

การเข้าสู่ยุคของวิทยาศาสตร์ข้อมูล Entering into the Era of Data Science

ประภาพร กุลลิ้มรัตน์ชัย

Prapaporn Kullimratchai

คณะวิศวกรรมศาสตร์ มหาวิทยาลัยอีสเทิร์นเอเซีย

Faculty of Engineering, Eastern Asia University

บทคัดย่อ

Data Science หรือวิทยาศาสตร์ข้อมูล คือสาขาสหวิทยาการใหม่ ที่สังเคราะห์และสร้างขึ้นบนพื้นฐานของ สถิติ สารสนเทศ คอมพิวเตอร์ การสื่อสาร และการจัดการ เพื่อแปลงข้อมูลให้เป็นข้อมูลเชิงลึกและนำไปใช้เพื่อช่วยในการตัดสินใจ และใช้ประโยชน์ในงานด้านต่าง ๆ บทความนี้ได้กล่าวถึงความรู้พื้นฐานของ Big Data คุณลักษณะของ Big Data นิยามของ Data Science ทักษะที่สำคัญต่อการทำงานด้าน Data Science ขั้นตอนการทำ Data Science อาชีพ Data Scientist และการทำงานด้าน Data Science ในกลุ่มธุรกิจประเภทต่าง ๆ

คำสำคัญ: ข้อมูลขนาดใหญ่, วิทยาศาสตร์ข้อมูล, นักวิทยาศาสตร์ข้อมูล

Abstract

Data Science is a new interdisciplinary field that synthesizes and builds on statistics, informatics, computing, communication and management in order to transform data to insights for beneficial decision making and other application. This article describes its meaning and characteristics of Big Data, definition of Data Science, technical skill needed for Data Science, the main steps in Data Science, definition of Data Scientist and Data Science in various types of business.

Keywords: big data, data science, data scientist



บทนำ

สื่อสังคมออนไลน์ (social media) ได้เข้ามา มีบทบาทในชีวิตประจำวันมากขึ้น ข้อมูลพฤติกรรมของผู้บริโภคคือสิ่งที่หลาย ๆ องค์กรต่างต้องการ เพื่อนำไป วิเคราะห์หาข้อมูลเชิงลึกที่จะนำไปใช้ประโยชน์ได้ จาก

การเพิ่มขึ้นของข้อมูลจำนวนมากมหาศาลที่อยู่ในหลากหลาย รูปแบบและมีการเปลี่ยนแปลงอย่างรวดเร็ว ซึ่งมาพร้อมกับเทคนิคและเครื่องมือต่าง ๆ ที่ใช้เพื่อวิเคราะห์ข้อมูล ขนาดใหญ่ หรือ Big Data ส่งผลให้เกิดศาสตร์ที่เรียกว่า Data Science หรือ วิทยาศาสตร์ข้อมูล โดยบุคลากร

ในสายงานนี้จำเป็นต้องมีทักษะและความเชี่ยวชาญในการทำงานกับข้อมูล อาจเรียกว่า นักวิทยาศาสตร์ข้อมูล (data scientist) และกำลังเป็นอาชีพที่องค์กรต่าง ๆ ต้องการเป็นจำนวนมาก ในบทความนี้มีวัตถุประสงค์เพื่อนำเสนอให้เห็นถึงทักษะและความรู้พื้นฐานที่จำเป็นสำหรับนักวิทยาศาสตร์ข้อมูล เพื่อเป็นพื้นฐานที่สำคัญก่อนจะต่อยอดความรู้ความเชี่ยวชาญเฉพาะด้านต่อไปในอนาคต

Big Data คืออะไร

Big Data หมายถึง ปริมาณข้อมูลที่มีขนาดใหญ่ มหาศาล ซึ่งเกินกว่าขีดความสามารถในการประมวลผลของระบบฐานข้อมูลธรรมดาที่จะรองรับได้ ประโยชน์ของ Big Data มีหลายประการ ที่เด่นชัด 2 ประการ คือ (1) การวิเคราะห์ (analytical use) คือ การวิเคราะห์ข้อมูลที่มีปริมาณมหาศาลเหล่านั้นเพื่อให้เห็นความรู้ที่ซ่อนอยู่ลึก และ (2) การเกิดผลิตภัณฑ์ใหม่ (enabling new products) โดยการนำข้อมูลที่มีปริมาณมหาศาลนั้นมาใช้สร้างผลิตภัณฑ์ หรือพัฒนาการบริการ เพื่อให้เหมาะสมตามความต้องการของลูกค้าหรือผู้ใช้ (Kuchaisit, 2013)

Big Data เป็นเทคโนโลยีที่องค์กรหลายแห่งนำเข้ามาใช้ในการประมวลผล วิเคราะห์ข้อมูลต่าง ๆ โดยเฉพาะข้อมูลเกี่ยวกับลูกค้าที่มาจากรายการค้า การให้บริการลูกค้า การดำเนินงานด้านต่าง ๆ ส่งผลให้มีการปรับปรุงการบริหารความสัมพันธ์กับลูกค้า กลยุทธ์การให้บริการ รวมไปถึงสินค้าและผลิตภัณฑ์ (Intojunyong, 2015)

