

# พื้นฐานเหมืองข้อมูล

แนวทางการประยุกต์ใช้และการบูรณาการ  
งานเหมืองข้อมูลเบื้องต้น

พีระ ลีลอม



พื้นฐานเหมืองข้อมูล แนวทางการประยุกต์ใช้  
และการบูรณาการงานเหมืองข้อมูลเบื้องต้น



# พื้นฐานเหมืองข้อมูล แนวทางการประยุกต์ใช้ และการบูรณาการงานเหมืองข้อมูลเบื้องต้น

พีระ ลีวลม

 **สำนักพิมพ์**  
จุฬาลงกรณ์มหาวิทยาลัย

2566

225.-

พีระ ลีวลม

พื้นฐานเหมืองข้อมูล : แนวทางการประยุกต์ใช้และการบูรณาการงานเหมืองข้อมูลเบื้องต้น /

พีระ ลีวลม

1. เหมืองข้อมูล.

006.312

ISBN (e-book) 978-616-604-965-7

สปจ. Q40/10/2566



assคุณคำวิชาการ สู้สังคม  
Knowledge to All  
www.cupress.chula.ac.th

หนังสือเล่มนี้ผ่านการตรวจคุณภาพวิชาการจากผู้ทรงคุณวุฒิของสำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย

จัดทำโดย ผศ. ดร.พีระ ลีวลม

ภาควิชาวิทยาการคอมพิวเตอร์และสารสนเทศ คณะวิทยาศาสตร์และวิศวกรรมศาสตร์  
มหาวิทยาลัยเกษตรศาสตร์

โทร. 08-1391-2704

ครั้งที่ 1 พ.ศ. 2566 [CUP6612-019K]

บรรณาธิการอำนวยการ : ศาสตราจารย์ ดร.อรุณ หาญสืบสาย นางอรทัย นันทนาดิษฐ์

กองบรรณาธิการฝ่ายวิชาการ : ศาสตราจารย์กิตติคุณ ดร.ปิยนดา บุนนาค

รองศาสตราจารย์ ดร.พิมพ์พันธ์ เดชะคุปต์

รองศาสตราจารย์ นายแพทย์ชิษณุ พันธุ์เจริญ

รองศาสตราจารย์ ดร.วิมลวรรณ พิมพ์พันธุ์

ผู้ประสานงาน : วาสนา ชำเซ็น

พิสูจน์อักษร : ทศนีย์ ผิวขำ

ออกแบบปกและรูปเล่ม : สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย

สั่งซื้อได้ที่ : ศูนย์หนังสือแห่งจุฬาลงกรณ์มหาวิทยาลัย

ถนนพญาไท เขตปทุมวัน กรุงเทพฯ 10330

<http://www.chulabook.com>

โทร. 08-6323-3703-4

[customer@cubook.chula.ac.th](mailto:customer@cubook.chula.ac.th), [info@cubook.chula.ac.th](mailto:info@cubook.chula.ac.th)

Apps: CU-eBook Store

# คำนำ

หนังสือเล่มนี้เรียบเรียงขึ้นเพื่อใช้เป็นหนังสือเสริมความรู้ศาสตร์เหมืองข้อมูลในระดับอุดมศึกษา ซึ่งได้รับการยอมรับให้เป็นความรู้สำคัญในหลากหลายหลักสูตรทั้งในระดับปริญญาตรีและบัณฑิตศึกษา เฉพาะในมหาวิทยาลัยเกษตรศาสตร์ที่ผู้เขียนเป็นอาจารย์สอนอยู่มีการเรียนการสอนศาสตร์เหมืองข้อมูล ทั้งด้านเทคโนโลยีสารสนเทศ วิทยาการคอมพิวเตอร์ วิศวกรรมคอมพิวเตอร์ วิศวกรรมคอมพิวเตอร์ และสารสนเทศศาสตร์ วิศวกรรมซอฟต์แวร์และความรู้ วิศวกรรมไฟฟ้า สถิติ และการจัดการเทคโนโลยี อุตสาหกรรมเกษตร รวมแล้วไม่ต่ำกว่า 10 รายวิชา

การที่เหมืองข้อมูลมีการเรียนการสอนในหลากหลายสาขาวิชาเนื่องจากศาสตร์เหมืองข้อมูล มีศักยภาพในการค้นพบแพตเทิร์นน่าสนใจที่นำไปสู่ความรู้ที่แฝงอยู่ในชุดข้อมูลที่มาจากหลากหลายสาขาวิชา เหล่านั้น และยังสามารถประยุกต์ใช้ในชุดข้อมูลทางธุรกิจ เศรษฐกิจ สังคมศาสตร์ การท่องเที่ยว การเกษตร และการแพทย์ซึ่งพบการวิจัยประยุกต์ใช้ในประเทศไทยมากขึ้น

การตอบสนองความต้องการและผลักดันให้มีการประยุกต์ใช้ศาสตร์เหมืองข้อมูลอย่างกว้างขวาง ดังกล่าว ต้องการประมวลความรู้พื้นฐานเหมืองข้อมูลรวมถึงแนวทางประยุกต์ใช้ที่เพียงพอต่อการนำไป ปรับใช้หลากหลายตามความสนใจของผู้ใช้ อย่างไรก็ตาม หนังสือภาษาไทยที่เกี่ยวข้องกับศาสตร์เหมืองข้อมูล ปัจจุบันยังมีไม่มากนัก อีกทั้งส่วนใหญ่มุ่งเน้นไปที่การอธิบายแจกแจงแต่ละเทคนิคที่มีการใช้ในเหมืองข้อมูล เป็นหลัก การประยุกต์ใช้และบูรณาการเหมืองข้อมูลตามกระบวนการค้นพบความรู้มีการอธิบายแจกแจงอยู่น้อยมาก ซึ่งสิ่งที่ขาดนี้เป็นส่วนความรู้สำคัญที่นำไปสู่การประยุกต์ใช้เหมืองข้อมูลที่หลากหลาย

การประยุกต์และบูรณาการงานเหมืองข้อมูลดังกล่าวได้ ผู้ใช้จะต้องมีความรู้พื้นฐานในเทคนิคเหมืองข้อมูล มีความเข้าใจจุดเด่นและข้อจำกัดของแต่ละเทคนิคที่เพียงพอที่จะศึกษาค้นคว้าด้วยตนเองต่อไป มีความสามารถในการวิเคราะห์ลักษณะเฉพาะของชุดข้อมูลที่น่ามาใช่ว่าเหมาะสมกับเทคนิคเหมืองข้อมูลใด จากนั้นจึงเป็นการประยุกต์และบูรณาการงานเหมืองข้อมูลตามกระบวนการเหมืองข้อมูลและการค้นพบความรู้ โดยสิ่งเหล่านี้ได้แจกแจงเป็นรายละเอียดแต่ละบทของหนังสือเล่มนี้

ผู้เขียนคลุกคลีอยู่กับศาสตร์เหมืองข้อมูลมากกว่า 20 ปี ตั้งแต่ศึกษาระดับปริญญาเอกที่ห้องปฏิบัติการวิจัยวิเคราะห์และสืบค้นความรู้จากฐานข้อมูลขนาดใหญ่ (Data Analysis and Knowledge Discovery Laboratory, DAKDL) มหาวิทยาลัยเกษตรศาสตร์มาตั้งแต่ พ.ศ. 2546 ซึ่งท่าน รศ. ดร.กฤษณะ ไวยรัมย์เป็นผู้ควบคุมดูแลห้องปฏิบัติการนี้ กล่าวได้ว่าเป็นผู้บุกเบิกศาสตร์เหมืองข้อมูลในรุ่นแรก ๆ ของประเทศไทย การเรียนรู้จากต้นแบบผู้บุกเบิกศาสตร์เหมืองข้อมูลจึงเป็นสิ่งที่บ่มเพาะผู้เขียนที่มั่นคง ในการสร้างสรรค์งานเหมืองข้อมูล ทั้งนี้หลังสำเร็จการศึกษาผู้เขียนยังได้โอกาสในการสอนวิชาเหมืองข้อมูล ที่คณะวิทยาศาสตร์และวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์มาอย่างต่อเนื่องจนถึงปัจจุบัน ในหนังสือเล่มแรกนี้ผู้เขียนจึงตั้งใจอย่างยิ่งว่าความรู้และประสบการณ์เหล่านี้จะเป็นประโยชน์ต่อผู้อ่านพอสมควร

ในฐานะที่ผู้เขียนทำงานวิชาการในศาสตร์เหมืองข้อมูลมาอย่างยาวนาน จึงคาดหวังอย่างยิ่งว่า จะเป็นส่วนหนึ่งในการผลักดันศาสตร์เหมืองข้อมูลให้มีการประยุกต์ใช้งานที่กว้างขวางมากยิ่งขึ้นผ่านการเขียน หนังสือเล่มนี้ สามารถช่วยให้ผู้อ่านมีพื้นฐานความรู้ด้านเหมืองข้อมูลและมีทักษะเพิ่มเติมในการประยุกต์ บำรุงการงานเหมืองข้อมูล และสามารถช่วยผู้อ่านที่มีความเชี่ยวชาญในศาสตร์อื่น ให้มีความเข้าใจแนวทางการ ประยุกต์ใช้และบำรุงการงานเหมืองข้อมูลกับชุดข้อมูลที่ตนมีอยู่ ความเข้าใจซึ่งกันและกันระหว่าง ผู้เชี่ยวชาญเทคนิคเหมืองข้อมูลกับเจ้าของชุดข้อมูลที่เชี่ยวชาญในศาสตร์ที่หลากหลายนี้ เป็นสิ่งที่ผู้เขียน คาดหวังให้เป็นส่วนหนึ่งในการผลักดันให้เกิดการประยุกต์ใช้งานเหมืองข้อมูลที่กว้างขวางแพร่หลายมากยิ่งขึ้น

