การรวบรวมข้อมูล

ข้อมูลเป็นสิ่งที่มีมูลค่ามหาศาลในปัจจุบัน จึงมีการนำข้อมูลมาวิเคราะห์หรือประมวลผลให้เกิดประโยชน์กับบุคคล หรือองค์กร แต่การได้มาซึ่งข้อมูลที่เป็นประโยชน์นั้น กระบวนการในการเก็บรวบรวมข้อมูล นับว่าเป็นสิ่งสำคัญ ดังประโยคที่ว่า garbage in garbage out ซึ่งได้กล่าวไว้ในขั้นตอนของกระบวนการวิทยาการข้อมูล สำหรับขั้นตอนของการเก็บรวบรวมข้อมูลที่จะกล่าวถึงนั้น เป็นการเก็บรวบรวมข้อมูลทุติยภูมิ โดยต้องกำหนดเป้าหมายให้ชัดเจนว่า จะนำข้อมูลที่รวบรวมได้ไปใช้ในเรื่องใด และจะวิเคราะห์อย่างไร เพื่อให้ได้ผลลัพธ์ตามที่ต้องการ โดยเป้าหมายนั้นสามารถบอกได้ว่าข้อมูลที่ต้องการ รวบรวมได้จากที่ใด และวิธีการใด”

การรวบรวมข้อมูล

ทบทวนเรื่องข้อมูล

ข้อมูลแบ่งตามลักษณะของการได้มา ดังนี้

1. ข้อมูลปฐมภูมิ (primary data) — ข้อมูลที่ได้จากแหล่งกำเนิดข้อมูล หรือจุดเริ่มต้นของข้อมูล เช่น ข้อมูลจากการทดลอง การค้นพบทางวิทยาศาสตร์ เหตุการณ์ประวัติศาสตร์ ภูมิปัญญา ความคิดเห็นของผู้เชี่ยวชาญ โดยข้อมูลปฐมภูมิเป็นข้อมูลที่เกิดขึ้นจากการกระทำ หรือการจดบันทึกของผู้มีส่วนร่วมในเรื่องราวหรือเหตุการณ์เหล่านั้น
2. ข้อมูลทุติยภูมิ (secondary data) — ข้อมูลที่ไม่ได้มาจากแหล่งกำเนิดโดยตรง แต่ได้มาจากการอ้างอิงถึงข้อมูลปฐมภูมิ หรือนำข้อมูลปฐมภูมิมาวิเคราะห์ ประมวลผล ซึ่งอาจอยู่ในรูปสถิติ บทวิจารณ์ บทความ เอกสารต่างๆ

การนำข้อมูลทุติยภูมิที่มีการจัดเก็บรวบรวมไว้แล้วใช้งาน อาจมีค่าใช้จ่ายและใช้เวลาน้อยกว่าการใช้ข้อมูลปฐมภูมิ อย่างไรก็ตาม ข้อมูลทุติยภูมิที่มีการอ้างอิงหรือส่งต่อกันมาเป็นทอดๆ อาจมีความจริงบางส่วนถูกบิดเบือนไปทั้งโดยเจตนา หรือไม่เจตนา ดังนั้น ในการอ้างอิงข้อมูลทุติยภูมิ ต้องตรวจสอบความถูกต้องและแหล่งที่มาอย่างละเอียดถี่ถ้วน เพื่อให้เกิดความเชื่อมั่นในการนำข้อมูลไปใช้ เพื่อให้เกิดประโยชน์อย่างแท้จริง

