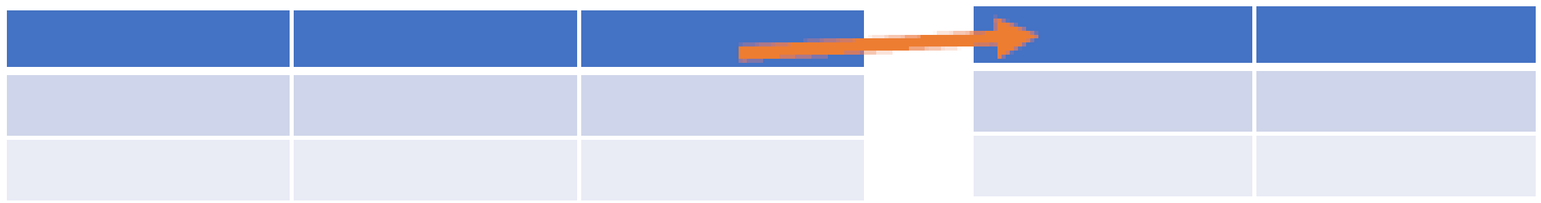
https://en.wikipedia.org/wiki/Cuneiform#/media/File:Xerxes_Cuneiform_Van.JPG

วิทยาการข้อมูลคืออะไร

- ประวัติศาสตร์ของวิทยาการข้อมูล (การเก็บข้อมูล)
- ตัวอักษระคูนิฟอร์มถูกใช้เพื่อการบันทึกข้อมูลธุรกรรม (Transactional Data)
  - ยอดขาย
  - ยอดค้างชำระ
  - เครดิต
  - ประกัน
- ข้อมูลที่ไม่ใช่ธุรกรรม (Non-transactional Data)
  - เช่น ข้อมูลประชากร
  - การสำรวจสำมะโนครั้งแรกเกิดในสมัยอียิปต์เพื่อประโยชน์ทางภาษีและการทหาร

วิทยาการข้อมูลคืออะไร

- ประวัติศาสตร์ของวิทยาการข้อมูล (การเก็บข้อมูล)
  - Edgar F. Codd นำเสนอตัวแบบการเก็บข้อมูลแบบความสัมพันธ์ (Relational Model) ใน งานวิจัย "A Relational Model of Data for Large Shared Data Banks" ซึ่งตีพิมพ์ในปี 1970



- เป็นที่มาของระบบฐานข้อมูลแบบสัมพันธ์ (Relational Database)
- ถูกใช้เพื่อเก็บข้อมูลธุรกรรมที่เกิดขึ้นจากการดำเนินการขององค์กร

วิทยาการข้อมูลคืออะไร

- ประวัติศาสตร์ของวิทยาการข้อมูล (การเก็บข้อมูล)
- ข้อมูลขนาดใหญ่ (Big Data) คือ ข้อมูลที่มีจำนวนมาก และหลากหลาย

3V :
- Variety
- Velocity
- Volume

4V :
- Variety
- Velocity
- Volume
- Veracity

5V :
- Variety
- Velocity
- Volume
- Veracity
- Value

วิทยาการข้อมูลคืออะไร

- ประวัติศาสตร์ของวิทยาการข้อมูล (การวิเคราะห์ข้อมูล)
- สถิติ (**Statistics**) คือ กระบวนการเก็บข้อมูล และวิเคราะห์ข้อมูลที่เกี่ยวข้องกับสถานะ
  - เช่น การสรุปข้อมูล การศึกษาความถี่ของข้อมูล การกระจายตัวของข้อมูล

วิทยาการข้อมูล

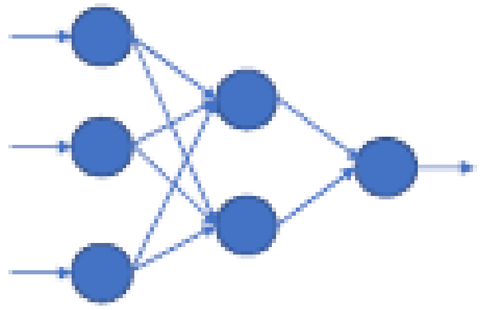
วิทยาการข้อมูล ≠ สถิติ

วิทยาการข้อมูลคืออะไร

- ประวัติศาสตร์ของวิทยาการข้อมูล (การวิเคราะห์ข้อมูล)

- การเรียนรู้ของเครื่อง (Machine Learning) คือ โปรแกรมชนิดหนึ่งที่ทำให้เครื่องคอมพิวเตอร์มีความสามารถที่จะเรียนรู้รูปแบบจากข้อมูล

- ตัวอย่างการเรียนรู้ของเครื่อง เช่น โครงข่ายประสาทเทียม การเรียนรู้เชิงลึก



ข้อมูลคืออะไร

ข้อมูล คือ ข้อเท็จจริง (Facts) ที่ถูกสังเกต (Observe) หรือ วัด (Measure)

จำนวนนับ

ความสว่าง

สี

รูปภาพ

ส่วนสูง

ความกดอากาศ

น้ำหนัก

อุณหภูมิ

ราคา

ข้อความ

ข้อมูลคืออะไร

- กรณีศึกษา

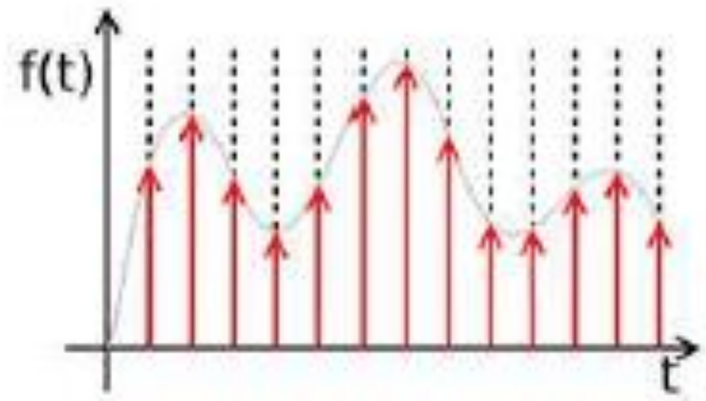
เสียงเป็นข้อมูลหรือไม่

ก.เป็น

ข.ไม่เป็น

ข้อมูลคืออะไร

- ข้อมูลสามารถถูกแบ่งออกได้เป็น 2 ประเภท
  - ข้อมูลเชิงประเภท Categorical Data
  - ข้อมูลเชิงปริมาณ Quantitative Data



ข้อมูลเชิงประเภท Categorical Data

- ข้อมูลเชิงประเภท คือ ข้อมูลเชิงคุณภาพที่มีการกำหนดค่าให้กับกลุ่ม
  - กรุ๊ปเลือด, เกรด, ชนิดสัตว์, อาชีพ
- ข้อมูลเชิงประเภท ไม่สามารถดำเนินการทางคณิตศาสตร์ได้
  - + - * / ไม่ได้
  - สามารถ นับได้
- ข้อมูลเชิงประเภทสามารถถูกแบ่งออกได้เป็น 2 ประเภท
  - ข้อมูลชื่อ Nominal Data
    - ไม่มีลำดับ
    - เช่น ชื่อ เบอร์โทรศัพท์ เพศ
  - ข้อมูลลำดับ Ordinal Data
    - มีลำดับ
    - เช่น เกรด ชั้นของดีกา

ข้อมูลเชิงปริมาณ Quantitative Data

- ข้อมูลเชิงปริมาณเกิดจากการ นับ หรือ วัด
- ข้อมูลเชิงปริมาณ**สามารถ**ดำเนินการทางคณิตศาสตร์ได้
  - + - * / ได้
- ข้อมูลเชิงปริมาณสามารถถูกแบ่งออกได้เป็น 2 ประเภท
  - ข้อมูลไม่ต่อเนื่อง **Discrete Data**
    - มีข้อมูลจำกัด
    - จำนวนสินค้า จำนวนนักศึกษา
  - ข้อมูลต่อเนื่อง **Continuous Data**
    - มีข้อมูลไม่จำกัดตามช่วงที่กำหนด
    - อุณหภูมิ น้ำหนัก ส่วนสูง

ข้อมูลคืออะไร

- กรณีศึกษา

This Photo by Unknown Author is licensed under CC BY

จำนวนเงินเป็นข้อมูลประเภทไหน

ก.ข้อมูลไม่ต่อเนื่อง

ข.ข้อมูลต่อเนื่อง

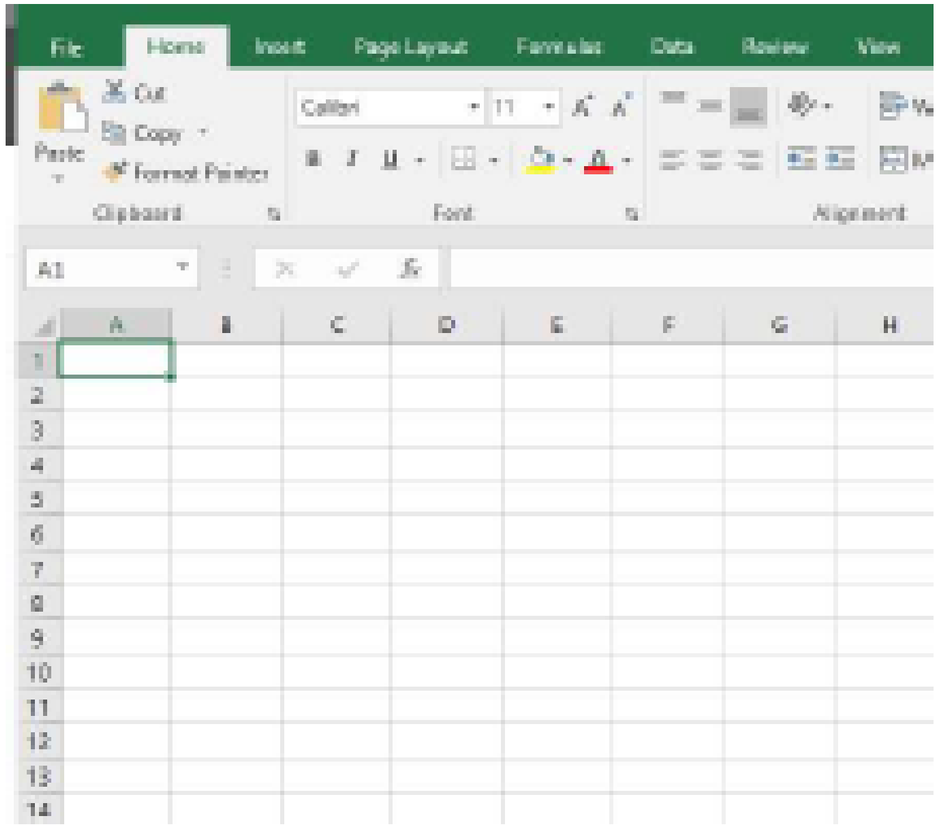
โครงสร้างของข้อมูล

โครงสร้างของข้อมูล: ข้อมูลแบบมีโครงสร้าง

- ข้อมูลถูกเก็บตามโครงสร้างของข้อมูลที่ถูกกำหนดไว้ก่อนหน้า
  - โดยปรกติ จะเป็นแบบตาราง
  - เปลี่ยนแปลงแก้ไข โครงสร้างไม่ได้

- คุณสมบัติ หรือ คุณลักษณะที่ถูกใช้เพื่ออธิบายข้อมูลจะถูกเรียกว่า “field” หรือ “attribute”.

- 1 ชุดข้อมูลจะถูกเรียกว่า 1 record


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Cut
Copy
Format Painter
Paste
Clipboard
Calibri
11
Font
Alignment
A1
A
B
C
D
E
F
G
H
1
2
3
4
5
6
7
8
9
10
11
12
13
14

Student ID	First name	Last name	Tel	Citizen ID	Address
590551020	Somchai	Somwang	1199	111111111	Lamphun
590551021	Somying	Maisomwang	9911	222222222	Lampang

Student ID	Faculty ID
590551020	10001
590551021	10002

Faculty ID	Faculty name
10001	Engineering
10002	CAMT

โครงสร้างของข้อมูล: ข้อมูลแบบไม่มีโครงสร้าง

- ข้อมูลถูกเก็บตามโครงสร้างของข้อมูล ตามการเก็บไม่มีโครงสร้างที่ถูกกำหนดล่วงหน้า
- เช่น ข้อความ ภาพ วีดีโอ

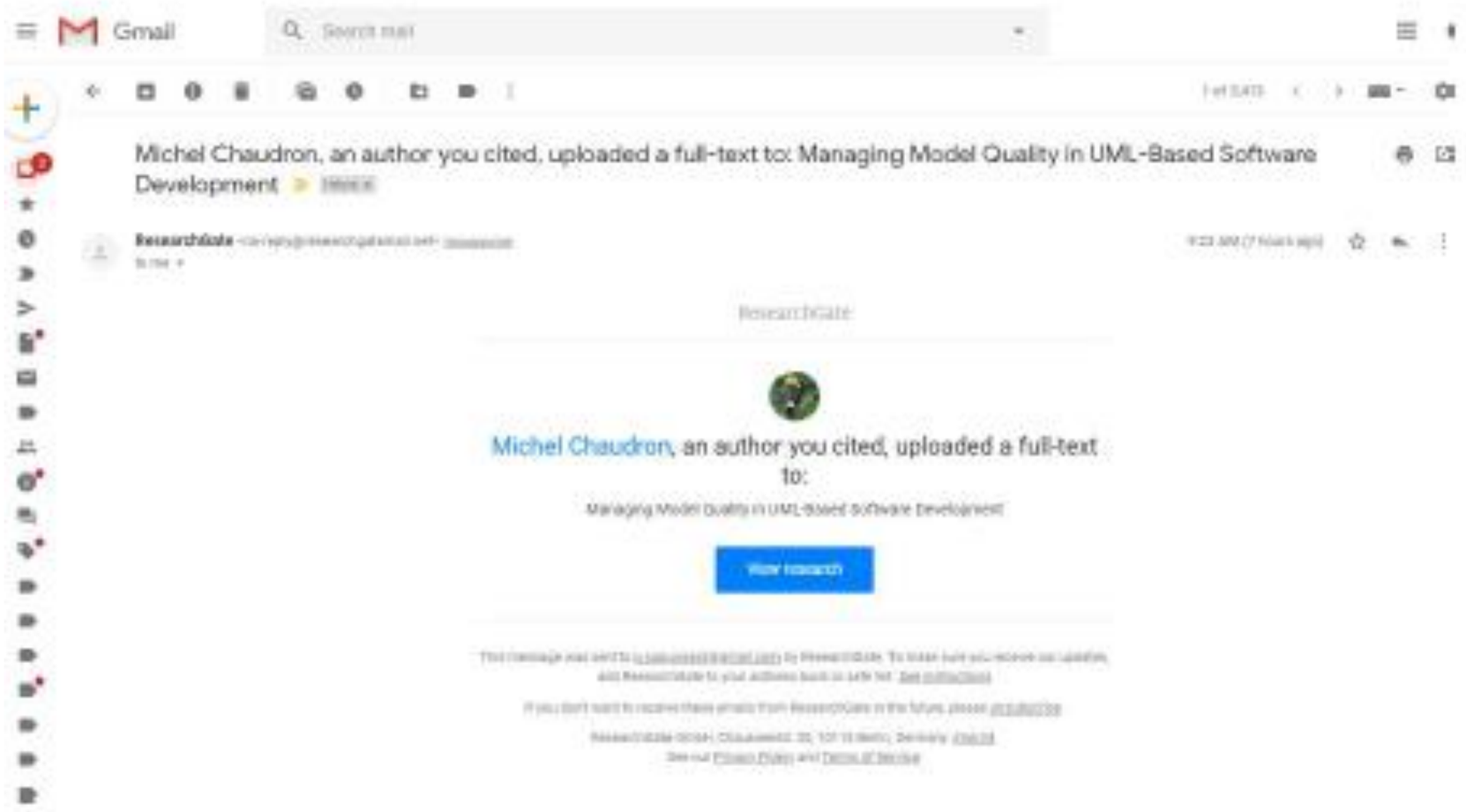
โครงสร้างของข้อมูล

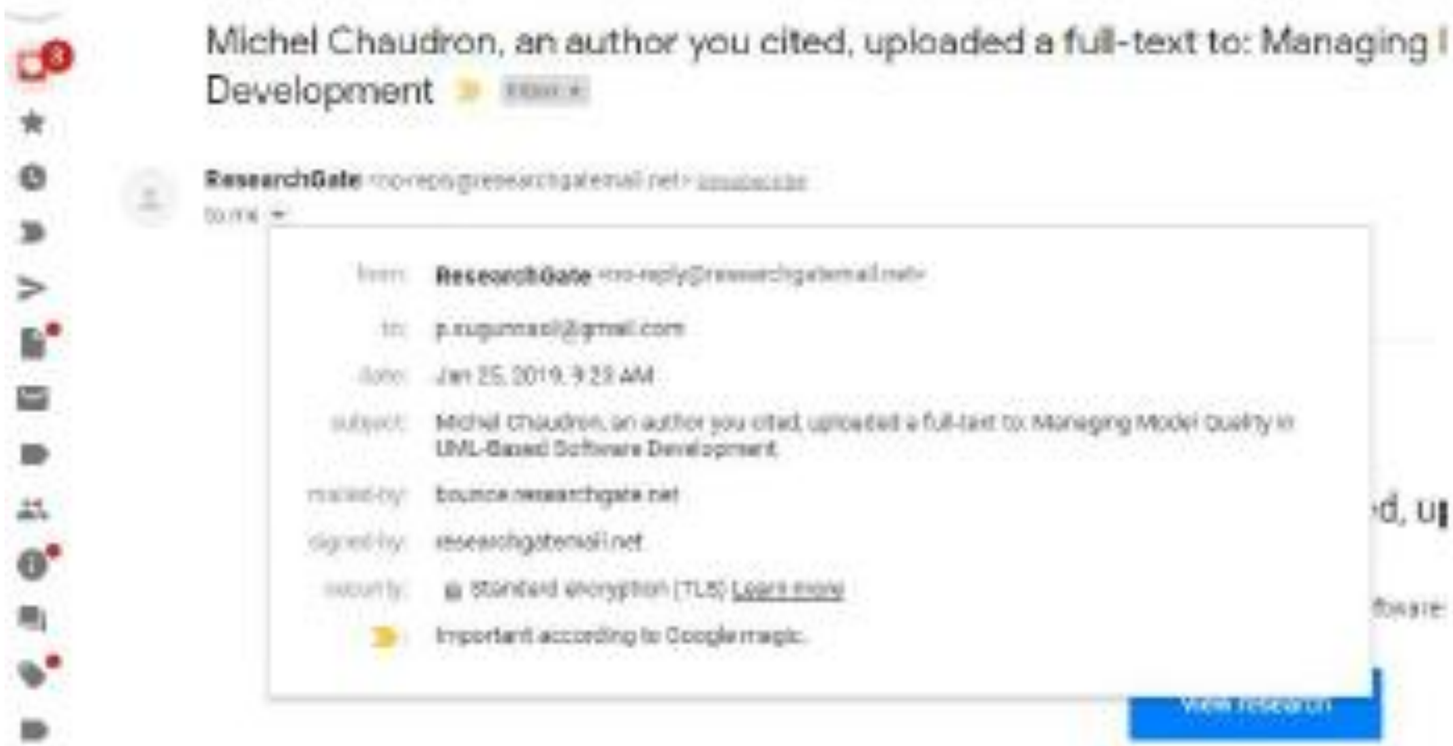
- กรณีศึกษา

Email เป็นมีโครงสร้างข้อมูลแบบไหน

ก.มีโครงสร้าง

ข.ไม่มีโครงสร้าง


Gmail
Search mail
Michel Chaudron, an author you cited, uploaded a full-text to: Managing Model Quality in UML-Based Software Development
ResearchGate
Michel Chaudron, an author you cited, uploaded a full-text to:
Managing Model Quality in UML-based software development
View research


Michel Chaudron, an author you cited, uploaded a full-text to: Managing I
Development
ResearchGate
from: ResearchGate
to: p.sugunasil@gmail.com
date: Jan 25, 2019, 9:23 AM
subject: Michel Chaudron, an author you cited, uploaded a full-text to: Managing Model Quality in UML-Based Software Development
mailed-by: bounce.researchgate.net
signed-by: researchgatemail.net
Standard encryption (TLS) Learn more
Important according to Google magic.

รูปแบบของข้อมูล Format of Data

- รูปแบบของข้อมูล คือ วิธีการเก็บข้อมูลถูกเข้ารหัสเพื่อการบันทึก หรือ ส่ง

รูปแบบของข้อมูล : Flat Files

- ไฟล์ข้อมูลที่ไม่จำเป็นต้องใช้โปรแกรมเฉพาะในการเปิด
  - เพียงข้อมูล
- รูปแบบมาตรฐานสำหรับข้อมูล
- **flat file** แบ่งออกได้เป็น 2 ประเภท
  - **Plain text –** ข้อความทั่วไป
  - **Delimited –** ข้อมูลแต่ละตัวถูกคั่นด้วยสัญลักษณ์พิเศษ

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
5.7,3.8,1.7,0.3,Iris-setosa
5.1,3.8,1.5,0.3,Iris-setosa
5.4,3.4,1.7,0.2,Iris-setosa
5.1,3.7,1.5,0.4,Iris-setosa
4.6,3.6,1.0,0.2,Iris-setosa
5.1,3.3,1.7,0.5,Iris-setosa
4.8,3.4,1.9,0.2,Iris-setosa
5.0,3.0,1.6,0.2,Iris-setosa
5.0,3.4,1.6,0.4,Iris-setosa
```

รูปแบบของข้อมูล : HTML

- หรือ **HyperText Markup Language**
- **Markup language** เป็นข้อความที่ถูกกำกับโดยสัญลักษณ์ หรือ คำสั่ง ที่ใช้ในการแปลผล
  - **Tags**
- เช่น การสร้าง **Web site**

```
<html>
<body>

<h1>My First Heading</h1>

<p>My first paragraph.</p>

</body>
</html>
```

My First Heading

My first paragraph.

รูปแบบของข้อมูล : XML

- หรือ **EXtensible Markup Language**

- โครงสร้างเหมือนกับ **HTML** แต่มีประโยชน์หลากหลายกว่าในการเก็บข้อมูล

```
<JOB>
      <NAME>Alison</NAME>
      <ID>1</ID>
      <COLOR>'blue'</COLOR>
      <DONE>FALSE</DONE>
</JOB>
<JOB>
      <NAME>Brian</NAME>
      <ID>2</ID>
      <COLOR>'red'</COLOR>
      <DONE>TRUE</DONE>
</JOB>
```

รูปแบบของข้อมูล : JSON

- หรือ JavaScript Object Notation
  - โครงสร้างคล้ายกับ XML
- เป็นโครงสร้างที่เหมาะกับการเก็บข้อมูลมากกว่า