ผู้เขียนขอขอบคุณท่านอาจารย์ รศ. ดร.กฤษณะ ไวยมัย ผู้ประสิทธิ์ประสาทศาสตร์ด้านเหมืองข้อมูล ให้กับผู้เขียน และขอเป็นกำลังใจให้เพื่อน ๆ พี่ ๆ น้อง ๆ ผู้ร่วมห้องปฏิบัติการ DAKDL ทุกท่านในการ สร้างสรรค์งานเหมืองข้อมูลเป็นกำลังหลักของศาสตร์นี้อย่างต่อเนื่อง

ผู้เขียนขอขอบคุณคณะวิทยาศาสตร์และวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขต เฉลิมพระเกียรติ จังหวัดสกลนคร และภาควิชาวิทยาการคอมพิวเตอร์และสารสนเทศ ที่ให้โอกาสผู้เขียน ในหน้าที่การงาน การใช้ความรู้ความสามารถในการเรียนการสอน ตลอดจนโอกาสในการสั่งสมประสบการณ์ งานเหมืองข้อมูลจนถึงปัจจุบัน

ท้ายสุดนี้ ผู้เขียนขอขอบคุณ คุณพ่อวัลลภและคุณแม่เสริมสุข ลีวลม ที่คอยสนับสนุนลูกคนนี ตลอดระยะเวลาที่ผ่านมา ขอขอบคุณภรรยา คุณสุรัสวดี แสนสุข และลูกชายทั้ง 2 คน ด.ช.พฤกษ์และ ด.ช.พริษฐ์ ลีวลม ที่ทำให้ชีวิตมีความสุขราบรื่น การสร้างสรรค์งานใด ๆ ของผู้เขียนจะเกิดขึ้นไม่ได้หากไม่ได้ กำลังใจเหล่านี้

ขอแสดงความนับถือ

ผศ. ดร.พีระ ลีวลม

ภาควิชาวิทยาการคอมพิวเตอร์และสารสนเทศ

คณะวิทยาศาสตร์และวิศวกรรมศาสตร์

มหาวิทยาลัยเกษตรศาสตร์

# สารบัญ

|                                                                                                         | หน้า      |
|---------------------------------------------------------------------------------------------------------|-----------|
| <b>บทที่ 1 บทนำ</b>                                                                                     | <b>1</b>  |
| 1.1 ความเป็นมาของเหมืองข้อมูล                                                                           | 2         |
| 1.2 นิยามเหมืองข้อมูล                                                                                   | 3         |
| 1.3 ลักษณะของชุดข้อมูลที่สามารถนำมาใช้งานเหมืองข้อมูล                                                   | 6         |
| 1.4 ศักยภาพในการใช้งานเหมืองข้อมูลในประเทศไทย                                                           | 7         |
| 1.4.1 ความเป็นมาของเหมืองข้อมูลในประเทศไทย                                                              | 7         |
| 1.4.2 สรรวจแนวโน้มงานเหมืองข้อมูล                                                                       | 8         |
| 1.4.3 ลักษณะการประยุกต์ใช้งานเหมืองข้อมูลในปัจจุบันของประเทศไทย                                         | 10        |
| 1.5 สรุปท้ายบท                                                                                          | 15        |
| <b>บทที่ 2 เหมืองข้อมูลในกระบวนการค้นพบความรู้</b>                                                      | <b>19</b> |
| 2.1 กระบวนการค้นพบความรู้จากฐานข้อมูล                                                                   | 19        |
| 2.2 บทบาทเหมืองข้อมูลในกระบวนการค้นพบความรู้                                                            | 21        |
| 2.3 ลักษณะงานเหมืองข้อมูลและผลลัพธ์ที่พาดพิงในกระบวนการค้นพบความรู้จากข้อมูล                            | 30        |
| 2.4 สรุปท้ายบท                                                                                          | 32        |
| <b>บทที่ 3 การทำความเข้าใจและเตรียมพร้อมชุดข้อมูลก่อนทำเหมืองข้อมูล</b>                                 | <b>33</b> |
| 3.1 ข้อมูลและการจัดเก็บ                                                                                 | 33        |
| 3.2 ลักษณะเฉพาะและปัญหาตามธรรมชาติของชุดข้อมูล (Data Characteristics)                                   | 40        |
| 3.3 การเตรียมพร้อมข้อมูล (Data Preprocessing)                                                           | 44        |
| 3.3.1 การทำความสะอาดข้อมูล (Data Cleaning)                                                              | 44        |
| 3.3.2 การบูรณาการข้อมูลเข้าด้วยกัน (Data Integration)                                                   | 49        |
| 3.3.3 การทำให้ชุดข้อมูลเป็นปรกติ (Normalization) และการทำให้ชุดข้อมูลเป็นมาตรฐาน (Data Standardization) | 50        |
| 3.3.4 การลดขนาดชุดข้อมูล (Data Reduction) และการลดจำนวนมิติของชุดข้อมูล                                 | 53        |
| 3.4 สรุปท้ายบท                                                                                          | 56        |

|                                                                                                 |           |
|-------------------------------------------------------------------------------------------------|-----------|
| <b>บทที่ 4 พื้นฐานเหมืองข้อมูลเชิงทำนาย</b>                                                     | <b>57</b> |
| 4.1 งานเหมืองข้อมูลเชิงทำนาย (Predictive Data Mining)                                           | 58        |
| 4.2 หลักการพื้นฐานเทคนิคการจำแนกประเภทข้อมูล                                                    | 59        |
| 4.3 การกำหนดอัตราส่วนชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ                                            | 61        |
| 4.4 แนะนำเทคนิคต้นไม้การตัดสินใจ (Decision Tree)                                                | 62        |
| 4.4.1 การวัดความสามารถของพีเชอร์ในการลดการเจือปนกันของคลาส                                      | 66        |
| 4.4.2 เฟสงานการสร้างหรือการอุปนัยต้นไม้การตัดสินใจ (Decision Tree Induction)                    | 74        |
| 4.4.3 เฟสงานการลิดกิ่งต้นไม้การตัดสินใจ (Decision Tree Pruning)                                 | 76        |
| 4.5 แนะนำแบบจำลองการจำแนกประเภทข้อมูลเชิงสถิติ<br>(Statistical Based Classification)            | 79        |
| 4.6 แนะนำแบบจำลองการจำแนกประเภทข้อมูลเชิงกฎ (Rule Based Classification)                         | 84        |
| 4.7 การวัดคุณภาพการจำแนกประเภทข้อมูลด้วยเมทริกซ์ความสับสน<br>(Confusion Matrix)                 | 90        |
| 4.8 สรุปท้ายบท                                                                                  | 92        |
| <b>บทที่ 5 พื้นฐานเหมืองข้อมูลเชิงบรรยาย</b>                                                    | <b>95</b> |
| 5.1 งานเหมืองข้อมูลเชิงบรรยาย (Descriptive Data Mining)                                         | 95        |
| 5.2 แนะนำเทคนิคการจับกลุ่มข้อมูล (Data Clustering)                                              | 96        |
| 5.2.1 ปัญหาการจับกลุ่มข้อมูลและการประยุกต์ใช้                                                   | 96        |
| 5.2.2 นิยามและหลักการพื้นฐานการจับกลุ่มข้อมูล                                                   | 98        |
| 5.2.3 ประเภทของพีเชอร์ (Feature Type)                                                           | 99        |
| 5.2.4 ฟังก์ชันระยะทาง (Distance Function)                                                       | 103       |
| 5.2.5 พีเชอร์ทวิภาคที่อสมมาตร (Asymmetric Binary Feature)                                       | 106       |
| 5.2.6 การรวมระยะทางจากทุกพีเชอร์เข้าด้วยกัน                                                     | 107       |
| 5.2.7 กลุ่มการจับกลุ่มข้อมูลด้วยค่าเฉลี่ยศูนย์กลางกลุ่ม $k$ กลุ่ม<br>( $k$ -Means Clustering)   | 108       |
| 5.2.8 การจับกลุ่มข้อมูลเชิงลำดับชั้นแบบเกาะกลุ่ม<br>(Agglomerative Hierarchical Clustering)     | 110       |
| 5.2.9 การจับกลุ่มข้อมูล DBSCAN (Density-Based Spatial Clustering of<br>Applications with Noise) | 112       |

