Data Scientist

อาชีพใหม่ของการเป็นนักวิทยาศาสตร์ข้อมูล

ศิระ ศิลานนท์
สถาบันวิจัยวิทยาศาสตร์และเทคโนโลยีแห่งประเทศไทย
35 หมู่ที่ 3 เทคโนธานี ตำบลคลองห้า อำเภอคลองหลวง จังหวัดปทุมธานี 12120

ในวารสารวิทยาศาสตร์และเทคโนโลยี วว. ปีที่ 32 ฉบับที่ 3 เราได้รู้จักกับ Big Data เบื้องต้นกันไปแล้ว มาในฉบับนี้เราจะมารู้จักเรื่องของ "ข้อมูลบิ๊กๆ" นี้กันให้มากขึ้นอีกกับคำว่า Data Science คืออะไร และถ้าจะทำงานทางด้านนี้ อาชีพที่กำลังฮอตฮิตเป็นที่ต้องการและกล่าวถึงกันมากในเวลานี้ เขาต้องทำอะไรกันบ้าง และประเทศไทยอยู่ตรงไหนแล้ว

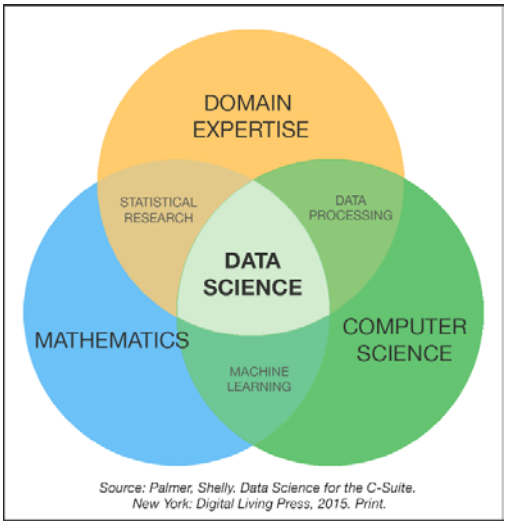
การจะนำ Big Data* มาใช้ให้เกิดประโยชน์ได้ จำเป็นต้องมี Data Science เข้ามา ซึ่งหากแปลตรงตัวว่า “วิทยาศาสตร์ข้อมูล” ก็อาจยังไม่เห็นภาพเท่ากับบอกว่า “วิทยาการข้อมูล” น่าจะเหมาะสมกว่า เช่นเดียวกับที่เราเรียก Computer Science ว่าวิทยาการคอมพิวเตอร์นั่นเอง ดังนั้น Data Science หรือวิทยาการข้อมูลก็คือ ศาสตร์หรือวิทยาการทางด้านการวิเคราะห์และจัดการข้อมูลให้เกิดประโยชน์ และผู้ทำงานด้านนี้จึงเรียกกันว่า Data Scientist

*** Big Data คืออะไร แค่ไหนจึงจะเข้าข่ายว่าบิ๊กจริง**

Big Data คือ ข้อมูลที่มีปริมาณมหาศาลและซับซ้อนมาก จนไม่สามารถจัดการได้ด้วยระบบฐานข้อมูลและซอฟต์แวร์ทั่วไปได้ ทำให้ต้องมีเทคนิคและซอฟต์แวร์เฉพาะในการจัดการข้อมูลขนาดใหญ่กว่าปกติขึ้นมา โดย Big Data มีหลักการพิจารณาความใหญ่ได้จาก หลัก 4Vs ซึ่งได้แก่

1. **Volume** ปริมาณข้อมูล ต้องมีจำนวนมหาศาลถึงระดับ Petabyte (10^{15} B) ในข้อมูลชุดเดียว และเมื่อชุดข้อมูลจากหลายแหล่งมารวมกันประมวลผลเข้าจะมีขนาดมหาศาลได้ถึงระดับ Yottabyte (10^{24} B) ลองจินตนาการถึงปริมาณเอกสารที่มีตัวอักษรอยู่ราวหนึ่งล้านล้านล้านล้านตัวอักษรดูนะครับ
2. **Variety** มีความหลากหลายของข้อมูลมากไม่มีรูปแบบที่ตายตัว
3. **Velocity** มีการเพิ่มขึ้นและเปลี่ยนแปลงของข้อมูลไปอย่างรวดเร็วและต่อเนื่องตลอดเวลา เช่น ข้อมูลธุรกรรมทางการเงินของธนาคาร ข้อมูลทรานแซกชันการซื้อขายของตลาดหลักทรัพย์ ข้อมูลสภาพอากาศ ข้อมูลบนโซเชียลมีเดียต่างๆ ที่มาเร็วไปเร็ว
4. **Veracity** มีความถูกต้องและครบถ้วนสมบูรณ์ของข้อมูล