การรวบรวมข้อมูล

วิธีการรวบรวมข้อมูล

- การสัมภาษณ์ (interview) — สัมภาษณ์โดยตรงหรือผ่านการสื่อสารอื่น เช่น โทรศัพท์ สื่อสังคมออนไลน์ ต้องใช้คำถามที่ชัดเจน ตรงประเด็น เป็นลักษณะคำถามปลายเปิด นิยมใช้รวบรวมข้อมูลเชิงคุณภาพ เช่น ความเห็นของนักเรียนต่อระเบียบปฏิบัติในห้องเรียน ความรู้สึกของผู้บริโภคเกี่ยวกับผลิตภัณฑ์ใหม่
- การสำรวจ (survey) — ใช้แบบสำรวจที่มีการกำหนดคำถาม เพื่อค้นหาข้อมูล หรือความเห็นที่ต้องการ เช่น ความพึงพอใจของการบริหารงานของสภานักเรียน แหล่งท่องเที่ยวที่นักท่องเที่ยวสนใจ
- การสังเกต (observe) — รวบรวมข้อมูลจากเหตุการณ์ สถานการณ์ หรือพฤติกรรมที่เปลี่ยนแปลงไป เช่น สังเกตพฤติกรรมของนักเรียนระหว่างรับประทานอาหาร พฤติกรรมการทิ้งขยะของคนในองค์กร
- การทดลอง (experiment) — รวบรวมข้อมูลจากการทดลองหรือทดสอบที่มีการควบคุมปัจจัยบางประการ เช่น การบันทึกผลการเจริญเติบโตของถั่วงอกเมื่อมีแสงแดดและไม่มีแสงแดด
- การทบทวนเอกสาร (document/literature review) — เป็นการรวบรวมข้อมูลจากเอกสาร รายงาน บทความ หรือแบบฟอร์มการรวมรวบข้อมูล เช่น แบบบันทึกการเข้าเรียนของนักเรียน รายงานประจำปี รายงานการประชุม จดหมายข่าว แบบฟอร์มลงเวลาปฏิบัติงาน
- การสำมะโน (census) — รวบรวมข้อมูลด้วยการสำรวจจากประชากรเกี่ยวกับเรื่องที่กำหนด เช่น สำนักงานสถิติแห่งชาติมีการสำมะโนประชากรและเคหะเป็นประจำทุกๆ 10 ปี

การรวบรวมข้อมูล

การเก็บรวบรวมข้อมูล (Data Collection)

ในปัจจุบัน แหล่งข้อมูลทุติยภูมิมีการเผยแพร่บนอินเทอร์เน็ตและอยู่ในหลายรูปแบบ (format) ในการนำไปใช้งานอาจมีวิธีจัดการข้อมูลที่แตกต่างกันขึ้นกับรูปแบบที่เผยแพร่ดังนี้

- ไฟล์ — ไฟล์ข้อมูล เช่น ไฟล์ที่ได้จากโปรแกรมตาราทำงาน (นามสกุล .xls, .xlsx, .odp) หรือไฟล์แบบข้อความ (text) (นามสกุล .csv) สามารถดาวน์โหลดไปใช้งานได้โดยไม่ต้องอาศัยขั้นต้อนซับซ้อนในการแปลงข้อมูล ส่วนไฟล์นามสกุล .pdf สามารถดาวน์โหลดได้แต่มีกระบวนการซับซ้อนในการแปลงข้อมูลให้อยู่ในรูปแบบที่นำไปใช้คำนวณ นอกจากนี้ ยังมีข้อมูลที่อยู่ในรูปแบบที่ต้องเขียนคำสั่งในการนำข้อมูลเหล่านั้นมาใช้งาน เช่น ข้อมูลจาก Facebook, Twitter ต้องเขียนคำสั่งผ่านวิธีการเชื่อมต่อเฉพาะ (API: Application Programming Interface)
- รายงานหรือตารางบนเว็บไซต์ — เป็นข้อมูลที่ผ่านการสรุปมาแล้ว ไม่มีข้อมูลดิบประกอบ ทำให้ยากในการนำข้อมูลไปวิเคราะห์ในประเด็นอื่น เช่น ข้อมูลสรุปจำนวนผู้ติดเชื้อและเสียชีวิตในช่วงการแพร่ระบาดของโรคโควิด-19 ซึ่งไม่มีรายละเอียดของแต่ละบุคคล แต่ละภูมิภาค ทำให้ไม่สามารถวิเคราะห์ถึงช่วงอายุ หรือภูมิภาคของผู้ติดเชื้อหรือเสียชีวิต

การรวบรวมข้อมูล

แหล่งข้อมูลทุติยภูมิ

data.go.th เป็นแหล่งข้อมูลทุติยภูมิสถิติจากศูนย์กลางข้อมูลภาครัฐ เพื่อประโยชน์ต่อสาธารณชนและหน่วยงานทั้งภาครัฐและเอกชน สามารถค้นหาและเข้าถึงข้อมูลที่มีคุณภาพของภาครัฐได้โดยสะดวก ซึ่งมีให้ดาวน์โหลดไฟล์ในรูปแบบ .xls และรูปแบบ .csv นอกจากนี้ยังสามารถดาวน์โหลดไฟล์คำอธิบายข้อมูล (metadata) ได้