|                                                                                                                | หน้า       |
|----------------------------------------------------------------------------------------------------------------|------------|
| 5.3 แนะนำเทคนิคการค้นพบกฎความสัมพันธ์ (Association Rules Discovery)                                            | 114        |
| 5.3.1 ความสัมพันธ์เชิงปริมาณในฐานข้อมูล                                                                        | 115        |
| 5.3.2 กฎความสัมพันธ์และคัพที่ที่เกี่ยวข้อง                                                                     | 117        |
| 5.3.3 ปัญหาพื้นฐานของการค้นพบกฎความสัมพันธ์                                                                    | 120        |
| 5.3.4 ขั้นตอนวิธี Apriori Algorithm                                                                            | 121        |
| 5.3.5 ตัวอย่างการคำนวณตามขั้นตอนวิธี Apriori Algorithm                                                         | 127        |
| 5.4 สรุปท้ายบท                                                                                                 | 130        |
| <b>บทที่ 6 กระบวนการทำเหมืองข้อมูลเบื้องต้น</b>                                                                | <b>133</b> |
| 6.1 กระบวนการทำเหมืองข้อมูลแบบ SEMMA                                                                           | 133        |
| 6.2 กระบวนการทำเหมืองข้อมูลแบบ CRISP                                                                           | 135        |
| 6.3 กระบวนการทำเหมืองข้อมูลกับการค้นพบความรู้                                                                  | 136        |
| 6.4 ภาพรวมกระบวนการทำเหมืองข้อมูล                                                                              | 137        |
| 6.5 สรุปท้ายบท                                                                                                 | 143        |
| <b>บทที่ 7 แนวทางการประยุกต์ใช้เหมืองข้อมูล</b>                                                                | <b>145</b> |
| 7.1 การทำงานเป็นทีมในการประยุกต์ใช้เหมืองข้อมูล                                                                | 146        |
| 7.2 ชุดข้อมูลแสดงคุณสมบัติเชิง “ข้อเท็จจริง” (Facts) หรือ “พฤติกรรม” (Behaviors)                               | 148        |
| 7.3 ชุดข้อมูลกับการก่อรูปเป้าหมายเหมืองข้อมูล                                                                  | 154        |
| 7.4 ชุดข้อมูลมีความเหมาะสมกับเทคนิคเหมืองข้อมูลเชิงทำนายหรือเชิงบรรยาย                                         | 156        |
| 7.5 ผลลัพธ์จากเทคนิคเหมืองข้อมูลเน้นความสัมพันธ์ระหว่างพีเชอร์ที่เข้าใจง่าย<br>หรือเน้นที่ประสิทธิภาพการใช้งาน | 157        |
| 7.6 ชุดข้อมูลมีพีเชอร์ที่มีโครงสร้างรากต้นไม้แฝงอยู่หรือไม่                                                    | 157        |
| 7.7 นำเสนอแนวทางการประยุกต์ใช้เหมืองข้อมูล                                                                     | 158        |
| 7.8 สรุปท้ายบท                                                                                                 | 162        |
| <b>บทที่ 8 แนวทางการบูรณาการงานเหมืองข้อมูล</b>                                                                | <b>165</b> |
| 8.1 วิเคราะห์แนวทางบูรณาการงานเหมืองข้อมูลในกระบวนการค้นพบความรู้                                              | 165        |
| 8.2 มุมมองเชิงบูรณาการในแนวทางประยุกต์ใช้งานเหมืองข้อมูล<br>เพื่อการออกแบบกระบวนการเหมืองข้อมูลเชิงบูรณาการ    | 168        |
| 8.3 การบูรณาการงานเหมืองข้อมูลในขั้นตอนแบบจำลองเหมืองข้อมูล                                                    | 170        |

|                                                                                                                                                           |            |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| 8.4 การบูรณาการงานเหมืองข้อมูลในขั้นตอนที่เปรียบเสมือนการเตรียมพร้อม<br>สำหรับเหมืองข้อมูลรอบใหม่                                                         | 173        |
| 8.5 การบูรณาการงานเหมืองข้อมูลในขั้นตอนที่เปรียบเสมือนการผสมผสานผลลัพธ์ที่ได้<br>จากเหมืองข้อมูลเข้าด้วยกัน เป็นการระดมองค์ความรู้ในกระบวนการค้นพบความรู้ | 175        |
| 8.6 สรุปท้ายบท                                                                                                                                            | 177        |
| <b>บทที่ 9 กรณีศึกษาการประยุกต์และบูรณาการเทคนิคเหมืองข้อมูล</b>                                                                                          | <b>179</b> |
| 9.1 กรณีชุดข้อมูลข้าวพันธุ์ Cammeo และ Osmancik                                                                                                           | 180        |
| 9.1.1 แนวทางประยุกต์เทคนิคเหมืองข้อมูลกับชุดข้อมูลข้าวพันธุ์ Cammeo<br>และ Osmancik                                                                       | 181        |
| 9.1.2 แนวทางบูรณาการเทคนิคเหมืองข้อมูลกับชุดข้อมูลข้าวพันธุ์ Cammeo<br>และ Osmancik                                                                       | 191        |
| 9.2 ชุดข้อมูลการใช้จ่ายของลูกค้า                                                                                                                          | 196        |
| 9.3 ชุดข้อมูลการศึกษานอกโรงเรียนจังหวัดสกลนคร                                                                                                             | 201        |
| 9.4 งานวิจัยบูรณาการเหมืองข้อมูลเชิงบรรยายด้วยผังก้างปลา                                                                                                  | 205        |
| 9.5 กรณีศึกษาเทคนิคบูรณาการถ่วงข้อมูล Bagging                                                                                                             | 208        |
| 9.6 การจำแนกประเภทข้อมูลด้วยกฎความสัมพันธ์คลาส                                                                                                            | 210        |
| 9.7 ต้นไม้กฎความสัมพันธ์                                                                                                                                  | 210        |
| 9.8 ป่าบรรยาย                                                                                                                                             | 214        |
| 9.9 สรุปท้ายเล่ม                                                                                                                                          | 215        |
| บรรณานุกรม                                                                                                                                                | 217        |
| ดรรชนี                                                                                                                                                    | 221        |

# สารบัญภาพ

|                                                                                                                                                                                                     | หน้า |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| ภาพที่ 1-1 แสดงจำนวนงานวิจัยเหมืองข้อมูลในฐานข้อมูล scopus ในช่วง ค.ศ. 2000-2022 สืบค้นเมื่อวันที่ 4 มีนาคม ค.ศ. 2023 (เฉพาะงานวิจัยที่ปรากฏคำว่า “data mining” ใน หัวเรื่อง, บทคัดย่อ, และคำสำคัญ) | 8    |
| ภาพที่ 1-2 แสดงจำนวนงานวิจัยเหมืองข้อมูลในประเทศไทย สืบค้นจาก scholar.google.com ในช่วง พ.ศ. 2543-2565 (เฉพาะงานวิจัยที่ปรากฏคำว่า “เหมืองข้อมูล”) สืบค้นเมื่อวันที่ 4 มีนาคม พ.ศ. 2566             | 9    |
| ภาพที่ 1-3 แสดงสัดส่วนการประยุกต์ใช้งานเหมืองข้อมูลด้านต่าง ๆ ในประเทศไทย สืบค้นจาก scholar.google.com ในช่วง พ.ศ. 2563-2565                                                                        | 10   |
| ภาพที่ 1-4 แสดงลำดับความสัมพันธ์ของเนื้อหาในแต่ละบทของหนังสือเล่มนี้                                                                                                                                | 17   |
| ภาพที่ 2-1 แสดงบทบาทของเหมืองข้อมูลในกระบวนการค้นพบความรู้ ปรับปรุงจากงานของ Usama Fayyad et al. ใน ค.ศ. 1996                                                                                       | 22   |
| ภาพที่ 2-2 กระบวนการค้นพบความรู้จากข้อมูลที่ปรับปรุงจากงานของ Han et al. (2022)                                                                                                                     | 23   |
| ภาพที่ 2-3 กระบวนการทำเหมืองข้อมูลปรับแต่งจาก Tan et al. (2019)                                                                                                                                     | 25   |
| ภาพที่ 2-4 แสดงกระบวนการค้นพบความรู้จากขั้นตอนวิธีวิจัย ปรับปรุงจาก Chakravanti Rajagopalachari Kothari                                                                                             | 26   |
| ภาพที่ 2-5 นำเสนอการปรับปรุงภาพรวมกระบวนการค้นพบความรู้จากข้อมูล                                                                                                                                    | 29   |
| ภาพที่ 3-1 แสดงข้อมูลกราฟแสดงความสัมพันธ์เว็บลิงก์ของเว็บกลุ่มหนึ่ง                                                                                                                                 | 39   |
| ภาพที่ 3-2 แสดงตัวอย่างธรรมชาติมิติข้อมูลจากชุดข้อมูลที่มีวัตถุข้อมูล 10 จุด เมื่อค่อย ๆ เพิ่มมิติข้อมูลจาก 1 มิติไป 3 มิติ                                                                         | 42   |
| ภาพที่ 3-3 แสดงตัวอย่างการปรับลดข้อมูลรบกวนด้วยสมการถดถอยเชิงเส้น                                                                                                                                   | 48   |
| ภาพที่ 3-4 แสดงการกำจัดข้อมูลที่ไม่สนใจด้วยการทำคลัสเตอร์                                                                                                                                           | 49   |
| ภาพที่ 3-5 แสดงแพตเทิร์นของอายุคนร้อยคนเปรียบเทียบกันระหว่างชุดข้อมูลตามปกติ กับชุดข้อมูลที่ทำค่าต่อเนื่องเป็นค่าไม่ต่อเนื่อง                                                                       | 52   |
| ภาพที่ 4-1 แสดงขั้นตอนการเลือกและจัดการข้อมูลที่นำมาพัฒนาแบบจำลองการจำแนกประเภทข้อมูล                                                                                                               | 60   |
| ภาพที่ 4-2 ขั้นตอนการพัฒนาแบบจำลองการจำแนกประเภทข้อมูล                                                                                                                                              | 61   |
| ภาพที่ 4-3 ตัวอย่างชุดข้อมูลและต้นไม้การตัดสินใจที่พบจากชุดข้อมูล                                                                                                                                   | 63   |
| ภาพที่ 4-4 ตัวอย่างการใช้ต้นไม้การตัดสินใจในการทำนายคลาสของระเบียบใหม่ที่ยังไม่ทราบคลาส                                                                                                             | 65   |

