Data Mining

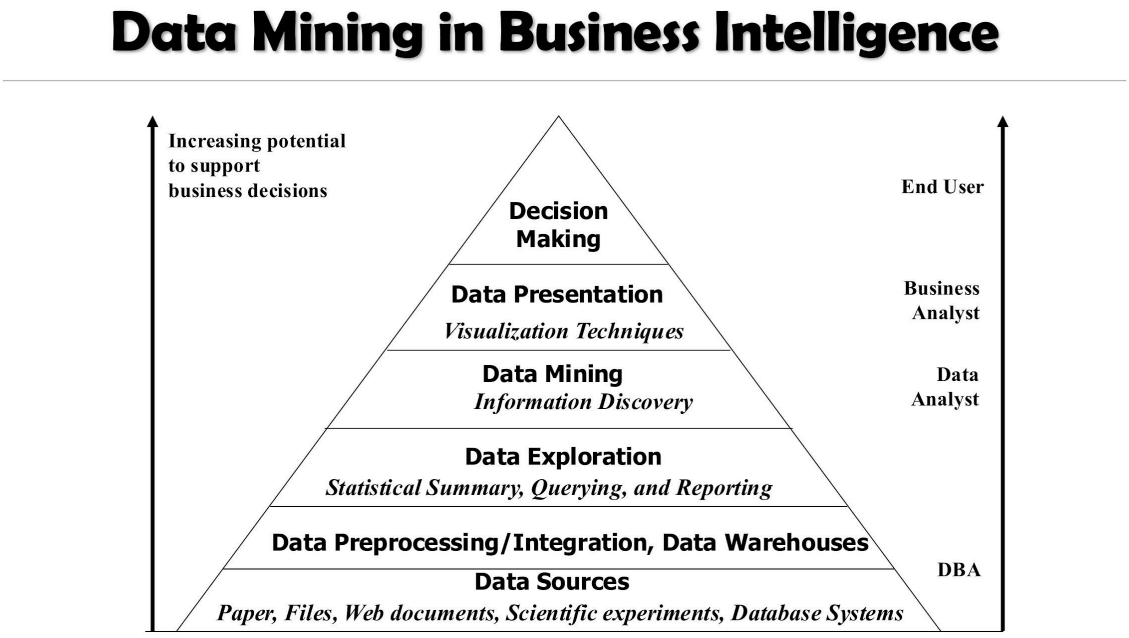
ทำไมต้องทำ Data Mining?

ต้องทำ Data Mining เพื่อให้คอมพิวเตอร์ช่วยอ่าน และ เข้าใจ ข้อมูลจำนวนมหาศาลแทนเรา เพื่อเปลี่ยนกองข้อมูลดิบให้เป็นข้อมูลเชิงทำนาย ที่ใช้ในการตัดสินใจได้อย่างแม่นยำและรวดเร็ว

- ในทางธุรกิจ: ช่วยให้องค์กรเข้าใจพฤติกรรมลูกค้าได้ดีขึ้น (เช่น การแบ่งกลุ่มลูกค้า, การแนะนำสินค้า) การคาดการณ์ยอดขาย และการลดความเสี่ยง
- ในทางวิทยาศาสตร์: ช่วยในการค้นพบใหม่ๆ (เช่น การจัดจำแนกยีน, การวิเคราะห์ภาพถ่ายดาวเทียม)

Data Mining คืออะไร?

Data Mining (การทำเหมืองข้อมูล) คือ กระบวนการสกัดรูปแบบ (patterns) หรือความรู้ที่มีประโยชน์ ไม่ธรรมดา และไม่เคยรู้มาก่อน ออกมาจากข้อมูลขนาดใหญ่จำนวนมหาศาล (Big Data) ด้วยระบบอัตโนมัติ Data Mining in Business Intelligence



1. ฐานราก: Data Sources (แหล่งข้อมูล)

- **คำอธิบาย:** เป็นรากฐานของระบบทั้งหมด ประกอบด้วยข้อมูลดิบที่หลากหลายรูปแบบ
- **ตัวอย่าง: Paper** (เอกสาร), **Files** (ไฟล์), **Web documents** (เอกสารเว็บ), **Scientific experiments** (การทดลองทางวิทยาศาสตร์), และ **Database Systems** (ระบบฐานข้อมูล)
- **ผู้ที่เกี่ยวข้อง:** โดยทั่วไปเกี่ยวข้องกับ **DBA** (Database Administrator - ผู้ดูแลระบบฐานข้อมูล) หรือผู้ที่จัดการข้อมูลต้นทาง