Data Scientist หรือนักวิทยาศาสตร์ข้อมูล คือ ผู้เชี่ยวชาญด้านการวิเคราะห์หาความสัมพันธ์ของข้อมูลขนาดใหญ่ ซึ่งต้องมีทักษะความรู้หลากหลายสาขามาประกอบกัน ทั้งฝั่งของไอทีและฝั่งของธุรกิจงานองค์กรนั้นๆ โดยอาจจะเป็นบุคลากรที่มาจากสายงาน โปรแกรมเมอร์ นักวิเคราะห์ระบบ นักวิเคราะห์ธุรกิจ นักสถิติ ที่ต่อยอดมาสู่การเป็นนักวิทยาศาสตร์ข้อมูลก็ได้ ซึ่งเป็นตำแหน่งงานที่ค่อนข้างใหม่ในประเทศไทย และกำลังเป็นที่ต้องการสูงทั่วโลก ส่งผลให้เป็นอาชีพที่มีรายได้ค่าตอบแทนสูงขึ้นตามไปด้วยในเวลานี้



องค์ความรู้ที่จำเป็นของนักวิทยาศาสตร์ข้อมูล

1. Domain Expertise

ความเชี่ยวชาญในขอบข่ายงานของข้อมูลธุรกิจที่จะนำมาทำด้านนั้นๆ ต้องมีความรู้ ความเข้าใจในกระบวนการและเป้าหมายของธุรกิจเป็นอย่างดี

2. Computer Science

วิทยาการคอมพิวเตอร์ ซึ่งก็คือ ต้องมีทักษะความสามารถในการเขียนโปรแกรมและใช้เครื่องมือเพื่อวิเคราะห์และจัดการข้อมูลขนาดใหญ่ได้ เช่น SQL, Python, R

3. Mathematics

คณิตศาสตร์ ทักษะการคำนวณและเทคนิคต่างๆ ความน่าจะเป็น สถิติ ตรรกศาสตร์

รวมไปถึงองค์ความรู้ที่เฉพาะทางมากขึ้น อาทิ Statistical research, Machine learning, Data processing, Data mining (การทำเหมืองข้อมูล), Data visualization (การแสดงผลข้อมูลให้เป็นภาพ), Artificial Intelligence (AI; ปัญญาประดิษฐ์), ดังนั้นการก้าวมาเป็น Data Scientist ได้ ก็ควรจะจบการศึกษาที่เกี่ยวข้องกับทักษะองค์ความรู้เหล่านี้มา เช่น วิศวกรรมศาสตร์คอมพิวเตอร์ วิทยาการคอมพิวเตอร์ เทคโนโลยีสารสนเทศ สถิติศาสตร์ หรือเศรษฐศาสตร์ และมีประสบการณ์การเขียนโปรแกรมการวิเคราะห์และแสดงผลข้อมูลให้เป็นภาพเข้าใจได้โดยง่าย สูงพอสมควร

หน้าที่ของนักวิทยาศาสตร์ข้อมูล

กระบวนการทำงานหลักๆ มีอยู่ 5 ขั้นตอน เริ่มจาก 1) การตั้งคำถามที่น่าสนใจ นำเอาข้อมูลมาช่วยแก้ปัญหาให้องค์กรได้ 2) การค้นหาและจัดเก็บข้อมูล 3) การสำรวจข้อมูลหาแบบแผนความเชื่อมโยงระหว่างกัน 4) การสร้างแบบจำลองเพื่อวิเคราะห์ข้อมูล และ 5) การสื่อสารผลลัพธ์ที่ได้ให้เห็นภาพและเข้าใจง่าย

หลังจากนำเสนอผลลัพธ์แล้ว ภาครัฐหรือธุรกิจอาจตัดสินใจนำไปใช้สร้างนวัตกรรมให้กับองค์กรต่อไป หรือปรับปรุงการวิเคราะห์เพิ่มเติมและกลับมาเริ่มที่กระบวนการแรกคือการตั้งคำถามใหม่ๆ อีกครั้ง