|            |                                                                                                                            |     |
|------------|----------------------------------------------------------------------------------------------------------------------------|-----|
| ภาพที่ 4-5 | แสดงความสัมพันธ์ระหว่างค่าพีเชอร์เพคกับค่าคลาส A กับ B                                                                     | 66  |
| ภาพที่ 4-6 | ตัวอย่างการเลือกพีเชอร์เป็นค่าประจำโหนดแรกจากข้อมูลในภาพที่ 4-3                                                            | 75  |
| ภาพที่ 4-7 | แสดงตัวอย่างปัญหาโอเวอร์ฟิตติ้ง                                                                                            | 77  |
| ภาพที่ 4-8 | แสดงโครงสร้างแบบจำลองโดยใช้สถิติความน่าจะเป็นแบบเบย์<br>จากตัวอย่างในตารางที่ 4-1                                          | 82  |
| ภาพที่ 5-1 | แสดงการแปลงพีเชอร์ “สี” ที่บันทึกค่ามาตฐานามบัญญัติเป็น “สีขาว” “สีดำ”<br>และ “สีน้ำเงิน” ที่บันทึกค่าในลักษณะข้อมูลทวิภาค | 103 |
| ภาพที่ 5-2 | แสดงตัวอย่างการจับกลุ่มข้อมูลด้วยค่าเฉลี่ยศูนย์กลางกลุ่ม $k$ กลุ่ม<br>( $k$ -MEANS Clustering)                             | 109 |
| ภาพที่ 5-3 | แสดงตัวอย่างการจับกลุ่มข้อมูลเชิงลำดับชั้นแบบเกาะกลุ่ม<br>(Agglomerative Hierarchical Clustering)                          | 112 |
| ภาพที่ 5-4 | แสดงตัวอย่างการจับกลุ่มข้อมูลด้วยวิธี DBSCAN                                                                               | 113 |
| ภาพที่ 5-5 | แสดงโครงสร้างความสัมพันธ์ตามหลัก Anti-Monotone ของไอเทมเซต<br>ที่มีสมาชิก 5 ตัวและสับเซตของมัน                             | 122 |
| ภาพที่ 5-6 | แสดงไอเทมเซตที่คัดกรองออกหรือ Pruned Itemsets ตามหลัก Anti-Monotone<br>ของตัวอย่างในภาพที่ 5-5                             | 123 |
| ภาพที่ 5-7 | แสดงการคัดกรองกฎตามหลัก Anti-Monotone                                                                                      | 124 |
| ภาพที่ 6-1 | ภาพรวมกระบวนการค้นพบความรู้                                                                                                | 137 |
| ภาพที่ 7-1 | นำเสนอแนวทางการประยุกต์ใช้งานเหมืองข้อมูล                                                                                  | 159 |
| ภาพที่ 8-1 | แสดงการบูรณาการงานเหมืองข้อมูลในภาพรวมกระบวนการค้นพบความรู้<br>ที่ผู้เขียนปรับปรุงจากที่นำเสนอในภาพที่ 2-5                 | 166 |
| ภาพที่ 8-2 | แสดงมุมมองเชิงบูรณาการงานเหมืองข้อมูลจากกลุ่มเทคนิคประยุกต์ใช้งาน<br>เหมืองข้อมูล                                          | 169 |
| ภาพที่ 9-1 | แสดงต้นไม้มากการตัดสินใจที่ได้จากชุดข้อมูลข่าว                                                                             | 183 |
| ภาพที่ 9-2 | แสดงการกระจายตัวของพันธุ์ข่าวในชุดข้อมูลข่าว Cammeo และ Osmancik                                                           | 184 |
| ภาพที่ 9-3 | แสดงต้นไม้มากการตัดสินใจที่พบจากการปรับแต่งค่าพีเชอร์ในชุดข้อมูลข่าว                                                       | 185 |
| ภาพที่ 9-4 | แสดงการจับกลุ่มข้อมูลจากชุดข้อมูลข่าวด้วยเทคนิค Expectation-Maximization                                                   | 190 |
| ภาพที่ 9-5 | แสดงพีเชอร์ที่มีศักยภาพน้อยในการจำแนกพันธุ์ข่าวในกลุ่ม Cluster 3, 28 และ 30                                                | 194 |
| ภาพที่ 9-6 | แสดงต้นไม้มากการตัดสินใจในการอธิบายความสัมพันธ์ระหว่างพีเชอร์ที่พัฒนาขึ้นใหม่<br>กับพันธุ์ข่าว Cammeo และ Osmancik         | 195 |

|                                                                                                                                                                                                      | หน้า |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| ภาพที่ 9-7 แสดงการกระจายตัวของพีเชอร์ SpendingScore                                                                                                                                                  | 198  |
| ภาพที่ 9-8 แสดงตำแหน่งการจัดกลุ่มด้วยค่าเฉลี่ยศูนย์กลางกลุ่มจากพีเชอร์ SpendingScore<br>ที่ $k = 3$                                                                                                  | 199  |
| ภาพที่ 9-9 แสดงตำแหน่งการจัดกลุ่มด้วยค่าเฉลี่ยศูนย์กลางกลุ่มจากพีเชอร์ SpendingScore<br>ที่ $k = 6$                                                                                                  | 199  |
| ภาพที่ 9-10 แสดงตำแหน่งการจัดกลุ่มด้วยค่าเฉลี่ยศูนย์กลางกลุ่มจากพีเชอร์ SpendingScore<br>ที่ $k = 3$ และ $k = 6$                                                                                     | 200  |
| ภาพที่ 9-11 แสดงต้นไม้การตัดสินใจที่ได้จากชุดข้อมูลการใช้จ่ายของลูกค้า                                                                                                                               | 201  |
| ภาพที่ 9-12 แสดงกลุ่มผู้เรียนที่ได้จากเทคนิคการจับกลุ่มข้อมูล Expectation-Maximization<br>ตามมิติความสัมพันธ์ระหว่างระดับ GPA ของกลุ่มวิชาวุฒนธรรมและระดับ GPA รวม<br>ของผู้สำเร็จการศึกษา พ.ศ. 2563 | 203  |
| ภาพที่ 9-13 แสดงกลุ่มผู้เรียนที่ได้จากเทคนิคการจับกลุ่มข้อมูล Expectation-Maximization<br>ตามมิติความสัมพันธ์ระหว่างอายุผู้เรียนจากฐาน พ.ศ. 2564 และระดับ GPA รวม<br>ของผู้สำเร็จการศึกษา พ.ศ. 2563  | 203  |
| ภาพที่ 9-14 ผังก้างปลาแสดงเค้าโครงพีเชอร์ที่สัมพันธ์กับโรคหัวใจและหลอดเลือด<br>ที่ได้จากเทคนิคการจับกลุ่มข้อมูล, การค้นพบกฎความสัมพันธ์, ต้นไม้กฎความสัมพันธ์<br>และต้นไม้การตัดสินใจ                | 207  |
| ภาพที่ 9-15 แสดงตัวอย่างการทำงานของเทคนิคบูรณาการข้อมูล                                                                                                                                              | 209  |

# สารบัญตาราง

|              |                                                                                                                                                                                                       | หน้า |
|--------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| ตารางที่ 2-1 | ศัพท์ที่ใช้ในกระบวนการค้นพบความรู้จากฐานข้อมูล                                                                                                                                                        | 21   |
| ตารางที่ 3-1 | เปรียบเทียบลักษณะค่าแอตทริบิวต์ทั้ง 4 ระดับ                                                                                                                                                           | 35   |
| ตารางที่ 3-2 | แสดงตัวอย่างการใช้งานฟังก์ชันในการแปลงค่าแอตทริบิวต์ทั้ง 4 ระดับ                                                                                                                                      | 36   |
| ตารางที่ 3-3 | แสดงตัวอย่างการจัดเก็บเป็นข้อมูลระเบียบ                                                                                                                                                               | 37   |
| ตารางที่ 3-4 | แสดงตัวอย่างของเมทริกซ์ข้อมูลการจัดเก็บค่าอุณหภูมิเป็นองศาเซลเซียส<br>ทุก 4 ชั่วโมงในแต่ละวันเป็นจำนวน 4 วัน                                                                                          | 38   |
| ตารางที่ 3-5 | แสดงตัวอย่างเวกเตอร์แสดงความรู้การปรากฏชื่อนายกรัฐมนตรี<br>ในบทความทางการเมือง 5 ชิ้น                                                                                                                 | 38   |
| ตารางที่ 3-6 | แสดงการซื้อสินค้าแต่ละตะกร้าสินค้าของร้าน ABC                                                                                                                                                         | 39   |
| ตารางที่ 3-7 | แสดงพฤติกรรมกรซื้อสินค้าร้าน ABC ของลูกค้า 4 คนที่ซื้อใน 7 วันล่าสุด                                                                                                                                  | 40   |
| ตารางที่ 4-1 | แสดงข้อมูลการเล่นเทนนิสตามสภาพภูมิอากาศของนักเทนนิสผู้หนึ่งเป็นเวลา 14 วัน                                                                                                                            | 82   |
| ตารางที่ 4-2 | แสดงชุดข้อมูลสิ่งมีชีวิต                                                                                                                                                                              | 85   |
| ตารางที่ 4-3 | แสดงตัวอย่างข้อมูลสิ่งมีชีวิต                                                                                                                                                                         | 87   |
| ตารางที่ 4-4 | แสดงตัวอย่างข้อมูลสิ่งมีชีวิต                                                                                                                                                                         | 87   |
| ตารางที่ 4-5 | แสดงค่าความคุ้มค่าและความแม่นยำของกฎตัวอย่างที่ได้จากชุดข้อมูลในตารางที่ 4-2                                                                                                                          | 88   |
| ตารางที่ 4-6 | แสดงลักษณะของเมทริกซ์ความสับสน (Confusion Matrix)                                                                                                                                                     | 90   |
| ตารางที่ 4-7 | แสดงตัวอย่างเมทริกซ์ความสับสน (Confusion Matrix) สำหรับฝึกคำนวณ                                                                                                                                       | 92   |
| ตารางที่ 5-1 | แสดงตัวอย่างชุดข้อมูลที่มีปะปนกันของฟีเจอร์เชิงตัวเลขและเชิงประเภ                                                                                                                                     | 100  |
| ตารางที่ 5-2 | แสดงตัวอย่างชุดข้อมูลจากตารางที่ 5-1 ที่ปรับช่วงค่าฟีเจอร์เชิงตัวเลข<br>ด้วยการทำให้ชุดข้อมูลเป็นมาตรฐาน (Data Standardization) โดยตัดฟีเจอร์<br>“รหัสลูกค้า” ที่ไม่มีประโยชน์ต่อการจับกลุ่มข้อมูลออก | 102  |
| ตารางที่ 5-3 | แสดงตัวอย่างชุดข้อมูลจากตารางที่ 5-2 ที่เลือกเฉพาะฟีเจอร์เชิงประเภ<br>ที่แปลงเป็นฟีเจอร์ทวิภาค                                                                                                        | 104  |
| ตารางที่ 5-4 | แสดงรายละเอียดนิสิตเก่าของมหาวิทยาลัยแห่งหนึ่ง                                                                                                                                                        | 115  |
| ตารางที่ 5-5 | แสดงตารางความสัมพันธ์ทั่วไปของชุดข้อมูลในตารางที่ 5-4                                                                                                                                                 | 116  |
| ตารางที่ 5-6 | แสดงตารางไขว้ของคู่ฟีเจอร์เพศและสถานที่เกิดของชุดข้อมูลในตารางที่ 5-5                                                                                                                                 | 117  |
| ตารางที่ 5-7 | แสดงชุดข้อมูล D : การสั่งซื้ออาหาร                                                                                                                                                                    | 127  |
| ตารางที่ 5-8 | แสดง Candidate $C_1$                                                                                                                                                                                  | 128  |

