วารสารธรรมศาสตร์ ปีที่ 39 ฉบับที่ 2 พ.ศ.2563 (167-185)

Big Data: ภูมิทัศน์ใหม่ในการศึกษาวิจัยด้านสังคมศาสตร์ในประเทศไทย

อัครนัย ขวัญอยู่ *

คณะสังคมวิทยาและมานุษยวิทยา, มหาวิทยาลัยธรรมศาสตร์

วันที่รับบทความ 30 มีนาคม พ.ศ.2563
วันที่แก้ไขบทความ 25 มิถุนายน พ.ศ.2563
วันที่ตอบรับบทความ 1 กรกฎาคม พ.ศ.2563

บทคัดย่อ

Big Data กำลังถูกกล่าวถึงอย่างมากในฐานะที่เป็นนวัตกรรมเปลี่ยนโลก และกำลังก้าวเข้ามามีบทบาทอย่างสำคัญในการศึกษาวิจัยด้านสังคมศาสตร์ และมนุษยศาสตร์ในประเทศไทย องค์ประกอบที่สำคัญ 6 ประการของ Big Data ได้แก่ 1) ความหลากหลายของข้อมูล (Variety) 2) ขนาดของข้อมูล (volume) 3) ความเป็นปัจจุบันทันด่วน (velocity) 4) ความปรับเปลี่ยนได้ของข้อมูล (Variability) 5) ความจริงแท้แน่นอน (Veracity) และ 6) ความมีคุณค่า (Value) มีส่วนเกื้อหนุนให้การศึกษาทางสังคมศาสตร์แบบภววิสัย (Objectivity) หรือการศึกษาสังคมโดยอาศัยกระบวนการทางวิทยาศาสตร์ใกล้ถึงจุดที่เป็นอุดมคติมากขึ้น โดยในปัจจุบันมีความพยายามของนักวิจัยด้านสังคมศาสตร์ในประเทศไทยที่ใช้ฐานข้อมูล Big Data ในการค้นหาความรู้และความจริงเกี่ยวกับปรากฏการณ์ต่าง ๆ ซึ่งในบทความนี้จะทำหน้าที่ในการสำรวจภูมิทัศน์ใหม่ของการศึกษาสังคมศาสตร์ไทยในยุค Big Data และแสดงให้เห็นถึงข้อได้เปรียบ ในการใช้งานฐานข้อมูล ตลอดจนการเตรียมความพร้อมของนักสังคมศาสตร์ในการใช้งาน Big Data

คำสำคัญ: ข้อมูลขนาดใหญ่, การวิจัยทางสังคมศาสตร์, ภววิสัย

*ผู้รับผิดชอบบทความ: akkaranai@gmail.com DOI: 10.14456/tujournal.2020.18

Thammasat Journal, Volume 39. No 2, 2020 (167-185)

Big Data: The New Landscape in Social Science Research in Thailand

Akkaranai Kwanyou*

Faculty of Sociology and Anthropology, Thammasat University

Received 30 March 2020

Received in revised 25 June 2020

Accepted 1 July 2020

Abstract

Big Data are being greatly discussed as the innovation that changes the world and it is starting to have an important role in the social science research and humanities in Thailand. There are 6 important parts of Big Data including 1) Variety 2) Volume 3) velocity 4) Variability of data 5) Veracity and 6) Value. These things support the objective study of social science or social study by using scientific process which gets near the ideal point more. Nowadays, there has attempt of social science researchers in Thailand who implement Big Data in searching for knowledge and truth regarding various phenomenon. This paper would serve to survey the new area of social study in Big Data era and shows the advantage in using database and preparation of social scientists in implementing Big Data.

Keyword: Big Data, Social Research, Objectivism

***Corresponding author: akkaranai@gmail.com** **DOI: 10.14456/tujournal.2020.18**

บทนำ

กว่าศตวรรษที่ผ่านมา Big Data หรือ “ข้อมูลเกินคณานับ” ได้เข้ามามีบทบาทสำคัญในการเปลี่ยนแปลงโลกแห่งวิทยาการข้อมูล และกลายเป็นกำลังหลักในการปฏิวัติอุตสาหกรรมครั้งที่ 4 (The Fourth Industrial Revolution, 4IR) หรือที่เรียกกันว่าโลกยุค 4.0 ที่ทุกอย่างขับเคลื่อนไปด้วยองคาพยพของข้อมูลสารสนเทศ นับตั้งแต่การผลิตในภาคอุตสาหกรรม การวางแผนหรือการตัดสินใจทางธุรกิจ การออกแบบนโยบายหรือการบริหารจัดการภาครัฐ หรือแม้แต่การศึกษาวิจัยในโลกทางวิชาการ ไม่เพียงเท่านั้น Big Data ยังแทรกซึมเข้ามาเกี่ยวข้องกับโลกในชีวิตประจำวันของผู้คนผ่านอุปกรณ์อิเล็กทรอนิกส์ (Smart Device) ต่างๆ ที่อยู่รอบตัว ทั้งโทรศัพท์มือถือ เครื่องคอมพิวเตอร์ หรืออุปกรณ์อิเล็กทรอนิกส์ขนาดพกพาอย่างสายรัดข้อมืออัจฉริยะ (Smart watch) ฯลฯ สิ่งเหล่านี้ได้ทำหน้าที่เป็นตัวกลางในการล้วงเอาข้อมูลพฤติกรรม ลักษณะนิสัย ความชื่นชอบ ฯลฯ ของผู้คนไปใช้ประโยชน์ทางด้านการตลาดและการศึกษาวิจัยเกี่ยวกับพฤติกรรมผู้บริโภค ในทุกครั้งที่เรา “ค้นหาคำ” ในโปรแกรมค้นหา (Search Engines) หรือโพสต์ข้อความและรูปภาพลงในสื่อสังคมออนไลน์ ข้อมูลดังกล่าวถูกนำไปจัดเก็บอย่างเป็นระบบในฐานข้อมูลขนาดใหญ่ เพื่อรอไว้ใช้ประโยชน์ในทางธุรกิจ ด้วยเหตุนี้ จึงไม่น่าแปลกใจที่ผู้ให้บริการทางเทคโนโลยีอย่าง Facebook Google หรือ Line Application ฯลฯ จะมีความเข้าอกเข้าใจผู้ใช้งาน และสามารถนำเสนอเนื้อหา โฆษณา หรือออกแบบบริการที่ตอบโจทย์ความต้องการของผู้ใช้ได้อย่างมีประสิทธิภาพ

ในขณะที่ Big Data พยายามแทรกซึมเข้าไปในทุกอณูของชีวิตมนุษย์ และกำลังคืบคลานเข้าสู่วงการธุรกิจ และอุตสาหกรรมประเภทต่างๆ อย่างเต็มรูปแบบนั้น วงการด้านการศึกษาวิจัยโดยเฉพาะในทางสังคมศาสตร์ก็เริ่มมีความตื่นตัว และให้ความสนใจกับการนำข้อมูลลักษณะดังกล่าวมาใช้ประโยชน์ในทางวิชาการมากขึ้นทุกขณะ โดยจากผลการศึกษาในบทความเรื่อง Theme Mapping and Bibliometrics Analysis of One Decade of Big Data Research in the Scopus Database ของ Parlina, Ramli , & Murfi (2020) ได้แสดงให้เห็นว่าในวงการวิชาการในระดับโลกได้มีการนำ Big Data มาใช้เป็นข้อมูลหลักในการศึกษาวิจัยมากขึ้น จากในปี 2009 ที่มีงานศึกษาวิจัยที่ใช้ Big Data เป็นฐานข้อมูลหลักในการวิเคราะห์วิจัย ตีพิมพ์ในฐานข้อมูล Scopus เพียง 1 รายการ กลับเพิ่มขึ้นเป็น 2,236 รายการ ในปี 2018 และมีมากถึงร้อยละ 30 ของงานทั้งหมดที่เป็นงานศึกษาในทางสังคมศาสตร์ มนุษย์ศาสตร์ และบริหารธุรกิจ ซึ่งจากสถิติดังที่ได้กล่าวมานี้ แสดงให้เห็นว่า Big Data เริ่มถูกนำมาใช้ในการศึกษาวิจัยในศาสตร์ทางสังคมเป็นจำนวนไม่น้อย มิได้จำกัดอยู่เพียงแค่วงการวิทยาศาสตร์และเทคโนโลยีเพียงอย่างเดียวเท่านั้น

อย่างไรก็ตาม การก้าวเข้ามาของ Big Data ในอาณาบริเวณของการศึกษาทางสังคมศาสตร์ ได้ส่งผลกระทบในทางบวกอย่างมีนัยยะอย่างสำคัญต่อการศึกษาสังคมแบบ ภววิสัย

(Objectivity) หรืออาจกล่าวได้ว่า Big Data เข้ามาเติมเต็มจินตนาการในการศึกษาสังคมโดยอาศัยระเบียบวิธีวิจัยแบบวิทยาศาสตร์ให้ใกล้จุดสูงสุดของความสมบูรณ์แบบ และหลุดออกจากกรอบหรือข้อจำกัดของการวิจัยเชิงปริมาณดั้งเดิมที่เคยมีมาในอดีต โดยในบทความนี้ ผู้เขียนจะทำการสำรวจภูมิทัศน์ใหม่ของการศึกษาสังคมศาสตร์ไทยในยุค Big Data และสร้างการรับรู้ให้แก่นักวิชาการทางสังคมศาสตร์ ถึงพลังของข้อมูลที่สามารถสร้างความเปลี่ยนแปลงให้กับระเบียบวิธีวิจัยเชิงปริมาณที่ใช้กันอยู่ในปัจจุบัน โดยเริ่มจากการฉายภาพให้เห็นคุณสมบัติอันโดดเด่น และข้อได้เปรียบในการใช้งานฐานข้อมูล Big Data (ในหัวข้อที่ 2 และ 3) ตลอดจนยกตัวอย่างความพยายามในการใช้ Big Data ในการศึกษาวิจัยทางสังคมศาสตร์ในประเทศไทย (ในหัวข้อที่ 4) ข้อจำกัด และการเตรียมความพร้อมในการใช้งาน Big Data (ในหัวข้อที่ 5)