```
[
    {
        NAME: "Alison",
        ID: 1,
        COLOR: 'blue',
        DONE: False
    },
    {
        NAME: "Brian",
        ID: 2,
        COLOR: 'red',
        DONE: True
    }
]
```

การเลือกใช้รูปแบบการเก็บข้อมูล

ชนิดข้อมูล	รูปแบบการเก็บข้อมูล
Tabular data, small data	Delimited flat file
Tabular data, large amount with lots of searching/querying	Relational database
Plain text, small amount	Flat file
Plain text, large amount	Non-relational database
Transmitting data between components	JSON
Transmitting document	XML

การพัฒนาทางด้านทรัพยากรบุคคล

กรณีศึกษาการใช้วิทยาการข้อมูลกับภาคธุรกิจ

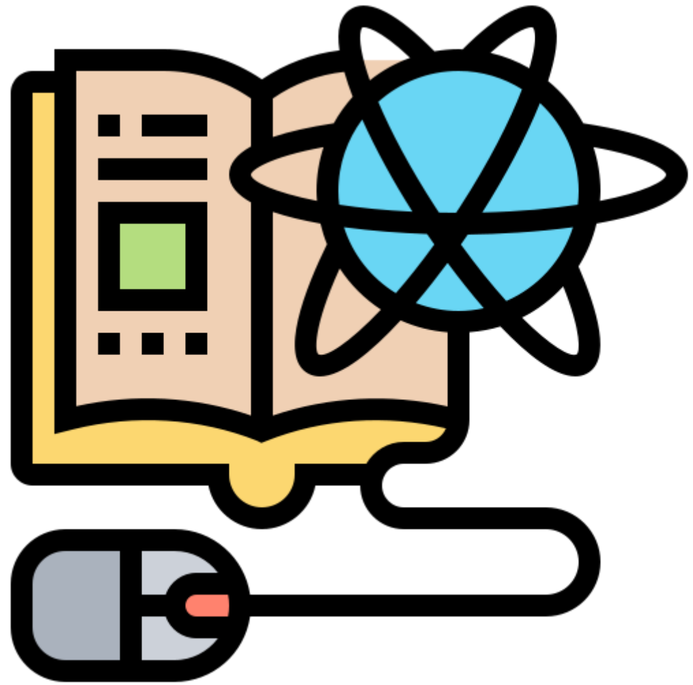
การแบ่งกลุ่มลูกค้า

การวิเคราะห์ตะกร้าสินค้า

การพยากรณ์ยอดขาย

การจัดสรรทรัพยากร

3.2 บทที่ 2 :
CRISP-DM



CRISP-DM

Agenda

- DS Process
- KDD review
- CRISP-DM
- CRISP-DM in action 1
- CRISP-DM in action 2
- On horizontal of CRIPS-DM

Why Process really does matter?

- Data Science is the **art of turning data** into impactful insights and then turning those insights into beneficial, meaningful actions.
- Being successful in Data Science **requires a team**, i.e., success comes from a cross-functional team that can work collaboratively; each member applies her/his expertise as to let the team achieve the common goal set.

Why Process really does matter?

- More and more big data or analytic projects are carried out with participation of more than one divisions within an organization.
- The lack of maturity to manage that kind of projects may result in a high failure rate.

e.g.,

Lack of reproducibility

Lack of reusability

Lack of collaboration

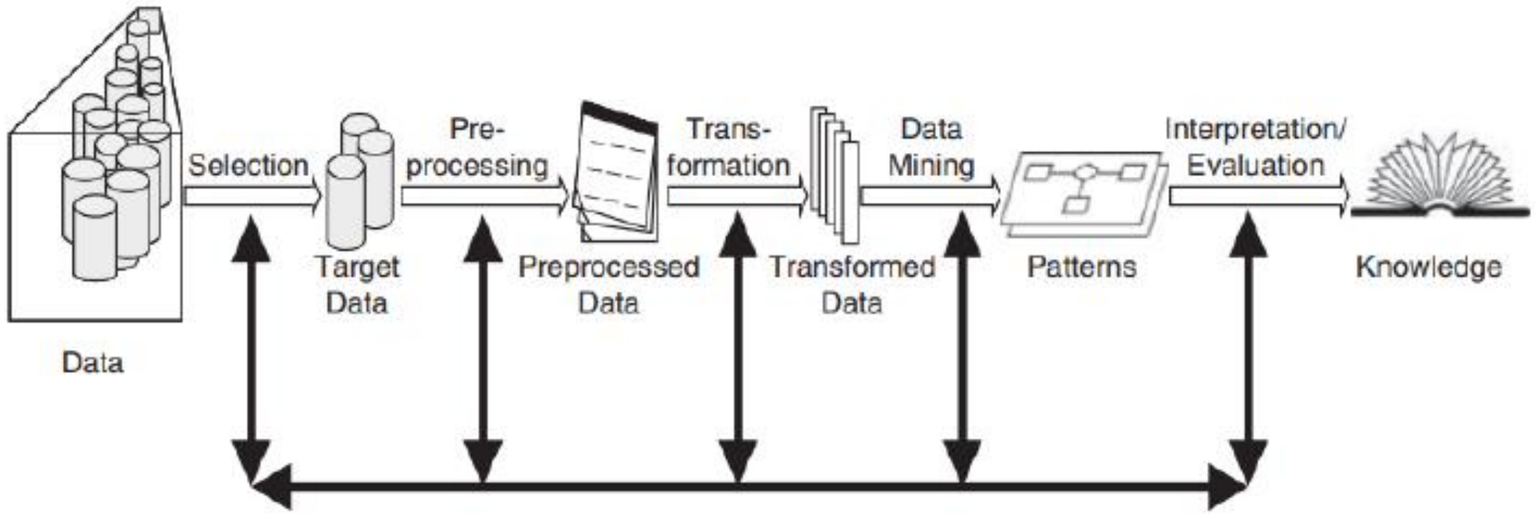
Evidence

- If one looks through analytics projects in open repositories, such as, GitHub and GitLab, likely to see a lot of **creative hacking of scripts**
  - In different programming languages
  - On different machine-learning framework
  - Nearly 0% are scalable and effortlessly deployment

Technology Blindness

- From a data scientist perspective, technology does not seem to matter too much from a functional perspective
  - The models and algorithms that are used are defined mathematically
  - Numerous tools are available, e.g., R, Python, Node-RED, KNIME, RapidMiner, Weka, SPSS, …
- **However, when it comes to maintenance and deployment, technology has a major impact on a project's success.**

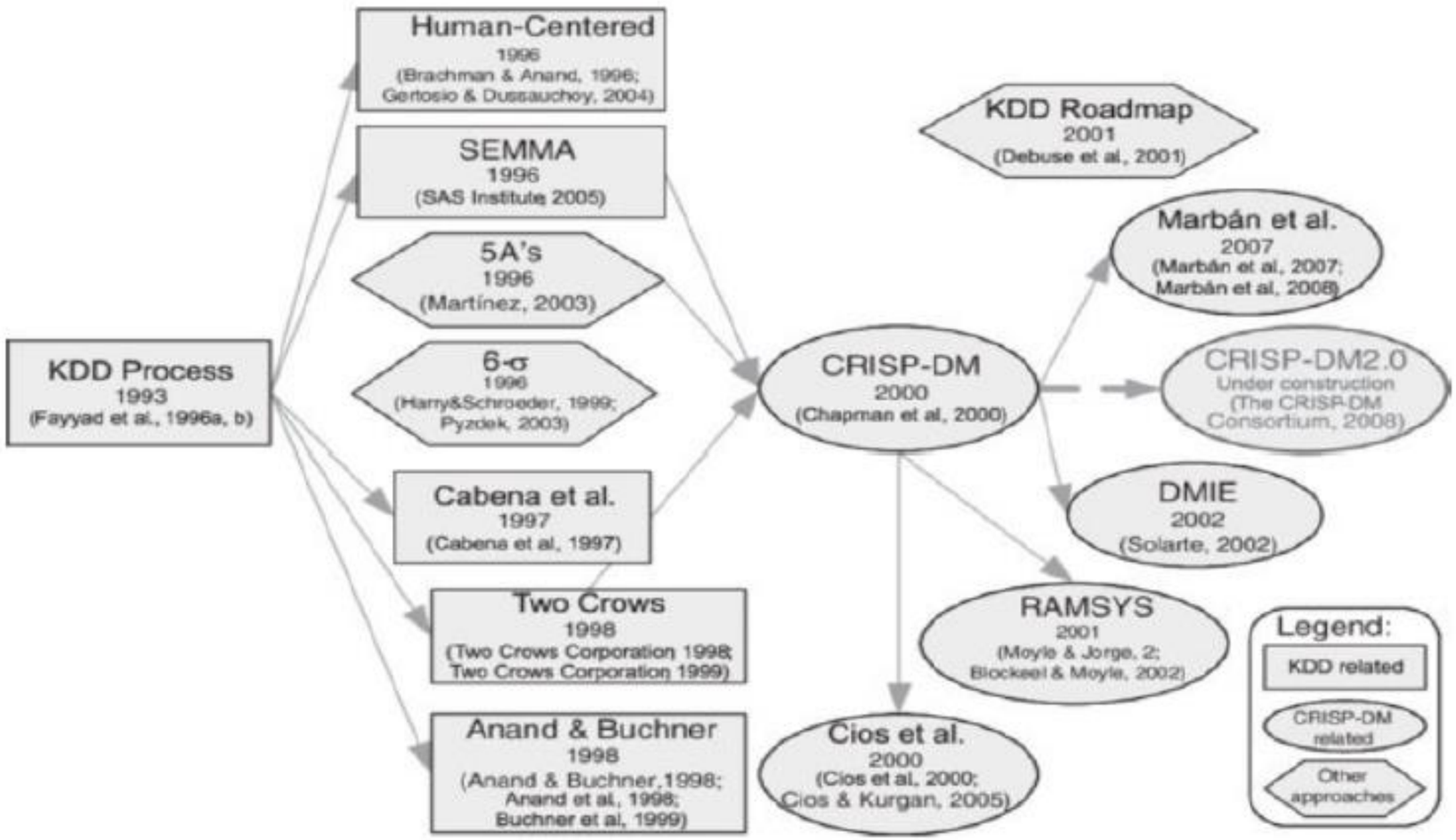
Recall KDD


Selection
Pre-
processing
Trans-
formation
Data
Mining
Interpretation/
Evaluation
Data
Target
Data
Preprocessed
Data
Transformed
Data
Patterns
Knowledge

The problem with KDD

- Lack of the business perspective
- No deployment stage
- Difficult to adapt with requirement changes
- Source is from data only, no explicit needs for data warehousing or data marts

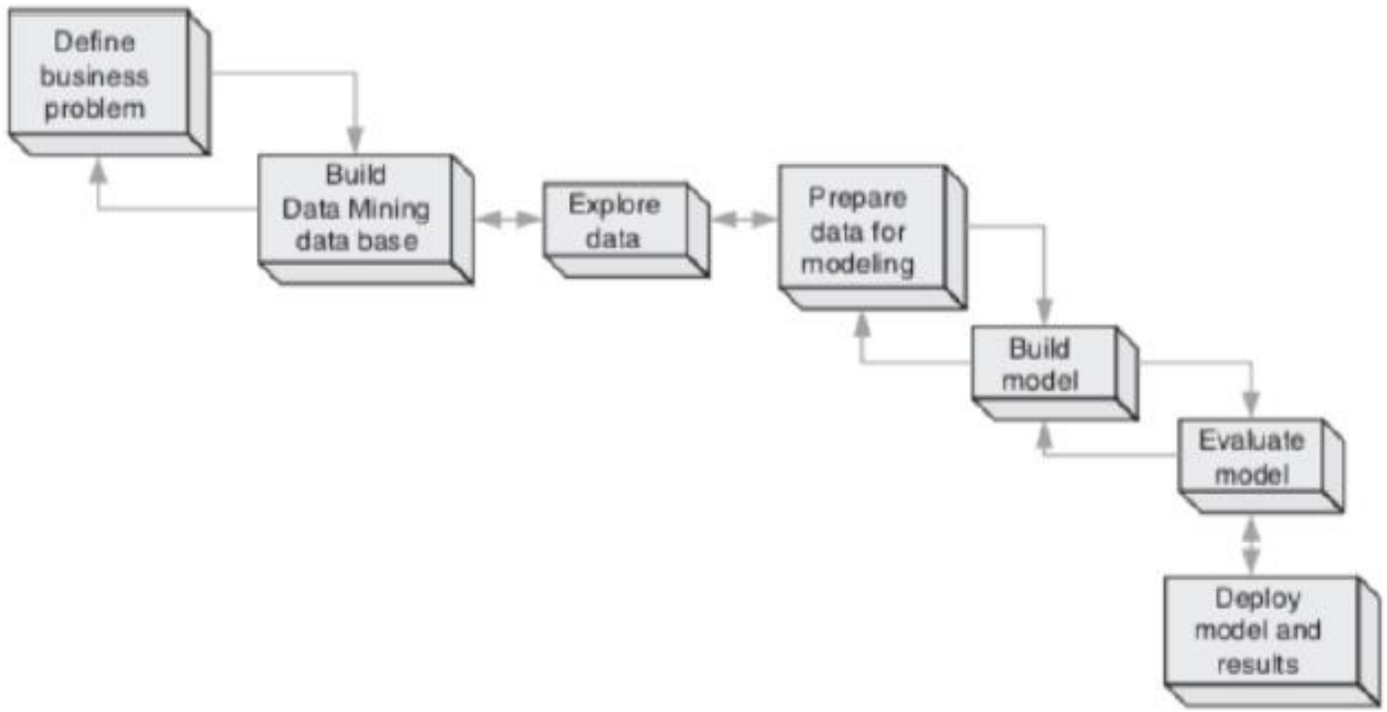
Before Data Science



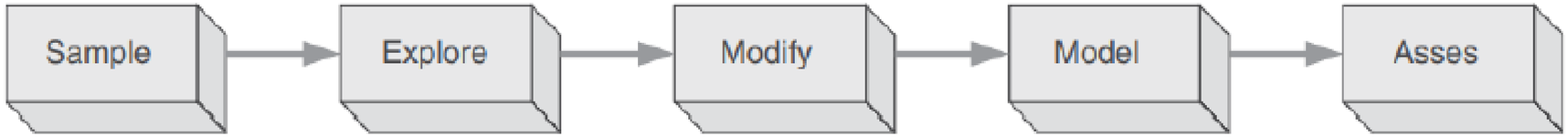
Mariscal, G., Marbán, Ó., & Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, *25*(2), 137–166. https://doi.org/10.1017/S0269888910000032

Two crows

- The very first non-linear process model



SEMMA



- Defined by SAS enterprise miner (integrated to SAS)
- Focused on the model development aspects of data mining
- SEMMA skips the first step of KDD (learning the application domain)
- SEMMA skips the discover knowledge step

CRoss Industry Standard Process for Data Mining (CRISP-DM)

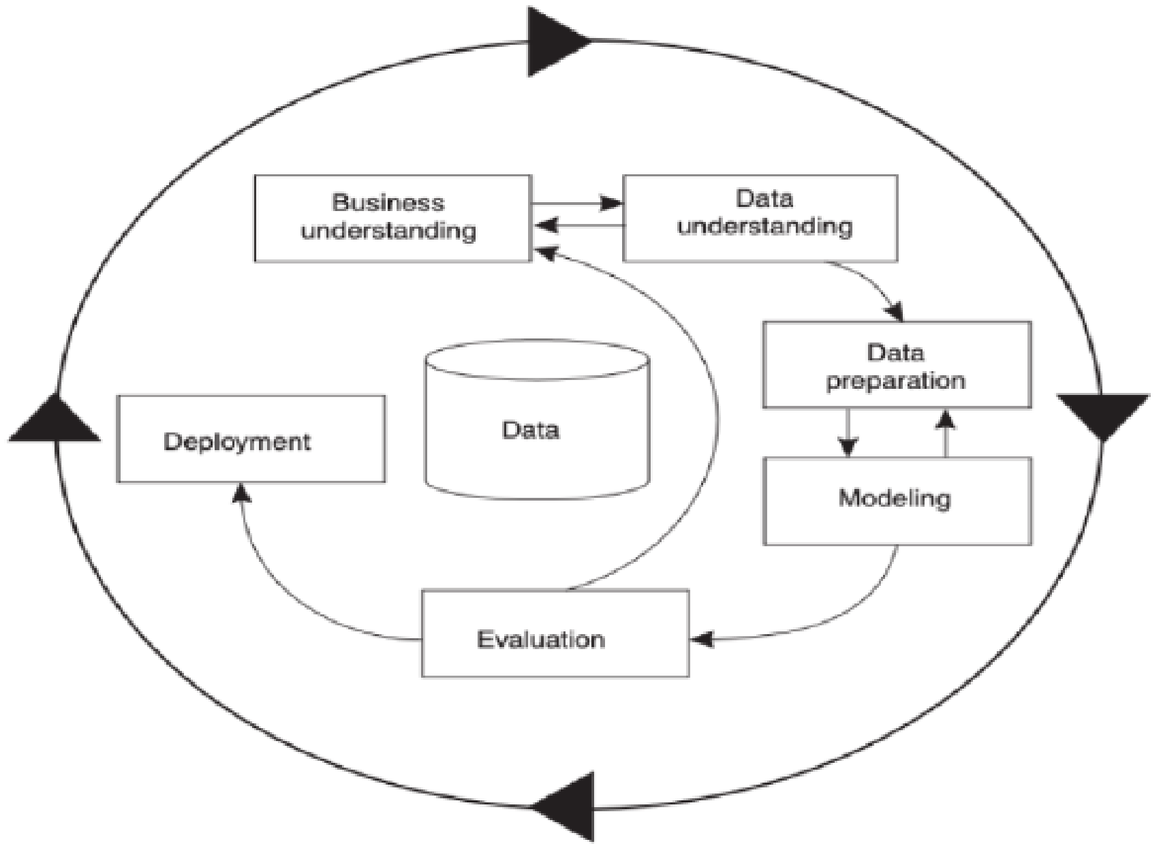
- Most commonly used methodology by data miners and data scientists. (KdNuggets.Com, 2007b)
- Released back in 1999 by SPSS, Teradata, Daimler-Chryler and OHRA
- Ensure quality of DM projects results
- Robust
- Capture experience for reuse
- Up to now, CRISP-DM has been considered the **de-facto**

standard for developing data mining and knowledge discovery projects

CRoss Industry Standard Process for Data Mining (CRISP-DM)

- Business Understanding
  - Define the problem
- Data Understanding
  - Collect the data and determine the underlying characteristics of the data
- Data Preparation
  - Clean the data, feature selection and dimension reduction
- Modeling
  - Implement the data mining tool
- Evaluation
  - Verify the result with the business objectives
- Deployment

CRoss Industry Standard Process for Data Mining (CRISP-DM)



Comparison between academic paper vs CRISP-DM

1. Set goals for the Big Data project	1. Business Understanding
2. Set the data and data sources	2. Data Understanding
3. Check if the available data can meet the objectives of the project and establish how you will meet the objectives	2. Data Understanding
4. The Data is transformed, in the cases that is necessary for the Big Data process	3. Data Preparation
5. Execute the algorithms that satisfies the project objectives	4. Modeling
6. The results are presented, analyzed and disseminated.	5. Evaluation and 6 .Deployment
7. Depending on the results, strategic decisions are taken to follow.	5. Evaluation and 6 .Deployment

CRISP-DM in detail

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion/ Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

CRIPS-DM vs KDD-related approaches

Methodology	Phases					
CRISP-DM	Business understanding	Data understanding	Data preparation	Modeling	Evaluation	Deployment
Human-Centered	Task discovery Data discovery	Task discovery Data discovery Data cleaning	Data cleaning Model development	Model development Data analysis	Data analysis	Output generation
SEMMA		Sample Explore	Explore Modify	Model Assess	Assess	
Cabena et al.	Select	Preprocess	Pre-process Transform	Mine	Analyse & assimilate	Analyse & assimilate
Two Crows	Define business problem	Build DM data base Explore data	Explore data for modeling Prepare data	Build model	Evaluate model	Deploy model and results
Anand & Buchner	Domain knowledge elicitation Human resource identification Problem specification	Domain knowledge elicitation Data prospecting	Methodology identification Data pre-processing	Methodology identification Pattern discovery	Knowledge post-processing	

Process in action

The Census data on CRISP-DM

Dataset

- **Census Income Data Set**
- https://archive.ics.uci.edu/ml/datasets/census+income

CRISP-DM : Business Understanding

- Goal
  - Understand the objectives and requirement
  - Define the data mining problem
- Tasks
  - Determine the business objectives
  - Assess the situation
  - Determine data mining goals
  - Produce project plan

CRISP-DM : Business Understanding (The census dataset)

CRISP-DM : Business Understanding

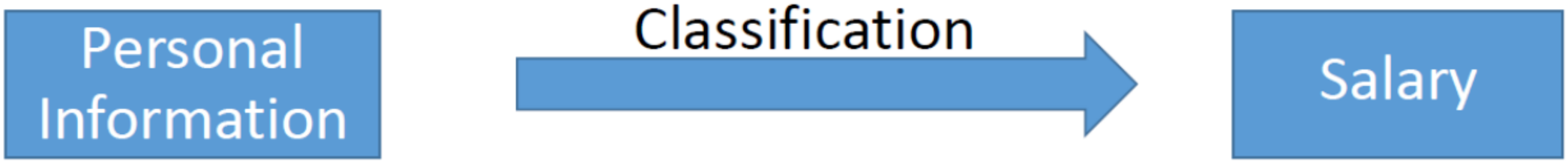
- Support the marketing of a new service targeted at potential customers with medium to high level salaries

- Setup cost: $18,000
- Cost per offer: $125
- Return per accepted offer: $500

- Currently, has achieved a high acceptance rate in the past of seventy five percent by individuals whose salaries exceed fifty thousand US dollars.

The data mining problem

Need a classification model to predict individuals whose salary exceeds $50 by mining by taking demographic information such as age, gender, education level and employment type.



Project plan

Project plan

Phase	*Date*	*Details*	*Status*
A	20/2/2011	Start date	*Closed*
B	28/2/2011	Project proposal compete	*Closed*
C	5/3/2011	*Crisp 1*: Business understanding	*Closed*
D	12/3/2011	*Crisp 2*: Initial data understanding	*Closed*
E	20/3/2011	*Crisp 3*: Initial data preparation (and testing with early modelling)	*Closed*
F	30/03/2011	Interim presentation based on phases A to D	*Closed*
		Crisp 4: Creation/test of models (update results in D and E as required)	*Closed*
G	06/04/2011	*Crisp 4*: Ongoing testing of data preparation (update results in D to F as required)	*Closed*
H	20/04/2011	*Crisp 5*: Evaluation of proposed models	*Closed*
I	1/5/2011	Final presentation of work	*Closed*
		Create project report	*Closed*
J	16/5/2011	Completion date	

CRISP-DM : Data Understanding

- Goal
  - Data collection and data understanding (initially)
  - Evaluation the data quality
- Tasks
  - Collect initial data and visualize the data
  - Verify data quality
  - Collect the meaning of each attribute
  - Identify missing data and its interpretation
  - Formulating data mining problem requires some understanding of the data.

Example

- The target output is >50k or <= 50k

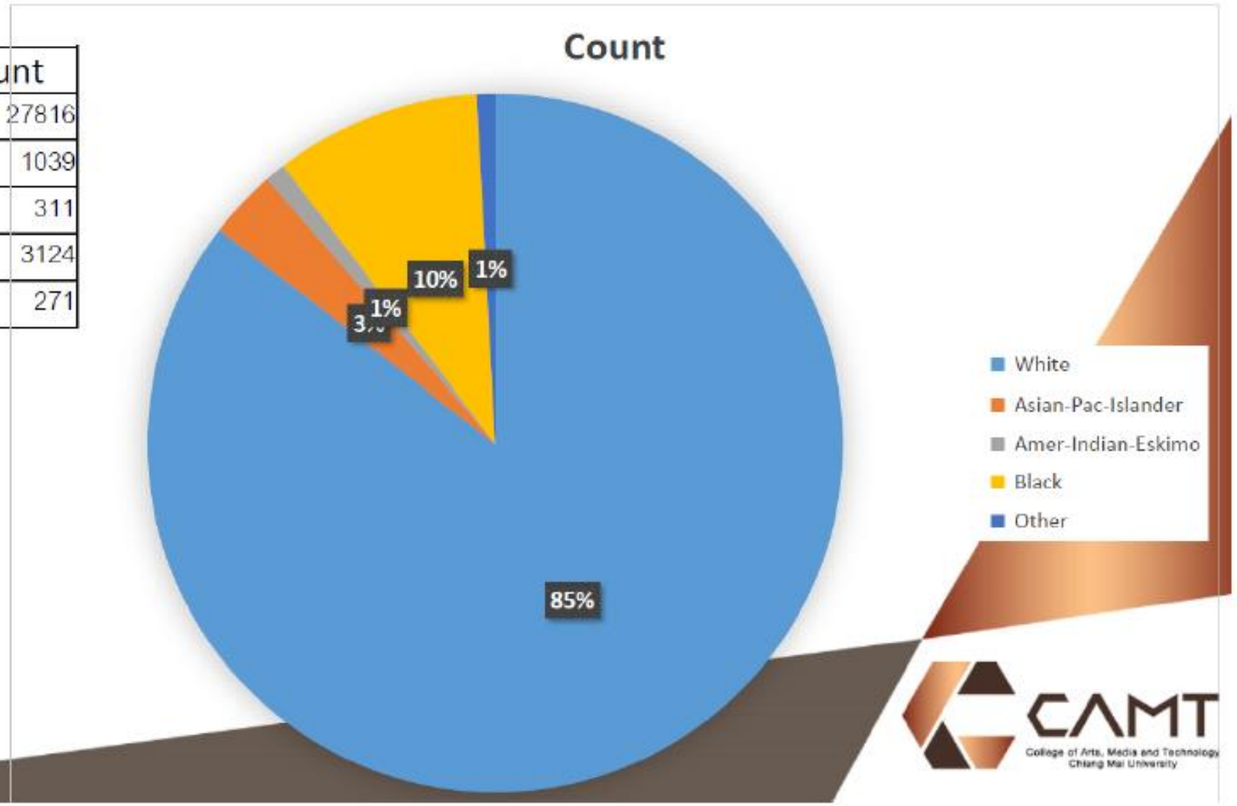
Example

Workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

- **Education**: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- **Marital-status**: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
- **Occupation**: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
- **Relationship**: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
- **Race**: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
- **Sex**: Female, Male
- **Native-country**: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

Describe

Race	Count
White	27816
Asian-Pac-Islander	1039
Amer-Indian-Eskimo	311
Black	3124
Other	271

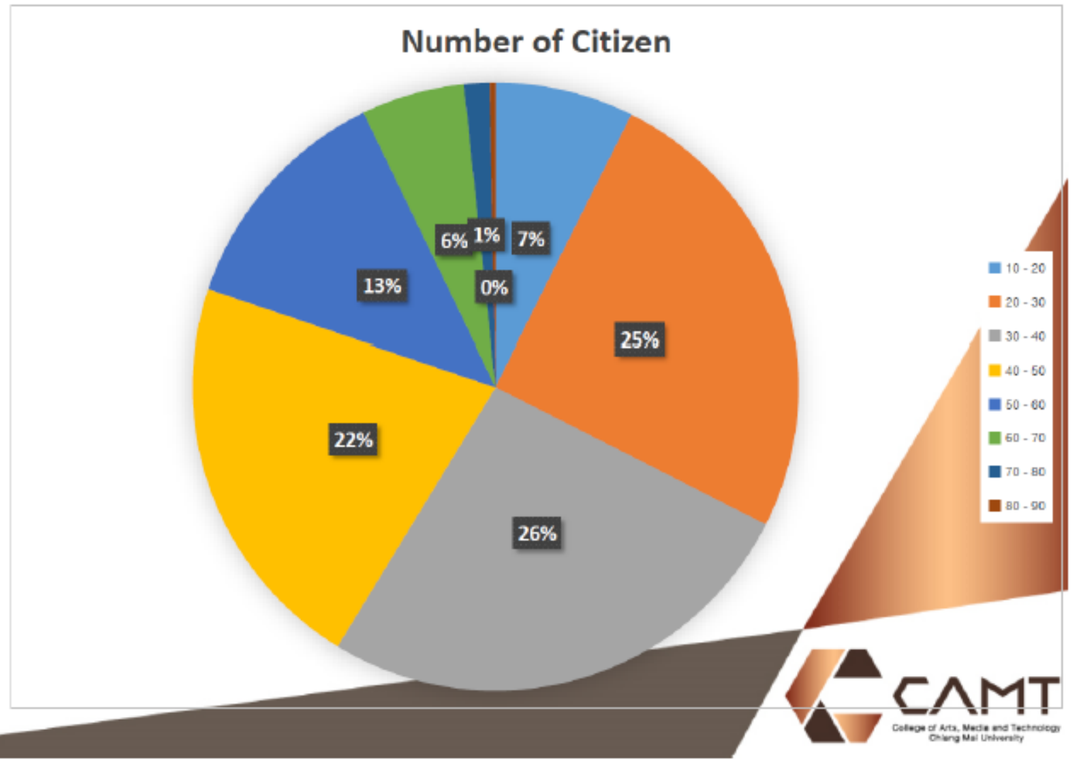


Describe (cont.)

Example

Age Range	Number of Citizen
10 - 20	2410
20 - 30	8162
30 - 40	8546
40 - 50	6983
50 - 60	4128
60 - 70	1792
70 - 80	441
80 - 90	99

Minimum age	17
Maximum age	90
Average age	38.58165

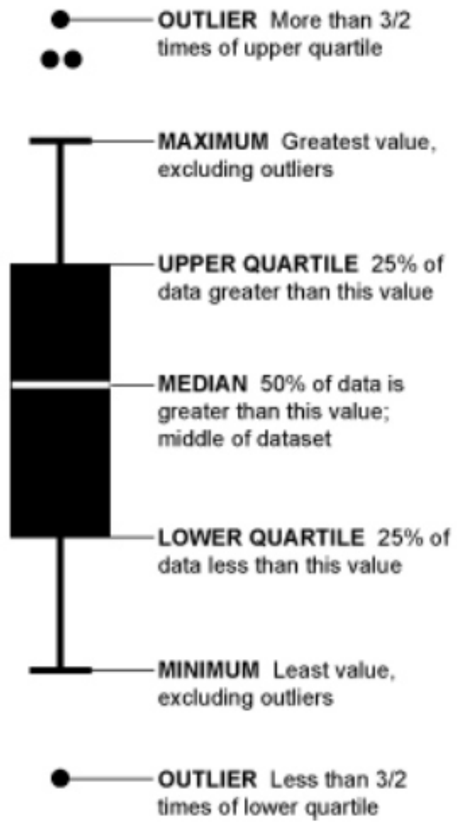


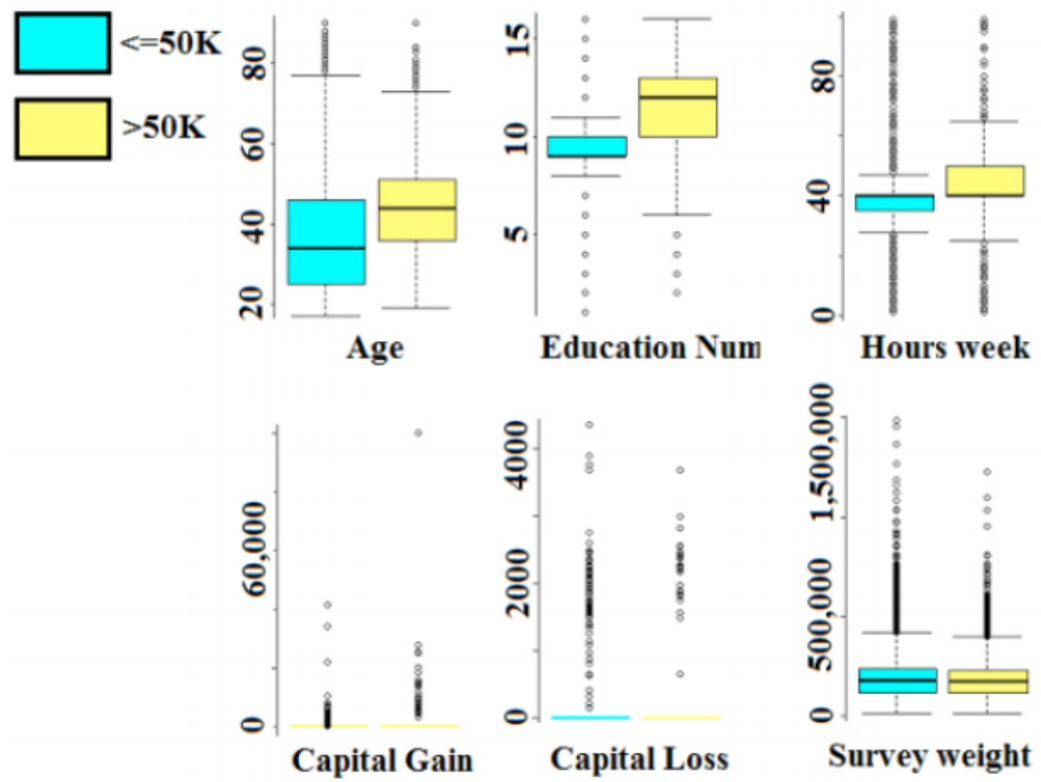
Explore (cont.)

	Attribute	Values	Missing
Polynomials	Employment Class	Private (68%), Self employed 1 (8%), Local Gov(6%), State Gov(4%), Unknown (5%), Self employed 2 (3%), Federal Gov(3%), No Pay(1.5%), Never Worked (0.5%)	1836
	Education Level	High School (32%), Some college (22%), Bachelors (16%), Masters (5%), Vocational (4%), 11th (4%), Assoc Academic (3%), 10th (3%), 7-8th (2%), Professional School (2%), 9th (2%), 12th (2%), Doctorate (1%), 5-6th (1%), 1-4th (1%), Preschool (1%)	0
	Relationship	Husband (41%), Not-in-family (26%), Own child (16%), Unmarried (11%), Wife (4%), Other relative (2%)	0
	Race	White (85%), Black (10%), Asian / Pacific Islander (3%), American Indian / Eskimo (1%), Other (1%)	0
	Marital Status	Married-civ-spouse (46%), Never-married (33%), Divorced (14%) Separated (3%), Widowed (2%), Married-AF-spouse (1%), Married-spouse-absent (1%)	0
	Occupation	15 categories	1843
	Country	42 categories: USA (90%)	583
Binomials	Salary [Label]	<=$50K (76%), >$50K (24%)	0
	Gender	Male (67%), Female (33%)	0

Real	Mean	Median	Std Dev	Skewness	Kurtosis	Range	
Age	38.58	37	13.64	0.56	2.83	17 - 90	0
Hours worked per week	40.44	40	12.35	0.23	5.92	1 - 99	0
Education Number	10.08	10	2.57	-0.31	3.62	1 - 16	0
Captial Gain	1078	0	7385	11.95	157.77	0 - 99999	0
Capital Loss	87.3	0	403	4.59	23.37	0 - 4356	0
Survey Weight	189778	178356	105550	1.45	9.22	12285 - 1484705	0

Explore (cont.)





Verify data quality

- Look for missing values %
- Look for outlier/extreme values
- Look for inconsistencies (e.g., male gender with a relationship value of wife or unmarried marital status with a relationship value of husband/wife etc.)
- Look for duplication
- Look for imbalance (should not be more than 1:3)

CRISP-DM : Data Pre-preparation

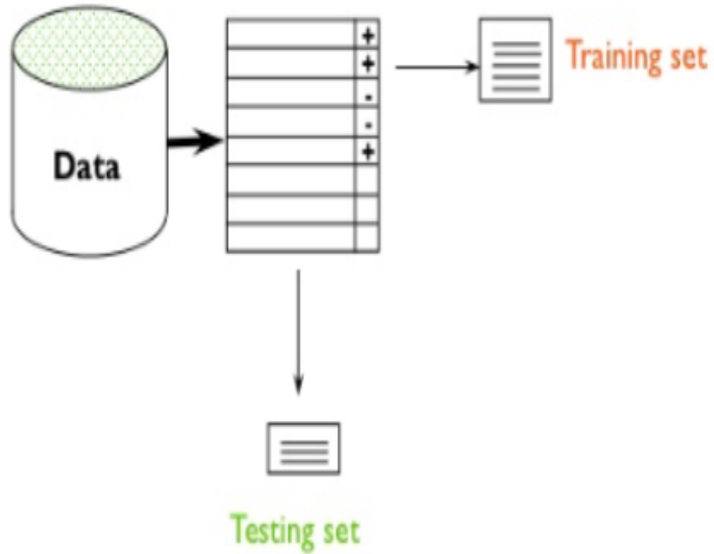
- Data Cleaning
- Missing value
- Outliers/ Extreme values
- Discretization/ Transform
- Normalization
- **Decide which portion of data that you have is actually going to used for data mining / machine learning**
- **Sampling techniques**

Training and Testing Dilemma

- What we usually expect
  - A large training data set
  - A large testing data set
- More often, we don’t have enough quality and quantity of data when doing analysis.

Example: Hold-out method

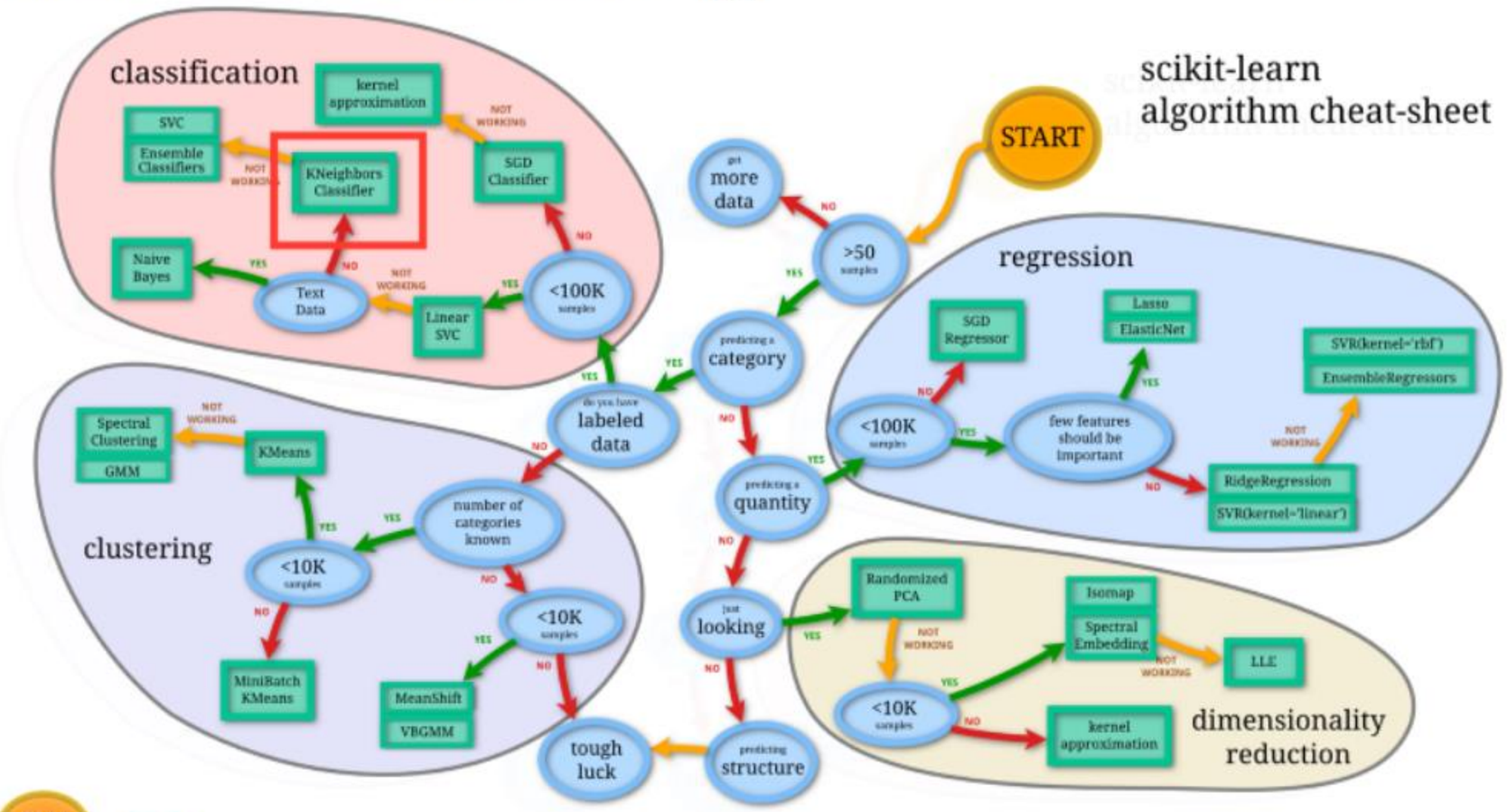
- Good approach for a large data set, if we have more than 1,000 samples, including several hundred instances from each class.
- Split data into training data and testing data
- 80% for train, 20% for test or
- Build classifier using the train data
- And test with the test data



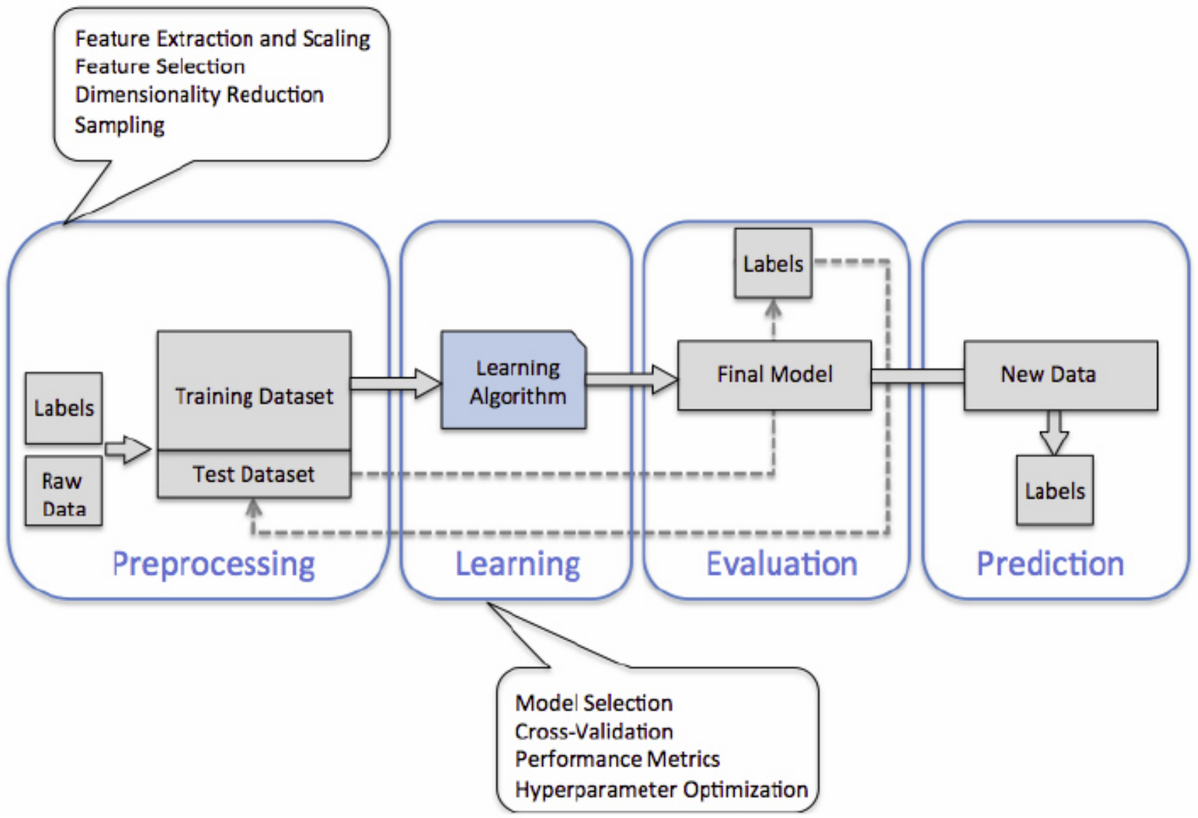
CRISP-DM : Modeling

- Goal
  - Construct the optimal model
- Tasks
  - Selecting data mining/machine learning techniques
  - Generate test design
  - Optimize the model

CRISP-DM : Modeling



Road map to Data Mining/ Machine learning



CRIPS-DM Evaluation

- Goal

  Evaluation the result model if it meets with the objectives (not only the accuracy)

- Tasks
  - Evaluate the result
  - Interpret the result
  - Review the process

Metric

- Metric : Nomenclature

True Positives (TP) -number of predicted positive in the actual positive group

True Negatives (TN) -number of predicted negative in the actual negative group

False Positives (FP) -number of predicted positive in the actual negative group

False Negatives (FN) -number of predicted negativein the actual positive group

Accuracy rate

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Confusion matrix

n = 165	Predicted: No	Predicted: Yes
Actual: No	50	10
Actual: Yes	5	100

n = 165	Predicted: No	Predicted: Yes	
Actual: No	Tn =50	FP=10	60
Actual: Yes	Fn=5	Tp=100	105
	55	110	

Confusion matrix

	Predicted Positive	Predicted Negative
Actual Positive	10 (TP)	15 (FN)
Actual Negative	25 (FP)	100 (TN)

$$\text{accuracy} = (10 + 100) / (10 + 100 + 25 + 15) = 73.3\%$$

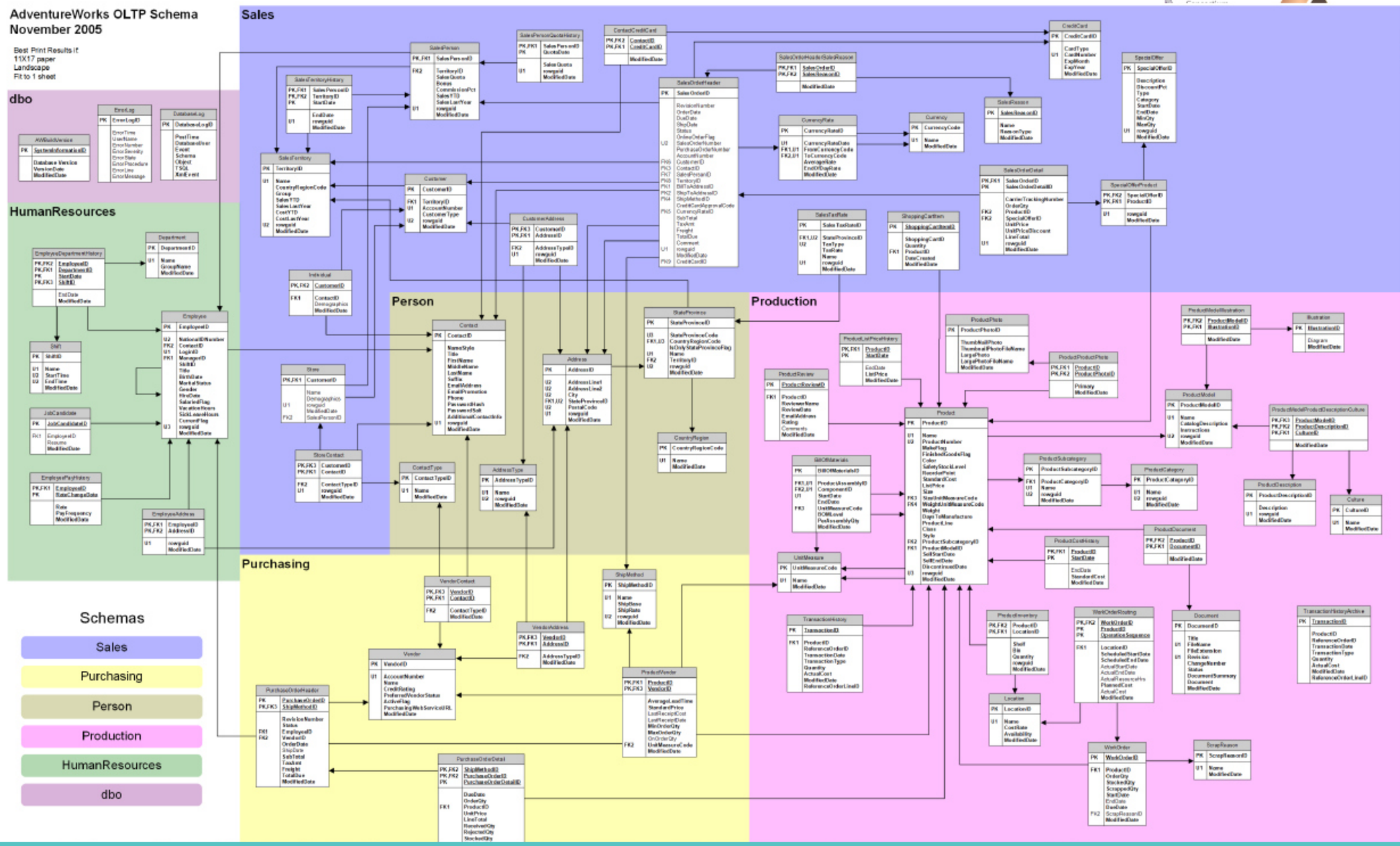
CRISP-DM - deployment

- End point of the project life cycle
  - Plan deployment
  - Plan monitoring and maintenance
  - Generation of final report
  - Review of the process sub-steps.

CRIPS DM in Action 2

Dataset

- AdventureWorks sample databases
- https://docs.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-ver15&tabs=ssms


AdventureWorks OLTP Schema
November 2005
Best Print Results if
11X17 paper
Landscape
Fit to 1 sheet
dbo
HumanResources
Sales
Person
Purchasing
Production
Schemas
Sales
Purchasing
Person
Production
HumanResources
dbo

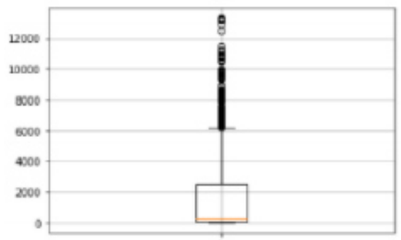
Business understanding

- Business operates by production bikes and export to all reseller around the world and online channel to distribute our product
- Aim to Maximize our profit by gain newer customer to by our product.

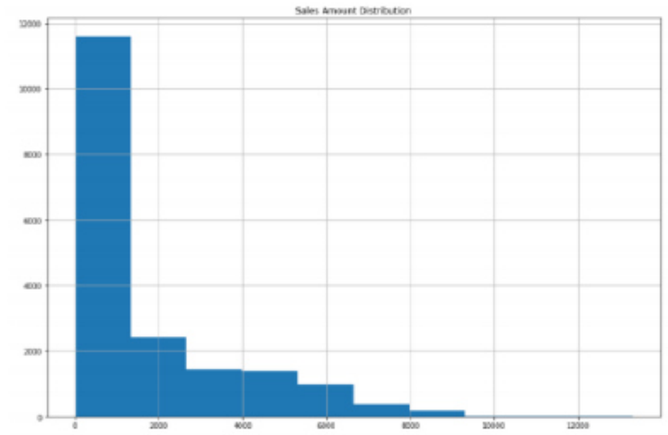
Phase	Timing	Details	Status
Business understanding	4 Day	Explore objective. Try to understand data base schema. Understand which customer is our target.	Close
Data understanding	5 Day	EDA: Categorical data group by with target. EDA: Numerical data group by with target. Find correlation between feature and target.	Close
Data preparation	2 Day	Train Data: Remove feature that does not use in model. Remove feature that have same meaning. Create one hot encoding. Test Data or Prospective buyer: Remove feature that does not use in model. Remove feature that have same meaning. Create one hot encoding.	Close
Modeling	1 Day	Using Logistic regression. Using k-Nearest Neighbor. Testing model with cross validation.	Close
Evaluation	1 Day	Compare Model: Using accuracy rate. Using precision rate. If model accuracy rate lower than 70% turn back to phase business understanding.	Close
Deployment	1 Day	Concatenate array that already predict from model with prospective buyer. Export file and send to marketing department.	Close

Data understanding

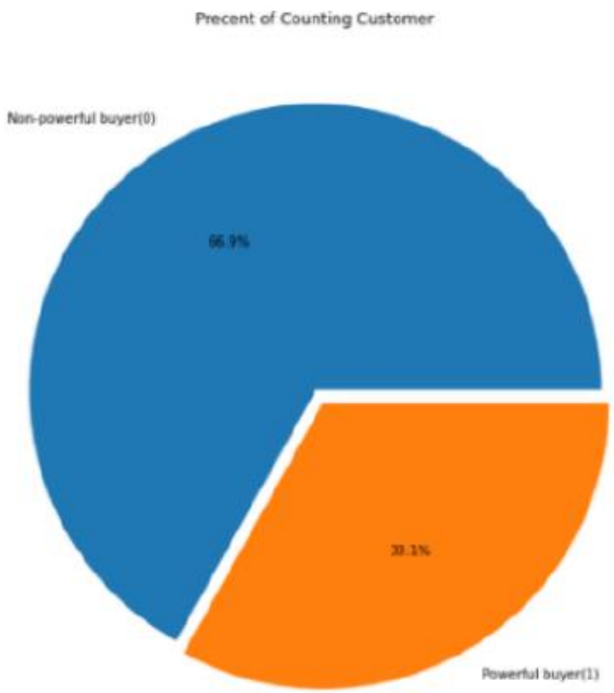
- .Now we visualize distribution on sales amount group by customer

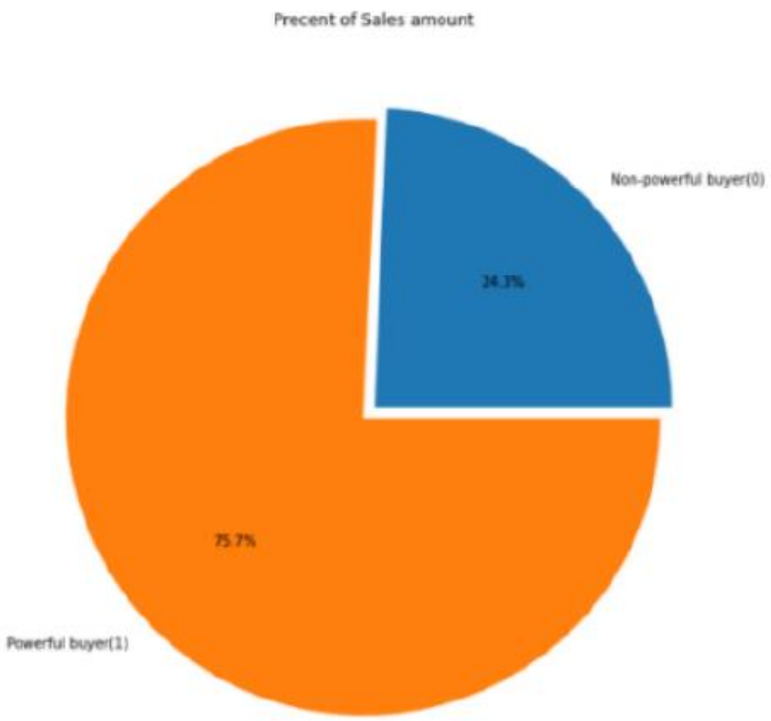


	SalesAmount
count	18484.000000
mean	1588.329216
std	2124.231717
min	2.290000
25%	49.970000
50%	270.265000
75%	2511.275000
max	13295.380000



Data understanding





Data understanding

- Select 6 table from Adventurework2019-Datawarehouse including ‘FactinternetCSale’, ‘DimCustomer’, ‘DimGeogrphy’, ‘Dimproduct’ , ‘DimProductSubCategory’ and ‘DimProductCategory’.
- These tables provide information of customer characteristic.

Data understanding

Features Description

In our dataset, it has 18,484 customer individuals, 24 Features & 1 Class-Label

Categorical Data

- CustomerKey
- MaritalStatus
- Gender
- EnglishEducation
- EnglishOccupation
- HouseOwnerFlag
- CommuteDistance
- EnglishCountryRegionName
- City
- PostalCode
- PromotionKey
- SalesOrderNumber
- EnglishProductCategoryName
- StateProvinceName

Continuous Data

- YearlyIncome
- LastDateBuy
- Age
- SalesAmount

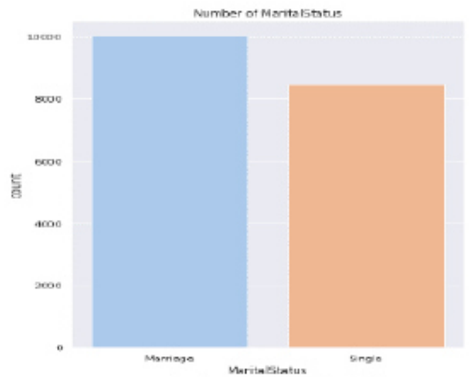
Class Label

- powerful_buyer

Data understanding (discrete)

Marital Status

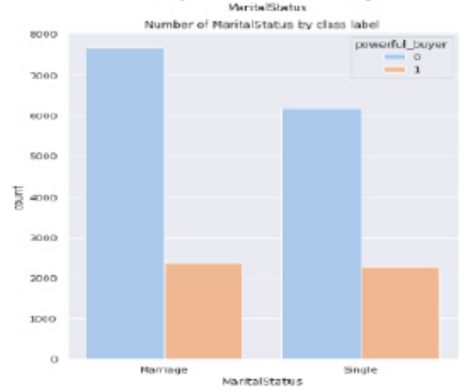
Marital Status = marriage status of each customer



From Number of MaritalStatus Chart we found that there is not much different between 'Marriage' & 'Single'

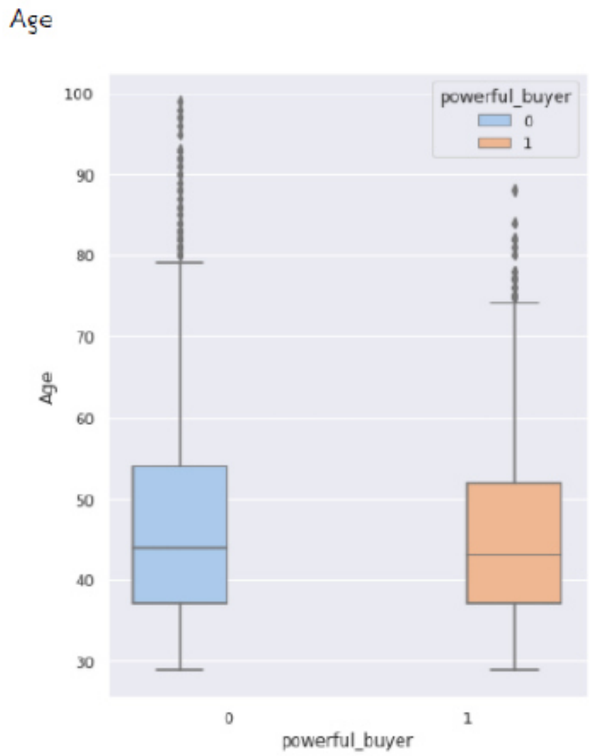
Marriage = 10,011 customers

Single = 8,473 customers



After Separate, the data of MaritalStatus there is not much different between 'Marriage' and 'Single' in Powerful_buyer Class (Class 1) as well.

Data understanding (continuous)



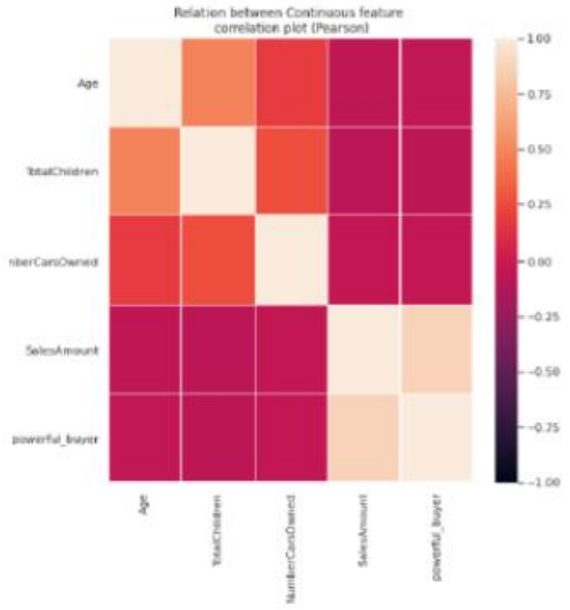
As we seen from the age distribution, we found that we have a widely range of customer age, but most of our customer are in range between 30-45.

Data understanding (continuous)

Relation of features = Finding Relation between each continuous feature.

	Age	TotalChildren	NumberCarsOwned	SalesAmount	powerful_buyer
0	44	3	1	8114.04	1
1	35	0	1	5976.32	1
2	55	3	1	5997.33	1
3	48	1	1	6090.59	1
4	57	2	2	14.98	0
...	...	...	...	...	...
18479	44	1	2	32.27	0
18480	31	4	2	39.98	0
18481	80	2	1	12.98	0
18482	57	3	2	69.99	0
18483	32	0	1	24.99	0

18484 rows × 5 columns



Data understanding (discrete)

- Chi test can be use to find correlation for categoriacal data and target

Data preparation

- Missing value
- Drop features/column
- Normalization
- Minmax
- One-hot

Data preparation

Clean Test data set

Clean Incontinence Value

In data test prospective buyer, it has incontinence in feature education in tuple:

High Schoo = High School

Partial Hi = Partial High School

Partial Co = Partial College

Graduate D = Graduate Degree