คุณลักษณะของ Big Data

Big Data ไม่ได้หมายถึงข้อมูลขนาดใหญ่เท่านั้น แต่จะต้องประกอบไปด้วยคุณลักษณะต่าง ๆ อย่างน้อย 3 ลักษณะ (3V) อันได้แก่ Volume, Velocity และ Variety ในบางบทความอาจกล่าวว่าคุณลักษณะของ Big Data มี 5 ลักษณะ (5V) ดังแสดงในภาพ 1 ซึ่งสามารถอธิบายคุณลักษณะต่าง ๆ ได้ดังนี้ (Pinijkitcharoenkul, 2016; Intojunyong, 2015; Kuchaisit, 2013)

1. Volume (ปริมาณ) หมายถึง ข้อมูลที่มีปริมาณมากในระดับ Terabyte หรือ Petabyte หรือ Exabyte ขึ้นไป
2. Velocity (ความเร็ว) หมายถึง ข้อมูลที่มีการเปลี่ยนแปลงตลอดเวลาและรวดเร็ว หรืออาจหมายถึงความเร็วของการสร้างข้อมูลและประมวลผลข้อมูล
3. Variety (ความหลากหลาย) หมายถึง ลักษณะรูปแบบของข้อมูลที่แตกต่างกัน อันเนื่องมาจากแหล่งข้อมูลที่ต่างประเภทกัน ลักษณะของข้อมูลอาจเป็นแบบมีโครงสร้าง แบบกึ่งโครงสร้าง หรือแบบไร้โครงสร้าง
4. Veracity (ความถูกต้อง) หมายถึง คุณภาพและความน่าเชื่อถือของข้อมูล เพราะหากนำข้อมูลที่ไม่มีความน่าเชื่อถือมาวิเคราะห์ย่อมได้ผลลัพธ์ที่ไม่น่าเชื่อถือ
5. Value (คุณค่า) หมายถึง ข้อมูลที่มีคุณค่า ซึ่งสามารถนำไปใช้ประโยชน์ได้จริง หรือมีมูลค่าและความสำคัญต่อองค์กร

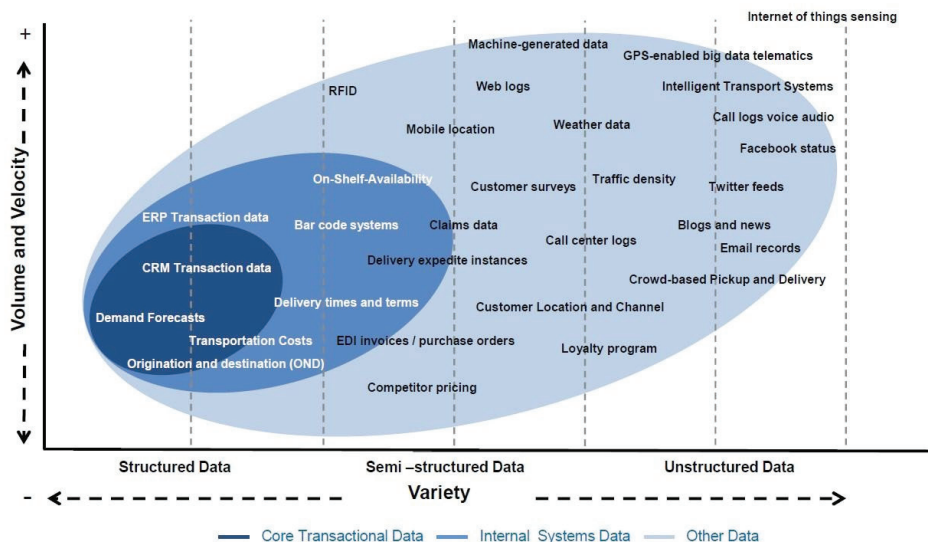
VOLUME	VELOCITY	VARIETY	VERACITY	VALUE
• Terabyte • Records/Arch • Transactions • Tables, Files	• Batch • Real-Time • Processes • Streams	• Structured • Unstructured • Multi-factor • Probabilistic	• Trustworthiness • Authenticity • Availability • Accountability	• Statistical • Events • Correlations • Hypothetical

ภาพ 1 คุณลักษณะของ Big Data

ที่มาของ Data Science

จากการเพิ่มขึ้นของการใช้งานอินเทอร์เน็ต และเทคโนโลยี Internet of Things ทำให้ปริมาณข้อมูลเพิ่มขึ้นอย่างมหาศาล ซึ่งมีทั้งข้อมูลแบบมีโครงสร้าง (structured data) แบบกึ่งโครงสร้าง (semi-structured data) และแบบไร้โครงสร้าง (unstructured data) ในภาพ 2 เป็น

แบบจำลองแสดงปริมาณ (volume) และความเร็ว (velocity) โดยเฉลี่ย เมื่อเทียบกับแหล่งข้อมูลที่หลากหลาย (variety) ซึ่งข้อมูลส่วนใหญ่อยู่ในรูปแบบไร้โครงสร้าง ซึ่งเกิดจากการใช้งาน Social Media เช่น Facebook Twitter YouTube E-mail ไฟล์เอกสาร ไฟล์ภาพ เสียง วิดีโอ รวมไปถึงข้อมูลจาก Sensor ต่าง ๆ ซึ่งข้อมูลมากมายเหล่านี้ สามารถนำไปวิเคราะห์ เพื่อใช้ประโยชน์ต่อไปได้