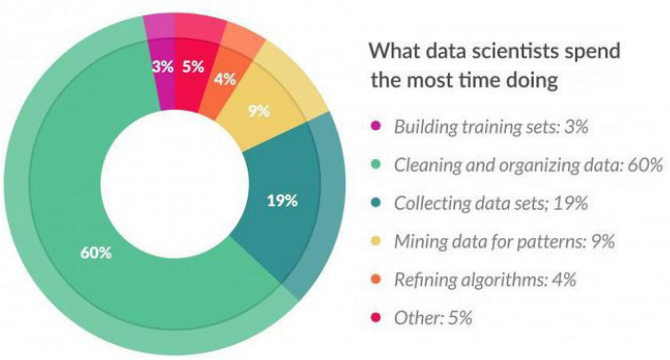
ดังนั้นหัวใจสำคัญที่สุดของนักวิทยาศาสตร์ข้อมูล ก็คือ การวิเคราะห์ข้อมูลขนาดใหญ่และสร้างแบบจำลองชุดข้อมูลเพื่อตอบโจทย์ปัญหาทางธุรกิจที่ต้องการ แต่ก่อนจะมาถึงตรงนี้ได้นั้น งานของนักวิทยาศาสตร์ข้อมูลราว 60% ถูกใช้ไปกับการคลีนข้อมูลและจัดระเบียบข้อมูล (cleaning and organizing data) ซึ่งต้องทำการปรับแก้ไขข้อมูลที่ผิดพลาดจากบุคคล (human error) ให้ข้อมูลกลับมาสมบูรณ์ก่อนจะนำไปใช้ ตัวอย่างเช่น

• ข้อมูลที่คลาดเคลื่อนผิดไปจากความเป็นจริง เช่น ข้อมูลหมายเลขโทรศัพท์มือถือของไทยก็ต้องมี 10 หลัก หมายเลขบัตรประชาชนยืนยันตัวบุคคลก็ต้องมี 13 หลัก ถ้ามีมาแค่ 9 หลัก นั่นคือไม่ถูกต้อง ไม่ครบถ้วน ใช้ไม่ได้จริงนั่นเอง หรือมีข้อมูลระบุว่า นาย ก. อายุ 217 ปี เด็กชาย ข. สูง 240 เซนติเมตร ซึ่งทั้งอายุและความสูงนี้อยู่นอกช่วงที่จะเป็นจริงได้ ศัพท์ทางสถิติเรียกข้อมูลแบบนี้ว่า outliers หรือค่าผิดปกติ ที่จะผิดแผกแตกต่างโดดเป็นแกะดำจากค่าอื่นๆ นั่นเอง

• ข้อมูลคนละหน่วยวัดกัน เช่น ข้อมูลทั้งหมดต้องใส่เป็นหน่วยกิโลกรัม แต่บางรายการกลับใส่ข้อมูลเป็นหน่วยกรัม หรือค่าอุณหภูมิลงเป็นหน่วยเซลเซียสบ้าง ฟาเรนไฮต์บ้าง ซึ่งจะส่งผลให้การวิเคราะห์ในขั้นต่อไปผิดพลาดคลาดเคลื่อนไปได้อย่างมาก

• ข้อมูลไม่เป็นตามรูปแบบเดียวกัน ตัวอย่างที่พบได้บ่อยคือ การระบุชื่อจังหวัดกรุงเทพฯ เพราะสามารถมีได้ทั้ง กรุงเทพฯ กรุงเทพ กรุงเทพมหานคร และ กทม.

• ข้อมูลที่ซ้ำซ้อนกัน เช่น พบว่ามีรายการสมาชิกของนาย ก. ซ้ำกัน อาจจะข้อมูลซ้ำเหมือนกันหมด หรือต่างกันเพียงบางข้อมูลอย่างอีเมลหรือเบอร์โทรศัพท์ กรณีแบบนี้ก็จำเป็นต้องคลีนด้วย โดยอาจจะต้องเลือกลบทิ้งข้อมูลที่เก่ากว่า