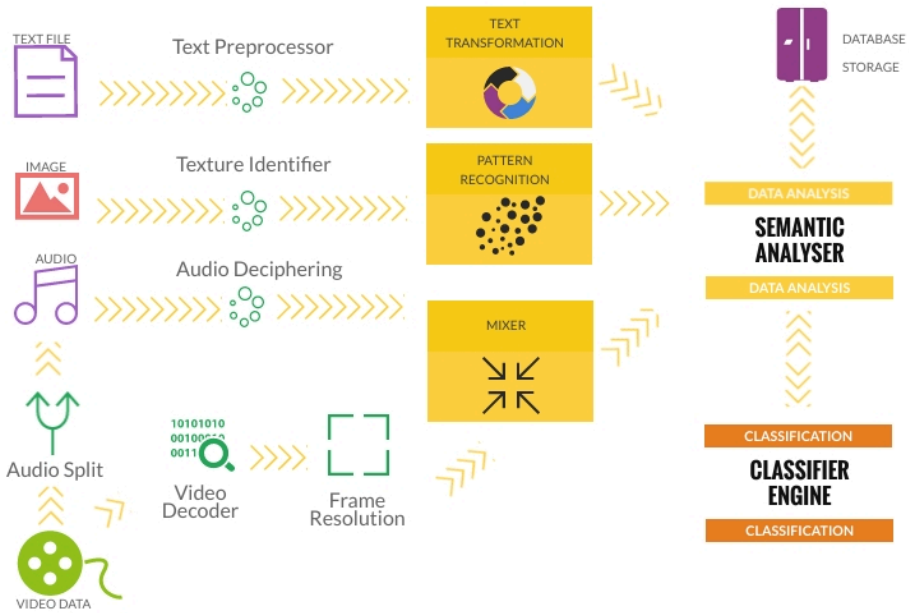
องค์ประกอบของ Big Data ในการศึกษาทางสังคมศาสตร์

หากแปลความหมายของคำว่า Big Data ตรงตัว อาจทำให้เกิดความเข้าใจที่คลาดเคลื่อนได้ว่า Big Data หมายถึง ลักษณะของข้อมูลที่มี “ขนาดใหญ่” เพียงอย่างเดียว และอาจชวนให้นึกไปว่าฐานข้อมูลบางฐานข้อมูลบางฐานที่มีการเก็บข้อมูลเป็นจำนวนมาก เช่น ฐานข้อมูลสำมะโนประชากร หรือฐานข้อมูลระดับเศรษฐสังคมของครัวเรือนไทย (Socio-economic Survey) ที่จัดเก็บโดยสำนักงานสถิติแห่งชาติ เป็นหนึ่งในฐานข้อมูล Big Data ซึ่งในความเป็นจริงแล้วองค์ประกอบของ Big Data ไม่ได้มีเพียงเรื่องของขนาดเท่านั้น แต่ยังหมายรวมถึงกระบวนการการได้มาของข้อมูล คุณสมบัติ และการใช้ประโยชน์จากข้อมูลอีกด้วย ทั้งนี้ ตามนิยามของ Moura และ Serrao (2015) อธิบายว่าข้อมูลที่มีลักษณะเป็น Big Data ต้องมีองค์ประกอบหลัก 6V ซึ่งสามารถนำมาปรับใช้กับข้อมูลทางสังคมศาสตร์ ได้ดังต่อไปนี้

1. ความหลากหลายของข้อมูล (Variety) ในที่นี้ไม่ได้หมายถึงเฉพาะความหลากหลายในแง่ของประเด็น หรือเนื้อหาของข้อมูลที่จัดเก็บเพียงอย่างเดียวเท่านั้น แต่ยังหมายรวมถึงประเภทของข้อมูล (Data Types) ที่จัดเก็บด้วย โดยปกติการศึกษาวิจัยในทางสังคมศาสตร์ด้วยระเบียบวิธีวิจัยเชิงปริมาณ ผู้วิจัยส่วนใหญ่จะคุ้นชินกับข้อมูลที่มีลักษณะเป็นข้อมูลเชิงโครงสร้าง (Structured data) หรือข้อมูลที่มีรูปแบบการจัดเก็บที่ตายตัวตามโครงสร้างของแบบสอบถาม หรือแบบสัมภาษณ์ ซึ่งผู้วิจัยสามารถจัดจำแนกออกมาเป็นตัวแปรต่างๆ ที่ใช้ในการศึกษา หรือจำแนกตามกรอบแนวคิดการวิจัยได้อย่างชัดเจน แต่สำหรับข้อมูล Big Data อาจไม่ได้มีเฉพาะข้อมูลเชิงโครงสร้างเพียงอย่างเดียว ในหนึ่งฐานข้อมูลอาจมีข้อมูลในลักษณะกึ่งโครงสร้าง หรือไร้โครงสร้าง (Unstructured data) เช่น ข้อมูลที่เป็นรูปภาพ ไฟล์เสียง ไฟล์วิดีโอ หรือข้อมูลในรูปของประโยคและข้อความ ฯลฯ อยู่รวมในฐานข้อมูลด้วย ดังนั้น นักวิจัยจึงจำเป็นต้องอาศัยองค์ความรู้ และเครื่องมือในการช่วยสังเคราะห์ หรือแปลงข้อมูลไร้โครงสร้างให้อยู่ในรูปแบบที่มีโครงสร้างมากขึ้น เพื่อง่ายต่อการวิเคราะห์วิจัยต่อไป ในภาพประกอบ 1 แสดงรูปแบบของการสังเคราะห์ข้อมูล

ไร้โครงสร้างที่มีอยู่หลากหลายรูปแบบ ซึ่งแต่ละรูปแบบมีเทคนิควิธีในการสังเคราะห์แตกต่างกัน เช่น ข้อมูลที่อยู่ในรูปของประโยค หรือข้อความ (Text File) จะต้องใช้การเทคนิคประมวลผลภาษาธรรมชาติหรือภาษามนุษย์ (Natural Language Processing) ในการตีความข้อความหรือรูปประโยคเพื่อดึงเอาข้อมูลที่จำเป็น หรือมีนัยยะต่อการศึกษาออกมา หรือในกรณีของข้อมูลที่อยู่ในลักษณะของรูปภาพ (Image File) จำเป็นต้องใช้การวิเคราะห์ด้วยโครงข่ายประสาทแบบคอนโวลูชัน Convolutional Neural Network (CNN) ซึ่งเป็นเทคนิคที่ช่วยในการวิเคราะห์ และจัดจำแนกประเภทของรูปภาพ ว่ามีความเกี่ยวข้องกับประเด็นใด หรือภาพดังกล่าวสื่อถึงเรื่องอะไร เป็นต้น

หากพิจารณาในแง่หนึ่งแล้วเทคนิคการสังเคราะห์ข้อมูลแบบไร้โครงสร้าง นับเป็นจินตนาการใหม่ของการวิจัยทางสังคมศาสตร์เช่นกัน เนื่องด้วย ข้อมูลที่ใช้ในการศึกษาทางสังคมจำนวนไม่น้อยอยู่ในรูปของภาพถ่าย วิดีโอบันทึกเหตุการณ์ หรือข้อมูลที่ได้จากการบันทึกภาคสนาม ฯลฯ ซึ่งจำเป็นต้องอาศัยการวิเคราะห์และตีความอย่างเป็นระบบ โดยในปัจจุบันความก้าวหน้าทางเทคโนโลยีทำให้เทคนิคการวิเคราะห์เหล่านี้สามารถกระทำได้โดยง่าย แม้ว่าผู้วิจัยจะไม่ได้มีฐานความรู้ในทางโปรแกรมมิ่งมาก่อนก็ตาม



ภาพประกอบ 1 แสดงรูปแบบการสังเคราะห์ข้อมูลไร้โครงสร้างเพื่อใช้ประโยชน์ในการศึกษาวิจัย

ที่มา: Pegasus One (2019)