|                                                                                                                                                                                           | หน้า |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| ตารางที่ 5-9 แสดง Large Itemset $L_1$                                                                                                                                                     | 129  |
| ตารางที่ 5-10 แสดงไอเทมเซตที่ค้นพบย่อยทั้งหมดจากตัวอย่างในตารางที่ 5-7                                                                                                                    | 129  |
| ตารางที่ 6-1 แสดงความสอดคล้องกระบวนการทำเหมืองข้อมูลแบบ SEMMA และ CRISP กับขั้นตอนทำเหมืองข้อมูลในภาพรวมกระบวนการค้นพบความรู้                                                             | 138  |
| ตารางที่ 7-1 แสดงการเปรียบเทียบความเหมาะสมต่อการนำไปประยุกต์ใช้งานเหมืองข้อมูลของชุดข้อมูลเชิงข้อเท็จจริงและชุดข้อมูลเชิงพฤติกรรมในประเด็นความแม่นยำของแบบจำลอง (Accuracy)                | 149  |
| ตารางที่ 7-2 แสดงการเปรียบเทียบความเหมาะสมต่อการนำไปประยุกต์ใช้งานเหมืองข้อมูลของชุดข้อมูลเชิงข้อเท็จจริงและชุดข้อมูลเชิงพฤติกรรมในประเด็นความต้องกันของชุดข้อมูลที่ใช่ (Consistency)     | 150  |
| ตารางที่ 7-3 แสดงการเปรียบเทียบความเหมาะสมต่อการนำไปประยุกต์ใช้งานเหมืองข้อมูลของชุดข้อมูลเชิงข้อเท็จจริงและชุดข้อมูลเชิงพฤติกรรมในประเด็นความทันเวลาของการได้มาซึ่งแบบจำลอง (Timeliness) | 151  |
| ตารางที่ 7-4 แสดงการเปรียบเทียบความเหมาะสมต่อการนำไปประยุกต์ใช้งานเหมืองข้อมูลของชุดข้อมูลเชิงข้อเท็จจริงและชุดข้อมูลเชิงพฤติกรรมในประเด็นความน่าเชื่อถือของแบบจำลอง (Believability)      | 152  |
| ตารางที่ 7-5 แสดงการเปรียบเทียบความเหมาะสมต่อการนำไปประยุกต์ใช้งานเหมืองข้อมูลของชุดข้อมูลเชิงข้อเท็จจริงและชุดข้อมูลเชิงพฤติกรรมในประเด็นความสามารถตีความของแบบจำลอง (Interpretability)  | 152  |
| ตารางที่ 7-6 แสดงการเปรียบเทียบความเหมาะสมต่อการนำไปประยุกต์ใช้งานเหมืองข้อมูลของชุดข้อมูลเชิงข้อเท็จจริงและชุดข้อมูลเชิงพฤติกรรมในประเด็นความทนทานต่อการใช้งานของแบบจำลอง (Robustness)   | 153  |
| ตารางที่ 9-1 แสดงลักษณะทั่วไปของพันธุ์ข้าว Cammeo และ Osmanic จากเทคนิคการจับกลุ่มข้อมูล                                                                                                  | 186  |
| ตารางที่ 9-2 แสดงกลุ่ม cluster ของชุดข้อมูลข้าวที่คัดกรองพีเอชอาร์ก่อนใช้เทคนิค Expectation-Maximization รวมถึงอัตราส่วนและเปอร์เซ็นต์เจือปนพันธุ์ข้าวในแต่ละกลุ่มย่อยนั้น                | 193  |





เหมืองข้อมูลเป็นศาสตร์ที่มีการประยุกต์ใช้ในประเทศไทยมากกว่า 20 ปีแล้ว แต่กล่าวได้ว่าคนที่รู้จักศาสตร์ด้านเหมืองข้อมูลยังมีไม่มากนัก และในกลุ่มคนที่รู้จักเหมืองข้อมูลเองก็ไม่น้อยที่ใช้งานเหมืองข้อมูลเป็นเพียงเครื่องมือทดลองใช้กับชุดข้อมูลที่มีอยู่ว่าจะได้ผลลัพธ์ที่น่าสนใจเกิดขึ้นหรือไม่เพียงเท่านั้น ซึ่งโดยศักยภาพของศาสตร์เหมืองข้อมูลเองมีความสามารถมากกว่านั้นอย่างมาก เนื่องจากเหมืองข้อมูลเป็นส่วนสำคัญหนึ่งในกระบวนการค้นพบความรู้จากชุดข้อมูลหรือฐานข้อมูลขนาดใหญ่

ปัจจุบันการบันทึกข้อมูลทางวิทยาศาสตร์ การบันทึกข้อมูลทางสังคมศาสตร์มีพร้อมอยู่เป็นจำนวนมาก โดยภายหลังสถานการณ์การแพร่กระจายของโควิด-19 กล่าวได้ว่ากิจกรรมจำนวนมากของมนุษย์มีการโอนย้ายมายังแพลตฟอร์มที่ดำเนินการผ่านอินเทอร์เน็ต ในกิจกรรมที่สำคัญ เช่น การเงิน การซื้อขาย การทำธุรกรรม ฯลฯ ล้วนมีการบันทึกเก็บเอาไว้ ทำให้มีชุดข้อมูลขนาดใหญ่ที่ต้องการเครื่องมือที่เหมาะสมในการค้นพบสิ่งที่น่าสนใจหรือความรู้ที่แฝงอยู่ ซึ่งเครื่องมือหนึ่งที่สามารถตอบสนองความต้องการดังกล่าวได้ก็คือเหมืองข้อมูล

นอกจากนี้ ความก้าวหน้าทางเทคโนโลยีการสื่อสารที่รวดเร็ว ครอบคลุม ทำให้เกิดการประยุกต์ใช้มากมายในเทคโนโลยีเครือข่าย เช่น การคำนวณแบบคลาวด์ (Cloud Computing) หรืออินเทอร์เน็ตประสาทรพสิ่ง (Internet of Things, IoT) ทำให้มีการไหลของข้อมูลผ่านเครือข่ายเพื่อบันทึกและควบคุมใช้งานอยู่เป็นจำนวนมาก ซึ่งการประยุกต์ใช้อย่างมีประสิทธิภาพต้องการเครื่องมือในการวิเคราะห์ข้อมูลที่เกิดขึ้นและต้องการเครื่องมือสำหรับอำนวยความสะดวกเพื่องานควบคุมผ่านเครือข่ายเหล่านี้

การตอบโจทย์ความต้องการวิเคราะห์ข้อมูลและเครื่องมืออำนวยความสะดวกนี้ทำให้เกิดความนิยมในศาสตร์เกิดใหม่อย่างวิทยาการข้อมูล (Data Science) ที่มีชื่อศาสตร์เข้าใจง่ายและตอบโจทย์ตรงตัวกับความต้องการวิเคราะห์ข้อมูล นอกจากนี้ยังมีการกลับมาได้รับความนิยมของปัญญาประดิษฐ์ (Artificial Intelligence, AI) ซึ่งเป็นที่ยอมรับและรู้จักในประเทศไทยมาอย่างยาวนาน

อย่างไรก็ตาม ผู้เขียนมีความเห็นว่าการตอบโจทย์ความต้องการวิเคราะห์ข้อมูลและเครื่องมืออำนวยความสะดวกที่เกิดขึ้นนี้ เป้าหมายสำคัญยังคงเป็นการค้นพบความรู้ที่แฝงอยู่ในชุดข้อมูล ซึ่งศาสตร์เหมืองข้อมูล