```
0                 Bachelors
1                 Bachelors
2           Graduate Degree
3               High School
4                 Bachelors
                 ...
2659    Partial High School
2660        Partial College
2661            High School
2662              Bachelors
2663        Partial College
Name: Education, Length: 2664, dtype: object
```

Modeling- LR

Logistic Regression.

Logistic Regression with 10 cross validation

	Model	AccuracyScore
0	LogisticRegression(C=1.0, class_weight=None, d...	0.753478
1	LogisticRegression(C=1.0, class_weight=None, d...	0.756569
2	LogisticRegression(C=1.0, class_weight=None, d...	0.761978
3	LogisticRegression(C=1.0, class_weight=None, d...	0.766615
4	LogisticRegression(C=1.0, class_weight=None, d...	0.741113
5	LogisticRegression(C=1.0, class_weight=None, d...	0.763524
6	LogisticRegression(C=1.0, class_weight=None, d...	0.767388
7	LogisticRegression(C=1.0, class_weight=None, d...	0.756569
8	LogisticRegression(C=1.0, class_weight=None, d...	0.754834
9	LogisticRegression(C=1.0, class_weight=None, d...	0.752514

Modeling - LR

LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, l1_ratio=None, max_iter=10000, multi_class='auto', n_jobs=None, penalty='l2', random_state=None, solver='lbfgs', tol=0.0001, verbose=0, warm_start=False)

The result from evaluation in 30% testing data is.

classification_report

	precision	recall	f1-score	support
0	0.77	0.97	0.86	4128
1	0.60	0.15	0.24	1418
accuracy			0.76	5546
macro avg	0.68	0.56	0.55	5546
weighted avg	0.72	0.76	0.70	5546

Modeling - KNN

KNN.

9	0.76596	KNeighborsClassifier(algorithm='auto', leaf_si...	10
10	0.76217	KNeighborsClassifier(algorithm='auto', leaf_si...	11
11	0.76611	KNeighborsClassifier(algorithm='auto', leaf_si...	12
12	0.76457	KNeighborsClassifier(algorithm='auto', leaf_si...	13
13	0.76411	KNeighborsClassifier(algorithm='auto', leaf_si...	14
14	0.76356	KNeighborsClassifier(algorithm='auto', leaf_si...	15
15	0.76612	KNeighborsClassifier(algorithm='auto', leaf_si...	16
16	0.76519	KNeighborsClassifier(algorithm='auto', leaf_si...	17
17	0.76488	KNeighborsClassifier(algorithm='auto', leaf_si...	18
18	0.76295	KNeighborsClassifier(algorithm='auto', leaf_si...	19

Modeling - KNN

In cross validation we use CV = 10 and group by fold, use mean to accuracy score so the best model that we got is k = 16, then test with 30%.

Modeling

classification_report

	precision	recall	f1-score	support
0	0.78	0.96	0.86	4128
1	0.66	0.22	0.33	1418
accuracy			0.77	5546
macro avg	0.72	0.59	0.60	5546
weighted avg	0.75	0.77	0.73	5546

conclusion

We select KNN model with becasue it has high precicion more than logistic model, so we select KNN model.

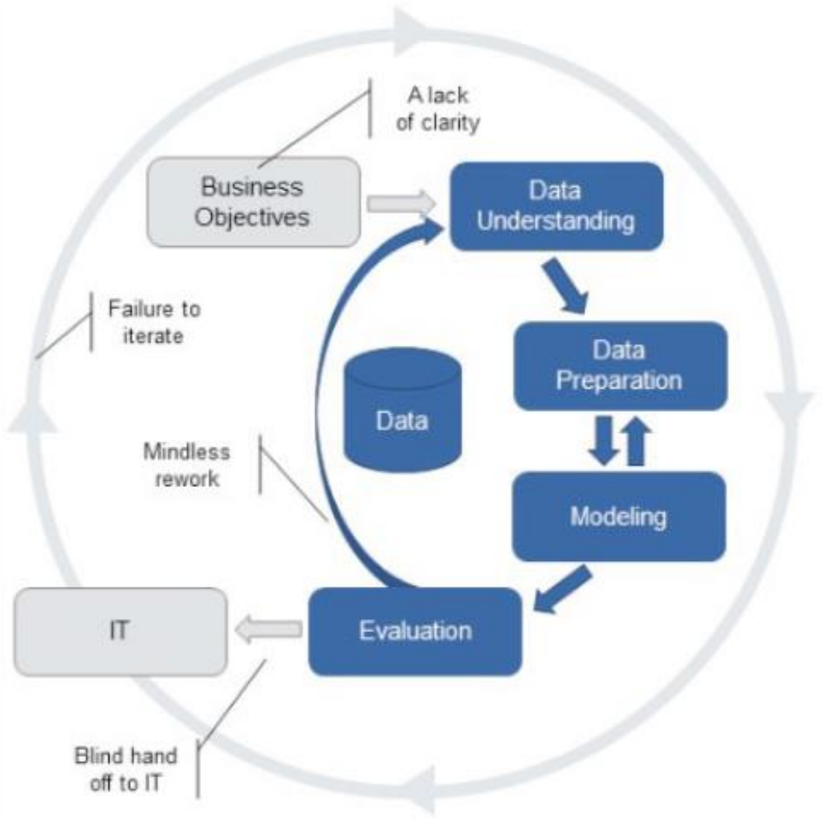
Deployment

- Generated Model report that can predict power buyer to the marketing department.

On continual to CRISP-DM

Furthermore the CEO from Decision Management Solutions
criticized the model upon the four problems:

- Skip the business analysis
- Bring more data instead of carefully developing the business understanding
- Technology blindness (discussed earlier)
- Fail to maintain and manage
- Deployment stage is a dead point (final report)
- Knowledge discovered is not stored anywhere and not used in future investigation
- Source is from data only, no explicit needs for data warehousing or data marts



On continual to CRISP-DM

- In the field of data science, there is no “best” process to do a data science project.
- Have a well-defined repeatable process can help data science teams for challenges such as:
  - Stakeholder in the process
  - Appropriate data architecture/infrastructure
  - Determining the appropriate analytical techniques and validation

Stakeholders

- Data science spend 80-90% of their times internally focused.
- Problem identify and managing key stakeholders.
- Stakeholder
  - A person, group, or organization that is actively involved in a project is affected by its outcome or can influence its outcome.

Stakeholder concerns

- What do they need to know about the project?
  - Most commons are:
    - What input will I be required to provide to the project team?
    - How can I make my needs known?
    - When will the project be done?
    - How will it affect me?
    - Will I be replaced?
    - How will I learn how to use the deliverables?

Stakeholder categories

- Those who are affected by the project and who will use its artifacts, including the results.
  - Customers
  - Heads and employees of functional units
  - End users
- Those who are not involved in the project, but because of their position or activities can influence it.
  - Top-manager of the company
  - Owners of the company
  - Shareholders and creditors
- Those who are involved in the project and work on it
  - Project team (e.g., developers, business analyst)
  - Management team (e.g., project manager)
  - Third-party companies.

References:

- https://www.linkedin.com/pulse/data-science-like-team-sport-you-need-strategy-execution-m%C3%A4%C3%A4tt%C3%A4/
- Architectural thinking in the Wild West of data science. (2018, December 5). Retrieved September 21, 2019, from IBM Developer website: https://developer.ibm.com/articles/architectural-thinking-in-the-wild-west-of-data-science/

3.3 บทที่ 3 :
Introduction to Python Programming

กระบวนการทางข้อมูล

Data Process

ผู้ช่วยศาสตราจารย์ ดร. พรัอมเพงศ์ สุภัณศีล

Introduction to Python

- Python is a general purpose programming language.
- Python provides a lot of useful library for data science.

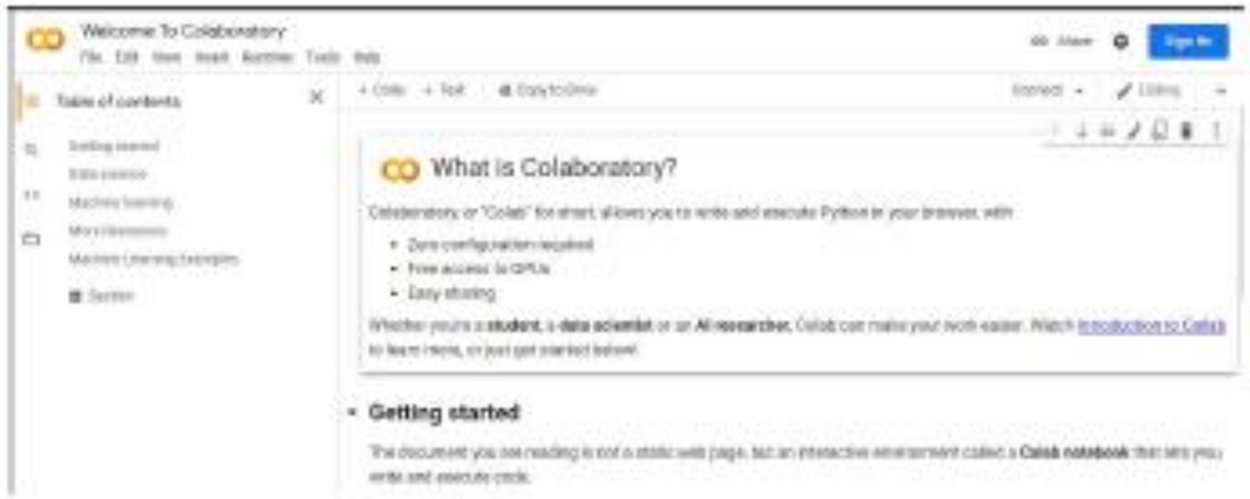
Google Colab

- Free Jupyter notebook
- Runs on cloud
  - Does not need to install any program.
- Allows collaboration

Google Colab

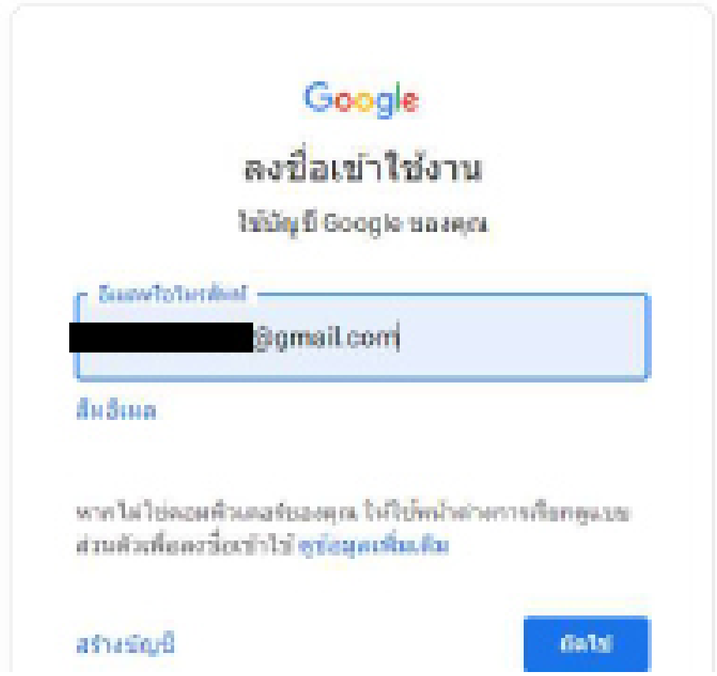
- Go to website

https://colab.research.google.com/



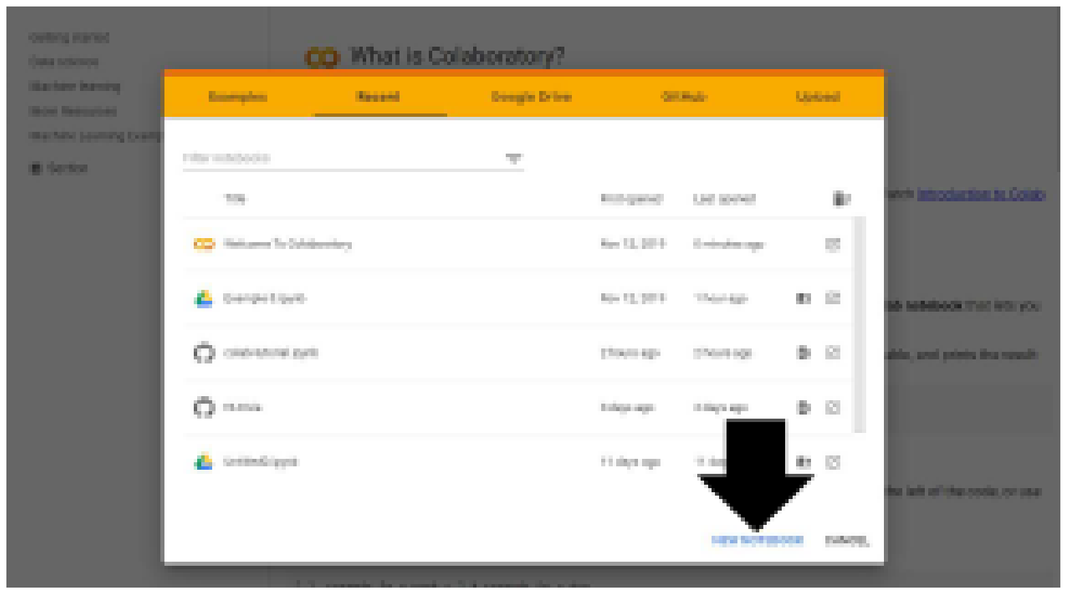
Google Colab

- Log in to the Colab using your gmail account



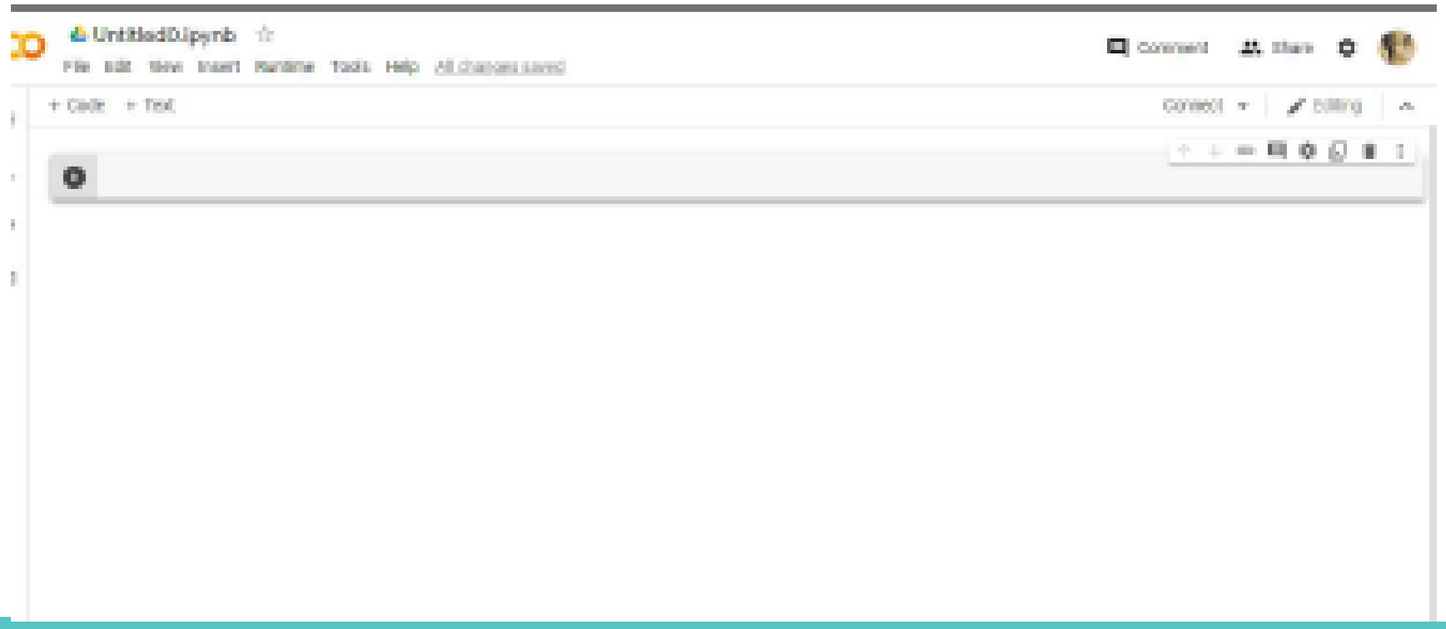
Google Colab

- List of project will appear.
- Choose “New Notebook”



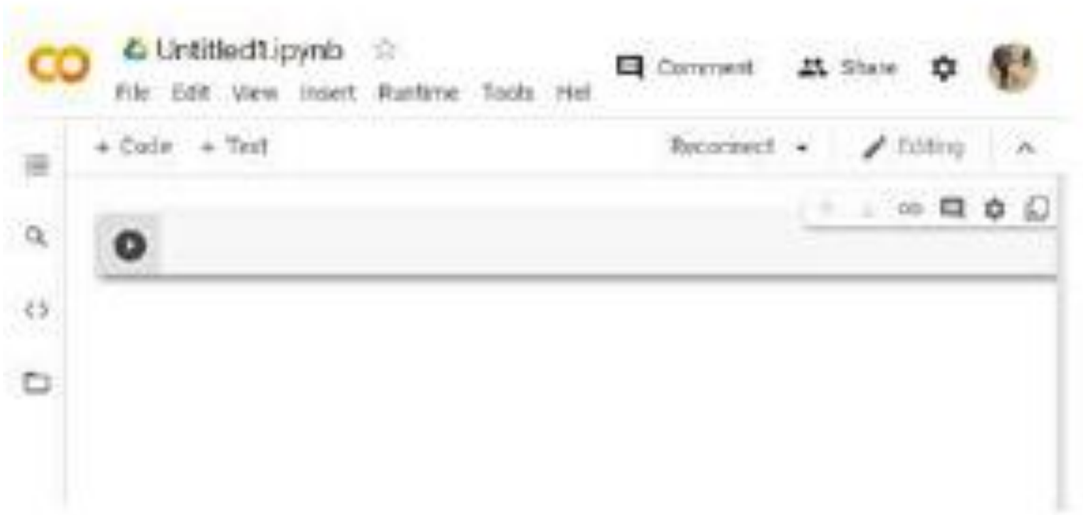
Google Colab

- The coding panel will appear.



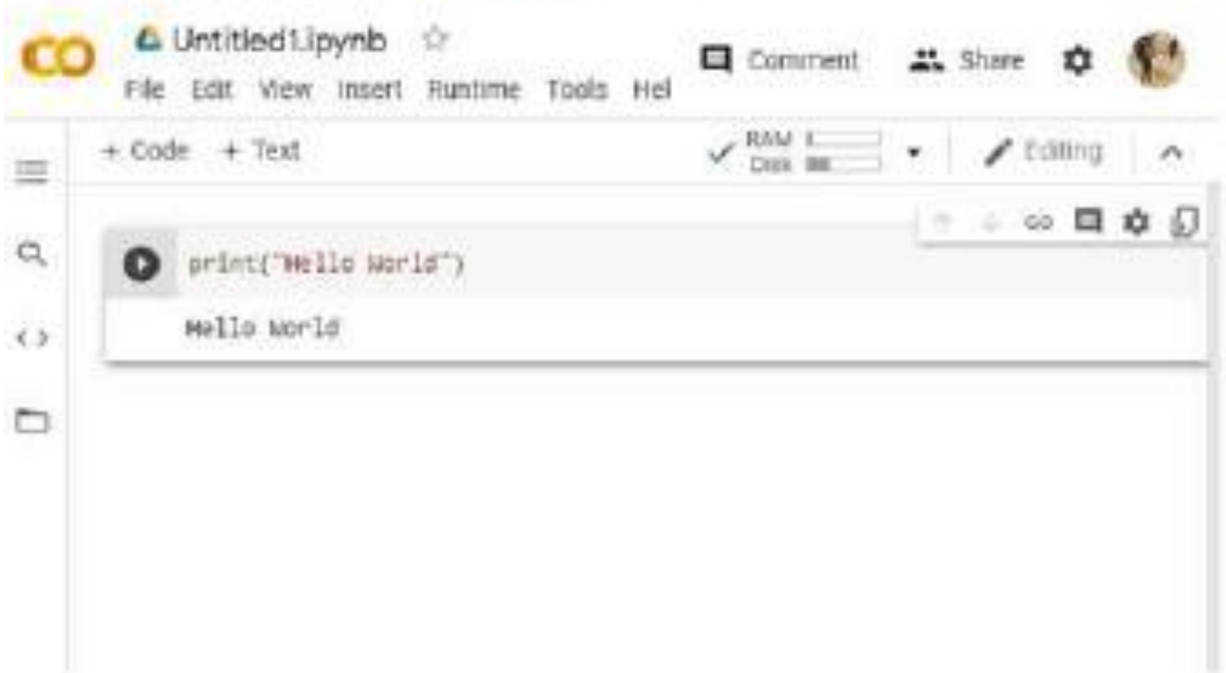
Google Colab

- Your first program “Hello World”



Google Colab

- Arithmetic Operation

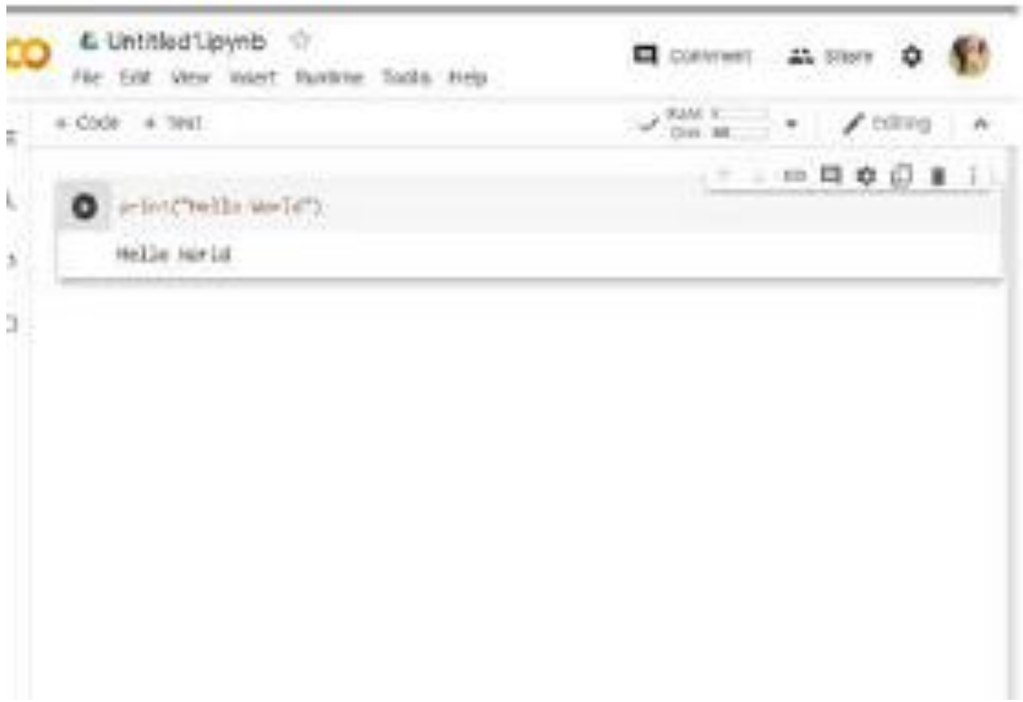


Google Colab

- Find the result of the following operation
- 1/2 = ?
- 2*5.5 =?
- 1 + “Hello” =?

Google Colab

- Answer

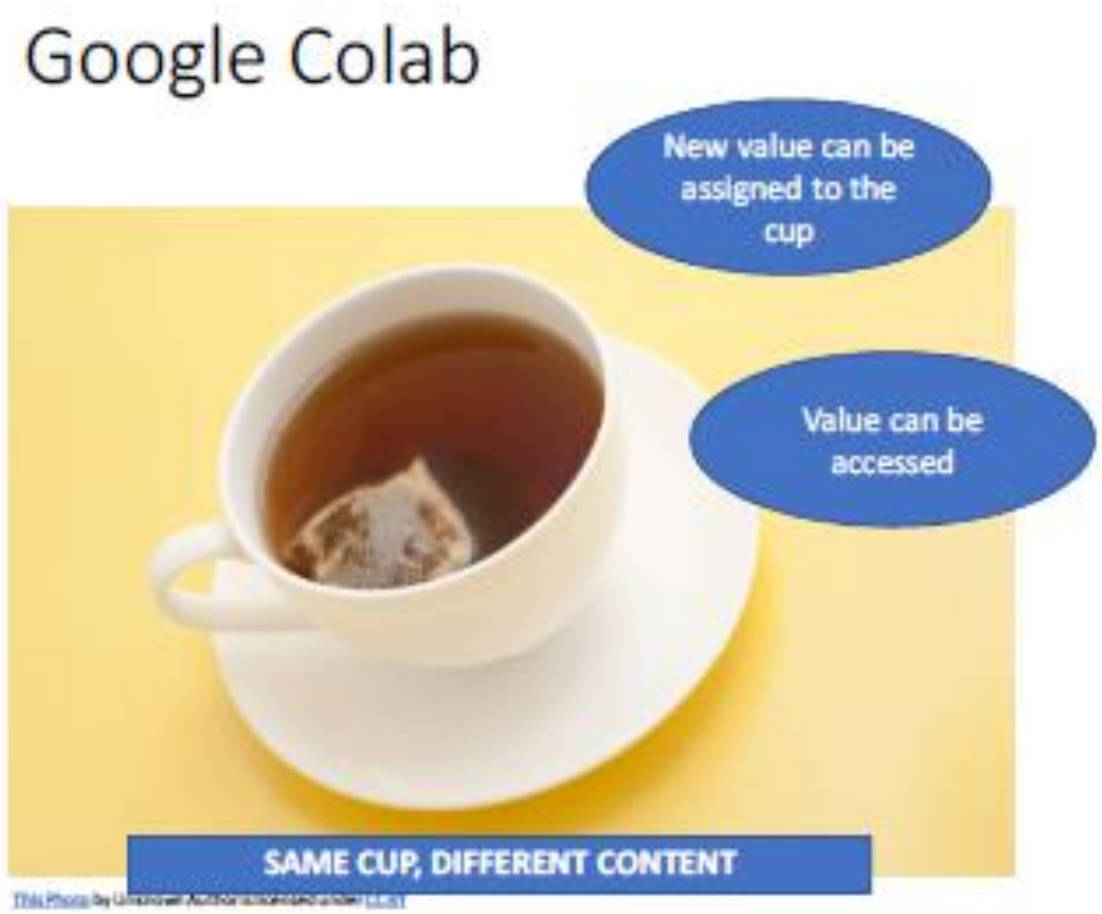


Google Colab

- We need a data container.
- Similar to the cup.
  - If the water is an data.

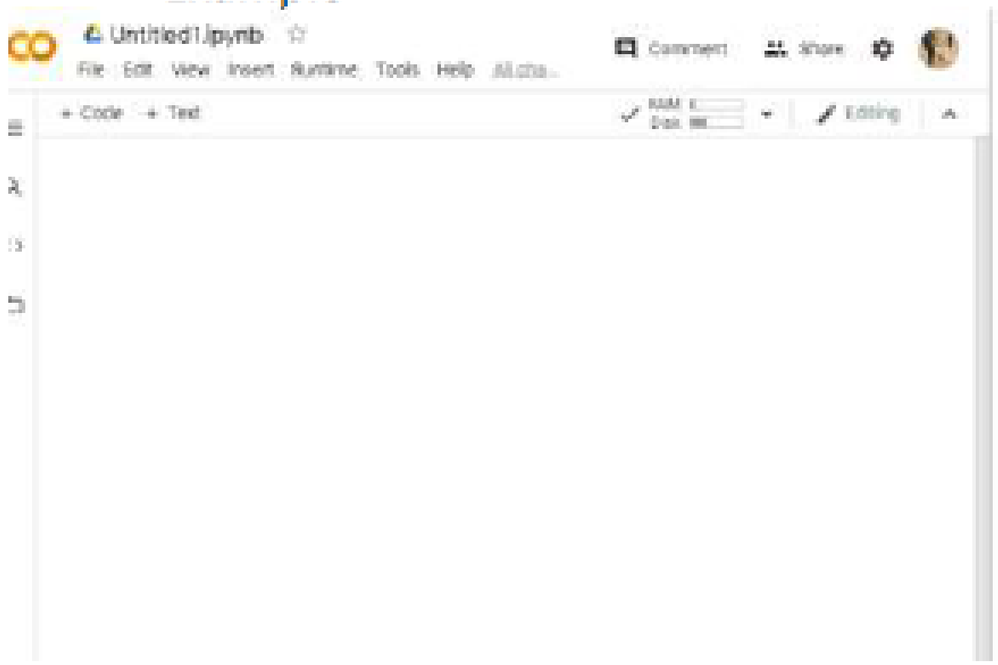
VARIABLE

143


Google Colab
New value can be assigned to the cup
Value can be accessed
SAME CUP, DIFFERENT CONTENT

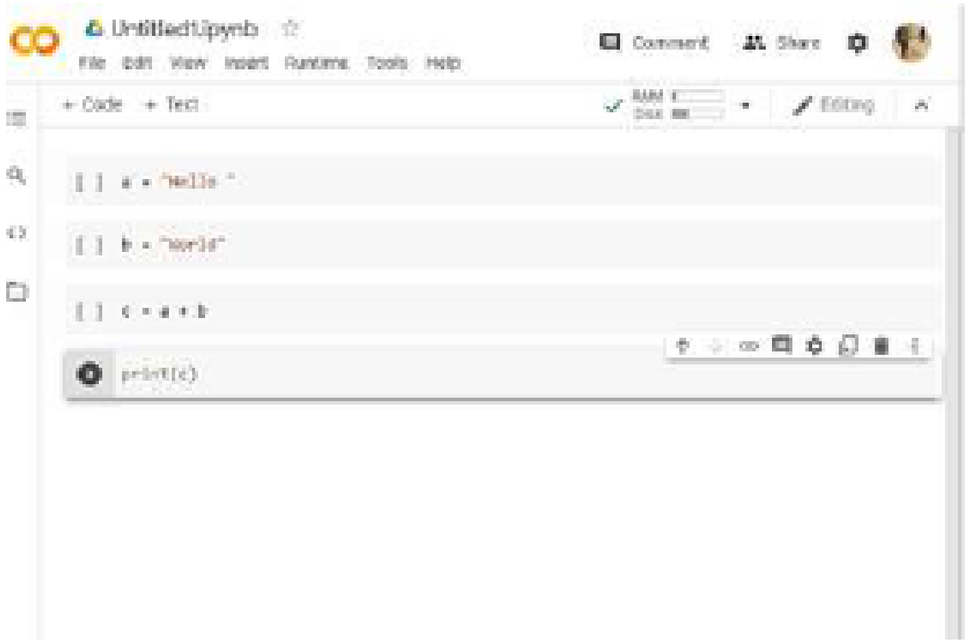
Google Colab

- Example



Google Colab

- Beware of the computation order !!!



Google Colab

- What the variable can store
- Different programming languages
  - Different types and names

- There are 3 major types
  - String
  - Number
    - Fraction
    - Integer
  - Boolean

Google Colab

- Integer
  - a number without fractions 1, 2, 7, - 11

- Decimal Number
  - A decimal number is a number with one or more digits to the right of the decimal point
  - a number that uses a decimal point followed by digits that show a value smaller than one

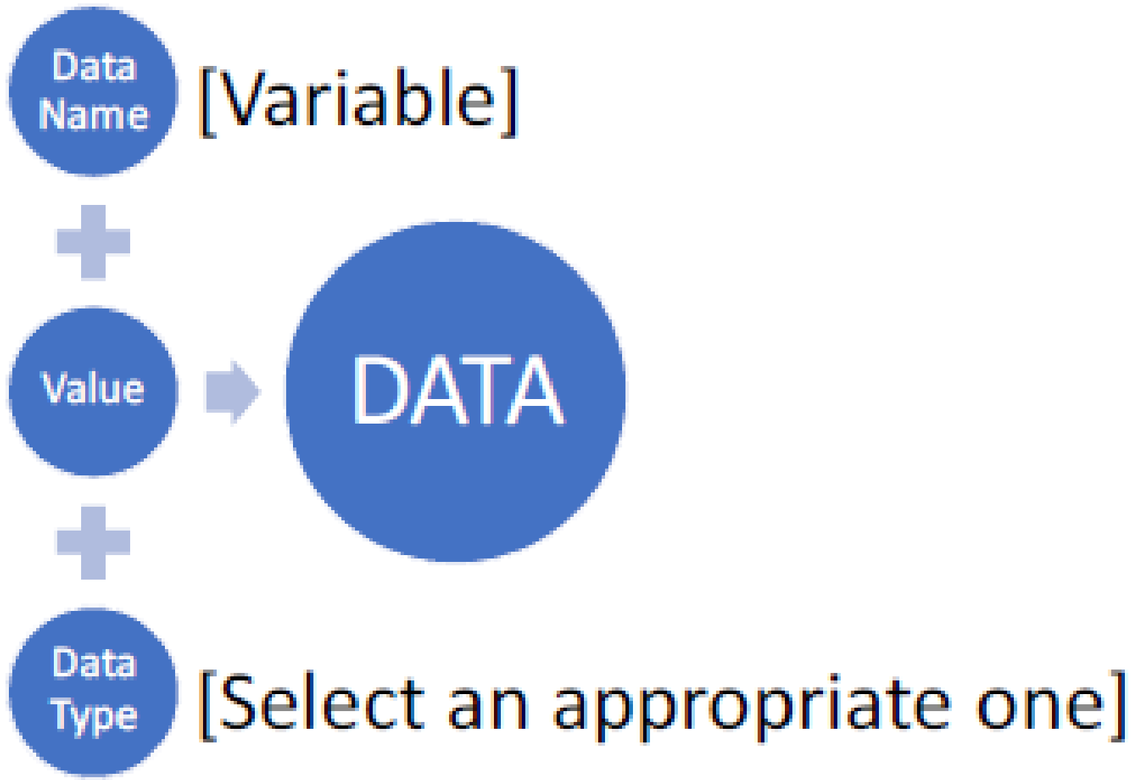
College of Arts, Media, and Technology

Google Colab

- Boolean
  - Logical value
  - Represent the truth values of logic and Boolean algebra
  - Two values
    - True, false

- String
  - a sequence of characters
  - Words or sentences
  - Always surround by double quote (")

College of Arts, Media, and Technology


Data Name
[Variable]
Value
DATA
Data Type
[Select an appropriate one]

What is Programming Thinking?

- Computer is not a human
  - Can manipulate data only one thing at a time

What is Programming Thinking?

- Changing the Mind set
  - Serialize your solution
- Break down the tasks
  - Into a step of operation
- Use the computational syntax to develop a solution

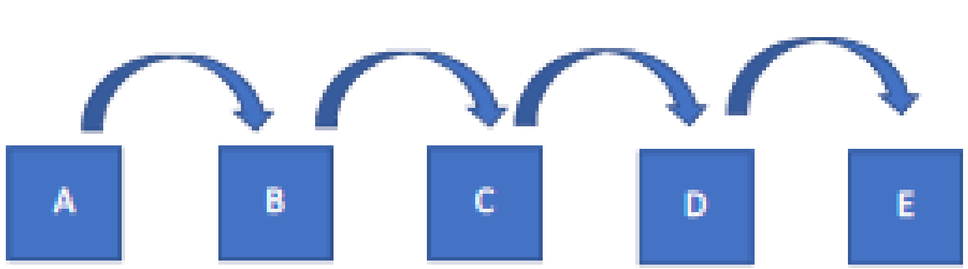
What will be inside?

- Sequential structure
  - Do operations from the first operation till the last operation

- Selection structure
  - Select to do something depended on the decision making

- Repetitive structure
  - Repeat operations until some criteria are satisfied

Sequential block

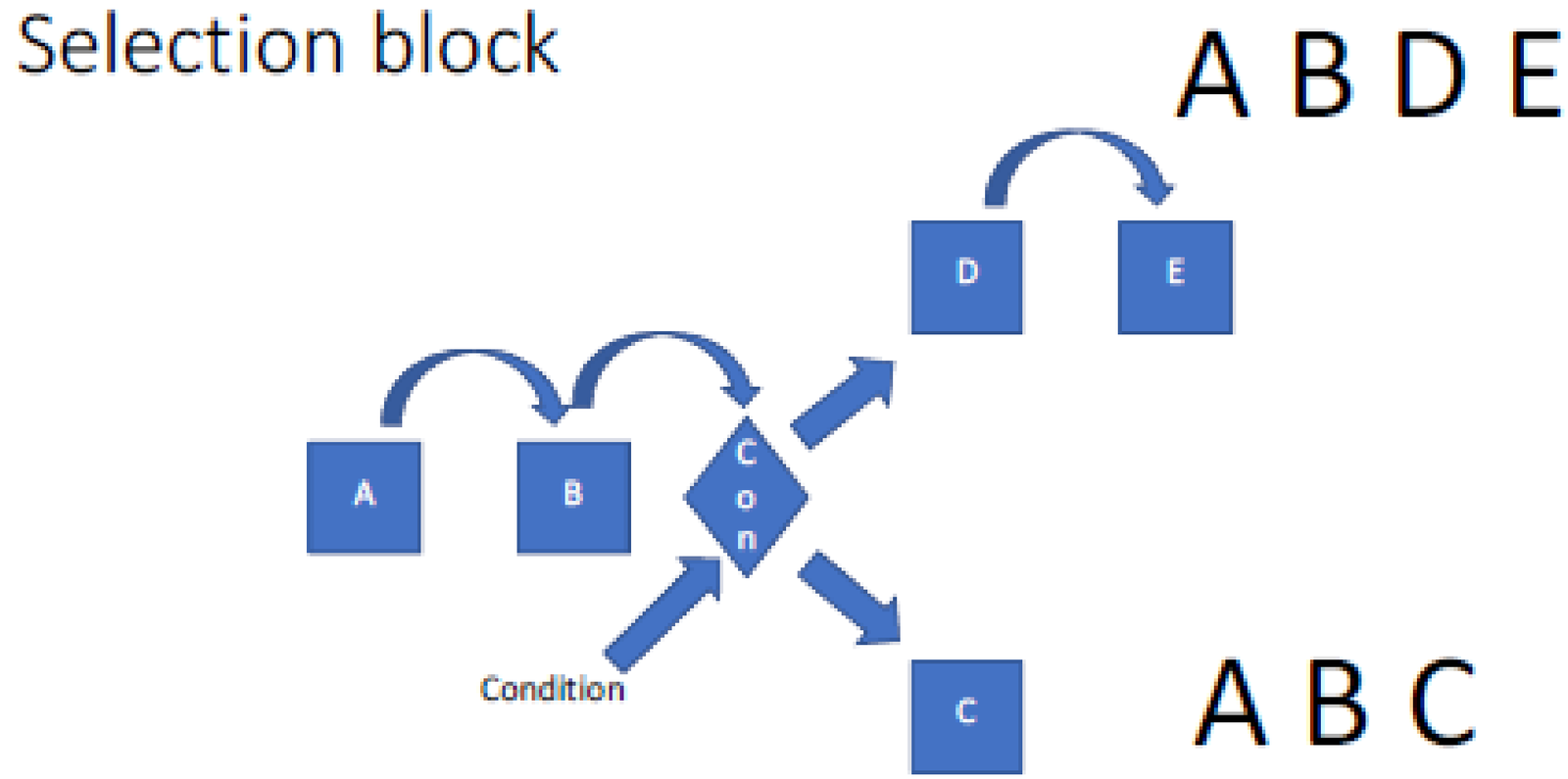


A B C D E

Sequential block

- Example of sequential block in Python

```
print("Hello")
print("Programming")
print("World")
```


Selection block
A B D E
D
E
A
B
Con
Condition
C
A B C

Selection block

- If-statement

```
if condition :
      statement 11
      statement 12
      …
else :
      statement 21
      statement 22
      …
```

Selection block

- When the condition is evaluated as True

```
[6]  a = 10
     if a > 5 :
       print("Hello World")
```

```
a = 10
if a > 5 :
  print("Hello World")
else :
  print("Not Hello World")
```

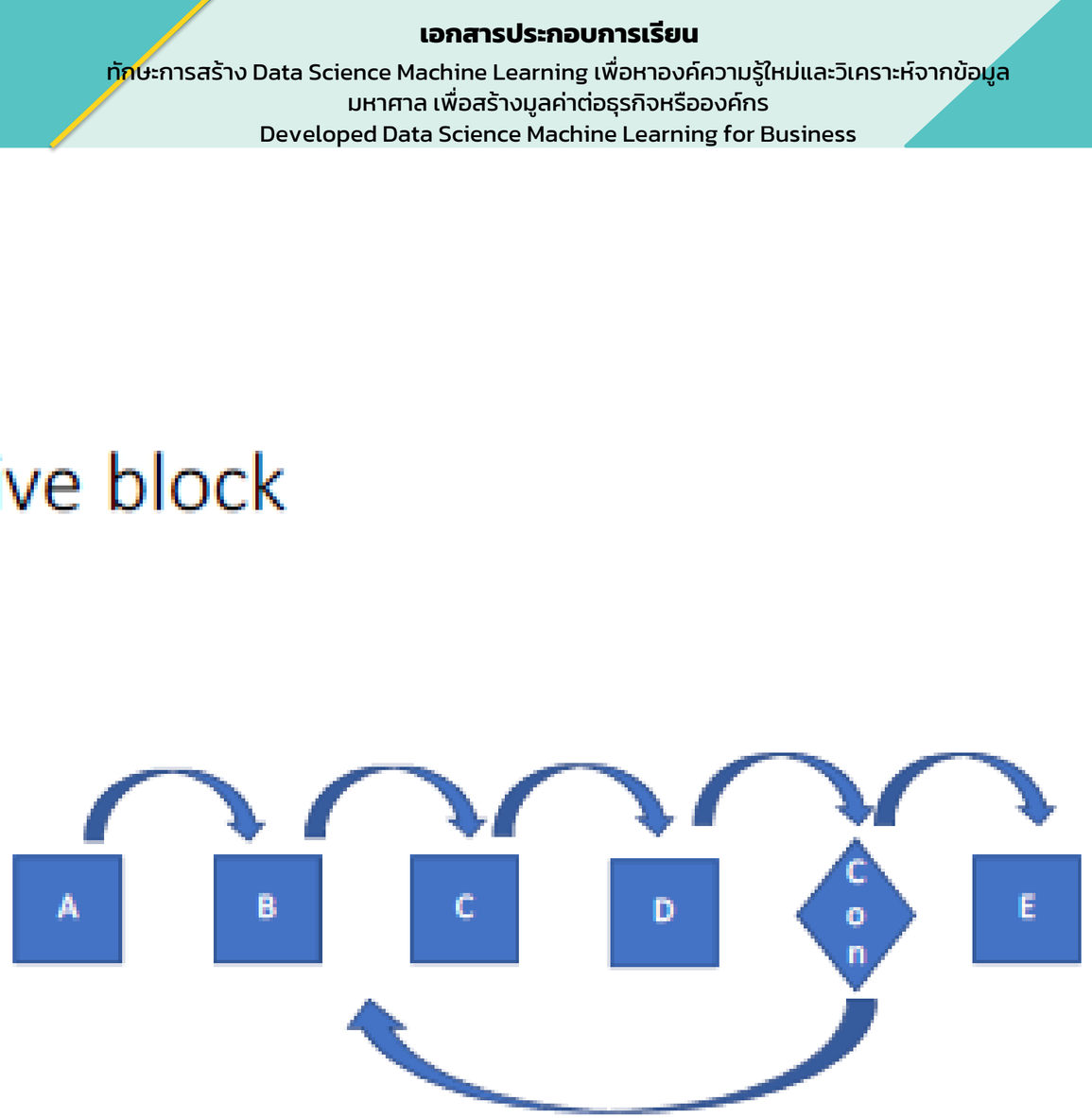
Selection block

- When the condition is evaluated as False