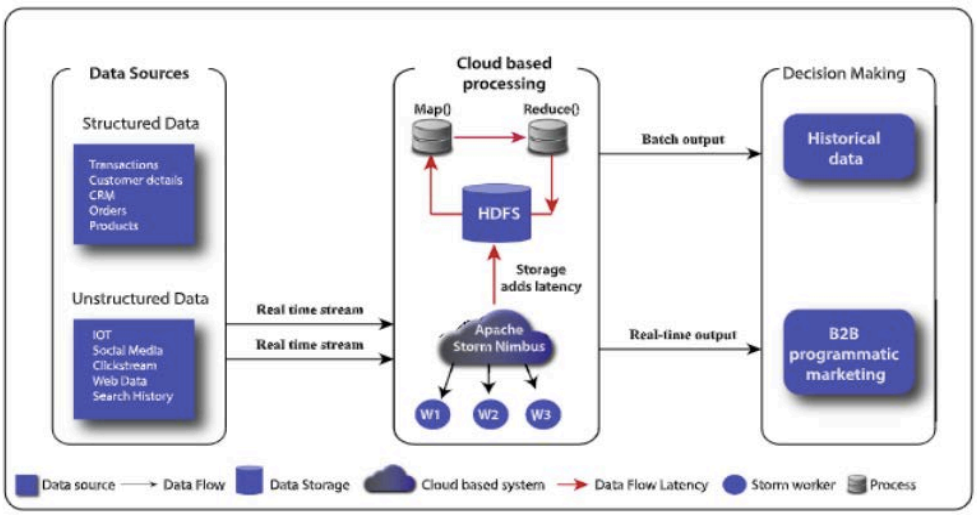
2. ขนาดของข้อมูล (volume) ตามชื่อที่ใช้ในการเรียกขานฐานข้อมูล Big Data ก็แสดงให้เห็นได้ว่าหัวใจหลักของฐานข้อมูลดังกล่าวอยู่ที่ขนาด และความใหญ่โตของข้อมูล ซึ่งไม่สามารถจัดเก็บในหน่วยบันทึกข้อมูลของเครื่องคอมพิวเตอร์โดยทั่วไป (Local Storage) จำเป็นต้องอาศัยระบบจัดเก็บข้อมูลขนาดใหญ่ เช่น การจัดเก็บข้อมูลบนระบบออนไลน์ (Cloud Storage) ซึ่งมีพื้นที่จัดเก็บมากกว่า และมีระบบรักษาความปลอดภัยของข้อมูลที่มีประสิทธิภาพ โดยในปัจจุบัน โลกที่กำลังก้าวเข้าสู่สังคมดิจิทัล ทุกอย่างกำลังถูกเคลื่อนจากโลกออฟไลน์ (Offline) หรือโลกที่ไม่เชื่อมต่อกับอินเตอร์เน็ต ไปสู่โลกออนไลน์ (Online) มากขึ้น สังเกตได้จากสิ่งที่อยู่ใกล้ตัว ยกตัวอย่างเช่น การซื้อสินค้าและบริการ ที่แต่ก่อนใช้วิธีการซื้อขายโดยใช้เงินสดผ่านการติดต่อที่หน้าร้าน (Walk-in) หรือเคาน์เตอร์ให้บริการ (Counter Service) แต่ในปัจจุบันการซื้อขายจำนวนมากถูกดำเนินการบนโลกอินเตอร์เน็ต ผ่านแพลตฟอร์ม (Platform) ต่างๆ ซึ่งนอกจากจะทำหน้าที่เสมือนหนึ่งเป็นหน้าร้านในการซื้อขายสินค้า และชำระค่าบริการแล้ว ยังทำหน้าที่ในการจัดเก็บข้อมูลการทำธุรกรรมตั้งแต่ ชื่อ ที่อยู่ ช่องทางการติดต่อของผู้ซื้อ ประเภทสินค้า จำนวนสินค้า และช่วงเวลาที่ซื้อ วิธีการชำระค่าสินค้าและบริการ ตลอดจนการตอบสนองต่อนโยบายส่งเสริมการตลาด เป็นต้น ซึ่งจินตนาการว่าในวันหนึ่งๆ การทำธุรกรรมซื้อขายสินค้าและบริการผ่านทางแพลตฟอร์มเกิดขึ้นนับล้านครั้ง และเกิดขึ้นทุกวัน ปริมาณข้อมูลที่ถูกจัดเก็บจะมีขนาดที่ใหญ่โตเพียงใด

โดยบทความชิ้นหนึ่งชื่อ “Big Data At Tesco: Real Time Analytics At The UK Grocery Retail Giant” โดย Marr (2016) ได้ทำการศึกษาการใช้ข้อมูล Big Data ในห้างสรรพสินค้าเทสโก้ ในสหราชอาณาจักร ซึ่งในปัจจุบันมี 3,500 สาขา และมีสินค้าเฉลี่ย 40,000 รายการในแต่ละสาขา ด้วยเหตุนี้ ทุกวันจะมีข้อมูลเกินกว่า 100 ล้านหน่วย แค่เฉพาะสำหรับการติดตามว่าสินค้าแต่ละชิ้นถูกซื้อขายไปแล้วหรือยัง อย่างไรก็ตาม ข้อมูลต่างๆ ที่ได้กล่าวถึงมานี้ล้วนแต่มีคุณูปการเป็นอย่างยิ่งต่อการศึกษาวิจัยทางด้านการตลาด และพฤติกรรมผู้บริโภค ซึ่งหากย้อนกลับไปในอดีตที่นักวิจัยทำการเก็บข้อมูลโดยใช้การสัมภาษณ์ด้วยแบบสอบถาม แทบจะเป็นไปไม่ได้เลยที่นักวิจัยจะสามารถเข้าถึงข้อมูลจำนวนมาก และมีความคลอบคลุมประชากรในการศึกษาได้ทั้งหมด หรือหากทำได้ก็จำเป็นต้องใช้งบประมาณในการจ้างบุคลากรในการเก็บข้อมูลมหาศาล และใช้เวลาในการประมวลผลยาวนานกว่าจะได้ผลการศึกษาออกมา

3. ความเป็นปัจจุบันทันด่วน (velocity) เป็นอีกหนึ่งคุณสมบัติที่สำคัญของข้อมูล Big Data ด้วยเหตุที่ข้อมูลในลักษณะดังกล่าวไม่ได้ใช้มนุษย์เป็นผู้จัดเก็บ หรือสังเกต (Observe) หากแต่ใช้ระบบคอมพิวเตอร์ในการจัดเก็บและประมวลผลแบบอัตโนมัติ ข้อมูลที่ได้จึงไม่มีความหน่วงด้านเวลา และมีความเป็นปัจจุบันอยู่เสมอ ในภาพประกอบที่ 2 แสดงระบบการจัดเก็บและประมวลผลข้อมูลตามเวลาจริง (Real-Time) ของฐานข้อมูลหนึ่ง ที่ใช้ในการศึกษาด้าน

การตลาด และพฤติกรรมผู้บริโภค ซึ่งสังเกตว่าแหล่งข้อมูล (Data Sources) ที่ไหลเข้าสู่ระบบมีทั้งข้อมูลในลักษณะของข้อมูลเชิงโครงสร้าง (เช่น ข้อมูลของลูกค้า ข้อมูลรายการสินค้าที่สั่ง ฯลฯ) และข้อมูลแบบไร้โครงสร้าง (เช่น ประวัติการค้นหาสินค้า ประวัติการเชื่อมต่อและการใช้งานเครือข่ายสังคมออนไลน์ ฯลฯ) โดยข้อมูลดังกล่าวจะถูกจัดเก็บและส่งเข้าสู่ระบบประมวลผล (Cloud Based Processing) ตามจุดเวลาที่เหตุการณ์เกิดขึ้นจริง (Real Time Stream) เช่น ทันทีที่ลูกค้ามีการสั่งสินค้าเข้ามาในระบบ ข้อมูลการสั่งสินค้าจะวิ่งเข้าสู่ระบบคลังข้อมูลในทันที ซึ่งต่างจากการเก็บข้อมูลด้วยแบบสอบถามที่ผู้เก็บข้อมูลจะรอจนกว่าจะทำการสัมภาษณ์และจัดเก็บข้อมูลจนเสร็จสิ้นกระบวนก่อน แล้วจึงนำข้อมูลทั้งหมดที่จัดเก็บได้ไปบันทึกลงระบบภายหลัง โดยสำหรับข้อมูลที่จำเป็นต้องอาศัยการสังเคราะห์ด้วยระบบ AI เช่น ข้อมูลแบบไร้โครงสร้าง ระบบจะดำเนินการสังเคราะห์และประมวลผลในทันที เพื่อนำข้อมูลทั้งหมด เข้าสู่กระบวนการวิเคราะห์ และตัดสินใจ (Decision Making) เพื่อให้ได้ผลรับแบบปัจจุบันทันด่วน (Real Time Output) เช่นกัน

แม้ว่าในการวิจัยทางสังคมศาสตร์ส่วนใหญ่จะไม่ได้เรียกร้องความเป็นปัจจุบันทันด่วนของข้อมูลแบบวินาทีต่อวินาทีมากนัก แต่อย่างน้อยที่สุด คุณสมบัติด้านดังกล่าวนี้สามารถช่วยการันตีได้ว่าข้อมูลที่ถูกจัดเก็บในฐานข้อมูล ถูกบันทึกตรงตามช่วงเวลาที่เหตุการณ์นั้นเกิดขึ้นจริง และสามารถลดปัญหาความคลาดเคลื่อนจากความทรงจำ (Recall Bias)[1] ได้อีกทางหนึ่งด้วย



ภาพประกอบที่ 2 แสดงระบบการจัดเก็บและประมวลผลข้อมูลตามเวลาจริง (Real-Time)
ที่มา: Jabbar, Akhtar, & Dani (2019)

[1] การสอบถาม หรือการสัมภาษณ์ถึงเหตุการณ์เกิดขึ้นมานานแล้ว อาจนำมาซึ่งข้อมูลที่คลาดเคลื่อนจากความเป็นจริง เพราะผู้ให้ข้อมูลหลงลืมข้อมูล หรือจำได้ไม่ถนัดนัก

4. ความปรับเปลี่ยนได้ของข้อมูล (Variability) เกิดขึ้นจากการที่ข้อมูลมีรูปแบบการจัดเก็บที่หลากหลาย และมีความสามารถในการเชื่อมโยงกับมิติต่างๆ ได้มากกว่า 1 มิติ จึงทำให้ผู้วิจัยสามารถปรับเปลี่ยนมุมมองในการวิเคราะห์ได้ตามความสนใจ ยกตัวอย่างเช่น หากผู้วิจัยทำงานศึกษาด้านศึกษาศาสตร์ และจัดเก็บข้อมูลการสอบวัดผลความรู้ด้านการเขียนเรียงความของนักเรียนในโรงเรียนแห่งหนึ่ง ผู้วิจัยสามารถนำข้อสอบเรียงความที่มีรูปแบบการตอบเชิงอัตนัย (Unstructured Data) ของเด็กนักเรียนมาทำการวิเคราะห์โครงสร้างการเขียนเรียงความว่ามีความสมเหตุสมผล ความสมบูรณ์ของประโยค ความถูกต้องของการสะกดคำอยู่ในระดับใดมากไปกว่านั้น ผู้วิจัยอาจทำการศึกษาว่าเด็กส่วนใหญ่เลือกหัวข้อเรียงความเกี่ยวกับเรื่องใด เช่น เรื่องสิ่งแวดล้อม เรื่องประชาธิปไตย เรื่องสิทธิและความเท่าเทียมทางเพศ ฯลฯ เพื่อหาความเชื่อมโยง หรือความสัมพันธ์กับคุณลักษณะส่วนบุคคลของเด็ก เป็นต้น ซึ่งจากตัวอย่างดังกล่าวนี้จะเห็นได้ว่าข้อมูลชุดเดียวกันสามารถนำไปสู่ผลการศึกษาหรือข้อสรุปในหลายประเด็น นอกจากนี้ข้อมูลแต่ละฐานข้อมูลสามารถเชื่อมโยงกับฐานข้อมูลอื่นๆ (Data integration) เช่น นำข้อมูลผลการสอบเขียนเรียงความของเด็ก มารวมกับฐานข้อมูลประวัติการยืมหนังสือจากห้องสมุดของโรงเรียน หรือข้อมูลการเข้าชั้นเรียนของเด็ก เป็นต้น ซึ่งคุณสมบัติในข้อนี้ สามารถทำให้นักวิจัยทำการศึกษาประเด็นต่างๆ ได้อย่างรอบด้านมากขึ้น ซึ่งต่างจากการสำรวจด้วยแบบสอบถามที่ผู้วิจัยต้องออกแบบเครื่องมือในการเก็บข้อมูลที่มีลักษณะตายตัว และไม่สามารถปรับเปลี่ยนได้ภายหลังจากการเก็บข้อมูล