เป็นศาสตร์ที่มีกรอบคิดและรากฐานมั่นคงผูกติดอยู่กับกระบวนการค้นพบความรู้ โดยหากผู้ต้องการตอบโจทย์วิเคราะห์ข้อมูลได้ทำความเข้าใจเหมืองข้อมูลมากขึ้นแล้ว ย่อมทำให้มีศักยภาพมากขึ้นในการค้นพบความรู้ที่แฝงอยู่ในชุดข้อมูล ซึ่งก็คือเนื้อหาและวัตถุประสงค์หลักของหนังสือเล่มนี้ที่จะได้นำเสนอให้ผู้อ่านเรียนรู้ตามลำดับ

ในบทนำนี้จะได้นำเสนอเหมืองข้อมูลให้ผู้อ่านรู้จักศาสตร์เหมืองข้อมูลยิ่งขึ้น รวมถึงมองเห็นศักยภาพในการนำไปประยุกต์ใช้กับความต้องการวิเคราะห์ข้อมูลด้านต่าง ๆ ของมนุษย์ ซึ่งรายละเอียดเชิงเทคนิคพื้นฐานของเหมืองข้อมูล การประยุกต์ และงานบูรณาการเหมืองข้อมูลจะได้นำเสนอในบทถัดไปตามลำดับ

## 1.1 ความเป็นมาของเหมืองข้อมูล

คำว่า “เหมืองข้อมูล” ปรากฏขึ้นครั้งแรกที่ผู้เขียนพบก็คือในบทความเรื่อง Data Mining ในวารสาร The Review of Economics and Statistics ฉบับเดือนกุมภาพันธ์ ค.ศ. 1983 เขียนโดย Michael C. Lovell ซึ่งไม่ได้ให้ความหมายเป็นนิยามที่ชัดเจน แต่กล่าวถึงความต้องการวิเคราะห์ข้อมูลเพื่อค้นพบความรู้ที่แฝงอยู่ในชุดข้อมูลทางเศรษฐศาสตร์ โดยเสนอกระบวนทัศน์ (Paradigm) ของการใช้ประโยชน์จากสิ่งที่เรียกว่าธนาคารข้อมูล (Data Bank) ที่ใช้กับข้อมูลทางเศรษฐศาสตร์ จากนั้นใช้เทคนิคทางสถิติค้นพบแบบจำลอง (Model) ที่แฝงอยู่เป็น “การเพิ่มความรู้จากเดิมที่ทำให้ทราบว่าเศรษฐกิจมีการทำงานแฝงอยู่อย่างไร” หรือ “a proportional increase in our knowledge of how the economy actually works” (Lovell, 1983)

แนวคิดกระบวนการในการค้นพบความรู้จากชุดข้อมูลถือเป็นอุดมการณ์หลักของการเกิดศาสตร์ที่เรียกว่าเหมืองข้อมูล อย่างไรก็ตาม ในการก่อรูปศาสตร์เหมืองข้อมูลจากวิธีการทางสถิติในงานด้านเศรษฐศาสตร์ระยะแรกแม้จะบ่งชี้ความต้องการได้ชัดเจน แต่รูปธรรมของวิธีการและผลลัพธ์จากเหมืองข้อมูลยังไม่สามารถระบุได้ชัดเจนนักโดยเฉพาะอย่างยิ่งในการประยุกต์ใช้ในงานด้านวิทยาการคอมพิวเตอร์

แนวคิด นิยาม วิธีการและผลลัพธ์ของเหมืองข้อมูลในงานด้านวิทยาการคอมพิวเตอร์เริ่มพัฒนาขึ้นมาเป็นรูปธรรมที่ชัดเจนในช่วงต้นทศวรรษ 1990 (Piatetsky-Shapiro, 1990) ที่ Gregory Piatetsky-Shapiro เสนอแนวคิดกระบวนการ “ค้นพบความรู้จากฐานข้อมูล” หรือ “Knowledge Discovery from Database” พร้อม ๆ กับการเริ่มปรากฏของนิยามเหมืองข้อมูลที่ชัดเจน โดยยังคงมีผู้ให้นิยามใหม่ในศาสตร์เหมืองข้อมูลจนถึงปัจจุบัน การปรับตัวของนิยามความหมายของเหมืองข้อมูลบ่งชี้ว่าศาสตร์เหมืองข้อมูลยังมีวิวัฒนาการได้อีกมาก เป็นศาสตร์ที่ยังไม่ตายและพร้อมปรับเปลี่ยนตามยุคสมัย ซึ่งเราสามารถศึกษา นิยามที่หลากหลายนี้ได้ในหัวข้อถัดไป

หลายเทคนิคที่ใช้ในเหมืองข้อมูลเป็นเทคนิคที่พัฒนาก่อนมีเหมืองข้อมูล เช่น การจำแนกประเภทข้อมูล (Data Classification) และการจับกลุ่มข้อมูล (Data Clustering) แต่เทคนิคเหล่านี้เมื่อนำมาใช้ด้วยอุดมการณ์เหมืองข้อมูลคือ “การค้นพบความรู้จากชุดข้อมูล” ซึ่งมีลักษณะสหวิทยาการ ทำให้ศาสตร์เหมืองข้อมูลสามารถหลอมรวมหลายเทคนิคที่พัฒนาขึ้นก่อนหน้านี้รวมเป็นเทคนิคที่ใช้งานในเหมืองข้อมูลได้

อย่างไรก็ตาม แม้อุดมการณ์ของศาสตร์เหมืองข้อมูลคือการค้นพบความรู้จากชุดข้อมูล แต่กรอบการทำงานของเหมืองข้อมูลเป็นเพียงส่วนหนึ่งของกระบวนการค้นพบความรู้จากชุดข้อมูลเท่านั้น ผลลัพธ์ของการทำเหมืองข้อมูลยังมีใช้ความรู้ เป็นเพียง “แพตเทิร์นที่น่าสนใจ” หรือ “Interesting Patterns” (Han, Pei, & Tong, 2022) สำหรับตรวจสอบ ดีความ และสรุปผลลัพธ์ที่นำไปสู่ความรู้ตามขั้นตอนกระบวนการค้นพบความรู้ในท้ายสุด

ทั้งนี้ เราจะได้ทำความเข้าใจเหมืองข้อมูลให้มากขึ้นผ่านหลากหลายนิยามเหมืองข้อมูลดังแสดงในหัวข้อถัดไป

## 1.2 นิยามเหมืองข้อมูล

นิยามเป็นเครื่องมือสำคัญสำหรับผู้เริ่มต้นศึกษาในศาสตร์ใดก็ตามเพื่อใช้สำรวจว่าตนศึกษาครบถ้วนแล้วหรือไม่ หรือเพื่อไม่ให้เรียนรู้อย่างความออกนอกขอบเขตไปไกลเกินไป การเลือกนิยามมาใช้เป็นเครื่องนำทางการเรียนรู้จึงเป็นเรื่องสำคัญ

สำหรับเหมืองข้อมูลที่เป็นศาสตร์ใหม่ นอกจากการเลือกนิยามที่เป็นมาตรฐานตั้งแต่เริ่มต้นก่อรูปศาสตร์เหมืองข้อมูลเพื่อเป็นเครื่องมือนำทางการเรียนรู้แล้ว การกำหนดนิยามใหม่ยังเป็นเครื่องมือปรับตัวการวิจัยเรียนรู้ในศาสตร์เหมืองข้อมูลให้มีความก้าวหน้าทันยุคสมัยมากยิ่งขึ้น ความใหม่ของศาสตร์เหมืองข้อมูลทำให้มีโอกาสพัฒนาความเป็นศาสตร์อีกมาก เราจึงพบการกำหนดนิยามใหม่ให้กับเหมืองข้อมูลอยู่เสมอจนถึงปัจจุบัน

ในหัวข้อนี้ เราจะได้ยกตัวอย่างนิยามเหมืองข้อมูลที่มีลักษณะมาตรฐาน นิยามที่มีความครอบคลุม และนิยามที่กำหนดองค์ประกอบให้มีความก้าวหน้าตามยุคสมัยดังนี้

หนังสือ Data Mining: Concepts and Techniques (Han et al., 2022) ที่มีผู้นิยมอย่างกว้างขวางได้ให้นิยามเหมืองข้อมูลที่ปรับปรุงในฉบับปรับปรุงครั้งที่ 4 นี้ว่า “เหมืองข้อมูลคือกระบวนการค้นพบแพตเทิร์นที่น่าสนใจ, แบบจำลอง และความรู้ลักษณะอื่นจากชุดข้อมูลขนาดใหญ่” หรือ “Essentially, data mining is the process of discovering interesting patterns, models, and other kinds of knowledge in large data sets.” โดยมีความเห็นเพิ่มเติมว่า เหมืองข้อมูลควรใช้ชื่อใหม่ที่เหมาะสมว่าเหมืองความรู้จากข้อมูล แต่ชื่อนี้เป็นคำที่ยาวเกินไป หรือ “Data mining should have been more

appropriately named ‘knowledge mining from data,’ which is unfortunately somewhat long.” (Han et al., 2022)

ในสารานุกรมบริแทนนิกา หรือ Encyclopedia Britannica (Clifton, 2010) ให้ความหมายของเหมืองข้อมูลไว้ว่า “เหมืองข้อมูลคือกระบวนการค้นพบแพตเทิร์นและความสัมพันธ์ที่น่าสนใจและมีประโยชน์จากชุดข้อมูลขนาดใหญ่ เป็นศาสตร์ที่รวมเครื่องมือจากศาสตร์ทางด้านสถิติและปัญญาประดิษฐ์ (เช่น โครงข่ายประสาทเทียมและเครื่องจักรเรียนรู้) ในการประยุกต์ใช้งานกับการจัดการฐานข้อมูลเพื่อวิเคราะห์ชุดข้อมูลดิจิทัลขนาดใหญ่” หรือ “the process of discovering interesting and useful patterns and relationships in large volumes of data. The field combines tools from statistics and artificial intelligence (such as neural networks and machine learning) with database management to analyze large digital collections, known as data sets.” (Clifton, 2010)