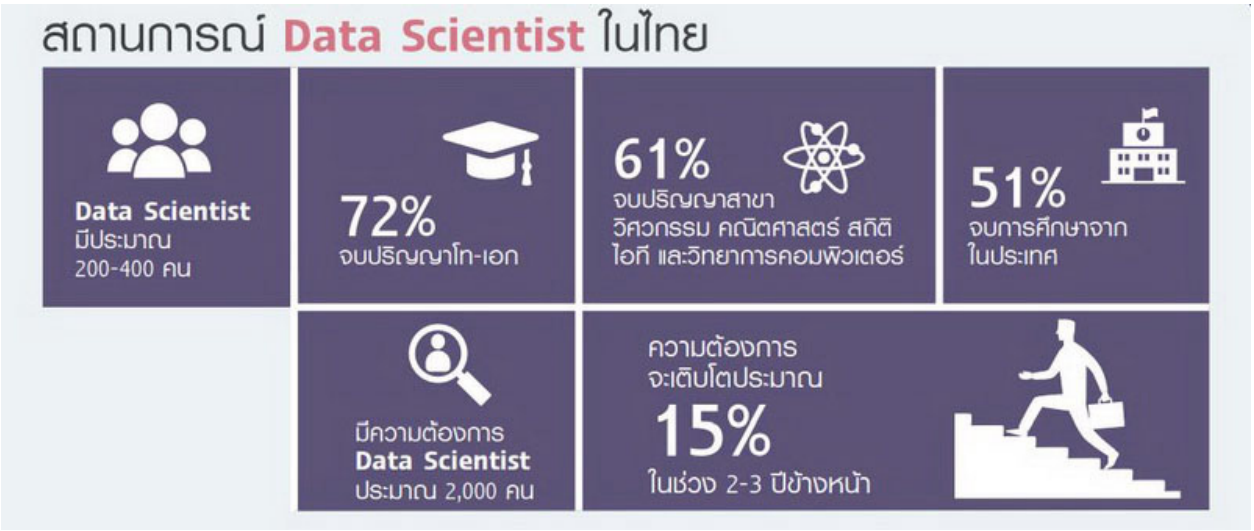


ภาพจาก Forbes

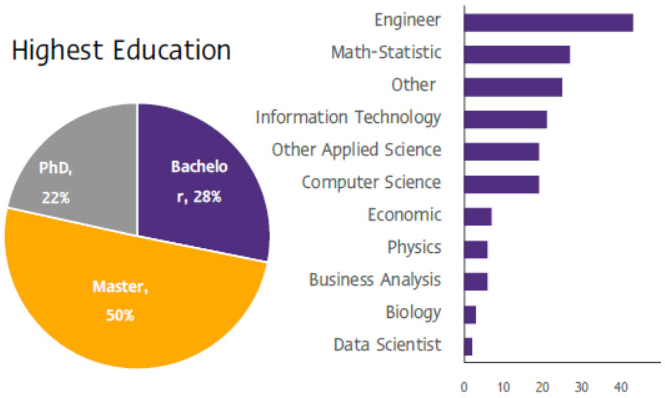
ดังนั้นกระบวนการทำความสะอาดข้อมูลเหล่านี้บนปริมาณข้อมูลระดับ Big Data ด้วยแล้วจึงค่อนข้างใช้เวลามาก และต้องอาศัยทักษะประสบการณ์สูงเพื่อจะหาวิธีตรวจสอบและจัดการแก้ไขข้อผิดพลาดที่มีหลากหลายรูปแบบนั้นให้หมดไปได้อย่างถูกต้อง

สถานการณ์ Data Scientist ในประเทศไทย

ศูนย์วิจัยเศรษฐกิจและธุรกิจ ธนาคารไทยพาณิชย์ (SCB EIC) ประเมินสถานการณ์ Data Scientist ในประเทศไทยไว้ว่า มีความต้องการราว 2,000 อัตรา และคาดว่ามีการเติบโตประมาณ 15% ในระยะเวลา 2-3 ปีข้างหน้า แต่ปัจจุบันมี Data Scientist จริงในประเทศไทยเพียงประมาณ 200-400 คน



จากการสำรวจระดับการศึกษาของ Data Scientist ในไทย ณ ปี ค.ศ. 2017 พบว่าส่วนใหญ่มักจบการศึกษาระดับปริญญาโทถึง 50% และปริญญาเอก 22% โดยจบการศึกษาจากคณะวิศวกรรมศาสตร์ คณิตศาสตร์ เทคโนโลยีสารสนเทศ วิทยาการคอมพิวเตอร์ เป็นส่วนใหญ่



หมายเหตุ : ประวัติการศึกษาของ Data Scientist ในไทยมาจากจำนวนผู้ที่เป็น Data Scientist ในไทยทั้งหมดใน LinkedIn สำรวจ ณ พฤศจิกายน 2017