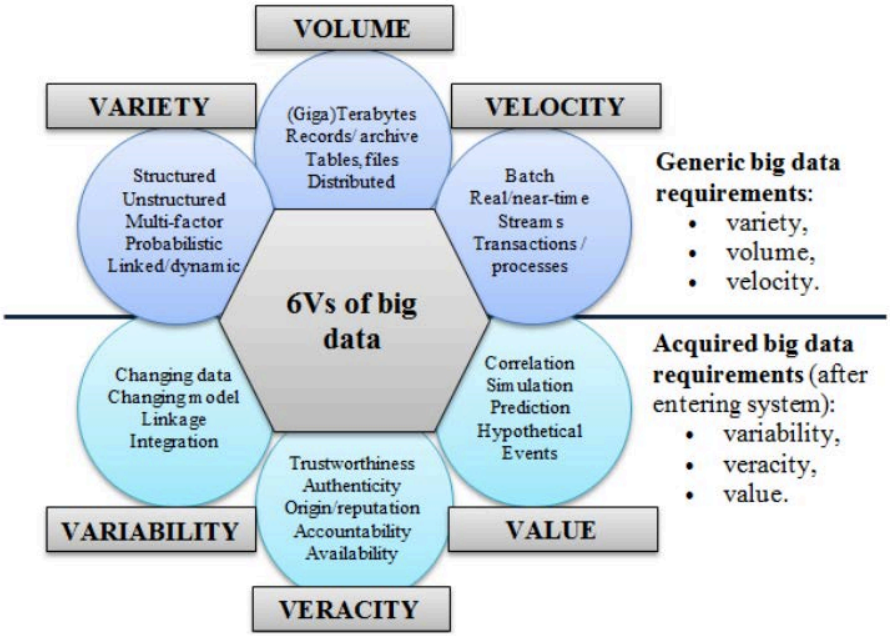
5. ความจริงแท้แน่นอน (Veracity) ข้อมูลที่จัดเก็บในฐานข้อมูล Big Data ล้วนเป็นเหตุการณ์ที่เกิดขึ้นจริง หรือเป็นพฤติกรรมที่ปรากฏให้เห็น ณ ช่วงเวลาหนึ่งๆ โดยมีเครื่องมือที่ใช้ในการวัด หรือจัดเก็บที่มีความเที่ยงตรง (Validity) และเชื่อถือได้ (Reliability) โดยในงานเขียนเรื่อง “Big-Data Analytics for Cloud, IoT and Cognitive Computing” โดย Hwang & Chen (2017) ได้ยกตัวอย่างการเก็บข้อมูลเกี่ยวกับพฤติกรรมการดูแลสุขภาพ เช่น พฤติกรรมการเคลื่อนไหว และการออกกำลังกาย ฯลฯ โดยใช้เครื่องมืออิเล็กทรอนิกส์แบบพกพาอย่างสายรัดข้อมืออัจฉริยะ (Smart watch) ที่ทำหน้าที่ในการเก็บสถิติการเคลื่อนไหวร่างกาย เช่น การเดิน การวิ่ง การนั่ง และการนอน ในแต่ละช่วงเวลาของวัน เพื่อนำข้อมูลดังกล่าวมาใช้ในการศึกษา และปรับเปลี่ยนพฤติกรรมการดูแลสุขภาพตนเองของผู้สูงอายุ เป็นต้น หรือในกรณีของการเก็บข้อมูลการซื้อขายสินค้าผ่านแพลตฟอร์มออนไลน์ การใช้งานเครือข่ายสังคมออนไลน์ ฯลฯ ก็ล้วนแล้วแต่มีระบบที่ทำหน้าที่ในการเก็บรวบรวมข้อมูลที่มีความเที่ยงตรงความแม่นยำในหน่วยวินาที และมีโอกาสในการผิดพลาดคลาดเคลื่อนได้น้อยมาก ถ้าเทียบกับการสัมภาษณ์ด้วยแบบสอบถามที่ผู้ตอบอาจใช้การประมาณการในการตอบหรือให้ข้อมูล ซึ่งความถูกต้องแม่นยำอย่างสมบูรณ์แบบของ Big Data นี้เองที่เป็นหัวใจหลักที่สำคัญประการหนึ่งของการศึกษาสังคม

แบบเป็นวิทยาศาสตร์ ที่ต้องการชั่งตวงวัดปรากฏการณ์ต่างๆ ออกมาได้อย่างถูกต้องเช่นเดียวกับที่นักวิทยาศาสตร์ทำการชั่งตวงวัดสิ่งต่างๆ ในห้องปฏิบัติการณ์ มากไปกว่านั้น ข้อมูลในฐานข้อมูล Big Data ทั้งหมด จะไม่นับรวมเอาสิ่งที่คาดเดาได้ยาก หรือมีลักษณะที่ไม่แน่นอนตายตัว เช่น ทัศนคติ ความคิดเห็น หรือมุมมองที่มีต่อเรื่องบางอย่าง เข้ามาปะปนในฐานข้อมูล ดังนั้น จึงอาจกล่าวได้ว่าข้อมูลที่ถูกจัดเก็บในฐานข้อมูล Big Data มีความใกล้เคียงกับคำว่า “จริงแท้แน่นอน” มากที่สุด หรือ หากมีความไม่จริงแท้ปะปนอยู่ก็สามารถประมาณการ และจำกัดความไม่จริงแท้ดังกล่าวให้ลดลงได้

6. ความมีคุณค่า (Value) โดยข้อมูลที่ถูกจัดเก็บในฐานข้อมูล Big Data จะต้องสามารถนำมาใช้ประโยชน์ในทางใดทางหนึ่ง เช่น สามารถนำมาใช้ในการวิเคราะห์วิจัยเพื่อออกแบบนโยบาย หรือนำมาซึ่งมาตรการบางสิ่งบางอย่างในการแก้ไขปัญหาขององค์กร หรือของสังคมได้อย่างมีประสิทธิภาพ ที่สำคัญข้อมูลในฐานข้อมูลควรมีคุณสมบัติที่สามารถนำไปวิเคราะห์นัยยะสำคัญทางสถิติ เช่น การหาความสัมพันธ์ (Correlation) การพยากรณ์ (Prediction) หรือการทดสอบสมมุติฐาน (Hypothesis Testing) ไม่ทางใดก็ทางหนึ่ง

ทั้งนี้ ในมุมของนักวิจัยทางสังคมศาสตร์ ข้อมูลที่จัดเก็บต้องสามารถตอบคำถามการวิจัย (Research Question) ได้อย่างครบถ้วน และตรงประเด็น ซึ่งในจุดนี้เองอาจเป็นข้อจำกัดของการใช้ฐานข้อมูล Big Data ในการศึกษาทางสังคมศาสตร์อยู่บ้าง เนื่องจาก ในบางกรณีนักวิจัยไม่ได้เป็นเจ้าของฐานข้อมูล และไม่ได้มีบทบาทในการกำหนดรูปแบบการเก็บข้อมูล จึงไม่สามารถออกแบบฐานข้อมูลที่ตอบโจทย์การวิจัยได้อย่างเต็มที่ อย่างไรก็ตาม ได้มีความพยายามของหน่วยงานด้านการศึกษาวิจัยในประเทศไทยที่มุ่งพัฒนาระบบจัดเก็บข้อมูล Big Data ขึ้นมาเพื่อใช้สำหรับการศึกษาวิจัยเป็นการเฉพาะ และในปัจจุบันสามารถออกแบบคลังข้อมูลให้สามารถตอบโจทย์ความต้องการของหน่วยงานได้อย่างมีประสิทธิภาพ (รายละเอียดในหัวข้อที่ 4)

อย่างไรก็ตาม ในมุมมองของ Moura & Serrao (2015) “3V แรก” ซึ่งประกอบด้วย ความหลากหลายของข้อมูล (Variety) ขนาดของข้อมูล (volume) และความเป็นปัจจุบันทันด่วน (velocity) นับเป็นองค์ประกอบพื้นฐาน (Generic Big Data Requirement) ที่มีความสำคัญต่อการเก็บรวบรวมข้อมูล หากการจัดเก็บข้อมูลไม่เป็นไปตามองค์ประกอบทั้ง 3 ประการนี้ ก็ยากที่ “3V หลัง” ซึ่งประกอบด้วย ความปรับเปลี่ยนได้ของข้อมูล (Variability) ความจริงแท้แน่นอน (Veracity) และความมีคุณค่า (Value) ของข้อมูลจะสามารถเกิดขึ้นตามมาได้