ภาพ 2 SCM Data Volume and Velocity vs. Variety

ที่มา จาก *Big data analytics in supply chain management: Trends and related research*, โดย Rozados and Tjahjono, 2014, สืบค้นจาก https://www.researchgate.net/publication/270506965_Big_Data_Analytics_in_Supply_Chain_Management_Trends_and_Related_Research

พุทธศักดิ์ ตันติสุทธิเวท (Tantisuttivet, 2016) กล่าวว่า การที่มีข้อมูลมากมายมหาศาลในรูปแบบที่หลากหลายไหลเวียนอยู่ และถ้าสามารถจัดการกับข้อมูลที่ไหลเวียนอยู่เหล่านี้ได้ จะสามารถนำมาใช้ประโยชน์ได้อีกมหาศาล และความพยายามที่จะนำข้อมูลเหล่านี้มาใช้ประโยชน์ก็คือศาสตร์ Data Science นั่นเอง

อภิพจน์ เพี้ยศักดิ์ (Piasak, 2017) กล่าวว่า ข้อมูล หรือ Data จากการใช้งาน Social Media คือแหล่งน้ำมัน

ดิบที่มีมูลค่ามหาศาล เนื่องจากสามารถนำข้อมูลเหล่านี้มาวิเคราะห์เพื่อหา Insight หรือ Recommendation สนับสนุนการตัดสินใจในการดำเนินธุรกิจ ซึ่งเป็นที่มาของกระบวนการหรือศาสตร์ (science) ในการกลั่นข้อมูลที่มีอยู่ออกมาให้เกิดประโยชน์มากที่สุด โดยกระบวนการทั้งหมดจะผสมผสานทั้งในด้าน Statistic, Mathematic, Programming และ Technology เพื่อตั้งสมมติฐาน ทดลอง และหาผลลัพธ์จากข้อมูลที่มีอยู่ ทั้งหมดนี้คือ Data Science

Data Science คืออะไร

ไต้ ปริญา (Parinya, 2015) กล่าวว่า Data Science คือ การรวมวิชาและเทคโนโลยีทางด้าน Cloud Computing, Big Data, Machine Learning, Data Mining, Statistics และ Internet of Things เข้าไว้ด้วยกัน มีจุดประสงค์เพื่อบริหารจัดการข้อมูลขนาดใหญ่ และสกัดความรู้จากข้อมูลขนาดใหญ่ ซึ่งมีลำดับขั้นตอนประกอบด้วย การนำเข้าข้อมูลหลายประเภท จากหลากหลายอุปกรณ์ตามนิยามของ Internet of Things ข้อมูลถูกจัดเก็บไว้บน Cloud Computing ซึ่งบริหารจัดการตามนิยามของ Big Data จากนั้นข้อมูลจะถูกนำมาตัดสินใจด้วย Machine Learning ถูกนำมาค้นหาความรู้ด้วย Data Mining และถูกนำมาชี้วัดสมมติฐานด้วย Statistics

All Data Science (2018) กล่าวว่า Data Science เป็นสาขาสหวิทยาการเกี่ยวกับกระบวนการและระบบเพื่อดึงความรู้ (knowledge) หรือข้อมูลเชิงลึก (insights) จากข้อมูลในรูปแบบต่าง ๆ ทั้งที่มีโครงสร้างหรือไม่มีโครงสร้าง

Cao (2017) ได้นิยาม Data Science และ Data Products ไว้ว่า Data Science เป็นวิทยาศาสตร์ของข้อมูล หรือ Data Science คือการศึกษาข้อมูล

Data Science คือ สาขาสหวิทยาการใหม่ที่สังเคราะห์และสร้างขึ้นบนพื้นฐานของสถิติ สารสนเทศ คอมพิวเตอร์ การสื่อสาร การจัดการ และสังคมวิทยา เพื่อการศึกษาข้อมูลและสภาพแวดล้อม เพื่อแปลงข้อมูลให้เป็นข้อมูลเชิงลึกและใช้ช่วยในการตัดสินใจ ตามแนวความคิดที่ว่า ข้อมูลสู่ความรู้สู่ภูมิปัญญา

ส่วนผลลัพธ์ของการทำ Data Science คือ Data Products ซึ่งนิยามไว้ว่า Data Products หรือ ผลิตภัณฑ์ข้อมูล เป็นผลที่ส่งมาจากข้อมูลหรือขับเคลื่อนด้วยข้อมูล และสามารถทำได้ ข้อค้นพบ คำแนะนำ การบริการ การทำนาย การตัดสินใจ กรอบความคิด แบบจำลอง รูปแบบ กระบวนการทัศน์ เครื่องมือ หรือระบบ ซึ่ง Data Products ที่มีค่าที่สุด นั่นคือ องค์กรความรู้ สถิติปัญญา ภูมิปัญญา และการตัดสินใจ

Weihs and Ickstadt (2018) กล่าวว่า หนึ่งในคำจำกัดความที่ครอบคลุมที่สุดของ Data Science ซึ่งนำเสนอโดย Cao สามารถเขียนได้ดังสมการ 1 ดังนี้

$$\text{Data Science} = \frac{(\text{statistics} + \text{informatics} + \text{computing} + \text{communication} + \text{sociology} + \text{management})}{(\text{data} + \text{environment} + \text{thinking})} \quad (1)$$