```
[10] a = 4
     if a > 5 :
       print("Hello World")
```

```
a = 4
if a > 5 :
  print("Hello World")
else :
  print("Not Hello World")
```

Repetitive block



A B C D C D ... C D E

Repetitive block

- While-statement

```
while condition :
    statement 11
    statement 12
    ...
```

```
[14] counter = 0
     while counter < 5 :
       print(counter)
       counter = counter + 1
```

Repetitive block

- For-statement

```
for index in list :
      statement 11
      statement 12
      …
```

Repetitive block

- Example of repetition block in Python

```
[14] counter = 0
     while counter < 5 :
       print(counter)
       counter = counter + 1
```

```
for i in [1,2,3,4,5]:
  print(i)
```

Try it yourself

- Calculate grade

Grade	Lower	Upper
A	75.00	100.00
B	50.00	74.99
F	0.00	49.99

- Print the following pattern

1

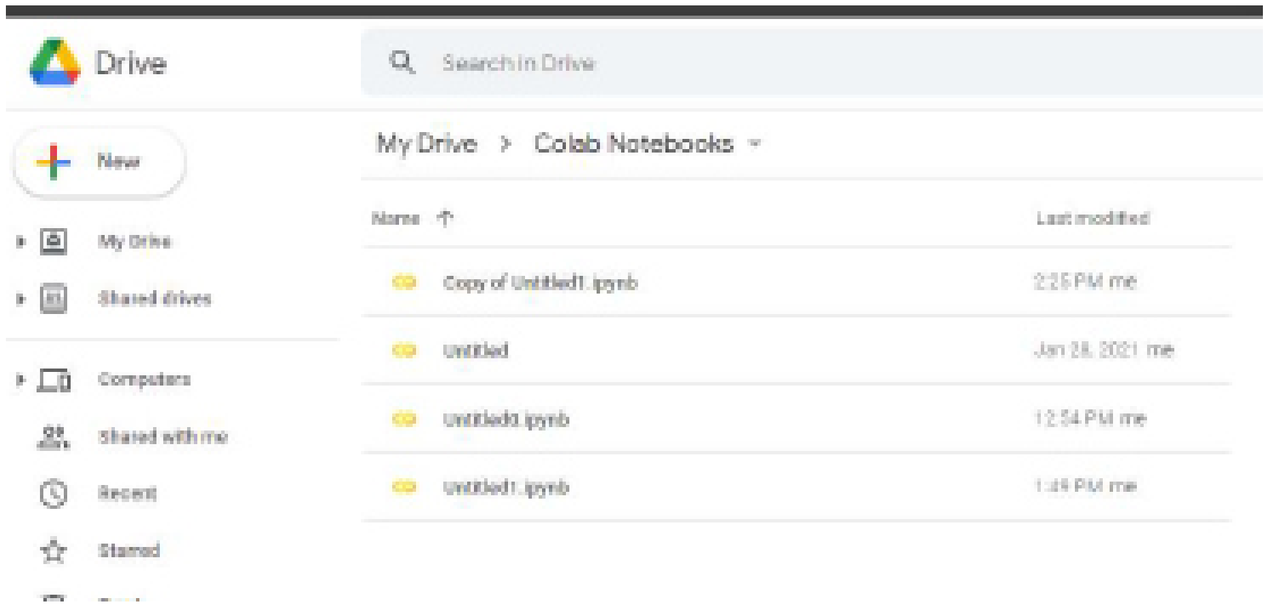
12

123

1234

Google Colab

- Colab Workbook is saved to your google drive.



Access to your Google Drive file from Colab

Colab

- Import the necessary library

```
!pip install PyDrive
```

```
Requirement already satisfied: PyDrive in /usr/local/lib/python3.6/dist-pa
Requirement already satisfied: google-api-python-client>=1.2 in /usr/local
Requirement already satisfied: oauth2client>=4.0.0 in /usr/local/lib/pytho
Requirement already satisfied: PyYAML>=3.0 in /usr/local/lib/python3.6/dis
Requirement already satisfied: google-auth>=1.4.1 in /usr/local/lib/python
Requirement already satisfied: uritemplate<4dev,>=3.0.0 in /usr/local/lib,
Requirement already satisfied: httplib2<1dev,>=0.17.0 in /usr/local/lib/py
Requirement already satisfied: google-auth-httplib2>=0.0.3 in /usr/local/l
Requirement already satisfied: six<2dev,>=1.6.1 in /usr/local/lib/python3
Requirement already satisfied: rsa>=3.1.4 in /usr/local/lib/python3.6/dist
Requirement already satisfied: pyasn1>=0.1.7 in /usr/local/lib/python3.6/d
Requirement already satisfied: pyasn1-modules>=0.0.5 in /usr/local/lib/pyt
Requirement already satisfied: cachetools<5.0,>=2.0.0 in /usr/local/lib/py
Requirement already satisfied: setuptools>=40.3.0 in /usr/local/lib/python
```

```
from pydrive.auth import GoogleAuth
from pydrive.drive import GoogleDrive
from google.colab import auth
from oauth2client.client import GoogleCredentials
```

Colab

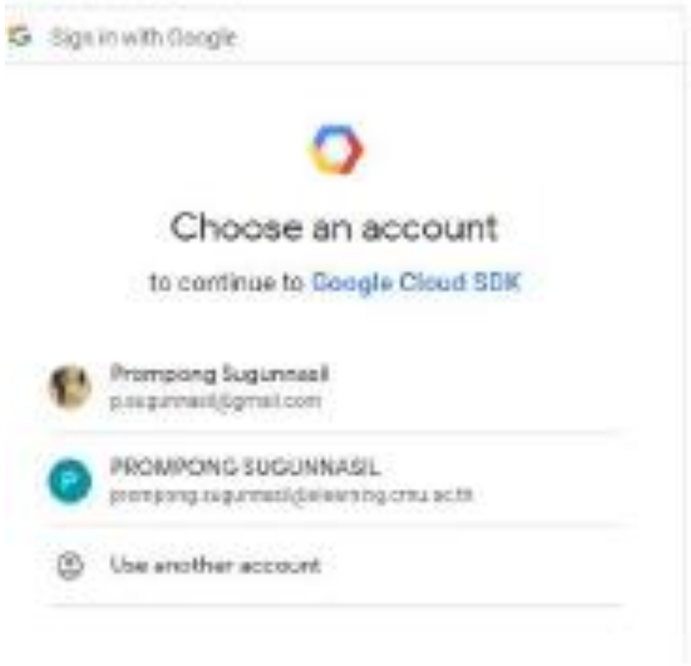
- Authenticate with google drive

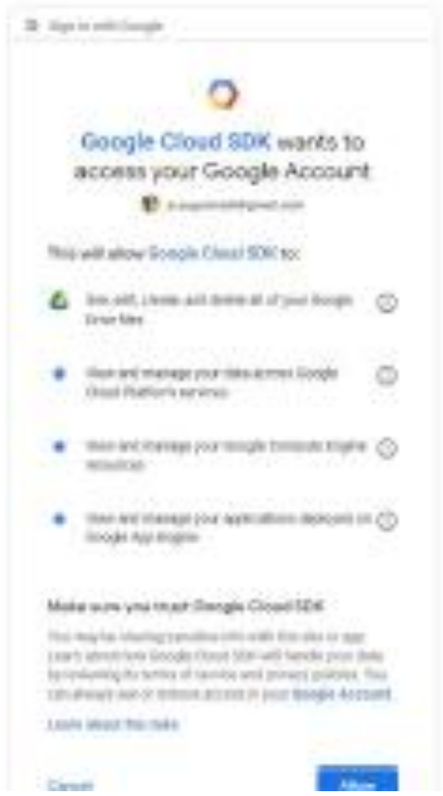
```
auth.authenticate_user()
gauth = GoogleAuth()
gauth.credentials = GoogleCredentials.get_application_default()
drive = GoogleDrive(gauth)
```

```
Go to the following link in your browser:

    https://accounts.google.com/o/oauth2/auth?response_type=code&client_id=32555940559.apps.googleusercontent.com&redirect_uri=urn%3Aiet

Enter verification code:
```

Coll

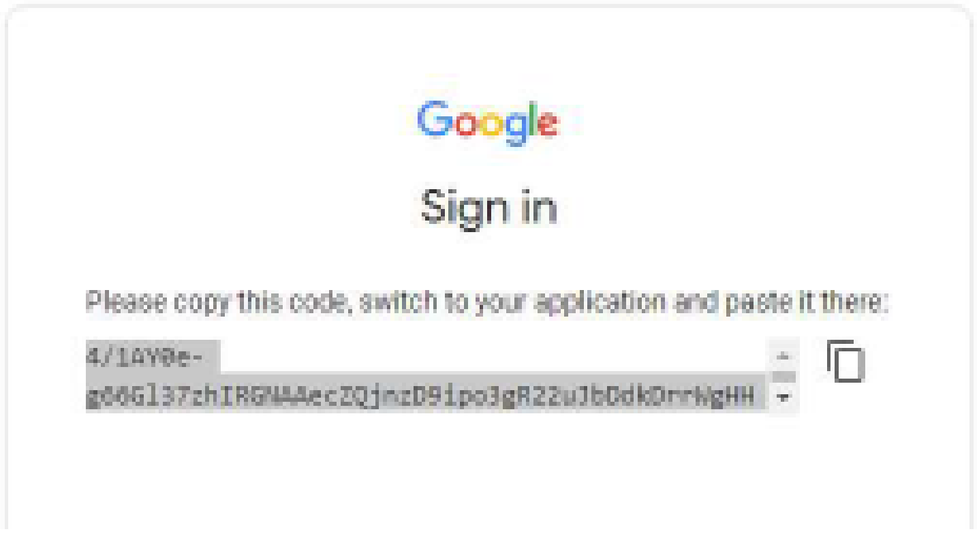
- Allow access to file





Colab

- Google will generate the authentication code



Colab

- Authenticate with google drive

```
auth.authenticate_user()
gauth = GoogleAuth()
gauth.credentials = GoogleCredentials.get_application_default()
drive = GoogleDrive(gauth)
```

```
Go to the following link in your browser:

    https://accounts.google.com/o/oauth2/auth?response_type=code&client_id=32555940559.apps.googleusercontent.com&redirect_uri=urn%3Aietf%

Enter verification code: ··········
```

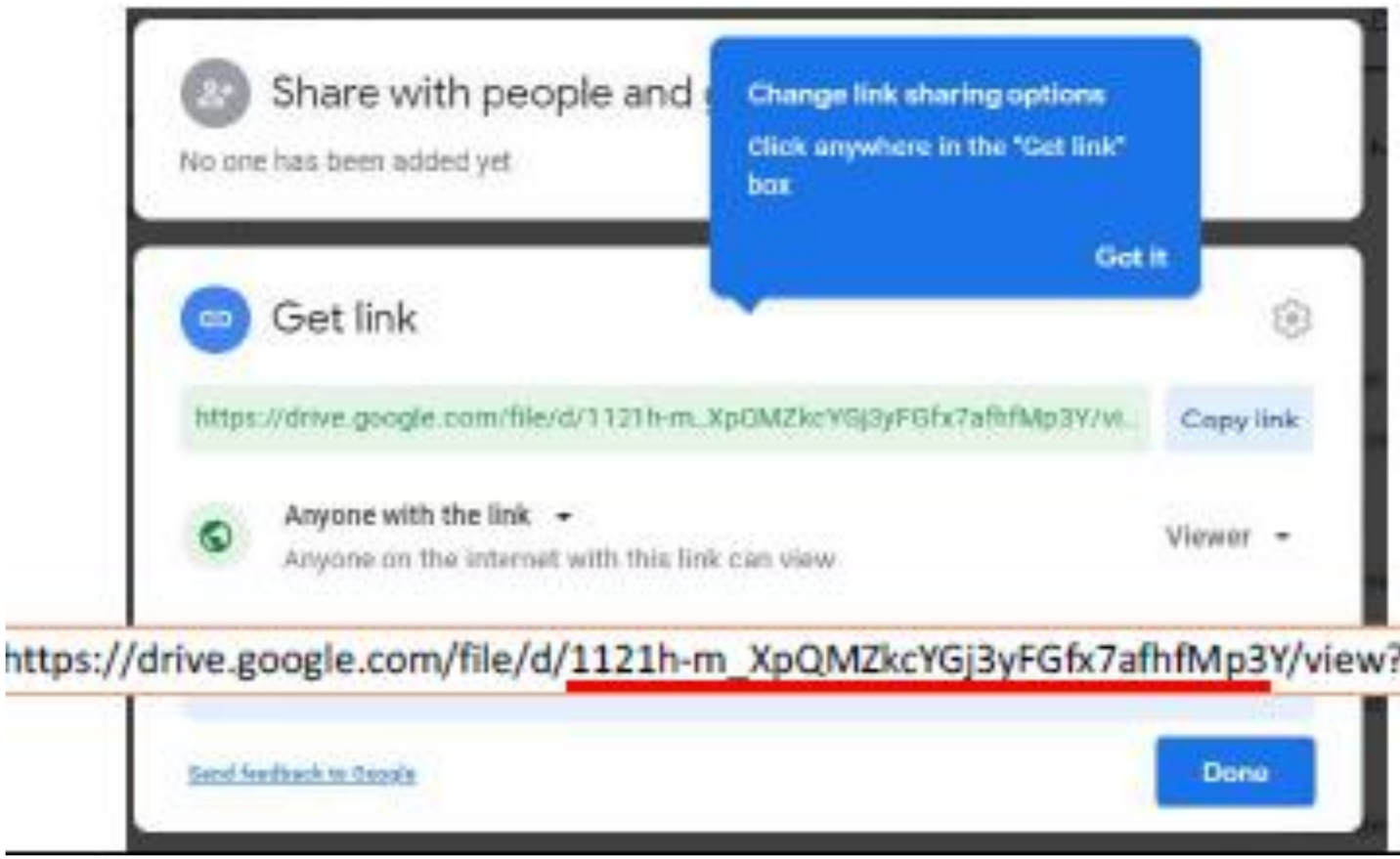
Colab

- Download the file from Google Drive to Colab environment

File ID

```
downloaded = drive.CreateFile({'id':"1121h-m_XpQMZkcYGj3yFGfx7afhfMp3Y"})
downloaded.GetContentFile('Churn_Modelling.csv')
```

Colab


Share with people and
No one has been added yet
Change link sharing options
Click anywhere in the "Get link" box
Got it
Get link
https://drive.google.com/file/d/1121h-m_XpQMZkcYGj3yFGfx7afhfMp3Y/vi...
Copy link
Anyone with the link
Anyone on the internet with this link can view
Viewer
https://drive.google.com/file/d/1121h-m_XpQMZkcYGj3yFGfx7afhfMp3Y/view?usp=sharing
Send feedback to Google
Done

Colab

- Access the files

```
import pandas as pd
data = pd.read_csv('Churn_Modelling.csv')
```

```
print(data)
```

```
      RowNumber  CustomerId    Surname  ...  IsActiveMember EstimatedSalary Exited
0             1    15634602   Hargrave  ...               1       101348.88      1
1             2    15647311       Hill  ...               1       112542.58      0
2             3    15619304       Onio  ...               0       113931.57      1
3             4    15701354       Boni  ...               0        93826.63      0
4             5    15737888   Mitchell  ...               1        79084.10      0
...         ...         ...        ...  ...             ...             ...    ...
9995       9996    15606229   Obijiaku  ...               0        96270.64      0
9996       9997    15569892  Johnstone  ...               1       101699.77      0
9997       9998    15584532        Liu  ...               1        42085.58      1
9998       9999    15682355   Sabbatini  ...               0        92888.52      1
9999      10000    15628319     Walker  ...               0        38190.78      0

[10000 rows x 14 columns]
```

Customer Segmentation

EDA

- Study the data
- Check all the column

```
print(data.columns)
```

```
Index(['RowNumber', 'CustomerId', 'Surname', 'CreditScore', 'Geography',
       'Gender', 'Age', 'Tenure', 'Balance', 'NumOfProducts', 'HasCrCard',
       'IsActiveMember', 'EstimatedSalary', 'Exited'],
      dtype='object')
```

EDA

- Study the data
- Check all the column

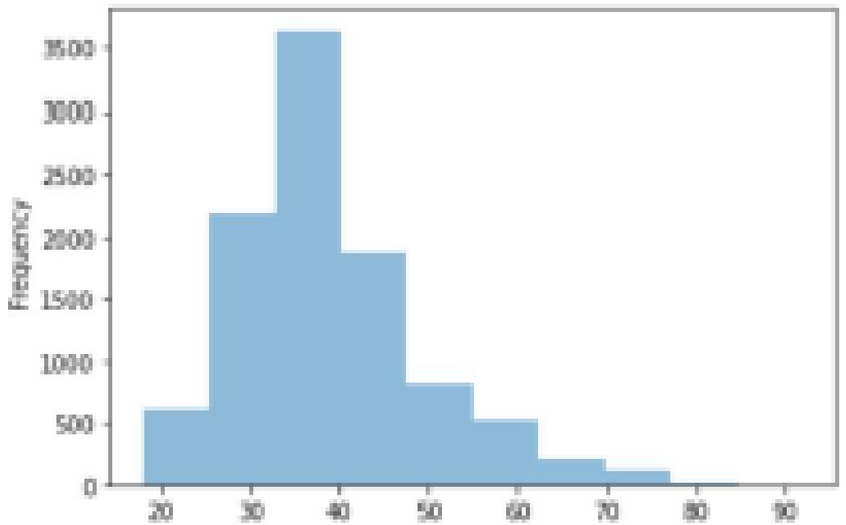
```
[8] print(data)

    data.describe()
```

EDA

- Investigate the data

```
data.Age.plot.hist(bins=10, alpha=0.5)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2f636bc748>
```

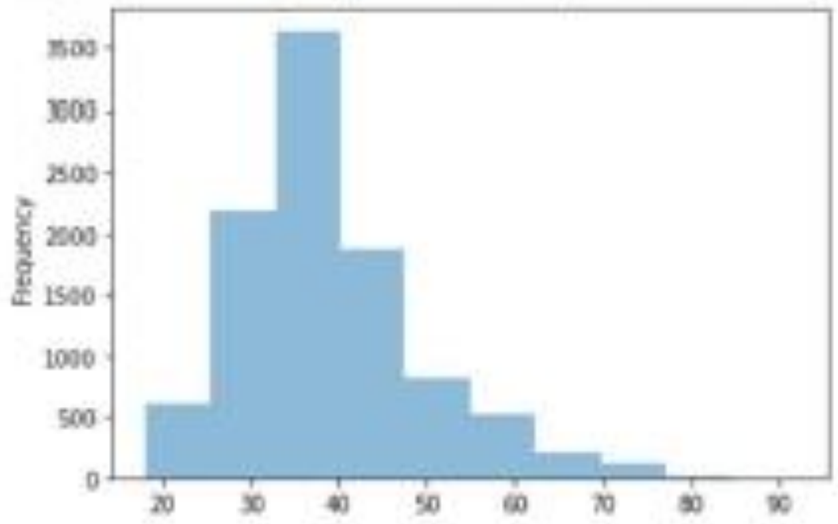


EDA

- Investigate the data

```
data.Age.plot.hist(bins=10, alpha=0.5)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2f636bc748>
```



EDA

- Remove all of the unnecessary data

```
temp_list =['RowNumber','Surname','CustomerId','Geography','Gender','Exited','HasCrCard', 'IsActiveMember','NumOfProducts','Balance','Tenure','Age']

data.drop(temp_list, axis=1, inplace=True)
```

```
print(data)
```

EDA

- Check the value

```
print(data)
```

```
      CreditScore  EstimatedSalary
0             619        101348.88
1             608        112542.58
2             502        113931.57
3             699         93826.63
4             850         79084.10
...           ...              ...
9995          771         96270.64
9996          516        101699.77
9997          709         42085.58
9998          772         92888.52
9999          792         38190.78

[10000 rows x 2 columns]
```

Data Preprocessing

- Transform the data

```
data.isna().sum()
data.isnull().sum()
```

```
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
transformed_data=scaler.fit_transform(data)
```

Construct Model

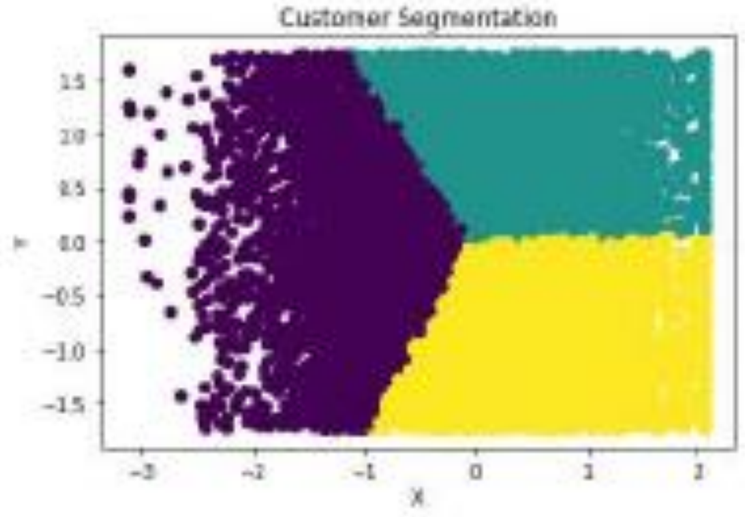
- Create a machine learning model

- K-means algorithm

```
from sklearn.cluster import KMeans

kmeans = KMeans(n_clusters=3, random_state=170)

y_pred = kmeans.fit_predict(transformed_data)
```

Display the result

```
import matplotlib.pyplot as plt

plt.scatter(transformed_data[:,0], transformed_data[:,1],c=y_pred)
plt.title('Customer Segmentation')
plt.xlabel('X')
plt.ylabel('Y')
plt.show()
```



Display the result

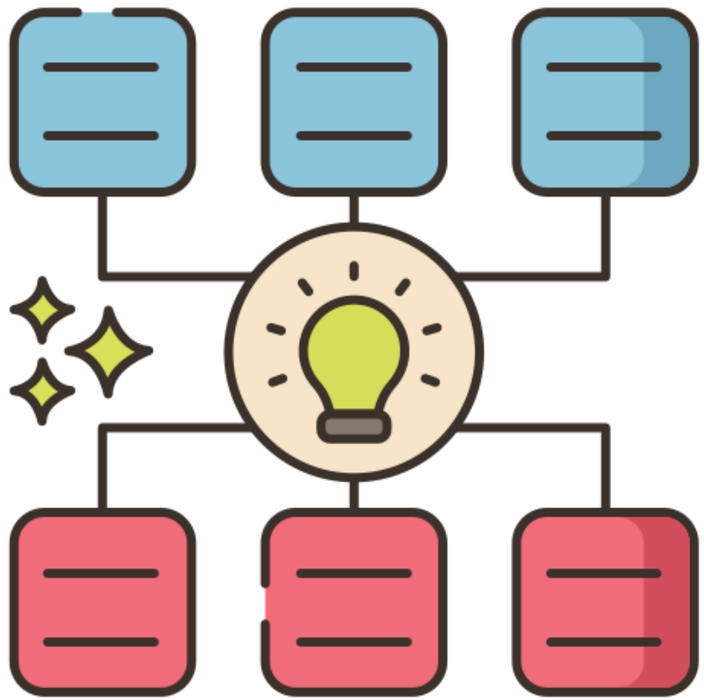
```
print(kmeans.cluster_centers_)
```

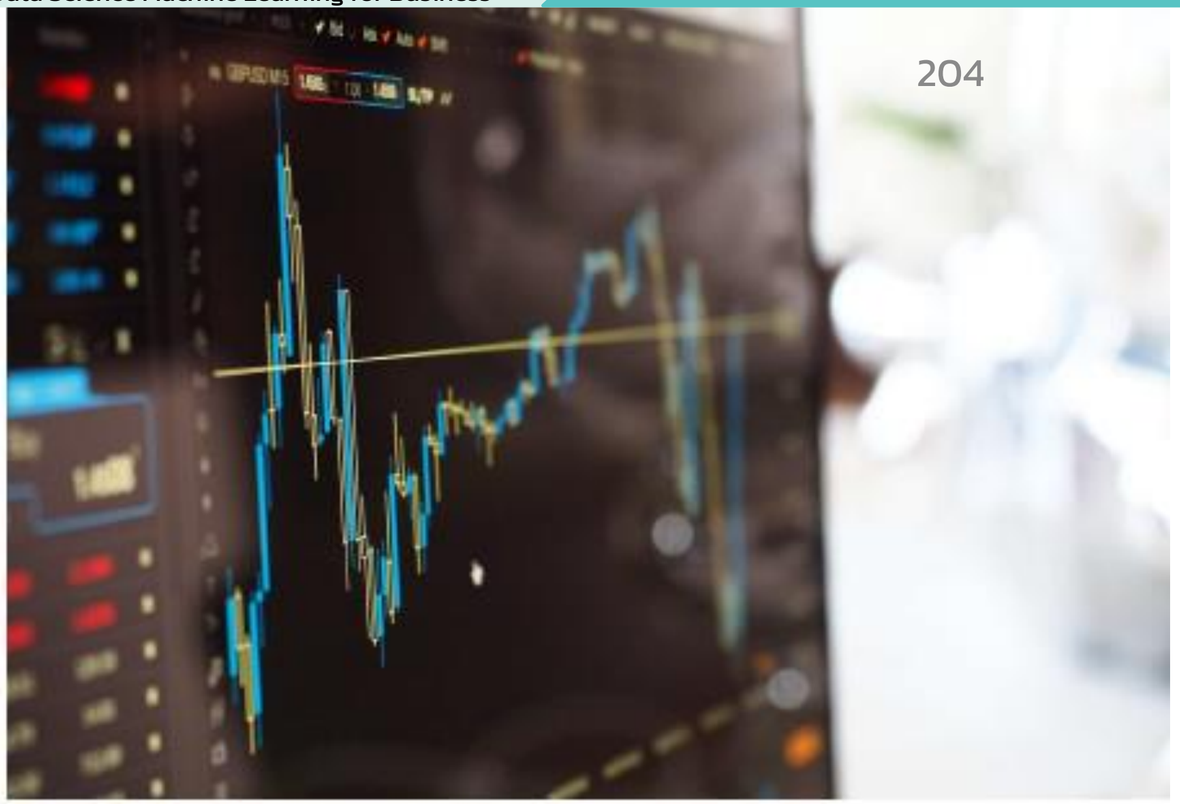
```
[[-1.12517304 -0.04669341]
 [ 0.46570969  0.95784833]
 [ 0.52125059 -0.91495691]]
```

```
[22] center = kmeans.cluster_centers_

     plt.scatter(transformed_data[:,0], transformed_data[:,1],c=y_pred)
     plt.scatter(center[:, 0], center[:, 1], c='red', alpha=0.6)
     plt.title('Customer Segmentation')
     plt.xlabel('X')
     plt.ylabel('Y')
```

```
Text(0, 0.5, 'Y')
```

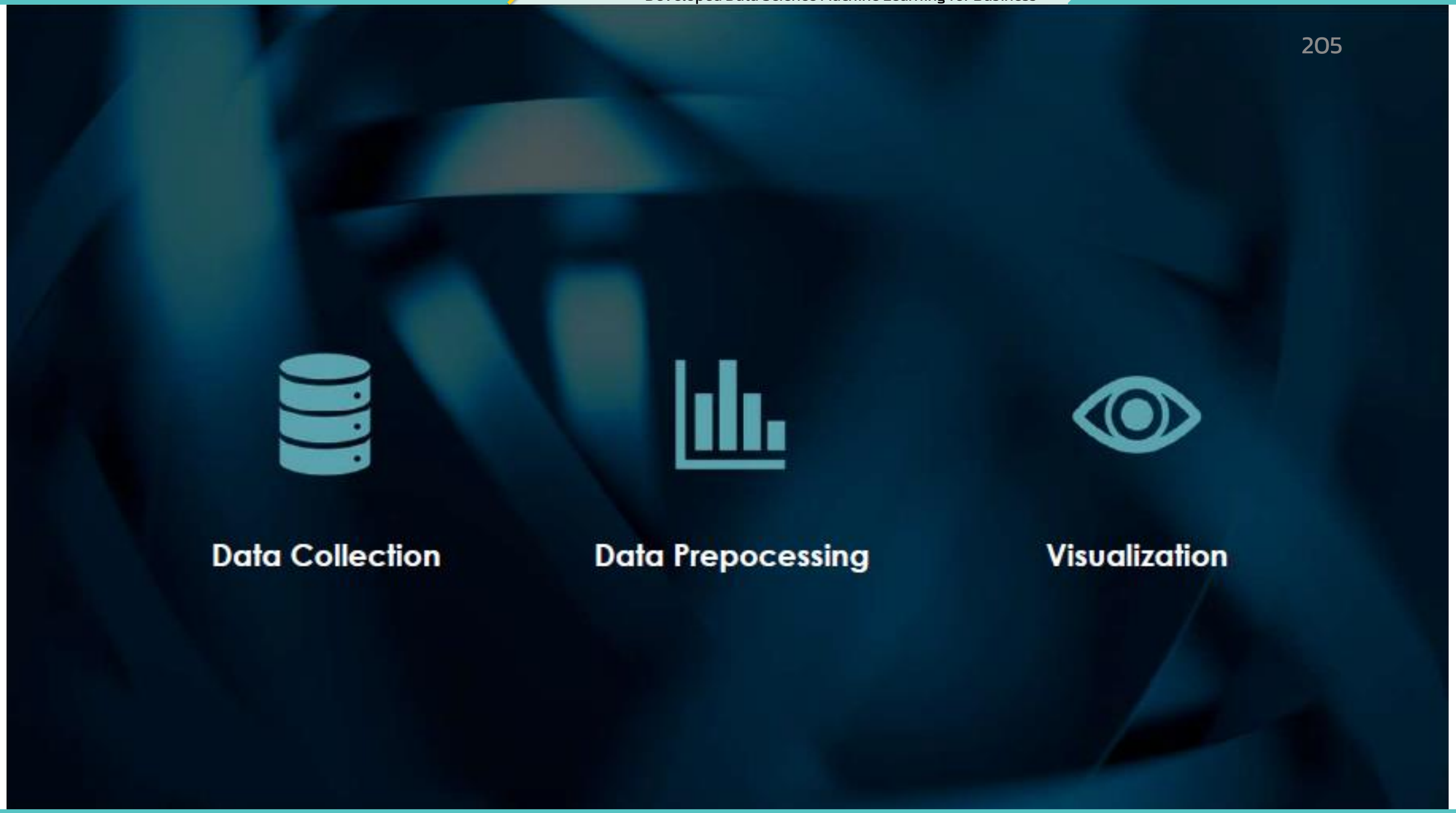
3.4 บทที่ 4 :
Data Preparation





Python Programming for Financial Analysis

Data Preparation (Financial News)

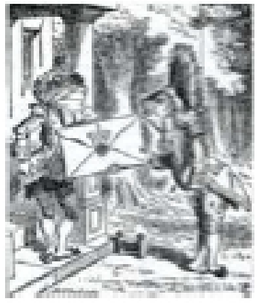

Data Collection
Data Prepocessing
Visualization

Recap: Data Collection

Identify Data Sources

Web scraping using Python 3

Library BeautifulSoup
Library Request



CRISP-DM (Cross-industry standard process for data mining)

CRISP-DM (Cross-industry standard process for data mining)

Data Preparation: the objective is to develop the final data set(s) for modeling.

Select	Clean	Construct	Integrate	Format
Select data: Determine which data sets will be used and document reasons for inclusion/exclusion.	**Clean data:** Often this is the lengthiest task. Without it, you'll likely fall victim to garbage-in, garbage-out. A common practice in this task is to correct, impute, or remove erroneous values.	**Construct data:** Derive new attributes that will be helpful. For example, derive someone's body mass index from height and weight fields.	**Integrate data:** Create new data sets by combining data from multiple sources.	**Format data:** Re-format data as necessary. For example, you might convert string values that store numbers to numeric values so that you can perform mathematical operations.

https://www.datascience-pm.com/crisp-dm-2/

Libraries

```
import pandas as pd
import ast
from collections import Counter
import matplotlib.pyplot as plt
```

Python's Library for Text Processing

```
!pip install pythainlp
from pythainlp import word_tokenize, Tokenizer, sent_tokenize
from pythainlp.corpus.common import thai_words, provinces
from pythainlp.tokenize import word_tokenize
from pythainlp.corpus import thai_stopwords, get_corpus
from pythainlp.util import dict_trie, normalize
from nltk.stem.porter import PorterStemmer
```

PyThaiNLP

Thai natural language processing in Python.

Function for Text Processing

```
#Name Entity
new_words = {'สตาร์ทอัพ','ซอมสา','พลังสา','กลิ้งไหบ','ไทยพาณิชย์'}
words = new_words.union(thai_words())
custom_dictionary_trie = dict_trie(words)

#Stopword
vowels = ('“','”','”','_','!!','-','-','(',')','!','"')
stopword_set = frozenset(vowels)
TH_stopword = thai_stopwords().union(stopword_set)

p_stemmer = PorterStemmer()

def clean(word):
    dfTitle = word.strip('!()/\|}"’‘{_<>[]')
    tokendfTitle = word_tokenize(normalize(dfTitle), custom_dict=custom_dictionary_trie,
    keep_whitespace=False, engine='newmm')
    Word_in_Title = [word for word in tokendfTitle if not word in TH_stopword]
    Word_in_Title = [p_stemmer.stem(i) for i in Word_in_Title]
    return Word_in_Title
```

Import Data

```
df = pd.read_excel('/content/finance_news_final.xlsx', sheet_name='Sheet1',index_col=0)
company = pd.read_csv('/content/Symbol_companies.csv')
```

```
[11] df.head(2)
```

	Date	Title	Summary	view	timestamp	link	Detail	Tags
0	21 กันยายน 2563	นักลงทุนทั่วโลกผวา สารพัดข่าวลบ แห่ขาย 'หุ้น-ทอ...	นักลงทุนตื่นขายทุก สินทรัพย์การลงทุนทั่ว โลก หลั...	83	2020-09-22 01:29:20.484985	https://www.bangkokbiznews.com/news/detail/898...	นักลงทุนทั่วโลก ต่างพากันเทขาย "สินทรัพย์" การล...	['ดาวโจนส์', 'หุ้นโลก', 'ทองคำ', 'น้ำมัน', 'เท...
1	21 กันยายน 2563	3 พันธมิตรตั้งทีมที่ ปรึกษาสตาร์ทอัพ	เมนเทอร์สตาร์ทอัพ ร่วม กับจัดงานสัมมนา "Ready, ...	86	2020-09-22 01:29:20.485048	https://www.bangkokbiznews.com/news/detail/898...	เมื่อวันที่ 17 กันยายน ที่ผ่านมา บริษัท หลักทร...	['สตาร์ทอัพ', 'Ready', 'Get Set', 'Go', 'บริษั...

```
[12] company.head(2)
```

	Symbol	Company	Market	Industry	Sector
0	2S	2S METAL PUBLIC COMPANY LIMITED	mai	Industrial	NaN
1	7UP	SEVEN UTILITIES AND POWER PUBLIC COMPANY LIMITED	SET	Resources	Energy & Utilities

Data Format

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4200 entries, 0 to 4199
Data columns (total 8 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   Date       4200 non-null   object
 1   Title      4200 non-null   object
 2   Summary    4200 non-null   object
 3   view       4200 non-null   object
 4   timestamp  4200 non-null   object
 5   link       4200 non-null   object
 6   Detail     4108 non-null   object
 7   Tags       4200 non-null   object
dtypes: object(8)
memory usage: 295.3+ KB
```

```
[14] company.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 793 entries, 0 to 792
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Symbol    793 non-null    object
 1   Company   793 non-null    object
 2   Market    793 non-null    object
 3   Industry  793 non-null    object
 4   Sector    622 non-null    object
dtypes: object(5)
memory usage: 31.1+ KB
```

Missing Data

```
[15] df.isnull().sum()

Date          0
Title         0
Summary       0
view          0
timestamp     0
link          0
Detail       92
Tags          0
dtype: int64
```

```
company.isnull().sum()

Symbol         0
Company        0
Market         0
Industry       0
Sector       171
dtype: int64
```

Data Type in python

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

NumPy

python™

Pandas dtype	Python type	NumPy type	Usage
object	str or mixed	string_, unicode_ mixed types	Text or mixed numeric and non-numeric values
int64	int	int_, int8, int16, int32, int64, uint8, uint16, uint32, uint64	Integer numbers
float64	float	float_, float16, float32, float64	Floating point numbers
bool	bool	bool_	True/False values
datetime64	NA	datetime64[ns]	Date and time values
timedelta[ns]	NA	NA	Differences between two datetimes
category	NA	NA	Finite list of text values

source: https://pbpython.com/pandas_dtypes.html

Re-format

```
df['view'] = df['view'].str.replace(',', '')
df['view'] = df.view.astype(int)

df['timestamp'] = pd.to_datetime(df['timestamp'])

df['Tags'] = df['Tags'].apply(lambda x: ast.literal_eval(x))
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4200 entries, 0 to 4199
Data columns (total 8 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Date       4200 non-null   object
 1   Title      4200 non-null   object
 2   Summary    4200 non-null   object
 3   view       4200 non-null   object
 4   timestamp  4200 non-null   object
 5   link       4200 non-null   object
 6   Detail     4108 non-null   object
 7   Tags       4200 non-null   object
dtypes: object(8)
memory usage: 295.3+ KB
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4108 entries, 0 to 4199
Data columns (total 8 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Date       4108 non-null   object
 1   Title      4108 non-null   object
 2   Summary    4108 non-null   object
 3   view       4108 non-null   int64
 4   timestamp  4108 non-null   datetime64[ns]
 5   link       4108 non-null   object
 6   Detail     4108 non-null   object
 7   Tags       4108 non-null   object
dtypes: datetime64[ns](1), int64(1), object(6)
memory usage: 288.8+ KB
```

```
print(type(df['Tags'][0]))
df['Tags'][0]
```

```
<class 'list'>
['ดาวโจนส์',
 'หุ้นโลก',
 'ทองคำ',
 'น้ำมัน',
 'เทขาย',
 'โควิด-19',
 'อังกฤษ',
 'ล็อกดาวน์']
```

Derive New Attributes

From 'Date'
To
- Day
- Month
- Year

```
[163] # new data frame with split value columns
      new = df['Date'].str.split(" ", n = 2, expand = True)
      new.head(2)
```

	0	1	2
0	21	กันยายน	2563
1	21	กันยายน	2563

```
df['Day'] = new[0]
df['Month'] = new[1]
df['Year'] = new[2]
df['Year'] = df['Year'].astype(int)
df['Year'] = df['Year'] - 543
m = {'มกราคม':'01', 'กุมภาพันธ์':'02', 'มีนาคม':'03', 'เมษายน':'04','พฤษภาคม':'05','มิถุนายน':'06',
     'กรกฎาคม':'07','สิงหาคม':'08','กันยายน':'09','ตุลาคม':'10','พฤศจิกายน':'11','ธันวาคม':'12'}
df['Month'] = df.Month.map(m)
df.head(2)
```

	Date	Title	Summary	view	timestamp	link	Detail	Tags	Day	Month	Year
0	21 กันยายน 2563	นักลงทุนทั่วโลกผวา สารพัดข่าวลบ หุ้น-ทอ...	นักลงทุนตื่นขายทุกสินทรัพย์ การลงทุนทั่วโลกพลิ...	83	2020-09-22 01:29:20.464985	https://www.bangkokbiznews.com/news/detail/898...	นักลงทุนทั่วโลกต่างพากันเทขาย "สินทรัพย์" การล...	[ดาวโจนส์, หุ้นโลก, ทองคำ, น้ำมัน, เทขาย, เครื่...	21	09	2020
1	21 กันยายน 2563	3 พันธมิตรตั้งทีมที่ปรึกษาสตาร์ท	เมนเทอร์สตาร์ทอัพร่วมกันจัดงานสัมมนา	86	2020-09-22 01:29:20.485048	https://www.bangkokbiznews.com/news/detail/898...	เมื่อวันที่ 17 กันยายนที่ผ่านมา บริษัท	[สตาร์ทอัพ, Ready, Get Set, Go, บริษัท	21	09	2020

Derive New Attributes (cont.)