ทว่าในหลายองค์กรยังคงหลงอยู่กับเพียงเป้าหมายปลายทาง เพื่อจะพูดให้ได้ว่ามี Big Data มี Data Scientist ทำงานอยู่ ทั้งที่ความเป็นจริงแล้ว องค์กรต้องพิจารณาความพร้อมของตัวเองก่อนว่ามาถึงตรงจุดไหนแล้ว มีปริมาณข้อมูล ระบบข้อมูล นโยบายและกฎเกณฑ์ กระบวนการทำงาน เทคโนโลยีต่างๆ ที่สนับสนุนและสอดคล้องซึ่งกันและกันทั้งระบบ พร้อมที่จะเป็นองค์กรขับเคลื่อนด้วยข้อมูล (data driven organization) แล้วหรือยัง? ซึ่งจะโยงสัมพันธ์ไปถึงเรื่องวัฒนธรรมองค์กร ทรัพยากรบุคคล ที่ต้องมีทัศนคติและทักษะถึงพร้อมแล้วอีกด้วย

แต่ถ้าหากองค์กรยังมีระบบข้อมูลต่างๆ แตกแยกกระจัดกระจายกันคนละทิศคนละทาง เก็บข้อมูลกันคนละชนิดคนละรูปแบบไปในแต่ละฝ่าย หรือแม้กระทั่งยังไม่เคยรวบรวมข้อมูลอะไรไว้ให้เป็นที่เป็นทางเลย ก็คงยังมาไม่ถึงมือของ Data Scientist แน่นอน เพราะสิ่งสำคัญที่ต้องทำเป็นอย่างแรกก่อนนั้นต้องจากฝ่าย IT และมี Data Engineer* วางระบบให้เรียบร้อยก่อนไหม? ควบคู่ไปกับการสร้างคนในองค์กรให้พร้อมใช้เทคโนโลยีสารสนเทศและเก็บรวบรวมข้อมูลเป็น จนเป็นวัฒนธรรมพื้นฐานการทำงานในองค์กร

* **Data Engineer** หรือ วิศวกรข้อมูล มีบทบาทหน้าที่ในการสร้างและจัดการระบบข้อมูลทั้งหมด โดยออกแบบสถาปัตยกรรมการไหลของข้อมูล (data flow architectures) ให้เชื่อมโยงกันทั้งระบบ พัฒนาและปรับปรุงโครงสร้างพื้นฐานด้านไอที ด้านฮาร์ดแวร์และซอฟต์แวร์ ที่เกี่ยวข้องต่างๆ เช่น Data security, Data protection เพื่อจัดการให้ระบบข้อมูลพร้อมสำหรับ Data Scientist ใช้วิเคราะห์ในขั้นต่อไป

จะเห็นได้ว่าการทำงานกับข้อมูลขนาดใหญ่ จำเป็นต้องบูรณาการร่วมกันจากบุคลากรอีกหลากหลายสาขาอาชีพ และไม่ได้เป็นเรื่องของฝ่ายใดฝ่ายหนึ่ง หรืออาชีพใดอาชีพหนึ่งแต่เพียงเท่านั้น องค์กรที่จะไปถึงผลสำเร็จทางธุรกิจได้นั้น ต้องอาศัยทักษะและประสบการณ์ของบุคลากรทุกส่วนงานตั้งแต่ต้นทางจนถึงปลายทางในการขับเคลื่อน ดังนั้นการทำงานเป็นทีม เข้าใจบริบทหน้าที่รับผิดชอบของตน และมีเป้าหมายเพื่อความสำเร็จไปในทิศทางเดียวกันนั้น ก็เป็นเรื่องบิ๊กๆ เรื่องหนึ่งขององค์กร ไม่แพ้กันกับการสนใจ Big Data เช่นกัน

เอกสารอ้างอิง

ศิลานนท์, ศิระ. 2560. ข้อมูลขนาดใหญ่ ต้องแค่ไหน..อะไรคือ Big Data. *วารสารวิทยาศาสตร์และเทคโนโลยี,* **32**(3), หน้า 19-24.

ศูนย์วิจัยเศรษฐกิจและธุรกิจ ธนาคารไทยพาณิชย์ (SCB EIC). 2561. In Focus : Data Scientist อาชีพมาแรงแห่งยุค. *ใน: Outlook ไตรมาส 1/2018*. กรุงเทพฯ: ธนาคารไทยพาณิชย์, หน้า 59-74.

Benjamin Aunkofer. 2017. Data Science vs Data Engineering. [online]. Available at: https://data-science-blog.com/blog/2017/10/22/data-science-vs-data-engineering-difference/, [accessed 29 May 2018].

Forbes. 2016. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. [online]. Available at: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#14b846d36f63, [accessed 29 May 2018].

Imarticus.org. 2018. What IT Companies Look at While Hiring a Data Scientist?. [online]. Available at: https://imarticus.org/what-it-companies-look-at-while-hiring-a-data-scientist/, [accessed 29 May 2018].