ดังนั้น Data Science ในภาษาไทยอาจใช้คำว่า วิทยาศาสตร์ข้อมูล หรือ วิทยาการข้อมูล ซึ่งหมายถึง การใช้กระบวนการทางวิทยาศาสตร์เข้ามาจัดการกับข้อมูล

ในบทความของ Cao (2017) ได้นำเสนอคำศัพท์ที่น่าสนใจเกี่ยวข้องกับงานด้าน Data Science ซึ่งผู้เขียนได้เรียบเรียงไว้ ดังตาราง 1

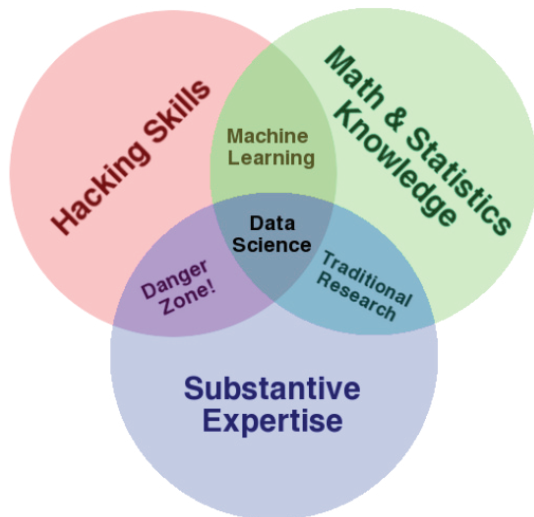
ตาราง 1

คำศัพท์ใน Data Science

คำศัพท์	คำอธิบาย
Advanced analytics	ทฤษฎี เทคโนโลยี เครื่องมือ และกระบวนการ ที่ช่วยให้เข้าใจอย่างลึกซึ้ง และค้นพบข้อมูลเชิงลึกจาก Big data ที่สามารถนำไปใช้งานได้ (actionable insights) ซึ่งไม่สามารถทำได้โดยการวิเคราะห์และประมวลผลข้อมูล ด้วยทฤษฎี เทคโนโลยี เครื่องมือ และกระบวนการ รูปแบบเดิม
Big data	ข้อมูลที่มีขนาดใหญ่มาก และ/หรือ มีความซับซ้อนมาก ที่จะจัดการให้มีประสิทธิภาพ และ/หรือประสิทธิผล ด้วยทฤษฎีที่เกี่ยวข้องกับข้อมูล เทคโนโลยี และ เครื่องมือ รูปแบบเดิม
Data analysis	การประมวลผลข้อมูลด้วยรูปแบบดั้งเดิม ทั้งทฤษฎี เทคโนโลยี และเครื่องมือ (เช่น สถิติแบบดั้งเดิม คณิตศาสตร์ หรือตรรกศาสตร์) เพื่อให้ได้รับข้อมูลที่เป็นประโยชน์ และเป้าหมายในการปฏิบัติ
Data analytics	ทฤษฎี เทคโนโลยี เครื่องมือ และกระบวนการ ที่ทำให้เกิดความเข้าใจในเชิงลึก และการค้นพบข้อมูลเชิงลึกที่สามารถนำไปใช้งานได้ (actionable insight) มีรูปแบบการวิเคราะห์ข้อมูล ประกอบไปด้วย การวิเคราะห์เชิงพรรณนา (descriptive analytics) การวิเคราะห์เชิงทำนาย (predictive analytics) และ การวิเคราะห์แบบให้คำแนะนำ (prescriptive analytics)
Data science	ศาสตร์ที่เกี่ยวข้องกับข้อมูล
Data scientist	บุคคลที่มีบทบาทอย่างมากกับข้อมูล
Descriptive analytics	รูปแบบการวิเคราะห์ข้อมูลโดยใช้สถิติเพื่ออธิบายผลจากข้อมูลที่ได้รับ เป็นรูปแบบการวิเคราะห์ข้อมูลแบบพื้นฐานที่สุด
Predictive analytics	รูปแบบการวิเคราะห์ข้อมูลที่เป็นการทำนาย พยากรณ์ เหตุการณ์ที่จะเกิดขึ้นในอนาคต โดยใช้ข้อมูลจากอดีต เป็นรูปแบบการวิเคราะห์ข้อมูลที่มีความซับซ้อนมากขึ้น
Prescriptive analytics	รูปแบบการวิเคราะห์ข้อมูลและคาดการณ์ผลลัพธ์ที่จะเกิดขึ้น และให้คำแนะนำทางเลือกในการตัดสินใจที่เหมาะสมที่สุด เป็นรูปแบบการวิเคราะห์ข้อมูลที่มีความซับซ้อนและยากที่สุด
Explicit analytics	มุ่งเน้นไปที่การวิเคราะห์เชิงพรรณนา (descriptive analytics) โดยการรายงาน การแจ้งเตือน และการคาดการณ์
Implicit analytics	มุ่งเน้นไปที่การวิเคราะห์เชิงลึก (deep analytics) โดยใช้โมเดลเชิงทำนาย (predictive modeling) การเพิ่มประสิทธิภาพ การวิเคราะห์แบบให้คำแนะนำ (prescriptive analytics) และส่งความรู้ที่สามารถนำไปใช้งานได้ (actionable knowledge)
Deep analytics	การวิเคราะห์ข้อมูลที่ทำให้ได้ความเข้าใจอย่างลึกซึ้ง ว่าเพราะเหตุใดเหตุการณ์จึงเกิดขึ้น และเกิดขึ้นได้อย่างไร เหตุการณ์จะเกิดขึ้นหรือไม่ ซึ่งไม่สามารถระบุได้โดยใช้การวิเคราะห์เชิงพรรณนา