From
- Day
- Month
- Year

To
'DatePublish'

```
cols = ['Year', 'Month','Day']
df['DatePublish'] = df[cols].apply(lambda x: '-'.join(x.values.astype(str)), axis=1)
df['DatePublish'] = pd.to_datetime(df['DatePublish'])
df.head(2)
```

Title	Summary	view	timestamp	link	Detail	Tags	Day	Month	Year	DatePublish
กลุ่มทุนทั่วโลกผวาสารพัดข่าวลบแห่ขาย "หุ้น-ทอ...	นักลงทุนสื่นขายทุกสินทรัพย์ การลงทุนทั่วโลกหลั...	83	2020-09-22 01:29:20.484985	https://www.bangkokbiznews.com/news/detail/898...	นักลงทุนทั่วโลกต่างพากันเทขาย "สินทรัพย์" การล...	[ดาวโจนส์, หุ้นโลก, ทองคำ, น้ำมัน, เทขาย, โควิ...	21	09	2020	2020-09-21
3 นูธมิตรตั้งทีมที่ปรึกษาสตาร์ทอัพ	เมนเทอร์สตาร์ทอัพร่วมกันจัดงานสัมมนา "Ready, ...	86	2020-09-22 01:29:20.485048	https://www.bangkokbiznews.com/news/detail/898...	เมื่อวันที่ 17 กันยายนที่ผ่านมา บริษัทหลักทร...	[สตาร์ทอัพ, Ready, Get Set, Go, บริษัทหลักทรั...	21	09	2020	2020-09-21

Derive New Attributes (cont.)

From 'Tags'
To
'CountTags'

```
[205] df['CountTags'] = df['Tags'].apply(lambda x: len(x))
      df.head(2)
```

ary	view	timestamp	link	Detail	Tags	Day	Month	Year	DatePublish	CountTags
งทุน ยทุก รัพย์ งทุน โลก ผลั...	83	2020-09-22 01:29:20.484985	https://www.bangkokbiznews.com/news/detail/898...	นักลงทุนทั่วโลกต่างพากันเทขาย "สินทรัพย์" การล...	[ดาวโจนส์, หุ้นโลก, ทองคำ, น้ำมัน, เทขาย, โควิ...	21	09	2020	2020-09-21	8
ทยร์ ทยัพ น่จัด งาน มนา ady, ...	86	2020-09-22 01:29:20.485048	https://www.bangkokbiznews.com/news/detail/898...	เมื่อวันที่ 17 กันยายน ที่ผ่านมา บริษัท หลักทร...	[สตาร์ท อัพ, Ready, Get Set, Go, บริษัท หลักทรั...	21	09	2020	2020-09-21	6

```
df['Tags'][0]
```

```
['ดาวโจนส์',
 'หุ้นโลก',
 'ทองคำ',
 'น้ำมัน',
 'เทขาย',
 'โควิด-19',
 'อังกฤษ',
 'ล็อกดาวน์']
```

Derive New Attributes (cont.)

From 'timestamp' & 'DatePublish'
To
'CountDays' & 'AvgViesperDay'

```
df['CountTime'] = df['timestamp'] - df['DatePublish']
CountDay = df['CountTime'].dt.days
df['CountDays'] = CountDay

df['AvgViewsperDay'] = df['view']/df['CountDays']

df.head(2)
```

	Date	Title	Summary	view	timestamp	link	Detail	Tags	Day	Month	Year	DatePublish	CountTags	CountTime	CountDays	AvgViewsperDay
0	21 กันยายน 2563	[illegible]	[illegible]	83	2020-09-22 01:29:20.484985	https://www.bangkokbiznews.com/news/detail/898...	[illegible]	[illegible]	21	09	2020	2020-09-21	8	1 days 01:29:20.484985	1	83.0
1	21 กันยายน 2563	[illegible]	[illegible]	86	2020-09-22 01:29:20.485048	https://www.bangkokbiznews.com/news/detail/898...	เมื่อวันที่ 17 กันยายน ที่ผ่านมา บริษัท [illegible]	[illegible]	21	09	2020	2020-09-21	6	1 days 01:29:20.485048	1	86.0

Derive New Attributes (cont.)

From 'Title'
To 'title_token','Count_title_token','title_token_stop','Count_title_token_stop'

```
[207] df['title_token'] = df['Title'].apply(lambda x: word_tokenize(normalize(x),keep_whitespace=False))
      df['Count_title_token'] = df['title_token'].apply(lambda x: len(x))

      df['title_token_stop'] = df['Title'].apply(lambda x: clean(x))
      df['title_token_stop'] = df['title_token_stop'].apply(lambda x: [val for val in x if val not in ['',' ','']])
      df['Count_title_token_stop'] = df['title_token_stop'].apply(lambda x: len(x))
```

```
df[['Title','title_token','Count_title_token','title_token_stop','Count_title_token_stop']]
```

	Title	title_token	Count_title_token	title_token_stop	Count_title_token_stop
0	นักลงทุนทั่วโลกผวาสารพัดข่าวลบ แห่ขาย 'หุ้น-ทอ...	[นักลงทุน, ทั่วโลก, ผวา, สารพัด, ข่าว, ลบ, แห่...	22	[นักลงทุน, ทั่วโลก, ผวา, สารพัด, ข่าว, ลบ, แห่...	18
1	3 พันธมิตรตั้งทีมที่ปรึกษาสตาร์ทอัพ	[3, พันธมิตร, ตั้ง, ทีม, ที่ปรึกษา, สตาร์ท, อัพ]	7	[3, พันธมิตร, ทีม, ที่ปรึกษา, สตาร์ทอัพ]	5
2	ข้อมูล "Warrant" (21 ก.ย.63)	[ข้อมูล, ", Warrant, ", (, 21, ก.ย., 63,)]	9	[ข้อมูล, warrant, 21, ก.ย., 63]	5
3	"อัตราแลกเปลี่ยน"เงินตราต่างประเทศ (21 ก.ย.63)	['', อัตราแลกเปลี่ยน, ", เงินตราต่างประเทศ, (, ...	9	[อัตราแลกเปลี่ยน, เงินตราต่างประเทศ, 21, ก.ย.,...	5
4	'เงินบาท' ปิดตลาดวันนี้ 'อ่อนค่า' ที่31.22บาท...	['', เงินบาท, ', ปิด, ตลาด, วันนี้, ', อ่อน, ค...	15	[เงินบาท, ตลาด, อ่อนค่า, 31.22, บาท, ดอลลาร์]	6

Derive New Attributes (cont.)

สร้าง **Attribute** ใหม่ โดยอาศัยข้อมูลชื่อหลักทรัพย์ เพื่อระบุว่าข่าวใดกล่าวถึงหลักทรัพย์ของบริษัทใด

class `collections.` **`Counter`**([*iterable-or-mapping*])

A `Counter` is a `dict` subclass for counting hashable objects. It is an unordered collection where elements are stored as dictionary keys and their counts are stored as dictionary values. Counts are allowed to be any integer value including zero or negative counts. The `Counter` class is similar to bags or multisets in other languages.

1. นับว่าแต่ละ **Tag** ปรากฏจำนวนกี่ครั้ง

```
[209] #Combine every tags
      TagsList = df.Tags.sum()

[210] #นับว่าแต่ละ Tag ปรากฏกี่ครั้ง
      count_tag = Counter(TagsList)
      count_tag_list = count_tag.most_common() #convert counter object to list

      count_tag_list
```

```
[('โควิด-19', 412),
 ('หุ้นไทย', 343),
 ('เงินบาท', 227),
 ('ธปท.อัตราแลกเปลี่ยน', 225),
 ('หุ้น', 162),
 ('เรารักษ์สิ่งนี้', 157),
 ('ราคาทอง', 139),
 ('ธปท.', 129),
 ('หุ้นไทยภาคบ่าย', 111),
 ('หุ้นไทยภาคเช้า', 110),
 ('อัตราแลกเปลี่ยน', 106),
 ('ศัพท์', 98),
 ('ออฟเวิคคอร์ด', 88),
 ('เอียวยา', 78),
 ('ก.ล.ต.', 76),
 ('สินเชื่อ', 69),
```

Derive New Attributes (cont.)

2. แสดง Tag ที่เป็นชื่อหลักทรัพย์ ที่ถูกติด Tag มากที่สุด

```
#check if any company symbol exist in TagsList
S1 = set(company.Symbol)
S2 = set(TagsList)
Symbol = S1.intersection(S2)
len(Symbol)
```

```
200
```

```
#create list of symbol that existed
symbol = [i for i in company.Symbol if i in TagsList]
```

```
symbol_count = []
for i in range(len(count_tag_list)):
    if count_tag_list[i][0] in symbol:
        symbol_count.append(count_tag_list[i])
symbol_count[:10]
```

```
[('KBANK', 50),
 ('SCB', 38),
 ('BBL', 30),
 ('THAI', 29),
 ('KTB', 25),
 ('SSF', 21),
 ('BAY', 13),
 ('TMB', 13),
 ('CIMBT', 13),
 ('GULF', 11)]
```

3. สร้าง attribute ใหม่เพื่อแสดงว่าข่าวนั้น ๆ มี Tag เป็นชื่อหลักทรัพย์หรือไม่ โดย
1 = มี
0 = ไม่มี

```
#เลือกมาเฉพาะ Symbol ที่ปรากฏมากที่สุด 10 อัน
symbol_top = [i[0] for i in symbol_count]
symbol_top = symbol_top[:10]
symbol_top
```

```
['KBANK', 'SCB', 'BBL', 'THAI', 'KTB', 'SSF', 'BAY', 'TMB', 'CIMBT', 'GULF']
```

```
from sklearn.preprocessing import MultiLabelBinarizer
mlb = MultiLabelBinarizer()
symbol_dummy = pd.DataFrame(mlb.fit_transform(df['Tags']),columns=mlb.classes_, index=df.index)
symbol_dummy = symbol_dummy[symbol_top]
symbol_dummy.head(3)
```

	KBANK	SCB	BBL	THAI	KTB	SSF	BAY	TMB	CIMBT	GULF
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0

sklearn.preprocessing.MultiLabelBinarizer

class sklearn.preprocessing.**MultiLabelBinarizer**(*, classes=None, sparse_output=False) [source]

Transform between iterable of iterables and a multilabel format

Although a list of sets or tuples is a very intuitive format for multilabel data, it is unwieldy to process. This transformer converts between this intuitive format and the supported multilabel format: a (samples x classes) binary matrix indicating the presence of a class label.

Derive New Attributes (cont.)

เพิ่ม columns หลักทรัพย์ที่สร้างขึ้นมาใหม่เข้าไปใน dataframe

```
df = pd.concat([df, symbol_dummy], axis=1, sort=False)
df.head(3)
```

untDays	AvgViewsperDay	title_token	Count_title_token	title_token_stop	Count_title_token_stop	KBANK	SCB	BBL	THAI	KTB	SSF	BAY	TMB	CIMBT	GULF
1	83.0	[นักลงทุน, ทั่วโลก, ผวา, สาวพัด, ข่าว, ลบ, แห่...	22	[นักลงทุน, ทั่วโลก, ผวา, สาวพัด, ข่าว, ลบ, แห่...	18	0	0	0	0	0	0	0	0	0	0
1	86.0	[3, พันธมิตร, ตั้ง, ทีม, ที่ปรึกษา, สตาร์ท, อัพ]	7	[3, พันธมิตร, ทีม, ที่ปรึกษา, สตาร์ทอัพ]	5	0	0	0	0	0	0	0	0	0	0
1	43.0	[ข้อมูล, ', Warrant, ', (, 21, ก.ย., 63,)]	9	[ข้อมูล, warrant, 21, ก.ย., 63]	5	0	0	0	0	0	0	0	0	0	0

Top common tags

นับจำนวน Tag ที่ซ้ำ โดยเลือกที่มีจำนวนซ้ำมากที่สุด 100 อันดับแรก และ export เป็นไฟล์ csv

Tag

```
[ ]  count_tag_top = count_tag.most_common(100)
```

```
#export top tags to csv
toptag = []
counttag = []
for item in count_tag_top:
    toptag.append(item[0])
    counttag.append(item[1])
df_toptag = pd.DataFrame(zip(toptag, counttag),columns=['toptag','counttoptag'])
df_toptag.index.name = 'index'
df_toptag.to_csv('df_toptag.csv',encoding='utf-8')
```

index	toptag	counttoptag
0	โควิด-19	412
1	หุ้นไทย	343
2	เงินบาท	227
3	อปท.สัดราแลกเปลี่ยน	225
4	หุ้น	162
5	เราไม่ทิ้งกัน	157
6	ราคาทอง	139
7	อปท.	129
8	หุ้นไทยภาคบ่าย	111
9	หุ้นไทยภาคเช้า	110
10	อัตราแลกเปลี่ยน	106
11	ตัดผ้า	98
12	ออฟโรดคลอร์ส	88
13	เสียงยา	79
14	ก.ล.ต.	76
15	สินเชื่อ	69

Data preparation for WordCloud

```
[171] df['Summary_new'] = df['Summary'].apply(lambda x: x.strip('\n'))
```

```
[172] dfSummary = df['Summary_new'].str.cat(sep='')
      dfSummary
```

→ นำเนื้อหาสรุปข่าวของทุกข่าวมาต่อกัน

```
Word_in_Summary = clean(dfSummary)
Word_in_Summary = [x for x in Word_in_Summary if x != '\xa0']
Word_in_Summary = [x for x in Word_in_Summary if x != '/']
Word_in_Summary = [x for x in Word_in_Summary if x != "'"]
Word_in_Summary = [x for x in Word_in_Summary if x != "."]
Word_in_Summary = [x for x in Word_in_Summary if x != '"']
Word_in_Summary = [x for x in Word_in_Summary if x != '‘']
Word_in_Summary = [x for x in Word_in_Summary if x != '’']
```

→ ลบ element ใน list ที่เป็น Stopword

```
len(Word_in_Summary)
```

```
count_Word_in_Summary = Counter(Word_in_Summary)
TopNwords = count_Word_in_Summary.most_common(200)

topword = []
counttopword = []
for item in TopNwords:
    topword.append(item[0])
    counttopword.append(item[1])

df_topword = pd.DataFrame(zip(topword, counttopword),columns=['topword','counttopword'])
df_topword.index.name = 'index'
df_topword.to_csv('df_topword.csv',encoding='utf-8')
```

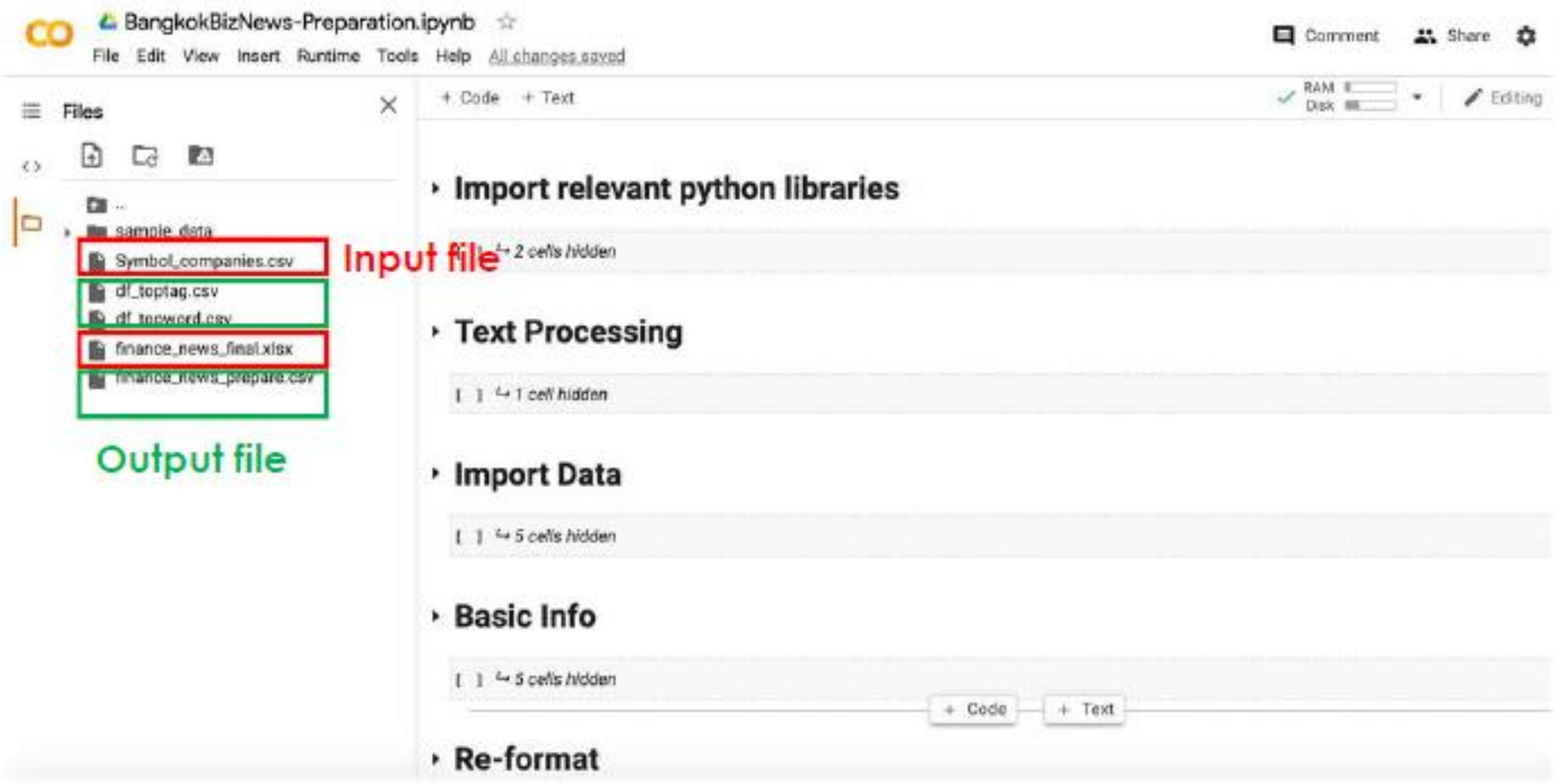
→ นับจำนวนคำที่ซ้ำ

→ เลือกเฉพาะคำที่ปรากฏซ้ำมากที่สุด 200 ลำดับแรก และ export เป็นไฟล์ csv

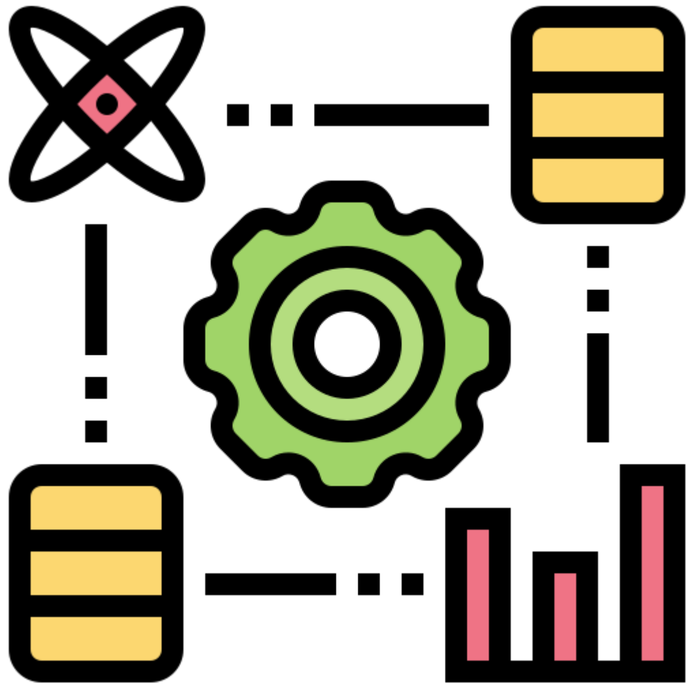
df_topword

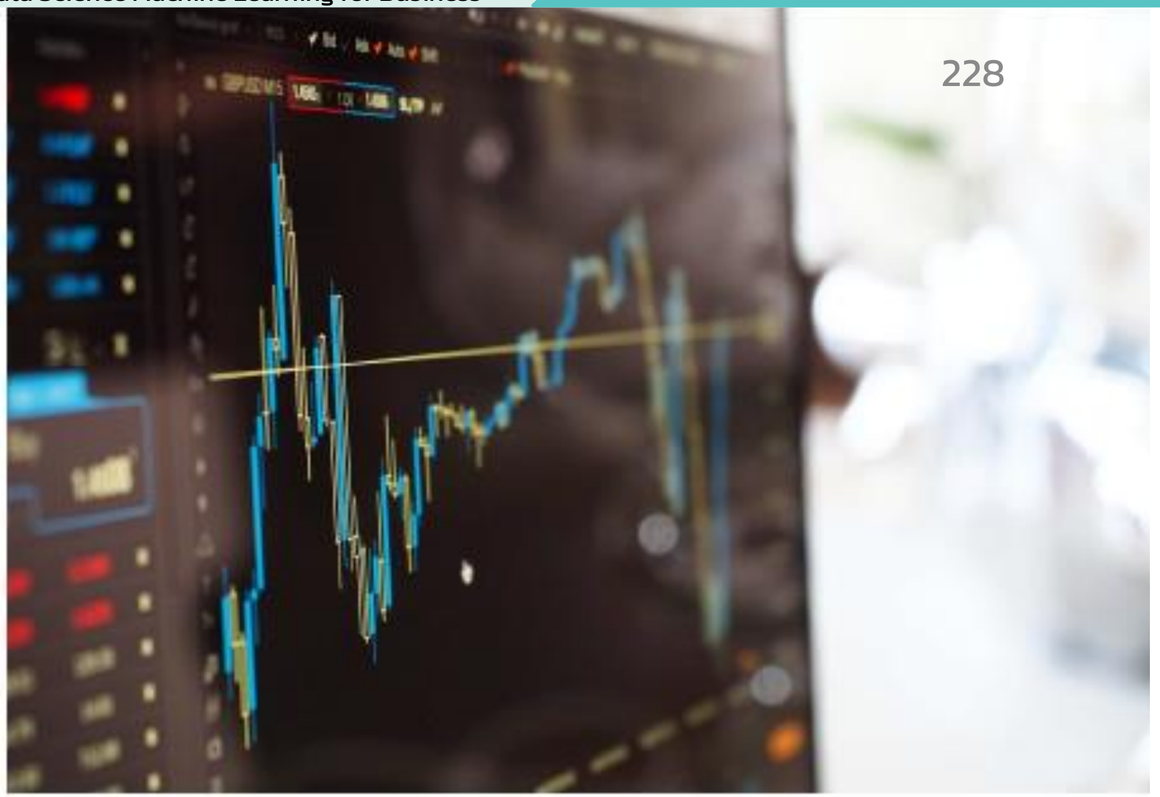
index	topword	counttopword
0	บาท	2308
1	ล้าน	1621
2	%	1447
3	หุ้น	1181
4	จุด	1088
5	ธุรกิจ	985
6	ไทย	975
7	ขาย	958

Prepared Data as output



3.5 บทที่ 5 :
Data Visualization





Python Programming for Financial Analysis

Data Visualization using MS Power BI (Financial News)

"Main goal of data visualization is to communicate information clearly and effectively through graphical means"

Vitaly Friedman (2008)

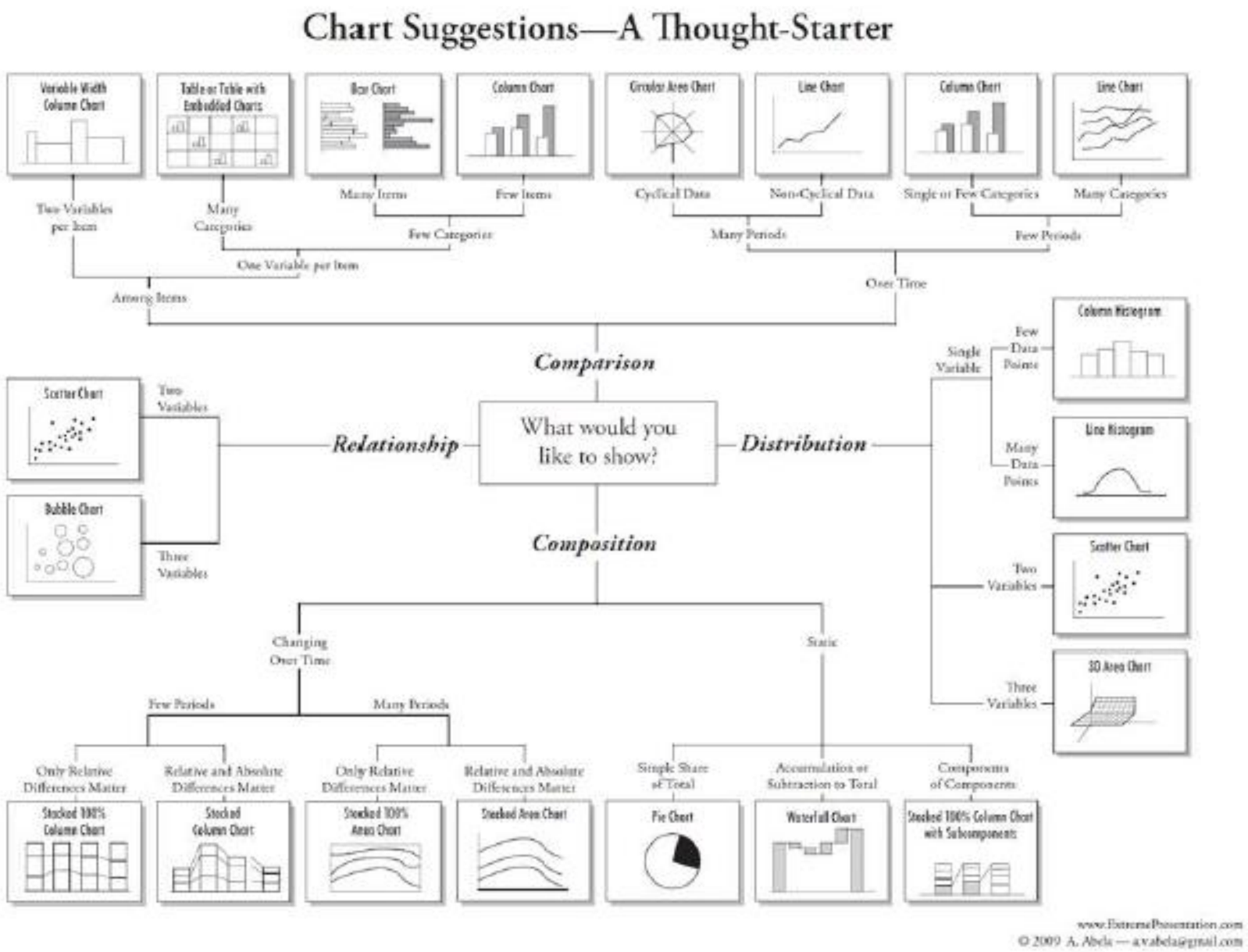
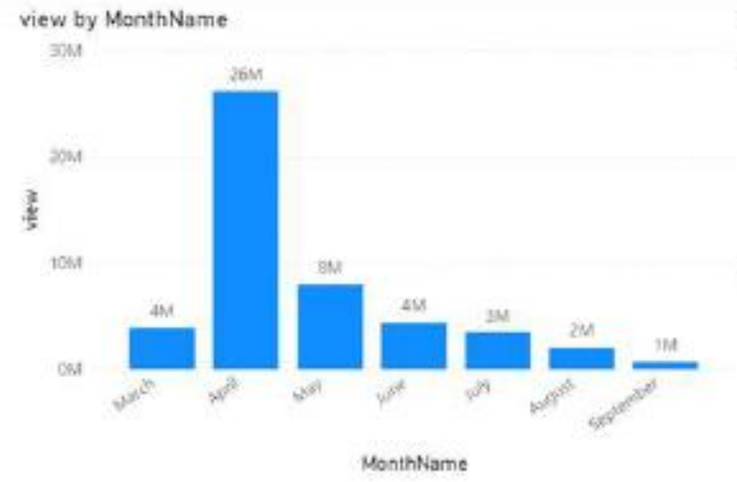


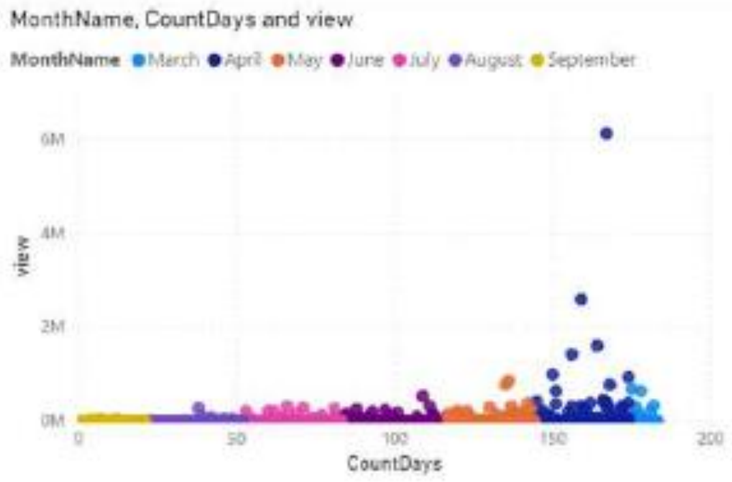
Chart Suggestions—A Thought-Starter
Variable Width Column Chart
Table or Table with Embedded Charts
Bar Chart
Column Chart
Circular Area Chart
Line Chart
Column Chart
Line Chart
Two Variables per Item
Many Categories
Many Items
Few Items
Few Categories
One Variable per Item
Cyclical Data
Non-Cyclical Data
Many Periods
Single or Few Categories
Many Categories
Few Periods
Over Time
Among Items
Comparison
Scatter Chart
Two Variables
Bubble Chart
Three Variables
Relationship
What would you like to show?
Distribution
Composition
Single Variable
Few Data Points
Column Histogram
Many Data Points
Line Histogram
Two Variables
Scatter Chart
Three Variables
3D Area Chart
Changing Over Time
Static
Few Periods
Many Periods
Only Relative Differences Matter
Relative and Absolute Differences Matter
Only Relative Differences Matter
Relative and Absolute Differences Matter
Simple Share of Total
Accumulation or Subtraction to Total
Components of Components
Stacked 100% Column Chart
Stacked Column Chart
Stacked 100% Area Chart
Stacked Area Chart
Pie Chart
Waterfall Chart
Stacked 100% Column Chart with Subcomponents
www.ExtremePresentation.com
© 2009 A. Abela — a.v.abela@gmail.com

Power BI

Example of Power BI Report

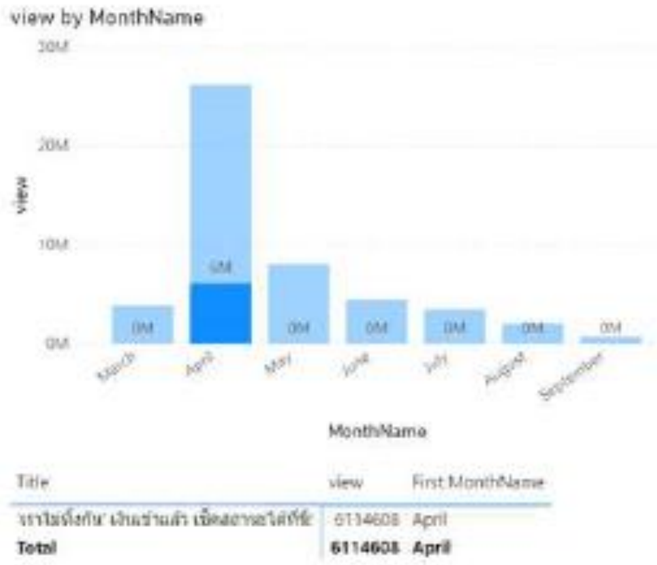
Visualization Report

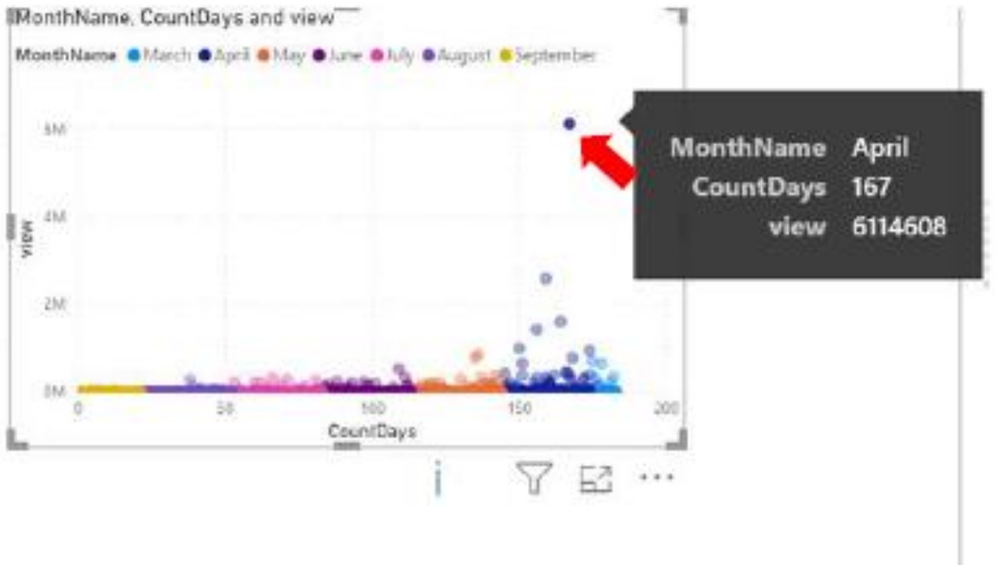




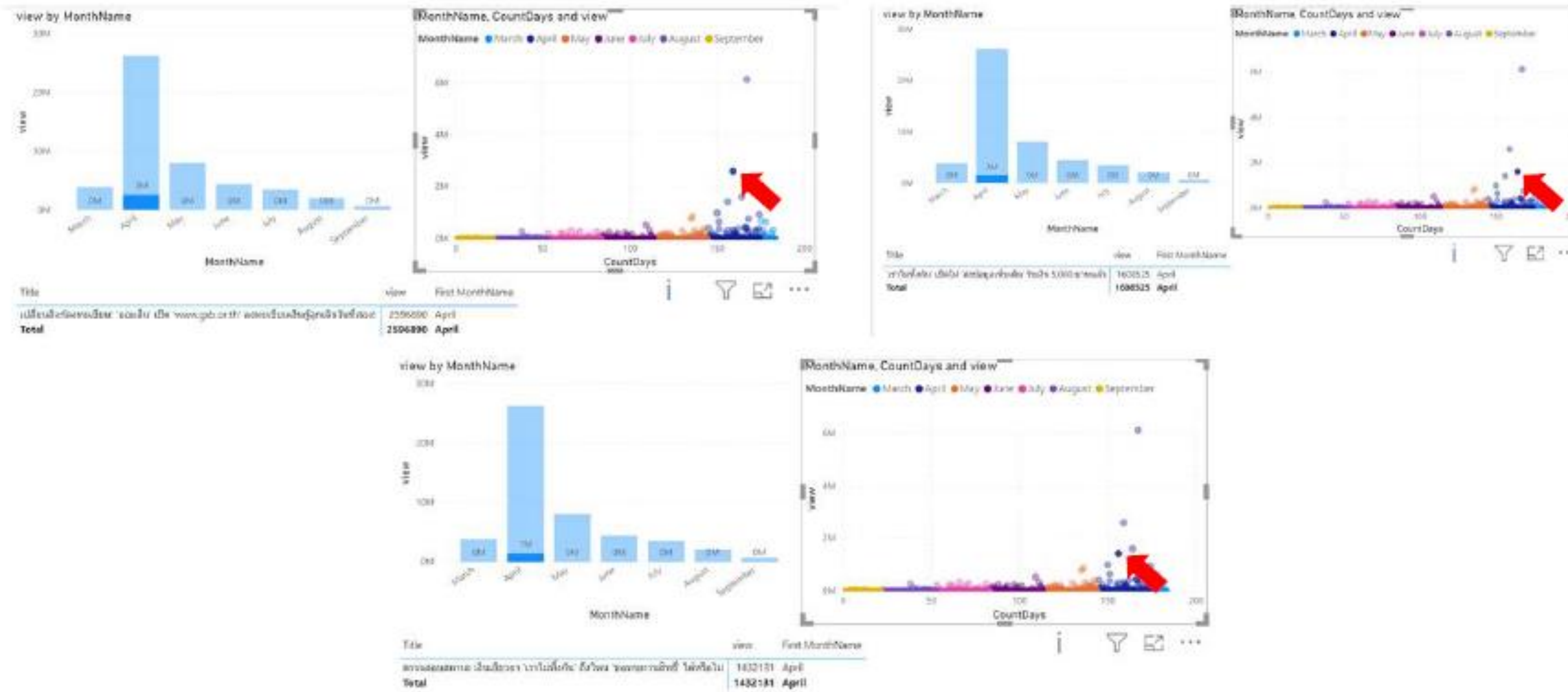
Title	view	First MonthName
'เขย่าค่าที่ธรกิจโบรกเกอร์ สะท้อนดีกว่าการลงชื่อเพิ่มขึ้น	186	September
'เคทีซี-อิออน ' จอดคลอดเบี้ยต่ำ ช่วยลูกค้าภาวะวิกฤติ	2273	April
'เงินบาท' เปิดตลาดเช้านี้ แข็งค่า ที่ 30.89 บาทต่อดอลลาร์	291	June
'เงินบาท' เปิดตลาดเช้านี้ แข็งค่า ที่ 32.63 บาทต่อดอลลาร์	529	March
'เงินบาท' เปิดตลาดเช้านี้ อ่อนค่า ที่ 30.93 บาทต่อดอลลาร์	295	June
'เงินบาท' ทุบสถิติใหม่ 'อ่อนค่าสุด' ในรอบ16ด. ทะลุ 33บาทต่อดอลลาร์	961	March
'เงินบาท' ปิดตลาด แข็งค่าที่ 31.23 บาทต่อดอลลาร์	250	July
'เงินบาท' ปิดตลาด แข็งค่าที่ 32.33 บาทต่อดอลลาร์	608	April
'เงินบาท' ปิดตลาด แข็งค่าที่ 32.40 บาทต่อดอลลาร์	411	April
'เงินบาท' ปิดตลาด ทรงตัวที่ 32.47 บาทต่อดอลลาร์	355	April
'เงินบาท' ปิดตลาด อ่อนค่าที่ 31.14 บาทต่อดอลลาร์	236	August
'เงินบาท' ปิดตลาด อ่อนค่าที่ 31.22 บาทต่อดอลลาร์	236	July
'เงินบาท' ปิดตลาด แข็งค่าที่31.01 บาทต่อดอลลาร์	307	June
'เงินบาท' ปิดตลาด ทรงตัวที่ 31.11บาทต่อดอลลาร์	271	June
Total	**48841931**	**April**

Visualization Report





Visualization Report



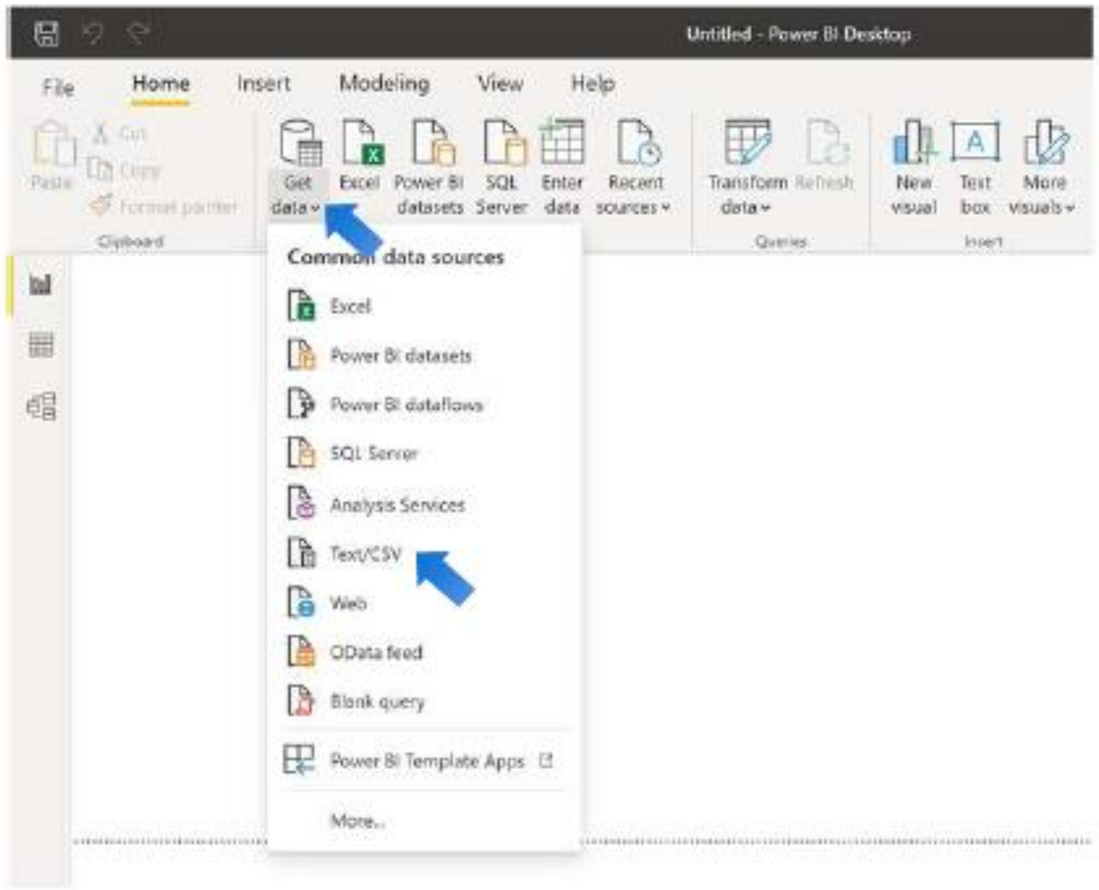
Visualization Report

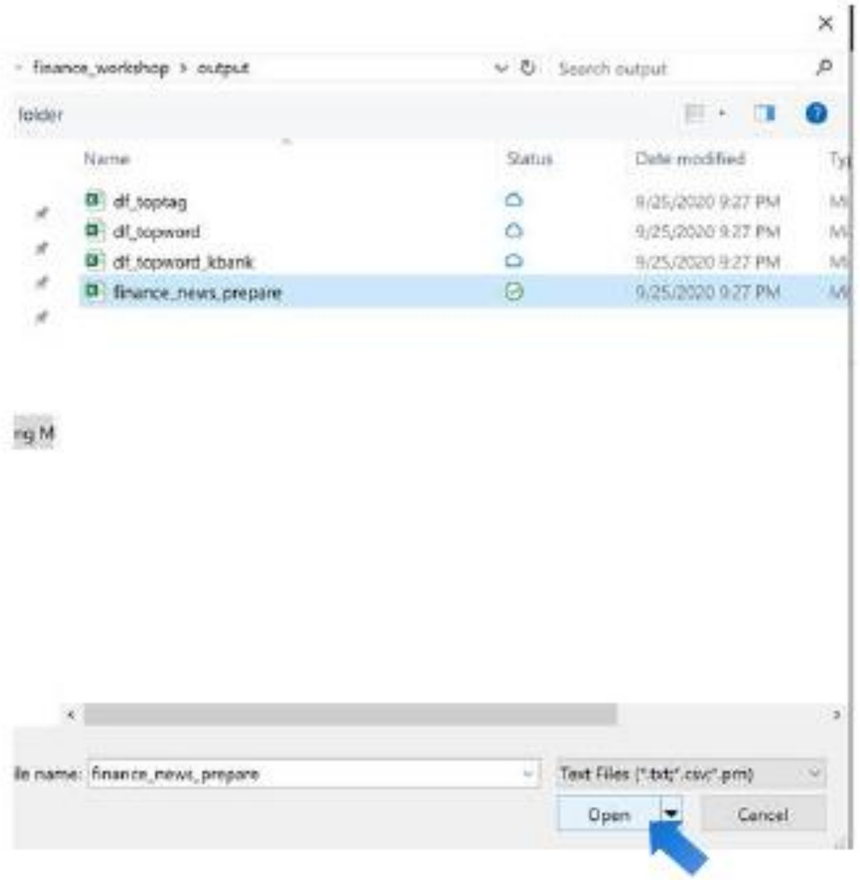
จากสไลด์ก่อนหน้า Power BI report แสดงจำนวนการเข้าชมข่าว ใน Visualization แบบต่าง ๆ ซึ่งระบุว่า