2. ชั้นที่สอง: Data Preprocessing/Integration, Data Warehouses (การเตรียมข้อมูล/การรวมข้อมูล, คลังข้อมูล)

- **คำอธิบาย:** เป็นขั้นตอนการเตรียมข้อมูลให้พร้อมสำหรับการวิเคราะห์ รวมถึงการทำความสะอาด (Cleaning), การแปลง (Transformation), และการรวมข้อมูลจากแหล่งต่างๆ เข้าด้วยกัน มักจะเก็บข้อมูลเหล่านี้ไว้ใน **Data Warehouses** (คลังข้อมูล)
- **ผู้ที่เกี่ยวข้อง:** ยังคงเกี่ยวข้องกับ **DBA** หรือวิศวกรข้อมูล (Data Engineers)

3. ชั้นที่สาม: Data Exploration (การสำรวจข้อมูล)

- **คำอธิบาย:** เป็นขั้นตอนการทำความเข้าใจภาพรวมของข้อมูล
- **เทคนิค:** ได้แก่ **Statistical Summary** (สรุปทางสถิติ), **Querying** (การสอบถามข้อมูล), และ **Reporting** (การรายงานผล)
- **ผู้ที่เกี่ยวข้อง:** เริ่มเกี่ยวข้องกับ **Data Analyst** (นักวิเคราะห์ข้อมูล)

4. ชั้นที่สี่: Data Mining (การทำเหมืองข้อมูล)

- **คำอธิบาย:** เป็นหัวใจสำคัญของ Business Intelligence ในระดับนี้ เป็นการใช้เทคนิคและอัลกอริทึมเพื่อค้นหารูปแบบ ความสัมพันธ์ และความรู้ที่ซ่อนอยู่ในข้อมูลขนาดใหญ่โดยอัตโนมัติ
- **วัตถุประสงค์: Information Discovery** (การค้นพบสารสนเทศ/ความรู้ใหม่)
- **ผู้ที่เกี่ยวข้อง: Data Analyst** (นักวิเคราะห์ข้อมูล) หรือ **Data Scientist** (นักวิทยาศาสตร์ข้อมูล)

5. ชั้นที่ห้า: Data Presentation (การนำเสนอข้อมูล)

- **คำอธิบาย:** เป็นการนำผลลัพธ์ที่ได้จากการทำเหมืองข้อมูลมาจัดรูปแบบให้เข้าใจง่ายและสื่อสารได้อย่างมีประสิทธิภาพ
- **เทคนิค:** เน้นที่ **Visualization Techniques** (เทคนิคการแสดงผลด้วยภาพ) เช่น การสร้างแผนภูมิ, แดชบอร์ด (Dashboards)
- **ผู้ที่เกี่ยวข้อง: Business Analyst** (นักวิเคราะห์ธุรกิจ) หรือผู้ที่ต้องสื่อสารข้อมูลให้ผู้บริหารรับทราบ

6. ยอดปิรามิด: Decision Making (การตัดสินใจ)

- **คำอธิบาย:** เป็นจุดสูงสุดและวัตถุประสงค์สุดท้ายของกระบวนการทั้งหมด ข้อมูลเชิงลึกที่ได้ถูกนำไปใช้ในการกำหนดกลยุทธ์, การวางแผน, และการดำเนินการทางธุรกิจ
- **เป้าหมาย:** การสนับสนุน **End User** (ผู้ใช้งานขั้นสุดท้าย) ซึ่งมักจะเป็น **ผู้บริหาร** หรือ **ผู้มีอำนาจตัดสินใจ** ในองค์กร

Data Mining vs. Data Exploration

Data Mining: มุ่งค้นหารูปแบบและความรู้ที่ซ่อนอยู่ในข้อมูล

Data Exploration / Business Intelligence: เน้น การสำรวจ, รายงาน, และนำเสนอข้อมูล มากกว่าการค้นหารูปแบบ