ทักษะที่สำคัญต่อการทำงานด้าน Data Science

แผนภาพ Data Science ที่สร้างขึ้นโดย Drew Conway เป็นแผนภาพที่ถูกอ้างอิงถึงมากที่สุด เพื่อใช้อธิบายถึงความรู้และทักษะที่จำเป็นต่อการทำงานด้าน Data Science



ภาพ 3 แผนภาพ Data Science

จากภาพ 3 มีประเด็นสำคัญที่อธิบายถึง 3 ทักษะหลัก ดังนี้

1. Hacking Skills หมายถึง ทักษะเกี่ยวกับการใช้เทคโนโลยีและเทคนิคเพื่อทำงานในด้านต่าง ๆ ได้แก่ การเขียนโปรแกรม ด้วยภาษา Python หรือ ภาษา R การจัดการฐานข้อมูลโดยใช้ MySQL, PostgreSQL, Cassandra, MongoDB หรือ CouchDB การสร้างรายงานและแสดงผลข้อมูล (visualization) โดยใช้เครื่องมือ เช่น D3 Tableau, Qlikview หรือ R Markdown การใช้เทคโนโลยีเพื่อจัดการกับ Big Data เช่น Hadoop, MapReduce หรือ Spark เป็นต้น

2. Math & Statistics Knowledge หมายถึง ทักษะทางคณิตศาสตร์และสถิติ ซึ่งเป็นสิ่งจำเป็นสำหรับนักวิทยาศาสตร์ข้อมูล สถิติเป็นส่วนสำคัญของการวิเคราะห์ข้อมูล ในขณะที่พีชคณิตเชิงเส้นและแคลคูลัสเป็นพื้นฐานในการทำความเข้าใจ และการประยุกต์ใช้อัลกอริทึมของ Machine Learning

3. Substantive Expertise หรือ Domain Knowledge หมายถึง ความรู้ความเชี่ยวชาญเฉพาะด้าน ซึ่งจะช่วยให้ นักวิทยาศาสตร์ข้อมูลสามารถเริ่มโครงการได้โดยมีเป้าหมายที่ชัดเจน รวมถึงความสามารถในการสื่อสารอย่างมีประสิทธิภาพ การตั้งคำถามที่ดีต้องอาศัยความเข้าใจในเรื่องนั้นจริง ๆ

จุดตัดของ Hacking skills และ Substantive expertise ในแผนภาพ Data Science ส่วนที่เรียกว่า Danger Zone! หมายถึง การทำงานโดยไม่มีความรู้ด้านคณิตศาสตร์และสถิติ อาจทำให้เลือกใช้อัลกอริทึมที่ไม่เหมาะสมกับงาน ซึ่งทำให้ได้ผลลัพธ์ที่ไม่ถูกต้อง (All Data Science, 2018)

Rohrer (2016) กล่าวว่า Machine Learning เป็นกลไกสำคัญในการขับเคลื่อน Data science ซึ่งมีเพียง 5 คำถามเท่านั้น ที่ต้องสามารถตอบได้ โดยใช้ อัลกอริทึมต่าง ๆ นั่นคือ (1) Is this A or B? (two-class classification) หรือ Is this A or B or C or D? (multi-class classification) (2) Is this weird? (anomaly detection) (3) How much/how many? (regression) (4) How is it organized? (clustering) และ (5) What should I do next? (reinforcement learning)

ขั้นตอนการทำ Data Science

กระบวนการหนึ่งที่ยอมรับใช้ในการวิเคราะห์ข้อมูลคือการวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining ตามมาตรฐาน CRISP-DM (Cross Industry Standard Process for Data Mining) ซึ่งมีทั้งหมด 6 ขั้นตอน ได้แก่ (1) Business understanding (2) Data understanding (3) Data preparation (4) Modeling (5) Evaluation และ (6) Deployment

Weih and Ickstadt (2018) ได้นำเสนอแนวทางในการทำ Data Science โดยประยุกต์มาจากกระบวนการวิเคราะห์ข้อมูลตามมาตรฐาน CRISP-DM ซึ่งมีลำดับขั้นตอนทั้งหมด 8 ขั้นตอน นำเสนอดังตาราง 2