- ข่าวที่มีจำนวนผู้เข้าชมมาก เป็นข่าวในเดือนเมษายน
- ข่าวที่มีจำนวนผู้ชมสูงแบบโดดเด่น เป็นข่าวที่มีเนื้อหาเกี่ยวกับนโยบาย เราไม่ทิ้งกัน

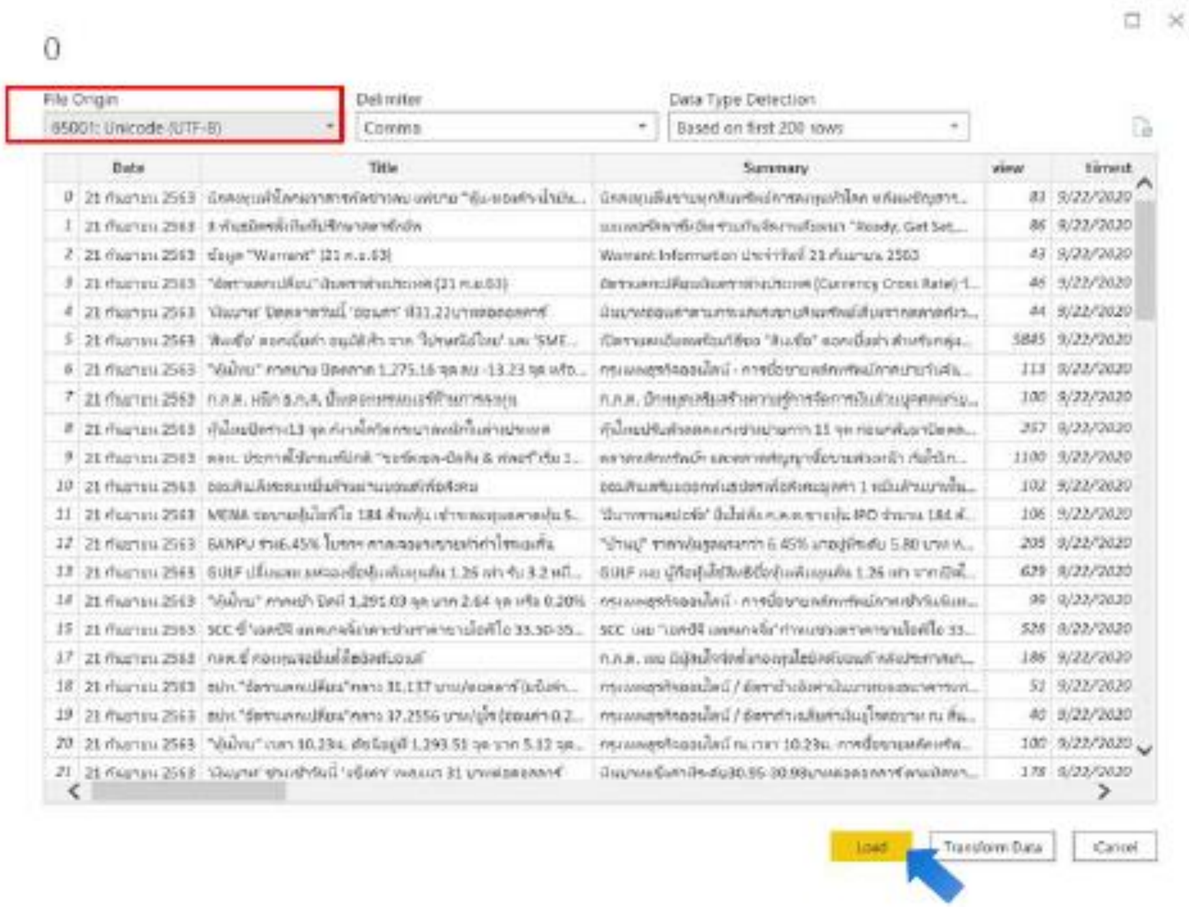
Let's create Power BI Report

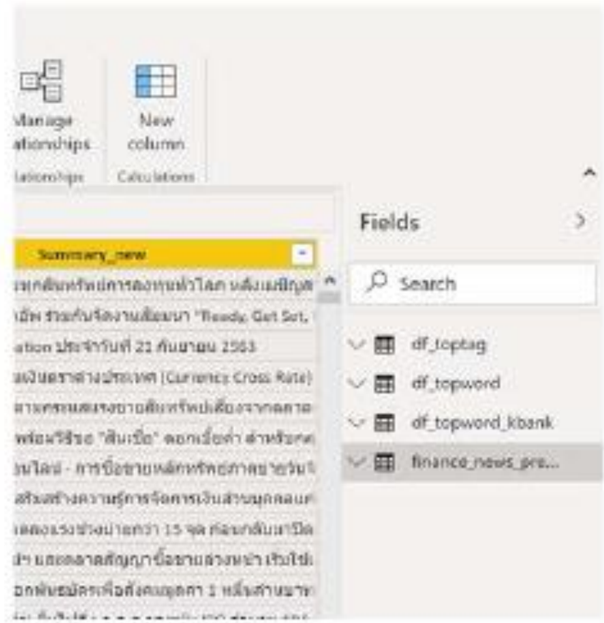
Get Data


Untitled - Power BI Desktop
File
Home
Insert
Modeling
View
Help
Paste
Cut
Copy
Format painter
Clipboard
Get data
Excel
Power BI datasets
SQL Server
Enter data
Recent sources
Transform data
Refresh
Queries
New visual
Text box
More visuals
Insert
Common data sources
Excel
Power BI datasets
Power BI dataflows
SQL Server
Analysis Services
Text/CSV
Web
OData feed
Blank query
Power BI Template Apps
More...

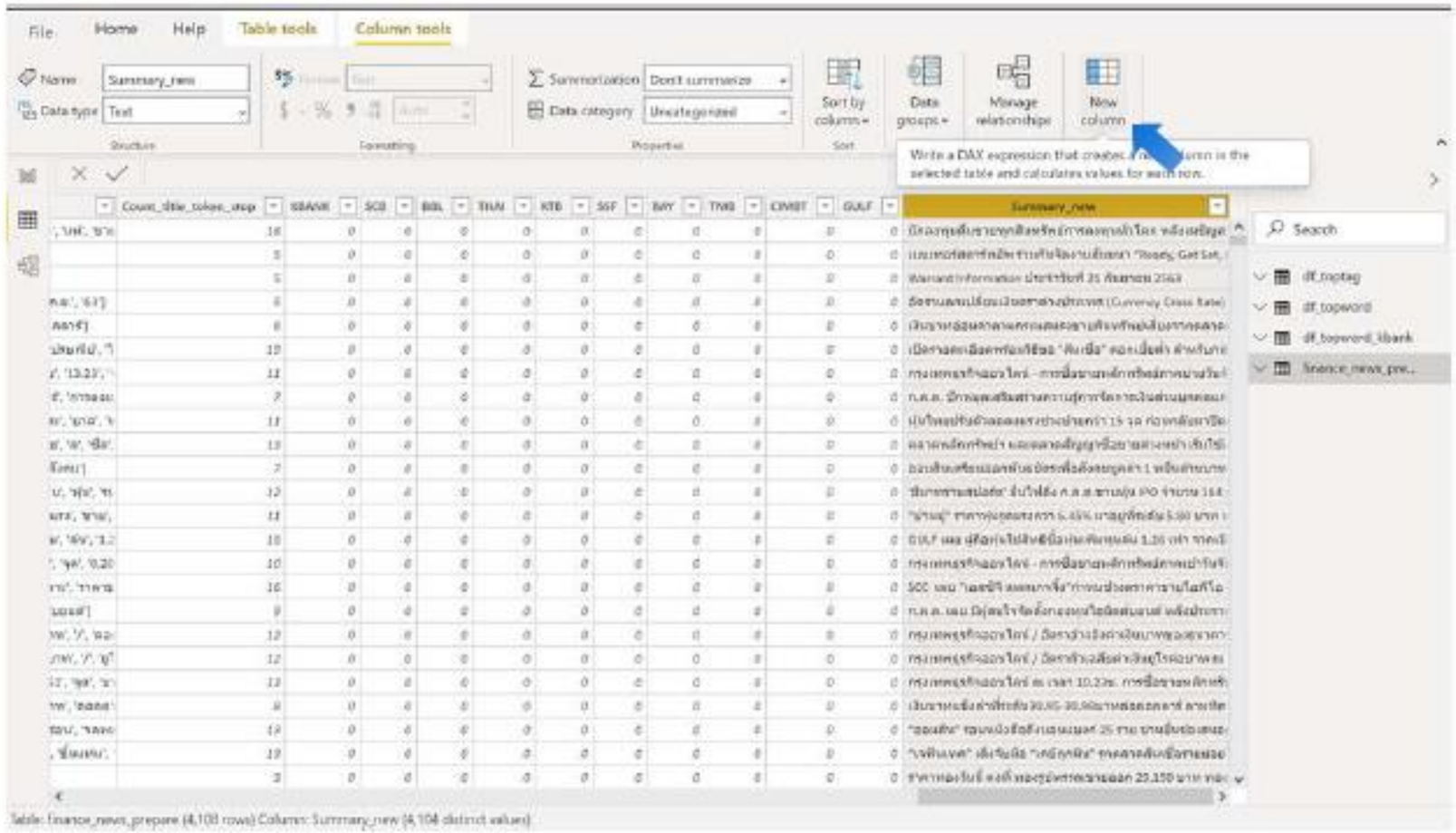

finance_workshop > output
Search output
folder
Name
Status
Date modified
df_toptag
df_topword
df_topword_kbank
finance_news_prepare
9/25/2020 9:27 PM
File name: finance_news_prepare
Text Files (*.txt;*.csv;*.prn)
Open
Cancel

Get Data


0
File Origin
65001: Unicode (UTF-8)
Delimiter
Comma
Data Type Detection
Based on first 200 rows
Load
Transform Data
Cancel

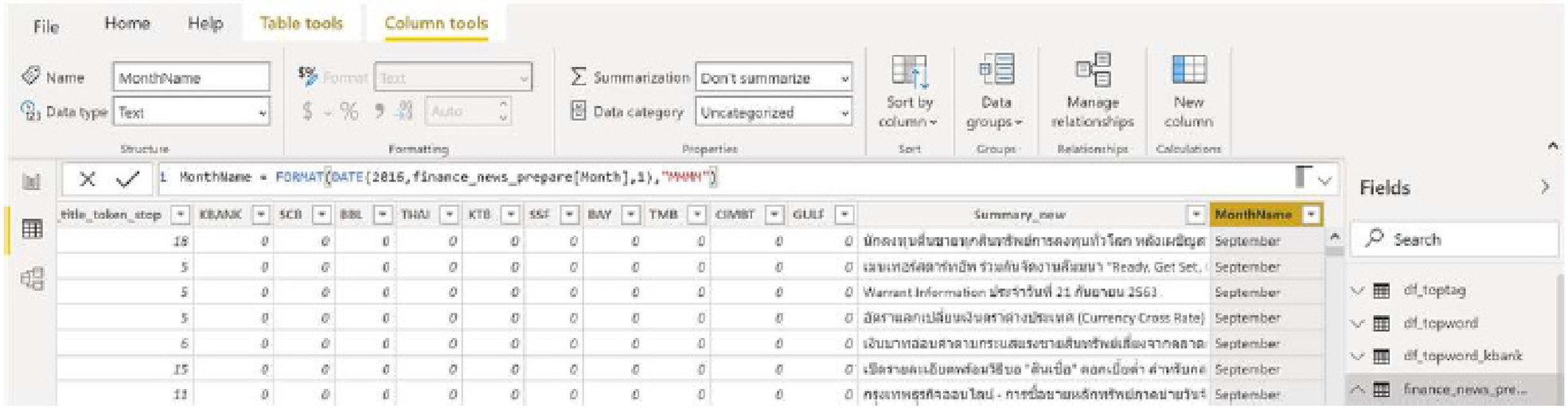

Manage relationships
New column
Calculations
Fields
Search
df_toptag
df_topword
df_topword_kbank
finance_news_pre...

Create MonthName Column



Create MonthName Column (cont.)

```
MonthName = FORMAT(DATE(2016,finance_news_prepare[Month],1),"MMMM")
```



_title_token_stop	KBANK	SCB	BBL	THAI	KTB	SSF	BAY	TMB	CIMBT	GULF	Summary_new	MonthName
18	0	0	0	0	0	0	0	0	0	0	นักลงทุนสถาบันชาวยุโรปมีมุมมองเชิงบวกต่อการลงทุนในตลาดหุ้นไทย หลังเหตุการณ์	September
5	0	0	0	0	0	0	0	0	0	0	เลขาธิการ ก.ล.ต. ร่วมกับ จัดงานสัมมนา "Ready, Get Set,	September
5	0	0	0	0	0	0	0	0	0	0	Warrant Information ประจำวันที่ 21 กันยายน 2563	September
5	0	0	0	0	0	0	0	0	0	0	อัตราแลกเปลี่ยนเงินตราต่างประเทศ (Currency Cross Rate)	September
6	0	0	0	0	0	0	0	0	0	0	เงินบาทอ่อนค่าตามกระแสแรงขายสินทรัพย์เสี่ยงจากกลยุทธ์	September
15	0	0	0	0	0	0	0	0	0	0	เปิดรายละเอียดพร้อมวิธีขอ "คืนเงิน" ดอกเบี้ยต่ำ ค่าจับกุ	September
11	0	0	0	0	0	0	0	0	0	0	กรุงเทพธุรกิจออนไลน์ - การซื้อขายหลักทรัพย์ภาคบ่ายวันที่	September

FORMAT

Converts a value to text according to the specified format.

Syntax

```
FORMAT(<value>, <format_string>)
```

DATE

Returns the specified date in datetime format.

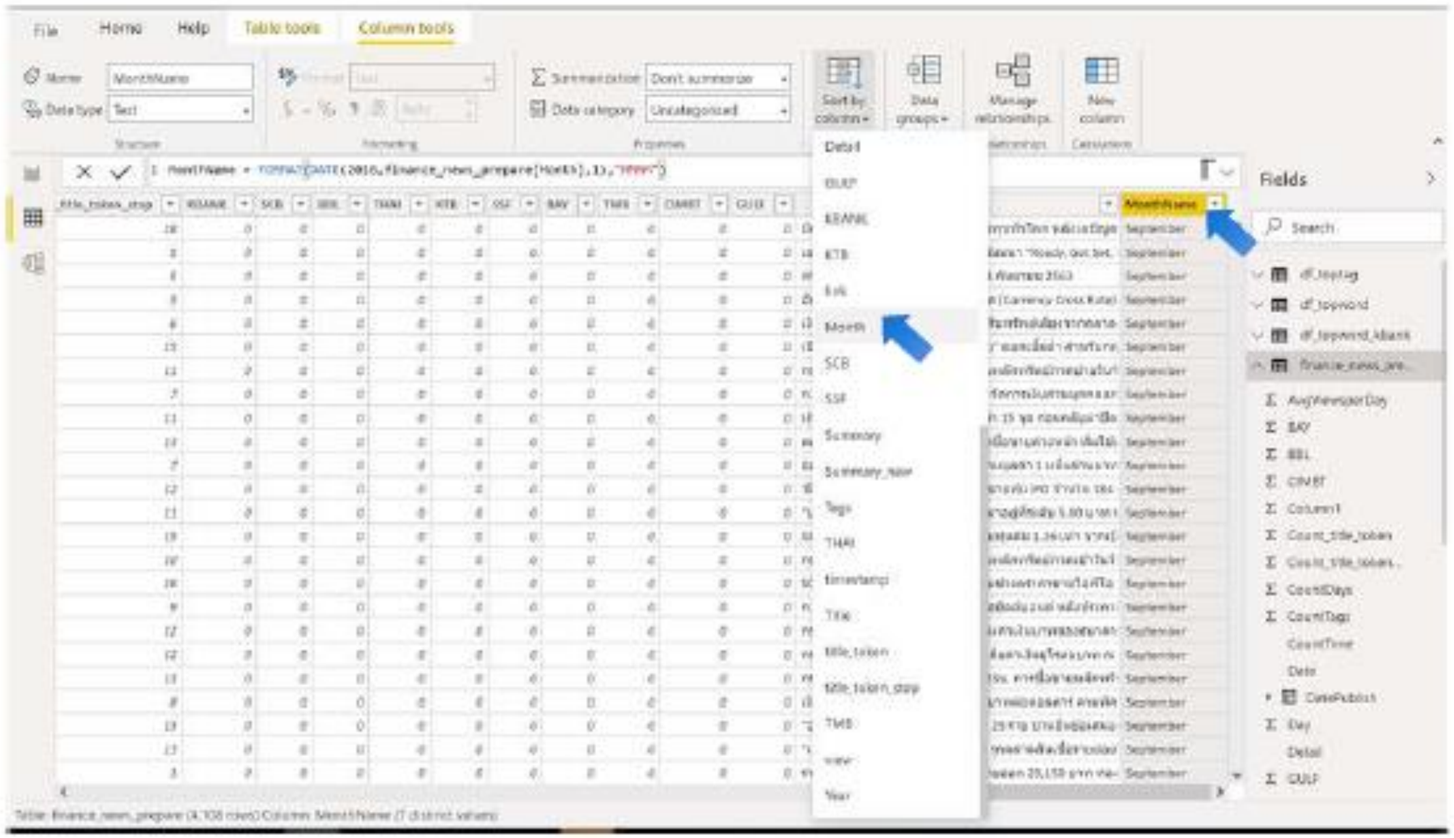
Syntax

```
DATE(<year>, <month>, <day>)
```

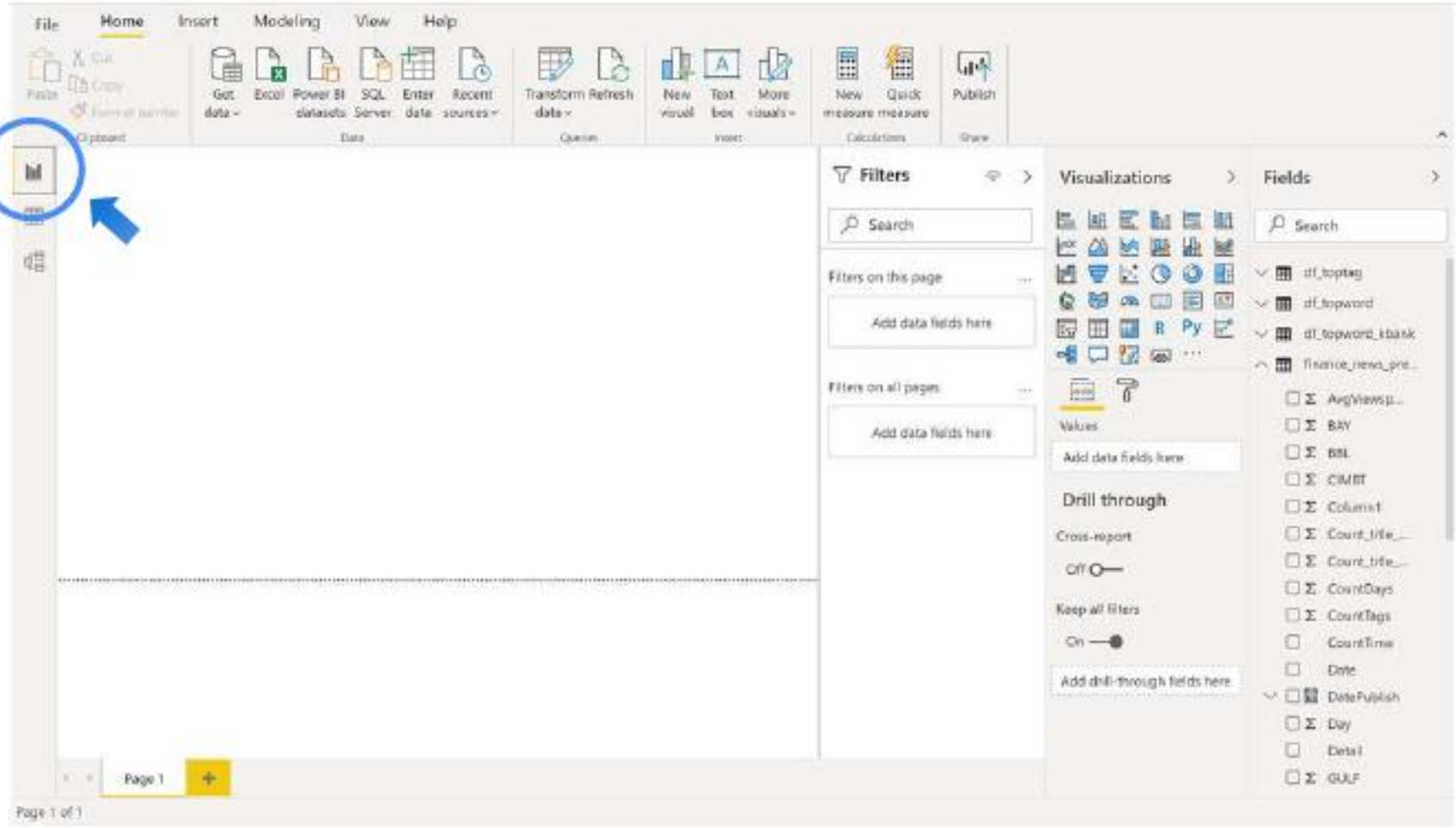
Format Date/Time: https://docs.microsoft.com/en-us/system-center/orchestrator/standard-activities/format-date-time?view=sc-orch-2019

Create MonthName Column (cont.)

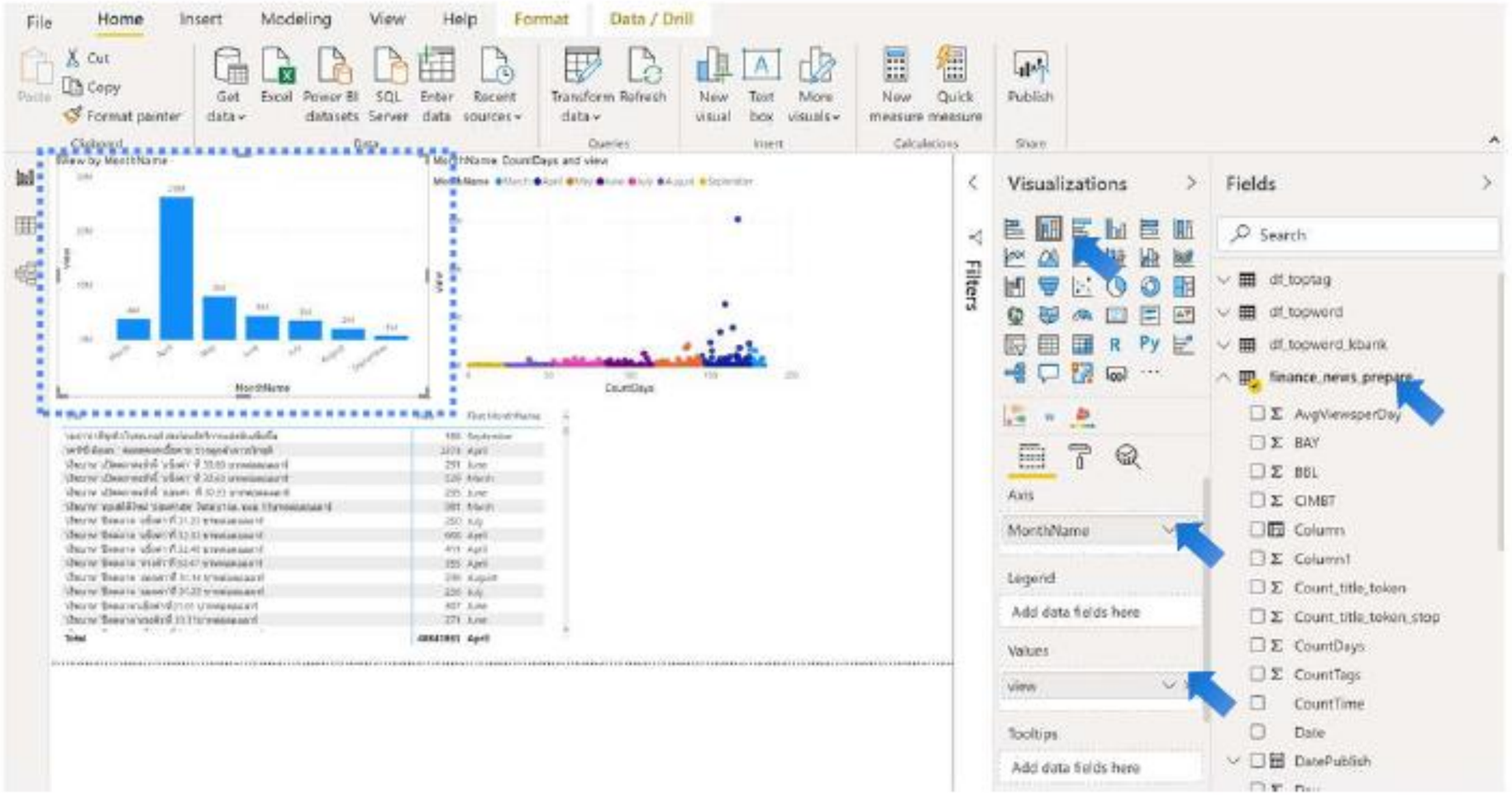
Sort column MonthName by Month



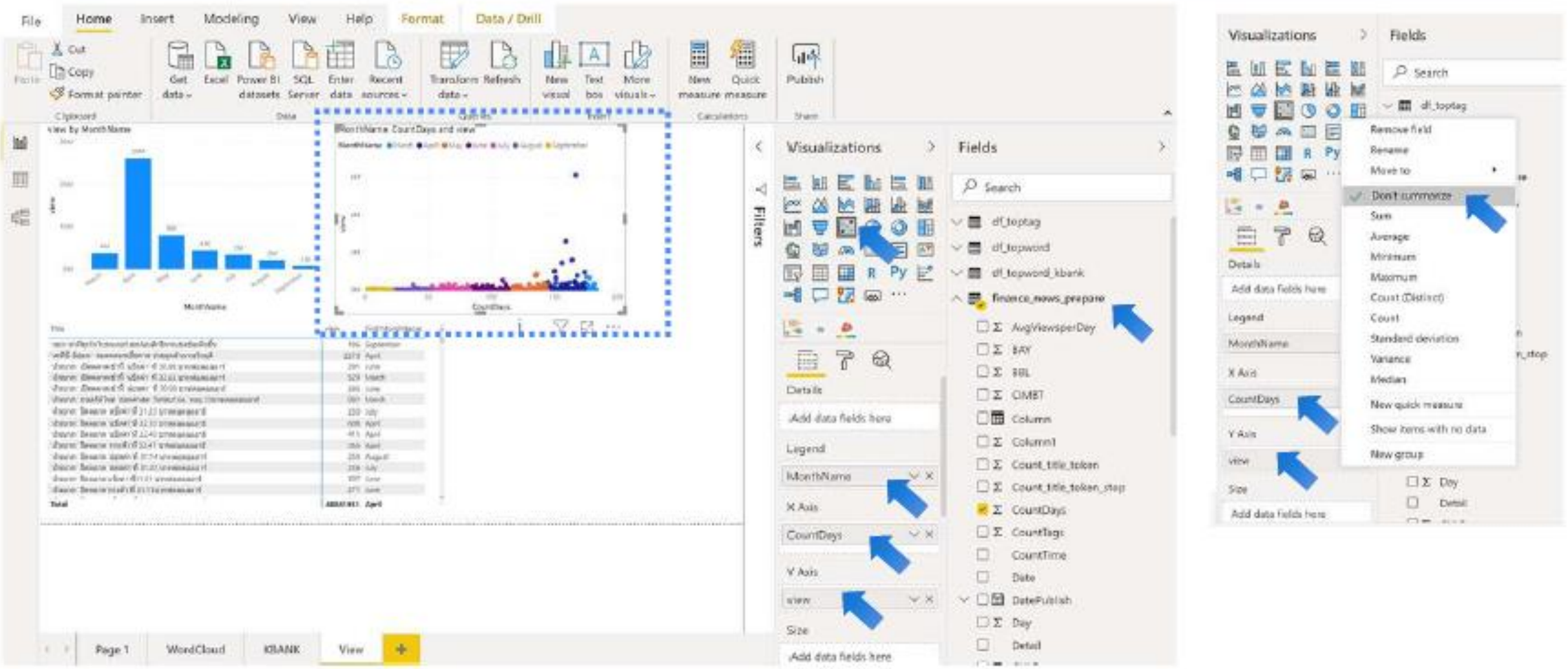
Create Visualization Report



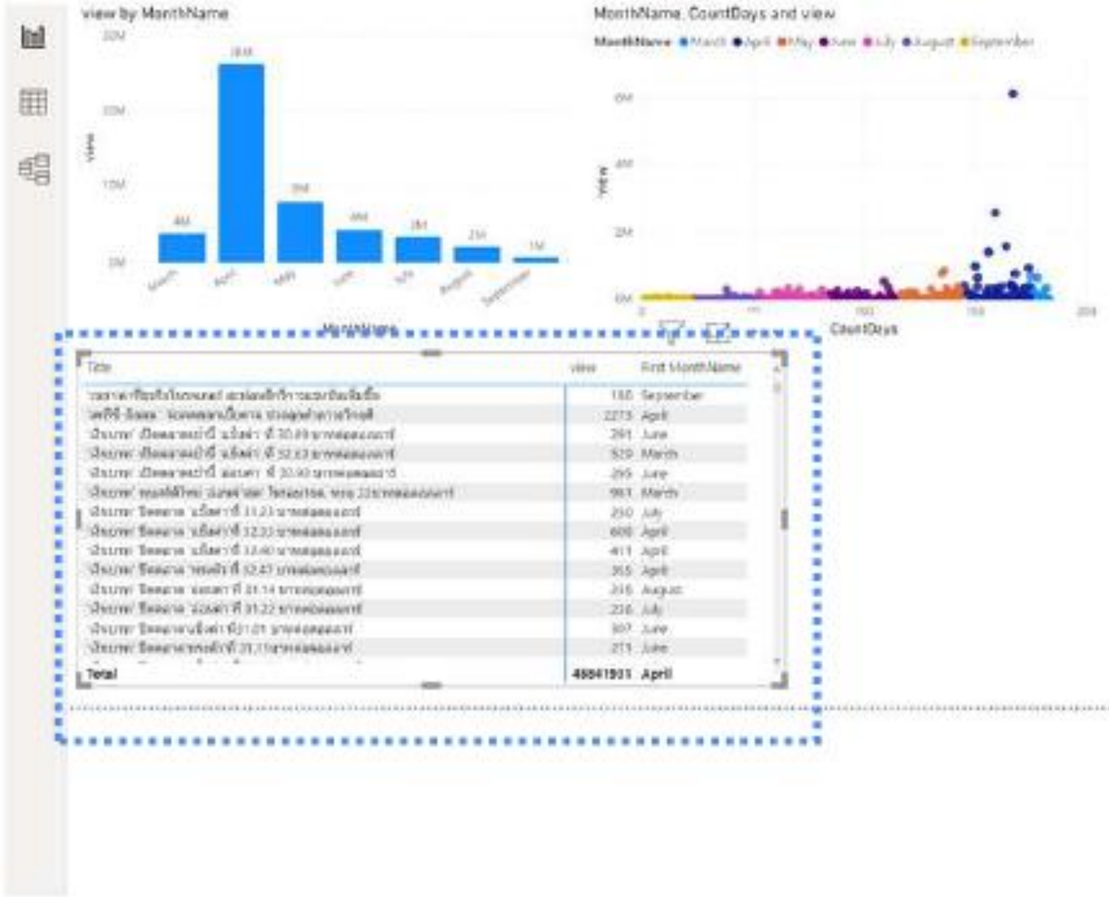
Create Visualization Report

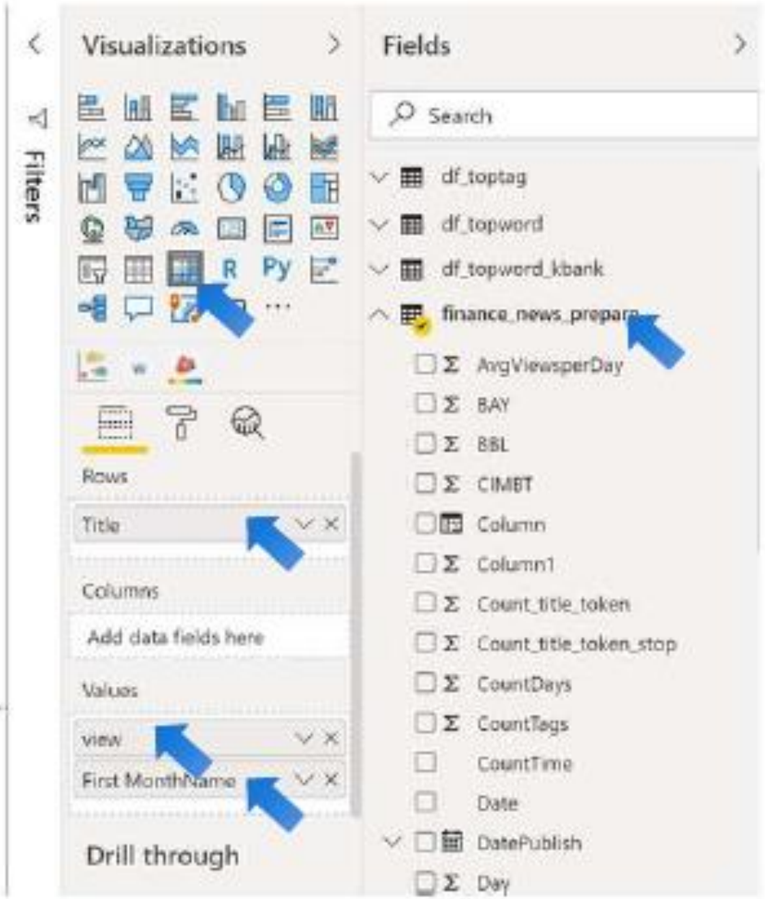


Create Visualization Report



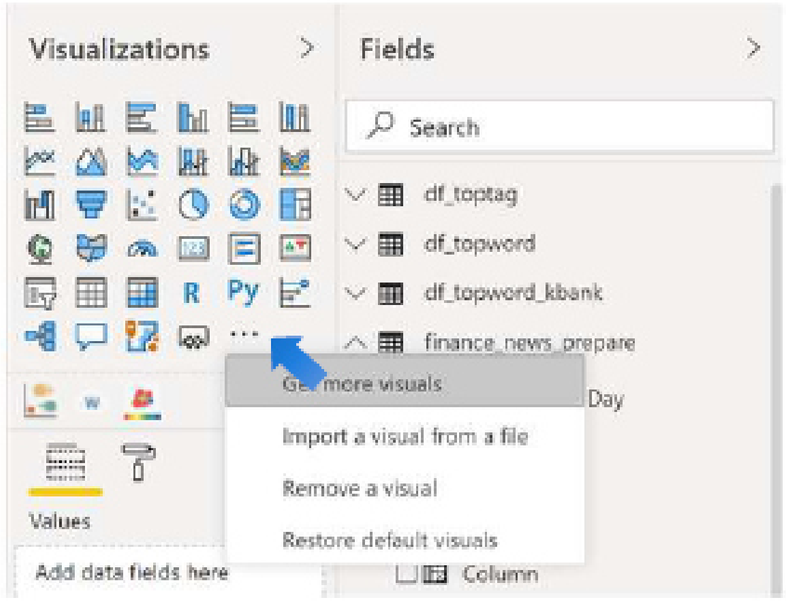
Create Visualization Report

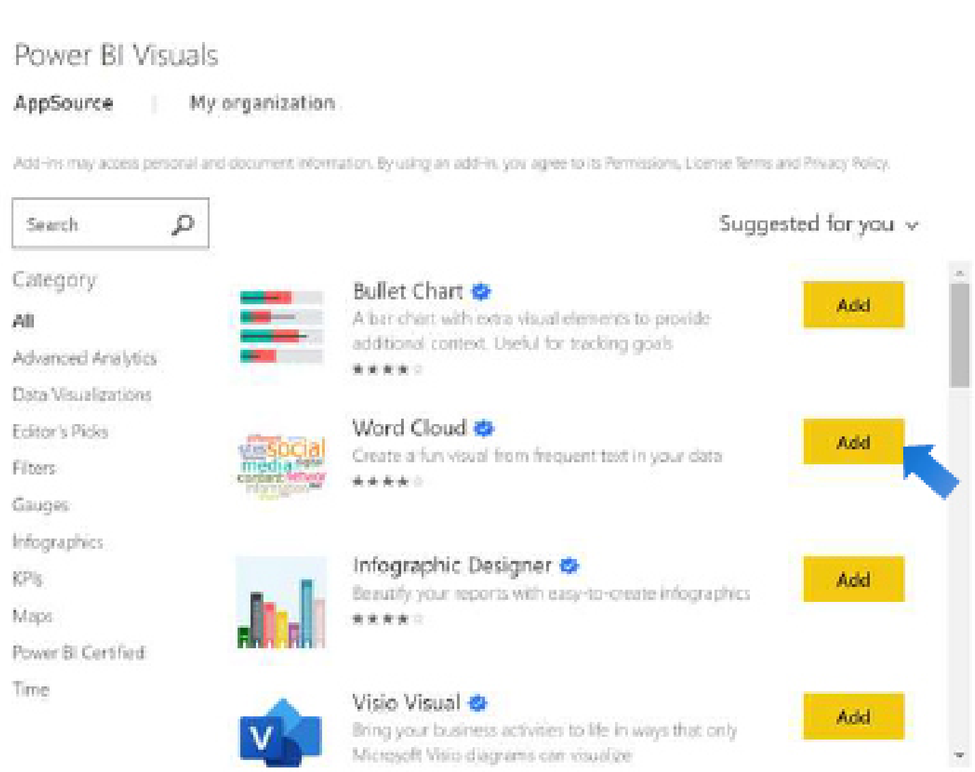




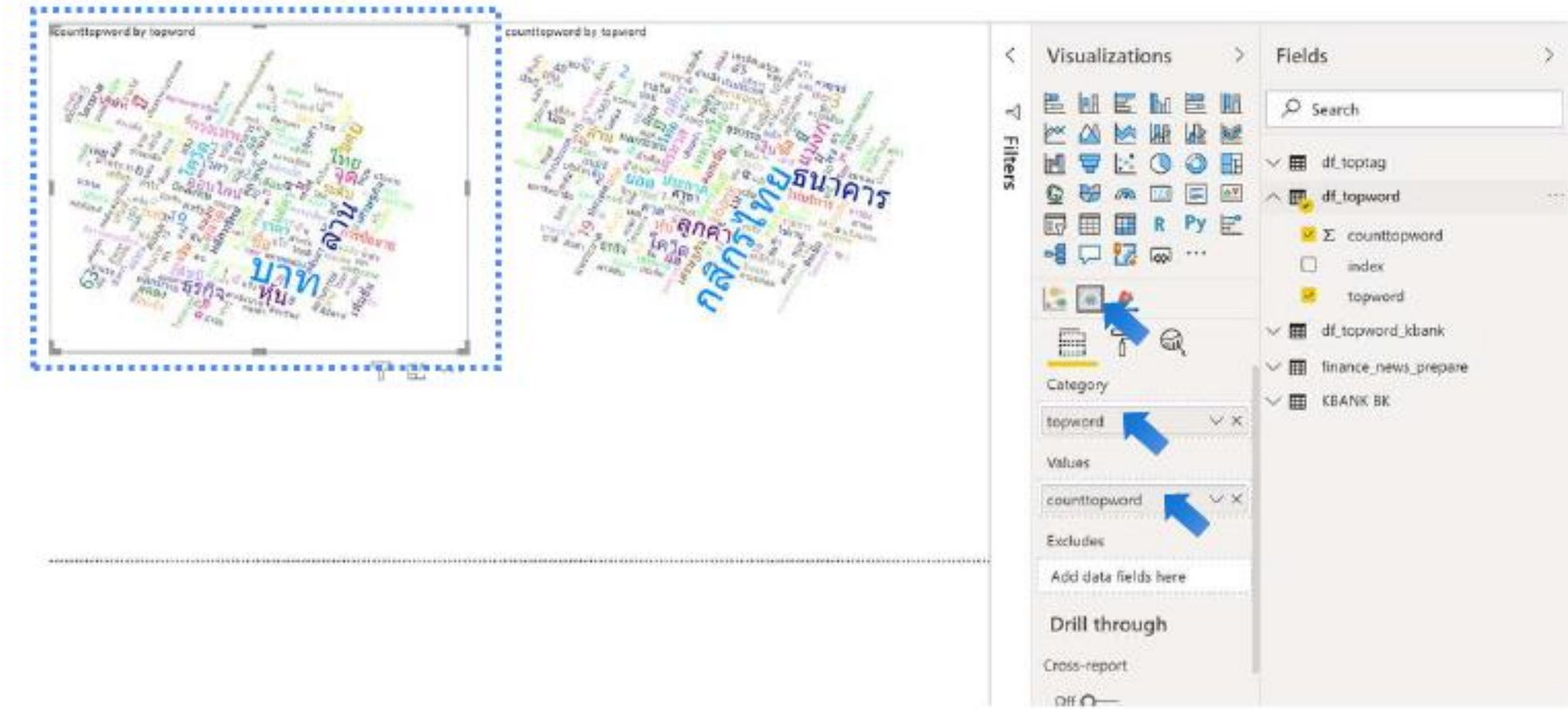
WordCloud in Power BI

Create Visualization Report

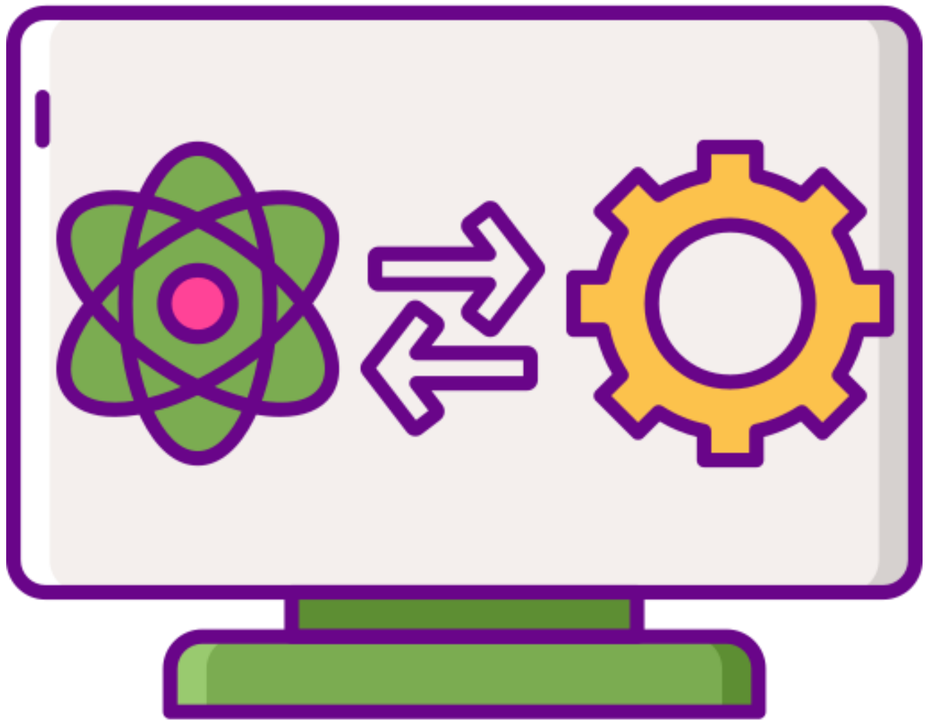




Create Visualization Report



3.6 บทที่ 6 :
Introduction to Natural Language Processing



Introduction to Data Analytics

Natural Language Processing (NLP)

- How can we make a computer understand the following text.