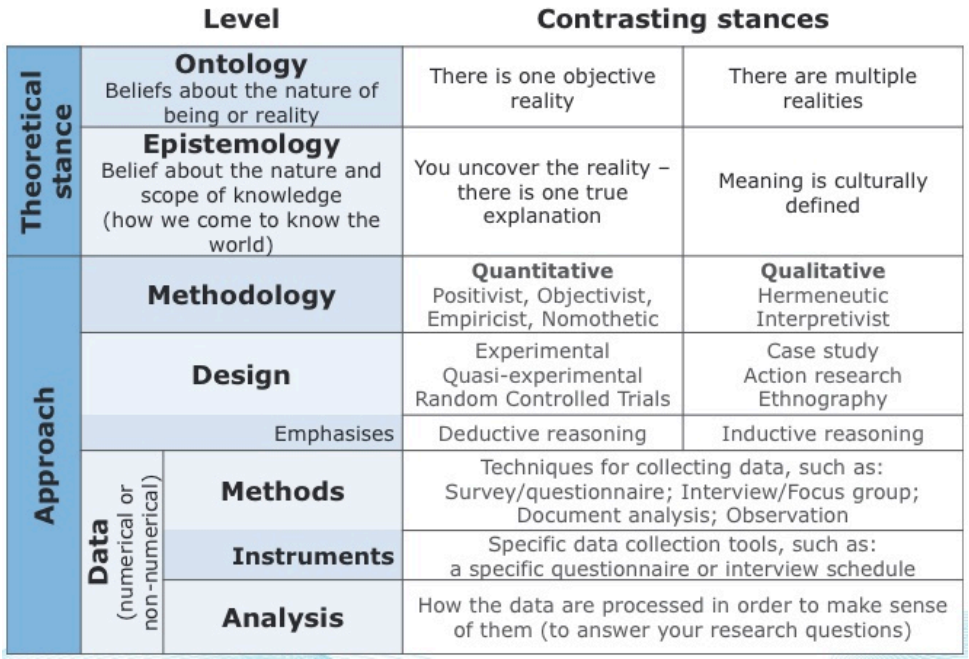
ภาพประกอบที่ 3 องค์ประกอบของฐานข้อมูล Big Data

ที่มา: Moura & Serrao (2015)

Big Data ภูมิทัศน์ใหม่ในการศึกษาสังคมศาสตร์ไทย

กระบวนทัศน์การวิจัยในทางสังคมศาสตร์ (Research Paradigm) ถูกแบ่งออกเป็น 2 สายหลัก ได้แก่ สายปฏิฐานนิยม (Positivism) และสายปรากฏการณ์นิยม (Phenomenology) โดยสำหรับสายปฏิฐานนิยมมีฐานความเชื่อในเชิงภววิทยา (Ontology) ว่าความรู้และความจริงในโลกมีลักษณะเป็นสากล อยู่นอกเหนือความรู้สึกนึกรู้ของบุคคล เป็นวัตถุที่สามารถชั่งตวงวัดและรอให้ผู้วิจัยเข้าไปค้นพบและศึกษา ซึ่งแตกต่างจากสายปรากฏการณ์นิยมที่มองว่า ความรู้และความจริงผูกติดกับความเชื่อ และวัฒนธรรมของบุคคลในลักษณะเป็นเนื้อเดียวกัน ดังนั้น การทำความเข้าใจมนุษย์คือการเข้าถึงความรู้และความจริงในเวลาเดียวกัน

จากจุดยืนทางภววิทยาที่แตกต่างกันนำไปสู่ญาณวิทยา (Epistemology) และวิธีวิทยา (Methodology) ในการศึกษาสังคมที่แตกต่างกันตามไปด้วย โดยสำหรับสายปฏิฐานนิยม ที่เชื่อในเรื่องความเป็นสากลของความรู้และความจริง ได้อาศัยวิธีการศึกษาเชิงปริมาณ (Quantitative Method) เป็นแกนกลางในการเข้าถึงข้อมูลที่ใช้ในการศึกษา ในขณะที่สายปรากฏการณ์นิยมอาศัยระเบียบวิธีวิจัยในเชิงคุณภาพ (Qualitative Method) เป็นแกนหลัก (รายละเอียดตามภาพประกอบที่ 4)



ภาพประกอบที่ 4 กระบวนทัศน์ในการศึกษาวิจัยทางสังคมศาสตร์

ที่มา: Peter, Rachelle, Miguel & Chin (2017)

อย่างไรก็ดี ทั้งสองสายนี้มีความเห็นแย้ง และความไม่ลงรอยกันกันเสมอมา จนนำไปสู่การต่อสู้แย่งชิงพื้นที่ทางวิชาการในนาม “สงครามกระบวนทัศน์ (Paradigm Wars)” แต่กระนั้นภายใต้สงครามระหว่างสองฝากฝั่ง ได้ก่อให้เกิดพัฒนาการทางด้านวิธีวิทยา และเทคนิควิธีต่างๆ ที่มีคุณูปการต่อการศึกษาวิจัยทางสังคมศาสตร์เป็นเงาตามไปด้วย ในฝั่งของนักวิจัยสายปฏิฐานนิยม ได้หยิบเอาเทคนิค Big Data มาช่วยเสริมหนุนให้กระบวนการวิจัยมีความเป็นเที่ยงตรง (Validity) น่าเชื่อถือ (Reliability) และปราศจากซึ่งอคติ ด้วยการลดความผิดพลาดคลาดเคลื่อน (Errors) ต่างๆ ที่อยู่ในกระบวนการศึกษา ซึ่งในมุมมองของผู้เขียน Big Data สามารถแก้ไขปัญหาและจัดการกับจุดอ่อนในกระบวนการวิจัยเชิงปริมาณที่ใช้กันอยู่ในปัจจุบันได้อย่างน้อย 4 ประการ ประกอบด้วย

ประการที่ 1 ลดความคลาดเคลื่อนที่เกิดจากการสุ่มตัวอย่าง (Sampling Error: SE)

ด้วยข้อจำกัดด้านงบประมาณ และระยะเวลา ทำให้นักวิจัยเชิงปริมาณไม่สามารถเก็บข้อมูลจากประชากร (Population) ที่ใช้ในการศึกษาได้ทั้งหมด จึงจำเป็นต้องเก็บข้อมูลจากกลุ่มตัวอย่าง (Sample) ซึ่งถือเป็นตัวแทนของประชากร (Representative) แทน อย่างไรก็ดี การเก็บ

ข้อมูลจากกลุ่มตัวอย่างแม้จะช่วยประหยัดเวลาและงบประมาณในการเก็บข้อมูล แต่ก็นำมาซึ่งปัญหาความคลาดเคลื่อนจากการสุ่ม (Sampling Error) ซึ่งสามารถคำนวณได้จาก สมการ (1)

สมการ (1) ---------- $$SE = z \times \sqrt{\frac{P(1-p)}{n}} \times 1 - \sqrt{\frac{n}{N}}$$

Z คือ ค่า Z-score ในระดับนัยยะสำคัญที่ระดับนัยยะสำคัญทางสถิติ α = 0.05

n คือ จำนวนกลุ่มตัวอย่าง

N คือ จำนวนประชากร

p คือ สัดส่วนของกลุ่มตัวอย่างที่คาดหวังว่าจะให้ข้อมูล

จะเห็นได้ว่าความคลาดเคลื่อนในการสุ่มตัวอย่าง จะมากหรือน้อยนั้นขึ้นอยู่กับจำนวนของกลุ่มตัวอย่าง (ค่า n) เป็นสำคัญ (ภายใต้ข้อสมมุติให้ค่า α = 0.05 และ p มีค่าคงที่ ณ ระดับใดระดับหนึ่งเสมอ) อย่างไรก็ตาม ในการจัดเก็บข้อมูลแบบ Big Data เป็นเทคนิคที่มีศักยภาพในการเข้าถึงข้อมูลของกลุ่มตัวอย่างได้มากกว่า หรือในบางกรณี Big Data อาจสามารถเก็บข้อมูลจากประชากรได้ทั้งหมด ซึ่งหากเป็นเช่นนั้น จะส่งผลทำให้ สมการ $1 - \sqrt{\frac{n}{N}}$ มีค่าเท่ากับ 0 ซึ่งส่งผลให้ค่าความคลาดเคลื่อนจากการสุ่มตัวอย่างเท่ากับ 0 ตามไปด้วย

ประการที่ 2 ลดความผิดพลาดในการทำงานของมนุษย์ (Human Error)

Big Data เป็นระบบจัดเก็บข้อมูลอัตโนมัติ ด้วยเหตุนี้ ผู้วิจัยจึงไม่มีความเสี่ยงจากการได้มาซึ่งข้อมูลที่ผิดพลาดอันเนื่องมาจากความประมาท หรือความไม่รอบคอบ เช่น การลงรหัส หรือการบันทึกข้อมูลผิดพลาด ฯลฯ ซึ่งถือเป็นธรรมชาติในการทำงานของมนุษย์ ต่างจากการวิธีการศึกษาเชิงปริมาณในปัจจุบันที่อาศัยแรงงานมนุษย์มากกว่าเทคโนโลยี ดังนั้น ย่อมมีโอกาสที่จะเกิดความผิดพลาดจากกระบวนการจัดกระทำข้อมูลมากกว่าเป็นปกติ

ประการที่ 3 ลดความคลาดเคลื่อนจากความทรงจำ (Recall Bias)

ตามธรรมชาติของการวิจัยเชิงปริมาณ มักมีการสอบถามหรือสัมภาษณ์ข้อมูลย้อนหลังหรือเหตุการณ์ที่เกิดขึ้นมานานแล้ว เช่น จำนวนเงินที่ใช้ในการบริโภคสินค้าและบริการประเภทต่างๆ ในรอบ 1 เดือนที่ผ่านมา ระยะเวลาที่ใช้ในการออกกำลังกายในแต่ละวัน หรือจำนวนครั้งในการทำธุรกรรมทางการเงินในแต่ละเดือน ฯลฯ ซึ่งข้อคำถามในลักษณะดังกล่าว เป็นข้อคำถามที่ยากต่อการตอบหรือให้ข้อมูล เนื่องจากเป็นเหตุการณ์ที่เกิดขึ้นมานานพอสมควร จึงเกิดการหลงลืม หรือมีความทรงจำที่คลาดเคลื่อนไปจากความเป็นจริง ซึ่งอาจนำมาซึ่งปัญหาเรื่องความถูกต้องแม่นยำของข้อมูล และมีผลต่อการวิเคราะห์ข้อมูลตามมา อย่างไรก็ตาม ปัญหาดังกล่าวจะไม่มีทางเกิดขึ้นในฐานข้อมูล Big Data เนื่องจากข้อมูลที่เกิดขึ้นจะไหลเข้าสู่คลังเก็บข้อมูลทันที ณ