ตาราง 2

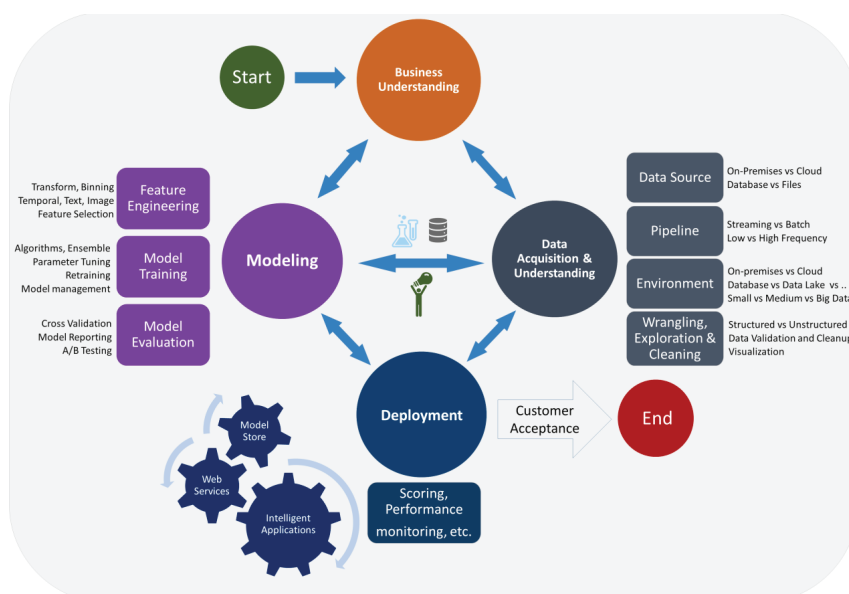
ขั้นตอนการทำ Data Science: เปรียบเทียบระหว่าง CRISP-DM และข้อเสนอของ Weihs and Ickstadt

CRISP-DM	Weihs and Ickstadt's Proposal
Business Understanding	Data Acquisition and Enrichment, Data Storage and Access
Data Understanding, Data Preparation	Data Exploration
Modeling	Data Analysis and Modeling, Optimization of Algorithms
Evaluation	Model Validation and Selection
Deployment	Representation and Reporting of Results, Business Deployment of Results

กลุ่ม TDSP (The Team Data Science Process) ได้นำเสนอวงจรชีวิต ซึ่งสามารถใช้เป็นโครงสร้างเพื่อจัดการโครงการเกี่ยวกับวิทยาศาสตร์ข้อมูล แต่ทั้งนี้อาจใช้วงจรชีวิตตามมาตรฐานอื่น เช่น การวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining (Cross Industry Standard Process for Data Mining--CRISP-DM) หรือ การค้นหาคำรู้ในฐานข้อมูล (Knowledge Discovery in Databases --KDD) วงจรชีวิตนี้ถูกออกแบบมาสำหรับโครงการวิทยาศาสตร์ข้อมูล โดยมีวัตถุประสงค์เพื่อใช้เป็นส่วนหนึ่งของแอปพลิเคชันอัจฉริยะ (intelligent applications)

ซึ่งแอปพลิเคชันเหล่านี้นำไปปรับใช้กับการเรียนรู้ด้วยเครื่อง (machine learning) หรือแบบจำลองปัญญาประดิษฐ์ (artificial intelligence) เพื่อการวิเคราะห์ข้อมูลเชิงทำนาย (predictive analytics) ต่อไป (Thakurta et al., 2017)

วงจรชีวิตที่นำเสนอโดย TDSP ประกอบด้วย 5 ขั้นตอนหลัก ได้แก่ (1) Business understanding (2) Data acquisition and understanding (3) Modeling (4) Deployment และ (5) Customer acceptance แสดงดังภาพ 4



ภาพ 4 วงจรชีวิตในการทำ Data Science

Data Scientist คือใคร

Donoho (2015) กล่าวว่า Data Scientist หมายถึง ผู้เชี่ยวชาญที่ใช้วิธีการทางวิทยาศาสตร์เพื่อปลดปล่อยและสร้างความหมายจากข้อมูลดิบ

ไท่ ปริญญา (Parinya, 2015) กล่าวว่า บุคลากรที่ทำงานด้านวิทยาศาสตร์ข้อมูลเหล่านี้ ต้องรู้จักใช้เครื่องมือ เช่น รู้จักใช้ Amazon Web Services เพื่อบริหารจัดการ Public Cloud Computing หรือ รู้จักการติดตั้งปรับแต่ง Hadoop เพื่อบริหารจัดการ Distributed Computing ให้อยู่ในรูปของ Private Cloud Computing รู้จักนิยาม NoSQL เพื่อจัดหาฐานข้อมูลสำหรับบริหารจัดการ Big Data รู้จักใช้ MaLab หรือ ภาษา R หรือ Weka หรือ RapidMiner เพื่อทำ Machine Learning หรือ Data Mining รู้จักใช้ SPSS เพื่อวิเคราะห์ Statistics เป็นต้น รู้จักการปรับแต่งและสั่งงานหน่วยประมวลผล เช่น Raspberry Pi หรือ หน่วยควบคุม เช่น Arduino เป็นต้น ซึ่งงานของนักวิทยาศาสตร์ข้อมูล จำเป็นต้องใช้ความรู้และทักษะหลายด้าน จึงทำให้เป็นอาชีพที่เป็นที่นิยมในขณะนี้