ประเทศไทยรวมเลือดเนื้อชาติเชื้อไทย

เป็นประชารัฐ ไผทของไทยทุกส่วน

อยู่ดำรงคงไว้ได้ทั้งมวล

ด้วยไทยล้วนหมาย รักสามัคคี

ไทยนี้รักสงบแต่ถึงรบไม่ขลาด

เอกราชจะไม่ให้ใครข่มขี่

- How can we communicate with computer using natural language?

Natural Language Processing (NLP)

- NLP is a discipline concerning the interaction between computer and human language using computer science and artificial intelligence

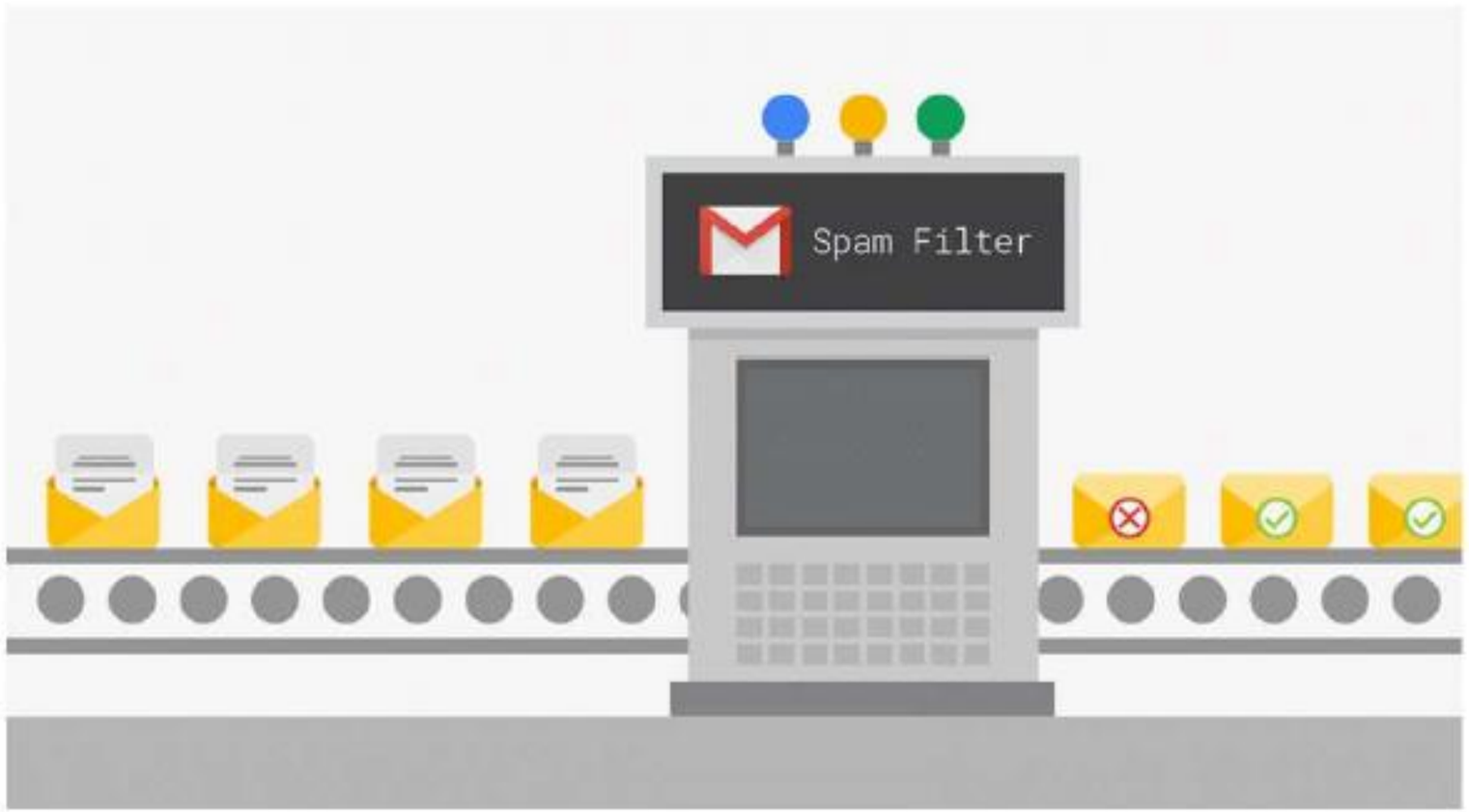
- Generally, it is an application of machine learning on text and speech

- Example of task
  - Text summarization
  - Machine translation
  - Question answering
  - Autocomplete

Example of usage

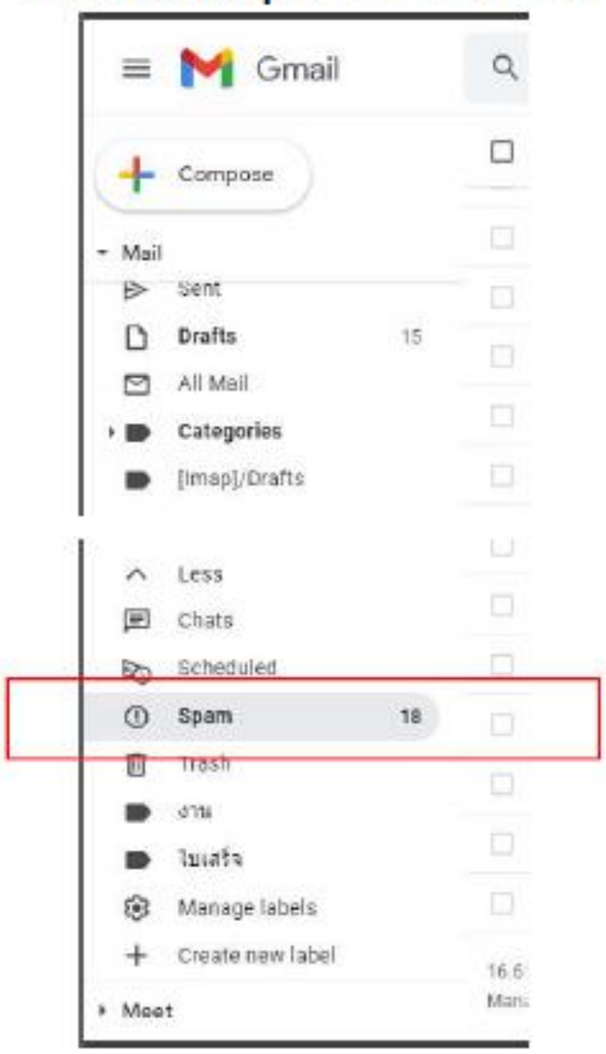
https://www.youtube.com/watch?v=equL39AYANo

Example of usage

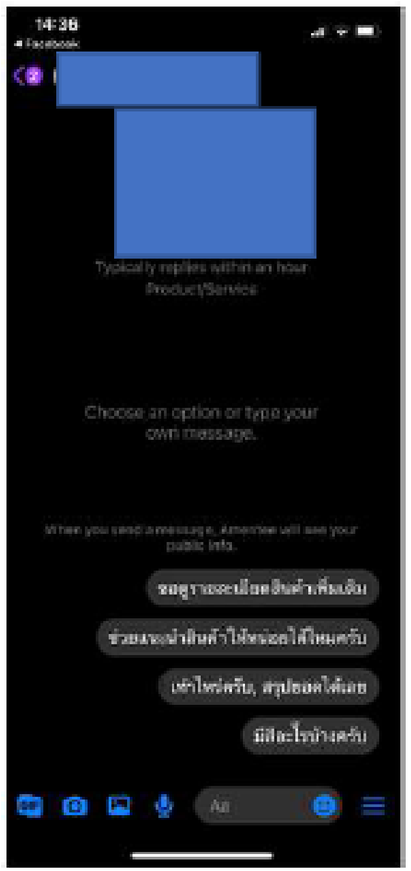


https://i.gifer.com/Ou1t.gif

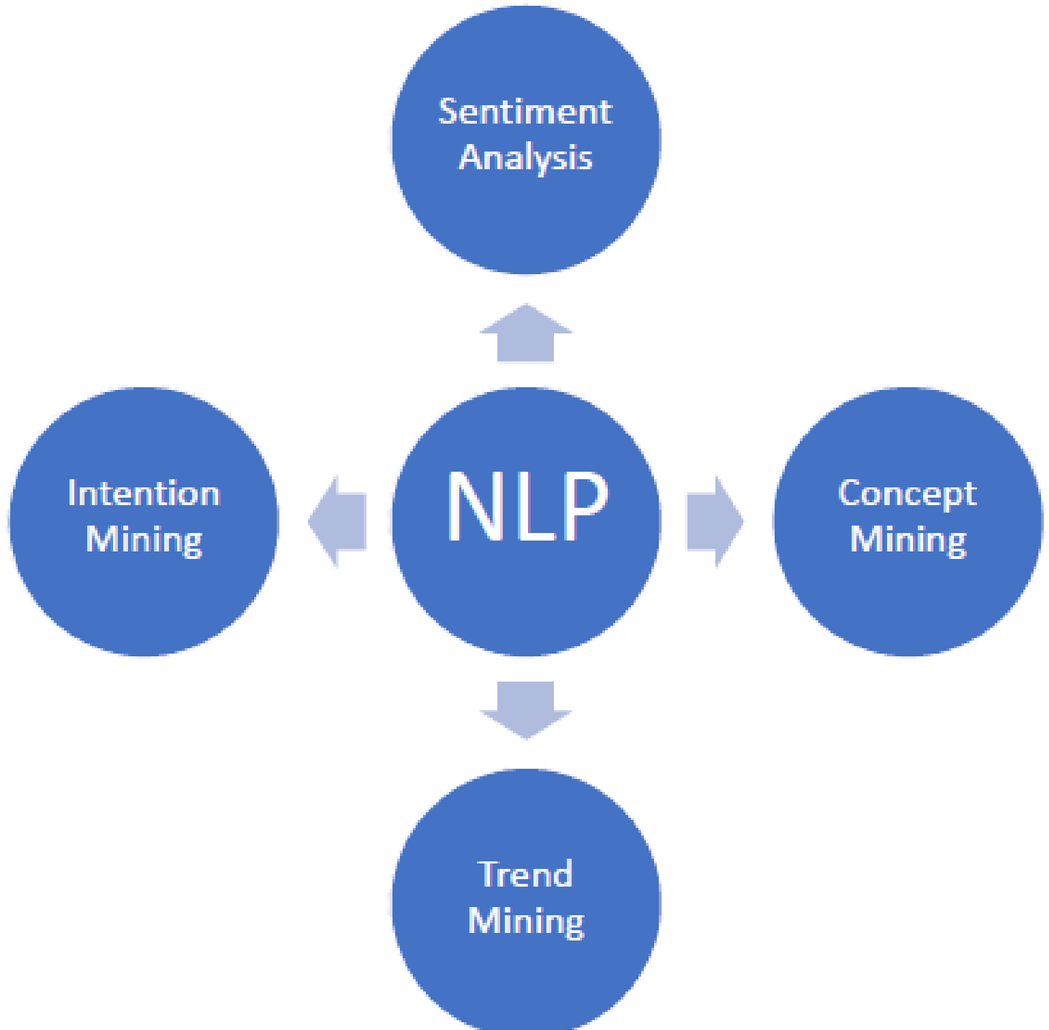
Example of usage



Example of usage



Purpose of natural language processing



Purpose of natural language processing

- Intention Mining
  - Aims to discover the user intention based on comment, review, tweets or blog.
  - Company use this technique to find new potential customers
- Trends Mining
  - Use historical data and social media data to predict future event.
  - Newly emerge

Purpose of Text Analytics

- Concept Mining
  - Extract ideas and concept from large static social media
  - E.g. clustering document

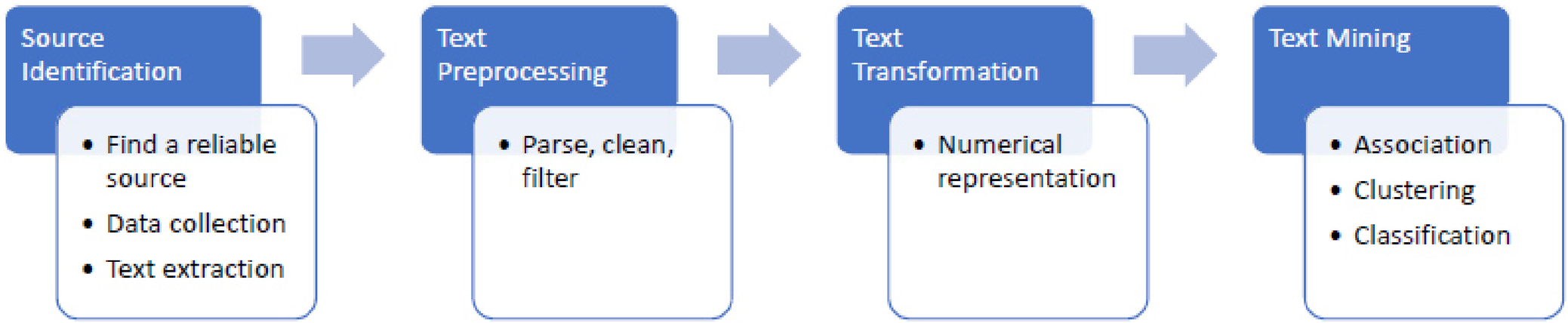
- Sentiment Analysis
  - Categorize the text to be positive (+), negative (-) or neutral (0).
  - The company often use this technique to determine how the customer feel about the product.

Types of Text

- Static Text
  - Purpose is to inform, educate or elaborate.
  - Typically, static text is very large.
  - E.g. Wiki, Blog, Email

- Dynamic Text
  - Generated by user to express an opinion
  - Shorter in length
  - E.g. Comment

Social Media Text Analytics



Sentiment Analysis

Opinion extraction, Opinion mining, Sentiment mining, Subjectivity analysis

Sentiment Analysis

- A data mining process to classify a text into categories : “Positive”, “Negative” or “Neutral”.

- Not 100% accurate marker.

- E.g. what is your temperature for “cold” ?

Sentiment Analysis

- Customers
  - To research products before purchasing.
- Marketers
  - To research opinion of their company or products
  - Analyze customer satisfaction

Sentiment Analysis

โดย Lazada Customer ตรวจสอบว่ามีการสั่งซื้อแล้ว

สั่งตอนโปร 9.09 ได้ของวันที่ 18.09 แพ็คอย่างดี เครื่องใช้งายสะดวก ลิงค์กับแอ๊พ Xiaomi Home แล้วยิ่งสะดวก เปิด-ปิดเครื่องผ่านแอ๊พในมือถือได้เลย



มาส่งเร็วเกิน เกินที่แจ้งไว้ 2 วัน เตรียมเงินเกือบไม่พอ แต่ก็ดี อยากทดลองเร็วๆ ขนส่งแพคมาดี ไม่บุบสลาย สภาพสินค้า คุณภาพดี การใช้งานครั้งแรกค่อนข้างง่าย รอดูผลการกรองอากาศระยะยาวอีกที

Color Family:ขาว

1






โดย Lazada Customer ตรวจสอบว่ามีการสั่งซื้อแล้ว

ได้มากล่องขาด เปิดใช้งานไม่ได้ เครื่องพัง แจ้งแอดมินไป
เห็นบอกว่าถ้ามีเครื่องใหม่จะเปลี่ยนให้
แต่ถึงตอนนี้ยังไม่ได้คำตอบเลย
สุดท้ายต้องคืนทาง LAZADA ไป


สี:"ขาว"

โดย Lazada Customer ตรวจสอบว่ามีการสั่งซื้อแล้ว

ได้ของช้ามาก สั่งวันที่ 10 วันที่ 17 ยังไม่ได้ คนส่งโทรมาบอกว่าต้องขอเซนรับแทนก่อน เดี๋ยวฟัสดุดิกลับ แล้วจะมาส่งวันที่ 18

สี:"ขาว"

0

Sentiment Analysis: Example

" jaws " is a rare film that grabs your attention before it shows you a single image on screen . the movie opens with blackness , and only distant , alien-like underwater sounds . then it comes , the first ominous bars of composer john williams' now infamous score . dah-dum .

Data : http://www.cs.cornell.edu/people/pabo/movie-review-data,

Sentiment Analysis: Example

you know what to expect from this movie : lots of shots where the camera is the eyes of the predator (croc cam) swimming toward someone's dangling legs while " jaws " -like music plays , one character (hector) who's obsessed with the croc and stupidly endangers the rest , another character who insists that the predator can't possibly exist . unlike its slippery cousin " anaconda , " " lake placid " wants to present its formulaic plot tongue-in-cheek , which is self-defeating . the result is neither scary nor funny ; it's just tedious .

Data : http://www.cs.cornell.edu/people/pabo/movie-review-data/

Approaches to the Sentimental Analysis

Approaches to the Sentimental Analysis

- Machine Learning
  - Classification

- General Approach
  1) Get annotated documents from your domain
  2) Convert he document to bag-of-words
  3) Train the model to recognize the class
  4) Apply the train classifier to the test case or application

Data Preparation

String Tokenization

"A process to divide a sequence of entities of a written language into entities"

- It can be applied with both articles and sentences.
  - Article ===> sentences
  - Sentence ===> words
- English is simpler than Thai.

Example of sentence tokenizer

Thailand, officially the Kingdom of Thailand and formerly known as Siam, is a country in Southeast Asia. Located at the centre of the Indochinese Peninsula, it is composed of 76 provinces spanning 513,120 square kilometres (198,120 sq mi), with a population of over 66 million people; Thailand is the world's 50th-largest country by land area and the 22nd-most-populous. The capital and largest city is Bangkok, a special administrative area. Thailand is bordered to the north by Myanmar and Laos, to the east by Laos and Cambodia, to the south by the Gulf of Thailand and Malaysia, and to the west by the Andaman Sea and the southern extremity of Myanmar. Its maritime boundaries include Vietnam in the Gulf of Thailand to the southeast, and Indonesia and India on the Andaman Sea to the southwest. Nominally, Thailand is a constitutional monarchy and parliamentary democracy; however, in recent history, its government has experienced multiple coups and periods of military dictatorships.

https://en.wikipedia.org/wiki/Thailand

How Many Sentences Are There?

Example of sentence tokenizer

1) Thailand, officially the Kingdom of Thailand and formerly known as Siam, is a country in Southeast Asia.

2) Located at the centre of the Indochinese Peninsula, it is composed of 76 provinces spanning 513,120 square kilometres (198,120 sq mi), with a population of over 66 million people; Thailand is the world's 50th-largest country by land area and the 22nd-most-populous.

3) The capital and largest city is Bangkok, a special administrative area.

...

Example of sentence tokenizer

```
import nltk

nltk.download('punkt')

string = "Thailand,  officially the Kingdom of Thailand and formerly known as Siam,  is a

sentences = nltk.sent_tokenize(string)

for sentence in sentences:
  print(sentence + "\n")
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
Thailand,  officially the Kingdom of Thailand and formerly known as Siam,  is a country in

Located at the centre of the Indochinese Peninsula, it is composed of 76 provinces spannin

The capital and largest city is Bangkok, a special administrative area.

Thailand is bordered to the north by Myanmar and Laos, to the east by Laos and Cambodia, t

Its maritime boundaries include Vietnam in the Gulf of Thailand to the southeast, and Indo

Nominally, Thailand is a constitutional monarchy and parliamentary democracy; however, in
```

How Many Words Are There?

Example of sentence tokenizer

Thailand, officially the Kingdom of Thailand and formerly known as Siam, is a country in Southeast Asia.

Example of words tokenizer

```
test_sentence = sentences[0]

words = nltk.word_tokenize(test_sentence)

for word in words:
  print(word + "\n")
```

```
Thailand

,

officially

the

Kingdom

of

Thailand

and

formerly

known

as

Siam

,
```

Example of sentence tokenizer

The capital and largest city is Bangkok, a special administrative area.

Similar Meaning

Text Lemmatization and Stemming

- Stemming

"A process to reduce into their root form by cutting off the end or the beginner of the word."

- The goal of the stemming is to by removing the prefix of suffix of the word.
- The result is not guarantee to match with the root form.
- Different algorithms produce a different result.

Text Lemmatization and Stemming

Words	Stem*
Studying	Stude
Studies	Studi
Studied	Studi

*This stemming process is based on Porter Stemming. Further information can be found at

https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Text Lemmatization and Stemming

```
test_sentence = sentences[0]

words = nltk.word_tokenize(test_sentence)

for word in words:
  print(word + "\n")
```

```
from nltk import PorterStemmer

stemmer = PorterStemmer()

print("Stem of studing : ",stemmer.stem("Studing"))
print("Stem of studies : ",stemmer.stem("Studies"))
print("Stem of studied : ",stemmer.stem("Studied"))

print("Stem of students : ",stemmer.stem("Students"))
```

```
Stem of studing :  stude
Stem of studies :  studi
Stem of studied :  studi
Stem of students :  student
```

Text Lemmatization and Stemming

- Lemmatization

"A process to reduce into their root form using morphological analysis and context analysis."

- The dictionary is required.
- The base form is referred as "lemma".

Text Lemmatization and Stemming

Words	Lemma
Studying	Study
Studies	Study
Studied	Study

*This lemmatizing process is based on WordNet.

Text Lemmatization and Stemming

```
from nltk.stem import WordNetLemmatizer
from nltk.corpus import wordnet
#nltk.download('wordnet')
lemmatizer = WordNetLemmatizer()
print("Lemmatized Word of studying :", lemmatizer.lemmatize("studying",wordnet.VERB))
print("Lemmatized Word of studyies :", lemmatizer.lemmatize("studies",wordnet.VERB))
print("Lemmatized Word of studyied :", lemmatizer.lemmatize("studied",wordnet.VERB))
print("Lemmatized Word of student :", lemmatizer.lemmatize("student",wordnet.NOUN))
print("Lemmatized Word of students :", lemmatizer.lemmatize("students",wordnet.NOUN))
```

```
Lemmatized Word of studying : study
Lemmatized Word of studyies : study
Lemmatized Word of studyied : study
Lemmatized Word of student : student
Lemmatized Word of students : student
```

Text Lemmatization and Stemming

```
**********************************************************************
*********
  Resource wordnet not found.
  Please use the NLTK Downloader to obtain the resource:

  >>> import nltk
  >>> nltk.download('wordnet')

  Attempted to load corpora/wordnet

  Searched in:
```

Word Representation

Bag-of-words

- Machine learning can not work with the string.
  - It works with numerical data.

- To apply the machine learning with the string, we need to convert the string into computable data.

- Bag-of-word is a representation of a string based on frequency of entities.

Bag-of-words

- Architecture

	Cat	Dog
John buys a cat.	1	
Jane gets a dog.		1
James has a dog.		1
Friends buy a cat and a dog.	1	1

Bag-of-words

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(vocabulary=["cat","dog"])
data_corpus = ["John buys a cat.",
"Jane gets a dog.",
"James has a dog.",
"Friends buy a cat and a dog."]
X = vectorizer.fit_transform(data_corpus)
print(X.toarray())
print(vectorizer.get_feature_names())
```

```
[[1 0]
 [0 1]
 [0 1]
 [1 1]]
['cat', 'dog']
```

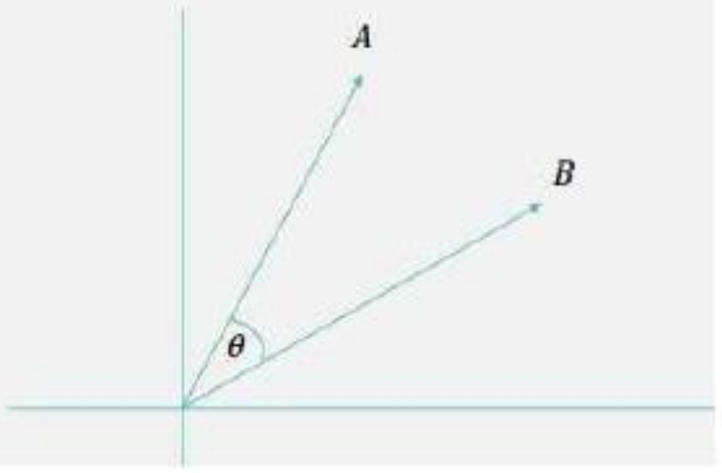
Cosine Similarity

- Cosine similarity is a measurement of similarity between 2 vectors.
- The angle denotes the similarity.
  - If two vectors are pointing to the similar direction, the angle between them is close.
  - If two vectors are pointing to the different direction, the angle between them is wide.

Cosine Similarity

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

- A and B denotes two vectors.



Cosine Similarity

id	Sentences
S00	John buys a cat.
S01	Jane gets a dog.
S02	James has a dog.
S03	Friends buy a cat and a dog.

Sentences Similarity

	S00	S01	S02	S03
S00		0	0	0.71
S01			1	0.71
S02				0.71
S03				

```
from sklearn.metrics.pairwise import cosine_similarity

vector_array = X.toarray()



for i in [0,1,2,3]:
  print(i)
  print(cosine_similarity(vector_array[i].reshape(1, -1),vector_array[0].reshape(1, -1)))
  print(cosine_similarity(vector_array[i].reshape(1, -1),vector_array[1].reshape(1, -1)))
  print(cosine_similarity(vector_array[i].reshape(1, -1),vector_array[2].reshape(1, -1)))
  print(cosine_similarity(vector_array[i].reshape(1, -1),vector_array[3].reshape(1, -1)))
```

```
0
[[1.]]
[[0.]]
[[0.]]
[[0.70710678]]
1
[[0.]]
[[1.]]
[[1.]]
[[0.70710678]]
2
[[0.]]
[[1.]]
[[1.]]
[[0.70710678]]
3
[[0.70710678]]
[[0.70710678]]
[[0.70710678]]
[[1.]]
```

Limitation of Bag-Of-Words

- Need training dataset
- Domain sensitivity
- Poor performance

Naïve Method

Technologies

Overall Process

1) Scrape pantips

```
import sys

sys.path.insert(0,'/usr/lib/chromium-browser/chromedriver')

from selenium import webdriver
import bs4 as BeautifulSoup
import bleach
from pythainlp import Tokenizer

chrome_options = webdriver.ChromeOptions()
chrome_options.add_argument('--headless')
chrome_options.add_argument('--no-sandbox')
chrome_options.add_argument('--disable-dev-shm-usage')
driver = webdriver.Chrome('chromedriver',chrome_options=chrome_options)
driver.get('https://pantip.com/topic/40268797')

html = driver.page_source
soup = BeautifulSoup.BeautifulSoup(html,'lxml')
driver.close()

mydivs = soup.find_all("div", class_="display-post-story")
```

2) Extract comments

```
comment = []
for div in mydivs:
    text = bleach.clean(div.text.strip(), tags=[], attributes={}, styles=[], strip=True)
    clean_text=' '.join(text.split())
    if len(clean_text)>0 :
        comment.append(clean_text)
        #print(clean_text+"\n\n")
```

3) Analyze comments

```
pos = ["หล่อ","ชอบ","ดี","ทีม","สเปก","ประสบความสำเร็จ","เชียร์","ท็อป"]
neg = ["แย่","ห่วย","ไม่เอา","ไม่ค่อย"]

mydict = pos+neg

tokenizer = Tokenizer(custom_dict=mydict, engine='newmm',keep_whitespace=False)



for com in comment:
  pos_count = 0
  neg_count = 0
  print("\n\n"+com)

  text = tokenizer.word_tokenize(com)

  for word in text :
    if word in pos :
      pos_count = pos_count + 1
    if word in neg :
      neg_count = neg_count + 1