- ขึ้นอยู่กับ ประเภทข้อมูล, จุดประสงค์, และแอปพลิเคชัน

  ตัวอย่าง: ใน Supply Chain

  Data Mining ค้นหารูปแบบและแนวโน้ม

  OLAP วิเคราะห์ข้อมูลเชิงลึกแบบหลายมิติ

  Presentation Tools แสดงผลข้อมูลให้ผู้ใช้เข้าใจง่าย

Multi-Dimensional View of Data Mining

Data Mining ไม่ใช่แค่การใช้วิธีเดียวกับข้อมูลชนิดเดียว แต่เป็นการผสมผสานเครื่องมือและข้อมูลหลากหลายรูปแบบเพื่อวัตถุประสงค์ที่แตกต่างกัน

1. ข้อมูลที่ต้องถูกทำเหมือง (วัตถุดิบ)

นี่คือ แหล่งข้อมูล ที่เรานำมาขุดค้น ซึ่งมีความหลากหลายและซับซ้อนมาก:

- ข้อมูลดั้งเดิม: ฐานข้อมูลทั่วไป (Relational) และคลังข้อมูล (Data Warehouse) ที่เป็นโครงสร้าง
- ข้อมูลเฉพาะทาง:
  - ธุรกรรม: ข้อมูลการซื้อขายแบบรายการ (เช่น ใบเสร็จ)

- เชิงเวลา: ข้อมูลที่มาเป็นชุดตามลำดับเวลา (Time-series) หรือสตรีมข้อมูลที่ไหลเข้ามาเรื่อย ๆ (Stream Data)
- ไม่เป็นโครงสร้าง/กึ่งโครงสร้าง: ข้อความ, เว็บไซต์, มัลติมีเดีย (รูปภาพ, วิดีโอ)
- ความสัมพันธ์: กราฟ, เครือข่ายสังคม (Social Networks)

2. ความรู้ที่จะถูกสกัดออกมา (เป้าหมาย)

นี่คือ วัตถุประสงค์ หรือ ชนิดของความรู้ ที่เราต้องการค้นหาจากข้อมูล ซึ่งแบ่งได้เป็น 2 ประเภทหลัก:

- Descriptive (บรรยาย): อธิบายว่า "เกิดอะไรขึ้น" ในอดีต
  - Characterization: สรุปลักษณะของกลุ่มข้อมูล (เช่น ลูกค้ากลุ่ม VIP มีลักษณะอย่างไร)
  - Association: ค้นหาความสัมพันธ์ร่วมกัน (เช่น คนที่ซื้อขนมปัง มักจะซื้อเนย)
  - Clustering: จัดกลุ่มข้อมูลที่มีคุณสมบัติคล้ายกันเข้าด้วยกันโดยไม่ได้กำหนดกลุ่มไว้ก่อน (เช่น จัดกลุ่มลูกค้าตามพฤติกรรม)
- Predictive (ทำนาย): ทำนายว่า "จะเกิดอะไรขึ้น" ในอนาคต
  - Classification: จัดหมวดหมู่ จำแนกข้อมูลใหม่ให้อยู่ในกลุ่มที่กำหนดไว้แล้ว (เช่น การทำนายว่าอีเมลนี้เป็น Spam หรือไม่)
  - Trend / Deviation Analysis: ค้นหาแนวโน้มในอนาคต หรือการเบี่ยงเบนจากแนวโน้มปกติ
  - Outlier Analysis: ค้นหาค่าที่ผิดปกติหรือเหตุการณ์ที่แตกต่างจากข้อมูลส่วนใหญ่ (ใช้ในการตรวจจับการทุจริต/Fraud)

3. เทคนิคที่ใช้ (เครื่องมือ)

นี่คือ วิธีการทางเทคนิค ที่ใช้ในการสกัดความรู้ ประกอบด้วยสาขาวิชาต่าง ๆ:

- Machine Learning (ML): อัลกอริทึมที่ทำให้คอมพิวเตอร์เรียนรู้จากข้อมูลได้เอง
- สถิติ (Statistics): หลักการทางคณิตศาสตร์ที่ใช้ในการวิเคราะห์ข้อมูลและการทำนาย
- การจัดการข้อมูล: การใช้เทคนิค Data Warehouse / OLAP และการประมวลผลประสิทธิภาพสูง (High-performance computing) เพื่อจัดการข้อมูลจำนวนมหาศาล
- การแสดงผล (Visualization): การนำเสนอรูปแบบที่ค้นพบให้อยู่ในรูปภาพที่เข้าใจง่าย

4. การประยุกต์ใช้งาน (การนำไปใช้จริง)