- ตัวอย่างข้อมูลรายได้เฉลี่ยต่อเดือนต่อครัวเรือน จาก www.data.go.th นี้ จำแนกตามภาค และจังหวัด ซึ่งข้อมูลดังกล่าว มีคุณลักษณะหรือแอตทริบิวต์ (attribute) ได้แก่ รายได้เฉลี่ยต่อเดือนต่อครัวเรือน ซึ่งได้ทำการเก็บรวบรวมเป็นรายปี ตั้งแต่ปี พ.ศ.2541 ถึง พ.ศ.2558 (18 ปี) สามารถนำมาประมวลผลเพื่อแบ่งกลุ่มจังหวัดที่มีรายได้เฉลี่ยมาก ปานกลาง หรือน้อย เพื่ออธิบายภาพรวมรายได้เฉลี่ยประชากรของประเทศ ทำให้สามารถวางนโยบายที่เหมาะสมในการบริหารงาน หรือพัฒนาจังหวัดต่างๆ

การรวบรวมข้อมูล

แหล่งข้อมูลทุติยภูมิ

แหล่งข้อมูลทุติยภูมิที่เผยแพร่ของประเทศไทย

สำนักงานสถิติแห่งชาติ — ข้อมูลสถิติประชากร แรงงาน การศึกษา ศาสนา ศิลปวัฒนธรรม สุขภาพ

สำนักงานพัฒนารัฐบาลดิจิทัล (องค์การมหาชน) — ข้อมูลที่รวบรวมจากแหล่งต่างๆ จัดเป็นหมวดหมู่

สำนักงานคณะกรรมการพัฒนาการเศรษฐกิจและสังคมแห่งชาติ — ข้อมูลด้านเศรษฐกิจและสังคม ทรัพยากรธรรมชาติและสิ่งแวดล้อม

แหล่งข้อมูลทุติยภูมิที่เผยแพร่ของต่างประเทศ

Kaggle

Data World

UCI Machine Learning Repository

Google Dataset Search

การรวบรวมข้อมูล

ความเหมาะสมของแหล่งข้อมูล

การเลือกใช้แหล่งข้อมูลที่มีการบิดเบือน ขาดความน่าเชื่อถือ อาจทำให้ข้อสรุปที่ได้เกิดความผิดพลาดหรือชี้นำไปในทางที่ผิด นอกจากนี้อาจเกิดอันตรายและสร้างความเสียหาย ดังนั้นก่อนเลือกใช้แหล่งข้อมูล ควรพิจารณาความเหมาะสมของแหล่งข้อมูลตามมุมมองดังนี้

- จุดมุ่งหมายของแหล่งข้อมูล (purpose) — ข้อมูลถูกพัฒนาขึ้นเพื่อเป้าหมายใด
- ความทันสมัยของข้อมูล (currency) — ข้อมูลเผยแพร่เมื่อใด
- ความสอดคล้องกับการใช้งาน (relevance) — ข้อมูลเกี่ยวข้องกับปัญหาที่ต้องการหรือไม่
- ความน่าเชื่อถือของแหล่งข้อมูล (authority) — แหล่งข้อมูลหรือผู้เผยแพร่น่าเชื่อถือหรือไม่
- ความถูกต้องแม่นยำ (accuracy) — ข้อมูลมีการยืนยันความถูกต้อง มีการถูกอ้างอิงถึงหรือไม่

การเตรียมข้อมูล (Data Preparation)

การเตรียมข้อมูล (Data Preparation)

หลังจากเลือกแหล่งข้อมูลและรวบรวมข้อมูลเรียบร้อยแล้ว ขั้นตอนต่อไปคือการเตรียมข้อมูล เพื่อทำให้ข้อมูลมีความถูกต้อง ครบถ้วน สมบูรณ์ ไม่มีค่าผิดปกติ เพื่อเตรียมพร้อมสำหรับการประมวลผลข้อมูล

การทำความสะอาดข้อมูล (Data Cleansing)

ข้อมูลที่รวบรวมมานั้น อาจมีข้อผิดพลาดซึ่งไม่เหมาะต่อการนำไปประมวลผล ได้แก่

- มีค่าว่าง
- มีค่าที่อยู่นอกขอบเขตจากค่าที่เป็นไปได้
- ใช้หน่วยนับผิด
- เป็นค่าผิดปกติ (outlier)
- ใช้รูปแบบข้อมูลแตกต่างกัน
- พิมพ์ผิด