เว็บไซต์ MGR Online (2018) ได้นำเสนอผลวิจัยของบริษัทแมคคินซี ดิจิทัล เอเชีย เมื่อเดือนพฤศจิกายน 2560 ระบุว่า ภาพรวมของทั่วโลกมีข้อมูลที่สามารถปรับให้เป็นระบบอัตโนมัติได้ 50% แต่ที่จะทำได้จริงในปี 2573 หรือ อีกกว่า 10 ปี ข้างหน้า มีเพียง 15% ขณะที่ประเทศไทยมีข้อมูลที่สามารถปรับให้เป็นระบบอัตโนมัติได้ 55% แต่จะทำได้จริง 17% ในปี 2573 เมื่อ AI เข้ามามีบทบาท จึงจำเป็นต้องมีการสร้างทักษะของคนเพื่อทำงานร่วมกันได้ ดังนั้นอาชีพนักวิทยาศาสตร์ข้อมูล และวิศวกรด้านข้อมูล จึงยังขาดแคลนอยู่ ซึ่งตัวเลขทั่วโลกพบว่าขาดแคลนถึง 30 ล้านคน ดังนั้นภาครัฐจึงควรเริ่มต้นด้วยการส่งเสริมการศึกษา และภาคเอกชนควรสร้างคอร์สอบรมหลักสูตรให้กับพนักงาน เพื่อเพิ่มทักษะให้สอดคล้องกับวิชาชีพที่ขาดแคลน

ดังนั้น Data Scientist จึงหมายถึง นักวิทยาศาสตร์ข้อมูล หรือ นักวิทยาการข้อมูล ซึ่งเป็นผู้ทำหน้าที่ในการวิเคราะห์ข้อมูล และใช้เทคโนโลยีเพื่อจัดการกับข้อมูลขนาดใหญ่

Data Scientist ทำงานที่ไหน

ข้อมูลจาก Big Data Analytics Thailand (2016) ได้วิเคราะห์ว่าในเมืองไทย บริษัทที่มีฐานข้อมูลขนาดใหญ่ และมีความต้องการในการจ้าง Data Scientist มาทำงาน จะอยู่ใน 5 กลุ่มธุรกิจ ดังนี้

1. ธนาคารพาณิชย์

ลักษณะงาน Data Science ในธนาคารแบ่งเป็นสองประเภทหลัก ๆ คือ ประเภทแรกเรียกว่า Credits ซึ่งจะเกี่ยวกับการวิเคราะห์ความเสี่ยงที่อยู่ในพอร์ตของธนาคาร รวมถึง Predictive analytics ทำนายว่าลูกค้าที่มาขอกู้เงินจะเป็นหนี้เสียหรือไม่ ส่วนงานประเภทที่สองเรียกว่า Customer analytics ซึ่งเป็นการวิเคราะห์พฤติกรรมลูกค้า ทำการแบ่งกลุ่มลูกค้า หารูปแบบของลูกค้า รวมถึงทำโมเดล Predictive analytics ทำนายพฤติกรรมลูกค้า ในลักษณะต่าง ๆ

2. E-Commerce และ เว็บไซต์

บริษัท E-Commerce ต้องการ Data Scientist เพื่อทำการวิเคราะห์ข้อมูลจากการที่ลูกค้ามาเข้าชมเว็บไซต์และเลือกซื้อสินค้า ซึ่งการวิเคราะห์ทำได้ตั้งแต่แบบพื้นฐาน เช่น การแบ่งกลุ่มลูกค้า การดูยอดขายรายสินค้า ไปจนถึงการวิเคราะห์แบบขั้นสูง เช่น การทำโมเดล Recommendation engine ซึ่งเป็นการแนะนำสินค้าที่ลูกค้าอาจถูกใจโดยดูจากพฤติกรรมที่ผ่านมาของลูกค้า

3. ธุรกิจค้าปลีกค้าส่ง

บริษัทในธุรกิจค้าปลีกค้าส่งหลายแห่งมีการออก Loyalty programs เพื่อให้ลูกค้าแสดงบัตรสะสมแต้มต่าง ๆ เมื่อซื้อสินค้า ทำให้บริษัทเหล่านี้เริ่มสามารถสร้างฐานข้อมูลลูกค้า และจะมีความต้องการบุคลากรทาง Data Science เพื่อทำการวิเคราะห์ข้อมูลลูกค้าต่อยอดต่อไปได้

4. ธุรกิจโทรคมนาคม

ลักษณะงาน Data Science ของธุรกิจโทรคมนาคม จะมีส่วนของ Customer analytics เพื่อนำไปใช้ในการออกแคมเปญเชิงการจัดการลูกค้าสัมพันธ์ (CRM) ข้อมูลโทรคมนาคมมักจะเป็นข้อมูลขนาดใหญ่และจะมีข้อมูลในรูปแบบไม่มีโครงสร้าง ดังนั้นสายงานโทรคมนาคมจึงต้องการ

นักวิเคราะห์ที่มีความสามารถในการจัดการและวิเคราะห์ข้อมูลขนาดใหญ่ได้ดีและสามารถทำงานกับข้อมูลแบบไม่มีโครงสร้างได้

5. ธุรกิจการผลิตภาคอุตสาหกรรม

งาน Data Science ในธุรกิจ Manufacturing มีทั้งการวิเคราะห์ข้อมูลเพื่อนำไปสู่การวางแผน รวมถึงการวางแผนกลยุทธ์ วิเคราะห์ข้อมูลเพื่อหาโอกาสทางธุรกิจใหม่ นอกจากนี้บริษัทที่มีหน่วยงานปลายทางสินค้าให้กับผู้บริโภคโดยตรงและมีการสร้างฐานข้อมูลลูกค้า จะมีงาน Customer analytics เช่นกัน