เวลาที่เกิดเหตุการณ์ ตามองค์ประกอบที่ว่าด้วย “ความเป็นปัจจุบันทันด่วน (velocity)” ของข้อมูล ดังนั้น การได้มาซึ่งข้อมูลจึงมีความคลาดเคลื่อนจากความเป็นจริงน้อยกว่าโดยเปรียบเทียบ

ประการที่ 4 ลดปัญหาที่เกิดจากตัวแปรแทรกซ้อน หรือตัวแปรรบกวน (Confounding Bias)

ตัวแปรแทรกซ้อน หรือตัวแปรรบกวน เป็นตัวแปรจากภายนอก (Extraneous Variable) ที่มีสหสัมพันธ์โดยตรงหรือโดยผกผันกับทั้งตัวแปรตาม (Dependent Variable) และตัวแปรอิสระ (Independent Variable) ในงานศึกษา โดยตัวแปรดังกล่าวส่งผลทำให้เกิดความสัมพันธ์ลวง (Spurious Relationship) และการอนุมานทางสถิติที่ผิดพลาด ทั้งนี้ สาเหตุของการเกิดตัวแปรแทรกซ้อนอาจมาจากการที่ผู้วิจัยไม่ได้มีการทบทวนวรรณกรรมและงานศึกษาวิจัยที่เกี่ยวข้องให้ดีพอ ซึ่งปัญหาดังกล่าวเมื่อเกิดขึ้นแล้วจะสามารถแก้ไข หรือบรรเทาผลกระทบลงได้ยาก อย่างไรก็ตาม ปัญหาดังกล่าวนี้จะมีโอกาสเกิดขึ้นได้น้อยกว่าในกรณีที่นักวิจัยใช้ข้อมูล Big Data เนื่องจากองค์ประกอบด้าน “ความปรับเปลี่ยนได้ของข้อมูล (Variability)” ทำให้ผู้วิจัยสามารถบูรณาการ (Integrate) ฐานข้อมูลอื่นๆ มาร่วมใช้ในการวิเคราะห์ได้ หรืออาจนำข้อมูลอื่นๆ ในฐานข้อมูลมาใช้เป็นตัวแปรควบคุม เพื่อลดผลกระทบจากตัวแปรแทรกซ้อนอีกทางหนึ่ง

จากเหตุผล 4 ประการดังที่ได้กล่าวมาแล้วนี้ ทำให้ผู้เขียนเห็นว่า การใช้ข้อมูล Big Data สามารถลดปัญหา และข้อจำกัดที่เกิดจากการวิจัยเชิงปริมาณลงได้ โดยในแง่หนึ่ง Big Data ทำให้กระบวนการศึกษาวิจัยแบบภววิสัยมีความปลอดจากอคติมากขึ้น และลดปัญหาความคลาดเคลื่อนต่างๆ ที่เกิดขึ้นระหว่างกระบวนการได้มาซึ่งข้อมูล ซึ่งสิ่งเหล่านี้ช่วยผลักดันให้การวิจัยทางสังคมศาสตร์ด้วยระเบียบวิธีเชิงปริมาณมีเข้าใกล้เป้าหมายในเชิงอุดมคติมากขึ้น

ตัวอย่างการใช้ Big Data ในการวิจัยทางสังคมศาสตร์ในประเทศไทย

สำหรับวงวิชาการด้านสังคมศาสตร์ของไทยได้มีความพยายามในการใช้ข้อมูล Big Data ในการศึกษาวิจัยในวงจำกัด และอยู่ในช่วงการเริ่มต้นของการลองผิดลองถูก ซึ่งจากการศึกษาหาข้อมูลของผู้เขียน พบว่า มีหน่วยงานอย่างน้อย 2 แห่งที่กำลังใช้ฐานข้อมูล Big Data เป็นฐานข้อมูลหลักในการศึกษาวิจัยอยู่ในขณะนี้ ที่แรกคือ มูลนิธิสดศรี-สฤษดิ์วงศ์ ซึ่งร่วมกับภาคธุรกิจเพื่อสังคม (Social Enterprise) ในการพัฒนาแพลตฟอร์มสำหรับการเรียนการสอนในชั้นเรียน ชื่อ “Class Start” และถูกนำไปใช้งานในโรงเรียนกว่า 38 แห่งทั่วประเทศ มีนักเรียนที่ใช้งานในฐานะ “ผู้ใช้หลัก (User)” จำนวนทั้งสิ้น 4,945 คน โดยแพลตฟอร์มดังกล่าวจะเป็นเสมือนสื่อกลางในการติดต่อสื่อสารระหว่างคุณครู และเด็กนักเรียน และถูกใช้ในการประเมินผลสัมฤทธิ์ทางการศึกษาของเด็กนักเรียนอีกทางหนึ่งด้วย

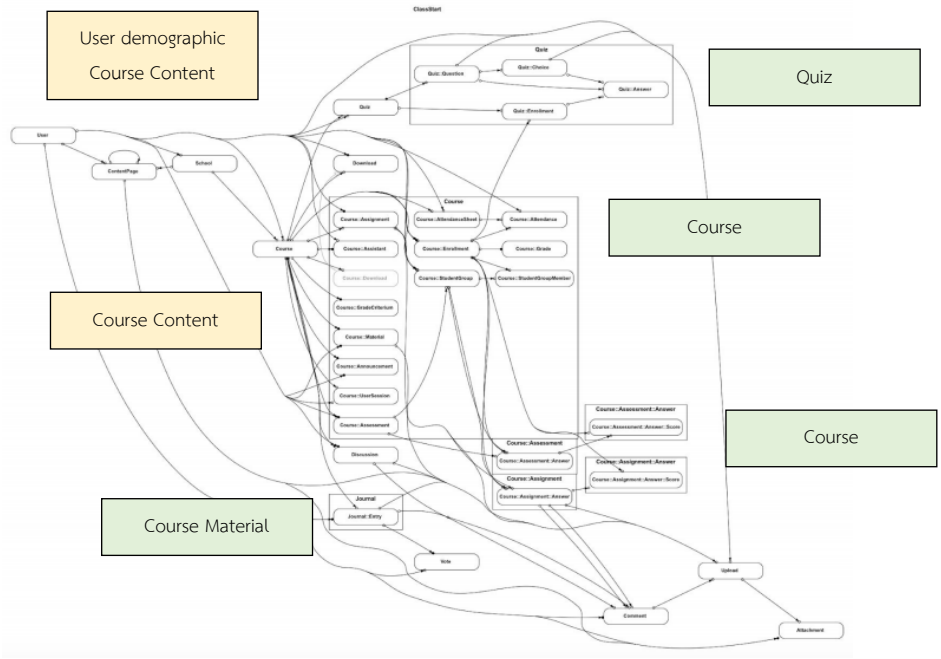
โดยในแพลตฟอร์มจะประกอบไปด้วยระบบการทำงาน 2 ส่วน ได้แก่ ส่วนที่หนึ่งข้อมูลทั่วไปของรายวิชา และประวัติผู้เรียน (User demographic Course Content) และส่วนที่สองกิจกรรมในรายวิชา (Course Content) ซึ่งประกอบด้วย ส่วนย่อยอีก 4 ส่วน ได้แก่ ส่วนชิ้นงานที่คุณครูมอบหมายให้นักเรียน (Course Assignment) ส่วนการประเมินผลเรียน (Course Assessment) ส่วนการทดสอบย่อย (Quiz) และส่วนของทรัพยากรสารสนเทศต่างๆ ที่ใช้ในการเรียน (Course Material) (รายละเอียดตามภาพประกอบที่ 5) นอกจากแพลตฟอร์มนี้จะช่วยให้ครูกับนักเรียนสามารถติดต่อสื่อสาร ในเรื่องการให้คำปรึกษา การรับและตรวจงานที่มอบหมายตลอดจนการจัดการสอบได้อย่างมีประสิทธิภาพ และรวดเร็วแล้ว แพลตฟอร์มยังทำหน้าที่เป็นตัวกลางในการจัดเก็บข้อมูลการใช้งานต่างๆ เพื่อนำมาใช้ในการวิเคราะห์วิจัย และปรับปรุงคุณภาพการสอนต่อไปได้อีกทางหนึ่งด้วย โดยผู้สอนสามารถเห็นความเคลื่อนไหวของนักเรียน ความรับผิดชอบในการส่งงาน พัฒนาการของคะแนนวัดผลสัมฤทธิ์ทางการศึกษา ซึ่งมีส่วนอย่างสำคัญในการปรับเปลี่ยนกลยุทธ์ในการสอนของคุณครู

นอกจาก Class Start แล้ว ทางมูลนิธิฯ ยังมีการจัดเก็บข้อมูลของ “โครงสร้างหลักสูตร” ที่ใช้สอนในโรงเรียนซึ่งอยู่ในโครงการเพาะพันธุ์ปัญญาทุกระดับชั้น ไว้ในฐานข้อมูล “21st Century Skill” และใช้เทคนิคการประมวลผลภาษาธรรมชาติหรือภาษามนุษย์ (Natural Language Processing) ในการวิเคราะห์ว่าหลักสูตรการเรียนการสอนของแต่ละโรงเรียนให้ความสำคัญกับทักษะแห่งศตวรรษที่ 21 หรือไม่ และให้ความสำคัญในระดับใด โดยมีเทคโนโลยีปัญญาประดิษฐ์ (Artificial Intelligence หรือ AI) ทำหน้าที่ในการวิเคราะห์ข้อมูลหลักสูตรการสอนของโรงเรียน ซึ่งเป็นข้อมูลแบบมีโครงสร้างไม่แน่นอน (Unstructured data) [2] เพื่อประเมินผลออกมาให้อยู่ในรูปของค่าทางสถิติ และแผนภาพ (Visualization)