บริษัทไอบีเอ็ม (IBM) ที่เป็นหนึ่งในผู้บุกเบิกเหมืองข้อมูล ได้ให้คำนิยามเหมืองข้อมูลล่าสุดใน ค.ศ. 2021 (IBM Cloud Education, 2021) ไว้ว่า “เหมืองข้อมูลหรือที่รู้จักกันอีกชื่อหนึ่งว่าการค้นพบความรู้จากชุดข้อมูล เป็นกระบวนการเปิดเผยแพตเทิร์นและสารสนเทศที่มีค่าที่แฝงอยู่ในชุดข้อมูลขนาดใหญ่” หรือ “Data mining, also known as knowledge discovery in data (KDD), is the process of uncovering patterns and other valuable information from large data sets.” (IBM Cloud Education, 2021)

จากตัวอย่างนิยามเหล่านี้ สามารถอธิบายความหมายของเหมืองข้อมูลที่ตรงกันได้ว่า “เป็นกระบวนการสำหรับค้นพบแพตเทิร์นที่น่าสนใจและมีประโยชน์จากข้อมูลขนาดใหญ่” ในบางนิยามเน้นไปที่ความสามารถประยุกต์ใช้งานจากแหล่งข้อมูลหลากหลายลักษณะ บางนิยามเน้นไปที่เทคนิคขั้นตอนวิธีแบบสหวิทยาการ บางความหมายให้ความสำคัญกับหาแพตเทิร์นที่เรียกว่าแบบจำลองในการอธิบายสิ่งที่สนใจที่แฝงอยู่ในชุดข้อมูล และบางนิยามเน้นที่การเป็นส่วนหนึ่งของกระบวนการค้นพบความรู้จากข้อมูลขนาดใหญ่

สำหรับผู้เขียนที่สอนและวิจัยงานเหมืองข้อมูลมากกว่า 20 ปี ล้วนประยุกต์ใช้และบูรณาการเทคนิคทางด้านเหมืองข้อมูลกับชุดข้อมูลที่หลากหลายขนาดเพื่อสร้างเครื่องมือเหมืองข้อมูลใหม่ ๆ อย่าง ต้นไม้ กฎความสัมพันธ์, พื้นที่ความน่าสนใจ, ความลาดชันของความน่าสนใจ (Liewlom, 2021a), ผังก้างปลา บูรณาการเหมืองข้อมูล (Sangsuriyun, Liewlom, Tangsakul, & Suchaiya, 2022) และป่าบรรยาย (Liewlom, 2023)

ในความเห็นของผู้เขียน เหมืองข้อมูลจึงไม่จำเป็นต้องประยุกต์ใช้กับชุดข้อมูลขนาดใหญ่เท่านั้น แต่มีขนาดของชุดข้อมูลที่พอเพียงให้ค้นพบสิ่งที่น่าสนใจที่แฝงอยู่ เปรียบเสมือนกับการทำเหมืองเพชรพลอยขนาดใหญ่ในพื้นที่ภูเขาหรือเหมืองบ่อพลอยพื้นบ้านในขนาดพื้นที่ที่กว้างยาวไม่กี่เมตร ก็ล้วนเป็นการทำเหมือง

ทั้งสิ้น หากการทำเหมืองนั้นเรารู้ว่าทำบนพื้นที่ที่มีศักยภาพที่จะค้นพบผลลัพธ์ที่มีค่า ความเห็นของผู้เขียนนี้สอดคล้องกับคำนิยามชิ้นใหม่ของ Jiawei Han et al. (2022) ที่แม้จะยังคงเน้นที่ชุดข้อมูลขนาดใหญ่ แต่ได้ให้ความเห็นเพิ่มเติมเกี่ยวกับชื่อเหมืองข้อมูล (Data Mining) ที่ควรใช้ชื่อ “knowledge mining from data” ซึ่งการกำกับวัตถุประสงค์ “เพื่อการค้นพบความรู้” เปิดทางให้เกิดการประยุกต์ใช้เหมืองข้อมูลกับชุดข้อมูลทั่วไปที่ไม่จำเป็นต้องเป็นชุดข้อมูลขนาดใหญ่ด้วย

ตัวอย่างของชุดข้อมูลขนาดใหญ่สามารถค้นพบความรู้ที่มีค่าที่แฝงอยู่ก็คือชุดข้อมูลทางวิทยาศาสตร์ แพตเทิร์นที่แฝงอยู่ในชุดข้อมูลทางวิทยาศาสตร์แม้ในขนาดที่ไม่ใหญ่มากมักเป็นแพตเทิร์นที่มีค่าหรือมีศักยภาพในการค้นพบความรู้ทางวิทยาศาสตร์ สามารถตรวจสอบและติดตามในระยะยาวเพื่อค้นพบความรู้ที่แฝงอยู่ ยกตัวอย่างเช่น ชุดข้อมูลของผู้ป่วยโรคที่หายาก ทั่วโลกอาจมีไม่ถึง 300 คน แต่ชุดข้อมูลจำนวนนี้ย่อมเป็นชุดข้อมูลที่มีค่าอย่างยิ่งในการค้นพบความรู้ที่เกี่ยวกับโรคดังกล่าว โดยเราสามารถเลือกใช้เทคนิคเหมืองข้อมูลในการนำมาประยุกต์ใช้กับชุดข้อมูลเพื่อค้นพบสิ่งมีค่าที่แฝงอยู่นี้

นอกจากนี้ ในงานวิจัยด้านเหมืองข้อมูลซึ่งสามารถสืบค้นได้ทั้งจาก scopus.com และ scholar.google.com มีงานวิจัยจำนวนมากโดยเฉพาะอย่างยิ่งการนำเสนอเครื่องมือใหม่ ๆ ในการทำเหมืองข้อมูล มีการใช้ขนาดชุดข้อมูลที่ไม่ใหญ่มากนัก และแม้แต่เฟรมเวิร์กด้านเหมืองข้อมูลที่มีชื่อเสียงอย่างซอฟต์แวร์ WEKA (Frank, Hall, & Witten, 2016) ก็มีข้อจำกัดทำงานได้กับชุดข้อมูลขนาดเล็กเท่านั้นเนื่องจากการประมวลผลทั้งหมดต้องอยู่ในหน่วยความจำ เป็นความจริงเชิงประจักษ์ที่ศาสตร์เหมืองข้อมูลไม่จำเป็นต้องผูกติดอยู่กับเฉพาะชุดข้อมูลขนาดใหญ่

อย่างไรก็ตาม เป็นความจริงอีกเช่นกันที่ผลลัพธ์ของการทำเหมืองข้อมูลคือ “สิ่งมีค่า” ดังนั้นจึงมีความต้องการใช้งานซอฟต์แวร์ด้านเหมืองข้อมูลเป็นตลาดซอฟต์แวร์ที่มีมูลค่ามหาศาล โดยเฉพาะอย่างยิ่งกับชุดข้อมูลขนาดใหญ่ที่จัดเก็บและจัดการโดยซอฟต์แวร์ฐานข้อมูลที่เดิมทีก็มีมูลค่ามหาศาลอยู่แล้ว ซอฟต์แวร์เหมืองข้อมูลที่มีศักยภาพทำงานกับชุดข้อมูลขนาดใหญ่มาได้จึงเป็นตลาดซอฟต์แวร์ระดับพรีเมียม

สำหรับผู้เขียนที่ทำงานประยุกต์เหมืองข้อมูลอย่างอิสระ ย่อมไม่เห็นด้วยว่าเหมืองข้อมูลเป็นเทคนิคที่เหมาะสมกับองค์กรขนาดใหญ่ที่มีฐานข้อมูลขนาดใหญ่เท่านั้น หากบริษัทหรือองค์กรขนาดเล็กสามารถใช้ซอฟต์แวร์ทั่วไปหรือเฟรมเวิร์กเพื่อประยุกต์ดำเนินการเหมืองข้อมูลกับชุดข้อมูลที่ไม่ใหญ่มากของตนได้ ย่อมเป็นการเปิดโอกาสให้เกิดการประยุกต์ใช้งานเหมืองข้อมูลได้กว้างขวางขึ้น ผู้เขียนจึงกำหนดนิยามเหมืองข้อมูลเพื่อส่งเสริมให้เกิดงานวิจัยและประยุกต์ใช้งานด้านเหมืองข้อมูลที่กว้างขวางยิ่งขึ้นดังนี้

“เหมืองข้อมูล เป็นกระบวนการค้นพบแพตเทิร์นที่น่าสนใจจากชุดข้อมูลที่มีศักยภาพและรู้ว่ามีค่าแฝงอยู่ โดยผลลัพธ์เหมืองข้อมูลนี้เป็นสิ่งที่มีค่าในกระบวนการค้นพบความรู้” โดยนิยามนี้ผู้เขียนขยายความจากอีกชื่อหนึ่งของเหมืองข้อมูลคือ “การทำเหมืองความรู้จากข้อมูล” (knowledge mining from data) ของ Jiawei Han et al. (2022) ที่เป็นการปรับปรุงเกี่ยวกับนิยามเหมืองข้อมูลล่าสุด

