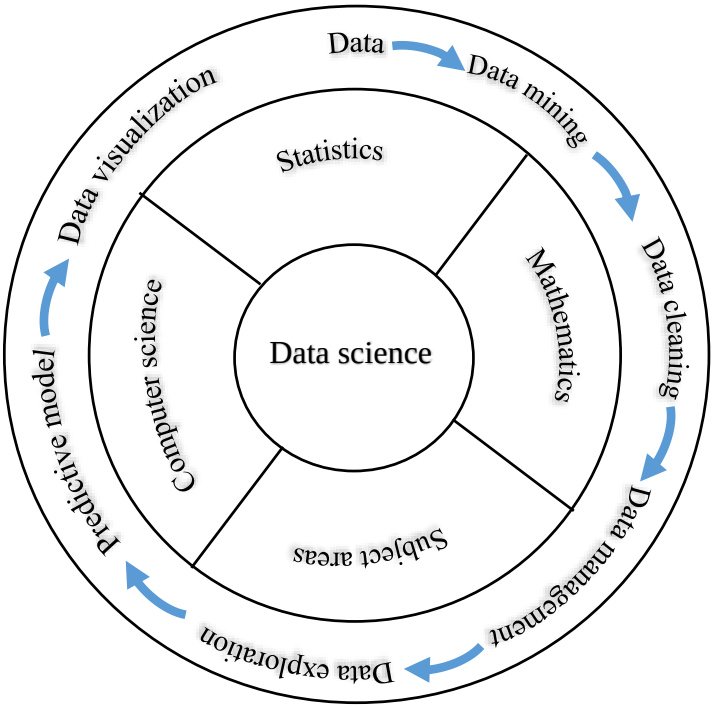
วิทยาศาสตร์ข้อมูล (Data Science)

รศ.ดร.อภิรดี แซ่ลิ่ม

ในปัจจุบันมีการเพิ่มขึ้นของปริมาณข้อมูลอย่างรวดเร็วและเป็นจำนวนมหาศาลในหลากหลายสาขา ประมาณร้อยละ 90 ของข้อมูลในโลกปัจจุบันเกิดขึ้นมาจากในช่วงระยะเวลา 2 ปี ที่ผ่านมา และมีการคาดการณ์กันว่าจะมีการเพิ่มขึ้นถึงร้อยละ 40 ต่อปี วิธีการทางสถิติแบบดั้งเดิมอาจไม่เพียงพอต่อการสกัดองค์ความรู้จากข้อมูลขนาดใหญ่ (Big data) ซึ่งประกอบไปด้วย ปริมาณที่มหาศาล (Volume) ที่มีความรวดเร็ว (Velocity) มีความหลากหลาย (Variety) มีคุณภาพถูกต้อง (Veracity) และ มีคุณค่าต่อการนำไปใช้ (Value) ทั้งนี้การจัดเก็บข้อมูลจากแหล่งต่าง ๆ มีทั้งข้อมูลที่มีโครงสร้าง (Structure data) และข้อมูลที่ไม่มีโครงสร้าง (Unstructured data) แต่การใช้ประโยชน์จากข้อมูลจำนวนมหาศาลดังกล่าวยังมีอยู่จำกัดในการเปลี่ยนข้อมูลให้เป็นสารสนเทศที่ง่ายต่อการเข้าใจ และใช้งานได้ เพื่อนำองค์ความรู้ใหม่ ๆ ที่ได้จากข้อมูลมาใช้ประโยชน์ในทางการบริหาร การจัดการต่าง ๆ การวางนโยบาย การแก้ปัญหาที่ตรงประเด็น และสามารถนำไปสู่การต่อยอดเพื่อสร้างนวัตกรรมใหม่ ๆ ได้

ทั้งนี้ ข้อมูลขนาดใหญ่ ส่วนใหญ่มักมีปัญหาในเรื่องของความกระจัดกระจายของข้อมูล และมีข้อมูลสูญหาย ซึ่งมักจะพบเจอว่าไม่เป็นไปตามข้อตกลงของการใช้วิธีการสถิติดั้งเดิม (Traditional statistics) ในการวิเคราะห์ข้อมูล ด้วยความก้าวหน้าขึ้นอย่างรวดเร็วของเทคโนโลยีในโลกยุคปัจจุบัน การบูรณาการความรู้ด้านวิทยาการคอมพิวเตอร์ สถิติ และคณิตศาสตร์เข้าด้วยกัน เพื่อสร้างวิธีการในการรวบรวม จัดเตรียม จัดการวิเคราะห์ และนำเสนอข้อมูลเพื่อให้ได้ผลลัพธ์ที่ต้องการ โดยบูรณาการตามแต่ละศาสตร์สาขา จึงเป็นกระบวนการของวิทยาการข้อมูล (Data science) ดังแสดงในแผนภาพที่ 1 ปัจจุบันข้อมูลได้เข้ามามีบทบาทสำคัญ ทั้งในด้านสุขภาพ สิ่งแวดล้อม การเงิน การตลาด ธุรกิจ การศึกษา อุตสาหกรรม และสังคม เป็นต้น การพัฒนาทั้งความรู้ กระบวนการ และผลิตภัณฑ์สามารถทำได้แบบก้าวกระโดดโดยอาศัยการตัดสินใจที่ถูกต้องจากผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูล



แผนภาพที่ 1 ความหมายและกระบวนการของวิทยาศาสตร์ข้อมูล

แหล่งข้อมูล

แหล่งข้อมูลมีโครงสร้าง (Structure data)

1. การเปลี่ยนแปลงสภาพภูมิอากาศ แหล่งข้อมูลขนาดใหญ่ คือ ข้อมูลจากดาวเทียมระบบ MODIS ทั้งดาวเทียม Terra และ Aqua ภาพถ่ายจากดาวเทียมระบบ MODIS ครอบคลุมทั่วโลก ข้อมูลมีความละเอียดอยู่ที่ 250x250 ตารางเมตร และ 1x1 ตารางกิโลเมตร เป็นข้อมูลที่แจกจ่ายข้อมูลให้ฟรี โดยผู้ใช้ต้องสมัครลงทะเบียนเพื่อใช้สำหรับการขอข้อมูลออนไลน์ผ่านเว็บไซต์ https://landweb.modaps.eosdis.nasa.gov/cgi-bin/developer/tilemap.cgi โดยสามารถขอข้อมูลย้อนหลังไปได้เป็นระยะเวลา 20 ปี
2. ข้อมูลดัชนีตลาดหลักทรัพย์ สามารถดาวน์โหลดได้จากเว็บไซต์ https://www.set.or.th/th/market/market_statistics.html และ http://siamchart.com/stock/ เป็นข้อมูลขนาดใหญ่ที่สามารถนำมาวิเคราะห์หาแนวโน้ม และการพยากรณ์ ความผันผวนของหุ้นในตลาดหลักทรัพย์ ด้วยการใช้กระบวนการเรียนรู้ด้วยเครื่องที่ใช้กับข้อมูลขนาดใหญ่
3. ข้อมูลคุณภาพน้ำ และคุณภาพอากาศ จากกรมควบคุมมลพิษ มีข้อมูลจากแต่ละสถานีครอบคลุมทั่วประเทศไทย ดาวน์โหลดได้จากเว็บไซต์

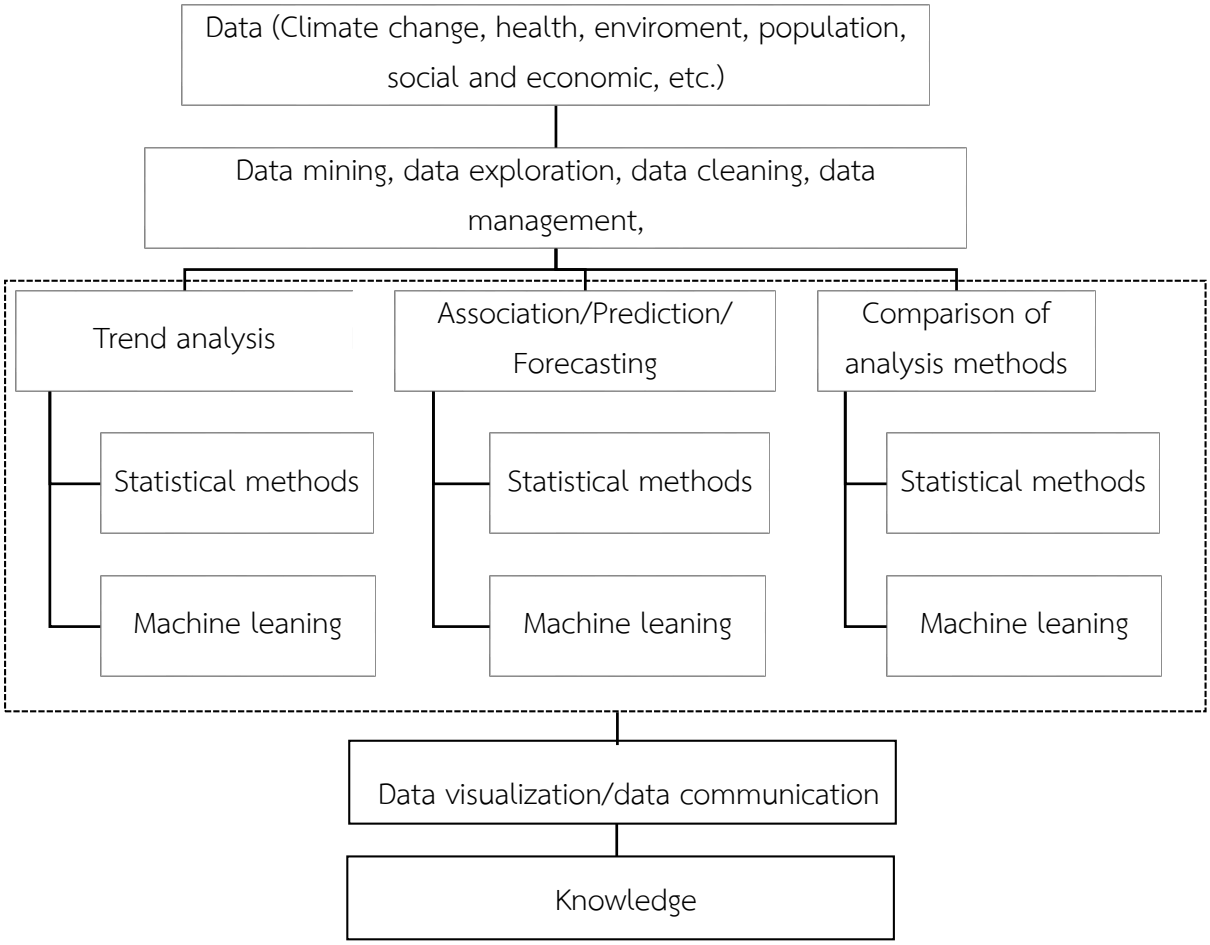
http://air4thai.pcd.go.th/webV2/download.php?grpIndex=3 ซึ่งมีข้อมูลค่าเฉลี่ยคุณภาพน้ำ คุณภาพอากาศรายเดือนให้ดาวน์โหลดมากกว่า 20 ปี