การเข้ามาของ Big Data ในการวิจัยทางการศึกษา ส่งผลให้นักวิจัยสามารถเข้าถึงข้อมูลที่ไม่เคยเข้าถึงมาก่อน หรือเข้าถึงได้ยาก เช่น ข้อมูลเกี่ยวกับวินัยในการส่งงานของนักเรียนตลอดการเรียนในโรงเรียน หรือข้อมูลการเข้าถึงสารสนเทศทางการศึกษาของนักเรียนทั้งจากที่บ้านและที่โรงเรียน เป็นต้น ซึ่งข้อมูลเหล่านี้ จะช่วยให้นักวิจัยสามารถทำความเข้าใจกับปรากฏการณ์ต่างๆ ที่ไม่เคยมีการศึกษามาก่อนได้ หรือสามารถทำการศึกษาประเด็นใดประเด็นหนึ่งในระยะยาว (longitudinal study) เพราะข้อมูลถูกจัดเก็บแบบอนุกรมเวลา (Time Series) นับตั้งแต่เด็กนักเรียนเริ่มเข้าศึกษาในโรงเรียน จนกระทั่งจบการศึกษาจากโรงเรียน นับตั้งแต่ก้าวเท้าเข้าโรงเรียนในเวลาเช้า และก้าวออกจากโรงเรียนในเวลาเย็น นอกจากนี้ ความก้าวหน้าของ AI ซึ่งเติบโตขึ้นมา

[2] ข้อมูลแบบมีโครงสร้างไม่แน่นอน หมายถึง ข้อมูลที่อยู่ในรูปของ ข้อความ (Text) รูปภาพ (Picture) แผนภาพ (Diagram) ซึ่งมีรูปแบบของข้อมูลแตกต่างหลากหลาย ไม่ได้มีโครงสร้างที่ชัดเจน

พร้อมกับ Big Data ทำให้นักวิจัยสามารถศึกษา หรือตีความ (Interpret) ข้อมูลที่มีอยู่มหาศาล โดยไม่จำเป็นต้องอาศัยแรงงานนักวิจัย หรือผู้เชี่ยวชาญ



ภาพประกอบที่ 5 ระบบหลังบ้านของ แพลตฟอร์มสำหรับการเรียนการสอนในชั้นเรียน “Class Start”

ที่มา: เอกสารประกอบการประชุม มูลนิธิสดศรี-สฤษดิ์วงศ์ (2563)

นอกจากวงการด้านการศึกษาไทยแล้ว เทคนิคการวิเคราะห์ข้อมูล Big Data ยังถูกนำไปใช้ในการวิจัยทางเศรษฐศาสตร์ และตลาดแรงงาน โดยหน่วยงานที่ริเริ่มนำเทคนิคดังกล่าวมาใช้ คือ มูลนิธิสถาบันวิจัยเพื่อการพัฒนาประเทศไทย (TDRI) ซึ่งได้อาศัยข้อมูลจากเว็บไซต์จัดหางานทั่วประเทศมาใช้ในการวิเคราะห์สถานะของตลาดแรงงานทั้งด้านอุปสงค์ (Labor Supply) และอุปทาน (Labor Demand) ของตลาด เช่น ทักษะที่กำลังเป็นที่ต้องการของตลาด อาชีพที่กำลังขาดแคลน สถานะทักษะของแรงงาน และกำลังแรงงานในปัจจุบัน เป็นต้น

แม้ว่าในปัจจุบัน ประเทศไทยจะมีฐานข้อมูลที่รวบรวมสถิติแรงงาน เช่น ฐานข้อมูลการสำรวจภาวะการทำงานของประชากร (Labor Force Survey: LFS) ที่จัดเก็บสำนักงานสถิติแห่งชาติ โดยมีวงรอบการจัดเก็บทุกไตรมาสของทุกปี แต่กระนั้น ฐานข้อมูลดังกล่าวเป็นการสำรวจเฉพาะด้านอุปทานในตลาด หรือฝั่งแรงงาน ยังไม่ได้ครอบคลุมถึงฝั่งอุปสงค์ หรือฝั่งนายจ้าง และยังมีข้อจำกัดในด้านจำนวนของตัวอย่างเนื่องจากมีงบประมาณการเก็บข้อมูลที่สูง ซึ่งเป็นปัญหาเดียวกันกับฐานข้อมูลแรงงานอื่นๆ ในประเทศ แต่สำหรับการศึกษาข้อมูลแรงงานจาก

เว็บไซต์จัดหางาน สามารถแก้ไขข้อจำกัดดังกล่าวได้ และมีข้อได้เปรียบกว่าการสำรวจข้อมูลด้วยวิธีการสัมภาษณ์หลายประการ ดังต่อไปนี้

1) ข้อมูลในเว็บไซต์จัดหางานมีทั้งข้อมูลฝั่งอุปสงค์ หรือนายจ้าง และอุปทาน หรือแรงงาน โดยในส่วนของนายจ้างซึ่งอาจอยู่ในรูปของบริษัท ห้างร้าน หรือหน่วยงานราชการ จะดำเนินการลงประกาศรับสมัครงานซึ่งมีรายละเอียดครบถ้วนทั้งตำแหน่งที่เปิดรับ ทักษะที่ต้องการ อัตราค่าตอบแทน เกณฑ์การคัดเลือก หรือแม้แต่สถานที่ตั้งของหน่วยงาน ฯลฯ ในขณะที่ฝั่งแรงงานหรือผู้สมัครงาน จะมีการลงประวัติส่วนบุคคล เช่น เพศ อายุ ภูมิลำเนา สถานะทางครอบครัว ตลอดจนประวัติการศึกษา และทักษะความชำนาญต่างๆ เพื่อให้เว็บไซต์ทำการจับคู่งานที่เหมาะสม (Job matching) ได้อย่างรวดเร็วและตรงตามความต้องการ ซึ่งกลไกดังกล่าวนี้ ทำให้ฝั่งนายจ้าง และผู้สมัครงานมีแรงจูงใจในการให้ข้อมูลเชิงลึกเพื่อให้การจับคู่เป็นไปอย่างมีประสิทธิภาพ ข้อมูลที่เกิดขึ้นระหว่างกระบวนการจึงกลายมาเป็นทรัพยากรในการวิจัย และสามารถทำความเข้าใจตลาดแรงงานทั้งฝั่งอุปสงค์ และอุปทานได้เป็นอย่างดี

2) ข้อมูลในเว็บไซต์จัดหางานครอบคลุมตำแหน่งงานส่วนใหญ่ในประเทศ เพราะในปัจจุบันไม่ว่าจะเป็นบริษัท หรือหน่วยงานของภาครัฐ ต่างใช้ช่องทางออนไลน์เป็นช่องทางหลักในการประกาศรับสมัครงาน มิได้ประกาศตามสื่อสิ่งพิมพ์ เช่น หนังสือพิมพ์ หรือวารสารจัดหางาน ซึ่งเป็นสื่อออฟไลน์เช่นแต่ก่อน ดังนั้น จึงมั่นใจได้ว่าตำแหน่งงานแทบทั้งหมดอยู่ในฐานข้อมูลของเว็บไซต์จัดหางาน

3) ข้อมูลในเว็บไซต์จัดหางานเป็นข้อมูลที่มีความเป็นปัจจุบันทันด่วน (Real Time) และมีเจ้าหน้าที่ของบริษัทหรือหน่วยงานตรวจสอบประกาศให้มีความทันสมัยอยู่เสมอ (UpToDate) ต่างจากการสำรวจข้อมูลโดยการสัมภาษณ์ หรือใช้แบบสอบถาม ซึ่งมีกระบวนการในการลงพื้นที่ การขออนุญาตเข้าเก็บข้อมูล การลงรหัสข้อมูล และการตรวจสอบความถูกต้องของข้อมูลที่ค่อนข้างยาวนาน กว่าที่กระบวนการจะสิ้นสุด สภาพของตลาดแรงงานอาจเปลี่ยนแปลงไปจากเดิม และทำให้ข้อมูลที่เก็บรวบรวมมาได้นั้นไม่เป็นปัจจุบันอีกต่อไป

4) การใช้ข้อมูลแรงงานจากเว็บไซต์จัดหางานมีต้นทุนในการได้มาซึ่งข้อมูลต่ำกว่าการสำรวจโดยการสัมภาษณ์ หรือใช้แบบสอบถาม เนื่องจากเว็บไซต์มีลักษณะเป็นผู้รับข้อมูล ไม่ใช่ผู้เก็บข้อมูล จึงไม่มีต้นทุนในการค้นหาและสัมภาษณ์กลุ่มเป้าหมาย

แต่กระนั้น การใช้ข้อมูล Big Data จากเว็บไซต์จัดหางานก็มีข้อจำกัดหลายประการ เช่น ผู้ใช้ข้อมูลจำเป็นที่จะต้องมีความรู้ในการเขียนโปรแกรม (Programming) เพื่อดึงข้อมูลจากฐานข้อมูลเว็บไซต์ออกมาทำการวิเคราะห์ และมีความรู้ด้านเทคโนโลยีปัญญาประดิษฐ์ หรือ AI เนื่องจาก ข้อมูลในเว็บไซต์จัดหางานอาจมีทั้งข้อมูลเชิงโครงสร้าง (Structure Data) เช่น จำนวน