  if pos_count > neg_count:
    print("positive")
  if pos_count < neg_count:
    print("negative")
  if pos_count == neg_count:
    print("neutral")
```

3.7 บทที่ : 7
Supply Chain Management

Data Science Machine Learning for Business

กรณีศึกษาการประยุกต์ใช้วิทยาการข้อมูล กรณีศึกษา Supply Chain Management

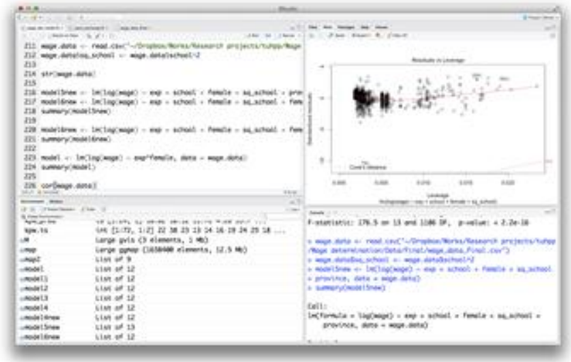
ดร.ไพรัช พิบูลย์รุ่งโรจน์
คณะเศรษฐศาสตร์ มหาวิทยาลัยเชียงใหม่

www.scerc.net | me@pairach.com | www.pairach.com

ลักสูตร Develop Data Science Machine Learning for Business ทักษะการสร้าง Data Science Machine Learning เพื่อหาองค์ความรู้ใหม่และวิเคราะห์ จากข้อมูลมหาศาล เพื่อสร้างมูลค่าต่อธุรกิจหรือองค์กร
โครงการพัฒนาความสามารถทางเทคโนโลยีของบุคลากรภาคอุตสาหกรรม (Brain Power Skill Up) ภายใต้โครงการสร้างกำลังคนและทักษะแห่งอนาคตในภูมิภาคเพื่อตอบโจทย์การพัฒนานวัตกรรมของประเทศประจำปีงบประมาณ พ.ศ.2563

Today Topics

1. Supply Chain Analytics (SCA)
2. How to use R!
3. Basic SCA
4. Case Study



What is Supply Chain Analytics ?

Supply Chain Analytics – A Buzz Words

- DPB: Data Science, Predictive Analytics and Big data

JOURNAL OF BUSINESS LOGISTICS
STRATEGIC SUPPLY CHAIN RESEARCH
Journal of Business Logistics, 2013, 34(2): 77–84

Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management

Matthew A. Waller[1] and Stanley E. Fawcett[2]

[1]*University of Arkansas*
[2]*Weber State University*

Skills needed by SCM Analytics

Discipline	SCM data scientist skill set: More important	SCM data scientist skill set: Less important
Statistics	Broad *awareness* of many different methods of estimation and sampling	Derivations of methods and proofs of maximum likelihood estimation
Forecasting	Understanding *application* of qualitative and quantitative methods of forecasting	Understanding of underlying stochastic processes
Optimization	Numerical methods of optimization	Finding global optimal solutions
Discrete event simulation	Quick design and implementation of discrete event simulation models	Queuing theory
Applied probability	Using probability theory with actual data to estimate the expected value of random variables of interest	The theory of stochastic processes
Analytical mathematical modeling	Using numerical methods to estimate functions relating independent variables to dependent variables	Proving theorems
Finance	Capital budgeting	Efficient market theory
Economics	Determining opportunity cost	Macroeconomic theory
Marketing	Marketing science	Semiotics
Accounting	Managerial accounting	Debits and credits journal entries

Source: Waller and Fawcett (2013), JBL

Research in SC Analytics

Comparative discipline	Dimension of interest	Predictive analytics research (examples)	
		Relevant	Less relevant
Statistics	Quantitative	Integrating quantitative and qualitative analysis	Improving Lagrange Multiplier tests for autocorrelation
Forecasting	Predicting the future	Using forecasting techniques for evaluating what would have happened under different circumstances	Deriving generalized estimators of seasonal factors
Optimization	Minimization and maximization	Assessment of the quality of the optimal solution and the ability to implement it versus near optimal solutions	Use of polyhedral functions in linear programming
Discrete event simulation	Quantitative analysis of a system in a stochastic setting	Discrete event simulation in a business process reengineering setting	Random number generation for discrete event simulation
Applied probability	Description of stochastic variables, expected values, and uncertainty	Applied probability along with application anchoring and framing affects from psychology	Asymptotic properties of Gaussian processes
Data mining	Search for patterns and relationships between a large number of variables with lots of data	Data mining preceded by logical and theoretical descriptions of possible relationships and patterns	Gibbs posterior for variable selection in data mining
Analytical mathematical modeling	Precise analysis using artificial and unrealistic assumptions for theorems and proofs	Methods of quickly and inexpensively modeling approximate relationships between variables while still using deductive mathematical methods	Proving inventory theorems that assume known, continuous demand with perfect information

Source: Waller and Fawcett (2013), JBL

The 3 V: Causes of Big Data

Type of data	Volume	Velocity	Variety
Sales	More detail around the sale, including price, quantity, items sold, time of day, date, and customer data	From monthly and weekly to daily and hourly	Direct sales, sales of distributors, Internet sales, international sales, and competitor sales
Consumer	More detail regarding decision and purchasing behavior, including items browsed and bought, frequency, dollar value, and timing	From click through to card usage	Face profiling data for shopper identification and emotion detection; eye-tracking data; customer sentiment about products purchased based on "Likes," "Tweets," and product reviews
Inventory	Perpetual inventory at more locations, at a more disaggregate level (e.g., style/color/size)	From monthly updates to hourly updates	Inventory in warehouses, stores, Internet stores, and a wide variety of vendors online
Location and time	Sensor data to detect location in store, including misplaced inventory, in distribution center (picking, racks, staging, etc.), in transportation unit	Frequent updates for new location and movement	Not only where it is, but what is close to it, who moved it, its path to get there, and its predicted path forward; location positions that are time stamped from mobile devices

Source: Waller and Fawcett (2013), JBL

Potential Applications of SC Analytics

User	Forecasting	Inventory management	Transportation management	Human resources
Carrier	Time of delivery, factoring in weather, driver characteristics, time of day and date	Real time capacity availability	Optimal routing, taking into account weather, traffic congestion, and driver characteristics	Reduction in driver turnover, driver assignment, using sentiment data analysis
Manufacturer	Early response to extremely negative or positive customer sentiment	Reduction in shrink, efficient consumer response, quick response, and vendor managed inventory	Improved notification of delivery time, and availability; surveillance data for improved yard management	More effective monitoring of productivity; medical sensors for safety of labor in factories
Retailer	Customer sentiment data and use of mobile devices in stores	Improvement in perpetual inventory system accuracy	Linking local traffic congestion and weather to store traffic	Reduction in labor due to reduction in misplaced inventory

Source: Waller and Fawcett (2013), JBL

Potential Applications of SC Analytics

Type of data	Inventory management	Transportation management	Customer and supplier relationship management
Sales	How can sales data be used with detailed customer data to improve inventory management either in terms of forecasting or treating some inventory as "committed" based on specific shoppers requirements?	How can more current sales data be used to re-direct shipments in transit? How can sales data, integrated with detailed customer data, be used for more efficient and effective merge-in-transit operations?	How can more granular sales " data from the wide variety of sources that exist be used to improve visibility on the one hand and trust on the other, between trading partners?
Consumer	How can face profiling data for shopper identification, emotion detection, and eye-tracking data be used to determine which items to carry and stock at particular shelf locations?	How can delivery preferences captured in online purchases be used to manage transportation mode and carrier selection decisions?	How can customer sentiment about products purchased based on "Likes," "Tweets," and product reviews be used to collaborate on forecasts?
Location and time	How can sensor data used to detect location in store, be used to improve inventory management, including departmental merchandising decisions?	How can sensor data in the distribution center be used to anticipate transportation requirements?	How can location and time-stamp data of shoppers be used for collaborative assortment and merchandising decisions?

SCERC Supply Chain Economics Research Centre

Source: Waller and Fawcett (2013), JBL

Debates in Supply Chain Analytics

Source: Deloitte (2010)

Debates in Supply Chain Analytics

Deloitte.

Deloitte Debates

Supply Chain Analytics: How Hard Should You Squeeze?

SCERC Supply Chain Economics Research Centre
Faculty of Economics | Chiang Mai University
Delivering world-class quality research and analytics through state-of-the-art techniques.

Source: Deloitte (2010)

Debates in Supply Chain Analytics (1)

	Point	Counterpoint
It's a gold mine. *Investing in advanced analytics will save money and solve problems you didn't know you had.*	You have plenty of data, but you need a different way of looking at it – one that can turn historical data captured in transactional systems into predictive insights.	Our company doesn't have the capabilities to do these kinds of analyses – and the resources aren't available to build them. This is nice to have, but not necessary.
	Supply chain cost is often the largest component of my cost structure and ultimately determines profits. It's the biggest opportunity to extract value.	It's also one of the most transparent inputs to my balance sheet and P&L. If I were overpaying, I'd know it already.
	The supply chain merits the same analysis as other parts of your business. You manage pricing to the fourth decimal point, but you're winging it on this?	Supply chain categories are not as sensitive or finely tuned as pricing. Why should we apply the same approach across the entire enterprise?

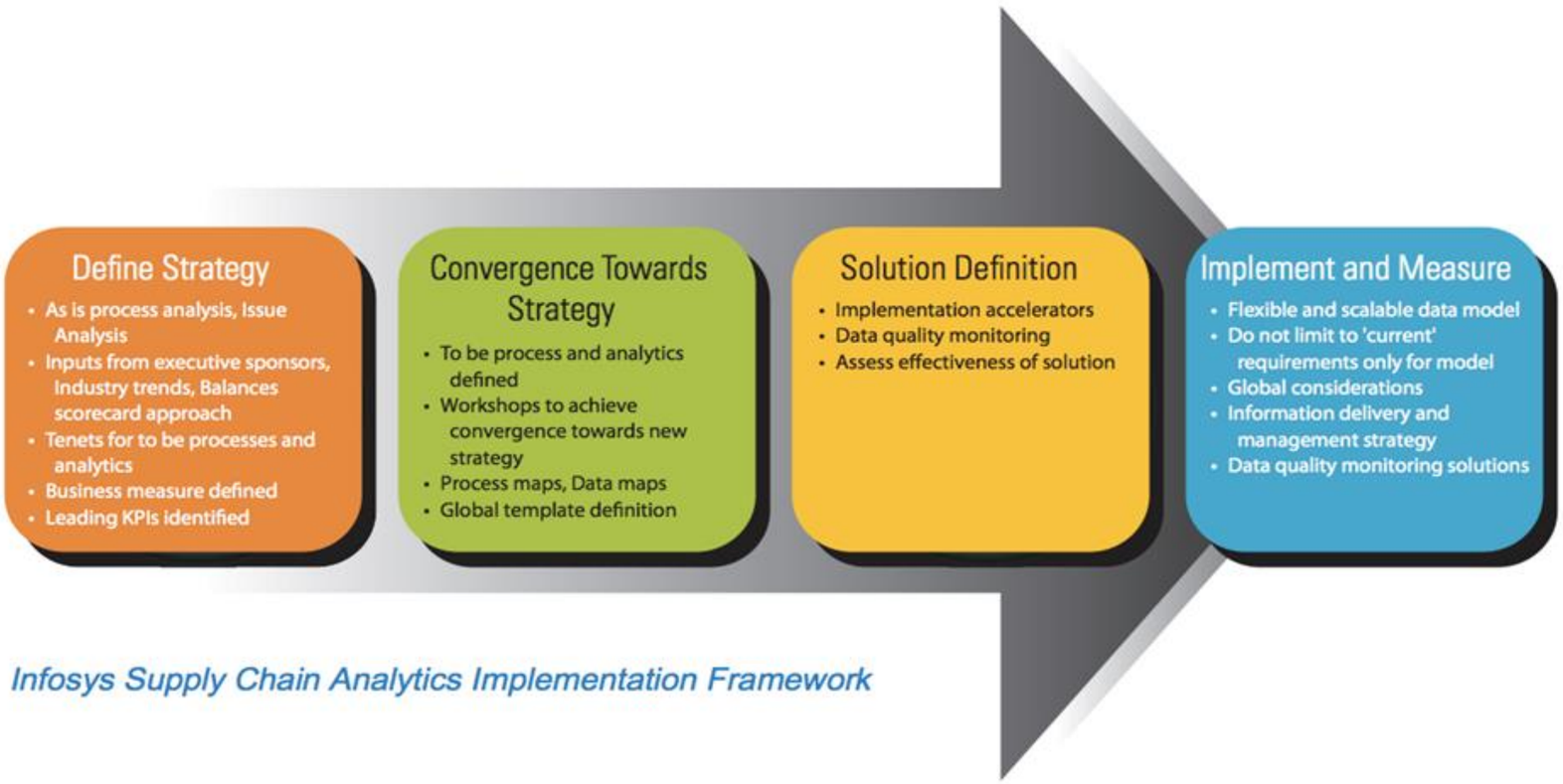
Source: Deloitte (2010)

Debates in Supply Chain Analytics (2)

	Point	Counterpoint
It's a rabbit hole.	Analytics may generate savings, but the cost of getting there may be too high. The low-hanging fruit has a better ROI.	Going the extra mile differentiates winners from losers. That's what competitive advantage is all about.
Beware of any plan that has you spending money just to save other money.	I'm in the _____ business, not the analytics business. Every time I stray from my core competency, it ends up costing me.	You're not in the floor sweeping business either, but you make sure that it's taken care of. If something outside your bull's-eye is a potential problem – or a potential source of profit – all the more reason to take action.
	Benchmarking to peers is enough to keep from losing competitive advantage.	Benchmarking is a valuable starting point, but in many cases it only identifies the gaps. Analysis helps you close them.
	Fads come and go, but operations management is operations management. Every element of supply chain management uses disciplines honed over many years.	New volumes of data and new abilities to mine value from it are an advancement, not a fad.

Source: Deloitte (2010)

Implementation of SCA

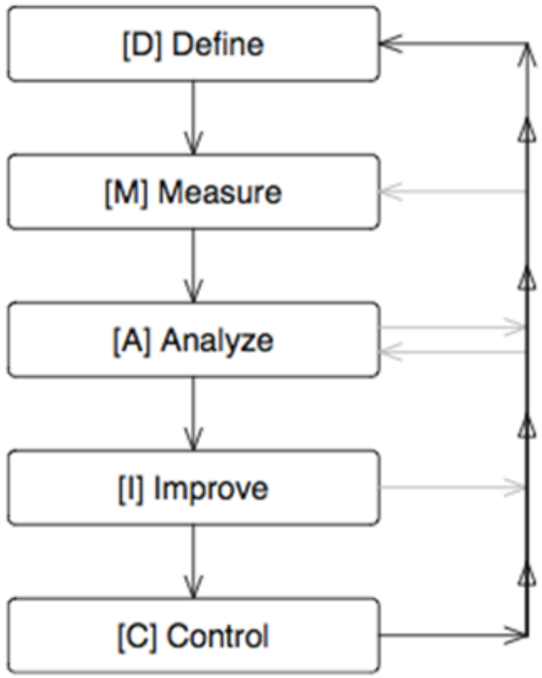


Infosys Supply Chain Analytics Implementation Framework

Source: Infosys (2011)

http://www.infosys.com/SAP/success-stories/Documents/supply-chain-analytics-implementation.pdf

DMAIC Steps in SCM analysis

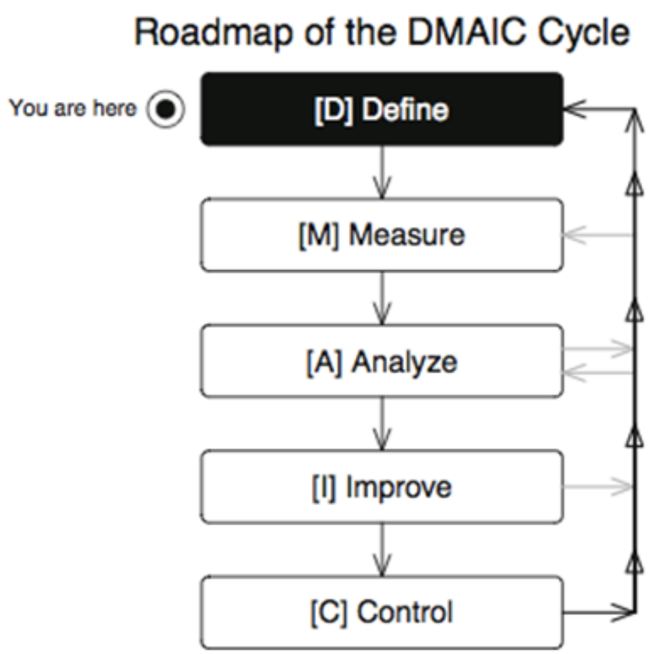


DMAIC Cycle	Scientific Method
Define	Ask a Question
Measure	Do some background research
Analyze	Construct a hypothesis
Improve	Test the hypothesis with an experiment
Control	Analyze the data and draw conclusions
	Communicate results

Process Mapping
with DMAIC Framework

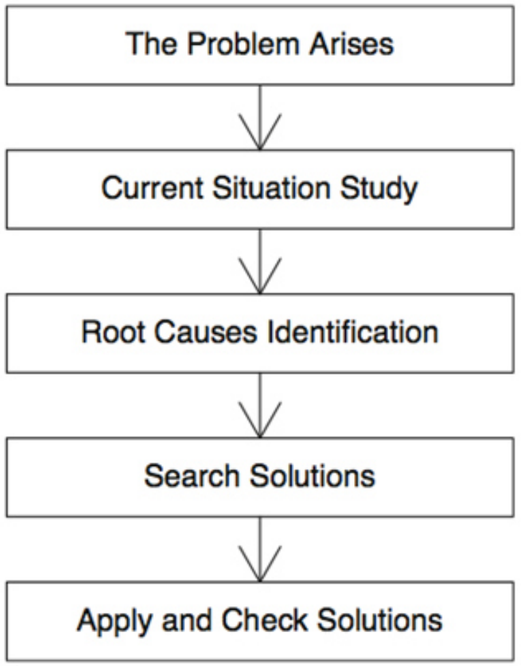
"A problem well stated is a problem half solved."
Charles Franklin Kettering

D for Define [DMAIC]



DMAIC Cycle	Scientific Method
Define	Ask a Question
Measure	Do some background research
Analyze	Construct a hypothesis
Improve	Test the hypothesis with an experiment
Control	Analyze the data and draw conclusions
	Communicate results

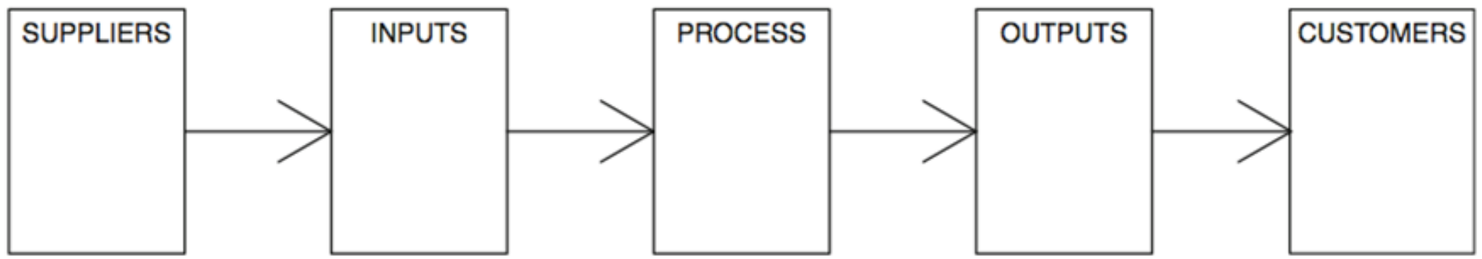
Process Mapping



- A problem-solving flow.
- A systematic method helps to get rid of the problems' causes.
- The process mapping includes mainly the first three stages and,
- in some cases, part of the fourth stage

Define Phrase

SIPOC Flow



SIPOC flow. The SIPOC flow chart represents the natural flow of a process or service from supplier to customer

Classifying the parameter

While detecting the parameters in a step of the process, we must assess their influence in the features and how this influence is produced. Thus, the parameters must be classified in one of the following groups:

1. N C P Cr
2. Noise: Noncontrollable factors
3. Controllable factors: May be varied during the process
4. Procedure: Controllable factors through a standard procedure
5. Critical: Those with more inflence on the process

Example: Pizza Process

The process of making and serving a pizza can be broken down into the following steps:

1. Prepare the dough.
2. Spread the toppings.
3. Bake the pizza.
4. Deliver the pizza to the customer.

Example: Pizza Process

We assume that the inputs of each step are the output of the previous step. The inputs for the first stage are the x_s defined previously (ingredients, cook, oven, and plates). Next, we describe in detail the parameters and outputs corresponding to each step (with the classification of the parameters in brackets).

The process of making and serving a pizza can be broken down into the following steps:

1. Prepare the dough.
2. Spread the toppings.
3. Bake the pizza.
4. Deliver the pizza to the customer.

Practices

- Draw the main process of your supply chain e.g., production or procurement
- You need inputs and outputs
- Make the process map of your own is the key

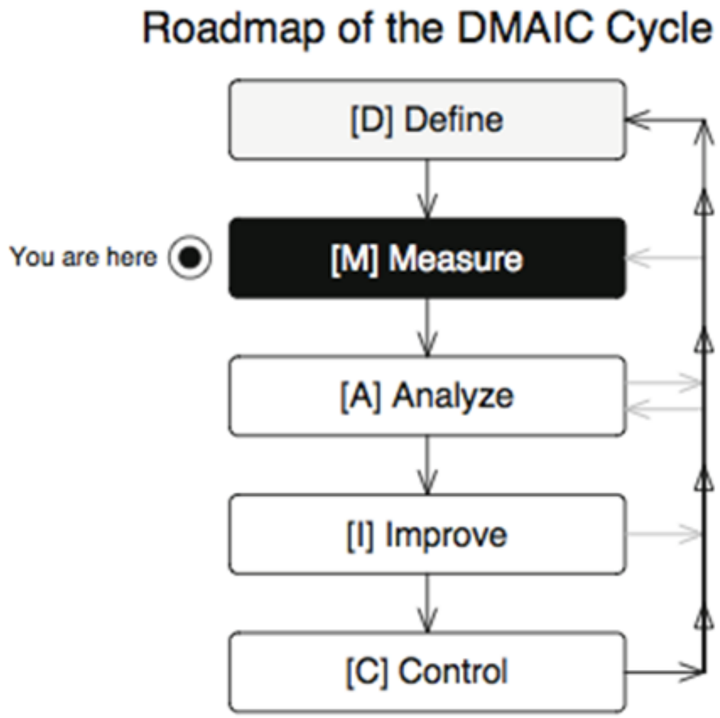
Pareto Analysis with R

Causa latet: vis est notissima.

[The cause is hidden, but the result is known.]

Ovid

M for Measure [DMAIC]



DMAIC Cycle	Scientific Method
Define	Ask a Question
Measure	Do some background research
Analyze	Construct a hypothesis
Improve	Test the hypothesis with an experiment
Control	Analyze the data and draw conclusions
	Communicate results

Pareto Principle

- Vilfredo Pareto (1848–1923) was an Italian economist whose most famous contribution was the principle known by his name.
- He was also a philosopher, engineer, sociologist, and political scientist.
- The Pareto principle was a result of Pareto's observations about the distribution of wealth in the 19th century.

Pareto Principle

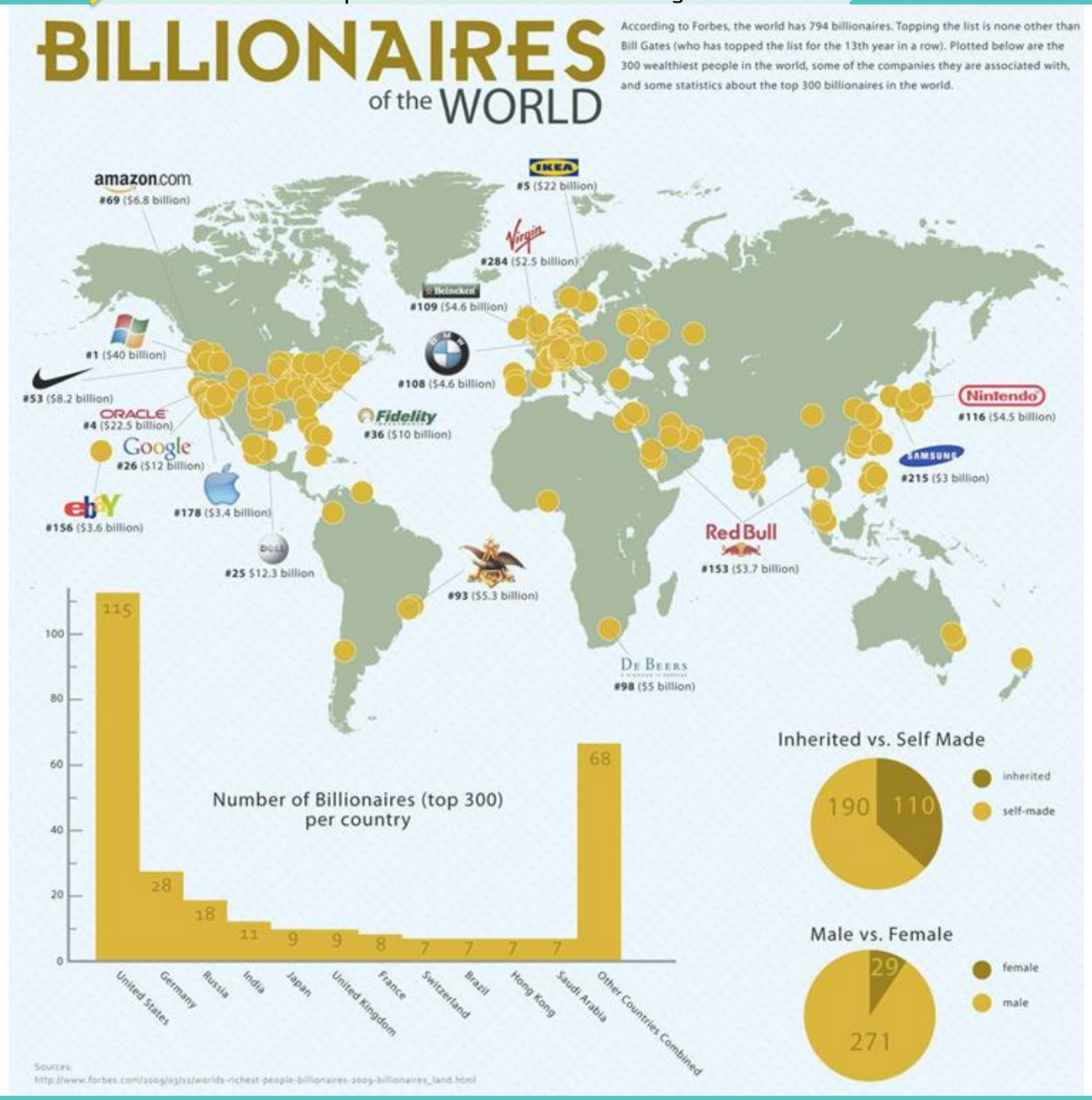
Quintile of population	Income
Richest 20%	82.70%
Second 20%	11.75%
Third 20%	2.30%
Forth 20%	1.85%
Poorest 20%	1.40%

Distribution of world GDP, 1989

1. He observed that 80% of the wealth was owned by 20% of the population.
2. Hence, the Pareto principle is also known as the 80/20 rule.
3. It appears in many real-life situations, and therefore it is sometimes considered a natural principle.

Koch (2001) The 80/20 Principle: The Secret with Achieving More with Less

Pareto Principle in Actions

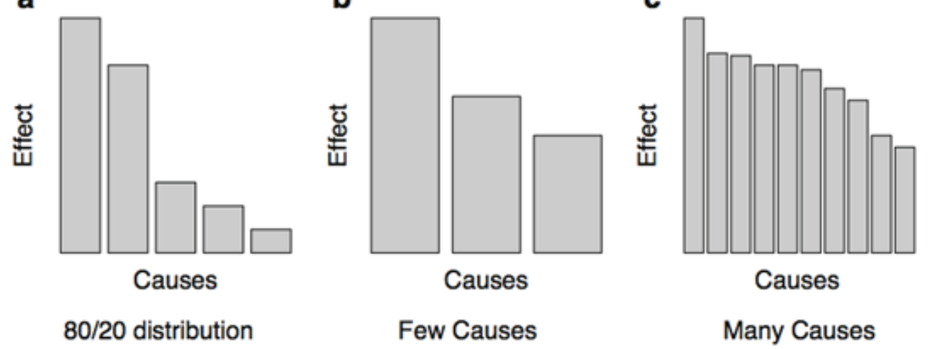


SCERC Supply Chain Economics Research Centre
Faculty of Economics | Chiang Mai University
delivering world-class quality research and analytics through state-of-the-art techniques.

Pareto Analysis

- The Pareto chart is considered one of the seven basic tools for quality control in the supply chain:

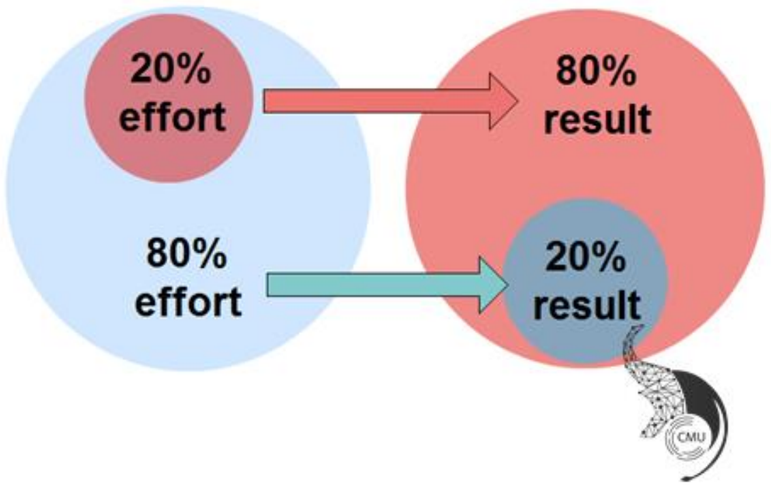
1. Histogram
2. Check sheet
3. Pareto chart
4. Cause-and-effect diagram
5. Defect concentration diagram
6. Scatter diagram
7. Control chart



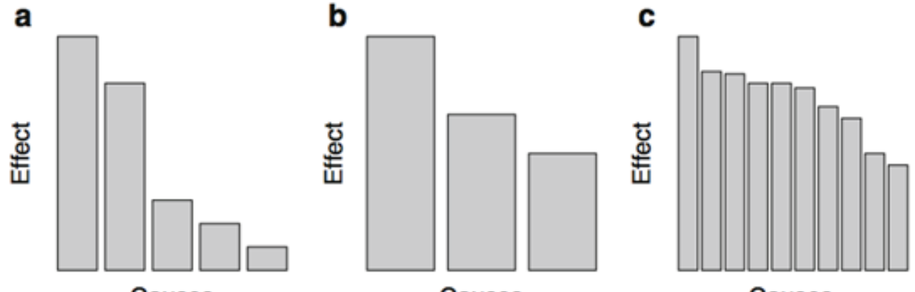
Pareto: 20% -> 80%

- Only 20% that affect 80%
- Effort -> Results or Cause -> Effects

- In a supply chain, there are many items/causes.
- Only 20% that's really matter



Case Study



For example:

- 80% of the profits comes from 20% of the customers.
- 20% of the employees do 80% of the work.
- 20% of patients use 80% of care resources.
- 80% of cost of quality is produced by 20% of the sources of error.
- The last example corresponds to the main application of the Pareto principle in Six Sigma. We will explain how to use it in the following sections.

Example – Pareto - Construction project

- The Black Belt in the construction company has investigated why a sampling of deadlines on projects developed in the last 2 years went unfulfilled.
- He has also estimated the cost of these delays for the company (larger labor force, extra payments, etc.).
- We will save the data in a data frame with a factor whose levels are the possible causes (we use the b.causes list created previously to draw the cause-and-effect diagram), and with two variables, (namely: number of unfulfilled deadlines and estimated cost).

Building a Pareto Chart – a step-by-step approach

1. Identify the causes.
2. Choose the appropriate measurement units.
3. Obtain the data.
4. Sort by importance.
5. Figure the cumulative percentage.
6. Plot a bar chart for the measurements.
7. Plot a line chart for the cumulative percentages.
8. Find the causes responsible for 80% of the effect.

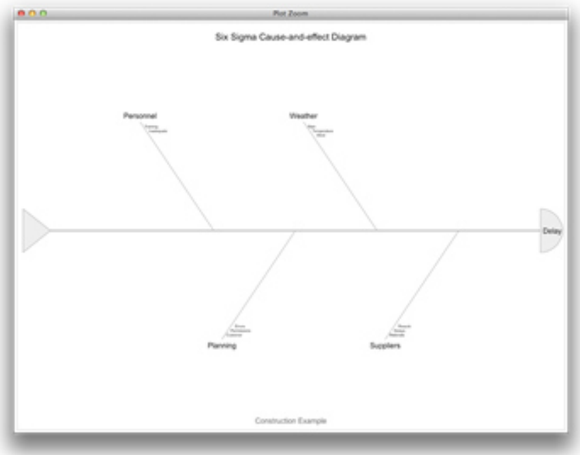
Measure Phrase

Example – Pareto - Construction project

Visualise cause-and-effects

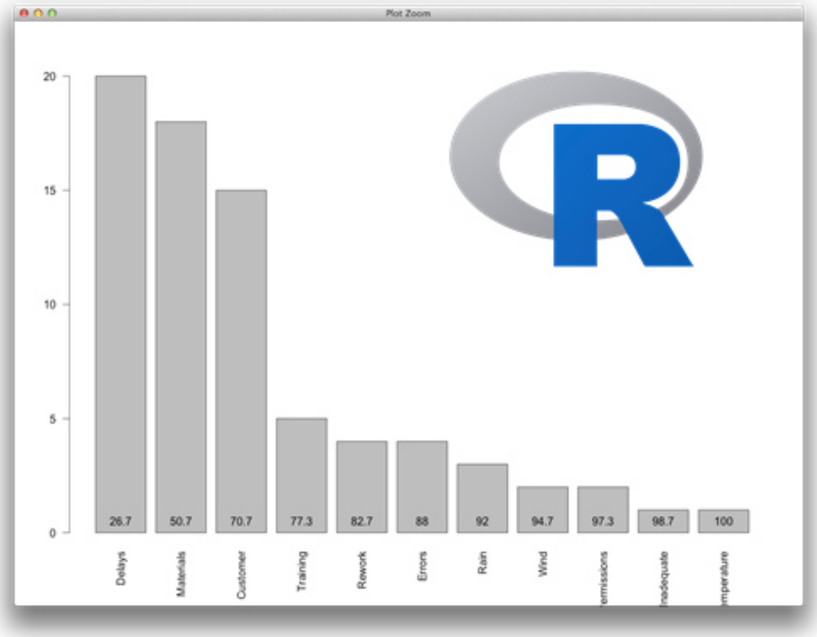
using fish bone-diagram

```
# Cause and Effect Diagram
b.effect <- "Delay"
b.groups <- c("Personnel", "Weather", "Suppliers", "Planning")
b.causes <- vector(mode = "list", length = length(b.groups))
b.causes[1] <- list(c("Training", "Inadequate"))
b.causes[2] <- list(c("Rain", "Temperature", "Wind"))
b.causes[3] <- list(c("Materials", "Delays", "Rework"))
b.causes[4] <- list(c("Customer", "Permissions", "Errors"))
ss.ceDiag(b.effect, b.groups, b.causes, sub = "Construction Example")
```



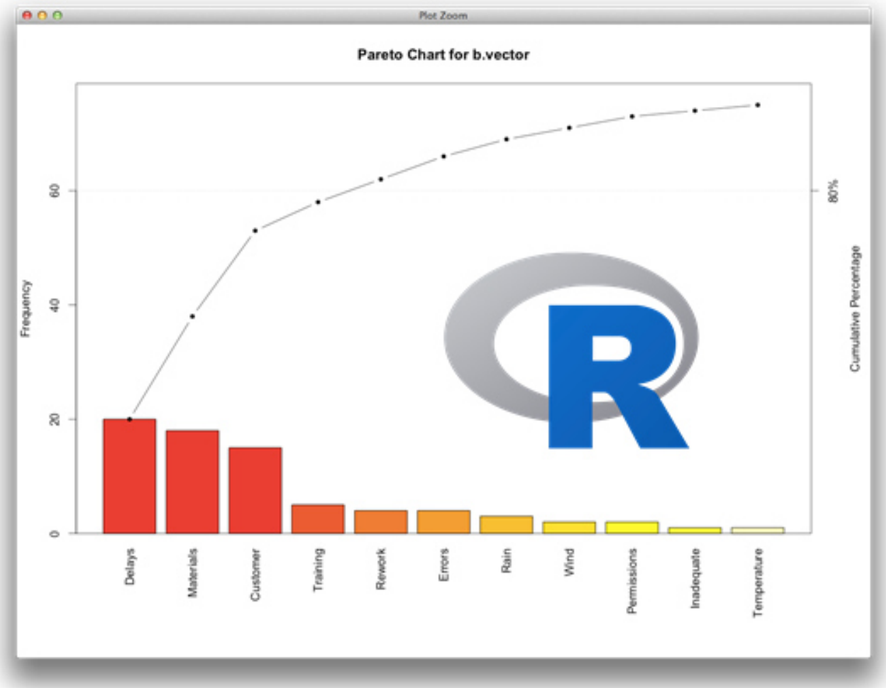
Measure Phrase

Example – Pareto - Construction project



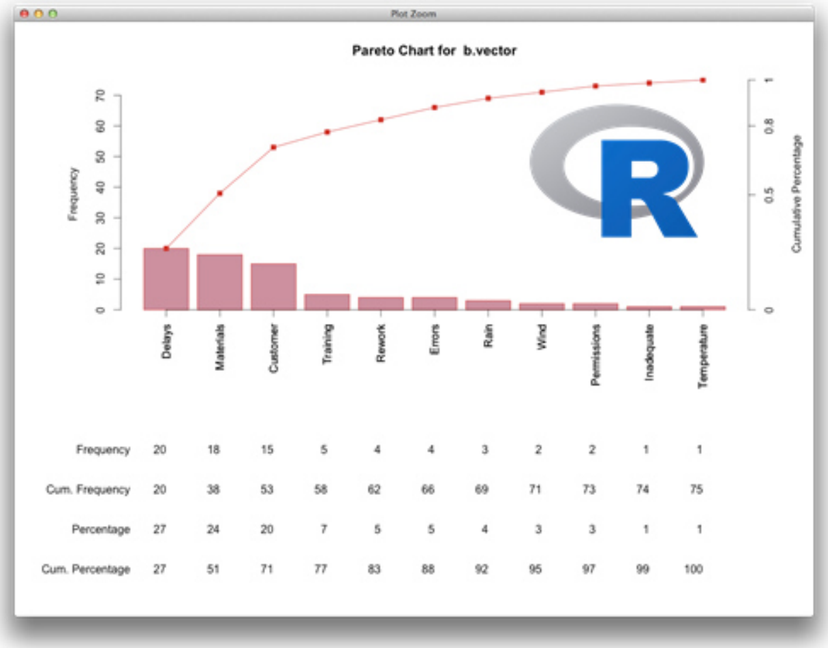
- Pareto chart with base graphics.
- Ordering the values in the data frame makes it easy to plot a Pareto chart with the barplot function.
- We can annotate the plot with the cumulative percentages using the text function

Example – Pareto - Construction project



- A secondary y-axis on the left from 0 to 100, representing the cumulative percentage of the effect measurement.
- A line chart linking the cumulative percentages of each cause.
- Auxiliary lines from the axis identifying the 80/20 rule, that is, which causes are responsible for 80% of the effect.

Example – Pareto - Construction project



- Pareto chart in qualityTools package.
- You can choose the ticks for the right axis (cumulative percentages).
- It plots a table below the chart with the values and the percentages

Your Turn

[15 minute break]

Your Turn - Practices – Pareto

- Think about your FOCAL SUPPLY CHAIN, and try to list the factors that can cause defects in the products, leading to worse outcomes
- Group the causes and create a cause-and-effect diagram. You can use just paper and pencil, but once you have finished with it, try to plot it with R.
- Then measure the errors making lots of THE PRODUCT with different conditions (economic recession, natural disaster, etc.). You can also invent or simulate the data.
- Save the data into an R data frame and plot a Pareto chart. Analyze the results, and figure out where you will have to focus your attention to produce optimal improvements.

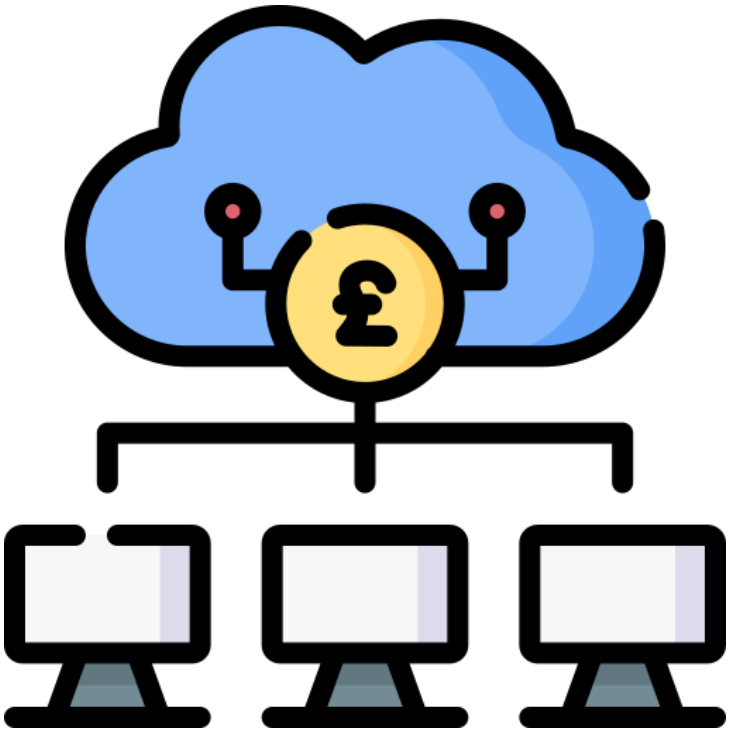
Q & A

me@pairach.com

Supply Chain Economics Research Centre
Faculty of Economics | Chiang Mai University
Delivering world-class quality research and analytics through state-of-the-art techniques.

บทที่ 4 : แบบทดสอบและประเมินผลหลังเรียน

4.1 แบบทดสอบหลังพัฒนาทักษะ (Post-Test)



ส่วนที่ 1 ลงทะเบียน

1. กรุณากรอกชื่อ-นามสกุล.....

2. สถานประกอบการ.....

3. Email.....

4. เบอร์โทร.....

ส่วนที่ 2 แบบทดสอบหลังพัฒนาทักษะ (Post-Test)

คำชี้แจง 1. แบบทดสอบฉบับนี้เป็นแบบอัตนัย จำนวน 4 ข้อ 5 คะแนน

2. จงเลือกคำตอบที่ถูกต้องที่สุดเพียงข้อเดียว

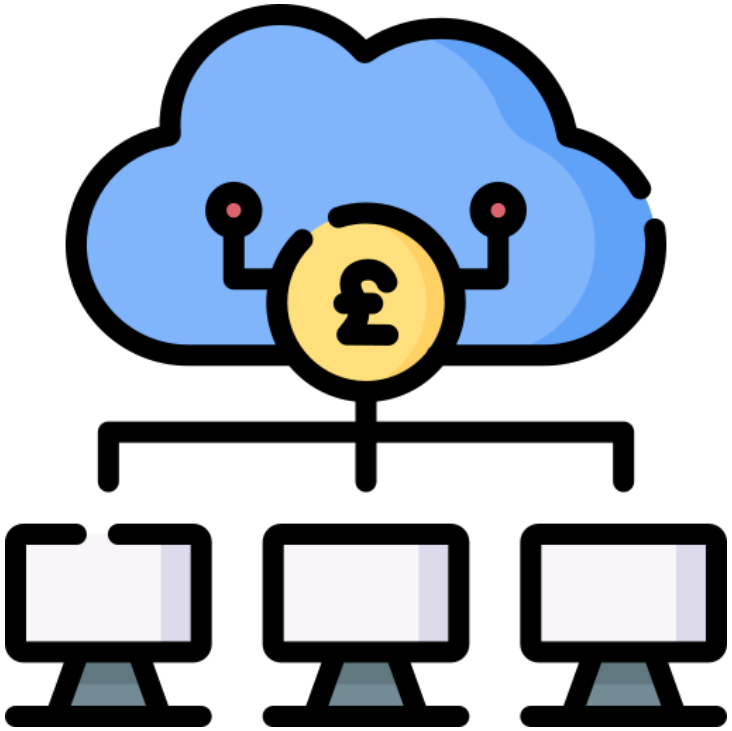
1. อะไรคือ CRISP-DM และทำไมต้องใช้ CRISP-DM จงอธิบายพร้อมยกตัวอย่างตามที่ท่านเข้าใจ

2. จงอธิบายความต่างระหว่างการเข้าใจบริบททางธุรกิจ (Business Understanding) และการเข้าใจข้อมูล (Data Understanding) จงอธิบายพร้อมยกตัวอย่างตามที่ท่านเข้าใจ

3. จงอธิบายปัญหาที่จะเกิดกับข้อมูล 1 ปัญหา พร้อม วิธีแก้ไข ตามที่ท่านเข้าใจ

4. จงอธิบายแนวทางการประยุกต์ใช้ วิทยาการข้อมูล กับ องค์กรของท่านตามแนวทาง CRISP-DM

4.2 แบบประเมินทักษะหลังการพัฒนาทักษะ (Post-Embedded Skill)



ส่วนที่ 1 สำหรับ ผู้เรียน

1.1 ข้อมูลทั่วไป

ชื่อ-นามสกุล.....

ชื่อสถานประกอบการ .....

1.2 เปรียบเทียบความรู้และทักษะที่ได้รับหลังเข้าร่วมพัฒนาทักษะ กับ พื้นฐานความรู้เดิม

( ) ได้พัฒนาทักษะใหม่ที่เพิ่มเติมและเป็นประโยชน์ มากกว่าความรู้เดิม ( ) ไม่ได้รับการพัฒนาทักษะ

1.3 ความพึงพอใจต่อหลักสูตรพัฒนาทักษะ

( ) พึงพอใจมากที่สุด ( ) พึงพอใจมาก ( ) พึงพอใจปานกลาง ( ) พึงพอใจน้อย

ส่วนที่ 2 สำหรับ เจ้าของกิจการ หรือ หัวหน้างาน

2.1 การประเมินผู้เรียน

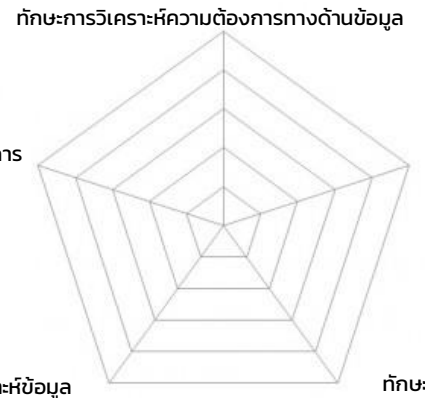
ความหมายระดับคะแนน

0 = Beginner	ไม่มีความรู้ ไม่มีทักษะ
1 = Learner	มีความเข้าใจในทฤษฎีเบื้องต้น
2 = Practitioner	มีความเข้าใจในทฤษฎีอย่างเต็มที่ มีความรู้ด้านปฏิบัติเล็กน้อย สามารถตอบคำถามหรือแก้ไขปัญหาที่ไม่ซับซ้อนได้
3 = Experienced	มีความเข้าใจในทฤษฎีและปฏิบัติอย่างเต็มที่ สามารถประยุกต์ใช้ความรู้เพื่อแก้ไขปัญหาซับซ้อนปานกลางได้
4 = Embedded	เกิดทักษะติดตัว สามารถเชื่อมโยงความรู้ในการแก้ไขปัญหาที่ซับซ้อนมากได้ และสามารถกำหนดแผนเพื่อปรับปรุงและพัฒนาประสิทธิภาพการทำงานในองค์กรได้และนำไปสู่การต่อยอดเพื่อลงมือทำจริง
5 = Broaden	เกิดทักษะอย่างท่องแท้ในระดับผู้เชี่ยวชาญ และสามารถถ่ายทอดทักษะให้แก่ผู้อื่นได้

กรุณา (√) ในช่องระดับคะแนน

ผลลัพธ์ทักษะ	ระดับคะแนน					
	0	1	2	3	4	5
1. ทักษะการวิเคราะห์ความต้องการทางด้านข้อมูล						
2. ทักษะด้านการออกแบบกระบวนการการวิเคราะห์ข้อมูล						
3. ทักษะในด้านการวิเคราะห์ข้อมูลเชิงสำรวจ (Exploratory Data Analysis)						
4. ทักษะพื้นฐานการออกแบบการวิเคราะห์ข้อมูลด้วยเทคนิคภาษาธรรมชาติ						
5. ทักษะพื้นฐานการเขียนโปรแกรมในด้านวิทยาการข้อมูล						

การวิเคราะห์ผลการพัฒนาทักษะด้วยกราฟเรดาร์ (Radar Chart)

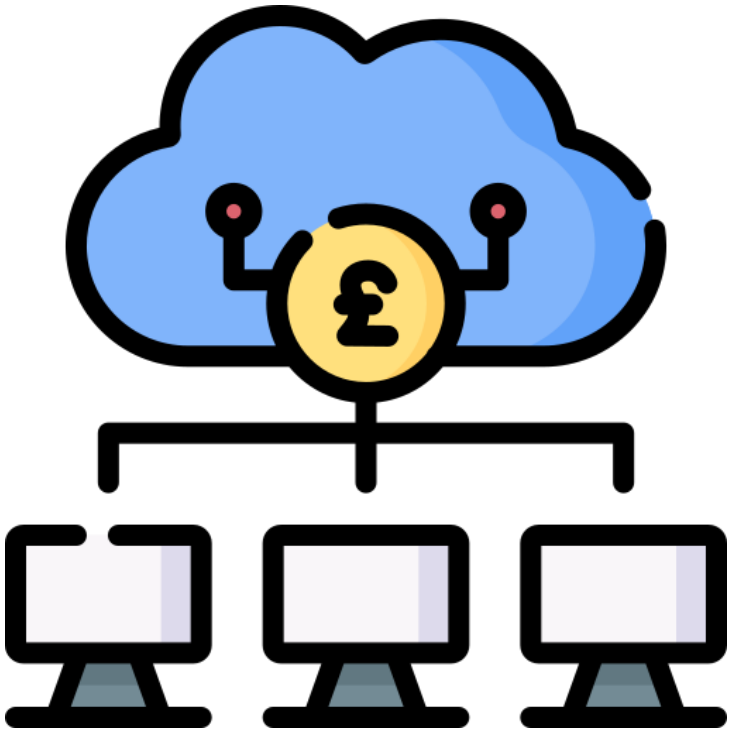


บทที่ 5 : แผนงาน

(Action Plan)

5.1 แบบฟอร์มแผนงาน

(Action Plan)



ส่วนที่ 1 **สำหรับผู้เข้าร่วมพัฒนาทักษะ**

ชื่อ-นามสกุล.....**ชื่อสถานประกอบการ**.....

ชื่อแผนงาน / ความต้องการ.....

วัตถุประสงค์.....

ที่	เป้าหมาย/ความต้องการ/ปัญหา	กลยุทธ์/แนวทางการแก้ไข	วิธีการดำเนินงาน (ระบุอย่างละเอียด)	ตัวชี้วัด	ระยะเวลา	ทรัพยากรที่มี
						งบประมาณ

ส่วนที่ 2 **สำหรับหัวหน้างาน หรือ เจ้าของกิจการ**

พิจารณาแผนที่ผู้เรียนนำเสนอ

.....

.....

แผนการต่อยอดหรือลงทุนจากทักษะที่ได้รับ

ที่	รายการ	พร้อมดำเนินงานทันที	มีแผนการดำเนินงานในอนาคต	โปรดอธิบายเพิ่มเติม	หมายเหตุ
1	ท่านสนใจลงทุนใน เครื่องจักร	☐	☐		
2	ท่านสนใจลงทุนใน กำลังคน เช่น มีการจ้างงานเพิ่มขึ้นเพื่อควบคุมเครื่องจักรที่ได้ลงทุนเพิ่ม	☐	☐		
3	ท่านสนใจต่อยอดและลงทุน ในด้านอื่น ๆ โปรดระบุ.....	☐	☐		

ความพึงพอใจต่อหลักสูตรพัฒนาทักษะ

( ) พึงพอใจมากที่สุด **( )** พึงพอใจมาก **( )** พึงพอใจปานกลาง **( )** พึงพอใจน้อย

KNOWLEDGE MANAGEMENT

หลักสูตรทักษะการสร้าง Data Science Machine Learning เพื่อหาองค์ความรู้ใหม่และวิเคราะห์จากข้อมูลมหาศาล เพื่อสร้างมูลค่าต่อธุรกิจหรือองค์กร

Developed Data Science Machine Learning for Business

ภายใต้แผนงานพัฒนาความสามารถทางเทคโนโลยีของบุคลากรภาคอุตสาหกรรม
โครงการสร้างกำลังคนและทักษะแห่งอนาคตในภูมิภาคเพื่อตอบโจทย์การพัฒนานวัตกรรมของประเทศ
ประจำปีงบประมาณ 2563

จัดทำโดย
อุทยานวิทยาศาสตร์ภาคเหนือ