นี่คือ อุตสาหกรรม ที่นำ Data Mining ไปใช้เพื่อสร้างมูลค่าเพิ่ม:

- ธุรกิจ: ค้าปลีก, โทรคมนาคม, ธนาคาร (เช่น การตลาดแบบตรง, การรักษาลูกค้า, การวิเคราะห์ความเสี่ยง)

- การเงิน: การวิเคราะห์หุ้นและการตรวจจับการทุจริต (Fraud Analysis)
- วิทยาศาสตร์: Bio-data mining (เช่น การวิเคราะห์ลำดับพันธุกรรม), Text/Web Mining

Data Mining Functions

1. Generalization (การทำให้เป็นทั่วไป)

- สรุปและทำให้ข้อมูลจำนวนมากเข้าใจง่าย
- ขั้นตอน: ทำความสะอาด, แปลง, รวมข้อมูล, สร้างคลังข้อมูลหลายมิติ (Data Warehouse & Data Cube)
- ใช้ OLAP และการอธิบายลักษณะข้อมูลหลายมิติ (Characterization & Discrimination)
- ตัวอย่าง: เปรียบเทียบ ภูมิภาคแห้ง vs. ชื้น

2. Pattern Discovery (ค้นหารูปแบบ)

- ค้นหารูปแบบหรือสินค้าที่มักปรากฏพร้อมกัน (Frequent Itemsets)
- วิเคราะห์ ความสัมพันธ์และความเชื่อมโยง (Association & Correlation)
- ตัวอย่าง: Diaper ⟶ Beer [Support 0.5%, Confidence 75%]
- ใช้ในการจัดประเภท (Classification), จัดกลุ่ม (Clustering) และแอปอื่น ๆ

3. Classification (การจัดประเภท)

- สร้างโมเดลจากข้อมูลตัวอย่างเพื่อทำนายคลาสหรือ Label
- ตัวอย่าง: แบ่งประเทศตาม ภูมิอากาศ, แบ่งรถตาม อัตราการใช้น้ำมัน
- วิธีการ: Decision Trees, Naïve Bayes, SVM, Neural Networks, Rule-based, Logistic Regression ฯลฯ
- ใช้งาน: ตรวจจับการทุจริตบัตรเครดิต, Direct Marketing, จัดประเภทดาว, โรค, เว็บเพจ

4. Cluster Analysis (การจัดกลุ่ม)

- Unsupervised Learning – ไม่มี Label ล่วงหน้า
- จัดกลุ่มข้อมูลเป็นคลัสเตอร์ใหม่
- หลักการ: เพิ่มความคล้ายภายในกลุ่ม & ลดความคล้ายระหว่างกลุ่ม
- ตัวอย่าง: จัดกลุ่มบ้านเพื่อวิเคราะห์รูปแบบการกระจาย

5. Outlier Analysis (วิเคราะห์ค่าผิดปกติ)

- หาค่าที่ไม่สอดคล้องกับพฤติกรรมทั่วไปของข้อมูล
- อาจเป็น Noise หรือ Exception
- ใช้วิธีจาก Clustering หรือ Regression

- ประโยชน์: ตรวจจับการทุจริต, วิเคราะห์เหตุการณ์หายาก

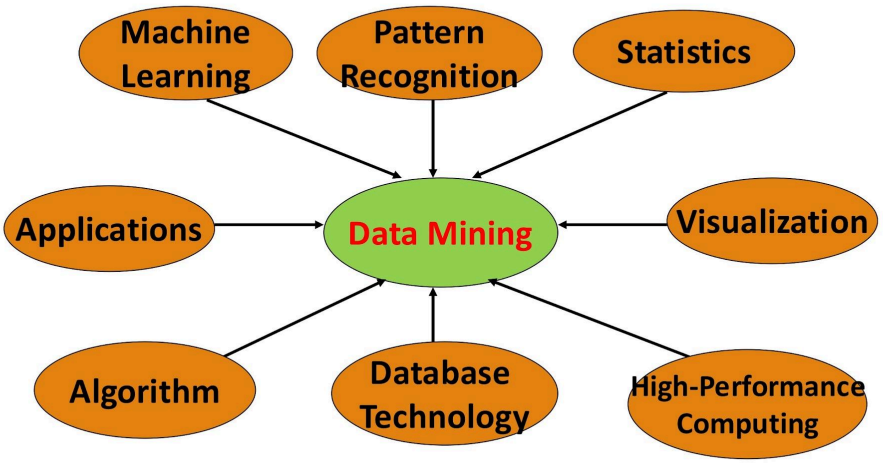
6. Time & Ordering: Sequential Pattern, Trend & Evolution Analysis