ของตำแหน่งงานที่เปิดรับสมัคร อัตราค่าตอบแทน ฯลฯ หรือข้อมูลอื่นใดที่มีรูปแบบของข้อมูลตายตัว เช่น เพศ อายุ ระดับการศึกษาของผู้สมัคร ฯลฯ ซึ่งข้อมูลเหล่านี้สามารถนำมาจัดกระทำ (Data manipulation) และประมวลผลทางสถิติได้โดยง่าย และข้อมูลที่ไม่มีโครงสร้าง หรือกึ่งโครงสร้าง (Unstructured Data) เช่น ข้อมูลประสบการณ์การทำงาน หรือข้อมูลทักษะแรงงานที่ต้องการ ซึ่งข้อมูลดังกล่าวมีรูปแบบ หรือความเป็นไปได้ที่หลากหลาย เช่น เมื่อกล่าวถึงทักษะที่จำเป็นต้องใช้ในการทำงาน อาจมีทั้ง “ทักษะการใช้โปรแกรมคอมพิวเตอร์” “ทักษะการพูดหรือฟังภาษาจีน” “ทักษะการทำงานเป็นทีม” หรือ “ทักษะด้านการเข้าสังคม” ฯลฯ ซึ่งในประกาศรับสมัครงานของแต่ละบริษัทอาจมีการใช้ “คำ” ที่กล่าวถึงทักษะเหล่านี้ที่หลากหลาย หรือซ่อนอยู่ในโครงสร้างของประโยคที่ซับซ้อน ดังนั้น การดึงเอา “รายการของทักษะ” ออกมาจากประกาศรับสมัครงานที่มีจำนวนมากมายมหาศาลจำเป็นต้องใช้เทคนิคการจัดจำแนกคำ (Text Classification) ซึ่งเป็นแขนงหนึ่งของการประมวลผลภาษาธรรมชาติ เพื่อให้สามารถล้วงเอาข้อเท็จจริงจากประกาศออกมาให้ได้มากที่สุด หรือแม้แต่ในกรณีที่ประกาศรับสมัครงาน หรือใบสมัครงานมีการสะกดคำผิด (Misspelled or Unseen Word) และมีการใช้คำที่ผันรูปได้หลากหลาย (Word Morphology) เช่นคำว่า “Stream” กับ “Streaming” ผู้ใช้ข้อมูลจะต้องมีการวางระบบการวิเคราะห์ที่สามารถแก้ไขปัญหาดังกล่าวนี้ได้

อย่างไรก็ดี การใช้เทคโนโลยี AI ในการจัดจำแนกคำ และการวิเคราะห์ข้อมูลจากประกาศรับสมัครงาน และใบสมัครงาน อาจมีความคลาดเคลื่อน หรือการตีความหมายคลาดเคลื่อน แต่กระนั้น AI สามารถเรียนรู้ความผิดพลาดที่เกิดขึ้น และสามารถลดความคลาดเคลื่อนในการจัดจำแนกลงได้ในระยะยาว เนื่องจากมีรูปแบบข้อมูลที่แตกต่างหลากหลายมากพอที่ AI จะสามารถเรียนรู้และปรับปรุงกระบวนการวิเคราะห์ได้ ในปัจจุบัน หน่วยงานของรัฐหลายแห่งเช่น สำนักงานปลัดกระทรวงแรงงาน มีแนวคิดที่จะนำเอาการจัดเก็บข้อมูลแบบ Big Data มาใช้ในการศึกษากลุ่มแรงงานนอกระบบ ซึ่งอยู่ระหว่างการศึกษาความเป็นไปได้

การเตรียมความพร้อมการใช้งาน Big Data ในงานศึกษาวิจัยทางสังคมศาสตร์

ในกรณีที่นักวิจัยทางสังคมศาสตร์มีความสนใจในการใช้ฐานข้อมูล Big Data ในการศึกษาวิจัยทางสังคม จำเป็นที่จะต้องมีการวางบทบาท หรือตำแหน่งแห่งที่ของตนเองให้ชัดเจนก่อนว่าจะอยู่ในสถานะของ “ผู้ใช้ข้อมูลในการศึกษาวิจัย” หรือ “ผู้สร้างฐานข้อมูลเพื่อการศึกษาวิจัย” ทั้งนี้ หากผู้วิจัยรู้จักฐานข้อมูลที่ตอบโจทย์ความต้องการ หรือตอบโจทย์การวิจัยได้อย่างครบถ้วนสมบูรณ์ และสามารถเข้าถึงฐานข้อมูลดังกล่าวได้ ผู้วิจัยก็สามารถจัดตนเองอยู่ในกลุ่ม “ผู้ใช้ข้อมูลในการศึกษาวิจัย” ซึ่งสำหรับในกลุ่มนี้ จำเป็นต้องอาศัยความรู้ในการใช้ซอฟต์แวร์ด้านการวิเคราะห์ข้อมูล Big Data ซึ่งในขณะนี้ มีซอฟต์แวร์ดังกล่าวให้เลือกใช้เป็นจำนวนมาก เช่น Hadoop ที่ออกแบบมาให้สามารถใช้งานการวิเคราะห์ข้อมูลขนาดใหญ่ (Large Dataset,

Internet Scale Dataset) และสามารถติดตั้งบนระบบปฏิบัติการณ์ได้หลายรูปแบบ นอกจากนี้ ผู้วิจัยจำเป็นต้องศึกษาเทคนิคการวิเคราะห์ข้อมูลทั้งในลักษณะของข้อมูลโครงสร้าง และไร้โครงสร้าง ซึ่งหากผู้วิจัยมีพื้นฐานความรู้ในทางสถิติมาก่อนแล้ว ย่อมสามารถเรียนรู้ทักษะในการวิเคราะห์ข้อมูลดังกล่าวได้โดยไม่ยาก เนื่องจากใช้ฐานคิดในการวิเคราะห์แบบเดียวกัน

แต่ถ้าหากผู้วิจัยไม่สามารถหาฐานข้อมูลที่ตอบโจทย์การวิจัยของตนเองได้อย่างครบถ้วน และจำเป็นต้องอยู่ในกลุ่ม “ผู้สร้างฐานข้อมูลเพื่อการศึกษาวิจัย” ผู้วิจัยต้องเริ่มต้นจากการประเมินความพร้อมและความต้องการของตนอย่างรอบด้านเสียก่อน เช่น งบประมาณ บุคลากรด้าน IT และเป้าหมายในการสร้างฐานข้อมูล ซึ่งในปัจจุบันมีบริษัทเอกชนที่ให้คำปรึกษา และบริการจัดทำฐานข้อมูลเป็นจำนวนมากในประเทศ ผู้วิจัยสามารถขอคำปรึกษา และประมาณการณ์ด้านงบประมาณที่ใช้ ตลอดจนประเมินความเป็นไปได้ในการจัดทำฐานข้อมูลจากที่ปรึกษาดังกล่าวได้เช่นกัน

อย่างไรก็ตาม การจัดทำคลังข้อมูล Big Data มีประเด็นเกี่ยวกับความเป็นส่วนตัว และการปกปิดข้อมูลของกลุ่มเป้าหมายที่เป็นข้อพึงระวัง ผู้วิจัยควรมีการให้ข้อมูลกับกลุ่มเป้าหมายถึงวัตถุประสงค์ของการเก็บข้อมูล ตลอดจนกระบวนการในการเก็บข้อมูลว่ามีการทำงานอย่างไร เช่น มีการเปิดระบบระบุพิกัด หรือเครื่องติดตามกลุ่มเป้าหมาย หรือมีการเข้าถึงข้อมูลในอุปกรณ์อิเล็กทรอนิกส์หรือไม่ และที่สำคัญกลุ่มเป้าหมาย หรือผู้ถูกเก็บข้อมูลต้องมีการลงนามในแบบการขอความยินยอม (Consent Form) ให้ถูกต้องตามหลักจริยธรรมการวิจัยในมนุษย์ เพื่อสร้างความมั่นใจให้กับผู้ถูกศึกษาว่าข้อมูลของตนจะไม่ถูกนำไปใช้ประโยชน์ในวัตถุประสงค์อื่นนอกเหนือจากการศึกษาวิจัยในประเด็นที่ตกลงไว้

สำหรับในส่วนของผู้ใช้ข้อมูลจากฐานข้อมูล Big Data ก็จำเป็นต้องมีความรับผิดชอบต่อกลุ่มผู้ถูกศึกษา และระมัดระวังในการวิเคราะห์ หรือนำเสนอผลการศึกษา มิให้สามารถระบุถึงตัวตนของกลุ่มเป้าหมายได้ เนื่องจากข้อมูล Big Data บางชุดข้อมูลอาจมีความละเอียดอ่อน และสามารถบ่งบอกถึงพฤติกรรม ลักษณะนิสัย หรือแม้แต่รสนิยมของกลุ่มเป้าหมาย ซึ่งกลุ่มเป้าหมายไม่พึงให้ข้อมูลส่วนบุคคลเหล่านี้หลุดรอดออกไปสู่สาธารณะ

เอกสารอ้างอิง

ภาษาอังกฤษ

Hwang, K., & Chen, M. (2017). *Big-Data Analytics for Cloud, IoT and Cognitive Computing.* New Jersey: Willey.

Jabbar, A., Akhtar, P., & Dani, S. (2019). Real-time big data processing for instantaneous marketing decisions: A problematization approach. *Industrial Marketing Management*, 1-12.

Malik, L., & Sangwan, S. (2015). Mapreduce Algorithms Optimizes the Potential of Big Data. *International Journal of Computer Science and Mobile Computing*, 663-674.

Marr, B. (2016, November 17). *Big Data At Tesco: Real Time Analytics At The UK Grocery Retail Giant.* Retrieved from Forbes, Inc: https://www.forbes.com/sites/bernardmarr/2016/11/17/big-data-at-tesco-real-time-analytics-at-the-uk-grocery-retail-giant/

Moura, J. A., & Serrao, C. (2015). Handbook of Research on Trends and Future Directions in Big Data and Web Intelligence. In N. Zaman, M. E. Seliaman, M. F. Hassan, & F. P. Marquez, *Security and Privacy Issues of Big Data* (p. 23). Pennsylvania: Information Science Reference.

Parlina, A., Ramli , K., & Murfi , H. (2020). Theme Mapping and Bibliometrics Analysis of One Decade of Big Data Research in the Scopus Database. *Information Journal*, 1-26.

Pegasus One. (2019, December 20). *Making Better Business Decisions with AI: Techniques to Analyze Unstructured Data.* Retrieved from Pegasus One, Inc.: https://www.pegasusone.com/making-better-business-decisions-with-ai-and-unstructured-data/

Peter, T., Rachelle, S. H., Miguel, N., & Chin, C, T,. (2017). Some guidance on conducting and reporting qualitative studies. *Computers & Education*,

Reichert, J.-R. &. (2017). *A supervised machine learning study of online discussion forums about type-2 diabetes.* Illinois: HealthCom.