นิยามของผู้เขียนยังคงยึดตามอุดมการณ์หลักของเหมืองข้อมูลที่เป็น “กระบวนการ” ที่เป็นส่วนหนึ่งของ “การค้นพบความรู้จากชุดข้อมูล” ที่ก่อรูปศาสตร์เหมืองข้อมูลในระยะแรกที่เน้นการค้นพบความรู้จากฐานข้อมูลโดยไม่จำเป็นต้องเป็นฐานข้อมูลขนาดใหญ่ เป็นนิยามที่ไม่ขัดกับนิยามเหมืองข้อมูลก่อนหน้านี้ เนื่องจากชุดข้อมูลที่มีศักยภาพและรู้ว่ามีความรู้แฝงอยู่มีความครอบคลุมกว้างกว่าคำว่าชุดข้อมูลขนาดใหญ่ นิยามของผู้เขียนนี้เป็นนิยามที่เปิดทางให้การประยุกต์ศาสตร์ด้านเหมืองข้อมูลเพื่อใช้กับชุดข้อมูลขนาดใหญ่ไม่ใหญ่มาก แต่ต้องการเครื่องมือค้นพบแพตเทิร์นที่มีค่าหรือความรู้ที่แฝงอยู่ ซึ่งในประเทศไทยมีชุดข้อมูลลักษณะนี้อยู่เป็นจำนวนมากเช่นในองค์กรขนาด SMEs

ในการเขียนหนังสือเล่มนี้ ยึดตามนิยามของผู้เขียนเป็นแนวทางหลักสำหรับอธิบายและสรุปเป็นผังแนวทางสำหรับการประยุกต์และบูรณาการงานเหมืองข้อมูล รายละเอียดดังนำเสนอในบทที่ 7 และ 8 โดยยังคงมีความสอดคล้องถูกต้องกับนิยามเหมืองข้อมูลมาตรฐาน

### 1.3 ลักษณะของชุดข้อมูลที่สามารถนำมาใช้งานเหมืองข้อมูล

เราสามารถพบเห็นชุดข้อมูลที่มีลักษณะหลากหลายเพื่อใช้งานเหมืองข้อมูล เช่น ฐานข้อมูล, ไฟล์, ภาพ, เสียง, กระแสข้อมูล (Streaming Data), แผนที่ ฯลฯ โดยหากเรามีหรือเป็นเจ้าของชุดข้อมูลใด ๆ และมีความต้องการประยุกต์ใช้เทคนิคเหมืองข้อมูล เราก็ควรทราบว่าชุดข้อมูลที่เราอยู่นี้สามารถประยุกต์ใช้กับงานเหมืองข้อมูลได้หรือไม่

ชุดข้อมูลที่สามารถใช้งานเหมืองข้อมูลได้ ก็คือชุดข้อมูลที่ตอบโจทย์งานเหมืองข้อมูล นั่นคือเป็นชุดข้อมูลที่มีศักยภาพ (ในการค้นพบแพตเทิร์น) และรู้ว่ามีความรู้แฝงอยู่ ดังนั้นชุดข้อมูลที่มีลักษณะดังกล่าวย่อมมีวิธีประยุกต์ใช้กับเหมืองข้อมูลโดยไม่จำกัดว่าต้องอยู่ในรูปใด เป็นไฟล์หรือเป็นกระแสข้อมูล (Streaming Data) เป็นตัวอักษรหรือรูปภาพหรือเสียง จัดเก็บเป็นฐานข้อมูลหรือจัดเก็บแบบกึ่งโครงสร้าง เป็นข้อมูลเว็บ ฯลฯ

ชุดข้อมูลที่ไม่สามารถทำเหมืองข้อมูลได้ก็คือชุดข้อมูลที่ไม่มีศักยภาพในการค้นพบแพตเทิร์นที่แฝงอยู่ ยกตัวอย่างเช่น ชุดข้อมูลที่มีขนาดเล็กเกินไปจนสามารถมองเห็นทุกแพตเทิร์นที่มีอยู่ในชุดข้อมูลได้ชัดเจนโดยไม่ต้องใช้เครื่องมือใด หรือชุดข้อมูลที่ไม่สามารถพบแพตเทิร์นได้ เช่น ไฟล์ที่มีเพียงรหัสสินค้าที่สร้างอัตโนมัติด้วยระบบฐานข้อมูล ชุดข้อมูลลักษณะนี้มีไว้เพื่อให้มีรหัสและจำแนกสินค้าโดยไม่ให้มีการซ้ำซ้อน ไม่ได้มีไว้สำหรับวิเคราะห์ข้อมูล ไม่มีแพตเทิร์นแฝงอยู่ หรือหากมีแพตเทิร์นแฝงอยู่ก็เป็นแพตเทิร์นที่ไม่มีค่า เช่น รหัสสินค้า A000-A099 จำนวน 100 ระเบียบ มีรหัสที่ประกอบด้วยเลข 9 อยู่ทั้งหมด 18 ระเบียบ แต่ไม่ได้หมายความว่าสินค้าที่มีรหัส A099 มีความเหมือนหรือคล้ายคลึงกับสินค้านี้ A098 มากกว่าสินค้านี้ A009 รวมถึงไม่ได้มีความหมายว่าสินค้านี้รหัสเลข 9 เหล่านี้จะมีความสัมพันธ์กันเป็นแพตเทิร์นที่มีค่า

ชุดข้อมูลที่มีศักยภาพในการค้นพบแพตเทิร์นและรู้ว่ามีข้อมูลแฝงอยู่ สามารถประยุกต์ใช้กับเทคนิคเหมืองข้อมูลได้อย่างหลากหลายวัตถุประสงค์งาน ดังจะเห็นตัวอย่างในหัวข้อถัดไป

## 1.4 ศักยภาพในการใช้งานเหมืองข้อมูลในประเทศไทย

ในหัวข้อนี้เราจะได้เข้าใจการนำเหมืองข้อมูลมาประยุกต์ใช้งานให้มากขึ้นผ่านความเป็นมาของการประยุกต์ใช้งานเหมืองข้อมูลในประเทศไทย โดยผู้เขียนทำการสำรวจแนวโน้มงานวิจัยเหมืองข้อมูลทั้งในระดับสากลและในประเทศไทย จากนั้นประมวลสรุปลักษณะการประยุกต์ใช้เหมืองข้อมูลในประเทศไทยนำเสนอตามลำดับดังนี้

### 1.4.1 ความเป็นมาของเหมืองข้อมูลในประเทศไทย

เหมืองข้อมูลถือเป็นศาสตร์ใหม่อย่างยิ่งเมื่อเริ่มเข้ามาผลักดันใช้งานในประเทศไทย กลุ่มผู้ที่นำมาใช้จึงเป็นกลุ่มนักวิจัยจากมหาวิทยาลัยเป็นหลัก โดยมีการเริ่มต้นงานในระยะเวลาที่ใกล้เคียงกันมากจาก 4 มหาวิทยาลัย คือ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี, มหาวิทยาลัยมหิดล, จุฬาลงกรณ์มหาวิทยาลัย และมหาวิทยาลัยเกษตรศาสตร์ ในช่วง พ.ศ. 2543 และ พ.ศ. 2544

พ.ศ. 2543 เหมืองข้อมูลปรากฏครั้งแรกในบทความวิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของ นางสาวรัตนา จรัสสุวัฒน์ (รัตนา จรัสสุวัฒน์, 2543) โดยมีอาจารย์ที่ปรึกษา คือ ดร.กิตติชัย ล้วนยานนท์ จากมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี คณะเทคโนโลยีสารสนเทศ ในหัวข้อเรื่อง Synergistic Genetic Algorithms for Inductive Learning (SynGAIL) โดยอ้างถึงการประยุกต์ใช้เทคนิคจากศาสตร์เหมืองข้อมูลปรากฏขึ้นเป็นครั้งแรก

พ.ศ. 2544 วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์ 2 ฉบับแรกที่มีคำว่าเหมืองข้อมูลปรากฏในชื่อวิทยานิพนธ์ก็คือ วิทยานิพนธ์เรื่อง การทำเหมืองข้อมูลบนเว็บ (สุทธศักดิ์ อินทวดี, 2544) ของ นายสุทธศักดิ์ อินทวดี มีอาจารย์ที่ปรึกษา คือ ดร.ดำรง วงศ์สว่าง จากมหาวิทยาลัยมหิดล และวิทยานิพนธ์เรื่อง การออกแบบและพัฒนาขั้นตอนวิธีการวิเคราะห์บันทึกการเข้าใช้โดยใช้เทคนิคการทำเหมืองข้อมูล ของ นายวิญญู ละอองสุวรรณ มีอาจารย์ที่ปรึกษา คือ ผศ. ดร.บุญเสริม กิจศิริกุล จากจุฬาลงกรณ์มหาวิทยาลัย

พ.ศ. 2544 ปีเดียวกันนี้ ทีมวิจัยจากห้องปฏิบัติการวิจัยวิเคราะห์และสืบค้นความรู้จากฐานข้อมูลขนาดใหญ่ (Data Analysis and Knowledge Discovery Laboratory, DAKDL) มหาวิทยาลัยเกษตรศาสตร์ นำเสนองานวิจัยในวารสาร NECTEC Technical Journal ปีที่ 3 เล่มที่ 11 เรื่อง การใช้เทคนิคดาต้าไมน์นิ่งเพื่อพัฒนาคุณภาพการศึกษานิสิตคณะวิศวกรรมศาสตร์ (กฤษณะ ไวยมัย, ชิตชนก ส่งศิริ และธนาวิรินทร์ รักรัชมานนท์, 2544) โดย ผศ. ดร.กฤษณะ ไวยมัย (ตำแหน่งทางวิชาการขณะนั้น), ชิตชนก ส่งศิริ และ