- วิเคราะห์ ลำดับเหตุการณ์, แนวโน้ม, และวิวัฒนาการของข้อมูล
- เช่น Trend Analysis, Time-series, Deviation, Regression, Value Prediction
- Sequential Pattern: ตัวอย่าง ซื้อกล้อง ⟶ ซื้อเมมโมรี่การ์ด
- Periodicity, Motifs (รวม biological sequences), Similarity-based Analysis
- Mining Data Streams: ข้อมูลเรียงลำดับ, เปลี่ยนตามเวลา, อาจไม่มีที่สิ้นสุด

7. Structure & Network Analysis (วิเคราะห์โครงสร้างและเครือข่าย)

- Graph Mining: หากราฟย่อยที่เกิดบ่อย เช่น สารเคมี, XML, เว็บ
- Information & Social Network Analysis: วิเคราะห์ Nodes และ Edges เช่น เครือข่ายผู้เขียน, เครือข่ายผู้ก่อการร้าย
- รองรับหลายเครือข่ายผสม (Heterogeneous Networks)
- Link Mining: วิเคราะห์ความสัมพันธ์เชิงความหมาย
- Web Mining: เว็บเป็นเครือข่ายข้อมูลขนาดใหญ่ เช่น PageRank, Community Discovery, Opinion Mining, Usage Mining

Data Mining: Confluence of Multiple Disciplines



28

การประยุกต์ใช้งานของ Data Mining)

- วิเคราะห์ เว็บเพจ: จัดประเภท, จัดกลุ่ม, จัดอันดับ
- ระบบ แนะนำและวิเคราะห์ร่วมกัน (Recommender & Collaborative Systems)
- วิเคราะห์ ข้อมูลตะกร้าสินค้า เพื่อทำการตลาดแบบเจาะจง
- วิเคราะห์ ข้อมูลชีวภาพและการแพทย์
- ใช้กับ วิศวกรรมซอฟต์แวร์ และ วิเคราะห์ข้อความ (Text Analysis)
- วิเคราะห์ เครือข่ายสังคมและข้อมูล

- ฟังก์ชันฝังในระบบใหญ่ เช่น Google, Microsoft, Yahoo!, LinkedIn, Facebook

- เครื่องมือเฉพาะทาง เช่น SAS, MS SQL Server Analysis Manager, Oracle Data Mining Tools

สรุป Data Mining

- Data Mining คือ การค้นหารูปแบบและความรู้ที่น่าสนใจจากข้อมูลจำนวนมหาศาล

- เป็น วิวัฒนาการตามธรรมชาติของวิทยาศาสตร์และเทคโนโลยีสารสนเทศ ที่มีความต้องการสูง และมีการประยุกต์ใช้งานอย่างกว้างขวาง

- กระบวนการ KDD (Knowledge Discovery in Databases) ประกอบด้วย:

    - ทำความสะอาดข้อมูล (Data Cleaning)

    - รวมข้อมูล (Data Integration)

    - เลือกข้อมูล (Data Selection)

    - แปลงข้อมูล (Transformation)

    - ทำเหมืองข้อมูล (Data Mining)

    - ประเมินรูปแบบ (Pattern Evaluation)

    - นำเสนอความรู้ (Knowledge Presentation)

- การทำเหมืองสามารถทำได้กับ ข้อมูลหลายประเภท

- ฟังก์ชันหลักของ Data Mining:

    - Characterization (สรุปลักษณะ)

    - Discrimination (แยกแยะ)

    - Association (วิเคราะห์ความสัมพันธ์)

    - Classification (จัดประเภท)

    - Clustering (จัดกลุ่ม)

    - Trend & Outlier Analysis (วิเคราะห์แนวโน้มและค่าผิดปกติ)

- มี เทคโนโลยีและการประยุกต์ใช้งาน มากมาย

- มี ประเด็นสำคัญที่ต้องพิจารณา เช่น คุณภาพข้อมูล, ประสิทธิภาพ, ความหลากหลายของข้อมูล, ผลกระทบต่อสังคม