ซึ่งสาเหตุเกิดจากผู้ให้ข้อมูลกรอกข้อมูลไม่ครบถ้วน ผู้บันทึกข้อมูลพิมพ์ข้อมูลผิดพลาด หรือการขาดข้อกำหนดในการบันทึกข้อมูล

การแก้ไขข้อมูลเมื่อพบว่ามีข้อผิดพลาด สามารถทำได้โดยการแก้ไขให้ถูกต้อง หรือลบข้อมูลที่ไม่ส่งผลกระทบต่อการประมวลผล หากข้อมูลมีจำนวนไม่มาก สามารถใช้คนดำเนินการตรวจสอบและแก้ไขข้อมูล แต่หากข้อมูลมีจำนวนมาก ต้องอาศัยโปรแกรมคอมพิวเตอร์ในการดำเนินการจัดเตรียมข้อมูลให้สอดคล้องกับเงื่อนไข และรูปแบบข้อมูลที่กำหนดในโปรแกรม

การเตรียมข้อมูล (Data Preparation)

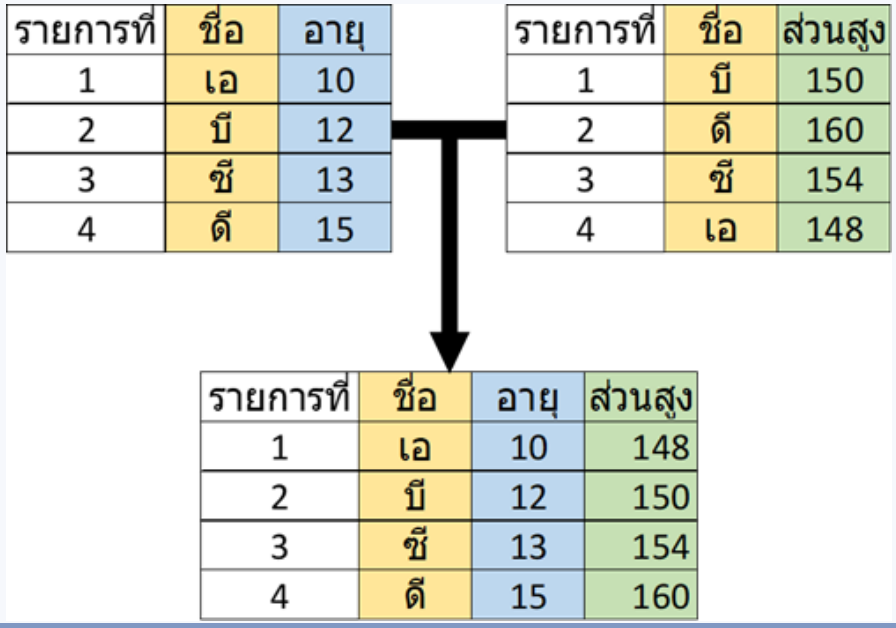
การแปลงข้อมูล (Data Transformation)

เป็นการเตรียมข้อมูลให้อยู่ในรูปแบบที่พร้อมสำหรับการประมวลผล โดยรูปแบบของข้อมูลที่พร้อมประมวลผลในโปรแกรมตารางทำงานนั้น แต่ละแถว (บรรทัด) คือข้อมูล 1 รายการ และแต่ละคอลัมน์ (หลัก) คือ คุณลักษณะ หรือแอตทริบิวต์

รายการที่	ชื่อ	อายุ
1	เอ	10
2	บี	12
3	ซี	13
4	ดี	15

การเชื่อมโยงข้อมูล (Data Combining)

กรณีที่ต้องการใช้ข้อมูลของกลุ่มตัวอย่างที่มีการเผยแพร่จากหลายแหล่ง หรือมีหลายไฟล์ข้อมูล ต้องทำการเชื่อมโยงข้อมูลจากหลายแหล่งเข้าด้วยกัน โดยใช้คุณลักษณะหรือแอตทริบิวต์ ที่มีอยู่รวมกันของหลายแหล่งข้อมูล เป็นตัวเชื่อมโยง



รายการที่	ชื่อ	อายุ
1	เอ	10
2	บี	12
3	ซี	13
4	ดี	15

รายการที่	ชื่อ	ส่วนสูง
1	บี	150
2	ดี	160
3	ซี	154
4	เอ	148

รายการที่	ชื่อ	อายุ	ส่วนสูง
1	เอ	10	148
2	บี	12	150
3	ซี	13	154
4	ดี	15	160

การเตรียมข้อมูล (Data Preparation)