บทสรุป

จากการเพิ่มขึ้นของข้อมูลหลากหลายรูปแบบอย่างรวดเร็ว ทำให้เกิดข้อมูลขนาดมหาศาลที่สามารถนำข้อมูลเหล่านั้นมาใช้ให้เกิดประโยชน์ต่อองค์กรได้ จึงนำไปสู่กระบวนการวิเคราะห์ข้อมูลขนาดใหญ่ (big data analytics) โดยใช้เครื่องมือ และเทคโนโลยีต่าง ๆ ผสมผสานกัน เพื่อให้ได้ผลลัพธ์ที่ต้องการ ทำให้เกิดศาสตร์ด้าน Data Science หรือ วิทยาศาสตร์ข้อมูล ซึ่งผลลัพธ์ที่ได้จากการทำวิทยาศาสตร์ข้อมูลนั้น เรียกว่า Data Product หรือ ผลิตภัณฑ์ข้อมูล ซึ่งผลิตภัณฑ์ข้อมูลนี้อาจ

อยู่ในรูปแบบของข้อมูลเพื่อช่วยสนับสนุนการตัดสินใจ เพื่อเพิ่มประสิทธิภาพการทำงาน เพื่อให้ได้แนวทางในการปรับปรุงการให้บริการ หรือเพื่อประโยชน์ในด้านอื่น ๆ ซึ่งการทำวิทยาศาสตร์ข้อมูลสามารถนำไปประยุกต์ใช้ได้ทั้งในหน่วยงานภาครัฐและภาคเอกชน ที่มีความต้องการวิเคราะห์พฤติกรรมของลูกค้า วิเคราะห์ข้อมูลลูกค้า เช่น งานธนาคาร ธุรกิจโทรคมนาคม ธุรกิจค้าปลีกค้าส่ง ธุรกิจการผลิตภาคอุตสาหกรรม กลุ่มธุรกิจ E-Commerce อินเทอร์เน็ต และเว็บไซต์

ดังนั้นบุคลากรที่ทำงานด้านนี้ หรือนักวิทยาศาสตร์ข้อมูล จึงจำเป็นต้องมีทักษะ ความรู้ ความเชี่ยวชาญ และมีความสามารถในการใช้เทคโนโลยีสำหรับจัดการกับข้อมูลขนาดใหญ่ การใช้ประโยชน์จากข้อมูลที่มีการเพิ่มขึ้นอย่างรวดเร็วและความหลากหลาย จำเป็นต้องผ่านกระบวนการวิเคราะห์ สังเคราะห์ สรุปผล และนำเสนอในรูปแบบที่เหมาะสม ซึ่งผู้ที่ทำงานในสายงานดังกล่าว จึงจำเป็นต้องมีความรู้พื้นฐานในด้านคณิตศาสตร์ สถิติ และการเขียนโปรแกรม ซึ่งการส่งเสริมและพัฒนาบุคลากรให้มีทักษะทางด้านนี้จึงมีความสำคัญในยุคที่มีการใช้ประโยชน์จากข้อมูลสารสนเทศ



References

- All Data Science. (2018). *What is data science?*. Retrieve form <https://alldatascience.com/2018/01/02/what-is-data-science/>
- Big Data Analytics Thailand. (2016). *Where is the Data Scientist work*. Retrieved from <https://www.facebook.com/truebigdata/posts/678151292336509> (in Thai)
- Cao, L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys*, 50(3), 43-48.
- Donoho, D. (2015). *50 years of data science*. Retrieved from <http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf>
- Intojunyong, S. (2015). Whether big data is essential for all organizations or not. *Executive Journal*, 38(146), 1-8. (in Thai)
- Kuchaisit, S. (2013). Transforms a new era of data “big data”. *Executive Journal*, 33(1), 22-28. (in Thai)
- MGR Online. (2018). *‘Data scientist - data engineer’ high demands*. Retrieved from <https://mgronline.com/cyberbiz/detail/9610000024219> (in Thai)
- Parinya, T. (2015). *What is data science and data scientist*. Retrieved from <http://www.parinya.net/node/2251> (in Thai)
- Piasak, A. (2017). *What is data science*. Retrieved from <https://blog.derlivery.com/data-science-คืออะไร-57d7530554bf> (in Thai)
- Pinijkitcharoenkul, S. (2016). *Get to know big data*. Retrieved from http://km.psru.ac.th/wp-content/uploads/2016/11/big-data_v2.pdf (in Thai)
- Rohrer, B. (2016). *What questions can data science answer?*. Retrieved from <https://www.kdnuggets.com/2016/01/questions-data-science-answer.html>
- Rozados, I. V., & Tjahjono, B. (2014). *Big data analytics in supply chain management: Trends and related research*. Retrieved from https://www.researchgate.net/publication/270506965_Big_Data_Analytics_in_Supply_Chain_Management_Trends_and_Related_Research
- Tantisuttivet, P. (2016). *What is data science that digital businesses need to know*. Retrieved from <https://brandinside.asia/data-science-sexiest-job/> (in Thai)
- Thakurta, D. G., Ericson, G., Martens., J., Severtson, R. B., & Rohm, W. A. (2017). *The team data science process lifecycle*. Retrieved from <https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle>
- Weihs, C., & Ickstadt, K. (2018). *Data science: The impact of statistics*. Retrieved from <https://link.springer.com/content/pdf/10.1007%2Fs41060-018-0102-5.pdf>