4. ข้อมูลประชากรศาสตร์ เช่น จำนวนประชากร อัตราการตาย อัตราการเกิด ภาวะเจริญพันธุ์ อัตราการป่วยด้วยโรคต่าง ๆ จากการสำมะโนประชากร และการสำรวจต่าง ๆ ดาวน์โหลดได้จากเว็บไซต์ http://statbbi.nso.go.th/staticreport/page/sector/th/index.aspx หรือสามารถขอข้อมูลโดยตรงจากสำนักงานสถิติแห่งชาติ
5. ข้อมูลการรักษาพยาบาลจากโรงพยาบาล และสถานบริการด้านสุขภาพ ประกอบด้วยฐานข้อมูล 43 แฟ้ม ซึ่งเป็นแหล่งข้อมูลด้านสุขภาพที่ใช้ในการเบิกค่ารักษาพยาบาล ที่เป็นข้อมูลขนาดใหญ่ในระดับประเทศ สามารถขอข้อมูลจากสำนักงานหลักประกันสุขภาพแห่งชาติ (สปสช)
6. ข้อมูลด้านอื่น ๆ เช่น การศึกษา แรงงาน การท่องเที่ยว การเกษตร การอุตสาหกรรม พลังงาน การขนส่งและโลจิสติกส์ การประกันภัย ทรัพยากรธรรมชาติ เป็นต้น

แหล่งข้อมูลไม่มีโครงสร้าง (Unstructured data)

ข้อมูลอีกแหล่งที่สำคัญ คือ ข้อมูล ประเภทไม่มีโครงสร้าง เช่น ข้อมูลรูปภาพ ข้อมูลเสียง ข้อมูลที่บันทึกเป็นข้อความต่าง ๆ และ internet of things ประกอบด้วย การซื้อของออนไลน์ การใช้บริการจองห้องพัก รถแท็กซี่ จองตั๋วโดยสาร เป็นต้น ข้อมูลเหล่านี้เป็นข้อมูลที่มีจำนวนมหาศาล แต่ยังขาดการวิเคราะห์ข้อมูล เพื่อสกัดองค์ความรู้ใหม่จาก จากข้อมูลเหล่านี้ โดยข้อมูลมีจำนวนเพิ่มขึ้นอย่างมหาศาล แต่ศักยภาพในการนำมาใช้ประโยชน์ยังมีอยู่จำกัด

กรอบแนวคิดการทำวิจัย



แนวทางการจัดทำข้อเสนอโครงการวิจัย

1. การศึกษาแนวทางการประยุกต์ใช้ทฤษฎีต่าง ๆ สำหรับ การจัดการข้อมูล เช่น เทคนิคการปรับเรียบ (Smoothing) การประมาณค่าสูญหายของข้อมูล (Missing data estimation methods) การทำจัดกลุ่ม (Clustering) เพื่อหาวิธีที่เหมาะสมที่สุด สำหรับการจัดเตรียมข้อมูลสำหรับการวิเคราะห์ในขั้นถัดไป
2. การศึกษาวิธีการวิเคราะห์ข้อมูลที่เหมาะสมสำหรับข้อมูลประเภทต่าง ๆ เช่น การสร้างตัวแบบทางสถิติ (Statistical modeling) การเรียนรู้ด้วยเครื่อง (Machine learning) การเรียนรู้เชิงลึก (Deep learning) เพื่อใช้สำหรับการทำนาย การคาดการณ์ และการพยากรณ์ และอธิบายเหตุการณ์ต่าง ๆ
3. ศึกษาวิธีการการนำผลการวิเคราะห์มานำเสนอด้วยวิธีการที่เหมาะสม (Data visualization and data communication) เพื่อแปลงเป็นสารสนเทศที่เข้าใจได้ง่าย สามารถสื่อสารองค์ความรู้ที่ได้สู่สาธารณะ อันจะนำไปสู่การนำผลการศึกษาที่ได้ ไปใช้ประโยชน์ในการแก้ปัญหา การบริหารจัดการ และการวางนโยบาย