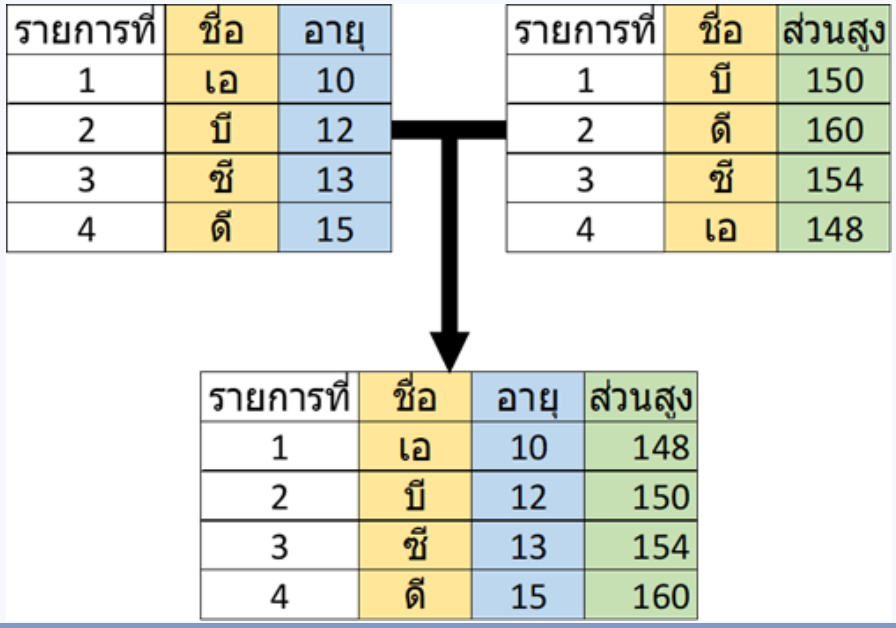
การแปลงข้อมูล (Data Transformation)

เป็นการเตรียมข้อมูลให้อยู่ในรูปแบบที่พร้อมสำหรับการประมวลผล โดยรูปแบบของข้อมูลที่พร้อมประมวลผลในโปรแกรมตารางทำงานนั้น แต่ละแถว (บรรทัด) คือข้อมูล 1 รายการ และแต่ละคอลัมน์ (หลัก) คือ คุณลักษณะ หรือแอตทริบิวต์

รายการที่	ชื่อ	อายุ
1	เอ	10
2	บี	12
3	ซี	13
4	ดี	15

การเชื่อมโยงข้อมูล (Data Combining)

กรณีที่ต้องการใช้ข้อมูลของกลุ่มตัวอย่างที่มีการเผยแพร่จากหลายแหล่ง หรือมีหลายไฟล์ข้อมูล ต้องทำการเชื่อมโยงข้อมูลจากหลายแหล่งเข้าด้วยกัน โดยใช้คุณลักษณะหรือแอตทริบิวต์ ที่มีอยู่รวมกันของหลายแหล่งข้อมูล เป็นตัวเชื่อมโยง



รายการที่	ชื่อ	อายุ
1	เอ	10
2	บี	12
3	ซี	13
4	ดี	15

รายการที่	ชื่อ	ส่วนสูง
1	บี	150
2	ดี	160
3	ซี	154
4	เอ	148

รายการที่	ชื่อ	อายุ	ส่วนสูง
1	เอ	10	148
2	บี	12	150
3	ซี	13	154
4	ดี	15	160

ที่มา

- หนังสือเรียน รายวิชาพื้นฐานวิทยาศาสตร์ เทคโนโลยี (วิทยาการคำนวณ) ชั้นมัธยมศึกษาปีที่ 4 ของสถาบันส่งเสริมการสอนวิทยาศาสตร์และเทคโนโลยี
- สถาบันส่งเสริมการสอนวิทยาศาสตร์และเทคโนโลยี, "เทคโนโลยี(วิทยาการคำนวณ)", โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, ศูนย์หนังสือแห่งจุฬาลงกรณ์มหาวิทยาลัย, 2561 หน้า 36 44
- ครูไอที – ฟรี บทเรียนออนไลน์ที่กระชับ และเข้าใจง่าย